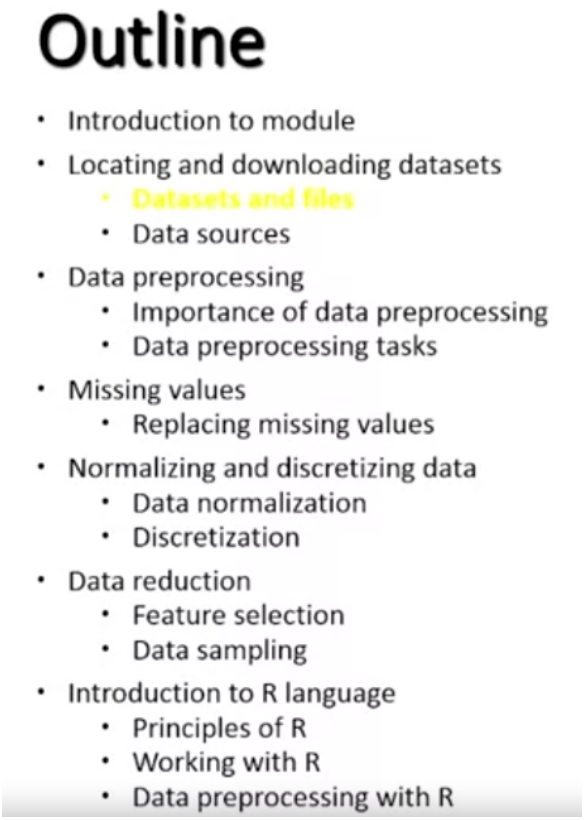
BIG DATA AND MEDICINE

Objetivos de aprendizagem

- Demonstrate how to locate and download files for data analysis involving genes and medicine.
- Select datasets, open files and preprocess data using R language.
- Develop and write R scripts to replace missing values, normalize data, discretize data, and sample data.



Datasources

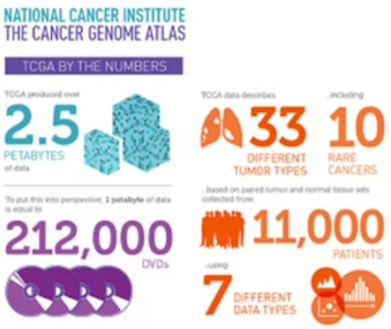
- Where do we find biomedical Big Data sources ?
  - Data in existing repositories (proprietary):
    - Electronic Medical Records (EMRs).
    - Clinical studies.
  - Open data sources (public).
    - List in resources
    - The Cancer Genome Atlas (TCGA)
    - Alzheimer's Disease Neuroimaging Initiative (ADNI)
    - Health and Retirement Study (HRS)
    - UK Biobank
    - Millennium Cohort Study
    - CALIBER (EHR and admin data)
    - UCI Machine Learning Repository

Machine learning

https://archive.ics.uci.edu/ml/datasets/Heart+Disease

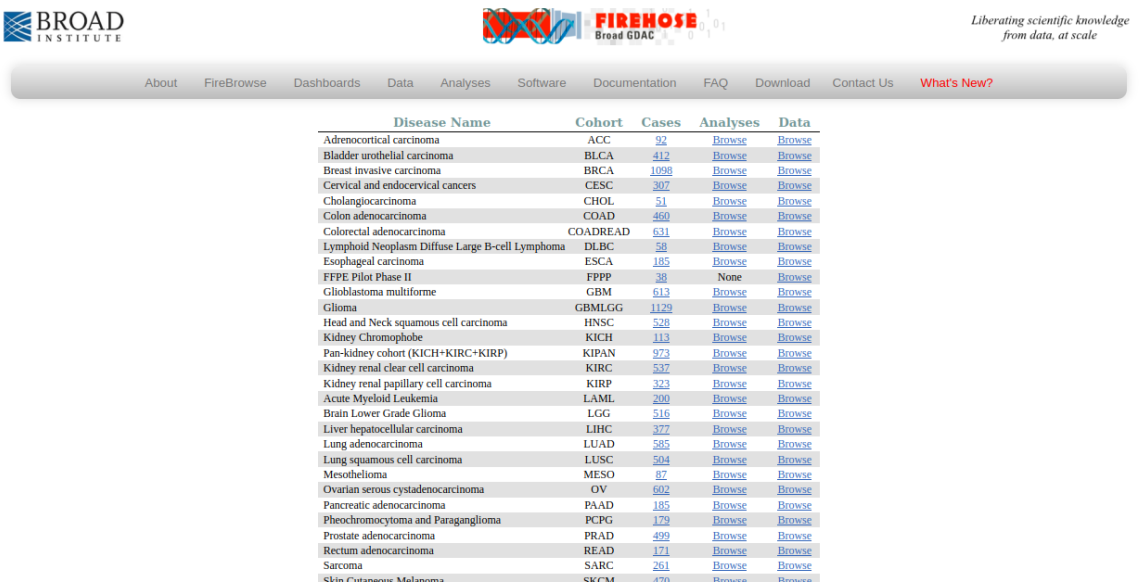
TGCA

- TCGA data
  - The cancer Genome Atlas (TCGA) provides multi-dimensional data about 33 types of cancers.
  - It provides 2.5 petabytes of data from over 11,000 patients including both tumor tissue and matched normal tissue.
  - For each type of cancer, several datasets are available with 7 different types of data – many being public, others requiring authorization to be accessed.



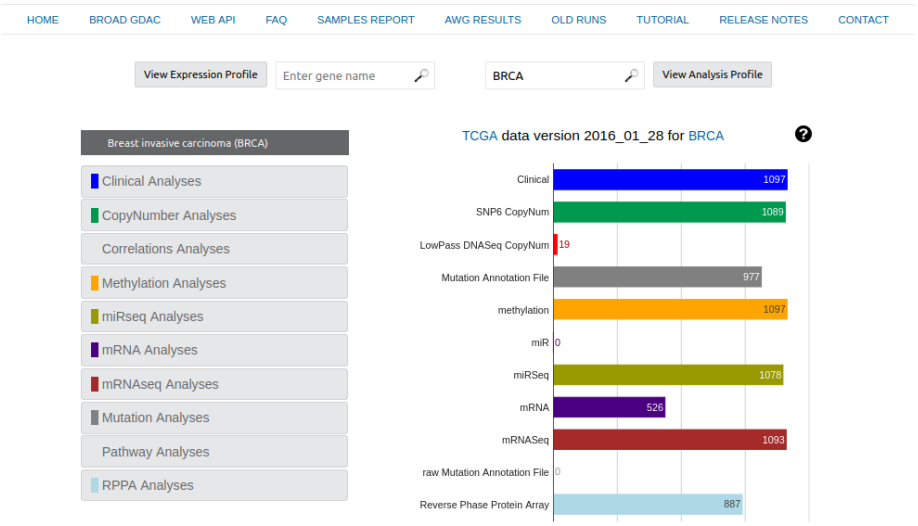
Broad Institute

- Broad Institute
  - TCGA provides data at the patient level. Each patient in a separate file.
  - The Broad Institute from Harvard, a famous bioinformatics research center, provides merged datasets from TCGA data in its Firehose Web-site (http://gdac.broadinstitute.org/).



Disease Name	Cohort	Cases	Analyses	Data
Adrenocortical carcinoma	ACC	92	Browse	Browse
Bladder urothelial carcinoma	BLCA	412	Browse	Browse
Breast invasive carcinoma	BRCA	1098	Browse	Browse
Cervical and endocervical cancers	CESC	307	Browse	Browse
Cholangiocarcinoma	CHOL	51	Browse	Browse
Colon adenocarcinoma	COAD	460	Browse	Browse
Colorectal adenocarcinoma	COADREAD	631	Browse	Browse
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	58	Browse	Browse
Esophageal carcinoma	ESCA	185	Browse	Browse
FFPE Pilot Phase II	FPPP	38	None	Browse
Glioblastoma multiforme	GBM	613	Browse	Browse
Glioma	GBMLGG	1129	Browse	Browse
Head and Neck squamous cell carcinoma	HNSC	528	Browse	Browse
Kidney Chromophobe	KICH	113	Browse	Browse
Pan-kidney cohort (KICH+KIRC+KIRP)	KIPAN	973	Browse	Browse
Kidney renal clear cell carcinoma	KIRC	537	Browse	Browse
Kidney renal papillary cell carcinoma	KIRP	323	Browse	Browse
Acute Myeloid Leukemia	LAML	200	Browse	Browse
Brain Lower Grade Glioma	LGG	516	Browse	Browse
Liver hepatocellular carcinoma	LIHC	377	Browse	Browse
Lung adenocarcinoma	LUAD	585	Browse	Browse
Lung squamous cell carcinoma	LUSC	504	Browse	Browse
Mesothelioma	MESO	87	Browse	Browse
Ovarian serous cystadenocarcinoma	OV	602	Browse	Browse
Pancreatic adenocarcinoma	PAAD	185	Browse	Browse
Pheochromocytoma and Paraganglioma	PCPG	179	Browse	Browse
Prostate adenocarcinoma	PRAD	499	Browse	Browse
Rectum adenocarcinoma	READ	171	Browse	Browse
Sarcoma	SARC	261	Browse	Browse
Skin Cutaneous Melanoma	SKCM	470	Browse	Browse

http://firebrowse.org/?cohort=BRCA



Microarray

https://learn.genetics.utah.edu/content/labs/microarray/

- Microarray
  - A microarray contains a set of measurements of gene expressions (mRNA – other types of RNA are said to be non coding).
  - Nucleus contains between 2,000 and 60,000 protein-coding genes, only a subset being expressed at a certain point in time in the form of mRNA.
  - A microarray is a solid support on which DNA of known sequence is deposited in a regular grid-like manner (cDNA).

RNASeq

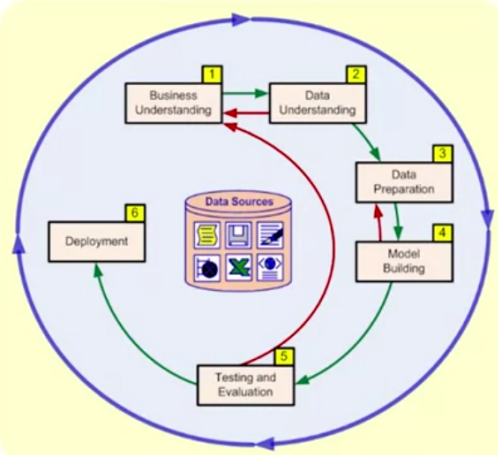
Data Sources

- RNA seq
  - A more advanced method for quantifying gene expressions.
  - Part of Next Generation Sequencing (NGS) methods – microarray being the first generation.
  - Methods based on synthesis chemistry of individual nucleotides performed in massively parallel form.
  - Generates vast amounts of data.
  - Affords for much higher sensitivity than previous methods.

Data Preprocessing

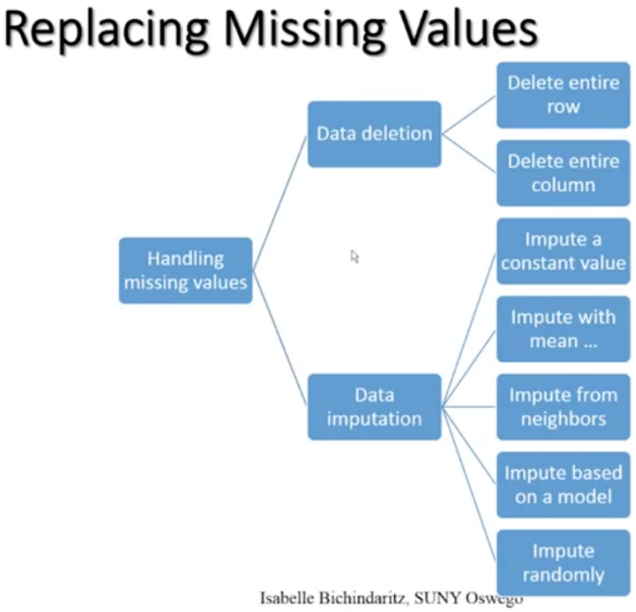
Data analysis process:

- Data analytics is actually a process involving several steps:
  - Application domain understanding.
  - Data understanding.
  - Data preprocessing.
  - Model building.
  - Training and evaluation.
  - Deployment.
- At this stage, we are going to build models for prediction tasks.



- Data preprocessing involves several tasks
  - Data cleaning
    - Dealing with missing values
    - Dealing with erroneous data and outliers
  - Data transformation
    - Changing data types (discretization)
    - Changing range of data values (normalization)
    - Adding variables
  - Data reduction
    - Feature selection
    - Sampling

Missing Values



Normalization and Discretization

Discretization

- Discretization transforms data from numeric into nominal data type.
- Effects of discretization:
  - Smooths data.
  - Reduces noise.
  - Reduces data size.
  - Enables specific methods using nominal data.

- Discretization methods
  - Manual methods:
    - Distribution analysis.
  - Automatic methods:
    - Binning.
      - Equal-width binning
      - Equal-depth binning
    - Regression analysis.
    - Cluster analysis.
    - Natural partitioning.
- Advantage of each method:
  - Equal-width binning is more simple however very sensitive to outliers in the data.
  - Equal-depth binning scales well by keeping the distribution of the data however the bin values may be more difficult to interpret.
- Smoothing of data can be accomplished by replacing the values in a bin by statistic such as average (numeric data), median (numeric data), or mode (categorical data).

Binning equal-width binning

- Equal-width binning
  - Given a range of values [min, max], we divide in intervals of approximately same width; either we set the width arbitrarily to w, or we set the desired number of bins to n, in which case w is calculated as:

    $w = max - min / n$

  - Ex: if the range is [0, 100] and we want 4 bins, each bin will have a width of
    $100 - 0 / 4 = 25$
    the bins will be: [0, 24], [25, 49], [50, 74], [75, 100].

Binning equal-depth binning

Os intervalos podem não ter o mesmo tamanho.

- Equal-depth binning
  - Given a range of values [min, max], we place approximately the same number of instances in each bin by dividing the total number of samples nb by the desired number of samples in each bin (depth) d, in which case the number of bins n is calculated as:

    n = nb / d

  - Ex: if the range is [0, 100] for 100 samples of different values (for example 99 is missing), we want 20 samples in each bin, the number of bins will be:
    100 / 20 = 5
    the bins will be: [0, 19], [20, 39], [40, 59], [60, 79], [80, 100].

Normalization

Data Normalization

- Normalization consists changing the scale in the data.
- When having data of mixed scale, some data analytics methods do not behave well (Ex: age and income have widely different ranges).
- For example, it is frequent to scale all data between the range [-1, 1] or [0, 1].
- Generally, data are scaled into a smaller range.
- Methods include:
  - Min-max normalization
  - Z-score normalization
  - Decimal scaling

- Comparison between the methods
  - The method that preserves the original data distribution is decimal scaling, therefore it preserves more than the others the shape of the data repartition. It acts similarly to image resizing in photo editing software (shrink / magnify).
  - Z-score normalization is the most used because the resulting distribution is going to be normal, which is advantageous with certain statistical methods. However it distorts the natural shape of the data distribution.
  - Min-max normalization can accommodate any new range we want, not only [0, 1] and [-1, 1] like the other ones.

