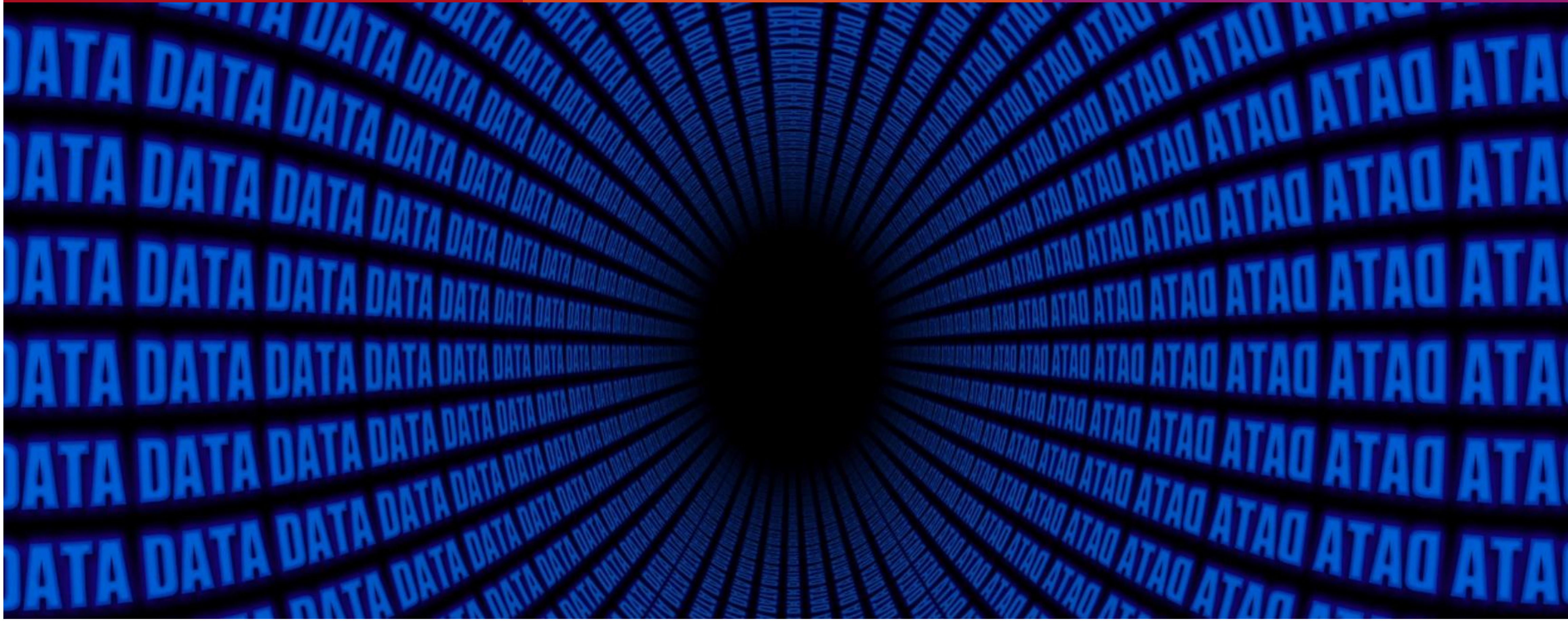
Data Mining SS25, Data Visualization

Fakultät für Informatik und Ingenieurwissenschaften
Institut für Data Science, Engineering, and Analytics

Technology
Arts Sciences
TH Köln

Agenda

- About Data
- Data Analysis:
  - IDA & EDA
- Data Visualization
- Additional Resources



About Data

Zühlke & Bader

Fakultät für Informatik und Ingenieurwissenschaften
Institut für Data Science, Engineering, and Analytics

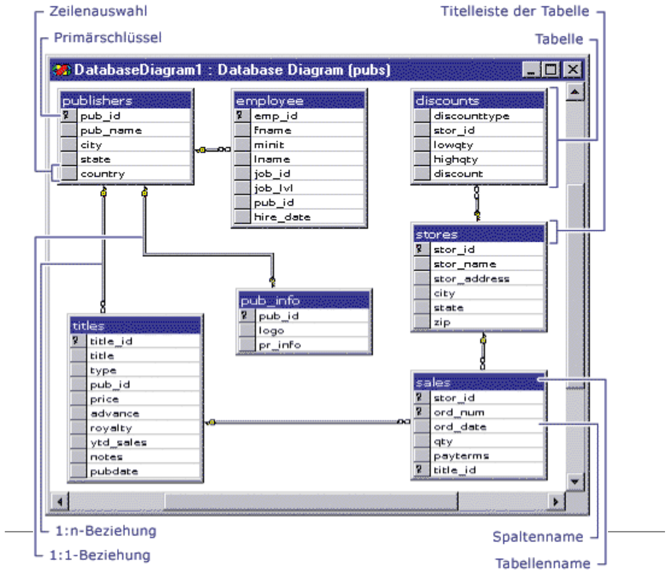
Technology
Arts Sciences
TH Köln

How Data exists

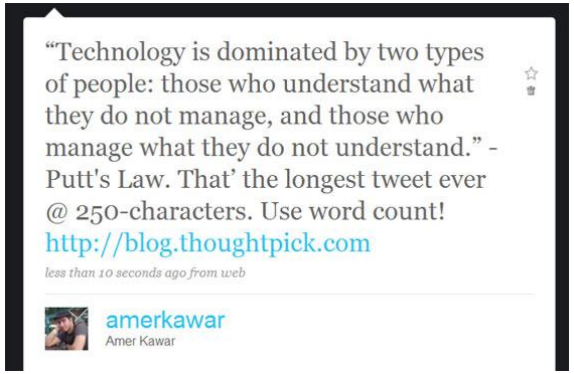
File (e.g. TXT, CSV, Excel, XML)

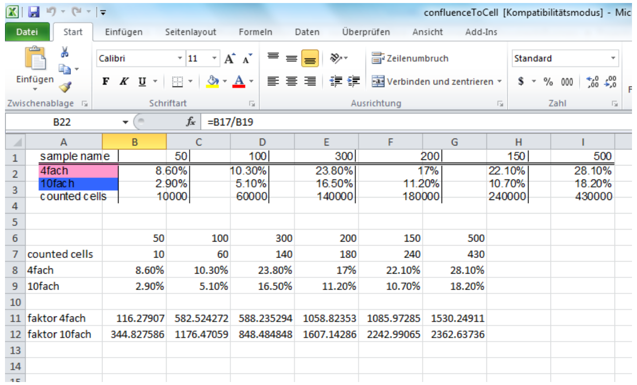
Database (e.g. Access, Oracle, RDF)

Tweets



```xml
<?xml version="1.0" encoding="UTF-8"?>
- <java class="java.beans.XMLDecoder" version="1.6.0">
  - <array class="java.lang.String" length="2">
    - <void index="0">
        <string>Cell</string>
      </void>
    - <void index="1">
        <string>Non-Cell</string>
      </void>
    </array>
  - <object class="de.fraunhofer.fit.lcm.imageanalysis.WindowSizeModel">
    - <void property="maximum">
        <int>200</int>
      </void>
    - <void property="minimum">
        <int>0</int>
      </void>
    - <void property="value">
        <int>35</int>
      </void>
    </object>
  - <object class="java.util.ArrayList">
    - <void method="add">
      - <object class="de.fraunhofer.fit.lcm.glyph.CrossGlyph">
        - <void property="centerPosition">
          - <object class="java.awt.Point">
              <int>423</int>
              <int>290</int>
            </object>
          </void>
        </object>
      </void>
    - <void method="add">
      - <object class="de.fraunhofer.fit.lcm.glyph.CrossGlyph">
        - <void property="centerPosition">
          - <object class="java.awt.Point">
              <int>398</int>
              <int>296</int>
            </object>
```



	A	B	C	D	E	F	G
1	sample name	50	100	300	200	150	500
2	4fach	8.60%	10.30%	23.80%	17%	22.10%	28.10%
3	10fach	2.90%	5.10%	16.50%	11.20%	10.70%	18.20%
4	counted cells	10000	60000	140000	180000	240000	430000
6		50	100	300	200	150	500
7	counted cells	10	60	140	180	240	430
8	4fach	8.60%	10.30%	23.80%	17%	22.10%	28.10%
9	10fach	2.90%	5.10%	16.50%	11.20%	10.70%	18.20%
11	faktor 4fach	116.27907	582.524272	588.235294	1058.82353	1085.97285	1530.24911
12	faktor 10fach	344.827586	1176.47059	848.484848	1607.14286	2242.99065	2362.63736

Types of data sets

Record

- Data Matrix
- Document Data
- Transaction Data

Graph

- World Wide Web
- Molecular Structures

Ordered

- Spatial Data
- Temporal Data
- Sequential Data
- Genetic Sequence Data

Record Data

Data that consists of a collection of records, each of which consists of a fixed set of attributes

Relational Databases

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Matrix

If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

Such data set can be represented by a m x n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data: Document Term Matrix

Each document becomes a `term' vector,

- each term is a component (attribute) of the vector,
- the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	pla y	ball	score	game	wi n	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction Data

A special type of record data, where

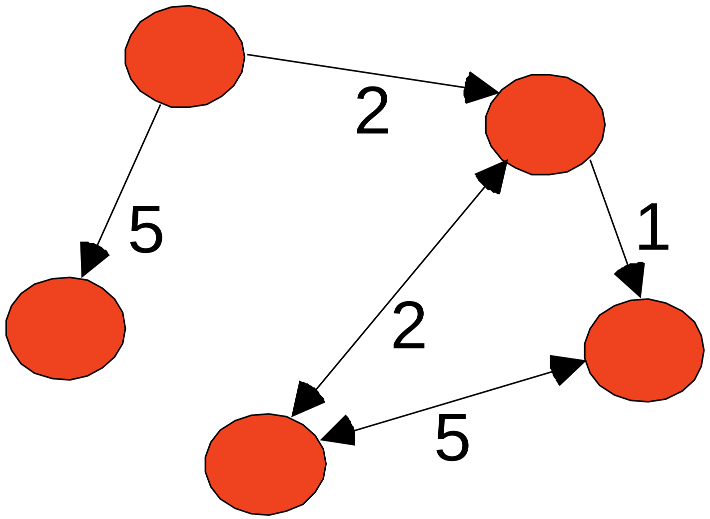
- Each record (transaction) involves a set of items.
- For example, consider a grocery store.
  - The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph Data

