Data Science Process

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*

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)

What is Data Science?

[Wikipedia quoting Dhar 13, Leek 13]

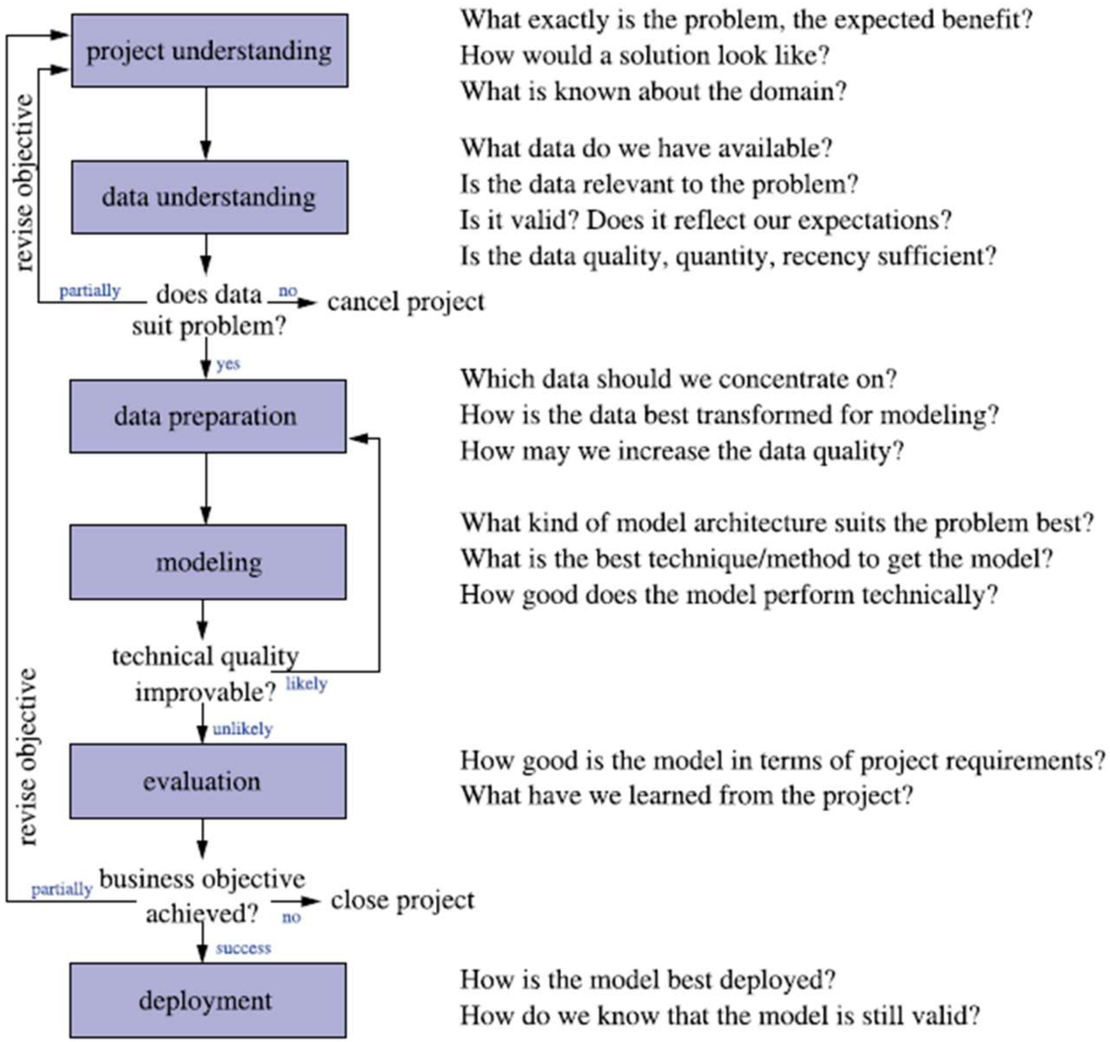
Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to **extract knowledge and insights** from structured and unstructured data.

[Fayyad, Piatetsky-Shapiro & Smyth 96]

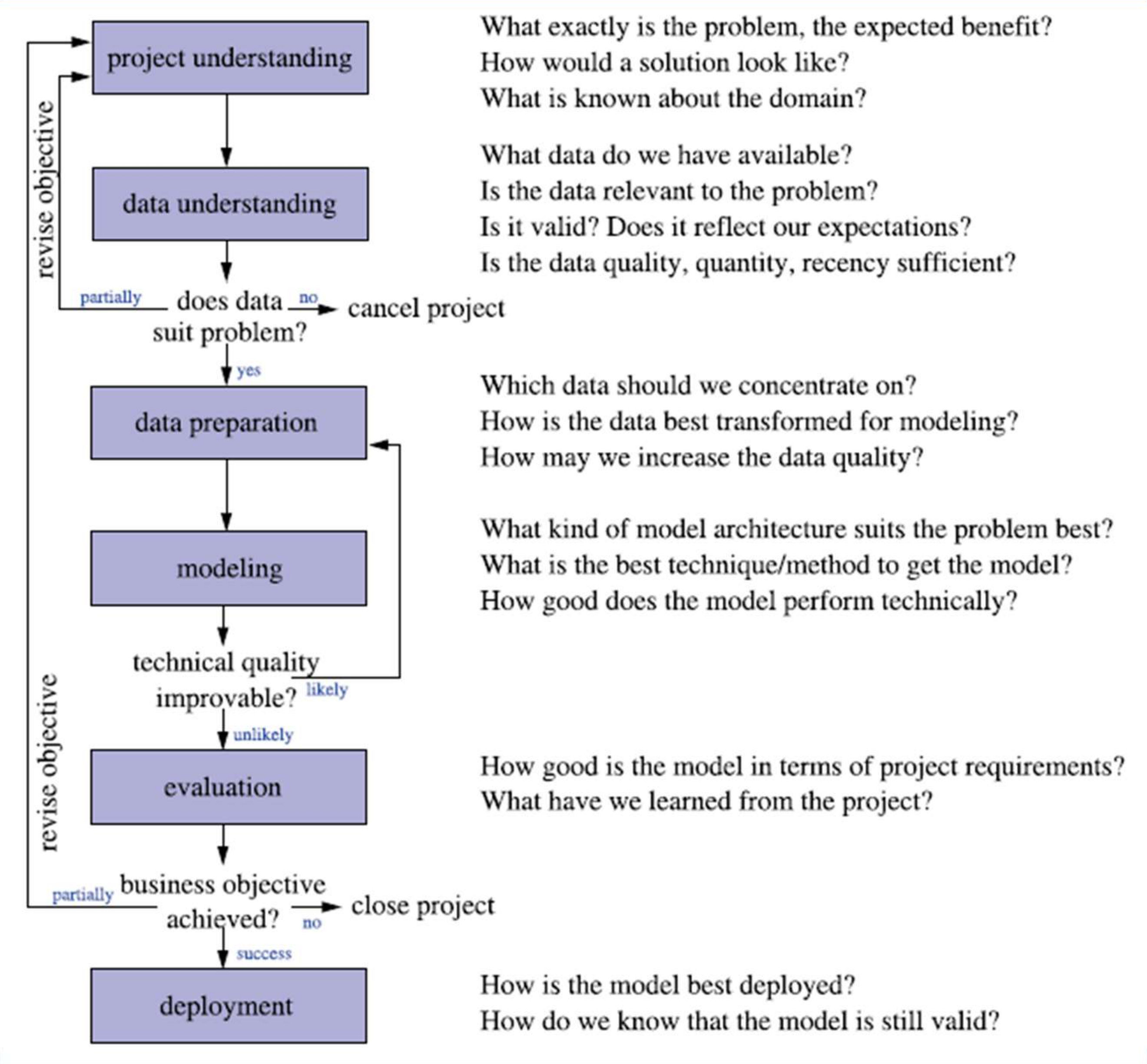
Knowledge discovery in databases (*KDD*) is the process of (semi-)automatic **extraction of knowledge** from databases which is *valid*, *previously unknown*, and *potentially useful.*

