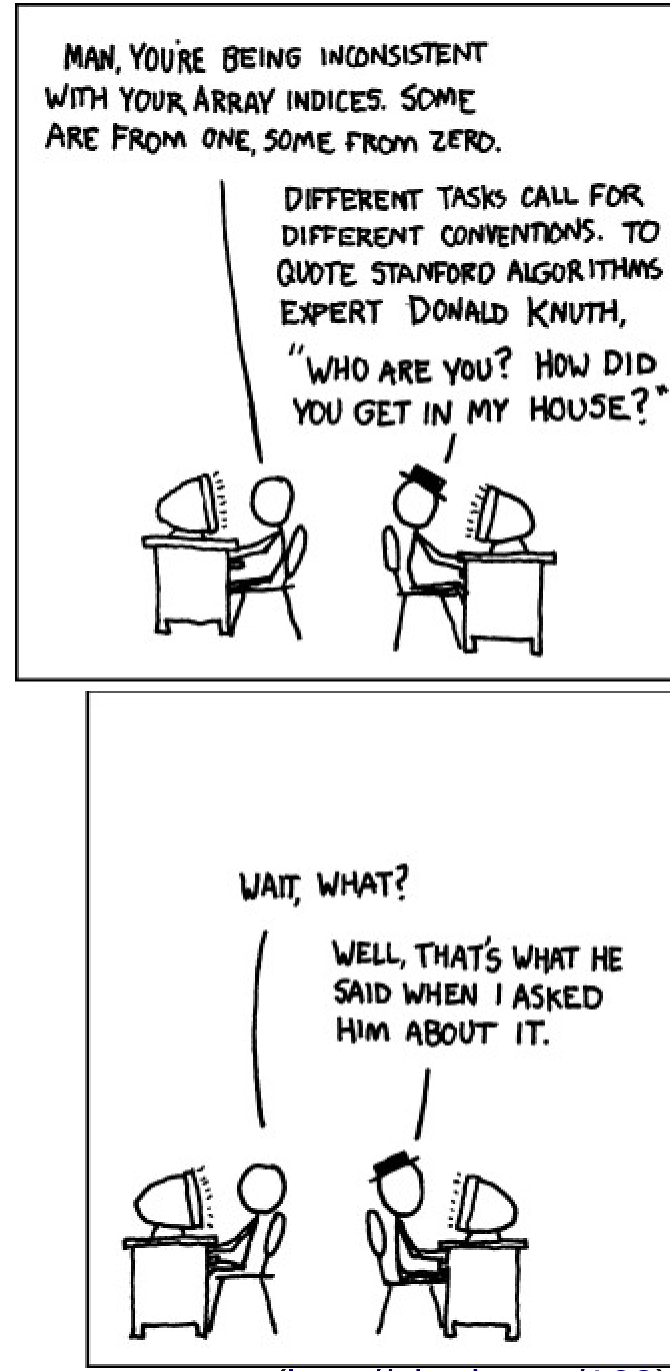
Programming Concepts: Containers

Paulo Penteado

http://www.ppenteado.net/pc/



(http://xkcd.com/163)

Containers

A single value in a variable is not enough.

Containers – variables that hold several values (the elements)

There are many ways to organize the elements: arrays are just one of them

- Each way is implementing some *data structure**.

There is no “best container”:

- Each is best suited to different problems

The 3 main properties of containers:

- **Homogeneous X heterogeneous**: whether all elements are the same type.
- **Static X dynamic**: whether the number of elements is fixed.
- **Sequentiality**:
  - Sequential containers: elements stored by order, and are accessed by indices.
  - Non-sequential containers: elements stored by name or through relationships.

**A data structure* is a way of organizing data; a *structure* is just one of them.

Containers

The most common types (names vary among languages; some have several implementations for the same type)*:

- **Array / vector / matrix (1D or MD)**: C, C++, Fortran, IDL, Java, Python+Numpy, R
- **List**: C++, Python, IDL (≥8), Java, R, Perl**
- **Map / hash / hashtable / associative array / dictionary**: C++, Python, IDL (≥8), Java, R***, Perl
- **Set**: C++, Python, Java, R
- **Tree / heap**: C++, Python, Java
- **Stack**: C++, Python, Java
- **Queue**: C++, Python, Java

*Listed only when the structure is part of a language's standard library.

**A Perl array is more like a list than an array.

***Which in R are also called *named lists.*

Arrays - definition

The simplest container.

A sequential set o elements, organized **regularly**, in 1D or more (MD).

Not natively present in some recent languages (Perl, Python without Numpy).

Sometimes called **array** only when more than 1D, being called **vector** in the 1D case.

2D sometimes called tables or **matrices**

- In some languages (ex: R, Python+Numpy), ***matrix*** is different from a generic array.

Arrays - characteristics

Homogeneous (all elements must be the same **type**)

Static (cannot change the number of elements)

- “Dynamic arrays” are actually creating new arrays, and throwing away the old ones on resize (which is inefficient).

Sequential (elements stored by an order)

Organized in 1D or more (MD).

Element access through their indices (sequential integer numbers).

Usually, **the most efficient container for random and sequential access.**

Provide the means to do vectorization (do operations on the whole array, or parts of the array, with a single statement).

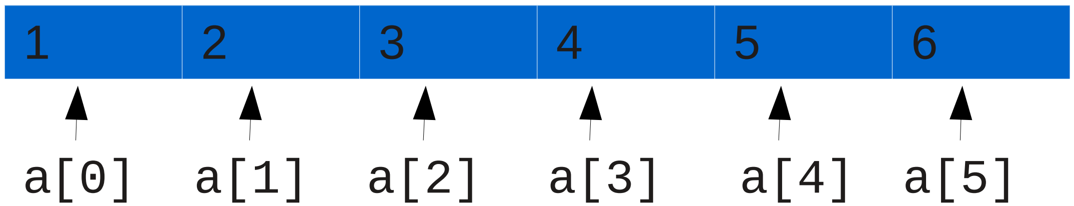
- 1D arrays are common.
- MD arrays are often awkward (2D may not be so bad): **IDL and Python+Numpy have high level MD operations**.

Internally all elements are **stored as a 1D array, even when there are more dimensions** (memory and files are 1D).

- **When over 1D, they are always regular** (each dimension has a constant number of elements).

Arrays

1D



Ex. (IDL):

```
IDL> a=bindgen(6)+1
```

→ Generates an array of type ***byte***, with 6 elements, valued 1 to 6.