Examples: Generic graph and HTML Links

- Usefull for calculating the page rank of a document



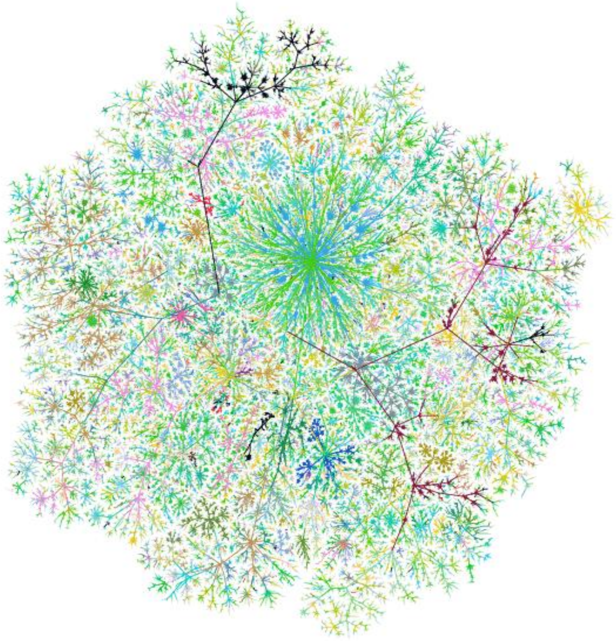
Data Mining

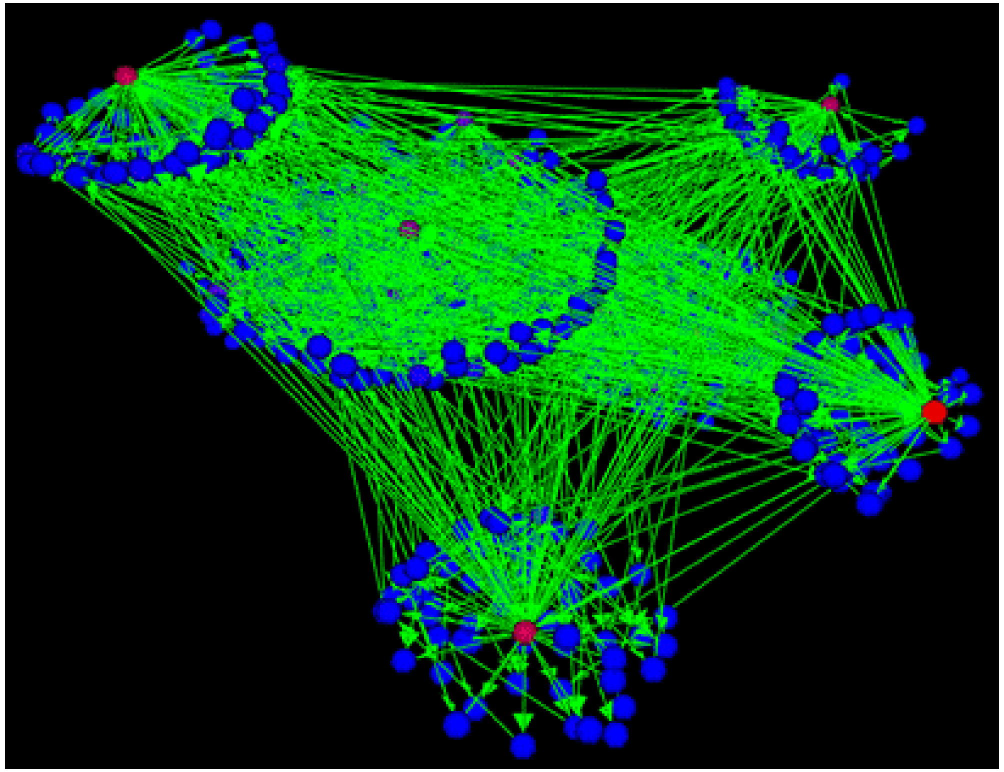
Graph Partitioning

Parallel Solution of Sparse Linear System of Equations

<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers

Internet-Graphs: Server which are connected



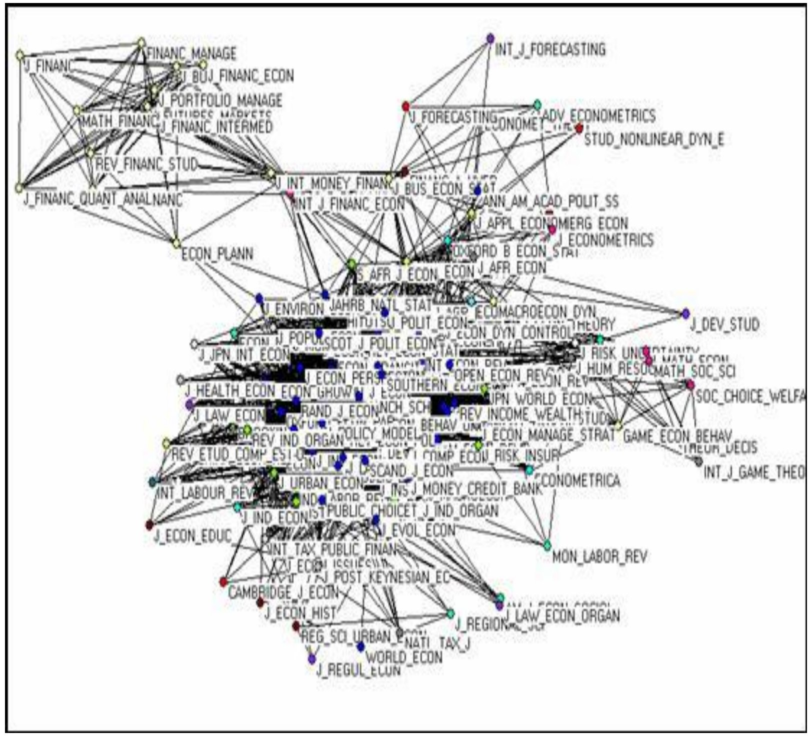


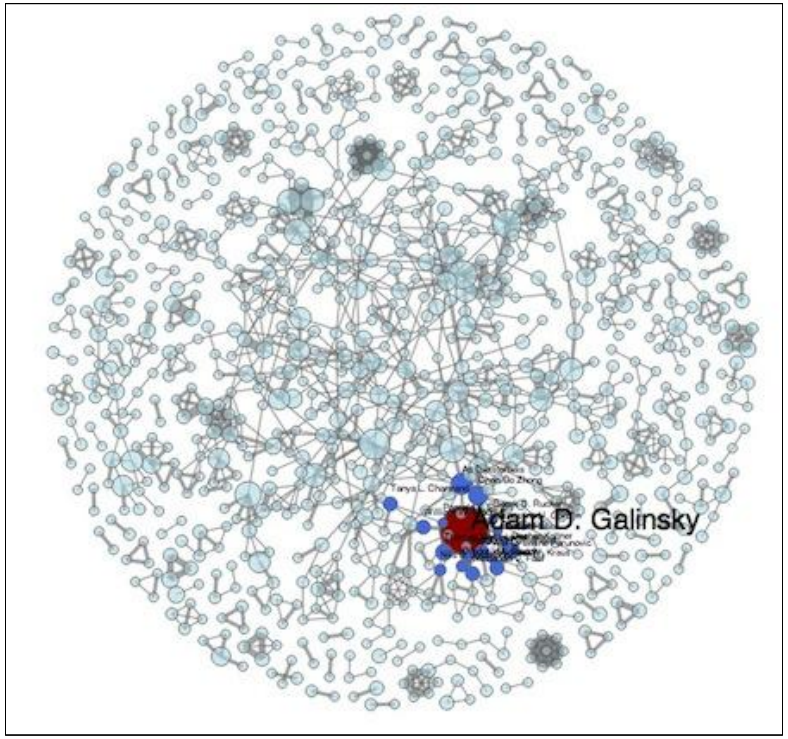
http://www.maths.dur.ac.uk/users/andrew.wade/research/web.gif

