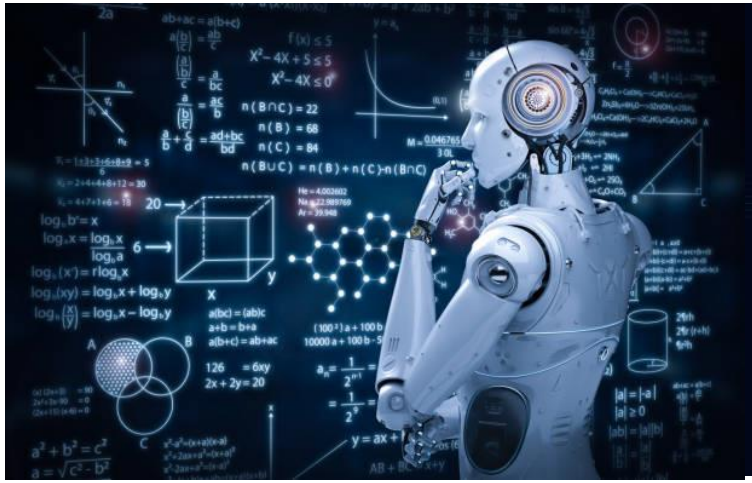
Welcome…

- Introduction to Data Science
- Process of Data Science Projects

CS 797Q
Fall 2024

08/26/2024



Outline

- Introduction to Big Data
- Data Science and Business Intelligence
- The Skillset of Data Scientists
- Summary

What is „Big Data“?!?

Naive Definition

- Naive definition:
  - Big data only depends on the data size
  - 1 Gigabyte? 1 Terabyte? 1 Petabyte?

- Naive interpretation misses important aspects
  - Time:
    - Analyzing 1 Gigabyte of data per day is different from analyzing 1 Gigabyte of data per second
  - Diversity:
    - Analyzing spread sheets with numeric data is different from analyzing Web pages that contain a mixture of text and images
  - Distribution:
    - Analyzing data from a single source is different from analyzing data from multiple sources

Definition of Big Data

- Following Gartner‘s IT Glossary:
  - Big data is high-**volume**, high-**velocity** and/or high-**variety** information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

- The three Vs
  - Volume
  - Velocity
  - Variety

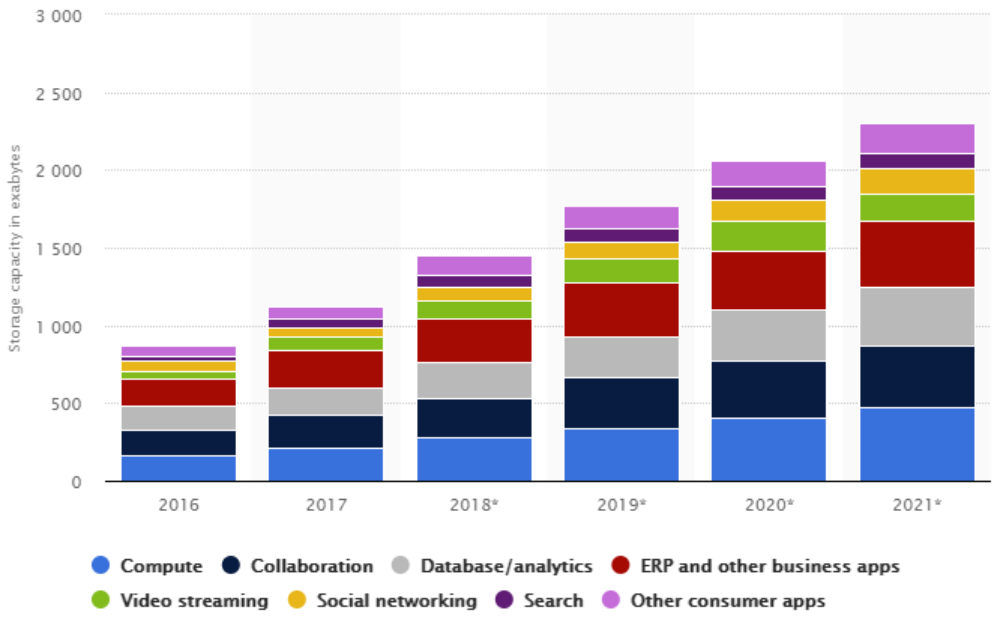
Some people actually use 10 Vs to define big data!

- Variability
- Veracity
- Validity
- Vulnerability
- Volatility
- Visualization
- Value

The 3 Vs: Volume

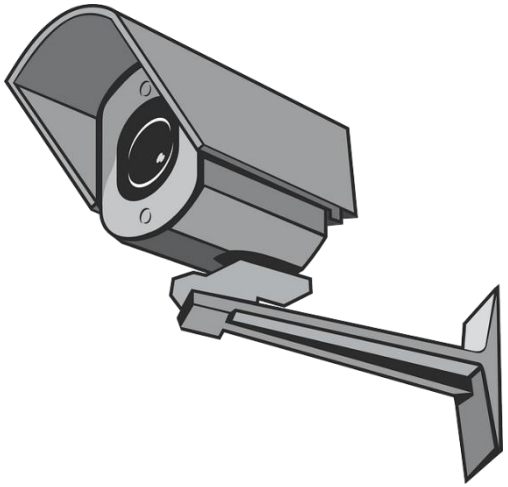
- Scale of the data must be „big“
  - No clear definition
  - „that demand […] innovative forms of information processing“ (Gartner)

