



Object Oriented Programming

Dr Andrew Rice

IA CST and NST (CS)
Michaelmas 2018/19

The Course

The OOP Course

- So far you have studied some **procedural programming** in Java and **functional programming** in ML
- Here we take your procedural Java and build on it to get object-oriented Java
- You have ticks in Java
 - This course **complements** the practicals
 - Some material appears only here
 - Some material appears only in the practicals
 - Some material appears in both: **deliberately***

* Some material may be repeated unintentionally. If so I will claim it was deliberate.

With thanks to Dr Robert Harle, who designed this course and who wrote these course materials.

So far in this term you have been taught to program using the functional programming language ML. There are many reasons we started with this, chief among them being that everything is a well-formed *function*, by which we mean that the output is dependent *solely* on the inputs (arguments). This generally makes understanding easier since it maps directly to the functions you are so familiar with from maths. In fact, if you try any other functional language (e.g. Haskell) you'll probably discover that it's very similar to ML in many respects and translation is very easy. This is a consequence of functional languages having very carefully defined features and rules.

However, if you have any real-world experience of programming, you're probably aware that functional programming is a niche choice. It is growing in popularity, but the dominant paradigm is undoubtedly *imperative* programming. Unlike their functional equivalents, imperative languages can look quite different to each other, although as time goes on there does seem to be more uniformity arising. Imperative programming is much more flexible¹ and, crucially, not all imperative languages support all of the same language concepts in the same way. So, if you just learn one language (e.g. Java) you'll probably struggle to separate the underlying programming concepts from the

¹some would say it gives you more rope to hang yourself with!

Java-specific quirks. This means that when you want to jump ship to the latest and greatest language, you may have serious conceptual difficulties.

This is quite similar to learning 'natural' languages (i.e. the languages we write and speak in). We are all proficient, if not expert, in our mother tongue. Oddly we are very good at spotting text that breaks the 'rules' of that language (grammar, spelling, etc.), but almost hopeless at identifying the rules themselves. Becoming fluent in a new language forces you to break language down into its constituent rules, and you often become better at your original language. Those who are multilingual often comment that once you know two languages well, picking up more is trivial: it's just a case of figuring out which rules to apply, possibly adding a few new rules, and learning the vocabulary. So it is with programming: once you've gone through the effort of learning a couple of languages picking up new ones is easy.

Outline

1. Types, Objects and Classes
2. Designing Classes
3. Pointers, References and Memory
4. Inheritance
5. Polymorphism
6. Lifecycle of an Object
7. Error Handling
8. Copying Objects
9. Java Collections
10. Object Comparison
11. Design Patterns
12. Design Pattern (cont.)

Books and Resources I

- OOP Concepts
 - Look for books for those learning to first program in an OOP language (Java, C++, Python)
 - *Java: How to Program* by Deitel & Deitel (also C++)
 - *Thinking in Java* by Eckels
 - *Java in a Nutshell* (O' Reilly) if you already know another OOP language
 - Java specification book: <http://java.sun.com/docs/books/jls/>
 - Lots of good resources on the web
- Design Patterns
 - *Design Patterns* by Gamma et al.
 - Lots of good resources on the web

Books and Resources II

- Also check the course web page
 - Updated notes (with annotations where possible)
 - Code from the lectures
 - Sample tripos questions

<http://www.cl.cam.ac.uk/teaching/current/OOProg/>

- **And the Moodle site "Computer Science Paper 1 (1A)"**
- **Watch for course announcements**

There are many books and websites describing the basics of OOP. The concepts themselves are quite abstract, although most texts will use a specific language to demonstrate them. The books I've listed favour Java but you shouldn't see that as a dis-recommendation for other books. In terms of websites, Oracle produce a series of tutorials for Java, which cover OOP: <http://java.sun.com/docs/books/tutorial/> but you'll find lots of other good resources if you search.

0.1 Languages and Exams

The 'examinable' imperative language for this Paper 1 course is Java, and you **won't be required or expected to program in anything else**. However, Java doesn't support all the interesting concepts found in imperative programming (yet) so other languages may be used to demonstrate certain features. The languages are non-examinable insofar as they won't feature in the OOP questions. The *concepts* they illustrate might, though.

This year I've added in some extra stuff on C++ and Python. This is because those in the Natural Sciences Tripos are likely to be looking at C++ next year. Similarly, Computer Science Tripos students will be using python for their Scientific Computing course over the Christmas vacation. I've clearly marked out the C++ and python bits in coloured boxes. E.g.

C++

Things in these boxes will describe non-examinable C++ or C-like languages

Python

Things in these boxes will describe non-examinable python

0.2 Ticks

There are five OOP ticks, all in Java. They follow on from the work you did in module three of the Pre-arrival course, using the concepts from lectures to build ever-better Game of Life implementations. Four of the ticks have deadlines this term. The fifth you all do over the vacation.

Lecture 1

Languages, Types, Objects and Classes

1.1 Imperative, Procedural, Object Oriented

Types of Languages

- **Declarative** - specify what to do, not how to do it. i.e.
 - E.g. HTML describes what should appear on a web page, and not how it should be drawn to the screen
 - E.g. SQL statements such as "select * from table" tell a program to get information from a database, but not how to do so
- **Imperative** – specify both what and how
 - E.g. "triple x" might be a declarative instruction that you want the variable x tripled in value. Imperatively we would have "x=x*3" or "x=x+x*x"

Firstly a recap on declarative vs imperative:

Declarative languages specify *what* should be done but not necessarily *how* it should be done. In a functional language such as ML you specify what you want to happen essentially by providing an example of how it can be achieved. The ML compiler/interpreter can do exactly that or something equivalent (i.e. it can do something else but it *must* give the same output or result for a given input).

Imperative languages specify exactly *how* something should be done. You can consider an imperative compiler to act very robotically—it does *exactly* what you tell it to and you can easily map your code to what goes on at a machine code level;

There's a nice meme¹ that helps here:

Functional programming is like describing your problem to a mathematician. Imper-

¹attributed to arcus, #scheme on Freenode

ative programming is like giving instructions to an idiot.

Those of you who have done the Databases course will have encountered SQL. Even those that haven't can probably decipher the language to a point—here's a trivial example:

```
select * from person_table where name="Bob";
```

This gets all entries from the database `person_table` where the name column contains "Bob". This language is highly functional: I have specified what I want to achieve but given no indication as to *how* it should be done. The point is that the SQL language simply doesn't have any way to specify how to look something up—that functionality is built into the underlying database and we (as users of the database) shouldn't concern ourselves.

On the other hand, the machine and assembly code you saw in the pre-arrival course are arguably the ultimate imperative languages. However, as you can imagine, programming in them directly isn't much fun. Other imperative languages have evolved from these, each attempting to make the coding process more human-friendly.

A key innovation was the use of procedures (equate these to functions for now) to form *procedural* programming. The idea is to group statements together into procedures/functions that can be called to manipulate the state. Code ends up as a mass of functions plus some logic that calls them in an appropriate way to achieve the desired result. That's exactly how you used Java in the pre-arrival course.

OOP is an extension to procedural programming (so still imperative) where we recognise that these procedures/functions can themselves be usefully grouped (e.g. all procedures that update a `PackedLong` vs all procedures that draw to screen) *and* furthermore that it often makes sense to group the state they affect with

them (e.g. group the `PackedLong` methods with the underlying `long`).

Wait...

You might be struggling with the fact you specified functions in ML that appear to fit the imperative mould: there were statements expressing *how* to do something. Think of these as specifying the desired result by giving an *example* of how it might be obtained. Exactly what the compiler does may or may not be the same as the example—so long as it gives the same outputs for all inputs, it doesn't matter.

ML as a Functional Language

- **Functional** languages are a subset of declarative languages
 - ML is a functional language
 - It may appear that you tell it how to do everything, but you should think of it as providing an explicit example of what should happen
 - The compiler may **optimise** i.e. replace your implementation with something entirely different but 100% equivalent.

Although it's useful to paint languages with these broad strokes, the truth is today's high-level languages should be viewed more as a collection of features. ML is a good example: it is certainly viewed as a functional language but it also supports all sorts of imperative programming constructs (e.g. references). Similarly, the compilers for most imperative languages support *optimisations* where they analyse small chunks of code and implement something different at machine-level to increase performance—this is of course a trait of declarative programming². So the boundaries are blurred, but ML is predominantly functional and Java predominantly imperative, and we'll take those narrow views in this course.

1.1.1 Java as a Procedural Language

We used Java to introduce you to general procedural programming in the pre-arrival course. This is not ideal since Java is designed as an object oriented lan-

²Note that we need a way to switch off optimisations because they don't always work due to the presence of side effects in functions. Tracking down an error in an optimisation is painful: the 'bug' isn't in the code you've written..!

guage through and through. Trying to force it to act entirely procedurally required a number of ugly hacks that you may have noticed:

- all the functions had to be **static** (we'll explain that shortly);
- we had to create placeholder classes (`public class TinyLife...`); and.
- we had a lot of annoying 'boilerplate' code that seemed rather unnecessary (e.g. `public static void main(...`

Over the next few lectures the reason why these irritations are necessary should become apparent. Ideally, we would have used a purely procedural language that wouldn't require such hacks, but we didn't want to force you to setup and learn yet another language in one year.

1.2 Procedures, Functions, Methods etc

Up to now, we've treated procedures as the same as functions. Herein it's useful for us to start making some distinctions. One of the key properties of functional languages is that they use *proper functions*. By this I mean functions that have the same properties as those you find in maths:

- All functions return a (non-void) result;
- the result is *only* dependent on the inputs (arguments); and
- no state outside of the function can be modified (i.e. no "side effects").

Restricting ourselves to proper functions makes it easier for the compiler to assert declarative-like optimisations: a given function is completely standalone (i.e. dependent on no state). You can pluck out an arbitrary function without reference to anything else in the program and optimise it as you see fit.

Function Side Effects

- Functions in imperative languages can use or alter larger system state → *procedures*

Maths: $m(x,y) = xy$

ML: `fun m(x,y) = x*y;`

Java: `int m(int x, int y) = x*y;`

```
int y = 7;
int m(int x) {
    y=y+1;
    return x*y;
}
```

Procedures have similarities to (proper) functions but permit side effects and hence break the three rules given above. Here's another example:

```
fun add(x,y)=x+y;
add(1,2);
```

This ML (proper) function will always return 3 for the input (1,2) regardless of the statements before or after it. In Java, there is a direct equivalent:

```
static int addimp(int x, int y) {
    return x+y;
}
addimp(1,2);    // 3
```

but we could also write something a bit more nefarious:

```
static int z=0; // this is some global state
static int addimp(int x, int y) {
    z=z+1;
    return x+y+z;
}
addimp(1,2);    // 4
addimp(1,2);    // 5
addimp(1,2);    // 6
```

Eeek! Three calls with the same arguments gives three different answers. You certainly don't see that in maths. The problem is that the output is dependent on some other state in the system (z), which it changes (a *side effect* of calling it). Given only the procedure name and its arguments, we cannot predict what the state of the system will be after calling it without reading the full procedure definition and analysing the current state of the computer. One way of looking at this is that the procedure is a proper function that takes

as input the argument you explicitly supply *plus* implicitly all of the system state.

To really hammer this home, you can now have useful functions with no arguments that return nothing (indicated as `void` in Java)—both rather useless in maths:

void Procedures

- A `void` procedure returns nothing:

```
int count=0;

void addToCount() {
    count=count+1;
}
```

Health warning: Many imperative programmers use the word 'function' as a synonym for 'procedure'. Even in these lectures I will use 'function' loosely. You will have to use your intelligence when you hear the words.

Procedures are much more powerful, but as that awful line in Spiderman goes, "with great power comes great responsibility". Now, that's not to say that imperative programming makes you into some superhuman freak who runs around in his pyjamas climbing walls and battling the evil functionals. It's just that it introduces a layer of complexity into programming that *might* make the job run faster, but almost certainly makes the job harder.

1.3 Recap: Control Flow

You've covered java's control flow in the pre-arrival course but it seems wrong not to at least mention it here (albeit the coverage in lectures will be very brief). The associated statements are classified into *decision-making*, *looping* and *branching*.

Decision-making is quite simple: `if (...) {...} else {...}` is the main thing we care about. Java steals this syntax from C-like languages, so it's the same there.

Looping doesn't require recursion (yay!) since we have `for` and `while`:

Control Flow: Looping

for(*initialisation; termination; increment*)

```
for (int i=0; i<8; i++) ...
```

```
int j=0; for(; j<8; j++) ...
```

```
for(int k=7; k>=0; k--) ...
```

while(*boolean_expression*)

```
int i=0; while (i<8) { i++; ...}
```

```
int j=7; while (j>=0) { j--; ...}
```

These examples all loop eight times. The following code loops over the entirety of an array (the `for` approach is more usual for this task—why do you think `while` is considered bad form here?):

Control Flow: Looping Examples

```
int arr[] = {1,2,3,4,5};
```

```
for (int i=0; i<arr.length; i++) {  
    System.out.println(arr[i]);  
}
```

```
int i=0;  
while (i<arr.length) {  
    System.out.println(arr[i]);  
    i=i+1;  
}
```

For branching, we mainly care about `return`, `break` and `continue`:

Control Flow: Branching I

- Branching statements interrupt the current control flow
- **return**
 - Used to return from a function at any point

```
boolean linearSearch(int[] xs, int v) {  
    for (int i=0; i<xs.length; i++) {  
        if (xs[i]==v) return true;  
    }  
    return false;  
}
```

Control Flow: Branching II

- Branching statements interrupt the current control flow
- **break**
 - Used to jump out of a loop

```
boolean linearSearch(int[] xs, int v) {  
    boolean found=false;  
    for (int i=0; i<xs.length; i++) {  
        if (xs[i]==v) {  
            found=true;  
            break; // stop looping  
        }  
    }  
    return found;  
}
```

Control Flow: Branching III

- Branching statements interrupt the current control flow
- **continue**
 - Used to skip the current iteration in a loop

```
void printPositives(int[] xs) {  
  
    for (int i=0; i<xs.length; i++) {  
        if (xs[i]<0) continue;  
        System.out.println(xs[i]);  
    }  
}
```

In passing, a reminder about scopes since this caused issues for some in the pre-arrival course. In Java a scope is defined using curly braces—i.e. `{...}`. The definition of a method used a scope, loops used a scope, etc:

```
//...  
public static void someMethod { // scope starts  
  
} // scope ends  
  
//...
```

In fact, you can create a scope anywhere in Java using the braces. Recall that any variable you *declared* inside a scope did not last beyond it. So:

```
//...  
  
int x=1;  
  
{ // Create an arbitrary scope
```

```

int y=2;
x=3;

System.out.println(x); // Fine: x is 3
System.out.println(y); // Fine: y is 2
} // End of scope

System.out.println(x); // Fine: x is 3
System.out.println(y); // Error: y is undefined

//...

```

C++

C++ (and all C-like languages) perform control flow and define scopes in the same way.

Python

Python looks quite different, although the concepts are all the same. Python defines the start of a scope using a colon. It is notorious for using white space (spaces, tabs, etc) to define the extent of the scope. Each scope has an indent; when the indent goes, the scope is over. So you get code like:

```

for i in range(0,10): # start scope 1
    print i
    for j in range(0,5): # start scope 2
        print i*10+j
    print "loop"

```

Here, scopes close when the associated indent is lost. Scope 1 finishes just after the line `print "loop"`. Scope 2 finishes just before it. The whitespace really matters here: one misplaced or missing space and it won't compile.. This is very much a 'marmite' feature of python.

1.4 Values, Variables and Types

1.4.1 State Mutability

Immutable to Mutable Data

```

ML
- val x=5;
> val x = 5 : int
- x=7;
> val it = false : bool
- val x=9;
> val x = 9 : int

```

```

Java
int x=5;
x=7;

int x=9;

```

In ML you had *values* and in Java you have *variables*. A simple way to ensure that functions in ML do not depend on external state is to make all state constant, or *immutable*. In ML you could write:

```

val x = 5;
x=7;
val x=9;

```

But these almost certainly didn't do what you expected the first time you tried them. The first line (`val x = 7`) creates a chunk of memory, sets the contents to 7 and associates it with a label `x`. The second (`x = 5`) you probably wanted to reassign the value, but—since the values are constant—it actually performed a comparison of `x` and 5, giving `false`! The third line `val x=5` actually creates another value in memory, sets the value to 5 and reassigns the *label* `x` to point to it. The original value of 7 remains untouched in memory: you just can't update it. So now you should be able to understand the behaviour of:

```

val x = 7;
fun f(a)=x*a;
f(3);
val x=5;
f(3);

```

Java doesn't have the shackles of proper functions so we can have variables that can be updated (i.e. *mutable* state)—this is actually the essence of imperative programming:

```

int x = 7; // create x and init to 7

```



```
x=5;           // change the value of x
x==5;         // compare the value of x
int x = 9;    // error: x already created
```

1.4.2 Explicit Types vs Type Inference

Types and Variables

- Most imperative languages don't have type inference

```
int x = 512;
int y = 200;
int z = x+y;
```

- The high-level language has a series of *primitive* (built-in) types that we use to signify what's in the memory
 - The compiler then knows what to do with them
 - E.g. An "int" is a primitive type in C, C++, Java and many languages. It's usually a 32-bit signed integer
- A variable is a name used in the code to refer to a specific instance of a type
 - x,y,z are variables above
 - They are all of type int

In ML you created values of various types (*real*, *int*, etc). You were (correctly) taught to avoid explicitly assigning types wherever possible: ML had the capability to infer types. This was particularly useful because it allowed polymorphism to avoid writing separate functions for integers, reals, etc. Every now and then you had to help it out and manually specify a type, but ML's *type inference* is essentially a really nice feature to have baked in. (I acknowledge that ML's error messages about type errors could be a little less... cryptic).