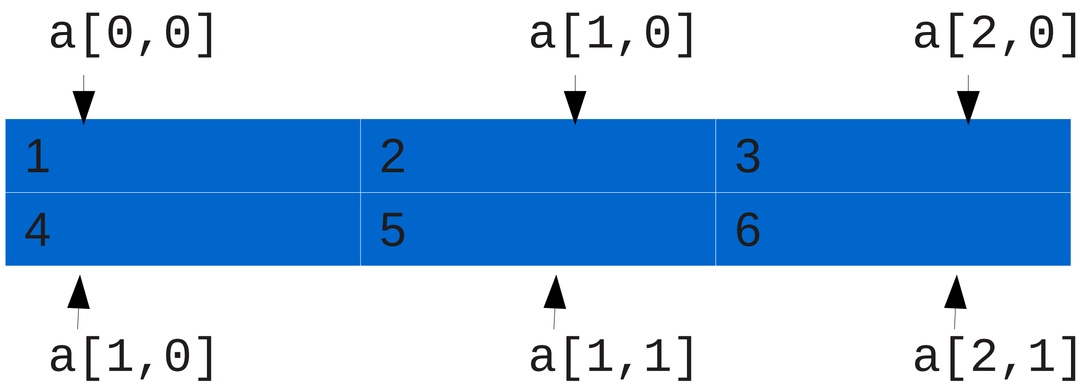
```
IDL> help,a
A               INT       = Array[6]
```

```
IDL> print,a
       1       2       3       4       5       6
```

Most often, indexes start at 0. In some languages, the start index can be chosen.

Arrays

2D



Ex. (IDL):

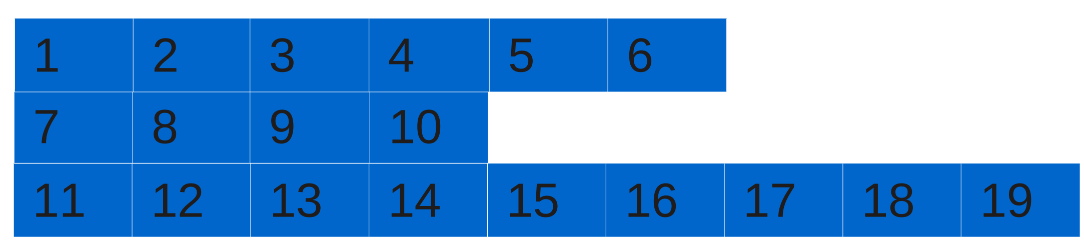
```
IDL> a=bindgen(3,2)+1
```

Generates an array of type ***byte***, with 6 elements, in 3 columns by 2 rows, valued 1 to 6.

```
IDL> help,a
A               INT       = Array[3, 2]

IDL> print,a
       1       2       3
       4       5       6
```

Must be regular: cannot be like



Arrays

3D is usually thought, graphically, as pile of “pages”, each page being a 2D table. Or as a brick. Ex. (IDL):

```
IDL> a=bindgen(4,3,3)
```

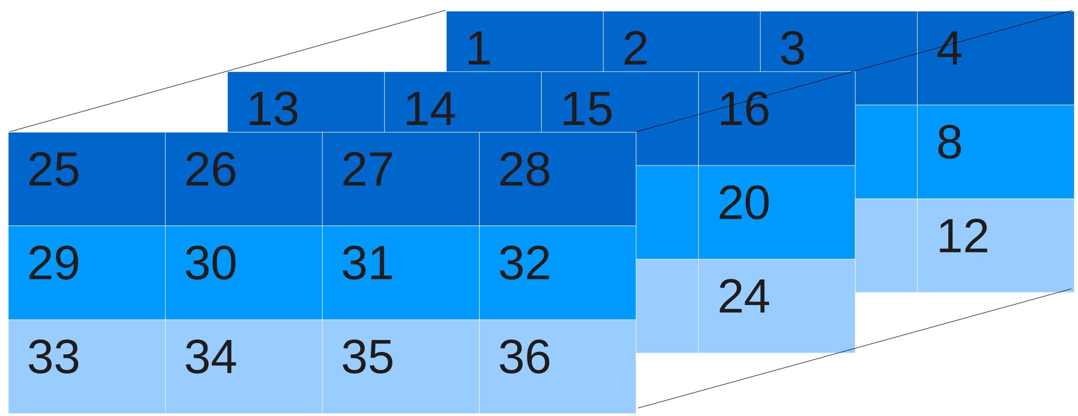
Generates an array of type ***byte***, with 36 elements, over 4 columns, 3 rows, 3 “pages”, valued 0 to 35.

```
IDL> help,a
A                BYTE      = Array[4, 3, 3]

IDL> print,a
   0   1   2   3
   4   5   6   7
   8   9  10  11

  12  13  14  15
  16  17  18  19
  20  21  22  23

  24  25  26  27
  28  29  30  31
  32  33  34  35
```



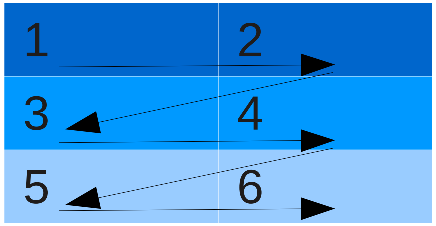
Beyond 3D, graphical representations get awkward (sets of 3D arrays for 4D, sets of 4D for 5D, etc.)

Arrays – MD storage

Internally, they are always 1D

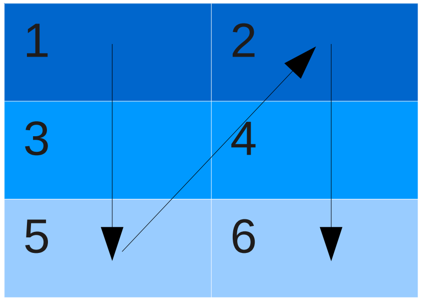
The dimensions are scanned sequentially. Ex (2D): a[2,3] - 6 elements:

1)
`a[0,0] a[1,0] a[2,0] a[0,1] a[1,1] a[2,1]`
Memory position:
`0      1      2      3      4      5`



or

2)
`a[0,0] a[0,1] a[1,0] a[1,1] a[2,0] a[2,1]`
Memory position:
`0      1      2      3      4      5`



Each language has its choice of dimension order:

Column major – first dimension is contiguous (1 above): IDL, Fortran, R, **Python+Numpy**
Row major – last dimension is contiguous (2 above): C, C++, Java, **Python+Numpy**

Note that languages / people may differ in the use of the terms ***row*** and ***column***.

