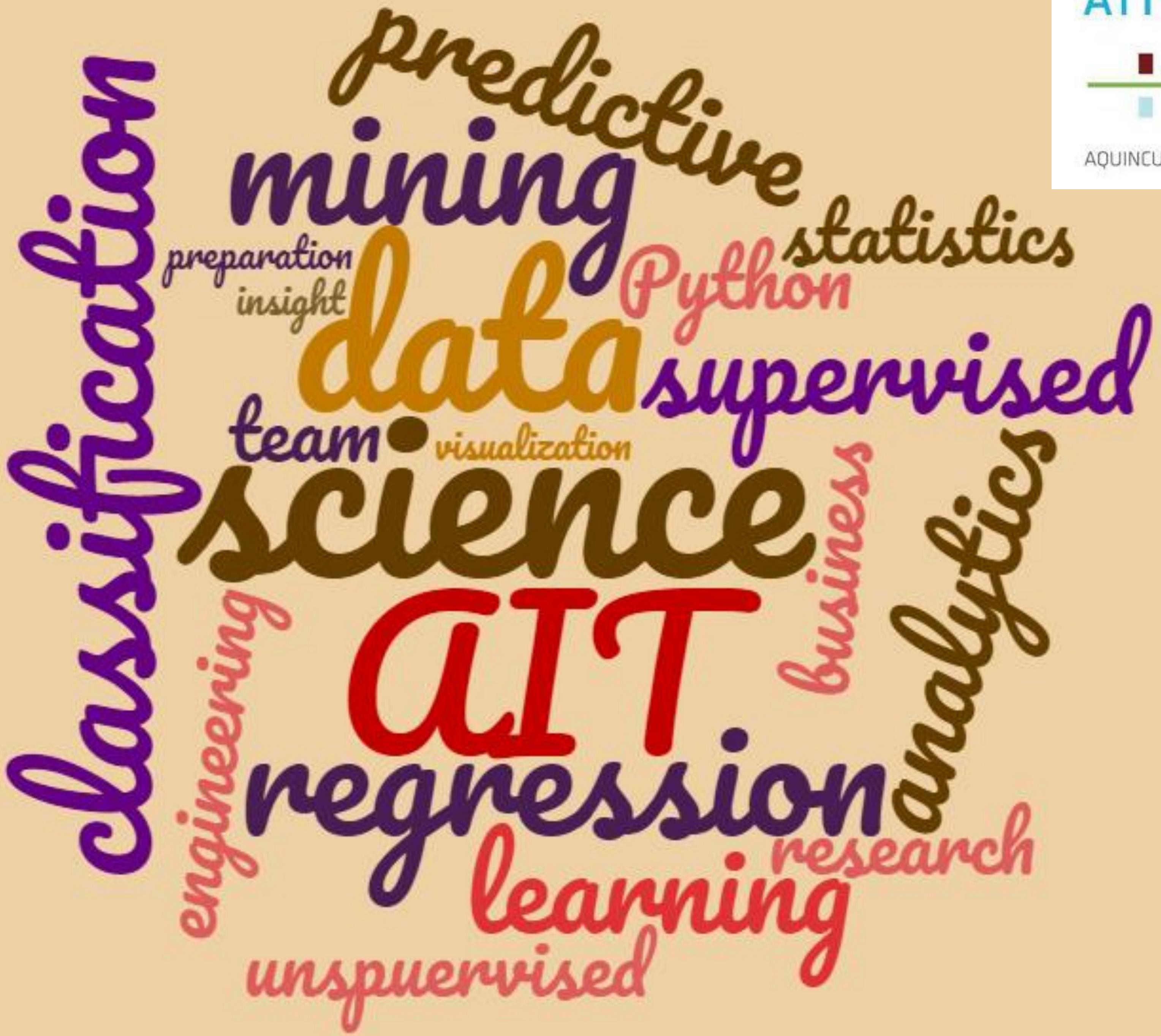
Data Science

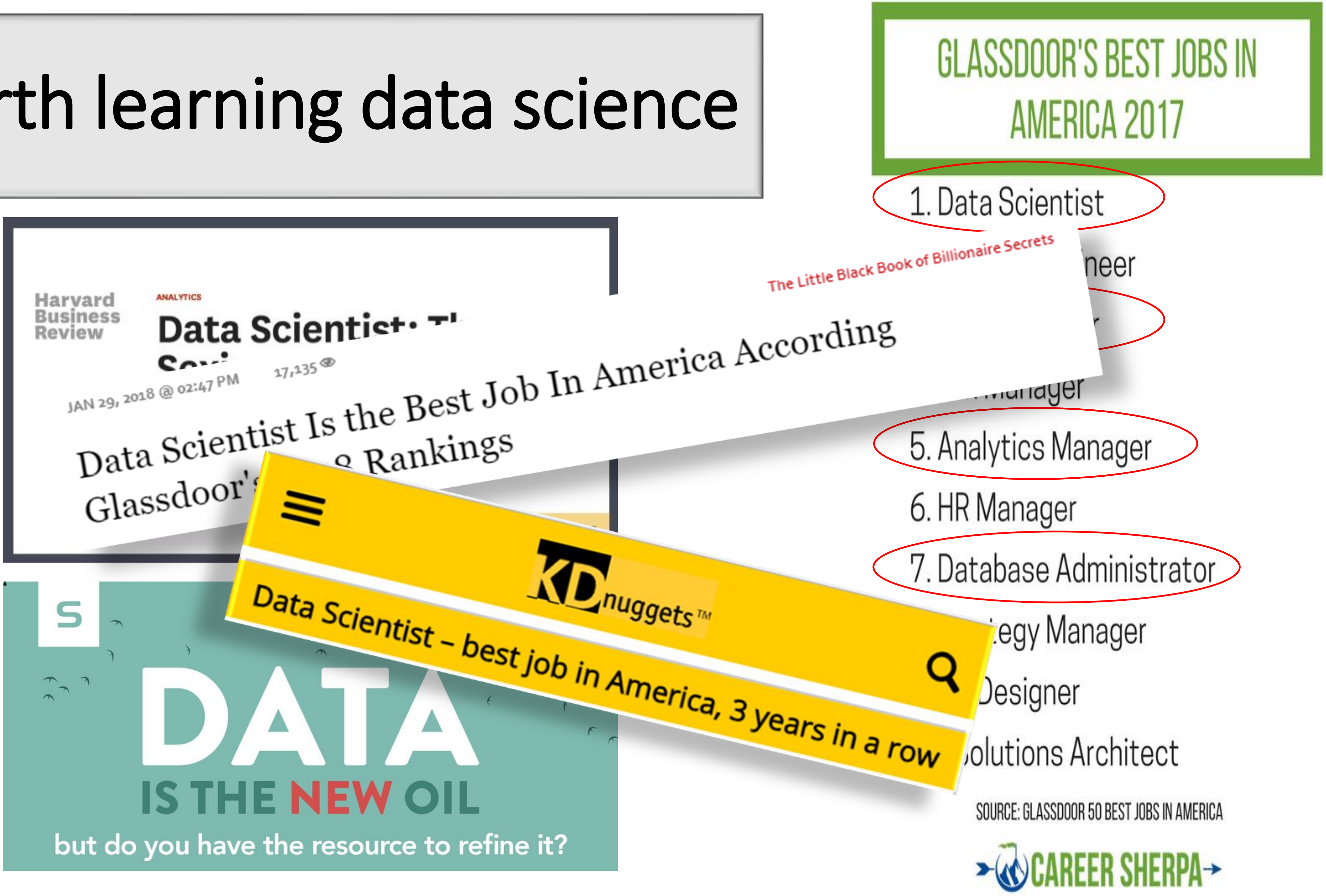
AIT-BUDAPEST
AQUINCUM INSTITUTE OF TECHNOLOGY



February 7, 2023.
Introduction

Dr. Roland Molontay

Worth learning data science



50 Best Jobs in America for 2022

Best Places to Work | Top CEOs | **Best Jobs** | Best Cities for Jobs | Highest Paying Jobs | Share

2022 | United States

	Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1	Enterprise Architect	$144,997	4.1/5	14,021	View Jobs
#2	Full Stack Engineer	$101,794	4.3/5	11,252	View Jobs
#3	Data Scientist	$120,000	4.1/5	10,071	View Jobs
#4	Devops Engineer	$120,095	4.2/5	8,548	View Jobs
#5	Strategy Manager	$140,000	4.2/5	6,977	View Jobs
#6	Machine Learning Engineer	$130,489	4.3/5	6,801	View Jobs
#7	Data Engineer	$113,960	4.0/5	11,821	View Jobs
#8	Software Engineer	$116,638	3.9/5	64,155	View Jobs
#9	Java Developer	$107,099	4.1/5	10,201	View Jobs
#10	Product Manager	$125,317	4.0/5	17,725	View Jobs

Forbes ADVISOR

Advisor › Education

Advertiser Disclosure

What Are The Fastest-Growing Jobs Of 2023?

