Einführung in die Datenanalyse

Introduction to Data Science

Max Heimel, MSc

Prof. Dr. Volker Markl

Fachgebiet Datenbanksysteme und Informationsmanagement
Technische Universität Berlin

http://www.dima.tu-berlin.de/

Course Overview

1. **What is Data Science?**

2. Data: An Overview.

3. Exploratory Data Analysis

What is Data Science? (I)

- *„The extraction of knowledge from data.“*
  -- Wikipedia

- *„A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning.”*
  -- Daniel Tukelang (LinkedIn)

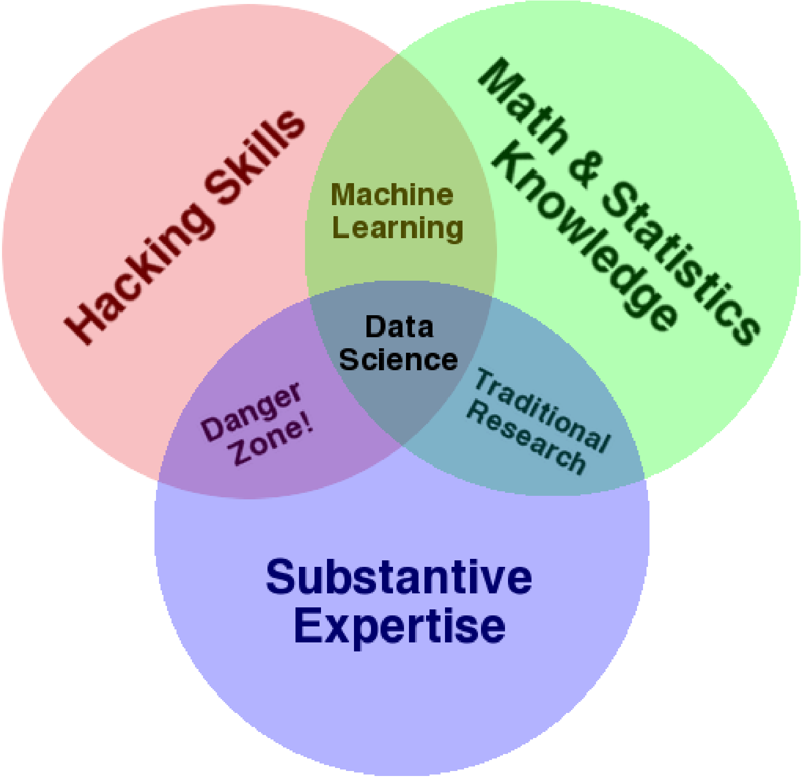
- *„The sexiest career of the 21st century.”*
  -- Harvard Business Review

- *„By 2018 the United States will experience a shortage of 190,000 skilled data scientists.”*
  -- McKinsey

- *„A buzzword without clear definition [that] has simply replaced Business Analytics in contexts such as graduate degree programs.“*
  -- Gil Press (Forbes)

- *„A sexed up term for a statistician.“*
  -- Nate Silver (FiveThirtyEight.com)

What is Data Science? (II)

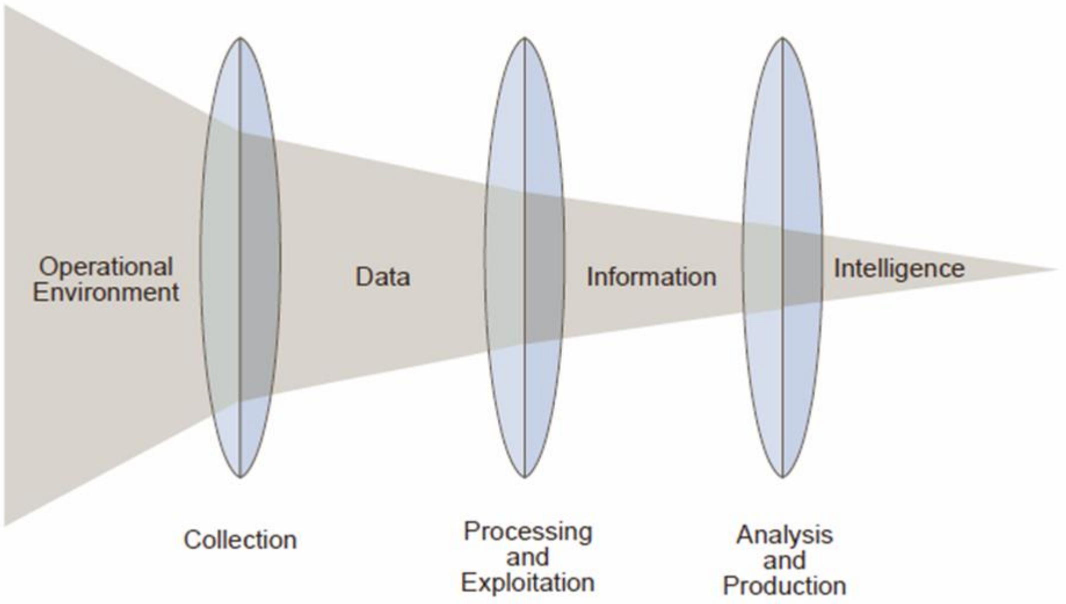


Source: http://www.niemanlab.org/images/drew-conway-data-science-venn-diagram.jpg

What is Data Science? (III)

How to get from here ....

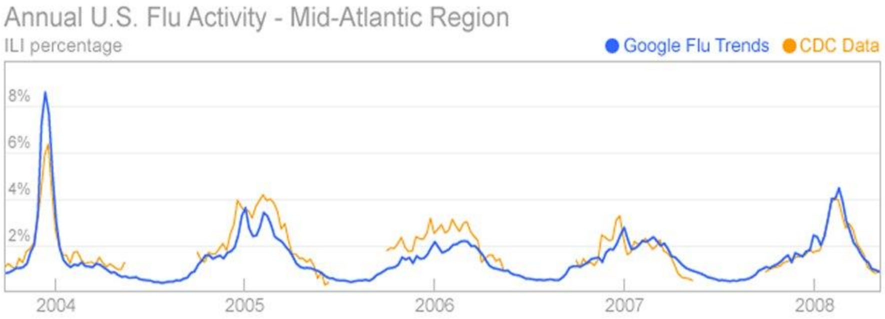
Relationship of Data, Information and Intelligence



... to here.