Java and C-like languages are *statically typed*. This means that every variable name must be bound to a type, which is given when it is declared. So you must say `int x=1`; and not `x=1`—the latter only works if you are updating a previously-declared `x`. You can't redeclare the type:

```
int x = 1;
String x = "Hi"; // Compile error
```

For methods, you must specify the type of the output (its 'return type') *and* the type(s) of its argument(s).³

Python

Python is *dynamically typed*, which means each variable name is *not* assigned a type. It can be

³Soon we meet Generics, where the type is left more open. However, there is a type assigned to everything, even if it's just a placeholder.

bound to something that does have a type. E.g.

```
x = 1;
x = "Hi";
```

This is fine. The first line creates a name `x` and binds it to an integer. The second line binds it to a string. Similarly methods don't need to have declared types for the return or the arguments. This gives us interesting ML-like flexibility:

```
def f(a,b):
    return a+b

f(1,2)           # return 3
f(1.0,2.0)      # returns 3.0
f("He","llo")  # returns "Hello"
```

However, not needing to declare a name can catch you out:

```
myVariable=7
myvariable=myVariable*myVariable # Typo!!
```

Here I made a typo: `myvariable` not `myVariable`. Java would spot this and complain. But python just assumes I wanted a new variable with the name `myvariable`...

If you're sticking around in CST next yer, you'll learn a lot more about this concept.

You have already met the primitive (built-in) types of Java in the pre-arrival course, but here's a recap

E.g. Primitive Types in Java

- "Primitive" types are the built in ones.
 - They are building blocks for more complicated types that we will be looking at soon.
- boolean – 1 bit (true, false)
- char – 16 bits
- byte – 8 bits as a signed integer (-128 to 127)
- short – 16 bits as a signed integer
- int – 32 bits as a signed integer
- long – 64 bits as a signed integer
- float – 32 bits as a floating point number
- double – 64 bits as a floating point number

C++

For any C/C++ programmers out there: yes, Java looks a lot like the C syntax. But watch out for the obvious gotcha—a char in C is a byte (an ASCII character), whilst in Java it is two bytes (a Unicode character). If you have an 8-bit number in Java you may want to use a byte, but you also need to be aware that a byte is *signed*..!

1.4.3 Polymorphism vs Overloading

Overloading Functions

- Same function name
- Different arguments
- Possibly different return type

```
int myfun(int a, int b) {...}
float myfun(float a, float b) {...}
double myfun(double a, double b) {...}
```

- But not just a different return type

```
int myfun(int a, int b) {...}
float myfun(int a, int b) {...}
```

X

Since Java demands that all types are explicit (disregarding Generics, which we'll come to soon), we rather lose the ability to write one function that can be applied to multiple types—the cool⁴ polymorphism you saw in ML. Instead we can make use of procedure *overloading*. This allows you to write multiple methods with the same name but different argument types and/or numbers:

```
int myfun(int a,int b,int c) {
    // blah blah blah
}

int myfun(double a, double b, double c) {
    // blah blah blah
}
```

When you call `myfun` the compiler looks at the argument types and picks the function that best matches. This is nowhere near as elegant as in ML (I have to write out a separate function for every argument set) but at least when it's being used there is naming consistency.

⁴Your definition of 'cool' may vary

Python

Python's dynamic typing means that you don't get method overloading:

```
def func(a,b):
    return a+b

def func(a,b,c):
    return a+b+c

func(1,2) # Error - func needs
          # 3 srguments

func(1,2,3) # returns 6
```

That said, since you don't specify the types, you get some ML-like polymorphism (note that the mechanism by which this is achieved is different to ML, which is statically typed—see Part IB):

```
def func(a,b):
    return a+b

func(1,2) # returns 3
func(1.0,2.0) # returns 3.0
func("A","B") # returns "AB"
func("A",1) # Error
```

In passing, we note that we talk about function *prototypes* or *signatures* to mean the combination of return type, function name and argument set—i.e. the first line such as `int myfun(double a, double b, double c)`.

Function Prototypes

- Functions are made up of a **prototype** and a **body**
 - Prototype specifies the function name, arguments and possibly return type
 - Body is the actual function code

```
fun myfun(a,b) = ...;

int myfun(int a, int b) {...}
```

1.5 Classes and Objects

Sooner or later, using just the built-in primitive types becomes restrictive. You saw this in ML, where you could create your own types. This is also possible in imperative programming and is, in fact, the crux of object oriented programming.

```
Custom Types

datatype 'a seq = Nil
  | Cons of 'a * (unit -> 'a seq);

public class Vector3D {
  float x;
  float y;
  float z;
}
```

Let's take a simple example: representing 3D vectors (x,y,z). We could keep independent variables in our code. e.g.

```
float x3=0.0;
float y3=0.0;
float z3=0.0;

void add_vec(float x1, float y1, float z1,
             float x2, float y2, float z2) {

  x3=x1+x2;
  y3=y1+y2;
  z3=z1+z2;
}
```

Clearly, this is not very elegant code. Note that, because I can only return one thing from a function, I can't return all three components of the answer (ML's tuples don't exist here—sorry!). Instead, I had to manipulate external state. You see a lot of this style of coding in procedural C coding. Yuk.

We would rather create a new type (call it `Vector3D`) that contains all three components, as per the slide. In OOP languages, the definition of such a type is called a *class*. But it goes further than just being a grouping of variables...

1.5.1 State *and* Behaviour

```
State and Behaviour

datatype 'a seq = Nil
  | Cons of 'a * (unit -> 'a seq);

fun hd (Cons(x,_)) = x;

public class Vector3D {
  float x;
  float y;
  float z;

  void add(float vx, float vy, float vz) {
    x=x+vx;
    y=y+vy;
    z=z+vz;
  }
}
```

What we've done so far looks a lot like procedural programming. There you create custom types to hold your data or state, then write a ton of functions/procedures to manipulate that state, and finally create your program by sequencing the various procedure calls appropriately. ML was similar: each time you created a new type (such as sequences), you also had to construct a series of helper functions to manipulate it (e.g. `hd()`, `tail()`, `merge()`, etc.). There was an implicit link between the data type and the helper functions, since one was useless without the other.

OOP goes a step further, making the link explicit by having the class represent the type *and* the helper functions that manipulate it. OOP classes therefore glue together the state (i.e. variables) and the behaviour (i.e. functions or procedures) to create a complete unit of the type.

```
Loose Terminology (again!)

State
Fields
Instance Variables
Properties
Variables
Members

Behaviour
Functions
Methods
Procedures
```

Having made all that fuss about 'function' and 'procedure', it only gets worse here: when we're talking about a procedure inside a class, it's more properly

called a *method*. In the wild, you'll find people use 'function', 'procedure' and 'method' interchangeably. Thankfully you're all smart enough to cope!

1.5.2 Instantiating classes: Objects

Classes, Instances and Objects

- Classes can be seen as templates for representing various **concepts**
- We create **instances** of classes in a similar way. e.g.

```
MyCoolClass m = new MyCoolClass();
MyCoolClass n = new MyCoolClass();
```

makes two instances of class MyCoolClass.

- An instance of a class is called an **object**

So a class is a grouping of state (data/variables) and behaviour (methods). Whenever we create an *instance* of a class, we call it an *object*. The difference between a class and an object is thus very simple, but you'd be surprised how much confusion it can cause for novice programmers. *Classes* define what properties and procedures every object of the type should have (a template if you like), while each *object* is a specific implementation with particular values. So a *Person* class might specify that a *Person* has a name and an age. Our program may instantiate two *Person* objects—one might represent 40-year old Bob; another might represent 20 year-old Alice. Programs are made up of lots of objects, which we manipulate to get a result (hence "object oriented programming").

We've already seen how to create (define) objects in the very last lesson of the pre-arrival course. There we had:

```
// Define p as a new Vector3 object
Vector3 p = new Vector3();
// Reassign p to a new Vector3 object
p = new Vector3()
```

The things to note are that we needed a special *new* keyword to instantiate an object; and that we pass it what looks to be a method call (*Vector3()*). Indeed, it *is* a method, but a special one: it's called a *constructor* for the class.

1.5.3 Defining Classes

Defining a Class

```
public class Vector3D {
    float x;
    float y;
    float z;

    void add(float vx, float vy, float vz) {
        x=x+vx;
        y=y+vy;
        z=z+vz;
    }
}
```

To define a class we need to declare the state and define the methods it is to contain. Here's a very simple Java class containing one integer as its state and a single method:

```
class MyShinyClass {
    int x;

    void setX(int xinput) {
        x=xinput;
    }
}
```

You were defining classes like this in your Pre-arrival course code. Except there it was peppered with the words *public* and *static*—we'll look at both shortly.

1.5.4 Constructors

Constructors

```
MyObject m = new MyObject();
```

- You will have noticed that the RHS looks rather like a function call, and that's exactly what it is.
- It's a method that gets called when the object is constructed, and it goes by the name of a **constructor** (it's not rocket science). It maps to the datatype constructors you saw in ML.
- We use constructors to initialise the state of the class in a convenient way
 - A constructor has **the same name** as the class
 - A constructor has **no return type**

You can define one or more *constructors* for a class.

These are simply methods that are run when the object is created. As with many OOP features, not all languages support it. Python, for example, doesn't have constructors.

Python

Python classes *do* have a single `__init__` method in each class that acts a bit like a constructor but technically isn't. Python fully constructs the object, then passes it to the `__init__` method if it exists—similar, but not quite the same thing as in Java. In particular, you can only have one `init` method in a class.

In Java, C++ and most other OOP languages, constructors have two properties:

1. they have the same name as the class; and
2. they have *no* return type.

You can't specify a return type for a constructor because it is always called using the special `new` keyword, which must return a reference to the newly constructed object. You can, however, specify arguments for a constructor in the usual way for a method:

```
class MyShinyClass {
    int x;

    MyShinyClass(int x_init) {
        setX(x_init);
    }

    void setX(int xinput) {
        x=xinput;
    }
}
```

Here, the constructor is used to initialise the member variable `x` to a value passed in. We would specify this when using the `new` keyword to create an object. e.g.

```
MyShinyClass m = new MyShinyClass(42);
```

C++

In C++, objects can be instantiated in two ways:

```
// Assume we have defined a class MyClass

MyClass x=MyClass();
MyClass *y = new MyClass();
```

The difference is subtle but important. The `x` is associated with an object. When `x` dies (goes out of scope), so does the object. `y` is actually a pointer to a newly created object of type `MyClass`. We discuss pointers soon, but for now you can just note that when `y` goes out of scope, the object it points too sticks around. We must manually delete it (forgetting to do so is a classic C++ error).

Overloaded Constructors

```
public class Vector3D {
    float x;
    float y;
    float z;

    Vector3D(float xi, float yi, float zi) {
        x=xi;
        y=yi;
        z=zi;
    }

    Vector3D() {
        x=0.f;
        y=0.f;
        z=0.f;
    }

    // ...
    Vector3D v = new Vector3D(1.f,0.f,2.f);
    Vector3D v2 = new Vector3D();
}
```

You can have multiple constructors in Java by overloading the method:

```
class MyShinyClass {
    int x;

    MyShinyClass() {
        set
    }

    MyShinyClass(int x_init) {
        setX(x_init);
    }

    void setX(int xinput) {
        x=xInput;
    }
}

//...

// An object with x set to 42
MyShinyClass m = new MyShinyClass(42);

// And object with x set to 0
MyShinyClass m2 = new MyShinyClass();
```

Default Constructor

```
public class Vector3D {
    float x;
    float y;
    float z;
}

Vector3D v = new Vector3D();
```

- No constructor provided
- So blank one generated with no arguments

If you don't specify any constructor at all, Java fills in a *default constructor* for you. This takes no arguments and does nothing, other than allowing you to make objects. i.e.

```
class MyShinyClass {
}
```

is converted to

```
class MyShinyClass {
    MyShinyClass() { }
}
```

so you can write `MyShinyClass m = new MyShinyClass();`.

1.5.5 Static

Sometimes there is state that is more logically associated with a class than an object. An example may help here:

Class-Level Data and Functionality I

- A **static** field is created only once in the program's execution, despite being declared as part of a class

```
public class ShopItem {
    float mVATRate;
    static float sVATRate;
    ....
}
```

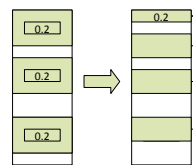
One of these created every time a new ShopItem is instantiated. Nothing keeps them all in sync.

Only one of these created ever. Every ShopItem object references it.

So a **static** variable is only instantiated once per *class* not per *object*. You don't even need to create an object to access a static variable. Just writing `ShopItem.sVATRate` would give you access.

You see examples of this in the `Math` class provided by Java: you can just call `Math.PI` to get the value of pi, rather than creating a `Math` object first. The same goes for methods: you write `Math.sqrt(...)` rather than having to first instantiate a `Math` object.

Class-Level Data and Functionality II



- Auto synchronised across instances
- Space efficient

- Also static methods:

```
public class Whatever {
    public static void main(String[] args) {
        ...
    }
}
```

Methods can also be **static**. In this case they must not use anything other than local or static variables. So it can't use anything that is instance-specific (i.e. non-static member variables are out).

Looking back at the pre-arrival course, we really wanted something that had no class notion at all. The closest we could get in Java was to make everything static so there weren't any objects floating around. Not pretty, but it got the job done.

Why use Static Methods?

- Easier to debug (only depends on static state)
- Self documenting
- Groups related methods in a Class without requiring an object
- The compiler can produce more efficient code since no specific object is involved

```
public class Math {
    public float sqrt(float x) {...}
    public double sin(float x) {...}
    public double cos(float x) {...}
}
```

```
...
Math mathobject = new Math();
mathobject.sqrt(9.0);
...
```

vs

```
public class Math {
    public static float sqrt(float x) {...}
    public static float sin(float x) {...}
    public static float cos(float x) {...}
}
```

```
...
Math.sqrt(9.0);
...
```

Lecture 2

Designing Classes

2.1 Identifying Classes

What Not to Do

- Your ML has doubtless been one big file where you threw together all the functions and value declarations
- Lots of C programs look like this :-(
- We *could* emulate this in OOP by having one class and throwing everything into it

- We can do (much) better

Having one massive class, `MyApplication` perhaps, with all the state and behaviour in it, is a surprisingly common novice error. This achieves nothing (in fact it just adds boilerplate code). Instead we aim to have multiple classes, each embodying a well-defined *concept*.

Identifying Classes

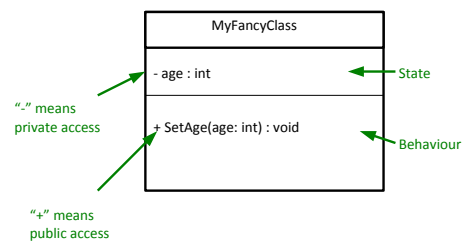
- We want our class to be a **grouping of conceptually-related state and behaviour**
 - One popular way to group is using grammar
 - **Noun** → **Object**
 - **Verb** → **Method**
- “A simulation of the Earth's orbit around the Sun”

Very often classes follow naturally from the problem domain. So, if you are making a snooker game, you might have an object to represent the table; to repre-

sent each ball; to represent the cue; etc. Identifying the best possible set of classes for your program is more of an art than a science and depends on many factors. However, it is usually straightforward to develop sensible classes, and then keep on refining them (“refactoring”) until we have something better.

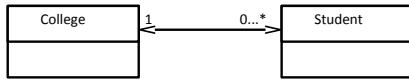
A helpful way to break your program down is in terms of tangible things—represented by the nouns you would use when describing the program. Similarly, the verbs often map well to the behaviour required of your classes. Think of these as guidelines or rules of thumb, not rules.

UML: Representing a Class Graphically



The graphical notation used here is part of UML (Unified Modeling Language). UML is a standardised set of diagrams that can be used to describe software independently of any programming language used to implement it. UML contains many different diagrams. In this course we will only be interested in basic *UML class diagrams* such as the one in the slide.

The has-a Association



- Arrow going left to right says “a College has zero or more students”
- Arrow going right to left says “a Student has exactly 1 College”
- What it means in real terms is that the College class will contain a variable that somehow links to a set of Student objects, and a Student will have a variable that references a College object.
- Note that we are only linking *classes*: we don't start drawing arrows to primitive types.

Note that the arrowhead must be ‘open’. It is normal to annotate the head with the multiplicity (i.e. how many of these something has), but some programmers are lax on this (for examination purposes, you *are* expected to annotate the heads). I've shown a dual-headed arrow; if the multiplicity value is zero, you can leave off the arrowhead and annotation entirely.

2.2 OOP Concepts

OOP Concepts

- OOP provides the programmer with a number of important concepts:
 - Modularity
 - Code Re-Use
 - Encapsulation
 - Inheritance
 - Polymorphism
- Let's look at these more closely...

Let's be clear here: OOP doesn't *enforce* the correct usage of the ideas we're about to look at. Nor are the ideas exclusively found in OOP languages. The main point is that OOP *encourages* the use of these concepts, which we believe is good for software design.

2.2.1 Modularity and Code Re-Use

Modularity and Code Re-Use

- You've long been taught to break down complex problems into more tractable sub-problems.
- Each class represents a sub-unit of code that (if written well) can be **developed, tested and updated independently** from the rest of the code.
- Indeed, two classes that achieve the same thing (but perhaps do it in different ways) can be swapped in the code
- Properly developed classes can be used in other programs without modification.

Modularity is extremely important in OOP. It's a common Computer Science trick: break big problems down into chunks and solve each chunk. In this case, we have large programs, meaning scope for lots of coding bugs. By identifying objects in our problem, we can write classes that represent them. Each class can be developed, tested and maintained independently of the others. Then, when we sequence them together to make our larger program, there are far fewer places where it can go wrong.

There is a further advantage to breaking a program down into self-contained objects: those objects can be ripped from the code and put into other programs. So, once you've developed and tested a class that embodies everything about donkey, say, you can use it in lots of other programs with minimal effort. Even better, the classes can be distributed to other programmers so they don't have to reinvent the wheel. Therefore OOP strongly encourages software *re-use*.

As an aside, modularity often goes further than the classes/objects. Java has the notion of **packages** to group together classes that are conceptually linked. We use them in the ticks to group together the code you write. It allows us to distinguish between the `ArrayLife` class you write and the `ArrayLife` class someone else writes (e.g. `uk.ac.cam.rkh23.ArrayLife` rather than the ambiguous `ArrayLife`).

C++

C++ has namespaces, which are very similar to Java's packages.

Python

Python has modules and packages.

