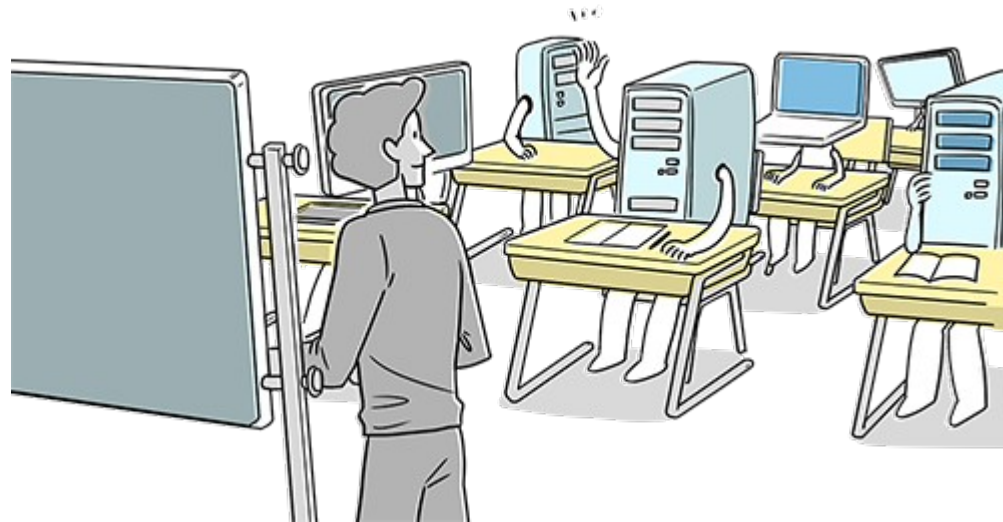


Machine learning

Lecture 1



Marcin Wolter

IFJ PAN

28 February 2018

- Machine learning: what does it mean?
- Software to work with and literature.
- A little bit of mathematics and examples of simple linear classifiers.
- Some examples

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 (unfortunately in Polish):

- 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 for real analysis, but we can learn together!

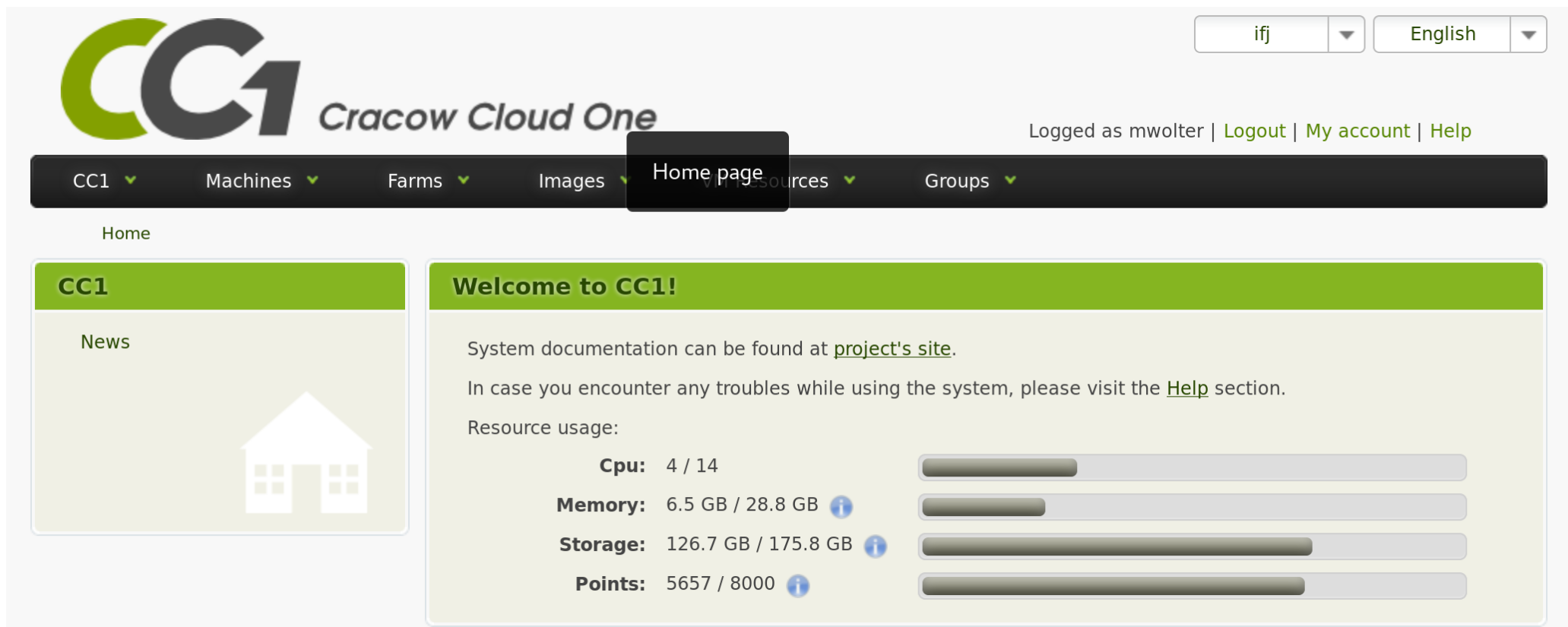
- <https://keras.io/>

Keras: The Python Deep Learning library

Emulates Deep Neural Network, uses google TensorFlow software

Computing

- <https://www.cloud.ifj.edu.pl/>
- Register, you can create your virtual linux box and play with it.
- Install root together with TMVA



The screenshot shows the CC1 Cracow Cloud One website. At the top right, there are dropdown menus for 'ifj' and 'English'. Below these, it says 'Logged as mwolter | Logout | My account | Help'. A dark navigation bar contains links for 'CC1', 'Machines', 'Farms', 'Images', 'Home page' (highlighted with a tooltip), 'Virtual Resources', and 'Groups'. The main content area is divided into two columns. The left column has a 'CC1' header and a 'News' section with a house icon. The right column has a 'Welcome to CC1!' header, followed by text about system documentation and a 'Help' section. Below this, it shows 'Resource usage' with four rows: 'Cpu: 4 / 14', 'Memory: 6.5 GB / 28.8 GB', 'Storage: 126.7 GB / 175.8 GB', and 'Points: 5657 / 8000'. Each row has a corresponding progress bar.

Statistics

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



١٢٠
 ١٢١
 ١٢٢
 ١٢٣
 ١٢٤
 ١٢٥
 ١٢٦
 ١٢٧
 ١٢٨
 ١٢٩
 ١٣٠
 ١٣١
 ١٣٢
 ١٣٣
 ١٣٤
 ١٣٥
 ١٣٦
 ١٣٧
 ١٣٨
 ١٣٩
 ١٤٠
 ١٤١
 ١٤٢
 ١٤٣
 ١٤٤
 ١٤٥
 ١٤٦
 ١٤٧
 ١٤٨
 ١٤٩
 ١٥٠
 ١٥١
 ١٥٢
 ١٥٣
 ١٥٤
 ١٥٥
 ١٥٦
 ١٥٧
 ١٥٨
 ١٥٩
 ١٦٠
 ١٦١
 ١٦٢
 ١٦٣
 ١٦٤
 ١٦٥
 ١٦٦
 ١٦٧
 ١٦٨
 ١٦٩
 ١٧٠
 ١٧١
 ١٧٢
 ١٧٣
 ١٧٤
 ١٧٥
 ١٧٦
 ١٧٧
 ١٧٨
 ١٧٩
 ١٨٠
 ١٨١
 ١٨٢
 ١٨٣
 ١٨٤
 ١٨٥
 ١٨٦
 ١٨٧
 ١٨٨
 ١٨٩
 ١٩٠
 ١٩١
 ١٩٢
 ١٩٣
 ١٩٤
 ١٩٥
 ١٩٦
 ١٩٧
 ١٩٨
 ١٩٩
 ٢٠٠
 ٢٠١
 ٢٠٢
 ٢٠٣
 ٢٠٤
 ٢٠٥
 ٢٠٦
 ٢٠٧
 ٢٠٨
 ٢٠٩
 ٢١٠
 ٢١١
 ٢١٢
 ٢١٣
 ٢١٤
 ٢١٥
 ٢١٦
 ٢١٧
 ٢١٨
 ٢١٩
 ٢٢٠
 ٢٢١
 ٢٢٢
 ٢٢٣
 ٢٢٤
 ٢٢٥
 ٢٢٦
 ٢٢٧
 ٢٢٨
 ٢٢٩
 ٢٣٠
 ٢٣١
 ٢٣٢
 ٢٣٣
 ٢٣٤
 ٢٣٥
 ٢٣٦
 ٢٣٧
 ٢٣٨
 ٢٣٩
 ٢٤٠
 ٢٤١
 ٢٤٢
 ٢٤٣
 ٢٤٤
 ٢٤٥
 ٢٤٦
 ٢٤٧
 ٢٤٨
 ٢٤٩
 ٢٥٠
 ٢٥١
 ٢٥٢
 ٢٥٣
 ٢٥٤
 ٢٥٥
 ٢٥٦
 ٢٥٧
 ٢٥٨
 ٢٥٩
 ٢٦٠
 ٢٦١
 ٢٦٢
 ٢٦٣
 ٢٦٤
 ٢٦٥
 ٢٦٦
 ٢٦٧
 ٢٦٨
 ٢٦٩
 ٢٧٠
 ٢٧١
 ٢٧٢
 ٢٧٣
 ٢٧٤
 ٢٧٥
 ٢٧٦
 ٢٧٧
 ٢٧٨
 ٢٧٩
 ٢٨٠
 ٢٨١
 ٢٨٢
 ٢٨٣
 ٢٨٤
 ٢٨٥
 ٢٨٦
 ٢٨٧
 ٢٨٨
 ٢٨٩
 ٢٩٠
 ٢٩١
 ٢٩٢
 ٢٩٣
 ٢٩٤
 ٢٩٥
 ٢٩٦
 ٢٩٧
 ٢٩٨
 ٢٩٩
 ٣٠٠
 ٣٠١
 ٣٠٢
 ٣٠٣
 ٣٠٤
 ٣٠٥
 ٣٠٦
 ٣٠٧
 ٣٠٨
 ٣٠٩
 ٣١٠
 ٣١١
 ٣١٢
 ٣١٣
 ٣١٤
 ٣١٥
 ٣١٦
 ٣١٧
 ٣١٨
 ٣١٩
 ٣٢٠
 ٣٢١
 ٣٢٢
 ٣٢٣
 ٣٢٤
 ٣٢٥
 ٣٢٦
 ٣٢٧
 ٣٢٨
 ٣٢٩
 ٣٣٠
 ٣٣١
 ٣٣٢
 ٣٣٣
 ٣٣٤
 ٣٣٥
 ٣٣٦
 ٣٣٧
 ٣٣٨
 ٣٣٩
 ٣٤٠
 ٣٤١
 ٣٤٢
 ٣٤٣
 ٣٤٤
 ٣٤٥
 ٣٤٦
 ٣٤٧
 ٣٤٨
 ٣٤٩
 ٣٥٠
 ٣٥١
 ٣٥٢
 ٣٥٣
 ٣٥٤
 ٣٥٥
 ٣٥٦
 ٣٥٧
 ٣٥٨
 ٣٥٩
 ٣٦٠
 ٣٦١
 ٣٦٢
 ٣٦٣
 ٣٦٤
 ٣٦٥
 ٣٦٦
 ٣٦٧
 ٣٦٨
 ٣٦٩
 ٣٧٠
 ٣٧١
 ٣٧٢
 ٣٧٣
 ٣٧٤
 ٣٧٥
 ٣٧٦
 ٣٧٧
 ٣٧٨
 ٣٧٩
 ٣٨٠
 ٣٨١
 ٣٨٢
 ٣٨٣
 ٣٨٤
 ٣٨٥
 ٣٨٦
 ٣٨٧
 ٣٨٨
 ٣٨٩
 ٣٩٠
 ٣٩١
 ٣٩٢
 ٣٩٣
 ٣٩٤
 ٣٩٥
 ٣٩٦
 ٣٩٧
 ٣٩٨
 ٣٩٩
 ٤٠٠
 ٤٠١
 ٤٠٢
 ٤٠٣
 ٤٠٤
 ٤٠٥
 ٤٠٦
 ٤٠٧
 ٤٠٨
 ٤٠٩
 ٤١٠
 ٤١١
 ٤١٢
 ٤١٣
 ٤١٤
 ٤١٥
 ٤١٦
 ٤١٧
 ٤١٨
 ٤١٩
 ٤٢٠
 ٤٢١
 ٤٢٢
 ٤٢٣
 ٤٢٤
 ٤٢٥
 ٤٢٦
 ٤٢٧
 ٤٢٨
 ٤٢٩
 ٤٣٠
 ٤٣١
 ٤٣٢
 ٤٣٣
 ٤٣٤
 ٤٣٥
 ٤٣٦
 ٤٣٧
 ٤٣٨
 ٤٣٩
 ٤٤٠
 ٤٤١
 ٤٤٢
 ٤٤٣
 ٤٤٤
 ٤٤٥
 ٤٤٦
 ٤٤٧
 ٤٤٨
 ٤٤٩
 ٤٥٠
 ٤٥١
 ٤٥٢
 ٤٥٣
 ٤٥٤
 ٤٥٥
 ٤٥٦
 ٤٥٧
 ٤٥٨
 ٤٥٩
 ٤٦٠
 ٤٦١
 ٤٦٢
 ٤٦٣
 ٤٦٤
 ٤٦٥
 ٤٦٦
 ٤٦٧
 ٤٦٨
 ٤٦٩
 ٤٧٠
 ٤٧١
 ٤٧٢
 ٤٧٣
 ٤٧٤
 ٤٧٥
 ٤٧٦
 ٤٧٧
 ٤٧٨
 ٤٧٩
 ٤٨٠
 ٤٨١
 ٤٨٢
 ٤٨٣
 ٤٨٤
 ٤٨٥
 ٤٨٦
 ٤٨٧
 ٤٨٨
 ٤٨٩
 ٤٩٠
 ٤٩١

