Data Science and Basic Algorithms

Bioinformatics for Systems and Synthetic Biology

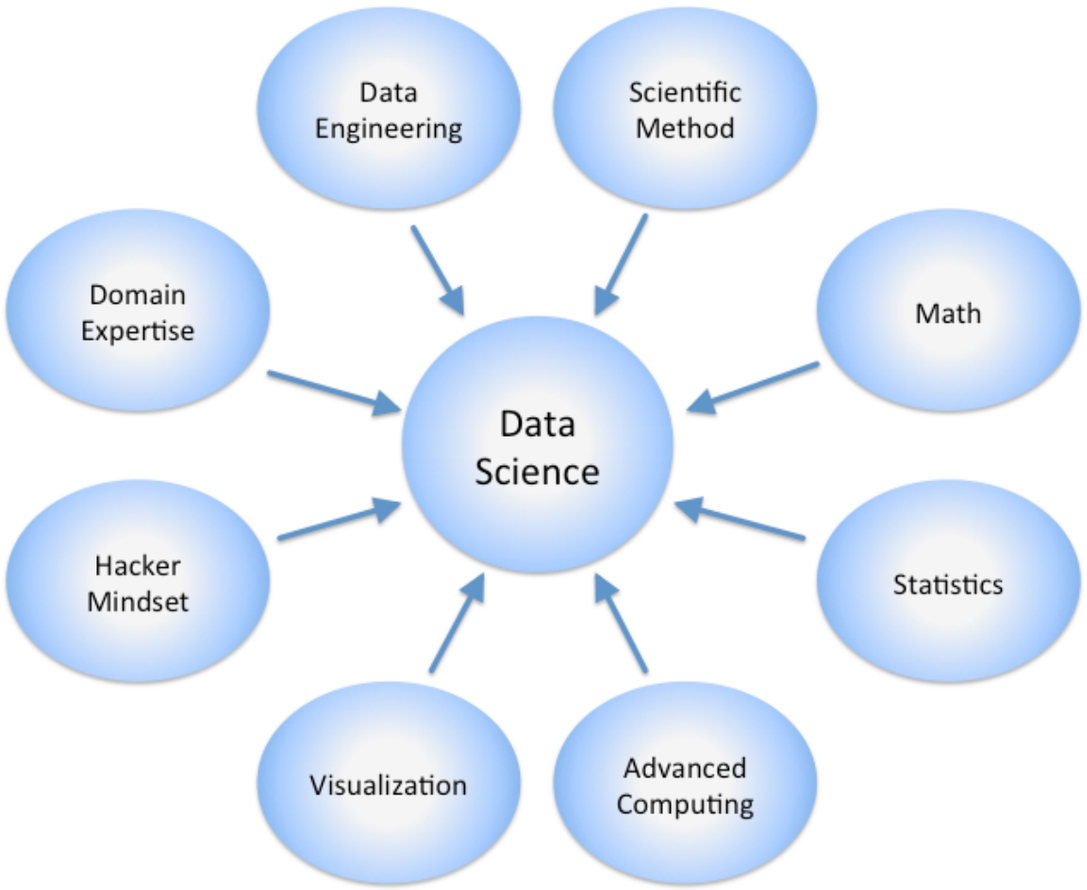
Emidio Capriotti
http://biofold.org/

Biomolecules Folding and Disease

Department of Pharmacy and Biotechnology (FaBiT)
University of Bologna

What is Data Science?

Data Science is an interdisciplinary field about processes and systems to extract knowledge or insights from large volumes of data in various forms, either structured or unstructured.

