Modern Methods in Data Analysis

Lecture 1: Intro

Boris Mirkin

Борис Григорьевич Миркин

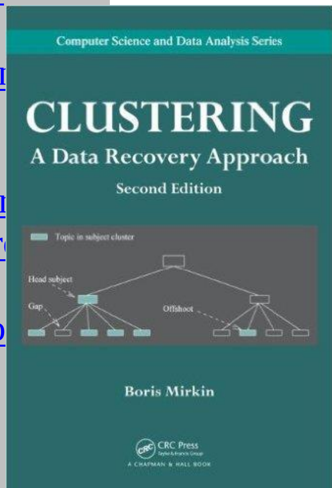
Professor, Data Analysis and AI, NRU HSE, Moscow, bmirkin@hse.ru, **8(963)-7234021**

Professor Emeritus, Computer Science, Birkbeck UL, London UK, mirkin@dcs.bbk.ac.uk

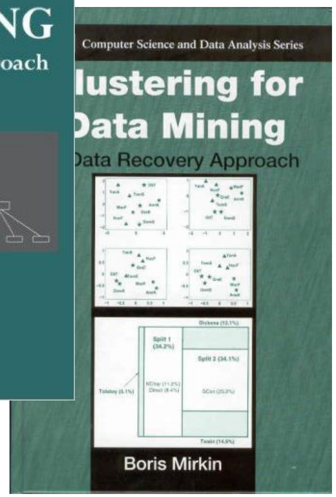
Boris Mirkin

- Home
- Biography
- Classes
- Publicatio
- Academic
- Interests
- Other Inte
- Current Pr
- Contact
- Informatio
- Resume

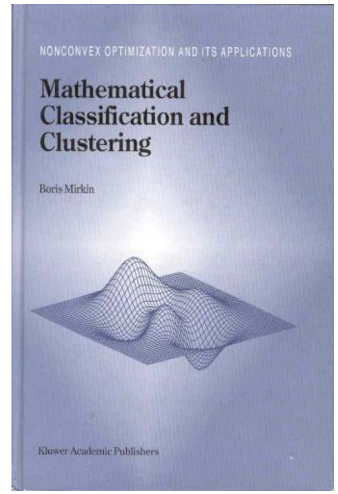
Welcome to my homepage. Here you will find information about me a
current activities.



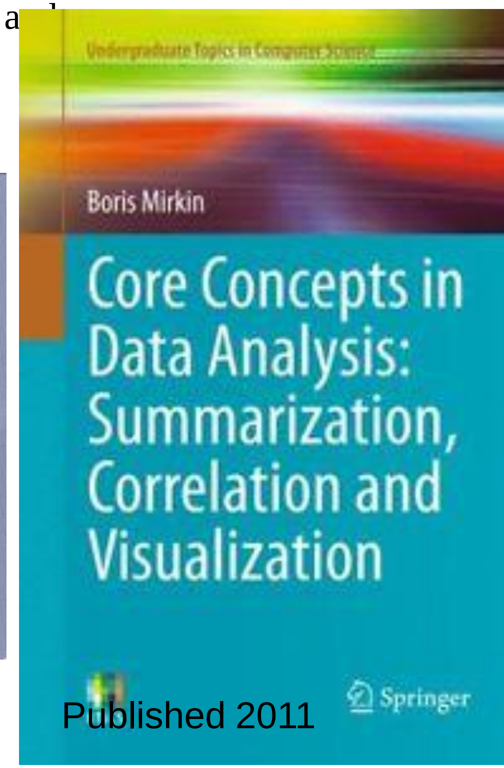
Monographs available in English



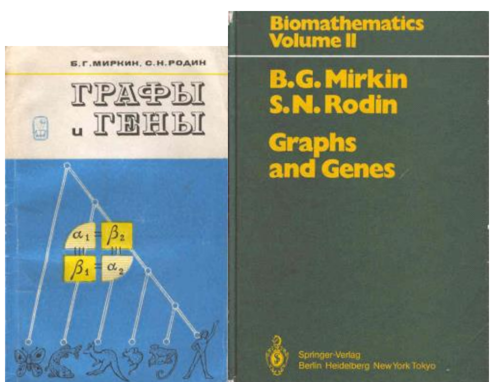
Published in 2005



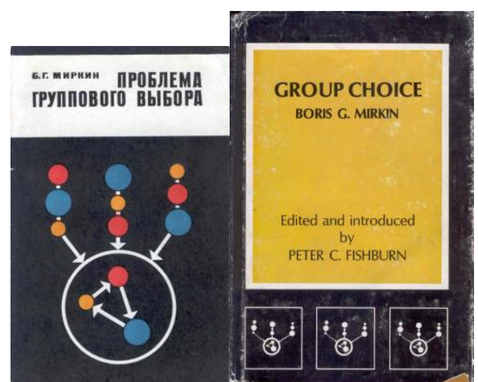
Published in 1996



Published 2011

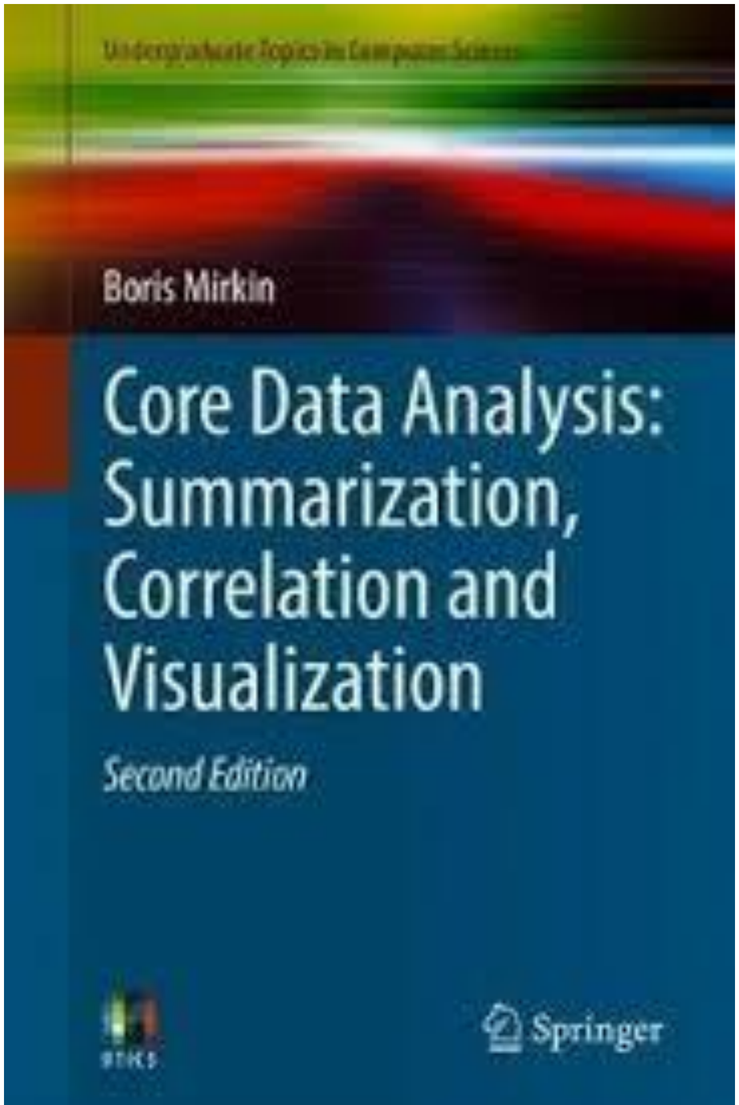


Translated in 1984



Translated in 1979

Main text for the class

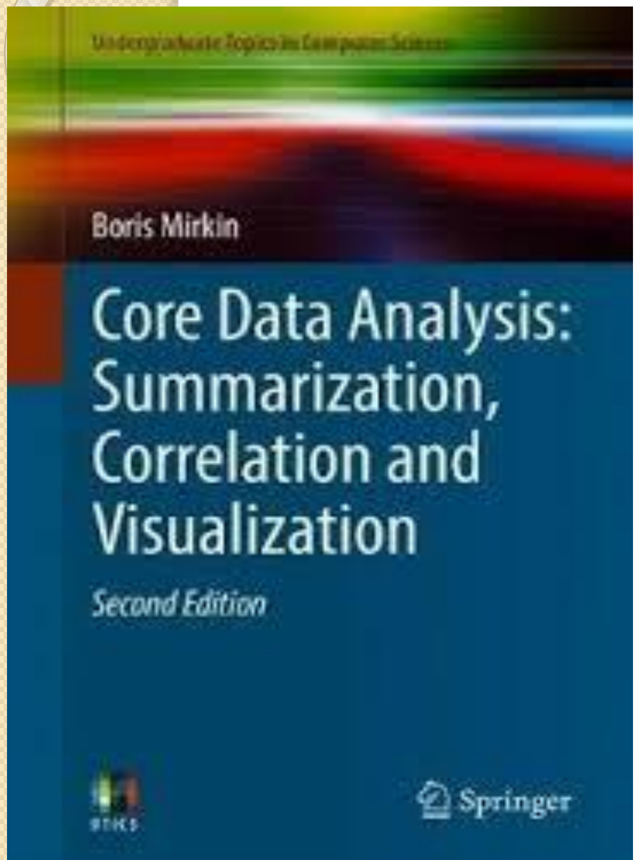


Boris Mirkin,

Core Data Analysis,

Springer, UTiCS Series, 2019, 527 p.

Book contents



1. **Topics in Substance of Data Analysis**
2. **Quantitative Summarization**

3 Learning Correlations

4 Core partitioning: K-Means and similarity clustering

5 Divisive and Separate Cluster Structures

Appendix. Basic Math and MatLab Code

Lecture 1 Contents

- Administration
- Brief history of Data Science
- Three examples of data analysis: two successful and one not
- Goal and contents of the class
- Data and metadata: Iris dataset and problems of its analysis
- Two formalizations of the concept of feature: vector and random variable

Lecture 1 Contents

- **Administration**
- Brief history of Data Science
- Three examples of data analysis: one successful and two not
- Goal and contents of the class
- Data and metadata: Iris dataset and problems of its analysis
- Two formalizations of the concept of feature: vector and random variable
- Feature scales: quantitative, ranking, nominal, and binary

Administration: Lectures and Labs

- Two modules (all of the Fall 2019)
- **In-class Exam Paper (EP)** in the end of December
- **Individual home-work (HW)**:
  - **Aa assignment in the end of each lecture**
  - **A report, in mid-December,** over
    - A dataset of at least 100 objects and 7 features taken from Internet or any other way (source must be indicated)- must be approved by me. Not necessarily one; may be a team of two **individuals**.
- **The final mark:**
  - **M=0.7EP+0.3HW**

Homework

- Individual home-work (HW):
  - A dataset of at least 100 objects and 7 features taken from Internet or any other way (source must be indicated)- must be approved by me. Not necessarily one; may be a team of two **individuals** .
  - About six Home assignments based on lectures including code (in **any language, including libraries**), computational application of a method and comments/interpretation of the result(s).

Generic Home-Work (in parentheses, share in mark %)

- **A1: Shaping of report including Data Description (10%)**
- **A2: K-means clustering (10%)**
- **A3: Cluster Interpretation (15%)**
- **A4: Bootstrap for comparing within-group averages (15%)**
- **A5: Contingency Table Analysis (20%)**
- **A6: PCA: Hidden Factor & Data Visualization (15%)**
- **A7: 2D Regression and correlation (15%)**

Finding a dataset of your liking, 1:

- Choose a subject of your liking, say «banking» или «global warming»
- Google that like "banking datasets", "global warming datasets"
- Take a look at the first one to five pages and click on a site of your liking; if this does not show anything interesting, repeat the attempt at a different site. Otherwise, go to the next item.
- If the data set is too large (say more than a thousand objects), select a smaller subset over a convenient feature.
- Select a few features (less than a dozen but more than four) and develop the corresponding data table
- Demonstrate that to the Instructor for approval.