By Cecilia Seiter
Contributor

Reviewed By Veronica Freeman
Editor

Published: Feb 1, 2023, 10:30am

Forbes

Forbes ADVISOR

Fastest-Growing Tech Careers

Data Scientists

Growth Rate (2021-31): +36%
Median Pay: $100,910 per year
Education Requirements: Bachelor's degree
Career Overview: Data scientists extract insights and knowledge from large, complex data sets. They leverage that data to make intelligent, informed decisions to help organizations improve their performance and achieve their goals.

Conducting surveys or scraping the web to collect data is a key component of a data scientist's job. From there, data scientists clean and classify raw data, using machine learning and data visualization software to demonstrate their findings. It's paramount that data scientists know how to communicate their findings effectively and in a way that's accessible to a general audience.

Information Security Analysts

Growth Rate (2021-31): +35%
Median Pay: $102,600 per year
Education Requirements: Bachelor's degree in cybersecurity or a related field
Career Overview: Information security analysts are responsible for ensuring the safety and security of an organization's sensitive information and computer systems. They rigorously monitor networks for security breaches and investigate any attacks that may occur.

Information security analysts use software like firewalls and data encryption programs to safeguard sensitive assets. They are also responsible for documenting metrics and reporting attempted attacks. Information security analysts recommend security enhancements to management or senior IT staff, and they help other employees gain their footing with new security products and procedures.

Cybersecurity analysts are a type of information security analyst. For more information, check out our guides on information security vs. cybersecurity and how to become a cybersecurity analyst.

Web Developers

Growth Rate (2021-31): +30%
Median Pay: $77,030 per year
Education Requirements: Bachelor's degree
Career Overview: What is web development? Web development is a multidisciplinary field that involves a combination of technical and creative skills.

Web developers build websites that align with a client's vision and business goals. These professionals must understand how to write code using programming languages such as HTML or XML. They also create and test website applications, interfaces and navigation menus, as well as collaborate with designers to determine a website's layout and functionality.

Some web developers build the entire site; others specialize in building out particular components. For example, back-end web developers create the basic architecture of a site. Front-end developers are responsible for a site's layout and visual features. For more information, see our guide on how to become a web developer.

Basic course information

- Credits: 4
- Contact hours: TU 4-6 (pm) + FR 9-11 (am)
- Instructor: Dr. Roland Molontay
  - 2015: MSc in Applied Mathematics (BME)
  - 2015 - 18: PhD student in Network Theory (BME)
  - 2016: visiting PhD student at Brown University
  - 2021 - : founder and leader of HSDSLab
  - 2022: visiting researcher at Indiana University
  - 2018 - 2020 : research fellow at BME
  - 2021 - 2022: assistant professor at BME
  - 2023 - ssociate professor at BMR
- E-mail: molontayr@gmail.com, data.science.ait@gmail.com
- Teaching lab session + grading homework: Kate Barnes
- Team project mentors: Donát Köller, Marcell Nagy, József Pintér

Donát Köller

Marcell Nagy

József Pintér

Kate Barnes

Teaching asssistants / graders / mentors

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 sessions, 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 14
  - You can use your own „cheat sheets”/ formula sheets /notes
- HOMEWORK problems (25%)
  - There will be 4 HW sheets (roughly in every two weeks)
  - Programming problems
- PROJECT 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
      - The same applies to ChatGPT or similar
  - 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
- Homework related questions:
  data.science.ait@gmail.com
  (Kate, Donát or József will help you!)

Homework – late submission policy

- Homework may be submitted after the deadline, but late submission will result in the following points deductions:
  - within 2 hours: 100%
  - within 24 hours: 95%
  - within 48 hours: 90%
  - within 72 hours: 85%
  - within 120 hours: 70%
  - within 168 hours: 60%
  - within 240 hours: 50%
  - otherwise: 0%

Midterm / Final

- On week 7 / on week 14
- 100 minutes long each
- 25% each
- Theoretical questions and numerical examples
- You can use your own „cheat sheets”/ formula sheets
  - As many of **your own** 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/
    - It also looks nice from a recruiter's point of view
  - A teaching assistant (mentor) will be assigned for each team to help you/ provide guidance.
- A list of potenital project ideas
  - Coming soon
  - Encouraged to choose from this list
  - If you don't find a topic that interests you, you can also come up with another project that all the team members are interested in
- Schedule
  - W4: forming teams
  - W5: project is chosen, you have talked to the assigned TA, the project plan is ready
  - W9: milestone 1
  - W12: milestone 2
  - W14/15: classroom presentation

Expectations

- Delivering a sophisticated enough data science project
- Studying related work (related papers, projects)
  - TAs will help you finding the relevant literature
  - Goal: Understand what others have done, attempt to not only reproduce the results but improve them in some respects
- 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 + indicating project preference
- W5: **Project plan**
  - After consulting with assigned TA
  - 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.

- W9: **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

- W14/15: 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

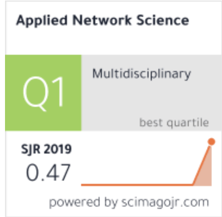
Final deliverable II.: codes

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

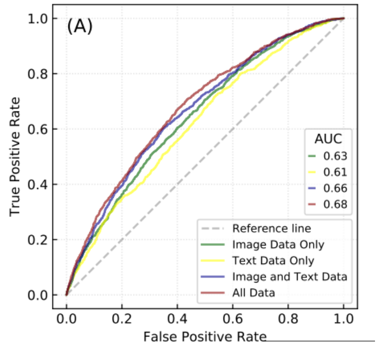
www.LeFunny.net

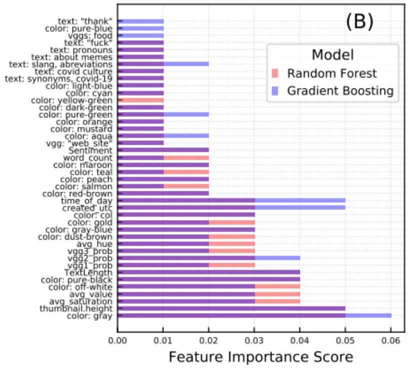
Projects from previous years

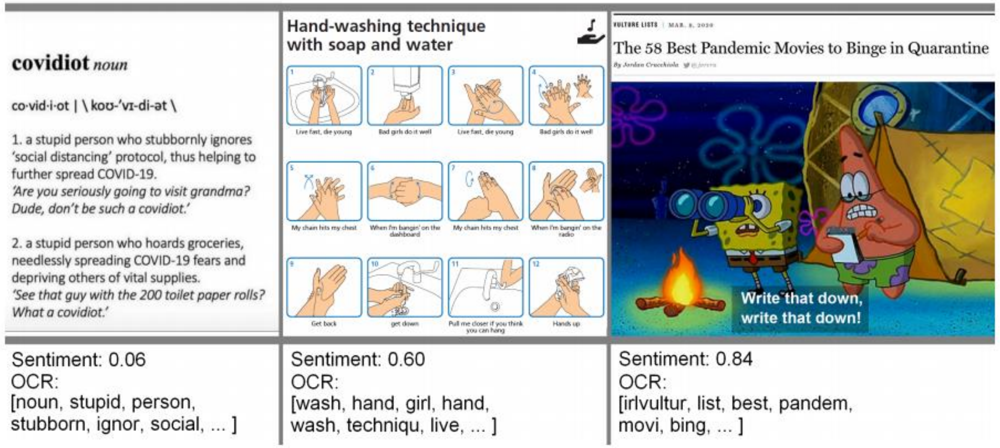
- Content-based analyis of memes
  - Predicting virality
  - Engineering image-based and text-based features



Barnes,, Riesenmy, Trinh,, Lleshi, Balogh, & Molontay, R. (2021). Dank or Not?--Analyzing and Predicting the Popularity of Memes on Reddit.Applied Network Science







Projects from previous years II.