Data center storage worldwide



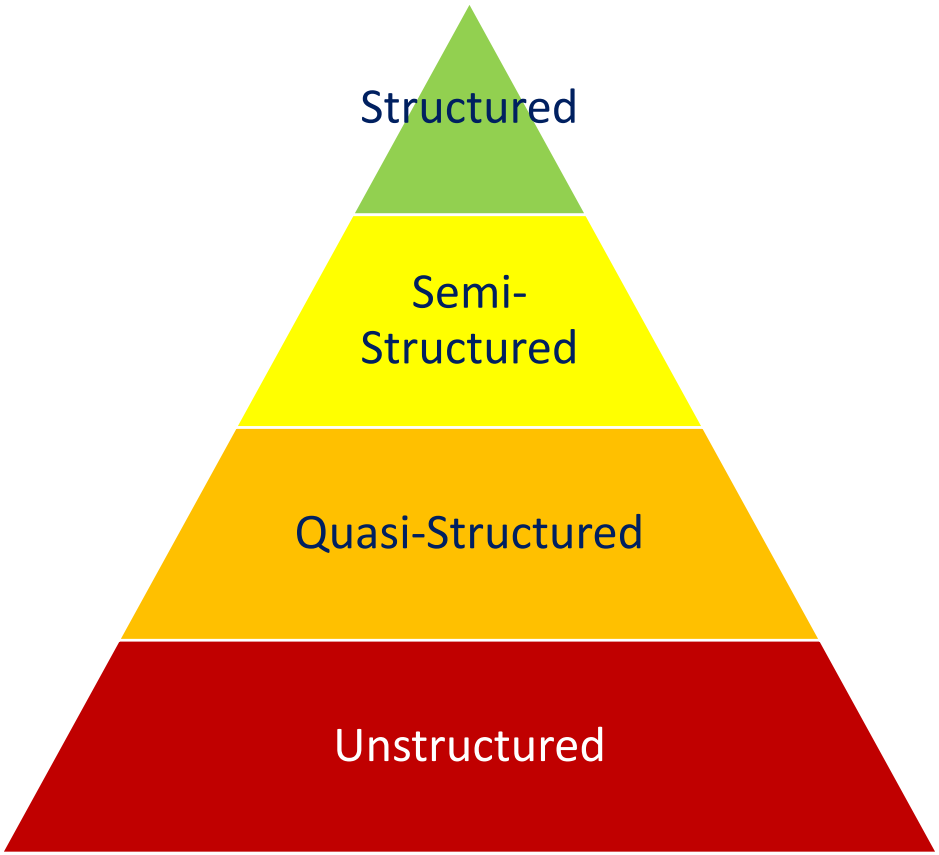
The 3 Vs: Velocity

- Speed at which new data is created
- Speed at which data must be processed and analyzed
  - Often close to real-time



The 3 Vs: Variety

- Diversity in data types and data sources



- Data with defined types and structure
- Example: comma separated values

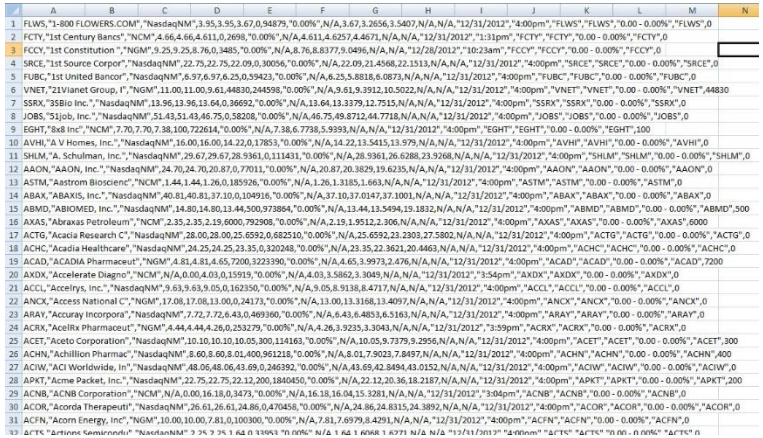
- Textual data with parseable pattern
- Example: XML files with schema

- Textual data with erratic formats that can be formated with effort
- Example: Clickstream data

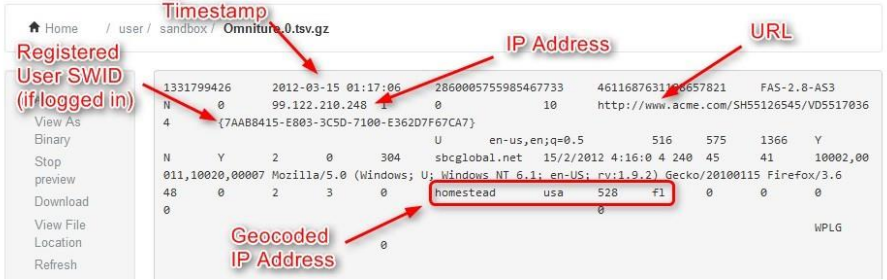
- Data that has no inherent structure, often with multiple formats
- Example: Web site, videos

Examples for data types

Structured



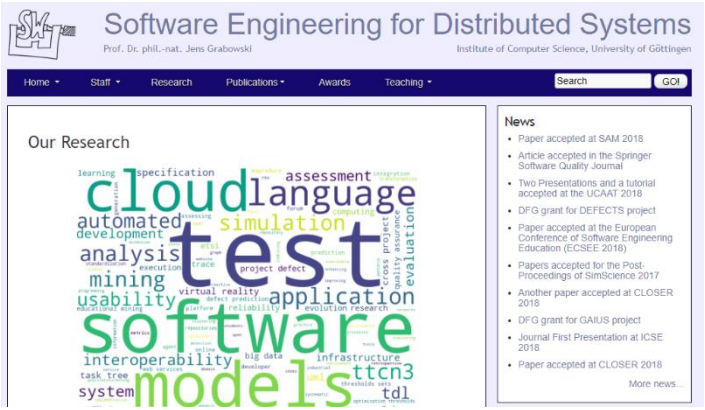
Quasi-Structured



Semi-Structured

```
<?xml version="1.0" encoding="iso-8859-8" standalone="yes" ?>
<CURRENCIES>
  <LAST_UPDATE>2004-07-29</LAST_UPDATE>
  <CURRENCY>
    <NAME>dollar</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>USD</CURRENCYCODE>
    <COUNTRY>USA</COUNTRY>
    <RATE>4.527</RATE>
    <CHANGE>0.044</CHANGE>
  </CURRENCY>
  <CURRENCY>
    <NAME>euro</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>EUR</CURRENCYCODE>
    <COUNTRY>European Monetary Union</COUNTRY>
    <RATE>5.4417</RATE>
    <CHANGE>-0.013</CHANGE>
  </CURRENCY>
</CURRENCIES>
```