Source: Joint Intelligence / Joint Publication 2-0 (Joint Chiefs of Staff)

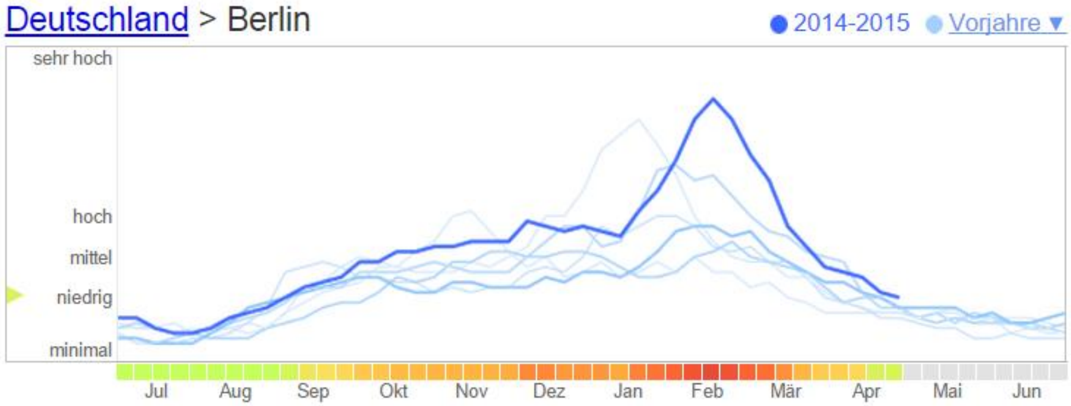
Data Science Success Stories: Google Flu Trends



- In 2008, Data Scientists at Google found that they can predict Flu seasons by monitoring the frequency of Flu-related search terms.
- Today, Google is able to predict an incoming Flu season about two weeks before it arrives!

Google offers its flu-related predictions and observations at:

www.google.org/flutrends/

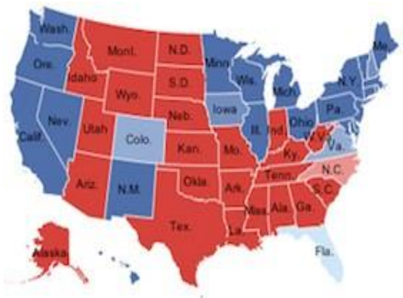


Data Science Success Stories: FiveThirtyEight.com

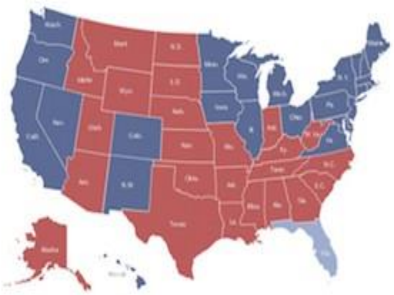
- A thorough statistical model based on polling data enabled Data Analyst & Blogger Nate Silver to accurately predict the outcome of the 2012 US presidential race with 96% accuracy.

- This came as a shock to the „traditional media", who previously called Nate a *„joke"* whose predictions were *„getting into silly land"*.

- Today, he offers statistical predictions for a variety of events from Economics, Sports & Politics at www.fivethirtyeight.com .



Nate Silver's Map



The Actual Map

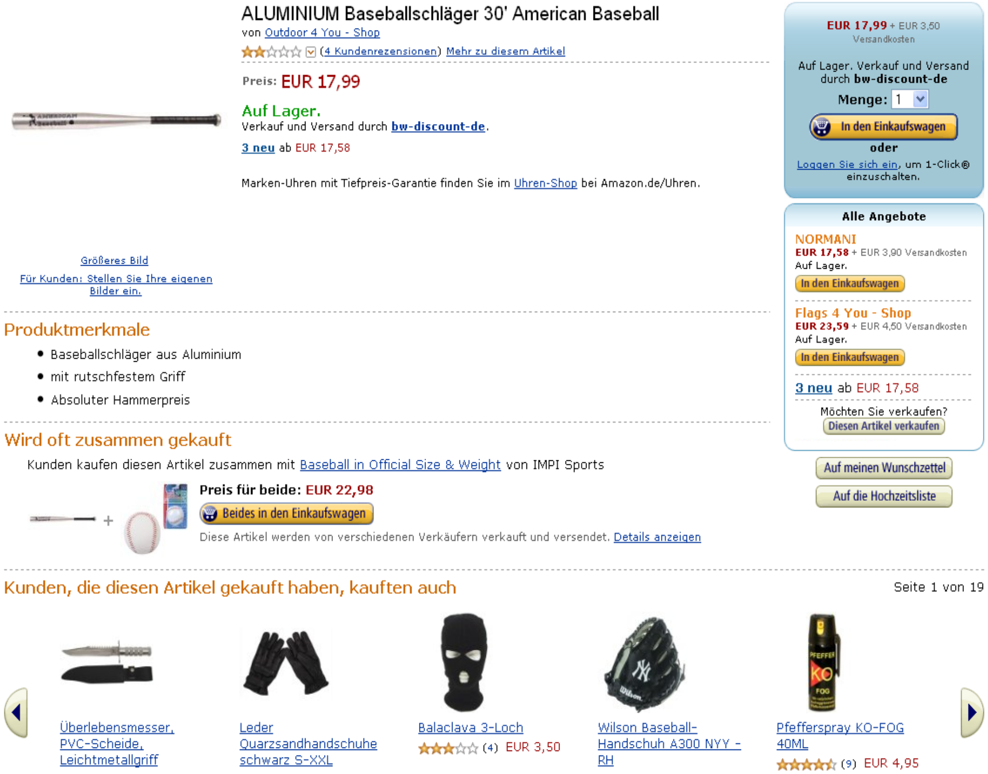
FiveThirtyEight

Data Science Success Stories: Amazon Recommendations

- Amazon is probably the best example for how data analysis can help a business to increase its revenue.

- By monitoring, and modelling the buying behavior of their users, Amazon was able to build a hugely successful product recommendation engine.

- Today, around 20-30% of Amazon‘s revenue can be traced back to product recommendations.



ALUMINIUM Baseballschläger 30' American Baseball
von Outdoor 4 You - Shop
★★☆☆☆ (4 Kundenrezensionen) Mehr zu diesem Artikel

Preis: EUR 17,99

Auf Lager.
Verkauf und Versand durch bw-discount-de.

3 neu ab EUR 17,58

Marken-Uhren mit Tiefpreis-Garantie finden Sie im Uhren-Shop bei Amazon.de/Uhren.

EUR 17,99 + EUR 3,50 Versandkosten

Auf Lager. Verkauf und Versand durch **bw-discount-de**

Menge: 1

oder

Loggen Sie sich ein, um 1-Click® einzuschalten.

Größeres Bild

Für Kunden: Stellen Sie Ihre eigenen Bilder ein.

Alle Angebote

NORMANI
EUR 17,58 + EUR 3,90 Versandkosten
Auf Lager.

Flags 4 You - Shop
EUR 23,59 + EUR 4,50 Versandkosten
Auf Lager.

3 neu ab EUR 17,58

Möchten Sie verkaufen?

Produktmerkmale

- Baseballschläger aus Aluminium
- mit rutschfestem Griff
- Absoluter Hammerpreis

Wird oft zusammen gekauft

Kunden kaufen diesen Artikel zusammen mit Baseball in Official Size & Weight von IMPI Sports

Preis für beide: EUR 22,98

Diese Artikel werden von verschiedenen Verkäufern verkauft und versendet. Details anzeigen

Kunden, die diesen Artikel gekauft haben, kauften auch

Seite 1 von 19

Überlebensmesser, PVC-Scheide, Leichtmetallgriff

Leder Quarzsandhandschuhe schwarz S-XXL

Balaclava 3-Loch
★★★☆☆ (4) EUR 3,50

Wilson Baseball-Handschuh A300 NYY - RH

Pfefferspray KO-FOG 40ML
★★★★☆ (9) EUR 4,95

Data Science Success Stories: IBM Watson

- In 2011, IBM’s Big Data Knowledge System Watson managed to beat human competitors in Jeopardy.

