

Practical Bioinformatics

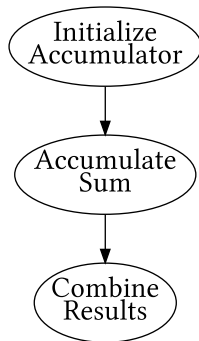
Mark Voorhies

4/25/2017

Mean

```
def mean(x):  
    s = 0.0  
    for i in x:  
        s += i  
    return s/len(x)
```

```
def mean(x):  
    return sum(x)/float(len(x))
```



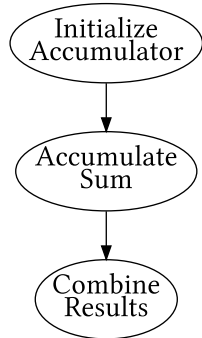
Standard Deviation

$$\sigma_x = \sqrt{\frac{\sum_i^N (x_i - \bar{x})^2}{N - 1}}$$

Standard Deviation

$$\sigma_x = \sqrt{\frac{\sum_i^N (x_i - \bar{x})^2}{N - 1}}$$

```
def stdev(x):  
    m = mean(x)  
    s = 0.0  
    for i in x:  
        s += (i - m)**2  
    return (s/(len(x) - 1))**.5
```



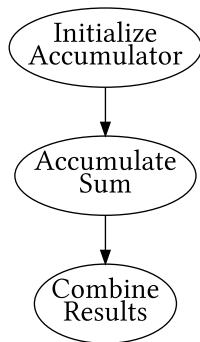
Pearson's Correlation Coefficient

$$r(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

Pearson's Correlation Coefficient

```
def pearson(x, y):  
    mx = mean(x)  
    my = mean(y)  
    sxy = 0.0  
    ssx = 0.0  
    ssy = 0.0  
    for i, j in zip(x, y):  
        dx = i - mx  
        dy = j - my  
        sxy += dx*dy  
        ssx += dx**2  
        ssy += dy**2  
    return sxy / ((ssx*ssy)**.5)
```

$$r(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$



[T]he relational graphic – in its barest form, the scatterplot and its variants – is the greatest of all graphical designs. It links at least two variables, encouraging and even imploring the viewer to assess the possible causal relationship between the plotted variables.

–Edward Tufte

Collections of objects

```
# A list is a mutable sequence of objects
mylist = [1, 3.1415926535, "GATACA", 4, 5]
# Indexing
mylist[0] == 1
mylist[-1] == 5
# Assigning by index
mylist[0] = "ATG"
# Slicing
mylist[1:3] == [3.1415926535, "GATACA"]
mylist[:2] == [1, 3.1415926535]
mylist[3:] == [4, 5]
# Assigning a second name to a list
also_mylist = mylist
# Assigning to a copy of a list
my_other_list = mylist[:]
```


Subject, verb that noun!

`return_value = object.function(parameter, ...)`

“Object, do *function* to *parameter*”

- `file = open("myfile.txt")`
- `file.read()`
- `file.readlines()`
- for line in file:
- `string.split()` and `string.join()`
- `file.write()`

