Data Science Survival Skills

Introduction
WS 2023/2024

Who we are

Andreas Kist

René Groh

Hernan Aguilera

Luisa Neubig

Marion Dörrich

Nina Goes

What to expect

Lectures: We explain how things work

Exercises: You experience how things work

Homework: You get in touch with the content

Administration stuff

- Please subscribe to the **StudOn** course! (slides, exercises, homework...)

- Register for the exam on **campo**!

- Attendance in lecture and exercise is not mandatory, but strongly encouraged.

- Homework is not mandatory, but strongly encouraged.
  ➜ you get access to the solutions, but if you don't understand them, you should have asked in the exercise!

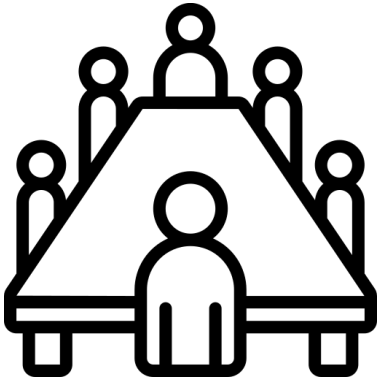
- Each **successfully** submitted and graded homework gives up to 1 bonus point

Lectures + Exercises

Lectures are Fridays 12-14 c.t.

Exercises are Wednesdays 16-18 c.t.

Homework is provided on Lecture Friday and due to the next Monday (10 days later)



All in this lecture hall!

Homework:
Task is given on a slide. Submission is due to the next Monday.

Please submit homework **until Monday 23:59 PM** to get potentially the bonus point

Content

This is the master slide!

Friday	Sat/Sun	Mon		Tue	Wednesday	Thu	Topic
20/10/2023					25/10/2023		
Introduction					Soft exercise		What is Data Science?
27/10/2023		29/10/2023			01/11/2023		
Lecture (Groh)		Voluntary homework			Online Exercise		What are computers?
3/11/2023		5/11/2023			08/11/2023		
Lecture		Homework due from	27/10/2023		Exercise		Programming 1on1
10/11/2023		12/11/2023			15/11/2023		
Lecture		Homework due from	3/11/2023		Exercise		What is actually data
17/11/2023		19/11/2023			22/11/2023		
Lecture		Homework due from	10/11/2023		Exercise		Data exploration
24/11/2023		26/11/2023			29/11/2023		
Lecture		Homework due from	17/11/2023		Exercise		Statistics
1/12/2023		3/12/2023			6/12/2023		
Lecture		Homework due from	24/11/2023		Exercise		From baselines to data imputation
8/12/2023		10/12/2023			13/12/2023		
Lecture		Homework due from	1/12/2023		Exercise		Machine Learning I
15/12/2023		17/12/2023			20/12/2023		
Lecture		Homework due from	8/12/2023		Exercise		Machine Learning II
22/12/2023		24/12/2023			27/12/2023		
Nothing		Nothing			Nothing		
29/12/2023		31/12/2023			3/1/2024		XMAS HOLIDAYS
Nothing		Nothing			Nothing		
5/1/2024		7/1/2024			10/1/2024		
Nothing		Nothing			Nothing		
12/1/2024		14/1/2024			17/1/2024		
Lecture		Homework due from	15/12/2023		Exercise		How to process natural language
19/1/2024		21/1/2024			24/1/2024		
Lecture		Homework due from	12/1/2024		Exercise		How to make code faster
26/1/2024		28/1/2024			31/1/2024		
Lecture		Homework due from	19/1/2024		Exercise		Graphical User Interfaces
2/2/2024		4/2/2024			7/2/2024		
Lecture		Homework due from	26/1/2024		Exercise		Deploying code
9/2/2024		End of semester					
Recap		Homework due from 2/2/2024					

Students

- We planned with ~ 20
- We have a room for ~ 50
- We have 135 registered students

Winter term 2022/2023:
Registered ~ 120
We have a room for ~ 100
We have another ~ 300 on the waiting list...
300 took the exam