Finding a dataset of your liking, 2: Example

Google Banking datasets

[banking data on data.world | 25 datasets available](https://data.world/datasets/banking)
https://data.world › datasets › banking
There are 25 **banking datasets** available on data.world.

Dataset Gallery: Banking & Finance | BigML.com
https://bigml.com › gallery › datasets › banking_...
BBVA Innova challenge Big Data

World Bank Data - Awesome Datasets - DataHub - Frictionless Data
https://datahub.io › collections › world-bank

Lecture 1 Contents

- Administration
- **Brief history of Data Science**
- Three examples of data analysis: two successful and one not
- Goal and contents of the class
- Data and metadata: Iris dataset and problems of its analysis
- Two formalizations of the concept of feature: vector and random variable

Brief history of Data Science 1

Period	Title	Contents
17-18 cent.	State statistics	Emerges in states of Northern Italy and Germany as «statista» from Italian stata=state
Begin-ning of 19 century	Methods for data averaging Social statistics	Astronomy: Least squares (K. Gauss, 1777-1855) and Least moduli (P-S. Laplace, 1749-1827) Analysis of mass phenomena (frequency = probability, A. Quetelet, 1796-1884)
End of 19 cen. –begin-ning of 20 cent	Multivariate statistics	Regression, correlation, variance, principal component, factor analyses in "hereditary genius" research (F. Galton, 1822-1911, K. Pearson, 1857-1936)

Brief history of Data Science, II

Period	Title	Contents
Beginning of 20 cen.	Classical mathematical statistics	Formulation of statistics within the probability theory as part of the theory of measurable sets and functions (A.N. Kolmogorov, 1903-1987, H. Kramer,1893-1984, R. Fisher,1890-1962)
Mid-20 century	Pattern recognition and Machine learning	Methods for developing classifiers (F. Rosenblatt, 1928-1971, E.M. Braverman, 1931-1976, V.N. Vapnik, 1936-)
End of 20 cen.	Data mining	Finding associations in big databases

Brief history of Data Science, III

Period	Title	Contents
Beginning of 21 century	Data analysis	Forming a system of methods related to data interpretation, structuration, summarization, correlation, and visualization.
Beginning of 21 century	Big data analysis	Realization of the fact that the current level of digitalization allows to move from the analysis of individual data tables to combined data and text analyses in real time, leading to a new quality – Artificial Intelligence

Lecture 1 Contents

- Administration
- Brief history of Data Science
- **Three examples of data analysis: two successful and one not**
- Goal and contents of the class
- Data and metadata: Iris dataset and problems of its analysis
- Two formalizations of the concept of feature: vector and random variable

Two examples of successful data analysis

- Pluto: a Planet?
- Planetary motion: Johann Kepler's 3d law

Planets: Is any of them a planet indeed?

Example of a good cluster structure: W. Jevons (1835-1882), updated in Mirkin 1996

	Planet	Distance kilomile	Diameter mile	Period year	Day	Moons amount	Matter	EBalance
	Mercury	36	3000	0.24	59	0	Solid	Negative
Cl. 1	Venus	67	7500	0.62	243	0	Solid	Negative
	Earth	93	7900	1	1	1	Solid	Negative
	Mars	142	4200	1.88	1	2	Solid	Negative
	Jupiter	483	89000	12	0.42	17	Liquid	Positive
Cl. 2	Saturn	885	74600	30	0.42	22	Liquid	Positive
	Uranus	1800	32200	84	0.67	15	Mixed	Positive
	Neptune	2800	30800	165	0.75	8	Liquid	Positive
???	Pluto	3660	1620	248	6.40	1	Solid	Negative

Table 1: **Planets:** Planets in the Solar system; EBalance is the difference between the received and emitted energies.

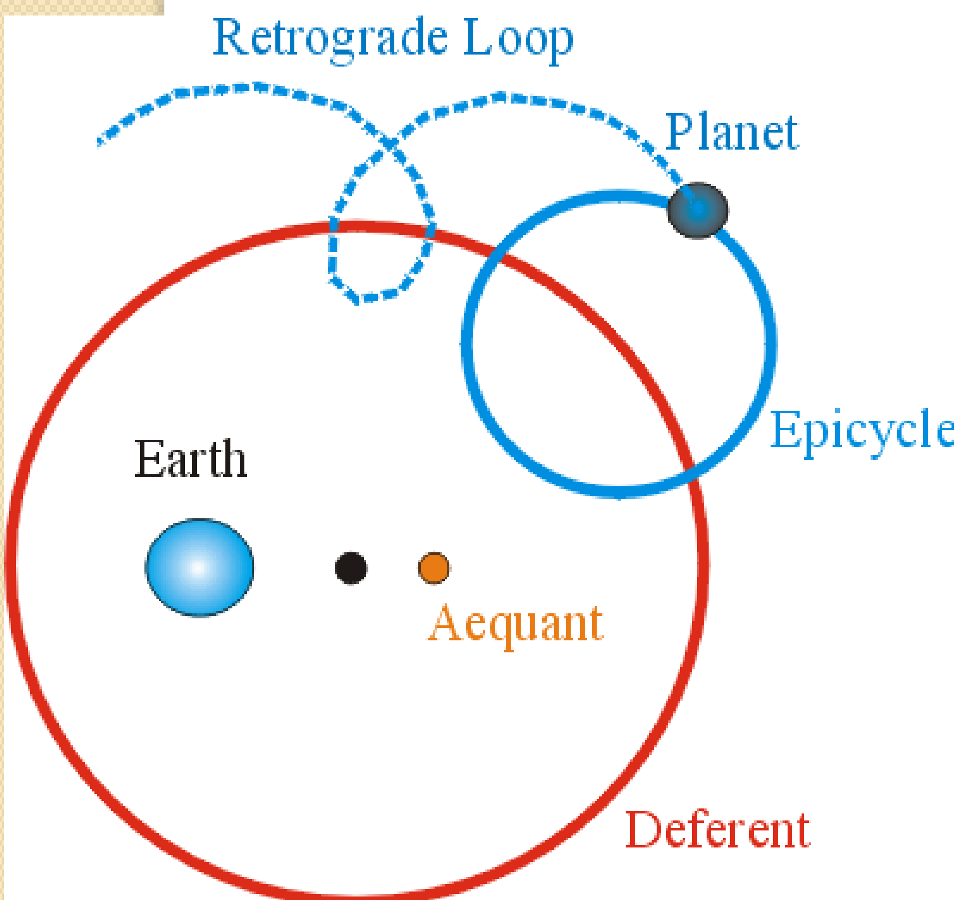
Pluto doesn't fit in the two clusters of planets: started a new cluster in 2006.

Planetary motion:
A much successful example of small data analysis

Double success 1

-

A History of Laws for planetary motion



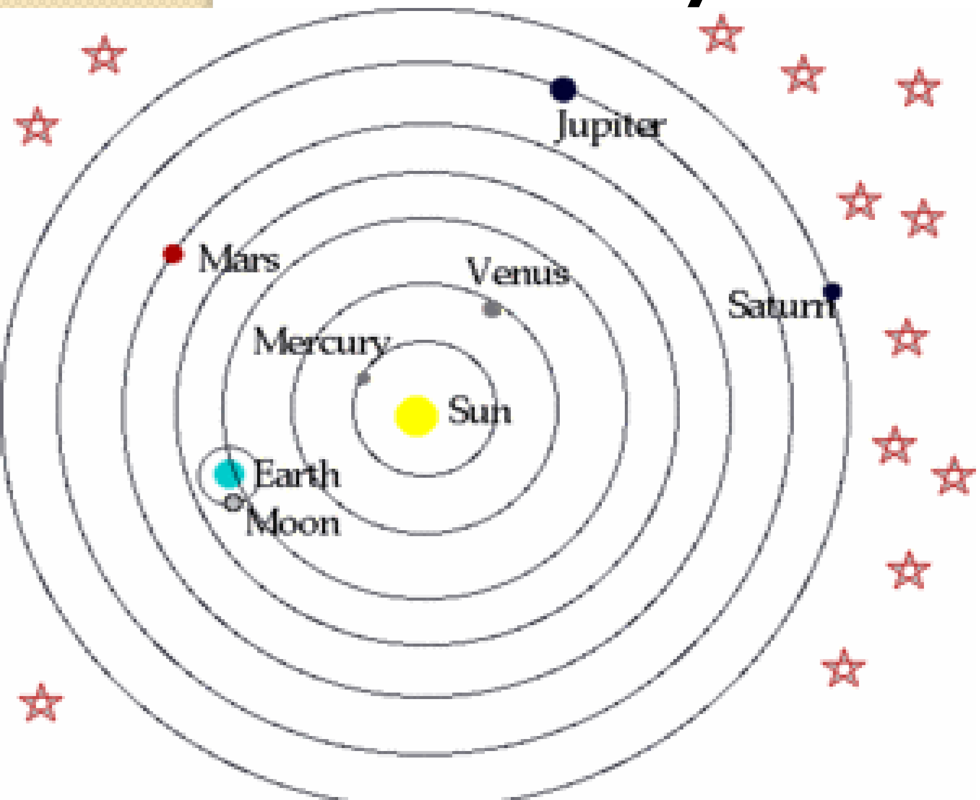
Ptolemy (c. 150 a. d.):

Sun and planets circle Earth

Does not match data well

Double success 2

- The History of Laws for planetary motion



Copernicus

(c. 1540):

Planets circle Sun

Does not match data well either

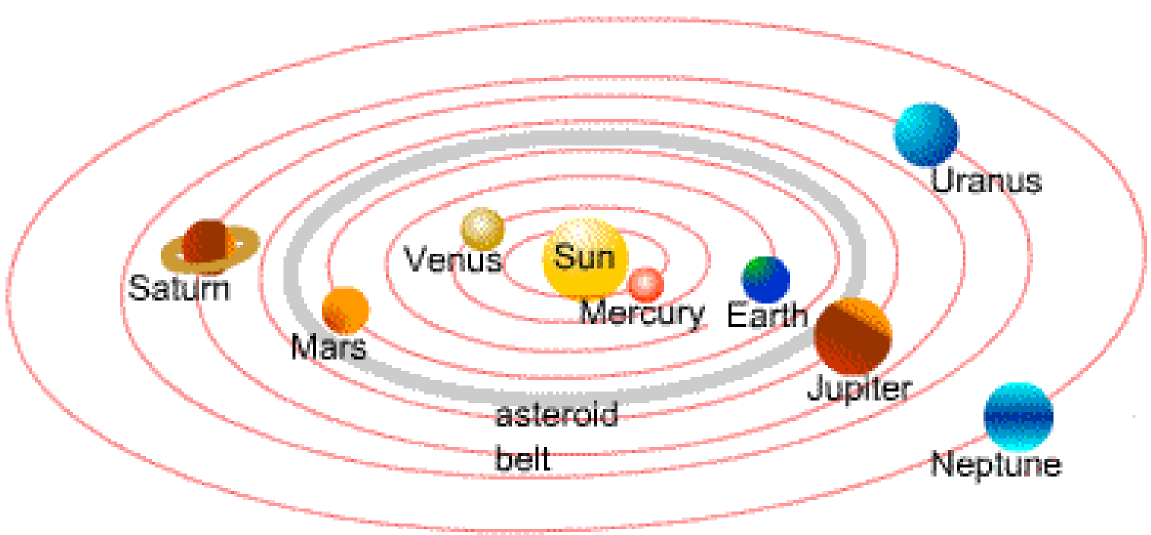
-

Double success 3

- 0^{th} Law: All planets move in the same plane

Laws for planetary motion: J. Kepler (c. 1605):