- March Madness bracket predictions
- Detect Parkinson Disease form Voice Recording
- Song popularity prediction on Spotify
- Sarcasm Detection

List of project ideas: coming soon

- Next week I will present some ideas
- Encouraged to choose from this list
- BUT: you may choose your own project
  - Find your own data set that you are interested in
    - Various data sources available online (e.g Kaggle)
  - Use own measurements
  - Independent data collection (e.g. web scraping techniques)

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…

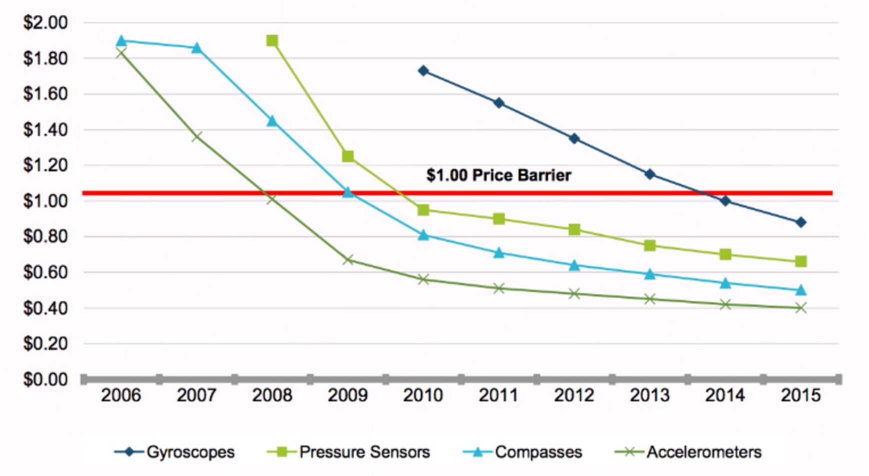
Schedule of the semester

	Monday midnight	*Tuesday class*	*Friday class*
W1 (02/06)			
W2 (02/13)		HW1 out	
W3 (02/20)			
W4 (02/27)	HW1 deadline + TEAMS	HW2 out	
W5 (03/06)	PROJECT PLAN		
W6 (03/13)	HW2 deadline	HW3 out	
W7 (03/20)			MIDTERM
SPRING BREAK		SPRING BREAK	SPRING BREAK
W8 (04/03)	HW3 deadline	HW4 out	GOOD FRIDAY
W9 (04/10)	MILESTONE 1		
W10 (04/17)	HW4 deadline		
W11 (04/24)			
W12 (05/01)	MILESTONE 2		
W13 (05/08)			
W14 (05/15)		FINAL	PROJECT presentations
W15 (05/22)		PROJECT presentations	

Historical overview

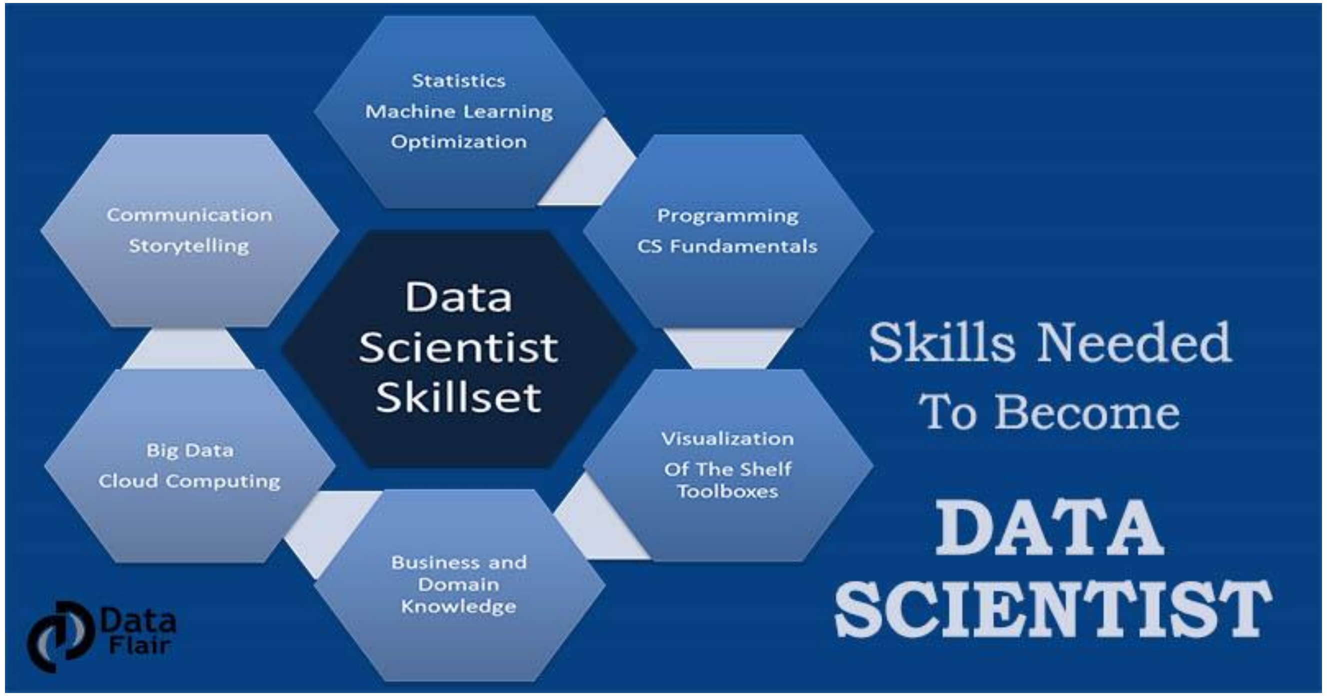
- John Tukey: The Future of Data Analysis, *Annals of Mathematical Statistics,* 1962
  - Before his time, he predicted the emergence of a new scientific discipline about data
- 80s, 90s: storage capacity increases rapidly + prices decrease ➔ data accumulation (data tomb)
  - Even exceeding Moore's law (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 + large text corpora ➔ 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 on 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."

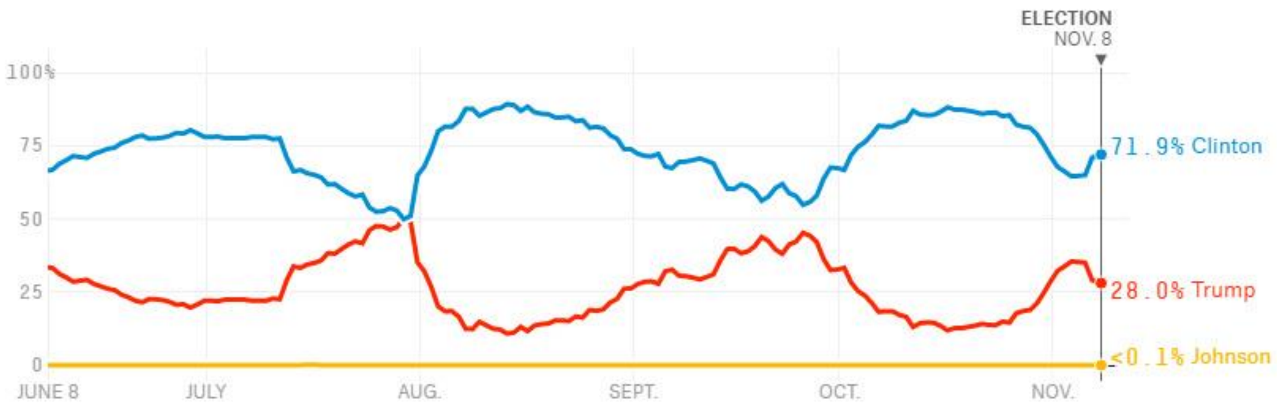
„For Nate Silver, asymptotic theory kicks in at N=1."

