Data mining in a few lines

Fülöp Bazsó
Wigner RCP, Institute for Particle and Nuclear Physics
SOTE Neuroinformatics course

19.-th and 26.-th,- of September, 2018.

## 1 Setting the scene

# Data ≠ information, knowledge!

John Naisbitt:
We are drowning in information but starving for knowledge.

Paradoxically, it is difficult to infer or make conclusion from large datasets.

Gaining information should be made objective, at the same time because of complex nature and amount of data information gain should be automatised.

Several reasons why data may be abundant:

- Experimental techniques are advanced, we can make more measurements for a lower price.
- Data storage price is lower, data availability is easier, we have faster access to data.
- Data processing price is lower.
- Within informatics and related areas new sciences emerged, data science, data mining, machine learning, etc. so we are not "afraid" of abundant data.

Data *per se* are not useful.
Several possible definitions of data mining:

- The nontrivial extraction of implicit, previously unknown, and potentially useful information from data. (Piatetsky-Shapiro)
- The automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large database, data ware houses, the Web ... or data streams. (Han)
- ... the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meningful. (... finding hidden information in a database.) (Dunham)
- ... the process of employing one or more computer learning techniques to automatically analyse and extract knowledge from data contained in a database. (Roiger)

## 2 Generalities

Data mining is multidisciplinary, it includes following fields (and other):

- Mathematics,
- Statistics,
- Probability theory,
- Machine learning,
- Informatics,
- Graph- and Network theory.

Data mining involves six common classes of tasks:

- Clustering is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. Number of clusters may be unknown in advance, sometimes we prescribe it.
- Classification is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam". Clusters are already known, and we update the database with new data.
- Anomaly detection (outlier/change/deviation detection) The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (dependency modelling) is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.
- Detecting ,,frequent" patterns or regularities.

- Regression  attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.
- Summarization  providing a more compact representation of the data set, including visualization and report generation.

Beware!
Do not be fooled with data analysis, perform sanity checks.

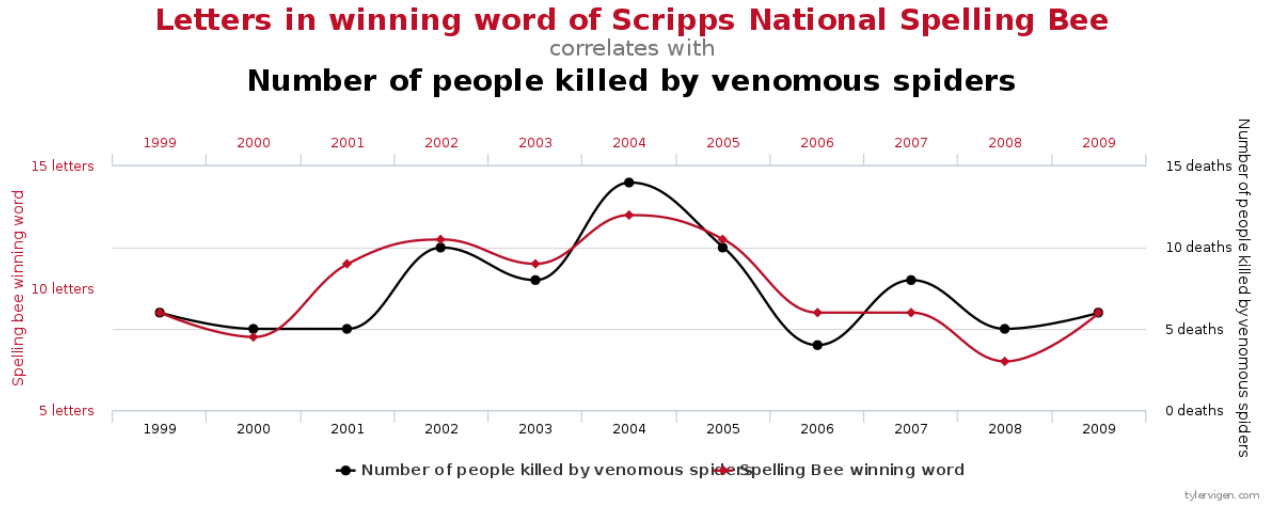



figure source: http://www.tylervigen.com/spurious-correlations:
Check the page for many other insane correlations.

Table 1: Anscombe's quartet properties

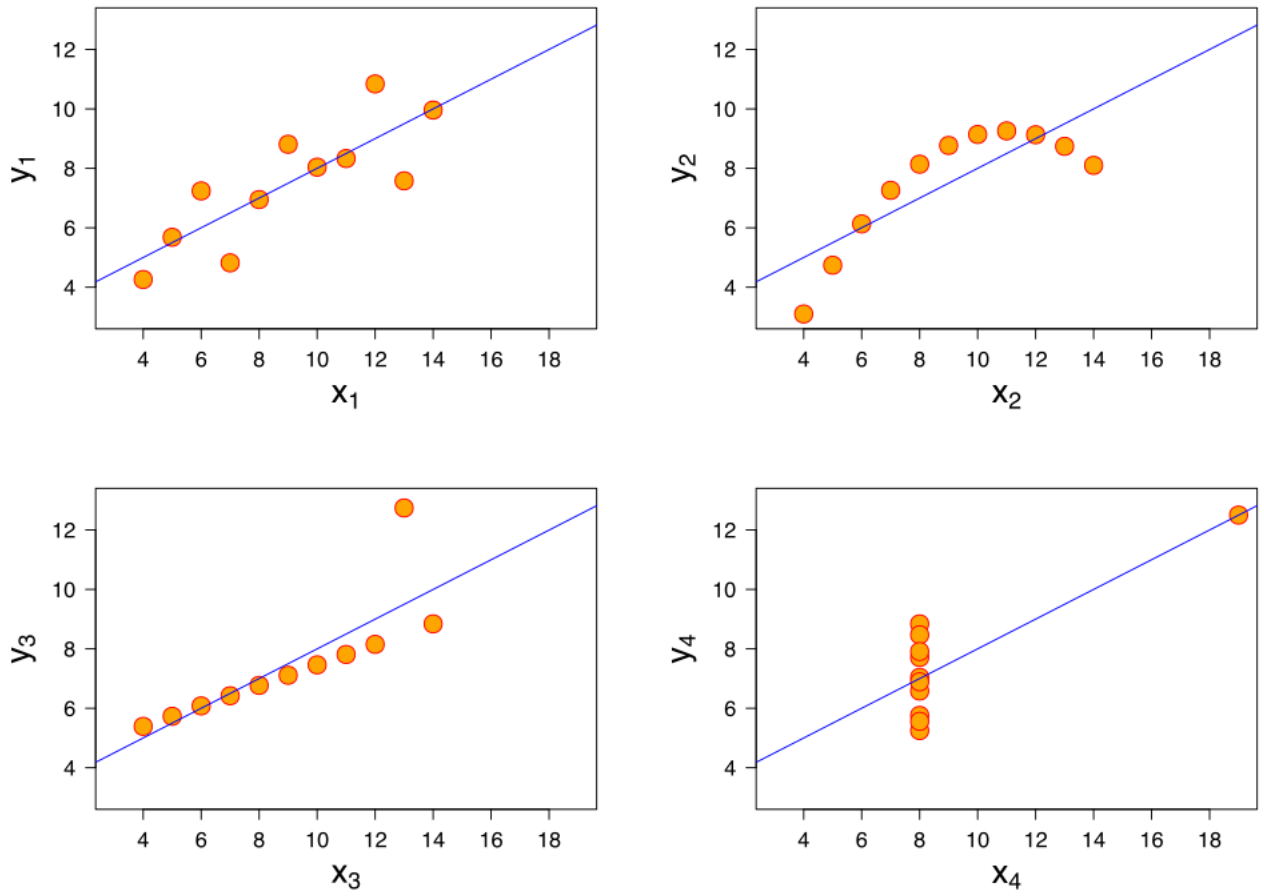
| | | |
|---|---|---|
| Mean of x | 9 | exact |
| Sample variance of x | 11 | exact |
| Mean of y | 7.50 | to 2 decimal places |
| Sample variance of y | 4.125 | ±0.003 |
| Correlation between x and y | 0.816 | to 3 decimal places |
| Linear regression line | y = 3.00 + 0.500x | to 2 and 3 decimal places, respectively |
| Coefficient of determination of the linear regression | 0.67 | to 2 decimal places |

