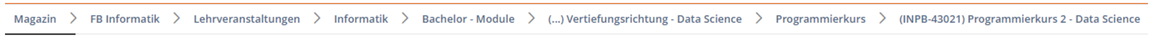
Introduction

Programmierkurs 2 Data Science
WS23/24

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:

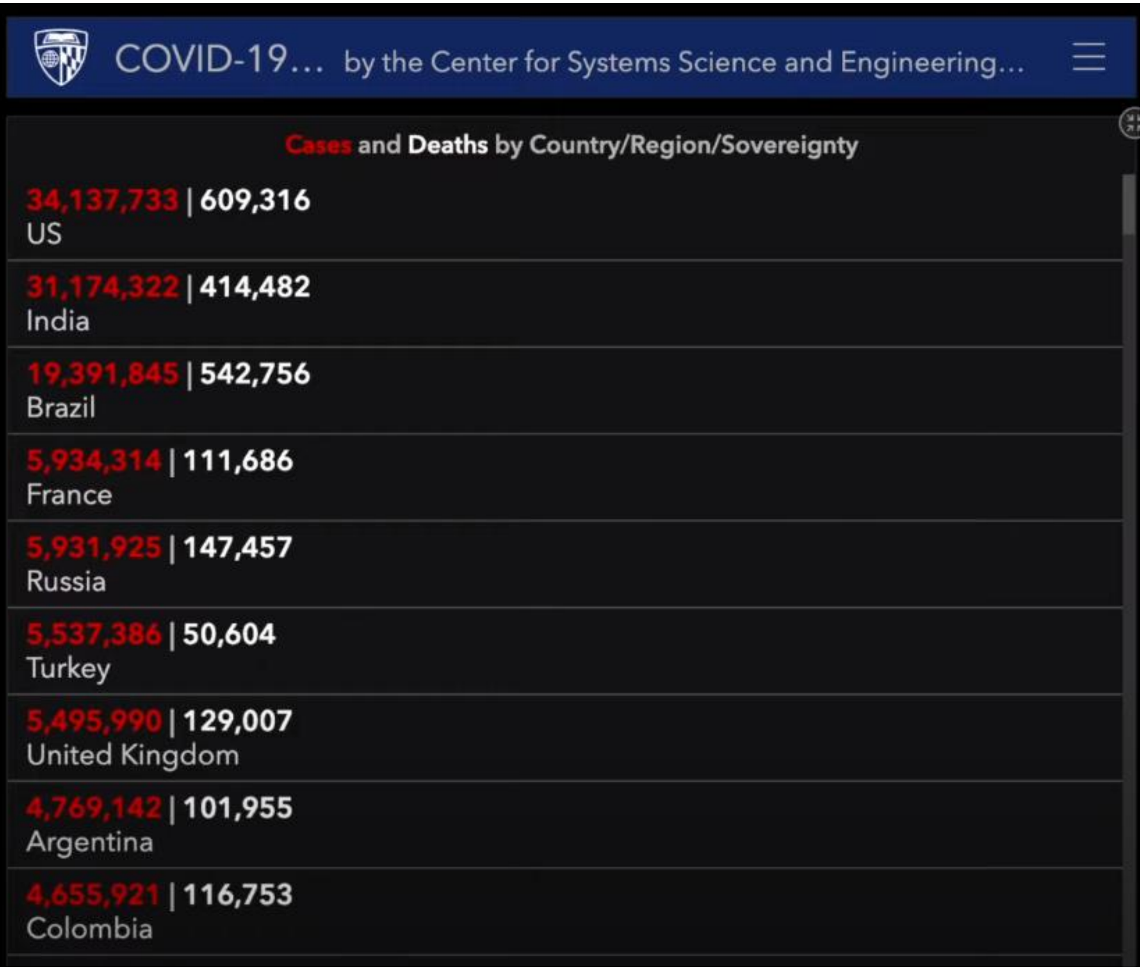
- The **assignments** will be in German.
- The **lecture itself** will be (mostly) in German!
- You can still **answer** the **questions** in the assignments in **German**.

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 projects.
- 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.

Historical Moment of Data Science



COVID-19… by the Center for Systems Science and Engineering…

Cases and Deaths by Country/Region/Sovereignty

34,137,733 | 609,316
US

31,174,322 | 414,482
India

19,391,845 | 542,756
Brazil

5,934,314 | 111,686
France

5,931,925 | 147,457
Russia

5,537,386 | 50,604
Turkey

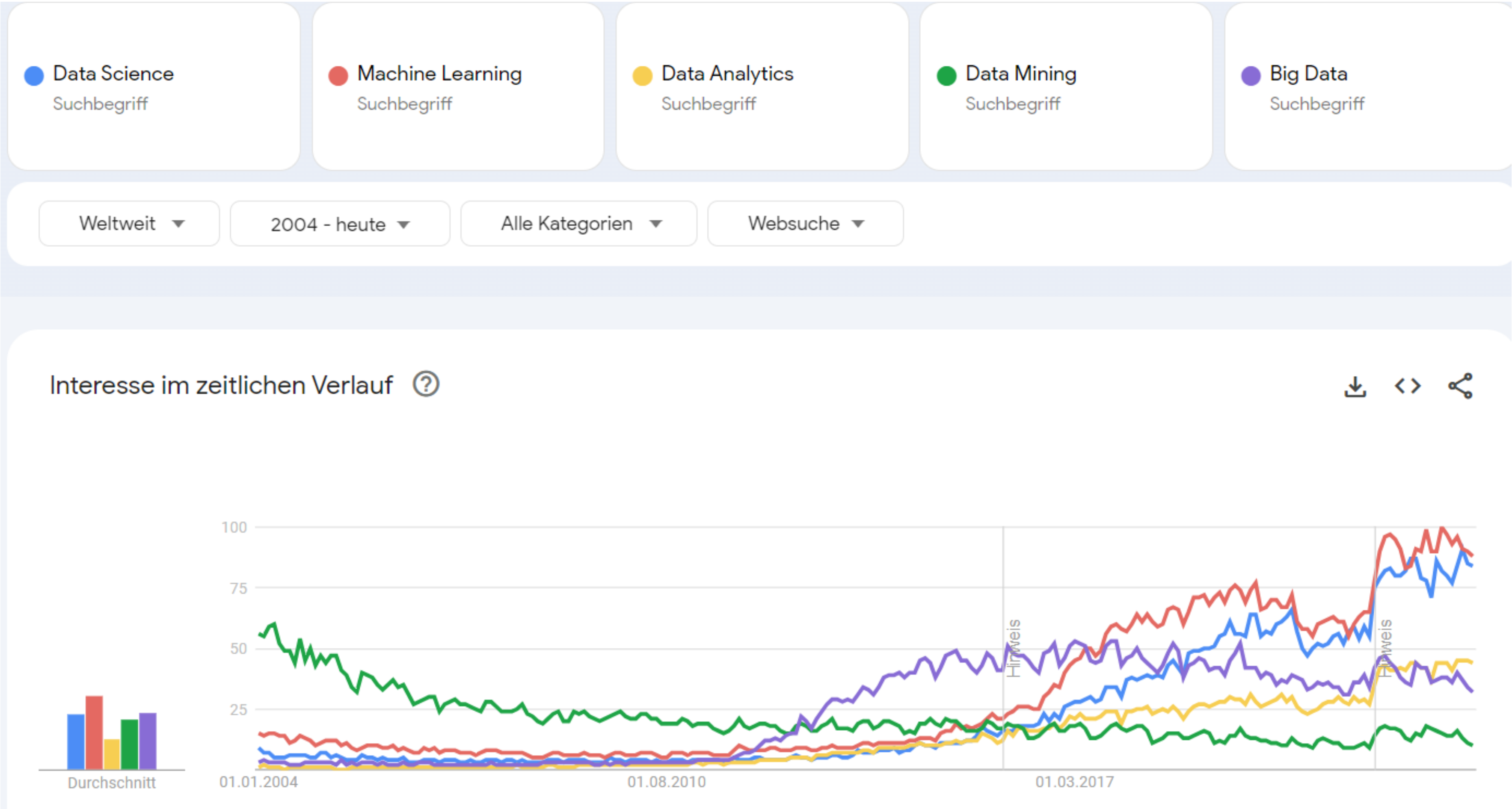
5,495,990 | 129,007
United Kingdom

4,769,142 | 101,955
Argentina

4,655,921 | 116,753
Colombia

https://coronavirus.jhu.edu/map.html - Screenshot taken on July, 20th 2021 6:00 PM CEST