Min-max Normalization

- Min-max normalization transforms data from range [m, M] into range [m', M'] using the formula

  val' = (val – m) / (M – m) * (M' – m') + m'

- Example: normalizing into [0, 1] age values between [0, 150]
  age 50 → 0.33 (intuitively)
  check val' = (50 – 0) / (150 – 0) * (1 – 0) + 0
  = 50 / 150 = 1/3 = 0.33

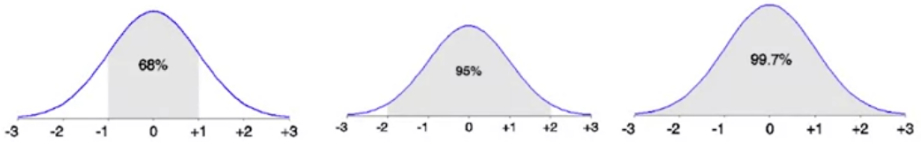
Z-score Normalization

- Z-score normalization

  val' = val – mean / std

- Ex: normalizing age values between [0, 150]
  where mean age in the population is 36.8 and standard deviation is 12
  age = 50 → val' = 50 – 36.8 / 12 = 1.1

- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it



Decimal scaling

- Decimal scaling

  val' = val / 10^n

  where n is determined such as the largest val' would be less than 1

  this formula transforms the values into interval [-1, 1] is there are negative values, and into [0, 1] otherwise.

- Ex: normalizing age values between [0, 150]

  we want the highest age to be less than 1, therefore divide by 1,000 = 10^3

  age = 50 → val' = 50 / 10^3 = 0.05

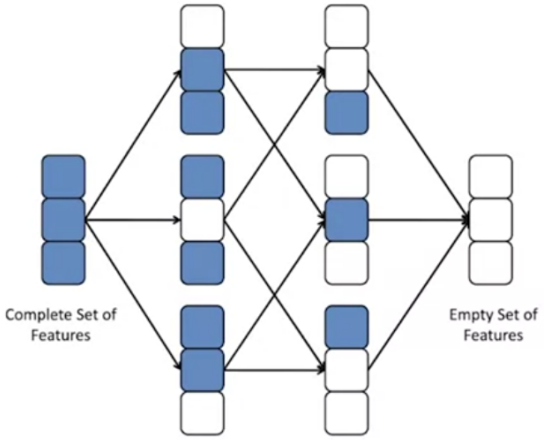
Data Reduction

- Data reduction can take several forms:
  - Feature selection.
  - Sampling.
  - Data compression.
  - Data aggregation.
  - etc.

Feature Selection

- Feature selection is also called dimensionality reduction.
- A feature is also called a variable (or a column).
- It is very important in biomedical data due to an often large number of features available – the curse of dimensionality (Ex: number of gene expressions).
- It will be studied in a future module.

Overview of Feature Selection Methods



Search problem – N features → 2^N subsets possible (exponential problem)

Overview of Feature Selection Methods

- More generally, feature selection methods belong to several families:
  - Filter methods where features are selected independently from the data analysis task
  - Wrapper methods where features are selected based on the resulting performance in the data analysis task
  - Embedded methods where feature selection is performed during the data analysis task (no as preprocessing).

- Some methods map the present features to another space where they may become more meaningful.

- This method is at the basis of such methods as
  - super vector machines (SVMs),
  - signal processing methods such as Fourier transformation or wavelet transformation,
  - statistical methods such as factor analysis or principal component analysis.

Filter Methods

- Filter methods for features and classes
  - A class is a predefined label associated to a particular sample.
  - Ex: two classes: duck / swan

  - They may measure uncertainty, distance or similarity, dependence, or consistency between classes.

  - They are independent from the data analytics task and therefore can be applied to several different tasks:
    - In this course, you will learn about several tasks, among which prediction and clustering are the main ones.

  - They are more simple than the wrapper methods and consequently are more efficient (take less time and resources).

- Filter methods examples (see module Resources for statistics background)
  - T-test of a feature between two classes (p-value < 0.05, numeric features).
  - Correlation coefficient (numeric features) between a feature and a class.
  - Chi-square (nominal features).
  - ...

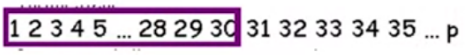
- They can yield a ranking of features, where features are ranked based on their differences between two classes.

- BSS/WSS (Dudoit et al. 2001)
- Between Sum of Squares (BSS) / Within Sum of Squares (WSS)
  - Two classes A and B, n_a samples in A, n_b samples in B, m the overall mean, m_a the mean in A, m_b the mean in B.
  - $BSS = n_a (m_a - m)^2 + n_b (m_b - m)^2$
  - $WSS = sum_{i \text{ in } A} (x_i - m_a)^2 + sum_{i \text{ in } B} (x_i - m_b)^2$
  - BSS is a measure of separation between the classes, while WSS is a measure of cohesion on each class.
  - BSS/WSS – ratio between to within groups sum of squares - can be used for ranking features by their discrimination power between two classes.

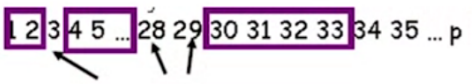
Wrapper Methods

- Wrapper methods select best features for a prediction method.
- Example:
  - Bayesian Model Averaging (BMA) selects best features using Bayesian (probabilistic) method / algorithm.
  - It is a multivariate variable selection technique.
  - Typical model selection approaches select a model and then proceed as if the selected model has generated the data --> overconfident inferences
  - Some advantages of BMA are:
    - Fewer selected genes
    - Can be generalized to any number of classes
    - Posterior probabilities for selected genes and selected models.

- BMA averages predictions from several models (Yeung, Bumgarner, Raftery 2005)
  - Pre-processing step: Rank genes using BSS/WSS ratio.
  - Initial step:



  - Repeat until all genes are processed:



- Output: selected genes and models with their posterior probabilities.

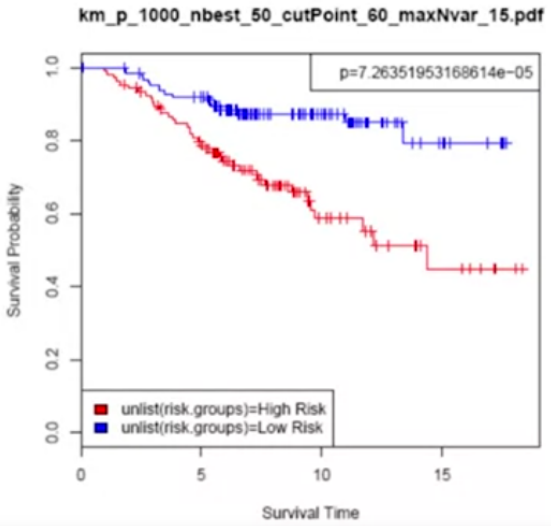
- Wrapper can search for best features according to several search methods involving search direction, search strategy, and evaluation measure:
  - Exhaustive tests all combinations of features.
  - Best first selects first the best overall features and builds on these.
  - Step wise adds or removes features one by one as long as each improves on the prediction task.
  - Step wise can go forward (starts from an empty set and adds features one by one) or backward (starts from all features and removes them one by one).

Evaluation

Evaluation of Feature Selection Methods

- Evaluation of the best feature subset can take several approaches, the main ones being:
  - Comparison of prediction performance.
  - Robustness testing.
  - Efficiency (time and resource comparison).
  - Value added (feature ranking, feature posterior probability, feature weighting).
- Prediction performance
  - Compare the predictive power of the subset of predictive genes using the predictive task – or several of them - and determine
    - Which is the best feature subset, or
    - Whether the feature subsets perform better than the complete set of features.

- Prediction performance (Annest, Bumgarner, Yeung 2009)
  - Genes selected for a high-risk group (red) in comparison with a low-risk group of breast cancer patients
  - Survival curve between the two groups is significant (p-value < 0.05)
  - We conclude that the genes / features selected add a value to the task of predicting the severity of this disease.
  - cBioPortal can be used to perform this analysis once genes have been selected.



- Prediction performance (Bichindaritz, 2010)
  - Often feature selection methods compare themselves with one another by selecting the smallest feature subset that maximizes performance.

Prediction method (algorithm)	#errors Leukemia2 (/34)	# errors Leukemia3 (/34)	Average accuracy
BMA+KNN	1	1	**97%**
BMA+DT	3	5	88%
BMA+LR	1	4	93%
BMA+NB	4	3	89%
BMA+NN	1	2	96%
BMA+SVM	1	2	96%
BSS/WSS +KNN	1	1	**97%**
BSS/WSS +DT	3	3	91%
BSS/WSS +LR	2	4	91%
BSS/WSS +NB	3	5	88%
BSS/WSS +NN	2	1	96%
BSS/WSS +SVM	0	1	**99%**

- Robustness testing
  - Consists in testing how well the feature set selected resists noise.
  - Obtained by adding noise (slightly modifying) the feature values and seeing whether the feature set remains unchanged.
  - Can participate in avoiding overfitting.

Sampling

Data Sampling

- Data sampling refers to creating a subset or sample of the complete dataset.
- The sample needs to be representative.
- Main methods:
  - Simple random sampling with replacement.
  - Simple random sampling without replacement.
  - Stratified sampling.

R

Useful commands

- **Important commands.** Some important commands include:

ls() to list the content of the memory.
rm() to empty the memory.
rm(object) to remove an object from memory.
q() to quit.
summary(object) to display summary characteristics of an object.
class(object) to display the class (type) of an object.

Packages

- Packages are libraries of functions to use in addition to the standard functions. They need to be loaded specifically. There are two types of packages:
- Standard packages, which can be installed from the *Package* menu, choosing *Load package in the graphical user interface (GUI)*.
- Packages to install from a local zip file, which can be installed from the *Package* menu, choosing *Install package(s) from local zip files…*, which proposes to load a zipped package from the working directory.
- Packages can also be installed with install.packages().
- Once packages are installed, they can be loaded with library().

Several repositories host contributed packages. Notably, Bioconductor hosts a large number of biology specific packages and datasets. Installation typically takes the form of:

```
source("https://bioconductor.org/biocLite.R")biocLite("examplePackage")
```

Gene Alterations with R

- We discussed several packages for gene alterations analysis with R, and there are many more, new packages are added every day:
  - Mutations
    - CancerMutationAnalysis
  - Copy number variations
    - seqCNA
    - R-Gada
    - CNV-seq
  - Methylation
    - methyAnalysis
    - methylPipeline.

Distributions (Bioconductor x Anaconda)

- Some distributions are available combining several tools for bioinformatics, such as
  - Bioconductor (http://bioconductor.org) for bioinformatics packages.
  - Anaconda (https://www.continuum.io/downloads) for data Science includes Python, R, and Scala with their most popular packages (including Bioconductor).

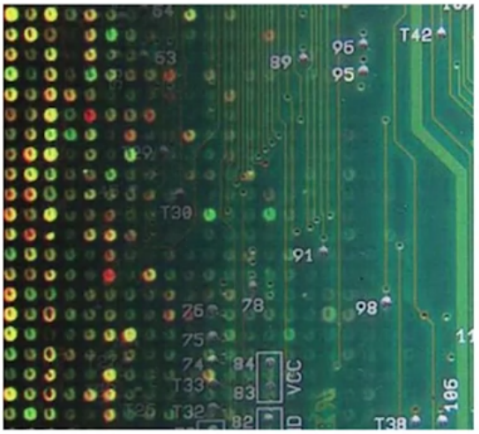
R is a package that can be added to Anaconda. However, by default, it is provided with Python only.
Para instalar o R no Anaconda:

conda install -c r r-irkernel

https://datatofish.com/r-jupyter-notebook/

Differentially Expressed Genes

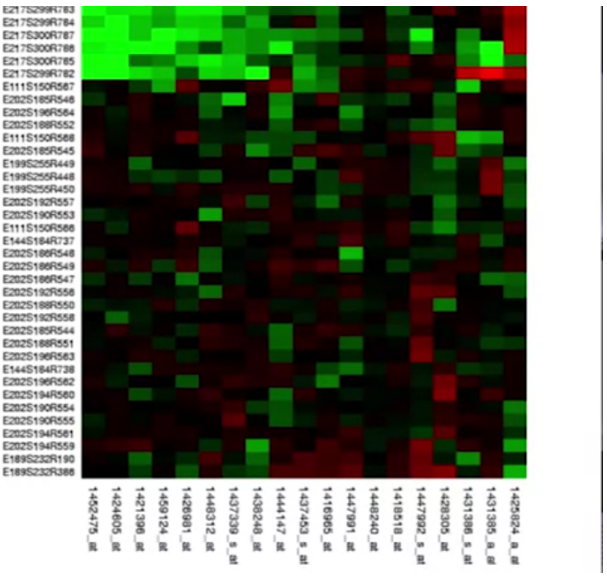
- Methods for selecting differentially expressed genes
  - Calculate expression level difference with normal tissue and apply feature selection method
  - DESeq2
  - EdgeR
  - Cuttdiff
  - Limma
  - etc.



HeatMap

Heatmaps

- A heatmap is a visualization of expression levels of features (genes etc.) using a color scale.
- In general, features are in rows and sample in columns.
- They are often combined with some clustering to group either patients in classes or features in groups, which is represented by a tree at the top and/or the side.



https://warwick.ac.uk/fac/sci/moac/people/students/peter_cock/r/heatmap
https://stat.ethz.ch/R-manual/R-devel/library/stats/html/heatmap.html

Tutorial

https://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/BRCA/20160128/gdac.broadinstitute.org_BRCA.Merge_rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes_normalized__data.Level_3.2016012800.0.0.tar.gz

```
tania@POLINESIA: ~/R_COURSERA/data
Arquivo  Editar  Ver  Pesquisar  Terminal  Ajuda
tania@POLINESIA:~$ cd R_COURSERA/
tania@POLINESIA:~/R_COURSERA$ cd data/
tania@POLINESIA:~/R_COURSERA/data$ head BRCA.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes_normalized__data.data.txt
Hybridization REF       TCGA-3C-AAAU-01A-11R-A41B-07    TCGA-3C-AALI-01A-11R-A41B-07    TCGA-3C-AALJ-01A-31R-A41B-07    TCGA-3C-AALK-01A-11R-A41B-07    TCGA-4H-AAAK-01A-12R-A41B-07    TCGA-5L-AAT0-01A-12R-A41B-07    TCGA-5L-AAT1-01A-12R-A41B-07    TCGA-5T-A9QA-01A-11R-A41B-07    TCGA-A1-A0SB-01A-11R-A144-07    TCGA-A1-A0SD-01A-11R-A115-07    TCGA-A1-A0SE-01A-11R-A084-07    TCGA-A1-A0SF-01A-11R-A144-07    TCGA-A1-A0SG-01A-11R-A144-07    TCGA-A1-A0SH-01A-11R-A084-07    TCGA-A1-A0SI-01A-11R-A144-07    TCGA-A1-A0SJ-01A-11R-A084-07    TCGA-A1-A0SK-01A-12R-A084-07    TCGA-A1-A0SM-01A-11R-A084-07    TCGA-A1-A0SN-01A-11R-A144-07    TCGA-A1-A0SO-01A-22R-A084-07    TCGA-A1-A0SP-01A-11R-A084-07    TCGA-A1-A0SQ-01A-21R-A144-07    TCGA-A2-A04N-01A-11R-A115-07    TCGA-A2-A04P-01A-31R-A034-07    TCGA-A2-A04Q-01A-21R-A034-07    TCGA-A2-A04R-01A-41R-A109-07    TCGA-A2-A04T-01A-21R-A034-07    TCGA-A2-A04U-01A-11R-A115-07    TCGA-A2-A04V-01A-21R-A034-07    TCGA-A2-A04W-01A-31R-A115-07    TCGA-A2-A04X-01A-21R-A034-07    TCGA-A2-A04Y-01A-21R-A034-07    TCGA-A2-A0CK-01A-11R-A22K-07    TCGA-A2-A0CL-01A-11R-A115-07    TCGA-A2-A0CM-01A-31R-A034-07    TCGA-A2-A0CO-01A-13R-A22K-07    TCGA-A2-A0CP-01A-11R-A034-07    TCGA-A2-A0CQ-01A-21R-A034-07    TCGA-A2-A0CR-01A-11R-A22K-07    TCGA-A2-A0CS-01A-11R-A115-07    TCGA-A2-A0CT-01A-31R-A056-07    TCGA-A2-A0CU-01A-12R-A034-07    TCGA-A2-A0CV-01A-31R-A115-07    TCGA-A2-A0CW-01A-21R-A115-07    TCGA-A2-A0CX-01A-21R-A00Z-07    TCGA-A2-A0CY-01A-12R-A034-07    TCGA-A2-A0CZ-01A-11R-A034-07    TCGA-A2-A0D0-01A-11R-A00Z-07    TCGA-A2-A0D1-01A-11R-A034-07    TCGA-A2-A0D2-01A-21R-A034-07    TCGA-A2-A0D3-01A-11R-A115-07    TCGA-A2-A0D4-01A-11R-A00Z-07    TCGA-A2-A0EM-01A-11R-A034-07    TCGA-A2-A0EN-01A-13R-A084-07    TCGA-A2-A0EO-01A-11R-A034-07    TCGA-A2-A0EP-01A-52R-A22U-07    TCGA-A2-A0EQ-01A-11R-A034-07    TCGA-A2-A0ER-01A-21R-A034-07    TCGA-A2-A0ES-01A-11R-A115-07    TCGA-A2-A0ET-01A-31R-A034-07    TCGA-A2-A0EU-01A-22R-A056-07    TCGA-A2-A0EV-01A-11R-A034-07    TCGA-A2-A0EW-01A-21R-A115-07    TCGA-A2-A0EX-01A-21R-A034-07    TCGA-A2-A0EY-01A-11R-A034-07    TCGA-A2-A0ST-01A-12R-A084-07    TCGA-A2-A0SU-01A-11R-A084-07    TCGA-A2-A0SV-01A-11R-A084-07    TCGA-A2-A0SW-01A-11R-A084-07    TCGA-A2-A0SX-01A-12R-A084-07    TCGA-A2-A0SY-01A-31R-A084-07    TCGA-A2-A0T0-01A-22R-A084-07    TCGA-A2-A0T1-01A-21R-A084-07    TCGA-A2-A0T2-01A-11R-A084-07    TCGA-A2-A0T3-01A-21R-A115-07    TCGA-A2-A0T4-01A-31R-A084-07
```

```
: head(mrnaIDs)
```

