

Advanced R

Hadley Wickham

Contents

Welcome	9
Other books	9
Preface	11
rlang	11
Foundations	11
Programming paradigms	12
Techniques	12
Removals	12
1 Introduction	13
1.1 Who should read this book	14
1.2 Related work	15
1.3 What you will get out of this book	15
1.4 Meta-techniques	15
1.5 Recommended reading	16
1.6 Getting help	16
1.7 Acknowledgments	16
1.8 Conventions	18
1.9 Colophon	19
I Foundations	23
Introduction	25
2 Names and values	27
2.1 Introduction	27
2.2 Binding basics	28
2.3 Copy-on-modify	30
2.4 Object size	35
2.5 Modify-in-place	36
2.6 Unbinding and the garbage collector	39
2.7 Answers	41
3 Vectors	43
3.1 Introduction	43
3.2 Atomic vectors	44
3.3 Attributes	46
3.4 S3 atomic vectors	49
3.5 Lists	53
3.6 Data frames and tibbles	55

3.7	NULL	63
3.8	Answers	64
4	Subsetting	65
4.1	Introduction	65
4.2	Selecting multiple elements	66
4.3	Selecting a single element	71
4.4	Subsetting and assignment	74
4.5	Applications	75
4.6	Answers	81
5	Functions	83
5.1	Introduction	83
5.2	Function fundamentals	84
5.3	Function composition	87
5.4	Lexical scoping	88
5.5	Lazy evaluation	92
5.6	... (dot-dot-dot)	97
5.7	Exiting a function	99
5.8	Function forms	103
5.9	Invoking a function	107
5.10	Quiz answers	108
6	Environments	109
6.1	Introduction	109
6.2	Environment basics	110
6.3	Recursing over environments	116
6.4	Special environments	117
6.5	The call stack	125
6.6	As data structures	128
6.7	<<-	129
6.8	Quiz answers	129
7	Conditions	131
7.1	Introduction	131
7.2	Signalling conditions	132
7.3	Ignoring conditions	136
7.4	Handling conditions	137
7.5	Custom conditions	143
7.6	Applications	146
7.7	Quiz answers	153
8	Connections	155
8.1	Basics	155
8.2	Reading and writing binary data	155
8.3	Reading and writing text data	155
II	Functional programming	157
Introduction		159
Functional programming languages		159
Functional style		160
9	Functionals	161

9.1	Introduction	161
9.2	My first functional: <code>map()</code>	162
9.3	Purrr style	169
9.4	Map variants	170
9.5	Reduce	178
9.6	Predicate functionals	183
9.7	Base functionals	184
10	Function factories	189
10.1	Introduction	189
10.2	Factory fundamentals	190
10.3	Graphical factories	195
10.4	Statistical factories	199
10.5	Function factories + functionals	204
11	Function operators	207
11.1	Introduction	207
11.2	Existing FOs	208
11.3	Case study: creating your own FOs	212
III	Object oriented programming	215
Introduction		217
11.4	OOP Systems	217
11.5	OOP in R	218
11.6	Field guide	218
12	Base types	221
12.1	Introduction	221
12.2	Base objects vs OO objects	221
12.3	Base types	222
12.4	The <code>is</code> functions	223
13	S3	225
13.1	Introduction	225
13.2	Basics	225
13.3	Classes	227
13.4	Generics and methods	233
13.5	Method dispatch	236
13.6	Inheritance	238
13.7	Dispatch details	241
14	S4	247
14.1	Introduction	247
14.2	Classes	248
14.3	Generics and methods	252
14.4	Method dispatch	255
14.5	S4 and existing code	261
15	R6	263
15.1	Introduction	263
15.2	Classes and methods	264
15.3	Controlling access	268
15.4	Reference semantics	271

16 Trade-offs	275
16.1 Introduction	275
16.2 S4 vs S3	275
16.3 R6 vs S3	276
IV Metaprogramming	281
Introduction	283
16.4 Domain specific languages	284
16.5 Overview	284
17 Expressions	287
17.1 Introduction	287
17.2 Abstract syntax trees	288
17.3 R's grammar	291
17.4 Data structures	294
17.5 Parsing and deparsing	299
17.6 Case study: Walking the AST with recursive functions	300
18 Quasiquotation	307
18.1 Introduction	307
18.2 Motivation	307
18.3 Quotation	310
18.4 Evaluation	313
18.5 Unquotation	313
18.6 Case studies	319
18.7 Dot-dot-dot (...)	324
19 Evaluation	329
19.1 Introduction	329
19.2 Evaluation basics	330
19.3 Quosures	334
19.4 Tidy evaluation	339
19.5 Wrapping quoting functions	346
20 Translating R code	353
20.1 Introduction	353
20.2 HTML	354
20.3 LaTeX	360
V Techniques	367
21 Debugging	369
21.1 Introduction	369
21.2 Techniques	369
21.3 Tools	370
22 Performance	377
22.1 Why is R slow?	377
22.2 Microbenchmarking	378
22.3 Language performance	379
22.4 Implementation performance	383
22.5 Alternative R implementations	385

23 Optimising code	389
23.1 Introduction	389
23.2 Measuring performance	390
23.3 Memory profiling with lineprof	393
23.4 Improving performance	395
23.5 Code organisation	396
23.6 Has someone already solved the problem?	397
23.7 Do as little as possible	397
23.8 Vectorise	402
23.9 Avoid copies	404
23.10 Byte code compilation	405
23.11 Case study: t-test	405
23.12 Parallelise	407
23.13 Other techniques	409
24 Rewriting R code in C++	411
24.1 Introduction	411
24.2 Getting started with C++	412
24.3 Attributes and other classes	419
24.4 Missing values	421
24.5 Rcpp sugar	424
24.6 The STL	426
24.7 Case studies	430
24.8 Using Rcpp in a package	433
24.9 Learning more	434
24.10 Acknowledgments	435
References	437

Welcome

This is the website for work-in-progress 2nd edition of “**Advanced R** (<http://amzn.com/1466586966?tag=devtools-20>)”, a book in Chapman & Hall’s R Series. The book is designed primarily for R users who want to improve their programming skills and understanding of the language. It should also be useful for programmers coming to R from other languages, as it explains some of R’s quirks and shows how some parts that seem horrible do have a positive side.

This edition is a work in progress. If you’re looking for the electronic version of the 1st edition, you can find it online at <http://adv-r.had.co.nz/>.

Other books

You may also be interested in:

- “**R for Data Science** (<http://r4ds.had.co.nz/>)” which introduces you to R as a tool for doing data science, focussing on a consistent set of packages known as the tidyverse.
- “**R Packages** (<http://r-pkgs.had.co.nz/>)” which teaches you how to make the most of R’s fantastic package system.

Preface

Welcome to the work-in-progress 2nd edition of **Advanced R**. This preface describes the major changes that I have made to the book.

The 2nd edition has been published in colour, which as well as improving the syntax highlighting of the code chunks, has considerably increased the scope for helpful diagrams. I have taken advantage of this and included many more diagrams throughout the book.

rlang

A big change since the first edition of the book is the creation of the rlang (<http://rlang.r-lib.org>) package, written primarily by Lionel Henry. The goal of this package is to provide a clean interface to low-level data structures and operations. I use this package in favour of base R because I believe it makes easier to understand how the R language works. Instead of struggling with the incidentals of functions that evolved organically over many years, the more consistent rlang API makes it easier to focus on the big ideas.

In each section, I'll briefly outline the base R equivalents to rlang code. But if you want to see the purest base R expression of these ideas, I recommend reading the first edition of the book, which you can find online at <http://adv-r.had.co.nz>.

Overall, rlang is still a work in progress, and much of the API continues to mature. However, the code used in this book is part of the rlang's testing process and will continue to work in the future. You can also see our confidence in the stability of rlang functions with the lifecycle badges at the documentation.

Foundations

- Environments: more pictures. Much improved discussion of frames and how they relate to the call stack.
- New chapter on “Names and values” that helps you form a better mental model of `<-`, and to better understand when R makes copies of existing data structures. Understanding the distinction between names and values is important for functional programming, and understanding when R makes copies is critical for accurate performance predictions.
- Vectors (previously data structures) has been rewritten with more diagrams to focus on vector types. More information about other important S3 vectors, and information about tibbles, a modern re-imagining of data frames.
- Exceptions and debugging has been split into two chapters, “debugging” and “conditions”. The contents of conditions has been expanded. The section on defensive programming has been removed, because discussing type stability is more natural in the context of functional programming, and programming with NSE (non-standard evaluation) is not the challenge it once was (now that tidy evaluation exists).

Programming paradigms

After foundations, the book is now organised around the three most important programming paradigms in R:

- Functional programming has been updated to focus on the tools provided by the purrr package. The greater consistency in the purrr package makes it possible to focus more on the underlying ideas without being distracted by incidental details. Divided more cleanly into functionals, function factories, and function operators. Greater focus on what time has shown to be important in practice. Less math + stat, and more data science. Focus on composition with the pipe, rather than more esoteric forms.
- Object oriented programming (OOP) now forms a major section of the book with individual chapters on base types, S3, S4, R6, and the tradeoffs between the systems.
- Metaprogramming, formerly computing on the language, describes the suite of tools that you can use to generate code with code. Compared to the first edition it has been substantially expanded (from three chapters to five) and reorganised. More diagrams.

Techniques

Final section discusses programming techniques, including both debugging, profiling, improving performance, and connecting R and C++.

Removals

- Chapter on base R vocabulary was removed.
- The style guide has moved to <http://style.tidyverse.org/>. It is now paired with the styler (<http://styler.r-lib.org/>) package which can automatically apply many of the rules.
- R's C interface moving to the work-in-progress <https://github.com/hadley/r-internals>
- Memory chapter either integrated in names and values, or removed because it's excessively technical and not that important to understand (unless you're working with C code in which case it belongs in internals).

Chapter 1

Introduction

With more than 10 years experience programming in R, I've had the luxury of being able to spend a lot of time trying to figure out and understand how the language works. This book is my attempt to pass on what I've learned so that you can quickly become an effective R programmer. Reading it will help you avoid the mistakes I've made and dead ends I've gone down, and will teach you useful tools, techniques, and idioms that can help you to attack many types of problems. In the process, I hope to show that, despite its frustrating quirks, R is, at its heart, an elegant and beautiful language, well tailored for data analysis and statistics.

If you are new to R, you might wonder what makes learning such a quirky language worthwhile. To me, some of the best features are:

- It's free, open source, and available on every major platform. As a result, if you do your analysis in R, anyone can easily replicate it.
- A massive set of packages for statistical modelling, machine learning, visualisation, and importing and manipulating data. Whatever model or graphic you're trying to do, chances are that someone has already tried to do it. At a minimum, you can learn from their efforts.
- Cutting edge tools. Researchers in statistics and machine learning will often publish an R package to accompany their articles. This means immediate access to the very latest statistical techniques and implementations.
- Deep-seated language support for data analysis. This includes features like missing values, data frames, and subsetting.
- A fantastic community. It is easy to get help from experts on the R-help mailing list (<https://stat.ethz.ch/mailman/listinfo/r-help>), stackoverflow (<http://stackoverflow.com/questions/tagged/r>), RStudio Community (<https://community.rstudio.com/>), or subject-specific mailing lists like R-SIG-mixed-models (<https://stat.ethz.ch/mailman/listinfo/r-sig-mixed-models>) or ggplot2 (<https://groups.google.com/forum/#!forum/ggplot2>). You can also connect with other R learners via twitter (<https://twitter.com/search?q=%23rstats>), linkedin (<http://www.linkedin.com/groups/R-Project-Statistical-Computing-77616>), and through many local user groups (<https://jumpingrivers.github.io/meetingsR/>).
- Powerful tools for communicating your results. R packages make it easy to produce HTML or PDF reports (<http://yihui.name/knitr/>), or create interactive websites (<http://www.rstudio.com/shiny/>).
- A strong foundation in functional programming. The ideas of functional programming are well suited to solving many of the challenges of data analysis. R provides a powerful and flexible toolkit which allows you to write concise yet descriptive code.

- An IDE (<http://www.rstudio.com/ide/>) tailored to the needs of interactive data analysis and statistical programming.
- Powerful metaprogramming facilities. R is not just a programming language, it is also an environment for interactive data analysis. Its metaprogramming capabilities allow you to write magically succinct and concise functions and provide an excellent environment for designing domain-specific languages.
- Designed to connect to high-performance programming languages like C, Fortran, and C++.

Of course, R is not perfect. R's biggest challenge is that most R users are not programmers. This means that:

- Much of the R code you'll see in the wild is written in haste to solve a pressing problem. As a result, code is not very elegant, fast, or easy to understand. Most users do not revise their code to address these shortcomings.
- Compared to other programming languages, the R community tends to be more focussed on results instead of processes. Knowledge of software engineering best practices is patchy: for instance, not enough R programmers use source code control or automated testing.
- Metaprogramming is a double-edged sword. Too many R functions use tricks to reduce the amount of typing at the cost of making code that is hard to understand and that can fail in unexpected ways.
- Inconsistency is rife across contributed packages, even within base R. You are confronted with over 20 years of evolution every time you use R. Learning R can be tough because there are many special cases to remember.
- R is not a particularly fast programming language, and poorly written R code can be terribly slow. R is also a profligate user of memory.

Personally, I think these challenges create a great opportunity for experienced programmers to have a profound positive impact on R and the R community. R users do care about writing high quality code, particularly for reproducible research, but they don't yet have the skills to do so. I hope this book will not only help more R users to become R programmers but also encourage programmers from other languages to contribute to R.

1.1 Who should read this book

This book is aimed at two complementary audiences:

- Intermediate R programmers who want to dive deeper into R and learn new strategies for solving diverse problems.
- Programmers from other languages who are learning R and want to understand why R works the way it does.

To get the most out of this book, you'll need to have written a decent amount of code in R or another programming language. You might not know all the details, but you should be familiar with how functions work in R and although you may currently struggle to use them effectively, you should be familiar with the `apply` family (like `apply()` and `lapply()`).

This book walks the narrow line between being a reference book (primarily used for lookup), and being linearly readable. This involves some tradeoffs, because it's difficult to linearise material while still keeping related materials together, whereas some concepts are much easier to explain if you're already familiar with specific technically vocabulary. I've tried to use footnotes and cross-references to make sure you can still make sense even if you just dip your toes in the occasional chapter.

1.2 Related work

Tidyverse + R4DS

R packages

1.3 What you will get out of this book

This book describes the skills I think an advanced R programmer should have: the ability to produce quality code that can be used in a wide variety of circumstances.

After reading this book, you will:

- Be familiar with the fundamentals of R. You will understand complex data types and the best ways to perform operations on them. You will have a deep understanding of how functions work, and be able to recognise and use the four object systems in R.
- Understand what functional programming means, and why it is a useful tool for data analysis. You'll be able to quickly learn how to use existing tools, and have the knowledge to create your own functional tools when needed.
- Appreciate the double-edged sword of metaprogramming. You'll be able to create functions that use non-standard evaluation in a principled way, saving typing and creating elegant code to express important operations. You'll also understand the dangers of metaprogramming and why you should be careful about its use.
- Have a good intuition for which operations in R are slow or use a lot of memory. You'll know how to use profiling to pinpoint performance bottlenecks, and you'll know enough C++ to convert slow R functions to fast C++ equivalents.
- Be comfortable reading and understanding the majority of R code. You'll recognise common idioms (even if you wouldn't use them yourself) and be able to critique others' code.

1.4 Meta-techniques

There are two meta-techniques that are tremendously helpful for improving your skills as an R programmer: reading source code and adopting a scientific mindset.

Reading source code is important because it will help you write better code. A great place to start developing this skill is to look at the source code of the functions and packages you use most often. You'll find things that are worth emulating in your own code and you'll develop a sense of taste for what makes good R code. You will also see things that you don't like, either because its virtues are not obvious or it offends your sensibilities. Such code is nonetheless valuable, because it helps make concrete your opinions on good and bad code.

A scientific mindset is extremely helpful when learning R. If you don't understand how something works, develop a hypothesis, design some experiments, run them, and record the results. This exercise is extremely useful since if you can't figure something out and need to get help, you can easily show others what you tried. Also, when you learn the right answer, you'll be mentally prepared to update your world view. When I clearly describe a problem to someone else (the art of creating a reproducible example (<http://stackoverflow.com/questions/5963269>)), I often figure out the solution myself.

1.5 Recommended reading

R is still a relatively young language, and the resources to help you understand it are still maturing. In my personal journey to understand R, I've found it particularly helpful to use resources from other programming languages. R has aspects of both functional and object-oriented (OO) programming languages. Learning how these concepts are expressed in R will help you leverage your existing knowledge of other programming languages, and will help you identify areas where you can improve.

To understand why R's object systems work the way they do, I found *The Structure and Interpretation of Computer Programs* (<https://mitpress.mit.edu/sites/default/files/sicp/full-text/book/book.html>) (SICP) by Harold Abelson and Gerald Jay Sussman, particularly helpful. It's a concise but deep book. After reading it, I felt for the first time that I could actually design my own object-oriented system. The book was my first introduction to the generic function style of OO common in R. It helped me understand its strengths and weaknesses. SICP also talks a lot about functional programming, and how to create simple functions which become powerful when combined.

To understand the trade-offs that R has made compared to other programming languages, I found *Concepts, Techniques and Models of Computer Programming* (<http://amzn.com/0262220695?tag=devtools-20>) by Peter van Roy and Sef Haridi extremely helpful. It helped me understand that R's copy-on-modify semantics make it substantially easier to reason about code, and that while its current implementation is not particularly efficient, it is a solvable problem.

If you want to learn to be a better programmer, there's no place better to turn than *The Pragmatic Programmer* (<http://amzn.com/020161622X?tag=devtools-20>) by Andrew Hunt and David Thomas. This book is language agnostic, and provides great advice for how to be a better programmer.

1.6 Getting help

Currently, there are three main venues to get help when you're stuck and can't figure out what's causing the problem: RStudio Community (<https://community.rstudio.com/>), stackoverflow (<http://stackoverflow.com>) and the R-help mailing list (<https://stat.ethz.ch/mailman/listinfo/r-help>). You can get fantastic help in each venue, but they do have their own cultures and expectations. It's usually a good idea to spend a little time lurking, learning about community expectations, before you put up your first post.

Some good general advice:

- Make sure you have the latest version of R and of the package (or packages) you are having problems with. It may be that your problem is the result of a recently fixed bug.
- Spend some time creating a reproducible example (<http://stackoverflow.com/questions/5963269>). This is often a useful process in its own right, because in the course of making the problem reproducible you often figure out what's causing the problem. The `reprex` (<https://reprex.tidyverse.org/>) package can help you create a **reproducible example** that can easily be run by people trying to help you. There are several resources available (<https://community.rstudio.com/t/faq-whats-a-reproducible-example-reprex-and-how-do-i-do-one/5219>) to help you create a successful `reprex`.
- Look for related problems before posting. If someone has already asked your question and it has been answered, it's much faster for everyone if you use the existing answer.

1.7 Acknowledgments

I would like to thank the tireless contributors to R-help and, more recently, stackoverflow (<http://stackoverflow.com/questions/tagged/r>). There are too many to name individually, but I'd particularly

like to thank Luke Tierney, John Chambers, Dirk Eddelbuettel, JJ Allaire and Brian Ripley for generously giving their time and correcting my countless misunderstandings.

This book was written in the open (<https://github.com/hadley/adv-r/>), and chapters were advertised on twitter (<https://twitter.com/hadleywickham>) when complete. It is truly a community effort: many people read drafts, fixed typos, suggested improvements, and contributed content. Without those contributors, the book wouldn't be nearly as good as it is, and I'm deeply grateful for their help. Special thanks go to Peter Li, who read the book from cover-to-cover and provided many fixes. Other outstanding contributors were Aaron Schumacher, @crtahlin, Lingbing Feng, @juancentro, and @johnbaums.

Thanks go to all contributers in alphabetical order: Aaron Wolen (@aaronwolen), @absolutelyNoWarranty, Adam Hunt (@adamphunt), @agrabovsky, Alexander Grueneberg (@agrueneberg), Anthony Damico (@ajdamico), James Manton (@ajdm), Aaron Schumacher (@ajschumacher), Alan Dipert (@alandipert), Alex Brown (@alexbrown), @alexpperrone, Alex Whitworth (@alexWhitworth), Alexandros Kokkalis (@alko989), @amarchin, Amelia McNamara (@AmeliaMN), Bryce Mecum (@amoeba), Andrew Laucius (@andrewla), Andrew Bray (@andrewpbray), Andrie de Vries (@andrie), @aranlunzer, Ari Lamstein (@arilamstein), @asnR, Andy Teucher (@ateucher), Albert Vilella (@avilella), baptiste (@baptiste), Brian G. Barkley (@BarkleyBG), Mara Averick (@batpigandme), Barbara Borges Ribeiro (@bborgesr), Brandon Greenwell (@bgreenwell), Brandon Hurr (@bhive01), Jason Knight (@binarybana), Brett Klamer (@bklamer), Jesse Anderson (@blindjesse), Brian Mayer (@blmayer), Benjamin L. Moore (@blmoore), Brian Diggs (@BrianDiggs), Brian S. Yandell (@byandell), @carey1024, Chip Hogg (@chiphogg), Chris Muir (@ChrisMuir), Christopher Gandrud (@christophergandrud), Clay Ford (@clayford), Colin Fay (@ColinFay), @cortinah, Cameron Plouffe (@cplouffe), Carson Sievert (@cpsievert), Craig Citro (@craigcitro), Craig Grabowski (@craiggrabowski), Christopher Roach (@croach), Peter Meilstrup (@crowding), Crt Ahlin (@crtahlin), Carlos Scheidegger (@cscheid), Colin Gillespie (@csgillespie), Christopher Brown (@ctbrown), Davor Cubranic (@cubranic), Darren Cusanovich (@cusanovich), Christian G. Warden (@cwarden), Charlotte Wickham (@cwickham), Dean Attali (@daattali), Dan Sullivan (@dan87134), Daniel Barnett (@danielbarnett), Kenny Darrell (@darrkj), Tracy Nance (@datapixie), Dave Childers (@davechilders), David Rubinger (@davidrubinger), David Chudzicki (@dchudz), Daisuke ICHIKAWA (@dichika), david kahle (@dkahle), David LeBauer (@dlebauer), David Schweizer (@dlschweizer), David Montaner (@dmontaner), Zhuoer Dong (@dongzhuoer), Doug Mitarotonda (@dougmitarotonda), Jonathan Hill (@Dripdrop12), Julian During (@duju211), @duncanwadsworth, @eaurele, Dirk Eddelbuettel (@eddelbuettel), @EdFineOKL, Edwin Thoen (@EdwinTh), Ethan Heinzen (@eheinzen), @eijoac, Joel Schwartz (@eipi10), Eric Ronald Legrand (@elegrand), Ellis Valentiner (@ellisvalentiner), Emil Hvitfeldt (@EmilHvitfeldt), Emil Rehnberg (@EmilRehnberg), Daniel Lee (@erget), Eric C. Anderson (@eriqande), Enrico Spinielli (@espinelli), @etb, David Hajage (@eusebe), Fabian Scheipl (@fabian-s), @flammy0530, François Michonneau (@fmichonneau), Francois Pepin (@fpepin), Frank Farach (@frankfarach), @freezby, Frans van Dunné (@FvD), @fyears, @gagnagaman, Garrett Grolemund (@garrettgman), Gavin Simpson (@gavinsimpson), @gezakiss7, @gggtest, Gökçen Eraslan (@gokceneraslan), Georg Russ (@gr650), @grasshoppermouse, Gregor Thomas (@gregorp), Garrett See (@gsee), Ari Friedman (@gsk3), Gunnlaugur Thor Briem (@gthb), Hadley Wickham (@hadley), Hamed (@hamedbh), Harley Day (@harleyday), @hassaad85, @helmingstay, Henning (@henningsway), Henrik Bengtsson (@HenrikBengtsson), Ching Boon (@hoscb), Iain Dillingham (@iaindillingham), @IanKopacka, Ian Lytle (@ijlyttle), Ilan Man (@ilanman), Imanuel Costigan (@imanuelcostigan), Thomas Bürli (@initdch), Os Keyes (@Ironholds), @irudnyts, i (@isomorphisms), Irene Steves (@isteves), Jan Gleixner (@jan-glx), Jason Asher (@jasonasher), Jason Davies (@jasondavies), Chris (@jastingo), jcborras (@jcborras), John Blischak (@jdbblischak), @jeharmse, Lukas Burk (@jemu42), Jennifer (Jenny) Bryan (@jennybc), Justin Jent (@jentjr), Jeston (@JestonBlu), Jim Hester (@jimhester), @JimInNashville, @jimmyliu2017, Jim Vine (@jimvine), Jinlong Yang (@jinlong25), J.J. Allaire (@jjallaire), @JMHay, Jochen Van de Velde (@jochenvdv), Johann Hibschman (@johannh), John Baumgartner (@johnbaums), John Horton (@johnjosephhorton), @johnthomas12, Jon Calder (@jonmcalder), Jon Harmon (@jonthegeek), Julia Gustavsen (@jooolia), JorneBicler (@JorneBicler), Jeffrey Arnold (@jrnold), Joyce Robbins (@jtr13), Juan Manuel Truppia (@juancentro), Kevin Markham (@justmarkham), john verzani (@jverzani), Michael Kane (@kaneplusplus), Bart Kastermans (@kasterma), Kevin D'Auria (@kdauria), Karandeep Singh (@kdpsingh), Ken Williams (@kenahoo), Kendon Bell (@kendonB), Kent Johnson (@kent37), Kevin Ushey (@kevinushey), (@kfeng123), Karl Forner (@kforner), Kirill Sevastyanenko

(@kirillseva), Brian Knaus (@knausb), Kirill Müller (@krlmlr), Kriti Sen Sharma (@ksens), Kevin Wright (@kwstat), suo.lawrence.liu@gmail.com (mailto:suo.lawrence.liu@gmail.com) (@Lawrence-Liu), @ldfmrails, Rachel Severson (@leighseverson), Laurent Gatto (@lgatto), C. Jason Liang (@liangcj), Steve Lianoglou (@lianost), @lindbrook, Lingbing Feng (@Lingbing), Marcel Ramos (@LiNk-NY), Zhongpeng Lin (@linzhp), Lionel Henry (@lionel-), myq (@lrcg), Luke W Johnston (@lwjohnst86), Kevin Lynagh (@lynaghk), Malcolm Barrett (@malcolmbarrett), @mannyishere, Matt (@mattbaggott), Matthew Grogan (@mattgrogan), @matthewhillary, Matthieu Gomez (@matthiegomez), Matt Malin (@mattmalin), Mauro Lepore (@maurolepore), Max Ghenis (@MaxGhenis), Maximilian Held (@maxheld83), Michal Bojanowski (@mbojan), Mark Rosenstein (@mbrmbr), Michael Sumner (@mdsumner), Jun Mei (@meijun), merkliopas (@merkliopas), mfrasco (@mfrasco), Michael Bach (@michaelbach), Michael Bishop (@MichaelMBishop), Michael Buckley (@michaelmikebuckley), Michael Quinn (@michaelquinn32), @miguelmorin, Michael (@mikekaminsky), Mine Cetinkaya-Rundel (@mine-cetinkaya-rundel), @mjsduncan, Mamoun Benghezal (@MoBeng), Matt Pettis (@mpettis), Martin Morgan (@mtmorgan), Guy Dawson (@Mullefa), Nacho Caballero (@nachocab), Natalya Rapstine (@natalya-patrikeeva), Nick Carchedi (@ncarchedi), Noah Greifer (@ngreifer), Nicholas Vasile (@nickv9), Nikos Ignatiadis (@ignatiadis), Xavier Laviron (@norival), Nick Pullen (@nstjhp), Oge Nnadi (@ogennadi), Oliver Paisley (@oliverpaisley), Pariksheat Nanda (@omsai), Øystein Sørensen (@osorensen), Paul (@otepoti), Otho Mantegazza (@othomanmantegazza), Dewey Dunnington (@paleolimbot), Parker Abercrombie (@parkerabercrombie), Patrick Hausmann (@patperu), Patrick Miller (@patr1ckm), Patrick Werkmeister (@Patrick01), @paulponcet, @pdb61, Tom Crockett (@pelotom), @pengyu, Jeremiah (@perryjer1), Peter Hickey (@PeteHaitch), Phil Chalmers (@philchalmers), Jose Antonio Magaña Mesa (@picarus), Pierre Casadebaig (@picasa), Antonio Piccolboni (@piccolbo), Pierre Roudier (@pierreroudier), Poor Yorick (@pooryorick), Marie-Helene Burle (@prosoitos), Peter Schulam (@pschulam), John (@quantbo), Quyu Kong (@qykong), Ramiro Magno (@ramiromagno), Ramnath Vaidyanathan (@ramnathy), Kun Ren (@renkun-ken), Richard Reeve (@richardreeve), Richard Cotton (@richierocks), Robert M Flight (@rmflight), R. Mark Sharp (@rmsharp), Robert Krzyzanowski (@robertzk), @robiRagan, Romain François (@romainfrancois), Ross Holmberg (@rossholmberg), Ricardo Pietrobon (@rpietro), @rrunner, Ryan Walker (@rtwalker), @rubenfcasal, Rob Weyant (@rweyant), Rumen Zarev (@rzarev), Nan Wang (@sailingwave), @sbgraves237, Scott Kostyshak (@scottkosty), Scott Leishman (@scctl), Sean Hughes (@seaaan), Sean Anderson (@seananderson), Sean Carmody (@seancarmody), Sebastian (@sebastian-c), Matthew Sedaghatfar (@sedaghatfar), @see24, Sven E. Templer (@setempler), @sflippl, @shabbybanks, Steven Pav (@shabbychef), Shannon Rush (@shannonrush), S'busiso Mkhondwane (@sibusiso16), Sigfried Gold (@Sigfried), Simon O'Hanlon (@simonohanlon101), Simon Potter (@sjp), Steve (@SplashDance), Scott Ritchie (@sritchie73), Tim Cole (@statist7), @ste-fan, @stephens999, Steve Walker (@stevencarlislewalker), Stefan Widgren (@stewid), Homer Strong (@strongh), Dirk (@surmann), Sebastien Vigneau (@svigneau), Scott Warchal (@Swarchal), Steven Nydick (@swnydick), Taekyun Kim (@taekyunk), Tal Galili (@talgalili), @Tazinho, Tom B (@tbuckl), @tdenes, @thomasherbig, Thomas (@thomaskern), Thomas Lin Pedersen (@thomas85), Thomas Zumbrunn (@thomaszumbrunn), Tim Waterhouse (@timwaterhouse), TJ Mahr (@tjmahr), Anton Antonov (@tonytonov), Ben Torvaney (@Torvaney), Jeff Allen (@trestletech), Terence Teo (@tteo), Tim Triche, Jr. (@ttriche), @tyhenkaline, Tyler Ritchie (@tylerritchie), Varun Agrawal (@varun729), Vijay Barve (@vijaybarve), Victor (@vkryukov), Vaidotas Zemlys-Balevičius (@vzemlys), Winston Chang (@wch), Linda Chin (@wchi144), Welliton Souza (@Welliton309), Gregg Whitworth (@whitwort), Will Beasley (@wibeasley), William R Bauer (@WilCrofter), William Doane (@WilDoane), Sean Wilkinson (@wilkinson), Christof Winter (@winterschlaefer), Bill Carver (@wmc3), Wolfgang Huber (@wolfganghuber), Krishna Sankar (@xsankar), Yihui Xie (@yihui), yang (@yiluhehei), Yoni Ben-Meshulam (@yoni), @yuchouchen, @zachcp, @zackham, Edward Cho (@zerokarmaleft), Albert Zhao (@zxzb).

1.8 Conventions

Throughout this book I use `f()` to refer to functions, `g` to refer to variables and function parameters, and `h/` to paths.

Larger code blocks intermingle input and output. Output is commented so that if you have an electronic version of the book, e.g., <https://adv-r.hadley.nz/>, you can easily copy and paste examples into R. Output

comments look like `#>` to distinguish them from regular comments.

Many examples use random numbers. These are made reproducible by `set.seed(1014)` which is run at the start of each chapter.

1.9 Colophon

This book was written in bookdown (<http://bookdown.org/>) inside RStudio (<http://www.rstudio.com/ide/>). The website (<https://adv-r.hadley.nz/>) is hosted with netlify (<http://netlify.com/>), and automatically updated after every commit by travis-ci (<https://travis-ci.org/>). The complete source is available from github (<https://github.com/hadley/adv-r>).

Code in the printed book is set in inconsolata (<http://levien.com/type/myfonts/inconsolata.html>).

setting	value
version	R version 3.5.0 (2017-01-27)
os	Ubuntu 14.04.5 LTS
system	x86_64, linux-gnu
ui	X11
language	(EN)
collate	en_US.UTF-8
tz	UTC
date	2018-09-04

package	version	source
assertthat	0.2.0	cran (\@0.2.0)
backports	1.1.2	cran (\@1.1.2)
base64enc	0.1-3	cran (\@0.1-3)
BH	1.66.0-1	cran (\@1.66.0-)
bindr	0.1.1	cran (\@0.1.1)
bindrcpp	0.2.2	cran (\@0.2.2)
bit	1.1-14	cran (\@1.1-14)
bit64	0.9-7	cran (\@0.9-7)
blob	1.1.1	cran (\@1.1.1)
bookdown	0.7	cran (\@0.7)
cli	1.0.0	cran (\@1.0.0)
clisymbols	1.2.0	cran (\@1.2.0)
codetools	0.2-15	CRAN (R 3.5.0)
colorspace	1.3-2	cran (\@1.3-2)
crayon	1.3.4	cran (\@1.3.4)
curl	3.2	CRAN (R 3.5.0)
DBI	1.0.0	cran (\@1.0.0)
dbplyr	1.2.2	cran (\@1.2.2)
devtools	1.13.6	CRAN (R 3.5.0)
digest	0.6.16	CRAN (R 3.5.0)
dplyr	0.7.6	cran (\@0.7.6)
emo	0.0.0.9000	Github (hadley/emo\@02a5206)
evaluate	0.11	cran (\@0.11)
fansi	0.3.0	cran (\@0.3.0)
furrr	0.1.0	cran (\@0.1.0)
future	1.9.0	cran (\@1.9.0)
ggplot2	3.0.0	cran (\@3.0.0)
git2r	0.23.0	CRAN (R 3.5.0)
globals	0.12.2	cran (\@0.12.2)
glue	1.3.0	Github (tidyverse/glue\@4e74901)
gttable	0.2.0	cran (\@0.2.0)
highr	0.7	cran (\@0.7)
hms	0.4.2	cran (\@0.4.2)
htmltools	0.3.6	cran (\@0.3.6)
httpuv	1.4.5	cran (\@1.4.5)
httr	1.3.1	CRAN (R 3.5.0)
inline	0.3.15	cran (\@0.3.15)
jsonlite	1.5	CRAN (R 3.5.0)
knitr	1.20	cran (\@1.20)
labeling	0.3	cran (\@0.3)
later	0.7.4	cran (\@0.7.4)
lattice	0.20-35	CRAN (R 3.5.0)
lazyeval	0.2.1	cran (\@0.2.1)
lineprof	0.1.9001	Github (hadley/lineprof\@972e71d)
listenv	0.7.0	cran (\@0.7.0)
lobstr	0.0.0.9000	Github (hadley/lobstr\@530db70)
lubridate	1.7.4	cran (\@1.7.4)
magrittr	1.5	cran (\@1.5)
markdown	0.8	cran (\@0.8)
MASS	7.3-49	CRAN (R 3.5.0)
Matrix	1.2-14	CRAN (R 3.5.0)
memoise	1.1.0	CRAN (R 3.5.0)
mgcv	1.8-23	CRAN (R 3.5.0)
microbenchmark	1.4-4	cran (\@1.4-4)
mime	0.5	CRAN (R 3.5.0)
munsell	0.5.0	cran (\@0.5.0)
nlme	3.1-137	CRAN (R 3.5.0)
ps	1.3.2	CRAN (R 3.5.0)

```
ruler()  
#> -----1-----2-----3-----4-----5-----6-----+  
#> 12345678901234567890123456789012345678901234567890123456789
```


Part I

Foundations

Introduction

To start your journey in mastering R, the following six chapters will help you learn what I consider to be the foundational components of R. I expect that you’re already seen many of these pieces before, but you probably have not studied them deeply. To help check your existing knowledge, each chapter starts with a quiz; if you get all the questions right, feel free to skip to the next chapter!

1. In Chapter 2, you’ll learn about one of the most important distinctions you haven’t previously needed to grapple with: the difference between an object and its name. Improving your mental model here will help you make better predictions about when R copies data and hence which basic operations are cheap and which are expensive.
2. Every day you’ve used R, you’ve used vectors, so Chapter 3 will dive into the details, helping you learn how the different types of vector fit together. You’ll also learn about attributes, which allow you to store arbitrary metadata, and form the basis for two of R’s object oriented programming toolkits
3. To write concise and performance R code it is important to fully appreciate the power of subsetting with `[`, `[[` and `$`, as described in Chapter 4. Understanding the fundamental components of subsetting will allow you to solve new problems by combining the building blocks in novel ways.
4. Functions are the most important building block of R code, and in Chapter 5, you’ll learn exactly how they work, including the **scoping** rules, which govern how R looks up values from names. You’ll also learn more of the details behind R’s lazy evaluation, and how you can control what happens when you exit a function.
5. In Chapter 6, you’ll learn about a data structure that is crucial for understanding how R works, but quite unimportant for data analysis: the environment. Environments are the data structure that binds names to values, and they power tools like package namespaces. Unlike most programming languages, environments in R are “first class” which means that you can manipulate them just like other objects.
6. Chapter 7 concludes this section of the book with a discussion of “conditions”, the umbrella term used to describe errors, warnings, and messages. You’ve certainly encountered these before, so in this chapter you learn how to signal them appropriately in your own functions, and how to handle them when signalled elsewhere.

Chapter 2

Names and values

2.1 Introduction

In R, it is important to understand the distinction between an object and its name. A correct mental model is important because it will help you:

- More accurately predict performance and memory usage of R code.
- Write faster code because accidental copies are a major cause of slow code.
- Better understand R’s functional programming tools.

The goal of this chapter is to help you understand the distinction between names and values, and when R will copy an object.

Quiz

Answer the following questions to see if you can safely skip this chapter. You can find the answers at the end of the chapter in Section 2.7.

1. Given the following data frame, how do I create a new column called “3” that contains the sum of 1 and 2? You may only use \$, not [[. What makes 1, 2, and 3 challenging as variable names?

```
df <- data.frame(runif(3), runif(3))
names(df) <- c(1, 2)
```

2. In the following code, how much memory does y occupy?

```
x <- runif(1e6)
y <- list(x, x, x)
```

3. On which line does a get copied in the following example?

```
a <- c(1, 5, 3, 2)
b <- a
b[[1]] <- 10
```

Outline

- Section 2.2 introduces you to the distinction between names and values, and discusses how <- creates a binding, or reference, between a name and a value.

- Section 2.3 describes when R makes a copy; whenever you modify vector, you’re almost always actually create a new, modified vector. You’ll learn how to use `tracemem()` to figure out when a copy actually occurs, and then explore the implications as they apply to function calls, lists, data frames, and character vectors.
- Section 2.4 explores the implications of the previous two sections on how much memory an object occupies. You’ll learn to use `lobstr::obj_size()` as your intuition may be profoundly wrong, and the base `object.size()` is unfortunately inaccurate.
- Section 2.5 describes the two important exceptions to copy-on-modify: values with a single name, and environments. In these two special cases, objects are actually modified in place.
- Section 2.6 closes out the chapter with a discussion of the garbage collector, which frees up memory used by objects that are no longer referenced by a name.

Prerequisites

We’ll use the development version of lobstr (<https://github.com/r-lib/lobstr>) to dig into the internal representation of R objects.

```
# devtools::install_github("r-lib/lobstr")
library(lobstr)
```

Sources

The details of R’s memory management are not documented in a single place. Much of the information in this chapter was gleaned from a close reading of the documentation (particularly `?Memory` and `?gc`), the memory profiling (<http://cran.r-project.org/doc/manuals/R-exts.html#Profiling-R-code-for-memory-use>) section of “Writing R extensions” (R Core Team 2018b), and the SEXPs (<http://cran.r-project.org/doc/manuals/R-ints.html#SEXP>) section of “R internals” (R Core Team 2018a). The rest I figured out by reading the C source code, performing small experiments, and asking questions on R-devel. Any mistakes are entirely mine.

2.2 Binding basics

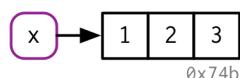
Take this code:

```
x <- c(1, 2, 3)
```

It’s easy to read it as: “create an object named ‘x’, containing the values 1, 2, and 3”. Unfortunately, that’s a simplification that will lead to you make inaccurate predictions about what R is actually doing behind the scenes. It’s more accurate to think about this code as doing two things:

- Creating an object, a vector of values, `c(1, 2, 3)`.
- Binding the object to a name, `x`.

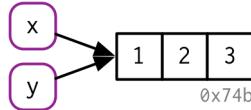
Note that the object, or value, doesn’t have a name; it’s the name that has a value. To make that distinction more clear, I’ll draw diagrams like this:



The name, `x`, is drawn with a rounded rectangle, and it has an arrow that points to, binds, or references, the value, the vector `1:3`. Note that the arrow points in opposite direction to the assignment arrow: `<-` creates a binding from the name on the left-hand side to the object on the right-hand side.

You can think of a name as a reference to a value. For example, if you run this code, you don't get another copy of the value `1:3`, you get another binding to the existing object:

```
y <- x
```



You might have noticed the value `1:3` has a label: `0x74b`. While the vector doesn't have a name, I'll occasionally need to refer to objects independent of their bindings. To make that possible, I'll label values with a unique identifier. These unique identifiers have a special form that looks like the object's memory "address", i.e. the location in memory in which the object is stored. It doesn't make sense to use the actual memory address because that changes every time the code is run.

You can access the address of an object with `lobstr::obj_addr()`. This allows us to see that `x` and `y` both point to the same location in memory:

```
obj_addr(x)
#> [1] "0x1fc3888"
obj_addr(y)
#> [1] "0x1fc3888"
```

These identifiers are long, and change every time you restart R.

It takes some time to get your head around the distinction between names and values, but it's really helpful for functional programming when you start to work with functions that have different names in different contexts.

2.2.1 Non-syntactic names

R has strict rules about what constitutes a valid name. A **syntactic** name must consist of letters¹, digits, `.` and `_`, and can't begin with `_` or a digit. Additionally, it can not be one of a list of **reserved words** like `TRUE`, `NULL`, `if`, and `function` (see the complete list in `?Reserved`). Names that don't follow these rules are called **non-syntactic** names, and if you try to use them, you'll get an error:

```
_abc <- 1
#> Error: unexpected input in "_"

if <- 10
#> Error: unexpected assignment in "if <-"
```

It's possible to override the usual rules and use a name with any sequence of characters by surrounding the name with backticks:

```
`_abc` <- 1
`_abc`
#> [1] 1

`if` <- 10
```

¹Surprisingly, what constitutes a letter is determined by your current locale. That means that the syntax of R code actually differs from computer to computer, and it's possible for a file that works on one computer to not even parse on another!

```
`if`  
#> [1] 10
```

Typically, you won't deliberately create such crazy names. Instead, you need to understand them because you'll be subjected to the crazy names created by others. This happens most commonly when you load data that has been created outside of R.

You *can* create non-syntactic bindings using single or double quotes (e.g. `"_abc" <- 1`) instead of backticks, but you shouldn't, because you'll have to use a different syntax to retrieve the values. The ability to use strings on the left hand side of the assignment arrow is a historical artefact, used before R supported backticks.

2.2.2 Exercises

1. Explain the relationship between `a`, `b`, `c` and `d` in the following code:

```
a <- 1:10  
b <- a  
c <- b  
d <- 1:10
```

2. The following code accesses the `mean` function in multiple different ways. Do they all point to the same underlying function object? Verify with `lobstr::obj_addr()`.

```
mean  
base::mean  
get("mean")  
evalq(mean)  
match.fun("mean")
```

3. By default, base R data import functions, like `read.csv()`, will automatically convert non-syntactic names to syntactic names. Why might this be problematic? What option allows you to suppress this behaviour?
4. What rules does `make.names()` use to convert non-syntactic names into syntactic names?
5. I slightly simplified the rules that govern syntactic names. Why is `.123e1` not a syntactic name? Read `?make.names` for the full details.

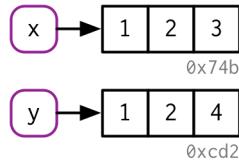
2.3 Copy-on-modify

Consider the following code, which binds `x` and `y` to the same underlying value, then modifies² `y`.

```
x <- c(1, 2, 3)  
y <- x  
  
y[[3]] <- 4  
x  
#> [1] 1 2 3
```

Modifying `y` clearly doesn't modify `x`, so what happened to the shared binding? While the value associated with `y` changes, the original object does not. Instead, R creates a new object, `0xcd2`, a copy of `0x74b` with one value changed, then rebinds `y` to that object.

²You may be surprised to see `[[` used with a numeric vector. We'll come back to this in Section 4.3, but in brief, I think you should use `[[` whenever you are getting or setting a single element.



This behaviour is called **copy-on-modify**, and understanding it makes your intuition for the performance of R code radically better. A related way to describe this phenomenon is to say that R objects are **immutable**, or unchangeable. However, I'll generally avoid that term because there are a couple of important exceptions to copy-on-modify that you'll learn about in Section 2.5.

2.3.1 `tracemem()`

You can see when an object gets copied with the help of `base::tracemem()`. You call it with an object and it returns the current address of the object:

```
x <- c(1, 2, 3)
cat(tracemem(x), "\n")
#> <0x7f80c0e0ffc8>
```

Whenever that object is copied in the future, `tracemem()` will print out a message telling you which object was copied, what the new address is, and the sequence of calls that lead to the copy:

```
y <- x
y[[3]] <- 4L
#> tracemem[0x7f80c0e0ffc8 -> 0x7f80c4427f40]:
```

Note that if you modify y again, it doesn't get copied. That's because the new object now only has a single name binding to it, so R can apply a modify-in-place optimisation. We'll come back to that shortly.

```
y[[3]] <- 5L

untracemem(y)
```

`untracemem()` is the opposite of `tracemem()`; it turns tracing off.

2.3.2 Function calls

The same rules for copying also apply to function calls. Take this code:

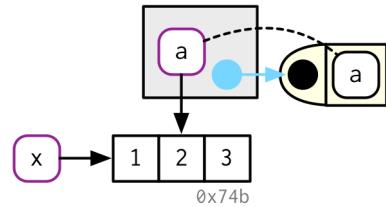
```
f <- function(a) {
  a
}

x <- c(1, 2, 3)
cat(tracemem(x), "\n")
#> <0x438ecf8>

z <- f(x)
# there's no copy here!

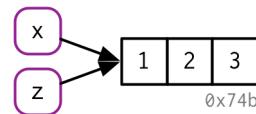
untracemem(x)
```

While `f()` is running, `a` inside the function will point to the same value as `x` does outside of it:



(You'll learn more about the conventions used in this diagram in Execution environments.)

And once complete, `x` and `z` will point to the same object. `0x74b` never gets copied because it never gets modified. If `f()` did modify `x`, `R` would create a new copy, and then `z` would bind that object.

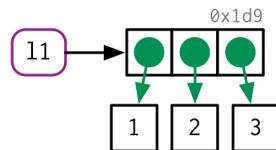


2.3.3 Lists

It's not just names (i.e. variables) that point to values; the elements of lists do too. Take this list, which superficially is very similar to the vector above:

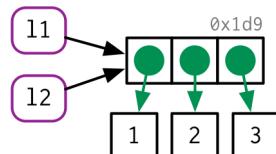
```
11 <- list(1, 2, 3)
```

The internal representation of the list is actually quite different to that of a vector. A list is really a vector of references:

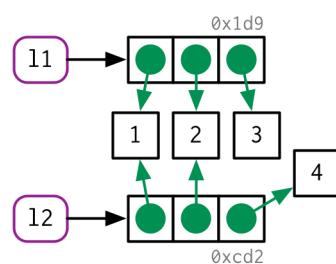


This is particularly important when we modify a list:

```
12 <- 11
```



```
12[[3]] <- 4
```



Like vectors, lists are copied-on-modify; the original list is left unchanged, and R creates a modified copy. This is a **shallow** copy: the list object and its bindings are copied, but the values pointed to by the bindings are not. The opposite of a shallow copy is a deep copy, where the contents of every reference are also copied. Prior to R 3.1.0, copies were always deep copies.

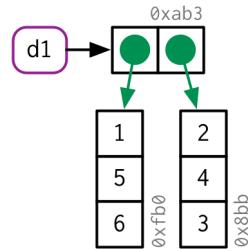
You can use `lobstr::ref()` to see values that are shared across lists. `ref()` prints the memory address of each object, along with a local id so that you can easily cross-reference shared components.

```
ref(11, 12)
#> [1:0x1ea4e18] <list>
#> [2:0x4509208] <dbl>
#> [3:0x45091d0] <dbl>
#> [4:0x4509198] <dbl>
#>
#> [5:0x75c12e8] <list>
#> [2:0x4509208]
#> [3:0x45091d0]
#> [6:0x76be2f8] <dbl>
```

2.3.4 Data frames

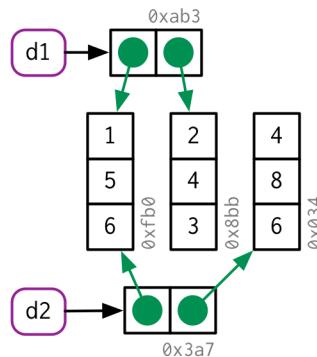
Data frames are lists of vectors, so copy-on-modify has important consequences when you modify a data frame. Take this data frame as an example:

```
d1 <- data.frame(x = c(1, 5, 6), y = c(2, 4, 3))
```



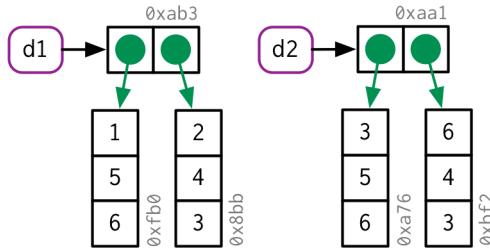
If you modify a column, only that column needs to be modified; the others can continue to point to the same place:

```
d2 <- d1
d2[, 2] <- d2[, 2] * 2
```



However, if you modify a row, there is no way to share data with the previous version of the data frame, and every column must be copied-and-modified.

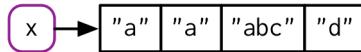
```
d3 <- d1
d3[1, ] <- d3[1, ] * 3
```



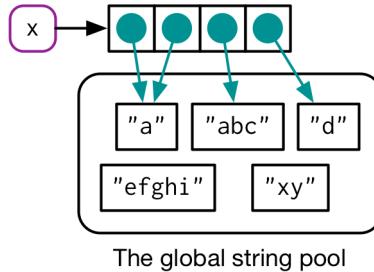
2.3.5 Character vectors

The final place that R uses references is in character vectors. I usually draw character vectors like this:

```
x <- c("a", "a", "abc", "d")
```



But this is a polite fiction, because R has a **global string pool**. Each element of a character vector is actually a pointer to a unique string in that pool:



You can request that `ref()` show these references:

```
ref(x, character = TRUE)
#> [1:0x35da988] <chr>
#> [2:0x149f6e8] <string: "a">
#> [2:0x149f6e8]
#> [3:0x2fccb88] <string: "abc">
#> [4:0x1955a58] <string: "d">
```

This has a profound impact on the amount of memory a character vector takes, but is otherwise not generally important, so elsewhere in the book I'll draw character vectors as if the strings live inside the vector.

2.3.6 Exercises

1. Why is `tracemem(1:10)` not useful?
2. Explain why `tracemem()` shows two copies when you run this code. Hint: carefully look at the difference between this code and the code shown earlier in the section.

```
x <- c(1L, 2L, 3L)
tracemem(x)

x[[3]] <- 4
```

3. Sketch out the relationship between the following objects:

```
a <- 1:10
b <- list(a, a)
c <- list(b, a, 1:10)
```

4. What happens when you run this code?

```
x <- list(1:10)
x[[2]] <- x
```

Draw a picture.

2.4 Object size

You can find out how much space an object occupies in memory with `lobstr::obj_size()`³:

```
obj_size(letters)
#> 1,792 B
obj_size(ggplot2::diamonds)
#> 3,457,048 B
```

Since the elements of lists are references to values, the size of a list might be much smaller than you expect:

```
x <- runif(1e6)
obj_size(x)
#> 8,000,048 B

y <- list(x, x, x)
obj_size(y)
#> 8,000,128 B
```

`y` is only 72 bytes⁴ bigger than `x`. That's the size of an empty list with three elements:

```
obj_size(list(NULL, NULL, NULL))
#> 80 B
```

Similarly, the global string pool means that character vectors take up less memory than you might expect: repeating a string 1000 times does not make it take up 1000 times as much memory.

```
banana <- "bananas bananas bananas"
obj_size(banana)
#> 272 B
obj_size(rep(banana, 100))
#> 1,064 B
```

References also make it challenging to think about the size of individual objects. `obj_size(x) + obj_size(y)` will only equal `obj_size(x, y)` if there are no shared values. Here, the combined size of `x` and `y` is the same as the size of `y`:

³Beware of the base `utils::object.size()` function. It does not correctly account for shared references and will return sizes that are too large.

⁴If you're running 32-bit R you'll see slightly different sizes.

```
obj_size(x, y)
#> 8,000,128 B
```

2.4.1 Exercises

1. In the following example, why are `object.size(y)` and `obj_size(y)` so radically different? Consult the documentation of `object.size()`.

```
y <- rep(list(runif(1e4)), 100)

object.size(y)
#> 8005648 bytes
obj_size(y)
#> 80,896 B
```

2. Take the following list. Why is its size somewhat misleading?

```
x <- list(mean, sd, var)
obj_size(x)
#> 17,664 B
```

3. Predict the output of the following code:

```
x <- runif(1e6)
obj_size(x)

y <- list(x, x)
obj_size(y)
obj_size(x, y)

y[[1]][[1]] <- 10
obj_size(y)
obj_size(x, y)

y[[2]][[1]] <- 10
obj_size(y)
obj_size(x, y)
```

2.5 Modify-in-place

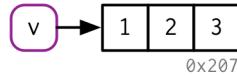
As we've seen above, modifying an R object will usually create a copy. There are two exceptions that we'll explore below:

- Objects with a single binding get a special performance optimisation.
- Environments are a special type of object that is always modified in place.

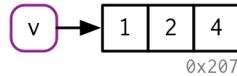
2.5.1 Objects with a single binding

If an object only has a single name that binds it, R will modify it in place:

```
v <- c(1, 2, 3)
```



```
v[[3]] <- 4
```



(Carefully note the object ids here: v continues to bind to the same object, 0x207.)

It's challenging to predict exactly when R applies this optimisation because of two complications:

- When it comes to bindings, R can currently⁵ only count 0, 1, and many. That means if an object has two bindings, and one goes away, the reference count does not go back to 1 (because one less than many is still many).
- Whenever you call any regular function, it will make a reference to the object. The only exception are specially written C functions. These occur mostly in the base package.

Together, this makes it hard to predict whether or not a copy will occur. Instead, it's better to determine it empirically with `tracemem()`. Let's explore the subtleties with a case study using for loops. For loops have a reputation for being slow in R, but often that slowness is because every iteration of the loop is creating a copy.

Consider the following code. It subtracts the median from each column of a large data frame:

```
x <- data.frame(matrix(runif(5 * 1e4), ncol = 5))
medians <- vapply(x, median, numeric(1))

for (i in seq_along(medians)) {
  x[[i]] <- x[[i]] - medians[[i]]
}
```

This loop is surprisingly slow because every iteration of the loop copies the data frame, as revealed by using `tracemem()`:

```
cat(tracemem(x), "\n")
#> <0x7f80c429e020>

for (i in 1:5) {
  x[[i]] <- x[[i]] - medians[[i]]
}
#> tracemem[0x7f80c429e020 -> 0x7f80c0c144d8]:
#> tracemem[0x7f80c0c144d8 -> 0x7f80c0c14540]: [[<- .data.frame [[<-
#> tracemem[0x7f80c0c14540 -> 0x7f80c0c145a8]: [[<- .data.frame [[<-
#> tracemem[0x7f80c0c145a8 -> 0x7f80c0c14610]:
#> tracemem[0x7f80c0c14610 -> 0x7f80c0c14678]: [[<- .data.frame [[<-
#> tracemem[0x7f80c0c14678 -> 0x7f80c0c146e0]: [[<- .data.frame [[<-
#> tracemem[0x7f80c0c146e0 -> 0x7f80c0c14748]:
#> tracemem[0x7f80c0c14748 -> 0x7f80c0c147b0]: [[<- .data.frame [[<-
#> tracemem[0x7f80c0c147b0 -> 0x7f80c0c14818]: [[<- .data.frame [[<-
#> tracemem[0x7f80c0c14818 -> 0x7f80c0c14880]:
#> tracemem[0x7f80c0c14880 -> 0x7f80c0c148e8]: [[<- .data.frame [[<-
#> tracemem[0x7f80c0c148e8 -> 0x7f80c0c14950]: [[<- .data.frame [[<-
```

⁵By the time you read this, that may have changed, as plans are afoot to improve reference counting: <https://developer.r-project.org/Refcnt.html>

```
#> tracemem[0x7f80c0c14950 -> 0x7f80c0c149b8]:
#> tracemem[0x7f80c0c149b8 -> 0x7f80c0c14a20]: [[<- .data.frame [[<-
#> tracemem[0x7f80c0c14a20 -> 0x7f80c0c14a88]: [[<- .data.frame [[<-
untracemem(x)
```

In fact, each iteration copies the data frame not once, not twice, but three times! Two copies are made by `[[.data.frame`, and a further copy⁶ it made because `[[.data.frame` is a regular function and hence increments the reference count of `x`.

We can reduce the number of copies by using a list instead of a data frame. Modifying a list uses internal C code, so the refs are not incremented and only a single copy is made:

```
y <- as.list(x)
cat(tracemem(y), "\n")
#> <0x7f80c5c3de20>

for (i in 1:5) {
  y[[i]] <- y[[i]] - medians[[i]]
}
#> tracemem[0x7f80c5c3de20 -> 0x7f80c48de210]:
```

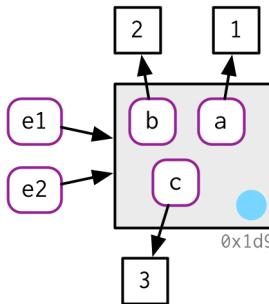
While it's not hard to determine when copies are made, it is hard to prevent them. If you find yourself resorting to exotic tricks to avoid copies, it may be time to rewrite your function in C++, as described in Chapter 24.

2.5.2 Environments

You'll learn more about environments in Chapter 6, but it's important to mention them here because they behave differently to other objects: environments are always modified in place. This property is sometimes described as **reference semantics** because when you modify an environment all existing bindings to the environment continue to have the same reference.

Take this environment, which we bind to `e1` and `e2`:

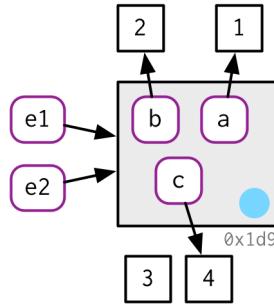
```
e1 <- rlang::env(a = 1, b = 2, c = 3)
e2 <- e1
```



If we change a binding, the environment is modified in place:

```
e1$c <- 4
e2$c
#> [1] 4
```

⁶Note that these copies are shallow, and only copy the reference to each individual column, not the contents. This means the performance isn't terrible, but it's obviously not as good as it could be.

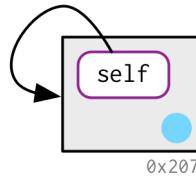


This basic idea can be used to create functions that “remember” their previous state. See Section 10.2.3 for more details.

One consequence of this is that environments can contain themselves:

```
e <- rlang::env()
e$self <- e

ref(e)
#> [1:0x1cbfe80] <env>
#> self = [1:0x1cbfe80]
```



This is a unique property of environments!

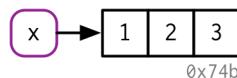
2.5.3 Exercises

1. Wrap the two methods for subtracting medians into two functions, then use the `bench` (Hester 2018) package to carefully compare their speeds. How does performance change as the number of columns increase?
2. What happens if you attempt to use `tracemem()` on an environment?

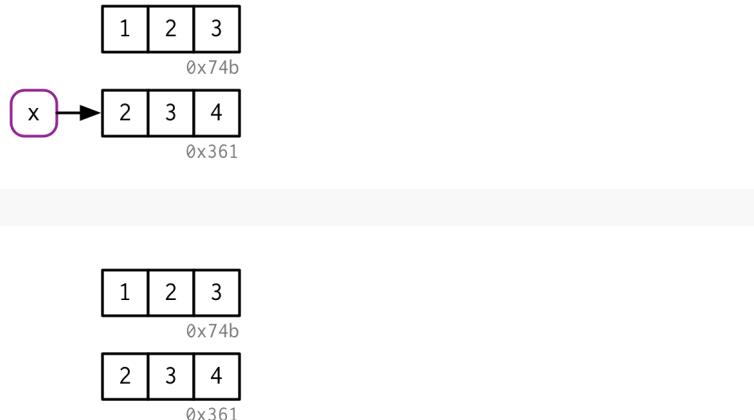
2.6 Unbinding and the garbage collector

Consider this code:

```
x <- 1:3
```



```
x <- 2:4
```



We create two objects, but by the end of code neither object is bound to a name. How do these objects get deleted? That's the job of the **garbage collector**, or GC, for short. The GC creates more memory by deleting R objects that are no longer used, and if needed, requesting more memory from the operating system.

R uses a **tracing** GC. That means it traces every object reachable from the global⁷ environment, and all the objects reachable from those objects (i.e. the references in lists and environments are searched recursively). The garbage collector does not use the reference count used for the modify-in-place optimisation described above. The two ideas are closely related but the internal data structures have been optimised for different use cases.

The garbage collector (GC) is run automatically whenever R needs more memory to create a new object. From the outside, it's basically impossible to predict when the GC will run, and indeed, you shouldn't try. Instead, if you want to find out when the GC runs, call `gcinfo(TRUE)`: the the GC will print a message to the console every time it runs.

You can force the garbage collector to run by calling `gc()`. Despite what you might have read elsewhere, there's never any *need* to call `gc()` yourself. You may *want* to call `gc()` to ask R to return memory to your operating system, or for its side-effect of telling you how much memory is currently being used:

```
gc()
#>      used (Mb) gc trigger (Mb) max used (Mb)
#> Ncells  675818 36.1    1284422 68.6  1284422 68.6
#> Vcells 3680210 28.1   11791146 90.0 11788009 90.0
```

`lobstr::mem_used()` is a wrapper around `gc()` that just prints the total number of bytes used:

```
mem_used()
#> 67,271,208 B
```

This number won't agree with the amount of memory reported by your operating system for three reasons:

1. It only includes objects created by R, not the R interpreter itself.
2. Both R and the operating system are lazy: they won't reclaim memory until it's actually needed. R might be holding on to memory because the OS hasn't yet asked for it back.
3. R counts the memory occupied by objects but there may be gaps due to deleted objects. This problem is known as memory fragmentation.

⁷And every environment on the current call stack.

2.7 Answers

1. You must surround non-syntactic names in ``. The variables 1, 2, and 3 have non-syntactic names, so must always be quoted with backticks.

```
df <- data.frame(runif(3), runif(3))
names(df) <- c(`1`, `2`)

df$`3` <- df$`1` + df$`2`
```

2. It occupies about 8 MB.

```
x <- runif(1e6)
y <- list(x, x, x)
obj_size(y)
#> 8,000,128 B
```

3. a is copied when b is modified, b[[1]] <- 10.

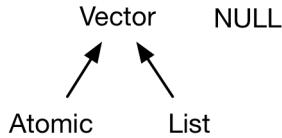
Chapter 3

Vectors

3.1 Introduction

This chapter discusses the most important family of data types in base R: the vector types¹. You've probably used many (if not all) of the vectors before, but you may not have thought deeply about how they are interrelated. In this chapter, I won't cover individual vectors types in too much depth. Instead, I'll show you how they fit together as a whole. If you need more details, you can find them in R's documentation.

Vectors come in two flavours: atomic vectors and lists². They differ in the types of their elements: all elements of an atomic vector must be the same type, whereas the elements of a list can have different types. Closely related to vectors is `NULL`; `NULL` is not a vector, but often serves the role of a generic 0-length vector. Throughout this chapter we'll expand on this diagram:



Every vector can also have **attributes**, which you can think of as a named list containing arbitrary metadata. Two attributes are particularly important because they create important vector variants. The **dimension** attribute turns vectors into matrices and arrays. The **class** attribute powers the S3 object system. You'll learn how to use S3 in Chapter 13, but here, you'll learn about a handful of the most important S3 vectors: factors, date/times, data frames, and tibbles. Matrices and data frames are not necessarily what you think of as vectors, so you'll learn why these 2d structures are considered to be vectors in R.

Quiz

Take this short quiz to determine if you need to read this chapter. If the answers quickly come to mind, you can comfortably skip this chapter. You can check your answers in answers.

1. What are the four common types of atomic vectors? What are the two rare types?
2. What are attributes? How do you get them and set them?

¹Collectively, all other data types are known as the “node” data types, and includes things like functions and environments. This is a highly technical term used in only a few places. The place where you're most likely to encounter it is the output of `gc()`: the “N” in `Ncells` stands for nodes, and the “V” in `Vcells` stands for vectors.

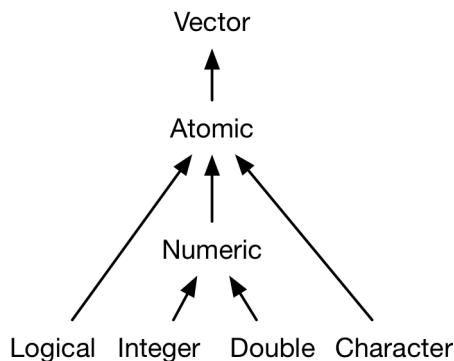
²A few places in R's documentation call lists generic vectors to emphasise their difference from atomic vectors.

3. How is a list different from an atomic vector? How is a matrix different from a data frame?
4. Can you have a list that is a matrix? Can a data frame have a column that is a matrix?
5. How do tibbles behave differently from data frames?

Outline

3.2 Atomic vectors

There are four common types of atomic vectors: logical, integer, double, and character. Collectively integer and double vectors are known as numeric vectors³. There are two rare types that I won't discuss further: complex and raw. Complex numbers are rarely needed for statistics, and raw vectors are a special type only needed when handling binary data.



3.2.1 Scalars

Each of the four primary atomic vectors has special syntax to create an individual value, aka a **scalar**⁴, and its own missing value.:

- Strings are surrounded by " ("hi") or ' ('bye'). The string missing value is `NA_character_`. Special characters are escaped with \\; see `?Quotes` for full details.
- Doubles can be specified in decimal (0.1234), scientific (1.23e4), or hexadecimal (0xcafe) forms. There are three special values unique to doubles: `Inf`, `-Inf`, and `NaN`. The double missing value is `NA_real_`.
- Integers are written similarly to doubles but must be followed by L⁵ (1234L, 1e4L, or 0xcafeL), and can not include decimals. The integer missing value is `NA_integer_`.
- Logicals can be spelled out (`TRUE` or `FALSE`), or abbreviated (`T` or `F`). The logical missing value is `NA`.

³This is a slight simplification as R does not use “numeric” consistently, which we’ll come back to in Section 12.4.1.

⁴Technically, the R language does not possess scalars, and everything that looks like a scalar is actually a vector of length one. This however, is mainly a theoretical distinction, and blurring the distinction between scalar and length-1 vector is unlikely to harm your code.

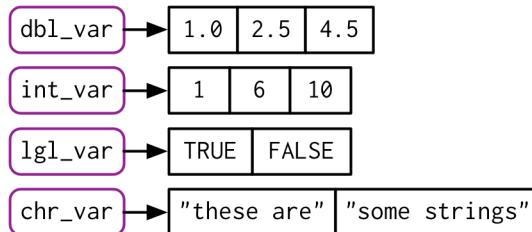
⁵L is not intuitive, and you might wonder where it comes from. At the time L was added to R, R’s integer type was equivalent to a long integer in C, and C code could use a suffix of L to force a number to be a long integer. It was decided that 1 was too visually similar to i (used for complex numbers in R), leaving L.

3.2.2 Making longer vectors with `c()`

To greater longer vectors from shorter vectors, use `c()`:

```
dbl_var <- c(1, 2.5, 4.5)
int_var <- c(1L, 6L, 10L)
lgl_var <- c(TRUE, FALSE)
chr_var <- c("these are", "some strings")
```

In diagrams, I'll depict vectors as connected rectangles, so the above code could be drawn as follows:



You can determine the type of a vector with `typeof()` and its length with `length()`.

```
typeof(dbl_var)
#> [1] "double"
typeof(int_var)
#> [1] "integer"
typeof(lgl_var)
#> [1] "logical"
typeof(chr_var)
#> [1] "character"
```

3.2.3 Testing and coercing

Generally, you can **test** if a vector is of a given type with an `is.` function, but they need to be used with care. `is.character()`, `is.double()`, `is.integer()`, and `is.logical()` do what you might expect: they test if a vector is a character, double, integer, or logical. Beware `is.vector()`, `is.atomic()`, and `is.numeric()`: they don't test if you have a vector, atomic vector, or numeric vector! We'll come back to what they actually do in Section 12.4.

The type is a property of the entire atomic vector, so all elements of an atomic must be the same type. When you attempt to combine different types they will be **coerced** to the most flexible one (character \gg double \gg integer \gg logical). For example, combining a character and an integer yields a character:

```
str(c("a", 1))
#> chr [1:2] "a" "1"
```

Coercion often happens automatically. Most mathematical functions (`+`, `log`, `abs`, etc.) will coerce to numeric. This coercion is particularly useful for logical vectors because `TRUE` becomes 1 and `FALSE` becomes 0.

```
x <- c(FALSE, FALSE, TRUE)
as.numeric(x)
#> [1] 0 0 1

# Total number of TRUES
sum(x)
#> [1] 1
```

```
# Proportion that are TRUE
mean(x)
#> [1] 0.333
```

Vectorised logical operations (`&`, `|`, `any`, etc) will coerce to a logical, but since this might lose information, it's always accompanied by a warning.

Generally, you can deliberately coerce by using an `as.` function, like `as.character()`, `as.double()`, `as.integer()`, or `as.logical()`. Failed coercions from strings generate a warning and a missing value:

```
as.integer(c("1", "1.5", "a"))
#> Warning: NAs introduced by coercion
#> [1] 1 1 NA
```

3.2.4 Exercises

1. How do you create scalars of type `raw` and `complex`? (See `?raw` and `?complex`)
 2. Test your knowledge of vector coercion rules by predicting the output of the following uses of `c()`:
- ```
c(1, FALSE)
c("a", 1)
c(TRUE, 1L)
```
3. Why is `1 == "1"` true? Why is `-1 < FALSE` true? Why is `"one" < 2` false?
  4. Why is the default missing value, `NA`, a logical vector? What's special about logical vectors? (Hint: think about `c(FALSE, NA_character_)`.)

## 3.3 Attributes

You might have noticed that the set of atomic vectors does not include a number of important data structures like matrices and arrays, factors and date/times. These types are built on top of atomic vectors by adding attributes. In this section, you'll learn the basics of attributes, and how the `dim` attribute makes matrices and arrays. In the next section you'll learn how the `class` attribute is used to create S3 vectors, including factors, dates, and date-times.

### 3.3.1 Getting and setting

You can think of attributes as a named list<sup>6</sup> used to attach metadata to an object. Individual attributes can be retrieved and modified with `attr()`, or retrieved en masse with `attributes()`, and set en masse with `structure()`.

```
a <- 1:3
attr(a, "x") <- "abcdef"
attr(a, "x")
#> [1] "abcdef"

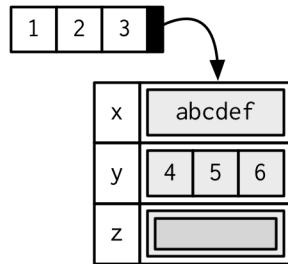
attr(a, "y") <- 4:6
str(attributes(a))
#> List of 2
```

---

<sup>6</sup>The reality is a little more complicated: attributes are actually stored in pairlists. Pairlists are functionally indistinguishable from lists, but are profoundly different under the hood, and you'll learn more about them in Section 17.4.5.

```
#> $ x: chr "abcdef"
#> $ y: int [1:3] 4 5 6

Or equivalently
a <- structure(
 1:3,
 x = "abcdef",
 y = 4:6
)
str(attributes(a))
#> List of 2
#> $ x: chr "abcdef"
#> $ y: int [1:3] 4 5 6
```



Attributes should generally be thought of as ephemeral. For example, most attributes are lost by most operations:

```
attributes(a[1])
#> NULL
attributes(sum(a))
#> NULL
```

There are only two attributes that are routinely preserved:

- **names**, a character vector giving each element a name.
- **dim**, short for dimensions, an integer vector, used to turn vectors into matrices and arrays.

To preserve additional attributes, you'll need to create your own S3 class, the topic of Chapter 13.

### 3.3.2 Names

You can name a vector in three ways:

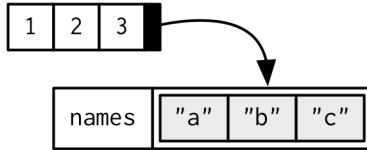
```
When creating it:
x <- c(a = 1, b = 2, c = 3)

By assigning names() to an existing vector:
x <- 1:3
names(x) <- c("a", "b", "c")

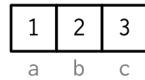
Inline, with setNames():
x <- setNames(1:3, c("a", "b", "c"))
```

Avoid using `attr(x, "names")` as it requires more typing and is less readable than `names(x)`. You can remove names from a vector by using `unname(x)` or `names(x) <- NULL`.

To be technically correct, when drawing the named vector `x`, I should draw it like so:



However, names are so special and so important, that unless I'm trying specifically to draw attention to the attributes data structure, I'll use them to label the vector directly:



To be maximally useful for character subsetting (e.g. Section 4.5.1) names should be unique, and non-missing, but this is not enforced by R. Depending on how the names are set, missing names may be either "" or `NA_character_`. If all names are missing, `names()` will return `NULL`.

### 3.3.3 Dimensions

Adding a `dim` attribute to a vector allows it to behave like a 2-dimensional **matrix** or multi-dimensional **array**. Matrices and arrays are primarily a mathematical/statistical tool, not a programming tool, so will be used infrequently in this book, and only covered briefly. Their most important feature is multidimensional subsetting, which is covered in Section 4.2.3.

You can create matrices and arrays with `matrix()` and `array()`, or by using the assignment form of `dim()`:

```
Two scalar arguments specify row and column sizes
a <- matrix(1:6, nrow = 2, ncol = 3)
a
#> [,1] [,2] [,3]
#> [1,] 1 3 5
#> [2,] 2 4 6

One vector argument to describe all dimensions
b <- array(1:12, c(2, 3, 2))
b
#> , , 1
#>
#> [,1] [,2] [,3]
#> [1,] 1 3 5
#> [2,] 2 4 6
#>
#> , , 2
#>
#> [,1] [,2] [,3]
#> [1,] 7 9 11
#> [2,] 8 10 12

You can also modify an object in place by setting dim()
c <- 1:6
dim(c) <- c(3, 2)
c
#> [,1] [,2]
#> [1,] 1 4
```

```
#> [2,] 2 5
#> [3,] 3 6
```

Many of the functions for working with vectors have generalisations for matrices and arrays:

| Vector          | Matrix                 | Array          |
|-----------------|------------------------|----------------|
| names()         | rownames(), colnames() | dimnames()     |
| length()        | nrow(), ncol()         | dim()          |
| c()             | rbind(), cbind()       | abind::abind() |
| —               | t()                    | aperm()        |
| is.null(dim(x)) | is.matrix()            | is.array()     |

A vector without `dim` attribute set is often thought of as 1-dimensional, but actually has a `NULL` dimensions. You also can have matrices with a single row or single column, or arrays with a single dimension. They may print similarly, but will behave differently. The differences aren't too important, but it's useful to know they exist in case you get strange output from a function (`tapply()` is a frequent offender). As always, use `str()` to reveal the differences.

```
str(1:3) # 1d vector
#> int [1:3] 1 2 3
str(matrix(1:3, ncol = 1)) # column vector
#> int [1:3, 1] 1 2 3
str(matrix(1:3, nrow = 1)) # row vector
#> int [1, 1:3] 1 2 3
str(array(1:3, 3)) # "array" vector
#> int [1:3(1d)] 1 2 3
```

### 3.3.4 Exercises

- How is `setNames()` implemented? How is `unname()` implemented? Read the source code.
- What does `dim()` return when applied to a 1d vector? When might you use `NROW()` or `NCOL()`?
- How would you describe the following three objects? What makes them different to `1:5`?

```
x1 <- array(1:5, c(1, 1, 5))
x2 <- array(1:5, c(1, 5, 1))
x3 <- array(1:5, c(5, 1, 1))
```

- An early draft used this code to illustrate `structure()`:

```
structure(1:5, comment = "my attribute")
#> [1] 1 2 3 4 5
```

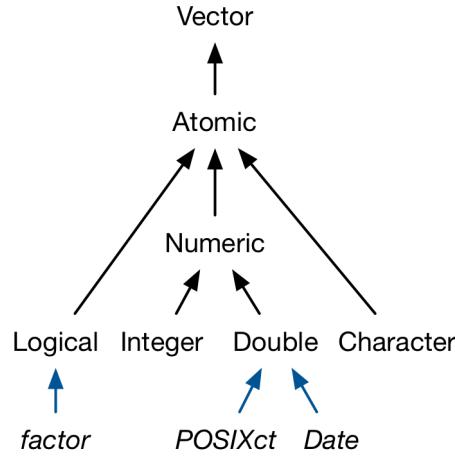
But when you print that object you don't see the comment attribute. Why? Is the attribute missing, or is there something else special about it? (Hint: try using `help`.)

## 3.4 S3 atomic vectors

One of the most important attributes is `class`, which defines the S3 object system. Having a class attribute makes an object an **S3 object**, which means that it will behave differently when passed to a **generic** function. Every S3 object is built on top of a base type, and often stores additional information in other attributes. You'll learn the details of the S3 object system, and how to create your own S3 classes, in Chapter 13.

In this section, we'll discuss three important S3 vectors used in base R:

- Categorical data, where values can only come from a fixed set of levels, are recorded in **factor** vectors.
- Dates (with day resolution) are recorded in **Date** vectors.
- Date-times (with second or sub-second resolution) are stored in **POSIXct** vectors.

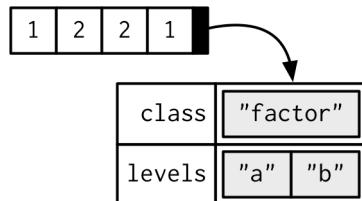


### 3.4.1 Factors

A factor is a vector that can contain only predefined values, and is used to store categorical data. Factors are built on top of integer vectors with two attributes: the `class`, “factor”, which makes them behave differently from regular integer vectors, and the `levels`, which defines the set of allowed values.

```
x <- factor(c("a", "b", "b", "a"))
x
#> [1] a b b a
#> Levels: a b

typeof(x)
#> [1] "integer"
attributes(x)
#> $levels
#> [1] "a" "b"
#>
#> $class
#> [1] "factor"
```



Factors are useful when you know the set of possible values, even if you don't see them all in a given dataset. Compared to a character vector, this means that tabulating a factor can yield counts of 0:

```
sex_char <- c("m", "m", "m")
sex_factor <- factor(sex_char, levels = c("m", "f"))

table(sex_char)
#> sex_char
#> m
#> 3
table(sex_factor)
#> sex_factor
#> m f
#> 3 0
```

A minor variation of factors is **ordered** factors, which generally behave similarly, but declare that the order of the levels is meaningful (a fact which is used automatically in some models and visualisations).

```
grade <- ordered(c("b", "b", "a", "c"), levels = c("c", "b", "a"))
grade
#[1] b b a c
#> Levels: c < b < a
```

With base R<sup>7</sup> you tend to encounter factors very frequently, because many base R functions (like `read.csv()` and `data.frame()`) automatically convert character vectors to factors. This is suboptimal, because there's no way for those functions to know the set of all possible levels or their optimal order: the levels are a property of the experimental design, not the data. Instead, use the argument `stringsAsFactors = FALSE` to suppress this behaviour, and then manually convert character vectors to factors using your knowledge of the data. To learn about the historical context of this behaviour, I recommend *stringsAsFactors: An unauthorized biography* (<http://simplystatistics.org/2015/07/24/stringsasfactors-an-unauthorized-biography/>) by Roger Peng, and *stringsAsFactors = <sigh>* (<http://notstatschat.tumblr.com/post/124987394001/stringsasfactors-sigh>) by Thomas Lumley.

While factors look like (and often behave like) character vectors, they are built on top of integers. Be careful when treating them like strings. Some string methods (like `gsub()` and `grep()`) will coerce factors to strings automatically, while others (like `nchar()`) will throw an error, and still others (like `c()`) will use the underlying integer values. For this reason, it's usually best to explicitly convert factors to character vectors if you need string-like behaviour.

### 3.4.2 Dates

Date vectors are built on top of double vectors. They have class “Date” and no other attributes:

```
today <- Sys.Date()

typeof(today)
#[1] "double"
attributes(today)
#> $class
#[1] "Date"
```

The value of the double (which can be seen by stripping the class), represents the number of days since 1970-01-01:

```
date <- as.Date("1970-02-01")
unclass(date)
```

---

<sup>7</sup>The tidyverse never automatically coerce characters to factor, and provides theforcats (Wickham 2018) package specifically for working with factors.

```
#> [1] 31
```

### 3.4.3 Date-times

Base R<sup>8</sup> provides two ways of storing date-time information, POSIXct, and POSIXlt. These are admittedly odd names: “POSIX” is short for Portable Operating System Interface which is a family of cross-platform standards. “ct” stands for calendar time (the `time_t` type in C), and “lt” for local time (the `struct tm` type in C). Here we’ll focus on `POSIXct`, because it’s the simplest, is built on top of an atomic vector, and is most appropriate for use in data frames. `POSIXct` vectors are built on top of double vectors, where the value represents the number of days since 1970-01-01.

```
now_ct <- as.POSIXct("2018-08-01 22:00", tz = "UTC")
now_ct
#> [1] "2018-08-01 22:00:00 UTC"

typeof(now_ct)
#> [1] "double"
attributes(now_ct)
#> $class
#> [1] "POSIXct" "POSIXt"
#>
#> $tzone
#> [1] "UTC"
```

The `tzone` attribute controls how the date-time is formatted, not the instant of time represented by the vector. Note that the time is not printed if it is midnight.

```
structure(now_ct, tzone = "Asia/Tokyo")
#> [1] "2018-08-02 07:00:00 JST"
structure(now_ct, tzone = "America/New_York")
#> [1] "2018-08-01 18:00:00 EDT"
structure(now_ct, tzone = "Australia/Lord_Howe")
#> [1] "2018-08-02 08:30:00 +1030"
structure(now_ct, tzone = "Europe/Paris")
#> [1] "2018-08-02 CEST"
```

### 3.4.4 Exercises

1. What sort of object does `table()` return? What is its type? What attributes does it have? How does the dimensionality change as you tabulate more variables?
2. What happens to a factor when you modify its levels?

```
f1 <- factor(letters)
levels(f1) <- rev(levels(f1))
```

3. What does this code do? How do `f2` and `f3` differ from `f1`?

```
f2 <- rev(factor(letters))

f3 <- factor(letters, levels = rev(letters))
```

---

<sup>8</sup>The tidyverse provides the lubridate (Grolemund and Wickham 2011) package for working with date-times. It provides a number of convenient helpers all which work with the base `POSIXct` type.

## 3.5 Lists

Lists are a step up in complexity from atomic vectors because an element of a list can be any type (not just vectors). An element of a list can even be another list!

### 3.5.1 Creating

Construct lists with `list()`:

```
11 <- list(
 1:3,
 "a",
 c(TRUE, FALSE, TRUE),
 c(2.3, 5.9)
)

typeof(11)
#> [1] "list"

str(11)
#> List of 4
#> $: int [1:3] 1 2 3
#> $: chr "a"
#> $: logi [1:3] TRUE FALSE TRUE
#> $: num [1:2] 2.3 5.9
```

As described in Section 2.3.3, the elements of a list are references. Creating a list does not copy the components in, so the total size of a list might be smaller than you expect.

```
lobstr::obj_size(mtcars)
#> 7,792 B

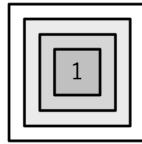
12 <- list(mtcars, mtcars, mtcars, mtcars)
lobstr::obj_size(12)
#> 7,872 B
```

Lists can contain complex objects so it's not possible to pick one visual style that works for every list. Generally I'll draw lists like vectors, using colour to remind you of the hierarchy.



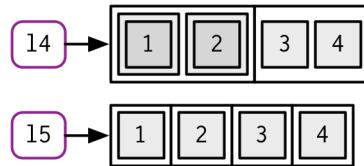
Lists are sometimes called **recursive** vectors, because a list can contain other lists. This makes them fundamentally different from atomic vectors.

```
13 <- list(list(list(1)))
str(13)
#> List of 1
#> $:List of 1
#> ..$:List of 1
#> ...$: num 1
```



`c()` will combine several lists into one. If given a combination of atomic vectors and lists, `c()` will coerce the vectors to lists before combining them. Compare the results of `list()` and `c()`:

```
14 <- list(list(1, 2), c(3, 4))
15 <- c(list(1, 2), c(3, 4))
str(14)
#> List of 2
#> $: List of 2
#> ..$: num 1
#> ..$: num 2
#> $: num [1:2] 3 4
str(15)
#> List of 4
#> $: num 1
#> $: num 2
#> $: num 3
#> $: num 4
```



### 3.5.2 Testing and coercing

The `typeof()` of a list is `list`. You can test for a list with `is.list()`.

And coerce to a list with `as.list()`.

```
list(1:3)
#> [[1]]
#> [1] 1 2 3
as.list(1:3)
#> [[1]]
#> [1] 1
#>
#> [[2]]
#> [1] 2
#>
#> [[3]]
#> [1] 3
```

You can turn a list into an atomic vector with `unlist()`. The rules for the resulting type are complex, not well documented, and not always equivalent to `c()`.

### 3.5.3 Matrices and arrays

While atomic vectors are most commonly turned into matrices, the dimension attribute can also be set on lists to make list-matrices or list-arrays:

```
l <- list(1:3, "a", TRUE, 1.0)
dim(l) <- c(2, 2)
l
#> [,1] [,2]
#> [1,] Integer,3 TRUE
#> [2,] "a" 1
l[[1, 1]]
#> [1] 1 2 3
```

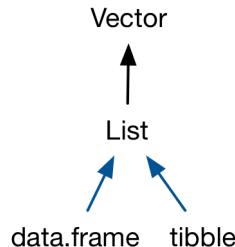
These are relatively esoteric data structures, but can be useful if you want to arrange objects into a grid-like structure. For example, if you’re running models on a spatio-temporal grid, it might be natural to preserve the grid structure by storing the models in a 3d array.

### 3.5.4 Exercises

1. List all the ways that a list differs from an atomic vector.
2. Why do you need to use `unlist()` to convert a list to an atomic vector? Why doesn’t `as.vector()` work?
3. Compare and contrast `c()` and `unlist()` when combining a date and date-time into a single vector.

## 3.6 Data frames and tibbles

There are two important S3 vectors that are built on top of lists: data frames and tibbles.



A data frame is the most common way of storing data in R, and is crucial for effective data analysis. A data frame is a named list of equal-length vectors. It has attributes providing the (column) `names`, `row.names`<sup>9</sup>, and a class of “data.frame”:

```
df1 <- data.frame(x = 1:2, y = 2:1)
typeof(df1)
#> [1] "list"
```

<sup>9</sup>Row names are one of the most surprisingly complex data structures in R, because they’ve been a persistent performance issue over many years. The most straightforward representations are character or integer vectors, with one element for each row. There’s also a compact representation for “automatic” row names (consecutive integers), created by `.set_row_names()`. R 3.5 has a special way of deferring integer to character conversions specifically to speed up `lm()`; see [https://svn.r-project.org/R/branches/ALTREP/ALTREP.html#deferred\\_string\\_conversions](https://svn.r-project.org/R/branches/ALTREP/ALTREP.html#deferred_string_conversions) for details.

```
attributes(df1)
#> $names
#> [1] "x" "y"
#>
#> $class
#> [1] "data.frame"
#>
#> $row.names
#> [1] 1 2
```

Because each element of the list has the same length, data frames have a rectangular structure, and hence shares properties of both the matrix and the list:

- A data frame has 1d `names()`, and 2d `colnames()` and `rownames()`<sup>10</sup>. The `names()` and `colnames()` are identical.
- A data frame has 1d `length()`, and 2d `ncol()` and `nrow()`. The `length()` is the number of columns.

Data frames are one of the biggest and most important ideas in R, and one of the things that makes R different from other programming languages. However, in the over 20 years since their creation, the ways people use R have changed, and some of the design decisions that made sense at the time data frames were created now cause frustration.

This frustration lead to the creation of the tibble (Müller and Wickham 2018), a modern reimagining of the data frame. Tibbles are designed to be (as much as possible) drop-in replacements for data frames, while still fixing the greatest frustrations. A concise, and fun, way to summarise the main differences is that tibbles are lazy and surly: they tend to do less and complain more. You'll see what that means as you work through this section.

Tibbles are provided by the tibble package and share the the same structure as a data frame. The only difference is that the class vector is longer, and includes `tbl_df`. This allows tibbles to behave differently in the key ways which we'll discuss below.

```
library(tibble)

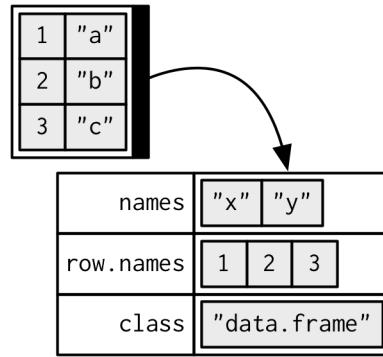
df2 <- tibble(x = 1:2, y = 2:1)
typeof(df2)
#> [1] "list"

attributes(df2)
#> $names
#> [1] "x" "y"
#>
#> $row.names
#> [1] 1 2
#>
#> $class
#> [1] "tbl_df" "tbl" "data.frame"
```

When drawing data frames and tibbles, rather than focussing on the implementation details, i.e. the attributes:

---

<sup>10</sup>Technically, you are encouraged to use `row.names()`, not `rownames()` with data frames, but this distinction is rarely important.



I'll draw them in the same way as a named list, but arranged to emphasised their columnar structure.

| x | y     |
|---|-------|
| 1 | \"a\" |
| 2 | \"b\" |
| 3 | \"c\" |

### 3.6.1 Creating

You create a data frame by supplying name-vector pairs to `data.frame()`:

```
df <- data.frame(
 x = 1:3,
 y = c("a", "b", "c")
)
str(df)
#> 'data.frame': 3 obs. of 2 variables:
#> $ x: int 1 2 3
#> $ y: Factor w/ 3 levels "a","b","c": 1 2 3
```

Beware the default conversion of strings to factors. Use `stringsAsFactors = FALSE` to suppress it and keep character vectors as character vectors:

```
df1 <- data.frame(
 x = 1:3,
 y = c("a", "b", "c"),
 stringsAsFactors = FALSE
)
str(df1)
#> 'data.frame': 3 obs. of 2 variables:
#> $ x: int 1 2 3
#> $ y: chr "a" "b" "c"
```

Creating a tibble is similar, but tibbles never coerce their input (this is one feature that makes them lazy):

```
df2 <- tibble(
 x = 1:3,
 y = c("a", "b", "c")
)
str(df2)
```

```
#> Classes 'tbl_df', 'tbl' and 'data.frame': 3 obs. of 2 variables:
#> $ x: int 1 2 3
#> $ y: chr "a" "b" "c"
```

Additionally, while data frames automatically transform non-syntactic names (unless `check.names = FALSE`); tibbles do not (although they do print non-syntactic names surrounded by `).

```
names(data.frame(`1` = 1))
#> [1] "X1"

names(tibble(`1` = 1))
#> [1] "1"
```

While every element of a data frame (or tibble) must have the same length, both `data.frame()` and `tibble()` can recycle shorter inputs. Data frames automatically recycle columns that are an integer multiple of the longest column; tibbles only ever recycle vectors of length 1.

```
data.frame(x = 1:4, y = 1:2)
#> x y
#> 1 1 1
#> 2 2 2
#> 3 3 1
#> 4 4 2
data.frame(x = 1:4, y = 1:3)
#> Error in data.frame(x = 1:4, y = 1:3):
#> arguments imply differing number of rows: 4, 3

tibble(x = 1:4, y = 1)
#> # A tibble: 4 x 2
#> x y
#> <int> <dbl>
#> 1 1 1
#> 2 2 1
#> 3 3 1
#> 4 4 1
tibble(x = 1:4, y = 1:2)
#> Error: Column `y` must be length 1 or 4, not 2
```

There is one final difference: `tibble()` allows you to refer to newly created variables:

```
tibble(
 x = 1:3,
 y = x * 2
)
#> # A tibble: 3 x 2
#> x y
#> <int> <dbl>
#> 1 1 2
#> 2 2 4
#> 3 3 6
```

### 3.6.2 Row names

Data frames allow you to label each row with a “name”, a character vector containing only unique values:

```
df3 <- data.frame(
 age = c(35, 27, 18),
 hair = c("blond", "brown", "black"),
 row.names = c("Bob", "Susan", "Sam")
)
df3
#> age hair
#> Bob 35 blond
#> Susan 27 brown
#> Sam 18 black
```

You can get and set row names with `rownames()`, and you can use them to subset rows:

```
rownames(df3)
#> [1] "Bob" "Susan" "Sam"

df3["Bob",]
#> age hair
#> Bob 35 blond
```

Row names arise naturally if you think of data frames as 2d structures like matrices: the columns (variables) have names so the rows (observations) should too. Most matrices are numeric, so having a place to store character labels is important. But this analogy to matrices is misleading because matrices possess an important property that data frames do not: they are transposable. In matrices the rows and columns are interchangeable, and transposing a matrix gives you another matrix (and transposing again gives you back the original matrix). With data frames, however, the rows and columns are not interchangeable, and the transpose of a data frame is not a data frame.