Meta Data Science



● Data Science
Suchbegriff

● Machine Learning
Suchbegriff

● Data Analytics
Suchbegriff

● Data Mining
Suchbegriff

● Big Data
Suchbegriff

Weltweit ▾ | 2004 - heute ▾ | Alle Kategorien ▾ | Websuche ▾

Interesse im zeitlichen Verlauf

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

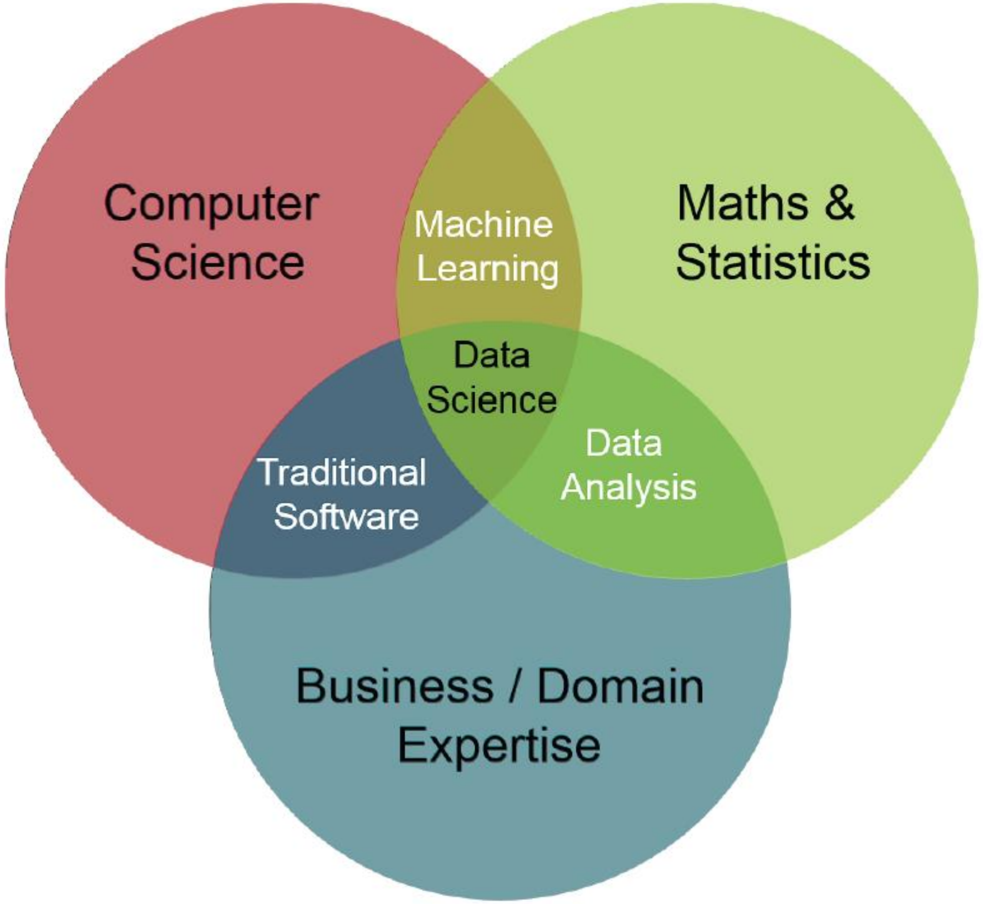
Data Science

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

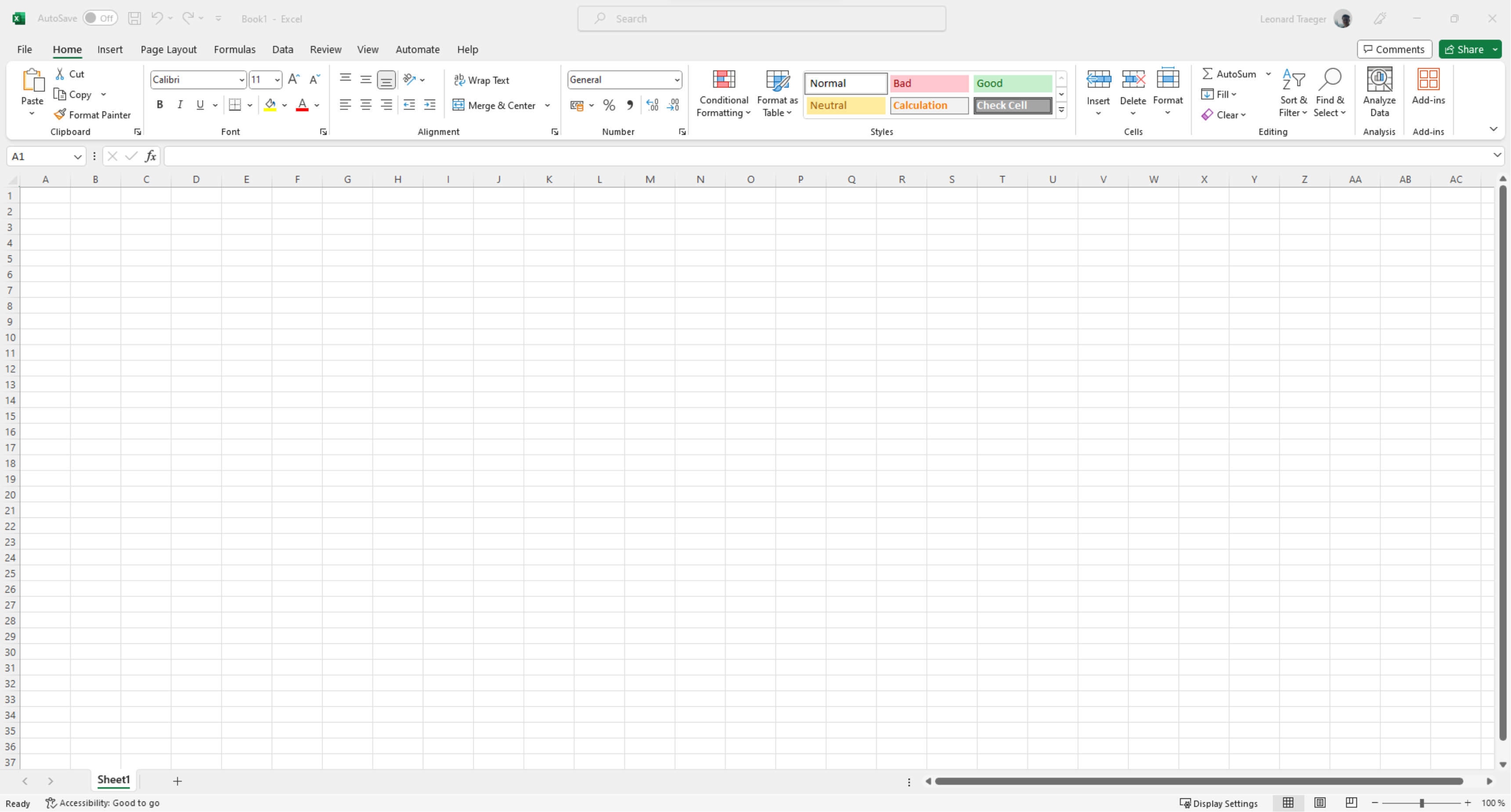
“In der Wissenschaft beschäftigt sich Data Science mit **unterschiedlichen Bereichen** und kann daher verschiedene akademische Hintergründe haben: *Informatik, Statistik, Mathematik, Natur- oder Wirtschaftswissenschaften, Machine Learnings, des statistischen Lernens, der Programmierung, der Datentechnik, 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

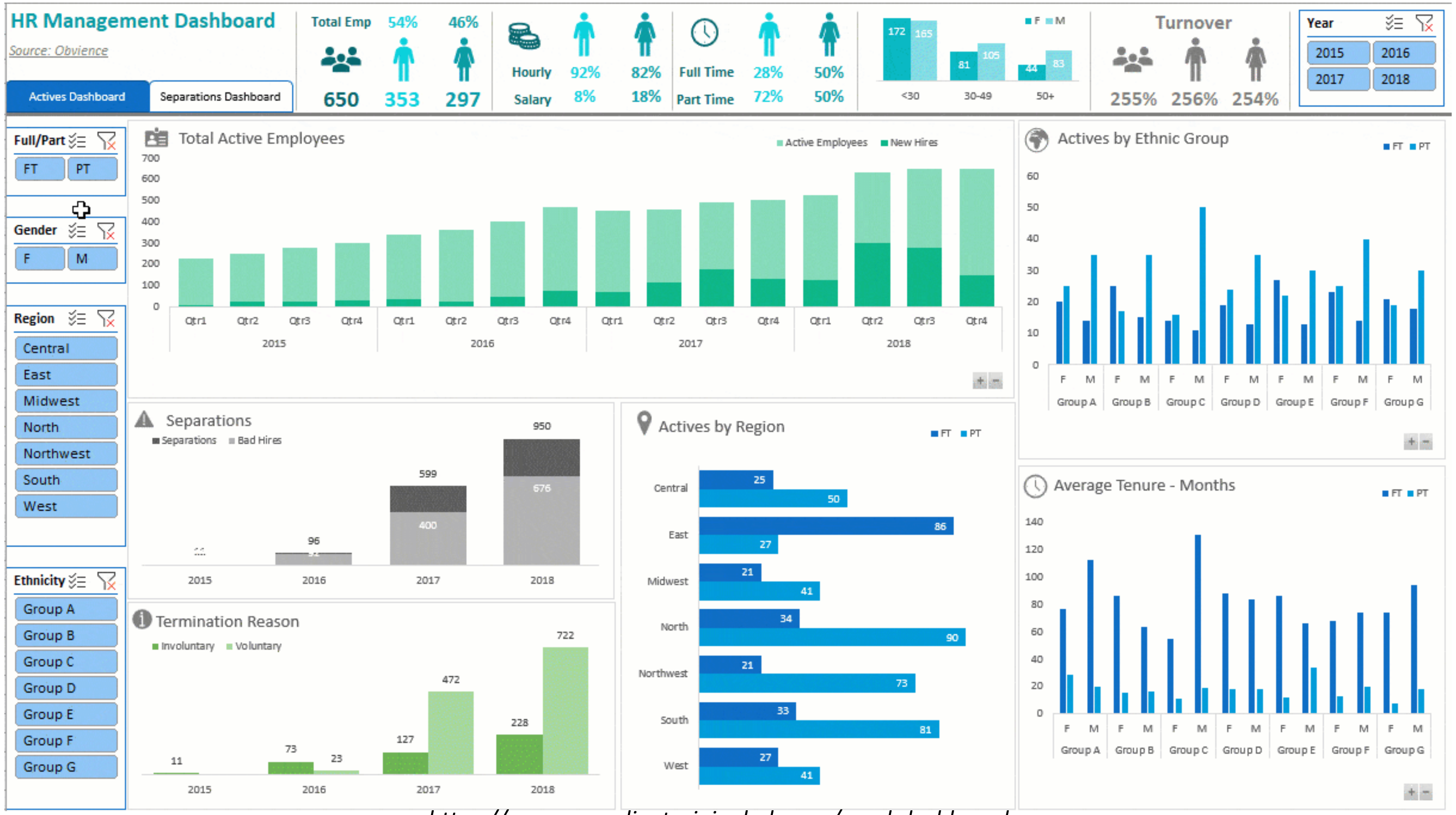
What is Data Science?



