CS 538

Introduction to the Theory and Design of Programming Languages

Charles N. Fischer

Fall 2004

http://www.cs.wisc.edu/~fischer/cs538.html

Class Meets
Mondays, Wednesdays & Fridays,
11:00 — 11:50
103 Psychology

Instructor
Charles N. Fischer
5397 Computer Sciences
Telephone: 262-6635
E-mail: fischer@cs.wisc.edu
Office Hours:
10:30 - Noon, Tuesdays & Thursdays, or by appointment

Teaching Assistant
Kevin Roundy
3360 Computer Sciences
Telephone: 262-9275
E-mail: roundy@cs.wisc.edu
Office Hours:
9:30 - 11:00, Mondays &
12:30 - 1:45 Wednesdays,
or by appointment

Key Dates
- October 8: Homework #1 (tentative)
- October 29: Programming Assignment #1 - Scheme (tentative)
- November 8: Midterm Exam (tentative)
- November 19: Programming Assignment #2 - Standard ML (tentative)
- December 8: Programming Assignment #3 - Prolog (tentative)
- December 15: Programming Assignment #4 - Tiger Java and Python
- December 20: Final Exam 2:45pm-4:45pm
**Class Text**
- Required text:
- Suggested supplemental Class Text:
- Handouts and Web-based reading will also be used.

**Reading Assignment**
- Roosta: Chapters 1-3 (as background)
- Webber: Chapters 1, 10, 18 (as background)

**Class Notes**
- Each lecture will be made available prior to, and after, that lecture on the class Web page (under the “Lecture Nodes” link).

**Instructional Computers**
Departmental Linux Machines (royal1-royal30) have been assigned to CS 538. All necessary compiler, interpreters and tools will be loaded onto these machines.

You may also use your own PC or workstation. It will be your responsibility to load needed software (instructions on where to find needed software are included on the class web page).

The Systems Lab teaches brief tutorials on Linux if you are unfamiliar with that OS.

**Academic Misconduct Policy**
- You must do your own assignments — **NO** copying or sharing of solutions.
- You may discuss general concepts and ideas.
- All cases of misconduct must be reported to the Dean's office.
- Penalties may be severe.
**Program & Homework Late Policy**

- An assignment may be handed in one, two, or three class periods late, but not any later.
- One late period will be debited 10%, two late periods will be debited 20%, three late periods will be debited 30%.
- All students are given 4 “free” late periods. That is, the first 40% in late debits will be automatically forgiven.
- Your 4 free late periods may be used at any time, and in any combination.

**Approximate Grade Weights**

- Homework 1: 10%
- Program 1 - Scheme: 16%
- Program 2 - ML: 16%
- Program 3 - Prolog: 12%
- Program 4 - Tiger Java & Python: 6%
- Midterm Exam: 20%
- Final Exam (non-cumulative): 20%

**Programming Languages to be Considered in Detail**

1. Scheme

   A modern variant of Lisp.

   **A Functional Language**: Functions are “first class” data values.

   Dynamically Typed: A variable’s type may change during execution; no type declarations are needed.

   All memory allocation and deallocation is automatic.

   Primary data structures, lists and numbers, are unlimited in size and may grow without bound.

   Continuations provide a novel way to suspend and “re-execute” computations.

2. ML (“Meta Language”)

   Strong, compile-time type checking. Types are determined by inference rather than declaration.

   Naturally polymorphic (one function declaration can be used with many different types).

   Pattern-directed programming (you define patterns that are automatically matched during a call).

   Typed exceptions are provided.

   Abstract data types, with constructors, are included.
3. **Prolog (Programming in Logic)**

Programs are Facts and Rules.

Programmers are concerned with definition, not execution.

Execution order is automatically determined.

4. **Pizza**

Extends a popular Object-oriented language, Java, to include

- Parametric polymorphism (similar to C++’s templates).
- First-class functional objects.
- Algebraic data types, including patterns.

5. **C#**

Microsoft’s answer to Java. In most ways it is very similar to Java, with some C++ concepts reintroduced and some useful additions.

- Events and delegates are included to handle asynchronous actions (like keyboard or mouse actions).
- Properties allow user-defined read and write actions for fields.
- Indexers allow objects other than arrays to be indexed.
- Collection classes may be directly enumerated:

```
foreach (int i in array) ...
```

- Structs and classes co-exist and may be inter-converted (boxed and unboxed).

6. **Java 1.5 (Tiger Java)**

Extends current definition of Java to include:

- Parametric polymorphism (collection types may be parameterized).
- Enhanced loop iterators.
- Automatic boxing and unboxing of wrapper classes.
- Typesafe enumerations.
- **Static imports** (out.println rather than System.out.println).
- Variable argument methods.
- Formatted output using `printf`:

```
out.printf("Ans = %3d",a+b);
```
7. Python
A simple, efficient scripting language that quickly builds new programs out of existing applications and libraries. It cleanly includes objects. It scales nicely into larger applications.

Evolution of Programming Languages

In the beginning, ...
programs were written in absolute machine code—a sequence of bits that encode machine instructions.
Example:

```
34020005
0000000c
3c011001
ac220000
```

This form of programming is
• Very detailed
• Very tedious
• Very error-prone
• Very machine specific

Symbolic Assemblers
Allow use of symbols for operation codes and labels.
Example:

```
li $v0,5
syscall
sw $v0,a
```

Far more readable, but still very detailed, tedious and machine-specific.
Types are machine types.
Control structures are conditional branches.
Subprograms are blocks of code called via a “subroutine branch” instruction.
All labels are global.

Fortran (Formula Translator)
Example:

```
do 10 i=1,100
  10 a(i)=0
```

Developed in the mid-50s.
A major step forward:
• Programming became more “problem oriented” and less “machine oriented.”
• Notions of control structures (ifs and do loops) were introduced.
• Subprograms, calls, and parameters were made available.
• Notions of machine independence were introduced.
• Has evolved into many new variants, including Fortran 77, Fortran 90 and HPF (High Performance Fortran).
**Cobol (Common Business Oriented Language)**

Example:

- multiply $i$ by 3 giving $j$.
- move $j$ to $k$.
- write line 1 after advancing 1 lines.

Developed in the early 60s.
The first widely-standardized programming language.
Once dominant in the commercial world; still important.
Wordy in structure; designed for non-scientific users.
Raised the issue of who should program and how important readability and maintainability are.

**Algol 60 (Algorithmic Language)**

Example:

```plaintext
real procedure cheb(x, n);
value x, n;
real x; integer n;
cheb := if n = 0 then 1
else if n = 1 then x
else 2 * x * 
    cheb(x, n - 1) - cheb(x, n - 2);
```

Developed about 1960.
A direct precursor of Pascal, C, C++ and Java.
Introduced many ideas now in wide use:
- Blocks with local declarations and scopes.
- Nested declarations and control structures.

**Lisp (List Processing Language)**

Example:

```plaintext
((lambda (x) (* x x)) 10)
```

Developed in the early 60s.
A radical departure from earlier programming languages.
Programs and data are represented in a uniform list format.
Types are a property of data values, not variables or parameters.
A program can build and run new functions as it executes.
Data values were not fixed in size.
Memory management was automatic.
A formal semantics was developed to define precisely what a program means.

- Parameter passing
- Automatic recursion.

But,
- I/O wasn’t standardized.
- IBM promoted Fortran and PL/I.
Simula 67 (Simulation Algol)

Example:

```plaintext
Class Rectangle (Width, Height);
Real Width, Height;
Boolean Procedure IsSquare;
   IsSquare := Width=Height;
End of Rectangle;

Developed about 1967.

Introduced the notion of a class (for simulation purposes).

Included objects, a garbage collector, and notions of extending a class.

C++ was originally C with classes (as Simula was Algol with classes).
```

C and C++

C was developed in the early 70’s; C++ in the mid-80s.

These languages have a concise, expressive syntax; they generate high quality code sufficient for performance-critical applications.

C, along with Unix, provided the viability of platform-independent languages and applications.

C and C++ allow programmers a great deal of freedom in bending and breaking rules.

Raises the issue of whether one language can span both novice and expert programmers.

Interesting issue—if most statements and expressions are meaningful, can errors be readily detected?

```plaintext
if (a=b)
   a=0;
else a = 1;
```

Java

Developed in the late 90s.

Cleaner object-oriented language than C++.

Introduced notions of dynamic loading of class definitions across the Web.

Much stronger emphasis on secure execution and detection of run-time errors.

Extended notions of platform independence to system independence.
What Drives Research into New Programming Languages?

Why isn’t C or C++ or C+++/ enough?

1. Curiosity
   What other forms can a programming language take?
   What other notions of programming are possible?

2. Productivity
   Procedural languages, including C, C++ and Java, are very detailed.
   Many source lines imply significant development and maintenance expenses.

3. Reliability
   Too much low-level detail in programs greatly enhances the chance of minor errors. Even minor errors can raise significant problems in applications.

4. Security
   Computers are entrusted with ever-increasing responsibilities. How can we know that a program is safe and reliable enough to trust?

5. Execution speed
   Procedural languages are closely tied to the standard sequential model of instruction execution. We may need radically different programming models to fully exploit parallel and distributed computers.

Desirable Qualities in a Programming Language

Theoretically, all programming languages are equivalent (Why?)
If that is so, what properties are desirable in a programming language?

- It should be easy to use.
  Programs should be easy to read and understand.
  Programs should be simple to write, without unexpected pitfalls.
  It should be orthogonal, providing only one way to do each step or computation.
  Its notation should be natural for the application being programmed.

- The language should support abstraction.
  You can’t anticipate all needed data structures and operations, so adding new definitions easily and efficiently should be allowed.
• The language should support testing, debugging and verification.
• The language should have a good development environment.
  Integrated editors, compilers, debuggers, and version control are a big plus.
• The language should be portable, spanning many platforms and operating systems.

• The language should be inexpensive to use:
  Execution should be fast.
  Memory needs should be modest.
  Translation should be fast and modular.
  Program creation and testing should be easy and cheap.
  Maintenance should not be unduly cumbersome.
  Components should be reusable.

Programming Paradigms

Programming languages naturally fall into a number of fundamental styles or paradigms.

Procedural Languages

Most of the widely-known and widely-used programming languages (C, Fortran, Pascal, Ada, etc.) are procedural in nature.

Programs execute statement by statement, reading and modifying a shared memory.
This programming style closely models conventional sequential processors linked to a random access memory (RAM).

Question:
Given
\[
a = a + 1;
\]
\[
if \ (a > 10) \ \\
\qquad b = 10;
\]
\[
else \ b = 15;
\]
\[
a = a \times b;
\]

Why can't 5 processors each execute one line to make the program run 5 times faster?
Functional Languages
Lisp, Scheme and ML are functional in nature.
Programs are expressions to be evaluated.
Language design aims to minimize side-effects, including assignment.
Alternative evaluation mechanisms are possible, including
Lazy (Demand Driven)
Eager (Data Driven or Speculative)

Object-Oriented Languages
C++, Java, Smalltalk, Pizza and Python are object-oriented.
Data and functions are encapsulated into Objects.
Objects are active, have persistent state, and uniform interfaces (messages or methods).
Notions of inheritance and common interfaces are central.
All objects that provide the same interface can be treated uniformly. In Java you can print any object that provides the toString method. You can iterate through the elements of any Java object that implements the Enumeration interface.

Subclassing allows to you extend or redefine part of an object’s behavior without reprogramming all of the object’s definition. Thus in Java, you can take a Hashtable class (which is fairly elaborate) and create a subclass in which an existing method (like toString) is redefined, or new operations are added.

Logic Programming Languages
Prolog notes that most programming languages address both the logic of a program (what is to be done) and the control flow of a program (how to do what you want).
A logic programming language, like Prolog, lets programmers focus on a program’s logic without concern for control issue.
These languages have no real control structures, and little notion of “flow of control.”
What results are programs that are unusually succinct and focused.
Example:

\[ \text{inOrder( [ ] ).} \]
\[ \text{inOrder( [ _ ] ).} \]
\[ \text{inOrder([a,b|c]) :- (a<b),} \]
\[ \text{inOrder([b|c]).} \]

This is a complete, executable function that determines if a list is in order. It is naturally polymorphic, and is not cluttered with declarations, variables or explicit loops.

Review of Concepts from Procedural Programming Languages

Declarations/Scope/Lifetime/Binding

Static/Dynamic

- Identifiers are declared, either explicitly or implicitly (from context of first use).
- Declarations bind type and kind information to an identifier. Kind specifies the grouping of an identifier (variable, label, function, type name, etc.)
- Each identifier has a scope (or range) in a program—that part of the program in which the identifier is visible (i.e., may be used).

Data objects have a lifetime—the span of time, during program execution, during which the object exists and may be used.

- Lifetimes of data objects are often tied to the scope of the identifier that denotes them. The objects are created when its identifier’s scope is entered, and they may be deleted when the identifier’s scope is exited. For example, memory for local variables within a function is usually allocated when the function is called (activated) and released when the call terminates.
- In Java, a method may be loaded into memory when the object it is a member of is first accessed.

Properties of an identifier (and the object it represents) may be set at

- Compile-time
  
  These are static properties as they do not change during execution. Examples include the type of a variable, the value of a constant, the initial value of a variable, or the body of a function.

- Run-time

  These are dynamic properties. Examples include the value of a variable, the lifetime of a heap object, the value of a function’s parameter, the number of times a while loop iterates, etc.
Example: In Fortran

- The scope of an identifier is the whole program or subprogram.
- Each identifier may be declared only once.
- Variable declarations may be implicit. (Using an identifier implicitly declares it as a variable.)
- The lifetime of data objects is the whole program.

Block Structured Languages

- Include Algol 60, Pascal, C and Java.
- Identifiers may have a non-global scope. Declarations may be local to a class, subprogram or block.
- Scopes may nest, with declarations propagating to inner (contained) scopes.
- The lexically nearest declaration of an identifier is bound to uses of that identifier.

Binding of an identifier to its corresponding declaration is usually static (also called lexical), though dynamic binding is also possible. Static binding is done prior to execution—at compile-time.

Example (drawn from C):

```c
int x, z;
void A() {
    float x, y;
    print(x, y, z);
}
void B() {
    print(x, y, z)
}
```

Block Structure Concepts

- Nested Visibility
  - No access to identifiers outside their scope.
- Nearest Declaration Applies
  - Static name scoping.
- Automatic Allocation and Deallocation of Locals
  - Lifetime of data objects is bound to the scope of the Identifiers that denote them.
Variations in these rules of name scoping are possible. For example, in Java, the lifetime of all class objects is from the time of their creation (via `new`) to the last visible reference to them. Thus

```java
... Object o; ...
```
creates an object reference but does not allocate any memory space for `o`. You need

```java
... Object o = new Object(); ...
```
to actually create memory space for `o`.

**Dynamic Scoping**

An alternative to static scoping is dynamic scoping, which was used in early Lisp dialects (but not in Scheme, which is statically scoped).

Under dynamic scoping, identifiers are bound to the dynamically closest declaration of the identifier. Thus if an identifier is not locally declared, the call chain (sequence of callers) is examined to find a matching declaration.

**Example:**

```java
int x;
void print() {
    write(x); }
main () {
    bool x;
    print();
}
```

Under static scoping the `x` written in `print` is the lexically closest declaration of `x`, which is as an `int`. Under dynamic scoping, since `print` has no local declaration of `x`, `print`'s caller is examined. Since `main` calls `print`, and it has a declaration of `x` as a `bool`, that declaration is used.

Dynamic scoping makes type checking and variable access harder and more costly than static scoping. (Why?) However, dynamic scoping does allow a notion of an "extended scope" in which declarations extend to subprograms called within that scope. Though dynamic scoping may seem a bit bizarre, it is closely related to virtual functions used in C++ and Java.
Virtual Functions

A function declared in a class, C, may be redeclared in a class derived from C. Moreover, for uniformity of redeclaration, it is important that all calls, including those in methods within C, use the new declaration.

Example:
```java
class C {
    void DoIt() {
        PrintIt();
    }
    void PrintIt()
    {
        println("C rules!");
    }
}
class D extends C {
    void PrintIt()
    {
        println("D rules!");
    }
    void TestIt() {DoIt();}
}
D dvar = new D();
dvar.TestIt();
```

Scope vs. Lifetime

It is usually required that the lifetime of a run-time object at least cover the scope of the identifier. That is, whenever you can access an identifier, the run-time object it denotes better exist.

But, it is possible to have a run-time object’s lifetime exceed the scope of its identifier. An example of this is static or own variables.

In C:
```c
void p() {
    static int i = 0;
    print(i++);
}
```

Each call to p prints a different value of i (0, 1, ...). Variable i retains its value across calls.

Some languages allow an explicit binding of an identifier for a fixed scope:

Let
```c
id = val
in
    type id = val;
    statements
end;
```

A declaration may appear wherever a statement or expression is allowed. Limited scopes enhance readability.

Structs vs. Blocks

Many programming languages, including C, C++, C#, Pascal and Ada, have a notion of grouping data together into structs or records.

For example:
```c
struct complex { float re, im; }
```

There is also the notion of grouping statements and declarations into blocks:
```c
{ float re, im;
    re = 0.0; im = 1.0;
}
```
Blocks and structs look similar, but there are significant differences:

**Structs** are data,
- As originally designed, structs contain only data (no functions or methods).
- Structs can be dynamically created, in any number, and included in other data structures (e.g., in an array of structs).
- All fields in a struct are visible outside the struct.

**Blocks** are code,
- They can contain both code and data.
- Blocks can’t be dynamically created during execution; they are “built into” a program.
- Locals in a block aren’t visible outside the block.

By adding functions and initialization code to structs, we get classes—a nice blend of structs and blocks.

For example:
```c
class complex{
    float re, im;
    complex (float v1, float v2){
        re = v1; im = v2; }
}
```

**Classes**
- Class objects can be created as needed, in any number, and included in other data structure.
- They include both data (fields) and functions (methods).
- They include mechanisms to initialize themselves (constructors) and to finalize themselves (destructors).
- They allow controlled access to members (private and public declarations).

**Type Equivalence in Classes**

In C, C++ and Java, instances of the same struct or class are type-equivalent, and mutually assignable.

For example:
```c
class MyClass { ... }
MyClass v1, v2;
v1 = v2; // Assignment is OK
```

We expect to be able to assign values of the same type, including class objects.

However, sometimes a class models a data object whose size or shape is set upon creation (in a constructor). Then we may not want assignment to be allowed.
class Point {
    int dimensions;
    float coordinates[];
    Point () {
        dimensions = 2;
        coordinates = new float[2];
    }
    Point (int d) {
        dimensions = d;
        coordinates = new float[d];
    }
    Point plane = new Point();
    Point solid  = new Point(3);
    plane = solid; //OK in Java
}

This assignment is allowed, even though the two objects represent points in different dimensions.

Subtypes
In C++, C# and Java we can create subclasses—new classes derived from an existing class.
We can use subclasses to create new data objects that are similar (since they are based on a common parent), but still type-inequivalent.
Example:
class Point2 extends Point {
    Point2() {super(2); }
}
class Point3 extends Point {
    Point3() {super(3); }
}
Point2 plane  = new Point2();
Point3 solid  = new Point3();
plane = solid; //Illegal in Java

Parametric Polymorphism
We can create distinct subclasses based on the values passed to constructors. But sometimes we want to create subclasses based on distinct types, and types can’t be passed as parameters. (Types are not values, but rather a property of values.)
We see this problem in Java, which tries to create general purpose data structures by basing them on the class Object. Since any object can be assigned to Object (all classes must be a subclass of Object), this works—at least partially.

class LinkedList {
    Object value;
    LinkedList next;
    Object head() {return value;}
    LinkedList tail(){return next;}
    LinkedList(Object O) {
        value = O; next = null;
    }
    LinkedList(Object O, LinkedList L){
        value = O; next = L;
    }
}
Using this class, we can create a linked list of any subtype of Object. But,
• We can’t guarantee that linked lists are type homogeneous (contain only a single type).
• We must cast Object types back into their “real” types when we extract list values.
We must use wrapper classes like `Integer` rather than `int` because primitive types like `int` aren't objects, and aren't subclass of `Object`.

For example, to use `LinkedList` to build a linked list of `int` s we do the following:

```java
LinkedList l = new LinkedList(new Integer(123));
int i = ((Integer) l.head()).intValue();
```

This is pretty clumsy code. We'd prefer a mechanism that allows us to create a "custom version" of `LinkedList`, based on the type we want the list to contain.

We can't just call something like `LinkedList(int)` or `LinkedList(Integer)` because types can't be passed as parameters. Parametric polymorphism is the solution. Using this mechanism, we can use type parameters to build a "custom version" of a class from a general purpose class.

C++ allows this using its template mechanism. Tiger Java, the newest version of Java, also allows type parameters.

In both languages, type parameters are enclosed in "angle brackets" (e.g., `LinkedList<T>` passes `T`, a type, to the `LinkedList` class).

Thus we have

```java
class LinkedList<T> {
    T value; LinkedList<T> next;
    T head() { return value; }
    LinkedList<T> tail() {
        return next;
    }
    LinkedList(T o) {
        value = o; next = null;
    }
    LinkedList(T o, LinkedList<T> l) {
        value = o; next = l;
    }
}
LinkedList<int> l = new LinkedList(123);
int i = l.head();
```

**Overloading and Ad-hoc Polymorphism**

Classes usually allow overloading of method names, if only to support multiple constructors.

That is, more than one method definition with the same name is allowed within a class, as long as the method definitions differ in the number and/or types of the parameters they take.

For example,

```java
class MyClass {
    int f(int i) { ... }
    int f(float g) { ... }
    int f(int i, int j) { ... }
}
```
Overloading is sometimes called “ad hoc” polymorphism, because, to the programmer, it appears that one method can take a variety of different parameter types. This isn’t true polymorphism because the methods have different bodies; there is no sharing of one definition among different parameter types. There is no guarantee that the different definitions do the same thing, even though they share a common name.

Issues in Overloading

Though many languages allow overloading, few allow overloaded methods to differ only on their result types. (Neither C++ nor Java allow this kind of overloading, though Ada does). For example,

```cpp
class MyClass {
  int f() { ... }
  float f() { ... }
}
```

is illegal. This is unfortunate, since methods with the same name and parameters, but different result types, could be used to automatically convert result values to the type demanded by the context of call.

Why is this form of overloading usually disallowed?

It’s because overload resolution (deciding which definition to use) becomes much harder. Consider:

```cpp
class MyClass {
  int f(int i, int j) { ... }
  float f(float i, float j) { ... }
  float f(int i, int j) { ... }
}
in
int a = f( f(1,2), f(3,4) );
```

which definitions of f do we use in each of the three calls? Getting the correctly answer can be tricky, though solution algorithms do exist.

Operator Overloading

Some languages, like C++ and C#, allow operators to be overloaded. You may add new definitions to existing operators, and use them on your own types. For example,

```cpp
class MyClass {
  int i;
  public:
    int operator+(int j) {
      return i+j;
    }
}
```

```cpp
MyClass c;
int i = c+10;
int j = c.operator+(10);
int k = 10+c; // Illegal!
```
The expression $10+c$ is illegal because there is no definition of $+$ for the types `int` and `MyClass&`. We can create one by using C++'s friend mechanism to insert a definition into MyClass that will have access to MyClass's private data:

```cpp
class MyClass {
    int i;
    public:
        int operator+(int j) {
            return i+j; }
        friend int operator+(int j, MyClass& v) {
            return j+v.i; }
    }
MyClass c;
int k = 10+c; // Now OK!
```

C++ limits operator overloading to existing predefined operators. A few languages, like Algol 68 (a successor to Algol 60, developed in 1968), allow programmers to define brand new operators.

In addition to defining the operator itself, it is also necessary to specify the operator's precedence (which operator is to be applied first) and its associativity (does the operator associate from left to right, or right to left, or not at all). Given this extra detail, it is possible to specify something like

```cpp
op +++ prec = 8;
int op +++(int& i, int& j) {
    return (i++)+(j++); }
```

(Why is `int&` used as the parameter type rather than `int`?)

---

**Parameter Binding**

Almost all programming languages have some notion of binding an actual parameter (provided at the point of call) to a formal parameter (used in the body of a subprogram).

There are many different, and inequivalent, methods of parameter binding. Exactly which is used depends upon the programming language in question.

Parameter Binding Modes include:

- **Value**: The formal parameter represents a local variable initialized to the value of the corresponding actual parameter.
- **Result**: The formal parameter represents a local variable. Its final value, at the point of return, is copied into the corresponding actual parameter.
- **Value/Result**: A combination of the value and results modes. The formal parameter is a local variable initialized to the value of the corresponding actual parameter. The formal's final value, at the point of return, is copied into the corresponding actual parameter.
- **Reference**: The formal parameter is a pointer to the corresponding actual parameter. All references to the formal parameter indirectly access the corresponding actual parameter through the pointer.
What Parameter Modes do Programming Languages Use?

- C: Value mode except for arrays which pass a pointer to the start of the array.
- C++: Allows reference as well as value modes. E.g.,
  ```c++
  int f(int a, int & b)
  ```
- C#: Allows result (out) as well as reference and value modes. E.g.,
  ```csharp
  int g(int a, out int b)
  ```
- Java: Scalar types (int, float, char, etc.) are passed by value; objects are passed by reference (references to objects are passed by value).
- Fortran: Reference (even for constants!)
- Ada: Value/result, reference, and readonly are used.

Why are there so Many Different Parameter Modes?

Parameter modes reflect different views on how parameters are to be accessed, as well as different degrees of efficiency in accessing and using parameters.

- Call by value protects the actual parameter value. No matter what the subprogram does, the parameter can’t be changed.
- Call by reference allows immediate updates to the actual parameter.
- Call by readonly protects the actual parameter and emphasizes the “constant” nature of the formal parameter.

Example

```c
void p(value int a, 
       reference int b, 
       name int c) { 
  a=1; b=2; print(c)
}
int i=3, j=3, k[10][10];
p(i, j, k[i][j]);
```

What element of \( k \) is printed?

- The assignment to \( a \) does not affect \( i \), since \( a \) is a value parameter.
- The assignment to \( b \) does affect \( j \), since \( b \) is a reference parameter.
- \( c \) is a name parameter, so it is evaluated whenever it is used. In the print statement \( k[i][j] \) is printed. At that point \( i=3 \) and \( j=2 \), so \( k[3][2] \) is printed.
Call by value/result allows actual parameters to change, but treats a call as a single step (assign parameter values, execute the subprogram's body, update parameter values).

Call by name delays evaluation of an actual parameter until it is actually needed (which may be never).

Call by Name

Call by name is a special kind of parameter passing mode. It allows some calls to complete that otherwise would fail. Consider

\( f(i, j/0) \)

Normally, when \( j/0 \) is evaluated, a divide fault terminates execution. If \( j/0 \) is passed by name, the division is delayed until the parameter is needed, which may be never.

Call by name also allows programmers to create some interesting solutions to hard programming problems.

Consider the conditional expression found in C, C++, and Java:

\( (\text{cond} ? \text{value1} : \text{value2}) \)

What if we want to implement this as a function call:

```c
condExpr(cond, value1, value2) {
    if (cond)
        return value1;
    else return value2;
}
```

With most parameter passing modes this implementation won't work! (Why?)

But if \text{value1} and \text{value2} are passed by name, the implementation is correct.

Call by Name and Lazy Evaluation

Call by name has much of the flavor of lazy evaluation. With lazy evaluation, you don’t compute a value but rather a suspension—a function that will provide a value when called.

This can be useful when we need to control how much of a computation is actually performed.

Consider an infinite list of integers. Mathematically it is represented as

\[ 1, 2, 3, \ldots \]

How do we compute a data structure that represents an infinite list?
The obvious computation

```java
infList(int start) {
    return list(start, infList(start+1));
}
```
doesn't work. (Why?)

A less obvious implementation, using suspensions, does work:

```java
infList(int start) {
    return list(start, function() {
        return infList(start+1);
    });
}
```

Now, whenever we are given an infinite list, we get two things: the first integer in the list and a suspension function. When called, this function will give you the rest of the infinite list (again, one more

---

Eager Parameter Evaluation

Sometimes we want parameters evaluated eagerly—as soon as they are known.

Consider a sorting routine that breaks an array in half, sorts each half, and then merges together the two sorted halves (this is a merge sort).

In outline form it is:

```java
sort(inputArray) {
    ...
    merge(sort(leftHalf(inputArray)), sort(rightHalf(inputArray)));
}
```

This definition lends itself nicely to parallel evaluation: The two halves of an input array can be sorted in parallel. Each of these two halves can again be split in two, allowing parallel sorting of four quarter-sized arrays, then leading to 8 sorts of 1/8 sized arrays, etc.

But, to make this all work, the two parameters to merge must be evaluated eagerly, rather than in sequence.
Type Equivalence

Programming languages use types to describe the values a data object may hold and the operations that may be performed.

By checking the types of values, potential errors in expressions, assignments and calls may be automatically detected. For example, type checking tells us that

\[ 123 + "123" \]

is illegal because addition is not defined for an integer, string combination.

Type checking is usually done at compile-time; this is static typing.

Type checking may also be done at run-time; this is dynamic typing.

A program is type-safe if it is impossible to apply an operation to a value of the wrong type. In a type-safe language, plus is never told to add an integer to a string, because its definition does not allow that combination of operands. In type-safe programs an operator can still see an illegal value (e.g., a division by zero), but it can’t see operands of the wrong type.

A strongly-typed programming language forbids the execution of type-unsafe programs.

Weakly-typed programming languages allow the execution of potentially type-unsafe programs.

The question reduces to whether the programming language allows programmers to “break” the type rules, either knowingly or unknowingly.

Java is strongly typed; type errors preclude execution. C and C++ are weakly typed; you can break the rules if you wish. For example:

```c
int i; int* p;
p = (int *) i * i;
```

Now \( p \) may be used as an integer pointer though multiplication need not produce valid integer pointers.

If we are going to do type checking in a program, we must decide whether two types, \( T_1 \) and \( T_2 \) are equivalent; that is, whether they be used interchangeably.

There are two major approaches to type equivalence:

Name Equivalence:

Two types are equivalent if and only if they refer to exactly the same type declaration.

For example,

```c
type PackerSalaries = int[100];
type AssemblySizes = int[100];
PackerSalaries salary;
AssemblySizes size;
```
Is
\[ \text{sal} = \text{size}; \]
allowed?
Using name equivalence, no. That is, salary \( \neq_N \) size since these two variables have different type declarations (that happen to be identical in structure).
Formally, we define \( \equiv_N \) (name type equivalence) as:
(a) \( T \equiv_N T \)
(b) Given the declaration
\[
\text{Type } T_1 = T_2; \\
T_1 \equiv_N T_2
\]

We treat anonymous types (types not given a name) as an abbreviation for an implicit declaration of a new and unique type name.
Thus
\[
\text{int } A[10];
\]
is an abbreviation for
\[
\text{Type } T_{\text{new}} = \text{int}[10]; \\
T_{\text{new}} \ A;
\]

**Structural Equivalence**

An alternative notion of type equivalence is structural equivalence (denoted \( \equiv_S \)). Roughly, two types are structurally equivalent if the two types have the same definition, independent of where the definitions are located. That is, the two types have the same definitional structure.

