

Advanced R

Hadley Wickham

2018-04-22

Contents

| | |
|---|-----------|
| Welcome | 5 |
| Other books | 5 |
| 1 Preface | 7 |
| 1.1 rlang | 7 |
| 1.2 Foundations | 7 |
| 1.3 Programming paradigms | 8 |
| 1.4 Removals | 8 |
| 2 Introduction | 9 |
| 2.1 Who should read this book | 10 |
| 2.2 Related work | 10 |
| 2.3 What you will get out of this book | 10 |
| 2.4 Meta-techniques | 11 |
| 2.5 Recommended reading | 11 |
| 2.6 Getting help | 12 |
| 2.7 Acknowledgments | 12 |
| 2.8 Conventions | 13 |
| 2.9 Colophon | 13 |
| I Foundations | 15 |
| 3 Names and values | 17 |
| 3.1 Introduction | 17 |
| 3.2 Binding basics | 17 |
| 3.3 Copy-on-modify | 20 |
| 3.4 Object size | 25 |
| 3.5 Modify-in-place | 26 |
| 3.6 Unbinding and the garbage collector | 29 |
| 4 Vectors | 33 |
| 4.1 Introduction | 33 |
| 4.2 Vectors | 34 |
| 4.3 Attributes | 38 |
| 4.4 Matrices and arrays | 41 |
| 4.5 Data frames | 44 |
| 4.6 Answers | 48 |
| 5 Subsetting | 51 |
| 5.1 Introduction | 51 |
| 5.2 Selecting multiple elements | 52 |

| | | |
|-----------|---|------------|
| 5.3 | Selecting a single element | 57 |
| 5.4 | Subsetting and assignment | 60 |
| 5.5 | Applications | 61 |
| 5.6 | Answers | 67 |
| 6 | Functions | 69 |
| 6.1 | Introduction | 69 |
| 6.2 | Function components | 70 |
| 6.3 | Lexical scoping | 71 |
| 6.4 | Every operation is a function call | 76 |
| 6.5 | Function arguments | 77 |
| 6.6 | Special calls | 83 |
| 6.7 | Return values | 86 |
| 6.8 | Quiz answers | 89 |
| 7 | Environments | 91 |
| 7.1 | Introduction | 91 |
| 7.2 | Environment basics | 92 |
| 7.3 | Recursing over environments | 98 |
| 7.4 | Special environments | 100 |
| 7.5 | The call stack | 107 |
| 7.6 | As data structures | 110 |
| 7.7 | <<- | 111 |
| 7.8 | Quiz answers | 111 |
| 8 | Debugging | 113 |
| 8.1 | Introduction | 113 |
| 8.2 | Techniques | 114 |
| 8.3 | Tools | 115 |
| 8.4 | Quiz answers | 119 |
| 9 | Conditions | 121 |
| 9.1 | Introduction | 121 |
| 9.2 | Signalling conditions | 122 |
| 9.3 | Ignoring conditions | 125 |
| 9.4 | Condition handlers | 126 |
| 9.5 | Use cases | 129 |
| 9.6 | Custom condition classes | 131 |
| 9.7 | Quiz answers | 133 |
| II | Functional programming | 135 |
| 10 | Functional programming | 137 |
| 10.1 | Introduction | 137 |
| 10.2 | Motivation | 138 |
| 10.3 | Anonymous functions | 141 |
| 10.4 | Closures | 142 |
| 10.5 | Lists of functions | 146 |
| 10.6 | Case study: numerical integration | 149 |
| 11 | Functionals | 153 |
| 11.1 | Introduction | 153 |
| 11.2 | My first functional: lapply() | 154 |
| 11.3 | For loop functionals: friends of lapply() | 157 |

| | |
|--|------------|
| 11.4 Manipulating matrices and data frames | 163 |
| 11.5 Manipulating lists | 167 |
| 11.6 Mathematical functionals | 169 |
| 11.7 Loops that should be left as is | 170 |
| 11.8 A family of functions | 172 |
| 12 Function operators | 177 |
| 12.1 Introduction | 177 |
| 12.2 Behavioural FOs | 178 |
| 12.3 Output FOs | 185 |
| 12.4 Input FOs | 187 |
| 12.5 Combining FOs | 190 |
| III Object oriented programming | 195 |
| 13 Introduction | 197 |
| 13.1 OOP Systems | 197 |
| 13.2 OOP in R | 198 |
| 13.3 Field guide | 198 |
| 14 Base types | 201 |
| 14.1 Introduction | 201 |
| 14.2 Base objects vs OO objects | 201 |
| 14.3 Base types | 202 |
| 14.4 The is functions | 203 |
| 15 S3 | 205 |
| 15.1 Introduction | 205 |
| 15.2 Basics | 205 |
| 15.3 Classes | 207 |
| 15.4 Generics and methods | 213 |
| 15.5 Method dispatch | 216 |
| 15.6 Inheritance | 218 |
| 15.7 Dispatch details | 221 |
| 16 S4 | 227 |
| 16.1 Introduction | 227 |
| 16.2 Classes | 228 |
| 16.3 Generics and methods | 231 |
| 16.4 Method dispatch | 235 |
| 16.5 S4 and existing code | 241 |
| 17 R6 | 243 |
| 17.1 Introduction | 243 |
| 17.2 Classes and methods | 244 |
| 17.3 Controlling access | 248 |
| 17.4 Reference semantics | 251 |
| 18 Trade-offs | 255 |
| 18.1 Introduction | 255 |
| 18.2 S4 vs S3 | 255 |
| 18.3 R6 vs S3 | 256 |

| | |
|---|------------|
| IV Metaprogramming | 261 |
| 19 Introduction | 263 |
| 19.1 Domain specific languages | 264 |
| 19.2 Overview | 264 |
| 20 Expressions | 267 |
| 20.1 Introduction | 267 |
| 20.2 Abstract syntax trees | 268 |
| 20.3 R's grammar | 271 |
| 20.4 Data structures | 274 |
| 20.5 Parsing and deparsing | 278 |
| 20.6 Case study: Walking the AST with recursive functions | 280 |
| 21 Quasiquotation | 287 |
| 21.1 Introduction | 287 |
| 21.2 Motivation | 288 |
| 21.3 Quotation | 290 |
| 21.4 Evaluation | 293 |
| 21.5 Unquotation | 294 |
| 21.6 Case studies | 300 |
| 21.7 Dot-dot-dot (...) | 305 |
| 22 Evaluation | 311 |
| 22.1 Introduction | 311 |
| 22.2 Evaluation basics | 312 |
| 22.3 Quosures | 316 |
| 22.4 Tidy evaluation | 321 |
| 22.5 Wrapping quoting functions | 329 |
| 23 Translating R code | 335 |
| 23.1 Introduction | 335 |
| 23.2 HTML | 336 |
| 23.3 LaTeX | 342 |
| V Performance | 349 |
| 24 Performance | 351 |
| 24.1 Why is R slow? | 351 |
| 24.2 Microbenchmarking | 352 |
| 24.3 Language performance | 353 |
| 24.4 Implementation performance | 357 |
| 24.5 Alternative R implementations | 359 |
| 25 Optimising code | 363 |
| 25.1 Introduction | 363 |
| 25.2 Measuring performance | 364 |
| 25.3 Memory profiling with lineprof | 367 |
| 25.4 Improving performance | 369 |
| 25.5 Code organisation | 370 |
| 25.6 Has someone already solved the problem? | 371 |
| 25.7 Do as little as possible | 371 |
| 25.8 Vectorise | 377 |
| 25.9 Avoid copies | 378 |

| | |
|--|------------|
| 25.10 Byte code compilation | 379 |
| 25.11 Case study: t-test | 379 |
| 25.12 Parallelise | 381 |
| 25.13 Other techniques | 383 |
| 26 High performance functions with Rcpp | 385 |
| 26.1 Introduction | 385 |
| 26.2 Getting started with C++ | 386 |
| 26.3 Attributes and other classes | 393 |
| 26.4 Missing values | 395 |
| 26.5 Rcpp sugar | 398 |
| 26.6 The STL | 400 |
| 26.7 Case studies | 404 |
| 26.8 Using Rcpp in a package | 407 |
| 26.9 Learning more | 408 |
| 26.10 Acknowledgments | 409 |

Welcome

This is the website for work-in-progress 2nd edition of “**Advanced R**”, 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**” 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**” which teaches you how to make the most of R’s fantastic package system.

Chapter 1

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.

1.1 rlang

A big change since the first edition of the book is the creation of the `rlang` package, written primarily by Lionel Henry. This goal of this package is to provide 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.

1.2 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 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.
- Exceptions and debugging has been split into two chapters, “debugging” and “conditions”. The contents of conditions has been expanded. The section of defensive programming has been removed, because discussing type stability is more natural in the context of functional programming, and programming with NSE is not the challenge it once was (now that tidy evaluation exists).

1.3 Programming paradigms

The meat of 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 focus more on the underlying ideas without being distracted by incidental details.
- 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 has been substantially expanded (from three chapters to five) and reorganised. More diagrams.

1.4 Removals

- Chapter of base R vocabulary was removed.
- The style guide has moved to <http://style.tidyverse.org/>. It is now paired with the `styler` 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>

Chapter 2

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, stackoverflow, or subject-specific mailing lists like R-SIG-mixed-models or ggplot2. You can also connect with other R learners via twitter, linkedin, and through many local user groups.
- Powerful tools for communicating your results. R packages make it easy to produce html or pdf reports, or create interactive websites.
- 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 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.

2.1 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()`).

2.2 Related work

Tidyverse + R4DS

R packages

2.3 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.

2.4 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), I often figure out the solution myself.

2.5 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* (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 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 by Andrew Hunt and David Thomas. This book is language agnostic, and provides great advice for how to be a better programmer.

2.6 Getting help

Currently, there are two main venues to get help when you're stuck and can't figure out what's causing the problem: stackoverflow and the R-help mailing list. You can get fantastic help in both venues, 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. 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.
- 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.

2.7 Acknowledgments

I would like to thank the tireless contributors to R-help and, more recently, stackoverflow. 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, and chapters were advertised on twitter 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 Schumacher, Aaron Wolen, @aaronwolen, @absolutelyNoWarranty, Adam Hunt, @agrabovsky, @ajdm, Alan Dipert, Alexander Grueneberg, @alexbrown, @alko989, @allegretto, @amarchin, @AmeliaMN, Andrew Bray, @andrewla, Andy Teucher, Anthony Damico, Anton Antonov, @aranlunzer, @arilamstein, @asnr, @avilella, @baptiste, Bart Kastermans, @blindjesse, @blmoore, @bnjmn, Brandon Greenwell, Brandon Hurr, Brett Klamer, Brian G. Barkley, Brian Knaus, @BrianDiggs, @Bryce, C. Jason Liang, @carey1024, @Carson, @cdrv, Ching Boon, @chiphogg, @ChrisMuir, Christopher Brown, @christophergandrud, Clay Ford, Colin Fay, @cornelius1729, @cplouffe, Craig Citro, @crossfitAL, @crowding, Crt Ahlin, @crtahlin, @cscheid, @csgillespie, @cusanovich, @cwarden, @cwickham, Daisuke Ichikawa, Daniel Lee, @darrkj, @Dasonk,

Dave Childers, David Hajage, David LeBauer, Davor Cubranic, @dchudz, Dean Attali, dennis feehan, Dewey Dunnington, @dfeehan, Dirk Eddelbuettel, @dkahle, @dlebauer, @dlschweizer, @dmontaner, @dougmitarotonda, @dpatschke, @duncandonutz, @eaurele, @EdFineOKL, @EDiLD, Edwin Thoen, @eijoac, @eipi10, @elegrand, Ellis Valentiner, @EmilRehnberg, Eric C. Anderson, @etb, @fabian-s, Facundo Muñoz, @flammy0530, @fpepin, Francois Michonneau, Frank Farach, Frans van Dunné, @freezby, @fyears, Garrett Grolemund, @garrettgman, @gavinsimpson, @gezakiss7, @gggtest, Gökçen Eraslan, @gr650, Gregg Whitworth, @gregorp, @gsee, @gsk3, @gthb, Guy Dawson, Harley Day, @hassaad85, @helmingstay, Henrik Bengtsson, @i, Iain Dillingham, Ian Lyttle, @IanKopacka, @ijlyttle, Ilan Man, @imanuelcostigan, @initdch, @irudnyts, Jason Asher, Jason Knight, @jasondavies, @jastingo, @jborras, Jeff Allen, @jeharmse, Jennifer (Jenny) Bryan, @jennybc, @jentjr, @Jeremiah, @JestonBlu, Jim Hester, Jim Vine, @JimInNashville, @jinlong25, JJ Allaire, @JMHay, Jochen Van de Velde, Johann Hibschman, John Blischak, john verzani, @johnbaums, @johnjosephhorton, @johnthomas12, Jon Calder, Joris Muller, @JorneBicler, Jose Antonio Magaña Mesa, Joseph Casillas, @juancentro, Julia Gustavsen, @kdauria, Ken Williams, @kenahoo, Kenny Darrell, @kent37, Kevin Markham, Kevin Ushey, @kforner, Kirill Müller, Kirill Sevastyanenko, Krishna Sankar, Kun Ren, Laurent Gatto, @Lawrence-Liu, @ldfmrails, @lgatto, @liangcj, Lingbing Feng, Lionel Henry, @lynaghk, Maarten Kruijver, Mamoun Benghezal, @mannyishere, Marcel Ramos, Mark Rosenstein, Martin Morgan, Matt Pettis, @mattbaggott, Matthew Grogan, Matthew Sedaghatfar, Matthieu Gomez, @mattmalin, Mauro Lepore, Max Ghenis, @Michael, Michael Bishop, Michael Buckley, Michael Kane, Michael Quinn, @michaelbach, Michał Bojanowski, @mjsduncan, @Mullefa, @myqlarson, Nacho Caballero, Nick Carchedi, @nignatiadis, @nstjhp, @ogennadi, Oliver Keyes, Oliver Paisley, @otepoti, Pariksheat Nanda, Parker Abercrombie, @patperu, Patrick Miller, @pavel-vodrazka, @pdb61, @pengyu, Peter F Schulam, Peter Lindbrook, Peter Meilstrup, @philchalmers, @picasa, @piccolbo, @pierreroudier, @polmath, @pooryorick, @quantbo, R. Mark Sharp, Ramnath Vaidyanathan, @ramnathv, @Rappster, Ricardo Pietrobon, Richard Cotton, @richardreeve, @rmflight, @rmsharp, Rob Weyant, Robert Krzyzanowski, Robert M Flight, @RobertZK, @robiRagan, Romain François, @rrunner, @rubenfcasal, Rumen Zarev, @sailingwave, @sarunasmerkliopas, @sbgraves237, Scott Ritchie, @scottko, @scottl, @seaaan, Sean Anderson, Sean Carmody, Sean Wilkinson, @Sebastian, @sebastian-c, Sébastien Vigneau, @shabbychef, Shannon Rush, Simon O'Hanlon, Simon Potter, @SplashDance, @ste-fan, Stefan Widgren, @stephens999, Steve Lianoglou, Steve Walker, Steven Pav, @strongh, @stuttungur, @surmann, Sven E. Templer, @Swarchal, @swnydick, @taekyunk, Tal Galili, @talgalili, @Tazinho, @tdenes, Terence Teo, @Thomas, Thomas Lin Pedersen, @thomasherbig, @thomaszumbrunn, Tim Cole, tj mahr, @tjmahr, Tom Buckley, Tom Crockett, @triche, @twjacobs, @tyhenkaline, @tylerrickie, @ulrichatz, @varun729, @victorkryukov, @vijaybarve, @vzemlys, @wchi144, Welliton Souza, @wibeasley, @WilCrofter, William Doane, Winston Chang, @winterschlaefer, @wmc3, Wolfgang Huber, @wordnerd, Yoni Ben-Meshulam, @yuchouchen, @zachcp, @zackham, @zerokarmaleft, Zhongpeng Lin, @ZhongpengLin.

2.8 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., <http://adv-r.had.co.nz>, you can easily copy and paste examples into R. Output comments look like `#>` to distinguish them from regular comments.

2.9 Colophon

This book was written in Rmarkdown inside Rstudio. knitr and pandoc converted the raw Rmarkdown to html and pdf. The website was made with jekyll, styled with bootstrap, and automatically published to Amazon's S3 by travis-ci. The complete source is available from [github](#).

Code is set in inconsolata.

Part I

Foundations

Chapter 3

Names and values

3.1 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.

Prerequisites

We'll use the development version of lobstr to dig into the memory 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. Most of the information in this chapter was gleaned from a close reading of the documentation (particularly `?Memory` and `?gc`), the memory profiling section of R-exts, and the SEXPs section of R-ints. 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.

3.2 Binding basics

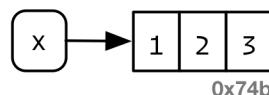
Take this code:

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

It's easy to read this code as: "create an object named 'x', containing the values 1, 2, and 3". But 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, 1:3.
- Binding the object to a name, x.

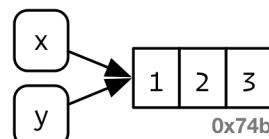
Note that the object, or value, doesn't have a name; it's the name that has a value. To make that distinction more clear, I'll draw diagrams like this:



The name, x, is drawn with a rounded rectangle, and it has an arrow that points to 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 independently of their bindings. To make that possible, I'll label values with a unique identifier. These unique identifiers have a special form that look like the object's memory "address", i.e. the location in memory in which the object is stored.

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

```
obj_addr(x)
#> [1] "0x563f0702d338"
obj_addr(y)
#> [1] "0x563f0702d338"
```

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.

3.2.1 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 _. 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:

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!

```
_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
`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, and doesn't follow R's rules.

::: sidebar You can also create non-syntactic bindings using single and double quotes (i.e. "a + b" <- 3) instead of backticks, but I don't recommend it 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. :::

3.2.2 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`.

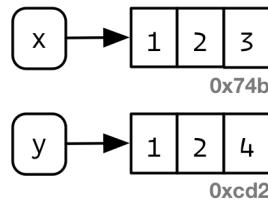
3.3 Copy-on-modify

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

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

y[[3]] <- 4
x
#> [1] 1 2 3
```

Clearly modifying `y` doesn't also modify `x`, so what happens 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.



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**; I'll generally avoid that term because there are a couple of important exceptions to copy-on-modify that you'll learn about in modify-in-place.

3.3.1 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 <- 1:3
cat(tracemem(x), "\n")
#> <0x563f06eb5d78>
```

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[0x563f06eb5d78 -> 0x563f0770dd28]: eval eval withVisible withCallingHandlers handle timing...
```

Figure out how to make results nicer inside RMarkdown

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 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.

3.3.2 Function calls

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

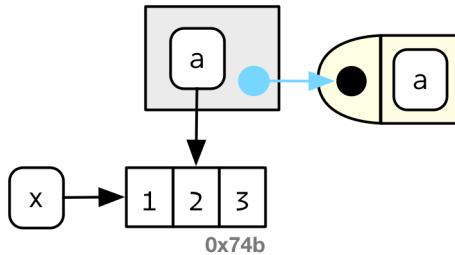
```
f <- function(a) {
  a
}

x <- 1:3
cat(tracemem(x), "\n")
#> <0x563f07702ce0>

z <- f(x)

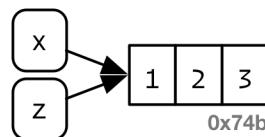
untracemem(x)
```

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, `z` will point to the same object.



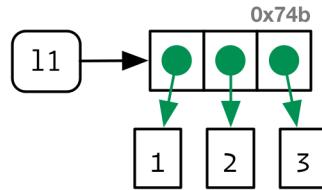
`0x74b` never gets copied because it never gets modified. If `f()` did modify `x`, R would create a new copy, and then `z` would bind that object.

3.3.3 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:

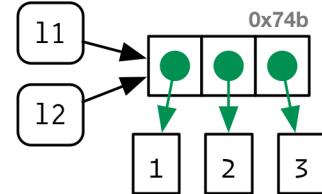
```
l1 <- list(1, 2, 3)
```

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

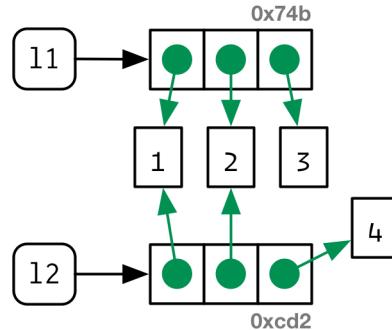


This is particularly important when we modify a list:

```
12 <- 11
```



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



Like vectors, lists are copied-on-modify; the original list is left unchanged, and R creates a modified copy. Note that the copy is **shallow**: the list object and its bindings are copied, but the values pointed to by the bindings are not. This behaviour was added in R 3.1.0 and had a big impact on performance.

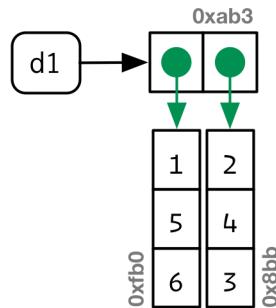
You can use `lobstr::ref()` to see values that are shared across lists. `ref()` prints the memory address of each object, along with a local id so that you can easily cross-reference shared components.

```
ref(11, 12)
#> <1:0x563f072ac820> list
#> <2:0x563f073ae248> dbl
#> <3:0x563f073ae278> dbl
#> <4:0x563f073ae2d8> dbl
#>
#> <5:0x563f0759c6b8> list
#> <2:0x563f073ae248>
#> <3:0x563f073ae278>
#> <6:0x563f07497a68> dbl
```

3.3.4 Data frames

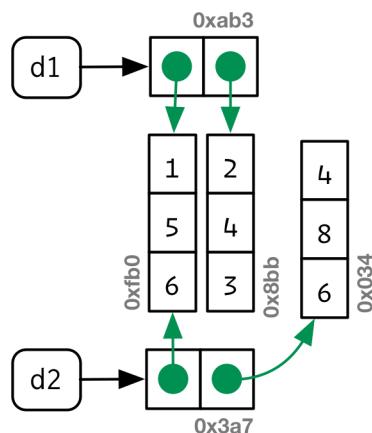
Data frames are lists, so copy-on-modify has important consequences when you modify a data frame. Take this data frame as an example:

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



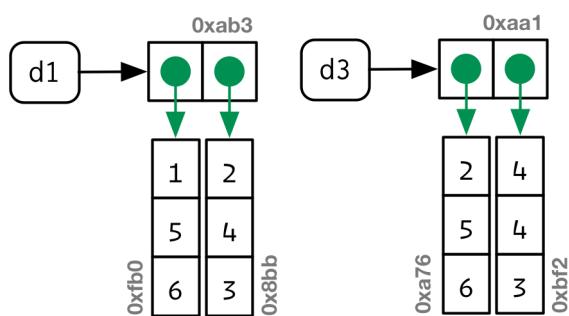
If you modify a column, only that column needs to be modified; the others can continue to point to the same place:

```
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.

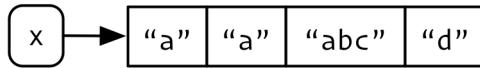
```
d3 <- d1
d3[1, ] <- d3[1, ] * 2
```



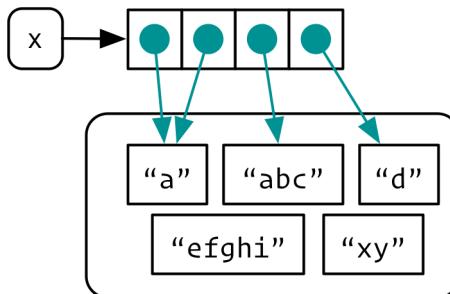
3.3.5 Character vectors

The final place that R uses references is in character vectors. In the previous chapter, we drew character vectors like this:

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



This is 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:



The global string pool

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

```
ref(x, character = TRUE)
#> <1:0x563f07f2e488> chr
#> <2:0x563f04f53868> string: 'a'
#> <2:0x563f04f53868>
#> <3:0x563f07436ac8> string: 'abc'
#> <4:0x563f052ee388> string: 'd'
```

Generally, however, this detail is not important, so elsewhere in the book I'll draw character vectors as if the individual strings live inside the vector.

3.3.6 Exercises

1. Why is `tracemem(1:10)` not useful?
2. Explain why `tracemem()` records 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 <- 1:3
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.

3.4 Object size

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

```
obj_size(letters)
#> 1,496 B
obj_size(ggplot2::diamonds)
#> 3,455,904 B
```

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

```
x <- 1:1e6
obj_size(x)
#> 4,000,040 B

y <- list(x, x, x)
obj_size(y)
#> 4,000,112 B
```

`y` is only 72 bytes² bigger than `x`. That's the size of an empty list with three elements:

```
obj_size(list(NULL, NULL, NULL))
#> 72 B
```

::: base We need to use `lobstr::obj_size()` here because the base equivalent, `utils::object.size()`, incorrectly counts `x` three times when computing the size of `y`. :::

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)
#> 120 B
obj_size(rep(banana, 100))
#> 912 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`:

```
obj_size(x, y)
#> 4,000,112 B
```

3.4.1 Exercises

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

²If you're running 32-bit R you'll see slightly different sizes.

```
x <- list(mean, sd, var)
obj_size(x)
#> 16,928 B
```

2. Predict the output of the following code:

```
x <- 1:1e6
obj_size(x)
#> 4,000,040 B

y <- list(x, x)
obj_size(y)
#> 4,000,096 B
obj_size(x, y)
#> 4,000,096 B

y[[1]][[1]] <- 10
obj_size(y)
#> 12,000,136 B
obj_size(x, y)
#> 12,000,136 B

y[[2]][[1]] <- 10
obj_size(y)
#> 16,000,136 B
obj_size(x, y)
#> 20,000,176 B
```

3.5 Modify-in-place

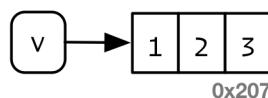
Most of the time, modifying an R object will create a copy. There are two exceptions:

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

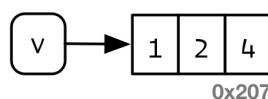
3.5.1 Objects with a single binding

If an object only has a single binding to it, R will modify it in place:

```
v <- 1:3
```



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



(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 currently³ only count 0, 1, and many. That means if an object has two bindings, and one goes away, the reference count does not get decremented (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, which 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")
#> <0x563f07b29580>

for (i in 1:5) {
  x[[i]] <- x[[i]] - medians[[i]]
}
#> tracemem[0x563f07b29580 -> 0x563f08d2e4f8]: eval eval withVisible withCallingHandlers handle timing_
#> tracemem[0x563f08d2e4f8 -> 0x563f08d2e490]: [[<- .data.frame [[<- eval eval withVisible withCallingHa
#> tracemem[0x563f08d2e490 -> 0x563f08d2e428]: [[<- .data.frame [[<- eval eval withVisible withCallingHa
#> tracemem[0x563f08d2e428 -> 0x563f08d2e3c0]: eval eval withVisible withCallingHandlers handle timing_
#> tracemem[0x563f08d2e3c0 -> 0x563f08d2e358]: [[<- .data.frame [[<- eval eval withVisible withCallingHa
#> tracemem[0x563f08d2e358 -> 0x563f08d2e288]: [[<- .data.frame [[<- eval eval withVisible withCallingHa
#> tracemem[0x563f08d2e288 -> 0x563f088eab58]: eval eval withVisible withCallingHandlers handle timing_
#> tracemem[0x563f088eab58 -> 0x563f088eaa88]: [[<- .data.frame [[<- eval eval withVisible withCallingHa
#> tracemem[0x563f088eaa88 -> 0x563f088ea9b8]: [[<- .data.frame [[<- eval eval withVisible withCallingHa
#> tracemem[0x563f088ea9b8 -> 0x563f088ea950]: eval eval withVisible withCallingHandlers handle timing_
#> tracemem[0x563f088ea950 -> 0x563f088ea8e8]: [[<- .data.frame [[<- eval eval withVisible withCallingHa
#> tracemem[0x563f088ea8e8 -> 0x563f088ea880]: [[<- .data.frame [[<- eval eval withVisible withCallingHa
#> tracemem[0x563f088ea880 -> 0x563f088ea818]: eval eval withVisible withCallingHandlers handle timing_
#> tracemem[0x563f088ea818 -> 0x563f088ea7b0]: [[<- .data.frame [[<- eval eval withVisible withCallingHa
#> tracemem[0x563f088ea7b0 -> 0x563f088ea748]: [[<- .data.frame [[<- eval eval withVisible withCallingHa

untracemem(x)
```

In fact, each iteration copies the data frame not once, not twice, but three times! We get two copies inside of `[[.data.frame`, and a further copy because `[[.data.frame` is a regular function and hence increments the reference count of `x`. (Note that these copies will be shallow so they are not too expensive, but they obviously make the loop slower than you might hope).

³By the time you read this, that may have changed, as plans are afoot to improve reference counting: <https://developer.r-project.org/RefCnt.html>

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")
#> <0x563f0c01ea38>

for (i in 1:5) {
  y[[i]] <- y[[i]] - medians[[i]]
}
#> tracemem[0x563f0c01ea38 -> 0x563f08653890]: eval eval withVisible withCallingHandlers handle timing_
```

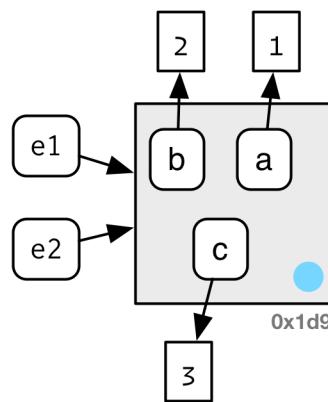
While determining that copies are being made is not hard, preventing such behaviour is. If you find yourself resorting to exotic tricks to avoid copies, it may be time to rewrite your function in C++, as described in Rcpp.

3.5.2 Environments

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

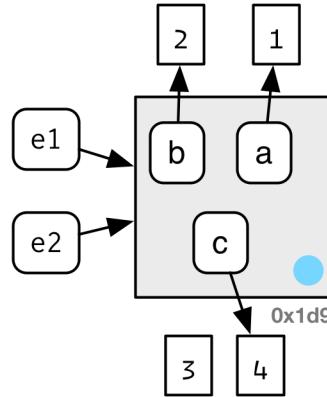
Take this environment, which we bind to e1 and e2:

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



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

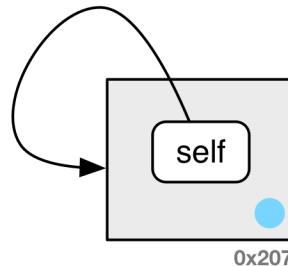
```
e1$c <- 4
e2$c
#> [1] 4
```



One consequence of this is that environments can contain themselves:

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

ref(e)
#> <1:0x563f06997a38> env
#> self = <1:0x563f06997a38>
```



This is a unique property of environments!

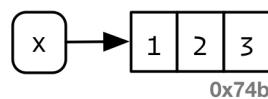
3.5.3 Exercises

1. Wrap the two methods for subtracting medians into two functions, then use the microbenchmark 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?

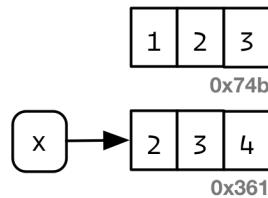
3.6 Unbinding and the garbage collector

Consider this code:

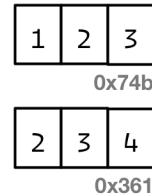
```
x <- 1:3
```



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



```
rm(x)
```



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.

R uses a **tracing** garbage collector. That means it traces every object reachable from the global⁴ 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.

3.6.1 When does the garbage collector run?

The garbage collector is run automatically whenever a new R object is created and R needs more memory. If you want to see when that occurs, call `gcinfo(TRUE)`: it will print a message to the console every time the garbage collector runs. Running the GC creates more memory by deleting R objects that are no longer used, and if needed, requesting more memory from the operating system.

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  619424 33.1    1168576 62.5   1168576 62.5
#> Vcells 1574489 12.1    7913801 60.4   9884807 75.5
```

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

```
mem_used()
#> 47,287,936 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.

⁴And every environment on the current call stack.

- 3. R counts the memory occupied by objects but there may be gaps due to deleted objects. This problem is known as memory fragmentation.

3.6.2 Memory leaks

The GC takes care of deleting all objects that do not have bindings. But you are still at risk for a **memory leak**, which occurs when you keep a binding to an object without realising it. In R, the two main causes of memory leaks are formulas and functions because they both capture the enclosing environment. The following code illustrates the problem. In `f1()`, `1:1e6` is only referenced inside the function, so when the function completes the memory is returned and the net memory change is 0. `f2()` and `f3()` both return objects that capture environments, so that `x` is not freed when the function completes.

```
f1 <- function() {
  x <- 1:1e6
  10
}
x <- f1()
obj_size(x)
#> 48 B

f2 <- function() {
  x <- 1:1e6
  a ~ b
}
y <- f2()
obj_size(y)
#> 4,000,864 B

f3 <- function() {
  x <- 1:1e6
  function() 10
}
z <- f3()
obj_size(z)
#> 4,012,192 B
```


Chapter 4

Vectors

4.1 Introduction

This chapter summarises the most important data structures in base R: the vector types. You've probably used many (if not all) of them before, but you may not have thought deeply about how they are interrelated. In this brief overview, I won't discuss individual types in 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.

R's vectors can be organised by their dimensionality (1d, 2d, or nd) and whether they're homogeneous (all contents must be of the same type) or heterogeneous (the contents can be of different types). This gives rise to the five data types most often used in data analysis:

| | Homogeneous | Heterogeneous |
|----|---------------|---------------|
| 1d | Atomic vector | List |
| 2d | Matrix | Data frame |
| nd | Array | |

Almost all other objects are built upon these foundations. In base types, you'll learn more about that foundation, and then in S3 you'll see how you can make your own extensions.

Note that R has no 0-dimensional, or scalar types. Individual numbers or strings, which you might think would be scalars, are actually vectors of length one.

Given an object, the best way to understand what data structures its composed of is to use `str()`. `str()` is short for structure and it gives a compact, human readable description of any R data structure.

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 three properties of a vector, other than its contents?
2. What are the four common types of atomic vectors? What are the two rare types?
3. What are attributes? How do you get them and set them?
4. How is a list different from an atomic vector? How is a matrix different from a data frame?

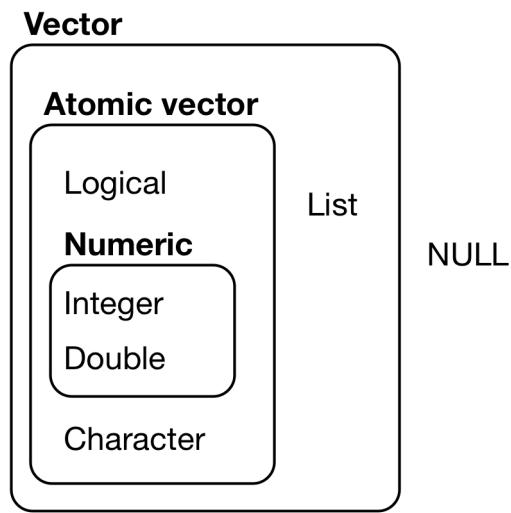
5. Can you have a list that is a matrix? Can a data frame have a column that is a matrix?
6. How do tibbles behave differently from data frames?