- Watson’s core algorithm utilizes Natural Language Processing, Information Extraction & Statistics to infer knowledge from textual data.

- IBM expects Watson to generate around 100 million USD in revenue, primarily in Healthcare, strategic business consulting & Pharmaceutical Research.

So, is Data everything we need?

DILBERT
BY SCOTT ADAMS
CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
DilbertCartoonist@gmail.com
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
©2012 Scott Adams, Inc./Dist. by Universal Uclick
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.
IN THE PAST, OUR COMPANY DID MANY EVIL THINGS.
BUT IF WE ACCEPT BIG DATA IN OUR SERVERS, WE WILL BE SAVED FROM BANKRUPTCY.
LET US PAY.
$
www.dilbert.com
7-29-12
IS IT TOO LATE TO SIDE WITH EVIL?
SHHHH! IT HEARS YOU.

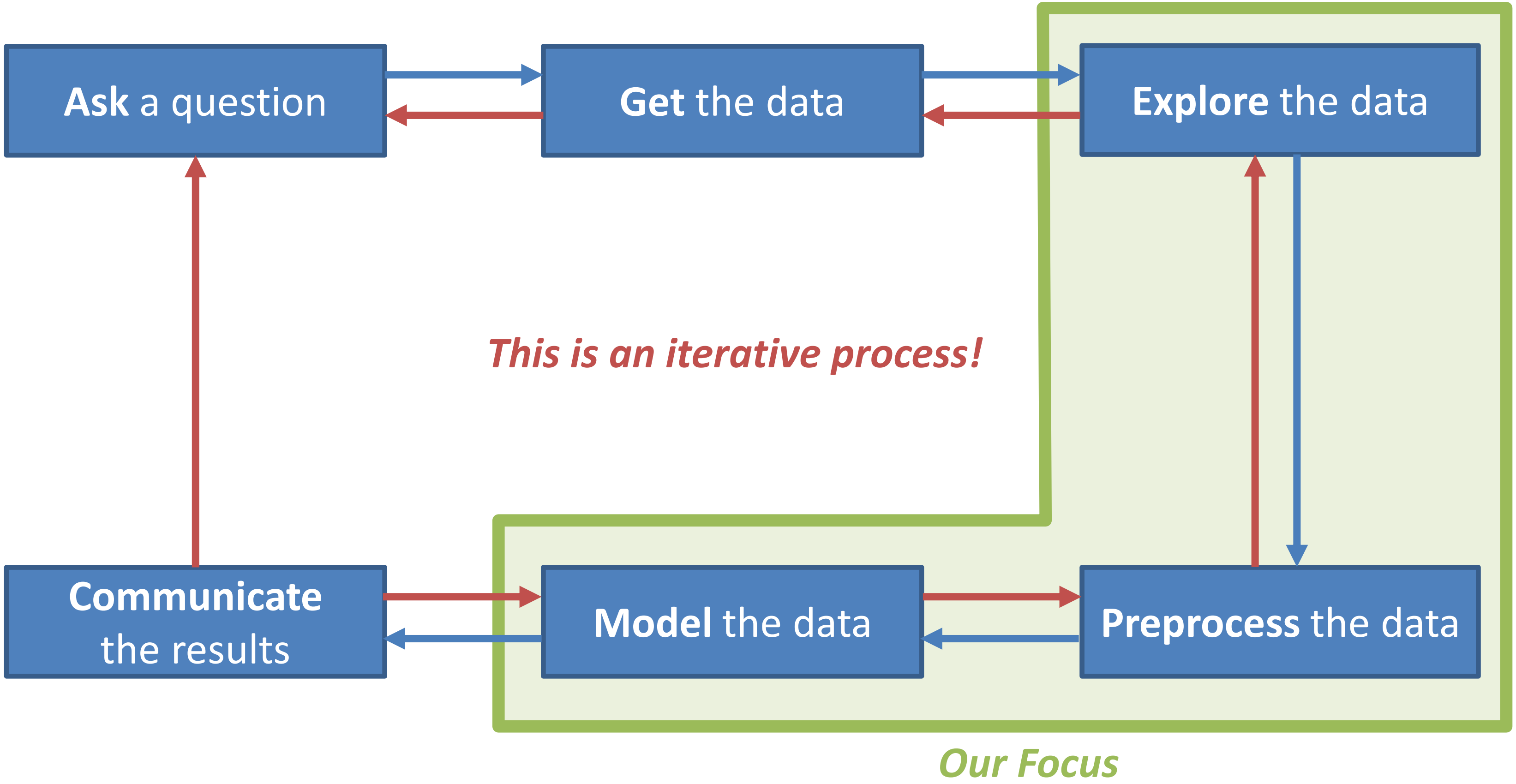
Avoiding Big Data Hubris

- At the moment, Data Science & Big Data are very hip topics.
  - □ Several big companies, research labs & government agencies are successfully applying it.
  - □ **However:** This success has also led to the topics becoming somewhat over-hyped.
  - □ ➔ People often put a lot of trust into results obtained from data analysis ("Big Data Hubris").

- However, always remember: **Above all, data analysis is a tool!**
  - □ It can help to prove assumptions, find new insights, understand problems.
  - □ But it cannot (and should not) replace experimentation, scientific modelling, applied domain knowledge, and (above all) human insight.

- Furthermore: Lying with data & misinterpreting results is incredibly easy!
  - □ Biased data sources, incorrect analysis methods, wrong model assumptions, implementation errors, misunderstood theory, deceiving representations, …
  - □ ➔ Always double-check results & insights coming from data analysis!
    - *"The only statistics you can trust are those you falsified yourself."*

The Data Analysis Process


Ask a question

Get the data

Explore the data

This is an iterative process!