Formally,
(a) \( T \equiv_S T \)
(b) Given the declaration
\[
\text{Type } T = Q; \\
T \equiv_S Q
\]
(c) If \( T \) and \( Q \) are defined using the same type constructor and corresponding parameters in the two definitions are equal or structurally equivalent then \( T \equiv_S Q \)
Returning to our previous example,

```plaintext
type PackerSalaries = int[100];
type AssemblySizes = int[100];
PackerSalaries salary;
AssemblySizes size;
```

\(\text{salary} \equiv_s \text{size}\) since both are arrays and \(100 = 100\) and \(\text{int} \equiv_s \text{int}\).

**Which notion of Equivalence do Programming Languages Use?**

C and C++ use structural equivalence except for structs and classes (where name equivalence is used). For arrays, size is ignored.

Java uses structural equivalence for scalars. For arrays, it requires name equivalence for the component type, ignoring size. For classes it uses name equivalence except that a subtype may be used where a parent type is expected. Thus given

```plaintext
void subr(Object O) { ... };
```

the call

```plaintext
subr(new Integer(100));
```

is OK since `Integer` is a subclass of `Object`.

**Automatic Type Conversions**

C, C++ and Java also allow various kinds of automatic type conversions.

In C, C++ and Java, a float will be automatically created from an int:

```plaintext
float f = 10; // No type error
```

Also, an integer type (char, short, int, long) will be widened:

```plaintext
int i = 'x';
```

In C and C++ (but not Java), an integer value can also be narrowed, possibly with the loss of significant bits:

```plaintext
char c = 1000000;
```

**Reading Assignment**

- Roosta: Sections 13.1, 13.2
- The Scheme Language Definition (linked from class web page)
Lisp & Scheme

Lisp (List Processing Language) is one of the oldest programming languages still in wide use. It was developed in the late 50s and early 60s by John McCarthy. Its innovations include:
• Support of symbolic computations.
• A functional programming style without emphasis on assignments and side-effects.
• A naturally recursive programming style.
• Dynamic (run-time) type checking.
• Dynamic data structures (lists, binary trees) that grow without limit.
• Automatic garbage collection to manage memory.
• Functions are treated as “first class” values; they may be passed as arguments, returned as result values, stored in data structures, and created during execution.
• A formal semantics (written in Lisp) that defines the meaning of all valid programs.
• An Integrated Programming Environment to create, edit and test Lisp programs.

Scheme

Scheme is a recent dialect of Lisp. It uses lexical (static) scoping. It supports true first-class functions. It provides program-level access to control flow via continuation functions.

Atomic (Primitive) Data Types

Symbols:
Essentially the same form as identifiers. Similar to enumeration values in C and C++.
Very flexible in structure; essentially any sequence of printable characters is allowed; anything that starts a valid number (except + or -) may not start a symbol.
Valid symbols include:
  abc  hello-world   +  <=!

Integers:
Any sequence of digits, optionally prefixed with a + or -. Usually unlimited in length.
Reals:
A floating point number in a decimal format (123.456) or in exponential format (1.23e45). A leading sign and a signed exponent are allowed (-12.3, 10.0e-20).

Rationals:
Rational numbers of the form integer/integer (e.g., 1/3 or 9/7) with an optional leading sign (-1/2, +7/8).

Complex:
Complex numbers of the form num+num i or num-num i, where num is an integer or real number. Example include 1+3i, -1.5–2.5i, 0+1i).

String:
A sequence of characters delimited by double quotes. Double quotes and backslashes must be escaped using a backslash. For example
"Hello World" "\Wow!"

Character:
A single character prefixed by \\#. For example, \a, \0, \\, \#. Two special characters are \space and \newline.

Boolean:
True is represented as \t and false is represented as \f.

Binary Trees

Binary trees are also called S-Expressions in Lisp and Scheme.
They are of the form
( item . item )
where item is any atomic value or any S-Expression. For example:
( A . B )
(1.2 . "xyz")
((A . B) . C)
(A . (B . C))

S-Expressions are linearizations of binary trees:

```
      C
 A      B
      1.2 "xyz"
```

S-Expressions are built and accessed using the predefined functions cons, car and cdr.

cons builds a new S-Expression from two S-Expressions that represent the left and right children.

\cons(E1,E2) = (E1 . E2)
car returns are left subtree of an S-Expression.

\car (E1 . E2) = E1
cdr returns are right subtree of an S-Expression.

cdr (E1 . E2) = E2
Lists

In Lisp and Scheme lists are a special, widely-used form of S-Expressions.

() represents the empty or null list

(A) represents the list containing A.
By definition, (A) ≡ (A . () )

(A B) represents the list containing A and B.
By definition,

(A B) ≡ (A . (B . () ) )

In general, (A B C ... Z) ≡

(A . (B . (C . ... (Z . () ) ... )))

(A B C )≡

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>()</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Function Calls

In List and Scheme, function calls are represented as lists.

(A B C) means:
Evaluate A (to a function)
Evaluate B and C (as parameters)
Call A with B and C as its parameters
Then use the value returned by the call as the “meaning” of (A B C).

cons, car and cdr are predefined symbols bound to built-in functions that build and access lists and S-Expressions.

Literals (of type integer, real, rational, complex, string, character and boolean) evaluate to themselves.

For example (⇒ means “evaluates to”)

(cons 1 2) ⇒ (1 . 2)
(cons 1 () ) ⇒ (1)
(car (cons 1 2)) ⇒ 1
(cdr (cons 1 ())) ⇒ ()

But,

(car (1 2)) fails during execution!
Why?
The expression (1 2) looks like a call, but 1 isn’t a function! We need some way to “quote” symbols and lists we don’t want evaluated.

(quote arg)
is a special function that returns its argument unevaluated.

Thus (quote (1 2)) doesn’t try to evaluate the list (1 2); it just returns it.

Since quotation is so often used, it may be abbreviated using a single quote. That is

(quote arg) ≡ 'arg

Thus

(car ' (a b c)) ⇒ a
(cdr ' ( (A) (B) (C)) ) ⇒
( (B) (C) )
(cons 'a '1) ⇒ (a . 1)
But,

('cdr ' (A B)) fails!
Why?
User-defined Functions

The list

(lambda (args) (body))

evaluates to a function with (args) as its argument list and (body) as
the function body.
No quotes are needed for (args) or (body).
Thus

(llambda (x) (+ x 1)) evaluates
to the increment function.
Similarly,

((lambda (x) (+ x 1)) 10) \Rightarrow 11

Since functions are frequently
defined, we may abbreviate
(define id
  (lambda (args) (body)) )
as
(define (id args) (body) )
Thus

(define (plus1 x) (+ x 1))
For example,

\[
(\text{if } (= 1 (+ 0 1))
\begin{array}{l}
\quad \text{'Yes} \\
\quad \text{'No}
\end{array}
\)
\]

\[
(\text{define}
\begin{array}{c}
(\text{fact } n) \\
(\text{if } (= n 0) \\
\quad 1 \\
\quad (* n (\text{fact } (- n 1)))
\end{array}
\)
\]

### Generalized Conditional

This is similar to a switch or case:

\[
(\text{cond }
\begin{array}{c}
(p_1 \ e_1) \\
(p_2 \ e_2) \\
\ldots \\
(\text{else } e_n)
\end{array})
\]

Each of the predicates \((p_1, p_2, \ldots)\) is evaluated until one is true \((\neq \text{#f})\). Then the corresponding expression \((e_1, e_2, \ldots)\) is evaluated and returned as the value of the \text{cond}. \text{else} acts like a predicate that is always true.

Example:

\[
(\text{cond}
\begin{array}{c}
((= a 1) \ 2) \\
((= a 2) \ 3) \\
(\text{else } \ 4)
\end{array})
\]

### Recursion in Scheme

Recursion is widely used in Scheme and most other functional programming languages.

Rather than using a loop to step through the elements of a list or array, recursion breaks a problem on a large data structure into a simpler problem on a smaller data structure.

A good example of this approach is the \text{append} function, which joins (or appends) two lists into one larger list containing all the elements of the two input lists (in the correct order).

Note that \text{cons} is not \text{append}. \text{cons} adds one element to the head of an existing list.

Thus

\[
(\text{cons } '(a b) \ '(c d)) \Rightarrow
\quad ((a b) \ c \ d)
\]

\[
(\text{append } '(a b) \ '(c d)) \Rightarrow
\quad (a \ b \ c \ d)
\]

The \text{append} function is predefined in Scheme, as are many other useful list-manipulating functions (consult the Scheme definition for what's available).

It is instructive to define \text{append} directly to see its recursive approach:

\[
(\text{define}
\begin{array}{c}
(\text{append } L_1 \ L_2) \\
(\text{if } (\text{null? } L_1) \\
\quad L_2 \\
\quad (\text{cons } (\text{car } L_1) \\
\quad (\text{append } (\text{cdr } L_1) \ L_2))
\end{array})
\)
\]
Let's trace \((\text{append } '(a \ b) \ '(c \ d))\)

Our definition is

\[
\text{(define (append L1 L2)} \\
\text{\quad (if (null? L1) \ L2)} \\
\text{\quad \quad (cons (car L1) \ (append (cdr L1) \ L2))}}
\]

Now \(L1 = (a \ b)\) and \(L2 = (c \ d)\).

\((\text{null? } L1)\) is false, so we evaluate

\((\text{cons (car L1) (append (cdr L1) \ L2)}) = (\text{cons 'a (append '(b) \ '(c \ d))}) = (\text{cons 'a (append '(b) \ '(c \ d))})\)

We need to evaluate

\((\text{append 'b (append 'c (append 'd))})\)

In this call, \(L1 = (b)\) and \(L2 = (c \ d)\).

\(L1\) is not null, so we evaluate

\((\text{cons (car L1) (append (cdr L1) \ L2)}) = (\text{cons 'b (append 'c (append 'd))})\)

As an example, consider

\((\text{rev '(1 2 3)}) = (3 2 1)\)

The definition of \text{rev} is straightforward:

\[
\text{(define (rev L)} \\
\text{\quad (if (null? L) \ L)} \\
\text{\quad \quad (append \ (rev \ (cdr L))) \ (list \ (car L))}}
\]

Reversing a List

Another useful list-manipulation function is \text{rev}, which reverses the members of a list. That is, the last element becomes the first element, the next-to-last element becomes the second element, etc.

For example,

\((\text{rev '}(1 \ 2)) = (2 \ 1)\)

We need to evaluate

\((\text{rev '}(1 \ 2))\)

In this call, \(L = (1 \ 2)\).

\(L\) is not null so we evaluate

\((\text{append (rev (cdr L)) \ (list (car L)))}) = \)

As an example, consider

\((\text{rev '}(1 \ 2))\)

Here \(L = (1 \ 2)\). \(L\) is not null so we evaluate

\((\text{append (rev (cdr L)) \ (list (car L)))}) = \)

We must evaluate \((\text{rev '}(2))\)

Here \(L = (2)\). \(L\) is not null so we evaluate

\((\text{append (rev (cdr L)) \ (list (car L)))}) = \)

We must evaluate \((\text{rev '}(2))\)

Here \(L = (2)\). \(L\) is not null so \((\text{rev '}(2)) = ()\)
Thus \((\text{append} \ (\text{rev} \ ())\ '(2))\) = \\
(\text{append} \ () \ '(2)) = (2) = \(\text{rev} \ '(2))\)

Finally, recall \(\text{rev} \ '(1\ 2))\) =
(\text{append} \ (\text{rev} \ '(2)) \ '(1)) = \\
(\text{append} \ '(2) \ '(1)) = (2\ 1)

As constructed, \(\text{rev}\) only reverses the “top level” elements of a list. That is, members of a list that themselves are lists aren’t reversed.

For example,
\(\text{rev} \ '\ (\ 1\ 2\ )\ (3\ 4)\) = \\
((3\ 4)\ (1\ 2))

We can generalize \(\text{rev}\) to also reverse list members that happen to be lists.

To do this, it will be convenient to use Scheme’s let construct.

The Let Construct

Scheme allows us to create local names, bound to values, for use in an expression.

The structure is
(\text{let} \ (\ (id1\ \ val1)\ (id2\ \ val2)\ ... \ ) \ nam
expr )

In this construct, \(val1\) is evaluated and bound to \(id1\), which will exist only within this let expression. If \(id1\) is already defined (as a global or parameter name) the existing definition is hidden and the local definition, bound to \(val1\), is used. Then \(val2\) is evaluated and bound to \(id2\), .... Finally, \(expr\) is evaluated in a scope that includes \(id1, id2, ...

For example,
(\text{let} \ (\ (a\ 10)\ (b\ 20)\ ) \ nam
(+ a b)) \Rightarrow 30

Using a let, the definition of \(\text{revall}\), a version of \(\text{rev}\) that reverses all levels of a list, is easy:

\[
\begin{align*}
\text{(define} \ (\text{revall} \ L) \\
(\text{if} \ (\text{null?} \ L) \ nam
L \\
(\text{let} \ ((E \ (\text{if} \ (\text{list?} \ (\text{car} \ L)) \ nam
(\text{revall} \ (\text{car} \ L)) \ nam
(\text{car} \ L) \ ))) \ nam
(\text{append} \ (\text{revall} \ (\text{cdr} \ L)) \ nam
(\text{list} \ E))
\)) \ nam
\)) \ nam
(\text{revall} \ '\ (\ 1\ 2\ )\ (3\ 4)\)) \Rightarrow \\
((3\ 4)\ (2\ 1))
\end{align*}
\]

Subsets

Another good example of Scheme’s recursive style of programming is subset computation.

Given a list of distinct atoms, we want to compute a list of all subsets of the list values.

For example,

(\text{subsets} \ '\ (1\ 2\ 3)) \Rightarrow \\
(\ ()\ (1)\ (2)\ (3)\ (1\ 2)\ (1\ 3)\ (2\ 3)\ (1\ 2\ 3))

The order of atoms and sublists is unimportant, but all possible subsets of the list values must be included.

Given Scheme’s recursive style of programming, we need a recursive definition of subsets.
That is, if we have a list of all subsets of \( n \) atoms, how do we extend this list to one containing all subsets of \( n+1 \) values?

First, we note that the number of subsets of \( n+1 \) values is exactly twice the number of subsets of \( n \) values.

For example,

\[
\text{(subsets '(1 2)) } \Rightarrow ( () (1) (2) (1 2)), \text{ which contains 4 subsets.}
\]

\[
\text{(subsets '(1 2 3)) contains 8 subsets (as we saw earlier).}
\]

Moreover, the extended list (of subsets for \( n+1 \) values) is simply the list of subsets for \( n \) values plus the result of “distributing” the new value into each of the original subsets.

Thus

\[
\text{(subsets '(1 2 3)) } \Rightarrow ( () (1) (2) (1 2) (1 3) (2 3) (1 2 3)) =
\]

\[
( () (1) (2) (1 2) ) \text{ plus (3) (1 3) (2 3) (1 2 3) }
\]

This insight leads to a concise program for subsets.

We will let \( \text{distrib L E} \) be a function that “distributes” \( E \) into each list in \( L \).

For example,

\[
\text{(distrib '((()) (1) (2) (1 2))) 3) = ( (3) (3 1) (3 2) (3 1 2) )}
\]

\[
\text{(define (distrib L E)}
\]

\[
\text{(if (null? L)}
\]

\[
\text{()}
\]

\[
\text{(cons (cons E (car L))}
\]

\[
\text{(distrib (cdr L) E))}
\]

\[
\text{)}
\]

\[
\text{)
}\]

We will let \( \text{extend L E} \) extend a list \( L \) by distributing element \( E \) through \( L \) and then appending this result to \( L \).

For example,

\[
\text{(extend '(() (a) ) 'b) } \Rightarrow ( () (a) (b) (b a))
\]

\[
\text{(define (extend L E)}
\]

\[
\text{(append L (distrib L E))}
\]

\[
\text{)}
\]

Now subsets is easy:

\[
\text{(define (subsets L)}
\]

\[
\text{(if (null? L)}
\]

\[
\text{(list ()])}
\]

\[
\text{(extend (subsets (cdr L))}
\]

\[
\text{(car L))}
\]

\[
\text{)}
\]

\[
\text{)}
\]

\[
\text{)}
\]

\[
\text{)}
\]

\[
\text{\text{Data Structures in Scheme}}
\]

In Scheme, lists and S-expressions are basic. Arrays can be simulated using lists, but access to elements “deep” in the list can be slow (since a list is a linked structure).

To access an element deep within a list we can use:

- \( \text{(list-tail L k)} \)
  
  This returns list \( L \) after removing the first \( k \) elements. For example,

  \[
  \text{(list-tail ' (1 2 3 4 5) 2) } \Rightarrow (3 4 5)
  \]

- \( \text{(list-ref L k)} \)
  
  This returns the \( k \)-th element in \( L \) (counting from 0). For example,

  \[
  \text{(list-ref ' (1 2 3 4 5) 2) } \Rightarrow 3
  \]
Vectors in Scheme

Scheme provides a vector type that directly implements one dimensional arrays.

Literals are of the form #( ... )

For example, #(1 2 3) or #(1 2.0 "three")

The function (vector? val) tests whether val is a vector or not.

(vector? 'abc) ⇒ #f
(vector? '(a b c)) ⇒ #f
(vector? #(a b c)) ⇒ #t

The function (vector v1 v2 ...) evaluates v1, v2, ... and puts them into a vector.

(vector 1 2 3) ⇒ #(1 2 3)

The function (make-vector k val) creates a vector composed of k copies of val. Thus

(make-vector 4 (/ 1 2)) ⇒ #(1/2 1/2 1/2 1/2)

The function (vector-ref vect k) returns the k-th element of vect, starting at position 0. It is essentially the same as vect[k] in C or Java. For example,

(vector-ref #(2 4 6 8 10) 3) ⇒ 8

The function (vector-set! vect k val) sets the k-th element of vect, starting at position 0, to be val. It is essentially the same as vect[k]=val in C or Java. The value returned by the function is unspecified. The suffix “!” indicates that the function has a side-effect. For example,

(define v #(1 2 3 4 5))
(vector-set! v 2 0)
v ⇒ #(1 2 0 4 5)

Vectors aren’t lists (and lists aren’t vectors).

Thus (car #(1 2 3)) doesn’t work.

There are conversion routines:

• (vector->list v) converts vector v to a list containing the same values as v. For example,
  (vector->list #(1 2 3)) ⇒ (1 2 3)

• (list->vector L) converts list L to a vector containing the same values as L. For example,
  (list->vector '(1 2 3)) ⇒ #(1 2 3)

In general Scheme names a conversion function from type T to type Q as T->Q. For example,

string->list converts a string into a list containing the characters in the string.
Records and Structs

In Scheme we can represent a record, struct, or class object as an association list of the form
(((obj1 val1) (obj2 val2) ...)

In the association list, which is a list of (object value) sublists, object serves as a “key” to locate the desired sublist.

For example, the association list
((A 10) (B 20) (C 30))

serves the same role as

```
struct
  { int a = 10;
  int b = 20;
  int c = 30;}
```

The predefined Scheme function
(assoc obj alist)

checks alist (an association list) to see if it contains a sublist with obj as its head. If it does, the list starting with obj is returned; otherwise #f (indicating failure) is returned.

For example,
(define L
  '( (a 10) (b 20) (c 30) )
(assoc 'a L) ⇒ (a 10)
(assoc 'b L) ⇒ (b 20)
(assoc 'x L) ⇒ #f

We can use non-atomic objects as keys too!
(define price-list
  ' ( ((bmw m5)  71095)
      ((bmw z4)  40495)
      ((jag xj8) 56975)
      ((mb sl500) 86655)
  )
(assoc '(bmw z4) price-list) ⇒ ((bmw z4) 40495)

Using assoc, we can easily define a structure function:
(structure key alist) will return the value associated with key in alist; in C or Java notation, it returns alist.key.
(define
  (structure key alist)
    (if (assoc key alist)
        (car (cdr (assoc key alist)))
        #f
    )
)

We can improve this function in two ways:
- The same call to assoc is made twice; we can save the value computed by using a let expression.
- Often combinations of car and cdr are needed to extract a value. Scheme
has a number of predefined functions that combine several calls to \texttt{car} and \texttt{cdr} into one function. For example,
\begin{align*}
\text{(caar x)} & \equiv (\text{car} (\text{car} x)) \\
\text{(cadr x)} & \equiv (\text{car} (\text{cdr} x)) \\
\text{(cdar x)} & \equiv (\text{cdr} (\text{car} x)) \\
\text{(cddr x)} & \equiv (\text{cdr} (\text{cdr} x))
\end{align*}

Using these two insights we can now define a better version of \texttt{structure}
\begin{verbatim}
(define (structure key alist)
  (let ((p (assoc key alist)))
    (if p
      (cadr p)
      #f))
)
\end{verbatim}

What does \texttt{assoc} do if more than one sublist with the same key exists? It returns the first sublist with a matching key. In fact, this property can be used to make a simple and fast function that updates association lists:
\begin{verbatim}
(define (set-structure key alist val)
  (cons (list key val) alist))
\end{verbatim}

If we want to be more space-efficient, we can create a version that updates the internal structure of an association list, using \texttt{set-cdr!} which changes the \texttt{cdr} value of a list:
\begin{verbatim}
(define (set-structure! key alist val)
  (let ((p (assoc key alist)))
    (if p
      (begin
        (set-cdr! p (list val))
        alist)
      (cons (list key val) alist)))
)
\end{verbatim}

\textbf{Functions are First-class Objects}

Functions may be passed as parameters, returned as the value of a function call, stored in data objects, etc.

This is a consequence of the fact that \((\text{lambda} (\text{args}) (\text{body}))\) evaluates to a function just as \((+ 1 1)\) evaluates to an integer.
**Scoping**

In Scheme scoping is static (lexical). This means that non-local identifiers are bound to containing lambda parameters, or let values, or globally defined values. For example,

\[
\begin{align*}
&(\text{define} \ (f \ x) \\
& \quad (\lambda (y) (+ x y)))
\end{align*}
\]

Function \( f \) takes one parameter, \( x \). It returns a function (of \( y \)), with \( x \) in the returned function bound to the value of \( x \) used when \( f \) was called.

Thus

\[
(f \ 10) \equiv (\lambda (y) (+ 10 y))
\]

\[
((f \ 10) \ 12) \Rightarrow 22
\]

Unbound symbols are assumed to be globals; there is a run-time error if an unbound global is referenced. For example,

\[
\begin{align*}
&(\text{define} \ (p \ y) (+ x y)) \\
&(p \ 20) ; \text{error} -- x \text{ is unbound}
\end{align*}
\]

We can use let bindings to create private local variables for functions:

\[
\begin{align*}
&(\text{define} \ F \\
& \quad (\text{let} \ ( (X \ 1) ) \\
& \quad \quad (\lambda () X) \\
& \quad )
\end{align*}
\]

\( F \) is a function (of no arguments).

\( (F) \) calls \( F \).

\[
\begin{align*}
&(\text{define} \ x \ 10) \\
&(p \ 20) \Rightarrow 30
\end{align*}
\]

Let Bindings can be Subtle

You must check to see if the let-bound value is created when the function is created or when it is called.

Compare

\[
\begin{align*}
&(\text{define} \ cnt \\
& \quad (\text{let} \ ( (I \ 0) ) \\
& \quad \quad (\lambda () \ \\
& \quad \quad \quad (\text{set!} \ I (+ I 1)) \ I) \\
& \quad )
\end{align*}
\]

Now,

\[
\begin{align*}
&(\text{cnt}) \Rightarrow 1 \\
&(\text{cnt}) \Rightarrow 2 \\
&(\text{cnt}) \Rightarrow 3
\end{align*}
\]

etc.

We can encapsulate internal state with a function by using private, let-bound variables:

\[
\begin{align*}
&(\text{define} \ cnt \\
& \quad (\text{let} \ ( (I \ 0) ) \\
& \quad \quad (\lambda () \ \\
& \quad \quad \quad (\text{set!} \ I (+ I 1)) \ I)
\end{align*}
\]

VS.

\[
\begin{align*}
&(\text{define} \ reset \\
& \quad (\lambda () \ \\
& \quad \quad (\text{let} \ ( (I \ 0) ) \\
& \quad \quad \quad (\text{set!} \ I (+ I 1)) \ I)
\end{align*}
\]

\( (\text{reset}) \Rightarrow 1, (\text{reset}) \Rightarrow 1, \text{etc.} \)
Simulating Class Objects

Using association lists and private bound values, we can encapsulate data and functions. This gives us the effect of class objects.

\[(\text{define (point } x \ y)\)\]
\[(\text{list}\)\]
\[(\text{list 'rect}\)\]
\[(\text{lambda () (list } x \ y))\)
\[(\text{list 'polar}\)\]
\[(\text{lambda ()}\)\]
\[(\text{list}\)\]
\[(\text{sqrt (+ (* } x \ x) (* } y \ y))\)
\[(\text{atan (/ } x \ y))\)
\[(\text{)}\)
\[(\text{)}\)

A call \[(\text{point } 1 \ 1)\] creates an association list of the form \[(\text{rect funct) (polar funct) )}\]

We can use structure to access components:

\[(\text{define p (point } 1 \ 1) )\]
\[(\text{ (structure 'rect p) ) } \Rightarrow \ (1 \ 1)\]
\[(\text{ (structure 'polar p) ) } \Rightarrow \ (\sqrt{2} \ \frac{\pi}{4})\]

We can add new functionality by just adding new \[(\text{id function) pairs to the association list.}\]

\[(\text{define (point } x \ y)\)\]
\[(\text{list}\)\]
\[(\text{list 'rect}\)\]
\[(\text{lambda () (list } x \ y))\)
\[(\text{list 'polar}\)\]
\[(\text{lambda ()}\)\]
\[(\text{list}\)\]
\[(\text{sqrt (+ (* } x \ x) (* } y \ y))\)
\[(\text{atan (/ } x \ y))\)
\[(\text{)}\)
\[(\text{)}\)
\[(\text{list 'set-rect!}\)\]
\[(\text{lambda (newx newy)}\)
\[(\text{set! } x \text{ newx})\]
\[(\text{set! } y \text{ newy})\]
\[(\text{list } x \ y)\]
\[(\text{)}\)
\[(\text{list 'set-polar!}\)\]
\[(\text{lambda (r theta)}\)
\[(\text{set! } x \text{ (* r (sin theta))})\]
\[(\text{set! } y \text{ (* r (cos theta))})\]
\[(\text{list } r \ \theta)\]
\[(\text{)}\)
\[(\text{)}\]

Now we have

\[(\text{define p (point } 1 \ 1) )\]
\[(\text{ (structure 'rect p) ) } \Rightarrow \ (1 \ 1)\]
\[(\text{ (structure 'polar p) ) } \Rightarrow \ (\sqrt{2} \ \frac{\pi}{4})\]
\[(\text{ (structure 'set-polar! p) } 1 \ \pi/4)\]
\[(\Rightarrow (1 \ \pi/4)\]
\[(\text{ (structure 'rect p) ) } \Rightarrow \]
\[\frac{1}{\sqrt{2}} \ \frac{1}{\sqrt{2}}\]

\[(\text{)}\)
\[(\text{)}\]
Limiting Access to Internal Structure

We can improve upon our association list approach by returning a single function (similar to a C++ or Java object) rather than an explicit list of (id function) pairs. The function will take the name of the desired operation as one of its arguments.

First, let’s differentiate between

(define def1
  (let ((I 0))
    (lambda () (set! I (+ I 1)) I))
)

and

(define (def2)
  (let ((I 0))
    (lambda () (set! I (+ I 1)) I))
)

def1 is a function of zero arguments that increments a local variable and returns its updated value.
def2 is a function (of zero arguments) that generates a function of zero arguments (that increments a local variable and returns its updated value). Each call to def2 creates a different function.

Stack Implemented as a Function

(define (stack)
  (let ((s () ))
    (lambda (op . args) ; var # args
      (cond
        ((equal? op 'push!)
          (set! s (cons (car args) s))
          (car s))
        ((equal? op 'pop!)
          (if (null? s)
            #f
            (let ((top (car s))
                  (set! s (cdr s))
                  top)))
        ((equal? op 'empty?)
          (null? s))
        (else #f))
      )
    ))
)

(define stk (stack)); new empty stack

(stk 'push! 1) ⇒ 1 ; s = (1)
(stk 'push! 3) ⇒ 3 ; s = (3 1)
(stk 'push! 'x) ⇒ x ; s = (x 3 1)
(stk 'pop!) ⇒ x ; s = (x 3 1)
(stk 'empty?) ⇒ #f ; s = (3 1)
(stk 'dump) ⇒ #f ; s = (3 1)
Higher-Order Functions

A higher-order function is a function that takes a function as a parameter or one that returns a function as its result.

A very important (and useful) higher-order function is map, which applies a function to a list of values and produces a list or results:

(define (map f L)
  (if (null? L)
      ()
      (cons (f (car L))
            (map f (cdr L)))))

Note: In Scheme's built-in implementation of map, the order of function application is unspecified.

(map sqrt '(1 2 3 4 5)) ⇒ (1 1.414 1.732 2 2.236)
(map (lambda(x) (* x x)) '(1 2 3 4 5)) ⇒ (1 4 9 16 25)

Map may also be used with multiple argument functions by supplying more than one list of arguments:

(map + '(1 2 3) '(4 5 6)) ⇒ (5 7 9)

The Reduce Function

Another useful higher-order function is reduce, which reduces a list of values to a single value by repeatedly applying a binary function to the list values.

This function takes a binary function, a list of data values, and an identity value for the binary function:

(define (reduce f L id)
  (if (null? L)
      id
      (f (car L)
          (reduce f (cdr L) id))
  )

(reduce append '((1 2 3) (4 5 6) (7 8)) ()) ⇒ (1 2 3 4 5 6 7 8)
(reduce expt '(2 2 2 2) 1) ⇒ \(2^{2^2} = 65536\)

(reduce expt '(2 2 2 2 2) 1) ⇒ \(2^{65536}\)

(string-length (number->string (reduce expt '(2 2 2 2 2 2 2) 1))) ⇒ 19729 ; digits in \(2^{65536}\)
Sharing vs. Copying

In languages without side-effects an object can be copied by copying a pointer (reference) to the object; a complete new copy of the object isn’t needed.