الحمد لله - ولله العاقبة والآخرين صلوات الله على سيدنا محمد وآله

بسم الله الرحمن الرحيم
 رسالة الى من يشاء من عباده
 في بيان ما هو الحق وما هو الباطل
 وما هو الخير وما هو الشر
 وما هو النجاة وما هو الضلال
 وما هو العلم وما هو الجهل
 وما هو القوة وما هو الضعف
 وما هو الشجاعة وما هو الخوف
 وما هو الشرف وما هو الذل
 وما هو العزة وما هو الهوان
 وما هو الكرامة وما هو المذلة
 وما هو النجاة وما هو الضلال
 وما هو العلم وما هو الجهل
 وما هو القوة وما هو الضعف
 وما هو الشجاعة وما هو الخوف
 وما هو الشرف وما هو الذل
 وما هو العزة وما هو الهوان
 وما هو الكرامة وما هو المذلة

Caesar cipher

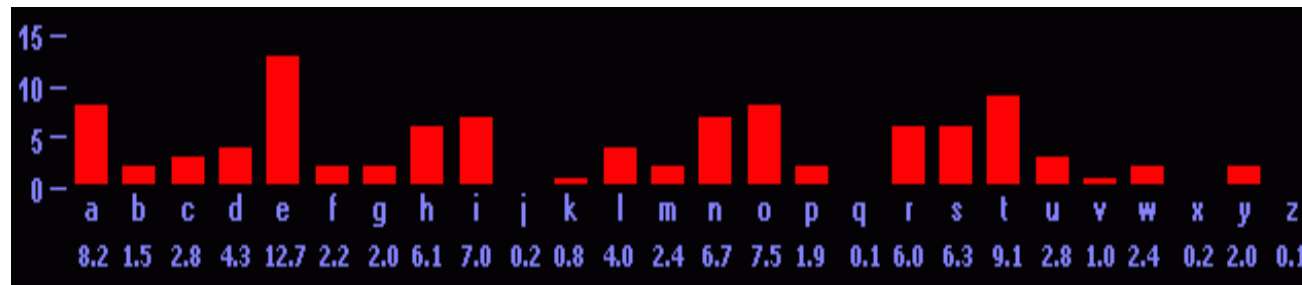
- Each letter of a text is replaced by another, shifted by n letters:

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

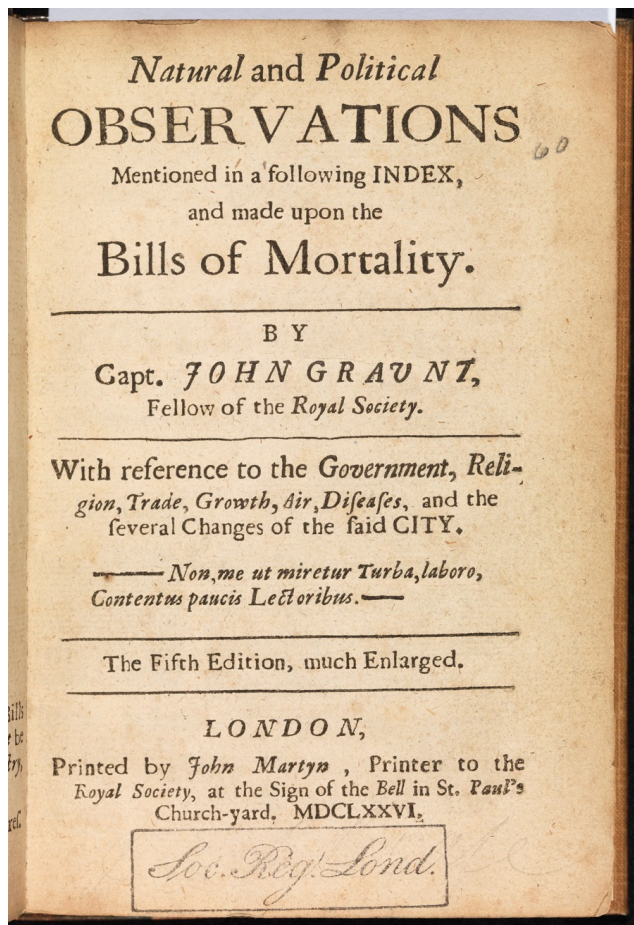
- Frequency analysis – the frequency of appearance of different letters is investigated.



- 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*“).



The Table of CASUALTIES.																																	1629	1633	1647	1651	1655	In 2																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																											
																																	1630	1634	1648	1652	1656	1629	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648	1652	1656	1630	1634	1648</

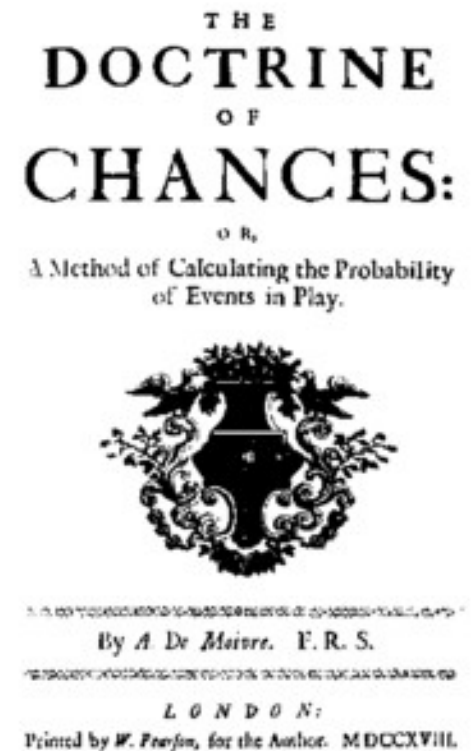
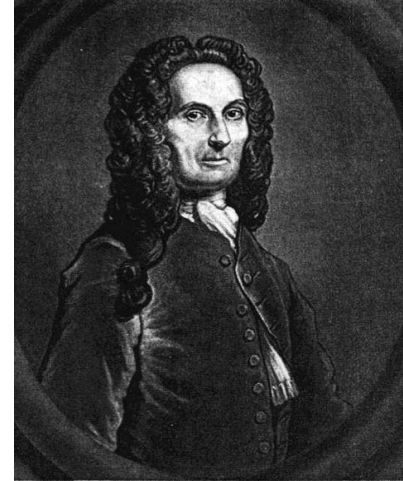
How to define the probability?

- 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:**

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

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.

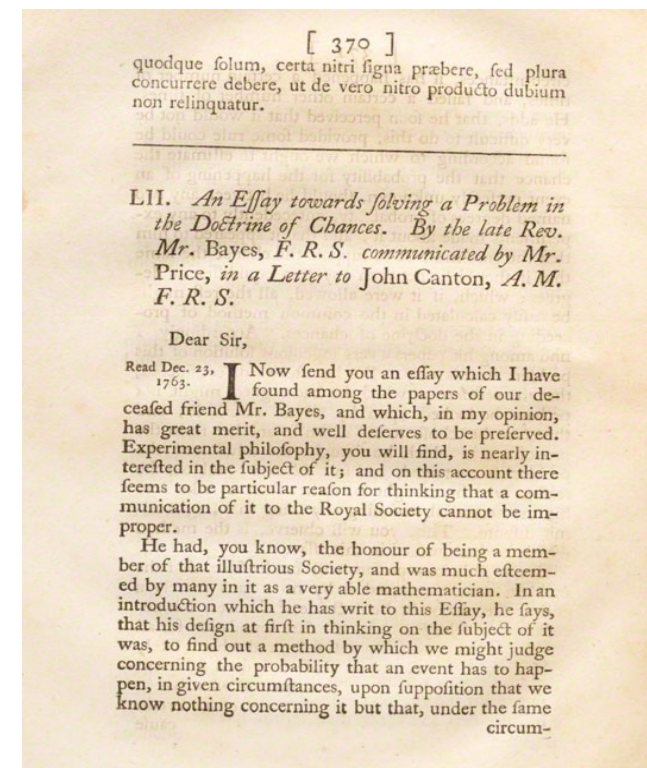
- The definition comes from a text book of Abraham de Moivre (1667-1754) – a text book on statistics „*The Doctrine of Chances*” (1718).