The Data Science Process

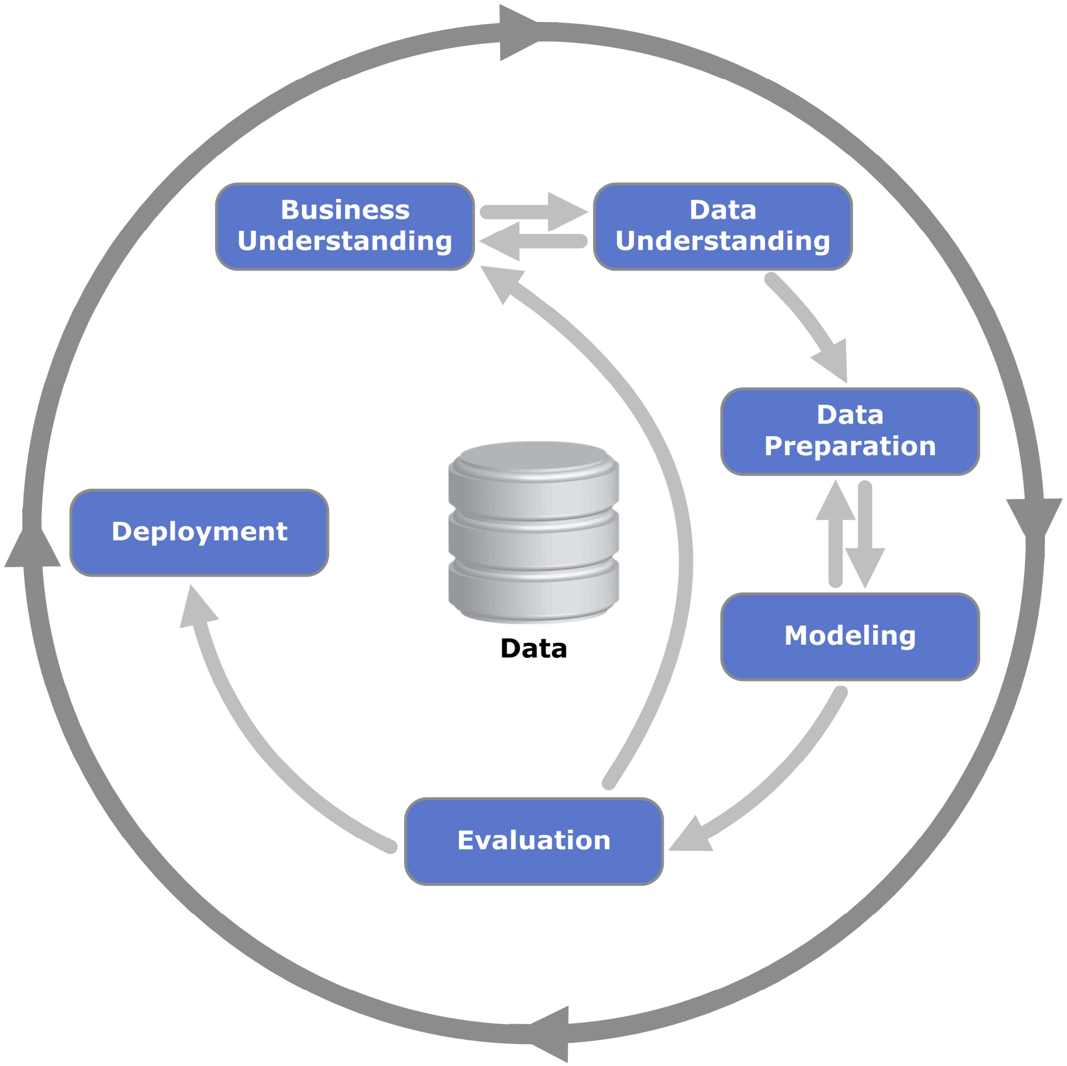
- **SEMMA**
  - Sample, Explore, Modify, Model, Assess
- **CRISP-DM**
  - Cross Industry Standard Process for Data Mining
- **KDD**
  - Knowledge Discovery in Databases