- 1^{st} Law: Planets revolve Sun in ellipses (ovals)
- 2^{d} Law: Speed changes – the further away from Sun, the slower (equal sectors in time unit)



Double success 4: 3d Kepler's Law:

Kepler's thinking after 1605:

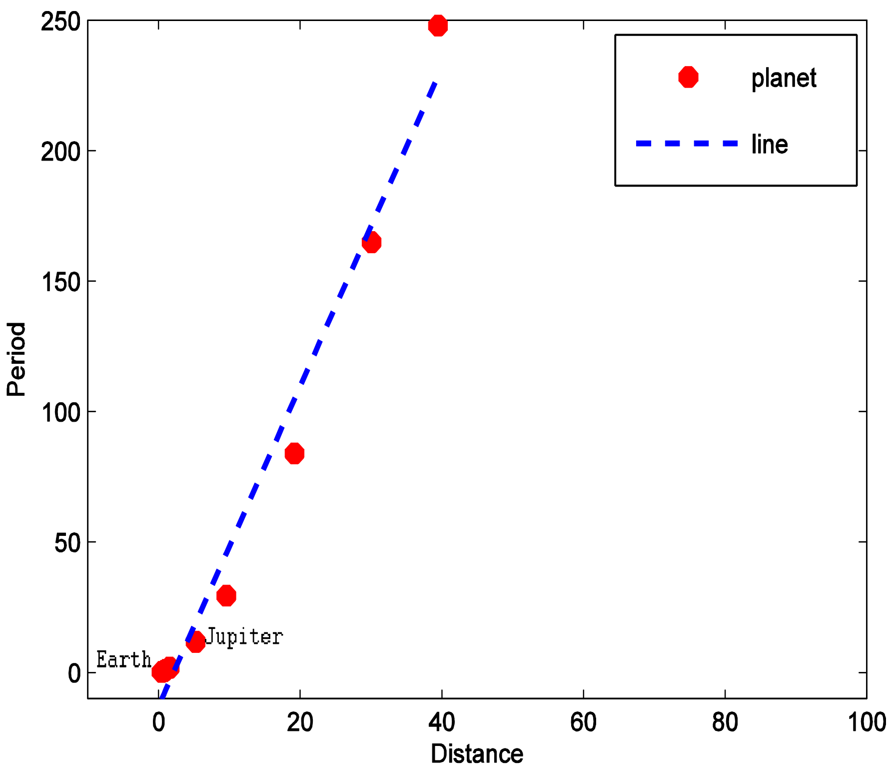
Planet	Period (year)	Distance (average, relative to that of Earth)
Mercury	0.241	0.39
Venus	0.615	0.72
Earth	1.00	1.00
Mars	1.88	1.52
Jupiter	11.8	5.20
Saturn	29.5	9.54
Uranus	84.0	19.18
Neptune	165	30.06
Pluto	248	39.44

It should be a relation between speed/period and distance;

which one?

Double success 5

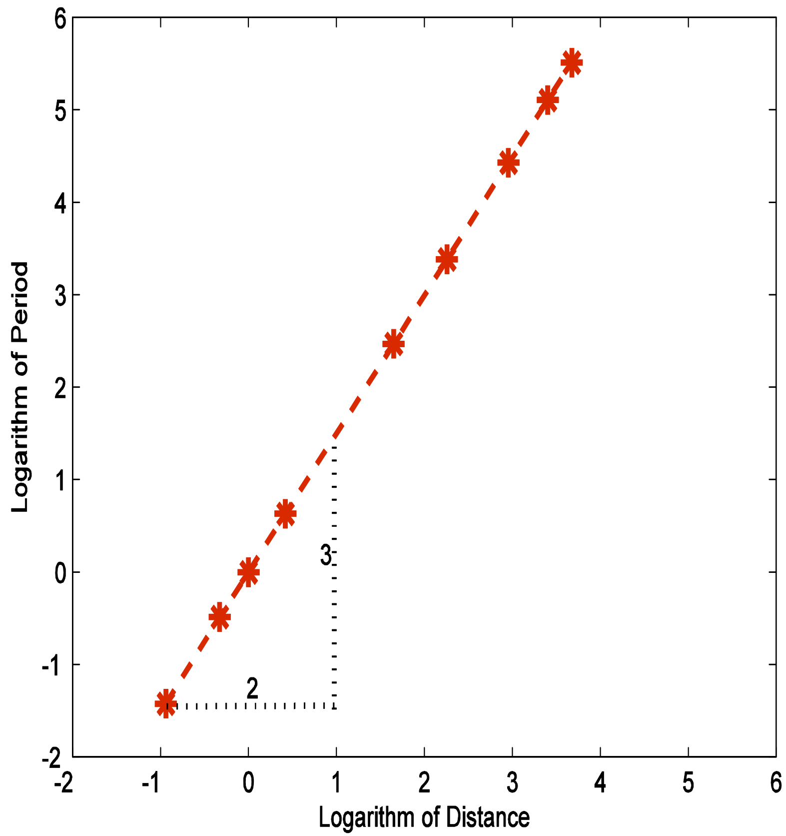
3^d Kepler's Law:



Is there any relation between speed/period and distance?

Points on the plane "Distance-Period" fit no line…

Example of Small Data Analysis **Double success 6**



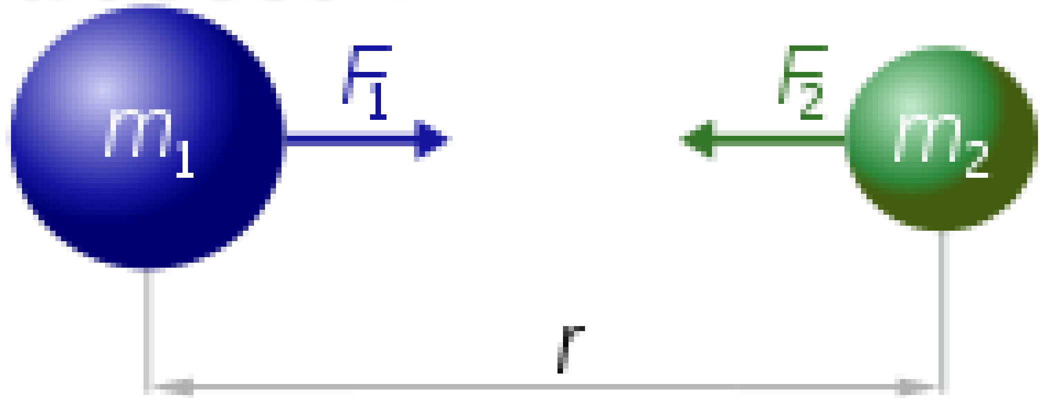
3d Kepler's Law (1619):

[J. Napier invented logarithm (1614)]

$$\mathrm{Log}(P) = \frac{3}{2}\,\mathrm{Log}(D)$$

$$P^2 = D^3$$

Double success 7



$$F_1 = F_2 = G\,\frac{m_1 \times m_2}{r^2}$$

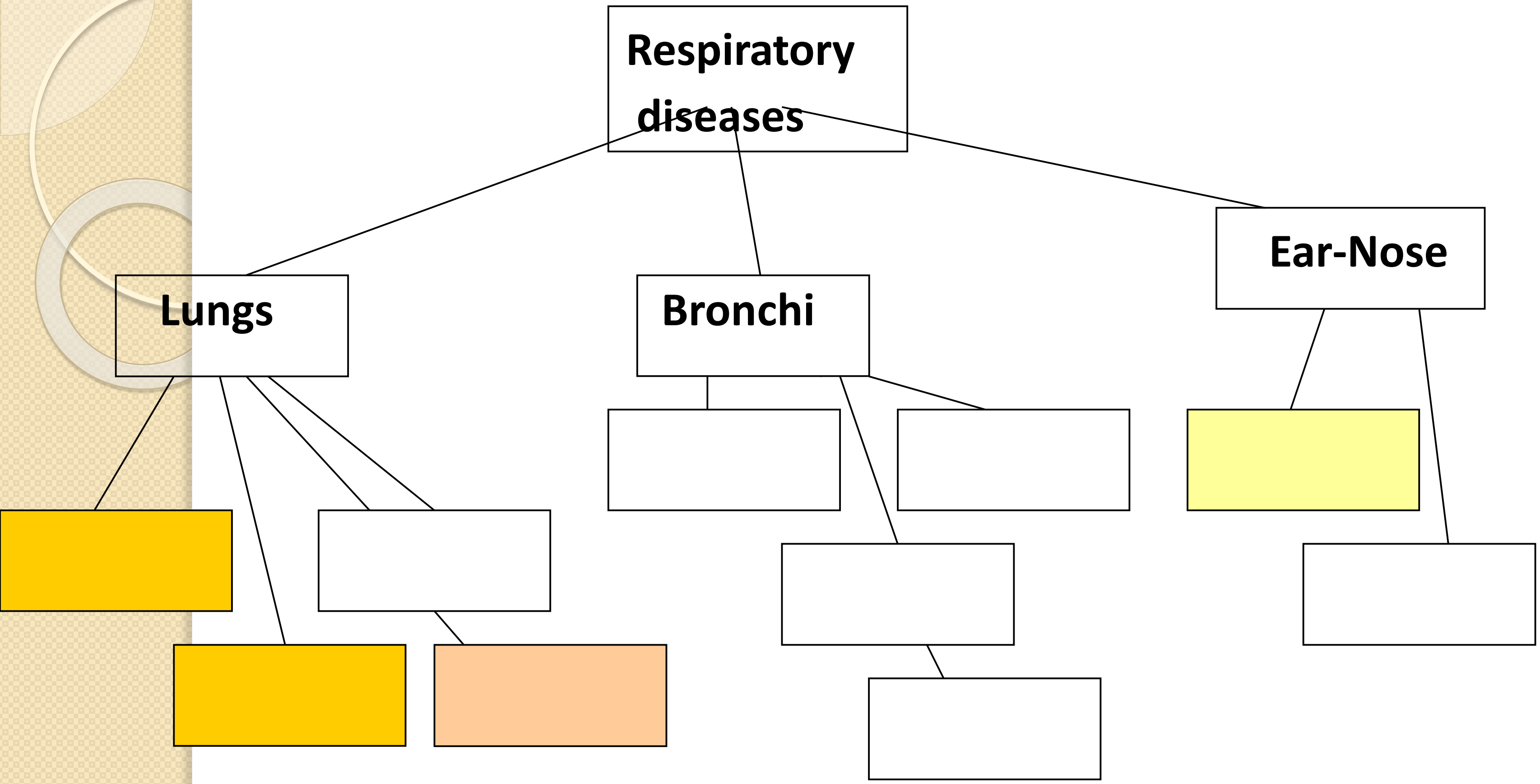
Three Kepler's Laws: What is so grand?