Bayes definition



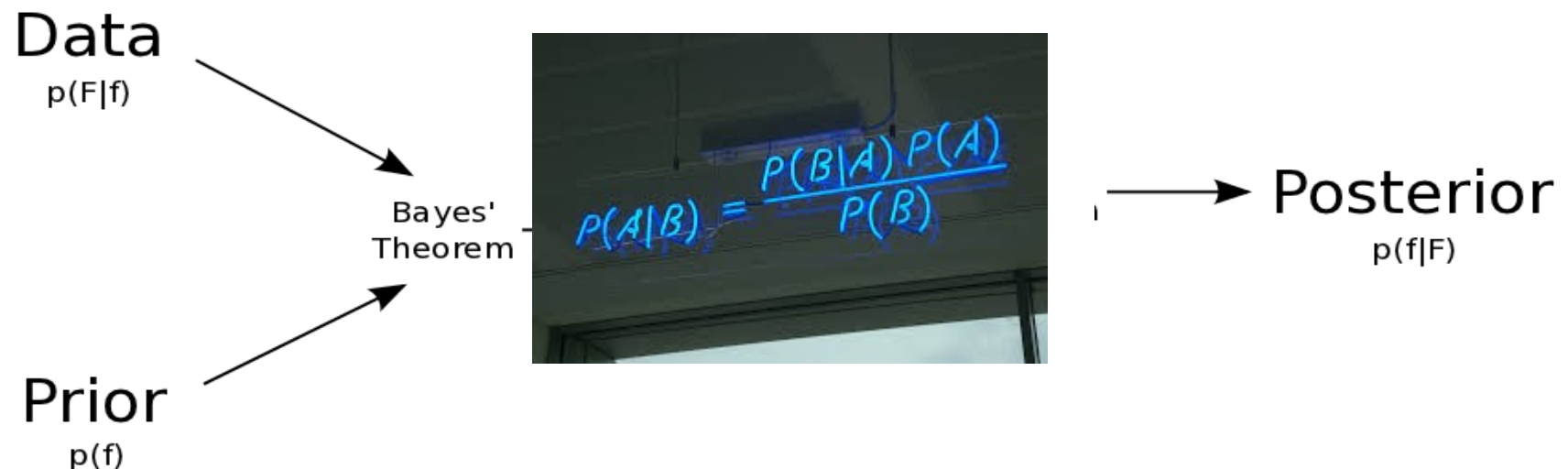
- Probability “a priori”, i.e. unconditional, is understood as a measure of belief, based on rational evidence, that such an event will happen.
- In the next step we make an experiment, called “observations”, and their results modify the probability. We get the probability “a posteriori”, which is a 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. They stated, that this is a way we recognize and study the nature.



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: $\frac{1}{2}$ (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: a cancer test

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

- $\Pr(A|X)$ = 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, let say 80% in our case.
- $\Pr(A)$ = Chance of having cancer (1%).
- $\Pr(\text{not } A)$ = Chance of not having cancer (99%).
- $\Pr(X|\text{not } A)$ = Chance of a positive test (X) given that you didn't have cancer (not 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.

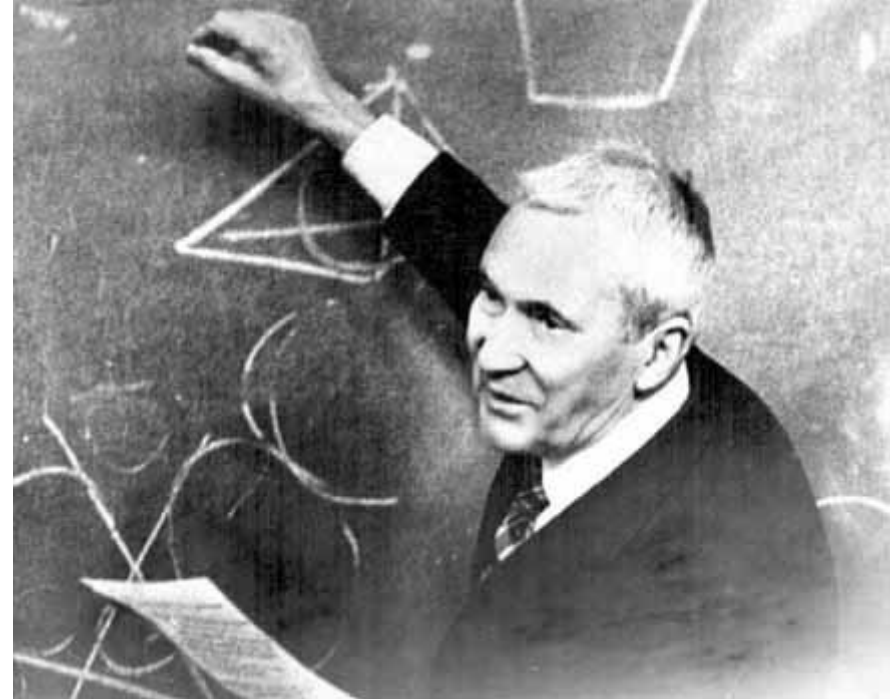
“We don’t know all about the world to start with; our knowledge by experience consists simply of a rather scattered lot of sensations, and we cannot get any further without some a priori postulates. My problem is to get these stated as clearly as possible.”

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...



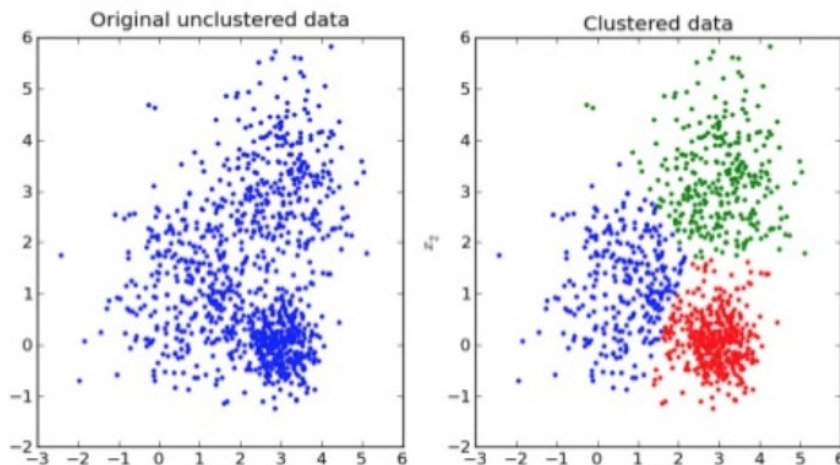
Андрей Никола́евич Колмогоров (1903-1987)

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:

- $P(A) \geq 0$ 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.

What does “machine learning” mean?

- **Machine learning** is a field of computer science that gives computer systems the ability to "learn" (i.e. progressively improve performance on a specific task) with data, without being explicitly programmed.
- Problems:
 - Supervised learning (classification & regression)
 - Clustering (unsupervised learning)
 - Dimensionality reduction
 - Reinforcement learning
 - Many others....



➤ Unsupervised Learning

- ❑ Technique of trying to find hidden structure in unlabeled data

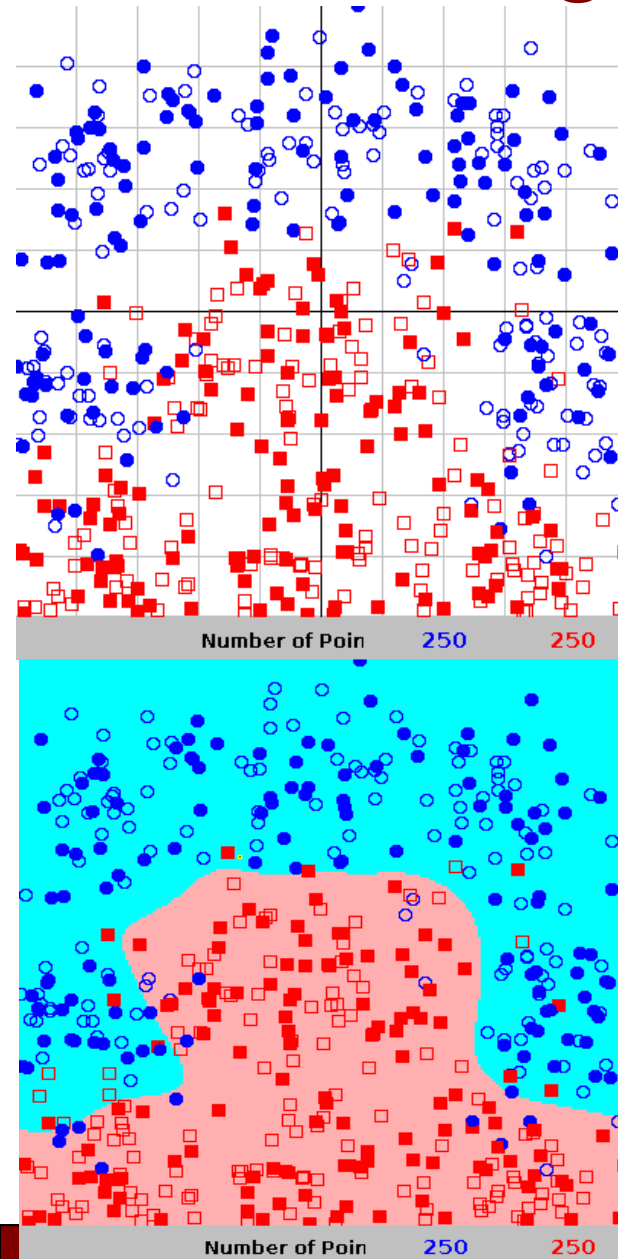
➤ Supervise Learning

- ❑ Technique for creating a function from training data. The training data consist of pairs of input objects (typically vectors), and desired outputs.

How do the (supervised) machine learning algorithms work?

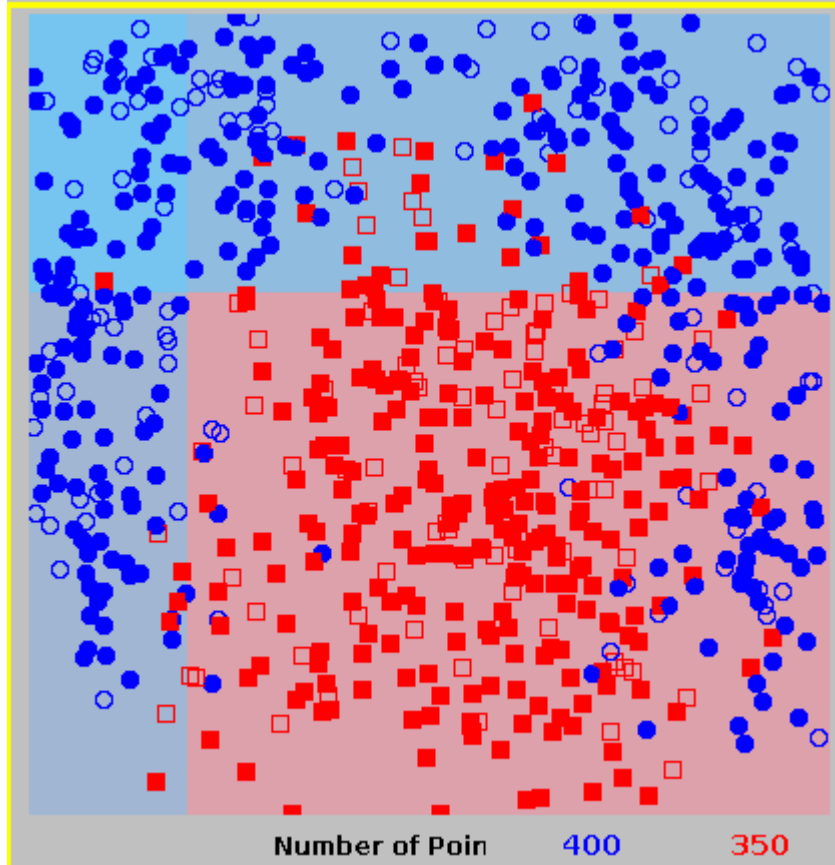
- We need **training data**, for which we know the correct answer, whether it's a signal or background. We divide the data into two samples: training and test.
- We find the best function $f(\mathbf{x})$ which describes the probability, that a given event belongs to the class "signal". This is done by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