Outline

- Vectors introduces you to atomic vectors and lists, R's 1d data structures.
- Attributes takes a small detour to discuss attributes, R's flexible metadata specification. Here you'll learn about factors, an important data structure created by setting attributes of an atomic vector.
- Matrices and arrays introduces matrices and arrays, data structures for storing 2d and higher dimensional data.
- Data frames teaches you about the data frame, the most important data structure for storing data in R. Data frames combine the behaviour of lists and matrices to make a structure ideally suited for the needs of statistical data.

4.2 Vectors

The most common data structure in R is the vector. Vectors come in two flavours: atomic vectors and lists.



They have three common properties:

- Type, `typeof()`, what it is.
- Length, `length()`, how many elements it contains.
- Attributes, `attributes()`, additional arbitrary metadata.

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.

4.2.1 Atomic vectors

There are four common types of atomic vectors that I'll discuss in detail: logical, integer, double, and character. Collectively integer and double vectors are known as numeric (. There are two rare types that I will not discuss further: complex and raw.

Atomic vectors are usually created with `c()`, short for combine:

```
dbl_var <- c(1, 2.5, 4.5)
# With the L suffix, you get an integer rather than a double
int_var <- c(1L, 6L, 10L)
# Use TRUE and FALSE (or T and F) to create logical vectors
log_var <- c(TRUE, FALSE, T, F)
chr_var <- c("these are", "some strings")
```

Throughout the book, I'll draw vectors as connected boxes:

| | | | |
|-------------|-------|----------------|-------|
| 1.0 | 2.5 | 4.5 | |
| 1 | 6 | 10 | |
| TRUE | FALSE | TRUE | FALSE |
| “these are” | | “some strings” | |

Atomic vectors are always flat, even if you nest `c()`'s:

```
c(1, c(2, c(3, 4)))
#> [1] 1 2 3 4
# the same as
c(1, 2, 3, 4)
#> [1] 1 2 3 4
```

Missing values are specified with `NA`, which is a logical vector of length 1. `NA` will always be coerced to the correct type if used inside `c()`, or you can create `NAs` of a specific type with `NA_real_` (a double vector), `NA_integer_` and `NA_character_`.

4.2.1.1 Types and tests

Given a vector, you can determine its type with `typeof()`.

Use “is” functions with care. `is.character()`, `is.double()`, `is.integer()`, `is.logical()` are ok. The following are surprising:

- `is.vector()` tests for vectors with no attributes apart from names
- `is.atomic()` tests for atomic vectors or `NULL`
- `is.numeric()` tests for the numerical-ness of a vector, not whether it's built on top of an integer or double.

4.2.1.2 Coercion

All elements of an atomic vector must be the same type, so when you attempt to combine different types they will be **coerced** to the most flexible type. Types from least to most flexible are: logical, integer, double, and character.

For example, combining a character and an integer yields a character:

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

When a logical vector is coerced to an integer or double, TRUE becomes 1 and FALSE becomes 0. This is very useful in conjunction with `sum()` and `mean()`:

```
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
```

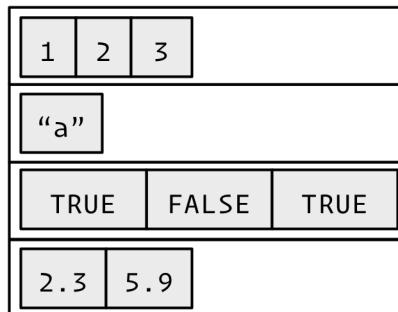
Coercion often happens automatically. Most mathematical functions (+, log, abs, etc.) will coerce to a double or integer, and most logical operations (&, |, any, etc) will coerce to a logical. You will usually get a warning message if the coercion might lose information. If confusion is likely, explicitly coerce with `as.character()`, `as.double()`, `as.integer()`, or `as.logical()`.

4.2.2 Lists

Lists are different from atomic vectors because their elements can be of any type, including lists. You construct lists by using `list()` instead of `c()`:

```
x <- list(1:3, "a", c(TRUE, FALSE, TRUE), c(2.3, 5.9))
str(x)
#> List of 4
#> $ : int [1:3] 1 2 3
#> $ : chr "a"
#> $ : logi [1:3] TRUE FALSE TRUE
#> $ : num [1:2] 2.3 5.9
```

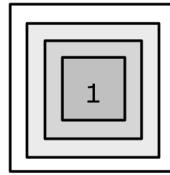
Lists can contain complex objects so it's not possible to pick one visual style that works for every list. Generally I'll draw lists like vectors, using colour to remind you of the hierarchy.



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

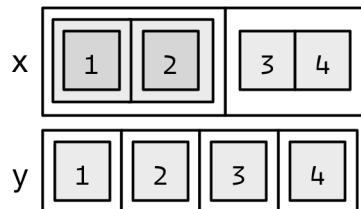
```
x <- list(list(list(1))))
str(x)
#> List of 1
#> $ :List of 1
#>   ..$ :List of 1
```

```
#> ... $ :List of 1
#> ... ... $ : num 1
is.recursive(x)
#> [1] TRUE
```



`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()`:

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



The `typeof()` of a list is `list`. You can test for a list with `is.list()` and coerce to a list with `as.list()`. You can turn a list into an atomic vector with `unlist()`. If the elements of a list have different types, `unlist()` uses the same coercion rules as `c()`.

Lists are used to build up many of the more complicated data structures in R. For example, both data frames (described in data frames) and linear models objects (as produced by `lm()`) are lists:

```
is.list(mtcars)
#> [1] TRUE

mod <- lm(mpg ~ wt, data = mtcars)
is.list(mod)
#> [1] TRUE
```

You'll learn more about that in S3.

4.2.3 NULL

Closely related to vectors is `NULL`, a singleton object often used to represent a vector of length 0.

```
typeof(NULL)
#> [1] "NULL"
length(NULL)
#> [1] 0
```

4.2.4 Exercises

1. What are the six types of atomic vector? How does a list differ from an atomic vector?
2. What makes `is.vector()` and `is.numeric()` fundamentally different to `is.list()` and `is.character()`?
3. Test your knowledge of vector coercion rules by predicting the output of the following uses of `c()`:

```
c(1, FALSE)
c("a", 1)
c(list(1), "a")
c(TRUE, 1L)
```

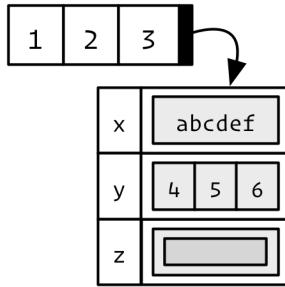
4. Why do you need to use `unlist()` to convert a list to an atomic vector? Why doesn't `as.vector()` work?
5. Why is `1 == "1"` true? Why is `-1 < FALSE` true? Why is `"one" < 2` false?
6. Why is the default missing value, `NA`, a logical vector? What's special about logical vectors? (Hint: think about `c(FALSE, NA_character_)`.)

4.3 Attributes

All objects can have arbitrary additional attributes, used to store metadata about the object. Attributes can be thought of as a named list¹ (with unique names). Attributes can be accessed individually with `attr()` or all at once (as a list) with `attributes()`.

```
a <- 1:3
attr(a, "x") <- "abcdef"
attr(a, "y") <- 4:6
attr(a, "z") <- list(list())
str(attributes(a))
#> List of 3
#> $ x: chr "abcdef"
#> $ y: int [1:3] 4 5 6
#> $ z:List of 1
#> ..$ : list()
```

¹The reality is a little more complicated: attributes are actually stored in pairlists, which can you learn more about in pairlists. This is why I used a slightly different convention for drawing attributes compared to regular lists.



The `structure()` function returns a new object with modified attributes:

```
structure(1:10, my_attribute = "This is a vector")
#> [1] 1 2 3 4 5 6 7 8 9 10
#> attr(,"my_attribute")
#> [1] "This is a vector"
```

By default, most attributes are lost when modifying a vector:

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

The only attributes not lost are the three most important:

- Names, a character vector giving each element a name, described in names.
- Dimensions, used to turn vectors into matrices and arrays, described in matrices and arrays.
- Class, used to implement the S3 object system, which we will discuss in detail in S3.

Each of these attributes has a specific accessor function to get and set values. When working with these attributes, use `names(x)`, `dim(x)`, and `class(x)`, not `attr(x, "names")`, `attr(x, "dim")`, and `attr(x, "class")`.

4.3.0.1 Names

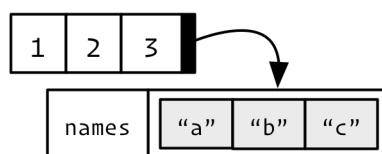
You can name a vector in three ways:

- When creating it: `x <- c(a = 1, b = 2, c = 3)`.
- By modifying an existing vector in place: `x <- 1:3; names(x) <- c("a", "b", "c")`.

Or: `x <- 1:3; names(x)[[1]] <- c("a")`.

- By creating a modified copy of a vector: `x <- setNames(1:3, c("a", "b", "c"))`.

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 the label the vector directly:

| | | |
|---|---|---|
| 1 | 2 | 3 |
| a | b | c |

Names don't have to be unique. However, character subsetting, described in subsetting, is the most important reason to use names and it is most useful when the names are unique.

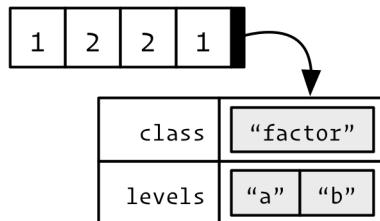
Not all elements of a vector need to have a name. Depending on how you create the vector the missing names will either have value "" or NA_character_. If all names are missing, names() will return NULL. You remove names from an existing vector using unname(x) or names(x) <- NULL.

4.3.1 Factors

One important use of attributes is to define 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 using 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 possible values a variable may take, even if you don't see all values in a given dataset. Using a factor instead of a character vector makes it obvious when some groups contain no observations:

```
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
```

Unfortunately, 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. Instead, use the argument `stringsAsFactors = FALSE` to suppress this behaviour, and then manually convert character vectors to factors using your knowledge of the data. A global option, `options(stringsAsFactors = FALSE)`, is available to control this behaviour, but I don't recommend using it. Changing a global option may have unexpected consequences when combined with other code (either from packages, or code that you're `source()`ing), and global options make code harder to understand because they increase the number of lines you need to read to understand how a single line of code will behave. Instead you might want to consider packages from the tidyverse: they never automatically convert strings to factors.

While factors look like (and often behave like) character vectors, they are actually integers. Be careful when treating them like strings. Some string methods (like `gsub()` and `grepl()`) will coerce factors to strings, 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.

4.3.2 Exercises

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()`.)

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))
```

4.4 Matrices and arrays

Adding a `dim` attribute to an atomic vector allows it to behave like a multi-dimensional **array**. A special case of the array is the **matrix**, which has two dimensions. Matrices are used commonly as part of the mathematical machinery of statistics. Arrays are much rarer, but worth being aware of.

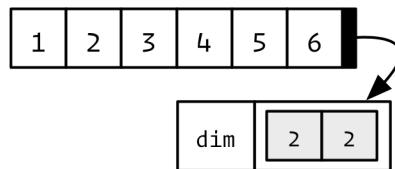
Matrices and arrays are created with `matrix()` and `array()`, or by using the assignment form of `dim()`:

```
# Two scalar arguments to specify rows and columns
a <- matrix(1:6, ncol = 3, nrow = 2)
# One vector argument to describe all dimensions
b <- array(1:12, c(2, 3, 2))

# 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
dim(c) <- c(2, 3)
c
#>      [,1] [,2] [,3]
#> [1,]    1    3    5
#> [2,]    2    4    6
```

To be technically correct, when drawing the matrix `a`, I should draw it like so:



However, dimensions, like names, are special, so it's usually easier to elide this detail and draw matrices as 2d structures:

| | | |
|---|---|---|
| 1 | 3 | 5 |
| 2 | 4 | 6 |

It's really hard to draw arrays, but fortunately they're not used in this book. Matrices and arrays are most useful for mathematical calculations (particularly when fitting models); lists are a better fit for most other programming tasks in R.

`length()` and `names()` have high-dimensional generalisations:

- `length()` generalises to `nrow()` and `ncol()` for matrices, and `dim()` for arrays.
- `names()` generalises to `rownames()` and `colnames()` for matrices, and `dimnames()`, a list of character vectors, for arrays.

```
length(a)
#> [1] 6
nrow(a)
#> [1] 2
ncol(a)
#> [1] 3
rownames(a) <- c("A", "B")
colnames(a) <- c("a", "b", "c")
a
#>   a b c
#> A 1 3 5
#> B 2 4 6

length(b)
#> [1] 12
dim(b)
```

```
#> [1] 2 3 2
dimnames(b) <- list(c("one", "two"), c("a", "b", "c"), c("A", "B"))
b
#> , , A
#>
#>     a b c
#> one 1 3 5
#> two 2 4 6
#>
#> , , B
#>
#>     a   b   c
#> one 7   9   11
#> two 8   10  12
```

`c()` generalises to `cbind()` and `rbind()` for matrices, and to `abind::abind()` for arrays. You can transpose a matrix with `t()`; the generalised equivalent for arrays is `aperm()`.

You can test if an object is a matrix or array using `is.matrix()` and `is.array()`, or by looking at the length of the `dim()`. `as.matrix()` and `as.array()` make it easy to turn an existing vector into a matrix or array.

Vectors are not the only 1-dimensional data structure. You 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
```

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
```

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.

4.4.1 Exercises

1. What does `dim()` return when applied to a vector?
2. When might you use `NROW()` or `NCOL()`?
3. If `is.matrix(x)` is `TRUE`, what will `is.array(x)` return?

4. 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))
```

4.5 Data frames

A data frame is the most common way of storing data in R, and it used systematically makes data analysis easier. Under the hood, a data frame is a list of equal-length vectors. This makes it a 2-dimensional structure, so it shares properties of both the matrix and the list. This means that a data frame has `names()`, `colnames()`, and `rownames()`, although `names()` and `colnames()` are the same thing. The `length()` of a data frame is the length of the underlying list and so is the same as `ncol()`; `nrow()` gives the number of rows. As described in subsetting, you can subset a data frame like a 1d structure (where it behaves like a list), or a 2d structure (where it behaves like a matrix).

4.5.1 Creation

You create a data frame using `data.frame()`, which takes named vectors as input:

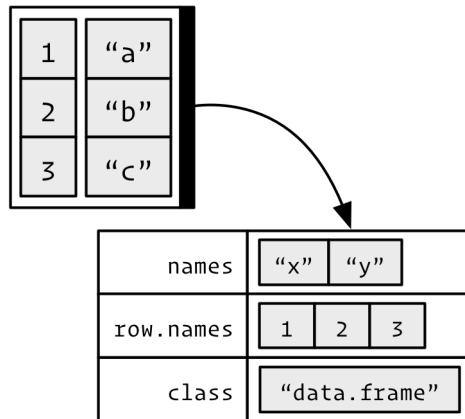
```
df <- data.frame(x = 1:3, y = c("a", "b", "c"))
str(df)
#> 'data.frame': 3 obs. of 2 variables:
#> $ x: int 1 2 3
#> $ y: Factor w/ 3 levels "a","b","c": 1 2 3
```

Beware `data.frame()`'s default behaviour which turns strings into factors. Use `stringsAsFactors = FALSE` to suppress this behaviour:

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

Data frames are named lists with attributes providing the (column) `names`, `row.names`, and a class of “`data.frame`”:

```
typeof(df)
#> [1] "list"
attributes(df)
#> $names
#> [1] "x" "y"
#>
#> $row.names
#> [1] 1 2 3
#>
#> $class
#> [1] "data.frame"
```



But usually these details are not important so I'll draw data frames in the same way as a named list, but arranged to emphasised the columnar structure.

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

4.5.2 Testing and coercion

Because a `data.frame` is an S3 class, its type reflects the underlying vector used to build it: the list. To check if an object is a data frame, use `is.data.frame()`:

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

You can coerce an object to a data frame with `as.data.frame()`:

- A vector will create a one-column data frame.
- A list will create one column for each element; it's an error if they're not all the same length.
- A matrix will create a data frame with the same number of columns and rows as the matrix.

4.5.3 Combining data frames

You can combine data frames using `cbind()` and `rbind()`:

```
cbind(df, data.frame(z = 3:1))
#>   x y z
#> 1 1 a 3
#> 2 2 b 2
#> 3 3 c 1
rbind(df, data.frame(x = 10, y = "z"))
#>   x y
#> 1 1 a
```

```
#> 2 2 b
#> 3 3 c
#> 4 10 z
```

When combining column-wise, the number of rows must match, but row names are ignored. When combining row-wise, both the number and names of columns must match. Use `dplyr::bind_rows()`, `data.table::rbindlist()`, or similar to combine data frames that don't have the same columns.

It's a common mistake to try and create a data frame by `cbind()`ing vectors together. This is unlikely to do what you want because `cbind()` will create a matrix unless one of the arguments is already a data frame. Instead use `data.frame()` directly:

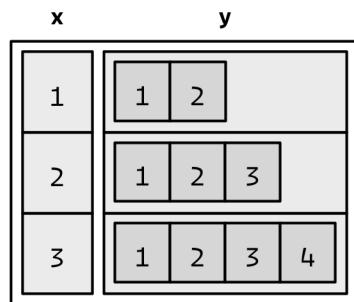
```
# This is always a mistake
bad <- data.frame(cbind(a = 1:2, b = c("a", "b")))
str(bad)
#> 'data.frame': 2 obs. of 2 variables:
#> $ a: Factor w/ 2 levels "1","2": 1 2
#> $ b: Factor w/ 2 levels "a","b": 1 2

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

4.5.4 List and matrix 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 a powerful technique because a list can contain any other R object. This means that you can have a column of data frames, or model objects, or even functions!

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



However, when a list is given to `data.frame()`, it tries to put each item of the list into its own column, so this fails:

```
data.frame(x = 1:3, y = list(1:2, 1:3, 1:4))
#> Error in (function (... , row.names = NULL, check.rows = FALSE, check.names = TRUE, : arguments imply
```

A workaround is to use `I()`, which causes `data.frame()` to treat the list as one unit:

```
df1 <- data.frame(x = 1:3, y = I(list(1:2, 1:3, 1:4)))
str(df1)
#> 'data.frame': 3 obs. of 2 variables:
#> $ x: int 1 2 3
#> $ y:List of 3
#> ..$ : int 1 2
#> ..$ : int 1 2 3
#> ..$ : int 1 2 3 4
#> ...- attr(*, "class")= chr "AsIs"
```

`I()` adds the `AsIs` class to its input, but this can usually be safely ignored.

Similarly, 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:

```
dfm <- data.frame(x = 1:3 * 10, y = I(matrix(1:9, nrow = 3)))
str(dfm)
#> 'data.frame': 3 obs. of 2 variables:
#> $ x: num 10 20 30
#> $ y: 'AsIs' int [1:3, 1:3] 1 2 3 4 5 6 7 8 9
```

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

Use list and array columns with caution. Many functions that work with data frames assume that all columns are atomic vectors, and the printed display can be confusing.

```
df1[2, ]
#>   x     y
#> 2 2 1, 2, 3
dfm[2, ]
#>   x y.1 y.2 y.3
#> 2 20    2    5    8
```

4.5.5 Tibbles

Data frames have a number of frustrating behaviours; things that made sense at the time data frames were created but now cause friction. To reduce that frustration, the tidyverse provides a modern reimagining of a data frame, called the tibble.

```
library(tibble)
```

Tibbles behave as similarly as possible to data frames (so you can use them with existing code), but tibbles:

- Never coerce their inputs. This makes them easier to use with character vectors and lists.

```
tibble(
  x = c("one", "two", "three"),
  y = list(1:3, letters, list())
)
#> # A tibble: 3 x 2
#>   x     y
#>   <chr> <list>
#> 1 one   <int [3]>
#> 2 two   <chr [26]>
#> 3 three <list [0]>
```

- Have a better print method which (by default) only shows the first 10 rows, prints the column types, has better defaults for list columns, and thoughtfully format columns for improved readability.

```
ggplot2::diamonds
#> # A tibble: 53,940 x 10
#>   carat cut      color clarity depth table price     x     y     z
#>   <dbl> <ord>    <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
#> 1 0.230 Ideal    E     SI2     61.5  55.   326   3.95  3.98  2.43
#> 2 0.210 Premium  E     SI1     59.8  61.   326   3.89  3.84  2.31
#> 3 0.230 Good    E     VS1     56.9  65.   327   4.05  4.07  2.31
#> 4 0.290 Premium I     VS2     62.4  58.   334   4.20  4.23  2.63
#> 5 0.310 Good    J     SI2     63.3  58.   335   4.34  4.35  2.75
#> 6 0.240 Very Good J    VVS2    62.8  57.   336   3.94  3.96  2.48
#> 7 0.240 Very Good I    VVS1    62.3  57.   336   3.95  3.98  2.47
#> 8 0.260 Very Good H    SI1     61.9  55.   337   4.07  4.11  2.53
#> 9 0.220 Fair     E     VS2     65.1  61.   337   3.87  3.78  2.49
#> 10 0.230 Very Good H   VS1     59.4  61.   338   4.00  4.05  2.39
#> # ... with 53,930 more rows
```

- Tibbles tweak the behaviour of [and \$ to be more consistent: [will always return another tibble, and \$ will warn if a column does not exist.

At time of writing, tibbles do not support matrix columns.

4.5.6 Exercises

1. What attributes does a data frame possess?
2. What does as.matrix() do when applied to a data frame with columns of different types? How does it differ from data.matrix()?
3. Can you have a data frame with 0 rows? What about 0 columns?

4.6 Answers

1. The three properties of a vector are type, length, and attributes.
2. The four common types of atomic vector are logical, integer, double (sometimes called numeric), and character. The two rarer types are complex and raw.
3. 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().

4. 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.
5. 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()))`.
6. Tibbles have an enhanced print method, never coerce strings to factors, and provide stricter subsetting methods.

Chapter 5

Subsetting

5.1 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 hard to learn because you need to master a number of interrelated concepts:

- The three subsetting operators.
- The six types of subsetting.
- Important differences in behaviour for different objects (e.g., vectors, lists, factors, matrices, and data frames).
- The use of subsetting in conjunction 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.

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 answers.

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

- Data types 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, data frames, and S3 objects.
- Subsetting operators expands your knowledge of subsetting operators to include `[[` and `$`, focussing on the important principles of simplifying vs. preserving.
- In Subsetting and assignment you'll learn the art of subassignment, combining subsetting and assignment to modify parts of an object.
- Applications leads you through eight important, but not obvious, applications of subsetting to solve problems that you often encounter in a data analysis.

5.2 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. Selecting a single element will cover `[[` and `$`, used to extra a single element from a data structure.

5.2.1 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 five 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 write the 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 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 vectors but 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
```

5.2.2 Lists

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

Matrices and arrays {#matrix-subsetting}

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 (>2d) 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. See [simplifying vs. preserving](#) to learn how to avoid this.

Because matrices and arrays are implemented as vectors with special attributes, you can subset them with a single vector. In that case, they will behave like a vector. Arrays in R are stored in column-major order:

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

```
vals <- outer(1:5, 1:5, FUN = "paste", sep = ",")
select <- matrix(ncol = 2, byrow = TRUE, c(
  1, 1,
  3, 1,
  2, 4
))
vals[select]
#> [1] "1,1" "3,1" "2,4"
```

5.2.3 Data frames

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:
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
```

5.2.4 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 as described below. 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
```

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
```

5.2.5 S3 objects

S3 objects are made up of atomic vectors, arrays, and lists, so you can always pull apart an S3 object using the techniques described above and the knowledge you gain from `str()`.

5.2.6 S4 objects

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.

5.2.7 Exercises

- 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, ]
```

- Why does `x <- 1:5; x[NA]` yield five missing values? (Hint: why is it different from `x[NA_real_]`?)
- 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?

5.3 Selecting a single element

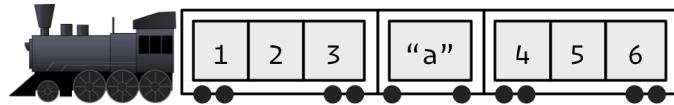
There are two other subsetting operators: `[[` and `$`. `[[` is used for extracting single values, and `$` is a useful shorthand for `[[` combined with character subsetting. `[[` is most important working with lists because subsetting a list with `[` always returns a smaller list. To help make this easier 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.”

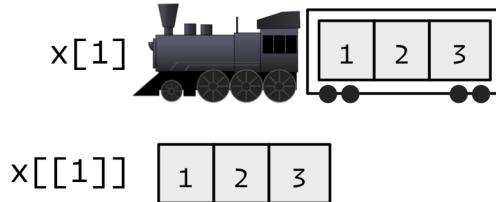
— @RLangTip, <https://twitter.com/RLangTip/status/268375867468681216>

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

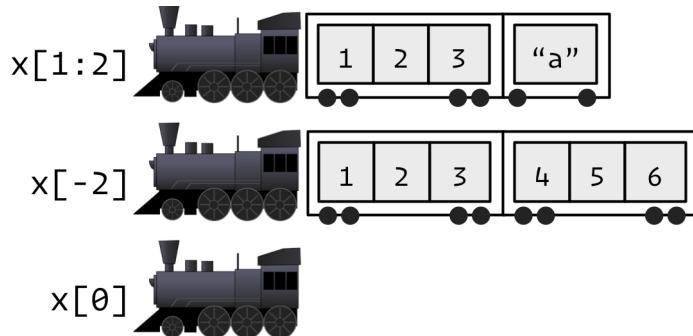
```
x <- list(1:3, "a", 4:5)
```



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 value, you must use `[[` with either a single positive integer or a string. Because data frames are lists of columns, you can use `[[` to extract a column from data frames: `mtcars[[1]]`, `mtcars[["cyl"]]`. S3 and S4 objects can override the standard behaviour of `[` and `[[` so they behave differently for different types of objects. 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 value. That frequently arises in for loops, i.e. 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]])
}
```

5.3.1 \$

`$` is a shorthand operator: `x$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 try and 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 specifically, you can avoid this problem by using tibbles instead: they never do partial matching.)

5.3.2 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 |
|----------|-------------------------|-------------------------|-------------------------|
| NULL | NULL | NULL | NULL |
| Logical | <code>logical(0)</code> | NA | NA |
| List | <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 |
|------------|-------------|-----------|-----------|---------|
| NULL | NULL | NULL | NULL | NULL |
| Atomic | Error | Error | Error | Error |
| List | Error | Error | NULL | NULL |

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

5.3.3 `pluck()`

The inconsistency of the `[[` table above lead to the development of `purrr::pluck()`, which solves the inconsistency by always returning `NULL`:

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

(A future function will solve the inconsistency in the other direction: by consistently throwing an error whenever the component is absent.)

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 (this 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
```

5.3.4 Exercises

1. Come up with 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))

5.4 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 original object class and structure. 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
```

5.5 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 that are not dealt with by existing functions.

5.5.1 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.

5.5.2 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()`.

5.5.3 Random samples/bootstrap (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 = rep(1:3, each = 2), y = 6:1, z = letters[1:6])

# Randomly reorder
df[sample(nrow(df)), ]
#>   x y z
#> 1 1 6 a
#> 5 3 2 e
#> 3 2 4 c
#> 6 3 1 f
#> 4 2 3 d
#> 2 1 5 b
#> 2 1 5 b

# Select 3 random rows
df[sample(nrow(df), 3), ]
#>   x y z
#> 3 2 4 c
#> 2 1 5 b
#> 6 3 1 f

# Select 6 bootstrap replicates
df[sample(nrow(df), 6, rep = TRUE), ]
#>   x y z
#> 5 3 2 e
#> 6 3 1 f
#> 2 1 5 b
#> 1 1 6 a
#> 2.1 1 5 b
#> 3 2 4 c

```

The arguments of `sample()` control the number of samples to extract, and whether sampling is performed

with or without replacement.

5.5.4 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
#> 2 b 5 1
#> 3 c 4 2
#> 1 a 6 1
#> 6 f 1 3
#> 5 e 2 3
#> 4 d 3 2

df2[order(df2$x), ]
#>   z y x
#> 2 b 5 1
#> 1 a 6 1
#> 3 c 4 2
#> 4 d 3 2
#> 6 f 1 3
#> 5 e 2 3
df2[, order(names(df2))]
#>   x y z
#> 2 1 5 b
#> 3 2 4 c
#> 1 1 6 a
#> 6 3 1 f
#> 5 3 2 e
#> 4 2 3 d
```

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

5.5.5 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
```

5.5.6 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 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
```

5.5.7 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, 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`.

`subset()` is a specialised shorthand function for subsetting data frames, and saves some typing because you don't need to repeat the name of the data frame. You'll learn how it works in metaprogramming.

```
subset(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

subset(mtcars, gear == 5 & 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
```

5.5.8 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] 3 6 7

unwhich <- function(x, n) {
  out <- rep_len(FALSE, n)
  out[x] <- TRUE
  out
}
unwhich(which(x), 10)
#> [1] FALSE FALSE  TRUE FALSE FALSE  TRUE  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 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. First, when the logical vector contains NA, logical subsetting replaces these values by NA while `which()` drops these values. Second, `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.

5.5.9 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?

5.6 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 6

Functions

6.1 Introduction

Functions are a fundamental building block of R: to master many of the more advanced techniques in this book, you need a solid foundation in how functions work. You've probably already created many R functions, and you're familiar with the basics of how they work. The focus of this chapter is to turn your existing, informal knowledge of functions into a rigorous understanding of what functions are and how they work. You'll see some interesting tricks and techniques in this chapter, but most of what you'll learn will be more important as the building blocks for more advanced techniques.

The most important thing to understand about R is that functions are objects in their own right. You can work with them exactly the same way you work with any other type of object. This theme will be explored in depth in functional programming.

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 answers.

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 more typically 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 function throw an error when called? 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. What function do you use to ensure that a cleanup action occurs regardless of how a function terminates?

Outline

- Function components describes the three main components of a function.
- Lexical scoping teaches you how R finds values from names, the process of lexical scoping.
- Every operation is a function call shows you that everything that happens in R is a result of a function call, even if it doesn't look like it.
- Function arguments discusses the three ways of supplying arguments to a function, how to call a function given a list of arguments, and the impact of lazy evaluation.
- Special calls describes two special types of function: infix and replacement functions.
- Return values discusses how and when functions return values, and how you can ensure that a function does something before it exits.

Prerequisites

The only package you'll need is `pryr`, which is used to explore what happens when modifying vectors in place. Install it with `install.packages("pryr")`.

6.2 Function components

All R functions have three parts:

- the `body()`, the code inside the function.
- the `formals()`, the list of arguments which controls how you can call the function.
- the `environment()`, the “map” of the location of the function’s variables.

When you print a function in R, it shows you these three important components. If the environment isn’t displayed, it means that the function was created in the global environment.

```
f <- function(x) x^2
f
#> function(x) x^2

formals(f)
#> $x
body(f)
#> x^2
environment(f)
#> <environment: R_GlobalEnv>
```

The assignment forms of `body()`, `formals()`, and `environment()` can also be used to modify functions.

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, which points to the source code used to create the function. Unlike `body()`, this contains code comments and other formatting. You can also add attributes to a function. For example, you can set the `class()` and add a custom `print()` method.

6.2.1 Primitive functions

There is one exception to the rule that functions have three components. Primitive functions, like `sum()`, call C code directly with `.Primitive()` and contain no R code. Therefore their `formals()`, `body()`, and `environment()` are all `NULL`:

```
sum
#> function (... , na.rm = FALSE) .Primitive("sum")
formals(sum)
#> NULL
body(sum)
#> NULL
environment(sum)
#> NULL
```

Primitive functions are only found in the `base` package, and since they operate at a low level, they can be more efficient (primitive replacement functions don’t have to make copies), and can have different rules for argument matching (e.g., `switch` and `call`). This, however, comes at a cost of behaving differently from all other functions in R. Hence the R core team generally avoids creating them unless there is no other option.

6.2.2 Exercises

1. 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?

2. This code makes a list of all functions in the base package.

```
objs <- mget(ls("package:base"), inherits = TRUE)
fun <- 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?
3. What are the three important components of a function?
 4. When does printing a function not show what environment it was created in?

6.3 Lexical scoping

Assignment is the act of binding a name to a value. Scoping is the opposite; finding a value given a name.

Scoping is the set of rules that govern how R looks up the value of a symbol. In the example below, scoping is the set of rules that R applies to go from the symbol `x` to its value 10:

```
x <- 10
x
#> [1] 10
```

Understanding scoping allows you to:

- build tools by composing functions, as described in functional programming.
- overrule the usual evaluation rules and do non-standard evaluation, as described in non-standard evaluation.

R has two types of scoping: **lexical scoping**, implemented automatically at the language level, and **dynamic scoping**, used in select functions to save typing during interactive analysis. We discuss lexical scoping here because it is intimately tied to function creation. Dynamic scoping is described in more detail in scoping issues.

Lexical scoping looks up symbol values based on how functions were nested when they were created, not how they are nested when they are called. With lexical scoping, you don't need to know how the function is called to figure out where the value of a variable will be looked up. You just need to look at the function's definition.

The “lexical” in lexical scoping doesn't correspond to the usual English definition (“of or relating to words or the vocabulary of a language as distinguished from its grammar and construction”) but comes from the computer science term “lexing”, which is part of the process that converts code represented as text to meaningful pieces that the programming language understands.

There are four basic principles behind R's implementation of lexical scoping:

- name masking
- functions vs. variables
- a fresh start
- dynamic lookup

You probably know many of these principles already, although you might not have thought about them explicitly. Test your knowledge by mentally running through the code in each block before looking at the answers.

6.3.1 Name masking

The following example illustrates the most basic principle of lexical scoping, and you should have no problem predicting the output.

```
f <- function() {
  x <- 1
  y <- 2
  c(x, y)
}
f()
rm(f)
```

If a name isn't defined inside a function, R will look one level up.

```
x <- 2
g <- function() {
  y <- 1
  c(x, y)
}
```

```
g()
rm(x, g)
```

The same rules apply if a function is defined inside another function: look inside the current function, then where that function was defined, and so on, all the way up to the global environment, and then on to other loaded packages. Run the following code in your head, then confirm the output by running the R code.

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

The same rules apply to closures, functions created by other functions. Closures will be described in more detail in functional programming; here we'll just look at how they interact with scoping. The following function, `j()`, returns a function. What do you think this function will return when we call it?

```
j <- function(x) {
  y <- 2
  function() {
    c(x, y)
  }
}
k <- j(1)
k()
rm(j, k)
```

This seems a little magical (how does R know what the value of `y` is after the function has been called). It works because `k` preserves the environment in which it was defined and because the environment includes the value of `y`. Environments gives some pointers on how you can dive in and figure out what values are stored in the environment associated with each function.

6.3.2 Functions vs. variables

The same principles apply regardless of the type of associated value — finding functions works exactly the same way as finding variables:

```
l <- function(x) x + 1
m <- function() {
  l <- function(x) x * 2
  l(10)
}
m()
#> [1] 20
rm(l, m)
```

For functions, there is one small tweak to the rule. If you are using a name in a context where it's obvious that you want a function (e.g., `f(3)`), R will ignore objects that are not functions while it is searching. In

the following example `n` takes on a different value depending on whether R is looking for a function or a variable.

```
n <- function(x) x / 2
o <- function() {
  n <- 10
  n(n)
}
o()
#> [1] 5
rm(n, o)
```

However, using the same name for functions and other objects will make for confusing code, and is generally best avoided.

