Data Science Survival Skills

Introduction
WS 2024/2025

Who we are

Andreas Kist

René Groh

and
Selina Baumgart
Florian Gritsch
Reem Hasheem

Luisa Neubig

Nina Goes

What to expect

https://www.adaptnetwork.com/outdoors/how-to-teach-wilderness-surviv
al-skills-to-children/

DSSS is **hard** work

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 Mondays 16-18 c.t.

Exercises are Wednesdays 16-18 c.t.

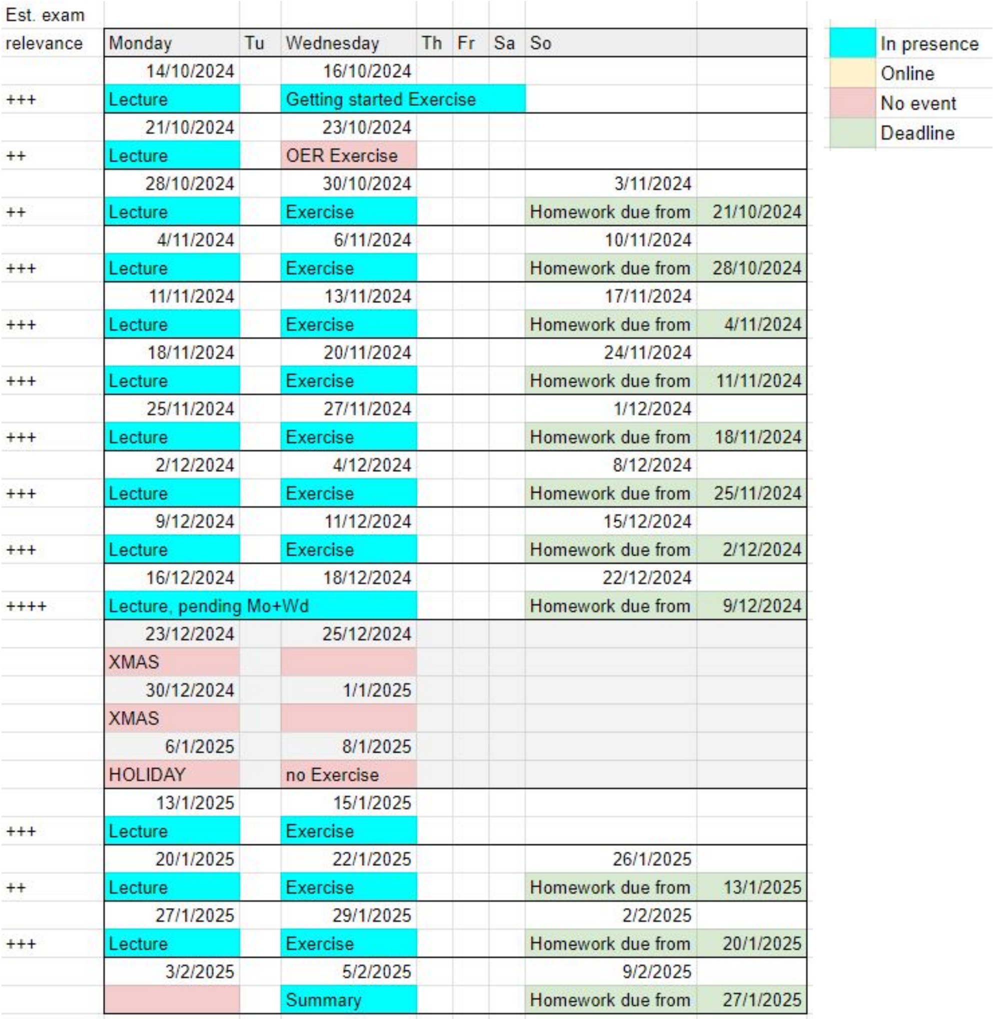
Homework is provided on Lecture Monday and due Sunday the week after (almost 14 days later)

All in this lecture hall!

<u>Homework</u>:
Task is given on a slide.

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

When And Where



Est. exam relevance	Monday	Tu	Wednesday	Th	Fr	Sa	So	
	14/10/2024		16/10/2024					
+++	Lecture		Getting started Exercise					
	21/10/2024		23/10/2024					
++	Lecture		OER Exercise					
	28/10/2024		30/10/2024				3/11/2024	
++	Lecture		Exercise				Homework due from	21/10/2024
	4/11/2024		6/11/2024				10/11/2024	
+++	Lecture		Exercise				Homework due from	28/10/2024
	11/11/2024		13/11/2024				17/11/2024	
+++	Lecture		Exercise				Homework due from	4/11/2024
	18/11/2024		20/11/2024				24/11/2024	
+++	Lecture		Exercise				Homework due from	11/11/2024
	25/11/2024		27/11/2024				1/12/2024	
+++	Lecture		Exercise				Homework due from	18/11/2024
	2/12/2024		4/12/2024				8/12/2024	
+++	Lecture		Exercise				Homework due from	25/11/2024
	9/12/2024		11/12/2024				15/12/2024	
+++	Lecture		Exercise				Homework due from	2/12/2024
	16/12/2024		18/12/2024				22/12/2024	
++++	Lecture, pending Mo+Wd						Homework due from	9/12/2024
	23/12/2024		25/12/2024					
	XMAS							
	30/12/2024		1/1/2025					
	XMAS							
	6/1/2025		8/1/2025					
	HOLIDAY		no Exercise					
	13/1/2025		15/1/2025					
+++	Lecture		Exercise					
	20/1/2025		22/1/2025				26/1/2025	
++	Lecture		Exercise				Homework due from	13/1/2025
	27/1/2025		29/1/2025				2/2/2025	
+++	Lecture		Exercise				Homework due from	20/1/2025
	3/2/2025		5/2/2025				9/2/2025	
			Summary				Homework due from	27/1/2025

🟦	In presence
🟨	Online
🟥	No event
🟩	Deadline

This is the master slide!