http://www.netdimes.org/new/?q=node/17

Technology
Arts Sciences
TH Köln

Citation-Graphs

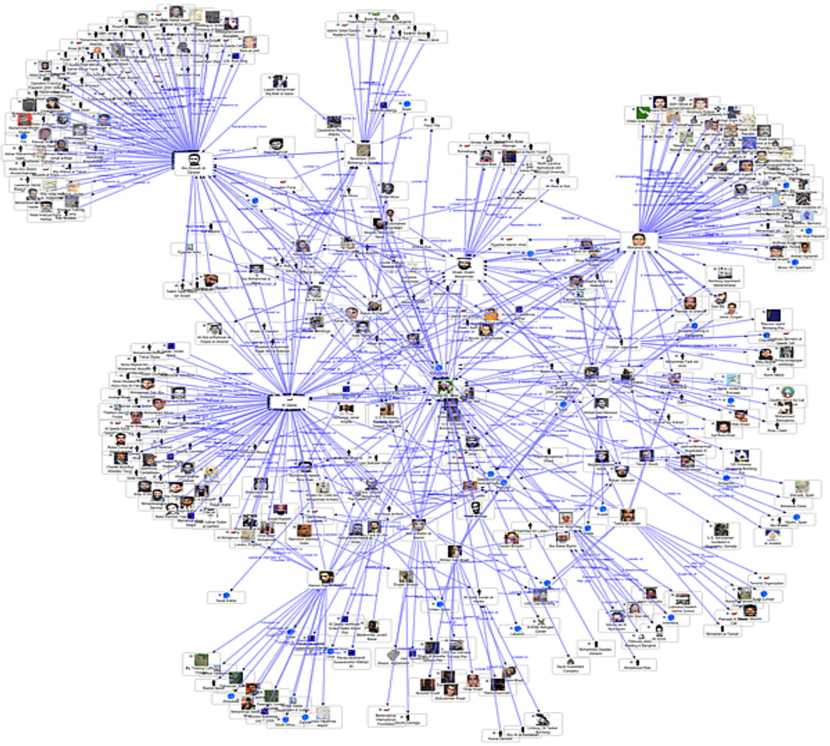


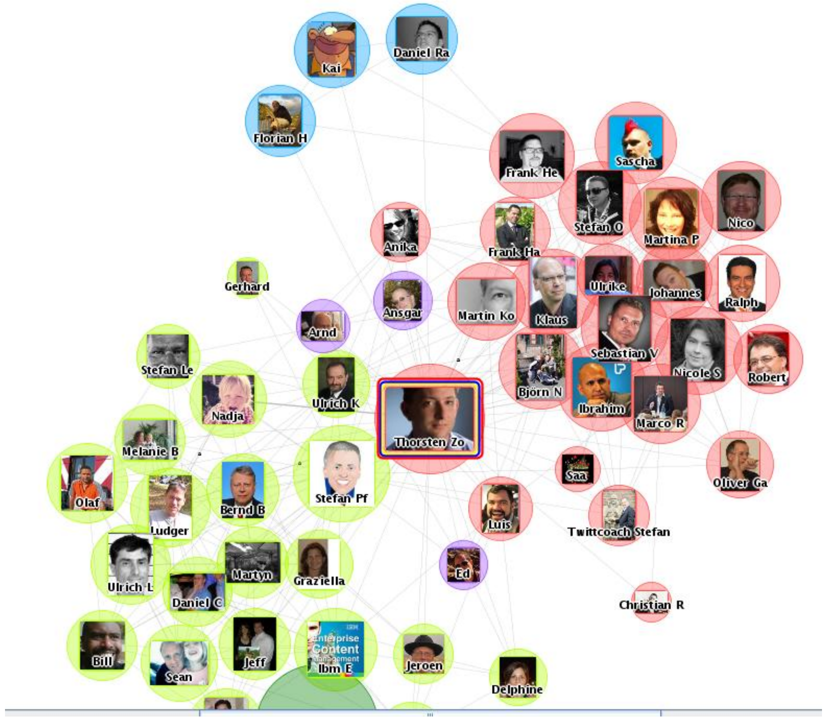


http://www.netdimes.org/new/?q=node/17

http://www.talyarkoni.org/blog/tag/social-graph/

Friendship-Graphs in Social Networks





http://www.fmsasg.com/SocialNetworkAnalysis/

http://www.cyber-junk.de/wp-content/uploads/2010/05/touchgraph.png

What data do we use for modelling? => Record data

Collection of data objects and their attributes

An attribute is a property or characteristic of an object

- Examples: eye color of a person, temperature, etc.
- Attribute is also known as variable, field, characteristic or feature

A collection of attributes describes an object

- An Object is also known as record, point, case, sample, entity or instance

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

Attribute values are numbers or symbols assigned to an attribute.

Distinction between attributes and attribute values:

- Same attribute can be mapped to different attribute values
  - Example: height can be measured in feet or meters
- Different attributes can be mapped to the same set of values
  - Example: Attribute values for ID and age are integers
- But properties of attribute values can be different
  - ID has no limit but age has a maximum and minimum value

Data Types / Types of attributes

Three main data types

- **Nominal**: values are equal or not, but nothing more can be said
- **Ordinal**: values have an order
- **Numeric**: one can do calculations with the values

Algorithms depend on the correct data type

Convert nominal values with an order into ordinal (or numbers)

- Weight: under, normal, over, very_over → 1,2,3,4

Convert numbers where math does not make sense into nominal

- ID: 11 → „11“

Properties of Attribute Values

The type of an attribute depends on which of the following methods can be applied:

- Distinctness: $= \; \neq$
- Order: $< \; >$
- Addition: $+ \; -$
- Multiplication: $* \; /$

- Nominal attribute: distinctness (colors: red, blue, yellow, … / gender: male, female, other)
- Ordinal attribute: distinctness & order (school grades: A, B, C / satisfaction: low, medium, high)
- Interval attribute: distinctness, order & addition (temperature in °C or °F / calendar years)
- Ratio attribute: all four methods (Income: 0$, 2000$ /weight / height / time duration / age)

Other Categorization:
Discrete & Continuous Attributes

Discrete Attribute

- Has only a finite or countable infinite set of values
  - Examples: zip codes, counts or the set of words in a collection of documents
- Often represented as **integer** variables.
- Note: binary attributes are a special case of discrete attributes

Continuous Attribute

- Has real numbers as attribute values
  - Examples: temperature, height, or weight.
- Practically, real values can only be measured and represented using a finite number of digits.
- Continuous attributes are typically represented as **floating-point** variables.

1. Excercise about different types of attribute values

Which operations are allowed ?

Attribut Type	Mathem. Operation (two values)	Aggregation (many values)
nominal		
ordinal		
interval		
ratio		

1. Excercise about different types of attribute values

Which operations are allowed ?

Attribut Type	Mathem. Operation (two values)	Aggregation (many values)
nominal	= ≠	Mode, count
ordinal	all from nominal & <>	Median, count
interval	all from ordinal & +, -	Mean, sum
ratio	all from intervall & *, /	Mean, sum, division

Data Analysis

Data Analysis

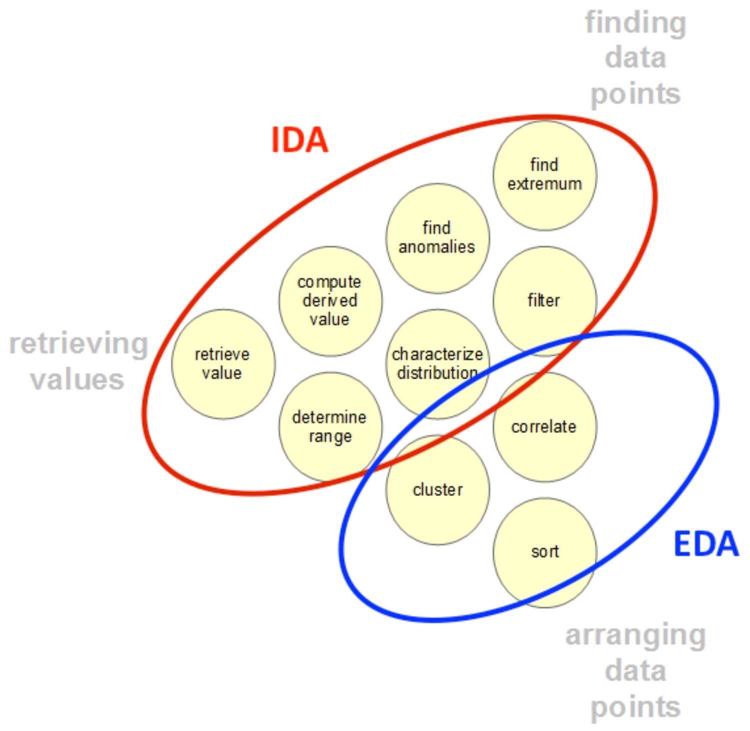
Initial Data Analysis (IDA) vs. Exploratory Data Analysis (EDA)

IDA

- Uncover underlying structure of the dataset.
- Detect outliers and anomalies.
- Test any necessary underlying assumptions.
- Treatment of problems (typically through transformations or imputations)

EDA

- Maximize insight into a data set.
- Understand and rank features by importance.
- Evaluate trade-offs between model simplicity and performance, and identify optimal parameter settings to enhance statistical methods



Source: Avornyo, E. T. (n.d.). *IDA – EDA Method*. Retrieved from https://etav.github.io/articles/ida_eda_method.html

Data Analysis

Initial Data Analysis (IDA) / Structure of the dataset

Understand underlying data structure

Goal: Understand the content, structure, origin and quality of the data.

1. Check the Quality of the Data (Important this happens first)

- Overview of variables: Types (categorical, numerical, etc.), units, meaning.
- Descriptive Summary Statistics (numerically): mean, median, standard deviation, size of the data set (number of features & observations)

1.2 Check the Quality of Data Collection Method

- Check metadata, data description, data collection methods. (Do we have any reason to believe the collection process could lead to systematic errors within the data?)