The Data Science Process



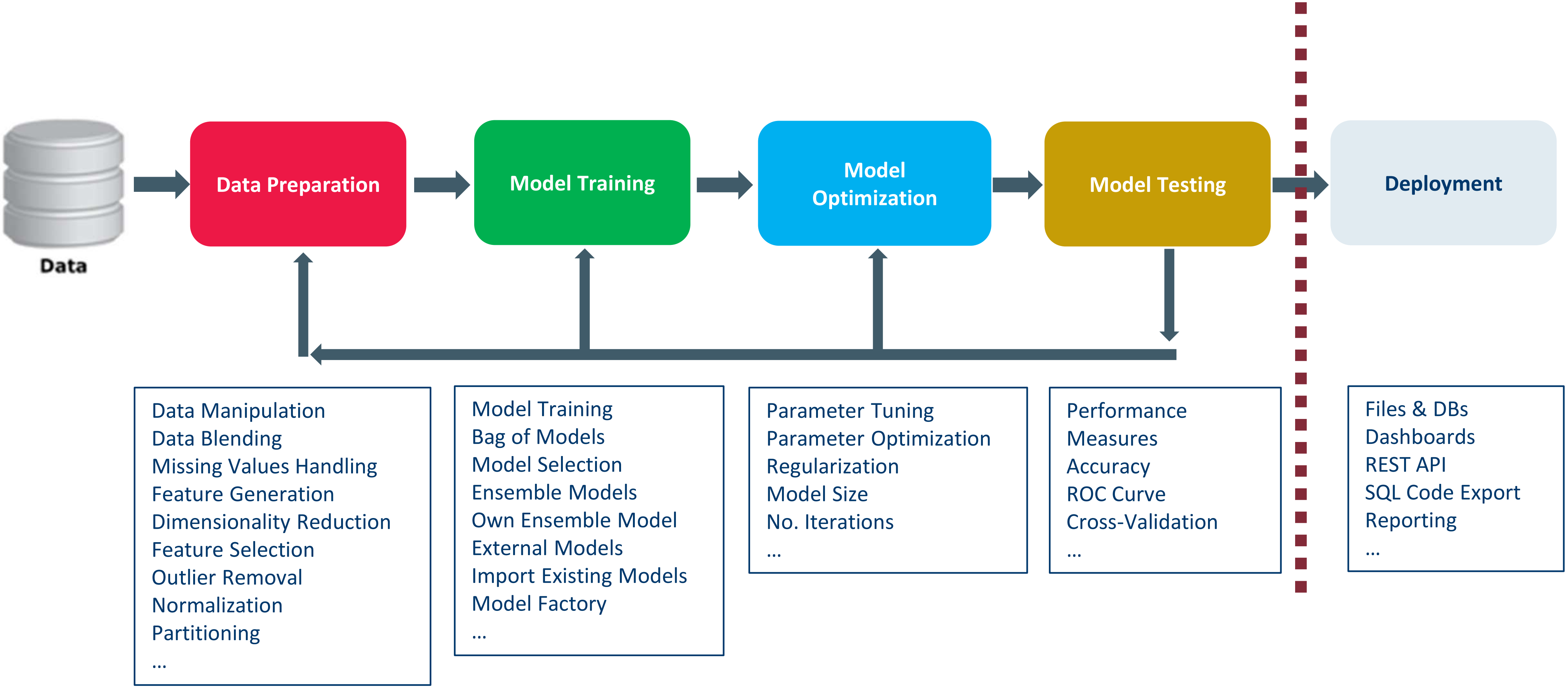
CRISP-DM



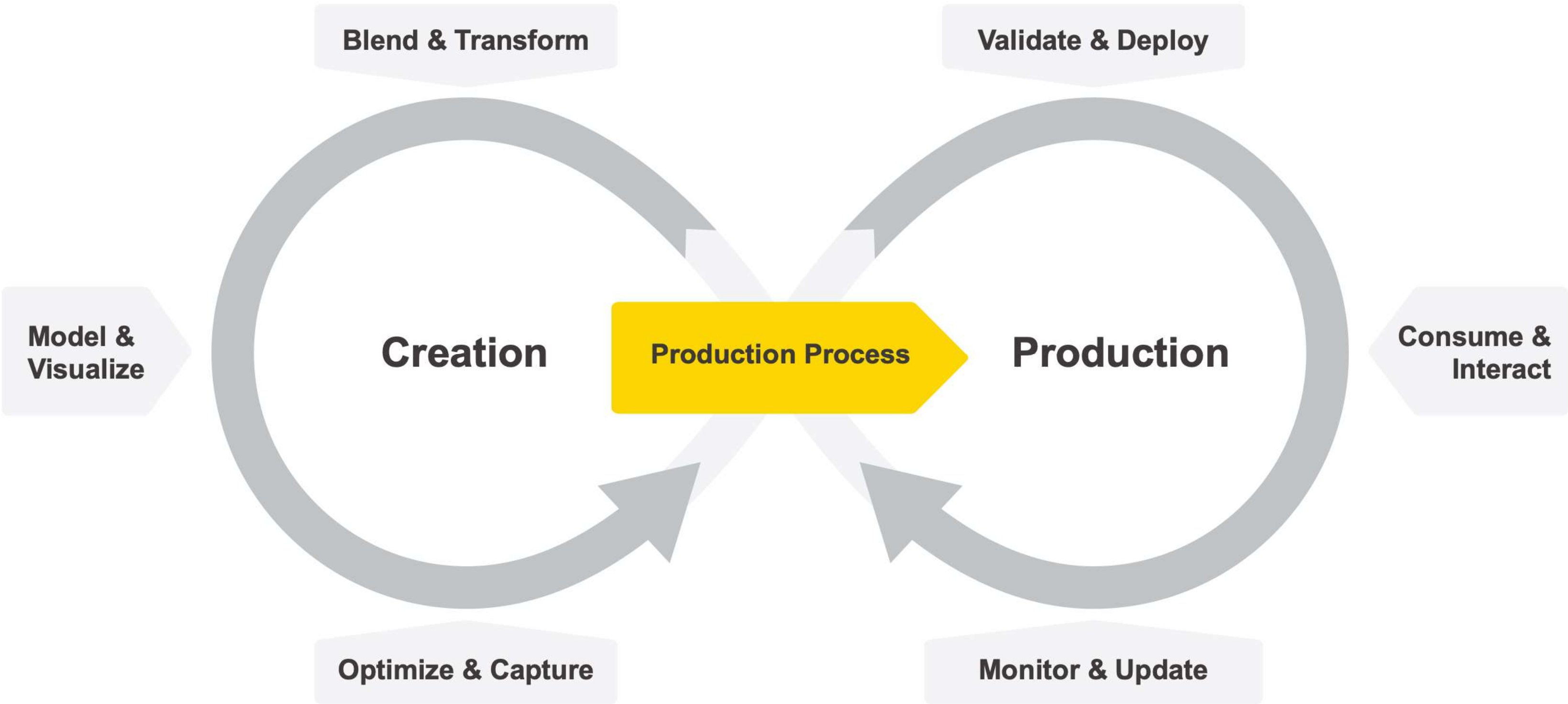
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