Binary files are like genomic DNA

```
hexdump -C computers.png
```

```
00000000 89 50 4e 47 0d 0a 1a 0a 00 00 00 0d 49 48 44 52 |.PNG.....IHDR|
00000010 00 00 03 5f 00 00 02 cc 08 06 00 00 00 1b c3 08 |.....|..d|
00000020 30 00 00 00 04 73 42 49 54 08 08 08 08 7c 08 64 |...sBIT....|..d|
00000030 88 00 00 00 09 70 48 59 73 00 00 2e 23 00 00 2e |...pHYs...#...|
00000040 23 01 78 a5 3f 76 00 00 00 19 74 45 58 74 53 6f |#.x.?v...tEXtSo|
00000050 66 74 77 61 72 65 00 77 77 77 2e 69 6e 6b 73 63 |ftware.www.inksc|
00000060 61 70 65 2e 6f 72 67 9b ee 3c 1a 00 00 20 00 49 |ape.org.<....I|
00000070 44 41 54 78 9c ec 9d 79 9c 25 57 59 fe bf cf 39 |DATx...y.%WY...9|
00000080 75 6f 2f 33 93 cc 92 c9 1e 48 42 08 01 45 92 a0 |uo/3.....HB..E..|
00000090 04 c2 26 88 08 8a 80 0a b2 28 18 14 54 14 45 04 |..&.....(..T.E.|
000000a0 7f 02 a2 2c b2 aa 2c 0a 28 22 3b ca 26 20 8b b2 |.....(";&..|
000000b0 08 c8 26 9b 61 4d 08 6b 08 d9 c8 be cd 4c 4f 77 |..&aM,k.....LOW|
000000c0 df 7b eb 9c f7 f7 c7 7b aa fb e6 ce bd 3d dd 93 |.{.....{.....=..|
000000d0 59 32 a4 9e fe d4 e7 76 55 9d 53 75 ea d4 a9 aa |Y2.....vU,Su...|
000000e0 77 7d 8e cc 8c 16 2d 5a b4 68 d1 a2 c5 8f 27 24 |w}.....-Z.h.....$|
000000f0 75 81 00 f4 cc cc 24 45 a0 03 d4 66 56 8f 94 ed |u.....$E...fV...|
00000100 00 b1 ac ee b2 7f 42 d9 15 cb ed 2f 48 da 0a 9c |.....B...../H...|
00000110 08 1c 0d 5c 0f 5c 05 9c 6f 66 fd 03 da b0 9b 29 |...\\.\\.of.....)|
00000120 24 4d 03 66 66 bd b2 5e 01 15 30 30 b3 b4 86 e3 |$M,ff.^..00...|
00000130 3c 1c 78 2a f0 25 33 7b f2 1a db b0 01 f8 58 59 |<.x*.%3{.....XY|
00000140 7d a0 99 5d bf 96 fa 2d f6 0c 92 8e 01 9e 08 dc |}.).}.....|
00000150 01 38 0a 10 f0 7b 66 fe 8d 03 d4 9e 67 01 0f 02 |.8.....{f.....g...|
00000160 de 69 66 2f 3f 10 6d d8 9f 68 07 ba 01 2d 5a b4 |.if/?m.....-Z..|
00000170 68 d1 a2 45 8b 7d 8a af 01 0b c0 ed cb fa 6f 97 |h..E.}.....o..|
```

```
fp = open("computers.png")
```

```
fp.read(50)
```

```
fp.close()
```

Text files are like ORFs

hexdump -C 3_4_2010.txt

```
00000000 4d 65 65 74 20 77 2f 20 4a 6f 65 20 72 65 3a 20 |Meet w/ Joe re:|
00000010 77 69 72 65 6c 65 73 73 20 74 68 65 72 6d 6f 73 |wireless thermos|
00000020 74 61 74 73 0a 20 20 20 2d 2d 3e 20 64 6f 6e 65 |tats. --> done|
00000030 0a 20 20 20 20 20 20 42 75 79 20 74 68 65 72 6d |. Buy therm|
00000040 6f 73 74 61 74 73 20 66 72 6f 6d 20 68 74 74 70 |ostats from http|
00000050 3a 2f 2f 77 77 77 2e 6f 6d 65 67 61 2e 63 6f 6d |://www.omega.com|
00000060 0a 20 20 20 20 20 20 20 20 20 20 53 74 61 72 74 |. Start|
00000070 20 77 69 74 68 3a 0a 20 20 20 20 20 20 20 20 20 |with:|
00000080 20 20 20 20 52 6f 75 74 65 72 20 55 57 54 43 52 | Router UWTCR|
00000090 45 43 33 20 28 61 62 6f 75 74 20 24 31 32 30 29 |EC3 (about $120)|
000000a0 0a 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 |. Can receive fro|
000000b0 20 43 61 6e 20 72 65 63 65 69 76 65 20 66 72 6f |m 12 transmitter|
000000c0 6d 20 31 32 20 74 72 61 6e 73 6d 69 74 74 65 72 |s. Can push confi|
000000d0 73 0a 20 20 20 20 20 20 20 20 20 20 20 20 20 20 |guration to tran|
000000e0 20 20 43 61 6e 20 70 75 73 68 20 63 6f 6e 66 69 |smitters. Communi|
000000f0 67 75 72 61 74 69 6f 6e 20 74 6f 20 74 72 61 6e |cate via etherne|
00000100 73 6d 69 74 74 65 72 73 0a 20 20 20 20 20 20 20 |t port and embed|
00000110 20 20 20 20 20 20 20 20 20 43 6f 6d 6d 75 6e 69 |ded web server. A|
00000120 63 61 74 65 20 76 69 61 20 65 74 68 65 72 6e 65 |ssign static IP|
00000130 74 20 70 6f 72 74 20 61 6e 64 20 65 6d 62 65 64 |address and conn|
00000140 64 65 64 20 77 65 62 20 73 65 72 76 65 72 0a 20 |
00000150 20 20 20 20 20 20 20 20 20 20 20 20 20 20 41 |
00000160 73 73 69 6f 6e 20 73 74 61 74 69 63 20 49 50 20 |
00000170 61 64 64 72 65 73 73 20 61 6e 64 20 63 6f 6e 6e |
```