Anscombe's quartet is an example when totaly different datasets have same statistics.

Read more at wikipedia: https://en.wikipedia.org/wiki/Anscombe's_quartet

Four datasets are given as respepctive columns in the table below.

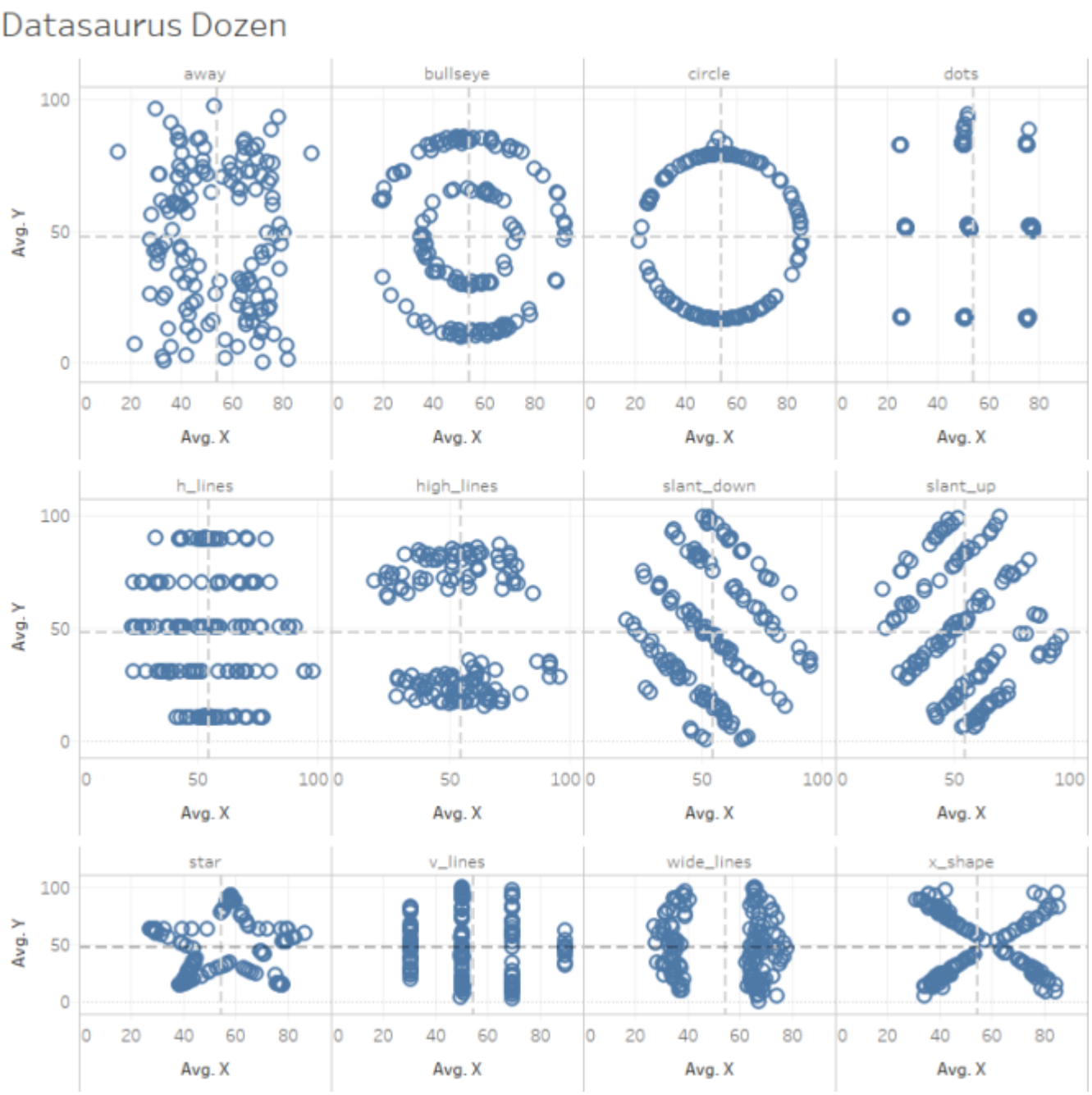
| x | y | x | y | x | y | x | y |
|---|---|---|---|---|---|---|---|
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |




Data processing does include sanity checks.