„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

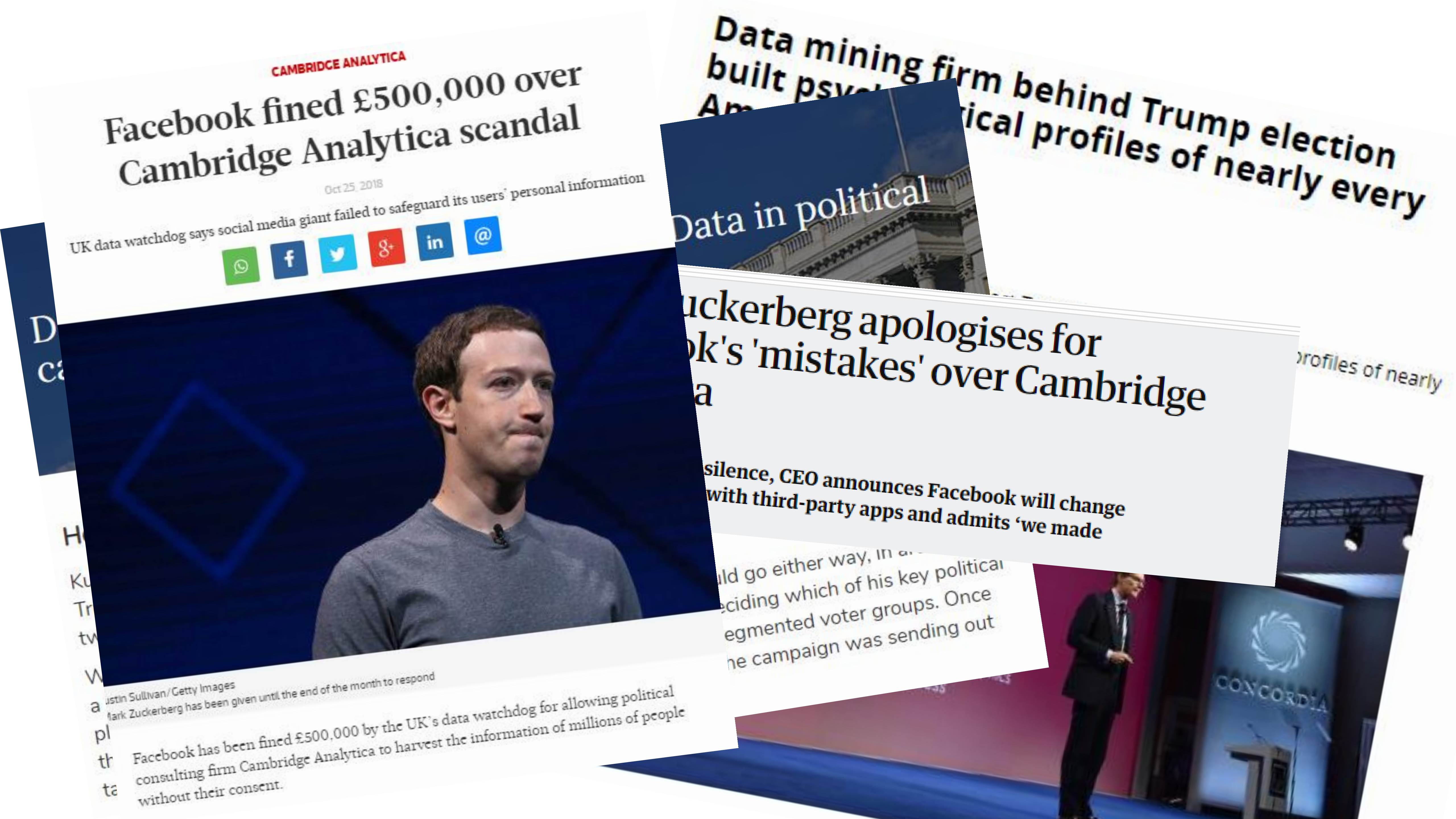
- 2016: Nate Silver claims that Clinton has much more chance to win (he gives 72% chance to this scenario)
    - Other statisticians gave even less chance for Trump
    - Trump won the election
- Big data also played a big role in the presidential campaign



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 psy
ical profiles of nearly every
Am
Data in political
uckerberg apologises for
ok's 'mistakes' over Cambridge
a
silence, CEO announces Facebook will change
with third-party apps and admits 'we made
uld go either way, in a
eciding which of his key political
egmented voter groups. Once
he campaign was sending out
profiles of nearly
CONCORDIA

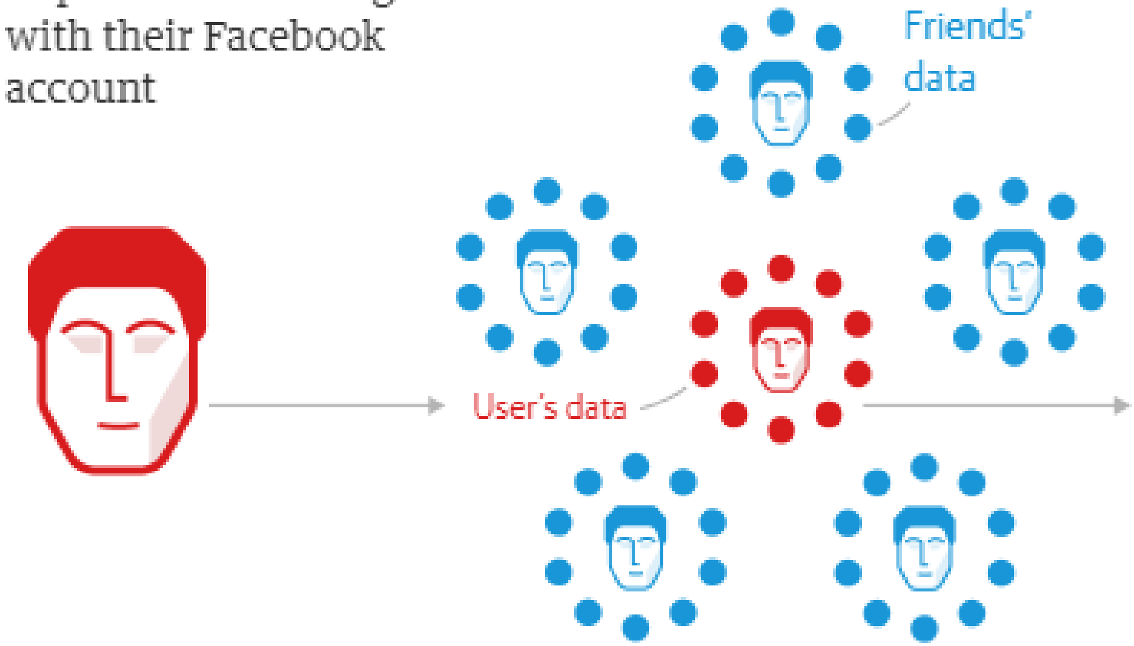
Cambridge Analytica: how 50m Facebook records were hijacked

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

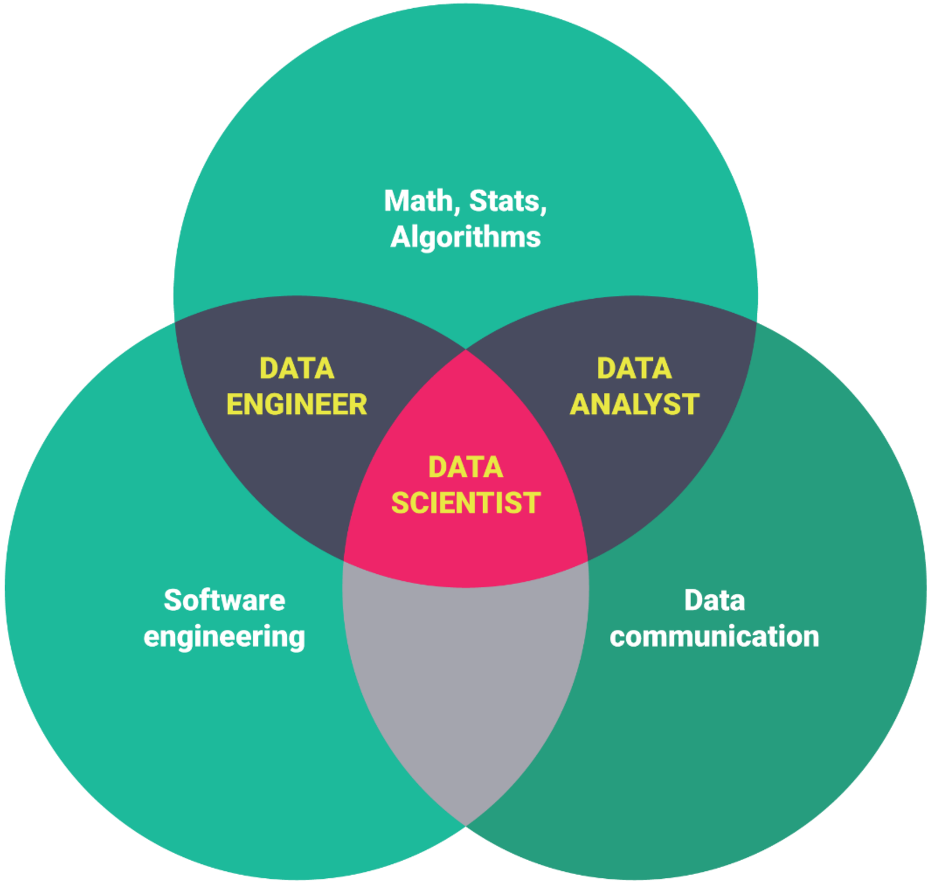
Guardian graphic. *Arkansas, Colorado, Florida, Iowa, Louisiana, Nevada, New Hampshire, North Carolina, Oregon, South Carolina, West Virginia