There are three reasons that row names are suboptimal:

- Metadata is data, so storing it in a different way to the rest of the data is fundamentally a bad idea. It also means that you need to learn a new set of tools to work with row names; you can't use what you already know about manipulating columns.
- Row names are poor abstraction for labelling rows because they only work when a row can be identified by a single string. This fails in many cases, for example when you want to identify a row by a non-character vector (e.g. a time point), or with multiple vectors (e.g. position, encoded by latitude and longitude).
- Row names must be unique, so any replication of rows (e.g. from bootstrapping) will create new row names. If you want to match rows from before and after the transformation you'll need to perform complicated string surgery.

```
df3[c(1, 1, 1),]
#> age hair
#> Bob 35 blond
#> Bob.1 35 blond
#> Bob.2 35 blond
```

For these reasons, tibbles do not support row names. Instead the `tibble` package provides tools to easily convert row names into a regular column with either `rownames_to_column()`, or the `rownames` argument to `as_tibble()`:

```
as_tibble(df3, rownames = "name")
#> # A tibble: 3 x 3
#> name age hair
#> <chr> <dbl> <fct>
#> 1 Bob 35 blond
```

```
#> 2 Susan 27 brown
#> 3 Sam 18 black
```

### 3.6.3 Printing

One of the most obvious differences between tibbles and data frames is how they are printed. I assume that you're already familiar with how data frames are printed, so here I'll highlight some of the biggest differences using an example dataset included in the dplyr package:

```
dplyr::starwars
#> # A tibble: 87 x 13
#> name height mass hair_color skin_color eye_color birth_year
#> <chr> <int> <dbl> <chr> <chr> <chr> <dbl>
#> 1 Luke~ 172 77 blond fair blue 19
#> 2 C-3PO 167 75 <NA> gold yellow 112
#> 3 R2-D2 96 32 <NA> white, bl~ red 33
#> 4 Dart~ 202 136 none white yellow 41.9
#> 5 Leia~ 150 49 brown light brown 19
#> 6 Owen~ 178 120 brown, gr~ light blue 52
#> 7 Beru~ 165 75 brown light blue 47
#> 8 R5-D4 97 32 <NA> white, red red NA
#> 9 Bigg~ 183 84 black light brown 24
#> 10 Obi-- 182 77 auburn, w~ fair blue-gray 57
#> # ... with 77 more rows, and 6 more variables: gender <chr>,
#> # homeworld <chr>, species <chr>, films <list>, vehicles <list>,
#> # starships <list>
```

Tibbles:

- Only show the first 10 rows and all the columns that will fit on screen. Additional columns are shown at the bottom.
- Each column is labelled with its type, abbreviated to three or four letters.
- Wide columns are truncated to avoid a single long string occupying an entire row. (This is still a work in progress: it's tricky to get the tradeoff right between showing as many columns as possible and showing a single wide column fully.)
- When used in console environments that support it, colour is used judiciously to highlight important information, and de-emphasise supplemental details.

### 3.6.4 Subsetting

As you will learn in Chapter 4, you can subset a data frame or a tibble like a 1d structure (where it behaves like a list), or a 2d structure (where it behaves like a matrix).

In my opinion, data frames have two suboptimal subsetting behaviours:

- When you subset columns with `df[, vars]`, you will get a vector if `vars` selects one variable, otherwise you'll get a data frame. This is a frequent source of bugs when using `[` in a function, unless you always remember to do `df[, vars, drop = FALSE]`.
- When you attempt to extract a single column with `df$x` and there is no column `x`, a data frame will instead select any variable that starts with `x`. If no variable starts with `x`, `df$x` will return `NULL`. This makes it easy to select the wrong variable or to select a variable that doesn't exist.

Tibbles tweak these behaviours so that `[` always returns a tibble, and `$` doesn't partial match, and warns if it can't find a variable (this is what makes tibbles surly).

```
df1 <- data.frame(xyz = "a")
df2 <- tibble(xyz = "a")

str(df1$x)
#> Factor w/ 1 level "a": 1
str(df2$x)
#> Warning: Unknown or uninitialised column: 'x'.
#> NULL
```

A tibble's insistence on returning a data frame from `[` can cause problems with legacy code, which often uses `df[, "col"]` to extract a single column. To fix this, use `df[["col"]]` instead; this is more expressive (since `[` always extracts a single element) and works with both data frames and tibbles.

### 3.6.5 Testing and coercing

To check if an object is a data frame or tibble, use `is.data.frame()`:

```
is.data.frame(df1)
#> [1] TRUE
is.data.frame(df2)
#> [1] TRUE
```

Typically, it should not matter if you have a tibble or data frame, but if you do need to distinguish, use `is_tibble()`:

```
is_tibble(df1)
#> [1] FALSE
is_tibble(df2)
#> [1] TRUE
```

You can coerce an object to a data frame with `as.data.frame()` or to a tibble with `as_tibble()`.

### 3.6.6 List columns

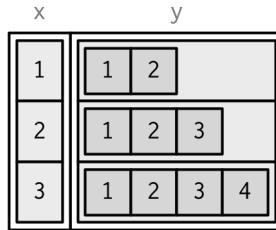
Since a data frame is a list of vectors, it is possible for a data frame to have a column that is a list. This is very useful because a list can contain any other object, which means that you can put any object in a data frame. This allows you to keep related objects together in a row, no matter how complex the individual objects are. You can see an application of this in the “Many Models” chapter of “R for Data Science”, <http://r4ds.had.co.nz/many-models.html>.

List-columns are allowed in data frames but you have to do a little extra work, either adding the list-column after creation, or wrapping the list in `I()`.

```
df <- data.frame(x = 1:3)
df$y <- list(1:2, 1:3, 1:4)

data.frame(
 x = 1:3,
 y = I(list(1:2, 1:3, 1:4))
)
#> x y
#> 1 1 1, 2
```

```
#> 2 2 1, 2, 3
#> 3 3 1, 2, 3, 4
```



List columns are easier to use with tibbles because you can provide them inside `tibble()`, are they are handled specially when printing:

```
tibble(
 x = 1:3,
 y = list(1:2, 1:3, 1:4)
)
#> # A tibble: 3 x 2
#> x y
#> <int> <list>
#> 1 1 <int [2]>
#> 2 2 <int [3]>
#> 3 3 <int [4]>
```

### 3.6.7 Matrix and data frame columns

It's also possible to have a column of a data frame that's a matrix or array, as long as the number of rows matches the data frame. (This requires a slight extension to our definition of a data frame: it's not the `length()` of each column that must be equal; but the `NROW()`.) Like with list-columns, you must either add after creation, or wrap in `I()`.

```
dfm <- data.frame(
 x = 1:3 * 10
)
dfm$y <- matrix(1:9, nrow = 3)
dfm$z <- data.frame(a = 3:1, b = letters[1:3], stringsAsFactors = FALSE)

str(dfm)
#> 'data.frame': 3 obs. of 3 variables:
#> $ x: num 10 20 30
#> $ y: int [1:3, 1:3] 1 2 3 4 5 6 7 8 9
#> $ z:'data.frame': 3 obs. of 2 variables:
#> ..$ a: int 3 2 1
#> ..$ b: chr "a" "b" "c"
```

| x  | y     | z     |
|----|-------|-------|
| 10 | 1 4 7 | a b   |
| 20 | 2 5 8 | 3 "a" |
| 30 | 3 6 9 | 2 "b" |

Matrix and data frame columns require a little caution. Many functions that work with data frames assume that all columns are vectors, and the printed display can be confusing.

```
dfm[1,]
#> x y.1 y.2 y.3 z.a z.b
#> 1 10 1 4 7 3 a
```

### 3.6.8 Exercises

1. Can you have a data frame with 0 rows? What about 0 columns?
2. What happens if you attempt to set rownames that are not unique?
3. If `df` is a data frame, what can you say about `t(df)`, and `t(t(df))`? Perform some experiments, making sure to try different column types.
4. What does `as.matrix()` do when applied to a data frame with columns of different types? How does it differ from `data.matrix()`?

## 3.7 NULL

To finish up the chapter, I wanted to talk about a final important data structure that's closely related to vectors: `NULL`. `NULL` is special because it has a unique type, is always length 0, and can't have any attributes:

```
typeof(NULL)
#> [1] "NULL"

length(NULL)
#> [1] 0

x <- NULL
attr(x, "y") <- 1
#> Error in attr(x, "y") <- 1:
#> attempt to set an attribute on NULL
```

You can test for `NULL`s with `is.null()`:

```
is.null(NULL)
#> [1] TRUE
```

There are two common uses of `NULL`:

- To represent an empty vector (a vector of length 0) of arbitrary type. For example, if you use `c()` but don't include any arguments, you get `NULL`, and concatenating `NULL` to a vector leaves it unchanged<sup>11</sup>:

<sup>11</sup>Algebraically, this makes `NULL` the identity element under vector concatenation.

```
c()
#> NULL
```

- To represent an absent vector. For example, `NULL` is often used as a default function argument, when the argument is optional but the default value requires some computation (see Section 5.5.4 for more on this idea). Contrast this with `NA` which is used to indicate that an *element* of a vector is absent.

If you’re familiar with SQL, you know about relational `NULL` and might expect it to be the same as R’s. However, the database `NULL` is actually equivalent to `NA`.

## 3.8 Answers

1. The four common types of atomic vector are logical, integer, double and character. The two rarer types are complex and raw.
2. Attributes allow you to associate arbitrary additional metadata to any object. You can get and set individual attributes with `attr(x, "y")` and `attr(x, "y") <- value`; or get and set all attributes at once with `attributes()`.
3. The elements of a list can be any type (even a list); the elements of an atomic vector are all of the same type. Similarly, every element of a matrix must be the same type; in a data frame, the different columns can have different types.
4. You can make “list-array” by assigning dimensions to a list. You can make a matrix a column of a data frame with `df$x <- matrix()`, or using `I()` when creating a new data frame `data.frame(x = I(matrix()))`.
5. Tibbles have an enhanced print method, never coerce strings to factors, and provide stricter subsetting methods.

# Chapter 4

## Subsetting

### 4.1 Introduction

R’s subsetting operators are powerful and fast. Mastery of subsetting allows you to succinctly express complex operations in a way that few other languages can match. Subsetting is easy to learn but hard to master because you need to internalise a number of interrelated concepts:

- The six types of thing that you can subset with.
- The three subsetting operators, `[`, `[[`, and `$`.
- How the subsetting operators interact with vector types (e.g., atomic vectors, lists, factors, matrices, and data frames).
- The use of subsetting together with assignment.

This chapter helps you master subsetting by starting with the simplest type of subsetting: subsetting an atomic vector with `[`. It then gradually extends your knowledge, first to more complicated data types (like arrays and lists), and then to the other subsetting operators, `[[]` and `$`. You’ll then learn how subsetting and assignment can be combined to modify parts of an object, and, finally, you’ll see a large number of useful applications.

Subsetting is a natural complement to `str()`. `str()` shows you the structure of any object, and subsetting allows you to pull out the pieces that you’re interested in. For large, complex objects, I also highly recommend the interactive RStudio Viewer, which you can activate with `View(my_object)`.

### Quiz

Take this short quiz to determine if you need to read this chapter. If the answers quickly come to mind, you can comfortably skip this chapter. Check your answers in Section 4.6.

1. What is the result of subsetting a vector with positive integers, negative integers, a logical vector, or a character vector?
2. What’s the difference between `[`, `[[]`, and `$` when applied to a list?
3. When should you use `drop = FALSE`?
4. If `x` is a matrix, what does `x[] <- 0` do? How is it different to `x <- 0`?
5. How can you use a named vector to relabel categorical variables?

## Outline

- Section 4.2 starts by teaching you about `[`. You'll start by learning the six types of data that you can use to subset atomic vectors. You'll then learn how those six data types act when used to subset lists, matrices, and data frames.
- Section 4.3 expands your knowledge of subsetting operators to include `[[` and `$`, focussing on the important principles of simplifying vs. preserving.
- In Section 4.4 you'll learn the art of subassignment, combining subsetting and assignment to modify parts of an object.
- Section 4.5 leads you through eight important, but not obvious, applications of subsetting to solve problems that you often encounter in a data analysis.

## 4.2 Selecting multiple elements

It's easiest to learn how subsetting works for atomic vectors, and then how it generalises to higher dimensions and other more complicated objects. We'll start with `[`, the most commonly used operator which allows you to extract any number of elements. Section 4.3 will cover `[[` and `$`, used to extra a single element from a data structure.

### 4.2.1 Atomic vectors

Let's explore the different types of subsetting with a simple vector, `x`.

```
x <- c(2.1, 4.2, 3.3, 5.4)
```

Note that the number after the decimal point gives the original position in the vector.

There are six things that you can use to subset a vector:

- **Positive integers** return elements at the specified positions:

```
x[c(3, 1)]
#> [1] 3.3 2.1
x[order(x)]
#> [1] 2.1 3.3 4.2 5.4

Duplicated indices yield duplicated values
x[c(1, 1)]
#> [1] 2.1 2.1

Real numbers are silently truncated to integers
x[c(2.1, 2.9)]
#> [1] 4.2 4.2
```

- **Negative integers** omit elements at the specified positions:

```
x[-c(3, 1)]
#> [1] 4.2 5.4
```

You can't mix positive and negative integers in a single subset:

```
x[c(-1, 2)]
#> Error in x[c(-1, 2)]:
#> only 0's may be mixed with negative subscripts
```

- **Logical vectors** select elements where the corresponding logical value is TRUE. This is probably the most useful type of subsetting because you can write an expression that creates the logical vector:

```
x[c(TRUE, TRUE, FALSE, FALSE)]
#> [1] 2.1 4.2
x[x > 3]
#> [1] 4.2 3.3 5.4
```

If the logical vector is shorter than the vector being subsetted, it will be silently **recycled** to be the same length.

```
x[c(TRUE, FALSE)]
#> [1] 2.1 3.3
Equivalent to
x[c(TRUE, FALSE, TRUE, FALSE)]
#> [1] 2.1 3.3
```

A missing value in the index always yields a missing value in the output:

```
x[c(TRUE, TRUE, NA, FALSE)]
#> [1] 2.1 4.2 NA
```

- **Nothing** returns the original vector. This is not useful for 1d vectors, as you'll see shortly, is very useful for matrices, data frames, and arrays. It can also be useful in conjunction with assignment.

```
x[]
#> [1] 2.1 4.2 3.3 5.4
```

- **Zero** returns a zero-length vector. This is not something you usually do on purpose, but it can be helpful for generating test data.

```
x[0]
#> numeric(0)
```

- If the vector is named, you can also use **character vectors** to return elements with matching names.

```
(y <- setNames(x, letters[1:4]))
#> a b c d
#> 2.1 4.2 3.3 5.4
y[c("d", "c", "a")]
#> d c a
#> 5.4 3.3 2.1

Like integer indices, you can repeat indices
y[c("a", "a", "a")]
#> a a a
#> 2.1 2.1 2.1

When subsetting with [, names are always matched exactly
z <- c(abc = 1, def = 2)
z[c("a", "d")]
#> <NA> <NA>
#> NA NA
```

## 4.2.2 Lists

Subsetting a list works in the same way as subsetting an atomic vector. Using `[` will always return a list; `[[` and `$`, as described in Section 4.3, let you pull out the components of the list.

## 4.2.3 Matrices and arrays

You can subset higher-dimensional structures in three ways:

- With multiple vectors.
- With a single vector.
- With a matrix.

The most common way of subsetting matrices (2d) and arrays ( $>2$ d) is a simple generalisation of 1d subsetting: you supply a 1d index for each dimension, separated by a comma. Blank subsetting is now useful because it lets you keep all rows or all columns.

```
a <- matrix(1:9, nrow = 3)
colnames(a) <- c("A", "B", "C")
a[1:2,]
#> A B C
#> [1,] 1 4 7
#> [2,] 2 5 8
a[c(TRUE, FALSE, TRUE), c("B", "A")]
#> B A
#> [1,] 4 1
#> [2,] 6 3
a[0, -2]
#> A C
```

By default, `[` will simplify the results to the lowest possible dimensionality. You'll learn how to avoid this in Section 4.2.5.

Because matrices and arrays are just vectors with special attributes, you can subset them with a single vector, as if they were a 1d vector. Arrays in R are stored in column-major order:

```
vals <- outer(1:5, 1:5, FUN = "paste", sep = ", ")
vals
#> [,1] [,2] [,3] [,4] [,5]
#> [1,] "1,1" "1,2" "1,3" "1,4" "1,5"
#> [2,] "2,1" "2,2" "2,3" "2,4" "2,5"
#> [3,] "3,1" "3,2" "3,3" "3,4" "3,5"
#> [4,] "4,1" "4,2" "4,3" "4,4" "4,5"
#> [5,] "5,1" "5,2" "5,3" "5,4" "5,5"

vals[c(4, 15)]
#> [1] "4,1" "5,3"
```

You can also subset higher-dimensional data structures with an integer matrix (or, if named, a character matrix). Each row in the matrix specifies the location of one value, where each column corresponds to a dimension in the array being subsetted. This means that you use a 2 column matrix to subset a matrix, a 3 column matrix to subset a 3d array, and so on. The result is a vector of values:

```
select <- rbind(
 c(1, 1),
 c(3, 1),
 c(2, 4)
```

```
)
vals[select]
#> [1] "1,1" "3,1" "2,4"
```

#### 4.2.4 Data frames and tibbles

Data frames possess the characteristics of both lists and matrices: if you subset with a single vector, they behave like lists; if you subset with two vectors, they behave like matrices.

```
df <- data.frame(x = 1:3, y = 3:1, z = letters[1:3])

df[df$x == 2,]
#> x y z
#> 2 2 2 b
df[c(1, 3),]
#> x y z
#> 1 1 3 a
#> 3 3 1 c

There are two ways to select columns from a data frame
Like a list, which
df[c("x", "z")]
#> x z
#> 1 1 a
#> 2 2 b
#> 3 3 c
Like a matrix
df[, c("x", "z")]
#> x z
#> 1 1 a
#> 2 2 b
#> 3 3 c

There's an important difference if you select a single
column: matrix subsetting simplifies by default, list
subsetting does not.
str(df["x"])
#> 'data.frame': 3 obs. of 1 variable:
#> $ x: int 1 2 3
str(df[, "x"])
#> int [1:3] 1 2 3
```

Subsetting a tibble with [ always returns a tibble:

```
df <- tibble::tibble(x = 1:3, y = 3:1, z = letters[1:3])

str(df["x"])
#> Classes 'tbl_df', 'tbl' and 'data.frame': 3 obs. of 1 variable:
#> $ x: int 1 2 3
str(df[, "x"])
#> Classes 'tbl_df', 'tbl' and 'data.frame': 3 obs. of 1 variable:
#> $ x: int 1 2 3
```

### 4.2.5 Preserving dimensionality

By default, any subsetting 2d data structures with a single number, single name, or a logical vector containing a single TRUE will simplify the returned output, i.e. it will return an object with lower dimensionality. To preserve the original dimensionality, you must use `drop = FALSE`

- For matrices and arrays, any dimensions with length 1 will be dropped:

```
a <- matrix(1:4, nrow = 2)
str(a[1,])
#> int [1:2] 1 3

str(a[1, , drop = FALSE])
#> int [1, 1:2] 1 3
```

- Data frames with a single column will return just that column:

```
df <- data.frame(a = 1:2, b = 1:2)
str(df[, "a"])
#> int [1:2] 1 2

str(df[, "a", drop = FALSE])
#> 'data.frame': 2 obs. of 1 variable:
#> $ a: int 1 2
```

- Tibbles default to `drop = FALSE`, and `[` will never return a single vector.

The default `drop = TRUE` behaviour is a common source of bugs in functions: you check your code with a data frame or matrix with multiple columns, and it works. Six months later you (or someone else) uses it with a single column data frame and it fails with a mystifying error. When writing functions, get in the habit of always using `drop = FALSE` when subsetting a 2d object.

Factor subsetting also has a `drop` argument, but the meaning is rather different. It controls whether or not levels are preserved (not the dimensionality), and it defaults to `FALSE` (levels are preserved, not simplified by default). If you find you are using `drop = TRUE` a lot it's often a sign that you should be using a character vector instead of a factor.

```
z <- factor(c("a", "b"))
z[1]
#> [1] a
#> Levels: a b
z[1, drop = TRUE]
#> [1] a
#> Levels: a
```

### 4.2.6 Exercises

1. Fix each of the following common data frame subsetting errors:

```
mtcars[mtcars$cyl = 4,]
mtcars[-1:4,]
mtcars[mtcars$cyl <= 5]
mtcars[mtcars$cyl == 4 | 6,]
```

2. Why does the following code yield five missing values? (Hint: why is it different from `x[NA_real_]`?)

```
x <- 1:5
x[NA]
#> [1] NA NA NA NA NA
```

3. What does `upper.tri()` return? How does subsetting a matrix with it work? Do we need any additional subsetting rules to describe its behaviour?

```
x <- outer(1:5, 1:5, FUN = "*")
x[upper.tri(x)]
```

4. Why does `mtcars[1:20]` return an error? How does it differ from the similar `mtcars[1:20, ]`?
5. Implement your own function that extracts the diagonal entries from a matrix (it should behave like `diag(x)` where `x` is a matrix).
6. What does `df[is.na(df)] <- 0` do? How does it work?

## 4.3 Selecting a single element

There are two other subsetting operators: `[[` and `$`. `[[` is used for extracting single items, and `x$y` is a useful shorthand for `x[["y"]]`.

### 4.3.1 [[

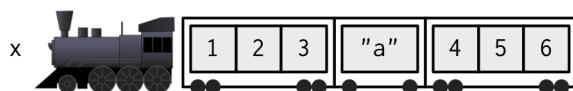
`[[` is most important working with lists because subsetting a list with `[` always returns a smaller list. To help make this easier to understand we can use a metaphor:

“If list `x` is a train carrying objects, then `x[[5]]` is the object in car 5; `x[4:6]` is a train of cars 4-6.”

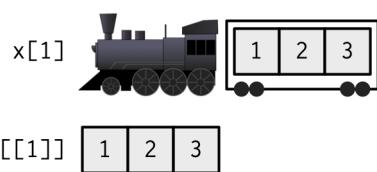
— @RLangTip, <https://twitter.com/RLangTip/status/268375867468681216>

Let's make a simple list and draw it as a train:

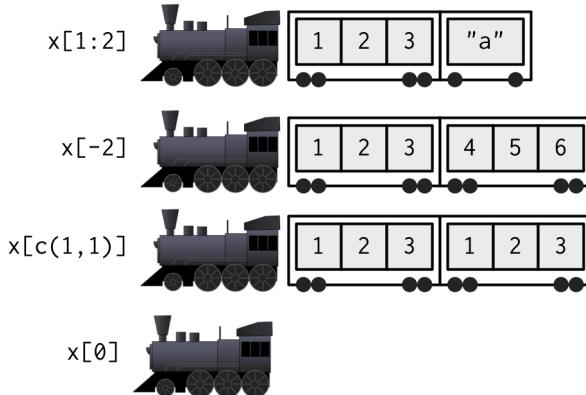
```
x <- list(1:3, "a", 4:6)
```



When extracting a single element, you have two options: you can create a smaller train, or you can extract the contents of a carriage. This is the difference between `[` and `[[`:



When extracting multiple elements (or zero!), you have to make a smaller train:



Because it can return only a single item, you must use `[[` with either a single positive integer or a string. If you use a vector with `[`, it will subset recursively:

```
b <- list(a = list(b = list(c = list(d = 1))))
b[[c("a", "b", "c", "d")]]
#> [1] 1

Equivalent to
b[["a"]][["b"]][["c"]][["d"]]
#> [1] 1
```

`[[` is crucial for working with lists, but I recommend using it whenever you want your code to clearly express that it's working with a single item. That frequently arises in for loops, e.g., instead of writing:

```
for (i in 2:length(x)) {
 out[i] <- fun(x[i], out[i - 1])
}
```

It's better to write:

```
for (i in 2:length(x)) {
 out[[i]] <- fun(x[[i]], out[[i - 1]])
}
```

That reinforces to the reader that you expect to get and set individual values.

### 4.3.2 \$

`$` is a shorthand operator: `x$y` is roughly equivalent to `x[["y"]]`. It's often used to access variables in a data frame, as in `mtcars$cyl` or `diamonds$carat`. One common mistake with `$` is to use it when you have the name of a column stored in a variable:

```
var <- "cyl"
Doesn't work - mtcars$var translated to mtcars[["var"]]
mtcars$var
#> NULL

Instead use [[
mtcars[[var]]
#> [1] 6 6 4 6 8 6 8 4 4 6 6 8 8 8 8 8 4 4 4 4 8 8 8 8 4 4 4 8 6 8 4
```

There's one important difference between `$` and `[[`. `$` does partial matching:

```
x <- list(abc = 1)
x$a
#> [1] 1
x[["a"]]
#> NULL
```

To help avoid this behaviour I highly recommend setting the global option `warnPartialMatchDollar` to `TRUE`:

```
options(warnPartialMatchDollar = TRUE)
x$a
#> Warning in x$a: partial match of 'a' to 'abc'
#> [1] 1
```

(For data frames, you can also avoid this problem by using tibbles instead: they never do partial matching.)

### 4.3.3 Missing/out of bounds indices

It's useful to understand what happens with `[` and `[[` when you use an "invalid" index. The following tables summarise what happen when you subset a logical vector, list, and `NULL` with an out-of-bounds value (OOB), a missing value (i.e `NA_integer_`), and a zero-length object (like `NULL` or `logical()`) with `[` and `[[`. Each cell shows the result of subsetting the data structure named in the row by the type of index described in the column. I've only shown the results for logical vectors, but other atomic vectors behave similarly, returning elements of the same type.

| row[col]             | Zero-length             | OOB                     | Missing                 |
|----------------------|-------------------------|-------------------------|-------------------------|
| <code>NULL</code>    | <code>NULL</code>       | <code>NULL</code>       | <code>NULL</code>       |
| <code>Logical</code> | <code>logical(0)</code> | <code>NA</code>         | <code>NA</code>         |
| <code>List</code>    | <code>list()</code>     | <code>list(NULL)</code> | <code>list(NULL)</code> |

With `[`, it doesn't matter whether the OOB index is a position or a name, but it does for `[[`:

| row[[col]]          | Zero-length        | OOB (int)          | OOB (chr)          | Missing            |
|---------------------|--------------------|--------------------|--------------------|--------------------|
| <code>NULL</code>   | <code>NULL</code>  | <code>NULL</code>  | <code>NULL</code>  | <code>NULL</code>  |
| <code>Atomic</code> | <code>Error</code> | <code>Error</code> | <code>Error</code> | <code>Error</code> |
| <code>List</code>   | <code>Error</code> | <code>Error</code> | <code>NULL</code>  | <code>NULL</code>  |

If the input vector is named, then the names of OOB, missing, or `NULL` components will be "`<NA>`".

The inconsistency of the `[[` table above lead to the development of `purrr::pluck()` and `purrr::chuck()`. `pluck()` always returns `NULL` (or the value of the `.default` argument) when the element is missing; `chuck()` always throws an error:

| pluck(row, col)     | Zero-length       | OOB (int)         | OOB (chr)         | Missing           |
|---------------------|-------------------|-------------------|-------------------|-------------------|
| <code>NULL</code>   | <code>NULL</code> | <code>NULL</code> | <code>NULL</code> | <code>NULL</code> |
| <code>Atomic</code> | <code>NULL</code> | <code>NULL</code> | <code>NULL</code> | <code>NULL</code> |
| <code>List</code>   | <code>NULL</code> | <code>NULL</code> | <code>NULL</code> | <code>NULL</code> |

| chuck(row, col) | Zero-length | OOB (int) | OOB (chr) | Missing |
|-----------------|-------------|-----------|-----------|---------|
| NULL            | Error       | Error     | Error     | Error   |
| Atomic          | Error       | Error     | Error     | Error   |
| List            | Error       | Error     | Error     | Error   |

The behaviour of `pluck()` makes it well suited for indexing into deeply nested data structures where the component you want does not exist always exist (as is common when working with JSON data from web APIs). `pluck()` also allows you to mingle integer and character indexes, and to provide an alternative default value if the item does not exist:

```
x <- list(
 a = list(1, 2, 3),
 b = list(3, 4, 5)
)

purrr::pluck(x, "a", 1)
#> [1] 1

purrr::pluck(x, "c", 1)
#> NULL

purrr::pluck(x, "c", 1, .default = NA)
#> [1] NA
```

#### 4.3.4 `@` and `slot()`

There are also two additional subsetting operators that are needed for S4 objects: `@` (equivalent to `$`), and `slot()` (equivalent to `[[]]`). `@` is more restrictive than `$` in that it will return an error if the slot does not exist. These are described in more detail in S4.

#### 4.3.5 Exercises

1. Brainstorm as many ways as possible to extract the third value from the `cyl` variable in the `mtcars` dataset.
2. Given a linear model, e.g., `mod <- lm(mpg ~ wt, data = mtcars)`, extract the residual degrees of freedom. Extract the R squared from the model summary (`summary(mod)`)

## 4.4 Subsetting and assignment

All subsetting operators can be combined with assignment to modify selected values of the input vector.

```
x <- 1:5
x[c(1, 2)] <- 2:3
x
#> [1] 2 3 3 4 5

The length of the LHS needs to match the RHS
x[-1] <- 4:1
x
```

```
#> [1] 2 4 3 2 1

Duplicated indices go unchecked and may be problematic
x[c(1, 1)] <- 2:3
x
#> [1] 3 4 3 2 1

You can't combine integer indices with NA
x[c(1, NA)] <- c(1, 2)
#> Error in x[c(1, NA)] <- c(1, 2):
#> NAs are not allowed in subscripted assignments

But you can combine logical indices with NA
(where they're treated as false).
x[c(T, F, NA)] <- 1
x
#> [1] 1 4 3 1 1

This is mostly useful when conditionally modifying vectors
df <- data.frame(a = c(1, 10, NA))
df$a[df$a < 5] <- 0
df$a
#> [1] 0 10 NA
```

Subsetting with nothing can be useful in conjunction with assignment because it will preserve the structure of the original object. Compare the following two expressions. In the first, `mtcars` will remain as a data frame. In the second, `mtcars` will become a list.

```
mtcars[] <- lapply(mtcars, as.integer)
mtcars <- lapply(mtcars, as.integer)
```

With lists, you can use `[[ + assignment + NULL` to remove components from a list. To add a literal `NULL` to a list, use `[` and `list(NULL)`:

```
x <- list(a = 1, b = 2)
x[["b"]] <- NULL
str(x)
#> List of 1
#> $ a: num 1

y <- list(a = 1)
y[["b"]] <- list(NULL)
str(y)
#> List of 2
#> $ a: num 1
#> $ b: NULL
```

## 4.5 Applications

The basic principles described above give rise to a wide variety of useful applications. Some of the most important are described below. Many of these basic techniques are wrapped up into more concise functions (e.g., `subset()`, `merge()`, `dplyr::arrange()`), but it is useful to understand how they are implemented with basic subsetting. This will allow you to adapt to new situations not handled by existing functions.

### 4.5.1 Lookup tables (character subsetting)

Character matching provides a powerful way to make lookup tables. Say you want to convert abbreviations:

```
x <- c("m", "f", "u", "f", "f", "m", "m")
lookup <- c(m = "Male", f = "Female", u = NA)
lookup[x]
#> m f u f f m m
#> "Male" "Female" NA "Female" "Female" "Male" "Male"
unname(lookup[x])
#> [1] "Male" "Female" NA "Female" "Female" "Male" "Male"
```

If you don't want names in the result, use `unname()` to remove them.

### 4.5.2 Matching and merging by hand (integer subsetting)

You may have a more complicated lookup table which has multiple columns of information. Suppose we have a vector of integer grades, and a table that describes their properties:

```
grades <- c(1, 2, 2, 3, 1)

info <- data.frame(
 grade = 3:1,
 desc = c("Excellent", "Good", "Poor"),
 fail = c(F, F, T)
)
```

We want to duplicate the info table so that we have a row for each value in `grades`. An elegant way to do this is by combining `match()` and integer subsetting:

```
id <- match(grades, info$grade)
info[id,]
#> grade desc fail
#> 3 1 Poor TRUE
#> 2 2 Good FALSE
#> 2.1 2 Good FALSE
#> 1 3 Excellent FALSE
#> 3.1 1 Poor TRUE
```

If you have multiple columns to match on, you'll need to first collapse them to a single column (with e.g. `interaction()`), but typically you are better off switching to a function design specifically for joining multiple tables like `merge()`, or `dplyr::left_join()`.

### 4.5.3 Random samples/bootstraps (integer subsetting)

You can use integer indices to perform random sampling or bootstrapping of a vector or data frame. `sample()` generates a vector of indices, then subsetting accesses the values:

```
df <- data.frame(x = c(1, 2, 3, 1, 2), y = 5:1, z = letters[1:5])

Randomly reorder
df[sample(nrow(df)),]
#> x y z
#> 1 1 5 a
```

```
#> 4 1 2 d
#> 2 2 4 b
#> 5 2 1 e
#> 3 3 3 c

Select 3 random rows
df[sample(nrow(df), 3),]
#> x y z
#> 3 3 3 c
#> 2 2 4 b
#> 1 1 5 a

Select 6 bootstrap replicates
df[sample(nrow(df), 6, replace = TRUE),]
#> x y z
#> 4 1 2 d
#> 4.1 1 2 d
#> 5 2 1 e
#> 1 1 5 a
#> 1.1 1 5 a
#> 2 2 4 b
```

The arguments of `sample()` control the number of samples to extract, and whether sampling is performed with or without replacement.

#### 4.5.4 Ordering (integer subsetting)

`order()` takes a vector as input and returns an integer vector describing how the subsetted vector should be ordered:

```
x <- c("b", "c", "a")
order(x)
#> [1] 3 1 2
x[order(x)]
#> [1] "a" "b" "c"
```

To break ties, you can supply additional variables to `order()`, and you can change from ascending to descending order using `decreasing = TRUE`. By default, any missing values will be put at the end of the vector; however, you can remove them with `na.last = NA` or put at the front with `na.last = FALSE`.

For two or more dimensions, `order()` and integer subsetting makes it easy to order either the rows or columns of an object:

```
Randomly reorder df
df2 <- df[sample(nrow(df)), 3:1]
df2
#> z y x
#> 3 c 3 3
#> 1 a 5 1
#> 2 b 4 2
#> 4 d 2 1
#> 5 e 1 2

df2[order(df2$x),]
#> z y x
```

```
#> 1 a 5 1
#> 4 d 2 1
#> 2 b 4 2
#> 5 e 1 2
#> 3 c 3 3
df2[, order(names(df2))]
#> x y z
#> 3 3 3 c
#> 1 1 5 a
#> 2 2 4 b
#> 4 1 2 d
#> 5 2 1 e
```

You can sort vectors directly with `sort()`, or use `dplyr::arrange()` or similar to sort a data frame.

#### 4.5.5 Expanding aggregated counts (integer subsetting)

Sometimes you get a data frame where identical rows have been collapsed into one and a count column has been added. `rep()` and integer subsetting make it easy to uncollapse the data by subsetting with a repeated row index:

```
df <- data.frame(x = c(2, 4, 1), y = c(9, 11, 6), n = c(3, 5, 1))
rep(1:nrow(df), df$n)
#> [1] 1 1 1 2 2 2 2 2 3

df[rep(1:nrow(df), df$n),]
#> x y n
#> 1 2 9 3
#> 1.1 2 9 3
#> 1.2 2 9 3
#> 2 4 11 5
#> 2.1 4 11 5
#> 2.2 4 11 5
#> 2.3 4 11 5
#> 2.4 4 11 5
#> 3 1 6 1
```

#### 4.5.6 Removing columns from data frames (character subsetting)

There are two ways to remove columns from a data frame. You can set individual columns to `NULL`:

```
df <- data.frame(x = 1:3, y = 3:1, z = letters[1:3])
df$z <- NULL
```

Or you can subset to return only the columns you want:

```
df <- data.frame(x = 1:3, y = 3:1, z = letters[1:3])
df[c("x", "y")]
#> x y
#> 1 1 3
#> 2 2 2
#> 3 3 1
```

If you only know the columns you don't want, use set operations to work out which columns to keep:

```
df[setdiff(names(df), "z")]
#> x y
#> 1 1 3
#> 2 2 2
#> 3 3 1
```

#### 4.5.7 Selecting rows based on a condition (logical subsetting)

Because it allows you to easily combine conditions from multiple columns, logical subsetting is probably the most commonly used technique for extracting rows out of a data frame.

```
mtcars[mtcars$gear == 5,]
#> mpg cyl disp hp drat wt qsec vs am gear carb
#> 27 26.0 4 120.3 91 4.43 2.14 16.7 0 1 5 2
#> 28 30.4 4 95.1 113 3.77 1.51 16.9 1 1 5 2
#> 29 15.8 8 351.0 264 4.22 3.17 14.5 0 1 5 4
#> 30 19.7 6 145.0 175 3.62 2.77 15.5 0 1 5 6
#> 31 15.0 8 301.0 335 3.54 3.57 14.6 0 1 5 8

mtcars[mtcars$gear == 5 & mtcars$cyl == 4,]
#> mpg cyl disp hp drat wt qsec vs am gear carb
#> 27 26.0 4 120.3 91 4.43 2.14 16.7 0 1 5 2
#> 28 30.4 4 95.1 113 3.77 1.51 16.9 1 1 5 2
```

Remember to use the vector boolean operators `&` and `|`, not the short-circuiting scalar operators `&&` and `||` which are more useful inside if statements. Don't forget De Morgan's laws ([http://en.wikipedia.org/wiki/De\\_Morgan%27s\\_laws](http://en.wikipedia.org/wiki/De_Morgan%27s_laws)), which can be useful to simplify negations:

- `!(X & Y)` is the same as `!X | !Y`
- `!(X | Y)` is the same as `!X & !Y`

For example, `!(X & !(Y | Z))` simplifies to `!X | !!(Y|Z)`, and then to `!X | Y | Z`.

#### 4.5.8 Boolean algebra vs. sets (logical & integer subsetting)

It's useful to be aware of the natural equivalence between set operations (integer subsetting) and boolean algebra (logical subsetting). Using set operations is more effective when:

- You want to find the first (or last) TRUE.
- You have very few TRUEs and very many FALSEs; a set representation may be faster and require less storage.

`which()` allows you to convert a boolean representation to an integer representation. There's no reverse operation in base R but we can easily create one:

```
x <- sample(10) < 4
which(x)
#> [1] 2 5 8

unwhich <- function(x, n) {
 out <- rep_len(FALSE, n)
 out[x] <- TRUE
 out
}
```

```
unwhich(which(x), 10)
#> [1] FALSE TRUE FALSE FALSE TRUE FALSE FALSE TRUE FALSE FALSE
```

Let's create two logical vectors and their integer equivalents and then explore the relationship between boolean and set operations.

```
(x1 <- 1:10 %% 2 == 0)
#> [1] FALSE TRUE FALSE TRUE FALSE TRUE FALSE TRUE FALSE TRUE
(x2 <- which(x1))
#> [1] 2 4 6 8 10
(y1 <- 1:10 %% 5 == 0)
#> [1] FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE
(y2 <- which(y1))
#> [1] 5 10

X & Y <-> intersect(x, y)
x1 & y1
#> [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE
intersect(x2, y2)
#> [1] 10

X | Y <-> union(x, y)
x1 | y1
#> [1] FALSE TRUE FALSE TRUE TRUE TRUE FALSE TRUE FALSE TRUE
union(x2, y2)
#> [1] 2 4 6 8 10 5

X & !Y <-> setdiff(x, y)
x1 & !y1
#> [1] FALSE TRUE FALSE TRUE FALSE TRUE FALSE TRUE FALSE FALSE
setdiff(x2, y2)
#> [1] 2 4 6 8

xor(X, Y) <-> setdiff(union(x, y), intersect(x, y))
xor(x1, y1)
#> [1] FALSE TRUE FALSE TRUE TRUE FALSE TRUE FALSE FALSE
setdiff(union(x2, y2), intersect(x2, y2))
#> [1] 2 4 6 8 5
```

When first learning subsetting, a common mistake is to use `x[which(y)]` instead of `x[y]`. Here the `which()` achieves nothing: it switches from logical to integer subsetting but the result will be exactly the same. In more general cases, there are two important differences.

- When the logical vector contains `NA`, logical subsetting replaces these values by `NA` while `which()` drops these values. It's not uncommon to use `which()` for this side-effect, but that's
- `x[-which(y)]` is **not** equivalent to `x[!y]`: if `y` is all `FALSE`, `which(y)` will be `integer(0)` and `-integer(0)` is still `integer(0)`, so you'll get no values, instead of all values.

In general, avoid switching from logical to integer subsetting unless you want, for example, the first or last `TRUE` value.

### 4.5.9 Exercises

1. How would you randomly permute the columns of a data frame? (This is an important technique in random forests.) Can you simultaneously permute the rows and columns in one step?
2. How would you select a random sample of  $m$  rows from a data frame? What if the sample had to be contiguous (i.e., with an initial row, a final row, and every row in between)?
3. How could you put the columns in a data frame in alphabetical order?

## 4.6 Answers

1. Positive integers select elements at specific positions, negative integers drop elements; logical vectors keep elements at positions corresponding to TRUE; character vectors select elements with matching names.
2. `[` selects sub-lists. It always returns a list; if you use it with a single positive integer, it returns a list of length one. `[[` selects an element within a list. `$` is a convenient shorthand: `x$y` is equivalent to `x[["y"]]`.
3. Use `drop = FALSE` if you are subsetting a matrix, array, or data frame and you want to preserve the original dimensions. You should almost always use it when subsetting inside a function.
4. If `x` is a matrix, `x[] <- 0` will replace every element with 0, keeping the same number of rows and columns. `x <- 0` completely replaces the matrix with the value 0.
5. A named character vector can act as a simple lookup table: `c(x = 1, y = 2, z = 3)[c("y", "z", "x")]`



# Chapter 5

## Functions

### 5.1 Introduction

If you’re reading this book, you’ve probably already created many R functions and know how to use them to reduce duplication in your code. In this chapter, you’ll learn how to turn that informal, working knowledge into more rigorous, theoretical understanding. And while you’ll see some interesting tricks and techniques along the way, keep in mind that what you’ll learn here will be important for understanding the more advanced topics discussed later in the book.

### Quiz

Answer the following questions to see if you can safely skip this chapter. You can find the answers in Section 5.10.

1. What are the three components of a function?
2. What does the following code return?

```
x <- 10
f1 <- function(x) {
 function() {
 x + 10
 }
}
f1(1)()
```

3. How would you usually write this code?

```
^+(1, ^*(2, 3))
```

4. How could you make this call easier to read?

```
mean(, TRUE, x = c(1:10, NA))
```

5. Does the following code throw an error when executed? Why/why not?

```
f2 <- function(a, b) {
 a * 10
}
f2(10, stop("This is an error!"))
```

6. What is an infix function? How do you write it? What's a replacement function? How do you write it?
7. How do you ensure that cleanup action occurs regardless of how a function exits?

## Outline

- Section 5.2 describes the basics of creating a function, the three main components of a function, and the exception to many function rules: primitive functions (which are implemented in C, not R).
- Section 5.3 discusses the strengths and weaknesses of the three forms of function composition commonly used in R code.
- Section 5.4 shows you how R finds the value associated with a given name, i.e. the rules of lexical scoping.
- Section 5.5 is devoted to an important property of function arguments: they are only evaluated when used for the first time.
- Section 5.6 discusses the special ... argument, which allows you to pass on extra arguments to another function.
- Section 5.7 discusses the two primary ways that a function can exit, and how to define an exit handler, code that is run on exit, regardless of what triggers it.
- Section 5.8 shows you the various ways in which R disguises ordinary function calls, and how you can use the standard prefix form to better understand what's going on.

## 5.2 Function fundamentals

To understand functions in R you need to internalise two important ideas:

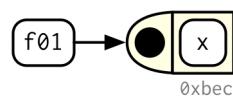
- Functions are objects, just as vectors are objects.
- Functions can be broken down into three components: arguments, body, and environment.

There are exceptions to every rule, and in this case, there is a small selection of “primitive” base functions that are implemented purely in C.

### 5.2.1 First-class functions

The most important thing to understand about R is that functions are objects in their own right, a language property often called “first-class functions”. Unlike in many other languages, there is no special syntax for defining and naming a function: you simply create a function object (with `function`) and bind it to a name with `<-`:

```
f01 <- function(x) {
 sin(1 / x ^ 2)
}
```



While you almost always create a function and then bind it to a name, the binding step is not compulsory. If you choose not to give a function a name, you get an **anonymous function**. This is useful when it's not worth the effort to figure out a name:

```
lapply(mtcars, function(x) length(unique(x)))
Filter(function(x) !is.numeric(x), mtcars)
integrate(function(x) sin(x) ^ 2, 0, pi)
```

A final option is to put functions in a list:

```
funcs <- list(
 half = function(x) x / 2,
 double = function(x) x * 2
)

funcs$double(10)
#> [1] 20
```

In R, you'll often see functions called **closures**. This name reflects the fact that R functions capture, or **enclose**, their environments.

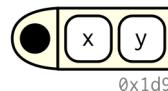
```
typeof(f01)
#> [1] "closure"
```

## 5.2.2 Function components

A function has three parts:

- The `formals()`, the list of arguments that control how you call the function.
- The `body()`, the code inside the function.
- The `environment()`, the data structure that determines how the function finds the values associated with the names.

I'll draw functions as in the following diagram. The black dot on the left is the environment. The two blocks to the right are the function arguments. I won't draw the body, because it's usually large, and doesn't help you understand the "shape" of the function.



While the formals and body are specified explicitly when you create a function, the environment is specified implicitly, based on *where* you defined the function. The function environment always exists, but it is only printed when the function isn't defined in the global environment.

```
f02 <- function(x) {
 # A comment
 x ^ 2
}

formals(f02)
#> $x

body(f02)
#> {
#> x^2
```

```
#> }

environment(f02)
#> <environment: R_GlobalEnv>
```

Like all objects in R, functions can also possess any number of additional `attributes()`. One attribute used by base R is “`srcref`”, short for source reference. It points to the source code used to create the function. The `srcref` is used for printing because, unlike `body()`, it contains code comments and other formatting.

```
attr(f02, "srcref")
#> function(x) {
#> # A comment
#> x ^ 2
#> }
```

### 5.2.3 Primitive functions

There is one exception to the rule that a function has three components. Primitive functions, like `sum()` and `[`, call C code directly.

```
sum
#> function (...) .Primitive("sum")
`[`
#> .Primitive("[")
```

They have type “builtin” or “special”:

```
typeof(sum)
#> [1] "builtin"
typeof(`[`)
#> [1] "special"
```

These functions exist primarily in C, not R, so their `formals()`, `body()`, and `environment()` are all `NULL`:

```
formals(sum)
#> NULL
body(sum)
#> NULL
environment(sum)
#> NULL
```

Primitive functions are only found in the base package. While they have certain performance advantages, this benefit comes at a price: they are harder to write. For this reason, R-core generally avoids creating them unless there is no other option.

### 5.2.4 Exercises

- Given a function, like “`mean`”, `match.fun()` lets you find a function. Given a function, can you find its name? Why doesn’t that make sense in R?
- It’s possible (although typically not useful) to call an anonymous function. Which of the two approaches below is correct? Why?

```
function(x) 3()
#> function(x) 3()
```

```
(function(x) 3)()
#> [1] 3
```

3. A good rule of thumb is that an anonymous function should fit on one line and shouldn't need to use {}.  
Review your code. Where could you have used an anonymous function instead of a named function?  
Where should you have used a named function instead of an anonymous function?
4. What function allows you to tell if an object is a function? What function allows you to tell if a function is a primitive function?
5. This code makes a list of all functions in the base package.

```
objs <- mget(ls("package:base"), inherits = TRUE)
funs <- Filter(is.function, objs)
```

Use it to answer the following questions:

- a. Which base function has the most arguments?
- b. How many base functions have no arguments? What's special about those functions?
- c. How could you adapt the code to find all primitive functions?
6. What are the three important components of a function?
7. When does printing a function not show the environment it was created in?

## 5.3 Function composition

Base R provides two ways to compose multiple function calls. For example, imagine you want to compute the population standard deviation using `sqrt()` and `mean()` as building blocks:

```
square <- function(x) x^2
deviation <- function(x) x - mean(x)
```

You either nest the function calls:

```
x <- runif(100)

sqrt(mean(square(deviation(x))))
#> [1] 0.274
```

Or you save the intermediate results as variables:

```
out <- deviation(x)
out <- square(out)
out <- mean(out)
out <- sqrt(out)
out
#> [1] 0.274
```

The magrittr package (Bache and Wickham 2014) provides a third option: the binary operator `%>%`, which is called the pipe and is pronounced as “and then”.

```
library(magrittr)

x %>%
 deviation() %>%
 square() %>%
```

```
mean() %>%
 sqrt()
#> [1] 0.274
```

`x %>% f()` is equivalent to `f(x)`; `x %>% f(y)` is equivalent to `f(x, y)`. The pipe is related to **tacit** or **point-free programming**<sup>1</sup>. In this style of programming, you don't explicitly refer to variables. Instead, you focus on the high-level composition of functions rather than the low-level flow of data; the focus is on what's being done (the verbs), rather than on what's being modified (the nouns). This style is common in Haskell and F#, the main inspiration for magrittr, and is the default style in stack based programming languages like Forth and Factor.

Each of the three options has its own strengths and weaknesses:

- Nesting, `f(g(x))`, is concise, and well suited for short sequences. But longer sequences are hard to read because they are read inside out and right to left. As a result, arguments can get spread out over long distances creating the “Dagwood sandwich ([https://en.wikipedia.org/wiki/Dagwood\\_sandwich](https://en.wikipedia.org/wiki/Dagwood_sandwich))” problem.
- Intermediate objects, `y <- f(x); g(y)`, requires you to name intermediate objects. This is a strength when objects are important, but a weakness when values are truly intermediate.
- Piping, `x %>% f() %>% g()`, allows you to read code in straightforward left-to-right fashion and doesn't require you to name intermediate objects. But you can only use it with linear sequences of transformations of a single object. It also requires an additional third party package and assumes that the reader understands piping.

Most code will use a combination of all three styles. Piping is more common in data analysis code, as much of an analysis consists of a sequence of transformations of an object (like a data frame or plot). I tend to use piping infrequently in packages; not because it is a bad idea, but because it's often a less natural fit.

## 5.4 Lexical scoping

In Names and values, we discussed assignment, the act of binding a name to a value. Here we'll discuss **scoping**, the act of finding the value associated with a name.

The basic rules of scoping are quite intuitive, and you've probably already internalised them, even if you never explicitly studied them. For example, what will the following code return, 10 or 20?

```
x <- 10
g01 <- function() {
 x <- 20
 x
}

g01()
```

In this section, you'll learn the formal rules of scoping as well as some of its more subtle details. A deeper understanding of scoping will help you to use more advanced functional programming tools, and eventually, even to write tools that translate R code into other languages.

R uses **lexical scoping**<sup>3</sup>: it looks up the values of names based on how a function is defined, not how it is

<sup>1</sup> Point-free programming is related to point-free topology. It's just a coincidence that many forms of point-free programming use `.` extensively. Point-free programming is sometimes humorously called pointless programming.

<sup>2</sup>I'll “hide” the answers to these challenges in the footnotes. Try solving them before looking at the answer; this will help you to better remember the correct answer. In this case, `g01()` will return 20.

<sup>3</sup> Functions that automatically quote one or more arguments (sometimes called NSE functions) can override the default scoping rules to implement other varieties of scoping. You'll learn more about that in metaprogramming.

called. “Lexical” here is not the English adjective “relating to words or a vocabulary”. It’s a technical CS term that tells us that the scoping rules use a parse-time, rather than a run-time structure.

R’s lexical scoping follows four primary rules:

- Name masking
- Functions vs. variables
- A fresh start
- Dynamic lookup

### 5.4.1 Name masking

The basic principle of lexical scoping is that names defined inside a function mask names defined outside a function. This is illustrated in the following example.

```
x <- 10
y <- 20
g02 <- function() {
 x <- 1
 y <- 2
 c(x, y)
}
g02()
#> [1] 1 2
```

If a name isn’t defined inside a function, R looks one level up.

```
x <- 2
g03 <- function() {
 y <- 1
 c(x, y)
}
g03()
#> [1] 2 1
```

The same rules apply if a function is defined inside another function. First, R looks inside the current function. Then, it looks where that function was defined (and so on, all the way up to the global environment). Finally, it looks in other loaded packages.

Run the following code in your head, then confirm the result by running the code.<sup>4</sup>

```
x <- 1
g04 <- function() {
 y <- 2
 i <- function() {
 z <- 3
 c(x, y, z)
 }
 i()
}
g04()
```

The same rules also apply to functions created by other functions, which we’ll call **closures**. Closures will be described in more detail in ??; here we’ll focus on how they interact with scoping. The following function, g05(), returns a function. What do you think this function will return when it’s called?<sup>5</sup>

---

<sup>4</sup>g04() returns c(1, 2, 3).

<sup>5</sup>g06() returns c(10, 2).

```
x <- 10
y <- 20

g05 <- function() {
 y <- 2
 function() {
 c(x, y)
 }
}
g06 <- g05()
g06()
```

This seems a little magical: how does R know what the value of `y` is after `g05()` is returned? R knows because `g06()` preserves the environment where it was defined and that environment includes the value of `y`. You'll learn more about how environments work in 6).

### 5.4.2 Functions vs. variables

In R, functions are ordinary objects. This means the same scoping principles that apply to other objects also apply to functions:

```
g07 <- function(x) x + 1
g08 <- function() {
 g07 <- function(x) x + 100
 g07(10)
}
g08()
#> [1] 110
```

The rule gets a little more complicated when a name is bound to a function and a non-function in different environments. When you use a name in function call, R will ignore non-function objects while looking for that value. For example, here `g9` takes on two different values:

```
g09 <- function(x) x + 100
g10 <- function() {
 g09 <- 10
 g09(g09)
}
g10()
#> [1] 110
```

But using the same name for two different things will make for confusing code, and is best avoided!

### 5.4.3 A fresh start

What happens to values between invocations of a function? Consider the example below. What will happen the first time you run this function? What will happen the second time?<sup>6</sup> (If you haven't seen `exists()` before, it returns `TRUE` if there's a variable with that name and returns `FALSE` if not.)

```
g11 <- function() {
 if (!exists("a")) {
 a <- 1
 } else {
```

---

<sup>6</sup>`g11()` returns 1 every time it's called.

```
a <- a + 1
}
a
}

g11()
g11()
```

You might be surprised that `g11()` always returns the same value. This happens because every time a function is called a new environment is created to host its execution. This means that a function has no way to tell what happened the last time it was run; each invocation is completely independent. (We'll see some ways to get around this in Section 10.2.3)

#### 5.4.4 Dynamic lookup

Lexical scoping determines where to look for values, not when to look for them. R looks for values when the function is run, not when it's created. This means that the output of a function can differ depending on objects outside its environment:

```
g12 <- function() x + 1
x <- 15
g12()
#> [1] 16

x <- 20
g12()
#> [1] 21
```

This behaviour can be quite annoying. If you make a spelling mistake in your code, you won't get an error when you create the function, and you might not even get one when you run the function, depending on what variables are defined in the global environment.

One way to detect this problem is to use `codetools::findGlobals()`. This function lists all the external dependencies (unbound symbols) within a function:

```
codetools::findGlobals(g12)
#> [1] "+" "x"
```

Another way to solve the problem would be to manually change the environment of the function to the `emptyenv()`, an environment which contains nothing:

```
environment(g12) <- emptyenv()
g12()
#> Error in x + 1:
#> could not find function "+"
```

Both of these approaches reveal why this undesirable behaviour exists: R relies on lexical scoping to find *everything*, even the `+` operator. This provides a rather beautiful simplicity to R's scoping rules.

#### 5.4.5 Exercises

- What does the following code return? Why? Describe how each of the three `c`'s is interpreted.

```
c <- 10
c(c = c)
```

2. What are the four principles that govern how R looks for values?
3. What does the following function return? Make a prediction before running the code yourself.

```
f <- function(x) {
 f <- function(x) {
 f <- function() {
 x ^ 2
 }
 f() + 1
 }
 f(x) * 2
}
f(10)
```

## 5.5 Lazy evaluation

In R, function arguments are **lazily evaluated**: they're only evaluated if accessed. For example, this code doesn't generate an error because `x` is never used:

```
h01 <- function(x) {
 10
}
h01(stop("This is an error!"))
#> [1] 10
```

This is an important feature because it allows you to do things like include potentially expensive computations in function arguments that will only be evaluated if needed.

### 5.5.1 Forcing evaluation

To **compel** the evaluation of an argument, use `force()`:

```
h02 <- function(x) {
 force(x)
 10
}
h02(stop("This is an error!"))
#> Error in force(x):
#> This is an error!
```

It's usually not necessary to force evaluation. However, it is important for certain functional programming techniques, like the one we'll cover in detail in function operators. Here, I'm just going to show you what the basic issue is.

Consider this small but surprisingly tricky function. It takes a single argument `x`, and returns a function that returns `x` when called.

```
capture1 <- function(x) {
 function() {
 x
 }
}
```

The subtlety here is that the value of `x` will be captured not when you call `capture1()`, but when you call the function that `capture1()` returns:

```
x <- 10
h03 <- capture1(x)
h04 <- capture1(x)

h03()
#> [1] 10

x <- 20
h04()
#> [1] 20
```

Even more confusingly this only happens once: the value is locked in after you have called `h03()`/`h04()` for the first time.

```
x <- 30
h03()
#> [1] 10
h04()
#> [1] 20
```

This behaviour is a consequence of lazy evaluation. The `x` argument is evaluated once `h03()`/`h04()` is called, and then its value is cached. We can avoid the confusion by forcing `x`:

```
capture2 <- function(x) {
 force(x)

 function() {
 x
 }
}

x <- 10
h05 <- capture2(x)

x <- 20
h05()
#> [1] 10
```

## 5.5.2 Promises

Lazy evaluation is powered by a data structure called a **promise**, or (less commonly) a thunk. We'll come back to this data structure in metaprogramming because it's one of the features of R that makes it most interesting as a programming language.

A promise has three components:

- The expression, like `x + y` which gives rise to the delayed computation.
- The environment where the expression should be evaluated.
- The value, which is computed and cached when the promise is first accessed by evaluating the expression in the specified environment.

The value cache ensures that accessing the promise multiple times always returns the same value. For example, you can see in the following code that `runif(1)` is only evaluated once:

```

h06 <- function(x) {
 c(x, x, x)
}

h06(runif(1))
#> [1] 0.806 0.806 0.806

```

You can also create promises “by hand” using `delayedAssign()`:

```

delayedAssign("x", {print("Executing code"); runif(1)})
x
#> [1] "Executing code"
#> [1] 0.814
x
#> [1] 0.814

```

You’ll see this idea again in advanced bindings.

### 5.5.3 Default arguments

Thanks to lazy evaluation, default value can be defined in terms of other arguments, or even in terms of variables defined later in the function:

```

h07 <- function(x = 1, y = x * 2, z = a + b) {
 a <- 10
 b <- 100

 c(x, y, z)
}

h07()
#> [1] 1 2 110

```

Many base R functions use this technique, but I don’t recommend it. It makes code harder to understand because it requires that you know exactly *when* default arguments are evaluated in order to predict *what* they will evaluate to.

The evaluation environment is slightly different for default and user supplied arguments, as default arguments are evaluated inside the function. This means that seemingly identical calls can yield different results. It’s easiest to see this with an extreme example:

```

h08 <- function(x = ls()) {
 a <- 1
 x
}

ls() evaluated inside f:
h08()
#> [1] "a" "x"

ls() evaluated in global environment:
h08(ls())
#> [1] "f"

```

### 5.5.4 Missing arguments

If an argument has a default, you can determine if the value comes from the user or the default with `missing()`:

```
h09 <- function(x = 10) {
 list(missing(x), x)
}
str(h09())
#> List of 2
#> $: logi TRUE
#> $: num 10
str(h09(10))
#> List of 2
#> $: logi FALSE
#> $: num 10
```

`missing()` is best used sparingly. Take `sample()`, for example. How many arguments are required?

```
args(sample)
#> function (x, size, replace = FALSE, prob = NULL)
#> NULL
```

It looks like both `x` and `size` are required, but in fact `sample()` uses `missing()` to provide a default for `size` if it's not supplied. If I was to rewrite `sample` myself<sup>7</sup>, I'd use an explicit `NULL` to indicate that `size` can be supplied, but it's not required:

```
sample <- function(x, size = NULL, replace = FALSE, prob = NULL) {
 if (is.null(size)) {
 size <- length(x)
 }

 x[sample.int(length(x), size, replace = replace, prob = prob)]
}
```

You can make that pattern even simpler with a small helper. The infix `%||%` function uses the LHS if it's not null, otherwise it uses the RHS:

```
%||% <- function(lhs, rhs) {
 if (!is.null(lhs)) {
 lhs
 } else {
 rhs
 }
}

sample <- function(x, size = NULL, replace = FALSE, prob = NULL) {
 size <- size %||% length(x)
 x[sample.int(length(x), size, replace = replace, prob = prob)]
}
```

Because of lazy evaluation, you don't need to worry about unnecessary computation: the RHS of `%||%` will only be evaluated if the LHS is null.

---

<sup>7</sup>Note that this only implements one way of calling `sample()`: you can also call it with a single integer, like `sample(10)`. This unfortunately makes `sample()` prone to silent errors in situations like `sample(x[i])`.

### 5.5.5 Exercises

1. What important property of `&&` make `x_ok()` work?

```
x_ok <- function(x) {
 !is.null(x) && length(x) == 1 && x > 0
}

x_ok(NULL)
#> [1] FALSE
x_ok(1)
#> [1] TRUE
x_ok(1:3)
#> [1] FALSE
```

What is different with this code? Why is this behaviour undesirable here?

```
x_ok <- function(x) {
 !is.null(x) & length(x) == 1 & x > 0
}

x_ok(NULL)
#> logical(0)
x_ok(1)
#> [1] TRUE
x_ok(1:3)
#> [1] FALSE FALSE FALSE
```

2. The definition of `force()` is simple:

```
force
#> function (x)
#> x
#> <bytecode: 0x870cc8>
#> <environment: namespace:base>
```

Why is it better to `force(x)` instead of just `x`?

3. What does this function return? Why? Which principle does it illustrate?

```
f2 <- function(x = z) {
 z <- 100
 x
}
f2()
```

4. What does this function return? Why? Which principle does it illustrate?

```
y <- 10
f1 <- function(x = {y <- 1; 2}, y = 0) {
 c(x, y)
}
f1()
y
```

5. In `hist()`, the default value of `xlim` is `range(breaks)`, the default value for `breaks` is "Sturges", and

```
range("Sturges")
#> [1] "Sturges" "Sturges"
```

Explain how `hist()` works to get a correct `xlim` value.

6. Explain why this function works. Why is it confusing?

```
show_time <- function(x = stop("Error!")) {
 stop <- function(...) Sys.time()
 print(x)
}
show_time()
#> [1] "2018-09-22 05:51:06 UTC"
```

7. How many arguments are required when calling `library()`?

## 5.6 ... (dot-dot-dot)

Functions can have a special argument `...` (pronounced dot-dot-dot). If a function has this argument, it can take any number of additional arguments. In other programming languages, this type of argument is often called a varargs, or the function is said to be variadic.

Inside a function, you can use `...` to pass those additional arguments on to another function.

```
i01 <- function(y, z) {
 list(y = y, z = z)
}

i02 <- function(x, ...) {
 i01(...)
}

str(i02(x = 1, y = 2, z = 3))
#> List of 2
#> $ y: num 2
#> $ z: num 3
```

It's possible (but rarely useful) to refer to elements of `...` by their position, using a special form:

```
i03 <- function(...) {
 list(first = ..1, third = ..3)
}
str(i03(1, 2, 3))
#> List of 2
#> $ first: num 1
#> $ third: num 3
```

More often useful is `list(...)`, which evaluates the arguments and stores them in a list:

```
i04 <- function(...) {
 list(...)
}
str(i04(a = 1, b = 2))
#> List of 2
#> $ a: num 1
#> $ b: num 2
```

(See also `rlang::list2()` to support splicing and to silently ignore trailing commas, and `rlang::enquos()` to capture the unevaluated arguments, the topic of quasiquotation.)

There are two primary uses of `...`, both of which we'll come back to later in the book:

- If your function takes a function as an argument, you want some way to pass on additional arguments to that function. In this example, `lapply()` uses `...` to pass `na.rm` on to `mean()`:

```
x <- list(c(1, 3, NA), c(4, NA, 6))
str(lapply(x, mean, na.rm = TRUE))
#> List of 2
#> $: num 2
#> $: num 5
```

We'll come back to this technique in Section ??.

- If your function is an S3 generic, you need some way to allow methods to take arbitrary extra arguments. For example, take the `print()` function. There are different options for printing types of object, so there's no way for the print generic to prespecify every possible argument. Instead, it uses `...` to allow individual methods to have different arguments:

```
print(factor(letters), max.levels = 4)

print(y ~ x, showEnv = TRUE)
```

We'll come back to this use of `...` in Section 13.4.2.

Using `...` comes with two downsides:

- When you use it to pass arguments on to another function, you have to carefully explain to the user where those arguments go. This makes it hard to understand what you can do with functions like `lapply()` and `plot()`.
- Any misspelled arguments will not raise an error. This makes it easy for typos to go unnoticed:

```
sum(1, 2, NA, na_rm = TRUE)
#> [1] NA
```

`...` is a powerful tool, but be aware of the downsides.

### 5.6.1 Exercises

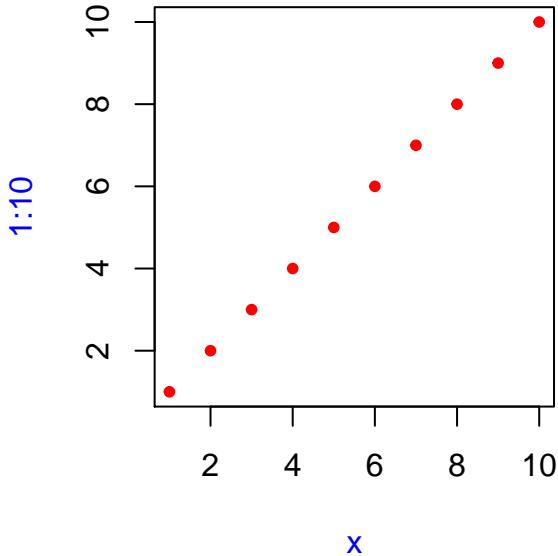
1. Explain the following results:

```
sum(1, 2, 3)
#> [1] 6
mean(1, 2, 3)
#> [1] 1

sum(1, 2, 3, na.omit = TRUE)
#> [1] 7
mean(1, 2, 3, na.omit = TRUE)
#> [1] 1
```

2. In the following call, explain how to find the documentation for the named arguments in the following function call:

```
plot(1:10, col = "red", pch = 20, xlab = "x", col.lab = "blue")
```



- 3. Why does `plot(1:10, col = "red")` only colour the points, not the axes or labels? Read the source code of `plot.default()` to find out.

## 5.7 Exiting a function

Most functions exit in one of two ways<sup>8</sup>: either returning a value, indicating successful completion, or throwing an error, indicating failure. This section describes return values (implicit vs. explicit; visible vs. invisible), briefly discusses errors, and introduces exit handlers, which allow you to run code when a function exits, regardless of how it exits.

### 5.7.1 Implicit vs. explicit returns

There are two ways that a function can return a value:

- Implicitly, where the last evaluated expression becomes the return value:

```
j01 <- function(x) {
 if (x < 10) {
 0
 } else {
 10
 }
}
j01(5)
#> [1] 0
j01(15)
#> [1] 10
```

- Explicitly, by calling `return()`:

```
j02 <- function(x) {
 if (x < 10) {
```

---

<sup>8</sup>Functions can exit in other more esoteric ways like signalling a condition that is caught by an exiting handler, invoking a restart, or pressing “Q” in an interactive browser.

```

 return(0)
} else {
 return(10)
}
}
```

### 5.7.2 Invisible values

Most functions return visibly: calling the function in an interactive context causes the result to be automatically printed.

```
j03 <- function() 1
j03()
#> [1] 1
```

However, it's also possible to return an `invisible()` value, which is not automatically printed.

```
j04 <- function() invisible(1)
j04()
```

You can verify that the value exists either by explicitly printing it or by wrapping in parentheses:

```
print(j04())
#> [1] 1

(j04())
#> [1] 1
```

Alternatively, use `withVisible()` to return the value and a visibility flag:

```
str(withVisible(j04()))
#> List of 2
#> $ value : num 1
#> $ visible: logi FALSE
```

The most common function that returns invisibly is `<-`:

```
a <- 2
(a <- 2)
#> [1] 2
```

And this is what makes it possible to chain assignment:

```
a <- b <- c <- d <- 2
```

In general, any function called primarily for its side effects (like `<-`, `print()`, or `plot()`) should return an invisible value (typically the value of the first argument).

### 5.7.3 Errors

If a function can not complete its assigned task, it should throw an error with `stop()`, which immediately terminates the execution of the function.

```
j05 <- function() {
 stop("I'm an error")
 return(10)
}
```

```
j05()
#> Error in j05():
#> I'm an error
```

Errors indicate that something has gone wrong, and force the user to handle them. Some languages (like C, go, and rust) rely on special return values to indicate problems, but in R you should always throw an error. You'll learn more about errors, and how to handle them, in Conditions.

### 5.7.4 Exit handlers

Sometimes a function needs to make a temporary change to global state and you want to ensure those changes are restored when the function completes. It's painful to make sure you cleanup before any explicit return, and what happens if there's an error? Instead, you can set up an **existing handler** that is called when the function terminates, regardless of whether it returns a value or throws an error.

To setup an existing handler, call `on.exit()` with the code to be run. It will execute when the function exits, regardless of what causes it to exit:

```
j06 <- function(x) {
 cat("Hello\n")
 on.exit(cat("Goodbye!\n"), add = TRUE)

 if (x) {
 return(10)
 } else {
 stop("Error")
 }
}

j06(TRUE)
#> Hello
#> Goodbye!
#> [1] 10

j06(FALSE)
#> Hello
#> Error in j06(FALSE):
#> Error
#> Goodbye!
```

Always set `add = TRUE` when using `on.exit()`. If you don't, each call to `on.exit()` will overwrite the previous existing handler. Even when only registering a single handler, it's good practice to set `add = TRUE` so that you don't get an unpleasant surprise if you later add more exit handlers

`on.exit()` is important because it allows you to place clean-up actions next to actions with their cleanup operations.

```
cleanup <- function(dir, code) {
 old_dir <- setwd(dir)
 on.exit(setwd(old_dir), add = TRUE)

 old_opt <- options(stringsAsFactors = FALSE)
 on.exit(options(old_opt), add = TRUE)
}
```

When coupled with lazy evaluation, this leads to a very useful pattern for running a block of code in an altered environment:

```
with_dir <- function(dir, code) {
 old <- setwd(dir)
 on.exit(setwd(old), add = TRUE)

 force(code)
}

getwd()
#> [1] "/home/travis/build/liao961120/adv-r"
with_dir("~/", getwd())
#> [1] "/home/travis"
```

See the `withr` package (<http://withr.r-lib.org>) for a collection of functions of this nature.

In R 3.4 and prior, `on.exit()` expressions are always run in the order in which they are created:

```
f <- function() {
 on.exit(message("a"), add = TRUE)
 on.exit(message("b"), add = TRUE)
}
f()
#> a
#> b
```

This can make cleanup a little tricky if some actions need to happen in a specific order; typically you want the most recent added expression to be run first. In R 3.5 and later, you can control this by setting `after = FALSE`:

```
f <- function() {
 on.exit(message("a"), add = TRUE, after = FALSE)
 on.exit(message("b"), add = TRUE, after = FALSE)
}
f()
#> b
#> a
```

### 5.7.5 Exercises

1. What does `load()` return? Why don't you normally see these values?
2. What does `write.table()` return? What would be more useful?
3. How does the `chdir` parameter of `source()` compare to `in_dir()`? Why might you prefer one approach to the other?
4. Write a function that opens a graphics device, runs the supplied code, and closes the graphics device (always, regardless of whether or not the plotting code worked).
5. We can use `on.exit()` to implement a simple version of `capture.output()`.

```
capture.output2 <- function(code) {
 temp <- tempfile()
 on.exit(file.remove(temp), add = TRUE, after = TRUE)

 sink(temp)
```

```

on.exit(sink(), add = TRUE, after = TRUE)

force(code)
readLines(temp)
}
capture.output2(cat("a", "b", "c", sep = "\n"))
#> [1] "a" "b" "c"

```

Compare `capture.output()` to `capture.output2()`. How do the functions differ? What features have I removed to make the key ideas easier to see? How have I rewritten the key ideas to be easier to understand?

## 5.8 Function forms

“To understand computations in R, two slogans are helpful:

- Everything that exists is an object.
- Everything that happens is a function call.”

— John Chambers

While everything that happens in R is a result of a function call, not all calls look the same. Function calls come in four varieties:

- In **prefix** form, the function name comes before its arguments, like `foofy(a, b, c)`. These constitute of the majority of function calls in R.
- In **infix** form, the function name comes inbetween its arguments, like `x + y`. Infix forms are used for many mathematical operators, as well as user-defined functions that begin and end with `%`.
- A **replacement** function assigns into what looks like a prefix function, like `names(df) <- c("a", "b", "c")`.
- **Special forms** like `[`, `if`, and `for`, don’t have a consistent structure and provide some of the most important syntax in R.

While four forms exist, you only need to use one, because any call can be written in prefix form. I’ll demonstrate this property, and then you’ll learn about each of the forms in turn.

### 5.8.1 Rewriting to prefix form

An interesting property of R is every infix, replacement, or special form can be rewritten in prefix form. Rewriting in prefix form is useful because it helps you better understand the structure of the language, and it gives you the real name of every function. Knowing the real name of non-prefix functions is useful because it allows you to modify them for fun and profit.

The following example shows three pairs of equivalent calls, rewriting an infix form, replacement form, and a special form into prefix form.

```

x + y
`+`(x, y)

names(df) <- c("x", "y", "z")
`names<-`(df, c("x", "y", "z"))

```

```
for(i in 1:10) print(i)
`for`(i, 1:10, print(i))
```

Knowing the function name of a non-prefix function allows you to override its behaviour. For example, if you're ever feeling particularly evil, run the following code while a friend is away from their computer. It will introduce a fun bug: 10% of the time, 1 will be added to any numeric calculation inside of parentheses.

Of course, overriding built-in functions like this is a bad idea, but, as you'll learn about in metaprogramming, it's possible to apply it only to selected code blocks. This provides a clean and elegant approach to writing domain specific languages and translators to other languages.

A more useful technique is to use this knowledge when using functional programming tools. For example, you could use `sapply()` to add 3 to every element of a list by first defining a function `add()`, like this:

```
add <- function(x, y) x + y
sapply(1:10, add, 3)
#> [1] 4 5 6 7 8 9 10 11 12 13
```

But we can also get the same effect more simply by relying on the existing `+` function:

```
sapply(1:5, `+`, 3)
#> [1] 4 5 6 7 8
```

We'll explore this idea in detail in functionals.

### 5.8.2 Prefix form

The prefix form is the most common form in R code, and indeed in the majority of programming languages. Prefix calls in R are a little special because you can specify arguments in three ways:

- By position, like `help(mean)`.
  - Using partial matching, like `help(to = mean)`.
  - By name, like `help(topic = mean)`.

As illustrated by the following chunk, arguments are matched by exact name, then with unique prefixes, and finally by position.

```
k01 <- function(abcdef, bcde1, bcde2) {
 list(a = abcdef, b1 = bcde1, b2 = bcde2)
}
str(k01(1, 2, 3))
#> List of 3
#> $ a : num 1
#> $ b1: num 2
#> $ b2: num 3
```

```

str(k01(2, 3, abcdef = 1))
#> List of 3
#> $ a : num 1
#> $ b1: num 2
#> $ b2: num 3

Can abbreviate long argument names:
str(k01(2, 3, a = 1))
#> List of 3
#> $ a : num 1
#> $ b1: num 2
#> $ b2: num 3

But this doesn't work because abbreviation is ambiguous
str(k01(1, 3, b = 1))
#> Error in k01(1, 3, b = 1):
#> argument 3 matches multiple formal arguments

```

Generally, only use positional matching for the first one or two arguments; they will be the most commonly used, and most readers will know what they are. Avoid using positional matching for less commonly used arguments, and never use partial matching. See the tidyverse style guide, <http://style.tidyverse.org/syntax.html#argument-names>, for more advice.

### 5.8.3 Infix functions

Infix functions are so called because the function name comes **in**between its arguments, and hence infix functions have two arguments. R comes with a number of built-in infix operators: `:`, `::`, `:::`, `$`, `@`, `^`, `*`, `/`, `+`, `-`, `>`, `>=`, `<`, `<=`, `==`, `!=`, `!`, `&`, `&&`, `|`, `||`, `~`, `<-`, and `<<-`. You can also create your own infix functions that start and end with `%`, and base R uses this to additionally define `%%`, `%*%`, `%/%`, `%in%`, `%o%`, and `%x%`.

Defining your own infix function is simple. You create a two argument function and bind it to a name that starts and ends with `%`:

```

`%+%` <- function(a, b) paste0(a, b)
"new" %+%"string"
#> [1] "new string"

```

The names of infix functions are more flexible than regular R functions: they can contain any sequence of characters except `"%"`. You will need to escape any special characters in the string used to define the function, but not when you call it:

```

`% %` <- function(a, b) paste(a, b)
`%/\\%` <- function(a, b) paste(a, b)

"a" % % "b"
#> [1] "a b"
"a" %/\\% "b"
#> [1] "a b"

```

R's default precedence rules mean that infix operators are composed from left to right:

```

`%-%` <- function(a, b) paste0("(", a, " %-% ", b, ")")
"a" %-% "b" %-% "c"
#> [1] "((a %-% b) %-% c)"

```

There are two special infix functions that can be called with a single argument: `+` and `-`.

```
-1
#> [1] -1
+10
#> [1] 10
```

### 5.8.4 Replacement functions

Replacement functions act like they modify their arguments in place, and have the special name `xxx<-`. They must have arguments named `x` and `value`, and must return the modified object. For example, the following function allows you to modify the second element of a vector:

```
`second<-` <- function(x, value) {
 x[2] <- value
 x
}
```

Replacement functions are used by placing the function call on the LHS of `<-`:

```
x <- 1:10
second(x) <- 5L
x
#> [1] 1 5 3 4 5 6 7 8 9 10
```

I say they “act” like they modify their arguments in place, because, as discussed in Modify-in-place, they actually create a modified copy. We can see that by using `tracemem()`:

```
x <- 1:10
tracemem(x)
#> <0x7ffae71bd880>

second(x) <- 6L
#> tracemem[0x7ffae71bd880 -> 0x7ffae61b5480]:
#> tracemem[0x7ffae61b5480 -> 0x7ffae73f0408]: second<-
```

If you want to supply additional arguments, they go inbetween `x` and `value`:

```
`modify<-` <- function(x, position, value) {
 x[position] <- value
 x
}
modify(x, 1) <- 10
x
#> [1] 10 5 3 4 5 6 7 8 9 10
```

When you write `modify(x, 1) <- 10`, behind the scenes R turns it into:

```
x <- `modify<-`(x, 1, 10)
```

Combining replacement with other functions requires more complex translation. For example, this:

```
x <- c(a = 1, b = 2, c = 3)
names(x)
#> [1] "a" "b" "c"

names(x)[2] <- "two"
names(x)
#> [1] "a" "two" "c"
```

Is translated into:

```
`*tmp*` <- x
x <- `names<-`(`*tmp*`, `[<-`(`names(`*tmp*`), 2, "two"))
rm(`*tmp*`)
```

(Yes, it really does create a local variable named *tmp*, which is removed afterwards.)

### 5.8.5 Special forms

Finally, there are a bunch of language features that are usually written in special ways, but also have prefix forms. These include parentheses:

- `(x)` (`^(`(x))`)
- `{x}` (`^`{`(^(`(x)))`)

The subsetting operators:

- `x[i]` (`^`(`(x, i))`)
- `x[[i]]` (`^`[`(`(`(x, i))`)

And the tools of control flow:

- `if (cond) true` (`^`if`(^(`(cond, true)))`)
- `if (cond) true else false` (`^`if`(^(`(cond, true, false)))`)
- `for(var in seq) action` (`^`for`(^(`(var, seq, action)))`)
- `while(cond) action` (`^`while`(^(`(cond, action)))`)
- `repeat expr` (`^`repeat`(^(`(expr)))`)
- `next` (`^`next`()`)
- `break` (`^`break`()`)

Finally, the most complex is the “function” function:

- `function(arg1, arg2) {body}` (`^`function`(^(`alist(arg1, arg2), body, env))`)

Knowing the name of the function that underlies the special form is useful for getting documentation. `?(`` is a syntax error; `?(`` will give you the documentation for parentheses.

Note that all special forms are implemented as primitive functions (i.e. in C); that means printing these functions is not informative:

```
`for`
#> .Primitive("for")
```

## 5.9 Invoking a function

Suppose you had a list of function arguments:

```
args <- list(1:10, na.rm = TRUE)
```

How could you then send that list to `mean()`? In base R, you need `do.call()`:

```
do.call(mean, args)
#> [1] 5.5
Equivalent to
mean(1:10, na.rm = TRUE)
#> [1] 5.5
```

### 5.9.1 Exercises

1. Rewrite the following code snippets into prefix form:

```
1 + 2 + 3
```

```
1 + (2 + 3)
```

```
if (length(x) <= 5) x[[5]] else x[[n]]
```

2. Clarify the following list of odd function calls:

```
x <- sample(replace = TRUE, 20, x = c(1:10, NA))
y <- runif(min = 0, max = 1, 20)
cor(m = "k", y = y, u = "p", x = x)
```

3. Explain why the following code fails:

```
modify(get("x"), 1) <- 10
#> Error: target of assignment expands to non-language object
```

4. Create a replacement function that modifies a random location in a vector.

5. Write your own version of `+` that will paste its inputs together if they are character vectors but behaves as usual otherwise. In other words, make this code work:

```
1 + 2
#> [1] 3

"a" + "b"
#> [1] "ab"
```

6. Create a list of all the replacement functions found in the base package. Which ones are primitive functions? (Hint use `apropos()`)

7. What are valid names for user-created infix functions?

8. Create an infix `xor()` operator.

9. Create infix versions of the set functions `intersect()`, `union()`, and `setdiff()`. You might call them `%n%`, `%u%`, and `%/%` to match conventions from mathematics.

## 5.10 Quiz answers

- The three components of a function are its body, arguments, and environment.
- `f1(1)()` returns 11.
- You'd normally write it in infix style: `1 + (2 * 3)`.
- Rewriting the call to `mean(c(1:10, NA), na.rm = TRUE)` is easier to understand.
- No, it does not throw an error because the second argument is never used so it's never evaluated.
- See infix and replacement functions.
- You use `on.exit()`; see `on.exit` for details.

# Chapter 6

# Environments

## 6.1 Introduction

The environment is the data structure that powers scoping. This chapter dives deep into environments, describing their structure in depth, and using them to improve your understanding of the four scoping rules described in lexical scoping. Understanding environments is not necessary for day-to-day use of R. But they are important to understand because they power many important R features like lexical scoping, namespaces, and R6 classes, and interact with evaluation to give you powerful tools for making domain specific languages, like dplyr and ggplot2.

## Quiz

If you can answer the following questions correctly, you already know the most important topics in this chapter. You can find the answers at the end of the chapter in answers.

1. List at least three ways that an environment is different to a list.
2. What is the parent of the global environment? What is the only environment that doesn't have a parent?
3. What is the enclosing environment of a function? Why is it important?
4. How do you determine the environment from which a function was called?
5. How are `<-` and `<<-` different?

## Outline

- Environment basics introduces you to the basic properties of an environment and shows you how to create your own.
- Recursing over environments provides a function template for computing with environments, illustrating the idea with a useful function.
- Explicit environments briefly discusses three places where environments are useful data structures for solving other problems.

## Prerequisites

This chapter will use rlang functions for working with environments, because it allows us to focus on the essence of environments, rather than the incidental details.

```
library(rlang)

Some API changes that haven't made it in rlang yet
search_envs <- function() {
 rlang:::new_environments(c(
 list(global_env()),
 head(env_parents(global_env()), -1)
))
}
```

Note that the `env_` functions in rlang are designed to work with the pipe: all take an environment as the first argument, and many also return an environment. I won't use the pipe in this chapter in the interest of keeping the code as simple as possible, but you should consider it for your own code.

## 6.2 Environment basics

Generally, an environment is similar to a named list, with four important exceptions:

- Every name must be unique.
- The names in an environment are not ordered (i.e. it doesn't make sense to ask what the first element of an environment is).
- An environment has a parent.
- Environments are not copied when modified.

Let's explore these ideas with code and pictures.

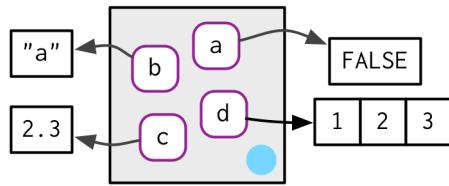
### 6.2.1 Basics

To create an environment, use `rlang:::env()`. It works like `list()`, taking a set of name-value pairs:

```
e1 <- env(
 a = FALSE,
 b = "a",
 c = 2.3,
 d = 1:3,
)
```

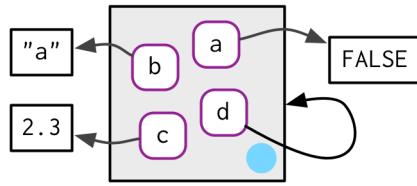
Use `new.env()` to creates a new environment. Ignore the `hash` and `size` parameters; they are not needed. Note that you can not simultaneously create and define values; use `$<-`, as shown below.

The job of an environment is to associate, or **bind**, a set of names to a set of values. You can think of an environment as a bag of names, with no implied order (i.e. it doesn't make sense to ask which is the first element in an environment). For that reason, we'll draw the environment as so:



As discussed in names and values, environments have reference semantics: unlike most R objects, when you modify them, you modify them in place, and don't create a copy. One important implication is that environments can contain themselves. This means that environments go one step further in their level of recursion than lists: an environment can contain any object, including itself!

```
e1$d <- e1
```



Printing an environment just displays its memory address, which is not terribly useful:

```
e1
#> <environment: 0x16b5408>
```

Instead, we'll use `env_print()` which gives us a little more information:

```
env_print(e1)
#> <environment: 0x16b5408>
#> parent: <environment: global>
#> bindings:
#> * a: <lgl>
#> * b: <chr>
#> * c: <dbl>
#> * d: <env>
```

You can use `env_names()` to get a character vector giving the current bindings

```
env_names(e1)
#> [1] "a" "b" "c" "d"
```

In R 3.2.0 and greater, use `names()` to list the bindings in an environment. If your code needs to work with R 3.1.0 or earlier, use `ls()`, but note that the default value of `all.names` is `FALSE` so you don't see any bindings that start with ..

## 6.2.2 Important environments

We'll talk in detail about special environments in Special environments, but for now we need to mention two. The current environment, or `current_env()` is the environment in which code is currently executing. When you're experimenting interactively, that's usually the global environment, or `global_env()`. The global environment is sometimes called your “workspace”, as it's where all interactive (i.e. outside of a function) computation takes place.

Note that to compare environments, you need to use `identical()` and not `==`:

```
identical(global_env(), current_env())
#> [1] TRUE

global_env() == current_env()
#> Error in global_env() == current_env():
#> comparison (1) is possible only for atomic and list types
```

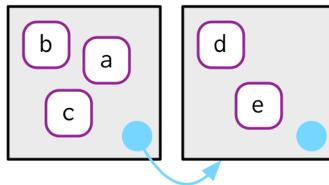
Access the global environment with `globalenv()` and the current environment with `environment()`. The global environment is printed as `Rf_GlobalEnv` and `.GlobalEnv`.

### 6.2.3 Parents

Every environment has a **parent**, another environment. In diagrams, the parent is shown as a small pale blue circle and arrow that points to another environment. The parent is what's used to implement lexical scoping: if a name is not found in an environment, then R will look in its parent (and so on).

You can set the parent environment by supplying an unnamed argument to `env()`. If you don't supply it, it defaults to the current environment.

```
e2a <- env(d = 4, e = 5)
e2b <- env(e2a, a = 1, b = 2, c = 3)
```



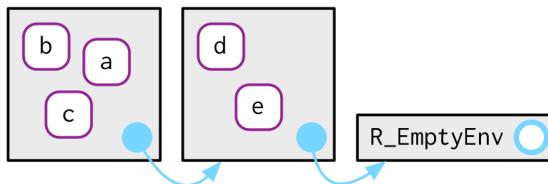
We use the metaphor of a family to name environments relative to one another. The grandparent of an environment is the parent's parent, and the ancestors include all parent environments up to the empty environment. To save space, I typically won't draw all the ancestors; just remember whenever you see a pale blue circle, there's a parent environment somewhere.

You can find the parent of an environment with `env_parent()`:

```
env_parent(e2b)
#> <environment: 0x4bb0348>
env_parent(e2a)
#> <environment: R_GlobalEnv>
```

Only one environment doesn't have a parent: the **empty** environment. I draw the empty environment with a hollow parent environment, and where space allows I'll label it with `R_EmptyEnv`, the name R uses.

```
e2c <- env(empty_env(), d = 4, e = 5)
e2d <- env(e2c, a = 1, b = 2, c = 3)
```



You'll get an error if you try to find the parent of the empty environment:

```
env_parent(empty_env())
#> Error: The empty environment has no parent
```

You can list all ancestors of an environment with `env_parents()`:

```
env_parents(e2b)
#> [[1]] <env: 0x4bb0348>
#> [[2]] $ <env: global>

env_parents(e2d)
#> [[1]] <env: 0x560cb58>
#> [[2]] $ <env: empty>
```

By default, `env_parents()` continues until it hits either the global environment or the empty environment. You can control this behaviour with the `last` environment.

Use `parent.env()` to find the parent of an environment. No base function returns all ancestors.

#### 6.2.4 Getting and setting

You can get and set elements of an environment with `$` and `[[` in the same way as a list:

```
e3 <- env(x = 1, y = 2)
e3$x
#> [1] 1
e3$z <- 3
e3[["z"]]
#> [1] 3
```

But you can't use `[[` with numeric indices, and you can't use `[`:

```
e3[[1]]
#> Error in e3[[1]]:
#> wrong arguments for subsetting an environment

e3[c("x", "y")]
#> Error in e3[c("x", "y")]:
#> object of type 'environment' is not subsettable
```

`$` and `[[` will return `NULL` if the binding doesn't exist. Use `env_get()` if you want an error:

```
e3$xyz
#> NULL

env_get(e3, "xyz")
#> Error in env_get(e3, "xyz"):
#> object 'xyz' not found
```

If you want to use a default value if the binding doesn't exist, you can use the `default` argument.

```
env_get(e3, "xyz", default = NA)
#> [1] NA
```

There are two other ways to add bindings to an environment:

- `env_poke()`<sup>1</sup> takes a name (as string) and a value:

---

<sup>1</sup>You might wonder why rlang has `env_poke()` instead of `env_set()`. This is for consistency: `_set()` functions return a

```
env_poke(e3, "a", 100)
e3$a
#> [1] 100
```

- `env_bind()` allows you to bind multiple values:

```
env_bind(e3, a = 10, b = 20)
env_names(e3)
#> [1] "x" "y" "z" "a" "b"
```

You can determine if an environment has a binding with `env_has()`:

```
env_has(e3, "a")
#> a
#> TRUE
```

Unlike lists, setting an element to `NULL` does not remove it. Instead, use `env_unbind()`:

```
e3$a <- NULL
env_has(e3, "a")
#> a
#> TRUE

env_unbind(e3, "a")
env_has(e3, "a")
#> a
#> FALSE
```

Unbinding a name doesn't delete the object. That's the job of the garbage collector, which automatically removes objects with no names binding to them. This process is described in more detail in GC.

See `get()`, `assign()`, `exists()`, and `rm()`. These are designed interactively for use with the current environment, so working with other environments is a little clunky. Also beware the `inherits` argument: it defaults to `TRUE` meaning that the base equivalents will inspect the supplied environment and all its ancestors.

## 6.2.5 Finalisers

Add something once rlang has an API. Also mention in data structures below

## 6.2.6 Advanced bindings

There are two more exotic variants of `env_bind()`:

- `env_bind_exprs()` creates **delayed bindings**, which are evaluated the first time they are accessed. Behind the scenes, delayed bindings create promises, so behave in the same way as function arguments.

```
env_bind_exprs(current_env(), b = {Sys.sleep(1); 1})

system.time(print(b))
#> [1] 1
#> user system elapsed
#> 0 0 1
system.time(print(b))
#> [1] 1
```

---

modified copy; `_poke()` functions modify in place.

```
#> user system elapsed
#> 0 0 0
```

Delayed bindings are used to implement `autoload()`, which makes R behave as if the package data is in memory, even though it's only loaded from disk when you ask for it.