Further examples can be seen here:

https://dabblingwithdata.wordpress.com/2017/05/03/the-datasaurus-a-monstrous-anscombe-for-the-21st-century/


Datasaurus Dozen
away
bullseye
circle
dots
h_lines
high_lines
slant_down
slant_up
star
v_lines
wide_lines
x_shape
Avg. Y
Avg. X
0
20
40
60
80
50
100

Based on https://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd

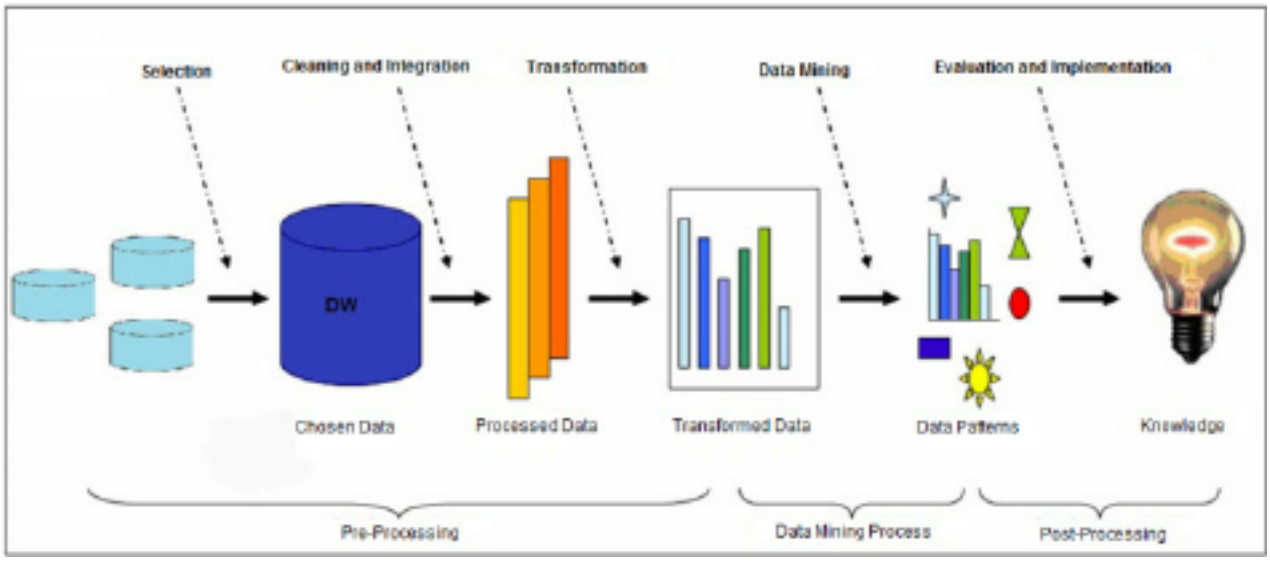
KDD (knowledge discovery in a database) includes multidisciplinary activities. This encompasses data storage and access, scaling algorithms to massive data sets and interpreting results. The data cleansing and data access process included in data warehousing facilitate the KDD process. Artificial intelligence also supports KDD by discovering empirical laws from experimentation and observations. The patterns recognized in the data must be valid on new data, and possess some degree of certainty. These patterns are considered new knowledge. Steps involved in the entire KDD process are:

1. Identify the goal of the KDD process from the customers perspective.
2. Understand application domains involved and the knowledge that's required
3. Select a target data set or subset of data samples on which discovery is be performed.
4. Cleanse and preprocess data by deciding strategies to handle missing fields and alter the data as per the requirements.
5. Simplify the data sets by removing unwanted variables. Then, analyze useful features that can be used to represent the data, depending on the goal or task.
6. Match KDD goals with data mining methods to suggest hidden patterns.
7. Choose data mining algorithms to discover hidden patterns. This process includes deciding which models and parameters might be appropriate for the overall KDD process.
8. Search for patterns of interest in a particular representational form, which include classification rules or trees, regression and clustering.
9. Interpret essential knowledge from the mined patterns.
10. Use the knowledge and incorporate it into another system for further action.

11. Validate your findnings.

12. Document it and make reports for interested parties.

Preprocessing (approx. first five steps) may consume up to 80% of your time and funds.



-what we expect from the knowledge gained:

- easily understandable – Knowledge gained should relate to the existing body of knowledge and interpretable by the experts – You need the aha feeling!
- valid – Often data is randomly separated to a "learning set" (typically 75-80% of the data) and a test set (20-25%). Information extracted from the learning set should be validated on the test set.
- useful. Otherwise who would care?
- novel. Otherwise why would you wish analyse the data?

Conditions-constraints of data mining applicability

- large datasets (up to a point, the more the better)
- many attributes (complex data, etc.)
- clean data, because GIGO: garbage in, garbage out
- unbiased data because BIBO: bias in, bias out
- the knowledge has to be applicable or interesting, (otherwise nothing happens)
- payoff (knowledge gained should materialise )

Based on https://en.wikipedia.org/wiki/Big_data

Big data is a term used to refer to the study and applications of data sets that are so big and complex that traditional data-processing application software are inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source. There are a number of concepts associated with big data: originally there were 3 concepts volume, variety, velocity. Other concepts later attributed with big data are veracity (i.e., how much noise is in the data) and value.

Lately, the term "big data" tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but thats not the most relevant characteristic of this new data ecosystem." Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on." Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet search, fintech, urban informatics, and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, connectomics, complex physics simulations, biology and environmental research.

In health and biology, conventional scientific approaches are based on experimentation. For these approaches, the limiting factor is the relevant data that can confirm or refute the initial hypothesis. A new postulate is accepted now in biosciences: the information provided by the data in huge volumes (omics) without prior hypothesis is complementary and sometimes necessary to conventional approaches based on experimentation. In the massive approaches it is the formulation of a relevant hypothesis to explain the data that is the limiting factor. The search logic is reversed and the limits of induction are to be considered.

## 3 Ethical and legal issues

Based on https://en.wikipedia.org/wiki/Big_data_ethics

Data collection involves many ethical and legal issues...

Check Big data ethics in wikipedisa, though what is written there applies to all data colletion, not necessarily big data.

In some cases consent is needed in order to collect, store and process data or make it available to third parties.

Data ownership must be clarified.

Who owns data? Ownership involves determining rights and duties over property. The concept of data ownership is linked to one's ability to exercise control over and limit the sharing of their own data. If one person records their observations on another person who owns those observations? The observer or the observed? What responsibilities do the observer and the observed have in relation to each other? Who owns a digital identity? European laws, the General Data Protection Regulation, indicate that individuals own their own personal data.

