Machine learning
Lecture 1

Marcin Wolter
IFJ PAN

27 February 2017

- Machine learning: what does it mean?
- What is a difference between “machine learning” and “bayesian learning”?
- A little bit of mathematics and examples of simple linear classifiers.
- Software to work with and literature.

Recommended books

- M. Krzyśko, *Systemy uczące się: rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*. WNT, 2008.
- C. Bishop, *Pattern recognition and machine learning*. Springer, 2009.

and maybe my thesis:

- M. Wolter, *Metody analizy wielu zmiennych w fizyce wysokich energii*

https://www.epnp.pl/ebook/metody_analizy_wielu_zmiennych_w_fizyce_wysokich_energii

Programs

- **TMVA – integrated with the ROOT package**

  http://tmva.sf.net

  Installs together with root

  Very popular at CERN

- http://scikit-learn.org

  scikit-learn - Machine Learning in Python

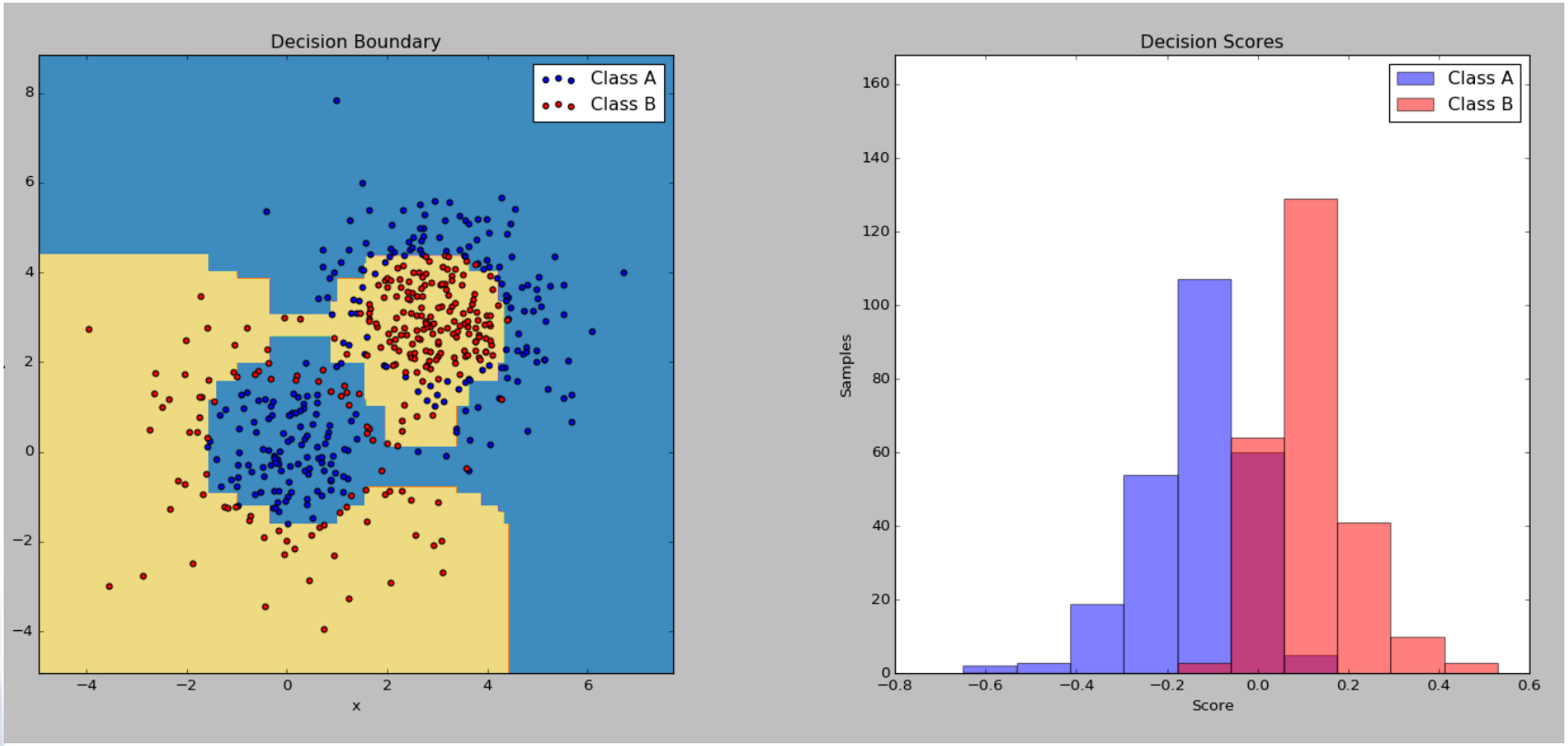
  Simple and efficient tools for data mining and data analysis

  Accessible to everybody, and reusable in various contexts

  Built on NumPy, SciPy, and matplotlib

  I have never used scikit, but we can learn together!

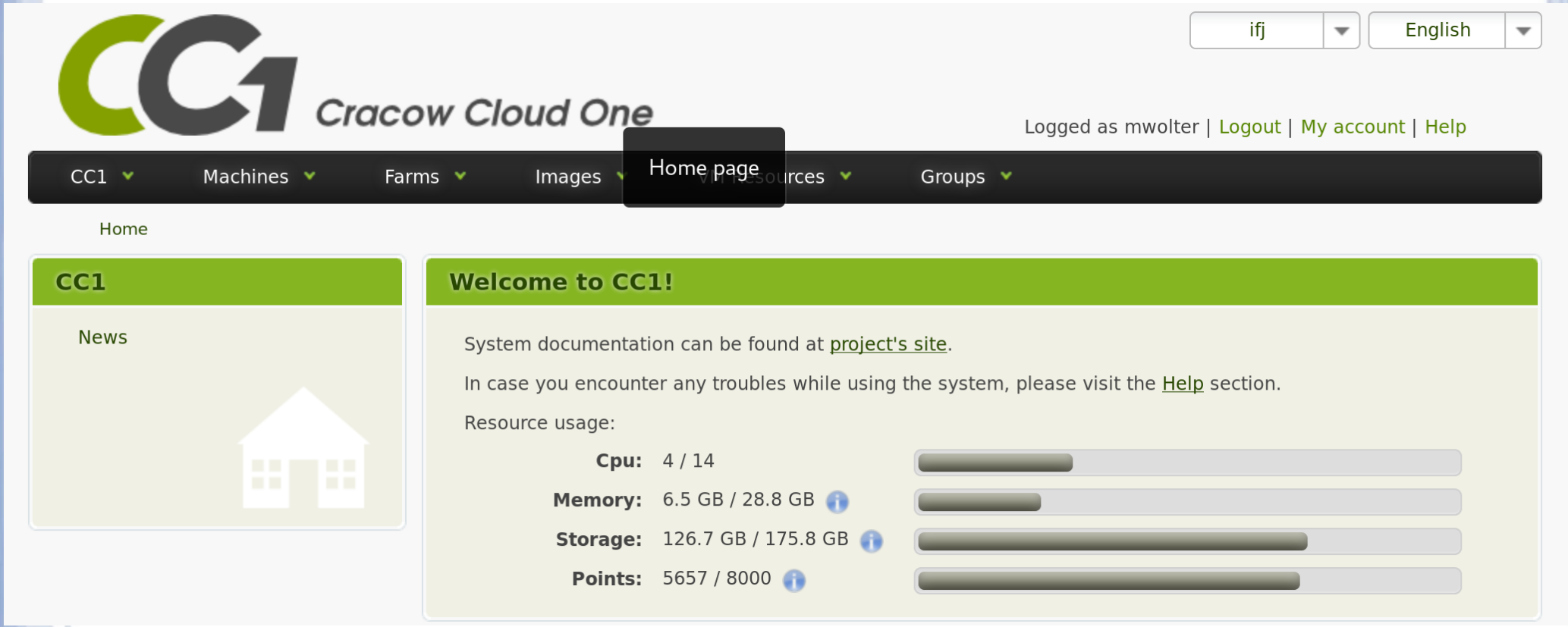
Scikit-learn example



- BDT AdaBoost separation example
- http://scikit-learn.org/stable/auto_examples/ensemble/plot_adaboost_twoclass.html#sphx-glr-auto-examples-ensemble-plot-adaboost-twoclass-py

Computing

- https://www.cloud.ifj.edu.pl/
- Register, you can create your virtual linux box and play with it.



Statistics

- **Statistics – describes random events.**
- First works – أبو يوسف يعقوب بن إسحاق الصبّاح الكندي Al-Kindi (801-873) used statistical methods to break the Ceasar cipher by investigating the frequency in which particular letters appear.

Caesar cipher

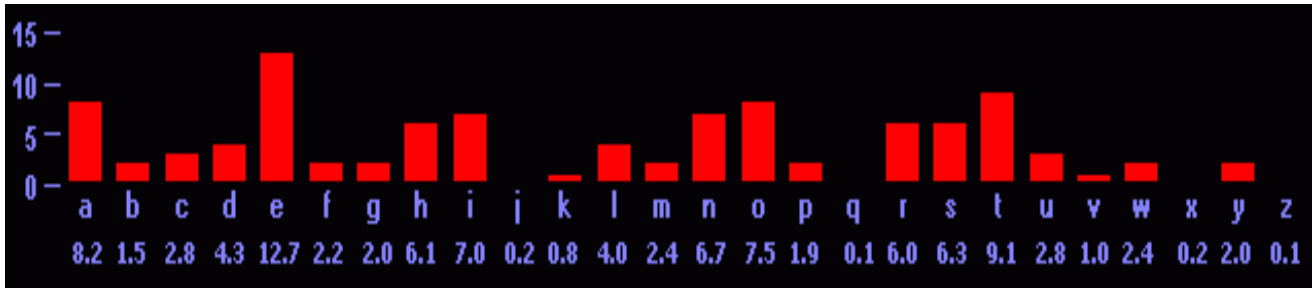
- Each letter of a text is replaced by another, shifted by n letters:

A	B	C	D	E
C	D	E	F	G

- In general Al-Kindi's method can be used to break any replacement cipher (each letter is replaced by another letter, always the same)

A	B	C	D	E
Z	D	P	G	T

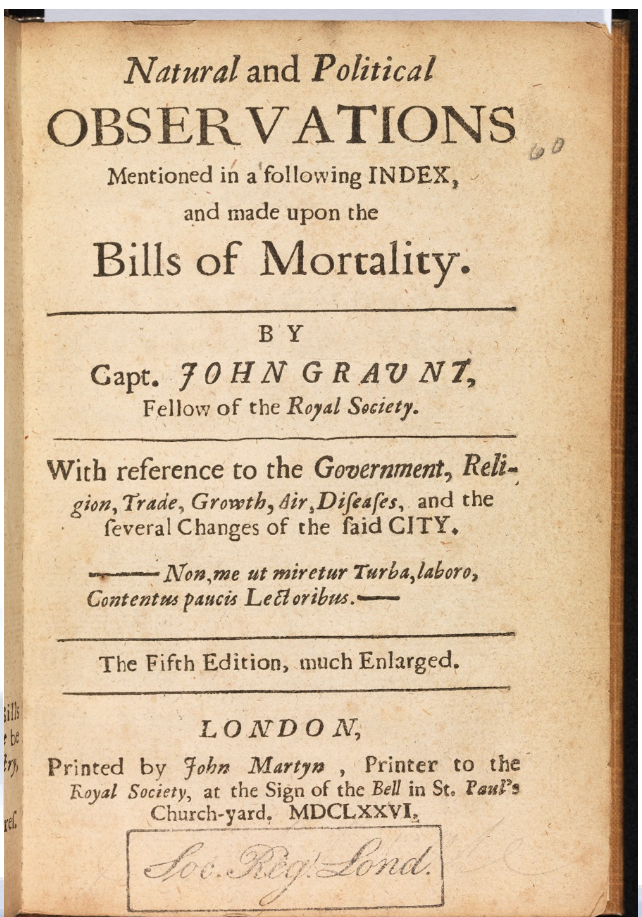
- Frequency analysis – the frequency of appearance of different letters is investigated.