- `env_bind_fns()` creates **active bindings** which are re-computed every time they're accessed:

```
env_bind_fns(current_env(), z1 = function(val) runif(1))

z1
#> [1] 0.0808
z1
#> [1] 0.834
```

The argument to the function allows you to also override behaviour when the variable is set:

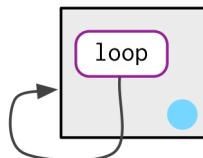
```
env_bind_fns(current_env(), z2 = function(val) {
 if (missing(val)) {
 2
 } else {
 stop("Don't touch z2!", call. = FALSE)
 }
})

z2
#> [1] 2
z2 <- 3
#> Error: Don't touch z2!
```

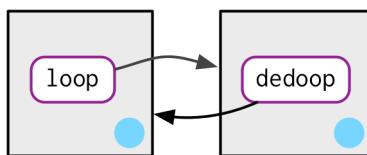
See `?delayedAssign()` and `?makeActiveBinding()`.

### 6.2.7 Exercises

1. List three ways in which an environment differs from a list.
2. Create an environment as illustrated by this picture.



3. Create a pair of environments as illustrated by this picture.



4. Explain why `e[[1]]` and `e[c("a", "b")]` don't make sense when `e` is an environment.
5. Create a version of `env_poke()` that will only bind new names, never re-bind old names. Some programming languages only do this, and are known as single assignment languages (<http://en.wikipedia.org>).

[org/wiki/Assignment\\_\(computer\\_science\)#Single\\_assignment\).](https://en.wikipedia.org/wiki/Assignment_(computer_science)#Single_assignment)

### 6.3 Recursing over environments

If you want to operate on every ancestor of an environment, it's often convenient to write a recursive function. This section shows you how, applying your new knowledge of environments to write a function that given a name, finds the environment `where()` that name is defined, using R's regular scoping rules.

The definition of `where()` is straightforward. It has two arguments: the name to look for (as a string), and the environment in which to start the search. (We'll learn why `caller_env()` is a good default in calling environments.)

```
where <- function(name, env = caller_env()) {
 if (identical(env, empty_env())) {
 # Base case
 stop("Can't find ", name, call. = FALSE)
 } else if (env_has(env, name)) {
 # Success case
 env
 } else {
 # Recursive case
 where(name, env_parent(env))
 }
}
```

There are three cases:

- The base case: we've reached the empty environment and haven't found the binding. We can't go any further, so we throw an error.
- The successful case: the name exists in this environment, so we return the environment.
- The recursive case: the name was not found in this environment, so try the parent.

These three cases are illustrated with these three examples:

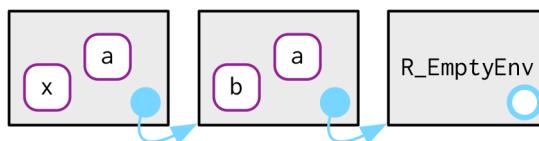
```
where("yyy")
#> Error: Can't find yyy

x <- 5
where("x")
#> <environment: R_GlobalEnv>

where("mean")
#> <environment: base>
```

It might help to see a picture. Imagine you have two environments, as in the following code and diagram:

```
e4a <- env(empty_env(), a = 1, b = 2)
e4b <- env(e4a, x = 10, a = 11)
```



- `where("a", e4b)` will find `a` in `e4b`.

- `where("b", e4b)` doesn't find `b` in `e4b`, so it looks in its parent, `e4a`, and finds it there.
- `where("c", e4b)` looks in `e4b`, then `e4a`, then hits the empty environment and throws an error.

It's natural to work with environments recursively, so `where()` provides a useful template. Removing the specifics of `where()` shows the structure more clearly:

```
f <- function(..., env = caller_env()) {
 if (identical(env, empty_env())) {
 # base case
 } else if (success) {
 # success case
 } else {
 # recursive case
 f(..., env = env_parent(env))
 }
}
```

## Iteration vs recursion

It's possible to use a loop instead of recursion. I think it's harder to understand than the recursive version, but I include it because you might find it easier to see what's happening if you haven't written many recursive functions.

```
f2 <- function(..., env = caller_env()) {
 while (!identical(env, empty_env())) {
 if (success) {
 # success case
 return()
 }
 # inspect parent
 env <- env_parent(env)
 }

 # base case
}
```

### 6.3.1 Exercises

1. Modify `where()` to return *all* environments that contain a binding for `name`. Carefully think through what type of object the function will need to return.
2. Write a function called `fget()` that finds only function objects. It should have two arguments, `name` and `env`, and should obey the regular scoping rules for functions: if there's an object with a matching name that's not a function, look in the parent. For an added challenge, also add an `inherits` argument which controls whether the function recurses up the parents or only looks in one environment.

## 6.4 Special environments

Most environments are not created by you (e.g. with `env()`) but are instead created by R. In this section, you'll learn about the most important environments, starting with the package environments. You'll then learn about the function environment bound to the function when it is created, and the (usually) ephemeral execution environment created every time the function is called. Finally, you'll see how the package and

function environments interact to support namespaces, which ensure that a package always behaves the same way, regardless of what other packages the user has loaded.

### 6.4.1 Package environments and the search path

Each package attached by `library()` or `require()` becomes one of the parents of the global environment. The immediate parent of the global environment is the last package you attached<sup>2</sup>:

```
env_parent(global_env())
#> <environment: package:rlang>
#> attr("name")
#> [1] "package:rlang"
#> attr("path")
#> [1] "/home/travis/R/Library/rlang"
```

And the parent of that package is the second to last package you attached:

```
env_parent(env_parent(global_env()))
#> <environment: package:stats>
#> attr("name")
#> [1] "package:stats"
#> attr("path")
#> [1] "/home/travis/R-bin/lib/R/library/stats"
```

If you follow all the parents back, you see the order in which every package has been attached. This is known as the **search path** because all objects in these environments can be found from the top-level interactive workspace.

```
search_envs()
#> [[1]] $ <env: global>
#> [[2]] $ <env: package:rlang>
#> [[3]] $ <env: package:stats>
#> [[4]] $ <env: package:graphics>
#> [[5]] $ <env: package:grDevices>
#> [[6]] $ <env: package:utils>
#> [[7]] $ <env: package:datasets>
#> [[8]] $ <env: package:methods>
#> [[9]] $ <env: Autoloads>
#> [[10]] $ <env: base>
```

You can access the names of the environments on the search path with `search()`

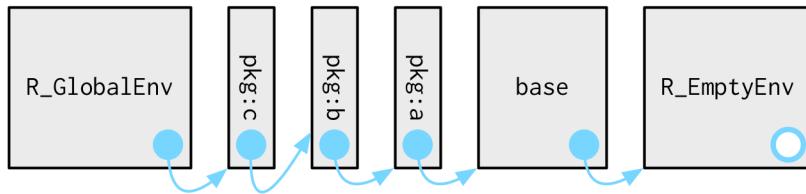
The last two environments on the search path are always the same:

- The `Autoloads` environment uses delayed bindings to save memory by only loading package objects (like big datasets) when needed.
- The base environment, `package:base` or sometimes just `base`, is the environment of the base package. It is special because it has to be able to bootstrap the loading of all other packages. You can access it directly with `base_env()`.

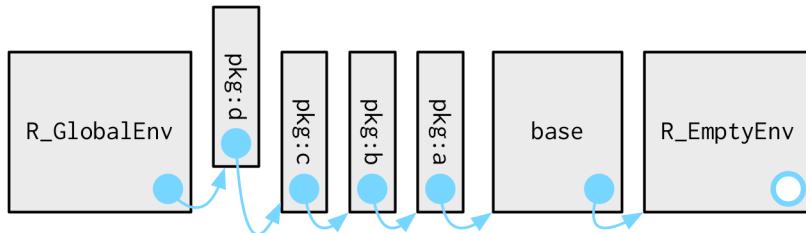
Graphically, the search path looks like this:

---

<sup>2</sup>Note the difference between attached and loaded. A package is loaded automatically if you access one of its functions using `:::`; it is only **attached** to the search path by `library()` or `require()`.



When you attach another package with `library()`, the parent environment of the global environment changes:



## 6.4.2 The function environment

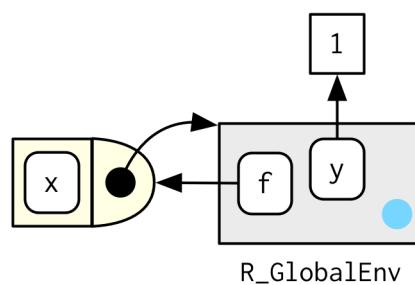
A function binds the current environment when it is created. This is called the **function environment**, and is used for lexical scoping. Across computer languages, functions that capture their environments are called **closures**, which is why this term is often used interchangeably with function in R's documentation.

You can get the function environment with `fn_env()`:

```
y <- 1
f <- function(x) x + y
fn_env(f)
#> <environment: R_GlobalEnv>
```

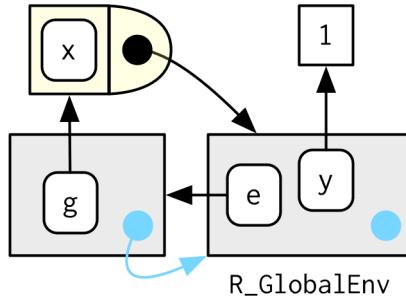
Use `environment(f)` to access the environment of function `f`.

In diagrams, I'll depict functions as rectangles with a rounded end that binds an environment.



In this case, `f()` binds the environment that binds the name `f` to the function. But that's not always the case: in the following example `g` is bound in a new environment `e`, but `g()` binds the global environment. The distinction between binding and being bound by is subtle but important; the difference is how we find `g` vs. how `g` finds its variables.

```
e <- env()
e$g <- function() 1
```



### 6.4.3 Namespaces

In the diagram above, you saw that the parent environment of a package varies based on what other packages have been loaded. This seems worrying: doesn't that mean that the package will find different functions if packages are loaded in a different order? The goal of **namespaces** is to make sure that this does not happen, and that every package works the same way regardless of what packages are attached by the user.

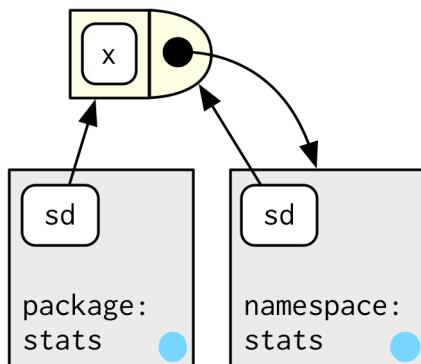
For example, take `sd()`:

```
sd
#> function (x, na.rm = FALSE)
#> sqrt(var(if (is.vector(x) || is.factor(x)) x else as.double(x),
#> na.rm = na.rm))
#> <bytecode: 0x64c5af8>
#> <environment: namespace:stats>
```

`sd()` is defined in terms of `var()`, so you might worry that the result of `sd()` would be affected by any function called `var()` either in the global environment, or in one of the other attached packages. R avoids this problem by taking advantage of the function vs. binding environment described above. Every function in a package is associated with a pair of environments: the package environment, which you learned about earlier, and the **namespace** environment.

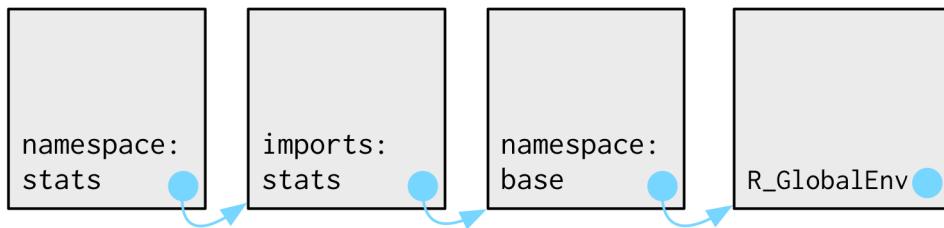
- The package environment is the external interface to the package. It's how you, the R user, find a function in an attached package or with `:::`. Its parent is determined by search path, i.e. the order in which packages have been attached.
- The namespace environment is the internal interface to the package. The package environment controls how we find the function; the namespace controls how the function finds its variables.

Every binding in the package environment is also found in the namespace environment; this ensures every function can use every other function in the package. But some bindings only occur in the namespace environment. These are known as internal or non-exported objects, which make it possible to hide internal implementation details from the user.

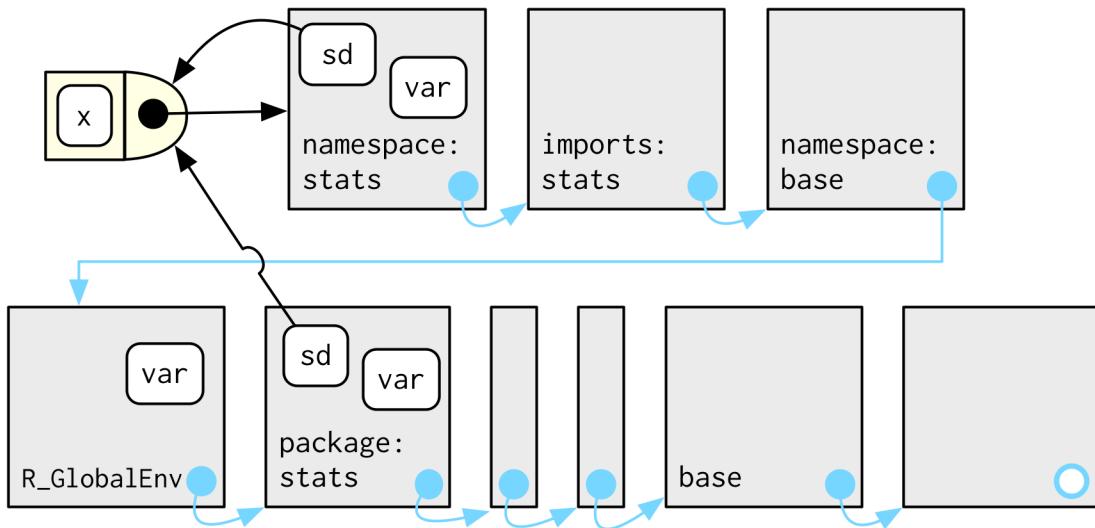


Every namespace environment has the same set of ancestors:

- Each namespace has an **imports** environment that contains bindings to all the functions used by the package. The imports environment is controlled by the package developer with the **NAMESPACE** file.
- Explicitly importing every base function would be tiresome, so the parent of the imports environment is the base **namespace**. The base namespace contains the same bindings as the base environment, but it has different parent.
- The parent of the base namespace is the global environment. This means that if a binding isn't defined in the imports environment the package will look for it in the usual way. This is usually a bad idea (because it makes code depend on other loaded packages), so R CMD check automatically warns about such code. It is needed primarily for historical reasons, particularly due to how S3 method dispatch works.



Putting all these diagrams together we get:



So when `sd()` looks for the value of `var` it always finds it in a sequence of environments determined by the package developer, but not by the package user. This ensures that package code always works the same way regardless of what packages have been attached by the user.

Note that there's no direct link between the package and namespace environments; the link is defined by the function environments.

#### 6.4.4 Execution environments

The last important topic we need to cover is the **execution environment**. What will the following function return the first time it's run? What about the second?

```

g <- function(x) {
 if (!env_has(current_env(), "a")) {
 message("Defining a")
 a <- 1
 } else {
 a <- a + 1
 }
 a
}

```

Think about it for a moment before you read on.

```

g(10)
#> Defining a
#> [1] 1
g(10)
#> Defining a
#> [1] 1

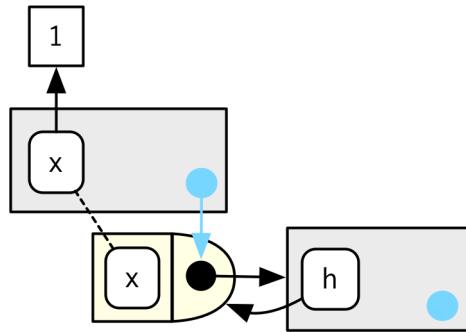
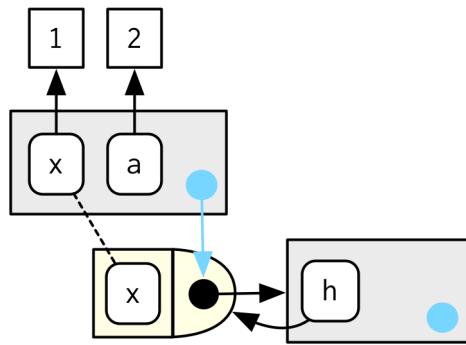
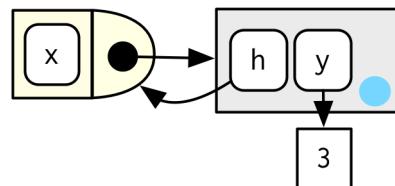
```

This function returns the same value every time because of the fresh start principle, described in a fresh start. Each time a function is called, a new environment is created to host execution. This is called the execution environment, and its parent is the function environment. Let's illustrate that process with a simpler function. I'll draw execution environments with an indirect parent; the parent environment is found via the function environment.

```

h <- function(x) {
 # 1.
 a <- 2 # 2.
 x + a
}
y <- h(1) # 3.

```

**1. Function called with  $x = 1$** **2. a bound to value 2****3. Function completes returning value 3.  
Execution environment goes away.**

An execution environment is usually ephemeral; once the function has completed, the environment will be GC'd. There are several ways to make it stay around for longer. The first is to explicitly return it:

```
h2 <- function(x) {
 a <- x * 2
 current_env()
}

e <- h2(x = 10)
env_print(e)
#> <environment: 0x5651058>
#> parent: <environment: global>
#> bindings:
```

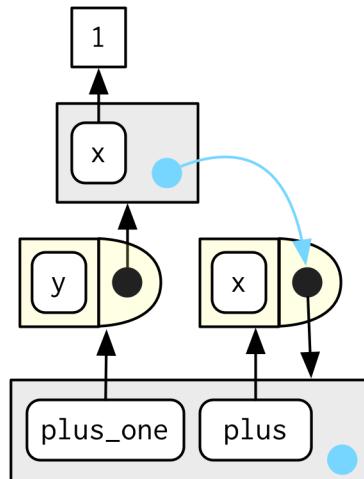
```
#> * a: <dbl>
#> * x: <dbl>
fn_env(h2)
#> <environment: R_GlobalEnv>
```

Another way to capture it is to return an object with a binding to that environment, like a function. The following example illustrates that idea with a function factory, `plus()`. We use that factory to create a function called `plus_one()`.

There's a lot going on in the diagram because the enclosing environment of `plus_one()` is the execution environment of `plus()`.

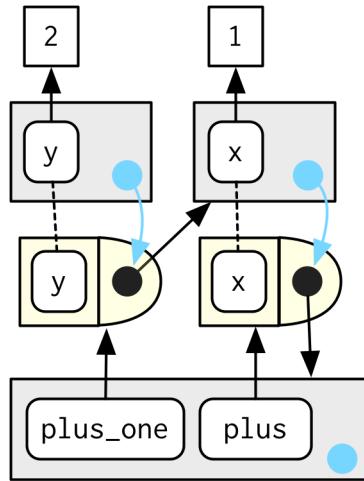
```
plus <- function(x) {
 function(y) x + y
}

plus_one <- plus(1)
plus_one
#> function(y) x + y
#> <environment: 0x593e550>
```



What happens when we call `plus_one()`? Its execution environment will have the captured execution environment of `plus()` as its parent:

```
plus_one(2)
#> [1] 3
```



You'll learn more about function factories in functional programming.

#### 6.4.5 Exercises

1. How is `search_envs()` different to `env_parents(global_env())`?
2. Draw a diagram that shows the enclosing environments of this function:

```
f1 <- function(x1) {
 f2 <- function(x2) {
 f3 <- function(x3) {
 x1 + x2 + x3
 }
 f3(3)
 }
 f2(2)
}
f1(1)
```

3. Write an enhanced version of `str()` that provides more information about functions. Show where the function was found and what environment it was defined in.

## 6.5 The call stack

There is one last environment we need to explain, the **caller** environment, accessed with `rlang::caller_env()`. This provides the environment from which the function was called, and hence varies based on how the function is called, not how the function was created. As we saw above this is a useful default whenever you write a function that takes an environment as an argument.

`parent.frame()` is equivalent to `caller_env()`; just note that it returns an environment, not a frame.

To fully understand the caller environment we need to discuss two related concepts: the **call stack**, which is made up of **frames**. Executing a function creates two types of context. You've learned about one already: the execution environment is a child of the function environment, which is determined by where the function was created. There's another type of context created by where the function was called: this is called the call stack.

There are also a couple of small wrinkles when it comes to custom evaluation. See environments vs. frames for more details.

### 6.5.1 Simple call stacks

Let's illustrate this with a simple sequence of calls: `f()` calls `g()` calls `h()`.

```
f <- function(x) {
 g(x = 2)
}
g <- function(x) {
 h(x = 3)
}
h <- function(x) {
 stop()
}
```

The way you most commonly see a call stack in R is by looking at the `traceback()` after an error has occurred:

```
f(x = 1)
#> Error:
traceback()
#> 4: stop()
#> 3: h(x = 3)
#> 2: g(x = 2)
#> 1: f(x = 1)
```

Instead of `stop()` + `traceback()` to understand the call stack, we're going to use `lobstr::cst()` to print out the call stack tree:

```
h <- function(x) {
 lobstr::cst()
}
f(x = 1)
#>
#> f(x = 1)
#> g(x = 2)
#> h(x = 3)
#> lobstr::cst()
```

This shows us that `cst()` was called from `h()`, which was called from `g()`, which was called from `f()`. Note that the order is the opposite from `traceback()`. As the call stacks get more complicated, I think it's easier to understand the sequence of calls if you start from the beginning, rather than the end (i.e. `f()` calls `g()`; rather than `g()` was called by `f()`).

### 6.5.2 Lazy evaluation

The call stack above is simple - while you get a hint that there's some tree-like structure involved, everything happens on a single branch. This is typical of a call stack when all arguments are eagerly evaluated.

Let's create a more complicated example that involves some lazy evaluation. We'll create a sequence of functions, `a()`, `b()`, `c()`, that pass along an argument `x`.

```
a <- function(x) b(x)
b <- function(x) c(x)
```

```
c <- function(x) x

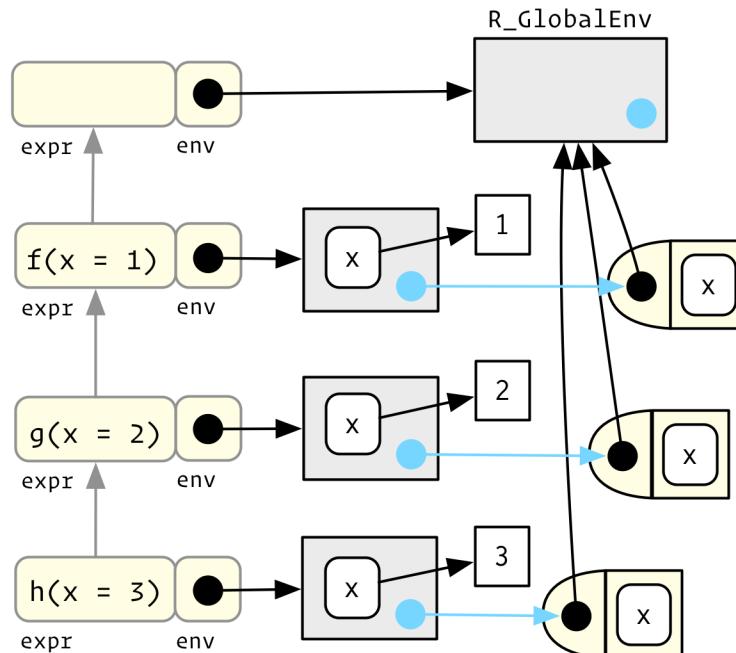
a(f())
#>
#> a(f())
#> b(x)
#> c(x)
#> f()
#> g(x = 2)
#> h(x = 3)
#> lobstr::cst()
```

`x` is lazily evaluated so this tree gets two branches. In the first branch `a()` calls `b()`, then `b()` calls `c()`. The second branch starts when `c()` evaluates its argument `x`. This argument is evaluated in a new branch because the environment in which it is evaluated is the global environment, not the environment of `c()`.

### 6.5.3 Frames

Each element of the call stack is a **frame**<sup>3</sup>, also known as an evaluation context. The frame is an extremely important internal data structure, and R code can only access a small part of the data structure because it's so critical. A frame has three main components that are accessible from R:

- An expression (labelled with `expr`) giving the function call. This is what `traceback()` prints out.
- An environment (labelled with `env`), which is typically the execution environment of a function. There are two main exceptions: the environment of the global frame is the global environment, and calling `eval()` also generates frames, where the environment can be anything.
- A parent, the previous call in the call stack (shown by a grey arrow).



<sup>3</sup>NB: `?environment` uses frame in a different sense: “Environments consist of a *frame*, or collection of named objects, and a pointer to an enclosing environment.” We avoid this sense of frame, which comes from S, because it’s very specific and not widely used in base R. For example, the “frame” in `parent.frame()` is an execution context, not a collection of named objects.

(To focus on the calling environments, I have omitted the bindings in the global environment from `f`, `g`, and `h` to the respective function objects.)

The frame also holds exit handlers created with `on.exit()`, restarts and handlers for the condition system, and which context to `return()` to when a function completes. These are important for the internal operation of R, but are not directly accessible.

### 6.5.4 Dynamic scope

Looking up variables in the calling stack rather than in the enclosing environment is called **dynamic scoping**. Few languages implement dynamic scoping (Emacs Lisp is a notable exception (<http://www.gnu.org/software/emacs/emacs-paper.html#SEC15>).) This is because dynamic scoping makes it much harder to reason about how a function operates: not only do you need to know how it was defined, you also need to know the context in which it was called. Dynamic scoping is primarily useful for developing functions that aid interactive data analysis. It is one of the topics discussed in non-standard evaluation.

### 6.5.5 Exercises

1. Write a function that lists all the variables defined in the environment in which it was called. It should return the same results as `ls()`.

## 6.6 As data structures

As well as powering scoping, environments are also useful data structures in their own right because they have reference semantics. There are three common problems that they can help solve:

- **Avoiding copies of large data.** Since environments have reference semantics, you'll never accidentally create a copy. This makes it a useful vessel for large objects. Bare environments are not that pleasant to work with; I recommend using R6 objects instead. Learn more in R6.
- **Managing state within a package.** Explicit environments are useful in packages because they allow you to maintain state across function calls. Normally, objects in a package are locked, so you can't modify them directly. Instead, you can do something like this:

```
my_env <- new.env(parent = emptyenv())
my_env$a <- 1

get_a <- function() {
 my_env$a
}

set_a <- function(value) {
 old <- my_env$a
 my_env$a <- value
 invisible(old)
}
```

Returning the old value from setter functions is a good pattern because it makes it easier to reset the previous value in conjunction with `on.exit()` (see more in on exit).

- **As a hashmap.** A hashmap is a data structure that takes constant,  $O(1)$ , time to find an object based on its name. Environments provide this behaviour by default, so can be used to simulate a hashmap. See the CRAN package hash for a complete development of this idea.

## 6.7 <<-

The ancestors of an environment have an important relationship to <<-. The regular assignment arrow, <- , always creates a variable in the current environment. The deep assignment arrow, <<-, never creates a variable in the current environment, but instead modifies an existing variable found by walking up the parent environments.

```
x <- 0
f <- function() {
 x <<- 1
}
f()
x
#> [1] 1
```

If <<- doesn't find an existing variable, it will create one in the global environment. This is usually undesirable, because global variables introduce non-obvious dependencies between functions. <<- is most often used in conjunction with a closure, as described in Closures.

### 6.7.1 Exercises

- What does this function do? How does it differ from <<- and why might you prefer it?

```
rebind <- function(name, value, env = caller_env()) {
 if (identical(env, empty_env())) {
 stop("Can't find `", name, "`, call. = FALSE")
 } else if (env_has(env, name)) {
 env_poke(env, name, value)
 } else {
 rebind(name, value, env_parent(env))
 }
}
rebind("a", 10)
#> Error: Can't find `a`
a <- 5
rebind("a", 10)
a
#> [1] 10
```

## 6.8 Quiz answers

- There are four ways: every object in an environment must have a name; order doesn't matter; environments have parents; environments have reference semantics.
- The parent of the global environment is the last package that you loaded. The only environment that doesn't have a parent is the empty environment.
- The enclosing environment of a function is the environment where it was created. It determines where a function looks for variables.
- Use `caller_env()` or `parent.frame()`.
- <- always creates a binding in the current environment; <<- rebinds an existing name in a parent of the current environment.



# Chapter 7

## Conditions

### 7.1 Introduction

The **condition** system provides a paired set of tools that allow the author of a function to indicate that something unusual is happening, and the user of that function to deal with it. The function author **signals** conditions with functions like `stop()` (for errors), `warning()` (for warnings), and `message()` (for messages), then the function user can handle them with functions like `tryCatch()` and `withCallingHandlers()`. Understanding the condition system is important because you'll often need to play both roles: signalling conditions from the functions you create, and handle conditions signalled by the functions you call.

R offers a very powerful condition system based on ideas from Common Lisp. Like R's approach to object oriented programming, it is rather different to currently popular programming languages so it is easy to misunderstand, and there has been relatively little written about how to use it effectively. Historically, this has lead to few people (including me!) taking full advantage of its power. The goal of this chapter is to remedy that situation. Here you will learn about the big ideas of R's conditional system, as well as learning a bunch of practical tools that will make your code stronger.

I found two resources particularly useful when writing this chapter. You may also want to read them if you want to learn more about the inspirations and motivations for the system:

- *A prototype of a condition system for R* (<http://homepage.stat.uiowa.edu/~luke/R/exceptions/simpcond.html>) by Robert Gentleman and Luke Tierney. This describes an early version of R's condition system. While the implementation has changed somewhat since this document was written, it provides a good overview of how the pieces fit together, and some motivation for its design.
- *Beyond exception handling: conditions and restarts* (<http://www.gigamonkeys.com/book/beyond-exception-handling-conditions-and-restarts.html>) by Peter Seibel. This describes exception handling in Lisp, which happens to be very similar to R's approach. It provides useful motivation and more sophisticated examples. I have provided an R translation of the chapter at <http://adv-r.had.co.nz/beyond-exception-handling.html>.

I also found it helpful to work through the underlying C code that implements these ideas. If you're interested in understanding how it all works, you might find my notes (<https://gist.github.com/hadley/4278d0a6d3a10e42533d59905fbed0ac>) to be useful.

### Quiz

Want to skip this chapter? Go for it, if you can answer the questions below. Find the answers at the end of the chapter in answers.

1. What are the three most important types of condition?
2. What function do you use to ignore errors in block of code?
3. What's the main difference between `tryCatch()` and `withCallingHandlers()`?
4. Why might you want to create a custom error object?

## Overview

- Signalling conditions introduces the basic tools for signalling conditions, and discusses when it is appropriate to use each type.
- Ignoring conditions teaches you about the simplest tools for handling conditions: functions like `try()` and `suppressMessages()` that swallow conditions and prevent them from getting to the top level.
- Handling conditions introduces the condition `object`, and the two fundamental tools of condition handling: `tryCatch()` for error conditions, and `withCallingHandlers()` for everything else.
- Custom conditions shows you how to extend the built-in condition objects to store useful data that condition handlers can use to make more informed decisions.
- Applications closes out the chapter with a grab bag of practical applications based on the low-level tools found in earlier sections.

### 7.1.1 Prerequisites

As well as base R functions, this chapter uses condition signalling and handling functions from `rlang`.

```
library(rlang)
```

## 7.2 Signalling conditions

There are three conditions that you can signal in code: errors, warnings, and messages.

- Errors are the most severe; they indicate that there is no way for a function to continue and execution must stop.
- Messages are the mildest; they are way of informing the user that some action has been performed on their behalf.
- Warnings fall somewhat in between, and typically indicate that something has gone wrong but the function has been able to at least partially recover.

There is a final condition that can only be generated interactively: an interrupt, which indicates that the user has “interrupted” execution by pressing Escape, Ctrl + Break, or Ctrl + C (depending on the platform).

Conditions are usually displayed prominently, in a bold font or coloured red, depending on the R interface. You can tell them apart because errors always start with “Error”, warnings with “Warning message”, and messages with nothing.

```
stop("This is what an error looks like")
#> Error in eval(expr, envir, enclos):
#> This is what an error looks like

warning("This is what a warning looks like")
#> Warning: This is what a warning looks like
```

```
message("This is what a message looks like")
#> This is what a message looks like
```

The following three sections describe errors, warnings, and message in more detail.

### 7.2.1 Errors

In base R, errors are signalled, or **thrown**, by `stop()`:

```
f <- function() g()
g <- function() h()
h <- function() stop("This is an error!")

f()
#> Error in h():
#> This is an error!
```

By default, the error message includes the call, but this is typically not useful (and recapitulates information that you can easily get from `traceback()`), so I think it's good practice to use `call. = FALSE`:

```
h <- function() stop("This is an error!", call. = FALSE)
f()
#> Error: This is an error!
```

The rlang equivalent to `stop()`, `rlang::abort()`, does this automatically. We'll use `abort()` throughout this chapter, but we won't get to its most compelling feature, the ability to add additional metadata to the condition object, until we're near the end of the chapter.

```
h <- function() abort("This is an error!")
f()
#> Error: This is an error!
```

(Note that `stop()` pastes together multiple inputs, while `abort()` does not. To create complex error messages with `abort`, I recommend using `glue::glue()`. This allows us to use other arguments to `abort()` for useful features that you'll learn about in custom conditions.)

The best error messages tell you what is wrong and point you in the right direction to fix the problem. Writing good error messages is hard because errors usually occur when the user has a flawed mental model of the function. As a developer, it's hard to imagine how the user might be thinking incorrectly about your function, and thus it's hard to write a message that will steer the user in the correct direction. That said, the tidyverse style guide discusses a few general principles that we have found useful: <http://style.tidyverse.org/error-messages.html>.

### 7.2.2 Warnings

Warnings, signalled by `warning()`, are weaker than errors: they signal that something has gone wrong, but the code has been able to recover and continue. Unlike errors, you can have multiple warnings from a single function call:

```
fw <- function() {
 cat("1\n")
 warning("W1")
 cat("2\n")
 warning("W2")
 cat("3\n")
```

```
warning("W3")
}
```

By default, warnings are cached and printed only when control returns to the top level:

```
fw()
#> 1
#> 2
#> 3
#> Warning messages:
#> 1: In f() : W1
#> 2: In f() : W2
#> 3: In f() : W3
```

You can control this behaviour with the `warn` option:

- To make warnings appear immediately, set `options(warn = 1)`.
- To turn warnings into errors, set `options(warn = 2)`. This is usually the easiest way to debug a warning, as once it's an error you can use tools like `traceback()` to find the source.
- Restore the default behaviour with `option(warn = 0)`.

Like `stop()`, `warning()` also has a call argument. It is slightly more useful (since warnings are often more distant from their source), but I still generally suppress it with `call. = FALSE`. Like `rlang::abort()`, the `rlang` equivalent of `warning()`, `rlang::warn()`, also suppresses the `call.` by default.

Warnings occupy a somewhat challenging place between messages (“you should know about this”) and errors (“you must fix this!”), and it’s hard to give precise advice on when to use them. Generally, be restrained, as warnings are easy to miss if there’s a lot of other output, and you don’t want your function to recover too easily from clearly invalid input. In my opinion, base R tends to overuse warnings, and many warnings in base R would be better off as errors. For example, I think these warnings would be more helpful as errors:

```
formals(1)
#> Warning in formals(fun): argument is not a function
#> NULL

file.remove("this-file-doesn't-exist")
#> Warning in file.remove("this-file-doesn't-exist"): cannot remove file
#> 'this-file-doesn't-exist', reason 'No such file or directory'
#> [1] FALSE

lag(1:3, k = 1.5)
#> Warning in lag.default(1:3, k = 1.5): 'k' is not an integer
#> [1] 1 2 3
#> attr(", "tsp")
#> [1] -1 1 1
```

There only a couple of cases where using a warning is clearly appropriate:

- When you **deprecate** a function you want to allow older code to continue to work (so ignoring the warning is ok) but you want to encourage the user to switch to a new function.
- When you are reasonably certain you can recover from a problem: If you were 100% certain that you could fix the problem, you wouldn’t need any message; if you were more uncertain that you could correctly fix the issue, you’d throw an error.

Otherwise use warnings with restraint, and carefully consider if an error would be more appropriate.

### 7.2.3 Messages

Messages, signalled by `message()`, are informational; use them to tell the user that you've done something on their behalf. Good messages are a balancing act: you want to provide just enough information so the user knows what's going on, but not so much that they're overwhelmed.

`messages()` are displayed immediately and do not have a `call.` argument:

```
fm <- function() {
 cat("1\n")
 message("M1")
 cat("2\n")
 message("M2")
 cat("3\n")
 message("M3")
}

fm()
#> 1
#> M1
#> 2
#> M2
#> 3
#> M3
```

Good places to use a message are:

- When a default argument requires some non-trivial amount of computation and you want to tell the user what value was used. For example, `ggplot2` reports the number of bins used if you don't supply a `binwidth`.
- In functions that are called primarily for their side-effects which would otherwise be silent. For example, when writing files to disk, calling a web API, or writing to a database, it's useful provide regular status messages telling the user what's happening.
- When you're about to start a long running process with no intermediate output. A progress bar (e.g. with `progress` (<https://github.com/r-lib/progress>)) is better, but a message is a good place start.
- When writing a package, you sometimes want to display a message when your package is loaded (i.e. in `.onAttach()`); here you must use `packageStartupMessage()`.

Generally any function that produces a message should have some way to suppress it, like a `quiet = TRUE` argument. It is possible to suppress all messages with `suppressMessages()`, as you'll learn shortly, but it is nice to also give finer grain control.

It's important to compare `message()` to the closely related `cat()`. In terms of usage and result, they appear quite similar<sup>1</sup>:

```
cat("Hi!\n")
#> Hi!

message("Hi!")
#> Hi!
```

However, the *purposes* of `cat()` and `message()` are different. Use `cat()` when the primary role of the function is to print to the console, like `print()` or `str()` methods. Use `message()` as a side-channel to print to the console when the primary purpose of the function is something else. In other words, `cat()` is for when the user *asks* for something to be printed and `message()` is for when developer *elects* to print something.

---

<sup>1</sup>But note that `cat()` requires an explicit trailing "\n" to print a new line.

### 7.2.4 Exercises

1. Write a wrapper around `file.remove()` that throws an error if the file to be deleted does not exist.
2. What does the `appendLF` argument to `message()` do? How is it related to `cat()`?
3. What does `options(error = recover)` do? Why might you use it?
4. What does `options(error = quote(dump.frames(to.file = TRUE)))` do? Why might you use it?

## 7.3 Ignoring conditions

The simplest way of handling conditions in R is to simply ignore them:

- Ignore errors with `try()`.
- Ignore warnings with `suppressWarnings()`.
- Ignore messages with `suppressMessages()`.

These functions are heavy handed as you can't use them to suppress a single type of condition that you know about, while allowing everything else to pass through. We'll come back to that challenge later in the chapter.

`try()` allows execution to continue even after an error has occurred. Normally if you run a function that throws an error, it terminates immediately and doesn't return a value:

```
f1 <- function(x) {
 log(x)
 10
}
f1("x")
#> Error in log(x):
#> non-numeric argument to mathematical function
```

However, if you wrap the statement that creates the error in `try()`, the error message will be displayed<sup>2</sup> but execution will continue:

```
f2 <- function(x) {
 try(log(x))
 10
}
f2("a")
#> Error in log(x) : non-numeric argument to mathematical function
#> [1] 10
```

It is possible, but not recommended, to save the result of `try()` and perform different actions based on whether or not the code succeed or failed<sup>3</sup>. Instead, it is better to use `tryCatch()` or a higher-level helper; you'll learn about those shortly. A simple, but useful, pattern is to do assignment inside the call: this lets you define a default value to be used if the code does not succeed.

```
default <- NULL
try(default <- read.csv("possibly-bad-input.csv"), silent = TRUE)
```

`suppressWarnings()` and `suppressMessages()` suppress all warnings and messages. Unlike errors, messages and warnings don't terminate execution, so there maybe multiple signalled in a single block.

---

<sup>2</sup>You can suppress the message with `try(..., silent = TRUE)`.

<sup>3</sup>You can tell if the expression failed because the result will have class `try-error`.

```

suppressWarnings({
 warning("Uhoh!")
 warning("Another warning")
 1
})
#> [1] 1

suppressMessages({
 message("Hello there")
 2
})
#> [1] 2

suppressWarnings({
 message("You can still see me")
 3
})
#> You can still see me
#> [1] 3

```

## 7.4 Handling conditions

Every condition has default behaviour: errors stop execution and return to the top level, warnings are captured and displayed in aggregate, and messages are immediately displayed. Condition **handlers** allow us to temporarily override or supplement the default behaviour.

Two functions, `tryCatch()` and `withCallingHandlers()`, allow us to register handlers, functions that take the signalled condition as their single argument. The registration functions have the same basic form:

```

tryCatch(
 error = function(cnd) {
 # code to run when error is thrown
 },
 code_to_run_while_handlers_are_active
)

withCallingHandlers(
 warning = function(cnd) {
 # code to run when warning is signalled
 },
 message = function(cnd) {
 # code to run when warning is signalled
 },
 code_to_run_while_handlers_are_active
)

```

They differ in the type of handlers that they create:

- `tryCatch()` defines **existing** handlers; after the condition is handled, control returns to the context where `tryCatch()` was called. This makes `tryCatch()` most suitable for working with errors and interrupts, as these have to exit anyway.
- `withCallingHandlers()` defines **calling** handlers; after the condition is captured control returns to the context where the condition was signalled. This makes it most suitable for working with non-error

conditions.

But before we can learn about and use these handlers, we need to talk a little bit about condition **objects**. These are created implicitly whenever you signal a condition, but become explicit inside the handler.

### 7.4.1 Condition objects

So far we've just signalled conditions, and not looked at the objects that are created behind the scenes. The easiest way to see a condition object is to catch one from a signalled condition. That's the job of `rlang::catch_cnd()`:

```
cnd <- catch_cnd(abort("An error"))
str(cnd)
#> List of 2
#> $ message: chr "An error"
#> $ call : NULL
#> - attr(*, "class")= chr [1:3] "rlang_error" "error" "condition"
```

Built-in conditions are lists with two elements:

- **message**, a length-1 character vector containing the text display to a user. To extract the message, use `conditionMessage(cnd)`.
- **call**, the call which triggered the condition. As described above, we don't use the call, so it will often be `NULL`. To extract it, use `conditionCall(cnd)`.

Custom conditions may contain other components, which we'll discuss in custom conditions.

Conditions also have a **class** attribute, which makes them S3 objects. We won't discuss S3 until S3, but fortunately, even if you don't know about S3, condition objects are quite simple. The most important thing to know is that the **class** attribute is a character vector, and it determines which handlers will match the condition.

### 7.4.2 Exiting handlers

`tryCatch()` registers exiting handlers, and is typically used to handle error conditions. It allows you to override the default error behaviour. For example, the following code will return 10 instead of displaying an error:

```
tryCatch(
 error = function(cnd) 10,
 stop("This is an error!")
)
#> [1] 10
```

If no conditions are signalled, or the class of the signalled condition does not match the handler name, the code executes normally:

```
tryCatch(
 error = function(cnd) 10,
 1 + 1
)
#> [1] 2

tryCatch(
 error = function(cnd) 10,
{
```

```

 message("Hi!")
 1 + 1
}
)
#> Hi!
#> [1] 2

```

The handlers set up by `tryCatch()` are called **exiting** handlers because after the condition is signalled, control passes to the handler and never returns to the original code, effectively meaning that the code “exits”:

```

tryCatch(
 message = function(cnd) "There",
{
 message("Here")
 stop("This code is never run!")
}
)
#> [1] "There"

```

Note that the code is evaluated in the environment of `tryCatch()`, but the handler code is not, because the handlers are functions. This is important to remember if you’re trying to modify objects in the parent environment.

The handler functions are called with a single argument, the condition object. I call this argument `cnd`, by convention. This value is only moderately useful for the base conditions because they contain relatively little data. It’s more useful when you make your own custom conditions, as you’ll see shortly.

```

tryCatch(
 error = function(cnd) {
 paste0("--", conditionMessage(cnd), "--")
 },
 stop("This is an error")
)
#> [1] "--This is an error--"

```

`tryCatch()` has one other argument: `finally`. It specifies a block of code (not a function) to run regardless of whether the initial expression succeeds or fails. This can be useful for clean up, like deleting files, or closing connections. This is functionally equivalent to using `on.exit()` (and indeed that’s how it’s implemented) but it can wrap smaller chunks of code than an entire function.

### 7.4.3 Calling handlers

The handlers set up by `tryCatch()` are called exiting handlers, because they cause code to exit once the condition has been caught. By contrast, `withCallingHandler()` sets up **calling** handlers: code execution continues normally once the handler returns. This tends to make `withCallingHandlers()` a more natural pairing with the non-error conditions.

Compare the results of `tryCatch()` and `withCallingHandlers()` in the example below. The message are not printed in the first case, because the code is terminated once the exiting handler completes. They are printed in the second case, because a calling handler does not exit.

```

tryCatch(
 message = function(cnd) cat("Caught a message!\n"),
{
 message("Someone there?")
}
)
#> [1] "Caught a message!"
#> [2] "Someone there?"

```

```

 message("Why, yes!")
 }
)
#> Caught a message!

withCallingHandlers(
 message = function(c) cat("Caught a message!\n"),
 {
 message("Someone there?")
 message("Why, yes!")
 }
)
#> Caught a message!
#> Someone there?
#> Caught a message!
#> Why, yes!

```

Handlers are applied in order, so you don't need to worry getting caught in an infinite loop:

```

withCallingHandlers(
 message = function(cnd) message("Second message"),
 message("First message")
)
#> Second message
#> First message

```

(But beware if you have multiple handlers, and some handlers signal conditions that could be captured by another handler: you'll need to think through the order carefully.)

The return value of a calling handler is ignored because the code continues to execute after the handler completes; where would the return value go? That means that calling handlers are only useful for their side-effects.

One important side-effect unique to calling handlers is the ability to **muffle** the signal. By default, a condition will continue to propagate to parent handlers, all the way up to the default handler (or an exiting handler, if provided):

```

Bubbles all the way up to default handler which generates the message
withCallingHandlers(
 message = function(cnd) cat("Level 2\n"),
 withCallingHandlers(
 message = function(cnd) cat("Level 1\n"),
 message("Hello")
)
)
#> Level 1
#> Level 2
#> Hello

Bubbles up to tryCatch
tryCatch(
 message = function(cnd) cat("Level 2\n"),
 withCallingHandlers(
 message = function(cnd) cat("Level 1\n"),
 message("Hello")
)
)
```

```
)
#> Level 1
#> Level 2
```

If you want to prevent the condition “bubbling up” but still run the rest of the code in the block, you need to explicitly muffle it with `rlang::cnd_muffle()`:

```
Muffles the default handler which prints the messages
withCallingHandlers(
 message = function(cnd) {
 cat("Level 2\n")
 cnd_muffle(cnd)
 },
 withCallingHandlers(
 message = function(cnd) cat("Level 1\n"),
 message("Hello")
)
)
#> Level 1
#> Level 2

Muffles level 2 handler and the default handler
withCallingHandlers(
 message = function(cnd) cat("Level 2\n"),
 withCallingHandlers(
 message = function(cnd) {
 cat("Level 1\n")
 cnd_muffle(cnd)
 },
 message("Hello")
)
)
#> Level 1
```

#### 7.4.4 Call stacks

To complete the section, there are some important differences between the call stacks of exiting and calling handlers. These differences are generally not important but I’m including it here because I’ve occassionally found it useful, and don’t want to forget about it!

It’s easiest to see the difference by setting up a small example that uses `lobstr::cst()`:

```
f <- function() g()
g <- function() h()
h <- function() message("!")
```

Calling handlers are called in the context of the call that signalled the condition:

```
withCallingHandlers(f(), message = function(cnd) {
 lobstr::cst()
 cnd_muffle(cnd)
})
#> x
#> +-withCallingHandlers(...
#> +-f()
```

```
#> / \-g()
#> / \-h()
#> / \-message("!")
#> / +-withRestarts(...)
#> / / \-withOneRestart(expr, restarts[[1L]])
#> / / \-doWithOneRestart(return(expr), restart)
#> / \-signalCondition(cond)
#> \-(function (cnd) ...
#> \-lobstr::cst()
```

Whereas exiting handlers are called in the context of the call to `tryCatch()`:

```
tryCatch(f(), message = function(cnd) lobstr::cst())
#> x
#> \-tryCatch(f(), message = function(cnd) lobstr::cst())
#> \-tryCatchList(expr, classes, parentenv, handlers)
#> \-tryCatchOne(expr, names, parentenv, handlers[[1L]])
#> \-value[[3L]](cond)
#> \-lobstr::cst()
```

### 7.4.5 Exercises

- Predict the results of evaluating the following code

```
show_condition <- function(code) {
 tryCatch(
 error = function(cnd) "error",
 warning = function(cnd) "warning",
 message = function(cnd) "message",
 {
 code
 NULL
 }
)
}

show_condition(stop("!"))
show_condition(10)
show_condition(warning("?!"))
show_condition({
 10
 message(?)
 warning("?!")
})
```

- Explain the results of running this code:

```
withCallingHandlers(
 message = function(cnd) message("b"),
 withCallingHandlers(
 message = function(cnd) message("a"),
 message("c")
)
)
#> b
```

```
#> a
#> b
#> c
```

3. Read the source code for `catch_cnd()` and explain how it works.
4. How could you rewrite `show_condition()` to use a single handler?

## 7.5 Custom conditions

One of the challenges of error handling in R is that most functions generate one of the built-in conditions, which contain only a `message` and a `call`. That means that if you want to detect a specific type of error, you can only work with the text of the error message. This is error prone, not only because the message might change over time, but also because messages can be translated into other languages.

Fortunately R has a powerful, but little used feature: the ability to create custom conditions that can contain additional metadata. Creating custom conditions is a little fiddly in base R, but `rlang::abort()` makes it very easy as you can supply a custom `.subclass` and additional metadata.

The following example shows the basic pattern. I recommend using the following call structure for custom conditions. This takes advantage of R's flexible argument matching so that the name of the “type” of error comes first, followed by the user facing text, followed by custom metadata.

```
abort(
 "error_not_found",
 message = "Path `blah.csv` not found",
 path = "blah.csv"
)
#> Error: Path `blah.csv` not found
```

Custom conditions work just like regular conditions when used interactively, but allow handlers to do much more.

### 7.5.1 Motivation

To explore these ideas in more depth, let's take `base::log()`. It does the minimum when throwing errors caused by invalid arguments:

```
log(letters)
#> Error in log(letters):
#> non-numeric argument to mathematical function
log(1:10, base = letters)
#> Error in log(1:10, base = letters):
#> non-numeric argument to mathematical function
```

I think we can do better by being explicit about which argument is the problem (i.e. `x` or `base`), and saying what the problematic input is (not just what it isn't).

```
my_log <- function(x, base = exp(1)) {
 if (!is.numeric(x)) {
 abort(paste0("`x` must be a numeric vector; not ", typeof(x), "."))
 }
 if (!is.numeric(base)) {
 abort(paste0("`base` must be a numeric vector; not ", typeof(base), "."))
 }
}
```

```
base::log(x, base = base)
}
```

This gives us:

```
my_log(letters)
#> Error: `x` must be a numeric vector; not character.
my_log(1:10, base = letters)
#> Error: `base` must be a numeric vector; not character.
```

This is an improvement for interactive usage as the error messages are more likely to guide the user towards a correct fix. However, they're no better if you want to programmatically handle the errors: all the useful metadata about the error is jammed into a single string.

## 7.5.2 Signalling

Let's build some infrastructure to improve this situation. We'll start by providing a custom `abort()` function for bad arguments. This is a little over-generalised for the example at hand, but it reflects common patterns that I've seen across other functions. The pattern is fairly simple. We create a nice error message for the user, using `glue::glue()`, and store metadata in the condition call for the developer.

```
abort_bad_argument <- function(arg, must, not = NULL) {
 msg <- glue::glue(`{arg}` must {must})
 if (!is.null(not)) {
 not <- typeof(not)
 msg <- glue::glue("{msg}; not {not}.")
 }

 abort("error_bad_argument",
 message = msg,
 arg = arg,
 must = must,
 not = not
)
}
```

If you want to throw a custom error without adding a dependency on `rlang`, you can create a condition object “by hand” and then pass it to `stop()`:

```
stop_custom <- function(.subclass, message, call = NULL, ...) {
 err <- structure(
 list(
 message = message,
 call = call,
 ...
),
 class = c(.subclass, "error", "condition")
)
 stop(err)
}

err <- catch_cnd(stop_custom("error_new", "This is a custom error", x = 10))
class(err)
err$x
```

We can now rewrite `my_log()` to use this new helper:

```
my_log <- function(x, base = exp(1)) {
 if (!is.numeric(x)) {
 abort_bad_argument("x", must = "be numeric", not = x)
 }
 if (!is.numeric(base)) {
 abort_bad_argument("base", must = "be numeric", not = base)
 }

 base::log(x, base = base)
}
```

`my_log()` itself is not much shorter, but is a little more meaningful, and it ensures that error messages for bad arguments are consistent across functions. It yields the same interactive error messages as before:

```
my_log(letters)
#> Error: `x` must be numeric; not character.
my_log(1:10, base = letters)
#> Error: `base` must be numeric; not character.
```

### 7.5.3 Handling

These structured condition objects are much easier to program with. The first place you might want to use this capability is when testing your function. Unit testing is not a subject of this book (see R packages (<http://r-pkgs.had.co.nz/>) for details), but the basics are easy to understand. The following code captures the error, and then asserts it has the structure that we expect.

```
library(testthat)

err <- catch_cnd(my_log("a"))
expect_s3_class(err, "error_bad_argument")
expect_equal(err$arg, "x")
expect_equal(err$not, "character")
```

We can also use the class (`error_bad_argument`) in `tryCatch()` to only handle that specific error:

```
tryCatch(
 error_bad_argument = function(cnd) "bad_argument",
 error = function(cnd) "other error",
 my_log("a")
)
#> [1] "bad_argument"
```

Note that when using `tryCatch()` with multiple handlers and custom classes, the first handler to match any class in the signal's class vector is called, not the best match. For this reason, you need to make sure to put the most specific handlers first. The following code does not do what you might hope:

```
tryCatch(
 error = function(cnd) "other error",
 error_bad_argument = function(cnd) "bad_argument",
 my_log("a")
)
#> [1] "other error"
```

### 7.5.4 Exercises

1. Inside a package, it's occassionally useful to check that a package is installed before using it. Write a function that checks if a package is installed (with `requireNamespace("pkg", quietly = FALSE)`) and if not, throws a custom condition that includes the package name in the metadata.
2. Inside a package you often need to stop with an error when something is not right. Other packages that depend on your package might be tempted to check these errors in their unit tests. How could you help these packages to avoid relying on the error message which is part of the user interface rather than the API and might change without notice?

## 7.6 Applications

Now that you've learned the basic tools of R's condition system, it's time to dive into some applications. The goal of this section is not to show every possible usage of `tryCatch()` and `withCallingHandlers()` but to illustrate some common patterns that frequently crop up. Hopefully these will get your creative juices flowing, so when you encounter a new problem you can come up with a useful solution.

### 7.6.1 Failure value

There are a few simple, but useful, `tryCatch()` patterns based on returning a value from the error handler. The simplest case is a wrapper to return a "default" value if an error occurs:

```
fail_with <- function(expr, value = NULL) {
 tryCatch(
 error = function(cnd) value,
 expr
)
}

fail_with(log(10), NA_real_)
#> [1] 2.3
fail_with(log("x"), NA_real_)
#> [1] NA
```

A more sophisticated application is `base::try()`. Below, `try2()` extracts the essense of `base::try()`; the real function is more complicated in order to make the error message look more like what you'd see if `tryCatch()` wasn't used.

```
try2 <- function(expr, silent = FALSE) {
 tryCatch(
 error = function(cnd) {
 msg <- conditionMessage(cnd)
 if (!silent) {
 message("Error: ", msg)
 }
 structure(msg, class = "try-error")
 },
 expr
)
}

try2(1)
```

```
#> [1] 1
try2(stop("Hi"))
#> Error: Hi
#> [1] "Hi"
#> attr(,"class")
#> [1] "try-error"
try2(stop("Hi"), silent = TRUE)
#> [1] "Hi"
#> attr(,"class")
#> [1] "try-error"
```

### 7.6.2 Success and failure values

We can extend this pattern to returns one value if the code evaluates successfully (`success_val`), and another if it fails (`error_val`). This pattern just requires one small trick: evaluating the user supplied code, then `success_val`. If the code throws an error, we'll never get to `success_val` and will instead return `error_val`.

```
foo <- function(expr) {
 tryCatch(
 error = function(cnd) error_val,
 {
 expr
 success_val
 }
)
}
```

We can use this to determine if an expression fails:

```
does_error <- function(expr) {
 tryCatch(
 error = function(cnd) TRUE,
 {
 expr
 FALSE
 }
)
}
```

Or to capture any condition, like just `rlang:::catch_cnd()`:

```
catch_cnd <- function(expr) {
 tryCatch(
 condition = function(cnd) cnd,
 {
 expr
 NULL
 }
)
}
```

We can also use this pattern to create a `try()` variant. One challenge with `try()` is that it's slightly challenging to determine if the code succeeded or failed. Rather than returning an object with a special class, I think it's slightly nicer to return a list with two components `result` and `error`.

```

safety <- function(expr) {
 tryCatch(
 error = function(cnd) {
 list(result = NULL, error = cnd)
 },
 list(result = expr, error = NULL)
)
}

str(safety(1 + 10))
#> List of 2
#> $ result: num 11
#> $ error : NULL
str(safety(abort("Error!")))
#> List of 2
#> $ result: NULL
#> $ error :List of 4
#> ...$ message: chr "Error!"
#> ...$ call : NULL
#> ...$ trace :List of 3
#>$ calls :List of 31
#>$: language local({ args = commandArgs(TRUE) ...
#>$: language eval.parent(substitute(eval(quote(expr), env..
#>$: language eval(expr, p)
#>$: language eval(expr, p)
#>$: language eval(quote({ args = commandArgs(TRUE) ...
#>$: language eval(quote({ args = commandArgs(TRUE) ...
#>$: language do.call(rmarkdown::render, c(args[1], readR..
#>$: language (function (input, output_format = NULL, outp..
#>$: language knitr::knit(knit_input, knit_output, envir =..
#>$: language process_file(text, output)
#>$: language withCallingHandlers(if (tangle) process_tang..
#>$: language process_group(group)
#>$: language process_group.block(group)
#>$: language call_block(x)
#>$: language block_exec(params)
#>$: language in_dir(input_dir(), evaluate(code, envir = e..
#>$: language evaluate(code, envir = env, new_device = FAL..
#>$: language evaluate::evaluate(...)
#>$: language evaluate_call(expr, parsed$src[[i]], envir =..
#>$: language timing_fn(handle(ev <- withCallingHandlers(w..
#>$: language handle(ev <- withCallingHandlers(withVisible..
#>$: language withCallingHandlers(withVisible(eval(expr, e..
#>$: language withVisible(eval(expr, envir, enclos))
#>$: language eval(expr, envir, enclos)
#>$: language eval(expr, envir, enclos)
#>$: language str(safety(abort("Error!")))
#>$: language safety(abort("Error!"))
#>$: language tryCatch(error = function(cnd) { list(re..
#>- attr(*, "srcref")= 'srcref' int [1:8] 2 3 7 3 3 3 2 7
#>- attr(*, "srcfile")=Classes 'srcfilecopy', 'srcfil..
#>$: language tryCatchList(expr, classes, parentenv, handl..
#>$: language tryCatchOne(expr, names, parentenv, handlers..

```

```
#>$.language doTryCatch(return(expr), name, parentenv, ha...
#>$.parents: int [1:31] 0 1 2 3 0 5 6 6 8 9 ...
#>$.envs : List of 31
#>$. : chr "0x36b77c0"
#>$. : chr "0x36b7670"
#>$. : chr "0x36bb158"
#>$. : chr "global"
#>$. : chr "0x36baac8"
#>$. : chr "0x36ba630"
#>$. : chr "0x36bc860"
#>$. : chr "0x3827f70"
#>$. : chr "0x2547c68"
#>$. : chr "0x2389e58"
#>$. : chr "0x2ab5fb0"
#>$. : chr "0x2ab5aa8"
#>$. : chr "0x2ab5808"
#>$. : chr "0x2ab5728"
#>$. : chr "0x1b4b120"
#>$. : chr "0x1aefb60"
#>$. : chr "0x1aee998"
#>$. : chr "0x1aed0d0"
#>$. : chr "0x2985d00"
#>$. : chr "0x24a5e18"
#>$. : chr "0x24a5d00"
#>$. : chr "0x24a58a0"
#>$. : chr "0x24a5280"
#>$. : chr "0x24a8e10"
#>$. : chr "global"
#>$. : chr "0x24a8a90"
#>$. : chr "0x24a8908"
#>$. : chr "0x1abc248"
#>$. : chr "0x1abc980"
#>$. : chr "0x1abcd00"
#>$. : chr "0x1abd080"
#> ... - attr(*, "class")= chr "rlang_trace"
#> ... $.parent : NULL
#> ... - attr(*, "class")= chr [1:3] "rlang_error" "error" "condition"
```

(This is closely related to `purrr::safely()`, a function operator, which we'll come back to in Section 11.2.1.)

### 7.6.3 Resignal

As well as returning default values when a condition is signalled, handlers can be used to make more informative error messages. One simple application is to make a function that works like `option(warn = 2)` for a single block of code. The idea is simple: we handle warnings by throwing an error:

```
warning2error <- function(expr) {
 withCallingHandlers(
 warning = function(cnd) abort(conditionMessage(cnd)),
 expr
)
}
```

```
warning2error({
 x <- 2 ^ 4
 warn("Hello")
})
#> Error: Hello
```

You could write a similar function if you were trying to find the source of an annoying message.

#### 7.6.4 Record

Another common pattern is to record conditions for later investigation. The new challenge here is that calling handlers are called only for their side-effects so we can't return values, but instead need to modify some object in place.

```
catch_cnds <- function(expr) {
 conds <- list()
add_cond <- function(cnd) {
 conds <-> append(conds, list(cnd))
 cnd_muffle(cnd)
}

withCallingHandlers(
 message = add_cond,
 warning = add_cond,
 expr
)

conds
}

catch_cnds({
 inform("a")
 warn("b")
 inform("c")
})
#> [[1]]
#> <message: a>
#> >
#>
#> [[2]]
#> <warning: b>
#>
#> [[3]]
#> <message: c>
#> >
```

What if you also want to capture errors? You'll need to wrap the `withCallingHandlers()` in a `tryCatch()`. If an error occurs, it will be the last condition.

```
catch_cnds <- function(expr) {
 conds <- list()
add_cond <- function(cnd) {
 conds <-> append(conds, list(cnd))
 cnd_muffle(cnd)
```

```

 }

tryCatch(
 error = function(cnd) {
 conds <- append(conds, list(cnd))
 },
 withCallingHandlers(
 message = add_cond,
 warning = add_cond,
 expr
)
)

conds
}

catch_cnds({
 inform("a")
 warn("b")
 abort("C")
})
#> [[1]]
#> <message: a
#> >
#>
#> [[2]]
#> <warning: b>
#>
#> [[3]]
#> <rlang_error: C>

```

This is the key idea underlying the evaluate (<https://github.com/r-lib/evaluate>) package which powers knitr: it captures every output into a special data structure so that it can be later replayed. As a whole, the evaluate package is quite a lot more complicated than the code here because it also needs to handle plots and text output.

### 7.6.5 No default behaviour

A final useful pattern is to signal a condition that doesn't inherit from `message`, `warning` or `error`. Because there is no default behaviour, this means the condition has no effect unless the user specifically requests it. For example, you could imagine a logging system based on conditions:

```

log <- function(message, level = c("info", "error", "fatal")) {
 level <- match.arg(level)
 signal(message, "log", level = level)
}

```

When you call `log()` a condition is signalled, but nothing happens because it has no default handler:

```
log("This code was run")
```

To “activate” logging you need a handler that does something with the `log` condition. Below I define a `record_log()` function that will record all logging messages to a path:

```
record_log <- function(expr, path = stdout()) {
 withCallingHandlers(
 log = function(cnd) {
 cat(
 "[", cnd$level, "] ", cnd$message, "\n", sep = "",
 file = path, append = TRUE
)
 },
 expr
)
}

record_log(log("Hello"))
#> [info] Hello
```

You could even imagine layering with another function that allows you to selectively suppress some logging levels.

```
ignore_log_levels <- function(expr, levels) {
 withCallingHandlers(
 log = function(cnd) {
 if (cnd$level %in% levels) {
 cnd_muffle(cnd)
 }
 },
 expr
)
}

record_log(ignore_log_levels(log("Hello"), "info"))
```

If you create a condition object by hand, and signal it with `signalCondition()`, `cnd_muffle()` will not work. Instead you need to call it with a muffle restart defined, like this:

```
withRestarts(signalCondition(cond), muffle = function() NULL)
```

Restarts are currently beyond the scope of the book, but I suspect will be included in the 3rd edition.

### 7.6.6 Exercises

1. Create `suppressConditions()` that works like `suppressMessages()` and `suppressWarnings()` but suppresses everything. Think carefully about how you should handle errors.
2. Compare the following two implementations of `message2error()`. What is the main advantage of `withCallingHandlers()` in this scenario? (Hint: look carefully at the traceback.)

```
message2error <- function(code) {
 withCallingHandlers(code, message = function(e) stop(e))
}
message2error <- function(code) {
 tryCatch(code, message = function(e) stop(e))
}
```

3. How would you modify the `catch_cnds()` defined if you wanted to recreate the original intermingling of warnings and messages?

4. Why is catching interrupts dangerous? Run this code to find out.

```
bottles_of_beer <- function(i = 99) {
 message("There are ", i, " bottles of beer on the wall, ", i, " bottles of beer.")
 while(i > 0) {
 tryCatch(
 Sys.sleep(1),
 interrupt = function(err) {
 i <- i - 1
 if (i > 0) {
 message(
 "Take one down, pass it around, ", i,
 " bottle", if (i > 1) "s", " of beer on the wall."
)
 }
 }
)
 }
 message("No more bottles of beer on the wall, no more bottles of beer.")
}
```

## 7.7 Quiz answers

1. `error`, `warning`, and `message`.
2. You could use `try()` or `tryCatch()`.
3. `tryCatch()` creates exiting handlers which will terminate the execution of wrapped code; `withCallingHandlers()` creates calling handlers which don't affect the execution of wrapped code.
4. Because you can then capture specific types of error with `tryCatch()`, rather than relying on the comparison of error strings, which is risky, especially when messages are translated.



# Chapter 8

## Connections

In R, every time you read data in or write data out, you are using a connection behind the scenes. Connections abstract away the underlying implementation so that you can read and write data the same way, regardless of whether you're writing to a file, an HTTP connection, a pipe, or something more exotic.

- <http://biostatmatt.com/R/R-conn-ints/index.html#Top>
- `?file`
- [https://cran.r-project.org/doc/Rnews/Rnews\\_2001-1.pdf](https://cran.r-project.org/doc/Rnews/Rnews_2001-1.pdf)
- <https://cran.r-project.org/doc/manuals/r-release/R-data.html#Connections>

### 8.1 Basics

- default connections: `stdin`, `stderr`, `stdout`
- `cat()` + `cat_line()`
- survey of base connections: file, compressed file, url, pipe, socket, text
- important packages: `curl`
- blocking vs non-blocking
- pattern: `close()` with `on.exit()` if you opened

### 8.2 Reading and writing binary data

- `raw()`
- `readBin()` vs `writeBin()`
- text vs binary (newlines and nulls)

### 8.3 Reading and writing text data

Reading and writing text is more complicated than reading and writing binary data because as soon as you move beyond regular ASCII characters (e.g. a-z, 0-9) there are many different ways of representing the same text. The way in which text data is stored in binary is known as the **encoding**.

- Encodings
  - <https://kevinushey.github.io/blog/2018/02/21/string-encoding-and-r/>
  - in general vs `Encoding`
  - `encoding` vs `fileEncoding`

- converting with iconv
- UTF-8 everywhere
- Reliably reading and writing UTF-8

## Part II

# Functional programming



# Introduction

R, at its heart, is a **functional** language. This means that it has certain technical properties, but more importantly that it lends itself to a style of problem solving centered on functions. Below I'll give a brief overview of the technical definition of a functional *language*, but in this book I will primarily focus on the functional *style* of programming, because I think it is an extremely good fit to the types of problem you commonly encounter when doing data analysis.

Recently, functional techniques have experienced a surge in interest because they can produce efficient and elegant solutions to many modern problems. A functional style tends to create functions that can easily be analysed in isolation (i.e. using only local information), and hence is often much easier to automatically optimise or parallelise. The traditional weaknesses of function languages, poorer performance and sometimes unpredictable memory usage, have been much reduced in recent years. Functional programming is complementary to object oriented programming, which has been the dominant programming paradigm for the last several decades.

## Functional programming languages

Every programming language has functions, so what makes a programming language functional? There are many definitions for precisely what makes a language “functional”, but there are two common threads.

Firstly, functional languages have **first class functions**, functions that behave like any other data structure. In R, this means that you can do anything with a function that you can do with a vector: you can assign them to variables, store them in lists, pass them as arguments to other functions, create them inside functions, and even return them as the result of a function.

Secondly, many functional languages require functions to be **pure**. A function is pure if it satisfies two properties:

- The output only depends on the inputs, i.e. if you call it again with the same inputs, you get the same outputs. This excludes functions like `runif()`, `read.csv()`, or `Sys.time()` that can return different values.
- The function has no side-effects, like changing the value of a variable, writing to disk, or displaying to the screen. This excludes functions like `print()`, `write.csv()` and `<-`.

Pure functions are much easier to reason about, but obviously have significant downsides: imagine doing a data analysis where you couldn't generate random numbers or read files from disk.

Strictly speaking, R isn't a functional programming *language* because it doesn't require that you write pure functions. However, you can certainly adopt a functional style in parts of your code: you don't *have* to write pure functions, but you often *should*. In my experience, partitioning code into functions that are either extremely pure and or extremely impure tends to lead to code that is easier to understand and extend to new situations.

## Functional style

It's hard to describe exactly what a functional *style* is, but generally I think it means decomposing a big problem into smaller pieces then solving each piece with a function or combination of functions. When using a functional style you strive to decompose components of the problem into isolated functions that operate independently. Each function taken by itself is simple and straightforward to understand; complexity is handled by composing functions in various ways.

The following three chapters discuss the three key functional techniques that help you to decompose problems into smaller pieces:

- Chapter 9 shows you how to replace many for loops with **functionals** which are functions (like `lapply()`) that take another function as an argument. Functionals allow you to take a function that solves the problem for a single input, and generalise it to handle any number of inputs. Functionals are by far and away the most important technique, and you'll use them all the time in data analysis.
- Chapter 10 introduces **function factories**, functions that create functions. Function factories are less useful than functionals, but often allow you elegantly partition work between different parts of your code.
- Chapter 11 shows you how to create **function operators**, functions that take functions as input and produce functions as output. They are like adverbs, because they typically modify the operation of a function.

Collectively, these types of function are called **higher-order functions**, and fill out a two-by-two table:

| <i>In</i> | <i>Out</i> |                   |
|-----------|------------|-------------------|
| Vector    | Vector     | Function          |
| Function  | Functional | Function operator |

# Chapter 9

## Functionals

### 9.1 Introduction

“To become significantly more reliable, code must become more transparent. In particular, nested conditions and loops must be viewed with great suspicion. Complicated control flows confuse programmers. Messy code often hides bugs.”

— Bjarne Stroustrup

A **functional** is a function that takes a function as an input and returns a vector as output. Here’s a simple functional: it calls the function provided as input with 1000 random uniform numbers.

```
randomise <- function(f) f(runif(1e3))
randomise(mean)
#> [1] 0.506
randomise(mean)
#> [1] 0.501
randomise(sum)
#> [1] 489
```

The chances are that you’ve already used a functional. You might have used for-loop replacement like base R’s `lapply()`, `apply()`, or `tapply()`, or maybe `purrr`’s `map()` or variant; or maybe you’ve used a mathematical functional like `integrate()` or `optim()`. All functionals take a function as input (among other things) and return a vector as output.

A common use of functionals is as an alternative to for loops. For loops have a bad rap in R. They have a reputation for being slow (although that reputation is only partly true, see Section 2.5.1 for more details). But the real downside of for loops is that they’re not very expressive. A for loop conveys that it’s iterating over something, but doesn’t clearly convey a high level goal. Instead of using a for loop, it’s better to use a functional. Each functional is tailored for a specific task, so when you recognise the functional you immediately know why it’s being used. Functionals play other roles as well as replacements for for-loops. They are useful for encapsulating common data manipulation tasks like split-apply-combine, for thinking “functionally”, and for working with mathematical functions.

Functionals reduce bugs in your code by better communicating intent. Functionals implemented in base R and `purrr` are well tested (i.e., bug-free) and efficient, because they’re used by so many people. Many are written in C, and use special tricks to enhance performance. That said, using functionals will not always produce the fastest code. Instead, it helps you clearly communicate and build tools that solve a wide range of problems. It’s a mistake to focus on speed until you know it’ll be a problem. Once you have clear, correct code you can make it fast using the techniques you’ll learn in Section 23.

Using functionals is a pattern matching exercise. You look at the for loop, and find a functional that matches the basic form. If one doesn't exist, don't try and torture an existing functional to fit the form you need. Instead, just leave it as a for loop!

It's not about eliminating for loops. It's about having someone else write them for you!

## Outline

### Prerequisites

This chapter will focus on functionals provided by the purrr package. These functions have a consistent interface that makes it easier to understand the key ideas than their base equivalents, which have grown organically over many years. I'll compare and contrast base R functions as we go, and then wrap up the chapter with a discussion of base functionals that don't have purrr equivalents.

```
library(purrr)
```

Many R users feel guilty about using for loops instead of apply functions. It's natural to blame yourself for failing to understand and internalise the apply family of functions. However, I think this is like blaming yourself when embarrass yourself by failing to pull open a door when it's supposed to be pushed open<sup>1</sup>. It's not actually your fault, because many people suffer the same problem; it's a failing of design. Similarly, I think the reason why the apply functions are so hard for so many people is because their design is suboptimal.

## 9.2 My first functional: `map()`

The most fundamental functional is `purrr::map()`<sup>2</sup>. It takes a vector and a function, calls the function once for each element of the vector, and returns the results in a list. In other words, `map(1:3, f)` yields `list(f(x[[1]]), f(x[[2]]), f(x[[3]]))`.

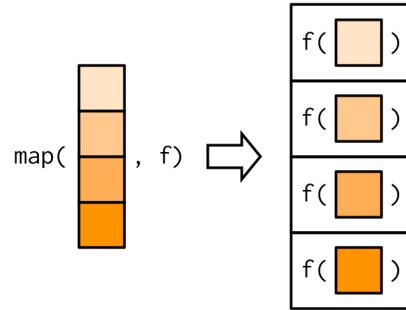
```
triple <- function(x) x * 3
map(1:3, triple)
#> [[1]]
#> [1] 3
#>
#> [[2]]
#> [1] 6
#>
#> [[3]]
#> [1] 9
```

Or, graphically:

---

<sup>1</sup>These are sometimes called Norman doors after Don Norman who described them in his book, “The Design of Everyday Things”. There’s a nice video about them at <https://99percentinvisible.org/article/norman-doors/>.

<sup>2</sup>Not to be confused with `base::Map()`, which is considerably more complex, and we’ll come back to in Section 9.4.5.



You might wonder why this function is called `map()`. What does it have to do with depicting physical features of land or sea? In fact, the meaning comes from mathematics where `map` refers to “an operation that associates each element of a given set with one or more elements of a second set”. This makes sense here because `map()` defines a mapping from one vector to another. (“Map” also has the nice property of being short, which is useful for such a fundamental building block.)

The implementation of `map()` is quite simple. We allocate a list the same length as the input, and then fill in the list with a for loop. The basic implementation is only a handful of lines of code:

```
simple_map <- function(x, f, ...) {
 out <- vector("list", length(x))
 for (i in seq_along(x)) {
 out[[i]] <- f(x[[i]], ...)
 }
 out
}
```

The real `purrr::map()` function has a few differences: it is written in C to eke out every last iota of performance, preserves names, and supports a few shortcuts that you’ll learn about shortly.

The base equivalent to `map()` is `lapply()`. The only difference is that `lapply()` does not support the helpers that you’ll learn about below, so if you’re only using `map()` from `purrr`, you can skip the additional package and use `base::lapply()` directly.

### 9.2.1 Producing atomic vectors

`map()` returns a list. This makes `map()` the most general of the “map” family because you can put anything in a list. There are four more specific variants, `map_lgl()`, `map_int()`, `map_dbl()` and `map_chr()`, that return atomic vectors:

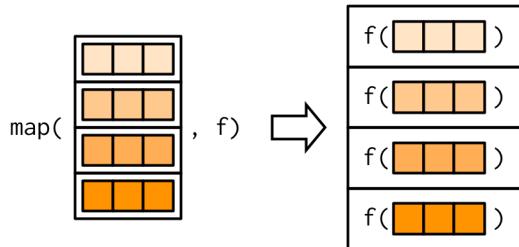
```
map_chr(mtcars, typeof)
#> mpg cyl disp hp drat wt qsec
#> "double" "double" "double" "double" "double" "double" "double"
#> vs am gear carb
#> "double" "double" "double" "double"

map_lgl(mtcars, is.double)
#> mpg cyl disp hp drat wt qsec vs am gear carb
#> TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

map_dbl(mtcars, mean)
#> mpg cyl disp hp drat wt qsec vs
#> 20.091 6.188 230.722 146.688 3.597 3.217 17.849 0.438
#> am gear carb
#> 0.406 3.688 2.812
```

```
n_unique <- function(x) length(unique(x))
map_int(mtcars, n_unique)
#> mpg cyl disp hp drat wt qsec vs am gear carb
#> 25 3 272 22 22 2.9 30 2 2 3 6
```

These examples rely on the fact that data frames are lists containing vectors of the same length:



Like `map()`, the input and the output must be the same length, so you can not return multiple values. When debugging problems like this, it's often useful to switch back to `map()` so you can see what the problematic output is.

```
pair <- function(x) c(x, x)
map_dbl(1:3, pair)
#> Error: Result 1 is not a length 1 atomic vector

map(1:3, pair)
#> [[1]]
#> [1] 1 1
#>
#> [[2]]
#> [1] 2 2
#>
#> [[3]]
#> [1] 3 3

simple_map_dbl <- function(x, f, ...) {
 out <- double(length(x))
 for (i in seq_along(x)) {
 val <- f(x[[i]], ...)
 if (length(val) != 1 || !is.numeric(out)) {
 stop("Result ", i, " is not a length 1 atomic vector", call. = FALSE)
 }
 out[[i]] <- val
 }
 out
}
```

Base R has two similar functions: `sapply()` and `vapply()`.

`sapply()` tries to simplify the result to an atomic vector, wherever possible. But this simplification depends on the input, so sometimes you'll get a list, sometimes a vector, and sometimes a matrix. This makes it difficult to program with.

`vapply()` allows you to provide a template that describes the output shape. If you want to stick to with base R code you should always use `vapply()` in your functions, not `sapply()`. The primary downside of `vapply()` is its verbosity: the equivalent to `map_dbl(x, mean, na.rm = TRUE)` is `vapply(x, mean, na.rm = TRUE, FUN.VALUE = double())`.

### 9.2.2 Anonymous functions and helpers

Instead of using `map()` with an existing function, you can create an inline anonymous function (as mentioned in Section ??first-class-functions)):

```
map_dbl(mtcars, function(x) length(unique(x)))
#> mpg cyl disp hp drat wt qsec vs am gear carb
#> 25 3 27 22 22 29 30 2 2 3 6
```

Anonymous functions are very useful, but the syntax is verbose. So `purrr` offers a shorthand:

```
map_dbl(mtcars, ~ length(unique(.x)))
#> mpg cyl disp hp drat wt qsec vs am gear carb
#> 25 3 27 22 22 29 30 2 2 3 6
```

That also makes for a handy way of generating random data:

```
x <- map(1:3, ~ runif(2))
str(x)
#> List of 3
#> $: num [1:2] 0.281 0.53
#> $: num [1:2] 0.433 0.917
#> $: num [1:2] 0.0275 0.8249
```

Reserve this syntax for short and simple functions. A good rule of thumb is that if your function involves spans lines or uses {}, it's time to name your function.

Inside all `purrr` functions you can create an anonymous function using a `~` (the usual formula operator, pronounced “twiddle”). You can see what happens by calling `as_mapper()`: the `map` functions normally do that for you, but it's useful to do it “by hand” to see what's going on:

```
as_mapper(~ length(unique(.x)))
#> function (... , .x = ..1, .y = ..2, . = ..1)
#> length(unique(.x))
```

The function arguments look a little quirky but allow you to refer to `.` for one argument functions, `.x` and `.y`. for two argument functions, and `..1`, `..2`, `..3`, etc, for functions with an arbitrary number of arguments.

`purrr` also provides helpers for extracting elements from a vector, powered by `purrr::pluck()`. You can use a character vector to select elements by name, an integer vector to select by position, or a list to select by both name and position. These are very useful for working with deeply nested lists, which often arise when working with JSON.

```
x <- list(
 list(-1, x = 1, y = c(2), z = "a"),
 list(-2, x = 4, y = c(5, 6), z = "b"),
 list(-3, x = 8, y = c(9, 10, 11))
)

Select by name
map_dbl(x, "x")
#> [1] 1 4 8

Or by position
map_dbl(x, 1)
#> [1] -1 -2 -3

Or by both
```

```
map_dbl(x, list("y", 1))
#> [1] 2 5 9

You'll get an error if a component doesn't exist:
map_chr(x, "z")
#> Error: Result 3 is not a length 1 atomic vector
Unless you supply a .default value
map_chr(x, "z", .default = NA)
#> [1] "a" "b" NA
```

In base R functions, like `lapply()`, you can provide the name of the function as a string. This isn't tremendously useful as most of the time `lapply(x, "f")` is exactly equivalent to `lapply(x, f)`, just more typing.

### 9.2.3 Passing arguments with ...

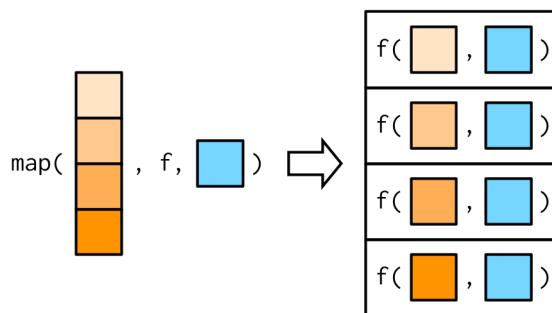
It's often convenient to pass on along additional arguments to the function that you're calling. For example, you might want to pass `na.rm = TRUE` along to `mean()`. One way to do that is with an anonymous function:

```
x <- list(1:5, c(1:10, NA))
map_dbl(x, ~ mean(.x, na.rm = TRUE))
#> [1] 3.0 5.5
```

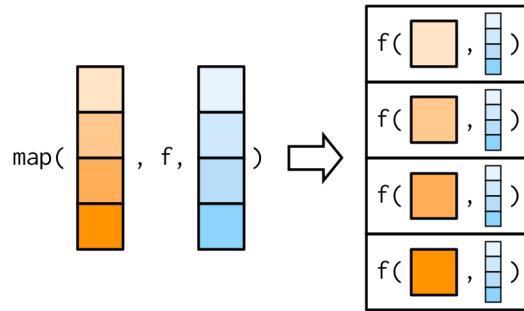
But because the map functions pass `...` along, there's a simpler form available:

```
map_dbl(x, mean, na.rm = TRUE)
#> [1] 3.0 5.5
```

This is easiest to understand with a picture: any arguments that come after `f` in the call to `map()` are inserted *after* the data in individual calls to `f()`:



It's important to note that these arguments are not decomposed; or said another way, `map()` is only vectorised over its first argument. If an argument after `f` is a vector, it will be passed along as is, not decomposed like the first argument:



Note there's a subtle difference between placing extra arguments inside an anonymous function compared with passing them to `map()`. Putting them in anonymous function means that they will be evaluated every time `f()` is executed, not just once when you call `map()`. This is easiest to see if we make the additional argument random:

```
plus <- function(x, y) x + y

x <- c(0, 0, 0, 0)
map_dbl(x, plus, runif(1))
#> [1] 0.0625 0.0625 0.0625 0.0625
map_dbl(x, ~ plus(.x, runif(1)))
#> [1] 0.903 0.132 0.629 0.945
```

#### 9.2.4 Argument names

In the diagrams, I've omitted argument names to focus on the overall structure. But I recommend writing out the full names in your code, as it makes it easier to read. `map(x, mean, 0.1)` is perfectly valid code, but it generates `mean(x[[1]], 0.1)` so it relies on the reader remembering that the second argument to `mean()` is `trim`. To avoid unnecessary burden on the brain of the reader<sup>3</sup>, be kind, and write `map(x, mean, trim = 0.1)`.

This is the reason why the arguments to `map()` are a little odd: instead of being `x` and `f`, they are `.x` and `.f`. It's easiest to the problem that leads to these names using `simple_map()` defined above. `simple_map()` has arguments `x` and `f` so you'll have problems whenever the function you are calling has arguments `x` or `f`:

```
bootstrap_summary <- function(x, f) {
 f(sample(x, replace = TRUE))
}

simple_map(mtcars, bootstrap_summary, f = mean)
#> Error in mean.default(x[[i]]):
#> 'trim' must be numeric of length one
```

The error is a little bewildering until you remember that the call to `simple_map()` is equivalent to `simple_map(x = mtcars, f = mean, bootstrap_summary)` because named matching beats positional matching.

`purrr` functions reduce the likelihood of such a clash by using `.f` and `.x` instead of the more common `f` and `x`. Of course this technique isn't perfect (because the function you are calling might still use `.f` and `.x`), but it avoids 99% of issues. The remaining 1% of the time, use an anonymous function.

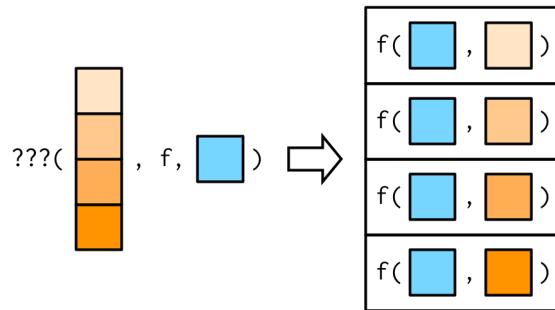
Base functions that pass along `...` use a variety of naming conventions to prevent undesired argument matching:

<sup>3</sup>Who is highly likely to be future you!

- The apply family mostly uses capital letters (e.g `X` and `FUN`).
- `transform()` uses more exotic prefix `_`: this makes the name non-syntactic so it must always be surrounded in ```, as described in Section 2.2.1. This makes undesired matches extremely unlikely.
- Other functional like `uniroot()` and `optim()` make no effort to avoid clashes; but they tend to be used with specially created functions so clashes are less likely.

### 9.2.5 Varying another argument

So far the first argument to `map()` has always become the first argument to the function. But what happens if the first argument should be constant, and you want to vary a different argument? How do you get the result in this picture?



It turns out that there's no way to do it directly, but there are two tricks you can use. To illustrate them, imagine I have a vector that contains a few unusual values, and I want to explore the effect of different amounts of trimming when computing the mean. In this case, the first argument to `mean()` will be constant, and I want to vary the second argument, `trim`.

```
trims <- c(0, 0.1, 0.2, 0.5)
x <- rcauchy(1000)
```

- The simplest technique is to use an anonymous function to rearrange the argument order:
- ```
map_dbl(trims, ~ mean(x, trim = .x))
#> [1] -0.3500  0.0434  0.0354  0.0502
```
- Sometimes, if you want to be (too) clever, you can take advantage of R's flexible argument matching rules (as described in Section 5.8.2). For example, in this example you can rewrite `mean(x, trim = 0.1)` as `mean(0.1, x = x)`, so you could write the call to `map_dbl()` as:
- ```
map_dbl(trims, mean, x = x)
#> [1] -0.3500 0.0434 0.0354 0.0502
```

I don't recommend this technique as it relies on the reader being very familiar with both the argument order to `.f`, and R's argument matching rules.

You'll see one more approach to this problem that in Section 9.4.5.

### 9.2.6 Exercises

1. Use `as_mapper()` to explore how purrr generates anonymous functions for the integer, character, and list helpers. What helper allows you to extract attributes? Read the documentation to find out.
2. `map(1:3, ~ runif(2))` is a useful pattern for generating random numbers, but `map(1:3, runif(2))` is not. Why not? Can you explain why it returns the result that it does?

3. Use the appropriate `map()` function to:
- Compute the standard deviation of every column in a numeric data frame.
  - Compute the standard deviation of every numeric column in a mixed data frame. (Hint: you'll need to do it in two steps.)
  - Compute the number of levels for in every factor in a data frame.
4. The following code simulates the performance of a t-test for non-normal data. Extract the p-value from each test, then visualise.

```
trials <- map(1:100, ~ t.test(rpois(10, 10), rpois(7, 10)))
```

5. The following code uses a `map` nested inside another `map` to apply a function to every element of a nested list. Why does it fail, and what do you need to do to make it work?

```
x <- list(
 list(1, c(3, 9)),
 list(c(3, 6), 7, c(4, 7, 6))
)

triple <- function(x) x * 3
map(x, map, .f = triple)
#> Error in .f(.x[[i]], ...):
#> unused argument (map)
```

6. Use `map()` to fit linear models to the `mtcars` using the formulas stored in this list:

```
formulas <- list(
 mpg ~ disp,
 mpg ~ I(1 / disp),
 mpg ~ disp + wt,
 mpg ~ I(1 / disp) + wt
)
```

7. Fit the model `mpg ~ disp` to each of the bootstrap replicates of `mtcars` in the list below, then extract the  $R^2$  of the model fit (Hint: you can compute the  $R^2$  with `summary()`)

```
bootstrap <- function(df) {
 df[sample(nrow(df), replace = TRUE), , drop = FALSE]
}

bootstraps <- map(1:10, ~ bootstrap(mtcars))
```

## 9.3 Purrr style

Before we go on to take a look at how you tend to use multiple purrr functions to solve a moderately realistic problem: fitting a model to each subgroups and extracting a coefficient of the model.

For this toy example, I'm going to break the `mtcars` data set down into groups defined by the number of cylinders, using the base `split` function:

```
by_cyl <- split(mtcars, mtcars$cyl)
```

Now imagine we want to fit a linear model, then extract the second coefficient (i.e. the `intern`). The following code shows how you might do that with purrr:

```
by_cyl %>%
 map(~ lm(mpg ~ wt, data = .x)) %>%
 map(coef) %>%
 map_dbl(2)
#> 4 6 8
#> -5.65 -2.78 -2.19
```

(If you haven't seen `%>%`, the pipe, before, it's described in Section 5.3.)

I think this code is easy to read because each line encapsulates a single step, you can easily distinguish the functional from what it does, and the purrr helpers allow us to very concisely describe what to do in each step.

How would you attack this problem with base R? You certainly *could* replace each purrr function with the equivalent base function:

```
by_cyl %>%
 lapply(function(data) lm(mpg ~ wt, data = data)) %>%
 lapply(coef) %>%
 vapply(function(x) x[[2]], double(1))
#> 4 6 8
#> -5.65 -2.78 -2.19
```

But this isn't really base R since we're using the pipe. To tackle purely in base I think you'd use an intermediate variable, and do more in each step:

```
models <- lapply(by_cyl, function(data) lm(mpg ~ wt, data = data))
vapply(models, function(x) coef(x)[[2]], double(1))
#> 4 6 8
#> -5.65 -2.78 -2.19
```

Or, of course, you could you use a for loop:

```
intercepts <- double(length(by_cyl))
for (i in seq_along(by_cyl)) {
 model <- lm(mpg ~ wt, data = by_cyl[[i]])
 intercepts[[i]] <- coef(model)[[2]]
}
intercepts
#> [1] -5.65 -2.78 -2.19
```

It's interesting to note that as you move from purrr to base apply functions to for loops you tend to do more and more in each iteration. In purrr we iterate 3 times (`map()`, `map()`, `map_dbl()`), with apply functions we iterate twice (`lapply()`, `vapply()`), and with a for loop we iterate once. The advantage of breaking the problem into smaller steps is that it's easier to understand and later modify as needs change.

## 9.4 Map variants

There are 23 primary variants of `map()`. So far, you've learned about five (`map()`, `map_lgl()`, `map_int()`, `map_dbl()` and `map_chr()`). That means that you've got 18 (!! more to learn. That sounds like a lot, but fortunately the design of purrr means that you only need to learn five new ideas:

- Output same type as input with `modify()`
- Iterate over two inputs with `map2()`.
- Iterate with an index using `imap()`
- Return nothing with `walk()`.

- Iterate over any number of inputs with `pmap()`.

The map family of functions has orthogonal input and outputs, meaning that we can organise all the family into a matrix, with inputs in the rows and outputs in the columns. Once you've mastered the idea in a row, you can combine it with any column; once you've mastered the idea in column, you can combine it with any row.

|                      | List                | Atomic                        | Same type              | Nothing              |
|----------------------|---------------------|-------------------------------|------------------------|----------------------|
| One argument         | <code>map()</code>  | <code>map_lgl()</code> , ...  | <code>modify()</code>  | <code>walk()</code>  |
| Two arguments        | <code>map2()</code> | <code>map2_lgl()</code> , ... | <code>modify2()</code> | <code>walk2()</code> |
| One argument + index | <code>imap()</code> | <code>imap_lgl()</code> , ... | <code>imodify()</code> | <code>iwalk()</code> |
| N arguments          | <code>pmap()</code> | <code>pmap_lgl()</code> , ... | —                      | <code>pwalk()</code> |

### 9.4.1 Same type of output as input: `modify()`

Imagine you wanted to double every column in a data frame. You might first try using `map()`, but `map()` always returns a list:

```
df <- data.frame(
 x = 1:3,
 y = 6:4
)

map(df, ~ .x * 2)
#> $x
#> [1] 2 4 6
#>
#> $y
#> [1] 12 10 8
```

If you want to keep the output as a data frame, you can use `modify()`, which always returns the same type of output as the input:

```
modify(df, ~ .x * 2)
#> x y
#> 1 2 12
#> 2 4 10
#> 3 6 8
```

Despite the name, `modify()` doesn't modify in place, it returns a modified copy, so if you wanted to permanently modify `df`, you'd need to assign it:

```
df <- modify(df, ~ .x * 2)
```

As usual, the basic implementation of `modify()` is simple, and in fact it's even simpler than `map()` because we don't need to create a new output vector; we can just progressively replace the input. The real code is a little complex to handle edge cases more gracefully.

```
simple_modify <- function(x, f, ...) {
 for (i in seq_along(x)) {
 x[[i]] <- f(x[[i]], ...)
 }
 x
}
```

In Section @(predicate-map) you'll learn about a very useful variant of `modify()`, called `modify_if()`. This

allows you to (e.g.) only double *numeric* columns of a data frame with `modify_if(df, is.numeric, ~ .x * 2)`.