Communicate the results

Model the data

Preprocess the data

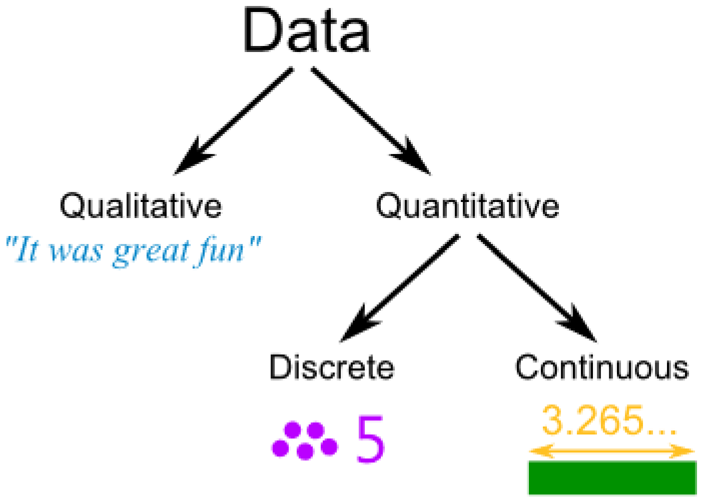
Our Focus

Course Overview

1. What is Data Science?
2. **Data: An Overview**
3. Exploratory Data Analysis

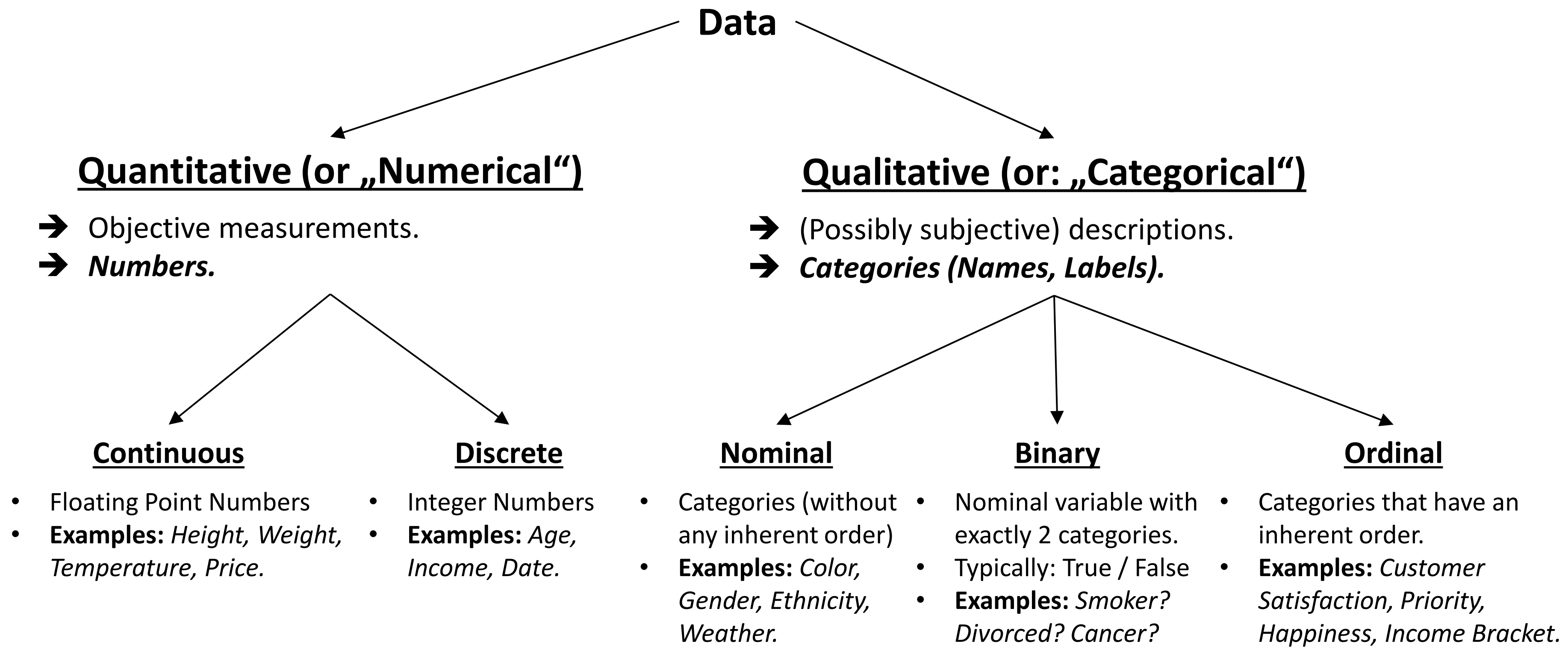
What is Data?

- Wikipedia: *„A set of values of qualitative or quantiative variables“.*



https://www.mathsisfun.com/data/images/data-types.gif

Classifying Variables



Data

Quantitative (or „Numerical“)

➔ Objective measurements.

➔ ***Numbers.***

Continuous

- Floating Point Numbers
- **Examples:** *Height, Weight, Temperature, Price.*

Discrete

- Integer Numbers
- **Examples:** *Age, Income, Date.*

Qualitative (or: „Categorical“)

➔ (Possibly subjective) descriptions.

➔ ***Categories (Names, Labels).***

Nominal

- Categories (without any inherent order)
- **Examples:** *Color, Gender, Ethnicity, Weather.*

Binary

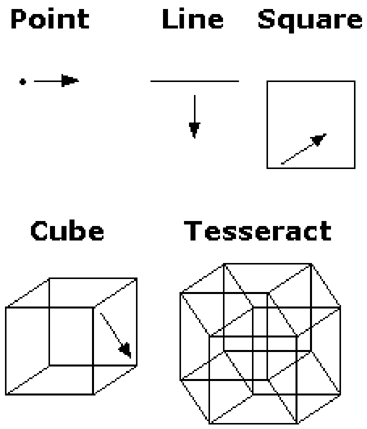
- Nominal variable with exactly 2 categories.
- Typically: True / False
- **Examples:** *Smoker? Divorced? Cancer?*

Ordinal

- Categories that have an inherent order.
- **Examples:** *Customer Satisfaction, Priority, Happiness, Income Bracket.*

Dimensionality

- The number of attributes (columns) in the dataset is called its **dimensionality**.
  - *Univariate data:* One dimension.
  - *Bivariate data:* Two dimensions.
  - *Multivariate data:* More than two dimensions.
    - ➔ This is the typical case!



- Data Analysis often gets very complicated for higher dimensions.
  - “Curse of Dimensionality”
  - Typical approaches: Visualize subspaces, Find structures (clustering), Project data into lower dimensional space (Dimensionality Reduction).

Structured, Unstructured, Semi-Structured

- We distinguish three primary data categories:

1. **Structured** data:
   - Follows a rigid pre-defined schema, consisting of multiple, well-defined variables.
   - **Examples:** *Relational databases (and everything that can be directly mapped to one).*

2. **Unstructured** data:
   - Does not follow any (apparent!) schema.
   - **Examples:** *Text, Images, Videos, Sound, CSV Files without metadata.*

3. **Semi-Structured** data:
   - Schema is encoded within the data (self-describing schema).
   - **Examples:** *JSON, XML.*