Substantiated theoretically by
R. Hooke (1635-1703) and I. Newton (1642-1727)
UNIVERSAL GRAVITATION LAW

Equation above, cornerstone of the science

An example of unsuccessful data analysis

- From my own data analysis experiences
- Risk factors of respiratory diseases in Akademgorodok, Novosibirsk, Russia (1981)



Rostovtsev, Mirkin, Shanin (1981 unpublished):

Investigation in the local respiratory diseases and risk factors for them

~50 000 respondents: 14 hierarchical clusters

Rostovtsev, Mirkin, Shanin (1981 unpublished), I: Respiratory diseases survey

Smoking

Drinking

Risk factors suggested according to the views of that time

Rostovtsev, Mirkin, Shanin (1981 unpublished), 2: Respiratory diseases survey

Risk factors according to the data:

The disease in family

Poor housing

Rostovtsev, Mirkin, Shanin (1981 unpublished), 3: Respiratory diseases

Risk factors according to data :

- The disease in family
- Poor housing

Smoking/Drinking:

Statistically independent, **not risk factors**

These conclusions, now a common place:

Rejected as contradicting to firmly established principles (1981) (like those by J. Snow 1854)

Lecture 1 Contents

- Administration
- Brief history of Data Science
- Three examples of data analysis: two successful and one not
- **Goal and contents of the class**
- Data and metadata: Iris dataset and problems of its analysis
- Two formalizations of the concept of feature: vector and random variable

Goal of the class

- **Mastering core concepts and approaches**
  - **To see through the structure of main methods and**
  - **To be able to extend them at**
    - **new data types or**
    - **new types of problems**

Core Data Analysis: 2 tasks x 3 forms

-
- Quantitative
  Principal component analysis
- **Summarization** Categorical
  Cluster analysis
  Ranking
- Google ranking/Consensus
-
- Quantitative
  Regression analysis
- **Correlation** Categorical
- Decision tree

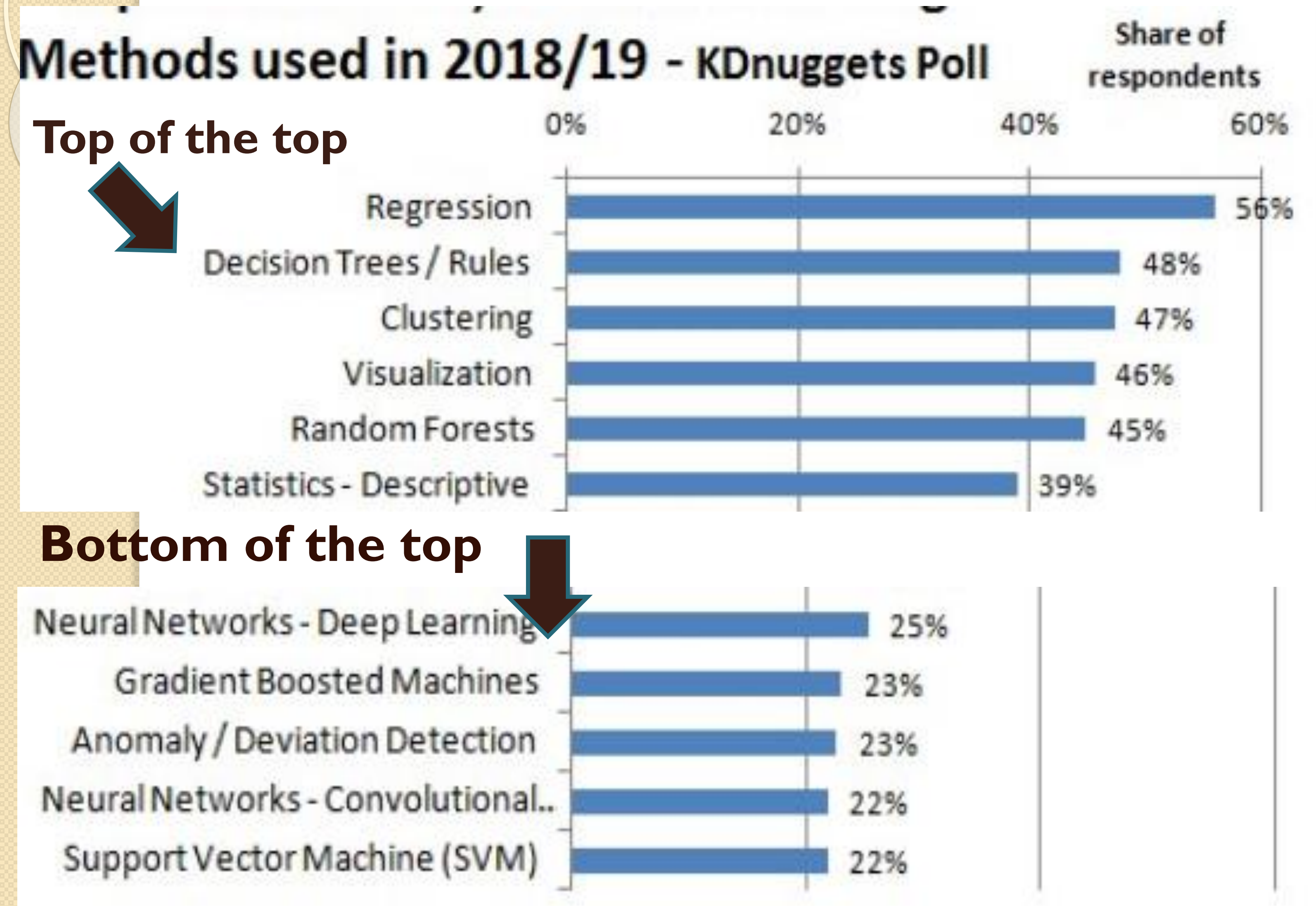
Plan of the class (bird's eye view)

- Clustering
  - Partition
  - Hierarchies (if time permits)
  - Methods for interpretation and data analytics
  - Comparing averages with bootstrap
  - Similarity data clustering
  - Consensus clustering
- SVD and Principal Component Analysis
  - Hidden factor
  - Data visualization
  - Ranking
- Regression, Decision Tree (if time permits), Naive Bayes

Top Data Science

Top of the top

Bottom of the top



Data Analysis versus Machine Learning:

- DA: Using data for enhancing knowledge of the domain
- ML: to equip computers with methods and rules to see the target by using data
- **HUGE OVERLAP**
- Example of difference: **Neural-Net∈ML–DA**
  - **Good** for robot to prevent an explosion
  - **Bad** for lawyer to build their case

Difficulty of this class, 1:

- Subject is yet in the **making**
- Spoken **English**
- Full of **mathematics** and **computation,** but differs from either by focusing on **data**
- Requires from the student not just thinking, but **decision making,** first of all

Difficulty of this class, II:

- **A method for computational tasks works always, but a data analysis method may bring forth an inconvenient solution (a failure):**
  - **If the solution does not help in making new conclusions of the object, it is inconvenient.**
  - **If the solution does not fit into existing knowledge, it is inconvenient.**
- **In such a case, the user has to revise their approach.**

Lecture 1 Contents

- Administration
- Brief history of Data Science
- Three examples of data analysis: two successful and one not
- Goal and contents of the class
- **Data and metadata: Iris dataset and problems of its analysis**
- Two formalizations of the concept of feature: vector and random variable

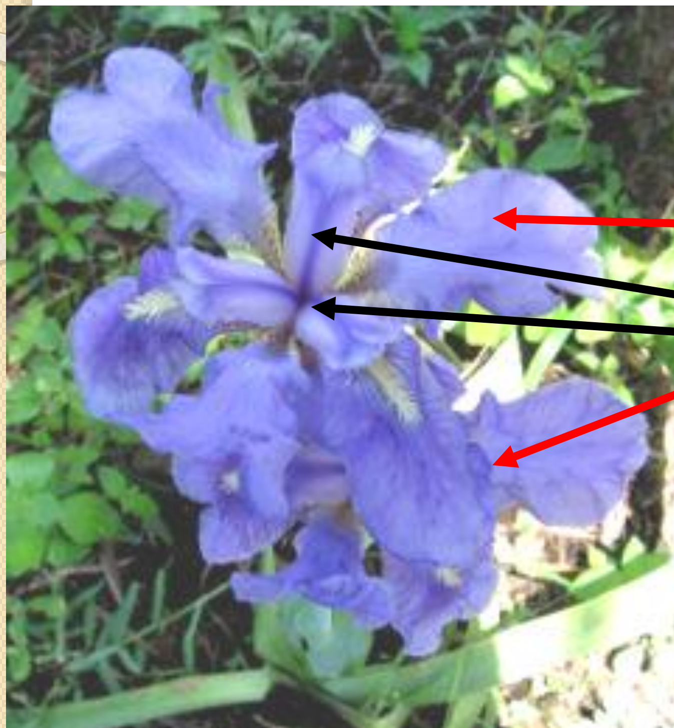
What is data: homogeneous information of a set of objects
Metadata: what is left outside of the data values

- Table
- Signal
- Text
- Sequence
- Map
- Image
- Video
- …….

- This class concentrates on **data tables** as
- generic, simplest, and best explored object

A typical dataset: Anderson–Fisher Iris

Iris flower



Sepal / **Чашелистик**

Petal / **Лепесток**

150×4 data of three taxa:

	Taxon
1-50	***Iris setosa* (diploid)**
51-100	***Iris versicolor* (tetraploid)**
101-150	***Iris virginica* (hexaploid)**

Features

W1	**Sepal length**
W2	**Sepal width**
W3	**Petal length**
W4	**Petal width**

} Metadata

Taxa

Three Iris taxa:

Setosa Virginica Versicolor

Data case : Iris, most popular dataset

#	Iris setosa	Iris versicolor	Iris virginica
	w1 w2 w3 w4	w1 w2 w3 w4	w1 w2 w3 w4
1	5.1 3.5 1.4 0.3	6.4 3.2 4.5 1.5	6.3 3.3 6.0 2.5
2	4.4 3.2 1.3 0.2	5.5 2.4 3.8 1.1	6.7 3.3 5.7 2.1
3	4.4 3.0 1.3 0.2	5.7 2.9 4.2 1.3	7.2 3.6 6.1 2.5
4	5.0 3.5 1.6 0.6	5.7 3.0 4.2 1.2	7.7 3.8 6.7 2.2
5	5.1 3.8 1.6 0.2	5.6 2.9 3.6 1.3	7.2 3.0 5.8 1.6
6	4.9 3.1 1.5 0.2	7.0 3.2 4.7 1.4	7.4 2.8 6.1 1.9
7	5.0 3.2 1.2 0.2	6.8 2.8 4.8 1.4	7.6 3.0 6.6 2.1
8	4.6 3.2 1.4 0.2	6.1 2.8 4.7 1.2	7.7 2.8 6.7 2.0
9	5.0 3.3 1.4 0.2	4.9 2.4 3.3 1.0	6.2 3.4 5.4 2.3
…			
50	5.1 3.5 1.4 0.2	6.0 2.2 4.0 1.0	6.5 3.2 5.1 2.0

What type of analysis to do?