### 9.4.2 Two inputs: `map2()` and friends

`map()` is vectorised over a single argument, `.x`. This means it only varies `.x` when calling `.f`, all other arguments are passed along unchanged. This makes it poorly suited for some problems. For example, how would you find a weighted mean when you have a list of observations and a list of weights? Imagine we have the following data:

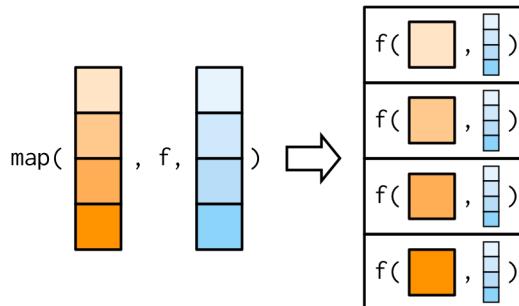
```
xs <- map(1:8, ~ runif(10))
xs[[1]][[1]] <- NA
ws <- map(1:8, ~ rpois(10, 5) + 1)
```

You can use `map_dbl()` to compute the unweighted means:

```
map_dbl(xs, mean)
#> [1] NA 0.463 0.551 0.453 0.564 0.501 0.371 0.443
```

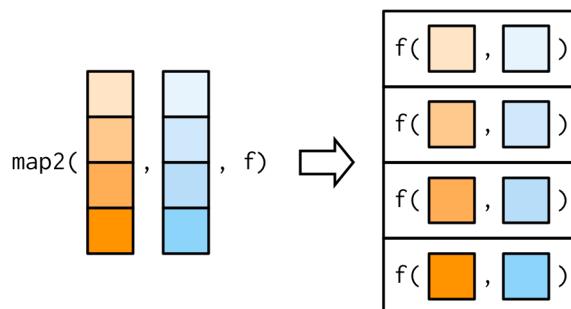
But passing `ws` as an additional argument doesn't work because arguments after `.f` are not transformed:

```
map_dbl(xs, weighted.mean, w = ws)
#> Error in weighted.mean.default(.x[[i]], ...):
#> 'x' and 'w' must have the same length
```



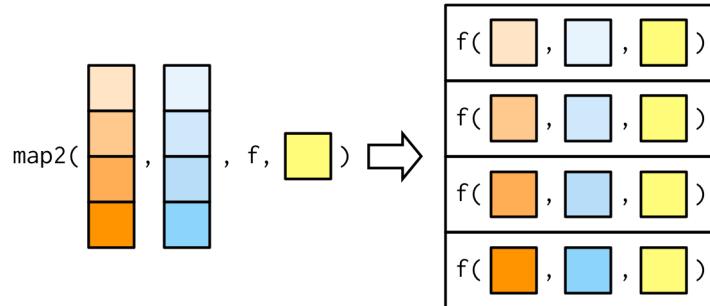
We need a new tool: a `map2()`, which is vectorised over two arguments. This means both `.x` and `.y` are varied in each call to `.f`:

```
map2_dbl(xs, ws, weighted.mean)
#> [1] NA 0.451 0.603 0.452 0.563 0.510 0.342 0.464
```



The arguments to `map2()` are slightly different to the arguments to `map()` as two vectors come before the function, rather than one. Additional arguments still go afterwards:

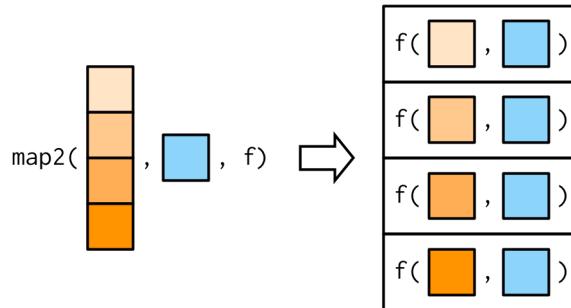
```
map2_dbl(xs, ws, weighted.mean, na.rm = TRUE)
#> [1] 0.504 0.451 0.603 0.452 0.563 0.510 0.342 0.464
```



The basic implementation of `map2()` is simple, and quite similar to that of `map()`. Instead of iterating over one vector, we iterate over two in parallel:

```
simple_map2 <- function(x, y, f, ...) {
 out <- vector("list", length(xs))
 for (i in seq_along(x)) {
 out[[i]] <- f(x[[i]], y[[i]], ...)
 }
 out
}
```

One of the big differences between `map2()` and the simple function above is that `map2()` recycles its inputs to make sure that they're the same length:



In other words, `map2(x, y, f)` will automatically behave like `map(x, f, y)` when needed. This is helpful when writing functions; in scripts you'd generally just use the simpler form directly.

The closest no base equivalent to `map2()` is `Map()`, which is discussed in Section 9.4.5.

### 9.4.3 No outputs: `walk()` and friends

Most functions are called for value that they return, so it makes sense to capture and store it with a `map()` function. But some functions are called primarily for their side-effects (e.g. `cat()`, `write.csv()`, or `ggsave()`) and it doesn't make sense to capture their results. Take this simple example that displays a welcome message using `cat()`. `cat()` returns `NULL`, so while `map` works (in the sense that it generates the desired welcomes), it also returns `list(NULL, NULL)`.

```
welcome <- function(x) {
 cat("Welcome ", x, "!\\n", sep = "")
```

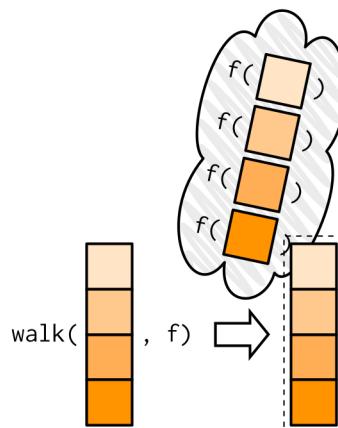
```
names <- c("Hadley", "Jenny")

As well as generate the welcomes, it also shows
the return value of cat()
map(names, welcome)
#> Welcome Hadley!
#> Welcome Jenny!
#> [[1]]
#> NULL
#>
#> [[2]]
#> NULL
```

You could avoid this problem by assigning the results of `map()` to a variable that you never use, but that would muddy the intent of the code. Instead, `purrr` provides the walk family of functions that ignore the return values of the `.f` and instead return `.x` invisibly<sup>4</sup>.

```
walk(names, welcome)
#> Welcome Hadley!
#> Welcome Jenny!
```

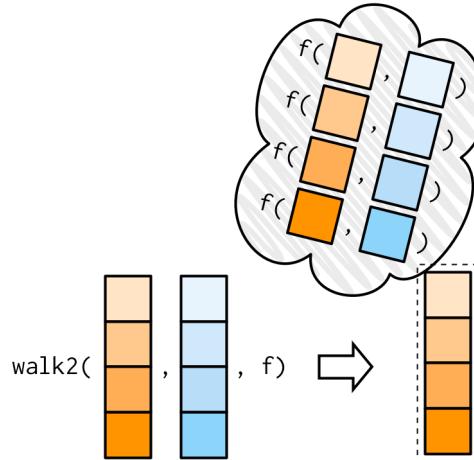
My visual depiction of walk attempts to capture the importance different from `map()`: the outputs are ephemeral, and the input is returned invisibly.



One of the most useful `walk()` variants is `walk2()` because a very common side-effect is saving something to disk, and when saving something to disk you always have a pair of values: the object and the path that you want to save it to.

---

<sup>4</sup>In brief, invisible values are only printed if you explicitly request it. This makes them well suited for functions called primarily for their side-effects, as it allows their output to be ignored by default, while still from an option to capture it. See Section 5.7.2 for more details.



For example, imagine you have a list of data frames (which I've created here using `split`), and you'd like to save each one to a separate csv file. That's easy with `walk2()`:

```
temp <- tempfile()
dir.create(temp)

cyls <- split(mtcars, mtcars$cyl)
paths <- file.path(temp, paste0("cyl-", names(cyls), ".csv"))
walk2(cyls, paths, write.csv)

dir(temp)
#> [1] "cyl-4.csv" "cyl-6.csv" "cyl-8.csv"
```

Here the `walk2()` is equivalent to `write.csv(cyls[[1]], paths[[1]]), write.csv(cyls[[2]], paths[[2]]), write.csv(cyls[[3]], paths[[3]])`.

There is no base equivalent to `walk()`; you can either wrap the result of `lapply()` in `invisible()` or save it to a variable that is never used.

#### 9.4.4 Iterating over values and indices

There are three basic ways to loop over a vector with a for loop:

- Loop over the elements: `for (x in xs)`
- Loop over the numeric indices: `for (i in seq_along(xs))`
- Loop over the names: `for (nm in names(xs))`

The first form is analogous to the `map()` family. The second and third forms are equivalent to the `imap()` family which allows you to iterate over the values and the indices of a vector in parallel.

`imap()` is like `map2()` in the sense that your `.f` gets called with two arguments, but here both are derived from the vector. `imap(x, f)` is equivalent to `map2(x, names(x), f)` if `x` has names, and `map2(x, seq_along(x), f)` if it does not.

`imap()` is often useful for constructing labels:

```
imap_chr(iris, ~ paste0("The first value of ", .y, " is ", .x[[1]]))
#> Sepal.Length
#> "The first value of Sepal.Length is 5.1"
#> Sepal.Width
#> "The first value of Sepal.Width is 3.5"
```

```
#> Petal.Length
#> "The first value of Petal.Length is 1.4"
#> Petal.Width
#> "The first value of Petal.Width is 0.2"
#> Species
#> "The first value of Species is setosa"
```

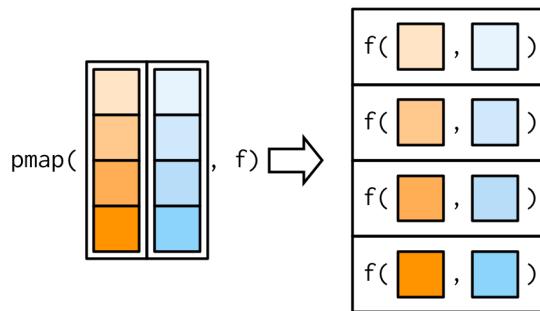
If the vector is unnamed, the second argument will be the index:

```
x <- map(1:6, ~ sample(1000, 10))
imap_chr(x, ~ paste0("The highest value at position ", .y, " is ", max(.x)))
#> [1] "The highest value at position 1 is 885"
#> [2] "The highest value at position 2 is 808"
#> [3] "The highest value at position 3 is 942"
#> [4] "The highest value at position 4 is 966"
#> [5] "The highest value at position 5 is 857"
#> [6] "The highest value at position 6 is 671"
```

`imap()` is a useful helper if you want to work the values in a vector along with their positions.

#### 9.4.5 Any number of inputs: `pmap()` and friends

Since we have `map()` and `map2()`, you might expect `map3()`, `map4()`, `map5()`, and so on. But where would you stop? Instead of generalising to an arbitrary number of arguments, purrr takes a slightly different tack with `pmap()`: you supply it a single list, which contains any number of arguments. In most cases, that will be a list of equal-length vectors, i.e. something very similar to a data frame. In diagrams, I'll emphasise that relationship by drawing the input similar to a data frame.

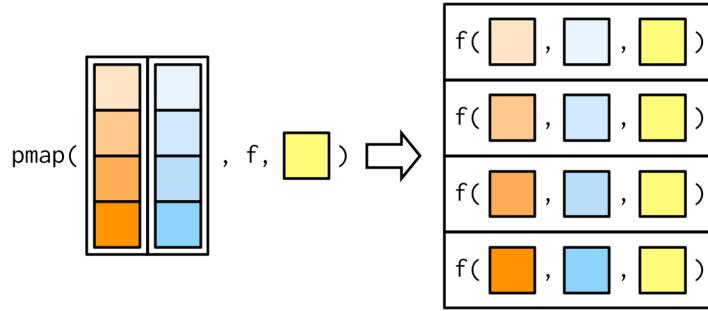


There's a simple equivalence between `map2()` and `pmap()`: `map2(x, y, f)` becomes `pmap(list(x, y), f)`. The `pmap()` equivalent to the `map2_dbl(xs, ws, weighted.mean)` used above is:

```
pmap_dbl(list(xs, ws), weighted.mean)
#> [1] NA 0.451 0.603 0.452 0.563 0.510 0.342 0.464
```

As before, the varying arguments come before `.f` (although now they must be wrapped in a list), and the constant arguments come afterwards.

```
pmap_dbl(list(xs, ws), weighted.mean, na.rm = TRUE)
#> [1] 0.504 0.451 0.603 0.452 0.563 0.510 0.342 0.464
```



A big difference between `pmap()` and the other map functions is that `pmap()` gives you much finer control over argument matching because you can name the components of the list. Returning to our example from Section ??, where we wanted to vary the `trim` argument to `x`, we could instead use `pmap()`:

```
trims <- c(0, 0.1, 0.2, 0.5)
x <- rcauchy(1000)

pmap_dbl(list(trim = trims), mean, x = x)
#> [1] -6.6754 0.0192 0.0228 0.0151
```

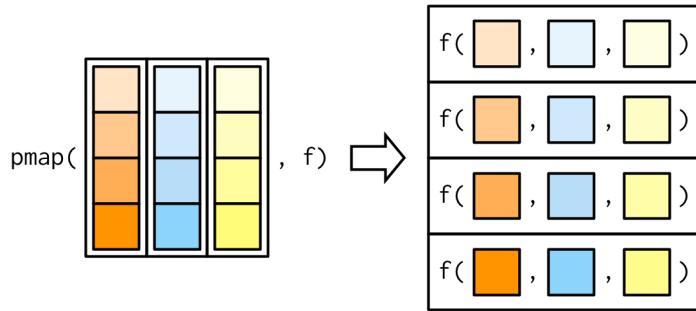
I think it's good practice to name the list to make it very clear how the function will be called.

It's often convenient to call `pmap()` with a data frame. A handy way to create that data frame is with `tibble::tribble()`, which allows you to describe a data frame row-by-row (rather than column-by-column, as usual): thinking about the parameters to a function as a data, is a very powerful pattern. The following example shows how you might draw random uniform numbers with varying parameters:

```
params <- tibble::tribble(
 ~ n, ~ min, ~ max,
 1L, 0, 1,
 2L, 10, 100,
 3L, 100, 1000
)

pmap(params, runif)
#> [[1]]
#> [1] 0.718
#>
#> [[2]]
#> [1] 19.5 39.9
#>
#> [[3]]
#> [1] 535 476 231
```

Here, the column names are critical: I've carefully chosen to match them to the arguments to `runif()`, so the `pmap(params, runif)` is equivalent to `runif(n = 1L, min = 0, max = 1)`, `runif(n = 2, min = 10, max = 100)`, `runif(n = 3L, min = 100, max = 1000)`.



There are two base equivalents to the `pmap()` family: `Map()` and `mapply()`. Both have significant drawbacks:

- `Map()` vectorises over all arguments so you can not supply arguments that do not vary.
- `mapply()` is the multidimensional version of `sapply()`; conceptually it takes the output of `Map()` and simplifies it if possible. This gives it similar issues to `sapply()`, and there's no multi-input equivalent of `vapply()`.

#### 9.4.6 Exercises

1. Explain the results of `modify(mtcars, 1)`.
2. Rewrite the following code to use `iwalk()` instead of `walk2()`. What are the advantages and disadvantages?

```
cyls <- split(mtcars, mtcars$cyl)
paths <- file.path(temp, paste0("cyl-", names(cyls), ".csv"))
walk2(cyls, paths, write.csv)
```
3. Explain how the following code transforms a data frame using functions stored in a list.

```
trans <- list(
 disp = function(x) x * 0.0163871,
 am = function(x) factor(x, labels = c("auto", "manual"))
)

vars <- names(trans)
mtcars[vars] <- map2(trans, mtcars[vars], function(f, var) f(var))
```

Compare and contrast the `map2()` approach to this `map()` approach:

```
mtcars[vars] <- map(vars, ~ trans[[.x]](mtcars[[.x]]))
```

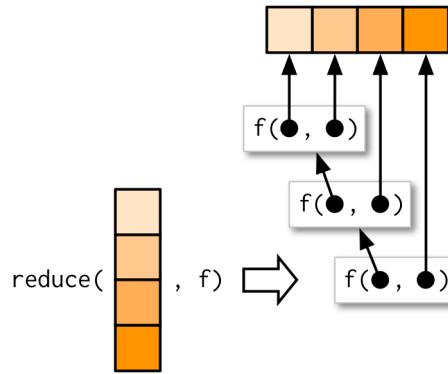
4. What does `write.csv()` return? i.e. what happens if you use it with `map2()` instead of `walk2()`?

## 9.5 Reduce

After the map family, the next most important family of functions is the reduce family. This family is much smaller, with only two main variants, and used less commonly, but it's a powerful idea, gives us the opportunity to discuss some useful algebra, and powers the map-reduce framework frequently used when working with large data.

### 9.5.1 Basics

`reduce()` takes a vector of length  $n$ , and produces a vector of length one, by calling a function with a pair of values at a time. In other words, `reduce(1:4, f)` is equivalent to `f(f(f(1, 2), 3), 4)`.



`reduce()` is a useful way to generalise a function that works with two inputs (a **binary** function) to work with any number of inputs. Imagine you have a list of numeric vectors, and you want to find the values that occur in every element:

```
l <- map(1:4, ~ sample(1:10, 15, replace = T))
str(l)
#> List of 4
#> $: int [1:15] 7 5 9 7 9 9 5 10 5 5 ...
#> $: int [1:15] 6 3 6 10 3 4 4 2 9 9 ...
#> $: int [1:15] 5 3 4 6 1 1 9 9 6 8 ...
#> $: int [1:15] 4 2 6 6 8 5 10 6 7 1 ...
```

To solve this challenge we need to use `intersect()` repeatedly:

```
out <- l[[1]]
out <- intersect(out, l[[2]])
out <- intersect(out, l[[3]])
out <- intersect(out, l[[4]])
out
#> [1] 5 1
```

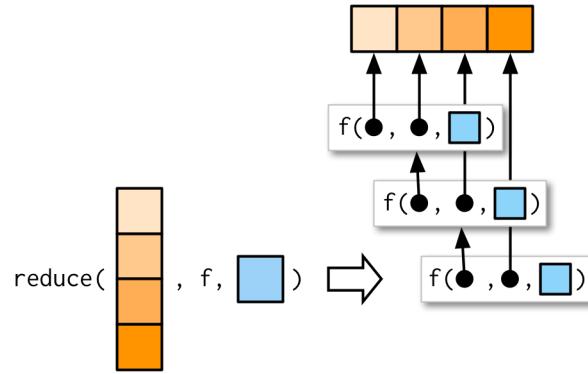
`reduce()` automates this solution for us, so we can write:

```
reduce(l, intersect)
#> [1] 5 1
```

We could apply the same idea if we wanted to list all the elements that appear in at least one entry. All we have to do is switch from `intersect()` to `union()`:

```
reduce(l, union)
#> [1] 7 5 9 10 1 6 3 4 2 8
```

Like the `map` family, you can also pass additional arguments. `intersect()` and `union()` don't take an extra argument so I can't demonstrate them here, but the principle is straight forward and I drew you a picture.



As usual, the essence of `reduce()` can be reduced to a simple wrapper around a for loop:

```
simple_reduce <- function(x, f) {
 out <- x[[1]]
 for (i in seq(2, length(x))) {
 out <- f(out, x[[i]])
 }
 out
}
```

The base equivalent is `Reduce()`. Note that the argument order is different: the function comes first, followed by the vector; there is no way to supply additional arguments.

### 9.5.2 Accumulate

The first `reduce()` variant, `accumulate()`, is useful for understanding how `reduce` works, because instead of return just the final result, it returns all the intermediate results as well:

```
accumulate(1, intersect)
#> [[1]]
#> [1] 7 5 9 7 9 9 5 10 5 5 5 10 9 9 1
#>
#> [[2]]
#> [1] 5 9 10 1
#>
#> [[3]]
#> [1] 5 9 1
#>
#> [[4]]
#> [1] 5 1
```

Another useful way to understand `reduce` is to think about `sum()`: `sum(x)` is equivalent to  $x[[1]] + x[[2]] + x[[3]] + \dots$ . And then `accumulate()` gives you the cumulative sum:

```
x <- c(4, 3, 10)
reduce(x, `+`)
#> [1] 17

accumulate(x, `+`)
#> [1] 4 7 17
```

### 9.5.3 Output types

In the above example using `+`, what should `reduce()` return when `x` is short, i.e. length 1 or 0? When `x` is length 1, `reduce` just returns it without applying the reduce function:

```
reduce(1, `+`)
#> [1] 1
```

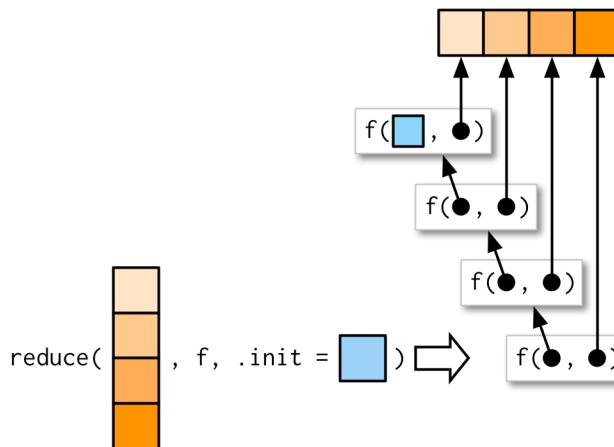
This means that `reduce()` has no way to check that the input is valid:

```
reduce("a", `+`)
#> [1] "a"
```

What if it's length 0? We get an error that suggest we need to use the `.init` argument:

```
reduce(integer(), `+`)
#> Error: `x` is empty, and no `.`init` supplied
```

What should `.init` be here? To figure that out, we need to see what happens when `.init` supplied:



So if we call `reduce(1, +, init)` the result will be `1 + init`. Now we know that the result should be just 1 one, so that suggests that `.init` should be 0:

```
reduce(integer(), `+`, .init = 0)
#> [1] 0
```

This also ensures that `reduce()` checks that length 1 inputs are valid for the function that you're calling:

```
reduce("a", `+`, .init = 0)
#> Error in .x + .y:
#> non-numeric argument to binary operator
```

If you want to get algebraic about it, 0 is called the **identity** of the numbers under the operation of addition: if you add a 0 to any number, you get the same number back. R applies the same principle to determine what a summary function with a zero length input should return:

```
sum(integer()) # x + 0 = x
#> [1] 0
prod(integer()) # x * 1 = x
#> [1] 1
min(integer()) # min(x, Inf) = x
#> [1] Inf
max(integer()) # max(x, -Inf) = x
```

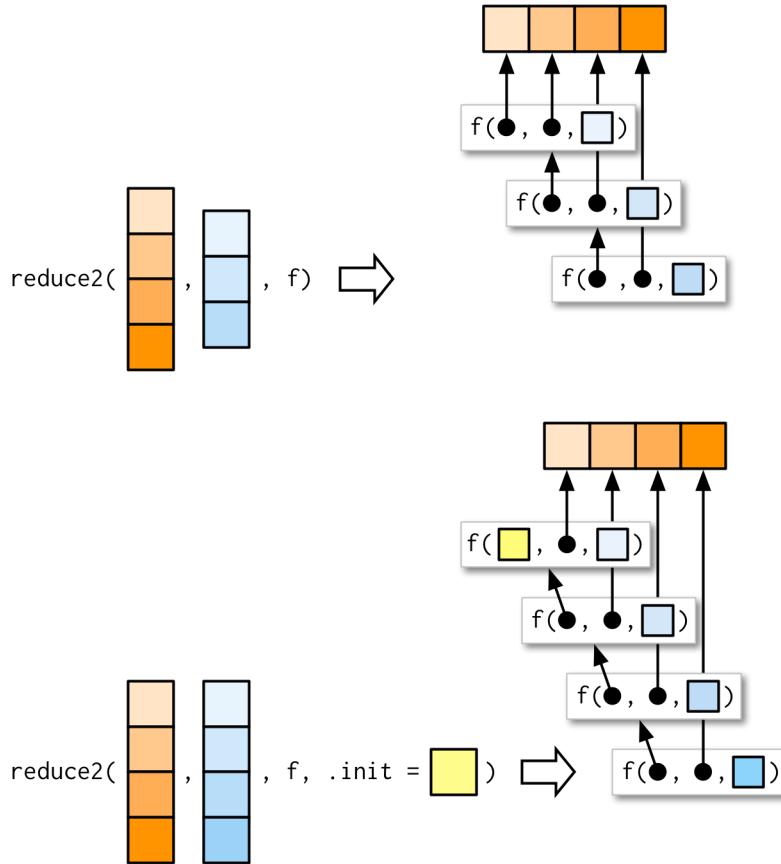
```
#> [1] -Inf
```

If you're using `reduce()` in a function, you should always supply `.init`. Think carefully about what your function should return when passed a vector of length zero or one, and make sure to test your implementation.

#### 9.5.4 Multiple inputs

Very occasionally you need to pass two arguments to the function that you're reducing. For example, you might have a list of data frames that you want to join together, and the variables that you are joining by vary from element to element. This is a very specialised scenario, so I don't want to spend much time on it, except to know that it exists.

Note that the length of the second argument varies based on whether or not `.init` is supplied: if you have four elements of `x`, `f` will only be called three times. If you supply `init`, `f` will be called four times.



#### 9.5.5 Map-reduce

You might have heard of map-reduce, the idea that powers technology like Hadoop. Now you can see how simple and powerful the underlying idea is: all map-reduce is a map combined with a reduce. The special idea for large data is that the data is spread over multiple computers. Each computer performs the map on the data that it has, then it sends the result to back to a coordinator which *reduces* the individual results back to a single result.

## 9.6 Predicate functionals

A **predicate** is a function that returns a single TRUE or FALSE, like `is.character()`, `is.null()`, or `all()`, and we say a predicate **matches** a vector if it returns TRUE.

### 9.6.1 Basics

A **predicate functional** applies a predicate to each element of a vector. purrr provides six useful functions which come in three pairs:

- `some(.x, .p)` returns TRUE if *any* element matches; `every(.x,, .p)` returns TRUE if *all* elements match.
- `detect(.x, .p)` returns the *value* of the first match; `detect_index(.x, .p)` returns the *location* of the first match.
- `keep(.x, .p)` *keeps* all matching elements; `discard(.x, .p)` *drops* all matching elements.

The following example shows how you might use these functionals with a data frame:

```
df <- data.frame(x = 1:3, y = c("a", "b", "c"))
detect(df, is.factor)
#> [1] a b c
#> Levels: a b c
detect_index(df, is.factor)
#> [1] 2

str(keep(df, is.factor))
#> 'data.frame': 3 obs. of 1 variable:
#> $ y: Factor w/ 3 levels "a","b","c": 1 2 3
str(discard(df, is.factor))
#> 'data.frame': 3 obs. of 1 variable:
#> $ x: int 1 2 3
```

All of these functions could be implemented by first computing a logical vector, e.g. `map_lgl(.x, .p)`, and then computing on that. However, that is a little inefficient because you can often exit early. For example, in

### 9.6.2 Map variants

`map()` and `modify()` come in variants that also take predicate functions, transforming only the elements of `.x` with `.p` is TRUE.

```
str(map_if(iris, is.numeric, mean))
#> List of 5
#> $ Sepal.Length: num 5.84
#> $ Sepal.Width : num 3.06
#> $ Petal.Length: num 3.76
#> $ Petal.Width : num 1.2
#> $ Species : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1..
str(modify_if(iris, is.numeric, mean))
#> 'data.frame': 150 obs. of 5 variables:
#> $ Sepal.Length: num 5.84 5.84 5.84 5.84 5.84 ...
#> $ Sepal.Width : num 3.06 3.06 3.06 3.06 3.06 ...
#> $ Petal.Length: num 3.76 3.76 3.76 3.76 3.76 ...
```

```
#> $ Petal.Width : num 1.2 1.2 1.2 1.2 1.2 ...
#> $ Species : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1..
str(map(keep(iris, is.numeric), mean))
#> List of 4
#> $ Sepal.Length: num 5.84
#> $ Sepal.Width : num 3.06
#> $ Petal.Length: num 3.76
#> $ Petal.Width : num 1.2
```

### 9.6.3 Exercises

1. Why isn't `is.na()` a predicate function? What base R function is closest to being a predicate version of `is.na()`?
2. What's the relationship between `which()` and `Position()`? What's the relationship between `where()` and `Filter()`?
3. `simple_reduce()` has a problem when `x` is length 0 or length 1. Describe the source of the problem and how you might go about fixing it.

```
simple_reduce <- function(x, f) {
 out <- x[[1]]
 for (i in seq(2, length(x))) {
 out <- f(out, x[[i]])
 }
 out
}
```

4. Implement the `span()` function from Haskell: given a list `x` and a predicate function `f`, `span()` returns the location of the longest sequential run of elements where the predicate is true. (Hint: you might find `rle()` helpful.)
5. Implement `arg_max()`. It should take a function and a vector of inputs, and return the elements of the input where the function returns the highest value. For example, `arg_max(-10:5, function(x) x ^ 2)` should return -10. `arg_max(-5:5, function(x) x ^ 2)` should return `c(-5, 5)`. Also implement the matching `arg_min()` function.
6. The function below scales a vector so it falls in the range [0, 1]. How would you apply it to every column of a data frame? How would you apply it to every numeric column in a data frame?

```
scale01 <- function(x) {
 rng <- range(x, na.rm = TRUE)
 (x - rng[1]) / (rng[2] - rng[1])
}
```

## 9.7 Base functionals

To finish up the chapter, here I provide a survey of important base functions that are not members of the map, reduce, or predicate families, and hence have no equivalent in purrr. This is not to say that they're not important, but they have more of a mathematical/statistical flavour, so they are generally less useful in data analyses.

### 9.7.1 Matrices and arrays

`map()` and friends are specialised to work with 1d vectors. `base::apply()` is specialised to work with 2d and higher vectors, i.e. matrices and arrays. You can think of `apply()` as an operation that summarises a matrix or array by collapsing each row or column to a single value. It has four arguments:

- `X`, the matrix or array to summarise.
- `MARGIN`, an integer vector giving the dimensions to summarise over, 1 = rows, 2 = columns, etc.
- `FUN`, a summary function.
- ... other arguments passed on to `FUN`.

A typical example of `apply()` looks like this

```
a <- matrix(1:20, nrow = 5)
apply(a, 1, mean)
#> [1] 8.5 9.5 10.5 11.5 12.5
apply(a, 2, mean)
#> [1] 3 8 13 18
```

You can specify multiple dimensions to `MARGINS`, which is useful for high-d arrays:

```
a <- array(1:24, c(2, 3, 4))
apply(a, 1, mean)
#> [1] 12 13
apply(a, c(1, 2), mean)
#> [,1] [,2] [,3]
#> [1,] 10 12 14
#> [2,] 11 13 15
```

There are two caveats to using `apply()`:

- Like `base::sapply()`, you have no control over the output type; it will automatically be simplified to a list, matrix, or vector. However, generally, you used `apply()` with a numeric arrays and numeric summary function so you are less likely to encounter a problem that with `sapply()`.
- `apply()` is also not idempotent in the sense that if the summary function is the identity operator, the output is not always the same as the input.

```
a1 <- apply(a, 1, identity)
identical(a, a1)
#> [1] FALSE
identical(a, t(a1))
#> [1] FALSE

a2 <- apply(a, 2, identity)
identical(a, a2)
#> [1] FALSE
```

- Never use `apply()` with a data frame. It always coerces `X` to a matrix, which will lead to undesirable results if your data frame contains anything other than numbers.

```
df <- data.frame(x = 1:3, y = c("a", "b", "c"))
apply(df, 2, mean)
#> Warning in mean.default(newX[, i], ...): argument is not numeric or
#> logical: returning NA

#> Warning in mean.default(newX[, i], ...): argument is not numeric or
#> logical: returning NA
```

```
#> x y
#> NA NA
```

### 9.7.2 “Ragged” arrays

You can think about `tapply()` as a generalisation to `apply()` that allows for “ragged” arrays, arrays where each row can have a different number of columns. This is often needed when you’re trying to summarise a data set. For example, imagine you’ve collected pulse rate data from a medical trial, and you want to compare the two groups:

```
pulse <- round(rnorm(22, 70, 10 / 3)) + rep(c(0, 5), c(10, 12))
group <- rep(c("A", "B"), c(10, 12))

tapply(pulse, group, length)
#> A B
#> 10 12
tapply(pulse, group, mean)
#> A B
#> 70.8 74.2
```

`tapply()` works by creating a “ragged” data structure from a set of inputs, and then applying a function to the individual elements of that structure. The first task is actually performed by `split()` function does. It takes two inputs and returns a list which groups elements together from the first vector according to elements, or categories, from the second vector:

```
split(pulse, group)
#> $A
#> [1] 68 70 71 72 72 70 73 70 70 72
#>
#> $B
#> [1] 76 74 75 74 75 73 73 73 70 64 78 85
```

Then `tapply()` is just the combination of `split()` and `sapply()`.

### 9.7.3 Mathematical

Functionals are very common in mathematics. The limit, the maximum, the roots (the set of points where  $f(x) = 0$ ), and the definite integral are all functionals: given a function, they return a single number (or vector of numbers). At first glance, these functions don’t seem to fit in with the theme of eliminating loops, but if you dig deeper you’ll find out that they are all implemented using an algorithm that involves iteration.

Base R provides a useful set:

- `integrate()` finds the area under the curve defined by `f()`
- `uniroot()` finds where `f()` hits zero
- `optimise()` finds the location of lowest (or highest) value of `f()`

The following example shows how they might be used with a simple function, `sin()`:

```
integrate(sin, 0, pi)
#> 2 with absolute error < 2.2e-14
str(uniroot(sin, pi * c(1 / 2, 3 / 2)))
#> List of 5
#> $ root : num 3.14
#> $ f.root : num 1.22e-16
```

```
#> $ iter : int 2
#> $ init.it : int NA
#> $ estim.prec: num 6.1e-05
str(optimise(sin, c(0, 2 * pi)))
#> List of 2
#> $ minimum : num 4.71
#> $ objective: num -1
str(optimise(sin, c(0, pi), maximum = TRUE))
#> List of 2
#> $ maximum : num 1.57
#> $ objective: num 1
```

#### 9.7.4 Exercises

1. How does `apply()` arrange the output? Read the documentation and perform some experiments.
2. There's no equivalent to `split() + vapply()`. Should there be? When would it be useful? Implement one yourself.
3. Implement a pure R version of `split()`. (Hint: use `unique()` and subsetting.) Can you do it without a for loop?
4. Challenge: read about the fixed point algorithm ([http://mitpress.mit.edu/sicp/full-text/book/book-Z-H-12.html#%\\_sec\\_1.3](http://mitpress.mit.edu/sicp/full-text/book/book-Z-H-12.html#%_sec_1.3)). Complete the exercises using R.



# Chapter 10

## Function factories

### 10.1 Introduction

A **function factory** is a function that makes functions. Here's a very simple example: we use a function factory (`power1()`) to make two child functions (`square()` and `cube()`):

```
power1 <- function(exp) {
 force(exp)

 function(x) {
 x ^ exp
 }
}

square <- power1(2)
cube <- power1(3)
```

I'll call `square()` and `cube()` **manufactured functions**, but this is just a term to ease communication with other humans: from R's perspective they are no different to functions created any other way.

```
square(3)
#> [1] 9
cube(3)
#> [1] 27
```

You have already learned about the individual components that make function factories possible:

- In Section 5.2.1, you learned about R's “first class” functions. In R, you bind a function to name in the same way as you bind any object to a name: with `<-`.
- In Section 6.4.2, you learned that a function captures (encloses) the environment in which it is created.
- In Section 6.4.4, you learned that a function creates a new execution environment every time it is run. This environment is usually ephemeral, but here it becomes the enclosing environment of the manufactured function.

In this chapter, you'll learn how the non-obvious combination of these three features lead to the function factory. You'll also see examples of their usage in visualisation and statistics.

Of the three main functional programming tools (functionals, function factories, and function operators), function factories are probably the least useful. Generally, they don't tend to reduce overall code complexity.

Instead, they tend to partition complexity into more easily digested chunks. Function factories are also an important building block for the very useful function operators, which you'll learn about in Chapter 11.

## Outline

- Section 10.2 begins the chapter with an explanation of how function factories work, pulling together ideas from scoping and environments. You'll also see how function factories can be used to implement a "memory" for functions, allowing data to persist across calls.
- Section 10.3 illustrates the use of function factories with examples from ggplot2. You'll see two examples of how ggplot2 works with user supplied function factories, and one example of where ggplot2 uses a function factory internally.
- Section 10.4 uses function factories to tackle three challenges from statistics: understanding the Box-Cox transform, solving maximum likelihood problems, and drawing bootstrap resamples.
- Section 10.5 shows how you can combine function factories and functionals to rapidly generate a family of functions from data.

## Prerequisites

Make sure you're familiar with the contents of Sections 5.2.1 (first class functions), 6.4.2 (function environments), and 6.4.4 (execution environments) mentioned above.

Function factories only need base R. We'll use a little rlang to peek inside of them more easily, and we'll use ggplot2 and scales to explore the use of function factories in visualisation.

```
The development version includes some printing tweaks that we need here
devtools::install_github("r-lib/rlang")
library(rlang)

library(ggplot2)
library(scales)
```

## 10.2 Factory fundamentals

The key idea that makes function factories work can be expressed very concisely:

The enclosing environment of the manufactured function is an execution environment of the function factory.

It only takes few words to express these big ideas, but it takes a lot more work to really understand what this means. This section will help you put the pieces together with interactive exploration and some diagrams.

### 10.2.1 Environments

Let's start by taking a look at `square()` and `cube()`:

```
square
#> function(x) {
#> x ^ exp
#> }
#> <environment: 0x4f79eb0>
```

```
cube
#> function(x) {
#> x ^ exp
#> }
#> <bytecode: 0x5326690>
#> <environment: 0x4fd1280>
```

Printing manufactured functions is not revealing because the bodies are identical; it's the contents of the enclosing environment that's important. We can get a little more insight by using `rlang::env_print()`. That shows us that we have two different environments (each of which was originally an execution environment of `power()`). The environments have the same parent, which is the enclosing environment of `power1()`, the global environment.

```
env_print(square)
#> <environment: 0x4f79eb0>
#> parent: <environment: global>
#> bindings:
#> * exp: <dbl>

env_print(cube)
#> <environment: 0x4fd1280>
#> parent: <environment: global>
#> bindings:
#> * exp: <dbl>
```

`env_print()` shows us that both environments have a binding to `exp`, but we want to see its value<sup>1</sup>. That's easily done with `env_get()`:

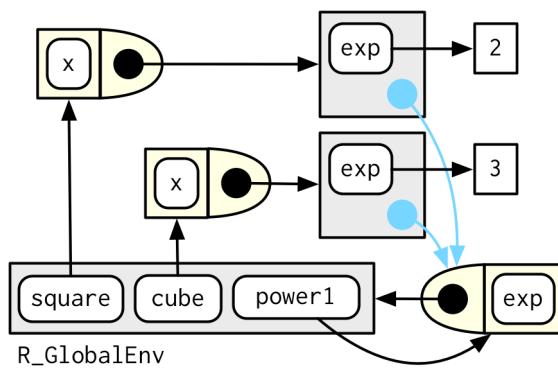
```
env_get(square, "exp")
#> [1] 2

env_get(cube, "exp")
#> [1] 3
```

This is what makes manufactured functions behave differently from one another: names in the enclosing environment are bound to different values.

### 10.2.2 Diagram conventions

We can also show these relationships in a diagram:

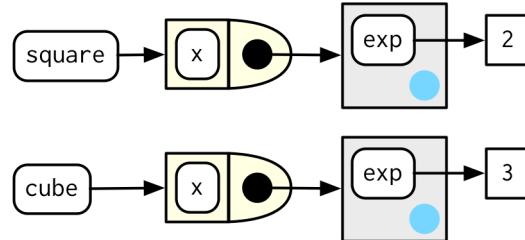



---

<sup>1</sup>A future version of `env_print()` is likely to do better at summarising the contents so you don't need this step.

There's a lot going on in this diagram and some of the details aren't that important. We can simplify considerably by using two conventions:

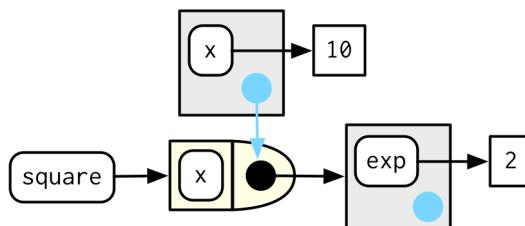
- Any free floating symbol lives in the global environment.
- Any environment without an explicit parent inherits from the global environment.



This view, which focuses on the environments, doesn't show any direct link between `cube()` and `square()`. That's because the link is through the body of the function, which is identical for both, but is not shown in this diagram.

To finish up, let's look at the execution environment of `square(10)`. When `square()` executes `x ^ exp` it finds `x` in the execution environment and `exp` in its enclosing environment.

```
square(10)
#> [1] 100
```



### 10.2.3 Stateful functions

Function factories also allow you to maintain state across function invocations, which is generally hard to do because of the fresh start principle described in Section 5.4.3.

There are two things that make this possible:

- The enclosing environment of the manufactured function is unique and constant.
- R has a special assignment operator, `<-<`, which modifies bindings in the enclosing environment.

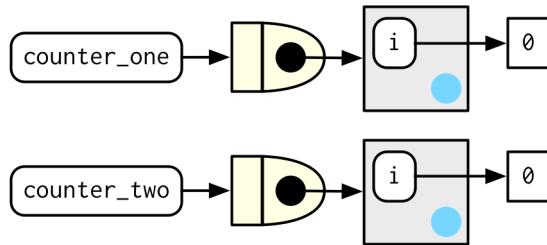
The usual assignment operator, `<-`, always creates a binding in the current environment. The **super assignment operator**, `<-<` rebinds an existing name found in a parent environment.

The following example shows how we can combine these ideas to create a function that records how many times it has been called:

```
new_counter <- function() {
 i <- 0

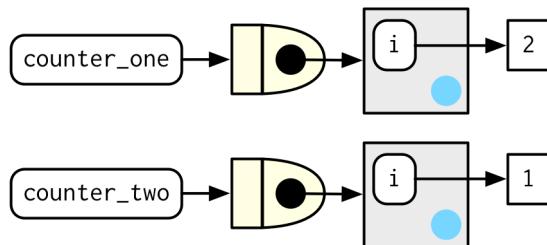
 function() {
 i <-< i + 1
 i
 }
}
```

```
counter_one <- new_counter()
counter_two <- new_counter()
```



When the manufactured function is run `i <= i + 1` will modify `i` in its enclosing environment. Because manufactured functions have independent enclosing environments, they have independent counts:

```
counter_one()
#> [1] 1
counter_one()
#> [1] 2
counter_two()
#> [1] 1
```



Stateful functions are best used in moderation. As soon as your function starts managing the state of multiple variables, it's better to switch to R6, the topic of Chapter 15.

#### 10.2.4 Potential pitfalls

There are two potential pitfalls to be aware of when creating your own function factories: forgetting to evaluate all inputs and accidentally capturing large objects.

Generally, you can rely on lazy evaluation to evaluate function inputs at the right time. However, there's a catch when it comes to function factories: if you don't eagerly evaluate every argument, it's possible to get confusing behaviour, as shown below.

```
power2 <- function(exp) {
 function(x) {
 x ^ exp
 }
}

exp2 <- 2
square2 <- power2(exp2)
exp2 <- 3
```

```
square2(2)
#> [1] 8
```

This is described in Section 5.5.1, and happens when a binding changes in between calling the factory function and calling the manufactured function. This is likely to only happen rarely, but when it does, it will lead to a real head-scratcher of a bug. Avoid future pain by ensuring every argument is evaluated, using `force()` if the argument is only used by the manufactured function.

With most functions, you can rely on the GC to clean up any large temporary objects created inside a function. However, manufactured functions hold on to the execution environment, so you'll need to explicitly unbind any large temporary objects with `rm()`. Compare the sizes of `g1()` and `g2()` in the example below:

```
f1 <- function(n) {
 x <- runif(n)
 m <- mean(x)
 function() m
}

g1 <- f1(1e6)
lobstr::obj_size(g1)
#> 8,013,720 B

f2 <- function(n) {
 x <- runif(n)
 m <- mean(x)
 rm(x)
 function() m
}

g2 <- f2(1e6)
lobstr::obj_size(g2)
#> 13,560 B
```

### 10.2.5 Exercises

1. Base R contains two function factories, `approxfun()` and `ecdf()`. Read their documentation and experiment to figure out what the functions do and what they return.
2. Create a function `pick()` that takes an index, `i`, as an argument and returns a function with an argument `x` that subsets `x` with `i`.

```
pick(1)(x)
should be equivalent to
x[[1]]

lapply(mtcars, pick(5))
should be equivalent to
lapply(mtcars, function(x) x[[5]])
```

3. Create a function that creates functions that compute the  $i^{\text{th}}$  central moment ([http://en.wikipedia.org/wiki/Central\\_moment](http://en.wikipedia.org/wiki/Central_moment)) of a numeric vector. You can test it by running the following code:

```
m1 <- moment(1)
m2 <- moment(2)
```

```
x <- runif(100)
stopifnot(all.equal(m1(x), 0))
stopifnot(all.equal(m2(x), var(x) * 99 / 100))
```

4. What happens if you don't use a closure? Make predictions then, verify with the code below.

```
i <- 0
new_counter2 <- function() {
 i <-> i + 1
 i
}
```

5. What happens if you use `<-` instead of `<->`? Make predictions, then verify with the code below.

```
new_counter3 <- function() {
 i <- 0
 function() {
 i <- i + 1
 i
 }
}
```

## 10.3 Graphical factories

We'll begin our exploration of useful function factories with a few examples from `ggplot2`.

### 10.3.1 Labelling

One of the goals of the `scales` (<http://scales.r-lib.org>) package is to make it easy to customise the labels on `ggplot2`. It provides many functions to control the fine details of axes and legends. One useful class of functions are the formatter functions<sup>2</sup> which make it easier to control the appearance of axis breaks. The design of these functions might initially seem a little odd: they all return a function, which you have to call in order to format a number.

```
y <- c(12345, 123456, 1234567)
comma_format()(y)
#> [1] "12,345" "123,456" "1,234,567"

number_format(scale = 1e-3, suffix = " K")(y)
#> [1] "12 K" "123 K" "1 235 K"
```

In other words, the primary interface is a function factory. At first glance, this seems to add extra complexity for little gain. But it enables a nice interaction with `ggplot2`'s scales, because they accept functions in the `label` argument:

```
df <- data.frame(x = 1, y = y)
core <- ggplot(df, aes(x, y)) +
 geom_point() +
 scale_x_continuous(breaks = 1, labels = NULL) +
 labs(x = NULL, y = NULL)
```

---

<sup>2</sup>It's an unfortunate accident of history that `scales` uses function suffixes instead of function prefixes. That's because it was written before I understood the autocomplete advantages to using common prefixes instead of common suffixes.

```
core
core + scale_y_continuous(label = comma_format())
core + scale_y_continuous(label = number_format(scale = 1e-3, suffix = " K"))
core + scale_y_continuous(label = scientific_format())
```



### 10.3.2 Histogram bins

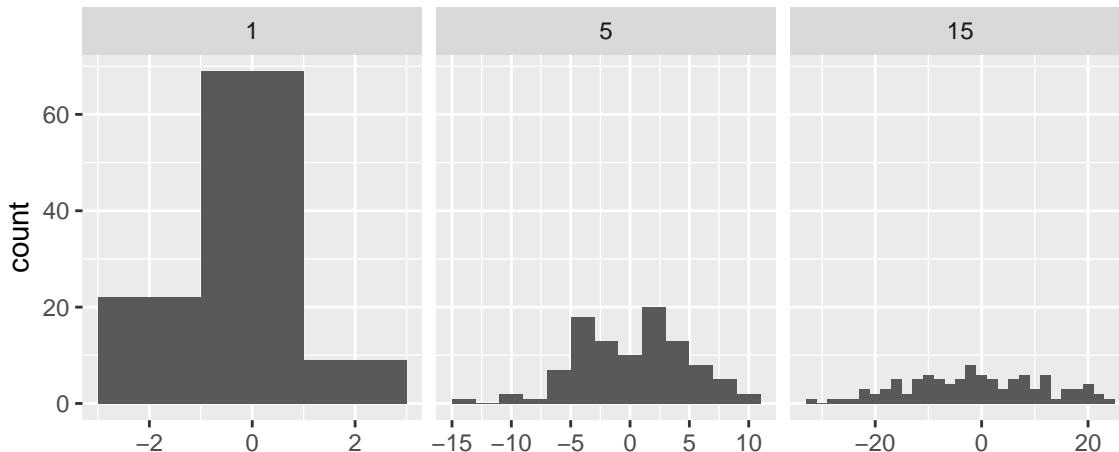
A little known feature of `geom_histogram()` is that the `binwidth` argument can be a function. This is particularly useful because the function is executed once for each group, which means you can have different binwidths in different facets, which is otherwise not possible.

To illustrate this idea, and see where variable binwidth might be useful, I'm going to construct an example where a fixed binwidth isn't great.

```
construct some sample data with very different numbers in each cell
sd <- c(1, 5, 15)
n <- 100

df <- data.frame(x = rnorm(3 * n, sd = sd), sd = rep(sd, n))

ggplot(df, aes(x)) +
 geom_histogram(binwidth = 2) +
 facet_wrap(~ sd, scales = "free_x") +
 labs(x = NULL)
```

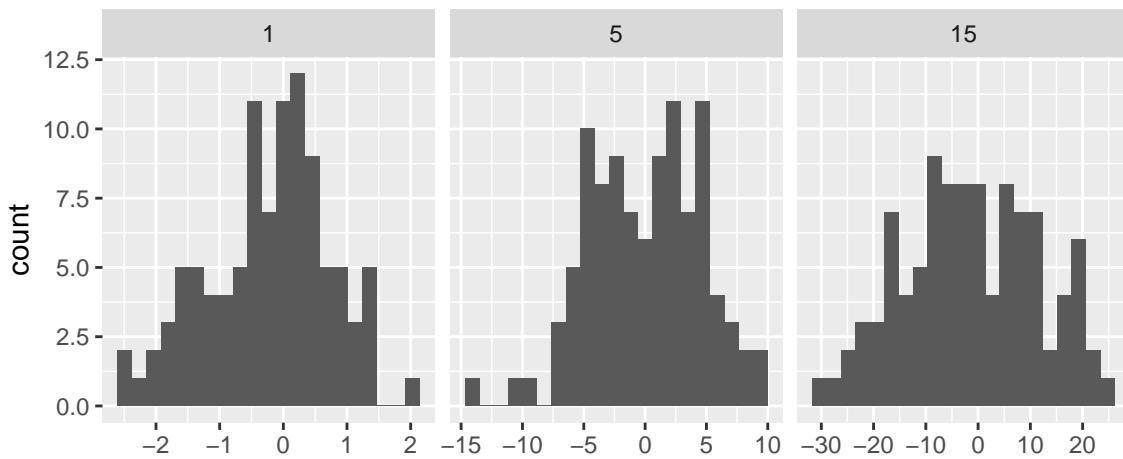


Here each facet has the same number of observations, but the variability is very different. It would be nice if we could request that the binwidths vary so we get approximately the same number of observations in each bin. One way to do that is with a function factory that inputs the desired number of bins (`n`), and outputs a function that takes a numeric vector and returns a binwidth:

```
binwidth_bins <- function(n) {
 force(n)

 function(x) {
 (max(x) - min(x)) / n
 }
}

ggplot(df, aes(x)) +
 geom_histogram(binwidth = binwidth_bins(20)) +
 facet_wrap(~ sd, scales = "free_x") +
 labs(x = NULL)
```



We could use this same pattern to wrap around the base R functions that automatically find the “optimal”<sup>3</sup> binwidth, `nclass.Sturges()`, `nclass.scott()`, and `nclass.FD()`:

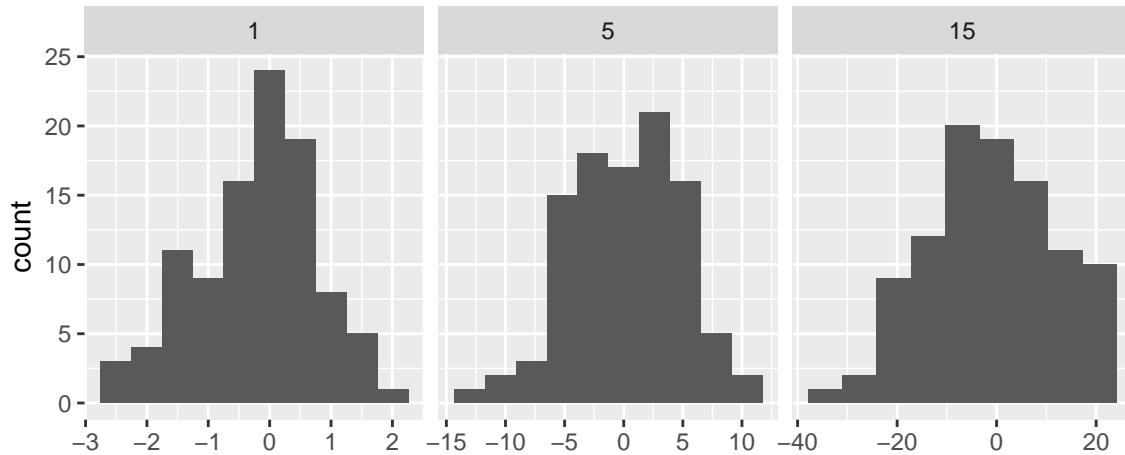
```
base_bins <- function(type) {
 fun <- switch(type,
 Sturges = nclass.Sturges,
 scott = nclass.scott,
 FD = nclass.FD,
 stop("Unknown type", call. = FALSE)
)

 function(x) {
 (max(x) - min(x)) / fun(x)
 }
}

ggplot(df, aes(x)) +
 geom_histogram(binwidth = base_bins("FD")) +
 facet_wrap(~ sd, scales = "free_x") +
 labs(x = NULL)
```

---

<sup>3</sup> ggplot2 doesn’t expose these functions directly because I don’t think the definition of optimality needed to make the problem mathematically tractable is a good match to the actual needs of data exploration.



### 10.3.3 ggsave

Finally, I want to show a function factory used internally by `ggplot2`. `ggplot2:::plot_dev()` is used by `ggsave()` to go from a file extension (e.g. `png`, `jpeg` etc) to a graphics device function (e.g. `png()`, `jpeg()`). The challenge here arises because the base graphics devices have some minor inconsistencies which we need to paper over:

- Most have `filename` as first argument but some have `file`.
- The `width` and `height` of raster graphic devices use pixels units by default, but the vector graphics use inches.

A mildly simplified version of `plot_dev()` is shown below:

```
plot_dev <- function(ext, dpi = 96) {
 force(dpi)

 switch(ext,
 eps = ,
 ps = function(filename, ...) {
 grDevices:::postscript(
 file = filename, ..., onefile = FALSE,
 horizontal = FALSE, paper = "special"
)
 },
 tex = function(filename, ...) grDevices:::pictex(file = filename, ...),
 pdf = function(filename, ...) grDevices:::pdf(file = filename, ...),
 svg = function(filename, ...) svglite:::svglite(file = filename, ...),
 emf = ,
 wmf = function(...) grDevices:::win.metafile(...),
 png = function(...) grDevices:::png(..., res = dpi, units = "in"),
 jpg = ,
 jpeg = function(...) grDevices:::jpeg(..., res = dpi, units = "in"),
 bmp = function(...) grDevices:::bmp(..., res = dpi, units = "in"),
 tiff = function(...) grDevices:::tiff(..., res = dpi, units = "in"),
 stop("Unknown graphics extension: ", ext, call. = FALSE)
)
}

plot_dev("pdf")
```

```
#> function(filename, ...) grDevices::pdf(file = filename, ...)
#> <bytecode: 0x43725e8>
#> <environment: 0x3e38250>
plot_dev("png")
#> function(...) grDevices::png(..., res = dpi, units = "in")
#> <bytecode: 0x457c3f0>
#> <environment: 0x49f6768>
```

### 10.3.4 Exercises

1. Compare and contrast `ggplot2::label_bquote()` with `scales::number_format()`

## 10.4 Statistical factories

More motivating examples for function factories come from statistics:

- The Box-Cox transformation.
- Bootstrap resampling.
- Maximum likelihood estimation.

All of these examples can be tackled without function factories, but I think function factories are a good fit for these problems and provide elegant solutions. These examples expect some statistical background, so feel free to skip if they don't make much sense to you.

### 10.4.1 Box-Cox transformation

The Box-Cox transformation is a flexible transformation often used to transform data towards normality. It has a single parameter,  $\lambda$  which controls the strength of the transformation. We could express the transformation as a simple two argument function:

```
boxcox1 <- function(x, lambda) {
 stopifnot(length(lambda) != 1)

 if (lambda == 0) {
 log(x)
 } else {
 (x ^ lambda - 1) / lambda
 }
}
```

But re-formulating as a function factory makes it easy to explore its behaviour with `stat_function()`:

```
boxcox2 <- function(lambda) {
 if (lambda == 0) {
 function(x) log(x)
 } else {
 function(x) (x ^ lambda - 1) / lambda
 }
}

stat_boxcox <- function(lambda) {
 stat_function(aes(colour = lambda), fun = boxcox2(lambda), size = 1)
```

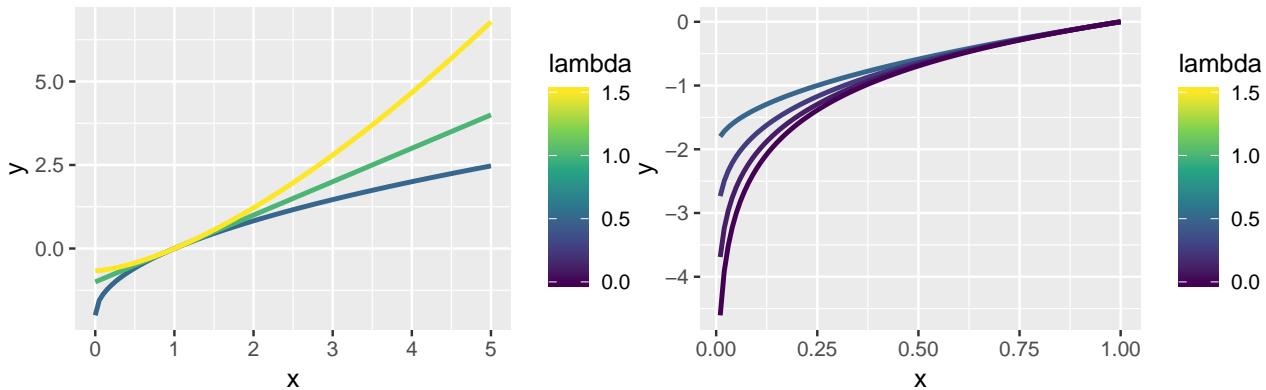
```

}

ggplot(data.frame(x = c(0, 5)), aes(x)) +
 lapply(c(0.5, 1, 1.5), stat_boxcox) +
 scale_colour_viridis_c(limits = c(0, 1.5))

visually, log() does seem to make sense as the limit as lambda -> 0
ggplot(data.frame(x = c(0.01, 1)), aes(x)) +
 lapply(c(0.5, 0.25, 0.1, 0), stat_boxcox) +
 scale_colour_viridis_c(limits = c(0, 1.5))

```



In general, this allows you to use a Box-Cox transformation with any function that accepts a unary transformation function: you don't have to worry about that function providing ... to pass along additional arguments. I also think that the partitioning of `lambda` and `x` into two different function arguments is natural since `lambda` plays quite a different role than `x`.

### 10.4.2 Bootstrap generators

Function factories are a useful approach for bootstrapping. Instead of thinking about a single bootstrap (you always need more than one!), you can think about a bootstrap **generator**, a function that yields a fresh bootstrap every time it is called:

```

boot_permute <- function(df, var) {
 n <- nrow(df)
 force(var)

 function() {
 df[[var]][sample(n, n, replace = TRUE)]
 }
}

boot_mtcars1 <- boot_permute(mtcars, "mpg")
head(boot_mtcars1())
#> [1] 18.1 22.8 21.5 14.7 21.4 17.3
head(boot_mtcars1())
#> [1] 19.2 19.2 14.3 21.0 13.3 21.4

```

The advantage of a function factory is more clear with a parametric bootstrap where we have to first fit a model. We can do this setup step once, when the factory is called, rather than once every time we generate the bootstrap:

```

boot_model <- function(df, formula) {
 mod <- lm(formula, data = df)
 fitted <- unname(fitted(mod))
 resid <- unname(resid(mod))
 rm(mod)

 function() {
 fitted + sample(resid)
 }
}

boot_mtcars2 <- boot_model(mtcars, mpg ~ wt)
head(boot_mtcars2())
#> [1] 23.1 24.3 23.0 19.1 19.1 16.2
head(boot_mtcars2())
#> [1] 30.2 17.4 31.3 26.1 17.8 16.7

```

I use `rm(mod)` because linear model objects are quite large (they include complete copies of the model matrix and input data) and I want to keep the manufactured function as small as possible.

### 10.4.3 Maximum likelihood estimation

The goal of maximum likelihood estimation (MLE) is to find the parameter values for a distribution that make the observed data “most likely”. To do MLE, you start with a probability function. For example, take the Poisson distribution. If we know  $\lambda$ , we can compute the probability of getting a vector  $\mathbf{x}$  of values  $(x_1, x_2, \dots, x_n)$  by multiplying the Poisson probability function as follows:

$$P(\lambda, \mathbf{x}) = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

In statistics, we almost always work with the log of this function. The log is a monotonic transformation which preserves important properties (i.e. the extrema occur in the same place), but has specific advantages:

- The log turns a product into a sum, which is easier to work with.
- Multiplying small numbers yields even smaller numbers, which makes the floating point approximation used by a computer less accurate.

Let’s apply a log transformation to this probability function and simplify it as much as possible:

$$\log(P(\lambda, \mathbf{x})) = \sum_{i=1}^n \log\left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right)$$

$$\log(P(\lambda, \mathbf{x})) = \sum_{i=1}^n (x_i \log(\lambda) - \lambda - \log(x_i!))$$

$$\log(P(\lambda, \mathbf{x})) = \sum_{i=1}^n x_i \log(\lambda) - \sum_{i=1}^n \lambda - \sum_{i=1}^n \log(x_i!)$$

$$\log(P(\lambda, \mathbf{x})) = \log(\lambda) \sum_{i=1}^n x_i - n\lambda - \sum_{i=1}^n \log(x_i!)$$

We can now turn this function into an R function. The R function is quite elegant because R is vectorised and, because it's a statistical programming language, R comes with built-in functions like the log-factorial (`lfactorial()`).

```
lprob_poisson <- function(lambda, x) {
 n <- length(x)
 (log(lambda) * sum(x)) - (n * lambda) - sum(lfactorial(x))
}
```

Consider this vector of observations:

```
x1 <- c(41, 30, 31, 38, 29, 24, 30, 29, 31, 38)
```

We can use `lprob_poisson()` to compute the (logged) probability of `x1` for different values of `lambda`.

```
lprob_poisson(10, x1)
#> [1] -184
lprob_poisson(20, x1)
#> [1] -61.1
lprob_poisson(30, x1)
#> [1] -31
```

So far we've been thinking of `lambda` as fixed and known and the function told us the probability of getting different values of `x`. But in real-life, we observe `x` and it is `lambda` that is unknown. The likelihood is the probability function seen through this lens: we want to find the `lambda` that makes the observed `x` the “most likely”. That is, given `x`, what value of `lambda` gives us the highest value of `lprob_poisson()`?

In statistics, we highlight this change in perspective by writing  $f_x(\lambda)$  instead of  $f(\lambda, x)$ . In R, we can use a function factory. We provide `x` and generate a function with a single parameter, `lambda`:

```
ll_poisson1 <- function(x) {
 n <- length(x)

 function(lambda) {
 log(lambda) * sum(x) - n * lambda - sum(lfactorial(x))
 }
}
```

One nice thing about this approach is that we can do some precomputation: any term that only involves `x` can be computed once in the factory. This is useful because we're going to need to call this function many times to find the best `lambda`.

```
ll_poisson2 <- function(x) {
 n <- length(x)
 sum_x <- sum(x)
 c <- sum(lfactorial(x))

 function(lambda) {
 log(lambda) * sum_x - n * lambda - c
 }
}
```

Now we can use this function to find the value of `lambda` that maximizes the (log) likelihood:

```
ll1 <- ll_poisson2(x1)

ll1(10)
#> [1] -184
ll1(20)
```

```
#> [1] -61.1
l11(30)
#> [1] -31
```

Rather than trial and error, we can automate the process of finding the best value with `optimise()`. It will evaluate `l11()` many times, using mathematical tricks to narrow in on the largest value as quickly as possible. The results tell us that the highest value is  $-30.27$  which occurs when `lambda = 32.1`:

```
optimise(l11, c(0, 100), maximum = TRUE)
#> $maximum
#> [1] 32.1
#>
#> $objective
#> [1] -30.3
```

Now, we could have solved this problem without using a function factory because `optimise()` passes ... on to the function being optimised. That means we could use the log-probability function directly:

```
optimise(lprob_poisson, c(0, 100), x = x1, maximum = TRUE)
#> $maximum
#> [1] 32.1
#>
#> $objective
#> [1] -30.3
```

The advantage of using a function factory here is fairly small, but there are two niceties:

- We can precompute some values in the factory itself, saving computation time in each iteration.
- I think the two-level design better reflects the mathematical structure of the underlying problem.

These advantages get bigger in more complex MLE problems, where you have multiple parameters and multiple data vectors.

#### 10.4.4 Exercises

1. In `boot_model()`, why don't I need to force the evaluation of `df` or `model`?
2. Why might you formulate the Box-Cox transformation like this?

```
boxcox3 <- function(x) {
 function(lambda) {
 if (lambda == 0) {
 log(x)
 } else {
 (x ^ lambda - 1) / lambda
 }
 }
}
```

3. Why don't you need to worry that `boot_permute()` stores a copy of the data inside the function that it generates?
4. How much time does `ll_poisson2()` save compared to `ll_poisson1()`. Use `bench::mark()` to see how much faster the optimisation occurs. How does changing the length of `x` change the results?

## 10.5 Function factories + functionals

To finish off the chapter, I'll show how you might combine functionals and function factories to turn data into many functions. The following code creates many specially named power functions by iterating over a list of arguments:

```
names <- list(
 square = 2,
 cube = 3,
 root = 1/2,
 cuberoot = 1/3,
 reciprocal = -1
)
funss <- purrr::map(names, power1)

funss$root(64)
#> [1] 8
funss$root
#> function(x) {
#> x ^ exp
#> }
#> <bytecode: 0x5326690>
#> <environment: 0x2f682f8>
```

This idea extends in a straightforward way if your function factory takes two (replace `map()` with `map2()`) or more (replace with `pmap()`) arguments.

### 10.5.1 Moving a lists to the global environment

One downside of the current construction is that you have to prefix every function call with `funss$`. There are three ways to eliminate this additional syntax:

- For a very temporary effect, you can use `with()`:

```
with(funss, root(100))
#> [1] 10
```

I recommend this because it makes it very clear when code is being executed in a special context and what that context is.

- For a longer effect, you can `attach()` the functions to the search path, then `detach()` when you're done:

```
attach(funss)
#> The following objects are masked _by_ .GlobalEnv:
#>
#> cube, square
root(100)
#> [1] 10
detach(funss)
```

You've probably been told to avoid using `attach()`, and that's generally good advice. However, the situation is a little different to the usual because we're attaching a list functions, not a data frame. It's less likely that you'll modify a function than a column in a data frame, so some of the worst problems with `attach()` don't apply.

- Finally, you could copy the functions to the global environment with `env_bind()`. This is mostly permanent:

```
rlang::env_bind(globalenv(), !!!fun)
root(100)
#> [1] 10
```

You can later unbind those same names, but there's no guarantee that they haven't been rebound in the meantime, and you might be deleting an object that someone else created.

```
rlang::env_unbind(globalenv(), names(fun))
```

### 10.5.2 Another approach

You'll learn an alternative approach to the same problem in Section 18.6.4. Instead of using a function factory, you could construct the function with quasiquotation. This requires additional knowledge, but generates functions with readable bodies, and avoids accidentally capturing large objects in the enclosing scope. The following code is a quick preview of how we could rewrite `power1()` to use quasiquotation:

```
power3 <- function(exponent) {
 new_function(
 exprs(x =),
 expr({
 x ^ !!exponent
 }),
 caller_env()
)
}
fun <- purrr::map(names, power3)

fun$root
#> function (x)
#> {
#> x^0.5
#> }
#> <environment: 0x3500300>
```

As well as 0.5 appearing directly in the body, note that the environment of the function is the global environment, not an execution environment of `power3()`.

### 10.5.3 Exercises

1. Which of the following commands is equivalent to `with(x, f(z))`?
  - `x$f(x$z)`.
  - `f(x$z)`.
  - `x$f(z)`.
  - `f(z)`.
  - It depends.
2. Compare and contrast the effects of `env_bind()` vs. `attach()` for the following code.

```
fun <- list(
 mean = function(x) mean(x, na.rm = TRUE),
 sum = function(x) sum(x, na.rm = TRUE)
)
```

```
attach(funns)
#> The following objects are masked from package:base:
#>
#> mean, sum
mean <- function(x) stop("Hi!")
detach(funns)

env_bind(globalenv(), !!!funns)
mean <- function(x) stop("Hi!")
env_unbind(globalenv(), names(funns))
```

# Chapter 11

## Function operators

### 11.1 Introduction

In this chapter, you'll learn about function operators (FOs). A function operator is a function that takes one (or more) functions as input and returns a function as output. The following code shows a simple function operator, `chatty()`. It wraps a function, making a new function that prints out its first argument. You might create a function like this because it gives you a window to see how functionals, like `map_int()`, work.

```
chatty <- function(f) {
 force(f)

 function(x, ...) {
 res <- f(x, ...)
 cat("Processing ", x, "\n", sep = "")
 res
 }
}
f <- function(x) x ^ 2
s <- c(3, 2, 1)

purrr::map_dbl(s, chatty(f))
#> Processing 3
#> Processing 2
#> Processing 1
#> [1] 9 4 1
```

Function operators are closely related to function factories; indeed they're just a function factory that takes a function as a input. As well as being built from the same building blocks, there's nothing you can't do without them, but they often allow you to factor out complexity in order to make your code more readable and reusable. Function operators are typically paired with functionals. If you're using a for-loop, there's rarely a reason to use a FO, as it will make your code more complex for little gain.

If you're familiar with Python, decorators are just another name for function operators.

### Outline

- Section 11.2 introduces you to two useful existing FOs, and show you how use them to solve real problems.

- Section 11.3 works through a case study where work through a problem amenable to function operators: downloading many web pages.

## Prerequisites

Function operators are a type of function factory, so make sure you're familiar with Section 5.2 before you go on.

We'll use a couple of functionals from purrr that you learned about in Chapter 9, as well as some function operators that you'll learn about below. We'll use the memoise package for a useful FO.

```
library(purrr)
#>
#> Attaching package: 'purrr'
#> The following objects are masked from 'package:pryr':
#>
#> compose, partial
library(memoise)
```

## 11.2 Existing FOs

There are two extremely useful function operators that will both help you solve common recurring problems, and give you a sense for what FOs can do: `purrr::safely()` and `memoise::memoise()`.

### 11.2.1 Capturing errors with `purrr::safely()`

One advantage of a for-loops is that if one of the iterations fails in a for-loop you can still access all the previous results:

```
x <- list(
 c(0.512, 0.165, 0.717),
 c(0.064, 0.781, 0.427),
 c(0.890, 0.785, 0.495),
 "oops"
)

out <- rep(NA_real_, length(x))
for (i in seq_along(x)) {
 out[[i]] <- sum(x[[i]])
}
#> Error in sum(x[[i]]):
#> invalid 'type' (character) of argument
out
#> [1] 1.39 1.27 2.17 NA
```

If you run the same code with a functional, you get no output and it can be hard to figure out where the problem lies:

```
map_dbl(x, sum)
#> Error in sum(..., na.rm = na.rm):
#> invalid 'type' (character) of argument
```

`purrr::safely()` provides a tool to help with this problem. `safely()` is a function operator that transforms a function to turn errors into data. (You can learn the basic idea that makes it work in Section 7.6.2). Let's start by taking a look at it outside of `map dbl()`:

```
safe_sum <- safely(sum)
str(safe_sum(x[[1]]))
#> List of 2
#> $ result: num 1.39
#> $ error : NULL
str(safe_sum(x[[4]]))
#> List of 2
#> $ result: NULL
#> $ error :List of 2
#> ..$ message: chr "invalid 'type' (character) of argument"
#> ..$ call : language sum(..., na.rm = na.rm)
#> ...- attr(*, "class")= chr [1:3] "simpleError" "error" "condition"
```

A function transformed by `safely()` always returns a list with two elements, `result` and `error`. If the function runs successfully, `error` is `NULL` and `result` contains the result; if the function fails, `result` is `NULL` and `error` contains the error.

```
out <- map(x, safely(sum))
str(out)
#> List of 4
#> $:List of 2
#> ..$ result: num 1.39
#> ..$ error : NULL
#> $:List of 2
#> ..$ result: num 1.27
#> ..$ error : NULL
#> $:List of 2
#> ..$ result: num 2.17
#> ..$ error : NULL
#> $:List of 2
#> ..$ result: NULL
#> ..$ error :List of 2
#> ...$ message: chr "invalid 'type' (character) of argument"
#> ...$ call : language sum(..., na.rm = na.rm)
#> ...- attr(*, "class")= chr [1:3] "simpleError" "error" "condit"...
```

The output is in a slightly inconvenient form, since we have four lists each containing a list containing the result and the error. We can make it more convenient by using `purrr::transpose()` to turn it “inside-out” so that we get a list of result and a list of errors:

```
out <- transpose(map(x, safely(sum)))
str(out)
#> List of 2
#> $ result:List of 4
#> ..$: num 1.39
#> ..$: num 1.27
#> ..$: num 2.17
#> ..$: NULL
#> $ error :List of 4
#> ..$: NULL
#> ..$: NULL
#> ..$: NULL
```

```
#> ...$:List of 2
#>$.message: chr "invalid 'type' (character) of argument"
#>$.call : language sum(..., na.rm = na.rm)
#>-. attr(*, "class")= chr [1:3] "simpleError" "error" "condit"..

```

Now we can easily find the results the worked, or the inputs that failed:

```
ok <- map_lgl(out$error, is.null)
ok
#> [1] TRUE TRUE TRUE FALSE

x[!ok]
#> [[1]]
#> [1] "oops"

out$result[ok]
#> [[1]]
#> [1] 1.39
#>
#> [[2]]
#> [1] 1.27
#>
#> [[3]]
#> [1] 2.17
```

You can use this same technique in many different situations. For example, imagine you're fitting a set of generalised linear models (GLMs) to a list of data frames. While GLMs can sometimes fail because of optimisation problems, you'd still want to be able to try to fit all the models, and later look back at those that failed:

```
fit_model <- function(df) {
 glm(y ~ x1 + x2 * x3, data = df)
}

models <- transpose(map(datasets, safely(fit_model)))
ok <- map_lgl(models$error, is.null)

which data failed to converge?
datasets[!ok]

which models were successful?
models[ok]
```

I think this is a great example of the power of combining functionals and function operators: it lets you succinctly express what you need to solve a common data analysis problem.

purrr comes with three other function operators in a similar vein:

- `possibly()`: returns a default value when there's an error.
- `quietly()`: turns output, messages, and warning side-effects in to `output`, `message`, and `warning` components of the output.
- `auto_browser()`: automatically executes `browser()` inside the function when there's an error.

See their documentation for more details.

### 11.2.2 Caching computations with `memoise::memoise()`

An extremely handy FO is `memoise::memoise()`. It **memoises** a function, meaning that the function will remember previous inputs and return a cache results. Memoisation is an example of the classic computer science tradeoff of memory versus speed. A memoised function can run much faster because it stores all of the previous inputs and outputs, using more memory.

Let's explore this idea with a toy function that simulates an expensive operation:

```
slow_function <- function(x) {
 Sys.sleep(1)
 x * 10 * runif(1)
}
system.time(print(slow_function(1)))
#> [1] 0.808
#> user system elapsed
#> 0 0 1

system.time(print(slow_function(1)))
#> [1] 8.34
#> user system elapsed
#> 0.004 0.000 1.003
```

When we memoise this function, it's slow when we call it with new arguments. But when we call it with arguments that it's seen before it's instantaneous: it retrieves the previous value of the computation.

```
fast_function <- memoise::memoise(slow_function)
system.time(print(fast_function(1)))
#> [1] 6.01
#> user system elapsed
#> 0 0 1

system.time(print(fast_function(1)))
#> [1] 6.01
#> user system elapsed
#> 0.016 0.000 0.016
```

A relatively realistic use of memoisation is computing the Fibonacci series. The Fibonacci series is defined recursively: the first two values are defined by convention,  $f(0) = 0$ ,  $f(n) = 1$ , and then  $f(n) = f(n - 1) + f(n - 2)$  (for any positive integer). A naive version is slow because, for example, `fib(10)` computes `fib(9)` and `fib(8)`, and `fib(9)` computes `fib(8)` and `fib(7)`, and so on.

```
fib <- function(n) {
 if (n < 2) return(1)
 fib(n - 2) + fib(n - 1)
}
system.time(fib(23))
#> user system elapsed
#> 0.04 0.00 0.04
system.time(fib(24))
#> user system elapsed
#> 0.064 0.000 0.064
```

Memoising `fib()` makes the implementation much faster because each value is computed only once:

```
fib2 <- memoise::memoise(function(n) {
 if (n < 2) return(1)
```

```

 fib2(n - 2) + fib2(n - 1)
})
system.time(fib2(23))
#> user system elapsed
#> 0.028 0.000 0.026

```

And future calls can rely on previous computations:

```

system.time(fib2(24))
#> user system elapsed
#> 0.004 0.000 0.000

```

This is an example of **dynamic programming**, where a complex problem can be broken down into many overlapping subproblems, and remembering the results of a subproblem considerably improves performance.

Think carefully before memoising a function. If the function is not **pure**, i.e. the output does not depend only on the input, you will get misleading and confusing results. I created a subtle bug in devtools because I memoised the results of `available.package()`, which is rather slow because it has to download a large file from CRAN. The available packages don't change that frequently, but if you have an R process that's been running for a few days, the changes can become important, and because the problem only arose in long-running R process, the bug was very painful to find.

### 11.2.3 Exercises

1. Base R provides a function operator in the form of `Vectorize()`. What does it do? When might you use it?

## 11.3 Case study: creating your own FOs

Imagine you have named vector of URLs and you'd like to download each one to disk. That's pretty simple with `walk2()` and `file.download()`:

```

urls <- c(
 "adv-r" = "https://adv-r.hadley.nz",
 "r4ds" = "http://r4ds.had.co.nz/"
 # and many many more
)
path <- paste(tempdir(), names(urls), ".html")

walk2(urls, path, download.file, quiet = TRUE)

```

This approach is fine for a handful of URLs, but as the vector gets longer, it'd be nice to add a couple more features:

- Add a small delay between each request to avoid hammering the server.
- Display a . every few URLs so that we know that the function is still working.

It's relatively easy to add these extra features if we're using a for loop:

```

for(i in seq_along(urls)) {
 Sys.sleep(0.1)
 if (i %% 10 == 0) cat(".")
 download.file(urls[[i]], paths[[i]])
}

```

But I think this for loop is suboptimal because it interleaves different concerns (iteration, printing, and downloading). This makes the code harder to read, and it makes the harder to components reuse in new situations. Instead, let's see if we can use function operators to extract out the two ideas and make them reusable.

First, let's write make an FO that adds a small delay. I'm going to call it `delay_by()` for reasons that will be clear more shortly, and it has two arguments: the function to wrap, and the amount of delay to add. The actual implementation is quite simple. The main trick is forcing evaluation of all arguments as described in Section 10.2.4, because function operators are a special type of function factory:

```
delay_by <- function(f, amount) {
 force(f)
 force(amount)

 function(...) {
 Sys.sleep(amount)
 f(...)
 }
}
system.time(runif(100))
#> user system elapsed
#> 0 0 0
system.time(delay_by(runif, 0.1)(100))
#> user system elapsed
#> 0.0 0.0 0.1
```

And we can use it with the original `walk2()`:

```
walk2(urls, path, delay_by(download.file, 0.1), quiet = TRUE)
```

Creating a function to display the occasional dot is a little harder, because we can no longer rely on the index from the loop. We could pass the index along as another argument, but that breaks encapsulation: now a concern of the progress function becomes a problem that the higher level wrapper needs to deal instead. Instead, we'll use another function factory trick (from Section 10.2.3), so that the progress wrapper can manage its own internal counter:

```
dot_every <- function(f, n) {
 force(f)
 force(n)

 i <- 1
 function(...) {
 if (i %% n == 0) cat(".")
 i <- i + 1
 f(...)
 }
}
walk(1:100, runif)
walk(1:100, dot_every(runif, 10))
#>
```

Now we can express our original goal as:

```
walk2(urls, path, dot_every(delay_by(download.file, 0.1), 10), quiet = TRUE)
```

This is starting to get a little hard to read because we are composing many function calls, and the arguments are getting spread out. One way to resolve that is to use the pipe:

```
walk2(
 urls, path,
 download.file %>% dot_every(10) %>% delay_by(0.1),
 quiet = TRUE
)
```

The pipe works well here because I've carefully chosen the function names to yield an (almost) readable sentence: take `download.file` then (add) a dot every 10 iterations, then delay by 0.1s. The more clearly you can express the intent of your code through function names, the more easily others (including future you!) can read and understand the code.

### 11.3.1 Exercises

1. Compare and contrast the for loop and `walk2()` approaches to downloading many urls. Which makes it easier to see the core objects and functions? Which requires more background knowledge? What are the advantages and disadvantages in factoring out components of the problem into independent functions?

```
for (i in seq_along(urls)) {
 Sys.sleep(0.1)
 if (i %% 10 == 0) cat(".")
 download.file(urls[[i]], paths[[i]])
}

walk2(
 urls, path,
 download.file %>% dot_every(10) %>% delay_by(0.1),
 quiet = TRUE
)
```

2. Create a FO that reports whenever a file is created or deleted in the working directory, using `dir()` and `setdiff()`. What other global function effects might you want to track?
3. Write a FO that logs a time stamp and message to a file every time a function is run.
4. Modify `delay_by()` so that instead of delaying by a fixed amount of time, it ensures that a certain amount of time has elapsed since the function was last called. That is, if you called `g <- delay_by(1, f); g(); Sys.sleep(2); g()` there shouldn't be an extra delay.

## **Part III**

# **Object oriented programming**



# Introduction

In the following five chapters you'll learn about **object oriented programming** (OOP) in R. OOP in R is a little more challenging than in other languages, because:

- There are multiple OOP systems to choose from. In this book, I'll focus on the three that I believe are most important: S3, S4, and R6.
- S3 and S4 come from a very different heritage than the OOP found in most other popular languages. This means your existing OOP skills are unlikely to be of much help.

Indeed, in day-to-day use FP is much more important than OOP in R. Nevertheless, there are three main reasons to learn OOP:

- S3 allows your functions to return richer results that are displayed in a user friendly way and that have programmer friendly internals. It also conforms to syntactic standards that apply across multiple packages. This is why S3 is used throughout base R.
- S4 can be helpful for building up large systems that will evolve over time and will be written by many programmers. This is why the Bioconductor project uses S4 as its fundamental infrastructure.
- R6 gives you a standardized way to escape R's copy-on-modify semantics. This is particularly important if you want to model real-world objects that change over time.

This chapter will give you a rough lay of the land, and a field guide to help you identify OOP systems in the wild. The following four chapters (Base types, S3, S4, and R6) will dive into the details, starting with R's base types. These are not technically an OOP system, but they're important to understand because they're the fundamental building block of the true OOP systems.

## 11.4 OOP Systems

We'll begin with an info dump of vocabulary and terminology. Don't worry if it doesn't stick. We'll come back to these ideas multiple times in the subsequent chapters.

Central to any OOP system are the concepts of class and method. A **class** defines the behaviour of a set of **objects**, or instances, by describing their attributes and their relationship to other classes. The class is also used when selecting **methods**, functions that behave differently depending on the class of their input. A class defines what something *is* and methods describe what something can *do*.

Classes are usually organised in a hierarchy: if a method does not exist for a child, then the parent's method is used instead. This means that a child class will **inherit** behaviour from the parent class. Inheritance is one of the most important parts of OOP because it allows you to reduce the amount of code you have to write.

Following the notation of *Extending R*, there are two main styles of OOP:

- In **encapsulated** OOP, methods belong to objects or classes. This is the most common paradigm in modern programming languages, and method calls typically look like `object.method`. This is called encapsulated because the object encapsulates all its metadata.
- In **functional** OOP, methods belong to functions called **generics**. Method calls look like ordinary function calls: `generic(object)`. This is called functional because from the outside it just looks like function calls.

## 11.5 OOP in R

Base R provides three OOP systems: S3, S4, and reference classes (RC):

- **S3** is R’s first OOP system, and is described in *Statistical Models in S* (1991). It informally implements the functional style. It provides no ironclad guarantees but instead relies on a set of conventions. This makes it easy to get started with, and a low cost way of solving many simple problems.
- **S4** is similar to S3, but much more formal. It was introduced in *Programming with Data* (1998). It requires more upfront work and in return provides greater consistency. S4 is implemented in the **methods** package, which is attached by default. The only package in base R to make use of S4 is `stats4`.

(You might wonder if S1 and S2 exist. They don’t: S3 and S4 were named according to the versions of S that they accompanied.)

- **RC** implements encapsulated OO. RC objects are also mutable: they don’t use R’s usual copy-on-modify semantics, but are modified in place. This makes them harder to reason about, but allows them to solve problems that are difficult to solve with S3 or S4.

There are a number of other OOP systems provided by packages. Three of the most popular are:

- **R6** implements encapsulated OOP like RC, but resolves some important issues. You’ll learn R6 instead of RC in this book. More on why later.
- **R.oo** provides some formalism on top of S3, and makes it possible to have mutable S3 objects.
- **proto** implements another style of OOP, called prototype based. It blurs the distinctions between classes and instances of classes (objects). There is some more information about prototype based programming <http://vita.had.co.nz/papers/mutatr.html>.

Most OO systems in external packages are primarily of academic interest: they will help you understand the spectrum of OOP better, and can make it easier to solve certain classes of problems. However, they come with a big drawback: few R users know and understand them, so it is hard for others to read and contribute to your code.

## 11.6 Field guide

Before we go on to discuss base types, S3, S4, and R6 in more detail I want to introduce the sloop package:

```
install_github("hadley/sloop")
library(sloop)
```

The sloop package (think sail the seas of OOP in R) provides a number of helpers to fill in missing pieces in base R. The first helper to know about is `sloop::otype()`. It makes it easy to figure what OOP system an object found in the wild uses:

```
otype(1:10)
#> [1] "base"

otype(mtcars)
#> [1] "S3"

mle_obj <- stats4::mle(function(x = 1) (x - 2)^2)
otype(mle_obj)
#> [1] "S4"
```

Without `otype()`, you need to work your way through the base functions:

- `is.object()` distinguishes between base types (`FALSE`) and everything else (`TRUE`).
- `isS4()` distinguishes between S3 and S4.
- `inherits()` lets you figure out if you have an R6 object (an S3 object that inherits from “R6”) or an RC object (an S4 object that inherits from “refClass”).



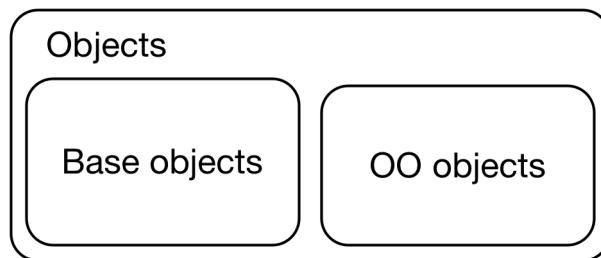
# Chapter 12

## Base types

### 12.1 Introduction

To talk about objects and OOP in R we need to first deal with a fundamental confusion: we use the word object to mean two different things. In this book so far, we've used object in a general sense, as captured by John Chambers' pithy quote: "Everything that exists in R is an object". However, while everything *is* an object, not everything is "object-oriented". This confusion arises because the base objects come from S, and were developed before anyone was thinking that S might need an OOP system. The tools and nomenclature evolved organically over many years without a single guiding principle.

Most of the time, the distinction between objects and object-oriented objects is not important. But here we need to get into the nitty gritty details so we'll use the terms **base objects** and **OO objects** to distinguish them.



We'll also discuss the `is.*` functions here. These functions are used for many purposes, but are commonly used to determine if an object has a specific base type.

### Outline

### 12.2 Base objects vs OO objects

To tell the difference between a base and an OO object, use `is.object()`:

```
A base object:
is.object(1:10)
#> [1] FALSE

An OO object
```

```
is.object(mtcars)
#> [1] TRUE
```

(This function would be better called `is.oo()` because it tells you if an object is a base object or a OO object.)

The primary attribute that distinguishes between base and OO object is the “class”. Base objects do not have a class attribute:

```
attr(1:10, "class")
#> NULL

attr(mtcars, "class")
#> [1] "data.frame"
```

Note that `attr(x, "class")` and `class(x)` do not always return the same thing, as `class()` returns a value, not `NULL`, for base objects. We’ll talk about exactly what it does return in the next chapter.

## 12.3 Base types

While only OO objects have a class attribute, every object has a **base type**:

```
typeof(1:10)
#> [1] "integer"

typeof(mtcars)
#> [1] "list"
```

Base types do not form an OOP system because functions that behave differently for different base types are primarily written in C, where dispatch occurs using switch statements. This means only R-core can create new types, and creating a new type is a lot of work. As a consequence, new base types are rarely added. The most recent change, in 2011, added two exotic types that you never see in R, but are needed for diagnosing memory problems (`NEWSXP` and `FREESXP`). Prior to that, the last type added was a special base type for S4 objects (`S4SXP`) added in 2005.

In total, there are 25 different base types. They are listed below, loosely grouped according to where they’re discussed in this book.

- The vectors: `NULL`, logical, integer, double, complex, character, list, raw.

```
typeof(1:10)
#> [1] "integer"

typeof(NULL)
#> [1] "NULL"

typeof(1i)
#> [1] "complex"
```

- Functions: closure (regular R functions), special (internal functions), builtin (primitive functions) and environment.

```
typeof(mean)
#> [1] "closure"

typeof(`[`)
#> [1] "special"

typeof(sum)
#> [1] "builtin"
```

```
typeof(globalenv())
#> [1] "environment"
```

- Language components: symbol (aka names), language (usually called calls), pairlist (used for function arguments).

```
typeof(quote(a))
#> [1] "symbol"
typeof(quote(a + 1))
#> [1] "language"
typeof(formals(mean))
#> [1] "pairlist"
```

“Expression” is a special purpose type that’s only returned by `parse()` and `expression()`. They are not needed in user code.

- There are a few esoteric types that are important for C code but not generally available at the R level: externalptr, weakref, bytecode, S4, promise, “...”, and any.

You may have heard of `mode()` and `storage.mode()`. I recommend ignoring these functions because they just provide S compatible aliases of `typeof()`. Read the source code if you want to understand exactly what they do.

## 12.4 The is functions

This is also a good place to discuss the `is` functions because they’re often used to check if an object has a specific type:

```
is.function(mean)
#> [1] TRUE
is.primitive(sum)
#> [1] TRUE
```

“Is” functions are often surprising because there are several different classes, they often have a few special cases, and their names are historical so don’t always reflect the usage in this book. They fall roughly into six classes:

- A specific value of `typeof()`: `is.call()`, `is.character()`, `is.complex()`, `is.double()`, `is.environment()`, `is.expression()`, `is.list()`, `is.logical()`, `is.name()`, `is.null()`, `is.pairlist()`, `is.raw()`, `is.symbol()`.

`is.integer()` is almost in this class, but it specifically checks for the absence of a class attribute containing “factor”. Also note that `is.vector()` belongs to the “attributes” class, and `is.numeric()` is described specially below.

- A set of possible base types:
  - `is.atomic()` = logical, integer, double, character, raw, and (surprisingly) NULL.
  - `is.function()` = special, builtin, closure.
  - `is.primitive()` = special, builtin.
  - `is.language()` = symbol, language, expression.
  - `is.recursive()` = list, language, expression.
- Attributes:

- `is.vector(x)` tests that `x` has no attributes apart from names. It does **not** check if an object is an atomic vector or list.
- `is.matrix(x)` tests if `length(dim(x))` is 2.
- `is.array(x)` tests if `length(dim(x))` is greater than zero.
- Has an S3 class: `is.data.frame()`, `is.factor()`, `is.numeric_version()`, `is.ordered()`, `is.package_version()`, `is.qr()`, `is.table()`.
- Vectorised mathematical operation: `is.finite()`, `is.infinite()`, `is.na()`, `is.nan()`.
- Finally there are a bunch of special purpose functions that don't fall into any other category:
  - `is.loaded()`: tests if a C/Fortran subroutine is loaded.
  - `is.object()`: discussed above.
  - `is.R()` and `is.single()`: are included for S+ compatibility
  - `is.unsorted()` tests if a vector is unsorted.
  - `is.element(x, y)` checks if `x` is an element of `y`: it's even more different as it takes two arguments, unlike every other `is.` function.