Winter term 2023/2024:
Registered ~ 724

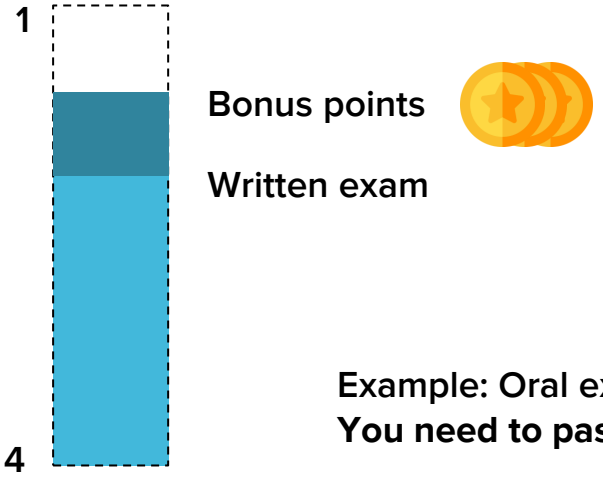
You are a lot - and we are not. Please be patient, as we need to handle all of you!

Exam

- Written Exam: 60 min
- Multiple choice + open questions (75:25)
- Content: Lectures + Exercises
- I am aiming for CONCEPTS and LOGICAL THINKING

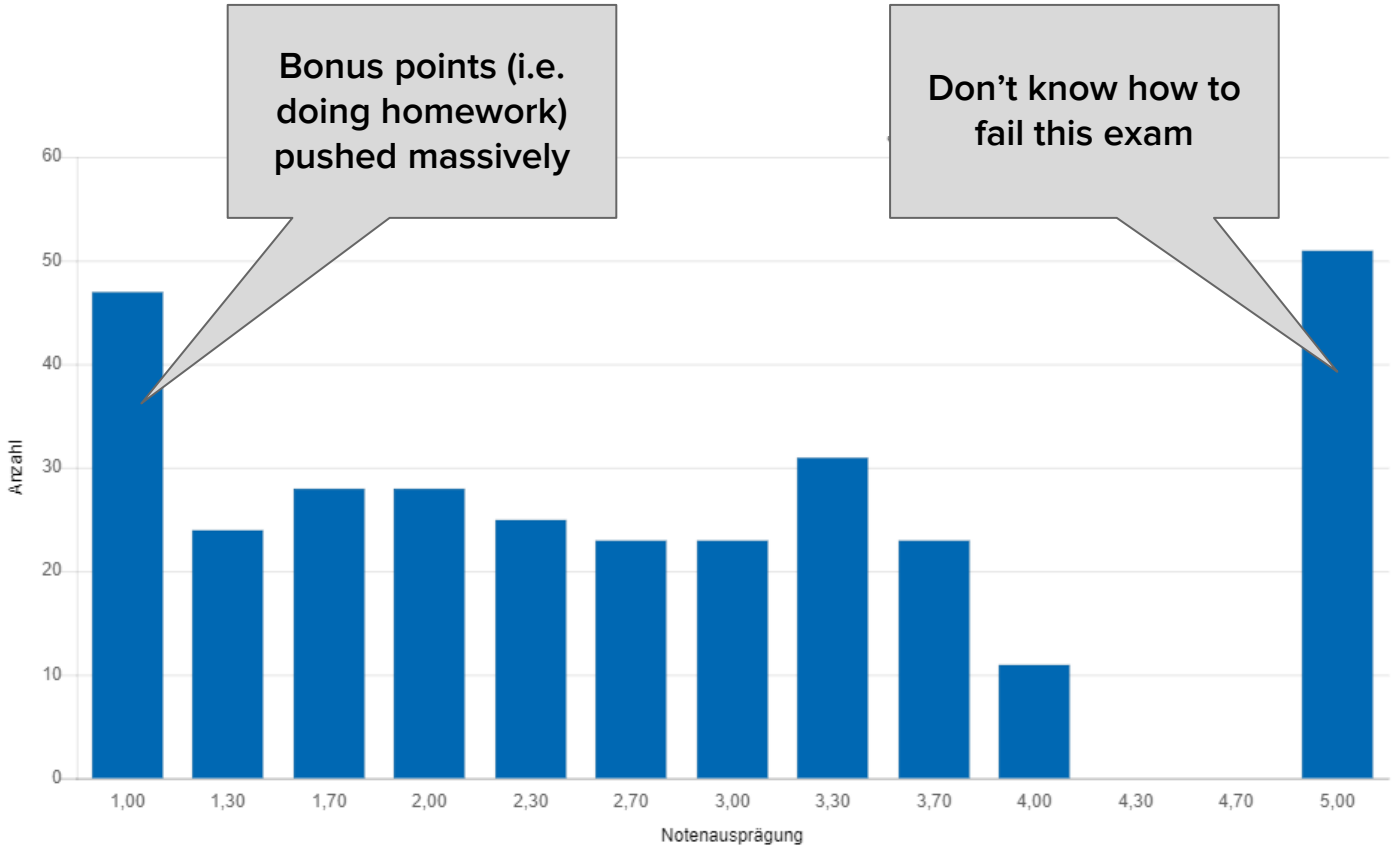
0-5 bonus points:	-0.0
6-9 bonus points:	-0.3
10+ bonus point:	-0.7