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



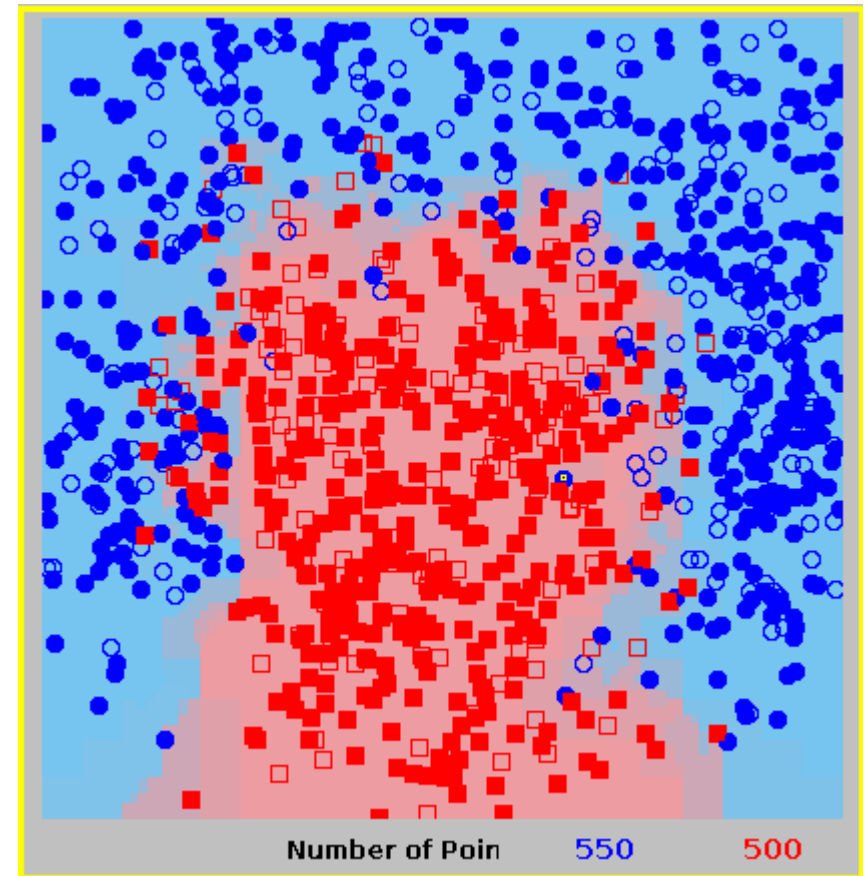
Cuts vs non-linear separation

Cuts



?

Non-linear separation



Neural Networks, boosted decision trees, and so on....

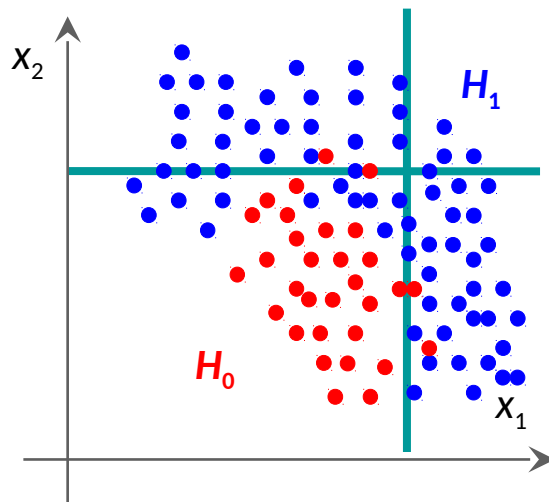
Types of algorithms



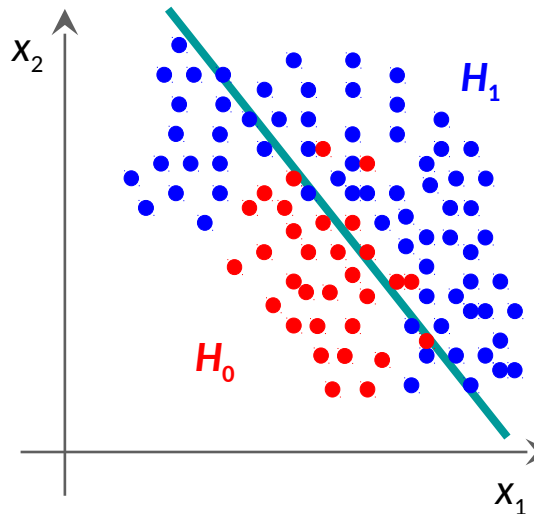
How to use the information available

- **Classification:** find a function $f(x_1, x_2)$ giving the probability, that a given data point belongs to a given class (signal vs background).

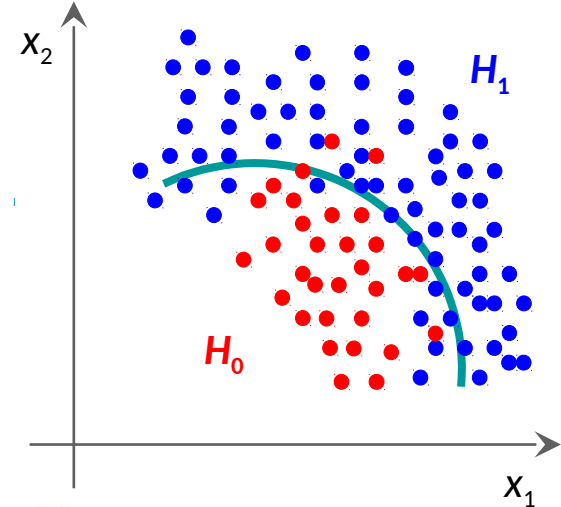
Simple cuts
(easy and intuitive)



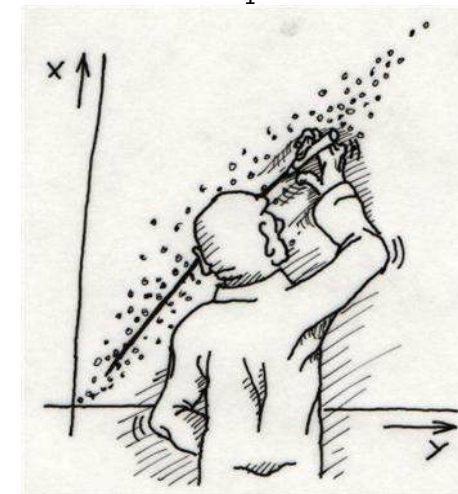
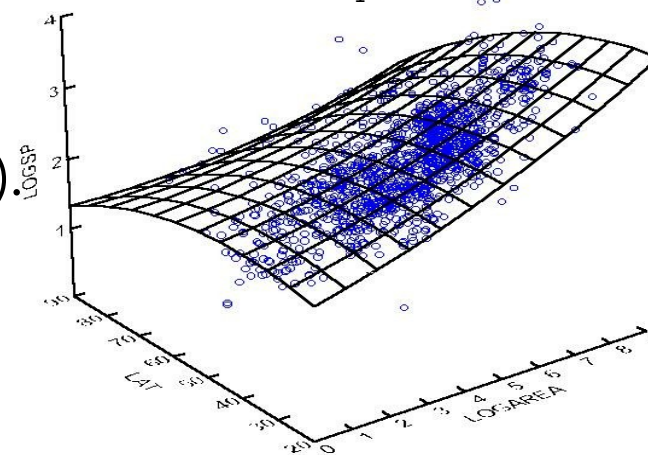
Linear
(fast and stable)



Non-linear
(most effective)



Regression: fit a continuous function
(find particle energy from calo readouts).



Classification

A Bayes classifier:

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

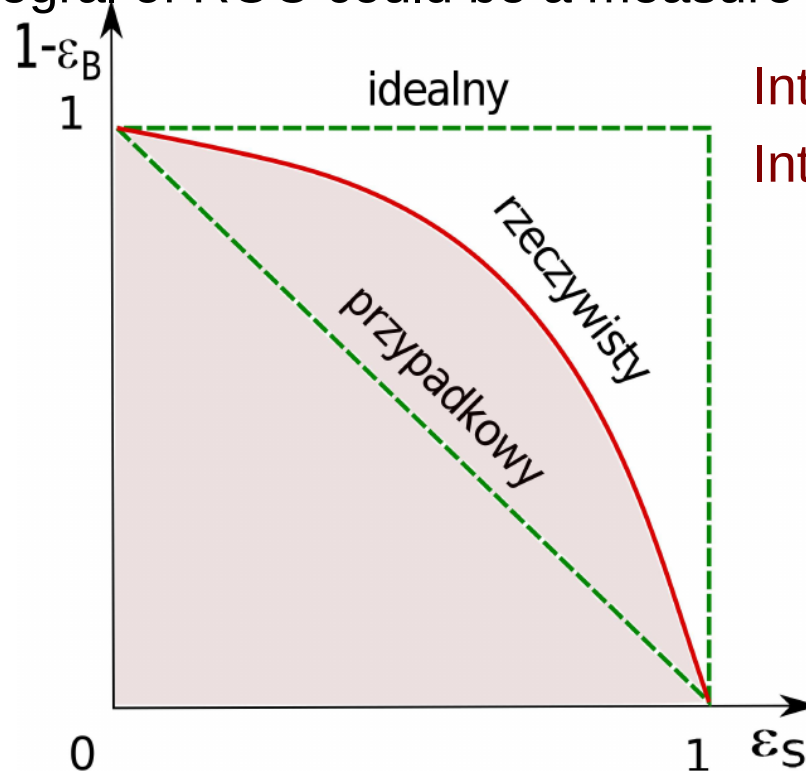
where **S** is associated with $y = 1$ and **B** with $y = 0$. **Bayes classifier** accepts events x if $p(\mathbf{S}|x) > \mathbf{cut}$ as belonging to **S**.

We need to approximate probability distributions $P(x|\mathbf{S})$ and $P(x|\mathbf{B})$.

- 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_S . Shows how good the classifier is.
- The integral of ROC could be a measure of the classifier quality:



Integral(ROC) = $\frac{1}{2}$ – random

Integral(ROC) = 1 – ideal

Practical applications

A Short List of Multivariate Methods

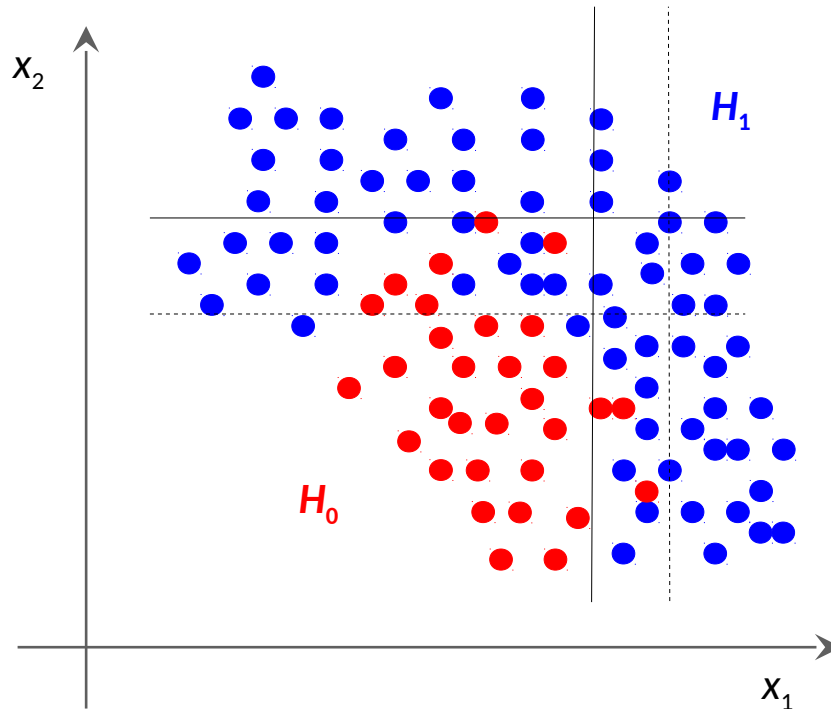
- Cuts
- Linear Discriminants (like Fisher)
- Support Vector Machines
- Naive Bayes (Likelihood Discriminant)
- Kernel Density Estimation
- Decision Trees
- Neural Networks
- Bayesian Neural Networks
- Genetic Algorithms