Unstructured

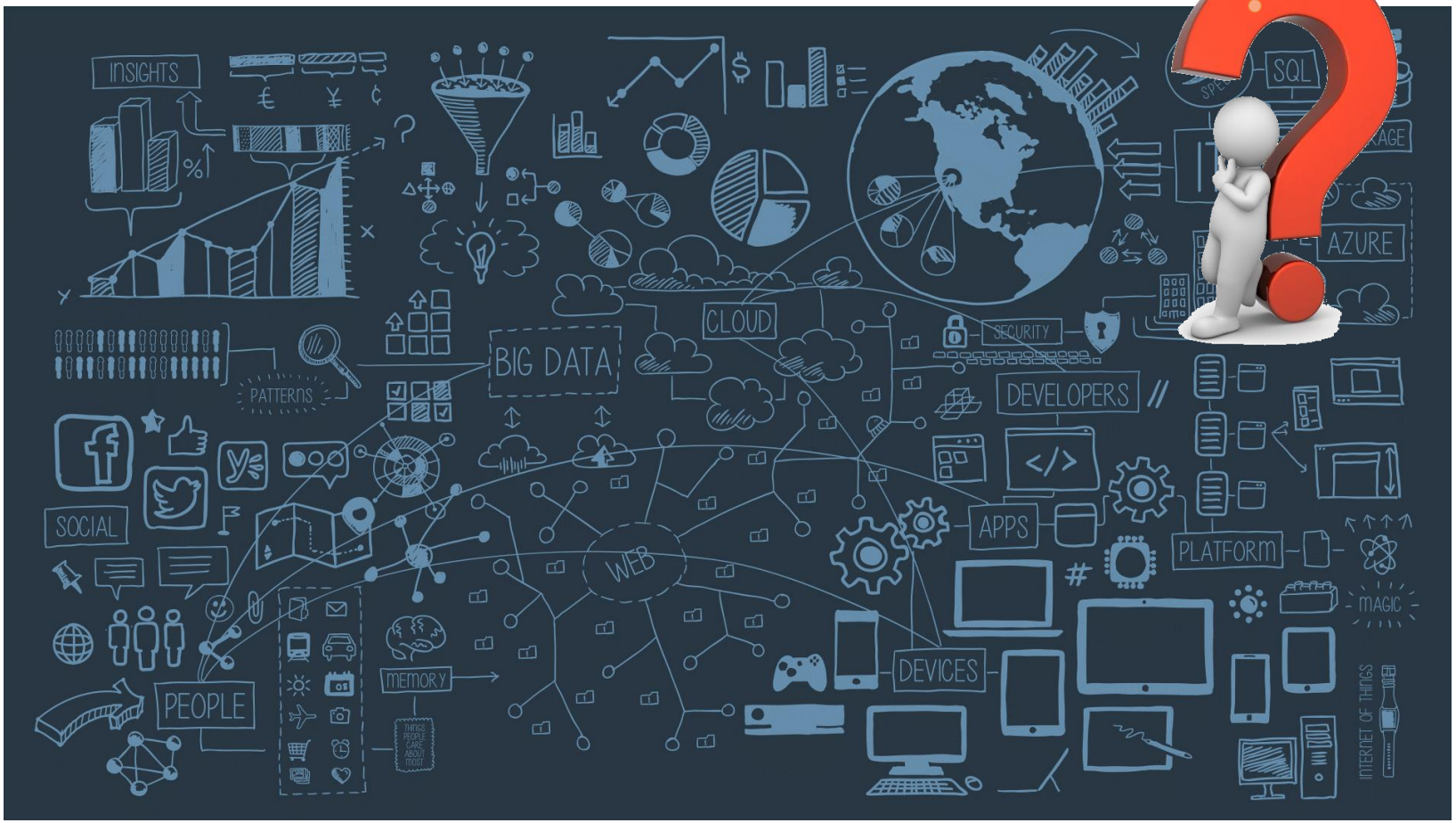


Defining Data Science

- Unfortunately, there is no clear definition (yet?)
- Goal is the extraction of knowledge from data
- Combination of techniques from different disciplines
- Scientific principles guide the data analysis

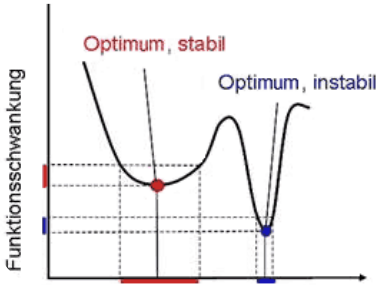
What is „Data Science“?!?

Tools? Big Data?
Machine Learning?