Grade:



Example: Oral exam 2,3 + 10 bonus points → 1,7
You need to pass the exam to receive bonus effect

Grades winter term (regular exam)



How does homework submission work?

A single pdf slide to be submitted

Homework:
Task A,
Task B

A passionate DSSS student

YOUR SOLUTION

Name:
Matriculation number:
IdM:

STUDON

Upload in time to StudOn submission folder.
Submission due Mondays 23:59 PM.
1 second too late is TOO LATE!
No late submissions accepted - no exceptions.

What to do when you have a (real!) problem

No E-Mails!

Please use the StudOn forum, such that anyone could answer!

Real (!) problems are:

- You have a question related to the lecture
  That you CANNOT FIND ANYWHERE ONLINE!!

- In your exam preparation you came across a problem re the content,
  That you CANNOT SOLVE USING THE LIBRARY or STACKOVERFLOW/GOOGLE.

And give us enough time,
e.g. two days before the exam is not the ideal moment!

Expectations

Student expectations

Please get in touch with your fellow students and ask yourself the following questions:

- What do I want from the course?
- How can I achieve this?
- How can I actively contribute to the course?
- What do expect from lecturers?

5 minutes

My/our expectations

- Be at and on time for lectures
- Do the exercises/homework
- Ask questions
- Use the course forum!

I will not answer E-mails when you can find the information online etc

Data Science

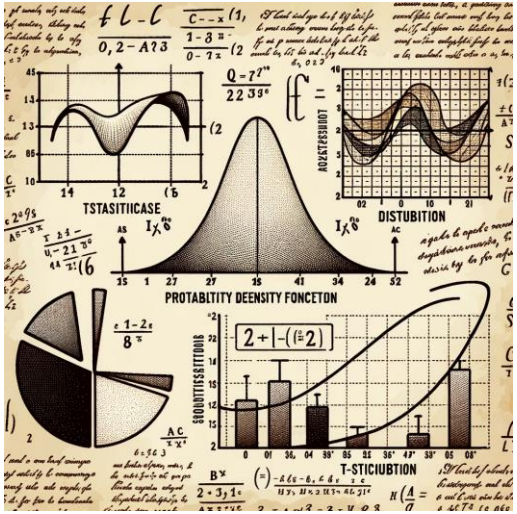
We live in a world of data

1900s

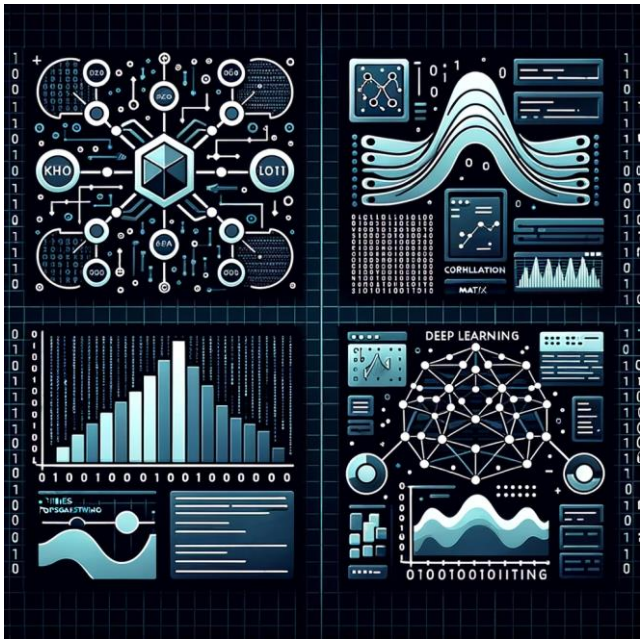
2020s

Why do we need “data science”?