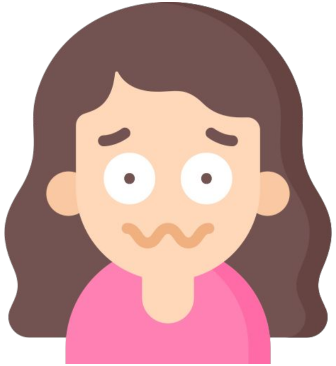
Topics

- Technical equipment/hardware basics (CPU, GPU, TPUs)
- From Clean Code to Version Control, Python Package Management, Documentation
- What is data? Differences in data and file types
- Data exploration
- Statistics
- Baselines, Data Mining
- Machine Learning, Deep Learning, (Meta)Heuristics
- Nature Language Processing/LLMs
- Multiprocessing/multithreading, Numba, Code vectorization
- GUIs and Code Deployment

Students

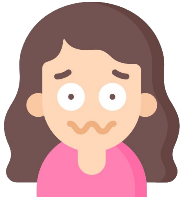
- We planned with ~ 20
- We have a room for ~ 50
- We have 135 registered students

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



Winter term 2024/2026:

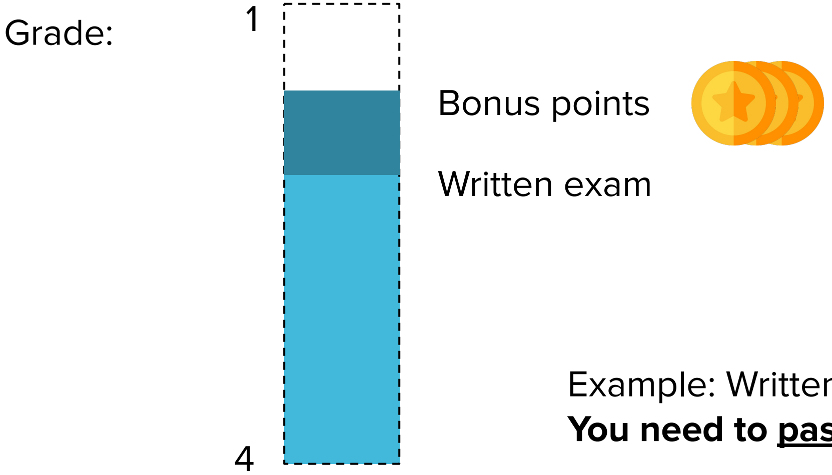
Registered ~ 330

Exam

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

max 11

- Centrally organized, we do not know when and where
- Written Exam: 60 min
- Single choice (Answers A-E, only one is correct)
- Content: Lectures + Exercises + Homework
- I am aiming for CONCEPTS and LOGICAL THINKING



Grade:

1

4

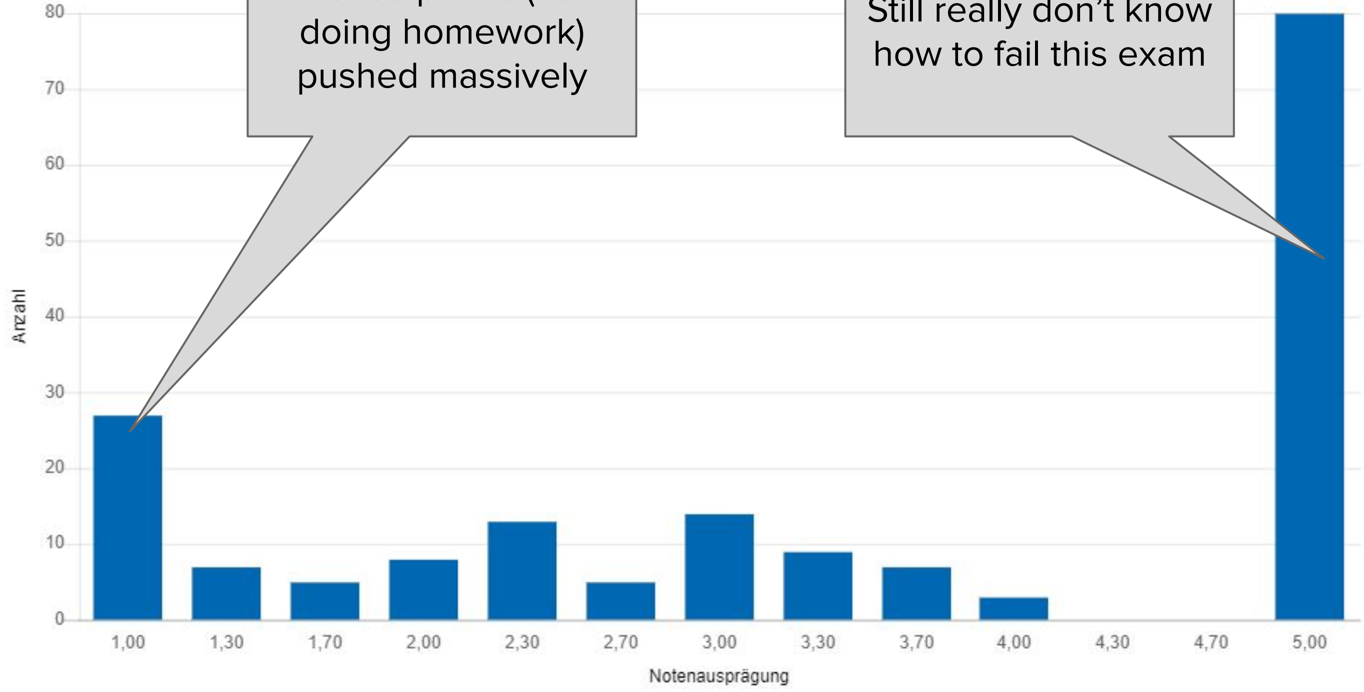
Bonus points

Written exam

Example: Written exam 2,3 + 10 bonus points → 1,7

You need to <u>pass</u> the exam to receive bonus effect

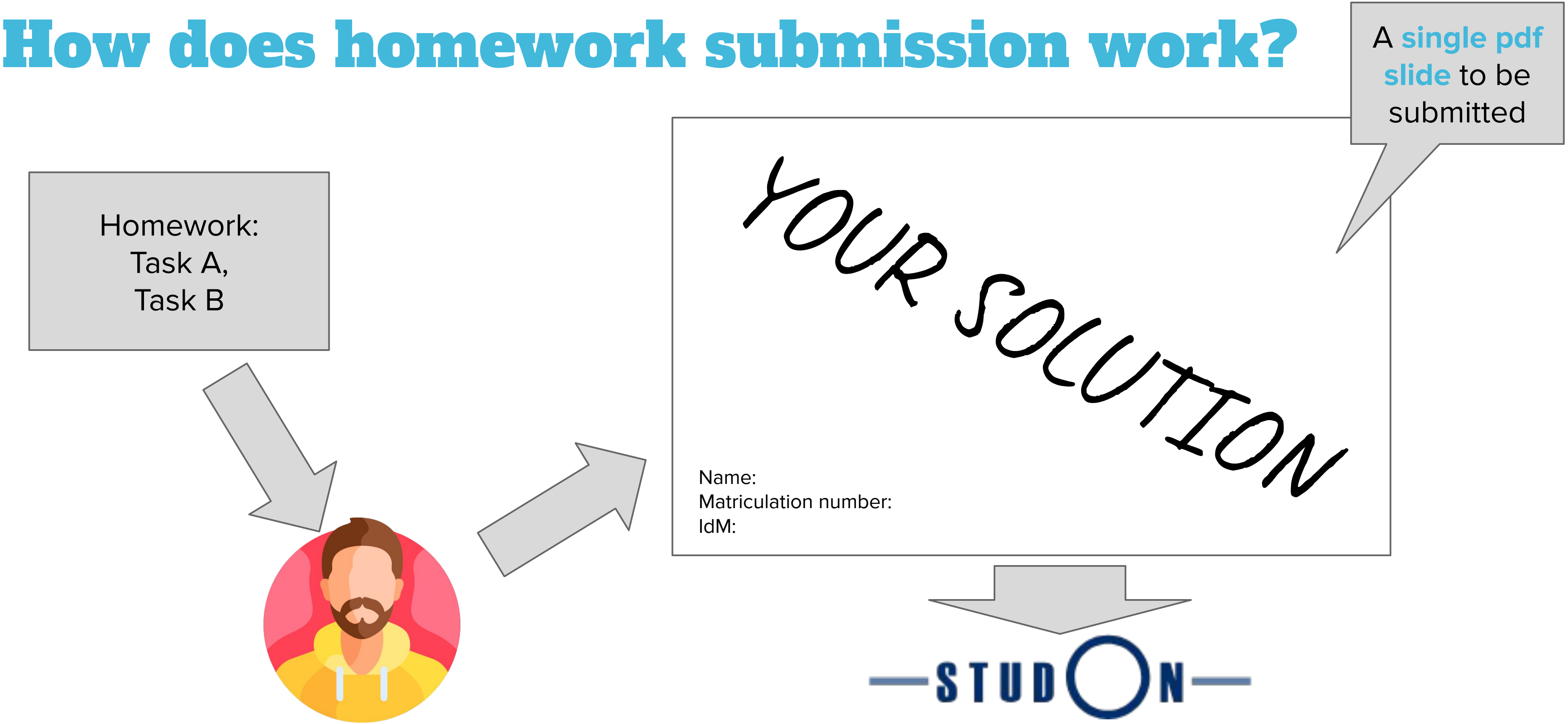
Grades summer term (repeaters)



Bonus points (i.e. doing homework) pushed massively

Still really don’t know how to fail this exam

How does homework submission work?



A single pdf slide to be submitted

Homework:
Task A,
Task B

YOUR SOLUTION

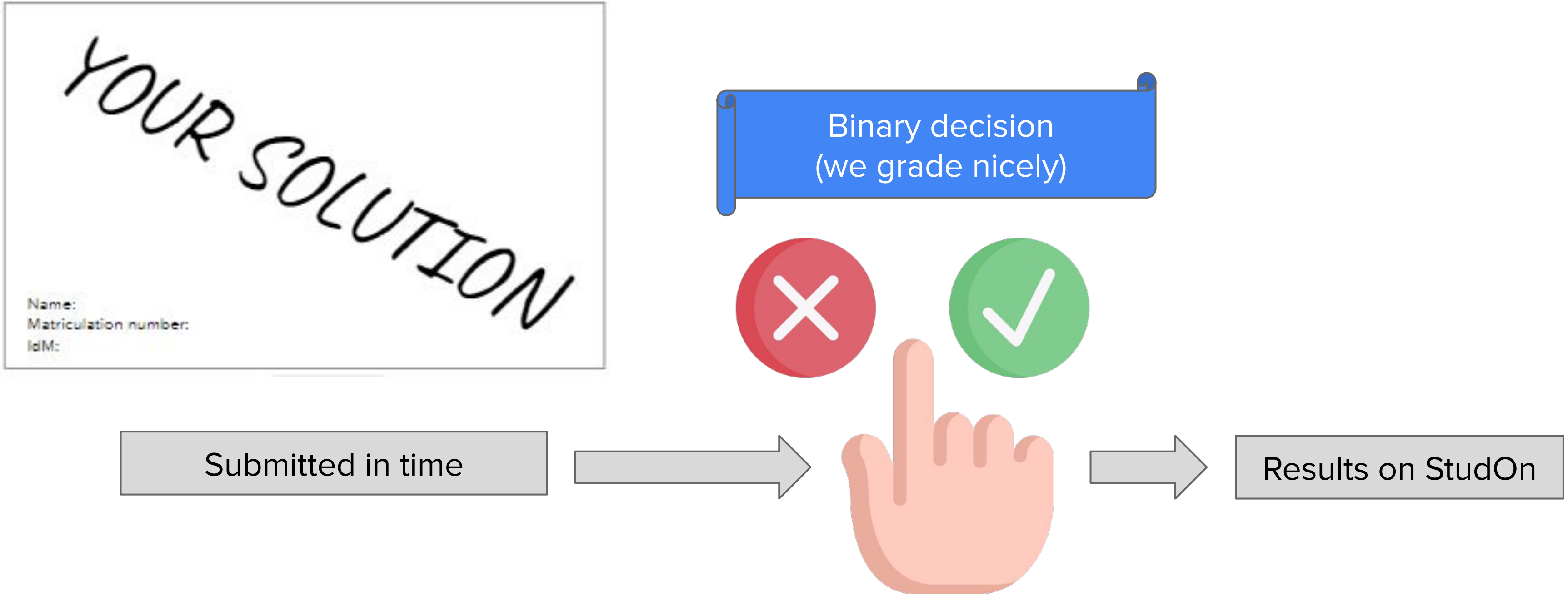
Name:
Matriculation number:
IdM:

STUDON

A passionate DSSS student

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

Homework grading



YOUR SOLUTION

Name:
Matriculation number:
IdM:

Submitted in time

Binary decision
(we grade nicely)

Results on StudOn

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

No E-Mails!
(or it is a highly personal matter)

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!