OS X sometimes uses CR newlines

hexdump -C macfile.txt

```
00000000 0d 65 65 74 20 77 2f 20 4a 6f 65 20 72 65 3a 20
00000010 77 69 72 65 6c 65 73 73 20 74 68 65 72 6d 6f 73
00000020 74 61 74 73 0d 20 20 20 2d 2d 3e 20 64 6f 6e 65
00000030 0d 20 20 20 20 20 20 42 75 79 20 74 68 65 72 6d
00000040 6f 73 74 61 74 73 20 66 72 6f 6d 20 68 74 74 70
00000050 3a 2f 2f 77 77 77 2e 6f 6d 65 67 61 2e 63 6f 6d
00000060 0d 20 20 20 20 20 20 20 20 20 20 53 74 61 72 74
00000070 20 77 69 74 68 3a 0d 20 20 20 20 20 20 20 20
00000080 20 20 20 20 52 6f 75 74 65 72 20 55 57 54 43 52
00000090 45 43 33 20 28 61 62 6f 75 74 20 24 31 32 30 29
000000a0 0d 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000000b0 20 43 61 6e 20 72 65 63 65 69 76 65 20 66 72 6f
000000c0 6d 20 31 32 20 74 72 61 6e 73 6d 69 74 74 65 72
000000d0 73 0d 20 20 20 20 20 20 20 20 20 20 20 20 20 20
000000e0 20 20 43 61 6e 20 70 75 73 68 20 63 6f 6e 66 69
000000f0 67 75 72 61 74 69 6f 6e 20 74 6f 20 74 72 61 6e
00000100 73 6d 69 74 74 65 72 73 0d 20 20 20 20 20 20 20
00000110 20 20 20 20 20 20 20 20 20 43 6f 6d 6d 75 6e 69
00000120 63 61 74 65 20 76 69 61 20 65 74 68 65 72 6e 65
00000130 74 20 70 6f 72 74 20 61 6e 64 20 65 6d 62 65 64
00000140 64 65 64 20 77 65 62 20 73 65 72 76 65 72 0d 20
00000150 20 20 20 20 20 20 20 20 20 20 20 20 20 20 41
00000160 73 73 69 67 6e 20 73 74 61 74 69 63 20 49 50 20
00000170 61 64 64 72 65 73 73 20 61 6e 64 20 63 6f 6e 6e
```