	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	...	V1205	V1206	V1207	V1208	V1209	V1210	V1211	V1212	V1213	V1214
	TCGA-3C-AAAU-01A-11R-A41B-07	TCGA-3C-AALI-01A-11R-A41B-07	TCGA-3C-AALJ-01A-31R-A41B-07	TCGA-3C-AALK-01A-11R-A41B-07	TCGA-4H-AAAK-01A-12R-A41B-07	TCGA-5L-AAT0-01A-12R-A41B-07	TCGA-5L-AAT1-01A-12R-A41B-07	TCGA-5T-A9QA-01A-11R-A41B-07	TCGA-A1-A0SB-01A-11R-A144-07	TCGA-A1-A0SD-01A-11R-A115-07	...	TCGA-UL-AAZ6-01A-11R-A41B-07	TCGA-UU-A93S-01A-21R-A41B-07	TCGA-V7-A7HQ-01A-11R-A33J-07	TCGA-W8-A86G-01A-21R-A36F-07	TCGA-WT-AB41-01A-11R-A41B-07	TCGA-WT-AB44-01A-11R-A41B-07	TCGA-XX-A899-01A-11R-A36F-07	TCGA-XX-A89A-01A-11R-A36F-07	TCGA-Z7-A8R5-01A-42R-A41B-07	TCGA-Z7-A8R6-01A-11R-A41B-07

[*]:
```
# cell #1
mrnaNorm <- read.table("BRCA.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes_normalized__data.data.txt",
                       header = F, fill = T, skip = 2)
class(mrnaNorm)
```

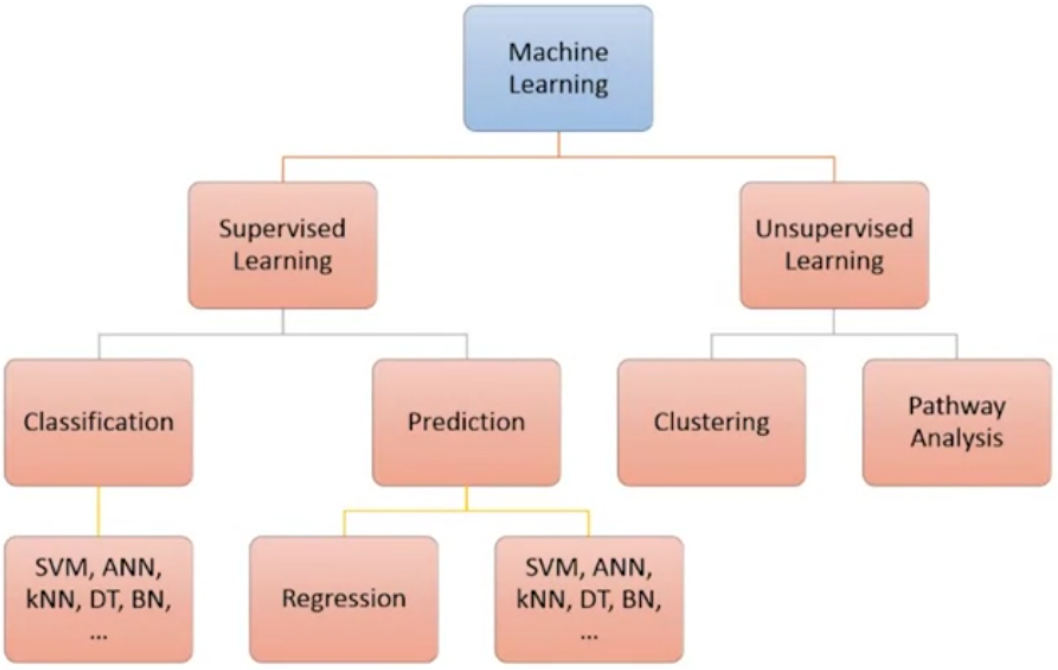
[2]:
```
head(mrnaNorm)
```

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	...	V1204	V1205	V1206	V1207	V1208	V1209	V1210	V1211	V1212	V1213
?|100130426	0.0000	0.0000	0.9066	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
?|100133144	16.3644	9.2659	11.6228	12.0894	6.8468	3.9889	0.0000	1.4644	15.3396	...	0.3992	4.3126	0.0000	5.5624	0.0000	0.0000	14.3858	22.3240	2.2638	6.8865
?|100134869	12.9316	17.3790	9.2294	11.0799	14.4298	13.6090	10.5949	8.9958	14.3935	...	14.3720	10.8828	3.0792	14.3711	6.3091	3.2580	21.4409	27.2744	7.2933	24.7795
?|10357	52.1503	69.7553	154.2974	143.8643	84.2128	114.2572	115.9984	107.5628	116.3870	...	135.6241	136.1288	29.9974	128.3151	53.6278	42.2643	137.7756	64.1427	85.0461	167.5511
?|10431	408.0760	563.8934	1360.8341	865.5358	766.3830	807.7431	1108.3945	1420.5021	657.2812	...	1570.1445	2886.3965	1721.8816	697.6744	1245.2681	1877.4180	652.7559	722.7208	1140.2801	1003.5668
?|136542	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

We also want the patient IDs in the first row, which is why we read only this first line and place it in a second dataframe callses 'mrnaIDs'. We also remove the first column with '[, -1]', the one entitled 'Hybridization REF', because it is not a patientID.

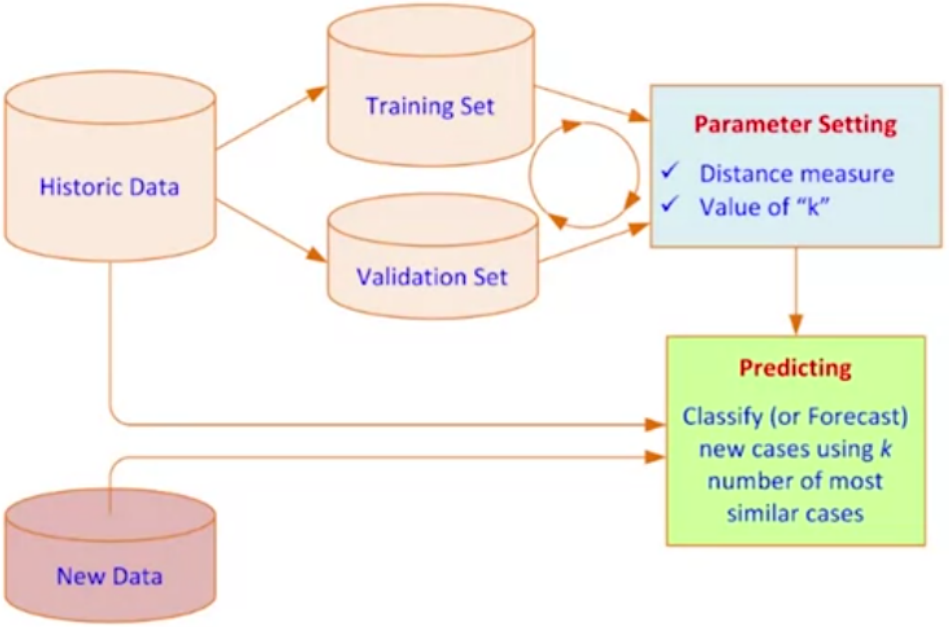
[3]:
```
# cell #2
```

Machine Learning



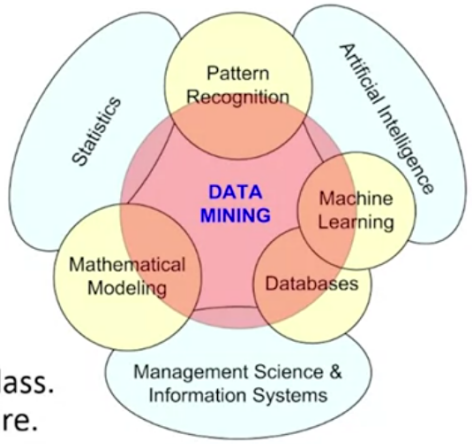
Supervised

Classification and Prediction Methods (Model Building)



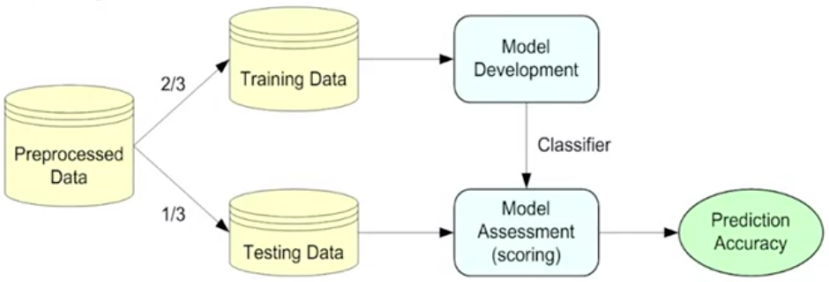
Overview of Classification and Prediction Methods

- Classification and prediction are the most frequently used tasks in data analytics.
- Methods to achieve this task are varied and span many disciplines, the most well known being machine learning and statistics.
- Classification predicts a categorical valued class. Prediction predicts a numerical valued feature.



- Classification and prediction tasks aim at building models that describe and distinguish classes or concept for future prediction.
- Ex: Diagnosing a disease is a typical classification task.
  Evaluating the risk or severity of a disease in a patient is a typical prediction task.
- What differentiates between the methods is the type of algorithm (I often call algorithms methods in this course), or process used to build the predictive models from data – whether it is based on analogies, rules, neural networks, probabilities, or statistics.

- This model building involves using a dataset for training, then another one for testing.



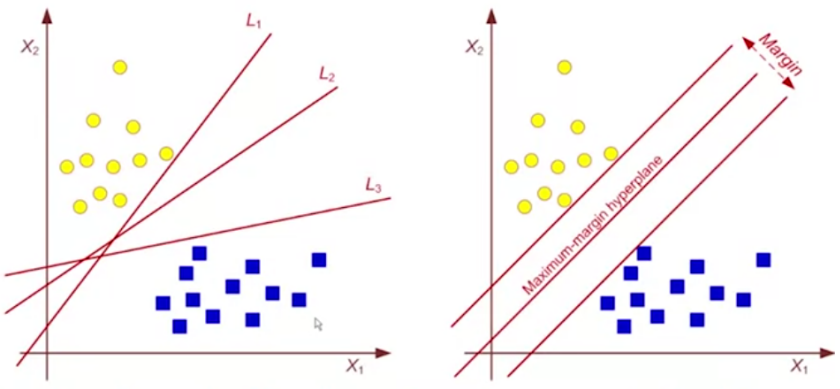
- Often, there is only one dataset to work from, and there are several strategies to create the training / testing framework.
- The end goal is to put the model in production, which is when it will be applied to any new data – a new patient coming for a visit, or online.

Classification Methods Based On Analogy

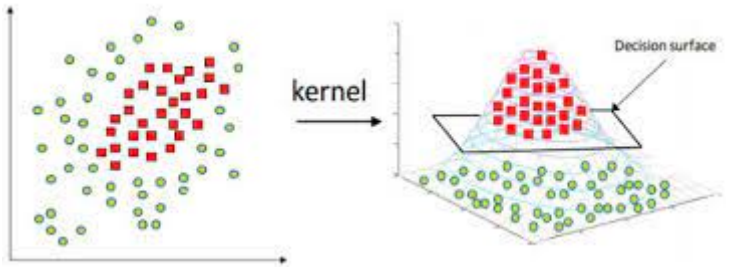
- Two main classification methods are based on analogy:
  - The support vector machine (SVM).
  - Nearest-neighbor classifier (kNN).

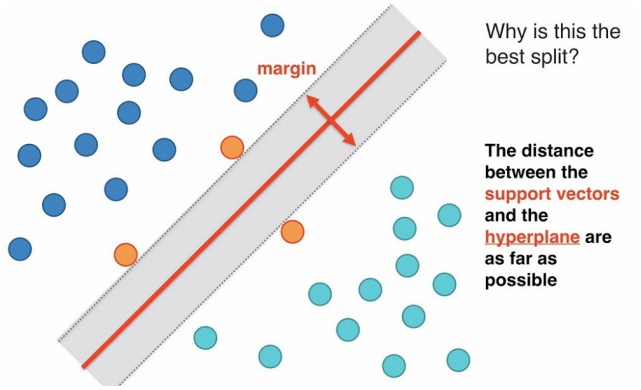
Support Vector Machine (SVM)

- Goal of SVM: to generate mathematical functions that map input variables to desired outputs for classification or regression type prediction problems.
  - First, SVM uses nonlinear kernel functions to transform non-linear relationships among the variables into linearly separable feature spaces.
  - Then, the maximum-margin hyperplanes are constructed to optimally separate different classes from each other based on the training dataset.
- SVM has solid mathematical foundation!
- A hyperplane is a geometric concept used to describe the separation surface between different classes of things.
  - In SVM, two parallel hyperplanes are constructed on each side of the separation space with the aim of maximizing the distance between them.
- A kernel function in SVM uses the kernel trick (a method for using a linear classifier algorithm to solve a nonlinear problem)
  - The most commonly used kernel function is the radial basis function (RBF).
- SVMs are the most widely used kernel-learning algorithms for wide range of classification and regression problems.
- SVMs represent the state-of-the-art by virtue of their excellent generalization performance, superior prediction power, ease of use, and rigorous theoretical foundation.
- Most comparative studies show its superiority in both regression and classification type prediction problems.



