Introduction

Programmierkurs 2 Data Science WS24/25

Leonard Traeger
M. Sc. Information Systems
leonard.traeger@fh-dortmund.de

Programmierkurs 2 Data Science

Dozent:	Leonard Traeger
Email:	*leonard.traeger@fh-dortmund.de*
Vorlesung und Praktikum:	Montags um 12-13:30 u. 14:15-15:50 Uhr; Vorlesungen werden nicht aufgezeichnet
Kurswebsite:	***http://leotraeg.github.io/me/I9PB-43021.html***
GitHub (Dateien)**:**	*https://github.com/leotraeg/FHDTM-P2DS-WS2324*
Ilias (Umfragen, Artefakte, QA):	Magazin > FB Informatik > Lehrveranstaltungen > Informatik > Bachelor - Module > (...) Vertiefungsrichtung - Data Science > Programmierkurs > (INPB-43021) Programmierkurs 2 - Data Science *https://www.ilias.fh-dortmund.de/ilias/goto_ilias-fhdo_crs_1334419.html*
Sprechstunde:	Online und über Kurswebsite buchbar
Raum:	C.2.32 (in der Regel **P**räsenzlehre) oder alternativ (**O**nline); Link siehe Kurswebsite

Some notes on language matters...

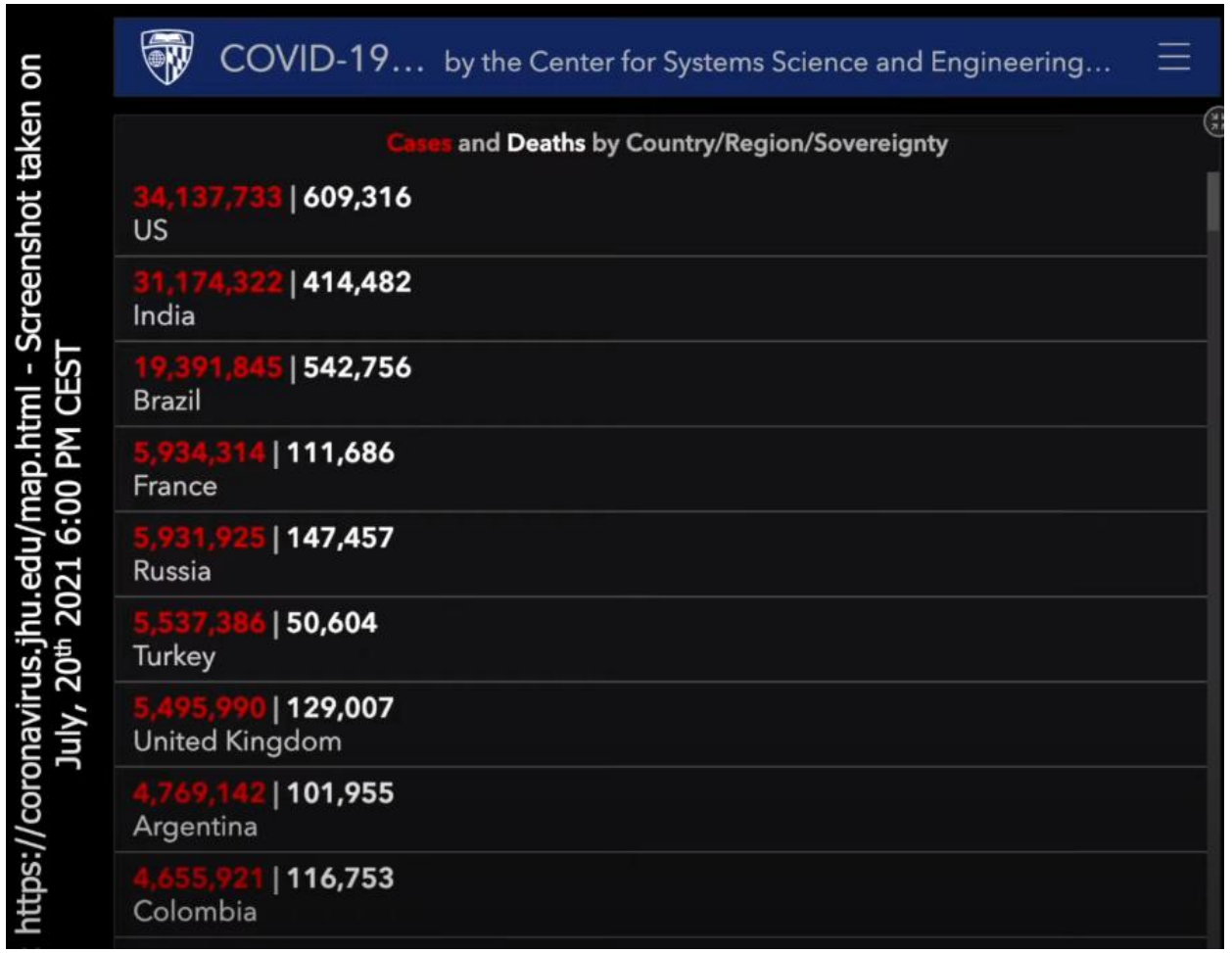
Most of the content of this course will be in English:

- The **slides** of this course will be in English.
- The **textbook** we will use, is freely available in English.
- **Additional materials** and **referenced web resources** are in English.

But:

- The **assignments** will be in German.
- The **lecture itself** will be (mostly) in German!
- You can still **answer** the **questions** in the assignments in **German**.

Historical Moment of Data Science

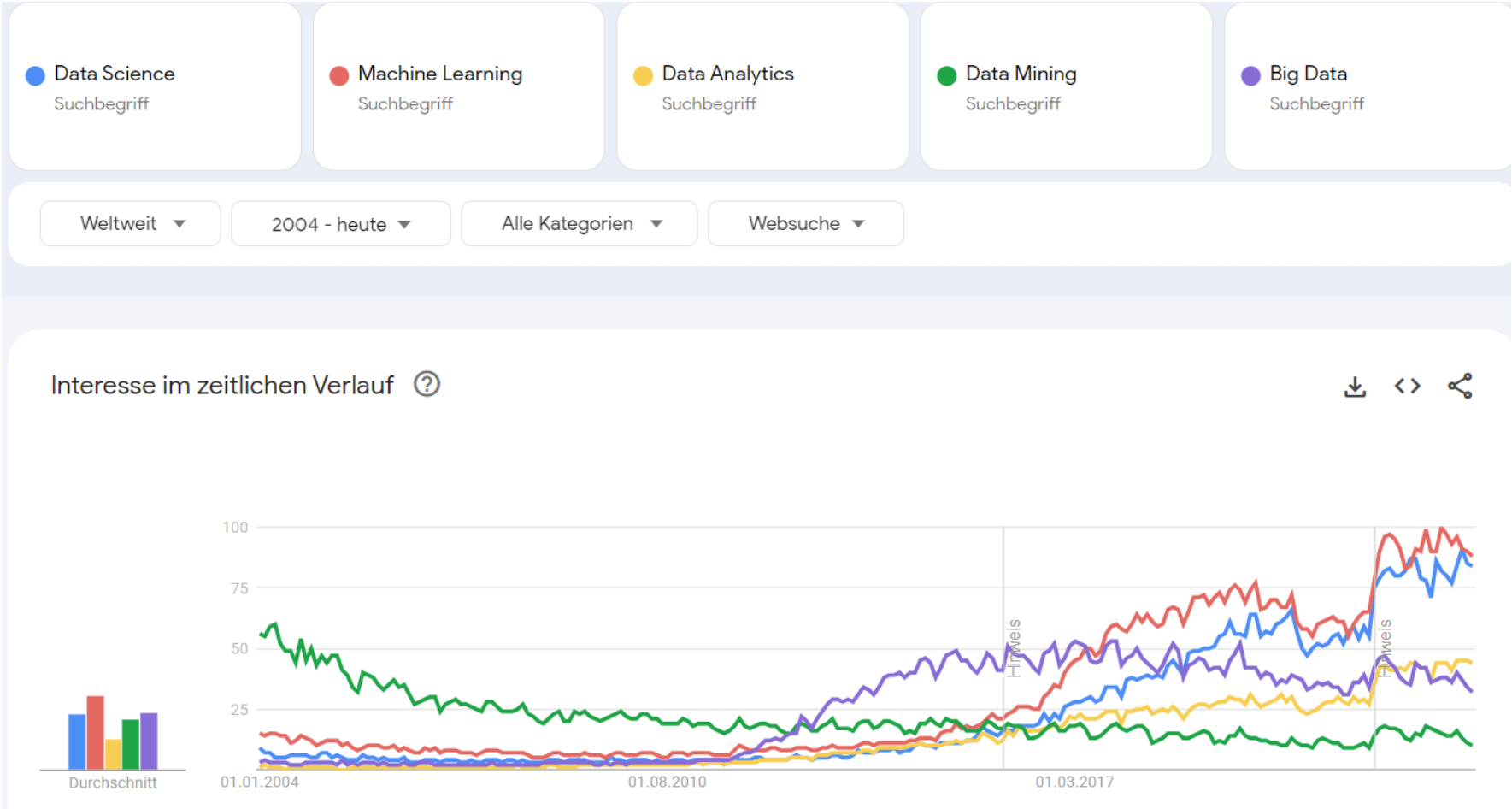


Overall Learning Goals

By the end of this course you will be able to:

- **Discuss** Data Science and its current trends.
- **Explain** the fundamentals of typical data science applications.
- For a variety of data science life cycle frameworks, be able to **explain, compare** and **contrast**, and **discuss** ethics, limitations, and applicability.
- **Apply** Data Science techniques in **Python** to solve real problems.