A Classic Data Science Project

It always starts with some data …



The Data Science Life Cycle



Problem Categories

- **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?*

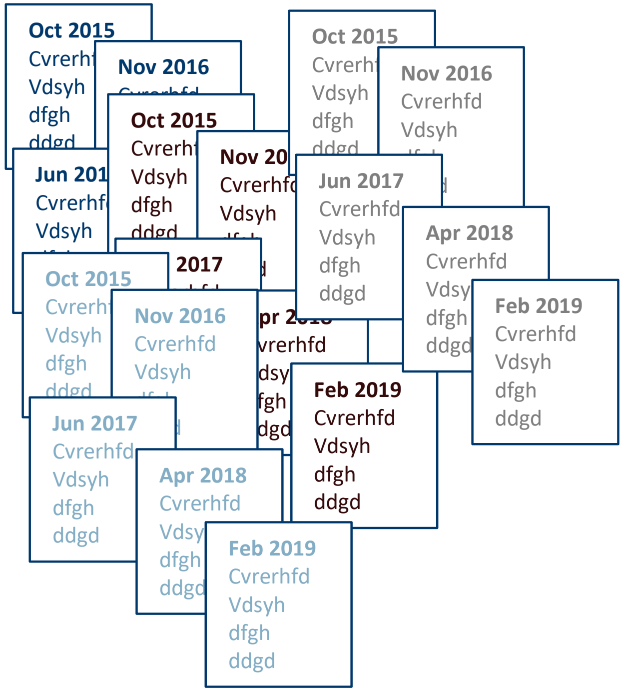
- **Deviation Analysis**
  - Knowing the trend of the data, find subgroups that behave differently
  - *Under which circumstances does the system behave differently?*
  - *Which properties do those customers - who do not follow the crowd - share?*

Some Classic Use Cases

Risk Assessment

– Risk Assessment: is this person going to repay the loan?

Customer History



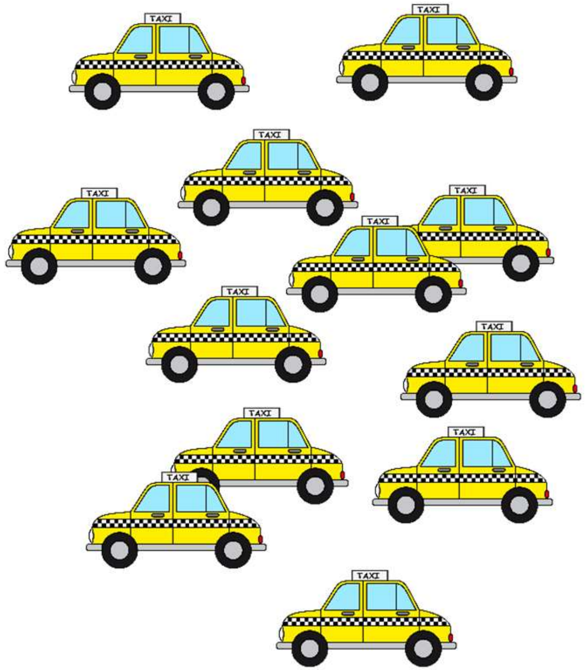
Model

Risk Prognosis

- High Risk
- Low Risk
- High Risk
- Very High Risk
- Very Low Risk
- Medium Risk
- …

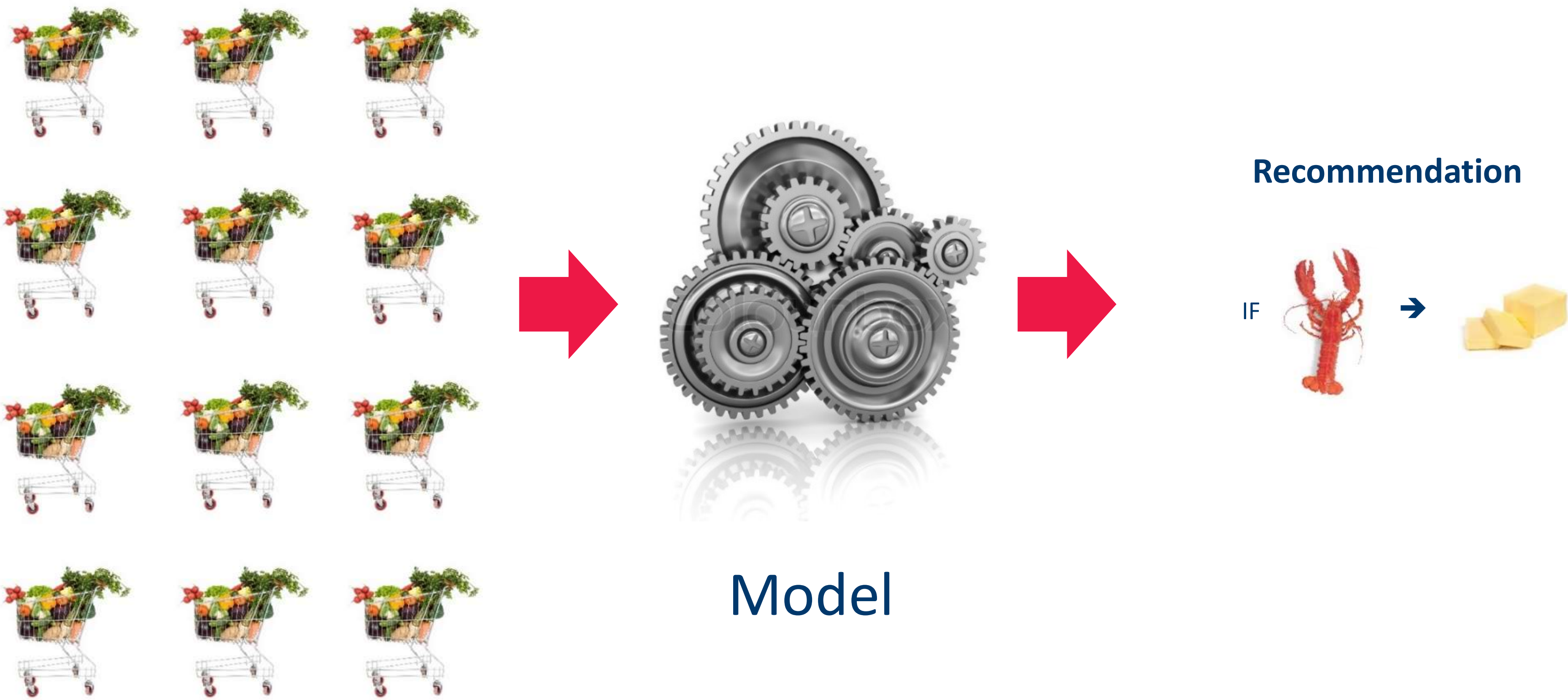
- How many taxis do I need in NYC on Wednesday at noon?
- Or how many kW will be required tomorrow at 6am in London?
- Or how many customers will come tonight to my restaurant?