Hence in Scheme (define A B) normally means

![Diagram of A and B]

But, if side-effects are possible we may need to force a physical copy of an object or structure:

```
(define (copy obj)
  (if (pair? obj)
      (cons (copy (car obj))
           (copy (cdr obj)))
      obj)
)
```

For example,

```
(define A '(1 2))
(define B (cons A A))
B = ( (1 2) 1 2)
```

```
(set-car! (car B) 10)
B = ( (10 2) 10 2)
```

```
(define C (cons (copy A) (copy A)))
```

```
(set-car! (car C) 20)
C = ((20 2) 10 2)
```

Similar concerns apply to strings and vectors, because their internal structure can be changed.
**Shallow & Deep Copying**

A copy operation that copies a pointer (or reference) rather than the object itself is a shallow copy. For example, in Java,

```java
Object O1 = new Object();
Object O2 = new Object();
O1 = O2; // shallow copy
```

If the structure within an object is physically copied, the operation is a deep copy.

In Java, for objects that support the clone operation,

```java
O1 = O2.clone(); // deep copy
```

Even in Java’s deep copy (via the clone() operation), objects referenced from within an object are shallow copied. Thus given

```java
class List {
    int value;
    List next;
}
List L, M;
M = L.clone();
```

L.value and M.value are independent, but L.next and M.next refer to the same List object.

A complete deep copy, that copies all objects linked directly or indirectly, is expensive and tricky to implement. (Consider a complete copy of a circular linked list).

---

**Equality Checking in Scheme**

In Scheme = is used to test for numeric equality (including comparison of different numeric types). Non-numeric arguments cause a run-time error. Thus

- `(= 1 1) ⇒ #t`
- `(= 1 1.0) ⇒ #t`
- `(= 1 2/2) ⇒ #t`
- `(= 1 1+0.0i) ⇒ #t`

To compare non-numeric values, we can use either:

- **pointer equivalence** (do the two operands point to the same address in memory)
- **structural equivalence** (do the two operands point to structures with the same size, shape and components, even if they are in different memory locations)

In general pointer equivalence is faster but less accurate.
Scheme implements both kinds of equivalence tests.

(eqv? obj1 obj2)
This tests if obj1 and obj2 are the exact same object. This works for atoms and pointers to the same structure.

(equal? obj1 obj2)
This tests if obj1 and obj2 are the same, component by component. This works for atoms, lists, vectors and strings.

(eqv? 1 1) ⇒ #t
(eqv? 1 (+ 0 1)) ⇒ #t
(eqv? 1/2 (- 1 1/2)) ⇒ #t
(eqv? (cons 1 2) (cons 1 2)) ⇒ #f
(eqv? "abc" "abc") ⇒ #f

In general it is wise to use equal? unless speed is a critical factor.

I/O in Scheme

Scheme provides simple read and write functions, directed to the “standard in” and “standard out” files.

(read)
Reads a single Scheme object (an atom, string, vector or list) from the standard in file. No quoting is needed.

(write obj)
Writes a single object, obj, to the standard out file.

(display obj)
Writes obj to the standard out file in a more readable format. (Strings aren’t quoted, and characters aren’t escaped.)

(newline)
Forces a new line on standard out file.

Ports

Ports are Scheme objects that interface with systems files. I/O to files is normally done through a port object bound to a system file.

(open-input-file "path to file")
This returns an input port associated with the "path to file" string (which is system dependent). A runtime error is signalled if "path to file" specifies an illegal or inaccessible file.

(read port)
Reads a single Scheme object (an atom, string, vector or list) from port, which must be an input port object.
(eof-object? obj)
When the end of an input file is reached, a special eof-object is returned. (eof-object? obj) tests whether an object is this special end-of-file marker.

(open-output-file "path to file")
This returns an output port associated with the "path to file" string (which is system dependent). A run-time error is signalled if "path to file" specifies an illegal or inaccessible file.

(write obj port)
 Writes a single object, obj, to the output port specified by port.

(display obj port)
 Writes obj to the output port specified by port. display uses a more readable format than write does. (Strings aren't quoted, and characters aren't escaped.)

(close-input-port port)
This closes the input port specified by port.

(close-output-port port)
This closes the output port specified by port.

Example—Reading & Echoing a File
We will iterate through a file, reading and echoing its contents. We need a good way to do iteration; recursion is neither natural nor efficient here.

Scheme provides a nice generalization of the let expression that is similar to C’s for loop.

(let X ( (id1 val1) (id2 val2) ... )
  ...)
  (X v1 v2 ...)
A name for the let (x in the example) is provided. As usual, val1 is evaluated and bound to id1, val2 is evaluated and bound to id2, etc. In the body of the let, the let may be "called" (using its name) with a fresh set of values for the let variables. Thus (X v1 v2 ...) starts the next iteration of the let with id1 bound to v1, id2 bound to v2, etc. The calls look like recursion, but they are implemented as loop iterations.

For example, in

(let loop ( (x 1) (sum 0) )
  (if (<= x 10)
    (loop (+ x 1) (+ sum x))
    sum
  )
)
we sum the values of x from 1 to 10. Compare it to

(for (x=1, sum=0; x <= 10;
     sum+=x, x+=1)
  )
Now a function to read and echo a file is straightforward:

```scheme
(define (echo filename)
  (let ((p (open-input-file filename)))
    (let loop ((obj (read p)))
      (if (eof-object? obj)
          #t ;normal termination
          (begin
            (write obj)
            (newline)
            (loop (read p)))
        ))))
```

We can create an alternative to `echo` that uses `call-with-input-file`:

```scheme
(define (echo2 filename)
  (call-with-input-file filename
    (lambda(port)
      (let loop ((obj (read port)))
        (if (eof-object? obj)
            #t
            (begin
              (write obj)
              (newline)
              (loop (read port)))
        ))))
```

Control Flow in Scheme

Normally, Scheme's control flow is simple and recursive:

- The first argument is evaluated to get a function.
- Remaining arguments are evaluated to get actual parameters.
- Actual parameters are bound to the function's formal parameters.
- The functions' body is evaluated to obtain the value of the function call.

This approach routinely leads to deeply nested expression evaluation.

As an example, consider a simple function that multiplies a list of integers:

```scheme
(define (*list L)
  (if (null? L)
      1
      (* (car L) (*list (cdr L)))
  ))
```

The call `(*list '(1 2 3 4 5))` expands to

`(* 1 (* 2 (* 3 (* 4 (* 5 1))))))`

But, what if we get clever and decide to improve this function by noting that if 0 appears anywhere in list L, the product must be 0?
Let's try

(define (*list0 L) (cond
  ((null? L) 1)
  ((= 0 (car L)) 0)
  (else (* (car L)
         (*list0 (cdr L))))))
)

This helps a bit—we never go past a zero in L, but we still unnecessarily do a sequence of pending multiplies, all of which must yield zero!
Can we escape from a sequence of nested calls once we know they’re unnecessary?

Exceptions

In languages like Java, a statement may throw an exception that’s caught by an enclosing exception handler. Code between the statement that throws the exception and the handler that catches it is abandoned.

Let’s solve the problem of avoiding multiplication of zero in Java, using its exception mechanism:

```java
class Node {
  int val;
  Node next;
}
class Zero extends Throwable {};```

```java
int mult (Node L) {
  try {
    return multNode(L);
  } catch (Zero z) {
    return 0;
  }
}
int multNode(Node L) throws Zero {
  if (L == null)
    return 1;
  else if (L.val == 0)
    throw new Zero();
  else return L.val * multNode(L.next);
}
```

In this implementation, no multiplies by zero are ever done.

Continuations

In our Scheme implementation of *list, we’d like a way to delay doing any multiplies until we know no zeros appear in the list. One approach is to build a continuation—a function that represents the context in which a function’s return value will be used:

```scheme
(define (*listC L con) (cond
  ((null? L) (con 1))
  ((= 0 (car L)) 0)
  (else
   (*listC (cdr L)
     (lambda (n) (* n (con (car L)))))
   ))
)}
```
The top-level call is
(*listC L (lambda (x) x))
For ordinary lists *listC expands to a series of multiplies, just like *list did.

(define (id x) x)
(*listC '(1 2 3) id) ⇒
(*listC '(2 3)
  (lambda (n) (* n (id 1))) ≡
  (*listC '(2 3)
    (lambda (n) (* n 1)) ⇒
    (*listC '3
      (lambda (n) (* n (* 2 1))) ≡
      (*listC '3
        (lambda (n) (* n 2)) ⇒
        (*listC()
          (lambda (n) (* n (* 3 2))) ≡
          (*listC () (lambda (n) (* n 6))) ⇒ (* 1 6) ⇒ 6

But for a list with a zero in it, we get a different execution path:
(*listC '(1 0 3) id) ⇒
(*listC '(0 3)
  (lambda (n) (* n (id 1))) ⇒ 0

No multiplies are done!

Another Example of Continuations

Let’s redo our list multiply example so that if a zero is seen in the list we return a function that computes the product of all the non-zero values and a parameter that is the “replacement value” for the unwanted zero value. The function gives the caller a chance to correct a probable error in the input data.

We create
(*list2 L) ≡
Product of all integers in L if no zero appears
else
  (lambda (n) (* n product-of-all-nonzeros-in-L)

(define (*list2 L) (*listE L id))
(define (*listE L con)
  (cond
    (((null? L) (con 1))
    ((= 0 (car L))
      (lambda (n)
        (* (con n)
          (*listE (cdr L) id))))
    else
      (*listE (cdr L)
        (lambda (m)
          (* m (con (car L)))))))
  )
)
In the following, we check to see if \( *\text{list2} \) returns a number or a function. If a function is returned, we call it with 1, effectively removing 0 from the list:

\[
\begin{align*}
\text{(let } (\text{V } (*\text{list2} \ L)) & \text{ )} \\
 & \text{(if (number? V)} \\
 & \quad \text{V} \\
 & \quad (\text{V 1}) \\
\text{)}
\end{align*}
\]

For ordinary lists \( *\text{list2} \) expands to a series of multiplies, just like \( *\text{list} \) did:

\[
\begin{align*}
(*\text{listE } '(1 2 3) \ \text{id}) & \Rightarrow \\
(*\text{listE } '(2 3) \ \text{id}) & \Rightarrow \\
(*\text{listE } '(3 2) \ \text{id}) & \Rightarrow \\
(*\text{listE } '(n) \ \text{id}) & \Rightarrow \\
(*\text{listE } () \ \text{id}) & \Rightarrow \\
\Rightarrow & 1 \Rightarrow 6
\end{align*}
\]

But for a list with a zero in it, we get a different execution path:

\[
\begin{align*}
(*\text{listE } '(1 0 3) \ \text{id}) & \Rightarrow \\
(*\text{listE } '(0 3) \ \text{id}) & \Rightarrow \\
(*\text{listE } '(0 3) \ \text{id}) & \Rightarrow \\
(*\text{listE } '(0 3) \ \text{id}) & \Rightarrow \\
(*\text{listE } () \ \text{id}) & \Rightarrow \\
\Rightarrow & (* 1 \Rightarrow 6) \Rightarrow 6
\end{align*}
\]

But note that only one zero value in the list is handled correctly! Why?

\[
\begin{align*}
\text{(define } (*\text{listE } L \ \text{con}) & \text{ )} \\
\text{(cond} & \text{ )} \\
((null? L) (con 1)) & \text{ )} \\
((0 \ \text{car L}) & \text{ )} \\
(* \ \text{con n}) & \text{ )} \\
(*\text{listE } (\text{cdr L}) \ \text{id}) & \text{ )} \\
\text{(else} & \text{ )} \\
(*\text{listE } (\text{cdr L}) & \text{ )} \\
(l\ \text{con (car L)} & \text{ )})
\end{align*}
\]

This function multiplies \( n \), the replacement value for 0, by 1 and 3, the non-zero values in the input list.
Continuations in Scheme

Scheme provides a built-in mechanism for creating continuations. It has a long name: `call-with-current-continuation` This name is usually abbreviated as `call/cc`

`call/cc` takes a single function as its argument. That function also takes a single argument. That is, we use `call/cc` as

```
(call/cc funct) where
```

```
funct ≡ (lambda (con) (body))
```

`call/cc` calls the function that it is given with the “current continuation” as the function’s argument.

Current Continuations

What is the current continuation?

It is itself a function of one argument. The current continuation function represents the execution context within which the `call/cc` appears. The argument to the continuation is a value to be substituted as the return value of `call/cc` in that execution context.

For example, given

```
(+ (fct n) 3)
```

the current continuation for `(fct n)` is

```
(lambada (x) (+ x 3))
```

Given

```
(* 2 (+ (fct z) 10))
```

the current continuation for `(fct z)` is

```
(lambada (m) (* 2 (+ m 10))
```

To use `call/cc` to grab a continuation in (say) `(+ (fct n) 3)` we make `(fct n)` the body of a function of one argument. Let’s call that argument `return`. We therefore create

```
(lambda (return) (fct n))
```

Then

```
(call/cc
  (lambda (return) (fct n)))
```

binds the current continuation to `return` and executes `(fct n)`.

We can ignore the current continuation bound to `return` and do a normal return or we can use `return` to force a return to the calling context of the `call/cc`.

The call `(return value)` forces value to be returned as the value of `call/cc` in its context of call.

Example:

```
(*
  (call/cc (lambda (return))
    (/ (g return) 0)))
```

Now during evaluation no divide by zero error occurs. Rather, when `(g return)` is called, 5 is passed to `con`, which is bound to `return`. Therefore 5 is used as the value of the call to `call/cc`, and 50 is computed.
Continuations are Just Functions

Continuations may be saved in variables or data structures and called in the future to “reactive” a completed or suspended computation.

```scheme
(define CC ())
(define (F)
  (let (
    (v (call/cc
        (lambda(here)
          (set! CC here)
          1))))
    (display "The ans is: ")
    (display v)
    (newline)
  ) )
)
This displays The ans is: 1
At any time in the future, (CC 10) will display The ans is: 10
```

List Multiplication Revisited

We can use call/cc to reimplement the original *list to force an immediate return of 0 (much like a throw in Java):

```scheme
(define (*listc L return)
  (cond
    ((null? L) 1)
    ((= 0 (car L)) (return 0))
    (else (* (car L)
          (*listc (cdr L) return)))))
)
(define (*list L)
  (call/cc
    (lambda (return)
      (*listc L return)
    )))
```

A 0 in L forces a call of (return 0) which makes 0 the value of call/cc.

Interactive Replacement of Error Values

Using continuations, we can also redo *listE so that zeroes can be replaced interactively! Multiple zeroes (in both original and replacement values) are correctly handled.

```scheme
(define (*liste L return)
  (if (null? L)
    1
    (let loop ((value (car L))
      (if (= 0 value)
        (loop
          (call/cc
            (lambda (x) (return x)))
          (* value
            (*liste (cdr L) return)))
        (*value
          (*liste (cdr L) return)))
      ) )

If a zero is seen, *liste passes back to the caller (via return) a continuation that will set the next value of value. This value is checked, so if it is itself zero, a substitute is requested. Each occurrence of zero forces a return to the caller for a substitute value.
Implementing Coroutines with call/cc

Coroutines are a very handy generalization of subroutines. A coroutine may suspend its execution and later resume from the point of suspension. Unlike subroutines, coroutines do not have to complete their execution before they return.

Coroutines are useful for computation of long or infinite streams of data, where we wish to compute some data, use it, compute additional data, use it, etc.

Subroutines aren’t always able to handle this, as we may need to save a lot of internal state to resume with the correct next value.

Producer/Consumer using Coroutines

The example we will use is one of a consumer of a potentially infinite stream of data. The next integer in the stream (represented as an unbounded list) is read. Call this value n. Then the next n integers are read and summed together. The answer is printed, and the user is asked whether another sum is required. Since we don’t know in advance how many integers will be needed, we’ll use a coroutine to produce the data list in segments, requesting another segment as necessary.

```
(define (consumer)
  (next 0) ; reset next function
  (let loop ((data (moredata)))
    (let {
      (sum+restoflist
       (sum-n-elems (car data)
          (cons 0 (cdr data)))))
      (display (car sum+restoflist))
      (newline)
      (display "more? ")
      (if (equal? (read) 'y)
        (if (= 1
            (length sum+restoflist))
          (loop (moredata))
          (loop (cdr sum+restoflist)))
        #t ; Normal completion
      )
    )
  )
)
```

Next, we’ll consider sum-n-elems, which adds the first element of list (a running sum) to the next n elements on the list. We’ll use moredata to extend the data list as needed.

```
(define (sum-n-elems n list)
  (cond
    ((= 0 n)   list)
    ((null? (cdr list))
      (sum-n-elems n
        (cons (car list) (moredata))))
    (else
      (sum-n-elems (- n 1)
        (cons (+ (car list)
                 (cadr list))
          (cddr list))))
  ))
)```
The function `moredata` is called whenever we need more data. Initially a `producer` function is called to get the initial segment of data. `producer` actually returns the next data segment plus a continuation (stored in `producer-cc`) used to resume execution of `producer` when the next data segment is required.

```scheme
(define moredata
  (let ( (producer-cc  () ) )
    (lambda ()
      (let ( 
        (data+cont
          (if (null? producer-cc)
            (call/cc (lambda (here)
              (producer here)))
            (call/cc (lambda (here)
              (producer-cc here))))
        ))
      (set! producer-cc
        (cdr data+cont))
      (car data+cont))
    )
  )
)
```

Function `(next z)` returns the next `z` integers in an infinite sequence that starts at 1. A value `z=0` is a special flag indicating that the sequence should be reset to start at 1.

```scheme
(define next
  (let ( (i 1))
    (lambda (z)
      (if (= 0 z)
        (set! i 1)
        (let loop
          ((cnt z) (val i) (ints () ))
          (if (> cnt 0)
            (loop (- cnt 1)
              (+ val 1)
              (append ints
                (list val)))
            (begin
              (set! i val)
              ints
              )
          )
        )
      )))
)
```

The function `producer` generates an infinite sequence of integers (1,2,3,...). It suspends every 5/10/15/25 elements and returns control to `moredata`.

```scheme
(define (producer initial-return)
  (let loop
    ( (return initial-return) )
    (set! return
      (call/cc (lambda (here)
        (return (cons (next 5)
          here)))))
    (set! return
      (call/cc (lambda (here)
        (return (cons (next 10)
          here)))))
    (set! return
      (call/cc (lambda (here)
        (return (cons (next 15)
          here)))))
    (loop
      (call/cc (lambda (here)
        (return (cons (next 25)
          here))))))
  )
)
```
Reading Assignment

- MULTILISP: a language for concurrent symbolic computation, by Robert H. Halstead (linked from class web page)

Lazy Evaluation

Lazy evaluation is sometimes called “call by need.” We do an evaluation when a value is used; not when it is defined.

Scheme provides for lazy evaluation:

(delay expression)
Evaluation of expression is delayed. The call returns a “promise” that is essentially a lambda expression.
(force promise)
A promise, created by a call to delay, is evaluated. If the promise has already been evaluated, the value computed by the first call to force is reused.

Example:

Though and is predefined, writing a correct implementation for it is a bit tricky.

The obvious program

(define (and A B)
  (if A
      B
      #f)
)

is incorrect since B is always evaluated whether it is needed or not. In a call like

(and (not (= i 0)) (> (/ j i) 10))

unnecessary evaluation might be fatal.

An argument to a function is strict if it is always used. Non-strict arguments may cause failure if evaluated unnecessarily.

With lazy evaluation, we can define a more robust and function:

(define (and A B)
  (if A
      (force B)
      #f)
)

This is called as:

(and (not (= i 0))
     (delay (> (/ j i) 10)))

Note that making the programmer remember to add a call to delay is unappealing.
Delayed evaluation also allows us a neat implementation of suspensions. The following definition of an infinite list of integers clearly fails:

\[
\text{(define (inflist i)}
\begin{array}{c}
\text{(cons i (inflist (+ i 1)))))}
\end{array}
\]

But with use of delays we get the desired effect in finite time:

\[
\text{(define (inflist i)}
\begin{array}{c}
\text{(cons i}
\begin{array}{c}
\text{(delay (inflist (+ i 1)))))}
\end{array}
\end{array}
\]

Now a call like \text{(inflist 1)} creates 1 promise for \text{(inflist 2)}

We need to slightly modify how we explore suspended infinite lists. We can't redefine \text{car} and \text{cdr} as these are far too fundamental to tamper with.

Instead we'll define \text{head} and \text{tail} to do much the same job:

\[
\text{(define head car)}
\begin{array}{c}
\text{(define (tail L)}
\begin{array}{c}
\text{(force (cdr L))})
\end{array}
\end{array}
\]

\text{head} looks at \text{car} values which are fully evaluated.

\text{tail} forces one level of evaluation of a delayed \text{cdr} and saves the evaluated value in place of the suspension (promise).

Exploiting Parallelism

Conventional procedural programming languages are difficult to compile for multiprocessors.

Frequent assignments make it difficult to find independent computations.

Consider (in Fortran):

\[
\begin{array}{c}
\text{do 10 I = 1,1000}
\text{X(I) = 0}
\text{A(I) = A(I+1) + 1}
\text{B(I) = B(I-1) - 1}
\text{C(I) = (C(I-2) + C(I+2))/2}
\text{10 continue}
\end{array}
\]

This loop defines 1000 values for arrays \text{X, A, B and C.}
Which computations can be done in parallel, partitioning parts of an array to several processors, each operating independently?

- \( X(I) = 0 \)
  Assignments to \( x \) can be readily parallelized.

- \( A(I) = A(I+1) + 1 \)
  Note that each computation of \( A(I) \) uses an \( A(I+1) \) value that is yet to be changed. Thus a whole array of new \( A \) values can be computed from an array of “old” \( A \) values in parallel.

- \( B(I) = B(I-1) - 1 \)
  This is less obvious. Each \( B(I) \) uses \( B(I-1) \) which is defined in terms of \( B(I-2) \), etc. Ultimately all new \( B \) values depend only on \( B(0) \) and \( I \). That is, \( B(I) = B(0) - I \). So this computation can be parallelized, but it takes a fair amount of insight to realize it.

- \( C(I) = (C(I-2) + C(I+2))/2 \)
  It is clear that even and odd elements of \( C \) don’t interact. Hence two processors could compute even and odd elements of \( C \) in parallel. Beyond this, since both earlier and later \( C \) values are used in each computation of an element, no further means of parallel evaluation is evident. Serial evaluation will probably be needed for even or odd values.

Exploiting Parallelism in Scheme

Assume we have a shared-memory multiprocessor. We might be able to assign different processors to evaluate various independent subexpressions.

For example, consider

\[
\text{(map (lambda(x) (* 2 x)) '(1 2 3 4 5))}
\]

We might assign a processor to each list element and compute the lambda function on each concurrently:

```
1  2  3  4  5
   ↑   ↑   ↑
Processor 1   Processor 5
2  4  6  8 10
```

How is Parallelism Found?

There are two approaches:

- We can use a “smart” compiler that is able to find parallelism in existing programs written in standard serial programming languages.
- We can add features to an existing programming language that allows a programmer to show where parallel evaluation is desired.
Concurrentization

Concurrentization (often called parallelization) is the process of automatically finding potential concurrent execution in a serial program.

Automatically finding concurrent execution is complicated by a number of factors:

- Data Dependence
  Not all expressions are independent. We may need to delay evaluation of an operator or subprogram until its operands are available. Thus in
  
  \[(+ (* x y) (* y z))\]
  
  we can't start the addition until both multiplications are done.

- Control Dependence
  Not all expressions need be (or should be) evaluated. In
  
  \[(if (= a 0) 0 (/ b a))\]
  
  we don't want to do the division until we know \(a \neq 0\).

- Side Effects
  If one expression can write a value that another expression might read, we probably will need to serialize their execution. Consider
  
  \[(define rand!
  \begin{align*}
  & (let ((seed 99)) \\
  & (lambda () \\
  & \begin{align*}
  & \text{(set! seed (mod (* seed 1001) 101101))} \\
  & \text{seed}
  \end{align*}
  \))
  \)
  \]
  
  Now in
  
  \[(+ (f (rand!)) (g (rand!)))\]
  
  we can't evaluate \((f (rand!))\) and \((g (rand!))\) in parallel, because of the side effect of \(\text{set! in rand!}.\) In fact if we did, \(f\) and \(g\) might see exactly the same “random” number! (Why?)

- Granularity
  Evaluating an expression concurrently has an overhead (to setup a concurrent computation). Evaluating very simple expressions (like \(\text{car } x\) or \((+ x 1)\)) in parallel isn't worth the overhead cost. Estimating where the “break even” threshold is may be tricky.

Utility of Concurrentization

Concurrentization has been most successful in engineering and scientific programs that are very regular in structure, evaluating large multidimensional arrays in simple nested loops. Many very complex simulations (weather, fluid dynamics, astrophysics) are run on multiprocessors after extensive concurrentization.

Concurrentization has been far less successful on non-scientific programs that don't use large arrays manipulated in nested for loops. A compiler, for example, is difficult to run (in parallel) on a multiprocessor.
Concurrentization within Processors

Concurrentization is used extensively within many modern uniprocessors. Pentium and PowerPC processors routinely execute several instructions in parallel if they are independent (e.g., read and write distinct registers). This are superscalar processors. These processors also routinely speculate on execution paths, “guessing” that a branch will (or won’t) be taken even before the branch is executed! This allows for more concurrent execution than if strictly “in order” execution is done. These processors are called “out of order” processors.

Adding Parallel Features to Programming Languages.

It is common to take an existing serial programming language and add features that support concurrent or parallel execution. For example versions for Fortran (like HPF—High Performance Fortran) add a parallel do loop that executes individual iterations in parallel. Java supports threads, which may be executed in parallel. Synchronization and mutual exclusion are provided to avoid unintended interactions.

Multilisp

Multilisp is a version of Scheme augmented with three parallel evaluation mechanisms:

- Pcall
  Arguments to a call are evaluated in parallel.

- Future
  Evaluation of an expression starts immediately. Rather than waiting for completion of the computation, a “future” is returned. This future will eventually transform itself into the result value (when the computation completes)

- Delay
  Evaluation is delayed until the result value is really needed.

The Pcall Mechanism

Pcall is an extension to Scheme’s function call mechanism that causes the function and its arguments to be all computed in parallel.

Thus

\[(\text{pcall} \ F \ X \ Y \ Z)\]

causes \(F, X, Y\) and \(Z\) to all be evaluated in parallel. When all evaluations are done, \(F\) is called with \(X, Y\) and \(Z\) as its parameters (just as in ordinary Scheme).

Compare

\[ (+ \ (\ast \ X \ Y) \ (\ast \ Y \ Z)) \]

with

\[ (\text{pcall} \ + \ (\ast \ X \ Y) \ (\ast \ Y \ Z)) \]
It may not look like `pcall` can give you that much parallel execution, but in the context of recursive definitions, the effect can be dramatic.

Consider `treemap`, a version of `map` that operates on binary trees (S-expressions).