Metadata

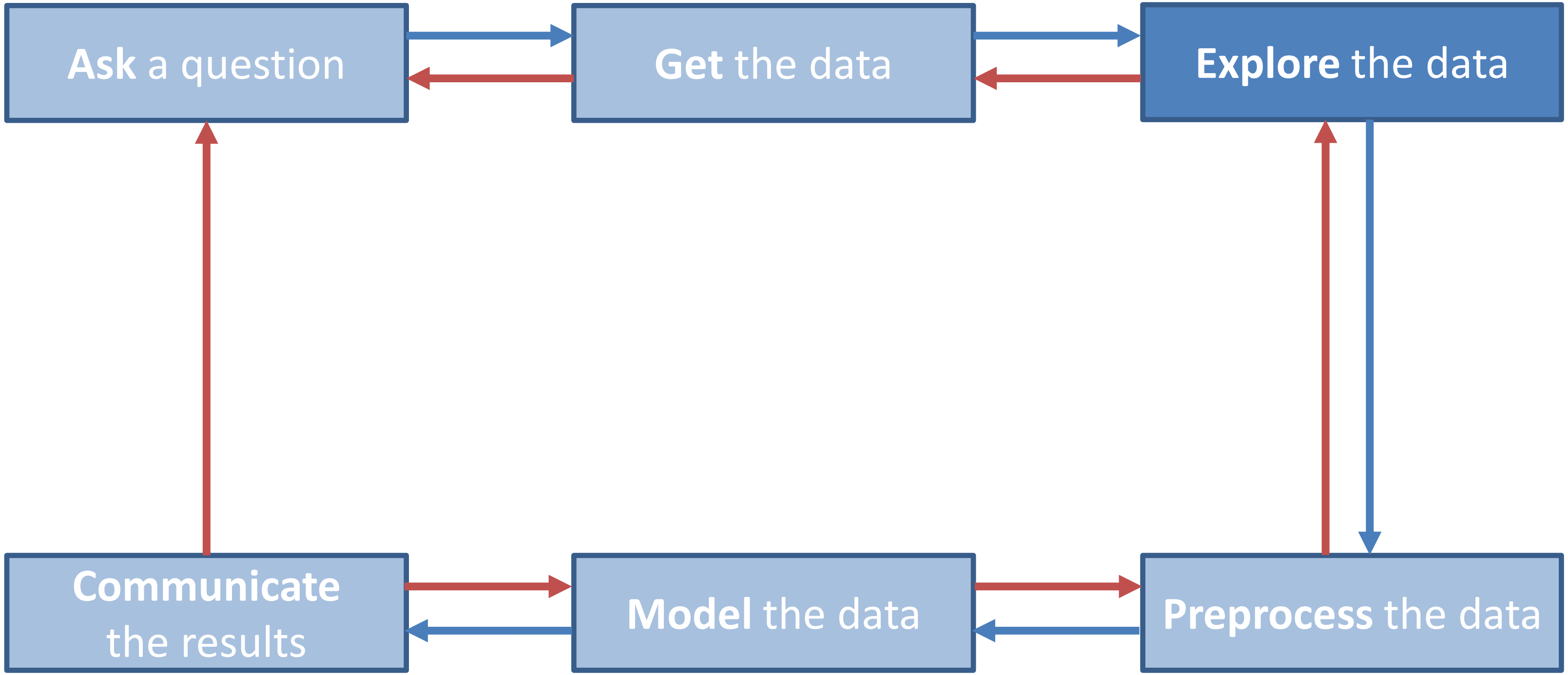
- Metadata is „data about data“ (Wikipedia).
  - Essentially, all information that describe the dataset.

- Some examples:
  - Name & Data type of the columns.
  - Length of the Dataset (# Tuples, Bytecount, …).
  - Lineage information (Author, Data Sources, Experimental Configuration, …).
  - Purpose of the dataset.
  - Statistical information (e.g. Measurement error).
  - Date of Creation / Modification / Last Access.
  - Encoding (Video Codec, UTF-8, …).
  - Access restrictions.
  - …

Course Overview

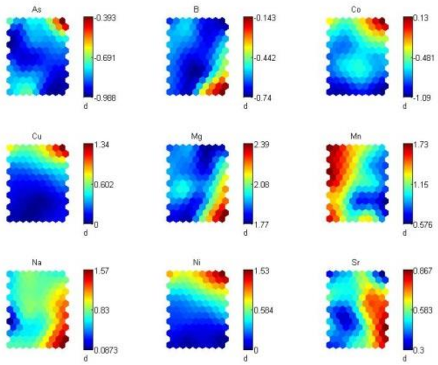
1. What is Data Science?
2. Data: An Overview.
3. **Exploratory Data Analysis**

Reminder: The Data Analysis Process

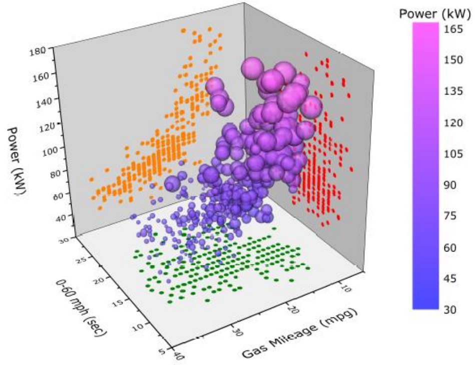


Exploratory Data Analysis

- Before we can perform any serious analysis tasks, we have to understand the data:
  - □ *“Listen”* to the data!
  - □ Investigate what is in the data / how it is structured / what are the interesting parts / are there anomalies / etc.
  - □ ➔ Helps to pick the right analysis methods & avoid costly mistakes.



- Exploratory Data Analysis (EDA):
  - □ „An approach of analyzing data to summarize their main characteristics without using a statistical model or having formulated a prior hypothesis.“
  - □ Done by inspecting & visualizing interesting data aspects.



The Exploratory Data Analysis Process

1. Look at the metadata!
   - □ Is the data structured or unstructured?
   - □ Which attributes are in the data? What are their datatypes?
     Are the attributes quantitative or qualitative?

2. Compute and inspect descriptive statistics for the attributes:
   - □ *Central tendency:* “What is a typical value for the attribute?”
   - □ *Variability measure:* “How are the values spread around the center?”
   - □ *Correlations:* “Do attributes influence each other?”

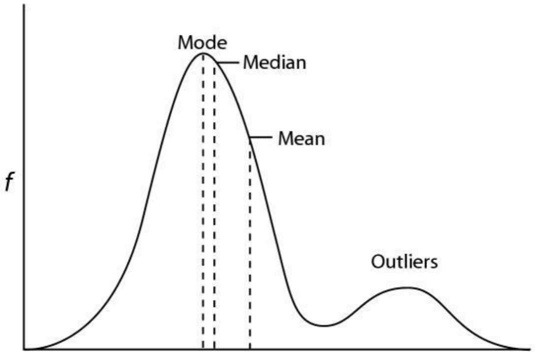
3. Plot data to visualize trends:
   - □ How is the data distributed? Can we see any relationships between attributes?
     Are there outliers or anomalies?

4. Rinse and Repeat:
   - □ While exploring the data, you will gain new insight that can
     be used to refine the process.

Central Tendencies

- Central Tendencies are descriptive statistics to describe the *typical values* of an attribute.

- The three most important tendencies are:
  - □ *Mean:* The *average value* in the attribute.
    - – Typically: Arithmetic Mean.
      - » $\mu_{ari} = \frac{1}{n} \sum_{i=1}^{n} x_i$
    - – Other means are: Weighted, Geometric, Harmonic.