Easy...?


AutoSave Off
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
Add-ins
A1
fx
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z AA AB AC
Sheet1
Ready
Accessibility: Good to go
Display Settings
100 %

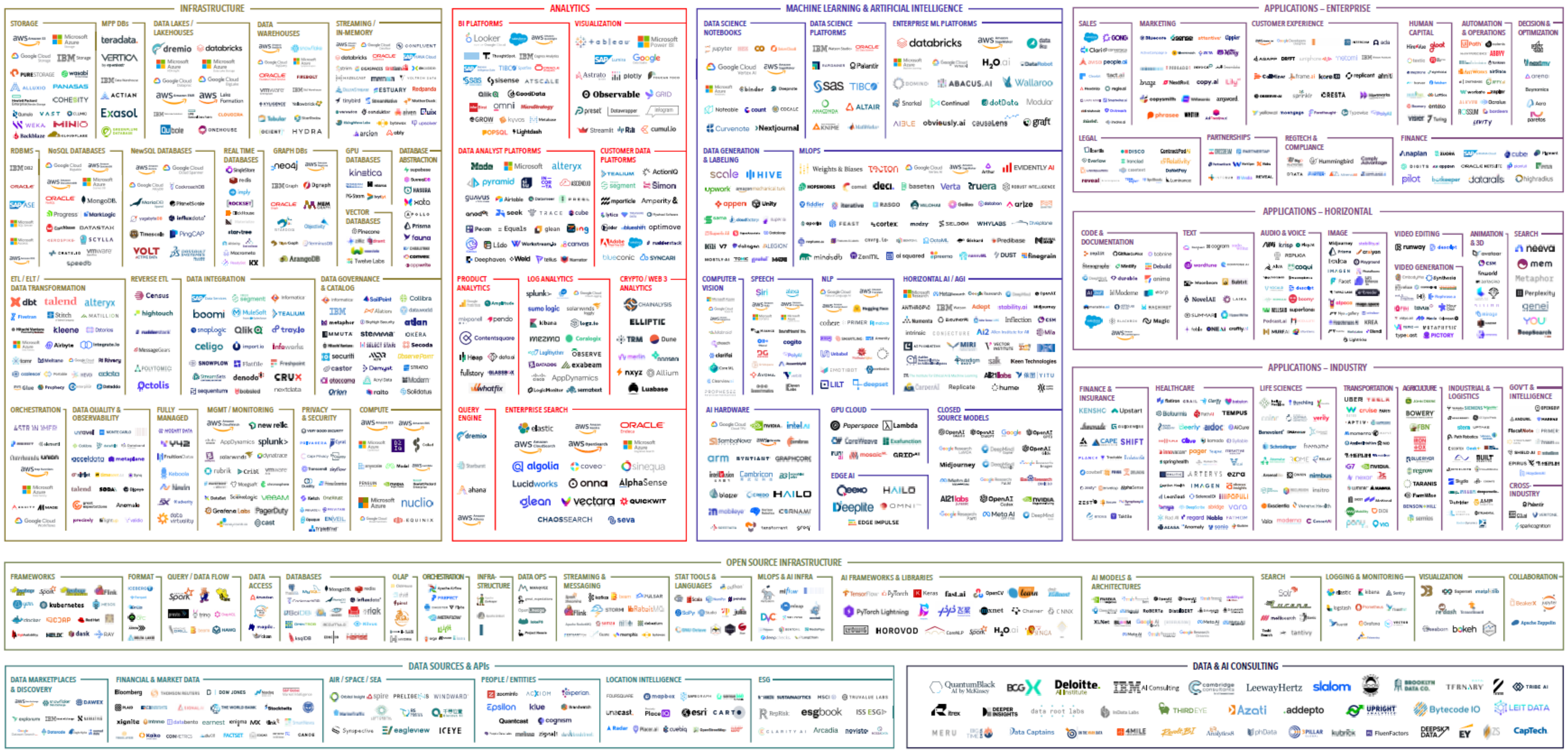
Wow…but isn‘t there much more?



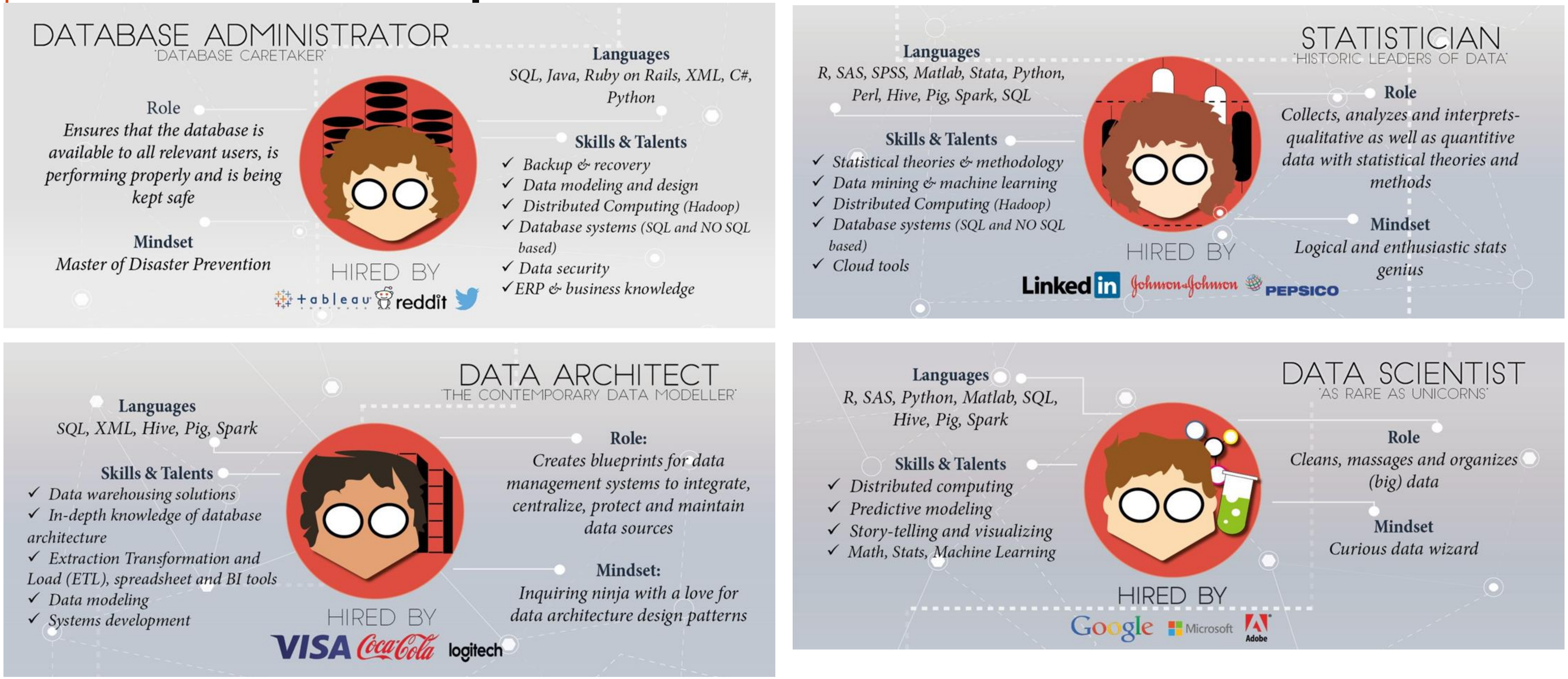
https://www.myonlinetraininghub.com/excel-dashboard-course

World of Data Science

THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE



Skills and Experience > Titles and Labels



DATABASE ADMINISTRATOR
'DATABASE CARETAKER'

Role
Ensures that the database is available to all relevant users, is performing properly and is being kept safe

Mindset
Master of Disaster Prevention

Languages
SQL, Java, Ruby on Rails, XML, C#, Python

Skills & Talents
- ✓ Backup & recovery
- ✓ Data modeling and design
- ✓ Distributed Computing (Hadoop)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Data security
- ✓ ERP & business knowledge

HIRED BY
tableau reddit

STATISTICIAN
'HISTORIC LEADERS OF DATA'

Languages
R, SAS, SPSS, Matlab, Stata, Python, Perl, Hive, Pig, Spark, SQL

Skills & Talents
- ✓ Statistical theories & methodology
- ✓ Data mining & machine learning
- ✓ Distributed Computing (Hadoop)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Cloud tools

Role
Collects, analyzes and interprets- qualitative as well as quantitive data with statistical theories and methods

Mindset
Logical and enthusiastic stats genius

HIRED BY
LinkedIn Johnson&Johnson PEPSICO

DATA ARCHITECT
'THE CONTEMPORARY DATA MODELLER'

