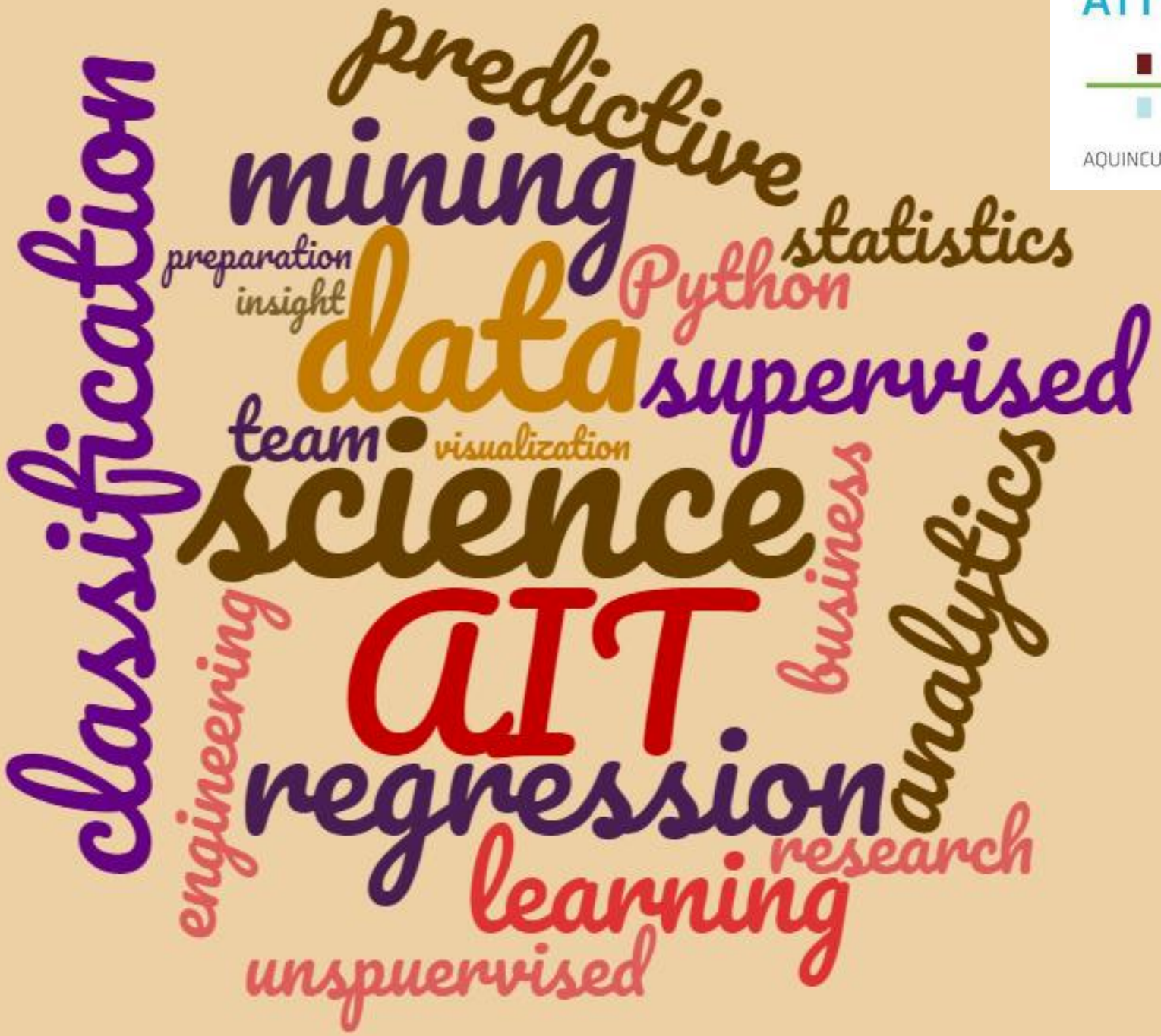
Data Science

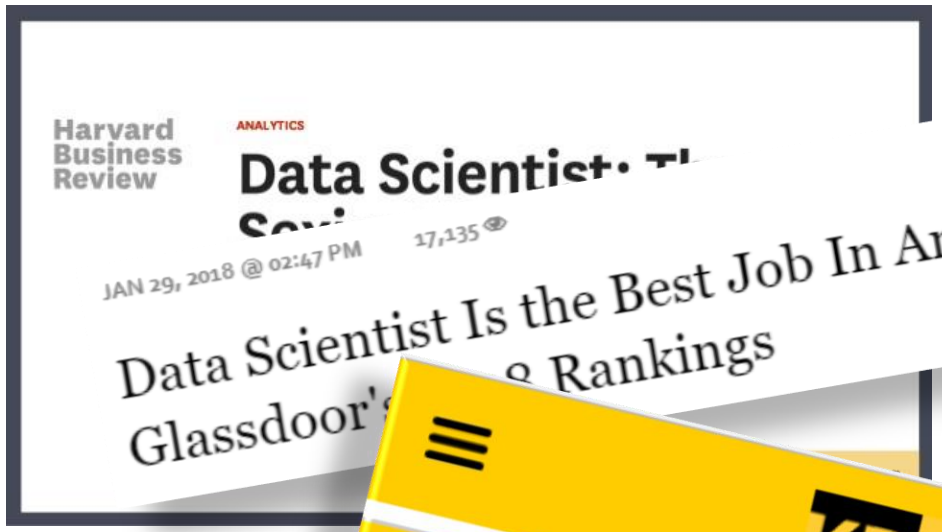


February 3, 2020.
Introduction

Roland Molontay

AIT-BUDAPEST
AQUINCUM INSTITUTE OF TECHNOLOGY

Worth learning data science



GLASSDOOR'S BEST JOBS IN AMERICA 2017

1. Data Scientist
5. Analytics Manager
6. HR Manager
7. Database Administrator

SOURCE: GLASSDOOR 50 BEST JOBS IN AMERICA

CAREER SHERPA

50 Best Jobs in America for 2019

Best Jobs | 2019 | United States

Share

	Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1	Data Scientist	$108,000	4.3/5	6,510	View Jobs
#2	Nursing Manager	$83,000	4/5	13,931	View Jobs
#3	Marketing Manager	$82,000	4.2/5	7,395	View Jobs
#4	Occupational Therapist	$74,000	4/5	17,701	View Jobs
#5	Product Manager	$115,000	3.8/5	11,884	View Jobs
#6	Devops Engineer	$106,000	4.1/5	4,657	View Jobs
#7	Program Manager	$87,000	3.9/5	14,753	View Jobs
#8	Data Engineer	$100,000	3.9/5	4,739	View Jobs
#9	HR Manager	$85,000	4.2/5	3,908	View Jobs
#10	Software Engineer	$104,000	3.6/5	49,007	View Jobs

1. **Data Scientist**
2. Site Reliability Engineer
3. Enterprise Account Executive
4. Product Designer
5. Product Owner
6. Customer Success Manager
7. Engagement Manager
8. Solutions Architect
9. Information Technology Lead
10. Scrum Master

2019's Most Promising Jobs in the U.S.

1. Data Scientist

Median Base Salary: $130,000

Job Openings (YoY Growth): 4,000+ (56%)

Career Advancement Score (out of 10): 9

Top Skills: Data Science, Data Mining, Data Analysis, Python, Machine Learning

Basic course information

- Data Science = Data Science 1 + Data Science 2 (for historical reasons)
- Credits: 2 + 2
- Contact hours: TU 16-18 + FR 16-18
- Instructor: Roland Molontay
  - 2015: MSc in Applied Mathematics (BME)
  - 2015 - 18: PhD studies in Network Theory (BME)
  - 2016: visiting PhD student at Brown University
  - 2018 - : junior research fellow at MTA-BME Stochastics Research Group

- E-mail: molontayr@gmail.com, data.science.spring.2020@gmail.com
- Teaching assistants/graders: Nóra Balogh, Nóra Stumphauser, Marcell Nagy

Syllabus

Some keywords: data types, data preparation, explanatory data analysis, supervised learning, classification, regression, model evaluation, clustering, recommender system, data visualization, case studies

Form of teaching: mostly lectures (with presentation), problem-solving session, computer-assisted problem solving (mostly in form of homework problems)

Course material: presented lecture slides (with oral explanation) + problem sheets + iPython notebooks

Plenty of useful materials are available online

Aim of the course

- Provide a broad overview of the field
- Learn about theory and use the methods in exciting real life datasets
- Excel in your interview for a junior data scientist position
  - Both in oral interview and take-home assignments

Recommended literature

- Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. 2005.
- James, Gareth, et al. *An introduction to statistical learning*. Vol. 112. New York: Springer, 2013.
- Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2014.
- Sammut, Claude, and Geoffrey I. Webb, eds. *Encyclopedia of machine learning and data mining*. Springer, 2016.

The books are uploaded in moodle in pdf form!

Requirements

- Attendance: there will be sign-up sheets every time
- MIDTERM (25%)
  - On Week 7
  - You can use your own „cheat sheets”/ formula sheets / notes
- FINAL (25%)
  - On Week 13
  - You can use your own „cheat sheets”/ formula sheets /notes
- HOMEWORK problems (25%)
  - There will be 4 HW sheets (in every two weeks)
  - Mostly programming problems
- PROJECTS in teams (25%)
  - In teams of 3 (2-4) students

Homework policy

- The homework must be your own work.
  - You can look up books / search for help online
    - If you use longer code snippets from online sources you must refer to the source
  - You are encouraged to help each other if one of you gets stuck, share some ideas with each other
    - You must not send your entire homework to your peers
    - Copying is forbidden but giving assistance is encouraged

Midterm / Final

- On week 7 / on week 13
- 90 minutes each
- 25% each
- Theoretical questions and numerical examples
- You can use your own „cheat sheets”/ formula sheets
  - As many hand-written notes as you wish

Projects

- A data science project in teams
  - Team: 3 (2-4) students
  - All the team members will get the same grade point
  - It is recommended to use a version control tool to share your codes with each other, e.g. https://github.com/
  - A teaching assistant will be assigned for each team to help you/ provide guidance.
- Choose a problem that all the team members are interested in
- Schedule
  - W4: forming teams
  - W6: project plan is ready
  - W8: milestone 1
  - W12: milestone 2
  - W14: classroom presentation

Expectations

- Delivering a sophisticated enough data science project
- Studying related work (related papers, projects)
  - Understand what others have done, attempt to not only reproduce the results but improve them in some aspects
- Implementing techniques that we have covered in class
- Try more models, evaluate them, find the best models
- Nice and shiny data visualization
- Optional but appreciated: using models/techniques that we have not covered in class

Deliverables

- W4: Team name + list of team members
- W6: **Project plan**
  - One page long report answering the following questions:
    - What is the vision? Why is the problem interesting?
    - What is the purpose of the project? What results do you expect?
    - What data do you plan to use? How do you plan to gather the data?
    - Are the data big enough and of suitable quality?
    - What data preparation steps do you plan to take?
    - What methodology, what models do you plan to use?
    - How would you visualize the results?

Deliverables II.

- W8: **Milestone 1**
  - Two page long report covering the followings:
    - Have you managed to gather the data? Do you have enough data of appropriate quality?
    - Did you collect the relevant related works? What useful information could you discover?
    - Initial data analysis steps
    - What next steps do you plan to take?
- W12: **Milestone 2**
  - Three page long report covering the followings:
    - Reviewing the related works
    - Data understanding and data preparation steps
    - More data analysis steps, implementing some models and evaluating them

Final deliverable: classroom presentation

- W14: oral presentation should include
  - Description of the problem, motivation
  - Some review of related works
  - Description of the data set
  - Data preparation steps
  - Modeling steps (what models, parameters of the models)
  - Evaluation of the models
  - Sophisticated visualization
  - Interpreting the results, conclusion

Final deliverable II.: codes

- W14: well-written codes (preferably an Ipython notebook)
  - Comments are necessary

