

# Statistical Data Analysis – a course map

---

**Jiří Kléma**

Department of Computer Science,  
Czech Technical University in Prague



<http://cw.felk.cvut.cz/wiki/courses/b4m36san/start>

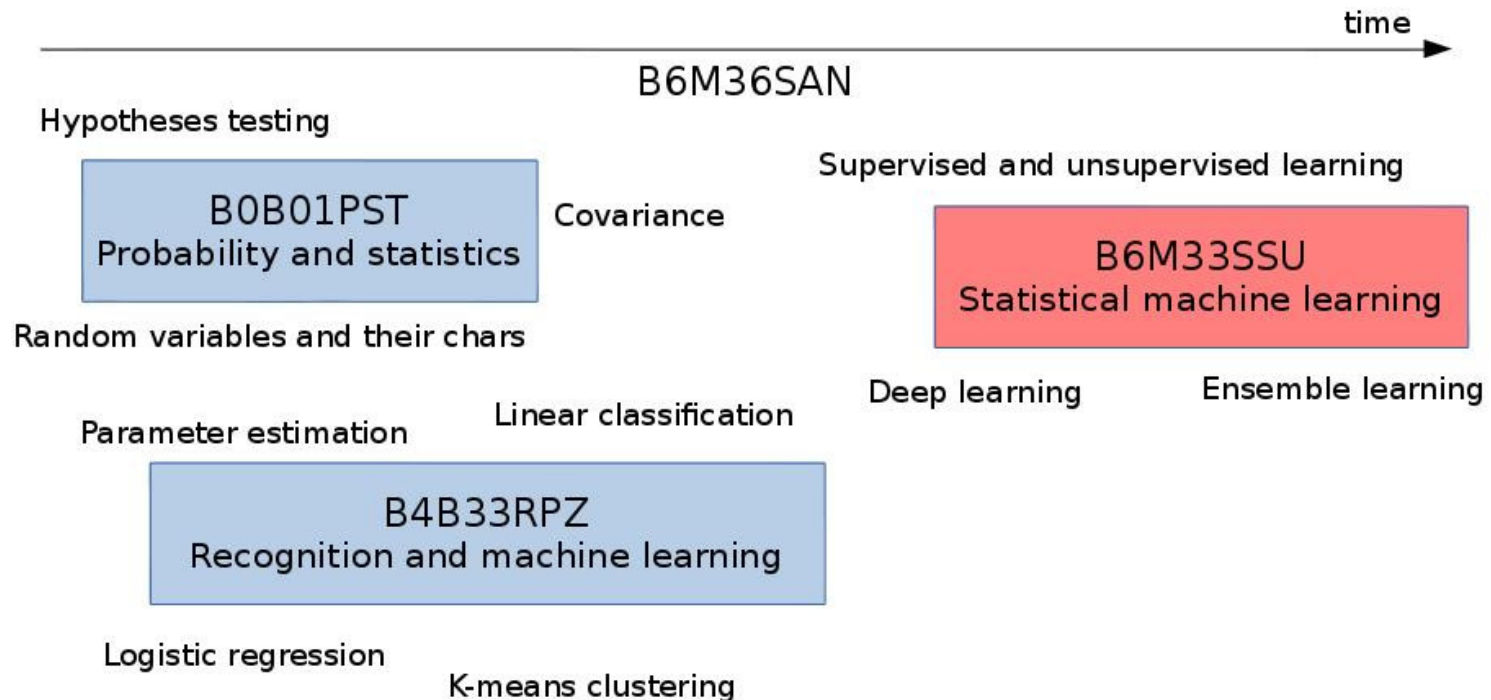
# B4M36SAN

---

- Purpose

- This course mainly aims at the statistical methods that help to understand, interpret, visualize and model potentially high-dimensional data. It works with R environment.

- Interactions with other courses



# Teachers

---



Doc. Jiří Kléma  
CTU, Dept. of Computer Science  
klema@fel.cvut.cz



Dr. Tomáš Pevný  
CTU, Dept. of Computer Science, CISCO Technical Leader  
pevnytom@fel.cvut.cz

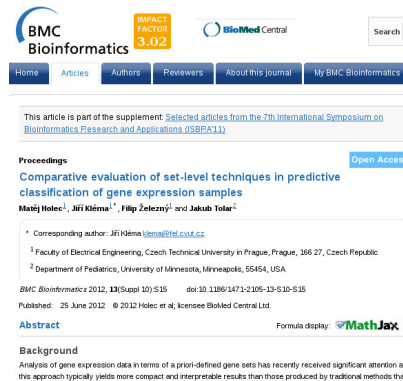


Doc. Zdeněk Míkovec  
CTU, Dept. of Computer Graphics and Interactions  
xmikovec@fel.cvut.cz



Ing. Anh Vu Le  
CTU, Dept. of Computer Science  
lequyanh@fel.cvut.cz

# IDA Highlights



publications



organizing conferences

SUPREME



software projects

# The key terms

---

- Multivariate statistical analysis
  - concerned with data that consists of **sets** of measurements on a number of individuals,
  - statistical approach based on **stochastic data models**
    - \* a certain model is assumed (a class of models),
    - \* its parameters are learned based on data,
  - more than independent testing of the individual variables (i.e., univariate tests known from introductory statistical courses),
  - intertwined variables, **examined simultaneously**,
  - not only the extensions of univariate and bivariate procedures,
  - examples: multivariate analysis of variance, multivariate discriminant analysis.