Statistics vs. Data science

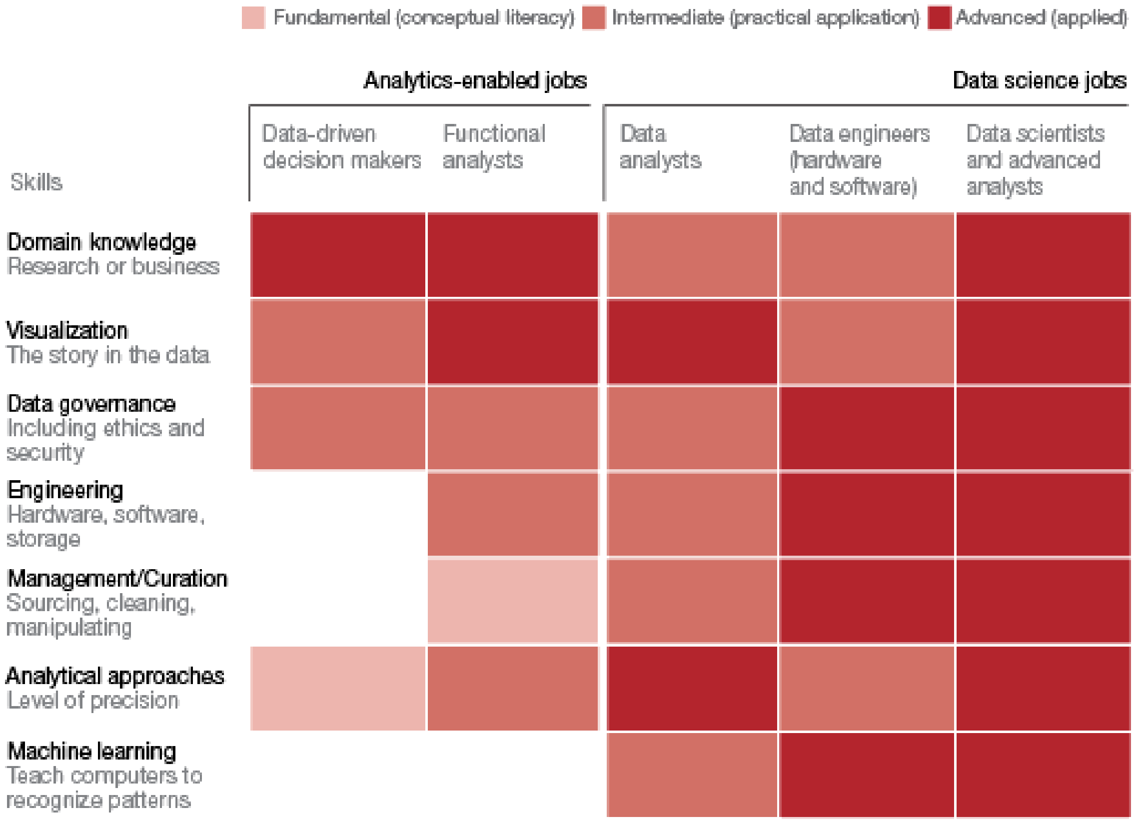
- Different aspects/approaches: testing hypothesis (statistical tests) vs. Finding hypothesis (more general question)
- Studying DNA sequences
  - Statistician: Is there a significant connection between a certain DNA subsequent and a certain disease?
  - Data scientist: What are the connections between 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?

Roles in data science

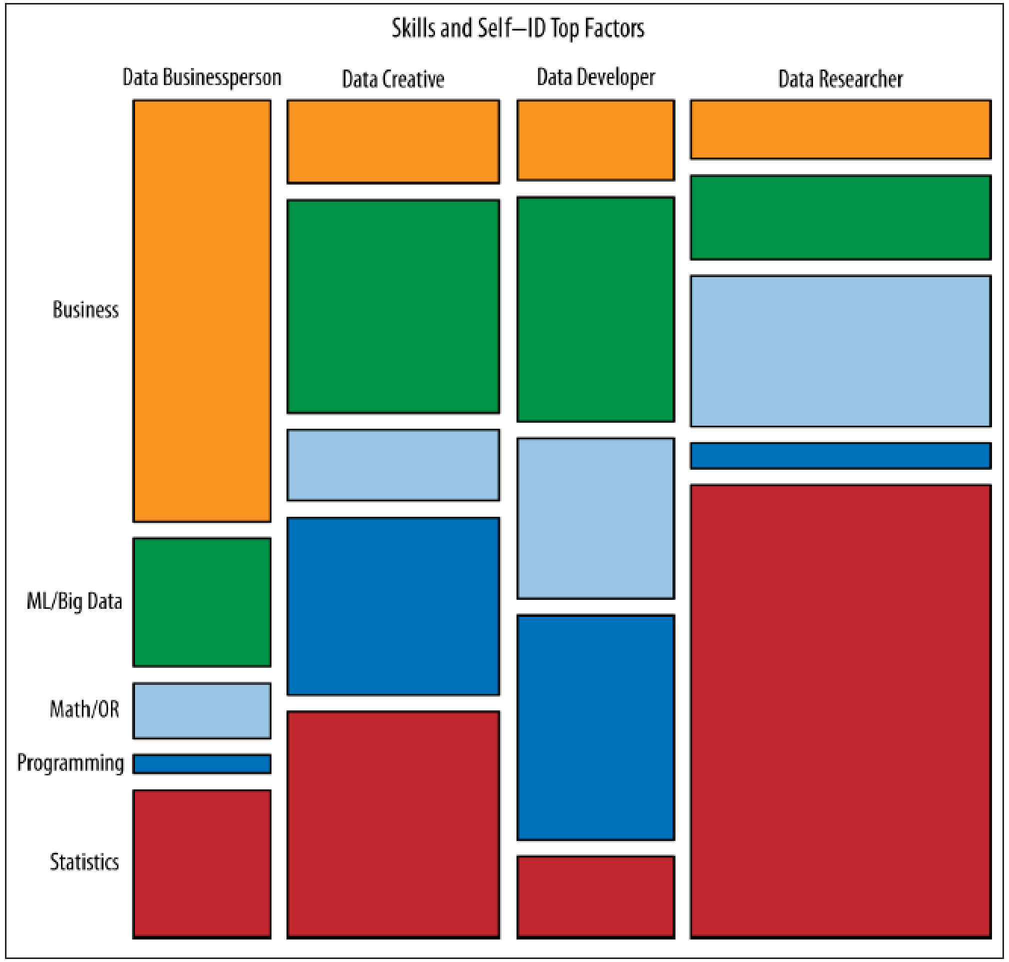
- There are no strict roles, it depends on the company, on the project
- Job listings are also ambiguous
  - Same positions may cover totally different job descriptions