### 12.4.1 Numeric type

One function, `is.numeric()`, is sufficiently complicated and important that it needs a little extra discussion. The complexity comes about because R uses “numeric” to mean three slightly different things:

1. In some places it's used as an alias for “double”. For example `as.numeric()` is identical to `as.double()`, and `numeric()` is identical to `double()`.  
R also occasionally uses “real” instead of double; `NA_real_` is the one place that you're likely to encounter this in practice.
2. In S3 and S4 it is used to mean either integer or double. We'll talk about `s3_class()` in the next chapter:

```
sloop::s3_class(1)
#> [1] "double" "numeric"
sloop::s3_class(1L)
#> [1] "integer" "numeric"
```

3. In `is.numeric()` it means an object built on a base type of integer or double that is not a factor, i.e. an object that behaves like a number.

```
is.numeric(1)
#> [1] TRUE
is.numeric(1L)
#> [1] TRUE
is.numeric(factor("x"))
#> [1] FALSE
```

# Chapter 13

## S3

### 13.1 Introduction

S3 is R’s first and simplest OO system. S3 is informal and ad hoc, but it has a certain elegance in its minimalism: you can’t take away any part of it and still have a useful OO system. Because of these reasons, S3 should be your default choice for OO programming: you should use it unless you have a compelling reason otherwise. S3 is the only OO system used in the base and stats packages, and it’s the most commonly used system in CRAN packages.

S3 is a very flexible system: it allows you to do a lot of things that are quite ill-advised. If you’re coming from a strict environment like Java, this will seem pretty frightening (and it is!) but it does give R programmers a tremendous amount of freedom. While it’s very difficult to prevent someone from doing something you don’t want them to do, your users will never be held back because there is something you haven’t implemented yet. Since S3 has few built-in constraints, the key to its successful use is applying the constraints yourself. This chapter will teach you the conventions you should (almost) always adhere to in order to use S3 safely.

#### Outline

#### Prerequisites

We’ll use the sloop package to fill in some missing pieces when it comes to S3.

```
install_github("hadley/sloop")
library(sloop)
```

### 13.2 Basics

An S3 object is built on top of a base type with the “class” attribute set. The base type is typically a vector, although we will see later that it’s possible to use other types of classes. For example, take the factor. It is built on top of an integer vector, and the value of the class attribute is “factor”. It stores information about the “levels” in another attribute.

```
f <- factor("a")

typeof(f)
#> [1] "integer"
```

```
attributes(f)
#> $levels
#> [1] "a"
#>
#> $class
#> [1] "factor"
```

An S3 object behaves differently from its underlying base type because of **generic functions**, or generics for short. A generic executes different code depending on the class of one of its arguments, almost always the first. You can see this difference with the most important generic function: `print()`.

```
print(f)
#> [1] a
#> Levels: a
print(unclass(f))
#> [1] 1
#> attr(levels")
#> [1] "a"
```

`unclass()` strips the class attribute from its input, so it is a useful tool for seeing what special behaviour an S3 class adds.

`str()` shows the internal structure of S3 objects. Be careful when using `str()`: some S3 classes provide a custom `str()` method which can hide the underlying details. For example, take the `POSIXlt` class, which is one of the two classes used to represent date-time data:

```
time <- strftime("2017-01-01", "%Y-%m-%d")
str(time)
#> POSIXlt[1:1], format: "2017-01-01"
str(unclass(time), list.len = 5)
#> List of 9
#> $ sec : num 0
#> $ min : int 0
#> $ hour : int 0
#> $ mday : int 1
#> $ mon : int 0
#> [list output truncated]
#> - attr(*, "tzone")= chr "UTC"
```

A **generic** and its **methods** are functions that operate on classes. The role of a generic is to find the right method for the arguments that it is provided, the process of **method dispatch**. A method is a function that implements the generic behaviour for a specific class. In other words the job of the generic is to find the right method; the job of the method is to do the work.

S3 methods are functions with a special naming scheme, `generic.class()`. For example, the Date method for the `mean()` generic is called `mean.Date()`, and the factor method for `print()` is called `print.factor()`. This is the reason that most modern style guides discourage the use of `.` in function names: it makes them look like S3 methods. For example, is `t.test()` the t method for `test` objects?

You can find some S3 methods (those in the base package and those that you've created) by typing their names. However, this will not work with most packages because S3 methods are not exported: they live only inside the package, and are not available from the global environment. Instead, you can use `getS3method()`, which will work regardless of where the method lives:

```
Only works because the method is in the base package
mean.Date
#> function (x, ...)
```

```
#> .Date(mean(unclass(x), ...))
#> <bytecode: 0x4e41410>
#> <environment: namespace:base>

Always works
getS3method("mean", "Date")
#> function (x, ...)
#> .Date(mean(unclass(x), ...))
#> <bytecode: 0x4e41410>
#> <environment: namespace:base>
```

### 13.2.1 Exercises

1. The most important S3 objects in base R are factors, data frames, and date/times (Dates, POSIXct, POSIXlt). You've already seen the attributes and base type that factors are built on. What base types and attributes are the others built on?
2. Describe the difference in behaviour in these two calls.

```
set.seed(1014)
some_days <- as.Date("2017-01-31") + sample(10, 5)

mean(some_days)
#> [1] "2017-02-05"
mean(unclass(some_days))
#> [1] 17202
```

3. Draw a Venn diagram illustrating the relationships between functions, generics, and methods.
4. What does the `as.data.frame.data.frame()` method do? Why is it confusing? How should you avoid this confusion in your own code?
5. What does the following code return? What base type is it built on? What attributes does it use?

```
x <- ecdf(rpois(100, 10))
x
#> Empirical CDF
#> Call: ecdf(rpois(100, 10))
#> x[1:18] = 2, 3, 4, ..., 2e+01, 2e+01
```

## 13.3 Classes

S3 is a simple and ad hoc system, and has no formal definition of a class. To make an object an instance of a class, you simply take an existing object and set the **class attribute**. You can do that during creation with `structure()`, or after the fact with `class<-()`:

```
Create and assign class in one step
foo <- structure(list(), class = "foo")

Create, then set class
foo <- list()
class(foo) <- "foo"
```

You can determine the class of any object using `class(x)`, and see if an object inherits from a specific class using `inherits(x, "classname")`.

```
class(foo)
#> [1] "foo"
inherits(foo, "foo")
#> [1] TRUE
```

The class name can be any character vector, but I recommend using only letters and `_`. Avoid `..`. Opinion is mixed whether to use underscores (`my_class`) or CamelCase (`MyClass`) for multi-word class names. Pick one convention and stick with it.

It's possible to provide a vector of class names, which allows S3 to implement a basic style of inheritance. This allows you to reduce your workload by allowing classes to share code where possible. We'll come back to this idea in inheritance.

S3 has no checks for correctness. This means you can change the class of existing objects:

```
Create a linear model
mod <- lm(log(mpg) ~ log(disp), data = mtcars)
class(mod)
#> [1] "lm"
print(mod)
#>
#> Call:
#> lm(formula = log(mpg) ~ log(disp), data = mtcars)
#>
#> Coefficients:
#> (Intercept) log(disp)
#> 5.381 -0.459

Turn it into a data frame (?!)
class(mod) <- "data.frame"

Unsurprisingly this doesn't work very well
print(mod)
#> [1] coefficients residuals effects rank
#> [5] fitted.values assign qr df.residual
#> [9] xlevels call terms model
#> <0 rows> (or 0-length row.names)
```

If you've used other OO languages, this might make you feel queasy. But surprisingly, this flexibility causes few problems: while you *can* change the type of an object, you never *should*. R doesn't protect you from yourself: you can easily shoot yourself in the foot. As long as you don't aim the gun at your foot and pull the trigger, you won't have a problem.

To avoid foot-bullet intersections when creating your own class, you should always provide:

- A **constructor**, `new_x()`, that efficiently creates new objects with the correct structure.

For more complicated classes, you may also want to provide:

- A **validator**, `validate_x()`, that performs more expensive checks that the object has correct values.
- A **helper**, `x()`, that provides a convenient and neatly parameterised way for others to construct and validate (create) objects of this class.

### 13.3.1 Constructors

S3 doesn't provide a formal definition of a class, so it has no built-in way to ensure that all objects of a given class have the same structure (i.e. same attributes with the same types). Instead, you should enforce a consistent structure yourself by using a **constructor**. A constructor is a function whose job is to create objects of a given class, ensuring that they always have the same structure.

There are three rules that a constructor should follow. It should:

1. Be called `new_class_name()`.
2. Have one argument for the base object, and one for each attribute. (More if the class can be subclassed, see inheritance.)
3. Check the types of the base object and each attribute.

Base R generally does not provide constructors (three exceptions are the internal `.difftime()`, `.POSIXct()`, and `.POSIXlt()`) so we'll demonstrate constructors by filling in some missing pieces in base. (If you want to use these constructors in your own code, you can use the versions exported by the sloop package, which complete a few details that we skip here in order to focus on the core issues.)

We'll start with one of the simplest S3 classes in base R: Date, which is just a double with a class attribute. The constructor rules lead to the slightly awkward name `new_Date()`, because the existing base class uses a capital letter. I recommend using lower case class names to avoid this problem.

```
new_Date <- function(x) {
 stopifnot(is.double(x))
 structure(x, class = "Date")
}

new_Date(c(-1, 0, 1))
#> [1] "1969-12-31" "1970-01-01" "1970-01-02"
```

You can use the `new_s3_*`() helpers provided by the sloop to make this even simpler. They are wrappers around `structure` that require a class argument, and check the base type of `x`.

```
new_Date <- function(x) {
 sloop::new_s3_dbl(x, class = "Date")
}
```

The purpose of the constructor is to help the developer (you). That means you can keep them simple, and you don't need to optimise the error messages for user friendliness. If you expect others to create your objects, you should also create a friendly helper function, called `class_name()`, that we'll describe shortly.

A slightly more complicated example is `POSIXct`, which is used to represent date-times. It is again built on a double, but has an attribute that specifies the time zone, a length 1 character vector. R defaults to using the local time zone, which is represented by the empty string. To create the constructor, we need to make sure each attribute of the class gets an argument to the constructor. This gives us:

```
new_POSIXct <- function(x, tzone = "") {
 stopifnot(is.double(x))
 stopifnot(is.character(tzone), length(tzone) == 1)

 structure(x,
 class = c("POSIXct", "POSIXt"),
 tzone = tzone
)
}

new_POSIXct(1)
```

```
#> [1] "1970-01-01 00:00:01 UTC"
new POSIXct(1, tzone = "UTC")
#> [1] "1970-01-01 00:00:01 UTC"
```

The constructor checks that `x` is a double, and that `tzone` is a length 1 character vector. We use `stopifnot()` here since the constructor is a developer focussed function so error messages don't need to be that friendly. Note that `POSIXct` uses a class `vector`; we'll come back to what that means in inheritance.

Generally, the constructor should not check that the values are valid because such checks are often expensive. For example, our `new POSIXct()` constructor does not check that `tzone` is a valid value, and we get a warning when the object is printed.

```
x <- new POSIXct(1, "Auckland NZ")
x
#> [1] "1970-01-01 00:00:01 Auckland"
```

### 13.3.2 Validators

More complicated classes will require more complicated checks for validity. Take factors, for example. The constructor function only checks that the structure is correct:

```
new_factor <- function(x, levels) {
 stopifnot(is.integer(x))
 stopifnot(is.character(levels))

 structure(
 x,
 levels = levels,
 class = "factor"
)
}
```

So it's possible to use this to create invalid factors:

```
new_factor(1:5, "a")
#> Error in as.character.factor(x):
#> malformed factor
new_factor(0:1, "a")
#> Error in as.character.factor(x):
#> malformed factor
```

Rather than encumbering the constructor with complicated checks, it's better to put them in a separate function. This is a good idea because it allows you to cheaply create new objects when you know that the values are correct, and to re-use the checks in other places.

```
validate_factor <- function(x) {
 values <- unclass(x)
 levels <- attr(x, "levels")

 if (!all(!is.na(values) & values > 0)) {
 stop(
 "All `x` values must be non-missing and greater than zero",
 call. = FALSE
)
 }
}
```

```

if (length(levels) < max(values)) {
 stop(
 "There must at least as many `levels` as possible values in `x`",
 call. = FALSE
)
}

x
}

validate_factor(new_factor(1:5, "a"))
#> Error: There must at least as many `levels` as possible values in `x`
validate_factor(new_factor(0:1, "a"))
#> Error: All `x` values must be non-missing and greater than zero

```

This function is called primarily for its side-effects (throwing an error if the object is invalid) so you'd expect it to invisibly return its primary input. However, unlike most functions called for their side effects, its useful for validation methods to return visibly, as we'll see next.

### 13.3.3 Helpers

If you want others to construct objects from your class, you should also provide a helper method that makes their life as easy as possible. This should have the same name as the class, and should be parameterised in a convenient way. `factor()` is a good example of this as well: you want to automatically derive the internal representation from a vector. The simplest possible implementation looks something like this:

```

factor <- function(x, levels = unique(x)) {
 ind <- match(x, levels)
 validate_factor(new_factor(ind, levels))
}
factor(c("a", "a", "b"))
#> [1] a a b
#> Levels: a b

```

The validator prevents the construction of invalid objects, but for a real helper you'd spend more time creating user friendly error messages.

```

factor(c("a", "a", "b"), levels = "a")
#> Error: All `x` values must be non-missing and greater than zero

```

In base R, neither `Date` nor `POSIXct` has a helper function. Instead there are two ways to construct them:

- By coercing from another type with `as.Date()` and `as.POSIXct()`. These functions should be S3 generics, so we'll come back to them in coercion.
- With a helper function that either parses a string (`strptime()`) or creates a date from individual components (`ISODate(datetime())`).

These missing helpers mean that there's no obvious default way to create a date or date-time in R. We can fill in those missing pieces with a couple of helpers:

```

Date <- function(year, month, day) {
 as.Date(ISOdate(year, month, day, tz = ""))
}

POSIXct <- function(year, month, day, hour, minute, sec, tzzone = "") {

```

```
ISOdatetime(year, month, day, hour, minute, sec, tz = tzone)
}
```

These helpers fill a useful role, but are not computationally efficient: behind the scenes `ISOdatetime()` works by pasting the components into a string and then using `strptime()`. More efficient equivalents are `lubridate::make_datetime()` and `lubridate::make_date()`.

### 13.3.4 Object styles

S3 gives you the freedom to build a new class on top of any existing base type. So far, we've focussed on vector-style where you take an existing vector type and add some attributes. Importantly, a single vector-style object represents multiple values. There are two other important styles: scalar-style and data-frame-style.

Each **scalar**-style object represents a single “value”, and are built on top of named lists. This is the style that you are most likely to use in practice. The constructor for the scalar type is slightly different because the arguments become named elements of the list, rather than attributes.

```
new_scalar_class <- function(x, y, z) {
 structure(
 list(
 x = x,
 y = y,
 z = z
),
 class = "scalar_class"
)
}
```

(For a real constructor, you'd also check that the `x`, `y`, and `z` fields are the types that you expect.)

In base R, the most important example of this style is `lm`, the class returned when you fit a linear model:

```
mod <- lm(mpg ~ wt, data = mtcars)
typeof(mod)
#> [1] "list"
names(mod)
#> [1] "coefficients" "residuals" "effects" "rank"
#> [5] "fitted.values" "assign" "qr" "df.residual"
#> [9] "xlevels" "call" "terms" "model"
```

The **data-frame-style** builds on top of a data frame (a named list where each element is a vector of the same length), and adds additional attributes to store important metadata. A data-frame-style constructor looks like:

```
new_df_class <- function(df, attr1, attr2) {
 stopifnot(is.data.frame(df))

 structure(
 df,
 attr1 = attr1,
 attr2 = attr2,
 class = c("df_class", "data.frame")
)
}
```

The most common data-frame-style class is the tibble, a modern reimagining of the data frame provided by the tibble package, and used extensively within the tidyverse.

Collectively, we'll call the attributes of a vector-style or data-frame-style class and the names of a list-style class the **fields** of an object.

When creating your own classes, you should pick the vector style if your class closely resembles an existing vector type. Otherwise, use a scalar (list) style. The scalar type is generally easier to work with because implementing a full range of convenient vectorised methods is usually a lot of work. It's typically obvious when you need to use a data-frame-style.

### 13.3.5 Exercises

1. Categorise the objects returned by `lm()`, `factor()`, `table()`, `as.Date()`, `ecdf()`, `ordered()`, `I()` into “vector”, “scalar”, and “other”.
2. Write a constructor for `difftime` objects. What base type are they built on? What attributes do they use? You'll need to consult the documentation, read some code, and perform some experiments.
3. Write a constructor for `data.frame` objects. What base type is a data frame built on? What attributes does it use? What are the restrictions placed on the individual elements? What about the names?
4. Enhance our `factor()` helper to have better behaviour when one or more `values` is not found in `levels`. What does `base::factor()` do in this situation?
5. Carefully read the source code of `factor()`. What does it do that our constructor does not?
6. What would a constructor function for `lm` objects, `new_lm()`, look like? Why is a constructor function less useful for linear models?

## 13.4 Generics and methods

The job of an S3 generic is to perform method dispatch, i.e. find the function designed to work specifically for the given class. S3 generics have a simple structure: they call `UseMethod()`, which then calls the right method. `UseMethod()` takes two arguments: the name of the generic function (required), and the argument to use for method dispatch (optional). If you omit the second argument it will dispatch based on the first argument, which is what I generally advise.

```
Dispatches on x
generic <- function(x, y, ...) {
 UseMethod("generic")
}

Dispatches on y
generic2 <- function(x, y, ...) {
 UseMethod("generic2", y)
}
```

Note that you don't pass any of the arguments of the generic to `UseMethod()`; it uses black magic to pass them on automatically. Generally, you should avoid doing any computation in a generic, because the semantics are complicated and few people know the details. In general, any modifications to the arguments of the generic will be undone, leading to much confusion.

A generic isn't useful without some methods, which are just functions that follow a naming scheme (`generic.class`). Because a method is just a function with a special name, you *can* call methods directly,

but you generally *shouldn't*. The main reason to call the method directly is that it sometimes leads to considerable performance improvements. See `performance` for an example.

```
generic.foo <- function(x, y, ...) {
 message("foo method")
}

generic(new_s3_scalar(class = "foo"))
#> foo method
```

You can see all the methods defined for a generic with `s3_methods_generic()`:

```
s3_methods_generic("generic")
#> # A tibble: 2 x 4
#> generic class visible source
#> <chr> <chr> <lgl> <chr>
#> 1 generic foo TRUE .GlobalEnv
#> 2 generic skeleton TRUE methods
```

Note the false positive: `generic.skeleton()` is not a method for our generic but an existing function in the `methods` package. It's picked up because method definition relies only on a naming convention. This is another reason that you should avoid using `.` in non-method function names.

Remember that apart from methods that you've created, and those defined in the base package, most S3 methods will not be directly accessible. You'll need to use `getS3method("generic", "class")` to see their source code.

### 13.4.1 Coercion

Many S3 objects can be naturally created from an existing object through **coercion**. If this is the case for your class, you should provide a coercion function, an S3 generic called `as_class_name`. Base R generally does not follow this convention, which can cause problems as illustrated by `as.factor()`:

- The name is confusing, since `as.factor()` is not the `factor` method of the `as()` generic.
- `as.factor()` is not a generic, which means that if you create a new class that could be usefully converted to a factor, you can not extend `as.factor()`.

We can fix these issues by creating a new generic coercion function and providing it with some methods:

```
as_factor <- function(x, ...) {
 UseMethod("as_factor")
}
```

Every `as_y()` generic should have a `y` method that returns its input unchanged:

```
as_factor.factor <- function(x, ...) x
```

This ensures that `as_factor()` works if the input is already a factor.

Two useful methods would be for character and integer vectors.

```
as_factor.character <- function(x, ...) {
 factor(x, levels = unique(x))
}
as_factor.integer <- function(x, ...) {
 factor(x, levels = as.character(unique(x)))
}
```

Typically the coercion methods will either call the constructor or the helper; pick the function that makes the code simpler. Here the helper is simplest. If you use the constructor, remember to also call the validator function.

If you think your coercion function will be frequently used, it's worth providing a default method that gives a better error message. Default methods are called when no other method is appropriate, and are discussed in more detail in inheritance.

```
as_factor(1)
#> Error in UseMethod("as_factor"):
#> no applicable method for 'as_factor' applied to an object of class "c('double', 'numeric')"

as_factor.default <- function(x, ...) {
 stop(
 "Don't know how to coerce object of class ",
 paste(class(x), collapse = "/"), " into a factor",
 call. = FALSE
)
}
as_factor(1)
#> Error: Don't know how to coerce object of class numeric into a factor
```

### 13.4.2 Arguments

Methods should always have the same arguments as their generics. This is not usually enforced, but it is good practice because it will avoid confusing behaviour. If you do eventually turn your code into a package, R CMD check will enforce it, so it's good to get into the habit now.

There is one exception to this rule: if the generic has ..., the method must still have all the same arguments (including ...), but can also have its own additional arguments. This allows methods to take additional arguments, which is important because you don't know what additional arguments that a method for someone else's class might need. The downside of using ..., however, is that any misspelled arguments will be silently swallowed.

### 13.4.3 Exercises

1. Read the source code for t() and t.test() and confirm that t.test() is an S3 generic and not an S3 method. What happens if you create an object with class test and call t() with it? Why?

```
x <- structure(1:10, class = "test")
t(x)
#>
#> One Sample t-test
#>
#> data: x
#> t = 6, df = 9, p-value = 3e-04
#> alternative hypothesis: true mean is not equal to 0
#> 95 percent confidence interval:
#> 3.33 7.67
#> sample estimates:
#> mean of x
#> 5.5
```

2. Carefully read the documentation for `UseMethod()` and explain why the following code returns the results that it does. What two usual rules of function evaluation does `UseMethod()` violate?

```
g <- function(x) {
 x <- 10
 y <- 10
 UseMethod("g")
}
g.default <- function(x) c(x = x, y = y)

x <- 1
y <- 1
g(x)
#> x y
#> 1 10
```

## 13.5 Method dispatch

At a high-level, S3 method dispatch is simple, and revolves around two functions, `UseMethod()` and `NextMethod()`. You'll learn about these two functions below, and then we'll come back to some of the additional wrinkles in dispatch details.

### 13.5.1 `UseMethod()`

The purpose of `UseMethod()` is to find the appropriate method to call given a generic and a class. It does this by creating a vector of function names, `paste0("generic", ".", c(class(x), "default"))`, and looking for each method in turn. As soon as it finds a matching method, it calls it. If no matching method is found, it throws an error. To explore dispatch, we'll use `sloop::s3_dispatch()`. You give it a call to an S3 generic, and it lists all the possible methods, noting which ones exist. For example, what happens when you try to print a `POSIXct` object?

```
x <- Sys.time()
s3_dispatch(print(x))
#> -> print.POSIXct
#> print.POSIXt
#> * print.default
```

`print()` will look for three possible methods, of which two exist, and one, `print.POSIXct()`, will be called. The last method is always the “default” method. This doesn't correspond to a specific class, so is a useful catch all.

### 13.5.2 `NextMethod()`

Method dispatch usually terminates as soon as a matching method is found. However, methods can explicitly choose to call the next available method using `NextMethod()`. This is useful because it allows you to rely on code that others have already written, which we'll come back to in inheritance. Let's make `NextMethod()` concrete with an example. Here, I define a new generic (“showoff”) with three methods. Each method signals that it's been called, and then calls the “next” method:

```
showoff <- function(x) {
 UseMethod("showoff")
}
```

```
showoff.default <- function(x) {
 message("showoff.default")
 TRUE
}
showoff.a <- function(x) {
 message("showoff.a")
 NextMethod()
}
showoff.b <- function(x) {
 message("showoff.b")
 NextMethod()
}
```

Let's create a dummy object with classes "b" and "a". `s3_dispatch()` shows that all three potential methods are available:

```
x <- new_s3_scalar(class = c("b", "a"))
s3_dispatch(showoff(x))
#> -> showoff.b
#> * showoff.a
#> * showoff.default
```

When you call `NextMethod()` it finds and calls the next available method in the dispatch list. When we call `showoff()`, the method for `b` forwards to the method for `a`, which forwards to the default method.

```
showoff(x)
#> showoff.b
#> showoff.a
#> showoff.default
#> [1] TRUE
```

Like `UseMethod()`, the precise semantics of `NextMethod()` are complex. It doesn't actually work with the class attribute of the object, but instead uses a special global variable (`.Class`) to keep track of which method to call next. This means that modifying the argument that is dispatched upon has no impact, and you should avoid modifying the object that is being dispatched on.

Generally, you call `NextMethod()` without any arguments. However, if you do give arguments, they are passed on to the next method, as if they'd been supplied to the generic.

### 13.5.3 Exercises

1. Which base generic has the greatest number of defined methods?
2. Explain what is happening in the following code.

```
generic2 <- function(x) UseMethod("generic2")
generic2.a1 <- function(x) "a1"
generic2.a2 <- function(x) "a2"
generic2.b <- function(x) {
 class(x) <- "a1"
 NextMethod()
}

generic2(new_s3_scalar(class = c("b", "a2")))
#> [1] "a2"
```

## 13.6 Inheritance

The class attribute is not limited to a single string, but can be a character vector. This, along with S3 method dispatch and `NextMethod()`, gives a surprising amount of flexibility that can be used creatively to reduce code duplication. However, this flexibility can also lead to code that is hard to understand or reason about, so you are best constraining yourself to simple styles of inheritance. Here we will focus on defining subclasses that inherit their fields, and some behaviour, from a parent class.

Subclasses use a character `vector` for the class attribute. There are two examples of subclasses that you might have come across in base R:

- Generalised linear models are a generalisation of linear models that allow the error term to belong to a richer set of distributions, not just the normal distribution like the linear model. This is a natural case for the use of inheritance and indeed, in R, `glm()` returns objects of class `c("glm", "lm")`.
- Ordered factors are used when the levels of a factor have some intrinsic ordering, like `c("Good", "Better", "Best")`. Ordered factors are produced by `ordered()` which returns an object with class `c("ordered", "factor")`.

You can think of the `glm` class “inheriting” behaviour from the `lm` class, and the ordered class inheriting behaviour from the factor class because of the way method dispatch works. If there is a method available for the subclass, R will use it, otherwise it will fall back to the “parent” class. For example, if you “plot” a `glm` object, it falls back to the `lm` method, but if you compute the ANOVA, it uses a `glm`-specific method.

```
mod1 <- glm(mpg ~ wt, data = mtcars)

s3_dispatch(plot(mod1))
#> plot.glm
#> -> plot.lm
#> * plot.default
s3_dispatch(anova(mod1))
#> -> anova.glm
#> * anova.lm
#> anova.default
```

### 13.6.1 Constructors

There are three principles to adhere to when creating a subclass:

- A subclass should be built on the same base type as a parent.
- The `class()` of the subclass should be of the form `c(subclass, parent_class)`
- The fields of the subclass should include the fields of the parent.

And these properties should be enforced by the constructor.

When you create a class, you need to decide if you want to allow subclasses, because it requires changes to the constructor and careful thought in your methods. To allow subclasses, the parent constructor needs to have `...` and `subclass` arguments:

```
new_my_class <- function(x, y, ..., subclass = NULL) {
 stopifnot(is.numeric(x))
 stopifnot(is.logical(y))

 structure(
 x,
 y = y,
```

```

 ...
 class = c(subclass, "my_class")
}
}

```

Then the implementation of the subclass constructor is simple: it checks the types of the new fields, then calls the parent constructor.

```

new_subclass <- function(x, y, z) {
 stopifnot(is.character(z))
 new_my_class(x, y, z = z, subclass = "subclass")
}

```

If you wanted to allow this subclass to be further subclassed, you'd need to include ... and subclass arguments:

```

new_subclass <- function(x, y, z, ..., subclass = NULL) {
 stopifnot(is.character(z))

 new_my_class(x, y, z = z, ..., subclass = c(subclass, "subclass"))
}

```

If your subclass is more complicated, you'd also provide validator and helper functions, as described previously.

### 13.6.2 Coercion

You also need to make sure that there's some way to convert the subclass back to the parent class. The best way to do that is to add a method to the coercion generic. Generally, this method should call the parent constructor:

```

as_my_class.sub_class <- function(x) {
 new_my_class(attr(x, "x"), attr(x, "y"))
}

```

### 13.6.3 Methods

The goal of creating a subclass is to reuse as much code as possible from the parent class. This means that you should not have to define every method that the parent class provides (if you do, reconsider if you actually need a subclass!). Generally, defining new methods is straightforward: you simply create a new method (`generic.subclass`) whenever the parent method doesn't do quite the right thing. In many cases, the new method will be able to call `NextMethod()` in order to take advantage of the computation done in the parent.

One wrinkle arises when you have methods that return the same type of object as the primary input. For example, dplyr has many functions (`arrange()`, `summarise()`, `mutate()`, ...) that input a data frame (or data frame-like object) and output a modified version of that data frame. Imagine you want to store the provenance of each data frame, i.e. who created it and when. To do so, you might create a data frame subclass called `provenance`:

```

new_provenance <- function(data, author, date = Sys.Date()) {
 stopifnot(is.data.frame(data))
 stopifnot(is.character(author), length(author) == 1)
 stopifnot(is.Date(date), length(date) == 1)
}

```

```

structure(
 data,
 author = author,
 date = date,
 class = c("provenance", "data.frame")
)
}

```

And now you want to make this class work with dplyr. The class doesn't change any of the computation related to the data frame, it just needs to preserve the attributes, which dplyr doesn't know anything about. That means you need to provide a method for each dplyr generic. The computation is unchanged, so you can use `NextMethod()` to do all the hard work, but you need to manually reconstruct the provenance object.

```

arrange.provenance <- function(.data, ...) {
 new_provenance(
 NextMethod(),
 author = attr(.data, "author"),
 date = attr(.data, "date")
)
}

mutate.provenance <- function(.data, ...) {
 new_provenance(
 NextMethod(),
 author = attr(.data, "author"),
 date = attr(.data, "date")
)
}

```

To do this for all the dplyr generics would require a lot of copying and pasting. Let's reduce some of that duplication by taking advantage of `sloop::reconstruct()`. `reconstruct()` is a generic function designed to reconstruct a subclass from an instance of the parent class, typically created by `NextMethod()`, and the original subclass. In other words, the job of a reconstructor is to take an object from a parent class, and copy over attributes from the subclass. (Note that `reconstruct()` is unusual in that it dispatches on the second argument. This allows a more natural specification.)

```

reconstruct.provenance <- function(new, old) {
 new_provenance(
 new,
 author = attr(old, "author"),
 date = attr(old, "date")
)
}

```

Now we can rewrite the methods to minimise the amount of duplicated code:

```

arrange.provenance <- function(.data, ...) {
 reconstruct(NextMethod(), .data)
}

mutate.provenance <- function(.data, ...) {
 reconstruct(NextMethod(), .data)
}

```

This duplicated code could be avoided completely if `arrange.data.frame()`, provided by dplyr, called `reconstruct()` for you. And indeed, a future version of that function will.

When designing a class that can be subclassed, you need to carefully think through these issues. Generally, whenever you implement a method that returns the same type of object as the primary input, you should call `reconstruct()` to ensure that it also works for subclasses. That way implementors of a subclass will only need to provide methods when the computation is actually different.

### 13.6.4 Exercises

1. The `ordered` class is a subclass of `factor`, but it's implemented in a very ad hoc way in base R. Implement it in a principled way by building a constructor and an `as_ordered` generic.

```
f1 <- factor("a", c("a", "b"))
as.factor(f1)
#> [1] a
#> Levels: a b
as.ordered(f1) # loses levels
#> [1] a
#> Levels: a
```

2. What classes have a method for the `Math` group generic in base R? Read the source code. How do the methods work?
3. R has two classes for representing date time data, `POSIXct` and `POSIXlt`, which both inherit from `POSIXt`. Which generics have different behaviours for the two classes? Which generics share the same behaviour?

## 13.7 Dispatch details

This chapter concludes with a few additional details about method dispatch that is not well documented elsewhere. It is safe to skip these details if you're new to S3.

### 13.7.1 Environments and namespaces

The precise rules for where a generic looks for the methods are a little complicated because there are two paths for discovery:

1. In the calling environment of the function that called the generic.
2. In the special `__S3MethodsTable__` object in the function environment of the generic. Every package has an `__S3MethodsTable__` which lists all the S3 methods exported by the package.

These details are not usually important, but are necessary in order for S3 generics to find the correct method when the generic and method are in different packages.

### 13.7.2 S3 and base types

What happens when you call an S3 generic with a non-S3 object, i.e. an object that doesn't have the `class` attribute set? You might think it would dispatch on what `class()` returns:

```
class(matrix(1:5))
#> [1] "matrix"
```

But unfortunately dispatch actually occurs on the **implicit class**, which has three components:

- “array” or “matrix” (if the object has dimensions).

- `typeof()` (with a few minor tweaks).
- If it's "integer" or "double", "numeric".

There is no base function that will compute the implicit class, but you can use a helper from the sloop package:

```
s3_class(matrix(1:5))
#> [1] "matrix" "integer" "numeric"
```

`s3_dispatch()` knows about the implicit class, so use it if you're ever in doubt about method dispatch:

```
s3_dispatch(print(matrix(1:5)))
#> print.matrix
#> print.integer
#> print.numeric
#> -> print.default
```

Note that this can lead to different dispatch for objects that look similar:

```
x1 <- 1:5
class(x1)
#> [1] "integer"
s3_dispatch(mean(x1))
#> mean.integer
#> mean.numeric
#> -> mean.default

x2 <- structure(x1, class = "integer")
class(x2)
#> [1] "integer"
s3_dispatch(mean(x2))
#> mean.integer
#> -> mean.default
```

### 13.7.3 Internal generics

Some S3 generics, like `[`, `sum()`, and `cbind()`, don't call `UseMethod()` because they are implemented in C. Instead, they call the C functions `DispatchGroup()` or `DispatchOrEval()`. These functions are called **internal generics**, because they do dispatch internally, in C code. Internal generics only exist in base R, so you can not create an internal generic in a package.

`s3_dispatch()` shows internal generics by including the name of the generic at the bottom of the method class. If this method is called, all the work happens in C code, typically using `[switchpatch]`.

```
s3_dispatch(Sys.time() [1])
#> -> [.POSIXct
#> [.POSIXt
#> [.default
#> * [
```

For performance reasons, internal generics do not dispatch to methods unless the class attribute has been set (`is.object()` is true). This means that internal generics do not use the implicit class. Again, if you're confused, rely on `s3_dispatch()` to show you the difference.

```
x <- sample(10)
class(x)
#> [1] "integer"
```

```
s3_dispatch(x[1])
#> [.integer
#> [.numeric
#> [.default
#> -> [

class(y)
#> [1] "numeric"
s3_dispatch(mtcars[1])
#> -> [.data.frame
#> [.default
#> * [
```

### 13.7.4 Group generics

Group generics are the most complicated part of S3 method dispatch because they involve both `NextMethod()` and internal generics. Group generics are worth learning about, however, because they allow you to implement a whole swath of methods with one function. Like internal generics, they only exist in base R, and you can not define your own group generic.

Base R has four group generics, which are made up of the following generics:

- **Math**: `abs`, `sign`, `sqrt`, `floor`, `cos`, `sin`, `log`, `exp`, ...
- **Ops**: `+`, `-`, `*`, `/`, `^`, `%%`, `%/%`, `&`, `|`, `!`, `==`, `!=`, `<`, `<=`, `>`, `>=`
- **Summary**: `all`, `any`, `sum`, `prod`, `min`, `max`, `range`
- **Complex**: `Arg`, `Conj`, `Im`, `Mod`, `Re`

Defining a single group generic for your class overrides the default behaviour for all of the members of the group. Methods for group generics are looked for only if the methods for the specific generic do not exist:

```
s3_dispatch(sum(Sys.time()))
#> sum.POSIXct
#> sum.POSIXt
#> sum.default
#> -> Summary.POSIXct
#> Summary.POSIXt
#> Summary.default
#> * sum
```

Most group generics involve a call to `NextMethod()`. For example, take `difftime()` objects. If you look at the method dispatch for `abs()`, you'll see there's a `Math` group generic defined.

```
y <- as.difftime(10, units = "mins")
s3_dispatch(abs(y))
#> abs.difftime
#> abs.default
#> -> Math.difftime
#> Math.default
#> * abs
```

`Math.difftime` basically looks like this:

```
Math.difftime <- function(x, ...) {
 new_difftime(NextMethod(), units = attr(x, "units"))
}
```

It dispatches to the next method, here the internal default, to perform the actual computation, then copies back over the the class and attributes.

Note that inside a group generic function a special variable `.Generic` provides the actual generic function called. This can be useful when producing error messages, and can sometimes be useful if you need to manually re-call the generic with different arguments.

### 13.7.5 Double dispatch

Generics in the “Ops” group, which includes the two-argument mathematical and logical operators like `-` and `&`, implement a special type of method dispatch. They dispatch on the type of *both* of the arguments, so called **double dispatch**. This is necessary to preserve the commutative property of many operators, i.e. `a + b` should equal `b + a`. Take the following simple example:

```
date <- as.Date("2017-01-01")
integer <- 1L

date + integer
#> [1] "2017-01-02"
integer + date
#> [1] "2017-01-02"
```

If `+` dispatched only on the first argument, it would return different values for the two cases. To overcome this problem, generics in the Ops group use a slightly different strategy from usual. Rather than doing a single method dispatch, they do two, one for each input. There are three possible outcomes of this lookup:

- The methods are the same, so it doesn’t matter which method is used.
- The methods are different, and R falls back to the internal method with a warning.
- One method is internal, in which case R calls the other method.

For the example above, we can look at the possible methods for each argument, taking advantage of the fact that we can call `+` with a single argument. In this case, the second argument would dispatch to the internal `+` function, so R will call `+.Date`.

```
s3_dispatch(+date)
#> -> +.Date
#> +.default
#> * Ops.Date
#> Ops.default
#> * +
s3_dispatch(+integer)
#> +.integer
#> +.numeric
#> +.default
#> Ops.integer
#> Ops.numeric
#> Ops.default
#> -> +
```

Let’s take a look at another case. What happens if you try to add a date to a factor? There is no method in common, so R calls the internal `+` method (which preserves the attributes of the LHS), with a warning.

```
factor <- factor("a")
s3_dispatch(+factor)
#> +.factor
#> +.default
```

```
#> -> Ops.factor
#> Ops.default
#> * +
#
date + factor
#> Warning: Incompatible methods ("+.Date", "Ops.factor") for "+"
#> [1] "2017-01-02"
factor + date
#> Warning: Incompatible methods ("Ops.factor", "+.Date") for "+"
#> Error in as.character.factor(x):
#> malformed factor
```

Finally, what happens if we try to subtract a `POSIXct` from a `POSIXlt`? A common `-.``POSIXt` method is found and called.

```
dt1 <- as.POSIXct(date)
dt2 <- as.POSIXlt(date)

s3_dispatch(-dt1)
#> -.POSIXct
#> -> -.POSIXt
#> -.default
#> Ops.POSIXct
#> * Ops.POSIXt
#> Ops.default
#> * -
s3_dispatch(-dt2)
#> -.POSIXlt
#> -> -.POSIXt
#> -.default
#> Ops.POSIXlt
#> * Ops.POSIXt
#> Ops.default
#> * -

dt1 - dt2
#> Time difference of 0 secs
```

### 13.7.6 Exercises

1. `Math.difftime()` is more complicated than I described. Why?



# Chapter 14

## S4

### 14.1 Introduction

Like S3, S4 implements functional OOP, but is much more rigorous and strict. There are three main differences between S3 and S4:

- S4 classes have formal definitions provided by a call to `setClass()`. An S4 class can have multiple parents (multiple inheritance).
- The fields of an S4 object are not attributes or named elements, but instead are called **slots** and are accessed with the special `@` operator.
- Methods are not defined with a naming convention, but are instead defined by a call to `setMethod()`. S4 generics can dispatch on multiple arguments (multiple dispatch).

A good overview of the motivation of S4 and its historical context can be found in Chambers and others (2014), [https://projecteuclid.org/download/pdfview\\_1/euclid.ss/1408368569](https://projecteuclid.org/download/pdfview_1/euclid.ss/1408368569).

S4 is a rich system, and it's not possible to cover all of it in one chapter. Instead, we'll focus on what you need to know to read most S4 code, and write basic S4 components. Unfortunately there is not one good reference for S4 and as you move towards more advanced usage, you will need to piece together needed information by carefully reading the documentation and performing experiments. Some good places to start are:

- Bioconductor course materials (<https://bioconductor.org/help/course-materials/>), a list of all courses taught by Bioconductor, a big user of S4. One recent (2017) course by Martin Morgan and Hervé Pagès is S4 classes and methods (<https://bioconductor.org/help/course-materials/2017/Zurich/S4-classes-and-methods.html>).
- S4 questions on stackoverflow (<http://stackoverflow.com/search?tab=votes&q=user%3a547331%20%5bs4%5d%20is%3aanswe>) answered by Martin Morgan.
- *Software for Data Analysis* (<http://amzn.com/0387759352?tag=devtools-20>), a book by John Chambers.

### Outline

#### Prerequisites

All S4 related functions live in the `methods` package. This package is always available when you're running R interactively, but may not be available when running R in batch mode (i.e. from `Rscript`). For this reason,

it's a good idea to call `library(methods)` whenever you use S4. This also signals to the reader that you'll be using the S4 object system.

```
library(methods)
```

## 14.2 Classes

Unlike S3, S4 classes have a formal definition. To define an S4 class, you must define three key properties:

- The class **name**. By convention, S4 class names use UpperCamelCase.
- A named character vector that describes the names and classes of the **slots** (fields). For example, a person might be represented by a character name and a numeric age: `c(name = "character", age = "numeric")`. The pseudo-class “ANY” allows a slot to accept objects of any type.
- The name of a class (or classes) to inherit behaviour from, or in S4 terminology, the classes that it **contains**.

Slots and contains can specify the names of S4 classes, S3 classes (if registered), and base types. We'll go into more detail about non-S4 classes at the end of the chapter, in S4 and existing code.

To create a class, you call `setClass()`, supplying these three properties. Lets make this concrete with an example. Here we create two classes: a person with character `name` and numeric `age`, and an `Employee` that inherits slots and methods from `Person`, adding an additional `boss` slot that must be a `Person`. `setClass()` returns a low-level constructor function, which should be given the class name with a `.` prefix.

```
.Person <- setClass("Person",
 slots = c(
 name = "character",
 age = "numeric"
)
)
.Employee <- setClass("Employee",
 contains = "Person",
 slots = c(
 boss = "Person"
)
)
```

`setClass()` has 10 other arguments, but they are all either deprecated or not recommended. If you have existing S4 code that uses them, I'd recommend carefully reading the documentation and upgrading to modern practice.

We can now use the constructor to create an object from that class:

```
hadley <- .Person(name = "Hadley", age = 37)
hadley
#> An object of class "Person"
#> Slot "name":
#> [1] "Hadley"
#>
#> Slot "age":
#> [1] 37
```

It's also possible to create an instance using `new()` and the name of the class. This is not recommended because it introduces some ambiguity. What happens if there are two packages that both define the `Person` class?

```
hadley2 <- new("Person", name = "Hadley", age = 37)
```

In most programming languages, class definition occurs at compile-time, and object construction occurs later, at run-time. In R, however, both definition and construction occur at run time. When you call `setClass()`, you are registering a class definition in a (hidden) global variable. As with all state-modifying functions you need to use `setClass()` with care. It's possible to create invalid objects if you redefine a class after already having instantiated an object:

```
.A <- setClass("A", slots = c(x = "numeric"))
a <- .A(x = 10)

.A <- setClass("A", slots = c(a_different_slot = "numeric"))
a
#> An object of class "A"
#> Slot "a_different_slot":
#> Error in slot(object, what):
#> no slot of name "a_different_slot" for this object of class "A"
```

This isn't usually a problem, because you'll define a class once, then leave the definition alone. If you want to enforce a single class definition, you can “seal” it:

```
setClass("Sealed", sealed = TRUE)
setClass("Sealed")
#> Error in setClass("Sealed"):
#> "Sealed" has a sealed class definition and cannot be redefined
```

### 14.2.1 Slots

You can access the slots with `@` or `slot()`: `@` is equivalent to `$`, and `slot()` to `[[]`.

```
hadley@age
#> [1] 37
slot(hadley, "age")
#> [1] 37
```

You can list all available slots with `slotNames()`:

```
slotNames(hadley)
#> [1] "name" "age"
```

Slots should be considered an internal implementation detail. That means:

- As a user, you should not reach into someone else's object with `@`, but instead, look for a method that provides the information you want.
- As a developer, you should make sure that all public facing slots have their own accessor methods.

We'll come back how to implement accessors in [Accessors], once you've learned how S4 generics and methods work.

### 14.2.2 Helper

The result of `setClass()` is a low-level constructor, which means that don't need to write one yourself. However, this default constructor has three drawbacks:

- The constructor takes ..., not individual named slots. This mean that printing the function is not revealing, and autocomplete doesn't have the data it needs to be helpful.

```
.Person
#> class generator function for class "Person" from package '.GlobalEnv'
#> function (...)

#> new("Person", ...)
```

- If you don't supply values for a slot, the constructor will automatically supply a default value:

```
.Person()
#> An object of class "Person"
#> Slot "name":
#> character(0)
#>
#> Slot "age":
#> numeric(0)
```

Here, you might prefer that `name` is required, or that `age` defaults to NA.

- While it's not possible to create an S4 object with the wrong slots or slots of the wrong type:

```
.Person(name = "Hadley", age = "thirty")
#> Error in validObject(.Object):
#> invalid class "Person" object: invalid object for slot "age" in class "Person": got class "character", expected "double"
#>
#> .Person(name = "Hadley", sex = "male")
#> Error in initialize(value, ...):
#> invalid name for slot of class "Person": sex
```

It is possible to create slots with the wrong lengths, or otherwise invalid values:

```
.Person(name = "Hadley", age = c(37, 99))
#> An object of class "Person"
#> Slot "name":
#> [1] "Hadley"
#>
#> Slot "age":
#> [1] 37 99
```

Like with S3, we resolve these issues by writing a helper function.

```
Person <- function(name, age = NULL, ...) {
 if (is.null(age)) {
 age <- rep(NA_real_, length(name))
 }

 stopifnot(length(name) == length(age))
 .Person(name = name, age = age)
}
```

This provides the behaviour that we want:

```
Name is now required
Person()
#> Error in Person():
#> argument "name" is missing, with no default

And name and age must have same length
```

```

Person("Hadley", age = c(30, 37))
#> Error in Person("Hadley", age = c(30, 37)):
#> length(name) == length(age) is not TRUE

And if not supplied, age gets a default value of NA
Person("Hadley")
#> An object of class "Person"
#> Slot "name":
#> [1] "Hadley"
#>
#> Slot "age":
#> [1] NA

```

It is *possible* to achieve the same effect by implementing an `initialize()` method, but the `initialize()` generic has a complicated contract and it is very hard to get all the details right.

To re-use checking code in a subclass, you can take advantage of a detail of the constructor: an unnamed argument is interpreted as predefined object from the parent class. For example, to define a constructor for the `Employee` class that reuses the `Person` helper, you first create a `Person()`, then pass that to the `.Employee` constructor.

```

Employee <- function(name, age, boss) {
 person <- Person(name = name, age = age)
 .Employee(person, boss = boss)
}

```

As with S3, if the validity checking code is lengthy or expensive, you should pull it out into a separate function which the helper calls.

### 14.2.3 Introspection

To determine what classes an object inherits from, use `is()`:

```

is(hadley)
#> [1] "Person"

```

To test if an object inherits from a specific class, use the second argument of `is()`:

```

is(hadley, "person")
#> [1] FALSE

```

If you are using a class provided by a package you can get help on it with `class?Person`.

### 14.2.4 Exercises

1. What happens if you define a new S4 class that doesn’t “contain” an existing class? (Hint: read about virtual classes in `?setClass`.)
2. Imagine you were going to reimplement ordered factors, dates, and data frames in S4. Sketch out the `setClass()` calls that you would use to define the classes. What should they inherit from? What slots should they use?

## 14.3 Generics and methods

The job of a generic is to perform method dispatch, i.e. find the method designed to handle the combination of classes passed to the generic. Here you'll learn how to define S4 generics and methods, then in the next section we'll explore precisely how S4 method dispatch works.

S4 generics have a similar structure to S3 generics, but are a little more formal. To create a new S4 generic, you call `setGeneric()` with a function that calls `standardGeneric("myGeneric")`.

```
setGeneric("myGeneric", function(x) standardGeneric("myGeneric"))
```

Note that it is bad practice to use `{` in the generic function. This triggers a special case that is more expensive, and generally best avoided.

Like `setClass()`, `setGeneric()` has many other arguments. There is only one that you need to know about: `signature`. This allows you to control the arguments that are used for method dispatch. If `signature` is not supplied, all arguments (apart from `...`) are used. It is occasionally useful to remove arguments from dispatch. This allows you to require that methods provide arguments like `verbose = TRUE` or `quiet = FALSE`, but they don't take part in dispatch.

A generic isn't useful without some methods, and in S4 you add methods with `setMethod()`. There are three important arguments: the name of the generic, the name of the class, and the method itself.

```
setMethod("myGeneric", "Person", function(x) {
 # method implementation
})
```

(Again, just like `setClass()`, `setMethod()` has other arguments, but you should never use them.)

### 14.3.1 Show method

As with S3, the most commonly defined S4 method controls printing, but in S4 we use a different generic: `show()`.

When defining a method for an existing generic, you need to first determine the arguments. You can get those from the documentation or by looking at the formals of the generic:

```
names(formals(getGeneric("show")))
#> [1] "object"
```

Our show method needs to have a single argument `object`:

```
setMethod("show", "Person", function(object) {
 cat(is(object)[[1]], "\n",
 " Name: ", object@name, "\n",
 " Age: ", object@age, "\n",
 sep = ""
)
})
hadley
#> Person
#> Name: Hadley
#> Age: 37
```

More formally, the second argument to `setMethod()` is called the `signature`. In S4, unlike S3, the signature can include multiple arguments. This makes method dispatch in S4 substantially more complicated, but avoids having to implement double-dispatch as a special case. We'll talk more about multiple dispatch in the next section.

### 14.3.2 Accessor methods

Slots are generally considered to be an internal implementation detail: they can change without warning and user code should avoid accessing them directly. Instead, all user-readable slots should get an **accessor**. If the slot is unique to the class, this can just be a function:

```
person_name <- function(x) x@name
```

But typically, you will want to define a generic and provide a method for your class:

```
setGeneric("name", function(x) standardGeneric("name"))
setMethod("name", "Person", function(x) x@name)

name(hadley)
#> [1] "Hadley"
```

If the slot is also writeable, you should provide a setter function. Typically this function will be more complicated than the getter because you'll need to check that the new value is valid, or you may need to modify other slots. Here we make sure that this functions only allows changing the values, not the length:

```
`person_name<-` <- function(x, value) {
 stopifnot(length(x@name) == length(value))
 x@name <- value
 x
}
```

Again, you'll typically want to do this with a method:

```
setGeneric("name<-", function(x, value) standardGeneric("name<-"))
setMethod("name<-", "Person", function(x, value) {
 stopifnot(length(x@name) == length(value))
 x@name <- value
 x
})

name(hadley) <- "Hadley Wickham"
name(hadley)
#> [1] "Hadley Wickham"
```

### 14.3.3 Coercion methods

To coerce S4 object from one class to another, use `as()`. One nice feature of S4 is that it provides default coercion methods for you:

```
mary <- new("Person", name = "Mary", age = 34)
roger <- new("Employee", name = "Roger", age = 36, boss = mary)

as(roger, "Person")
#> Person
#> Name: Roger
#> Age: 36
```

The defaults are not always quite right. For example, what happens if we try to coerce a Person to an Employee? The coercion succeeds because the `boss` slot is “helpfully” filled in with a default object:

```
mary_employee <- as(mary, "Employee")
mary_employee@boss
```

```
#> Person
#> Name:
#> Age:
```

We can override the default coercion to supply an informative error.

```
setAs("Person", "Employee", function(from) {
 stop("Can not coerce an Person to an Employee", call. = FALSE)
})
as(mary, "Employee")
#> Error: Can not coerce an Person to an Employee
```

#### 14.3.4 Introspection

To list all the methods that belong to a generic, or that are associated with a class, use `sloop::s4_methods_generic()` and `s4_methods_class()`:

```
library(sloop)
s4_methods_generic("initialize")
#> # A tibble: 14 x 4
#> generic class visible source
#> <chr> <chr> <lgl> <chr>
#> 1 initialize .environment TRUE ""
#> 2 initialize ANY TRUE methods
#> 3 initialize array TRUE ""
#> 4 initialize environment TRUE ""
#> 5 initialize envRefClass TRUE methods
#> 6 initialize externalRefMethod TRUE ""
#> 7 initialize matrix TRUE ""
#> 8 initialize MethodsList TRUE ""
#> 9 initialize Module TRUE Rcpp
#> 10 initialize mts TRUE ""
#> 11 initialize oldClass TRUE ""
#> 12 initialize signature TRUE ""
#> 13 initialize traceable TRUE ""
#> 14 initialize ts TRUE ""
s4_methods_class("Person")
#> # A tibble: 7 x 4
#> generic class visible source
#> <chr> <chr> <lgl> <chr>
#> 1 coerce Person TRUE R_GlobalEnv
#> 2 coerce<- Person TRUE R_GlobalEnv
#> 3 coerce<-> Person TRUE R_GlobalEnv
#> 4 myGeneric Person TRUE R_GlobalEnv
#> 5 name Person TRUE R_GlobalEnv
#> 6 name<-> Person TRUE R_GlobalEnv
#> 7 show Person TRUE R_GlobalEnv
```

If you're looking for the implementation of a specific method, you can use `selectMethod()`. You give it the name of the generic and the class (or classes) that it's called with:

```
selectMethod("show", "Person")
#> Method Definition:
#>
```

```
#> function (object)
#> {
#> cat(is(object)[[1]], "\n", " Name: ", object@name, "\n",
#> " Age: ", object@age, "\n", sep = "")
#> }
#> <bytecode: 0x6a124e0>
#>
#> Signatures:
#> object
#> target "Person"
#> defined "Person"
```

If you're using a method defined in a package, the easiest way to get help on it is to construct a valid call, and then put `?` in front of it. `?` will use the arguments to figure out which help file you need:

```
?show(hadley)
```

### 14.3.5 Exercises

1. In the definition of the generic, why is it necessary to repeat the name of the generic twice?
2. What's the difference between the generics generated by these two calls?

```
setGeneric("myGeneric", function(x) standardGeneric("myGeneric"))
setGeneric("myGeneric", function(x) {
 standardGeneric("myGeneric")
})
```
3. What happens if you define a method with different argument names to the generic?
4. What other ways can you find help for a method? Read `??"?"` and summarise the details.

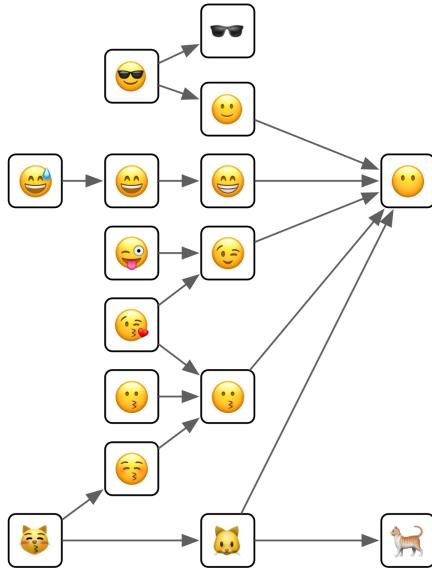
## 14.4 Method dispatch

S4 dispatch is complicated because S4 has two important features:

- Multiple inheritance, i.e. a class can have multiple parents,
- Multiple dispatch, i.e. a generic can use multiple arguments to pick a method.

These features make S4 very powerful, but can also make it hard to understand which method will get selected for a given combination of inputs.

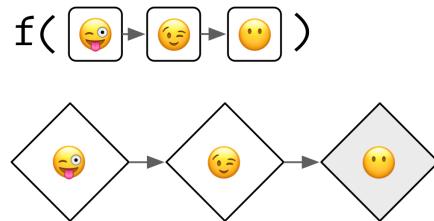
To explain method dispatch, we'll start simple with single inheritance and single dispatch, and work our way up to the more complicated cases. To illustrate the ideas without getting bogged down in the details, we'll use an imaginary **class graph** based on emoji:



Emoji give us very compact class names (just one symbol) that evoke the relationships between the classes. It should be straightforward to remember that `inherits` from `which` inherits from `,`, and that `inherits` from both `and`

#### 14.4.1 Single dispatch

Let's start with the simplest case: a generic function that dispatches on a single class with a single parent. The method dispatch here is quite simple, and the same as S3, but this will serve to define the graphical conventions we'll use for the more complex cases.



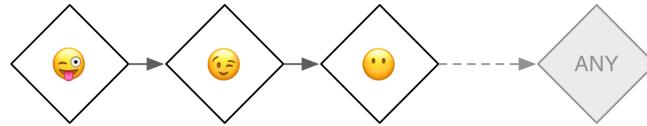
There are two parts to this diagram:

- The top part, `f(...)`, defines the scope of the diagram. Here we have a generic with one argument, and we're going to explore method dispatch for a class hierarchy that is three levels deep. We'll only ever look at a small fragment of the complete class graph. This keeps individual diagrams simple while helping you build intuition that you apply to more complex class graphs.
- The bottom part is the **method graph** and displays all the possible methods that could be defined. Methods that have been defined (i.e. with `setMethod()`) have a grey background.

To find the method that gets called, you start with the class of the actual arguments, then follow the arrows until you find a method that exists. For example, if you called the function with an object of class `you` you would follow the arrow right to find the method defined for the more general `class`. If no method is found, method dispatch has failed and you get an error. For this reason, class graphs should usually have methods defined for all the terminal nodes, i.e. those on the far right.

There are two pseudo-classes that you can define methods for. These are called pseudo-classes because they don't actually exist, but allow you to define useful behaviours. The first pseudo-class is "ANY". This matches

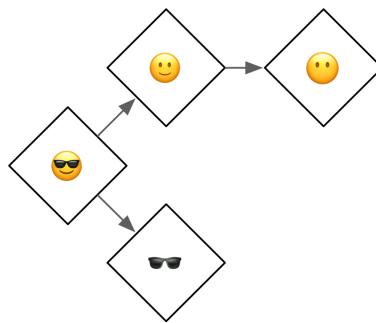
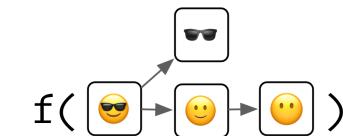
any class, and plays the same role as the `default` pseudo-class in S3. For technical reasons that we'll get to later, the link to the “ANY” method is longer than the links between the other classes:



The second pseudo-class is “MISSING”. If you define a method for this “class”, it will match whenever the argument is missing. It's generally not useful for functions that take a single argument, but can be used for functions like `+` and `-` that behave differently depending on whether they have one or two arguments.

#### 14.4.2 Multiple inheritance

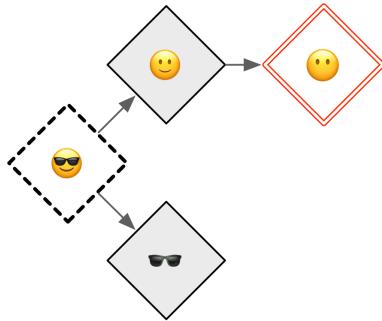
Things get more complicated when the class has multiple parents.



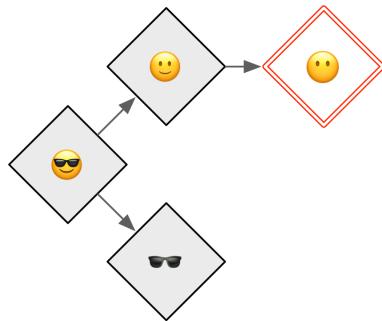
The basic process remains the same: you start from the actual class supplied to the generic, then follow the arrows until you find a defined method. The wrinkle is now that there are multiple arrows to follow, so you might find multiple methods. If that happens, you pick the method that is closest, i.e. requires travelling the fewest arrows.

(The method graph is a powerful metaphor that helps you understand how method dispatch works. However, implementing method dispatch in this way would be rather inefficient so the actual approach that S4 uses is somewhat different. You can read the details in [?Methods\\_Details](#))

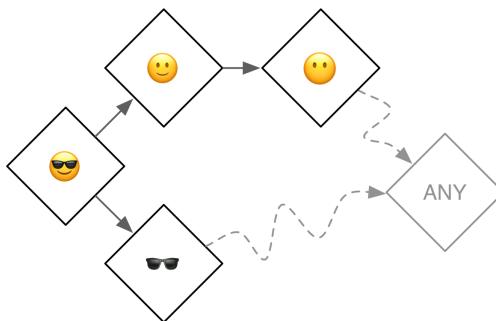
What happens if methods are the same distance? For example, imagine we've defined methods for `cool` and `smile`, and we call the generic with `.Note that there's no implementation for the cool class, as indicated by the red double outline.`



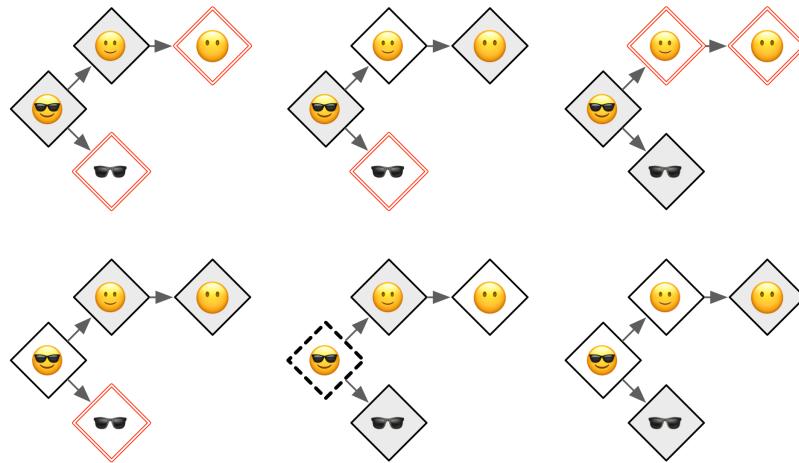
This is called an **ambiguous** method, and in diagrams I'll illustrate it with a thick dotted border. When this happens in R, you'll get a warning, and one of the two methods is basically picked at random (it uses the method that comes first in the alphabet). When you discover ambiguity you should always resolve it by providing a more precise method:



The fallback “ANY” method still exists but the rules are little more complex. As indicated by the wavy dotted lines, the “ANY” method is always considered further away than a method for a real class. This means that it will never contribute to ambiguity.



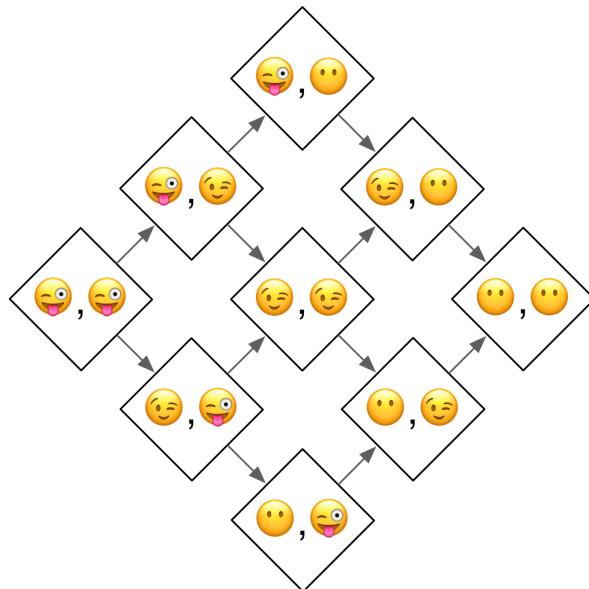
It is hard to simultaneously prevent ambiguity, ensure that every terminal method has an implementation, and minimise the number of defined methods (in order to benefit from OOP). For example, of the six ways to define only two methods for this call, only one is free from problems. For this reason, I recommend using multiple inheritance with extreme care: you will need to carefully think about the method graph and plan accordingly.



### 14.4.3 Multiple dispatch

Once you understand multiple inheritance, understanding multiple dispatch is straightforward. You follow multiple arrows in the same way as previously, but now each method is specified by two classes (separated by a comma).

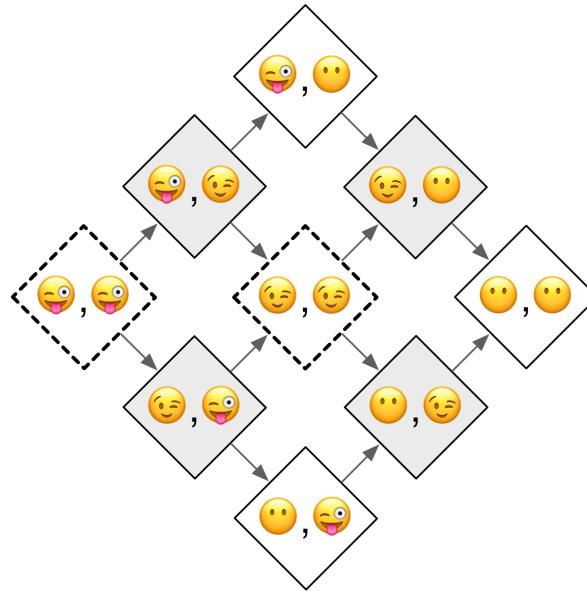
$f( \text{ } \text{ } \text{ } \text{ } \rightarrow \text{ } \text{ } \text{ } \text{ } , \text{ } \text{ } \text{ } \text{ } \rightarrow \text{ } \text{ } \text{ } \text{ } )$



I'm not going to show examples of dispatching on more than two arguments, but you can follow the basic principles to generate your own method graphs.

The main difference between multiple inheritance and multiple dispatch is that there are many more arrows to follow. The following diagram shows four defined methods which produce two ambiguous cases:

`f( [ ]->[ ]->[ ], [ ]->[ ]->[ ] )`

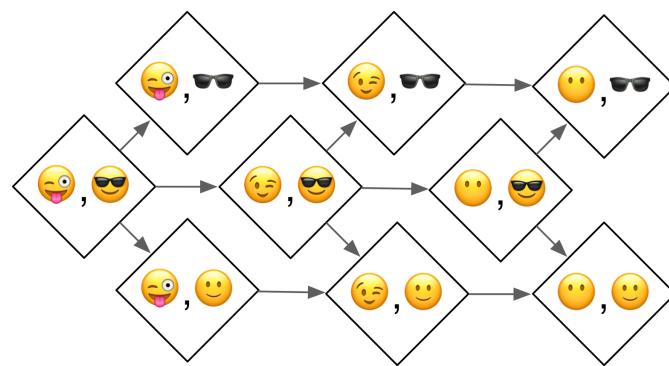


Multiple dispatch tends to be less tricky to work with than multiple inheritance because there are usually fewer terminal class combinations. In this example, there's only one. That means, at a minimum, you can define a single method and have default behaviour for all inputs.

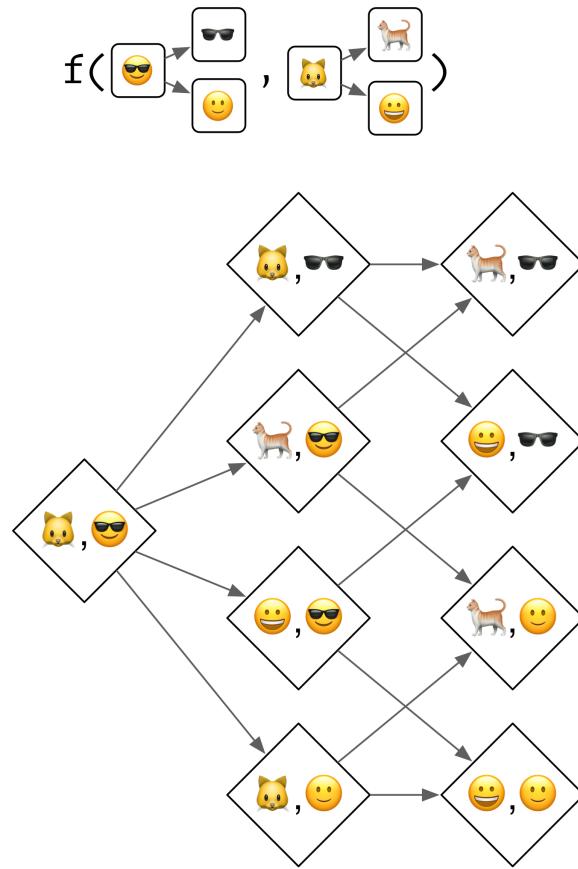
#### 14.4.4 Multiple dispatch and multiple inheritance

Of course you can combine multiple dispatch with multiple inheritance:

`f( [ ]->[ ]->[ ], [ ]->[ ]->[ ] )`



A still more complicated case dispatches on two classes, both of which have multiple inheritance:



However, as the method graph gets more and more complicated it gets harder and harder to predict which actual method will get called given a combination of inputs, and it gets harder and harder to make sure that you haven't introduced ambiguity. I highly recommend avoiding the combination of the two. There are some techniques (like mixins) that allow you to tame this complexity, but I am not aware of a detailed treatment as applied to S4.

#### 14.4.5 Exercises

- Take the last example which shows multiple dispatch over two classes that use multiple inheritance. What happens if you define a method for all terminal classes? Why does method dispatch not save us much work here?

## 14.5 S4 and existing code

Even when writing new S4 code, you'll still need to interact with existing S3 classes and functions, including existing S3 generics. This section describes how S4 classes, methods, and generics interact with existing code.

### 14.5.1 Classes

In `slots` and `contains` you can use S4 classes, S3 classes, or the implicit class of a base type. To use an S3 class, you must first register it with `setOldClass()`. You call this function once for each S3 class, giving it the class attribute. For example, the following definitions are already provided by base R:

```
setOldClass("data.frame")
setOldClass(c("ordered", "factor"))
setOldClass(c("glm", "lm"))
```

Generally, these definitions should be provided by the creator of the S3 class. If you’re trying to build an S4 class on top of a S3 class provided by a package, it is better to request that the package maintainer add this call to the package, rather than running it yourself.

If an S4 object inherits from an S3 class or a base type, it will have a special virtual slot called `.Data`. This contains the underlying base type or S3 object:

```
RangedNumeric <- setClass(
 "RangedNumeric",
 contains = "numeric",
 slots = c(min = "numeric", max = "numeric"))
)
rn <- RangedNumeric(1:10, min = 1, max = 10)
rn@min
#> [1] 1
rn@Data
#> [1] 1 2 3 4 5 6 7 8 9 10
```

It is possible to define S3 methods for S4 generics, and S4 methods for S3 generics (provided you’ve called `setOldClass()`). However, it’s more complicated than it might appear at first glance, so make sure you thoroughly read `?Methods_for_S3`.

### 14.5.2 Generics

As well as creating a new generic from scratch (as shown in `generics` and `methods`), it’s also possible to convert an existing function to a generic.

```
sides <- function(object) 0
setGeneric("sides")
```

In this case, the existing function becomes the default (“ANY”) method:

```
selectMethod("sides", "ANY")
#> Method Definition (Class "derivedDefaultMethod"):
#>
#> function (object)
#> 0
#>
#> Signatures:
#> object
#> target "ANY"
#> defined "ANY"
```

Note that `setMethod()` will automatically call `setGeneric()` if the first argument isn’t already a generic, enabling you to turn any existing function into an S4 generic. I think it is ok to convert an existing S3 generic to S4, but you should avoid converting regular functions because it makes code harder to use (and requires coordination if done by multiple packages).

### 14.5.3 Exercises

# Chapter 15

## R6

### 15.1 Introduction

This chapter describes the R6 object system. Unlike S3 and S4, it provides encapsulated OO, which means that:

- R6 methods belong to objects, not generics.
- R6 objects are mutable: the usual copy-on-modify semantics do not apply.

These properties make R6 objects behave more like objects in programming languages such as Python, Ruby and Java. This does not mean that R6 is good, and S3 and S4 are bad, it just means that R has a different heritage than most modern mainstream programming languages.

R6 is very similar to a built-in OO system called **reference classes**, or RC for short. I'm going to teach you R6 instead of RC for four reasons:

- R6 is much simpler. Both R6 and RC are built on top of environments, but while R6 uses S3, RC uses S4. R6 is only ~500 lines of R code (and ~1700 lines of tests!). We're not going to discuss the implementation in depth here, but if you've mastered the contents of this book, you should be able to read the source code and figure out how it works.
- RC mingles variables and fields in the same stack of environments so that you get (`field`) and set fields (`field <- value`) like regular values. R6 puts fields in a separate environment so you get (`self$field`) and set (`self$field <- value`) with a prefix. The R6 approach is more verbose but is worth the tradeoff because it makes code easier to understand. It also makes inheritance across packages simpler and more robust.
- R6 is much faster than RC. Generally, the speed of method dispatch is not important outside of microbenchmarks but R6 is substantially better than RC. Switching from RC to R6 yielded substantial performance in shiny. `vignette("Performance", "R6")` provides more details on the performance.
- Because the ideas that underlie R6 and RC are similar, it will only require a small amount of additional effort to learn RC if you need to.

### Outline

#### Prerequisites

Because R6 is not built into base R, you'll need to install and load a package in order to use it:

```
library(R6)
```

If you'd like to learn more about R6 after reading this chapter, the best place to start is the vignettes included in the package. You can list them by calling `browseVignettes(package = "R6")`.

## 15.2 Classes and methods

R6 only needs a single function call to create both the class and its methods: `R6::R6Class()`. And this is the only function from the package that you'll ever use! The following example shows the two most important arguments:

- The first argument is the `classname`. It's not strictly needed, but it improves error messages and makes it possible to also use R6 objects with S3 generics. By convention, R6 classes use UpperCamelCase.
- The second argument, `public`, supplies a list of methods (functions) and fields (anything else) that make up the public interface of the object. By convention, methods and fields use `snake_case`. Methods can access the methods and fields of the current object via `self$`.

```
Accumulator <- R6Class("Accumulator", list(
 sum = 0,
 add = function(x = 1) {
 self$sum <- self$sum + x
 invisible(self)
 }
))
```

You should always assign the result of `R6Class()` into a variable with the same name as the class. This creates an R6 object that defines the R6 class:

```
Accumulator
#> <Accumulator> object generator
#> Public:
#> sum: 0
#> add: function (x = 1)
#> clone: function (deep = FALSE)
#> Parent env: <environment: R_GlobalEnv>
#> Locked objects: TRUE
#> Locked class: FALSE
#> Portable: TRUE
```

You construct a new object from the class by calling the `new()` method. Methods belong to R6 objects so you use `$` to access `new()`:

```
x <- Accumulator$new()
```

You can then call methods and access fields with `$`:

```
x$add(4)
x$sum
#> [1] 4
```

In this class, the fields and methods are public which means that you can get or set the value of any field. Later, we'll see how to use private fields and methods to prevent casual access to the internals of your class.

To make it clear when we're talking about fields and methods as opposed to variables and functions, when referring to them in text, we'll prefix with `$`. For example, the `Accumulate` class has field `$sum` and method `$add()`.

### 15.2.1 Method chaining

`$add()` is called primarily for its side-effect of updating `$sum`.

```
Accumulator <- R6Class("Accumulator", list(
 sum = 0,
 add = function(x = 1) {
 self$sum <- self$sum + x
 invisible(self)
 })
)
```

Side-effect R6 methods should always return `self` invisibly. This returns the “current” object and makes it possible to chain together multiple method calls:

```
x$add(10)$add(10)$sum
#> [1] 24
```

Alternatively, for long chains, you can spread the call over multiple lines:

```
x$
 add(10)$
 add(10)$
 sum
#> [1] 44
```

This technique is called **method chaining** and is commonly used in encapsulated OO languages (like Python and JavaScript) to create fluent interfaces. Method chaining is deeply related to the pipe, and we’ll discuss the pros and cons of each approach in pipe vs method-chaining tradeoffs.

### 15.2.2 Important methods

There are two important methods that will be defined for most classes: `$initialize()` and `$print()`. You don’t have to provide them, but it’s a good idea to do so because they will make your class easier to use.

`$initialize()` overrides the default behaviour of `$new()`. For example, the following code defines an R6 Person class, similar to the S4 equivalent in S4. Unlike S4, R6 provides no checks for object type by default. `$initialize()` is a good place to check that `name` and `age` are the correct types.

```
Person <- R6Class("Person", list(
 name = NULL,
 age = NA,
 initialize = function(name, age = NA) {
 stopifnot(is.character(name), length(name) == 1)
 stopifnot(is.numeric(age), length(age) == 1)

 self$name <- name
 self$age <- age
 }
))

hadley <- Person$new("Hadley", age = 37)
```

If you have more expensive validation requirements, implement them in a separate `$validate()` and only call when needed.

Defining `$print()` allows you to override the default printing behaviour. As with any R6 method called for its side effects, `$print()` should return `invisible(self)`.

```
Person <- R6Class("Person", list(
 name = NULL,
 age = NA,
 initialize = function(name, age = NA) {
 self$name <- name
 self$age <- age
 },
 print = function(...) {
 cat("Person: \n")
 cat(" Name: ", self$name, "\n", sep = "")
 cat(" Age: ", self$age, "\n", sep = "")
 invisible(self)
 }
))

hadley2 <- Person$new("Hadley")
hadley2
#> Person:
#> Name: Hadley
#> Age: NA
```

This code illustrates an important aspect of R6. Because methods are bound to individual objects, the previously created `hadley` does not get this new method:

```
hadley
#> <Person>
#> Public:
#> age: 37
#> clone: function (deep = FALSE)
#> initialize: function (name, age = NA)
#> name: Hadley
```

Indeed, from the perspective of R6, there is no relationship between `hadley` and `hadley2`. This can make interactive experimentation with R6 confusing. If you're changing the code and can't figure out why the results of method calls aren't changed, make sure you've re-constructed R6 objects with the new class.

There's a useful alternative to `$print()`: implement `$format()`, which should return a character vector. This will automatically be used by both `print()` and `format()` S3 generics.

```
Person <- R6Class("Person", list(
 age = NA,
 name = NULL,
 initialize = function(name, age = NA) {
 self$name <- name
 self$age <- age
 },
 format = function(...) {
 # The first `paste0()` is not necessary but it lines up
 # with the subsequent lines making it easier to see how
 # it will print
 c(
 paste0("Person:"),
 paste0(" Name: ", self$name),
```

```

 paste0(" Age: ", self$age)
)
}
))

hadley3 <- Person$new("Hadley")
format(hadley3)
#> [1] "Person:" " Name: Hadley" " Age: NA"
hadley3
#> Person:
#> Name: Hadley
#> Age: NA

```

### 15.2.3 Adding methods after creation

Instead of continuously creating new classes, it's also possible to modify the methods of an existing class. This is useful when exploring interactively, and when you have a class with many functions that you'd like to break up into pieces.

Once the class has been defined, you can add elements to it with `$set()`, supplying the visibility (more on that below), the name, and the component.

```

Accumulator <- R6Class("Accumulator")
Accumulator$set("public", "sum", 0)
Accumulator$set("public", "add", function(x = 1) {
 self$sum <- self$sum + x
 invisible(self)
})

```

`$set()` will not overwrite an existing method unless you explicitly ask for it:

```

Accumulator$set("public", "sum", 1)
#> Error in Accumulator$set("public", "sum", 1):
#> Can't add sum because it already present in Accumulator generator.
Accumulator$set("public", "sum", 1, overwrite = TRUE)

```

Also note that adding methods will only affect new objects generated from the class. It does not retroactively apply to existing objects:

```

x1 <- Accumulator$new()
Accumulator$set("public", "hello", function() message("Hi!"))
x1$hello()
#> Error in eval(expr, envir, enclos):
#> attempt to apply non-function

x2 <- Accumulator$new()
x2$hello()
#> Hi!

```

### 15.2.4 Inheritance

To inherit behaviour from an existing class, provide the class object to the `inherit` argument:

```

AccumulatorChatty <- R6Class("AccumulatorChatty",
 inherit = Accumulator,
 public = list(
 add = function(x = 1) {
 cat("Adding ", x, "\n", sep = "")
 super$add(x = x)
 }
)
)

x2 <- AccumulatorChatty$new()
x2$add(10)$add(1)$sum
#> Adding 10
#> Adding 1
#> [1] 12

```

Note that `$add()` overrides the implementation in the superclass, but we can access the previous implementation through `super$`. Any methods which are overridden will automatically call the implementation in the parent class.

Like S3, R6 only supports single inheritance: you cannot supply a vector of classes to inherit.

### 15.2.5 Introspection

Every R6 object has an S3 class that reflects the hierarchy of R6 classes. This means that the easiest way to determine the class (and all classes it inherits from) is to use `class()`:

```

class(hadley3)
#> [1] "Person" "R6"

```

The S3 hierarchy includes the base “R6” class. This provides common behaviour, including an `print.R6()` method which calls `$print()` or `$format()`, as described above.

You can list all methods and fields with `names()`:

```

names(hadley3)
#> [1] ".__enclos_env__" "name" "age"
#> [4] "clone" "format" "initialize"

```

There’s one method that we haven’t defined: `$clone()`. It’s provided by R6 and we’ll come back to it in reference semantics.

### 15.2.6 Exercises

1. Can subclasses access private fields/methods from their parent? Perform an experiment to find out.

## 15.3 Controlling access

`R6Class()` has two other arguments that work similarly to `public`: `private` and `active`. `private` allows you to create components that the user can not easily access, and `active` allows you to use accessor functions to define dynamic, or active, fields.

### 15.3.1 Privacy

With R6 you can define **private** fields and methods, elements that can only be accessed from within the class, not from the outside. There are two things that you need to know to take advantage of private elements:

- The **private** argument works in the same way as the **public** argument: you give it a named list of methods (functions) and fields (everything else).
- Fields and methods defined in **private** are available within the methods with **private\$** instead of **self\$**. You cannot access private fields or methods outside of the class.

To make this concrete, we could make **\$age** and **\$name** fields of the Person class private. With this definition of Person we can only set **\$age** and **\$name** during object creation, and we cannot access their values from outside of the class.

```
Person <- R6Class("Person",
 public = list(
 initialize = function(name, age = NA) {
 private$name <- name
 private$age <- age
 },
 print = function(...) {
 cat("Person: \n")
 cat(" Name: ", private$name, "\n", sep = "")
 cat(" Age: ", private$age, "\n", sep = "")
 }
),
 private = list(
 age = NA,
 name = NULL
)
)

hadley4 <- Person$new("Hadley")
hadley4$name
#> NULL
```

The distinction between public and private fields is important when you create complex networks of classes, and you want to make it as clear as possible what it's ok for others to access. Anything that's private can be more easily refactored because you know others aren't relying on it. Private methods tend to be less important in R compared to other programming languages because the object hierarchies in R tend to be simpler.

### 15.3.2 Active fields

Active fields may allow you to define components that look like fields from the outside, but are defined with functions, like methods. For example, we can define an active field **x** that returns a different value every time you access it:

```
Rando <- R6::R6Class("Rando", active = list(
 random = function(value) {
 runif(1)
 }
))
x <- Rando$new()
x$random
```

```
#> [1] 0.0808
x$random
#> [1] 0.834
x$random
#> [1] 0.601
```

Active fields are particularly useful in conjunction with privacy, because they make it possible to implement components that work like fields from the outside but provide additional checks. For example, you can use them to implement read-only fields or fields that validate their inputs.

Active fields are implemented using active bindings from base R. Each active binding is a function that takes a single argument: `value`. If the argument is `missing()`, the value is being retrieved; otherwise it's being modified. We can use that idea to make a read-only `age` field, and to ensure that `name` is a length 1 character vector.

```
Person <- R6Class("Person",
 private = list(
 .age = NA,
 .name = NULL
),
 active = list(
 age = function(value) {
 if (missing(value)) {
 private$.age
 } else {
 stop("`$age` is read only", call. = FALSE)
 }
 },
 name = function(value) {
 if (missing(value)) {
 private$.name
 } else {
 stopifnot(is.character(value), length(value) == 1)
 private$.name <- value
 self
 }
 }
),
 public = list(
 initialize = function(name, age = NA) {
 private$.name <- name
 private$.age <- age
 }
)
)

hadley5 <- Person$new("Hadley")
hadley5$name
#> [1] "Hadley"
hadley5$name <- 10
#> Error in (function (value) :
#> is.character(value) is not TRUE

#> Error in {:
```

```
#> Error in if (missing(value)) {:
#> is.character(value) is not TRUE

#> Error in private$name:
#> is.character(value) is not TRUE

#> Error in }:
#> is.character(value) is not TRUE

#> Error in else {:
#> is.character(value) is not TRUE

#> Error in stopifnot(is.character(value), length(value) == 1):
#> is.character(value) is not TRUE

#> Error in private$name <- value:
#> is.character(value) is not TRUE

#> Error in self:
#> is.character(value) is not TRUE

#> Error in }:
#> is.character(value) is not TRUE

#> Error in })(quote(10)):
#> is.character(value) is not TRUE

hadley5$age
#> [1] NA
hadley5$age <- 20
#> Error: `age` is read only
```

### 15.3.3 Exercises

1. How would you define a write-only field?

## 15.4 Reference semantics

One of the big differences between R6 and most other objects in R is that they have reference semantics. This is because they are S3 objects built on top of environments:

```
typeof(x2)
#> [1] "environment"
```

The main consequence of reference semantics is that objects are not copied when modified:

```
y1 <- Accumulator$new()
y2 <- y1

y1$add(10)
c(y1 = y1$sum, y2 = y2$sum)
```

```
#> y1 y2
#> 11 11
```

Instead, if you want a copy, you'll need to explicitly `$clone()` the object:

```
y1 <- Accumulator$new()
y2 <- y1$clone()

y1$add(10)
c(y1 = y1$sum, y2 = y2$sum)
#> y1 y2
#> 11 1
```

(Note that `$clone()` does not recursively clone nested R6 objects. If you want that, you'll need to use `$clone(deep = TRUE)`. Note that this only clones R6 objects: if you have other fields with reference semantics (e.g. environments) you'll need to define your own `$clone()`.)

There are three other less obvious consequences:

- It is harder to reason about code that uses R6 objects because you need to understand more context.
- It makes sense to think about when an R6 object is deleted, and you can write a `finalizer()` to complement the `initializer()`.
- If one of the fields is an R6 class, you must call `$new()` inside `$initialize()` not inside `R6Class()`.

These are described in more detail below.

### 15.4.1 Reasoning

Generally, reference semantics makes code harder to reason about. Take this very simple example:

```
x <- list(a = 1)
y <- list(b = 2)

z <- f(x, y)
```

For the vast majority of functions, you know that the final line only modifies `z`.

Take a similar equivalent that uses an imaginary `List` reference class:

```
x <- List$new(a = 1)
y <- List$new(b = 2)

z <- f(x, y)
```

The final line is much harder to reason about - it's completely possible that `f()` calls methods of `x` or `y`, modifying them in place. This is the biggest potential downside of R6. The best way to ameliorate this problem is to avoid writing functions that both return a value and modify R6 inputs.

That said, modifying R6 inputs can lead to substantially simpler code in some cases. One challenge of working with immutable data is known as **threading state**: if you want to return a value that's modified in a deeply nested function, you need to return the modified value up through every function. This can complicate code, particularly if you need to modify multiple values. For example, ggplot2 uses R6 objects for scales. Scales are complex because they need to combine data across every facet and every layer. Using R6 makes the code substantially simpler, at the cost of introducing subtle bugs. Fixing those bugs required careful placement of calls to `$clone()` to ensure that independent plots didn't accidentally share scale data. We'll come back to this idea in [oo-tradeoffs].

### 15.4.2 Finalizer

One useful property of reference semantics is that it makes sense to think about when an R6 object is **finalised**, i.e. when it's deleted. This doesn't make sense for S3 and S4 objects because copy-on-modify semantics mean that there may be many transient versions of an object. For example, in the following code, there are actually two factor objects: the second is created when the levels are modified, leaving the first to be destroyed at the next garbage collection.

```
x <- factor(c("a", "b", "c"))
levels(x) <- c("c", "b", "a")
```

Since R6 objects are not copied-on-modify they will only get deleted once, and it makes sense to think about `$finalize()` as a complement to `$initialize()`. Finalizers usually play a similar role to `on.exit()`, cleaning up any resources created by the initializer. For example, the following class wraps up a temporary file, automatically deleting it when the class is finalised.

```
TemporaryFile <- R6Class("TemporaryFile", list(
 path = NULL,
 initialize = function() {
 self$path <- tempfile()
 },
 finalize = function() {
 message("Cleaning up ", self$path)
 unlink(self$path)
 }
))

tf <- TemporaryFile$new()
```

The `finalize` method will be run when R exits, or by the first garbage collection after the object has been removed. Generally, this will happen when it happens, but it can occasionally be useful to force a run with an explicit call to `gc()`.

```
rm(tf)
invisible(gc())
```

### 15.4.3 R6 fields

A final consequence of reference semantics can crop up where you don't expect it. Beware of setting a default value to an R6 class: it will be shared across all instances of the object. This is because the child object is only initialized once, when you defined the class, not each time you call `new`.

```
TemporaryDatabase <- R6Class("TemporaryDatabase", list(
 con = NULL,
 file = TemporaryFile$new(),
 initialize = function() {
 DBI::dbConnect(RSQLite::SQLite(), path = file$path)
 }
))

db_a <- TemporaryDatabase$new()
db_b <- TemporaryDatabase$new()

db_a$file$path == db_b$file$path
#> [1] TRUE
```

You can fix this by creating the object in `$initialize()`:

```
TemporaryDatabase <- R6Class("TemporaryDatabase", list(
 con = NULL,
 file = NULL,
 initialize = function() {
 self$file <- TemporaryFile$new()
 DBI::dbConnect(RSQLite::SQLite(), path = file$path)
 }
))

db_a <- TemporaryDatabase$new()
db_b <- TemporaryDatabase$new()

db_a$file$path == db_b$file$path
#> [1] FALSE
```

#### 15.4.4 Exercises

# Chapter 16

## Trade-offs

### 16.1 Introduction

You now know about the three most important OOP toolkits available in R. Now that you understand their basic operation and the principles that underlie them, we can start to compare and contrast the systems in order to understand their strengths and weaknesses. This will help you pick the system that is most likely to solve new problems.

When picking an OO system, I recommend that you default to S3. S3 is simple, and widely used throughout base R and CRAN. While it's far from perfect, its idiosyncracies are well understood and there are known approaches to overcome most shortcomings. If you have an existing background in programming you are likely to lean towards R6 because it will feel familiar. I think you should resist this tendency for two reasons. Firstly, if you use R6 it's very easy to create an non-idiomatic API that will feel very odd to native R users, and will have surprising pain points because of the reference semantics. Secondly, if you stick to R6, you'll lose out on learning a new way of thinking about OOP that gives you a new set of tools for solving problems.

### Outline

This chapter is divided into two parts. S4 vs S3 compares S3 and S4. In brief, S4 is more formal and tends to require more upfront planning. That makes it more suitable for big projects developed by teams, not individuals. R6 vs S3 compares S3 and R6. This section is quite long because these two systems are fundamentally different and there are a number of tradeoffs that you need to consider.

### 16.2 S4 vs S3

Once you've mastered S3, S4 is relatively easy to pick up: the underlying ideas are the same, S4 is just more formal, more strict, and more verbose. The strictness and formality of S4 make it well suited for large teams. Since more structure is provided by the system itself, there is less need for convention, and new contributors don't need as much training. S4 tends to require more upfront design than S3, and this investment tends to be more likely to pay off on larger projects because greater resources are available.

One large team effort where S4 is used to good effect is Bioconductor. Bioconductor is similar to CRAN: it's a way of sharing packages amongst a wider audience. Bioconductor is smaller than CRAN (~1,300 vs ~10,000 packages, July 2017) and the packages tend to be more tightly integrated because of the shared domain and because Bioconductor has a stricter review process. Bioconductor packages are not required to

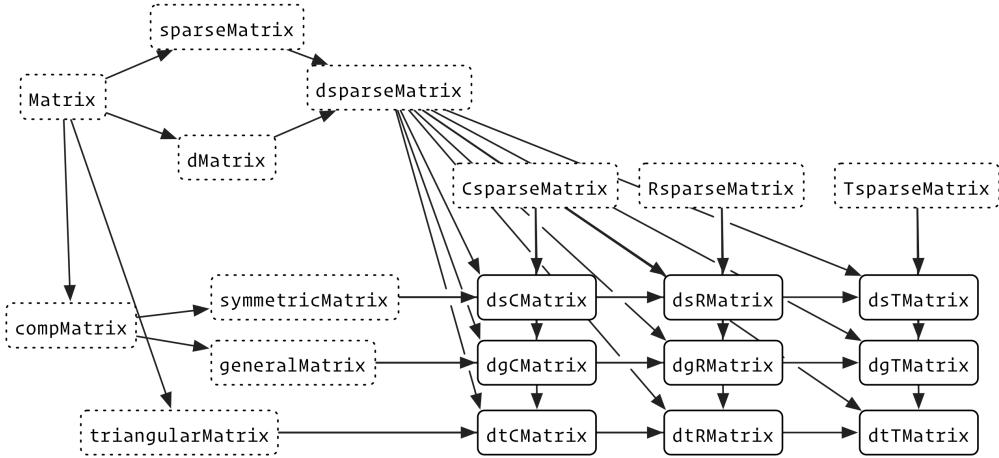


Figure 16.1: A small subset of the `Matrix` class graph showing the inheritance of sparse matrices. Each concrete class inherits from two virtual parents: one that describes how the data is stored (`C` = column oriented, `R` = row oriented, `T` = tagged) and one that describes any restriction on the matrix (`s` = symmetric, `t` = triangle, `g` = general)

use S4, but most will because the key data structures (e.g. `SummarizedExperiment`, `IRanges`, `DNAStringSet`) are built using S4.