# The key terms

---

- Applied statistics
  - in general, rather a branch of study than a course,
  - in here, the course could be understood as an opportunity to bring the (previously learned) methods to practice,
  - in labs, stress on applications and their implementation in R.
- Statistical inference/learning
  - close interaction with (statistical) machine learning,
  - sometimes it is difficult to distinguished these two fields
    - \* as their goals are interchangeable,
  - the most striking distinctions
    - \* different schools – statistics is a subfield of mathematics, machine learning is a subfield of computer science,
    - \* different eras – for centuries versus modern,
    - \* different degree of assumptions – larger versus smaller.



## B4M36SAN and quotes/jokes

---

:: Data do not give up their secrets easily. They must be tortured to confess.  
Jeff Hopper, Bell Labs

:: All models are wrong, but some models are useful.  
George Box, Princeton University

:: There are two kinds of statistics ...  
... the kind you look up and the kind you make up.  
Rex Stout, writer

:: What is the difference between statistics, ML, AI and data mining?  
unknown author

# Changes in this year's run

---

- Mainly as a reaction to feedback from students,
- Lectures
  - the conceptual change in 2020: no lectures aimed at a particular branch of study.
- Labs
  - before 2020: 9 small assignments, nearly all the labs aimed at them
    - \* strength was that every (major) topic has been touched,
    - \* weakness was that a reasonable part of students felt lost,
  - now: only 4+1 assignments
    - \* there will be more supervised exercises in the labs,
    - \* weakness is that the assignments do not cover all the course content,
    - \* the final assignment can be motivated by your branch of study (basically, data science, cybersecurity, bioinformatics, HCI).
  - labs concluded with a solved problem
    - \* exactly the same form as exam questions.



# Syllabus

---

#	Lect	Content
1.	JK	Introduction, course map, review of the basic stat terms/methods.
2.	JK	Multivariate regression (continuous, linear regression, p-vals, overfitting)
3.	JK	Multivariate regression (non-linear, polynomial and local regression).
4.	JK	Multivariate confirmation analysis (ANOVA and MANOVA).
5.	JK	Discriminant analysis (categorical, LDA, logistic regression).
6.	JK	Dimension reduction (PCA and kernel PCA).
7.	JK	Dimension reduction (other non-linear methods).
8.	JK	Spare lecture.
9.	TP	Anomaly detection.
10.	TP	Robust statistics.
11.	ZM	Empirical studies, their design and evaluation. Power analysis.
12.	JK	Clustering (basic methods).
13.	JK	Clustering (advanced methods, spectral clustering).

# R package

---

- **R – the platform selected for labs**

- the leading tool for statistics,
- one of the main tools in data analysis and machine learning,
- it is free, open-source and platform independent,
- a large community of developers and users
  - a great variety of libraries, tutorials, mailing lists,
- easy to integrate with other languages (C, Java, Python),
- we actually use it,
- bottlenecks in memory management, speed, and efficiency,

- alternatives

- **Python** with its data analysis libraries (more general use),
- **Matlab** (popular at FEL for its forte in control, Simulink etc.).

# The key prerequisites – a brief review

---

- probability, independence, conditional probability, Bayes theorem,
- random variables, random vector,
- their description, distribution function, quantile function,
- categorical and continuous random variables,
- characteristics of random variables,
- the most common probability distributions,
- random vector characteristics, covariance, correlation, central limit theorem,
- measures of central tendency and dispersion, sample mean and variance,
- point and interval estimates of population mean and variance,
- maximum likelihood estimation, EM algorithm,
- statistical hypotheses testing,
- parametric and non-parametric tests,
- multiple comparisons problem, family wise error rate and false discovery rate.

## Exam – the prerequisites make a part of it

---

- Sample questions (see the course web page for a larger list)
  - Explain in your own words the meaning of *p-value*. Assume that a p-value of a test is 0.028. What is the probability that its  $H_0$  does not hold? Does it have any connection with the level of significance  $\alpha$ ?

## Exam – the prerequisites make a part of it

---

- Sample questions (see the course web page for a larger list)
  - Explain in your own words the meaning of *p-value*. Assume that a p-value of a test is 0.028. What is the probability that its  $H_0$  does not hold? Does it have any connection with the level of significance  $\alpha$ ?
  - $p = P(\text{observation like this or more extreme} | \text{null}) = P(o|H_0)$
  - $P(H_0|o) = \frac{P(o|H_0)P(H_0)}{P(o)} = \frac{P(o|H_0)P(H_0)}{P(o|H_0)P(H_0)+P(o|H_a)(1-P(H_0))}$
  - $H_0$  probability decreases with decreasing p-value of a correct statistical test, however, it is also a function of unexpectedness of the alternative hypothesis and the effect size (both can be hidden variables),
- an illustrative example: Did the sun just explode?
  - <https://xkcd.com/1132/>
  - $H_0$ : the sun did not change,  $H_a$ : the sun has gone nova,
  - $P(H_0) = .999$ ,  $P(o|H_0) = .028$ ,  $P(o|H_a) = 0.972 \dots P(H_0|o) = .97$ .

# The plan for the first half of the course ...

---

- assess strength of relationship between a pair of variables based on their type
  - independent (causal) and dependent (effect) variable,
  - rejection of null hypothesis does not imply causal relationship,
- all of them can be generalized towards multivariate statistics.

		dependent variable	
		categorical	continuous
independent variable	categorical	contingency table chi-square test	analysis of variance
	continuous	LDA logistic regression	correlation regression

# The main references

---

:: Resources (slides, scripts, tasks) and reading

- G. James, D. Witten, T. Hastie and R. Tibshirani: **An Introduction to Statistical Learning with Applications in R.** Springer, 2014.
- A. C. Rencher, W. F. Christensen: **Methods of Multivariate Analysis.** 3rd Edition, Wiley, 2012.
- T. Hastie, R. Tibshirani and J. Friedman: **The Elements of Statistical Learning: Data Mining, Inference, and Prediction.** Springer, 2009.