Model



Recommendation Engines / Market Basket Analysis

– Recommendation Engines: People who bought this item were often interested in this other items.



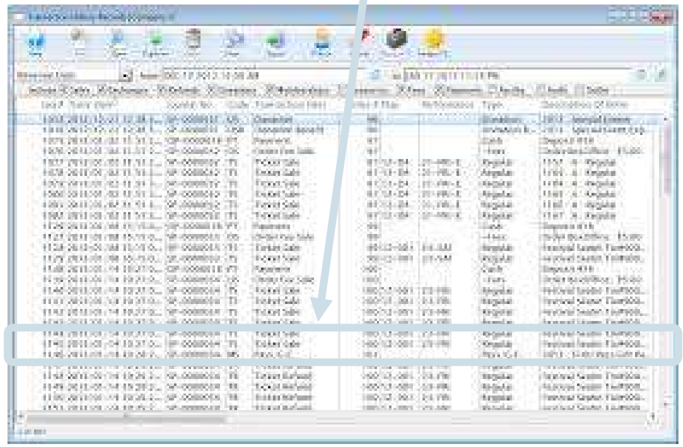
– Fraud Detection: Is this transaction legitimate or is it a fraud?

Transactions

- Trx 1
- Trx 2
- Trx 3
- Trx 4
- Trx 5
- Trx 6
- …

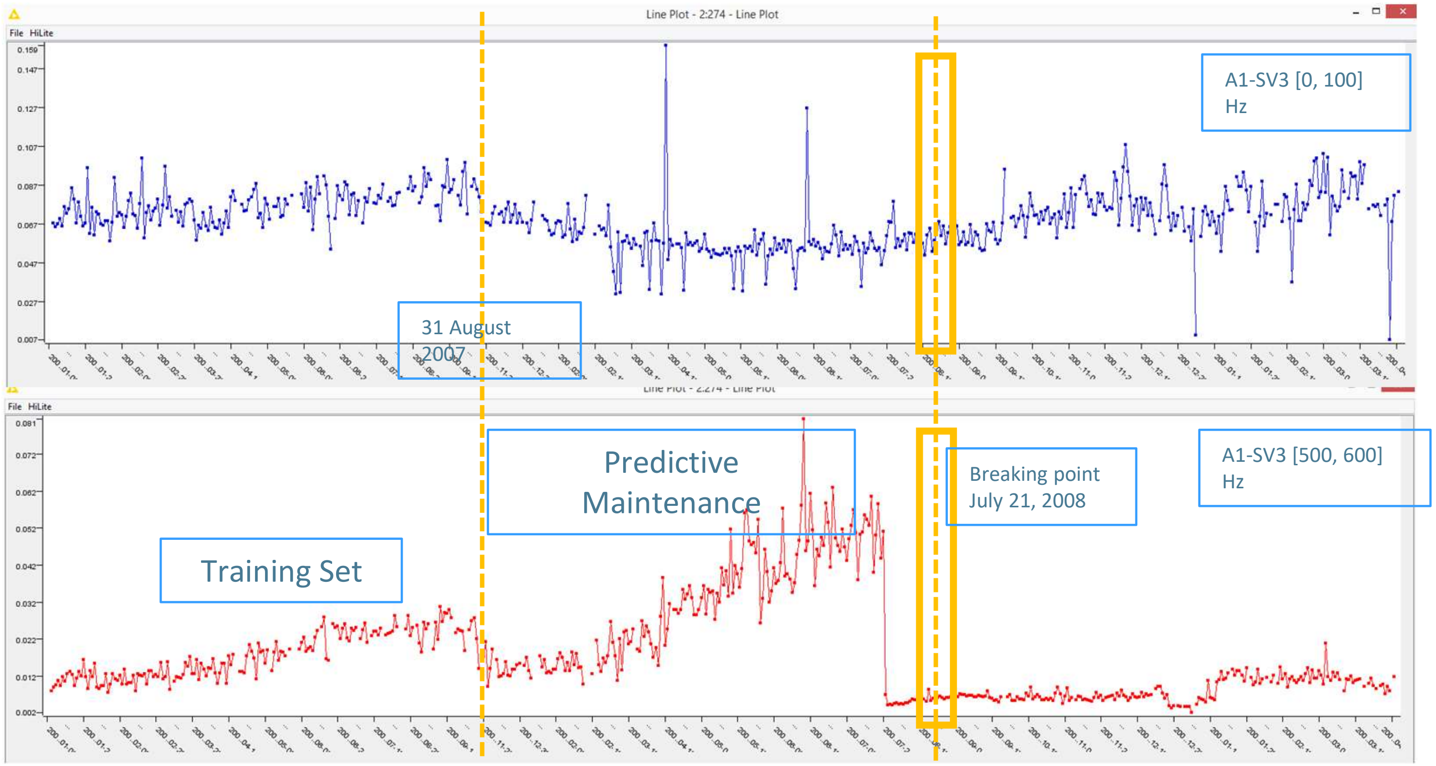
Model

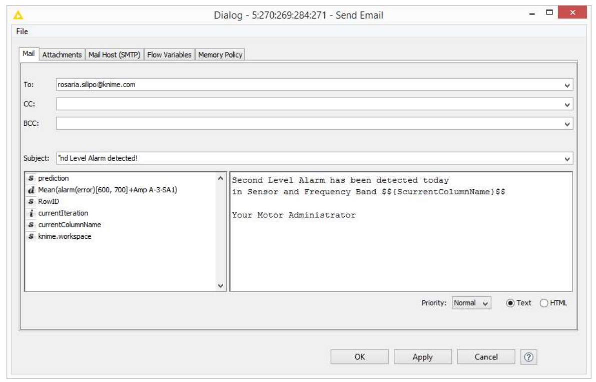
Suspicious Transaction



– Sentiment Analysis: how can I know what people are thinking?

