Real World OCaml

Jason Hickey, Anil Madhavapeddy, and Yaron Minsky

Table of Contents

Preface

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic
> Indicates new terms, URLs, email addresses, filenames, and file extensions.

`Constant width`
> Used for program listings, as well as within paragraphs to refer to program elements such as variable or function names, databases, data types, environment variables, statements, and keywords.

`Constant width bold`
> Shows commands or other text that should be typed literally by the user.

`Constant width italic`
> Shows text that should be replaced with user-supplied values or by values determined by context.

> This icon signifies a tip, suggestion, or general note.
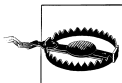
> This icon indicates a warning or caution.

Using Code Examples

This book is here to help you get your job done. In general, you may use the code in this book in your programs and documentation. You do not need to contact us for permission unless you're reproducing a significant portion of the code. For example, writing a program that uses several chunks of code from this book does not require permission. Selling or distributing a CD-ROM of examples from O'Reilly books does

require permission. Answering a question by citing this book and quoting example code does not require permission. Incorporating a significant amount of example code from this book into your product's documentation does require permission.

We appreciate, but do not require, attribution. An attribution usually includes the title, author, publisher, and ISBN. For example: "*Book Title* by Some Author (O'Reilly). Copyright 2012 Some Copyright Holder, 978-0-596-xxxx-x."

If you feel your use of code examples falls outside fair use or the permission given above, feel free to contact us at *permissions@oreilly.com*.

Safari® Books Online

Safari Books Online (*www.safaribooksonline.com*) is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business. Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of product mixes and pricing programs for organizations, government agencies, and individuals. Subscribers have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and dozens more. For more information about Safari Books Online, please visit us online.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

> O'Reilly Media, Inc.
> 1005 Gravenstein Highway North
> Sebastopol, CA 95472
> 800-998-9938 (in the United States or Canada)
> 707-829-0515 (international or local)
> 707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at:

> *http://www.oreilly.com/catalog/<catalog page>*

To comment or ask technical questions about this book, send email to:

bookquestions@oreilly.com

For more information about our books, courses, conferences, and news, see our website at *http://www.oreilly.com*.

Find us on Facebook: *http://facebook.com/oreilly*

Follow us on Twitter: *http://twitter.com/oreillymedia*

Watch us on YouTube: *http://www.youtube.com/oreillymedia*

Prologue

(yminsky: this is something of a placeholder. We need a real introduction that should talk, amongst other things, about what kinds of applications OCaml is good for and why one should want to learn it. Also, some coverage of who uses OCaml successfully now.)

Why OCaml?

Programming languages matter.

The programming languages that you use affect your productivity. They affect how reliable your software is, how efficient it is, how easy it is to read, to refactor, and to extend. And the programming languages you know and use can deeply affect how you think about programming and software design.

But not all ideas about how to design a programming language are created equal. Over the last 40 years, a few key language features have emerged that together form a kind of sweet-spot in language design. These features include:

- Garbage collection
- First-class and higher-order functions
- Static type-checking
- Parametric polymorphism
- Support for programming with immutable values
- Algebraic datatypes and pattern-matching
- Type inference

Some of these features you already know and love, and some are probably new to you. But as we hope to demonstrate over the course of this book, it turns out that there is something transformative about having them all together and able to interact with each other in a single language.

Despite their importance, these ideas have made only limited inroads into mainstream languages. And when they do arrive there, like higher-order functions in C# or parametric polymorphism in Java, it's typically in a limited and awkward form. The only languages that support these ideas well are statically-typed functional programming languages like OCaml, F#, Haskell, Scala and Standard-ML.

Among this worthy set of languages, OCaml stands apart because it manages to provide a great deal of power while remaining highly pragmatic, highly performant, and comparatively simple to use and understand. It is this that makes OCaml a great choice for programmers who want to step up to a better programming language, and at the same time want to get practical work done.

Why Core?

A language on its own isn't enough. You also need a rich set of libraries to base your applications on. A common source of frustration for those learning OCaml is that the standard library that ships with the OCaml compiler is not ideal. While it's well implemented, it covers only a small subset of the functionality you expect from a standard library, and the interfaces are idiosyncratic and inconsistent.

But all is not lost! There is an effective alternative to the OCaml standard library called Core. Jane Street, a company that has been using OCaml for nearly a decade, developed Core for its own internal use, but it was designed from the start with an eye towards being a general-purpose standard library. Core is also distributed with syntax-extensions which provide essential new functionality to OCaml; and there are additional libraries, like Core_extended and Async, that provide even more useful functionality.

We believe that Core makes OCaml a better tool, and that's why we'll present OCaml and Core together.

About the Authors

Jason Hickey

Jason Hickey is a Software Engineer at Google Inc. in Mountain View, California. He is part of the team that designs and develops the global computing infrastructure used to support Google services, including the software systems for managing and scheduling massively distributed computing resources.

Prior to joining Google, Jason was an Assistant Professor of Computer Science at Caltech, where his research was in reliable and fault-tolerant computing systems, including programming language design, formal methods, compilers, and new models of distributed computation. He obtained his PhD in Computer Science from Cornell University, where he studied programming languages. He is the author of the MetaPRL

system, a logical framework for design and analysis of large software systems; OMake, an advanced build system for large software projects. He is the author of the textbook, *An Introduction to Objective Caml* (unpublished).

Anil Madhavapeddy

Anil Madhavapeddy is a Senior Research Fellow at the University of Cambridge, based in the Systems Research Group. He was on the original team that developed the Xen hypervisor, and helped develop an industry-leading cloud management toolstack written entirely in OCaml. This XenServer product has been deployed on hundreds of thousands of physical hosts, and drives critical infrastructure for many Fortune 500 companies.

Prior to obtaining his PhD in 2006 from the University of Cambridge, Anil had a diverse background in industry at Network Appliance, NASA and Internet Vision. In addition to professional and academic activities, he is an active member of the open-source development community with the OpenBSD operating system, is co-chair of the Commercial Uses of Functional Programming workshop, and serves on the boards of startup companies such as Ashima Arts where OCaml is extensively used.

Yaron Minsky

Yaron Minsky heads the Technology group at Jane Street, a proprietary trading firm that is the largest industrial user of OCaml. He was responsible for introducing OCaml to the company and for managing the company's transition to using OCaml for all of its core infrastructure. Today, billions of dollars worth of securities transactions flow each day through those systems.

Yaron obtained his PhD in Computer Science from Cornell University, where he studied distributed systems. Yaron has lectured, blogged and written about OCaml for years, with articles published in Communications of the ACM and the Journal of Functional Programming. He chairs the steering committee of the Commercial Users of Functional Programming, and is a member of the steering committee for the International Conference on Functional Programming.

A Guided Tour

This chapter is going to give an overview of OCaml by walking through a series of small examples that cover most of the major features. This should give a sense of what OCaml can do, without going into too much detail about any particular topic.

We'll present this guided tour using the OCaml toplevel, an interactive shell that lets you type in expressions and evaluate them interactively. When you get to the point of running real programs, you'll want to leave the toplevel behind, but it's a great tool for getting to know the language.

You should have a working toplevel as you go through this chapter, so you can try out the examples as you go. There is a zero-configuration browser-based toplevel that you can use for this, which you can find here:

```
http://realworldocaml.org/core-top
```

Or you can install OCaml and Core on your computer directly. Instructions for this are found in Appendix {???}.

OCaml as a calculator

Let's spin up the toplevel and open the `Core.Std` module, which gives us access to Core's libraries, and then try out a few simple numerical calculations.

```
$ rlwrap ocaml
        Objective Caml version 3.12.1

# open Core.Std;;
# 3 + 4;;
- : int = 7
# 8 / 3;;
- : int = 2
# 3.5 +. 6.;;
- : float = 9.5
```

```
# sqrt 9.;;
- : float = 3.
```

This looks a lot what you'd expect from any language, but there are a few differences that jump right out at you.

- We needed to type `;;` in order to tell the toplevel that it should evaluate an expression. This is a pecularity of the toplevel that is not required in compiled code.
- After evaluating an expression, the toplevel spits out both the type of the result and the result itself.
- Function application in OCaml is syntactically unusual, in that function arguments are written out separated by spaces, rather than being demarcated by parens and commas.
- OCaml carefully distinguishes between `float`, the type for floating point numbers and `int`. The types have different literals (`6.` instead of `6`) and different infix operators (`+.` instead of `+`), and OCaml doesn't do any automated casting between the types. This can be a bit of a nuisance, but it has its benefits, since it prevents some classes of bugs that arise from confusion between the semantics of `int` and `float`.

We can also create variables to name the value of a given expression, using the `let` syntax.

```
# let x = 3 + 4;;
val x : int = 7
# let y = x + x;;
val y : int = 14
```

After a new variable is created, the toplevel tells us the name of the variable, in addition to its type and value.

Functions and Type Inference

The `let` syntax can also be used for creating functions:

```
# let square x = x * x ;;
val square : int -> int = <fun>
# square (square 2);;
- : int = 16
```

Now that we're creating more interesting values, the types have gotten more interesting too. `int -> int` is a function type, in this case indicating a function that takes an `int` and returns an `int`. We can also write functions that take multiple arguments:

```
# let abs_diff x y =
    abs (x - y) ;;
val abs_diff : int -> int -> int = <fun>
```

and even functions that take other functions as arguments:

```
# let abs_change f x =
    abs_diff (f x) x ;;
val abs_change : (int -> int) -> int -> int = <fun>
# abs_change square 10;;
- : int = 90
```

This notation for multi-argument functions may be a little surprising at first, but we'll explain where it comes from when we get to function currying in Chapter {{{VARI-ABLES}}}. For the moment, think of the arrows as separating different arguments of the function, with the type after the final arrow being the return value of the function. Thus,

```
int -> int -> int
```

describes a function that takes two `int` arguments and returns an `int`, while

```
(int -> int) -> int -> int
```

describes a function of two arguments where the first argument is itself a function.

As the types we encounter get more complicated, you might ask yourself how OCaml is able to determine these types, given that we didn't write down any explicit type information. It turns out that OCaml is able to determine the type of a new expression using a technique called *type-inference*, by which it infers the type of a new expression based on what it already knows about the types of other related variables. For example, in `abs_change` above, the fact that `abs_diff` is already known to take two integer arguments lets the compiler infer that `x` is an `int` and that `f` returns an `int`.

Sometimes, there isn't enough information to fully determine the concrete type of a given value. Consider this function.

```
# let first_if_true test x y =
    if (test x) then x else y;;
```

`first_if_true` takes as its arguments a function `test`, and two values, `x` and `y`, where `x` is to be returned if `(test x)` evaluates to `true`, and `y` otherwise. So what's the type of `first_if_true`? There are no obvious clues such as arithmetic operators to tell you what the type of `x` and `y` are. Indeed, it seems like one could use this `first_if_true` on values of any type. Indeed, if we look at the type returned by the toplevel:

```
val first_if_true : ('a -> bool) -> 'a -> 'a -> 'a = <fun>
```

we see that rather than choose a single concrete type, OCaml has introduced a *type variable* `'a` to express that the type is generic. In particular, the type of the `test` argument is (`'a -> bool`), which means that test is a one-argument function whose return value

is bool, and whose argument could be of any type 'a. But, whatever type 'a is, it has to be the same as the type of the other two arguments, x and y.

This genericity means that we can write:

```
# let long_string s = String.length s > 6;;
val long_string : string -> bool = <fun>
# first_if_true long_string "short" "loooooong";;
- : string = "loooooong"
```

And we can also write:

```
# let big_number x = x > 3;;
val big_number : int -> bool = <fun>
# first_if_true big_number 4 3;;
- : int = 4
```

But we can't mix and match two different concrete types for 'a in the same use of first_if_true:

```
# first_if_true big_number "short" "loooooong";;
Characters 25-30:
  first_if_true big_number "short" "loooooong";;
                           ^^^^^^^
Error: This expression has type string but
   an expression was expected of type int
```

While the 'a in the type of first_if_true can be instantiated as any concrete type, it has to be the same concrete type in all of the different places it appears. This kind of genericity is called *parametric polymorphism*, and is very similar to generics in C# and Java.

Type errors vs exceptions

There's a big difference in OCaml (and really in any compiled language) between errors that are caught at compile time and those that are caught at run-time. It's better to catch errors as early as possible in the development process, and compilation time is best of all.

Working in the top-level somewhat obscures the difference between run-time and compile time errors, but that difference is still there. Generally, type errors, like this one:

```
# 3 + "potato";;
Characters 4-12:
  3 + "potato";;
      ^^^^^^^^
Error: This expression has type string but an expression was expected of type
       int
```

are compile-time errors, whereas division by zero, is a runtime error, or an exception:

```
# 3 / 0;;
Exception: Division_by_zero.
```

One important distinction is that type errors will stop you whether or not the offending code is ever actually executed. Thus, you get an error from typing in this code:

```
# if 3 < 4 then 0 else 3 + "potato";;
Characters 25-33:
  if 3 < 4 then 0 else 3 + "potato";;
                          ^^^^^^^^
Error: This expression has type string but an expression was expected of type
       int
```

but this code works fine.

```
# if 3 < 4 then 0 else 3 / 0;;
- : int = 0
```

Tuples, Lists, Options and Pattern-matching

Tuples

So far we've encountered a handful of basic types like `int`, `float` and `string` as well as function types like `string -> int`. But we haven't yet talked about any data structures. We'll start by looking at a particularly simple data structure, the tuple. You can create a tuple by joining values together with a comma:

```
# let tup = (3,"three");;
val tup : int * string = (3, "three")
```

The type `int * string` corresponds to the set of pairs of `int`s and `string`s. For the mathematically inclined, the * character is used because the space of all 2-tuples of type `t * s` corresponds to the Cartesian product of `t` and `s`.

You can extract the components of a tuple using OCaml's pattern-matching syntax. For example:

```
# let (x,y) = tup;;
val x : int = 3
val y : string = "three"
```

`(x,y)` is the pattern, and it's used as the right-hand side of the `let` binding. Note that this operation lets us mint the new variables `x` and `y`, each bound to different components of the value being matched.

This is just a first taste of pattern matching. Pattern matching shows up in many contexts, and is a surprisingly powerful and pervasive tool.

Here's another example: a function for computing the distance between two points on the plane, where each point is represented as a pair of floats.

```
# let distance p1 p2 =
    let (x1,y1) = p1 in
    let (x2,y2) = p2 in
    sqrt ((x1 -. x2) ** 2. +. (y1 -. y2) ** 2)
;;
val distance : float * float -> float * float -> float = <fun>
```

We can make this code more concise by doing the pattern matching on the arguments to the function directly.

```
# let distance (x1,y1) (x2,y2) =
    sqrt ((x1 -. x2) ** 2. +. sqr (y1 -. y2) ** 2.)
;;
```

Lists

Where tuples let you combine a fixed number of items, potentially of different types, lists let you hold any number of items of the same type. For example:

```
# let languages = ["OCaml";"Perl";"C"];;
val languages : string list = ["OCaml"; "Perl"; "C"]
```

Note that you can't mix elements of different types on the same list, as we did with tuples.

```
# let numbers = [3;"four";5];;
Characters 17-23:
  let numbers = [3;"four";5];;
                   ^^^^^^
Error: This expression has type string but an expression was expected of type
         int
```

We can also use the operator :: for adding elements to the front of a list

```
# "French" :: "Spanish" :: languages;;
- : string list = ["French"; "Spanish"; "OCaml"; "Perl"; "C"]
```

This is creating a new list, not changing the list we started with, so the definition of languages is unchanged.

```
# languages;;
- : string list = ["OCaml"; "Perl"; "C"]
```

The bracket notation for lists is just syntactic sugar for ::. Thus, the following declarations are all equivalent. Note that [] is used to represent the empty list.

```
# [1; 2; 3];;
- : int list = [1; 2; 3]
# 1 :: (2 :: (3 :: []));;
- : int list = [1; 2; 3]
# 1 :: 2 :: 3 :: [];;
- : int list = [1; 2; 3]
```

Basic list patterns

The elements of a list can be accessed through pattern-matching. List patterns have two key components: [], which represents the empty list, and ::, which connects an element at the head of a list to the remainder of the list, as you can see below.

```
# let (my_favorite :: the_rest) = languages ;;
val my_favorite : string = "OCaml"
val the_rest : string list = ["Perl"; "C"]
```

By pattern matching using ::, we've broken off the first element of languages from the rest of the list. If you know Lisp or Scheme, what we've done is the equivalent of using car and cdr to break down a list.

If you tried the above example in the toplevel, you probably noticed that I omitted a warning generated by the compiler. Here's the full output:

```
# let (my_favorite :: the_rest) = languages ;;
Characters 5-28:
  let (my_favorite :: the_rest) = languages ;;
       ^^^^^^^^^^^^^^^^^^^^^^^^
Warning 8: this pattern-matching is not exhaustive.
Here is an example of a value that is not matched:
[]
val my_favorite : string = "OCaml"
val the_rest : string list = ["Perl"; "French"; "C"]
```

The warning comes because the compiler can't be certain that the pattern match won't lead to a runtime error, and the warnings gives an example of the problem, the empty list, []. Indeed, if we try to use such a pattern-match on the empty list:

```
# let (my_favorite :: the_rest) = [];;
Characters 5-28:
  let (my_favorite :: the_rest) = [];;
       ^^^^^^^^^^^^^^^^^^^^^^^^
Warning 8: this pattern-matching is not exhaustive.
Here is an example of a value that is not matched:
[]
Exception: (Match_failure "" 1 5).
```

we get a runtime error in addition to the compilation warning.

You can avoid these warnings, and more importantly make sure that your code actually handles all of the possible cases, by using a match statement. Here's an example:

```
# let my_favorite_language languages =
    match languages with
    | first :: the_rest -> first
    | [] -> "OCaml" (* A good default! *)
 ;;
val my_favorite_language : string list -> string = <fun>
# my_favorite_language ["English";"Spanish";"French"];;
- : string = "English"
# my_favorite_language [];;
- : string = "OCaml"
```

This is the syntax of a match statement.

```
match <expr> with
| <pattern1> -> <expr1>
| <pattern2> -> <expr2>
| ...
```

The return value of the match is the result of evaluating the right-hand side of the first pattern that matches the value of <expr>. As with print_log_entry, the patterns can mint new variables, giving names to sub-components of the data structure being matched.

Recursive list functions

(yminsky: maybe we should kick this subsection to the full list chapter? This is getting long...)

If we combine pattern matching with a recursive function call, we can do things like define a function for summing the elements of a list.

```
# let rec sum l =
    match l with
    | [] -> 0
    | hd :: tl -> hd + sum tl
  ;;
val sum : int list -> int
# sum [1;2;3;4;5];;
- : int = 15
```

We had to add the rec keyword in the definition of sum to allow for sum to refer to itself. We can introduce more complicated list patterns as well. Here's a function for destuttering a list, *i.e.*, for removing sequential duplicates.

```
# let rec destutter list =
    match list with
    | [] -> []
    | hd1 :: (hd2 :: tl) ->
      if hd1 = hd2 then destutter (hd2 :: tl)
      else hd1 :: destutter (hd2 :: tl)
```

Actually, the code above has a problem. If you type it into the top-level, you'll see this error:

```
Warning 8: this pattern-matching is not exhaustive.
Here is an example of a value that is not matched:
_::[]
```

This is warning you that we've missed something, in particular that our code doesn't handle one-element lists. That's easy enough to fix by adding another case to the match:

```
# let rec destutter list =
    match list with
    | [] -> []
    | [hd] -> [hd]
    | hd1 :: (hd2 :: tl) ->
      if hd1 = hd2 then destutter (hd2 :: tl)
      else hd1 :: destutter (hd2 :: tl)
val destutter : 'a list -> 'a list = <fun>
# destutter ["hey";"hey";"hey";"man!"];;
- : string list = ["hey"; "man!"]
```

Note that in the above, we used another variant of the list pattern, [hd], to match a list with a single element. We can do this to match a list with any fixed number of elements, *e.g.*, [x;y;z] will match any list with exactly three elements, and will bind those elements to the variables x, y and z.