Graphically, usually the "horizontal" dimension (shown over a line) can be either the first of the last. Usually the horizontal dimension is the contiguous.

Arrays – basic usage

Access to individual elements, through the M indices (MD), or single index (MD or 1D). Ex. (IDL):

Return arrays of **doubles** where each element has the value of its index.

```
IDL> a=dindgen(4)
IDL> b=dindgen(2,3)
IDL> help,a
A               DOUBLE    = Array[4]
IDL> help,b
B               DOUBLE    = Array[2, 3]
IDL> print,a
       0.0000000       1.0000000       2.0000000       3.0000000
IDL> print,b
       0.0000000       1.0000000
       2.0000000       3.0000000
       4.0000000       5.0000000
IDL> print,a[2]
       2.0000000
IDL> print,a[-1]
       3.0000000
IDL> print,a[-2]
       2.0000000
IDL> print,a[n_elements(a)-2]
       2.0000000
IDL> print,b[1,2]
       5.0000000
IDL> print,array_indices(b,5)
           1           2
IDL> print,b[5]
       5.0000000
```

Negative indices are counted from the end (Python+Numpy, R, IDL≥8): -1 is the last element, -2 the one before the last, etc.

Elements in MD arrays can also be accessed through their 1D index (IDL).

Arrays – basic usage

Accessing slices: regular subsets*, 1D or MD, contiguous or not. Ex. (IDL):

```
IDL> b=bindgen(4,5)
IDL> print,b
   0   1   2   3
   4   5   6   7
   8   9  10  11
  12  13  14  15
  16  17  18  19
IDL> c=b[1:2,2:4]
IDL> help,c
C               BYTE      = Array[2, 3]
IDL> print,c
   9  10
  13  14
  17  18
IDL> print,b[*,0:2]
   0   1   2   3
   4   5   6   7
   8   9  10  11
IDL> print,b[1:2,0:-1:2]
   1   2
   9  10
  17  18
IDL> print,b[1,2:0:-1]
   9
   5
   1
```

- `c=b[1:2,2:4]` → Elements from columns **1 to 2**, from lines 2 to 4
- `print,b[*,0:2]` → **All** columns, lines 0 to 2
- `print,b[1:2,0:-1:2]` → Columns **1 to 2**, lines **0 to last (-1)**, **every second line** (stride 2)
- `print,b[1,2:0:-1]` → Stride can be negative, to take elements in reverse order.

*In Numpy there are non-regular slices

Arrays – should I care whether they are row/column major?

For most light, simple use, it does not matter.

When does it matter?

1) **Vector operations**: to select contiguous elements, to use single index for MD arrays.

2) **Mixed language / data sources:**

- When calling a function from another language, accessing files / network connections between different languages.

Arrays – should I care whether they are row/column major?

3) Efficiency:

If an array has to be scanned, it is more efficient (**specially in disk**) to do it in the same order used internally.
Ex: to run through all the elements of this column major array:

a[0,0] (a[0]) : 1	a[1,0] (a[1]) : 2
a[0,1] (a[2]) : 3	a[1,1] (a[3]) : 4
a[0,2] (a[4]) : 5	a[1,2] (a[5]) : 6

In the same order used internally:

```
for j=0,2 do begin
  for i=0,1 do begin
    k=i+j*2
    print,i,j,k,a[i,j]
    do_some_stuff,a[i,j]
  endfor
endfor
```

i	j	**k**	a[i,j]
0	0	**0**	1
1	0	**1**	2
0	1	**2**	3
1	1	**3**	4
0	2	**4**	5
1	2	**5**	6

No going back and forth (shown by variable **k**).

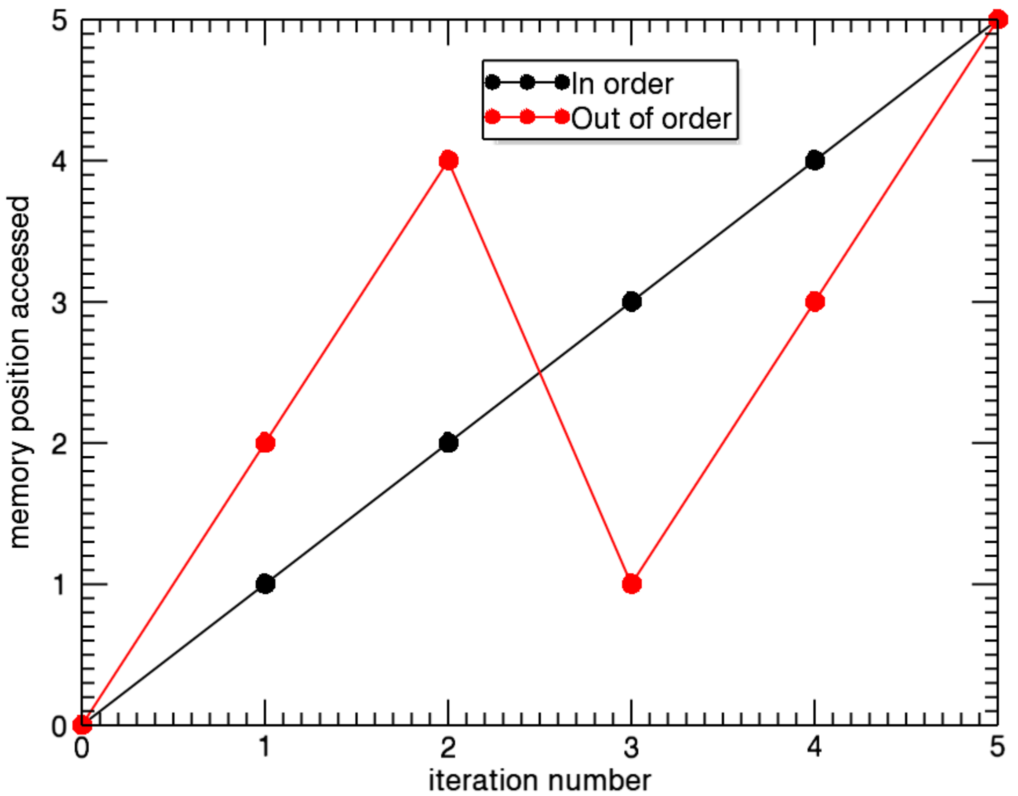
Arrays – should I care whether they are row/column major?

