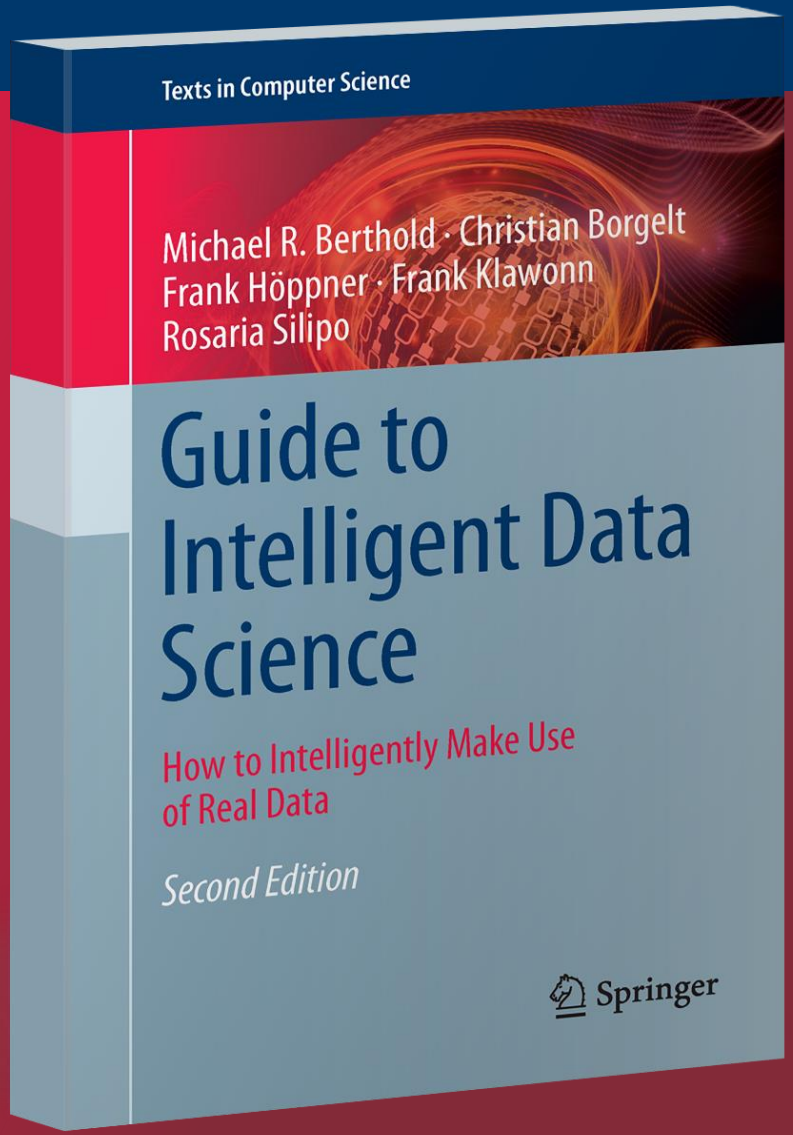
Introduction to Data Science



Summary of this lesson

„We are drowning in information, but starving for knowledge”

-John Naisbett

What is **knowledge**?

**This lesson refers to chapters 1 and 2 of the GIDS book*

Content of this lesson

- What is Data Science?
- The Data Science Process
- Data Science: An Example

What is Data Science?

Data vs. Knowledge

Data

- refer to single instances (single objects, people, events, points in time, etc.)
- describe individual properties
- are often available in large amounts (databases, archives)
- are often easy to collect or to obtain (e.g., scanner cashiers in supermarkets, Internet)
- do not allow us to make predictions or forecasts

Knowledge

- refers to *classes* of instances (*sets* of objects, people, events, points in time, etc.)
- describes general patterns, structures, laws, principles, etc.
- consists of as few statements as possible
- is often difficult and time consuming to find or to obtain (e.g., natural laws, education)
- allows us to make predictions and forecasts

Criteria to assess knowledge

- **correctness** (probability, success in tests)
- **generality** (domain and conditions of validity)
- **usefulness** (relevance, predictive power)
- **comprehensibility** (simplicity, clarity, parsimony)
- **novelty** (previously unknown, unexpected)