The List module

So far, we've built up all of our list functions using pattern matching and recursion. But in practice, this isn't usually necessary. OCaml libraries are organized into *modules*, and there is a module in Core called List which contains many useful functions for dealing with lists. For example:

```
# List.map ~f:String.length languages;;
- : int list = [5; 4; 6; 1]
```

Here, we use the dot-notation to reference elements of the List and String module. List.map in particular is a function that takes a list and a function for transforming elements of that list, and returns to us a new list with the transformed elements.

There's another new piece of syntax to learn here: labeled arguments. String.length is passed with the label, ~f. Labeled arguments are arguments that are specified by name rather than position, which means they can be passed in any order. Thus, we could have written List.map languages ~f:String.length instead of List.map ~f:String.length languages. We'll see why labels are important in Chapter {{{Functions}}}.

Options

Another common data structure in OCaml is the `option`. An `option` is used to express that a value that might or might not be present. For example,

```
# let divide x y =
    if y = 0 then None else Some (x/y)
val divide : int -> int -> int option = <fun>
```

Here, `Some` and `None` are explicit tags, caled *type constructors* that are used to build an optional value. You can think of an `option` as a specialized list that can only have zero or one element.

To get a value out of an option, we use pattern matching, as we did with tuples and lists. Consider the following simple function for printing a log entry given an optional time and a message. If no time is provided (*i.e.*, if the time is `None`), the current time is computed and used in its place.

```
# let print_log_entry maybe_time message =
    let time =
      match maybe_time with
      | Some x -> x
      | None -> Time.now ()
    in
    printf "%s: %s\n" (Time.to_string time) message
val print_log_entry : Time.t option -> string -> unit
```

Here, we again use a match statement for handling the two possible states of an option. It's worth noting that we don't necessarily need to use an explicit `match` statement in this case. We can instead use some built in functions from the `Option` module, which, like the `List` module for lists, is a place where you can find a large collection of useful functions for owrking with options.

In this case, we can rewrite `print_log_entry` using `Option.value`, which either returns the content of an option if the option is `Some`, or a default value if the option is `None`.

```
# let print_log_entry maybe_time message =
    let time = Option.value ~default:(Time.now ()) maybe_time in
    printf "%s: %s\n" (Time.to_string time) message
```

Options are important because they are the standard way in OCaml to encode a value that might not be there. Values in OCaml are non-nullable, so if you have a function that takes an argument of type `string`, then the compiler guarantees that, if the code compiles successfully, then at run-time, that function will only be called with awell-defined value of type `string`. This is different from most other languages, including Java and C#, where objects are by default nullable, and whose type systems do little to defend from null-pointer exceptions at runtime.

Given that in OCaml ordinary values are not nullable, you need some other way of representing values that might not be there, and the `option` type is the standard solution.

Records and Variants

So far, we've only looked at data structures that were pre-defined in the language, like lists and tuples. But OCaml also allows us to define new datatypes. Here's a toy example of a datatype representing a point in 2-dimensional space:

```
# type point2d = { x : float; y : float };;
type point2d = { x : float; y : float; }
```

`point2d` is a *record* type, which you can think of as a tuple where the individual fields are named, rather than being defined positionally. Record types are easy enough to construct:

```
# let p = { x = 3.; y = -4. };;
val p : point2d = {x = 3.; y = -4.}
```

And we can get access to the contents of these types using pattern matching:

```
# let magnitude { x = x; y = y } = sqrt (x ** 2. +. y ** 2.);;
val magnitude : point2d -> float = <fun>
```

In the case where we want to name the value in a record field after the name of that field, we can write the pattern match even more tersely. Instead of writing { x = x } to name a variable x for the value of field x, we can write { x }. Using this, we can rewrite the magnitude function as follows.

```
# let magnitude { x; y } = sqrt (x ** 2. +. y ** 2.);;
```

We can also use dot-notation for accessing record fields:

```
# let distance v1 v2 =
      magnitude { x = v1.x -. v2.x; y = v1.y -. v2.y };;
val distance : point2d -> point2d -> float = <fun>
```

And we can of course include our newly defined types as components in larger types, as in the following types, each of which representing a different geometric object.

```
# type circle = { center: point2d; radius: float } ;;
# type rect = { lower_left: point2d; width: float; height: float } ;;
# type segment = { endpoint1: point2d; endpoint2: point2d } ;;
```

Now, imagine that you want to combine multiple of these scene objects together, say as a description scene containing multiple objects. You need some unified way of representing these objects together in a single type. One way of doing this is using a *variant* type:

```
# type shape = | Circle of circle
               | Rect of rect
               | Segment of segment;;
```

You can think of a variant as a way of combining different types as different possibilities. The | character separates the different cases of the variant (the first | is optional), and each case has a tag (like `Circle`, `Rect` and `Segment`) to distinguish each case from the other. Here's how we might write a function for testing whether a point is in the interior of one of a `shape list`.

```
# let is_inside_shape point shape =
    match shape with
    | Circle { center; radius } ->
      distance center point < radius
    | Rect { lower_left; width; height } ->
      point.x > lower_left.x && point.x < lower_left.x +. width
      && point.y > lower_left.y && point.y < lower_left.y +. height
    | Segment _ -> false
    ;;
val is_inside_shape : point2d -> shape -> bool = <fun>
# let is_inside_shapes point shapes =
    let point_is_inside_shape shape =
      is_inside_shape point shape
    in
    List.for_all shapes ~f:point_is_inside_shape
val is_inside_shapes : point2d -> shape list -> bool = <fun>
```

You might at this point notice that the use of `match` here is reminiscent of how we used `match` with `option` and `list`. This is no accident: `option` and `list` are really just examples of variant types that happen to be important enough to be defined in the standard library (and in the case of lists, to have some special syntax).

Mutation

All of our examples so far have been examples of mutation-free, or *pure* code. This is typical of code in functional languages, which tend to have a focus on so-called *pure* code. That said, OCaml has good support for mutation, including standard mutable data structures like arrays and hashtables. For example:

```
# let numbers = [| 1;2;3;4 |];;
val numbers : int array = [|1; 2; 3; 4|]
# numbers.(2) <- 4;;
- : unit = ()
# numbers;;
- : int array = [|1; 2; 4; 4|]
```

In the above, the `ar.(i)` syntax is used for referencing the element of an array, and the `<-` syntax is used for setting a mutable value.

Variable bindings in OCaml are always immutable, but datastructures like arrays can be mutable. In addition, record fields, which are immutable by default can be declared as mutable. Here's a small example of a datastructure for mutable storing a running sum. Here, we've declared all the record fields as mutable. Here

```
# type running_sum = { mutable sum: float;
                       mutable sum_sq: float; (* sum of squares, for stdev *)
                       mutable samples: float; }
  let empty () = { sum = 0.; sum_sq = 0.; samples = 0. }
  let mean rsum = rsum.sum /. rsum.samples
  let stdev rsum =
      let square x = x *. x in
      sqrt (rsum.sum_sq /. rsum.samples -. square (rsum.sum /. rsum.samples))
  let update rsum x =
      rsum.sum <- rsum.sum +. x;
      rsum.sum_sq <- rsum.sum_sq +. x *. x;
      rsum.samples <- rsum.samples +. 1.
  ;;
# let rsum = empty ();;
val rsum : running_sum = {sum = 0.; sum_sq = 0.; samples = 0}
# List.iter [1.;3.;2.;-7.;4.;5.] ~f:(fun x -> update rsum x);;
- : unit = ()
# mean rsum;;
- : float = 1.33333333333333326
# stdev rsum;;
- : float = 1.61015297179882655
```

We can declare a single mutable value by using a `ref`, which is a record type with a single mutable field that is defined in the standard library.

```
# let x = { contents = 0 };;
val x : int ref = {contents = 0}
# x.contents <- x.contents + 1;;
- : unit = ()
# x;;
- : int ref = {contents = 1}
```

There are a handful of useful functions and operators defined for refs to make them more convenient to work with.

```
# let x = ref 0 ;; (* create a ref, i.e., { contents = 0 } *)
val x : int ref = {contents = 0}
# !x ;;              (* get the contents of a ref, i.e., x.contents *)
- : int = 0
# x := !x + 1 ;;    (* assignment, i.e., x.contents <- ... *)
- : unit = ()
# incr x ;;          (* increment, i.e., x := !x + 1 *)
- : unit = ()
# !x ;;
- : int = 2
```

I/O

Variables and Functions

Variables are a fundamental concept in programming, one that comes up in all but the simplest of examples. Indeed, we encountered OCaml's variables multiple times in chapter {{TOUR}}. But while variables are no doubt a familiar topic, variables in OCaml are different in subtle but important ways from what you find in most other languages. Accordingly we're going to spend a some time diving into the details of how variables work in OCaml.

At its simplest, a variable is an identifier whose meaning is bound to a particular value. In OCaml these bindings are usually introduced using the `let` keyword, which at the top-level of a module has the following syntax.

```
let <identifier> = <expr>
```

Every variable binding has a *scope*, which is the portion of the code that considers the given variable binding during its evaluation. The scope of a top-level let binding is everything that follows it in that module (or, the remainder of the session if you're using the top-level.)

Here's a simple example.

```
# let x = 3;;
val x : int = 3
# let y = 4;;
val y : int = 4
# let z = x + y;;
val z : int = 7
```

`let` can also be used to create a variable binding whose scope is limited to a particular, bounded expression, using the following syntax.

```
let <identifier> = <expr1> in <expr2>
```

This first evaluates `<expr1>`, and then evaluates `<expr2>`, with `<identifier>` bound to whatever value was produced by the evaluation of `<expr1>`. For example, `let x = 3 + 1 in x * 2` evaluates to `8`.

In this form, multiple let bindings can be nested, like this:

```
# let x = 3 in
  let y = 4 in
  x + y
;;
- : int = 7
```

Note that nested bindings can *shadow*, or hide, previous bindings. Thus, we can write

```
# let x = 1 in
  let x = x + x in
  let x = Int.to_string (x + x) in
  x ^ x;;
- : string = "44"
```

It's important not to confuse shadowing of variables with assignment, *i.e.*, mutation. Consider the following function.

```
# let x = 3 in
  let add_to_x y = x + y in
  let x = 4 in
  add_to_x 0
;;
- : int = 3
```

If the second let binding of x were in fact an assignment, then you would expect `add_to_x 0` to return `4`. Instead, it returns `3`, because the x that `add_to_x` refers to is still there, unchanged, even after the new binding of x to `4` is created.

Here's another demonstration of how let bindings differ from assignment. In the following example, the second binding of x is only visible within the scope of a fixed subexpression, in particular, the sub-expression that makes up the right-hand side of the definition of y. When the definition of y is complete, we see that the inner definition disappears, and the original definition of x shows up again, unaffected.

```
# let x = 3 in
  let y =
    let x = 2 in
    x + x
  in
  x + y
- : int = 7
```

In OCaml, let bindings are always immutable. As we'll see in chapter {{MUTABIL-ITY}}, there are mutable values in OCaml, but no mutable variables.

Pattern matching and `let`

Another useful feature of let bindings is that they support the use of patterns on the left-hand side of the bind. Consider the following code, which uses `List.unzip`, a function for converting a list of pairs to a pair of lists.

```
# let (ints,strings) = List.unzip [(1,"one"); (2,"two"); (3,"three")]
val ints : int Core.Std.List.t = [1; 2; 3]
val strings : string Core.Std.List.t = ["one"; "two"; "three"]
```

This actually binds two variables, one for each element of the pair. Using a pattern in a let-binding makes the most sense for a pattern that is *irrefutable*, i.e., where any value of the type in question is guaranteed to match the pattern. Tuple and record patterns are irrefutable, but list patterns are not. Here's an example of a list pattern match that generates a warning because not all cases are covered.

```
# let (hd::tl) = [1;2;3];;
Characters 4-12:
  let (hd::tl) = [1;2;3];;
      ^^^^^^^^
Warning 8: this pattern-matching is not exhaustive.
Here is an example of a value that is not matched:
[]
```

As a general matter, inexhaustive matches like the one above should be avoided.

`let`/`and` bindings

Another form of let binding that comes up on occasion is where you bind multiple arguments in a single declaration. For example, we can write:

```
# let x = 100 and y = 3.5;;
val x : int = 100
val y : float = 3.5
```

This can be useful when you want to create a number of new let bindings at once, without having each definition affect the next. So, if we wanted to create new bindings that swapped the values of x and y, we could write:

```
# let x = y and y = x ;;
val x : float = 3.5
val y : int = 100
```

This use-case doesn't come up that often. Most of the time that `and` comes into play, it's used to define multiple mutually recursive values, which we'll learn about later in the chapter.

Note that when doing a `let`/`and` style declaration, the order of execution of the right-hand side of the binds is undefined.

Functions

OCaml function declarations come in multiple styles. The most basic form is to create an *anonymous* function using the `fun` keyword:

```
# (fun x -> x + 1);;
- : int -> int = <fun>
```

The above expression creates a one-argument function, which can straightforwardly be applied to an argument:

```
# (fun x -> x + 1) 7;;
- : int = 8
```

Anonymous functions are quite convenient, particularly in a higher-order context, *e.g.*, when constructing a function to be passed as an argument to another function.

We can create a named function using a let binding.

```
# let plusone = (fun x -> x + 1);;
val plusone : int -> int = <fun>
# plusone 3;;
- : int = 4
```

The declaration of `plusone` above is equivalent to the following form, which we already saw in chapter {{TOUR}}:

```
# let plusone x = x + 1;;
```

This is the most common and convenient way to declare a function, but syntatic niceties aside, the two forms are entirely equivalent.

---

let and fun

Functions and let bindings have a lot to do with each other. In some sense, you can think of the argument of a function as a variable being bound to its argument. Indeed, the following two expressions are nearly equivalent:

```
# (fun x -> x + 1) 7;;
- : int = 8
# let x = 7 in x + 1;;
- : int = 8
```

This connection is important, and will come up more when programming in a monadic style, as we'll see in chapter {{ASYNC}}.

Multi-argument functions

OCaml of course also supports multi-argument functions. Here's an example that came up in chapter {{TOUR}}.

```
# let abs_diff x y = abs (x - y);;
val abs_diff : int -> int -> int = <fun>
# abs_diff 3 4;;
- : int = 1
```

You may find the type signature of `abs_diff` a bit obscure at first. To understand what's going on, let's rewrite `abs_diff` in an equivalent form, using the `fun` keyword:

```
# let abs_diff =
    (fun x -> (fun y -> abs (x - y)));;
val abs_diff : int -> int -> int = <fun>
```

This rewrite makes it explicit that `abs_diff` is actually a function of one argument that returns another function of one argument, which itself returns the absolute difference between the argument given to the first function and the argument given to the second. In other words, `abs_diff` is a nested, or *curried* function. (Currying is named after Haskell Curry, a famous logician who had a significant impact on the design and theory of programming languages.)

The key to interpreting the type signature of a curried function is the observation that `->` is right-associative. The type signature of `abs_diff` can therefore be parenthesized as follows to make the currying more obvious without changing the meaning of the signature.

```
val abs_diff : int -> (int -> int)
```

Currying is more than just a theoretical curiosity. Here's an example of how you can make use of currying.

```
# let dist_from_3 = abs_diff 3;;
val dist_from_3 : int -> int = <fun>
# dist_from_3 8;;
- : int = 5
# dist_from_3 (-1);;
- : int = 4
```

The practice of applying some of the arguments of a curried function to get a new function is called *partial application*, and it is a convenient way to mint new, specialized functions from more general ones.

Note that the `fun` keyword supports its own syntactic sugar for currying, so we could also have written `abs_diff` as follows.

```
# let abs_diff = (fun x y -> abs (x - y));;
```

You might worry that curried functions are terribly expensive, but this is not an issue. In OCaml, there is no penalty for calling a curried function with all of its arguments. (Partial application, unsurprisingly, does have a small cost.)

Currying is the standard way in OCaml of writing a multi-argument function, but it's not the only way. It's also possible to use the different arms of a tuple as different arguments. So, we could write:

```
# let abs_diff (x,y) = abs (x - y)
val abs_diff : int * int -> int = <fun>
# abs_diff (3,4);;
- : int = 1
```

OCaml handles this calling convention efficiently as well. In particular it does not generally have to allocate a tuple just for the purpose of sending arguments to a tuple-style function.

There are small tradeoffs between these two styles, but most of the time, once should stick to currying, since it's the default style in the OCaml world.

Recursive functions

In order to define a recursive function, you need to mark the let binding as recursive with the rec keyword, as shown in this example:

```
# let rec find_first_stutter = function
    | [] | [_] ->
      (* only zero or one elements, so no repeats *)
      None
    | x :: y :: tl ->
      if x = y then Some x else find_first_stutter (y::tl)
val find_first_stutter : 'a list -> 'a option = <fun>
```

We can also define multiple mutually recursive values by using let rec and and together, as in this (gratuitiously inefficient) example.

```
# let rec is_even x =
    x = 0 || is_odd (x - 1)
  and is_odd x =
    is_even (x - 1)
val is_even : int -> bool = <fun>
val is_odd : int -> bool = <fun>
```

Note that in the above example, we take advantage of the fact that the right hand side of the || is only evaluated if the left hand side evaluates to false.

Prefix and Infix operators

So far, we've seen examples of functions used in both prefix and infix style:

```
# Int.max 3 4;;  (* prefix *)
- : int = 4
# 3 + 4;;        (* infix  *)
- : int = 7
```

In OCaml, functions can only be used infix if the name of the function is chosen from one of a specialized set of identifiers called *operators*. An operator is any identifier that is a sequence of characters from the following set

```
! $ % & * + - . / : < = > ? @ ^ | ~
```

or is one of a handful of pre-determined strings, including things like mod, the modulus operator, and lsl, for "logical shift right", which is a bit-shifting operation.

We can define (or redefine) the meaning of an operator as follows:

```
# let (+!) (x1,y1) (x2,y2) = (x1 + x2, y1 + y2)
val ( +! ) : int * int -> int * int -> int *int = <fun>
# (3,2) +! (-2,4);;
- : int * int = (1,6)
```

Note that operators can be used in prefix style as well, if they are put in parens:

```
# (+!) (3,2) (-2,4);;
- : int = (1,6)
```

The details of how the operator works are determined by the first character of the operator. This table describes how, and lists the operators from highest to lowest precedence.