2.2.2 Encapsulation and Information Hiding

Encapsulation I

```
class Student {
    int age;
};

void main() {
    Student s = new Student();
    s.age = 21;

    Student s2 = new Student();
    s2.age=-1;

    Student s3 = new Student();
    s3.age=10055;
}
```

This code defines a basic `Student` class, with only one piece of state per `Student`. In the `main()` method we create three instances of `Students`. We observe that nothing stops us from assigning nonsensical values to the age.

Encapsulation II

```
class Student {
    private int age;

    boolean setAge(int a) {
        if (a>=0 && a<130) {
            age=a;
            return true;
        }
        return false;
    }

    int getAge() {return age;}
}

void main() {
    Student s = new Student();
    s.setAge(21);
}
```

Here we have assigned an *access modifier* called `private` to the age variable. This means nothing external to the class (i.e. no piece of code defined outside of the class definition) can read or write the age variable directly.

Another name for encapsulation is *information hiding* or even *implementation hiding* in some texts. The idea is that a class should expose a clean interface that allows full interaction with it, but should expose nothing about its internal state or how it manipulates it.

From now on, apply this rule: **all state is private unless there is a very good reason for it not to be.**

To get access to the age variable we define a `getAge()` and a `setAge()` method to allow read and write, respectively. On the face of it, this is just more code to achieve the same thing. However, we have new options: by omitting `setAge()` altogether we can prevent anyone modifying the age (thereby adding immutability!); or we can provide sanity checks in the `setAge()` code to ensure we can only ever store sensible values.

Encapsulation III

```
class Location {
    private float x;
    private float y;

    float getX() {return x;}
    float getY() {return y;}

    void setX(float nx) {x=nx;}
    void setY(float ny) {y=ny;}
}

class Location {
    private Vector2D v;

    float getX() {return v.getX();}
    float getY() {return v.getY();}

    void setX(float nx) {v.setX(nx);}
    void setY(float ny) {v.setY(ny);}
}
```

Here we have a simple example where we wish to change the underlying representation of a co-ordinate (x,y) from raw primitives to a custom `Vector2D` object. We can do this without changing the public interface to the class and hence without having to update any piece of code that uses the `Location` class.

You may hear people talking about *coupling* and *cohesion*. Coupling refers to how much one class depends on another. High coupling is bad since it means changing one class will require you to fix up lots of others. Cohesion is a qualitative measure of how strongly related everything in the class is—we strive for high cohesion. Encapsulation helps to minimise coupling and maximise cohesion.

Access Modifiers

	Everyone	Subclass	Same package (Java)	Same Class
private				X
package (Java)			X	X
protected		X	X	X
public	X	X	X	X

OOP languages feature some set of access modifiers that allow us to do various levels of data hiding. Java uses `public`, `protected`, `private` and `package`. We haven't yet talked about subclassing, so don't worry about that column yet.

C++

C++ has `public`, `protected`, and `private`, with the same meanings as Java. There's no `package` equivalent.

Python

Here we see a big departure. Python doesn't have access modifiers, despite being considered an OOP language! Everything in a python class can be directly accessed by anything external to it (i.e. the equivalent of `public` for everything).

Instead, there is a *convention* that variables or functions starting with an underscore (e.g. `_myvar` or `_myfunc` should be treated as private—i.e. you shouldn't touch them directly. The philosophy is “we're all adults here”. You can decide yourself whether you think this is a good idea. My betting is that you will be as suspicious of this philosophy as I am once you have developed anything substantial in a team!

2.3 Immutability

The discussion of access modifiers leads us naturally to talk about immutability. You should recall from FoCS that every value in ML is immutable: once it's set, it can't be changed. From a low-level perspective, writing `val x=7`; allocates a chunk of memory and sets it to the value 7. Thereafter you can't change that chunk of memory. You *could* reassign the *label* by writing `val x=8`; but this sets a new chunk of memory to the value 8, rather than changing the original chunk (which sticks around, but can't be addressed directly now since `x` points elsewhere).

It turns out that immutability has some serious advantages when concurrency is involved—knowing that nothing can change a particular chunk of memory means we can happily share it between threads without worry of contention issues. It also has a tendency to make code less ambiguous and more readable. It is, however, more efficient to manipulate previously allocated memory rather than constantly allocate new chunks. In OOP, we can have the best of both worlds.

Immutability

- Everything in ML was immutable (ignoring the reference stuff). Immutability has a number of advantages:
 - Easier to construct, test and use
 - Can be used in concurrent contexts
 - Allows lazy instantiation
- We can use our access modifiers to create immutable classes

To make a class immutable:

- Make sure all state is `private`.
- Consider making state `final` (this just tells the compiler that the value never changes once constructed).
- Make sure no method tries to change any internal state.

To quote *Effective Java* by Joshua Bloch:

“Classes should be immutable unless there's a very good reason to make them mutable... If a class cannot be made immutable, limit its mutability as much as possible.”

C++

Java's `final` is close in spirit to C++'s `const` keyword. However, C++ lets you take a mutable class and make a particular *instance* immutable using the `const` keyword. This can be useful, but it also tends to cause confusion: sometimes you are mutating a variable without realising it is `const` and you just see a host of compiler errors. The problem isn't applicable to Java due to the way Java only deals with references (we'll come back to this soon).

2.4 Parameterised Types

We commented earlier that Java lacked the nice polymorphism that type inference gave us. Languages evolve, however, and it has been retrofitted to the language via something called *Generics*. It's not quite the same (it would have been too big a change to put

in full type inference), but it does give us similar flexibility.

Parameterised Classes

- ML's polymorphism allowed us to specify functions that could be applied to multiple types

```
> fun self(x)=x;
val self = fn : 'a -> 'a
```
- In Java, we can achieve something similar through *Generics*; C++ through *templates*
 - Classes are defined with placeholders (see later lectures)
 - We fill them in when we create objects using them

```
LinkedList<Integer> = new LinkedList<Integer>()
LinkedList<Double> = new LinkedList<Double>()
```

Initially, you will most likely encounter Generics when using Java's built-in data structures such as `LinkedList`, `ArrayList`, `Map`, etc. For example, say you wanted a linked list of integers or `Vector3D` objects. You would declare:

```
LinkedList<Integer> lli = new LinkedList<Integer>();
LinkedList<Vector3D> llv = new LinkedList<Vector3D>();
```

This was shoe-horned into Java relatively recently, so if you are looking at old code on the web or old books, you might see them using the non-Generics versions that ignore the type e.g. `LinkedList ll = new LinkedList()` allows you to throw almost anything into it (including a mix of types—a source of many bugs!).

The astute amongst you may have noted that I used `LinkedList<Integer>` and not `LinkedList<int>`—it turns out that, in order to keep old code working, we simply can't use primitive types directly in Generics classes. This is a java-specific irritation and we will be looking at why later on in the course. For now, just be aware that every primitive has associated with it an (immutable) *class* that wraps around a variable of that type. For example, `int` has `Integer`, `double` has `Double`, etc.

2.4.1 Creating Parameterised Types

Creating Parameterised Types

- These just require a placeholder type

```
class Vector3D<T> {
    private T x;
    private T y;

    T getX() {return x;}
    T getY() {return y;}

    void setX(T nx) {x=nx;}
    void setY(T ny) {y=ny;}
}
```

We already saw how to use Generics types in Java (e.g. `LinkedList<Integer>`). Declaring them is not much harder than a 'normal' class. The `T` is just a placeholder (and I could have used any letter or word—`T` is just the de-facto choice). Once declared we can create `Vector3D` objects with different underlying storage types, just like with `LinkedList`:

```
Vector3D<Integer> vi = new Vector3D<Integer>(); // Vector3D<Integer>
Vector3D<Float> vi = new Vector3D<Float>(); // Vector3D<Float>
Vector3D<Double> vi = new Vector3D<Double>(); // Vector3D<Double>
```

There is no problem having parameterised types as parameters—for example `LinkedList<Vector3D<Integer>>` declares a list of integer vector objects. And we can have multiple parameters in our definitions:

```
public class Pair<U,V> {
    private U mFirst;
    private V mSecond;
    ...
}
```

You see this most commonly with `Maps` in Java, which represent dictionaries, mapping keys of some type to values of (potentially) some other type. e.g. a `TreeMap<String,Integer>` could be used to map names to ages).

C++

If you've used C++ you might be familiar with template classes. These share a similar syntax with Java Generics and, for now, it won't hurt to think of them as equivalent. However, they

are very different under the hood as we will see.

Lecture 3

Pointers, References and Memory

Imperative languages manipulate state held in system memory. They more naturally extend from assembly and before we go any further we need a mental model of how the compiler uses all this memory.

3.1 Pointers and References

Memory and Pointers

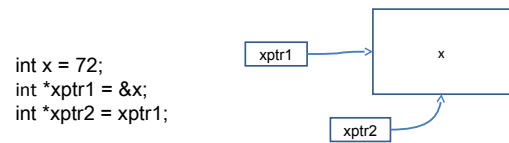
- In reality the compiler stores a mapping from variable name to a specific memory address, along with the type so it knows how to interpret the memory (e.g. “*x is an int so it spans 4 bytes starting at memory address 43526*”).
- Lower level languages often let us work with memory addresses directly. Variables that store memory addresses are called **pointers** or sometimes **references**
- Manipulating memory directly allows us to write fast, efficient code, but also exposes us to bigger risks
 - Get it wrong and the program 'crashes' .

The compiler must manipulate the computer’s memory, but the notion of type doesn’t exist at the lowest level. Memory is simply a vast pool of bits, grouped (usually) into bytes, and the compiler must manually specify the byte it wants to read or change using the memory address. This is little more than a number uniquely identifying that specific byte. So when you ask for an `int` to be created, the compiler knows to find a 4-byte chunk of memory that isn’t being used (assuming `ints` are 32 bits), mark it as used and set the bytes appropriately.

Some languages allow us, as programmers, to move beyond the abstraction of memory provided by explicit variable creation. They allow us to have variables that contain the actual memory addresses and even to manipulate them. We call such variables *pointers* and the traditional way to understand them is the “box and arrow” model:

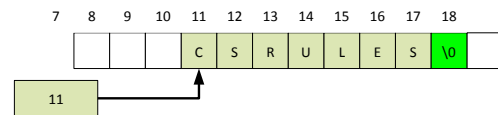
Pointers: Box and Arrow Model

- A pointer is just the memory address of the first memory slot used by the variable
- The pointer **type** tells the compiler how many slots the whole object uses



Example: Representing Strings I

- A single character is fine, but a text string is of variable length – how can we cope with that?
- We simply store the start of the string in memory and require it to finish with a special character (the NULL or terminating character, aka `\0`)
- So now we need to be able to store memory addresses → use **pointers**

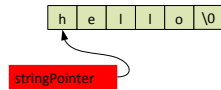


- We think of there being an **array** of characters (single letters) in memory, with the string pointer pointing to the first element of that array

Example: Representing Strings II

```
char letterArray[] = {'h','e','l','l','o','\0'};
char *stringPointer = &(letterArray[0]);
printf("%s\n",stringPointer);

letterArray[3]='0';
printf("%s\n",stringPointer);
```



Pointers are simply variables where the value is a memory address. We can arbitrarily modify them either accidentally or intentionally and this can lead to all sorts of problems. Although the symptom is usually the same: a hard program crash.

References

- A reference is an **alias** for another thing (object/array/etc)
- When you use it, you are 'redirected' somehow to the underlying thing
- Properties:
 - Either assigned or unassigned
 - If assigned, it is valid
 - You can easily check if assigned

References are *aliases* for other objects—i.e. they redirect to the 'real' object. You have of course met them in the last lecture on ML.

Implementing References

- A sane reference implementation in an imperative language is going to use pointers
- So each reference is the same as a pointer except that the compiler restricts operations that would violate the properties of references
- For this course, thinking of a reference as a restricted pointer is fine

Any sane implementation of a reference is likely to use pointers (and in FoCS they were directly equated). However, the concept of a reference forbids you from doing all of the things you can do with a pointer:

Distinguishing References and Pointers

	Pointers	References
Can be unassigned (null)	Yes	Yes
Can be assigned to established object	Yes	Yes
Can be assigned to an arbitrary chunk of memory	Yes	No
Can be tested for validity	No	Yes
Can perform arithmetic	Yes	No

The ability to test for validity is particularly important. A pointer points to something valid, something invalid, or *null* (a special zero-pointer that indicates it's not initialised). References, however, either point to something valid or to *null*. With a non-null reference, you know it's valid. With a non-null pointer, who knows? So references are going to be safer than pointers.

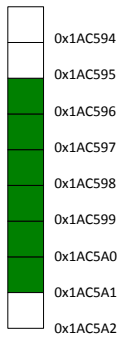
For those with experience with pointers, you might have found pointer arithmetic rather useful at times (e.g. incrementing a pointer to move one place forward in an array, etc). You can't do that with a reference since it would lead to you being able to reference arbitrary memory.

Languages and References

- Pointers are useful but dangerous
- C, C++: pointers *and* references
- Java: references *only*
- ML: references *only*

Arrays

```
byte[] arraydemo1 = new byte[6];
byte arraydemo2[] = new byte[6];
```



References in Java

- Declaring unassigned

```
SomeClass ref = null; // explicit
```

```
SomeClass ref2; // implicit
```

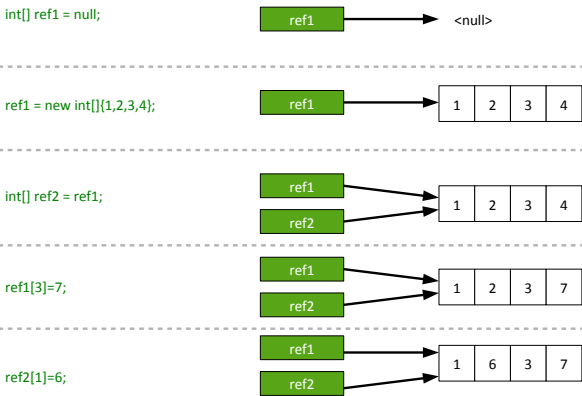
- Defining/assigning

```
// Assign
SomeClass ref = new ClassRef();
```

```
// Reassign to alias something else
ref = new ClassRef();
```

```
// Reference the same thing as another reference
SomeClass ref2 = ref;
```

References Example (Java)



Sun decided that Java would have *only* references and no explicit pointers. Whilst slightly limiting, this makes programming much safer (and it's one of the many reasons we teach with Java). Java has two classes of types: *primitive* and *reference*. A primitive type is a built-in type. **Everything** else is a reference type, including arrays and objects.

3.2 Keeping Track of Function Calls: The Call Stack

To this point we've been a bit woolly about what happens when we run ("call") functions. We have used nebulous terms like "the stack" or maybe "the call stack".

Keeping Track of Function Calls

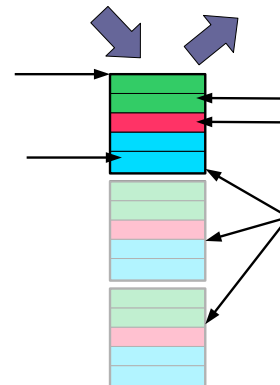
- We need a way of keeping track of which functions are currently running

```
public void a() {
    //...
}

public void b() {
    a();
}
```

When we call `b()`, the system must run `a()` while remembering that we return to `b()` afterwards. When a function is called from another, this is called *nesting*¹ (just as loops-within-loops are considered *nested*). The nesting can go arbitrarily deep (well, OK, until we run out of memory to keep track). The data structure widely used to keep track is the *call stack*

The Call Stack



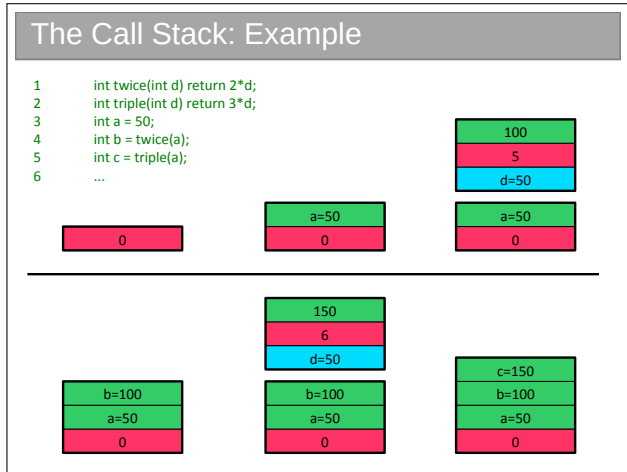
Remember the way the fetch-execute cycle handles procedure calls²: whenever a procedure is called we jump to the machine code for the procedure, execute it, and then jump back to where it was before and continue on. This means that, before it jumps to the

¹If the function is calling itself then it is of course *recursive*

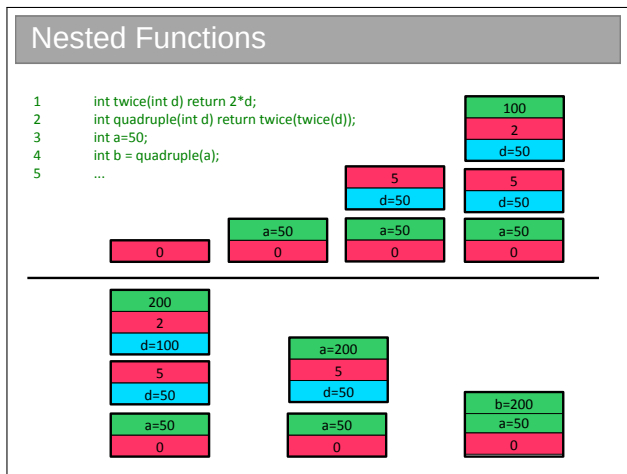
²Review the pre-arrival course if not

procedure code, it must save where it is.

We do this using a *call stack*. A stack is a simple data structure that is the digital analogue of a stack of plates: you add and take from the top of the pile *only*³. We say that we *push* new entries onto the stack and *pop* entries from its top. Here the ‘plates’ are called *stack frames* and they contain the function parameters, any local variables the function creates and, crucially, a return address that tells the CPU where to jump to when the function is done. When we finish a procedure, we delete the associated stack frame and continue executing from the return address it saved.