S4 is also a good fit when you have a complicated system of interrelated objects, and it's possible to minimise code duplication through careful implementation of methods. The best example of this use of S4 is the `Matrix` package by Douglas Bates and Martin Mächler. It is designed to efficiently store and compute with many different types of sparse and dense matrices. As of version 1.2.14, it defines 102 classes, 21 generic functions, and 1993 methods. To give you some idea of the complexity, a small subset of the class graph is shown in Figure 16.1.

This domain is a good fit for S4 because there are often computational shortcuts for specific types of sparse matrix. S4 makes it easy to provide a general method that works for all inputs, and then provide a more specialised methods where the pair of data structures allow for a more efficient implementation. This requires careful planning to avoid method dispatch ambiguity, but the planning pays off for complicated systems.

The biggest challenge to using S4 is the combination of increased complexity and absence of a single source of documentation. S4 is a complex system and it can be challenging to use effectively in practice. This wouldn't be such a problem if S4 documentation wasn't scattered through R documentation, books, and websites. S4 needs a book length treatment, but that book does not (yet) exist. (The documentation for S3 is no better, but the lack is less painful because S3 is much simpler.)

### 16.3 R6 vs S3

R6 is a profoundly different OO system from S3 and S4 because it is built on encapsulated objects, rather than generic functions. Additionally R6 objects have reference semantics, which means that they can be modified in place. These two big differences have a number of non-obvious consequences which we'll explore in this chapter:

- A generic is a regular function so lives in the global namespace. A R6 method belongs to an object so lives in a local namespace. This influences how we think about naming.
- R6's reference semantics allow methods to simultaneously return a value and update the object. This solves a painful problem called "threading state".

- You invoke an R6 method using \$, which is an infix operator. If you set up your methods correctly you can use chains of method calls as an alternative to the pipe.

(All these trade-offs apply in general to immutable functional OOP vs mutable encapsulated OOP so also serve as a discussion of the tradeoffs between S3 and reference classes, and S3 and OOP in languages like Python.)

### 16.3.1 Namespacing

One non-obvious difference between S3 and R6 is the “space” in which methods are found:

- Generic functions are global: all packages share the same namespace.
- Encapsulated methods are local: methods are bound to a single object.

The advantage of a global namespace is that multiple packages can use exactly the same verbs for working with different types of objects. Generic functions provide a uniform API that makes it easier to perform typical actions with a new object because there are strong naming conventions. This works well for data analysis because you often want to do the same thing to different types of objects. In particular, this is one reason that R’s modelling system is so useful: regardless of where the model has been implemented you always work with it using the same set of tools (`summary()`, `predict()`, ...).

The disadvantage of a global namespace is that forces you to think more deeply about naming. You want to avoid multiple generics with the same name in different packages because it requires the user to type `::` frequently. This can be hard because function names are usually English verbs, and verbs often have multiple meanings. Take `plot()` for example:

```
plot(data) # plot some data
plot(bank_heist) # plot a crime
plot(land) # create a new plot of land
plot(movie) # extract plot of a movie
```

Generally, you should avoid defining methods like this. Don’t use homonyms of the original generic, but instead define a new generic. This problem doesn’t occur with R6 methods because they are scoped to the object. The following code is fine, because there is no implication that the `plot` method of two different R6 objects has the same meaning:

```
data$plot()
bank_heist$plot()
land$plot()
movie$plot()
```

These considerations also apply to the arguments to the generic. S3 generics must have the same core arguments, which mean they generally have to have non-specific names like `x` or `.data`. S3 generics generally need `...` to pass on additional arguments to methods, but this has the downside that misspelled argument names will not create an error. In comparison, R6 methods can vary more widely and use more specific and evocative argument names.

A secondary advantage of local namespacing is that creating an R6 method is very cheap. Most encapsulated OO languages encourage you to create many small methods, each doing one thing well with an evocative name. Creating a new S3 method is more expensive, because you may also have to create a generic, and think about the naming issues described above. That means that the advice to create many small methods does not apply to S3. It’s still a good idea to break your code down into small, easily understood chunks, but they should generally just be regular functions, not methods.

### 16.3.2 Threading state

One challenge of programming with S3 is when you want to both return a value and modify the object. This violates our guideline that a function should either be called for its return value or for its side effects, but is necessary in a handful of cases. For example, imagine you want to create a **stack** of objects. A stack has two main methods:

- `push()` adds a new object to the top of the stack.
- `pop()` returns the top most value, and removes it from the stack.

The implementation of the constructor and the `push()` method is straightforward. A stack contains a list of items, and pushing an object to the stack simply appends to this list.

```
new_stack <- function(items = list()) {
 structure(list(items = items), class = "stack")
}

push <- function(x, y) {
 x$items <- c(x$items, list(y))
 x
}
```

(Note that I haven't created a real method for `push()` because making it generic would just make this example more complicated for no real benefit.)

Implementing `pop()` is more challenging because it has to both return a value (the object at the top of the stack), and have a side-effect (remove that object from that top). Since we can't modify the input object in S3 we need to return two things: the value, and the updated object.

```
pop <- function(x) {
 n <- length(x$items)

 item <- x$items[[n]]
 x$items <- x$items[-n]

 list(item = item, x = x)
}
```

This leads to rather awkward usage:

```
s <- new_stack()
s <- push(s, 10)
s <- push(s, 20)

out <- pop(s)
out$item
#> [1] 20
s <- out$x
s
#> $items
#> $items[[1]]
#> [1] 10
#>
#>
#> attr(,"class")
#> [1] "stack"
```

This problem is known as **threading state** or **accumulator programming**, because no matter how deeply

the `pop()` is called, you have to feed the modified stack object all the way back to where the stack lives.

One way that other FP languages deal with this challenge is to provide a “multiple assign” (or destructing bind) operator that allows you to assign multiple values in a single step. The `zeallot` R package, by Nathan and Paul Teator, provides multi-assign for R with `%<-%`. This makes the code more elegant, but doesn’t solve the key problem:

```
library(zeallot)

c(value, s) %<-% pop(s)
value
#> [1] 10
```

An R6 implementation of a stack is simpler because `$pop()` can modify the object in place, and return only the top-most value:

```
Stack <- R6:::R6Class("Stack", list(
 items = list(),
 push = function(x) {
 self$items <- c(self$items, x)
 invisible(self)
 },
 pop = function() {
 item <- self$items[[self$length()]]
 self$items <- self$items[-self$length()]
 item
 },
 length = function() {
 length(self$items)
 }
))
```

This leads to more natural code:

```
s <- Stack$new()
s$push(10)
s$push(20)
s$pop()
#> [1] 20
```

### 16.3.3 Method chaining

The pipe, `%>%`, is useful because it provides an infix operator that makes it easy to compose functions from left-to-right. Interestingly, the pipe is not so important for R6 objects because they already use an infix operator: `$`. This allows the user to chain together multiple method calls in a single expression, a technique known as **method chaining**:

```
s <- Stack$new()
s$push(10)$
push(20)$
pop()
#> [1] 20
```

This technique is commonly used in other programming languages, like Python and JavaScript, and is made possible with one convention: any R6 method that is primarily called for its side-effects (usually modifying

the object) should return `invisible(self)`.

The primary advantage of method chaining is that you can get useful autocomplete; the primary disadvantage is that only the creator of the class can add new methods (and there's no way to use multiple dispatch).

# Part IV

# Metaprogramming



# Introduction

“Flexibility in syntax, if it does not lead to ambiguity, would seem a reasonable thing to ask of an interactive programming language.”

— Kent Pitman

One of the most surprising things about R is its capability for metaprogramming: the ability of code to inspect and modify other code. In R, functions that use metaprogramming are commonly said to use **non-standard evalution**, or NSE for short. That’s because they evaluate one (or more) of their arguments in a non-standard way. As you might guess, defining these tools by what they are not (standard evaluation) is challenging, so you’ll learn more precise vocabulary as you work through these chapters.

Additionally, implementation of the underlying ideas has occurred piecemeal over the last twenty years. These two forces tend to make base R metaprogramming code harder to understand than it could be: the key ideas are obscured by unimportant details. To focus on the main ideas, the following chapters will start with functions from the `rlang` package, which have been developed more recently with an eye for consistency. Once you have the basic ideas with `rlang`, I’ll show you the equivalent with base R so you can use your knowledge to understand existing code.

Metaprogramming is particularly important in R because it is well suited to facilitating interactive data analysis. There are two primary uses of metaprogramming that you have probably already seen:

- It makes it possible to trade precision for concision in functions like `subset()` and `dplyr::filter()` that make interactive data exploration faster at the cost of introducing some ambiguity.
- It makes it possible build **domain specific languages** (DSLs) that tailor R’s semantics to specific problem domains like visualisation or data manipulation.

We’ll briefly illustrate these important concepts before diving into the details of how they work in the subsequent chapters.

## 16.3.4 Trading precision for concision

A common use of metaprogramming is to allow you to use names of variables in a dataframe as if they were objects in the environment. This makes interactive exploration more fluid at the cost of introducing some minor ambiguity. For example, take `base::subset()`. It allows you to pick rows from a dataframe based on the values of their observations:

```
data("diamonds", package = "ggplot2")
subset(diamonds, x == 0 & y == 0 & z == 0)
#> # A tibble: 7 x 10
#> carat cut color clarity depth table price x y z
#> <dbl> <ord> <ord> <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>
#> 1 1 Very Good H VS2 63.3 53 5139 0 0 0
#> 2 1.14 Fair G VS1 57.5 67 6381 0 0 0
#> 3 1.56 Ideal G VS2 62.2 54 12800 0 0 0
```

```
#> 4 1.2 Premium D VVS1 62.1 59 15686 0 0 0
#> 5 2.25 Premium H SI2 62.8 59 18034 0 0 0
#> 6 0.71 Good F SI2 64.1 60 2130 0 0 0
#> 7 0.71 Good F SI2 64.1 60 2130 0 0 0
```

(Base R functions like `subset()` and `transform()` inspired the development of dplyr.)

`subset()` is considerably shorter than the equivalent code using `[` and `$` because you only need to provide the name of the data frame once:

```
diamonds[diamonds$x == 0 & diamonds$y == 0 & diamonds$z == 0,]
```

## 16.4 Domain specific languages

More extensive use of metaprogramming leads to DSLs like `ggplot2` and `dplyr`. DSLs are particularly useful because they make it possible to translate R code into another language. For example, one of the headline features of `dplyr` is that you can write R code that is automatically translated into SQL:

```
library(dplyr)

con <- DBI::dbConnect(RSQLite::SQLite(), filename = ":memory:")
mtcars_db <- copy_to(con, mtcars)

mtcars_db %>%
 filter(cyl > 2) %>%
 select(mpg:hp) %>%
 head(10) %>%
 show_query()

#> <SQL>
#> SELECT `mpg`, `cyl`, `disp`, `hp`
#> FROM `mtcars`
#> WHERE (`cyl` > 2.0)
#> LIMIT 10

DBI::dbDisconnect(con)
```

This is a useful technique because it makes it possible to retrieve data from a database without paying the high cognitive overhead of switching between R and SQL.

`ggplot2` and `dplyr` are known as **embedded** DSLs, because they take advantage of R's parsing and execution framework, but tailor R's semantics for specific tasks. If you're interested in learning more, I highly recommend *Domain Specific Languages* (<http://amzn.com/0321712943?tag=devtools-20>) by Martin Fowler. It discusses many options for creating a DSL and provides many examples of different languages.

## 16.5 Overview

In the following chapters, you'll learn about the three big ideas that underpin metaprogramming:

- In **Expressions**, Expressions, you'll learn that all R code forms a tree. You'll learn how to visualise that tree, how the rules of R's grammar convert linear sequences of characters into a tree, and how to use recursive functions to work with code trees.

- In **Quotation**, Quotation, you’ll learn to use tools from rlang to capture (“quote”) unevaluated function arguments. You’ll also learn about quasiquotation, which provides a set of techniques for “unquoting” input that makes it possible to easily generate new trees from code fragments.
- In **Evaluation**, Evaluation, you’ll learn about the inverse of quotation: evaluation. Here you’ll learn about an important data structure, the quosure, which ensures correct evaluation by capturing both the code to evaluate, and the environment in which to evaluate it. This chapter will show you how to put all the pieces together to understand how NSE in base R works, and how to write your own functions that work like `subset()`.
- Finally, in **Translating R code**, [Translation], you’ll see how to combine first class environments, lexical scoping, and metaprogramming to translate R code in to other languages, namely HTML and LaTeX.

Each chapter follows the same basic structure. You’ll get the lay of the land in introduction, then see a motivating example. Next you’ll learn the big ideas using functions from rlang, and then we’ll circle back to talk about how those ideas are expressed in base R. Each chapter finishes with a case study, using the ideas to solve a bigger problem.



# Chapter 17

## Expressions

### 17.1 Introduction

To compute on the language, we first need to understand its structure. That requires some new vocabulary, some new tools, and some new ways of thinking about R code. The first thing you'll need to understand is the distinction between an operation and its result. Take this code, which takes a variable `x` multiplies it by 10 and saves the result to a new variable called `y`. It doesn't work because we haven't defined a variable called `x`:

```
y <- x * 10
#> Error in eval(expr, envir, enclos):
#> object 'x' not found
```

It would be nice if we could capture the intent of the code, without executing the code. In other words, how can we separate our description of the action from performing it? One way is to use `rlang::expr()`:

```
z <- expr(y <- x * 10)
z
#> y <- x * 10
```

`expr()` returns a quoted **expression**: the R code that captures our intent.

In this chapter, you'll learn about the structure of those expressions, which will also help you understand how R executes code. Later, we'll learn about `eval()` which allows you to take such an expression and perform, or **evaluate**, it:

```
x <- 4
eval(z)
y
#> [1] 40
```

### Outline

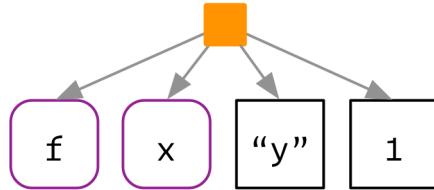
#### Prerequisites

Make sure you've installed `rlang` and `lobstr` from GitHub:

```
devtools::install_github("r-lib/rlang")
devtools::install_github("hadley/lobstr")
```

## 17.2 Abstract syntax trees

Quoted expressions are also called abstract syntax trees (AST) because the structure of code is hierarchical and can be naturally represented as a tree. To make that more obvious we're going to introduce some graphical conventions, illustrated with the very simple call `f(x, "y", 1)`.



- Function **calls** define the hierarchy of the tree. Calls are shown with an orange square. The first child (`f`) is the function that gets called; the second and subsequent children (`x`, `"y"`, and `1`) are the arguments.

**NB:** Unlike many tree diagrams the order of the children is important: `f(x, 1)` is not the same as `f(1, x)`.

- The leaves of the tree are either **symbols**, like `f` and `x`, or **constants** like `1` or `"y"`. Symbols have a purple border and rounded corners. Constants, which are atomic vectors of length one, have black borders and square corners. Strings are always surrounded in quotes to emphasise their difference from symbols — more on that later.

Drawing these diagrams by hand takes me some time, and obviously you can't rely on me to draw diagrams for your own code. I'll supplement the hand-drawn trees with trees drawn by `lobstr::ast()`. `ast()` tries to make trees as similar as possible to my hand-drawn trees, while respecting the limitations of the console. Let's use `ast()` to display the tree above:

```

lobstr::ast(f(x, "y", 1))
#> f
#> x
#> "y"
#> 1

```

Calls get an orange square, symbols are bold and purple, and strings are surrounded by quote marks. (The formatting is not currently shown in the book, but you can see it if you run the code yourself.)

`ast()` supports “unquoting” with `!!` (pronounced bang-bang). We'll talk about unquoting in detail in the next chapter; for now note that it's useful if you've already used `expr()` to capture the expression.

```

x <- expr(f(x, "y", 1))

not useful!
lobstr::ast(x)
#> x

what we want
lobstr::ast(!!x)
#> f
#> x
#> "y"
#> 1

```

For more complex code, you can also use RStudio's tree viewer to explore the AST interactively, e.g. `View(expr(y <- x * 10))`.

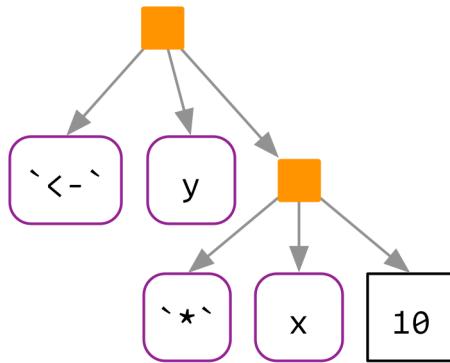
### 17.2.1 Infix vs. prefix calls

Every call in R can be written in tree form, even if it doesn't look like it at first glance. Take `y <- x * 10` again: what are the functions that are being called? It is not as easy to spot as `f(x, 1)` because this expression contains two calls in **infix** form: `<-` and `*`. Infix functions come **inbetween** their arguments (so an infix function can only have two arguments), whereas most functions in R are **prefix** functions where the name of the function comes first.<sup>1</sup>

In R, any infix call can be converted to a prefix call if you escape the function name with backticks. That means that these two lines of code are equivalent:

```
y <- x * 10
`<-`(y, `*(x, 10))
```

And they have this AST:



```
lobstr::ast(y <- x * 10)
#> `<-`
#> y
#> `*`
#> x
#> 10
```

You might remember that code like `names(x) <- y` ends up calling the `names<-` function. That is not reflected in the parse tree because the translation needs to happen later, due to the complexities of nested assignments like `names(x)[2] <- "z"`.

```
lobstr::ast(names(x) <- y)
#> `<-`
#> names
#> x
#> y
```

### 17.2.2 Special forms

R has a small number of other syntactical constructs that don't look like either prefix or infix function calls. These are called **special forms** and include `function`, the control flow operators (`if`, `for`, `while`, `repeat`), and parentheses (`{`, `(`, `[`, and `]`). These can also be written in prefix form, and hence appear in the same way in the AST:

---

<sup>1</sup>Some programming languages use **postfix** calls where the name of the function comes last. If you ever used an old HP calculator, you might have fallen in love with reverse Polish notation, postfix notation for algebra. There is also a family of "stack"-based programming languages descending from Forth which takes this idea as far as it might possibly go.

```
lobstr::ast(function(x, y) {
 if (x > y) {
 x
 } else {
 y
 }
})
#> `function`
#> x = ``
#> y = ``
#> `}`
#> `if`
#> `>`
#> x
#> y
#> `t`
#> x
#> `t`
#> y
#> <inline srcref>
```

Note that functions include a node `<inline srcref>`, this contains the source reference for the function, as mentioned in function components.

### 17.2.3 Function factories

Another small detail we need to consider are calls like `f()()`. The first component of the call is usually a symbol:

```
lobstr::ast(f(a, 1))
#> f
#> a
#> 1
```

But if you are using a function factory (as described in function factories), a function that returns another function, the first component might be another call:

```
lobstr::ast(f()(a, 1))
#> f
#> a
#> 1
```

And of course that function might also take arguments:

```
lobstr::ast(f(b, 2)(a, 1))
#> f
#> b
#> 2
#> a
#> 1
```

These forms are relatively rare, but it's good to be able to recognise them when they crop up.

### 17.2.4 Argument names

So far the examples have only used unnamed arguments. Named arguments don't change the parsing rules, but just add some additional metadata:

```
lobstr::ast(mean(x = mtcars$cyl, na.rm = TRUE))
#> mean
#> x = `$`
#> mtcars
#> cyl
#> na.rm = TRUE
```

(Note the appearance of another infix function: \$)

### 17.2.5 Exercises

1. Use `ast()` and experimentation to figure out the three arguments to `if()`. What would you call them? Which arguments are required and which are optional?
2. What does the call tree of an `if` statement with multiple `else if` conditions look like? Why?
3. What are the arguments to the `for()` and `while()` calls?
4. Two arithmetic operators can be used in both prefix and infix style. What are they?

## 17.3 R's grammar

The process by which a computer language takes a sequence of tokens (like `x`, `+`, `y`) and constructs a tree is called **parsing**, and it is governed by a set of rules known as a **grammar**. In this section, we'll use `lobstr::ast()` to explore some of the details of R's grammar.

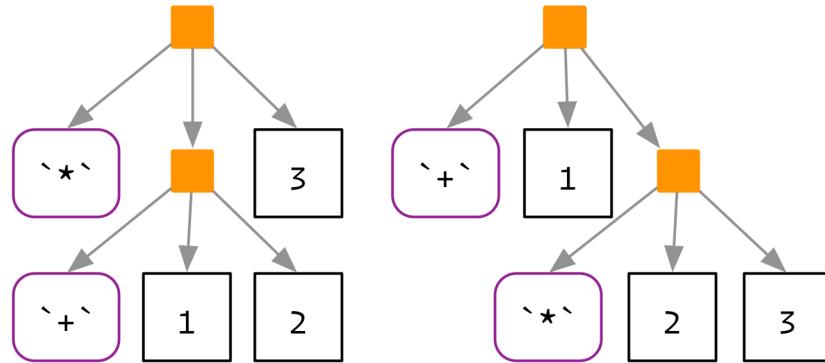
If this is your first reading of the metaprogramming chapters, now is a good time to read the first sections of the next two chapters in order to get the big picture. Come back and learn more of the details once you've seen how all the big pieces fit together.

### 17.3.1 Operator precedence

Infix functions introduce ambiguity in a way that prefix functions do not. The parser has to resolve two sources of ambiguity when parsing infix operators<sup>2</sup>. First, what does `1 + 2 * 3` yield? Do you get 9 (i.e. `(1 + 2) * 3`), or 7 (i.e. `1 + (2 * 3)`). Which of the two possible parse trees below does R use?

---

<sup>2</sup>These two sources of ambiguity do not exist without infix operators, which can be considered an advantage of purely prefix and postfix languages. It's interesting to compare a simple arithmetic operation in Lisp (prefix) and Forth (postfix). In Lisp you'd write `(+ (+ 1 2) 3)`; this avoids ambiguity by requiring parentheses everywhere. In Forth, you'd write `1 2 + 3 +`; this doesn't require any parentheses, but does require more thought when reading.



Programming languages use conventions called **operator precedence** to resolve this ambiguity. We can use `ast()` to see what R does:

```
lobstr::ast(1 + 2 * 3)
#> `+`
#> 1
#> `*`
#> 2
#> 3
```

Predicting the precedence of arithmetic operations is usually easy because it's drilled into you in school and is consistent across the vast majority of programming languages. Predicting the precedence of other operators is harder. There's one particularly surprising case in R: `!` has a much lower precedence (i.e. it binds less tightly) than you might expect. This allows you to write useful operations like:

```
lobstr::ast(!x %in% y)
#> `!`
#> `%in%`
#> x
#> y
```

R has over 30 infix operators divided into 18 precedence groups. While the details are described in `?Syntax`, very few people have memorised the complete ordering. Indeed, if there's any confusion, use parentheses! These also appear in the AST, like all other special forms:

```
lobstr::ast(1 + (2 + 3))
#> `+`
#> 1
#> `(`
#> `+`
#> 2
#> 3
```

### 17.3.2 Associativity

Another source of ambiguity is introduced by repeated usage of the same infix function. For example, is  $1 + 2 + 3$  equivalent to  $(1 + 2) + 3$  or to  $1 + (2 + 3)$ ? This normally doesn't matter because  $x + (y + z) == (x + y) + z$ , i.e. addition is associative, but is needed because some S3 classes define `+` in a non-associative way. For example, `ggplot2` overloads `+` to build up a complex plot from simple pieces; this usage is non-associative because earlier layers are drawn underneath later layers.

In R, most operators are **left-associative**, i.e. the operations on the left are evaluated first:

```
lobstr::ast(1 + 2 + 3)
#> `+`
#> `+`
#> 1
#> 2
#> 3
```

There are two exceptions: exponentiation and assignment.

```
lobstr::ast(2 ^ 2 ^ 3)
#> `^`
#> 2
#> `^`
#> 2
#> 3
lobstr::ast(x <- y <- z)
#> `<-`
#> x
#> `<-`
#> y
#> z
```

### 17.3.3 Whitespace

R, in general, is not sensitive to white space. Most white space is not significant and is not recorded in the AST: `x+y` yields exactly the same AST as `x + y`. This means that you're generally free to add whitespace to enhance the readability of your code. There's one major exception:

```
lobstr::ast(y <- x)
#> `<-`
#> y
#> x
lobstr::ast(y <- -x)
#> `<-`
#> y
#> `--`
#> x
```

### 17.3.4 Exercises

1. R uses parentheses in two slightly different ways as illustrated by these two calls:

```
f((1))
`(`(`(1 + 1)
```

Compare and contrast the two uses by referencing the AST.

2. `=` can also be used in two ways. Construct a simple example that shows both uses.
3. What does `!1 + !1` return? Why?
4. Why does `x1 <- x2 <- x3 <- 0` work? There are two reasons.
5. Compare the ASTs `x + y %+% z` and `x ^ y %+% z`. What does that tell you about the precedence of custom infix functions?

## 17.4 Data structures

Now that you have a good feel for ASTs and how R’s grammar helps to define them, it’s time to learn about the underlying implementation. In this section you’ll learn about the data structures that appear in the AST:

- Constants and symbols form the leaves of the tree.
- Calls form the branches of the tree.
- Pairlists are a largely historical data structure that are now only used for function arguments.

### 17.4.1 Naming conventions

Before we continue, a word of caution about the naming conventions used in this book. Because base R evolved organically, it does not have a set of names that are used consistently throughout all functions. Instead, we’ve adopted our own set of conventions, and used them consistently throughout the book and in `rlang`. You will need to remember some translations when reading base R documentation.

The biggest difference is the use of the term “expression”. We use expression to refer to the set containing constants, symbols, calls, and pairlists. In base R, “expression” is a special type that is basically equivalent to a list of what we call expressions. To avoid confusion we’ll call these **expression objects**, and we’ll discuss them in expression objects. Base R does not have an equivalent term for our “expression”. The closest is “language object”, which includes symbols and calls, but not constants or pairlists.

But note that `typeof()` and `str()` use “language” not for language objects, but instead to mean calls. Base R uses symbol and name interchangeably; we prefer symbol because “name” has other common meanings (e.g. the name of a variable).

### 17.4.2 Constants

Constants occurred in the leaves of the AST. They are the simplest data structure found in the AST because they are atomic vectors of length 1. Constants are “self-quoting” in the sense that the expression used to represent a constant is the constant itself:

```
identical(expr("x"), "x")
#> [1] TRUE
identical(expr(TRUE), TRUE)
#> [1] TRUE
identical(expr(1), 1)
#> [1] TRUE
identical(expr(2), 2)
#> [1] TRUE
```

### 17.4.3 Symbols

Symbols represent variable names. They are basically a single string stored in a special way. You can convert back and forth between symbols and the strings that represent them with `sym()` and `as_string()`:

```
"x"
#> [1] "x"
sym("x")
#> x
as_string(sym("x"))
#> [1] "x"
```

Symbols are scalars: if you want multiple symbols, you'll need to put them in a list. This is what `syms()` does:

```
syms(c("a", "bcd"))
#> [[1]]
#> a
#>
#> [[2]]
#> bcd
```

The big difference between strings and symbols is what happens when you evaluate them: evaluating a string returns the string; evaluating a symbol returns the value associated with the symbol in the current environment.

There's one special symbol that needs a little extra discussion: the empty symbol which is used to represent missing arguments (not missing values!). You can make it with `missing_arg()` (or `expr()`):

```
missing_arg()
typeof(missing_arg())
#> [1] "symbol"
as_string(missing_arg())
#> [1] ""
```

And see if you have a missing symbol with `rlang::is_missing()`:

```
is_missing(missing_arg())
#> [1] TRUE
```

This symbol has a peculiar property: if you bind it to a variable, then access that variable, you will get an error:

```
m1 <- missing_arg()
m1
#> Error in eval(expr, envir, enclos):
#> argument "m1" is missing, with no default
```

But you won't get an error if it's stored inside another data structure!

```
m2 <- list(missing_arg())
m2[[1]]
```

This is the magic that makes missing arguments work in functions. If you do need to work with a missing argument stored in a variable, you can use `rlang::maybe_missing()`:

```
maybe_missing(m1)
```

That prevents the error from occurring and instead returns another empty symbol.

You only need to care about the missing symbol if you're programmatically creating functions with missing arguments; we'll come back to that in the next chapter.

#### 17.4.4 Calls

Calls define the tree in AST. A call behaves similarly to a list:

- It has a `length()`.
- You can extract elements with `[[`, `[`, and `$`.
- Calls can contain other calls.

The main difference is that the first element of a call is special: it's the function that will get called. Let's explore these ideas with a simple example:

```
x <- expr(read.table("important.csv", row = FALSE))
lobstr::ast(!!x)
#> read.table
#> "important.csv"
#> row = FALSE
```

The length of a call minus one gives the number of arguments:

```
length(x) - 1
#> [1] 2
```

The names of a call are empty, except for named arguments:

```
names(x)
#> [1] "" "" "row"
```

You can extract the leaves of the call by position and by name using [[ and \$ in the usual way:

```
x[[1]]
#> read.table
x[[2]]
#> [1] "important.csv"

x$row
#> [1] FALSE
```

Extracting specific arguments from calls is challenging because of R's flexible rules for argument matching: it could potentially be in any location, with the full name, with an abbreviated name, or with no name. To work around this problem, you can use `rlang::call_standardise()` which standardises all arguments to use the full name:

```
rlang::call_standardise(x)
#> read.table(file = "important.csv", row.names = FALSE)
```

(Note that if the function uses ... it's not possible to standardise all arguments.)

You can use [ to extract multiple components, but if you drop the the first element, you're going to end up with a weird call:

```
x[2:3]
#> "important.csv"(row = FALSE)
```

If you do want to extract multiple elements in this way, it's good practice to coerce the results to a list:

```
as.list(x[2:3])
#> [[1]]
#> [1] "important.csv"
#>
#> $row
#> [1] FALSE
```

Calls can be modified in the same way as lists:

```
x$header <- TRUE
x
#> read.table("important.csv", row = FALSE, header = TRUE)
```

You can construct a call from its children by using `rlang::lang()`. The first argument should be the

function to be called (supplied either as a string or a symbol), and the subsequent arguments are the call to that function:

```
lang("mean", x = expr(x), na.rm = TRUE)
#> mean(x = x, na.rm = TRUE)
lang(expr(mean), x = expr(x), na.rm = TRUE)
#> mean(x = x, na.rm = TRUE)
```

### 17.4.5 Pairlists

There is one data structure we need to discuss for completeness: the pairlist. Pairlists are a remnant of R's past and have been replaced by lists almost everywhere. The only place you are likely to see pairlists in R is when working with function arguments:

```
f <- function(x = 10) x + 1
typeof(formals(f))
#> [1] "pairlist"
```

(If you're working in C, you'll encounter pairlists more often. For example, calls are also implemented using pairlists.)

Fortunately, whenever you encounter a pairlist, you can treat it just like a regular list:

```
pl <- pairlist(x = 1, y = 2)
length(pl)
#> [1] 2
pl$x
#> [1] 1
```

However, behind the scenes pairlists are implemented using a different data structure, a linked list instead of a vector. That means that subsetting is slower with pairlists, and gets slower the further along the pairlist you index. This has limited practical impacts, but it's a useful fact to know.

```
l1 <- as.list(1:100)
l2 <- as.pairlist(l1)

microbenchmark::microbenchmark(
 l1[[1]],
 l1[[100]],
 l2[[1]],
 l2[[100]]
)
#> Unit: nanoseconds
#> expr min lq mean median uq max neval
#> l1[[1]] 142 155 293 163 184 11201 100
#> l1[[100]] 146 160 196 166 222 437 100
#> l2[[1]] 1392 1677 1868 1810 2012 2887 100
#> l2[[100]] 1519 1841 2224 2008 2164 11465 100
```

### 17.4.6 Expression objects

Finally, we need to briefly discuss the expression object. Expression objects are produced by only two base functions: `expression()` and `parse()`:

```

exp1 <- parse(text = c(
 x <- 4
 x
))
exp2 <- expression(x <- 4, x)

typeof(exp1)
#> [1] "expression"
typeof(exp2)
#> [1] "expression"

exp1
#> expression(x <- 4, x)
exp2
#> expression(x <- 4, x)

```

Like calls and pairlists, expression objects behave like a list:

```

length(exp1)
#> [1] 2
exp1[[1]]
#> x <- 4

```

Conceptually, an expression object is just a list of expressions. The only difference is that calling `eval()` on an expression evaluates each individual expression. We don't believe this advantage merits introducing a new data structure, so instead of expression objects we always use regular lists of expressions.

#### 17.4.7 Exercises

1. Which two of the six types of atomic vector can't appear in an expression? Why? Why can't you create an expression that contains an atomic vector of length greater than one?
2. How is `rlang::maybe_missing()` implemented? Why does it work?
3. `rlang::call_standardise()` doesn't work so well for the following calls. Why? What makes `mean()` special?

```

call_standardise(quote(mean(1:10, na.rm = TRUE)))
#> mean(x = 1:10, na.rm = TRUE)
call_standardise(quote(mean(n = T, 1:10)))
#> mean(x = 1:10, n = T)
call_standardise(quote(mean(x = 1:10, , TRUE)))
#> mean(x = 1:10, , TRUE)

```

4. Why does this code not make sense?

```

x <- expr(foo(x = 1))
names(x) <- c("x", "")

```

5. Construct the expression `if(x > 1) "a" else "b"` using multiple calls to `lang()`. How does the structure code reflect the structure of the AST?

## 17.5 Parsing and deparsing

Most of the time you type code into the console, and R takes care of turning the characters you've typed into an AST. But occasionally you have code stored in a string, and you want to parse it yourself. You can do so using `rlang::parse_expr()`:

```
x1 <- "y <- x + 10"
lobstr::ast (!!x1)
#> "y <- x + 10"

x2 <- rlang::parse_expr(x1)
x2
#> y <- x + 10
lobstr::ast (!!x2)
#> `<-`
#> y
#> `+`
#> x
#> 10
```

If you have multiple expressions in a string, you'll need to use `rlang::parse_exprs()`. It returns a list of expressions:

```
x3 <- "a <- 1; a + 1"
rlang::parse_exprs(x3)
#> [[1]]
#> a <- 1
#>
#> [[2]]
#> a + 1
```

(If you find yourself working with strings containing code very frequently, you should reconsider your process. Read the next chapter and consider if you can instead more safely generate expressions using quasiquotation.)

The base equivalent to `parse_exprs()` is `parse()`. It is a little harder to use because it's specialised for parsing R code stored in files. That means you need supply your string to the `text` argument, and you get back an expression object:

```
parse(text = x1)[[1]]
#> y <- x + 10
```

The opposite of parsing is **deparsing**: you have an AST and you want a string that would generate it when parsed:

```
z <- expr(y <- x + 10)
expr_text(z)
#> [1] "y <- x + 10"
```

Parsing and deparsing are not perfectly symmetric because parsing throws away all information not directly related to the AST. This includes backticks around ordinary names, comments, and whitespace:

```
cat(expr_text(expr({
 # This is a comment
 x <- `x` + 1
})))
#> {
#> x <- x + 1
#> }
```

There are few other cases where parsing and deparsing is not symmetric. We'll encounter one in the next chapter:

```
expr_text(parse_expr("!!x"))
#> [1] "!!x"
```

Deparsing is often used to provide default names for data structures (like data frames), and default labels for messages or other output. rlang provides two helpers for those situations:

```
z <- expr(f(x, y, z))

expr_name(z)
#> [1] "f(x, y, z)"
expr_label(z)
#> [1] "`f(x, y, z)`"
```

Be careful when using the base R equivalent, `deparse()`: it returns a character vector with one element for each line. Whenever you use it, remember that the length of the output might be greater than one, and plan accordingly.

### 17.5.1 Exercises

1. What happens if you attempt to parse an invalid expression? e.g. "a +" or "f()".
2. `deparse()` produces vectors when the input is long. For example, the following call produces a vector of length two:

```
expr <- expr(g(a + b + c + d + e + f + g + h + i + j + k + l + m +
 n + o + p + q + r + s + t + u + v + w + x + y + z))

deparse(expr)
```

What do `expr_text()`, `expr_name()`, and `expr_label()` do with this input?

3. Why does `as.Date.default()` use `substitute()` and `deparse()`? Why does `pairwise.t.test()` use them? Read the source code.
4. `pairwise.t.test()` assumes that `deparse()` always returns a length one character vector. Can you construct an input that violates this expectation? What happens?

## 17.6 Case study: Walking the AST with recursive functions

To conclude the chapter I'm going to pull together everything that you've learned about ASTs and use that knowledge to solve more complicated problems. The inspiration comes from the base `codetools` package, which provides two interesting functions:

- `findGlobals()` locates all global variables used by a function. This can be useful if you want to check that your function doesn't inadvertently rely on variables defined in their parent environment.
- `checkUsage()` checks for a range of common problems including unused local variables, unused parameters, and the use of partial argument matching.

Getting all of the details of these functions correct is fiddly, so we won't explore their full expression. Instead we'll focus on the big underlying idea: recursion on the AST. Recursive functions are a natural fit to tree-like data structures because a recursive function is made up of two parts that correspond to the two parts of the tree:

- The **recursive case** handles the nodes in the tree. Typically, you'll do something to each child of node, usually calling the recursive function again, and then combine the results back together again. For expressions, you'll need to handle calls and pairlists (function arguments).
- The **base case** handles the leaves of the tree. The base cases ensure that the function eventually terminates, by solving the simplest cases directly. For expressions, you need to handle symbols and constants in the base case.

To make this pattern easier to see, we'll need two helper functions. First we define `expr_type()` which will return "constant" for constant, "symbol" for symbols, "call", for calls, "pairlist" for pairlists, and the "type" of anything else:

```
expr_type <- function(x) {
 if (rlang::is_syntactic_literal(x)) {
 "constant"
 } else if (is.symbol(x)) {
 "symbol"
 } else if (is.call(x)) {
 "call"
 } else if (is.pairlist(x)) {
 "pairlist"
 } else {
 typeof(x)
 }
}

expr_type(expr("a"))
#> [1] "constant"
expr_type(expr(f(1, 2)))
#> [1] "call"
```

We'll couple this with a wrapper around the `switch` function:

```
switch_expr <- function(x, ...) {
 switch(expr_type(x),
 ...,
 stop("Don't know how to handle type ", typeof(x), call. = FALSE)
)
}
```

With these two functions in hand, the basic template for any function that walks the AST is as follows:

```
recurse_call <- function(x) {
 switch_expr(x,
 # Base cases
 symbol = ,
 constant = ,

 # Recursive cases
 call = ,
 pairlist =
)
}
```

Typically, solving the base case is easy, so we'll do that first, then check the results. The recursive cases are a little more tricky. Typically you'll think about the structure of final output and then find the correct purrr function to produce it. To that end, make sure you're familiar with Functionals before continuing.

### 17.6.1 Finding F and T

We'll start simple with a function that determines whether a function uses the logical abbreviations T and F: it will return TRUE if it finds a logical abbreviation, and FALSE otherwise. Using T and F is generally considered to be poor coding practice, and is something that R CMD check will warn about.

Let's first compare the AST for T vs. TRUE:

```
ast(TRUE)
#> TRUE
ast(T)
#> T
```

TRUE is parsed as a logical vector of length one, while T is parsed as a name. This tells us how to write our base cases for the recursive function: a constant is never a logical abbreviation, and a symbol is an abbreviation if it's "F" or "T":

```
logical_abbr_rec <- function(x) {
 switch_expr(x,
 constant = FALSE,
 symbol = as_string(x) %in% c("F", "T")
)
}

logical_abbr_rec(expr(TRUE))
#> [1] FALSE
logical_abbr_rec(expr(T))
#> [1] TRUE
```

I've written `logical_abbr_rec()` function assuming that the input will be an expression as this will make the recursive operation simpler. However, when writing a recursive function it's common to write a wrapper that provides defaults or makes the function a little easier to use. Here we'll typically make a wrapper that quotes its input (we'll learn more about that in the next chapter), so we don't need to use `expr()` every time.

```
logical_abbr <- function(x) {
 logical_abbr_rec(enexpr(x))
}

logical_abbr(T)
#> [1] TRUE
logical_abbr(FALSE)
#> [1] FALSE
```

Next we need to implement the recursive cases. Here it's simple because we want to do the same thing for calls and for pairlists: recursively apply the function to each subcomponent, and return TRUE if any subcomponent contains a logical abbreviation. This is made easy by `purrr::some()`, which iterates over a list and returns TRUE if the predicate function is true for any element.

```
logical_abbr_rec <- function(x) {
 switch_expr(x,
 # Base cases
 constant = FALSE,
 symbol = as_string(x) %in% c("F", "T"),
 # Recursive cases
 call = ,
```

```

 pairlist = purrr::some(x, logical_abbr_rec)
)
}

logical_abbr(mean(x, na.rm = T))
#> [1] TRUE
logical_abbr(function(x, na.rm = T) FALSE)
#> [1] TRUE

```

## 17.6.2 Finding all variables created by assignment

`logical_abbr()` is very simple: it only returns a single `TRUE` or `FALSE`. The next task, listing all variables created by assignment, is a little more complicated. We'll start simply, and then make the function progressively more rigorous.

We start by looking at the AST for assignment:

```

ast(x <- 10)
#> `->`
#> x
#> 10

```

Assignment is a call where the first element is the symbol `<-`, the second is name of variable, and the third is the value to be assigned.

Next, we need to decide what data structure we're going to use for the results. Here I think it will be easiest if we return a character vector. If we return symbols, we'll need to use a `list()` and that makes things a little more complicated.

With that in hand we can start by implementing the base cases and providing a helpful wrapper around the recursive function. The base cases here are really simple!

```

find_assign_rec <- function(x) {
 switch_expr(x,
 constant = ,
 symbol = character()
)
}
find_assign <- function(x) find_assign_rec(enexpr(x))

find_assign("x")
#> character(0)
find_assign(x)
#> character(0)

```

Next we implement the recursive cases. This is made easier by a function that should exist in `purrr`, but currently doesn't. `flat_map_chr()` expects `.f` to return a character vector of arbitrary length, and flattens all results into a single character vector.

```

flat_map_chr <- function(.x, .f, ...) {
 purrr::flatten_chr(purrr::map(.x, .f, ...))
}

flat_map_chr(letters[1:3], ~ rep(., sample(3, 1)))
#> [1] "a" "a" "b" "b" "b" "c" "c" "c"

```

The recursive case for pairlists is simple: we iterate over every element of the pairlist (i.e. each function argument) and combine the results. The case for calls is a little bit more complex - if this is a call to `<-` then we should return the second element of the call:

```
find_assign_rec <- function(x) {
 switch_expr(x,
 # Base cases
 constant = ,
 symbol = character(),

 # Recursive cases
 pairlist = flat_map_chr(as.list(x), find_assign_rec),
 call = {
 if (is_call(x, "<-")) {
 as_string(x[[2]])
 } else {
 flat_map_chr(as.list(x), find_assign_rec)
 }
 }
)
}

find_assign(a <- 1)
#> [1] "a"
find_assign({
 a <- 1
 {
 b <- 2
 }
})
#> [1] "a" "b"
```

Now we need to make our function more robust by coming up with examples intended to break it. What happens when we assign to the same variable multiple times?

```
find_assign({
 a <- 1
 a <- 2
})
#> [1] "a" "a"
```

It's easiest to fix this at the level of the wrapper function:

```
find_assign <- function(x) unique(find_assign_rec(enexpr(x)))

find_assign({
 a <- 1
 a <- 2
})
#> [1] "a"
```

What happens if we have nested calls to `<-`? Currently we only return the first. That's because when `<-` occurs we immediately terminate recursion.

```
find_assign({
 a <- b <- c <- 1
})
```

```
#> [1] "a"
```

Instead we need to take a more rigorous approach. I think it's best to keep the recursive function focused on the tree structure, so I'm going to extract out `find_assign_call()` into a separate function.

```
find_assign_call <- function(x) {
 if (is.call(x, "<-") && is.symbol(x[[2]])) {
 lhs <- as.string(x[[2]])
 children <- as.list(x)[-1]
 } else {
 lhs <- character()
 children <- as.list(x)
 }

 c(lhs, flat_map_chr(children, find_assign_rec))
}

find_assign_rec <- function(x) {
 switch_expr(x,
 # Base cases
 constant = ,
 symbol = character(),

 # Recursive cases
 pairlist = flat_map_chr(x, find_assign_rec),
 call = find_assign_call(x)
)
}

find_assign(a <- b <- c <- 1)
#> [1] "a" "b" "c"
find_assign(system.time(x <- print(y <- 5)))
#> [1] "x" "y"
```

While the complete version of this function is quite complicated, it's important to remember we wrote it by working our way up by writing simple component parts.

### 17.6.3 Exercises

- `logical_abbr()` returns TRUE for `T(1, 2, 3)`. How could you modify `logical_abbr_rec()` so that it ignores function calls that use `T` or `F`?
- `logical_abbr()` works with expressions. It currently fails when you give it a function. Why not? How could you modify `logical_abbr()` to make it work? What components of a function will you need to recurse over?

```
f <- function(x = TRUE) {
 g(x + T)
}
logical_abbr(!f)
```

- Modify `find_assignment` to also detect assignment using replacement functions, i.e. `names(x) <- y`.
- Write a function that extracts all calls to a specified function.



# Chapter 18

## Quasiquotation

### 18.1 Introduction

Now that you understand the tree structure of R code, it's time to come back to one of the fundamental ideas that make `expr()` and `ast()` work: **quasiquotation**. There are two sides to quasiquotation:

- **Quotation** allows you to capture the AST associated with an argument. As a function author, this gives you a lot of power to influence how expressions are evaluated.
- **Unquotation** allows you to selectively evaluate parts of a quoted expression. This is a powerful tool that makes it easy to build up a complex AST from simpler fragments.

The combination of these two ideas makes it easy to compose expressions that are mixtures of direct and indirect specification, and helps to solve a wide variety of challenging problems.

Quoting functions have deep connections to Lisp **macros**. But macros are usually run at compile-time, which doesn't have any meaning in R, and they always input and output ASTs. (Lumley (2001) shows one way you might implement them in R). Quoting functions are more closely related to Lisp **fexprs** (<http://en.wikipedia.org/wiki/Fexpr>), functions where all arguments are quoted by default. These terms are useful to know when looking for related techniques in other programming languages.

### Outline

#### Prerequisites

Make sure you're familiar with the tree structure of code described in Abstract syntax trees.

You'll also need the development version of rlang:

```
if (packageVersion("rlang") < "0.2.0") {
 stop("This chapter requires rlang 0.2.0", call. = FALSE)
}
library(rlang)
```

### 18.2 Motivation

We'll start with a simple and concrete example that helps motivate the need for unquoting, and hence quasiquotation. Imagine you're creating a lot of strings by joining together words:

```
paste("Good", "morning", "Hadley")
#> [1] "Good morning Hadley"
paste("Good", "afternoon", "Alice")
#> [1] "Good afternoon Alice"
```

You are sick and tired of writing all those quotes, and instead you just want to use bare words. To that end, you've managed to write the following function:

```
cement <- function(...) {
 dots <- exprs(...)
 paste(purrr::map(dots, expr_name), collapse = " ")
}

cement(Good, morning, Hadley)
#> [1] "Good morning Hadley"
cement(Good, afternoon, Alice)
#> [1] "Good afternoon Alice"
```

(You'll learn what `exprs()` does shortly; for now just look at the results.)

Formally, this function **quotes** the arguments in .... You can think of it as automatically putting quotation marks around each argument. That's not precisely true as the intermediate objects it generates are expressions, not strings, but it's a useful approximation for now.

This function is nice because we no longer need to type quotes. The problem, however, comes when we want to use variables. It's easy to use variables with `paste()` as we just don't surround them with quotes:

```
name <- "Hadley"
time <- "morning"

paste("Good", time, name)
#> [1] "Good morning Hadley"
```

Obviously this doesn't work with `cement()` because every input is automatically quoted:

```
cement(Good, time, name)
#> [1] "Good time name"
```

We need some way to explicitly **unquote** the input, to tell `cement()` to remove the automatic quote marks. Here we need `time` and `name` to be treated differently to `Good`. Quasiquotation give us a standard tool to do so: `!!`, called “unquote”, and pronounced bang-bang. `!!` tells a quoting function to drop the implicit quotes:

```
cement(Good, !!time, !!name)
#> [1] "Good morning Hadley"
```

It's useful to compare `cement()` and `paste()` directly. `paste()` evaluates its arguments, so we need to quote where needed; `cement()` quotes its arguments, so we need to unquote where needed.

```
paste("Good", time, name)
cement(Good, !!time, !!name)
```

### 18.2.1 Vocabulary

The distinction between quoted and evaluated arguments is important:

- An **evaluated** argument obeys R's usual evaluation rules.

- A **quoted** argument is captured by the function and something unusual will happen.

If you’re even unsure about whether an argument is quoted or evaluated, try executing the code outside of the function. If it doesn’t work, then that argument is quoted. For example, you can use this technique to determine that the first argument to `library()` is quoted:

```
works
library(MASS)

fails
MASS
#> Error in eval(expr, envir, enclos):
#> object 'MASS' not found
```

Talking about whether an argument is quoted or evaluated is a more precise way of stating whether or not a function uses NSE. I will sometimes use “quoting function” as short-hand for a “function that quotes one or more arguments”, but generally, I’ll refer to quoted arguments since that is the level at which the difference occurs.

### 18.2.2 Theory

Now that you’ve seen the basic idea, it’s time to talk a little bit about the theory. The idea of quasiquotation is an old one. It was first developed by a philosopher, Willard van Orman Quine<sup>1</sup>, in the early 1940s. It’s needed in philosophy because it helps when precisely delineating the use and mention of words, i.e. between the object and the words we use to refer to that object.

Quasiquotation was first used in a programming language, LISP, in the mid-1970s (Bawden 1999). LISP has one quoting function ` , and uses , for unquoting. Most languages with a LISP heritage behave similarly. For example, racket (` and @), clojure (` and ~), and julia (: and @) all have quasiquotation tools that differ only slightly from LISP.

Quasiquotation has only come to R recently (2017). Despite its newness, I teach it in this book because it is a rich and powerful theory that makes many hard problems much easier. Quasiquotation in R is a little different to LISP and descendants. In LISP there is only one function that does quasiquotation (the quote function), and you must call it explicitly when needed. This makes these languages less ambiguous (because there’s a clear code signal that something odd is happening), but is less appropriate for R because quasiquotation is such an important part of DSLs for data analysis.

### 18.2.3 Exercises

1. For each function in the following base R code, identify which arguments are quoted and which are evaluated.

```
library(MASS)

mtcars2 <- subset(mtcars, cyl == 4)

with(mtcars2, sum(vs))
sum(mtcars2$am)

rm(mtcars2)
```

---

<sup>1</sup>You might be familiar with the name Quine from “quines”, computer programs that when run return a copy of their own source code.

2. For each function in the following tidyverse code, identify which arguments are quoted and which are evaluated.

```
library(dplyr)
library(ggplot2)

by_cyl <- mtcars %>%
 group_by(cyl) %>%
 summarise(mean = mean(mpg))

ggplot(by_cyl, aes(cyl, mean)) + geom_point()
```

## 18.3 Quotation

The first part of quasiquotation is quotation: capturing an AST without evaluating it. There are two components to this: capturing an expression directly, and capturing an expression from a lazily-evaluated function argument. We'll discuss two sets of tools for these two ways of capturing: those provided by rlang, and those provided by base R.

### 18.3.1 With rlang

There are four important quoting functions, broken down by whether they capture one or many expressions, and whether they capture the developer's or user's expression:

|      | Developer            | User                   |
|------|----------------------|------------------------|
| One  | <code>expr()</code>  | <code>enexpr()</code>  |
| Many | <code>exprs()</code> | <code>enexprs()</code> |

For interactive exploration, the most important quoting function is `expr()`. It captures its argument exactly as provided:

```
expr(x + y)
#> x + y
expr(1 / 2 / 3)
#> 1/2/3
```

(Remember that white space and comments are not part of the AST, so will not be captured by a quoting function.)

`expr()` is great for interactive exploration, because it captures what you, the developer, typed. It's not useful inside a function:

```
f1 <- function(x) expr(x)
f1(a + b + c)
#> x
```

Instead, we need another function: `enexpr()`. This captures what the user supplies to the function by looking at the internal promise object that powers lazy evaluation.

```
f2 <- function(x) enexpr(x)
f2(a + b + c)
#> a + b + c
```

(Occasionally you just want to capture symbols, and throw an error for other types of input. In that case you can use `ensym()`. In the next chapter, you'll learn about `enquo()` which also captures the environment and is needed for tidy evaluation.)

To capture multiple arguments, use `enexprs()`:

```
f <- function(...) enexprs(...)
f(x = 1, y = 10 * z)
#> $x
#> [1] 1
#>
#> $y
#> 10 * z
```

Finally, `exprs()` is useful interactively to make a list of expressions:

```
exprs(x = x ^ 2, y = y ^ 3, z = z ^ 4)
#> $x
#> x^2
#>
#> $y
#> y^3
#>
#> $z
#> z^4
shorthand for
list(x = expr(x ^ 2), y = expr(y ^ 3), z = expr(z ^ 4))
```

Note that it can return missing arguments:

```
val <- exprs(x =)
is_missing(val$x)
#> [1] TRUE
```

There's not much you can do with a list of expressions yet, but we'll see a few techniques later in case studies (quasi-case-studies): using `purrr` to work with lists of expressions turns out to be a surprisingly powerful tool.

Use `enexpr()` and `enexprs()` inside a function when you want to capture the expressions supplied as arguments *by the user* of that function. Use `expr()` and `exprs()` when you want to capture expressions that *you* supply.

### 18.3.2 With base R

The base equivalent of `expr()` is `quote()`:

```
quote(x + y)
#> x + y
quote(1 / 2 / 3)
#> 1/2/3
```

It is identical to `expr()` except that does not support unquoting, so it is a quoting function, not a quasiquoting function.

The base function closest to `enexpr()` is `substitute()`:

```
f3 <- function(x) substitute(x)
f3(x + y + z)
#> x + y + z
```

You'll most often see it used to capture unevaluated arguments; often in concert with `deparse()` to create labels for output. However, `substitute()` also does "substitution": if you give it an expression, rather than a symbol, it will substitute in values of symbols defined in the current environment.

```
f4 <- function(x) substitute(x * 2)
f4(a + b + c)
#> (a + b + c) * 2
```

`substitute()` provides a sort of automatic unquoting for any symbol that is bound to a value. However, making use of this behaviour can make for hard to read code, because for example, taken out of context, you can't tell if the goal of `substitute(x + y)` is to replace `x`, or, `y`, or both. If you do want to use `substitute()` in this way, I recommend that you use the 2nd argument to make it clear that is your goal:

```
substitute(x * y * z, list(x = 10, y = quote(a + b)))
#> 10 * (a + b) * z
```

The base equivalent to `exprs()` is `alist()`:

```
alist(x = 1, y = x + 2)
#> $x
#> [1] 1
#>
#> $y
#> x + 2
```

There are two other important base quoting functions that we'll cover elsewhere:

- `bquote()` provides a limited form of quasiquote, and is discussed in unquoting with base R.
- `~`, the formula, is a quoting function that also captures the environment. It's the inspiration for quosures, the topic of the next chapter, and is discussed in [formulas].

### 18.3.3 Exercises

1. What happens if you try to use `enexpr()` with an expression? What happens if you try to use `enexpr()` with a missing argument?
2. Compare and contrast the following two functions. Can you predict the output before running them?

```
f1 <- function(x, y) {
 exprs(x = x, y = y)
}
f2 <- function(x, y) {
 enexprs(x = x, y = y)
}
f1(a + b, c + d)
f2(a + b, c + d)
```

3. How are `exprs(a)` and `exprs(a = )` different? Think about both the input and the output.
4. What does the following command return? What information is lost? Why?

```
expr({
 x +
 y # comment
})
```

5. The documentation for `substitute()` says:

Substitution takes place by examining each component of the parse tree as follows: If it is not a bound symbol in env, it is unchanged. If it is a promise object, i.e., a formal argument to a function or explicitly created using `delayedAssign()`, the expression slot of the promise replaces the symbol. If it is an ordinary variable, its value is substituted, unless env is `.GlobalEnv` in which case the symbol is left unchanged.

Create four examples that illustrate each of the different cases.

## 18.4 Evaluation

Typically you have quoted a function argument for one of two reasons:

- You want to operate on the AST using the techniques described in the previous chapter.
- You want to run, or **evaluate** the code in a special context, as described in depth next chapter.

Evaluation is a rich topic, so we'll cover it in depth in the next chapter. Here I'll just illustrate the most important ideas.

The most important base R function is `base:::eval()`. Its first argument is the expression to evaluate:

```
ru5 <- expr(runif(5))
ru5
#> runif(5)

eval(ru5)
#> [1] 0.0808 0.8343 0.6008 0.1572 0.0074
eval(ru5)
#> [1] 0.466 0.498 0.290 0.733 0.773
```

Note that every time we evaluate this expression we get a different result.

The second argument to `eval()` is the environment in which the expression is evaluated. Manipulating this environment gives us amazing power to control the execution of R code. This is the basic technique that gives dbplyr the ability to turn R code into SQL.

```
x <- 9
fx <- expr(f(x))

eval(fx, env=f = function(x) x * 10))
#> [1] 90
eval(fx, env=f = function(x) x ^ 2))
#> [1] 81
```

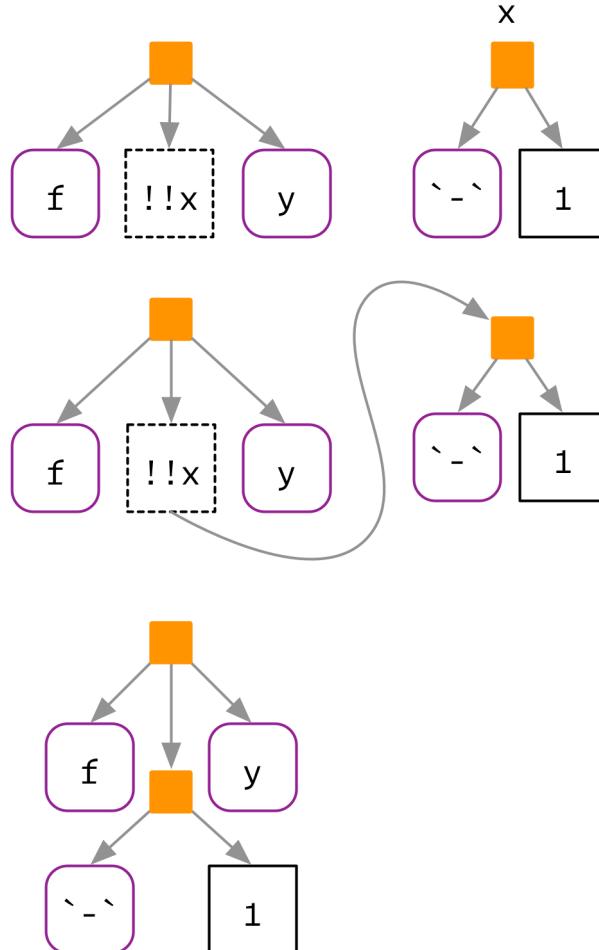
## 18.5 Unquotation

Evaluation is a developer tool: in combination with quoting, it allows the author of a function to capture an argument and evaluate it in a special way. Unquoting is related to evaluation, but it's a user tool: it allows the person calling the function to selectively evaluate parts of the expression that would otherwise be quoted.

### 18.5.1 With rlang

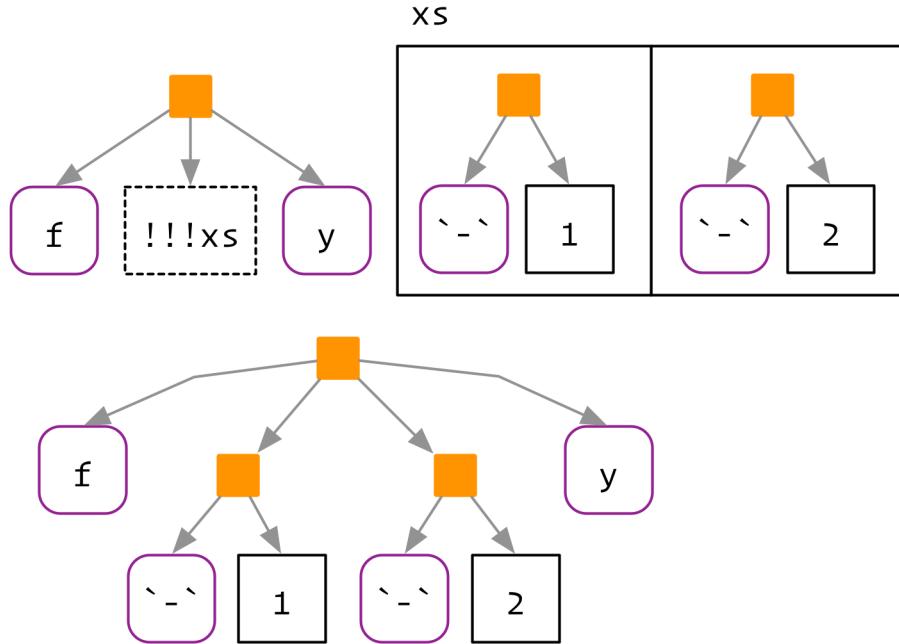
All quoting functions in rlang (`expr()`, `enexpr()`, and friends) support unquoting with `!!` (called “unquote”, and pronounced bang-bang) and `!!!` (called “unquote-splice”, and pronounced bang-bang-bang). They both replace nodes in the AST. `!!` is a one-to-one replacement. It takes a single expression and inlines the AST at the location of the `!!`.

```
x <- expr(a + b + c)
expr(f(!!x, y))
#> f(a + b + c, y)
```



`!!!` is a one-to-many replacement. It takes a list of expressions and inserts them at the location of the `!!!`:

```
x <- exprs(1, 2, 3, y = 10)
expr(f(!!!x, z = z))
#> f(1, 2, 3, y = 10, z = z)
```



### 18.5.2 The polite fiction of !!

So far we have acted as if `!!` and `!!!` are regular prefix operators like `+`, `-`, and `!`. They're not. Instead, from R's perspective, `!!` and `!!!` are simply the repeated application of `:`:

```
!!TRUE
#> [1] TRUE
!!!TRUE
#> [1] FALSE
```

`!!` and `!!!` have special behaviour inside all quoting functions powered by rlang, and the unquoting operators are given precedence similar to `+` and `-`, not `!`. We do this because the operator precedence for `!` is surprisingly low: it has lower precedence than that of the binary algebraic and logical operators. Most of the time this doesn't matter as it is unusual to mix `!` and binary operators (e.g. you typically would not write `!x + y` or `!x > y`). However, expressions like `!!x + !!!y` are not uncommon when unquoting, and requiring explicit parentheses, `(!!x) + (!!y)`, feels onerous. For this reason, rlang manipulates the AST to give the unquoting operators a higher, more natural, precedence.

You might wonder why rlang does not use a regular function call. Indeed, early versions of rlang provided `UQ()` and `UQS()` as alternatives to `!!` and `!!!`. However, these looked like regular function calls, rather than special syntactic operators, and evoked a misleading mental model, which made them harder to use correctly. In particular, function calls only happen (lazily) at evaluation time; unquoting always happens at quotation time. We adopted `!!` and `!!!` as the best compromise: they are strong visual symbols, don't look like existing syntax, and take over a rarely used piece of syntax. (And if for some reason you do need to doubly negate a value in a quasiquoting function, you can just add parentheses `!(!x)`.)

The biggest downside<sup>2</sup> to using a fake operator is that you might get silent errors when misusing `!!` outside of quasiquoting functions. Most of the time this is not an issue because `!!` is typically used to unquote

<sup>2</sup>Prior to R 3.5.1, there was another major downside: the R deparser was treating `!!x` as `!(!x)`. This is why in old versions of R you might see extra parentheses when printing tidyeval functions at the console. The good news is that these parentheses are not real and can be safely ignored most of the time. The bad news is that they will become real if you reparse that printed output to R code. These roundtripped functions will not work as expected since `!(!x)` does not unquote anything.

expressions or quosures. Since expressions are not supported by the negation operator, you will get an argument type error in this case:

```
x <- quote(variable)
!!x
#> Error in !x:
#> invalid argument type
```

However be extra careful when unquoting numeric values that can be negated silently:

```
x <- 100
with(mtcars, cyl + !!x)
#> [1] 7 7 5 7 9 7 9 5 5 7 7 9 9 9 9 9 5 5 5 5 9 9 9 9 5 5 5 5 9 7 9 5
```

Instead of adding the value of `x` to `cyl` as intended, we have in fact added the double negation of `x`:

```
!x
#> [1] FALSE
!!x
#> [1] TRUE
```

### 18.5.3 With base R

Base R has one function that implements quasiquotation: `bquote()`. It uses `.()` for unquoting:

```
xyz <- bquote((x + y + z))
bquote(-.(xyz) / 2)
#> -(x + y + z)/2
```

`bquote()` is a neat function, but is not used by any other function in base R. Instead functions that quote an argument use some other technique to allow indirect specification. There are four basic forms seen in base R:

- A pair of quoting and non-quoting functions. For example, `$` has two arguments, and the second argument is quoted. This is easier to see if you write in prefix form: `mtcars$cyl` is equivalent to ``$`(mtcars, cyl). If you want to refer to a variable indirectly, you use [, as it takes the name of a variable as a string.`

```
x <- list(var = 1, y = 2)
var <- "y"

x$var
#> [1] 1
x[[var]]
#> [1] 2
```

`<-/assign()` and `::/getExportedValue()` work similarly.

- A pair of quoting and non-quoting arguments. For example, `data()`, `rm()`, and `save()` allow you to provide bare variable names in `...`, or a character vector of variable names in `list`:

```
x <- 1
rm(x)

y <- 2
vars <- c("y", "vars")
rm(list = vars)
```

- An argument that controls whether a different argument is quoting or non-quoting. For example, in `library()`, the `character.only` argument controls the quoting behaviour of the first argument, `package`:

```
library(MASS)

pkg <- "MASS"
library(pkg, character.only = TRUE)
```

`demo()`, `detach()`, `example()`, and `require()` work similarly.

- Quoting if evaluation fails. For example, the first argument to `help()` is non-quoting if it evaluates to a string; if evaluation fails, the first argument is quoted.

```
Shows help for var
help(var)

var <- "mean"
Shows help for mean
help(var)

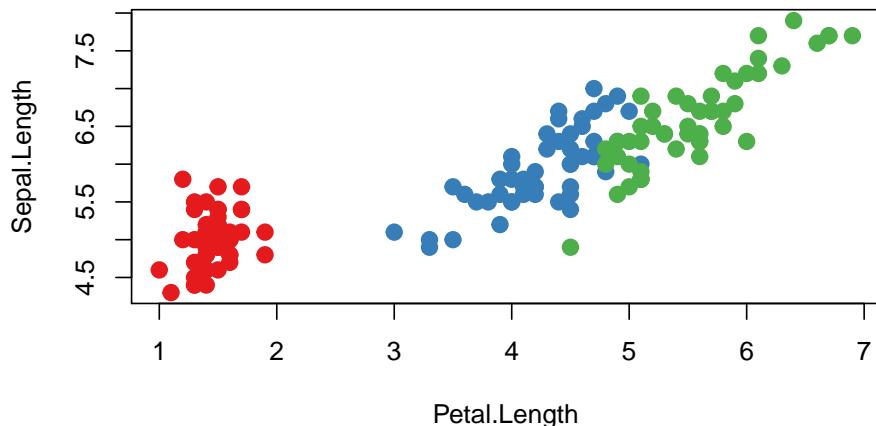
var <- 10
Shows help for var
help(var)
```

`ls()`, `page()`, and `match.fun()` work similarly.

Some quoting functions, like `subset()`, `transform()`, and `with()`, don't have a non-quoting form. This is because they are seen as wrappers around `[` and `[<-` that are only suitable for interactive use.

Another important class of quoting functions are the base modelling and plotting functions, which quote some of their arguments, and follow that so-called standard non-standard evaluation rules: <http://developer.r-project.org/nonstandard-eval.pdf>. For example, `lm()` quotes the `weight` and `subset` arguments, and when used with a formula argument, the plotting function quotes the aesthetic arguments (`col`, `cex`, etc):

```
palette(RColorBrewer::brewer.pal(3, "Set1"))
plot(Sepal.Length ~ Petal.Length, data = iris, col = Species, pch = 20, cex = 2)
```



In the next chapter, you'll learn how to simulate unquoting for these functions using tools from `rlang`.

### 18.5.4 Non-standard ASTs

Before we continue on to the case studies, we need to discuss a couple of technical issues. You might want to skip these sections on your first read through.

With unquoting, it is easy to create non-standard ASTs, i.e. ASTs that contain components that are not constants, symbols, or calls. (It is also possible to create non-standard ASTs by directly manipulating the underlying objects, but it's harder to do so accidentally.) These are valid, and occasionally useful, but their correct use is beyond the scope of this book. However, it's important to learn about them because they can be decompiled, and hence printed, in misleading ways.

For example, if you inline more complex objects, their attributes are not printed. This can lead to confusing output:

```
x1 <- expr(class (!!data.frame(x = 10)))
x1
#> class(list(x = 10))
lobstr::ast (!!x1)
#> class
#> <inline data.frame>
eval(x1)
#> [1] "data.frame"
```

In other cases, R will print parentheses that do not exist in the AST:

```
y2 <- expr(2 + 3)
x2 <- expr(1 + !!y2)
x2
#> 1 + (2 + 3)
lobstr::ast (!!x2)
#> `+`
#> 1
#> `+`
#> 2
#> 3
```

And finally, R will display integer sequences as if they were generated with `:`.

```
x3 <- expr(f (!!c(1L, 2L, 3L, 4L, 5L)))
x3
#> f(1:5)
lobstr::ast (!!x3)
#> f
#> <inline integer>
```

In general, if you're ever confused about what is actually in an AST, display the object with `lobstr::ast()`!

### 18.5.5 Missing arguments

Occasionally it is useful to unquote a missing argument, but the naive approach doesn't work:

```
arg <- missing_arg()
expr(foo (!!arg, !!arg))
#> Error in enexpr(expr):
#> argument "arg" is missing, with no default
```

You can either wrap in a list and use unquote-splice, or use the `maybe_missing()` helper:

```
args <- list(missing_arg(), missing_arg())
expr(foo(!!!args))
#> foo(,)

expr(foo(!!maybe_missing(arg), !!maybe_missing(arg)))
#> foo(,)
```

### 18.5.6 Exercises

- Given the following components:

```
xy <- expr(x + y)
xz <- expr(x + z)
yz <- expr(y + z)
abc <- exprs(a, b, c)
```

Use quasiquotation to construct the following calls:

```
(x + y) / (y + z)
-(x + z) ^ (y + z)
(x + y) + (y + z) - (x + y)
atan2(x + y, y + z)
sum(x + y, x + y, y + z)
sum(a, b, c)
mean(c(a, b, c), na.rm = TRUE)
foo(a = x + y, b = y + z)
```

- Explain why both `!0 + !0` and `!1 + !1` return `FALSE` while `!0 + !1` returns `TRUE`.
- Base functions `match.fun()`, `page()`, and `ls()` all try to automatically determine whether you want standard or non-standard evaluation. Each uses a different approach. Figure out the essence of each approach by reading the source code, then compare and contrast the techniques.
- The following two calls print the same, but are actually different:

```
(a <- expr(mean(1:10)))
#> mean(1:10)
(b <- expr(mean(!!(1:10))))
#> mean(1:10)
identical(a, b)
#> [1] FALSE
```

What's the difference? Which one is more natural?

## 18.6 Case studies

To make these ideas concrete, this section contains a few smaller case studies that show how quasiquotation can be used to solve real problems. Some of the case studies also use purrr: I find the combination of quasiquotation and functional programming to be particularly elegant.

```
library(purrr)
library(dplyr)
```

### 18.6.1 Map-reduce to generate code

Quasiquotation gives us powerful tools for generating code, particularly when combined with `purrr::map()` and `purr::reduce()`. For example, assume you have a linear model specified by the following coefficients:

```
intercept <- 10
coefs <- c(x1 = 5, x2 = -4)
```

And you want to convert it into an expression like  $10 + (5 * x1) + (-4 * x2)$ . The first thing we need to do is turn the character names vector into a list of symbols. `rlang::syms()` is designed precisely for this case:

```
coef_sym <- syms(names(coefs))
coef_sym
#> [[1]]
#> x1
#>
#> [[2]]
#> x2
```

Next we need to combine each variable name with its coefficient. We can do this by combining `expr()` with `map2()`:

```
summands <- map2(coef_sym, coefs, ~ expr (!!x * !!y))
summands
#> [[1]]
#> (x1 * 5)
#>
#> [[2]]
#> (x2 * -4)
```

In this case, the intercept is also a part of the sum, although it doesn't involve a multiplication. We can just add it to the start of the `summands` vector:

```
summands <- c(intercept, summands)
summands
#> [[1]]
#> [1] 10
#>
#> [[2]]
#> (x1 * 5)
#>
#> [[3]]
#> (x2 * -4)
```

Finally, we need to reduce the individual terms in to a single sum by adding the pieces together:

```
eq <- reduce(summands, ~ expr (!!x + !!y))
eq
#> 10 + (x1 * 5) + (x2 * -4)
```

This map-reduce pattern is an elegant way to solve many code generation problems.

Once you have this expression, you could evaluate it with new data, or turn it into a function:

```
df <- data.frame(x1 = runif(5), x2 = runif(5))
eval(eq, df)
#> [1] 13.59 9.26 9.92 10.05 8.11
```

```
args <- map(coefs, ~ missing_arg())
new_function(args, expr({!eq}))
#> function (x1, x2)
#> {
#> 10 + (x1 * 5) + (x2 * -4)
#> }
```

## 18.6.2 Partition

Imagine that you want to extend `dplyr::select()` to return two data frames: one with the variables you selected, and one with the variables that remain. (This problem was inspired by <https://stackoverflow.com/questions/46828296/>.) There are plenty of ways to attack this problem, but one way is to take advantage of `select()`'s ability to negate column selection expression in order to remove those columns.

We can capture the inputs with quasiquotation, then invert each selection call by negating it. We start by practicing interactively with a list of variables created with `exprs()`:

```
vars <- exprs(x, y, c(a, b), starts_with("x"))
map(vars, ~ expr(-!!.x))
#> [[1]]
#> -x
#>
#> [[2]]
#> -y
#>
#> [[3]]
#> -c(a, b)
#>
#> [[4]]
#> -starts_with("x")
```

Then turn it into a function:

```
partition_cols <- function(.data, ...) {
 included <- enexprs(...)
 excluded <- map(included, ~ expr(-!!.x))

 list(
 incl = select(.data, !!!included),
 excl = select(.data, !!!excluded)
)
}

df <- data.frame(x1 = 1, x2 = 3, y = "a", z = "b")
partition_cols(df, starts_with("x"))
#> $incl
#> x1 x2
#> 1 1 3
#>
#> $excl
#> y z
#> 1 a b
```

Note the name of the first argument: `.data`. This is a standard convention through the tidyverse because you don't need to explicitly name this argument (because it's always used), and it avoids potential clashes

with argument names in ....

### 18.6.3 Slicing an array

One occasionally useful tool that's missing from base R is the ability to extract a slice of an array given a dimension and an index. For example, we'd like to write `slice(x, 2, 1)` to extract the first slice along the second dimension, which you can write as `x[, 1, ]`.

We'll need to generate a call with multiple missing arguments. Fortunately is easy with `rep()` and `missing_arg()`. Once we have those arguments, we can unquote-splice them into a call:

```
indices <- rep(list(missing_arg()), 3)
expr(x[!!!indices])
#> x[, ,]
```

We then wrap this into a function, using subset-assignment to insert the index in the desired position:

```
slice <- function(x, along, index) {
 stopifnot(length(index) == 1)

 nd <- length(dim(x))
 indices <- rep(list(missing_arg()), nd)
 indices[along] <- index

 expr(x[!!!indices])
}

x <- array(sample(30), c(5, 2, 3))
slice(x, 1, 3)
#> x[3, ,]
slice(x, 2, 2)
#> x[, 2,]
slice(x, 3, 1)
#> x[, , 1]
```

A real `slice()` would evaluate the generated call, but here I think it's more illuminating to see the code that's generated, as that's the hard part of the challenge.

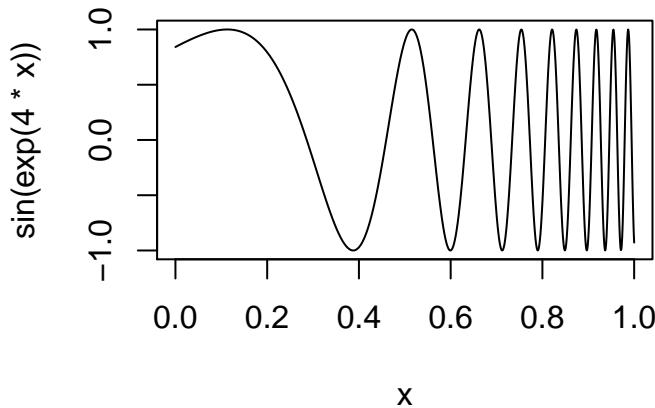
### 18.6.4 Creating functions

Another powerful function to use in combination with unquoting is `rlang::new_function()`: it allows us to create a function by supplying the arguments, the body, and (optionally) the environment:

```
new_function(
 exprs(x = , y =),
 expr({x + y})
)
#> function (x, y)
#> {
#> x + y
#> }
```

One application is to create functions that work like `graphics::curve()`. `curve()` allows you to plot a mathematical expression, without creating a function:

```
curve(sin(exp(4 * x)), n = 1000)
```



Here `x` is a pronoun. As with `.` in pipelines and `.x` and `.y` in purrr functionals, `x` doesn't represent a single concrete value, but is instead a placeholder that varies over the range of the plot. Functions, like `curve()`, that use an expression containing a pronoun are known as **anaphoric** functions<sup>3</sup>.

One way to implement `curve()` is to turn the expression into a function with a single argument, then call that function:

```
curve2 <- function(expr, xlim = c(0, 1), n = 100) {
 expr <- enexpr(expr)
 f <- new_function(exprs(x =), expr)

 x <- seq(xlim[1], xlim[2], length = n)
 y <- f(x)

 plot(x, y, type = "l", ylab = expr_text(expr))
}
curve2(sin(exp(4 * x)), n = 1000)
```

Another use for `new_function()` is as an alternative to simple function factories and function operators. The primary advantage is that the generated functions have readable source code:

```
negate1 <- function(f) {
 force(f)
 function(...) !f(...)
}

negate1(is.null)
#> function(...) !f(...)
#> <environment: 0x612d248>

negate2 <- function(f) {
 f <- enexpr(f)
 new_function(exprs(... =), expr(!(!f)(...)), caller_env())
}

negate2(is.null)
#> function ...
#> !is.null(...)
```

Note that this is often useful if the higher order function have arguments that are expressions: inlining more

<sup>3</sup>Anaphoric comes from the linguistics term “anaphora”, an expression that is context dependent. Anaphoric functions are found in Arc (<http://www.arcfn.com/doc/anaphoric.html>) (a LISP like language), Perl ([http://www.perlmonks.org/index.pl?node\\_id=666047](http://www.perlmonks.org/index.pl?node_id=666047)), and Clojure (<http://amalloy.hubpages.com/hub/Unhygienic-anaphoric-Clojure-macros-for-fun-and-profit>).

complex objects into the AST can yield confusing source code.

### 18.6.5 Exercises

1. Implement `arrange_desc()`, a variant of `dplyr::arrange()` that sorts in descending order by default.
2. Implement `filter_or()`, a variant of `dplyr::filter()` that combines multiple arguments using `|` instead of `&`.
3. Implement `partition_rows()` which, like `partition_cols()`, returns two data frames, one containing the selected rows, and the other containing the rows that weren't selected.
4. Add error handling to `slice()`. Give clear error messages if either `along` or `index` have invalid values (i.e. not numeric, not length 1, too small, or too big).
5. Re-implement the Box-Cox transform defined below using unquoting and `new_function()`:

```
bc <- function(lambda) {
 if (lambda == 0) {
 function(x) log(x)
 } else {
 function(x) (x ^ lambda - 1) / lambda
 }
}
```

6. Re-implement the simple `compose()` defined below using quasiquotation and `new_function()`:

```
compose <- function(f, g) {
 function(...) f(g(...))
}
```

## 18.7 Dot-dot-dot (...)

Quasiquotation ensures that every quoted argument has an escape hatch that allows the user to unquote, or evaluate, selected components, if needed. A similar and related need arises with functions that take arbitrary additional arguments with `....`. Take the following two motivating problems:

- What do you do if the elements you want to put in `....` are already stored in a list? For example, imagine you have a list of data frames that you want to `rbind()` together:

```
dfs <- list(
 a = data.frame(x = 1, y = 2),
 b = data.frame(x = 3, y = 4)
)
```

You could solve this specific case with `rbind(dfs$a, df$b)`, but how do you generalise that solution to a list of arbitrary length?

- What do you do if you want to supply the argument name indirectly? For example, imagine you want to create a single column data frame where the name of the column is specified in a variable:

```
var <- "x"
val <- c(4, 3, 9)
```

In this case, you could create a data frame and then change names (i.e. `setNames(data.frame(val), var)`), but this feels inelegant. How can we do better?

### 18.7.1 `do.call()`

Base R provides a swiss-army knife to solve these problems: `do.call()`. `do.call()` has two main arguments. The first argument, `what`, gives a function to call. The second argument, `args`, is a list of arguments to pass to that function, and so `do.call("f", list(x, y, z))` is equivalent to `f(x, y, z)`.

- `do.call()` gives a straightforward solution to `rbind()`ing together many data frames:

```
do.call("rbind", dfs)
#> x y
#> a 1 2
#> b 3 4
```

- With a little more work, we can use `do.call()` to solve the second problem. We first create a list of arguments, then name that, then use `do.call()`:

```
args <- list(val)
names(args) <- var

do.call("data.frame", args)
#> x
#> 1 4
#> 2 3
#> 3 9
```

### 18.7.2 The tidyverse approach

The tidyverse solves these problems in a different way to base R, by drawing parallel to quasiquotation:

- Row-binding multiple data frames is like unquote-splicing: we want to inline individual elements of the list into the call:

```
dplyr::bind_rows(!!!dfs)
#> x y
#> 1 1 2
#> 2 3 4
```

When used in this context, the behaviour of `!!!` is known as splicing in Ruby, Go, PHP, and Julia. It is closely related to `*args` (star-args) and `**kwargs` (star-star-kwargs) in Python, which are sometimes called argument unpacking.

- The second problem is like unquoting on the LHS of `=`: rather than interpreting `var` literally, we want to use the value stored in the variable called `var`:

```
tibble::tibble(!!var := val)
#> # A tibble: 3 x 1
#> x
#> <dbl>
#> 1 4
#> 2 3
#> 3 9
```

Note the use of `:=` (pronounced colon-equals) rather than `=`. Unfortunately we need this new operation because R's grammar does not allow expressions as argument names:

```
tibble::tibble(!!var = value)
#> Error: unexpected '=' in "tibble::tibble(!!var ="
```

`:=` is like a vestigial organ: it's recognised by R's parser, but it doesn't have any code associated with it. It looks like an `=` but allows expressions on either side, making it a more flexible alternative to `=`. It is used in `data.table` for similar reasons.

### 18.7.3 `list2()`

Both `dplyr::bind_rows()` and `tibble::tibble()` are powered by `rlang::list2(...)`. This function is very similar to `list(...)`, but it understands `!!!` and `!!`. If you want to take advantage of this behaviour in your own function, all you need to do is use `list2()` in your own code. For example, imagine you want to make a version of `structure()` that understands `!!!` and `!!`. We'll call it `set_attr()`:

```
set_attr <- function(.x, ...) {
 attr <- rlang::list2(...)
 attributes(.x) <- attr
 .x
}

attrs <- list(x = 1, y = 2)
attr_name <- "z"

1:10 %>%
 set_attr(w = 0, !!! attrs, !!attr_name := 3) %>%
 str()
#> int [1:10] 1 2 3 4 5 6 7 8 9 10
#> - attr(*, "w")= num 0
#> - attr(*, "x")= num 1
#> - attr(*, "y")= num 2
#> - attr(*, "z")= num 3
```

(`rlang` also provides a `set_attr()` function with a few extra conveniences, but the essence is the same.)

Note that we call the first argument `.x`: whenever you use `...` to take arbitrary data, it's good practice to give the other argument names a `.` prefix. This eliminates any ambiguity about who owns the argument, and in this case makes it possible to set the `x` attribute.

`list2()` provides one other handy feature: by default it will ignore any empty arguments at the end. This is useful in functions like `tibble::tibble()` because it means that you can easily change the order of variables without worrying about the final comma:

```
Can easily move x to first entry:
tibble::tibble(
 y = 1:5,
 z = 3:-1,
 x = 5:1,
)

Need to remove comma from z and add comma to x
data.frame(
 y = 1:5,
 z = 3:-1,
 x = 5:1
)
```

As well as `list2()`, `rlang` also provides `lg1()`, `int()`, `dbl()`, and `chr()` which create atomic vectors in the same way.

### 18.7.4 Application: `invoke()` and `lang()`

One useful application of `list2()` is `invoke()`:

```
invoke <- function(.f, ...) {
 do.call(.f, list2(...), envir = parent.frame())
}
```

(At time of writing, both `purrr::invoke()` and `rlang::invoke()` have somewhat different definitions because they were written before we understood how quasiquotation syntax and `...` intersected.)

As a wrapper around `do.call()`, `invoke()` gives powerful ways to call functions with arguments supplied directly (in `...`) or indirectly (in a list):

```
invoke("mean", x = 1:10, na.rm = TRUE)

Equivalent to
x <- list(x = 1:10, na.rm = TRUE)
invoke("mean", !!!x)
```

It also allows us to specify argument names indirectly:

```
arg_name <- "na.rm"
arg_val <- TRUE
invoke("mean", 1:10, !!arg_name := arg_val)
```

Closely related to `invoke()` is `rlang::call2()`. It constructs a call from its components:

```
call2("mean", 1:10, !!arg_name := arg_val)
#> mean(1:10, na.rm = TRUE)
```

The chief advantage of `call2()` over `expr()` is that it can use `:=`.

### 18.7.5 Other approaches

Apart from `rlang::list2()` there are several other techniques used to overcome the motivating challenges described above. One technique is to take `...` and a single unnamed argument that is a list, making `f(list(x, y, z))` equivalent to `f(x, y, z)`. The implementation looks something like this:

```
f <- function(...) {
 dots <- list(...)
 if (length(dots) == 1 && is.list(dots[[1]])) {
 dots <- dots[[1]]
 }

 # Do something
 ...
}
```

Base functions that use this technique include `interaction()`, `expand.grid()`, `options()`, and `par()`. Since these functions take either a list or `...`, but not both, they are slightly less flexible than functions powered by `list2()`.

Another related technique is used in the `RCurl::getURL()` function written by Duncan Temple Lang. `getURL()` takes both `...` and `.opts` which are concatenated together. This is useful when writing functions to call web APIs because you often have some options that need to be passed to every request. You put these in a common list and pass to `.opts`, saving `...` for the options unique for a given call.

I found this technique particularly compelling so you can see it used throughout the tidyverse. Now, however, `rlang::list2()` dots solves more problems, more elegantly, by using the ideas from tidy eval. The tidyverse is slowly migrating to `list2()` style for all functions that take ....

### 18.7.6 Exercises

1. Carefully read the source code for `interaction()`, `expand.grid()`, and `par()`. Compare and contrast the techniques they use for switching between dots and list behaviour.
2. Explain the problem with this definition of `set_attr()`

```
set_attr <- function(x, ...) {
 attr <- rlang::list2(...)
 attributes(x) <- attr
 x
}
set_attr(1:10, x = 10)
#> Error in attributes(x) <- attr:
#> attributes must be named
```

# Chapter 19

# Evaluation

## 19.1 Introduction

The user-facing opposite of quotation is unquotation: it gives the *user* the ability to selectively evaluate parts of an otherwise quoted argument. The developer-facing complement of quotation is evaluation: this gives the *developer* the ability to evaluate quoted expressions in custom environments to achieve specific goals.

This chapter begins with a discussion of evaluation in its purest form with `rlang::eval_bare()` which evaluates an expression in given environment. We'll then see how these ideas are used to implement a handful of base R functions, and then learn about the similar `base::eval()`.

The meat of the chapter focusses on extensions needed to implement evaluation robustly. There are two big new ideas:

- We need a new data structure that captures both the expression **and** the environment associated with each function argument. We call this data structure a **quosure**.
- `base::eval()` supports evaluating an expression in the context of a data frame and an environment. We formalise this idea by calling it **data mask** and to resolve the ambiguity it creates, introduce the idea of data pronouns.

Together, quasiquotation, quosures, data masks, and pronouns form what we call **tidy evaluation**, or tidy eval for short. Tidy eval provides a principled approach to NSE that makes it possible to use such functions both interactively and embedded with other functions. We'll finish off the chapter showing the basic pattern you use to wrap quasiquoting functions, and how you can adapt that pattern to base R NSE functions.

## Outline

### Prerequisites

Environments play a very important big role in evaluation, so make sure you're familiar with the basics in Environments.

```
library(rlang)
```

## 19.2 Evaluation basics

In the previous chapter, we briefly mentioned `eval()`. Here, however, we're going to start with `rlang::eval_bare()` which is the purest evocation of the idea of evaluation. The first argument, `expr` is an expression to evaluate. This will usually be either a symbol or expression:

```
x <- 10
eval_bare(expr(x))
#> [1] 10

y <- 2
eval_bare(expr(x + y))
#> [1] 12
```

Everything else yields itself when evaluated:

```
eval_bare(10)
#> [1] 10
```

The second argument, `env`, gives the environment in which the expression should be evaluated, i.e. where should the values of `x`, `y`, and `+` be looked for? By default, this is the current environment, i.e. the calling environment of `eval_bare()`, but you can override it if you want:

```
eval_bare(expr(x + y), env(x = 1000))
#> [1] 1002
```

Because R looks up functions in the same way as variables, we can also override the meaning of functions. This is a very useful technique if you want to translate R code into something else, as you'll learn about in the next chapter.

```
eval_bare(
 expr(x + y),
 env(`+` = function(x, y) paste0(x, " + ", y))
)
#> [1] "10 + 2"
```

Note that the first argument to `eval_bare()` (and to `base::eval()`) is evaluated, not quoted. This can lead to confusing results if you forget to quote the input:

```
eval_bare(x + y)
#> [1] 12
eval_bare(x + y, env(x = 1000))
#> [1] 12
```

Now that you've seen the basics, let's explore some applications. We'll focus primarily on base R functions that you might have used before; now you can learn how they work. To focus on the underlying principles, we'll extract out their essence, and rewrite to use `rlang` functions. Once you've seen some applications, we'll circle back and talk about more about `base::eval()`.

### 19.2.1 Application: `local()`

Sometimes you want to perform a chunk of calculation that creates a bunch of intermediate variables. The intermediate variables have no long-term use and could be quite large, so you'd rather not keep them around. One approach is to clean up after yourself using `rm()`; another approach is to wrap the code in a function, and just call it once. A more elegant approach is to use `local()`:

```
Clean up variables created earlier
rm(x, y)

foo <- local({
 x <- 10
 y <- 200
 x + y
})

foo
#> [1] 210
x
#> Error in eval(expr, envir, enclos):
#> object 'x' not found
y
#> Error in eval(expr, envir, enclos):
#> object 'y' not found
```

The essence of `local()` is quite simple. We capture the input expression, and create a new environment in which to evaluate it. This inherits from the caller environment so it can access the current lexical scope, but any intermediate variables will be GC'd once the function has returned.

```
local2 <- function(expr) {
 env <- child_env(caller_env())
 eval_bare(enexpr(expr), env)
}

foo <- local2({
 x <- 10
 y <- 200
 x + y
})

foo
#> [1] 210
x
#> Error in eval(expr, envir, enclos):
#> object 'x' not found
y
#> Error in eval(expr, envir, enclos):
#> object 'y' not found
```

Understanding how `base::local()` works is harder, as it uses `eval()` and `substitute()` together in rather complicated ways. Figuring out exactly what's going on is good practice if you really want to understand the subtleties of `substitute()` and the base `eval()` functions, so is included in the exercises below.

### 19.2.2 Application: `source()`

We can create a simple version of `source()` by combining `parse_expr()` and `eval_bare()`. We read in the file from disk, use `parse_expr()` to parse the string into a list of expressions, and then use `eval_bare()` to evaluate each component in turn. This version evaluates the code in the caller environment, and invisibly returns the result of the last expression in the file like `source()`.

```
source2 <- function(path, env = caller_env()) {
 file <- paste(readLines(path, warn = FALSE), collapse = "\n")
 exprs <- parse_exprs(file)

 res <- NULL
 for (i in seq_along(exprs)) {
 res <- eval_bare(exprs[[i]], env)
 }

 invisible(res)
}
```

The real `source()` is considerably more complicated because it can echo input and output, and has many other settings that control its behaviour.