First character	Usage
! ? ~	Prefix and unary
**	Infix, right associative
+ -	Infix, left associative
@ ^	Infix, right associative
= < > | & $	Infix, left associative

Declaring functions with function

Another way to define a function is using the function keyword. Instead of having syntactic support for declaring curried functions, function has built-in pattern matching. Here's an example:

```
# let some_or_zero = function
      | Some x -> x
      | None -> 0
  ;;
```

```
val some_or_zero : int option -> int = <fun>
# List.map ~f:some_or_zero [Some 3; None; Some 4];;
- : int list = [3; 0; 4]
```

We can also combine the different styles of function declaration together, as in the following example where we declare a two argument function with a pattern-match on the second argument.

```
# let some_or_default default = function
    | Some x -> x
    | None -> default
  ;;
# List.map ~f:(some_or_default 100) [Some 3; None; Some 4];;
- : int Core.Std.List.t = [3; 100; 4]
```

Labeled Arguments

Up until now, the different arguments to a function have been specified positionally, *i.e.*, by the order in which the arguments are passed to the function. OCaml also supports labeled arguments, which let you identify a function argument by name. Functions with labeled arguments can be declared by putting a tilde in front of the variable name in the definition of the function:

```
# let f ~foo:a ~bar:b = a + b
val f : foo:int -> bar:int -> int = <fun>
```

And the function can be called using the same convention:

```
# f ~foo:3 ~bar:10;;
- : int = 13
```

In addition, OCaml supports *label punning*, meaning that you get to drop the text after the : if the name of the label and the name of the variable being used are the same. Label punning works in both function declaration and function invocation, as shown in these examples:

```
# let f ~foo ~bar = foo + bar
val f : foo:int -> bar:int -> int = <fun>
# let foo = 3;;
# let bar = 4;;
# f ~foo ~bar;;
- : int = 7
```

Labeled arguments are useful in a few different cases:

• When defining a function with lots of arguments. When you have enough arguments, names are easier to remember than positions.

- For functions that have multiple arguments that might get confused with each other, particularly if they're of the same type. For example, consider this signature for a function for extracting a substring of another string.

  ```
  val substring: string -> int -> int -> string
  ```

  where the two ints are the starting position and length of the substring to extract. Labeled arguments can make this signature clearer:

  ```
  val substring: string -> pos:int -> len:int -> string
  ```

  This improves the readability of both the signature and of client code that makes use of substring, and makes it harder to accidentally swap the position and the length.

- Labeled arguments give you a way to assign a clear name and meaning to an argument whose type is otherwise less than informative. For example, consider a function for creating a hashtable where the first argument is the initial size of the table, and the second argument is a flag which, when true, indicates that the hashtable will adjust its size down when its size is small. The following signature is less than informative.

  ```
  val create_hashtable : int -> bool -> ('a,'b) Hashtable.t
  ```

  but with labeled arguments, we can make the intent much clearer:

  ```
  val create_hashtable : init_size:int -> allow_shrinking:bool -> ('a,'b) Hashtable.t
  ```

- Labeled arguments can be used to make a function signature more flexible. For example, labeled arguments make it possible for the caller to decide which argument of a function to partially apply, whereas in ordinary curried functions, you can only partially apply the arguments in order from first to last. Labeled arguments also make it possible to place the arguments to a function in different orders, which is useful for functions like List.map where you often want to partially apply List.map with just the function, and at the same time mapping over a large function is easier to read if the function is the last argument.

One surprising gotcha about labeled arguments is that while order doesn't matter when calling a function with labeled arguments, it does matter in a higher-order context, *i.e.*, when passing a labeled argument to another function. This is shown by the following example.

```
# let apply_to_tuple f (first,second) = f ~first ~second;;
val apply_to_tuple : (first:'a -> second:'b -> 'c) -> 'a * 'b -> 'c = <fun>
# let divide ~second ~first = first / second;;
val divide : second:int -> first:int -> int = <fun>
# apply_to_tuple divide 3 4;;
Characters 15-21:
```

```
    apply_to_tuple divide 3 4;;
                    ^^^^^^
  Error: This expression has type second:int -> first:int -> int
         but an expression was expected of type
            first:'a -> second:'b -> 'c -> 'd
```

Optional arguments

An optional argument is like a labeled argument that the caller can choose whether or not to provide. A function with an optional argument must define a default for when the argument is absent. Consider the following example of a string concatenation function with an optionally specified separator. Note that the ? in front of an argument is used to make the separator optional.

```
# let concat ?sep x y =
      let sep = match sep with None -> "" | Some x -> x in
      x ^ sep ^ y
  ;;
val concat : ?sep:string -> string -> string -> string = <fun>
# concat "foo" "bar";; (* without the optional argument *)
- : string = "foobar"
# concat ~sep:":" "foo" "bar";; (* with the optional argument *)
- : string = "foo:bar"
```

Optional arguments can be passed in using the same syntax as labeled arguments. Also, similarly to labeled arguments, optional arguments can be passed in in any order.

How are optional arguments inferred?

One tricky aspect of labeled and optional arguments is the way in which those arguments are inferred. Consider the following example:

```
# let foo g x y = g ~x ~y ;;
val foo : (x:'a -> y:'b -> 'c) -> 'a -> 'b -> 'c = <fun>
```

In principle, it seems like the type first argument of foo could have had a different order for the arguments (*e.g.* y:'b -> x:'a -> 'c) or could have optional instead of labeled arguments (*e.g.*, ?y:'a -> x:'b -> 'c). OCaml disambiguates between these cases by picking labeled arguments when it can, and by choosing the order based on the order that is actually used. If you try to use two different orders in the same context, you'll get a compilation errro:

```
# let foo g x y = g ~x ~y + g ~y ~x ;;
Characters 26-27:
  let foo g x y = g ~x ~y + g ~y ~x ;;
                            ^
Error: This function is applied to arguments
in an order different from other calls.
This is only allowed when the real type is known.
```

If, however, we put in an explicit type constraint, then we can specify any compatible type.

```
# let foo g x y = (g : ?y:'a -> x:'b -> int) ~x ~y + g ~y ~x;;
val foo : (?y:'a -> x:'b -> int) -> 'b -> 'a -> int = <fun>
```

Type constraints are discussed in more detail in chapter {{???}}.

The behavior of substituting in a default value is so common that it has its own syntax. Thus, we could rewrite the `concat` function as follows:

```
# let concat ?(sep="") x y = x ^ sep ^ y ;;
```

Explicit passing of an optional argument

Sometimes you want to explicitly invoke an optional argument with a concrete option, where `None` indicates that the argument won't be passed in, and `Some` indicates it will. You can do that as follows:

```
# concat ?sep:None "foo" "bar";;
- : string = "foobar"
```

This is particularly useful when you want to pass through an optional argument from one function to another, leaving the choice of default to the second function. For example:

```
# let uppercase_concat ?sep a b = concat ?sep (String.uppercase a) b ;;
val uppercase_concat : ?sep:string -> string -> string -> string =
  <fun>
# uppercase_concat "foo" "bar";;
- : string = "FOObar"
```

Erasure of optional arguments

One subtle aspect of optional arguments is the question of OCaml decides to *erase* an optional argument, *i.e.*, to give up waiting for an optional argument, and substitute in the default value? Note that, for ordinary labeled arguments, if you pass in all of the non-labeled arguments, you're left with a partially applied function that is still waiting for its labeled arguments. *e.g.*,

```
# let concat ~sep x y = x ^ sep ^ y ;;
val concat : sep:string -> string -> string -> string = <fun>
# concat "a" "b";;
- : sep:string -> string = <fun>
```

So when should an optional argument be erased?

OCaml's rule is: an optional argument is erased as soon as the first positional argument defined *after* the optional argument is passed in. Thus, the following partial application of concat causes the optional argument to disappear:

```
# let prepend_foo = concat "foo";;
val prepend_foo : string -> string = <fun>
# prepend_foo "bar";;
- : string = "foobar"
# prepend_foo "bar" ~sep:":";;
Characters 0-11:
  prepend_foo "bar" ~sep:":";;
  ^^^^^^^^^^^
Error: This function is applied to too many arguments;
maybe you forgot a `;'
```

But if we had instead defined concat with the optional argument in the second position:

```
# let concat x ?(sep="") y = x ^ sep ^ y ;;
val concat : string -> ?sep:string -> string -> string = <fun>
```

then application of the first argument would not cause the optional argument to be erased.

```
# let prepend_foo = concat "foo";;
val prepend_foo : ?sep:string -> string -> string = <fun>
# prepend_foo "bar";;;
- : string = "foobar"
# prepend_foo ~sep:"=" "bar";;;
- : string = "foo=bar"
```

One oddity is that, if all arguments to a function are presented at once, then erasure of optional arguments isn't applied until all of the arguments are passed in. Thus, this works:

```
# concat "a" "b" ~sep:"=";;
- : string = "a=b"
```

but a well-placed pair of parenthesis fails.

```
# (concat "a" "b") ~sep:"=";;
Characters 0-16:
  (concat "a" "b") ~sep:"=";;
  ^^^^^^^^^^^^^^^^
Error: This expression is not a function; it cannot be applied
```

The failure is a result of the fact that the expression (concat "a" "b") has erased concat's optional argument.

It's possible to define a function in such a way that the optional argument can never be erased, by having no positional arguments defined after the optional one. This leads to a compiler warning:

```
# let concat x y ?(sep="") = x ^ sep ^ y ;;
Characters 15-38:
  let concat x y ?(sep="") = x ^ sep ^ y ;;
                ^^^^^^^^^^^^^^^^^^^^^^^
Warning 16: this optional argument cannot be erased.
val concat : string -> string -> ?sep:string -> string = <fun>
```

When to use optional arguments

Optional arguments are very useful, but they're also easy to abuse. The key advantage of optional arguments is that they let you write functions with complex options that users can ignore most of the time, only needing to think about them when they specifically want to invoke those options.

The downside is that it's easy for the caller of a function to not be aware that there is a choice to be made, and as a result end up making the wrong choice by not doing anything. Optional arguments really only make sense when the extra concision of omitting the argument overwhelms the corresponding loss of explicitness.

This means that rarely used functions should not have optional arguments. A good rule of thumb for optional arguments is that you should never use an optional argument for internal functions of a module, only for functions that are exposed to users of a module.

Example: pretty-printing a table

One common programming task is displaying tabular data. In this example, will go over the design of a simple library to do just that.

We'll start with the interface. The code will go in a new module called `Text_table` whose `.mli` contains just the following function:

```
(* [render headers rows] returns a string containing a formatted
   text table, using unix-style newlines as separators *)
val render
   :  string list          (* header *)
   -> string list list     (* data *)
   -> string
```

If you invoke `render` as follows:

```
let () =
  print_string (Text_table.render
    ["language";"architect";"first release"]
    [ ["Lisp" ;"John McCarthy" ;"1958"] ;
      ["C"    ;"Dennis Ritchie";"1969"] ;
      ["ML"   ;"Robin Milner"  ;"1973"] ;
      ["OCaml";"Xavier Leroy"  ;"1996"] ;
    ])
```

you'll get the following output:

```
| language | architect      | first release |
|----------+----------------+---------------|
| Lisp     | John McCarthy  | 1958          |
| C        | Dennis Ritchie | 1969          |
| ML       | Robin Milner   | 1973          |
| OCaml    | Xavier Leroy   | 1996          |
```

Now that we know what `render` is supposed to do, let's dive into the implementation.

Computing the widths

To render the rows of the table, we'll first need the width of the widest entry in each column. The following function does just that.

```
let max_widths header rows =
  let to_lengths l = List.map ~f:String.length l in
  List.fold rows
    ~init:(to_lengths header)
    ~f:(fun acc row ->
      List.map2_exn ~f:Int.max acc (to_lengths row))
```

In the above we define a helper function, `to_lengths` which uses `List.map` and `String.length` to convert a list of strings to a list of string lengths. Then, starting with the lengths of the headers, we use `List.fold` to join in the lengths of the elements of each row by `max`'ing them together elementwise.

Note that this code will throw an exception if any of the rows has a different number of entries than the header. In particular, `List.map2_exn` throws an exception when its arguments have mismatched lengths.

Rendering the rows

Now we need to write the code to render a single row. There are really two different kinds of rows that need to be rendered; an ordinary row:

```
| Lisp     | John McCarthy  | 1962          |
```

and a separator row:

```
|----------+----------------+---------------|
```

Let's start with the separator row, which we can generate as follows:

```
let render_separator widths =
  let pieces = List.map widths
    ~f:(fun w -> String.make (w + 2) '-')
```

```
in
"|" ^ String.concat ~sep:"+" pieces ^ "|"
```

We need the extra two-characters for each entry to account for the one character of padding on each side of a string in the table.

Performance of `String.concat` and `^`

In the above, we're using two different ways of concatenating strings, `String.concat`, which operates on lists of strings, and `^`, which is a pairwise operator. You should avoid `^` for joining long numbers of strings, since, it allocates a new string every time it runs. Thus, the following code:

```
let s = "." ^ "."  ^ "."  ^ "."  ^ "."  ^ "."  ^ "."
```

will allocate a string of length 2, 3, 4, 5, 6 and 7, whereas this code:

```
let s = String.concat [".";".";".";".";".";".";"."]
```

allocates one string of size 7, as well as a list of length 7. At these small sizes, the differences don't amount to much, but for assembling of large strings, it can be a serious performance issue.

We can write a very similar piece of code for rendering the data in an ordinary row.

```
let pad s length =
  if String.length s >= length then s
  else s ^ String.make (length - String.length s) ' '

let render_row row widths =
  let pieces = List.map2 row widths
    ~f:(fun s width -> " " ^ pad s width ^ " ")
  in
  "|" ^ String.concat ~sep:"|" pieces ^ "|"
```

You might note that `render_row` and `render_separator` share a bit of structure. We can improve the code a bit by factoring that repeated structure out:

```
let decorate_row ~sep row = "|" ^ String.concat ~sep row ^ "|"

let render_row widths row =
  decorate_row ~sep:"|"
    (List.map2_exn row widths ~f:(fun s w -> " " ^ pad s w ^ " "))

let render_separator widths =
  decorate_row ~sep:"+"
    (List.map widths ~f:(fun width -> String.make (width + 2) '-'))
```

Bringing it all together

And now we can write the function for rendering a full table.

```
let render header rows =
  let widths = max_widths header rows in
  String.concat ~sep:"\n"
    (render_row widths header
     :: render_separator widths
     :: List.map rows ~f:(fun row -> render_row widths row)
    )
```

Now, let's think about how you might actually use this interface in practice. Usually, when you have data to render in a table, that data comes in the form of a list of objects of some sort, where you need to extract data from each record for each of the columns. So, imagine that you start off with a record type for representing information about a given programming language:

```
type style =
    Object_oriented | Functional | Imperative | Logic

type prog_lang = { name: string;
                   architect: string;
                   year_released: int;
                   style: style list;
                 }
```

If we then wanted to render a table from a list of languages, we might write something like this:

```
let print_langs langs =
    let headers = ["name";"architect";"year released"] in
    let to_row lang =
      [lang.name; lang.architect; Int.to_string lang.year_released ]
    in
    print_string (Text_table.render headers (List.map ~f:to_row langs))
```

This is OK, but as you consider more complicated tables with more columns, it becomes easier to make the mistake of having a mismatch in between headers and to_row. Also, adding, removing and reordering columns becomes awkward, because changes need to be made in two places.

We can improve the table API by adding a type which is a first-class representative for a column. We'd add the following to the interface of Text_table:

```
(** An ['a column] is a specification of a column for rending a table
    of values of type ['a] *)
type 'a column

(** [column header to_entry] returns a new column given a header and a
    function for extracting the text entry from the data associated
```

```
      with a row *)
val column : string -> ('a -> string) -> 'a column

(** [column_render columns rows] Renders a table with the specified
    columns and rows *)
val column_render :
  'a column list -> 'a list -> string
```

Thus, the `column` functions creates a `column` from a header string and a function for extracting the text for that column associated with a given row. Implementing this interface is quite simple:

```
type 'a column = string * ('a -> string)
let column header to_string = (header,to_string)

let column_render columns rows =
  let header = List.map columns ~f:fst in
  let rows = List.map rows ~f:(fun row ->
    List.map columns ~f:(fun (_,to_string) -> to_string row))
  in
  render header rows
```

And we can rewrite `print_langs` to use this new interface as follows.

```
let columns =
  [ Text_table.column "Name"      (fun x -> x.name);
    Text_table.column "Architect" (fun x -> x.architect);
    Text_table.column "Year Released"
        (fun x -> Int.to_string x.year_released);
  ]

let print_langs langs =
  print_string (Text_table.column_render columns langs)
```

The code is a bit longer, but it's also less error prone. In particular, several errors that might be made by the user are now ruled out by the type system. For example, it's no longer possible for the length of the header and the lengths of the rows to be mismatched.

The simple column-based interface described here is also a good starting for building a richer API. You could for example build specialized colums with different formatting and alignment rules, which is easier to do with this interface than with the original one based on passing in lists-of-lists.

Lists, Options and Pattern Matching

Error Handling

Nobody likes dealing with errors. It's tedious, it's easy to get wrong, and it's usually just not as fun as planning out how your program is going to succeed. But error handling is important, and however much you don't like thinking about it, having your software fail due to poor error handling code is worse.

Thankfully, OCaml has powerful tools for handling errors reliably and with a minimum of pain. In this chapter we'll discuss some of the different approaches in OCaml to handling errors, and give some advice on how to design interfaces that help rather than hinder error handling.

We'll start by describing the two basic approaches for reporting errors in OCaml: error-aware return types and exceptions.

Error-aware return types

The best way in OCaml to signal an error is to include that error in your return value. Consider the type of the `find` function in the list module.

```
# List.find;;
- : 'a list -> f:('a -> bool) -> 'a option
```

The option in the return type indicates that the function may not succeed in finding a suitable element, as you can see below.

```
# List.find [1;2;3] ~f:(fun x -> x >= 2) ;;
- : int option = Some 2
# List.find [1;2;3] ~f:(fun x -> x >= 10) ;;
- : int option = None
```

Having errors be explicit in the return values of your functions tells the caller that there is an error that needs to be handled. The caller can then handle the error explicitly, either recovering from the error or propagating it onward.

The function `compute_bounds` below is an example of how you can handle errors in this style. The function takes a list and a comparison function, and returns upper and lower bounds for the list by finding the smallest and largest element on the list. `List.hd` and `List.last`, which return `None` when they encounter an empty list, are used to extract the largest and smallest element of the list.

```
# let compute_bounds ~cmp list =
    let sorted = List.sort ~cmp list in
    let smallest = List.hd sorted in
    let largest = List.last sorted in
    match smallest, largest with
    | None,_ | _, None -> None
    | Some x, Some y -> Some (x,y)
  ;;
val compute_bounds :
  cmp:('a -> 'a -> int) -> 'a list -> ('a * 'a) option = <fun>
```

The match statement is used to handle the error cases, propagating an error in `hd` or `last` into the return value of `compute_bounds`. On the other hand, in `find_mismatches` below, errors encountered during the computation do not propagate to the return value of the function. `find_mismatches` takes two hashtables as its aruguments and tries to find keys that are stored in both. As such, a failure to find a key in one of the tables isn't really an error.

```
# let find_mismatches table1 table2 =
    Hashtbl.fold table1 ~init:[] ~f:(fun ~key ~data errors ->
      match Hashtbl.find table2 key with
      | Some data' when data' <> data -> key :: errors
      | _ -> errors
    )
  ;;
val find_mismatches :
  ('a, 'b) Hashtbl.t -> ('a, 'b) Hashtbl.t -> 'a list = <fun>
```

The use of options to encode errors underlines the fact that it's not clear whether a particular outcome, like not finding something on a list, is really an error, or just another valid outcome of your function. This turns out to be very context-dependent, and error-aware return types give you a uniform way of handling the result that works well for both situations.