Languages
SQL, XML, Hive, Pig, Spark

Skills & Talents
- ✓ Data warehousing solutions
- ✓ In-depth knowledge of database architecture
- ✓ Extraction Transformation and Load (ETL), spreadsheet and BI tools
- ✓ Data modeling
- ✓ Systems development

Role:
Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources

Mindset:
Inquiring ninja with a love for data architecture design patterns

HIRED BY
VISA Coca-Cola logitech

DATA SCIENTIST
'AS RARE AS UNICORNS'

Languages
R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

Skills & Talents
- ✓ Distributed computing
- ✓ Predictive modeling
- ✓ Story-telling and visualizing
- ✓ Math, Stats, Machine Learning

Role
Cleans, massages and organizes (big) data

Mindset
Curious data wizard

HIRED BY
Google Microsoft Adobe

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

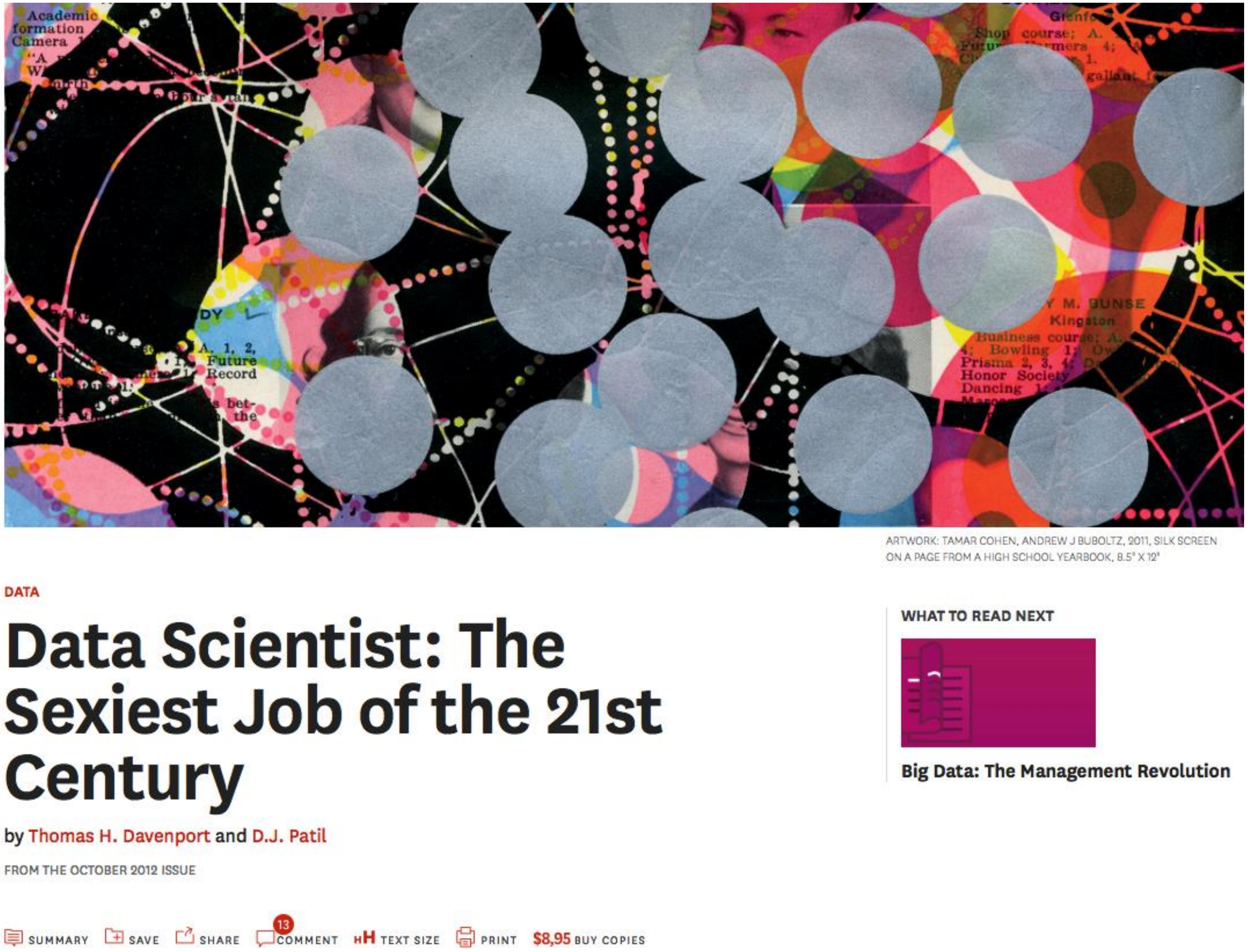
...do not get lost ☺

- 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

WHAT TO READ NEXT

Big Data: The Management Revolution

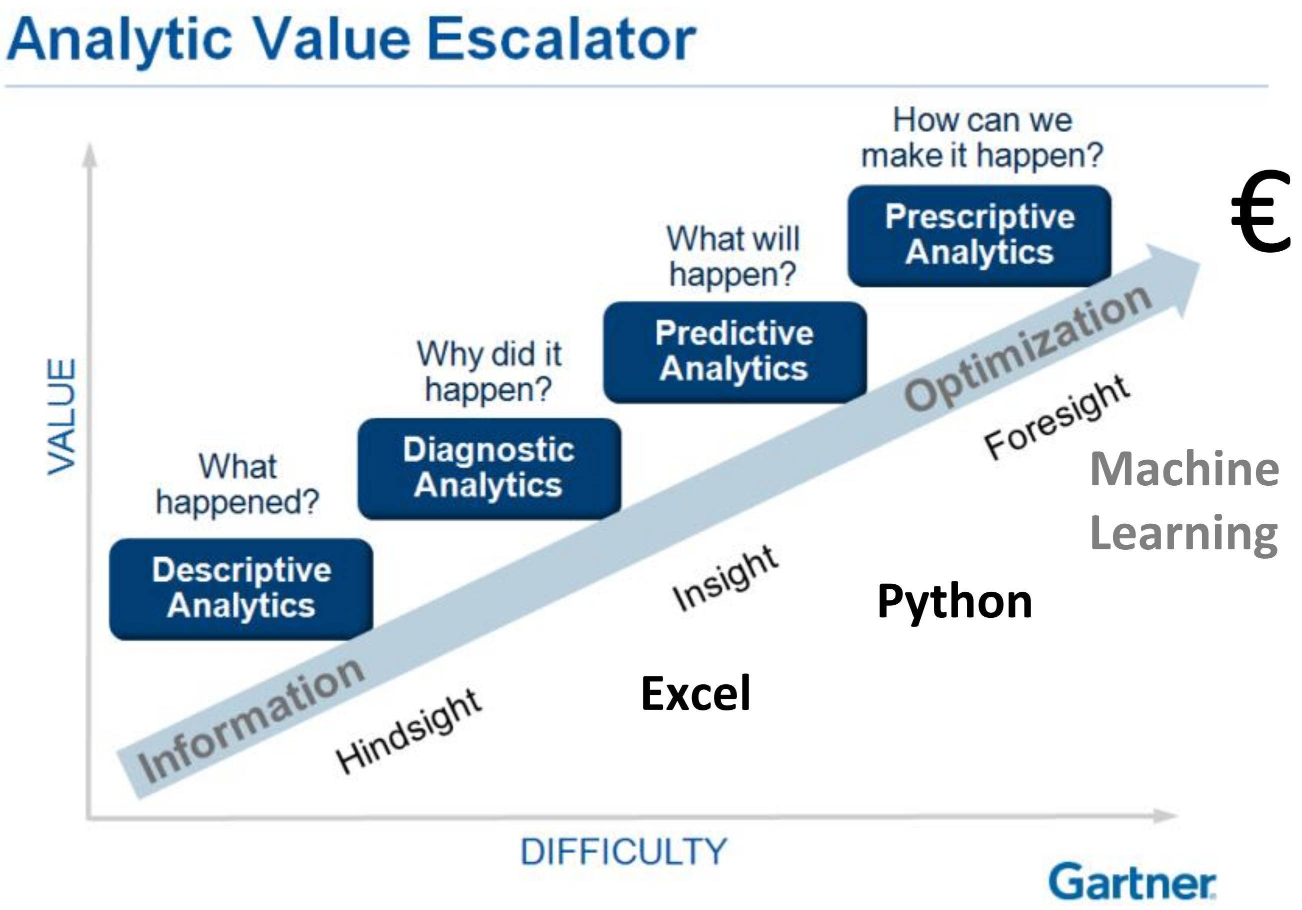
https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Why Learn Data Science?

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

Motives:

- 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

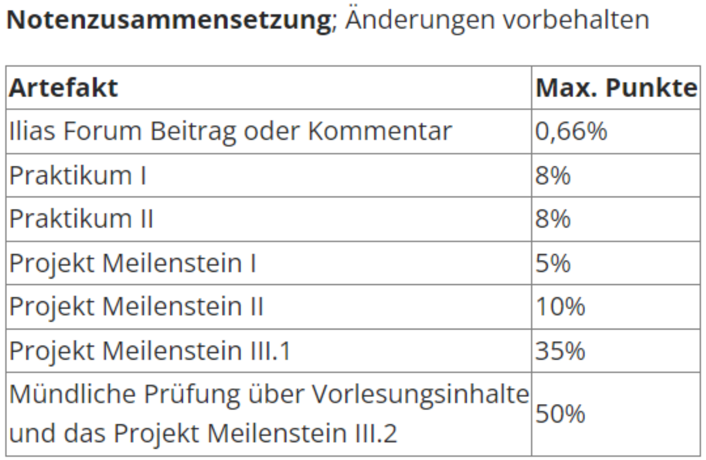
- As a data scientist, you'll often be **work**ing for **someone other** than yourself.
- Expect **under-specified requirements** from customers.
- Provide incomplete solutions (**Minimum Viable Product**) rather than waiting until the product is perfect.