### 19.2.3 Gotcha: `function()`

There's one small gotcha that you should be aware of if you're using `eval_bare()` and `expr()` to generate functions:

```
x <- 10
y <- 20
f <- eval_bare(expr(function(x, y) !!x + !!y))
f
#> function(x, y) !!x + !!y
```

This function doesn't look like it will work, but it does:

```
f()
#> [1] 30
```

This is because, if available, functions print their `srcref`. The source reference is a base R feature that doesn't know about quasiquotation. To work around this problem, I recommend using `new_function()` as shown in the previous chapter. Alternatively, you can remove the `srcref` attribute:

```
attr(f, "srcref") <- NULL
f
#> function (x, y)
#> 10 + 20
```

### 19.2.4 Advanced: environments vs. frames

Frame look up from environment

```
f <- function() g()
g <- function() h()
h <- function() eval(expr(lobstr::cst()), caller_env(2))
f()
#> x
#> \-f()
#> +-g()
#> / \-h()
#> / \-eval(expr(lobstr::cst()), caller_env(2))
#> / \-eval(expr(lobstr::cst()), caller_env(2))
#> / \-lobstr::cst()
```

### 19.2.5 Base R

The base function equivalent to `eval_bare()` is the two-argument form of `eval()`: `eval(expr, envir)`:

```
eval(expr(x + y), env(x = 1000, y = 1))
#> [1] 1001
```

The final argument, `enclos` provides support for data masks, which you'll learn about in tidy evaluation.

`eval()` is paired with two helper functions:

- `evalq(x, env)` quotes its first argument, and is hence a shortcut for `eval(quote(x), env)`.
- `eval.parent(expr, n)` is shortcut for `eval(expr, env = parent.frame(n))`.

`base::eval()` has special behaviour for expression **objects**, evaluating each component in turn. This makes for a very compact implementation of `source2()` because `base::parse()` also returns an expression object:

```
source3 <- function(file, env = parent.frame()) {
 lines <- parse(file)
 res <- eval(lines, envir = env)
 invisible(res)
}
```

While `source3()` is considerably more concise than `source2()`, this one use case is the strongest argument for expression objects, and overall we don't believe this one benefit outweighs the cost of introducing a new data structure. That's why this book has reneged expression objects to a secondary role.

### 19.2.6 Exercises

1. Carefully read the documentation for `source()`. What environment does it use by default? What if you supply `local = TRUE`? How do you provide a custom argument?

2. Predict the results of the following lines of code:

```
eval(quote(eval(quote(eval(quote(2 + 2))))))
eval(eval(quote(eval(quote(eval(quote(2 + 2)))))))
quote(eval(quote(eval(quote(eval(quote(2 + 2)))))))
```

3. Write an equivalent to `get()` using `sym()` and `eval_bare()`. Write an equivalent to `assign()` using `sym()`, `expr()`, and `eval_bare()`. (Don't worry about the multiple ways of choosing an environment that `get()` and `assign()` support; assume that the user supplies it explicitly.)

```
name is a string
get2 <- function(name, env) {}
assign2 <- function(name, value, env) {}
```

4. Modify `source2()` so it returns the result of *every* expression, not just the last one. Can you eliminate the for loop?

5. The code generated by `source2()` lacks source references. Read the source code for `sys.source()` and the help for `srcfilecopy()`, then modify `source2()` to preserve source references. You can test your code by sourcing a function that contains a comment. If successful, when you look at the function, you'll see the comment and not just the source code.

6. We can make `base::local()` slightly easier to understand by spreading out over multiple lines:

```
local3 <- function(expr, envir = new.env()) {
 call <- substitute(eval(quote(expr)), envir)
```

```

 eval(call, envir = parent.frame())
}

```

Explain how `local()` works in words. (Hint: you might want to `print(call)` to help understand what `substitute()` is doing, and read the documentation to remind yourself what environment `new.env()` will inherit from.)

## 19.3 Quosures

The simplest form of evaluation combines an expression and an environment. This coupling is so important that we need a data structure that can hold both pieces: we need a **quosure**, a portmanteau of quoting and closure. In this section, you'll learn about why quosures are important, how to create and manipulate them, and a little about how they are implemented. We'll finish off by discussing the few cases where you should work with expressions rather than quosures.

### 19.3.1 Motivation

Quosures are important when the distance between capturing and evaluating an expression grows. Take this simple, if somewhat contrived example:

```

foo <- function(x) {
 y <- 100
 x <- enexpr(x)

 eval_bare(x)
}

```

It appears to work for simple cases:

```

z <- 100
foo(z * 2)
#> [1] 200

```

But if our expression uses `y` it will find the wrong one:

```

y <- 10
foo(y * 2)
#> [1] 200

```

We could fix this by manually specifying the correct environment:

```

foo2 <- function(x) {
 y <- 100
 x <- enexpr(x)

 eval_bare(x, caller_env())
}

y <- 10
foo2(y * 2)
#> [1] 20

```

That works for this simple case, but does not generalise well. Take this more complicated example that uses `...`. Each argument to `f()` needs to be evaluated in a different environment:

```
f <- function(...) {
 x <- 1
 g(..., x = x)
}

g <- function(...) {
 x <- 2
 h(..., x = x)
}

h <- function(...) {
 exprs <- enexprs(...)
 purrr::map_dbl(exprs, eval_bare, env = caller_env())
}

x <- 0
f(x = x)
#> x x x
#> 2 2 2
```

We can overcome this problem by using two new tools that you'll learn about shortly: we capture with `enquo()` instead of `enexprs()`, and evaluate with `eval_tidy()` instead of `eval_bare()`:

```
h <- function(...) {
 exprs <- enquo(...)
 purrr::map_dbl(exprs, eval_tidy)
}

x <- 0
f(x = x)
#> x x x
#> 0 1 2
```

This ensures that each expression is evaluated in the correct environment.

### 19.3.2 Creating and manipulating

Each of the `expr()` functions that you learned about in the previous chapter has an equivalent `quo()` function that creates a quosure:

- Use `quo()` and `quos()` to capture your expressions.

```
quo(x + y + z)
#> <quosure>
#> expr: ~x + y + z
#> env: global
quos(x + 1, y + 2)
#> [[1]]
#> <quosure>
#> expr: ~x + 1
#> env: global
#>
#> [[2]]
#> <quosure>
```

```
#> expr: ^y + 2
#> env: global
```

- Use `enquo()` and `enquos()` to capture user-supplied expressions.

```
foo <- function(x) enquo(x)
foo(a + b)
#> <quosure>
#> expr: ^a + b
#> env: global
```

Note how quosures are printed: each quosure starts with `^`. This is a signal that you’re looking at something special, and is useful if you unquote a quosure inside another quosure. In the console, each quosure gets a different colour to help remind you that it has a different environment attached to it.

```
q2 <- quo(x + !!x)
q2
#> <quosure>
#> expr: ^x + 0
#> env: global
```

Finally, you can use `new_quosure()` to create a quosure from its components: an expression and an environment.

```
x <- new_quosure(expr(x + y), env(x = 1, y = 10))
x
#> <quosure>
#> expr: ^x + y
#> env: 0x5bb1f20
```

If you need to turn a quosure into text for output to the console you can use `quo_name()`, `quo_label()`, or `quo_text()`. `quo_name()` and `quo_label()` are guaranteed to be short; `quo_expr()` may span multiple lines.

```
y <- quo(long_function_name(
 argument_1 = long_argument_value,
 argument_2 = long_argument_value,
 argument_3 = long_argument_value,
 argument_4 = long_argument_value
))
quo_name(y) # e.g. for data frames
#> [1] "long_function_name(...)"
quo_label(y) # e.g. for error messages
#> [1] "`long_function_name(...)`"
quo_text(y) # for longer messages
#> [1] "long_function_name(argument_1 = long_argument_value, argument_2 = long_argument_value, \n argument_3 = long_argument_value, argument_4 = long_argument_value)"
```

### 19.3.3 Evaluating

You can evaluate a quosure with `eval_tidy()`:

```
x <- new_quosure(expr(x + y), env(x = 1, y = 10))
eval_tidy(x)
#> [1] 11
```

And you can extract its components with the `quo_get_` helpers:

```
quo_get_env(x)
#> <environment: 0x64d3a90>
quo_get_expr(x)
#> x + y
```

For this simple case, `eval_tidy()` is basically a wrapper around `eval_bare()`. In the next section, you'll learn about the `data` argument which makes `eval_tidy()` particularly powerful.

```
eval_bare(quo_get_expr(x), quo_get_env(x))
#> [1] 11
```

### 19.3.4 Implementation

Quosures rely on R's internal representation of function arguments as a special type of object called a **promise**. A promise captures the expression needed to compute the value and the environment in which to compute it. You're not normally aware of promises because the first time you access a promise its code is evaluated in its environment, yielding a value. This is what powers lazy evaluation. You cannot manipulate promises with R code. Promises are like a quantum state: any attempt to inspect them with R code will force an immediate evaluation, making the promise disappear. To work around this, rlang manipulates promises with C code, reifying them into an R object that you can work with.

There is one big difference between promises and quosures. A promise is evaluated once, when you access it for the first time. Every time you access it subsequently it will return the same value. A quosure must be evaluated explicitly, and each evaluation is independent of the previous evaluations.

```
The argument x is evaluated once, then reused
foo <- function(x_arg) {
 list(x_arg, x_arg)
}
foo(runif(3))
#> [[1]]
#> [1] 0.0808 0.8343 0.6008
#>
#> [[2]]
#> [1] 0.0808 0.8343 0.6008

The quosure x is evaluated afresh each time
x_quo <- quo(runif(3))
eval_tidy(x_quo)
#> [1] 0.1572 0.0074 0.4664
eval_tidy(x_quo)
#> [1] 0.498 0.290 0.733
```

Quosures are inspired by R's formulas, `~`, which, like quosures, capture both the expression and its environment:

```
f <- ~runif(3)
f
#> ~runif(3)

str(f)
#> Class 'formula' language ~runif(3)
#> ... - attr(*, ".Environment")=<environment: R_GlobalEnv>
```

Initial versions of rlang used formulas instead of quosures, as an attractive feature of `~` is that it provides

quoting with a single keystroke. Unfortunately, however, there is no way to add quasiquotation to `~`, so we decided to use a new function, `quo()`, instead.

### 19.3.5 When not to use quo<sub>sures</sub>

Almost all quoting functions should capture quo<sub>sures</sub> rather than expressions, and you should default to using `enquo()` and `enquos()` to capture arguments from the user. You should only use expressions if you have explicitly decided that the environment is not important. This tends to happen in three main cases:

- In code generation, such as you saw in Slicing an array.
- When you are wrapping a NSE function that doesn't use quo<sub>sures</sub>. We'll discuss this in detail in the case study at the end of the chapter.
- When you have carefully created a self-contained expression using unquoting. For example, instead of this quo<sub>sure</sub>:

```
base <- 2
quo(log(x, base = base))
#> <quosure>
#> expr: `log(x, base = base)`
#> env: global
```

You could create this self-contained expression:

```
expr(log(x, base = !!base))
#> log(x, base = 2)
```

(Assuming that `x` will be supplied in some other way)

### 19.3.6 Exercises

- Predict what evaluating each of the following quo<sub>sures</sub> will return.

```
q1 <- new_quosure(expr(x), env(x = 1))
q1
#> <quosure>
#> expr: `x
#> env: 0x5508638

q2 <- new_quosure(expr(x + !!q1), env(x = 10))
q2
#> <quosure>
#> expr: `x + (^x)
#> env: 0x5745ce8

q3 <- new_quosure(expr(x + !!q2), env(x = 100))
q3
#> <quosure>
#> expr: `x + (^x + (^x))
#> env: 0x5b08e88
```

- Write a function `enenv()` that captures the environment associated with an argument.

## 19.4 Tidy evaluation

In the previous section, you learned how to capture quosures, why they are important, and the basics of `eval_tidy()`. In this section, we'll go deep on `eval_tidy()` and talk more generally about the ideas of **tidy evaluation**. There are two big new concepts:

- A **data mask** is a data frame where the evaluated code will look first for variable definitions.
- A data mask introduces ambiguity, so to remove that ambiguity when necessary we introduce **pronouns**.

We'll explore tidy evaluation in the context of `base::subset()`, because it's a simple yet powerful function that encapsulates one of the central ideas that makes R so elegant for data analysis. Once we've seen the tidy implementation, we'll return to the base R implementation, learn how it works, and explore the limitations that make `subset()` suitable only for interactive usage.

### 19.4.1 Data masks

In the previous section, you learned that `eval_tidy()` is basically a wrapper around `eval_bare()` when evaluating a quosure. The real power of `eval_tidy()` comes with the second argument: `data`.<sup>1</sup> This lets you set up a **data mask**, where variables in the environment are potentially masked by variables in a data frame. This allows you to mingle variables from the environment and variables from a data frame:

```
x <- 10
df <- data.frame(y = 1:10)
q1 <- quo(x * y)

eval_tidy(q1, df)
#> [1] 10 20 30 40 50 60 70 80 90 100
```

The data mask is the key idea that powers base functions like `with()`, `subset()` and `transform()`, and that is used throughout tidyverse, in packages like `dplyr`.

How does this work? Unlike environments, data frames don't have parents, so we can effectively turn it into an environment using the environment of the quosure as its parent. The above code is basically equivalent to:

```
df_env <- as_environment(df, parent = quo_get_env(q1))
q2 <- quo_set_env(q1, df_env)

eval_tidy(q2)
#> [1] 10 20 30 40 50 60 70 80 90 100
```

`base::eval()` has similar functionality. If the 2nd argument is a data frame it becomes a data mask, and you provide the environment in the 3rd argument:

```
eval(quo_get_expr(q1), df, quo_get_env(q1))
#> [1] 10 20 30 40 50 60 70 80 90 100
```

### 19.4.2 Application: `subset()`

To see why the data mask is so useful, let's implement our own version of `subset()`. If you haven't used it before, `subset()` (like `dplyr::filter()`), provides a convenient way of selecting rows of a data frame using

---

<sup>1</sup>`eval_tidy()` has a `env` argument, but you only need this if you pass an expression to the first argument.

an expression that is evaluated in the context of the data frame. It allows you to subset without repeatedly referring to the name of the data frame:

```
sample_df <- data.frame(a = 1:5, b = 5:1, c = c(5, 3, 1, 4, 1))

Shorthand for sample_df[sample_df$a >= 4,]
subset(sample_df, a >= 4)
#> a b c
#> 4 4 2 4
#> 5 5 1 1

Shorthand for sample_df[sample_df$b == sample_df$c,]
subset(sample_df, b == c)
#> a b c
#> 1 1 5 5
#> 5 5 1 1
```

The core of our version of `subset()`, `subset2()`, is quite simple. It takes two arguments: a data frame, `df`, and an expression, `rows`. We evaluate `rows` using `df` as a data mask, then use the results to subset the data frame with `[`. I've included a very simple check to ensure the result is a logical vector; real code should do more work to create an informative error.

```
subset2 <- function(df, rows) {
 rows <- enquo(rows)

 rows_val <- eval_tidy(rows, df)
 stopifnot(is.logical(rows_val))

 df[rows_val, , drop = FALSE]
}

subset2(sample_df, b == c)
#> a b c
#> 1 1 5 5
#> 5 5 1 1
```

### 19.4.3 Application: `arrange()`

A slightly more complicated exercise is to implement the heart of `dplyr::arrange()`. The goal of `arrange()` is to allow you to sort a data frame by multiple variables, each evaluated in the context of the data frame. This is more challenging than `subset()` because we want to arrange by multiple variables captured in `....`.

```
arrange2 <- function(.df, ..., .na.last = TRUE) {
 # Capture all dots
 args <- enquos(...)

 # Create a call to order, using `!!!` to splice in the
 # individual expressions, and `!!` to splice in na.last
 order_call <- quo(order(!!!args, na.last = !!.na.last))

 # Evaluate the call to order with
 ord <- eval_tidy(order_call, .df)

 .df[ord, , drop = FALSE]
}
```

```
df <- data.frame(x = c(2, 3, 1), y = runif(3))

arrange2(df, x)
#> x y
#> 3 1 0.175
#> 1 2 0.773
#> 2 3 0.875
arrange2(df, -y)
#> x y
#> 2 3 0.875
#> 1 2 0.773
#> 3 1 0.175
```

#### 19.4.4 Ambiguity and pronouns

One of the downsides of the data mask is that it introduces ambiguity: when you say `x`, are you referring to a variable in the data or in the environment? This ambiguity is ok when doing interactive data analysis because you are familiar with the data, and if there are problems, you'll spot them quickly because you are looking at the data frequently. However, ambiguity becomes a problem when you start programming with functions that use tidy evaluation. For example, take this simple wrapper:

```
threshold_x <- function(df, val) {
 subset2(df, x >= val)
}
```

This function can silently return an incorrect result in two situations:

- If `df` does not contain a variable called `x` and `x` exists in the calling environment, `threshold_x()` will silently return an incorrect result:

```
x <- 10
no_x <- data.frame(y = 1:3)
threshold_x(no_x, 2)
#> y
#> 1 1
#> 2 2
#> 3 3
```

- If `df` contains a variable called `val`, the function will always return an incorrect answer:

```
has_val <- data.frame(x = 1:3, val = 9:11)
threshold_x(has_val, 2)
#> [1] x val
#> <0 rows> (or 0-length row.names)
```

These failure modes arise because tidy evaluation is ambiguous: each variable can be found in **either** the data mask **or** the environment. To make this function work we need to remove that ambiguity and ensure that `x` is always found in the data and `val` in the environment. To make this possible `eval_tidy()` provides `.data` and `.env` pronouns:

```
threshold_x <- function(df, val) {
 subset2(df, .data$x >= .env$val)
}

x <- 10
threshold_x(no_x, 2)
```

```
#> Error: Column `x` not found in `^.data`
threshold_x(has_val, 2)
#> x val
#> 2 2 10
#> 3 3 11
```

(NB: unlike indexing an ordinary list or environment with `$`, these pronouns will throw an error if the variable is not found)

Generally, whenever you use the `.env` pronoun, you can use unquoting instead:

```
threshold_x <- function(df, val) {
 subset2(df, ^.data$x >= !!val)
}
```

There are subtle differences in when `val` is evaluated. If you unquote, `val` will be evaluated by `enquo()`; if you use a pronoun, `val` will be evaluated by `eval_tidy()`. These differences are usually unimportant, so pick the form that looks most natural.

What if we generalise `threshold_x()` slightly so that the user can pick the variable used for thresholding. There are two basic approaches. Both start by capturing a *symbol*:

```
threshold_var1 <- function(df, var, val) {
 var <- ensym(var)
 subset2(df, `$`(.data, !!var) >= !!val)
}

threshold_var2 <- function(df, var, val) {
 var <- as.character(ensym(var))
 subset2(df, .data[[var]] >= !!val)
}
```

In `threshold_var1` we need to use the prefix form of `$`, because `.data$!!var` is not valid R syntax. Alternatively, we can convert the symbol to a string, and use `[[`.

Note that it is not always the responsibility of the function author to avoid ambiguity. Imagine we generalise further to allow thresholding based on any expression:

```
threshold_expr <- function(df, expr, val) {
 expr <- enquo(expr)
 subset2(df, !!expr >= !!val)
}
```

There's no way to ensure that `expr` is only evaluated in the data, and even if you could, you wouldn't want to because the data does not include any functions. For this function, it's the user's responsibility to avoid ambiguity. As a function author it's your responsibility to avoid ambiguity with any expressions that you create; it's the users responsibility to avoid ambiguity in expressions that they create.

Now that you've seen data masks and pronouns in action, we'll return to `base::subset()` to learn about its limitations.

### 19.4.5 Base `subset()`

The documentation of `subset()` includes the following warning:

This is a convenience function intended for use interactively. For programming it is better to use the standard subsetting functions like `[`, and in particular the non-standard evaluation of argument `subset` can have unanticipated consequences.

Why is `subset()` dangerous for programming and how does tidy evaluation help us avoid those dangers? First, let's implement the key parts of `subset()` using base R, following the same structure as `subset2()`. We convert `enquo()` to `substitute()` and `eval_tidy()` to `eval()`. We also need to supply a backup environment to `eval()`. There's no way to access the environment associated with an argument in base R, so we take the best approximation: the caller environment (aka parent frame).

```
subset_base <- function(data, rows) {
 rows <- substitute(rows)

 rows_val <- eval(rows, data, caller_env())
 stopifnot(is.logical(rows_val))

 data[rows_val, , drop = FALSE]
}
```

There are three problems with this implementation:

- `subset()` doesn't support unquoting, so wrapping the function is hard. First, you use `substitute()` to capture the complete expression, then you evaluate it. Because `substitute()` doesn't use a syntactic marker for unquoting, it is hard to see exactly what's happening here.

```
f1a <- function(df, expr) {
 call <- substitute(subset(df, expr))
 eval(call, caller_env())
}

df <- data.frame(x = 1:3, y = 3:1)
f1a(df, x == 1)
#> x y
#> 1 1 3
```

I think the tidy evaluation equivalent is easier to understand because the quoting and unquoting is explicit:

```
f1b <- function(df, expr) {
 expr <- enquo(expr)
 subset2(df, !!expr)
}

f1b(df, x == 1)
#> x y
#> 1 1 3
```

- `base::subset()` always evaluates `rows` in the parent frame, but if `...` has been used, then the expression might need to be evaluated elsewhere:

```
f <- function(df, ...) {
 xval <- 3
 subset(df, ...)
}

xval <- 1
f(df, x == xval)
#> x y
#> 3 3 1
```

Because `enquo()` captures the environment of the argument as well as its expression, this is not a problem with `subset2()`:

```
f <- function(df, ...) {
 xval <- 10
 subset2(df, ...)
}

xval <- 1
f(df, x == xval)
#> x y
#> 1 1 3
```

- Finally, `eval()` doesn't provide any pronouns so there's no way to write a safe version of `threshold_x()`.

You might wonder if all this rigamorale is worth it when you can just use `[`. Firstly, it seems unappealing to have functions that can only be used safely in an interactive context. That would mean that every interactive function needs to be paired with function suitable for programming. Secondly, even the simple `subset()` function provides two useful features compared to `[`:

- It sets `drop = FALSE` by default, so it's guaranteed to return a data frame
- It drops rows where the condition evaluates to `NA`.

That means `subset(df, x == y)` is not equivalent to `df[x == y, ]` as you might expect. Instead, it is equivalent to `df[x == y & !is.na(x == y), , drop = FALSE]`: that's a lot more typing!

#### 19.4.6 Performance

Note that there is some performance overhead when evaluating a quosure compared to evaluating an expression:

```
n <- 1000
x1 <- expr(runif(n))
e1 <- globalenv()
q1 <- quo(runif(n))

microbenchmark::microbenchmark(
 runif(n),
 eval_bare(x1, e1),
 eval_tidy(q1),
 eval_tidy(q1, mtcars)
)
#> Unit: microseconds
#> expr min lq mean median uq max neval
#> runif(n) 36.8 37.7 41.5 38.4 39.2 79.8 100
#> eval_bare(x1, e1) 37.5 38.7 42.7 39.1 40.1 107.2 100
#> eval_tidy(q1) 40.2 41.5 44.2 42.4 43.6 82.6 100
#> eval_tidy(q1, mtcars) 42.6 44.9 57.1 46.0 50.3 581.2 100
```

However, most of the overhead is due to setting up the data mask so if you need to evaluate code repeatedly, it's a good idea to define the data mask once then reuse it. This considerably reduces the overhead, with a small change in behaviour: if the code being evaluated creates objects in the "current" environment, those objects will persist across calls.

```
d_mtcars <- as_data_mask(mtcars)

microbenchmark::microbenchmark(
 as_data_mask(mtcars),
```

```

eval_tidy(q1, mtcars),
eval_tidy(q1, d_mtcars)
)
#> Unit: microseconds
#> expr min lq mean median uq max neval
#> as_data_mask(mtcars) 4.8 5.6 7.2 6.19 7.43 61.9 100
#> eval_tidy(q1, mtcars) 42.7 44.1 49.8 45.37 47.27 90.7 100
#> eval_tidy(q1, d_mtcars) 38.4 39.2 43.3 39.63 40.52 96.6 100

```

### 19.4.7 Exercises

1. Improve `subset2()` to make it more like `base::subset()`:
  - Drop rows where `subset` evaluates to `NA`.
  - Give a clear error message if `subset` doesn't yield a logical vector.
  - What happens if `subset` yields a vector that's not the same as the number rows in `data`? What do you think should happen?
2. The third argument in `base::subset()` allows you to select variables. It treats variable names as if they were positions. This allows you to do things like `subset(mtcars, , -cyl)` to drop the cylinder variable, or `subset(mtcars, , disp:drat)` to select all the variables between `disp` and `drat`. How does this work? I've made this easier to understand by extracting it out into its own function that uses tidy evaluation.

```

select <- function(df, vars) {
 vars <- enexpr(vars)
 var_pos <- set_names(as.list(seq_along(df)), names(df))

 cols <- eval_tidy(vars, var_pos)
 df[, cols, drop = FALSE]
}
select(mtcars, -cyl)

```

3. Here's an alternative implementation of `arrange()`:

```

invoke <- function(fun, ...) do.call(fun, dots_list(...))
arrange3 <- function(.data, ..., .na.last = TRUE) {
 args <- enquos(...)

 ords <- purrr::map(args, eval_tidy, data = .data)
 ord <- invoke(order, !!!ords, na.last = .na.last)

 .data[ord, , drop = FALSE]
}

```

Describe the primary difference in approach compared to the function defined in the text.

One advantage of this approach is that you could check each element of `...` to make sure that input is correct. What property should each element of `ords` have?

4. Here's an alternative implementation of `subset2()`:

```

subset3 <- function(data, rows) {
 eval_tidy(quo(data[!enquo(rows)], , drop = FALSE]), data = data)
}

```

Use intermediate variables to make the function easier to understand, then explain how this approach differs to the approach in the text.

5. Implement a form of `arrange()` where you can request a variable to be sorted in descending order using named arguments:

```
arrange(mtcars, cyl, desc = mpg, vs)
```

(Hint: The `decreasing` argument to `order()` will not help you. Instead, look at the definition of `dplyr::desc()`, and read the help for `xtfrm()`.)

6. Why do you not need to worry about ambiguous argument names with `...` in `arrange()`? Why is it a good idea to use the `.` prefix anyway?
7. What does `transform()` do? Read the documentation. How does it work? Read the source code for `transform.data.frame()`. What does `substitute(list(...))` do?
8. Use tidy evaluation to implement your own version of `transform()`. Extend it so that a calculation can refer to variables created by `transform`, i.e. make this work:

```
df <- data.frame(x = 1:3)
transform(df, x1 = x + 1, x2 = x1 + 1)
#> Error in x1 + 1:
#> non-numeric argument to binary operator
```

9. What does `with()` do? How does it work? Read the source code for `with.default()`. What does `within()` do? How does it work? Read the source code for `within.data.frame()`. Why is the code so much more complex than `with()`?
10. Implement a version of `within.data.frame()` that uses tidy evaluation. Read the documentation and make sure that you understand what `within()` does, then read the source code.

## 19.5 Wrapping quoting functions

Now we have all the tools we need to wrap a quoting function inside another function, regardless of whether the quoting function uses tidy evaluation or base R. This is important because it allows you to reduce duplication by turning repeated code into functions. It's straightforward to do this for evaluated arguments; now you'll learn the techniques that allow you to wrap quoted arguments.

### 19.5.1 Tidy evaluation

If you need to wrap a function that quasi-quotes one of its arguments, it's simple to wrap. You just need to quote and unquote. Take this repeat code:

```
df %>% group_by(x1) %>% summarise(mean = mean(y1))
df %>% group_by(x2) %>% summarise(mean = mean(y2))
df %>% group_by(x3) %>% summarise(mean = mean(y3))
```

If no arguments were quoted, we could remove the duplication with:

```
grouped_mean <- function(df, group_var, summary_var) {
 df %>%
 group_by(group_var) %>%
 summarise(mean = mean(summary_var))
}
```

However, both `group_by()` and `summarise()` quote their second and subsequent arguments. That means we need to quote `group_var` and `summary_var` and then unquote when we call `group_by()` and `summarise()`:

```
grouped_mean <- function(df, group_var, summary_var) {
 group_var <- enquo(group_var)
 summary_var <- enquo(summary_var)

 df %>%
 group_by (!!group_var) %>%
 summarise(mean = mean (!!summary_var))
}
```

Just remember that quoting is infectious, so whenever you call a quoting function you need to quote and then unquote.

### 19.5.2 Base R

Unfortunately, things are bit more complex if you want to wrap a base R function that quotes an argument. We can no longer rely on tidy evaluation everywhere, because the semantics of NSE functions are not quite rich enough, but we can use it to generate a mostly correct solution. The wrappers that we create can be used interactively, but can not in turn be easily wrapped. This makes them useful for reducing duplication in your analysis code, but not suitable for inclusion in a package.

We'll focus on wrapping models because this is a common need, and illustrates the spectrum of challenges you'll need to overcome for any other base funtion. Let's start with a very simple wrapper around `lm()`:

```
lm2 <- function(formula, data) {
 lm(formula, data)
}
```

This wrapper works, but is suboptimal because `lm()` captures its call, and displays it when printing:

```
lm2(mpg ~ disp, mtcars)
#>
#> Call:
#> lm(formula = formula, data = data)
#>
#> Coefficients:
#> (Intercept) disp
#> 29.5999 -0.0412
```

This is important because this call is the chief way that you see the model specification when printing the model. To overcome this problem, we need to capture the arguments, create the call to `lm()` using unquoting, then evaluate that call:

```
lm3 <- function(formula, data) {
 formula <- enexpr(formula)
 data <- enexpr(data)

 lm_call <- expr(lm (!!formula, data = !!data))
 eval_bare(lm_call, caller_env())
}

lm3(mpg ~ disp, mtcars)$call
#> lm(formula = mpg ~ disp, data = mtcars)
```

Note that we manually supply an evaluation environment, `caller_env()`. We'll discuss that in more detail shortly.

Note that this technique works for all the arguments, even those that use NSE, like `subset()`:

```
lm4 <- function(formula, data, subset = NULL) {
 formula <- enexpr(formula)
 data <- enexpr(data)
 subset <- enexpr(subset)

 lm_call <- expr(lm (!!formula, data = !!data, subset = !!subset))
 eval_bare(lm_call, caller_env())
}

coef(lm4(mpg ~ disp, mtcars))
#> (Intercept) disp
#> 29.5999 -0.0412
coef(lm4(mpg ~ disp, mtcars, subset = cyl == 4))
#> (Intercept) disp
#> 40.872 -0.135
```

Note that I've supplied a default argument to `subset`. I think this is good practice because it clearly indicates that `subset` is optional: arguments with no default are usually required. `NULL` has two nice properties here:

1. `lm()` already knows how to handle `subset = NULL`: it treats it the same way as a missing `subset`.
2. `expr(NULL)` is `NULL`; which makes it easier to detect programmatically.

However, the current approach has one small downside: `subset = NULL` is shown in the call.

```
lm4(mpg ~ disp, mtcars)$call
#> lm(formula = mpg ~ disp, data = mtcars, subset = NULL)
```

It's possible, if a little more work, to generate a call where `subset` is simply absent. There are two tricks needed to do this:

1. We use the `%||%` helper to replace a `NULL` subset with `missing_arg()`.
2. We use `maybe_missing()` in `expr()`: if we don't do that the essential weirdness of the missing argument crops up and generates an error.

This leads to `lm5()`:

```
lm5 <- function(formula, data, subset = NULL) {
 formula <- enexpr(formula)
 data <- enexpr(data)
 subset <- enexpr(subset) %||% missing_arg()

 lm_call <- expr(lm (!!formula, data = !!data, subset = !!maybe_missing(subset)))
 eval_bare(lm_call, caller_env())
}

lm5(mpg ~ disp, mtcars)$call
#> lm(formula = mpg ~ disp, data = mtcars)
```

Note that all these wrappers have one small advantage over `lm()`: we can use unquoting.

```
f <- mpg ~ disp
lm5 (!!f, mtcars)$call
#> lm(formula = mpg ~ disp, data = mtcars)

resp <- expr(mpg)
lm5 (!!resp ~ disp, mtcars)$call
#> lm(formula = mpg ~ disp, data = mtcars)
```

### 19.5.3 The evaluation environment

What if you want to mingle objects supplied by the user with objects that you create in the function? For example, imagine you want to make an auto-bootstrapping version of `lm()`. You might write it like this:

```
boot_lm0 <- function(formula, data) {
 formula <- enexpr(formula)
 boot_data <- data[sample(nrow(data), replace = TRUE), , drop = FALSE]

 lm_call <- expr(lm(!!formula, data = boot_data))
 eval_bare(lm_call, caller_env())
}

df <- data.frame(x = 1:10, y = 5 + 3 * (1:10) + rnorm(10))
boot_lm0(y ~ x, data = df)
#> Error in is.data.frame(data):
#> object 'boot_data' not found
```

Why doesn't this code work? It's because we're evaluating `lm_call` in the caller environment, but `boot_data` exists in the execution environment. We could instead evaluate in the execution environment of `boot_lm0()`, but there's no guarantee that `formula` could be evaluated in that environment.

There are two basic ways to overcome this challenge:

1. Unquote the data frame into the call. This means that no look up has to occur, but has all the problems of inlining expressions. For modelling functions this means that captured call is suboptimal:

```
boot_lm1 <- function(formula, data) {
 formula <- enexpr(formula)
 boot_data <- data[sample(nrow(data), replace = TRUE), , drop = FALSE]

 lm_call <- expr(lm(!!formula, data = !!boot_data))
 eval_bare(lm_call, caller_env())
}

boot_lm1(y ~ x, data = df)$call
#> lm(formula = y ~ x, data = list(x = c(3L, 6L, 7L, 9L, 9L,
#> 4L, 1L, 7L, 3L), y = c(14.6648781752736, 22.1126241808277, 25.9200218389642,
#> 33.1201105917209, 33.1201105917209, 33.1201105917209, 17.4530448354519,
#> 7.91010669552239, 25.9200218389642, 14.6648781752736)))
```

2. Alternatively you can create a new environment that inherits from the caller, and you can bind variables that you've created inside the function to that environment.

```
boot_lm2 <- function(formula, data) {
 formula <- enexpr(formula)
 boot_data <- data[sample(nrow(data), replace = TRUE), , drop = FALSE]

 lm_env <- child_env(caller_env(), boot_data = boot_data)
 lm_call <- expr(lm(!!formula, data = boot_data))
 eval_bare(lm_call, lm_env)
}

boot_lm2(y ~ x, data = df)
#>
#> Call:
#> lm(formula = y ~ x, data = boot_data)
#>
#> Coefficients:
```

```
#> (Intercept) x
#> 5.12 2.88
```

### 19.5.4 Making formulas

One final aspect to wrapping modelling functions is generating formulas. You just need to learn about one small wrinkle and then you can use the techniques you learned in Quotation. Formulas print the same when evaluated and unevaluated:

```
y ~ x
#> y ~ x
expr(y ~ x)
#> y ~ x
```

Instead, check the class to make sure you have an actual formula:

```
class(y ~ x)
#> [1] "formula"
class(expr(y ~ x))
#> [1] "call"
class(eval_bare(expr(y ~ x)))
#> [1] "formula"
```

Once you understand this, you can generate formulas with unquoting and `reduce()`. Just remember to evaluate the result before returning it. Like in another base NSE wrapper, you should use `caller_env()` as the evaluation environment.

Here's a simple example that generates a formula by combining a response variable with a set of predictors.

```
build_formula <- function(resp, ...) {
 resp <- enexpr(resp)
 preds <- enexprs(...)

 pred_sum <- purrr::reduce(preds, ~ expr (!!x + !!y))
 eval_bare(expr (!!resp ~ !!pred_sum), caller_env())
}
```

`build_formula(y, a, b, c)`

### 19.5.5 Exercises

- When model building, typically the response and data are relatively constant while you rapidly experiment with different predictors. Write a small wrapper that allows you to reduce duplication in this situation.

```
pred_mpg <- function(resp, ...) {

}
pred_mpg(~ disp)
pred_mpg(~ I(1 / disp))
pred_mpg(~ disp * cyl)
```

- Another way to write `boot_lm()` would be to include the bootstrapping expression `(data[sample(nrow(data), replace = TRUE), , drop = FALSE])` in the `data` argument. Implement that approach. What are the advantages? What are the disadvantages?

3. To make these functions some what more robust, instead of always using the `caller_env()` we could capture a quosure, and then use its environment. However, if there are multiple arguments, they might be associated with different environments. Write a function that takes a list of quosures, and returns the common environment, if they have one, or otherwise throws an error.
4. Write a function that takes a data frame and a list of formulas, fitting a linear model with each formula, generating a useful model call.
5. Create a formula generation function that allows you to optionally supply a transformation function (e.g. `log()`) to the response or the predictors.



# Chapter 20

## Translating R code

### 20.1 Introduction

The combination of first class environments, lexical scoping, and metaprogramming gives us a powerful toolkit for translating R code in to other languages. One fully-fledged example of this idea is dbplyr. dbplyr powers the database backends for dplyr, allowing to express data manipulation in R and automatically translating it in to SQL. An important part of dbplyr is `translate_sql()` which turns vector R code in to the equivalent SQL:

```
library(dbplyr)
translate_sql(x ^ 2)
#> <SQL> POWER("x", 2.0)
translate_sql(x < 5 & !is.na(x))
#> <SQL> "x" < 5.0 AND NOT((("x") IS NULL))
translate_sql(!first %in% c("John", "Roger", "Robert"))
#> <SQL> NOT("first" IN ('John', 'Roger', 'Robert'))
translate_sql(select == 7)
#> <SQL> "select" = 7.0
```

This chapter will develop two simple, but useful DSLs: one to generate HTML, and the other to turn mathematical expressions from R code into LaTeX.

### Outline

#### Prerequisites

This chapter together pulls together many techniques discussed elsewhere in the book. In particular, you'll need to understand environments, metaprogramming, and a little functional programming and S3. We'll use rlang for its metaprogramming tools, and purrr for its mapping functions

```
library(rlang)
library(purrr)
```

## 20.2 HTML

HTML (hypertext markup language) is the language that underlies the majority of the web. It's a special case of SGML (standard generalised markup language), and it's similar but not identical to XML (extensible markup language). HTML looks like this:

```
<body>
 <h1 id='first'>A heading</h1>
 <p>Some text < b >some bold text.</p>

</body>
```

Even if you've never looked at HTML before, you can still see that the key component of its coding structure is tags: `<tag></tag>`. Tags can be nested within other tags and intermingled with text. There are over 100 HTML tags, but in this chapter we'll focus on just a handful:

- `<body>` is the top-level tag that contains all content.
- `<h1>` defines a top level heading.
- `<p>` defines a paragraph.
- `<b>` emboldens text.
- `<img>` embeds an image.

Tags can also have named **attributes** which look like `<tag name1='value1' name2='value2'></tag>`. Two important attributes used with just about every tag are `id` and `class`. These are used in conjunction with CSS (cascading style sheets) in order to control the visual appearance of the page.

**Void tags**, like `<img>`, don't have any content, are written `<img />`, not `<img></img>`. Since they have no content, attributes are more important, and `img` has three that are used with almost every image: `src` (where the image lives), `width`, and `height`.

Because `<` and `>` have special meanings in HTML, you can't write them directly. Instead you have to use the HTML **escapes**: `&gt;` and `&lt;`. And, since those escapes use `&`, if you want a literal ampersand you have to escape it with `&amp;`.

### 20.2.1 Goal

Our goal is to make it easy to generate HTML from R. To give a concrete example, we want to generate the following HTML:

```
<body>
 <h1 id='first'>A heading</h1>
 <p>Some text < b >some bold text.</p>

</body>
```

And we want the structure of the R code to match the structure of the HTML as closely as possible. To that end, we will work our way up to the following DSL:

```
with_html(
 body(
 h1("A heading", id = "first"),
 p("Some text &", b("some bold text.")),
 img(src = "myimg.png", width = 100, height = 100)
)
)
```

This DSL has the following three properties:

- The nesting of function calls matches the nesting of tags.
- Unnamed arguments become the content of the tag, and named arguments become their attributes.
- We can automatically escape & and other special characters because tags and text are clearly distinct.

### 20.2.2 Escaping

Escaping is so fundamental to translation that it'll be our first topic. There are two related challenges:

- In user input, we need to automatically escape &, < and >.
- At the same time we need to make sure that the &, < and > we generate are not double-escaped (i.e. to &amp;amp;, &amp;lt; and &amp;gt;).

The easiest way to do this is to create an S3 class that distinguishes between regular text (that needs escaping) and HTML (that doesn't).

```
html <- function(x) structure(x, class = "advr_html")
cat_line <- function(...) cat(..., "\n", sep = "")

print.advr_html <- function(x, ...) {
 out <- paste0("<HTML> ", x)
 cat_line(paste(strwrap(out), collapse = "\n"))
}
```

We then write an escape method. It has two important methods:

- `escape.character()` takes a regular character vector and returns an HTML vector with special characters (&, <, >) escaped.
- `escape.advr_html()` which leaves already escaped HTML as is.

```
escape <- function(x) UseMethod("escape")

escape.character <- function(x) {
 x <- gsub("&", "&", x)
 x <- gsub("<", "<", x)
 x <- gsub(">", ">", x)

 html(x)
}

escape.advr_html <- function(x) x
```

Now we check that it works

```
escape("This is some text.")
#> <HTML> This is some text.
escape("x > 1 & y < 2")
#> <HTML> x > 1 & y < 2

Double escaping is not a problem
escape(escape("This is some text. 1 > 2"))
#> <HTML> This is some text. 1 > 2

And text we know is HTML doesn't get escaped.
escape(html("<hr />"))
#> <HTML> <hr />
```

Conveniently this also gives the user a way to opt-out of our escaping if they know the content is already escaped.

### 20.2.3 Basic tag functions

Next, we'll write a few simple tag functions then figure out how to generalise this function to cover all possible tags.

Let's start with `<p>`. HTML tags can have both attributes (e.g., id or class) and children (like `<b>` or `<i>`). We need some way of separating these in the function call. Given that attributes are named values and children don't have names, it seems natural to separate using named arguments from unnamed ones. For example, a call to `p()` might look like:

```
p("Some text. ", b(i("some bold italic text")), class = "mypara")
```

We could list all the possible attributes of the `<p>` tag in the function definition. But that's hard not only because there are many attributes, but also because it's possible to use custom attributes (<http://html5doctor.com/html5-custom-data-attributes/>). Instead, we'll just use `...` and separate the components based on whether or not they are named. With this in mind, we create a helper function that wraps around `rlang::dots_list()` (so we can use `!!` and `!!!`) and returns named and unnamed components separately:

```
dots_partition <- function(...) {
 dots <- dots_list(...)

 is_named <- names(dots) != ""
 list(
 named = dots[is_named],
 unnamed = dots[!is_named]
)
}

str(dots_partition(a = 1, 2, b = 3, 4))
#> List of 2
#> $ named :List of 2
#> ..$ a: num 1
#> ..$ b: num 3
#> $ unnamed:List of 2
#> ..$: num 2
#> ..$: num 4
```

We can now create our `p()` function. Notice that there's one new function here: `html_attributes()`. It takes a named list and returns the HTML attribute specification as a string. It's a little complicated (in part, because it deals with some idiosyncracies of HTML that I haven't mentioned.), but it's not that important and doesn't introduce any programming new ideas, so I won't discuss it here (you can find the source online (<https://github.com/hadley/adv-r/blob/master/dsl-html-attributes.r>)).

```
source("dsl-html-attributes.r", local = TRUE)
p <- function(...) {
 dots <- dots_partition(...)
 attrs <- html_attributes(dots$named)
 children <- map_chr(dots$unnamed, escape)

 html(paste0(
 "<p", attrs, ">",
```

```

 paste(children, collapse = ""),
 "</p>"
))
}

p("Some text")
#> <HTML> <p>Some text</p>
p("Some text", id = "myid")
#> <HTML> <p id='myid'>Some text</p>
p("Some text", class = "important", `data-value` = 10)
#> <HTML> <p class='important' data-value='10'>Some text</p>

```

## 20.2.4 Tag functions

It's straightforward to adapt `p()` to other tags: we just need to replace "p" with the name of the tag. One elegant way to do that is to manually create a function with `rlang::new_function()`, using unquoting with `paste0()` to generate the starting and ending tags.

```

tag <- function(tag) {
 new_function(
 exprs(... =),
 expr({
 dots <- dots_partition(...)
 attrbs <- html_attributes(dots$named)
 children <- map_chr(dots$unnamed, escape)

 html(paste0(
 !paste0("<", tag), attrbs, ">",
 paste(children, collapse = ""),
 !paste0("</", tag, ">")
))
 }),
 caller_env()
)
}

tag("b")
#> function (...){
#> dots <- dots_partition(...)
#> attrbs <- html_attributes(dots$named)
#> children <- map_chr(dots$unnamed, escape)
#> html(paste0("<b", attrbs, ">", paste(children, collapse = ""),
#> ""))
#> }

```

Now we can run our earlier example:

```

p <- tag("p")
b <- tag("b")
i <- tag("i")
p("Some text. ", b(i("some bold italic text")), class = "mypara")
#> <HTML> <p class='mypara'>Some text. <i>some bold italic
#> text</i></p>

```

Before we generate functions for every possible HTML tag, we need to create a variant of `tag()` for void tags. It's very similar to `tag()`, but it will throw an error if there are any unnamed tags, and the tag itself looks a little different.

```
void_tag <- function(tag) {
 new_function(
 exprs(... =),
 expr({
 dots <- dots_partition(...)
 if (length(dots$unnamed) > 0) {
 stop (!!paste0("<", tag, "> must not have unnamed arguments")), call. = FALSE)
 }
 attribs <- html_attributes(dots$named)

 html(paste0 (!!paste0("<", tag), attribs, " />"))
 }),
 caller_env()
)
}

img <- void_tag("img")
img(src = "myimage.png", width = 100, height = 100)
#> <HTML>
```

## 20.2.5 Processing all tags

Next we need a list of all the HTML tags:

```
tags <- c("a", "abbr", "address", "article", "aside", "audio",
 "b", "bdi", "bdo", "blockquote", "body", "button", "canvas",
 "caption", "cite", "code", "colgroup", "data", "datalist",
 "dd", "del", "details", "dfn", "div", "dl", "dt", "em",
 "eventsource", "fieldset", "figcaption", "figure", "footer",
 "form", "h1", "h2", "h3", "h4", "h5", "h6", "head", "header",
 "hgroup", "html", "i", "iframe", "ins", "kbd", "label",
 "legend", "li", "mark", "map", "menu", "meter", "nav",
 "noscript", "object", "ol", "optgroup", "option", "output",
 "p", "pre", "progress", "q", "ruby", "rp", "rt", "s", "samp",
 "script", "section", "select", "small", "span", "strong",
 "style", "sub", "summary", "sup", "table", "tbody", "td",
 "textarea", "tfoot", "th", "thead", "time", "title", "tr",
 "u", "ul", "var", "video")

void_tags <- c("area", "base", "br", "col", "command", "embed",
 "hr", "img", "input", "keygen", "link", "meta", "param",
 "source", "track", "wbr")
```

If you look at this list carefully, you'll see there are quite a few tags that have the same name as base R functions (`body`, `col`, `q`, `source`, `sub`, `summary`, `table`), and others that have the same name as popular packages (e.g., `map`). This means we don't want to make all the functions available by default, in either the global environment or in a package. Instead, we'll put them in a list and then provide a helper to make it easy to use them when desired. First, we make a named list:

```
html_tags <- c(
 tags %>% set_names() %>% map(tag),
 void_tags %>% set_names() %>% map(void_tag)
)
```

This gives us an explicit (but verbose) way to call tag functions:

```
html_tags$p(
 "Some text. ",
 html_tags$b(html_tags$i("some bold italic text")),
 class = "mypara"
)
#> <HTML> <p class='mypara'>Some text. <i>some bold italic
#> text</i></p>
```

We can then finish off our HTML DSL with a function that allows us to evaluate code in the context of that list. Here we slightly abuse the data mask, passing it a list of functions rather than a data frame. This is quick hack to mingle the execution environment of `code` with the functions in `html_tags`.

```
with_html <- function(code) {
 code <- enquo(code)
 eval_tidy(code, html_tags)
}
```

This gives us a succinct API which allows us to write HTML when we need it but doesn't clutter up the namespace when we don't.

```
with_html(
 body(
 h1("A heading", id = "first"),
 p("Some text &", b("some bold text.")),
 img(src = "myimg.png", width = 100, height = 100)
)
)
#> <HTML> <body><h1 id='first'>A heading</h1><p>Some text
#> &#amp;lt;b>some bold text.</p><img src='myimg.png'
#> width='100' height='100' /></body>
```

If you want to access the R function overridden by an HTML tag with the same name inside `with_html()`, you can use the full `package::function` specification.

## 20.2.6 Exercises

1. The escaping rules for `<script>` and `<style>` tags are different: you don't want to escape angle brackets or ampersands, but you do want to escape `</script>` or `</style>`. Adapt the code above to follow these rules.
2. The use of `...` for all functions has some big downsides. There's no input validation and there will be little information in the documentation or autocomplete about how they are used in the function. Create a new function that, when given a named list of tags and their attribute names (like below), creates functions which address this problem.

```
list(
 a = c("href"),
 img = c("src", "width", "height")
)
```

All tags should get `class` and `id` attributes.

3. Currently the HTML doesn't look terribly pretty, and it's hard to see the structure. How could you adapt `tag()` to do indenting and formatting?
4. Reason about the following code that calls `with_html()` referencing objects from the environment. Will it work or fail? Why? Run the code to verify your predictions.

```
greeting <- "Hello!"
with_html(p(greeting))

address <- "123 anywhere street"
with_html(p(address))
```

## 20.3 LaTeX

The next DSL will convert R expressions into their LaTeX math equivalents. (This is a bit like `?plotmath`, but for text instead of plots.) LaTeX is the lingua franca of mathematicians and statisticians: it's common to use LaTeX notation whenever you want to express an equation in text (e.g., in an email). Since many reports are produced using both R and LaTeX, it might be useful to be able to automatically convert mathematical expressions from one language to the other.

Because we need to convert both functions and names, this mathematical DSL will be more complicated than the HTML DSL. We'll also need to create a “default” conversion, so that functions we don't know about get a standard conversion. Like the HTML DSL, we'll also use metaprogramming to make it easier to generate the translators.

Can no longer just use `eval`: we also need to walk the tree. Ideally this would not be necessary. ObjectTables (see `objectable` package) almost make it possible to eliminate the tree walking but:

- They have currently have a big performance penalty
- There's no way to distinguish symbols used in function calls vs. other symbols.

Before we begin, let's quickly cover how formulas are expressed in LaTeX.

### 20.3.1 LaTeX mathematics

The full spectrum of LaTeX mathematical notation is complex. Fortunately, they are well documented (<http://en.wikibooks.org/wiki/LaTeX/Mathematics>), and the most common commands have a fairly simple structure:

- Most simple mathematical equations are written in the same way you'd type them in R: `x * y`, `z ^ 5`. Subscripts are written using `_` (e.g., `x_1`).
- Special characters start with a `\`: `\pi` =  $\pi$ , `\pm` =  $\pm$ , and so on. There are a huge number of symbols available in LaTeX. Googling for `latex math symbols` will return many lists (<http://www.sunilpatel.co.uk/latex-type/latex-math-symbols/>). There's even a service (<http://detexify.kirelabs.org/classify.html>) that will look up the symbol you sketch in the browser.
- More complicated functions look like `\name{arg1}{arg2}`. For example, to write a fraction you'd use `\frac{a}{b}`. To write a square root, you'd use `\sqrt{a}`.
- To group elements together use `{}`: i.e., `x ^ a + b` vs. `x ^ {a + b}`.
- In good math typesetting, a distinction is made between variables and functions. But without extra information, LaTeX doesn't know whether `f(a * b)` represents calling the function `f` with input `a`

\*  $b$ , or is shorthand for  $f * (a * b)$ . If  $f$  is a function, you can tell LaTeX to typeset it using an upright font with `\textrm{f}(a * b)`.

### 20.3.2 Goal

Our goal is to use these rules to automatically convert an R expression to its appropriate LaTeX representation. We'll tackle this in four stages:

- Convert known symbols:  $\pi \rightarrow \pi$
- Leave other symbols unchanged:  $x \rightarrow x$ ,  $y \rightarrow y$
- Convert known functions to their special forms: `sqrt(frac(a, b))`  $\rightarrow \sqrt{\frac{a}{b}}$
- Wrap unknown functions with `\textrm{f}(a) \rightarrow \textrm{f}(a)`

We'll code this translation in the opposite direction of what we did with the HTML DSL. We'll start with infrastructure, because that makes it easy to experiment with our DSL, and then work our way back down to generate the desired output.

### 20.3.3 `to_math`

To begin, we need a wrapper function that will convert R expressions into LaTeX math expressions. This will work similarly to `to_html()`: capture the unevaluated expression and evaluate it in a special environment. Two main differences:

- Environment is no longer constant. It will vary depending on the expression. We do this in order to be specially handle unknown symbols and functions
- Don't use quosure.

```
to_math <- function(x) {
 expr <- enexpr(x)
 out <- eval_bare(expr, latex_env(expr))

 latex(out)
}

latex <- function(x) structure(x, class = "advr_latex")
print.advr_latex <- function(x) {
 cat_line("<LATEX> ", x)
}
```

### 20.3.4 Known symbols

Our first step is to create an environment that will convert the special LaTeX symbols used for Greek, e.g.,  $\pi$  to `\pi`. We'll use the same basic trick as used by `subset` to make it possible to select column ranges by name (`subset(mtcars, , cyl:wt)`): bind a name to a string in a special environment.

We create that environment by naming a vector, converting the vector into a list, and converting the list into an environment.

```
greek <- c(
 "alpha", "theta", "tau", "beta", "vartheta", "pi", "upsilon",
 "gamma", "varpi", "phi", "delta", "kappa", "rho",
 "varphi", "epsilon", "lambda", "varkappa", "chi", "varepsilon",
```

```
"mu", "sigma", "psi", "zeta", "nu", "varsigma", "omega", "eta",
"xi", "Gamma", "Lambda", "Sigma", "Psi", "Delta", "Xi",
"Upsilon", "Omega", "Theta", "Pi", "Phi")
greek_list <- set_names(paste0("\\", greek), greek)
greek_env <- as_env(greek_list)
```

We can then check it:

```
latex_env <- function(expr) {
 greek_env
}

to_math(pi)
#> <LATEX> \pi
to_math(beta)
#> <LATEX> \beta
```

Looks good so far!

### 20.3.5 Unknown symbols

If a symbol isn't Greek, we want to leave it as is. This is tricky because we don't know in advance what symbols will be used, and we can't possibly generate them all. So we'll use the approach described in walking the tree. The `all_names` function takes an expression and does the following: if it's a name, it converts it to a string; if it's a call, it recurses down through its arguments.

```
all_names_rec <- function(x) {
 switch_expr(x,
 constant = character(),
 symbol = as.character(x),
 pairlist = ,
 call = flat_map_chr(as.list(x[-1]), all_names)
)
}

all_names <- function(x) {
 unique(all_names_rec(x))
}

all_names(expr(x + y + f(a, b, c, 10)))
#> [1] "x" "y" "a" "b" "c"
```

We now want to take that list of symbols, and convert it to an environment so that each symbol is mapped to its corresponding string representation (e.g., so `eval(quote(x), env)` yields "x"). We again use the pattern of converting a named character vector to a list, then converting the list to an environment.

```
latex_env <- function(expr) {
 names <- all_names(expr)
 symbol_env <- as_env(set_names(names))

 symbol_env

 to_math(x)
#> <LATEX> x
```

```
to_math(longvariablename)
#> <LATEX> longvariablename
to_math(pi)
#> <LATEX> pi
```

This works, but we need to combine it with the Greek symbols environment. Since we want to give preference to Greek over defaults (e.g., `to_math(pi)` should give "`\pi`", not "pi"), `symbol_env` needs to be the parent of `greek_env`. To do that, we need to make a copy of `greek_env` with a new parent.

This gives us a function that can convert both known (Greek) and unknown symbols.

```
latex_env <- function(expr) {
 # Unknown symbols
 names <- all_names(expr)
 symbol_env <- as_env(set_names(names))

 # Known symbols
 env_clone(greek_env, parent = symbol_env)
}

to_math(x)
#> <LATEX> x
to_math(longvariablename)
#> <LATEX> longvariablename
to_math(pi)
#> <LATEX> \pi
```

### 20.3.6 Known functions

Next we'll add functions to our DSL. We'll start with a couple of helper closures that make it easy to add new unary and binary operators. These functions are very simple: they only assemble strings. (Again we use `force()` to make sure the arguments are evaluated at the right time.)

```
unary_op <- function(left, right) {
 new_function(
 exprs(e1 =),
 expr(
 paste0(!left, e1, !right)
),
 caller_env()
)
}
```

```
binary_op <- function(sep) {
 new_function(
 exprs(e1 = , e2 =),
 expr(
 paste0(e1, !sep, e2)
),
 caller_env()
)
}
```

```
unary_op("\\sqrt{", "}")
```

```
#> function (e1)
#> paste0("\sqrt{", e1, "}")
binary_op("+")
#> function (e1, e2)
#> paste0(e1, "+", e2)
```

Using these helpers, we can map a few illustrative examples of converting R to LaTeX. Note that with R's lexical scoping rules helping us, we can easily provide new meanings for standard functions like `+`, `-`, and `*`, and even `(` and `{`.

```
Binary operators
f_env <- child_env(
 .parent = empty_env(),
 `+` = binary_op(" + "),
 ` - ` = binary_op(" - "),
 `*` = binary_op(" * "),
 `/` = binary_op(" / "),
 `^` = binary_op(" ^ "),
 `[_` = binary_op(" _"),

Grouping
`{` = unary_op("\left{ ", " \right}"),
`(` = unary_op("\left(", " \right)"),
paste = paste,

Other math functions
sqrt = unary_op("\sqrt{", "}",
sin = unary_op("\sin(", ")"),
log = unary_op("\log(", ")"),
abs = unary_op("\left| ", " \right| "),
frac = function(a, b) {
 paste0("\frac{", a, "} {", b, "}")
},
hat = unary_op("\hat{", "}"),
tilde = unary_op("\tilde{", "}")
)
```

We again modify `latex_env()` to include this environment. It should be the last environment R looks for names in: in other words, `sin(sin)` should work.

```
latex_env <- function(expr) {
 # Known functions
 f_env

 # Default symbols
 names <- all_names(expr)
 symbol_env <- as_env(set_names(names), parent = f_env)

 # Known symbols
 greek_env <- env_clone(greek_env, parent = symbol_env)
}

to_math(sin(x + pi))
```

```
#> <LATEX> \sin(x + \pi)
to_math(log(x_i ^ 2))
#> <LATEX> \log(x_i^2)
to_math(sin(sin))
#> <LATEX> \sin(\sin)
```

### 20.3.7 Unknown functions

Finally, we'll add a default for functions that we don't yet know about. Like the unknown names, we can't know in advance what these will be, so we again use a little metaprogramming to figure them out:

```
all_calls_rec <- function(x) {
 switch_expr(x,
 constant = ,
 symbol = character(),
 call = {
 fname <- as.character(x[[1]])
 children <- flat_map_chr(as.list(x[-1]), all_calls)
 c(fname, children)
 },
 pairlist = flat_map_chr(as.list(x[1]), all_calls)
)
}
all_calls <- function(x) {
 unique(all_calls_rec(x))
}

all_calls(expr(f(g + b, c, d(a))))
#> [1] "f" "+" "d"
```

And we need a closure that will generate the functions for each unknown call.

```
unknown_op <- function(op) {
 new_function(
 exprs(... =),
 expr({
 contents <- paste(..., collapse = ", ")
 paste0(!paste0("\\" \mathrm{", op, "})(\"", contents, "\""))
 })
)
}
unknown_op("foo")
#> function (...)

#> {
#> contents <- paste(..., collapse = ", ")
#> paste0("\\" \mathrm{foo})(\"", contents, "\""))
#> }
#> <environment: 0x2c80db0>
```

And again we update `latex_env()`:

```
latex_env <- function(expr) {
 calls <- all_calls(expr)
 call_list <- map(set_names(calls), unknown_op)
```

```
call_env <- as_environment(call_list)

Known functions
f_env <- env_clone(f_env, call_env)

Default symbols
names <- all_names(expr)
symbol_env <- as_env(set_names(names), parent = f_env)

Known symbols
greek_env <- env_clone(greek_env, parent = symbol_env)
}

to_math(f(a * b))
#> <LATEX> \mathrm{f}(a * b)
```

### 20.3.8 Exercises

1. Add escaping. The special symbols that should be escaped by adding a backslash in front of them are \, \$, and %. Just as with HTML, you'll need to make sure you don't end up double-escaping. So you'll need to create a small S3 class and then use that in function operators. That will also allow you to embed arbitrary LaTeX if needed.
2. Complete the DSL to support all the functions that `plotmath` supports.

# **Part V**

# **Techniques**



# Chapter 21

## Debugging

### 21.1 Introduction

What happens when something goes wrong with your R code? What do you do? What tools do you have to address the problem? This chapter will teach you how to fix unanticipated problems (debugging), show you how functions can communicate problems and how you can take action based on those communications (condition handling), and teach you how to avoid common problems before they occur (defensive programming).

Debugging is the art and science of fixing unexpected problems in your code. In this section you'll learn the tools and techniques that help you get to the root cause of an error. You'll learn general strategies for debugging, useful R functions like `traceback()` and `browser()`, and interactive tools in RStudio.

The chapter concludes with a discussion of “defensive” programming: ways to avoid common errors before they occur. In the short run you'll spend more time writing code, but in the long run you'll save time because error messages will be more informative and will let you narrow in on the root cause more quickly. The basic principle of defensive programming is to “fail fast”, to raise an error as soon as something goes wrong. In R, this takes three particular forms: checking that inputs are correct, avoiding non-standard evaluation, and avoiding functions that can return different types of output.

### Outline

1. Debugging techniques outlines a general approach for finding and resolving bugs.
2. Debugging tools introduces you to the R functions and RStudio features that help you locate exactly where an error occurred.
3. Defensive programming introduces you to some important techniques for defensive programming, techniques that help prevent bugs from occurring in the first place.

### 21.2 Techniques

“Finding your bug is a process of confirming the many things that you believe are true — until you find one which is not true.”

—Norm Matloff

Debugging code is challenging. Many bugs are subtle and hard to find. Indeed, if a bug was obvious, you probably would've been able to avoid it in the first place. While it's true that with a good technique, you can productively debug a problem with just `print()`, there are times when additional help would be welcome. In this section, we'll discuss some useful tools, which R and RStudio provide, and outline a general procedure for debugging.

While the procedure below is by no means foolproof, it will hopefully help you to organise your thoughts when debugging. There are four steps:

### 1. Realise that you have a bug

If you're reading this chapter, you've probably already completed this step. It is a surprisingly important one: you can't fix a bug until you know it exists. This is one reason why automated test suites are important when producing high-quality code. Unfortunately, automated testing is outside the scope of this book, but you can read more about it at <http://r-pkgs.had.co.nz/tests.html>.

### 2. Make it repeatable

Once you've determined you have a bug, you need to be able to reproduce it on command. Without this, it becomes extremely difficult to isolate its cause and to confirm that you've successfully fixed it.

Generally, you will start with a big block of code that you know causes the error and then slowly whittle it down to get to the smallest possible snippet that still causes the error. Binary search is particularly useful for this. To do a binary search, you repeatedly remove half of the code until you find the bug. This is fast because, with each step, you reduce the amount of code to look through by half.

If it takes a long time to generate the bug, it's also worthwhile to figure out how to generate it faster. The quicker you can do this, the quicker you can figure out the cause.

As you work on creating a minimal example, you'll also discover similar inputs that don't trigger the bug. Make note of them: they will be helpful when diagnosing the cause of the bug.

If you're using automated testing, this is also a good time to create an automated test case. If your existing test coverage is low, take the opportunity to add some nearby tests to ensure that existing good behaviour is preserved. This reduces the chances of creating a new bug.

### 3. Figure out where it is

If you're lucky, one of the tools in the following section will help you to quickly identify the line of code that's causing the bug. Usually, however, you'll have to think a bit more about the problem. It's a great idea to adopt the scientific method. Generate hypotheses, design experiments to test them, and record your results. This may seem like a lot of work, but a systematic approach will end up saving you time. I often waste a lot of time relying on my intuition to solve a bug ("oh, it must be an off-by-one error, so I'll just subtract 1 here"), when I would have been better off taking a systematic approach.

### 4. Fix it and test it

Once you've found the bug, you need to figure out how to fix it and to check that the fix actually worked. Again, it's very useful to have automated tests in place. Not only does this help to ensure that you've actually fixed the bug, it also helps to ensure you haven't introduced any new bugs in the process. In the absence of automated tests, make sure to carefully record the correct output, and check against the inputs that previously failed.

## 21.3 Tools

To implement a strategy of debugging, you'll need tools. In this section, you'll learn about the tools provided by R and the RStudio IDE. RStudio's integrated debugging support makes life easier by exposing existing R tools in a user friendly way. I'll show you both the R and RStudio ways so that you can work with

whatever environment you use. You may also want to refer to the official RStudio debugging documentation (<https://support.rstudio.com/hc/en-us/articles/205612627-Debugging-with-RStudio>) which always reflects the tools in the latest version of RStudio.

There are three key debugging tools:

- RStudio’s error inspector and  `traceback()` which list the sequence of calls that lead to the error.
- RStudio’s “Rerun with Debug” tool and  `options(error = browser())` which open an interactive session where the error occurred.
- RStudio’s breakpoints and  `browser()` which open an interactive session at an arbitrary location in the code.

I’ll explain each tool in more detail below.

You shouldn’t need to use these tools when writing new functions. If you find yourself using them frequently with new code, you may want to reconsider your approach. Instead of trying to write one big function all at once, work interactively on small pieces. If you start small, you can quickly identify why something doesn’t work. But if you start large, you may end up struggling to identify the source of the problem.

### 21.3.1 Determining the sequence of calls

The first tool is the **call stack**, the sequence of calls that lead up to an error. Here’s a simple example: you can see that `f()` calls `g()` calls `h()` calls `i()` which adds together a number and a string creating an error:

```
f <- function(a) g(a)
g <- function(b) h(b)
h <- function(c) i(c)
i <- function(d) "a" + d
f(10)
```

When we run this code in RStudio we see:

```
> f(10)
Error in "a" + d : non-numeric argument to binary operator
>Show Traceback
★ Rerun with Debug
```

Two options appear to the right of the error message: “Show Traceback” and “Rerun with Debug”. If you click “Show traceback” you see:

```
> f(10)
Error in "a" + d : non-numeric argument to binary operator
>Show Traceback
★ Rerun with Debug
4 i(c) at exceptions-example.R#3
3 h(b) at exceptions-example.R#2
2 g(a) at exceptions-example.R#1
1 f(10)
```

If you’re not using RStudio, you can use  `traceback()` to get the same information:

```
traceback()
4: i(c) at exceptions-example.R#3
3: h(b) at exceptions-example.R#2
2: g(a) at exceptions-example.R#1
1: f(10)
```

Read the call stack from bottom to top: the initial call is `f()`, which calls `g()`, then `h()`, then `i()`, which triggers the error. If you’re calling code that you `source()`d into R, the traceback will also display the

location of the function, in the form `filename.r#linenumber`. These are clickable in RStudio, and will take you to the corresponding line of code in the editor.

Sometimes this is enough information to let you track down the error and fix it. However, it's usually not. `traceback()` shows you where the error occurred, but not why. The next useful tool is the interactive debugger, which allows you to pause execution of a function and interactively explore its state.

### 21.3.2 Browsing on error

The easiest way to enter the interactive debugger is through RStudio's "Rerun with Debug" tool. This reruns the command that created the error, pausing execution where the error occurred. You're now in an interactive state inside the function, and you can interact with any object defined there. You'll see the corresponding code in the editor (with the statement that will be run next highlighted), objects in the current environment in the "Environment" pane, the call stack in a "Traceback" pane, and you can run arbitrary R code in the console.

As well as any regular R function, there are a few special commands you can use in debug mode. You can access them either with the RStudio toolbar ( ) or with the keyboard:

- Next, `n`: executes the next step in the function. Be careful if you have a variable named `n`; to print it you'll need to do `print(n)`.
- Step into, or `s`: works like next, but if the next step is a function, it will step into that function so you can work through each line.
- Finish, or `f`: finishes execution of the current loop or function.
- Continue, `c`: leaves interactive debugging and continues regular execution of the function. This is useful if you've fixed the bad state and want to check that the function proceeds correctly.
- Stop, `Q`: stops debugging, terminates the function, and returns to the global workspace. Use this once you've figured out where the problem is, and you're ready to fix it and reload the code.

There are two other slightly less useful commands that aren't available in the toolbar:

- Enter: repeats the previous command. I find this too easy to activate accidentally, so I turn it off using `options(browserNLdisabled = TRUE)`.
- `where`: prints stack trace of active calls (the interactive equivalent of  `traceback`).

To enter this style of debugging outside of RStudio, you can use the `error` option which specifies a function to run when an error occurs. The function most similar to RStudio's debug is `browser()`: this will start an interactive console in the environment where the error occurred. Use `options(error = browser)` to turn it on, re-run the previous command, then use `options(error = NULL)` to return to the default error behaviour. You could automate this with the `browseOnce()` function as defined below:

```
browseOnce <- function() {
 old <- getOption("error")
 function() {
 options(error = old)
 browser()
 }
}
options(error = browseOnce())

f <- function() stop("!"")
```

```
Enters browser
f()
Runs normally
f()
```

(You'll learn more about functions that return functions in Functional programming.)

There are two other useful functions that you can use with the `error` option:

- `recover` is a step up from `browser`, as it allows you to enter the environment of any of the calls in the call stack. This is useful because often the root cause of the error is a number of calls back.
- `dump.frames` is an equivalent to `recover` for non-interactive code. It creates a `last.dump.rda` file in the current working directory. Then, in a later interactive R session, you load that file, and use `debugger()` to enter an interactive debugger with the same interface as `recover()`. This allows interactive debugging of batch code.

```
In batch R process ----
dump_and_quit <- function() {
 # Save debugging info to file last.dump.rda
 dump.frames(to.file = TRUE)
 # Quit R with error status
 q(status = 1)
}
options(error = dump_and_quit)

In a later interactive session ----
load("last.dump.rda")
debugger()
```

To reset error behaviour to the default, use `options(error = NULL)`. Then errors will print a message and abort function execution.

### 21.3.3 Browsing arbitrary code

As well as entering an interactive console on error, you can enter it at an arbitrary code location by using either an RStudio breakpoint or `browser()`. You can set a breakpoint in RStudio by clicking to the left of the line number, or pressing Shift + F9. Equivalently, add `browser()` where you want execution to pause. Breakpoints behave similarly to `browser()` but they are easier to set (one click instead of nine key presses), and you don't run the risk of accidentally including a `browser()` statement in your source code. There are two small downsides to breakpoints:

- There are a few unusual situations in which breakpoints will not work: read breakpoint troubleshooting (<http://www.rstudio.com/ide/docs/debugging/breakpoint-troubleshooting>) for more details.
- RStudio currently does not support conditional breakpoints, whereas you can always put `browser()` inside an `if` statement.

As well as adding `browser()` yourself, there are two other functions that will add it to code:

- `debug()` inserts a `browser` statement in the first line of the specified function. `undebug()` removes it. Alternatively, you can use `debugonce()` to browse only on the next run.
- `utils::setBreakpoint()` works similarly, but instead of taking a function name, it takes a file name and line number and finds the appropriate function for you.

These two functions are both special cases of `trace()`, which inserts arbitrary code at any position in an existing function. `trace()` is occasionally useful when you're debugging code that you don't have the source

for. To remove tracing from a function, use `untrace()`. You can only perform one trace per function, but that one trace can call multiple functions.

### 21.3.4 The call stack: `traceback()`, `where`, and `recover()`

Unfortunately, the call stacks printed by  `traceback()`,  `browser()` +  `where`, and  `recover()` are not consistent. The following table shows how the call stacks from a simple nested set of calls are displayed by the three tools.

<code> traceback()</code>	<code> where</code>	<code> recover()</code>
4: <code>stop("Error")</code>	where 1: <code>stop("Error")</code>	1: <code>f()</code>
3: <code>h(x)</code>	where 2: <code>h(x)</code>	2: <code>g(x)</code>
2: <code>g(x)</code>	where 3: <code>g(x)</code>	3: <code>h(x)</code>
1: <code>f()</code>	where 4: <code>f()</code>	

Note that numbering is different between  `traceback()` and  `where`, and that  `recover()` displays calls in the opposite order, and omits the call to  `stop()`. RStudio displays calls in the same order as  `traceback()` but omits the numbers.

### 21.3.5 Other types of failure

There are other ways for a function to fail apart from throwing an error or returning an incorrect result.

- A function may generate an unexpected warning. The easiest way to track down warnings is to convert them into errors with `options(warn = 2)` and use the regular debugging tools. When you do this you'll see some extra calls in the call stack, like `doWithOneRestart()`, `withOneRestart()`, `withRestarts()`, and `.signalSimpleWarning()`. Ignore these: they are internal functions used to turn warnings into errors.
- A function may generate an unexpected message. There's no built-in tool to help solve this problem, but it's possible to create one:

```
message2error <- function(code) {
 withCallingHandlers(code, message = function(e) stop(e))
}

f <- function() g()
g <- function() message("Hi!")
f()
Hi!
message2error(f())
Error in message("Hi!"): Hi!
traceback()
11: stop(e) at #2
10: (function (e)
stop(e))(list(message = "Hi!\n", call = message("Hi!")))
9: signalCondition(cond)
8: doWithOneRestart(return(expr), restart)
7: withOneRestart(expr, restarts[[1L]])
6: withRestarts{
signalCondition(cond)
defaultHandler(cond)
```

```
}, muffleMessage = function() NULL)
5: message("Hi!") at #1
4: g() at #1
3: f() at #2
2: withCallingHandlers(code, message = function(e) stop(e)) at #2
1: message2error(f())
```

As with warnings, you'll need to ignore some of the calls on the traceback (i.e., the first two and the last six).

- A function might never return. This is particularly hard to debug automatically, but sometimes terminating the function and looking at the call stack is informative. Otherwise, use the basic debugging strategies described above.
- The worst scenario is that your code might crash R completely, leaving you with no way to interactively debug your code. This indicates a bug in the underlying C code. This is hard to debug. Sometimes an interactive debugger, like `gdb`, can be useful, but describing how to use it is beyond the scope of this book.

If the crash is caused by base R code, post a reproducible example to R-help. If it's in a package, contact the package maintainer. If it's your own C or C++ code, you'll need to use numerous `print()` statements to narrow down the location of the bug, and then you'll need to use many more print statements to figure out which data structure doesn't have the properties that you expect.



# Chapter 22

# Performance

R is not a fast language. This is not an accident. R was purposely designed to make data analysis and statistics easier for you to do. It was not designed to make life easier for your computer. While R is slow compared to other programming languages, for most purposes, it's fast enough.

The goal of this part of the book is to give you a deeper understanding of R's performance characteristics. In this chapter, you'll learn about some of the trade-offs that R has made, valuing flexibility over performance. The following four chapters will give you the skills to improve the speed of your code when you need to:

- In Profiling, you'll learn how to systematically make your code faster. First you figure what's slow, and then you apply some general techniques to make the slow parts faster.
- For really high-performance code, you can move outside of R and use another programming language. Rcpp will teach you the absolute minimum you need to know about C++ so you can write fast code using the Rcpp package.
- To really understand the performance of built-in base functions, you'll need to learn a little bit about R's C API. In R's C interface, you'll learn a little about R's C internals.

Let's get started by learning more about why R is slow.

## 22.1 Why is R slow?

To understand R's performance, it helps to think about R as both a language and as an implementation of that language. The R-language is abstract: it defines what R code means and how it should work. The implementation is concrete: it reads R code and computes a result. The most popular implementation is the one from r-project.org (<http://r-project.org>). I'll call that implementation GNU-R to distinguish it from R-language, and from the other implementations I'll discuss later in the chapter.

The distinction between R-language and GNU-R is a bit murky because the R-language is not formally defined. While there is the R language definition (<http://cran.r-project.org/doc/manuals/R-lang.html>), it is informal and incomplete. The R-language is mostly defined in terms of how GNU-R works. This is in contrast to other languages, like C++ (<http://isocpp.org/std/the-standard>) and javascript (<http://www.ecma-international.org/publications/standards/Ecma-262.htm>), that make a clear distinction between language and implementation by laying out formal specifications that describe in minute detail how every aspect of the language should work. Nevertheless, the distinction between R-language and GNU-R is still useful: poor performance due to the language is hard to fix without breaking existing code; fixing poor performance due to the implementation is easier.

In Language performance, I discuss some of the ways in which the design of the R-language imposes fundamental constraints on R's speed. In Implementation performance, I discuss why GNU-R is currently far

from the theoretical maximum, and why improvements in performance happen so slowly. While it's hard to know exactly how much faster a better implementation could be, a  $>10x$  improvement in speed seems achievable. In alternative implementations, I discuss some of the promising new implementations of R, and describe one important technique they use to make R code run faster.

Beyond performance limitations due to design and implementation, it has to be said that a lot of R code is slow simply because it's poorly written. Few R users have any formal training in programming or software development. Fewer still write R code for a living. Most people use R to understand data: it's more important to get an answer quickly than to develop a system that will work in a wide variety of situations. This means that it's relatively easy to make most R code much faster, as we'll see in the following chapters.

Before we examine some of the slower parts of the R-language and GNU-R, we need to learn a little about benchmarking so that we can give our intuitions about performance a concrete foundation.

## 22.2 Microbenchmarking

A microbenchmark is a measurement of the performance of a very small piece of code, something that might take microseconds (`μs`) or nanoseconds (`ns`) to run. I'm going to use microbenchmarks to demonstrate the performance of very low-level pieces of R code, which help develop your intuition for how R works. This intuition, by-and-large, is not useful for increasing the speed of real code. The observed differences in microbenchmarks will typically be dominated by higher-order effects in real code; a deep understanding of subatomic physics is not very helpful when baking. Don't change the way you code because of these microbenchmarks. Instead wait until you've read the practical advice in the following chapters.

The best tool for microbenchmarking in R is the `microbenchmark` (<http://cran.r-project.org/web/packages/microbenchmark/>) package. It provides very precise timings, making it possible to compare operations that only take a tiny amount of time. For example, the following code compares the speed of two ways of computing a square root.

```
library(microbenchmark)

x <- runif(100)
microbenchmark(
 sqrt(x),
 x ^ 0.5
)
#> Unit: nanoseconds
#> expr min lq mean median uq max neval
#> sqrt(x) 930 1,180 1452 1,330 1,500 8,940 100
#> x ^ 0.5 9,300 9,620 10624 9,920 10,800 49,400 100
```

By default, `microbenchmark()` runs each expression 100 times (controlled by the `times` parameter). In the process, it also randomises the order of the expressions. It summarises the results with a minimum (`min`), lower quartile (`lq`), median, upper quartile (`uq`), and maximum (`max`). Focus on the median, and use the upper and lower quartiles (`lq` and `uq`) to get a feel for the variability. In this example, you can see that using the special purpose `sqrt()` function is faster than the general exponentiation operator.

As with all microbenchmarks, pay careful attention to the units: here, each computation takes about 1,000 ns, 1,000 billionths of a second. To help calibrate the impact of a microbenchmark on run time, it's useful to think about how many times a function needs to run before it takes a second. If a microbenchmark takes:

- 1 ms, then one thousand calls takes a second
- 1  $\mu$ s, then one million calls takes a second
- 1 ns, then one billion calls takes a second

The `sqrt()` function takes about 1,000 ns, or 1  $\mu$ s, to compute the square root of 100 numbers. That means if you repeated the operation a million times, it would take 1 s. So changing the way you compute the square root is unlikely to significantly affect real code.

### 22.2.1 Exercises

- Instead of using `microbenchmark()`, you could use the built-in function `system.time()`. But `system.time()` is much less precise, so you'll need to repeat each operation many times with a loop, and then divide to find the average time of each operation, as in the code below.

```
n <- 1e6
system.time(for (i in 1:n) sqrt(x)) / n
system.time(for (i in 1:n) x ^ 0.5) / n
```

How do the estimates from `system.time()` compare to those from `microbenchmark()`? Why are they different?

- Here are two other ways to compute the square root of a vector. Which do you think will be fastest? Which will be slowest? Use microbenchmarking to test your answers.

```
x ^ (1 / 2)
exp(log(x) / 2)
```

- Use microbenchmarking to rank the basic arithmetic operators (+, -, \*, /, and  $\wedge$ ) in terms of their speed. Visualise the results. Compare the speed of arithmetic on integers vs. doubles.
- You can change the units in which the microbenchmark results are expressed with the `unit` parameter. Use `unit = "eps"` to show the number of evaluations needed to take 1 second. Repeat the benchmarks above with the eps unit. How does this change your intuition for performance?

## 22.3 Language performance

In this section, I'll explore three trade-offs that limit the performance of the R-language: extreme dynamism, name lookup with mutable environments, and lazy evaluation of function arguments. I'll illustrate each trade-off with a microbenchmark, showing how it slows GNU-R down. I benchmark GNU-R because you can't benchmark the R-language (it can't run code). This means that the results are only suggestive of the cost of these design decisions, but are nevertheless useful. I've picked these three examples to illustrate some of the trade-offs that are key to language design: the designer must balance speed, flexibility, and ease of implementation.

If you'd like to learn more about the performance characteristics of the R-language and how they affect real code, I highly recommend “Evaluating the Design of the R Language” (<http://r.cs.purdue.edu/pub/ecoop12.pdf>) by Floreal Morandat, Brandon Hill, Leo Osvald, and Jan Vitek. It uses a powerful methodology that combines a modified R interpreter and a wide set of code found in the wild.

### 22.3.1 Extreme dynamism

R is an extremely dynamic programming language. Almost anything can be modified after it is created. To give just a few examples, you can:

- Change the body, arguments, and environment of functions.
- Change the S4 methods for a generic.
- Add new fields to an S3 object, or even change its class.
- Modify objects outside of the local environment with `<->`.

Pretty much the only things you can't change are objects in sealed namespaces, which are created when you load a package.

The advantage of dynamism is that you need minimal upfront planning. You can change your mind at any time, iterating your way to a solution without having to start afresh. The disadvantage of dynamism is that it's difficult to predict exactly what will happen with a given function call. This is a problem because the easier it is to predict what's going to happen, the easier it is for an interpreter or compiler to make an optimisation. (If you'd like more details, Charles Nutter expands on this idea at On Languages, VMs, Optimization, and the Way of the World (<http://blog.headius.com/2013/05/on-languages-vms-optimization-and-way.html>).) If an interpreter can't predict what's going to happen, it has to consider many options before it finds the right one. For example, the following loop is slow in R, because R doesn't know that `x` is always an integer. That means R has to look for the right `+` method (i.e., is it adding doubles, or integers?) in every iteration of the loop.

```
x <- 0L
for (i in 1:1e6) {
 x <- x + 1
}
```

The cost of finding the right method is higher for non-primitive functions. The following microbenchmark illustrates the cost of method dispatch for S3, S4, and RC. I create a generic and a method for each OO system, then call the generic and see how long it takes to find and call the method. I also time how long it takes to call the bare function for comparison.

```
f <- function(x) NULL

s3 <- function(x) UseMethod("s3")
s3.integer <- f

A <- setClass("A", representation(a = "list"))
setGeneric("s4", function(x) standardGeneric("s4"))
setMethod(s4, "A", f)

B <- setRefClass("B", methods = list(rc = f))

a <- A()
b <- B$new()

microbenchmark(
 fun = f(),
 S3 = s3(1L),
 S4 = s4(a),
 RC = b$rc()
)
#> Unit: nanoseconds
#> expr min lq mean median uq max neval
#> fun 181 204 399 227 237 15,200 100
#> S3 985 1,100 10375 1,200 1,270 885,000 100
#> S4 12,000 12,400 28281 12,900 14,400 912,000 100
#> RC 8,200 8,500 41241 8,700 9,130 3,210,000 100
```

The bare function takes about 200 ns. S3 method dispatch takes an additional 1,000 ns; S4 dispatch, 10,000 ns; and RC dispatch, 8,000 ns. S3 and S4 method dispatch are expensive because R must search for the right method every time the generic is called; it might have changed between this call and the last. R could do better by caching methods between calls, but caching is hard to do correctly and a notorious source of bugs.

### 22.3.2 Name lookup with mutable environments

It's surprisingly difficult to find the value associated with a name in the R-language. This is due to combination of lexical scoping and extreme dynamism. Take the following example. Each time we print `a` it comes from a different environment:

```
a <- 1
f <- function() {
 g <- function() {
 print(a)
 assign("a", 2, envir = parent.frame())
 print(a)
 a <- 3
 print(a)
 }
 g()
}
f()
#> [1] 1
#> [1] 2
#> [1] 3
```

This means that you can't do name lookup just once: you have to start from scratch each time. This problem is exacerbated by the fact that almost every operation is a lexically scoped function call. You might think the following simple function calls two functions: `+` and `^`. In fact, it calls four because `{` and `(` are regular functions in R.

```
f <- function(x, y) {
 (x + y) ^ 2
}
```

Since these functions are in the global environment, R has to look through every environment in the search path, which could easily be 10 or 20 environments. The following microbenchmark hints at the performance costs. We create four versions of `f()`, each with one more environment (containing 26 bindings) between the environment of `f()` and the base environment where `+`, `^`, `(`, and `{` are defined.

```
random_env <- function(parent = globalenv()) {
 letter_list <- setNames(as.list(runif(26)), LETTERS)
 list2env(letter_list, envir = new.env(parent = parent))
}

set_env <- function(f, e) {
 environment(f) <- e
 f
}

f2 <- set_env(f, random_env())
f3 <- set_env(f, random_env(environment(f2)))
f4 <- set_env(f, random_env(environment(f3)))

microbenchmark(
 f(1, 2),
 f2(1, 2),
 f3(1, 2),
 f4(1, 2),
 times = 10000
)
#> Unit: nanoseconds
```

```
#> expr min lq mean median uq max neval
#> f(1, 2) 366 398 769 410 453 2,330,000 10000
#> f2(1, 2) 669 706 1181 721 802 2,810,000 10000
#> f3(1, 2) 691 739 915 756 841 38,300 10000
#> f4(1, 2) 735 776 975 796 876 37,000 10000
```

Each additional environment between `f()` and the base environment makes the function slower by about 30 ns.

It might be possible to implement a caching system so that R only needs to look up the value of each name once. This is hard because there are so many ways to change the value associated with a name: `<-`, `assign()`, `eval()`, and so on. Any caching system would have to know about these functions to make sure the cache was correctly invalidated and you didn't get an out-of-date value.

Another simple fix would be to add more built-in constants that you can't override. This, for example, would mean that R always knew exactly what `+`, `-`, `{`, and `(` meant, and you wouldn't have to repeatedly look up their definitions. That would make the interpreter more complicated (because there are more special cases) and hence harder to maintain, and the language less flexible. This would change the R-language, but it would be unlikely to affect much existing code because it's such a bad idea to override functions like `{` and `(`.

### 22.3.3 Lazy evaluation overhead

In R, function arguments are evaluated lazily (as discussed in lazy evaluation and capturing expressions). To implement lazy evaluation, R uses a promise object that contains the expression needed to compute the result and the environment in which to perform the computation. Creating these objects has some overhead, so each additional argument to a function decreases its speed a little.

The following microbenchmark compares the runtime of a very simple function. Each version of the function has one additional argument. This suggests that adding an additional argument slows the function down by ~20 ns.

```
f0 <- function() NULL
f1 <- function(a = 1) NULL
f2 <- function(a = 1, b = 1) NULL
f3 <- function(a = 1, b = 2, c = 3) NULL
f4 <- function(a = 1, b = 2, c = 4, d = 4) NULL
f5 <- function(a = 1, b = 2, c = 4, d = 4, e = 5) NULL
microbenchmark(f0(), f1(), f2(), f3(), f4(), f5(), times = 10000)
#> Unit: nanoseconds
#> expr min lq mean median uq max neval
#> f0() 143 168 240 174 180 421,000 10000
#> f1() 175 197 313 203 210 502,000 10000
#> f2() 210 228 359 234 241 474,000 10000
#> f3() 241 257 458 262 270 882,000 10000
#> f4() 270 287 485 294 307 545,000 10000
#> f5() 301 317 534 323 339 518,000 10000
```

In most other programming languages there is little overhead for adding extra arguments. Many compiled languages will even warn you if arguments are never used (like in the above example), and automatically remove them from the function.

### 22.3.4 Exercises

1. `scan()` has the most arguments (21) of any base function. About how much time does it take to make 21 promises each time `scan` is called? Given a simple input (e.g., `scan(text = "1 2 3", quiet = TRUE)`) what proportion of the total run time is due to creating those promises?
2. Read “Evaluating the Design of the R Language” (<http://r.cs.purdue.edu/pub/ecoop12.pdf>). What other aspects of the R-language slow it down? Construct microbenchmarks to illustrate.
3. How does the performance of S3 method dispatch change with the length of the class vector? How does performance of S4 method dispatch change with number of superclasses? How about RC?
4. What is the cost of multiple inheritance and multiple dispatch on S4 method dispatch?
5. Why is the cost of name lookup less for functions in the base package?

## 22.4 Implementation performance

The design of the R language limits its maximum theoretical performance, but GNU-R is currently nowhere near that maximum. There are many things that can (and will) be done to improve performance. This section discusses some aspects of GNU-R that are slow not because of their definition, but because of their implementation.

R is over 20 years old. It contains nearly 800,000 lines of code (about 45% C, 19% R, and 17% Fortran). Changes to base R can only be made by members of the R Core Team (or R-core for short). Currently R-core has twenty members (<http://www.r-project.org/contributors.html>), but only six are active in day-to-day development. No one on R-core works full time on R. Most are statistics professors who can only spend a relatively small amount of their time on R. Because of the care that must be taken to avoid breaking existing code, R-core tends to be very conservative about accepting new code. It can be frustrating to see R-core reject proposals that would improve performance. However, the overriding concern for R-core is not to make R fast, but to build a stable platform for data analysis and statistics.

Below, I'll show two small, but illustrative, examples of parts of R that are currently slow but could, with some effort, be made faster. They are not critical parts of base R, but they have been sources of frustration for me in the past. As with all microbenchmarks, these won't affect the performance of most code, but can be important for special cases.

### 22.4.1 Extracting a single value from a data frame

The following microbenchmark shows five ways to access a single value (the number in the bottom-right corner) from the built-in `mtcars` dataset. The variation in performance is startling: the slowest method takes 30x longer than the fastest. There's no reason that there has to be such a huge difference in performance. It's simply that no one has had the time to fix it.

```
microbenchmark(
 "[32, 11]" = mtcars[32, 11],
 "$carb[32]" = mtcars$carb[32],
 "[[c(11, 32)]]" = mtcars[[c(11, 32)]],
 "[[11]][32]" = mtcars[[11]][32],
 ".subset2" = .subset2(mtcars, 11)[32]
)
#> Unit: nanoseconds
#> expr min lq mean median uq max neval
#> [32, 11] 9,440 9,820 11034 10,100 10,500 53,200 100
#> $carb[32] 4,920 5,310 6221 5,540 6,000 21,600 100
```

```
#> [[c(11, 32)]] 4,180 4,440 9510 4,610 4,990 472,000 100
#> [[11]][32] 3,950 4,190 4967 4,380 4,810 37,500 100
#> .subset2 269 336 416 360 382 2,740 100
```

## 22.4.2 `ifelse()`, `pmin()`, and `pmax()`

Some base functions are known to be slow. For example, take the following three implementations of `squish()`, a function that ensures that the smallest value in a vector is at least `a` and its largest value is at most `b`. The first implementation, `squish_ife()`, uses `ifelse()`. `ifelse()` is known to be slow because it is relatively general and must evaluate all arguments fully. The second implementation, `squish_p()`, uses `pmin()` and `pmax()`. Because these two functions are so specialised, one might expect that they would be fast. However, they're actually rather slow. This is because they can take any number of arguments and they have to do some relatively complicated checks to determine which method to use. The final implementation uses basic subassignment.

```
squish_ife <- function(x, a, b) {
 ifelse(x <= a, a, ifelse(x >= b, b, x))
}

squish_p <- function(x, a, b) {
 pmax(pmin(x, b), a)
}

squish_in_place <- function(x, a, b) {
 x[x <= a] <- a
 x[x >= b] <- b
 x
}

x <- runif(100, -1.5, 1.5)
microbenchmark(
 squish_ife = squish_ife(x, -1, 1),
 squish_p = squish_p(x, -1, 1),
 squish_in_place = squish_in_place(x, -1, 1),
 unit = "us"
)
#> Unit: microseconds
#> expr min lq mean median uq max neval
#> squish_ife 19.7 22.40 53.6 28.50 33.30 2,380 100
#> squish_p 12.3 12.90 37.4 13.30 15.30 1,560 100
#> squish_in_place 2.9 3.17 33.8 3.72 4.58 2,960 100
```

Using `pmin()` and `pmax()` is about 2x faster than `ifelse()`, and using subsetting directly is about 4x as fast again. We can often do even better by using C++. The following example compares the best R implementation to a relatively simple, if verbose, implementation in C++. Even if you've never used C++, you should still be able to follow the basic strategy: loop over every element in the vector and perform a different action depending on whether or not the value is less than `a` and/or greater than `b`.

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
NumericVector squish_cpp(NumericVector x, double a, double b) {
 int n = x.length();
 NumericVector out(n);
```

```

for (int i = 0; i < n; ++i) {
 double xi = x[i];
 if (xi < a) {
 out[i] = a;
 } else if (xi > b) {
 out[i] = b;
 } else {
 out[i] = xi;
 }
}

return out;
}

```

(You'll learn how to access this C++ code from R in Rcpp.)