[Wikipedia quoting Dhar 13, Leek 13]

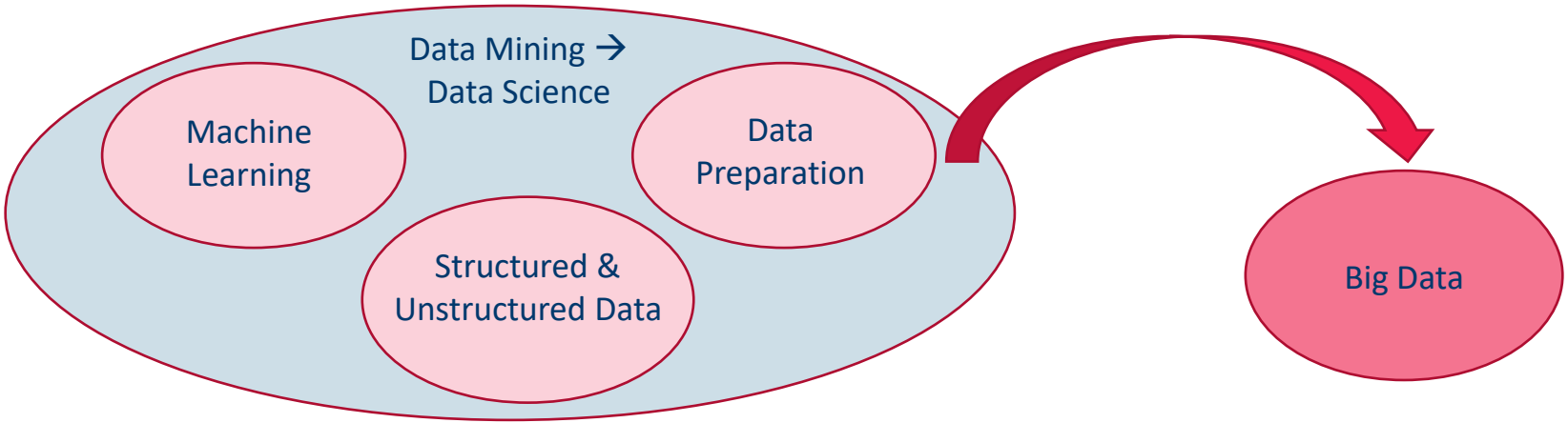
Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to **extract knowledge and insights** from structured and unstructured data.

[Fayyad, Piatetsky-Shapiro & Smyth 96]

Knowledge discovery in databases (*KDD*) is the process of (semi-)automatic **extraction of knowledge** from databases which is *valid*, *previously unknown*, and *potentially useful.*

Some Clarity about Words

- *(semi)-automatic*: no manual analysis, though some user interaction required
- *valid*: in the statistical sense
- *previously unknown*: not explicit, no „common sense knowledge“
- *potentially useful*: for a given application
- *structured data*: numbers
- *unstructured data*: everything else (images, texts, networks, chem. compounds, …)



Valid?

99.98%

Valid?

customer age $\in [18, 150]$

(in 9, 999 of 10, 000 cases)

Previously Unknown?

A => B (in 100% of all cases)

Previously Unknown?

Pregnant => Female

Useful?

A => B

(with s =0.81% and c = 21.3%)

Useful?

Beer => Diapers

(with s =0.81% and c = 21.3%)