6.3.3 A fresh start

What happens to the values in between invocations of a function? What will happen the first time you run this function? What will happen the second time? (If you haven't seen `exists()` before: it returns TRUE if there's a variable of that name, otherwise it returns FALSE.)

```
j <- function() {
  if (!exists("a")) {
    a <- 1
  } else {
    a <- a + 1
  }
  a
}
j()
rm(j)
```

You might be surprised that it returns the same value, 1, every time. This is because every time a function is called, a new environment is created to host execution. 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 mutable state.)

6.3.4 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 be different depending on objects outside its environment:

```
f <- function() x
x <- 15
f()
#> [1] 15

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

You generally want to avoid this behaviour because it means the function is no longer self-contained. This is a common error — 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 the `findGlobals()` function from `codetools`. This function lists all the external dependencies of a function:

```
f <- function() x + 1
codetools::findGlobals(f)
#> [1] "+" "x"
```

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

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

This doesn't work because R relies on lexical scoping to find everything, even the `+` operator. It's never possible to make a function completely self-contained because you must always rely on functions defined in base R or other packages.

You can use this same idea to do other things that are extremely ill-advised. For example, since all of the standard operators in R are functions, you can override them with your own alternatives. If you ever are feeling particularly evil, run the following code while your friend is away from their computer:

```
^(<- function(e1) {
  if (is.numeric(e1) && runif(1) < 0.1) {
    e1 + 1
  } else {
    e1
  }
}
replicate(50, (1 + 2))
#> [1] 4 3 3 3 4 3 3 3 3 3 3 3 3 4 3 3 3 3 3 3 3 3 3 3 3 3 4 3 3 3 4 3 3 3 3 3
#> [36] 3 3 3 3 3 3 3 3 3 4 3 3 3 3 3 3
rm("e")
```

This will introduce a particularly pernicious bug: 10% of the time, 1 will be added to any numeric calculation inside parentheses. This is another good reason to regularly restart with a clean R session!

6.3.5 Exercises

- What does the following code return? Why? What does each of the three `c`'s mean?

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

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

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

```
f(x) * 2
}
f(10)
```

6.4 Every operation is a function call

“To understand computations in R, two slogans are helpful:

- Everything that exists is an object.
- Everything that happens is a function call.”

— John Chambers

The previous example of redefining `(` works because every operation in R is a function call, whether or not it looks like one. This includes infix operators like `+`, control flow operators like `for`, `if`, and `while`, subsetting operators like `[]` and `$`, and even the curly brace `{`. This means that each pair of statements in the following example is exactly equivalent. Note that ```, the backtick, lets you refer to functions or variables that have otherwise reserved or illegal names:

```
x <- 10; y <- 5
x + y
#> [1] 15
`+`(x, y)
#> [1] 15

for (i in 1:2) print(i)
#> [1] 1
#> [1] 2
`for`(i, 1:2, print(i))
#> [1] 1
#> [1] 2

if (i == 1) print("yes!") else print("no.")
#> [1] "no."
`if`(i == 1, print("yes!"), print("no."))
#> [1] "no."

x[3]
#> [1] NA
`[`(x, 3)
#> [1] NA

{ print(1); print(2); print(3) }
#> [1] 1
#> [1] 2
#> [1] 3
`{`(`print(1), print(2), print(3))
#> [1] 1
#> [1] 2
#> [1] 3
```

It is possible to override the definitions of these special functions, but this is almost certainly a bad idea. However, there are occasions when it might be useful: it allows you to do something that would have otherwise been impossible. For example, this feature makes it possible for the `dplyr` package to translate

R expressions into SQL expressions. Domain specific languages uses this idea to create domain specific languages that allow you to concisely express new concepts using existing R constructs.

It's more often useful to treat special functions as ordinary functions. For example, we 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 using the built-in `+` function.

```
sapply(1:5, `+`, 3)
#> [1] 4 5 6 7 8
sapply(1:5, "+", 3)
#> [1] 4 5 6 7 8
```

Note the difference between ``+`` and `"+"`. The first one is the value of the object called `+`, and the second is a string containing the character `+`. The second version works because `sapply` can be given the name of a function instead of the function itself: if you read the source of `sapply()`, you'll see the first line uses `match.fun()` to find functions given their names.

A more useful application is to combine `lapply()` or `sapply()` with subsetting:

```
x <- list(1:3, 4:9, 10:12)
sapply(x, "[", 2)
#> [1] 2 5 11

# equivalent to
sapply(x, function(x) x[2])
#> [1] 2 5 11
```

Remembering that everything that happens in R is a function call will help you in metaprogramming.

6.5 Function arguments

It's useful to distinguish between the formal arguments and the actual arguments of a function. The formal arguments are a property of the function, whereas the actual or calling arguments can vary each time you call the function. This section discusses how calling arguments are mapped to formal arguments, how you can call a function given a list of arguments, how default arguments work, and the impact of lazy evaluation.

6.5.1 Calling functions

When calling a function you can specify arguments by position, by complete name, or by partial name. Arguments are matched first by exact name (perfect matching), then by prefix matching, and finally by position.

```
f <- function(abcdef, bcde1, bcde2) {
  list(a = abcdef, b1 = bcde1, b2 = bcde2)
}
str(f(1, 2, 3))
#> List of 3
#> $ a : num 1
#> $ b1: num 2
#> $ b2: num 3
```

```

str(f(2, 3, abcdef = 1))
#> List of 3
#> $ a : num 1
#> $ b1: num 2
#> $ b2: num 3

# Can abbreviate long argument names:
str(f(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(f(1, 3, b = 1))
#> Error in f(1, 3, b = 1): argument 3 matches multiple formal arguments

```

Generally, you only want to 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 only use readable abbreviations with partial matching. (If you are writing code for a package that you want to publish on CRAN you can not use partial matching, and must use complete names.) Named arguments should always come after unnamed arguments. If a function uses . . . (discussed in more detail below), you can only specify arguments listed after . . . with their full name.

These are good calls:

```

mean(1:10)
mean(1:10, trim = 0.05)

```

This is probably overkill:

```
mean(x = 1:10)
```

And these are just confusing:

```

mean(1:10, n = T)
mean(1:10, , FALSE)
mean(1:10, 0.05)
mean(, TRUE, x = c(1:10, NA))

```

6.5.2 Calling a function given a list of arguments

Suppose you had a list of function arguments:

```
args <- list(1:10, na.rm = TRUE)
```

How could you then send that list to `mean()`? You need `do.call()`:

```

do.call(mean, args)
#> [1] 5.5
# Equivalent to
mean(1:10, na.rm = TRUE)
#> [1] 5.5

```

6.5.3 Default and missing arguments

Function arguments in R can have default values.

```
f <- function(a = 1, b = 2) {
  c(a, b)
}
f()
#> [1] 1 2
```

Since arguments in R are evaluated lazily (more on that below), the default value can be defined in terms of other arguments:

```
g <- function(a = 1, b = a * 2) {
  c(a, b)
}
g()
#> [1] 1 2
g(10)
#> [1] 10 20
```

Default arguments can even be defined in terms of variables created within the function. This is used frequently in base R functions, but I think it is bad practice, because you can't understand what the default values will be without reading the complete source code.

```
h <- function(a = 1, b = d) {
  d <- (a + 1) ^ 2
  c(a, b)
}
h()
#> [1] 1 4
h(10)
#> [1] 10 121
```

You can determine if an argument was supplied or not with the `missing()` function.

```
i <- function(a, b) {
  c(missing(a), missing(b))
}
i()
#> [1] TRUE TRUE
i(a = 1)
#> [1] FALSE TRUE
i(b = 2)
#> [1] TRUE FALSE
i(1, 2)
#> [1] FALSE FALSE
```

Sometimes you want to add a non-trivial default value, which might take several lines of code to compute. Instead of inserting that code in the function definition, you could use `missing()` to conditionally compute it if needed. However, this makes it hard to know which arguments are required and which are optional without carefully reading the documentation. Instead, I usually set the default value to `NULL` and use `is.null()` to check if the argument was supplied.

6.5.4 Lazy evaluation

By default, R function arguments are lazy — they're only evaluated if they're actually used:

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

If you want to ensure that an argument is evaluated you can use `force()`:

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

The `apply` functions underwent this same change in R 3.2.0:

Higher order functions such as the `apply` functions and `Reduce()` now force arguments to the functions they apply in order to eliminate undesirable interactions between lazy evaluation and variable capture in closures.

So, as of R 3.2.0 (but not older versions), you can safely do:

```
add <- function(x) {
  function(y) x + y
}
adders <- lapply(1:10, add)
adders[[1]](10)
#> [1] 11
adders[[10]](10)
#> [1] 20
```

Fortunately, all good! The lesson here is that you need to keep lazy evaluation in mind when creating closures with a loop or any other construct (unless you know that these, like the `apply` family, force their functions' arguments). For example, here's a naive implementation that wants to achieve the same result as above, using a `for` loop instead of `lapply`:

```
add <- function(x) {
  function(y) x + y
}

adders <- list()
for (i in 1:10) {
  adders[[i]] <- add(i)
}

adders[[1]](10)
#> [1] 20
adders[[10]](10)
#> [1] 20
```

`x` is lazily evaluated the first time that you call one of the adder functions. At this point, the loop is complete and the final value of `x` is 10. Therefore all of the adder functions will add 10 on to their input, probably not what you wanted! Manually forcing evaluation inside `add()` fixes the problem:

```

add <- function(x) {
  force(x)
  function(y) x + y
}

adders <- list()
for (i in 1:10) {
  adders[[i]] <- add(i)
}

adders[[1]](10)
#> [1] 11
adders[[10]](10)
#> [1] 20

```

The add function is exactly equivalent to

```

add <- function(x) {
  x
  function(y) x + y
}

```

because the force function is defined as `force <- function(x) x`. However, using this function clearly indicates that you're forcing evaluation, not that you've accidentally typed `x`.

Default arguments are evaluated inside the function. This means that if the expression depends on the current environment the results will differ depending on whether you use the default value or explicitly provide one.

```

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

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

# ls() evaluated in global environment:
f(ls())
#> [1] "add"           "adders"        "args"          "begin_sidebar"
#> [5] "doc_type"      "end_sidebar"   "f"             "fun"
#> [9] "g"              "h"             "i"             "obj"
#> [13] "x"             "y"

```

More technically, an unevaluated argument is called a **promise**, or (less commonly) a thunk. A promise is made up of two parts:

- The expression which gives rise to the delayed computation. (It can be accessed with `substitute()`. See non-standard evaluation for more details.)
- The environment where the expression was created and where it should be evaluated.

The first time a promise is accessed the expression is evaluated in the environment where it was created. This value is cached, so that subsequent access to the evaluated promise does not recompute the value (but the original expression is still associated with the value, so `substitute()` can continue to access it). You can find more information about a promise using `pryr::promise_info()`. This uses some C++ code to extract

information about the promise without evaluating it, which is impossible to do in pure R code.

Laziness is useful in if statements — the second statement below will be evaluated only if the first is true. If it wasn't, the statement would return an error because `NULL > 0` is a logical vector of length 0 and not a valid input to `if`.

```
x <- NULL
if (!is.null(x) && x > 0) {
}
```

We could implement “`&&`” ourselves:

```
`&&` <- function(x, y) {
  if (!x) return(FALSE)
  if (!y) return(FALSE)

  TRUE
}
a <- NULL
!is.null(a) && a > 0
#> [1] FALSE
```

This function would not work without lazy evaluation because both `x` and `y` would always be evaluated, testing `a > 0` even when `a` was `NULL`.

Sometimes you can also use laziness to eliminate an if statement altogether. For example, instead of:

```
if (is.null(a)) stop("a is null")
#> Error in eval(expr, envir, enclos): a is null
```

You could write:

```
!is.null(a) || stop("a is null")
#> Error in eval(expr, envir, enclos): a is null
```

6.5.5 ...

There is a special argument called `...`. This argument will match any arguments not otherwise matched, and can be easily passed on to other functions. This is useful if you want to collect arguments to call another function, but you don't want to prespecify their possible names. `...` is often used in conjunction with S3 generic functions to allow individual methods to be more flexible.

One relatively sophisticated user of `...` is the base `plot()` function. `plot()` is a generic method with arguments `x`, `y` and `...`. To understand what `...` does for a given function we need to read the help: “Arguments to be passed to methods, such as graphical parameters”. Most simple invocations of `plot()` end up calling `plot.default()` which has many more arguments, but also has `...`. Again, reading the documentation reveals that `...` accepts “other graphical parameters”, which are listed in the help for `par()`. This allows us to write code like:

```
plot(1:5, col = "red")
plot(1:5, cex = 5, pch = 20)
```

This illustrates both the advantages and disadvantages of `...`: it makes `plot()` very flexible, but to understand how to use it, we have to carefully read the documentation. Additionally, if we read the source code for `plot.default`, we can discover undocumented features. It's possible to pass along other arguments to `Axis()` and `box()`:

```
plot(1:5, bty = "u")
plot(1:5, labels = FALSE)
```

To capture `...` in a form that is easier to work with, you can use `list(...)`. (See capturing unevaluated dots for other ways to capture `...` without evaluating the arguments.)

```
f <- function(...) {
  names(list(...))
}
f(a = 1, b = 2)
#> [1] "a" "b"
```

Using `...` comes at a price — any misspelled arguments will not raise an error, and any arguments after `...` must be fully named. This makes it easy for typos to go unnoticed:

```
sum(1, 2, NA, na.rm = TRUE)
#> [1] NA
```

It's often better to be explicit rather than implicit, so you might instead ask users to supply a list of additional arguments. That's certainly easier if you're trying to use `...` with multiple additional functions.

6.5.6 Exercises

- 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)
```

- What does this function return? Why? Which principle does it illustrate?

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

- What does this function return? Why? Which principle does it illustrate?

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

6.6 Special calls

R supports two additional syntaxes for calling special types of functions: infix and replacement functions.

6.6.1 Infix functions

Most functions in R are “prefix” operators: the name of the function comes before the arguments. You can also create infix functions where the function name comes in between its arguments, like `+` or `-`. All user-created infix functions must start and end with `%`. R comes with the following infix functions predefined:

`%%`, `%*%`, `%/%`, `%in%`, `%o%`, `%x%`. (The complete list of built-in infix operators that don't need `%` is: `:`, `::`, `:::`, `$`, `@`, `^`, `*`, `/`, `+`, `-`, `>`, `>=`, `<`, `<=`, `==`, `!=`, `!`, `&`, `&&`, `|`, `||`, `~`, `<-`, `<<-`)

For example, we could create a new operator that pastes together strings:

```
`%+%` <- function(a, b) paste0(a, b)
"new" %+% " string"
#> [1] "new string"
```

Note that when creating the function, you have to put the name in backticks because it's a special name. This is just a syntactic sugar for an ordinary function call; as far as R is concerned there is no difference between these two expressions:

```
"new" %+% " string"
#> [1] "new string"
`%+%`("new", " string")
#> [1] "new string"
```

Or indeed between

```
1 + 5
#> [1] 6
`+`(1, 5)
#> [1] 6
```

The names of infix functions are more flexible than regular R functions: they can contain any sequence of characters (except “%”, of course). 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)
`%/\\%` <- function(a, b) paste(a, b)

"a" % % "b"
#> [1] "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's one infix function that I use very often. It's inspired by Ruby's `||` logical or operator, although it works a little differently in R because Ruby has a more flexible definition of what evaluates to TRUE in an if statement. It's useful as a way of providing a default value in case the output of another function is NULL:

```
`%||%` <- function(a, b) if (!is.null(a)) a else b
function_that_might_return_null() %||% default_value
```

6.6.2 Replacement functions

Replacement functions act like they modify their arguments in place, and have the special name `xxx<-`. They typically have two arguments (`x` and `value`), although they can have more, and they 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
}
x <- 1:10
second(x) <- 5L
x
#> [1] 1 5 3 4 5 6 7 8 9 10
```

When R evaluates the assignment `second(x) <- 5`, it notices that the left hand side of the `<-` is not a simple name, so it looks for a function named `second<-` to do the replacement.

I say they “act” like they modify their arguments in place, because they actually create a modified copy. We can see that by using `pryr::address()` to find the memory address of the underlying object.

```
library(pryr)
x <- 1:10
address(x)
#> [1] "0x55ddb99f8670"
second(x) <- 6L
address(x)
#> [1] "0x55ddb9b22550"
```

Built-in functions that are implemented using `.Primitive()` will modify in place:

```
x <- 1:10
address(x)
#> [1] "0x103945110"

x[2] <- 7L
address(x)
#> [1] "0x103945110"
```

It’s important to be aware of this behaviour since it has important performance implications.

If you want to supply additional arguments, they go in between `x` and `value`:

```
`modify<-` <- function(x, position, value) {
  x[position] <- value
  x
}
modify(x, 1) <- 10
x
#> [1] 10 7 3 4 5 6 7 8 9 10
```

When you call `modify(x, 1) <- 10`, behind the scenes R turns it into:

```
x <- `modify<-`(x, 1, 10)
```

This means you can’t do things like:

```
modify(get("x"), 1) <- 10
```

because that gets turned into the invalid code:

```
get("x") <- `modify<-`(get("x"), 1, 10)
```

It’s often useful to combine replacement and subsetting:

6.6.3 Exercises

1. Create a list of all the replacement functions found in the base package. Which ones are primitive functions?
2. What are valid names for user-created infix functions?
3. Create an infix `xor()` operator.
4. Create infix versions of the set functions `intersect()`, `union()`, and `setdiff()`.
5. Create a replacement function that modifies a random location in a vector.

6.7 Return values

The last expression evaluated in a function becomes the return value, the result of invoking the function.

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

Generally, I think it's good style to reserve the use of an explicit `return()` for when you are returning early, such as for an error, or a simple case of the function. This style of programming can also reduce the level of indentation, and generally make functions easier to understand because you can reason about them locally.

```
f <- function(x, y) {
  if (!x) return(y)

  # complicated processing here
}
```

Functions can return only a single object. But this is not a limitation because you can return a list containing any number of objects.

The functions that are the easiest to understand and reason about are pure functions: functions that always map the same input to the same output and have no other impact on the workspace. In other words, pure functions have no **side effects**: they don't affect the state of the world in any way apart from the value they return.

R protects you from one type of side effect: most R objects have copy-on-modify semantics. So modifying a function argument does not change the original value:

```
f <- function(x) {
  x$a <- 2
  x
}
x <- list(a = 1)
f(x)
```

```
#> $a
#> [1] 2
x$a
#> [1] 1
```

(There are two important exceptions to the copy-on-modify rule: environments and reference classes. These can be modified in place, so extra care is needed when working with them.)

This is notably different to languages like Java where you can modify the inputs of a function. This copy-on-modify behaviour has important performance consequences which are discussed in depth in profiling. (Note that the performance consequences are a result of R's implementation of copy-on-modify semantics; they are not true in general. Clojure is a new language that makes extensive use of copy-on-modify semantics with limited performance consequences.)

Most base R functions are pure, with a few notable exceptions:

- `library()` which loads a package, and hence modifies the search path.
- `setwd()`, `Sys.setenv()`, `Sys.setlocale()` which change the working directory, environment variables, and the locale, respectively.
- `plot()` and friends which produce graphical output.
- `write()`, `write.csv()`, `saveRDS()`, etc. which save output to disk.
- `options()` and `par()` which modify global settings.
- S4 related functions which modify global tables of classes and methods.
- Random number generators which produce different numbers each time you run them.

It's generally a good idea to minimise the use of side effects, and where possible, to minimise the footprint of side effects by separating pure from impure functions. Pure functions are easier to test (because all you need to worry about are the input values and the output), and are less likely to work differently on different versions of R or on different platforms. For example, this is one of the motivating principles of ggplot2: most operations work on an object that represents a plot, and only the final `print` or `plot` call has the side effect of actually drawing the plot.

Functions can return `invisible` values, which are not printed out by default when you call the function.

```
f1 <- function() 1
f2 <- function() invisible(1)

f1()
#> [1] 1
f2()
f1() == 1
#> [1] TRUE
f2() == 1
#> [1] TRUE
```

You can force an invisible value to be displayed by wrapping it in parentheses:

```
(f2())
#> [1] 1
```

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

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

This is what makes it possible to assign one value to multiple variables:

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

because this is parsed as:

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

6.7.1 On exit

As well as returning a value, functions can set up other triggers to occur when the function is finished using `on.exit()`. This is often used as a way to guarantee that changes to the global state are restored when the function exits. The code in `on.exit()` is run regardless of how the function exits, whether with an explicit (early) return, an error, or simply reaching the end of the function body.

```
in_dir <- function(dir, code) {  
  old <- setwd(dir)  
  on.exit(setwd(old))  
  
  force(code)  
}  
getwd()  
#> [1] "/home/rstudio/adv-r"  
in_dir("~/", getwd())  
#> [1] "/home/rstudio"
```

The basic pattern is simple:

- We first set the directory to a new location, capturing the current location from the output of `setwd()`.
- We then use `on.exit()` to ensure that the working directory is returned to the previous value regardless of how the function exits.
- Finally, we explicitly force evaluation of the code. (We don't actually need `force()` here, but it makes it clear to readers what we're doing.)

Caution: If you're using multiple `on.exit()` calls within a function, make sure to set `add = TRUE`. Unfortunately, the default in `on.exit()` is `add = FALSE`, so that every time you run it, it overwrites existing exit expressions. Because of the way `on.exit()` is implemented, it's not possible to create a variant with `add = TRUE`, so you must be careful when using it.

6.7.2 Exercises

1. How does the `chdir` parameter of `source()` compare to `in_dir()`? Why might you prefer one approach to the other?
2. What function undoes the action of `library()`? How do you save and restore the values of `options()` and `par()`?
3. 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).
4. 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)
```

```
sink(temp)
on.exit(sink(), add = 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?

6.8 Quiz answers

1. The three components of a function are its body, arguments, and environment.
2. `f1(1)()` returns 11.
3. You'd normally write it in infix style: `1 + (2 * 3)`.
4. Rewriting the call to `mean(c(1:10, NA), na.rm = TRUE)` is easier to understand.
5. No, it does not throw an error because the second argument is never used so it's never evaluated.
6. See infix and replacement functions.
7. You use `on.exit()`; see `on.exit` for details.

Chapter 7

Environments

7.1 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.

7.2 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.

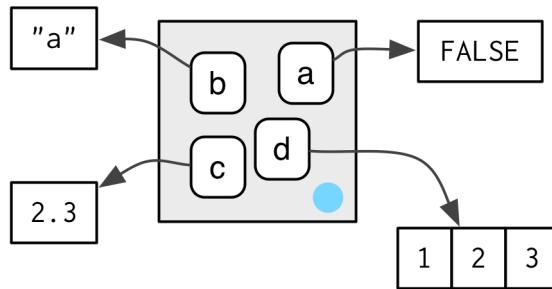
7.2.1 Basics

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,
)
```

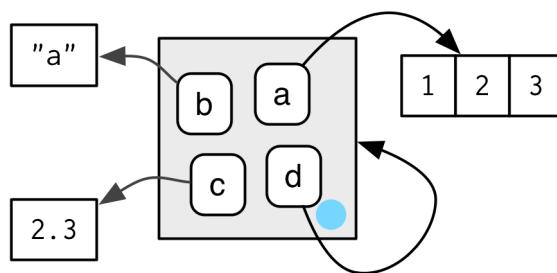
::: base Use `new.env()` to creates a new environment. Ignore the `hash` and `size` parameters; they are not needed. Note that you can not simultaneously create and define values; use `$<-`, as shown below. :::

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 them modify 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: 0x565357a57488>
```

Instead, we'll use `env_print()` which gives us a little more information:

```
env_print(e1)  
#> <environment: 0x565357a57488>  
#>   parent: <env: global>  
#>   bindings:  
#>     * a: <lgl>  
#>     * b: <chr>  
#>     * c: <dbl>  
#>     * d: <env: 0x565357a57488>
```

You can use `env_names()` to get a character vector giving the current bindings

```
env_names(e1)  
#> [1] "a" "b" "c" "d"
```

::: base 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 ... :::

7.2.2 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
```

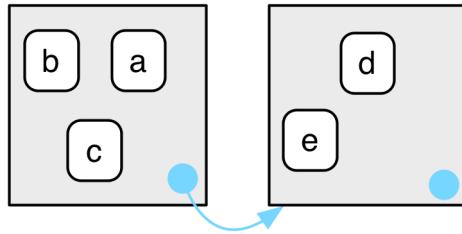
::base Access the global environment with `globalenv()` and the current environment with `environment()`. The global environment is printed as `Rf_GlobalEnv` and `.GlobalEnv`. :::

7.2.3 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)
```



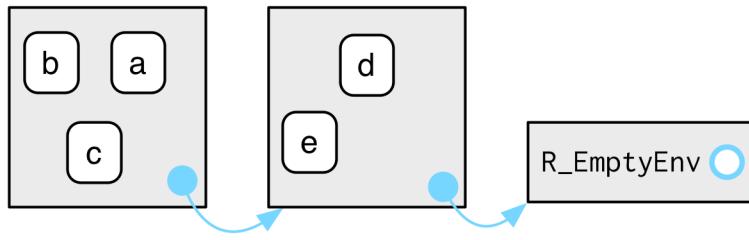
We use the metaphor of a family to name environments relative to one another. The grandparent of an environment is the parent's parent, and the ancestors include all parent environments up to the empty environment. To save space, I typically won't draw all the ancestors; just remember whenever you see 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: 0x5653586b4f30>
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 and 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: 0x5653586b4f30>
#> [[2]] $ <env: global>

env_parents(e2d)
#> [[1]] <env: 0x565357867fd8>
#> [[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.

::: base Use `parent.env()` to find the parent of an environment. No base function returns all ancestors. :::

7.2.4 Getting and setting

You can get and set elements of a 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 get(nm, envir = env, inherits = inherit): 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()`¹ takes a name (as string) and a value:

```
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.

::: base 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. :::

7.2.5 Finalisers

Add something once rlang has an API. Also mention in data structures below

7.2.6 Advanced bindings

There are two more exotic variants of `env_bind()`:

¹You might wonder why rlang has `env_poke()` instead of `env_set()`. This is for consistency: `_set()` functions return a modified copy; `_poke()` functions modify in place.

- `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
#>   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:

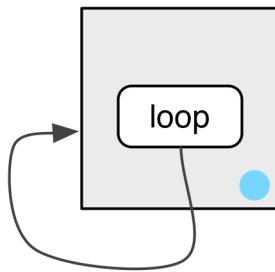
```
env_bind_fns(current_env(), z2 = function(val) {
  if (missing(val)) {
    2
  } else {
    stop("Don't touch z2!", call. = FALSE)
  }
})

z2
#> [1] 2
z2 <- 3
#> Error: Don't touch z2!
```

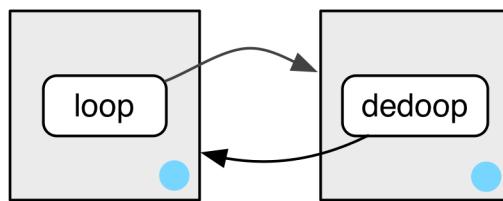
::: base See the `?delayedAssign()` and `?makeActiveBinding()`. :::

7.2.7 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.

7.3 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:

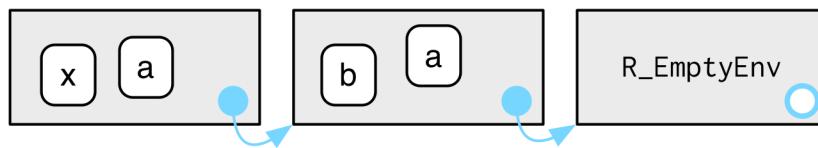
```
where("yyy")
#> Error: Can't find yyy

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, e4a)` will find `a` in `e4a`.
- `where("b", e4a)` doesn't find `b` in `e4a`, so it looks in its parent, `e4b`, and finds it there.
- `where("c", e4a)` looks in `e4a`, then `e4b`, 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))
  }
}
```

::: sidebar ### 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
}

:::
```

7.3.1 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.

7.4 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 then 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.

7.4.1 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 attached²:

```

env_parent(global_env())
#> <environment: package:rlang>
#> attr(", "name")
#> [1] "package:rlang"
#> attr(", "path")
#> [1] "/usr/local/lib/R/site-library/rlang"
```

And the parent of that package is the second to last package you attached:

```

env_parent(env_parent(global_env()))
#> <environment: package:purrr>
#> attr(", "name")
#> [1] "package:purrr"
#> attr(", "path")
#> [1] "/usr/local/lib/R/site-library/purrr"
```

²Note 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()`.

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.

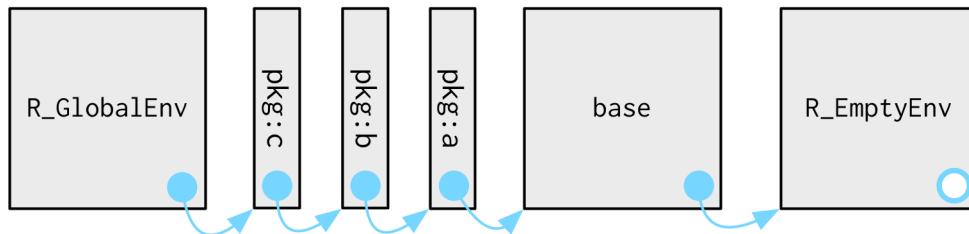
```
search_envs()
#> [[1]] $ <env: global>
#> [[2]] $ <env: package:rlang>
#> [[3]] $ <env: package:purrr>
#> [[4]] $ <env: package:methods>
#> [[5]] $ <env: package:stats>
#> [[6]] $ <env: package:graphics>
#> [[7]] $ <env: package:grDevices>
#> [[8]] $ <env: package:utils>
#> [[9]] $ <env: package:datasets>
#> [[10]] $ <env: Autoloads>
#> [[11]] $ <env: base>
```

:::base You can access the names of the environments on the search path with `search()` :::

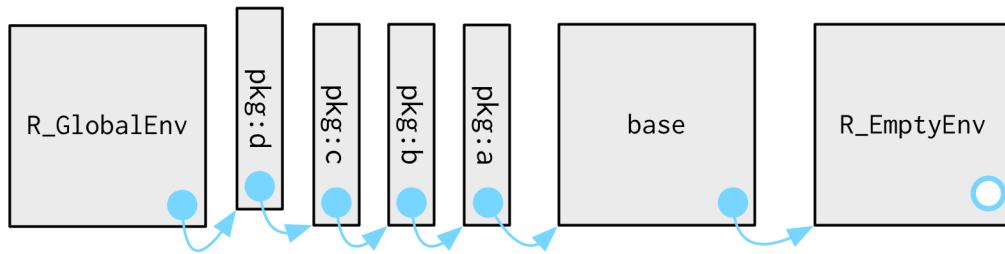
The last two environments on the search path are always the same:

- 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:



When you attach another package with `library()`, the parent environment of the global environment changes:



7.4.2 The function environment

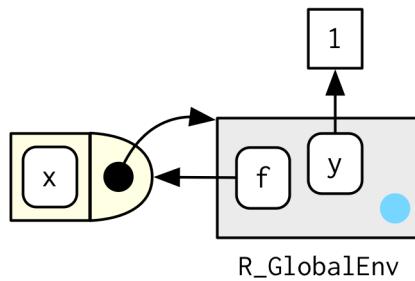
A function binds the current environment when it is created. This is called the **function environment**, and is used for lexical scoping. Across computer languages, functions that capture their environments are called **closures**, which is why this term is often used interchangeably with function in 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>
```

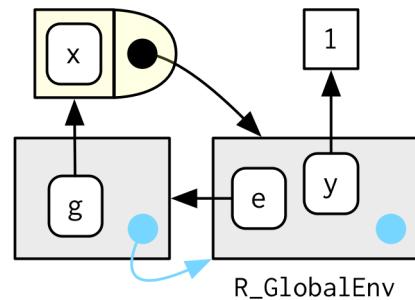
::: base Use `environment(f)` to access the environment of function `f`. :::

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 being 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
```



7.4.3 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.

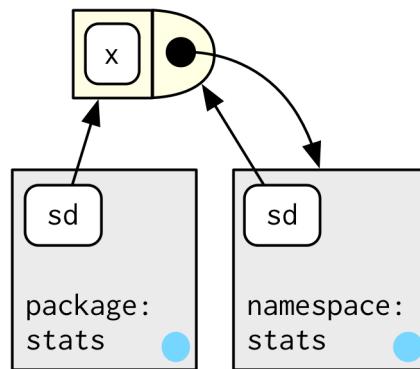
For example, take `sd()`:

```
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: 0x565358335f28>
#> <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.

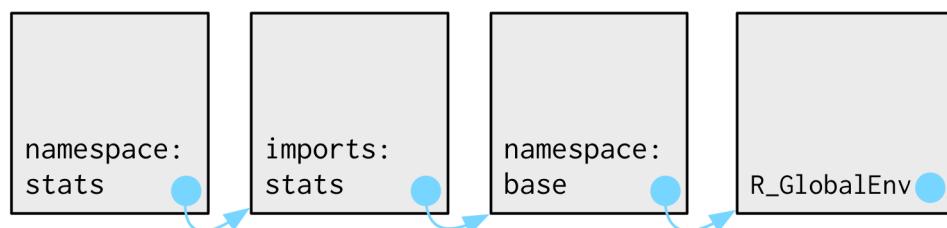
- The package environment is the external interface to the package. It's how you, the R user, find a function in an attached package or with `:::`. Its parent is determined by search path, i.e. the order in which packages have been attached.
- The namespace environment is the internal interface to the 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.

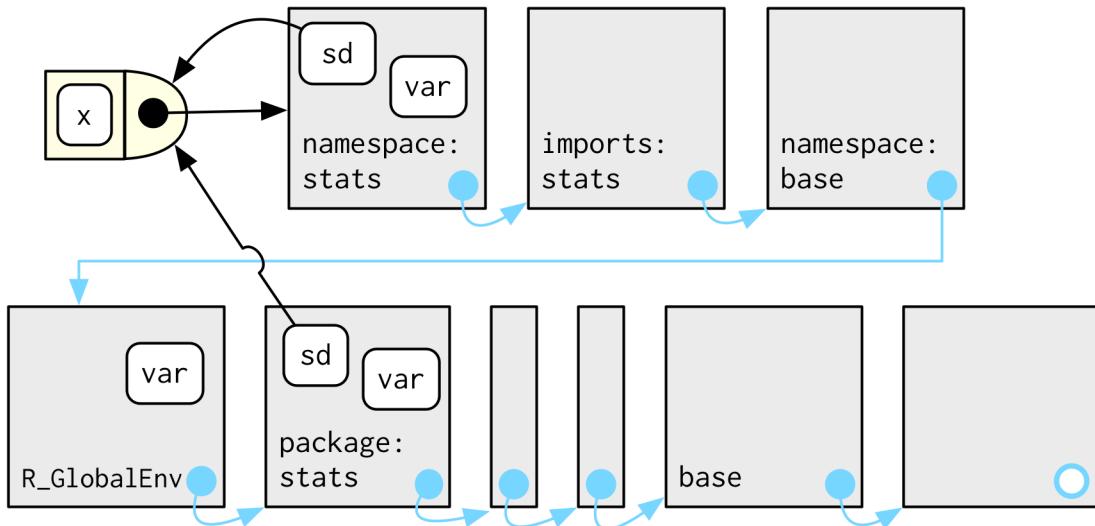


Every namespace environment has the same set of ancestors:

- 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 packge 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.