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

Literature

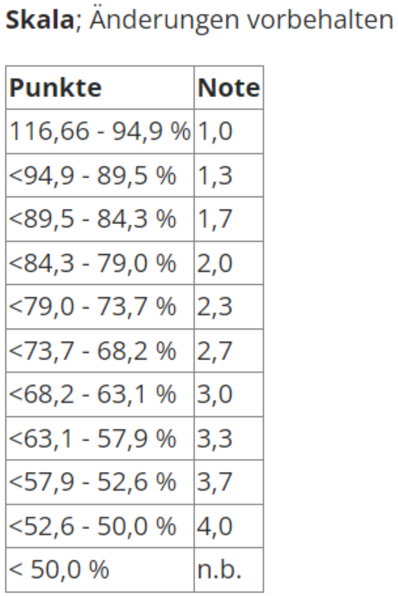
- VanderPlas, J., "Python Data Science Handbook", O'Reilly, 2017
  Digital free copy: https://jakevdp.github.io/PythonDataScienceHandbook/
- Fabio Nelli, "Python Data Analytics With Pandas, NumPy, and Matplotlib" (2nd edition), Apress (Springer), 2018
  Digital free copy via FH VPN
- Wickham, H. und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

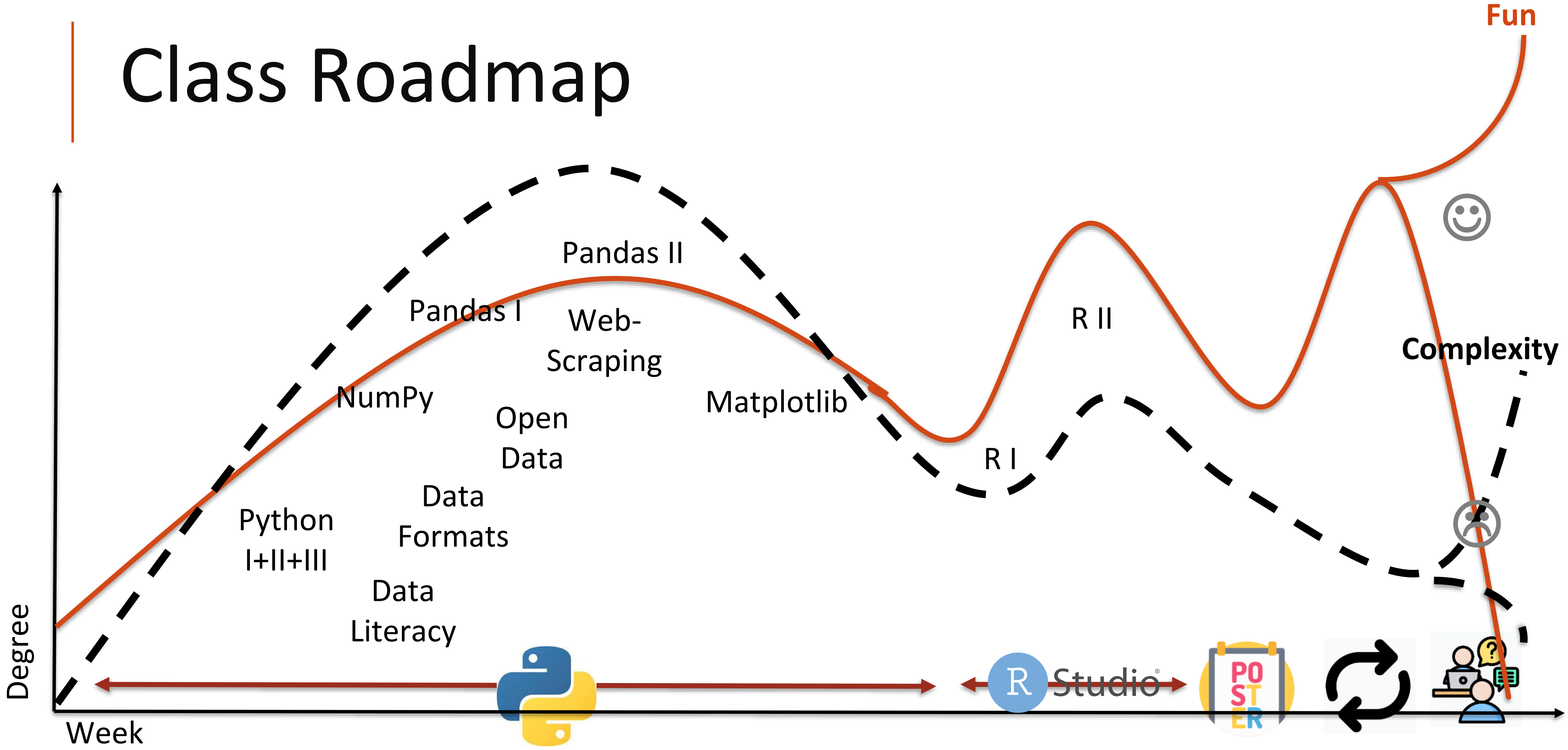


Skala; Änderungen vorbehalten

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

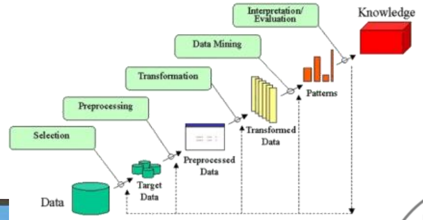
- Introduction Data Science
- Course logistics

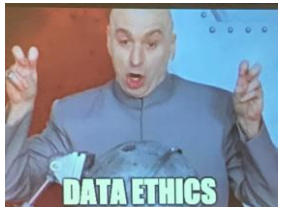
Python I

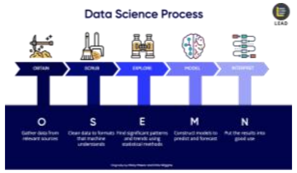
- Python set up
- Jupyter and Colab Notebooks
- Basic Data Types
- Random Numbers
- String methods

Week 3: Python II

- Data Literacy and Ethics
- Data Science Life Cycle







Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

- 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']]
```

veg
lettuce
lettuce
peppers
zucchini
lettuce
lettuce
peppers
zucchini
lettuce
cilantro
peppers
zucchini

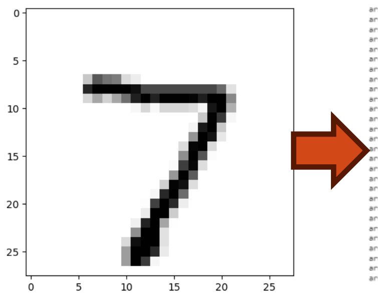
Week 4: Python III + Data Formats

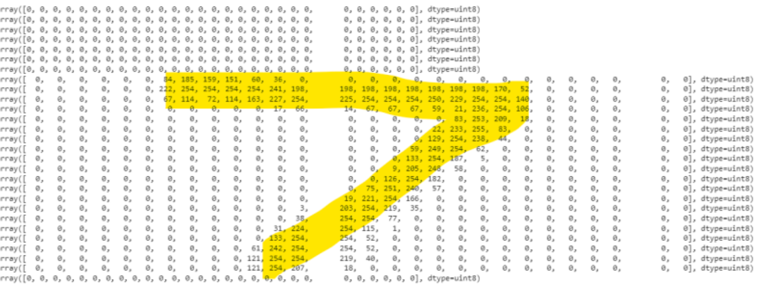
- Imperative and Declarative Paradigm
- Object-Oriented Programming
  - Constructor
  - Destructor
  - Decorator annotated and regular Class Methods
  - Inheritance

- 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



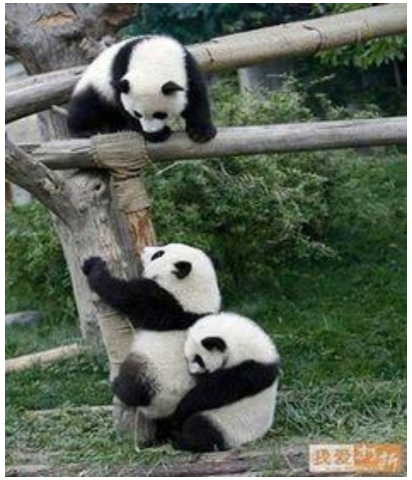