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

Lecture recordings

We are recording this lecture together with RRZE automatically.
This is **an additional service** for you and not guaranteed (that is happens and that the quality is sufficient for understanding and learning).

Recordings from older semesters can be used but content may likely have changed!

Data Science

We live in a world of data

1900s

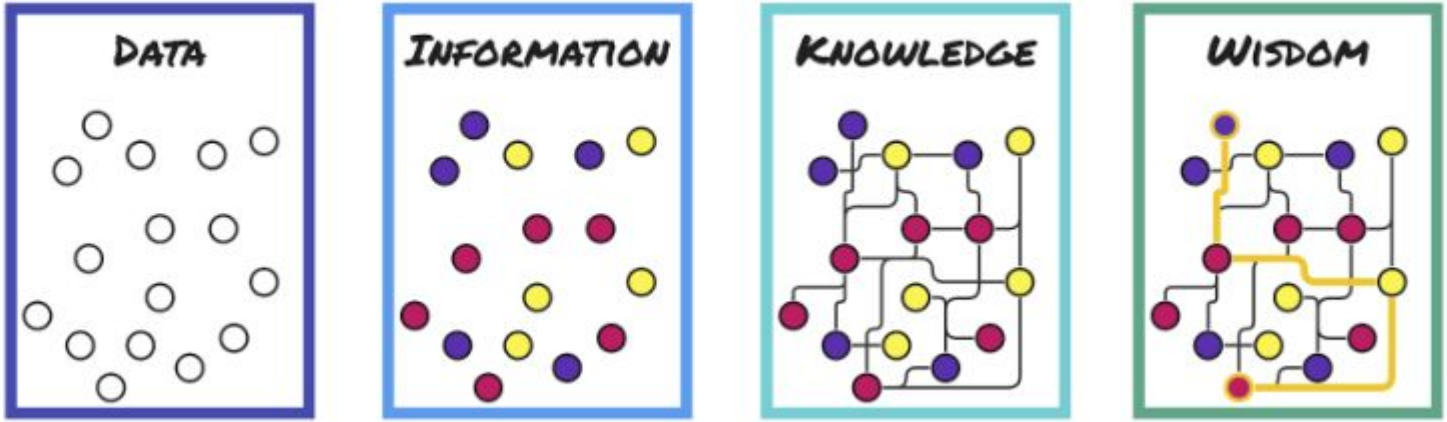
2020s

Data itself is nothing, context and interpretation is everything

Dickinson Animal Hospital

Ersteller: ERIK S. LESSER | Credit: EPA

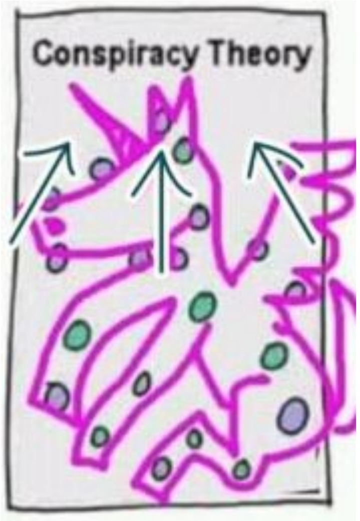
Data is only a starting point



WHO?
WHAT?
WHEN?
WHERE?

HOW?

WHY?

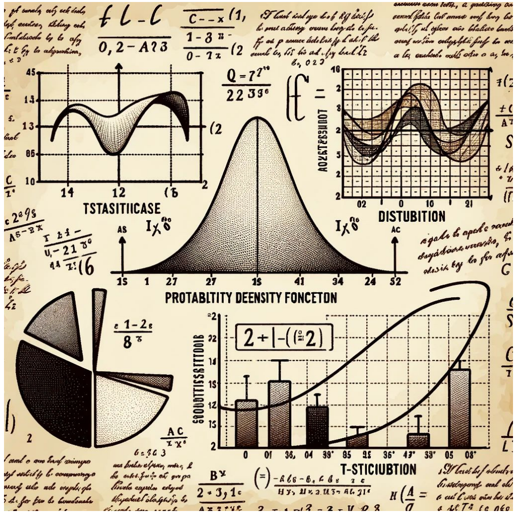


https://www.linkedin.com/posts/aksel_dikw-theory-conspiracy-activity-6701900565351407616-aPjg/

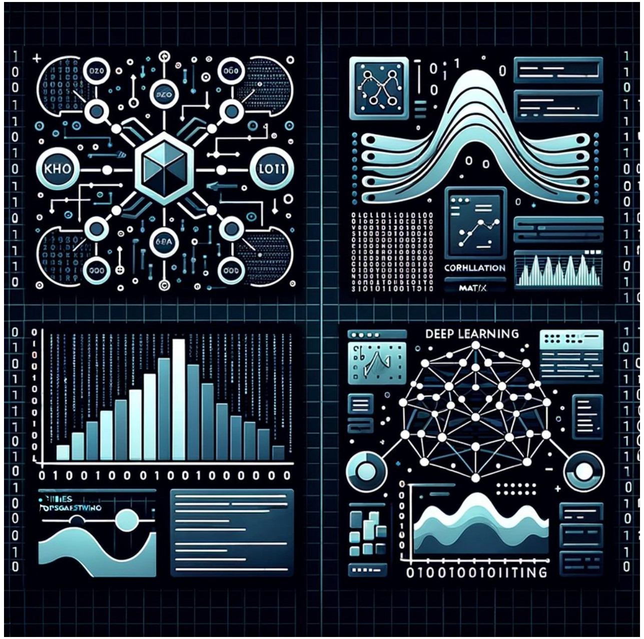
https://www.verses.ai/blogs/what-is-genius

Why do we need "data science"?