Meta Data Science



https://trends.google.com/trends/explore?date=all&q=Data%20Science,Machine%20Learning,Data%20Analytics,Data%20Mining,Big%20Data&hl=de

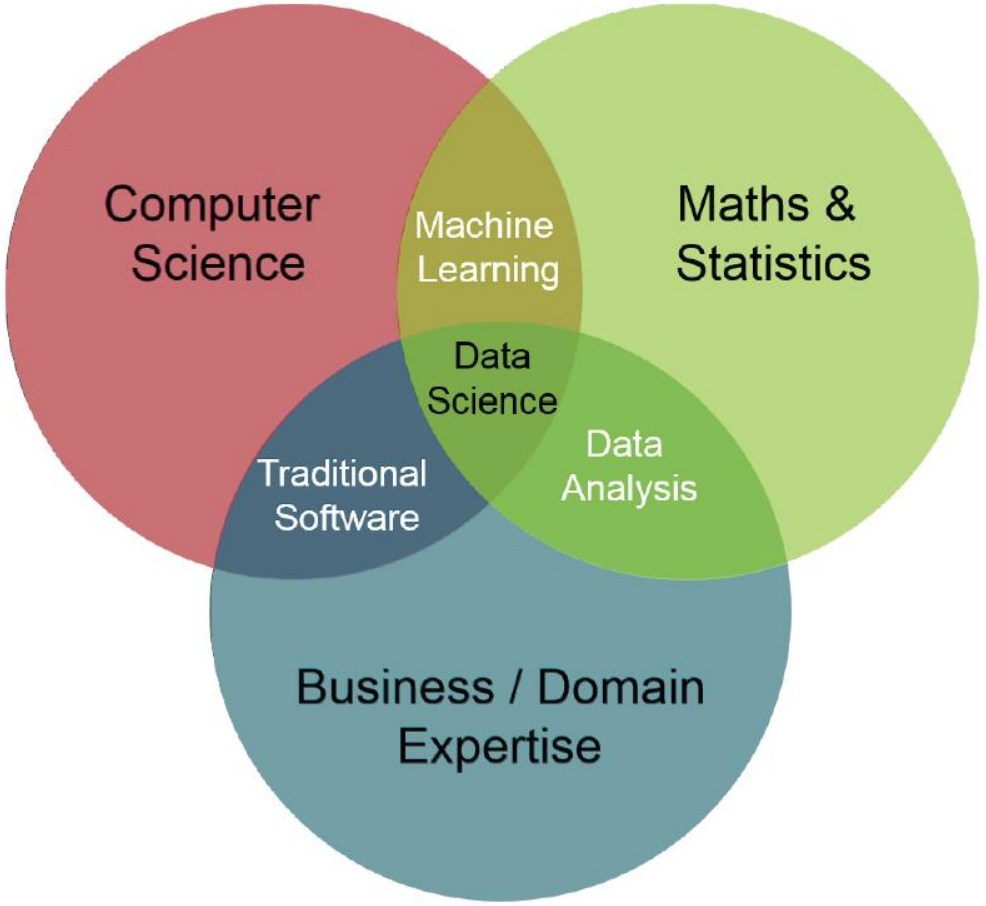
Data Science

"Data Science beschäftigt sich mit einer **zweckorientierten Datenanalyse** und der **systematischen Generierung** von **Entscheidungshilfen** und -grundlagen, um **Wettbewerbsvorteile** erzielen zu können."

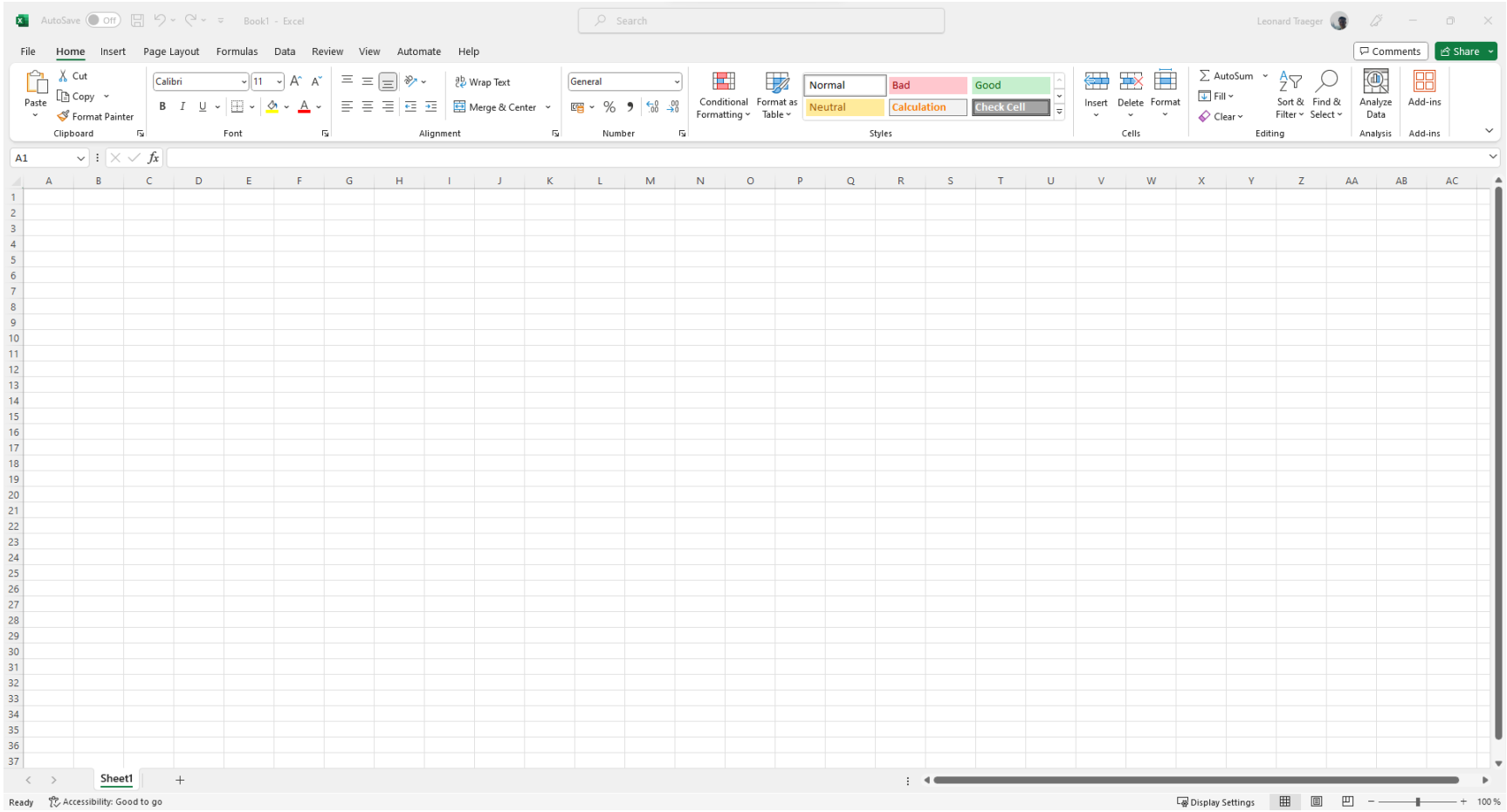
"In der Wissenschaft beschäftigt sich Data Science mit **unterschiedlichen Bereichen** und kann daher verschiedene akademische Hintergründe haben: *Informatik, Statistik, Mathematik, Natur- oder Wirtschaftswissenschaften, Machine Learnings, des statistischen Lernens, der Programmierung, der Datentechnik, der Mustererkennung, der Prognostik, der Modellierung von Unsicherheiten und der Datenlagerung*."

https://gi.de/themen/beitrag/data-literacy-und-data-science-education-digitale-kompetenzen-in-der-hochschulausbildung