Reading out of order:

```
for i=0,1 do begin
  for j=0,2 do begin
    k=i+j*2
    print,i,j,k,a[i,j]
    do_some_stuff,a[i,j]
  endfor
endfor
```

i	j	**k**	a[i,j]
0	0	**0**	1
0	1	**2**	3
0	2	**4**	5
1	0	**1**	2
1	1	**3**	4
1	2	**5**	6

Lots of going back and forth:

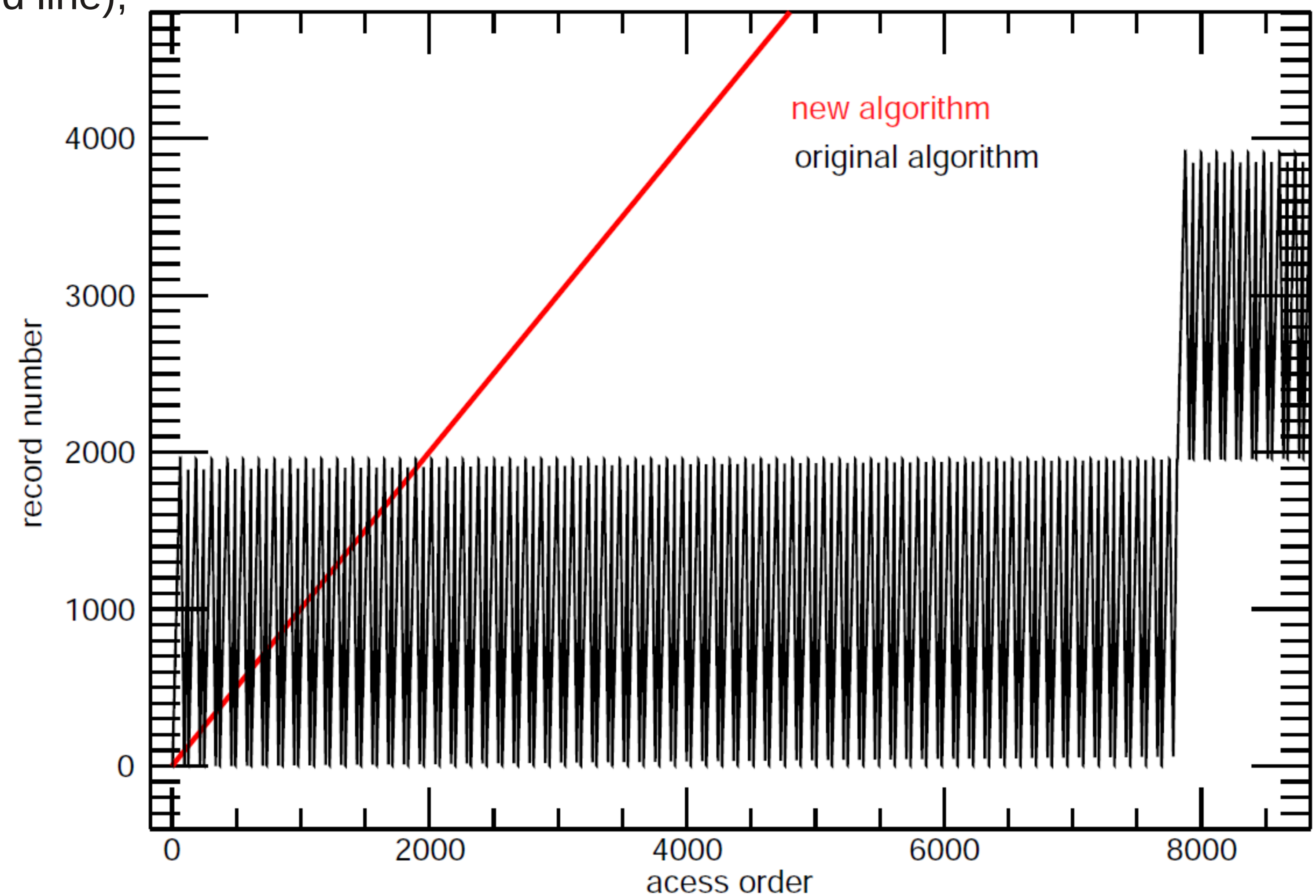


Arrays – should I care whether they are row/column major?

One real life example

The original code read through disk out of order, taking ~1h to run (black line).

When reading in order (red line), the code ran in ~3 min.



Lists - definition

Elements stored **sequentially**, accessed by their indices

- **Similar to 1D arrays.**

Unlike arrays, lists are dynamic, and, in some languages, heterogeneous (IDL, Python, R, Perl)*. Ex. (IDL):

```
IDL> l=list()
IDL> l.add,2
IDL> l.add,[5.9d0,7d0,12d0]
IDL> l.add,['one','two']
IDL> help,l
L               LIST  <ID=1  NELEMENTS=3>
IDL> print,l
       2
       5.9000000       7.0000000       12.000000
one two
IDL> l.remove,1
IDL> print,l
       2
one two
IDL> l.add,bindgen(3),1
IDL> print,l
       2
   0   1   2
one two
```

Creates an empty list

Elements added to the list

Removes element from position **1**. If position unspecified, the last element is removed.

Add element to position **1**. When position is unspecified, added to the end of the list.

Lists - characteristics

Efficient to add / remove elements, from any place in the list.

- Usually elements are added / removed to the end by default.

Most appropriate when

- The number of elements to be stored is not known in advance.
- The types / dimensions of the elements are not known in advance.
- When there will be many adds / removals of elements.

Lists – application examples

Easy storage of “non-regular” arrays.

Applications where each element in the list contains a different number of elements:

- Elements of
  - Asteroid families
  - Star / galaxy clusters
  - Planetary / stellar systems

- Neighbors of objects (from clustering / classification algorithms)
  - Observations / model results
  - Different number of observations for each object
  - Different number of sources found on each observation
  - Different number of objects used in each model

- Non regular grids
  - Model parameters (models are calculated for different values of each parameter)
  - Grids with non-regular spacing
  - Models with different numbers of objects / species

Lists – application examples

Easy storage of “non-regular” arrays.