```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,      0, 0, 0, 0, 0, 0], dtype=uint8)
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,      0, 0, 0, 0, 0, 0], dtype=uint8)
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,      0, 0, 0, 0, 0, 0], dtype=uint8)
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,      0, 0, 0, 0, 0, 0], dtype=uint8)
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,      0, 0, 0, 0, 0, 0], dtype=uint8)
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,      0, 0, 0, 0, 0, 0], dtype=uint8)
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,      0, 0, 0, 0, 0, 0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,  84, 185, 159, 151,  60,  36,   0,      0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0, 222, 254, 254, 254, 254, 241, 198,    198, 198, 198, 198, 198, 198, 198, 170,  52,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,  67, 114,  72, 114, 163, 227, 254,    225, 254, 254, 254, 250, 229, 254, 254, 140,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,  17,  66,     14,  67,  67,  67,  59,  21, 236, 254, 106,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,      0,   0,   0,   0,   0,  83, 253, 209,  18,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,      0,   0,   0,   0,  22, 233, 255,  83,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,      0,   0,   0,   0, 129, 254, 238,  44,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,      0,   0,   0,  59, 249, 254,  62,   0,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,      0,   0,   0, 133, 254, 187,   5,   0,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,      0,   0,   9, 205, 248,  58,   0,   0,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,      0,   0, 126, 254, 182,   0,   0,   0,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,      0,  75, 251, 240,  57,   0,   0,   0,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,     19, 221, 254, 166,   0,   0,   0,   0,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   3,    203, 254, 219,  35,   0,   0,   0,   0,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,  38,    254, 254,  77,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,  31, 224,    254, 115,   1,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0, 133, 254,    254,  52,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0,  61, 242, 254,    254,  52,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0, 121, 254, 254,    219,  40,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([  0,   0,   0,   0,   0,   0,   0,   0,   0,   0, 121, 254, 207,     18,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,        0,   0], dtype=uint8)
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,      0, 0, 0, 0, 0, 0], dtype=uint8)
```

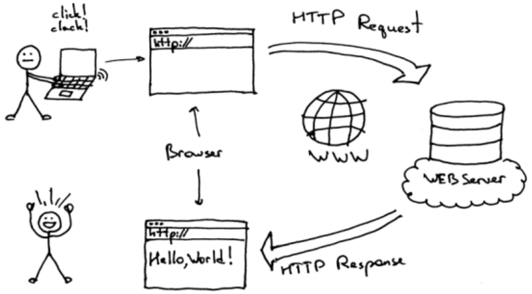
- Open Data and Principles

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: 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

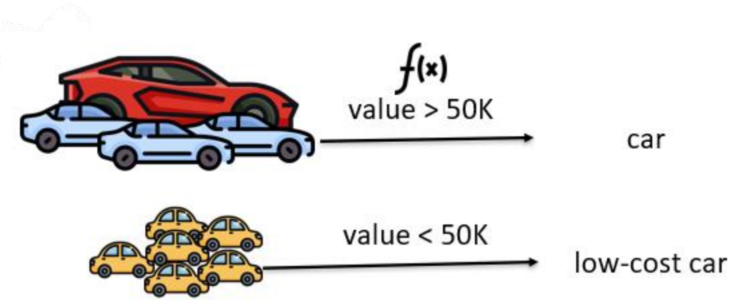
- Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies caused by data integration.

Data Integration

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

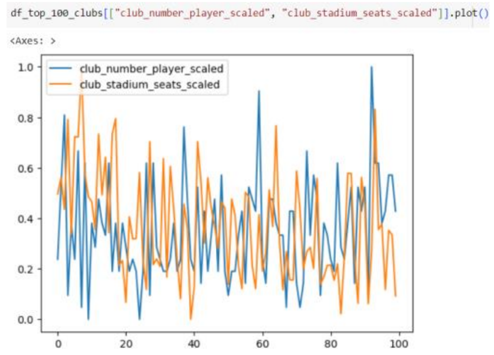
Data Transformation

- Aggregation, generalization, normalization and attribute construction.



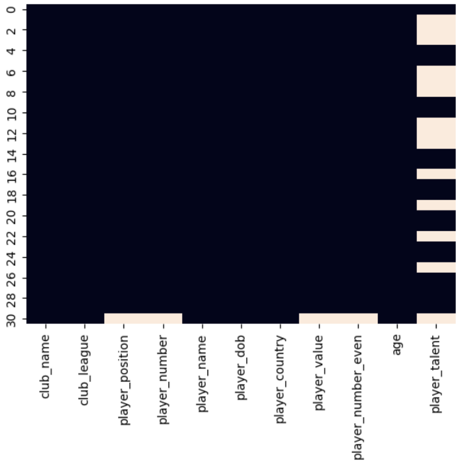
Week 9: Python Matplotlib

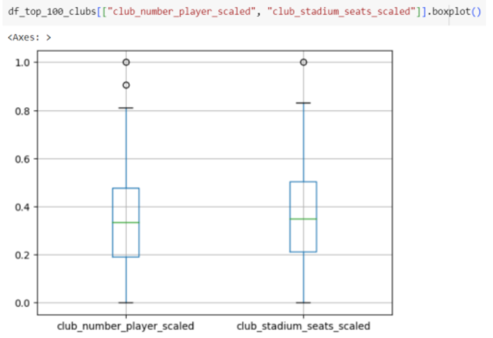
- Simple Plots: Bar, Pizza, Histogram,
- Text, Annotation, Color
- Data Summary Plots
- Meta Data Plots
- Encoder Decoder Design Guide



```
df_top_100_clubs[["club_number_player_scaled", "club_stadium_seats_scaled"]].plot()
```

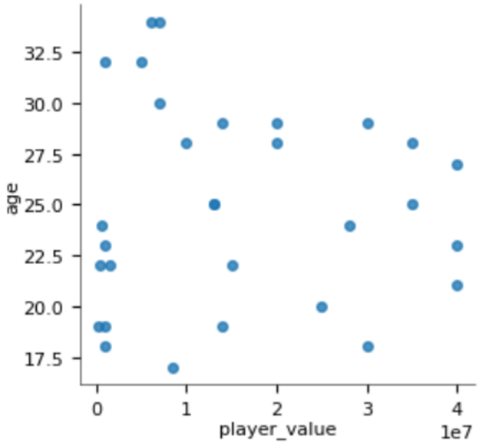
```
<Axes: >
```

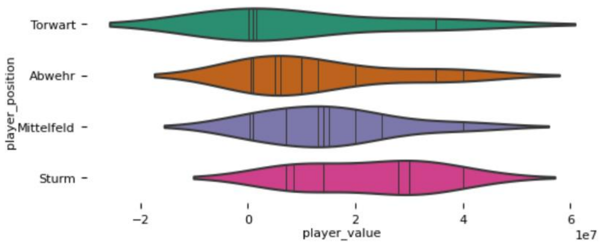


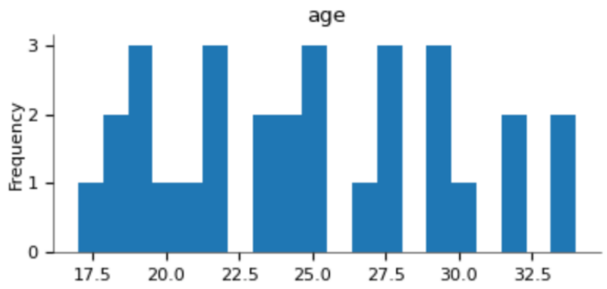


```
df_top_100_clubs[["club_number_player_scaled", "club_stadium_seats_scaled"]].boxplot()
```