Data Analysis

Initial Data Analysis (IDA) / Structure of the dataset

- Checking structure: df.shape
- Checking data types: df.info()
- Checking first & last rows: df.head() & df.tail()
- Checking unqiue values: df.nunique().sort_values()

- Checking data quality:
  - Missing values: df.isnull().sum().sort_values(ascending=False)
  - Latest year for features

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13692 entries, 0 to 13691
Data columns (total 17 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   Country                                       13692 non-null  object
 1   Code                                          13692 non-null  object
 2   ContinentCode                                 11932 non-null  object
 3   Year                                          13692 non-null  int64
 4   Industry value added billion USD              8324 non-null   float64
 5   Manufacturing value added billion USD         7498 non-null   float64
 6   Population size in millions                   12514 non-null  float64
 7   Capital investment as percent of GDP          8264 non-null   float64
 8   Economic growth: the rate of change of real GDP  10427 non-null  float64
 9   Economic decline index 0 (low) - 10 (high)    3148 non-null   float64
 10  Economic freedom overall index (0-100)        4941 non-null   float64
 11  Gross Domestic Product billions of U.S. dollars  10734 non-null  float64
 12  Population growth percent                     12316 non-null  float64
 13  Business freedom index (0-100)                4984 non-null   float64
 14  Trade freedom index (0-100)                   4928 non-null   float64
 15  Economic globalization index (0-100)          9064 non-null   float64
 16  Literacy rate                                 1044 non-null   float64
dtypes: float64(13), int64(1), object(3)
memory usage: 1.8+ MB
```

Indicator	2019	2020	2021	2022	2023	2024	2025	2026	2027
Business freedom index (0-100)	182	181	181	175	179	175	0	0	0
Capital investment as percent of GDP	165	163	162	154	128	0	0	0	0
Economic decline index 0 (low) - 10 (high)	175	175	172	176	176	174	0	0	0
Economic freedom overall index (0-100)	177	176	176	175	174	174	0	0	0
Economic globalization index (0-100)	182	181	181	181	0	0	0	0	0
Economic growth: the rate of change of real GDP	190	190	190	190	184	0	0	0	0
Gross Domestic Product billions of U.S. dollars	191	191	191	191	184	0	0	0	0
Industry value added billion USD	187	185	184	181	167	0	0	0	0
Literacy rate	40	32	48	54	5	0	0	0	0
Manufacturing value added billion USD	176	175	172	166	152	0	0	0	0
Population growth percent	196	196	196	195	196	0	0	0	0
Population size in millions	196	196	196	196	196	0	0	0	0
Trade freedom index (0-100)	178	177	177	173	174	175	0	0	0