7.4.4 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
```

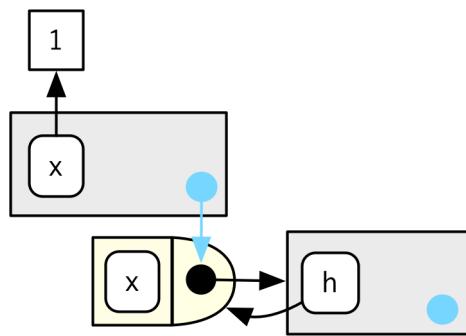
This function returns the same value every time because of the fresh start principle, described in a fresh start. Each time a function is called, a new environment is created to host execution. This is called the execution environment, and its parent is the function environment. Let's illustrate that process with a simpler function. I'll draw execution environments with an indirect parent; the parent environment is found via the function environment.

```

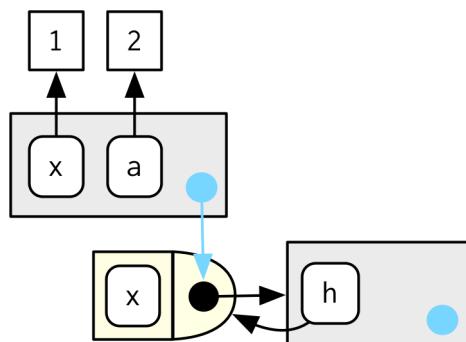
h <- function(x) {
  # 1.
  a <- 2 # 2.
  x + a
}
y <- h(1) # 3.

```

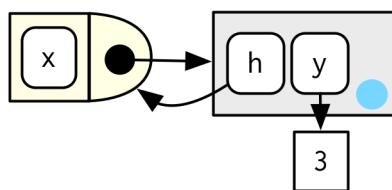
1. Function called with $x = 1$



2. a bound to value 2



**3. Function completes returning value 3.
Execution environment goes away.**



The 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: 0x5653587e9a90>
#>   parent: <env: 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()`.

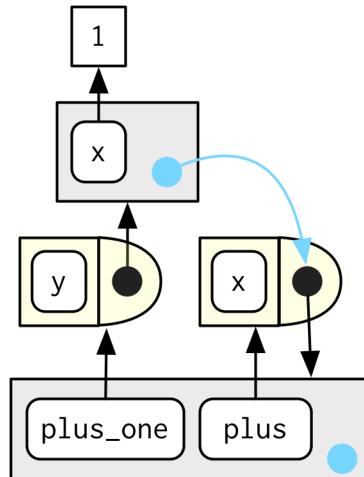
There's a lot going on in the diagram because the enclosing environment of `plus_one()` is the execution environment of `plus()`.

```

plus <- function(x) {
  function(y) x + y
}

plus_one <- plus(1)
plus_one
#> function(y) x + y
#> <environment: 0x5653583183c0>

```

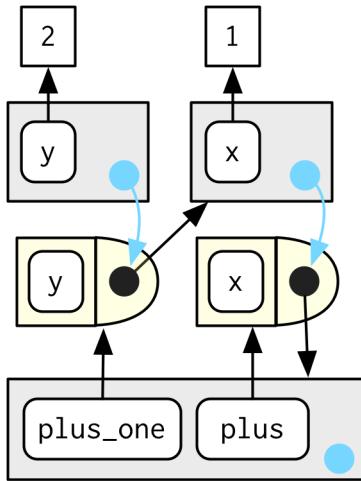


What happens when we call `plus_one()`? Its execution environment will have the captured execution environment of `plus()` as its parent:

```

plus_one(2)
#> [1] 3

```



You'll learn more about function factories in functional programming.

7.4.5 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.

7.5 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.

::: `base.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.

7.5.1 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()`).

7.5.2 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

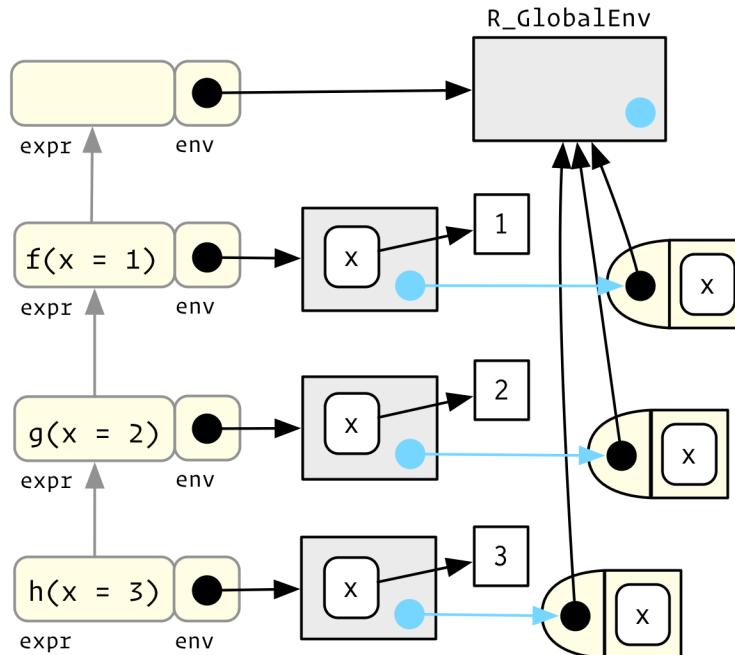
a(f())
#>
#>  a(f())
#>  b(x)
#>  c(x)
#>  f()
#>  g(x = 2)
#>  h(x = 3)
#>  lobstr::cst()
```

`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()`.

7.5.3 Frames

Each element of the call stack is a **frame**³, 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).



³NB: ?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.

7.5.4 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.) 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.

7.5.5 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()`.

7.6 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.

7.7 <<-

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.

7.7.1 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
```

7.8 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 8

Debugging

8.1 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.

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. How can you find out where an error occurred?
2. What does `browser()` do? List the five useful single-key commands that you can use inside of a `browser()` environment.

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.

8.2 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.

8.3 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 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.

8.3.1 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 a 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
Hide 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.

8.3.2 Browsing on error

The easiest way to enter the interactive debugger is through RStudio's "Rerun with Debug" tool. This re-runs 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 = browserOnce())

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.

8.3.3 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 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.

8.3.4 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.

8.3.5 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!")
g()
# Hi!
message2error(g())
# Error in message("Hi!"): Hi!
traceback()
# 10: stop(e) at #2
# 9: (function (e) stop(e))(list(message = "Hi!\n",
#       call = message("Hi!")))
# 8: signalCondition(cond)
```

```
# 7: doWithOneRestart(return(expr), restart)
# 6: withOneRestart(expr, restarts[[1L]])
# 5: withRestarts()
# 4: message("Hi!") at #1
# 3: g()
# 2: withCallingHandlers(code, message = function(e) stop(e))
#      at #2
# 1: message2error(g())
```

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.

8.4 Quiz answers

1. The most useful tool to determine where a error occurred is `traceback()`. Or use RStudio, which displays it automatically where an error occurs.
2. `browser()` pauses execution at the specified line and allows you to enter an interactive environment. In that environment, there are five useful commands: `n`, execute the next command; `s`, step into the next function; `f`, finish the current loop or function; `c`, continue execution normally; `Q`, stop the function and return to the console.

Chapter 9

Conditions

9.1 Introduction

Not all problems are unexpected. When writing a function, you can often anticipate potential problems (like a non-existent file or the wrong type of input). Communicating these problems to the user is the job of **conditions** such as errors (`stop()`), warnings (`warning()`), and messages (`message()`). Conditions are usually displayed prominently, in a bold font or coloured red depending on your R interface. You can tell them apart because errors always start with “Error” and warnings with “Warning message”.

Unexpected errors require interactive debugging to figure out what went wrong. Some errors, however, are expected, and you want to handle them automatically. In R, expected errors crop up most frequently when you’re fitting many models to different datasets, such as bootstrap replicates. Sometimes the model might fail to fit and throw an error, but you don’t want to stop everything. Instead, you want to fit as many models as possible and then perform diagnostics after the fact.

In R, there are three tools for handling conditions (including errors) programmatically:

- `try()` gives you the ability to continue execution even when an error occurs.
- `tryCatch()` lets you specify **handler** functions that control what happens when a condition is signalled.
- `withCallingHandlers()` is a variant of `tryCatch()` that establishes local handlers, whereas `tryCatch()` registers exiting handlers. Local handlers are called in the same context as where the condition is signalled, without interrupting the execution of the function. When an exiting handler from `tryCatch()` is called, the execution of the function is interrupted and the handler is called. `withCallingHandlers()` is rarely needed, but is useful to be aware of.

The following sections describe these tools in more detail.

Condition handling tools, like `withCallingHandlers()`, `tryCatch()`, and `try()` allow you as a user, to take specific actions when a condition occurs. For example, if you’re fitting many models, you might want to continue fitting the others even if one fails to converge. R offers an exceptionally powerful condition handling system based on ideas from Common Lisp, but it’s currently not very well documented or often used. This chapter will introduce you to the most important basics, but if you want to learn more, I recommend the following two sources:

- A prototype of a condition system for R 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 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>.

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 function do you use to ignore errors in block of code?
2. Why might you want to create an error with a custom S3 class?

9.1.1 Prerequisites

```
library(rlang)
#>
#> Attaching package: 'rlang'
#> The following objects are masked from 'package:purrr':
#>
#>     %>%, %/%, as_function, flatten, flatten_chr, flatten_dbl,
#>     flatten_int, flatten_lgl, invoke, list_along, modify, prepend,
#>     rep_along, splice
```

9.2 Signalling conditions

Collectively messages, warnings, and errors are known as conditions, and creating and sending them to the user is known as **signalling**. `stop()`, `warning()`, `message()`.

Also interrupts.

To help better understand conditions and the underlying object that defines their behaviour we will use `rlang::catch_cnd()`. This takes a block of code and returns the first condition signalled, or `NULL`.

```
# Captures error object
c <- catch_cnd(stop("An error"))
c
#> <simpleError in force(expr): An error>
str(c)
#> List of 2
#> $ message: chr "An error"
#> $ call    : language force(expr)
#> - attr(*, "class")= chr [1:3] "simpleError" "error" "condition"

# Captures first condition
c <- catch_cnd({
  warning("First")
  warning("Second")
})
c
#> <simpleWarning in force(expr): First>
```

```
# No condition, so returns NULL
catch_cnd(1 + 2)
#> NULL
```

9.2.1 Errors

Fatal errors are raised by `stop()` and force all execution to terminate. Errors are used when there is no way for a function to continue.

```
stop("This is an error message")
#> Error in eval(expr, envir, enclos): This is an error message
```

Style: <http://style.tidyverse.org/error-messages.html>

To learn more about the internal construction of the object, we need to capture it:

```
e <- catch_cnd(stop("Oops"))
str(e)
#> List of 2
#> $ message: chr "Oops"
#> $ call   : language force(expr)
#> - attr(*, "class")= chr [1:3] "simpleError" "error" "condition"
```

This shows us that the error object has class inherits from “condition”. And it has two components: the error message, and the call from which the error occurred.

The call is often not useful, so I think it's good practice to use `call. = FALSE`

```
stop("No call info", call. = FALSE)
#> Error: No call info

e <- catch_cnd(stop("Oops", call. = FALSE))
str(e)
#> List of 2
#> $ message: chr "Oops"
#> $ call   : NULL
#> - attr(*, "class")= chr [1:3] "simpleError" "error" "condition"
```

Something about rlang errors and capturing the traceback (when that actually works).

9.2.2 Warnings

Warnings are weaker than errors: they signal that something has gone wrong, but the code has been able to recover and continue. They are generated by `warning()`.

```
f <- function() {
  cat("1\n")
  warning("W1")
  cat("2\n")
  warning("W2")
  cat("3\n")
  warning("W3")
}
```

By defaults, warnings are cached and printed only when control returns to the top level.

```
f()
#> 1
#> 2
#> 3
#> Warning messages:
#> 1: In f() : W1
#> 2: In f() : W2
#> 3: In f() : W3
```

You can override this setting in two ways:

- To control someone else's warnings, set `options(warn = 1)`
- To control your own warnings, set `immediate. = TRUE`

Warning objects are very similar to error objects. They have `message` and `call`, and are inherit from the condition class.

```
e <- catch_cnd(warning("Oops"))
str(e)
#> List of 2
#> $ message: chr "Oops"
#> $ call   : language force(expr)
#> - attr(*, "class")= chr [1:3] "simpleWarning" "warning" "condition"
```

You should be cautious with your use of `warnings()`: 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 incorrect input. Reserve warnings for when you're almost sure that the result is correct, but there's something the user really should know. A good use of warnings is for deprecation: the code works, but will not work in the future, or generally a better method is available.

Base R tends to use warnings when only part of a vectorised input is invalid. However, I don't find these warnings terrifically informative: they don't tell you where the problem lies in the vector, and when embedded inside other code, it is challenging to figure the source of the warning. In fact, usually the best technique is to turn warnings into errors with `options(warn = 2)`. Then you can use your existing error diagnosis skills.

```
log(c(-1, 10, 100))
#> Warning in log(c(-1, 10, 100)): NaNs produced
#> [1] NaN 2.30 4.61

as.numeric(c("a", "1", "10"))
#> Warning: NAs introduced by coercion
#> [1] NA  1 10
```

9.2.3 Messages

Messages are generated by `message()` and are used to give informative output in a way that can easily be suppressed by the user (`?suppressMessages()`). I often use messages to let the user know what value the function has chosen for an important missing argument.

Messages are also important when developing packages. you need to print messages during startup, use `'packageStartupMessage()'`: that ensures `library(yourpackage, quietly = TRUE)` hides all your messages too.

9.2.4 Printed output

Function authors can also communicate with their users with `print()` or `cat()`, but I think that's a bad idea because it's hard to capture and selectively ignore this sort of output. Printed output is not a condition, so you can't use any of the useful condition handling tools you'll learn about below.

Generally, you should use `message()` rather than `cat()` or `print()` for informing the user about actions that your function has taken. This is useful, for example, if you've had to do non-trivial computation to determine the default value of an argument, and you want to let the user know exactly what you've done.

9.2.5 Interrupts

Interrupts can't be generated directly by the programmer, but are raised when the user attempts to terminate execution by pressing Ctrl + Break, Escape, or Ctrl + C (depending on the platform).

9.3 Ignoring conditions

Simplest way of handling conditions in R is to simply ignore them. These are the bluntest instruments, but can be convenient.

9.3.1 Ignoring errors

`try()` allows execution to continue even after an error has occurred. For example, 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 printed 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
```

You can suppress the message with `try(..., silent = TRUE)`.

You can also capture the output of the `try()` function. If successful, it will be the last result evaluated in the block (just like a function). If unsuccessful it will be an (invisible) object of class "try-error".

```
success <- try(1 + 2)
failure <- try("a" + "b")
class(success)
#> [1] "numeric"
```

```
class(failure)
#> [1] "try-error"
```

Generally, however, you should avoid switching between different behaviours based on the result of `try()`. Instead use `tryCatch()`, as described below. A useful `try()` pattern is to do assignment inside: 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)
```

9.3.2 Silencing messages and warnings

There are two functions that are sort of analogous to `try()` for `warnings()` and `messages()`: `suppressWarnings()` and `suppressMessages()`. These allow you to suppress all warnings and messages generated by a block of code.

```
suppressWarnings({
  warning("Uhoh!")
})

suppressMessages({
  message("Hello there")
})
```

Be aware that these functions are fairly heavy handed: you can't use them to suppress a single warning that you know about, while allowing other warnings that you don't know about to pass through.

The implementation of these functions are complex because they rely on the restart system. This is basically the only use of the restart system in base R (or pretty much any package) so we don't discuss here.

9.4 Condition handlers

`tryCatch()` and `withCallingHandlers()` are general tool for handling conditions. They allows you to map conditions to **handlers**, functions that are called with the condition as an input.

`tryCatch()` and `withCallingHandlers()` differ in the type of handlers they define;

- `tryCatch()` defines **exiting** handlers; after the condition is captured control returns to the context where `tryCatch()` was called. This makes it most suitable for working with errors, as errors have to exit the code anyway.
- `withCallingHandlers()` defines **in-place** handlers; after the condition is captured control returns to the context where the condition was signalled. This makes it most suitable for working with `warnings()`, `messages()`, and other conditions.

9.4.1 Exiting handlers

If a condition is signalled, `tryCatch()` will call the first handler whose name matches one of the classes of the condition. The names useful for built-in conditions are `error`, `warning`, `message`, `interrupt`, and the catch-all condition.

A handler function can do anything, but typically it will either return a value or create a more informative error message. For example, the `show_condition()` function below sets up handlers that return the type of condition signalled:

```

show_condition <- function(code) {
  tryCatch(
  {
    code
    NULL
  },
  error = function(c) "error",
  warning = function(c) "warning",
  message = function(c) "message"
)
}

show_condition(stop("!"))
#> [1] "error"
show_condition(warning("?!"))
#> [1] "warning"
show_condition(message("?"))
#> [1] "message"

# If no condition is captured, tryCatch returns NULL
show_condition(10)
#> NULL

```

9.4.2 In-place handlers

The primary difference from `tryCatch()` is execution continues normally when the handler returns. This includes the signalling function which continues its course after having called the handler (e.g., `stop()` will continue stopping the program and `message()` or `warning()` will continue signalling a message/warning).

```

message_handler <- function(c) cat("Caught a message!\n")

tryCatch(
  message = message_handler,
{
  message("Someone there?")
  message("Why, yes!")
}
)
#> Caught a message!

withCallingHandlers(
  message = message_handler,
{
  message("Someone there?")
  message("Why, yes!")
}
)
#> Caught a message!
#> Someone there?
#> Caught a message!
#> Why, yes!

```

`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 (e.g., deleting files, 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.

9.4.3 Differences

The handlers in `withCallingHandlers()` are called in the context of the call that generated the condition whereas the handlers in `tryCatch()` are called in the context of `tryCatch()`. We can see this most easily by using `calltrace()`

```
f <- function() g()
g <- function() h()
h <- function() stop("!")

tryCatch(f(), error = function(e) print(rlang::calltrace(globalenv())))
#> x
#> \-tryCatch(f(), error = function(e) print(rlang::calltrace(globalenv())))
#>   \-tryCatchList(expr, classes, parentenv, handlers)
#>     \-tryCatchOne(expr, names, parentenv, handlers[[1L]])
#>       \-value[[3L]](cond)
#>         \-print(rlang::calltrace(globalenv()))

withCallingHandlers(f(), error = function(e) print(rlang::calltrace(globalenv())))
#> x
#> +-withCallingHandlers(f(), error = function(e) print(rlang::calltrace(globalenv())))
#> +-f()
#> /  \-g()
#> /    \-h()
#> /      \-stop("!")
#> \-.handleSimpleError(...)
#>   \-h(simpleError(msg, call))
#>     \-print(rlang::calltrace(globalenv()))
#> Error in h(): !
```

Closely related is the return value of an inplace handler is effectively ignored, because control flow returns to the previous location. `withCallingHandlers()`:

```
f <- function() message("!")

tryCatch(f(), message = function(m) 1)
#> [1] 1

withCallingHandlers(f(), message = function(m) 1)
#> !
```

9.4.4 Exercises

1. Read the source code for `catch_cnd()` and explain how it works.
2. How could you rewrite `show_condition()` to use a single handler.
3. 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))
}

```

9.5 Use cases

What can you do with this tools? The following section exposes some come use cases.

9.5.1 Replacement value

You can use `tryCatch()` to implement `try()`. A simple implementation is shown below. `base::try()` is more complicated in order to make the error message look more like what you'd see if `tryCatch()` wasn't used. Note the use of `conditionMessage()` to extract the message associated with the original error.

```

fail_with <- function(expr, value = NULL) {
  tryCatch(expr, error = function(c) value)
}

try2 <- function(code, silent = FALSE) {
  tryCatch(code, error = function(c) {
    msg <- conditionMessage(c)
    if (!silent) {
      message(msg)
    }
    structure(msg, class = "try-error")
  })
}

try2(1)
#> [1] 1

try2(stop("Hi"))
#> Hi
#> [1] "Hi"
#> attr(,"class")
#> [1] "try-error"

try2(stop("Hi"), silent = TRUE)
#> [1] "Hi"
#> attr(,"class")
#> [1] "try-error"

```

9.5.2 Resignal

As well as returning default values when a condition is signalled, handlers can be used to make more informative error messages. For example, by modifying the message stored in the error condition object, the following function wraps `read.csv()` to add the file name to any errors:

```
read.csv2 <- function(file, ...) {
  tryCatch(read.csv(file, ...), error = function(c) {
    message <- paste0(c$message, " (in ", file, ")")
    abort(message)
  })
}
read.csv("code/dummy.csv")
#> Error in file(file, "rt"): cannot open the connection
read.csv2("code/dummy.csv")
#> Error: cannot open the connection (in code/dummy.csv)
```

Update to use whatever `rethrow()` becomes.

9.5.3 Record

This is what the `evaluate` package does. It powers `knitr`. (A little more complicated because it also has to handle output which uses a different system.)

9.5.4 Return early

```
try_parse_eval <- function(x, env = globalenv()) {
  expr <- tryCatch(parse(text = text), error = function(e) NULL)
  if (is.null(expr)) {
    return(NULL)
  }

  res <- tryCatch(eval(expr, env), error = function(e) NULL)
  if (is.null(res)) {
    return(res)
  }

  ...
}

try_parse_eval <- function(x, env = globalenv()) {
  expr <- tryCatch(parse(text = text), error = function(e) return_from(NULL))
  res <- tryCatch(eval(expr, env), error = function(e) return_from(NULL))
  ...
}
```

9.5.5 Muffle

Due to the way that restarts are implemented in R, the ability to muffle, or ignore a condition (so it doesn't bubble up to other handlers) is defined by the function that signals the condition. `message()` and `warning()` automatically setup muffle handlers, but `signalCondition()` does not.

`cnd_signal()` ensures that a muffler is always set up. `cnd_muffle(c)` always picks the right muffler depending on the class of the condition.

Log messages to disk example.

```

write_line <- function(path, ...) {
  cat(..., "\n", file = path, append = TRUE, sep = "")
}

log_messages <- function(expr, path) {

  withCallingHandlers(expr,
    message = function(c) {
      write_line(path, "[MESSAGE] ", conditionMessage(c))
      cnd_muffle(c)
    })
}

```

9.5.6 Exercises

1. Why is catching interrupts dangerous?

```

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.")
  }
}
```

9.6 Custom condition classes

One of the challenges of error handling in R is that most functions just call `stop()` with a string. That means if you want to figure out if a particular error occurred, you have to look at the text of the error message. This is error prone, not only because the text of the error might change over time, but also because many error messages are translated, so the message might be completely different to what you expect.

There are two reasons to create your own conditions:

- To make it easier to test your own code. Rather than relying on string matching on the text of the error, you can perform richer comparisons.
- To make it easier for the user to take different actions for different types of errors.

For example, “expected” errors (like a model failing to converge for some input datasets) can be silently ignored, while unexpected errors (like no disk space available) can be propagated to the user.

Base R doesn't make it easier to create your own classed conditions but the rlang equivalents provide some helpers.

```
abort(), warn(), inform().

abort <- function(.msg, .type = NULL, ...) {
  cnd <- error_cnd(.type = .type, ..., .msg = .msg)
  stop(cnd)
}

abort_bad_argument <- function(arg, must, not = NULL) {
  msg <- glue::glue(`{arg}` must {must})
  if (!is.null(not)) {
    msg <- glue::glue("{msg}; not {not}")
  }
  abort(msg, "error_bad_argument", arg = arg)
}

abort_bad_argument("x", must = "be numeric")
#> Error: `x` must be numeric
abort_bad_argument("x", must = "be numeric", not = "logical")
#> Error: `x` must be numeric; not logical

catch_cnd(abort_bad_argument("x", must = "be numeric"))$arg
#> [1] "x"
```

(Note that you can define a method for the `conditionMessage()` message generic instead of generating a message at creation time. This is usually of limited utility.)

```
my_log <- function(x, base = exp(1)) {
  if (!is.numeric(x)) {
    abort_bad_argument("x", must = "be numeric", not = typeof(x))
  }
  if (!is.numeric(base) && length(base) == 1) {
    abort_bad_argument("base", must = "be a single number")
  }

  log(x)
}

cnd <- catch_cnd(my_log("a"))
str(cnd)
#> List of 2
#> $ message:Classes 'glue', 'character' chr `x` must be numeric; not character"
#> $ arg : chr "x"
#> - attr(*, "class")= chr [1:3] "error_bad_argument" "error" "condition"
```

Note that when using `tryCatch()` with multiple handlers and custom classes, the first handler to match any class in the signal's class hierarchy is called, not the best match. For this reason, you need to make sure to put the most specific handlers first:

```
tryCatch(my_log("a"),
  error = function(c) "???", 
  error_bad_argument = function(c) "bad_argument"
)
#> [1] "???"
```

```
tryCatch(my_log("a"),
  error_bad_argument = function(c) "bad_argument",
  error = function(c) "???"
)
#> [1] "bad_argument"
```

9.7 Quiz answers

1. You could use `try()` or `tryCatch()`.
2. 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.

Part II

Functional programming

Chapter 10

Functional programming

10.1 Introduction

R, at its heart, is a functional programming (FP) language. This means that it provides many tools for the creation and manipulation of functions. In particular, R has what's known as first class functions. You can do anything with functions that you can do with vectors: 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.

The chapter starts by showing a motivating example, removing redundancy and duplication in code used to clean and summarise data. Then you'll learn about the three building blocks of functional programming: anonymous functions, closures (functions written by functions), and lists of functions. These pieces are twined together in the conclusion which shows how to build a suite of tools for numerical integration, starting from very simple primitives. This is a recurring theme in FP: start with small, easy-to-understand building blocks, combine them into more complex structures, and apply them with confidence.

The discussion of functional programming continues in the following two chapters: functionals explores functions that take functions as arguments and return vectors as output, and function operators explores functions that take functions as input and return them as output.

Outline

- Motivation motivates functional programming using a common problem: cleaning and summarising data before serious analysis.
- Anonymous functions shows you a side of functions that you might not have known about: you can use functions without giving them a name.
- Closures introduces the closure, a function written by another function. A closure can access its own arguments, and variables defined in its parent.
- Lists of functions shows how to put functions in a list, and explains why you might care.
- Numerical integration concludes the chapter with a case study that uses anonymous functions, closures and lists of functions to build a flexible toolkit for numerical integration.

Prequisites

You should be familiar with the basic rules of lexical scoping, as described in lexical scoping. Make sure you've installed the `pryr` package with `install.packages("pryr")`

10.2 Motivation

Imagine you've loaded a data file, like the one below, that uses `-99` to represent missing values. You want to replace all the `-99`s with NAs.

```
# Generate a sample dataset
set.seed(1014)
df <- data.frame(replicate(6, sample(c(1:10, -99), 6, rep = TRUE)))
names(df) <- letters[1:6]
df
#>     a   b   c   d   e   f
#> 1  1   6   1   5 -99  1
#> 2 10   4   4 -99   9   3
#> 3  7   9   5   4   1   4
#> 4  2   9   3   8   6   8
#> 5  1 10   5   9   8   6
#> 6  6   2   1   3   8   5
```

When you first started writing R code, you might have solved the problem with copy-and-paste:

```
df$a[df$a == -99] <- NA
df$b[df$b == -99] <- NA
df$c[df$c == -98] <- NA
df$d[df$d == -99] <- NA
df$e[df$e == -99] <- NA
df$f[df$g == -99] <- NA
```

One problem with copy-and-paste is that it's easy to make mistakes. Can you spot the two in the block above? These mistakes are inconsistencies that arose because we didn't have an authoritative description of the desired action (replace `-99` with NA). Duplicating an action makes bugs more likely and makes it harder to change code. For example, if the code for a missing value changes from `-99` to `9999`, you'd need to make the change in multiple places.

To prevent bugs and to make more flexible code, adopt the "do not repeat yourself", or DRY, principle. Popularised by the "pragmatic programmers", Dave Thomas and Andy Hunt, this principle states: "every piece of knowledge must have a single, unambiguous, authoritative representation within a system". FP tools are valuable because they provide tools to reduce duplication.

We can start applying FP ideas by writing a function that fixes the missing values in a single vector:

```
fix_missing <- function(x) {
  x[x == -99] <- NA
  x
}
df$a <- fix_missing(df$a)
df$b <- fix_missing(df$b)
df$c <- fix_missing(df$c)
df$d <- fix_missing(df$d)
df$e <- fix_missing(df$e)
df$g <- fix_missing(df$g)
```

This reduces the scope of possible mistakes, but it doesn't eliminate them: you can no longer accidentally type -98 instead of -99, but you can still mess up the name of variable. The next step is to remove this possible source of error by combining two functions. One function, `fix_missing()`, knows how to fix a single vector; the other, `lapply()`, knows how to do something to each column in a data frame.

`lapply()` takes three inputs: `x`, a list; `f`, a function; and `...`, other arguments to pass to `f()`. It applies the function to each element of the list and returns a new list. `lapply(x, f, ...)` is equivalent to the following for loop:

```
out <- vector("list", length(x))
for (i in seq_along(x)) {
  out[[i]] <- f(x[[i]], ...)
}
```

The real `lapply()` is rather more complicated since it's implemented in C for efficiency, but the essence of the algorithm is the same. `lapply()` is called a **functional**, because it takes a function as an argument. Functionals are an important part of functional programming. You'll learn more about them in functionals.

We can apply `lapply()` to this problem because data frames are lists. We just need a neat little trick to make sure we get back a data frame, not a list. Instead of assigning the results of `lapply()` to `df`, we'll assign them to `df[]`. R's usual rules ensure that we get a data frame, not a list. (If this comes as a surprise, you might want to read subsetting and assignment.) Putting these pieces together gives us:

```
fix_missing <- function(x) {
  x[x == -99] <- NA
  x
}
df[] <- lapply(df, fix_missing)
```

This code has five advantages over copy and paste:

- It's more compact.
- If the code for a missing value changes, it only needs to be updated in one place.
- It works for any number of columns. There is no way to accidentally miss a column.
- There is no way to accidentally treat one column differently than another.
- It is easy to generalise this technique to a subset of columns:

```
df[1:5] <- lapply(df[1:5], fix_missing)
```

The key idea is function composition. Take two simple functions, one which does something to every column and one which fixes missing values, and combines them to fix missing values in every column. Writing simple functions that can be understood in isolation and then composed is a powerful technique.

What if different columns used different codes for missing values? You might be tempted to copy-and-paste:

```
fix_missing_99 <- function(x) {
  x[x == -99] <- NA
  x
}
fix_missing_999 <- function(x) {
  x[x == -999] <- NA
  x
}
fix_missing_9999 <- function(x) {
  x[x == -9999] <- NA
  x
}
```

As before, it's easy to create bugs. Instead we could use closures, functions that make and return functions. Closures allow us to make functions based on a template:

```
missing_fixer <- function(na_value) {
  function(x) {
    x[x == na_value] <- NA
    x
  }
}
fix_missing_99 <- missing_fixer(-99)
fix_missing_999 <- missing_fixer(-999)

fix_missing_99(c(-99, -999))
#> [1] NA -999
fix_missing_999(c(-99, -999))
#> [1] -99 NA

#> NULL
```

In this case, you could argue that we should just add another argument:

```
fix_missing <- function(x, na_value) {
  x[x == na_value] <- NA
  x
}
```

That's a reasonable solution here, but it doesn't always work well in every situation. We'll see more compelling uses for closures in MLE.

```
#> NULL
```

Now consider a related problem. Once you've cleaned up your data, you might want to compute the same set of numerical summaries for each variable. You could write code like this:

```
mean(df$a)
median(df$a)
sd(df$a)
mad(df$a)
IQR(df$a)

mean(df$b)
median(df$b)
sd(df$b)
mad(df$b)
IQR(df$b)
```

But again, you'd be better off identifying and removing duplicate items. Take a minute or two to think about how you might tackle this problem before reading on.

One approach would be to write a summary function and then apply it to each column:

```
summary <- function(x) {
  c(mean(x), median(x), sd(x), mad(x), IQR(x))
}
lapply(df, summary)
```

That's a great start, but there's still some duplication. It's easier to see if we make the summary function more realistic:

```
summary <- function(x) {
  c(mean(x, na.rm = TRUE),
    median(x, na.rm = TRUE),
    sd(x, na.rm = TRUE),
    mad(x, na.rm = TRUE),
    IQR(x, na.rm = TRUE))
}
```

All five functions are called with the same arguments (`x` and `na.rm`) repeated five times. As always, duplication makes our code fragile: it's easier to introduce bugs and harder to adapt to changing requirements.

To remove this source of duplication, you can take advantage of another functional programming technique: storing functions in lists.

```
summary <- function(x) {
  funs <- c(mean, median, sd, mad, IQR)
  lapply(funs, function(f) f(x, na.rm = TRUE))
}
```

This chapter discusses these techniques in more detail. But before you can start learning them, you need to learn the simplest FP tool, the anonymous function.

10.3 Anonymous functions

In R, functions are objects in their own right. They aren't automatically bound to a name. Unlike many languages (e.g., C, C++, Python, and Ruby), R doesn't have a special syntax for creating a named function: when you create a function, you use the regular assignment operator to give it a name. If you choose not to give the function a name, you get an **anonymous function**.

You use an anonymous function when it's not worth the effort to give it a name:

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

Like all functions in R, anonymous functions have `formals()`, a `body()`, and a parent `environment()`:

```
formals(function(x = 4) g(x) + h(x))
#> $x
#> [1] 4
body(function(x = 4) g(x) + h(x))
#> g(x) + h(x)
environment(function(x = 4) g(x) + h(x))
#> <environment: R_GlobalEnv>
```

You can call an anonymous function without giving it a name, but the code is a little tricky to read because you must use parentheses in two different ways: first, to call a function, and second to make it clear that you want to call the anonymous function itself, as opposed to calling a (possibly invalid) function inside the anonymous function:

```
# This does not call the anonymous function.
# (Note that "3" is not a valid function.)
function(x) 3()
#> function(x) 3()

# With appropriate parenthesis, the function is called:
```

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

# So this anonymous function syntax
(function(x) x + 3)(10)
#> [1] 13

# behaves exactly the same as
f <- function(x) x + 3
f(10)
#> [1] 13
```

You can call anonymous functions with named arguments, but doing so is a good sign that your function needs a name.

One of the most common uses for anonymous functions is to create closures, functions made by other functions. Closures are described in the next section.

10.3.1 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?
- Use `lapply()` and an anonymous function to find the coefficient of variation (the standard deviation divided by the mean) for all columns in the `mtcars` dataset.
- Use `integrate()` and an anonymous function to find the area under the curve for the following functions. Use Wolfram Alpha to check your answers.
 - $y = x^2 - x$, $x \in [0, 10]$
 - $y = \sin(x) + \cos(x)$, $x \in [-\pi, \pi]$
 - $y = \exp(x) / x$, $x \in [10, 20]$
- 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?

10.4 Closures

"An object is data with functions. A closure is a function with data." — John D. Cook

One use of anonymous functions is to create small functions that are not worth naming. Another important use is to create closures, functions written by functions. Closures get their name because they **enclose** the environment of the parent function and can access all its variables. This is useful because it allows us to have two levels of parameters: a parent level that controls operation and a child level that does the work.