Applications where each element in the list contains a different number of elements:

- Elements of
  - Asteroid families
  - Star / galaxy clusters
  - Planetary / stellar systems

- Neighbors of objects (from clustering / classification algorithms)
  - Observations / model results
  - Different number of observations for each object
  - Different number of sources found on each observation
  - Different number of objects used in each model

- Non regular grids
  - Model parameters (models are calculated for different values of each parameter)
  - Grids with non-regular spacing
  - Models with different numbers of objects / species

Lists – application examples

Easy storage of "non-regular" arrays. Exs. (IDL):

```
IDL> l=list()
IDL> l.add,[1.0d0,9.1d0,-5.2d0]
IDL> l.add,[2.5d0]
IDL> l.add,[-9.8d0,3d2,54d1,7.8d-3]
IDL> print,l
       1.0000000       9.1000000      -5.2000000
       2.5000000
      -9.8000000       300.00000       540.00000    0.0078000000
IDL> a=l[2]
IDL> print,a
      -9.8000000       300.00000       540.00000    0.0078000000
```

Dictionaries - characteristics

Similar to structures: store **values** by names (**keys**).

Unlike structures, **keys can be any data type** (most often used: strings, integers, reals).

Unlike indices (arrays and lists), **keys are not sequential.**

Unlike structures, dictionaries are dynamic: elements can be freely and efficiently added / removed.

- Dictionaries are to structures as lists are to 1D arrays.

May be heterogeneous – both keys and values can have different types / dimensions.

Elements may not be stored in order:

- The order the keys are listed may not be the same order in which they were put into the dictionary.

Find out whether a key is present, and retrieve the value from a key are operations that take **constant time**: It does not matter (usually) whether the dictionary has 10 or 1 million elements.

Dictionaries - characteristics

Similar to structures: store **values** by names (**keys**).

Unlike structures, **keys can be any data type** (most often used: strings, integers, reals).

Unlike indices (arrays and lists), **keys are not sequential.**

Unlike structures, dictionaries are dynamic: elements can be freely and efficiently added / removed.

- Dictionaries are to structures as lists are to 1D arrays.

May be heterogeneous – both keys and values can have different types / dimensions.

Elements may not be stored in order:

- The order the keys are listed may not be the same order in which they were put into the dictionary.

Find out whether a key is present, and retrieve the value from a key are operations that take **constant time**: It does not matter (usually) whether the dictionary has 10 or 1 million elements.

- Key/value lookup does not involve searches.
- Like a paper dictionary, a paper phone book, or the index in a paper book.

Dictionaries – basic use (ex. IDL):

```
IDL> h=hash()
IDL> h['one']=[9.0,5.8]
IDL> h[18.7]=-45
IDL> h[10]=bindgen(3,2)
IDL> help,h
H               HASH  <ID=1  NELEMENTS=3>
IDL> print,h
10:    0   1   2 ...
one:       9.00000      5.80000
18.7000:      -45
IDL> print,h[10]
   0   1   2
   3   4   5
IDL> print,h.keys()
      10
one
      18.7000
IDL> print,h.values()
   0   1   2   3   4   5
      9.00000      5.80000
     -45
IDL> print,h.haskey('two')
           0
IDL> h.remove,'one'
IDL> print,h.haskey('one')
           0
```

Creates an empty dictionary (hash)

Add values to it

Dictionaries - examples

Storing elements by a useful name, to avoid keep searching for the element of interest. Ex. (IDL): Storing several spectra, by the target name:

```
spectra=hash()
foreach el, files do begin
  read_spectrum,el,spectrum_data
  spectra[spectrum_data.target]=spectrum_data
endforeach
```

Which would be convenient to use:

```
IDL> help,h
H                HASH  <ID=1  NELEMENTS=3>

IDL> print,h
HR21948: { HR21948       5428.1000       5428.1390    5428.1780   5428.2170 ...
HR5438: { HR5438         5428.0000       5428.0390    5428.0780   5428.1170 ...
HD205937: { HD205937     5428.1000       5428.1390    5428.1780   5428.2170 ...

IDL> help,h['HR5438']
** Structure <90013e58>, 7 tags, length=4213008, data length=4213008, refs=6:
   TARGET          STRING    'HR5438'
   WAVELENGTH      DOUBLE    Array[1024]
   FLUX            DOUBLE    Array[1024]
   DATE            STRING    '20100324'
   FILE            STRING    'spm_0049.fits'
   DATA            DOUBLE    Array[512, 1024]
   HEADER          STRING    Array[142]
```

Dictionaries - examples

A lot of freedom in key choice:

- Strings are arbitrary, without the character limitations in structure fields (which cannot have whitespace or special symbols): **`-+*/\()[]{} ,”'`**.

- Special characters commonly appear in useful keys:
  - File names (**`some-file.fits`**)
  - Object names (**`alpha centauri, 433 Eros, 2011 MD`**)
  - Catalog identifier (**`PNG 004.9+04.9`**)
  - Object classification (**`[WC6],R*`**), etc.

- Non-strings are often useful:
  - Doubles – Julian date, wavelength, coordinates, etc.
  - Non consecutive integers, not starting at 0: Julian day, catalog number, index number, etc.

Other containers

Structures are usually implemented as types, but are also containers – **heterogeneous, static and non sequential**:

```
** Structure <9019c628>, 6 tags, length=64, data length=58, refs=2:
   ELEMENT          STRING    'argon'
   INTENSITY        DOUBLE           98.735900
   WIDTH            DOUBLE        0.0087539000
   ENERGY           DOUBLE           12.983800
   IONIZATION       INT              3
   DATABASE         STRING    'NIST Catalog 12C'
   WAVELENGTH       DOUBLE           6398.9548
```

Dictionaries are to structures (both non sequential) as lists are to arrays (both sequential): **the former is the dynamic version of the latter.**

Arrays, lists, structures and dictionaries are the 4 basic containers.

- Most others are specializations of these 4.

Container choice – lists x arrays

Lists and arrays store elements ordered by index. They share many uses.

Differences:

- Lists are dynamic, 1D and may be heterogeneous.
- Arrays are static, homogeneous, and may be more than 1D.

Usually,

- Lists are chosen when one needs:
  - “non regular arrays”
  - add/remove elements (particulalry when the number of elements to store is not known in advance).
  - elements that are not scalar, or not of the same type.
- Arrays are more convenient when one needs:
  - More than 1D
  - vector operations
  - make sure that elements are scalar and of the same type

Container choice – structures x dictionaries

Structures and dictionaries store elements by name. They share many uses.