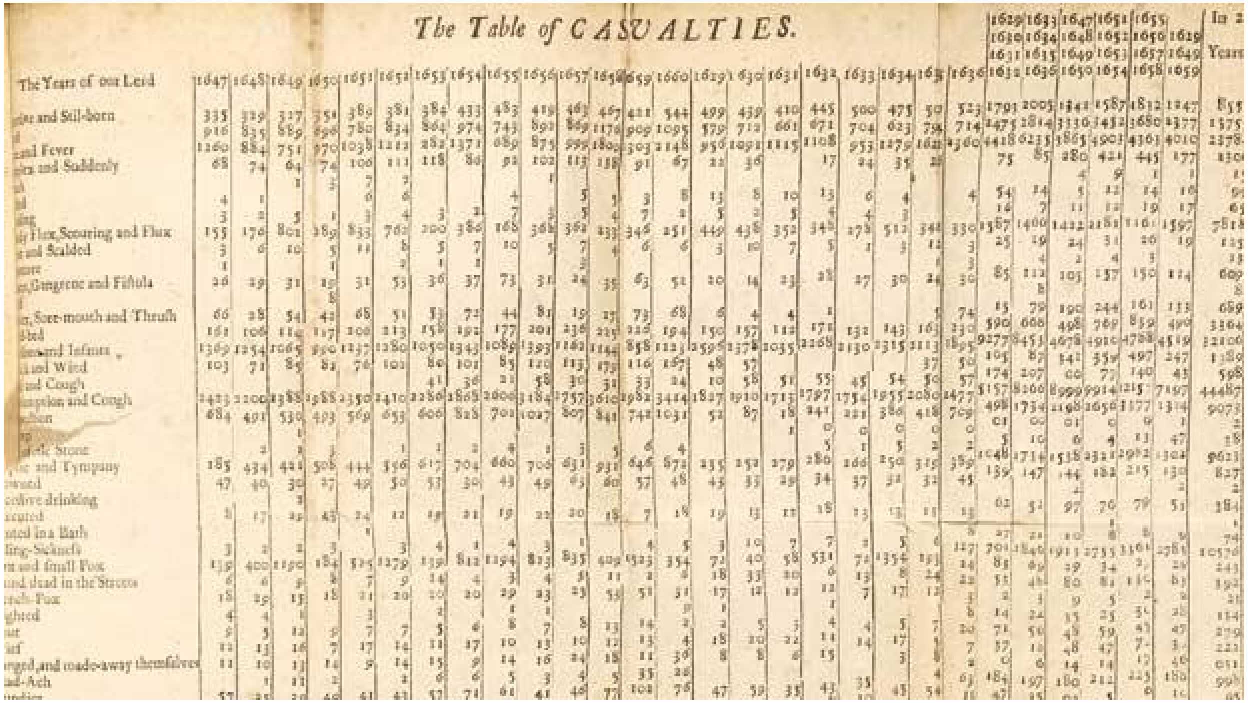


- Frequency of different letters in English.

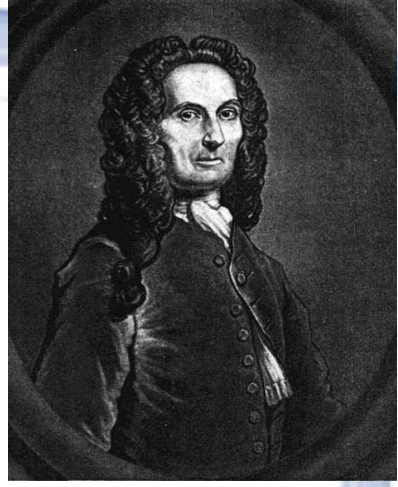
Statistics

- Important step in the development of statistics were the first studies of demography and the games of chance (1663 John Graunt „*Natural and Political Observations upon the Bills of Mortality*”).





How to define the probability?

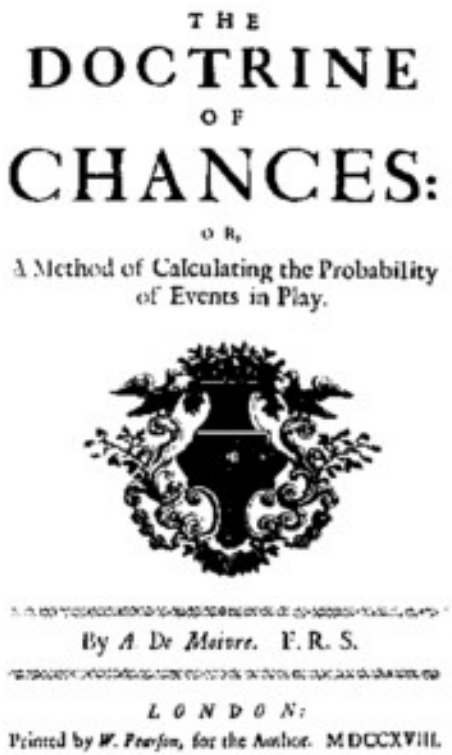


- In probability theory, the sample space of an experiment or random trial is the set of all possible outcomes or results of that experiment.
- **Probability of an event A (frequentist definition) is a limit by N going to infinity of n/N, where n is a number of successes and N is a number of trials:**

$$P(A) = \lim_{N \to \infty} \frac{n}{N}$$

What is a probability of getting “6” by throwing a dice? The same as the fraction of “6” results in the infinite number of trials.

- The definition comes from a text book of Abraham de Moivre (1667-1754) – a text book on statistics „*The Doctrine of Chances*” (1718).

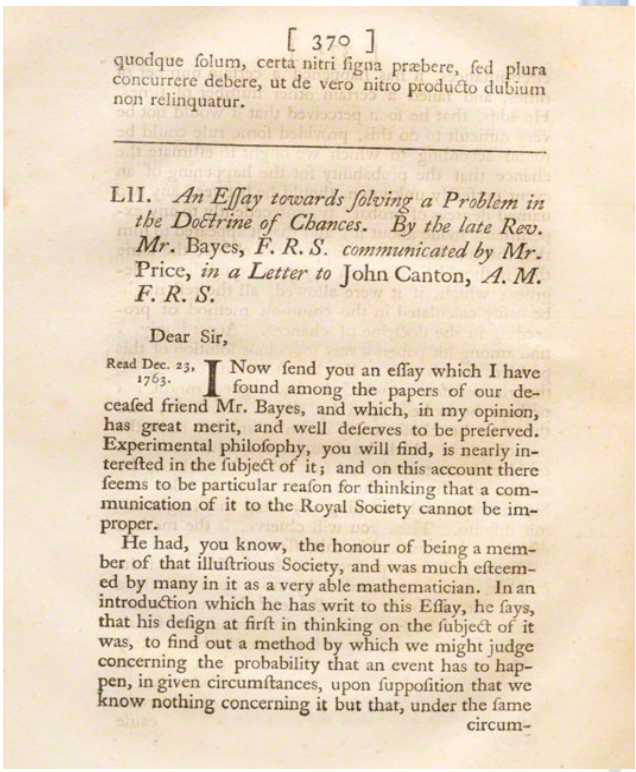

THE
DOCTRINE
OF
CHANCES:
OR,
A Method of Calculating the Probability
of Events in Play.

By *A. De Moivre.* F. R. S.

LONDON:
Printed by *W. Pearson*, for the Author. MDCCXVIII.

Bayes definition

- Probability “a priori”, i.e. unconditional, is understood as a measure of belief, based on rational evidence, that such an event will happen.
- In the next step we make an experiment, called “observations”, and their results modify the probability. We get the probability “a posteriori”, which is a measure of belief modified by the experiment.
- This idea of Thomas Bayes was supported by P. S. Laplace, H. Poincare or the well known economist John Keynes. The stated, that this is a way we recognize and study the nature.



[370]

quodque ſolum, certa nitri ſigna præbere, ſed plura concurrere debere, ut de vero nitro producto dubium non relinquatur.

LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

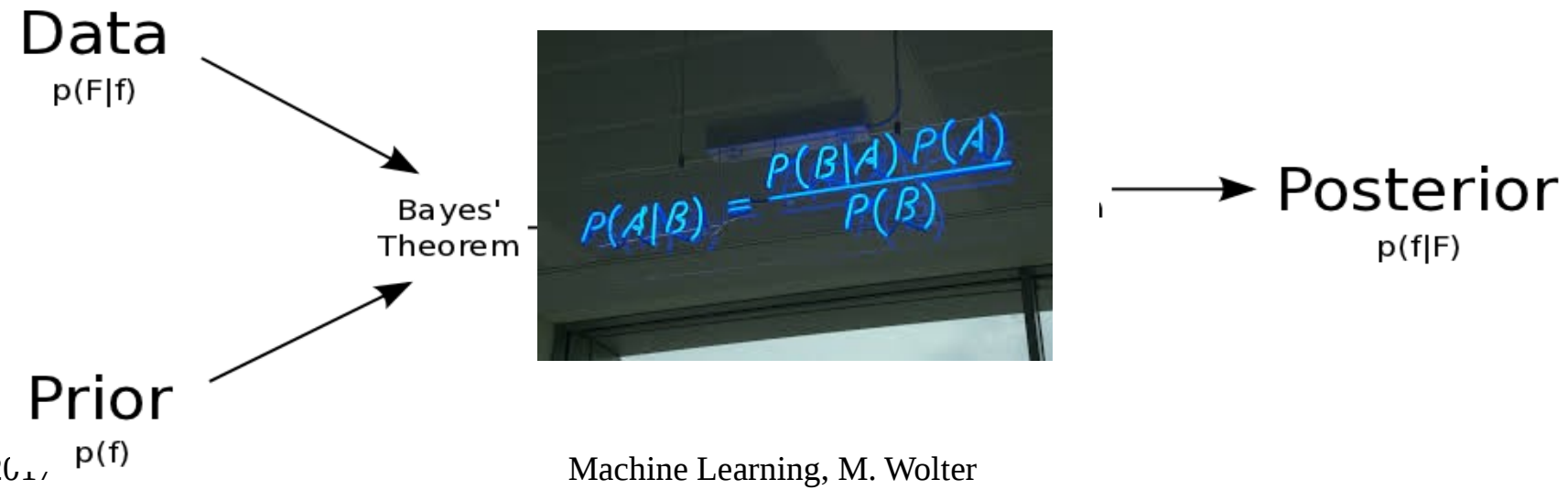
Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illuſtrious Society, and was much eſteemed by many in it as a very able mathematician. In an introduction which he has writ to this Eſſay, he ſays, that his deſign at firſt in thinking on the ſubject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumſtances, upon ſuppoſition that we know nothing concerning it but that, under the ſame circum-

Thomas Bayes (1702 - 1761) was a British mathematician and the Presbyterian pastor. His most important work is „Essay Towards Solving a Problem in the Doctrine of Chances”.

Bayes definition

- The experiment we can’t repeat many times: what is a probability to pass an exam?
- Based on our knowledge (we studied text books for few days), we estimate the probability to be: ½ (probability *a priori*).
- If all four people trying to pass the exam before us failed, and we know that their knowledge wasn’t much different from our, wouldn't we modify our estimation? In this way we get the probability *a posteriori*.



Bayes Theorem

- Bayes' theorem relates the conditional (posterior) and marginal (prior) probabilities of events A and B:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- **P(A)** is the prior probability or marginal probability of A. It is a "prior" in the sense that it does not take into account any information about B.
- **P(A|B)** is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.

- Intuitively, Bayes' theorem in this form describes the way in which one's beliefs about observing 'A' are updated by having observed 'B'.