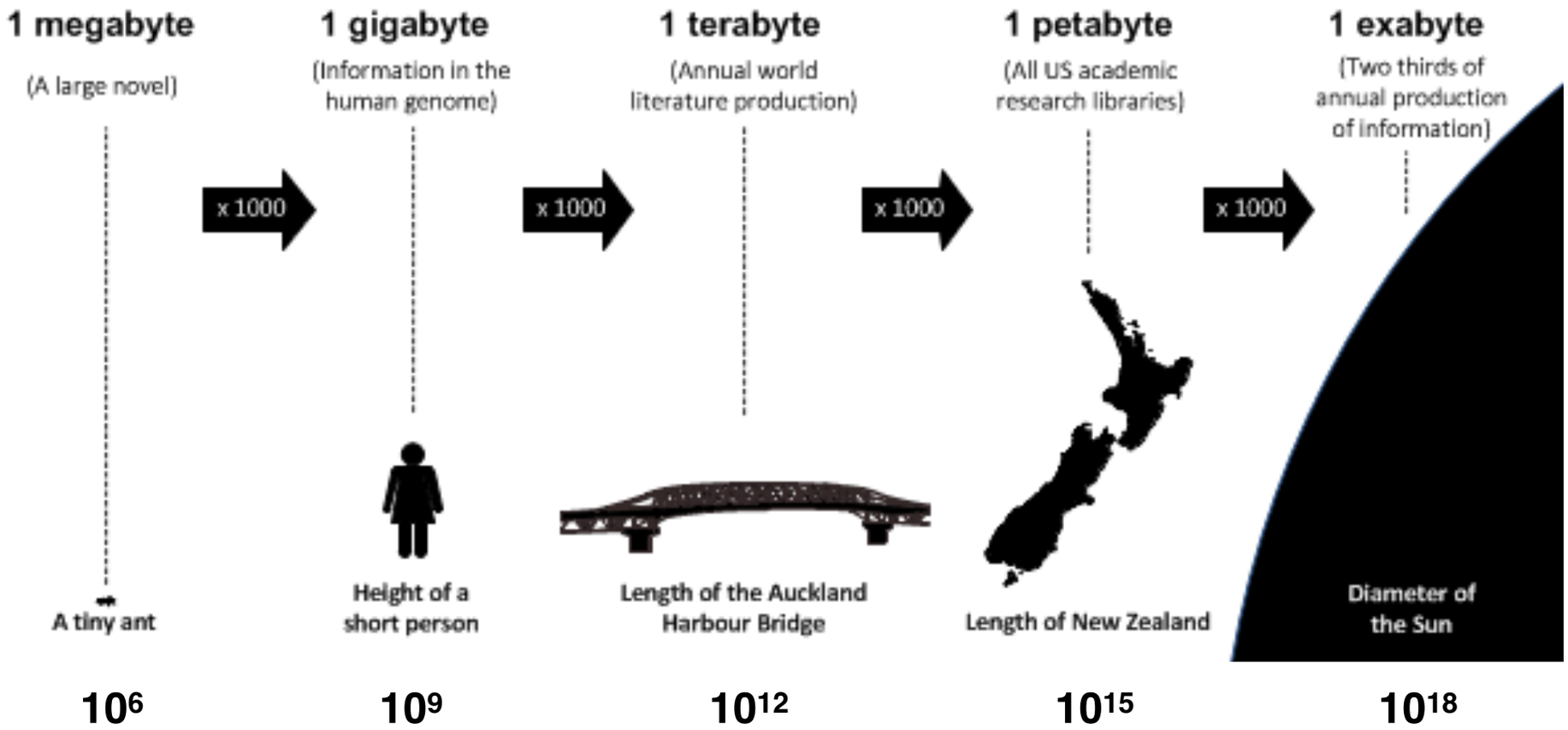


The need for data scientists emerged in response to the “data deluge” – the increasingly large amounts of data generated each year – and the realization that some of this data is uniquely valuable.

http://dreamhost.com

Data Deluge

Comparing data size with physical objects to visualize the magnitude of data growth.