Examples of personal data include: Genome data, GPS location, written communication, spoken communication, lists of contacts, internet browsing habits, financial transactions, supermarket spending, tax payments, criminal record, laptop and mobile phone camera lens recording, device microphone recordings, driving habits via car trackers, mobile and health records, fitness activity, nutrition, substance use, heartbeat, sleep patterns and other vital signs. The collective of one individual's personal data forms a digital identity (or perhaps digital alter ego).

A key component of personal data ownership is unique and controlled access i.e. exclusivity. Ownership implies exclusivity, particularly with abstract concepts like ideas or data points. It is not enough to simply have a copy of your own data. Others should be restricted in their access to what is yours.

In regard to personal data, the individual has the right to know:

Why the data is being collected?

How it is going to be used?

How long it will be stored?

How it can be amended by the individual concerned?

If consent is given, the consent applies to ...

If an individual or legal entity would like to use personal data, one needs informed and explicitly expressed consent of what personal data moves to whom, when, and for what purpose from the owner of the data.

The permission needs to be given in a format which is explicit, not implied.

While a person could give consent on a general topic to be continuous, it should always be possible to retract that permission for future transactions.

If data transactions occur all reasonable effort needs to be made to preserve privacy.

The idea of open data is centred around the argument that data should be freely available and should not have restrictions that would prohibit its use, such as copyright laws.

Data are often anonimised, in order to comply with ethical regulations, especially with sensitive data.

Ethical approval of relevant body may be needed prior to the data collection or processing.

Different worldwide practices. e.g. EU vs US

## 4 Modeling

Understanding often relates to a model.

Good modell has large explanatory power and uses few assumptions.
We keep only the relelvant variables and disregard everything else, i.e. separate relevant from irrelevant information.
Thus:
Learning is inseparable from an information loss!
We get rid of the irrelevant information and keep only the relevant one.

We perform gold washing - keep the gold particles and throw the sand away.

## 5 Problem of model selection, information criteria

Occam's or Occham's raisor

*Pluralitas non est ponenda sine necessitate* or *Numquam ponenda est pluralitas sine necessitate"* Plurality must never be posited without necessity.

William of Ockham (circa 1287 - 1347)

If we use less assumptions we have larger generalisation ability.

One justification of Occam's razor is a direct result of basic probability theory. By definition, all assumptions introduce possibilities for error; if an assumption does not improve the accuracy of a theory, its only effect is to increase the probability that the overall theory is wrong.

Richness of a model is limited by the availability of the data.

Model should be as complex as possible, but not more, as we will run into the overfitting (and loose generalisation ability).

## 6 Model selection criteria

Once we have data and a list of candidate models we should pick one. Model selection should be objective. So we need and objectivity criterion which embodies the Occam rasor in one form or another.

We wil need the following notion:

Likelihood function $L$ (often simply the likelihood) is a function of the parameters of a statistical model, given specific observed data. Likelihood functions play a key role in frequentist inference, especially methods of estimating a parameter from a set of statistics.

Likelihood expresses how probable is the observation given a model.

### 6.1 Akakike's information criterion AIC

Let $k$ denote the number of parameters in the model, and $L$ the likelihood, then

$$AIC = 2k - 2\log \hat{L} \tag{1}$$

where

$$\hat{L} = p(x|\hat{\theta}, M) \tag{2}$$

Remark: The amount of data plays no role.

AIC with correction for the data size (the size matters):

$$\text{AICc} \;=\; \text{AIC} + \frac{2k^2 + 2k}{n - k - 1} \tag{3}$$

### 6.2 Bayesian information criterion BIC

$$BIC = \ln(n)k - 2\ln(\hat{L}). \tag{4}$$

$x$= the observed data;

$n$ = the number of data points in $x$, the number of observations, or equivalently,

the sample size;
$k$ = the number of parameters estimated by the model.

Here we observe that larger datasets may allow more detailed models.

### 6.3 MDL (minimal description lenght) principle, Rissanen

two step approach:

1. we count the number of bits needed to describe the data within the model
2. we count the number of bits neede to describe the model itself
3. we add the two and select a model for which the total amount of information measured in bits is minimal.

$$L(D) = \min_{H \in \mathcal{H}} (L(H) + L(D|H)) \tag{5}$$

Here $\mathcal{H}$ is the set containing all the hypostheses (models), $L(H)$ is the number of bits needed to describe a particular model chosen from $\mathcal{H}$, and $L(D|H)$ is the probability of data $D$ given a model $H$.

MDL is an embodiment of the Ocham's raisor principle, we accept the most succint explanation.

MDL: The Basic Idea

The goal of statistical inference may be cast as trying to find regularity in the data. Regularity may be identified with ability to compress. MDL combines these two insights by viewing learning as data compression: it tells us that, for a given set of hypotheses H and data set D, we should try to find the hypothesis or combination of hypotheses in H that compresses D most.

For practical purposes one choses a particular information criterion based on model assumptions, nature of the data, personal bias, etc.

The ,,final answer" of model selection is in Kolmogorov structure functions , but unfortunately they are not very practical, at least for the time being, though things can change in the near future.

https://en.wikipedia.org/wiki/Kolmogorov_structure_function

# 7 Clustering

see https://en.wikipedia.org/wiki/K-means_clustering

Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a $d$-dimensional real vector, k-means clustering aims to partition the $n$ observations into $k$ ($\leq n$) sets $S = S_1, S_2, \ldots, S_k$ so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance). Formally, the objective is to find:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i \tag{6}$$

where $\mu_i$ is the mean of points in $S_i$. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{\mathbf{x},\mathbf{y} \in S_i} \|\mathbf{x} - \mathbf{y}\|^2 \tag{7}$$

k-means allways works except when it doesn't ... Anyhow, it is the working horse in the whole field, big data included.

## 7.1 Standard algorithm

The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the k-means algorithm; it is also referred to as Lloyd's algorithm, particularly in the computer science community.

Given an initial set of k means $m_1^{(1)}, \ldots, m_k^{(1)}$, the algorithm proceeds by alternating between two steps:

Assignment step: Assign each observation to the cluster whose mean has the least squared Euclidean distance, this is intuitively the "nearest" mean. (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means).

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k \right\} \tag{8}$$

where each $x_p$ is assigned to exactly one $S^{(t)}$, even if it could be assigned to two or more of them.
Update step: Calculate the new means to be the centroids of the observations in the new clusters.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{9}$$

The algorithm has converged when the assignments no longer change. There is no guarantee that the optimum is found using this algorithm.