Bayes Theorem – an example: cancer test

$$\Pr(\mathrm{A}|\mathrm{X}) = \frac{\Pr(\mathrm{X}|\mathrm{A})\Pr(\mathrm{A})}{\Pr(\mathrm{X})} = \frac{\Pr(\mathrm{X}|\mathrm{A})\Pr(\mathrm{A})}{\Pr(\mathrm{X}|\mathrm{A})\Pr(\mathrm{A}) + \Pr(\mathrm{X}|\mathrm{not\ A})\Pr(\mathrm{not\ A})}$$

- Pr(A|X) = Chance of having cancer (A) given a positive test (X). This is what we want to know: How likely is it to have cancer with a positive result? .
- Pr(X|A) = Chance of a positive test (X) given that you had cancer (A). This is the chance of a true positive, 80% in our case.
- Pr(A) = Chance of having cancer (1%).
- Pr(not A) = Chance of not having cancer (99%).
- Pr(X|not A) = Chance of a positive test (X) given that you didn't have cancer (~A). This is a false positive, 9.6% in our case.
- In our case Pr(A|X) is 7.8%

Bayesian vs. Frequentist approach

- **PROBABILITY: degree of belief** (Bayes, Laplace, Gauss, Jeffreys, de Finetti)
- **PROBABILITY: relative frequency** (Venn, Fisher, Neyman, von Mises).
- **Bayesian approach:** probability is degree of belief. Thus the probability *p* is our assessment of the probability of success at each trial, based on our current state of knowledge.

  If our assessment, initially, is incorrect? As our state of knowledge changes, our assessment of the probability of success changes accordingly.
- **Bayesian inference** is statistical inference in which **evidence or observations are used** to update or to newly infer the probability that a hypothesis may be true.
- This allows for a *cleaner* foundation than the frequentist interpretation.

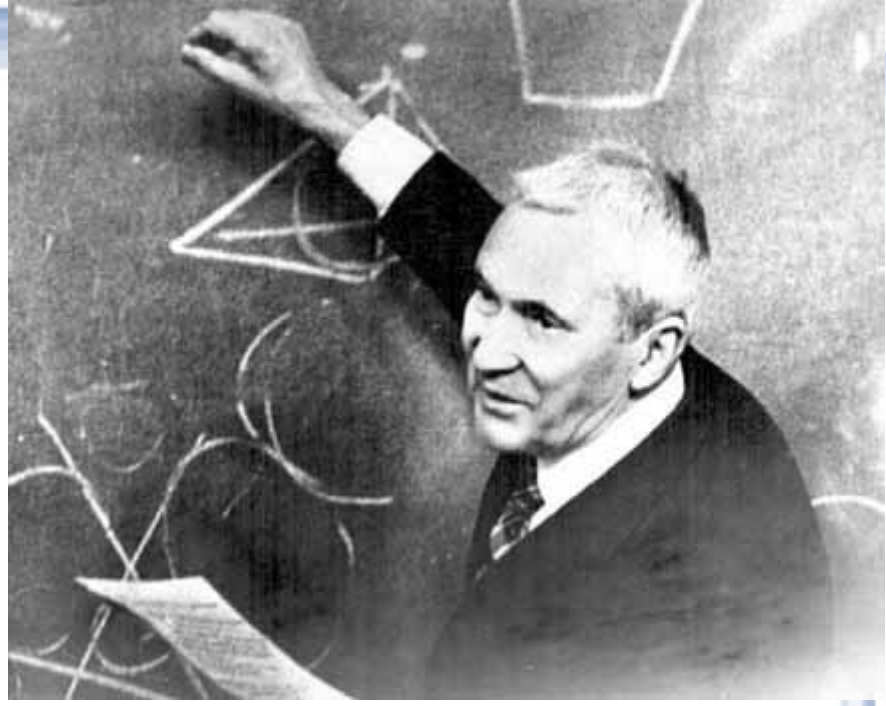
"*We don't know all about the world to start with; our knowledge by experience consists simply of a rather scattered lot of sensations, and we cannot get any further without some a priori postulates. My problem is to get these stated as clearly as possible.*"

Sir Harold Jeffreys, in a letter to Sir Ronald Fisher dated 1 March, 1934

H.B. Prosper,"Bayesian Analysis", arXiv:hep-ph/0006356v1 30 Jun 2000

Axiomatic definition

Probability could be defined in many ways, not necessarily like de Moivre...

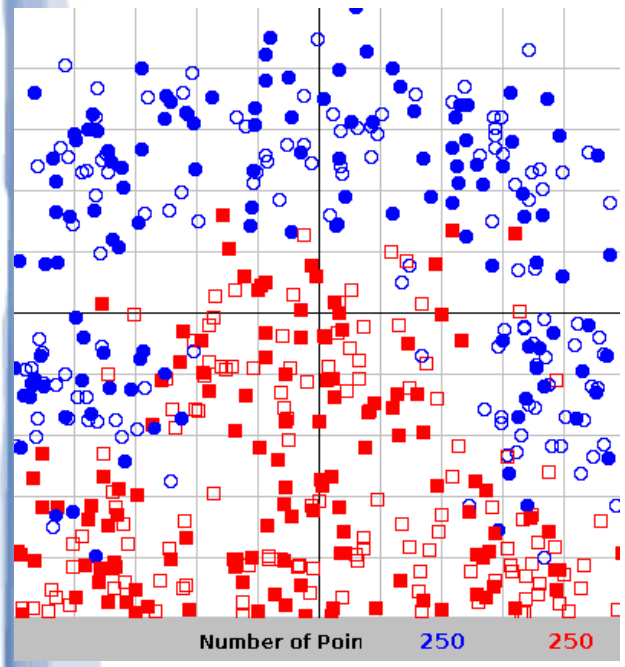


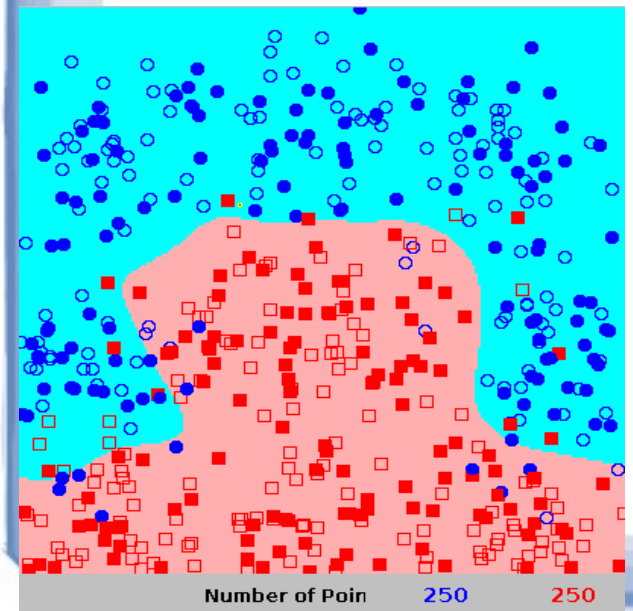
Андре́й Никола́евич Колмого́ров (1903-1987)

Axiomatic definition by Kolmogorov: probability is a function P defined on the space of elementary events, which assigns to each event A a number P(A) such, that:

- P(A)≥0 for each event A
- P(A)=1 for the sure event A
- $P(A \cup B) = P(A) + P(B)$ when the events A and B are mutually exclusive.

How do the machine learning algorithms work?



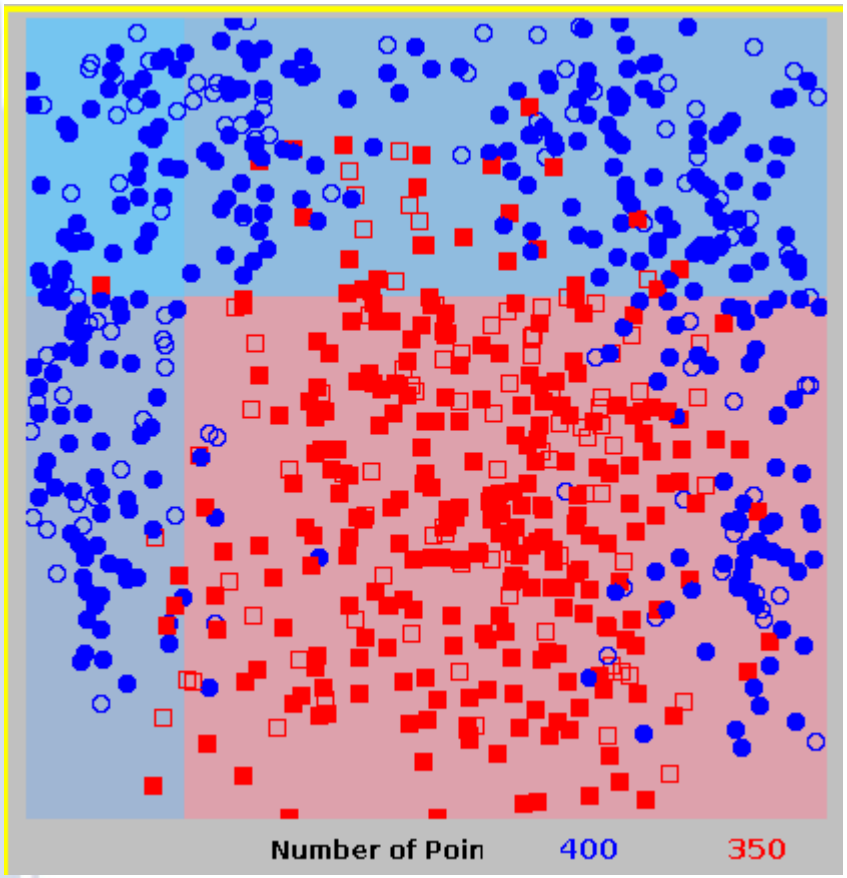


- We need **training data**, for which we know the correct answer, whether it's a signa or background. We divide the data into two samples: training and test.
- We find the best function ***f(x)*** which describes the probability, that a given event belongs to the class "signal". This is done by minimizing the loss function (for example χ^2).
- Different algorithms differ by: the class of function used as ***f(x)*** (linear, non-linear etc), loss function and the way it's minimized.
- All these algorithms try to approximate the unknown *Bayessian Decisive Function* (BDF) relying on the finit training sample.

BDF -an ideal classification function given by the unknown probability densities of signal and background.

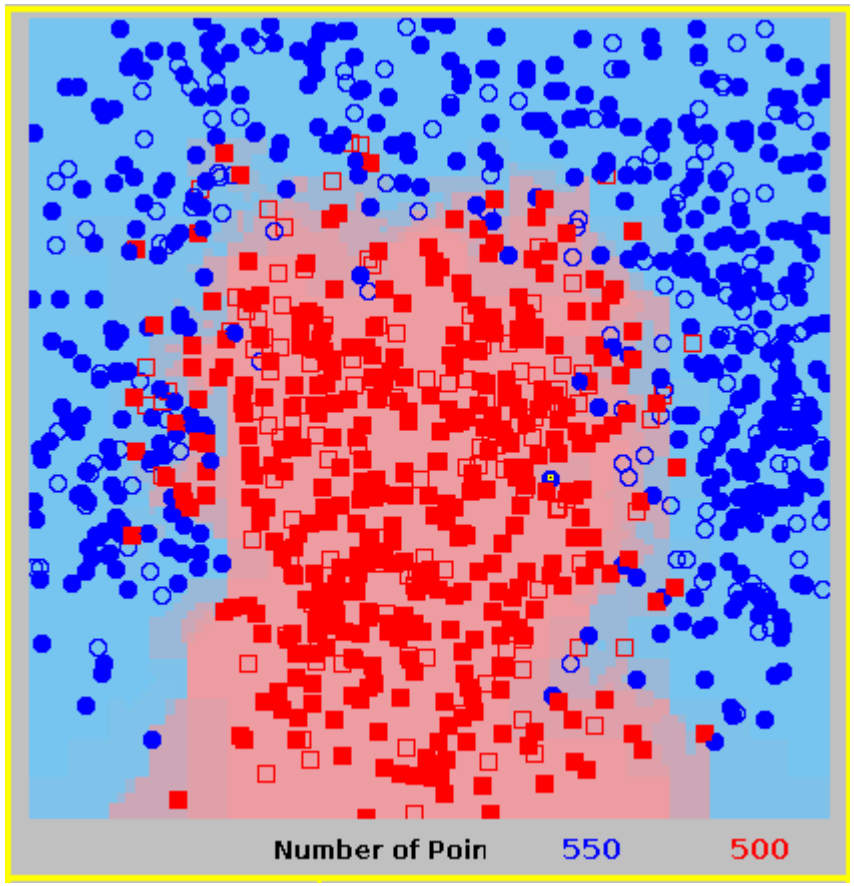
Cuts vs non-linear separation

Cuts



?