Data Analysis

Initial Data Analysis (IDA) / Structure of the dataset

- Basic statistics:
  df.describe(include='all').T

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Country	13692	202	Lithuania	69	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Code	13692	202	LTU	69	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ContinentCode	11932	5	AF	3653	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Year	13692.0	NaN	NaN	NaN	1993.928498	19.815637	1960.0	1977.0	1994.0	2011.0	2028.0
Industry value added billion USD	8324.0	NaN	NaN	NaN	60.273512	300.015825	0.0	0.47	2.94	23.185	7032.49
Manufacturing value added billion USD	7498.0	NaN	NaN	NaN	39.128145	210.326726	0.0	0.21	1.44	12.35	4909.02
Population size in millions	12514.0	NaN	NaN	NaN	27.664585	110.160467	0.01	1.0525	5.06	16.49	1428.63
Capital investment as percent of GDP	8264.0	NaN	NaN	NaN	23.147631	8.659905	-15.92	17.965	22.55	27.46	76.78
Economic growth: the rate of change of real GDP	10427.0	NaN	NaN	NaN	9637.170829	983697.553267	-64.05	1.27	3.83	6.31	100448000.0
Economic decline index 0 (low) - 10 (high)	3148.0	NaN	NaN	NaN	5.681226	1.963554	0.7	4.3	5.8	7.1	10.0
Economic freedom overall index (0-100)	4941.0	NaN	NaN	NaN	59.665857	11.681306	1.0	53.0	60.0	67.0	91.0
Gross Domestic Product billions of U.S. dollars	10734.0	NaN	NaN	NaN	197.469521	1078.606597	0.0	1.6125	8.785	56.0025	27720.71
Population growth percent	12316.0	NaN	NaN	NaN	1.731267	1.669597	-27.72	0.71	1.705	2.64	19.36
Business freedom index (0-100)	4984.0	NaN	NaN	NaN	63.899679	15.919347	5.0	55.0	65.0	74.0	100.0
Trade freedom index (0-100)	4928.0	NaN	NaN	NaN	69.99513	14.293197	13.0	62.0	72.0	80.0	95.0
Economic globalization index (0-100)	9064.0	NaN	NaN	NaN	50.264654	17.006363	11.12	37.6075	49.12	62.0	95.29
Literacy rate	1044.0	NaN	NaN	NaN	79.986705	21.092663	5.4	69.0	90.0	96.0	100.0

Data Analysis

Initial Data Analysis (IDA) / Detect outliers and anomalies

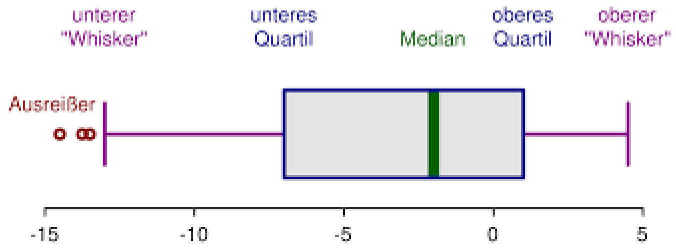
Detect outliers:

1. **Z-Score** (The **Z-score** indicates how many **standard deviations** a value is from the **mean**):
   - Assumption: normal distribution
   - Values with |Z| > 3 (or another threshold value) are considered outliers.

$$Z = \frac{x - \mu}{\sigma}$$

x = data point
μ = mean
σ = standard deviation

2. **Interquartile range (IQR):**
   - IQR is the range of the middle 50% of the data:
     - Q1 = 25th percentile (lower quartile)
     - Q3 = 75th percentile (upper quartile)
     - IQR=Q3-Q1
   - Outliers are outside the range:
     - Lower limit: Q1-1.5×IQR
     - Upper limit: Q3+1.5×IQR

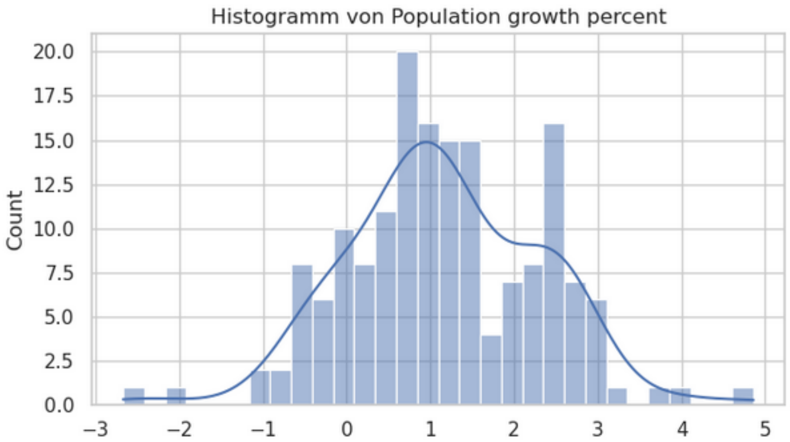


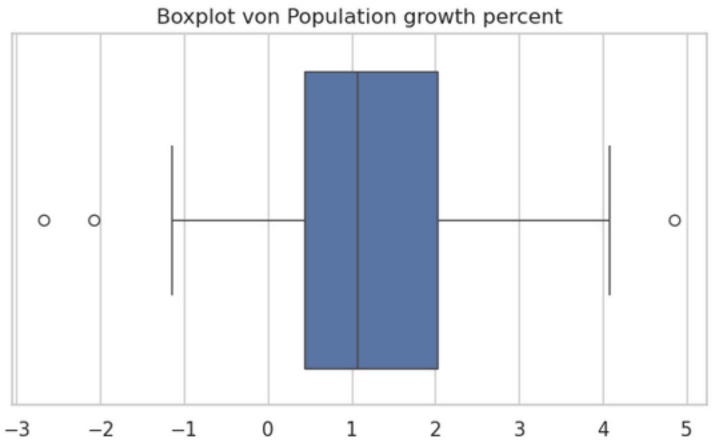