Statistics



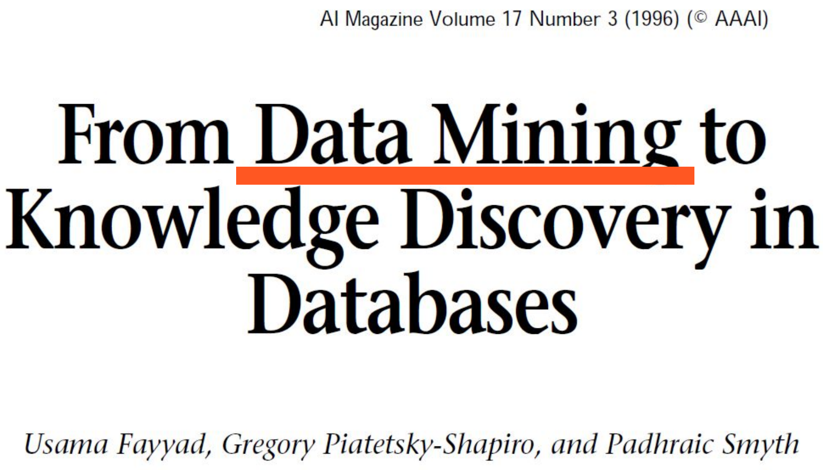
- Likelihood, Probabilities
- PDFs
- Descriptive statistics
- Explorative statistics



What we can't do with "classical" statistics alone:

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

A bit of history



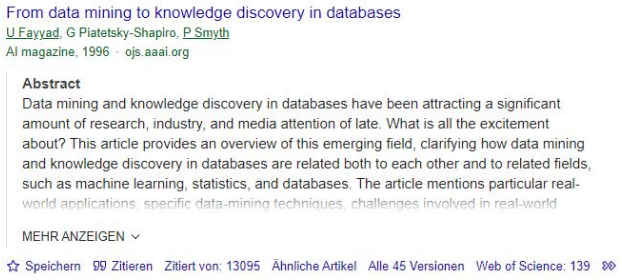
AI Magazine Volume 17 Number 3 (1996) (© AAAI)

Articles

From Data Mining to Knowledge Discovery in Databases

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.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth



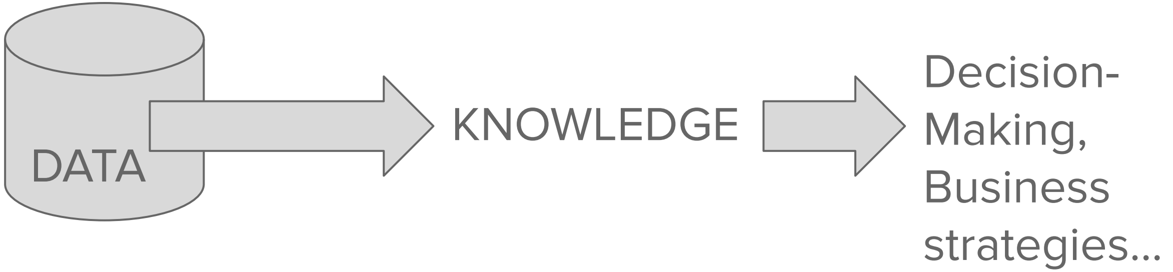
From data mining to knowledge discovery in databases

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

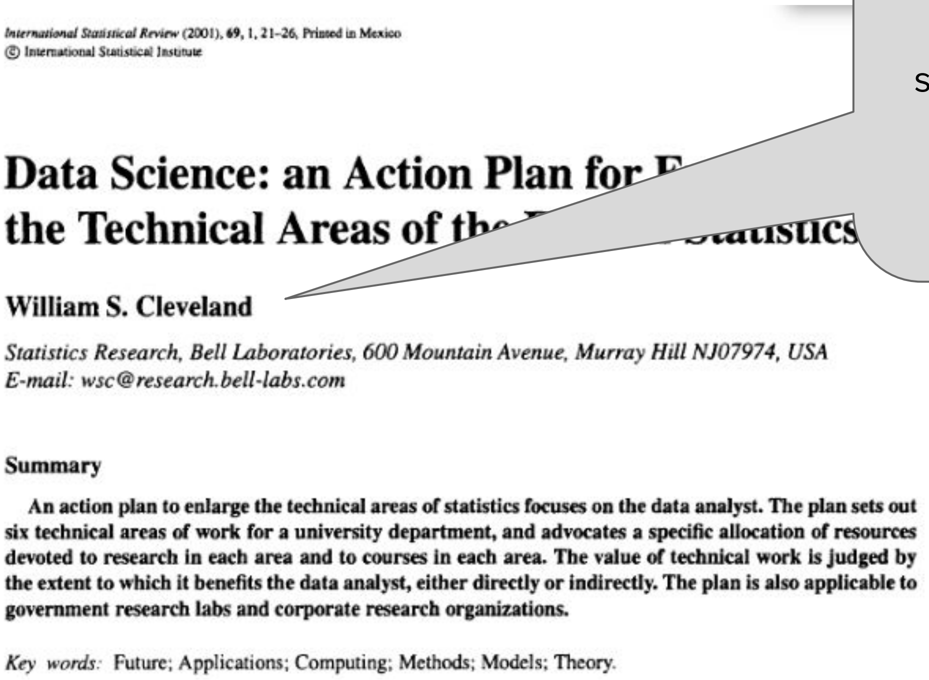


DATA

KNOWLEDGE

Decision-Making, Business strategies...