Non-linear separation



Neural Networks, boosted decision trees, itd

What are ML methods used for?

Multivariate methods (machine learning) can be used for:

- **Classification (select signal out of background)**
- **Function approximation (regression) –** for example find a particle energy using signals from the calorimeter cells.

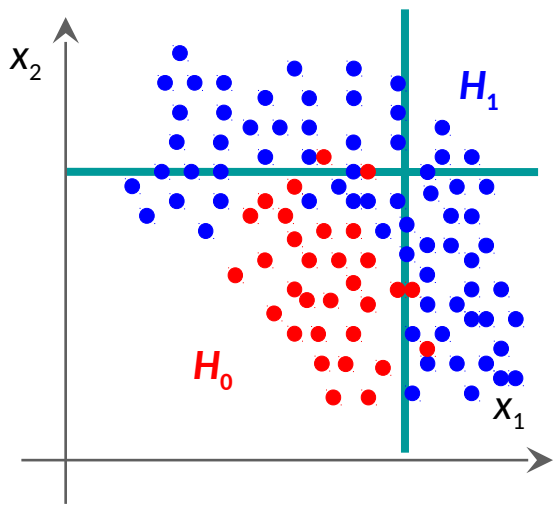
- Probability density estimation (estimate the probability distribution)
- Variable selection (find most important variables)
- Optimization (tuning, verification)
- Many others...

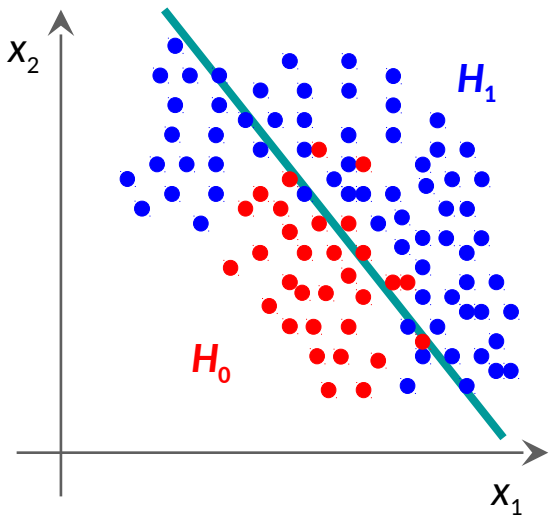
Types of algorithms

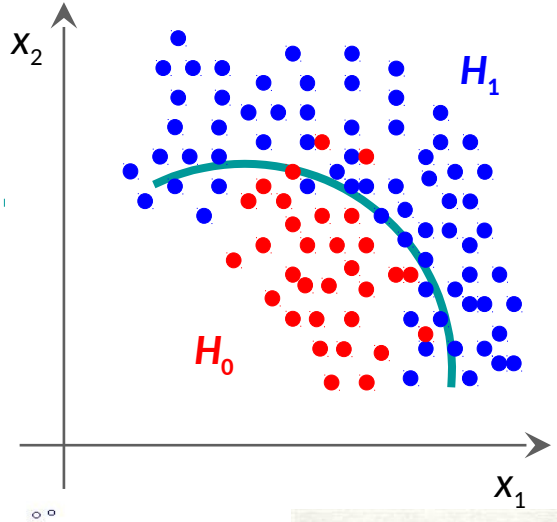
How to use the information available

•**Classification:** find a function ***f(x1,x2)*** giving the probability, that a given data point belongs to a given class (signal vs background).

Simple cuts (easy and intuitive)	Linear (fast and stable)	Non-linear (most effective)

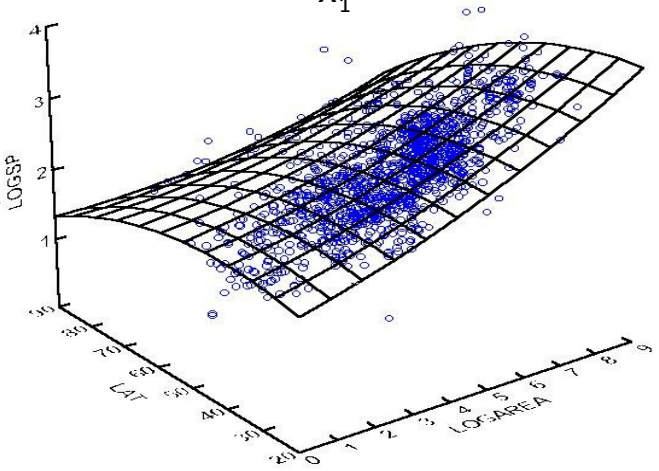


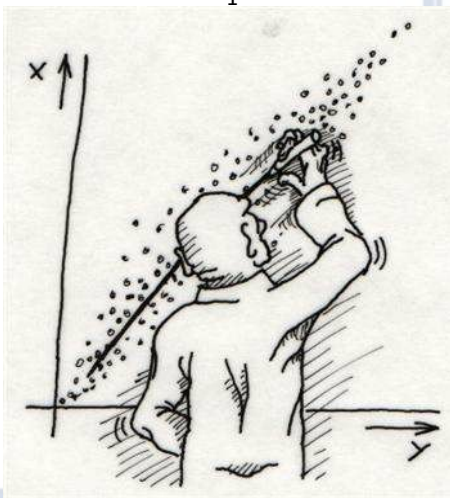




Regression: fit a continuous function

(find particle energy from calo readouts).





Multivariate Methods

Machine Learning

Teach a machine to learn $y = f(x)$ by feeding it **training data** $\boldsymbol{T} = (\boldsymbol{x}, \boldsymbol{y}) = (x,y)_1, (x,y)_2,\ldots(x,y)_N$ and a **constraint** on the class of functions.

Bayesian Learning

For each function f(x) in the function space F calculate the posterior probability $p(\boldsymbol{f} | \boldsymbol{T})$ using a given training sample T= $(\boldsymbol{x}, \boldsymbol{y})$.

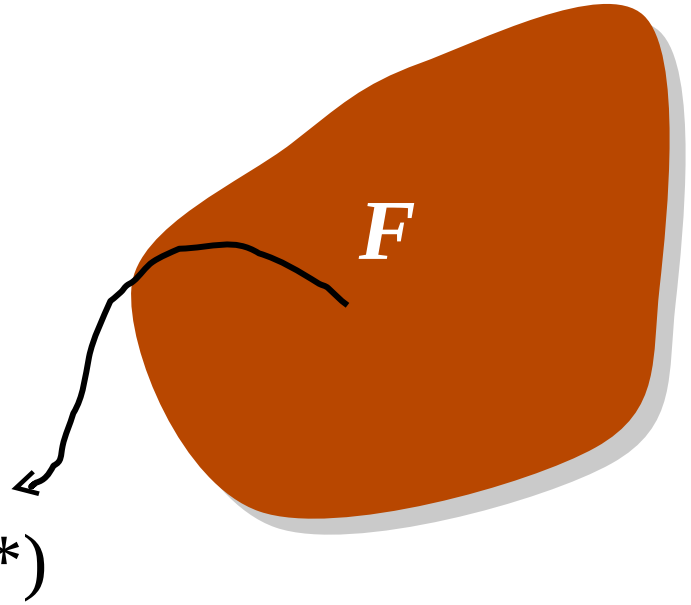
Don't use a single function, use a bunch of functions weighted by probabilities

- The posterior probability is the conditional probability that is assigned after the relevant evidence is taken into account.
- Training sample: $\boldsymbol{T} = (\boldsymbol{x}, \boldsymbol{y})$ set of input vectors $\boldsymbol{x}$ and desired outputs $\boldsymbol{y}$.

Machine Learning

We have to chose:

Function class	***F* = { *f*(*x*, *w*) }**
Constraint	***C***
Loss function	***L***
	example:
	$L=y_i^2 - f^2(x_i,w)$
Training data	***T*(*y*,*x*)**



Method

Find *f*(*x*, ***w***) by minimizing the **empirical risk *R***

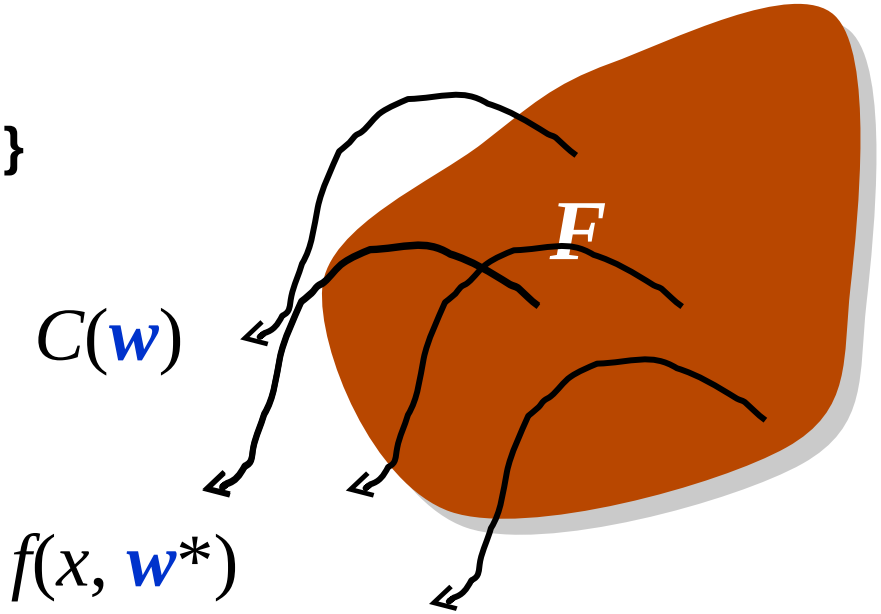
$$R(w)=\frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i, w))$$ subject to the constraint *C*(***w***)

We choose at the end a single “best” function *f*(*x*, ***w***)
(best single Neural Network, best likelihood etc.)

Bayesian Learning

Choose

Function class	***F* = { *f*(*x*, *w*) }**
Prior	*p*(*w*)
Likelihood	*p*(***y***|***x***, *w*)
Training data	***T*(*y*,*x*)**



We do not pick up a single function *f(x)*, instead *p*(*w*|*T*) assigns a probability density to every function in the function class.

Bayesian Learning: Why?

- **Probabilistic learning:** Calculate explicit probabilities for hypothesis, among the most practical approaches to certain types of learning problems.
- **Probabilistic prediction**: Predict multiple hypotheses, weighted by their probabilities.
- **Incremental:** Each training example can incrementally increase or decrease the probability that a hypothesis is correct. Prior knowledge can be combined with observed data.
- **Standard:** Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured.