Main difference:

- Dictionaries are dynamic
- Structures are static

Usually,

- Dictionaries are more convenient when:
  - The keys / types are not known in advance
  - The values may have to change type / dimensions
  - Adding removing fields will be necessary
  - Keys are not just simple strings
- Structures are more convenient:
  - To put them into arrays, to do vector operations
  - To enforce constant type / dimensions of values

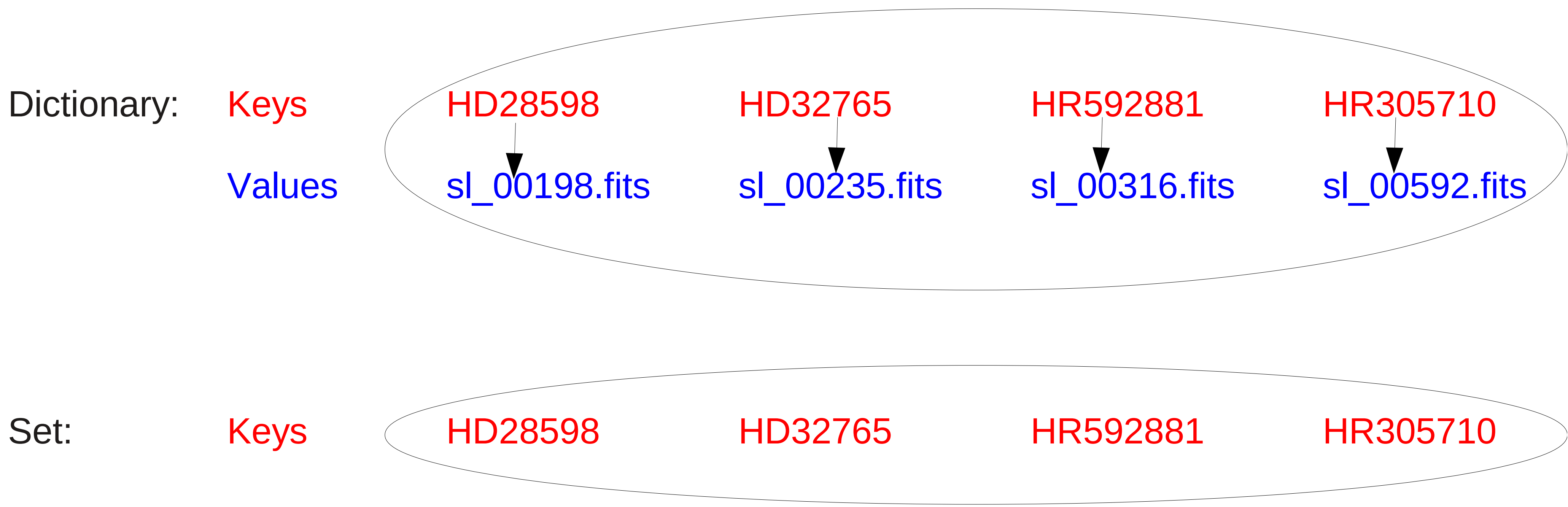
Other containers

***Sets* –** similar to dictionaries, but only store keys, without values. Like sets in mathematics.

- Common uses: sets of elements with no repetition: one can just add elements to the set, without having to check if already present.
  - Exs: Sets of: observed objects, files used, observation dates, etc.

- Important for usefulness of set operations: **union, intersection, difference**.

- Dictionaries may be used as sets, ignoring the values

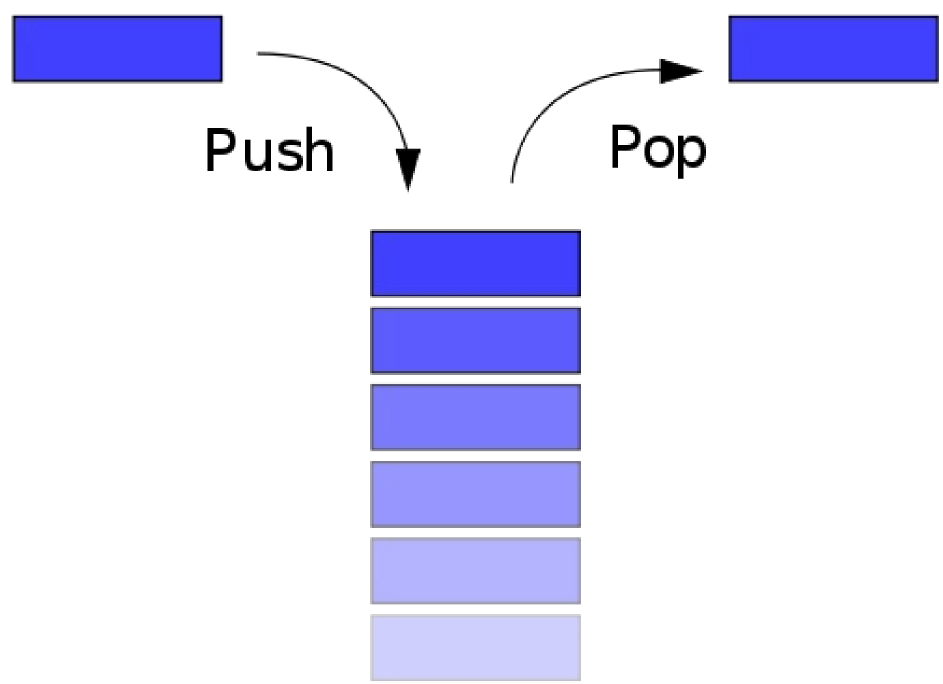


Other containers

Stacks – Lists where elements are only added / removed from the end.

- Like a physical stack: one cannot remove or add a book to the bottom or middle of a stack; only to the top.