Possible data sources

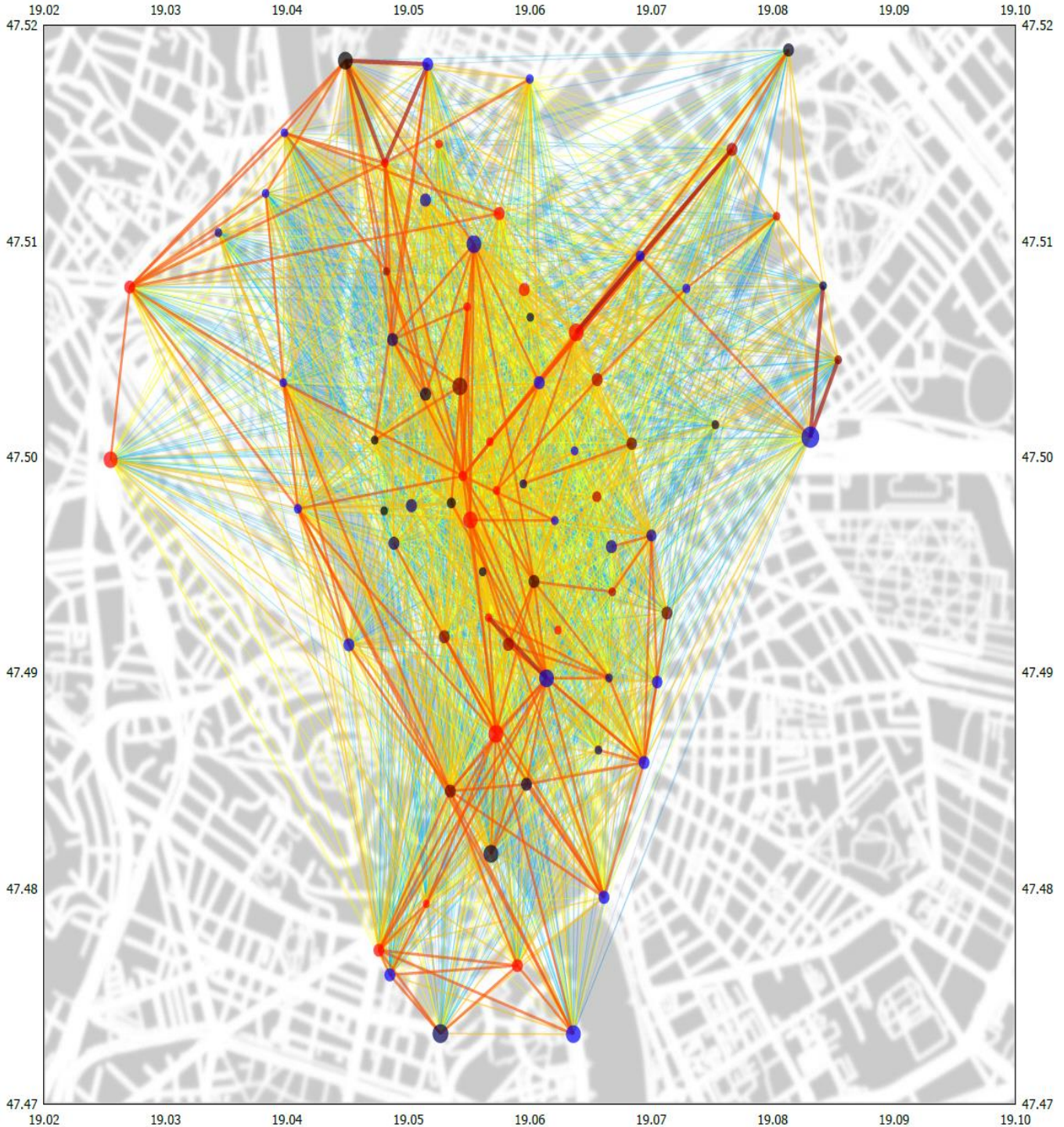
- Own measurements, independent data collection (e.g. using web scraping techniques)
- Using data repositories:
  - https://www.kaggle.com/competitions
  - https://www.kaggle.com/datasets
  - https://archive.ics.uci.edu/ml/index.php
  - http://www.kdnuggets.com/datasets/index.html https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research
  - https://www.google.com/publicdata/directory

IMPORTANT: USE DATA THAT YOU ARE INTERESTED IN!

Bike sharing system

- Busiest Route Prediction
- Docking Station Demand Prediction
- Several interesting projects:

https://dms.sztaki.hu/bubi/

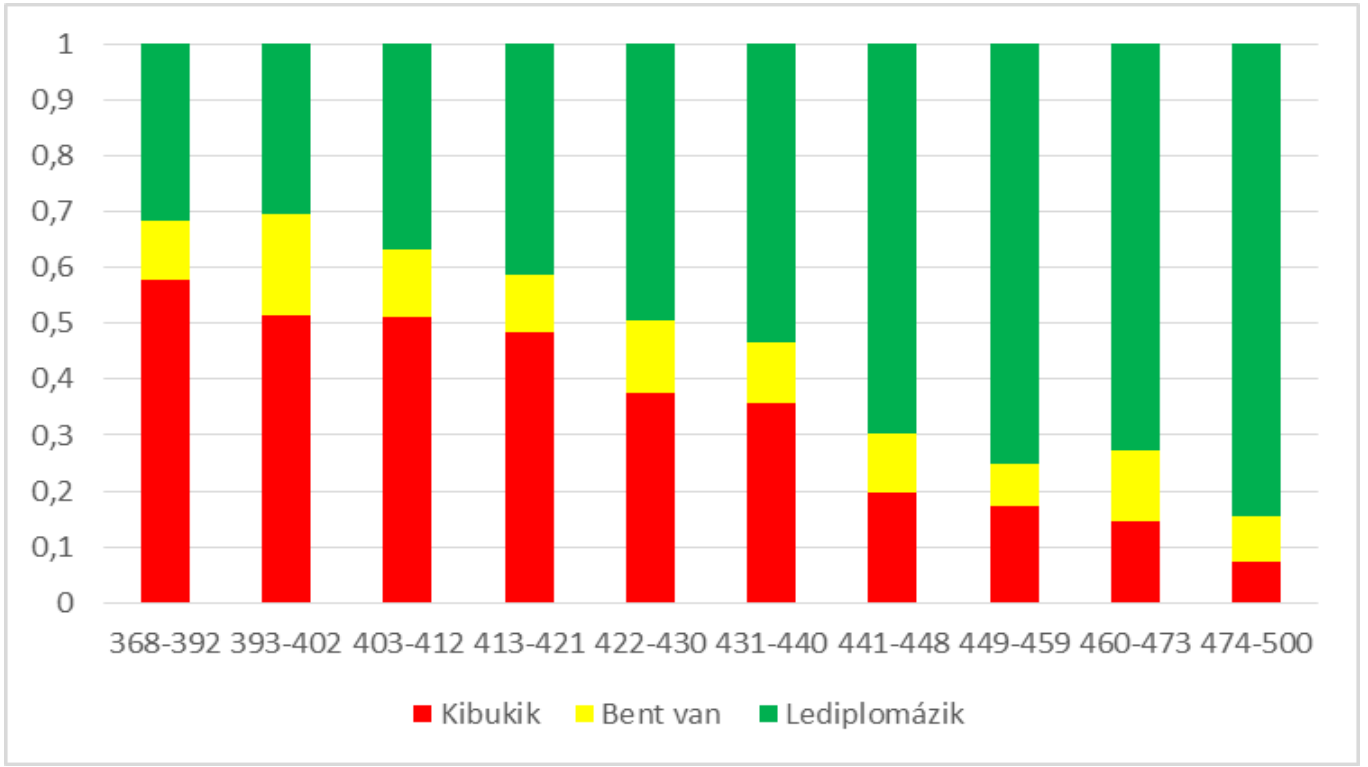


Marcell Stippinger

Predictive analytics of university performance

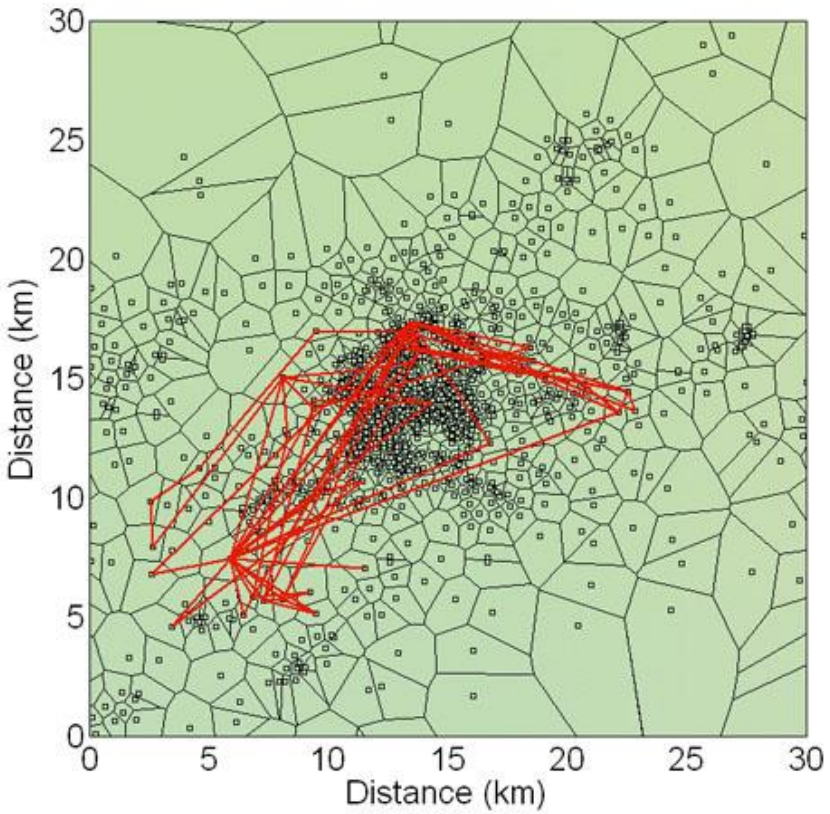
Math multiplicative	132.0
History multiplicative	82.0
Foreign language multiplicative	95.0
Program ID	9N-AM06
APS calculating method	SP+MP+EP
Financial status	Scholarship
Freshman or Re-enrolled	Freshman

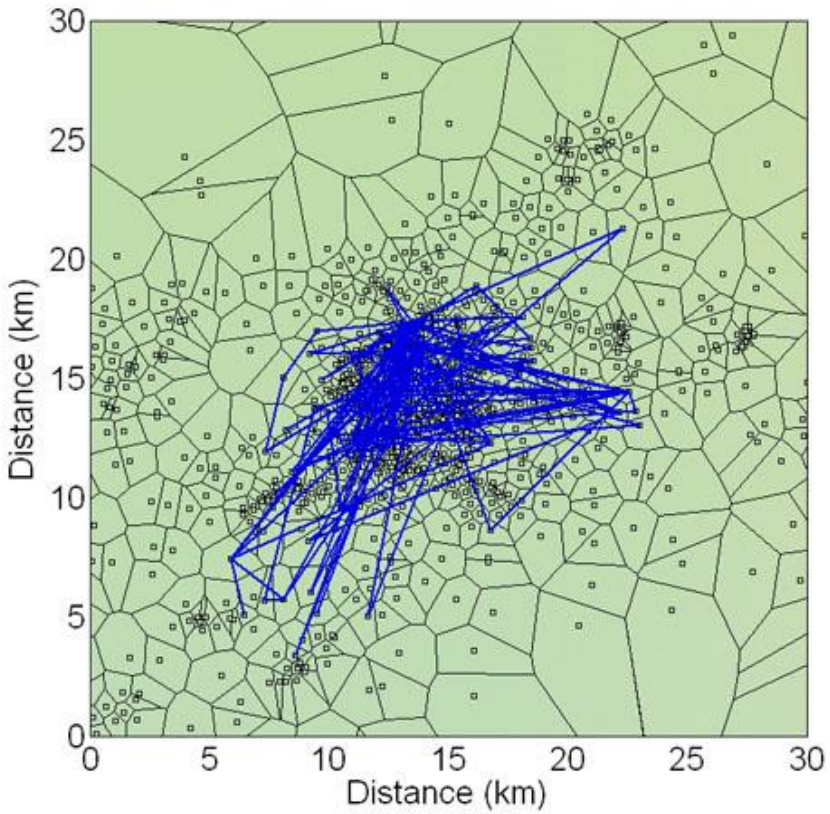
confidence(Graduated)	prediction(Status)
0.7056270445331028	Graduated



Marcell Nagy, Zoltán Barabás

Mobility patterns from cell phone data

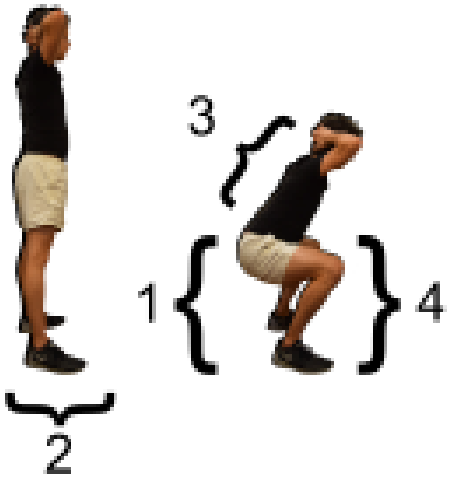




Barabási Lab

Automatic personal trainer for squatting

What Makes a Squat?