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

We will talk today about:

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

Cuts



Optimization of cuts:

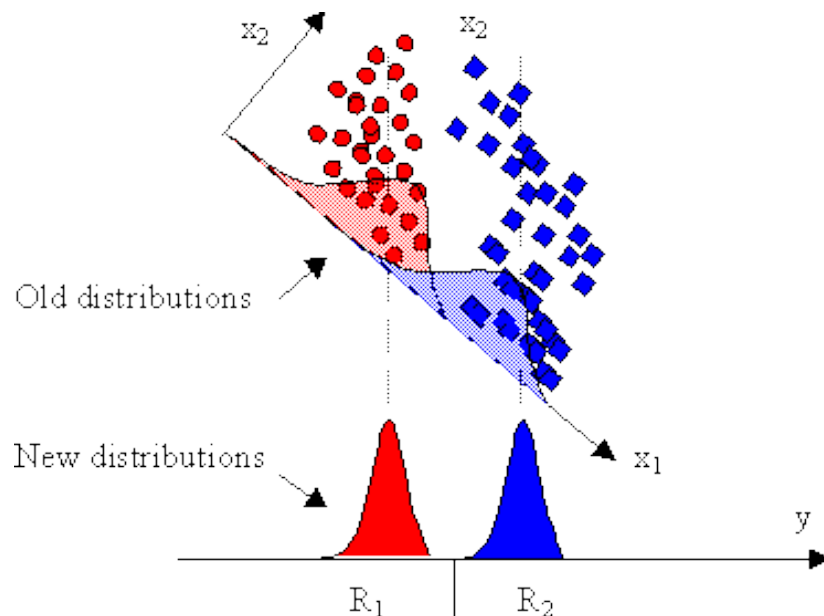
- Move cuts as long as we get the optimal signal vs. background selection. For a given signal efficiency we find the best background rejection → we get the entire ROC curve.
- Optimization methods:
 - Brute force
 - Genetic algorithms
 - Many others...

Fisher discriminants

LDA, Linear Discriminant Analysis

Projection to one dimension, then 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:

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

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

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

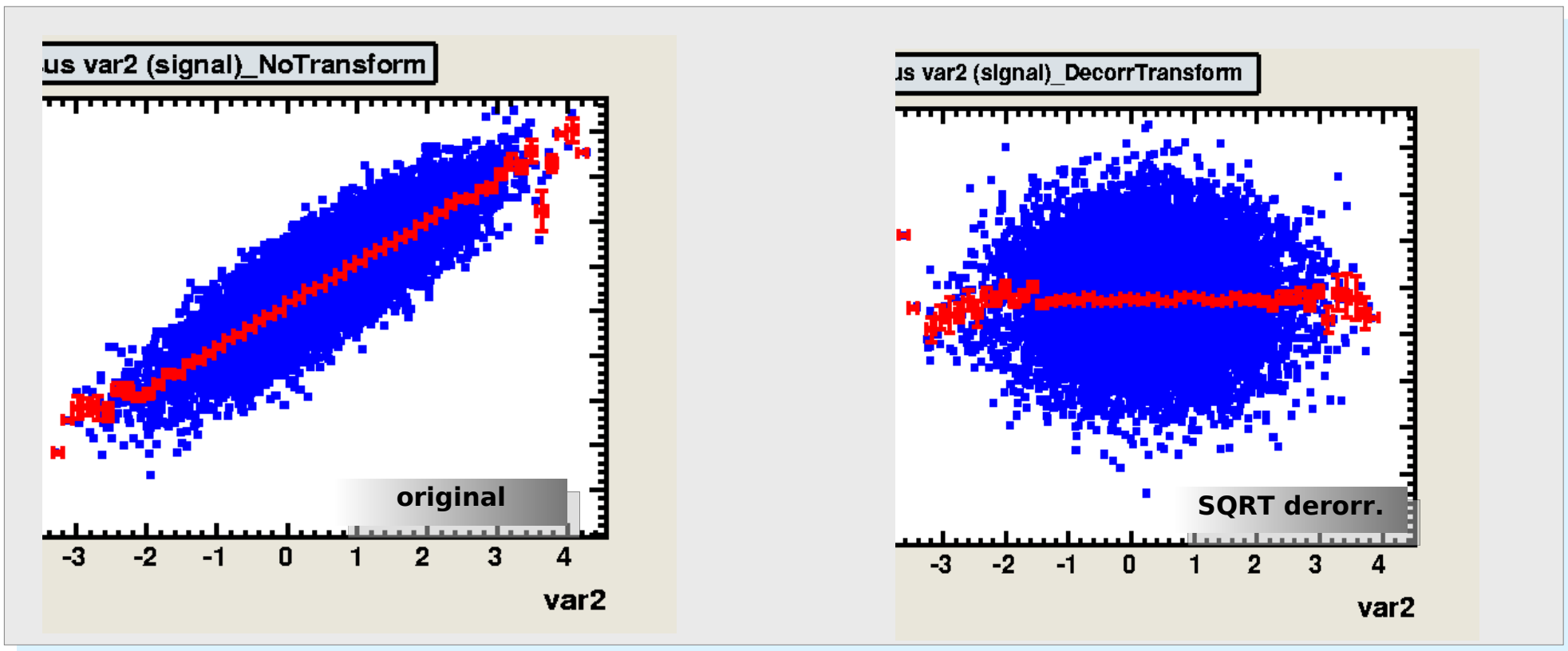
When the assumptions of LDA are satisfied, the above equation is equivalent to LDA.

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

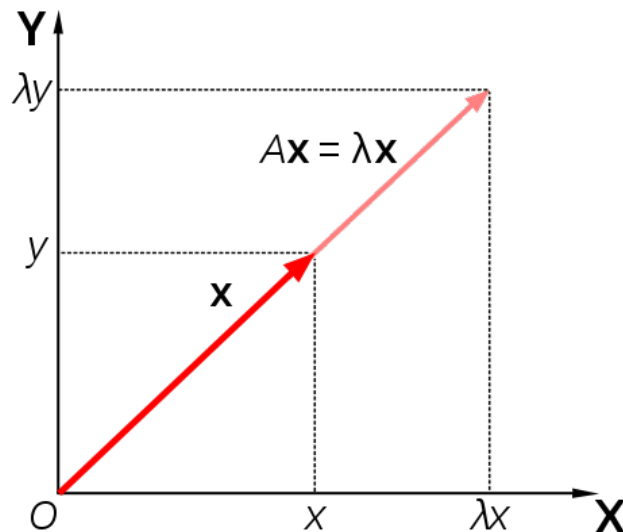


Eigenvalues and eigenvectors

In essence, an eigenvector \mathbf{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 \mathbf{x} , not changing its direction, so \mathbf{x} is an eigenvector of A .

$$\begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} = \underbrace{\begin{bmatrix} | & | & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \\ | & | & | \end{bmatrix}}_{\substack{\text{Eigen vectors} \\ \text{of} \\ A}} \underbrace{\begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}}_{\substack{\text{Eigen values} \\ \text{of} \\ A}} \underbrace{\begin{bmatrix} | & | & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \\ | & | & | \end{bmatrix}^{-1}}_{\substack{\text{Eigen vectors} \\ \text{of} \\ A}^{-1}}$$

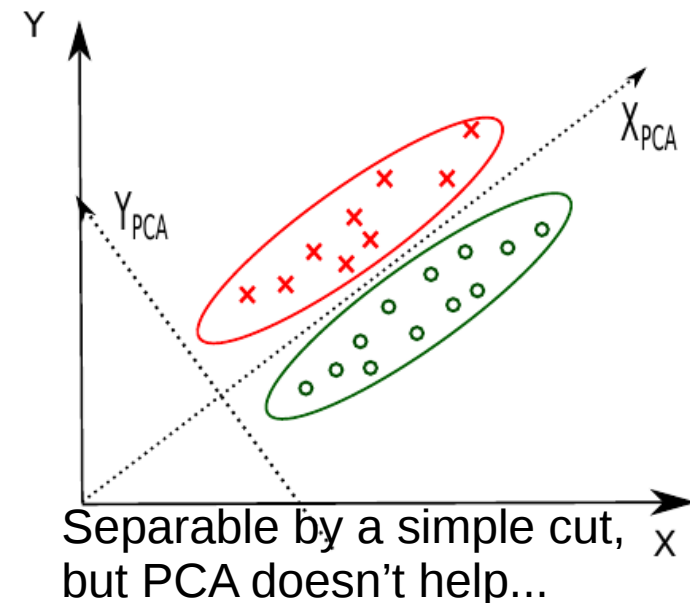
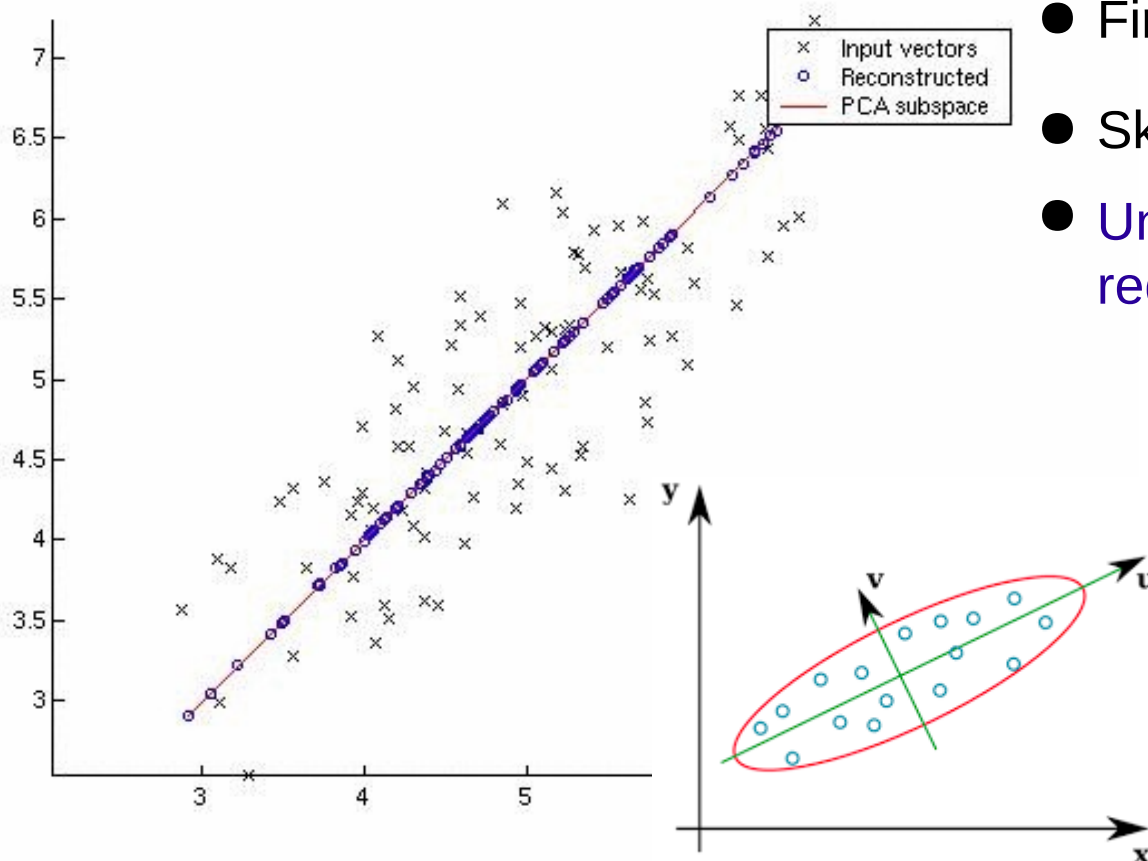
Eigendecomposition of a matrix