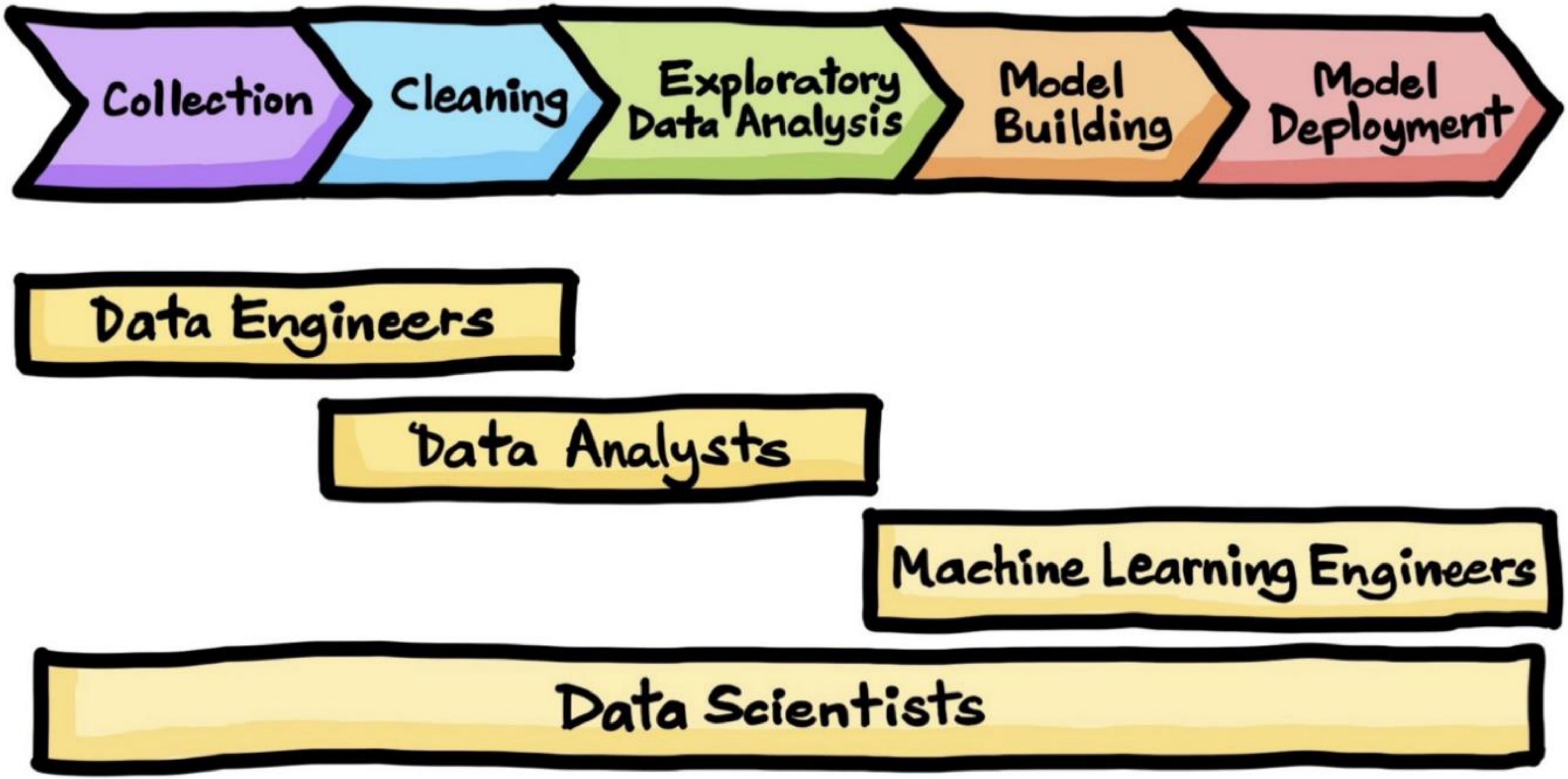


Positions and required skills



Source: PwC analysis based on Burning Glass Technologies data, January 2017.



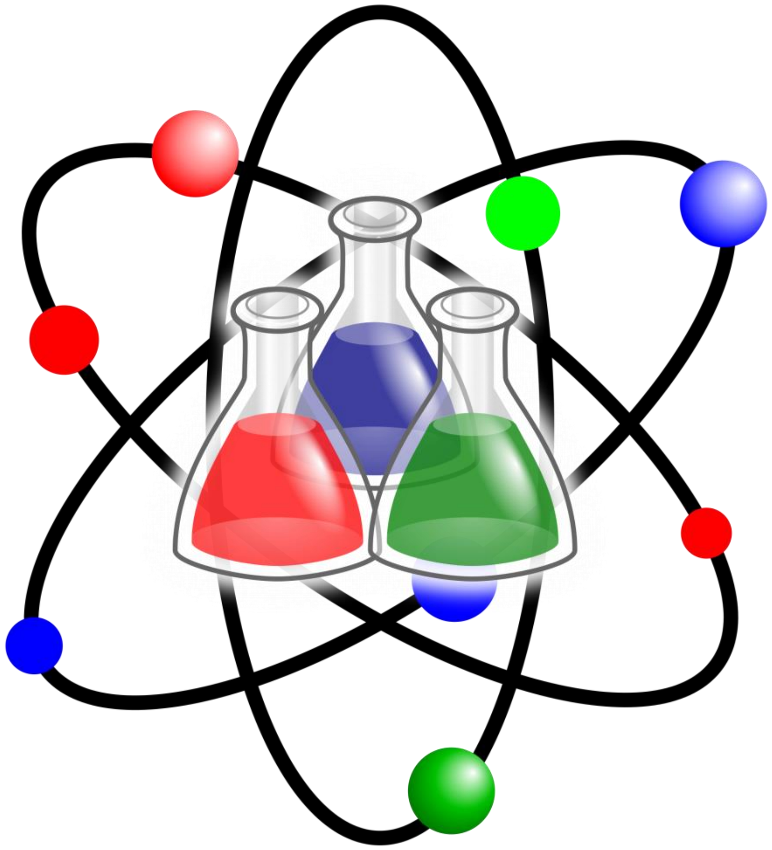

Collection
Cleaning
Exploratory Data Analysis
Model Building
Model Deployment
Data Engineers
Data Analysts
Machine Learning Engineers
Data Scientists

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)

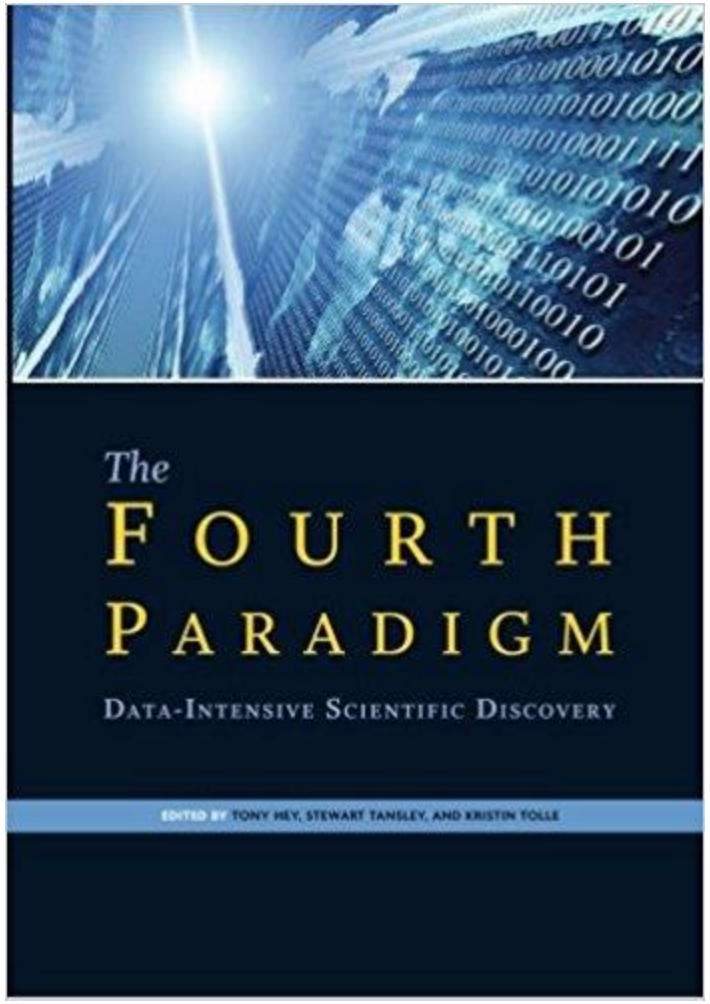
Some application areas– science

- 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

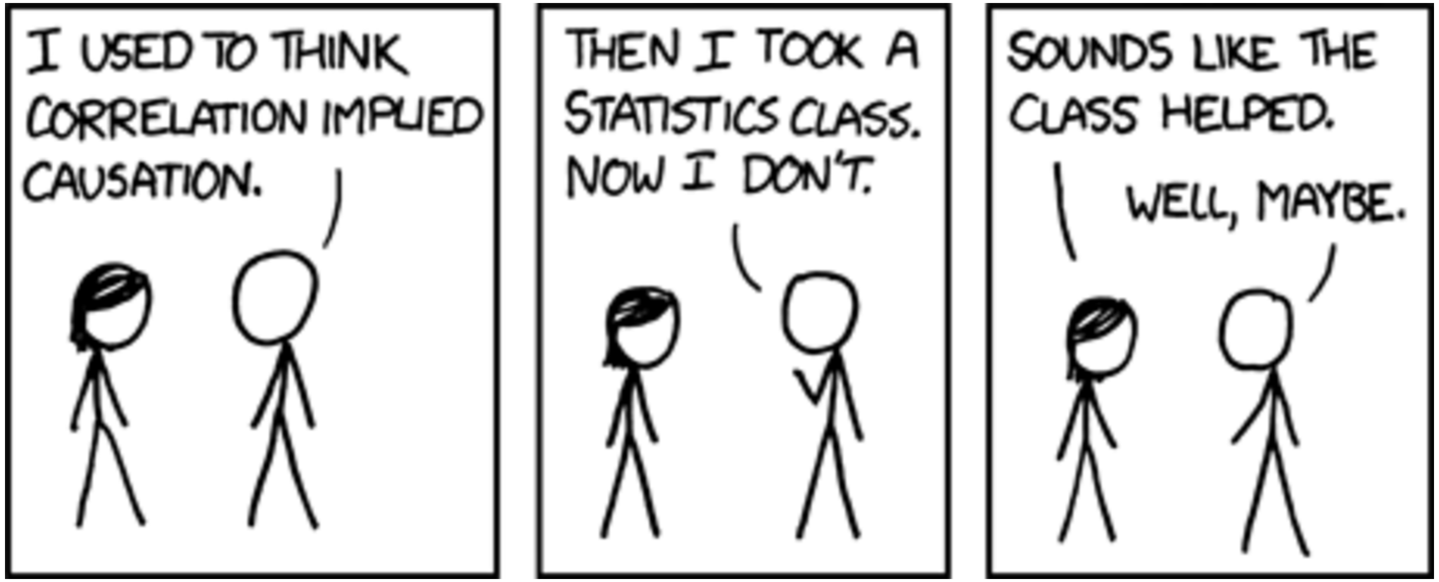


What is not data science?