```scheme
(define (treemap fct tree)
  (if (pair? tree)
      (pcall cons
        (treemap fct (car tree))
        (treemap fct (cdr tree))
      )
      (fct tree)
  ))
```

Look at the execution of `treemap` on the tree

```scheme
(((1 . 2) . (3 . 4)) .
 ((5 . 6) . (7 . 8)))
```

We start with one call that uses the whole tree. This splits into two parallel calls, one operating on

```scheme
(((1 . 2) . (3 . 4))
```

and the other operating on

```scheme
(((5 . 6) . (7 . 8))
```

Each of these calls splits into 2 calls, and finally we have 8 independent calls, each operating on the values 1 to 8.

---

**Futures**

Evaluation of an expression as a future is the most interesting feature of Multilisp.

The call

```scheme
(future expr)
```

begins the evaluation of `expr`. But rather than waiting for `expr`’s evaluation to complete, the call to `future` returns immediately with a new kind of data object—a future. This future is actually an “IOU.” When you try to use the value of the future, the computation of `expr` may or may not be completed. If it is, you see the value computed instead of the future—it automatically transforms itself. Thus evaluation of `expr` appears instantaneous.

If the computation of `expr` is not yet completed, you are forced to wait until computation is completed. Then you may use the value and resume execution.

But this is exactly what ordinary evaluation does anyway—you begin evaluation of `expr` and wait until evaluation completes and returns a value to you!
To see the usefulness of futures, consider the usual definition of Scheme's map function:

```
(define (map f L)
  (if (null? L)
      ()
      (cons (f (car L))
            (map f (cdr L)))))
```

If we have a call

```
(map slow-function long-list)
```

where slow-function executes slowly and long-list is a large data structure, we can expect to wait quite a while for computation of the result list to complete.

Now consider `fastmap`, a version of `map` that uses futures:

```
(define (fastmap f L)
  (if (null? L)
      ()
      (cons (future (f (car L)))
            (fastmap f (cdr L)))))
```

Now look at the call

```
(fastmap slow-function long-list)
```

We will exploit a useful aspect of futures—they can be cons'ed together without delay, even if the computation isn’t completed yet.

Why? Well a cons just stores a pair of pointers, and it really doesn’t matter what the pointers reference (a future or an actual result value).

The call to `fastmap` can actually return before any of the call to slow-function have completed:

![Diagram of futures and answers]

Eventually all the futures automatically transform themselves into data values:

```
(answer1
 answer2
 answer3 ...)
```

Note that `pcall` can be implemented using futures.

That is, instead of

```
(pcall F X Y Z)
```

we can use

```
((future F) (future X) (future Y) (future Z))
```

In fact the latter version is actually more parallel—execution of `F` can begin even if all the parameters aren’t completely evaluated.
Another Example of Futures

The following function, `partition`, will take a list and a data value (called pivot). `partition` will partition the list into two sublists:

(a) Those elements \( \leq \) pivot
(b) Those elements \( > \) pivot

```
(define (partition pivot L)
  (if (null? L)
      (cons () () )
      (let ((tail-part
              (partition pivot (cdr L))))
        (if (\(\leq\) (car L) pivot)
            (cons
             (cons (car L) (car tail-part))
             (cdr tail-part))
            (cons
             (car tail-part))
            (cons (car L) (cdr tail-part))))
    ) ) )
)
```

We want to add futures to `partition`, but where?

It makes sense to use a future when a computation may be lengthy and we may not need to use the value computed immediately.

What computation fits that pattern? The computation of `tail-part`. We’ll mark it in a blue box to show we plan to evaluate it using a future:

```
(define (partition pivot L)
  (if (null? L)
      (cons () () )
      (let ((tail-part
              (partition pivot (cdr L))))
        (if (\(\leq\) (car L) pivot)
            (cons
             (cons (car L) (car tail-part))
             (cdr tail-part))
            (cons
             (car tail-part))
            (cons (car L) (cdr tail-part))))
    ) ) )
)
```

But this one change isn’t enough! We soon access the `car` and `cdr` of `tail-part`, which forces us to wait for its computation to complete. To avoid this delay, we can place the four references to `car` or `cdr` of `tail-part` into futures too:
Now we can build the initial part of the partitioned list (that involving pivot and \((\text{car } L)\) independently of the recursive call of \text{partition}, which completes the rest of the list. For example,

\[
(\text{partition } 17 '(5 3 8 \ldots))
\]
creates a future (call it \text{future1}) to compute

\[
(\text{partition } 17 '(3 8 \ldots))
\]
It also creates \text{future2} to compute \((\text{car tail-part})\) and \text{future3} to compute \((\text{cdr tail-part})\). The call builds

\begin{center}
\begin{tikzpicture}

\node (future2) at (0,0) {future2};
\node (future3) at (1,0) {future3};
\node (5) at (-1,0.5) {5};

\draw[->] (future2) -- (5);
\draw[->] (future3) -- (5);
\end{tikzpicture}
\end{center}

\section*{M L — M et a L angu age}

\textbf{SML is Standard ML, a popular ML variant.}

ML is a functional language that is designed to be efficient and type-safe. It demonstrates that a functional language need not use Scheme's odd syntax and need not bear the overhead of dynamic typing.

SML's features and innovations include:

1. \textbf{Strong, compile-time typing.}
2. \textbf{Automatic type inference rather than user-supplied type declarations.}
3. \textbf{Polymorphism, including “type variables.”}

\begin{verbatim}
fun len([]) = 0 |
   len(a::b) = 1+len(b);
\end{verbatim}

\section*{Reading Assignment}

- Roosta: Section 13.3
- Introduction to Standard ML (linked from class web page)
- Webber: Chapters 5, 7, 9, 11

A good ML reference is

“Elements of ML Programming,”
by Jeffrey Ullman
(Prentice Hall, 1998)
SML is Interactive

You enter a definition or expression, and SML returns a result with an inferred type.
The command
use "file name";
loads a set of ML definitions from a file.
For example (SML responses are in blue):
21;
val it = 21 : int
(2 div 3);
val it = 0 : int
ture;
val it = true : bool
"xyz";
val it = "xyz" : string

Basic SML Predefined Types

• Unit
Its only value is (). Type unit is similar to void in C; it is used where a type is needed, but no "real" type is appropriate. For example, a call to a write function may return unit as its result.

• Integer
Constants are sequences of digits. Negative values are prefixed with a ~ rather than a – (~ is a binary subtraction operator). For example, ~123 is negative 123.
Standard operators include
\[ + \quad - \quad * \quad div \quad mod \]
\[ < \quad > \quad <= \quad >= \quad = \quad <> \]

• Real
Both fractional (123.456) and exponent forms (10e7) are allowed. Negative signs and exponents use ~ rather than – (~10.0e~12). Standard operators include
\[ + \quad - \quad * \quad / \]
\[ < \quad > \quad <= \quad >= \]
Note that = and <> aren't allowed! (Why?) Conversion routines include
real(int) to convert an int to a real,
floor(real) to take the floor of a real,
ceil(real) to take the ceiling of a real.
round(real) to round a real,
trunc(real) to truncate a real.

For example, real(3) returns 3.0,
floor(3.1) returns 3,
ceiling(3.3) returns 3,
round(~3.6) returns ~4,
trunc(3.9) returns 3.
Mixed mode expressions, like
1 + 2.5 aren't allowed; you must do explicit conversion, like
real(1) + 2.5

• Strings
Strings are delimited by double quotes. Newlines are \n, tabs are \t, and \" and \\ escape double quotes and backslashes. E.g. "Bye now\n" The ^ operator is concatenation.
"abc" ^ "def" = "abcdef"
The usual relational operators are provided: < > <= >= = <>
. Characters

Single characters are delimited by double quotes and prefixed by a #. For example, #"a" or #"\t". A character is not a string of length one. The `str` function may be used to convert a character into a string. Thus `str(#"a") = "a"

. Boolean

Constants are `true` and `false`. Operators include `andalso` (short-circuit and), `orelse` (short-circuit or), `not`, `=`, and `<>

A conditional expression,

(if boolval v_1 else v_2) is available.

Equality is checked componentwise:

(1, 2) = (0+1, 1+1);
val it = true : bool
(1, 2, 3) = (1, 2) causes a compile-time type error (tuples must be of the same length and have corresponding types to be compared).

#_ selects the _-th component of a tuple (counting from 1). Hence

#_2(1, 2, 3);
val it = 2 : int

Tuples

A tuple type, composed of two or more values of any type is available. Tuples are delimited by parentheses, and values are separated by commas. Examples include:

(1, 2);
val it = (1, 2) : int * int
("xyz", 1=2);
val it = ("xyz", false) : string * bool
(1, 3.0, false);
val it = (1, 3.0, false) : int * real * bool
(1, 2, (3, 4));
val it = (1, 2, (3, 4)) : int * int * (int * int)

Lists

Lists are required to have a single element type for all their elements; their length is unbounded. Lists are delimited by [ and ] and elements are separated by commas. Thus [1, 2, 3] is an integer list. The empty (or null) list is [] or nil.

The cons operator is ::

Hence [1, 2, 3] ≡ 1::2::3::[]

Lists are automatically typed by ML:

[1, 2];
val it = [1, 2] : int list
Cons

Cons is an infix operator represented as ::
The left operand of :: is any value of type T.
The right operand of :: is any list of type T list.
The result of :: is a list of type T list.
Hence :: is polymorphic.
[] is the empty list. It has a type 'a list. The symbol 'a, read as “alpha” or “tic a” is a type variable.
Thus [] is a polymorphic constant.

List Equality

Two lists may be compared for equality if they are of the same type. Lists L1 and L2 are considered equal if:
(1) They have the same number of elements
(2) Corresponding members of the two lists are equal.

List Operators

hd ≡ head of list operator ≈ car
tl ≡ tail of list operator ≈ cdr
null ≡ null list predicate ≈ null?
@ ≡ infix list append operator ≈ append

Records

Their general form is {name1=val1, name2=val2, ... }
Field selector names are local to a record.
For example:
{a=1,b=2};
val it = {a=1,b=2} :
{a:int, b:int}
{a=1,b="xyz"};
val it = {a=1,b="xyz"} :
{a:int, b:string}
{a=1.0,b={c=[1,2]});
val it = {a=1.0,b={c=[1,2]}} :
{a:real, b:{c:int list}}
The order of fields is irrelevant; equality is tested using field names.
{a=1,b=2}={b=2,a=2-1};
val it = true : bool
#id extracts the field named id from a record.
#b {a=1,b=2} ;
val it = 2 : int
Identifiers

There are two forms:
- Alphanumeric (excluding reserved words)
  Any sequence of letters, digits, single quotes and underscores; must begin with a letter or single quote.
  Case is significant. Identifiers that begin with a single quote are type variables.
  Examples include:
    abc a10 'polar sum_of_20
- Symbolic
  Any sequence (except predefined operators) of
    ! % & + - / : < = > ? @ \ ^ | # *
  Usually used for user-defined operators.
  Examples include: ++ <=> !=

Comments

Of form
(* text *)
May cross line boundaries.

Declaration of Values

The basic form is
val id = expression;
This defines id to be bound to expression; ML answers with the name and value defined and the inferred type.
For example
val x = 10*10;
val x = 100 : int

Redefinition of an identifier is OK, but this is redefinition not assignment;
Thus
val x = 100;
val x = (x=100);
is fine; there is no type error even though the first x is an integer and then it is a boolean.

val x = 100 : int
val x = true : bool

Examples

val x = 1;
val x = 1 : int
val z = (x,x,x);
val L = [z,z];
val L = [(1,1,1),(1,1,1)] :
  (int * int * int) list
val r = {a=L};
val r = {a={(1,1,1),(1,1,1)}} :
  {a:(int * int * int) list}

After rebinding, the “nearest” (most recent) binding is used.
The and symbol (not boolean and) is used for simultaneous binding:

```ml
val x = 10;
val x = 10 : int
val x = true and y = x;
val x = true : bool
val y = 10 : int
```

Local definitions are temporary value definitions:

```ml
local
  val x = 10
in
  val u = x*x;
end;
val u = 100 : int
```

Let bindings are used in expressions:

```ml
let
  val x = 10
in
  5*x
end;
val it = 50 : int
```

Patterns

Scheme (and most other languages) use access or decomposition functions to access the components of a structured object.

Thus we might write

```ml
(let ((h (car L) (t (cdr L)))) )
  body )
```

Here car and cdr are used as access functions to locate the parts of L we want to access.

In ML we can access components of lists (or tuples, or records) directly by using patterns. The context in which the identifier appears tells us the part of the structure it references.

```ml
val x = (1,2);
val x = (1,2) : int * int
val (h,t) = x;
val h = 1 : int
val t = 2 : int
val L = [1,2,3];
val L = [1,2,3] : int list
val [v1,v2,v3] = L;
val v1 = 1 : int
val v2 = 2 : int
val v3 = 3 : int
val [1,x,3] = L;
val x = 2 : int
val [1,rest] = L;
(* This is illegal. Why? *)
val yy::rest = L;
val yy = 1 : int
val rest = [2,3] : int list
```
**Wildcards**

An underscore (_ ) may be used as a "wildcard" or "don't care" symbol. It matches part of a structure without defining a new binding.

```ocaml
val zz::_ = L;
val zz = 1 : int
```

Pattern matching works in records too.

```ocaml
val r = {a=1,b=2};
val r = {a=1,b=2} :
  {a:int, b:int}
val {a=va,b=vb} = r;
val va = 1 : int
val vb = 2 : int
val {a=wa,b=_}=r;
val wa = 1 : int
val {a=za, ...}=r;
val za = 1 : int
```

**Functions**

Functions take a single argument (which can be a tuple).

Function calls are of the form

```ocaml
function_name argument;
```

For example

```ocaml
size "xyz";
cos 3.14159;
```

The more conventional form

```ocaml
size("xyz"); or cos(3.14159);
```

is OK (the parentheses around the argument are allowed, but unnecessary).

The form `(size "xyz") or (cos 3.14159)`

is OK too.

**Patterns can be nested too.**

```ocaml
val x = ((1, 3.0), 5);
val x = ((1, 3.0), 5) :
  (int * real) * int
val ((1, y), _) = x;
val y = 3.0 : real
```

**Note that the call**

```ocaml```

```ocaml
plus(1,2);
passes one argument, the tuple (1,2) to plus.
The call dummy();
passes one argument, the unit value, to dummy.
All parameters are passed by value.
```
Function Types

The type of a function in ML is denoted as \( T_1 \rightarrow T_2 \). This says that a parameter of type \( T_1 \) is mapped to a result of type \( T_2 \).

The symbol \( fn \) denotes a value that is a function.

Thus

\[
\begin{align*}
\text{size;} \\
\text{val it = fn : string \rightarrow int} \\
\text{not;} \\
\text{val it = fn : bool \rightarrow bool} \\
\text{Math.cos;} \\
\text{val it = fn : real \rightarrow real}
\end{align*}
\]

(Math is an ML structure—an external library member that contains separately compiled definitions).

User-Defined Functions

The general form is

fun name arg = expression;

ML answers back with the name defined, the fact that it is a function (the \( fn \) symbol) and its inferred type.

For example,

\[
\begin{align*}
\text{fun twice x = 2*x;} \\
\text{val twice = fn : int \rightarrow int} \\
\text{fun twotimes(x) = 2*x;} \\
\text{val twotimes = fn : int \rightarrow int} \\
\text{fun fact n =} \\
\text{\quad if n=0} \\
\text{\quad then 1} \\
\text{\quad else n*fact(n-1);} \\
\text{val fact = fn : int \rightarrow int}
\end{align*}
\]

Patterns In Function Definitions

The following defines a predicate that tests whether a list, \( L \) is null (the predefined null function already does this).

\[
\begin{align*}
\text{fun isNull L =} \\
\text{\quad if L=[] then true else} \\
\text{\quad false;} \\
\text{val isNull = fn : 'a list \rightarrow bool}
\end{align*}
\]

However, we can decompose the definition using patterns to get a simpler and more elegant definition:

\[
\begin{align*}
\text{fun isNull [] = true} \\
\text{\quad | isNull(_::_) = false;} \\
\text{val isNull = fn : 'a list \rightarrow bool}
\end{align*}
\]
The “|” divides the function definition into different argument patterns; no explicit conditional logic is needed. The definition that matches a particular actual parameter is automatically selected.

```latex
fun fact(1) = 1
    | fact(n) = n*fact(n-1);
val fact = fn : int -> int
```

If patterns that cover all possible arguments aren't specified, you may get a run-time Match exception. If patterns overlap you may get a warning from the compiler.

```latex
fun append([],L) = L
    | append(hd::tl,L) = hd::append(tl,L);
val append = fn :
    'a list * 'a list -> 'a list
```

If we add the pattern

```latex
append(L,[]) = L
```

we get a redundant pattern warning (Why?)

```latex
fun append([],L) = L
    | append(hd::tl,L) = hd::append(tl,L)
    | append(L,[]) = L;
```

```latex
stdIn:151.1-153.20 Error: match redundant
  (nil,L) => ...
  (hd :: tl,L) => ...
  --> (L,nil) => ...
```

But a more precise decomposition is fine:

```latex
fun append([],L) = L
    | append(hd::tl,hd2::tl2) = hd::append(tl,hd2::tl2)
    | append(hd::tl,[]) = hd::tl;
val append = fn :
    'a list * 'a list -> 'a list
```

```latex
Function Types Can be Polytypes
Recall that 'a, 'b, ... represent type variables. That is, any valid type may be substituted for them when checking type correctness.

ML said the type of append is

```latex
val append = fn :
    'a list * 'a list -> 'a list
```

Why does 'a appear in three places?

We can define eitherNull, a predicate that determines whether either of two lists is null as

```latex
fun eitherNull(L1,L2) = null(L1) orelse null(L2);
val eitherNull = fn :
    'a list * 'b list -> bool
```

Why are both 'a and 'b used in eitherNull's type?
Currying

ML chooses the most general (least-restrictive) type possible for user-defined functions.
Functions are first-class objects, as in Scheme.
The function definition
\[ \text{fun } f \ x \ y = \text{expression}; \]
defines a function \( f \) (of \( x \)) that returns a function (of \( y \)).
Reducing multiple argument functions to a sequence of one argument functions is called currying (after Haskell Curry, a mathematician who popularized the approach).

Thus
\[ \text{fun } f \ x \ y = x :: [y]; \]
\[ \text{val } f = \text{fn} : 'a -> 'a -> 'a list \]
says that \( f \) takes a parameter \( x \), of type \( 'a \), and returns a function (of \( y \), whose type is \( 'a \)) that returns a list of \( 'a \).
Contrast this with the more conventional
\[ \text{fun } g(x,y) = x :: [y]; \]
\[ \text{val } g = \text{fn} : 'a * 'a -> 'a list \]
Here \( g \) takes a pair of arguments (each of type \( 'a \)) and returns a value of type \( 'a \) list.
The advantage of currying is that we can bind one argument and leave the remaining argument(s) free.

For example
\[ f(1); \]
is a legal call. It returns a function of type
\[ \text{fn} : \text{int} -> \text{int list} \]
The function returned is equivalent to
\[ \text{fun } h \ b = 1 :: [b]; \]
\[ \text{val } h = \text{fn} : \text{int} -> \text{int list} \]

Map Revisited

ML supports the map function, which can be defined as
\[ \text{fun } \text{map}(f,[]) = []; \]
\[ \text{| } \text{map}(f,[x::y]) = \]
\[ (f\ x) :: \text{map}(f,y); \]
\[ \text{val } \text{map} = \]
\[ \text{fn} : ('a -> 'b) * 'a list -> 'b list \]
This type says that \( \text{map} \) takes a pair of arguments. One is a function from type \( 'a \) to type \( 'b \). The second argument is a list of type \( 'a \). The result is a list of type \( 'b \).
In curried form \( \text{map} \) is defined as
\[ \text{fun } \text{map} \ f \ [] = []; \]
\[ \text{| } \text{map} \ f \ [x::y] = \]
\[ (f\ x) :: \text{map} \ f \ y; \]
\[ \text{val } \text{map} = \]
\[ \text{fn} : ('a -> 'b) -> \]
\[ 'a list -> 'b list \]
This type says that \texttt{map} takes one argument that is a function from type 'a to type 'b. It returns a function that takes an argument that is a list of type 'a and returns a list of type 'b.

The advantage of the curried form of \texttt{map} is that we can now use \texttt{map} to create "specialized" functions in which the function that is mapped is fixed.

For example,
\begin{verbatim}
val neg = map not;
val neg =
  fn : bool list -> bool list
neg [true,false,true];
val it = [false,true,false] :
  bool list
\end{verbatim}

Power Sets Revisited

Let’s compute power sets in ML.

We want a function \texttt{pow} that takes a list of values, viewed as a set, and which returns a list of lists. Each sublist will be one of the possible subsets of the original argument.

For example,
\begin{verbatim}
pow [1,2] = [[1,2],[1],[2],[[]]]
\end{verbatim}

We first define a version of \texttt{cons} in curried form:

\begin{verbatim}
fun cons h t = h::t;
val cons = fn :
  'a -> 'a list -> 'a list
\end{verbatim}

Now we define \texttt{pow}. We define the powerset of the empty list, [], to be [[]]. That is, the power set of the empty set is the set that contains only the empty set.

For a non-empty list, consisting of \texttt{h::t}, we compute the power set of \texttt{t}, which we call \texttt{pset}. Then the power set for \texttt{h::t} is just \texttt{h} distributed through \texttt{pset} appended to \texttt{pset}.

We distribute \texttt{h} through \texttt{pset} very elegantly: we just map the function \texttt{(cons h)} to \texttt{pset}. \texttt{(cons h)} adds \texttt{h} to the head of any list it is given. Thus mapping \texttt{(cons h)} to \texttt{pset} adds \texttt{h} to all lists in \texttt{pset}.

The complete definition is simply
\begin{verbatim}
 fun pow [] = [[]]
 | pow (h::t) =
   let
     val pset = pow t
   in
     (map (cons h) pset) @ pset
   end;
val pow =
  fn : 'a list -> 'a list list
\end{verbatim}

Let’s trace the computation of \texttt{pow [1,2]}.

Here \texttt{h = 1} and \texttt{t = [2]}. We need to compute \texttt{pow [2]}.

Now \texttt{h = 2} and \texttt{t = []}.

We know \texttt{pow [1] = [[]]},
so \texttt{pow [2] =}
\begin{verbatim}
(map (cons 2) [[]])@[[]] =
([[[2]]])@[[]] = [2],[[]]
\end{verbatim}
Therefore \( \text{pow} \ [1,2] = (\text{map} \ (\text{cons} \ 1) \ [[2],[]]) \ @ [[2],[]] = [[1,2],[1],[2],[]] \)

### Composing Functions

We can define a composition function that composes two functions into one:

```sml
define comp (f,g) x = f(g(x));

val comp = fn :
('a -> 'b) * ('c -> 'a) -> 'c -> 'b

In curried form we have

```sml
define comp f g x = f(g(x));

val comp = fn :
('a -> 'b) -> ('c -> 'a) -> 'c -> 'b

For example,

```sml
define sqr x :int = x*x;

val sqr = fn : int -> int

comp sqr sqr 3;

val it = 81 : int
```

In SML \( o \) (lower-case \( O \)) is the infix composition operator.

Hence

\( \text{sqr} \circ \text{sqr} \equiv \text{comp} \ \text{sqr} \ \text{sqr} \)

### Lambda Terms

ML needs a notation to write down unnamed (anonymous) functions, similar to the lambda expressions Scheme uses.

That notation is

```sml
fn \( \text{arg} \) => \( \text{body} \);
```

For example,

```sml
val sqr = fn \( \text{x:int} \) => \( \text{x*x} \);

val sqr = fn : int -> int

comp sqr sqr;

val it = fn : int -> int
```

In fact the notation used to define functions,

```sml
fun \( \text{name} \ \text{arg} \) = \( \text{body} \);
```

is actually just an abbreviation for the more verbose

```sml
val name = fn \( \text{arg} \) => \( \text{body} \);
```
An anonymous function can be used wherever a function value is needed. For example,

```plaintext
code
map (fn x => [x]) [1,2,3]
val it = [[1],[2],[3]] : int list list
```

We can use patterns too:

```plaintext
code
(fn [] => [] | (h::t) => h::h::t)
val it = fn : 'a list -> 'a list
```

(What does this function do?)

Polymorphism vs. Overloading

ML supports polymorphism.

A function may accept a polytype (a set of types) rather than a single fixed type.

In all cases, the same function definition is used. Details of the supplied type are irrelevant and may be ignored.

For example,

```plaintext
code
fun id x = x;
val id = fn : 'a -> 'a
fun toList x = [x];
val toList = fn : 'a -> 'a list
```

Overloading, as in C++ and Java, allows alternative definitions of the same method or operator, with selection based on type.

Thus in Java + may represent integer addition, floating point addition or string concatenation, even though these are really rather different operations.

In ML +, -, *, and = are overloaded.

When = is used (to test equality), ML deduces that an equality type is required. (Most, but not all, types can be compared for equality).

When ML decides an equality type is needed, it uses a type variable that begins with two tics rather than one.

```plaintext
code
fun eq(x,y) = (x=y);
val eq = fn : ''a * ''a -> bool
```

Defining New Types in ML

We can create new names for existing types (type abbreviations) using

```plaintext
code
type id = def;
For example,

type triple = int*real*string;
type triple = int * real * string
type rec1= {a:int,b:real,c:string};
type rec1 = {a:int, b:real, c:string}
type 'a triple3  = 'a*'a*'a;
type 'a triple3 = 'a * 'a * 'a
```

These type definitions are essentially macro-like name substitutions.
The Datatype Mechanism

New types are defined using the datatype mechanism, which specifies new data value constructors. For example,

datatype color =
  red | blue | green;

datatype color =
  blue | green | red

Pattern matching works on user-defined types using their constructors:

fun translate red = "rot"
  | translate blue = "blau"
  | translate green = "gruen";
val translate =
  fn : color -> string

fun jumble red = blue
  | jumble blue = green
  | jumble green = red;
val jumble = fn : color -> color
translate (jumble green);
val it = "rot" : string

SML Examples

Source code for most of the SML examples presented here may be found in
~cs538-1/public/sml/class.sml

Parameterized Constructors

The constructors used to define data types may be parameterized:

datatype money =
  none
  | coin of int
  | bill of int
  | iou of real * string;

datatype money =
  bill of int | coin of int
  | iou of real * string | none

Now expressions like color(25) or bill(5) or iou(10.25, "Lisa") represent valid values of type money.

fun amount(none) = 0.0
  | amount(coin(cents)) = real(cents)/100.0
  | amount(bill(dollars)) = real(dollars)
  | amount(iou(amt, _)) = 0.5*amt;
val amount = fn : money -> real
Polymorphic Datatypes

A user-defined data type may be polymorphic. An excellent example is

datatype 'a option =
  none | some of 'a;

val zilch = none;
val mucho = some(10e10);

val newStudent =
  {name="Mystery Man", ssNumber=none}: studentInfo;

val newStudent =
  {name="Mystery Man", ssNumber=none} : studentInfo

type studentInfo =
  {name:string, ssNumber:int option};

type studentInfo = {name:string, ssNumber:int option}

Datatypes may be Recursive

Recursive datatypes allow linked structures without explicit pointers.

datatype binTree =
  null |
  leaf |
  node of binTree * binTree;

datatype binTree =
  leaf |
  node of binTree * binTree |
  null

fun size(null) = 0 |
    size(leaf) = 1 |
    size(node(t1,t2)) =
        size(t1)+size(t2) + 1

val size = fn : binTree -> int

Recursive Datatypes may be Polymorphic

datatype 'a binTree =
  null |
  leaf of 'a |
  node of 'a binTree * 'a binTree

datatype 'a binTree =
  leaf of 'a |
  node of 'a binTree * 'a binTree |
  null

fun frontier(null) = [] |
    frontier(leaf(v)) = [v] |
    frontier(node(t1,t2)) =
        frontier(t1) @ frontier(t2)

val frontier =
  fn : 'a binTree -> 'a list
We can model n-ary trees by using lists of subtrees:

```ml
datatype 'a Tree =
  null |
  leaf of 'a |
  node of 'a Tree list;
```

```ml
fun frontier(null) = []
  |  frontier(leaf(v)) = [v]
  |  frontier(node(h::t)) =
    frontier(h) @
    frontier(node(t))
  |  frontier(node([])) = []
val frontier = fn :
  'a Tree -> 'a list
```

### Abstract Data Types

ML also provides abstract data types in which the implementation of the type is hidden from users.

The general form is

```ml
abstype name = implementation
with
  val and fun definitions
end;
```

Users may access the name of the abstract type and the val and fun definitions that follow the with, but the implementation may be used only with the body of the abstype definition.

### Example

```ml
abstype 'a stack =
  stk of 'a list
with
  val Null = stk([])
  fun empty(stk([])) = true
    |  empty(stk(_::_)) = false
  fun top(stk(h::_)) = h
  fun pop(stk(_::t)) = stk(t)
  fun push(v, stk(L)) =
    stk(v::L)
end
```

### Local Definitions

Local value and function definitions, not to be exported to users of the type, can be created using the local definition mechanism described earlier:

```ml
local
  val and fun definitions in exported definitions
end;
```
abstype 'a stack =
  stk of 'a list
with
  local
    fun size(stk(L))=length(L);
  in
    val Null = stk([])
    fun empty(s) =
      (size(s) = 0)
    fun top(stk(h::_)) = h
    fun pop(stk(_::t)) = stk(t)
    fun push(v,stk(L)) =
      stk(v::L)
  end
end

val Null = - : 'a stack
val empty = fn : 'a stack -> bool
val top = fn : 'a stack -> 'a
val pop = fn :
  'a stack -> 'a stack
val push = fn :
  'a * 'a stack -> 'a stack

Why are abstract data types useful?
Because they hide an implementation
of a type from a user, allowing
implementation changes without any
impact on user programs.

Consider a simple implementation of
queues:
abstype 'a queue =
  q of 'a list
with
  val Null = q([])
  fun front(q(h::_)) = h
  fun rm(q(_::t)) = q(t)
  fun enter(v,q(L)) =
    q(rev(v::rev(L)))
end

type 'a queue
val Null = - : 'a queue
val front = fn : 'a queue -> 'a
val rm =
  fn : 'a queue -> 'a queue
val enter =
  fn : 'a * 'a queue -> 'a queue

This implementation of queues is
valid, but somewhat inefficient. In
particular to enter a new value onto
the rear end of a queue, we do the
following:
fun enter(v,q(L)) =
  q(rev(v::rev(L)))

We reverse the list that implements
the queue, add the new value to the
head of the reversed queue then
reverse the list a second time.

A more efficient (but less obvious)
implementation of a queue is to store
it as two lists. One list represents the
"front" of the queue. It is from this
list that we extract the front value,
and from which we remove elements.
The other list represents the "back" of
the queue (in reversed order). We add
elements to the rear of the queue by
adding elements to the front of the
list. From time to time, when the
front list becomes null, we "promote"
the rear list into the front list (by
reversing it). Now access to both the
front and the back of the queue is
fast and direct. The new
implementation is:
abstype 'a queue =
  q of 'a list * 'a list
with
  val Null = q([],[])
  fun front(q(h::_,_)) = h |
  front(q([],L)) =
    front(q(rev(L),[]))
  fun rm(q(_::t,L)) = q(t,L) |
  rm(q([],L)) =
    rm(q(rev(L),[]))
  fun enter(v,q(L1,L2)) =
    q(L1,v::L2)
end

type 'a queue
val Null = - : 'a queue
val front = fn :
  'a queue -> 'a
val rm = fn :
  'a queue -> 'a queue
val enter = fn :
  'a * 'a queue -> 'a queue

From the user’s point of view, the two implementations are identical (they export exactly the same set of values and functions). Hence the new implementation can replace the old implementation without any impact at all to the user (except, of course, performance!).

Exception Handling

Our definitions of stacks and queues are incomplete. Reconsider our definition of stack:
abstype 'a stack =
  stk of 'a list
with
  val Null = stk([])
  fun empty(stk([])) = true |
  empty(stk(_::_)) = false
  fun top(stk(h::_)) = h
  fun pop(stk(_::t)) = stk(t)
  fun push(v,stk(L)) =
    stk(v::L)
end

What happens if we evaluate

  top(Null);

We get a “match failure” because our definition of top is incomplete!

In ML we can raise an exception if an illegal or unexpected operation occurs. Asking for the top of an empty stack ought to raise an exception since the requested value does not exist.

ML contains a number of predefined exceptions, including

Match  Empty  Div  Overflow
(exception names by convention begin with a capital letter).

Predefined exception are raised by illegal values or operations. If they are not caught, the run-time system prints an error message.
fun \texttt{f}(1) = 2;
val \texttt{f} = \texttt{fn} : \texttt{int} \to \texttt{int}
\texttt{f}(2);
uncaught exception nonexhaustive
match failure
hd [];
uncaught exception Empty
1000000*1000000;
uncaught exception overflow
(1 \texttt{div} 0);
uncaught exception divide by zero
1.0/0.0;
val \texttt{it} = \texttt{inf} : \texttt{real}
\begin{center}(\texttt{inf} is the IEEE floating-point
standard “infinity” value)\end{center}

\textbf{User Defined Exceptions}

New exceptions may be defined as
exception \texttt{name};
or
exception \texttt{name of type};
\textbf{For example}
exception \texttt{IsZero};
exception \texttt{IsZero}
exception \texttt{NegValue of real};
exception \texttt{NegValue of real}

\textbf{Exceptions May be Raised}

The \texttt{raise} statement raises (throws) an exception:
raise \texttt{exceptionName};
or
raise \texttt{exceptionName(expr)};
\textbf{For example}
fun \texttt{divide}(a,0) = raise \texttt{IsZero}
| \texttt{divide}(a,b) = a \texttt{div} b;
val \texttt{divide} =
\texttt{fn} : \texttt{int} \times \texttt{int} \to \texttt{int}
divide(10,3);
val \texttt{it} = 3 : \texttt{int}
divide(10,0);
uncaught exception \texttt{IsZero}

val \texttt{sqrt} = \texttt{Real.Math.sqrt};
val \texttt{sqrt} = \texttt{fn} : \texttt{real} \to \texttt{real}
\textbf{fun \texttt{sqroot}(x) =}
\begin{center}if \texttt{x} < 0.0
\texttt{then} raise \texttt{NegValue(x)}
\texttt{else} \texttt{sqrt(x)};\end{center}
val \texttt{sqroot} = \texttt{fn} : \texttt{real} \to \texttt{real}
\texttt{sqroot}(2.0);
val \texttt{it} = 1.41421356237 : \texttt{real}
\texttt{sqroot}(-2.0);
uncaught exception \texttt{NegValue}
Exception Handlers

You may catch an exception by defining a handler for it:

```
(expr) handle exception1 => val1
    || exception2 => val2
    || ... ;
```

For example,

```
(sqroot ~100.0)
handle NegValue(v) =>
    (sqrt (~v));
val it = 10.0 : real
```

Stacks Revisited

We can add an exception, EmptyStk, to our earlier stack type to handle top or pop operations on an empty stack:

```
abstype 'a stack = stk of 'a list
with
val Null = stk(\[
exception EmptyStk
fun empty(stk([\[])) = true
    | empty(stk(_::_)) = false
fun top(stk(h::_)) = h
    | top(stk([\[])) =
        raise EmptyStk
fun pop(stk(_::t)) = stk(t)
    | pop(stk([\[])) =
        raise EmptyStk
fun push(v,stk(L)) =
    stk(v::L)
end
```

User-Defined Operators

SML allows users to define symbolic operators composed of non-alphanumeric characters. This means operator-like symbols can be created and used. Care must be taken to avoid predefined operators (like +, -, ^, @, etc.).

If we wish, we can redo our stack definition using symbols rather than identifiers.

We might choose the following symbols:

```
top  |=
pop  <==
push =>
null <@
empty <?>
```
Now we can have expressions like

```ml
<?> <@>;
val it = true : bool
|= (==> (1,<@>));
val it = 1 : int
```

Binary functions, like `==>` (push) are much more readable if they are infix. That is, we'd like to be able to write

```ml
1 ==> 2+3 ==> <@
```

which pushes `2+3`, then 1 onto an empty stack.

To make a function (either identifier or symbolic) infix rather than prefix we use the definition

```ml
infix level name
```

or

```ml
infixr level name
```

level is an integer representing the “precedence” level of the infix operator. 0 is the lowest precedence level; higher precedence operators are applied before lower precedence operators (in the absence of explicit parentheses).

**infix** defines a left-associative operator (groups from left to right).

**infixr** defines a right-associative operator (groups from right to left).

Thus

```ml
fun cat(L1,L2) = L1 @ L2;
infix 5 cat
```

makes `cat` a left associative infix operator at the same precedence level as `@`. We can now write

```ml
[1,2] cat [3,4,5] cat [6,7];
val it = [1,2,3,4,5,6,7] : int list
```

The standard predefined operators have the following precedence levels:

<table>
<thead>
<tr>
<th>Level</th>
<th>Operator</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>o</td>
</tr>
<tr>
<td>4</td>
<td>= &lt;&gt; &lt; &gt; &lt;= =&gt;</td>
</tr>
<tr>
<td>5</td>
<td>:: @</td>
</tr>
<tr>
<td>6</td>
<td>+ - ^</td>
</tr>
<tr>
<td>7</td>
<td>* / div mod</td>
</tr>
</tbody>
</table>

If we define `==>` (push) as

```ml
infixr 2 ==> |
```

then

```ml
1 ==> 2+3 ==> <@
```

will work as expected, evaluating expressions like `2+3` before doing any pushes, with pushes done right to left.