➢ Many linear classifiers (hyperplanes) may separate the data





Nearest Neighbor Classifier (kNN)

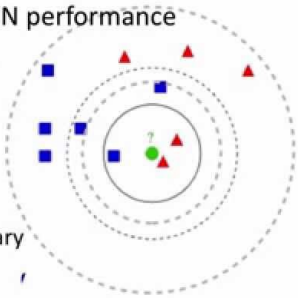
https://www.youtube.com/watch?v=PB4qATziTlQ

https://blogdozouza.wordpress.com/2019/04/11/introducao-ao-algoritmo-k-nearest-neighbour-codigo-python/

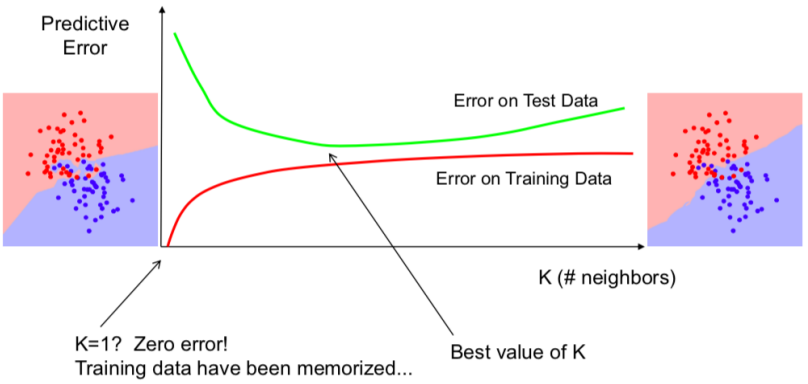
- SVMs → time-demanding, computationally intensive iterative derivations
- *k*-NN is a simplistic and logical prediction method, that produces very competitive results – in particular for Big Data
- *k*-NN is a prediction method for classification as well as regression types (similar to SVM)
- *k*-NN is a type of instance-based learning (or lazy learning) – most of the work takes place at the time of prediction (not at modeling)
- *k* : the number of neighbors used
- Data are kept after modeling (not only the model).

Choosing the value of k

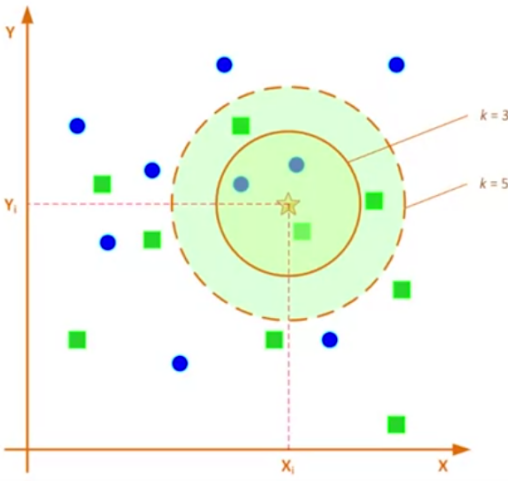
- Value of k has strong effect on kNN performance
  - large value → everything classified as the most probable class: P(y)
  - small value → highly variable, unstable decision boundaries
    - small changes to training set → large changes in classification
  - affects "smoothness" of the boundary
- Selecting the value of k
  - set aside a portion of the training data (validation set)
  - vary k, observe training → validation error
  - pick k that gives best generalization performance



Error rates and K

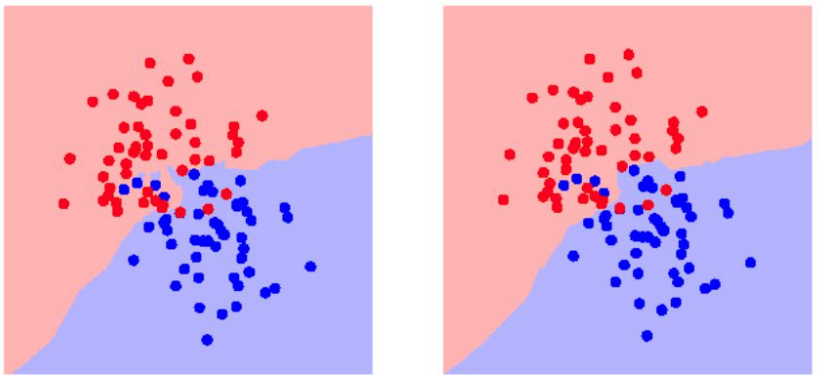


- How many neighbors ?
- The answer depends on the value of k.
- Which distance measure ?
- It depends on the data –
  - Euclidian distance or more complex.



- piecewise linear decision boundary
- Increasing k "simplifies" decision boundary
  - Majority voting means less emphasis on individual points

K = 5 K = 7



Classification Methods Based On Rules

Decision Tree

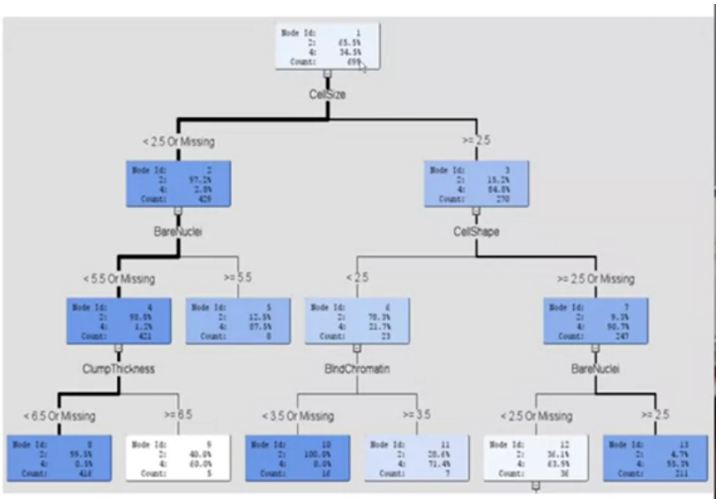
A general algorithm for decision tree building

- Employs the divide and conquer method
- Recursively divides a training set until each division consists of examples from one class

1. Create a root node and assign all of the training data to it
2. Select the best splitting attribute
3. Add a branch to the root node for each value of the split. Split the data into mutually exclusive subsets along the lines of the specific split
4. Repeat the steps 2 and 3 for each and every leaf node until the stopping criteria is reached.

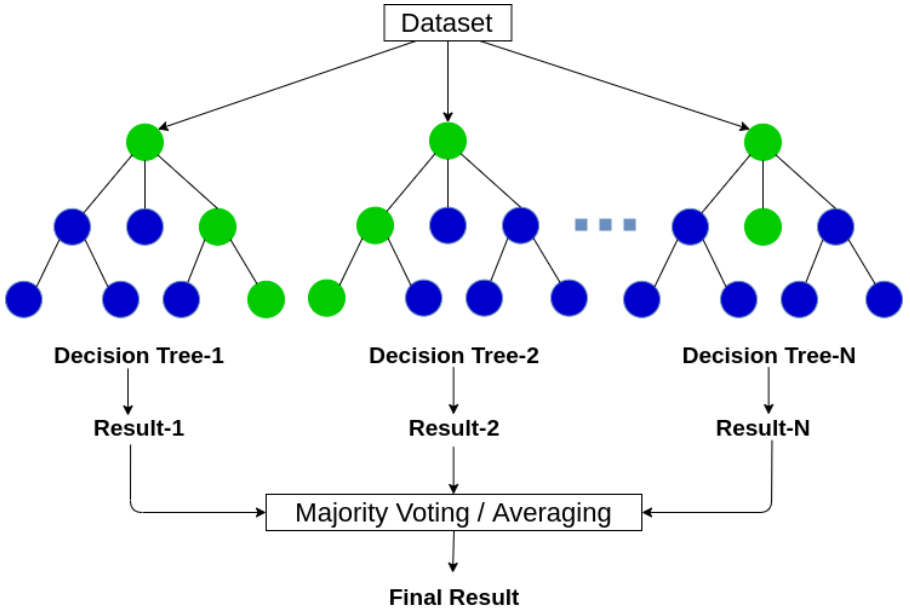
- Decision tree algorithms mainly differ on
  - Splitting criteria
    - Which variable to split first?
    - What values to use to split?
    - How many splits to form for each node?
  - Stopping criteria
    - When to stop building the tree
  - Pruning (generalization method)
    - Pre-pruning versus post-pruning
- Most popular decision tree algorithms include
  - ID3, C4.5, C5; CART; CHAID; M5

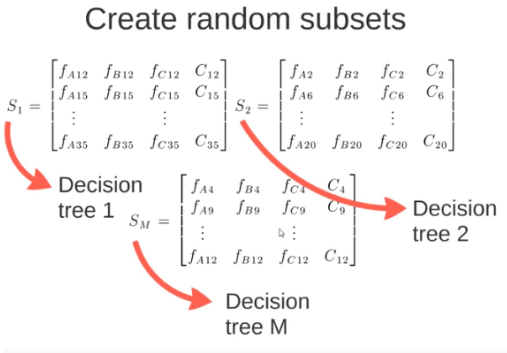
- Decision tree building process is very efficient for Big Data, and has the advantage of being easily understandable.
- For example, on this tree to classify sample between two classes 2 and 4, samples with Cellsize < 2.5 AND BareNuclei < 5.5 AND ClumpThickness < 6.5 belong to class 2 in 99.5% of the cases.



Random Forest

https://www.youtube.com/watch?v=loNcrMjYh64

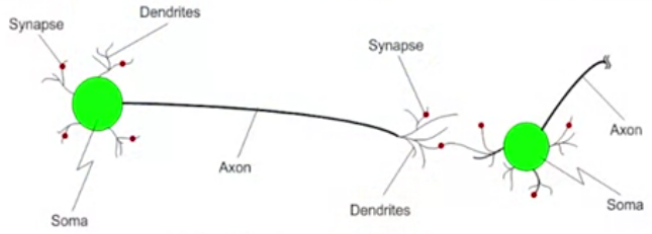




Classification Methods Based On Neural Networks

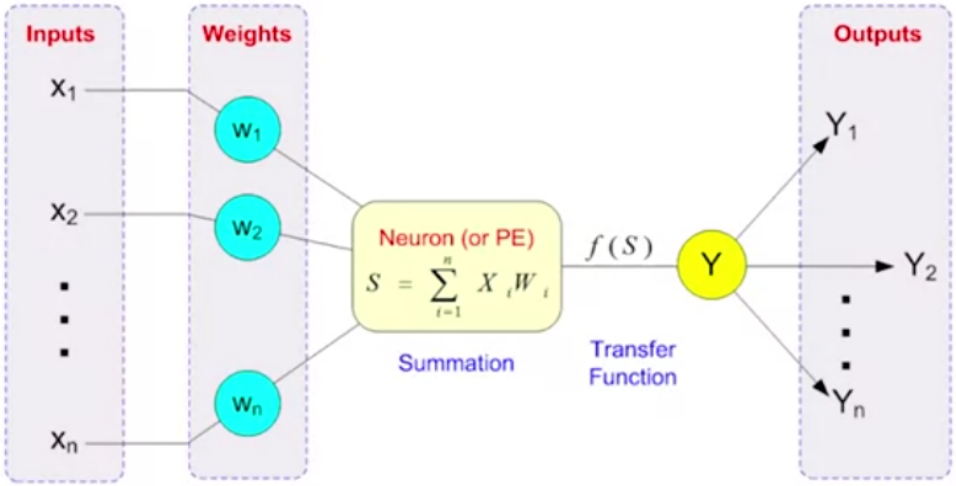
Neural Networks (NN)

- Neural networks (NN): a brain metaphor for information processing
- Neural computing
- Artificial neural network (ANN)
- Many uses for ANN for
  - pattern recognition, forecasting, prediction, and classification.



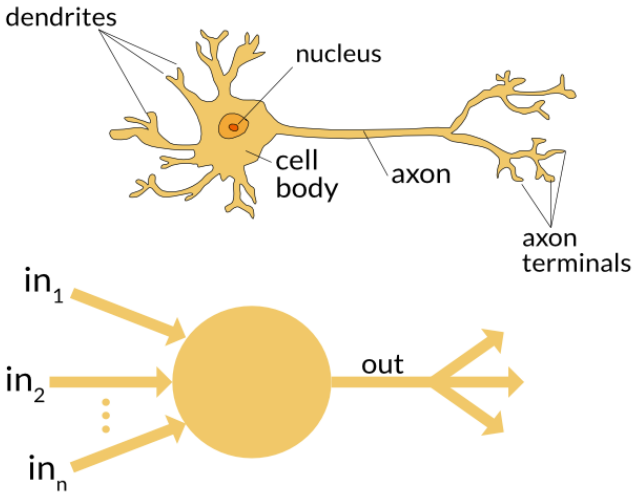
Isabelle Bichindaritz, SUNY Oswego

- A single neuron (processing element – PE) with inputs and outputs

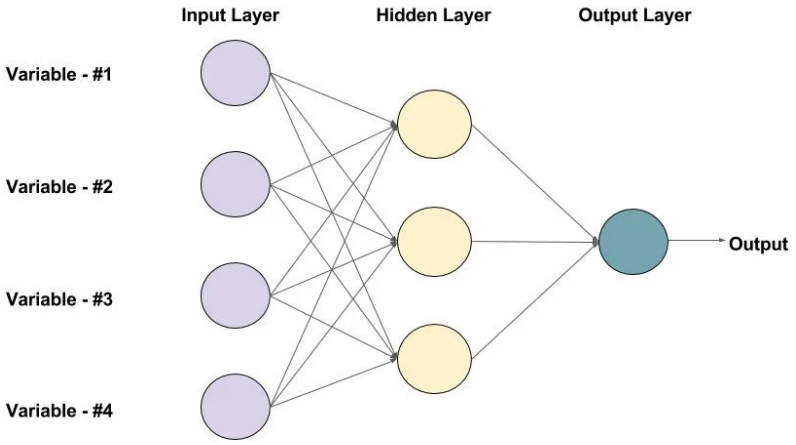


- Architecture of a neural network is driven by the task it is intended to address
  - Classification, regression, clustering, general optimization, association, ….
- Most popular architecture: Feedforward, multi-layered perceptron with backpropagation learning algorithm
  - Used for both classification and regression type problems
- Others – Recurrent, self-organizing feature maps, Hopfield networks, …

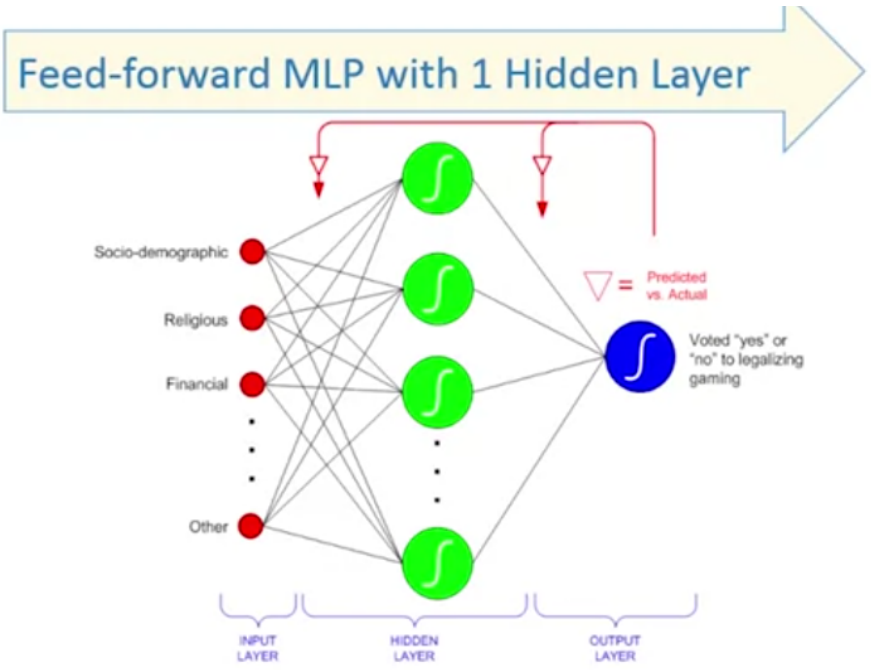
Perceptron:



Feed-forward: it is a multi-layered perceptron

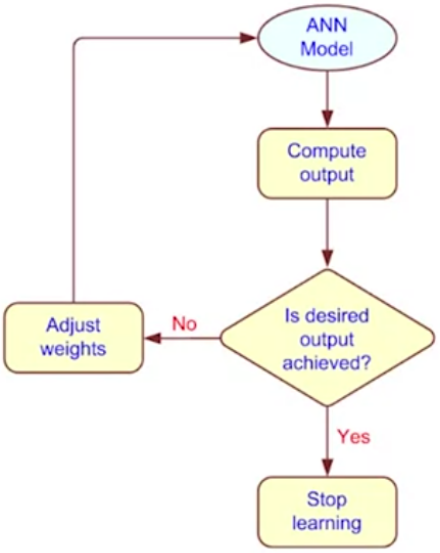


An example of a Feed-forward Neural Network with one hidden layer (with 3 neurons)

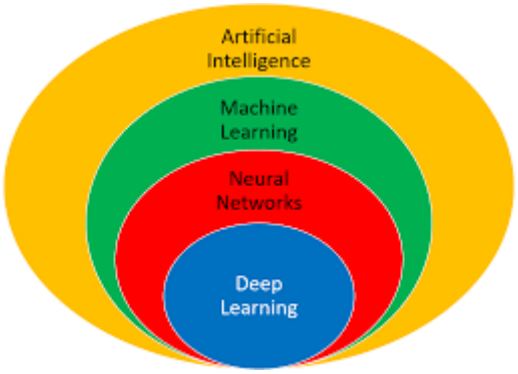


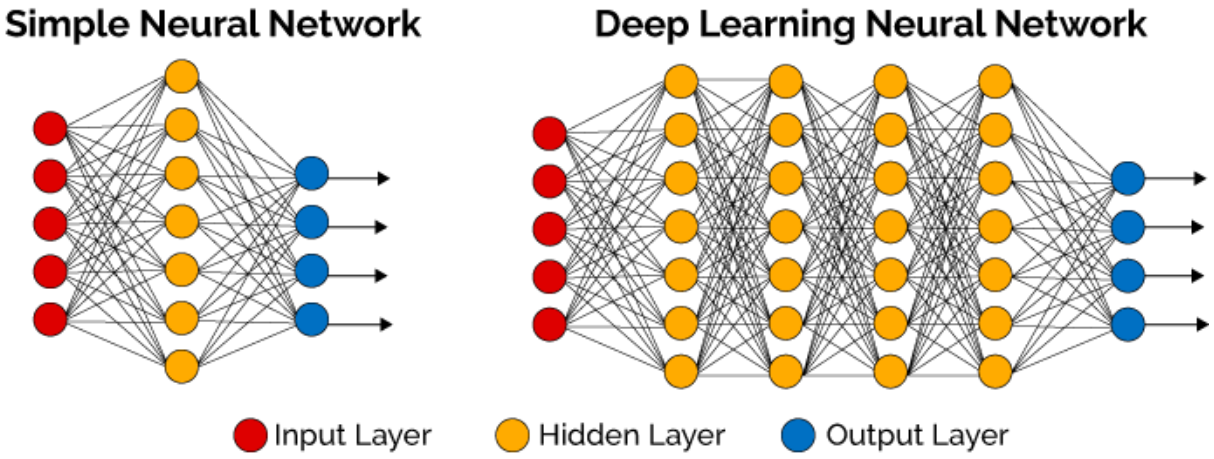
Isabelle Bichindaritz, SUNY Oswego

- Neural networks learn the prediction model, which is here a set of weights representing the underlying relationship between inputs and outputs, or just among the inputs.
- Once trained, the neural network stops learning and can be applied to new data for classification or prediction tasks.
- Deep learning is based on advanced neural networks.
- They provide excellent pattern recognition performance and are one of the major methods used in machine learning.
- Neural networks require important computing power, have a set of parameters to adjust, and are often not much understandable.



Deep Learning





Classification Methods Based On Statistics (Regressions and GLMs)

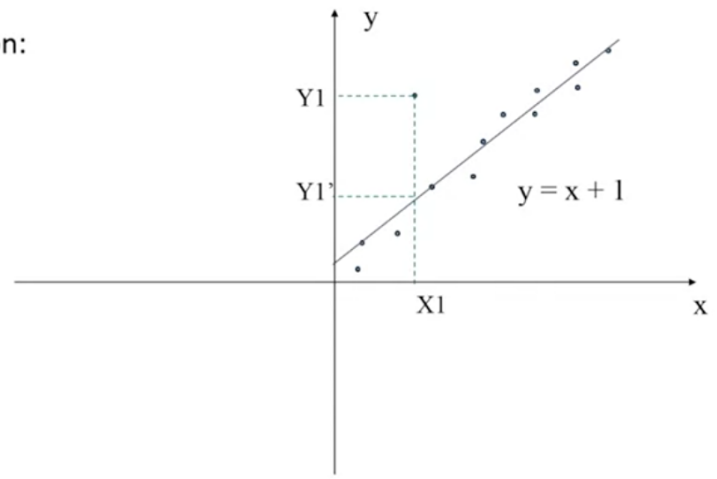
Regressão Linear e Logística

- Statistics provides many useful models for classification and prediction, among which are:
  - Regression models to predict a numeric feature.
  - Logistic regression models to predict a categorical feature (or class).
  - Generalized linear models (GLM) to predict a numeric or categorical feature with more flexibility of data distribution shape (does not need to be normal like in standard linear regression).
- They often require particular conditions to be used, such as a normal distribution, or independent features, which are often not completely met with Big Data.

- Linear regression for prediction:

  If X is fever, Y may calculate the severity of the flu.

- Linear regression learns weights associated with the features and a probability as to whether this feature is significantly important for the target feature (the one we want to calculate).

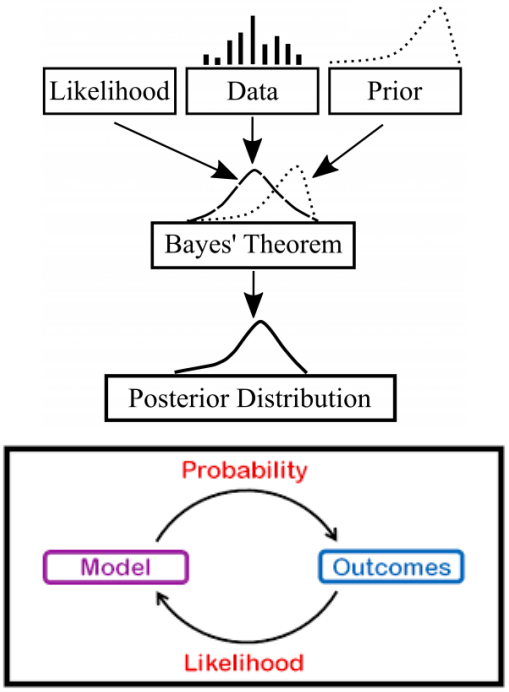


Modelos Lineares Generalizados (GLM)

- Statistical models like GLM do provide though some functions that others do not, for example the ability to consider confounding factors – or factors that may influence the classification and that we would like to isolate from the other features under consideration.

- Ex: age, gender, body mass index, socio-economic status.

Classification Methods Based On Probabilities (Bayesian)



However, we can calculate the probability of obtaining the results, given our model (scientific hypothesis (P(data|model).

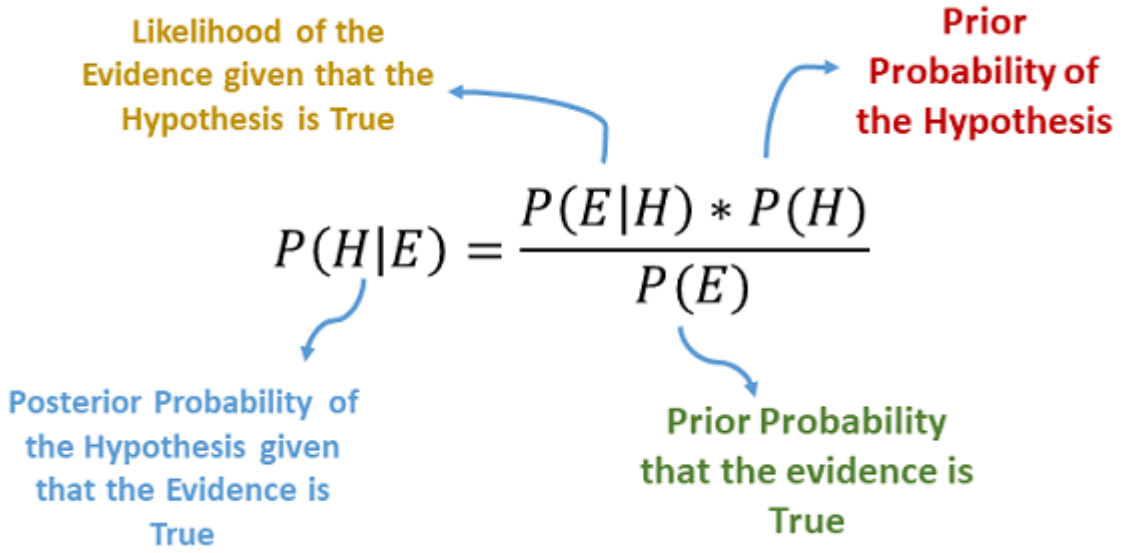
Ex: estou fazendo lançamento de dados e anoto quantas vezes Coroa aparece e me pergunto que dado os resultados obtidos qual probabilidade da moeda ser honesta (modelo).

- Bayesian networks, also called belief networks or graphical models, derive a predictive model from data based on Bayes theorem:

$$P(M \mid D) = \frac{P(D \mid M)P(M)}{P(D)}$$

where M is a model (hypotheses), D are the data

- P(M|D) is the posterior - updated belief that M is correct
- P(M) is our estimate that M is correct prior to any data
- P(D|M) is the likelihood.



$$P(A/B) = P(B/A) * P(A) / P(B)$$

P(A/B) : Probability of Purchasing Macbook after purchasing iPhone

P(A/B) : Probability of Purchasing iPhone after purchasing Macbook

P(A) : Probability of Purchasing Macbook

P(B) : Probability of Purchasing iPhone

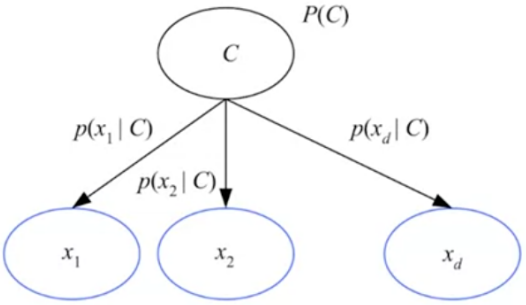
@ dataaspirant.com

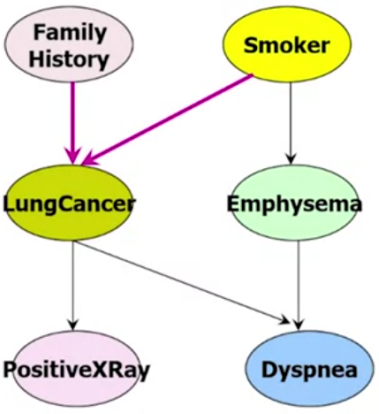
- To be able to infer a model, the model needs to evaluate:
  - The prior P(M)
  - The likelihood P(D/M).

- They allow to model situations in a probabilistic network but also to reason from the network – also called to perform inferences.

- Causal relationships are made explicit and can be used to propagate new facts or beliefs into the network.

- They are very important in biomedicine because they can calculate a probability associated with a diagnosis (or classification), for example in differential diagnosis.
- Graphical representations are highly valued for understandability – like in decision trees.
- There is a simplified version of this method called Naïve Bayes.
- These methods are based on the hypothesis of independence between features, which may require feature selection before the prediction task.





Bayesian Belief Networks

	(FH, S)	(FH, ~S)	(~FH, S)	(~FH, ~S)
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

The conditional probability table for the variable LungCancer

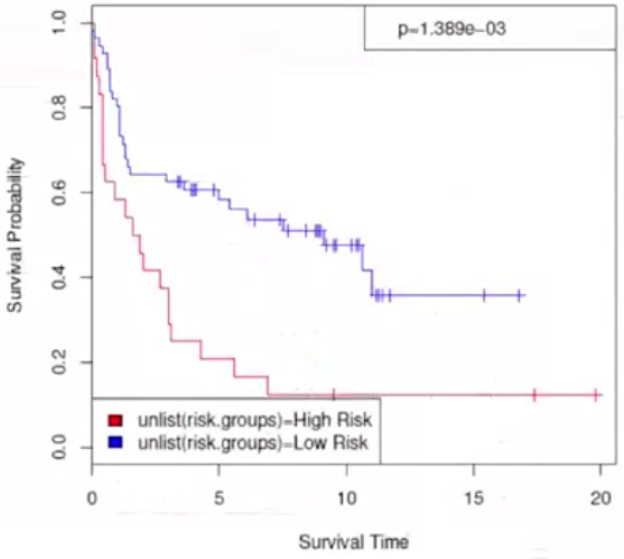
Naive Bayes

https://www.youtube.com/watch?v=IVKF_wmIdiI
https://www.youtube.com/watch?v=RftMipEbL48

Prediction Methods

- Survival analysis example (Annest et al. 2009)

  regression with Kaplan Meier survival (in years) curve showing the survival difference between two groups of patients: high-risk and low-risk.



Evaluating Models

- In order to evaluate a model's performance, specific measures have been designed.
- They also afford for the comparison between different models.
- The most important are accuracy, area under ROC curve, sensitivity, specificity, and error.

Classifiers

- Evaluation of classifiers
  - Predictive accuracy
    - Hit rate
  - Area under ROC curve
  - Speed
    - Model building; predicting
  - Robustness
  - Scalability
  - Interpretability
    - Transparency, explainability.