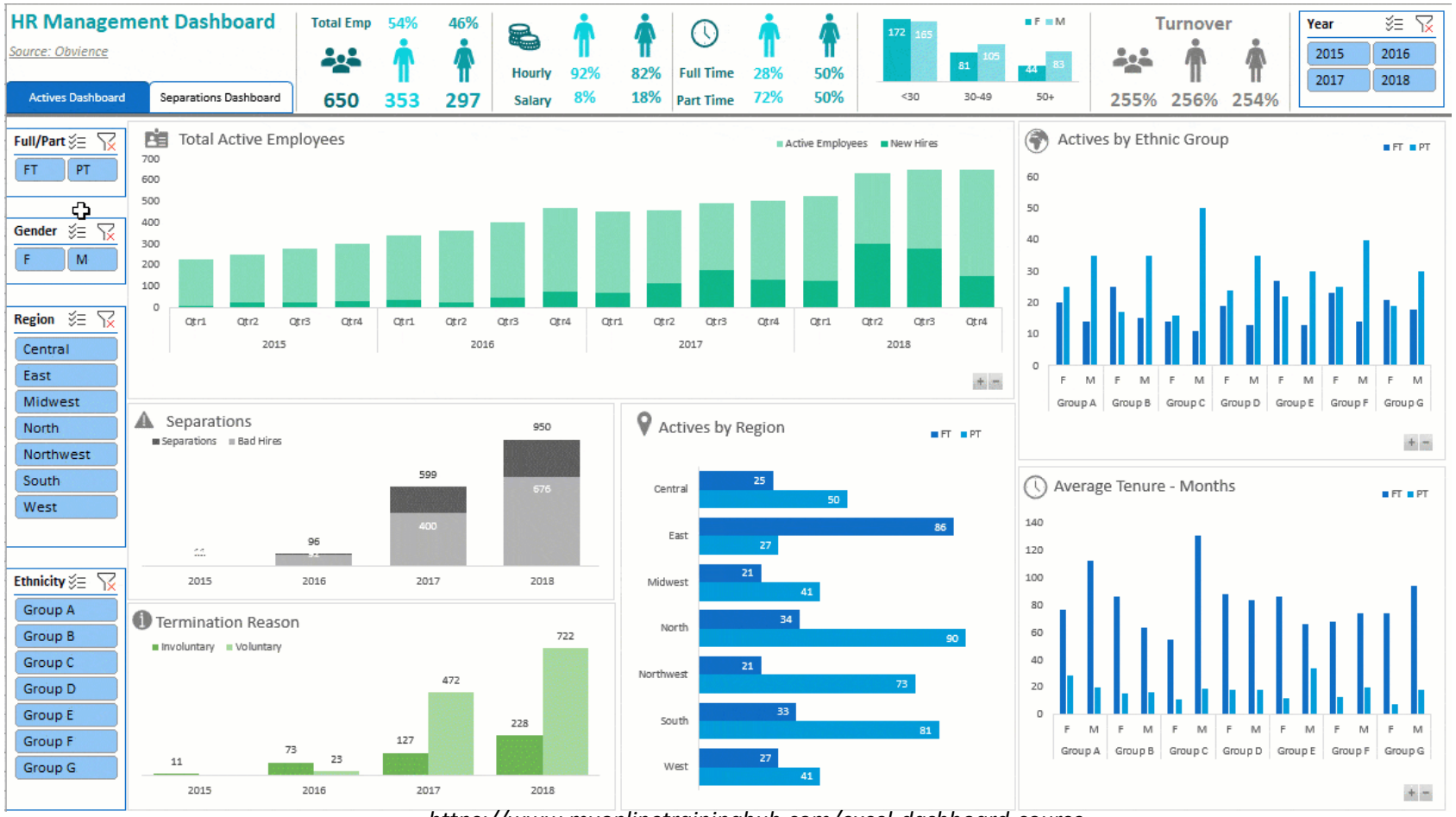
What is Data Science?



Easy...?


AutoSave
Book1 - Excel
Search
Leonard Traeger
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Automate
Help
Comments
Share
Paste
Cut
Copy
Format Painter
Clipboard
Calibri
11
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Normal
Bad
Good
Neutral
Calculation
Check Cell
Styles
Insert
Delete
Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing
Analyze Data
Analysis
Add-ins
A1
Sheet1
Ready
Accessibility: Good to go
Display Settings
100 %

Wow…but isn‘t there much more?