```ml
abstype 'a stack = stk of 'a list
with
val <@> = stk([])
exception emptyStk
fun ?>(stk([])) = true
    | ?>(stk(_::_)) = false
fun |=(stk(h::_)) = h
    | |=(stk([])) = raise emptyStk
fun <==(stk(_::t)) = stk(t)
    | <==(stk([])) = raise emptyStk
fun ==>v,stk(L)) = stk(v::L)
infixr 2 ==> |
```

end
type 'a stack
val <@> = - : 'a stack
exception emptyStk
val ?> = fn : 'a stack -> bool
val |# = fn : 'a stack -> 'a
val <=# = fn :
    'a stack -> 'a stack
val >=# = fn :
    'a stack -> 'a stack
infixr 2 ==#

Now we can write
val myStack =
    1 ==# 2+3 ==# <@>;
val myStack = - : int stack
|# myStack;
val it = 1 : int
|# (<=# myStack);
val it = 5 : int

Thus
dupl op ^ "abc";
val it = "abcabc" : string
works fine.

Using Infix Operators as Values
Sometimes we simply want to use an infix operator as a symbol whose value is a function.
For example, given
fun dupl f v = f(v,v);
val dupl =
    fn : ('a * 'a -> 'b) -> 'a -> 'b
we might try the call
dupl ^ "abc";
This fails because SML tries to parse dupl and "abc" as the operands of ^.
To pass an operator as an ordinary function value, we prefix it with op which tells the SML compiler that the following symbol is an infix operator.

The Case Expression
ML contains a case expression patterned on switch and case statements found in other languages.
As in function definitions, patterns are used to choose among a variety of values.
The general form of the case is
case expr of
    pattern_1 => expr_1 |
    pattern_n => expr_n |
    ...
    pattern_n => expr_n;
If no pattern matches, a Match exception is thrown.
It is common to use _ (the wildcard) as the last pattern in a case.
Examples include

```ml
case c of
  red   => "rot" |
  blue  => "blau" |
  green => "gruen";

case pair of
  (1,_) => "win" |
  (2,_) => "place" |
  (3,_) => "show" |
  (_,_) => "loser";

case intOption of
  none => 0 |
  some(v) => v;
```

Imperative Features of ML

ML provides references to heap locations that may be updated. This is essentially the same as access to heap objects via references (Java) or pointers (C and C++).

The expression

```
ref val
```

creates a reference to a heap location initialized to val. For example,

```
ref 0;
val it = ref 0 : int ref
```

The prefix operator `!` fetches the value contained in a heap location (just as `*` dereferences a pointer in C or C++).

Thus

```
! (ref 0);
val it = 0 : int
```

Sequential Composition

Expressions or statements are sequenced using ";;". Hence

```
val a = (1+2;3+4);
val a = 7 : int
(x:=1;!x);
val it = 1 : int
```

Iteration

```
while expr1 do expr2
```

implements iteration (and returns unit); Thus

```
(while false do 10);
val it = () : unit
while !x > 0 do x:= !x-1;
val it = () : unit
!x;
val it = 0 : int
```
Simple I/O

The function

\[
\text{print;} \quad \text{val it = fn : string -> unit}
\]

prints a string onto standard output. For example,

\[
\text{print("Hello World
");}
\]

Hello World

The conversion routines

\[
\text{Real.toString;} \quad \text{val it = fn : real -> string}
\]

\[
\text{Int.toString;} \quad \text{val it = fn : int -> string}
\]

\[
\text{Bool.toString;} \quad \text{val it = fn : bool -> string}
\]

cat a value (real, int or bool) into a string. Unlike Java, the call must be explicit.

For example,

\[
\text{print(\text{Int.toString}(123));}
\]

123

Also available are

\[
\text{Real.fromString;} \quad \text{val it = fn : string -> real option}
\]

\[
\text{Int.fromString;} \quad \text{val it = fn : string -> int option}
\]

\[
\text{Bool.fromString;} \quad \text{val it = fn : string -> bool option}
\]

which convert from a string to a real or int or bool if possible. (That's why the option type is used).

For example,

\[
\text{case (\text{Int.fromString}("123");}
\]

\[
\text{SOME(i) => i | NONE => 0;}
\]

val it = 123 : int

\[
\text{case (\text{Int.fromString("One two three")}) of}
\]

\[
\text{SOME(i) => i | NONE => 0;}
\]

val it = 0 : int

Text I/O

The structure TextIO contains a wide variety of I/O types, values and functions. You load these by entering:

\[
\text{open TextIO;}
\]

Among the values loaded are

- type instream
  This is the type that represents input text files.

- type outstream
  This is the type that represents output text files.

- type vector = string
  Makes vector a synonym for string.

- type elem = char
  Makes elem a synonym for char.
Predefined input and output streams.

- val stdIn : instream
- val stdOut : outstream
- val stdErr : outstream

Open an input or output stream.

- val openIn : string -> instream
- val openOut : string -> outstream

For example,

```
val out = openOut("/tmp/test1");
val out = -- : outstream
```

Read a line of input into a string.

- val input : instream -> vector

For example (user input is in red):

```
val s = input(stdIn);
Hello!
val s = "Hello!\n" : vector
```

Read the next \( N \) input characters into a string.

- val inputN : instream * int -> vector

For example,

```
val t = inputN(stdIn,3);
abcde
val t = "abc" : vector
```

Read the rest of the input file into a string (with newlines separating lines).

- val inputAll : instream -> vector

For example,

```
val u = inputAll(stdIn);
Four score and
seven years ago ...
val u = "Four score and\nseven years ago ...\n" : vector
```

Are we at the end of this input stream?

- val endOfStream : instream -> bool

Output a string on the specified output stream.

- val output : outstream * vector -> unit

```
val s = input(stdIn);
Hello!
val s = "Hello!\n" : vector
```

String Operations
ML provides a wide variety of string manipulation routines. Included are:

- The string concatenation operator, \(^\)
  
  "abc" \(^\) "def" = "abcdef"

- The standard 6 relational operators:
  
  \(<\) \(\leq\) \(\geq\) \(=\) \(<\) \(\geq\)

- The string size operator:

  \(\text{val size : string} \rightarrow \text{int}\)

  val size = "abcd"

  val it = 4 : int

- The string subscripting operator

  (indexing from 0):

  \(\text{val sub = fn : string} \rightarrow \text{char}\)

  val sub("abcde",2);

  val it = 

• The substring function
val substring :
string * int * int -> string
This function is called as
substring(string,start,len)
start is the starting position,
counting from 0.
len is the length of the desired
substring. For example,
substring("abcdefghij",3,4)
val it = "defg" : string

• Concatenation of a list of strings into
a single string:
concat :
string list -> string
For example,
concat ["What’s"," up","?"];
val it = "What’s up?" : string

• Convert a character into a string:
str : char -> string
For example,
str(#"x")
val it = "x" : string

• “Explode” a string into a list of
characters:
explode : string -> char list
For example,
explode("abcde");
val it = [#"a",[#"b",[#"c",[#"d",[#"e"]]]] : char list

• “Implode” a list of characters into a
string.
implode : char list -> string
For example,
implode [#"a",[#"b",[#"c",[#"d",[#"e"]]]];
val it = "abcde" : string

Structures and Signatures

In C++ and Java you can group
variable and function definitions into
classes. In Java you can also group
classes into packages.
In ML you can group value, exception
and function definitions into
structures.
You can then import selected
definitions from the structure (using
the notation structure.name) or
you can open the structure, thereby
importing all the definitions within
the structure.
(Examples used in this section may be
found at
~cs538-1/public/sml/struct.sml)

The general form of a structure
definition is
structure name =
struct
val, exception and
fun definitions
end
For example,
structure Mapping =
struct
exception NotFound;
val create = []; fun lookup(key,[]) = raise NotFound |
lookup(key,
(key1,value1)::rest) =
if key = key1
then value1
else lookup(key,rest);
fun insert(key, value, []) = [(key, value)]
  | insert(key, value, (key1, value1)::rest) = 
    if key = key1
    then (key, value)::rest
    else (key1, value1)::
      insert(key, value, rest);
end;

We can access members of this structure as Mapping.name. Thus
Mapping.insert(538, "languages", []);
val it = [(538, "languages")]:
(int * string) list
open Mapping;
exception NotFound
val create : 'a list
val insert : 'a * 'b * ('a * 'b) list -> ('a * 'b) list
val lookup : 'a * ('a * 'b) list -> 'b

You can define a signature as
signature name = sig
  type definitions for values, functions and exceptions
end
For example,
signature Str2IntMapping =
sig
  exception NotFound;
  val lookup: string * (string * int) list -> int;
end;

Signatures

Each structure has a signature, which is its type.

For example, Mapping’s signature is
structure Mapping :
sig
  exception NotFound
  val create : 'a list
  val insert : 'a * 'b * ('a * 'b) list -> ('a * 'b) list
  val lookup : 'a * ('a * 'b) list -> 'b
end

Signatures can be used to
• Restrict the type of a value or function in a structure.
• Hide selected definitions that appear in a structure
For example
structure Str2IntMap :
  Str2IntMapping = Mapping;
defines a new structure, Str2IntMap, created by restricting Mapping to the
Str2IntMapping signature. When we do this we get
open Str2IntMap;
  exception NotFound
  val lookup : string * (string * int) list -> int

Only lookup and NotFound are created, and lookup is limited to keys that are strings.
Extending ML’s Polymorphism

In languages like C++ and Java we must use types like `void*` or `Object` to simulate the polymorphism that ML provides. In ML whenever possible a general type (a polytype) is used rather than a fixed type. Thus in

```ml
fun len([]) = 0
  | len(a::b) = 1 + len(b);
```

we get a type of

`'a list -> int`

because this is the most general type possible that is consistent with `len`’s definition.

Is this form of polymorphism general enough to capture the general idea of making program definitions as type-independent as possible?

It isn’t, and to see why consider the following ML definition of a merge sort. A merge sort operates by first splitting a list into two equal length sublists. The following function does this:

```ml
fun split [] = ([], [])
  | split [a] = ([a], [])
  | split (a::b::rest) = let val (left, right) = split(rest) in (a::left, b::right) end;
```

After the input list is split into two halves, each half is recursively sorted, then the sorted halves are merged together into a single list.

The following ML function merges two sorted lists into one:

```ml
fun merge([], []) = []
  | merge([], hd::tl) = hd::tl
  | merge(hd::tl, []) = hd::tl
  | merge(hd::tl, h::t) = if hd <= h
    then hd::merge(tl, h::t)
    else h::merge(hd::tl, t)
```

This definition looks very general—it should work for a list of any type. Unfortunately, when ML types the functions we get a surprise:

```ml
val split = fn : 'a list -> 'a list * 'a list
val merge = fn : int list * int list -> int list
val sort = fn : int list -> int list
```

`split` is polymorphic, but `merge` and `sort` are limited to integer lists!

Where did this restriction come from?
The problem is that we did a comparison in `merge` using the `<=` operator, and ML typed this as an integer comparison.

We can make our definition of sort more general by adding a comparison function, `le(a,b)` as a parameter to `merge` and `sort`. If we curry this parameter we may be able to hide it from end users. Our updated definitions are:

```ml
fun merge(le, [], []) = []
  | merge(le, [], hd::tl) = hd::tl
  | merge(le, hd::tl, []) = hd::tl
  | merge(le, hd::tl, h::t) = 
    if le(hd, h)
    then hd::merge(le, tl, h::t)
    else h::merge(le, hd::tl, t)

fun sort le [] = []
  | sort le [a] = [a]
  | sort le (a::b::rest) = 
    let val (left, right) = 
      split(a::b::rest) in 
      merge(le, sort le left, sort le right) 
    end;
```

Now the types of `merge` and `sort` are:

```ml
val merge = fn : ('a * 'a -> bool) * 'a list * 'a list -> 'a list
val sort = fn : ('a * 'a -> bool) -> 'a list -> 'a list
```

We can now “customize” `sort` by choosing a particular definition for the `le` parameter:

```ml
fun le(a,b) = a <= b;
val le = fn : int * int -> bool
```

```ml
fun intsort L = sort le L;
val intsort = 
  fn : int list -> int list 
intsort(
  [4,9,0,2,111,~22,8,~123]);
val it = 
  [-123,~22,0,2,4,8,9,111] : int list

fun strle(a:string,b) = a <= b;
val strle = 
  fn : string * string -> bool 
strle(
  ["aac","aaa","ABC","123"]);
val it = 
  ["123","ABC","aaa","aac"] : string list
```

Making the comparison relation an explicit parameter works, but it is a bit ugly and inefficient. Moreover, if we have several functions that depend on the comparison relation, we need to ensure that they all use the same relation. Thus if we wish to define a predicate `inOrder` that tests if a list is already sorted, we can use:

```ml
fun inOrder le [] = true
  | inOrder le [a] = true
  | inOrder le (a::b::rest) = 
    le(a,b) andalso
    inOrder le (b::rest);
val inOrder = fn : ('a * 'a -> bool) -> 'a list -> bool
```

Now `sort` and `inOrder` need to use the same definition of `le`. But how can we enforce this?
The structure mechanism we studied earlier can help. We can put a single definition of \( \text{le} \) in the structure, and share it:

```
structure Sorting =
  struct
    fun le(a,b) = a <= b;

    fun split [] = ([],[])
    | split [a] = ([a],[]) 
    | split (a::b::rest) = 
        let val (left,right) = 
            split rest in 
            (a::left,b::right)
        end;
    
    fun merge([],[]) = []
    | merge([],hd::tl) = hd::tl
    | merge(hd::tl,[]) = hd::tl
    | merge(hd::tl,h::t) = 
        if le(hd,h) 
        then hd::merge(tl,h::t)
        else h::merge(hd::tl,t)
  end;

fun sort [] = []
| sort([a]) = [a]
| sort(a::b::rest) = 
    let val (left,right) = 
        split(a::b::rest) in 
        merge(sort(left),
               sort(right))
    end;

fun inOrder [] = true
| inOrder [a] = true
| inOrder (a::b::rest) = 
    le(a,b) andalso 
    inOrder (b::rest);
end;
structure Sorting :
  sig
    val inOrder : int list -> bool
    val le : int * int -> bool
    val merge : int list * 
               int list -> int list
    val sort : 
              int list -> int list
  end
```

To sort a type other than integers, we replace the definition of \( \text{le} \) in the structure. But rather than actually edit that definition, ML gives us a powerful mechanism to parameterize a structure. This is the functor, which allows us to use one or more structures as parameters in the definition of a structure.

---

**Functors**

The general form of a functor is

```
functor name
  (structName:signature) =
  structure definition;
```

This functor will create a specific version of the structure definition using the structure parameter passed to it.

For our purposes this is ideal—we pass in a structure defining an ordering relation (the \( \text{le} \) function). This then creates a custom version of all the functions defined in the structure body, using the specific \( \text{le} \) definition provided.
We first define
signature Order =
  sig
    type elem
    val le : elem*elem -> bool
  end;
This defines the type of a structure that defines a le predicate defined on a pair of types called elem.
An example of such a structure is
structure IntOrder:Order =
  struct
    type elem = int;
    fun le(a,b) = a <= b;
  end;
Now we just define a functor that creates a Sorting structure based on an Order structure:

functor MakeSorting(O:Order) =
  struct
    open O; (* makes le available*)
    fun split [] = ([],[])
    |  split [a] = ([a],[])
    |  split (a::b::rest) =
      let val (left,right) =
        split rest in
        (a::left,b::right)
    end;
    fun merge([],[]) = []
    |  merge([],hd::tl) = hd::tl
    |  merge(hd::tl,[]) = hd::tl
    |  merge(hd::tl,h::t) =
      if le(hd,h)
      then hd::merge(tl,h::t)
      else h::merge(hd::tl,t)
  end;

fun sort [] = []
  |  sort([a]) = [a]
  |  sort(a::b::rest) =
    let val (left,right) =
      split(a::b::rest) in
      merge(sort(left),
            sort(right))
    end;

fun inOrder [] = true
  |  inOrder [a] = true
  |  inOrder (a::b::rest) =
    le(a,b) andalso
    inOrder (b::rest);
  end;

functor MakeSorting(O:Order) =
  struct
    open O; (* makes le available*)
    fun split [] = ([],[])
    |  split [a] = ([a],[])
    |  split (a::b::rest) =
      let val (left,right) =
        split rest in
        (a::left,b::right)
    end;
    fun merge([],[]) = []
    |  merge([],hd::tl) = hd::tl
    |  merge(hd::tl,[]) = hd::tl
    |  merge(hd::tl,h::t) =
      if le(hd,h)
      then hd::merge(tl,h::t)
      else h::merge(hd::tl,t)
  end;

fun sort [] = []
  |  sort([a]) = [a]
  |  sort(a::b::rest) =
    let val (left,right) =
      split(a::b::rest) in
      merge(sort(left),
            sort(right))
    end;

fun inOrder [] = true
  |  inOrder [a] = true
  |  inOrder (a::b::rest) =
    le(a,b) andalso
    inOrder (b::rest);
  end;

Now
structure IntSorting =
  MakeSorting(IntOrder);
creates a custom structure for sorting integers:
  IntSorting.sort [3,0,~22,8];
  val it = [~22,0,3,8] : elem list
To sort strings, we just define a structure containing an le defined for strings with Order as its signature (i.e., type) and pass it to MakeSorting:
structure StrOrder:Order =
  struct
    type elem = string
    fun le(a:string,b) = a <= b;
  end;
structure StrSorting = 
  MakeSorting(StrOrder);
StrSorting.sort(
  ["cc","abc","xyz"]);
val it = ["abc","cc","xyz"] :
  StrOrder.elem list
StrSorting.inOrder(
  ["cc","abc","xyz"]);
val it = false : bool
StrSorting.inOrder(
  [3,0,~22,8]);
stdin:593.1-593.32 Error:  
operator and operand don't agree
[literal]  
operator domain: strOrder.elem list  
operand:         int list  
in expression:  
  StrSorting.inOrder (3 :: 0 :: 
~22 :: <exp> :: <exp>)

The SML Basis Library

SML provides a wide variety of useful types and functions, grouped into structures, that are included in the Basis Library.

A web page fully documenting the Basis Library is linked from the ML page that is part of the Programming Languages Links page on the CS 538 home page.

Many useful types, operators and functions are “preloaded” when you start the SML compiler. These are listed in the “Top-level Environment” section of the Basis Library documentation.

Many other useful definitions must be explicitly fetched from the structures they are defined in.

For example, the Math structure contains a number of useful mathematical values and operations. You may simply enter

open Math;
while will load all the definitions in Math. Doing this may load more definitions than you want. What’s worse, a definition loaded may redefine a definition you currently want to stay active. (Recall that ML has virtually no overloading, so functions with the same name in different structures are common.)

A more selective way to access a definition is to qualify it with the structure’s name. Hence

Math.pi;
val it = 3.14159265359 : real

gets the value of pi defined in Math. Should you tire of repeatedly qualifying a name, you can (of course) define a local value to hold its value. Thus

val pi = Math.pi;
val pi = 3.14159265359 : real

works fine.
An Overview of Structures in the Basis Library

The Basis Library contains a wide variety of useful structures. Here is an overview of some of the most important ones.

- Option
  Operations for the option type.
- Bool
  Operations for the bool type.
- Char
  Operations for the char type.
- String
  Operations for the string type.
- Byte
  Operations for the byte type.
- Int
  Operations for the int type.
- IntInf
  Operations for an unbounded precision integer type.
- Real
  Operations for the real type.
- Math
  Various mathematical values and operations.
- List
  Operations for the list type.
- ListPair
  Operations on pairs of lists.
- Vector
  A polymorphic type for immutable (unchangeable) sequences.
- IntVector, RealVector, BoolVector, CharVector
  Monomorphic types for immutable sequences.
- Array
  A polymorphic type for mutable (changeable) sequences.
- IntArray, RealArray, BoolArray, CharArray
  Monomorphic types for mutable sequences.
- Array2
  A polymorphic 2 dimensional mutable type.
- IntArray2, RealArray2, BoolArray2, CharArray2
  Monomorphic 2 dimensional mutable types.
- TextIO
  Character-oriented text IO.
- BinIO
  Binary IO operations.
- OS, Unix, Date, Time, Timer
  Operating systems types and operations.

ML Type Inference

One of the most novel aspects of ML is the fact that it infers types for all user declarations.

How does this type inference mechanism work?

Essentially, the ML compiler creates an unknown type for each declaration the user makes. It then solves for these unknowns using known types and a set of type inference rules. That is, for a user-defined identifier $i$, ML wants to determine $T(i)$, the type of $i$. 
The type inference rules are:

1. The types of all predefined literals, constants and functions are known in advance. They may be looked-up and used. For example,
   
   \(2 : \text{int}\)
   
   \(\text{true} : \text{bool}\)
   
   \([] : \text{'a list}\)
   
   \(:: : \text{'a *'a list -> 'a list}\)

2. All occurrences of the same symbol (using scoping rules) have the same type.

3. In the expression
   
   \(I = J\)
   
   we know \(T(I) = T(J)\).

4. In a conditional
   
   \((\text{if } E_1 \text{ then } E_2 \text{ else } E_3)\)
   
   we know that
   
   \(T(E_1) = \text{bool},\)
   
   \(T(E_2) = T(E_3) = T(\text{conditional})\)

5. In a function call
   
   \((f x)\)
   
   we know that if \(T(f) = \text{'a -> 'b}\)
   
   then \(T(x) = \text{'a}\) and \(T(f x) = \text{'b}\)

6. In a function definition
   
   \(\text{fun } f x = \text{expr; if } t(x) = \text{'a} \text{ and } T(\text{expr}) = \text{'b}\)
   
   then \(T(f) = \text{'a -> 'b}\)

7. In a tuple \((e_1, e_2, \ldots, e_n)\)
   
   if we know that \(T(e_i) = \text{'a}_i 1 \leq i \leq n\)
   
   then \(T(e_1, e_2, \ldots, e_n) = \text{'a}_1 * \text{'a}_2 * \ldots * \text{'a}_n\)

8. In a record
   
   \{\ a = e_1 , b = e_2 , \ldots \}\)
   
   if \(T(e_1) = \text{'a}_1 1 \leq i \leq n\) then
   
   the type of the record =
   
   \{a:'a_1 , b:'a_2 , \ldots\}

9. In a list \([v_1 , v_2 , \ldots, v_n]\)
   
   if we know that \(T(v_1) = \text{'a}_1 1 \leq i \leq n\)
   
   then we know that
   
   \(\text{'a}_1 = \text{'a}_2 = \ldots = \text{'a}_n\) and
   
   \(T([v_1 , v_2 , \ldots, v_n]) = \text{'a}_1 \text{ list}\)

To Solve for Types:

1. Assign each untyped symbol its own distinct type variable.
2. Use rules (1) to (9) to solve for and simplify unknown types.
3. Verify that each solution “works” (causes no type errors) throughout the program.

Examples

Consider

\(\text{fun fact(n)} =\)

\(\text{if n=1 then 1 else n*fact(n-1);}\)

To begin, we’ll assign type variables:

\(T(\text{fact}) = \text{'a} \rightarrow \text{'b}\)

(fact is a function)

\(T(n) = \text{'c}\)
Now we begin to solve for the types 'a, 'b and 'c must represent. We know (rule 5) that 'c = 'a since n is the argument of fact. We know (rule 3) that 'c = T(1) = int since n=1 is part of the definition. We know (rule 4) that T(1) = T(if expression)='b since the if expression is the body of fact. Thus, we have

'c = 'a = 'b = 'c = int, so 
T( fact) = int -> int
T( n) = int

These types are correct for all occurrences of fact and n in the definition.

A Polymorphic Function:

fun leng(L) =
  if L = []
  then 0
  else 1+len(tl L);
To begin, we know that
T([]) = 'a list and 
T(tl) = 'b list -> 'b list
We assign types to leng and L:
T(leng) = 'c -> 'd
T(L) = 'e
Since L is the argument of leng, 'e = 'c
From the expression L=[] we know 'e = 'a list

Type Inference for Patterns

Type inference works for patterns too.
Consider

fun leng [] = 0
    | leng (a::b) = 1 + leng b;
We first create type variables:
T(leng) = 'a -> 'b
T(a) = 'c
T(b) = 'd
From leng [] we conclude that
'a = 'e list
From leng [] = 0 we conclude that
'b = int
From leng (a::b) we conclude that
'c = 'e and 'd = 'e list
Thus we have
T(leng) = 'e list -> int
T(a) = 'e
T(b) = 'e list
This solution is type correct throughout the definition.

Not Everything can be Automatically Typed in ML

Let’s try to type

```ml
fun f x = (x x);
```

We assume

T(f) = 'a -> 'b
t(x) = 'c

Now (as usual) 'a = 'c since x is the argument of f.

From the call (x x) we conclude that 'c must be of the form 'd -> 'e (since x is being used as a function).
Moreover, 'c = 'd since x is an argument in (x x).
Thus 'c = 'd -> 'e = 'c -> 'e.
But 'c = 'c -> 'e has no solution, so in ML this definition is invalid. We
can’t pass a function to itself as an argument—the type system doesn’t allow it.
In Scheme this is allowed:

```scheme
(define (f x) (x x))
```

but a call like

(f f)
certainly doesn’t do anything good!

Type Unions

Let’s try to type

```ml
fun f g = ((g 3), (g true));
```

Now the type of g is 'a -> 'b since g is used as a function.
The call (g 3) says 'a = int and the call (g true) says 'a = boolean.
Does this mean g is polymorphic?
That is, is the type of f
f : ('a->'b)->'b*'b?
NO!
All functions have the type 'a -> 'b but not all functions can be passed to f.

Consider not: bool->bool.
The call (not 3) is certainly illegal.
What we’d like in this case is a union type. That is, we’d like to be able to type \( g \) as \((\text{int}|\text{bool})\rightarrow\text{'b}\) which ML doesn’t allow.

Fortunately, ML does allow type constructors, which are just what we need.

Given

```plaintext
datatype T =
  I of int|B of bool;
```

we can redefine \( f \) as

```plaintext
fun f g =
  (g (I(3)), g (B(true)));
val f = fn : (T -> 'a) -> 'a * 'a
```

Finally, note that in a definition like

```plaintext
let
  val f =
    fn x => x (* id function*)
in (f 3,f true)
end;
```

type inference works fine:

```plaintext
val it = (3,true) : int * bool
```

Here we define \( f \) in advance, so its type is known when calls to it are seen.

---

**Reading Assignment**

- Roosta: Chapters 10 and 11
- Webber: Chapters 19, 20 and 22

---

**Prolog**

Prolog presents a view of programming that is very different from most other programming languages.

A famous text book is entitled “Algorithms + Data Structures = Programs”

This formula represents well the conventional approach to programming that most programming languages support.

In Prolog there is an alternative rule of programming: “Algorithms = Logic + Control”

This rule encompasses a non-procedural view of programming.
Logic (what the program is to compute) comes first. Then control (how to implement the logic) is considered. In Prolog we program the logic of a program, but the Prolog system automatically implements the control. Logic is essential—control is just efficiency.

Logic Programming

Prolog implements logic programming. In fact Prolog means Programming in Logic. In Prolog programs are statements of rules and facts. Program execution is deduction—can an answer be inferred from known rules and facts. Prolog was developed in 1972 by Kowalski and Colmerauer at the University of Marseilles.

Elementary Data Objects

- In Prolog integers and atoms are the elementary data objects.
- Integers are ordinary integer literals and values.
- Atoms are identifiers that begin with a lower-case letter (much like symbolic values in Scheme).
- In Prolog data objects are called terms.
- In Prolog we define relations among terms (integers, atoms or other terms).
- A predicate names a relation. Predicates begin with lower-case letters.
- To define a predicate, we write clauses that define the relation.

- There are two kinds of program clauses, facts and rules.
- A fact is a predicate that prefixes a sequence of terms, and which ends with a period (".").

As an example, consider the following facts which define "fatherOf" and "motherOf" relations.

fatherOf(tom, dick).
fatherOf(dick, harry).
fatherOf(jane, harry).
motherOf(tom, judy).
motherOf(dick, mary).
motherOf(jane, mary).

The symbols fatherOf and motherOf are predicates. The symbols tom, dick, harry, judy, mary and jane are atoms.
Once we have entered rules and facts that define relations, we can make queries (ask the Prolog system questions).

Prolog has two interactive modes that you can switch between.

To enter definition mode (to define rules and facts) you enter

[ user ].

You then enter facts and rules, terminating this phase with ^D (end of file).

Alternatively, you can enter

[ 'filename' ].

to read in rules and facts stored in the file named filename.

When you start Prolog, or after you leave definitions mode, you are in query mode.

In query mode you see a prompt of the form

| ?- or ?- (depending on the system you are running).

In query mode, Prolog allows you to ask whether a relation among terms is true or false.

Thus given our definition of motherOf and fatherOf relations, we can ask:

| ?- fatherOf(tom,dick).

yes

A “yes” response means that Prolog is able to conclude from the facts and rules it has been given that the relation queried does hold.

| ?- fatherOf(georgeW,george).

no

A “no” response to a query means that Prolog is unable to conclude that the relation holds from what it has been told. The relation may actually be true, but Prolog may lack necessary facts or rules to deduce this.

Variables in Queries

One of the attractive features of Prolog is the fact that variables may be included in queries. A variable always begins with a capital letter.

When a variable is seen, Prolog tries to find a value (binding) for the variable that will make the queried relation true.

For example,

fatherOf(X,harry).

asks Prolog to find an value for x such that x's father is harry.

When we enter the query, Prolog gives us a solution (if one can be found):

?- fatherOf(X,harry).

X = dick
If no solution can be found, it tells us so:

```prolog
?- fatherOf(Y, jane).
no
```

Since solutions to queries need not be unique, Prolog will give us alternate solutions if we ask for them. We do so by entering a “;” after a solution is printed. We get a “no” when no more solutions can be found:

```prolog
?- fatherOf(X, harry).
X = dick ;
X = jane ;
no
```

Variables may be placed anywhere in a query. Thus we may ask

```prolog
?- fatherOf(jane, X).
X = harry ;
no
```

We may use more than one variable if we wish:

```prolog
?- fatherOf(X, Y).
X = tom, 
Y = dick ;
X = dick, 
Y = harry ;
X = jane, 
Y = harry ;
no
```

(This query displays all the fatherOf relations).

Conjunction of Goals

More than one relation can be included as the “goal” of a query. A comma (“,”) is used as an AND operator to indicate a conjunction of goals—all must be satisfied by a solution to the query.