Statistics



- Likelihood, Probabilities
- PDFs
- Descriptive statistics
- Explorative statistics



What we can’t do with statistics alone:

- Machine Learning
- Working with unstructured data (Deep Learning)
- Complex time-series forecasting
- Clustering

A bit of history

AI Magazine Volume 17 Number 3 (1996) (© AAAI)

There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data.

From Data Mining to Knowledge Discovery in Databases

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth

From data mining to knowledge discovery in databases

U Fayyad, G Piatetsky-Shapiro, P Smyth

AI magazine, 1996 · ojs.aaai.org

Abstract

Data mining and knowledge discovery in databases have been attracting a significant amount of research, industry, and media attention of late. What is all the excitement about? This article provides an overview of this emerging field, clarifying how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases. The article mentions particular real-world applications, specific data-mining techniques, challenges involved in real-world

MEHR ANZEIGEN

Coining the word “data science”

International Statistical Review (2001), 69, 1, 21–26, Printed in Mexico

Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland

Statistics Research, Bell Laboratories, 600 Mountain Avenue, Murray Hill
E-mail: wsc@research.bell-labs.com

Summary

An action plan to enlarge the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department, and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations.

Key words: Future; Applications; Computing; Methods; Models; Theory.

Why not using Data Mining (a common concept based around statistics) and Computer Science to take advantage of both
➜ Data Science.

What changed?

Wikipedia size and users

Update	
English articles	6,730,059
Total wiki pages	59,193,160
Article percentage	11.37%
Average revisions	19.86
Total admins	881
Total users	46,321,402

UTC time: 16:06 on 2023-Oct-17

Read only
“I am online”

Only consuming

Read+Write
“I am contributing”

- Social media
  - Myspace
  - Facebook
  - YouTube
- Communicate
- Spread information
- Wikipedia

Let's define the job of data science.

Tons of data, from shopping to trading, health-related information, email conversations, ...

Messy, unstructured, maybe totally irrelevant data

Taking messy data and creating/gaining insights

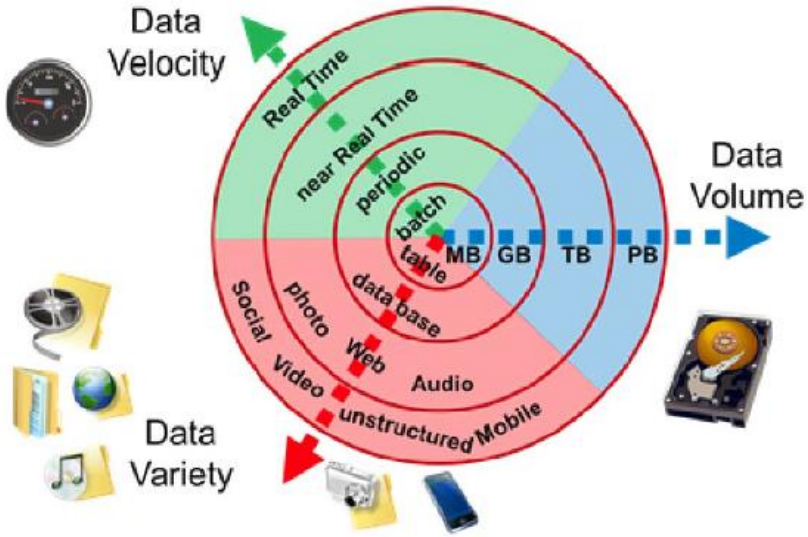
Takeaways, relevant variables, biomarkers, ...

How large is data?