```

microbenchmark(
 squish_in_place = squish_in_place(x, -1, 1),
 squish_cpp = squish_cpp(x, -1, 1),
 unit = "us"
)
#> Unit: microseconds
#> expr min lq mean median uq max neval
#> squish_in_place 3.60 4.26 5.07 4.72 5.11 29.2 100
#> squish_cpp 2.63 2.97 17.50 3.20 3.57 1,370.0 100

```

The C++ implementation is around 1x faster than the best pure R implementation.

### 22.4.3 Exercises

1. The performance characteristics of `squish_ife()`, `squish_p()`, and `squish_in_place()` vary considerably with the size of `x`. Explore the differences. Which sizes lead to the biggest and smallest differences?
2. Compare the performance costs of extracting an element from a list, a column from a matrix, and a column from a data frame. Do the same for rows.

## 22.5 Alternative R implementations

There are some exciting new implementations of R. While they all try to stick as closely as possible to the existing language definition, they improve speed by using ideas from modern interpreter design. The four most mature open-source projects are:

- pqR (<http://www.pqr-project.org/>) (pretty quick R) by Radford Neal. Built on top of R 2.15.0, it fixes many obvious performance issues, and provides better memory management and some support for automatic multithreading.
- Renjin (<http://www.renjin.org/>) by BeDataDriven. Renjin uses the Java virtual machine, and has an extensive test suite (<http://packages.renjin.org/>).
- FastR (<https://github.com/allr/fastr>) by a team from Purdue. FastR is similar to Renjin, but it makes more ambitious optimisations and is somewhat less mature.

- Riposte (<https://github.com/jtalbot/riposte>) by Justin Talbot and Zachary DeVito. Riposte is experimental and ambitious. For the parts of R it implements, it is extremely fast. Riposte is described in more detail in Riposte: A Trace-Driven Compiler and Parallel VM for Vector Code in R (<http://www.justintalbot.com/wp-content/uploads/2012/10/pact080talbot.pdf>).

These are roughly ordered from most practical to most ambitious. Another project, CXXR (<http://www.cs.kent.ac.uk/projects/cxxr/>) by Andrew Runnalls, does not provide any performance improvements. Instead, it aims to refactor R's internal C code in order to build a stronger foundation for future development, to keep behaviour identical to GNU-R, and to create better, more extensible documentation of its internals.

R is a huge language and it's not clear whether any of these approaches will ever become mainstream. It's a hard task to make an alternative implementation run all R code in the same way as GNU-R. Can you imagine having to reimplement every function in base R to be not only faster, but also to have exactly the same documented bugs? However, even if these implementations never make a dent in the use of GNU-R, they still provide benefits:

- Simpler implementations make it easy to validate new approaches before porting to GNU-R.
- Knowing which aspects of the language can be changed with minimal impact on existing code and maximal impact on performance can help to guide us to where we should direct our attention.
- Alternative implementations put pressure on the R-core to incorporate performance improvements.

One of the most important approaches that pqR, Renjin, FastR, and Riposte are exploring is the idea of deferred evaluation. As Justin Talbot, the author of Riposte, points out: "for long vectors, R's execution is completely memory bound. It spends almost all of its time reading and writing vector intermediates to memory". If we could eliminate these intermediate vectors, we could improve performance and reduce memory usage.

The following example shows a very simple example of how deferred evaluation can help. We have three vectors, `x`, `y`, `z`, each containing 1 million elements, and we want to find the sum of `x + y` where `z` is TRUE. (This represents a simplification of a pretty common sort of data analysis question.)

```
x <- runif(1e6)
y <- runif(1e6)
z <- sample(c(T, F), 1e6, rep = TRUE)

sum((x + y)[z])
```

In R, this creates two big temporary vectors: `x + y`, 1 million elements long, and `(x + y)[z]`, about 500,000 elements long. This means you need to have extra memory available for the intermediate calculation, and you have to shuttle the data back and forth between the CPU and memory. This slows computation down because the CPU can't work at maximum efficiency if it's always waiting for more data to come in.

However, if we rewrote the function using a loop in a language like C++, we only need one intermediate value: the sum of all the values we've seen:

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
double cond_sum_cpp(NumericVector x, NumericVector y,
 LogicalVector z) {
 double sum = 0;
 int n = x.length();

 for(int i = 0; i < n; i++) {
 if (!z[i]) continue;
 sum += x[i] + y[i];
 }
}
```

```

}

return sum;
}

cond_sum_r <- function(x, y, z) {
 sum((x + y)[z])
}

microbenchmark(
 cond_sum_cpp = cond_sum_cpp(x, y, z),
 cond_sum_r = cond_sum_r(x, y, z),
 unit = "ms"
)
#> Unit: milliseconds
#> expr min lq mean median uq max neval
#> cond_sum_cpp 5.21 5.24 5.28 5.25 5.29 6.7 100
#> cond_sum_r 12.30 13.70 15.47 14.20 14.50 146.0 100

```

On my computer, this approach is about 3x faster than the vectorised R equivalent, which is already pretty fast.

The goal of deferred evaluation is to perform this transformation automatically, so you can write concise R code and have it automatically translated into efficient machine code. Sophisticated translators can also figure out how to make the most of multiple cores. In the above example, if you have four cores, you could split `x`, `y`, and `z` into four pieces performing the conditional sum on each core, then adding together the four individual results. Deferred evaluation can also work with for loops, automatically discovering operations that can be vectorised.

This chapter has discussed some of the fundamental reasons that R is slow. The following chapters will give you the tools to do something about it when it impacts your code.



# Chapter 23

## Optimising code

### 23.1 Introduction

“Programmers waste enormous amounts of time thinking about, or worrying about, the speed of noncritical parts of their programs, and these attempts at efficiency actually have a strong negative impact when debugging and maintenance are considered.”

— Donald Knuth.

Optimising code to make it run faster is an iterative process:

1. Find the biggest bottleneck (the slowest part of your code).
2. Try to eliminate it (you may not succeed but that’s ok).
3. Repeat until your code is “fast enough.”

This sounds easy, but it’s not.

Even experienced programmers have a hard time identifying bottlenecks in their code. Instead of relying on your intuition, you should **profile** your code: use realistic inputs and measure the run-time of each individual operation. Only once you’ve identified the most important bottlenecks can you attempt to eliminate them. It’s difficult to provide general advice on improving performance, but I try my best with six techniques that can be applied in many situations. I’ll also suggest a general strategy for performance optimisation that helps ensure that your faster code will still be correct code.

It’s easy to get caught up in trying to remove all bottlenecks. Don’t! Your time is valuable and is better spent analysing your data, not eliminating possible inefficiencies in your code. Be pragmatic: don’t spend hours of your time to save seconds of computer time. To enforce this advice, you should set a goal time for your code and optimise only up to that goal. This means you will not eliminate all bottlenecks. Some you will not get to because you’ve met your goal. Others you may need to pass over and accept either because there is no quick and easy solution or because the code is already well optimised and no significant improvement is possible. Accept these possibilities and move on to the next candidate.

### Outline

- Measuring performance describes how to find the bottlenecks in your code using line profiling.
- Improving performance outlines seven general strategies for improving the performance of your code.
- Code organisation teaches you how to organise your code to make optimisation as easy, and bug free, as possible.

- Already solved reminds you to look for existing solutions.
- Do as little as possible emphasises the importance of being lazy: often the easiest way to make a function faster is to let it to do less work.
- Vectorise concisely defines vectorisation, and shows you how to make the most of built-in functions.
- Avoid copies discusses the performance perils of copying data.
- Byte code compilation shows you how to take advantage of R’s byte code compiler.
- Case study: t-test pulls all the pieces together into a case study showing how to speed up repeated t-tests by ~1000x.
- Parallelise teaches you how to use parallelisation to spread computation across all the cores in your computer.
- Other techniques finishes the chapter with pointers to more resources that will help you write fast code.

## Prerequisites

In this chapter we’ll be using the `lineprof` package to understand the performance of R code. Get it with:

```
devtools::install_github("hadley/lineprof")
```

## 23.2 Measuring performance

To understand performance, you use a profiler. There are a number of different types of profilers. R uses a fairly simple type called a sampling or statistical profiler. A sampling profiler stops the execution of code every few milliseconds and records which function is currently executing (along with which function called that function, and so on). For example, consider `f()`, below:

```
library(lineprof)
f <- function() {
 pause(0.1)
 g()
 h()
}
g <- function() {
 pause(0.1)
 h()
}
h <- function() {
 pause(0.1)
}
```

(I use `lineprof::pause()` instead of `Sys.sleep()` because `Sys.sleep()` does not appear in profiling outputs because as far as R can tell, it doesn’t use up any computing time.)

If we profiled the execution of `f()`, stopping the execution of code every 0.1 s, we’d see a profile like below. Each line represents one “tick” of the profiler (0.1 s in this case), and function calls are nested with `>`. It shows that the code spends 0.1 s running `f()`, then 0.2 s running `g()`, then 0.1 s running `h()`.

```
f()
f() > g()
f() > g() > h()
```

```
f() > h()
```

If we actually profile `f()`, using the code below, we're unlikely to get such a clear result.

```
tmp <- tempfile()
Rprof(tmp, interval = 0.1)
f()
Rprof(NULL)
```

That's because profiling is hard to do accurately without slowing your code down by many orders of magnitude. The compromise that `RProf()` makes, sampling, only has minimal impact on the overall performance, but is fundamentally stochastic. There's some variability in both the accuracy of the timer and in the time taken by each operation, so each time you profile you'll get a slightly different answer. Fortunately, pinpoint accuracy is not needed to identify the slowest parts of your code.

Rather than focussing on individual calls, we'll visualise aggregates using the `lineprof` package. There are a number of other options, like `summaryRprof()`, the `proftools` package, and the `profr` package, but these tools are beyond the scope of this book. I wrote the `lineprof` package as a simpler way to visualise profiling data. As the name suggests, the fundamental unit of analysis in `lineprof()` is a line of code. This makes `lineprof` less precise than the alternatives (because a line of code can contain multiple function calls), but it's easier to understand the context.

To use `lineprof`, we first save the code in a file and `source()` it. Here `profiling-example.R` contains the definition of `f()`, `g()`, and `h()`. Note that you *must* use `source()` to load the code. This is because `lineprof` uses `srcrefs` to match up the code to the profile, and the needed `srcrefs` are only created when you load code from disk. We then use `lineprof()` to run our function and capture the timing output. Printing this object shows some basic information. For now, we'll just focus on the `time` column which estimates how long each line took to run and the `ref` column which tells us which line of code was run. The estimates aren't perfect, but the ratios look about right.

```
library(lineprof)
source("profiling-example.R")
l <- lineprof(f())
l
#> time alloc release dups ref src
#> 1 0.074 0.001 0 0 profiling.R#2 f/pause
#> 2 0.143 0.002 0 0 profiling.R#3 f/g
#> 3 0.071 0.000 0 0 profiling.R#4 f/h
```

`lineprof` provides some functions to navigate through this data structure, but they're a bit clumsy. Instead, we'll start an interactive explorer using the `shiny` package. `shine(l)` will open a new web page (or if you're using RStudio, a new pane) that shows your source code annotated with information about how long each line took to run. `shine()` starts a shiny app which "blocks" your R session. To exit, you'll need to stop the process using escape or `ctrl + c`.

#	Source code	t	r	a	d
1	f <- function() {				
2	pause(0.1)				
3	g()				
4	h()				
5	}				
6	g <- function() {				
7	pause(0.1)				
8	h()				
9	}				
10	h <- function() {				
11	pause(0.1)				
12	}				

The **t** column visualises how much time is spent on each line. (You'll learn about the other columns in memory profiling.) While not precise, it allows you to spot bottlenecks, and you can get precise numbers by hovering over each bar. This shows that twice as much time is spent on `g()` as on `h()`, so it would make sense to drill down into `g()` for more details. To do so, click `g()`:

#	Source code	t	r	a	d
1	f <- function() {				
2	pause(0.1)				
3	g()				
4	h()				
5	}				
6	g <- function() {				
7	pause(0.1)				
8	h()				
9	}				
10	h <- function() {				
11	pause(0.1)				
12	}				

Then `h()`:

#	Source code	t	r	a	d
1	f <- function() {				
2	pause(0.1)				
3	g()				
4	h()				
5	}				
6	g <- function() {				
7	pause(0.1)				
8	h()				
9	}				
10	h <- function() {				
11	pause(0.1)				
12	}				

This technique should allow you to quickly identify the major bottlenecks in your code.

### 23.2.1 Limitations

There are some other limitations to profiling:

- Profiling does not extend to C code. You can see if your R code calls C/C++ code but not what functions are called inside of your C/C++ code. Unfortunately, tools for profiling compiled code are beyond the scope of this book (i.e., I have no idea how to do it).
- Similarly, you can't see what's going on inside primitive functions or byte code compiled code.
- If you're doing a lot of functional programming with anonymous functions, it can be hard to figure out exactly which function is being called. The easiest way to work around this is to name your functions.
- Lazy evaluation means that arguments are often evaluated inside another function. For example, in the following code, profiling would make it seem like `i()` was called by `j()` because the argument isn't evaluated until it's needed by `j()`.

```
i <- function() {
 pause(0.1)
 10
}
j <- function(x) {
 x + 10
}
j(i())
```

If this is confusing, you can create temporary variables to force computation to happen earlier.

## 23.3 Memory profiling with lineprof

`mem_change()` captures the net change in memory when running a block of code. Sometimes, however, we may want to measure incremental change. One way to do this is to use memory profiling to capture usage every few milliseconds. This functionality is provided by `utils::Rprof()` but it doesn't provide a very useful display of the results. Instead we'll use the `lineprof` (<https://github.com/hadley/lineprof>) package. It is powered by `Rprof()`, but displays the results in a more informative manner.

To demonstrate `lineprof`, we're going to explore a bare-bones implementation of `read.delim()` with only three arguments:

We'll also create a sample csv file:

```
library(ggplot2)
write.csv(diamonds, "diamonds.csv", row.names = FALSE)
```

Using `lineprof` is straightforward. `source()` the code, apply `lineprof()` to an expression, then use `shine()` to view the results. Note that you *must* use `source()` to load the code. This is because `lineprof` uses `srcrefs` to match up the code and run times. The needed `srcrefs` are only created when you load code from disk.

```
library(lineprof)

source("code/read-delim.R")
prof <- lineprof(read_delim("diamonds.csv"))
shine(prof)
```

#	Source code	t	r	a	d
1	# ---- read_delim				
2	read_delim <- function(file, header = TRUE, sep = ",") {				
3	# Determine number of fields by reading first line				
4	first <- scan(file, what = character(1), nlines = 1, se...				
5	p <- length(first)				
6					
7	# Load all fields as character vectors				
8	all <- scan(file, what = as.list(rep("character", p)), ...		█	█	
9	skip = if (header) 1 else 0, quiet = TRUE)		█		
10					
11	# Convert from strings to appropriate types (never to f...				
12	all[] <- lapply(all, type.convert, as.is = TRUE)			█	
13					
14	# Set column names				
15	if (header) {				
16	names(all) <- first				
17	} else {				
18	names(all) <- paste0("V", seq_along(all))				
19	}				
20					
21	# Convert list into data frame				
22	as.data.frame(all)		█	█	█
23	}				

`shine()` will also open a new web page (or if you're using RStudio, a new pane) that shows your source code annotated with information about memory usage. `shine()` starts a shiny app which will "block" your R session. To exit, press escape or ctrl + break.

Next to the source code, four columns provide details about the performance of the code:

- **t**, the time (in seconds) spent on that line of code (explained in measuring performance).
- **a**, the memory (in megabytes) allocated by that line of code.
- **r**, the memory (in megabytes) released by that line of code. While memory allocation is deterministic, memory release is stochastic: it depends on when the GC was run. This means that memory release only tells you that the memory released was no longer needed before this line.
- **d**, the number of vector duplications that occurred. A vector duplication occurs when R copies a vector as a result of its copy on modify semantics.

You can hover over any of the bars to get the exact numbers. In this example, looking at the allocations tells us most of the story:

- `scan()` allocates about 2.5 MB of memory, which is very close to the 2.8 MB of space that the file occupies on disk. You wouldn't expect the two numbers to be identical because R doesn't need to store the commas and because the global string pool will save some memory.
- Converting the columns allocates another 0.6 MB of memory. You'd also expect this step to free some memory because we've converted string columns into integer and numeric columns (which occupy less space), but we can't see those releases because GC hasn't been triggered yet.
- Finally, calling `as.data.frame()` on a list allocates about 1.6 megabytes of memory and performs over 600 duplications. This is because `as.data.frame()` isn't terribly efficient and ends up copying the

input multiple times. We'll discuss duplication more in the next section.

There are two downsides to profiling:

1. `read_delim()` only takes around half a second, but profiling can, at best, capture memory usage every 1 ms. This means we'll only get about 500 samples.
2. Since GC is lazy, we can never tell exactly when memory is no longer needed.

You can work around both problems by using `torture = TRUE`, which forces R to run GC after every allocation (see `gctorture()` for more details). This helps with both problems because memory is freed as soon as possible, and R runs 10–100x slower. This effectively makes the resolution of the timer greater, so that you can see smaller allocations and exactly when memory is no longer needed.

### 23.3.1 Exercises

1. When the input is a list, we can make a more efficient `as.data.frame()` by using special knowledge. A data frame is a list with class `data.frame` and `row.names` attribute. `row.names` is either a character vector or vector of sequential integers, stored in a special format created by `.set_row_names()`. This leads to an alternative `as.data.frame()`:

```
to_df <- function(x) {
 class(x) <- "data.frame"
 attr(x, "row.names") <- .set_row_names(length(x[[1]]))
 x
}
```

What impact does this function have on `read_delim()`? What are the downsides of this function?

2. Line profile the following function with `torture = TRUE`. What is surprising? Read the source code of `rm()` to figure out what's going on.

```
f <- function(n = 1e5) {
 x <- rep(1, n)
 rm(x)
}
```

## 23.4 Improving performance

“We should forget about small efficiencies, say about 97% of the time: premature optimization is the root of all evil. Yet we should not pass up our opportunities in that critical 3%. A good programmer will not be lulled into complacency by such reasoning, he will be wise to look carefully at the critical code; but only after that code has been identified.”

— Donald Knuth.

Once you've used profiling to identify a bottleneck, you need to make it faster. The following sections introduce you to a number of techniques that I've found broadly useful:

1. Look for existing solutions.
2. Do less work.
3. Vectorise.
4. Parallelise.
5. Avoid copies.
6. Byte-code compile.

A final technique is to rewrite in a faster language, like C++. That's a big topic and is covered in Rcpp.

Before we get into specific techniques, I'll first describe a general strategy and organisational style that's useful when working on performance.

## 23.5 Code organisation

There are two traps that are easy to fall into when trying to make your code faster:

1. Writing faster but incorrect code.
2. Writing code that you think is faster, but is actually no better.

The strategy outlined below will help you avoid these pitfalls.

When tackling a bottleneck, you're likely to come up with multiple approaches. Write a function for each approach, encapsulating all relevant behaviour. This makes it easier to check that each approach returns the correct result and to time how long it takes to run. To demonstrate the strategy, I'll compare two approaches for computing the mean:

```
mean1 <- function(x) mean(x)
mean2 <- function(x) sum(x) / length(x)
```

I recommend that you keep a record of everything you try, even the failures. If a similar problem occurs in the future, it'll be useful to see everything you've tried. To do this I often use R Markdown, which makes it easy to intermingle code with detailed comments and notes.

Next, generate a representative test case. The case should be big enough to capture the essence of your problem but small enough that it takes only a few seconds to run. You don't want it to take too long because you'll need to run the test case many times to compare approaches. On the other hand, you don't want the case to be too small because then results might not scale up to the real problem.

Use this test case to quickly check that all variants return the same result. An easy way to do so is with `stopifnot()` and `all.equal()`. For real problems with fewer possible outputs, you may need more tests to make sure that an approach doesn't accidentally return the correct answer. That's unlikely for the mean.

```
x <- runif(100)
stopifnot(all.equal(mean1(x), mean2(x)))
```

Finally, use the `microbenchmark` package to compare how long each variation takes to run. For bigger problems, reduce the `times` parameter so that it only takes a couple of seconds to run. Focus on the median time, and use the upper and lower quartiles to gauge the variability of the measurement.

```
microbenchmark(
 mean1(x),
 mean2(x)
)
#> Unit: nanoseconds
#> expr min lq mean median uq max neval
#> mean1(x) 3,060 3,210 14608 3,390 3,870 1,090,000 100
#> mean2(x) 657 738 24602 810 962 2,350,000 100
```

(You might be surprised by the results: `mean(x)` is considerably slower than `sum(x) / length(x)`. This is because, among other reasons, `mean(x)` makes two passes over the vector to be more numerically accurate.)

Before you start experimenting, you should have a target speed that defines when the bottleneck is no longer a problem. Setting such a goal is important because you don't want to spend valuable time over-optimising your code.

If you'd like to see this strategy in action, I've used it a few times on stackoverflow:

- <http://stackoverflow.com/questions/22515525#22518603>
- <http://stackoverflow.com/questions/22515175#22515856>
- <http://stackoverflow.com/questions/3476015#22511936>

## 23.6 Has someone already solved the problem?

Once you've organised your code and captured all the variations you can think of, it's natural to see what others have done. You are part of a large community, and it's quite possible that someone has already tackled the same problem. If your bottleneck is a function in a package, it's worth looking at other packages that do the same thing. Two good places to start are:

- CRAN task views (<http://cran.rstudio.com/web/views/>). If there's a CRAN task view related to your problem domain, it's worth looking at the packages listed there.
- Reverse dependencies of Rcpp, as listed on its CRAN page (<http://cran.r-project.org/web/packages/Rcpp>). Since these packages use C++, it's possible to find a solution to your bottleneck written in a higher performance language.

Otherwise, the challenge is describing your bottleneck in a way that helps you find related problems and solutions. Knowing the name of the problem or its synonyms will make this search much easier. But because you don't know what it's called, it's hard to search for it! By reading broadly about statistics and algorithms, you can build up your own knowledge base over time. Alternatively, ask others. Talk to your colleagues and brainstorm some possible names, then search on Google and stackoverflow. It's often helpful to restrict your search to R related pages. For Google, try rseek (<http://www.rseek.org/>). For stackoverflow, restrict your search by including the R tag, [R], in your search.

As discussed above, record all solutions that you find, not just those that immediately appear to be faster. Some solutions might be initially slower, but because they are easier to optimise they end up being faster. You may also be able to combine the fastest parts from different approaches. If you've found a solution that's fast enough, congratulations! If appropriate, you may want to share your solution with the R community. Otherwise, read on.

### 23.6.1 Exercises

1. What are faster alternatives to `lm`? Which are specifically designed to work with larger datasets?
2. What package implements a version of `match()` that's faster for repeated lookups? How much faster is it?
3. List four functions (not just those in base R) that convert a string into a date time object. What are their strengths and weaknesses?
4. How many different ways can you compute a 1d density estimate in R?
5. Which packages provide the ability to compute a rolling mean?
6. What are the alternatives to `optim()`?

## 23.7 Do as little as possible

The easiest way to make a function faster is to let it do less work. One way to do that is use a function tailored to a more specific type of input or output, or a more specific problem. For example:

- `rowSums()`, `colSums()`, `rowMeans()`, and `colMeans()` are faster than equivalent invocations that use `apply()` because they are vectorised (the topic of the next section).

- `vapply()` is faster than `sapply()` because it pre-specifies the output type.
- If you want to see if a vector contains a single value, `any(x == 10)` is much faster than `10 %in% x`. This is because testing equality is simpler than testing inclusion in a set.

Having this knowledge at your fingertips requires knowing that alternative functions exist: you need to have a good vocabulary. Start with the basics, and expand your vocab by regularly reading R code. Good places to read code are the R-help mailing list (<https://stat.ethz.ch/mailman/listinfo/r-help>) and stackoverflow (<http://stackoverflow.com/questions/tagged/r>).

Some functions coerce their inputs into a specific type. If your input is not the right type, the function has to do extra work. Instead, look for a function that works with your data as it is, or consider changing the way you store your data. The most common example of this problem is using `apply()` on a data frame. `apply()` always turns its input into a matrix. Not only is this error prone (because a data frame is more general than a matrix), it is also slower.

Other functions will do less work if you give them more information about the problem. It's always worthwhile to carefully read the documentation and experiment with different arguments. Some examples that I've discovered in the past include:

- `read.csv()`: specify known column types with `colClasses`.
- `factor()`: specify known levels with `levels`.
- `cut()`: don't generate labels with `labels = FALSE` if you don't need them, or, even better, use `findInterval()` as mentioned in the "see also" section of the documentation.
- `unlist(x, use.names = FALSE)` is much faster than `unlist(x)`.
- `interaction()`: if you only need combinations that exist in the data, use `drop = TRUE`.

Sometimes you can make a function faster by avoiding method dispatch. As we saw in (Extreme dynamism), method dispatch in R can be costly. If you're calling a method in a tight loop, you can avoid some of the costs by doing the method lookup only once:

- For S3, you can do this by calling `generic.class()` instead of `generic()`.
- For S4, you can do this by using `getMethod()` to find the method, saving it to a variable, and then calling that function.

For example, calling `mean.default()` is quite a bit faster than calling `mean()` for small vectors:

```
x <- runif(1e2)

microbenchmark(
 mean(x),
 mean.default(x)
)
#> Unit: microseconds
#> expr min lq median uq max neval
#> mean(x) 2.76 2.86 3.49 2.95 3.18 43.8 100
#> mean.default(x) 1.22 1.29 1.44 1.34 1.40 5.2 100
```

This optimisation is a little risky. While `mean.default()` is almost twice as fast, it'll fail in surprising ways if `x` is not a numeric vector. You should only use it if you know for sure what `x` is.

Knowing that you're dealing with a specific type of input can be another way to write faster code. For example, `as.data.frame()` is quite slow because it coerces each element into a data frame and then `rbind()`s them together. If you have a named list with vectors of equal length, you can directly transform it into a data frame. In this case, if you're able to make strong assumptions about your input, you can write a method that's about 20x faster than the default.

```

quickdf <- function(l) {
 class(l) <- "data.frame"
 attr(l, "row.names") <- .set_row_names(length(l[[1]]))
 l
}

l <- lapply(1:26, function(i) runif(1e3))
names(l) <- letters

microbenchmark(
 quick_df = quickdf(l),
 as.data.frame = as.data.frame(l)
)
#> Unit: microseconds
#> expr min lq mean median uq max neval
#> quick_df 6.51 9.45 49.6 16 22.1 3,300 100
#> as.data.frame 1,050.00 1,100.00 1231.1 1,140 1,190.0 5,600 100

```

Again, note the trade-off. This method is fast because it's dangerous. If you give it bad inputs, you'll get a corrupt data frame:

```

quickdf(list(x = 1, y = 1:2))
#> Warning in format.data.frame(x, digits = digits, na.encode = FALSE):
#> corrupt data frame: columns will be truncated or padded with NAs
#> x y
#> 1 1 1

```

To come up with this minimal method, I carefully read through and then rewrote the source code for `as.data.frame.list()` and `data.frame()`. I made many small changes, each time checking that I hadn't broken existing behaviour. After several hours work, I was able to isolate the minimal code shown above. This is a very useful technique. Most base R functions are written for flexibility and functionality, not performance. Thus, rewriting for your specific need can often yield substantial improvements. To do this, you'll need to read the source code. It can be complex and confusing, but don't give up!

The following example shows a progressive simplification of the `diff()` function if you only want computing differences between adjacent values. At each step, I replace one argument with a specific case, and then check to see that the function still works. The initial function is long and complicated, but by restricting the arguments I not only make it around twice as fast, I also make it easier to understand.

First, I take the code of `diff()` and reformat it to my style:

```

diff1 <- function (x, lag = 1L, differences = 1L) {
 ismat <- is.matrix(x)
 xlen <- if (ismat) dim(x)[1L] else length(x)
 if (length(lag) > 1L || length(differences) > 1L ||
 lag < 1L || differences < 1L)
 stop("'lag' and 'differences' must be integers >= 1")

 if (lag * differences >= xlen) {
 return(x[0L])
 }

 r <- unclass(x)
 i1 <- -seq_len(lag)
 if (ismat) {
 for (i in seq_len(differences)) {

```

```

 r <- r[i1, , drop = FALSE] -
 r[-nrow(r):-(nrow(r) - lag + 1L), , drop = FALSE]
 }
} else {
 for (i in seq_len(differences)) {
 r <- r[i1] - r[-length(r):-(length(r) - lag + 1L)]
 }
}
class(r) <- oldClass(x)
r
}

```

Next, I assume vector input. This allows me to remove the `is.matrix()` test and the method that uses matrix subsetting.

```

diff2 <- function (x, lag = 1L, differences = 1L) {
 xlen <- length(x)
 if (length(lag) > 1L || length(differences) > 1L ||
 lag < 1L || differences < 1L)
 stop("'lag' and 'differences' must be integers >= 1")

 if (lag * differences >= xlen) {
 return(x[0L])
 }

 i1 <- -seq_len(lag)
 for (i in seq_len(differences)) {
 x <- x[i1] - x[-length(x):-(length(x) - lag + 1L)]
 }
 x
}
diff2(cumsum(0:10))
#> [1] 1 2 3 4 5 6 7 8 9 10

```

I now assume that `differences = 1L`. This simplifies input checking and eliminates the for loop:

```

diff3 <- function (x, lag = 1L) {
 xlen <- length(x)
 if (length(lag) > 1L || lag < 1L)
 stop("'lag' must be integer >= 1")

 if (lag >= xlen) {
 return(x[0L])
 }

 i1 <- -seq_len(lag)
 x[i1] - x[-length(x):-(length(x) - lag + 1L)]
}
diff3(cumsum(0:10))
#> [1] 1 2 3 4 5 6 7 8 9 10

```

Finally I assume `lag = 1L`. This eliminates input checking and simplifies subsetting.

```

diff4 <- function (x) {
 xlen <- length(x)
 if (xlen <= 1) return(x[0L])

```

```

x[-1] - x[-xlen]
}
diff4(cumsum(0:10))
#> [1] 1 2 3 4 5 6 7 8 9 10

```

Now `diff4()` is both considerably simpler and considerably faster than `diff1()`:

```

x <- runif(100)
microbenchmark(
 diff1(x),
 diff2(x),
 diff3(x),
 diff4(x)
)
#> Unit: microseconds
#> expr min lq mean median uq max neval
#> diff1(x) 3.83 4.38 205.5 4.86 6.19 20,000.0 100
#> diff2(x) 3.13 3.68 5.0 4.09 5.27 46.3 100
#> diff3(x) 2.64 3.22 87.4 3.59 4.18 8,330.0 100
#> diff4(x) 2.03 2.49 40.0 2.76 3.23 3,700.0 100

```

You'll be able to make `diff()` even faster for this special case once you've read Rcpp.

A final example of doing less work is to use simpler data structures. For example, when working with rows from a data frame, it's often faster to work with row indices than data frames. For instance, if you wanted to compute a bootstrap estimate of the correlation between two columns in a data frame, there are two basic approaches: you can either work with the whole data frame or with the individual vectors. The following example shows that working with vectors is about twice as fast.

```

sample_rows <- function(df, i) sample.int(nrow(df), i,
 replace = TRUE)

Generate a new data frame containing randomly selected rows
boot_cor1 <- function(df, i) {
 sub <- df[sample_rows(df, i), , drop = FALSE]
 cor(subx, suby)
}

Generate new vectors from random rows
boot_cor2 <- function(df, i) {
 idx <- sample_rows(df, i)
 cor(df$x[idx], df$y[idx])
}

df <- data.frame(x = runif(100), y = runif(100))
microbenchmark(
 boot_cor1(df, 10),
 boot_cor2(df, 10)
)
#> Unit: microseconds
#> expr min lq mean median uq max neval
#> boot_cor1(df, 10) 85.0 90.0 133 95.9 105.0 2,820 100
#> boot_cor2(df, 10) 55.7 58.8 111 62.0 71.9 2,860 100

```

### 23.7.1 Exercises

- How do the results change if you compare `mean()` and `mean.default()` on 10,000 observations, rather than on 100?
- The following code provides an alternative implementation of `rowSums()`. Why is it faster for this input?

```
rowSums2 <- function(df) {
 out <- df[[1L]]
 if (ncol(df) == 1) return(out)

 for (i in 2:ncol(df)) {
 out <- out + df[[i]]
 }
 out
}

df <- as.data.frame(
 replicate(1e3, sample(100, 1e4, replace = TRUE)))
)
system.time(rowSums(df))
#> user system elapsed
#> 0.056 0.008 0.064
system.time(rowSums2(df))
#> user system elapsed
#> 0.032 0.000 0.032
```

- What's the difference between `rowSums()` and `.rowSums()`?
- Make a faster version of `chisq.test()` that only computes the chi-square test statistic when the input is two numeric vectors with no missing values. You can try simplifying `chisq.test()` or by coding from the mathematical definition ([http://en.wikipedia.org/wiki/Pearson%27s\\_chi-squared\\_test](http://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test)).
- Can you make a faster version of `table()` for the case of an input of two integer vectors with no missing values? Can you use it to speed up your chi-square test?
- Imagine you want to compute the bootstrap distribution of a sample correlation using `cor_df()` and the data in the example below. Given that you want to run this many times, how can you make this code faster? (Hint: the function has three components that you can speed up.)

```
n <- 1e6
df <- data.frame(a = rnorm(n), b = rnorm(n))

cor_df <- function(df, n) {
 i <- sample(seq(n), n, replace = TRUE)
 cor(df[i, , drop = FALSE])[2,1]
}
```

Is there a way to vectorise this procedure?

## 23.8 Vectorise

If you've used R for any length of time, you've probably heard the admonishment to “vectorise your code”. But what does that actually mean? Vectorising your code is not just about avoiding for loops, although

that's often a step. Vectorising is about taking a “whole object” approach to a problem, thinking about vectors, not scalars. There are two key attributes of a vectorised function:

- It makes many problems simpler. Instead of having to think about the components of a vector, you only think about entire vectors.
- The loops in a vectorised function are written in C instead of R. Loops in C are much faster because they have much less overhead.

Functionals stressed the importance of vectorised code as a higher level abstraction. Vectorisation is also important for writing fast R code. This doesn't mean simply using `apply()` or `lapply()`, or even `Vectorise()`. Those functions improve the interface of a function, but don't fundamentally change performance. Using vectorisation for performance means finding the existing R function that is implemented in C and most closely applies to your problem.

Vectorised functions that apply to many common performance bottlenecks include:

- `rowSums()`, `colSums()`, `rowMeans()`, and `colMeans()`. These vectorised matrix functions will always be faster than using `apply()`. You can sometimes use these functions to build other vectorised functions.

```
rowAny <- function(x) rowSums(x) > 0
rowAll <- function(x) rowSums(x) == ncol(x)
```

- Vectorised subsetting can lead to big improvements in speed. Remember the techniques behind lookup tables (lookup tables) and matching and merging by hand (matching and merging by hand). Also remember that you can use subsetting assignment to replace multiple values in a single step. If `x` is a vector, matrix or data frame then `x[is.na(x)] <- 0` will replace all missing values with 0.
- If you're extracting or replacing values in scattered locations in a matrix or data frame, subset with an integer matrix. See matrix subsetting for more details.
- If you're converting continuous values to categorical make sure you know how to use `cut()` and `findInterval()`.
- Be aware of vectorised functions like `cumsum()` and `diff()`.

Matrix algebra is a general example of vectorisation. There loops are executed by highly tuned external libraries like BLAS. If you can figure out a way to use matrix algebra to solve your problem, you'll often get a very fast solution. The ability to solve problems with matrix algebra is a product of experience. While this skill is something you'll develop over time, a good place to start is to ask people with experience in your domain.

The downside of vectorisation is that it makes it harder to predict how operations will scale. The following example measures how long it takes to use character subsetting to lookup 1, 10, and 100 elements from a list. You might expect that looking up 10 elements would take 10x as long as looking up 1, and that looking up 100 elements would take 10x longer again. In fact, the following example shows that it only takes about 9 times longer to look up 100 elements than it does to look up 1.

```
lookup <- setNames(as.list(sample(100, 26)), letters)

x1 <- "j"
x10 <- sample(letters, 10)
x100 <- sample(letters, 100, replace = TRUE)

microbenchmark(
 lookup[x1],
 lookup[x10],
 lookup[x100]
)
```

```
#> Unit: nanoseconds
#> expr min lq mean median uq max neval
#> lookup[x1] 531 570 781 672 915 2,080 100
#> lookup[x10] 1,350 1,430 1930 1,540 1,830 23,900 100
#> lookup[x100] 5,060 6,180 6777 6,650 7,150 16,900 100
```

Vectorisation won't solve every problem, and rather than torturing an existing algorithm into one that uses a vectorised approach, you're often better off writing your own vectorised function in C++. You'll learn how to do so in Rcpp.

### 23.8.1 Exercises

1. The density functions, e.g., `dnorm()`, have a common interface. Which arguments are vectorised over? What does `rnorm(10, mean = 10:1)` do?
2. Compare the speed of `apply(x, 1, sum)` with `rowSums(x)` for varying sizes of `x`.
3. How can you use `crossprod()` to compute a weighted sum? How much faster is it than the naive `sum(x * w)`?

## 23.9 Avoid copies

A pernicious source of slow R code is growing an object with a loop. Whenever you use `c()`, `append()`, `cbind()`, `rbind()`, or `paste()` to create a bigger object, R must first allocate space for the new object and then copy the old object to its new home. If you're repeating this many times, like in a for loop, this can be quite expensive. You've entered Circle 2 of the "R inferno" ([http://www.burns-stat.com/pages/Tutor/R\\_inferno.pdf](http://www.burns-stat.com/pages/Tutor/R_inferno.pdf)).

Here's a little example that shows the problem. We first generate some random strings, and then combine them either iteratively with a loop using `collapse()`, or in a single pass using `paste()`. Note that the performance of `collapse()` gets relatively worse as the number of strings grows: combining 100 strings takes almost 30 times longer than combining 10 strings.

```
random_string <- function() {
 paste(sample(letters, 50, replace = TRUE), collapse = "")
}
strings10 <- replicate(10, random_string())
strings100 <- replicate(100, random_string())

collapse <- function(xs) {
 out <- ""
 for (x in xs) {
 out <- paste0(out, x)
 }
 out
}

microbenchmark(
 loop10 = collapse(strings10),
 loop100 = collapse(strings100),
 vec10 = paste(strings10, collapse = ""),
 vec100 = paste(strings100, collapse = ""))
)
```

```
#> Unit: microseconds
#> expr min lq mean median uq max neval
#> loop10 20.50 21.10 22.78 22.00 23.50 34.3 100
#> loop100 803.00 816.00 859.76 820.00 831.00 4,030.0 100
#> vec10 5.12 5.31 6.43 5.64 6.05 60.8 100
#> vec100 40.80 41.30 43.33 41.70 43.60 59.6 100
```

Modifying an object in a loop, e.g., `x[i] <- y`, can also create a copy, depending on the class of `x`. Modification in place discusses this issue in more depth and gives you some tools to determine when you're making copies.

## 23.10 Byte code compilation

R 2.13.0 introduced a byte code compiler which can increase the speed of some code. Using the compiler is an easy way to get improvements in speed. Even if it doesn't work well for your function, you won't have invested a lot of time in the effort. The following example shows the pure R version of `lapply()` from functionals. Compiling it gives a considerable speedup, although it's still not quite as fast as the C version provided by base R.

```
lapply2 <- function(x, f, ...) {
 out <- vector("list", length(x))
 for (i in seq_along(x)) {
 out[[i]] <- f(x[[i]], ...)
 }
 out
}

lapply2_c <- compiler::cmpfun(lapply2)

x <- list(1:10, letters, c(F, T), NULL)
microbenchmark(
 lapply2(x, is.null),
 lapply2_c(x, is.null),
 lapply(x, is.null)
)
#> Unit: microseconds
#> expr min lq mean median uq max neval
#> lapply2(x, is.null) 1.80 1.86 43.29 1.99 2.28 4,100.00 100
#> lapply2_c(x, is.null) 1.80 1.84 2.15 1.94 2.28 7.12 100
#> lapply(x, is.null) 2.22 2.33 2.71 2.50 2.87 7.14 100
```

Byte code compilation really helps here, but in most cases you're more likely to get a 5-10% improvement. All base R functions are byte code compiled by default.

## 23.11 Case study: t-test

The following case study shows how to make t-tests faster using some of the techniques described above. It's based on an example in "Computing thousands of test statistics simultaneously in R" (<http://stat-computing.org/newsletter/issues/scgn-18-1.pdf>) by Holger Schwender and Tina Müller. I thoroughly recommend reading the paper in full to see the same idea applied to other tests.

Imagine we have run 1000 experiments (rows), each of which collects data on 50 individuals (columns). The first 25 individuals in each experiment are assigned to group 1 and the rest to group 2. We'll first generate some random data to represent this problem:

```
m <- 1000
n <- 50
X <- matrix(rnorm(m * n, mean = 10, sd = 3), nrow = m)
grp <- rep(1:2, each = n / 2)
```

For data in this form, there are two ways to use `t.test()`. We can either use the formula interface or provide two vectors, one for each group. Timing reveals that the formula interface is considerably slower.

```
system.time(for(i in 1:m) t.test(X[i,] ~ grp)$statistic)
#> user system elapsed
#> 0.908 0.000 0.908
system.time(
 for(i in 1:m) t.test(X[i, grp == 1], X[i, grp == 2])$statistic
)
#> user system elapsed
#> 0.204 0.000 0.201
```

Of course, a for loop computes, but doesn't save the values. We'll use `apply()` to do that. This adds a little overhead:

```
compT <- function(x, grp){
 t.test(x[grp == 1], x[grp == 2])$statistic
}
system.time(t1 <- apply(X, 1, compT, grp = grp))
#> user system elapsed
#> 0.212 0.000 0.212
```

How can we make this faster? First, we could try doing less work. If you look at the source code of `stats:::t.test.default()`, you'll see that it does a lot more than just compute the t-statistic. It also computes the p-value and formats the output for printing. We can try to make our code faster by stripping out those pieces.

```
my_t <- function(x, grp) {
 t_stat <- function(x) {
 m <- mean(x)
 n <- length(x)
 var <- sum((x - m) ^ 2) / (n - 1)

 list(m = m, n = n, var = var)
 }

 g1 <- t_stat(x[grp == 1])
 g2 <- t_stat(x[grp == 2])

 se_total <- sqrt(g1$var / g1$n + g2$var / g2$n)
 (g1$m - g2$m) / se_total
}

system.time(t2 <- apply(X, 1, my_t, grp = grp))
#> user system elapsed
#> 0.024 0.000 0.026
stopifnot(all.equal(t1, t2))
```

This gives us about a 6x speed improvement.

Now that we have a fairly simple function, we can make it faster still by vectorising it. Instead of looping over the array outside the function, we will modify `t_stat()` to work with a matrix of values. Thus, `mean()` becomes `rowMeans()`, `length()` becomes `ncol()`, and `sum()` becomes `rowSums()`. The rest of the code stays the same.

```
rowtstat <- function(X, grp){
 t_stat <- function(X) {
 m <- rowMeans(X)
 n <- ncol(X)
 var <- rowSums((X - m) ^ 2) / (n - 1)

 list(m = m, n = n, var = var)
 }

 g1 <- t_stat(X[, grp == 1])
 g2 <- t_stat(X[, grp == 2])

 se_total <- sqrt(g1$var / g1$n + g2$var / g2$n)
 (g1$m - g2$m) / se_total
}
system.time(t3 <- rowtstat(X, grp))
#> user system elapsed
#> 0.012 0.000 0.013
stopifnot(all.equal(t1, t3))
```

That's much faster! It's at least 40x faster than our previous effort, and around 1000x faster than where we started.

Finally, we could try byte code compilation. Here we'll need to use `microbenchmark()` instead of `system.time()` in order to get enough accuracy to see a difference:

```
rowtstat_bc <- compiler::cmpfun(rowtstat)

microbenchmark(
 rowtstat(X, grp),
 rowtstat_bc(X, grp),
 unit = "ms"
)
#> Unit: milliseconds
#>
#> expr min lq mean median uq max neval
#> rowtstat(X, grp) 0.514 0.543 0.708 0.821 0.837 0.911 100
#> rowtstat_bc(X, grp) 0.511 0.531 0.695 0.554 0.829 4.320 100
```

In this example, byte code compilation doesn't help at all.

## 23.12 Parallelise

Parallelisation uses multiple cores to work simultaneously on different parts of a problem. It doesn't reduce the computing time, but it saves your time because you're using more of your computer's resources. Parallel computing is a complex topic, and there's no way to cover it in depth here. Some resources I recommend are:

- *Parallel R* (<http://amzn.com/B005Z29QT4>) by Q. Ethan McCallum and Stephen Weston.
- *Parallel Computing for Data Science* (<http://amzn.com/1466587016>) by Norm Matloff.

What I want to show is a simple application of parallel computing to what are called “embarrassingly parallel problems”. An embarrassingly parallel problem is one that’s made up of many simple problems that can be solved independently. A great example of this is `lapply()` because it operates on each element independently of the others. It’s very easy to parallelise `lapply()` on Linux and the Mac because you simply substitute `mclapply()` for `lapply()`. The following code snippet runs a trivial (but slow) function on all cores of your computer.

```
library(parallel)

cores <- detectCores()
cores
#> [1] 2

pause <- function(i) {
 function(x) Sys.sleep(i)
}

system.time(lapply(1:10, pause(0.25)))
#> user system elapsed
#> 0.0 0.0 2.5
system.time(mclapply(1:10, pause(0.25), mc.cores = cores))
#> user system elapsed
#> 0.000 0.012 1.263
```

Life is a bit harder in Windows. You need to first set up a local cluster and then use `parLapply()`:

```
cluster <- makePSOCKcluster(cores)
system.time(parLapply(cluster, 1:10, function(i) Sys.sleep(i)))
#> user system elapsed
#> 0.004 0.000 40.063
```

The main difference between `mclapply()` and `makePSOCKcluster()` is that the individual processes generated by `mclapply()` inherit from the current process, while those generated by `makePSOCKcluster()` start with a fresh session. This means that most real code will need some setup. Use `clusterEvalQ()` to run arbitrary code on each cluster and load needed packages, and `clusterExport()` to copy objects in the current session to the remote sessions.

```
x <- 10
psock <- parallel::makePSOCKcluster(1L)
clusterEvalQ(psock, x)
#> Error: one node produced an error: object 'x' not found

clusterExport(psock, "x")
clusterEvalQ(psock, x)
#> [[1]]
#> [1] 10
```

There is some communication overhead with parallel computing. If the subproblems are very small, then parallelisation might hurt rather than help. It’s also possible to distribute computation over a network of computers (not just the cores on your local computer) but that’s beyond the scope of this book, because it gets increasingly complicated to balance computation and communication costs. A good place to start for more information is the high performance computing CRAN task view (<http://cran.r-project.org/web/views/HighPerformanceComputing.html>).

## 23.13 Other techniques

Being able to write fast R code is part of being a good R programmer. Beyond the specific hints in this chapter, if you want to write fast R code, you'll need to improve your general programming skills. Some ways to do this are to:

- Read R blogs (<http://www.r-bloggers.com/>) to see what performance problems other people have struggled with, and how they have made their code faster.
- Read other R programming books, like Norm Matloff's *The Art of R Programming* (<http://amzn.com/1593273843>) or Patrick Burns' *R Inferno* (<http://www.burns-stat.com/documents/books/the-r-inferno/>) to learn about common traps.
- Take an algorithms and data structure course to learn some well known ways of tackling certain classes of problems. I have heard good things about Princeton's Algorithms course (<https://www.coursera.org/course/algs4partI>) offered on Coursera.
- Read general books about optimisation like *Mature optimisation* (<http://carlos.bueno.org/optimization/mature-optimization.pdf>) by Carlos Bueno, or the *Pragmatic Programmer* (<http://amzn.com/020161622X>) by Andrew Hunt and David Thomas.

You can also reach out to the community for help. Stackoverflow can be a useful resource. You'll need to put some effort into creating an easily digestible example that also captures the salient features of your problem. If your example is too complex, few people will have the time and motivation to attempt a solution. If it's too simple, you'll get answers that solve the toy problem but not the real problem. If you also try to answer questions on stackoverflow, you'll quickly get a feel for what makes a good question.



# Chapter 24

## Rewriting R code in C++

### 24.1 Introduction

Sometimes R code just isn't fast enough. You've used profiling to figure out where your bottlenecks are, and you've done everything you can in R, but your code still isn't fast enough. In this chapter you'll learn how to improve performance by rewriting key functions in C++. This magic comes by way of the Rcpp (<http://www.rcpp.org/>) package, a fantastic tool written by Dirk Eddelbuettel and Romain Francois (with key contributions by Doug Bates, John Chambers, and JJ Allaire). Rcpp makes it very simple to connect C++ to R. While it is *possible* to write C or Fortran code for use in R, it will be painful by comparison. Rcpp provides a clean, approachable API that lets you write high-performance code, insulated from R's arcane C API.

Typical bottlenecks that C++ can address include:

- Loops that can't be easily vectorised because subsequent iterations depend on previous ones.
- Recursive functions, or problems which involve calling functions millions of times. The overhead of calling a function in C++ is much lower than that in R.
- Problems that require advanced data structures and algorithms that R doesn't provide. Through the standard template library (STL), C++ has efficient implementations of many important data structures, from ordered maps to double-ended queues.

The aim of this chapter is to discuss only those aspects of C++ and Rcpp that are absolutely necessary to help you eliminate bottlenecks in your code. We won't spend much time on advanced features like object oriented programming or templates because the focus is on writing small, self-contained functions, not big programs. A working knowledge of C++ is helpful, but not essential. Many good tutorials and references are freely available, including <http://www.learncpp.com/> and <http://www.cplusplus.com/>. For more advanced topics, the *Effective C++* series by Scott Meyers is a popular choice. You may also enjoy Dirk Eddelbuettel's *Seamless R and C++ integration with Rcpp* (<http://www.springer.com/statistics/computational+statistics/book/978-1-4614-6867-7>), which goes into much greater detail into all aspects of Rcpp.

### Outline

- Getting started with C++ teaches you how to write C++ by converting simple R functions to their C++ equivalents. You'll learn how C++ differs from R, and what the key scalar, vector, and matrix classes are called.

- Using `sourceCpp` shows you how to use `sourceCpp()` to load a C++ file from disk in the same way you use `source()` to load a file of R code.
- Attributes & other classes discusses how to modify attributes from Rcpp, and mentions some of the other important classes.
- Missing values teaches you how to work with R's missing values in C++.
- Rcpp sugar discusses Rcpp "sugar", which allows you to avoid loops in C++ and write code that looks very similar to vectorised R code.
- The STL shows you how to use some of the most important data structures and algorithms from the standard template library, or STL, built-in to C++.
- Case studies shows two real case studies where Rcpp was used to get considerable performance improvements.
- Putting Rcpp in a package teaches you how to add C++ code to a package.
- Learning more concludes the chapter with pointers to more resources to help you learn Rcpp and C++.

## Prerequisites

All examples in this chapter need version 0.10.1 or above of the `Rcpp` package. This version includes `cppFunction()` and `sourceCpp()`, which makes it very easy to connect C++ to R. Install the latest version of Rcpp from CRAN with `install.packages("Rcpp")`.

You'll also need a working C++ compiler. To get it:

- On Windows, install Rtools (<http://cran.r-project.org/bin/windows/Rtools/>).
- On Mac, install Xcode from the app store.
- On Linux, `sudo apt-get install r-base-dev` or similar.

## 24.2 Getting started with C++

`cppFunction()` allows you to write C++ functions in R:

```
library(Rcpp)
cppFunction('int add(int x, int y, int z) {
 int sum = x + y + z;
 return sum;
}')
add works like a regular R function
add
#> function (x, y, z)
#> .Call(<pointer: 0x7f98c1878f60>, x, y, z)
add(1, 2, 3)
#> [1] 6
```

When you run this code, Rcpp will compile the C++ code and construct an R function that connects to the compiled C++ function. We're going to use this simple interface to learn how to write C++. C++ is a large language, and there's no way to cover it all in just one chapter. Instead, you'll get the basics so that you can start writing useful functions to address bottlenecks in your R code.

The following sections will teach you the basics by translating simple R functions to their C++ equivalents. We'll start simple with a function that has no inputs and a scalar output, and then get progressively more complicated:

- Scalar input and scalar output
- Vector input and scalar output
- Vector input and vector output
- Matrix input and vector output

### 24.2.1 No inputs, scalar output

Let's start with a very simple function. It has no arguments and always returns the integer 1:

```
one <- function() 1L
```

The equivalent C++ function is:

```
int one() {
 return 1;
}
```

We can compile and use this from R with `cppFunction`

```
cppFunction('int one() {
 return 1;
}')
```

This small function illustrates a number of important differences between R and C++:

- The syntax to create a function looks like the syntax to call a function; you don't use assignment to create functions as you do in R.
- You must declare the type of output the function returns. This function returns an `int` (a scalar integer). The classes for the most common types of R vectors are: `NumericVector`, `IntegerVector`, `CharacterVector`, and `LogicalVector`.
- Scalars and vectors are different. The scalar equivalents of numeric, integer, character, and logical vectors are: `double`, `int`, `String`, and `bool`.
- You must use an explicit `return` statement to return a value from a function.
- Every statement is terminated by a `;`.

### 24.2.2 Scalar input, scalar output

The next example function implements a scalar version of the `sign()` function which returns 1 if the input is positive, and -1 if it's negative:

```
signR <- function(x) {
 if (x > 0) {
 1
 } else if (x == 0) {
 0
 } else {
 -1
 }
}

cppFunction('int signC(int x) {
 if (x > 0) {
 return 1;
 } else if (x == 0) {
```

```

 return 0;
} else {
 return -1;
}
}'')

```

In the C++ version:

- We declare the type of each input in the same way we declare the type of the output. While this makes the code a little more verbose, it also makes it very obvious what type of input the function needs.
- The `if` syntax is identical — while there are some big differences between R and C++, there are also lots of similarities! C++ also has a `while` statement that works the same way as R's. As in R you can use `break` to exit the loop, but to skip one iteration you need to use `continue` instead of `next`.

### 24.2.3 Vector input, scalar output

One big difference between R and C++ is that the cost of loops is much lower in C++. For example, we could implement the `sum` function in R using a loop. If you've been programming in R a while, you'll probably have a visceral reaction to this function!

```

sumR <- function(x) {
 total <- 0
 for (i in seq_along(x)) {
 total <- total + x[i]
 }
 total
}

```

In C++, loops have very little overhead, so it's fine to use them. In STL, you'll see alternatives to `for` loops that more clearly express your intent; they're not faster, but they can make your code easier to understand.

```

cppFunction('double sumC(NumericVector x) {
 int n = x.size();
 double total = 0;
 for(int i = 0; i < n; ++i) {
 total += x[i];
 }
 return total;
}')

```

The C++ version is similar, but:

- To find the length of the vector, we use the `.size()` method, which returns an integer. C++ methods are called with `.` (i.e., a full stop).
- The `for` statement has a different syntax: `for(init; check; increment)`. This loop is initialised by creating a new variable called `i` with value 0. Before each iteration we check that `i < n`, and terminate the loop if it's not. After each iteration, we increment the value of `i` by one, using the special prefix operator `++` which increases the value of `i` by 1.
- In C++, vector indices start at 0. I'll say this again because it's so important: **IN C++, VECTOR INDICES START AT 0!** This is a very common source of bugs when converting R functions to C++.
- Use `=` for assignment, not `<-`.

- C++ provides operators that modify in-place: `total += x[i]` is equivalent to `total = total + x[i]`. Similar in-place operators are `-=`, `*=`, and `/=`.

This is a good example of where C++ is much more efficient than R. As shown by the following microbenchmark, `sumC()` is competitive with the built-in (and highly optimised) `sum()`, while `sumR()` is several orders of magnitude slower.

```
x <- runif(1e3)
microbenchmark(
 sum(x),
 sumC(x),
 sumR(x)
)
#> Unit: microseconds
#> expr min lq mean median uq max neval
#> sum(x) 1.34 1.38 1.51 1.40 1.49 5.72 100
#> sumC(x) 3.15 3.44 14.14 3.68 4.22 996.00 100
#> sumR(x) 43.10 43.40 86.03 43.60 44.30 4,100.00 100
```

#### 24.2.4 Vector input, vector output

Next we'll create a function that computes the Euclidean distance between a value and a vector of values:

```
pdistR <- function(x, ys) {
 sqrt((x - ys) ^ 2)
}
```

It's not obvious that we want `x` to be a scalar from the function definition. We'd need to make that clear in the documentation. That's not a problem in the C++ version because we have to be explicit about types:

```
cppFunction('NumericVector pdistC(double x, NumericVector ys) {
 int n = ys.size();
 NumericVector out(n);

 for(int i = 0; i < n; ++i) {
 out[i] = sqrt(pow(ys[i] - x, 2.0));
 }
 return out;
}')
```

This function introduces only a few new concepts:

- We create a new numeric vector of length `n` with a constructor: `NumericVector out(n)`. Another useful way of making a vector is to copy an existing one: `NumericVector zs = clone(ys)`.
- C++ uses `pow()`, not `^`, for exponentiation.

Note that because the R version is fully vectorised, it's already going to be fast. On my computer, it takes around 8 ms with a 1 million element `y` vector. The C++ function is twice as fast, ~4 ms, but assuming it took you 10 minutes to write the C++ function, you'd need to run it ~150,000 times to make rewriting worthwhile. The reason why the C++ function is faster is subtle, and relates to memory management. The R version needs to create an intermediate vector the same length as `y` (`x - ys`), and allocating memory is an expensive operation. The C++ function avoids this overhead because it uses an intermediate scalar.

In the sugar section, you'll see how to rewrite this function to take advantage of Rcpp's vectorised operations so that the C++ code is almost as concise as R code.

### 24.2.5 Matrix input, vector output

Each vector type has a matrix equivalent: `NumericMatrix`, `IntegerMatrix`, `CharacterMatrix`, and `LogicalMatrix`. Using them is straightforward. For example, we could create a function that reproduces `rowSums()`:

```
cppFunction('NumericVector rowSumsC(NumericMatrix x) {
 int nrow = x.nrow(), ncol = x.ncol();
 NumericVector out(nrow);

 for (int i = 0; i < nrow; i++) {
 double total = 0;
 for (int j = 0; j < ncol; j++) {
 total += x(i, j);
 }
 out[i] = total;
 }
 return out;
}')
set.seed(1014)
x <- matrix(sample(100), 10)
rowSums(x)
#> [1] 458 558 488 458 536 537 488 491 508 528
rowSumsC(x)
#> [1] 458 558 488 458 536 537 488 491 508 528
```

The main differences:

- In C++, you subset a matrix with `()`, not `[]`.
- Use `.nrow()` and `.ncol()` *methods* to get the dimensions of a matrix.

### 24.2.6 Using sourceCpp

So far, we've used inline C++ with `cppFunction()`. This makes presentation simpler, but for real problems, it's usually easier to use stand-alone C++ files and then source them into R using `sourceCpp()`. This lets you take advantage of text editor support for C++ files (e.g., syntax highlighting) as well as making it easier to identify the line numbers in compilation errors.

Your stand-alone C++ file should have extension `.cpp`, and needs to start with:

```
#include <Rcpp.h>
using namespace Rcpp;
```

And for each function that you want available within R, you need to prefix it with:

```
// [[Rcpp::export]]
```

Note that the space is mandatory.

If you're familiar with roxygen2, you might wonder how this relates to `@export`. `Rcpp::export` controls whether a function is exported from C++ to R; `@export` controls whether a function is exported from a package and made available to the user.

You can embed R code in special C++ comment blocks. This is really convenient if you want to run some test code:

```
/** R
This is R code
*/
```

The R code is run with `source(echo = TRUE)` so you don't need to explicitly print output.

To compile the C++ code, use `sourceCpp("path/to/file.cpp")`. This will create the matching R functions and add them to your current session. Note that these functions can not be saved in a `.Rdata` file and reloaded in a later session; they must be recreated each time you restart R. For example, running `sourceCpp()` on the following file implements `mean` in C++ and then compares it to the built-in `mean()`:

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
double meanC(NumericVector x) {
 int n = x.size();
 double total = 0;

 for(int i = 0; i < n; ++i) {
 total += x[i];
 }
 return total / n;
}

/** R
library(microbenchmark)
x <- runif(1e5)
microbenchmark(
 mean(x),
 meanC(x)
)
*/
```

NB: if you run this code yourself, you'll notice that `meanC()` is much faster than the built-in `mean()`. This is because it trades numerical accuracy for speed.

For the remainder of this chapter C++ code will be presented stand-alone rather than wrapped in a call to `cppFunction`. If you want to try compiling and/or modifying the examples you should paste them into a C++ source file that includes the elements described above.

### 24.2.7 Exercises

With the basics of C++ in hand, it's now a great time to practice by reading and writing some simple C++ functions. For each of the following functions, read the code and figure out what the corresponding base R function is. You might not understand every part of the code yet, but you should be able to figure out the basics of what the function does.

```
double f1(NumericVector x) {
 int n = x.size();
 double y = 0;

 for(int i = 0; i < n; ++i) {
 y += x[i] / n;
 }
```

```

 return y;
}

NumericVector f2(NumericVector x) {
 int n = x.size();
 NumericVector out(n);

 out[0] = x[0];
 for(int i = 1; i < n; ++i) {
 out[i] = out[i - 1] + x[i];
 }
 return out;
}

bool f3(LogicalVector x) {
 int n = x.size();

 for(int i = 0; i < n; ++i) {
 if (x[i]) return true;
 }
 return false;
}

int f4(Function pred, List x) {
 int n = x.size();

 for(int i = 0; i < n; ++i) {
 LogicalVector res = pred(x[i]);
 if (res[0]) return i + 1;
 }
 return 0;
}

NumericVector f5(NumericVector x, NumericVector y) {
 int n = std::max(x.size(), y.size());
 NumericVector x1 = rep_len(x, n);
 NumericVector y1 = rep_len(y, n);

 NumericVector out(n);

 for (int i = 0; i < n; ++i) {
 out[i] = std::min(x1[i], y1[i]);
 }

 return out;
}

```

To practice your function writing skills, convert the following functions into C++. For now, assume the inputs have no missing values.

1. `all()`
2. `cumprod()`, `cummin()`, `cummax()`.
3. `diff()`. Start by assuming lag 1, and then generalise for lag `n`.

4. **range**.
5. **var**. Read about the approaches you can take on wikipedia ([http://en.wikipedia.org/wiki/Algorithms\\_for\\_calculating\\_variance](http://en.wikipedia.org/wiki/Algorithms_for_calculating_variance)). Whenever implementing a numerical algorithm, it's always good to check what is already known about the problem.

## 24.3 Attributes and other classes

You've already seen the basic vector classes (`IntegerVector`, `NumericVector`, `LogicalVector`, `CharacterVector`) and their scalar (`int`, `double`, `bool`, `String`) and matrix (`IntegerMatrix`, `NumericMatrix`, `LogicalMatrix`, `CharacterMatrix`) equivalents.

All R objects have attributes, which can be queried and modified with `.attr()`. Rcpp also provides `.names()` as an alias for the name attribute. The following code snippet illustrates these methods. Note the use of `::create()`, a *class* method. This allows you to create an R vector from C++ scalar values:

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
NumericVector attrs() {
 NumericVector out = NumericVector::create(1, 2, 3);

 out.names() = CharacterVector::create("a", "b", "c");
 out.attr("my-attr") = "my-value";
 out.attr("class") = "my-class";

 return out;
}
```

For S4 objects, `.slot()` plays a similar role to `.attr()`.

### 24.3.1 Lists and data frames

Rcpp also provides classes `List` and `DataFrame`, but they are more useful for output than input. This is because lists and data frames can contain arbitrary classes but C++ needs to know their classes in advance. If the list has known structure (e.g., it's an S3 object), you can extract the components and manually convert them to their C++ equivalents with `as()`. For example, the object created by `lm()`, the function that fits a linear model, is a list whose components are always of the same type. The following code illustrates how you might extract the mean percentage error (`mpe()`) of a linear model. This isn't a good example of when to use C++, because it's so easily implemented in R, but it shows how to work with an important S3 class. Note the use of `.inherits()` and the `stop()` to check that the object really is a linear model.

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
double mpe(List mod) {
 if (!mod.inherits("lm")) stop("Input must be a linear model");

 NumericVector resid = as<NumericVector>(mod["residuals"]);
 NumericVector fitted = as<NumericVector>(mod["fitted.values"]);

 int n = resid.size();
```

```

double err = 0;
for(int i = 0; i < n; ++i) {
 err += resid[i] / (fitted[i] + resid[i]);
}
return err / n;
}

mod <- lm(mpg ~ wt, data = mtcars)
mpe(mod)
#> [1] -0.0154

```

### 24.3.2 Functions

You can put R functions in an object of type `Function`. This makes calling an R function from C++ straightforward. We first define our C++ function:

```

#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
RObject callWithOne(Function f) {
 return f(1);
}

```

Then call it from R:

```

callWithOne(function(x) x + 1)
#> [1] 2
callWithOne(paste)
#> [1] "1"

```

What type of object does an R function return? We don't know, so we use the catchall type `RObject`. An alternative is to return a `List`. For example, the following code is a basic implementation of `lapply` in C++:

```

#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
List lapply1(List input, Function f) {
 int n = input.size();
 List out(n);

 for(int i = 0; i < n; i++) {
 out[i] = f(input[i]);
 }

 return out;
}

```

Calling R functions with positional arguments is obvious:

```
f("y", 1);
```

But to use named arguments, you need a special syntax:

```
f(_["x"] = "y", _["value"] = 1);
```

### 24.3.3 Other types

There are also classes for many more specialised language objects: `Environment`, `ComplexVector`, `RawVector`, `DottedPair`, `Language`, `Promise`, `Symbol`, `WeakReference`, and so on. These are beyond the scope of this chapter and won't be discussed further.

## 24.4 Missing values

If you're working with missing values, you need to know two things:

- how R's missing values behave in C++'s scalars (e.g., `double`).
- how to get and set missing values in vectors (e.g., `NumericVector`).

### 24.4.1 Scalars

The following code explores what happens when you take one of R's missing values, coerce it into a scalar, and then coerce back to an R vector. Note that this kind of experimentation is a useful way to figure out what any operation does.

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
List scalar_missings() {
 int int_s = NA_INTEGER;
 String chr_s = NA_STRING;
 bool lgl_s = NA_LOGICAL;
 double num_s = NA_REAL;

 return List::create(int_s, chr_s, lgl_s, num_s);
}

str(scalar_missings())
#> List of 4
#> $: int NA
#> $: chr NA
#> $: logi TRUE
#> $: num NA
```

With the exception of `bool`, things look pretty good here: all of the missing values have been preserved. However, as we'll see in the following sections, things are not quite as straightforward as they seem.

#### 24.4.1.1 Integers

With integers, missing values are stored as the smallest integer. If you don't do anything to them, they'll be preserved. But, since C++ doesn't know that the smallest integer has this special behaviour, if you do anything to it you're likely to get an incorrect value: for example, `evalCpp('NA_INTEGER + 1')` gives `-2147483647`.

So if you want to work with missing values in integers, either use a length one `IntegerVector` or be very careful with your code.

#### 24.4.1.2 Doubles

With doubles, you may be able to get away with ignoring missing values and working with NaNs (not a number). This is because R's NA is a special type of IEEE 754 floating point number NaN. So any logical expression that involves a NaN (or in C++, NAN) always evaluates as FALSE:

```
evalCpp("NAN == 1")
#> [1] FALSE
evalCpp("NAN < 1")
#> [1] FALSE
evalCpp("NAN > 1")
#> [1] FALSE
evalCpp("NAN == NAN")
#> [1] FALSE
```

But be careful when combining them with boolean values:

```
evalCpp("NAN && TRUE")
#> [1] TRUE
evalCpp("NAN || FALSE")
#> [1] TRUE
```

However, in numeric contexts NaNs will propagate NAs:

```
evalCpp("NAN + 1")
#> [1] NaN
evalCpp("NAN - 1")
#> [1] NaN
evalCpp("NAN / 1")
#> [1] NaN
evalCpp("NAN * 1")
#> [1] NaN
```

#### 24.4.2 Strings

`String` is a scalar string class introduced by Rcpp, so it knows how to deal with missing values.

#### 24.4.3 Boolean

While C++'s `bool` has two possible values (`true` or `false`), a logical vector in R has three (`TRUE`, `FALSE`, and `NA`). If you coerce a length 1 logical vector, make sure it doesn't contain any missing values otherwise they will be converted to `TRUE`.

#### 24.4.4 Vectors

With vectors, you need to use a missing value specific to the type of vector, `NA_REAL`, `NA_INTEGER`, `NA_LOGICAL`, `NA_STRING`:

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
List missing_sampler() {
 return List::create(
 NumericVector::create(NA_REAL),
 IntegerVector::create(NA_INTEGER),
 LogicalVector::create(NA_LOGICAL),
 CharacterVector::create(NA_STRING));
}

str(missing_sampler())
#> List of 4
#> $: num NA
#> $: int NA
#> $: logi NA
#> $: chr NA
```

To check if a value in a vector is missing, use the class method `::is_na()`:

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
LogicalVector is_naC(NumericVector x) {
 int n = x.size();
 LogicalVector out(n);

 for (int i = 0; i < n; ++i) {
 out[i] = NumericVector::is_na(x[i]);
 }
 return out;
}

is_naC(c(NA, 5.4, 3.2, NA))
#> [1] TRUE FALSE FALSE TRUE
```

Another alternative is the sugar function `is_na()`, which takes a vector and returns a logical vector.

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
LogicalVector is_naC2(NumericVector x) {
 return is_na(x);
}

is_naC2(c(NA, 5.4, 3.2, NA))
#> [1] TRUE FALSE FALSE TRUE
```

#### 24.4.5 Exercises

1. Rewrite any of the functions from the first exercise to deal with missing values. If `na.rm` is true, ignore the missing values. If `na.rm` is false, return a missing value if the input contains any missing values.

Some good functions to practice with are `min()`, `max()`, `range()`, `mean()`, and `var()`.

2. Rewrite `cumsum()` and `diff()` so they can handle missing values. Note that these functions have slightly more complicated behaviour.

## 24.5 Rcpp sugar

Rcpp provides a lot of syntactic “sugar” to ensure that C++ functions work very similarly to their R equivalents. In fact, Rcpp sugar makes it possible to write efficient C++ code that looks almost identical to its R equivalent. If there’s a sugar version of the function you’re interested in, you should use it: it’ll be both expressive and well tested. Sugar functions aren’t always faster than a handwritten equivalent, but they will get faster in the future as more time is spent on optimising Rcpp.

Sugar functions can be roughly broken down into

- arithmetic and logical operators
- logical summary functions
- vector views
- other useful functions

### 24.5.1 Arithmetic and logical operators

All the basic arithmetic and logical operators are vectorised: `+`, `*`, `-`, `/`, `pow`, `<`, `<=`, `>`, `>=`, `==`, `!=`, `!`. For example, we could use sugar to considerably simplify the implementation of `pdistC()`.

```
pdistR <- function(x, ys) {
 sqrt((x - ys) ^ 2)
}

#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
NumericVector pdistC2(double x, NumericVector ys) {
 return sqrt(pow((x - ys), 2));
}
```

### 24.5.2 Logical summary functions

The sugar function `any()` and `all()` are fully lazy so that `any(x == 0)`, for example, might only need to evaluate one element of a vector, and return a special type that can be converted into a `bool` using `.is_true()`, `.is_false()`, or `.is_na()`. We could also use this sugar to write an efficient function to determine whether or not a numeric vector contains any missing values. To do this in R, we could use `any(is.na(x))`:

```
any_naR <- function(x) any(is.na(x))
```

However, this will do the same amount of work regardless of the location of the missing value. Here’s the C++ implementation:

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
```

```

bool any_naC(NumericVector x) {
 return is_true(any(is_na(x)));
}

x0 <- runif(1e5)
x1 <- c(x0, NA)
x2 <- c(NA, x0)

microbenchmark(
 any_naR(x0), any_naC(x0),
 any_naR(x1), any_naC(x1),
 any_naR(x2), any_naC(x2)
)
#> Unit: microseconds
#> expr min lq mean median uq max neval
#> any_naR(x0) 497.00 615.00 726.30 625.00 632.00 5,580.0 100
#> any_naC(x0) 484.00 490.00 500.67 498.00 505.00 581.0 100
#> any_naR(x1) 464.00 617.00 622.15 624.00 629.00 724.0 100
#> any_naC(x1) 486.00 489.00 510.29 498.00 503.00 1,740.0 100
#> any_naR(x2) 335.00 471.00 494.35 481.00 493.00 1,740.0 100
#> any_naC(x2) 2.07 3.21 4.65 4.27 5.37 15.2 100

```

### 24.5.3 Vector views

A number of helpful functions provide a “view” of a vector: `head()`, `tail()`, `rep_each()`, `rep_len()`, `rev()`, `seq_along()`, and `seq_len()`. In R these would all produce copies of the vector, but in Rcpp they simply point to the existing vector and override the subsetting operator (`[]`) to implement special behaviour. This makes them very efficient: for instance, `rep_len(x, 1e6)` does not have to make a million copies of `x`.

### 24.5.4 Other useful functions

Finally, there’s a grab bag of sugar functions that mimic frequently used R functions:

- Math functions: `abs()`, `acos()`, `asin()`, `atan()`, `beta()`, `ceil()`, `ceiling()`, `choose()`, `cos()`, `cosh()`, `digamma()`, `exp()`, `expm1()`, `factorial()`, `floor()`, `gamma()`, `lbeta()`, `lchoose()`, `lfactorial()`, `lgamma()`, `log()`, `log10()`, `log1p()`, `pentagamma()`, `psigamma()`, `round()`, `signif()`, `sin()`, `sinh()`, `sqrt()`, `tan()`, `tanh()`, `tetragamma()`, `trigamma()`, `trunc()`.
- Scalar summaries: `mean()`, `min()`, `max()`, `sum()`, `sd()`, and (for vectors) `var()`.
- Vector summaries: `cumsum()`, `diff()`, `pmin()`, and `pmax()`.
- Finding values: `match()`, `self_match()`, `which_max()`, `which_min()`.
- Dealing with duplicates: `duplicated()`, `unique()`.
- d/q/p/r for all standard distributions.

Finally, `noNA(x)` asserts that the vector `x` does not contain any missing values, and allows optimisation of some mathematical operations. For example, when computing the mean of a vector with no missing values, Rcpp doesn’t need to check each value is not missing when computing the sum and the length.

## 24.6 The STL

The real strength of C++ shows itself when you need to implement more complex algorithms. The standard template library (STL) provides a set of extremely useful data structures and algorithms. This section will explain some of the most important algorithms and data structures and point you in the right direction to learn more. I can't teach you everything you need to know about the STL, but hopefully the examples will show you the power of the STL, and persuade you that it's useful to learn more.

If you need an algorithm or data structure that isn't implemented in STL, a good place to look is boost (<http://www.boost.org/doc/>). Installing boost on your computer is beyond the scope of this chapter, but once you have it installed, you can use boost data structures and algorithms by including the appropriate header file with (e.g.) `#include <boost/array.hpp>`.

### 24.6.1 Using iterators

Iterators are used extensively in the STL: many functions either accept or return iterators. They are the next step up from basic loops, abstracting away the details of the underlying data structure. Iterators have three main operators:

1. Advance with `++`.
2. Get the value they refer to, or **dereference**, with `*`.
3. Compare with `==`.

For example we could re-write our sum function using iterators:

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
double sum3(NumericVector x) {
 double total = 0;

 NumericVector::iterator it;
 for(it = x.begin(); it != x.end(); ++it) {
 total += *it;
 }
 return total;
}
```

The main changes are in the for loop:

- We start at `x.begin()` and loop until we get to `x.end()`. A small optimization is to store the value of the end iterator so we don't need to look it up each time. This only saves about 2 ns per iteration, so it's only important when the calculations in the loop are very simple.
- Instead of indexing into `x`, we use the dereference operator to get its current value: `*it`.
- Notice the type of the iterator: `NumericVector::iterator`. Each vector type has its own iterator type: `LogicalVector::iterator`, `CharacterVector::iterator`, etc.

Iterators also allow us to use the C++ equivalents of the apply family of functions. For example, we could again rewrite `sum()` to use the `accumulate()` function, which takes a starting and an ending iterator, and adds up all the values in the vector. The third argument to `accumulate` gives the initial value: it's particularly important because this also determines the data type that `accumulate` uses (so we use `0.0` and not `0` so that `accumulate` uses a `double`, not an `int`). To use `accumulate()` we need to include the `<numeric>` header.

```
#include <numeric>
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
double sum4(NumericVector x) {
 return std::accumulate(x.begin(), x.end(), 0.0);
}
```

`accumulate()` (along with the other functions in `<numeric>`, like `adjacent_difference()`, `inner_product()`, and `partial_sum()`) is not that important in Rcpp because Rcpp sugar provides equivalents.

## 24.6.2 Algorithms

The `<algorithm>` header provides a large number of algorithms that work with iterators. A good reference is available at <http://www.cplusplus.com/reference/algorithm/>. For example, we could write a basic Rcpp version of `findInterval()` that takes two arguments a vector of values and a vector of breaks, and locates the bin that each `x` falls into. This shows off a few more advanced iterator features. Read the code below and see if you can figure out how it works.

```
#include <algorithm>
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
IntegerVector findInterval2(NumericVector x, NumericVector breaks) {
 IntegerVector out(x.size());

 NumericVector::iterator it, pos;
 IntegerVector::iterator out_it;

 for(it = x.begin(), out_it = out.begin(); it != x.end();
 ++it, ++out_it) {
 pos = std::upper_bound(breaks.begin(), breaks.end(), *it);
 *out_it = std::distance(breaks.begin(), pos);
 }

 return out;
}
```

The key points are:

- We step through two iterators (input and output) simultaneously.
- We can assign into a dereferenced iterator (`out_it`) to change the values in `out`.
- `upper_bound()` returns an iterator. If we wanted the value of the `upper_bound()` we could dereference it; to figure out its location, we use the `distance()` function.
- Small note: if we want this function to be as fast as `findInterval()` in R (which uses handwritten C code), we need to compute the calls to `.begin()` and `.end()` once and save the results. This is easy, but it distracts from this example so it has been omitted. Making this change yields a function that's slightly faster than R's `findInterval()` function, but is about 1/10 of the code.

It's generally better to use algorithms from the STL than hand rolled loops. In *Effective STL*, Scott Meyers gives three reasons: efficiency, correctness, and maintainability. Algorithms from the STL are written by

C++ experts to be extremely efficient, and they have been around for a long time so they are well tested. Using standard algorithms also makes the intent of your code more clear, helping to make it more readable and more maintainable.

### 24.6.3 Data structures

The STL provides a large set of data structures: `array`, `bitset`, `list`, `forward_list`, `map`, `multimap`, `multiset`, `priority_queue`, `queue`, `deque`, `set`, `stack`, `unordered_map`, `unordered_set`, `unordered_multimap`, `unordered_multiset`, and `vector`. The most important of these data structures are the `vector`, the `unordered_set`, and the `unordered_map`. We'll focus on these three in this section, but using the others is similar: they just have different performance trade-offs. For example, the `deque` (pronounced "deck") has a very similar interface to vectors but a different underlying implementation that has different performance trade-offs. You may want to try them for your problem. A good reference for STL data structures is <http://www.cplusplus.com/reference/stl/> — I recommend you keep it open while working with the STL.

Rcpp knows how to convert from many STL data structures to their R equivalents, so you can return them from your functions without explicitly converting to R data structures.

### 24.6.4 Vectors

An STL vector is very similar to an R vector, except that it grows efficiently. This makes vectors appropriate to use when you don't know in advance how big the output will be. Vectors are templated, which means that you need to specify the type of object the vector will contain when you create it: `vector<int>`, `vector<bool>`, `vector<double>`, `vector<String>`. You can access individual elements of a vector using the standard `[]` notation, and you can add a new element to the end of the vector using `.push_back()`. If you have some idea in advance how big the vector will be, you can use `.reserve()` to allocate sufficient storage.

The following code implements run length encoding (`rle()`). It produces two vectors of output: a vector of values, and a vector `lengths` giving how many times each element is repeated. It works by looping through the input vector `x` comparing each value to the previous: if it's the same, then it increments the last value in `lengths`; if it's different, it adds the value to the end of `values`, and sets the corresponding length to 1.

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
List rleC(NumericVector x) {
 std::vector<int> lengths;
 std::vector<double> values;

 // Initialise first value
 int i = 0;
 double prev = x[0];
 values.push_back(prev);
 lengths.push_back(1);

 NumericVector::iterator it;
 for(it = x.begin() + 1; it != x.end(); ++it) {
 if (prev == *it) {
 lengths[i]++;
 } else {
 values.push_back(*it);
 lengths.push_back(1);
 prev = *it;
 }
 }
}
```

```

 i++;
 prev = *it;
 }
}

return List::create(
 _["lengths"] = lengths,
 _["values"] = values
);
}
}

```

(An alternative implementation would be to replace `i` with the iterator `lengths.rbegin()` which always points to the last element of the vector. You might want to try implementing that yourself.)

Other methods of a vector are described at <http://www.cplusplus.com/reference/vector/vector/>.

## 24.6.5 Sets

Sets maintain a unique set of values, and can efficiently tell if you've seen a value before. They are useful for problems that involve duplicates or unique values (like `unique`, `duplicated`, or `in`). C++ provides both ordered (`std::set`) and unordered sets (`std::unordered_set`), depending on whether or not order matters for you. Unordered sets tend to be much faster (because they use a hash table internally rather than a tree), so even if you need an ordered set, you should consider using an unordered set and then sorting the output. Like vectors, sets are templated, so you need to request the appropriate type of set for your purpose: `unordered_set<int>`, `unordered_set<bool>`, etc. More details are available at <http://www.cplusplus.com/reference/set/set/> and [http://www.cplusplus.com/reference/unordered\\_set/unordered\\_set/](http://www.cplusplus.com/reference/unordered_set/unordered_set/).

The following function uses an unordered set to implement an equivalent to `duplicated()` for integer vectors. Note the use of `seen.insert(x[i]).second`. `insert()` returns a pair, the `.first` value is an iterator that points to element and the `.second` value is a boolean that's true if the value was a new addition to the set.

```

// [[Rcpp::plugins(cpp11)]]
#include <Rcpp.h>
#include <unordered_set>
using namespace Rcpp;

// [[Rcpp::export]]
LogicalVector duplicatedC(IntegerVector x) {
 std::unordered_set<int> seen;
 int n = x.size();
 LogicalVector out(n);

 for (int i = 0; i < n; ++i) {
 out[i] = !seen.insert(x[i]).second;
 }

 return out;
}

```

Note that unordered sets are only available in C++ 11, which means we need to use the `cpp11` plugin, `[[Rcpp::plugins(cpp11)]]`.

## 24.6.6 Map

A map is similar to a set, but instead of storing presence or absence, it can store additional data. It's useful for functions like `table()` or `match()` that need to look up a value. As with sets, there are ordered (`std::map`) and unordered (`std::unordered_map`) versions. Since maps have a value and a key, you need to specify both types when initialising a map: `map<double, int>`, `unordered_map<int, double>`, and so on. The following example shows how you could use a `map` to implement `table()` for numeric vectors:

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
std::map<double, int> tableC(NumericVector x) {
 std::map<double, int> counts;

 int n = x.size();
 for (int i = 0; i < n; i++) {
 counts[x[i]]++;
 }

 return counts;
}
```

Note that unordered maps are only available in C++ 11, so to use them, you'll again need `[[Rcpp::plugins(cpp11)]]`.

## 24.6.7 Exercises

To practice using the STL algorithms and data structures, implement the following using R functions in C++, using the hints provided:

1. `median.default()` using `partial_sort`.
2. `%in%` using `unordered_set` and the `find()` or `count()` methods.
3. `unique()` using an `unordered_set` (challenge: do it in one line!).
4. `min()` using `std::min()`, or `max()` using `std::max()`.
5. `which.min()` using `min_element`, or `which.max()` using `max_element`.
6. `setdiff()`, `union()`, and `intersect()` for integers using sorted ranges and `set_union`, `set_intersection` and `set_difference`.

## 24.7 Case studies

The following case studies illustrate some real life uses of C++ to replace slow R code.

### 24.7.1 Gibbs sampler

The following case study updates an example blogged about (<http://dirk.eddelbuettel.com/blog/2011/07/14/>) by Dirk Eddelbuettel, illustrating the conversion of a Gibbs sampler in R to C++. The R and C++ code shown below is very similar (it only took a few minutes to convert the R version to the C++ version), but runs about 20 times faster on my computer. Dirk's blog post also shows another way to make it even faster: using the faster random number generator functions in GSL (easily accessible from R through the `RcppGSL` package) can make it another 2–3x faster.

The R code is as follows:

```
gibbs_r <- function(N, thin) {
 mat <- matrix(nrow = N, ncol = 2)
 x <- y <- 0

 for (i in 1:N) {
 for (j in 1:thin) {
 x <- rgamma(1, 3, y * y + 4)
 y <- rnorm(1, 1 / (x + 1), 1 / sqrt(2 * (x + 1)))
 }
 mat[i,] <- c(x, y)
 }
 mat
}
```

This is straightforward to convert to C++. We:

- add type declarations to all variables
- use `(` instead of `[` to index into the matrix
- subscript the results of `rgamma` and `rnorm` to convert from a vector into a scalar

```
#include <Rcpp.h>
using namespace Rcpp;

// [[Rcpp::export]]
NumericMatrix gibbs_cpp(int N, int thin) {
 NumericMatrix mat(N, 2);
 double x = 0, y = 0;

 for(int i = 0; i < N; i++) {
 for(int j = 0; j < thin; j++) {
 x = rgamma(1, 3, 1 / (y * y + 4))[0];
 y = rnorm(1, 1 / (x + 1), 1 / sqrt(2 * (x + 1)))[0];
 }
 mat(i, 0) = x;
 mat(i, 1) = y;
 }

 return(mat);
}
```

Benchmarking the two implementations yields:

```
microbenchmark(
 gibbs_r(100, 10),
 gibbs_cpp(100, 10)
)
#> Unit: microseconds
#> expr min lq mean median uq max neval
#> gibbs_r(100, 10) 5,320 5,410 6356 5,470 5,570 16,000 100
#> gibbs_cpp(100, 10) 313 340 372 355 372 1,730 100
```

### 24.7.2 R vectorisation vs. C++ vectorisation

This example is adapted from “Rcpp is smoking fast for agent-based models in data frames” (<https://gweissman.github.io/babelgraph/blog/2017/06/15/rcpp-is-smoking-fast-for-agent-based-models-in-data-frames.html>). The challenge is to predict a model response from three inputs. The basic R version of the predictor looks like:

```
vacc1 <- function(age, female, ily) {
 p <- 0.25 + 0.3 * 1 / (1 - exp(0.04 * age)) + 0.1 * ily
 p <- p * if (female) 1.25 else 0.75
 p <- max(0, p)
 p <- min(1, p)
 p
}
```

We want to be able to apply this function to many inputs, so we might write a vector-input version using a for loop.

```
vacc1 <- function(age, female, ily) {
 n <- length(age)
 out <- numeric(n)
 for (i in seq_len(n)) {
 out[i] <- vacc1(age[i], female[i], ily[i])
 }
 out
}
```

If you’re familiar with R, you’ll have a gut feeling that this will be slow, and indeed it is. There are two ways we could attack this problem. If you have a good R vocabulary, you might immediately see how to vectorise the function (using `ifelse()`, `pmin()`, and `pmax()`). Alternatively, we could rewrite `vacc1a()` and `vacc1()` in C++, using our knowledge that loops and function calls have much lower overhead in C++.

Either approach is fairly straightforward. In R:

```
vacc2 <- function(age, female, ily) {
 p <- 0.25 + 0.3 * 1 / (1 - exp(0.04 * age)) + 0.1 * ily
 p <- p * ifelse(female, 1.25, 0.75)
 p <- pmax(0, p)
 p <- pmin(1, p)
 p
}
```

(If you’ve worked R a lot you might recognise some potential bottlenecks in this code: `ifelse`, `pmin`, and `pmax` are known to be slow, and could be replaced with `p * 0.75 + p * 0.5 * female`, `p[p < 0] <- 0`, `p[p > 1] <- 1`. You might want to try timing those variations yourself.)

Or in C++:

```
#include <Rcpp.h>
using namespace Rcpp;

double vacc3a(double age, bool female, bool ily){
 double p = 0.25 + 0.3 * 1 / (1 - exp(0.04 * age)) + 0.1 * ily;
 p = p * (female ? 1.25 : 0.75);
 p = std::max(p, 0.0);
 p = std::min(p, 1.0);
 return p;
}
```

```
// [[Rcpp::export]]
NumericVector vacc3(NumericVector age, LogicalVector female,
 LogicalVector ily) {
 int n = age.size();
 NumericVector out(n);

 for(int i = 0; i < n; ++i) {
 out[i] = vacc3a(age[i], female[i], ily[i]);
 }

 return out;
}
```

We next generate some sample data, and check that all three versions return the same values:

```
n <- 1000
age <- rnorm(n, mean = 50, sd = 10)
female <- sample(c(T, F), n, rep = TRUE)
ily <- sample(c(T, F), n, prob = c(0.8, 0.2), rep = TRUE)

stopifnot(
 all.equal(vacc1(age, female, ily), vacc2(age, female, ily)),
 all.equal(vacc1(age, female, ily), vacc3(age, female, ily))
)
```

The original blog post forgot to do this, and introduced a bug in the C++ version: it used 0.004 instead of 0.04. Finally, we can benchmark our three approaches:

```
microbenchmark(
 vacc1 = vacc1(age, female, ily),
 vacc2 = vacc2(age, female, ily),
 vacc3 = vacc3(age, female, ily)
)
#> Unit: microseconds
#> expr min lq mean median uq max neval
#> vacc1 1,740.0 1,900.0 2119.3 2,000.0 2,050.0 5,840 100
#> vacc2 114.0 125.0 220.9 147.0 174.0 6,810 100
#> vacc3 30.1 31.6 51.3 36.3 40.2 1,510 100
```

Not surprisingly, our original approach with loops is very slow. Vectorising in R gives a huge speedup, and we can eke out even more performance (~10x) with the C++ loop. I was a little surprised that the C++ was so much faster, but it is because the R version has to create 11 vectors to store intermediate results, where the C++ code only needs to create 1.

## 24.8 Using Rcpp in a package

The same C++ code that is used with `sourceCpp()` can also be bundled into a package. There are several benefits of moving code from a stand-alone C++ source file to a package:

1. Your code can be made available to users without C++ development tools.
2. Multiple source files and their dependencies are handled automatically by the R package build system.
3. Packages provide additional infrastructure for testing, documentation, and consistency.

To add Rcpp to an existing package, you put your C++ files in the `src/` directory and modify/create the following configuration files:

- In `DESCRIPTION` add

```
LinkingTo: Rcpp
Imports: Rcpp
```

- Make sure your `NAMESPACE` includes:

```
useDynLib(mypackage)
importFrom(Rcpp, sourceCpp)
```

We need to import something (anything) from Rcpp so that internal Rcpp code is properly loaded.  
This is a bug in R and hopefully will be fixed in the future.

To generate a new Rcpp package that includes a simple “hello world” function you can use `Rcpp.package.skeleton()`:

```
Rcpp.package.skeleton("NewPackage", attributes = TRUE)
```

To generate a package based on C++ files that you’ve been using with `sourceCpp()`, use the `cpp_files` parameter:

```
Rcpp.package.skeleton("NewPackage", example_code = FALSE,
 cpp_files = c("convolve.cpp"))
```

Before building the package, you’ll need to run `Rcpp::compileAttributes()`. This function scans the C++ files for `Rcpp::export` attributes and generates the code required to make the functions available in R. Re-run `compileAttributes()` whenever functions are added, removed, or have their signatures changed. This is done automatically by the `devtools` package and by Rstudio.

For more details see the Rcpp package vignette, `vignette("Rcpp-package")`.

## 24.9 Learning more

This chapter has only touched on a small part of Rcpp, giving you the basic tools to rewrite poorly performing R code in C++. The Rcpp book (<http://www.rcpp.org/book>) is the best reference to learn more about Rcpp. As noted, Rcpp has many other capabilities that make it easy to interface R to existing C++ code, including:

- Additional features of attributes including specifying default arguments, linking in external C++ dependencies, and exporting C++ interfaces from packages. These features and more are covered in the Rcpp attributes vignette, `vignette("Rcpp-attributes")`.
- Automatically creating wrappers between C++ data structures and R data structures, including mapping C++ classes to reference classes. A good introduction to this topic is Rcpp modules vignette, `vignette("Rcpp-modules")`
- The Rcpp quick reference guide, `vignette("Rcpp-quicref")`, contains a useful summary of Rcpp classes and common programming idioms.

I strongly recommend keeping an eye on the Rcpp homepage (<http://www.rcpp.org>) and Dirk’s Rcpp page (<http://dirk.eddelbuettel.com/code/rcpp.html>) as well as signing up for the Rcpp mailing list (<http://lists.r-forge.r-project.org/cgi-bin/mailman/listinfo/rcpp-devel>). Rcpp is still under active development, and is getting better with every release.

Other resources I’ve found helpful in learning C++ are:

- *Effective C++* (<http://amzn.com/0321334876?tag=devtools-20>) and *Effective STL* (<http://amzn.com/0201749629?tag=devtools-20>) by Scott Meyers.

- *C++ Annotations* (<http://www.icce.rug.nl/documents/cplusplus/cplusplus.html>), aimed at “knowledgeable users of C (or any other language using a C-like grammar, like Perl or Java) who would like to know more about, or make the transition to, C++”.
- *Algorithm Libraries* (<http://www.cs.helsinki.fi/u/tpkarkka/alglib/k06/>), which provides a more technical, but still concise, description of important STL concepts. (Follow the links under notes).

Writing performance code may also require you to rethink your basic approach: a solid understanding of basic data structures and algorithms is very helpful here. That’s beyond the scope of this book, but I’d suggest the *Algorithm Design Manual* (<http://amzn.com/0387948600?tag=devtools-20>), MIT’s *Introduction to Algorithms* (<http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-046j-introduction-to-algorithms-sma-5503-fall-2005/>), *Algorithms* by Robert Sedgewick and Kevin Wayne which has a free online textbook (<http://algs4.cs.princeton.edu/home/>) and a matching coursera course (<https://www.coursera.org/course/algs4partI>).

## 24.10 Acknowledgments

I’d like to thank the Rcpp-mailing list for many helpful conversations, particularly Romain Francois and Dirk Eddelbuettel who have not only provided detailed answers to many of my questions, but have been incredibly responsive at improving Rcpp. This chapter would not have been possible without JJ Allaire; he encouraged me to learn C++ and then answered many of my dumb questions along the way.