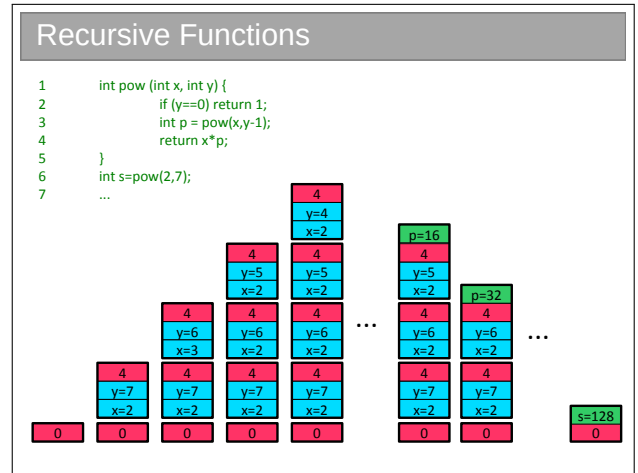


In this example I’ve avoided going down to assembly code and just assumed that the return address can be the code line number. This causes a small problem with e.g. line 4, which would be a couple of machine instructions (one to get the value of `twice()` and one to store it in `b`). I’ve just assumed the computer magically remembers to store the return value for brevity. This is all very simple and the stack never gets very big—things are more interesting if we start nesting functions:

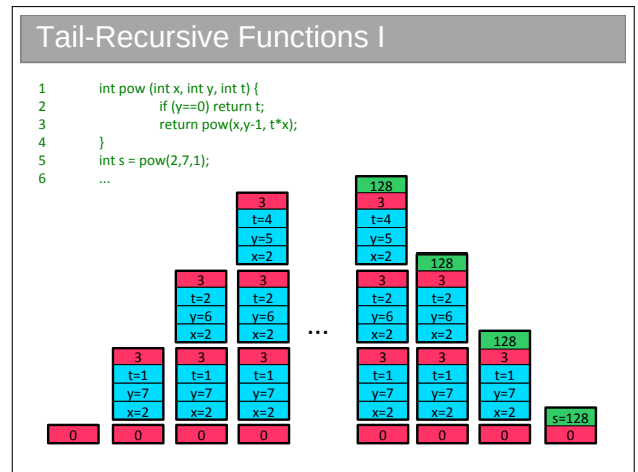


³See Algorithms next term for a full analysis

And even more interesting if we add nesting/recursion into the mix:



We immediately see a problem: computers only have finite memory so if our recursion is really deep, we’ll be throwing lots of stack frames into memory and, sooner or later, we will run out of memory. We call this *stack overflow* and it is an unrecoverable error that you’re almost certainly familiar with from ML. You know that tail-recursion does better, but:



If you’re in the habit of saying tail-recursive functions are better, be careful—they’re only better if the compiler/interpreter knows that it can optimise them to use $O(1)$ space. Java compilers don’t...⁴

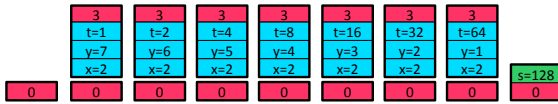
⁴Language designers usually speak of ‘tail-call optimisation’ since there is actually nothing special about recursion in this case: functions that call other functions may be written to use only tail calls, allowing the same optimisations.

Tail-Recursive Functions II

```

1  int pow (int x, int y, int t) {
2      if (y==0) return t;
3      return pow(x,y-1, t*x);
4  }
5  int s = pow(2,7,1);
6  ...

```



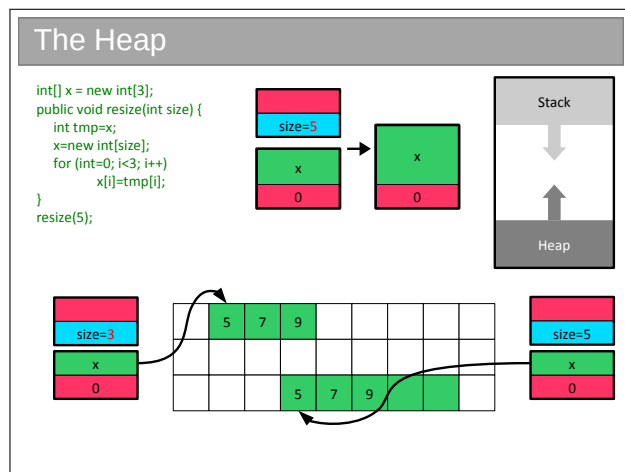
located in the stack. Pointers are of known size so won't ever increase. If we want to resize our array, we create a new, bigger array, copy the contents across and update the pointer within the stack. In the slide above the array exists on the stack and ends up being replaced with a new, bigger stack. The reference to it (x) is updated to point to the new one. Note: there's some missing Java code in the slide, since the contents of the initial array would need to be manually copied across to the new one.

For those who do the Paper 2 O/S course, you will find that the heap gets *fragmented*: as we create and delete stuff we leave holes in memory. Occasionally we have to spend time 'compacting' the holes (i.e. shifting all the stuff on the heap so that it's used more efficiently).

3.3 The Heap

There's a subtlety with the stack that we've passed over until now. What if we want a function to create something that sticks around after the it finishes? Or to resize something (say an array)? We talk of memory being *dynamically* allocated rather than *statically* allocated as per the stack.

Why can't we dynamically allocate on the stack? Well, imagine that we do everything on a stack and you have a function that resizes an array. We'd have to grow the stack, but not from the top, but where the stack was put. This rather invalidates our stack and means that every memory address we have will need to be updated if it comes after the array.



We avoid this by using a *heap*⁵. Quite simply we allocate the memory we need from some large pool of free memory, and store a pointer to the chunk we al-

⁵Note: you meet something called a 'heap' in Algorithms: it is NOT the same thing

3.4 Pass-by-value and Pass-by-reference

Argument Passing

- **Pass-by-value.** Copy the object into a new value in the stack

```

void test(int x) {...}
int y=3;
test(y);

```



- **Pass-by-reference.** Create a reference to the object and pass that.

```

void test(int &x) {...}
int y=3;
test(y);

```



Note I had to use C here since Java doesn't have a pass-by-reference operator such as &.

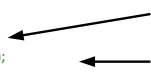
Pass-by-value. The value of the argument is copied into a new argument variable (this is what we assumed in the call stack earlier)

Pass-by-reference. Instead of copying the object (be it primitive or otherwise), we pass a reference to it. Thus the function can access the original and (potentially) change it.

When arguments are passed to java functions, you may hear it said that primitive values are "passed by value" and arrays are "passed by reference". I think this is misleading (and technically wrong).

Passing Procedure Arguments In Java

```
class Reference {  
  
    public static void update(int i, int[] array) {  
        i++;  
        array[0]++;  
    }  
  
    public static void main(String[] args) {  
        int test_i = 1;  
        int[] test_array = {1};  
        update(test_i, test_array);  
        System.out.println(test_i);  
        System.out.println(test_array[0]);  
    }  
}
```



This example is taken from your practicals, where you observed the different behaviour of `test_i` and `test_array`—the former being a primitive `int` and the latter being a reference to an array.

Let's create a model for what happens when we pass a primitive in Java, say an `int` like `test_i`. A new stack frame is created and the value of `test_i` is *copied* into the stack frame. You can do whatever you like to this copy: at the end of the function it is deleted along with the stack frame. The original is untouched.

Now let's look at what happens to the `test_array` variable. This is a *reference* to an array in memory. When passed as an argument, a new stack frame is created. The *value* of `test_array` (which is just a memory address) is copied into a *new* reference in the stack frame. So, we have two references pointing at the same thing. Making modifications through either changes the original array.

So we can see that Java *actually passes all arguments by value*, it's just that arguments are either primitives or references. i.e. Java is strictly pass-by-value⁶.

The confusion over this comes from the fact that many people view `test_array` to *be* the array and not a reference to it. If you think like that, then Java passes it by reference, as some texts (incorrectly) claim. The examples sheet has a question that explores this further.

C++

Passing Procedure Arguments In C++

```
void update(int i, int &iref){  
    i++;  
    iref++;  
}  
  
int main(int argc, char** argv) {  
    int a=1;  
    int b=1;  
    update(a,b);  
    printf("%d %d\n",a,b);  
}
```

Things are a bit clearer in other languages, such as C++. They may allow you to specify how something is passed. In this C++ example, putting an ampersand ('&') in front of the argument tells the compiler to pass by reference and not by value.

Having the ability to choose how you pass variables can be very powerful, but also problematic. Look at this code:

```
bool testA(HugeInt h) {  
    if (h > 1000) return TRUE;  
    else return FALSE;  
}  
  
bool testB(HugeInt &h) {  
    if (h > 1000) return TRUE;  
    else return FALSE;  
}
```

Here I have made a fictional type `HugeInt` which is meant to represent something that takes a lot of space in memory. Calling either of these functions will give the same answer, but what happens at a low level is quite different. In the first, the variable is copied (lots of memory copying required—bad) and then destroyed (ditto). In the second, only a reference is created and destroyed, and that's quick and easy.

So, even though both pieces of code work fine, if you miss that you should pass by reference (just one tiny ampersand's difference) you incur a large overhead and slow your program.

I see this sort of mistake a *lot* in C++ program-

⁶If your supervisor frowns at this, point them to the Java specification, section 8.4.1.

ming and I guess the Java designers did too—they stripped out the ability to specify pass by reference or value from Java!

Lecture 4

Inheritance

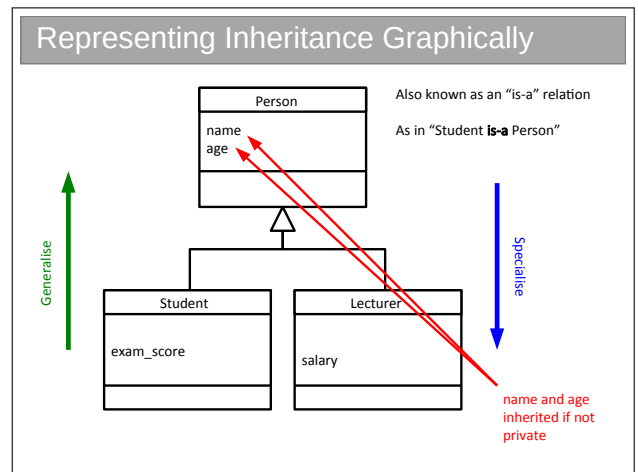
Inheritance I

```
class Student {
    public int age;
    public String name;
    public int grade;
}

class Lecturer {
    public int age;
    public String name;
    public int salary;
}
```

- There is a lot of duplication here
- Conceptually there is a hierarchy that we're not really representing
- Both Lecturers and Students are people (no, really).
- We can view each as a kind of specialisation of a general person
 - They have all the properties of a person
 - But they also have some extra stuff specific to them

(I should not have used public variables here, but I did it to keep things simple)



Inheritance II

```
class Person {
    public int age;
    public String name;
}

class Student extends Person {
    public int grade;
}

class Lecturer extends Person {
    public int salary;
}
```

- We create a *base class* (Person) and add a new notion: classes can *inherit* properties from it
 - Both state and functionality
- We say:
 - Person is the *superclass* of Lecturer and Student
 - Lecturer and Student *subclass* Person

Java uses the keyword `extends` to indicate inheritance of classes.

C++

In C++ it's a more opaque colon:

```
class Parent {...};
class Student : public Parent {...};
class Lecturer : public Parent {...};
```

Inheritance is an extremely powerful concept that is used extensively in good OOP. We discussed the “has-a” relation amongst classes; inheritance adds an “is-a” concept. E.g. A car *is a* vehicle that *has a* steering wheel.

We speak of an inheritance *tree* where moving down the tree makes things more specific and up the tree more general. Unfortunately, we tend to use an array of different names for things in an inheritance tree. For B extends A, you might hear any of:

- A is the superclass of B
- A is the parent of B
- A is the base class of B
- B is the child of A
- B derives from A
- B extends A
- B inherits from A
- B subclasses A

Many students confuse “is-a” and “has-a” arrows in their UML class diagrams: please make sure you don't! Inheritance has an empty triangle for the arrowhead, whilst association has two ‘wings’.

4.1 Casting

Casting

- Many languages support *type casting* between numeric types

```
int i = 7;
float f = (float) i; // f==7.0
double d = 3.2;
int i2 = (int) d; // i2==3
```

- With inheritance it is reasonable to type cast an object to any of the types above it in the inheritance tree...

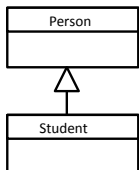
the chunk if we cast to a parent class (plus some extra stuff).

If we try to cast to a child class, there won't be all the necessary info in the memory so it will fail. *But* beware—you don't get a compiler error in the failed example above! The compiler is fine with the cast and instead the program chokes when we try to *run* that piece of code—a *runtime* error.

Note the example of casting primitive numeric types in the slide is a bit different, since a new variable of the primitive type is created and assigned the relevant value.

4.2 Shadowing

Widening



```
Student s = new Student();
Person p = (Person) s;
```

"Casting"

- Student is-a Person
- Hence we can use a Student object anywhere we want a Person object
- Can perform *widening* conversions (up the tree)

```
public void print(Person p) {...}
Student s = new Student();
print(s);
```

Implicit cast

Fields and Inheritance

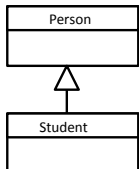
```
class Person {
  public String mName;
  protected int mAge;
  private double mHeight;
}
class Student extends Person {
  public void do_something() {
    mName="Bob";
    mAge=70;
    mHeight=1.70;
  }
}
```

Student inherits this as a public variable and so can access it

Student inherits this as a protected variable and so can access it

Student inherits this but as a **private** variable and so cannot access it directly

Narrowing



```
Person p = new Person();
Student s = (Student) p;
```

FAILS. Not enough info
In the real object to represent
a Student

- Narrowing conversions move down the tree (more specific)
- Need to take care...

```
Student s = new Student();
Person p = (Person) s;
Students s2 = (Student) p;
```

OK because underlying object
really is a Student

You will see that the **protected** access modifier can now be explained. A **protected** variable is exposed for read and write within a class, and *within all subclasses of that class*. Code outside the class or its subclasses can't touch it directly¹.

¹At least, that's how it is in most languages. Java actually allows any class in the same Java package to access protected variables as discussed previously.

When we create an object, a specific chunk of memory is allocated with all the necessary info and a reference to it returned (in Java). Casting just creates a new reference with a different type and points it to the same memory chunk. Everything we need will be in

Fields and Inheritance: Shadowing

```
class A { public int x; }

class B extends A {
  public int x;
}

class C extends B {
  public int x;

  public void action() {
    // Ways to set the x in C
    x = 10;
    this.x = 10;

    // Ways to set the x in B
    super.x = 10;
    ((B)this).x = 10;

    // Ways to set the x in A
    ((A)this).x = 10;
  }
}
```

What happens here?? There is an inheritance tree (A is the parent of B is the parent of C). Each of these declares an integer field with the name x. In memory, you will find three allocated integers for every object of type C. We say that variables in parent classes with the same name as those in child classes are *shadowed*.

Note that the variables are genuinely being shadowed and nothing is being replaced. This is in contrast to the behaviour with methods...

NB: A common novice error is to assume that we have to redeclare a field in its subclasses for it to be inherited: not so. *Every* non-private field is inherited by a subclass.

There are two new keywords that have appeared here: `super` and `this`. The `this` keyword can be used in any class method² and provides us with a reference to the current object. In fact, the `this` keyword is what you need to access anything within a class, but because we'd end up writing `this` all over the place, it is taken as implicit. So, for example:

```
public class A {
  private int x;
  public void go() {
    this.x=20;
  }
}
```

becomes:

```
public class A {
  private int x;
  public void go() {
    x=20;
  }
}
```

²By this I mean it cannot be used outside of a class, such as within a `static` method: see later for an explanation of these.

```
}
```

The `super` keyword gives us access to the direct parent (one step up in the tree). You've met both keywords in your Java practicals.

4.3 Overloading

We have already discussed function overloading, where we had multiple functions with the same name, but a different prototype (i.e. set of arguments). The same is possible within classes.

4.4 Overriding

The remaining question is what happens to methods when they are inherited and rewritten in the child class. The obvious possibility is that they are treated the same as fields, and shadowed. When this occurs we say that the method is *overridden*. As it happens, we can't do this in Java, but it is the default in C++ so we can use that to demonstrate:

Methods and Inheritance: Overriding

- We might want to require that every `Person` can dance. But the way a `Lecturer` dances is not likely to be the same as the way a `Student` dances...

```
class Person {
  public void dance() {
    jiggle_a_bit();
  }
}

class Student extends Person {
  public void dance() {
    body_pop();
  }
}

class Lecturer extends Person {
}
```

Person defines a 'default' implementation of dance()

Student overrides the default

Lecturer just inherits the default implementation and jiggles

Every object that has `Person` for a parent must have a `dance()` method since it is defined in the `Person` class and is inherited. If we override it in `Child` then `Child` objects will behave differently. There are some subtleties to this that we'll return to next lecture.

A useful habit to get into is to annotate every function you override using `@Override`. This serves two purposes: firstly it tells anyone reading the code that it's an overridden method; secondly it allows the compiler to check it really does override something. It's surprisingly easy to make a typo and think you've overridden but actually not. We'll see this later when we discuss

object comparison.

4.5 Abstract Methods and Classes

Abstract Methods

- Sometimes we want to force a class to implement a method but there isn't a convenient default behaviour
- An **abstract** method is used in a base class to do this
- It has no implementation whatsoever

```
class abstract Person {
    public abstract void dance();
}

class Student extends Person {
    public void dance() {
        body_pop();
    }
}

class Lecturer extends Person {
    public void dance() {
        jiggle_a_bit();
    }
}
```

An abstract method can be thought of as a contractual obligation: any non-abstract class that inherits from this class *will* have that method implemented.

Abstract Classes

- Note that I had to declare the class abstract too. This is because it has a method without an implementation so we can't directly instantiate a Person.

```
public abstract class Person {
    public abstract void dance();
}                                     Java

class Person {
    public:
    virtual void dance()=0;
}                                     C++
```

- All state and non-abstract methods are inherited as normal by children of our abstract class
- Interestingly, Java allows a class to be declared abstract even if it contains no abstract methods!