Matriz de Confusão

- In classification problems, the primary source for accuracy estimation is the confusion matrix

		True Class: Positive	True Class: Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
Predicted Class	Negative	False Negative Count (FN)	True Negative Count (TN)

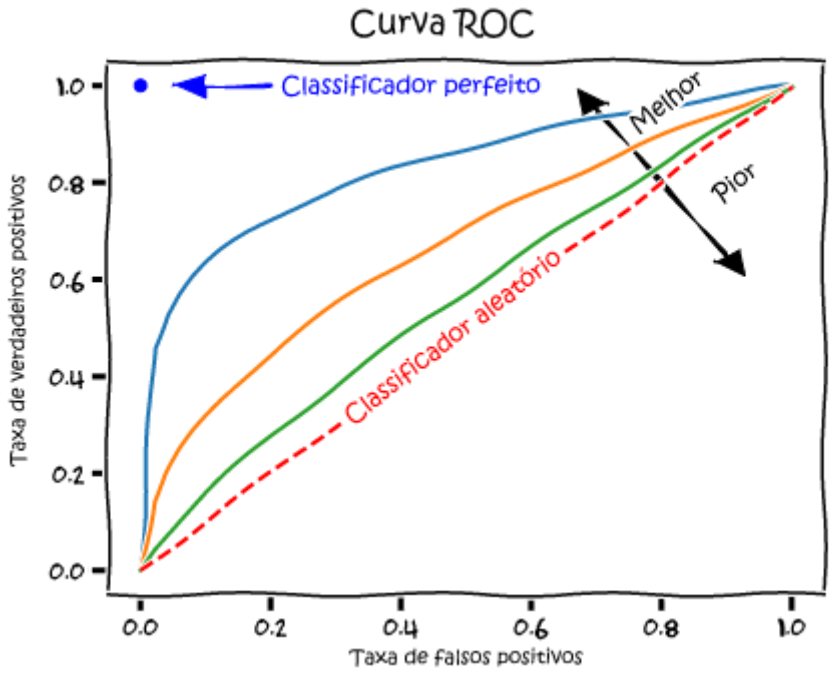
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$True\ Positive\ Rate = Sensitivity = \frac{TP}{TP + FN}$$

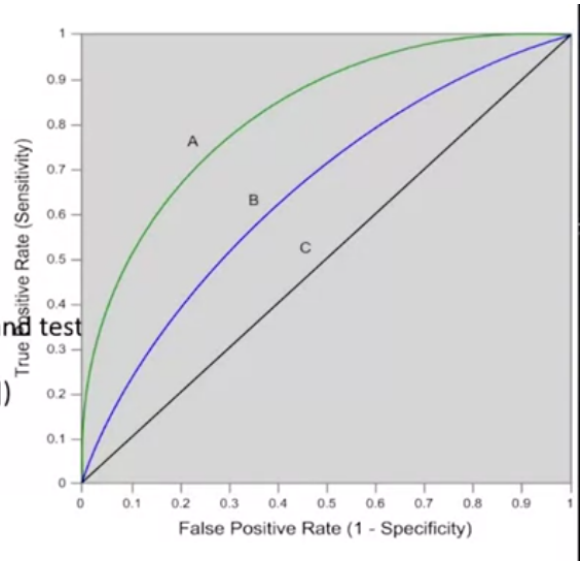
$$True\ Negative\ Rate = Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

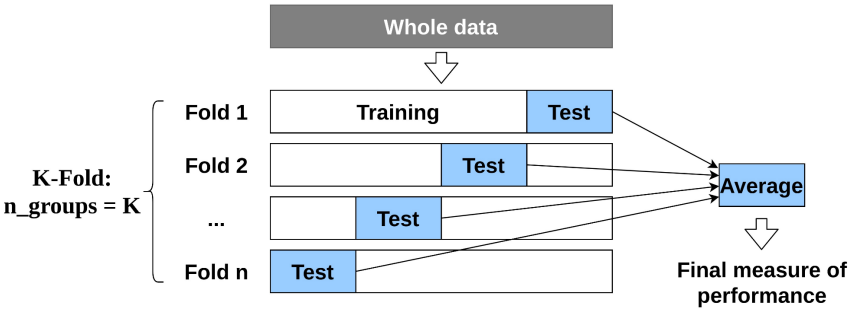
ROC Curve



- Independent training and test sets (preferred).
- *k*-Fold Cross Validation (rotation estimation)
  - Split the data into *k* mutually exclusive subsets
  - Use each subset as testing while using the rest of the subsets as training
  - Repeat the experimentation for *k* times
  - Aggregate the test results for true estimation of prediction accuracy training.
- Simple split (or holdout or test sample estimation)
  - Split the data into 2 mutually exclusive sets training (~70%) and test
  - For neural networks, the data is split into three sub-sets (training [~60%], validation [~20%], testing [~20%])
- Other estimation methodologies
  - Leave-one-out, bootstrapping, jackknifing
  - Area under the ROC curve.



Cross validation



- The classifier with least number of errors, or highest accuracy, is superior.
- Ex: BSS/WSS + SVM performs best due to its accuracy of 99%.

Prediction method (algorithm)	#errors Leukemia2 (/34)	# errors Leukemia3 (/34)	Average accuracy
BMA+KNN	1	1	97%
BMA+DT	3	5	88%
BMA+LR	1	4	93%
BMA+NB	4	3	89%
BMA+NN	1	2	96%
BMA+SVM	1	2	96%
BSS/WSS +KNN	1	1	97%
BSS/WSS +DT	3	3	91%
BSS/WSS +LR	2	4	91%
BSS/WSS +NB	3	5	88%
BSS/WSS +NN	2	1	96%
BSS/WSS +SVM	0	1	99%

Nesse exemplo, temos a combinação de feature selection e classification method

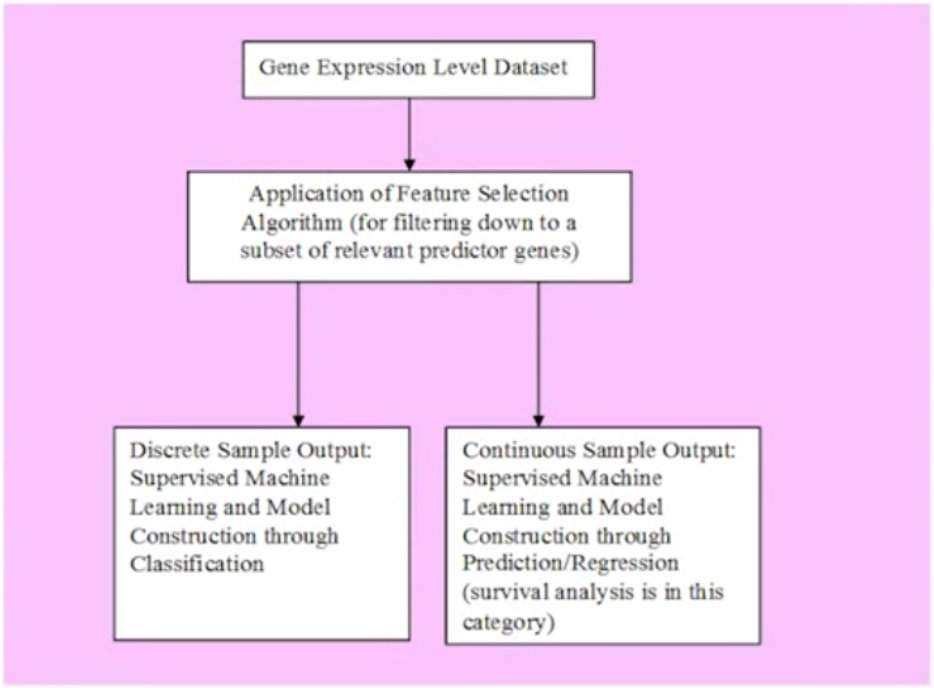
Prediction Methods

- In prediction problems, error is measured as the sum of differences between a predicted and given target value.
- There are several error measurements:
  - absolute mean,
  - root squared mean,
  - etc.

Combining Prediction Methods and Feature Selection

- In addition to improving the efficiency of classification / prediction models by reducing the number of features, feature selection generally improves the models' performance.
- Therefore feature selection is often a prerequisite step to classification / prediction, particularly in high dimensional domains such as bioinformatics.

Prediction Workflow



Isabelle Bichindaritz, SUNY Oswego

A combinação de feature selection e classification methods aumenta a acurácia.

Prediction Workflow

Algorithm	# errors Leukemia2 (/34)	Average accuracy
KNNV	2	94%
DT	3	91%
LR	1	97%
NB	1	97%
NN	-	-
SVM	2	94%

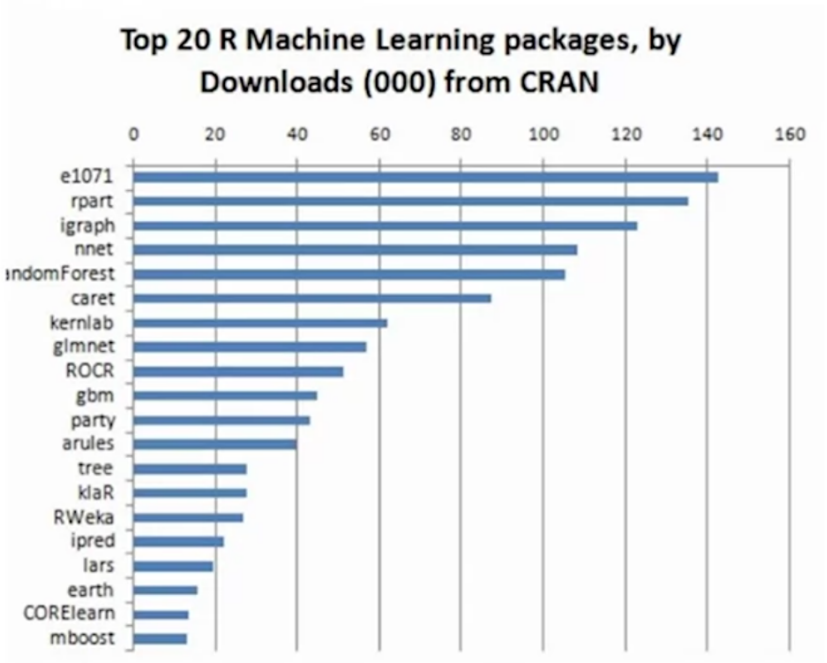
Prediction method (algorithm)	#errors Leukemia2 (/34)	# errors Leukemia3 (/34)	Average accuracy
BMA+KNN	1	1	97%
BMA+DT	3	5	88%
BMA+LR	1	4	93%
BMA+NB	4	3	89%
BMA+NN	1	2	96%
BMA+SVM	1	2	96%
BSS/WSS +KNN	1	1	97%
BSS/WSS +DT	3	3	91%
BSS/WSS +LR	2	4	91%
BSS/WSS +NB	3	5	88%
BSS/WSS +NN	2	1	96%
BSS/WSS +SVM	0	1	99%

Left: 3226 features

Right: feature selection of 16-20 features

Best accuracy for BSS/WSS feature selection + SVM → 99% (Bichindaritz 2010).

R packages



https://www.bioconductor.org/packages/release/bioc/vignettes/ReactomePA/inst/doc/ReactomePA.html
https://bioconductor.org/packages/release/bioc/html/ReactomePA.html
https://cran.r-project.org/web/packages/cluster/index.html

Exercício Prático - Predicting diseases from genes using RandomForest, KNN

https://radacad.com/prediction-via-knn-k-nearest-neighbours-r-codes-part-2

O exercício usa o mesmo dataset do módulo anterior.

Para instalar bibliotecas:
install.packages("randomForest")
install.packages("class")

Unsupervised

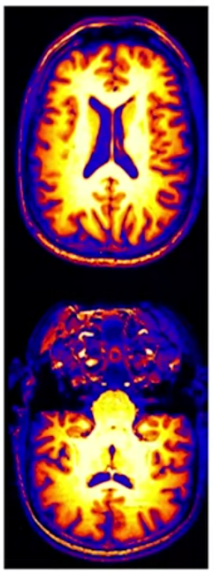
1. Find clusters in biomedical data involving genes.

2. Analyze and visualize biological pathways.

2. Write R scripts for clustering and for pathway analysis.

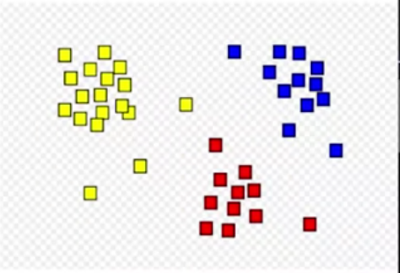
Clustering

Clustering Overview



- A cluster is a group of data within a dataset.
- Clustering analysis is closely related to pattern recognition.
- Clustering aims at finding clusters in which the objects are
  - As similar as possible to other objects in the same cluster.
  - As different as possible to objects in the other clusters.
- Clustering is an unsupervised data analytics task because the clusters are not known in advance – they have to be discovered automatically.

- Clustering is a search problem aiming at finding homogeneous subgroups of objects, aiming at:
  - Maximizing the overall similarity of objects within each cluster – intra-class similarity.
  - Minimizing the overall similarity of objects between clusters – called inter-class similarity.
  - One way to calculate these similarities is to add individual distances in each category after squaring them.



- Clustering performance measures:
  - Understandability.
  - Scalability.
  - Ability to discover clusters of any shape (not only circular).
  - Ability to cluster heterogeneous data (of different types).
  - Robustness to noise.
  - Robustness to outliers.
  - etc.

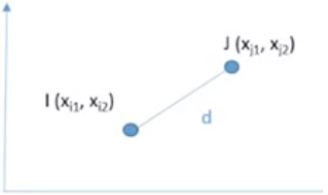
- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
  - Typical method: KMeans.
- Density-based approach:
  - Based on connectivity and density functions
  - Typical method: DBSCAN.
- Hierarchical approach:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical method: DIANA.

SImilaridade (Distância Euclidiana, Coeficiente de Jaccard e outras)

- Similarity measures the opposite of distance:
  high similarity = low distance
  low similarity = high distance.
- Normalization is usually applied before calculating similarity so that variables are within the same scale.
- Distance between:
  - numeric variables
  - nominal variables
  - ordinal variables
  - mixed types of variables.

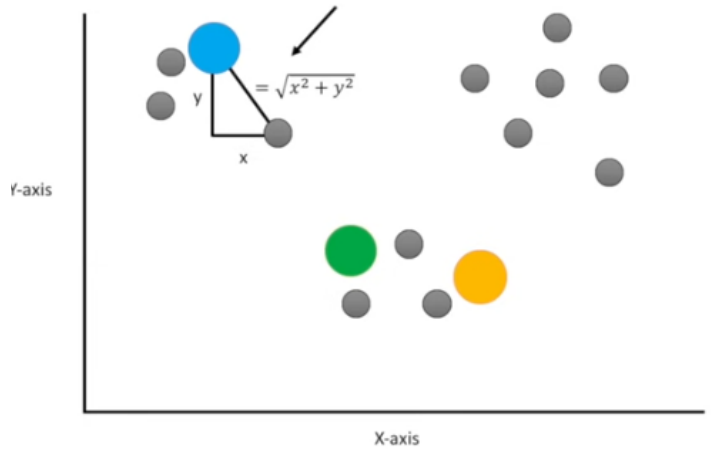
- Distance between numeric variables:
  - Euclidean distance $$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$
  - Any distance function can be used
    (a distance is a mathematical function with following properties:
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$ *(Pythagorean theorem)*

).



- We can also use correlation coefficient or any other measure of dissimilarity.

And we use the Euclidean distance. In 2 dimensions, the Eucli distance is the same thing as the Pythagorean theorem.



Note: We don't actually need to plot the data in order to cluster it. We just need to calculate the distances between things.

When we have 2 samples, or 2 axes, the Euclidean distance is:

$$\sqrt{x^2 + y^2}$$

When we have 3 samples, or 3 axes, the Euclidean distance is:

$$\sqrt{x^2 + y^2 + z^2}$$

When we have 4 samples, or 4 axes, the Euclidean distance is:

$$\sqrt{x^2 + y^2 + z^2 + a^2}$$

- Distance between nominal variables:
  - For binary variables:

    if symmetric binary variables

$$d(i, j) = \frac{b + c}{a + b + c + d}$$

Object *i* \ Object *j*	1	0	*sum*
1	*a*	*b*	*a+b*
0	*c*	*d*	*c+d*
sum	*a+c*	*b+d*	*p*

    if asymmetric binary variables (Jaccard coefficient)

$$sim_{Jaccard}(i, j) = \frac{a}{a + b + c} = 1 - \frac{b + c}{a + b + c}$$

  - For non-binary, d(I, j) = #mismatches / #variables

- Distance between ordinal variables:
  - Ordinal variables can be processed as
    - first map each value into a number: a → 1, b → 2, etc.
    - then calculate distance as for a numeric variable (1 < 2 < 3 …).
      - Distance between mixed types variables:
        - Add all the respective distances according to the types present in the object.
        - Ultimately a final distance is calculated over all the objects.
      - Distance between mixed types variables:
        - Add all the respective distances according to the types present in the object.
        - Ultimately a final distance is calculated over all the objects.

Partitioning (K-Means)

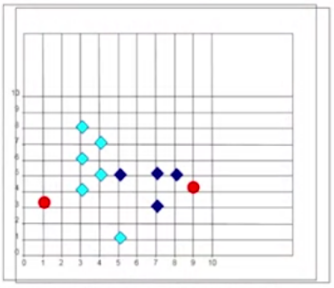
Algorithm 1 *k*-means algorithm

1: Specify the number k of clusters to assign.
2: Randomly initialize k centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** The centroid positions do not change.

The initialization of the centroids is an important step. It also highlights the use of SSE as a measure of clustering performance. After choosing a number of clusters and the initial centroids, the expectation-maximization step is repeated until the centroid positions reach convergence and are unchanged. Researchers commonly run several initializations of the entire *k*-means algorithm and choose the cluster assignments from the initialization with the lowest SSE.