Source: RobSeb – Eigenes Werk, *„Elements of a boxplot“*, Wikipedia

Data Analysis

Initial Data Analysis (IDA) / Detect outliers and anomalies

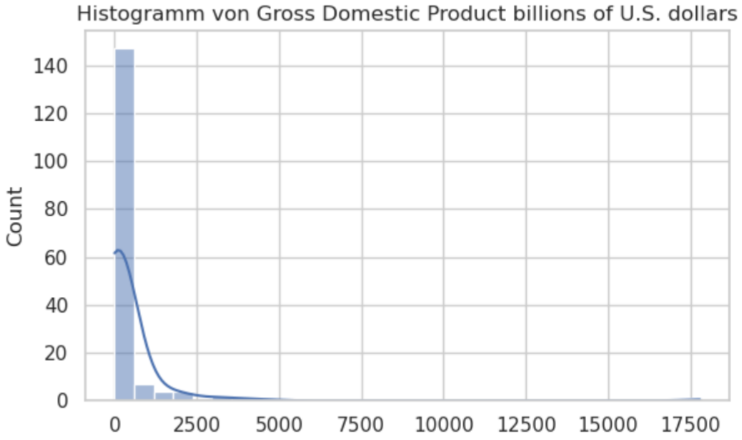
3. **Visually: histograms / boxplots**

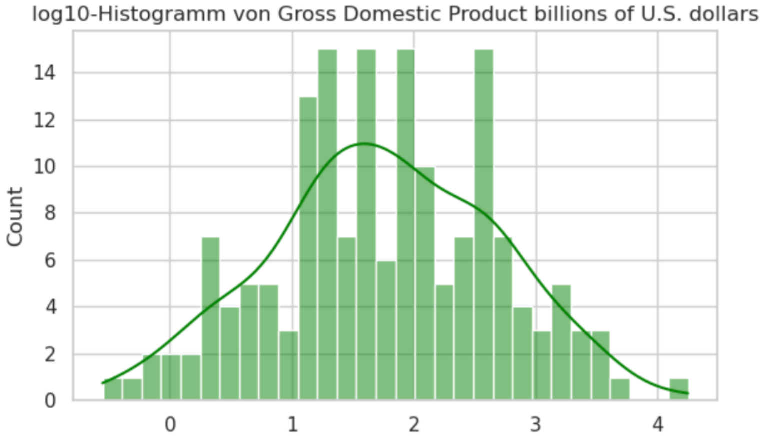




Check the normality of the dataset

- **Visually: histograms**



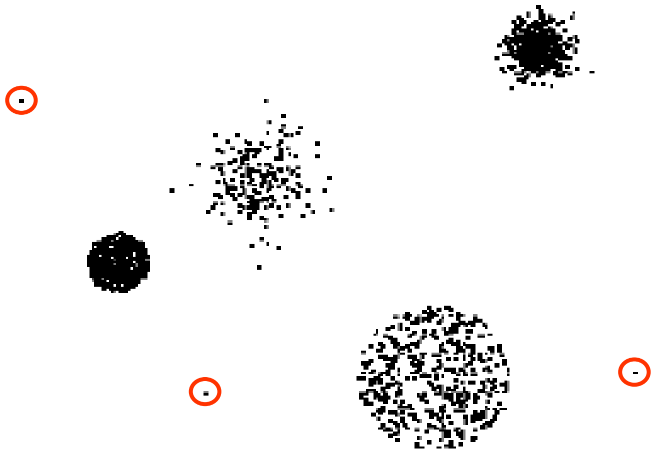


Zühlke & Bader
Data Mining SS25
Fakultät für Informatik und Ingenieurwissenschaften
Institut für Data Science, Engineering, and Analytics
Technology Arts Sciences TH Köln

Data Analysis

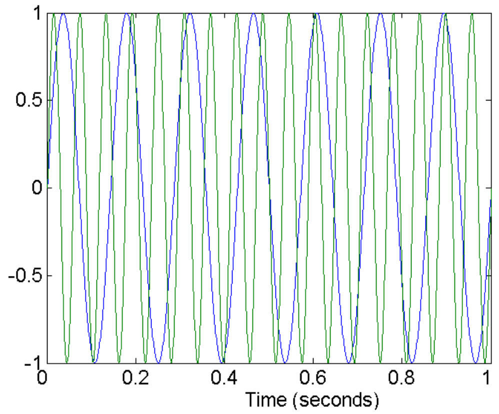
Initial Data Analysis (IDA) / Detect outliers and anomalies

Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set

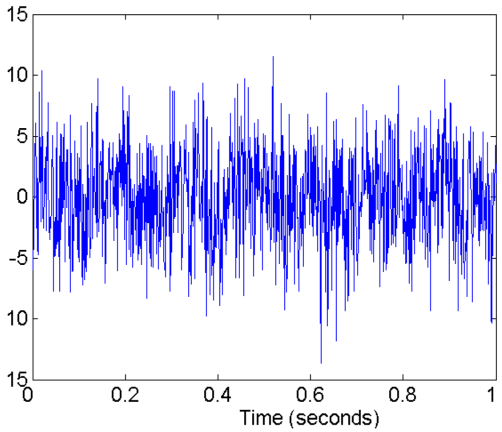


Noise refers to modification of original values

Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen



Two Sine Waves



Two Sine Waves + Noise

Data Analysis

Exploratory Data Analysis (EDA)

Exploratory data analysis techniques are designed to for **open-minded exploration** and not guided by a research question. EDA should not be thought of as an exhaustive set of steps to be strictly followed but rather a mindset or philosophy the analyst brings with her to guide her exploration.

The analyst uses EDA techniques to “tease out” the underlying structure of the data and manipulate it in ways that will reveal otherwise hidden patterns, relationships and features. EDA techniques are **primarily graphical** because humans have innate pattern recognition abilities which we utilize to synthesize complex conclusions from visual cues. [1]

[1] Völkl, T. (n.d.). *Initial Data Analysis (IDA) & Exploratory Data Analysis (EDA) – Method*. etav.github.io. Retrieved April 30, 2025, from https://etav.github.io/articles/ida_eda_method.html

Data Analysis

Exploratory Data Analysis (EDA) / Univariate vs. Bivariate / Multivariate Analysis

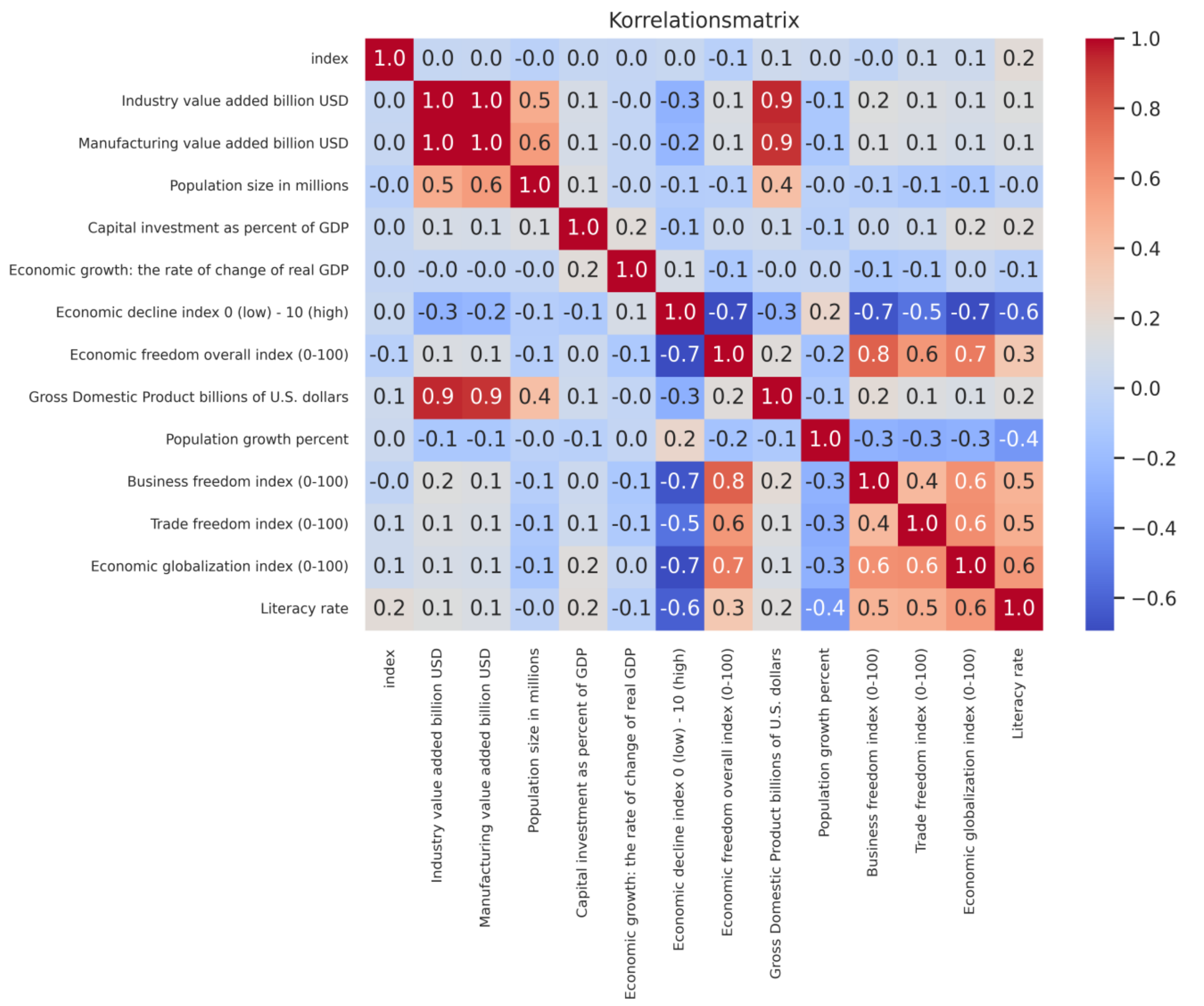
	Univariate Analysis	**Bivariate / Multivariate Analysis**
Purpose	Examine single variables individually and understand their distribution	Analyze relationships between two or more variables
Typical Questions	What is the distribution? Are there outliers?	Are variables correlated? Are there patterns or associations?
Common Tools	▪ Histograms ▪ Boxplots	▪ Scatter Plots ▪ Pair Plots ▪ Correlation Heatmaps
Usefulness	Assess transformations or effects of imputations	Detect relationships, multicollinearity, or confounding variables (relevant for modeling)
Python Tools	seaborn.histplot(), matplotlib.pyplot.hist()	seaborn.scatterplot(), seaborn.pairplot(), seaborn.heatmap()

Data Analysis

Exploratory Data Analysis (EDA)

Correlation heatmap