Abstract classes allow us to partially define a type. Because it's not fully defined, you can't make an object from an abstract class (try it). Only once *all* of the 'blanks' have been filled in can we create an object from it. This is particularly useful when we want to represent high level concepts that do not exist in isolation.

Depending on who you're talking to, you'll find different terminology for the initial declaration of the abstract function (e.g. the `public abstract void dance()` bit). Common terms include *method prototype* and *method stub*.

C++

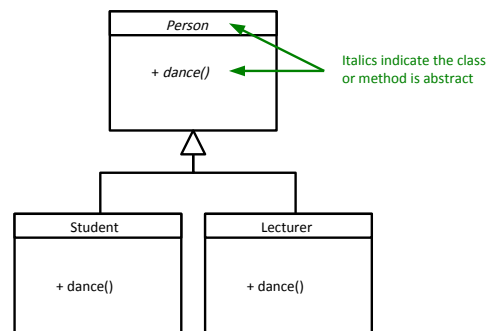
In C++, the syntax for abstract is a bit different. Firstly you define the method as `virtual` to indicate you want dynamic polymorphism (see next lecture) and you use `=0` to indicate you aren't filling it in here:

```
class Person {
    public:
        virtual void dance()=0;
}
```

Python

Abstract methods/base classes in python were only added fairly recently. And they look very much like the afterthought they are.

Representing Abstract Classes



You have to look at UML diagrams carefully since the italics that represent abstract methods or classes aren't always obvious on a quick glance.

Lecture 5

Polymorphism

You should be comfortable with the polymorphism¹ that you met in FoCS, where you wrote functions that could operate on multiple types. It turns out that is just one type of polymorphism in programming, and it isn't the form that most programmers mean when they use the word. To understand that, we should look back at our overridden methods:

Polymorphic Methods

```
Student s = new Student();
Person p = (Person)s;
p.dance();
```

- Assuming Person has a default dance() method, what should happen here??
- General problem: when we refer to an object via a parent type and both types implement a particular method: which method should it run?

Polymorphic Concepts I

- Static** polymorphism
 - Decide at compile-time
 - Since we don't know what the true type of the object will be, we just run the parent method
 - Type errors give compile errors

```
Student s = new Student();
Person p = (Person)s;
p.dance();
```

- Compiler says "p is of type Person"
- So p.dance() should do the default dance() action in Person

If we can get different method implementations by casting the same object to different types, we have

¹The etymology of the word polymorphism is from the ancient Greek: *poly* (many)–*morph* (form)–ism

static polymorphism. In general static polymorphism refers to anything where decisions are made at compile-time (so-called "early binding"). You may realise that all the polymorphism you saw in ML was static polymorphism. The shadowing of fields also fits this description.

Polymorphic Concepts II

- Dynamic** polymorphism
 - Run the method in the child
 - Must be done at run-time since that's when we know the child's type
 - Type errors cause run-time faults (crashes!)

```
Student s = new Student();
Person p = (Person)s;
p.dance();
```

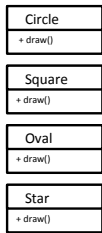
- Compiler looks in memory and finds that the object is really a Student
- So p.dance() runs the dance() action in Student

Here we get the same method implementation regardless of what we cast the object to. In order to be sure that it gets this right, we can't figure out which method to run when we are compiling. Instead, the system has to run the program and, when a decision needs to be made about which method to run, it must look at the actual object in memory (regardless of the type of the reference, which may be a cast) and act appropriately.

This form of polymorphism is OOP-specific and is sometimes called *sub-type* or *ad-hoc* polymorphism. It's crucial to good, clean OOP code. Because it must check types at run-time (so-called "late binding") there is a performance overhead associated with dynamic polymorphism. However, as we'll see, it gives us much more flexibility and can make our code more legible.

Beware: Most programmers use the word 'polymorphism' to refer to dynamic polymorphism.

The Canonical Example I

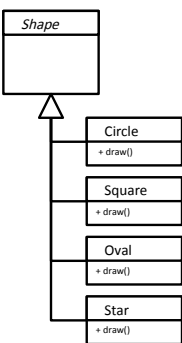


- A drawing program that can draw circles, squares, ovals and stars
- It would presumably keep a list of all the drawing objects
- **Option 1**
 - Keep a list of Circle objects, a list of Square objects,...
 - Iterate over each list drawing each object in turn
 - What has to change if we want to add a new shape?

Implementations

- Java
 - All methods are dynamic polymorphic.
- Python
 - All methods are dynamic polymorphic.
- C++
 - Only functions marked *virtual* are dynamic polymorphic
- Polymorphism in OOP is an extremely important concept that you need to make sure you understand...

The Canonical Example II



- **Option 2**
 - Keep a single list of Shape references
 - Figure out what each object really is, narrow the reference and then draw()
- ```

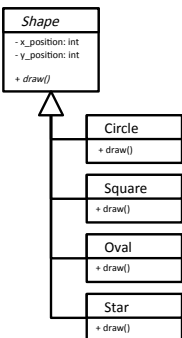
for every Shape s in myShapeList
 if (s is really a Circle)
 Circle c = (Circle);
 c.draw();
 else if (s is really a Square)
 Square sq = (Square);
 sq.draw();
 else if...

```
- What if we want to add a new shape?

C++ allows you to choose whether methods are inherited statically (default) or dynamically (explicitly labelled with the keyword *virtual*). This can be good for performance (you only incur the dynamic overhead when you need to) but gets complicated, especially if the base method isn't dynamic but a derived method is...

The Java designers avoided the problem by enforcing dynamic polymorphism. You may find reference to final methods being Java's static polymorphism since this gives a compile error if you try to override it in subclasses. But really, this isn't the same: the compiler isn't choosing between multiple implementations but rather enforcing that there can only be one implementation.

## The Canonical Example III



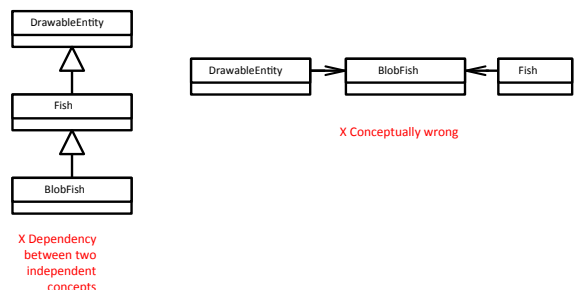
- **Option 3 (Polymorphic)**
    - Keep a single list of Shape references
    - Let the compiler figure out what to do with each Shape reference
- ```

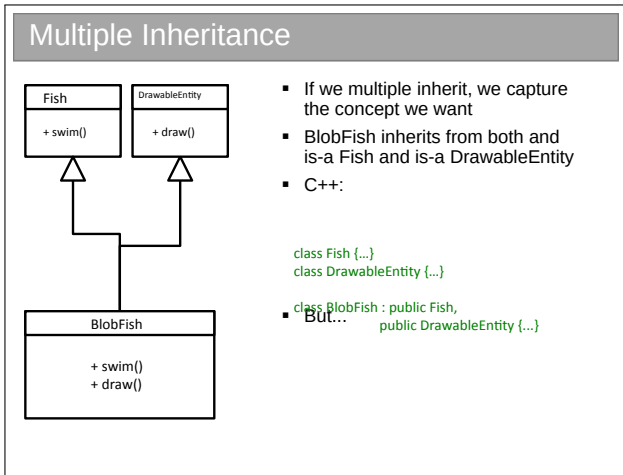
For every Shape s in myShapeList
  s.draw();
  
```
- What if we want to add a new shape?

5.1 Multiple Inheritance and Interfaces

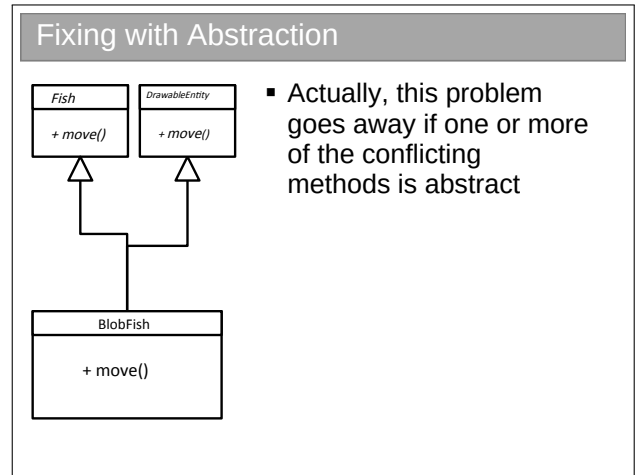
Harder Problems

- Given a class Fish and a class DrawableEntity, how do we make a BlobFish class that is a drawable fish?

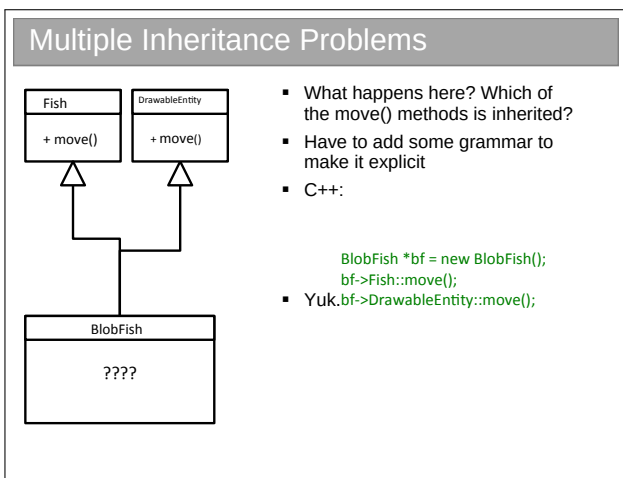




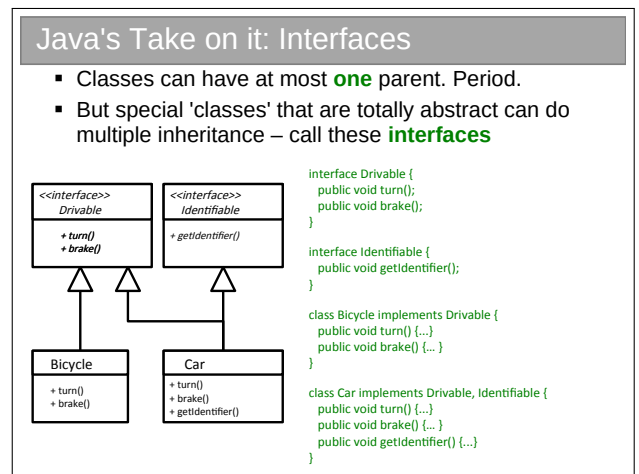
This is the obvious and (perhaps) sensible option that manages to capture the concept nicely.



The problem goes away here because the methods are abstract and hence have no implementation that can conflict.



Many texts speak of the “dreaded diamond”. This occurs when a base class has two children who are the parents of another class through multiple inheritance (thereby forming a diamond in the UML diagram). If the two classes in the middle independently override a method from the top class, the bottom class suffers from the problem in this slide.



So Java allows you to inherit from one class *only* (which may itself inherit from one other, which may itself...). Many programmers coming from C++ find this limiting, but it just means you have to think of another way to represent your classes (often a better way, although not always!).

A Java *interface* is essentially just a class that has:

- No state whatsoever; and
- All methods abstract.

This is a greatly simplified concept that allows for multiple inheritance without any chance of conflict. Interfaces are represented in our UML class diagram with a preceding <<interface>> label and inheritance occurs via the implements keyword rather than through extends.

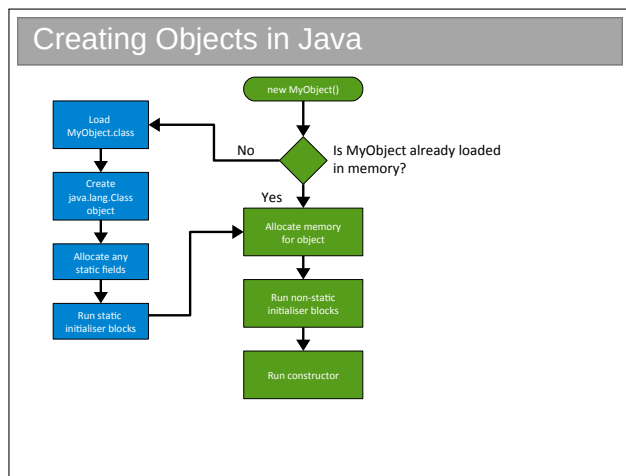
Interfaces are so important in Java they are considered

to be the third reference type (the other two being classes and arrays). Using interfaces encourages a high level of abstraction in code, which is generally a good thing since it makes the code more flexible/portable. However, it is possible to overdo it, ending up with 20 files where just one would do...

Lecture 6

Lifecycle of an Object

We met constructors earlier in the course as methods that initialise objects. We can now add a bit more detail. When you request a new object, Java will do quite a lot of work:



Note that Java maintains a `java.lang.Class` object for every class it loads into memory from a `.class` file. This object actually allows you query things about the class, such as its name or to list all the methods it has. The ability to do inspect (and possibly modify!) a program's structure is a feature called *reflection*. It's quite a powerful feature that exists in some (but certainly not all) languages. It's out of scope here but worth exploring if you're interested.

Initialisation Example

```
public class Blah {
    private int mX = 7;
    public static int sX = 9;

    {
        mX=5;
    }

    static {
        sX=3;
    }

    public Blah() {
        mX=1;
        sX=9;
    }
}
```

`Blah b = new Blah();`
`Blah b2 = new Blah();`

1. Blah loaded
2. sX created
3. sX set to 9
4. sX set to 3
5. Blah object allocated
6. mX set to 7
7. mX set to 5
8. Constructor runs (mX=1, sX=9)
9. b set to point to object
10. Blah object allocated
11. mX set to 7
12. mX set to 5
13. Constructor runs (mX=1, sX=9)
14. b2 set to point to object

Things get even more complex when we throw in some inheritance:

Constructor Chaining

- When you construct an object of a type with parent classes, we call the constructors of all of the parents in sequence

```
Student s = new Student();
```

```
graph BT
    Animal --> Person
    Person --> Student
```

1. Call Animal()
2. Call Person()
3. Call Student()

In reality, Java asserts that the first line of a constructor *always* starts with `super()`, which is a call to the parent constructor (which itself starts with `super()`, etc.). If it does not, the compiler adds one for you:

```
public class Person {
    public Person() {
```

```

}
}

```

becomes:

```

public class Person {
    public Person() {
        super();
    }
}

```

C++

In other languages that support multiple inheritance, this becomes more complex since there may be more than one parent and a simple keyword like `super` isn't enough. Instead they support manually specifying the constructor parameters for the parents. E.g. for C++:

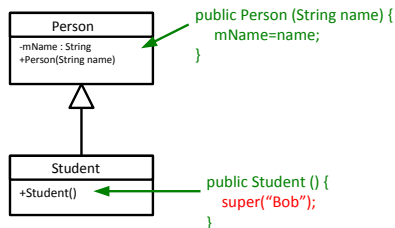
```

class Child : public Parent1, Parent2 {
public:
    Child() : Parent1("Alice"),
            Parent2("Bob") {...}
}

```

Chaining without Default Constructors

- What if your classes have explicit constructors that take arguments? You need to explicitly chain
- Use `super` in Java:



Deterministic Destruction

- Objects are created, used and (eventually) destroyed. Destruction is very language-specific
- Deterministic destruction is what you would expect
 - Objects are deleted at predictable times
 - Perhaps manually deleted (C++):

```

void UseRawPointer()
{
    MyClass *mc = new MyClass();
    // ...use mc...
    delete mc;
}

```

- Or auto-deleted when out of scope (C++):

```

void UseSmartPointer()
{
    unique_ptr<MyClass> *mc = new MyClass();
    // ...use mc...
} // mc deleted here

```

Destructors

- Most OO languages have a notion of a destructor too
 - Gets run when the object is destroyed
 - Allows us to release any resources (open files, etc) or memory that we might have created especially for the object

```

class FileReader {
public:
    // Constructor
    FileReader() {
        f = fopen("myfile", "r");
    }
    // Destructor
    ~FileReader() {
        fclose(f);
    }
private:
    FILE *file;
}

int main(int argc, char ** argv) {
    // Construct a FileReader Object
    FileReader *f = new FileReader();
    // Use object here
    ...
    // Destruct the object
    delete f;
}

```

It will shortly become apparent why I used C++ and not Java for this example.

Non-Deterministic Destruction

- Deterministic destruction is easy to understand and seems simple enough. But it turns out we humans are rubbish at keeping track of what needs deleting when
- We either forget to delete (→ memory leak) or we delete multiple times (→ crash)
- We can instead leave it to the system to figure out when to delete
 - "Garbage Collection"
 - The system somehow figures out when to delete and does it for us
 - In reality it needs to be cautious and sure it can delete. This leads to us not being able to predict exactly when something will be deleted!!
- This is the Java approach!!

What about Destructors?

- Conventional destructors don't make sense in non-deterministic systems
 - When will they run?
 - Will they run at all??
- Instead we have **finalisers**: same concept but they only run when the system deletes the object (which may be never!)

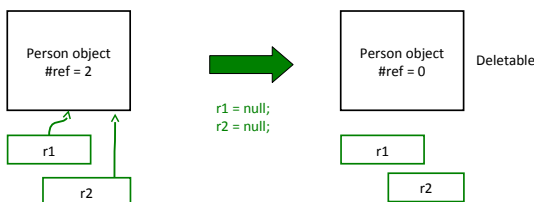
OK, so a finalizer is just a rebadged destructor, but the rebadging is important. It reminds us as programmers that it won't run deterministically. Because you can't tell when finalizer methods will get called in Java, their value is greatly reduced. It's actually quite rare to see them in Java in my experience.

