RWTH AACHEN UNIVERSITY

WS 2018/19

Fundamentals of Big Data Analytics

Prof. Dr. Rudolf Mathar
Dr. Arash Behboodi
Dipl.-Ing. Markus Rothe

- Register at L2P / RWTH online
- Dates Lectures Friday 8:30 – 10:00
  Exercises Friday 10:30 – 11:15

Fr. 12.10.18
Fr. 19.10.18
Fr. 26.10.18 (AB)
Fr. 2.11.18 (AB)
Fr. 9.11.18
Fr. 16.11.18
Fr. 23.11.18 (AB)
Fr. 30.11.18 (AB)
Fr. 7.12.18 (AB)
[Fr. 14.12.18] ← Tag der ET&IT
Fr. 21.12.18
Fr. 11.01.19
Fr. 18.01.19
Fr. 25.01.19
Fr. 01.02.19

- Written examination
  7.3.2019, 10:30 – 12:00, Aula 1
  Register at RWTHonline 3.12.18 – 10.1.19
  (Check!)

Thinking the Future
Zukunft denken

- web page: www.ti.rwth-aachen.de
- Material:
  - Lectures notes: → teaching → FBDA
  - Handwritten sheets will be published
  - Recording (video) available from WS 17/18
- Students from which course of studies?
  Attend exam?
- List of textbooks (see separate sheet)
- Consultation hours
  - upon agreement
  - bigdata@ti.rwth-aachen.de

Fundamentals of Big Data Analytics

References

[1] Charu C. Aggarwal. *Data Mining*. Springer International Publishing, 2015.

[2] Afonso S. Bandeira. Ten Lectures and Forty-Two Open Problems in the Mathematics of Data Science. `http://www.cims.nyu.edu/~bandeira/TenLecturesFortyTwoProblems.pdf`, 2008. [Online].

[3] Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, New York, NY, 2nd edition, 2009.

[4] Jurij Leskovec, Anand Rajaraman, and Jeffrey D. Ullman. *Mining of massive datasets*. Cambridge University Press, Cambridge, second edition edition, 2014.

[5] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Probability and mathematical statistics. Academic Press, London ; New York, 1979.

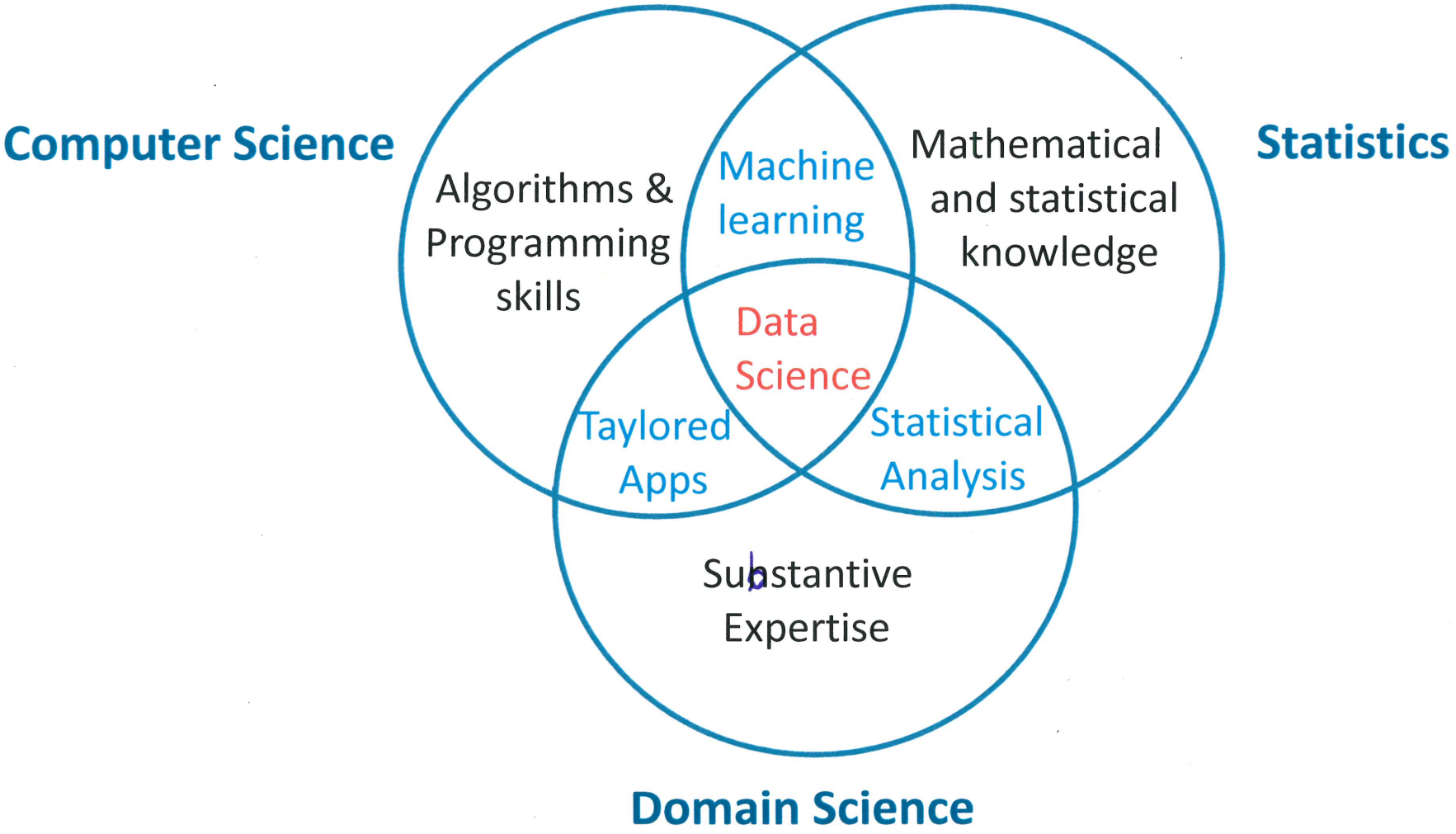
[6] C. D. Meyer. *Matrix analysis and applied linear algebra*. SIAM, Philadelphia, 2000.