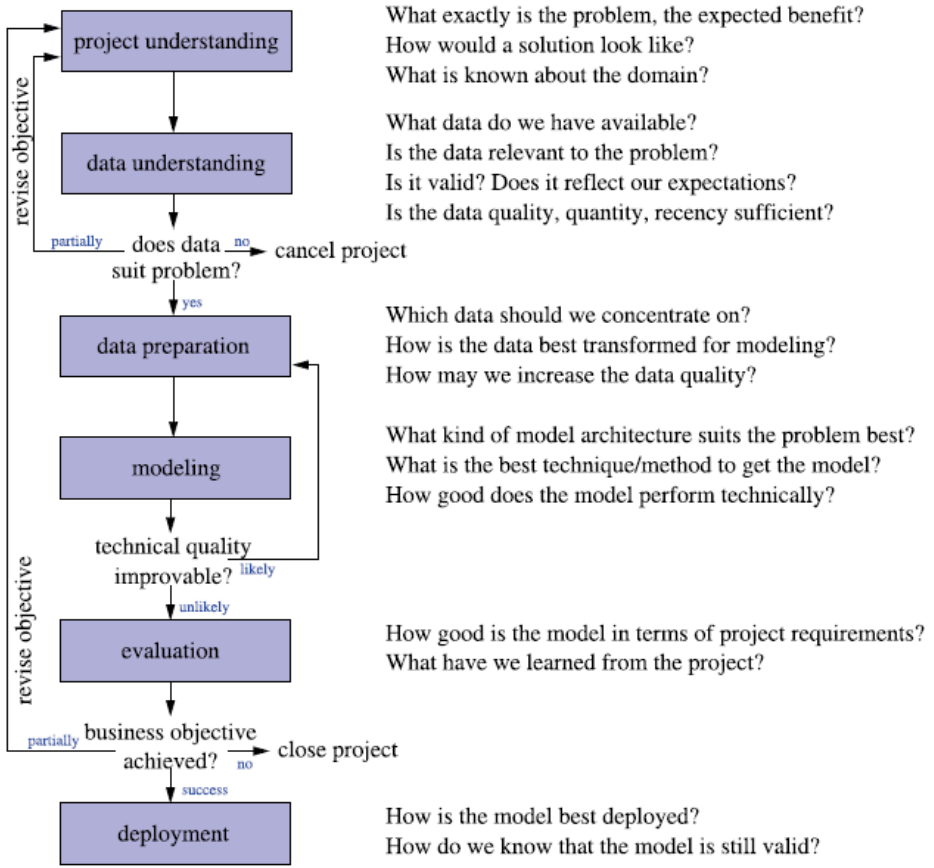
Valid, Interesting, and Useful?

Books A and B => Book C

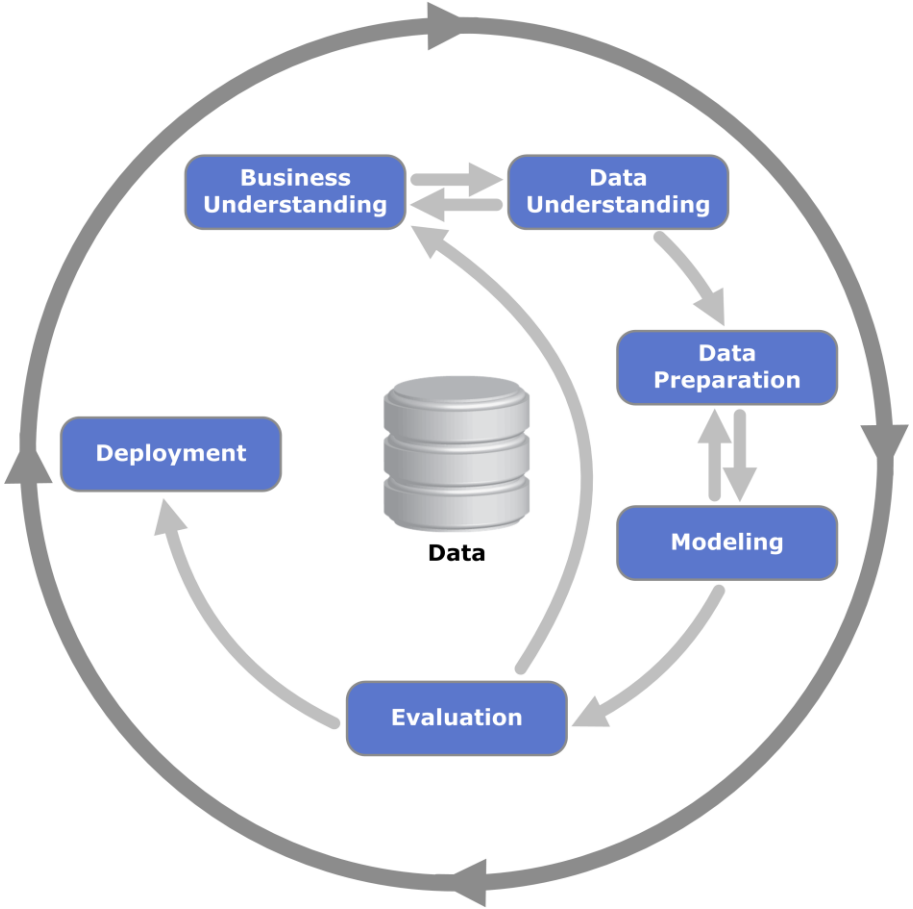
(with s =0.81% and c = 21.3%)