The algorithm is often presented as assigning objects to the nearest cluster by distance. Using a different distance function other than (squared) Euclidean distance may stop the algorithm from converging. Various modifications of k-means such as spherical k-means and k-medoids have been proposed to allow using other distance measures.

### 7.2 Initialization methods

Commonly used initialization methods are Forgy and Random Partition. The Forgy method randomly chooses k observations from the data set and uses these as the initial means. The Random Partition method first randomly assigns a cluster to each observation and then proceeds to the update step, thus computing the initial mean to be the centroid of the cluster's randomly assigned points. The Forgy method tends to spread the initial means out, while Random Partition places all of them close to the center of the data set.

## 8 (Relational) Data, Networks and Graphs

Many datasets are not numerical, rather relational. They are usually represented with graphs of networks.

We have entities (nodes) and relations between entities (edges).

Keep in mind some facts:

(binary) Relations need not be symmetric.

Sometimes relations can be quantified - we have edge weights.

Sometimes nodes can be quantified - we have node weights.

Relations need not be binary!

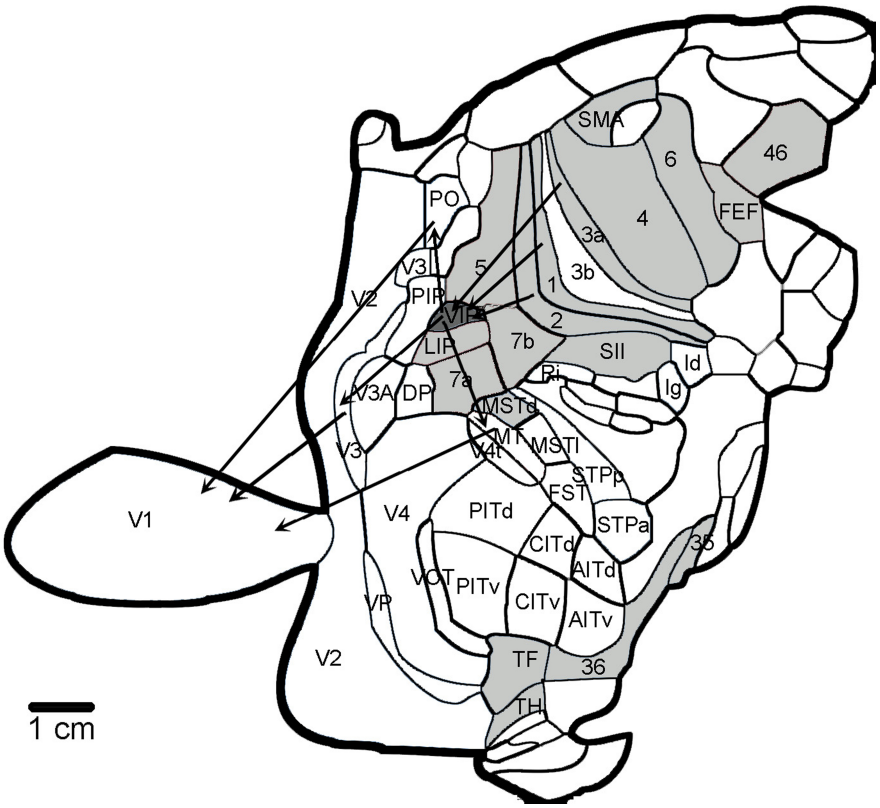
Real world graphs are sparse! We have far more nodes than edges in a graph.

Real world graphs are often poorly known!

## 9 Data representable as networks

Cortical connectivity data
macaque visuo-tactile cortex

Nodes are cortical areas, oriented edges are connections between areas.




Flattened hemisphere of the macaque visuo-tactile cortex.