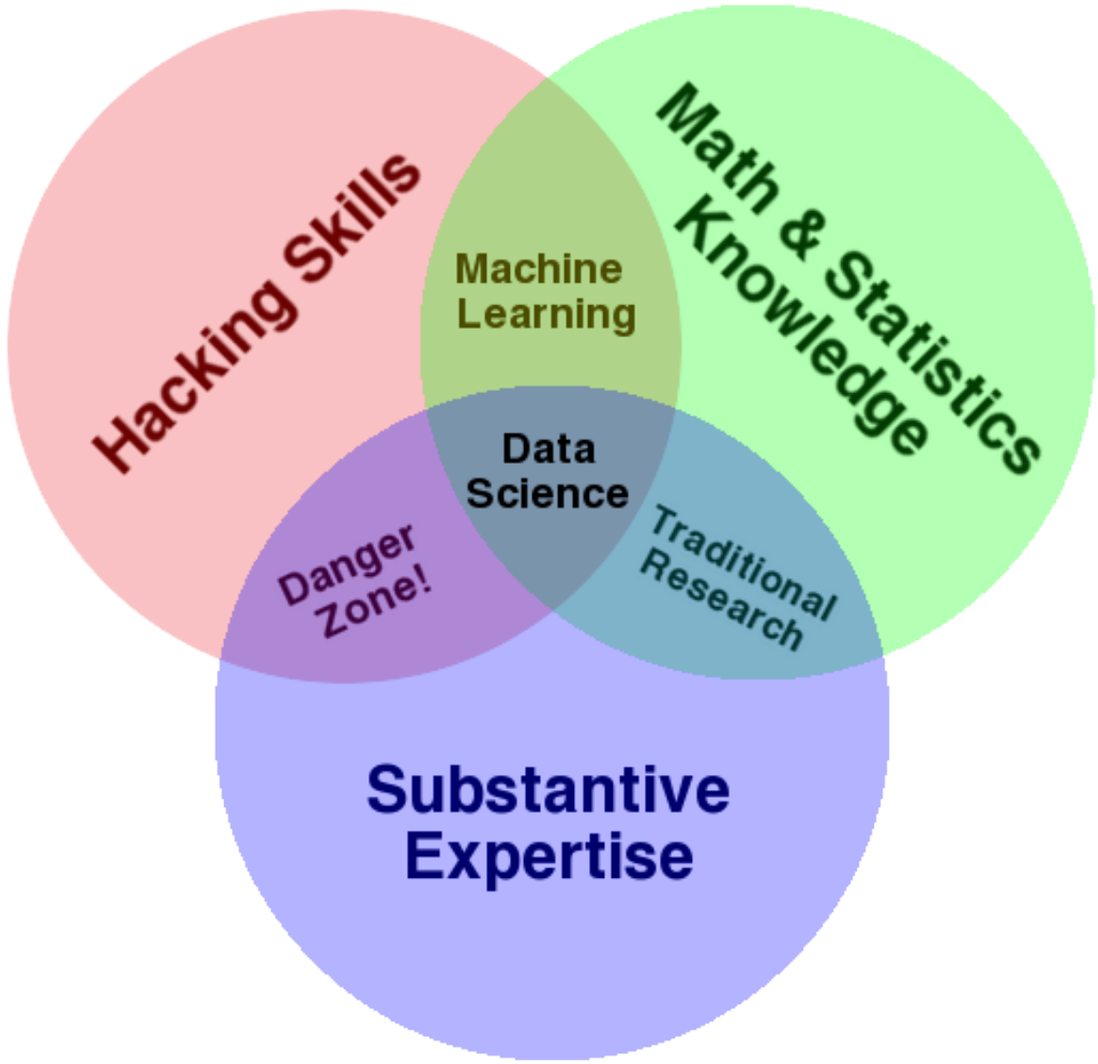


Length of a tiny ant	1.4 mm
Height of a short person	1.4 m
Length of the Auckland Harbor Bridge	1,020 m
Length of New Zealand	1,600 km
Diameter of the Sun	1,390,000 km

http://seradigm.co.nz

Data Science Venn Diagram

The primary skills in data science are hacking, math and stats knowledge, and substantive expertise.



http://http://drewconway.com/

Required Skills

Data scientists use their data and analytical ability to:

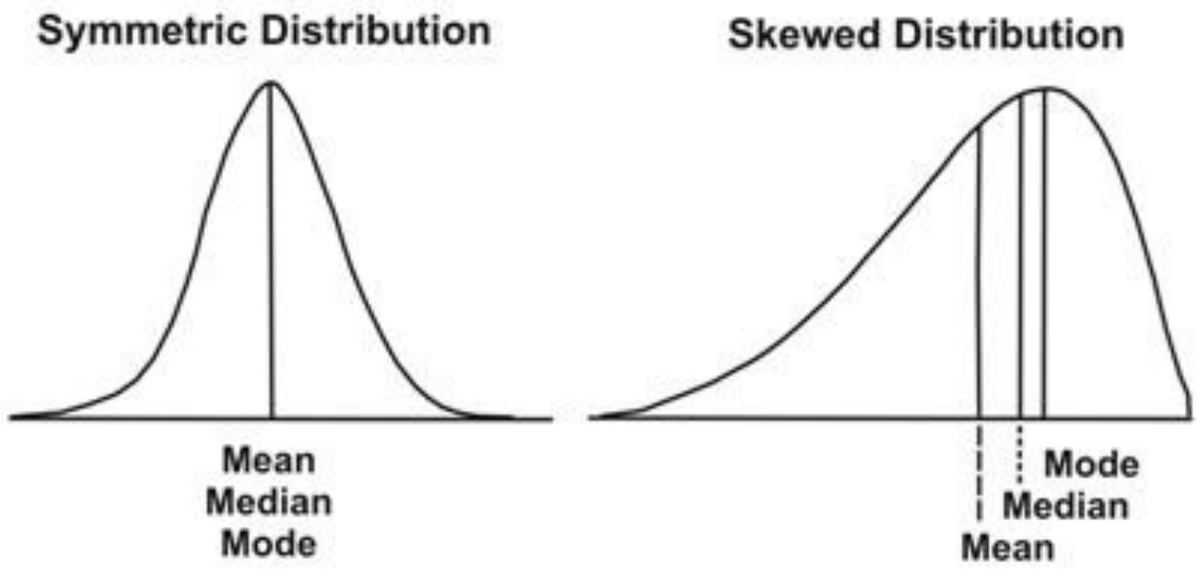
- find and interpret rich data sources;
- manage large amounts of data;
- merge data sources;
- ensure consistency of datasets;
- build mathematical models using the data;
- visualize and communicate the data insights/findings.