- 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

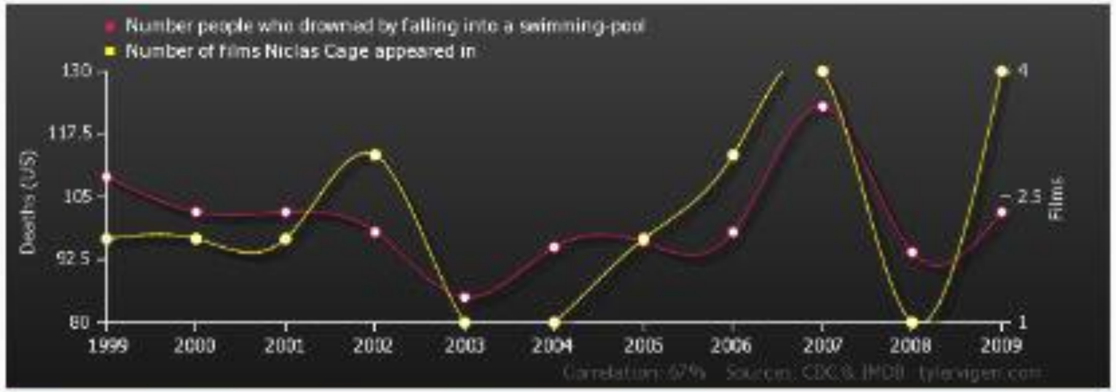
Correlation ≠ causation

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

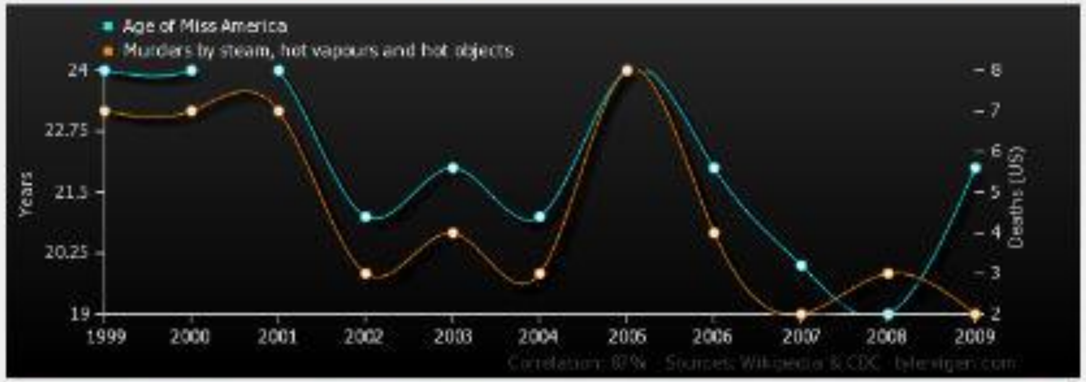


- Be careful! From data one can retrieve connections that is just there due to chance
  - You can't generalize them!
  - See next slides!

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

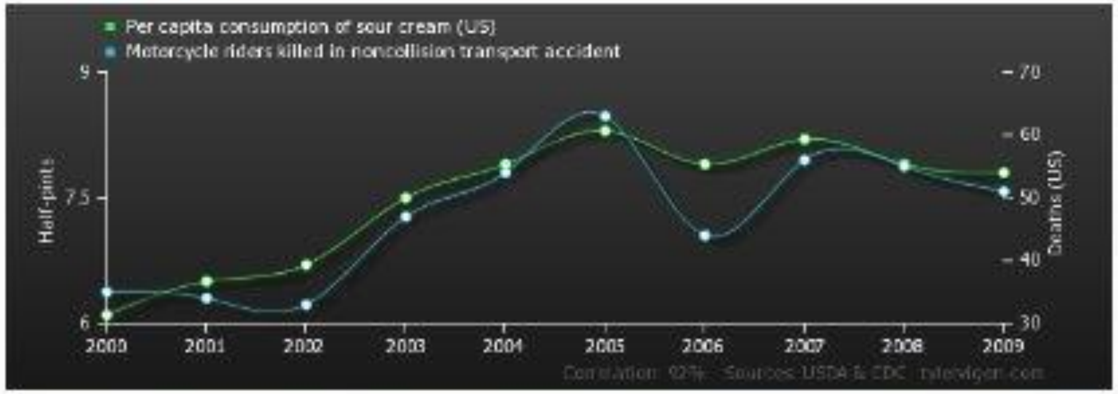


Age of Miss America
correlates with
Murders by steam, hot vapours and hot objects



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009

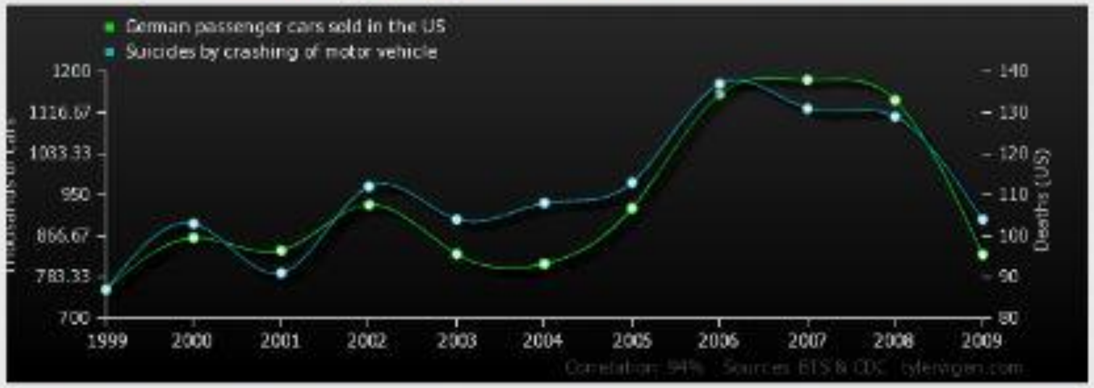
Per capita consumption of sour cream (US)
correlates with
Motorcycle riders killed in noncollision transport accident



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of sour cream (US) Half-pints (USDA)	6.1	6.5	6.7	7.5	7.9	8.3	7.9	8.2	7.9	7.8
Motorcycle riders killed in noncollision transport accident Deaths (US) (CDC)	35	34	33	47	54	63	44	56	55	51