Encoding errors with `Result`

Options aren't always a sufficiently expressive way to report errors. Specifically, when you encode an error as `None`, there's nowhere to say anything about the nature of the error.

`Result.t` is meant to address this deficiency. Here's the definition:

```
module Result : sig
```

```
    type ('a,'b) t = | Ok of 'a
                     | Error of 'b
end
```

A `Result.t` is essentially an option augmented with the ability to store other information in the error case. Like `Some` and `None` for options, the constructors `Ok` and `Error` are promoted to the top-level by `Core.Std`. As such, we can write:

```
# [ Ok 3; Error "abject failure"; Ok 4 ];;
[Ok 3; Error "abject failure"; Ok 4]
- : (int, string) Result.t list =
[Ok 3; Error "abject failure"; Ok 4]
```

without first opening the `Result` module.

`Error` and `Or_error`

`Result.t` gives you complete freedom to choose the type of value you use to represent errors, but it's often useful to standardize on an error type. Among other things, this makes it easier to write utility functions to automate common error handling patterns.

But which type to choose? Is it better to represent errors as strings? Or S-expressions? Or something else entirely?

Core's answer to this question is the `Error.t` type, which tries to forge a good compromise between efficiency, convenience and control over the presentation of errors.

It might not be obvious at first why efficiency is an issue at all. But generating error messages is an expensive business. An ASCII representation of a type can be quite time-consuming to construct, particularly if it includes expensive-to-convert numerical datatypes.

`Error` gets around this issue through laziness. In particular, an `Error.t` allows you to put off generation of the actual error string until you actually need, which means a lot of the time you never have to construct it at all. You can of course construct an error directly from a string:

```
# Error.of_string "something went wrong";;
- : Core.Std.Error.t = "something went wrong"
```

A more interesting construction message from a performance point of view is to construct an `Error.t` from a thunk:

```
# Error.of_thunk (fun () ->
    sprintf "something went wrong: %f" 32.3343);;
  - : Core.Std.Error.t = "something went wrong: 32.334300"
```

In this case, we can benefit from the laziness of `Error`, since the thunk won't be called until the `Error.t` is converted to a string.

We can also create an `Error.t` based on an s-expression converter. This is probably the most common idiom in Core.

```
# Error.create "Something failed a long time ago" Time.epoch Time.sexp_of_t;;
- : Core.Std.Error.t =
"Something failed a long time ago: (1969-12-31 19:00:00.000000)"
```

Here, the value `Time.epoch` is included in the error, but `Time.sexp_of_t`, which is used for converting the time to an s-expression, isn't run until the error is converted to a string. Using the Sexplib syntax-extension, which is discussed in more detail in chapter {{SYNTAX}}, we can inline create an s-expression converter for a collection of types, thus allowing us to register multiple pieces of data in an `Error.t`.

```
# Error.create "Something went terribly wrong"
    (3.5, ["a";"b";"c"],6034)
    <:sexp_of<float * string list * int>> ;;
- : Core.Std.Error.t = "Something went terribly wrong: (3.5(a b c)6034)"
```

Here, the declaration `<:sexp_of<float * string list * int>>` asks Sexplib to generate the sexp-converter for the tuple.

Error also has operations for transforming errors. For example, it's often useful to augment an error with some extra information about the context of the error, or to combine multipler errors together. `Error.of_list` and `Error.tag` fill these roles.

The type `'a Or_error.t` is just a shorthand for `('a,Error.t) Result.t`, and it is, after `option`, the most common way of returning errors in Core.

`bind` and other error-handling idioms

As you write more error handling code, you'll discover that certain patterns start to emerge. A number of these common patterns been codified in the interfaces of modules like `Option` and `Result`. One particularly useful one is built around the function `bind`, which is both an ordinary function and an infix operator `>>=`, both with the same type signature:

```
val (>>=) : 'a option -> ('a -> 'b option) -> 'b option
```

`bind` is a way of sequencing together error-producing functions so that that the first one to produce an error terminates the computation. In particular, `None >>= f` returns `None` without calling `f`, and `Some x >>= f` returns `f x`. We can use a nested sequence of these binds to express a multi-stage computation that can fail at any stage. Here's a rewrite `compute_bounds` in this style.

```
# let compute_bounds ~cmp list =
    let open Option.Monad_infix in
    let sorted = List.sort ~cmp list in
    List.hd sorted >>= (fun first ->
```

```
      List.last sorted >>= (fun last ->
        Some (first,last)))
```

Note that we locally open the `Option.Monad_infix` module to get access to the infix operator `>>=`. The module is called `Monad_infix` because the bind operator is part of a sub-interface called `Monad`, which we'll talk about more in chapter {{{ASYNC}}}.

This is a bit easier to read if we write it with fewer parens and less indentation, as follows.

```
# let compute_bounds ~cmp list =
    let open Option.Monad_infix in
    let sorted = List.sort ~cmp list in
    List.hd sorted   >>= fun first ->
    List.last sorted >>= fun last  ->
    Some (first,last)
```

There are other usuful idioms encoded in the functions in `Option`. Another example is `Option.both`, which takes two optional values and produces a new optional pair that is `None` if either of its arguments are `None`. Using `Option.both`, we can make `com pute_bounds` even shorter.

```
# let compute_bounds ~cmp list =
    let sorted = List.sort ~cmp list in
    Option.both (List.hd sorted) (List.last sorted)
```

These error-handling functions are valuable because they let you express your error handling both explicitly and concisely. We've only discussed these functions in the context of the `Option` module, but similar functionality is available in both `Result` and `Or_error`.

Exceptions

Exceptions in OCaml are not that different from exceptions in many other languages, like Java, C# and Python. In all these cases, exceptions are a way to terminate a computation and report an error, while providing a mechanism to catch and handle (and possibly recover from) exceptions that are triggered by sub-computations.

We'll see an exception triggered in OCaml if, for example, we try to divide an integer by zero:

```
# 3 / 0;;
Exception: Division_by_zero.
```

And an exception can terminate a computation even if it happens nested a few levels deep in a computation.

```
# List.map ~f:(fun x -> 100 / x) [1;3;0;4];;
Exception: Division_by_zero.
```

In addition to built-in exceptions like `Divide_by_zero`, OCaml lets you define your own.

```
# exception Key_not_found of string;;
exception Key_not_found of string
# Key_not_found "a";;
- : exn = Key_not_found("a")
```

Here's an example of a function for looking up a key in an *association list*, *i.e.* a list of key/value pairs which uses this newly-defined exception:

```
# let rec find_exn alist key = match alist with
    | [] -> raise (Key_not_found key)
    | (key',data) :: tl -> if key = key' then data else find_exn tl key
  ;;
val find_exn : (string * 'a) list -> string -> 'a = <fun>
# let alist = [("a",1); ("b",2)];;
val alist : (string * int) list = [("a", 1); ("b", 2)]
# find_exn alist "a";;
- : int = 1
# find_exn alist "c";;
Exception: Key_not_found("c").
```

Note that we named the function `find_exn` to warn the user that the function routinely throws exceptions, a convention that is used heavily in Core.

In the above example, `raise` throws the exception, thus terminating the computation. The type of raise is a bit surprising when you first see it:

```
# raise;;
- : exn -> 'a = <fun>
```

Having the return type be an otherwise unused type variable `'a` suggests that `raise` could return a value of any type. That seems impossible, and it is. `raise` has this type because it never returns at all. This behavior isn't restricted to functions like `raise` that terminate by throwing exceptions. Here's another example of a function that doesn't return a value.

```
# let rec forever () = forever ();;
val forever : unit -> 'a = <fun>
```

`forever` doesn't return a value for a different reason: it is an infinite loop.

This all matters because it means that the return type of `raise` can be whatever it needs to be to fit in to the context it is called in. Thus, the type system will let us throw an exception anywhere in a program.

Declaring exceptions with `with sexp`

OCaml can't always generate a useful textual representation of your exception, for example:

```
# exception Wrong_date of Date.t;;
exception Wrong_date of Date.t
# Wrong_date (Date.of_string "2011-02-23");;
- : exn = Wrong_date(_)
```

But if you declare the exception using `with sexp` (and the constituent types have sexp converters), we'll get something with more information.

```
# exception Wrong_date of Date.t with sexp;;
exception Wrong_date of Core.Std.Date.t
# Wrong_date (Date.of_string "2011-02-23");;
- : exn = (.Wrong_date 2011-02-23)
```

The period in front of `Wrong_date` is there because the representation generated by `with sexp` includes the full module path of the module where the exception in question is defined. This is quite useful in tracking down which precise exception is being reported. In this case, since we've declared the exception at the toplevel, that module path is trivial.

Exception handlers

So far, we've only seen exceptions fully terminate the execution of a computation. But often, we want a program to be able to respond to and recover from an exception. This is achieved through the use of *exception handlers*.

In OCaml, an exception handler is declared using a `try`/`with` statement. Here's the basic syntax.

```
try <expr> with
| <pat1> -> <expr1>
| <pat2> -> <expr2>
...
```

A `try`/`with` clause would first evaluate `<expr>`, and if that evaluation completes without returning an exception, then the value of the overall expression is the value of `<expr>`.

But if evaluating `<expr>` leads to an exception being thrown, then the exception will be fed to the pattern match statements following the `with`. If the exception matches a pattern, then the expression on the right hand side of that pattern will be evaluated. Otherwise, the original exception continues up the call stack, to be handled by the next outer exception handler, or terminate the program if there is none.

Cleaning up in the presence of exceptions

One headache with exceptions is that they can terminate your execution at unexpected places, leaving your program in an awkward state. Consider the following code snippet:

```
let load_config filename =
```

```
let inc = In_channel.create filename in
let config = Config.t_of_sexp (Sexp.input_sexp inc) in
In_channel.close inc;
config
```

The problem with this code is that the function that loads the s-expression and parses it into a `Config.t` might thrown an exception if the config file in question is malformed. Unfortunately, that means that the `In_channel.t` that was opened will never be closed, leading to a file-descriptor leak.

We can fix this using Core's `protect` function. The basic purpose of `protect` is to ensure that the `finally` thunk will be called when `f` exits, whether it exited normally or with an exception. Here's how it could be used to fix `load_config`.

```
let load_config filename =
  let inc = In_channel.create filename in
  protect ~f:(fun () -> Config.t_of_sexp (Sexp.input_sexp inc))
    ~finally:(fun () -> In_channel.close inc)
```

Catching specific exceptions

OCaml's exception-handling system allows you to tune your error-recovery logic to the particular error that was thrown. For example, `List.find_exn` always throws `Not_found`. You can take advantage of this in your code, for example, let's define a function called `lookup_weight`, with the following signature:

```
(** [lookup_weight ~compute_weight alist key] Looks up a
    floating-point weight by applying [compute_weight] to the data
    associated with [key] by [alist].  If [key] is not found, then
    return 0.
*)
val lookup_weight :
  compute_weight:('data -> float) -> ('key * 'data) list -> 'key -> float
```

We can implement such a function using exceptions as follows:

```
# let lookup_weight ~compute_weight alist key =
    try
      let data = List.Assoc.find_exn alist key in
      compute_weight data
    with
      Not_found -> 0. ;;
val lookup_weight :
  compute_weight:('a -> float) -> ('b * 'a) list -> 'b -> float =
  <fun>
```

This implementation is more problematic than it looks. In particular, what happens if `compute_weight` itself throws an exception? Ideally, `lookup_weight` should propagate that exception on, but if the exception happens to be `Not_found`, then that's not what will happen:

```
# lookup_weight ~compute_weight:(fun _ -> raise Not_found)
    ["a",3; "b",4] "a" ;;
- : float = 0.
```

This kind of problem is hard to detect in advance, because the type system doesn't tell us what kinds of exceptions a given function might throw. Because of this kind of confusion, it's usually better to avoid catching specific exceptions. In this case, we can improve the code by catching the exception in a narrower scope.

```
# let lookup_weight ~compute_weight alist key =
    match
      try Some (List.Assoc.find_exn alist key) with
      | Not_found -> None
    with
    | None -> 0.
    | Some data -> compute_weight data ;;
```

At which point, it makes sense to simply use the non-exception throwing function, `List.Assoc.find`, instead.

Backtraces

A big part of the point of exceptions is to give useful debugging information. But at first glance, OCaml's exceptions can be less than informative. Consider the following simple program.

```
(* exn.ml *)

open Core.Std
exception Empty_list

let list_max = function
  | [] -> raise Empty_list
  | hd :: tl -> List.fold tl ~init:hd ~f:(Int.max)

let () =
  printf "%d\n" (list_max [1;2;3]);
  printf "%d\n" (list_max [])
```

If we build and run this program, we'll get a pretty uninformative error:

```
$ ./exn
3
Fatal error: exception Exn.Empty_list
```

The example in qeustion is short enough that it's quite easy to see where the error came from. But in a complex program, simply knowing which exception was thrown is usually not enough information to figure out what went wrong.

We can get more information from OCaml if we turn on stack traces. This can be done by setting the `OCAMLRUNPARAM` environment variable, as shown:

```
exn $ export OCAMLRUNPARAM=b
exn $ ./exn
3
Fatal error: exception Exn.Empty_list
Raised at file "exn.ml", line 7, characters 16-26
Called from file "exn.ml", line 12, characters 17-28
```

Backtraces can also be obtained at runtime. In particular, `Exn.backtrace` will return the backtrace fo the most recently thrown exception.

Exceptions for control flow

From exceptions to error-aware types and back again

Both exceptions and error-aware types are necessary parts of programming in OCaml. As such, you often need to move between these two worlds. Happily, Core comes with some useful helper functions to help you do just that. For example, given a piece of code that can throw an exception, you can capture that exception into an option as follows:

```
# let find alist key =
    Option.try_with (fun () -> find_exn alist key) ;;
val find : (string * 'a) list -> string -> 'a option = <fun>
# find ["a",1; "b",2] "c";;
- : int Core.Std.Option.t = None
# find ["a",1; "b",2] "b";;
- : int Core.Std.Option.t = Some 2
```

And `Result` and `Or_error` have similar `try_with` functions. So, we could write:

```
# let find alist key =
    Result.try_with (fun () -> find_exn alist key) ;;
val find : (string * 'a) list -> string -> ('a, exn) Result.t = <fun>
# find ["a",1; "b",2] "c";;
- : (int, exn) Result.t = Result.Error Key_not_found("c")
```

And then we can re-raise that exception:

```
# Result.ok_exn (find ["a",1; "b",2] "b");;
- : int = 2
# Result.ok_exn (find ["a",1; "b",2] "c");;
Exception: Key_not_found("c").
```

Files, Modules and Programs

We've so far experienced OCaml only through the toplevel. As you move from exercises to real-world programs, you'll need to leave the toplevel behind and start building programs from files. Files are more than just a convenient way to store and manage your code; in OCaml, they also act as abstraction boundaries that divide your program into conceptual components.

In this chapter, we'll show you how to build an OCaml program from a collection of files, as well as the basics of working with modules and module signatures.

Single File Programs

We'll start with an example: a utility that reads lines from `stdin`, computing a frequency count of the lines that have been read in. At the end, the 10 lines with the highest frequency counts are written out. Here's a simple implementation, which we'll save as the file `freq.ml`. Note that we're using several functions from the `List.Assoc` module, which provides utility functions for interacting with association lists, *i.e.*, lists of key/value pairs.

```
(* file.ml: basic implementation *)

open Core.Std

(* build_counts recursively builds up a mapping from lines to
   number of occurences of that line. *)
let rec build_counts counts =
  match In_channel.input_line stdin with
  | None -> counts (* EOF, so return the counts accummulated so far *)
  | Some line ->
    (* get the number of times this line has been seen before,
       inferring 0 if the line doesn't show up in [counts] *)
    let count =
      match List.Assoc.find counts line with
      | None -> 0
      | Some x -> x
```

```
      in
      (* increment the count for line by 1, and recurse *)
      build_counts (List.Assoc.add counts line (count + 1))

  let () =
    (* Compute the line counts *)
    let counts = build_counts [] in
    (* Sort the line counts in descending order of frequency *)
    let sorted_counts = List.sort ~cmp:(fun (_,x) (_,y) -> descending x y) counts  in
    (* Print out the 10 highest frequency entries *)
    List.iter (List.take 10 sorted_counts) ~f:(fun (line,count) ->
      printf "%3d: %s\n" count line)
```

Where is the main function?

Unlike C, programs in OCaml do not have a unique `main` function. When an OCaml program is evaluated, all the statements in the implementation files are evaluated in order. These implementation files can contain arbitrary expressions, not just function definitions. In this example, the role of the `main` function is played by the expression `let () = process_lines []`, which kicks off the actions of the program. But really the entire file is evaluated at startup, and so in some sense the full codebase is one big `main` function.

If we weren't using Core or any other external libraries, we could build the executable like this:

```
ocamlc freq.ml -o freq
```

But in this case, this command will fail with the error `Unbound module Core`. We need a somewhat more complex invocation to get Core linked in:

```
ocamlfind ocamlc -linkpkg -thread -package core freq.ml -o freq
```

Here we're using `ocamlfind`, a tool which itself invokes other parts of the ocaml toolchain (in this case, `ocamlc`) with the appropriate flags to link in particular libraries and packages. Here, `-package core` is asking `ocamlfind` to link in the Core library, `-linkpkg` is required to do the final linking in of packages for building a runnable executable, and `-thread` turns on threading support, which is required for Core.

While this works well enough for a one-file project, more complicated builds will require a tool to orchestrate the build. One great tool for this task is `ocamlbuild`, which is shipped with the OCaml compiler. We'll talk more about `ocamlbuild` in chapter {{{OCAMLBUILD}}}, but for now, we'll just walk through the steps required for this simple application. First, create a `_tags` file, containing the following lines.

```
true:package(core)
true:thread,annot,debugging
```

The purpose of the `_tags` file is to specify which compilation options are required for which files. In this case, we're telling `ocamlbuild` to link in the `core` package and to turn on threading, output of annotation files, and debugging support for all files (the pattern `true` matches every file in the project.)

We then create a build script `build.sh` that invokes `ocamlbuild`:

```
#!/usr/bin/env bash

TARGET=freq
ocamlbuild -use-ocamlfind $TARGET.byte && cp $TARGET.byte $TARGET
```

If you invoke `build.sh`, you'll get a bytecode executable. If we'd used a target of `unique.native` in `build.sh`, we would have gotten native-code instead.

Whichever way you build the application, you can now run it from the command-line. The following line extracts strings from the `ocamlopt` executable, and then reports the most frequently occurring ones.

```
$ strings `which ocamlopt` | ./freq
 13: movq
 10: cmpq
  8: ", &
  7: .globl
  6: addq
  6: leaq
  5: ", $
  5: .long
  5: .quad
  4: ", '
```

Byte-code vs native-code

OCaml ships with two compilers---the `ocamlc` byte-code compiler, and the `ocamlopt` native-code compiler. Programs compiled with `ocamlc` are interpreted by a virtual machine, while programs compiled with `ocamlopt` are compiled to native machine code to be run on a specific operating system and processor architecture.