- Identify strong relationships: Easily spot variables with high positive or negative correlation.
- Detect multicollinearity: Helps to identify redundant features that may affect models.
- Simplify feature selection: Supports decisions on which variables to keep or drop.
- Get a quick overview: Offers a compact visual summary of all pairwise correlations in the dataset.

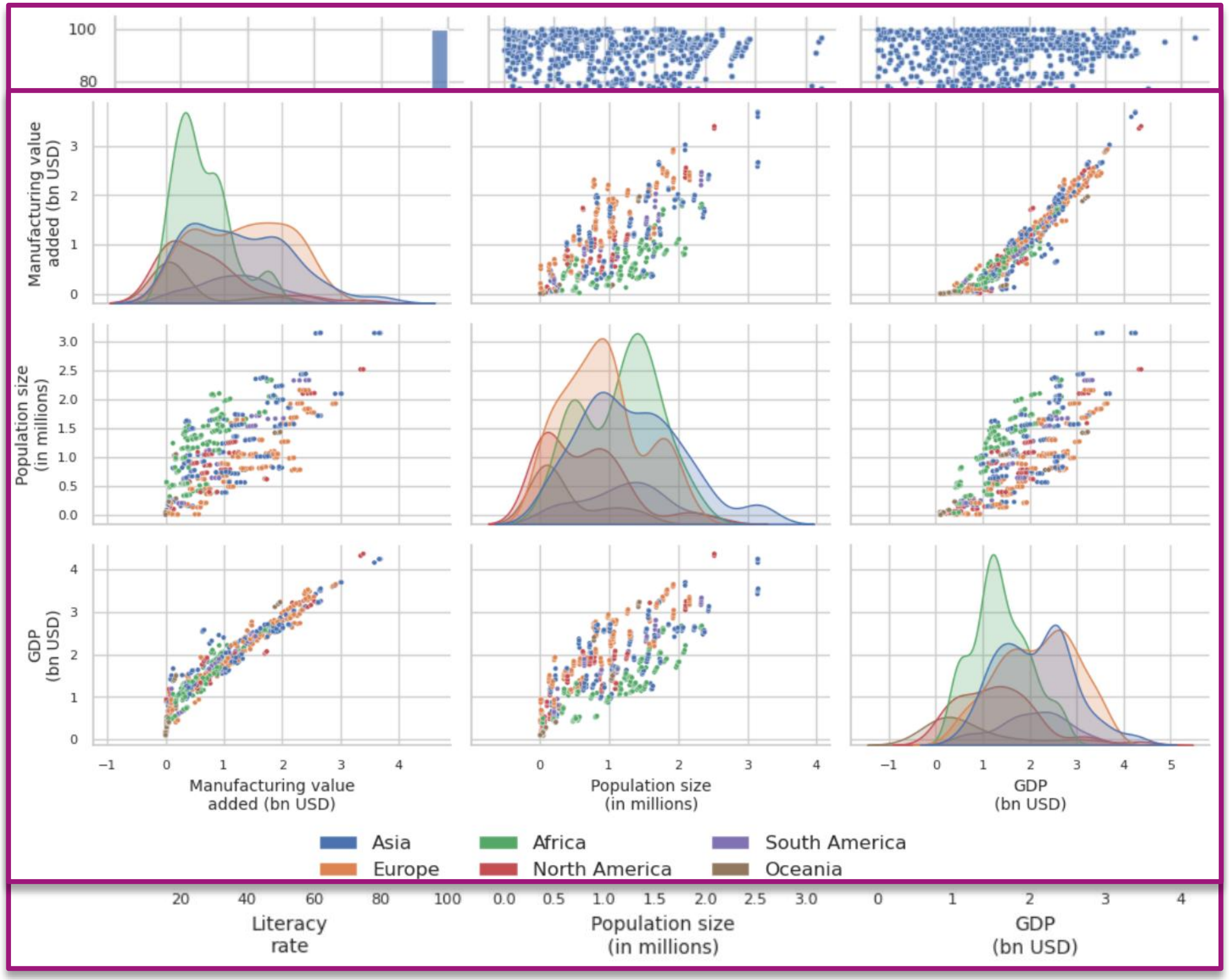


Data Analysis

Exploratory Data Analysis (EDA)

Pairplot

- Recognize correlations: Visualize linear and non-linear relationships between variables.
- See distributions: Individual distributions of each variable (e.g. skewness, outliers, multi-peakedness).
- Scatter & cluster: Indications of natural groups or separability in scatterplots.
- Make class differences visible: With hue, differences between categories can be visualized.

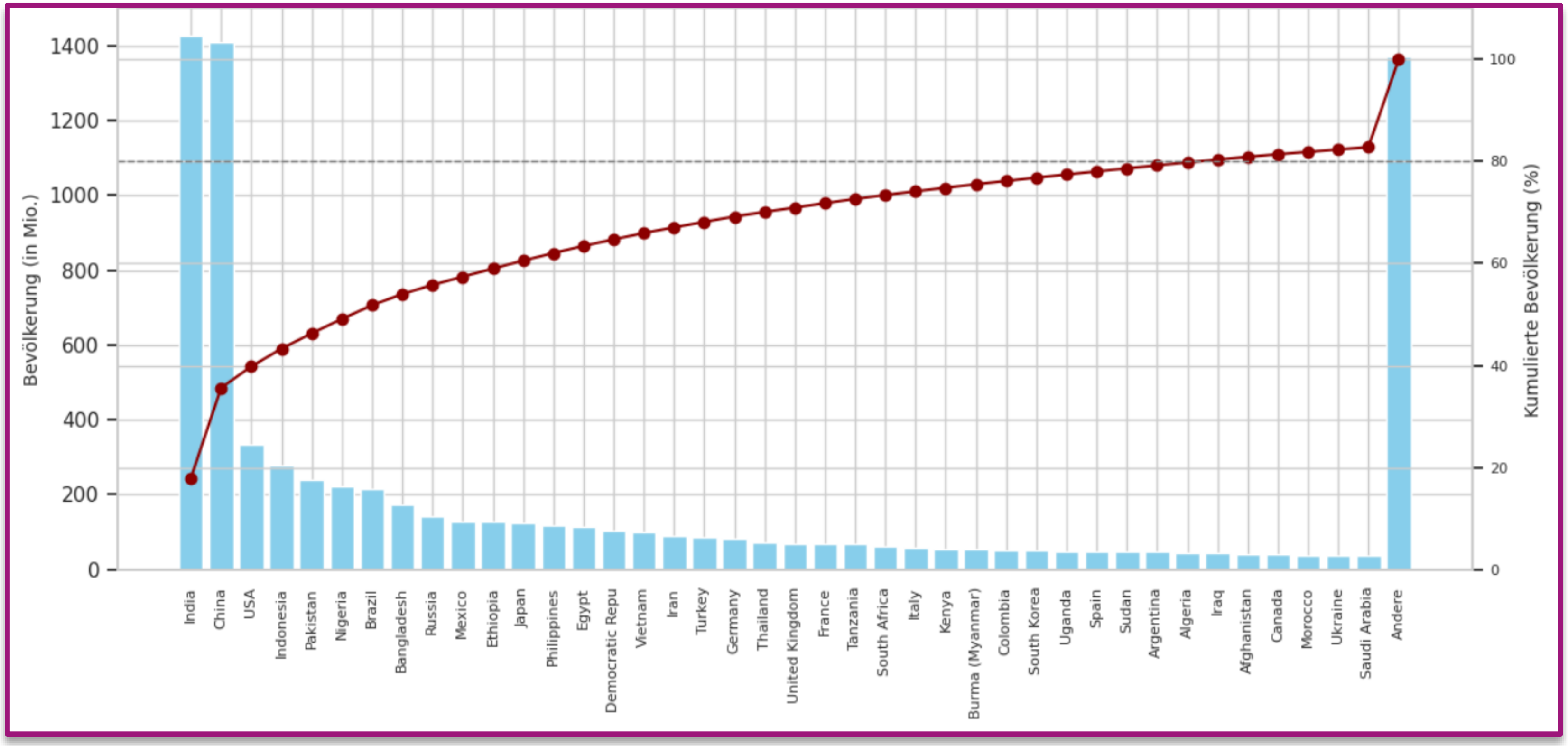


Data Analysis

Exploratory Data Analysis (EDA)

Paretoplot

- Shows top contributors clearly by sorting values.
- Reveals 80/20 patterns (Pareto principle).
- Helps prioritize efforts and focus areas.
- Improves clarity with descending bars and cumulative line.



Zühlke & Bader
Data Mining SS25
Fakultät für Informatik und Ingenieurwissenschaften
Institut für Data Science, Engineering, and Analytics

Data Analysis

Exploratory Data Analysis (EDA)

- **Visualization overview: [Link]**

Data Visualization

Zühlke & Bader

Fakultät für Informatik und Ingenieurwissenschaften
Institut für Data Science, Engineering, and Analytics

Data Visualization

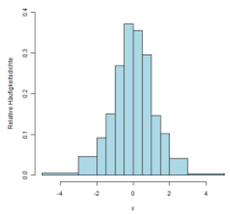
Data Visualization in the CRISP-DM

Data Understanding

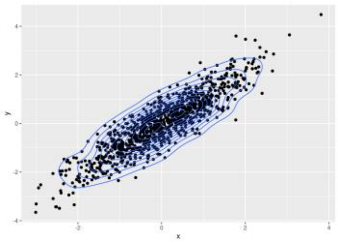
Use Cases:

- Overview of the data structure (e.g. distributions, correlations, gaps)
- Recognize anomalies (outliers, incorrect entries) (IDA)
- Forming initial hypotheses through exploratory data analysis (EDA)

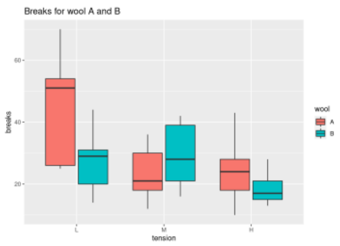
Typical visualization methods:



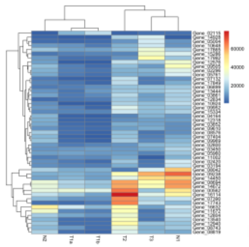
Histogramm



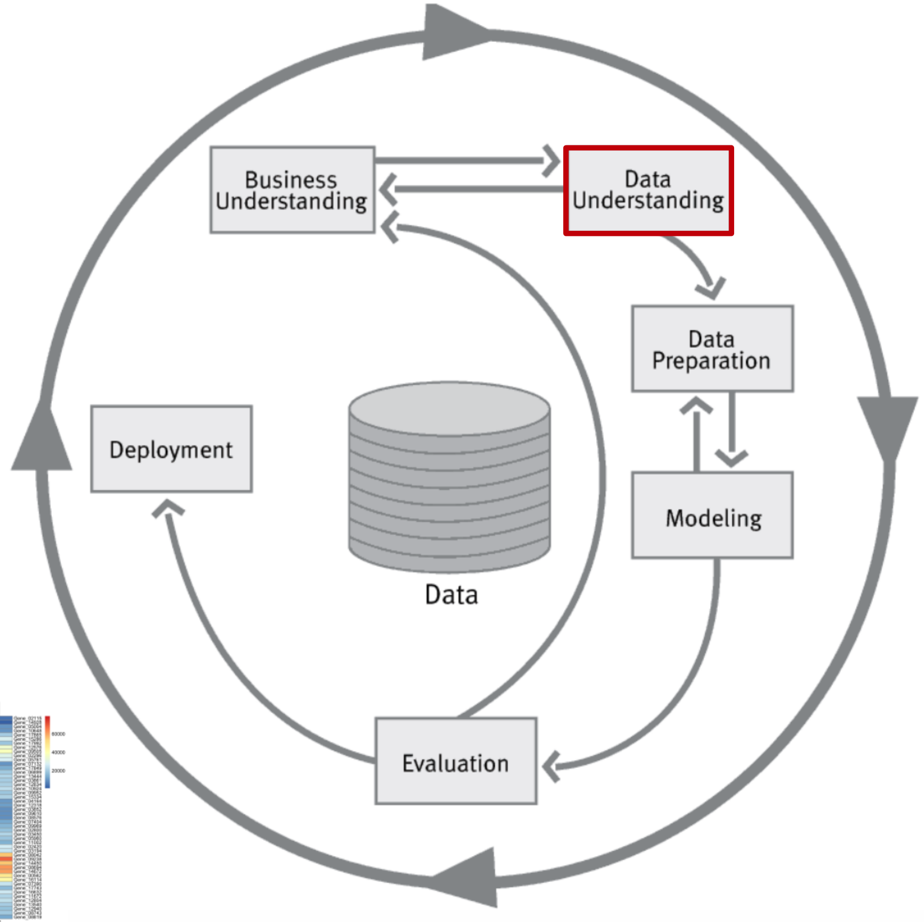
Scatterplot



Boxplots

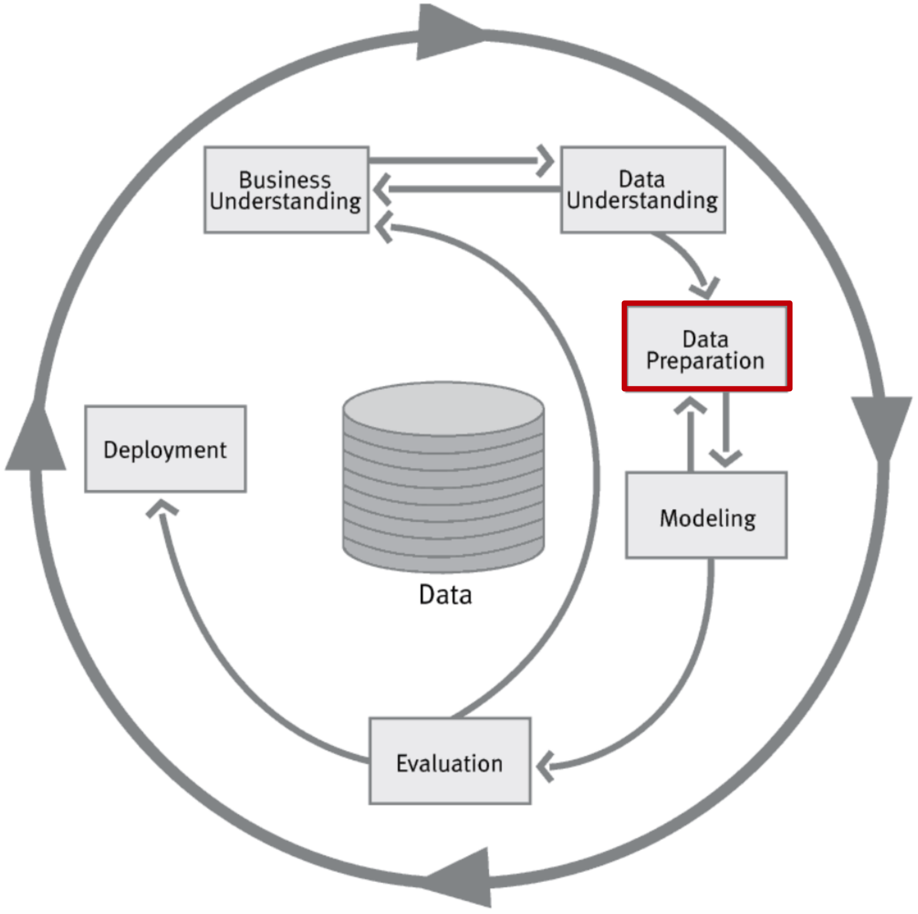


Heatmaps



Data Visualization

Data Visualization in the CRISP-DM

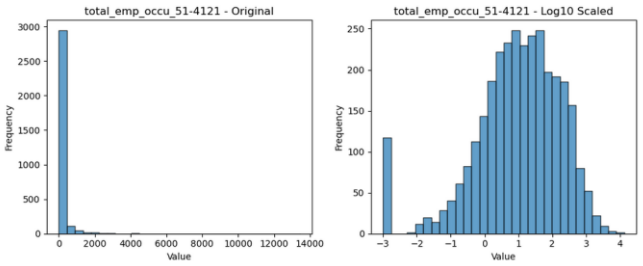


Data Preparation

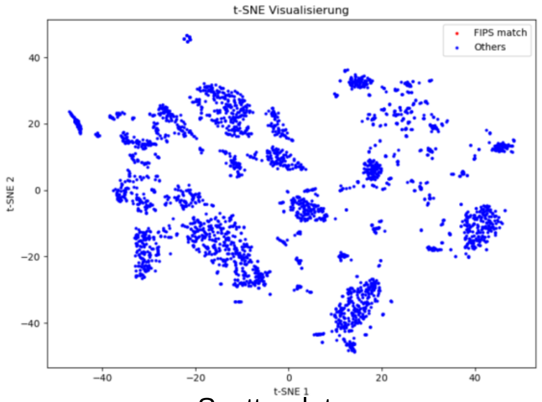
Use Cases:

- Visual control of transformations:
  - Before and after comparisons
  - Ensure that scaling, coding, feature engineering, etc. have been carried out correctly
  - Cluster or dimension reductions can be evaluated visually (e.g. PCA, t-SNE, …)

Typical visualization methods:



Histogram



Scatterplots

Data Visualization

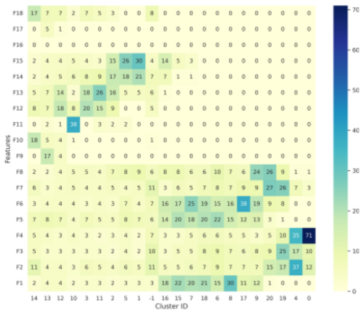
Data Visualization in the CRISP-DM

Modeling