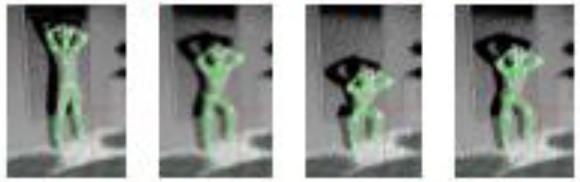


Squat components:

1. Depth: Angle formed between the ankle knee and hip while squatting
2. Stance: distance between feet and center of mass over your ankles
3. Back-Hip Angle: Angle formed between the knees hips and back throughout the squat
4. Knee-Toe Alignment: Location and movement of knees throughout the squat

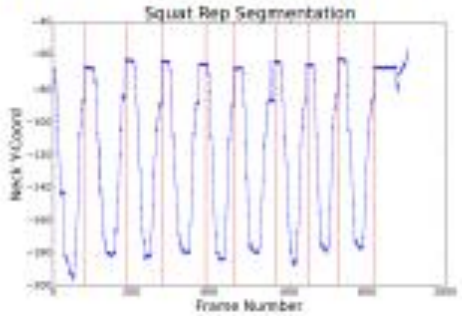
Classification Pipeline

1. Observe data with Kinect (10 reps)

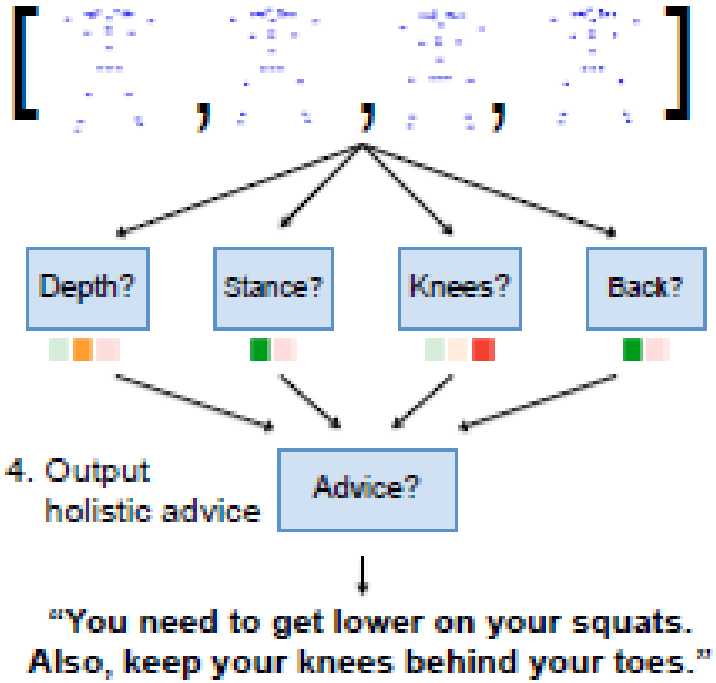


Example squat rep with resolution τ = 4

2. Segment data into discrete reps



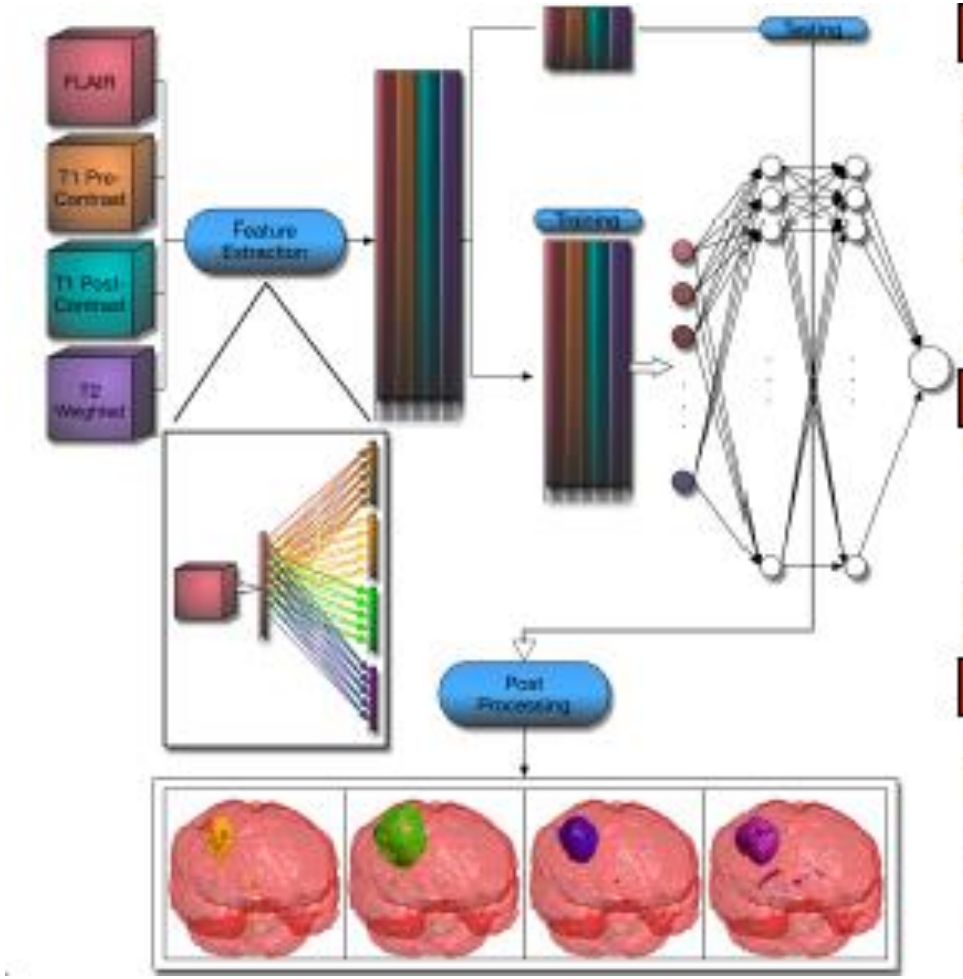
3. Run reps through component classifiers



Brandon Garcia, Russel Kaplan, Aditya Viswanathan

Check out their poster:
http://cs229.stanford.edu/proj2015/183_poster.pdf

Classification of types of brain tumor using MR images



Feature Space

Data Problem: Each patient has millions of voxels of information but only a few hundred patients. Consider voxels independently to increase n. (Bias ↑ Variance ↓)

$X (\in \mathbb{R}^{276 \times (\text{Vars} \cdot \text{Pools})}) \rightarrow y (\in \{0,1\}^{276 \times \text{Vars}})$

$X (\in \mathbb{R}^{(276 \cdot \text{Vars}) \times \text{Pools}}) \rightarrow y (\in \{0,1\}^{276 \cdot \text{Vars}})$

Convolution

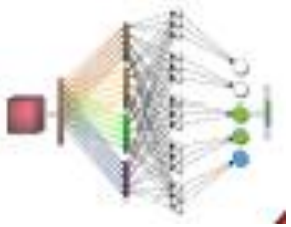
Convolution with filters to get features.

$$[f * g](t) = \int_0^t d\tau\, f(\tau)\, g(t - \tau)$$

Even though we consider each voxel completely independently, we still couple environmental information.

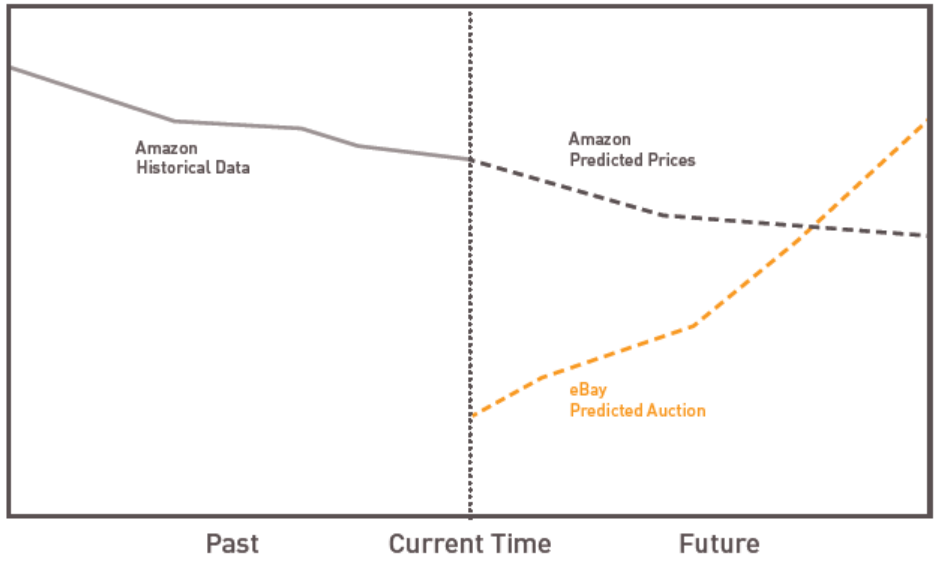
Neural Network

- 72 Input
- Binary Output
- 2 Hidden Layer
- 10 Neurons Ea.
- 4 Nets for Each Subregion

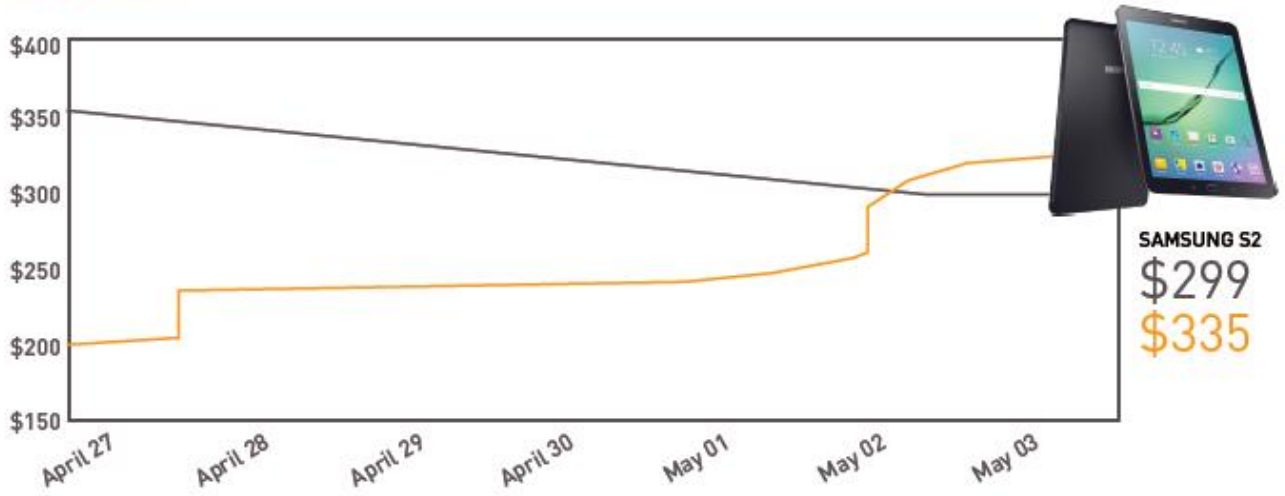


Zhao Chen
Darvin Yi
Tianmin Liu
Check out their poster: http://cs229.stanford.edu/proj2015/277_poster.pdf

Revealing arbitrage possibilites on Amazon and eBay markets



Results



SAMSUNG S2
$299
$335

GOPRO HERO SILVER
$329.99
$360

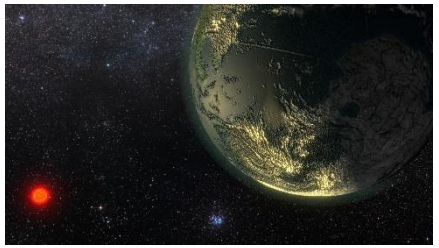
LEGO STAR WARS
$55.99
$62.9

THE MARTIAN
$12.99
$14.99

Angelia Wang, Chris Grimm, Adam Hoff, and Athyuttam Eleti · CSCI1951A

Some other project ideas

- Finding exoplanets using machine learning techniques
- Prediciting the duration of used car advertisements
- Predicting Customer Lifetime Value
- Recognizing the movement of a dog based on sensor data



Grading

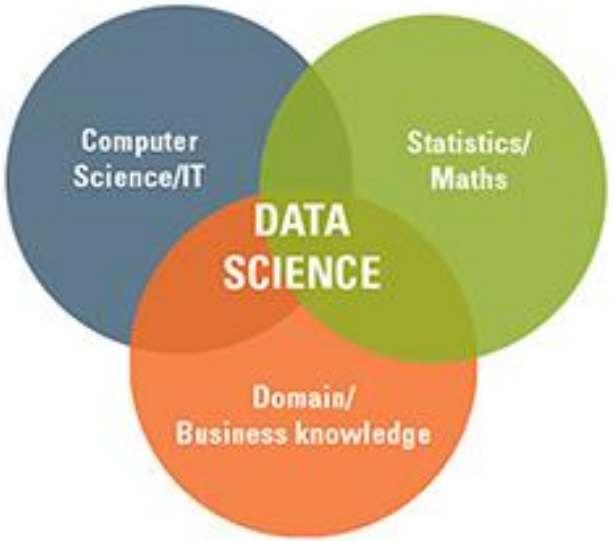
- MIDTERM (25%) + FINAL (25%) + HOMEWORK (25%) + PROJECT (25%)
- Cutoffs for letter grades:
  - 90% - A+
  - 85% - A
  - 80% - A-
  - 75% - B+
  - 70% - B
  - 65% - B-
  - 60% - C+
  - 55% - C
  - And so on…

Schedule of the semester

	Tuesday class	*Thursday*	*Friday class*	*Sunday*
W1 (02/03)				
W2 (02/10)			HW1 out	
W3 (02/17)				
W4 (02/24)		HW1 deadline	HW2 out	Forming teams
W5 (03/02)				
W6 (03/09)		HW2 deadline	HW3 out	Project plan
W7 (03/16)	MIDTERM			
SPRING BREAK	SPRING BREAK		SPRING BREAK	
W8 (03/30)		HW3 deadline	HW4 out	MILESTONE 1
W9 (04/06)			GOOD FRIDAY	
W10 (04/13)		HW4 deadline		
W11 (04/20)				
W12 (04/27)			LABOR DAY	MILESTONE 2
W13 (05/04)			FINAL	
W14 (05/11)			PROJECT presentations	

What is data science?