Correlation: 0.916391

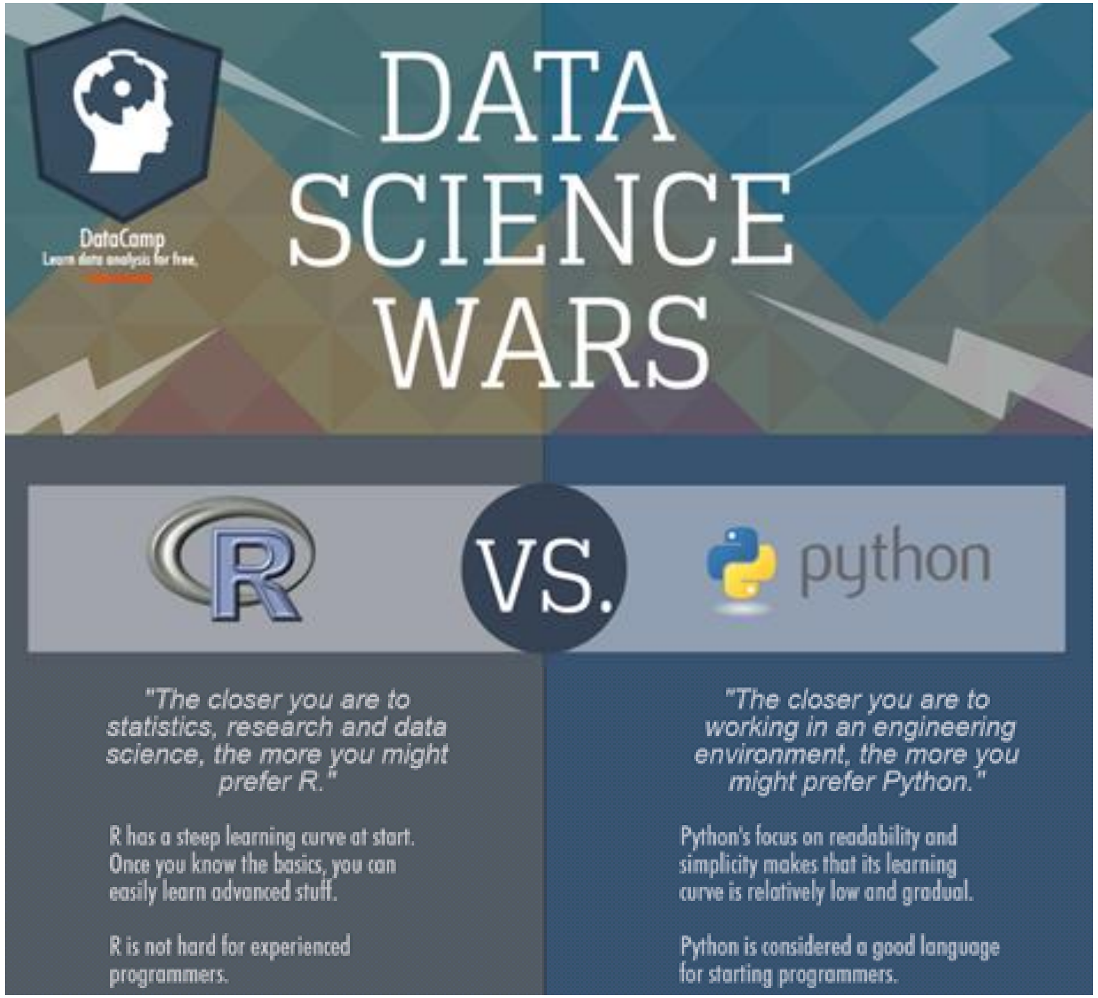
German passenger cars sold in the US
correlates with
Suicides by crashing of motor vehicle

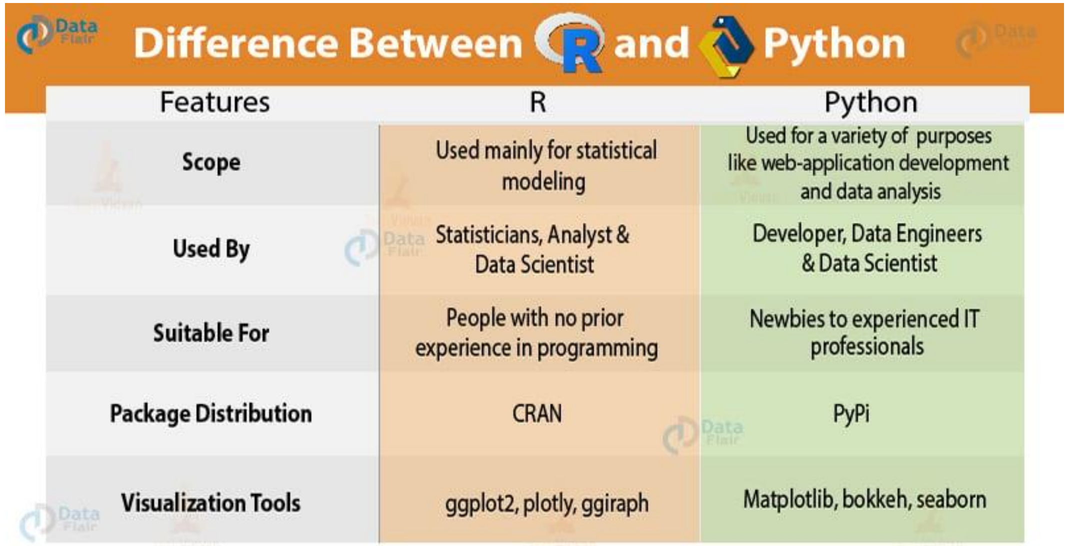


	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
German passenger cars sold in the US Thousands of cars (BTS)	758	863	837	930	830	810	923	1,154	1,183	1,142	829
Suicides by crashing of motor vehicle Deaths (US) (CDC)	87	103	91	112	104	108	113	137	131	129	104

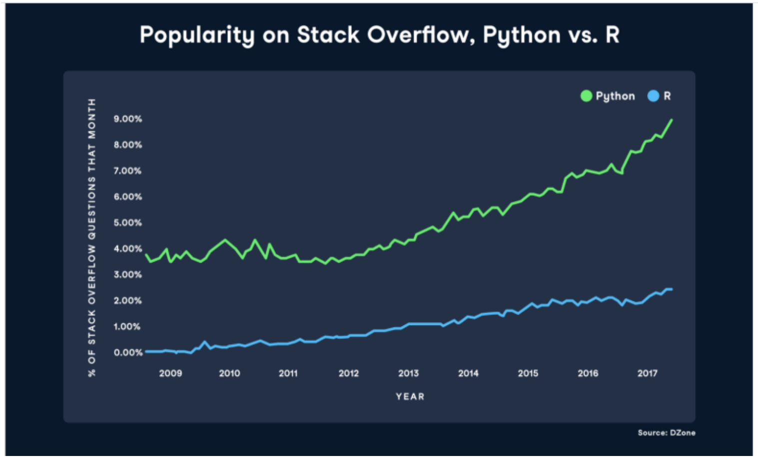
Correlation: 0.935701

Python vs. R




Difference Between R and Python

Features	R	Python
Scope	Used mainly for statistical modeling	Used for a variety of purposes like web-application development and data analysis
Used By	Statisticians, Analyst & Data Scientist	Developer, Data Engineers & Data Scientist
Suitable For	People with no prior experience in programming	Newbies to experienced IT professionals
Package Distribution	CRAN	PyPi
Visualization Tools	ggplot2, plotly, ggiraph	Matplotlib, bokkeh, seaborn



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 their 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/
- Other useful links:
  - https://ipython.org/install.html
  - https://www.continuum.io/downloads#windows

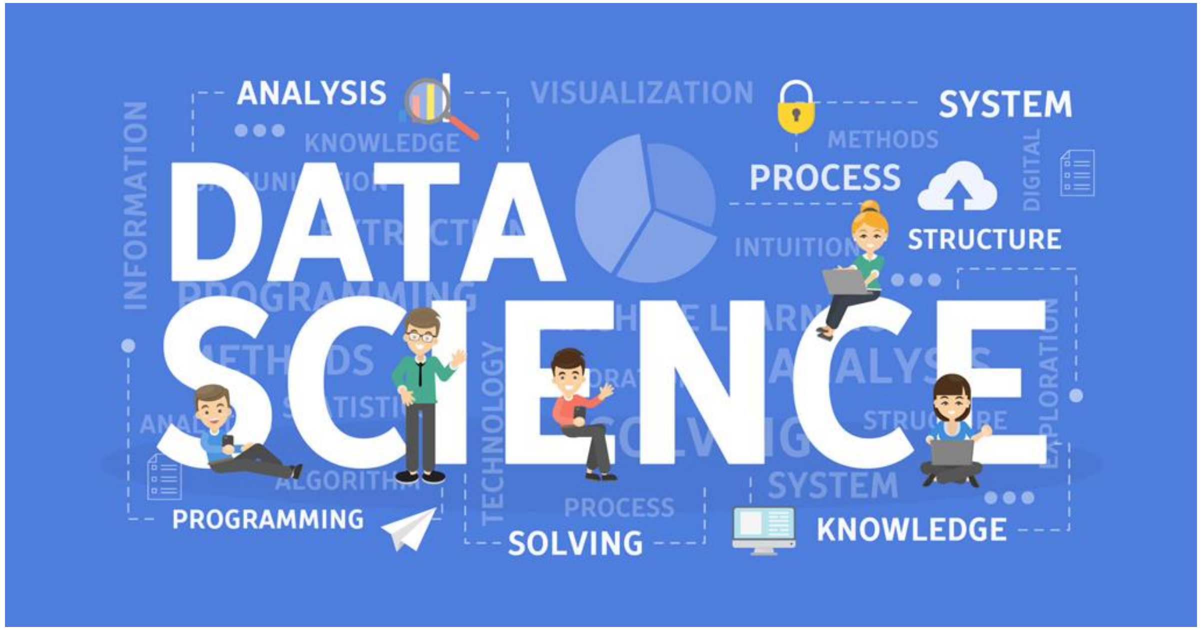
Using Ipython notebooks

- 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 (kNN, 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