https://www.myonlinetraininghub.com/excel-dashboard-course

A few current Data Science problems

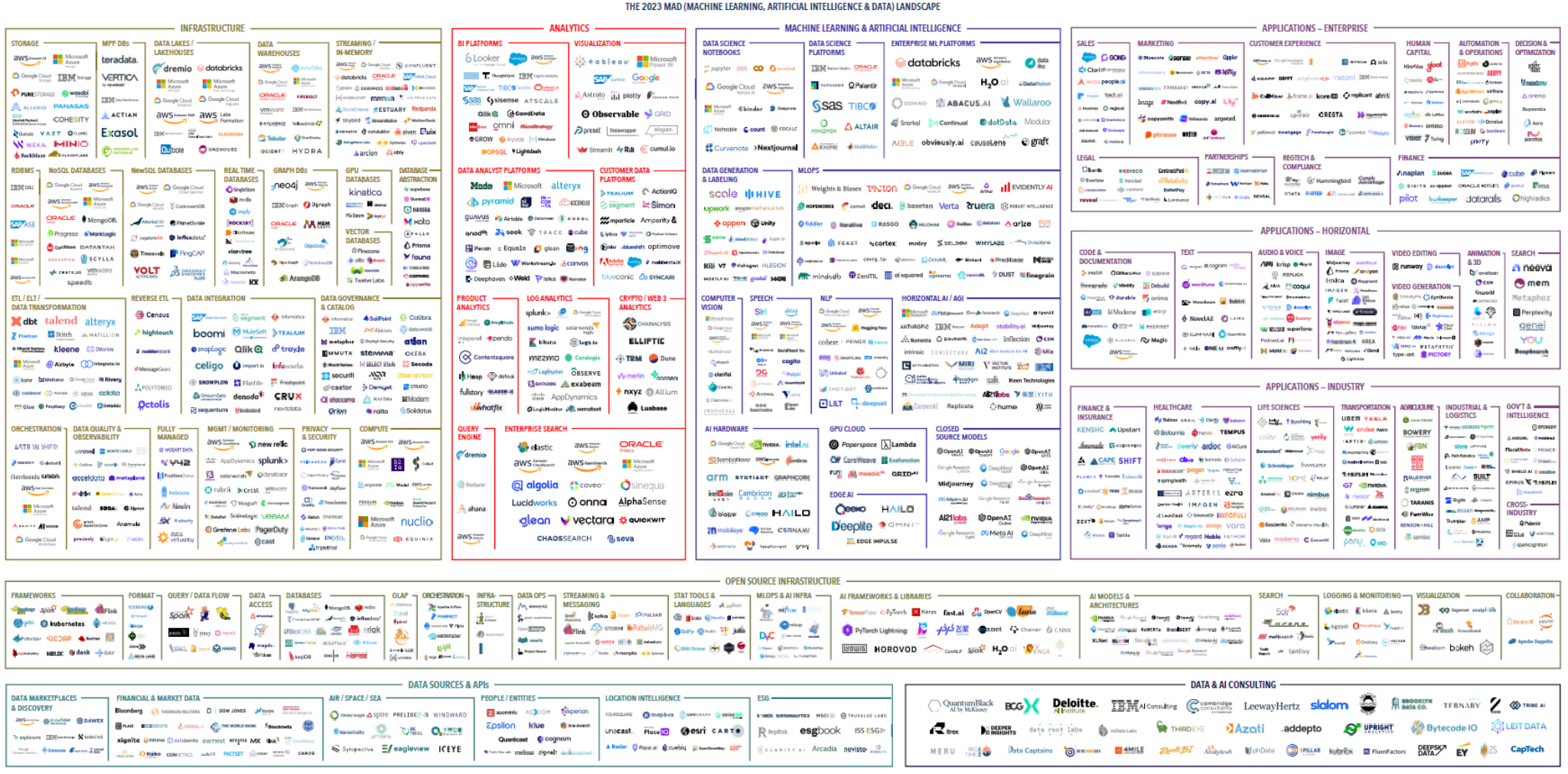
Data Discovery

Natural Data Interfaces

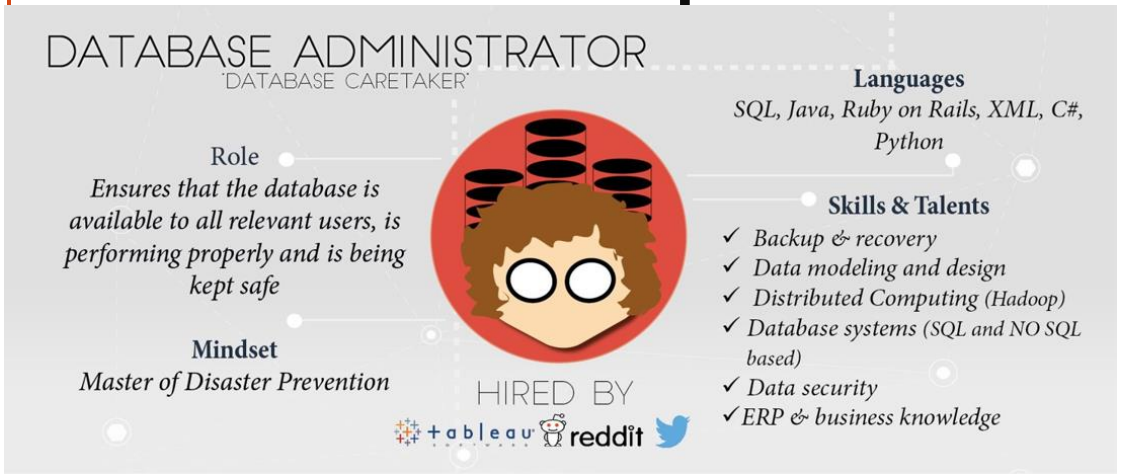
Data Integration

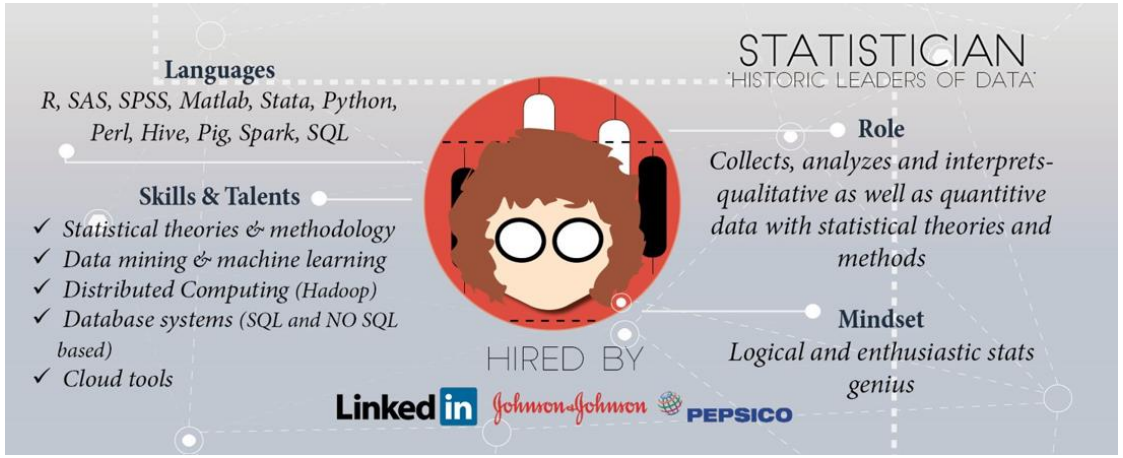
Data Quality and Trust

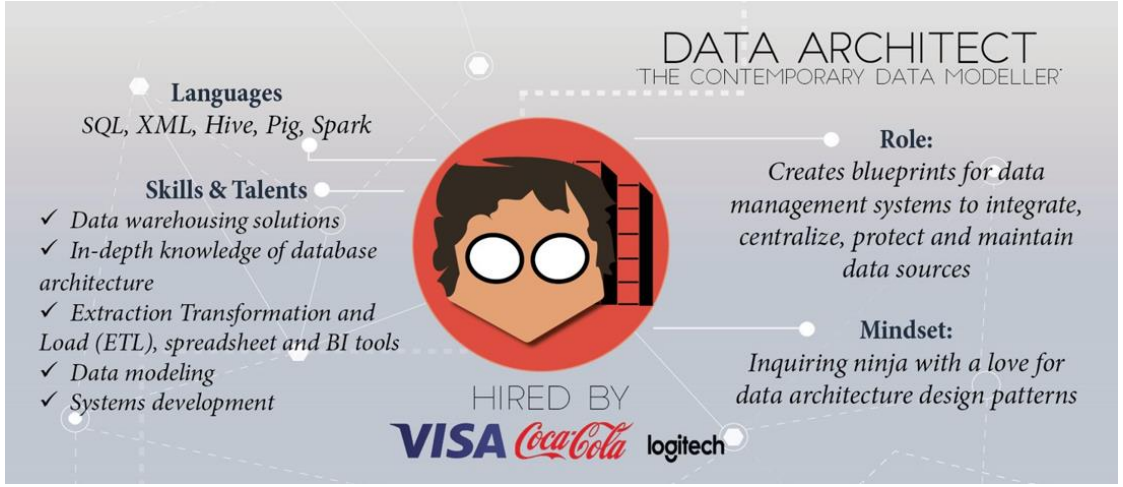
World of Data Science

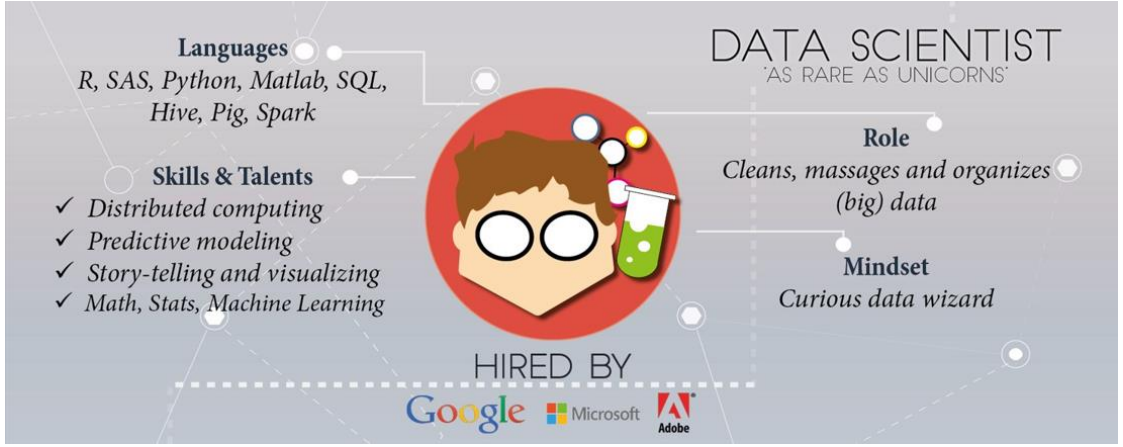


Skills and Experience > Titles and Labels









https://www.datacamp.com/community/tutorials/data-science-industry-infographic

...try not to get lost ☺

- There will always be something you haven't heard of before.
- Research concepts before using them.
- Be curious about new topics.

- Use glossaries and read documentations in the beginning!

https://swcarpentry.github.io/python-novice-inflammation/reference.html#glossary

Harvard Business Review

ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE 13 COMMENT HH TEXT SIZE PRINT $8,95 BUY COPIES

WHAT TO READ NEXT

Big Data: The Management Revolution

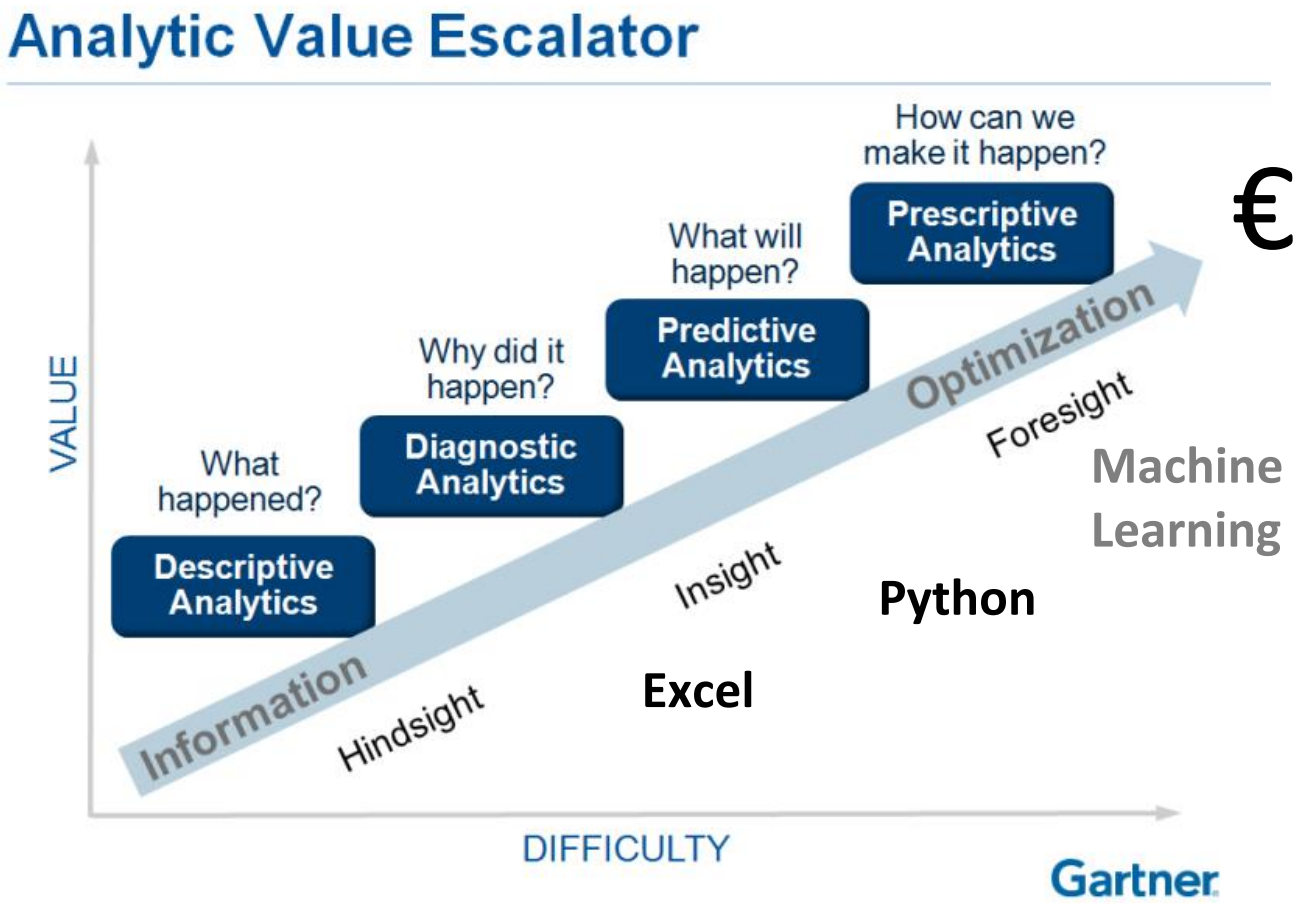
https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Why Learn Data Science?