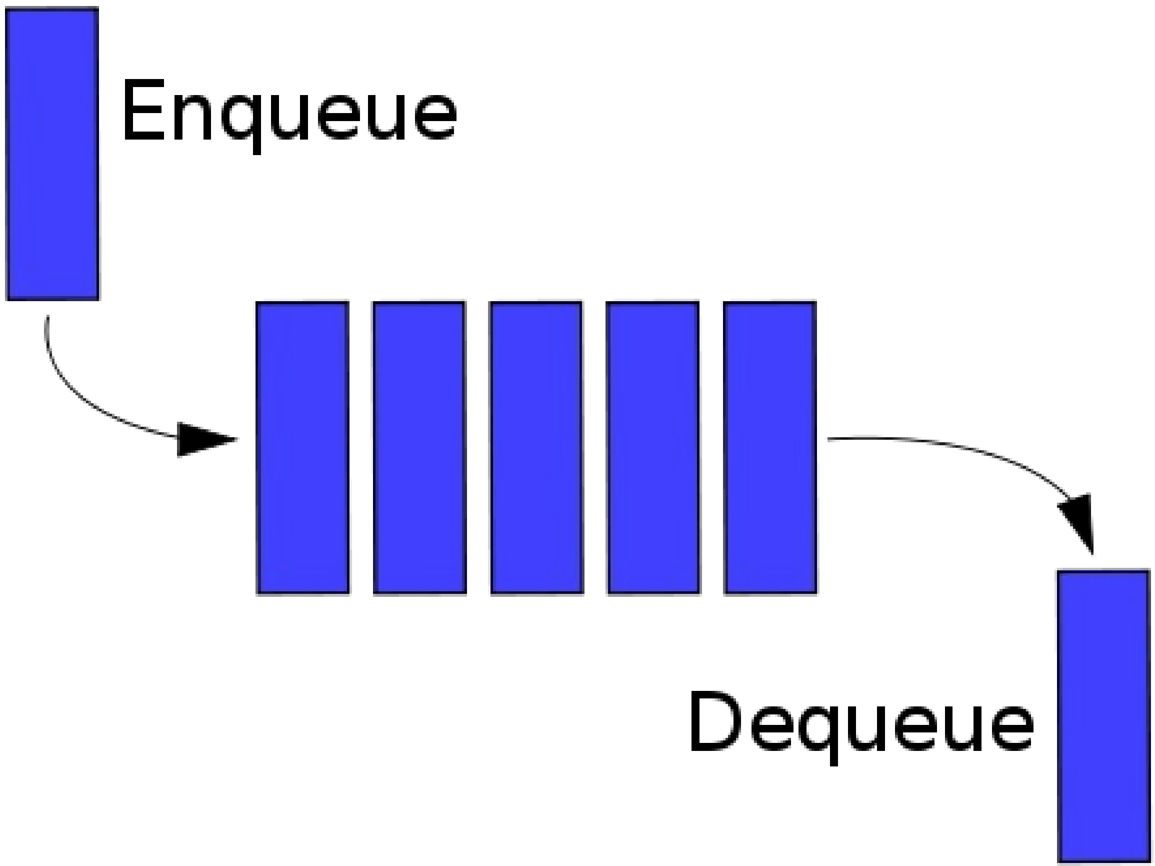
- ***LIFO – Last In, First Out***.



Other containers

Queues – **FIFO (*First In, First Out*) lists**: elements are only added to the end, and only removed from the beginning.

- Like a queue of people waiting to enter some place.



Other containers

Trees / heaps – non sequential containers where access is not by order, nor by name. A hierarchical structures is used:

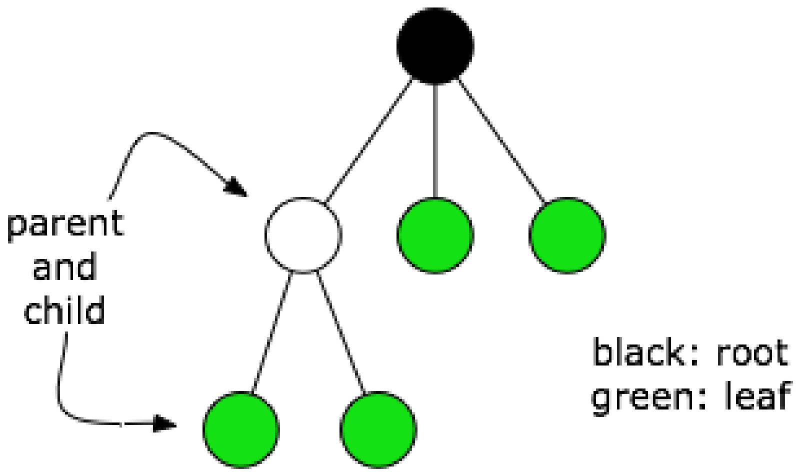
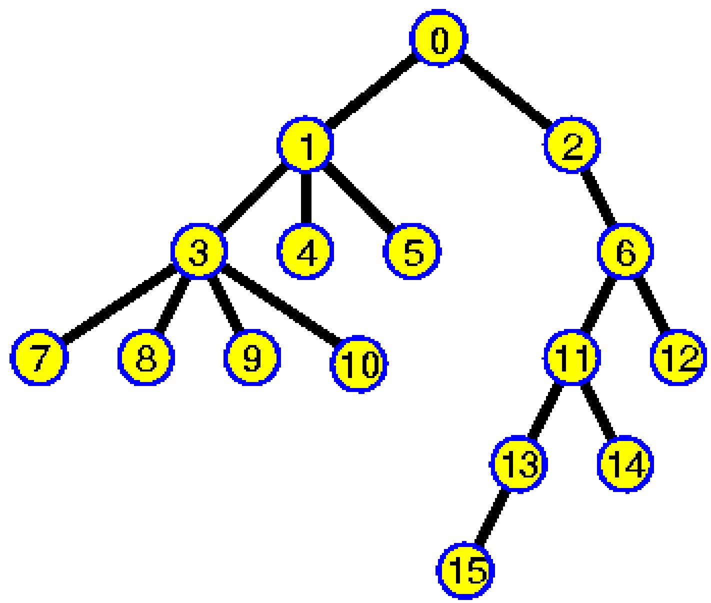