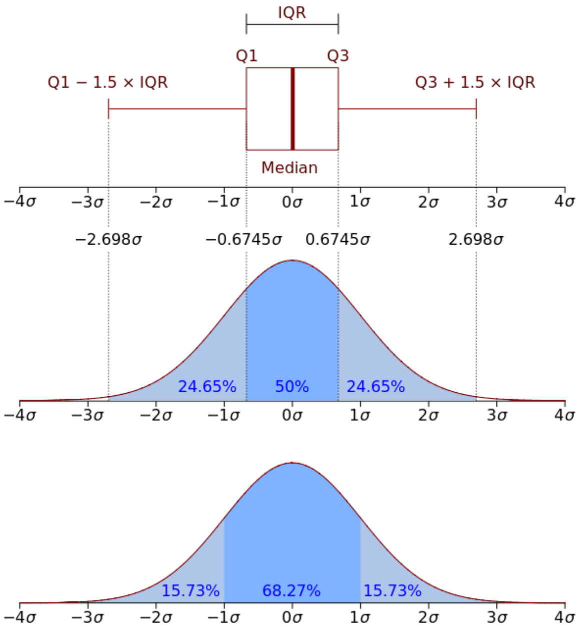


  - □ *Median:* The *middle value* in the attribute (half of all values are larger / smaller).
  - □ *Mode:* The *most common value* in the attribute.
    - – The mode is the only central tendency that is well-defined for qualitative variables.

Variability Measures

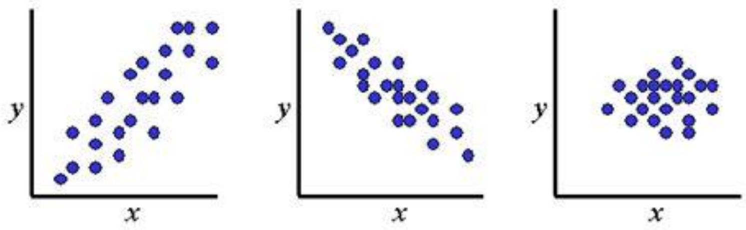
- Variability Measures are descriptive statistics to describe how the data is distributed around the central value.

- The most important variability measures are:
  - □ Range:
    - Difference between largest and smallest values.
  - □ Interquartile Range:
    - Difference between third and first quartile.
    - The three quartiles (Q1, median, Q3) divide the data set into four sets of equal magnitude.
  - □ Standard Deviation :
    - Average distance from the mean.
    - $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\mu - x_i)^2}$



Correlation

- Correlation is the (statistical) dependence between two attributes.
  - □ Roughly: Changes in attribute A also appear in attribute B.

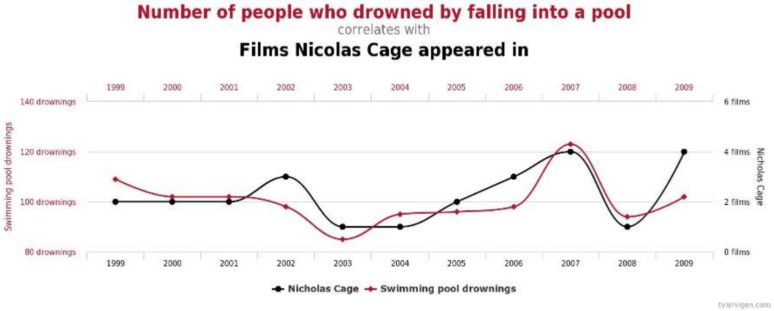


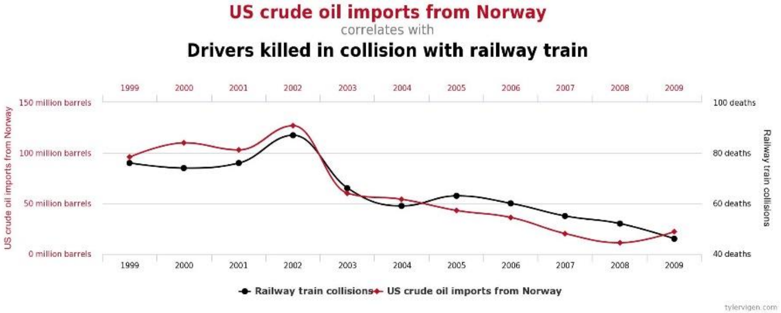
Positive correlation Negative correlation No correlation

- Correlation is typically measured via the Pearson coefficient:
  - □ $Cor(x,y) = \frac{\sum_{i=1}^{n}(x_i-\mu_x)\cdot(y_i-\mu_y)}{\sqrt{\sum_{n}^{i=1}(x_i-\mu_x)^2\cdot(y_i-\mu_y)^2}}$
  - □ Captures linear dependence between the two attributes x and y.

Correlation vs. Causation

- Always remember: **Correlation does not imply causation!**
  - □ Correlation may hint at causation, but you should always verify this externally.
  - □ There are several potential reasons why two variables A and B are correlated:
    - A causes B; B causes A; A causes B and B causes A.
    - A and B are both caused by a different variable C.
    - A causes C, which causes B.
    - Pure coincidence.



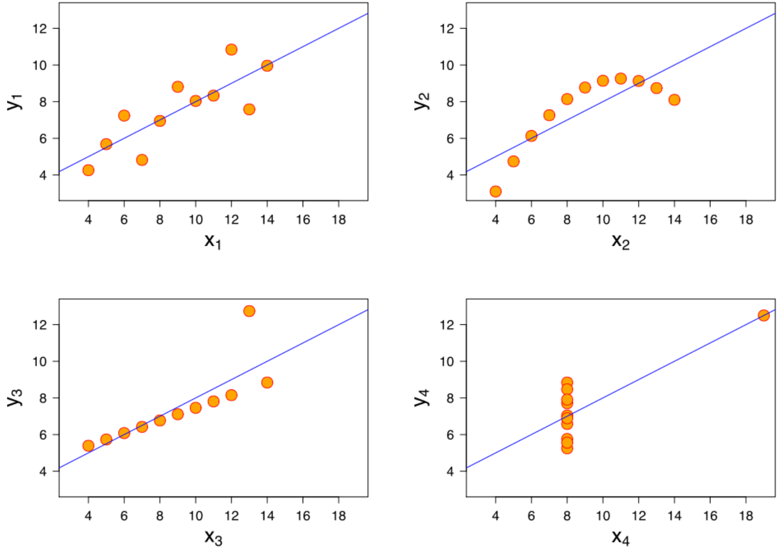


Source: http://www.tylervigen.com/spurious-correlations

Visualization is key!

- Descriptive statistics can give an important first look at the data.
  - □ However, they can be deceiving (and don’t tell the whole picture).

- Example: Anscombe’s quartet.
  - □ Four different, bivariate datasets that have the same:
    - – Average value.
    - – Standard deviation.
    - – Correlation coefficient.
  - □ ➔ We can see the differences only by visualizing the datasets!