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 might be skipped

Procedure:

- Find the covariance matrix $\text{Cov}(X)$
- Find eigenvalues λ_i and eigenvectors v_i
- Skip smallest λ_i
- Unsupervised learning & dimensionality reduction



Independent Component Analysis ICA

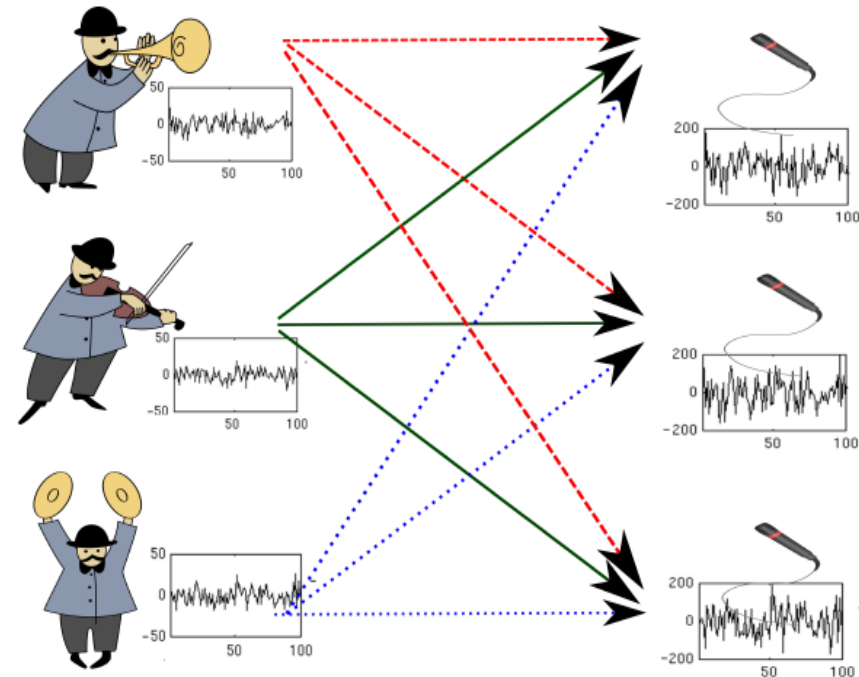
Developed at Helsinki University of Technology <http://www.cis.hut.fi/projects/ica/>

● Problem:

- Assume, that signal \mathbf{X} is a linear combination $\mathbf{X} = \mathbf{A}\mathbf{S}$ of independent sources \mathbf{S} . The mixing matrix \mathbf{A} and vector of sources \mathbf{S} are unknown.
- **Task:** find a matrix \mathbf{T} (inverted \mathbf{A}), such that elements of vector $\mathbf{U} = \mathbf{T}\mathbf{X}$ are statistically independent. \mathbf{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.





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, $\sigma=1$).

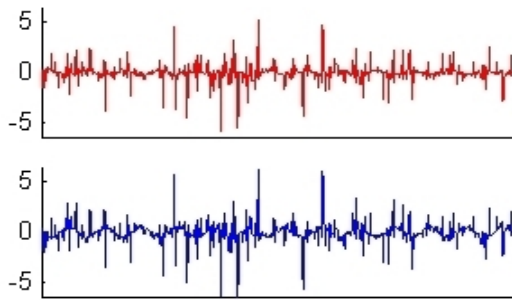
Superposition of many independent distributions gives Gaussian in the limit.

- ICA – rotation, signals should be maximally non-Gaussian (measure of non-Gaussianity might be kurtosis).

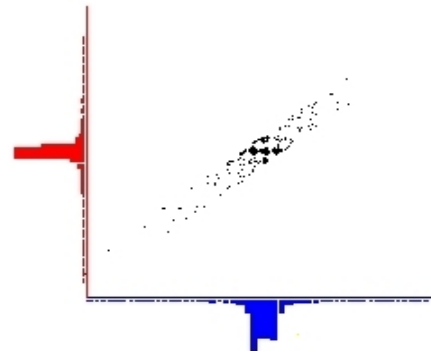
- *Curtosis:*
$$\text{Kurt} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^4}{\sigma^4} - 3$$

where μ is the mean of the distribution and σ is a standard deviation.

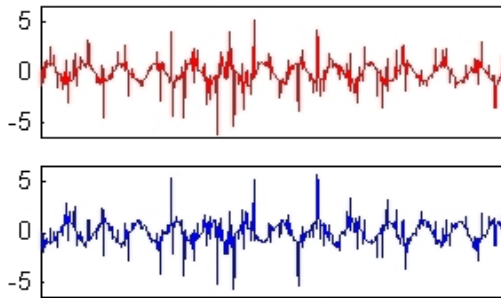
SIGNALS



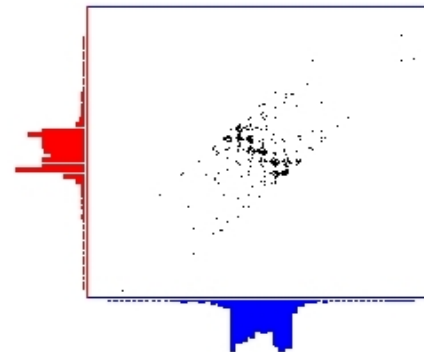
JOINT DENSITY



SIGNALS

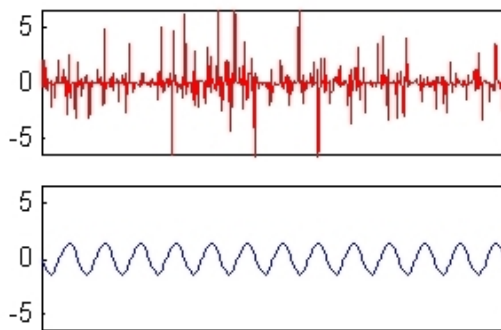


JOINT DENSITY

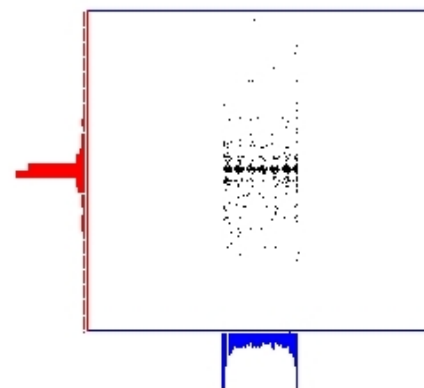


Whitened signals and density

SIGNALS

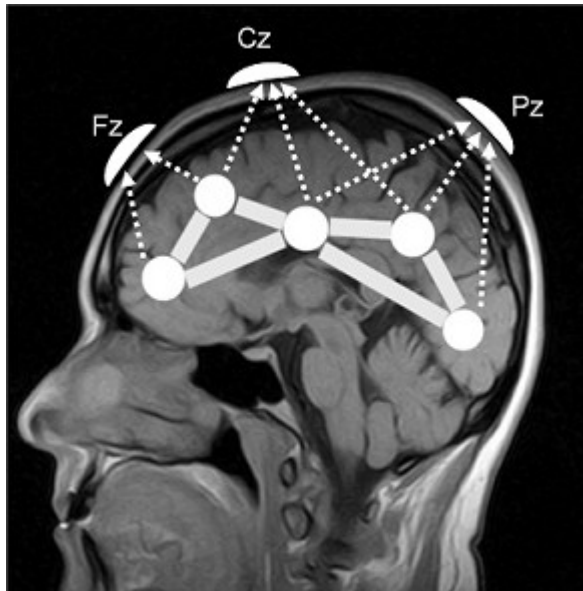


JOINT DENSITY

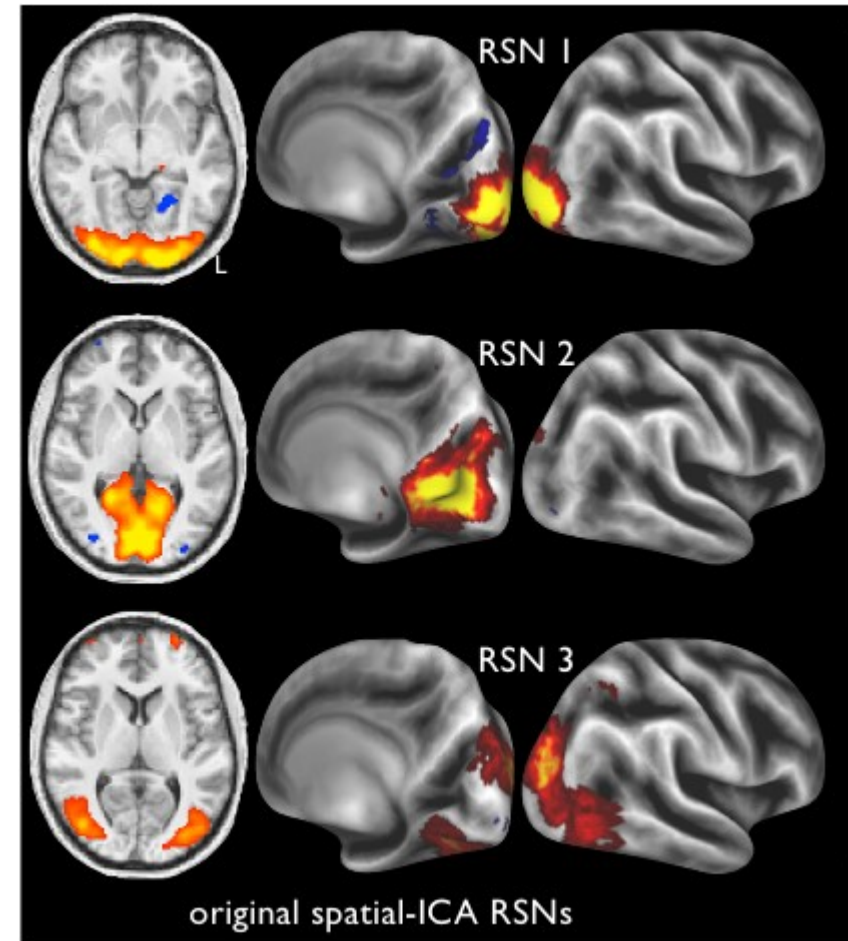
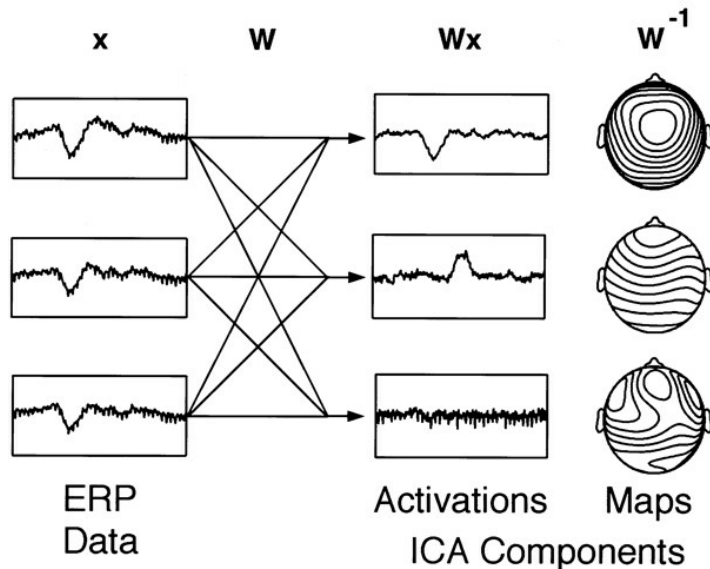