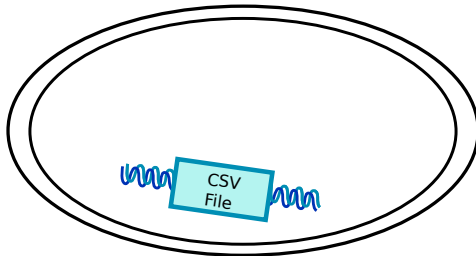
```
Meet w/ Joe re:
wireless thermos
tats. --> done
Buy therm
ostats from http
://www.omega.com
Start
with:
Router UWTCR
EC3 (about $120)
Can receive fro
m 12 transmitter
s.
Can push confi
guration to tran
smitters.
Communi
cate via etherne
t port and embed
ded web server.
A
Assign static IP
address and conn
```

```
tr '\r' '\n' < macfile.txt > unixfile.txt
```

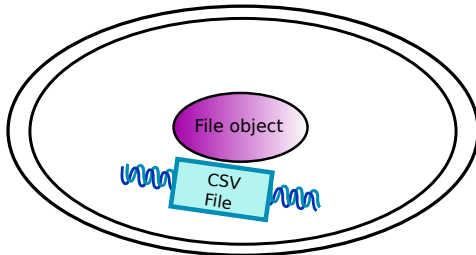
Windows uses CRLF newlines

hexdump -C dosfile.txt

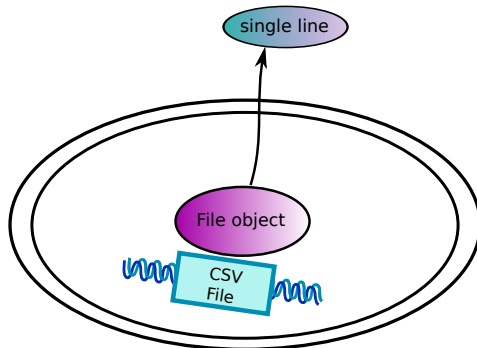
```
00000000 4d 65 65 74 20 77 2f 20 4a 6f 65 20 72 65 3a 20 |Meet w/ Joe re:|
00000010 77 69 72 65 6c 65 73 73 20 74 68 65 72 6d 6f 73 |wireless thermos|
00000020 74 61 74 73 0d 0a 20 20 20 2d 2d 3e 20 64 6f 6e |tats.. --> don|
00000030 65 0d 0a 20 20 20 20 20 20 42 75 79 20 74 68 65 |e.. Buy the|
00000040 72 6d 6f 73 74 61 74 73 20 66 72 6f 6d 20 68 74 |rmstats from ht|
00000050 74 70 3a 2f 2f 77 77 77 2e 6f 6d 65 67 61 2e 63 |tp://www.omega.c|
00000060 6f 6d 0d 0a 20 20 20 20 20 20 20 20 20 20 53 74 |om.. St|
00000070 61 72 74 20 77 69 74 68 3a 0d 0a 20 20 20 20 20 |art with:..|
00000080 20 20 20 20 20 20 20 20 52 6f 75 74 65 72 20 55 | Router U|
00000090 57 54 43 52 45 43 33 20 28 61 62 6f 75 74 20 24 |WTCREC3 (about $|
000000a0 31 32 30 29 0d 0a 20 20 20 20 20 20 20 20 20 20 |120)..|
000000b0 20 20 20 20 20 20 43 61 6e 20 72 65 63 65 69 76 | Can receiv|
000000c0 65 20 66 72 6f 6d 20 31 32 20 74 72 61 6e 73 6d |e from 12 transm|
000000d0 69 74 74 65 72 73 0d 0a 20 20 20 20 20 20 20 20 |itters..|
000000e0 20 20 20 20 20 20 43 61 6e 20 70 75 73 68 | Can push|
000000f0 20 63 6f 6e 66 69 67 75 72 61 74 69 6f 6e 20 74 |configuration t|
00000100 6f 20 74 72 61 6e 73 6d 69 74 74 65 72 73 0d 0a |o transmitters..|
00000110 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 | |
00000120 43 6f 6d 6d 75 6e 69 63 61 74 65 20 76 69 61 20 |Communicate via|
00000130 65 74 68 65 72 6e 65 74 20 70 6f 72 74 20 61 6e |ethernet port an|
00000140 64 20 65 6d 62 65 64 64 65 64 20 77 65 62 20 73 |d embedded web s|
00000150 65 72 76 65 72 0d 0a 20 20 20 20 20 20 20 20 |erver..|
00000160 20 20 20 20 20 20 73 73 69 67 6e 20 20 73 74 | Assign st|
00000170 61 74 69 63 20 49 50 20 61 64 64 72 65 73 73 20 |atic IP address|
```