Coining the word “data science”



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

Data Science: an Action Plan for E
the Technical Areas of th
tatistics

William S. Cleveland

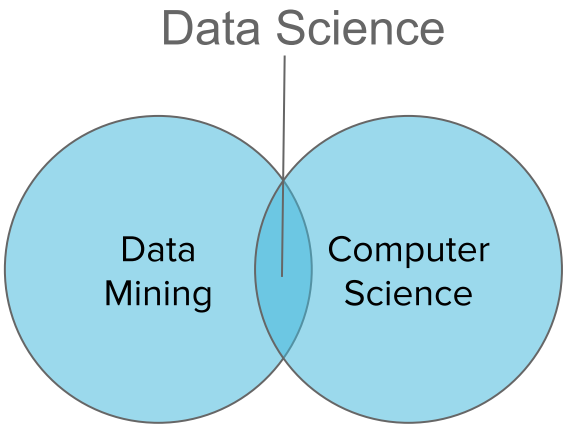
Statistics Research, Bell Laboratories, 600 Mountain Avenue, Murray Hill NJ07974, USA
E-mail: wsc@research.bell-labs.com

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.

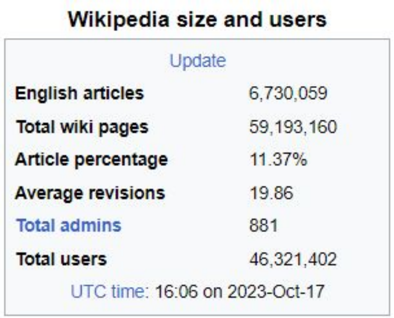


What changed?

Read only
“I am online”

Only consuming

WHEN I WAS A KID, THERE WERE NO PHONES OR TABLETS. WE READ CEREAL BOXES AT BREAKFAST



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+Write
“I am **contributing**”

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

Let's define the job of data science.

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

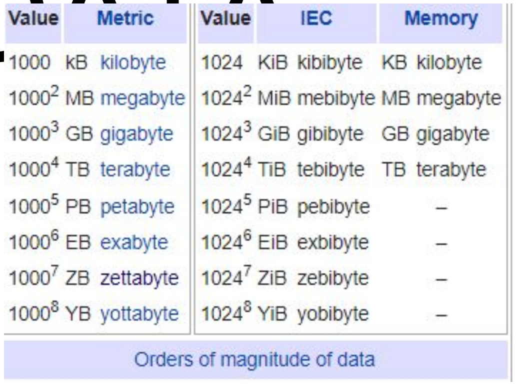
Messy, unstructured, maybe totally irrelevant data

Taking messy data and creating/gaining insights

Takeaways, relevant variables, biomarkers, …

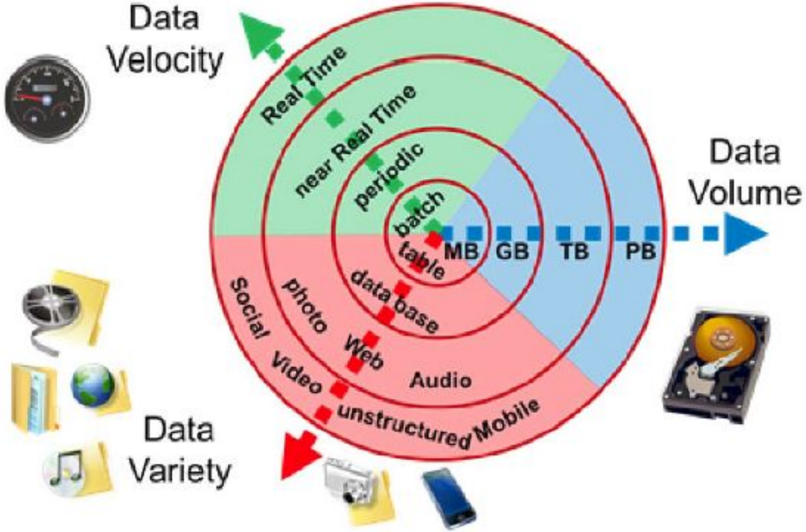
How large is data?

BIG DATA



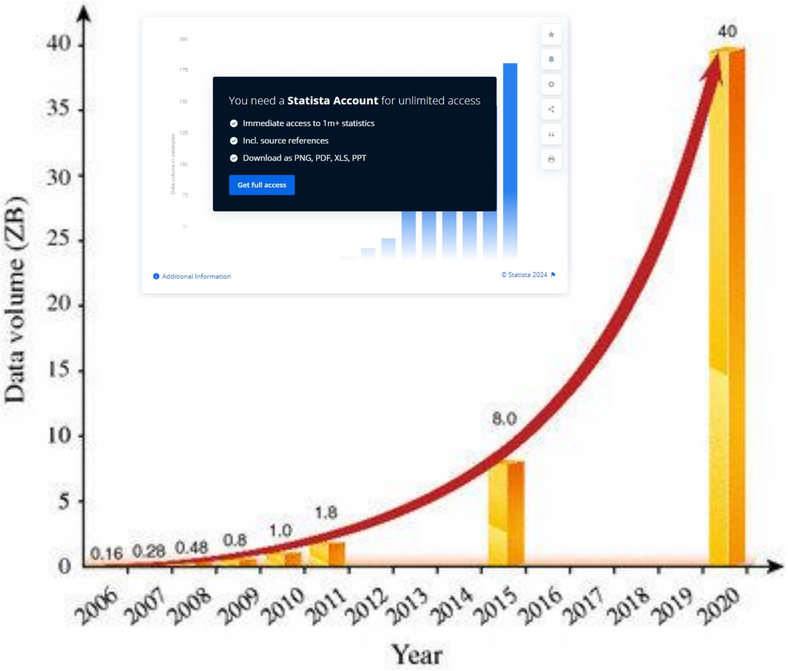
Value	Metric		Value	IEC		Memory	
1000	kB	kilobyte	1024	KiB	kibibyte	KB	kilobyte
1000^2	MB	megabyte	1024^2	MiB	mebibyte	MB	megabyte
1000^3	GB	gigabyte	1024^3	GiB	gibibyte	GB	gigabyte
1000^4	TB	terabyte	1024^4	TiB	tebibyte	TB	terabyte
1000^5	PB	petabyte	1024^5	PiB	pebibyte	–	
1000^6	EB	exabyte	1024^6	EiB	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?