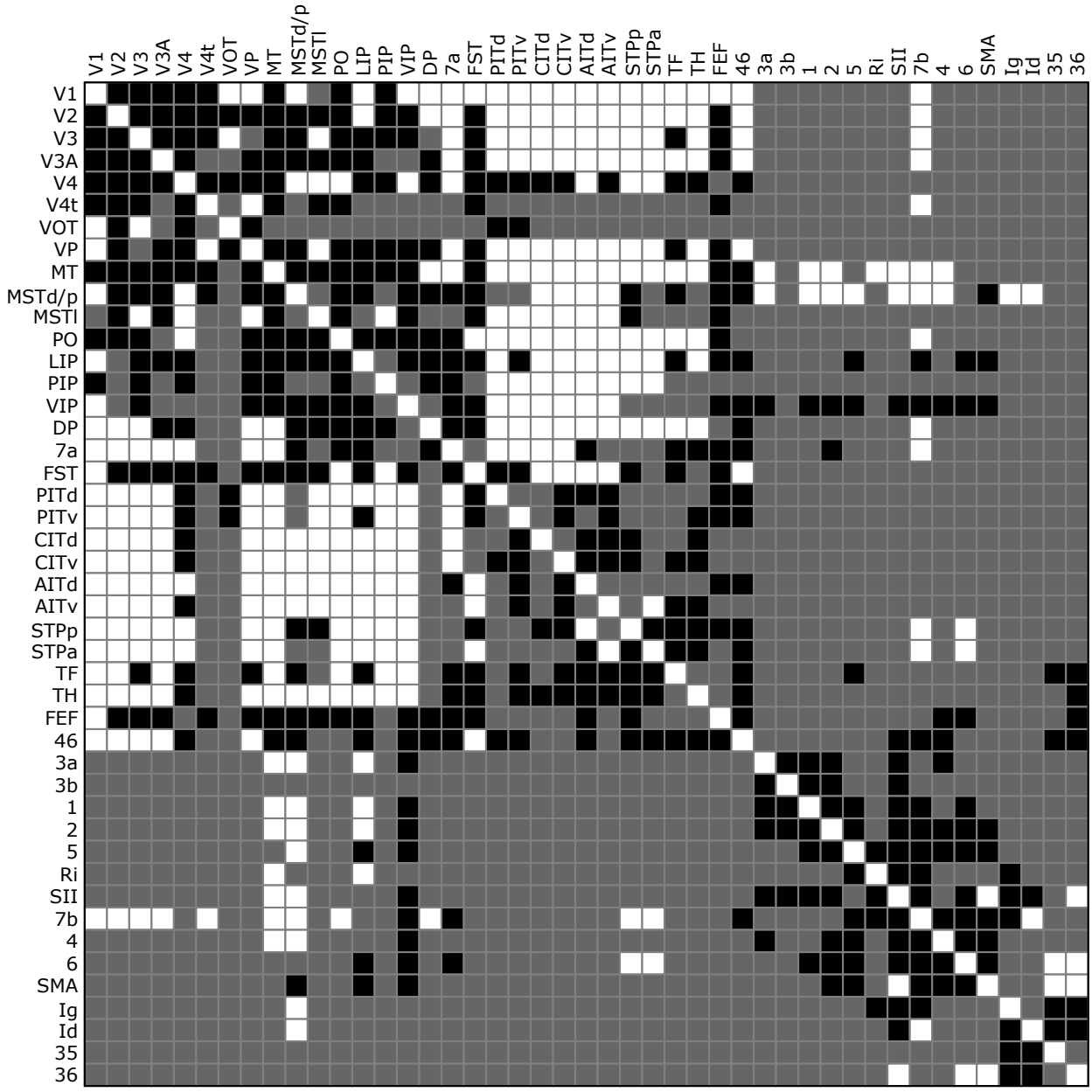
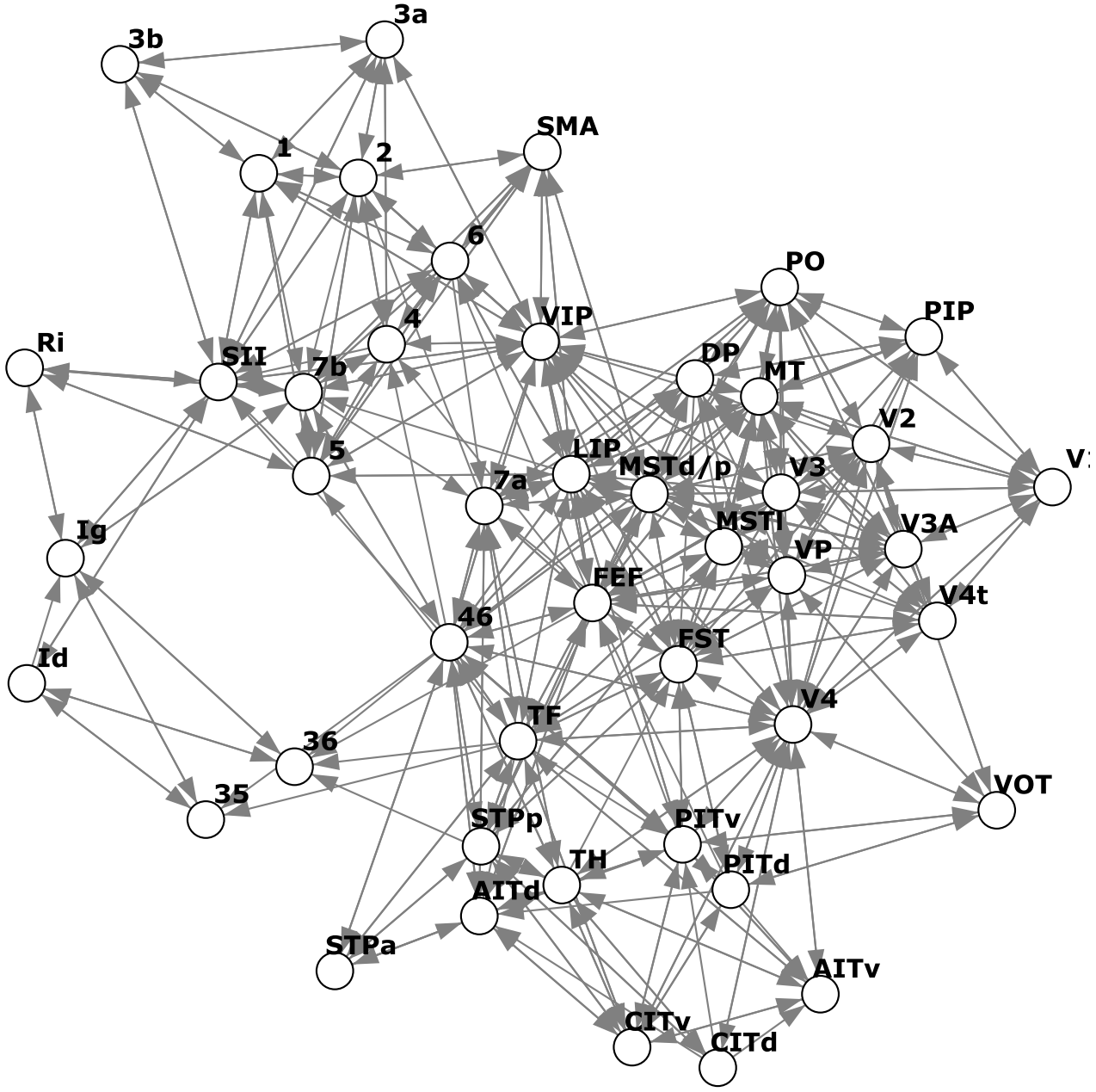
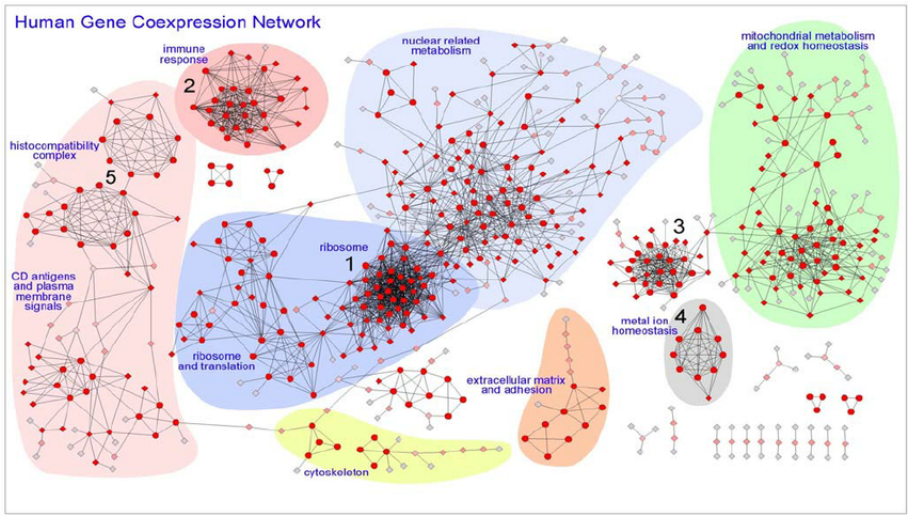



Table showing the status of an edge: black - an edge is there, white - there is no edge, gray - unknown.