Aside from performance, executables generated by thet two compilers have nearly identical behavior. There are a few things to be aware of. First, the byte-code compiler can be used on more architectures, and has some better tool support; in particular, the OCaml debugger only works with byte-code. Also, the byte-code compiler compiles faster than the native code compiler.

As a general matter, production executables should usually be built using the native-code compiler, but it sometimes makes sense to use bytecode for development builds. And, of course, bytecode makese sense when targetting a platform not supported by the native code compiler.

Multi-file programs and modules

Source files in OCaml are tied into the module system, with each file compiling down into a module whose name is derived from the name of the file. We've encountered modules before, for example, when we used functions like `find` and `add` from the `List.Assoc` module. At it's simplest, you can think of a module as a collection of definitions that are stored within a namespace.

Let's consider how we can use modules to refactor the implementation of `freq.ml`. Remember that the variable `counts` contains an association list representing the counts of the lines seen so far. But updating an association list takes time linear in the length of the list, meaning that the time complexity of processing a file is quadratic in the number of distinct lines in the file.

We can fix this problem by replacing association lists with a more efficient datastructure. To do that, we'll first factor out the key functionality into a separate module with an explicit interface. We can consider alternative (and more efficient) implementations once we have a clear interface to program against.

We'll start by creating a file, `counter.ml`, that contains the logic for maintaining the association list used to describe the counts. The key function, called `touch`, updates the association list with the information that a given line should be added to the frequency counts.

```
(* counter.ml: first version *)

open Core.Std

let touch t s =
  let count =
    match List.Assoc.find t s with
    | None -> 0
    | Some x -> x
  in
  List.Assoc.add t s (count + 1)
```

We can now rewrite `freq.ml` to use `Counter`. Note that the resulting code can still be built with `build.sh`, since `ocambuild` will discover dependencies and realize that `counter.ml` needs to be compiled.

```
(* freq.ml: using Counter *)

open Core.Std

let rec build_counts counts =
  match In_channel.input_line stdin with
  | None -> counts
  | Some line -> build_counts (Counter.touch counts line)

let () =
```

```
let counts = build_counts [] in
let sorted_counts = List.sort counts
  ~cmp:(fun (_,x) (_,y) -> Int.descending x y)
in
List.iter (List.take sorted_counts 10)
  ~f:(fun (line,count) -> printf "%3d: %s\n" count line)
```

Signatures and Abstract Types

While we've pushed some of the logic to the `Counter` module, the code in `freq.ml` can still depend on the details of the implementation of `Counter`. Indeed, if you look at the invocation of `build_counts`:

```
let counts = build_counts [] in
```

you'll see that it depends on the fact that the empty set of frequency counts is represented as an empty list. We'd like to prevent this kind of dependency, so that we can change the implementation of `Counter` without needing to change client code like that in `freq.ml`.

The first step towards hiding the implementation details of `Counter` is to create an interface file, `counter.mli`, which controls how `counter` is accessed. Let's start by writing down a simple descriptive interface, *i.e.*, an interface that describes what's currently available in `Counter` without hiding anything. We'll use `val` declarations in the `mli`, which have the following syntax

```
val <identifier> : <type>
```

and are used to expose the existence of a given value in the module. Here's an interface that describes the current contents of `Counter`. We can save this as `counter.mli` and compile, and the program will build as before.

```
(* counter.mli: descriptive interface *)

val touch : (string * int) list -> string -> (string * int) list
```

To actually hide the fact that frequency counts are represented as association lists, we need to make the type of frequency counts *abstract*. A type is abstract if its name is exposed in the interface, but its definition is not. Here's an abstract interface for `Counter`:

```
(* counter.mli: abstract interface *)

open Core.Std

type t

val empty : t
val to_list : t -> (string * int) list
val touch : t -> string -> t
```

Note that we needed to add `empty` and `to_list` to `Counter`, since otherwise, there would be no way to create a `Counter.t` or get data out of one.

Here's a rewrite of `counter.ml` to match this signature.

```
(* counter.ml: implementation matching abstract interface *)

open Core.Std

type t = (string * int) list

let empty = []

let to_list x = x

let touch t s =
  let count =
    match List.Assoc.find t s with
    | None -> 0
    | Some x -> x
  in
  List.Assoc.add t s (count + 1)
```

If we now try to compile `freq.ml`, we'll get the following error:

```
File "freq.ml", line 11, characters 20-22:
Error: This expression has type 'a list
       but an expression was expected of type Counter.t
```

This is because `freq.ml` depends on the fact that frequency counts are represented as association lists, a fact that we've just hidden. We just need to fix the code to use `Counter.empty` instead of `[]` and `Counter.to_list` to get the association list out at the end for processing and printing.

Now we can turn to optimizing the implementation of `Counter`. Here's an alternate and far more efficient implementation, based on the `Map` datastructure in Core.

```
(* counter.ml: efficient version *)

open Core.Std

type t = (string,int) Map.t

let empty = Map.empty

let touch t s =
  let count =
    match Map.find t s with
    | None -> 0
    | Some x -> x
  in
  Map.add t s (count + 1)
```

```
let to_list t = Map.to_alist t
```

More on modules and signatures

Concrete types in signatures

In our frequency-count example, the module `Counter` had an abstract type `Counter.t` for representing a collection of frequency counts. Sometimes, you'll want to make a type in your interface *concrete*, by including the type definition in the interface.

For example, imagine we wanted to add a function to `Counter` for returning the line with the median frequency count. If the number of lines is even, then there is no precise median, so the function would return the two lines before and after the median instead. We'll use a custom type to represent the fact that there are two possible possible return values. Here's a possible implementation.

```
type median = | Median of string
              | Before_and_after of string * string

let median t =
  let sorted_strings = List.sort (Map.to_alist t)
      ~cmp:(fun (_,x) (_,y) -> Int.descending x y)
    in
  let len = List.length sorted_strings in
  if len = 0 then failwith "median: empty frequency count";
  let nth n = List.nth_exn sorted_strings n in
  if len mod 2 = 1
  then Median (nth (len/2))
  else Before_and_after (nth (len/2) (len/2 + 1))
```

Now, to expose this usefully in the interface, we need to expose both the function and the type `median` with its definition. We'd do that by adding these lines to the `counter.mli`:

```
type median = | Median of string
              | Before_and_after of string * string

val get_median : t -> median
```

The decision of whether a given type should be abstract or concrete is an important one. Abstract types give you more control over how values are created and accessed, and makes it easier to enforce invariants beyond the what's enforced by the type itself; concrete types let you expose more detail and structure to client code in a lightweight way. The right choice depends very much on the context.

The `include` directive

OCaml provides a number of tools for manipulating modules. One particularly useful one is the `include` directive, which is used to include the contents of one module into another.

One natural application of `include` is to create one module which is an extension of another one. For example, imagine you wanted to build an extended version of the `List` module, where you've added some functionality not present in the module as distributed in Core. We can do this easily using `include`:

```
(* ext_list.ml: an extended list module *)

open Core.Std

(* The new function we're going to add *)
let rec intersperse list el =
  match list with
  | [] | [ _ ]   -> list
  | x :: y :: tl -> x :: el :: intersperse (y::tl) el

(* The remainder of the list module *)
include List
```

Now, what about the interface of this new module? It turns out that include works on the signature language as well, so we can pull essentially the same trick to write an `mli` for this new module. The only trick is that we need to get our hands on the signature for the list module, which can be done using `module type of`.

```
(* ext_list.mli: an extended list module *)

open Core.Std

(* Include the interface of the list module from Core *)
include (module type of List)

(* Signature of function we're adding *)
val intersperse : 'a list -> 'a -> 'a list
```

And we can now use `Ext_list` as a replacement for `List`. If we want to use `Ext_list` in preference to `List` in our project, we can create a file of common definitions:

```
(* common.ml *)

module List = Ext_list
```

And if we then put `open Common` after `open Core.Std` at the top of each file in our project, then references to `List` will automatically go to `Ext_list` instead.

Modules within a file

Up until now, we've only considered modules that correspond to files, like `counter.ml`. But modules (and module signatures) can be nested inside other modules. As a simple example, consider a program that needs to deal with some class of identifier like a username. Rather than just keeping usernames as strings, you might want to mint an abstract type, so that the type-system will help you to not confuse usernames with other string data that is floating around your program.

Here's how you might create such a type, within a module:

```
open Core.Std

module Username : sig
  type t
  val of_string : string -> t
  val to_string : t -> string
end = struct
  type t = string
  let of_string x = x
  let to_string x = x
end
```

The basic structure of a module declaration like this is:

```
module <name> : <signature> = <implementation>
```

We could have written this slightly differently, by giving the signature its own top-level `module type` declaration, making it possible to in a lightweight way create multiple distinct types with the same underlying implementation.

```
module type ID = sig
  type t
  val of_string : string -> t
  val to_string : t -> string
end

module String_id = struct
  type t = string
  let of_string x = x
  let to_string x = x
end

module Username : ID = String_id
module Hostname : ID = String_id

(* Now the following buggy code won't compile *)
type session_info = { user: Username.t;
                      host: Hostname.t;
                      when_started: Time.t;
                    }
```

```
let sessions_have_same_user s1 s2 =
  s1.user = s1.host
```

Opening modules

One useful primitive in OCaml's module language is the open directive. We've seen that already in the open `Core.Std` that has been at the top of our source files.

The basic purpose of open is to extend the namespaces that OCaml searches when trying to resolve an identifier. Roughly, if you open a module M, then every subsequent time you look for an identifier foo, the module system will look in M for a value named foo. This is true for all kinds of identifiers, including types, type constructors, values and modules.

open is essential when dealing with something like a standard library, but it's generally good style to keep opening of modules to a minimum. Opening a module is basically a tradeoff between terseness and explicitness - the more modules you open, the harder it is to look at an identifier and figure out where it's defined.

Here's some general advice on how to deal with opens.

- Opening modules at the top-level of a module should be done quite sparingly, and generally only with modules that have been specifically designed to be opened, like `Core.Std` or `Option.Monad_infix`.

- One alternative to local opens that makes your code terser without giving up on explicitness is to locally rebind the name of a module. So, instead of writing:

```
let print_median m =
  match m with
  | Counter.Median string -> printf "True median:\n    %s\n"
  | Counter.Before_and_after of before * after ->
    printf "Before and after median:\n   %s\n   %s\n" before after
```

  you could write

```
let print_median m =
  let module C = Counter in
  match m with
  | C.Median string -> printf "True median:\n    %s\n"
  | C.Before_and_after of before * after ->
    printf "Before and after median:\n   %s\n   %s\n" before after
```

  Because the module name C only exists for a short scope, it's easy to read and remember what C stands for. Rebinding modules to very short names at the top-level of your module is usually a mistake.

- If you do need to do an open, it's better to do a *local open*. There are two syntaxes for local opens. For example, you can write:

```
let average x y =
  let open Int64 in
  x + y / of_int 2
```

In the above, `of_int` and the infix operators are the ones from `Int64` module.

There's another even more lightweight syntax for local opens, which is particularly useful for small expressions:

```
let average x y =
  Int64.(x + y / of_int 2)
```

Common errors with modules

When OCaml compiles a program with an `ml` and an `mli`, it will complain if it detects a mismatch between the two. Here are some of the common errors you'll run into.

Type mismatches

The simplest kind of error is where the type specified in the signature does not match up with the type in the implementation of the module. As an example, if we replace the `val` declaration in `counter.mli` by swapping the types of the first two arguments:

```
val touch : string -> t -> t
```

and then try to compile `Counter` (by writing `ocamlbuild -use-ocamlfind counter.cmo`), we'll ge the following error:

```
File "counter.ml", line 1, characters 0-1:
Error: The implementation counter.ml
       does not match the interface counter.cmi:
       Values do not match:
         val touch :
           ('a, int) Core.Std.Map.t -> 'a -> ('a, int) Core.Std.Map.t
       is not included in
         val touch : string -> t -> t
```

This error message is a bit intimidating at first, and it takes a bit of thought to see where the first type, which is the type of [touch] in the implementation, doesn't match the second one, which is the type of [touch] in the interface. You need to recognize that [t] is in fact a [Core.Std.Map.t], and the problem is that in the first type, the first argument is a map while the second is the key to that map, but the order is swapped in the second type.

Missing definitions

We might decide that we want a new function in `Counter` for pulling out the frequency count of a given string. We can update the `mli` by adding the following line.

```
val count : t -> string -> int
```

Now, if we try to compile without actully adding the implementation, we'll get this
error:

```
File "counter.ml", line 1, characters 0-1:
Error: The implementation counter.ml
       does not match the interface counter.cmi:
       The field `count' is required but not provided
```

A missing type definition will lead to a similar error.

Type definition mismatches

Type definitions that show up in an `mli` need to match up with corresponding defini-
tions in the `ml`. Consider again the example of the type `median`. The order of the decla-
ration of variants matters to the OCaml compiler so, if the definition of `median` in the
implementation lists those options in a different order:

```
type median = | Before_and_after of line * line
              | Median of line
```

that will lead to a compilation error:

```
File "counter.ml", line 1, characters 0-1:
Error: The implementation counter.ml
       does not match the interface counter.cmi:
       Type declarations do not match:
         type median = Before_and_after of string * string | Median of string
       is not included in
         type median = Median of string | Before_and_after of string * string
       Their first fields have different names, Before_and_after and Median.
```

Order is similarly important in other parts of the signature, including the order in which
record fields are declared and the order of arguments (including labelled and optional
arguments) to a function.

Cyclic dependencies

In most cases, OCaml doesn't allow circular dependencies, *i.e.*, a collection of defini-
tions that all refer to each other. If you want to create such definitions, you typically
have to mark them specially. For example, when defining a set of mutually recursive
values, you need to define them using `let rec` rather than ordinary `let`.

The same is true at the module level. By default, circular dependencies between mod-
ules is not allowed, and indeed, circular dependencies among files is never allowed.

The simplest case of this is that a module can not directly refer to itself (although definitions within a module can refer to each other in the ordinary way). So, if we tried to add a reference to Counter from within counter.ml:

```
let singleton l = Counter.touch Counter.empty
```

then when we try to build, we'll get this error:

```
File "counter.ml", line 17, characters 18-31:
Error: Unbound module Counter
Command exited with code 2.
```

The problem manifests in a different way if we create circular references between files. We could create such a situation by adding a reference to Freq from counter.ml, *e.g.*, by adding the following line:

```
let build_counts = Freq.build_counts
```

In this case, ocamlbuild will notice the error and complain:

```
Circular dependencies: "freq.cmo" already seen in
  [ "counter.cmo"; "freq.cmo" ]
```

Functors and First-class modules

Up until now, modules have played a limited role, used as the way for organizing the system into different components with well-defined interfaces. This is of course important, but it is only part of what modules can do. OCaml's module system is part of a powerful toolset for organizing and putting together complex functionality.

Functors

A functor in OCaml is essentially a function from modules to modules. Let's walk through a small, complete example of how to use functors. In this case, we're going to show how you can use a functor build up a larger module interface from a few key components. First, let's start with the interface, in particular, an interface for something that supports the comparison operator.

```
module type Comparable = sig
  type t
  val compare : t -> t -> int
  val (=)  : t -> t -> bool
  val (<>) : t -> t -> bool
  val (>)  : t -> t -> bool
  val (<)  : t -> t -> bool
  val (>=) : t -> t -> bool
  val (<=) : t -> t -> bool
  val min  : t -> t -> t
  val max  : t -> t -> t
end
```

This is in fact a subset of the `Comparable` interface in Core.

Now, suppose we want to implement multiple modules that support this interface. Implementing all of these individual functions for each module is going to be a lot of painful boilerplate, since it's essentially the same logic that ties it all together.

By using a functor, however, we can derive an implementation that satisfies this interface provided with just the `compare` function. In particular, we'll require as input a module that satisfies this signature:

```
module type Comparable_input = sig
  type t
  val compare : t -> t -> int
end
```

and we'll create a new module that satisfies the `Comparable` signature. Here's a functor which does just that.

```
module Make(M:Comparable_input) : Comparable with type t = M.t = struct
  type t = M.t
  let compare = M.compare
  let (=) x y = compare x y = 0
  let (<>) x y = compare x y <> 0
  let (>) x y = compare x y > 0
  let (<) x y = compare x y < 0
  let (>=) x y = compare x y >= 0
  let (<=) x y = compare x y <= 0
  let min x y = if x < y then x else y
  let max x y = if x > y then x else y
end
```

Syntax Extensions

(yminsky: still very very rough)

This chapter convers several extensions to OCaml's syntax that are distributed with Core. Before diving into the details of the syntax extensions, let's take a small detour that will explain the motivation behind creating them in the first place.

Serialization with s-expressions

Serialization, *i.e.* reading and writing program data to a sequence of bytes, is an important and common programming task. To this end, Core comes with good support for *s-expressions*, which are a convenient general-purpose serialization format. The type of an s-expression is as follows:

```
module Sexp : sig
  type t = Atom of string | List of t list
end
```

An s-expression is in essence a nested parenthetical list whose atomic values are strings. The Sexp module comes with functionality for parsing and printing s-expressions.

```
# let sexp =
    let a x = Sexp.Atom x and l x = Sexp.List x in
    l [a "this";l [a "is"; a "an"]; l [a "s"; a "expression"]];;
val sexp : Sexp.t = (this (is an) (s expression))
```

In addition, most of the base types in Core support conversion to and from s-expressions. For example, we can write:

```
# Int.sexp_of_t 3;;
- : Sexp.t = 3
# List.sexp_of_t;;
- : ('a -> Sexp.t) -> 'a List.t -> Sexp.t = <fun>
# List.sexp_of_t Int.sexp_of_t [1;2;3];;
- : Sexp.t = (1 2 3)
```

Notice that `List.sexp_of_t` is polymorphic, and takes as its first argument another conversion function to handle the elements of the list to be converted. Core uses this scheme more generally for defining sexp-converters for polymorphic types.

But what if you want a function to convert some brand new type to an s-expression? You can of course write it yourself:

```
# type t = { foo: int; bar: float };;
# let sexp_of_t t =
    let a x = Sexp.Atom x and l x = Sexp.List x in
    l [ l [a "foo"; Int.sexp_of_t t.foo  ];
        l [a "bar"; Float.sexp_of_t t.bar]; ]
  ;;
val sexp_of_t : t -> Core.Std.Sexp.t = <fun>
# sexp_of_t { foo = 3; bar = -5.5 };;
- : Core.Std.Sexp.t = ((foo 3) (bar -5.5))
```

This is somewhat tiresome to write, and it gets more so when you consider the parser, *i.e.*, `t_of_sexp`, which is considerably more complex. Writing this kind of parsing and printing code by hand is mechanical and error prone, not to mention a drag.

Given how mechanical the code is, you could imagine writing a program that inspected the type definition and auto-generated the conversion code for you. That is precisely where syntax extensions come in. Using Sexplib and adding `with sexp` as an annotation to our type definition, we get the functions we want for free.

```
# type t = { foo: int; bar: float } with sexp;;
type t = { foo : int; bar : float; }
val t_of_sexp__ : Sexplib.Sexp.t -> t = <fun>
val t_of_sexp : Sexplib.Sexp.t -> t = <fun>
val sexp_of_t : t -> Sexplib.Sexp.t = <fun>
# t_of_sexp (Sexp.of_string "((bar 35) (foo 3))");;
- : t = {foo = 3; bar = 35.}
```

(You can ignore `t_of_sexp__`, which is a helper function that is needed in very rare cases.)