BIG DATA

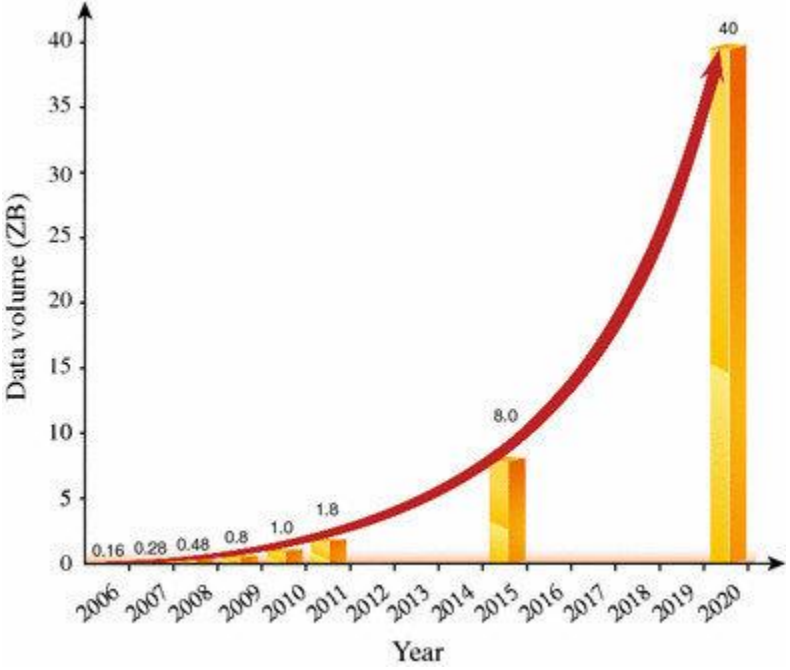
Value	Metric		Value	IEC		Memory	
1000	kB	kilobyte	1024	KiB	kibibyte	KB	kilobyte
1000^2	MB	megabyte	1024^2	MiB	mebibyte	MB	megabyte
1000^3	GB	gigabyte	1024^3	GiB	gibibyte	GB	gigabyte
1000^4	TB	terabyte	1024^4	TiB	tebibyte	TB	terabyte
1000^5	PB	petabyte	1024^5	PiB	pebibyte		–
1000^6	EB	exabyte	1024^6	EiB	exbibyte		–
1000^7	ZB	zettabyte	1024^7	ZiB	zebibyte		–
1000^8	YB	yottabyte	1024^8	YiB	yobibyte		–

Orders of magnitude of data



By Ender005 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=49888192

How much data is around?



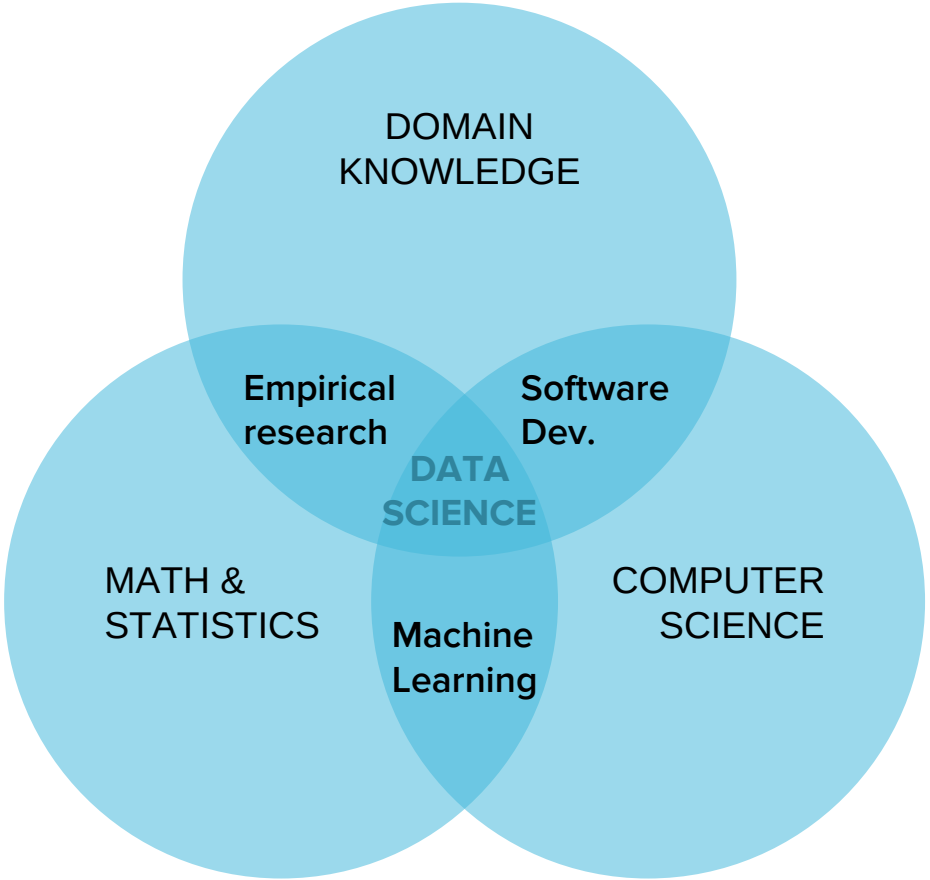
Global growth trend of data volume, 2006–2020 (based on “The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east”)