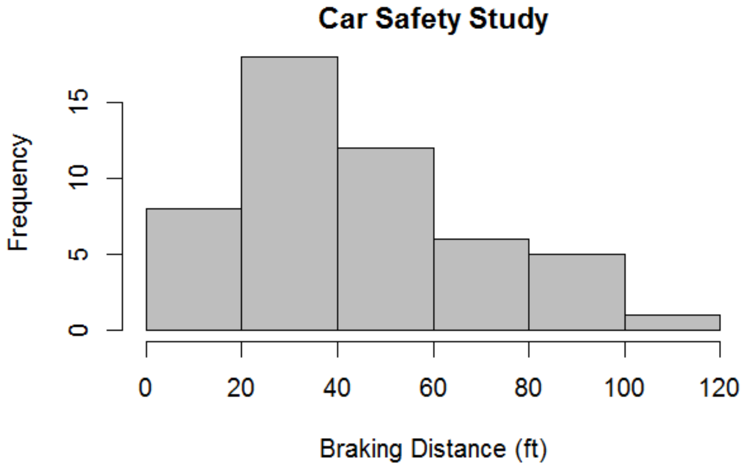
- “A picture is worth a 1000 words!”

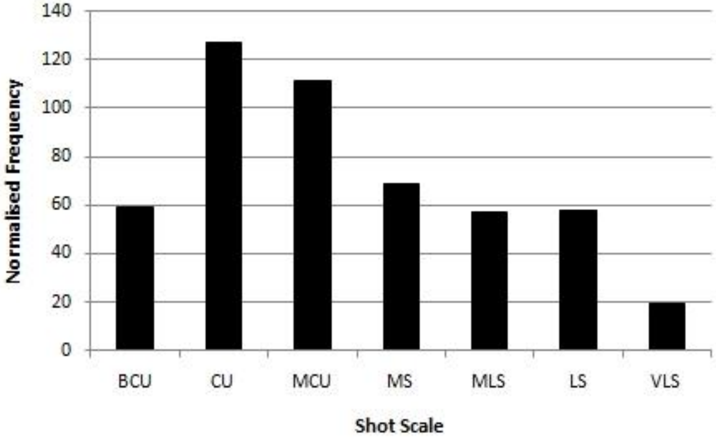
Visualization methods

- The choice of visualization method depends on the dimensionality.
- Univariate (Single Attributes):
  - Histograms.
  - Boxplots.
- Bivariate (Two Attributes):
  - Scatterplots.
- Multivariate (Multiple Attributes):
  - Scatterplots for 3D-data.
  - Scattermatrix.
  - Parallel Coordinate Plots.

Histograms

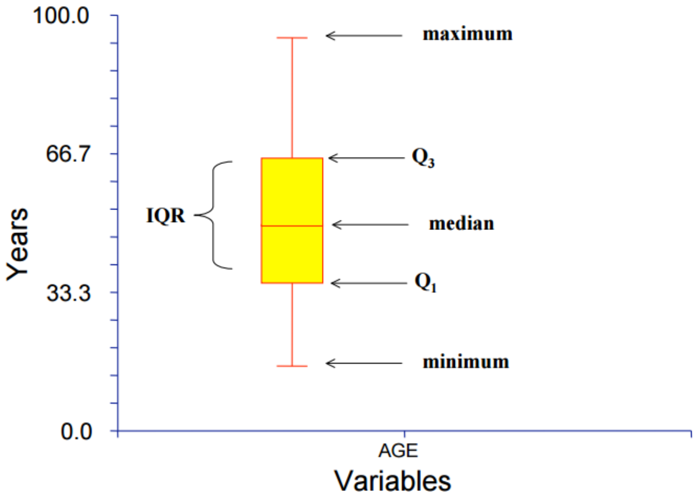
- Histograms are a quick way to visualize the data distribution of univariate qualitative and quantitative attributes:
  - □ *Qualitative attributes:* Count (or frequency) per distinct value.
  - □ *Quantitative attributes:* Discretization (binning) of neighboring values, then count the frequency count per bin.

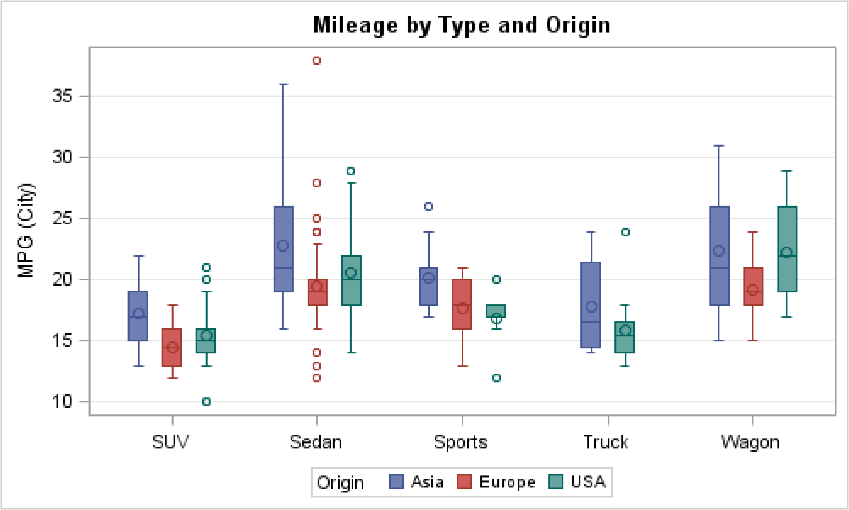




Boxplots

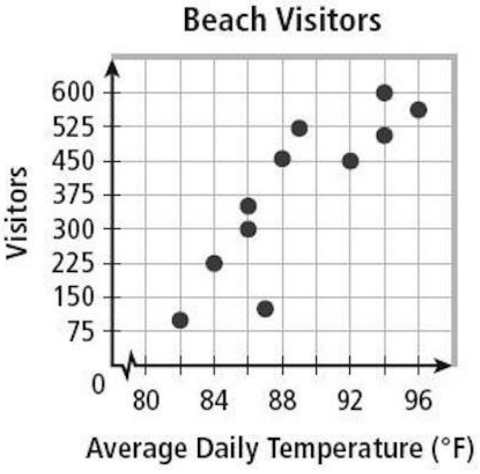
- Boxplots are a compact representation of important descriptive statistics for univariate quantitative attributes.
  - Typically: Median, first & third quartile, minimum & maximum, (outliers).
  - Boxplots can also visualize dependencies between a quantitative variable and one or two qualitative ones by grouping the data according to their labels.