The syntax-extensions in Core that we're going to discuss all have this same basic structure: they auto-generate code based on type definitions, implementing functionality that you could in theory have implemented by hand, but with far less programmer effort.

There are several syntax extensions distributed with Core, including:

- **Sexplib**: provides serialization for s-expressions.
- **Bin_prot**: provides serialization to an efficient binary format.
- **Fieldslib**: generates first-class values that represent fields of a record, as well as accessor functions and setters for mutable record fields.
- **Variantslib**: like Fieldslib for variants, producing first-class variants and other helper functions for interacting with variant types.

- **Pa_compare**: generates efficient, type-specialized comparison functions.
- **Pa_typehash**: generates a hash value for a type definition, *i.e.*, an integer that is highly unlikely to be the same for two distinct types.

We'll discuss each of these syntax extensions in detail, starting with Sexplib.

Sexplib

Formatting of s-expressions

Sexplib's format for s-expressions is pretty straightforward: an s-expression is written down as a nested parenthetical expression, with whitespace-separated strings as the atoms. Quotes are used for atoms that contain parenthesis or spaces themselves, backslash is the escape character, and semicolons are used to introduce comments. Thus, if you create the following file:

```
;; foo.scm

((foo 3.3) ;; Shall I compare thee  to a summer's dream?
 (bar "this is () an \" atom"))
```

we can load it up and print it back out again:

```
# Sexp.load_sexp "foo.scm";;
- : Sexp.t = ((foo 3.3) (bar "this is () an \" atom"))
```

Note that the comments were dropped from the file upon reading. This is expected, since there's no place in the `Sexp.t` type to store comments.

If we introduce an error into our s-expression, by, say, deleting the open-paren in front of `bar`, we'll get a parse error:

```
# Exn.handle_uncaught ~exit:false (fun () ->
    ignore (Sexp.load_sexp "foo.scm"));;
  Uncaught exception:

  (Sexplib.Sexp.Parse_error
   ((location parse) (err_msg "unexpected character: ')'") (text_line 4)
    (text_char 29) (global_offset 94) (buf_pos 94)))
```

(In the above, we use `Exn.handle_uncaught` to make sure that the exception gets printed out in full detail.)

Sexp converters

The most important functionality provided by Sexplib is the auto-generation of converters for new types. We've seen a bit of how this works already, but let's walk through

a complete example. Here's the source for the beginning of a library for representing integer intervals.

```
(* file: int_interval.ml *)
(* Module for representing closed integer intervals *)

open Core.Std

(* Invariant: For any Range (x,y), y > x *)
type t = | Range of int * int
         | Empty
with sexp

let is_empty = function Empty -> true | Range _ -> false
let create x y = if x > y then Empty else Range (x,y)
let contains i x = match i with
   | Empty -> false
   | Range (low,high) -> x >= low && x <= high
```

We can now use this module as follows:

```
(* file: test_interval.ml *)

open Core.Std

let intervals =
  let module I = Int_interval in
  [ I.create 3 4;
    I.create 5 4; (* should be empty *)
    I.create 2 3;
    I.create 1 6;
  ]

let () =
  intervals
  |! List.sexp_of_t Int_interval.sexp_of_t
  |! Sexp.to_string_hum
  |! print_endline
```

But we're still missing something: we haven't created an mli for Int_interval yet. Note that we need to explicitly export the s-expression converters that were created within the ml. If we don't:

```
(* file: int_interval.mli *)
(* Module for representing closed integer intervals *)

type t

val is_empty : t -> bool
val create : int -> int -> t
val contains : t -> int -> bool
```

then we'll get the following error:

```
File "test_interval.ml", line 15, characters 20-42:
Error: Unbound value Int_interval.sexp_of_t
Command exited with code 2.
```

We could export the types by hand:

```
type t
val sexp_of_t : Sexp.t -> t
val t_of_sexp : t -> Sexp.t
```

But Sexplib has a shorthand for this as well, so that we can instead write simply:

```
type t with sexp
```

at which point `test_interval.ml` will compile again, and if we run it, we'll get the following output:

```
$ ./test_interval.native
((Range 3 4) Empty (Range 2 3) (Range 1 6))
```

Preserving invariants

One easy mistake to make when dealing with sexp converters is to ignore the fact that those converters can violate the invariants of your code. For example, the `Int_inter val` module depends for the correctness of the `is_empty` check on the fact that for any value `Range (x,y)`, `y` is greater than or equal to `x`. The `create` function preserves this invariant, but the `t_of_sexp` function does not.

We can fix this problem by writing a custom sexp-converter, in this case, using the sexp-converter that we already have:

```
type t = | Range of int * int
         | Empty
with sexp

let create x y = if x > y then Empty else Range (x,y)

let t_of_sexp sexp =
  let t = t_of_sexp sexp in
  begin match t with
  | Range (x,y) when y < x ->
    of_sexp_error "Upper and lower bound of Range swapped" sexp
  | Empty | Range _ -> ()
  end;
  t
```

We call the function `of_sexp_error` to raise an exception because that improves the error reporting that Sexplib can provide when a conversion fails.

Getting good error messages

There are two steps to deserializing a type from an s-expression: first, converting the bytes in a file to an s-expression, and the second, converting that s-expression into the type in question. One problem with this is that it can be hard to localize errors to the right place using this scheme. Consider the following example:

```
(* file: read_foo.ml *)

open Core.Std

type t = { a: string; b: int; c: float option } with sexp

let run () =
  let t =
    Sexp.load_sexp "foo.scm"
    |! t_of_sexp
  in
  printf "b is: %d\n%!" t.b

let () =
  Exn.handle_uncaught ~exit:true run
```

If you were to run this on a malformatted file, say, this one:

```
;; foo.scm
((a not-an-integer)
 (b not-an-integer)
 (c ()))
```

you'll get the following error:

```
read_foo $ ./read_foo.native
Uncaught exception:

  (Sexplib.Conv.Of_sexp_error
   (Failure "int_of_sexp: (Failure int_of_string)") not-an-integer)
```

If all you have is the error message and the string, it's not terribly informative. In particular, you know that the parsing errored out on the atom "not-an-integer", but you don't know which one! In a large file, this kind of bad error message can be pure misery.

But there's hope! If we make s amll change to the `run` function as follows:

```
let run () =
  let t = Sexp.load_sexp_conv_exn "foo.scm" t_of_sexp in
  printf "b is: %d\n%!" t.b
```

and run it again, we'll get the following much more helpful error message:

```
read_foo $ ./read_foo.native
Uncaught exception:
```

```
(Sexplib.Conv.Of_sexp_error
 (Sexplib.Sexp.Annotated.Conv_exn foo.scm:3:4
  (Failure "int_of_sexp: (Failure int_of_string)"))
 not-an-integer)
```

In the above error, "foo.scm:3:4" tells us that the error occurred on "foo.scm", line 3, character 4, which is a much better start for figuring out what has gone wrong.

Sexp-conversion directives

Sexplib supports a collection of directives for modifying the default behavior of the autogenerated sexp-converters. These directives allow you to customize the way in which types are represented as s-expressions without having to write a custom parser. We describe these directives below.

sexp-opaque

The most commonly used directive is **sexp_opaque**, whose purpose is to mark a given component of a type as being unconvertable. Anything marked with **sexp_opaque** will be presented as the atom **<opaque>** by the to-sexp converter, and will trigger an exception from the from-sexp converter. Note that the type of a component marked as opaque doesn't need to have a sexp-converter defined. Here, if we define a type without a sexp-converter, and then try to use it another type with a sexp-converter, we'll error out:

```
# type no_converter = int * int;;
type no_converter = int * int
# type t = { a: no_converter; b: string } with sexp;;
Characters 14-26:
  type t = { a: no_converter; b: string } with sexp;;
             ^^^^^^^^^^^^
Error: Unbound value no_converter_of_sexp
```

But with **sexp_opaque**, we won't:

```
# type t = { a: no_converter sexp_opaque; b: string } with sexp;;
type t = { a : no_converter Core.Std.sexp_opaque; b : string; }
val t_of_sexp__ : Sexplib.Sexp.t -> t = <fun>
val t_of_sexp : Sexplib.Sexp.t -> t = <fun>
val sexp_of_t : t -> Sexplib.Sexp.t = <fun>
```

And if we now convert a value of this type to an s-expression, we'll see the contents of field **a** marked as opaque:

```
# sexp_of_t { a = (3,4); b = "foo" };;
- : Sexp.t = ((a <opaque>) (b foo))
```

sexp_option

Another common directive is `sexp_opaque`, which is used to make an optional field in a record. Ordinary optional values are represented either as `()` for `None`, or as `(x)` for `Some x`. If you put an option in a record field, then the record field will always be required, and its value will be presented in the way an ordinary optional value would. For example:

```
# type t = { a: int option; b: string } with sexp;;
# sexp_of_t { a = None; b = "hello" };;
- : Sexp.t = ((a ()) (b hello))
# sexp_of_t { a = Some 3; b = "hello" };;
- : Sexp.t = ((a (3)) (b hello))
```

But what if we want a field to be optional, *i.e.*, we want to allow it to be omitted from the record entirely? In that case, we can mark it with `sexp_option`:

```
# type t = { a: int sexp_option; b: string } with sexp;;
# sexp_of_t { a = Some 3; b = "hello" };;
- : Sexp.t = ((a 3) (b hello))
# sexp_of_t { a = None; b = "hello" };;
- : Sexp.t = ((b hello))
```

sexp_list

One problem with the autogenerated sexp-converters is that they can have more parens than one would ideally like. Consider, for example, the following variant type:

```
# type compatible_versions = | Specific of string list
                             | All
  with sexp;;
# sexp_of_compatible_versions (Specific ["3.12.0"; "3.12.1"; "3.13.0"]);;
- : Sexp.t = (Specific (3.12.0 3.12.1 3.13.0))
```

You might prefer to make the syntax a bit less parenthesis-laden by dropping the parens around the list. `sexp_list` gives us this alternate syntax:

```
# type compatible_versions = | Specific of string sexp_list
                             | All
  with sexp;;
# sexp_of_compatible_versions (Specific ["3.12.0"; "3.12.1"; "3.13.0"]);;
- : Sexp.t = (Specific 3.12.0 3.12.1 3.13.0)
```

Bin_prot

S-expressions are a good serialization format when you need something machine-parseable as well as human readable and editable. But Sexplib's s-expressions are not particularly performant. There are a number of reasons for this. For one thing, s-expression serialization goes through an intermediate type, `Sexp.t`, which must be allo-

cated and is then typically thrown away, putting non-trivial pressure on the GC. In addition, parsing and printing to strings in an ASCII format can be expensive for types like `ints`, `floats` and `Time.t`s where some real computation needs to be done to produce or parse the ASCII representation.

Bin_prot is a library designed to address these issues by providing fast serialization in a compact binary format. Kicking off the syntax extension is done by putting `with bin_io`. (This looks a bit unsightly in the top-level because of all the definitions that are generated. We'll elide those definitions here, but you can see it for yourself in the top-level.)

Here's a small complete example of a program that can read and write values using bin-io. Here, the serialization is of types that might be used as part of a message-queue, where each message has a topic, some content, and a source, which is in turn a hostname and a port.

```
(* file: message_example.ml *)

open Core.Std

(* The type of a message *)
module Message = struct
  module Source = struct
    type t = { hostname: string;
               port: int;
             }
    with bin_io
  end

  type t = { topic: string;
             content: string;
             source: Source.t;
           }
  with bin_io
end

(* Create the 1st-class module providing the binability of messages *)
let binable = (module Message : Binable.S with type t = Message.t)

(* Saves a message to an output channel.  The message is serialized to
   a bigstring before being written out to the channel.  Also, a
   binary encoding of an integer is written out to tell the reader how
   long of a message to expect.  *)
let save_message outc msg =
  let s = Binable.to_bigstring binable msg in
  let len = Bigstring.length s in
  Out_channel.output_binary_int outc len;
  Bigstring.really_output outc s

(* Loading the message is done by first reading in the length, and by
   then reading in the appropriate number of bytes into a Bigstring
   created for that purpose. *)
```

```
let load_message inc =
  match In_channel.input_binary_int inc with
  | None -> failwith "Couldn't load message: length missing from header"
  | Some len ->
    let buf = Bigstring.create len in
    Bigstring.really_input ~pos:0 ~len inc buf;
    Binable.of_bigstring binable buf

(* To generate some example messages *)
let example content =
  let source =
    { Message.Source.
      hostname = "ocaml.org"; port = 2322 }
  in
  { Message.
    topic = "rwo-example"; content; source; }

(* write out three messages... *)
let write_messages () =
  let outc = Out_channel.create "tmp.bin" in
  List.iter ~f:(save_message outc) [
    example "a wonderful";
    example "trio";
    example "of messages";
  ];
  Out_channel.close outc

(* ... and read them back in *)
let read_messages () =
  let inc = In_channel.create "tmp.bin" in
  for i = 1 to 3 do
    let msg = load_message inc in
    printf "msg %d: %s\n" i msg.Message.content
  done

let () =
  write_messages (); read_messages ()
```

Fieldslib

One common idiom when using records is to provide field accessor functions for a particular record.

```
type t = { topic: string;
           content: string;
           source: Source.t;
         }

let topic   t = t.topic
let content t = t.content
let source  t = t.source
```

Similarly, sometimes you simultaneously want an accessor to a field of a record and a textual representation of the name of that field. This might come up if you were validating a field and needed the string representation to generate an error message, or if you wanted to scaffold a form in a GUI automatically based on the fields of a record. Fieldslib provides a module `Field` for this purpose. Here's some code for creating `Field.t`'s for all the fields of our type `t`.

```
# module Fields = struct
    let topic =
      { Field.
        name   = "topic";
        setter = None;
        getter = (fun t -> t.topic);
        fset   = (fun t topic -> { t with topic });
      }
    let content =
      { Field.
        name   = "content";
        setter = None;
        getter = (fun t -> t.content);
        fset   = (fun t content -> { t with content });
      }
    let source =
      { Field.
        name   = "source";
        setter = None;
        getter = (fun t -> t.source);
        fset   = (fun t source -> { t with source });
      }
  end ;;
module Fields :
  sig
    val topic : (t, string list) Core.Std.Field.t
    val content : (t, string) Core.Std.Field.t
    val source : (t, Source.t) Core.Std.Field.t
  end
```

Object Oriented Programming

(yminsky: If we don't feel like these are "great" tools, maybe we shouldn't say it!)

(yminsky: I wonder if it's worth emphasizing what makes objects unique early on. I think of them as no better of an encapsulation tool than closures. What makes them unique in my mind is that they are some combination of lighter weight and more dynamic than the alternatives (modules, records of closures, etc.))

(yminsky: I'm not sure where we should say it, but OCaml's object system is strikingly different from those that most people are used to. It would be nice if we could call those differences out clearly somewhere. The main difference I see is the fact that subtyping and inheritance are not tied together, and that subtyping is structural.)

We've already seen several tools that OCaml provides for organizing programs, particularly first-class modules. In addition, OCaml also supports object-oriented programming. There are objects, classes, and their associated types. Objects are good for encapsulation and abstraction, and classes are good for code re-use.

When to use objects

You might wonder when to use objects. First-class modules are more expressive (a module can include types, classes and objects cannot), and modules, functors, and algebraic data types offer a wide range of ways to express program structure. In fact, many seasoned OCaml programmers rarely use classes and objects, if at all.

What exactly is object-oriented programming? Mitchell [6] points out four fundamental properties.

- *Abstraction*: the details of the implementation are hidden in the object; the interface is just the set of publically-accessible methods.
- *Subtyping*: if an object **a** has all the functionality of an object **b**, then we may use **a** in any context where **b** is expected.

- *Dynamic lookup*: when a message is sent to an object, the method to be executed is determined by the implementation of the object, not by some static property of the program. In other words, different objects may react to the same message in different ways.
- *Inheritance*: the definition of one kind of object can be re-used to produce a new kind of object.

Modules already provide these features in some form, but the main focus of classes is on code re-use through inheritance and late binding of methods. This is a critical property of classes: the methods that implement an object are determined when the object is instantiated, a form of *dynamic* binding. In the meantime, while classes are being defined, it is possible (and necessary) to refer to methods without knowing statically how they will be implemented.

In contrast, modules use static (lexical) scoping. If you want to parameterize your module code so that some part of it can be implemented later, you would write a function/functor. This is more explicit, but often more verbose than overriding a method in a class.

In general, a rule of thumb might be: use classes and objects in situations where dynamic binding is a big win, for example if you have many similar variations in the implementation of a concept. Real world examples are fairly rare, but one good example is Xavier Leroy's Cryptokit (*http://gallium.inria.fr/~xleroy/software.html#cryptokit*), which provides a variety of cryptographic primitives that can be combined in building-block style.

OCaml objects

If you already know about object oriented programming in a language like Java or C++, the OCaml object system may come as a surprise. Foremost is the complete separation of subtyping and inheritance in OCaml. In a language like Java, a class name is also used as the type of objects created by instantiating it, and the subtyping rule corresponds to inheritance. For example. if we implement a class `Stack` in Java by inheriting from a class `Deque`, we would be allowed to pass a stack anywhere a deque is expected (this is a silly example of course, practitioners will point out that we shouldn't do it).

OCaml is entirely different. Classes are used to construct objects and support inheritance, including non-subtyping inheritance. Classes are not types. Instead, objects have *object types*, and if you want to use objects, you aren't required to use classes at all. Here is an example of a simple object.

```
# let p =
  object
    val mutable x = 0
    method get = x
    method set i = x <- i
```

```
      end;;
    val p : < get : int; set : int -> unit > = <obj>
```

The object has an integer value x, a method **get** that returns x, and a method **set** that updates the value of x.

The object type is enclosed in angle brackets < ... >, containing just the types of the methods. Fields, like x, are not part of the public interface of an object. All interaction with an object is through its methods. The syntax for a method invocation (also called "sending a message" to the object) uses the # character.

```
# p#get;
- : int = 0
# p#set 17;;
- : unit = ()
# p#get;;
- : int = 17
```

Objects can also be constructed by functions. If we want to specify the initial value of the object, we can define a function that takes the initial value and produces an object.

```
# let make i =
    object
      val mutable x = i
      method get = x
      method set y = x <- y
    end;;
val make : 'a -> < get : 'a; set : 'a -> unit > = <fun>
# let p = make 5;;
val p : < get : int; set : int -> unit > = <obj>
# p#get;;
- : int = 5
```

Note that the types of the function **make** and the returned object now use the polymorphic type **'a**. When make is invoked on a concrete value **5**, we get the same object type as before, with type **int** for the value.

Object Polymorphism

(yminsky: Maybe this is a good time to talk about the nature of object subtyping?)