Data Science can be used in different domains. Some example are:

- Public Health
- Consumer target models

Basics in Statistics

Given a set of values $A = [x_1, x_2, x_3, \ldots x_N]$ that defines a probability distribution p we can calculate the following measures



Mean: $$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Standard Deviation: $$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Median: $$m \mid \ \mathrm{P}(X \leq m) \geq \frac{1}{2} \text{ and } \mathrm{P}(X \geq m) \geq \frac{1}{2}$$

where P is the probability.

Statistics with Python

Scipy is the most popular library of scientific tools in python that implements many statistical functions. Basic statistical functions are also available in numpy.

```
>>> import scipy
>>> A=[2,2,3,4,5,10,7]
>>> scipy.mean(A)
4.714285714285714
>>> scipy.std(A)
2.9277002188455996
>>> scipy.median(A)
4.0
```

Using the module stats, scipy can be used to generate random normal distributions with given mean and standard deviation:

```
>>> import scipy.stats as stats
>>> dist=stats.norm(10,1)
>>> stats.rvs(size=10)
array([  9.63517267,   8.88827343,   9.01236771,   9.40981448,
        10.25279873,  11.09357001,  10.20825657,   8.52022001,
         9.99897057,   9.07548534])
```

Some exercises

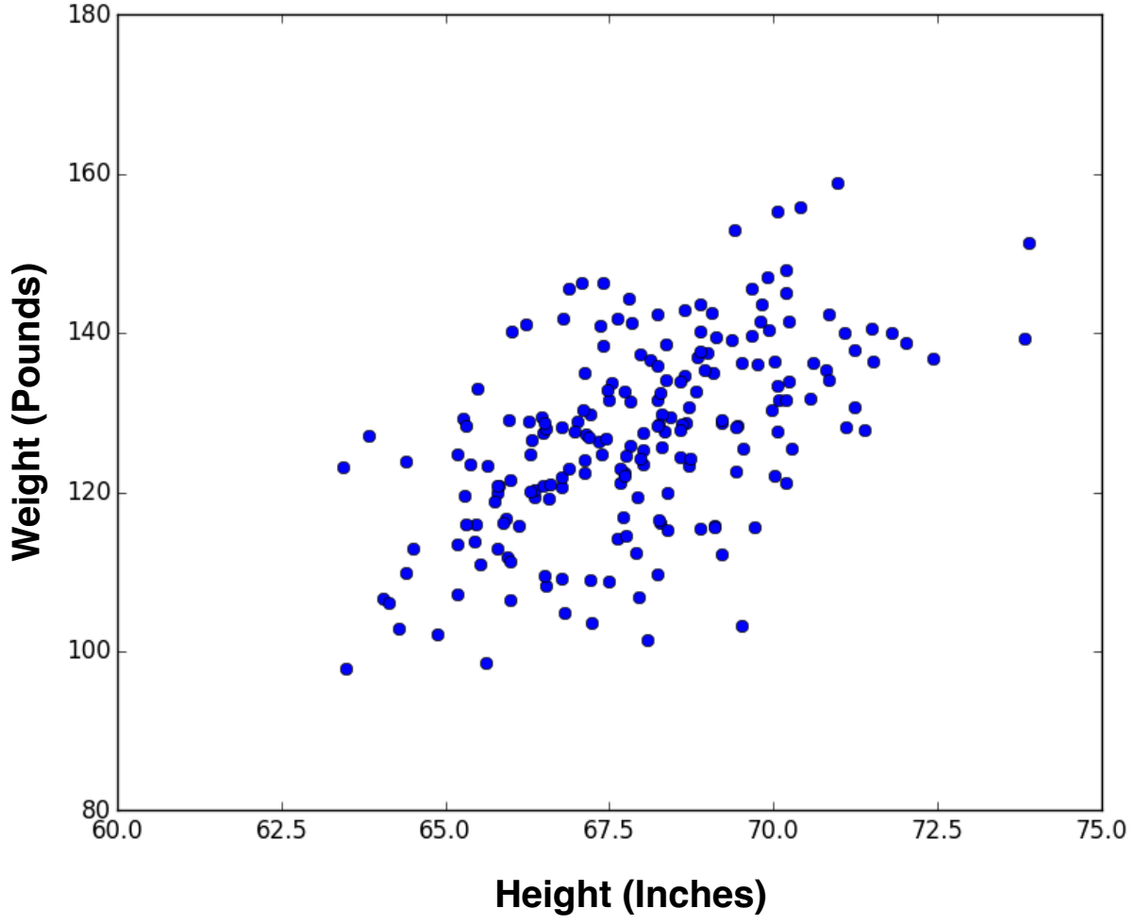
1. Write a python script that takes in input a fasta file of only one protein sequence and calculate the frequency of each amino acid. Check the output of for the files:
   http://www.uniprot.org/uniprot/P53_HUMAN.fasta
   http://www.uniprot.org/uniprot/BRCA1_HUMAN.fasta