Garbage Collection

- So how exactly does garbage collection work? How can a system know that something can be deleted?
- The garbage collector is a separate process that is constantly monitoring your program, looking for things to delete
- Running the garbage collector is obviously not free. If your program creates a lot of short-term objects, you will soon notice the collector running
 - Can give noticeable pauses to your program!
 - But minimises memory leaks (it does not prevent them...)
- There are various algorithms: we'll look at two that can be found in Java
 - Reference counting
 - Tracing

Reference Counting

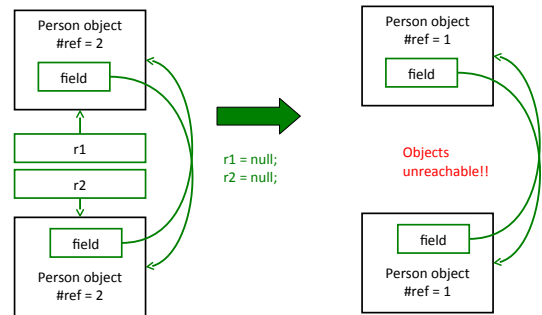
- Java's original GC. It keeps track of how many references point to a given object. If there are none, the programmer can't access that object ever again so it can be deleted



- every object needs more memory (to store the reference count) and we have to monitor changes to all references to keep the counts up to date.

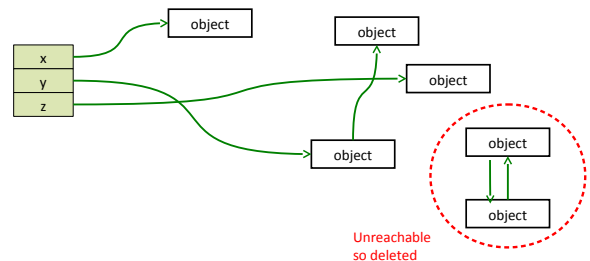
Reference Counting Gotcha

- Circular references are a pain



Tracing

- Start with a list of all references you can get to
- Follow all references recursively, marking each object
- Delete all objects that were not marked



Note that reference counting has an associated cost

Lecture 7

Java Collections and Object Comparison

Java Class Library

- Java the platform contains around 4,000 classes/interfaces
 - Data Structures
 - Networking, Files
 - Graphical User Interfaces
 - Security and Encryption
 - Image Processing
 - Multimedia authoring/playback
 - And more...
- All neatly(ish) arranged into packages (see API docs)

Remember Java is a *platform*, not just a programming language. It ships with a huge *class library*: that is to say that Java itself contains a big set of built-in classes for doing all sorts of useful things like:

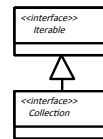
- Complex data structures and algorithms
- I/O (input/output: reading and writing files, etc)
- Networking
- Graphical interfaces

Of course, most programming languages have built-in classes, but Java has a big advantage. Because Java code runs on a virtual machine, the underlying platform is abstracted away. For C++, for example, the compiler ships with a fair few data structures, but things like I/O and graphical interfaces are completely different for each platform (Windows, OSX, Linux, whatever). This means you usually end up using lots of third-party libraries to get such extras—not so in Java.

There is, then, good reason to take a look at the Java class library to see how it is structured.

7.1 Collections and Generics

Java's Collections Framework



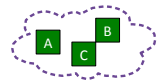
- Important chunk of the class library
- A collection is some sort of grouping of things (objects)
- Usually when we have some grouping we want to go through it (“*iterate* over it”)
- The Collections framework has two main interfaces: *Iterable* and *Collection*. They define a set of operations that all classes in the Collections framework support
- `add(Object o)`, `clear()`, `isEmpty()`, etc.

The Java Collections framework is a set of interfaces and classes that handles groupings of objects and allows us to implement various algorithms invisibly to the user (you’ll learn about the algorithms themselves next term).

Sets

<<interface>> Set

- A collection of elements with no duplicates that represents the mathematical notion of a set
- TreeSet: objects stored in order
- HashSet: objects in unpredictable order but fast to operate on (see Algorithms course)

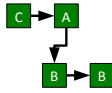


```
TreeSet<Integer> ts = new TreeSet<Integer>();
ts.add(15);
ts.add(12);
ts.contains(7); // false
ts.contains(12); // true
ts.first(); // 12 (sorted)
```

Lists

<<interface>> List

- An ordered collection of elements that may contain duplicates
- LinkedList: linked list of elements
- ArrayList: array of elements (efficient access)
- Vector: Legacy, as ArrayList but threadsafe



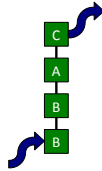
```

LinkedList<Double> ll = new LinkedList<Double>();
ll.add(1.0);
ll.add(0.5);
ll.add(3.7);
ll.add(0.5);
ll.get(1); // get element 2 (==3.7)
  
```

Queues

<<interface>> Queue

- An ordered collection of elements that may contain duplicates and supports removal of elements from the head of the queue
- offer() to add to the back and poll() to take from the front
- LinkedList: supports the necessary functionality
- PriorityQueue: adds a notion of priority to the queue so more important stuff bubbles to the top



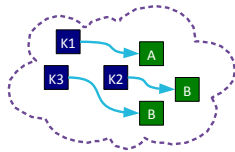
```

LinkedList<Double> ll = new LinkedList<Double>();
ll.offer(1.0);
ll.offer(0.5);
ll.poll(); // 1.0
ll.poll(); // 0.5
  
```

Maps

<<interface>> Map

- Like dictionaries in ML
- Maps key objects to value objects
- Keys must be unique
- Values can be duplicated and (sometimes) null.
- TreeMap: keys kept in order
- HashMap: Keys not in order, efficient (see Algorithms)



```

TreeMap<String, Integer> tm = new TreeMap<String, Integer>();
tm.put("A", 1);
tm.put("B", 2);
tm.get("A"); // returns 1
tm.get("C"); // returns null
tm.contains("G"); // false
  
```

There are other interfaces in the Collections class, and you may want to poke around in the API documentation. In day-to-day programming, however, these are likely to be the interfaces you use.

Now, don't worry about too much what's going on behind the scenes (that comes in the Algorithms course), just recognise that there are a series of implementa-

tions in the class library that you can use, and that each has different properties. You should get into the habit of reading the API descriptions to find best choice for your specific problem.

Iteration

▪ for loop

```

LinkedList<Integer> list = new LinkedList<Integer>();
...
for (int i=0; i<list.size(); i++) {
    Integer next = list.get(i);
}
  
```

▪ foreach loop (Java 5.0+)

```

LinkedList list = new LinkedList();
...
for (Integer i : list) {
    ...
}
  
```

The foreach notation works for arrays too and it's particularly neat when we have nested iteration. E.g. iteration over all students and their subjects:

```

for (Student stu : studentlist)
    for (Subject sub : subjectlist)
        getMarks(stu, sub);
  
```

versus:

```

for (int i=0; i<studentlist.size(); i++) {
    Student stu = studentlist.get(i);
    for (int j=0; j<subjectlist.size(); j++) {
        Subject sub = subjectlist.get(j);
        getMarks(stu, sub);
    }
}
  
```

Iterators

▪ What if our loop changes the structure?

```

for (int i=0; i<list.size(); i++) {
    If (i==3) list.remove(i);
}
  
```

▪ Java introduced the Iterator class

```

Iterator<Integer> it = list.iterator();
while(it.hasNext()) {Integer i = it.next();}
for (; it.hasNext(); ) {Integer i = it.next();}
  
```

▪ Safe to modify structure

```

while(it.hasNext()) {
    it.remove();
}
  
```


Note that the foreach structure isn't useful with Iterators. So we sacrifice some code readability for the ability to adjust the Collection's structure as we go.

7.2 Comparing Objects

Comparing Objects

- You often want to impose orderings on your data collections
- For TreeSet and TreeMap this is automatic

```
TreeMap<String, Person> tm = ...
```

- For other collections you may need to explicitly sort

```
LinkedList<Person> list = new LinkedList<Person>();
//...
Collections.sort(list);
```

- For numeric types, no problem, but how do you tell Java how to sort Person objects, or any other custom class?

Collections are great, but often you end up needing to impose orderings (i.e. sort). Examples include printing users by surname, or computing numerical metrics such as the median.

Comparing Primitives

- > Greater Than
- >= Greater than or equal to
- == Equal to
- != Not equal to
- < Less than
- <= Less than or equal to

- Clearly compare the value of a primitive
- But what does (ref1==ref2) do??
 - Test whether they point to the same object?
 - Test whether the objects they point to have the same state?

The problem is that we deal with references to objects, not objects. So when we compare two things, do we compare the references of the objects they point to? As it turns out, both can be useful so we want to support both.

7.2.1 Object Equality

Reference Equality

- $r1==r2$, $r1!=r2$
- These test *reference equality*
- i.e. do the two references point to the same chunk of memory?

```
Person p1 = new Person("Bob");
Person p2 = new Person("Bob");
```

```
(p1==p2); ← False (references differ)
```

```
(p1!=p2); ← True (references differ)
```

```
(p1==p1); ← True
```

Value Equality

- Use the equals() method in Object
- Default implementation just uses reference equality (==) so we have to override the method

```
public EqualsTest {
    public int x = 8;

    @Override
    public boolean equals(Object o) {
        EqualsTest e = (EqualsTest)o;
        return (this.x==e.x);
    }

    public static void main(String args[]) {
        EqualsTest t1 = new EqualsTest();
        EqualsTest t2 = new EqualsTest();
        System.out.println(t1==t2);
        System.out.println(t1.equals(t2));
    }
}
```

I find this mildly irritating: every class you use will support equals() but you'll have to check whether or not it has been overridden to do something other than ==. Personally, I try to limit my use of equals() to objects from core Java classes, where I trust it to have been done properly.

Aside: Use The Override Annotation

- It's so easy to mistakenly write:

```
public EqualsTest {
    public int x = 8;

    public boolean equals(EqualsTest e) {
        return (this.x==e.x);
    }

    public static void main(String args[]) {
        EqualsTest t1 = new EqualsTest();
        EqualsTest t2 = new EqualsTest();
        Object o1 = (Object) t1;
        Object o2 = (Object) t2;
        System.out.println(t1.equals(t2));
        System.out.println(o1.equals(o2));
    }
}
```

Aside: Use The Override Annotation II

- Annotation would have picked up the mistake:

```
public EqualsTest {
    public int x = 8;

    @Override
    public boolean equals(EqualsTest e) {
        return (this.x==e.x);
    }

    public static void main(String args[]) {
        EqualsTest t1 = new EqualsTest();
        EqualsTest t2 = new EqualsTest();
        Object o1 = (Object) t1;
        Object o2 = (Object) t2;
        System.out.println(t1.equals(t2));
        System.out.println(o1.equals(o2));
    }
}
```

What's happening here is that the signature of our overriding method doesn't match the one in `Object`. So, Java actually *overloads* it, keeping both methods. By using `@Override` when we mean to override not overload, the compiler will spot our error.

For the geeks out there (i.e. non-examinable), we could write a compiler that spots that `EqualsTest` is a subclass of `Object` and therefore do overriding. This is called *covariant parameter types* and is *not* supported by Java.

Java Quirk: hashCode()

- Object also gives classes `hashCode()`
- Code assumes that if `equals(a,b)` returns true, then `a.hashCode()` is the same as `b.hashCode()`
- So you should override `hashCode()` at the same time as `equals()`

I don't want to go into this in too much detail since you haven't yet met hashes (it's in the Algorithms course next term). For now, just accept that a hash is a function that takes in chunks of information (e.g. all the fields in an object) and spits out a number. Java uses this in its `HashMap` implementation and other places as a shortcut to having to sequentially compare each field. I mention it here really for completeness so that if any of you override `equals()` in production code then you know you should also override `hashCode()`. Details of doing so are easily found on the web and in books (because it's a very common mistake to make!).

7.3 Less Than and Greater Than

In order to sort your classes using the built in classes, you need to write something that allows two objects to be ordered. Often our classes have a *natural ordering* e.g. people are usually sorted first by surname and then by forename. We can build-in natural ordering to our classes using the `Comparable` interface:

Comparable<T> Interface I

```
int compareTo(T obj);
```

- Part of the Collections Framework
- Doesn't just tell us true or false, but smaller, same, or larger: useful for sorting.
- Returns an integer, `r`:
 - `r<0` This object is less than obj
 - `r==0` This object is equal to obj
 - `r>0` This object is greater than obj

Comparable<T> Interface II

```
public class Point implements Comparable<Point> {
    private final int mX;
    private final int mY;
    public Point (int, int y) { mX=x; mY=y; }

    // sort by y, then x
    public int compareTo(Point p) {
        if (mY>p.mY) return 1;
        else if (mY<p.mY) return -1;
        else {
            if (mX>p.mX) return 1;
            else if (mX<p.mX) return -1;
            else return 0.
        }
    }
}

// This will be sorted automatically by y, then x
Set<Point> list = new TreeSet<Point>();
```

This is all very well, but sometimes we might want to sort with a different ordering (e.g. sort just by forename). Java Collections lets us do this by supplying a custom piece of code for the ordering: a *Comparator*:

is at this stage.

Comparator<T> Interface I

```
int compare(T obj1, T obj2)
```

- Also part of the Collections framework and allows us to specify a specific ordering for a particular job
- E.g. a Person might have natural ordering that sorts by surname. A Comparator could be written to sort by age instead...

Comparator<T> Interface II

```
public class Person implements Comparable<Person> {
    private String mSurname;
    private int mAge;
    public int compareTo(Person p) {
        return mSurname.compareTo(p.mSurname);
    }
}

public class AgeComparator implements Comparator<Person> {
    public int compare(Person p1, Person p2) {
        return (p1.mAge-p2.mAge);
    }
}

...
ArrayList<Person> plist = ...;
...
Collections.sort(plist); // sorts by surname
Collections.sort(plist, new AgeComparator()); // sorts by age
```

Note that a natural ordering uses `compareTo()` whilst a comparator uses `compare()`.

7.4 Operator Overloading

Operator Overloading

- Some languages have a neat feature that allows you to overload the comparison operators. e.g. in C++

```
class Person {
public:
    int mAge;
    bool operator==(Person &p) {
        return (p.mAge==mAge);
    };
}

Person a, b;
b == a; // Test value equality
```

Java doesn't have this, but it's good to know what it

Lecture 8

Error Handling Revisited

As you have almost certainly discovered, errors crop up all over the place when developing software. We see various types:

Syntactic errors (missing brackets or whatever) are usually quite easy to spot because you get a nice explanatory compiler warning (ahem... unless you're using poly/ML...).

Logical errors (i.e. bugs) are more problematic, not least because comprehensive testing (checking the output for every possible input *and* system state) is usually infeasible for anything but toy programs.

External errors occur for processes our code relies on but we don't control. Examples might be a failing hard disk or an overheating CPU causing them to do things that shouldn't be possible.

So what do you do? Firstly you do what you can to minimise the chance of bugs. Secondly, you accept that there will still be problems (if nothing else the external errors will persist) and you use techniques to handle them. You've already seen the latter with ML's exceptions: we'll look at Java's exceptions here too, but first let's consider ways to reduce the bugs in the code you deliver.

8.1 Minimising Bugs

8.1.1 Modular (Unit) Testing

OOP (strongly) pushes you to develop uncoupled chunks of code in the form of classes. Each class should be testable (mostly) independently of the others. It is much easier to comprehensively test lots of small bits of code and then stitch them together than the stitched result!

8.1.2 Using Assertions

When you are debugging an algorithm, it can be useful to use *assertions* at various stages to mark invariants (things that should be true if your algorithm is working). You'll see these next term in the Algorithms course.

8.1.3 Defensive Programming Styles

You can also learn useful habits for each language that can reduce errors. In C for example, if (`exp`) is true whenever `exp` is greater than 0. The problem with this is that you can accidentally do an assignment without realising it:

```
if (x=5) {...}
else {...}
```

Here the programmer presumably wanted to test whether `x` is 5. What actually happens is `x` is *set* to 5 and the expression itself is evaluated as 5, or always true. All because they used `=` and not `==` by accident.

But you can remove this (very common) error altogether by always writing `(5==x)` and not `(x==5)`. Then the error will be caught by the compiler because `(5=x)` is not valid syntax!

8.1.4 Pair Programming etc.

Another quite effective way to spot bugs is via pair programming. Here you program in pairs insofar as one person writes code, while the other watches over their shoulder, looking for errors or bugs. The writer and the watcher switch roles regularly. Various other such agile programming techniques exist.

8.2 Dealing with Errors

8.2.1 Return Codes

Return Codes

- The traditional imperative way to handle errors is to return a value that indicates success/failure/error

```
public int divide(double a, double b) {  
    if (b==0.0) return -1; // error  
    double result = a/b;  
    return 0; // success  
}
```

- Problems:
 - `if (divide(x,y)<0) System.out.println("Failure!!");`
 - Could ignore the return value
 - Have to keep checking what the return values are meant to signify, etc.
 - The actual result often can't be returned in the same way

Many older languages (C included) have no explicit mechanism for error handling, Instead the common approach is to return the error status via the normal return type: a *return code*. If your function isn't proper and its 'result' is a side effect (i.e. it has a void return type in Java) then we can just return the error code:

```
int setValue(LinkedList<int> list,  
            int element, int value) {  
    if (list.size()>element) {  
        list.set(element,value);  
        return 0; // no error to signal  
    }  
    else return -1; // this element doesn't exist  
}
```

Here the function can only return 0 or -1, the latter being a signal that there was an error (the element didn't exist).

If you have a function that naturally returns a result, you can pick some result values that are used (only) to signal errors:

```
float sqrt(float a) {  
    if (a<0.0) return -1.0;  
    else {  
        ...  
    }  
}
```

Here the `sqrt` function only returns positive roots so we can repurpose all negative floats to signal errors.

```
float sqrt(float a) {  
    if (a<0.0) return -1.0;  
    else {  
        ...  
    }  
}
```

If the return type isn't something we can repurpose (e.g. a custom class) then we can instead pass the output by reference and have the function return an integer to indicate the error state. E.g,

```
SomeCustomClass sqrt(float a) {  
    return new SomeCustomClass(...);  
}
```

becomes

```
int func(float a, SomeCustomClass result ) {  
    if (a<0.0) return -1.0;  
    else result.set(...);  
    return 0;  
}
```

You might see functions that return `null` if they have an error. This is a very bad practice since it relies on the programmer using the function to check for `null`. If they don't, they'll likely try to dereference `null` and their program will die...