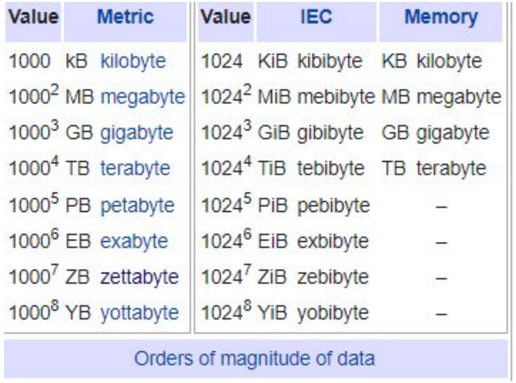


You need a **Statista Account** for unlimited access

- Immediate access to 1m+ statistics
- Incl. source references
- Download as PNG, PDF, XLS, PPT

Get full access

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”)



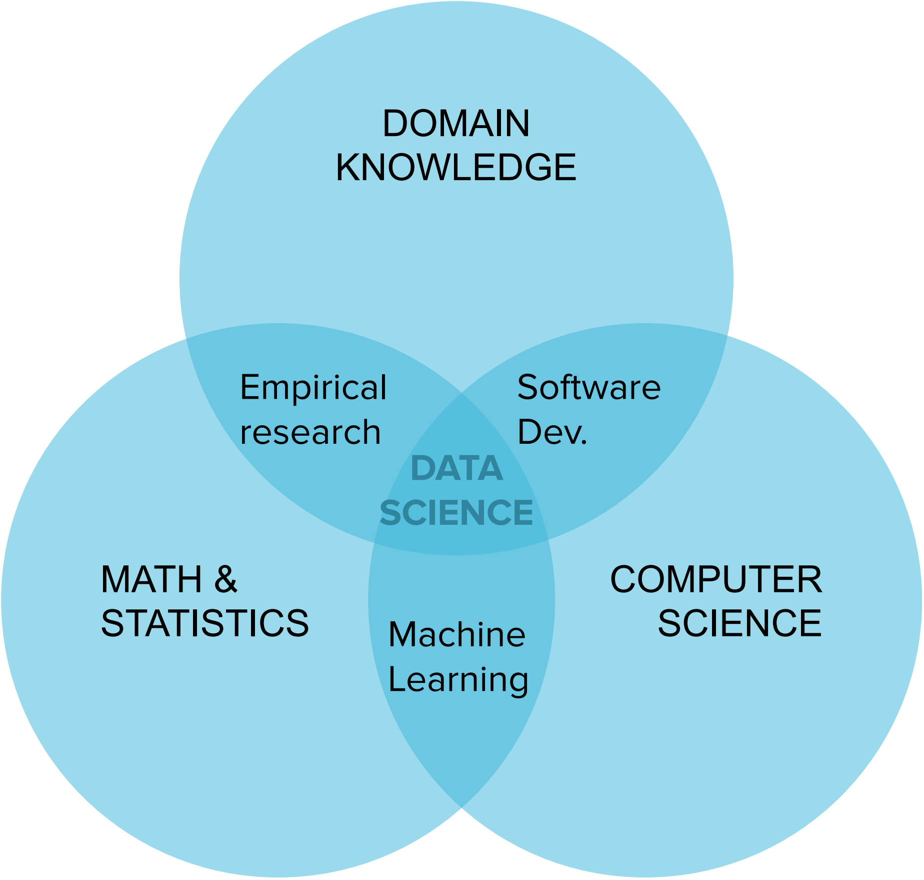
Value	Metric		Value	IEC		Memory	
1000	kB	kilobyte	1024	KiB	kibibyte	KB	kilobyte
1000^2	MB	megabyte	1024^2	MiB	mebibyte	MB	megabyte
1000^3	GB	gigabyte	1024^3	GiB	gibibyte	GB	gigabyte
1000^4	TB	terabyte	1024^4	TiB	tebibyte	TB	terabyte
1000^5	PB	petabyte	1024^5	PiB	pebibyte	–	
1000^6	EB	exabyte	1024^6	EiB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