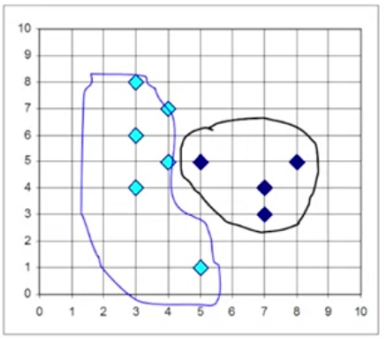
https://realpython.com/k-means-clustering-python/
https://www.youtube.com/watch?v=4b5d3muPQmA

- Algorithm proceeds in several steps.
  Suppose that we want to create 2 clusters.
- Step 1:
  We first create two separate clusters and assign arbitrarily a center object (red point) – called centroid.

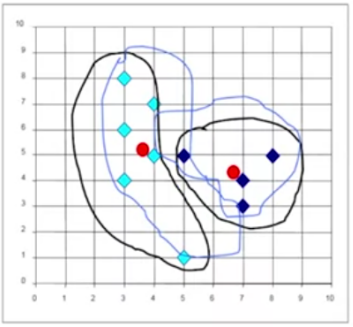


From Han and Kamber

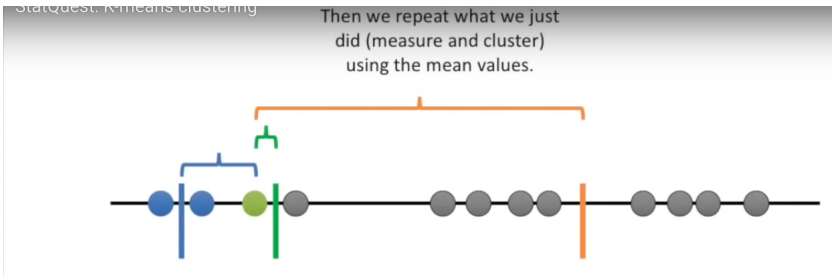
- Step 2:
  We assign each object to the cluster formed around each centroid.



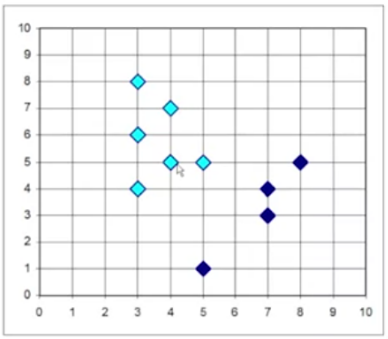
- Repeat Step 1:
  We calculate the centroid in each cluster.

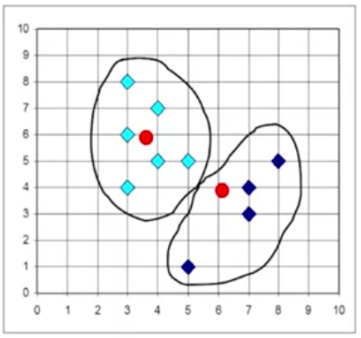


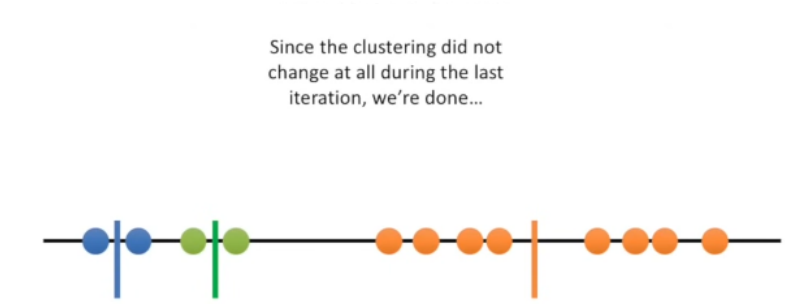
To do this, calculate the mean in each cluster and the means will be the next centroids.



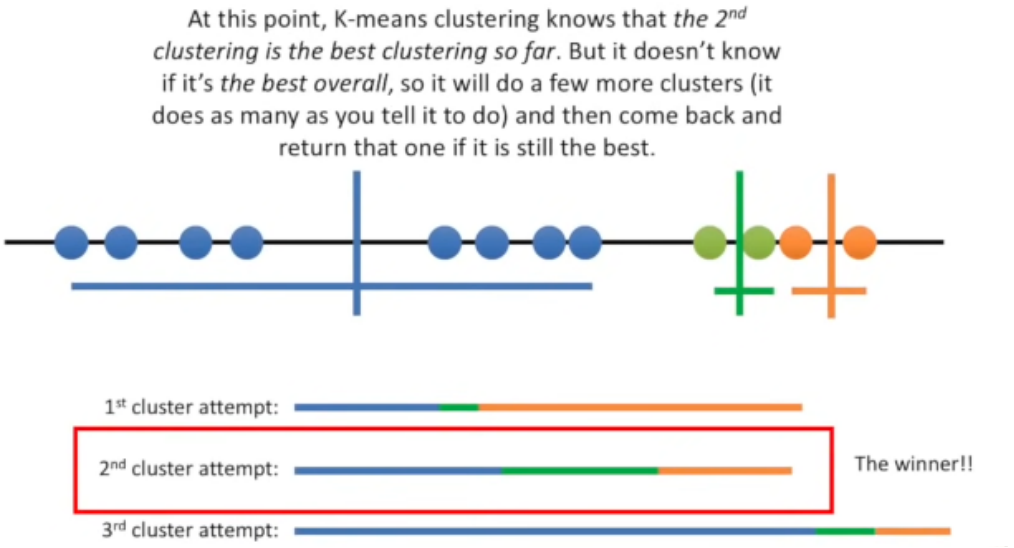
- Repeat Step 2:
  We assign each object to the cluster formed around each centroid. We continue until the clusters do not change.



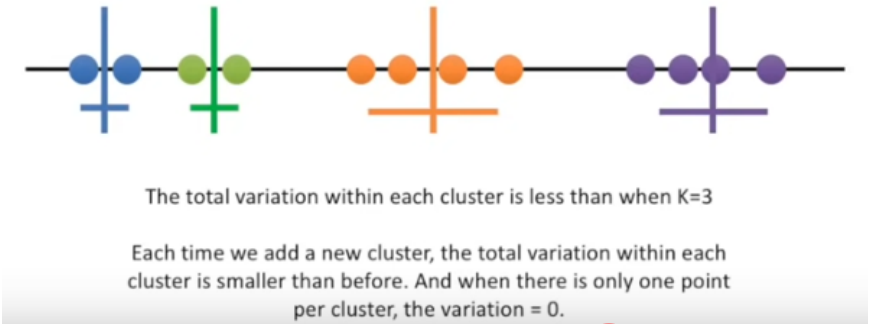


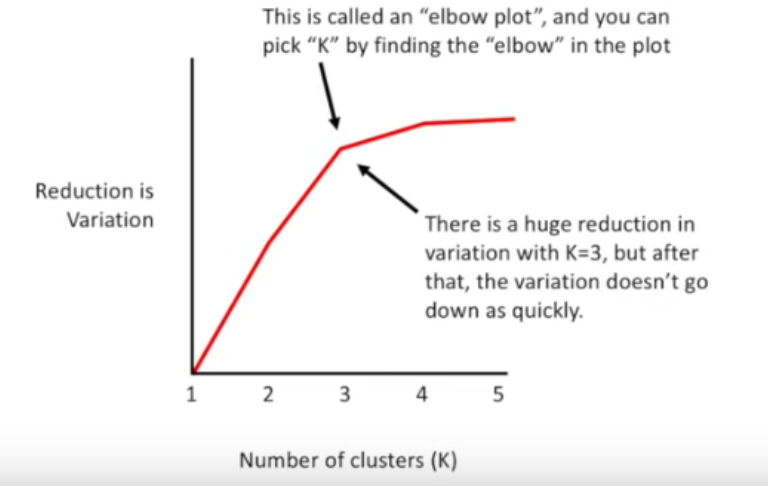


The best cluster attempt is that with the lowest variation:

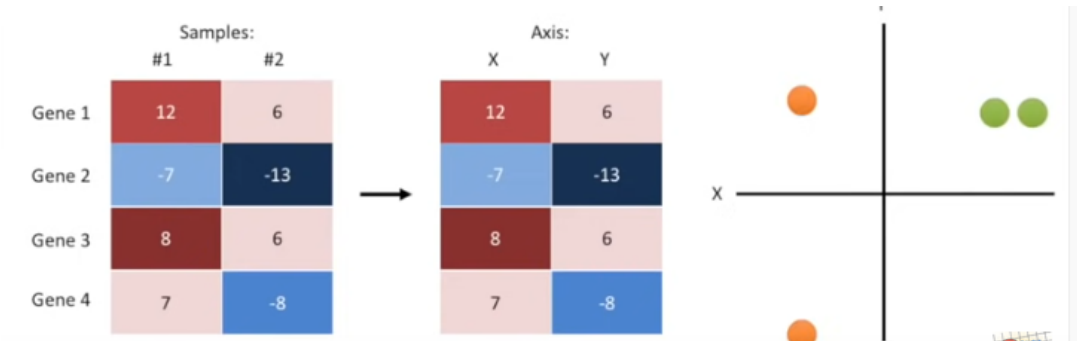


How to assess the best K value:





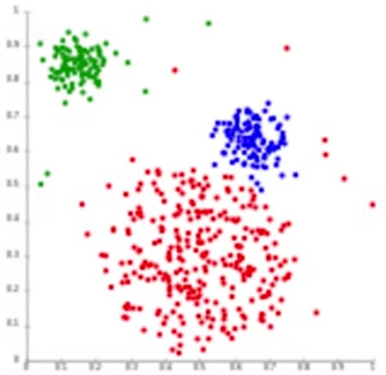
Uma ideia de aplicação:

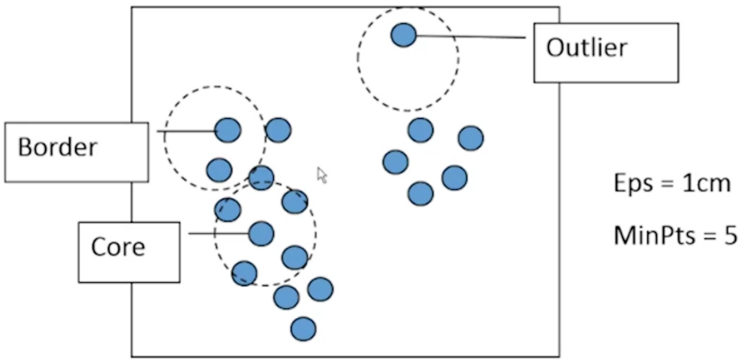


Density-based

Não é preciso saber de antemão a quantidade de clusters.
Identifica clusters de qualquer forma e não é sensível a outliers.

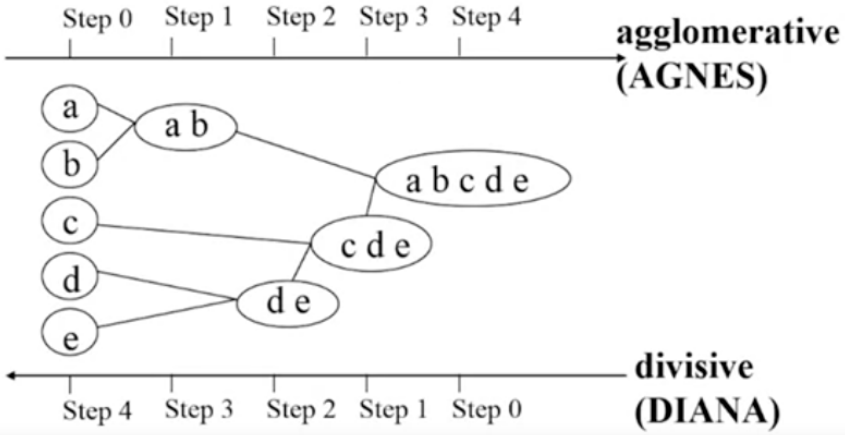
- Density-based clustering forms clusters according to the density of objects in regions.
- DBSCAN is the most famous example.
- Advantage:
  - It looks for clusters of any shape.
  - It is little sensitive to outliers or noise.
  - It is efficient.





Hierarchical

- This method can find any number of clusters and organizes them in a tree with sub-clusters.
- The tree is called a dendrogram.
- It is the representation of evolutionary trees.
- There are two main methods – divisive and agglomerative.



Pathway Analysis

- Ultimately the goal is to move from biological pathways to gene regulatory networks.
- Gene transcription depends on
  - Trans regulatory circuits:
    - The trans-regulatory circuit is considered as the regulatory interactions between upstream regulatory genes and transcription factor binding site motifs or cis elements → transcription factors.
  - Cis-regulatory circuit:
    - The cis-regulatory circuit can be viewed as a dynamic interactive circuit among binding site motifs with their effective action on the expression scheme of the target gene → binding site motifs.
  - Another mechanism is to add/remove genes to/from a network to change its behavior.

- Next step in genomics: assemble functional components of genomes and cells into the circuits that transform signals into cellular responses.

- Two approaches:
  - Small-scale approaches develop detailed quantitative models of the regulatory circuits controlling one or very few genes.
  - Large-scale approaches use computational algorithms to reconstruct genome-scale circuits.

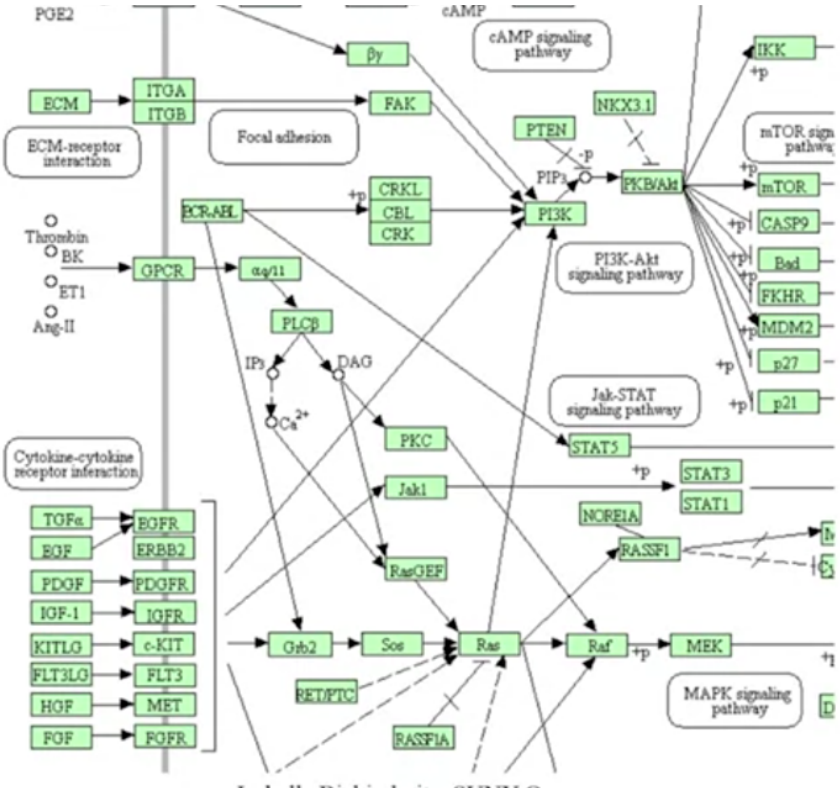
- They are likely to merge to produce genome-wide models.

- A large number of pathways is already known and can be retrieved from databases:

  - KEGG (Kyoto Encyclopedia of Genes and Genomes) has known 494 pathways.

  - It has both healthy pathways and disease-specific pathways.