In fact, this is a larger problem with the general approach. We are dependent on the programmer testing the return value. Two problems arise: firstly, they could neglect to check (really common); secondly, they end up with really nasty looking code such as:

```
int retval = somefunc();  
  
if (retval==1) {  
    // handle error type 1  
}  
else if (retval==2) {  
    // handle error type 2  
}  
else if (retval==3) {  
    // handle error type 3  
}
```

Here, just writing one line to call one function results in a screen-worth of error handling code. This constant mixing of code and error handling makes the code all but unreadable.

8.2.2 Deferred Error Handling

Deferred Error Handling

- A similar idea (with the same issues) is to set some state in the system that needs to be checked for errors.
- C++ does this for streams:

```
ifstream file( "test.txt" );
if ( file.good() )
{
    cout << "An error occurred opening the file" << endl;
}
```

8.2.3 Exceptions

Exceptions

- An exception is an object that can be *thrown* or *raised* by a method when an error occurs and *caught* or *handled* by the calling code
- Example usage:

```
try {
    double z = divide(x,y);
}
catch(DivideByZeroException d) {
    // Handle error here
}
```

Of course, you already met Java's exceptions in the pre-arrival course, as well as ML's in FoCS. We'll cover the Java concepts in a little more depth here, whilst recapping the content you've done. Note there is a tendency to use the terminology throw/catch rather than raise/handle in OOP languages—I don't know why. First some recap:

Flow Control During Exceptions

- When an exception is thrown, any code left to run in the try block is skipped

```
double z=0.0;
boolean failed=false;
try {
    z = divide(5,0);
    z = 1.0;
}
catch(DivideByZeroException d) {
    failed=true;
}
z=3.0;
System.out.println(z+" "+failed);
```

Throwing Exceptions

- An exception is an object that has Exception as an ancestor
- So you need to create it (with new) before throwing

```
double divide(double x, double y) throws DivideByZeroException {
    if (y==0.0) throw new DivideByZeroException();
    else return x/y;
}
```

Multiple Handlers

- A try block can result in a range of different exceptions. We test them in sequence

```
try {
    FileReader fr = new FileReader("somefile");
    int r = fr.read();
}
catch(FileNotFoundException fnf) {
    // handle file not found with FileReader
}
catch(IOException d) {
    // handle read() failed
}
```

finally

- With resources we often want to ensure that they are closed whatever happens

```
try {
    fr.read();
    fr.close();
}
catch(IOException ioe) {
    // read() failed but we must still close the FileReader
    fr.close();
}
```

Exception Hierarchies

- You can use inheritance hierarchies

```
public class MathException extends Exception {...}
public class InfiniteResult extends MathException {...}
public class DivByZero extends MathException {...}
```

- And catch parent classes

```
try {
    ...
}
catch(InfiniteResult ir) {
    // handle an infinite result
}
catch(MathException me) {
    // handle any MathException or DivByZero
}
```

finally II

- The finally block is added and will *always* run (after any handler)

```
try {
    fr.read();
}
catch(IOException ioe) {
    // read() failed
}
finally {
    fr.close();
}
```

Checked vs Unchecked Exceptions

- **Checked:** must be handled or passed up.
 - Used for recoverable errors
 - Java requires you to declare checked exceptions that your method throws
 - Java requires you to catch the exception when you call the function

```
double somefunc() throws SomeException {}
```

- **Unchecked:** not expected to be handled. Used for programming errors
 - Extends RuntimeException
 - Good example is NullPointerException

Note that once any catch block is matched, the remaining catch blocks are skipped. Whilst you already know about the flow control, you hadn't considered creating your own exceptions:

Creating Exceptions

- Just extend Exception (or RuntimeException if you need it to be unchecked). Good form to add a detail message in the constructor but not required.

```
public class DivideByZero extends Exception {}

public class ComputationFailed extends Exception {
    public ComputationFailed(String msg) {
        super(msg);
    }
}
```

- You can also add more data to the exception class to provide more info on what happened (e.g. store the numerator and denominator of a failed division)

There is an ongoing debate about the value of checked exceptions and they feature in some OOP languages but not others. Most of the time you'll be writing and dealing with checked exceptions in Java. You'll encounter unchecked exceptions only when you mess up in your code.

Aside: It turns out with Java they decided that RuntimeException should inherit from Exception. This means that if you ever write `catch(Exception e) {...}` then you will also catch the unchecked exceptions. So don't ever write that unless you know what you are doing!

Evil I: Exceptions for Flow Control

- At some level, throwing an exception is like a GOTO
- Tempting to exploit this

```
try {
    for (int i=0; ; i++) {
        System.out.println(myarray[i]);
    }
}
catch (ArrayOutOfBoundsException ae) {
    // This is expected
}
```
- This is not good. Exceptions are for exceptional circumstances only
 - Harder to read
 - May prevent optimisations

The code readability argument should be obvious but the second argument warrants more discussion. If you Google the notion of flow control with exceptions, you will probably find many comments that suggest exception throwing is very slow compared to ‘normal’ code execution. This is attributed variously to the need to create an Exception object; the need to create a stack trace; or even just the need to create a message string. Some people report Exception handling was 50 times slower on the first JVMs!

Now, you *could* write a JVM that handled exception throwing efficiently, such that code like that in the slide would carry little performance penalty. But the crucial point is that there is no guarantee that a JVM will do so (and many still don’t). Exceptions are intended to be rare occurrences and it is perfectly reasonable (if not natural) for a JVM creator to assume this and therefore not need to worry about optimising exception handling. Bottom line: this smells bad.

Evil II: Blank Handlers

- Checked exceptions must be handled
- Constantly having to use try..catch blocks to do this can be annoying and the temptation is to just gaffer-tape it for now

```
try {
    FileReader fr = new FileReader(filename);
}
catch (FileNotFoundException fnf) {
}
```

- ...but we never remember to fix it and we could easily be missing serious errors that manifest as bugs later on that are extremely hard to track down

This is a bad habit that novices tend to adopt—try not to develop it yourself. Eclipse at least discourages blank handlers, automatically filling in `e.printStackTrace()` so there’s some record of the

problem printed to the screen. However, in large programs, where there’s often lots of debug output flowing to the console, these messages are easily missed... Better to fill in your handlers!

Evil III: Circumventing Exception Handling

```
try{
    // whatever
}
catch(Exception e) {}
```

- Just don't.

Advantages of Exceptions

- Advantages:
 - Class name can be descriptive (no need to look up error codes)
 - Doesn't interrupt the natural flow of the code by requiring constant tests
 - The exception object itself can contain state that gives lots of detail on the error that caused the exception
 - Can't be ignored, only **handled**

<http://java.sun.com/docs/books/tutorial/essential/exceptions/>

8.2.4 Assertions

Assertions

- Assertions are a form of error checking designed for **debugging** (only)
- They are a simple statement that evaluates a boolean: if it's true nothing happens, if it's false, the program ends.
- In Java:

```
assert (x>0);  
  
// or  
  
assert (a==0) : "Some error message here";
```

Assertions are a simple addition to many languages that can really help development, but they complement exceptions (or other error handling techniques) rather than replace them.

Assertions are NOT for Production Code!

- Assertions are there to help you check the logic of your code is correct i.e. when you're trying to get an algorithm working
- **They should be switched OFF** for code that gets released ("production code")
- In Java, the JVM takes a parameter that enables (-ea) or disables (-da) assertions. The default is for them to be **disabled**.

```
> java -ea SomeClass  
> java -da SomeClass
```

This is important: assertions will kill your program if they detect an error. There's no opportunity to handle the error so they're just for development, not production.

As Oracle Puts It

"Assertions are meant to require that the program be consistent with itself, not that the user be consistent with the program"

Great for Postconditions

- **Postconditions** are things that must be true at the end of an algorithm/function if it is functioning correctly
- E.g.

```
public float sqrt(float x) {  
    float result = ....  
    // blah  
    assert(result>=0.f);  
}
```

Sometimes for Preconditions

- **Preconditions** are things that are assumed true at the start of an algorithm/function
- E.g.

```
private void method(SomeObject so) {  
    assert (so!=null);  
    //...  
}
```

- **BUT you shouldn't** use assertions to check for **public** preconditions

```
public float method(float x) {  
    assert (x>=0);  
    //...
```

- (you should use exceptions for this)

If a use of your method provides bad (nonsensical) inputs, you should offer them the chance to remedy the mistake by throwing an exception. Assertions would just kill the program (if enabled for release, which they shouldn't be), or not catch the error because they are disabled!

Sqrt Example

```
public float method(float x) throws InputException {
    // Input sanitisation (precondition)
    if (x<0.f) throw new InputException();

    float result=0.f;
    // compute sqrt and store in result

    // Postcondition
    assert (result>=0);

    return result;
}
```

For the Last Word on Assertions...

<http://www.oracle.com/technetwork/articles/javase/javapch06.pdf>

The distinction is subtle but important. The 'assert' is only used to test the correctness of the algorithm output when given a valid (positive) input. If the assertion fires, it's programmer error and not user error.

Assertions can be Slow if you Like

```
public int[] sort(int[] arr) {
    int[] result = ...
    // blah
    assert(isSorted(result));
}
```

- Here, isSorted() is presumably quite costly (at least O(n)).
- That's OK for debugging (it's checking the sort algorithm is working, so you can accept the slowdown)
- And will be turned off for production so that's OK
- (*but your assertion shouldn't have side effects*)

NOT for Checking your Compiler/Computer

```
public void method() {
    int a=10;
    assert (a==10);
    //...
}
```

- If this isn't working, there is something much bigger wrong with your system!
- It's pointless putting in things like this

Lecture 9

Copying Objects

Cloning I

- Sometimes we really do want to copy an object

- Java calls this **cloning**
- We need special support for it

Shallow and Deep Copies

```
public class MyClass {  
    private MyOtherClass moc;  
}
```

- Shallow
- Deep

Cloning II

- Every class in Java ultimately inherits from the **Object** class
 - This class contains a clone() method so we just call this to clone an object, right?
 - This can go horribly wrong if our object contains reference types (objects, arrays, etc)

Java Cloning

- So do you want shallow or deep?
 - The default implementation of clone() performs a **shallow copy**
 - But Java developers were worried that this might not be appropriate: they decided they wanted to know for sure that we'd thought about whether this was appropriate
- Java has a **Cloneable** interface
 - If you call clone on anything that doesn't extend this interface, it fails

Java is unusual in that it really, really wants you to use OOP. In your practicals you will have noticed that, even to do simple procedural stuff, you had to encase everything in a class—even the main() method. A further decision they made is that ultimately *all* classes will inherit from a special Object class. i.e. the top of all inheritance trees is Object even though we never explicitly say so in code...

Clone Example I

```
public class Velocity {
    public float vx;
    public float vy;
    public Velocity(float x, float y) {
        vx=x;
        vy=y;
    }
};

public class Vehicle {
    private int age;
    private Velocity vel;
    public Vehicle(int a, float vx, float vy) {
        age=a;
        vel = new Velocity(vx,vy);
    }
};
```

Clone Example II

```
public class Vehicle implements Cloneable {
    private int age;
    private Velocity vel;
    public Vehicle(int a, float vx, float vy) {
        age=a;
        vel = new Velocity(vx,vy);
    }

    public Object clone() {
        return super.clone();
    }
};
```

Here we fill in the `clone()` method using `super.clone()`. You can think of this as doing a byte-for-byte copy of an object in memory. Any primitive types (such as `age`) will therefore be copied. And references will also be copied, but not the objects they point to. Hence this much gets us a shallow copy.

Clone Example III

```
public class Velocity implements Cloneable {
    ....
    public Object clone() {
        return super.clone();
    }
};

public class Vehicle implements Cloneable {
    private int age;
    private Velocity v;
    public Student(int a, float vx, float vy) {
        age=a;
        vel = new Velocity(vx,vy);
    }

    public Object clone() {
        Vehicle cloned = (Vehicle) super.clone();
        cloned.vel = (Velocity)vel.clone();
        return cloned;
    }
};
```

A deep clone requires that we clone the objects that

are referenced (and they, in turn clone any objects they reference, and so on). Here we make `Velocity` cloneable and make sure to clone the member variable that `Vehicle` has.

Cloning Arrays

- Arrays have built in cloning but the contents are only cloned *shallowly*

```
int intarray[] = new int[100];
Vector3D vecarray = new Vector3D[10];

...

int intarray2[] = intarray.clone();
Vector3D vecarray2 = vecarray.clone();
```

Covariant Return Types

- The need to cast the clone return is annoying

```
public Object clone() {
    Vehicle cloned = (Vehicle) super.clone();
    cloned.vel = (Velocity)vel.clone();
    return cloned;
}
```

- Recent versions of Java allow you to override a method in a subclass and change its return type to a subclass of the original's class

```
class A {}
class B extends A {}

class C {
    A mymethod() {}
}

class D extends C {
    B mymethod() {}
}
```

This is a similar concept to the covariant parameter types we met briefly in lecture 7. We saw Java does *not* support that, but it does support this. So if we have:

```
public class A {
    Object work(Object o) {...}
}
```

then the following is not allowed (covariant parameter types):

```
public class B extends A {
    @Override
    public Object work(Person p) {...}
}
```

but this is (covariant return types):

```
public class C extends A {
    @Override
    public Person work(Object o) {...}
}
```

Marker Interfaces

- If you look at what's in the `Cloneable` interface, you'll find it's empty!! What's going on?
- Well, the `clone()` method is already inherited from `Object` so it doesn't need to specify it
- This is an example of a **Marker Interface**
 - A marker interface is an empty interface that is used to label classes
 - This approach is found occasionally in the Java libraries

Copy Constructors II

- Now we can create copies by:

```
Vehicle v = new Vehicle(5, 0.f, 5.f);
```

```
Vehicle vcopy = new Vehicle(v);
```

- This is quite a neat approach, but has some drawbacks which are explored on the Examples Sheet

I won't go into detail on these here. Instead they are on the examples sheet.

You might also see these marker interfaces referred to as *tag interfaces*. They are simply a way to label or tag a class. They can be very useful, but equally they can be a pain (you can't dynamically tag a class, nor can you prevent a tag being inherited by all subclasses).

The `clone()` approach is unique to Java. It can be a bit of a headache, but it was meant to address the shortcomings of the de-facto copying approach in OOP, which is the use of copy constructors:

Copy Constructors I

- Another way to create copies of objects is to define a **copy constructor** that takes in an object of the same type and manually copies the data

```
public class Vehicle {
    private int age;
    private Velocity vel;
    public Vehicle(int a, float vx, float vy) {
        age=a;
        vel = new Velocity(vx,vy);
    }
    public Vehicle(Vehicle v) {
        age=v.age;
        vel = v.vel.clone();
    }
}
```

Lecture 10

Language Evolution

Evolve or Die

- Modern languages start out as a programmer “scratching an itch”: they create something that is particularly suitable for some niche
- If the language is to 'make it' then it has to evolve to incorporate both new paradigms and also the old paradigms that were originally rejected but turn out to have value after all
- The challenge is backwards compatibility: you don't want to break old code or require programmers to relearn your language (they'll probably just jump ship!)
- Let's look at some examples for Java...

Ostensibly this course is about OOP, but in reality few languages can claim to be a pure implementation of any particular paradigm. Even ML offers you imperative programming. Actual languages are a mish-mash of concepts, some of which are inevitably retrofitted. This retrofitting tends to produce ugly syntax and unexpected quirks, so it's good to explore some examples (in Java of course).

Vector

- The original Java included the `Vector` class, which was an expandable array

```
Vector v = new Vector();
v.add(x);
```
- They chose to make it *synchronised*, which just means it is safe to use with multi-threaded programs
- When they introduced Collections, they decided everything should *not* be synchronised
- Created `ArrayList`, which is just an unsynchronised (=better performing) `Vector`
- Had to retain `Vector` for backwards compatibility!

`Vector` has no place in modern Java really, and if you are using it you should stop doing so, in favour of using `ArrayList`. If you need it to be synchronised, this can

be done (see next year for those sticking around in the CST). The only reason `Vector` remains is backwards compatibility. It's handy to know about it though, since it features in a lot of legacy code.

10.1 Generics

The Origins of Generics

```
// Make a TreeSet object
TreeSet ts = new TreeSet();

// Add integers to it
ts.add(new Integer(3));

// Loop through
iterator it = ts.iterator();
while(it.hasNext()) {
    Object o = it.next();
    Integer i = (Integer)o;
}
```

- The original Collections framework just dealt with collections of Objects
 - Everything in Java “is-a” Object so that way our collections framework will apply to any class
 - But this leads to:
 - Constant casting of the result (ugly)
 - The need to know what the return type is
 - Accidental mixing of types in the collection

The Origins of Generics II

```
// Make a TreeSet object
TreeSet ts = new TreeSet();

// Add integers to it
ts.add(new Integer(3));
ts.add(new Person("Bob"));

// Loop through
iterator it = ts.iterator();
while(it.hasNext()) {
    Object o = it.next();
    Integer i = (Integer)o;
}
```

Going to fail for the second element!
(But it will compile: the error will be at runtime)

This is pretty nasty. The OOP paradigm has let us write a flexible data structure that can handle us wrap-

ping around various types, but it can't apply the restriction that all the types in one object should be the same. Additionally, all this casting makes for ugly code. This is what convinced the Java designers that parameterised types (Generics) were needed. But it was already a bit late: there was tons of established code using Collections (and still is). The Java designers were faced with the problem of updating the language to support parameterised types without breaking everything that went before.

Generics and SubTyping

```

classDiagram
    class Animal
    class Person
    Animal <|-- Person
    
```

```

// Object casting
Person p = new Person();
Animal o = (Animal) p;

// List casting
List<Person> plist = new LinkedList<Person>();
List<Animal> alist = (List<Animal>)plist;
    
```

So a list of **Persons** is a list of **Animals**, yes?

The Generics Solution

- Java implements *type erasure*
 - Compiler checks through your code to make sure you only used a single type with a given Generics object
 - Then it deletes all knowledge of the parameter, converting it to the old code invisibly

```

LinkedList<Integer> ll =
new LinkedList<Integer>();
...
for (Integer i : ll) {
do_something(i);
}
    
```

➔

```

LinkedList ll =
new LinkedList();
...
for (Object i : ll) {
do_something( (Integer)i );
}
    
```

So now we see why we can't use primitives as parameters: whatever we put there must be castable to Object, which primitives simply aren't.