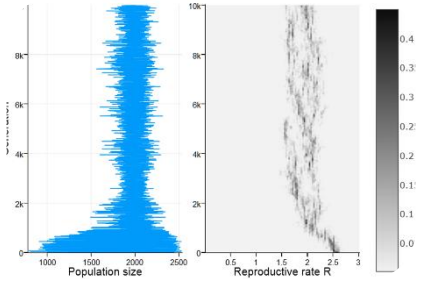
Mathematical Aspects

Computational Geometry



Optimization

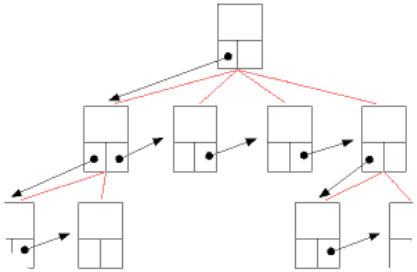
Stochastics



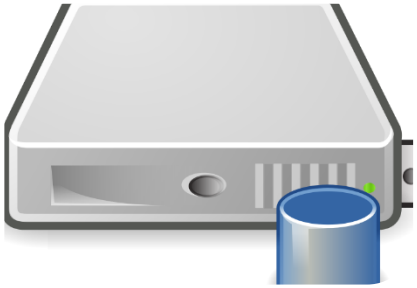
Scientific Computing

Machine Learning

Computer Science Aspects



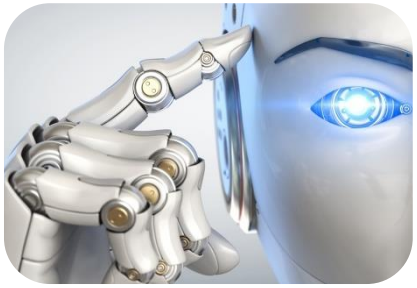
Data Structures and Algorithms



Databases

Distributed Computing

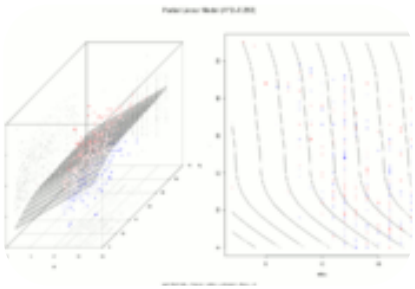
Software Engineering



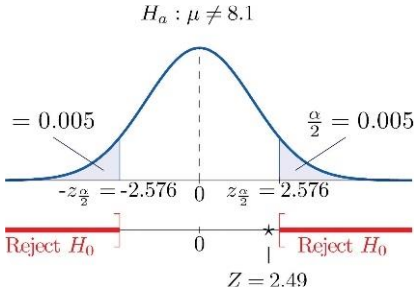
Artificial Intelligence

Machine Learning

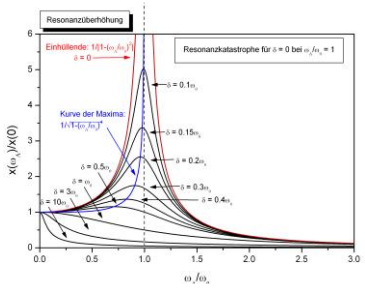
Statistical Aspects



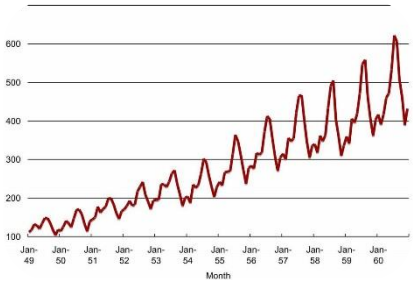
Linear Models



Statistical Tests



Inference



Time Series Analysis

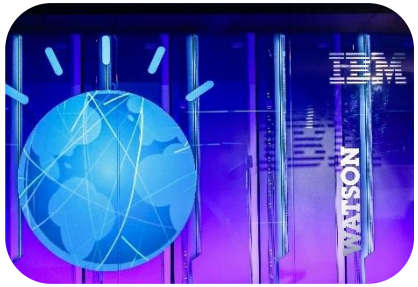
Machine Learning

Applications

Intelligent Systems

Robotics

Marketing



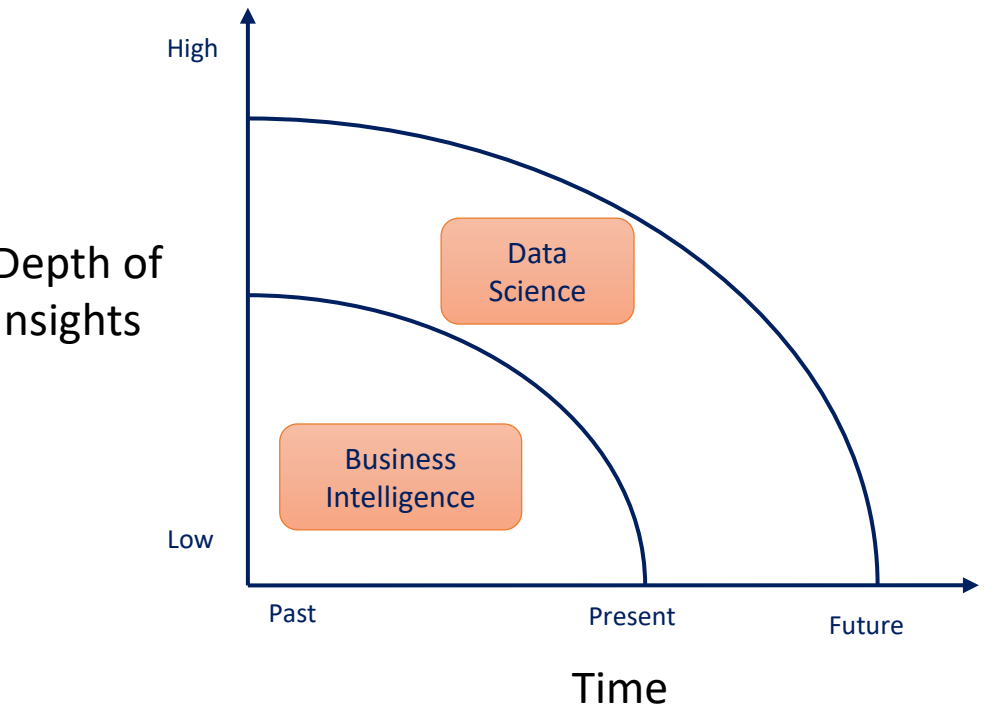
Medicine

Autonomous Driving

Social Networks

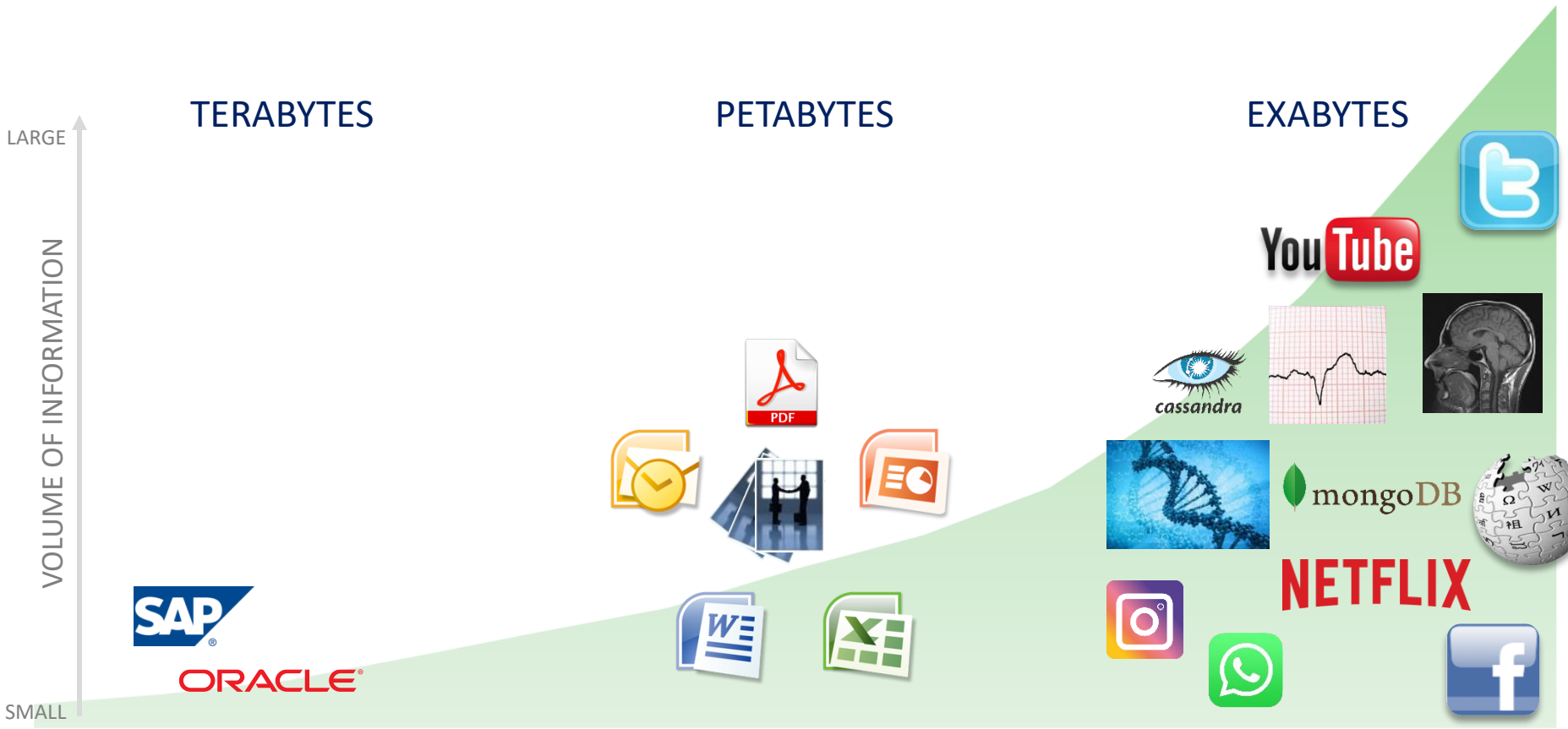
Data Science vs. Business Intelligence

- Business Intelligence (Gartner IT Glossary)
  - […] best practices that enable access to and analysis of information to improve and optimize decisions and performance.