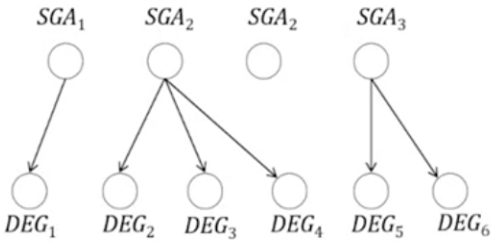
  - Fee-based however the previous version is freely available.



pathways in cancer (www.genome.jp/kegg

- Pathway enrichment uses known pathways from a pathway database and links them to differentially expressed genes.
  - Pathway databases such as KEGG, REACTOME, BIOCARTA etc. are used to map differentially expressed genes.
  - The association between these genes and pathways is quantified to statistically determine which pathways are significantly involved.
  - Main methods involved:
    - Causal discovery.
    - De-convolution methods.
    - Metabolomics.

  - Graphical models support causal discovery and the creation of models linking gene alterations, gene expression data, and other types of data (clinical, environmental, metabolomics etc.).
  - Ex: Tetrad (http://www.phil.cmu.edu/tetrad/).



- Cytoscape (http://www.cytoscape.org/).

Isabelle Bichindaritz, SUNY Oswego

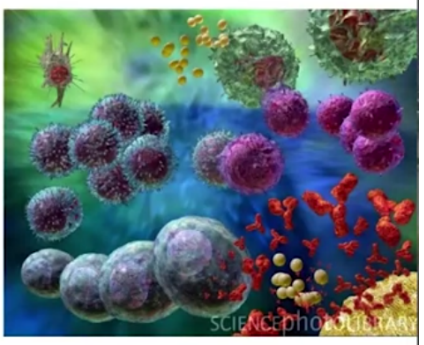
Alterações genômicas

As principais alterações são mutações somáticas, variações no número de cópia e metilação.

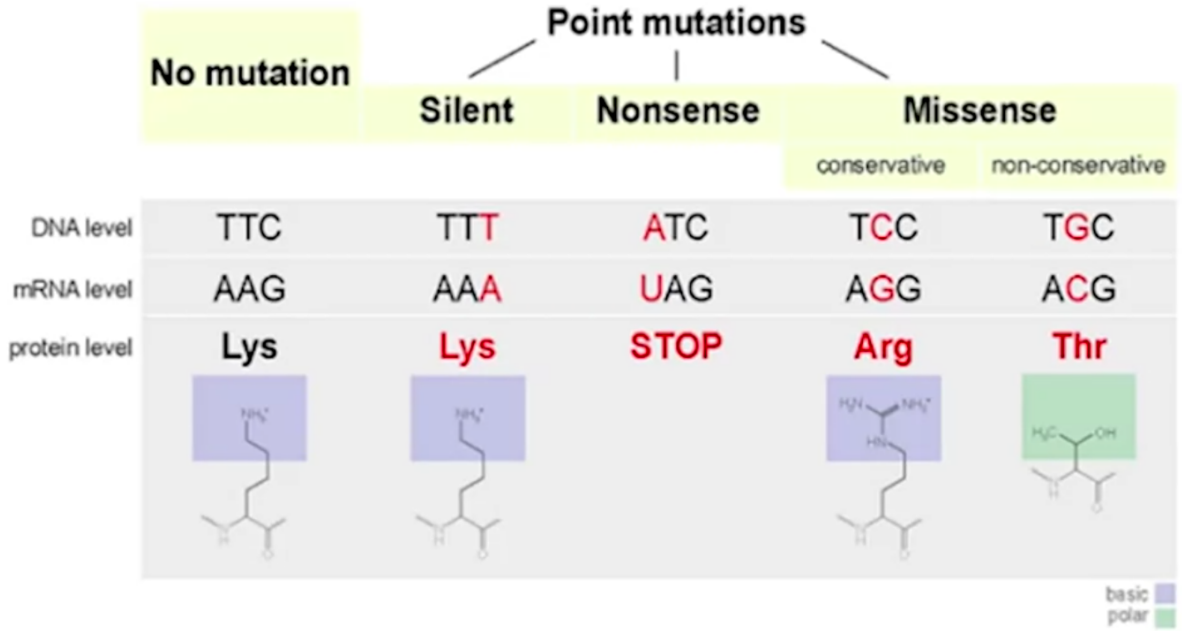
Overview of Gene Alterations

- Major gene alterations reported in the DNA are:
  - Gene mutations refers to changes in the DNA sequence of a gene. Most are harmless and single base changes (point mutations).
  - Chromosomal instability refers to the loss of portions of chromosomes , or rearrangements of part of chromosome such as inversion, translocation, deletion, or insertion.
  - Copy number variation refers to sections of the genome being repeated and the number of repeats varying between individuals in a population.
  - Gene amplification refers to duplication of genes in the DNA.
  - Methylation refers to the addition of methyl groups on some nucleotides in the DNA.

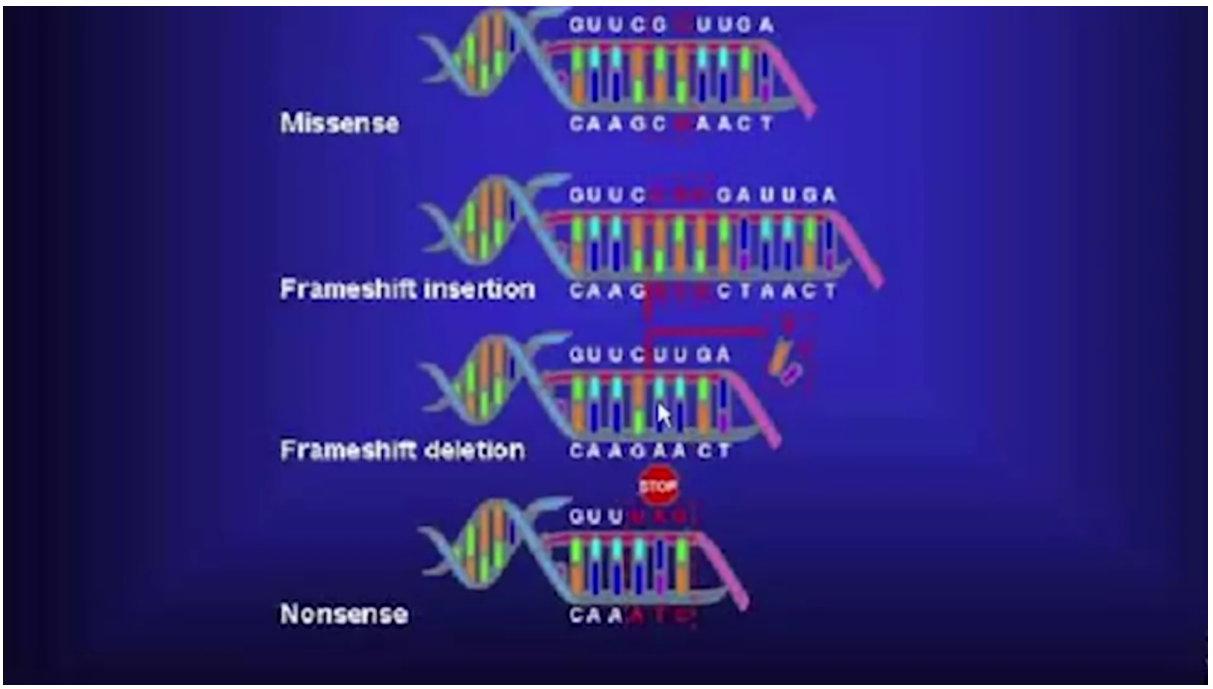
- The effects of gene alterations may be:
  - Overexpression or underexpression of the corresponding gene.
    - Ex: gene amplification generates an overexpression of the corresponding mRNA.
  - Translation into new forms of proteins, which may lead more easily to a disease, or less so through the modification of their original function and regulation.
    - Ex: ABL-BCR protein is a different form of ABL; ABL is an oncogene, normally regulated, however ABL-BCR is not regulated in the same way, leading to an oncogene.
  - Suppression of a protein.



Mutations



Conservative: another aminoacid, but with same properties
Non-conservative: another aminoacid that has different properties

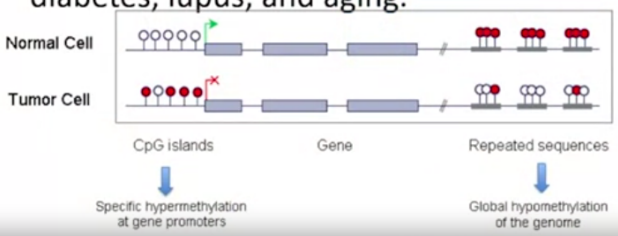


- Human beings have 2 copies of the same chromosome – one from their father and one from their mother – except for Y chromosome. They are called diploid.
- However a particular gene may not exist in the same form, called allele, in both chromosomes – these individuals are said to be heterozygous for the gene.
- Somatic mutations generally affect only one version of a gene.
- Many common diseases originate in a mutation in a single gene passed on to future generations.
- Ex: sickle-cell anemia is caused by a single missense mutation at codon 6 of the β-globin gene and affects 1 in 500 individuals of African descent.

Methilation

Methylation

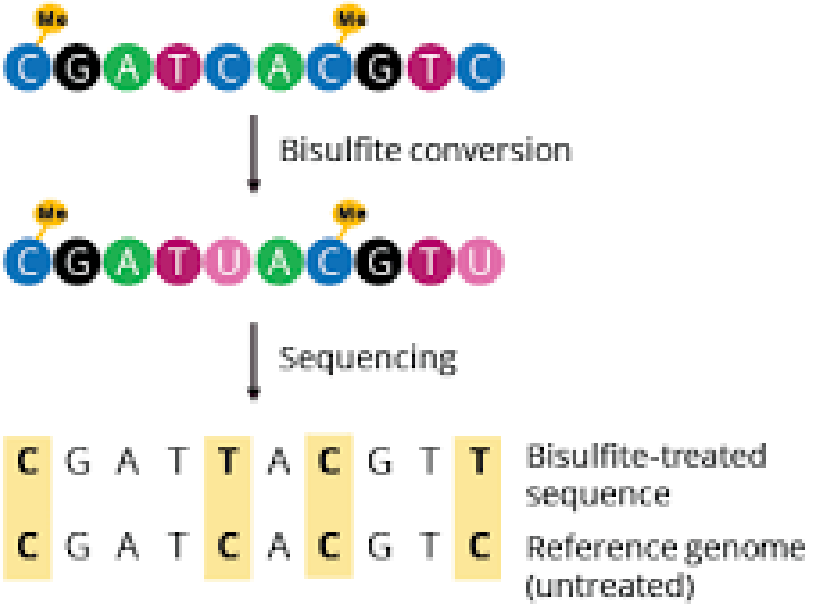
- Methylation is the addition of methyl groups (-CH3) on nucleotides in the DNA. It affects only C and A nucleotides.
- Methylation has for effect to repress gene transcription and plays an important role in normal cell functioning.
- However a change in the methylation process has been linked to many diseases, including cancer, schizophrenia, lupus, heart disease, type II diabetes, lupus, and aging.

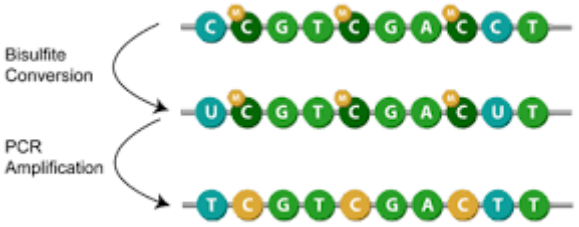


From http://atlasgeneticsoncology.org/

- DNA methylation is critical for gene expression such that a specific methylation signature can be associated with diseases in a way similar to gene expression signatures (differentially expressed genes).
- DNA methylation plays a role in the silencing of tumor suppressor genes in cancer and can help in cancer staging.
- Whole genome sequencing data (methylome-seq) give us the ability to classify different subtypes of cancer based on methylation biomarkers and are more stable than RNA or proteins.
- Methylation is linked to epigenetic modifications – those that are not part of the genetic material in chromosomes however affect it and can be passed onto offspring.

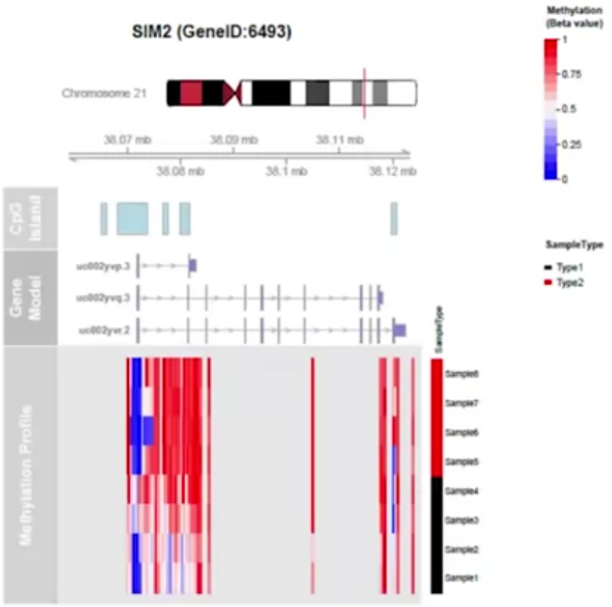
- Methylated genes are determined by analyzing bisulfite sequencing (BS-seq) data.
- TCGA provides a number of methylation data for cancer, coming mainly from two methylation platforms:
  - Methylation450 (higher resolution and greater sample count), and
  - Methylation27.
- Data can be analyzed to compare methylation between groups or between patients and normal.





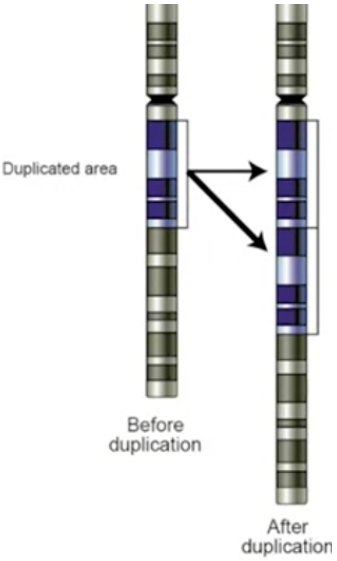
Finding Methylated Genes

Methylation heatmap for chromosome 21 (methyAnalysis)



Copy Number Variation

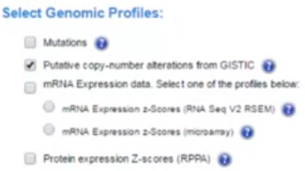
- Copy number alterations refer to changes in the number of copies:
  - Gene amplification is when a gene has multiple copies resulting from aberrant DNA replication. It is also called gene duplication (in principle, any number of the same gene greater than two in humans and diploid species is an amplification). The number of copies can be very large and has clinical impact.
  - Gene copy number gain is when the number of copies of a gene is greater than 2 and is present for example in cancer cells (low gain, high gain). Gain is considered an amplification with few copies generated.
  - Gene deletion is when a gene or part of a gene is deleted from a chromosome:
    - Deep deletion or deletion below minimum median chromosomal arm copy number for that sample by at least 0.1.
    - Shallow deletion or deletion from 1.9 to deep deletion threshold..
    - Homozygous deletion or deletion of a gene on a single chromosome in considered a deep loss.



From Human Genome Research Institute.

Finding Copy Number Alterations

- cBioPortal may be used to study copy number alterations in cancer patients under the Oncoprint tab after option: Putative copy-number alterations from GISTIC.
- Example: EGFR ERBB2 ERBB3 ERBB4 in Breast cancer shows amplification and deep deletion.



33

- Software to identify copy number alterations include:
  - GISTIC (from Broad Institute, Linux-based).
  - Many others (https://omictools.com/cnv-detection2-category).
- TCGA provides files already processed by GISTIC and ready for further analysis and quantification.

Genomic Alterations and Gene Expression

- There is a lot of evidence that genomic alterations are associated with dysregulated genes (upregulated or downregulated).
- cBioPortal provides a tool of choice to observe this phenomenon on cancer data.
- Even though both disturbances – gene alterations and gene expression modifications – are observed in many diseases, showing the causal effect is more difficult than showing that there is an association. Several clinical experiments have been designed to demonstrate this.