Classification

A Bayes classifier:

$$p(S|x) = \frac{p(x|S)\,p(S)}{p(x|S)\,p(S) + p(x|B)\,p(B)}$$

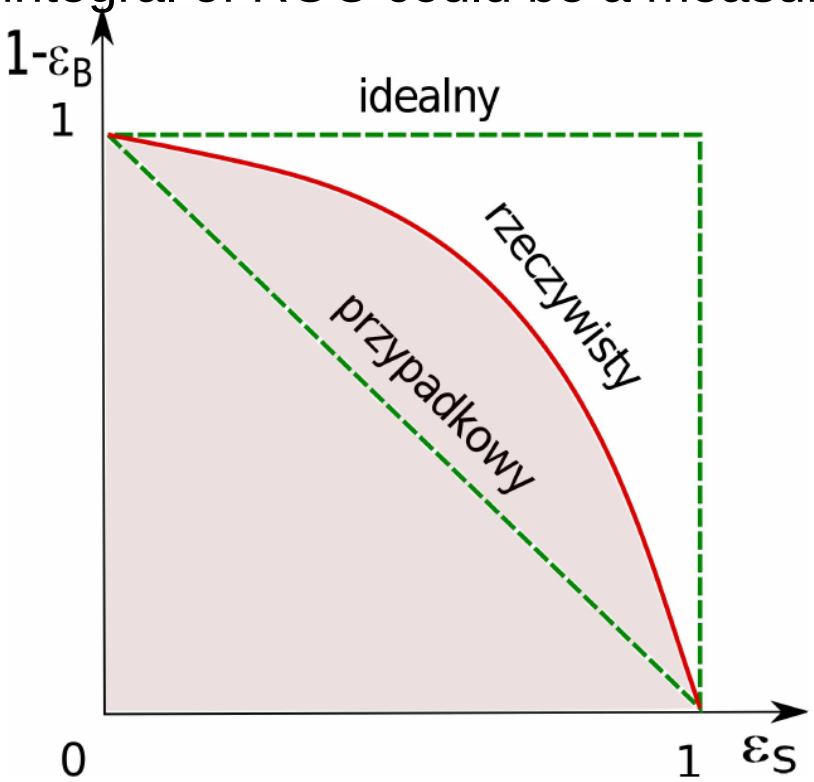
where ***S*** is associated with *y* = **1** and ***B*** with *y* = **0**. **Bayes classifier** *accepts events x if p(**S**|x) > **cut** as belonging to **S**.*

We need to approximate probability distributions *P(**x**|**S**)* and *P(**x**|**B**).*

- *If your goal is to **classify objects** with the fewest errors, then the **Bayes classifier** is the **optimal** solution.*
- Consequently, if you have a classifier known to be ***close*** to the **Bayes limit**, then *any* other classifier, *however sophisticated*, can ***at best*** be only marginally better than the one you have.
  - =>If your problem is **linear** you don't gain anything by using sophisticated **Neural Network**
- *All* classification methods, such as the ones in TMVA, are different numerical approximations of the Bayes classifier.

ROC curve

- ROC (Receiver Operation Characteristic) curve was first used to calibrate radars.
- Shows the background rejection ($1-\varepsilon_B$) vs signal efficiency ε_B. Shows how good the classifier is.
- The integral of ROC could be a measure of the classifier quality:



Integral(ROC) = ½ – random
Integral(ROC) = 1 - ideal

Practical applications

A Short List of Multivariate Methods

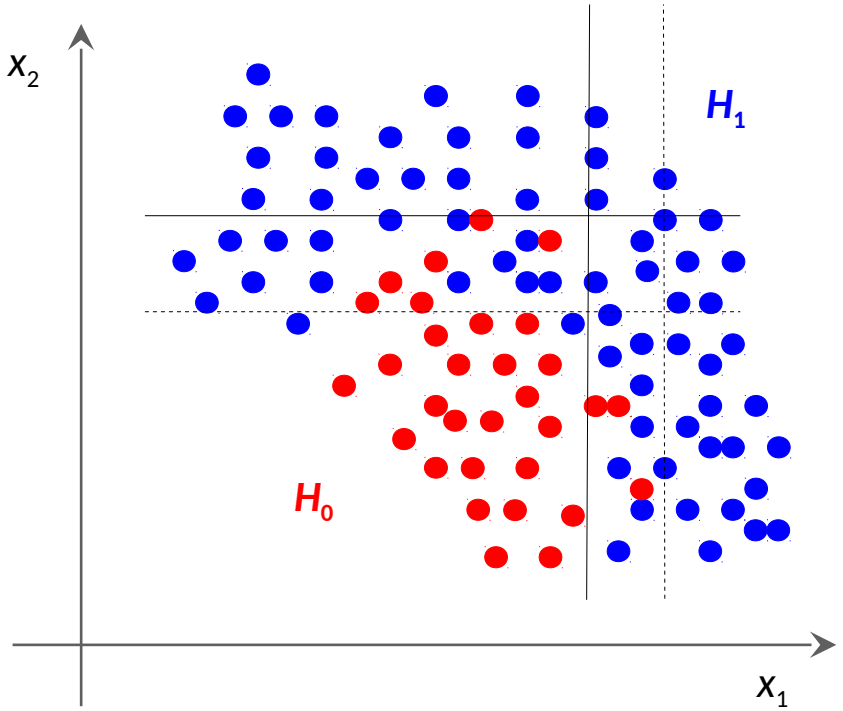
- Cuts
- Linear Discriminants (like Fisher)
- Support Vector Machines
- Naive Bayes (Likelihood Discriminant)
- Kernel Density Estimation
- Decision Trees
- Neural Networks
- Bayesian Neural Networks
- Genetic Algorithms

- And many, many others..... I want to present briefly just few of them.

We describe now

- Simple ML linear methods:
  - Cuts
  - Fisher linear discriminant
  - Principal Component Analysis, PCA
  - Independent Component Analysis, ICA

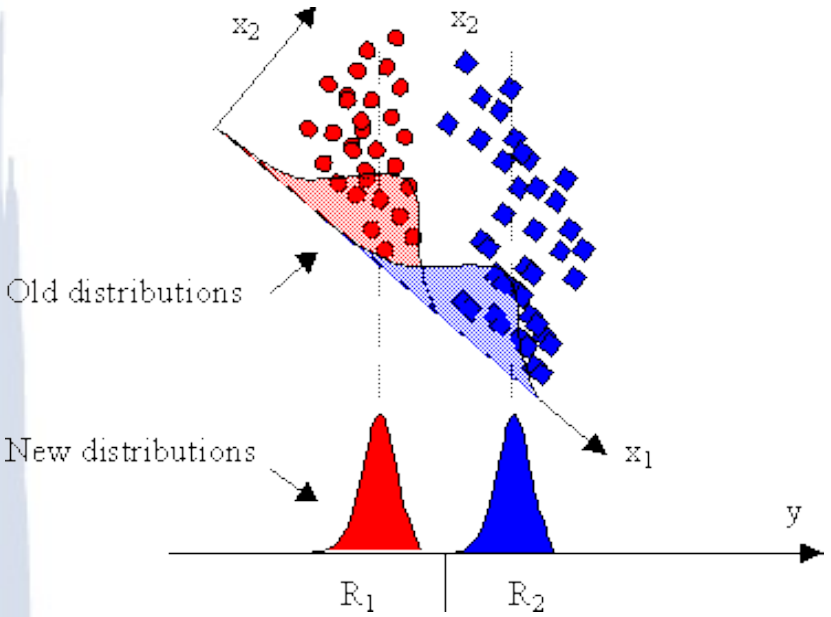
Cuts



- Optimization of cuts:
- Move cuts as long as we get the optimal signal vs. background selection. For a given signal efficiency we find the best background rejection → we get the entire ROC curve.
- Optimization methods:
  - Brute force
  - Genetic algorithms
  - Many others...

Fisher discriminants

Projection to one dimension, than discrimination



Equivalent to linear separation

We choose a projection vector in such a way, that the separation is maximized.

Method introduced by Fisher in 1936.
Optimal separation for Gaussian distributions.

Fisher's linear discriminant

The terms *Fisher's linear discriminant* and *LDA* are often used interchangeably, although Fisher's original article *The Use of Multiple Measures in Taxonomic Problems* (1936) actually describes a slightly different discriminant, which does not make some of the assumptions of LDA such as normally distributed classes or equal class covariances.

Suppose two classes of observations have means $\vec{\mu}_{y=0}, \vec{\mu}_{y=1}$ and covariances $\Sigma_{y=0}, \Sigma_{y=1}$. Then the linear combination of features $\vec{w} \cdot \vec{x}$ will have means $\vec{w} \cdot \vec{\mu}_{y=i}$ and variances $\vec{w}^T \Sigma_{y=i} \vec{w}$ for $i = 0,1$. Fisher defined the separation between these two distributions to be the ratio of the variance between the classes to the variance within the classes:

$$S = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{(\vec{w} \cdot \vec{\mu}_{y=1} - \vec{w} \cdot \vec{\mu}_{y=0})^2}{\vec{w}^T \Sigma_{y=1} \vec{w} + \vec{w}^T \Sigma_{y=0} \vec{w}} = \frac{(\vec{w} \cdot (\vec{\mu}_{y=1} - \vec{\mu}_{y=0}))^2}{\vec{w}^T (\Sigma_{y=0} + \Sigma_{y=1}) \vec{w}}$$

This measure is, in some sense, a measure of the signal-to-noise ratio for the class labelling. It can be shown that the maximum separation occurs when

$$\vec{w} = (\Sigma_{y=0} + \Sigma_{y=1})^{-1} (\vec{\mu}_{y=1} - \vec{\mu}_{y=0})$$

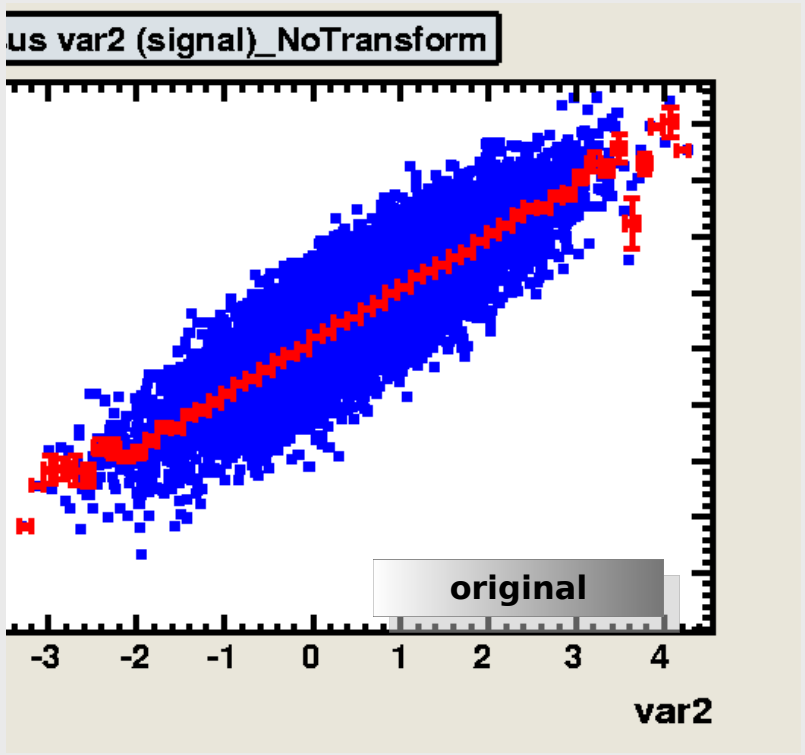
When the assumptions of LDA are satisfied, the above equation is equivalent to LDA.

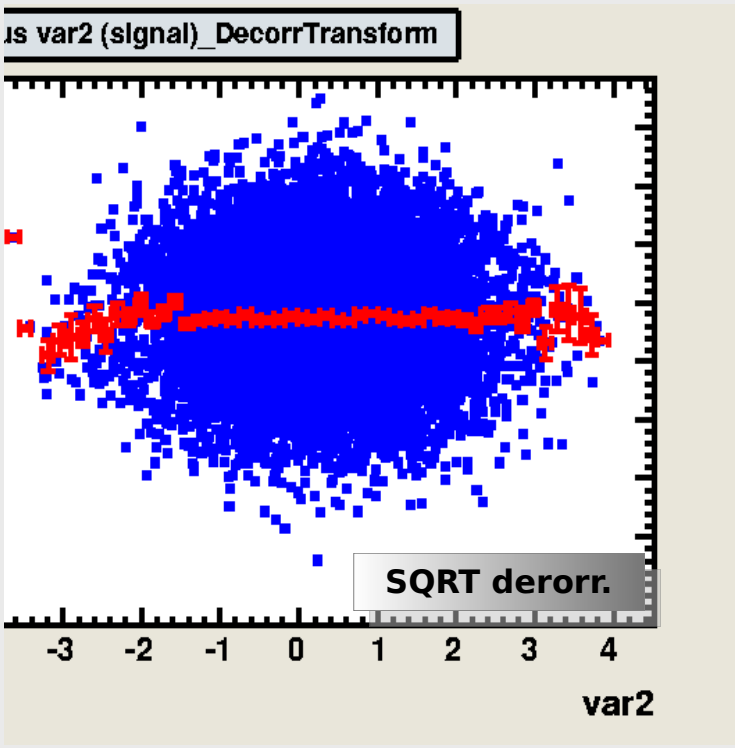
Be sure to note that the vector $\vec{w}$ is the normal to the discriminant hyperplane. As an example, in a two dimensional problem, the line that best divides the two groups is perpendicular to $\vec{w}$.