[7] Kevin P. Murphy. *Machine learning: a probabilistic perspective*. Adaptive computation and machine learning series. MIT Press, Cambridge, MA, 2012.

RWTH AACHEN UNIVERSITY

Table of Contents

Skills for Data Science



What is Data Science ?

As an emerging field not yet well defined, but incorporates elements of

- Exploratory data analysis and visualization
- Machine learning and statistics
- High performance computing technologies for dealing with scale

1. Introduction

(Big) data analytics:

"The discovery of models for data to extract information, draw conclusions and make decisions."

"Model" can be one of several things:

- Statistical model, underlying distribution from which the data is drawn.

  Ex. Given a set of numbers, assume Gaussian, estimate the mean and variance.

  Model: $N(\mu, \sigma^2)$, independent observations.

- Use data as a training set for algorithm of machine learning, e.g., Bayes nets, support vector machine, decision tree, etc.

  Ex. "Netflix challenge": devise an alg. to predict the rating of movies.

- Extract the most prominent features of the data and ignore the rest.

  Ex. Feature extraction, similarity, PCA

- Summarization of features

  Ex. – Page rank (Google's web mining)

  – Clustering (assign points to cluster heads.)

In large/huge random samples unusual features occur which you deem unusual but are purely random.

Example:

Find evil-doers by looking for people who both were in the same hotel on two different days.

Assumptions:

1. 10^5 hotels
2. Everyone goes to a hotel one day in 100.
3. 10^9 people
4. People pick hotels and days at random (independently.
5. Examine hotel records for 1000 days.

Prob. that any 2 people visit a hotel on any site day $= \frac{1}{100}\frac{1}{100}$

Prob. that they pick the same hotel: $\frac{1}{10^4} \cdot \frac{1}{10^5} = \frac{1}{10^9} = 10^{-9}$

Prob. that 2 people visit the same hotel on 2 diff. days:

$$(10^{-9})^2 = 10^{-18}$$

Cardinality of the event space:

Pairs of people : $\binom{10^9}{2}$

Pairs of days : $\binom{10^3}{2}$

Expected no. of evil-doing events (use $\binom{n}{2} \sim \frac{n^2}{2}$)

$$\binom{10^9}{2} \cdot \binom{10^3}{2} \cdot 10^{-18} \approx 5 \cdot 10^{17} \cdot 5 \cdot 10^5 \cdot 10^{-18}$$
$$= 25 \cdot 10^4 = 250\,000.$$

$\longrightarrow$ Bonferroni's principle