The Data Science Process

- **SEMMA**
  - Sample, Explore, Modify, Model, Assess
- **CRISP-DM**
  - Cross Industry Standard Process for Data Mining
- **KDD**
  - Knowledge Discovery in Databases

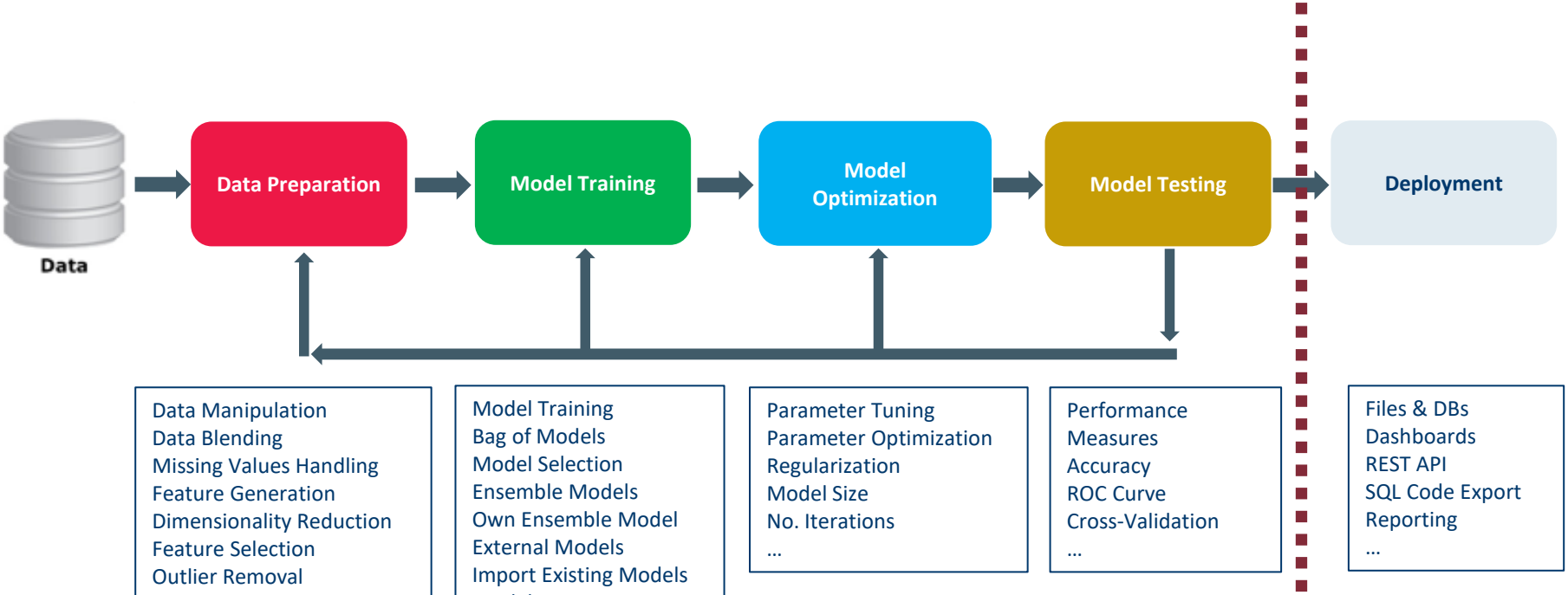


CRISP-DM

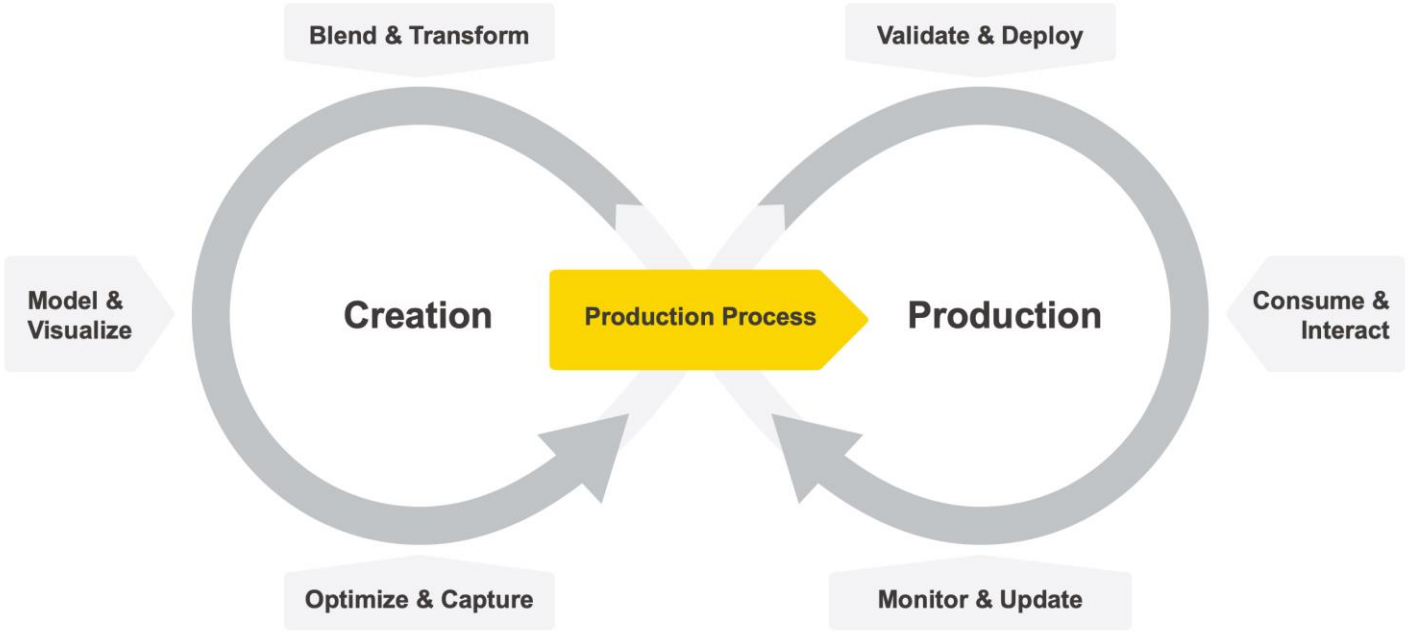

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

A Classic Data Science Project

It always starts with some data …



The Data Science Life Cycle



- **Classification**
  - Predict experiment outcome falling into a finite number of possible results
  - *How credit-worthy is this customer ? Very / Enough / Not enough / Absolutely not*
  - *Will this customer respond to our mailing? Yes / No*

- **Regression**
  - Predict numeric values
  - *How will the EUR/USD exchange rate develop?*
  - *What will be the price of this washing machine next week?*

- **Clustering, Segmentation**
  - Group similar cases in order to get overview, detect outliers, or get insights on the data structure
  - *Do my customers separate into different groups?*
  - *How many operating points does the machine have, and what do they look like?*

- **Association Analysis**
  - Find correlations to better understand the interdependencies of all the attributes
  - Focus in the full record (all the attributes) rather than on a single target variable
  - *Which optional equipment of a car often goes together?*
  - *How do the various qualities in a car influence each other?*