- New concept without a canonical definition
- Multidisciplinary nature (mathematics, statistics, computer science, etc.)
- Related concepts: data mining, statistics, machine learning, predictive analytics, big data
- **Purpose: finding novel, non-trivial, useful insights in big data**



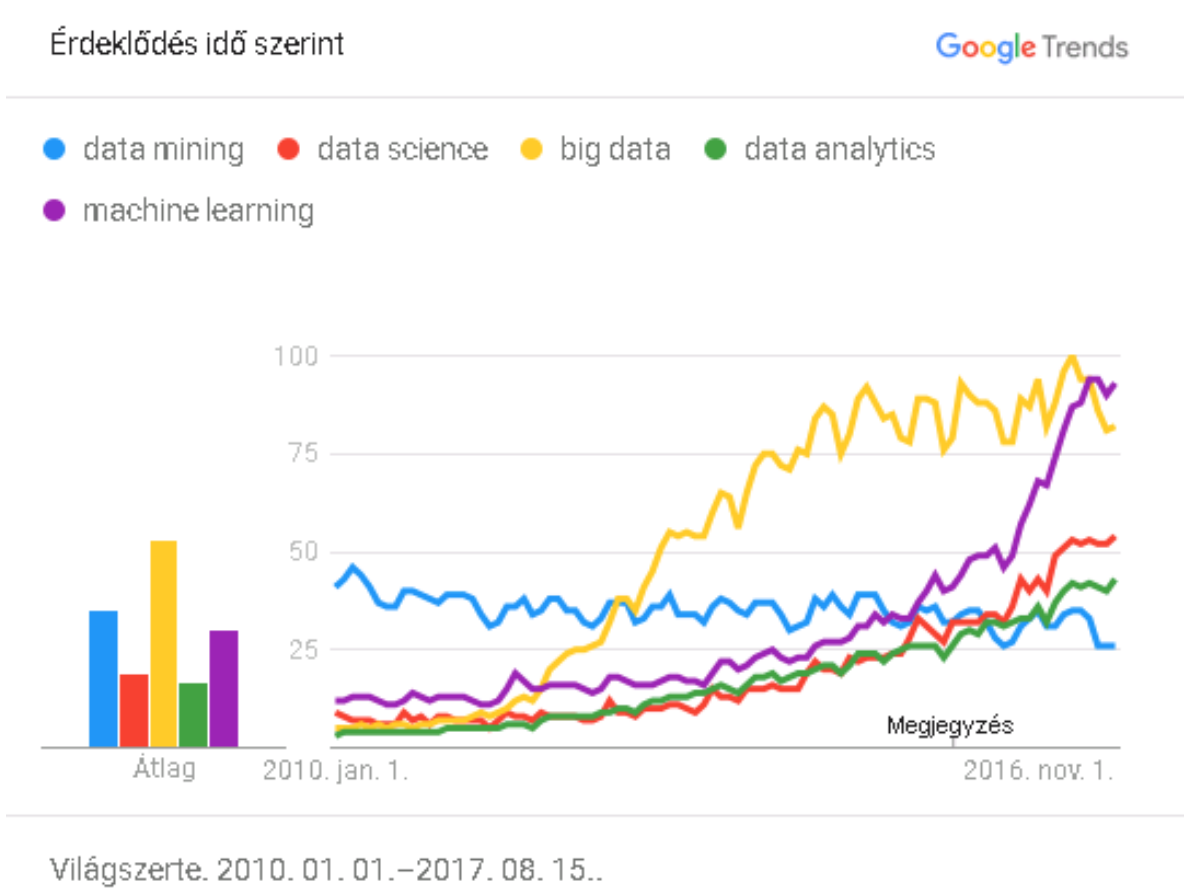
Other definitions

„Data Science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, data mining, and predictive analytics.”
Wikipedia

“Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management and preservation of large collections of information.”
Jeffrey Stanton, An Introduction to Data Science

Data science and related notions

- Data science
- Data mining
- Data analytics
- Data engineering
- Machine learning
- Statistics
- Knowledge discovery
- Business intelligence
- Predictive analytics
- Big Data

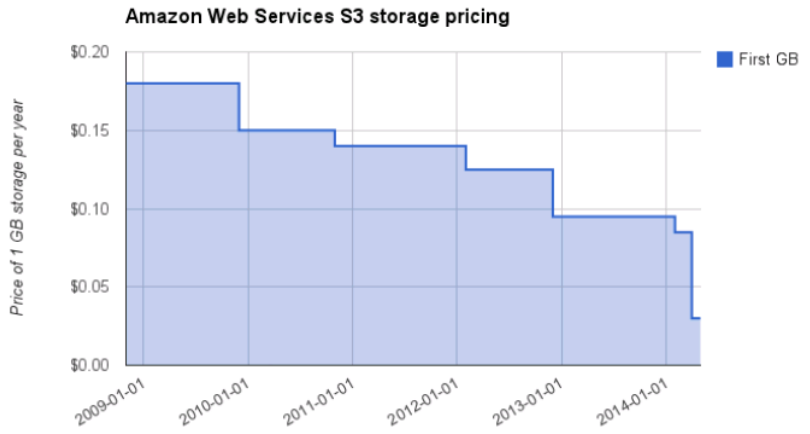


Machine learning

- Purpose: to develop and study algorithms that are able to learn from data and able to make predictions
- Artificial intelligence (AI): the agent is able to make decisions independently (autonomously), it is able to respond to its environment
  - Machine learning: learn the decision from big data

Big data

- „Big Data is when the size of the data itself becomes part of the problem”
  Mike Loukides, O’Reilly Media
- What size is considered to be „big” data?
  - It depends… When traditional data-processing methods are not able to hande them
- „Between the dawn of civilization and 2003, we only created five exabytes of information; now we’re creating that amount every two days.”
  Eric Schmidt, Google
- 1 EB = 1 billion GB

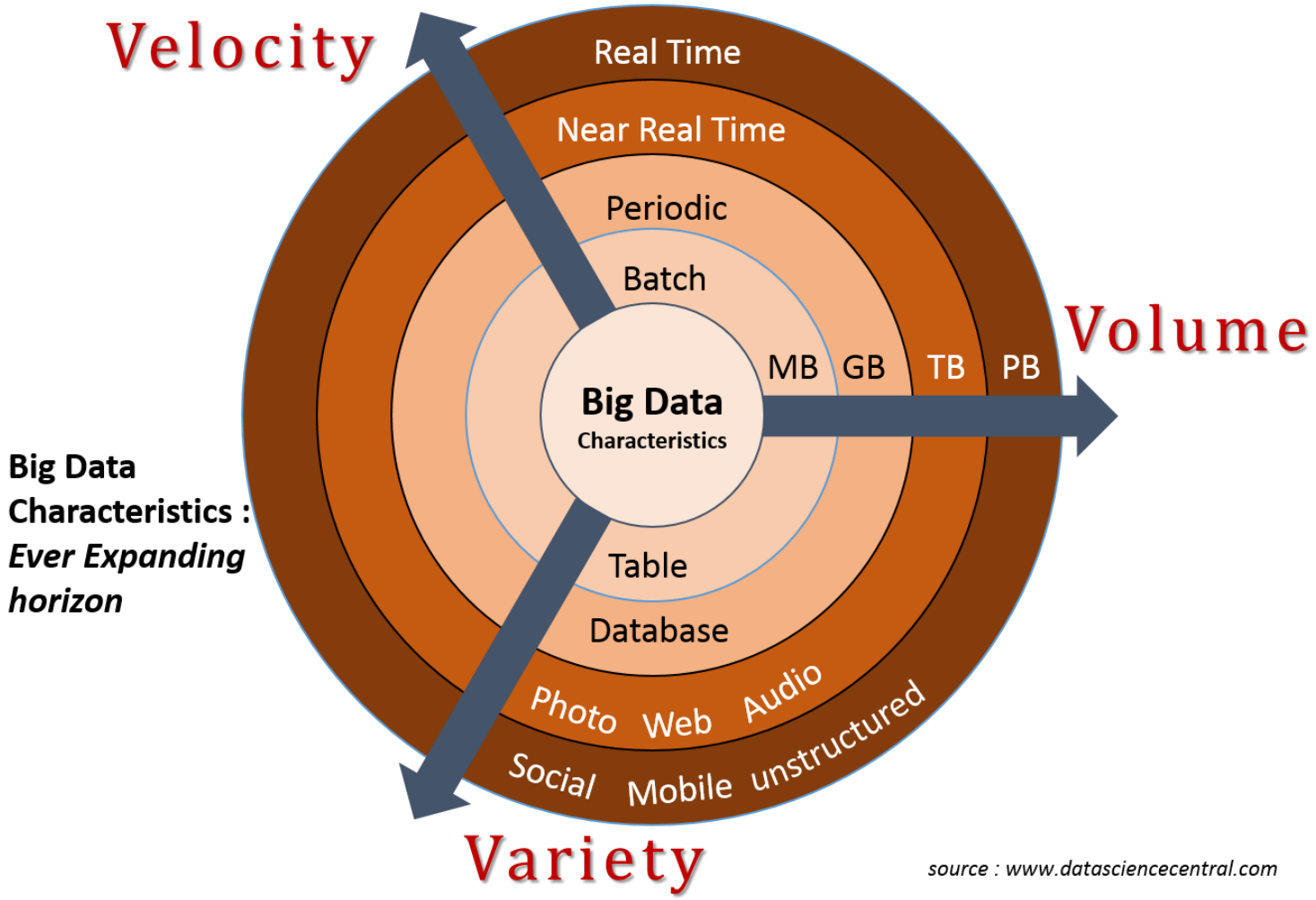


Solution: distributed systems, parallelization, Hadoop, MapReduce

➔ we probably won’t have time to cover that in this course

Big data – 3V

- 3Vs model of Doug Laney
- 4th V: Variability
  - Temporal inconsistency with peaks
- 5th V: Veracity
  - The quality of the data is changing, it worsens reliability



Big Data Characteristics : ***Ever Expanding horizon***

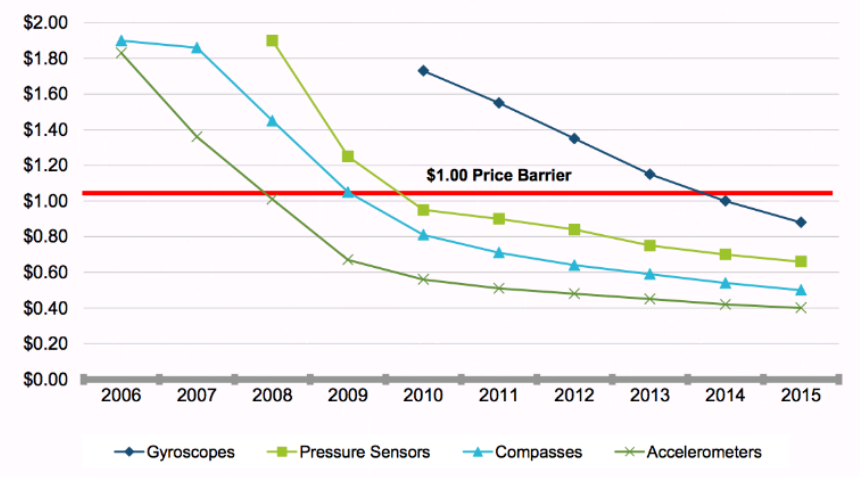
source : www.datasciencecentral.com

Historical overview

- John Tukey: The Future of Data Analysis, *Annals of Mathematical Statistics,* 1962
  - Before his time, he predicted the emergence of a new scientific discipline about data
- 80s, 90s: storage capacity increases rapidly + prices decrease ➔ data accumulation (data tomb)
  - Even exceeding Moore's law (the number of transistors in a dense integrated circuit doubles about every two years)– a similar observation is true for storage capacity