Value	Metric		Value	IEC		Memory	
1000	kB	kilobyte	1024	KiB	kibibyte	KB	kilobyte
1000^2	MB	megabyte	1024^2	MiB	mebibyte	MB	megabyte
1000^3	GB	gigabyte	1024^3	GiB	gibibyte	GB	gigabyte
1000^4	TB	terabyte	1024^4	TiB	tebibyte	TB	terabyte
1000^5	PB	petabyte	1024^5	PiB	pebibyte		–
1000^6	EB	exabyte	1024^6	EiB	exbibyte		–
1000^7	ZB	zettabyte	1024^7	ZiB	zebibyte		–
1000^8	YB	yottabyte	1024^8	YiB	yobibyte		–

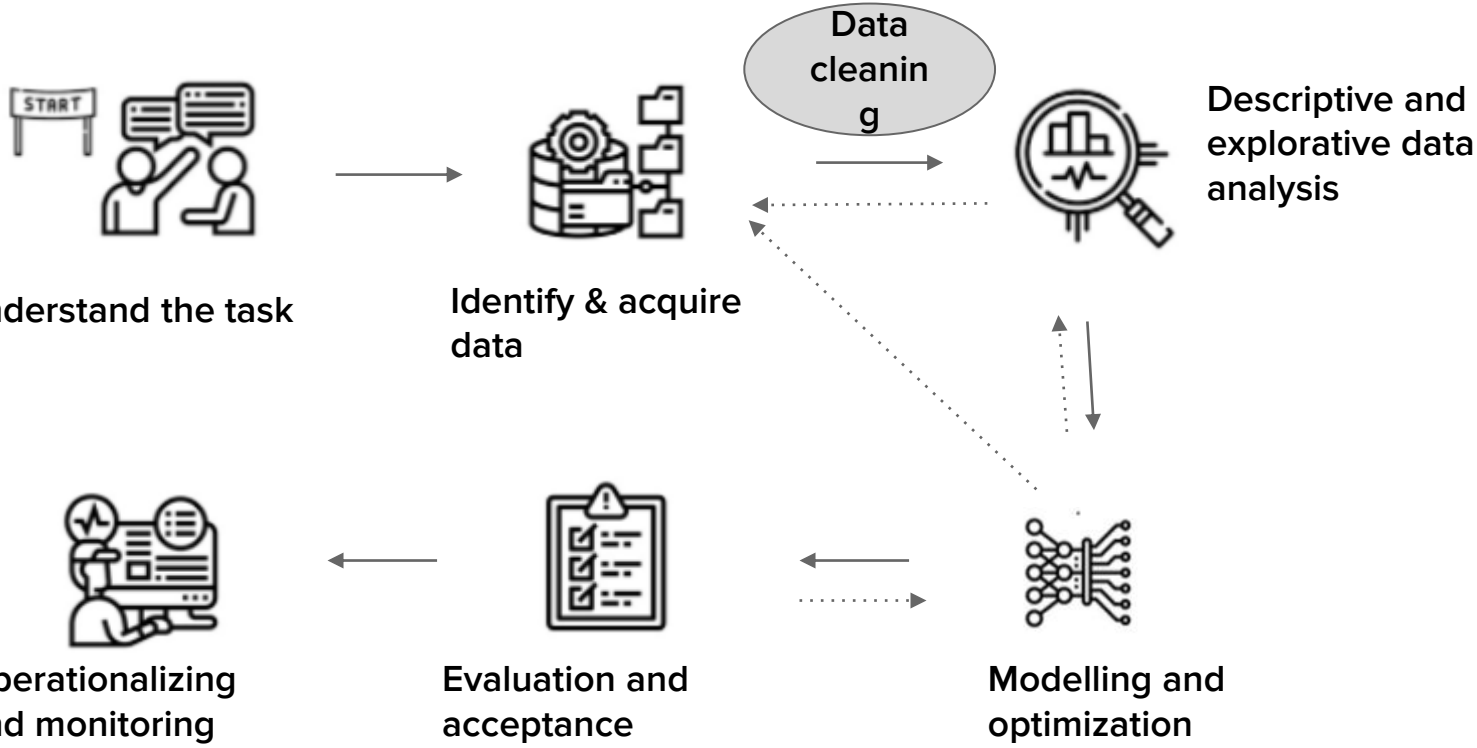
Orders of magnitude of data

Exponential growth of data!

What is “Data Science”?