- **Deviation Analysis**
  - Knowing the trend of the data, find subgroups that behave differently
  - *Under which circumstances does the system behave differently?*
  - *Which properties do those customers - who do not follow the crowd - share?*

Data Science: an Example

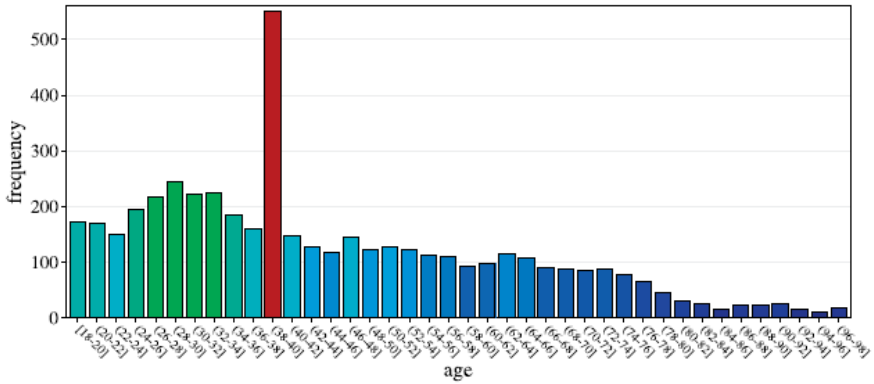
Example

- Dataset from a hypotetical supermarket chain
  - Customers
  - Products
  - Purchases
- Three tasks
  - Divide customers into different groups according to their purchase behaviour
  - Identify connections between products to implement cross-selling campaigns
  - Helping design a marketing campaign to increase purchases
- Two approaches
  - Naive approach lead by common sense
  - Sound approach using DS techniques

- Naive Approach
  - Aggregate purchases to respective customer
  - Join with the customer details
  - No interesting relations highlighted

Cluster id	Age	Customer revenue
1	46.5	€ 1,922.07
2	39.4	€ 11,162.20
3	39.1	€ 7,279.59
4	46.3	€ 419.23
5	39.0	€ 4,459.30

- Sound Approach
  - Check values for the string attributes (name, employment..)
  - Check and add constraints to numeric attributes (e.g. Age between 18-100)
  - Look for misleading information (e.g. In the dataset a missing birthdate was by default set to 1970. If not handled properly, this information can lead to errors)
  - Use average basket price as estimator for the value of a customer
  - Use average number of purchases per month as further estimator
  - Apply normalization to average attributes magnitudes



Cluster	Age	Avg. cart price	Avg. purchase/month
1	75.3	€ 19.-	5.6
2	42.1	€ 78.-	7.8
3	38.1	€ 112.-	9.3
4	30.6	€ 16.-	4.8
5	44.7	€ 45.-	3.7

Explanation Finding: Find Product Dependencies

- Naive Approach
  - Run Association Rule Mining algorithm with default setting
  - Consider Product ID (differenciating each product)
  - Unintuitive and unuseful result
  - Rules have high confidence but low support values

- Sound Approach
  - Consider product categories
  - Rules match with well-known facts
  - Monitor combinations on regular basis

‘foie gras’ (p1231) ← ‘champagne Don Huberto’ (p2149),
‘truffle oil de Rossini’ (p578) [s=1E-5, c=75%]

‘Tortellini De Cecco 500g’ (p3456)
← ‘De Cecco Sugo Siciliana’ (p8764) [s=1E-5, c=60%]

tomatoes ← capers, pasta [s=0.007, c=32%]

tomatoes ← apples [s=0.013, c=22%]

Predicting the Future: Forecast customers reactions to coupon mailings

- Naive Approach
  - No detailed analysis
  - Send coupon with discounts after a certain purchase amount
  - Just monitor the results
  - Fail: customers only combine shopping trips, no additional revenues
  - The data analyst is in the end fired
- Sound Approach
  - Discriminate valuable customers => exploit earlier segmentation
  - Derive meaningful attributes, e.g. Customers underperforming on specific category, distance
  - Build black box classifier model

Thank you

Guide to Intelligent Data Science Second Edition, 2020