Functions can also take object arguments. Let's construct a new object **average** that's the average of any two objects with a **get** method.

```
# let average p1 p2 =
    object
      method get = (p1#get + p2#get) / 2
    end;;
val average : < get : int; .. > -> < get : int; .. > -> < get : int > = <fun>
# let p1 = make 5;;
# let p2 = make 15;;
```

```
# let a = average p1 p2;;
# a#get;;
- : int = 10
# p2#set 25;;
# a#get;;
- : int = 15
```

Note that the type for average uses the object type < get : int; .. >. The .. are ellipsis, standing for any other methods. The type < get : int; .. > specifies an object that must have at least a get method, and possibly some others as well. If we try using the exact type < get : int > for an object with more methods, type inference will fail.

```
# let (p : < get : int >) = make 5;;
Error: This expression has type < get : int; set : int -> unit >
        but an expression was expected of type < get : int >
        The second object type has no method set
```

Elisions are polymorphic

The .. in an object type is an elision, standing for "possibly more methods." It may not be apparent from the syntax, but an elided object type is actually polymorphic. If we try to write a type definition, we get an obscure error.

```
# type point = < get:int; .. >;;
Error: A type variable is unbound in this type declaration.
In type < get : int; .. > as 'a the variable 'a is unbound
```

A .. in an object type is called a *row variable* and this typing scheme is called *row polymorphism*. Even though .. doesn't look like a type variable, it actually is. The error message suggests a solution, which is to add the as 'a type constraint.

```
# type 'a point = < get:int; .. > as 'a;;
type 'a point = 'a constraint 'a = < get : int; .. >
```

In other words, the type 'a point is equal to 'a, where 'a = < get : int; .. >. That may seem like an odd way to say it, and in fact, this type definition is not really an abbreviation because 'a refers to the entire type.

An object of type < get:int; .. > can be any object with a method get:int, it doesn't matter how it is implemented. So far, we've constructed two objects with that type; the function make constructed one, and so did average. When the method #get is invoked, the actual method that is run is determined by the object.

```
# let print_point p = Printf.printf "Point: %d\n" p#get;;
val print_point : < get : int; .. > -> unit = <fun>
# print_point (make 5);;
Point: 5
# print_point (average (make 5) (make 15));;
Point: 10
```

Classes

Programming with objects directly is great for encapsulation, but one of the main goals of object-oriented programming is code re-use through inheritance. For inheritance, we need to introduce *classes*. In object-oriented programming, a class is a "recipe" for creating objects. The recipe can be changed by adding new methods and fields, or it can be changed by modifying existing methods.

In OCaml, class definitions must be defined as top-level statements in a module. A class is not an object, and a class definition is not an expression. The syntax for a class definition uses the keyword `class`.

```
# class point =
  object
    val mutable x = 0
    method get = x
    method set y = x <- y
  end;;
class point :
  object
    val mutable x : int
    method get : int
    method set : int -> unit
  end
```

The type `class point : ... end` is a *class type*. This particular type specifies that the `point` class defines a mutable field `x`, a method `get` that returns an `int`, and a method `set` with type `int -> unit`.

To produce an object, classes are instantiated with the keyword `new`.

```
# let p = new point;;
val p : point = <obj>
# p#get;;
- : int = 0
# p#set 5;;
- : unit = ()
# p#get;;
- : int = 5
```

(yminsky: You say that inheritance uses an existing class to define a new one, but the example below looks like using an existing class to define a new module. Is that what's going on? Or is a new class being created implicitly? If the latter, it might be better to be more explicit in this example and name the new class.)

Inheritance uses an existing class to define a new one. For example, the following class definition supports an addition method `moveby` that moves the point by a relative amount. This also makes use of the `(self : 'self)` binding after the `object` keyword. The variable `self` stands for the current object, allowing self-invocation, and the type

variable `'self` stands for the type of the current object (which in general is a subtype of `movable_point`).

```
# class movable_point =
  object (self : 'self)
    inherit point
    method moveby dx = self#set (self#get + dx)
  end;;
```

Class parameters and polymorphism

A class definition serves as the *constructor* for the class. In general, a class definition may have parameters that must be provided as arguments when the object is created with `new`.

Let's build an example of an imperative singly-linked list using object-oriented techniques. First, we'll want to define a class for a single element of the list. We'll call it a `node`, and it will hold a value of type `'a`. When defining the class, the type parameters are placed in square brackets before the class name in the class definition. We also need a parameter `x` for the initial value.

```
class ['a] node x =
object
  val mutable value : 'a = x
  val mutable next_node : 'a node option = None

  method get = value
  method set x = value <- x

  method next = next_node
  method set_next node = next_node <- node
end;;
```

The `value` is the value stored in the node, and it can be retrieved and changed with the `get` and `set` methods. The `next_node` field is the link to the next element in the stack. Note that the type parameter `['a]` in the definition uses square brackets, but other uses of the type can omit them (or use parentheses if there is more than one type parameter).

The type annotations on the `val` declarations are used to constrain type inference. If we omit these annotations, the type inferred for the class will be "too polymorphic," `x` could have some type `'b` and `next_node` some type `'c option`.

```
class ['a] node x =
object
  val mutable value = x
  val mutable next_node = None

  method get = value
  method set x = value <- x
```

```
      method next = next_node
      method set_next node = next_node <- node
   end;;
Error: Some type variables are unbound in this type:
        class ['a] node :
          'b ->
          object
            val mutable next_node : 'c option
            val mutable value : 'b
            method get : 'b
            method next : 'c option
            method set : 'b -> unit
            method set_next : 'c option -> unit
          end
      The method get has type 'b where 'b is unbound
```

In general, we need to provide enough constraints so that the compiler will infer the correct type. We can add type constraints to the parameters, to the fields, and to the methods. It is a matter of preference how many constraints to add. You can add type constraints in all three places, but the extra text may not help clarity. A convenient middle ground is to annotate the fields and/or class parameters, and add constraints to methods only if necessary.

Next, we can define the list itself. We'll keep a field `head` the refers to the first element in the list, and `last` refers to the final element in the list. The method `insert` adds an element to the end of the list.

```
class ['a] slist =
object
   val mutable first : ('a) node option = None
   val mutable last : ('a) node option = None

   method is_empty = first = None

   method insert x =
      let new_node = Some (new node x) in
      match last with
         Some last_node ->
            last_node#set_next new_node;
            last <- new_node
       | None ->
            first <- new_node;
            last <- new_node
end;;
```

Object types

This definition of the class `slist` is not complete, we can construct lists, but we also need to add the ability to traverse the elements in the list. One common style for doing this is to define a class for an `iterator` object. An iterator provides a generic mechanism

to inspect and traverse the elements of a collection. This pattern isn't restricted to lists, it can be used for many different kinds of collections.

There are two common styles for defining abstract interfaces like this. In Java, an iterator would normally be specified with an interface, which specifies a set of method types. In languages without interfaces, like C++, the specification would normally use *abstract* classes to specify the methods without implementing them (C++ uses the "= 0" definition to mean "not implemented").

```
// Java-style iterator, specified as an interface.
interface <T> iterator {
  T Get();
  boolean HasValue();
  void Next();
};

// Abstract class definition in C++.
template<typename T>
class Iterator {
 public:
  virtual ~Iterator() {}
  virtual T get() const = 0;
  virtual bool has_value() const = 0;
  virtual void next() = 0;
};
```

OCaml support both styles. In fact, OCaml is more flexible than these approaches because an object type can be implemented by any object with the appropriate methods, it does not have to be specified by the object's class *a priori*. We'll leave abstract classes for later. Let's demonstrate the technique using object types.

First, we'll define an object type `iterator` that specifies the methods in an iterator.

```
type 'a iterator = < get : 'a; has_value : bool; next : unit >;;`
```

Next, we'll define an actual iterator for the class `slist`. We can represent the position in the list with a field `current`, following links as we traverse the list.

```
class ['a] slist_iterator cur =
object
  val mutable current : 'a node option = cur

  method has_value = current <> None

  method get =
    match current with
        Some node -> node#get
      | None -> raise (Invalid_argument "no value")

  method next =
    match current with
        Some node -> current <- node#next
```

```
          | None -> raise (Invalid_argument "no value")
    end;;
```

Finally, we add a method `iterator` to the slist class to produce an iterator. To do so, we construct an `slist_iterator` that refers to the first node in the list, but we want to return a value with the object type `iterator`. This requires an explicit coercion using the `:>` operator.

```
class ['a] slist = object
...
    method iterator = (new slist_iterator first :> 'a iterator)
end

# let l = new slist;;
# l.insert 5;;
# l.insert 4;;
# let it = l#iterator;;
# it#get;;
- : int = 5
# it#next;;
- : unit = ()
# it#get;;
- : int = 4
# it#next;;
- : unit = ()
# it#has_value;;
- : bool = false
```

We may also wish to define functional-style methods, `iter f` takes a function `f` and applies it to each of the elements of the list.

```
method iter f =
   let it = self#iterator in
   while it#has_value do
     f it#get
     it#next
   end
```

What about functional operations similar to `List.map` or `List.fold`? In general, these methods take a function that produces a value of some other type than the elements of the set. For example, the function `List.fold` has type `'a list -> ('b -> 'a -> 'b) -> 'b -> 'b`, where `'b` is an arbitrary type. To replicate this in the `slist` class, we need a method type `('b -> 'a -> 'b) -> 'b -> 'b`, where the method type is polymorphic over `'b`.

The solution is to use a type quantifier, as shown in the following example. The method type must be specified directly after the method name, which means that method parameters must be expressed using a `fun` or `function` expression.

```
method fold : 'b. ('b -> 'a -> 'b) -> 'b -> 'b =
   (fun f x ->
```

```
            let y = ref x in
            let it = self#iterator in
            while it#has_value do
                y := f !y it#get;
                it#next
            done;
            !y)
```

Immutable objects

Many people consider object-oriented programming to be intrinsically imperative, where an object is like a state machine. Sending a message to an object causes it to change state, possibily sending messages to other objects.

Indeed, in many programs, this makes sense, but it is by no means required. Let's define an object-oriented version of lists similar to the imperative list above. We'll implement it with a regular list type `'a list`, and insertion will be to the beginning of the list instead of to the end.

```
class ['a] flist =
object (self : 'self)
    val elements : 'a list = []

    method is_empty = elements = []

    method insert x : 'self = {< elements = x :: elements >}

    method iterator =
        (new flist_iterator elements :> 'a iterator)

    method iter (f : 'a -> unit) = List.iter f elements

    method fold : 'b. ('b -> 'a -> 'b) -> 'b -> 'b =
        (fun f x -> List.fold_left f x elements)
end;;
```

A key part of the implementation is the definition of the method `insert`. The expression `{< ... >}` produces a copy of the current object, with the same type, and the specified fields updated. In other words, the `new_fst new_x` method produces a copy of the object, with x replaced by `new_x`. The original object is not modified, and the value of y is also unaffected.

There are some restriction on the use of the expression `{< ... >}`. It can be used only within a method body, and only the values of fields may be updated. Method implementations are fixed at the time the object is created, they cannot be changed dynamically.

We use the same object type `iterator` for iterators, but implement it differently.

```
class ['a] flist_iterator l =
```

```
object
   val mutable elements : 'a list = l

   method has_value = l <> []

   method get =
      match l with
         h :: _ -> h
       | [] -> raise (Invalid_argument "list is empty")

   method next =
      match l with
         _ :: l -> elements <- l
       | [] -> raise (Invalid_argument "list is empty")
end;;
```

Class types

Once we have defined the list implementation, the next step is to wrap it in a module or `.ml` file and give it a type so that it can be used in the rest of our code. What is the type?

Before we begin, let's wrap up the implementation in an explicit module (we'll use explicit modules for illustration, but the process is similar when we want to define a `.mli` file). In keeping with the usual style for modules, we define a type `'a t` to represent the type of list values.

```
module SList = struct
   type 'a iterator = < get : 'a; has_value : bool; next : unit >
   type 'a t = < is_empty : bool; insert : 'a -> unit; iterator : 'a iterator >

   class ['a] node x = object ... end
   class ['a] slist_iterator cur = object ... end
   class ['a] slist = object ... end

   let make () = new slist
end;;
```

We have multiple choices in defining the module type, depending on how much of the implementation we want to expose. At one extreme, a maximally-abstract signature would completely hide the class definitions.

```
module AbstractSList : sig
   type 'a iterator = < get : 'a; has_value : bool; next : unit >
   type 'a t = < is_empty : bool; insert : 'a -> unit; iterator : 'a iterator >

   val make : unit -> 'a t
end = SList
```

The abstract signature is simple because we ignore the classes. But what if we want to include them in the signature, so that other modules can inherit from the class definitions? For this, we need to specify types for the classes, called *class types*. Class types

do not appear in mainstream object-oriented programming languages, so you may not be familiar with them, but the concept is pretty simple. A class type specifies the type of each of the visible parts of the class, including both fields and methods. Just like for module types, you don't have to give a type for everything; anything you omit will be hidden.

```
module VisibleSList : sig
  type 'a iterator = < get : 'a; has_value : bool; next : unit >
  type 'a t = < is_empty : bool; insert : 'a -> unit; iterator : 'a iterator >

  class ['a] node : 'a ->
  object
     method get : 'a
     method set : 'a -> unit
     method next : 'a node option
     method set_next : 'a node option -> unit
  end

  class ['a] slist_iterator : 'a node option ->
  object
     method has_value : bool
     method get : 'a
     method next : unit
  end

  class ['a] slist :
  object
    val mutable first : 'a node option
    val mutable last : 'a node option
    method is_empty : bool
    method insert : 'a -> unit
    method iterator : 'a iterator
  end

  val make : unit -> 'a slist
end = SList
```

In this signature, we've chosen to make nearly everything visible. The class type for `slist` specifies the types of the fields `first` and `last`, as well ad the types of each of the methods. We've also included a class type for `slist_iterator`, which is of somewhat more questionable value, since the type doesn't appear in the type for `slist` at all.

One more thing, in this example the function `make` has type `unit -> 'a slist`. But wait, we've stressed *classes are not types*, so what's up with that? In fact, what we've said is entirely true, classes and class names *are not* types. However, class names can be used to stand for types. When the compiler sees a class name in type position, it automatically constructs an object type from it by erasing all the fields and keeping only the method types. In this case, the type expression `'a slist` is exactly equivalent to `'a t`.

Subtyping

Subtyping is a central concept in object-oriented programming. It governs when an object with one type *A* can be used in an expression that expects an object of another type *B*. When this is true, we say that *A* is a *subtype* of *B*. Actually, more concretely, subtyping determines when the coercion operator `e :> t` can be applied. This coercion works only if the expression `e` has some type `s` and `s` is a subtype of `t`.

To explore this, let's define some simple classes for geometric shapes. The generic type `shape` has a method to compute the area, and a `square` is a specific kind of shape.

```
type shape = < area : float >;;

class square w =
object (self : 'self)
  method area = self#width *. self#width
  method width = w
end;;
```

A `square` has a method `area` just like a `shape`, and an additional method `width`. Still, we expect a `square` to be a `shape`, and it is. The coercion `:>` must be explicit.

```
# let new_square x : shape = new square x;;
Characters 27-39:
  let new_square x : shape = new square x;;
                            ^^^^^^^^^^^^
Error: This expression has type square but an expression was expected of type shape
The second object type has no method width
# let new_square x : shape = (new square x :> shape);;
val new_square : float -> shape = <fun>
```

What are the rules for subtyping? In general, object subtyping has two general forms, called *width* and *depth* subtyping. Width subtyping means that an object type *A* is a a a subtype of *B*, if *A* has all of the methods of *B*, and possibly more. A `square` is a subtype of `shape` because it implements all of the methods of `shape` (the `area` method).

The subtyping rules are purely technical, they have no relation to object semantics. We can define a class `rectangle` that has all of the methods of a `square`, so it is a subtype of square and can be used wherever a `square` is expected.

```
# class rectangle h w =
    object (self : 'self)
      inherit square w
      method area = self#width *. self#height
      method height = h
    end;;
# let square_rectangle h w : square = (new rectangle h w :> square);;
val square_rectangle : float -> float -> square = <fun>
```

This may seem absurd, but this concept is expressible in all object-oriented languages. The contradiction is semantic -- we know that in the real world, not all rectangles are squares; but in the programming world, rectangles have all of the features of squares (according to our definition), so they can be used just like squares. Suffice it to say that it is usually better to avoid such apparent contradictions.

Next, let's take a seemingly tiny step forward, and start building collections of shapes. It is easy enough to define a `slist` of squares.

```
# let squares =
     let l = SList.make () in
     l#insert (new square 1.0);
     l#insert (new square 2.0);
     l;;
val squares : square slist = <obj>
```

We can also define a function to calculate the total area of a list of shapes. There is no reason to restrict this to squares, it should work for any list of shapes with type `shape slist`. The problem is that doing so raises some serious typing questions -- can a `square slist` be passed to a function that expects a `shape slist`? If we try it, the compiler produces a verbose error message.

```
# let total_area (l : shape slist) : float =
     let total = ref 0.0 in
     let it = l#iterator in
     while it#has_value do
        total := !total +. it#get#area;
        it#next
     done;
     !total;;
val total_area : shape slist -> float = <fun>
# total_area squares;;
Characters 11-18:
  total_area squares;;
           ^^^^^^^
Error: This expression has type
         square slist =
           < insert : square -> unit; is_empty : bool;
             iterator : square iterator >
       but an expression was expected of type
         shape slist =
           < insert : shape -> unit; is_empty : bool;
             iterator : shape iterator >
       Type square = < area : float; width : float >
       is not compatible with type shape = < area : float >
       The second object type has no method width
```

It might seem tempting to give up at this point, especially because the subtyping is not even true -- the type `square slist` is not a subtype of `shape slist`. The problem is with the `insert` method. For `shape slist`, the `insert` method takes an arbitrary `shape` and

inserts it into the list. So if we could coerce a `square slist` to a `shape slist`, then it would be possible to insert an arbitrary shape into the list, which would be an error.

Using more precise types to address subtyping problems

Still, the `total_area` function should be fine, in principle. It doesn't call `insert`, so it isn't making that error. To make it work, we need to use a more precise type that indicates we are not going to be mutating the list. We define a type `readonly_shape_slist` and confirm that we can coerce the list of squares.

```
# type readonly_shape_slist = < iterator : shape iterator >;;
type readonly_shape_slist = < iterator : shape iterator >
# (squares :> readonly_shape_slist);;
- : readonly_shape_slist = <obj>
# let total_area (l : readonly_shape_slist) : float = ...;;
val total_area : readonly_shape_slist -> float = <fun>
#   total_area (squares :> readonly_shape_slist);;
- : float = 5.
```

Why does this work, why is a `square slist` a subtype of `readonly_shape_slist`. The reasoning is in two steps. First, the easy part is width subtyping: we can drop the other methods to see that `square slist` is a subtype of `< iterator : square iterator >`. The next step is to use *depth* subtyping, which, in its general form, says that an object type `< m : t1 >` is a subtype of a type `< m : t2>` iff `t1` is a subtype of `t2`. In other words, instead of reasoning about the number of methods in a type (the width), the number of methods is fixed, and we look within the method types themselves (the "depth").

In this particular case, depth subtyping on the `iterator` method requires that `square iterator` be a subtype of `shape iterator`. Expanding the type definition for the type `iterator`, we again invoke depth subtyping, and we need to show that the type `< get : square >` is a subtype of `<get : shape >`, which follows because `square` is a subtype of `shape`.