```prolog
?- fatherOf(jane, X), motherOf(jane, Y).
X = harry, 
Y = mary ;
no
```

A given variable may appear more than once in a query. The same value of the variable must be used in all places in which the variable appears (this is called unification).

For example,

```prolog
?- fatherOf(tom, X), fatherOf(X, harry).
X = dick ;
no
```
Rules in Prolog

Rules allow us to state that a relation will hold depending on the truth (correctness) of other relations. In effect a rule says, “If I know that certain relations hold, then I also know that this relation holds.”

A rule in Prolog is of the form
\[ \text{rel}_1 :\!-\! \text{rel}_2, \text{rel}_3, \ldots, \text{rel}_n. \]

This says \( \text{rel}_1 \) can be assumed true if we can establish that \( \text{rel}_2 \) and \( \text{rel}_3 \) and all relations to \( \text{rel}_n \) are true. \( \text{rel}_1 \) is called the head of the rule. \( \text{rel}_2 \) to \( \text{rel}_n \) form the body of the rule.

Example

The following two rules define a \text{grandMotherOf} relation using the \text{motherOf} and \text{fatherOf} relations:

\begin{verbatim}
grandMotherOf(X,GM) :-
motherOf(X,M),
motherOf(M,GM).

grandMotherOf(X,GM) :-
fatherOf(X,F),
motherOf(F,GM).
\end{verbatim}

\begin{verbatim}
| ?- grandMotherOf(tom,GM).
GM = mary ;
no
| ?- grandMotherOf(dick,GM).
no
| ?- grandMotherOf(X,mary).
X = tom ;
no
\end{verbatim}

As is the case for all programming, in all languages, you must be careful when you define a rule that it correctly captures the idea you have in mind.

Consider the following rule that defines a sibling relation between two people:

\begin{verbatim}
sibling(X,Y) :-
motherOf(X,M), motherOf(Y,M),
fatherOf(X,F), fatherOf(Y,F).
\end{verbatim}

This rule says that \( X \) and \( Y \) are siblings if each has the same mother and the same father.

But the rule is wrong! Why?

Let’s give it a try:

\begin{verbatim}
| ?- sibling(X,Y).
X = Y = tom
Darn! That’s right, you can’t be your own sibling. So we refine the rule to force \( X \) and \( Y \) to be distinct:

sibling(X,Y) :-
motherOf(X,M), motherOf(Y,M),
fatherOf(X,F), fatherOf(Y,F),
not(X=Y).
(A few Prolog systems use “\( \not+ \)” for not; but most include a not relation.)

| ?- sibling(X,Y).
X = dick,
Y = jane ;
X = jane,
Y = dick ;
no
\end{verbatim}
Note that distinct but equivalent solutions (like \( X = \text{dick}, Y = \text{jane} \) vs. \( X = \text{jane}, Y = \text{dick} \)) often appear in Prolog solutions. You may sometimes need to “filter out” solutions that are effectively redundant (perhaps by formulating stricter or more precise rules).

\[ \text{How Prolog Solves Queries} \]

The unique feature of Prolog is that it automatically chooses the facts and rules needed to solve a query. But how does it make its choice? It starts by trying to solve each goal in a query, left to right (recall goals are connected using “,” which is the and operator). For each goal it tries to match a corresponding fact or the head of a corresponding rule.

A fact or head of rule matches a goal if:

- Both use the same predicate.
- Both have the same number of terms following the predicate.
- Each term in the goal and fact or rule head match (are equal), possibly binding a free variable to force a match.

For example, assume we wish to match the following goal:

\( x(a, B) \)

This can match the fact

\( x(a, b) \).

or the head of the rule

\( x(Y, Z) :- Y = Z. \)

But \( x(a, B) \) can’t match

\( y(a, b) \) (wrong predicate name) or
\( x(b, d) \) (first terms don’t match) or
\( x(a, b, c) \) (wrong number of terms).

If we succeed in matching a rule, we have solved the goal in question; we can go on to match any remaining goals.

If we match the head of a rule, we aren’t done—we add the body of the rule to the list of goals that must be solved.

Thus if we match the goal \( x(a, B) \) with the rule

\( x(Y, Z) :- Y = Z. \)

then we must solve \( a = B \) which is done by making \( B \) equal to \( a \).
Backtracking

If we reach a point where a goal can’t be matched, or the body of a rule can’t be matched, we backtrack to the last (most recent) spot where a choice of matching a particular fact or rule was made. We then try to match a different fact or rule. If this can’t be done, we go back to the next previous place where a choice was made and try a different match there. We try alternatives until we are able to solve all the goals in our query or until all possible choices have been tried and found to fail. If this happens, we answer “no” the query can’t be solved.

As we try to match facts and rules we try them in their order of definition.

Example

Let’s trace how

?- grandMotherOf(tom,GM).

is solved.

Recall that

grandMotherOf(X,GM) :-
    motherOf(X,M),
    motherOf(M,GM).

grandMotherOf(X,GM) :-
    fatherOf(X,F),
    motherOf(F,GM).

fatherOf(tom,dick).
fatherOf(dick,harry).
fatherOf(jane,harry).
motherOf(tom,judy).
motherOf(dick,mary).
motherOf(jane,mary).

We try the first grandMotherOf rule first.
This forces X = tom. We have to solve
    motherOf(tom,M),
    motherOf(M,GM).

We now try to solve
    motherOf(tom,M)
This forces M = judy.
We then try to solve
    motherOf(judy,GM)
None of the motherOf rules match this goal, so we backtrack. No other motherOf rule can solve
    motherOf(tom,M)
so we backtrack again and try the second grandMotherOf rule:

grandMotherOf(X,GM) :-
    fatherOf(X,F),
    motherOf(F,GM).

This matches, forcing X = tom.
We have to solve
    fatherOf(tom,F),
    motherOf(F,GM).
We can match the first goal with
    fatherOf(tom,dick).
This forces F = dick.
We then must solve
    motherOf(dick,GM)
which can be matched by
    motherOf(dick,mary).
We have matched all our goals, so we know the query is true, with
    GM = mary.
List Processing in Prolog

Prolog has a notation similar to “cons cells” of Lisp and Scheme.

The “.” functor (predicate name) acts like cons.

Hence .(a, b) in Prolog is essentially the same as (a . b) in Scheme.

Lists in Prolog are formed much the same way as in Scheme and ML:

[] is the empty list

[1, 2, 3] is an abbreviation for .(1, .(2, .(3, [])))

just as (1, 2, 3) in Scheme is an abbreviation for

(cons 1 (cons 2 (cons 3 [])))

The notation [H|T] represents a list with H matching the head of the list and T matching the rest of the list.

Thus [1, 2, 3] ≡ [1| [2, 3]] ≡ [1, 2| [3]] ≡ [1, 2, 3| []]

As in ML, “_” (underscore) can be used as a wildcard or “don’t care” symbol in matches.

Given the fact

p([1, 2, 3, 4]).

The query

?- p([X|Y]).

answers

X = 1,
Y = [2, 3, 4]

The query

p([_, _, X|Y]).

answers

X = 3,
Y = [4]

List Operations in Prolog

List operations are defined using rules and facts. The definitions are similar to those used in Scheme or ML, but they are non-procedural.

That is, you don’t give an execution order. Instead, you give recursive rules and non-recursive “base cases” that characterize the operation you are defining.

Consider append:

append([], L, L).

append([H|T1], L2, [H|T3]) :- append(T1, L2, T3).

The first fact says that an empty list (argument 1) appended to any list L (argument 2) gives L (argument 3) as its answer.
The rule in line 2 says that if you take a list that begins with \( H \) and has \( T_1 \) as the rest of the list and append it to a list \( L \) then the resulting appended list will begin with \( H \).

Moreover, the rest of the resulting list, \( T_3 \), is the result of appending \( T_1 \) (the rest of the first list) with \( L_2 \) (the second input list).

The query

\[
?- \text{append([1], [2, 3], [1, 2, 3]).}
\]

answers

\text{Yes}

because with \( H=1, T_1=[], L_2=[2, 3] \) and \( T_3=[2, 3] \) it must be the case that \( \text{append([1], [2, 3], [2, 3])} \) is true and fact (1) says that this is so.

How does it choose this answer?
First Prolog tries to match the query against fact (1) or rule (2).
Fact (1) doesn’t match (the first arguments differ) so we match rule (2).
This gives us \( H=1, T_1=[], L_2=x \) and \( T_3=[2, 3] \).
We next have to solve the body of rule (2) which is

\[
\text{append([], L_2, [2, 3])}.
\]

Fact (1) matches this, and tells us that \( L_2=[2, 3]=x \), and that’s our answer!

Inverting Inputs and Outputs

In Prolog the division between “inputs” and “outputs” is intentionally vague. We can exploit this. It is often possible to “invert” a query and ask what inputs would compute a given output. Few other languages allow this level of flexibility.

Consider the query

\[
\text{append([1], X, [1, 2, 3]).}
\]

This asks Prolog to find a list \( x \) such that if we append \([1]\) to \( x \) we will get \([1, 2, 3]\).
Prolog answers

\( x = [2, 3] \)

The Member Relation

A common predicate when manipulating lists is a membership test—is a given value a member of a list?

An “obvious” definition is a recursive one similar to what we might program in Scheme or ML:

\[
\text{member(X, [X|_]).}
\]
\[
\text{member(X, [_|Y]) :- member(X, Y).}
\]

This definition states that the first argument, \( x \), is a member of the second argument (a list) if \( x \) matches the head of the list or if \( x \) is (recursively) a member of the rest of the list.
Note that we don’t have to “tell” Prolog that \( x \) can’t be a member of an empty list—if we don’t tell Prolog that something is true, it automatically assumes that it must be false.

Thus saying nothing about membership in an empty list is the same as saying that membership in an empty list is impossible.

Since inputs and outputs in a relation are blurred, we can use \( \text{member} \) in an unexpected way—to iterate through a list of values.

If we want to know if any member of a list \( L \) satisfies a predicate \( p \), we can simply write:

\[
\text{member}(X, L), p(X).
\]

There is no explicit iteration or searching. We simply ask Prolog to find an \( x \) such that \( \text{member}(X, L) \) is true (\( x \) is in \( L \)) and \( p(X) \) is true. Backtracking will find the “right” value for \( x \) (if any such \( x \) exists).

This is sometimes called the “guess and verify” technique.

Thus we can query:

\[
\text{member}(X, [3, -3, 0, 10, -10]), (X > 0).
\]

This asks for an \( x \) in the list \([3, -3, 0, 10, -10]\) which is greater than 0.

Prolog answers

\[
X = 3;
X = 10;
\]

Note too that our “obvious” definition of member is not the only one possible.

An alternative definition (which is far less obvious) is

\[
\text{member}(X, L) :-
\text{append}(_, [X | _], L).
\]

This definition says \( x \) is a member of \( L \) if I can take some list (whose value I don’t care about) and append it to a list that begins with \( x \) (and which ends with values I don’t care about) and get a list equal to \( L \).

Said more clearly, \( x \) is a member of \( L \) if \( x \) is anywhere in the “middle” of \( L \).

Prolog solves a query involving \( \text{member} \) by partitioning the list \( L \) in all possible ways, and checking to see if \( x \) ever is the head of the second list. Thus for \( \text{member}(X, [1, 2, 3]) \), it tries the partition \([\] \) and \([1, 2, 3]\) (exposing 1 as a possible \( x \)), then \([1]\) and \([2, 3]\) (exposing 2) and finally \([1, 2]\) and \([3]\) (exposing 3).
Sorting Algorithms

Sorting algorithms are good examples of Prolog's definitional capabilities. In a Prolog definition the "logic" of a sorting algorithm is apparent, stripped of the cumbersome details of data structures and control structures that dominate algorithms in other programming languages.

Consider the simplest possible sort imaginable, which we'll call the "naive sort."

At the simplest level a sorting of a list \( L \) requires just two things:

- The sorting is a permutation (a reordering) of the values in \( L \).
- The values are "in order" (ascending or descending).

We can implement this concept of a sort directly in Prolog. We
(a) permute an input list
(b) check if it is in sorted order
(c) repeat (a) & (b) until a sorting is found.

Permutations

Let's first look at how permutations are defined in Prolog. In most languages generating permutations is non-trivial—you need data structures to store the permutations you are generating and control structures to visit all permutations in some order.

In Prolog, permutations are defined quite concisely, though with a bit of subtlety:

\[ \text{perm}(X,Y) \text{ will be true if list } Y \text{ is a permutation of list } X. \]

Only two definitions are needed:

\[ \text{perm}([],[]). \]

\[ \text{perm}(L,[H|T]) \text{ :- } \]
- \( \text{append}(V,[H|U],L), \text{ append}(V,U,W), \text{ perm}(W,T). \)

The first definition, \( \text{perm}([],[]). \)

is trivial. An empty list may only be permuted into another empty list.

The second definition is rather more complex:

\[ \text{perm}(L,[H|T]) \text{ :- } \]
- \( \text{append}(V,[H|U],L), \text{ append}(V,U,W), \text{ perm}(W,T). \)

This rule says a list \( L \) may be permuted in to a list that begins with \( H \) and ends with list \( T \) if:

1. \( L \) may be partitioned into two lists, \( V \) and \( [H|U] \). (That is, \( H \) is somewhere in the "middle" of \( L \)).
2. Lists \( V \) and \( U \) (all of \( L \) except \( H \)) may be appended into list \( w \).
3. List \( w \) may be permuted into \( T \).
Let's see `perm` in action:

```
?- perm([1,2,3],X).
X = [1,2,3] ;
X = [1,3,2] ;
X = [2,1,3] ;
X = [2,3,1] ;
X = [3,1,2] ;
X = [3,2,1] ;
no
```

We'll trace how the first few answers are computed. Note though that all permutations are generated, and with no apparent data structures or control structures.

We start with \( L = [1,2,3] \) and \( X = [H|T] \).

We first solve \( \text{append}(V, [H|U], L) \), which simplifies to \( \text{append}(V, [H|U], [1,2,3]) \).

One solution to this goal is

\[ V = [], \ H = 1, \ U = [2,3] \]

We next solve \( \text{append}(V, U, W) \) which simplifies to \( \text{append}([1], [2,3], W) \).

The only solution for this is \( W = [2,3] \).

Finally, we solve \( \text{perm}(W, T) \), which simplifies to \( \text{perm}([2,3], T) \).

One solution to this is \( T = [2,3] \).

This gives us our first solution:

\[ [H|T] = [1,2,3] \]

To get our next solution we backtrack. Where is the most recent place we made a choice of how to solve a goal? It was at \( \text{perm}([2,3], T) \). We chose \( T = [2,3] \), but \( T = [3,2] \) is another solution. Using this solution, we get out next answer \( [H|T] = [1,3,2] \).

Let's try one more. We backtrack again. No more solutions are possible for \( \text{perm}([2,3], T) \), so we backtrack to an earlier choice point.

At \( \text{append}(V, [H|U], [1,2,3]) \) another solution is

\[ V = [1], \ H = 2, \ U = [3] \]

Using this binding, we solve \( \text{append}(V, U, W) \) which simplifies to \( \text{append}([1], [3], W) \). The solution to this must be \( W = [1,3] \).

We then solve \( \text{perm}(W, T) \) which simplifies to \( \text{perm}([1,3], T) \). One solution to this is \( T = [1,3] \). This makes our third solution for \( [H|T] = [2,1,3] \).

You can check out the other bindings that lead to the last three solutions.

### A Permutation Sort

Now that we know how to generate permutations, the definition of a permutation sort is almost trivial.

We define an `inOrder` relation that characterizes our notion of when a list is properly sorted:

```
inOrder([]).
inOrder([_]).
inOrder([A,B|T]) :-
    A <= B, inOrder([B|T]).
```

These definitions state that a null list, and a list with only one element are always in sorted order. Longer lists are in order if the first two elements are in proper order. \((A =< B)\) checks this and then the rest of the list, excluding the first element, is checked.
Now our naive permutation sort is only one line long:

```prolog
naiveSort(L1,L2) :-
    perm(L1,L2), inOrder(L2).
```

And the definition works too!

```prolog
| ?- naiveSort([1,2,3],[3,2,1]).
no
?- naiveSort([3,2,1],L).
L = [1,2,3] ;
no
| ?- naiveSort([7,3,88,2,1,6,77, -23,5],L).
L = [-23,1,2,3,5,6,7,77,88]
```

Though this sort works, it is hopelessly inefficient—it repeatedly “shuffles” the input until it happens to find an ordering that is sorted. The process is largely undirected. We don’t “aim” toward a correct ordering, but just search until we get lucky.

---

A Bubble Sort

Perhaps the best known sorting technique is the interchange or “bubble” sort. The idea is simple. We examine a list of values, looking for a pair of adjacent values that are “out of order.” If we find such a pair, we swap the two values (placing them in correct order). Otherwise, the whole list must be in sorted order and we are done.

In conventional languages we need a lot of code to search for out-of-order pairs, and to systematically reorder them. In Prolog, the whole sort may be defined in a few lines:

```prolog
bubbleSort(L,L) :- inOrder(L).
bubbleSort(L1,L2) :-
    append(X,[A,B|Y],L1), A > B,
    append(X,[B,A|Y],T),
    bubbleSort(T,L2).
```

The first line says that if $L$ is already in sorted order, we are done.

The second line is a bit more complex. It defines what it means for a list $L_2$ to be a sorting for list $L_1$, using our insight that we should swap out-of-order neighbors. We first partition list $L_1$ into two lists, $X$ and $[A,B|Y]$. This “exposes” two adjacent values in $L$, $A$ and $B$. Next we verify that $A$ and $B$ are out-of-order ($A>B$). Next, in $append(X,[B,A|Y],T)$, we determine that list $T$ is just our input $L$, with $A$ and $B$ swapped into $B$ followed by $A$. 

Finally, we verify that \(\text{bubbleSort}(T,L_2)\) holds. That is, \(T\) may be bubble-sorted into \(L_2\).

This approach is rather more directed than our permutation sort—we look for an out-of-order pair of values, swap them, and then sort the “improved” list. Eventually there will be no more out-of-order pairs, the list will be in sorted order, and we will be done.

**Merge Sort**

Another popular sort in the “merge sort” that we have already seen in Scheme and ML. The idea here is to first split a list of length \(L\) into two sublists of length \(L/2\). Each of these two lists is recursively sorted. Finally, the two sorted sublists are merged together to form a complete sorted list.

The bubble sort can take time proportional to \(n^2\) to sort \(n\) elements (as many as \(n^2/2\) swaps may be needed). The merge sort does better—it takes time proportional to \(n \log_2 n\) to sort \(n\) elements (a list of size \(n\) can only be split in half \(\log_2 n\) times).

We first need Prolog rules on how to split a list into two equal halves:

\[
\text{split([],[],[]).} \\
\text{split([A],[A],[]).} \\
\text{split([A,B|T],[A|P1],[B|P2]) :- split(T,P1,P2).}
\]

The first two lines characterize trivial splits. The third rule distributes one of the first two elements to each of the two sublists, and then recursively splits the rest of the list.

We also need rules that characterize how to merge two sorted sublists into a complete sorted list:

\[
\text{merge([],L,L).} \\
\text{merge(L,[],L).} \\
\text{merge([A|T1],[B|T2],[A|P1],[B|P2]) :- split(T,P1,P2).} \\
\text{merge([A|T1],[B|T2],[A|L2]) :- A =< B, merge(T1,[B|T2],L2).} \\
\text{merge([A|T1],[B|T2],[B|L2]) :- A > B, merge([A|T1],[T2,L2]).}
\]

The first 2 lines handle merging null lists. The third line handles the case where the head of the first sublist is \(\leq\) the head of the second sublist; the final rule handles the case where the head of the second sublist is smaller.
With the above definitions, a merge sort requires only three lines:

```prolog
mergeSort([],[]).
mergeSort([A],[A]).
mergeSort(L1,L2) :-
split(L1,P1,P2),
mergeSort(P1,S1),
mergeSort(P2,S2),
merge(S1,S2,L2).
```

The first two lines handle the trivial cases of lists of length 0 or 1. The last line contains the full “logic” of a merge sort: split the input list, \( L \) into two half-sized lists \( P_1 \) and \( P_2 \). Then merge sort \( P_1 \) into \( S_1 \) and \( P_2 \) into \( S_2 \). Finally, merge \( s_1 \) and \( S_2 \) into a sorted list \( L_2 \). That’s it!

---

**Quick Sort**

The merge sort partitions its input list rather blindly, alternating values between the two lists. What if we partitioned the input list based on values rather than positions?

The quick sort does this. It selects a “pivot” value (the head of the input list) and divides the input into two sublists based on whether the values in the list are less than the pivot or greater than or equal to the pivot. Next the two sublists are recursively sorted. But now, after sorting, no merge phase is needed. Rather, the two sorted sublists can simply be appended, since we know all values in the first list are less than all values in the second list.

---

With our notion of partitioning defined, the quicksort itself requires only 2 lines:

```prolog
qsort([],[]).
qsort([A|T],L) :-
pnartition(A,T,L1,L2),
qsort(L1,S1), qsort(L2,S2),
append(S1,[A|S2],L).
```

The first line defines a trivial sort of an empty list.

The second line says to sort a list that begins with \( A \) and ends with list \( T \), we partition \( T \) into sublists \( L_1 \) and \( L_2 \), based on \( A \). Then we recursively quick sort \( L_1 \) into \( S_1 \) and \( L_2 \) into \( S_2 \). Finally we append \( S_1 \) to \([A|S2]\) (\( A \) must be \( > \) all values in \( S_1 \) and \( A \) must be \( \leq \) all values in \( S_2 \)). The result is \( L \), a sorting of \([A|T]\).
**Arithmetic in Prolog**

The `=` predicate can be used to test bound variables for equality (actually, identity).

If one or both of `=`'s arguments are free variables, `=` forces a binding or an equality constraint.

Thus

<table>
<thead>
<tr>
<th><code>?- 1=2.</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
</tr>
<tr>
<td><code>?- X=2.</code></td>
</tr>
<tr>
<td><code>X = 2</code></td>
</tr>
<tr>
<td><code>?- Y=X.</code></td>
</tr>
<tr>
<td><code>Y = X = _10751</code></td>
</tr>
<tr>
<td><code>?- X=Y, X=joe.</code></td>
</tr>
<tr>
<td><code>X = Y = joe</code></td>
</tr>
</tbody>
</table>

**Hence**

<table>
<thead>
<tr>
<th><code>?- 2 is 1+1.</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
</tr>
<tr>
<td><code>?- X is 3*4.</code></td>
</tr>
<tr>
<td><code>X = 12</code></td>
</tr>
<tr>
<td><code>?- Y is Z+1.</code></td>
</tr>
<tr>
<td>! Instantiation error in argument 2 of is/2</td>
</tr>
<tr>
<td>! goal: _10712 is _10715+1</td>
</tr>
</tbody>
</table>

The requirement that the right-hand side of an `is` relation be ground is essentially procedural. It exists to avoid having to invert complex equations. Consider,

\(0 \text{ is } (I^{*N})+(J^{*N})-(K^{*N})\), \(N>2\).

**Arithmetic Terms are Symbolic**

Evaluation of an arithmetic term into a numeric value must be forced.

That is, `1+2` is an infix representation of the relation `+(1, 2)`. This term is not an integer!

Therefore

<table>
<thead>
<tr>
<th><code>?- 1+2=3.</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
</tr>
</tbody>
</table>

To force arithmetic evaluation, we use the infix predicate `is`.

The right-hand side of `is` must be all ground terms (literals or variables that are already bound). No free (unbound) variables are allowed.

**Counting in Prolog**

Rules that involve counting often use the `is` predicate to evaluate a numeric value.

Consider the relation `len(L, N)` that is true if the length of list \(L\) is \(N\).

\[
\text{len([], 0).}
\]

\[
\text{len([_|T], N) :-}
\]

\[
\text{len(T, M), N is M+1.}
\]

<table>
<thead>
<tr>
<th><code>?- len([1, 2, 3], X).</code></th>
</tr>
</thead>
<tbody>
<tr>
<td><code>X = 3</code></td>
</tr>
<tr>
<td><code>?- len(Y, 2).</code></td>
</tr>
<tr>
<td><code>Y = [__10903, __10905]</code></td>
</tr>
</tbody>
</table>

The symbols `_10903` and `_10905` are “internal variables” created as needed when a particular value is not forced in a solution.
Debugging Prolog

Care is required in developing and testing Prolog programs because the language is untyped; undeclared predicates or relations are simply treated as false.

Thus in a definition like

\[
\text{adj}([A,B|\_]) :- A=B.
\]

\[
\text{adj}([\_,B|T]) :- \text{adj}([B|T]).
\]

\[
?- \text{adj}([1,2,2]).
\]

no

(Some Prolog systems warn when an undefined relation is referenced, but many others don't).

Similarly, given

\[
\text{member}(A, [A|\_]).
\]

\[
\text{member}(A, [\_|T]) :- \text{member}(A, [T]).
\]

\[
?- \text{member}(2, [1,2]).
\]

Infinite recursion! (Why?)

If you're not sure what is going on, Prolog's trace feature is very handy. The command

\[
\text{trace}.
\]

turns on tracing. \texttt{(notrace turns tracing off)}.

Hence

\[
?- \text{trace}.
\]

\[
\text{yes}
\]

\[
\text{[trace]}
\]

\[
?- \text{member}(2, [1,2]).
\]

Termination Issues in Prolog

Searching infinite domains (like integers) can lead to non-termination, with Prolog trying every value.

Consider

\[
\text{odd}(1).
\]

\[
\text{odd}(N) :- \text{odd}(M), N \text{ is } M+2.
\]

\[
?- \text{odd}(X).
\]

\[
X = 1 ;
\]

\[
X = 3 ;
\]

\[
X = 5 ;
\]

\[
X = 7
\]
A query
\[ ?- \text{odd}(X), X=2. \]
going into an infinite search, generating each and every odd integer and finding none is equal to 2!
The obvious alternative, \( \text{odd}(2) \) (which is equivalent to \( X=2, \text{odd}(X) \)) also does an infinite, but fruitless search.
We'll soon learn that Prolog does have a mechanism to “cut off” fruitless searches.

Definition Order can Matter
Ideally, the order of definition of facts and rules should not matter.
But, in practice definition order can matter. A good general guideline is to define facts before rules. To see why, consider a very complete database of motherOf relations that goes back as far as motherOf(cain, eve).
Now we define
\[
\text{isMortal}(X) :- \\
\text{isMortal}(Y), \text{motherOf}(X,Y).
\]
isMortal(eve).

These definitions state that the first woman was mortal, and all individuals descended from her are also mortal.
But when we try as trivial a query as
\[ ?- \text{isMortal}(eve). \]
we go into an infinite search!
Why?
Let's trace what Prolog does when it sees
\[ ?- \text{isMortal}(eve). \]
It matches with the first definition involving isMortal, which is
\[
\text{isMortal}(X) :- \\
\text{isMortal}(Y), \text{motherOf}(X,Y).
\]
It sets \( X=\text{eve} \) and tries to solve
\[
\text{isMortal}(Y), \text{motherOf}(\text{eve},Y).
\]
It will then expand isMortal(Y) into
\[
\text{isMortal}(Z), \text{motherOf}(Y,Z).
\]
An infinite expansion ensues.
The solution is simple—place the “base case” fact that terminates recursion first.
If we use
\[
\text{isMortal}(\text{eve}).
\]
isMortal(X) :-
\[
\text{isMortal}(Y), \text{motherOf}(X,Y).
\]
yes
\[ ?- \text{isMortal}(eve). \]
yes
But now another problem appears!
If we ask
\[ ?- \text{isMortal}(\text{clarkKent}). \]
we go into another infinite search!
Why?
The problem is that Clark Kent is from the planet Krypton, and hence won’t appear in our motherOf database. Let’s trace the query. It doesn’t match isMortal(eve).

We next try:

\[
\text{isMortal(clarkKent) :- isMortal(Y), motherOf(clarkKent,Y).}
\]

We try \(Y = \text{eve}\), but eve isn’t Clark’s mother. So we recurse, getting:

\[
\text{isMortal(Z), motherOf(Y,Z), motherOf(clarkKent,Y).}
\]

But eve isn’t Clark’s grandmother either! So we keep going further back, trying to find a chain of descendants that leads from eve to clarkKent. No such chain exists, and there is no limit to how long a chain Prolog will try.

There is a solution though!

We simply rewrite our recursive definition to be:

\[
\text{isMortal(X) :- motherOf(X,Y), isMortal(Y).}
\]

This is logically the same, but now we work from the individual \(X\) back toward eve, rather than from eve toward \(X\). Since we have no motherOf relation involving clarkKent, we immediately stop our search and answer \(\text{no!}\).

---

Extra-logical Aspects of Prolog

To make a Prolog program more efficient, or to represent negative information, Prolog needs features that have a procedural flavor. These constructs are called “extra-logical” because they go beyond Prolog’s core of logic-based inference.

---

The Cut

The most commonly used extra-logical feature of Prolog is the “cut symbol,” “!.”

A ! in a goal, fact or rule “cuts off” backtracking. In particular, once a ! is reached (and automatically matched), we may not backtrack across it. The rule we’ve selected and the bindings we’ve already selected are “locked in” or “frozen.”

For example, given:

\[
\text{x(A) :- y(A, B), z(B), !, v(B, C).}
\]

Once the ! is hit we can’t backtrack to resatisfy \(y(A, B)\) or \(z(B)\) in some other way. We are locked into this
rule, with the bindings of A and B already in place.

We can backtrack to try various solutions to \( v(B, C) \).

It is sometimes useful to have several 's in a rule. This allows us to find a partial solution, lock it in, find a further solution, then lock it in, etc.

For example, in a rule

\[
a(X) - b(X), !, c(X, Y), !, d(Y).
\]

we first try to satisfy \( b(X) \), perhaps trying several facts or rules that define the \( b \) relation. Once we have a solution to \( b(X) \), we lock it in, along with the binding for \( X \).

Then we try to satisfy \( c(X, Y) \), using the fixed binding for \( X \), but perhaps trying several bindings for \( Y \) until \( c(X, Y) \) is satisfied.

We then lock in this match using another !.

Finally we check if \( d(Y) \) can be satisfied with the binding of \( Y \) already selected and locked in.

---

**When are Cuts Needed?**

A cut can be useful in improving efficiency, by forcing Prolog to avoid useless or redundant searches.

Consider a query like

\[
\text{member}(X, \text{list1}), \\
\text{member}(X, \text{list2}), \text{isPrime}(X).
\]

This asks Prolog to find an \( X \) that is in \( \text{list1} \) and also in \( \text{list2} \) and also is prime.

\( X \) will be bound, in sequence, to each value in \( \text{list1} \). We then check if \( X \) is also in \( \text{list2} \), and then check if \( X \) is prime.

Assume we find \( X = 8 \) is in \( \text{list1} \) and \( \text{list2} \), \( \text{isPrime}(8) \) fails (of course). We backtrack to \( \text{member}(X, \text{list2}) \) and try to resatisfy it with the same value of \( X \).

But clearly there is never any point in trying to resatisfy \( \text{member}(X, \text{list2}) \). Once we know a value of \( X \) is in \( \text{list2} \), we test it using \( \text{isPrime}(X) \).

If it fails, we want to go right back to \( \text{member}(X, \text{list1}) \) and get a different \( X \).

To create a version of \( \text{member} \) that never backtracks once it has been satisfied we can use 1.

We define

\[
\text{member1}(X, [X|\_]) :- !. \\
\text{member1}(X, [\_|Y]) :- \\
\qquad \text{member1}(X, Y).
\]

Our query is now

\[
\text{member}(X, \text{list1}), \\
\text{member1}(X, \text{list2}), \text{isPrime}(X).
\]

(Why isn't \( \text{member1} \) used in both terms?)
Expressing Negative Information

Sometimes it is useful to state rules about what can’t be true. This allows us to avoid long and fruitless searches.

fail is a goal that always fails. It can be used to represent goals or results that can never be true.

Assume we want to optimize our grandMotherOf rules by stating that a male can never be anyone’s grandmother (and hence a complete search of all motherOf and fatherOf relations is useless).

A rule to do this is

\[
\text{grandMotherOf}(X, \text{GM}) \leftarrow \text{male}(\text{GM}), \text{fail}.
\]

This rule doesn’t do quite what we hope it will!

Why?

The standard approach in Prolog is to try other rules if the current rule fails. Hence we need some way to “cut off” any further backtracking once this negative rule is found to be applicable.

This can be done using

\[
\text{grandMotherOf}(X, \text{GM}) \leftarrow \\
\quad \text{male}(\text{GM}), \text{!}, \text{fail}.
\]

Other Extra-Logical Operators

- assert and retract

These operators allow a Prolog program to add new rules during execution and (perhaps) later remove them. This allows programs to learn as they execute.

- findall

Called as findall(X, goal, List) where X is a variable in goal. All possible solutions for X that satisfy goal are found and placed in List.

For example,

\[
\text{findall}(X, \\
\quad (\text{append}(_, [X|_], [-1, 2, -3, 4]), (X < 0)), \\
\quad L).
\]

\[
L = [-1, -3]
\]

- var and nonvar

\[
\text{var}(X) \text{ tests whether } X \text{ is unbound (free).}
\]

\[
\text{nonvar}(Y) \text{ tests whether } Y \text{ is bound (no longer free).}
\]

These two operators are useful in tailoring rules to particular combinations of bound and unbound variables.

For example, the rule

\[
\text{grandMotherOf}(X, \text{GM}) \leftarrow \\
\quad \text{male} (\text{GM}), \text{!}, \text{fail}.
\]

might backfire if \text{GM} is not yet bound. We could set \text{GM} to a person for whom \text{male} (\text{GM}) is true, then fail because we don’t want grandmothers who are male!
To remedy this problem, we use the rule only when \( \text{GM} \) is bound. Our rule becomes

\[
\text{grandMotherOf}(X, \text{GM}) :- \\
\text{nonvar}(\text{GM}), \text{male}(\text{GM}), !, \text{fail}.
\]

An Example of Extra-Logical Programming

Factorial is a very common example program. It’s well known, and easy to code in most languages.

In Prolog the “obvious” solution is:

\[
\text{fact}(N, 1) :- N =< 1. \\
\text{fact}(N, F) :- N > 1, M \text{ is } N-1, \\
\text{fact}(M, G), F \text{ is } N*G.
\]

This definition is certainly correct. It mimics the usual recursive solution.

But, in Prolog “inputs” and “outputs” are less distinct than in most languages. In fact, we can envision 4 different combinations of inputs and outputs, based on what is fixed (and thus an input) and what is free (and hence is to be computed):

1. \( N \) and \( F \) are both ground (fixed). We simply must decide if \( F=N! \)
2. \( N \) is ground and \( F \) is free. This is how \text{fact} is usually used. We must compute an \( F \) such that \( F=N! \)
3. \( F \) is fixed and \( N \) is free. This is an uncommon usage. We must find an \( N \) such that \( F=N! \), or determine that no such \( N \) is possible.
4. Both \( N \) and \( F \) are free. We generate, in sequence, pairs of \( N \) and \( F \) values such that \( F=N! \)

Our solution works for combinations 1 and 2 (where \( N \) is fixed), but not combinations 3 and 4. (The problem is that \( N =< 1 \) and \( N > 1 \) can’t be satisfied when \( N \) is free).

We’ll need to use \text{nonvar} and \! to form a solution that works for all 4 combinations of inputs.

We first handle the case where \( N \) is ground:

\[
\text{fact}(1,1). \\
\text{fact}(N, 1) :- \text{nonvar}(N), N =< 1, !. \\
\text{fact}(N, F) :- \text{nonvar}(N), N > 1, !, \\
M \text{ is } N-1, \text{fact}(M, G), F \text{ is } N*G, !.
\]

The first rule handles the base case of \( N=1 \).
The second rule handles the case of \( N<1 \).
The third rule handles the case of \( N > 1 \). The value of \( F \) is computed recursively. The first ! in each of these rules forces that rule to be the only one used for the values of \( N \) that match. Moreover, the second ! in the third rule states that after \( F \) is computed, further backtracking is useless; there is only one \( F \) value for any given \( N \) value.

To handle the case where \( F \) is bound and \( N \) is free, we use

\[
\text{fact}(N,F) :- \text{nonvar}(F), !, \\
\text{fact}(M,G), N \text{ is } M+1, F2 \text{ is } N*G, \\
F =< F2, !, F=F2.
\]

In this rule we generate \( N, F2 \) pairs until \( F2 >= F \). Then we check if \( F=F2 \). If this is so, we have the \( N \) we want. Otherwise, no such \( N \) can exist and we fail (and answer no).

For the case where both \( N \) and \( F \) are free we use:

\[
\text{fact}(N,F) :- \text{fact}(M,G), N \text{ is } M+1, \\
F \text{ is } N*G.
\]

This systematically generates \( N, F \) pairs, starting with \( N=2, F=2 \) and then recursively building successor values (\( N=3, F=6 \), then \( N=4, F=24 \), etc.).

**Parallelism in Prolog**

One reason that Prolog is of interest to computer scientists is that its search mechanism lends itself to parallel evaluation.

In fact, it supports two different kinds of parallelism:

- AND Parallelism
- OR Parallelism

**And Parallelism**

When we have a goal that contains subgoals connected by the "\&" (And) operator, we may be able to utilize "and parallelism."

Rather than solve subgoals in sequence, we may be able to solve them in parallel if bindings can be properly propagated.

Thus in

\[
a(X), \ b(X,Y), \ c(X,Z), \ d(Y,Z).
\]

we may be able to first solve \( a(X) \), binding \( X \), then solve \( b(X,Y) \) and \( c(X,Z) \) in parallel, binding \( Y \) and \( Z \), then finally solve \( d(Y,Z) \).
An example of this sort of parallelism is:

\[
\text{member}(X, \text{list1}), \\
\text{member1}(X, \text{list2}), \text{isPrime}(X).
\]

Here we can let \text{member}(X, \text{list1}) select an \(X\) value, then test

\text{member1}(X, \text{list2}) \text{ and isPrime}(X)

in parallel. If one or the other fails, we just select another \(X\) from \text{list1}
and retest \text{member1}(X, \text{list2}) \text{ and isPrime}(X) in parallel.

\[\text{OR Parallelism}\]

When we match a goal we almost always have a choice of several rules
or facts that may be applicable. Rather than try them in sequence, we
can try several matches of different facts or rules in parallel. This is "or
parallelism."

Thus given:

\[
\begin{align*}
\text{a}(X) & : = \text{b}(X) . \\
\text{a}(Y) & : = \text{c}(Y) .
\end{align*}
\]

when we try to solve \(\text{a}(10)\),
we can simultaneously check both
\(\text{b}(10)\) and \(\text{c}(10)\).

Recall our definition of:

\[
\text{member}(X, \text{L}) : - \\
\text{append}(P, [X|S], \text{L}).
\]

where \text{append} is defined as:

\[
\text{append}([], L, L) . \\
\text{append}([X|L1], L2, [X|L3]) :- \\
\text{append}(L1, L2, L3).
\]

Assume we have the query

\[
| \ ? \text{member}(2, \{1, 2, 3\}).
\]

This immediately simplifies to:

\[
\text{append}(P, [2|S], [1, 2, 3]).
\]

Now there are two \text{append} definitions we can try in parallel:

1. Match \text{append}(P, [2|S], [1, 2, 3])
   with \text{append}([], L, L). This requires that \([2|S] = [1, 2, 3]\), which must fail.

2. Match \text{append}(P, [2|S], [1, 2, 3])
   with \text{append}([X|L1], L2, [X, L3]).

This requires that \(P=[X|L1],
\[2|S]=L2, [1, 2, 3]=[X, L3]\).

Simplifying, we require that \(X=1,
P=[1|L1], L3=[2, 3]\).

Moreover we must solve
\text{append}(L1, L2, L3) \text{ which simplifies to } 
\text{append}(L1, [2|S], [2, 3]).

We can match this call to \text{append} in
two different ways, so or parallelism
can be used again.

When we try matching
\text{append}(L1, [2|S], [2, 3]) \text{ against }
\text{append}([], L, L) \text{ we get }
\[2|S]=[2, 3]\), which is satisfiable if \(S\)
is bound to \([3]\). We therefore signal
back that the query is true.
Speculative Parallelism

Prolog also lends itself nicely to speculative parallelism. In this form of parallelism, we “guess” or speculate that some computation may be needed in the future and start it early. This speculative computation can often be done in parallel with the main (non-speculative) computation.

Recall our example of

\[\text{member}(X, \text{list1}), \text{member1}(X, \text{list2}), \text{isPrime}(X).\]

After \(\text{member}(X, \text{list1})\) has generated a preliminary solution for \(X\), it is tested (perhaps in parallel) by \(\text{member1}(X, \text{list2})\) and \(\text{isPrime}(X)\).

But this value of \(X\) may be rejected by one or both of these tests. If it is, we’ll ask \(\text{member}(X, \text{list1})\) to find a new binding for \(X\). If we wish, this next binding can be generated speculatively, while the current value of \(X\) is being tested. In this way if the current value of \(X\) is rejected, we’ll have a new value ready to try (or know that no other binding of \(X\) is possible).

If the current value of \(X\) is accepted, the extra speculative work we did is ignored. It wasn’t needed, but was useful insurance in case further \(X\) bindings were needed.

Reading Assignment

- Java for C++ Programmers (linked from class web page)
- Roosta: Section 9.3
- Webber: Chapters 13, 15 and 17

Java & Object-Oriented Programming

Java is a fairly new and very popular programming language designed to support secure, platform-independent programming.

It is a good alternative to C or C++, trading a bit of efficiency for easier programming, debugging and maintenance.

Java is routinely interpreted (at the byte-code level), making it significantly slower than compiled C or C++. However true Java compilers exist, and are becoming more widespread. (IBM’s Jalapeno project is a good example). When compiled, Java’s execution speed is close to that of C or C++.
Subclassing in Java

When a new class is defined in terms of an existing class, the new class extends the existing class. The new class inherits all public and protected members of its parent (or base) class.
The new class may add new methods or fields. It may also redefine inherited methods or fields.

```java
class Point {
    int x, y;
    Point(int xin, int yin) {
        x = xin; y = yin;
    }
    static float dist(Point P1, Point P2) {
        return (float) Math.sqrt((P1.x-P2.x)*(P1.x-P2.x) + (P1.y-P2.y)*(P1.y-P2.y));
    }
}
```

```java
class Point3 extends Point {
    int z;
    Point3(int xin, int yin, int zin) {
        super(xin,yin); z=zin;
    }
    static float dist(Point3 P1, Point3 P2) {
        float d=Point.dist(P1,P2);
        return (float) Math.sqrt((P1.z-P2.z)*(P1.z-P2.z) + d*d);
    }
}
```

Note that although Point3 redefines dist, the old definition of dist is still available by using the parent class as a qualifier (Point.dist).
The same is true for fields that are hidden when a field in a parent is redeclared.

Non-static methods are automatically virtual: a redefined method is automatically used in all inherited methods including those defined in parent classes that think they are using an earlier definition of the class.
Example:

```java
class C {
    void DoIt(){PrintIt();}
    void PrintIt()
        {println("C rules!");}
}
class D extends C {
    void PrintIt()
        {println("D rules!");}
    void TestIt() {DoIt();}
}
D dvar = new D();
dvar.TestIt();
D rules! is printed.
```

Static methods in Java are not virtual (this can make them easier to implement efficiently).
Abstract Classes and Methods

Sometimes a Java class is not meant to be used by itself because it is intentionally incomplete. Rather, the class is meant to be starting point for the creation (via subclassing) of more complete classes. Such classes are abstract.

Example:

```java
abstract class Shape {
    Point location;
}

class Circle extends Shape {
    float radius;
}
```

Methods can also be made abstract to indicate that their actual definition will appear in subclasses:

```java
abstract class Shape {
    Point location;
    abstract float area();
}
class Circle extends Shape {
    float radius;
    float area(){
        return Math.pi*radius*radius;
    }
}
```

Subtyping and Inheritance

We can use a subtyping mechanism, as found in C++ or Java, for two different purposes:

- We may wish to inherit the actual implementations of classes and members to use as the basis of a more complete or extended class. To inherit an implementation, we say a given class “extends” an existing class:
  ```java
class Derived extends Base {
    ...
};
```
  Class Derived contains all of the members of Base plus any others it cares to add.

- We may wish to inherit an interface—a set of method names and values that will be available for use. To inherit (or claim) an interface, we use a Java interface definition.
  ```java
  interface Shape {
    abstract float area();
  }
  ```
  An interface doesn't implement anything; rather, it gives a name to a set of operations or values that may be available within one or more classes.
Why are Interfaces Important?

Many classes, although very different, share a common subset of values or operations. We may be willing to use any such class as long as only interface values or operations are used.

For example, many objects can be ordered (or at least partially-ordered) using a “less than” operation.

If we always implement less than the same way, for example,

```java
boolean lessThan(Object o1, Object o2);
```

then we can create an interface that admits all classes that know about the `lessThan` function:

```java
interface Compare {
    boolean lessThan(Object o1, Object o2);
}
```

Now different classes can each implement the `Compare` interface, proclaiming to the world that they know how to compare objects of the class they define:

```java
class IntCompare implements Compare {
    public boolean lessThan(Object i1, Object i2) {
        return ((Integer)i1).intValue() < ((Integer)i2).intValue();
    }

class StringCompare implements Compare {
    public boolean lessThan(Object i1, Object i2) {
        return ((String)i1).compareTo((String)i2) < 0;
    }
}
```

The advantage of using interfaces is that we can now define a method or class that only depends on the given interface, and which will accept any type that implements that interface.

```java
class PrintCompare {
    public static void printAns(Object v1, Object v2, Compare c) {
        System.out.println(v1.toString() + " < " + v2.toString() + " is " + new Boolean(c.lessThan(v1, v2)).toString());
    }

class Test {
    public static void main(String args[]) {
        Integer i1 = new Integer(2);
        Integer i2 = new Integer(1);
        PrintCompare.printAns(i1, i2, new IntCompare());
        String s2 = "abcdef";
        String s1 = "xyzaa";
        PrintCompare.printAns(s1, s2, new StringCompare());
    }
}
```

Since classes may have many methods and modes of use or operation, a given class may implement many different interfaces. For example, many classes support the `Clonable` interface, which states that objects of the class may be duplicated (cloned).
Multiple Inheritance

We have seen that a class may be derived from a given parent class. It is sometimes useful to allow a class to be derived from more than one parent, inheriting members of all parents. This is multiple inheritance; it is allowed by C++ and Python, but not by Java or C#.

The basic idea is that sometimes we want a “composite” object formed from more than one source. Hence a Computer object can be viewed as both a PhysicalObject (with height, weight, color, cost, etc.) and also a CPUImplementation (with memory size, processor design, processor speed, I/O ports, etc.)

Using multiple inheritance we merge aspects of a PhysicalObject and a CPUImplementation, and perhaps add additional data:

```cpp
class PhysicalObject {
    float height, width, weight;
    Color outsideColor;
    ...
}
class CPUImplementation {
    CPUClass CPUKind;
    int memorySize, CPUSpeed;
    ...
}
class Computer: PhysicalObject, CPUImplementation {
    String myURL; ...
}
```

The advantages of multiple inheritance are obvious—you can build a class from many sources rather than just one.

There are problems though:

- If the same name appears in more than one parent, which is used? For example, if both parents contain a “copyright” field, which do you get? C++ forbids access to fields common to several parents, though accidental clashes of member names are certainly possible. Python relies on order of specification of the parents (which can be somewhat arbitrary).

- Access to fields and methods can be less efficient. A method in a parent class can’t know how fields in derived class will be allocated when multiple parents may exist. Hence some form of indirection may be needed. For example, in class PhysicalObject, we may believe

that `height` is the first field allocated, while in CPUImplementation we may believe that `CPUKind` is allocated first. But in class Computer, which contains both `height` and `CPUKind`, both fields can’t come first.
Threads and Parallelism in Java

Java is one of the few “main stream” programming languages to explicitly provide for user-programmed parallelism in the form of threads.

A Java programmer may organize a program as several threads that may execute concurrently.

Even if the program is run on a uniprocessor, use of threads may improve performance. This is because the threads can be multi-programmed, with threads switched automatically on I/O delays, page faults or even cache misses.

A program that is designed to support multiple threads is also “prepared” for future upgrades to multiprocessors or multi-threaded processors.

Java Threads

In Java any class that implements the Runnable interface can be started as a concurrent thread:

```java
interface Runnable {
    public void run();
}
```

When the thread is started, the method run begins to execute (perhaps concurrently with other threads of the main program).

You create a thread using the Thread constructor:

```java
new Thread(RunnableObject)
```

Creating a thread does not start it; you must execute the start method within the Thread object.

When this is done, the run method immediately starts, and continues until that method terminates normally, or it throws an uncaught exception, or it is explicitly stopped, or the main program stops.

On uniprocessors, thread execution is interleaved; on multiprocessors or multithreaded architectures execution can be concurrent.

```java
class DoSort implements Runnable {
    int [] data;
    DoSort(int[] in) {data=in;}
    public void run(){
        // sort the data array;
    }
}
```

class Test {
    public static void main(String args[]){
        DoSort d = new DoSort(new int[1000]);
        Thread t1 = new Thread(d);
        t1.start();
        // We can continue while t1
        // does its sort
    }
}
We can start multiple threads, and the threads can use `sleep` to delay or synchronize their execution:

```java
class PingPong implements Runnable {
    int delay; String word;
    PingPong(String s, int i) {
        delay = i; word = s;
    }
    public void run() {
        try {
            while (true) {
                System.out.print(word + " ");
                Thread.sleep(delay);
            }
        } catch (InterruptedException e) {
        }
    }
}
```

```java
public static void main(String args[]) {  
    Thread t1 = new Thread(new PingPong("ping", 33));
    Thread t2 = new Thread(new PingPong("PONG", 100));
    t1.start();
    t2.start();
}
```

ping PONG ping ping PONG ping ping ping PONG ping ping PONG ping ...

---

**Synchronization in Java**

We often want threads to co-operate, typically in how they access shared data structures.

Since thread execution is asynchronous, the details of how threads interact can be unpredictable.

Consider a method

```java
    update() {
        n = n + 1;
        val = f(n);
    }
```

that updates fields of an object. If two or more threads execute `update` concurrently, we might get unexpected or even illegal behavior. (Why?)

A Java method may be synchronized, which guarantees that at most one thread can execute the method at a time. Other threads wishing access are forced to wait until the currently executing thread completes.

Thus

```java
    void synchronized update() { ... }
```

can safely be used to update an object, even if multiple threads are active.

There is also a synchronized statement in Java that forces threads to execute a block of code sequentially.

```java
    synchronized(obj) {
        obj.n = obj.n + 1;
        obj.val = f(obj.n);
    }
```
Synchronization Primitives

The following operations are provided to allow threads to safely interact:

- `wait()`  
  Sleep until awakened
- `wait(n)`  
  Sleep until awakened or until n milliseconds pass
- `notify()`  
  Wake up one sleeping thread
- `notifyAll()`  
  Wake up all sleeping threads

Using these primitives, correct concurrent access to a shared data structure can be programmed.

Locks, Semaphores and Monitors

Java's synchronization mechanisms are based upon the notion of a lock. A lock is a special value that can be held by at most one thread.

If a thread holds a lock, it has permission to do some “critical” operation like writing a shared variable or restructuring a shared data object.

If a thread wants to do an operation but doesn’t hold the necessary lock, it must wait until it gets the lock.

In Java there is a lock associated with each run-time object.

Lock Granularity and Access

Though each Java object has a lock, you often don’t want to lock and unlock each object you access.

If you are manipulating a large data structure (like a tree or hash table), acquiring the lock for each object in the tree or table can be costly and error-prone.

Instead, it is common to create a lock corresponding to a group of objects. Hence holding the lock to the root of a tree may give you permission to access the whole tree.

There is a danger though—if all or most of a large structure is held by one thread, then other threads won’t be able to access that structure concurrently.

Consider a Buffer class in which independent threads may try to store or fetch data objects:

```java
class Buffer {
    private Queue q;
    Buffer() { q = new Queue(); }
    public synchronized void put(Object obj) {
        q.enqueue(obj);
        notify(); //Why is this needed?
    }
    public synchronized Object get() {
        while (q.isEmpty()) {
            //Why a while loop?
            wait();
        }
        return q.dequeue();
    }
}
```
For example, a large shared data base (like one used to record current bookings for an airline) shouldn’t be held exclusively by one thread—hundreds of concurrent threads may want to access it at any time. An intermediate lock, like all reservations for a single fight, is more reasonable. There is also an issue of how long you hold a lock. The ideal is to have exclusive access for as short a period as is necessary. Work that is not subject to interference by other threads (like computations using local variables) should be done before a lock is obtained. Hence Java’s synchronized statement allows a method to get exclusive access to an object for a limited region, enhancing shared access.

Deadlock

A variety of programming problems appear in concurrent programs that don’t exist in ordinary sequential programs. The most serious of these is deadlock: Two or more threads hold locks that other threads require. Each waits for the other thread to release a needed lock, and no thread is able to execute. As an example of how deadlock may occur, consider two threads, $t_1$ and $t_2$. Each requires two files, a master file and a log file. Since these files are shared, each has a lock. Assume $t_1$ gets the lock for the master file while $t_2$ (at the same instant) gets the lock for the log file.

Now each is stuck. Each has one file, and will wait forever for the other file to be released.

In Java deadlock avoidance is wholly up to the programmer. There are no language-level guarantees that deadlock can’t happen. Some languages have experimented with ways to help programmers avoid deadlock:

- If all locks must be claimed at once, deadlock can be avoided. You either get all of them or none, but you can’t block other threads while making no progress yourself.
- Locks (and the resources they control) can be ordered, with the rule that you must acquire locks in the proper order. Now two threads can’t each hold locks the other needs.
- The language can require that the largest set of locks ever needed be declared in advance. When locks are requested, the operating system can track what’s claimed and what may be needed, and refuse to honor unsafe requests.
Fairness & Starvation

When one thread has a lock, other threads who want the lock will be suspended until the lock is released. It can happen that a waiting thread may be forced to wait indefinitely to acquire a lock, due to an unfair waiting policy. A waiting thread that never gets a lock it needs due to unfair lock allocation faces starvation.

As an example, if we place waiting threads on a stack, newly arrived threads will get access to a lock before earlier arrivals. This can lead to starvation. Most thread managers try to be fair and guarantee that all waiting threads have a fair chance to acquire a lock.

to wait until the lock is released. If the lock value is > 0 you get it and all others are “locked out.”

When you want to release a lock, you do up(lock), which makes lock non-zero and eligible for another thread to claim.

In fact, since only one thread will ever have a lock, the lock value needs to be only one bit, with 1 meaning currently free and unlocked and 0 meaning currently claimed and locked.

How are Locks Implemented?

Internally, Java needs operations to acquire a lock and release a lock. These operations can be implemented using the notion of a semaphore.

A semaphore is an integer value (often just a single bit) with two atomic operations: up and down.

up(s) increments s atomically.
down(s) decrements s atomically.

But if s is already zero, the process doing the down operation is put in a wait state until s becomes positive (eventually some other process should do an up operation).

Now locks are easy to implement. You do a down(lock) to claim a lock. If someone else has it, you are forced

Monitors

Direct manipulation of semaphores is tedious and error-prone. If you acquire a lock but forget to release it, threads may be blocked forever.

Depending on where down and up operations are placed, it may be difficult to understand how synchronization is being performed.

Few modern languages allow direct use of semaphores. Instead, semaphores are used in the implementation of higher-level constructs like monitors.

A monitor is a language construct that guarantees that a block of code will be executed synchronously (one thread at a time).
The Java `synchronized` statement is a form of monitor.

When

```java
synchronized(obj) { ... }
```

is executed, “invisible” `getLock` and `freeLock` operations are added:

```java
synchronized(obj) {
    getLock(obj)
    ...
    freeLock(obj);
}
```

This allows the body of the `synchronized` statement to execute only when it has the lock for `obj`. Thus two different threads can never simultaneously execute the body of a `synchronized` statement because two threads can’t simultaneously hold `obj`’s lock.

In fact, synchronized methods are really just methods whose bodies are enclosed in an invisible `synchronized` statement.

If we execute

```java
obj.method()
```

where `method` is synchronized, `method`’s body is executed as if it were of the form

```java
synchronized(obj) {
    body of method
}
```

Operations like `sleep`, `wait`, `notify` and `notifyAll` also implicitly cause threads to release locks, allowing other threads to proceed.

Java 1.5 (Tiger Java)

Recently, Java has been extended to include a variety of improvements, many drawn from functional languages.

Added features include:

- Parametric polymorphism.

  Classes and interfaces may be parameterized using a type parameter.

  ```java
  class List<T> {
      T head;
      List<T> tail;
  }
  ```

  Interfaces may also be parameterized.

- Enhanced loop iterators.

  ```java
  for (v : myArray) {
      // each element of myArray
      // appears as a value of v
  }
  ```

- Automatic boxing and unboxing of wrapper classes.

  Conversion from `int` to `Integer` or `Integer` to `int` is now automatic.

- Typesafe enumerations.