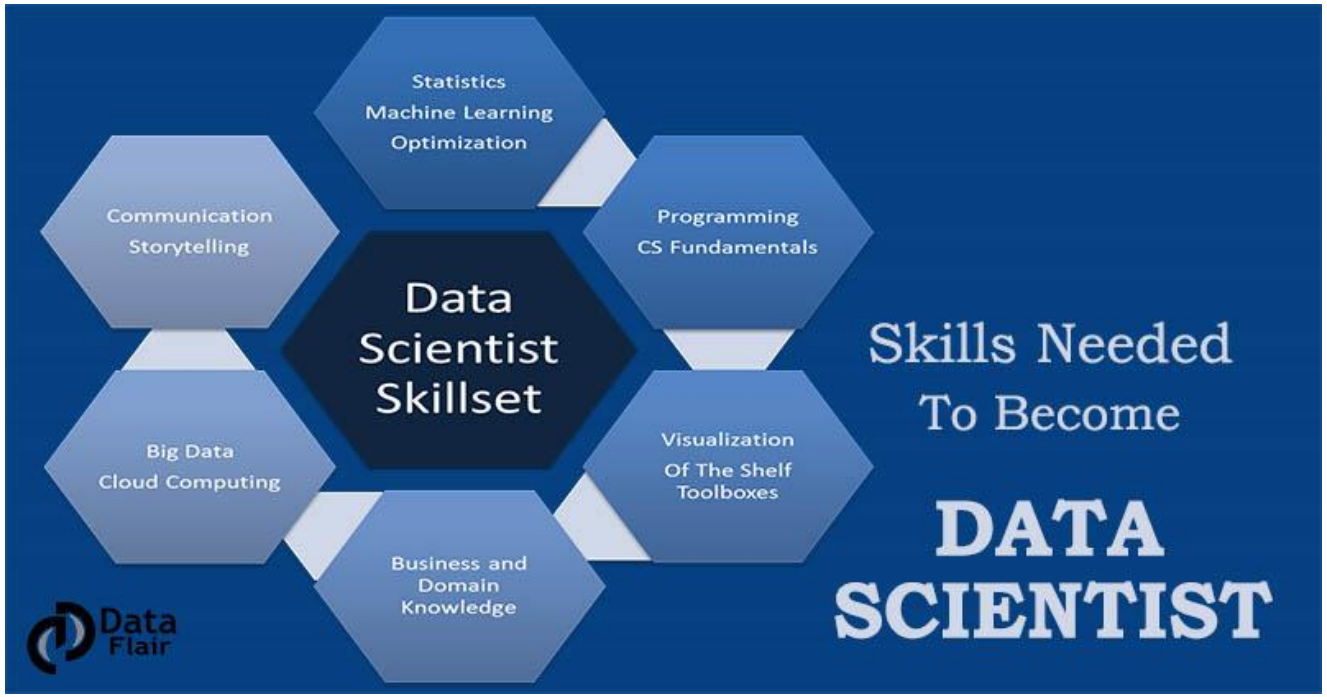
> „We are drowning in information,
> but starving for knowledge"
> *John Naisbitt, 1982*

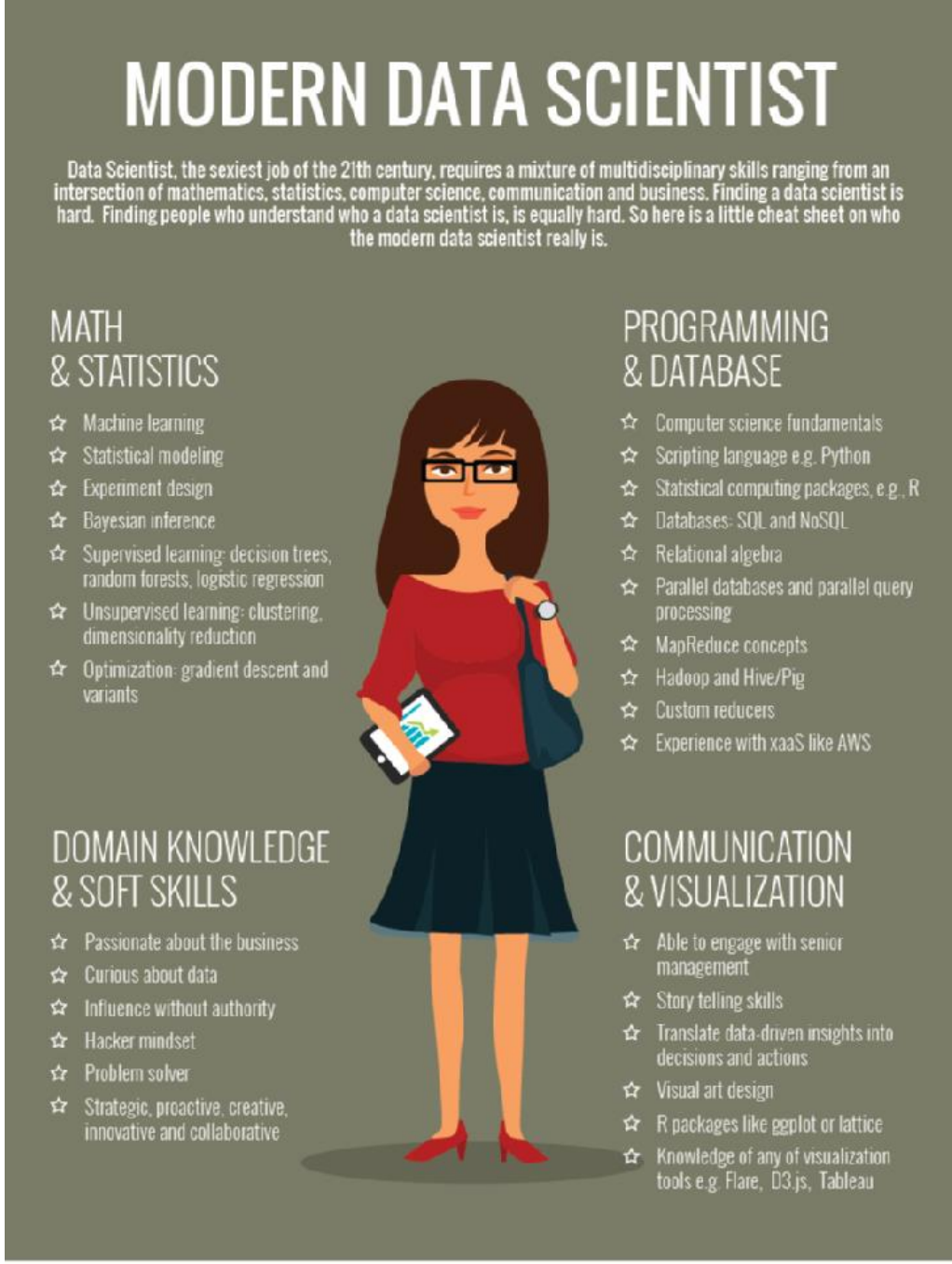
- New sophisticated methods were needed to retrieve information from large databases ➔ new algorithms
  - Initially heuristics (without proper theory)
  - In the new millennium it receives more research interest ➔ theoretical support



- Nowadays: the price of sensors are decreasing ➔ more data ➔ the challenge is continuous

What does a data scientist know?





Who is a data scientist?

„I think data scientist is a sexed-up term for a statistician.”
Nate Silver

„A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.”
Josh Willis

„A data scientist is a statistician who lives in San Francisco.”
„Data Science is statistics on a Mac”
Twitter

Outlook – Nate Silver

- American statistician, the founder and editor in chief of FiveThirtyEight
- He accurately predicted the result of 49 states at 2008 presidential election and got all the 50 states right in 2012

#natesilverfacts

„When Alexander Bell invented the telephone he had 3 missed calls from Nate Silver."

„Nate Silver can hear sign language."

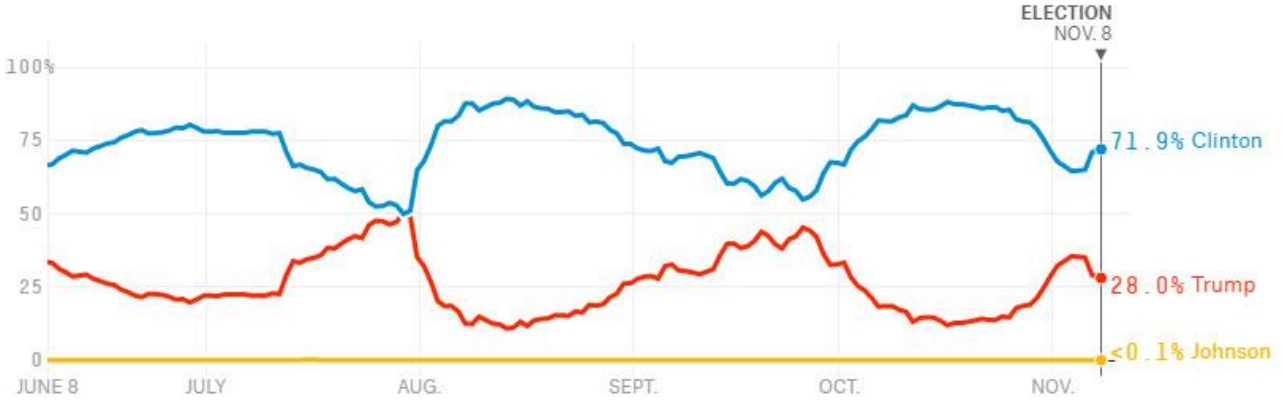
„For Nate Silver, asymptotic theory kicks in at N=1."

„Fearing Nate Silver, the Null Hypothesis rejected itself."

„Nate Silver's model fit the test data even better than the training data."

Outlook - USA elections, 2016

- 2016: Nate Silver claims that Clinton has much more chance to win (he gives 72% chance to this scenario)
  - Other statisticians give even less chance for Trump
  - Trump won the election
- Big data also played a big role in the presidential campaign



Donald Trump's campaign shifted odds by making big data personal

Social media surveys helped to target thousands of individuals in swing states

Gillian Tett

CAMBRIDGE ANALYTICA

Facebook fined £500,000 over Cambridge Analytica scandal

Oct 25, 2018

UK data watchdog says social media giant failed to safeguard its users' personal information

Justin Sullivan/Getty Images
Mark Zuckerberg has been given until the end of the month to respond

Facebook has been fined £500,000 by the UK's data watchdog for allowing political consulting firm Cambridge Analytica to harvest the information of millions of people without their consent.

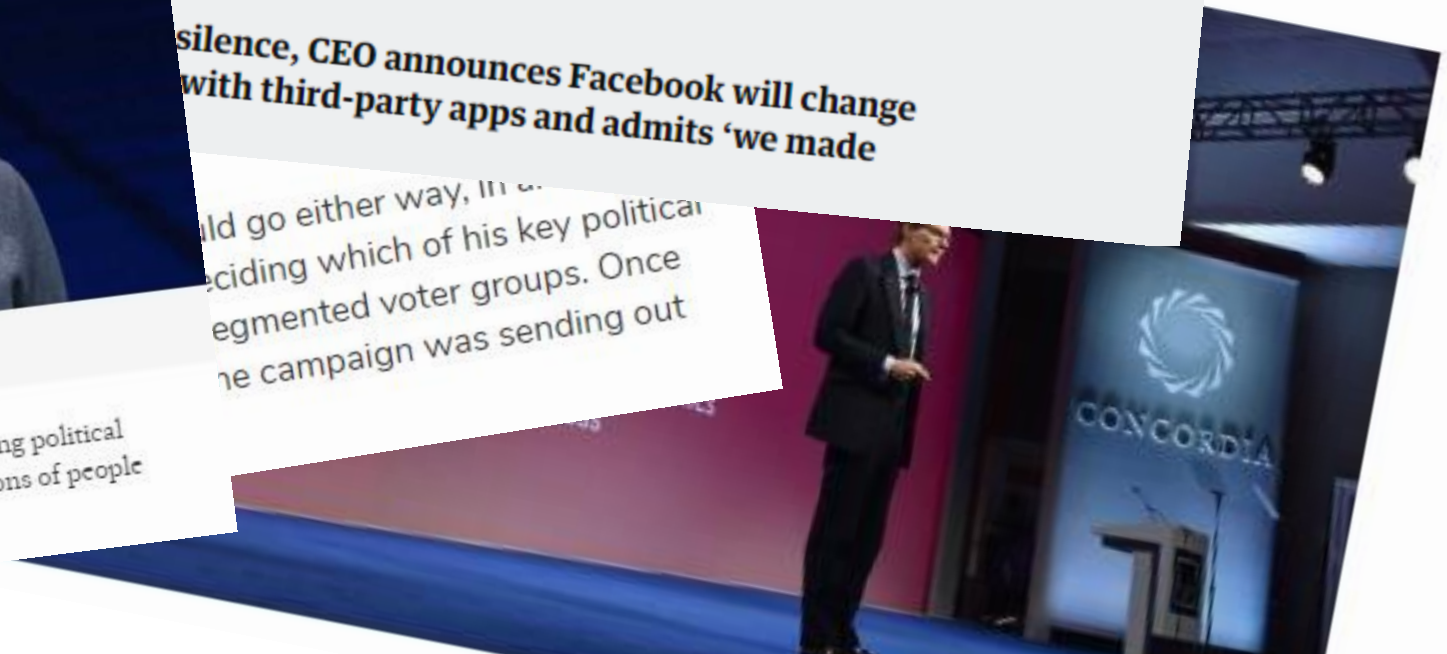
Data mining firm behind Trump election built psyc… Am… …ical profiles of nearly every

…uckerberg apologises for …k's 'mistakes' over Cambridge …a

…silence, CEO announces Facebook will change … with third-party apps and admits 'we made

…ld go either way, in a… …eciding which of his key political … egmented voter groups. Once … campaign was sending out