- **Explore**: identify patterns.
- **Predict**: make informed guesses.
- **Infer**: quantify what you know.

Motivation

- Gain new knowledge
- Help people
- Employment



Data Science is more than Math and CS

Human interaction - “**The best data scientists get out and talk to people**”:

- Discovering stakeholders.
- Negotiating with data owners.
- Customer engagement.

https://hbr.org/2017/01/the-best-data-scientists-get-out-and-talk-to-people

Iterative and **cross-disciplinary** process

- As a data scientist, you'll often be **work**ing for **someone other** than yourself.
- Expect **under-specified requirements** from customers.
- Provide incomplete solutions (**Minimum Viable Product**) rather than waiting until the product is perfect.

https://wirtschaftslexikon.gabler.de/definition/minimum-viable-product-mvp-119157

Literature

- VanderPlas, J., "Python Data Science Handbook", O'Reilly, 2017
  Digital free copy: https://jakevdp.github.io/PythonDataScienceHandbook/
- Fabio Nelli, "Python Data Analytics With Pandas, NumPy, and Matplotlib" (2nd edition), Apress (Springer), 2018
  Digital free copy via FH VPN
- Wickham, H. und Grolemund, G., "R für Data Science", Heidelberg, O'Reilly, 2017

Grading

Notenzusammensetzung; Änderungen vorbehalten

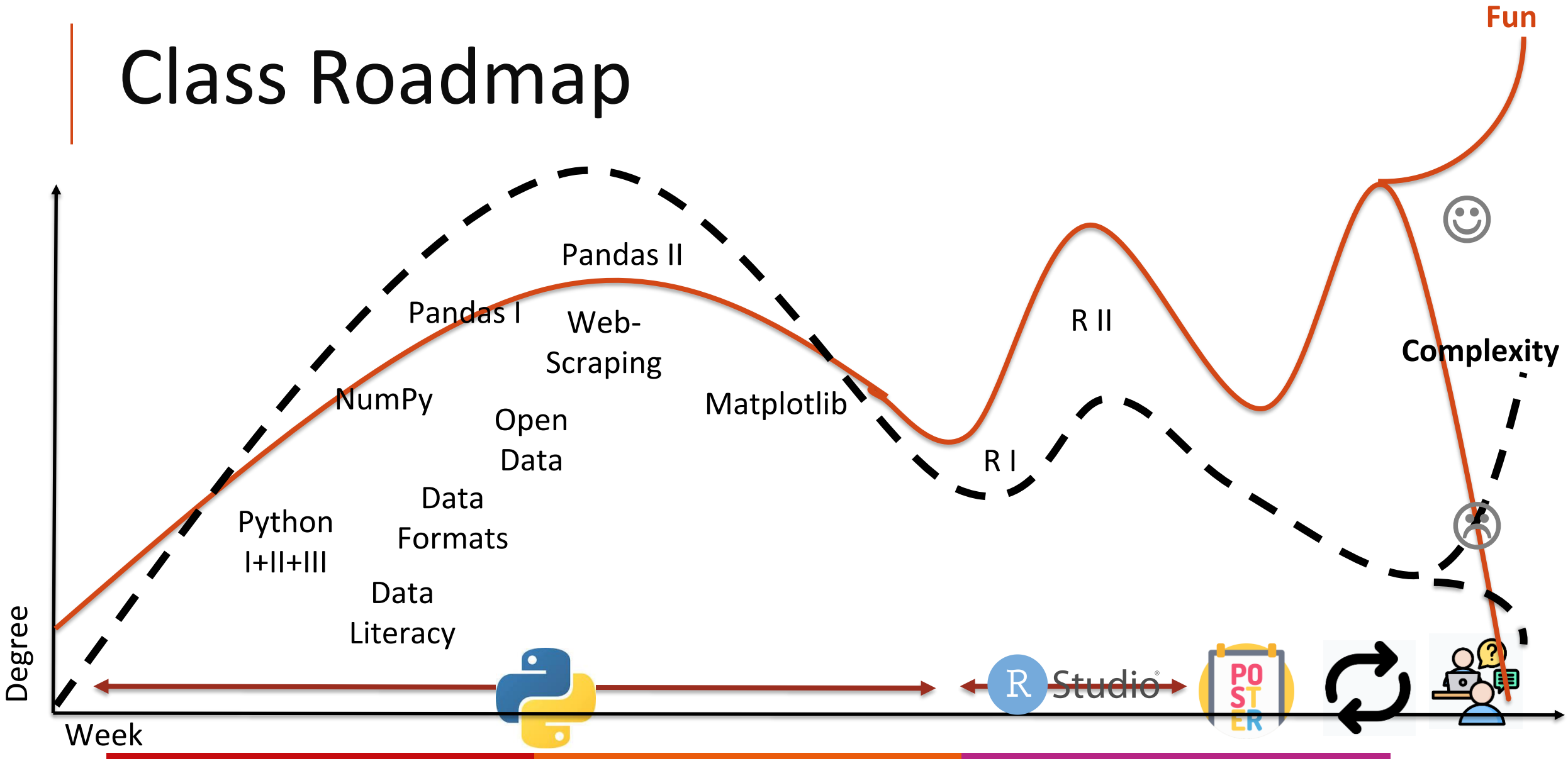
Artefakt	Max. Punkte
Ilias Forum Beitrag oder Kommentar	0,66%
Praktikum I	8%
Praktikum II	8%
Projekt Meilenstein I	5%
Projekt Meilenstein II	10%
Projekt Meilenstein III.1	35%
Mündliche Prüfung über Vorlesungsinhalte und das Projekt Meilenstein III.2	50%

Skala; Änderungen vorbehalten

Punkte	Note
116,66 - 94,9 %	1,0
<94,9 - 89,5 %	1,3
<89,5 - 84,3 %	1,7
<84,3 - 79,0 %	2,0
<79,0 - 73,7 %	2,3
<73,7 - 68,2 %	2,7
<68,2 - 63,1 %	3,0
<63,1 - 57,9 %	3,3
<57,9 - 52,6 %	3,7
<52,6 - 50,0 %	4,0
< 50,0 %	n.b.

- **Timely** submission of artefacts (lab work or project milestones) through **Ilias**.
- Copying, modifying, rewriting or not following citation rules is unacceptable
  (see falsification, fabrication, plagiarism, ...*www.niu.edu/academic-integrity/students/*).