Predicting mechanical failure as late as possible but before it happens





via REST

Only some Spectral Time Series show the break down

Project Understanding

Determine the Project Objective

- What is the primary objective?
- What are the criteria for success?
- These are difficult to define
  - The project owner & the analysis *speak different languages*

Problem source	Project owner perspective	Analyst perspective
Communication	Project owner does not understand the technical terms of the analyst	Analyst does not understand the terms of the domain of the project owner
Lack of understanding	Project owner was not sure what the analyst could do or achieve Models of analyst were different from what the project owner envisioned	Analyst found it hard to understand how to help the project owner
Organization	Requirements had to be adopted in later stages as problems with the data became evident	Project owner was an unpredictable group (not so concerned with the project)

Data Understanding

- **Goal of the Data Understanding phase**
  - Gain general insights about the data that will potentially be helpful for the further steps in the data analysis process
- **Reasons**
  - Never trust any data as long as you have not carried out some simple plausibility checks.
- **Results**
  - At the end of the data understanding phase, we know much better whether the assumptions we made during the project understanding phase concerning representativeness, informativeness, data quality, and the presence or absence of external factors are justified.

Attribute Understanding

No	Sex	Age	Blood pr.	Height	Drug
1	male	20	normal	175,0	A
2	female	73	normal	172,2	B
3	female	37	high	163,8	A
4	male	33	low	171,4	B
5	female	48	high	165,9	A
6	male	29	normal	182,3	A
7	female	52	normal	167,2	B
8	male	42	low	177,2	B
9	male	61	normal	168,4	B
10	female	30	normal	174,9	A

Attributes, features, variables…

Instances, records, data objects, entries…

- Data can usually be described in terms of table or matrices
- Sometimes data are spread among different table that need to be **joined**

Attribute Understanding

Categorical (Sex) · Ordinal (Blood pr.) · Numeric (Height)

No	Sex	Age	Blood pr.	Height	Drug
1	male	20	normal	175,0	A
2	female	73	normal	172,2	B
3	female	37	high	163,8	A
4	male	33	low	171,4	B
5	female	48	high	165,9	A
6	male	29	normal	182,3	A
7	female	52	normal	167,2	B
8	male	42	low	177,2	B
9	male	61	normal	168,4	B
10	female	30	normal	174,9	A

Numeric (Age) · Categorical (Drug)

- Attributes differ for their **scale type**, according to the type of values that they can assume
- Three scale types:
  - Categorical / Nominal
  - Ordinal
  - Numeric

Categorical Attributes

Categorical

No	Sex	Age	Blood pr.	Height	Drug
1	male	20	normal	175,0	A
2	female	73	normal	172,2	B
3	female	37	high	163,8	A
4	male	33	low	171,4	B
5	female	48	high	165,9	A
6	male	29	normal	182,3	A
7	female	52	normal	167,2	B
8	male	42	low	177,2	B
9	male	61	normal	168,4	B
10	female	30	normal	174,9	A

Categorical

- Categorical (or Nominal) attributes have a finite set of possible values
- Granularity must be taken into account
  - Hierarchical structure of the categories
  - e.g. shallow subdivision: *food, non-food, drinks…*
  - further subdivision for drinks: *water, beer, wine…*
  - Which level of granularity is appropriate?
- Dynamic Domain
  - Some attributes have a fixed domain (e.g. months)
  - For other attributes the domain can change over time (e.g. the products in a catalogue)
  - Those attributes must be identified and handled

Ordinal Attributes

Ordinal

No	Sex	Age	Blood pr.	Height	Drug
1	male	20	normal	175,0	A
2	female	73	normal	172,2	B
3	female	37	high	163,8	A
4	male	33	low	171,4	B
5	female	48	high	165,9	A
6	male	29	normal	182,3	A
7	female	52	normal	167,2	B
8	male	42	low	177,2	B
9	male	61	normal	168,4	B
10	female	30	normal	174,9	A

- Ordinal attributes have an additional linear ordering offered by the domain
- The ordering does not provide the distance between two object
- e.g. for an attribute containing university degrees, we can state that a *Ph.D* is an higher degree than a *M.Sc.* and that this is higher than a *B.Sc.*.

Attribute Understanding

Numeric continuous

No	Sex	Age	Blood pr.	Height	Drug
1	male	20	normal	175,0	A
2	female	73	normal	172,2	B
3	female	37	high	163,8	A
4	male	33	low	171,4	B
5	female	48	high	165,9	A
6	male	29	normal	182,3	A
7	female	52	normal	167,2	B
8	male	42	low	177,2	B
9	male	61	normal	168,4	B
10	female	30	normal	174,9	A

Numeric discrete

- The domain of numerical attributes are numbers. They can be
- **Discrete**
  - e.g. age, count…
  - Represented as integer values
- **Continuous**
  - e.g. height, weight, distance…
  - Represented as real values
  - Precision (rounding) has to be handled
- The scale of numeric attributes can be:
  - Interval e.g. date
  - Ratio Scale e.g. distance, with a canonical zero value
  - Absolute Scale e.g. counting

Data Quality

- Data quality refers to how well the data fit their intended use
- There are various data quality dimensions
  - Accuracy
  - Completeness
  - Unbalanced Data
  - Timeliness

Accuracy

Accuracy is defined as the closeness between the value in the data and the true value.

Syntactic

- The value might not be correct but it belongs at least to the domain of the corresponding attribute
- Easy to spot: verify values lying in the domain

e.g. “fmale” for the attribute Gender and “-15” for the attribute Weight violate the syntactic accuracy

Semantic

- The value might be in the domain of the corresponding attribute, but it is not correct
- Hard or impossible to spot: double check with other sources or check “business rules”

e.g. “*2090*” for the attribute *YearOfBirth* is (at least at the moment) surely incorrect, therefore violates the semantic accuracy

Completeness

- Completeness with respect to **attributes**
  - All the attributes have a value associated
  - i.e. Missing Values (coming soon in next lessons)
  - Missing values might not always be explicitly marked

- Completeness with respect to **records**
  - The data set contains the necessary information required for the analysis
  - Some rows might have been lost for various reasons (e.g. during DB migration)
  - Sometimes data about a certain situation simply does not exist (e.g. data about a failure that has never –yet- occurred)
  - It is hard to obtain a reasonably wide dataset containing all the possible combinations of data

Unbalanced Data

- Data regarding a certain situation might be underrepresented
- E.g. machine quality control: parts produced with flaws are – hopefully – lower than the correct ones, therefore the corresponding data will be way less

Timeliness

- Available data are too old to provide up to date information
- Often a problem in dynamically changing domains, where older data might indicate trends that have vanished

Describing your Data

Visual Inspection: Example

– Let’s look at our data

– Can we find some connections between age and shopping cart size?

– Anything else that looks a bit odd? (...the age distribution, maybe?)

– Visualizations are a good way for first sanity checks

– Interactivity on a plot or among plots is very helpful

Looking at the Data

Familiarize yourself with the data

- Identify trends
- strange patterns
- outliers
- ...