Generally, the data points are projected onto $\vec{w}$. However, to find the actual plane that best separates the data, one must solve for the bias term b in $w^T \mu_1 + b = -(w^T \mu_2 + b)$.

Decorrelation

- **Removes correlation between variables by a rotation in the space of variables**



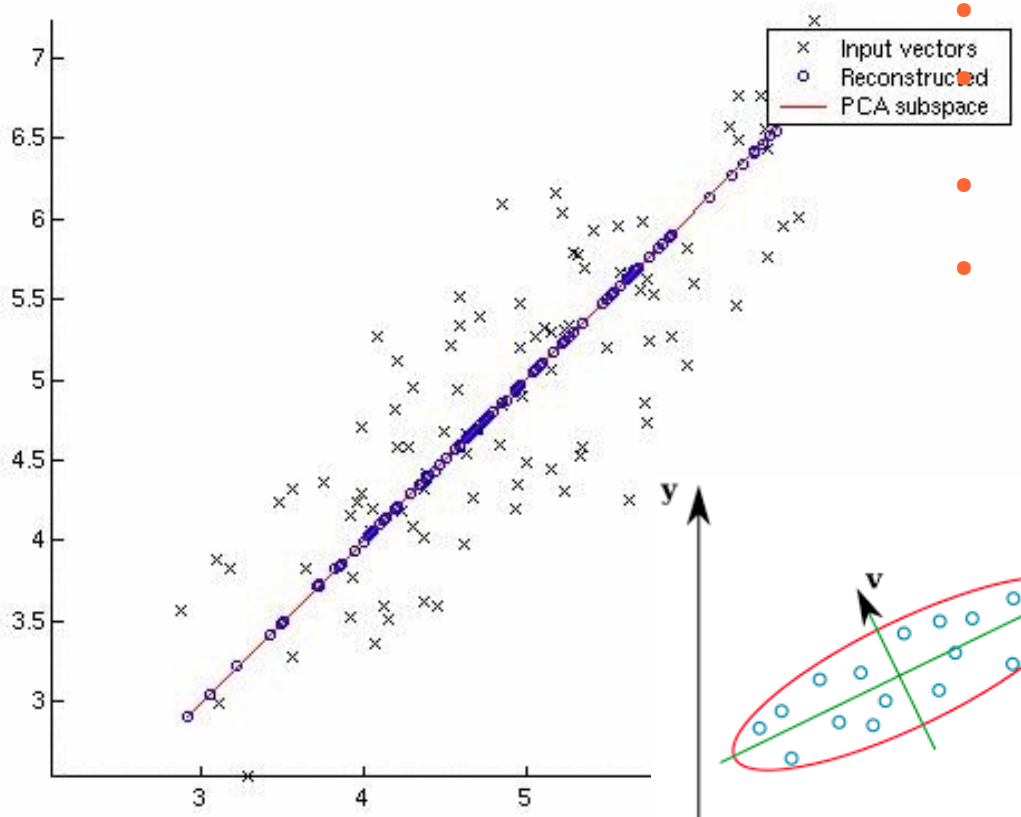


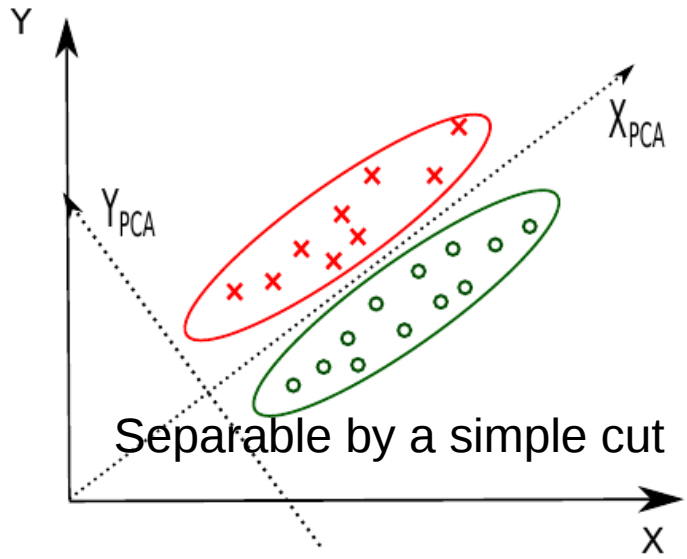
Principal Component Analysis - PCA

- Task: reduce the number of dimensions minimizing the loss of information
- Finds the orthogonal base of the covariance matrix , the eigenvectors with the smallest eigenvalues migh be skipped

Procedure:

- Find the covariance matrix Cov(X)
- Find eigenvalues λ_i and eigenvectors v_i
- Skip smallest λ_i
- Unsupervised learning!!!



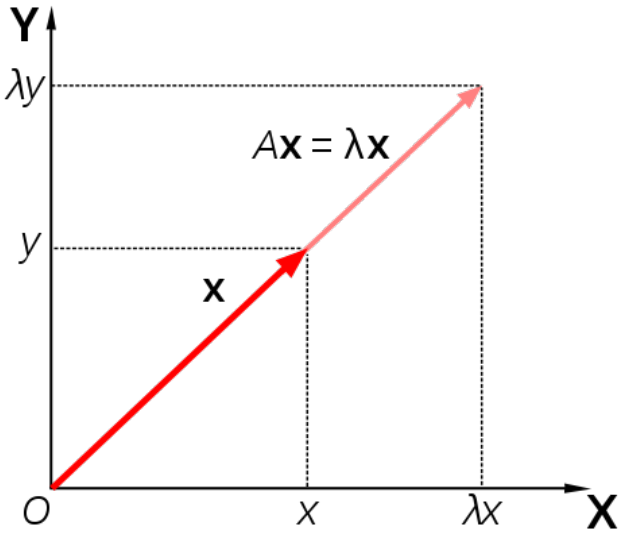


Eigenvalues and eigenvectors

In essence, an eigenvector v of a linear transformation T is a non-zero vector that, when T is applied to it, does not change direction. Applying T to the eigenvector only scales the eigenvector by the scalar value λ, called an eigenvalue. This condition can be written as the equation

$$T(\mathbf{v}) = \lambda \mathbf{v}$$

referred to as the eigenvalue equation or eigenequation. In general, λ may be any scalar. For example, λ may be negative, in which case the eigenvector reverses direction as part of the scaling, or it may be zero or complex.



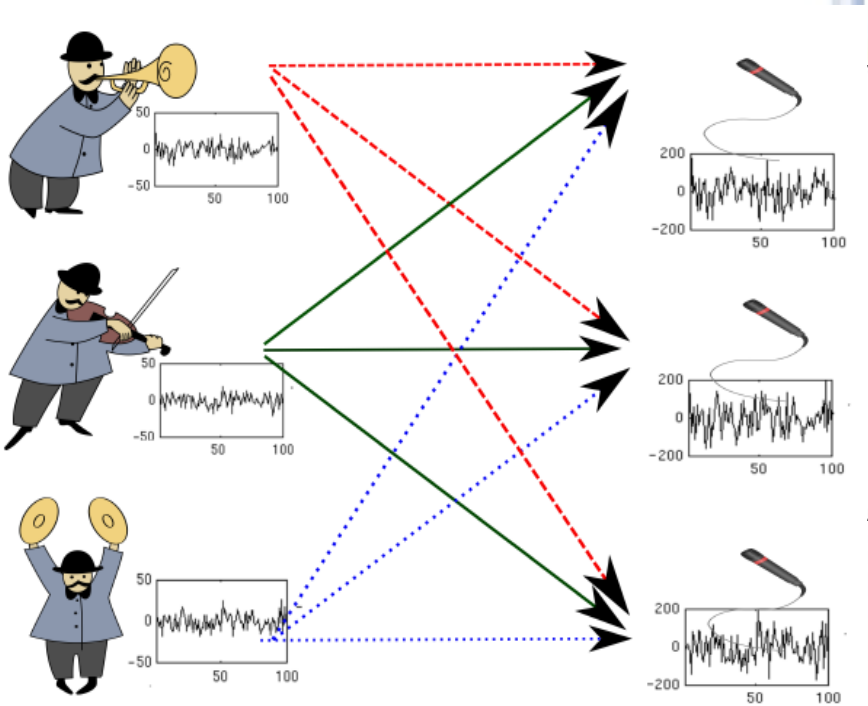
Matrix A acts by stretching the vector x, not changing its direction, so x is an eigenvector of A.

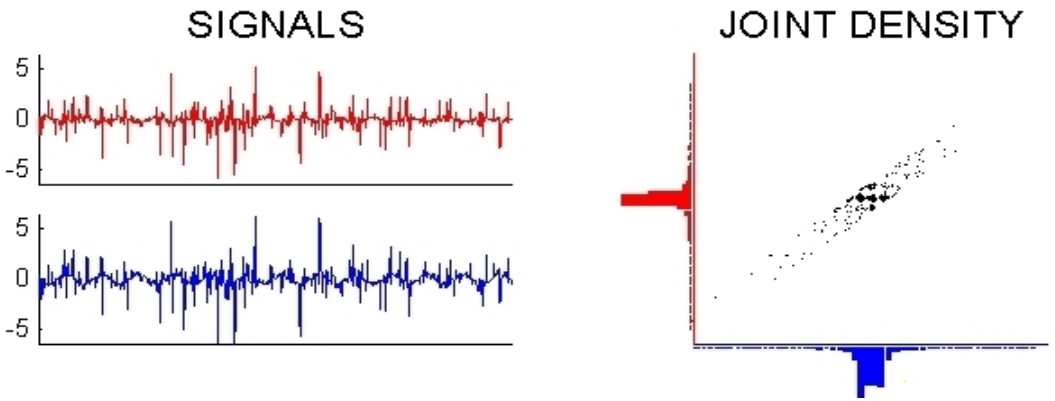
Independent Component Analysis ICA

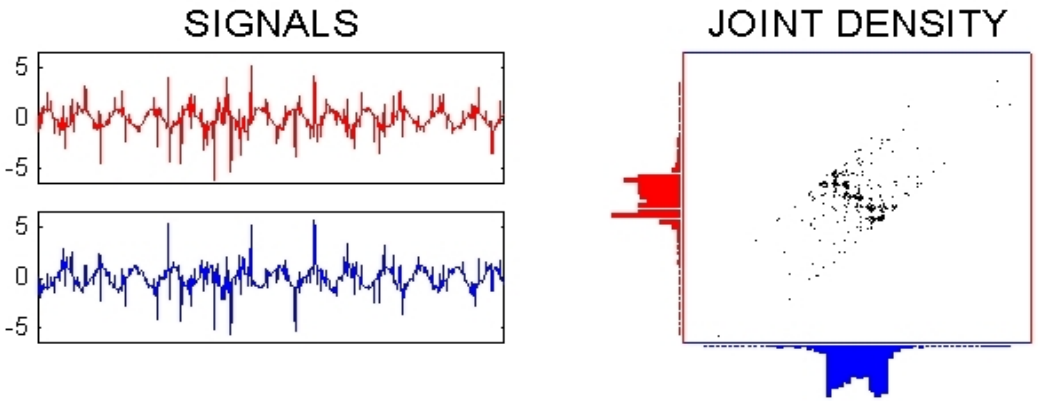
Developed at Helsinki University of Technology *http://www.cis.hut.fi/projects/ica/*

- **Problem:**
  - Assume, that signal **X** is a linear combination **X = AS** of independent sources **S**. The mixing matrix **A** and vector of sources **S** are unknown.
  - **Task:** find a matrix **T (inverted A)**, such that elements of vector **U = TX** are statistically independent. **T** is the matrix returning the original signals.