Separated signals after 5 steps of FastICA

ICA – brain research, signal separation



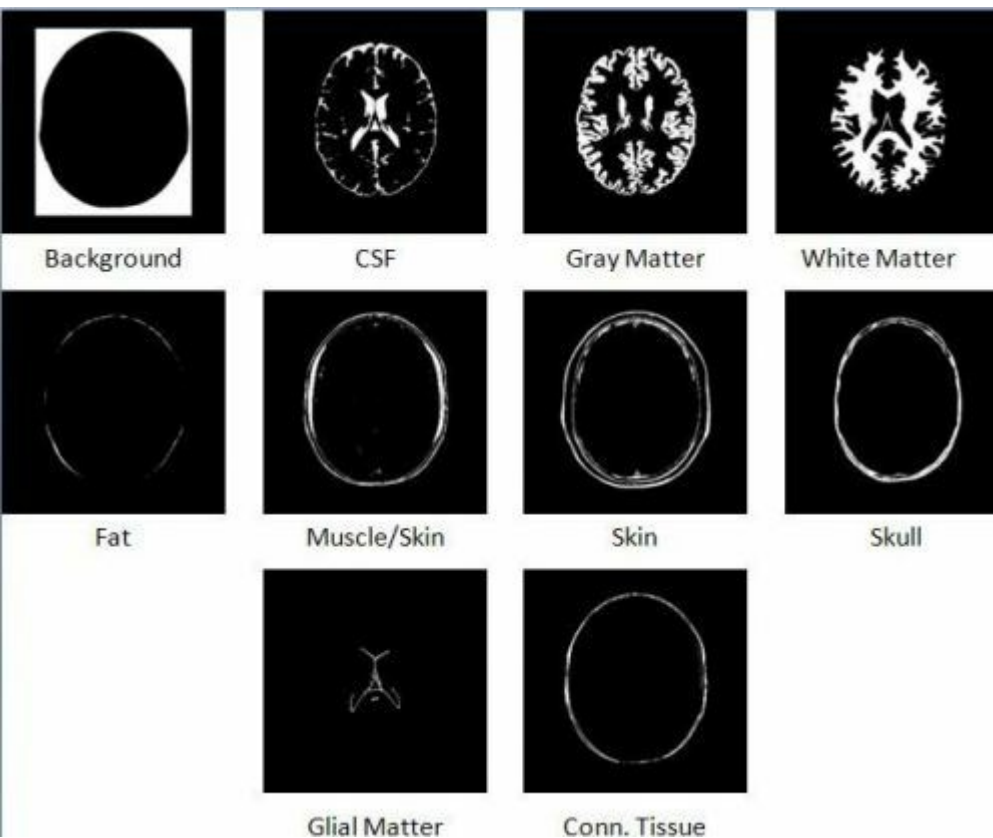
ICA Decomposition



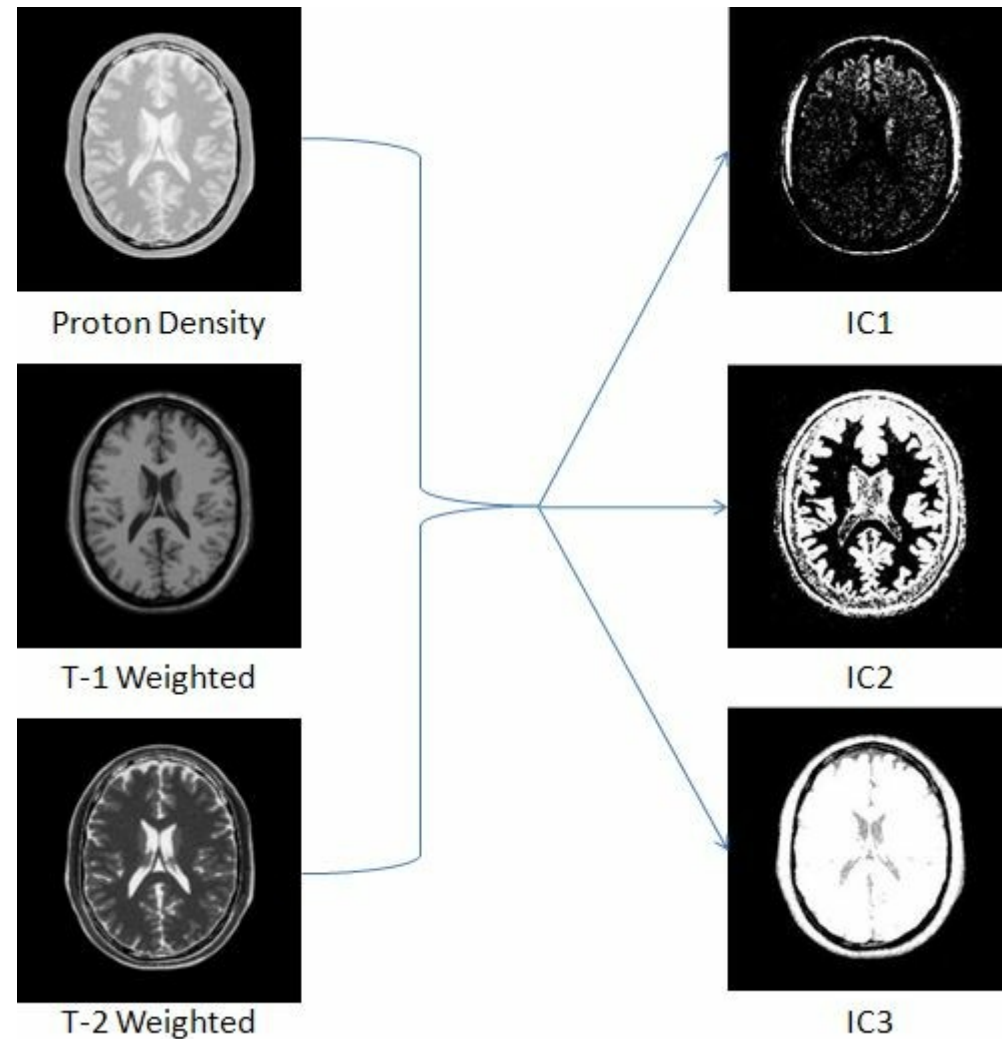
3 components from 21-dimensional decomposition using the "spatial-ICA" algorithm.

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

ICA and magnetic resonance



Sources of signals

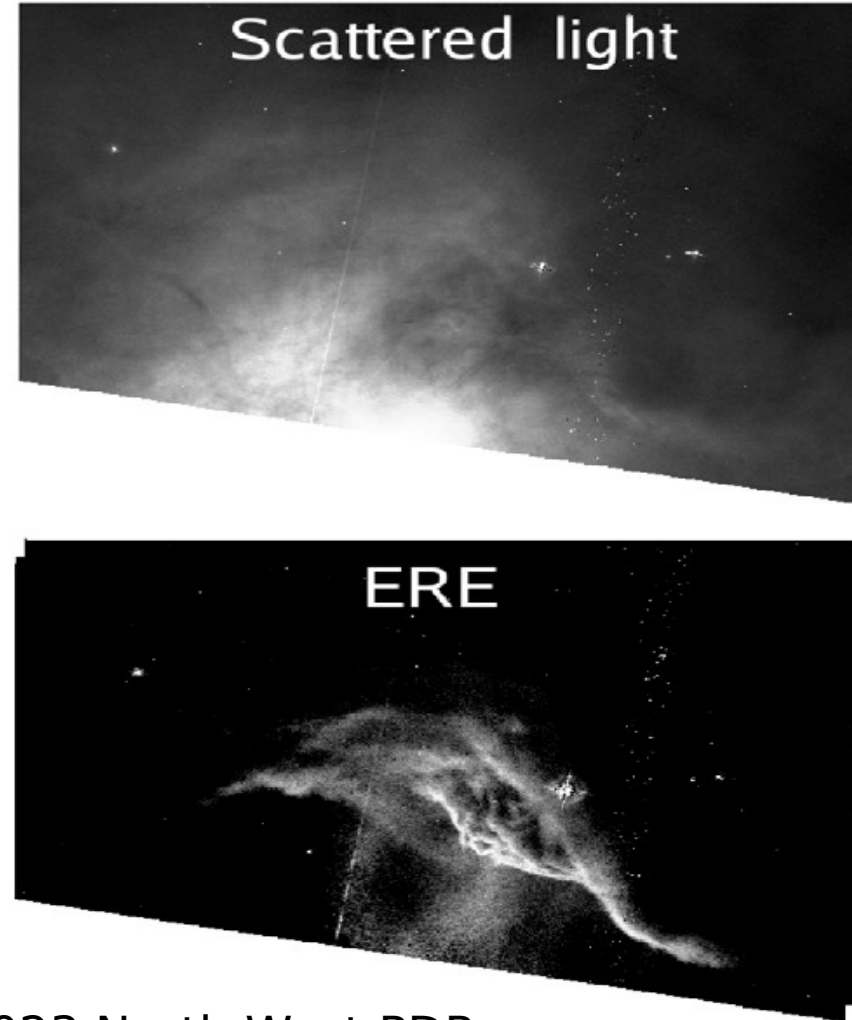
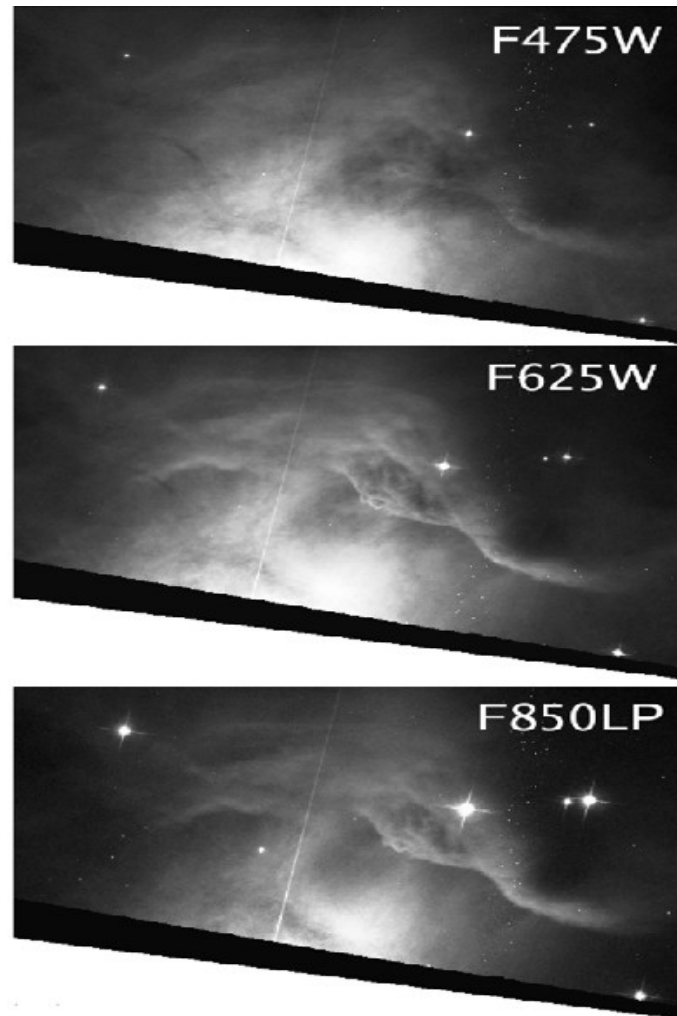