  ```java
  public enum Color {RED, BLUE, GREEN};
  ```

- Static imports.

  You may import all static members of a class and use them without qualification. Thus you may now write `System.out.println` rather than `System.out.println`.

- Variable argument methods.

- Formatted output using `printf`:

  ```java
  out.printf("Ans = %3d", a+b);
  ```
C# is Microsoft’s answer to Java. In most ways it is very similar to Java, with some C++ concepts reintroduced and some useful new features.

Similarities to Java include:
- C# is object-based, with all objects descended from class `Object`.
- Objects are created from classes using `new`. All objects are heap-allocated and garbage collection is provided.
- All code is placed within methods which must be defined within classes.
- Almost all Java reserved words have C# equivalents (many are identical).
- Classes have single inheritance.

C# Improves Upon Some Java Features
- Operators as well as methods can be overloaded:
  ```csharp
class Point {
    int x, y;
    static Point operator + (Point p1, Point p2) {
        return new Point(p1.x+p2.x, p1.y+p2.y);
    }
}
```
- Switch statements may be indexed by string literals.
- In a switch, fall throughs to the next case are disallowed (if non-empty).
- Goto’s are allowed.
- Virtual methods must be marked.
Persistent objects (that may be stored across executions) are available.

C# Adds Useful Features

- Events and delegates are included to handle asynchronous actions (like keyboard or mouse actions).
- Properties allow user-defined read and write actions for fields. You can add `get` and `set` methods to the definition of a field. For example,

```csharp
class Customer {
    private string name;
    public string Name {
        get { return name; }
        set { name = value; }
    }
}
```

Customer c; ...  
string s = c.Name;

- Indexers allow objects other than arrays to be indexed. The `[]` operator is overloadable. This allows you to define the meaning of `obj[123]` or `obj["abc"]` within any class definition.

- Collection classes may be directly enumerated:  
  `foreach (int i in array) ...`

- Fields, methods and constructors may be defined within a struct as well as a class. Structs are allocated within the stack instead of the heap, and are passed by value. For example:

```csharp
struct Point {
    int x, y;
    void reset () {
        x=0; y=0; }
}
```

- When an object is needed, a primitive (int, char, etc.) or a struct will be automatically boxed or unboxed without explicit use of a wrapper class (like `Integer` or `Character`). Thus if method `List.add` expects an object, you may write

```csharp
List.add(123);
```
and 123 will be boxed into an `Integer` object automatically.

- Enumerations are provided:

```csharp
enum Color {Red, Blue, Green};
```

- Rectangular arrays are provided:

```csharp
int[,] multi = new int[5,5];
```

- Reference, out and variable-length parameter lists are allowed.

- Pointers may be used in methods marked `unsafe`.
One of the newest and most innovative scripting languages is Python, developed by Guido van Rossum in the mid-90s. Python is named after the BBC “Monty Python” television series. Python blends the expressive power and flexibility of earlier scripting languages with the power of object-oriented programming languages. It offers a lot to programmers:

- An interactive development mode as well as an executable “batch” mode for completed programs.
- Very reasonable execution speed. Like Java, Python programs are compiled. Also like Java, the compiled code is in an intermediate language for which an interpreter is written. Like Java this insulates Python from many of the vagaries of the actual machines on which it runs, giving it portability of an equivalent level to that of Java. Unlike Java, Python retains the interactivity for which interpreters are highly prized.
- Python programs require no compilation or linking. Nevertheless, the semi-compiled Python program still runs much faster than its traditionally interpreted rivals such as the shells, awk and perl.
- Python is freely available on almost all platforms and operating systems (Unix, Linux, Windows, MacOS, etc.)

- Python is completely object oriented. It comes with a full set of object oriented features.
- Python presents a first class object model with first class functions and multiple inheritance. Also included are classes, modules, exceptions and late (run-time) binding.
- Python allows a clean and open program layout. Python code is less cluttered with the syntactic “noise” of declarations and scope definitions. Scope in a Python program is defined by the indentation of the code in question. Python thus breaks with current language designs in that white space has now once again acquired significance.

- Like Java, Python offers automated memory management through run-time reference counting and garbage collection of unreferenced objects.
- Python can be embedded in other products and programs as a control language.
- Python’s interface is well exposed and is reasonably small and simple.
- Python’s license is truly public. Python programs can be used or sold without copyright restrictions.
- Python is extendable. You can dynamically load compiled Python, Python source, or even dynamically load new machine (object) code to provide new features and new facilities.
• Python allows low-level access to its interpreter. It exposes its internal plumbing to a significant degree to allow programs to make use of the way the plumbing works.

• Python has a rich set of external library services available. This includes, network services, a GUI API (based on tcl/Tk), Web support for the generation of HTML and the CGI interfaces, direct access to databases, etc.

Using Python

Python may be used in either interactive or batch mode.

In interactive mode you start up the Python interpreter and enter executable statements. Just naming a variable (a trivial expression) evaluates it and echoes its value. For example (`>>>` is the Python interactive prompt):

```python
>>> 1
1
>>> a=1
>>> a
1
>>> b=2.5
>>> b
2.5
```

You can also incorporate Python statements into a file and execute them in batch mode. One way to do this is to enter the command `python file.py`

where `file.py` contains the Python code you want executed. Be careful though; in batch mode you must use a `print` (or some other output statement) to force output to be printed. Thus

```python
>>> a+b
3.5
>>> print a+b
3.5
```

when run in batch mode prints only `3.5` (the output of the `print` statement).

You can also run Python programs as Unix shell scripts by adding the line `#!/usr/bin/env python` to the head of your Python file.

(Since `#` begins Python comments, you can also feed the same augmented file directly to the Python interpreter)
Python Command Format

In Python, individual primitive commands and expressions must appear on a single line.
This means that

\[
a = 1 \\
+ b
\]

does not assign \(1+b\) to \(a\). Rather, it assigns 1 to \(a\), then evaluates \(+b\).
If you wish to span more than one line, you must use \(\backslash\) to escape the line:

\[
a = 1 \backslash \\
+ b
\]

is equivalent to

\[
a = 1 + b
\]

Compound statements, like if statements and while loops, can span multiple lines, but individual statements within an if or while (if they are primitive) must appear on a single line.

Why this restriction?

With it, ; ’s are mostly unnecessary!
A ; at the end of a statement is legal but usually unnecessary, as the end-of-line forces the statement to end.
You can use a ; to squeeze more than one statement onto a line, if you wish:

\[
a=1; b=2 ; c=3
\]

Identifiers and Reserved Words

Identifiers look much the same as in most programming languages. They are composed of letters, digits and underscores. Identifiers must begin with a letter or underscore. Case is significant. As in C and C++, identifiers that begin with an underscore often have special meaning.

Python contains a fairly typical set of reserved words:

- and
- del
- for
- is
- raise
- assert
- elif
- from
- lambda
- return
- break
- else
- global
- not
- try
- class
- except
- if
- or
- while
- continue
- exec
- import
- pass
- def
- finally
- in
- print

Numeric Types

There are four numeric types:

1. Integers, represented as a 32 bit (or longer) quantity. Digits sequences (possibly) signed are integer literals:

\[
1  \quad -123  \quad +456
\]

2. Long integers, of unlimited precision. An integer literal followed by an \(l\) or \(L\) is a long integer literal:

\[
12345678900000000000000L
\]

3. Floating point values, represented as a 64 bit floating point number. Literals are of fixed decimal or exponential form:

\[
123.456  \quad 1e10  \quad 6.0231023
\]

4. Complex numbers, represented as a pair of floating point numbers. In complex literals \(j\) or \(J\) is used to
denote the imaginary part of the complex value:
\[ 1.0+2.0j \quad -22.1j \quad 10e10j+20.0 \]

There is no character type. A literal like 'a' or "c" denotes a string of length one.

There is no boolean type. A zero numeric value (any form), or None (the equivalent of void) or an empty string, list, tuple or dictionary is treated as false; other values are treated as true.

Hence
"abc" and "def"
is treated as true in an if, since both strings are non-empty.

<table>
<thead>
<tr>
<th>Op</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>**</td>
<td>Exponentiation</td>
</tr>
<tr>
<td>+</td>
<td>Unary plus</td>
</tr>
<tr>
<td>-</td>
<td>Unary minus</td>
</tr>
<tr>
<td>~</td>
<td>Bit-wise complement (int or long only)</td>
</tr>
<tr>
<td>*</td>
<td>Multiplication</td>
</tr>
<tr>
<td>/</td>
<td>Division</td>
</tr>
<tr>
<td>%</td>
<td>Remainder</td>
</tr>
<tr>
<td>+</td>
<td>Binary plus</td>
</tr>
<tr>
<td>-</td>
<td>Binary minus</td>
</tr>
<tr>
<td>&lt;&lt;</td>
<td>Bit-wise left shift (int or long only)</td>
</tr>
<tr>
<td>&gt;&gt;</td>
<td>Bit-wise right shift (int or long only)</td>
</tr>
<tr>
<td>&amp;</td>
<td>Bit-wise and (int or long only)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>^</td>
<td>Bit-wise Xor (int or long only)</td>
</tr>
<tr>
<td>&lt;</td>
<td>Less than</td>
</tr>
<tr>
<td>&gt;</td>
<td>Greater than</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Operator Precedence Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listed from lowest to highest:</td>
</tr>
<tr>
<td>or</td>
</tr>
<tr>
<td>and</td>
</tr>
<tr>
<td>not</td>
</tr>
<tr>
<td>&lt;, &lt;=, &gt;, &gt;=, &lt;&gt;, !=, ==</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>^</td>
</tr>
<tr>
<td>&amp;</td>
</tr>
<tr>
<td>&lt;&lt;, &gt;&gt;</td>
</tr>
<tr>
<td>+, -</td>
</tr>
<tr>
<td>*, /, %</td>
</tr>
<tr>
<td>**</td>
</tr>
<tr>
<td>+, -</td>
</tr>
<tr>
<td>~</td>
</tr>
</tbody>
</table>
**Arithmetic Operator Use**

Arithmetic operators may be used with any arithmetic type, with conversions automatically applied. Bit-wise operations are restricted to integers and long integers. The result type is determined by the “generality” of the operands. (Long is more general than int, float is more general than both int and long, complex is the most general numeric type). Thus

```python
>>> 1+2
3
>>> 1+111L
112L
>>> 1+1.1
2.1
>>> 1+2.0j
(1+2j)
```

**Assignment Statements**

In Python assignment is a statement not an expression.

Thus

```python
a+(b=2)
```

is illegal.

Chained assignments are allowed:

```python
a = b = 3
```

Since Python is dynamically typed, the type (and value) associated with an identifier can change because of an assignment:

```python
>>> a = 0
>>> print a
0
>>> a = a + 0L
>>> print a
0L
>>> a = a + 0.0
>>> print a
0.0
>>> a = a + 0.0j
>>> print a
0j
```
If and While Statements

Python contains if and while statements that are fairly similar to those found in C and Java.

There are some significant differences though.

A line that contains an if, else or while ends in a "": Thus we might write:

```python
if a > 0:
    b = 1
```

Moreover the indentation of the then part is significant! You don't need { and } in Python because all statements indented at the same level are assumed to be part of the same block.

In the following Python statements

```python
if a>0:
    b=1
c=2
d=3
```

the assignments to b and c constitute then part; the assignment to a follows the if statement, and is independent of it. In interactive mode a blank line is needed to allow the interpreter to determine where the if statement ends; this blank line is not needed in batch mode.

The if Statement

The full form of the if statement is

```python
if expression:
    statement(s)
elif expression:
    statement(s)
...
else:
    statement(s)
```

Note those pesky :'s at the end of the if, elif and else lines. The expressions following the if and optional elif lines are evaluated until one evaluates to true. Then the following statement(s), delimited by indentation, are executed. If no expression evaluates to true, the statements following the else are executed.

Use of else and elif are optional; a "bare" if may be used.

If any of the lists of statements is to be null, use pass to indicate that nothing is to be done.

For example

```python
if a>0:
    b=1
elif a < 0:
    pass
else:
    b=0
```

This if sets b to 1 if a is > 0; it sets b to 0 if a == 0, and does nothing if a < 0.
### While Loops

Python contains a fairly conventional while loop:

```python
while expression:
    body
```

Note the ":" that ends the header line. Also, indentation delimits the body of the loop; no braces are needed. For example,

```python
>>> a=0; b=0
>>> while a < 5:
...     b = b+a**2
...     a= a+1
... >>> print a,b
5 30
```

### Break, Continue and Else in Loops

Like C, C++ and Java, Python allows use of break within a loop to force loop termination. For example,

```python
>>> a=1
>>> while a < 10:
...     if a+a == a**2:
...         break
...     else:
...         a=a+1
...
>>> print a
2
```

A `continue` may be used to force the next loop iteration:

```python
>>> a=1
>>> while a < 100:
...     a=a+1
...     if a%2==0:
...         continue
...     a=3*a
...
>>> print a
105
```

Python also allows you to add an else clause to a while (or for) loop. The syntax is

```python
while expression:
    body
else:
    statement(s)
```

The else statements are executed when the termination condition becomes false, but not when the loop is terminated with a break. As a result, you can readily program “search loops” that need to handle the special case of search failure:
>>> a=1
>>> while a < 1000:
    ... if a**2 == 3*a-1:
    ...     print "winner: ",a
    ...     break
    ... a=a+1
    ... else:
    ...     print "No match"
    ...
No match

Sequence Types

Python includes three sequence types: strings, tuples and lists. All sequence types may be indexed, using a very general indexing system.

Strings are sequences of characters; tuples and lists may contain any type or combination of types (like Scheme lists).

Strings and tuples are immutable (their components may not be changed). Lists are mutable, and be updated, much like arrays.

Strings may be delimited by either a single quote (') or a double quote ("), or even a triple quote (''' or """'). A given string must start and stop with the same delimiter. Triply quoted strings may span multiple lines. There is no character type or value; characters are simply strings of length 1. Legal strings include

'abc' "xyz" '''It's OK!'''

Lists are delimited by "[" and "]". Empty (or null lists) are allowed. Valid list literals include

[1,2,3] ['one',1]
[['a'],['b'],['c']] []

Tuples are a sequence of values separated by commas. A tuple may be enclosed within parentheses, but this isn't required. A empty tuple is (). A singleton tuple ends with a comma (to distinguish it from a simple scalar value).

Thus (1,) or just 1, is a valid tuple of length one.

Indexing Sequence Types

Python provides a very general and powerful indexing mechanism. An index is enclosed in brackets, just like a subscript in C or Java. Indexing starts at 0.

Thus we may have

```python
>>> 'abcde'[2]
'c'
>>> [1,2,3,4,5][1]
2
>>> (1.1,2.2,3.3)[0]
1.1
```

Using an index that's too big raises an IndexError exception:

```python
>>> 'abcde'[3]
IndexError: string index out of range
```
Unlike most languages, you can use negative index values; these simply index from the right:

```python
>>> 'abc'[-1]
'c'
>>> [5, 4, 3, 2, 1][-2]
2
>>> (1, 2, 3, 4)[-4]
1
```

You may also access a slice of a sequence value by supplying a range of index values. The notation is `data[i:j]`, which selects the values in `data` that are \( \geq i \) and \( < j \). Thus

```python
>>> 'abcde'[1:2]
'b'
>>> 'abcde'[0:3]
'abc'
```

You may omit a lower or upper bound on a range. A missing lower bound defaults to 0 and a missing upper bound defaults to the maximum legal index. For example,

```python
>>> [1, 2, 3, 4, 5][2:]
[3, 4, 5]
>>> [1, 2, 3, 4, 5][:3]
[1, 2, 3]
```

An upper bound that’s too large in a range is interpreted as the maximum legal index:

```python
>>> 'abcdef'[3:100]
'def'
```

You may use negative values in ranges too—they’re interpreted as being relative to the right end of the sequence:

```python
>>> 'abcde'[0:-2]
'abc'
>>> 'abcdefg'[-5:-2]
'cde'
>>> 'abcde'[-3:]
'cde'
>>> 'abcde'[:1]
'abc'
```

The length of the value assigned to a slice need not be the same size as the slice itself, so you can shrink or expand a list by assigning slices:

```python
>>> a=[1, 2, 3, 4, 5]
>>> a[2:3]=[3.1, 3.2]
>>> a
[1, 2, 3.1, 3.2, 4, 5]
>>> a[4:]=[]
>>> a
[1, 2, 3.1, 3.2]
>>> a[:0]=[-3, -2, -1]
>>> a
[-3, -2, -1, 1, 2, 3.1, 3.2]
```
Other Operations on Sequences

Besides indexing and slicing, a number of other useful operations are provided for sequence types (strings, lists and tuples).

These include:

- \texttt{+} (concatenation):
  
  \begin{verbatim}
  >>> [1, 2, 3] + [4, 5, 6]
  [1, 2, 3, 4, 5, 6]
  >>> (1, 2, 3) + (4, 5)
  (1, 2, 3, 4, 5)
  >>> (1, 2, 3) + [4, 5]
  TypeError: illegal argument type for built-in operation
  >>> "abc" + "def"
  'abcdef'
  \end{verbatim}

- \texttt{*} (Repetition):
  
  \begin{verbatim}
  >>> 'abc'*2
  'abcabc'
  >>> [3, 4, 5]*3
  [3, 4, 5, 3, 4, 5, 3, 4, 5]
  \end{verbatim}

- Membership (\texttt{in}, not \texttt{in}):
  
  \begin{verbatim}
  >>> 3 in [1, 2, 3, 4]
  1
  >>> 'c' in 'abcde'
  1
  \end{verbatim}

- \texttt{max} and \texttt{min}:
  
  \begin{verbatim}
  >>> max([3, 8, -9, 22, 4])
  22
  >>> min('aa', 'bb', 'abc')
  'aa'
  \end{verbatim}

Operations on Lists

As well as the operations available for all sequence types (including lists), there are many other useful operations available for lists. These include:

- \texttt{count} (Count occurrences of an item in a list):
  
  \begin{verbatim}
  >>> [1, 2, 3, 3, 21].count(3)
  2
  >>> [1, 2, 3, 3, 21].count(17)
  ValueError: list.count(x): x not in list
  \end{verbatim}

- \texttt{index} (Find first occurrence of an item in a list):
  
  \begin{verbatim}
  >>> [1, 2, 3, 3, 21].index(3)
  2
  >>> [1, 2, 3, 3, 21].index(17)
  ValueError: list.index(x): x not in list
  \end{verbatim}

- \texttt{remove} (Find and remove an item from a list):
  
  \begin{verbatim}
  >>> a=[1, 2, 3, 4, 5]
  >>> a.remove(4)
  >>> a
  [1, 2, 3, 5]
  >>> a.remove(17)
  ValueError: list.remove(x): x not in list
  \end{verbatim}

- \texttt{pop} (Fetch and remove \texttt{i}-th element of a list):
  
  \begin{verbatim}
  >>> a=[1, 2, 3, 4, 5]
  >>> a.pop(3)
  4
  >>> a
  [1, 2, 3, 5]
  >>> a.pop()
  5
  >>> a
  [1, 2, 3]
  \end{verbatim}
. reverse a list:
  >>> a=[1,2,3,4,5]
  >>> a.reverse()
  >>> a
  [5, 4, 3, 2, 1]
.
. sort a list:
  >>> a=[5,1,4,2,3]
  >>> a.sort()
  >>> a
  [1, 2, 3, 4, 5]
.
. Create a range of values:
  >>> range(1,5)
  [1, 2, 3, 4]
  >>> range(1,10,2)
  [1, 3, 5, 7, 9]

Dictionaries

Python also provides a dictionary type (sometimes called an associative array). In a dictionary you can use a number (including a float or complex), string or tuple as an index. In fact any immutable type can be an index (this excludes lists and dictionaries).

An empty dictionary is denoted `{ }`. A non-empty dictionary may be written as

```python
{ key_1: value_1, key_2: value_2, ... }
```

For example,

```python
c={ 'bmw':545, 'lexus':'sc 430', 'mercedes':'S 500'}
```

You can use a dictionary much like an array, indexing it using keys, and updating it by assigning a new value to a key:

```python
>>> c['bmw']
545
>>> c['bmw']='m5'
>>> c['honda']='accord'
You can delete a value using del:

```python
>>> del c['honda']
>>> c['honda']
KeyError: honda
```

You can also check to see if a given key is valid, and also list all keys, values, or key-value pairs in use:

```python
>>> c.has_key('edsel')
0
>>> c.keys()
['bmw', 'mercedes', 'lexus']
>>> c.values()
['m5', 'S 500', 'sc 430']
>>> c.items()
[('bmw', 'm5'), ('mercedes', 'S 500'), ('lexus', 'sc 430')]
```
For Loops

In Python's `for` loops, you don't explicitly control the steps of an iteration. Instead, you provide a sequence type (a string, list or sequence), and Python automatically steps through the values.

Like a `while` loop, you must end the `for` loop header with a `:` and the body is delimited using indentation. For example,

```python
>>> for c in 'abc':
...    print c
...a
b
c
```

You can use an `else` with `for` loops too. Once the values in the specified sequence are exhausted, the `else` is executed unless the `for` is exited using a `break`. For example,

```python
for i in a:
    if i < 0:
        print 'Neg val:', i
        break
    else:
        print 'No neg vals'
```

The `range` function, which creates a list of values in a fixed range is useful in `for` loops:

```python
>>> a=[5,2,1,4]
>>> for i in range(0,len(a)):
...    a[i]=2*a[i]
...
>>> print a
[10, 4, 2, 8]
```

Function Definitions

Function definitions are of the form

```python
def name(args):
    body
```

The symbol `def` tells Python that a function is to be defined. The function is called `name` and `args` is a tuple defining the names of the function's arguments. The `body` of the function is delimited using indentation. For example,

```python
def fact(n):
    if n<=1:
        return 1
    else:
        return n*fact(n-1)
>>> fact(5)
120
>>> fact(20L)
```
Examples of function calls in Python:

```python
>>> fact(2.5)
3.75
>>> fact(2+1j)
(1+3j)
```

Scalar parameters are passed by value; mutable objects are allocated in the heap and hence are passed (in effect) by reference:

```python
>>> def asg(ar):
...     a[1]=0
...     print ar
...     ...
>>> a=[1,2,3,4.5]
>>> asg(a)
[1, 0, 3, 4.5]
```

Arguments may be given a default value, making them optional in a call. Optional parameters must follow required parameters in definitions. For example,

```python
>>> def expo(val,exp=2):
...     return val**exp
...     ...
>>> expo(3,3)
27
>>> expo(3)
9
>>> expo()
TypeError: not enough arguments; expected 1, got 0
```

A variable number of arguments is allowed; you prefix the last formal parameter with a *; this parameter is bound to a tuple containing all the actual parameters provided by the caller:

```python
>>> def sum(*args):
...     sum=0
...     for i in args:
...         sum=sum+i
...     return sum
...     ...
>>> sum(1,2,3)
6
>>> sum(2)
2
>>> sum()
0
```

You may also use the name of formal parameters in a call, making the order of parameters less important:

```python
>>> def cat(left="[",body="",right="]"):
...     return left+body+right
...     ...
>>> cat(body='xyz');
'[xyz]'
>>> cat(body='hi there!',left='--[')
'--[hi there!]'
```
Scoping Rules in Functions

Each function body has its own local namespace during execution. An identifier is resolved (if possible) in the local namespace, then (if necessary) in the global namespace. Thus

```python
>>> def f():
...   a=11
...   return a+b
...
>>> b=2; f()
13
>>> a=22; f()
13
>>> b=33; f()
44
```

Assignments are to local variables, even if a global exists. To force an assignment to refer to a global identifier, you use the declaration `global id` which tells Python that in this function `id` should be considered global rather than local. For example,

```python
>>> a=1; b=2
>>> def f():
...   global a
...   a=111; b=222
...
>>> f(); print a, b
111 2
```

Other Operations on Functions

Since Python is interpreted, you can dynamically create and execute Python code.

The function `eval(string)` interprets `string` as a Python expression (in the current execution environment) and returns what is computed. For example,

```python
>>> a=1; b=2
>>> eval('a+b')
3
```

`exec(string)` executes `string` as arbitrary Python code (in the current environment):

```python
>>> a=1; b=2
>>> exec('for op in "+-*/":
\n    print(eval("a"+op+"b"))')
3
-1
2
0
```

`execfile(string)` executes the contents of the file whose pathname is specified by `string`. This can be useful in loading an existing set of Python definitions.
The expression

\[ \text{lambda args: expression} \]

creates an anonymous function with
\[ \text{args} \]
as its argument list and
\[ \text{expression} \]
as its body. For example,

\[
\text{>>> (lambda a:a+1)(2)}
\]

3

And there are definitions of \text{map}, \text{reduce} and \text{filter} to map a
function to a list of values, to reduce
a list (using a binary function) and to
select values from a list (using a
predicate):

\[
\text{>>> def double(a):}
... 
... \text{return 2*a;}
... 
\text{>>> map(double, [1,2,3,4])}
\]

[2, 4, 6, 8]

\[
\text{>>> def sum(a,b):}
... \text{return a+b}
... 
\text{>>> reduce(sum, [1,2,3,4,5])}
\]

15

\[
\text{>>> def even(a):}
... \text{return not(a%2)}
... 
\text{>>> filter(even, [1,2,3,4,5])}
\]

[2, 4]

I/O in Python

The easiest way to print information
in Python is the \text{print} statement.
You supply a list of values separated
by commas. Values are converted to
strings (using the \text{str()} function)
and printed to standard out, with a
terminating new line automatically
included. For example,

\[
\text{>>> print "1+1=",1+1}
1+1= 2
\]

If you don’t want the automatic end
of line, add a comma to the end of
the print list:

\[
\text{>>> for i in range(1,11):}
... \text{print i,}
... 
1 2 3 4 5 6 7 8 9 10
\]

For those who love C’s \text{printf},
Python provides a nice formatting
capability using a printf- like
notation. The expression

\[ \text{format % tuple} \]

formats a tuple of values using a
format string. The detailed formatting
rules are those of C’s \text{printf}. Thus

\[
\text{>>> print "%d+%d=%d" % (10,20,10+20)}
10+20=30
\]
**File-oriented I/O**

You open a file using
\[ \text{open(name, mode)} \]
which returns a “file object.”

- `name` is a string representing the file’s path name;
- `mode` is a string representing the desired access mode (‘r’ for read, ‘w’ for write, etc.).

Thus
\[ \text{f} = \text{open("/tmp/f1", "w")}; \]
\[ \text{f} \]
opens a temp file for writing.

The command
\[ \text{f.read(n)} \]
reads \(n\) bytes (as a string).

\[
\begin{align*}
\text{f.read()} & \text{ reads the whole file into a string. At end-of-file, f.read returns the null string:} \\
\text{f.readline()} & \text{ reads a whole line of input, and f.readlines() reads the whole input file into a list of strings:}
\end{align*}
\]

Classes in Python

Python contains a class creation mechanism that’s fairly similar to what’s found in C++ or Java.

There are significant differences though:
- All class members are public.
- Instance fields aren’t declared. Rather, you just create fields as needed by assignment (often in constructors).
- There are class fields (shared by all class instances), but there are no class methods. That is, all methods are instance methods.
Defining Classes

You define a class by executing a class definition of the form

```python
class name:
    statement(s)
```

A class definition creates a class object from which class instances may be created (just like in Java). The statements within a class definition may be data members (to be shared among all class instances) as well as function definitions (prefixed by a `def` command). Each function must take (at least) an initial parameter that represents the class instance within which the function (instance method) will operate. For example,

```python
class Example:
    cnt=1
    def msg(self):
        print "Bo"+"o"*Example.cnt+"!"*self.n

>>> Example.cnt
1
>>> Example.msg
<unbound method Example.msg>
Example.msg is unbound because we haven't created any instances of the Example class yet.

We create class instances by using the class name as a function:

```python
>>> e=Example()
>>> e.msg()
AttributeError: n
```

We get the `AttributeError` message regarding `n` because we haven't defined `n` yet! One way to do this is to just assign to it, using the usual field notation:

```python
>>> e.n=1
>>> e.msg()
Boo!
>>> e.n=2;Example.cnt=2
>>> e.msg()
Booo!!
```

We can also call an instance method by making the class object an explicit parameter:

```python
>>> Example.msg(e)
Booo!!
```

It's nice to have data members initialized when an object is created. This is usually done with a constructor, and Python allows this too.
A special method named __init__ is called whenever an object is created. This method takes self as its first parameter; other parameters (possibly made optional) are allowed. We can therefore extend our Example class with a constructor:

```python
class Example:
    cnt=1
    def __init__(self, nval=1):
        self.n = nval
    def msg(self):
        print "Bo" + "o" * Example.cnt + "!" * self.n
```

```python
>>> e = Example()
>>> e.n
1
>>> f = Example(2)
>>> f.n
2
```

You can also define the equivalent of Java's toString method by defining a member function named __str__(self).

For example, if we add

```python
def __str__(self):
    return "<%d>" % self.n
```
to Example,

then we can include Example objects in print statements:

```python
>>> e = Example(2)
>>> print e
<2>
```

### Inheritance

Like any language that supports classes, Python allows inheritance from a parent (or base) class. In fact, Python allows multiple inheritance in which a class inherits definitions from more than one parent. When defining a class you specify parents classes as follows:

```python
class name (parent classes):
    statement(s)
```

The subclass has access to its own definitions as well as those available to its parents. All methods are virtual, so the most recent definition of a method is always used.

```python
class C:
    def DoIt(self):
        self.PrintIt()
    def PrintIt(self):
        print "C rules!"

class D(C):
    def PrintIt(self):
        print "D rules!"
    def TestIt(self):
        self.DoIt()

dvar = D()
dvar.TestIt()
```

D rules!
If you specify more than one parent for a class, lookup is depth-first, left to right, in the list of parents provided. For example, given

class A(B,C): ...
we first look for a non-local definition in B (and its parents), then in C (and its parents).

Operator Overloading

You can overload definitions of all of Python's operators to apply to newly defined classes. Each operator has a corresponding method name assigned to it. For example, + uses __add__, - uses __sub__, etc.

Given

class Triple:
    def __init__(self,A=0,B=0,C=0):
        self.a=A
        self.b=B
        self.c=C
    def __str__(self):
        return "(%d,%d,%d)"% (self.a,self.b,self.c)
    def __add__(self,other):
        return Triple(self.a+other.a,
                      self.b+other.b,
                      self.c+other.c)

the following code

t1=Triple(1,2,3)
t2=Triple(4,5,6)
print t1+t2
produces
    (5, 7, 9)

Exceptions

Python provides an exception mechanism that's quite similar to the one used by Java.
You "throw" an exception by using a raise statement:

    raise exceptionValue

There are numerous predefined exceptions, including
OverflowError (arithmetic overflow), EOFError (when end-of-file is hit), NameError (when an undeclared identifier is referenced), etc.
You may define your own exceptions as subclasses of the predefined class `Exception`:
```python
class badValue(Exception):
    def __init__(self, val):
        self.value = val
```

You catch exceptions in Python's version of a `try` statement:
```python
try:
    statement(s)
except exceptionName_1, id_1:
    statement(s)
...
except exceptionName_n, id_n:
    statement(s)
```

As was the case in Java, an exception raised within the `try` body is handled by an `except` clause if the raised exception matches the class named in the `except` clause. If the raised exception is not matched by any `except` clause, the next enclosing `try` is considered, or the exception is reraised at the point of call.

For example, using our `badValue` exception class,
```python
def sqrt(val):
    if val < 0.0:
        raise badValue(val)
    else:
        return cmath.sqrt(val)
```
```
try:
    print "Ans =", sqrt(-123.0)
except badValue, b:
    print "Can't take sqrt of", b.value
```
When executed, we get

`Ans = Can't take sqrt of -123.0`

---

**Modules**

Python contains a module feature that allows you to access Python code stored in files or libraries. If you have a source file `mydefs.py` the command
```python
import mydefs
```
will read in all the definitions stored in the file. What's read in can be seen by executing
```python
dir(mydefs)
```
To access an imported definition, you qualify it with the name of the module. For example,
```python
mydefs.fct
```
accesses `fct` which is defined in module `mydefs`.

To avoid explicit qualification you can use the command
```python
from modulename import id_1, id_2, ...
```
This makes `id_1, id_2, ...` available without qualification. For example,
```python
>>> from test import sqrt
>>> sqrt(123)
(11.0905365064+0j)
```
You can use the command
```python
from modulename import *
```
to import (without qualification) all the definitions in modulename.
The Python Library

One of the great strengths of Python is that it contains a vast number of modules (at least several hundred) known collectively as the Python Library. What makes Python really useful is the range of prewritten modules you can access. Included are network access modules, multimedia utilities, data base access, and much more.

See

www.python.org/doc/lib

for an up-to-date listing of what’s available.