Data Science Workflow



Tools that people need in Data Science

Friday	Sat/Sun	Mon		Tue	Wednesday	Thu	Topic
20/10/2023					25/10/2023		
Introduction					Soft exercise		What is Data Science?
27/10/2023		29/10/2023			01/11/2023		
Lecture (Groh)		Voluntary homework			Online Exercise		What are computers?
3/11/2023		5/11/2023			08/11/2023		
Lecture		Homework due from	27/10/2023		Exercise		Programming 1on1
10/11/2023		12/11/2023			15/11/2023		
Lecture		Homework due from	3/11/2023		Exercise		What is actually data
17/11/2023		19/11/2023			22/11/2023		
Lecture		Homework due from	10/11/2023		Exercise		Data exploration
24/11/2023		26/11/2023			29/11/2023		
Lecture		Homework due from	17/11/2023		Exercise		Statistics
1/12/2023		3/12/2023			6/12/2023		
Lecture		Homework due from	24/11/2023		Exercise		From baselines to data imputation
8/12/2023		10/12/2023			13/12/2023		
Lecture		Homework due from	1/12/2023		Exercise		Machine Learning I
15/12/2023		17/12/2023			20/12/2023		
Lecture		Homework due from	8/12/2023		Exercise		Machine Learning II
22/12/2023		24/12/2023			27/12/2023		XMAS HOLIDAYS
Nothing		Nothing			Nothing		
29/12/2023		31/12/2023			3/1/2024		
Nothing		Nothing			Nothing		
5/1/2024		7/1/2024			10/1/2024		
Nothing		Nothing			Nothing		
12/1/2024		14/1/2024			17/1/2024		
Lecture		Homework due from	15/12/2023		Exercise		How to process natural language
19/1/2024		21/1/2024			24/1/2024		
Lecture		Homework due from	12/1/2024		Exercise		How to make code faster
26/1/2024		28/1/2024			31/1/2024		
Lecture		Homework due from	19/1/2024		Exercise		Graphical User Interfaces
2/2/2024		4/2/2024			7/2/2024		
Lecture		Homework due from	26/1/2024		Exercise		Deploying code
9/2/2024		End of semester					
Recap		Homework due from 2/2/2024					

- Python programming 2.0 (numpy, pandas, ...)
- What is data? File types and Co.
- Data exploration techniques + visualization
- Statistics
- Machine Learning
- NLP
- Code enhancement (numba, Cython)
- GUIs
- Deploying code

Roles in Data Science

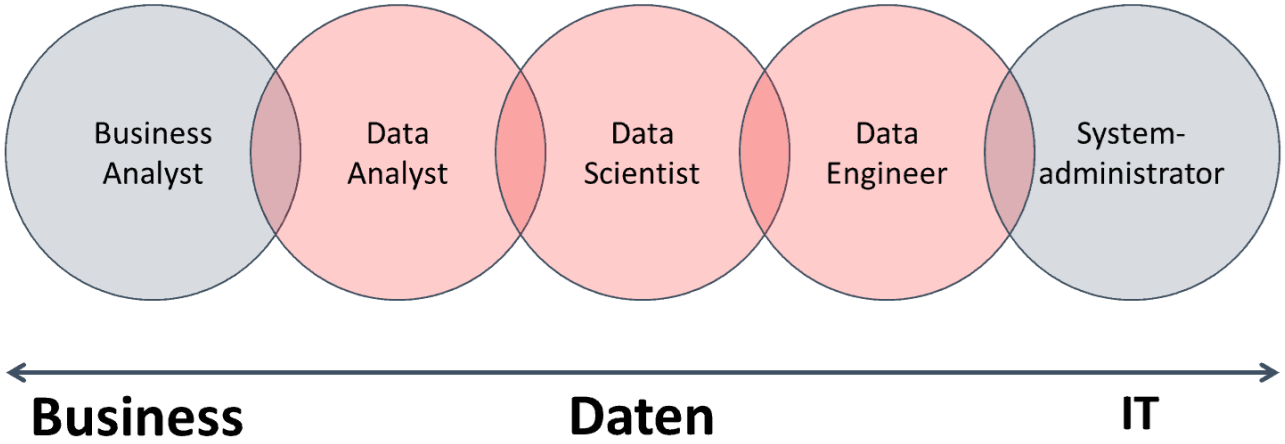
Data Analyst: Analyzes data to provide actionable insights.