2. Write a code that analyze a fasta file with multiple sequences and calculate the length of each sequence. Save the results in a file.

A basic predictive model

The linear regression model is one of the simplest predictive model that is used in different fields.

For example: it is known that the weight of a person is correlated with its height. Can we build a model to predict the weight of a new person for which the know the height?



Fitting a linear regression

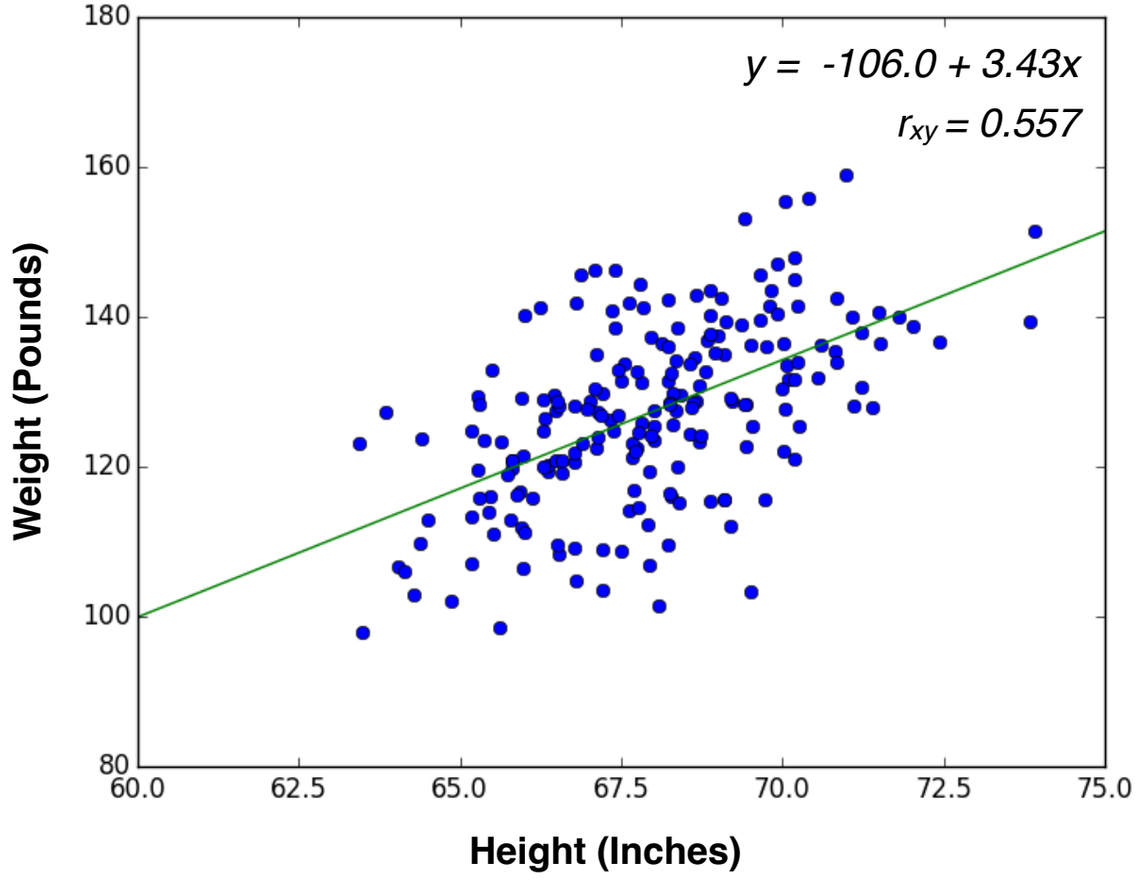
Given a list of points A=[(x_1,y_1), (x_2,y_2) ….. (x_n,y_n)], we calculate the parameters $\tilde{\alpha}$ *and* $\tilde{\beta}$ *of the* line $y = \alpha + \beta x$ that minimizes