Class Roadmap



Week 1: Intro + Python I

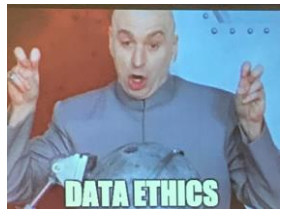
- Introduction Data Science
- Course logistics

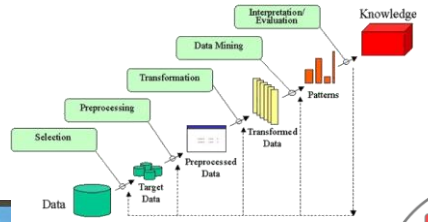
Python I

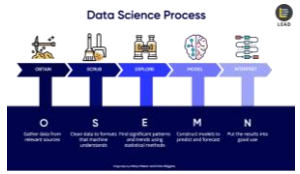
- Python set up
- Jupyter and Colab Notebooks
- Basic Data Types
- Random Numbers
- String methods

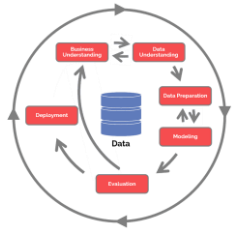
Week 2: Python II

- Data Literacy and Ethics
- Data Science Life Cycle









- Comparison and Logical Operators
- Control Statements, Containers (Lists, Dictionaries, Sets, Tuples)
- Functions
- Functional Programming incl. Map, Filter, Reduce
- List Comprehensions

```
veg = [['lettuce', 'lettuce', 'peppers', 'zucchini'],
       ['lettuce', 'lettuce', 'peppers', 'zucchini'],
       ['lettuce', 'cilantro', 'peppers', 'zucchini']]
```

Week 3: Python III + Data Formats

- Imperative and Declarative Paradigm
- Object-Oriented Programming
  - Constructor
  - Destructor
  - Decorator annotated and regular Class Methods
  - Inheritance
- CSV, JSON, and XML as Common Data Formats

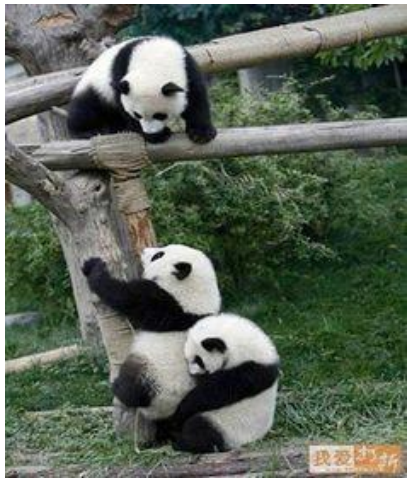
Week 5: Python NumPy + Open Data

- Containers versus NumPy, NumPy Datatypes, Booleans, Comparison
- Indexing / Slicing, Reshape, Copy()
- Vectorization (Ufuncs), SciPy
- Aggregation, Sorting, Broadcasting

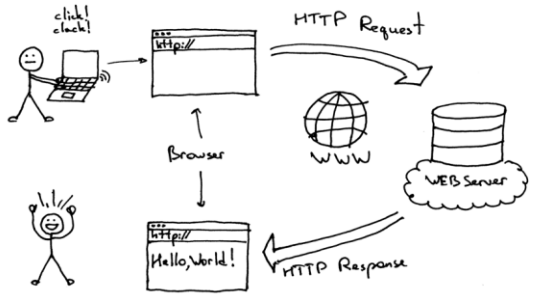
- Open Data and Principles

Week 6: Python Pandas I + Web-Scraping

- Data Series and Frames
- I/O: Read and Parse Different Data Formats
- Viewing Data, Indexing, Data Reduction (Selection and Deletion)
- Data Masking, Viewing Meta Data,

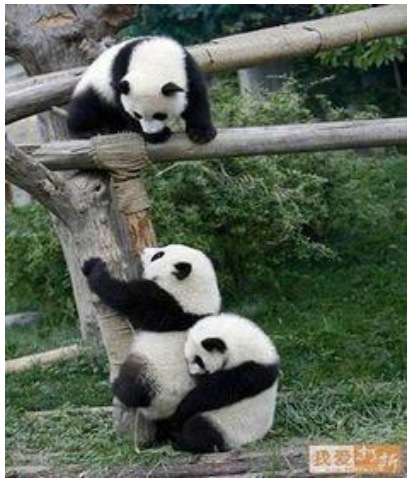


- Web Scraping with BeautifulSoup



Week 7 + 9: Python Pandas II

- NumPy and Pandas
- Data Preprocessing

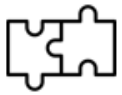


Data Reduction

- Obtains reduced representation in volume but produces the same or similar analytical results.

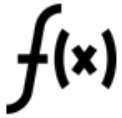
Data Cleaning

- Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies caused by data integration.



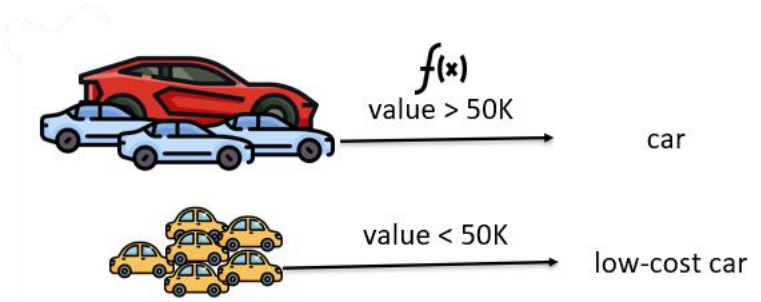
Data Integration

- **Integration of multiple tables**, databases, data cubes, or files.



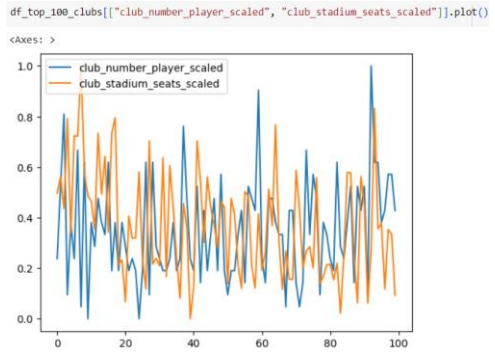
Data Transformation

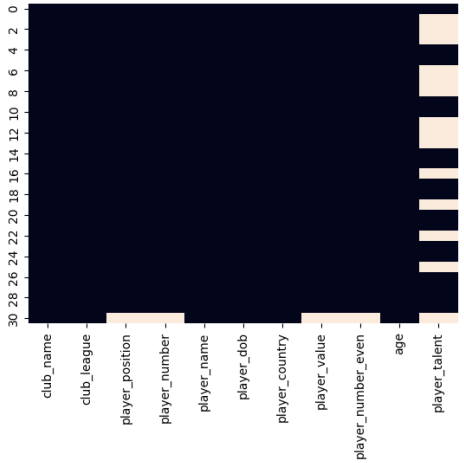
- Aggregation, generalization, normalization and attribute construction.

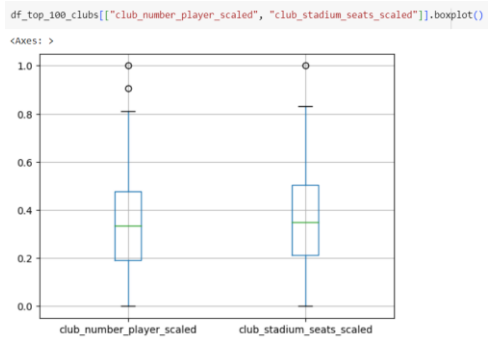


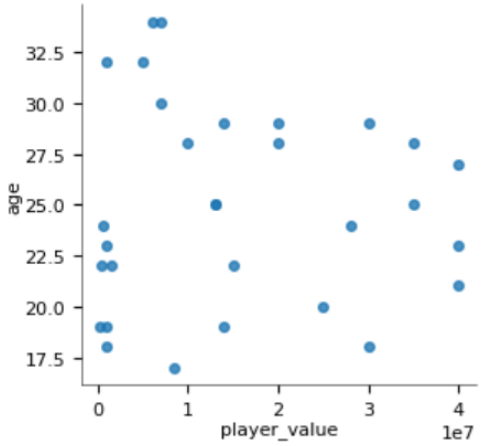
Week 10: Visualization

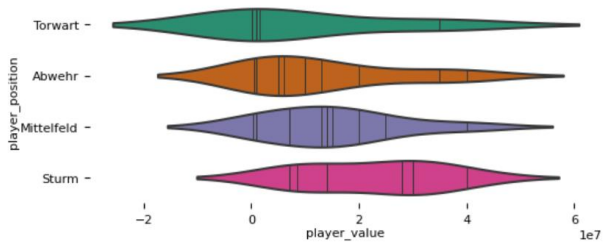
- Simple Plots: Bar, Pizza, Histogram, …
- Text, Annotation, Color
- Data Summary Plots
- Meta Data Plots
- Encoder Decoder Design Guide

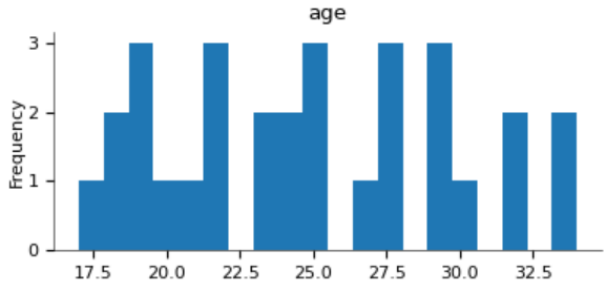








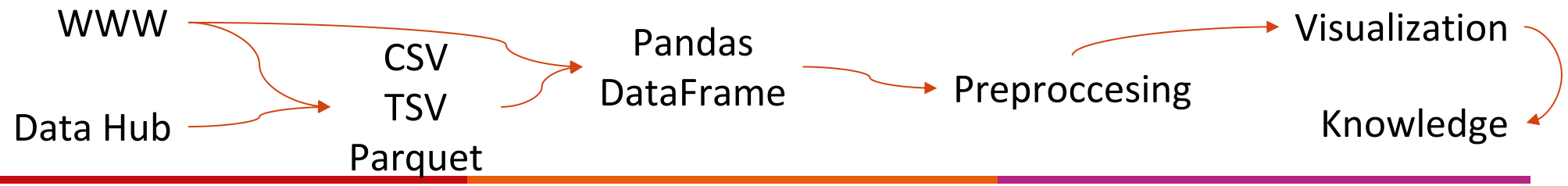




Week 11-12: **R** or **Big Data Issues**

Team Project

- Self-determined teams with **four** students.
- Runs **in parallel** to the entire semester.
- The goal is to carry out a practical data science project based on a team-determined **data set** to tackle some domain-problem**.**
- Core is the **programmatic implementation**.
- **Research** or **extract** a **dataset,** apply **preprocessing** techniques**,** run **analytical queries** and create **visualizations** so you gain **interpretable insights** for your domain problem.
- **Should** be related to your **interest**. **Can** be based on your **work** in **industry** or **science.**



Team Project: where to find data?