…rofiles of nearly



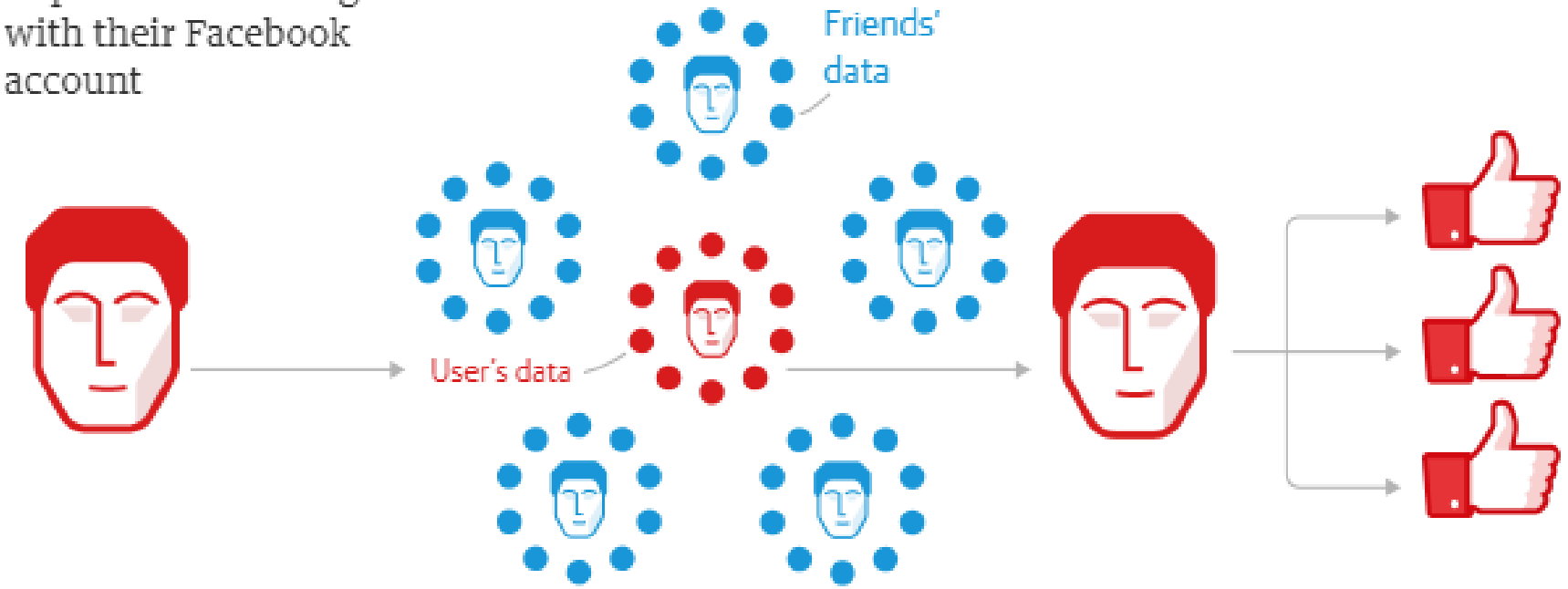
Cambridge Analytica: how 50m Facebook records were hijacked

1
Approx. 320,000 US voters ('seeders') were paid $2-5 to take a detailed personality/ political test that required them to log in with their Facebook account

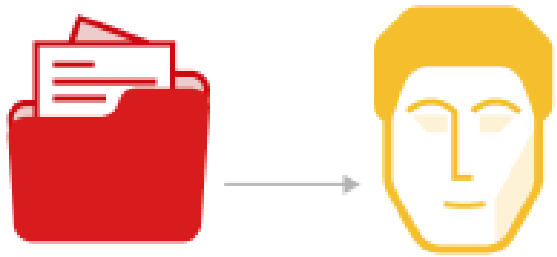
2
The app also collected data such as likes and personal information from the test-taker's Facebook account ...

3
The personality quiz results were paired with their Facebook data - such as likes - to seek out psychological patterns

4
Algorithms combined the data with other sources such as voter records to create a superior set of records (initially 2m people in 11 key states*), with hundreds of data points per person



... as well their friends' data, amounting to over 50m people's raw Facebook data



These individuals could then be targeted with highly personalised advertising based on their personality data

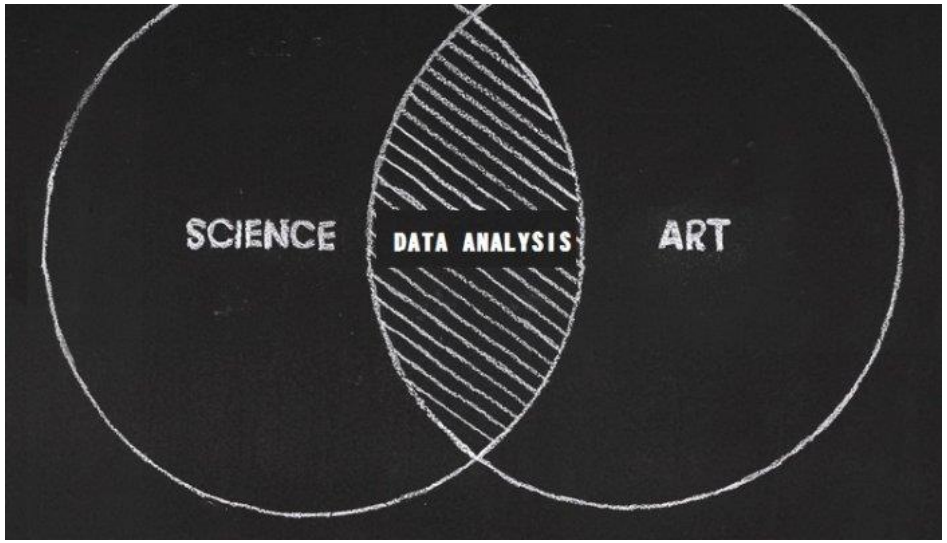
Guardian graphic. *Arkansas, Colorado, Florida, Iowa, Louisiana, Nevada, New Hampshire, North Carolina, Oregon, South Carolina, West Virginia

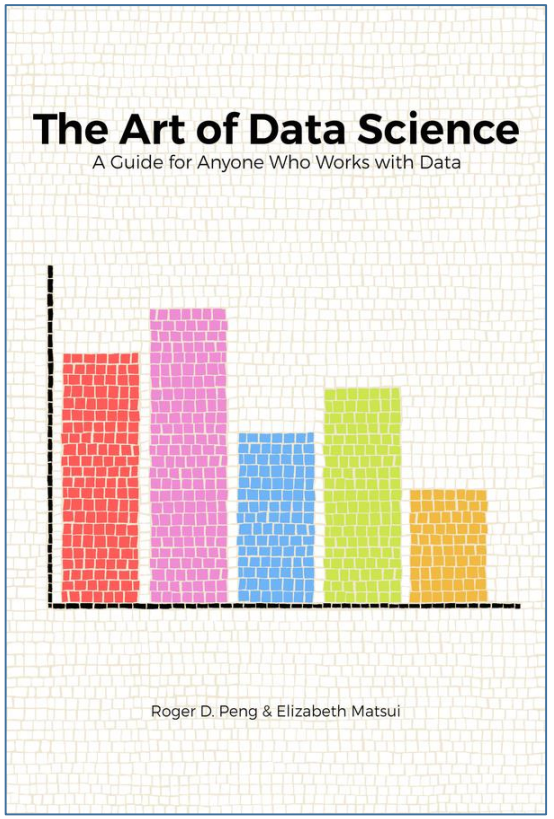
Statistics vs. Data science

- Different aspects/approaches: testing hypothesis (statistical tests) vs. Finding hypothesis (more general question)
- Studying DNA sequences
  - Statistician: Is there a significant connection between a certain DNA subsequent and a certain disease?
  - Data scientist: What are the connections between certain diseases and certain DNA subsequents?
- Studying smoking habits
  - Statistician: Is there a significant difference regarding smoking ratios between males and females?
  - Data scientist: What are the typical groups regarding smoking habits?

Is data science an art?

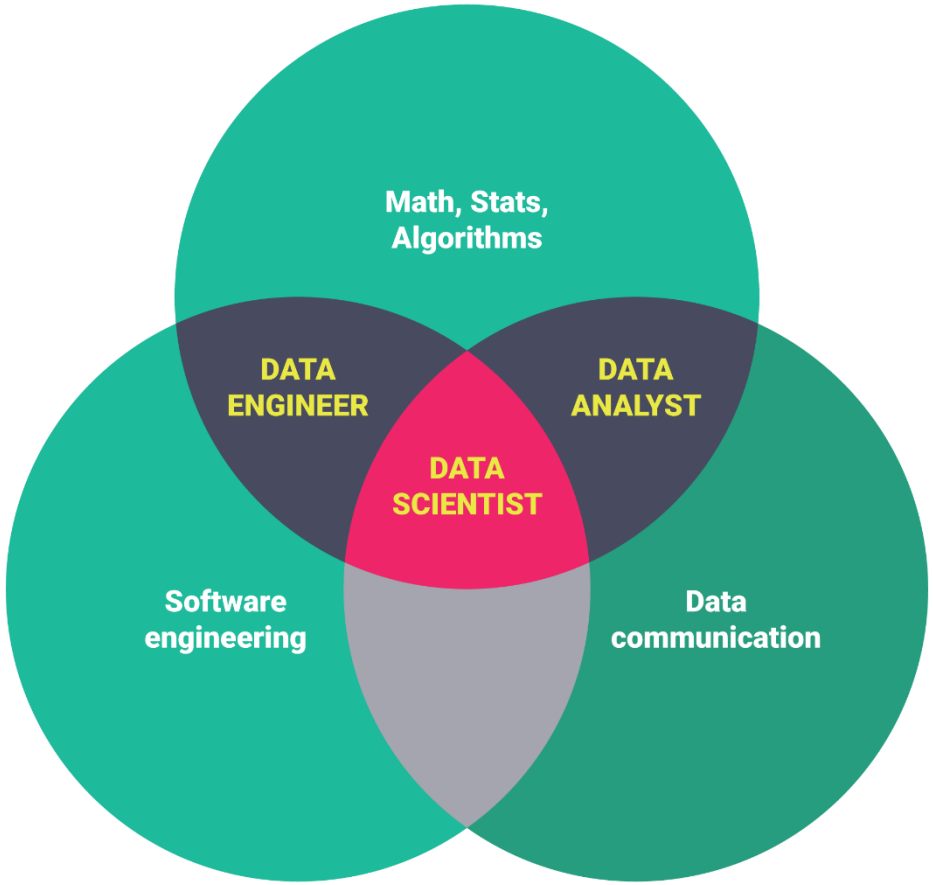
„Science is what we understand well enough to explain to a computer. Art is everything else we do.”
Donald Knuth





Roles in data science

- There are no strict roles, it depends on the company, on the project
- Job listings are also ambiguous
  - Same positions may cover totally different job descriptions

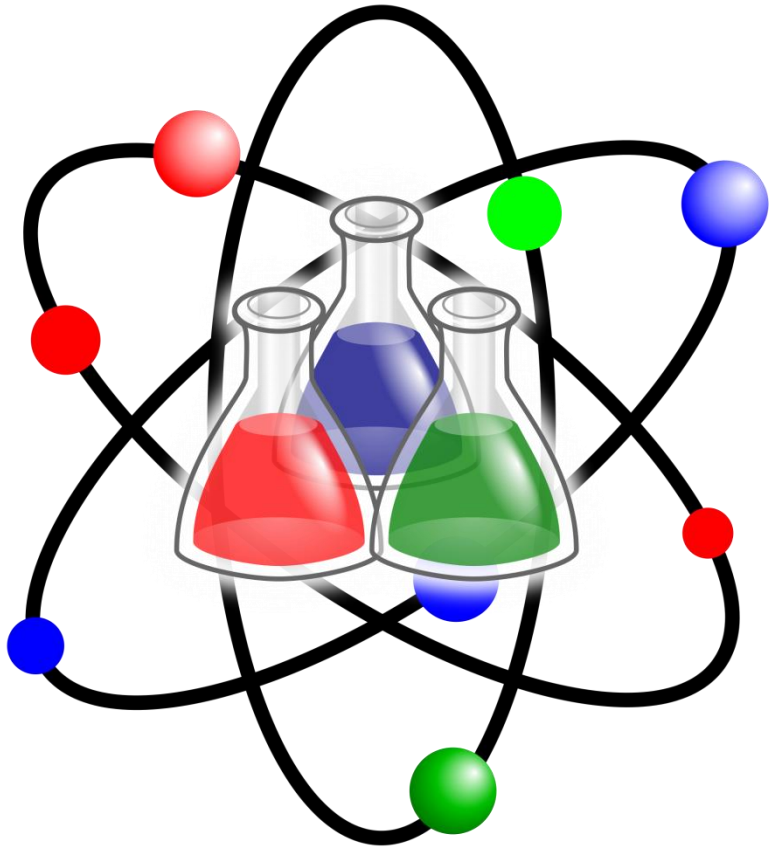


Some application areas– business world

- Telecommunication (optimal pricing, churn detection)
  - Customer history, phone logs
- Retail (up-selling, cross-selling, improving customer satisfaction)
  - Credit card transactions, data from online purchase
- Banks (credit assessment, fraud detection)
  - Customer history, credit card transactions
- Several other domains (stock exchange, social media, websites)