$$Q(\alpha, \beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

$$\tilde{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\tilde{\alpha} = \bar{y} - \tilde{\beta}\,\bar{x},$$

$$r_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}$$

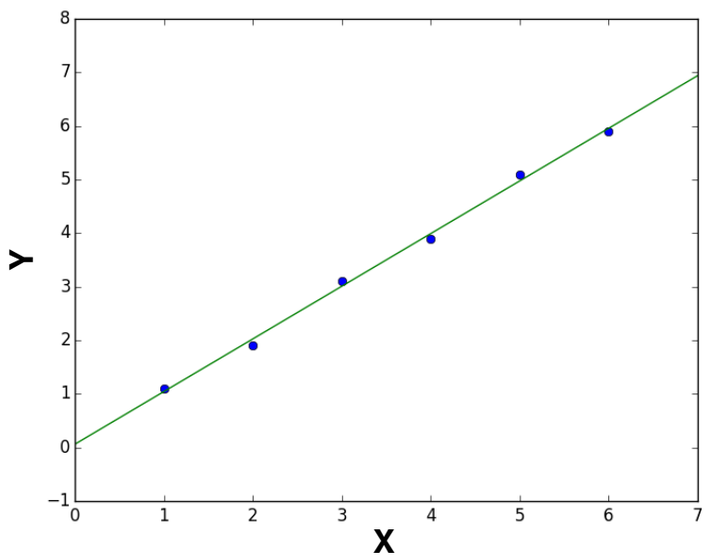


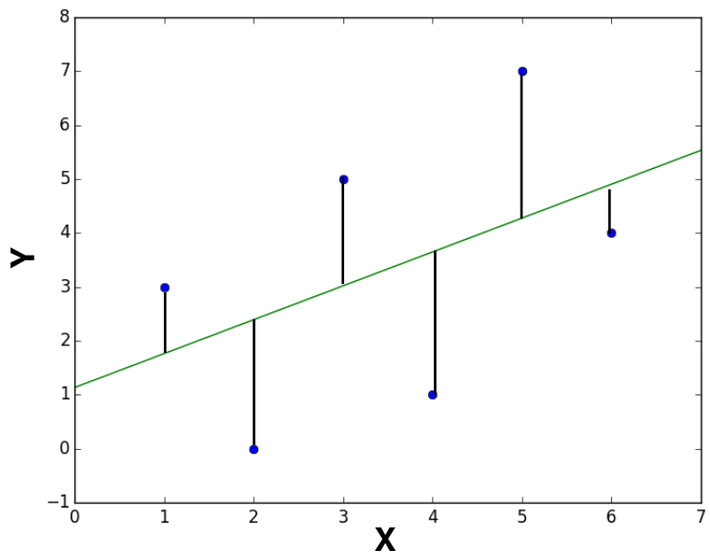
The linregress function

Import linregress from spicy.stats and calculate calculate the fitting curve

```
>>> from scipy.stats import linregress
>>> import numpy as np
>>> x = np.array([1,2,3,4,5,6])
>>> y = np.array([1.1,2.2,2.9,3.98,5.2,6.1])
>>> reg=linregress(x,y)
>>> print (reg)
LinregressResult(slope=0.9828571428571287,
intercept=0.06000000000000053, rvalue=0.9983814394570299,
pvalue=3.927487244422332e-06, stderr=0.02799416848895046)
```

what happen if y = [3,0,5,1,7,4]. is this a better fitting? why?