Data Engineer: Manages and optimizes databases to handle and query data.

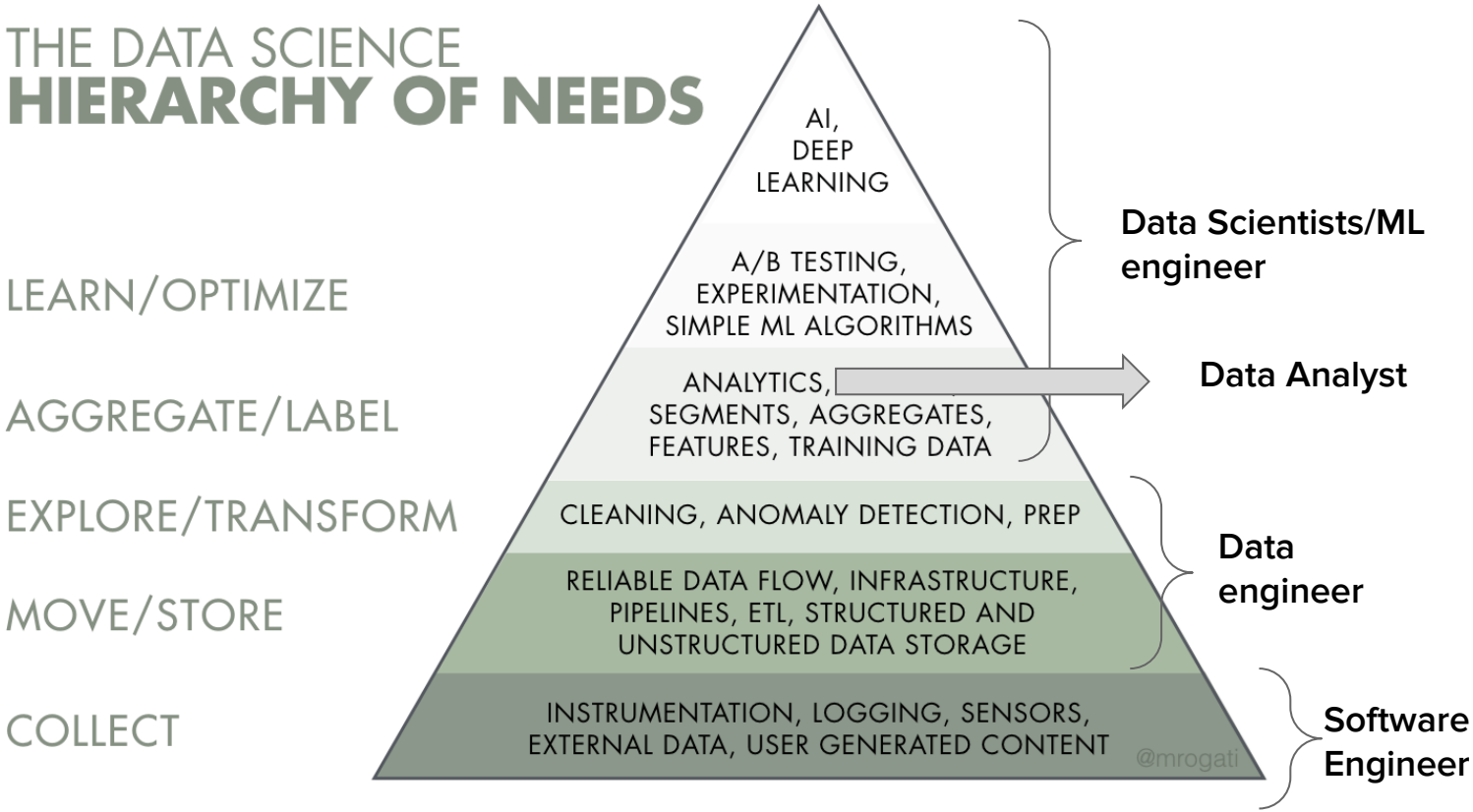
Machine Learning Engineer: Designs and implements machine learning models.

Data Scientist: Encompasses roles of data analyst and machine learning engineer, often with domain expertise.

Roles in Data Science



Data Science Pyramid of Needs



https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

Next week

How does a computer actually work? From transistors to ASICs.

Homework

Description of the homework

- We put an example Jupyter notebook on StudOn,
  That should help you get started **with Colab and numpy.**
  This is voluntary homework until next week.