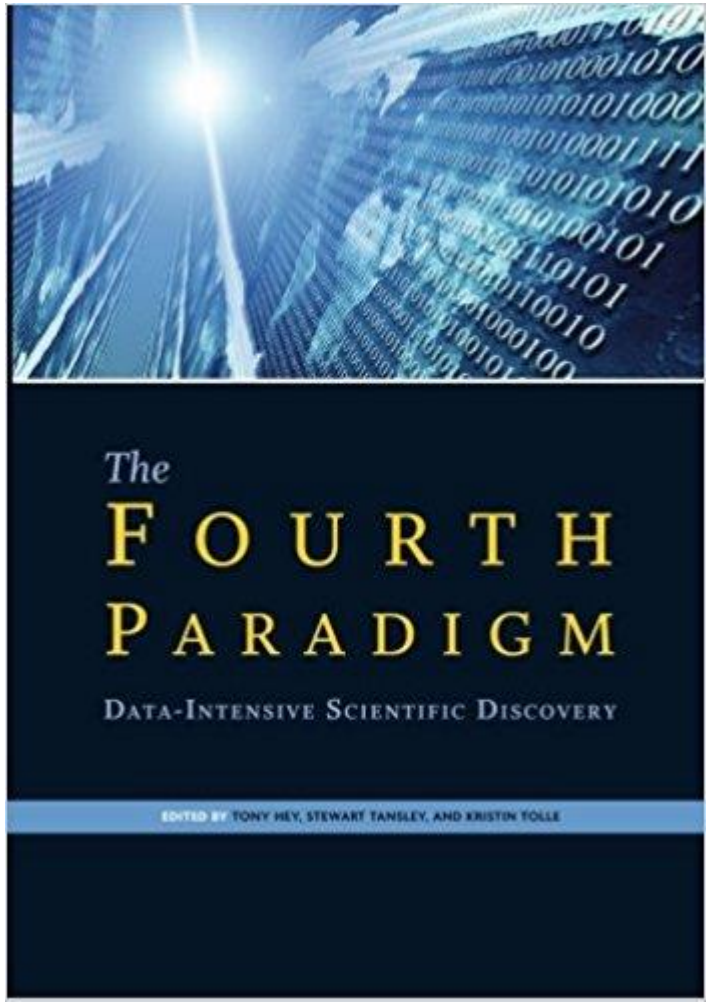
Some application areas– science

- Particle physics
  - Finding new phenomena, validating theories
- Astronomy
  - Analyzing data space telescopes
  - Classifying photos automatically (without a human)
- Drug development
  - Finding drug substance, take out experiments
- Medicine
  - Supporting diagnostic
  - Monitoring systems
- Several other areas (brain research, gene map)



Scientific paradigm shift?

- A thousand years ago: empirical science
  - Describing nature
- Last few hundred years: theoretical approach
  - Introducing models, generalizations
- Last few decades: computational approach
  - Simulation of complex phenomena
- Nowadays: data-driven approach
  - Merging experiments, theory and simulations

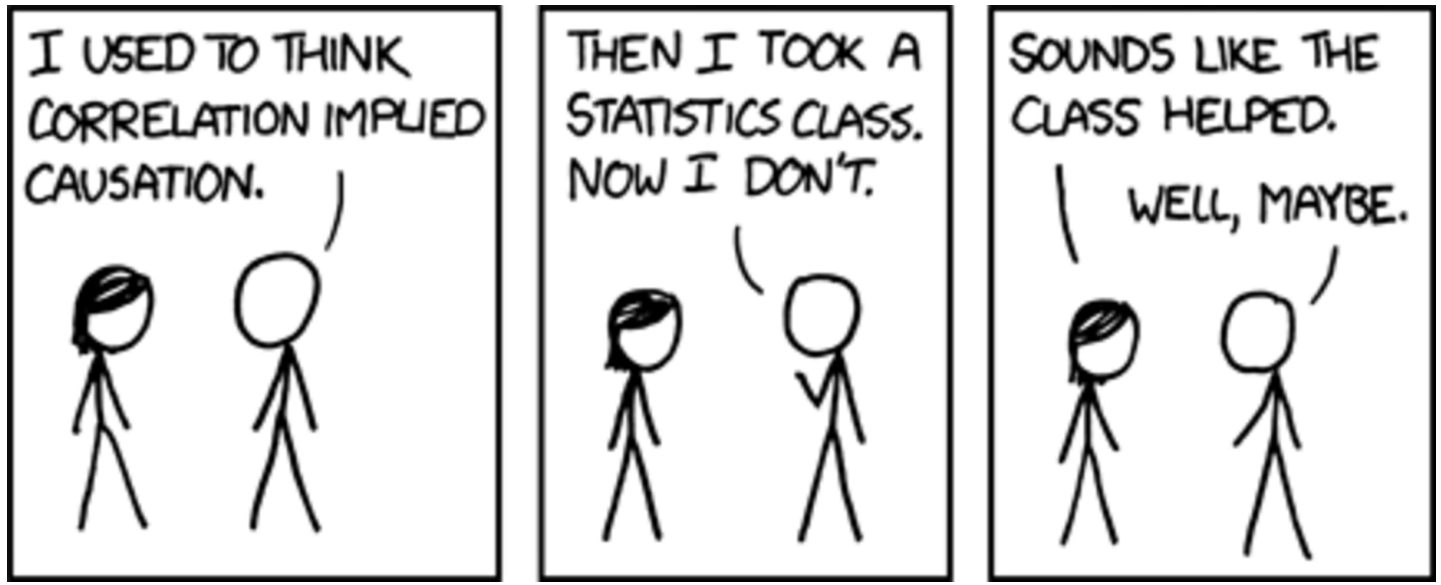


What is not data science?

- Processing and analyzing data are not always considered to be data science
  - Descriptive analysis (e.g. a summary of a population census) is not data science by itself
- Data science looks for patterns, correlations in data BUT correlation DOES NOT imply causation
- Exploring cause and effect relationship is usually out of scope of data science
  - Need for randomized groups
  - Using control groups in verified environments
  - Econometrics – finding causal relations in economic data

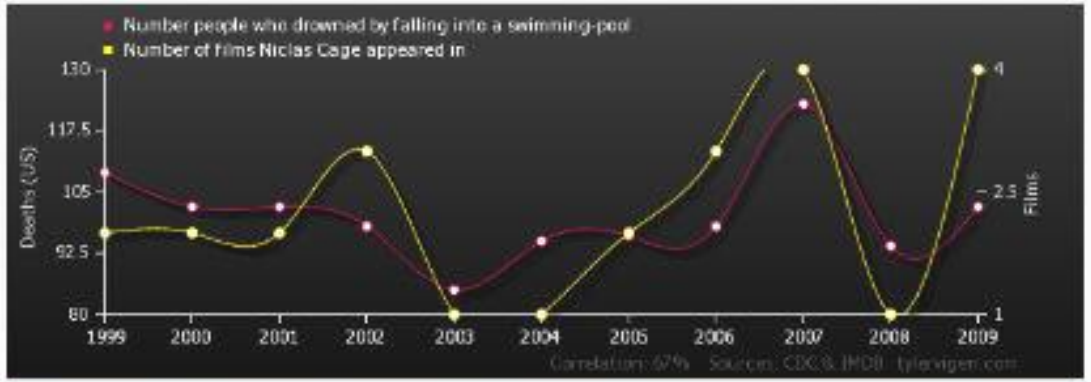
Correlation ≠ causation

- Strong correlation:
  - Height and hair length
  - Ice cream sales and number of drownings



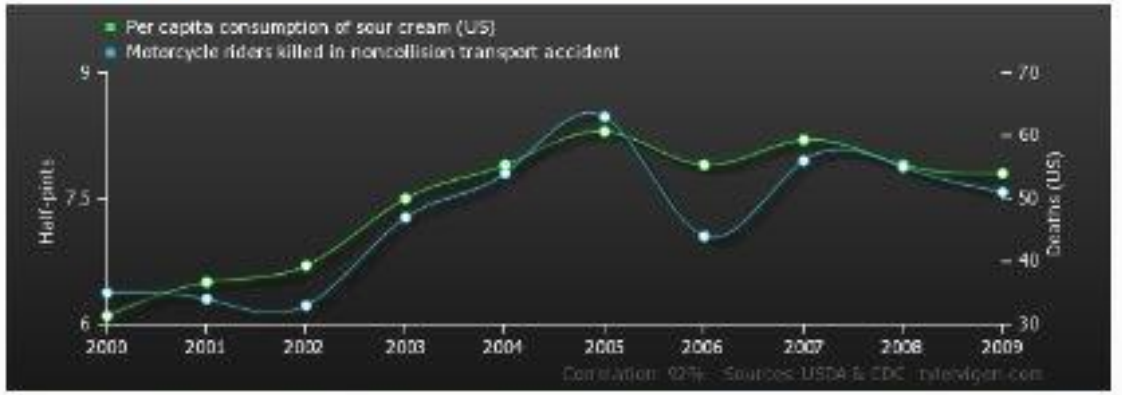
- Be careful! From data one can retrieve connections that is just there due to chance
  - You can't generalize them!
  - See next slides!

Number people who drowned by falling into a swimming-pool
correlates with
Number of films Nicolas Cage appeared in



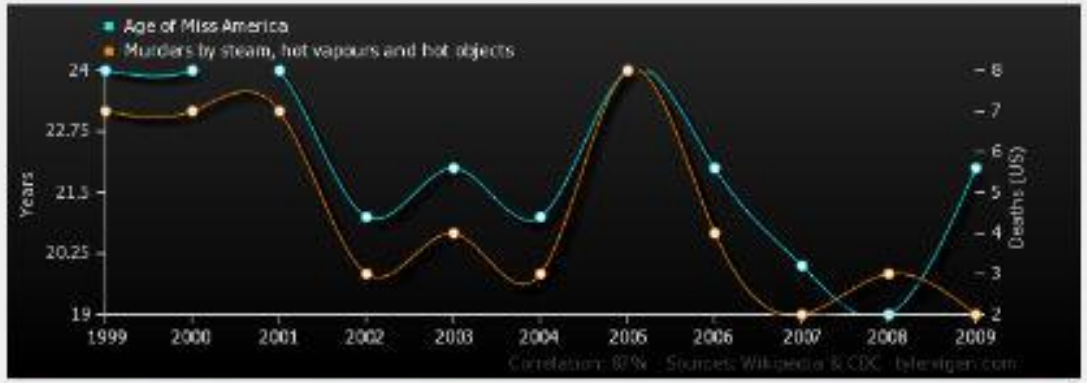
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number people who drowned by falling into a swimming-pool	109	102	102	98	85	95	96	98	123	94	102

Per capita consumption of sour cream (US)
correlates with
Motorcycle riders killed in noncollision transport accident



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009

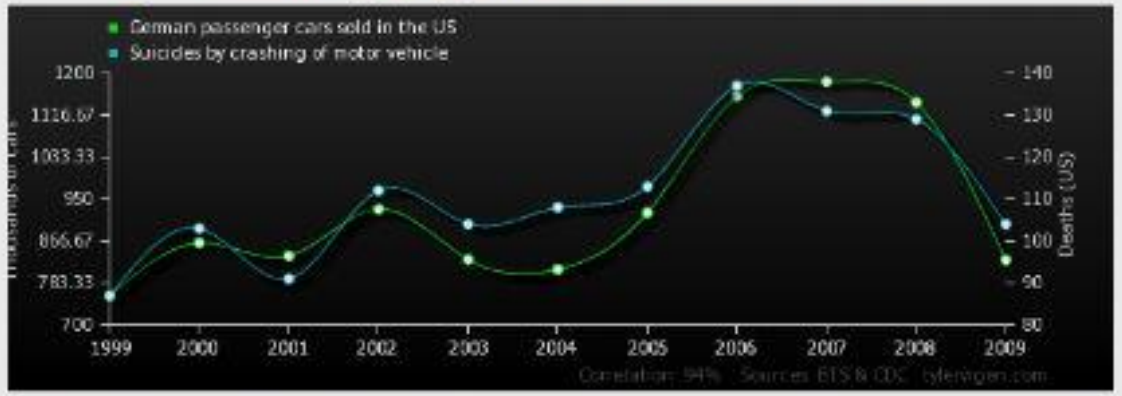
Age of Miss America
correlates with
Murders by steam, hot vapours and hot objects



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Age of Miss America Years (Wikipedia)	24	24	24	21	22	21	24	22	20	19	22
Murders by steam, hot vapours and hot objects Deaths (US) (CDC)	7	7	7	3	4	3	8	4	2	3	2