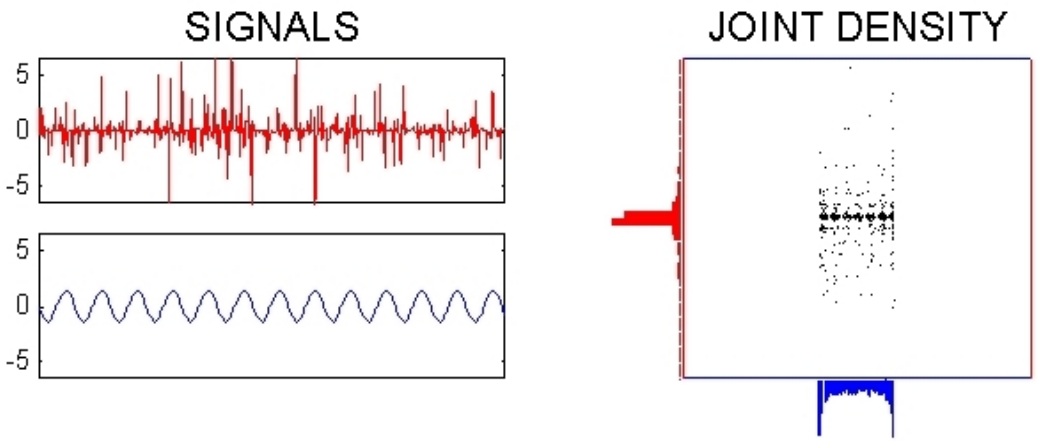
- **Applications:**
  - Filtering of one source out of many others,
  - Separation of signals in telecommunication,
  - Separation of signals from different regions of brain,
  - Signal separation in astrophysics,
  - Decomposition of signals in accelerator beam analysis in FERMILAB.







Whitened signals and density



Separated signals after 5 steps of FastICA

How does ICA work?

- We have two measured signals and we want to separate them into two independent sources.
- Preparing data - decorrelation (correlation coefficients equal zero, σ=1).

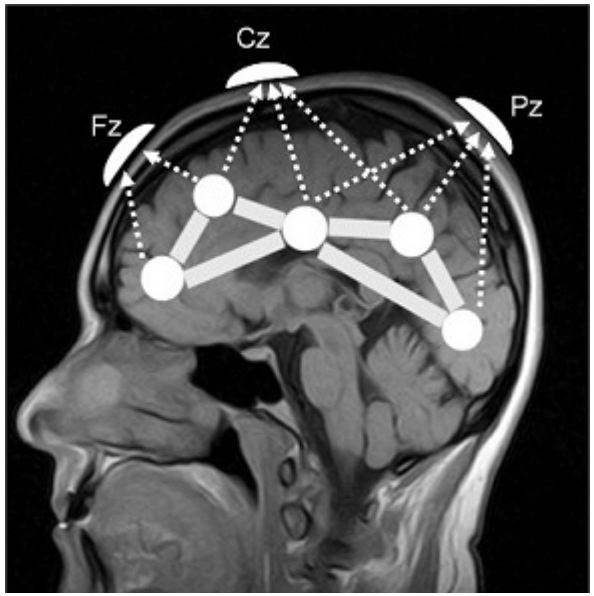
Superposition of many independent distributions gives Gaussian in the limit.

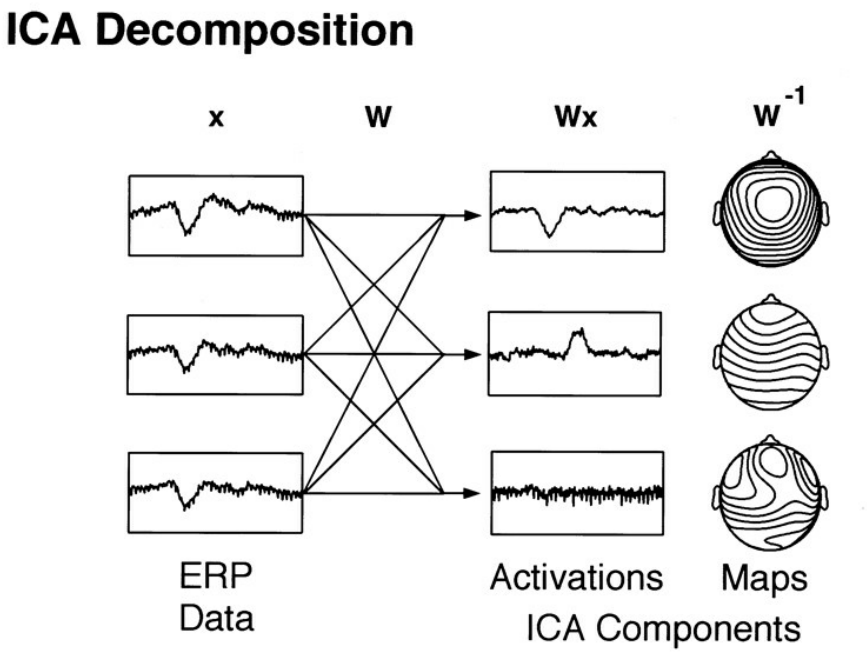
- ICA – rotation, signals should be maximally non-Gaussian (measure of non-Gaussianity might be curtosis).
- *Curtosis:*

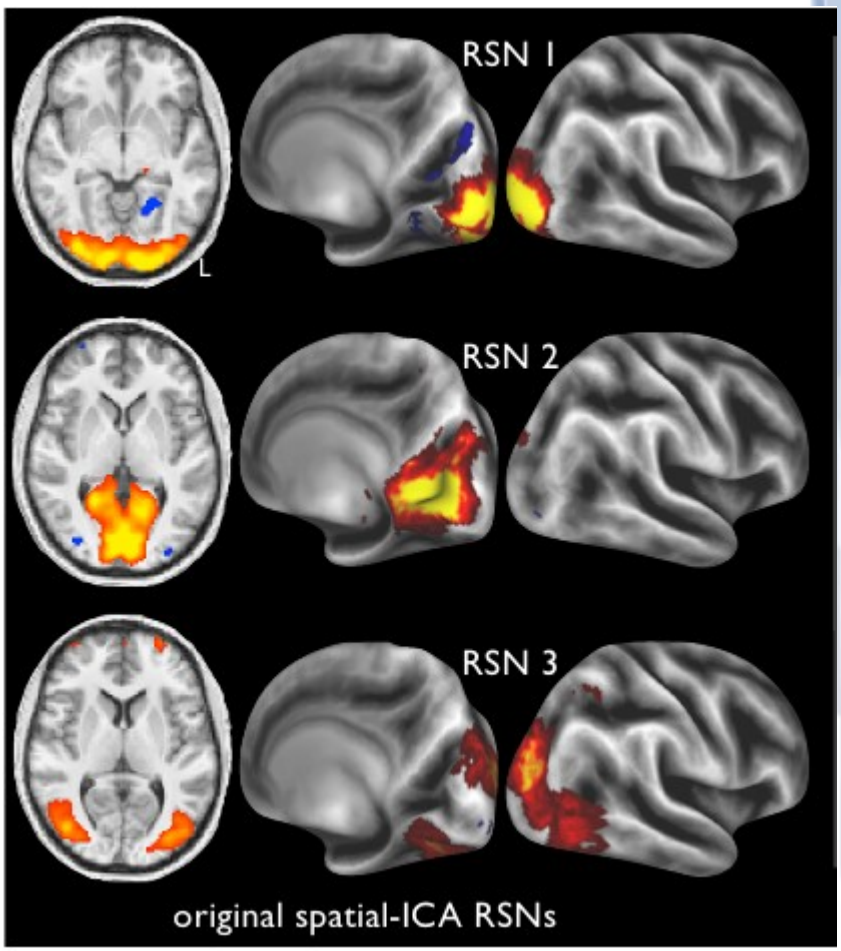
$$\mathrm{Kurt} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^4}{\sigma^4} - 3$$

where μ is i distribution and σ is a standard deviation.

ICA – brain research, signal separation



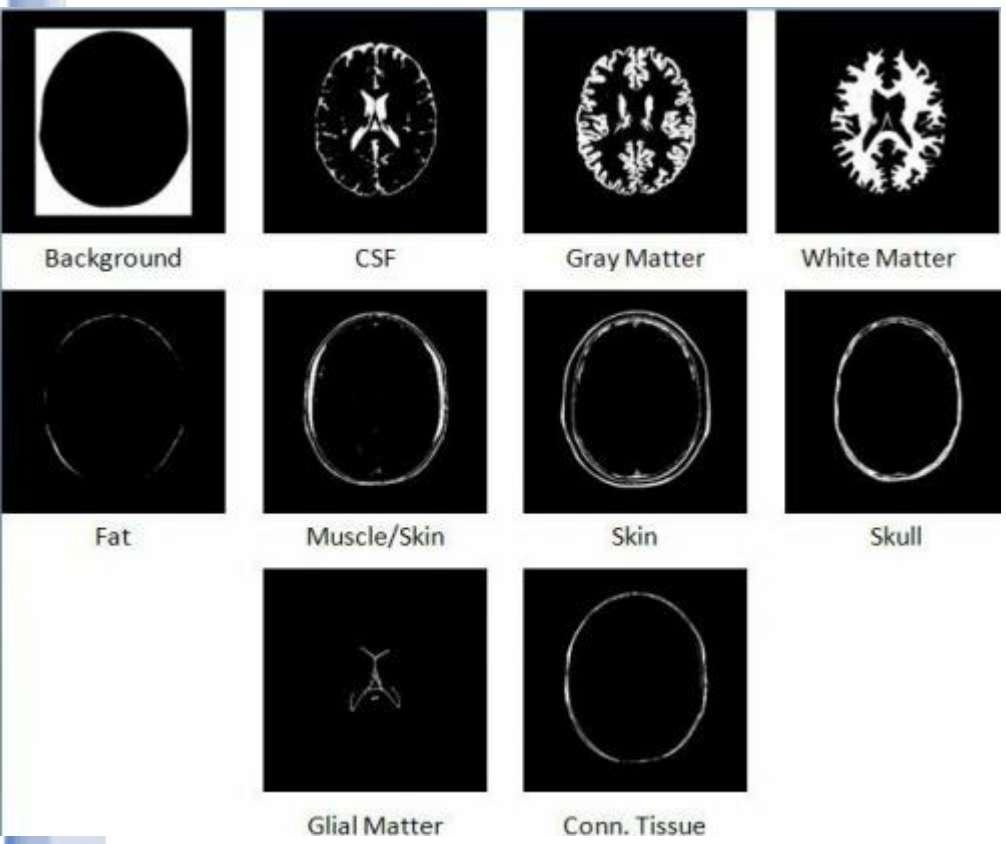




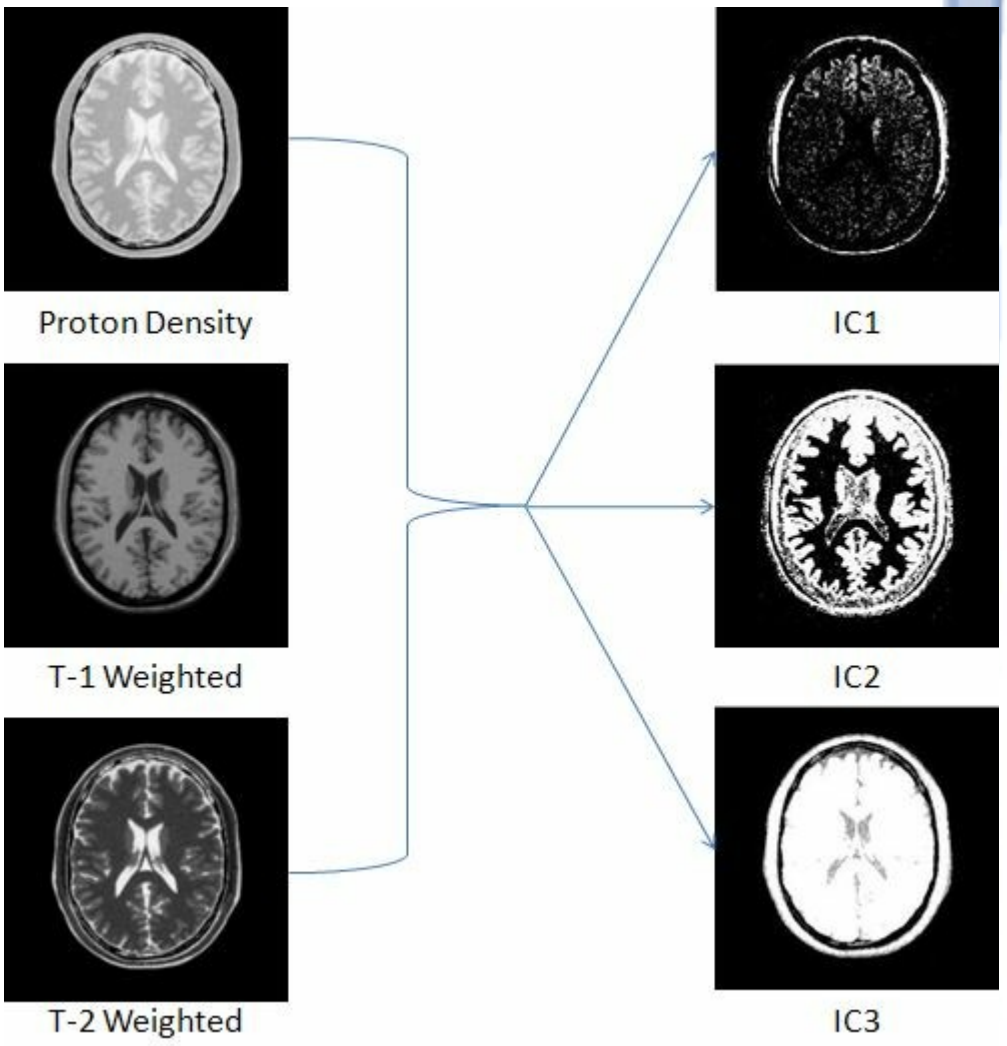
3 components from 21-dimensional decomposition using the “spatial-ICA'' algorithm.