This reasoning may seem fairly long and complicated, but it should be pointed out that this typing *works*, and in the end the type annotations are fairly minor. In most typed object-oriented languages, the coercion would simply not be possible. For example, in C++, a STL type `slist<T>` is invariant in `T`, it is simply not possible to use `slist<square>` where `slist<shape>` is expected (at least safely). The situation is similar in Java, although Java supports has an escape hatch that allows the program to fall back to dynamic typing. The situation in OCaml is much better; it works, it is statically checked, and the annotations are pretty simple.

Using elided types to address subtyping problems

Before we move to the next topic, there is one more thing to address. The typing we gave above, using `readonly_shape_slist`, requires that the caller perform an explicit

coercion before calling the `total_area` function. We would like to give a better type that avoids the coercion.

A solution is to use an elided type. Instead of `shape`, we can use the elided type `< area : float; .. >`. In fact, once we do this, it also becomes possible to use the `slist` type.

```
# let total_area (l : < area : float; .. > slist) : float = ...;;
val total_area : < area : float; .. > slist -> float = <fun>
# total_area squares;;
- : float = 5.
```

This works, and it removes the need for explicit coercions. This type is still fairly simple, but it does have the drawback that the programmer needs to remember that the types `< area : float; ..>` and `shape` are related.

OCaml supports an abbreviation in this case, but it works only for classes, not object types. The type expression `# classname` is an abbreviation for an elided type containing all of the methods in the named class, and more. Since `shape` is an object type, we can't write `#shape`. However, if a class definition is available, this abbreviation can be useful. The following definition is exactly equivalent to the preceding.

```
# class cshape = object method area = 0.0 end;;
class cshape : object method area : float end
# let total_area (l : #cshape list) : float = ...;;
val total_area : #cshape slist -> float = <fun>
# total_area squares;;
- : float = 5.
```

Narrowing

Narrowing, also called *down casting*, is the ability to coerce an object to one of its subtypes. For example, if we have a list of shapes `shape slist`, we might know (for some reason) what the actual type of each shape is. Perhaps we know that all objects in the list have type `square`. In this case, *narrowing* would allow the re-casting of the object from type `shape` to type `square`. Many languages support narrowing through dynamic type checking. For example, in Java, a coercion `(Square) x` is allowed if the value `x` has type `Square` or one of its subtypes; otherwise the coercion throws an exception.

Narrowing is *not permitted* in OCaml. Period.

Why? There are two reasonable explanations, one based on a design principle, and another technical (the technical reason is simple: it is hard to implement).

The design argument is this: narrowing violates abstraction. In fact, with a structural typing system like in OCaml, narrowing would essentially provide the ability to enumerate the methods in an object. To check whether an object `obj` has some method `foo : int`, one would attempt a coercion `(obj :> < foo : int >)`.

More commonly, narrowing leads to poor object-oriented style. Consider the following Java code, which returns the name of a shape object.

```java
String GetShapeName(Shape s) {
  if (s instanceof Square) {
    return "Square";
  } else if (s instanceof Circle) {
    return "Circle";
  } else {
    return "Other";
  }
}
```

Most programmers would consider this code to be "wrong." Instead of performing a case analysis on the type of object, it would be better to define a method to return the name of the shape. Instead of calling `GetShapeName(s)`, we should call `s.Name()` instead.

However, the situation is not always so obvious. The following code checks whether an array of shapes looks like a "barbell," composed to two `Circle` objects separated by a `Line`, where the circles have the same radius.

```java
boolean IsBarBell(Shape[] s) {
  return s.length == 3 && (s[0] instanceof Circle) &&
    (s[1] instanceof Line) && (s[2] instanceof Circle) &&
    ((Circle) s[0]).radius() == ((Circle) s[2]).radius();
}
```

In this case, it is much less clear how to augment the `Shape` class to support this kind of pattern analysis. It is also not obvious that object-oriented programming is well-suited for this situation. Pattern matching seems like a better fit.

```ocaml
let is_bar_bell = function
  | [Circle r1; Line _; Circle r2] when r1 == r2 -> true
  | _ -> false;;
```

Regardless, there is a solution if you find yourself in this situation, which is to augment the classes with variants. You can define a method `variant` that injects the actual object into a variant type.

```ocaml
type shape = < variant : repr; area : float>
and circle = < variant : repr; area : float; radius : float >
and line = < variant : repr; area : float; length : float >
and repr =
  | Circle of circle
  | Line of line;;

let is_bar_bell = function
  | [s1; s2; s3] ->
    (match s1#variant, s2#variant, s3#variant with
      | Circle c1, Line _, Circle c2 when c1#radius == c2#radius -> true
```

```
      | _ -> false)
    | _ -> false;;
```

This pattern works, but it has drawbacks. In particular, the recursive type definition should make it clear that this pattern is essentially equivalent to using variants, and that objects do not provide much value here.

Binary methods

A *binary method* is a method that takes an object of self type. One common example is defining a method for equality.

```
# class square w =
    object (self : 'self)
      method width = w
      method area = self#width * self#width
      method equals (other : 'self) = other#width = self#width
    end;;
class square : int ->
  object ('a)
    method area : int
    method equals : 'a -> bool
    method width : int
  end
# class rectangle w h =
    object (self : 'self)
      method width = w
      method height = h
      method area = self#width * self#height
      method equals (other : 'self) = other#width = self#width && other#height = self#height
    end;;
...
# (new square 5)#equals (new square 5);;
- : bool = true
# (new rectangle 5 6)#equals (new rectangle 5 7);;
- : bool = false
```

This works, but there is a problem lurking here. The method `equals` takes an object of the exact type `square` or `rectangle`. Because of this, we can't define a common base class `shape` that also includes an equality method.

```
# type shape = < equals : shape -> bool; area : int >;;
# let sq = new square 5;;
# (sq :> shape);;
Characters 0-13:
  (sq :> shape);;
  ^^^^^^^^^^^^^
Error: Type square = < area : int; equals : square -> bool; width : int >
       is not a subtype of shape = < area : int; equals : shape -> bool >
Type shape = < area : int; equals : shape -> bool > is not a subtype of
  square = < area : int; equals : square -> bool; width : int >
```

The problem is that a `square` expects to be compared with a `square`, not an arbitrary shape; similarly for `rectangle`.

This problem is fundamental. Many languages solve it either with narrowing (with dynamic type checking), or by method overloading. Since OCaml has neither of these, what can we do?

One proposal we could consider is, since the problematic method is equality, why not just drop it from the base type `shape` and use polymorphic equality instead? Unfortunately, the builtin equality has very poor behavior when applied to objects.

```
# (object method area = 5 end) = (object method area = 5 end);;
- : bool = false
```

The problem here is that the builtin polymorphic equality compares the method implementations, not their return values. The method implementations (the function values that implement the methods) are different, so the equality comparison is false. There are other reasons not to use the builtin polymorphic equality, but these false negatives are a showstopper.

If we want to define equality for shapes in general, the remaining solution is to use the same approach as we described for narrowing. That is, introduce a *representation* type implemented using variants, and implement the comparison based on the representation type.

```
type shape_repr =
  | Square of int
  | Circle of int
  | Rectangle of int * int;;

type shape = < repr : shape_repr; equals : shape -> bool; area : int >;;

class square w =
object (self : 'self)
  method width = w
  method area = self#width * self#width
  method repr = Square self#width
  method equals (other : shape) = self#repr = other#repr
end;;
```

The binary method `equals` is now implemented in terms of the concrete type `shape_repr`. In fact, the objects are now isomorphic to the `shape_repr` type. When using this pattern, you will not be able to hide the `repr` method, but you can hide the type definition using the module system.

```
module Shapes : sig
  type shape_repr
  type shape = < repr : shape_repr; equals : shape -> bool; area -> int >

  class square : int ->
```

```
object
   method width : int
   method area : int
   method repr : shape_repr
   method equals : shape -> bool
 end
end = struct
 type shape_repr = Square of int | Circle of int | Rectangle of int * int
 ...
end;;
```

Private methods

Methods can be declared *private*, which means that they may be called by subclasses, but they are not visible otherwise (similar to a *protected* method in C++).

To illustrate, let's build a class vector that contains an array of integers, resizing the storage array on demand. The field values contains the actual values, and the get, set, and length methods implement the array access. For clarity, the resizing operation is implemented as a private method ensure_capacity that resizes the array if necessary.

```
# class vector =
  object (self : 'self)
     val mutable values : int array = [|||]

     method get i = values.(i)
     method set i x =
        self#ensure_capacity i;
        values.(i) <- x
     method length = Array.length values

     method private ensure_capacity i =
        if self#length <= i then
           let new_values = Array.create (i + 1) 0 in
           Array.blit values 0 new_values 0 (Array.length values);
           values <- new_values
  end;;
# let v = new vector;;
# v#set 5 2;;
# v#get 5;;
- 2 : int
# v#ensure_capacity 10;;
Characters 0-1:
  v#ensure_capacity 10;;
  ^
Error: This expression has type vector
       It has no method ensure_capacity
```

To be precise, the method ensure_capacity is part of the class type, but it is not part of the object type. This means the object v has no method ensure_capacity. However, it is available to subclasses. We can extend the class, for example, to include a method swap that swaps two elements.

```
# class swappable_vector =
  object (self : 'self)
    inherit vector

    method swap i j =
        self#ensure_capacity (max i j);
        let tmp = values.(i) in
        values.(i) <- values.(j);
        values.(j) <- tmp
  end;;
```

Yet another reason for private methods is to factor the implementation and support recursion. Moving along with this example, let's build a binary heap, which is a binary tree in heap order: where the label of parent elements is smaller than the labels of its children. One efficient implementation is to use an array to represent the values, where the root is at index 0, and the children of a parent node at index i are at indexes `2 * i` and `2 * i + 1`. To insert a node into the tree, we add it as a leaf, and then recursively move it up the tree until we restore heap order.

```
class binary_heap =
object (self : 'self)
  val values = new swappable_vector

  method min =
      if values#length = 0 then
          raise (Invalid_argument "heap is empty");
      values#get 0

  method add x =
      let pos = values#length in
      values#set pos x;
      self#move_up pos

  method private move_up i =
      if i > 0 then
          let parent = (i - 1) / 2 in
              if values#get i < values#get parent then begin
                  values#swap i parent;
                  self#move_up parent
              end
  end;;
```

The method move_up implements the process of restoring heap order as a recursive method (though it would be straightforward avoid the recursion and use iteration here).

The key property of private methods is that they are visible to subclasses, but not anywhere else. If you want the stronger guarantee that a method is *really* private, not even accessible in subclasses, you can use an explicit typing that omits the method. In the following code, the move_up method is explicitly omitted from the object type, and it can't be invoked in subclasses.

```
# class binary_heap :
  object
    method min : int
    method add : int -> unit
  end =
  object (self : 'self) {
    ...
    method private move_up i = ...
  end;;
```

Virtual classes and methods

A *virtual* class is a class where some methods or fields are declared, but not imple-
mented. This should not be confused with the word "virtual" as it is used in C++. In
C++, a "virtual" method uses dynamic dispatch, regular non-virtual methods use static
dispatched. In OCaml, *all* methods use dynamic dispatch, but the keyword *virtual*
means the method or field is not implemented.

In the previous section, we defined a class swappable_vector that inherits from
array_vector and adds a swap method. In fact, the swap method could be defined for
any object with get and set methods; it doesn't have to be the specific class array_vec
tor.

One way to do this is to declare the swappable_vector abstractly, declaring the methods
get and set, but leaving the implementation for later. However, the swap method can
be defined immediately.

```
class virtual abstract_swappable_vector =
object (self : 'self)
  method virtual get : int -> int
  method virtual set : int -> int -> unit
  method swap i j =
    let tmp = self#get i in
    self#set i (self#get j);
    self#set j tmp
end;;
```

At some future time, we may settle on a concrete implementation for the vector. We
can inherit from the abstract_swappable_bvector to get the swap method "for free."
Here's one implementation using arrays.

```
class array_vector =
object (self : 'self)
  inherit abstract_swappable_vector

  val mutable values = [||]
  method get i = values.(i)
  method set i x =
    self#ensure_capacity i;
```

```
            values.(i) <- x
    method length = Array.length values

    method private ensure_capacity i =
        if self#length <= i then
            let new_values = Array.create (i + 1) 0 in
                Array.blit values 0 new_values 0 (Array.length values);
                values <- new_values
end
```

Here's a different implementation using `HashTbl`.

```
class hash_vector =
object (self : 'self)
    inherit abstract_swappable_vector

    val table = Hashtbl.create 19

    method get i =
        try Hashtbl.find table i with
            Not_found -> 0

    method set = Hashtbl.add table
end;;
```

One way to view a `virtual` class is that it is like a functor, where the "inputs" are the declared, but not defined, virtual methods and fields. The functor application is implemented through inheritance, when virtual methods are given concrete implementations.

We've been mentioning that fields can be virtual too. Here is another implementation of the swapper, this time with direct access to the array of values.

```
class virtual abstract_swappable_array_vector =
object (self : 'self)
    val mutable virtual values : int array
    method private virtual ensure_capacity : int -> unit

    method swap i j =
        self#ensure_capacity (max i j);
        let tmp = values.(i) in
        values.(i) <- values.(j);
        values.(j) <- tmp
end;;
```

This level of dependency on the implementation details is possible, but it is hard to justify the use of a virtual class -- why not just define the `swap` method as part of the concrete class? Virtual classes are better suited for situations where there are multiple (useful) implementations of the virtual parts. In most cases, this will be public virtual methods.

Multiple inheritance

When a class inherits from more than one superclass, it is using *multiple inheritance*. Multiple inheritance extends the variety of ways in which classes can be combined, and it can be quite useful, particularly with virtual classes. However, it can be tricky to use, particularly when the inheritance hierarchy is a graph rather than a tree, so it should be used with care.

How names are resolved

The main "trickiness" of multiple inheritance is due to naming -- what happens when a method or field with some name is defined in more than one class?

If there is one thing to remember about inheritance in OCaml, it is this: inheritance is like textual inclusion. If there is more than one definition for a name, the last definition wins. Let's look at some artificial, but illustrative, examples.

First, let's consider what happens when we define a method more than once. In the following example, the method `get` is defined twice; the second definition "wins," meaning that it overrides the first one.

```
# class m1 =
object (self : 'self)
    method get = 1
    method f = self#get
    method get = 2
end;;
class m1 : object method f : int method get : int end
# (new m1)#f;;
- : int = 2
```

Fields have similar behavior, though the compiler produces a warning message about the override.

```
# class m2 =
# class m2 =
  object (self : 'self)
      val x = 1
      method f = x
      val x = 2
  end;;
Characters 69-74:
      val x = 2
          ^^^^^
Warning 13: the instance variable x is overridden.
The behaviour changed in ocaml 3.10 (previous behaviour was hiding.)
class m2 : object val x : int method f : int end
# (new m2)#f;;
- : int = 2
```

Of course, it is unlikely that you will define two methods or two fields of the same name in the same class. However, the rules for inheritance follow the same pattern: the last definition wins. In the following definition, the `inherit` declaration comes last, so the method definition `method get = 2` overrides the previous definition, always returning 2.

```
# class m4 = object method get = 2 end;;
# class m5 =
  object
    val mutable x = 1
    method get = x
    method set x' = x <- x'
    inherit m4
  end;;
class m5 : object val mutable x : int method get : int method set : int -> unit end
# let x = new m5;;
val x : m5 = <obj>
# x#set 5;;
- : unit = ()
# x#get;;
- : int = 2
```

To reiterate, to understand what inheritance means, replace each `inherit` directive with its definition, and take the last definition of each method or field. This holds even for private methods. However, it does *not* hold for private methods that are "really" private, meaning that they have been hidden by a type constraint. In the following definitions, there are three definitions of the private method g. However, the definition of g in m8 is not overridden, because it is not part of the class type for m8.

```
# class m6 =
  object (self : 'self)
     method f1 = self#g
     method private g = 1
  end;;
class m6 : object method f1 : int method private g : int end
# class m7 =
  object (self : 'self)
     method f2 = self#g
     method private g = 2
  end;;
class m7 : object method f2 : int method private g : int end
# class m8 : object method f3 : int end =
  object (self : 'self)
     method f3 = self#g
     method private g = 3
  end;;
class m8 : object method f3 : int end
# class m9 =
  object (self : 'self)
     inherit m6
     inherit m7
     inherit m8
  end;;
```

```
# class m9 :
  object
    method f1 : int
    method f2 : int
    method f3 : int
    method private g : int
  end
# let x = new m9;;
val x : m9 = <obj>
# x#f1;;
- : int = 2
# x#f3;;
- : int = 3
```

Mixins

When should you use multiple inheritance? If you ask multiple people, you're likely to get multiple (perhaps heated) answers. Some will argue that multiple inheritance is overly complicated; others will argue that inheritance is problematic in general, and one should use object composition instead. But regardless of who you talk to, you will rarely hear that multiple inheritance is great and you should use it widely.

In any case, if you're programming with objects, there's one general pattern for multiple inheritance that is both useful and reasonably simple, the *mixin* pattern. Generically, a *mixin* is just a virtual class that implements a feature based on another one. If you have a class that implements methods *A*, and you have a mixin *M* that provides methods *B* from *A*, then you can inherit from *M* -- "mixing" it in -- to get features *B*.

That's too abstract, so let's give an example based on collections. In Section XXX:Objecttypes, we introduced the *iterator* pattern, where an *iterator* object is used to enumerate the elements of a collection. Lots of containers can have iterators, singly-linked lists, dictionaries, vectors, etc.

```
type 'a iterator = < get : 'a; has_value : bool; next : unit >;;
class ['a] slist : object ... method iterator : 'a iterator end;;
class ['a] vector : object ... method iterator : 'a iterator end;;
class ['a] deque : object ... method iterator : 'a iterator end;;
class ['a, 'b] map : object ... method iterator : 'b iterator end;;
...
```

The collections are different is some ways, but they share a common pattern for iteration that we can re-use. For a simple example, let's define a mixin that implements an arithmetic sum for a collection of integers.

```
# class virtual int_sum_mixin =
  object (self : 'self)
    method virtual iterator : int iterator
    method sum =
      let it = self#iterator in
      let total = ref 0 in
```

```
            while it#has_value do
                total := !total + it#get;
                it#next
            done;
            !total
    end;;
# class int_slist =
  object
      inherit [int] slist
      inherit int_sum_mixin
  end;;
# let l = new int_slist;;
val l : int_slist = <obj>
# l#insert 5;;
# l#insert 12;;
# l#sum;;
- : int = 17
# class int_deque =
  object
      inherit [int] deque
      inherit int_sum_mixin
  end;;
```

In this particular case, the mixin works only for a collection of integers, so we can't add the mixin to the polymorphic class definition `['a] slist` itself. However, the result of using the mixin is that the integer collection has a method `sum`, and it is done with very little of the fuss we would need if we used object composition instead.

The mixin pattern isn't limited to non-polymorphic classes, of course. We can use it to implement generic features as well. The following mixin defines functional-style iteration in terms of the imperative iterator pattern.

```
class virtual ['a] fold_mixin =
object (self : 'self)
   method virtual iterator : 'a iterator
   method fold : 'b. ('b -> 'a -> 'b) -> 'b -> 'b =
      (fun f x ->
            let y = ref x in
            let it = self#iterator in
            while it#has_value do
               y := f !y it#get;
               it#next
            done;
            !y)
end;;

class ['a] slist_with_fold =
object
   inherit ['a] slist
   inherit ['a] fold_mixin
end;;
```