10.2 Java 8

The C++ Templates Solution

- Compiler first generates the class definitions from the template

```

class MyClass<T> {
T membervar;
};
    
```

➔

```

class MyClass_float {
float membervar;
};
class MyClass_int {
int membervar;
};
class MyClass_double {
double membervar;
};
...
    
```

Adding Functional Elements...

- Java is undeniably imperative, but there is something seductive about some of the highly succinct and efficient syntax

```

result=map (fn x => (x+1)*(x+1)) numlist;

int[] result = new int[numlist.length];
for (int i=0; i<numlist.length; i++) {
result[i] = (numlist[i]+1)*(numlist[i]+1)
}
    
```

- Enter Java 8...

Java 8 is a major Java release. A lot has been added, some of it controversial. Not much of it relates to OOP but we discuss it partly for interest and partly because it emphasises how (big) languages tend to just subsume multiple paradigms, blurring the boundaries more and more

C++ doesn't suffer from the same problem since it just generates a special class for each instance you request.

Lambda Functions

- Supports anonymous functions

```
()->System.out.println("It's nearly over...");

s->s+"hello";

s->{s=s+"hi";
    System.out.println(s);}

(x,y)->x+y;
```

Functions as Values

```
// No arguments
Runnable r = ()->System.out.println("It's nearly over...");
r.run();

// No arguments, non-void return
Callable<Double> pi = ()->3.141;
pi.call();

// One argument, non-void return
Function<String,Integer> f = s->s.length();
f.apply("Seriously, you can go soon")
```

Method References

- Can use established functions too

```
System.out::println
Person::doSomething
Person::new
```

New forEach for Lists

```
List<String> list = new LinkedList<>();
list.add("Just a");
list.add("few more slides");

list.forEach(System.out::println);

list.forEach(s->System.out::println(s));

list.forEach(s->{s=s.toUpperCase();
    System.out::println(s);});
```

Note this is effectively our beloved ‘map’ function from ML!

Sorting

- Who needs Comparators?

```
List<String> list = new LinkedList<>();
....
Collections.sort(list, (s1, s2) -> s1.length() - s2.length());
```

Streams

- Collections can be made into streams (sequences)
- These can be **filtered** or **mapped**!

```
List<Integer> list = ...
list.stream().map(x->x+10).collect(Collectors.toList());
list.stream().filter(x->x>5).collect(Collectors.toList());
```

This is a more explicit introduction of the ‘filter’ and ‘map’ features seen in functional programming (you didn’t actually meet ‘filter’ formally in FoCS, but it just filters a list according to some supplied predicate). However, notice how ugly the syntax has become...

Lecture 11

Design Patterns

Design Patterns

- A **Design Pattern** is a general reusable solution to a commonly occurring problem in software design
- Coined by Erich Gamma in his 1991 Ph.D. thesis
- Originally 23 patterns, now many more. Useful to look at because they illustrate some of the power of OOP (and also some of the pitfalls)
- We will only consider a subset

Coding anything more complicated than a toy program usually benefits from forethought. After you've coded a few medium-sized pieces of object-oriented software, you'll start to notice the same general problems coming up over and over. And you'll start to automatically use the same solutions to them. We need to make sure that set of default solutions is a good one!

In his 1991 PhD thesis, Erich Gamma compared this to the field of architecture, where recurrent problems are tackled by using known good solutions. The follow-on book (**Design Patterns: Elements of Reusable Object-Oriented Software, 1994**) identified a series of commonly encountered problems in object-oriented software design and 23 solutions that were deemed elegant or good in some way. Each solution is known as a *Design Pattern*:

A Design Pattern is a general reusable solution to a commonly occurring problem in software design.

The modern list of design patterns is ever-expanding and there is no shortage of literature on them. In this course we will look at a few key patterns and how they are used.

11.0.1 So Design Patterns are like coding recipes?

No. Creating software by stitching together a series of Design Patterns is like painting by numbers — it's easy and it probably works, but it doesn't produce a Picasso! Design Patterns are about intelligent solutions to a series of generalised problems that you *may* be able to identify in your software. You might find they don't apply to your problem, or that they need adaptation. You simply can't afford to disengage your brain (sorry!).

11.0.2 Why Bother Studying Them?

Design patterns are useful for a number of things, not least:

1. They encourage us to identify the fundamental aims of given pieces of code
2. They save us time and give us confidence that our solution is sensible
3. They demonstrate the power of object-oriented programming
4. They demonstrate that naïve solutions are bad
5. They give us a common vocabulary to describe our code

The last one is important: when you work in a team, you quickly realise the value of being able to succinctly describe what your code is trying to do. If you can replace twenty lines of comments¹ with a single word, the code becomes more readable and maintainable. Furthermore, you can insert the word into the class name itself, making the class self-describing.

¹You are commenting your code liberally, aren't you?

11.0.3 The Open-Closed Principle

The Open-Closed Principle

***Classes should be open for extension
but closed for modification***

- i.e. we would like to be able to modify the behaviour without touching its source code
- This rule-of-thumb leads to more reliable large software and will help us to evaluate the various design patterns

To help understand why this is helpful, it's useful to think about multiple developers using a software library. If they want to alter one of the classes in the library, they could edit its source code. But this would mean they had a customised version of the library that they wouldn't be able to update when new (bug-reduced) versions appeared. A better solution is to use the library class as a base class and implement the minor changes that are desired in the custom child. So, if you're writing code that others will use (and you should *always* assume you are in OOP) you should make it easy for them to extend your classes and discourage direct editing of them.

11.0.4 The Decorator Pattern

Decorator

Abstract problem: How can we add state or methods at runtime?

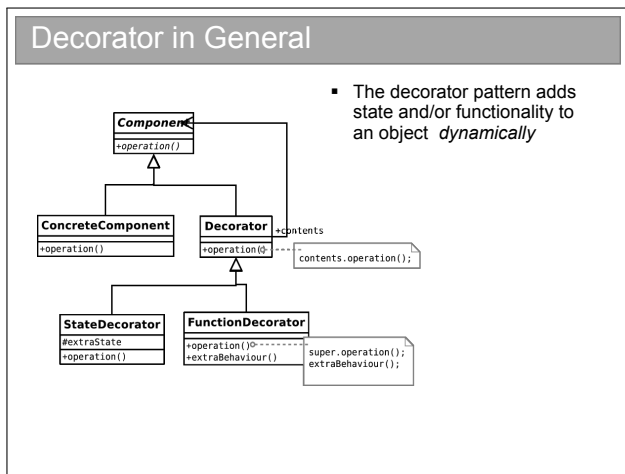
Example problem: How can we efficiently support gift-wrapped books in an online bookstore?

want. In the diagram above, I have explicitly allowed for both options by deriving `StateDecorator` and `FunctionDecorator`. This is usually unnecessary — in our book seller example we only want to decorate one thing so we might as well just put the code into `Decorator`.

Solution 1: Add variables to the established `Book` class that describe whether or not the product is to be gift wrapped.

Solution 2: Extend `Book` to create `WrappedBook`.

Solution 3: (Decorator) Extend `Book` to create `WrappedBook` and also add a member reference to a `Book` object. Just pass through any method calls to the internal reference, intercepting any that are to do with shipping or price to account for the extra wrapping behaviour.



So we take an object and effectively give it extra state or functionality. I say ‘effectively’ because the actual object in memory is untouched. Rather, we create a new, small object that ‘wraps around’ the original. To remove the wrapper we simply discard the wrapping object. Real world example: humans can be ‘decorated’ with contact lenses to improve their vision.

Note that we can use the pattern to add state (variables) or functionality (methods), or both if we

11.0.5 The Singleton Pattern

Singleton

Abstract problem: How can we ensure only one instance of an object is created by developers using our code?

Example problem: You have a class that encapsulates accessing a database over a network. When instantiated, the object will create a connection and send the query. Unfortunately you are only allowed one connection at a time.

A valid solution to this is to make sure you close the database connection after using it, so you can just create Database objects every time you have a query. However, what if you forgot to close it? And what if making the connection was slow (they always are in computer time...).

Instead we exploit our access modifiers and create a private constructor (to ensure no-one can create objects at will) and add in a static member (the only instance we will ever have). Finally, we include a static getter for this member.

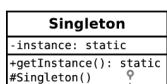
Ideally the instantiation of the Database should be *lazy*—i.e. only done on the first call to the getter.

Protected members are accessible to the class, any subclasses, and all classes in the same package. Therefore, any class in the same package as your base class will be able to instantiate Singleton objects at will, using the new keyword!

Additionally, we don't want a crafty user to subclass our singleton and implement Cloneable on their version. How could you ensure this doesn't happen?

Singleton in General

- The singleton pattern ensures a class has only one instance and provides global access to it



```
if (instance==null) instance=new Singleton();
return instance;
```

There is a caveat with Java. If you choose to make the constructor protected (this would be useful if you wanted a singleton base class for multiple applications of the singleton pattern, and is actually the 'official' solution) you have to be careful.

11.0.6 The State Pattern

State

Abstract problem: How can we let an object alter its behaviour when its internal state changes?

Example problem: Representing academics as they progress through the rank

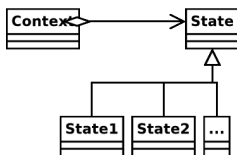
Solution 1: Have an abstract Academic class which acts as a base class for Lecturer, Professor, etc.

Solution 2: Make Academic a concrete class with a member variable that indicates rank. To get rank-specific behaviour, check this variable within the relevant methods.

Solution 3: (State) Make Academic a concrete class that has-a AcademicRank as a member. Use AcademicRank as a base for Lecturer, Professor, etc., implementing the rank-specific behaviour in each..

State in General

- The state pattern allows an object to cleanly alter its behaviour when internal state changes



11.0.7 The Strategy Pattern

Strategy

Abstract problem: How can we select an algorithm implementation at runtime?

Example problem: We have many possible change-making implementations. How do we cleanly change between them?

- Strategy is about encapsulating behaviour in a class. This behaviour does not depend on internal variables.
- Different concrete Strategies may produce exactly the same output, but do so in a different way. For example, we might have a new algorithm to compute the standard deviation of some variables. Both the old algorithm and the new one will produce the same output (hopefully), but one may be faster than the other. The Strategy pattern lets us compare them cleanly.
- Strategy in the strict definition usually assumes the class is selected at compile time and not changed during runtime.
- The usage of the Strategy pattern is normally visible to external classes. i.e. there will be a `set-Strategy(Strategy s)` function or it will be set in the constructor.

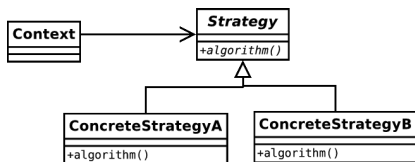
Solution 1: Use a lot of if...else statements in the `getChange(...)` method.

Solution 2: (Strategy) Create an abstract `ChangeFinder` class. Derive a new class for each of our algorithms.

However, the similarities do cause much debate and you will find people who do not differentiate between the two patterns as strongly as I tend to.

Strategy in General

- The strategy pattern allows us to cleanly interchange between algorithm implementations



Note that this is essentially the same UML as the State pattern! The *intent* of each of the two patterns is quite different however:

- State is about encapsulating behaviour that is linked to specific internal state within a class.
- Different states produce different outputs (externally the class behaves differently).
- State assumes that the state will continually change at run-time.
- The usage of the State pattern is normally invisible to external classes. i.e. there is no `set-State(State s)` function.

11.0.8 The Composite Pattern

Composite

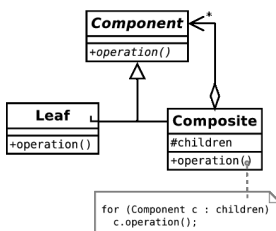
Abstract problem: How can we treat a group of objects as a single object?

Example problem: Representing a DVD box-set as well as the individual films without duplicating info and with a 10% discount

The solution is fairly straightforward. We want to be able to treat a group of DVDs to just like a single DVD, so `BoxSet` inherits from `DVD`. To avoid repeating the description information and to keep pricing in sync, `BoxSet` must also have access to the constituent `DVD` objects.

Composite in General

- The composite pattern lets us treat objects and groups of objects uniformly



If you're still awake, you may be thinking this looks like the `Decorator` pattern, except that the new class supports associations with multiple DVDs (note the * by the arrowhead). Plus the intent is different—we are not adding new functionality to objects but rather supporting the same functionality for groups of those objects.

If you try to make a graphical representation of composites, you'll end up with some form of tree with each composite a node and each single entity a leaf. Many texts use this terminology when discussing the composite pattern.

11.0.9 The Observer Pattern

Observer

Abstract problem: When an object changes state, how can any interested parties know?

Example problem: How can we write phone apps that react to accelerator events?

This pattern is used regularly, but is particularly useful for event-based programs. The process is analogous to a magazine subscription: you *subscribe* with the publisher in order to receive *publish* events (magazines) as soon as they are available. In design patterns parlance, you are an observer of the publisher, who is the subject. It should be clear that this is also a very important pattern for the various proxy implementations if the source information might change during use.

In an Android smartphone, the system provides a subject in the form of a `SensorManager` object, which is actually a singleton (only one manager at any time). So we get it by calling:

```
SensorManager sManager = (SensorManager)
    getSystemService(SENSOR_SERVICE);
```

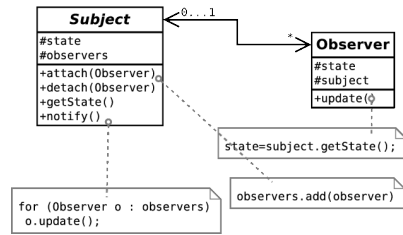
We then register with it with a line like:

```
sManager.registerListener(this,
    sManager.getDefaultSensor(
        Sensor.TYPE_ACCELEROMETER),
    SensorManager.SENSOR_DELAY_NORMAL);
```

Our class must implement `SensorEventListener`, which forces us to specify a `onSensorEvent()` method. Whenever the system gets a new accelerometer reading, it cycles over all the objects that have registered with it, feeding them the new reading.

Observer in General

- The observer pattern allows an object to have multiple dependents and propagates updates to the dependents automatically.



11.0.10 Classifying Patterns

Often patterns are classified according to what their intent is or what they achieve. The original book defined three classes:

Creational Patterns . Patterns concerned with the creation of objects (e.g. Singleton, Abstract Factory).

Structural Patterns . Patterns concerned with the composition of classes or objects (e.g. Composite, Decorator, Proxy).

Behavioural Patterns . Patterns concerned with how classes or objects interact and distribute responsibility (e.g. Observer, State, Strategy).

11.0.11 Other Patterns

You've now met a few Design Patterns. There are plenty more (23 in the original book and many, many more identified since), but this course will not cover them. What has been presented here should be sufficient to:

- Demonstrate that object-oriented programming is powerful.
- Provide you with (the beginnings of) a vocabulary to describe your solutions.
- Make you look critically at your code and your software architectures.
- Entice you to read further to improve your programming.

Of course, you probably won't get it right first time (if there even is a 'right'). You'll probably end up *refactoring* your code as new situations arise. However, if

a Design Pattern *is* appropriate, you should probably use it.

11.0.12 Performance

Note that all of the examples here have concentrated on structuring code to be more readable and maintainable, and to incorporate constraints structurally where possible. At no point have we discussed whether the solutions *perform* better. Many of the solutions exploit runtime polymorphic behaviour, for example, and that carries with it certain overheads.

This is another reason why you can't apply Design Patterns blindly. [This is a good thing since, if it wasn't true, programming wouldn't be interesting, and you wouldn't get jobs!].

Appendix I: Java, the JVM and Bytecode

Java is known for its cross-platform abilities, which has given it strong internet credentials. Being able to send a file compiled on one machine to another machine with a different architecture and have it run is a neat trick. It shouldn't work because the machine code for one machine shouldn't make sense to another.

Interpreter to Virtual Machine

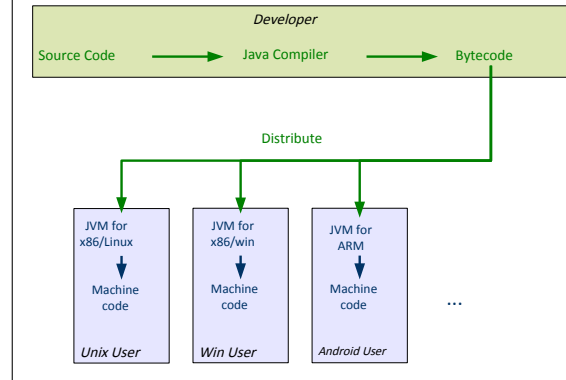
- Java was born in an era of internet connectivity. SUN wanted to distribute programs to internet machines
 - But many architectures were attached to the internet – how do you write one program for them all?
 - And how do you keep the size of the program small (for quick download)?
- Could use an interpreter (→ Javascript). But:
 - High level languages not very space-efficient
 - The source code would implicitly be there for anyone to see, which hinders commercial viability.
- Went for a clever hybrid interpreter/compiler

Java Bytecode I

- SUN envisaged a hypothetical **Java Virtual Machine (JVM)**. Java is compiled into machine code (called **bytecode**) for that (imaginary) machine. The bytecode is then distributed.
- To use the bytecode, the user must have a JVM that has been specially compiled for their architecture.
- The JVM takes in bytecode and spits out the correct machine code for the local computer. i.e. is a **bytecode interpreter**

So the trick is to *partially* compile the Java code to a machine code for a universal machine (that doesn't actually exist). To actually *use* this special machine code ("bytecode") a machine must translate from bytecode to its own local machine code. To that it must have a Java Virtual Machine (JVM) installed that knows the translation.

Java Bytecode II



Java Bytecode III

- + Bytecode is compiled so not easy to reverse engineer
- + The JVM ships with tons of libraries which makes the bytecode you distribute small
- + The toughest part of the compile (from human-readable to computer readable) is done by the compiler, leaving the computer-readable bytecode to be translated by the JVM (→ easier job → faster job)
- Still a performance hit compared to fully compiled ("native") code