Some problems for Iris data analysis (I):

Iris 2

#	I Iris setosa	II Iris versicolor	III Iris virginica
	w1 w2 w3 w4	w1 w2 w3 w4	w1 w2 w3 w4
1	5.1 3.5 1.4 0.3	6.4 3.2 4.5 1.5	6.3 3.3 6.0 2.5
2	4.4 3.2 1.3 0.2	5.5 2.4 3.8 1.1	6.7 3.3 5.7 2.1
3	4.4 3.0 1.3 0.2	5.7 2.9 4.2 1.3	7.2 3.6 6.1 2.5
4	5.0 3.5 1.6 0.6	5.7 3.0 4.2 1.2	7.7 3.8 6.7 2.2
5	5.1 3.8 1.6 0.2	5.6 2.9 3.6 1.3	7.2 3.0 5.8 1.6
6	4.9 3.1 1.5 0.2	7.0 3.2 4.7 1.4	7.4 2.8 6.1 1.9
7	5.0 3.2 1.2 0.2	6.8 2.8 4.8 1.4	7.6 3.0 6.6 2.1
8	4.6 3.2 1.4 0.2	6.1 2.8 4.7 1.2	7.7 2.8 6.7 2.0
9	5.0 3.3 1.4 0.2	4.9 2.4 3.3 1.0	6.2 3.4 5.4 2.3
50	5.1 3.5 1.4 0.2	6.0 2.2 4.0 1.0	6.5 3.2 5.1 2.0

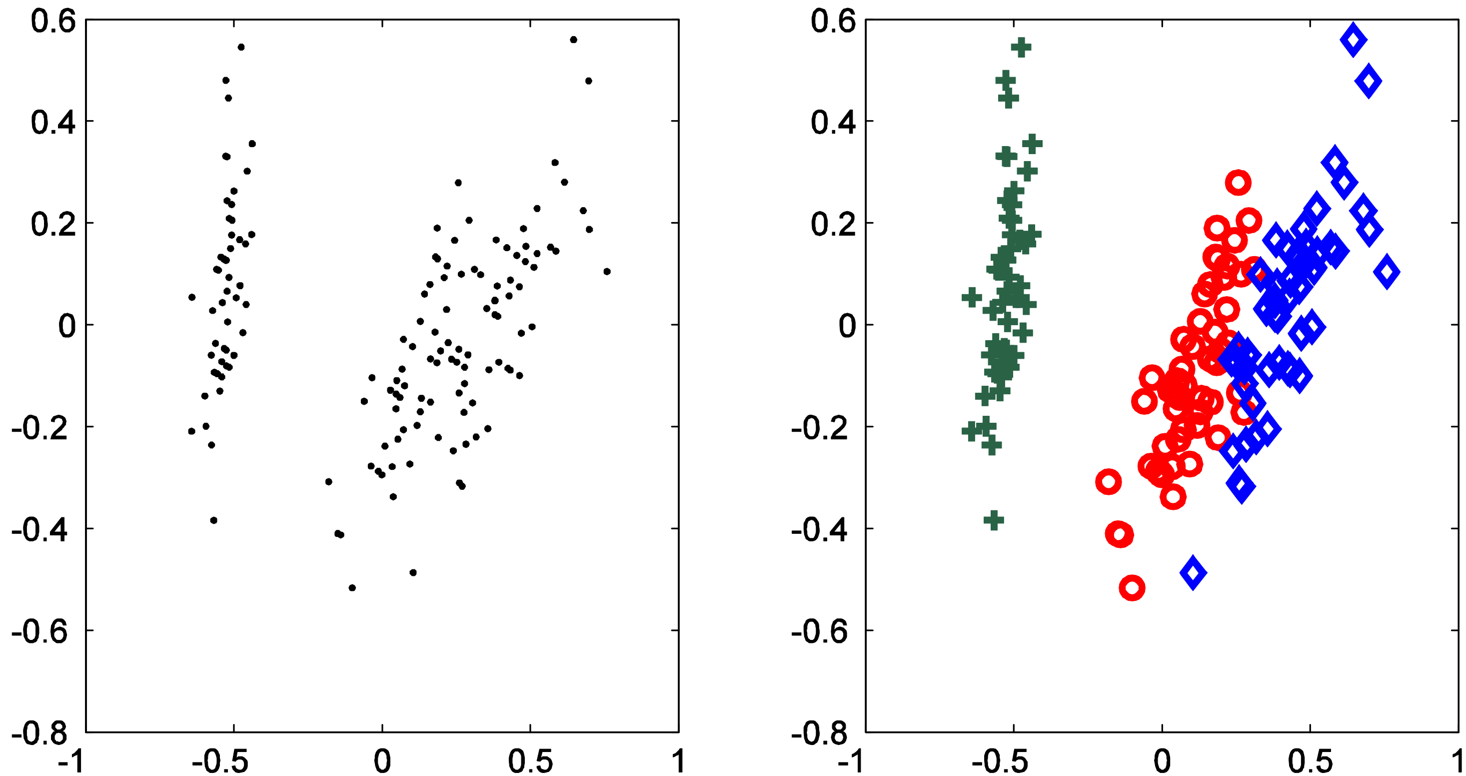
- **Visualise** data: map **similar** specimens at points **near** each other; **dissimilar** specimens, at **far away** points
- Build a **predictor of sepal sizes** from the petal sizes (say, to lessen the burden of measurement)

Iris dataset structure: 2D visualized with MATLAB

Left: >>subplot(1,2,1); plot(z1, z2, ‘k.’);

Right: >>subplot(1,2,2);

>>plot(z1(1:50),z2(1:50),'g+',z1(51:100),z2(51:100),'ro',z1(101:150),z2(101:150),'bd');



Some problems for Iris data analysis (II):

Iris 2

#	I Iris setosa	II Iris versicolor	III Iris virginica
	w1 w2 w3 w4	w1 w2 w3 w4	w1 w2 w3 w4
1	5.1 3.5 1.4 0.3	6.4 3.2 4.5 1.5	6.3 3.3 6.0 2.5
2	4.4 3.2 1.3 0.2	5.5 2.4 3.8 1.1	6.7 3.3 5.7 2.1
3	4.4 3.0 1.3 0.2	5.7 2.9 4.2 1.3	7.2 3.6 6.1 2.5
4	5.0 3.5 1.6 0.6	5.7 3.0 4.2 1.2	7.7 3.8 6.7 2.2
5	5.1 3.8 1.6 0.2	5.6 2.9 3.6 1.3	7.2 3.0 5.8 1.6
6	4.9 3.1 1.5 0.2	7.0 3.2 4.7 1.4	7.4 2.8 6.1 1.9
7	5.0 3.2 1.2 0.2	6.8 2.8 4.8 1.4	7.6 3.0 6.6 2.1
8	4.6 3.2 1.4 0.2	6.1 2.8 4.7 1.2	7.7 2.8 6.7 2.0
9	5.0 3.3 1.4 0.2	4.9 2.4 3.3 1.0	6.2 3.4 5.4 2.3
50	5.1 3.5 1.4 0.2	6.0 2.2 4.0 1.0	6.5 3.2 5.1 2.0

- Build a **predictor of taxa** (classifier) based on the petal/sepal sizes
- Check how much features W1—W4 are relevant (for example, by making **clusters** and comparing them to the taxa)

Lecture 1 Contents

- Administration
- Brief history of Data Science
- Three examples of data analysis: one successful and one not
- Goal and contents of the class
- Data and metadata: Iris dataset and problems of its analysis
- **Two formalizations of the concept of feature: vector and random variable**

Iris, most popular dataset, features w1, w2, w3, w4

#	Iris setosa	Iris versicolor	Iris virginica
	w1 w2 w3 w4	w1 w2 w3 w4	w1 w2 w3 w4
1	5.1 3.5 1.4 0.3	6.4 3.2 4.5 1.5	6.3 3.3 6.0 2.5
2	4.4 3.2 1.3 0.2	5.5 2.4 3.8 1.1	6.7 3.3 5.7 2.1
3	4.4 3.0 1.3 0.2	5.7 2.9 4.2 1.3	7.2 3.6 6.1 2.5
4	5.0 3.5 1.6 0.6	5.7 3.0 4.2 1.2	7.7 3.8 6.7 2.2
5	5.1 3.8 1.6 0.2	5.6 2.9 3.6 1.3	7.2 3.0 5.8 1.6
6	4.9 3.1 1.5 0.2	7.0 3.2 4.7 1.4	7.4 2.8 6.1 1.9
7	5.0 3.2 1.2 0.2	6.8 2.8 4.8 1.4	7.6 3.0 6.6 2.1
8	4.6 3.2 1.4 0.2	6.1 2.8 4.7 1.2	7.7 2.8 6.7 2.0
9	5.0 3.3 1.4 0.2	4.9 2.4 3.3 1.0	6.2 3.4 5.4 2.3
…			
50	5.1 3.5 1.4 0.2	6.0 2.2 4.0 1.0	6.5 3.2 5.1 2.0

What is feature w1? According to data analysis view, just the column w1's contents!

What is feature w1? According to data analysis view, just the column w1's contents:

- Index 1 through 9

5.1 4.4 4.4 5.0 5.1 4.9 5.0 4.6 5.0

. . . .

- Index 142 through 150

6.7 6.3 6.5 6.5 7.3 6.7 5.6 6.4 6.5

What is this as a mathematical object?

What is the column w1's contents as a mathematical object?

:

- Index 1 through 9

5.1 4.4 4.4 5.0 5.1 4.9 5.0 4.6 5.0

. . . .

- Index 142 through 150

6.7 6.3 6.5 6.5 7.3 6.7 5.6 6.4 6.5

Two different views co-exist (like the photon, unit of light, in quantum physics: both a particle and a wave)

Two different views at the w1 feature as a mathematical object::

- Index 1 through 9

5.1 4.4 4.4 5.0 5.1 4.9 5.0 4.6 5.0

. . . .

- Index 142 through 150

6.7 6.3 6.5 6.5 7.3 6.7 5.6 6.4 6.5

A) **Vector of 150x1 dimension**

B) **150-strong sample from a random variable**

A) Feature as vector, 1:

:

- Index 1 through 9

5.1 4.4 4.4 5.0 5.1 4.9 5.0 4.6 5.0

. . . .

- Index 142 through 150

6.7 6.3 6.5 6.5 7.3 6.7 5.6 6.4 6.5

Math: Given a set I of object indices or names, feature is a mapping $f: I \rightarrow R$ where R is the set of all reals, that is, $f=(f_i)$, $i \in I$, an $|I|$-dimensional vector

A) Feature as vector, 2:

- Index 1 through 9

5.1 4.4 4.4 5.0 5.1 4.9 5.0 4.6 5.0

. . . .