PNAS February 21, 2012 vol. 109 no. 8 3131-3136

ICA and magnetic resonance



Sources of signals

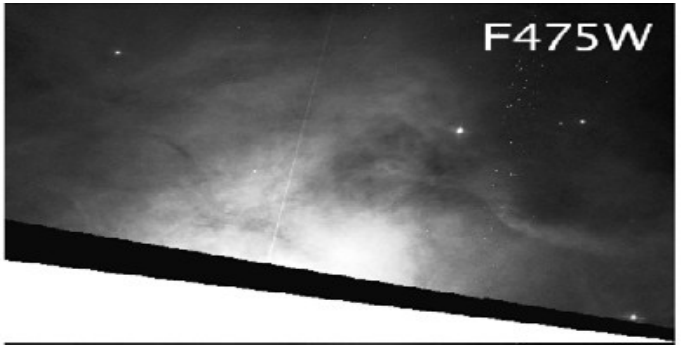


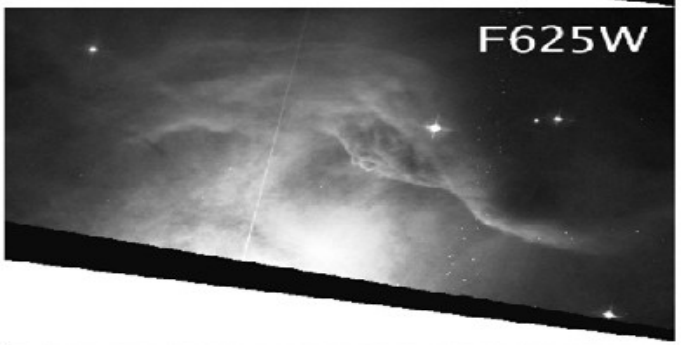
Measured

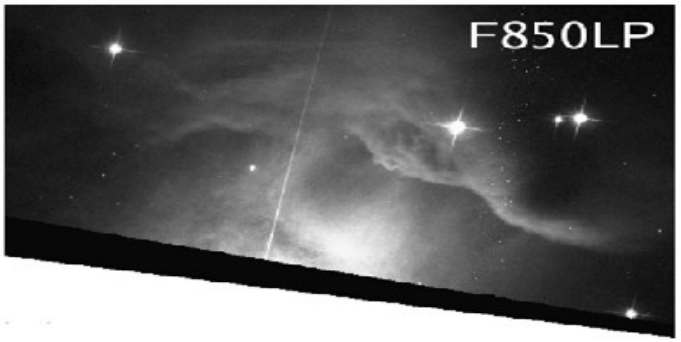
Separated components

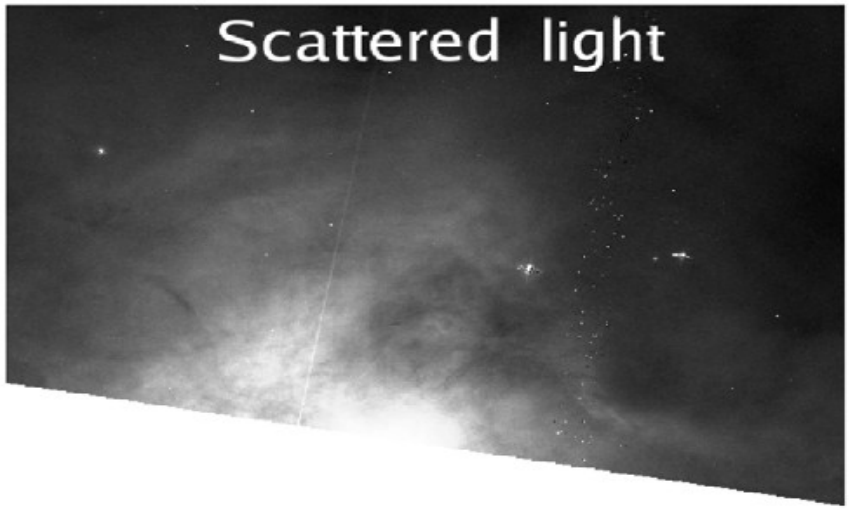
Blind Source Separation in Magnetic Resonance Images
January 30, 2010 by Shubhendu Trivedi

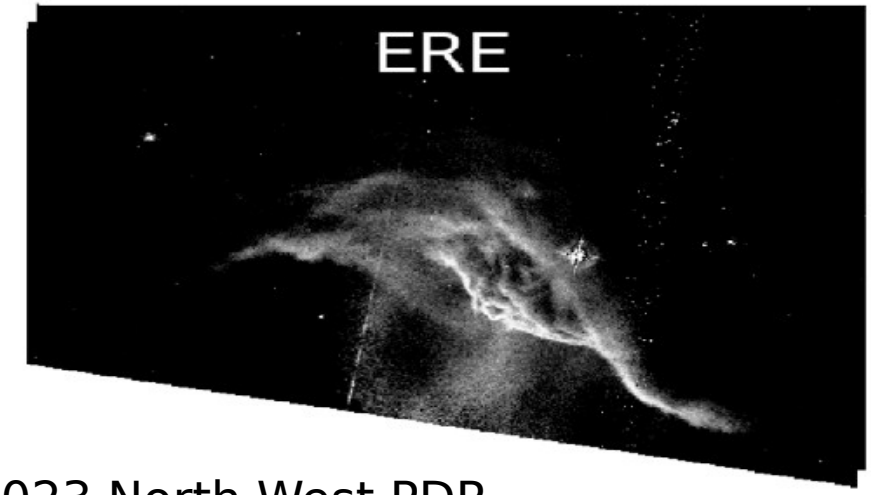
ICA – astronomy









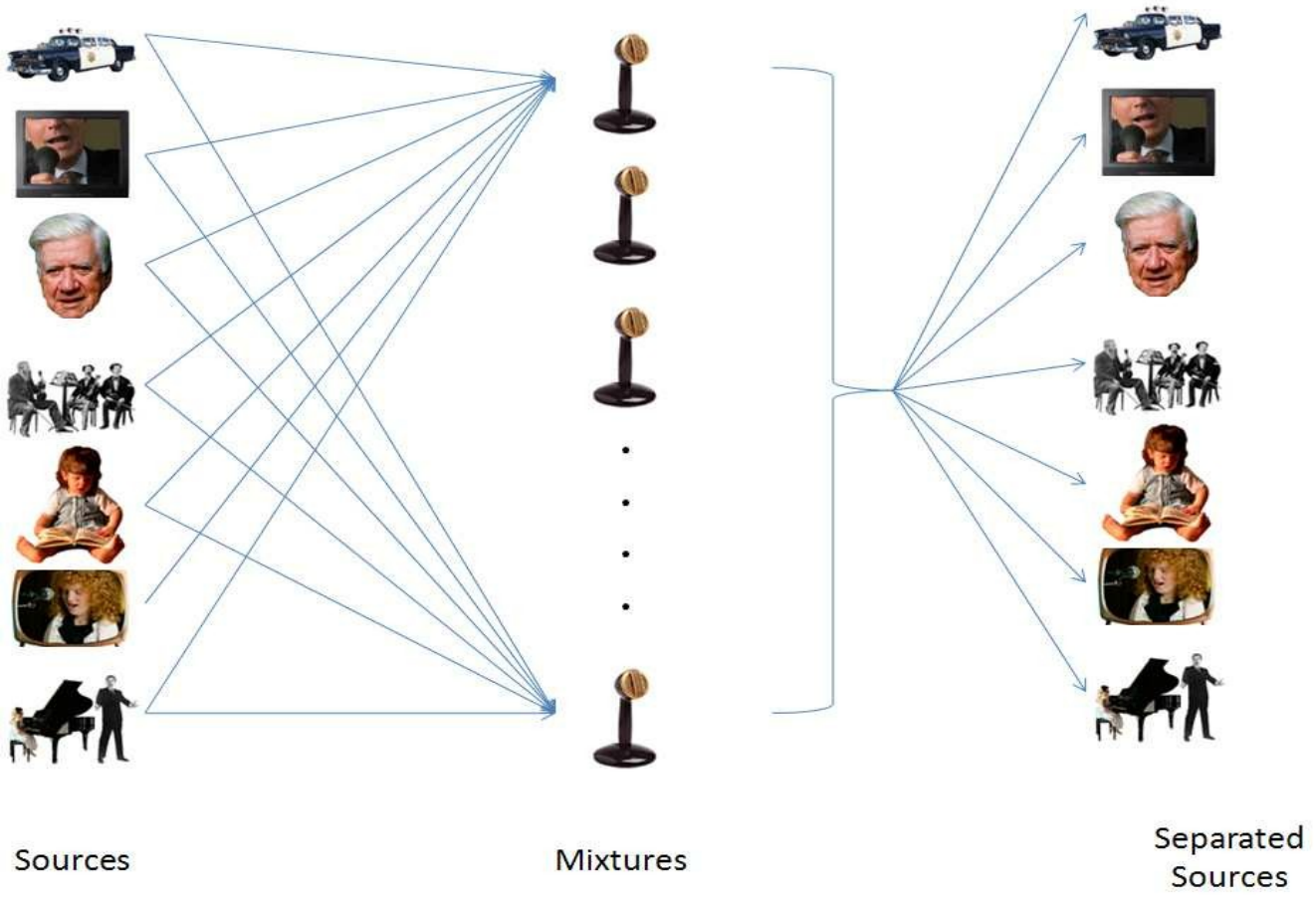


On the left: HST images of the NGC 7023 North-West PDR in three SDSS wide-band filters. On the right: scattered light and ERE (Extended Red Emission) images extracted with FastICA from the observations.

A&A 479, L41-L44 (2008)
DOI: 10.1051/0004-6361:20079158

Deser

- Cocktail Party Demo - applet pokazujący działanie algorytmu analizy składowych niezależnych na przykładzie problemu ,,przyjęcia cocktailowego".



Event Weighting

The probability ***p*(*S*|*x*)** is optimal in another sense:

If one ***weights*** an admixture of **signal** and **background** events by the weight function

$$W(x) = p(S|x)$$

then the ***signal*** strength will be extracted with ***zero bias*** and the ***smallest possible variance***, provided that our models describe the signal and background densities accurately and the signal to background ratio $p(S)/p(B)$ is equal to the true value.

We can recover signal by event weighting with the discriminant output without cutting on it.

Roger Barlow, J. Comp. Phys. 72, 202 (1987)