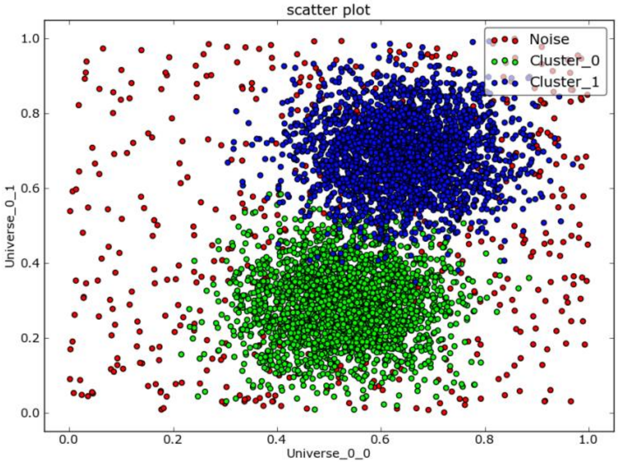


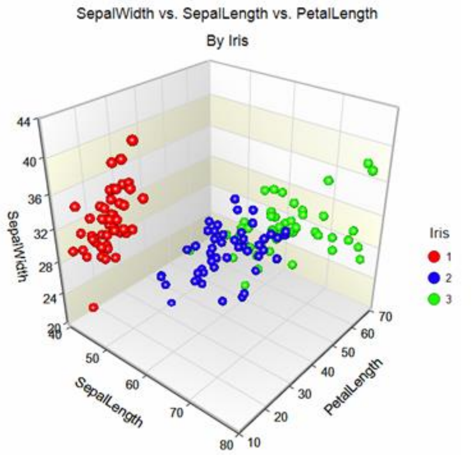


Scatterplots

- Scatterplots can be used to visualize the correlation & relationship between two (or three) quantitative attributes:
  - □ The attribute values are interpreted as (x,y)-coordinates and then drawn as points in a Cartesian coordinate system.
  - □ Additional coloring based on the label can be used to visualize dependencies on a qualitative attribute.

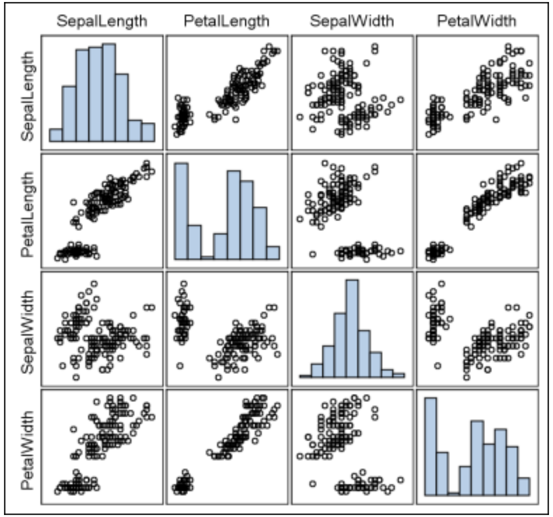


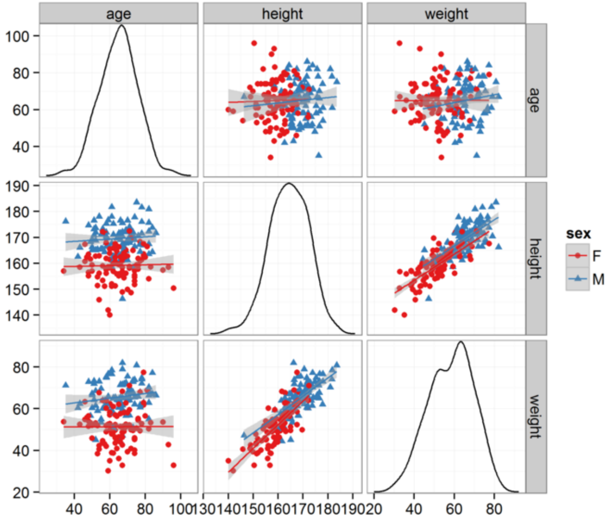




Scattermatrix

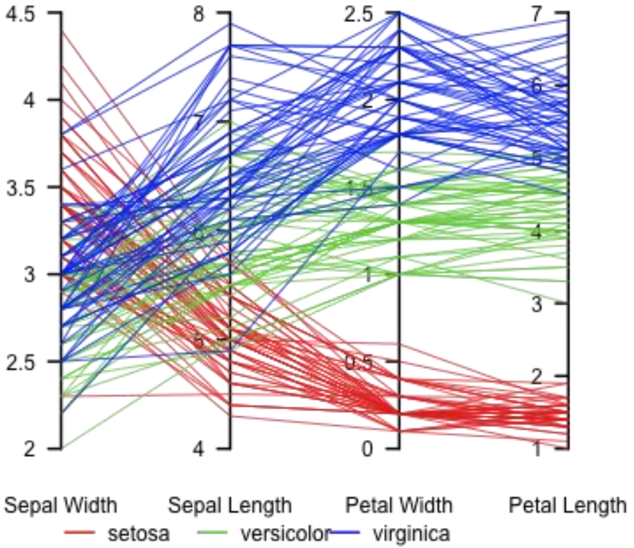
- For multivariate data (> three dimensions) a Scattermatrix can be used to visualize all pairwise correlations (relationships):
  - □ Draw all pairwise Scatterplots, align them in a grid according to the attributes.
  - □ Diagonal typically features Histograms or Density Plots for the single attributes.
  - □ Coloring based on label visualize dependence on qualitative attribute.

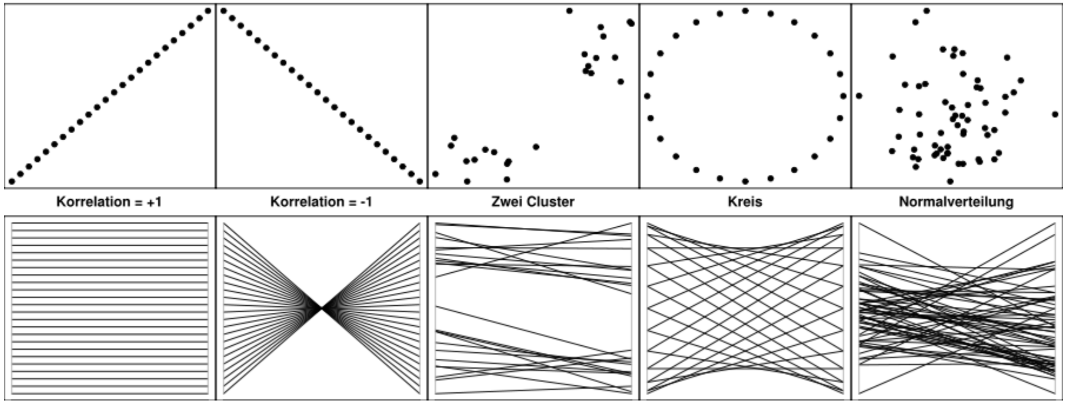




Parallel Coordinate Plots

- Visualizes multivariate data (both continuous and discrete) by aligning the attribute axes in parallel (rather than perpendicular in the scatter plot).
  - □ Points are interpreted as coordinates and illustrated as lines between the axes.
  - □ Can visualize very high-dimensional data. However: Ordering of the axes is important!





Outlook & Overview

- Today we discussed:
  - □ What is Data Science?
  - □ What is the Data Analysis Process?
  - □ How can we classify Data?
  - □ What is Exploratory Data Analysis?
  - □ What are the important statistical measures?
  - □ How can we visualize interesting data aspects?

- Next week:
  - □ Introduction to Machine Learning.
  - □ Machine Learning Methods for Data Analysis.