The following example uses this idea to generate a family of power functions in which a parent function (`power()`) creates two child functions (`square()` and `cube()`).

```
power <- function(exponent) {
  function(x) {
    x ^ exponent
  }
}
```

```
square <- power(2)
square(2)
#> [1] 4
square(4)
#> [1] 16

cube <- power(3)
cube(2)
#> [1] 8
cube(4)
#> [1] 64
```

When you print a closure, you don't see anything terribly useful:

```
square
#> function(x) {
#>   x ^ exponent
#> }
#> <environment: 0x562f452e8670>
cube
#> function(x) {
#>   x ^ exponent
#> }
#> <bytecode: 0x562f44b3bdc8>
#> <environment: 0x562f44cee438>
```

That's because the function itself doesn't change. The difference is the enclosing environment, `environment(square)`. One way to see the contents of the environment is to convert it to a list:

```
as.list(environment(square))
#> $exponent
#> [1] 2
as.list(environment(cube))
#> $exponent
#> [1] 3
```

Another way to see what's going on is to use `pryr::unenclose()`. This function replaces variables defined in the enclosing environment with their values:

```
library(pryr)
unenclose(square)
#> function (x)
#> {
#>   x^2
#> }
unenclose(cube)
#> function (x)
#> {
#>   x^3
#> }
```

The parent environment of a closure is the execution environment of the function that created it, as shown by this code:

```
power <- function(exponent) {
  print(environment())
  function(x) x ^ exponent
```

```

}
zero <- power(0)
#> <environment: 0x562f45c832c0>
environment(zero)
#> <environment: 0x562f45c832c0>
```

The execution environment normally disappears after the function returns a value. However, functions capture their enclosing environments. This means when function a returns function b, function b captures and stores the execution environment of function a, and it doesn't disappear. (This has important consequences for memory use, see [memory usage](#) for details.)

In R, almost every function is a closure. All functions remember the environment in which they were created, typically either the global environment, if it's a function that you've written, or a package environment, if it's a function that someone else has written. The only exception is primitive functions, which call C code directly and don't have an associated environment.

Closures are useful for making function factories, and are one way to manage mutable state in R.

10.4.1 Function factories

A function factory is a factory for making new functions. We've already seen two examples of function factories, `missing_fixer()` and `power()`. You call it with arguments that describe the desired actions, and it returns a function that will do the work for you. For `missing_fixer()` and `power()`, there's not much benefit in using a function factory instead of a single function with multiple arguments. Function factories are most useful when:

- The different levels are more complex, with multiple arguments and complicated bodies.
- Some work only needs to be done once, when the function is generated.

Function factories are particularly well suited to maximum likelihood problems, and you'll see a more compelling use of them in mathematical functionals.

10.4.2 Mutable state

Having variables at two levels allows you to maintain state across function invocations. This is possible because while the execution environment is refreshed every time, the enclosing environment is constant. The key to managing variables at different levels is the double arrow assignment operator (`<<-`). Unlike the usual single arrow assignment (`<-`) that always assigns in the current environment, the double arrow operator will keep looking up the chain of parent environments until it finds a matching name. (Binding names to values has more details on how it works.)

Together, a static parent environment and `<<-` make it possible to maintain state across function calls. The following example shows a counter that records how many times a function has been called. Each time `new_counter` is run, it creates an environment, initialises the counter `i` in this environment, and then creates a new function.

```

new_counter <- function() {
  i <- 0
  function() {
    i <<- i + 1
    i
  }
}
```

The new function is a closure, and its enclosing environment is the environment created when `new_counter()` is run. Ordinarily, function execution environments are temporary, but a closure maintains access to the environment in which it was created. In the example below, closures `counter_one()` and `counter_two()` each get their own enclosing environments when run, so they can maintain different counts.

```
counter_one <- new_counter()
counter_two <- new_counter()

counter_one()
#> [1] 1
counter_one()
#> [1] 2
counter_two()
#> [1] 1
```

The counters get around the “fresh start” limitation by not modifying variables in their local environment. Since the changes are made in the unchanging parent (or enclosing) environment, they are preserved across function calls.

What happens if you don’t use a closure? What happens if you use `<-` instead of `<<-`? Make predictions about what will happen if you replace `new_counter()` with the variants below, then run the code and check your predictions.

```
i <- 0
new_counter2 <- function() {
  i <<- i + 1
  i
}
new_counter3 <- function() {
  i <- 0
  function() {
    i <- i + 1
    i
  }
}
```

Modifying values in a parent environment is an important technique because it is one way to generate “mutable state” in R. Mutable state is normally hard because every time it looks like you’re modifying an object, you’re actually creating and then modifying a copy. However, if you do need mutable objects and your code is not very simple, it’s usually better to use reference classes, as described in RC.

The power of closures is tightly coupled with the more advanced ideas in functionals and function operators. You’ll see many more closures in those two chapters. The following section discusses the third technique of functional programming in R: the ability to store functions in a list.

10.4.3 Exercises

1. Why are functions created by other functions called closures?
2. What does the following statistical function do? What would be a better name for it? (The existing name is a bit of a hint.)

```
bc <- function(lambda) {
  if (lambda == 0) {
    function(x) log(x)
  } else {
```

```

    function(x) (x ^ lambda - 1) / lambda
  }
}

```

3. What does approxfun() do? What does it return?
4. What does ecdf() do? What does it return?
5. Create a function that creates functions that compute the ith 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))

```

6. 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.

```

lapply(mtcars, pick(5))
# should do the same as this
lapply(mtcars, function(x) x[[5]])

```

10.5 Lists of functions

In R, functions can be stored in lists. This makes it easier to work with groups of related functions, in the same way a data frame makes it easier to work with groups of related vectors.

We'll start with a simple benchmarking example. Imagine you are comparing the performance of multiple ways of computing the arithmetic mean. You could do this by storing each approach (function) in a list:

```

compute_mean <- list(
  base = function(x) mean(x),
  sum = function(x) sum(x) / length(x),
  manual = function(x) {
    total <- 0
    n <- length(x)
    for (i in seq_along(x)) {
      total <- total + x[i] / n
    }
    total
  }
)

```

Calling a function from a list is straightforward. You extract it then call it:

```

x <- runif(1e5)
system.time(compute_mean$base(x))
#>   user  system elapsed
#> 0.000 0.000 0.001
system.time(compute_mean[[2]](x))
#>   user  system elapsed
#> 0 0 0
system.time(compute_mean[["manual"]](x))

```

```
#>    user  system elapsed
#> 0.010  0.000  0.009
```

To call each function (e.g., to check that they all return the same results), use `lapply()`. We'll need either an anonymous function or a new named function, since there isn't a built-in function to handle this situation.

```
lapply(compute_mean, function(f) f(x))
#> $base
#> [1] 0.499
#>
#> $sum
#> [1] 0.499
#>
#> $manual
#> [1] 0.499

call_fun <- function(f, ...) f(...)
lapply(compute_mean, call_fun, x)
#> $base
#> [1] 0.499
#>
#> $sum
#> [1] 0.499
#>
#> $manual
#> [1] 0.499
```

To time each function, we can combine `lapply()` and `system.time()`:

```
lapply(compute_mean, function(f) system.time(f(x)))
#> $base
#>    user  system elapsed
#> 0.000  0.000  0.001
#>
#> $sum
#>    user  system elapsed
#>      0      0      0
#>
#> $manual
#>    user  system elapsed
#> 0.000  0.000  0.006
```

Another use for a list of functions is to summarise an object in multiple ways. To do that, we could store each summary function in a list, and then run them all with `lapply()`:

```
x <- 1:10
funs <- list(
  sum = sum,
  mean = mean,
  median = median
)
lapply(funs, function(f) f(x))
#> $sum
#> [1] 55
#>
#> $mean
```

```
#> [1] 5.5
#>
#> $median
#> [1] 5.5
```

What if we wanted our summary functions to automatically remove missing values? One approach would be to make a list of anonymous functions that call our summary functions with the appropriate arguments:

```
fun2 <- list(
  sum = function(x, ...) sum(x, ..., na.rm = TRUE),
  mean = function(x, ...) mean(x, ..., na.rm = TRUE),
  median = function(x, ...) median(x, ..., na.rm = TRUE)
)
lapply(fun2, function(f) f(x))
#> $sum
#> [1] 55
#>
#> $mean
#> [1] 5.5
#>
#> $median
#> [1] 5.5
```

This, however, leads to a lot of duplication. Apart from a different function name, each function is almost identical. A better approach would be to modify our `lapply()` call to include the extra argument:

```
lapply(fun2, function(f) f(x, na.rm = TRUE))
```

10.5.1 Moving lists of functions to the global environment

From time to time you may create a list of functions that you want to be available without having to use a special syntax. For example, imagine you want to create HTML code by mapping each tag to an R function. The following example uses a function factory to create functions for the tags `<p>` (paragraph), `` (bold), and `<i>` (italics).

```
simple_tag <- function(tag) {
  force(tag)
  function(...) {
    paste0("<", tag, ">", paste0(...), "</", tag, ">")
  }
}
tags <- c("p", "b", "i")
html <- lapply(setNames(tags, tags), simple_tag)
```

I've put the functions in a list because I don't want them to be available all the time. The risk of a conflict between an existing R function and an HTML tag is high. But keeping them in a list makes code more verbose:

```
html$p("This is ", html$b("bold"), " text.")
#> [1] "<p>This is <b>bold</b> text.</p>"
```

Depending on how long we want the effect to last, you have three options to eliminate the use of `html$`:

- For a very temporary effect, you can use `with()`:

```
with(html, p("This is ", b("bold"), " text."))
#> [1] "<p>This is <b>bold</b> text.</p>"
```

- For a longer effect, you can attach() the functions to the search path, then detach() when you're done:

```
attach(html)
p("This is ", b("bold"), " text.")
#> [1] "<p>This is <b>bold</b> text.</p>"
detach(html)
```

- Finally, you could copy the functions to the global environment with list2env(). You can undo this by deleting the functions after you're done.

```
list2env(html, environment())
#> <environment: R_GlobalEnv>
p("This is ", b("bold"), " text.")
#> [1] "<p>This is <b>bold</b> text.</p>"
rm(list = names(html), envir = environment())
```

I recommend the first option, using with(), because it makes it very clear when code is being executed in a special context and what that context is.

10.5.2 Exercises

1. Implement a summary function that works like base::summary(), but uses a list of functions. Modify the function so it returns a closure, making it possible to use it as a function factory.
2. Which of the following commands is equivalent to with(x, f(z))?
 - (a) x\$f(x\$z).
 - (b) f(x\$z).
 - (c) x\$f(z).
 - (d) f(z).
 - (e) It depends.

10.6 Case study: numerical integration

To conclude this chapter, I'll develop a simple numerical integration tool using first-class functions. Each step in the development of the tool is driven by a desire to reduce duplication and to make the approach more general.

The idea behind numerical integration is simple: find the area under a curve by approximating the curve with simpler components. The two simplest approaches are the **midpoint** and **trapezoid** rules. The midpoint rule approximates a curve with a rectangle. The trapezoid rule uses a trapezoid. Each takes the function we want to integrate, f , and a range of values, from a to b , to integrate over. For this example, I'll try to integrate $\sin x$ from 0 to π . This is a good choice for testing because it has a simple answer: 2.

```
midpoint <- function(f, a, b) {
  (b - a) * f((a + b) / 2)
}

trapezoid <- function(f, a, b) {
  (b - a) / 2 * (f(a) + f(b))
```

```

}

midpoint(sin, 0, pi)
#> [1] 3.14
trapezoid(sin, 0, pi)
#> [1] 1.92e-16

```

Neither of these functions gives a very good approximation. To make them more accurate using the idea that underlies calculus: we'll break up the range into smaller pieces and integrate each piece using one of the simple rules. This is called **composite integration**. I'll implement it using two new functions:

```

midpoint_composite <- function(f, a, b, n = 10) {
  points <- seq(a, b, length = n + 1)
  h <- (b - a) / n

  area <- 0
  for (i in seq_len(n)) {
    area <- area + h * f((points[i] + points[i + 1]) / 2)
  }
  area
}

trapezoid_composite <- function(f, a, b, n = 10) {
  points <- seq(a, b, length = n + 1)
  h <- (b - a) / n

  area <- 0
  for (i in seq_len(n)) {
    area <- area + h / 2 * (f(points[i]) + f(points[i + 1]))
  }
  area
}

midpoint_composite(sin, 0, pi, n = 10)
#> [1] 2.01
midpoint_composite(sin, 0, pi, n = 100)
#> [1] 2
trapezoid_composite(sin, 0, pi, n = 10)
#> [1] 1.98
trapezoid_composite(sin, 0, pi, n = 100)
#> [1] 2

```

You'll notice that there's a lot of duplication between `midpoint_composite()` and `trapezoid_composite()`. Apart from the internal rule used to integrate over a range, they are basically the same. From these specific functions you can extract a more general composite integration function:

```

composite <- function(f, a, b, n = 10, rule) {
  points <- seq(a, b, length = n + 1)

  area <- 0
  for (i in seq_len(n)) {
    area <- area + rule(f, points[i], points[i + 1])
  }

  area
}

```

```

}

composite(sin, 0, pi, n = 10, rule = midpoint)
#> [1] 2.01
composite(sin, 0, pi, n = 10, rule = trapezoid)
#> [1] 1.98

```

This function takes two functions as arguments: the function to integrate and the integration rule. We can now add even better rules for integrating over smaller ranges:

```

simpson <- function(f, a, b) {
  (b - a) / 6 * (f(a) + 4 * f((a + b) / 2) + f(b))
}

boole <- function(f, a, b) {
  pos <- function(i) a + i * (b - a) / 4
  fi <- function(i) f(pos(i))

  (b - a) / 90 *
    (7 * fi(0) + 32 * fi(1) + 12 * fi(2) + 32 * fi(3) + 7 * fi(4))
}

composite(sin, 0, pi, n = 10, rule = simpson)
#> [1] 2
composite(sin, 0, pi, n = 10, rule = boole)
#> [1] 2

```

It turns out that the midpoint, trapezoid, Simpson, and Boole rules are all examples of a more general family called Newton-Cotes rules. (They are polynomials of increasing complexity.) We can use this common structure to write a function that can generate any general Newton-Cotes rule:

```

newton_cotes <- function(coef, open = FALSE) {
  n <- length(coef) + open

  function(f, a, b) {
    pos <- function(i) a + i * (b - a) / n
    points <- pos(seq.int(0, length(coef) - 1))

    (b - a) / sum(coef) * sum(f(points) * coef)
  }
}

boole <- newton_cotes(c(7, 32, 12, 32, 7))
milne <- newton_cotes(c(2, -1, 2), open = TRUE)
composite(sin, 0, pi, n = 10, rule = milne)
#> [1] 1.99

```

Mathematically, the next step in improving numerical integration is to move from a grid of evenly spaced points to a grid where the points are closer together near the end of the range, such as Gaussian quadrature. That's beyond the scope of this case study, but you could implement it with similar techniques.

10.6.1 Exercises

1. Instead of creating individual functions (e.g., `midpoint()`, `trapezoid()`, `simpson()`, etc.), we could store them in a list. If we did that, how would that change the code? Can you create the list of functions from a list of coefficients for the Newton-Cotes formulae?
2. The trade-off between integration rules is that more complex rules are slower to compute, but need fewer pieces. For `sin()` in the range $[0, \pi]$, determine the number of pieces needed so that each rule will be equally accurate. Illustrate your results with a graph. How do they change for different functions? $\sin(1 / x^2)$ is particularly challenging.

Chapter 11

Functionals

11.1 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 higher-order function is a function that takes a function as an input or returns a function as output. We’ve already seen one type of higher order function: closures, functions returned by another function. The complement to a closure is a **functional**, 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: the three most frequently used are `lapply()`, `apply()`, and `tapply()`. All three 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 modification in place 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 know immediately 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 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 improving the speed of your code.

Outline

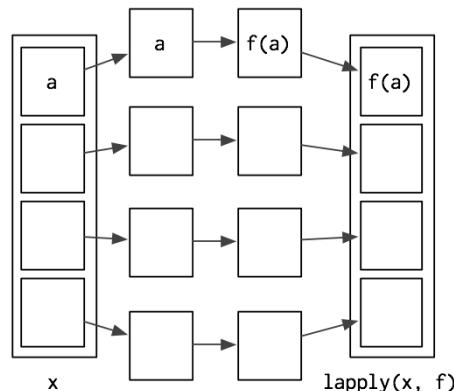
- My first functional: `lapply()` introduces your first functional: `lapply()`.
- For loop functionals shows you variants of `lapply()` that produce different outputs, take different inputs, and distribute computation in different ways.
- Data structure functionals discusses functionals that work with more complex data structures like matrices and arrays.
- Functional programming teaches you about the powerful `Reduce()` and `Filter()` functions which are useful for working with lists.
- Mathematical functionals discusses functionals that you might be familiar with from mathematics, like root finding, integration, and optimisation.
- Loops that shouldn't be converted to functions provides some important caveats about when you shouldn't attempt to convert a loop into a functional.
- A family of functions finishes off the chapter by showing you how functionals can take a simple building block and use it to create a set of powerful and consistent tools.

Prerequisites

You'll use closures frequently used in conjunction with functionals. If you need a refresher, review closures.

11.2 My first functional: `lapply()`

The simplest functional is `lapply()`, which you may already be familiar with. `lapply()` takes a function, applies it to each element in a list, and returns the results in the form of a list. `lapply()` is the building block for many other functionals, so it's important to understand how it works. Here's a pictorial representation:



`lapply()` is written in C for performance, but we can create a simple R implementation that does the same thing:

```
lapply2 <- function(x, f, ...) {
  out <- vector("list", length(x))
  for (i in seq_along(x)) {
    out[[i]] <- f(x[[i]], ...)
  }
  out
}
```

From this code, you can see that `lapply()` is a wrapper for a common for loop pattern: create a container for output, apply `f()` to each component of a list, and fill the container with the results. All other for loop functionals are variations on this theme: they simply use different types of input or output.

`lapply()` makes it easier to work with lists by eliminating much of the boilerplate associated with looping. This allows you to focus on the function that you're applying:

```
# Create some random data
l <- replicate(20, runif(sample(1:10, 1)), simplify = FALSE)

# With a for loop
out <- vector("list", length(l))
for (i in seq_along(l)) {
  out[[i]] <- length(l[[i]])
}
unlist(out)
#> [1] 3 1 1 2 2 10 5 9 7 2 4 10 8 2 9 7 3 2 2 8

# With lapply
unlist(lapply(l, length))
#> [1] 3 1 1 2 2 10 5 9 7 2 4 10 8 2 9 7 3 2 2 8
```

(I'm using `unlist()` to convert the output from a list to a vector to make it more compact. We'll see other ways of making the output a vector shortly.)

Since data frames are also lists, `lapply()` is also useful when you want to do something to each column of a data frame:

```
# What class is each column?
unlist(lapply(mtcars, class))
#>      mpg      cyl      disp       hp      drat       wt      qsec
#> "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
#>      vs       am      gear      carb
#> "numeric" "numeric" "numeric" "numeric"

# Divide each column by the mean
mtcars[] <- lapply(mtcars, function(x) x / mean(x))
```

The pieces of `x` are always supplied as the first argument to `f`. If you want to vary a different argument, you can use an anonymous function. The following example varies the amount of trimming applied when computing the mean of a fixed `x`.

```
trims <- c(0, 0.1, 0.2, 0.5)
x <- rcauchy(1000)
unlist(lapply(trims, function(trim) mean(x, trim = trim)))
#> [1] 0.2879 0.0790 0.0535 0.0502
```

11.2.1 Looping patterns

It's useful to remember that there are three basic ways to loop over a vector:

1. loop over the elements: `for (x in xs)`
2. loop over the numeric indices: `for (i in seq_along(xs))`
3. loop over the names: `for (nm in names(xs))`

The first form is usually not a good choice for a for loop because it leads to inefficient ways of saving output. With this form it's very natural to save the output by extending a data structure, like in this example:

```
xs <- runif(1e3)
res <- c()
for (x in xs) {
  # This is slow!
  res <- c(res, sqrt(x))
}
```

This is slow because each time you extend the vector, R has to copy all of the existing elements. Avoid copies discusses this problem in more depth. Instead, it's much better to create the space you'll need for the output and then fill it in. This is easiest with the second form:

```
res <- numeric(length(xs))
for (i in seq_along(xs)) {
  res[i] <- sqrt(xs[i])
}
```

Just as there are three basic ways to use a for loop, there are three basic ways to use lapply():

```
lapply(xs, function(x) {})
lapply(seq_along(xs), function(i) {})
lapply(names(xs), function(nm) {})
```

Typically you'd use the first form because lapply() takes care of saving the output for you. However, if you need to know the position or name of the element you're working with, you should use the second or third form. Both give you an element's position (`i`, `nm`) and value (`xs[[i]]`, `xs[[nm]]`). If you're struggling to solve a problem using one form, you might find it easier with another.

11.2.2 Exercises

1. Why are the following two invocations of lapply() equivalent?

```
trims <- c(0, 0.1, 0.2, 0.5)
x <- rcauchy(100)

lapply(trims, function(trim) mean(x, trim = trim))
lapply(trims, mean, x = x)
```

2. 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])
}
```

3. Use both for loops and lapply() 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
)
```

4. Fit the model `mpg ~ disp` to each of the bootstrap replicates of `mtcars` in the list below by using a for loop and `lapply()`. Can you do it without an anonymous function?

```
bootstraps <- lapply(1:10, function(i) {
  rows <- sample(1:nrow(mtcars), rep = TRUE)
  mtcars[rows, ]
})
```

5. For each model in the previous two exercises, extract R^2 using the function below.

```
rsq <- function(mod) summary(mod)$r.squared
```

11.3 For loop functionals: friends of lapply()

The key to using functionals in place of for loops is recognising that common looping patterns are already implemented in existing base functionals. Once you've mastered these existing functionals, the next step is to start writing your own: if you discover you're duplicating the same looping pattern in many places, you should extract it out into its own function.

The following sections build on `lapply()` and discuss:

- `sapply()` and `vapply()`, variants of `lapply()` that produce vectors, matrices, and arrays as **output**, instead of lists.
- `Map()` and `mapply()` which iterate over multiple **input** data structures in parallel.
- `mclapply()` and `mcMap()`, parallel versions of `lapply()` and `Map()`.
- Writing a new function, `rollapply()`, to solve a new problem.

11.3.1 Vector output: sapply and vapply

`sapply()` and `vapply()` are very similar to `lapply()` except they simplify their output to produce an atomic vector. While `sapply()` guesses, `vapply()` takes an additional argument specifying the output type. `sapply()` is great for interactive use because it saves typing, but if you use it inside your functions you'll get weird errors if you supply the wrong type of input. `vapply()` is more verbose, but gives more informative error messages and never fails silently. It is better suited for use inside other functions.

The following example illustrates these differences. When given a data frame, `sapply()` and `vapply()` return the same results. When given an empty list, `sapply()` returns another empty list instead of the more correct zero-length logical vector.

```
sapply(mtcars, is.numeric)
#> mpg cyl disp hp drat wt qsec vs am gear carb
#> TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
vapply(mtcars, is.numeric, logical(1))
#> mpg cyl disp hp drat wt qsec vs am gear carb
#> TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
sapply(list(), is.numeric)
#> list()
vapply(list(), is.numeric, logical(1))
#> logical(0)
```

If the function returns results of different types or lengths, `sapply()` will silently return a list, while `vapply()` will throw an error. `sapply()` is fine for interactive use because you'll normally notice if something goes wrong, but it's dangerous when writing functions.

The following example illustrates a possible problem when extracting the class of columns in a data frame: if you falsely assume that class only has one value and use `sapply()`, you won't find out about the problem until some future function is given a list instead of a character vector.

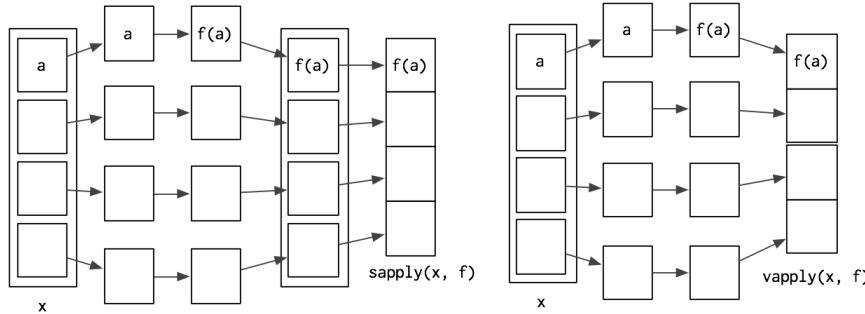
```
df <- data.frame(x = 1:10, y = letters[1:10])
sapply(df, class)
#>      x          y
#> "integer"  "factor"
vapply(df, class, character(1))
#>      x          y
#> "integer"  "factor"

df2 <- data.frame(x = 1:10, y = Sys.time() + 1:10)
sapply(df2, class)
#> $x
#> [1] "integer"
#>
#> $y
#> [1] "POSIXct" "POSIXt"
vapply(df2, class, character(1))
#> Error in vapply(df2, class, character(1)): values must be length 1,
#> but FUN(X[[2]]) result is length 2
```

`sapply()` is a thin wrapper around `lapply()` that transforms a list into a vector in the final step. `vapply()` is an implementation of `lapply()` that assigns results to a vector (or matrix) of appropriate type instead of as a list. The following code shows a pure R implementation of the essence of `sapply()` and `vapply()` (the real functions have better error handling and preserve names, among other things).

```
sapply2 <- function(x, f, ...) {
  res <- lapply2(x, f, ...)
  simplify2array(res)
}

vapply2 <- function(x, f, f.value, ...) {
  out <- matrix(rep(f.value, length(x)), nrow = length(f.value))
  for (i in seq_along(x)) {
    res <- f(x[[i]], ...)
    stopifnot(
      length(res) == length(f.value),
      typeof(res) == typeof(f.value)
    )
    out[, i] <- res
  }
  out
}
```



vapply() and sapply() have different outputs from lapply(). The following section discusses Map(), which has different inputs.

11.3.2 Multiple inputs: Map (and mapply)

With lapply(), only one argument to the function varies; the others are fixed. This makes it poorly suited for some problems. For example, how would you find a weighted mean when you have two lists, one of observations and the other of weights?

```
# Generate some sample data
xs <- replicate(5, runif(10), simplify = FALSE)
ws <- replicate(5, rpois(10, 5) + 1, simplify = FALSE)
```

It's easy to use lapply() to compute the unweighted means:

```
unlist(lapply(xs, mean))
#> [1] 0.678 0.445 0.427 0.469 0.560
```

But how could we supply the weights to weighted.mean()? lapply(x, means, w) won't work because the additional arguments to lapply() are passed to every call. We could change looping forms:

```
unlist(lapply(seq_along(xs), function(i) {
  weighted.mean(xs[[i]], ws[[i]])
}))
#> [1] 0.695 0.464 0.403 0.501 0.521
```

This works, but it's a little clumsy. A cleaner alternative is to use Map, a variant of lapply(), where all arguments can vary. This lets us write:

```
unlist(Map(weighted.mean, xs, ws))
#> [1] 0.695 0.464 0.403 0.501 0.521
```

Note that the order of arguments is a little different: function is the first argument for Map() and the second for lapply().

This is equivalent to:

```
stopifnot(length(xs) == length(ws))
out <- vector("list", length(xs))
for (i in seq_along(xs)) {
  out[[i]] <- weighted.mean(xs[[i]], ws[[i]])
}
```

There's a natural equivalence between Map() and lapply() because you can always convert a Map() to an lapply() that iterates over indices. But using Map() is more concise, and more clearly indicates what you're trying to do.

Map is useful whenever you have two (or more) lists (or data frames) that you need to process in parallel. For example, another way of standardising columns is to first compute the means and then divide by them. We could do this with `lapply()`, but if we do it in two steps, we can more easily check the results at each step, which is particularly important if the first step is more complicated.

```
mtmeans <- lapply(mtcars, mean)
mtmeans[] <- Map(`/`, mtcars, mtmeans)

# In this case, equivalent to
mtcars[] <- lapply(mtcars, function(x) x / mean(x))
```

If some of the arguments should be fixed and constant, use an anonymous function:

```
Map(function(x, w) weighted.mean(x, w, na.rm = TRUE), xs, ws)
```

We'll see a more compact way to express the same idea in the next chapter.

```
#> NULL
```

You may be more familiar with `mapply()` than `Map()`. I prefer `Map()` because:

- It's equivalent to `mapply` with `simplify = FALSE`, which is almost always what you want.
- Instead of using an anonymous function to provide constant inputs, `mapply` has the `MoreArgs` argument that takes a list of extra arguments that will be supplied, as is, to each call. This breaks R's usual lazy evaluation semantics, and is inconsistent with other functions.

In brief, `mapply()` adds more complication for little gain.

```
#> NULL
```

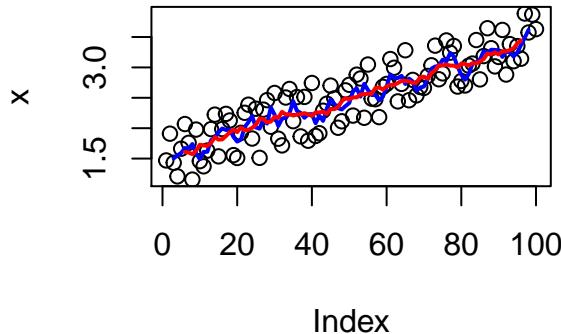
11.3.3 Rolling computations

What if you need a for loop replacement that doesn't exist in base R? You can often create your own by recognising common looping structures and implementing your own wrapper. For example, you might be interested in smoothing your data using a rolling (or running) mean function:

```
rollmean <- function(x, n) {
  out <- rep(NA, length(x))

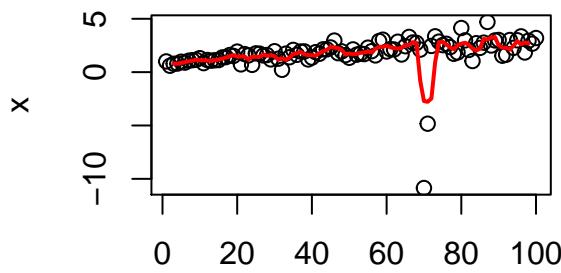
  offset <- trunc(n / 2)
  for (i in (offset + 1):(length(x) - n + offset + 1)) {
    out[i] <- mean(x[(i - offset):(i + offset - 1)])
  }
  out
}

x <- seq(1, 3, length = 1e2) + runif(1e2)
plot(x)
lines(rollmean(x, 5), col = "blue", lwd = 2)
lines(rollmean(x, 10), col = "red", lwd = 2)
```



But if the noise was more variable (i.e., it has a longer tail), you might worry that your rolling mean was too sensitive to outliers. Instead, you might want to compute a rolling median.

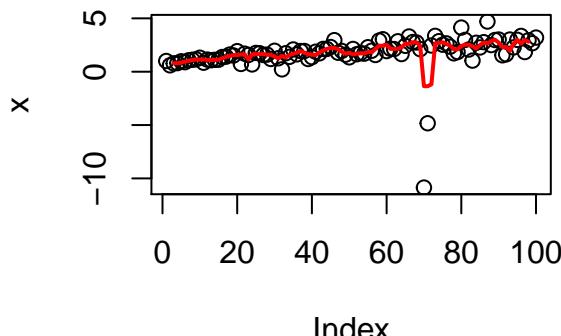
```
x <- seq(1, 3, length = 1e2) + rt(1e2, df = 2) / 3
plot(x)
lines(rollmean(x, 5), col = "red", lwd = 2)
```



To change `rollmean()` to `rollmedian()`, all you need to do is replace `mean` with `median` inside the loop. But instead of copying and pasting to create a new function, we could extract the idea of computing a rolling summary into its own function:

```
rollapply <- function(x, n, f, ...) {
  out <- rep(NA, length(x))

  offset <- trunc(n / 2)
  for (i in (offset + 1):(length(x) - n + offset + 1)) {
    out[i] <- f(x[(i - offset):(i + offset - 1)], ...)
  }
  out
}
plot(x)
lines(rollapply(x, 5, median), col = "red", lwd = 2)
```



You might notice that the internal loop looks pretty similar to a `vapply()` loop, so we could rewrite the function as:

```
rollapply <- function(x, n, f, ...) {
  offset <- trunc(n / 2)
  locs <- (offset + 1):(length(x) - n + offset + 1)
  num <- vapply(
    locs,
    function(i) f(x[(i - offset):(i + offset)]), ...
  , numeric(1)
  )

  c(rep(NA, offset), num)
}
```

This is effectively the same as the implementation in `zoo::rollapply()`, which provides many more features and much more error checking.

11.3.4 Parallelisation

One interesting thing about the implementation of `lapply()` is that because each iteration is isolated from all others, the order in which they are computed doesn't matter. For example, `lapply3()` scrambles the order of computation, but the results are always the same:

```
lapply3 <- function(x, f, ...) {
  out <- vector("list", length(x))
  for (i in sample(seq_along(x))) {
    out[[i]] <- f(x[[i]]), ...
  }
  out
}
unlist(lapply(1:10, sqrt))
#> [1] 1.00 1.41 1.73 2.00 2.24 2.45 2.65 2.83 3.00 3.16
unlist(lapply3(1:10, sqrt))
#> [1] 1.00 1.41 1.73 2.00 2.24 2.45 2.65 2.83 3.00 3.16
```

This has a very important consequence: since we can compute each element in any order, it's easy to dispatch the tasks to different cores, and compute them in parallel. This is what `parallel::mcLapply()` (and `parallel::mclapply()`) does. (These functions are not available in Windows, but you can use the similar `parLapply()` with a bit more work. See `parallelise` for more details.)

```
library(parallel)
unlist(mcLapply(1:10, sqrt, mc.cores = 4))
#> [1] 1.00 1.41 1.73 2.00 2.24 2.45 2.65 2.83 3.00 3.16
```

In this case, `mclapply()` is actually slower than `lapply()`. This is because the cost of the individual computations is low, and additional work is needed to send the computation to the different cores and to collect the results.

If we take a more realistic example, generating bootstrap replicates of a linear model for example, the advantages are clearer:

```
boot_df <- function(x) x[sample(nrow(x), rep = T), ]
rsquared <- function(mod) summary(mod)$r.square
boot_lm <- function(i) {
  rsquared(lm(mpg ~ wt + disp, data = boot_df(mtcars)))
```

```

}

system.time(lapply(1:500, boot_lm))
#>    user  system elapsed
#>  0.660   0.000   0.675
system.time(mclapply(1:500, boot_lm, mc.cores = 2))
#>    user  system elapsed
#>  0.310   0.030   0.346

```

While increasing the number of cores will not always lead to linear improvement, switching from `lapply()` or `Map()` to its parallelised forms can dramatically improve computational performance.

11.3.5 Exercises

1. Use `vapply()` to:
 - a) Compute the standard deviation of every column in a numeric data frame.
 - b) Compute the standard deviation of every numeric column in a mixed data frame. (Hint: you'll need to use `vapply()` twice.)
2. Why is using `sapply()` to get the `class()` of each element in a data frame dangerous?
3. The following code simulates the performance of a t-test for non-normal data. Use `sapply()` and an anonymous function to extract the p-value from every trial.