- Index 142 through 150

6.7 6.3 6.5 6.5 7.3 6.7 5.6 6.4 6.5

Pro: **a) Intuitive;**

b) Objects are explicit (rows)

c) Linear algebra applies

Con: **d) Empirical (depends on I, cannot be extended to the universe)**

B) Feature as random variable, 1:

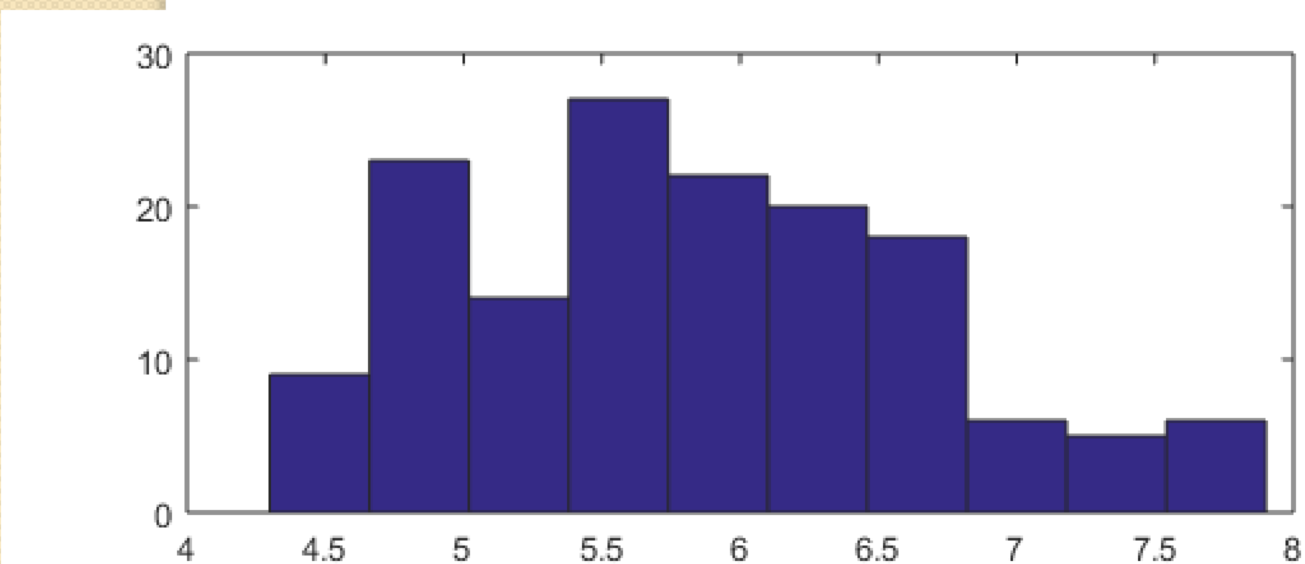
- Index 1 through 9

5.1 4.4 4.4 5.0 5.1 4.9 5.0 4.6 5.0

. . . .

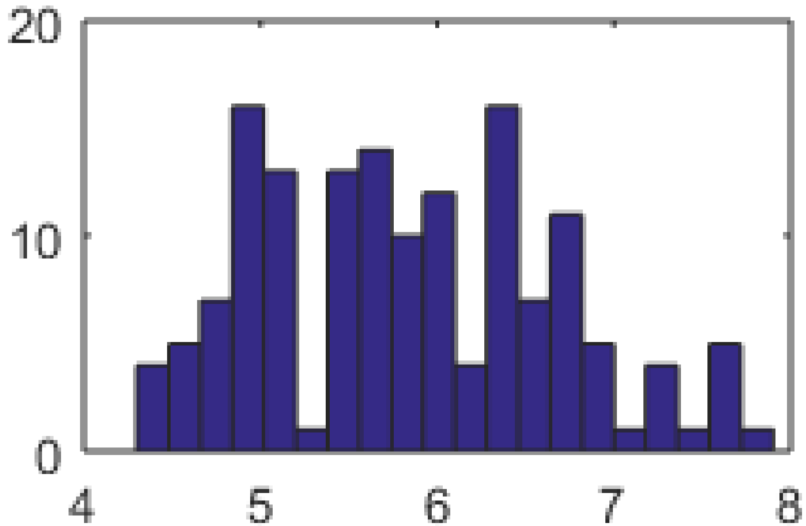
- Index 142 through 150

6.7 6.3 6.5 6.5 7.3 6.7 5.6 6.4 6.5

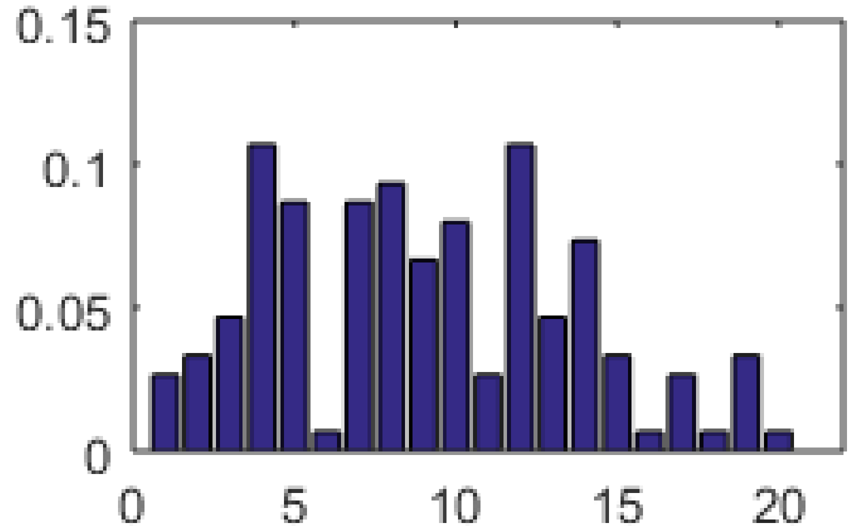


Histogram: range is divided in n(=10) bins; numbers of objects falling in bins are presented by bars.

B) Feature as random variable, 2:



(a)

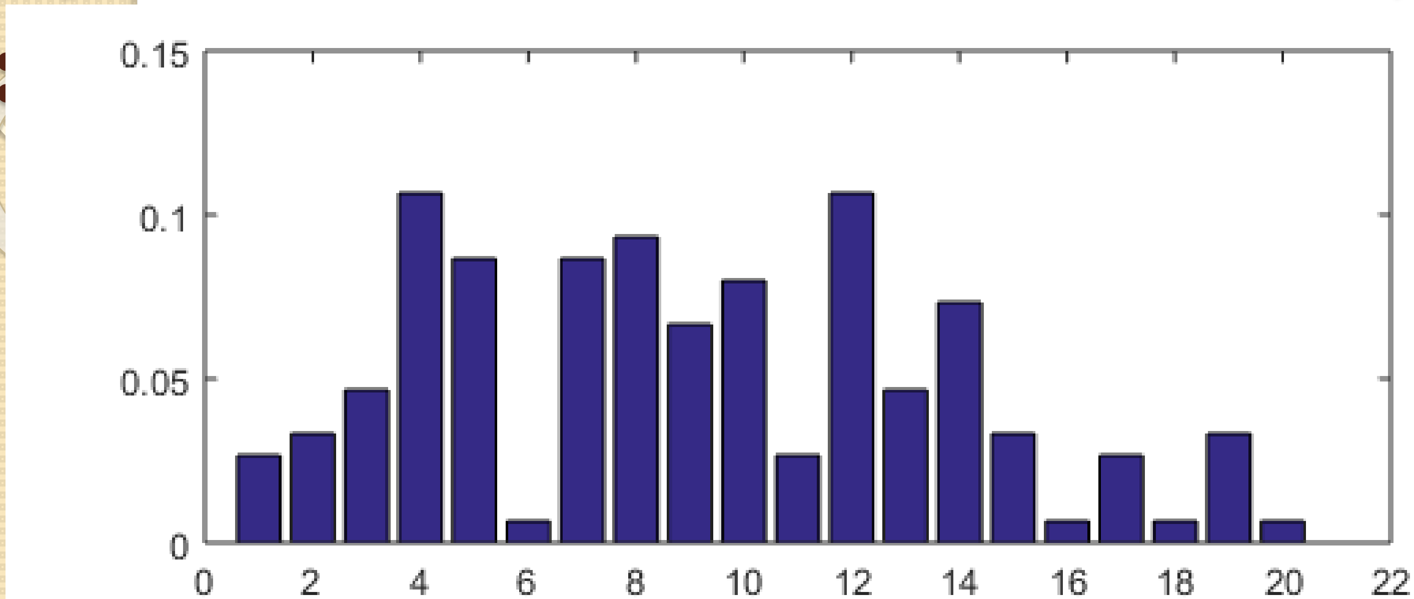


(b)

Histogram: (a) range is divided in n(=20) bins; numbers of objects falling in bins are presented by bars.

Relative histogram: **(b)** bars express proportions of objects in the bins (sum to 1).

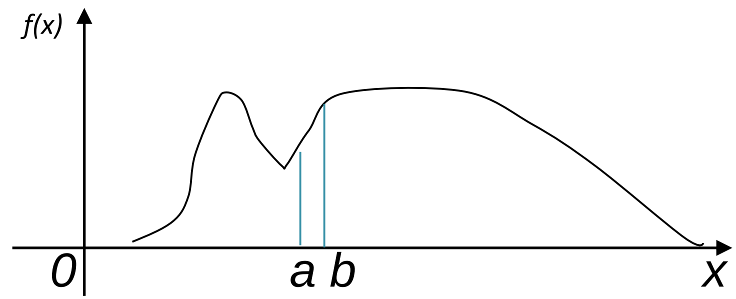
B) Feature as random variable, 3:



Relative histogram: bars express proportions of objects in the bins.

Density function, an abstraction of histogram at *N* and *n* tending to infinity: a measurable function (curve) *f*(*x*) such that $\int_{-\infty}^{+\infty} f(x)dx = 1$.

B) Feature as random variable, 4:

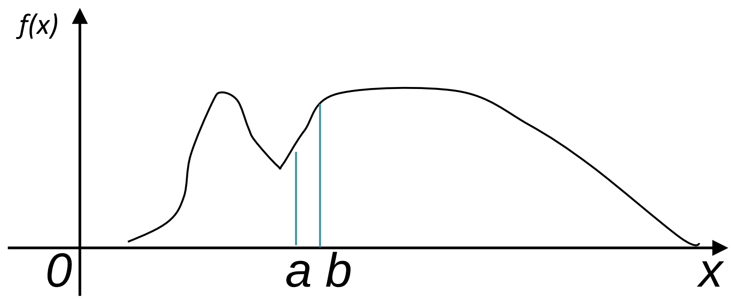


Density function, an abstraction of relative histogram at *N*, *n* tending to infinity: a measurable function *f(x)* such that

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

$$\int_a^b f(x)dx = \text{probability of the variable to fall in } [a,b]$$

B) Feature as random variable, 5:



Math: Random variable=Density function

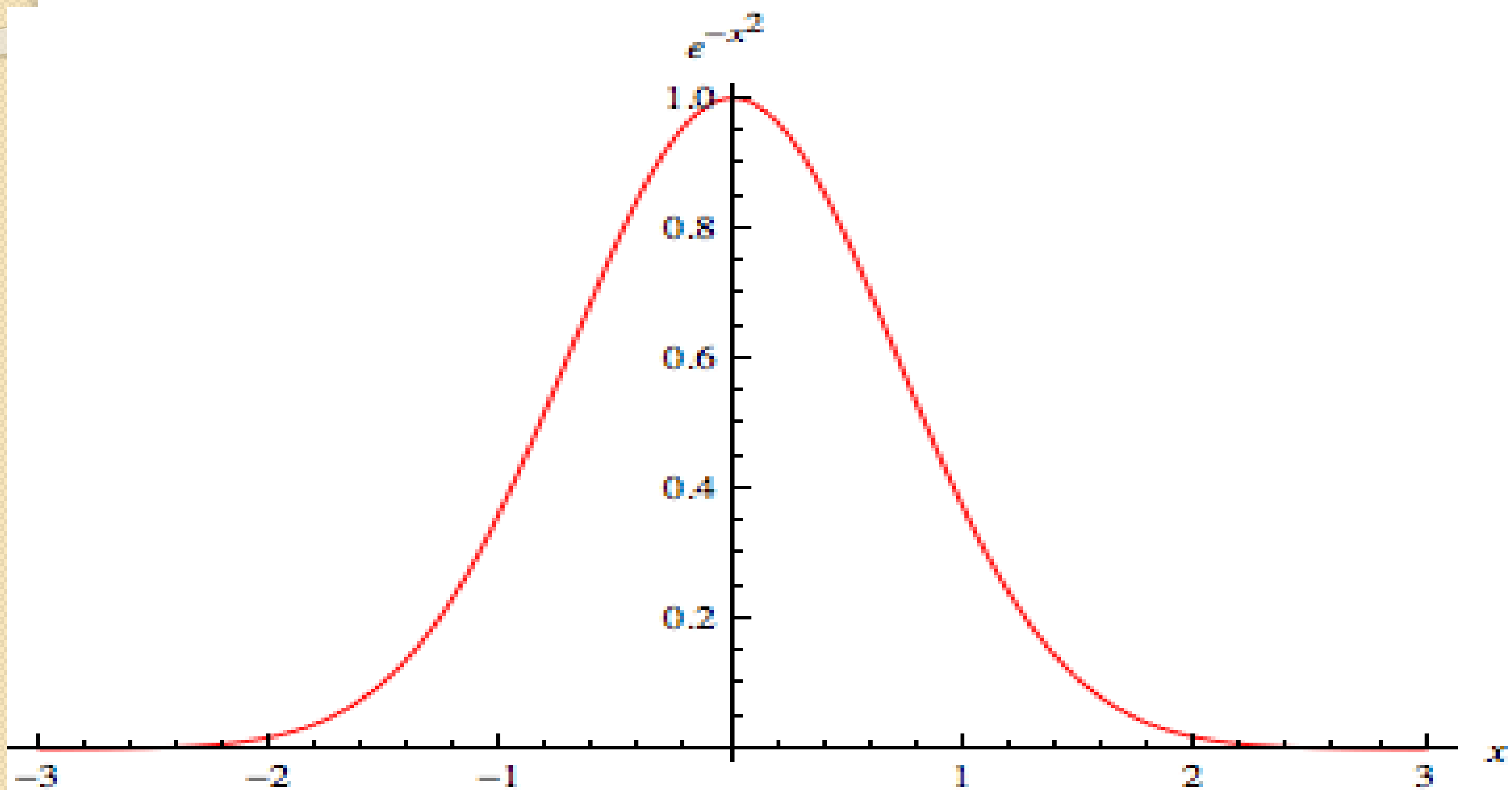
Pro: (a) Universal, does not depend on set I