Orders of magnitude of data

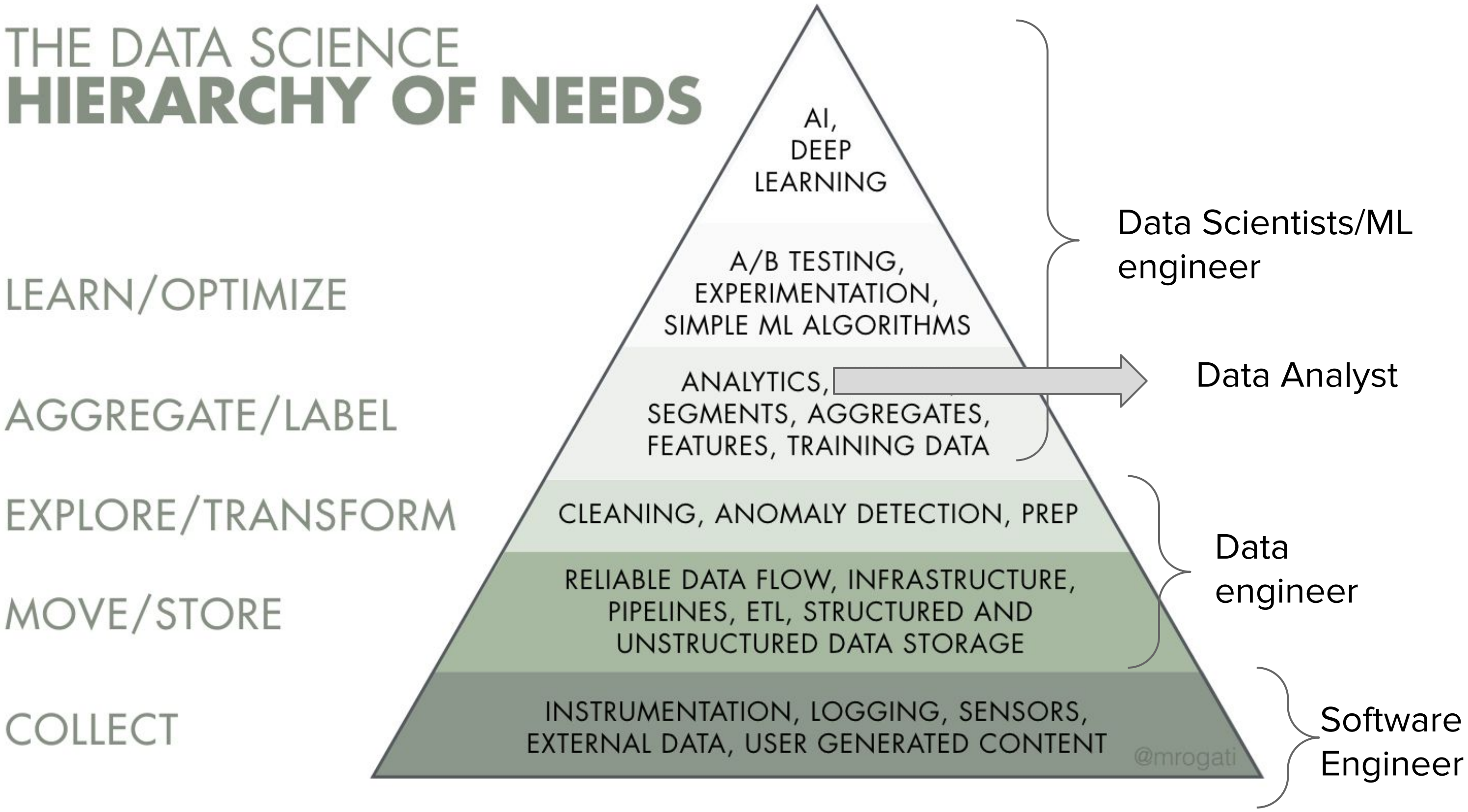
Exponential growth of data!

2021: we estimate 180 ZB in 2025

What is “Data Science”?



Data Science Pyramid of Needs



THE DATA SCIENCE
HIERARCHY OF NEEDS

LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

AI,
DEEP
LEARNING

A/B TESTING,
EXPERIMENTATION,
SIMPLE ML ALGORITHMS

ANALYTICS,
SEGMENTS, AGGREGATES,
FEATURES, TRAINING DATA

CLEANING, ANOMALY DETECTION, PREP

RELIABLE DATA FLOW, INFRASTRUCTURE,
PIPELINES, ETL, STRUCTURED AND
UNSTRUCTURED DATA STORAGE

INSTRUMENTATION, LOGGING, SENSORS,
EXTERNAL DATA, USER GENERATED CONTENT

@mrogati

Data Scientists/ML engineer

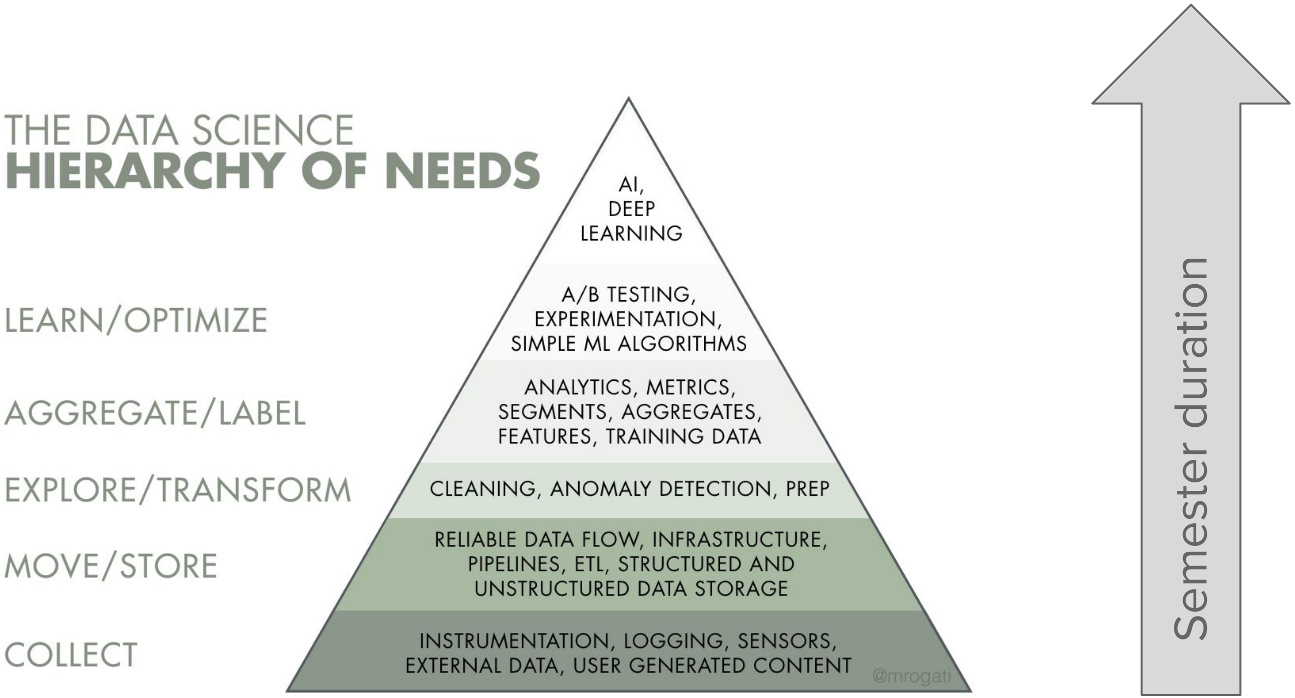
Data Analyst

Data engineer

Software Engineer

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

What we are tackling in this course



THE DATA SCIENCE
HIERARCHY OF NEEDS

LEARN/OPTIMIZE
AGGREGATE/LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
AI, DEEP LEARNING
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
@mrogati
Semester duration

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.

  We have an exercise on Wednesday as a small recap for everyone.

No (!) bonus points for this task