	Business Intelligence	Data Science
Techniques	Dashboards, alerts, queries	Optimization, predictive modelling, forecasting
Data Types	Structured, data warehouses	Any kind, often unstructured
Common questions	What happened…? How much did…? When did…?	What if…? What will…? How can we…?

More Data → More Opportunities



1990's
Relational Databases
& Data Warehouses

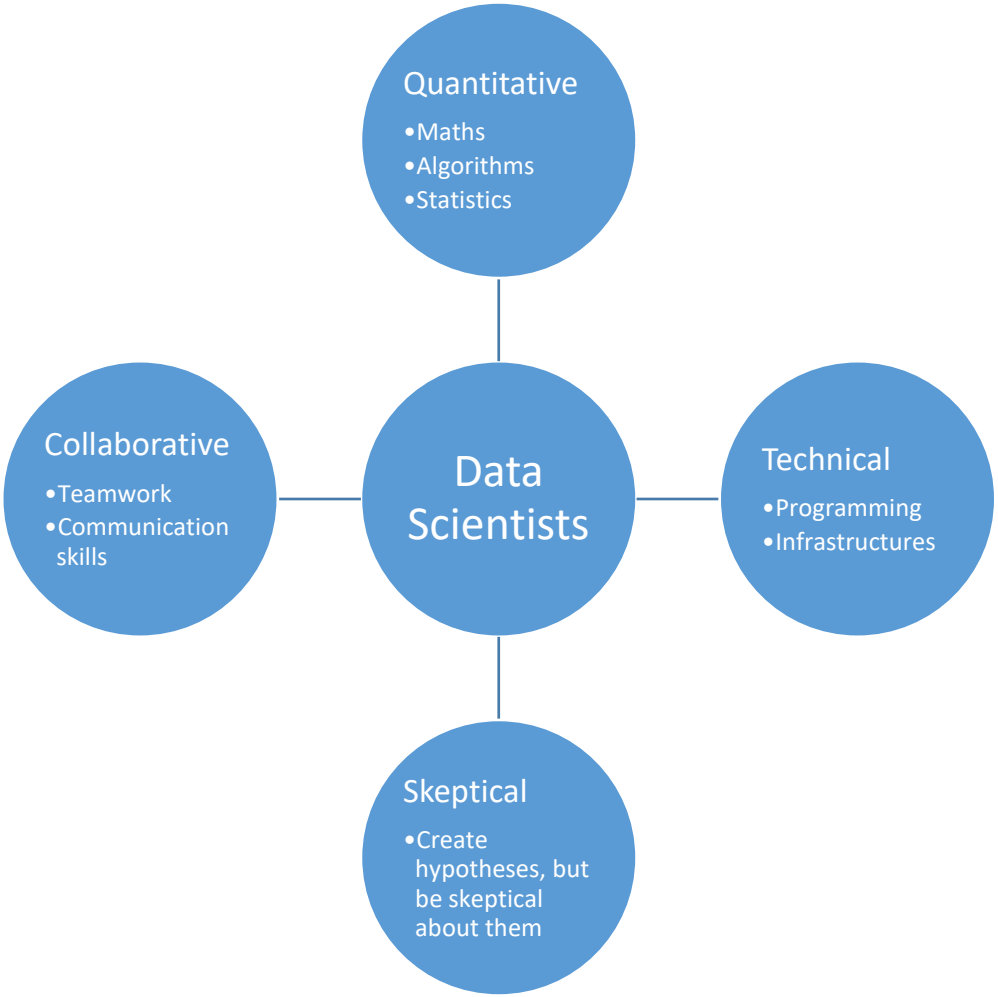
2000's
Content Management

2010's
Key-Value Storages
& Unstructured Data

What are Data Scientists?

- Not computer scientists
  - But should know about databases, data structures, algorithms, etc.
- Not mathematicians
  - But should know about optimization, stochastics, etc.
- Not statisticians
  - But should know about regression, statistical tests, etc.
- Not domain experts
  - But must work together with them

Skills of Data Scientists



A bit of everything

… but actually as much as possible of everything

Different types of Data Scientists

- According to Microsoft Research:
  - Polymath
    - „Do it all“
  - Data Evangelist
    - Data analysis, disseminating and acting on insights
  - Data Preparer
    - Querying existing data, preparing data for analysis
  - Data Shapers
    - Analyzing and preparing data
  - Data Analyzer
    - Analyzing data
  - Platform Builder
    - Collect data and create infrastructures
  - Moonlighters (50%/20%)
    - „Spare time“ data scientists
  - Insight Actors
    - Use the outcome and act on insights.

Miyung Kim, Thomas Zimmermann, Robert DeLine, Andrew Begel: Data Scientists in Software Teams: State of the Art and Challenges, IEEE Transactions on Software Engineering (Online First)

Summary

- Big data has a high volume, velocity, and variety
- Different data structures
  - Structured, semi-structured, quasi-structured, unstructured
- Data science is a very diverse discipline
  - Maths, computer science, statistics, applications