(b) Probability theory can be used

Con: (c) Objects are implicit

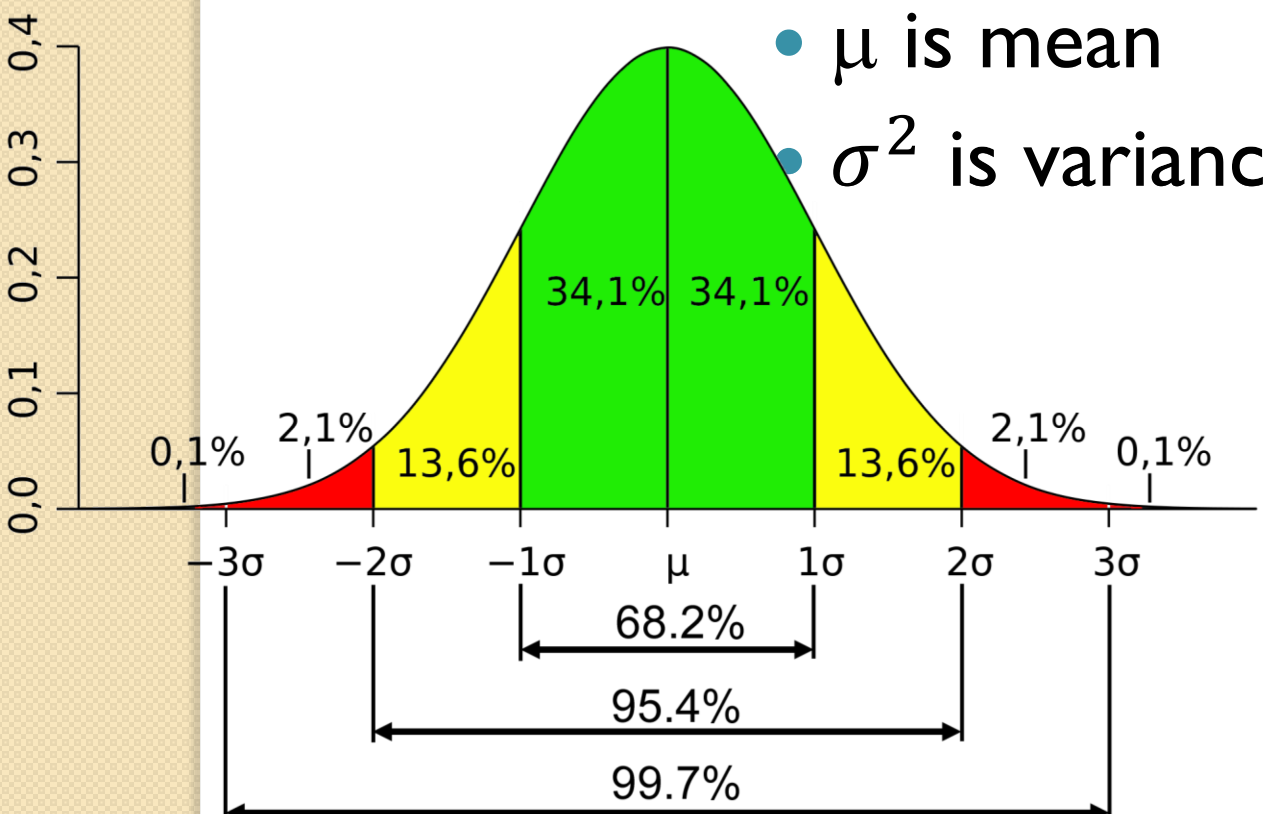
B) Popular density functions: Gaussian N(0,1)

$$f(x) = \exp\{-x^2\}$$



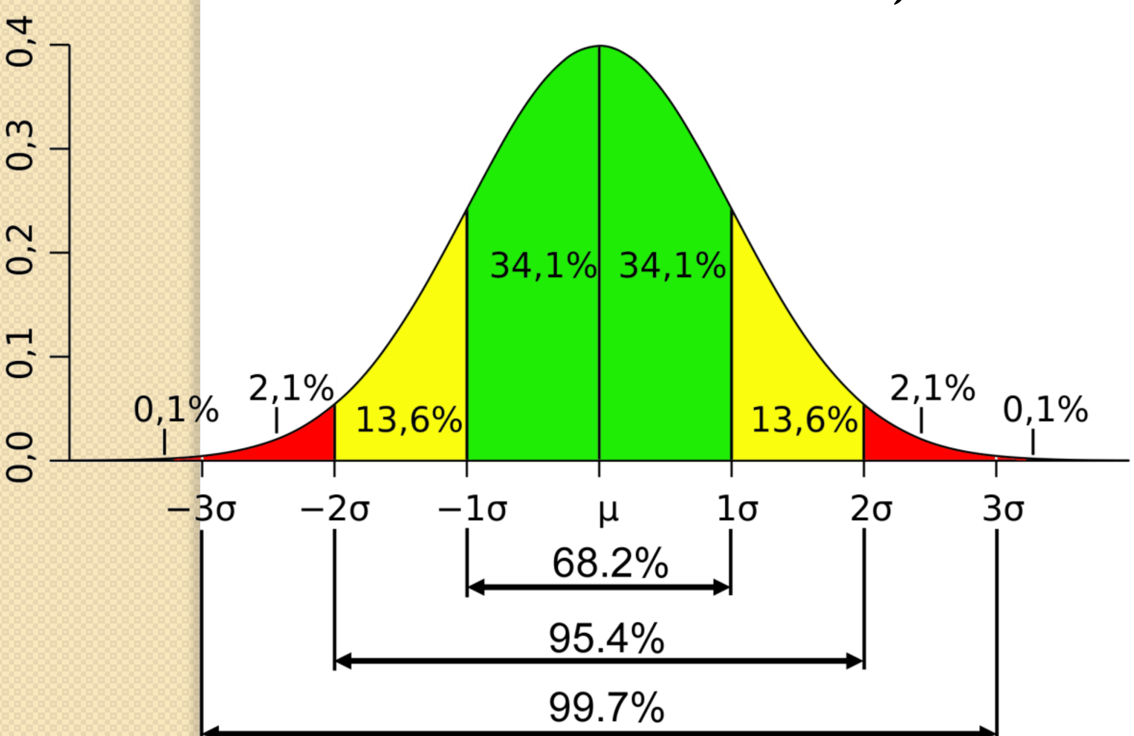
B) Popular density functions: general Gaussian N(μ,σ)

- $f(x)=\exp(\frac{-(x-\mu)^2}{2\sigma^2})/\sqrt{2\pi\sigma^2}$
- μ is mean
- σ^2 is variance



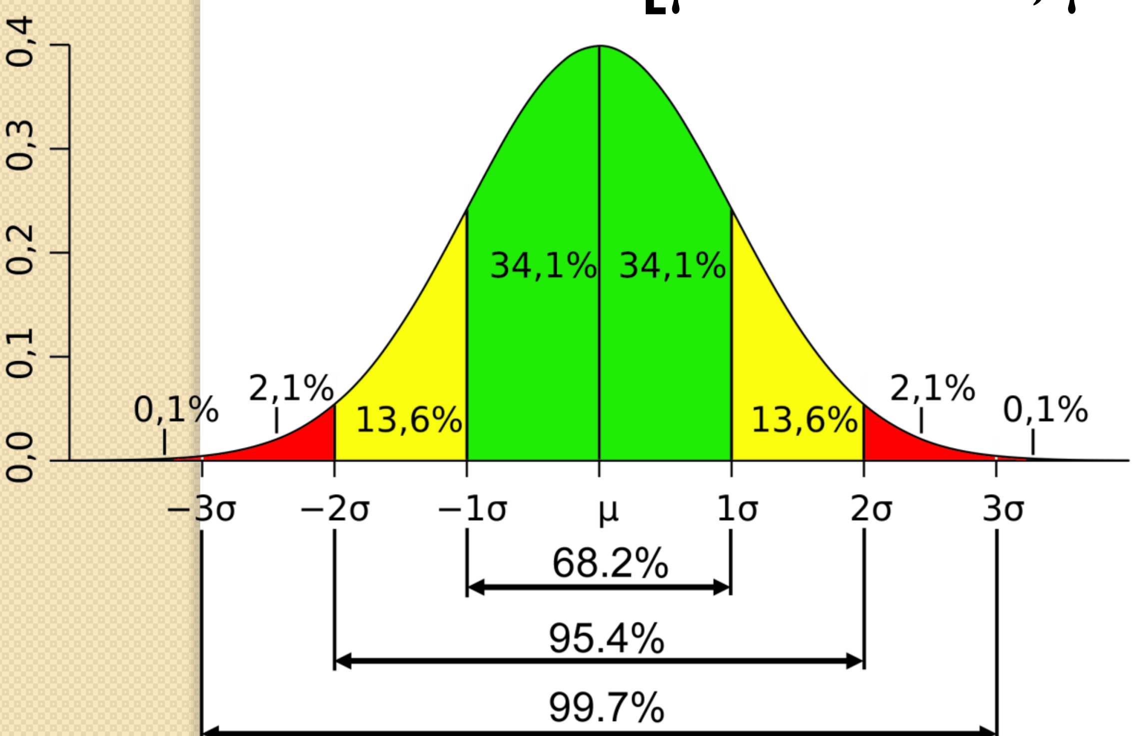
B) General Gaussian N(μ,σ)

- Bell curve (symmetric over μ)
- σ^2 is variance, σ is standard deviation (same scale)
- 2σ rule, 3σ rule