```

trials <- replicate(
  100,
  t.test(rpois(10, 10), rpois(7, 10)),
  simplify = FALSE
)

```

Extra challenge: get rid of the anonymous function by using `[[` directly.

4. What does `replicate()` do? What sort of for loop does it eliminate? Why do its arguments differ from `lapply()` and friends?
5. Implement a version of `lapply()` that supplies `FUN` with both the name and the value of each component.
6. Implement a combination of `Map()` and `vapply()` to create an `lapply()` variant that iterates in parallel over all of its inputs and stores its outputs in a vector (or a matrix). What arguments should the function take?
7. Implement `mcsapply()`, a multicore version of `sapply()`. Can you implement `mcvapply()`, a parallel version of `vapply()`? Why or why not?

11.4 Manipulating matrices and data frames

Functionals can also be used to eliminate loops in common data manipulation tasks. In this section, we'll give a brief overview of the available options, hint at how they can help you, and point you in the right direction to learn more. We'll cover three categories of data structure functionals:

- `apply()`, `sweep()`, and `outer()` work with matrices.
- `tapply()` summarises a vector by groups defined by another vector.

- the `plyr` package, which generalises `tapply()` to make it easy to work with data frames, lists, or arrays as inputs, and data frames, lists, or arrays as outputs.

11.4.1 Matrix and array operations

So far, all the functionals we've seen work with 1d input structures. The three functionals in this section provide useful tools for working with higher-dimensional data structures. `apply()` is a variant of `sapply()` that works with matrices and arrays. You can think of it as an operation that summarises a matrix or array by collapsing each row or column to a single number. 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
```

There are a few caveats to using `apply()`. It doesn't have a `simplify` argument, so you can never be completely sure what type of output you'll get. This means that `apply()` is not safe to use inside a function unless you carefully check the inputs. `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] TRUE
a2 <- apply(a, 2, identity)
identical(a, a2)
#> [1] TRUE
```

(You can put high-dimensional arrays back in the right order using `aperm()`, or use `plyr::aapply()`, which is idempotent.)

`sweep()` allows you to "sweep" out the values of a summary statistic. It is often used with `apply()` to standardise arrays. The following example scales the rows of a matrix so that all values lie between 0 and 1.

```
x <- matrix(rnorm(20, 0, 10), nrow = 4)
x1 <- sweep(x, 1, apply(x, 1, min), `/-`)
x2 <- sweep(x1, 1, apply(x1, 1, max), `/`)
```

The final matrix functional is `outer()`. It's a little different in that it takes multiple vector inputs and creates a matrix or array output where the input function is run over every combination of the inputs:

```
# Create a times table
outer(1:3, 1:10, "*")
#> [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
#> [1,]    1    2    3    4    5    6    7    8    9   10
#> [2,]    2    4    6    8   10   12   14   16   18   20
#> [3,]    3    6    9   12   15   18   21   24   27   30
```

Good places to learn more about `apply()` and friends are:

- “Using `apply`, `sapply`, `lapply` in R” by Peter Werner.
- “The infamous `apply` function” by Slawa Rokicki.
- “The R `apply` function - a tutorial with examples” by axiomOfChoice.
- The stackoverflow question “R Grouping functions: `sapply` vs. `lapply` vs. `apply` vs. `tapply` vs. `by` vs. `aggregate`”.

11.4.2 Group apply

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
#> 70.5 75.0
```

`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 what the `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] 69 71 74 66 71 67 73 69 73 72
#>
#> $B
#> [1] 73 79 74 72 74 76 76 68 77 74 79 78
```

Then `tapply()` is just the combination of `split()` and `sapply()`:

```
tapply2 <- function(x, group, f, ..., simplify = TRUE) {
  pieces <- split(x, group)
  sapply(pieces, f, simplify = simplify)
}
tapply2(pulse, group, length)
#> A   B
#> 10 12
tapply2(pulse, group, mean)
#>    A    B
#> 70.5 75.0
```

Being able to rewrite `tapply()` as a combination of `split()` and `sapply()` is a good indication that we’ve identified some useful building blocks.

11.4.3 The plyr package

One challenge with using the base functionals is that they have grown organically over time, and have been written by multiple authors. This means that they are not very consistent:

- With `tapply()` and `sapply()`, the `simplify` argument is called `simplify`. With `mapply()`, it's called `SIMPLIFY`. With `apply()`, the argument is absent.
- `vapply()` is a variant of `sapply()` that allows you to describe what the output should be, but there are no corresponding variants for `tapply()`, `apply()`, or `Map()`.
- The first argument of most base functionals is a vector, but the first argument in `Map()` is a function.

This makes learning these operators challenging, as you have to memorise all of the variations. Additionally, if you think about the possible combinations of input and output types, base R only covers a partial set of cases:

| | list | data frame | array |
|------------|-----------------------|------------|-----------------------|
| list | <code>lapply()</code> | | <code>sapply()</code> |
| data frame | <code>by()</code> | | |
| array | | | <code>apply()</code> |

This was one of the driving motivations behind the creation of the `plyr` package. It provides consistently named functions with consistently named arguments and covers all combinations of input and output data structures:

| | list | data frame | array |
|------------|----------------------|----------------------|----------------------|
| list | <code>llply()</code> | <code>ldply()</code> | <code>laply()</code> |
| data frame | <code>dlply()</code> | <code>ddply()</code> | <code>daply()</code> |
| array | <code>alply()</code> | <code>adply()</code> | <code>aaply()</code> |

Each of these functions splits up the input, applies a function to each piece, and then combines the results. Overall, this process is called “split-apply-combine”. You can read more about it and `plyr` in “The Split-Apply-Combine Strategy for Data Analysis”, an open-access article published in the Journal of Statistical Software.

11.4.4 Exercises

- How does `apply()` arrange the output? Read the documentation and perform some experiments.
- There's no equivalent to `split() + vapply()`. Should there be? When would it be useful? Implement one yourself.
- Implement a pure R version of `split()`. (Hint: use `unique()` and subsetting.) Can you do it without a for loop?
- What other types of input and output are missing? Brainstorm before you look up some answers in the `plyr` paper.

11.5 Manipulating lists

Another way of thinking about functionals is as a set of general tools for altering, subsetting, and collapsing lists. Every functional programming language has three tools for this: `Map()`, `Reduce()`, and `Filter()`. We've seen `Map()` already, and the following sections describe `Reduce()`, a powerful tool for extending two-argument functions, and `Filter()`, a member of an important class of functionals that work with predicates, functions that return a single TRUE or FALSE.

11.5.1 Reduce()

`Reduce()` reduces a vector, `x`, to a single value by recursively calling a function, `f`, two arguments at a time. It combines the first two elements with `f`, then combines the result of that call with the third element, and so on. Calling `Reduce(f, 1:3)` is equivalent to `f(f(1, 2), 3)`. `Reduce` is also known as `fold`, because it folds together adjacent elements in the list.

The following two examples show what `Reduce` does with an infix and prefix function:

```
Reduce(`+`, 1:3) # -> ((1 + 2) + 3)
Reduce(sum, 1:3) # -> sum(sum(1, 2), 3)
```

The essence of `Reduce()` can be described by a simple for loop:

```
Reduce2 <- function(f, x) {
  out <- x[[1]]
  for(i in seq(2, length(x))) {
    out <- f(out, x[[i]])
  }
  out
}
```

The real `Reduce()` is more complicated because it includes arguments to control whether the values are reduced from the left or from the right (`right`), an optional initial value (`init`), and an option to output intermediate results (`accumulate`).

`Reduce()` is an elegant way of extending a function that works with two inputs into a function that can deal with any number of inputs. It's useful for implementing many types of recursive operations, like merges and intersections. (We'll see another use in the final case study.) Imagine you have a list of numeric vectors, and you want to find the values that occur in every element:

```
1 <- replicate(5, sample(1:10, 15, replace = T), simplify = FALSE)
str(1)
#> List of 5
#> $ : int [1:15] 10 8 8 1 10 6 6 7 3 9 ...
#> $ : int [1:15] 5 1 2 10 4 1 1 9 9 6 ...
#> $ : int [1:15] 1 6 2 7 9 10 8 9 4 6 ...
#> $ : int [1:15] 9 4 6 10 1 6 9 3 4 4 ...
#> $ : int [1:15] 4 4 7 7 1 8 1 7 2 3 ...
```

You could do that by intersecting each element in turn:

```
intersect(intersect(intersect(intersect(1[[1]], 1[[2]]),
  1[[3]]), 1[[4]]), 1[[5]])
#> [1] 10 1 6 4 2
```

That's hard to read. With `Reduce()`, the equivalent is:

```
Reduce(intersect, 1)
#> [1] 10 1 6 4 2
```

11.5.2 Predicate functionals

A **predicate** is a function that returns a single TRUE or FALSE, like `is.character`, `all`, or `is.NULL`. A predicate functional applies a predicate to each element of a list or data frame. There are three useful predicate functionals in base R: `Filter()`, `Find()`, and `Position()`.

- `Filter()` selects only those elements which match the predicate.
- `Find()` returns the first element which matches the predicate (or the last element if `right = TRUE`).
- `Position()` returns the position of the first element that matches the predicate (or the last element if `right = TRUE`).

Another useful predicate functional is `where()`, a custom functional that generates a logical vector from a list (or a data frame) and a predicate:

```
where <- function(f, x) {
  vapply(x, f, logical(1))
}
```

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"))
where(is.factor, df)
#>   x     y
#> FALSE  TRUE
str(Filter(is.factor, df))
#> 'data.frame':   3 obs. of  1 variable:
#> $ y: Factor w/ 3 levels "a","b","c": 1 2 3
str(Find(is.factor, df))
#> Factor w/ 3 levels "a","b","c": 1 2 3
Position(is.factor, df)
#> [1] 2
```

11.5.3 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. Use `Filter()` and `vapply()` to create a function that applies a summary statistic to every numeric column in a data frame.
3. What's the relationship between `which()` and `Position()`? What's the relationship between `where()` and `Filter()`?
4. Implement `Any()`, a function that takes a list and a predicate function, and returns TRUE if the predicate function returns TRUE for any of the inputs. Implement `All()` similarly.
5. 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.)

11.6 Mathematical functionals

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.

In this section we'll use some of R's built-in mathematical functionals. There are three functionals that work with functions to return single numeric values:

- `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()`

Let's explore how these are 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
```

In statistics, optimisation is often used for maximum likelihood estimation (MLE). In MLE, we have two sets of parameters: the data, which is fixed for a given problem, and the parameters, which vary as we try to find the maximum. These two sets of parameters make the problem well suited for closures. Combining closures with optimisation gives rise to the following approach to solving MLE problems.

The following example shows how we might find the maximum likelihood estimate for λ , if our data come from a Poisson distribution. First, we create a function factory that, given a dataset, returns a function that computes the negative log likelihood (NLL) for parameter `lambda`. In R, it's common to work with the negative since `optimise()` defaults to finding the minimum.

```
poisson_nll <- function(x) {
  n <- length(x)
  sum_x <- sum(x)
  function(lambda) {
    n * lambda - sum_x * log(lambda) # + terms not involving lambda
  }
}
```

Note how the closure allows us to precompute values that are constant with respect to the data.

We can use this function factory to generate specific NLL functions for input data. Then `optimise()` allows us to find the best values (the maximum likelihood estimates), given a generous starting range.

```

x1 <- c(41, 30, 31, 38, 29, 24, 30, 29, 31, 38)
x2 <- c(6, 4, 7, 3, 3, 7, 5, 2, 2, 7, 5, 4, 12, 6, 9)
nll1 <- poisson_nll(x1)
nll2 <- poisson_nll(x2)

optimise(nll1, c(0, 100))$minimum
#> [1] 32.1
optimise(nll2, c(0, 100))$minimum
#> [1] 5.47

```

We can check that these values are correct by comparing them to the analytic solution: in this case, it's just the mean of the data, 32.1 and 5.467.

Another important mathematical functional is `optim()`. It is a generalisation of `optimise()` that works with more than one dimension. If you're interested in how it works, you might want to explore the `Rvmmin` package, which provides a pure-R implementation of `optim()`. Interestingly `Rvmmin` is no slower than `optim()`, even though it is written in R, not C. For this problem, the bottleneck lies not in controlling the optimisation but with having to evaluate the function multiple times.

11.6.1 Exercises

1. 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.
2. Challenge: read about the fixed point algorithm. Complete the exercises using R.

11.7 Loops that should be left as is

Some loops have no natural functional equivalent. In this section you'll learn about three common cases:

- modifying in place
- recursive functions
- while loops

It's possible to torture these problems to use a functional, but it's not a good idea. You'll create code that is harder to understand, eliminating the main reason for using functionals in the first case.

11.7.1 Modifying in place

If you need to modify part of an existing data frame, it's often better to use a for loop. For example, the following code performs a variable-by-variable transformation by matching the names of a list of functions to the names of variables in a data frame.

```

trans <- list(
  disp = function(x) x * 0.0163871,
  am = function(x) factor(x, labels = c("auto", "manual"))
)
for(var in names(trans)) {
  mtcars[[var]] <- trans[[var]](mtcars[[var]])
}

```

We wouldn't normally use `lapply()` to replace this loop directly, but it is possible. Just replace the loop with `lapply()` by using `<->`:

```
lapply(names(trans), function(var) {
  mtcars[[var]] <- trans[[var]](mtcars[[var]])
})
```

The for loop is gone, but the code is longer and much harder to understand. The reader needs to understand `<->` and how `x[[y]] <-> z` works (it's not simple!). In short, we've taken a simple, easily understood for loop, and turned it into something few people will understand: not a good idea!

11.7.2 Recursive relationships

It's hard to convert a for loop into a functional when the relationship between elements is not independent, or is defined recursively. For example, exponential smoothing works by taking a weighted average of the current and previous data points. The `exps()` function below implements exponential smoothing with a for loop.

```
exps <- function(x, alpha) {
  s <- numeric(length(x) + 1)
  for (i in seq_along(s)) {
    if (i == 1) {
      s[i] <- x[i]
    } else {
      s[i] <- alpha * x[i] + (1 - alpha) * s[i - 1]
    }
  }
  s
}
x <- runif(6)
exps(x, 0.5)
#> [1] 0.0518 0.3078 0.5284 0.6148 0.5978 0.3628      NA
```

We can't eliminate the for loop because none of the functionals we've seen allow the output at position `i` to depend on both the input and output at position `i - 1`.

One way to eliminate the for loop in this case is to solve the recurrence relation by removing the recursion and replacing it with explicit references. This requires a new set of mathematical tools, and is challenging, but it can pay off by producing a simpler function.

11.7.3 While loops

Another type of looping construct in R is the `while` loop. It keeps running until some condition is met. `while` loops are more general than `for` loops: you can rewrite every `for` loop as a `while` loop, but you can't do the reverse. For example, we could turn this `for` loop:

```
for (i in 1:10) print(i)
```

into this `while` loop:

```
i <- 1
while(i <= 10) {
  print(i)
  i <- i + 1
}
```

Not every while loop can be turned into a for loop because many while loops don't know in advance how many times they will be run:

```
i <- 0
while(TRUE) {
  if (runif(1) > 0.9) break
  i <- i + 1
}
```

This is a common problem when you're writing simulations.

In this case we can remove the loop by recognising a special feature of the problem. Here we're counting the number of successes before Bernoulli trial with $p = 0.1$ fails. This is a geometric random variable, so you could replace the code with `i <- rgeom(1, 0.1)`. Reformulating the problem in this way is hard to do in general, but you'll benefit greatly if you can do it for your problem.

11.8 A family of functions

To finish off the chapter, this case study shows how you can use functionals to take a simple building block and make it powerful and general. I'll start with a simple idea, adding two numbers together, and use functionals to extend it to summing multiple numbers, computing parallel and cumulative sums, and summing across array dimensions.

We'll start by defining a very simple addition function, one which takes two scalar arguments:

```
add <- function(x, y) {
  stopifnot(length(x) == 1, length(y) == 1,
            is.numeric(x), is.numeric(y))
  x + y
}
```

(We're using R's existing addition operator here, which does much more, but the focus here is on how we can take very simple building blocks and extend them to do more.)

I'll also add an `na.rm` argument. A helper function will make this a bit easier: if `x` is missing it should return `y`, if `y` is missing it should return `x`, and if both `x` and `y` are missing then it should return another argument to the function: `identity`. This function is probably a bit more general than what we need now, but it's useful if we implement other binary operators.

```
rm_na <- function(x, y, identity) {
  if (is.na(x) && is.na(y)) {
    identity
  } else if (is.na(x)) {
    y
  } else {
    x
  }
}
rm_na(NA, 10, 0)
#> [1] 10
rm_na(10, NA, 0)
#> [1] 10
rm_na(NA, NA, 0)
#> [1] 0
```

This allows us to write a version of `add()` that can deal with missing values if needed:

```

add <- function(x, y, na.rm = FALSE) {
  if (na.rm && (is.na(x) || is.na(y))) rm_na(x, y, 0) else x + y
}
add(10, NA)
#> [1] NA
add(10, NA, na.rm = TRUE)
#> [1] 10
add(NA, NA)
#> [1] NA
add(NA, NA, na.rm = TRUE)
#> [1] 0

```

Why did we pick an identity of 0? Why should `add(NA, NA, na.rm = TRUE)` return 0? Well, for every other input it returns a number, so even if both arguments are NA, it should still do that. What number should it return? We can figure it out because addition is associative, which means that the order of addition doesn't matter. That means that the following two function calls should return the same value:

```

add(add(3, NA, na.rm = TRUE), NA, na.rm = TRUE)
#> [1] 3
add(3, add(NA, NA, na.rm = TRUE), na.rm = TRUE)
#> [1] 3

```

This implies that `add(NA, NA, na.rm = TRUE)` must be 0, and hence `identity = 0` is the correct default.

Now that we have the basics working, we can extend the function to deal with more complicated inputs. One obvious generalisation is to add more than two numbers. We can do this by iteratively adding two numbers: if the input is `c(1, 2, 3)` we compute `add(add(1, 2), 3)`. This is a simple application of `Reduce()`:

```

r_add <- function(xs, na.rm = TRUE) {
  Reduce(function(x, y) add(x, y, na.rm = na.rm), xs)
}
r_add(c(1, 4, 10))
#> [1] 15

```

This looks good, but we need to test a few special cases:

```

r_add(NA, na.rm = TRUE)
#> [1] NA
r_add(numeric())
#> NULL

```

These are incorrect. In the first case, we get a missing value even though we've explicitly asked to ignore them. In the second case, we get `NULL` instead of a length one numeric vector (as we do for every other set of inputs).

The two problems are related. If we give `Reduce()` a length one vector, it doesn't have anything to reduce, so it just returns the input. If we give it an input of length zero, it always returns `NULL`. The easiest way to fix this problem is to use the `init` argument of `Reduce()`. This is added to the start of every input vector:

```

r_add <- function(xs, na.rm = TRUE) {
  Reduce(function(x, y) add(x, y, na.rm = na.rm), xs, init = 0)
}
r_add(c(1, 4, 10))
#> [1] 15
r_add(NA, na.rm = TRUE)
#> [1] 0
r_add(numeric())
#> [1] 0

```

`r_add()` is equivalent to `sum()`.

It would be nice to have a vectorised version of `add()` so that we can perform the addition of two vectors of numbers in element-wise fashion. We could use `Map()` or `vapply()` to implement this, but neither is perfect. `Map()` returns a list, instead of a numeric vector, so we need to use `simplify2array()`. `vapply()` returns a vector but it requires us to loop over a set of indices.

```
v_add1 <- function(x, y, na.rm = FALSE) {
  stopifnot(length(x) == length(y), is.numeric(x), is.numeric(y))
  if (length(x) == 0) return(numeric())
  simplify2array(
    Map(function(x, y) add(x, y, na.rm = na.rm), x, y)
  )
}

v_add2 <- function(x, y, na.rm = FALSE) {
  stopifnot(length(x) == length(y), is.numeric(x), is.numeric(y))
  vapply(seq_along(x), function(i) add(x[i], y[i], na.rm = na.rm),
    numeric(1))
}
```

A few test cases help to ensure that it behaves as we expect. We're a bit stricter than base R here because we don't do recycling. (You could add that if you wanted, but I find that recycling is a frequent source of silent bugs.)

```
# Both versions give the same results
v_add1(1:10, 1:10)
#> [1] 2 4 6 8 10 12 14 16 18 20
v_add1(numeric(), numeric())
#> numeric(0)
v_add1(c(1, NA), c(1, NA))
#> [1] 2 NA
v_add1(c(1, NA), c(1, NA), na.rm = TRUE)
#> [1] 2 0
```

Another variant of `add()` is the cumulative sum. We can implement it with `Reduce()` by setting the `accumulate` argument to `TRUE`:

```
c_add <- function(xs, na.rm = FALSE) {
  Reduce(function(x, y) add(x, y, na.rm = na.rm), xs,
    accumulate = TRUE)
}
c_add(1:10)
#> [1] 1 3 6 10 15 21 28 36 45 55
c_add(10:1)
#> [1] 10 19 27 34 40 45 49 52 54 55
```

This is equivalent to `cumsum()`.

Finally, we might want to define addition for more complicated data structures like matrices. We could create `row` and `col` variants that sum across rows and columns, respectively, or we could go the whole hog and define an array version that could sum across any arbitrary set of dimensions. These are easily implemented as combinations of `r_add()` and `apply()`.

```
row_sum <- function(x, na.rm = FALSE) {
  apply(x, 1, r_add, na.rm = na.rm)
}
col_sum <- function(x, na.rm = FALSE) {
```

```

apply(x, 2, r_add, na.rm = na.rm)
}
arr_sum <- function(x, dim, na.rm = FALSE) {
  apply(x, dim, r_add, na.rm = na.rm)
}

```

The first two are equivalent to `rowSums()` and `colSums()`.

If every function we have created has an existing equivalent in base R, why did we bother? There are two main reasons:

- Since all variants were implemented by combining a simple binary operator (`add()`) and a well-tested functional (`Reduce()`, `Map()`, `apply()`), we know that our variants will behave consistently.
- We can apply the same infrastructure to other operators, especially those that might not have the full suite of variants in base R.

The downside of this approach is that these implementations are not that efficient. (For example, `colSums(x)` is much faster than `apply(x, 2, sum)`.) However, even if they aren't that fast, simple implementations are still a good starting point because they're less likely to have bugs. When you create faster versions, you can compare the results to make sure your fast versions are still correct.

If you enjoyed this section, you might also enjoy “List out of lambda”, a blog article by Steve Losh that shows how you can produce high level language structures (like lists) out of more primitive language features (like closures, aka lambdas).

11.8.1 Exercises

1. Implement `smaller` and `larger` functions that, given two inputs, return either the smaller or the larger value. Implement `na.rm = TRUE`: what should the identity be? (Hint: `smaller(NA, NA, na.rm = TRUE)`, `na.rm = TRUE`) must be `x`, so `smaller(NA, NA, na.rm = TRUE)` must be bigger than any other value of `x`.) Use `smaller` and `larger` to implement equivalents of `min()`, `max()`, `pmin()`, `pmax()`, and new functions `row_min()` and `row_max()`.
2. Create a table that has and, or, add, multiply, smaller, and larger in the columns and binary operator, reducing variant, vectorised variant, and array variants in the rows.
 - a) Fill in the cells with the names of base R functions that perform each of the roles.
 - b) Compare the names and arguments of the existing R functions. How consistent are they? How could you improve them?
 - c) Complete the matrix by implementing any missing functions.
3. How does `paste()` fit into this structure? What is the scalar binary function that underlies `paste()`? What are the `sep` and `collapse` arguments to `paste()` equivalent to? Are there any `paste` variants that don't have existing R implementations?

Chapter 12

Function operators

12.1 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. In some ways, function operators are similar to functionals: there's nothing you can't do without them, but they can make your code more readable and expressive, and they can help you write code faster. The main difference is that functionals extract common patterns of loop use, where function operators extract common patterns of anonymous function use.

The following code shows a simple function operator, `chatty()`. It wraps a function, making a new function that prints out its first argument. It's useful because it gives you a window to see how functionals, like `vapply()`, work.

```
chatty <- function(f) {
  function(x, ...) {
    res <- f(x, ...)
    cat("Processing ", x, "\n", sep = "")
    res
  }
}
f <- function(x) x ^ 2
s <- c(3, 2, 1)
chatty(f)(1)
#> Processing 1
#> [1] 1

vapply(s, chatty(f), numeric(1))
#> Processing 3
#> Processing 2
#> Processing 1
#> [1] 9 4 1
```

In the last chapter, we saw that many built-in functionals, like `Reduce()`, `Filter()`, and `Map()`, have very few arguments, so we had to use anonymous functions to modify how they worked. In this chapter, we'll build specialised substitutes for common anonymous functions that allow us to communicate our intent more clearly. For example, in multiple inputs we used an anonymous function with `Map()` to supply fixed arguments:

```
Map(function(x, y) f(x, y, zs), xs, ys)
```

Later in this chapter, we'll learn about partial application using the `partial()` function. Partial application encapsulates the use of an anonymous function to supply default arguments, and allows us to write succinct code:

```
Map(partial(f, zs = zs), xs, ys)
```

This is an important use of FOs: by transforming the input function, you eliminate parameters from a functional. In fact, as long as the inputs and outputs of the function remain the same, this approach allows your functionals to be more extensible, often in ways you haven't thought of.

The chapter covers four important types of FO: behaviour, input, output, and combining. For each type, I'll show you some useful FOs, and how you can use as another to decompose problems: as combinations of multiple functions instead of combinations of arguments. The goal is not to exhaustively list every possible FO, but to show a selection that demonstrate how they work together with other FP techniques. For your own work, you'll need to think about and experiment with how function operators can help you solve recurring problems.

Outline

- Behavioural FOs introduces you to FOs that change the behaviour of a function like automatically logging usage to disk or ensuring that a function is run only once.
- Output FOs shows you how to write FOs that manipulate the output of a function. These can do simple things like capturing errors, or fundamentally change what the function does.
- Input FOs describes how to modify the inputs to a function using a FO like `Vectorize()` or `partial()`.
- Combining FOs shows the power of FOs that combine multiple functions with function composition and logical operations.

Prerequisites

As well as writing FOs from scratch, this chapter uses function operators from the `memoise`, `plyr`, and `pryr` packages. Install them by running `install.packages(c("memoise", "plyr", "pryr"))`.

12.2 Behavioural FOs

Behavioural FOs leave the inputs and outputs of a function unchanged, but add some extra behaviour. In this section, we'll look at functions which implement three useful behaviours:

- Add a delay to avoid swamping a server with requests.
- Print to console every n invocations to check on a long running process.
- Cache previous computations to improve performance.

To motivate these behaviours, imagine we want to download a long vector of URLs. That's pretty simple with `lapply()` and `download_file()`:

```
download_file <- function(url, ...) {
  download.file(url, basename(url), ...)
}
lapply(urls, download_file)
```

(`download_file()` is a simple wrapper around `utils::download.file()` which provides a reasonable default for the file name.)

There are a number of useful behaviours we might want to add to this function. If the list was long, we might want to print a `.` every ten URLs so we know that the function's still working. If we're downloading files over the internet, we might want to add a small delay between each request to avoid hammering the server. Implementing these behaviours in a `for` loop is rather complicated. We can no longer use `lapply()` because we need an external counter:

```
i <- 1
for(url in urls) {
  i <- i + 1
  if (i %% 10 == 0) cat(".")
  Sys.sleep(1)
  download_file(url)
}
```

Understanding this code is hard because different concerns (iteration, printing, and downloading) are interleaved. In the remainder of this section we'll create FOs that encapsulate each behaviour and allow us to write code like this:

```
lapply(urls, dot_every(10, delay_by(1, download_file)))
```

Implementing `delay_by()` is straightforward, and follows the same basic template that we'll see for the majority of FOs in this chapter:

```
delay_by <- function(delay, f) {
  function(...) {
    Sys.sleep(delay)
    f(...)
  }
}
system.time(runif(100))
#>   user  system elapsed
#>       0        0        0
system.time(delay_by(0.1, runif)(100))
#>   user  system elapsed
#>  0.000  0.000  0.101
```

`dot_every()` is a little bit more complicated because it needs to manage a counter. Fortunately, we saw how to do that in mutable state.

```
dot_every <- function(n, f) {
  i <- 1
  function(...) {
    if (i %% n == 0) cat(".")
    i <- i + 1
    f(...)
  }
}
x <- lapply(1:100, runif)
x <- lapply(1:100, dot_every(10, runif))
#> .....
```

Notice that I've made the function the last argument in each FO. This makes it easier to read when we compose multiple function operators. If the function were the first argument, then instead of:

```
download <- dot_every(10, delay_by(1, download_file))
```

we'd have

```
download <- dot_every(delay_by(download_file, 1), 10)
```

That's harder to follow because (e.g.) the argument of `dot_every()` is far away from its call. This is sometimes called the Dagwood sandwich problem: you have too much filling (too many long arguments) between your slices of bread (parentheses).

I've also tried to give the FOs descriptive names: `delay_by` 1 (second), `(print a)` `dot every` 10 (invocations). The more clearly the function names used in your code express your intent, the easier it will be for others (including future you) to read and understand the code.

12.2.1 Memoisation

Another thing you might worry about when downloading multiple files is accidentally downloading the same file multiple times. You could avoid this by calling `unique()` on the list of input URLs, or manually managing a data structure that mapped the URL to the result. An alternative approach is to use memoisation: modify a function to automatically cache its results.

```
library(memoise)
```

```
slow_function <- function() {
  Sys.sleep(1)
  10
}
system.time(slow_function())
#>    user  system elapsed
#>      0        0       1
system.time(slow_function())
#>    user  system elapsed
#>     0.01    0.00    1.00
fast_function <- memoise(slow_function)
system.time(fast_function())
#>    user  system elapsed
#>      0        0       1
system.time(fast_function())
#>    user  system elapsed
#>     0.020   0.000   0.013
```

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.

A realistic use of memoisation is computing the Fibonacci series. The Fibonacci series is defined recursively: the first two values are 1 and 1, then $f(n) = f(n - 1) + f(n - 2)$. A naive version implemented in R would be very slow because, for example, `fib(10)` computes `fib(9)` and `fib(8)`, and `fib(9)` computes `fib(8)` and `fib(7)`, and so on. As a result, the value for each value in the series gets computed many, many times. Memoising `fib()` makes the implementation much faster because each value is computed only once.

```
fib <- function(n) {
  if (n < 2) return(1)
  fib(n - 2) + fib(n - 1)
}
system.time(fib(23))
#>    user  system elapsed
```

```
#> 0.050 0.000 0.052
system.time(fib(24))
#> user system elapsed
#> 0.06 0.00 0.06

fib2 <- memoise(function(n) {
  if (n < 2) return(1)
  fib2(n - 2) + fib2(n - 1)
})
system.time(fib2(23))
#> user system elapsed
#> 0.020 0.000 0.028
system.time(fib2(24))
#> user system elapsed
#> 0.000 0.000 0.001
```

It doesn't make sense to memoise all functions. For example, a memoised random number generator is no longer random:

```
runifm <- memoise(runif)
runifm(5)
#> [1] 0.883 0.678 0.073 0.920 0.988
runifm(5)
#> [1] 0.883 0.678 0.073 0.920 0.988
```

Once we understand `memoise()`, it's straightforward to apply to our problem:

```
download <- dot_every(10, memoise(delay_by(1, download_file)))
```

This gives a function that we can easily use with `lapply()`. However, if something goes wrong with the loop inside `lapply()`, it can be difficult to tell what's going on. The next section will show how we can use FOs to pull back the curtain and look inside.

12.2.2 Capturing function invocations

One challenge with functionals is that it can be hard to see what's going on inside of them. It's not easy to pry open their internals like it is with a for loop. Fortunately we can use FOs to peer behind the curtain with `tee()`.

`tee()`, defined below, has three arguments, all functions: `f`, the function to modify; `on_input`, a function that's called with the inputs to `f`; and `on_output`, a function that's called with the output from `f`.

```
ignore <- function(...) NULL
tee <- function(f, on_input = ignore, on_output = ignore) {
  function(...) {
    on_input(...)
    output <- f(...)
    on_output(output)
    output
  }
}
```

(The function is inspired by the unix shell command `tee`, which is used to split up streams of file operations so that you can both display what's happening and save intermediate results to a file.)

We can use `tee()` to look inside the `uniroot()` functional, and see how it iterates its way to a solution. The following example finds where x and $\cos(x)$ intersect:

```

g <- function(x) cos(x) - x
zero <- uniroot(g, c(-5, 5))
show_x <- function(x, ...) cat(sprintf("%+.08f", x), "\n")

# The location where the function is evaluated:
zero <- uniroot(tee(g, on_input = show_x), c(-5, 5))
#> -5.00000000
#> +5.00000000
#> +0.28366219
#> +0.87520341
#> +0.72298040
#> +0.73863091
#> +0.73908529
#> +0.73902425
#> +0.73908529

# The value of the function:
zero <- uniroot(tee(g, on_output = show_x), c(-5, 5))
#> +5.28366219
#> -4.71633781
#> +0.67637474
#> -0.23436269
#> +0.02685676
#> +0.00076012
#> -0.00000026
#> +0.00010189
#> -0.00000026

```

`cat()` allows us to see what's happening as the function runs, but it doesn't give us a way to work with the values after the function as completed. To do that, we could capture the sequence of calls by creating a function, `remember()`, that records every argument called and retrieves them when coerced into a list. The small amount of S3 code needed is explained in S3.

```

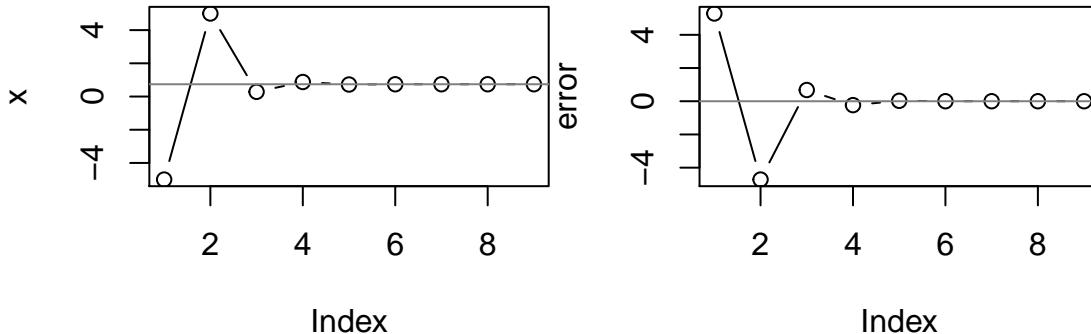
remember <- function() {
  memory <- list()
  f <- function(...) {
    # This is inefficient!
    memory <- append(memory, list(...))
    invisible()
  }

  structure(f, class = "remember")
}
as.list.remember <- function(x, ...) {
  environment(x)$memory
}
print.remember <- function(x, ...) {
  cat("Remembering...\n")
  str(as.list(x))
}

```

Now we can draw a picture showing how `uniroot` zeroes in on the final answer:

```
locs <- remember()
vals <- remember()
zero <- uniroot(tee(g, locs, vals), c(-5, 5))
x <- unlist(as.list(locs))
error <- unlist(as.list(vals))
plot(x, type = "b"); abline(h = 0.739, col = "grey50")
plot(error, type = "b"); abline(h = 0, col = "grey50")
```



12.2.3 Laziness

The function operators we've seen so far follow a common pattern:

```
funop <- function(f, otherargs) {
  function(...) {
    # maybe do something
    res <- f(...)
    # maybe do something else
    res
  }
}
```

Unfortunately there's a problem with this implementation because function arguments are lazily evaluated: `f()` may have changed between applying the FO and evaluating the function. This is a particular problem if you're using a for loop to create multiple function operators. In the following example, we take a list of functions and delay each one. But when we try to evaluate the mean, we get the sum instead.