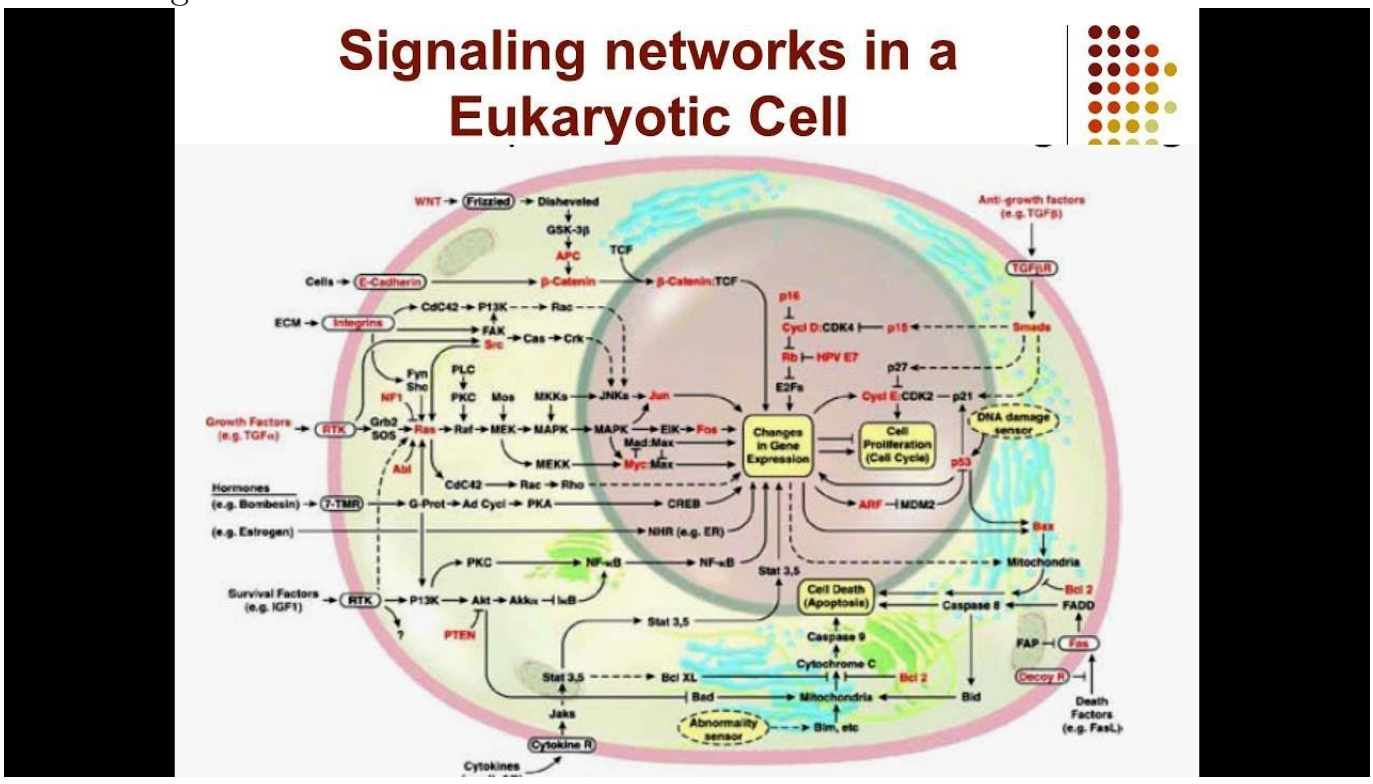

Result of node clustering.

Genne (co)expression data




http://bioinfow.dep.usal.es/coexpression/network.jpg

Cellular signal transduction network




https://www.youtube.com/watch?v=nMrKbRj_2-s

Nodes are reactants in the cellular methabolism. and there is an edge from A to B if A and B are on different sides in the equation describing a chemical reaction.

Other typical examples may be:

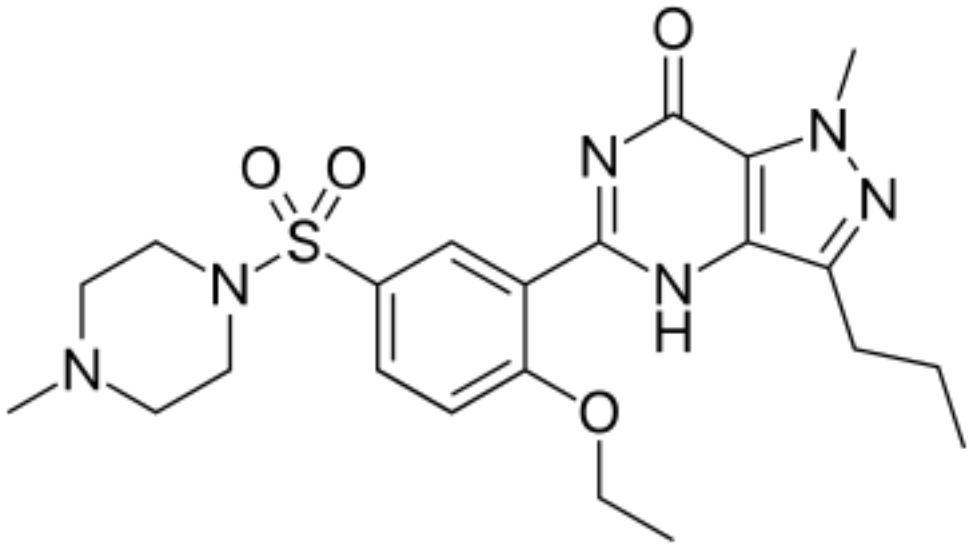
Social network data

Program flow data

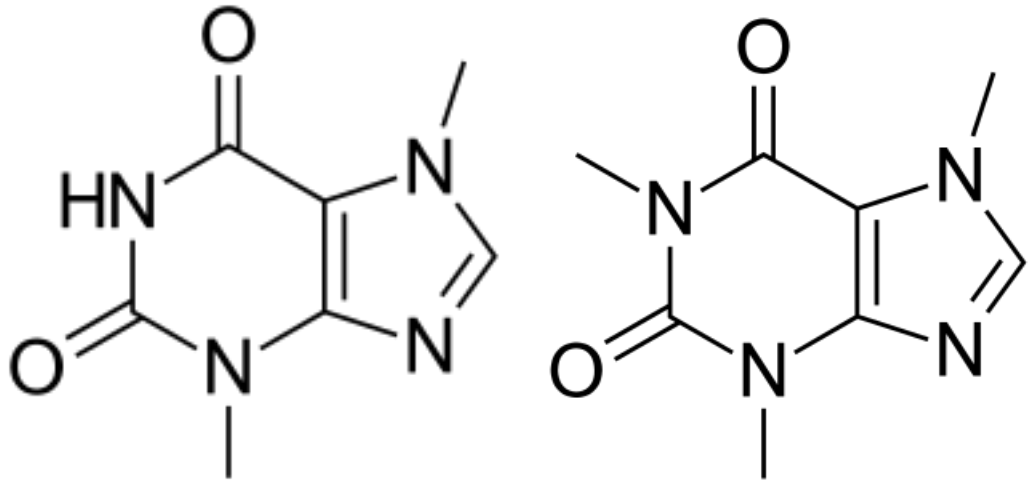
and many more

Some problems encountered in applications:

chemical compounds - finding frequent subgraphs

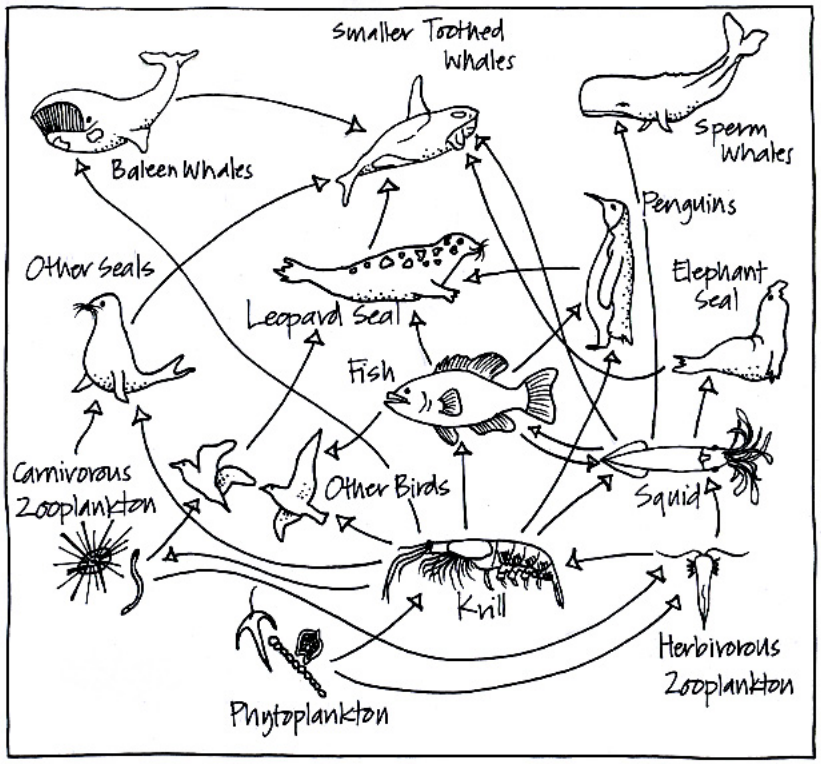



viagra, https://upload.wikimedia.org/wikipedia/commons/thumb/d/d3/Sildenafil_struct Sildenafil_structure.svg.png




caffeine
finding functional modules (in large molecules, say ...)
program control flow analysis
intrusion network analysis
mining commmunication network
anomaly detection
mining xml structures
etc

Some biomedical applications:
food webs




whale food web, http://albertcoward.co/wp-content/uploads/2018/08/schoolyard-food-webs-in-this-activity-students-explore-the-idea-chains-and-worksheets-ks2.jpg

mining biochemical patterns

finding conserved subnetworks

...

## 10 How to perform the network analysis?

First one should clarify whether the graph is connected, if not, determine the connectivity components.

In case of directed networks two types of connectivity are used: weak and strong, weak and strong, so connectivity components can be weak and strong.

Shortest paths are important as information transfere needs to be fast, therefore one wishes to minimise the number of steps in path connecting two nodes. Beware, the length on the shortest path is unique, the path itself need not be.

How to assign importance to a node based on the graph structure? One possible answer is betweenness.

Betweenness is a centrality measure of a vertex within a graph (there is also edge betweenness, which is not discussed here). It was introduced as a measure for quantifying the control of a human on the communication between other humans in a social network by Linton Freeman. In his conception, vertices that have a high probability to occur on a randomly chosen shortest path between two randomly chosen nodes have a high betweenness.

The betweenness of a vertex $v$ in a graph $G := (V, E)$ with $V$ vertices is computed as follows:

1. For each pair of vertices $(s, t)$, compute the shortest paths between them.
2. For each pair of vertices $(s, t)$, determine the fraction of shortest paths that pass through the vertex in question (here, vertex $v$).
3. Sum this fraction over all pairs of vertices $(s, t)$.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{10}$$

where $\sigma_{st}$ is total number of shortest paths from node $s$ to node $t$ and $\sigma_{st}(v)$ is the number of those paths that pass through $v$.

We can do the same for paths.

"Edge betweenness" of an edge is the number of shortest paths between pairs of nodes that run along it. If there is more than one shortest path between a pair

of nodes, each path is assigned equal weight such that the total weight of all of the paths is equal to unity.

## 11 Graph clustering and community detection

One often assumes that edge density between clusters is smaller than edge density within the clusters. This is useful, but not every meaningful structure can be described this way.

Especially for large graphs randomness has a (large) role, so we often use probabilistic algorithms.

## 12 Software

free, open source:
octave, scilab, sagemath, igraph, R, ...

Commercial software:
Matlab, Mathematica, S, ...

To have a hint how network analysis may look like a simple example is given.

To analys large graphs one is encouraged to use igraph:

In the following example you have your own favorite graph named yourgraph in graphml format and cluster it using a walktrap algorithm. Then you write the sizes of the clusters found.

```
python
from igraph import *
g=ReadGraph("yourgraph.graphml")
clusters=g.community_walktrap()
for i in range(len(clusters)):  len(clusters[i])
```