→ Data scientists require a diverse skillset

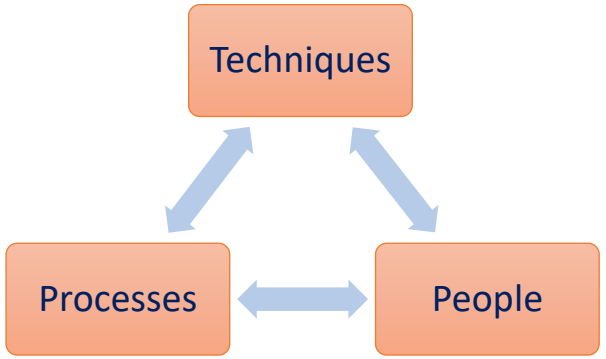
Process of Data Science Projects

Outline

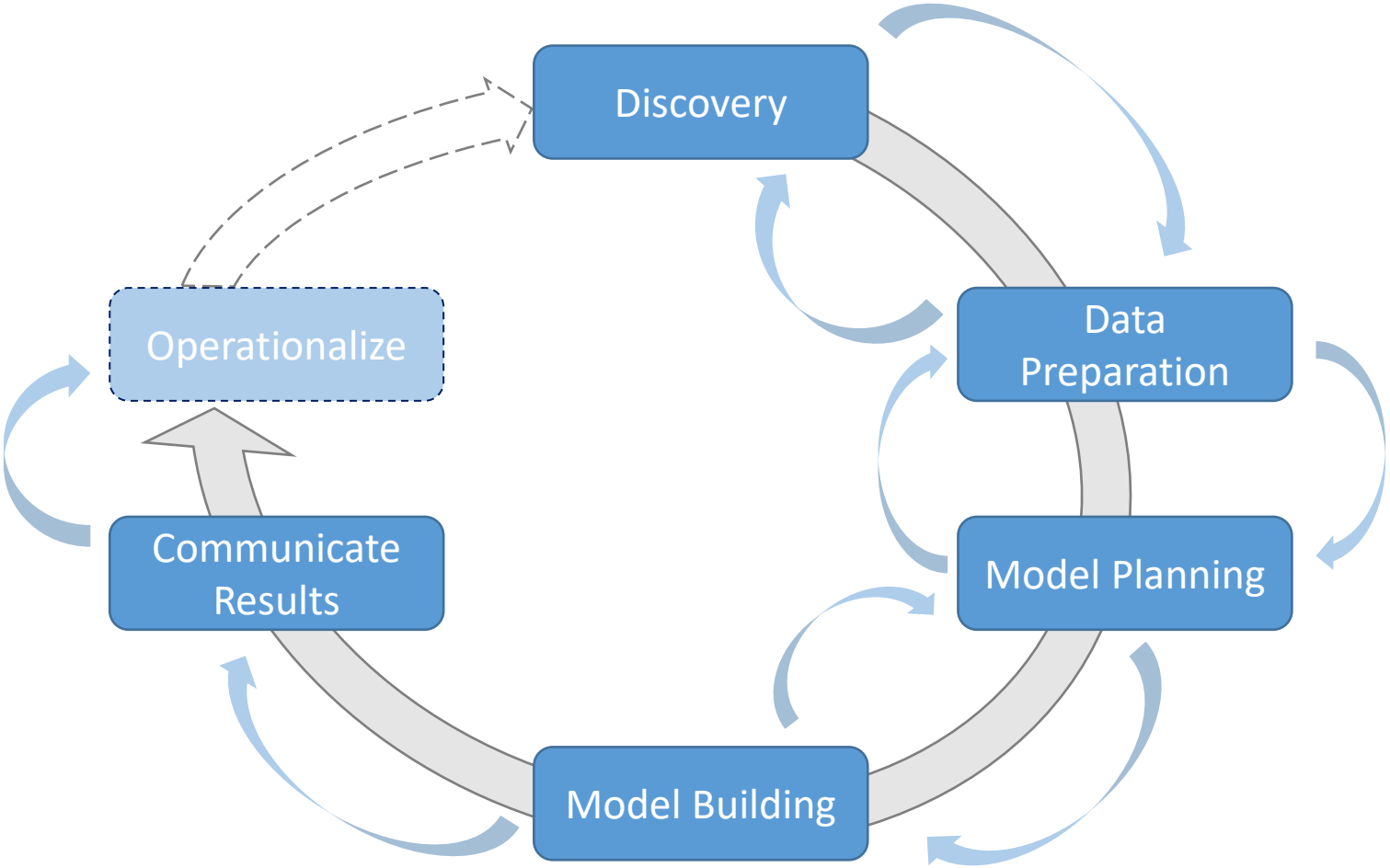
- Generic Process Model
- Roles
- Core Deliverables
- Summary

Processes are Important

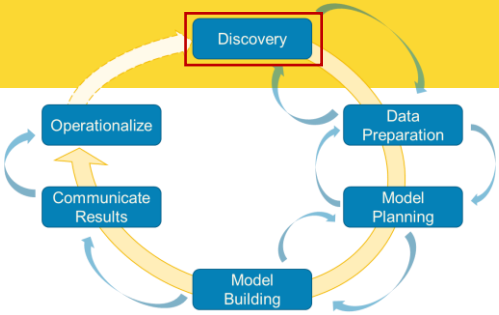
- Techniques
  - Languages, tools, and methods
  - Must be suited for the given problem
- People
  - Require training for the techniques
  - Should be guided through a project by a process
- Process
  - Supports the people
  - Must be accepted by the people
  - Should have a measurable positive effect



Process of Data Science Projects



Discovery

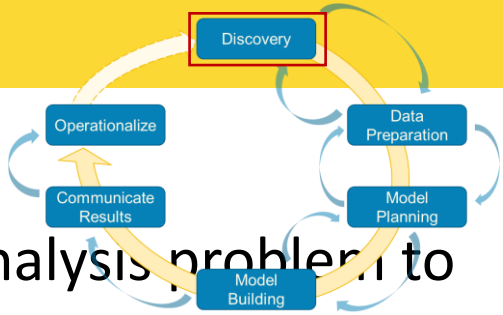


- Initial phase of the project

- Learn the domain
  - Knowledge for understanding the data and the use cases of the project
  - Knowledge for the interpretation of the results

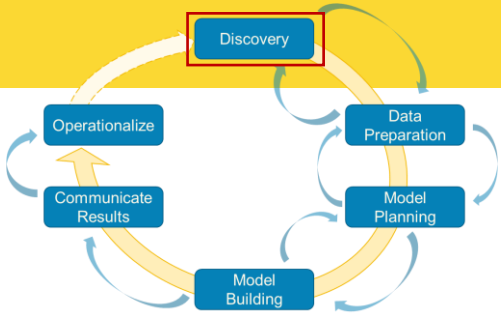
- Learn from the past
  - Identify past projects on similar issues
    - Differences, reasons for failures, weaknesses of past projects
  - Can also be projects of competitors, if reports are available

Discovery



- Frame the problem
  - Framing is the process of stating the data analysis problem to be solved
  - Why is the problem important?
  - Who are the key stakeholders and what are their interests in the project?
  - What is the current situation and what are pain points that motivate the project?
  - What are the objectives of the project?
    - Business needs
    - Research goals
  - What needs to be done to achieve the objectives?
  - What are success criteria for the project?
  - What are risks for the project?