Correlation: 0.870127

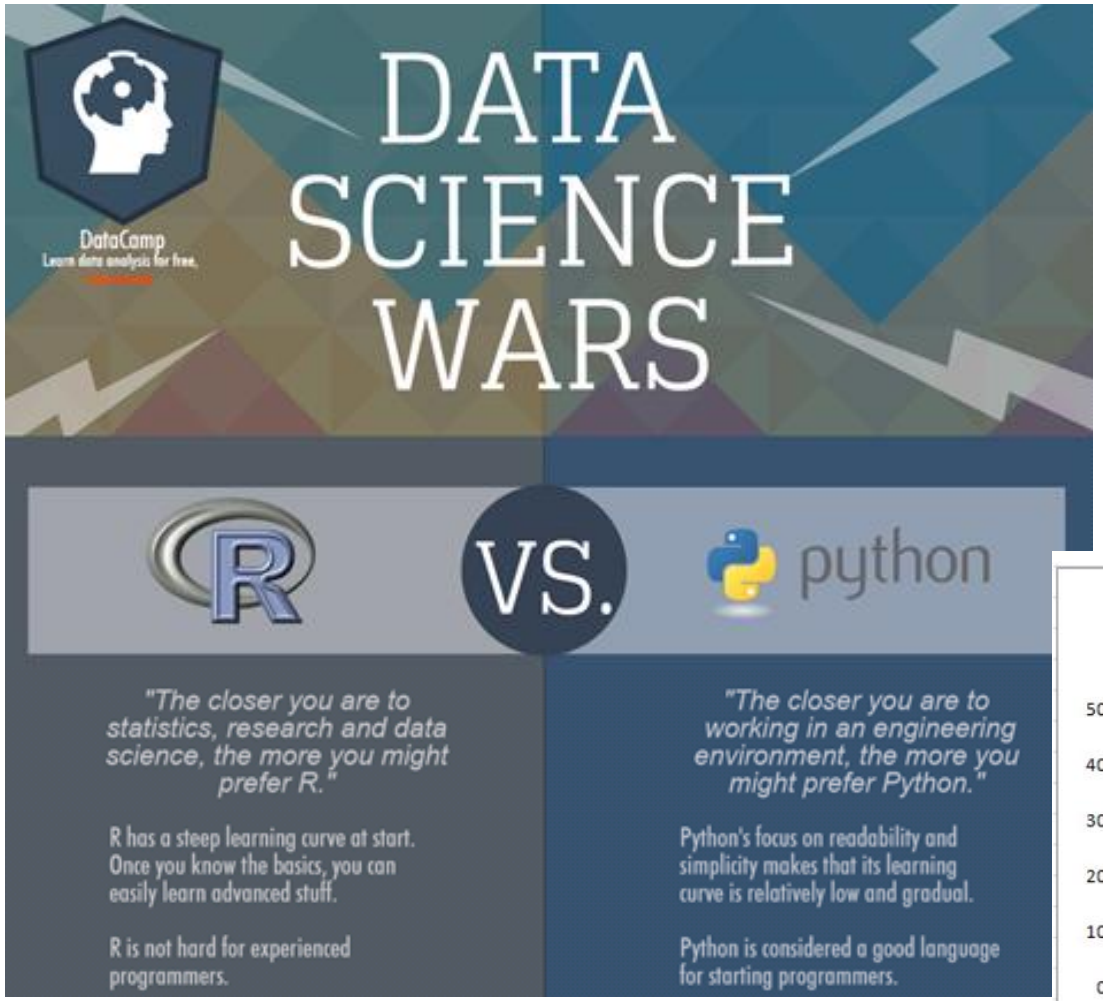
German passenger cars sold in the US
correlates with
Suicides by crashing of motor vehicle

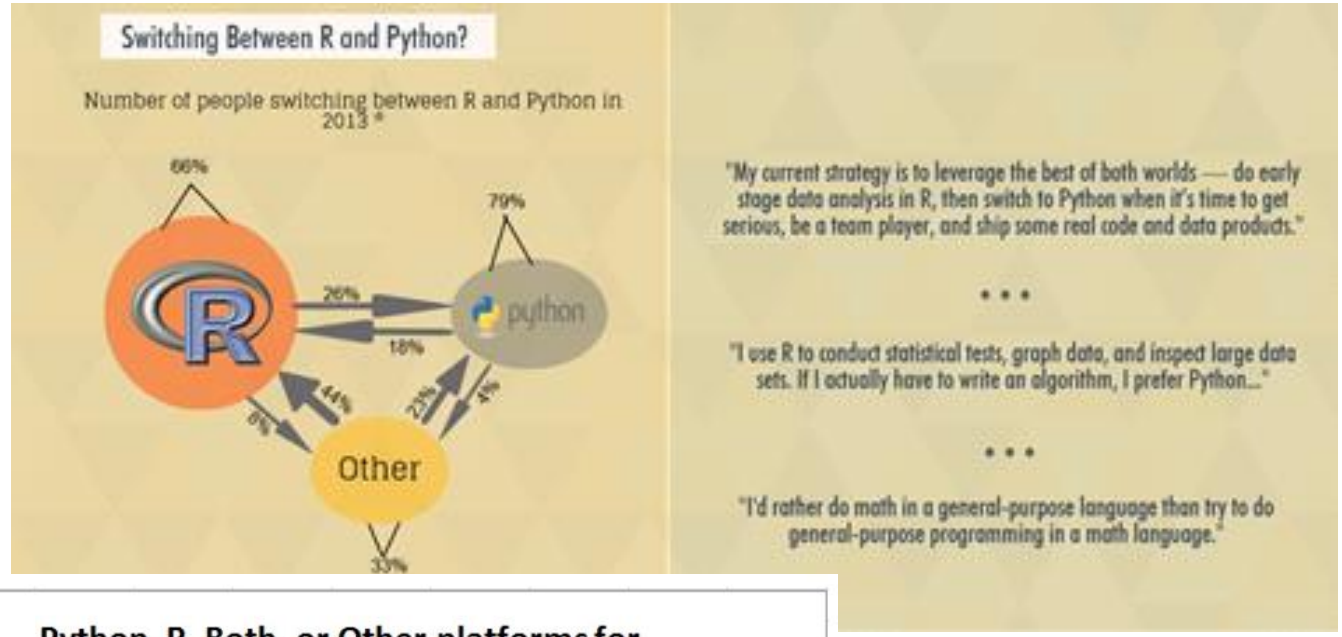


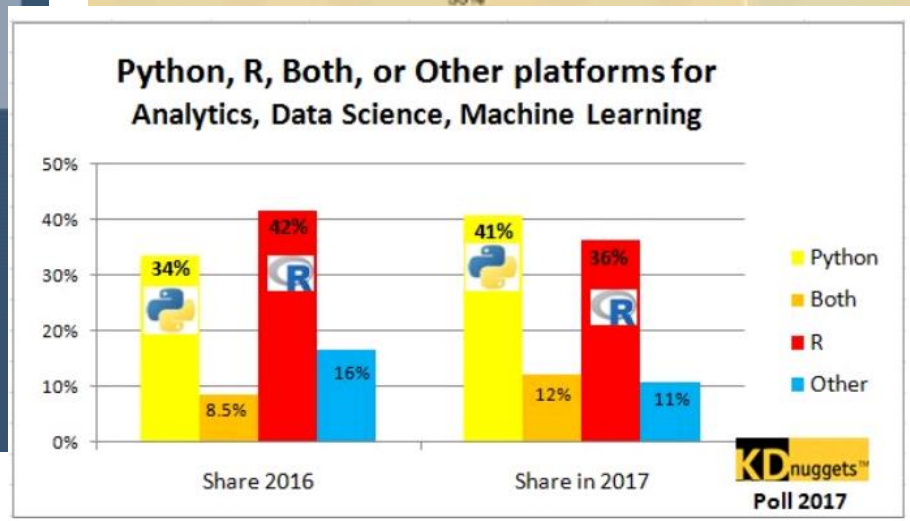
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
German passenger cars sold in the US Thousands of cars (BTS)	758	863	837	930	830	810	923	1,154	1,183	1,142	829
Suicides by crashing of motor vehicle Deaths (US) (CDC)	87	103	91	112	104	108	113	137	131	129	104

Correlation: 0.935701

Python vs. R







Some other interesting comparism: https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis#gs.fC_rDHo

Refreshing Python

- I really encourage everybody to refresh his/her knowledge in Python
  - A good tutorial
    - https://www.tutorialspoint.com/python/python_quick_guide.htm
      - Until „Directories in Python”
      - From „Creating Classes” to „Bulit-In Class Attributes”
  - If you need more than a quick refresh I recommend the following short, free interactive online courses:
    - https://www.datacamp.com/courses/intro-to-python-for-data-science/
    - https://www.codeschool.com/courses/try-python
  - Some other useful materials are uploaded in Moodle

Python preparation

- We will use Jupyter IPython with Anaconda distribution
  - Runs in a web browser
  - Interactive
  - Simple
  - Scenic

- Download it by following this link: https://www.anaconda.com/download/
- More useful links:
  - https://ipython.org/install.html
  - https://www.continuum.io/downloads#windows

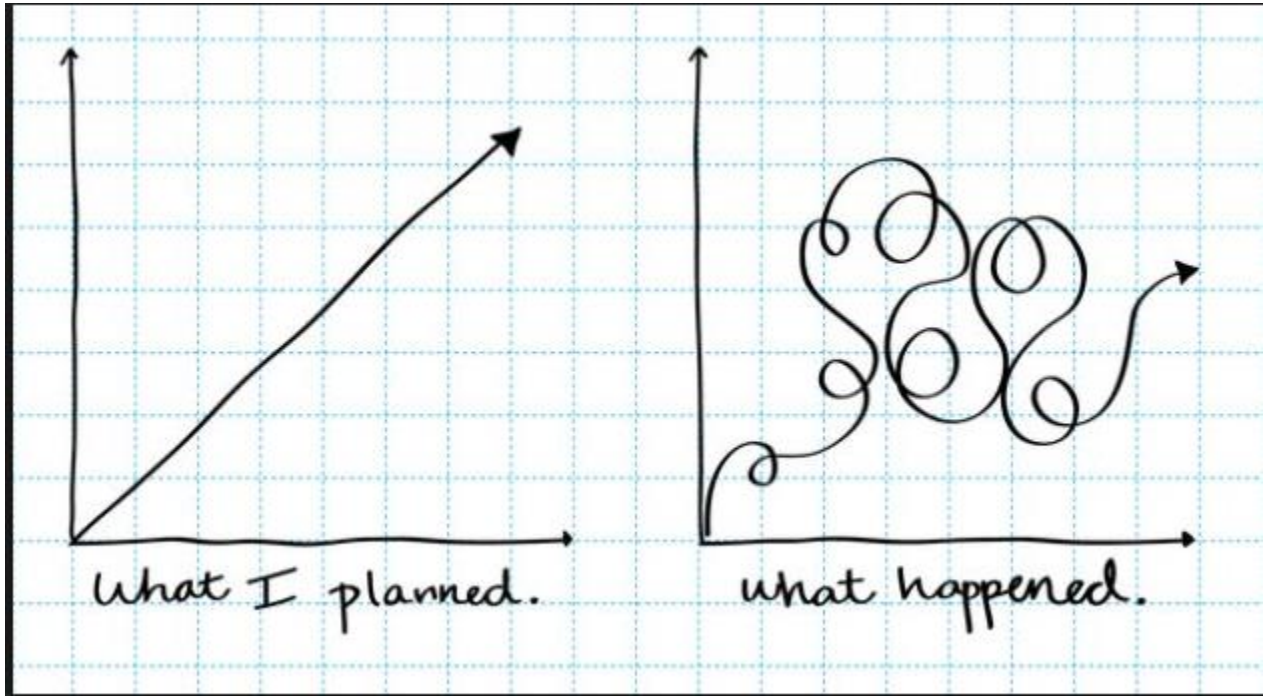
Using Ipython notebooks

- From Anaconda Navigator you can launch Jupyter Notebook or JupyterLab
  - The notebook will start in your default web browser

- More information: https://jupyter.readthedocs.io/en/latest/running.html#running

What are we going to learn about?

- Data types, data processing
- Classification (k NN, decision tree, naive Bayes, logistic regression, SVM, neural networks)
- Hybrid classification (bagging, boosting, ensemble)
- Regression (linear, polynomial)
- Evaluating models
- Clustering (k-means, hierarchical, density based)
- Recommender systems
- Networks, PageRank algorithm
- Data visualization
- Case studies



Acknowledgement

- András Benczúr, Róbert Pálovics, SZTAKI-AIT, DM1-2
- Krisztián Buza, MTA-BME, VISZJV68
- Bálint Daróczy, SZTAKI-BME, VISZAMA01
- Judit Csima, BME, VISZM185
- Gábor Horváth, Péter Antal, BME, VIMMD294, VIMIA313
- Lukács András, ELTE, MM1C1AB6E
- Tim Kraska, Brown University, CS195
- Dan Potter, Carsten Binnig, Eli Upfal, Brown University, CS1951A
- Erik Sudderth, Brown University, CS142
- Joe Blitzstein, Hanspeter Pfister, Verena Kaynig-Fittkau, Harvard University, CS109
- Rajan Patel, Stanford University, STAT202
- Andrew Ng, John Duchi, Stanford University, CS229