Recommended (but not limited to):

- Eine Sammlung von Datensätzen aus einer Vielzahl von Domänen, zu der über 100 Personen beigetragen haben: https://github.com/awesomedata/awesome-public-datasets
- UCI ML Repository: http://archive.ics.uci.edu/ml/
- Kaggle Datasets: https://www.kaggle.com/datasets
- Open Data Europa: https://data.europa.eu/en
- Datenportal für Deutschland: https://www.govdata.de/
- Landesdatenbank NRW: https://www.landesdatenbank.nrw.de/ldbnrw/online
- Open Data Dortmund: https://opendata.dortmund.de/Informationsportal/
- Web-Scraping: mehr dazu in Woche #6 mit einer Live-Demo (auf Anfrage stelle ich gerne die Demo Skripte vorab zur Verfügung).

I can recommend many interesting datasets in (my) research area on **Data Management**:

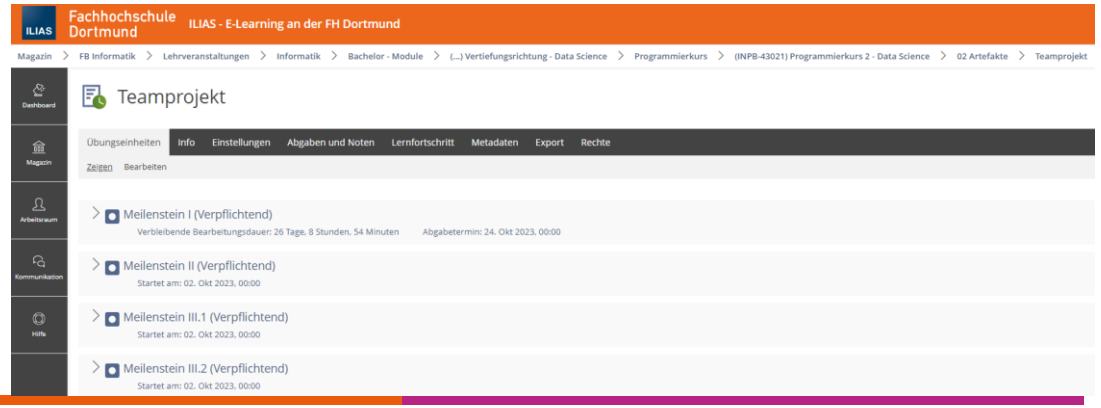
- 70 Million Queries on Snowflake: https://github.com/resource-disaggregation/snowset
- 8 Million Queries on AWS data centres: https://github.com/amazon-science/redset

Team Project: Milestones

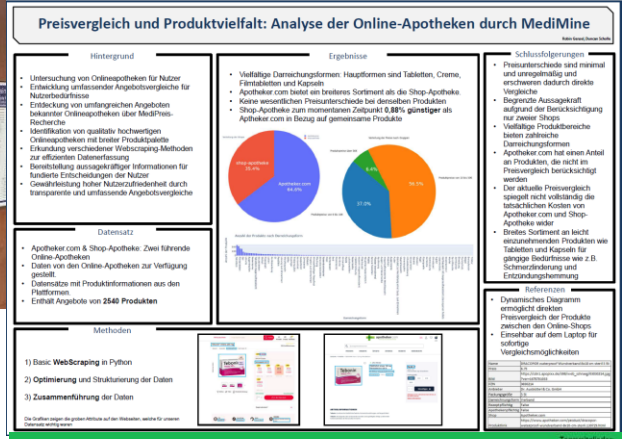
Milestones and artefact deliverables:

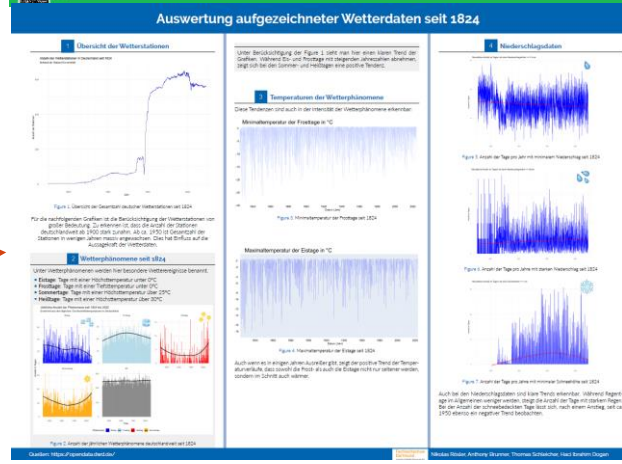
- Teams formed and sent via mail by one team-member by *30.09.24*
- Milestone I (.pdf file) due to 07.*10.24*
- Milestone II (.ipynb as file or link to file and print version) due to *25.11.24*
- Milestone III.1 (.pdf or .pptx) due to *06.01.25 (earlier for print via University)*
- Milestone III.2 (.ipynb as file or link to file and print version) due to *13.01.25*

Submitted only via Ilias.



Week 15: Project Day





First hour: students present their project (Milestone III.1) to each other.

Second hour: graded presentation à 10 minutes for each project.

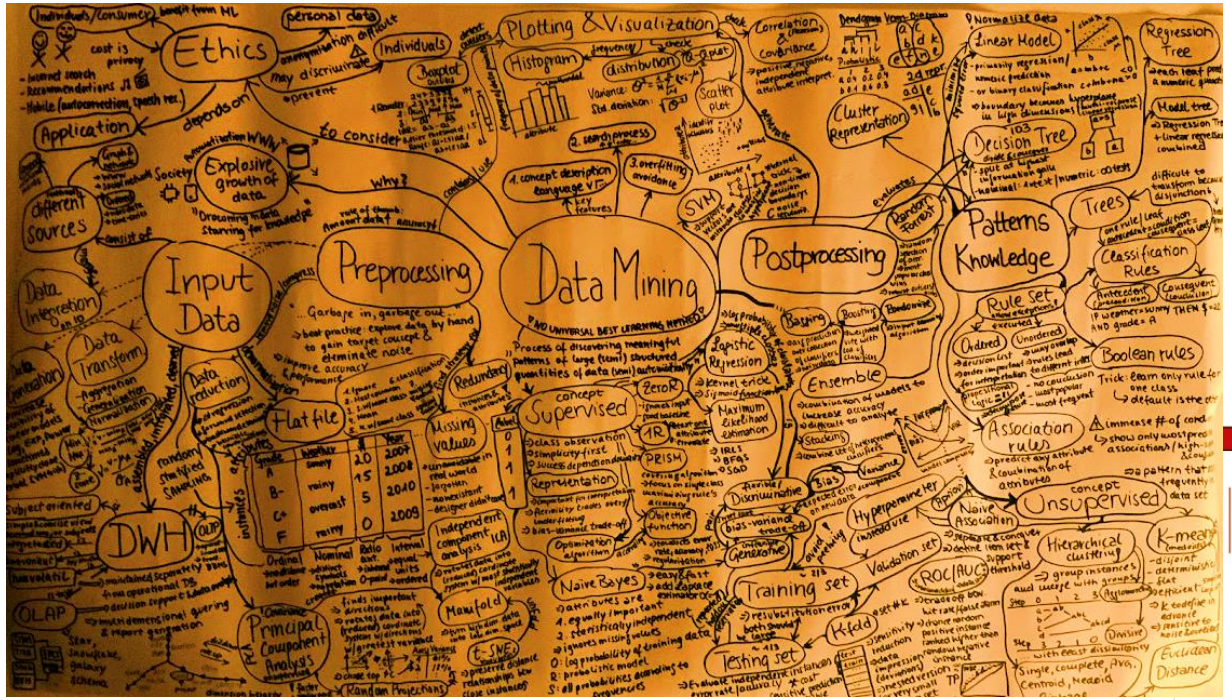
After lunch: feedback session.

project posters from previous years

Guidance:

- guides.nyu.edu/posters or
- colinpurrington.com/tips/poster-design/

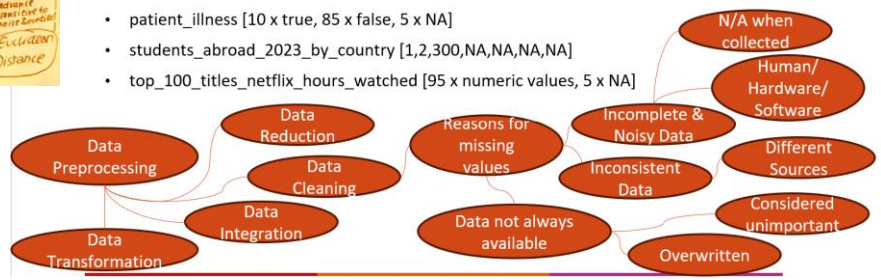
Week 16: Recap



Example 2: Conceptual Knowledge

Frage: Nennen Sie Gründe für fehlende Daten (NA) und schlagen Sie eine Strategie vor, um diese zu finden und sinnvoll zu ersetzen.