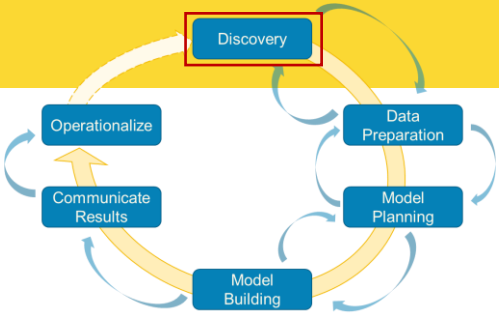
Discovery



- Begin learning the data
  - Get a high-level understanding of the data
    - Maybe even some initial statistics or visualizations of the data
  - Determine requirements for data structures and tools for processing the data

- Formulate hypothesis
  - Part of the „Science“ in „Data Science“
  - Should define expectations
    - „Feature X is well suited for the prediction of …“
    - „The following patterns will be found in the data: …“
    - „Deep learning will outperform …“
    - „Decision trees will perform well and allow insights into …“
  - Should be discussed with stakeholders

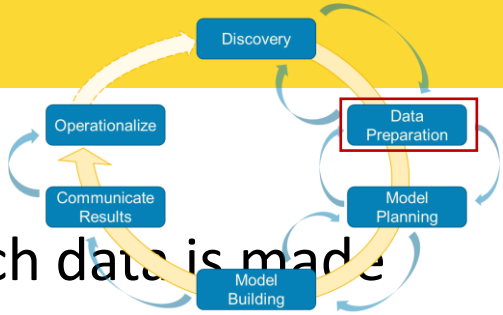
Discovery



- Analyze available resources
  - Technologies
    - Resources for computation and storage
    - Licenses for analysis frameworks
  - Data
    - Is the available data sufficient for the use case?
    - Would other data be required and could the additional data be collected within the scope of the project?
  - Timeframe
    - Scope in calendar time and person months
  - Human resources
    - Who is available for the project?
    - Is the skillset a good match for the tasks of the project?

→ Only start project if the resources are sufficient!

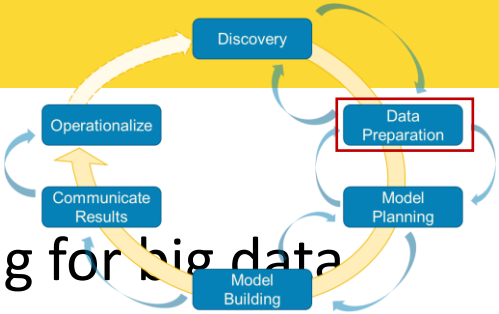
Data Preparation



- Create the infrastructure for the project
  - Usually different from infrastructure in which data is made available to you
  - Warehouse/csv-file/… ←→ distributed storage that enables analysis
    - Could also be simpler, for small data sizes

- Extract – Transform – Load (ETL) the data
  - Define how to query existing database to extract required data
  - Determine required transformations of the raw data
    - Quality checking (e.g., filtering of missing data, implausible data)
    - Structuring (e.g., for unstructured data, differences in data structures)
    - Conversions (e.g., timestamps, character encodings)
  - Load the data into your analysis environment

Data Preparation



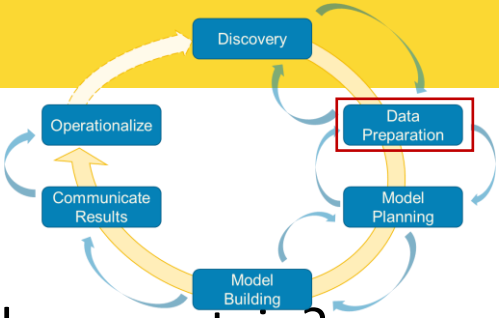
- ELT vs. ETL
  - Transformations can be very time-consuming for big data
  - Might not be possible without using the analysis infrastructure

→ Load raw data, transform afterwards → ELT!

- Also allows more flexibility with transformations
  - E.g., testing the effect of different transformations

- Allows access to raw data

Data Preparation

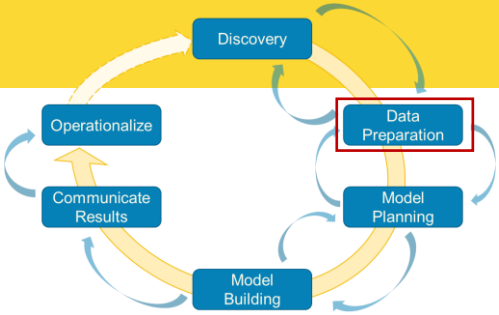


- Get a deep understanding of the data
  - Understand all data sources
  - E.g., what does each column in a relational database contain?
  - How can a structure be imposed on semi-/quasi-/unstructured data?

- Survey and visualize data
  - Descriptive statistics
  - Correlation analysis
  - Visualizations like histograms, density plots, pair-wise plots, etc.

- Clean and normalize data
  - Discard data that is not required
  - Normalize to remove scale effects

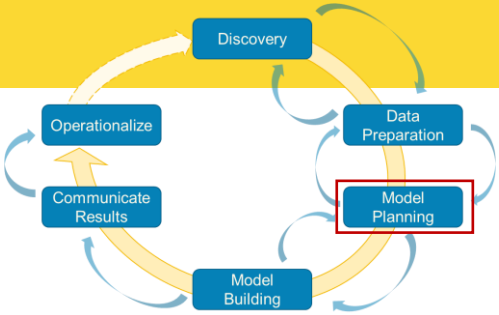
Data Preparation



- Clean data
  - Discard data that is not required
  - Can make the difference between a complex infrastructure and a single machine for analysis
  - Example:
    - 100 million measurements
    - 10 floating point features per measurement → 80 Bytes per measurement
    - 3 useful features ≈ 24 Bytes per measurement
    - 7.45 Gigabytes with all features, 2.23 Gigabytes with only useful features

  → Can use my laptop for cleaned data without problems

Model Planning



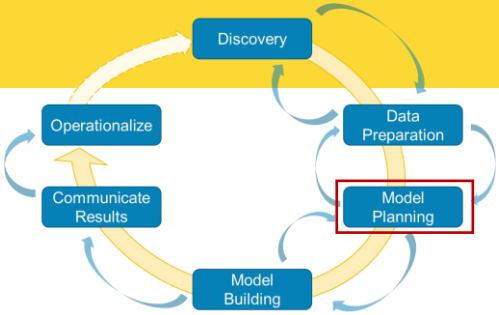
- Determine methods for data analysis

- Should be well-suited to meet objectives
  - Often determines the type of method
    - Classification, regression, clustering, association mining, …
  - Other factors can also restrict the available methods
    - For example, if insight is important, „blackbox“ methods cannot be used

- Should be well-suited for the available data
  - Volume, structure, …

A blackbox method is a method where you only get results, but do not really understand why the output is computed that way.
A whitebox method also explains why the output is as it is.