Use matplotlib to plot

Import matplotlib and plot the points and fitting curve.

```
>>> import matplotlib.pyplot as plt
>>> yp=reg[0]*x+reg[1]
>>> plt.plot(x,y,'o',x,yp,'-')
>>> plt.show()
```

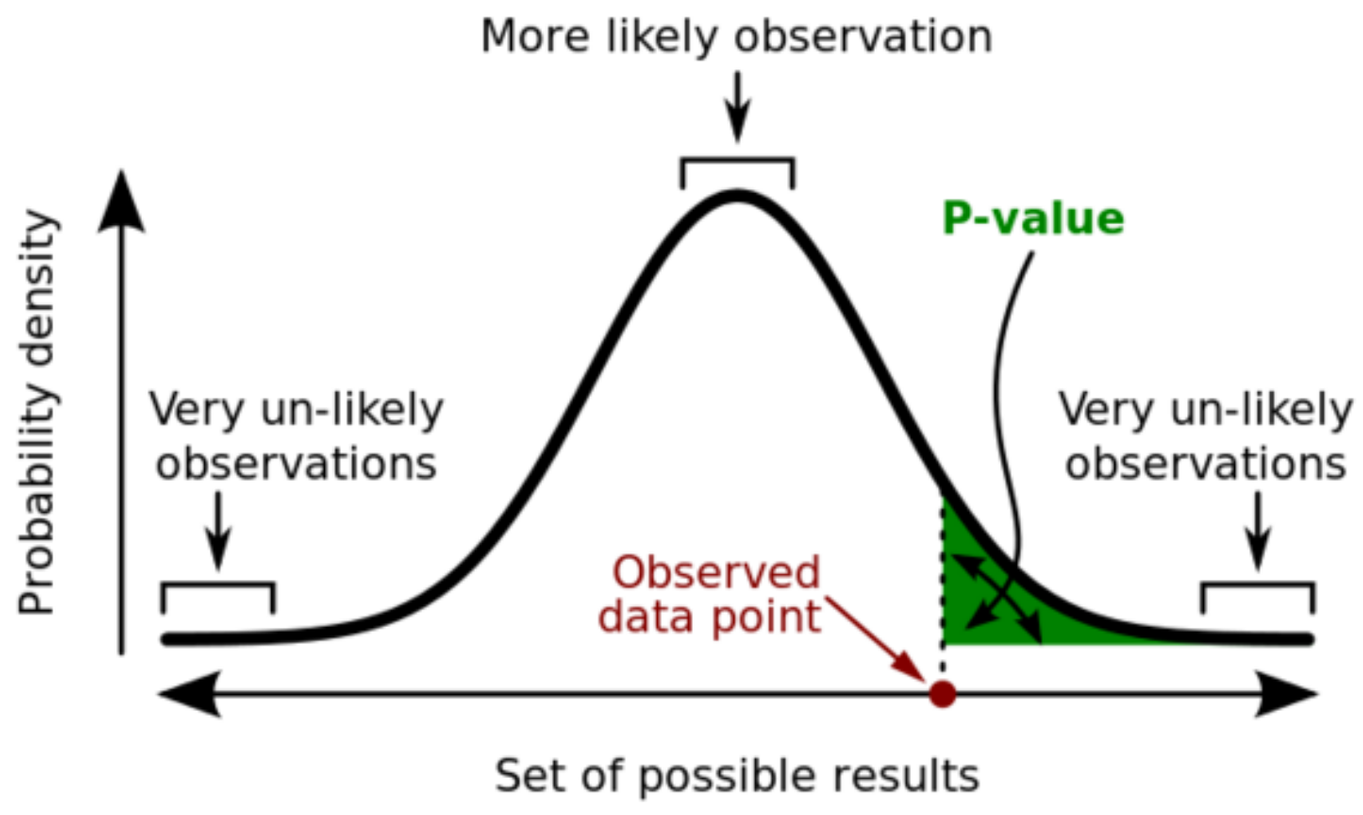
Exercise:

Write a python script that reads a file containing two columns of data (x,y) and calculate the linear regression curve and plots both the points an the regression curve.

For this exercise download the data using the command *wget* from http://biofold.org/pages/courses/docs/data_hw.txt

The p-value

In statistics, the p-value is a function of the observed results that is used for testing a statistical hypothesis. More specifically, the p-value is defined as the probability of obtaining a result equal to or "more extreme" than what was actually observed.