- patient_illness [10 x true, 85 x false, 5 x NA]
- students_abroad_2023_by_country [1,2,300,NA,NA,NA,NA]
- top_100_titles_netflix_hours_watched [95 x numeric values, 5 x NA]



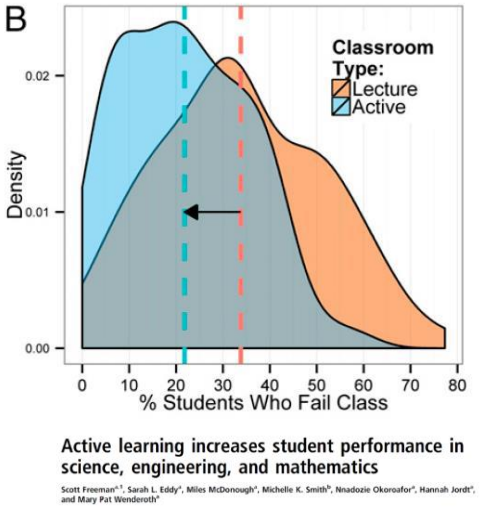
Week 17: Oral Exams

- Questions about lecture concepts (40%), coding problems (40%), and Data Science Life Cycle related to your project (20%).
- Imagine you are the expert providing consultancy to a potential customer ☺

Hierarchy of relevancy:

1. Slides including Training / Think-Pair-Share.
2. Your project.
3. Lab work.
4. Scripts and demos.
5. Books, articles, documentations
   (no readings are relevant if they are not covered in the slides).

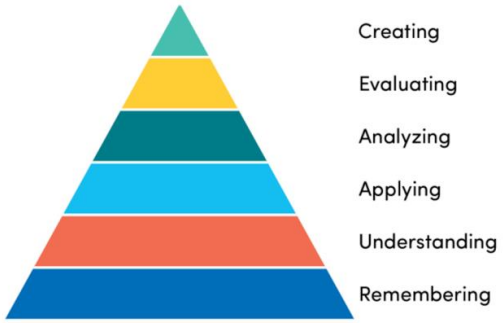
How to succeed in “Programmierkurs 2 Data Science“?



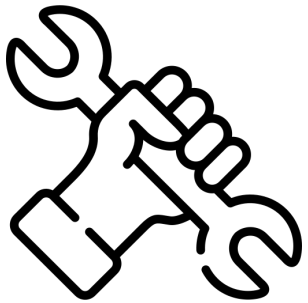
Active learning increases student performance in science, engineering, and mathematics

Scott Freeman[a,1], Sarah L. Eddy[a], Miles McDonough[a], Michelle K. Smith[b], Nnadozie Okoroafor[a], Hannah Jordt[a], and Mary Pat Wenderoth[a]

1. Follow each week‘s **learning goals** (in the beginning of the slides).



2. Participate in **Training** and **Think-Pair-Share**.

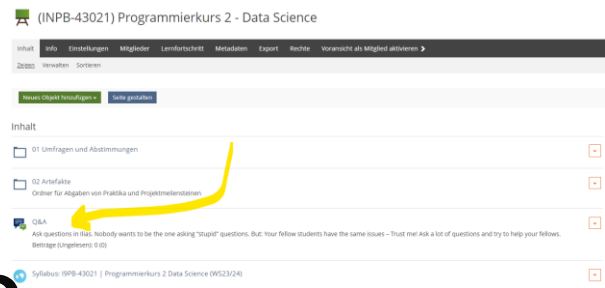


How to succeed in “Programmierkurs 2 Data Science“?

3. Lab Work “Praktikum”:

- Manifestation of conceptual and programming knowledge about frameworks and libraries.
- **Optional**.
- Split into two sections:
  - Lab I: Python I+II+III and NumPy
  - Lab II: Pandas and visualization
- **Preferably** submit in **pairs of two** through Ilias (individual also okay).
- To be completed over three to four weeks.
- Each section contributes up to 8% (total 16%) of additional percentage points towards the final grade.

How to succeed in “Programmierkurs 2 Data Science“?



4. Ask questions in Ilias:

- Nobody wants to be the one asking “stupid” questions.
- But: Your fellow students have the same issues – Trust me!
- Ask a lot of questions and try to **help your fellows**.

A single question or comment related to conceptual frameworks, coding problems, team project, exam preparation, or anything (in your opinion) useful for the class contributes to **additional 0.66%** towards your final grade.

Approaching Problem

- **Emotions in Data Science**

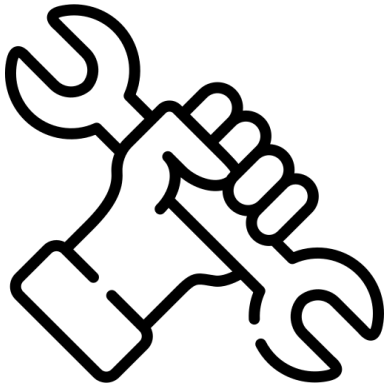
  As a data scientist, most of your time will be spent in a desert of uncertainty, frustration, and doubt.

  There will be rare short-lived interspersed spikes of excitement and happiness due to events like getting a *new dataset*, creating a *new analysis*, getting a *new result*, or being *thanked by a stakeholder*.

  This experience is normal and does not go away.

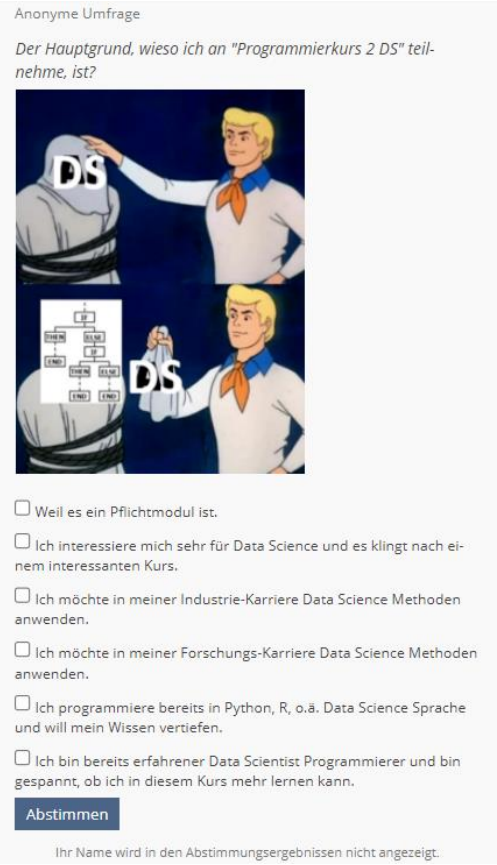
- **Pomodoro Technique:** Concquer issue for 30 minutes, then seek help or do something else.
- **Lesson I:** Ask for help with well-formed questions. *https://stackoverflow.com/help/how-to-ask*
- **Lesson II:** Regardless of how you implement best practices, avoid inventing solutions for which someone else already provided a path.

About you...



Talk to your seating neighbour and ask for their

1. **Motivation** for joining the Data Science program.
2. **Expectation** from this class.
3. **Interest** or hobby.



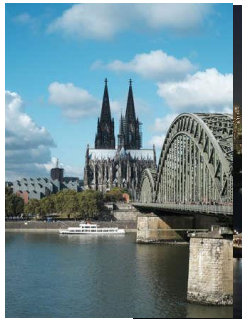
...you are going to shortly introduce your mate afterwards ☺

About me...

2022-now Research in Big Data Analytics (Data Integration)

2019-2022 Data Warehousing

2015-2019 Software Development & Support



Köln Tourismus GmbH/Dieter Jakobi

By Diliff - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=5420726

https://www.visittheusa.de/experience/baltimore-maryland-altbewahrte-tradition-trifft-auf-trendige-stadtviertel

See you after lunch at 14:15!

Questions?