```
<Axes: >
```

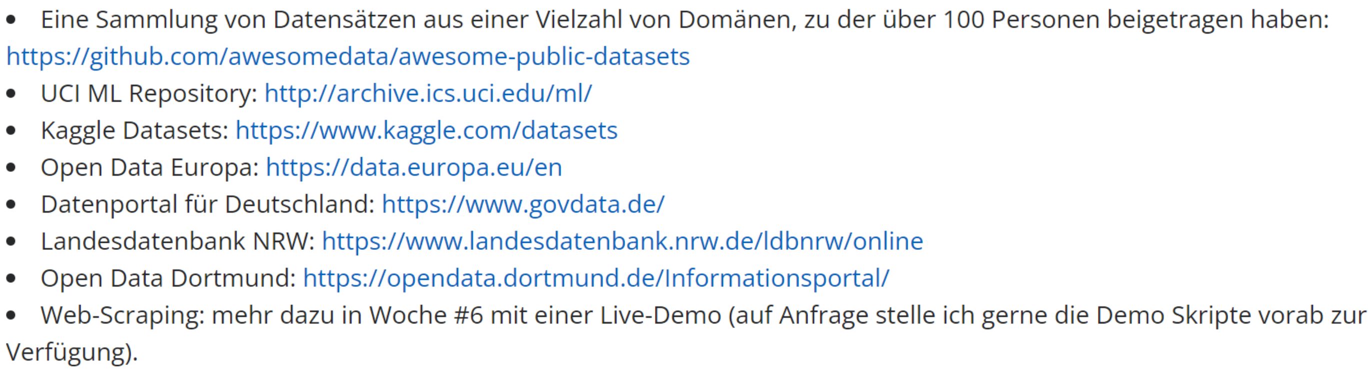






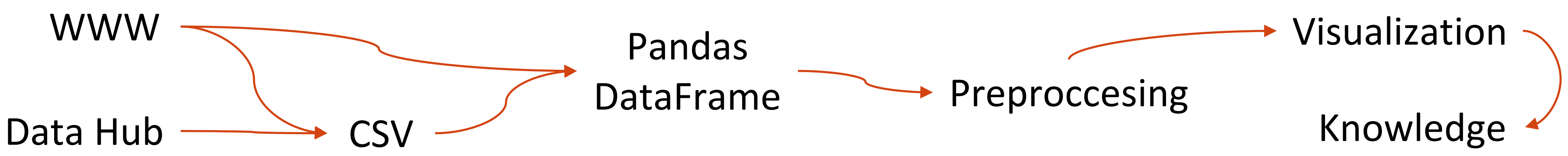
Week 10/11 - 12: R

Team Project



- 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).

- 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.**



WWW

Data Hub

CSV

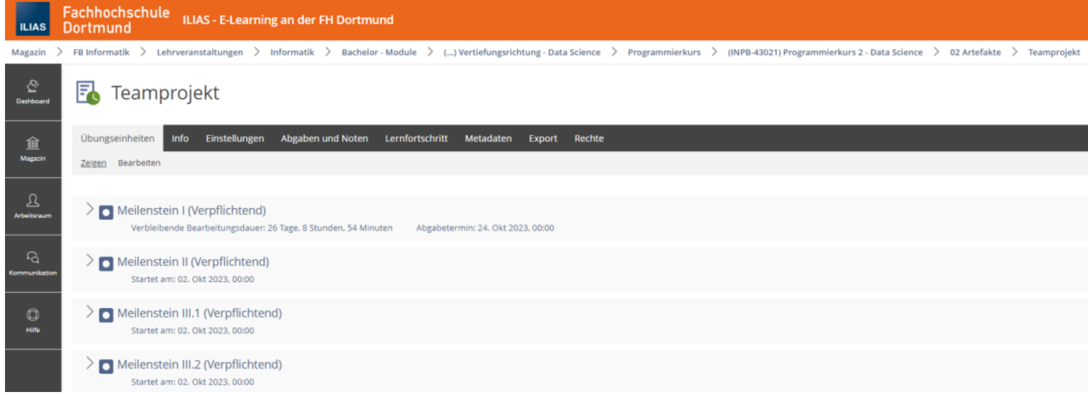
Pandas DataFrame

Preproccesing

Visualization

Knowledge

Team Project (cont.)



ILIAS
Fachhochschule Dortmund
ILIAS - E-Learning an der FH Dortmund
Magazin > FB Informatik > Lehrveranstaltungen > Informatik > Bachelor - Module > (...) Vertiefungsrichtung - Data Science > Programmierkurs > (INPB-43021) Programmierkurs 2 - Data Science > 02 Artefakte > Teamprojekt
Dashboard
Magazin
Arbeitsraum
Kommunikation
Hilfe
Übungseinheiten
Info
Einstellungen
Abgaben und Noten
Lernfortschritt
Metadaten
Export
Rechte
Zeigen
Bearbeiten

Teamprojekt

Meilenstein I (Verpflichtend)
Verbleibende Bearbeitungsdauer: 26 Tage, 8 Stunden, 54 Minuten Abgabetermin: 24. Okt 2023, 00:00

Meilenstein II (Verpflichtend)
Startet am: 02. Okt 2023, 00:00

Meilenstein III.1 (Verpflichtend)
Startet am: 02. Okt 2023, 00:00

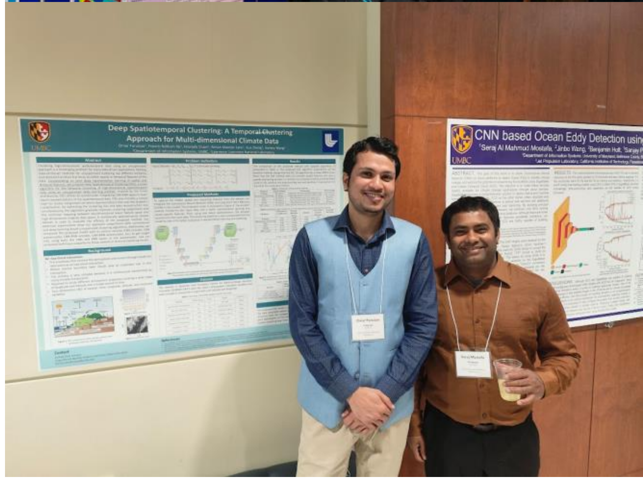
Meilenstein III.2 (Verpflichtend)
Startet am: 02. Okt 2023, 00:00

Milestones and artefact deliverables:

- Teams formed and sent via mail by one team-member by *16.10.23*
- Milestone I (.pdf file) due to *23.10.23*
- Milestone II (.ipynb as file or link to file and print version) due to *27.11.23*
- Milestone III.1 (.pdf or .pptx) due to *08.01.24 (05.01.24 for print via University)*
- Milestone III.2 (.ipynb as file or link to file and print version) due to *16.01.24*

Submitted only via Ilias.

Week 15: Project Day


Deep Spatiotemporal Clustering: A Temporal Clustering Approach for Multi-dimensional Climate Data
CNN based Ocean Eddy Detection usin

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

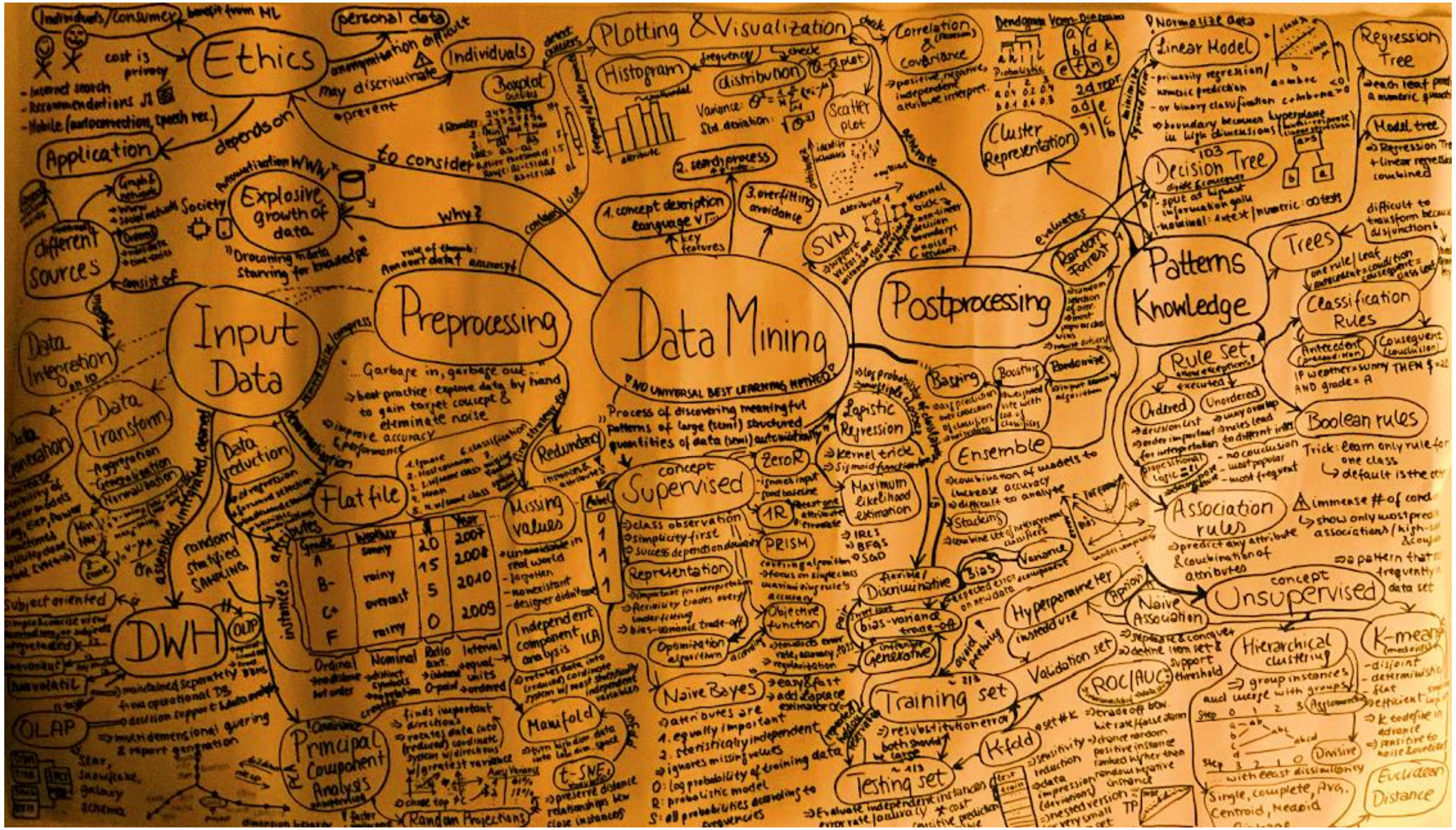
Second hour: graded presentation à 10 minutes for each project.

After lunch: feedback session.

Guidance:

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

Week 16: Recap


Individuals/Consumer
benefit from ML
cost is privacy
- Internet search
- Recommendations
- Mobile (autocorrection, speech rec.)
Application
depends on
Ethics
personal data
anonymization difficult
may discriminate
Individuals
• prevent
to consider
Explosive growth of data
automatization WWW
Society
"Drowning in data Starving for knowledge"
why?
different sources
Graph & network
consist of
Data Integration
Data Transform
- Aggregation
- Generalization
- Normalization
Min Max
z-score
Data Generation
Input Data
assembled, integrated, cleaned
Data reduction
random stratified SAMPLING
Subject oriented
DWH
OLTP
non volatile
maintained separately from operational DB
decision support & data analysis
OLAP
multidimensional quering & report generation
Star, Snowflake, Galaxy schema
dimension hierarchy
attributes
instances
Flat file
Grade
Weather
Year
A
sunny
20
2007
B-
rainy
15
2008
C+
overcast
5
2010
F
rainy
0
2009
Ordinal
Nominal
Ratio
Interval
equal units
ordered
0-point
Principal Component Analysis
PCA
finds important directions
rotates data into (reduced) coordinate system w/ directions w/ greatest variance
choose top PC
Random Projections
Preprocessing
... Garbage in, garbage out ...
best practice: explore data by hand to gain target concept & eliminate noise
improve accuracy & performance
rule of thumb: Amount data ↑ accuracy ↑
1. Ignore
2. Most common
3. 1. w/ same class
4. Mean
5. w/ same class
6. classification
Redundancy
Missing values
unavoidable in real world
- forgotten
- nonexistant
- designer didn't care
Label
0
1
1
1
Independent component analysis
ICA
rotates data into (reduced) coordinate system w/ most statistically independent variables
Manifold
turn high dim data into low dim. space
t-SNE
preserve distance relationships btw close instances
Plotting & Visualization
detect outliers
Boxplot
outliers
Histogram
frequency
distribution
Q-Q plot
check
Variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$
Std. deviation: $\sqrt{\sigma^2}$
Scatter plot
attribute
Correlation (Pearson) & Covariance
positive, negative, independent attribute interpret.
Dendogram
Venn-Diagram
Probabilistic
2d repr.
Cluster Representation
1. concept description language
key features
2. search process
3. overfitting avoidance
SVM
support vectors are instances closest to hyperplane
kernel trick
non-linear decision boundarys
C noise
Data Mining
NO UNIVERSAL BEST LEARNING METHOD!
"Process of discovering meaningful patterns of large (semi) structured quantities of data (semi) automatically"
concept Supervised
class observation
simplicity first
success depends on domain
Representation
important for interpretation
flexibility trades overfitting underfitting
bias-variance trade-off
Optimization algorithm
ZeroR
- ignores input
- good baseline
1R
PRISM
covering algorithm
focus on single class
maximizing rule's accuracy
Objective function
regularization
Naive Bayes
easy & fast
add Laplace estimator
attributes are
1. equally important
2. statistically independent
ignores missing values
O: log probability of training data