```
funcs <- list(mean = mean, sum = sum)

funcs_m <- vector("list", length(funcs))
for (fun in names(funcs)) {
  funcs_m[[fun]] <- delay_by(funcs[[fun]], delay = 0.1)
}
funcs_m$mean(1:10)
#> [1] 55
```

We can avoid that problem by explicitly forcing the evaluation of `f()`:

```
delay_by <- function(delay, f) {
  force(f)
  function(...) {
    Sys.sleep(delay)
    f(...)
  }
}
```

```

}

funsm_m <- vector("list", length(funsm))
for (fun in names(funsm)) {
  funsm_m[[fun]] <- delay_by(funsm[[fun]], delay = 0.1)
}
funsm_m$mean(1:10)
#> [1] 5.5

```

Note that `lapply()` and friends have special “non-lazy” behaviour so you don’t see this problem:

```

delay_by <- function(delay, f) {
  function(...) {
    Sys.sleep(delay)
    f(...)
  }
}

funsm_m <- lapply(funsm, delay_by, delay = 0.1)
funsm_m$mean(1:10)
#> [1] 5.5

```

12.2.4 Exercises

1. Write a FO that logs a time stamp and message to a file every time a function is run.
2. What does the following function do? What would be a good name for it?

```

f <- function(g) {
  force(g)
  result <- NULL
  function(...) {
    if (is.null(result)) {
      result <- g(...)
    }
    result
  }
}
runif2 <- f(runif)
runif2(5)
#> [1] 0.128 0.756 0.459 0.877 0.618
runif2(10)
#> [1] 0.128 0.756 0.459 0.877 0.618

```

3. 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.
4. Write `wait_until()` which delays execution until a specific time.
5. There are three places we could have added a memoise call: why did we choose the one we did?

```

download <- memoise(dot_every(10, delay_by(1, download_file)))
download <- dot_every(10, memoise(delay_by(1, download_file)))
download <- dot_every(10, delay_by(1, memoise(download_file)))

```

6. Why is the `remember()` function inefficient? How could you implement it in more efficient way?
7. Why does the following code, from stackoverflow, not do what you expect?

```
# return a linear function with slope a and intercept b.
f <- function(a, b) function(x) a * x + b

# create a list of functions with different parameters.
fs <- Map(f, a = c(0, 1), b = c(0, 1))

fs[[1]](3)
#> [1] 0
# should return 0 * 3 + 0 = 0
```

How can you modify `f` so that it works correctly?

12.3 Output FOs

The next step up in complexity is to modify the output of a function. This could be quite simple, or it could fundamentally change the operation of the function by returning something completely different to its usual output. In this section you'll learn about two simple modifications, `Negate()` and `failwith()`, and two fundamental modifications, `capture_it()` and `time_it()`.

12.3.1 Minor modifications

`base::Negate()` and `plyr::failwith()` offer two minor, but useful, modifications of a function that are particularly handy in conjunction with functionals.

`Negate()` takes a function that returns a logical vector (a predicate function), and returns the negation of that function. This can be a useful shortcut when a function returns the opposite of what you need. The essence of `Negate()` is very simple:

```
Negate <- function(f) {
  force(f)
  function(...) !f(...)}
}

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

I often use this idea to make a function, `compact()`, that removes all null elements from a list:

```
compact <- function(x) Filter(Negate(is.null), x)
```

`plyr::failwith()` turns a function that throws an error into a function that returns a default value when there's an error. Again, the essence of `failwith()` is simple; it's just a wrapper around `try()`, the function that captures errors and allows execution to continue.

```
failwith <- function(default = NULL, f, quiet = FALSE) {
  force(f)
  function(...) {
    out <- default
    try(out <- f(...), silent = quiet)
    out
  }
}
```

```
log("a")
#> Error in log("a"): non-numeric argument to mathematical function

failwith(NA, log)("a")
#> Error in f(...) : non-numeric argument to mathematical function
#> NA

failwith(NA, log, quiet = TRUE) ("a")
#> [1] NA
```

(If you haven't seen `try()` before, it's discussed in more detail in exceptions and debugging.)

`failwith()` is very useful in conjunction with functionals: instead of the failure propagating and terminating the higher-level loop, you can complete the iteration and then find out what went wrong. 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:

```
# If any model fails, all models fail to fit:
models <- lapply(datasets, glm, formula = y ~ x1 + x2 * x3)
# If a model fails, it will get a NULL value
models <- lapply(datasets, failwith(NULL, glm),
  formula = y ~ x1 + x2 * x3)

# remove failed models (NULLs) with compact
ok_models <- compact(models)
# extract the datasets corresponding to failed models
failed_data <- datasets[vapply(models, is.null, logical(1))]
```

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.

12.3.2 Changing what a function does

Other output function operators can have a more profound effect on the operation of the function. Instead of returning the original return value, we can return some other effect of the function evaluation. Here are two examples:

- Return text that the function `print()`ed:

```
capture_it <- function(f) {
  force(f)
  function(...) {
    capture.output(f(...))
  }
}
str_out <- capture_it(str)
str(1:10)
#> int [1:10] 1 2 3 4 5 6 7 8 9 10
str_out(1:10)
#> [1] "int [1:10] 1 2 3 4 5 6 7 8 9 10"
```

- Return how long a function took to run:

```
time_it <- function(f) {
  force(f)
```

```

function(...) {
  system.time(f(...))
}
}

```

`time_it()` allows us to rewrite some of the code from the functionals chapter:

```

compute_mean <- list(
  base = function(x) mean(x),
  sum = function(x) sum(x) / length(x)
)
x <- runif(1e6)

# Previously we used an anonymous function to time execution:
# lapply(compute_mean, function(f) system.time(f(x)))

# Now we can compose function operators:
call_fun <- function(f, ...) f(...)
lapply(compute_mean, time_it(call_fun), x)
#> $base
#>   user  system elapsed
#>   0.010  0.000  0.002
#>
#> $sum
#>   user  system elapsed
#>   0.010  0.000  0.002

```

In this example, there's not a huge benefit to using function operators, because the composition is simple and we're applying the same operator to each function. Generally, using function operators is most effective when you are using multiple operators or if the gap between creating them and using them is large.

12.3.3 Exercises

1. Create a `negative()` FO that flips the sign of the output of the function to which it is applied.
2. The `evaluate` package makes it easy to capture all the outputs (results, text, messages, warnings, errors, and plots) from an expression. Create a function like `capture_it()` that also captures the warnings and errors generated by a function.
3. Create a FO that tracks files created or deleted in the working directory (Hint: use `dir()` and `setdiff()`.) What other global effects of functions might you want to track?

12.4 Input FOs

The next step up in complexity is to modify the inputs of a function. Again, you can modify how a function works in a minor way (e.g., setting default argument values), or in a major way (e.g., converting inputs from scalars to vectors, or vectors to matrices).

12.4.1 Prefilling function arguments: partial function application

A common use of anonymous functions is to make a variant of a function that has certain arguments “filled in” already. This is called “partial function application”, and is implemented by `pryr::partial()`. Once

you have read metaprogramming, I encourage you to read the source code for `partial()` and figure out how it works — it's only 5 lines of code!

`partial()` allows us to replace code like

```
f <- function(a) g(a, b = 1)
compact <- function(x) Filter(Negate(is.null), x)
Map(function(x, y) f(x, y, zs), xs, ys)
```

with

```
f <- partial(g, b = 1)
compact <- partial(Filter, Negate(is.null))
Map(partial(f, zs = zs), xs, ys)
```

We can use this idea to simplify the code used when working with lists of functions. Instead of:

```
fun2 <- list(
  sum = function(...) sum(..., na.rm = TRUE),
  mean = function(...) mean(..., na.rm = TRUE),
  median = function(...) median(..., na.rm = TRUE)
)
```

we can write:

```
library(pryr)
#>
#> Attaching package: 'pryr'
#> The following object is masked _by_ '.GlobalEnv':
#>
#>     f
#> The following objects are masked from 'package:purrr':
#>
#>     compose, partial
fun2 <- list(
  sum = partial(sum, na.rm = TRUE),
  mean = partial(mean, na.rm = TRUE),
  median = partial(median, na.rm = TRUE)
)
```

Using `partial` function application is a straightforward task in many functional programming languages, but it's not entirely clear how it should interact with R's lazy evaluation rules. The approach `pryr::partial()` takes is to create a function that is as similar as possible to the anonymous function that you'd create by hand. Peter Meilstrup takes a different approach in his `ptools` package. If you're interested in the topic, you might want to read about the binary operators he created: `%()%`, `%>>%`, and `%<<%`.

12.4.2 Changing input types

It's also possible to make a major change to a function's input, making a function work with fundamentally different types of data. There are a few existing functions that work along these lines:

- `base::Vectorize()` converts a scalar function to a vector function. It takes a non-vectorised function and vectorises it with respect to the arguments specified in the `vectorize.args` argument. This doesn't give you any magical performance improvements, but it's useful if you want a quick and dirty way of making a vectorised function.

A mildly useful extension to `sample()` would be to vectorize it with respect to `size`. Doing so would allow you to generate multiple samples in one call.

```
sample2 <- Vectorize(sample, "size", SIMPLIFY = FALSE)
str(sample2(1:5, c(1, 1, 3)))
#> List of 3
#> $ : int 2
#> $ : int 1
#> $ : int [1:3] 2 1 5
str(sample2(1:5, 5:3))
#> List of 3
#> $ : int [1:5] 5 1 2 3 4
#> $ : int [1:4] 2 3 5 1
#> $ : int [1:3] 5 4 2
```

In this example we have used `SIMPLIFY = FALSE` to ensure that our newly vectorised function always returns a list. This is usually what you want.

- `splat()` converts a function that takes multiple arguments to a function that takes a single list of arguments.

```
splat <- function (f) {
  force(f)
  function(args) {
    do.call(f, args)
  }
}
```

This is useful if you want to invoke a function with varying arguments:

```
x <- c(NA, runif(100), 1000)
args <- list(
  list(x),
  list(x, na.rm = TRUE),
  list(x, na.rm = TRUE, trim = 0.1)
)
lapply(args, splat(mean))
#> [[1]]
#> [1] NA
#>
#> [[2]]
#> [1] 10.4
#>
#> [[3]]
#> [1] 0.478
```

- `plyr::colwise()` converts a vector function to one that works with data frames:

```
median(mtcars)
#> Error in median.default(mtcars): need numeric data
median(mtcars$mpg)
#> [1] 19.2
plyr::colwise(median)(mtcars)
#>   mpg cyl disp  hp drat    wt qsec vs am gear carb
#> 1 19.2   6 196 123 3.7 3.33 17.7 0 0     4    2
```

12.4.3 Exercises

1. Our previous `download()` function only downloads a single file. How can you use `partial()` and `lapply()` to create a function that downloads multiple files at once? What are the pros and cons of using `partial()` vs. writing a function by hand?
2. Read the source code for `plyr::colwise()`. How does the code work? What are `colwise()`'s three main tasks? How could you make `colwise()` simpler by implementing each task as a function operator? (Hint: think about `partial()`.)
3. Write FOs that convert a function to return a matrix instead of a data frame, or a data frame instead of a matrix. If you understand S3, call them `as.data.frame.function()` and `as.matrix.function()`.
4. You've seen five functions that modify a function to change its output from one form to another. What are they? Draw a table of the various combinations of types of outputs: what should go in the rows and what should go in the columns? What function operators might you want to write to fill in the missing cells? Come up with example use cases.
5. Look at all the examples of using an anonymous function to partially apply a function in this and the previous chapter. Replace the anonymous function with `partial()`. What do you think of the result? Is it easier or harder to read?

12.5 Combining FOs

Besides just operating on single functions, function operators can take multiple functions as input. One simple example of this is `plyr::each()`. It takes a list of vectorised functions and combines them into a single function.

```
summaries <- plyr::each(mean, sd, median)
summaries(1:10)
#>   mean      sd median
#>  5.50    3.03  5.50
```

Two more complicated examples are combining functions through composition, or through boolean algebra. These capabilities are the glue that allow us to join multiple functions together.

12.5.1 Function composition

An important way of combining functions is through composition: `f(g(x))`. Composition takes a list of functions and applies them sequentially to the input. It's a replacement for the common pattern of anonymous function that chains multiple functions together to get the result you want:

```
sapply(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
```

A simple version of compose looks like this:

```
compose <- function(f, g) {
  function(...) f(g(...))
}
```

(`pryr::compose()` provides a more full-featured alternative that can accept multiple functions and is used for the rest of the examples.)

This allows us to write:

```
sapply(mtcars, compose(length, unique))
#>   mpg cyl disp  hp drat    wt  qsec vs am gear carb
#> 25   3 27 22 22 29 30 2 2 3 6
```

Mathematically, function composition is often denoted with the infix operator, \circ , $(f \circ g)(x)$. Haskell, a popular functional programming language, uses $.$ to the same end. In R, we can create our own infix composition function:

```
"%o%" <- compose
sapply(mtcars, length %o% unique)
#>   mpg cyl disp  hp drat    wt  qsec vs am gear carb
#> 25   3 27 22 22 29 30 2 2 3 6

sqrt(1 + 8)
#> [1] 3
compose(sqrt, `+`)(1, 8)
#> [1] 3
(sqrt %o% `+`)(1, 8)
#> [1] 3
```

Compose also allows for a very succinct implementation of Negate, which is just a partially evaluated version of compose().

```
Negate <- partial(compose, `!`)
```

We could implement the population standard deviation with function composition:

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

sd2 <- sqrt %o% mean %o% square %o% deviation
sd2(1:10)
#> [1] 2.87
```

This type of programming is called tacit or point-free programming. (The term point-free comes from the use of “point” to refer to values in topology; this style is also derogatorily known as pointless). 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, not on objects it’s being done to. Since we’re using only functions and not parameters, we use verbs and not nouns. This style is common in Haskell, and is the typical style in stack based programming languages like Forth and Factor. It’s not a terribly natural or elegant style in R, but it is fun to play with.

compose() is particularly useful in conjunction with partial(), because partial() allows you to supply additional arguments to the functions being composed. One nice side effect of this style of programming is that it keeps a function’s arguments near its name. This is important because as the size of the chunk of code you have to hold in your head grows code becomes harder to understand.

Below I take the example from the first section of the chapter and modify it to use the two styles of function composition described above. Both results are longer than the original code, but they may be easier to understand because the function and its arguments are closer together. Note that we still have to read them from right to left (bottom to top): the first function called is the last one written. We could define compose() to work in the opposite direction, but in the long run, this is likely to lead to confusion since we’d create a small part of the language that reads differently from every other part.

```
download <- dot_every(10, memoise(delay_by(1, download_file)))

download <- pryr::compose(
  partial(dot_every, 10),
```

```

memoise,
partial(delay_by, 1)
)(download_file)

download <- (partial(dot_every, 10) %o%
  memoise %o%
  partial(delay_by, 1))(download_file)

```

12.5.2 Logical predicates and boolean algebra

When I use `Filter()` and other functionals that work with logical predicates, I often find myself using anonymous functions to combine multiple conditions:

```
Filter(function(x) is.character(x) || is.factor(x), iris)
```

As an alternative, we could define function operators that combine logical predicates:

```

and <- function(f1, f2) {
  force(f1); force(f2)
  function(...) {
    f1(...) && f2(...)
  }
}

or <- function(f1, f2) {
  force(f1); force(f2)
  function(...) {
    f1(...) || f2(...)
  }
}

not <- function(f) {
  force(f)
  function(...) {
    !f(...)
  }
}

```

This would allow us to write:

```
Filter(or(is.character, is.factor), iris)
Filter(not(is.numeric), iris)
```

And we now have a boolean algebra on functions, not on the results of functions.

12.5.3 Exercises

1. Implement your own version of `compose()` using `Reduce` and `%o%`. For bonus points, do it without calling `function`.
2. Extend `and()` and `or()` to deal with any number of input functions. Can you do it with `Reduce()`? Can you keep them lazy (e.g., for `and()`, the function returns once it sees the first `FALSE`)?
3. Implement the `xor()` binary operator. Implement it using the existing `xor()` function. Implement it as a combination of `and()` and `or()`. What are the advantages and disadvantages of each approach?

Also think about what you'll call the resulting function to avoid a clash with the existing `xor()` function, and how you might change the names of `and()`, `not()`, and `or()` to keep them consistent.

4. Above, we implemented boolean algebra for functions that return a logical function. Implement elementary algebra (`plus()`, `minus()`, `multiply()`, `divide()`, `exponentiate()`, `log()`) for functions that return numeric vectors.

Part III

Object oriented programming

Chapter 13

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 between. In this book, I'll focus on the three that are most important in my opinion: 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, for day-to-day use of R, FP is much more important than OOP. There are three main reasons to learn OOP:

- Learning a little S3 allows your functions to return richer results that have a user friendly display and programmer friendly internals. It also defines syntactic standards can apply across multiple packages. This is why S3 is used throughout base R.
- Investing in S4 can be helpful for building up large systems that evolve over many years and are written by many programmers. This is why the Bioconductor project uses S4 as fundamental infrastructure.
- Mastering R6 gives you a standard way to escape R's copy-on-modify semantics when needed. 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.

13.1 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.

13.2 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 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 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.

13.3 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”).

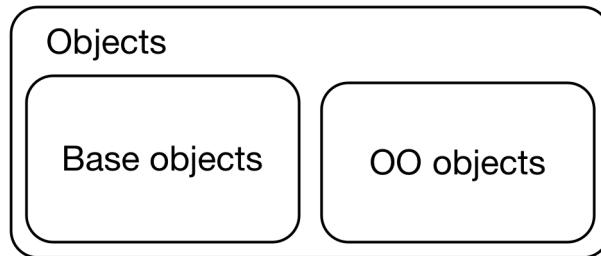
Chapter 14

Base types

14.1 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

14.2 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.

14.3 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.

14.4 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 of 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 1 or 3+.
- 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.

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. it is a number and behaves like a number).

```
is.numeric(1)
#> [1] TRUE
is.numeric(1L)
#> [1] TRUE
is.numeric(factor("x"))
#> [1] FALSE
```

Chapter 15

S3

15.1 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)
```

15.2 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 <- strptime("2017-01-01", "%Y-%m-%d")
str(time)
#> POSIXlt[1:1], format: "2017-01-01"
str(unclass(time), list.len = 5)
#> List of 11
#> $ sec    : num 0
#> $ min    : int 0
#> $ hour   : int 0
#> $ mday   : int 1
#> $ mon    : int 0
#> [list output truncated]
```

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, ...)
```

```
#> structure(mean(unclass(x), ...), class = "Date")
#> <bytecode: 0x557c68bda3a0>
#> <environment: namespace:base>

# Always works
getS3method("mean", "Date")
#> function (x, ...)
#> structure(mean(unclass(x), ...), class = "Date")
#> <bytecode: 0x557c68bda3a0>
#> <environment: namespace:base>
```

15.2.1 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
```

15.3 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        fitted.values
#> [6] assign        qr          df.residual  xlevels    call
#> [11] 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.

15.3.1 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"
```

15.3.2 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(
      "The number of levels must be at least as large as the maximum value",
      call. = FALSE
    )
  }
}
```

```

    "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.

15.3.3 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/ctime()`).

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 = tzzone)
}

```

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()`.

15.3.4 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.

15.3.5 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?

15.4 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.

15.4.1 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

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
```

15.4.2 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.

15.4.3 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

```

15.5 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.

15.5.1 `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 and 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.

15.5.2 `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.

15.5.3 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"

```

15.6 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
```

15.6.1 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, 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, ..., subclass = c(subclass, "subclass"))
}

```

If your subclass is more complicated, you'd also provide validator and helper functions, as described previously.

15.6.2 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"))
}

```

15.6.3 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.

15.6.4 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?

15.7 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.

15.7.1 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.

15.7.2 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
```

15.7.3 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
#> * [

```

15.7.4 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 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.

15.7.5 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 calls the first 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 and 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

```

15.7.6 Exercises

1. `Math.difftime()` is more complicated than I described. Why?

Chapter 16

S4

16.1 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.

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, 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.
- S4 questions on stackoverflow answered by Martin Morgan.
- Software for Data Analysis, 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)
```

16.2 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
```

16.2.1 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.

16.2.2 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  
.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: 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.

16.2.3 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`.

16.2.4 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?

16.3 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()`.

```
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.)

16.3.1 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.

16.3.2 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"
```

16.3.3 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 and 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

```

16.3.4 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>    <lgsl>  <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>    <lgsl>  <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: 0x55a8389b8588>
#>
#> 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)
```

16.3.5 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.

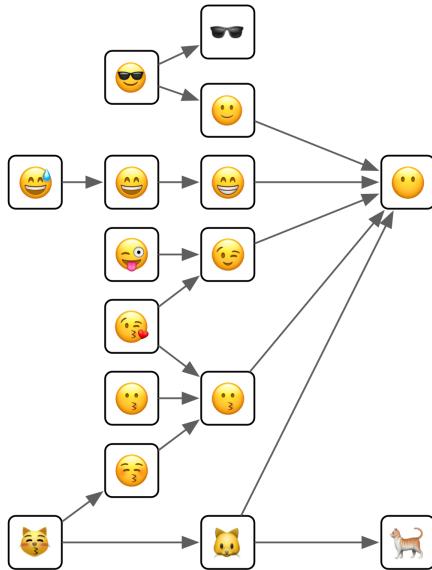
16.4 Method dispatch

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.

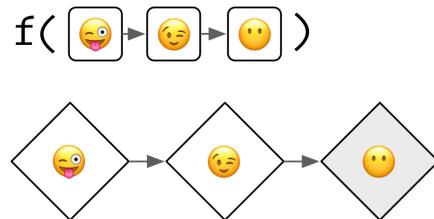
To explain method dispatch, we'll start simple with single inheritance and single dispatch, and work our way up to the more complicated cases. To illustrate the ideas without getting bogged down in the details, we'll use an imaginary **class graph** based on emoji:



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 ☻

16.4.1 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.



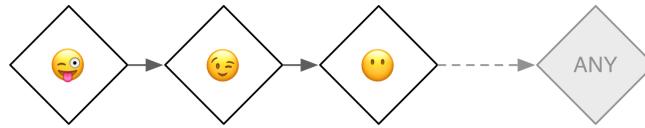
There are two parts to this diagram:

- 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 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

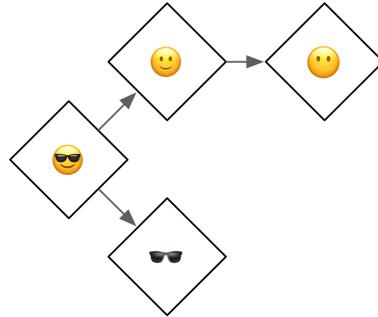
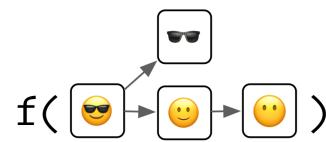
matches any class, and plays the same role as the `default` pseudo-class in S3. For technical reasons that we'll get to later, the link to the "ANY" method is longer than the links between the other classes:



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.

16.4.2 Multiple inheritance

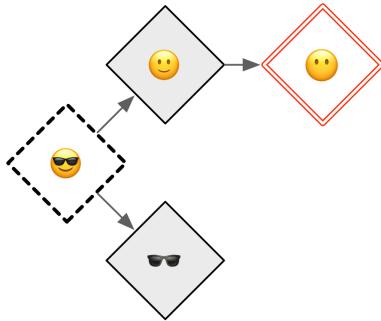
Things get more complicated when the class has multiple parents.



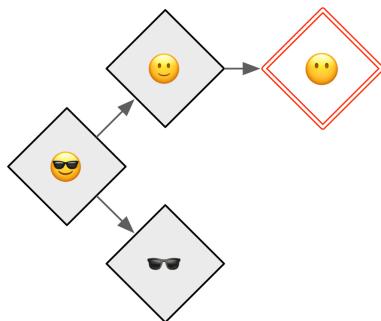
The basic process remains the same: you start from the actual class supplied to the generic, then follow the arrows until you find a defined method. The wrinkle is now that there are multiple arrows to follow, so you might find multiple methods. If that happens, you pick the method that is closest, i.e. 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](#))

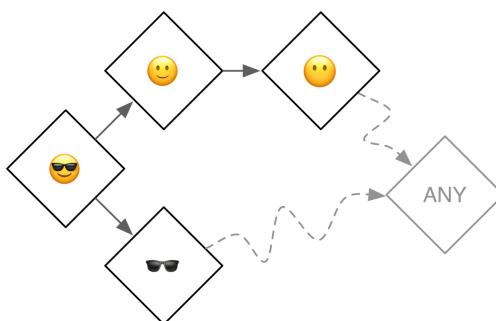
What happens if methods are the same distance? For example, imagine we've defined methods for `⊤` and `⊥`, and we call the generic with `⊤`. Note that there's no implementation for the `⊥` class, as indicated by the red double outline.



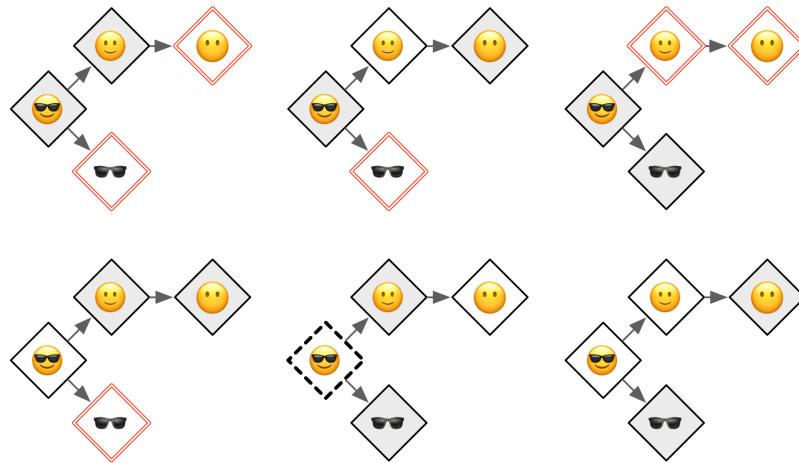
This is called an **ambiguous** method, and in diagrams I'll illustrate it with a thick dotted border. When this happens in R, you'll get a warning, and one of the two methods is basically picked at random (it uses the method that comes first in the alphabet). When you discover ambiguity you should always 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.



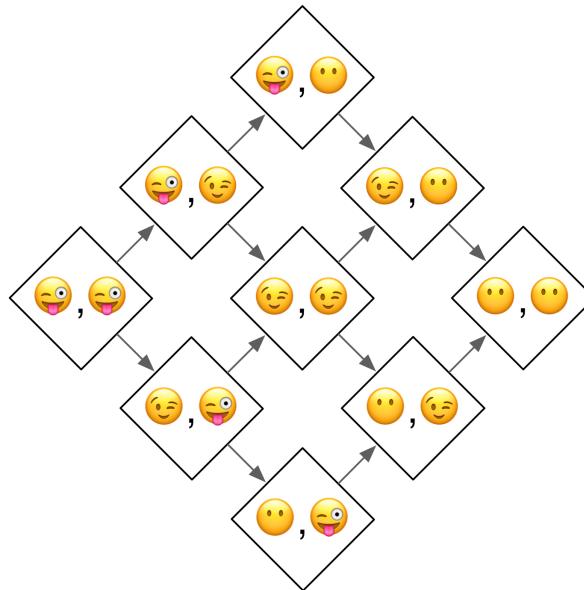
It is hard to simultaneously prevent ambiguity, ensure that every terminal method has an implementation, and minimise the number of defined methods (in order to benefit from OOP). For example, of the six ways to define only two methods for this call, only one is free from problems. For this reason, I recommend using multiple inheritance with extreme care: you will need to carefully think about the method graph and plan accordingly.



16.4.3 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).

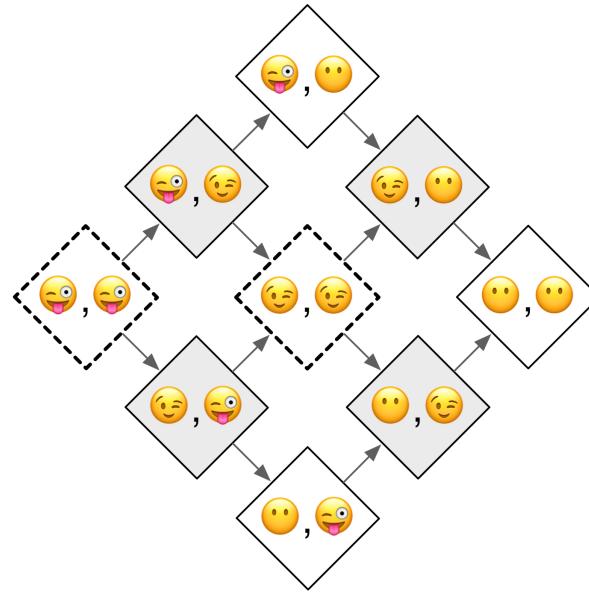
$f(\text{ } \square \rightarrow \text{ } \square \rightarrow \text{ } \square , \text{ } \square \rightarrow \text{ } \square \rightarrow \text{ } \square)$



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([]->[]->[], []->[]->[])`

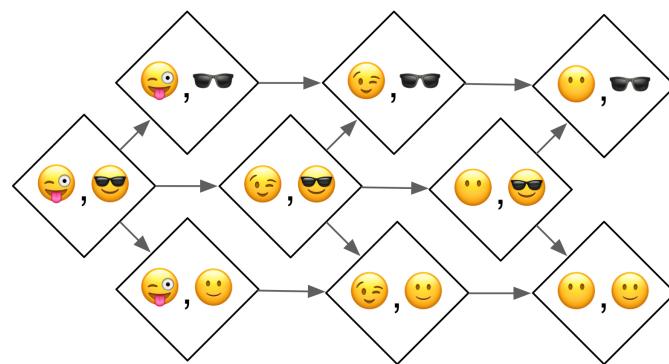


Multiple dispatch tends to be less tricky to work with than multiple inheritance because there are usually fewer terminal class combinations. In this example, there's only one. That means, at a minimum, you can define a single method and have default behaviour for all inputs.

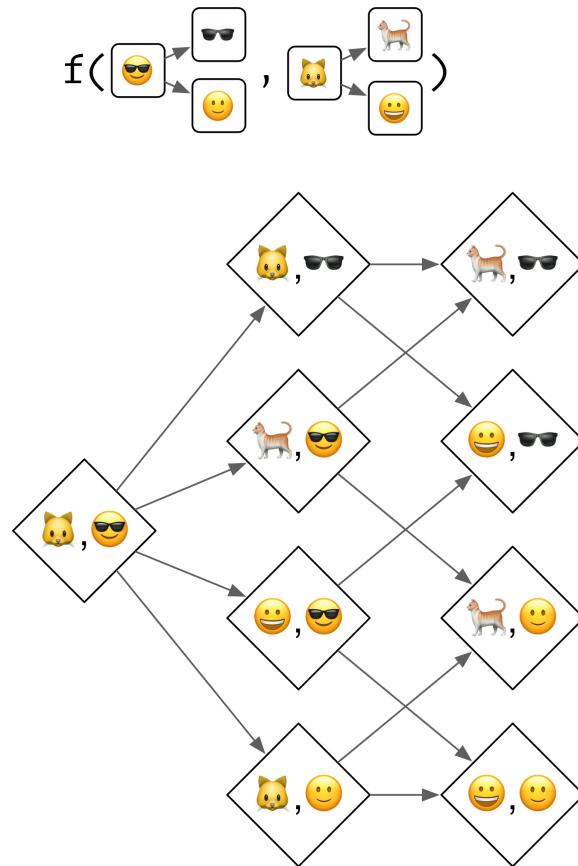
16.4.4 Multiple dispatch and multiple inheritance

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.

16.4.5 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?

16.5 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.

16.5.1 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`.

16.5.2 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).

16.5.3 Exercises

Chapter 17

R6

17.1 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")`.

17.2 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()`.

17.2.1 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 message-chaining tradeoffs.

17.2.2 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

```

17.2.3 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
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!

```

17.2.4 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.

17.2.5 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"           "clone"
#> [5] "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.

17.2.6 Exercises

1. Can subclasses access private fields/methods from their parent? Perform an experiment to find out.

17.3 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.

17.3.1 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 more important in other programming languages compared to R because the object hierarchies in R tend to be simpler.

17.3.2 Active fields

Active fields make 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: is.character(value) is not TRUE

hadley5$age
#> [1] NA
hadley5$age <- 20
```

```
#> Error: `\$age` is read only
```

17.3.3 Exercises

- How would you define a write-only field?

17.4 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.

17.4.1 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].

17.4.2 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 finalise 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())
```

17.4.3 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
```

17.4.4 Exercises

Chapter 18

Trade-offs

18.1 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 idiosyncrasies 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.

18.2 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 use S4, but

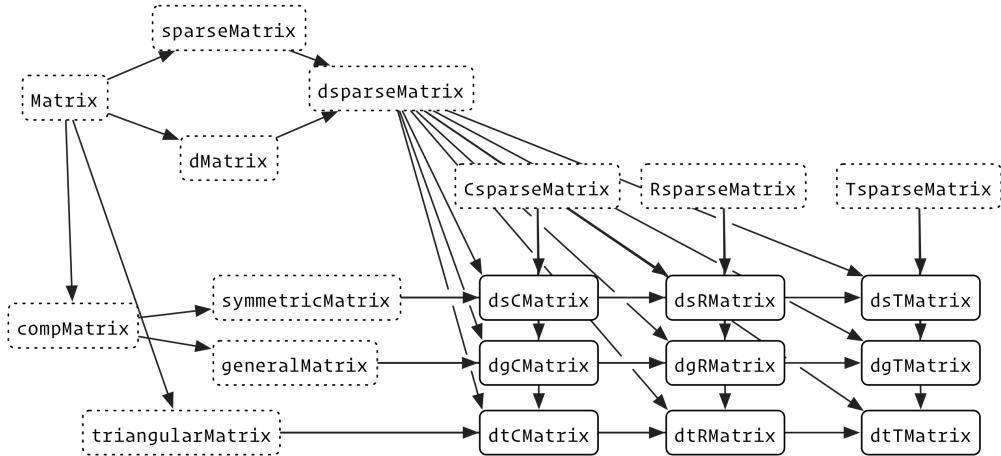


Figure 18.1: A small subset of the `Matrix` class graph showing the inheritance of sparse matrices. Each concrete class inherits from two virtual parents: one that describes how the data is stored (C = column oriented, R = row oriented, T = tagged) and one that describes any restriction on the matrix (s = symmetric, t = triangle, g = general)

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 18.1.

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.)

18.3 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.)

18.3.1 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.

18.3.2 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
```

18.3.3 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

Chapter 19

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 evaluation**, 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.