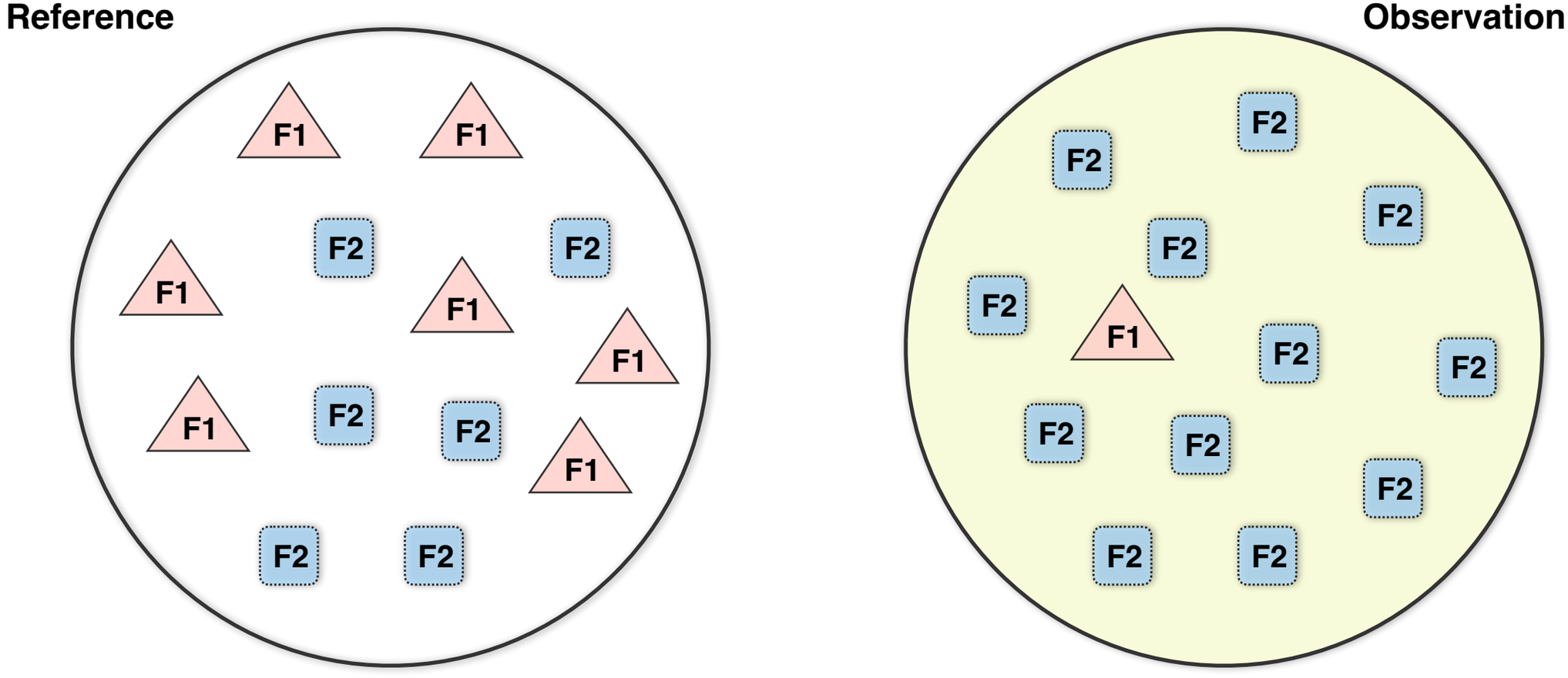


https://en.wikipedia.org/

The enrichment analysis

It is a method to identify classes of genes or proteins or functions that are over-represented in a large set of genes or proteins.

The gene set enrichment analysis is used to understand the functional profile of a set of genes.



Contingency table

In statistics, a contingency table is a type of table in a matrix format that displays the frequency distribution of the variables.

They are heavily used in survey research, business intelligence, engineering and scientific research.

	Function 1	Function 2	Row Total
Reference	7	6	13
Observation	1	12	13
Column Total	8	18	26

Fisher's exact test

Fisher's exact test is a statistical significance test used in the analysis of contingency tables.

It is one of a class of exact tests, so called because the significance of the deviation from a null hypothesis (p-value) can be calculated exactly.

For a generalized contingency tables:

	Function 1	Function 2	Row Total
Reference	a	b	a+b
Observation	c	d	c+d
Column Total	a+c	b+d	a+b+c+d=n

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

Fisher's test in python

The fisher_exact function is contained in the spicy.stats module and takes in input a contingency matrix. The function returns *odd ratio* and *p-value.*

```
>>> from scipy.stats import fisher_exact
>>> import numpy as np
>>> cm=np.array([[6,7],[12,1]])
>>> ft=fisher_exact(cm)
>>> print (ft)
(0.07142857142857425, 0.03020594965675050 1)
```

Exercise:

Write a python script that reads two file containing a columns with alleles carried by each individual. Use Fisher's exact test verify if an allele is over represented in one of the populations.

For this exercise download the data using the command *wget* from
http://biofold.org/pages/courses/docs/pop1_allele.txt
http://biofold.org/pages/courses/docs/pop2_allele.txt