R: probabilistic model
S: all probabilities according to frequencies
Logistic Regression
kernel trick
Sigmoid function
Maximum likelihood estimation
IRLS
BFGS
SGD
Flexible / Discriminative
bias-variance trade-off
Generative
Bias
Variance
expected error on new data
Training set
~2/3
resubstitution error
both should be large
Testing set
~1/3
Evaluate independent instances
error rate / accuracy
cost sensitive prediction
K-fold
Validation set
Hyperparameter
instead use
avoid peeking
Postprocessing
Bagging
Boosting
Randomize
Random Forrest
Ensemble
combination of models to increase accuracy
difficult to analyze
Stacking
combine set of heterogeneous classifiers
evaluates
ROC / AUC
trade off btw. hit rate / false alarm
sensitivity
Patterns Knowledge
Normalize data
Linear Model
- primarily regression / numeric prediction
- or binary classification
boundary becomes hyperplane in high dimensions
Decision Tree
divide & conquer
- split at highest information gain
- nominal: 1 test / numeric: ∞ tests
Regression Tree
each leaf predicts a numeric quantity
Model tree
Regression Tree + linear regression combined
Trees
difficult to transform because disjunctions
one rule / leaf
antecedent = condition
consequent = class leaf
Classification Rules
Antecedent (precondition)
Consequent (conclusion)
IF weather = sunny THEN $ = 22 AND grade = A
Rule set
allow exceptions!
executed
Ordered
Unordered
decision list
order important for interpretation
may overlap
rules lead to different infer.
- no conclusion
- most popular
- most frequent
propositional logic
Boolean rules
Trick: learn only rule for one class
default is the other
Association rules
predict any attribute & combination of attributes
immense # of cond.
show only most predictive associations / high-support & confidence
a pattern that frequently in data set
Apriori
Naive Association
separate & conquer
define item set & support threshold
concept Unsupervised
Hierarchical clustering
group instances and merge with groups
Agglomerative
Divisive
Step 0 1 2 3
Step 3 2 1 0
with least dissimilarity
Single, complete, Avg, Centroid, Medoid
K-means (medoids)
- disjoint
- deterministic
- flat
- efficient
k to define in advance
sensitive to noise & outliers
Euclidean Distance

Week 17: Oral Exams

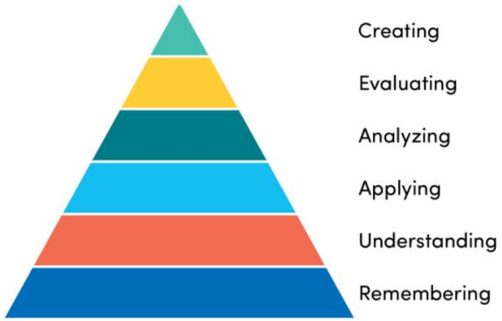
- About lecture contents and project.
- Questions about both conceptual and coding problems.
- Imagine you are the expert providing consultancy to a potential customer ☺

Hierarchy of relevancy:

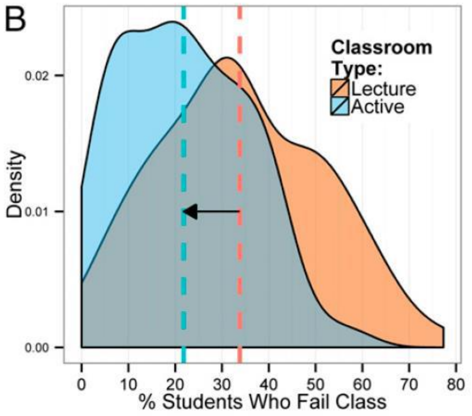
1. Slides including Training / Think-Pair-Share.
2. Your project documents.
3. Lab work.
4. Scripts and demos.
5. Books, articles, documentations
   (no readings are relevant if they are not covered in the slides).

How to succeed in “Programmierkurs 2 Data Science“?

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



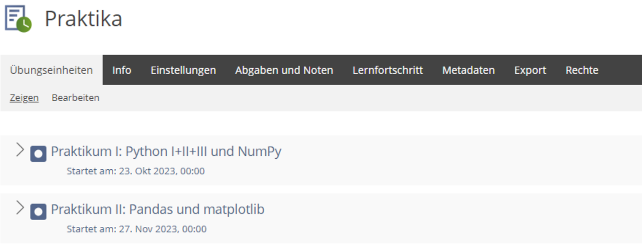
2. Participate in **Training** and **Think-Pair-Share**.



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]

How to succeed in “Programmierkurs 2 Data Science“?



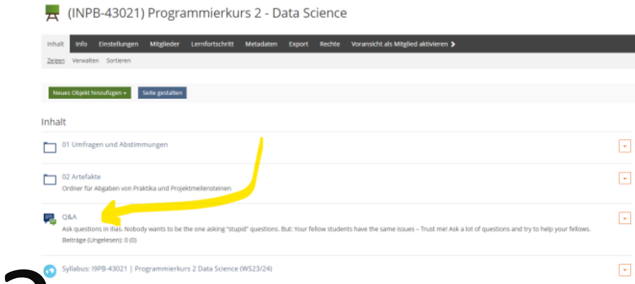
Praktikum I: Python I+II+III und NumPy
Startet am: 23. Okt 2023, 00:00

Praktikum II: Pandas und matplotlib
Startet am: 27. Nov 2023, 00:00

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 matplotlib
- Submit individually or (preferably) in **pairs of two** through Ilias.
- To be completed over three 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“?



(INPB-43021) Programmierkurs 2 - Data Science

Inhalt

01 Umfragen und Abstimmungen

02 Artefakte
Ordner für Abgaben von Praktika und Projektmeilensteinen

Q&A
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.
Beiträge (Ungelesen): 0 (0)

Syllabus: I9PB-43021 | Programmierkurs 2 Data Science (WS23/24)

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.

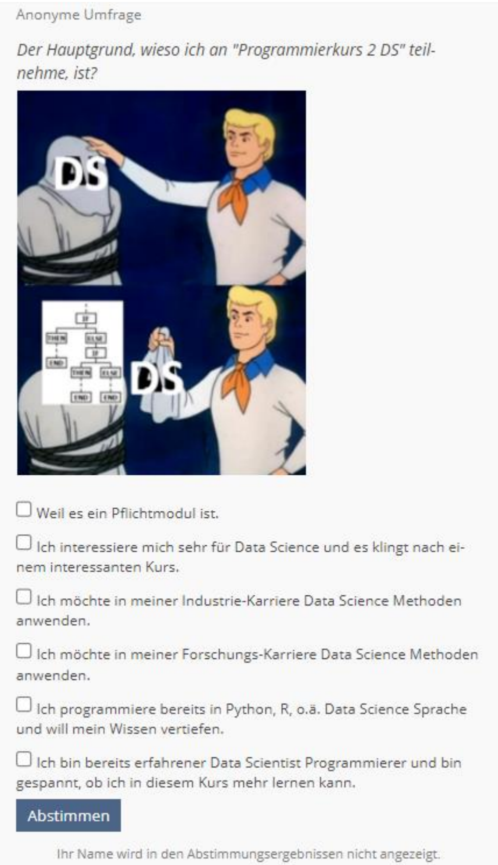
- **Pomodoro Technique:** Concquer issue for 30 minutes, then seek help or do something else.
- **Lesson I:** Ask for help with well-formed questions. *https://stackoverflow.com/help/how-to-ask*
- **Lesson II:** Regardless of how you implement best practices, avoid inventing solutions for which someone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

About you...

Get to know your seating neighbour and ask for their

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

...you are going to introduce your mate to the class afterwards ☺



Anonyme Umfrage

Der Hauptgrund, wieso ich an "Programmierkurs 2 DS" teilnehme, ist?

□ Weil es ein Pflichtmodul ist.

□ Ich interessiere mich sehr für Data Science und es klingt nach einem interessanten Kurs.

□ Ich möchte in meiner Industrie-Karriere Data Science Methoden anwenden.

□ Ich möchte in meiner Forschungs-Karriere Data Science Methoden anwenden.

□ Ich programmiere bereits in Python, R, o.ä. Data Science Sprache und will mein Wissen vertiefen.

□ Ich bin bereits erfahrener Data Scientist Programmierer und bin gespannt, ob ich in diesem Kurs mehr lernen kann.

Ihr Name wird in den Abstimmungsergebnissen nicht angezeigt.

About me...

2022-now Scientific 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?