Figure: tree data structure

TREE DEFINITIONS

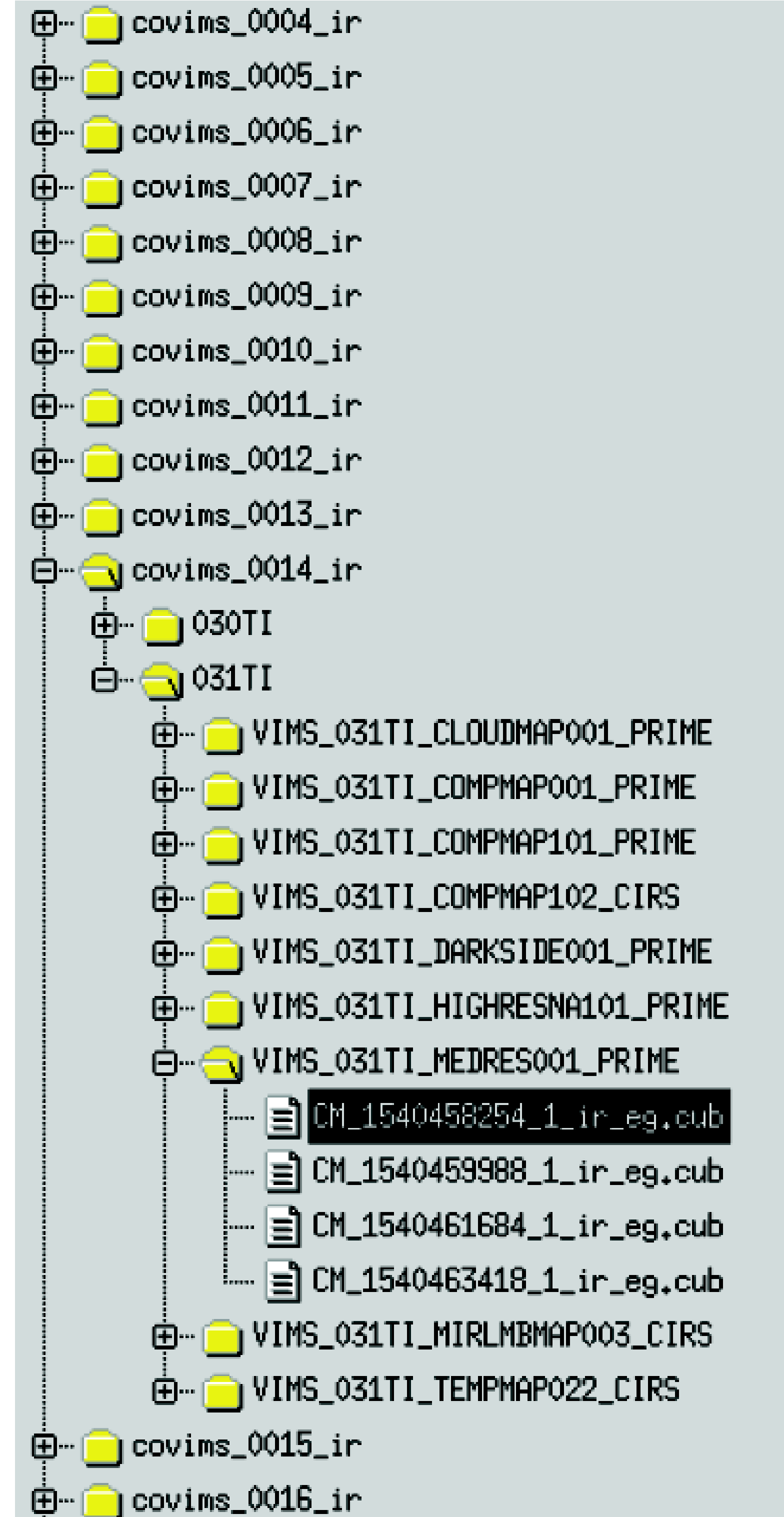


Tree has 16 nodes
Tree has degree 4
Tree has depth 5
Node 0 is the root
Node 1 is internal
Node 4 is a leaf
4 is a child of 1
1 is the parent of 4
0 is grandparent of 4
3, 4 and 5 are siblings

Other containers - trees

Exs:

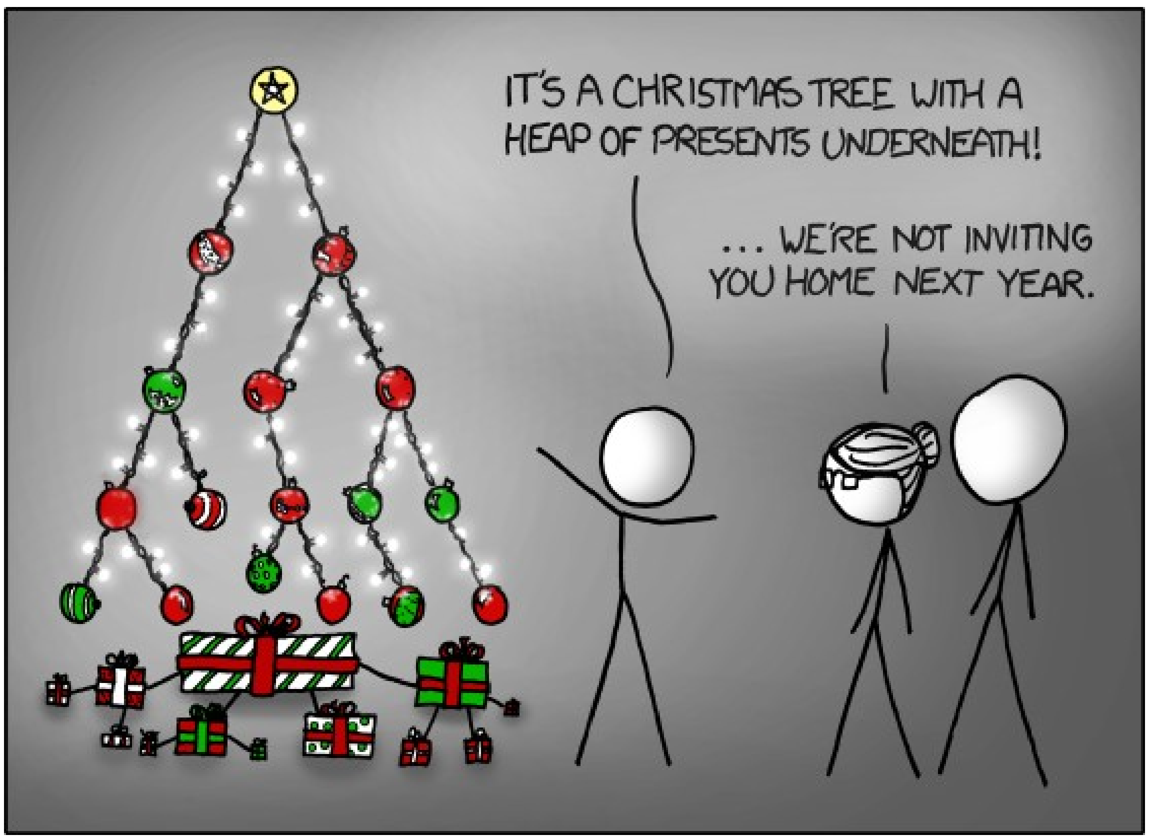
- Directory tree in a disk.
- Hierarchical classifications



Other containers - trees

Exs:

- Directory tree in a disk.
- Hierarchical classifications



(http://www.xkcd.org/835)