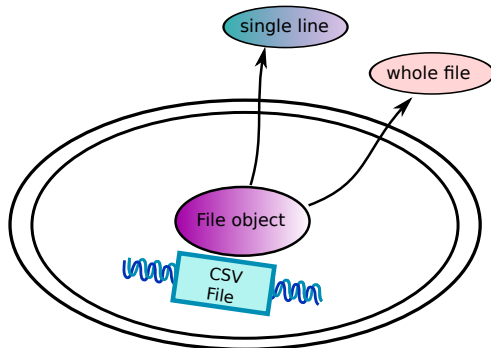
```
open("supp2data.csv")
```



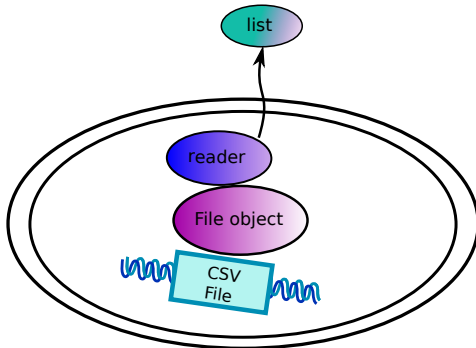
```
open("supp2data.csv").next()
```



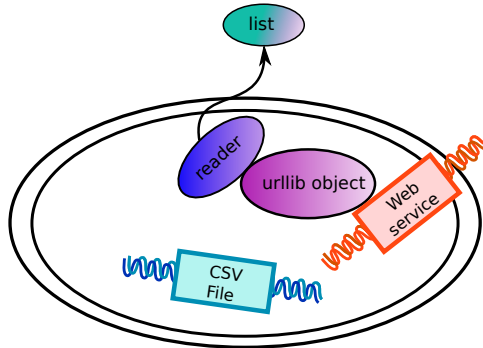

```
open("supp2data.csv").read()
```



```
csv.reader(open("supp2data.csv")).next()
```

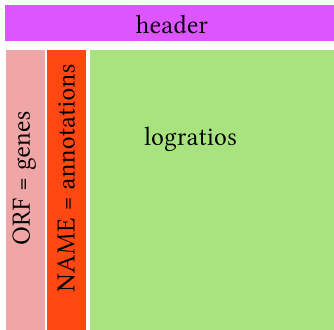


```
csv.reader(urlopen("http://example.com/csv")).next()
```

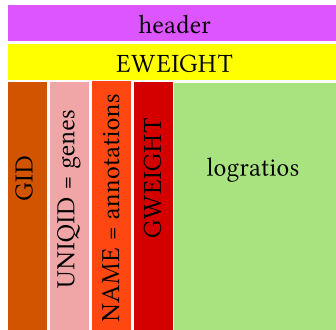


The CDT file format

Minimal CLUSTER input



Cluster3 CDT output



- Tab delimited (`\t`)
- UNIX newlines (`\n`)
- Missing values \rightarrow empty cells

Homework

- 1 Download and install JavaTreeView
- 2 Try reading the first few bytes of different files on your computer. Can you distinguish binary files from text files?
- 3 Create a simple data table in your favorite spreadsheet program and save it in a text format (e.g., save as CSV or tab-delimited text from Excel¹). Practice reading the data from Python.
- 4 Write a function to dissect `supp2data.cdt` into three lists of strings (gene names, gene annotations, and experimental conditions) and one matrix (list of lists) of log ratio values (as floats, using *None* or *0.* to represent missing values).
- 5 If you are familiar with Python classes, write a CDT class based on the parse in the previous exercise. Provide methods for looking up annotations and log ratios by gene name.

¹Note for Mac users: Excel will offer you Macintosh and DOS/Windows text formats. Choose *DOS/Windows*; otherwise, Python will think that the entire file is a single line.