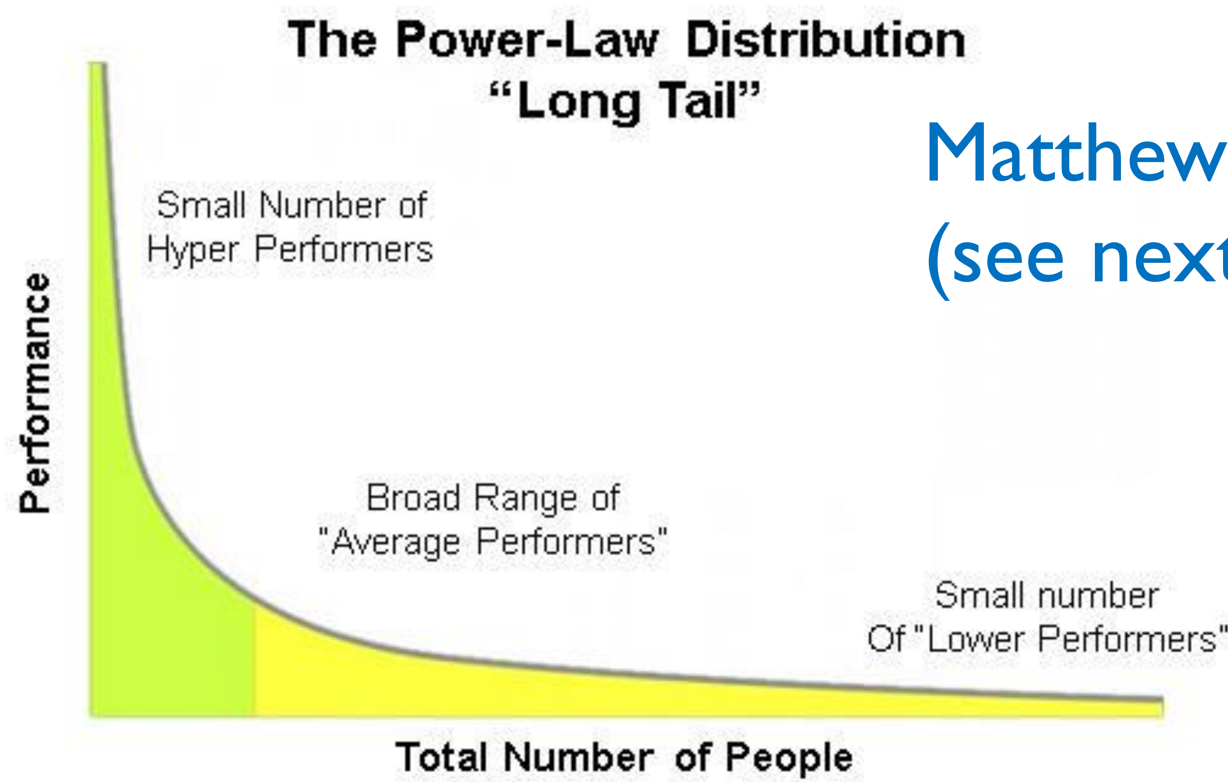
B) General Gaussian N(μ,σ)

- Bell curve (symmetric over μ)
- Central interval to account for 0.95=95% of the area:
- $[\mu - 1.96\sigma, \mu + 1.96\sigma]$



B) Popular density functions: **power law**

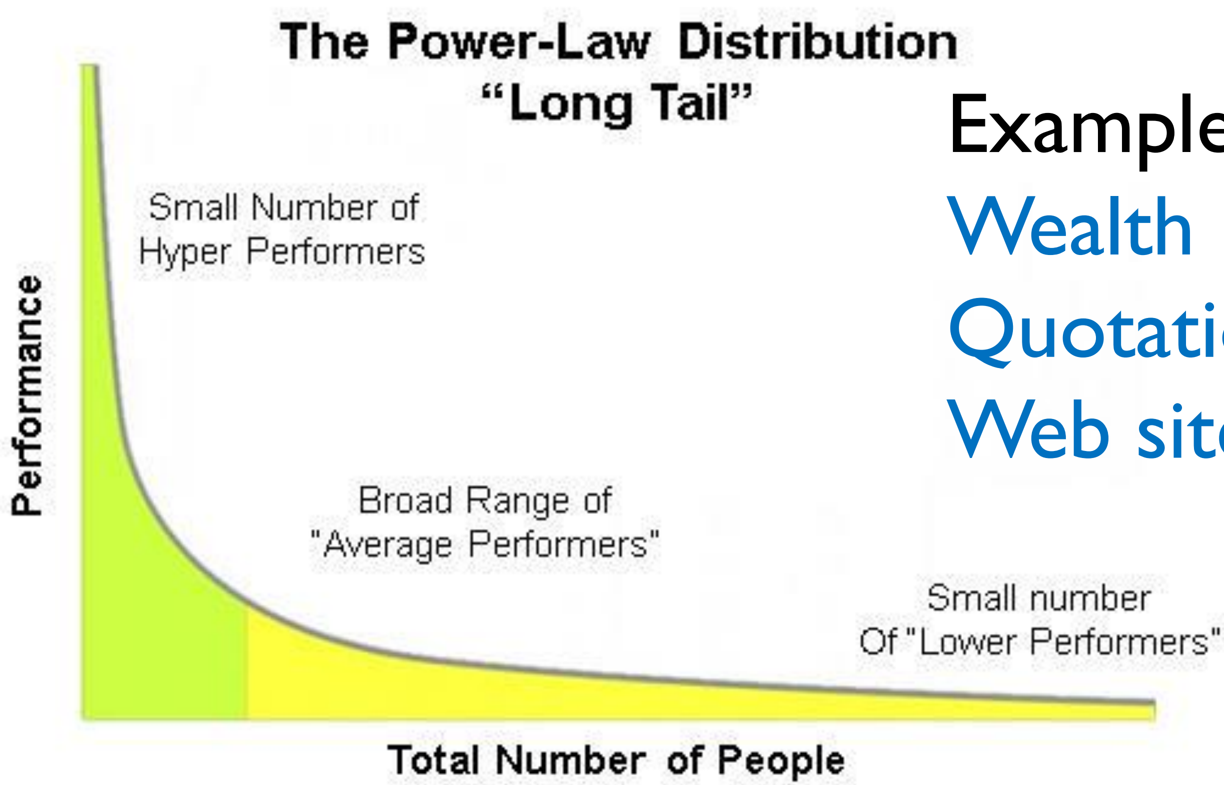
- $f(x)=cx^{-\alpha}$
- α the steepness
- Scale-free (why? Can you tell?)



Matthew effect
(see next slide)

B) Power law: Matthew effect

- **For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken even that which he hath.** Matthew Gospel 25:29



Examples:

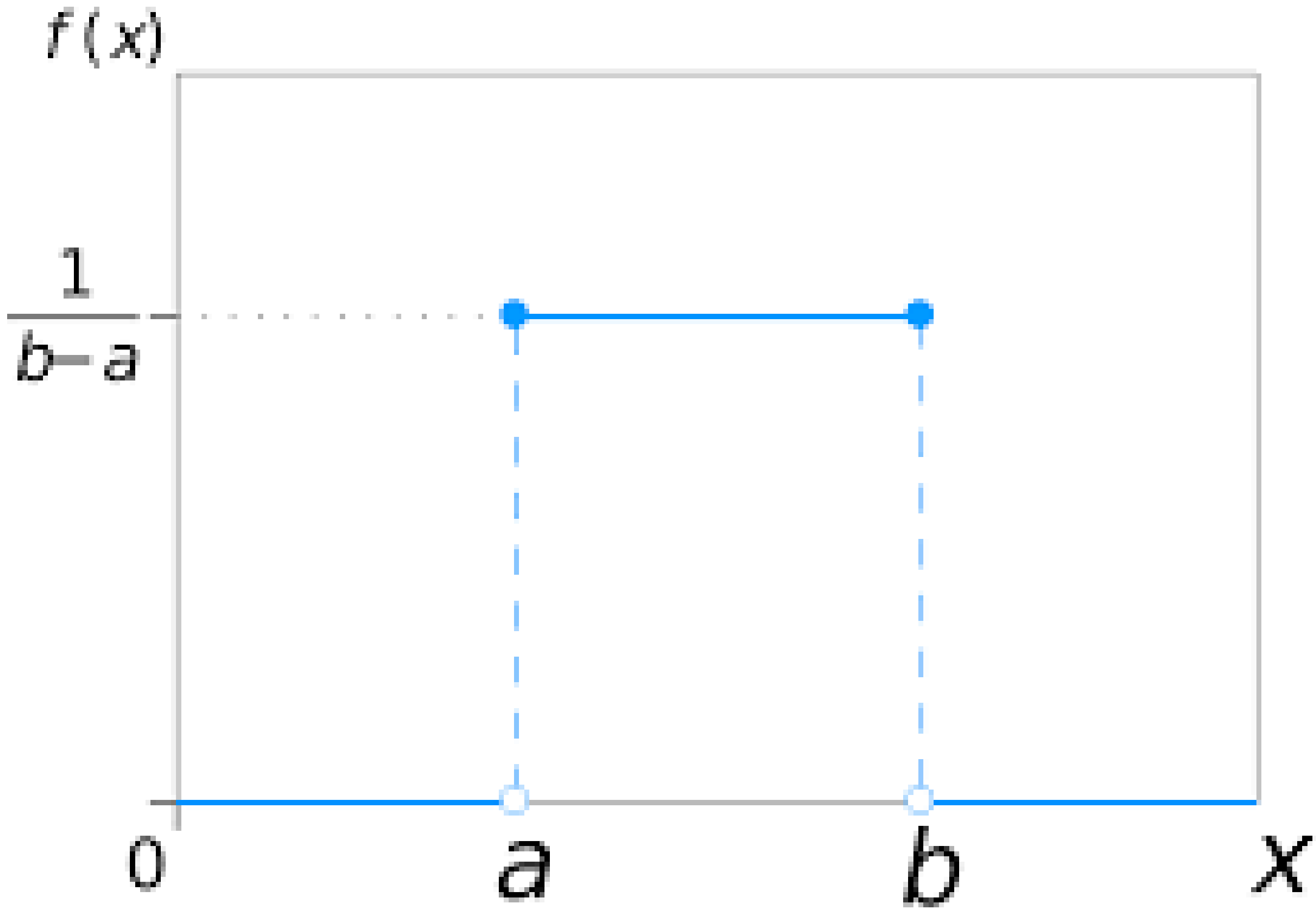
Wealth

Quotations

Web site popularity

B) Popular density functions: **uniform** distribution over [*a*, *b*] interval

Why is that?



Review of Lecture 1

- Administration
- Brief **history** of Data Science
- Three examples of **data analysis**: two successful and one not
- **Goal** and contents of the class
- **Data and metadata**: Iris dataset and problems of its analysis
- Two formalizations of the concept of feature: **vector/mapping** and **random variable/density function**

Keeping up: How to prepare yourself to the next lecture:

- After the lecture, put down **main concepts** that have been discussed in the lecture and think a few minutes of what do they mean

- Just before the next lecture: Take a few minutes and look through the slides of the previous lecture

Home work 1:

- **1.** Each to form/join a team of one, two or three; the team **finds a meaningful dataset** of their liking **on the internet**: say, by Googling "data analysis dataset"
- Number of entities ≥ 100, of features ≥ 7
- **No missings**
- **No Irivine ML repository**
- **The dataset is to be approved by me.**

- **2. Start writing a team's report file**
- Project title page
- Section 1.
  - Explanation of the choice of the dataset
  - Information of the dataset: features, number of entities, source address, examples of problems