Types of views

- Basic Statistics
- 1D: Histograms
- 2D: Scatterplots, Scatter Matrix, Multi Dimensional Scaling
- 3D Scatterplots
- 3D: Parallel Coordinates

Simple Descriptors

- Simple statistical descriptors, such as:
  - range
  - mean/median
  - standard deviation
  - nominal values and their frequencies
  - ...
- can help to sanity check your data (and find dependencies that otherwise might surprise you quite a bit afterwards!)

- Can we look at the range and other simple 1D descriptors?
- How about 2D correlations between attributes?

Finding Patterns

Finding Patterns

- Finding (significant?) patterns in data may reveal interesting connections:

- Global patterns: groups of customers or products
  - Clusters

- Local patterns: connections between products, sub populations of customers (recommendation engines!)
  - Subgroups
  - Association Rules

Example

- Can we find groups of similar customers?
- (and what does similarity mean, anyway?)

- **Similarity**
- Finding the right similarity metric is an art.
- (and what is a cluster anyway?)
- Distance based methods in high dimensions offer all sorts of interesting surprises...

Finding Models

- Deriving models that describe (aspects of) the data:
  - Rules
  - Trees
  - Typical (or really odd!) examples
  - ...
- Models attempt to describe what is going on in the system that “generated” the data.

- Example:
  - Can we find a decision tree describing why certain customers buy so much?

Types of Data Processing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Reduce number of attributes

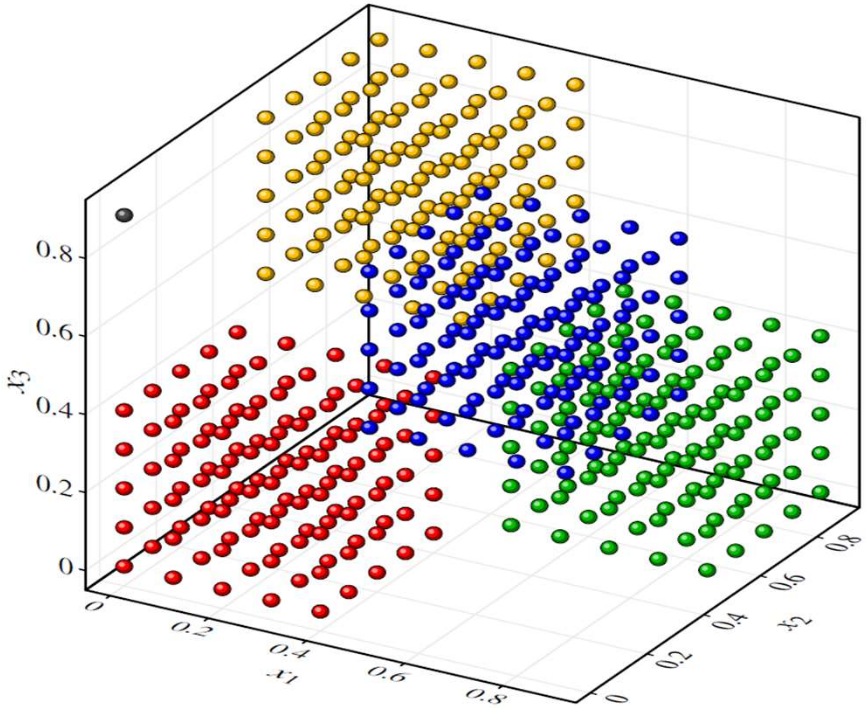
Data Preparation

- Feature Selection
- Dimensionality Reduction
- Sampling
- Outlier Detection
- Missing Value Imputation
- Data Cleaning & Standardization (domain dependent)
- Aggregations (often domain dependent)
- Normalization
- Feature Engineering
- Integration of multiple Data Sources

Select the Model

- Methods for One and Two Attributes
  - Barchart and Histogram
  - Boxplot
  - Scatter plot and density plot

- Methods for Higher-dimensional Data
  - Principal Component Analysis (PCA)
  - Multidimensional Scaling (MDS)
  - t-distributed Stochastic Neighbor Embedding
  - Parallel Coordinates
  - Radar and Star Plots
  - Sunburst Chart
  - Correlation Analysis



What‘s the best model to use?

Finding the “best” model is not a trivial task at all, since the question what a good (or best) model means is not always easy to answer.

From the Data:

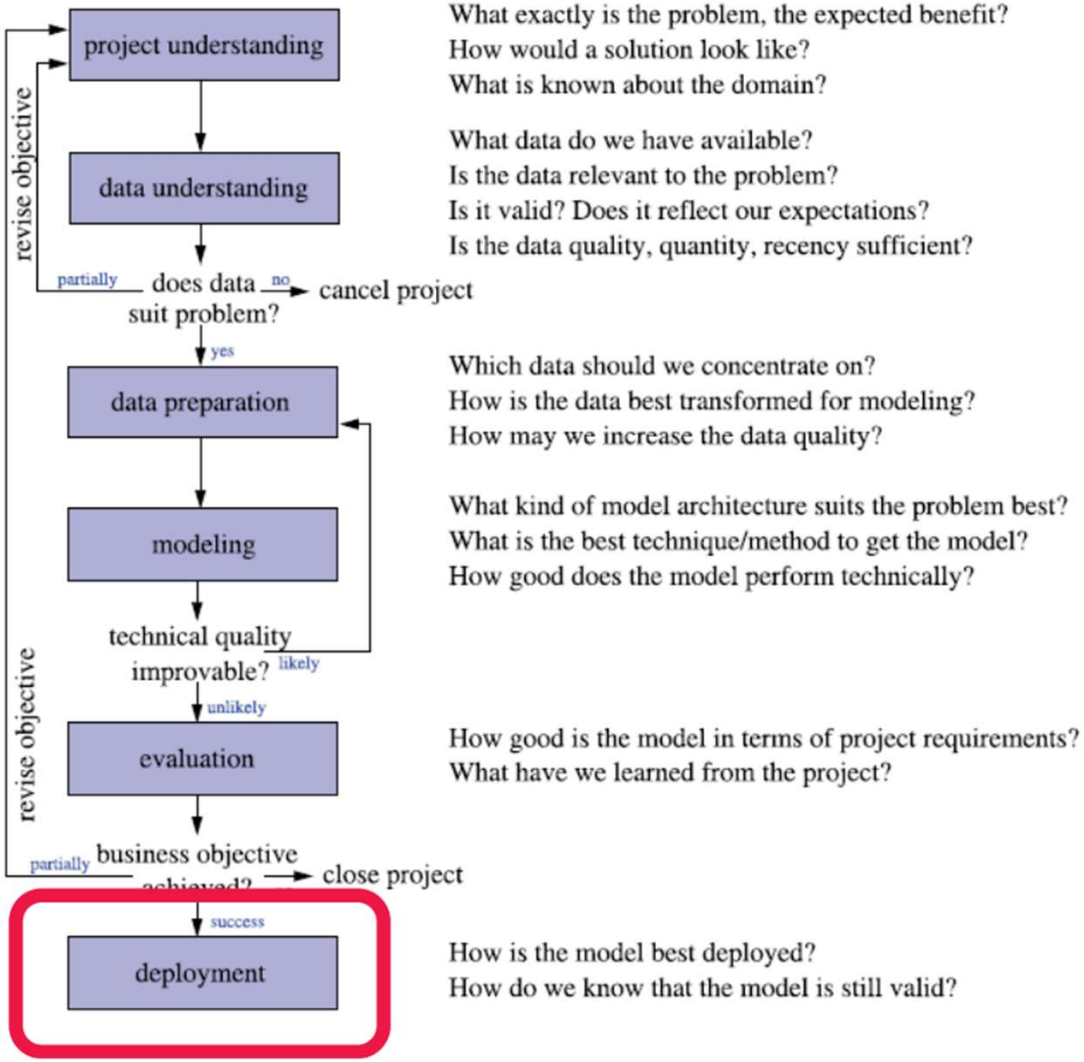
- Classification vs. Numerical
- Supervised vs. Unsupervised