Use Cases:

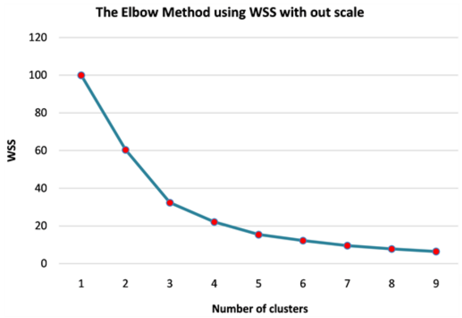
- Feature imports (e.g. with random forest)
- Model behavior in the scatter plot
- Visualization of decision boundaries
- Hyperparameter tuning can be supported by heat maps or line plots

Typical visualization methods:

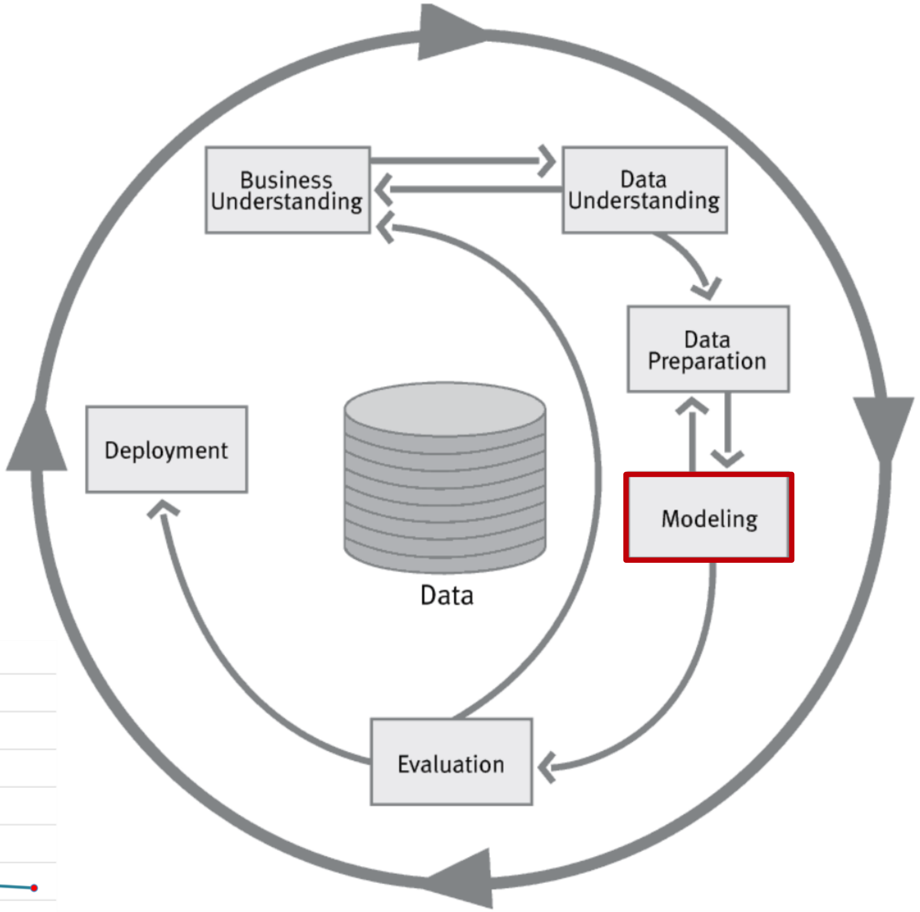


Heatmap

Scatterplot



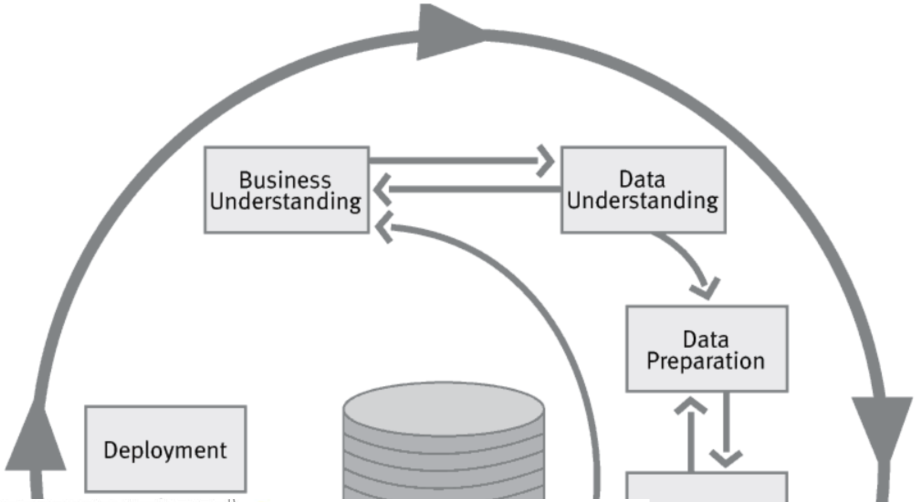
Lineplot



Zühlke & Bader
Data Mining SS25
Fakultät für Informatik und Ingenieurwissenschaften
Institut für Data Science, Engineering, and Analytics
Technology Arts Sciences
TH Köln

Data Visualization

Data Visualization in the CRISP-DM

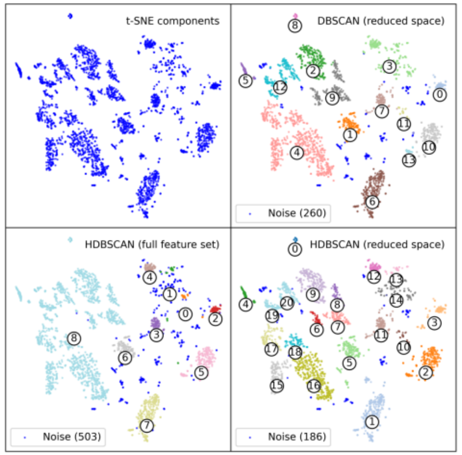


Modeling

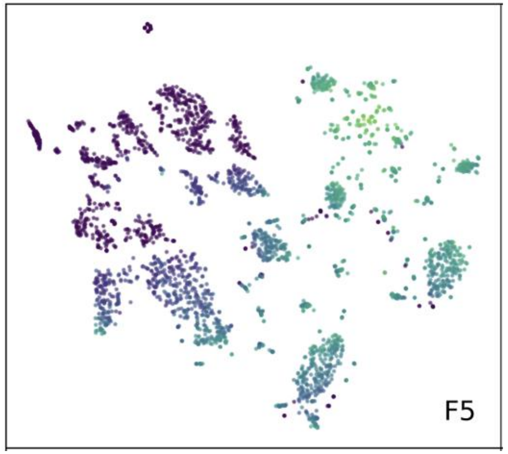
Use Cases:

- Visualization of specific model-dependent metrics
- Comparison of several models (e.g. DBSCAN vs. K-Means vs. HDBSCAN)
- Visual Interactive Evaluation: Domain experts can validate visual clusters/results

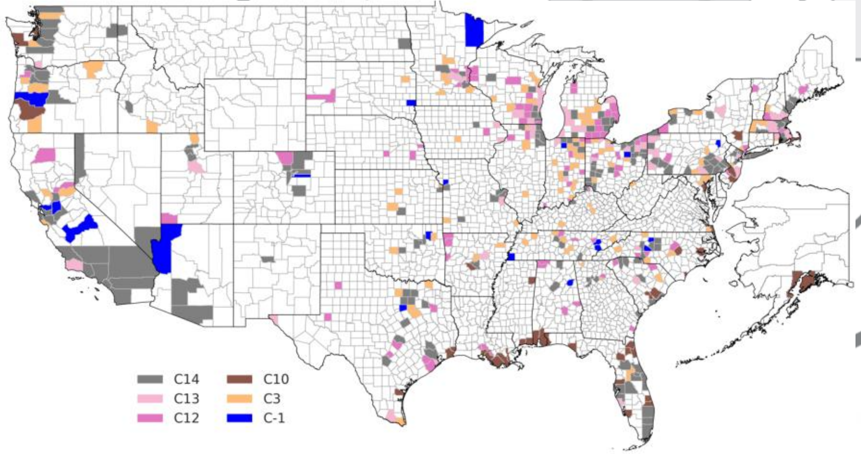
Typical visualization methods:



Scatterplot



Scatterplot



Geographic Mapplot

Data Visualization

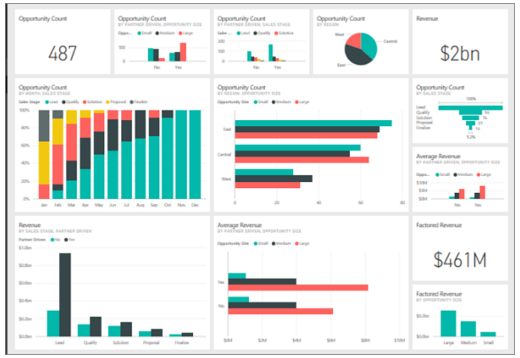
Data Visualization in the CRISP-DM

Deployment

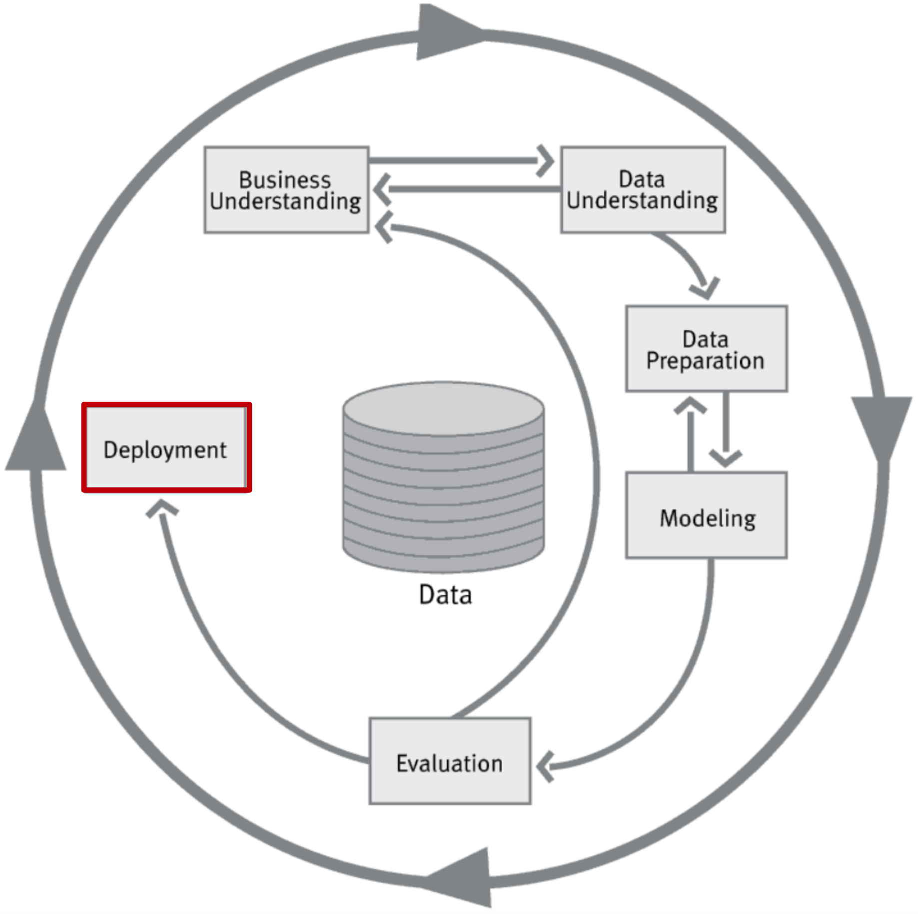
Use Cases:

- Visualization for stakeholders or monitoring:
  - Dashboards for presenting results (e.g. in Power BI, Tableau, Streamlit)
  - Visualization of KPIs
  - Regular monitoring of model performance (drift detection, performance over time)

Typical visualization methods:



Dashboards



Zühlke & Bader
Data Mining SS25
Fakultät für Informatik und Ingenieurwissenschaften
Institut für Data Science, Engineering, and Analytics

Additional Resources

Additional Resources

Inspiring Examples:

- PBS Digital Studios. The Art of Data Visualization | Off Book_ [video]. YouTube, 2012-11-28 [accessed 2025-04-24]. Available at: [https://www.youtube.com/watch?v=AdSZJzb-aX8]
- Ribecca, Severino. The Data Visualisation Catalogue [online]. 2014 [accessed 2025-04-24]. Available at: [https://datavizcatalogue.com]

Literature:

- KNAFLIC, Cole Nussbaumer. Storytelling with data: A data visualization guide for business professionals. John Wiley & Sons, 2015. (Link)
- Few, Stephen. Information Dashboard Design: Displaying Data for At-a-Glance Monitoring. 2nd ed. Burlingame: Analytics Press, 2013. (Link)
- Bolten, Randall. Painting with Numbers: Presenting Financials and Other Numbers So People Will Understand You. Hoboken: Wiley, 2012. (Link)
- Wexler, Steve; Shaffer, Jeffrey; Cotgreave, Andy. The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios. Hoboken: Wiley, 2017. (Link)

Podcasts:

- Data Stories (Link)
- Storytelling with data podcast (Link)
- Data Viz Today (Link)