Measured

Separated components

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

ICA – astronomy



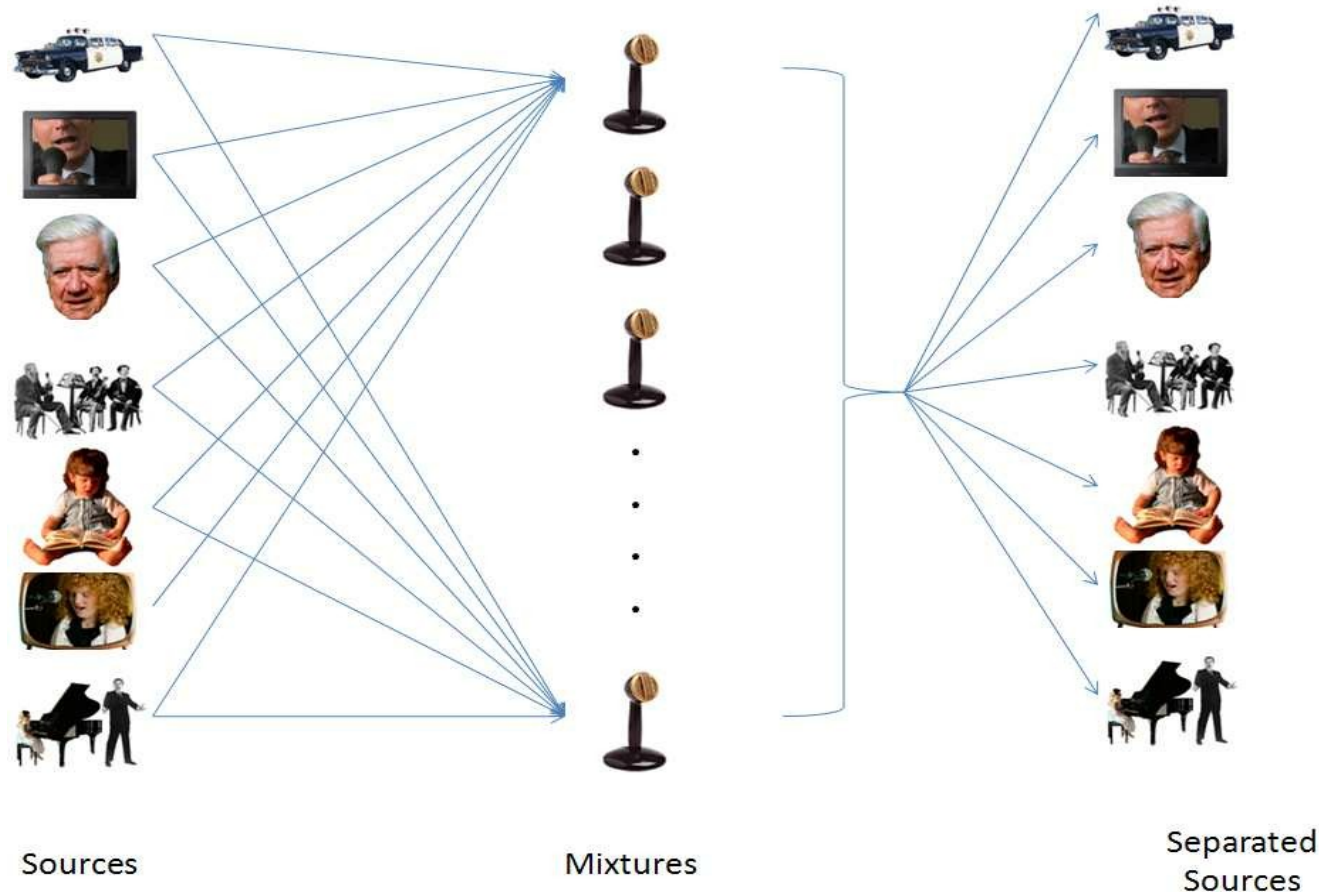
On the left: HST images of the NGC 7023 North-West PDR in three SDSS wide-band filters. On the right: scattered light and ERE (Extended Red Emission) images extracted with FastICA from the observations.

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

Desert

http://cni.salk.edu/~tewon/Blind/blind_audio.html

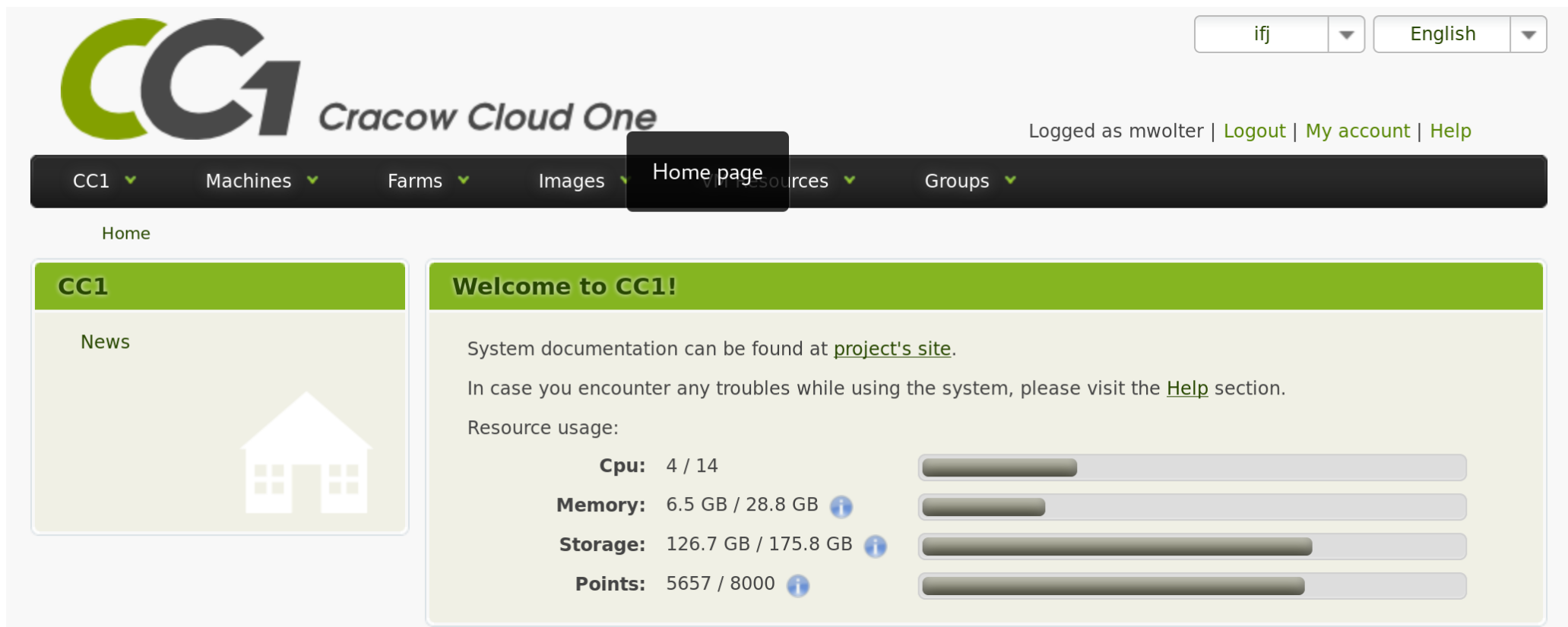
- Cocktail Party Demo - applet showing how the ICA algorithm works – blind separation of sound sources.



Practical exercises

How to get started with ROOT & TMVA?

- <https://www.cloud.ifj.edu.pl/>
- Register, you create your virtual UBUNTU linux box and play with it.
- Install root together with TMVA



CC1 Cracow Cloud One

Logged as mwolter | [Logout](#) | [My account](#) | [Help](#)

CC1 ▾ Machines ▾ Farms ▾ Images ▾ Home page ▾ Virtual Resources ▾ Groups ▾

Home

CC1

News

Welcome to CC1!

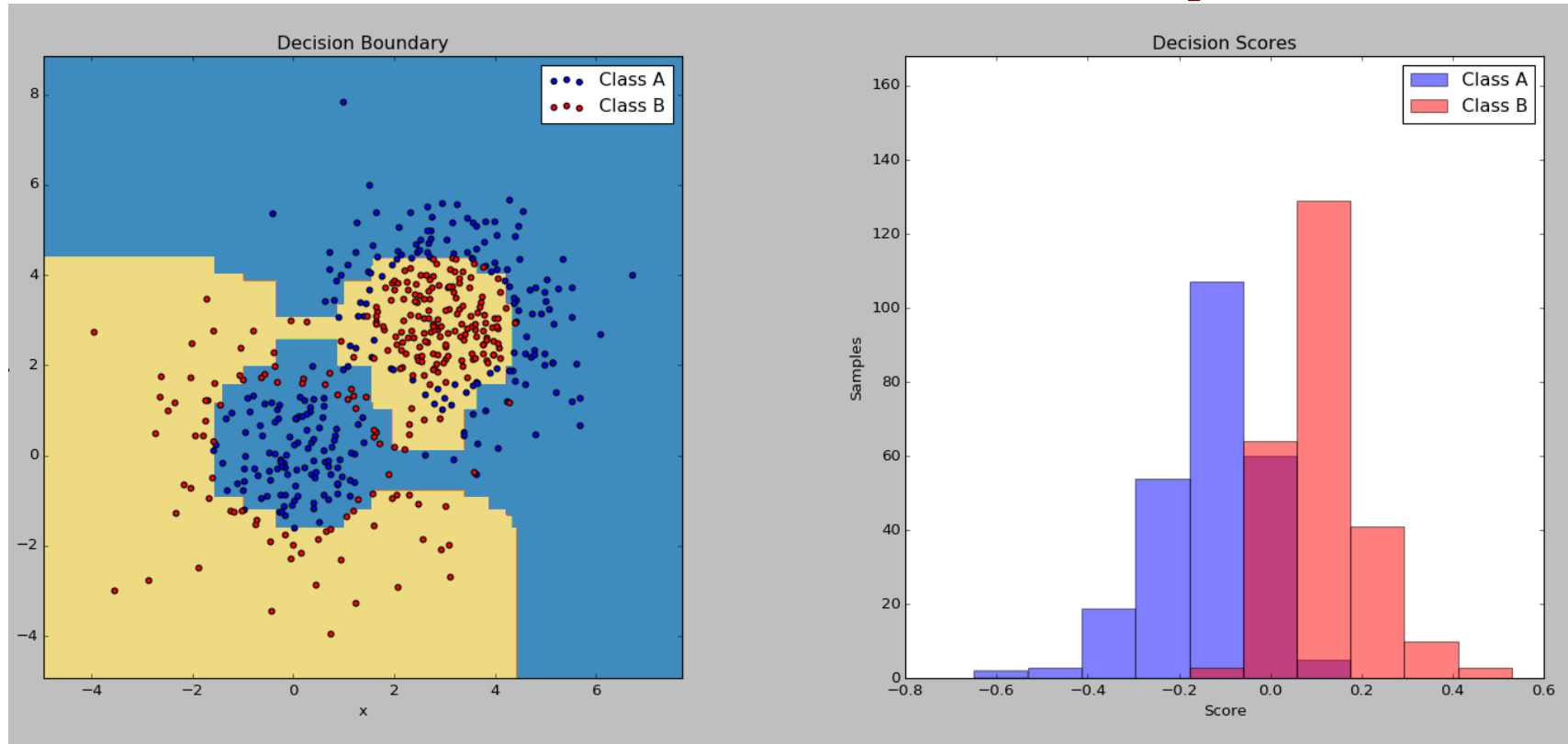
System documentation can be found at [project's site](#).

In case you encounter any troubles while using the system, please visit the [Help](#) section.

Resource usage:

Cpu:	4 / 14	<div><div></div></div>
Memory:	6.5 GB / 28.8 GB i	<div><div></div></div>
Storage:	126.7 GB / 175.8 GB i	<div><div></div></div>
Points:	5657 / 8000 i	<div><div></div></div>

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

ROOT exercises

- Scripts are attached to this talk.