19.0.1 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 1.00 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 1.20 Premium    D    VVS1      62.1  59. 15686    0.    0.    0.
#> 5 2.25 Premium    H    SI2      62.8  59. 18034    0.    0.    0.
#> 6 0.710 Good      F    SI2      64.1  60. 2130    0.    0.    0.
#> 7 0.710 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, ]
```

19.1 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 by Martin Fowler. It discusses many options for creating a DSL and provides many examples of different languages.

19.2 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 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 20

Expressions

20.1 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

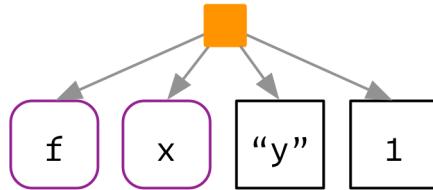
Prerequisites

Make sure you've installed `rlang` and `lobstr` from GitHub:

```
devtools::install_github("r-lib/rlang")
devtools::install_github("hadley/lobstr")
```

20.2 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))`.

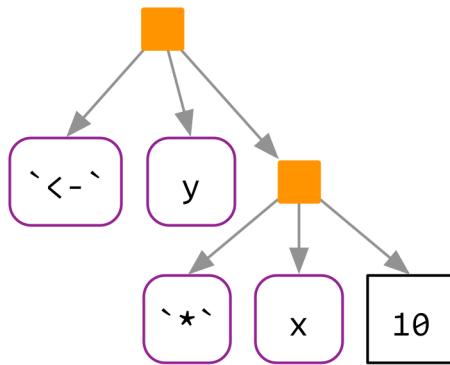
20.2.1 Infix vs. prefix calls

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's 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.¹

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
```

20.2.2 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:

¹Some 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` `{
#>     `x` `y` `{
#>       `x` `y` `<inline srcref>

```

Note that functions include a node `<inline srcref>`, this contains the source reference for the function, as mentioned in function components.

20.2.3 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.

20.2.4 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: \$)

20.2.5 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?

20.3 R's grammar

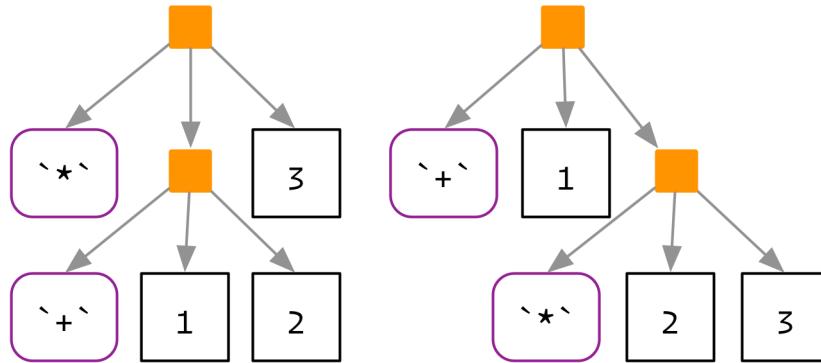
The process by which a computer language takes a sequence of tokens (like `x`, `+`, `y`) and constructs a tree is called **parsing**, and it is governed by a set of rules known as a **grammar**. In this section, we'll use `lobstr::ast()` to explore some of the details of R's grammar.

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.

20.3.1 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 operators². 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?

²These 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
```

20.3.2 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
```

20.3.3 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
```

20.3.4 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?

20.4 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.

20.4.1 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 . 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).

20.4.2 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
```

20.4.3 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.

20.4.4 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::lang_standardise()` which standardises all arguments to use the full name:

```
rlang::lang_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)
```

20.4.5 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 cld
#>  l1[[1]] 158 194 335    208 230 7983    100  a
#>  l1[[100]] 161 196 304    216 244 7991    100  a
#>  l2[[1]] 1017 1139 1251   1210 1310 2427    100  b
#>  l2[[100]] 1113 1218 1576   1330 1480 11754   100   c
```

20.4.6 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.

20.4.7 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?

20.5 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.

20.5.1 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?

20.6 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.

20.6.1 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

```

20.6.2 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 purrrr, 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"  "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 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.

20.6.3 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 21

Quasiquotation

21.1 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**, 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)  
#>  
#> Attaching package: 'rlang'  
#> The following objects are masked from 'package:purrr':  
#>  
#>     %%, %/%, as_function, flatten, flatten_chr, flatten_dbl,
```

```
#>     flatten_int, flatten_lgl, invoke, list_along, modify, prepend,
#>     rep_along, splice
```

21.2 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)
```

21.2.1 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.

21.2.2 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 Quine¹, 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.

21.2.3 Exercises

1. For each function in the following base R code, identify which arguments are quoted and which are evaluated.

¹You might be familiar with the name Quine from “quines”, computer programs that when run return a copy of their own source code.

```
library(MASS)

mtcars2 <- subset(mtcars, cyl == 4)

with(mtcars2, sum(vs))
sum(mtcars2$am)

rm(mtcars2)
```

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()
```

21.3 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.

21.3.1 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 users' expression:

| | Developer | User |
|------|-----------|-----------|
| One | expr() | enexpr() |
| Many | exprs() | enexprs() |

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 an 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: using `purrr` to work with list 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.

21.3.2 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 a quoting function, not a quasiquotting 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 quasiquotation, 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].

21.3.3 Exercises

1. What happens if you try and use `enexpr()` with an expression? What happens if you try and use `enexpr()` with a missing argument?
2. Compare and contrast the following two functions. Can you predict the ouput 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)
```

```
#> $x
#> x
#>
#> $y
#> y
f2(a + b, c + d)
#> $x
#> a + b
#>
#> $y
#> 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.

21.4 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 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 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
```

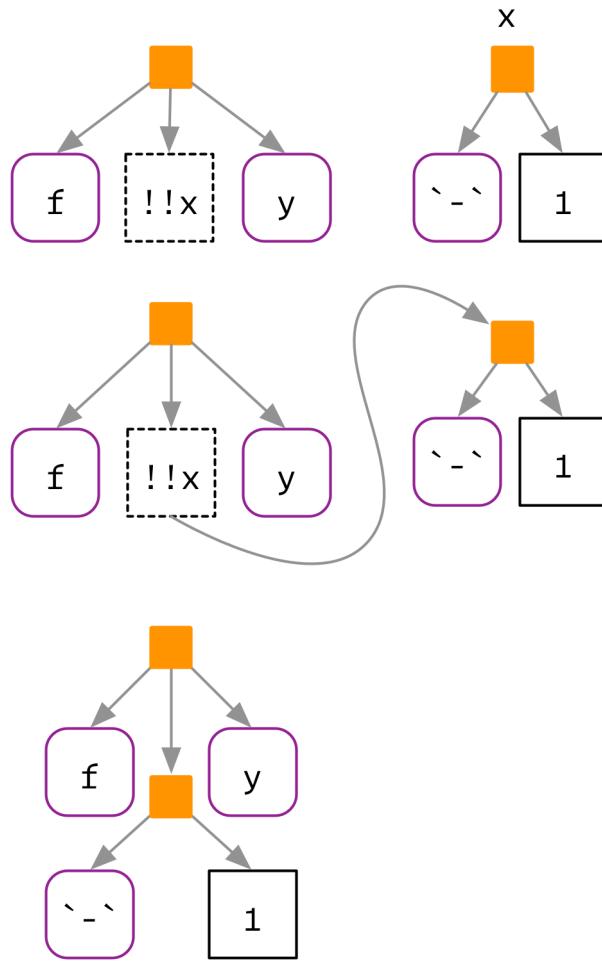
21.5 Unquotation

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.

21.5.1 With rlang

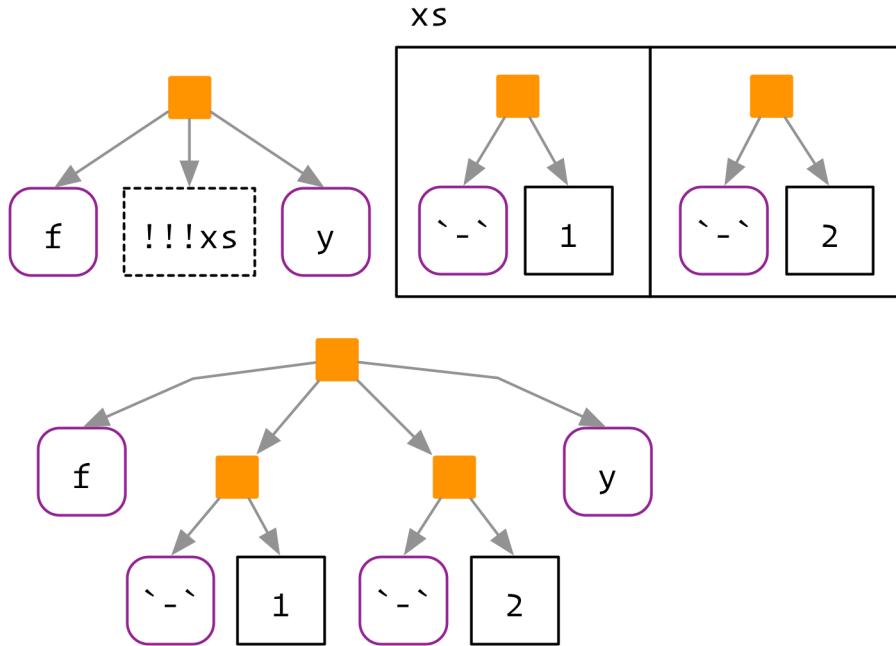
All quoting functions in rlang (`expr()`, `enexpr()`, and friends) supporting unquoting with `!!` (called “unquote”, and pronounced bang-bang) and `!!!` (called “unquote-splice”, and pronounced bang-bang-bang). They both replace nodes in the AST. `!!` is a one-to-one replacement. It takes a single expression 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)
```



21.5.2 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)`.)

One place where the illusion currently breaks down is `base::deparse()`:

```
x <- quote(!!x + !!y)
deparse(x)
#> [1] "!(!x + !(!y))"
```

Although the R parser can distinguish between `!(x)` and `!x`, the deparser currently does not. You are most

likely to see this when printing the source for a function in another package, where the source references have been lost. `rlang::expr_deparse()` works around this problem if you need to manually deparse an expression, but often this does not help because the deparsing occurs outside of your control, as during debugging.

```
expr_deparse(x)
#> [1] "!!x + (!!y)"
```

Hopefully this will be resolved in a future version of R, but for now, you'll need to watch out for this problem.

21.5.3 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)

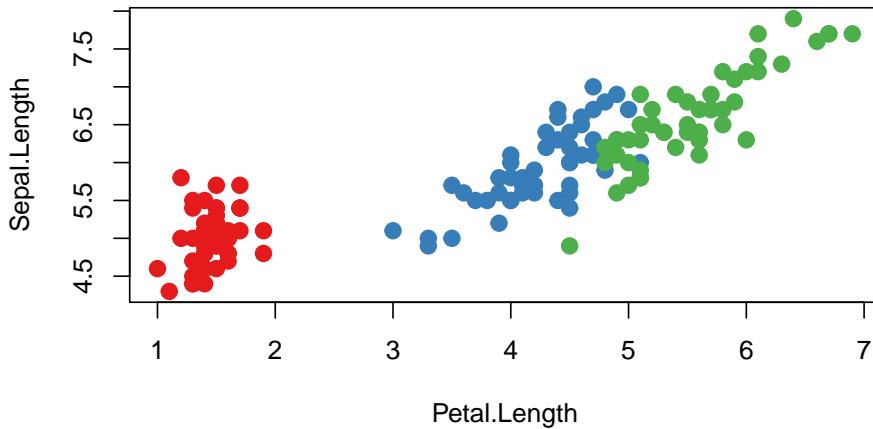
var <- 10
# Shows help for var
help(var)
```

`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`.

21.5.4 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 deparsed, 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()`!

21.5.5 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(, )
```

21.5.6 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)
```

2. Explain why both `!0 + !0` and `!1 + !1` return FALSE while `!0 + !1` returns TRUE.
3. 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.
4. 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?

21.6 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)
```

21.6.1 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 turn 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)
#> }
```

21.6.2 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 `...`.

21.6.3 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.

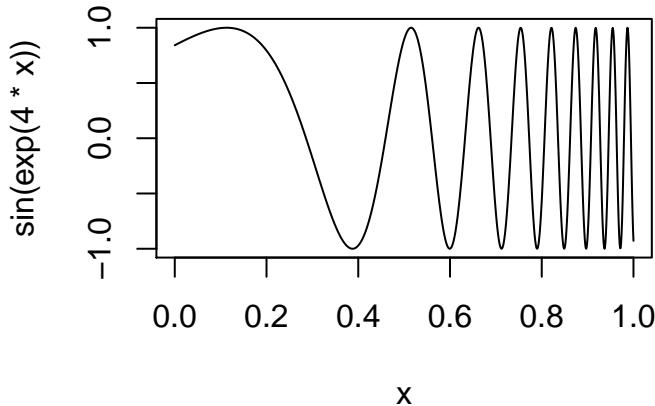
21.6.4 Creating functions

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 functions, `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** functions².

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: 0x55b5ffddaa90>

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 complex objects into the AST can yield confusing source code.

²Anaphoric comes from the linguistics term “anaphora”, an expression that is context dependent. Anaphoric functions are found in Arc (a LISP like language), Perl, and Clojure.

21.6.5 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(...))
}
```

21.7 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(dfsa, dfb)`, 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 (ie. `setNames(data.frame(val), var)`), but this feels inelegant. How can we do better?

21.7.1 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
#> 1 2
#> 2 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
```

21.7.2 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 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.

21.7.3 `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()
#> atomic [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 column 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.

21.7.4 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 `:=`.

21.7.5 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 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 particular 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

21.7.6 Exercises

1. Carefully read the source code for `interaction()`, `expand.grid()`, and `par()`. Compare and construct 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 22

Evaluation

22.1 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** then 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 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)
#>
#> Attaching package: 'rlang'
#> The following objects are masked from 'package:purrr':
#>
```

```
#>     %0%, %/1%, as_function, flatten, flatten_chr, flatten_dbl,
#>     flatten_int, flatten_lgl, invoke, list_along, modify, prepend,
#>     rep_along, splice
```

22.2 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 = env)
#> [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()`.

22.2.1 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.

22.2.2 Application: source()

We can create a simple version of `source()` by combining `expr_text()` 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.

22.2.3 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
```

22.2.4 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()
#>
#> +-local(...)
#> / \ -eval.parent(substitute(eval(quote(expr), envir)))
#> /   \ -eval(expr, p)
#> /     \ -eval(expr, p)
```

```

#> +-eval(...)
#> / \-eval(...)
#> /   +-do.call(...)
#> /     \-(function (input, output_format = NULL, output_file = NULL, output_dir = NULL, ...
#> /       \-knitr::knit(...)
#> /         \-process_file(text, output)
#> /           +-withCallingHandlers(...)
#> /             +-process_group(group)
#> /               \-process_group.block(group)
#> /                 \-call_block(x)
#> /                   \-block_exec(params)
#> /                     +-in_dir(...)
#> /                       \-evaluate(...)
#> /                         \-evaluate::evaluate(...)
#> /                           \-evaluate_call(...)
#> /                             +-timing_fn(...)
#> /                               +-handle(...)
#> /                                 +-withCallingHandlers(...)
#> /                                   +-withVisible(expr, envir, enclos))
#> /                                     \-eval(expr, envir, enclos)
#> /                                       \-eval(expr, envir, enclos)
#> \-f()
#>   +-g()
#>   / \-h()
#>   /   \-eval(expr(lobstr::cst()), caller_env(2))
#>   /     \-eval(expr(lobstr::cst()), caller_env(2))
#>   \-lobstr::cst()
#'
#'   f()
#'   g()
#'   h()
#'     eval(expr(lobstr::cst()), caller_env(2))
#'     eval(expr(lobstr::cst()), caller_env(2))
#'   lobstr::cst()

```

22.2.5 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.

22.2.6 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.)

22.3 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.

22.3.1 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.

22.3.2 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:  0x5588391aa408
```

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)"
```

22.3.3 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: 0x558839acec90>
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
```

22.3.4 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 reuses
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.

22.3.5 When not to use quosures

Almost all quoting functions should capture quosures 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 quosures. 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 quosure:

```
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)

22.3.6 Exercises

- Predict what evaluating each of the following quosures will return.

```
q1 <- new_quosure(expr(x), env(x = 1))
q1
#> <quosure>
#>   expr: ^x
#>   env: 0x558839f8d690

q2 <- new_quosure(expr(x + !!q1), env(x = 10))
q2
#> <quosure>
#>   expr: ^x + (^x)
#>   env: 0x558839573f08

q3 <- new_quosure(expr(x + !!q2), env(x = 100))
q3
#> <quosure>
#>   expr: ^x + (^x + (^x))
#>   env: 0x5588390bd420
```

- Write a function `enenv()` that captures the environment associated with an argument.

22.4 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.

22.4.1 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`.¹ This lets

¹`eval_tidy()` has a `env` argument, but you only need this if you pass an expression to the first argument.

you set up a **data mask**, where variables in the environment are potentially masked by variables in 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_env(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
```

22.4.2 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 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
```

22.4.3 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
```

22.4.4 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 `$`, because `df$!!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.

22.4.5 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!

22.4.6 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 cld
#>    runif(n) 26.9 27.4 30.5  27.9 28.3  81.1   100   a
#> eval_bare(x1, e1) 27.5 28.4 31.0  28.7 29.2 101.4   100   a
#> eval_tidy(q1) 29.6 30.4 31.9  30.8 31.7  61.3   100   a
#> eval_tidy(q1, mtcars) 32.3 33.4 41.2  33.9 34.8 512.8   100   b
```

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 cld
#>    as_data_mask(mtcars) 4.91  5.68 12.0   6.26  7.83 199   100   a
#> eval_tidy(q1, mtcars) 32.86 35.29 44.5  37.08 40.92 144   100   b
#> eval_tidy(q1, d_mtcars) 29.03 30.39 37.5  31.47 33.96 243   100   b
```

22.4.7 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 sorted in descending order using named arguments:

```
arrange(mtcars, cyl, desc = mpg, vs)
```

(Hint: The `descreasing` 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.

22.5 Wrapping quoting functions

Now we have all the tools need to wrap a quoting function inside another function, regardless of the 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 argument; now you'll learn the techniques that allow you to wrap quoted arguments.

22.5.1 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, summar_vary) {
  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, summar_vary) {
  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.

22.5.2 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 function. 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)
```

22.5.3 The evaluation environment

What if you want to mingle object supplied by the user with objects that you create in the function? For example, imagine you want to make an auto-booztrapping 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 way 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, 9L, 7L, 4L, 1L, 2L,
#> 2L, 4L, 5L, 5L), y = c(14.1527498283867, 31.3577210908734, 25.590418059691,
#> 16.2118102327009, 8.48257501673016, 11.0606709535963, 11.0606709535963,
#> 16.2118102327009, 19.9122584164126, 19.9122584164126)))
```

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
#>      5.71            2.83
```

22.5.4 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 they 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)`

#> $y \sim a + b + c$

22.5.5 Exercises

- When model building, typically the predictor 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?
- 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.
- 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.
- Create a formula generation function that allows you to optionally supply a transformation function (e.g. `log()`) to the response or the predictors.

Chapter 23

Translating R code

23.1 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 maniplation 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

Prequisites

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)
#>
#> Attaching package: 'rlang'
#> The following objects are masked from 'package:purrr':
#>
#>     %@%, %//%, as_function, flatten, flatten_chr, flatten_dbl,
#>     flatten_int, flatten_lgl, invoke, list_along, modify, prepend,
```

```
#>     rep_along, splice
library(purrr)
```

23.2 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 &lt; b >some bold text.</b></p>
  <img src='myimg.png' width='100' height='100' />
</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.
- `` emboldens text.
- `` 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 ``, don't have any content, are written ``, not ``. 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**: `>` and `<`. And, since those escapes use `&`, if you want a literal ampersand you have to escape it with `&`.

23.2.1 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 &lt; b >some bold text.</b></p>
  <img src='myimg.png' width='100' height='100' />
</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 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.

23.2.2 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 & ; &lt; ; and '>').

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.html()` which leaves already escaped HTML as is.

```

escape <- function(x) UseMethod("escape")

escape.character <- function(x) {
  x <- gsub("&", "&amp;", x)
  x <- gsub("<", "&lt;", x)
  x <- gsub(">", "&gt;", 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 &gt; 1 &amp; y &lt; 2

```

```
# Double escaping is not a problem
escape(escape("This is some text. 1 > 2"))
#> <HTML> This is some text. 1 &gt; 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.

23.2.3 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 `` 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. 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 idiosyncrasies 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).

```
source("dsl-html-attributes.r", local = TRUE)
p <- function(...) {
```

```

dots <- dots_partition(...)
attribs <- html_attributes(dots$named)
children <- map_chr(dots$unnamed, escape)

html(paste0(
  "<p", attribs, ">",
  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>

```

23.2.4 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(...)
      attribs <- html_attributes(dots$named)
      children <- map_chr(dots$unnamed, escape)

      html(paste0(
        !!paste0("<", tag), attribs, ">",
        paste(children, collapse = ""),
        !!paste0("</", tag, ">")
      ))
    }),
    caller_env()
  )
}

tag("b")
#> function (...)

#> {
#>   dots <- dots_partition(...)
#>   attribs <- html_attributes(dots$named)
#>   children <- map_chr(dots$unnamed, escape)
#>   html(paste0("<b", attribs, ">", paste(children, collapse = ""), 
#>             "</b>"))
#> }

```

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. <b><i>some bold italic
#> text</i></b></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> <img src='myimage.png' width='100' height='100' />
```

23.2.5 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. <b><i>some bold italic
#> text</i></b></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
#> &lt;b>some bold text.</b></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.

23.2.6 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))
```

23.3 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 `objectionable` 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 for function calls vs. other symbols.

Before we begin, let's quickly cover how formulas are expressed in LaTeX.

23.3.1 LaTeX mathematics

The full spectrum of LaTeX mathematical notation is complex. Fortunately, they are well documented, 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` = π , `\pm` = \pm , and so on. There are a huge number of symbols available in LaTeX. Googling for `latex math symbols` will return many lists. There's even a service 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 `\text{f}(a * b)`.

23.3.2 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 `\text{f}(a) \rightarrow \text{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.

23.3.3 `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)
}
```

23.3.4 Known symbols

Our first step is to create an environment that will convert the special LaTeX symbols used for Greek, e.g., π to π . 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!

23.3.5 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

```

23.3.6 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, "}")
  },

  # Labelling
  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)

```

23.3.7 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: 0x5565041653d0>

```

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)
```

23.3.8 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

Performance

Chapter 24

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.

24.1 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. 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, 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++ and javascript, 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.

24.2 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` 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 cld
#>  sqrt(x)  588  623  892   762  850 11,000   100    a
#>  x ^ 0.5 6,490 6,640 7574  6,760 6,880 40,200   100    b
```

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 800 ns, 800 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 μs , then one million calls takes a second
- 1 ns, then one billion calls takes a second

The `sqrt()` function takes about 800 ns, or 0.8 μs , to compute the square root of 100 numbers. That means if you repeated the operation a million times, it would take 0.8 s. So changing the way you compute the square root is unlikely to significantly affect real code.

24.2.1 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?

24.3 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" 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.

24.3.1 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](#).) 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 cld
#>   fun    233    265   322    300    338   2,020   100    a
#>   S3   870 1,130  8098   1,250   1,410 658,000   100    a
#>   S4  7,560 9,090 18579   9,600 10,300 707,000   100    a
#>   RC  5,450 5,950 10284   6,440  6,880 361,000   100    a
```

The bare function takes about 300 ns. S3 method dispatch takes an additional 1,000 ns; S4 dispatch, 9,000 ns; and RC dispatch, 6,000 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.

24.3.2 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 cld
#>      f(1, 2) 363 420   821    495 579 1,700,000 10000    a
#>     f2(1, 2) 604 677   978    770 872   43,600 10000    a
#>     f3(1, 2) 633 705   986    800 903   48,800 10000    a
#>     f4(1, 2) 667 737  1069    831 939   48,400 10000    a
```

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 `(`.

24.3.3 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 cld
#>   f0() 159 206 439    221 276 468,000 10000    a
#>   f1() 210 251 516    267 334 700,000 10000   ab
#>   f2() 228 270 539    327 398 438,000 10000   ab
#>   f3() 240 294 586    366 454 430,000 10000   ab
#>   f4() 262 319 618    415 506 382,000 10000   ab
#>   f5() 289 350 685    464 560 447,000 10000     b
```

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.

24.3.4 Exercises

- `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 = T)`) what proportion of the total run time is due to creating those promises?
- Read “Evaluating the Design of the R Language”. 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?

24.4 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, 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.

24.4.1 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 cld
#>      [32, 11] 9,510 10,400 12487 10,700 11,300 58,900   100    b
#>      $carb[32] 4,960  5,980 11910  6,370  6,750 488,000   100    b
#> [[c(11, 32)]] 4,150  4,690  5728  5,240  5,670 34,600   100   ab
#> [[11]][32]    4,120  4,660  5766  4,920  5,240 45,900   100   ab
#> .subset2     313    433   624    464    524 10,500   100    a
```

24.4.2 `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 cld
#>    squish_ife 18.70 20.10 47.0  20.80 22.7 2,310   100    a
#>    squish_p   10.60 11.50 31.3  12.30 12.9 1,340   100    a
#>  squish_in_place  1.82  2.25 23.4   2.44  2.6 2,090   100    a

```

Using `pmin()` and `pmax()` is about 2x faster than `ifelse()`, and using subsetting directly is about 5x 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 cld
#>  squish_in_place 1.88 2.52  2.94   2.76 3.04   10.9   100    a
#>  squish_cpp     1.88 2.07 14.64   2.20 2.38 1,170.0   100    a

```

The C++ implementation is around 1x faster than the best pure R implementation.

24.4.3 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.

24.5 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 (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 by BeDataDriven. Renjin uses the Java virtual machine, and has an extensive test suite.
- FastR by a team from Purdue. FastR is similar to Renjin, but it makes more ambitious optimisations and is somewhat less mature.
- 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](#).

These are roughly ordered from most practical to most ambitious. Another project, 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 cld
#> cond_sum_cpp 3.88  4.08  4.39  4.17  4.35  7.8   100    a
#> cond_sum_r  9.93 10.70 13.47 12.30 13.30 108.0  100    b
```

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 25

Optimising code

25.1 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")
```

25.2 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 `screfs` to match up the code to the profile, and the needed `screfs` 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(1)` 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.

25.2.1 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.

25.3 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` 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.

25.3.1 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)
}
```

25.4 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.

25.5 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 cld
#> mean1(x) 3,010 3,150 12906  3,250 3,470  920,000    100    a
#> mean2(x)   649   778 18065     834   935 1,700,000    100    a
```

(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>

25.6 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. 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. 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. 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.

25.6.1 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()`?

25.7 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 and stackoverflow.

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()` quite a bit faster than calling `mean()` for small vectors:

```
x <- runif(1e2)

microbenchmark(
  mean(x),
  mean.default(x)
)
#> Unit: microseconds
#>          expr   min    lq    mean   median    uq    max neval cld
#>        mean(x) 2.71 2.90  4.09   3.05 3.39 41.4    100    b
#> mean.default(x) 1.23 1.37  1.99   1.48 1.66 23.6    100    a
```

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 cld
#>   quick_df     7.27    8.34   38.1     14   14.8  2,480    100     a
#> as.data.frame 1,130.00 1,160.00 1222.4   1,190  1,230.0  2,830    100     b
```

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 `difference = 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 cld
#> diff1(x) 3.88 4.15 184.35  4.31 4.67 18,000.0   100   a
#> diff2(x) 3.29 3.59  4.66  3.76 4.08     35.1   100   a
#> diff3(x) 2.74 3.07 78.50  3.24 3.50  7,490.0   100   a
#> diff4(x) 2.08 2.33 32.88  2.51 2.74   3,030.0   100   a
```

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(sub$x, sub$y)
}

# 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 cld
#> boot_cor1(df, 10) 2.08 2.33 32.88  2.51 2.74   3,030.0   100   a
#> boot_cor2(df, 10) 4.15 4.66 184.35  4.31 4.67 18,000.0   100   a
```

```
#> boot_cor1(df, 10) 76.9 79.1 124.6 81.0 89.3 2,310 100 a
#> boot_cor2(df, 10) 45.6 47.0 80.7 47.9 49.3 2,520 100 a
```

25.7.1 Exercises

1. How do the results change if you compare `mean()` and `mean.default()` on 10,000 observations, rather than on 100?
2. 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
#>  0.060  0.000  0.055
system.time(rowSums2(df))
#>   user  system elapsed
#>  0.040  0.000  0.038
```

3. What's the difference between `rowSums()` and `.rowSums()`?
4. 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.
5. 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?
6. 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?

25.8 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 look up 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 cld
#>  lookup[x1] 526  632 1304    690  778 21,800   100   a
#>  lookup[x10] 1,470 1,630 2161   1,740 1,850 12,100   100   b
#>  lookup[x100] 5,420 5,730 6377   6,130 6,510 18,700   100   c

```

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.

25.8.1 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)`?

25.9 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".

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 cld
#>  loop10 19.80 22.50 55.80 23.50 25.80 2,830.0   100   a
#> loop100 682.00 716.00 768.50 729.00 794.00 1,330.0   100   b
#>  vec10   4.95  5.52  7.05  6.05  6.84   52.1   100   a
#> vec100  35.30 36.70 42.36 38.20 43.20   102.0   100   a

```

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.

25.10 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 cld
#>  lapply2(x, is.null) 1.73 1.85 38.10  1.94 2.07 3,600.0   100   a
#> lapply2_c(x, is.null) 1.74 1.88  2.57  1.98 2.10    26.0   100   a
#>    lapply(x, is.null) 2.26 2.44  3.26  2.62 2.84    33.5   100   a

```

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.

25.11 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" 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)$statistic)
#>   user  system elapsed
#> 0.770  0.000  0.774
system.time(
  for(i in 1:m) t.test(X[i, grp == 1], X[i, grp == 2])$statistic
)
#>   user  system elapsed
#> 0.150  0.000  0.158
```

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
#> 0.160  0.000  0.161
```

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
#> 0.030  0.000  0.025
```

```
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
#> 0.010  0.000  0.011
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 cld
#>  rowtstat(X, grp) 0.563 0.602 0.710 0.749 0.790 0.981    100    a
#> rowtstat_bc(X, grp) 0.564 0.598 0.708 0.710 0.778 2.540    100    a
```

In this example, byte code compilation doesn't help at all.

25.12 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 by Q. Ethan McCallum and Stephen Weston.
- Parallel Computing for Data Science 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] 3

pause <- function(i) {
  function(x) Sys.sleep(i)
}

system.time(lapply(1:10, pause(0.25)))
#>    user  system elapsed
#>    0.00    0.00   2.51
system.time(mclapply(1:10, pause(0.25), mc.cores = cores))
#>    user  system elapsed
#>    0.00    0.01   1.02
```

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
#>    0.00    0.01   27.06
```

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.

25.13 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 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* or Patrick Burns' *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 offered on Coursera.
- Read general books about optimisation like *Mature optimisation* by Carlos Bueno, or the *Pragmatic Programmer* 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 26

High performance functions with Rcpp

26.1 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 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, 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.
- On Mac, install Xcode from the app store.
- On Linux, `sudo apt-get install r-base-dev` or similar.

26.2 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: 0x7f485261b1f0>, 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

26.2.1 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 `;`.

26.2.2 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`.

26.2.3 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 cld
#>  sum(x) 1.24  1.28  1.39  1.33  1.38   3.77   100   a
#> sumC(x) 2.14  2.25  9.39  2.38  2.73  666.00   100   a
#> sumR(x) 45.30 45.40 74.41 45.50 45.90 2,690.00   100   b
```

26.2.4 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.

26.2.5 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.

26.2.6 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.

26.2.7 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. Whenever implementing a numerical algorithm, it's always good to check what is already known about the problem.

26.3 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()`.

26.3.1 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

```

26.3.2 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);
```

26.3.3 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.

26.4 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`).

26.4.1 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.

26.4.1.1 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.

26.4.1.2 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
```

26.4.2 Strings

`String` is a scalar string class introduced by Rcpp, so it knows how to deal with missing values.

26.4.3 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`.

26.4.4 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
```

26.4.5 Exercises

- 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()`.

- Rewrite `cumsum()` and `diff()` so they can handle missing values. Note that these functions have slightly more complicated behaviour.

26.5 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

26.5.1 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));
}
```

26.5.2 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 cld
#>  any_naR(x0) 135.0 141.00 229.12 147.00 186.0 1,860.0   100   b
#>  any_naC(x0) 202.0 205.00 222.39 212.00 234.0   379.0   100   b
#>  any_naR(x1) 134.0 141.00 192.40 149.00 178.0 1,620.0   100   b
#>  any_naC(x1) 202.0 205.00 225.54 211.00 221.0 1,040.0   100   b
#>  any_naR(x2)  64.7  73.10 189.86  92.50 170.0 1,480.0   100   b
#>  any_naC(x2)   1.4   1.88   2.93   2.22   3.1    26.3   100   a

```

26.5.3 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`.

26.5.4 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.

26.6 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. 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>`.

26.6.1 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.

26.6.2 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 an 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.

26.6.3 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.

26.6.4 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);
      i++;
      prev = *it;
    }
  }
}
```

```

    lengths.push_back(1);

    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/>.

26.6.5 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/.

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)]]`.

26.6.6 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)]]`.

26.6.7 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`.

26.7 Case studies

The following case studies illustrate some real life uses of C++ to replace slow R code.

26.7.1 Gibbs sampler

The following case study updates an example blogged about 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 cld
#>  gibbs_r(100, 10) 3,710 3,880 4,771  4,140 5,560 13,600   100    a
#>  gibbs_cpp(100, 10) 1,560 1,660 3554  3,100 3,470 94,500   100    a
```

26.7.2 R vectorisation vs. C++ vectorisation

This example is adapted from “Rcpp is smoking fast for agent-based models in data frames”. The challenge is to predict a model response from three inputs. The basic R version of the predictor looks like:

```
vacc1a <- 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] <- vacc1a(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 cld
#>   vacc1 1,580.0 1,640.0 1808.7 1,710.0 1,850.0 3,680    100     c
#>   vacc2   98.4  105.0  190.3  121.0  140.0  6,480    100     b
#>   vacc3   25.0   27.8   39.3   29.8   31.7   807    100     a
```

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.

26.8 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")`.

26.9 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 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 and Dirk’s Rcpp page as well as signing up for the Rcpp mailing list. 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++ and Effective STL by Scott Meyers.
- C++ Annotations, 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, 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, MIT's Introduction to Algorithms, Algorithms by Robert Sedgewick and Kevin Wayne which has a free online textbook and a matching coursera course.

26.10 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.

Bawden, Alan. 1999. “Quasiquotation in Lisp.” In PEPM ’99, 4–12. <http://repository.readscheme.org/ftp/papers/pepm99/bawden.pdf>.

Chambers, John M, and others. 2014. “Object-Oriented Programming, Functional Programming and R.” *Statistical Science* 29 (2). Institute of Mathematical Statistics: 167–80.

Lumley, Thomas. 2001. “Programmer’s Niche: Macros in R.” *R News* 1 (3): 11–13. https://www.r-project.org/doc/Rnews/Rnews_2001-3.pdf.