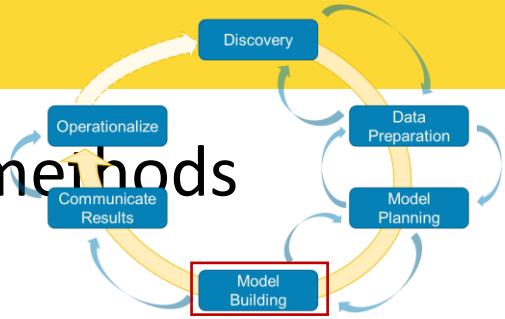
Model Planning



- Methods for data analysis may cover
  - Feature modeling, e.g., for text mining
  - Feature selection, e.g., based on information gain, correlations, etc.
  - Model creation, e.g., different models that may address the use case
  - Statistical methods, e.g., for the comparison of results
  - Visualizations, e.g., for the presentation of results

- Split data into different data sets
  - Training data, validation data, test data
  - „Toy“ data for local use in case of big data
    - Same structure, but very small

Model Building

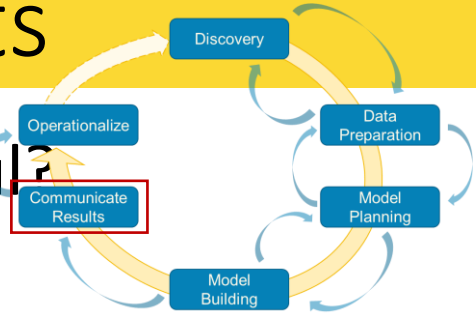


- Perform the analysis using the planned methods
  - Often iterative process!

- Separate phase, because this can be VERY time consuming
  - Use toy examples for model planning
  - Use real big data set with potentially lots of hyper parameters for tuning during model building

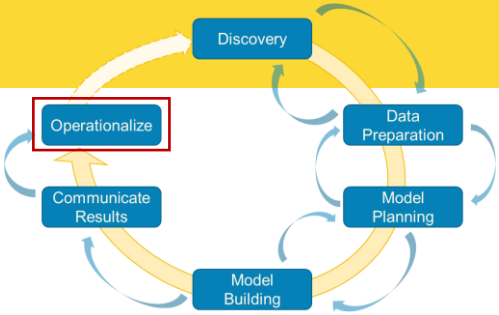
- Includes the calculation of performance indicators

Communicate Results



- Main question: Was the project successful?
- Compare results to hypothesis from the discovery phase
- Identify the key findings
- Try to quantify the value of your results
  - Business value, e.g., the expected Return On Investment (ROI)
  - Advancement of the state of the art
- Summarize findings for different audiences

Operationalize



- Implement results in operation
  - Only in case of successful projects
- Should run a pilot first
  - Determine if expectations hold during the practical application
  - All kinds of reasons for failures
    - Rejection by users, shift in data reduces model performance, ...
- Define a process to update and retrain model
  - Data gets older, models get outdated
  - Data driven models should be updated regularly
  - Process is required

Outline

- Generic Process Model
- **Roles**
- Core Deliverables
- Summary

Roles within Projects

- A role is „a function or part performed especially in a particular operation or process” (Merriam-Webster)

- Role ≠ Person
  - One role can be fulfilled by multiple persons
  - One person can fulfill multiple roles

- Roles assign responsibilities within processes
  - In practice, roles are often related to job titles
    - „Software Developer“, „Database Administrator“, „Project Manager“, …

Roles for Data Science Projects

Role	Description
Business User	• Someone who uses the end results • Can consult and advise project team on value of end results and how these will be operationalized
Project Sponsor	• Responsible for the genesis of the project • Generally provides the funding • Gauge the value from the final outputs
Project Manager	• Ensure key milestones and objectives are met on time and at expected quality • Plans and manages resources
Business Intelligence Analyst	• Business domain expertise with deep understanding of the data • Understands reporting in the domain, e.g., Key Performance Indicators (KPIs)
Data Engineer	• Deep technical skills to assist with data management and ETL/ELT
Database Administrator	• Provisions and configures database environment to support the analytical needs of the project
Data Scientist	• Expert on analytical techniques and data modeling • Applies valid analytical techniques to given business problems • Ensures analytical objectives are met

Outline

- Generic Process Model
- Roles
- **Core Deliverables**
- Summary

Deliverables

- A deliverable is a tangible or intangible good or service produced as a result of a project.
  - Are often parts of contracts
  - Should meet stakeholder‘s needs and expectations

- Four core deliverables for data science projects
  - Sponsor presentation
  - Analyst presentation
  - Code
  - Technical specifications

Sponsor Presentation

- „Big Picture“ of the project
- Clear takeaway messages
  - Highlight KPIs
  - Should aid decision making
- Should address a non-technical audience
- Clean and simple visualizations
  - For example, bar charts, line charts, …

Analyst Presentation

- Describe analysis methods and data
  - General approach
  - Interesting insights, unexpected situations
- Details on how results change current status
  - Business process changes
  - Advancement of the state of the art
- May use more complex visualizations
  - For example, density plots, histograms, boxplots, ROC curves, …
  - Should still be clean and not overloaded

Code and Technical Specification

- All available code of the project
  - Often code is prototypical („hacky“) because results are more important than clean code
- Enables operationalization
  - May re-use code as is
  - May adopt code or clean up code
  - May rewrite same functionality in a different language/for a different environment
- Technical specification should be provided as well
  - Description of the environment
  - Description of how to invoke code

Expected Deliverables by Role

Role	Deliverable
Business User	Expects a sponsor presentation: ➢ Are the results good for me? ➢ What are the benefits for me? ➢ What are the implications for me?
Project Sponsor	Expects a sponsor presentation: ➢ What is the impact of operationalizing the results? ➢ What are the risk and what is the potential ROI? ➢ How can this be evangelized within the organization (and beyond)?
Project Manager	• Responsible for the timely availability of all deliverables • Responsible for the sponsor presentations
Business Intelligence Analyst	Expects an analyst presentation: ➢ Which data was used? ➢ How will reporting change? ➢ How will KPIs change?
Data Engineer	• Responsible for data engineering code and technical documentation
Database Administrator	• Responsible for infrastructure code and technical documentation
Data Scientist	• May be the target audience for analyst presentations. • Responsible for data analysis code and technical documentation • Responsible for the analyst presentation • Support of the project management with the sponsor presentation

Data as Deliverable

- Only applicable if new data was collected/generated
- Sharing the data may be very important
  - Especially in research to enable reproducible and replicable research
- Sharing may be internal (industry) or public (research)
  - Use stable links for references to prevent link rot
  - Ideally Digital Object Identifiers (DOIs)
- Should not only contain the data, but also metadata and tools for collecting the data

Outline

- Generic Process Model
- Roles
- Core Deliverables
- **Summary**

Summary

- Generic process for data science projects with six phases
  - Discovery, data preparation, model planning, model building, communication of results, and operationalization
- Different actors in different roles involved in project
  - Expectations depend on role
- Four core deliverables fulfill most stakeholder needs
  - Sponsor presentation, analyst presentation, code, technical specification
- Data may also be a deliverable