From the business case:

- Performances: what is acceptable?
- Simplicity: do not use a cannon for a simple problem
- Interpretability: do I need to know the decision process?
- Computational costs: it must be trainable and applicable in a reasonable time with reasonable hardware

The Data Science Process

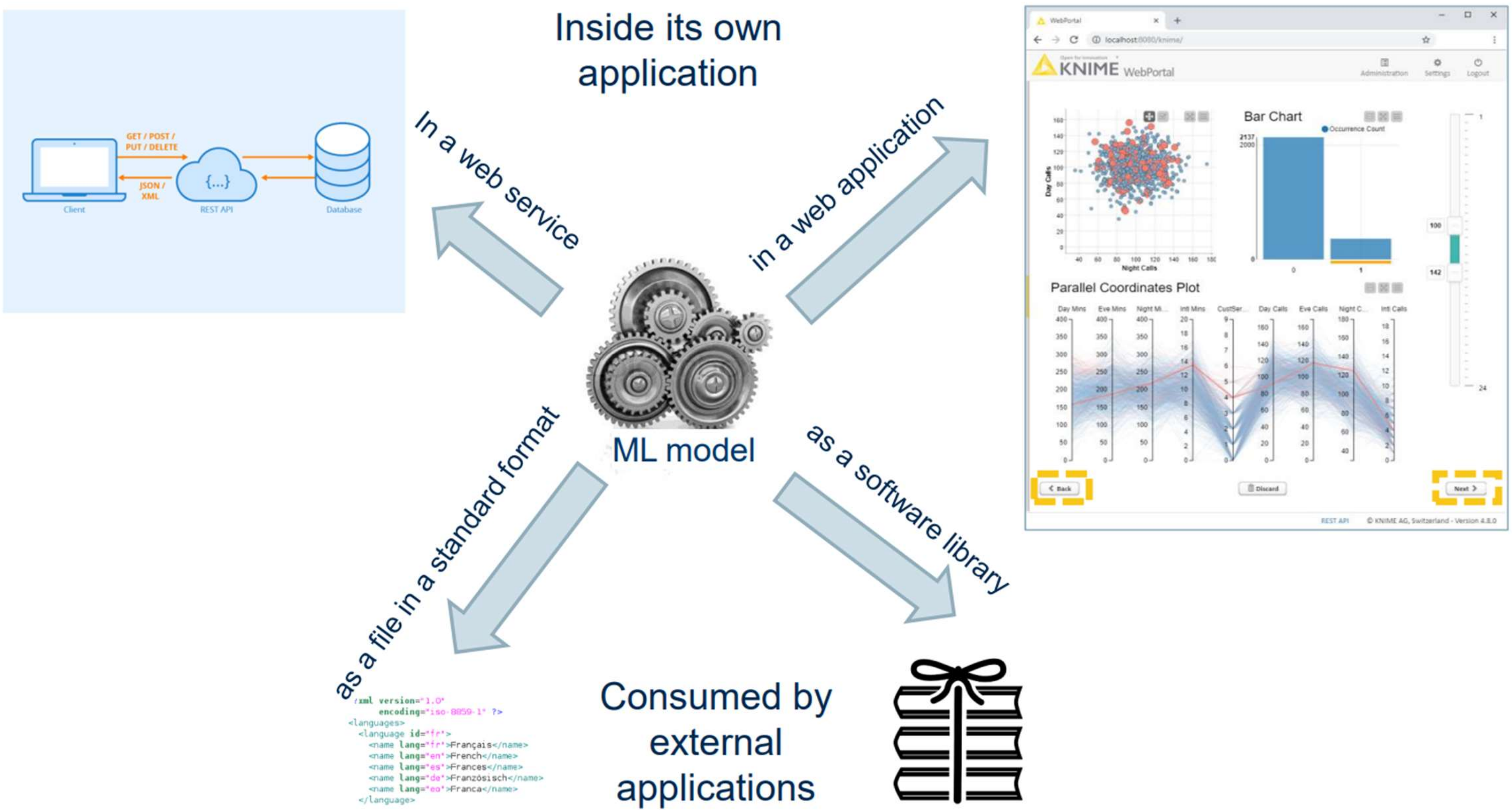
- **SEMMA**
  - Sample, Explore, Modify, Model, Assess
- **CRISP-DM**
  - Cross Industry Standard Process for Data Mining
- **KDD**
  - Knowledge Discovery in Databases



What is model deployment?

- Notice the dashed line between model testing and model deployment?
- This is where the jump **from the lab to the real world** happens
- Eventually a trained model must be included in a final application to be used **by external applications and/or end users**
- The final application is the deployment application
- The step of building the application around the trained model is called **deployment**
- Notice that the deployment application must be developed and finally put into production like all pieces of software
- When the deployment application is moved into production, so is the trained model

Deploying the ML Model



Deployment in a web application

- Interactive plots and charts
- Data selection across plots, charts, and tables
- Items such as: range slider, selection bullets, menus, …