Standard processes for data science & AI

Cosmin Lazar, EDS/EDS3 (Data Science and AI Group)

BOSCH

Industrial Standards for Data Science
Why Should There be a Standard Process?

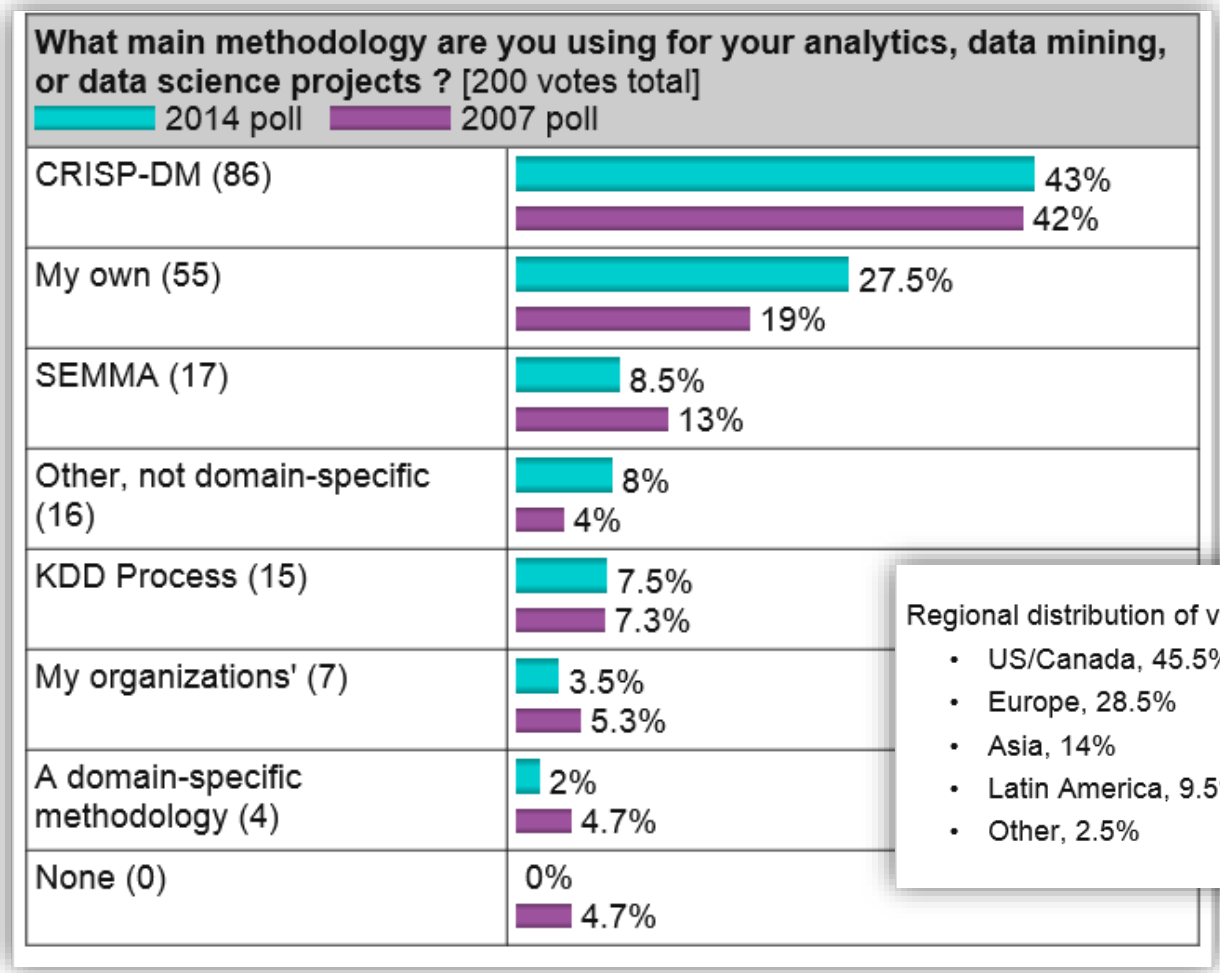
- The data science process must be reliable and repeatable - also by people with little data science background

- Framework for recording experience
  - Allows projects to be replicated
- Aid to project planning and management
- "Comfort factor" for new adopters
  - Demonstrates maturity of Data Science

Industrial Standards for Big Data Analytics
Poll Results: CRISP-DM still the top methodology

- CRISP-DM - top methodology for analytics, data mining, or data science projects

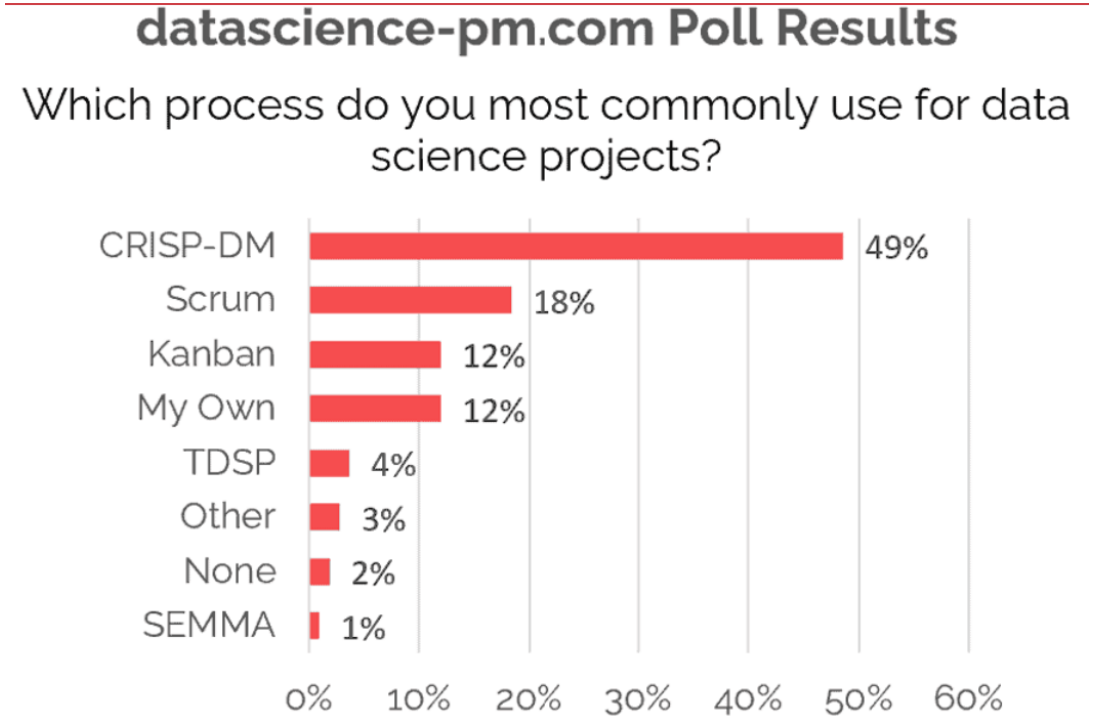
- Source: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html



Industrial Standards for Big Data Analytics
Poll Results: CRISP-DM still the top methodology

- **2022**: CRISP-DM – still top methodology for analytics, data mining, or data science projects

- Source: https://www.datascience-pm.com/crisp-dm-still-most-popular/

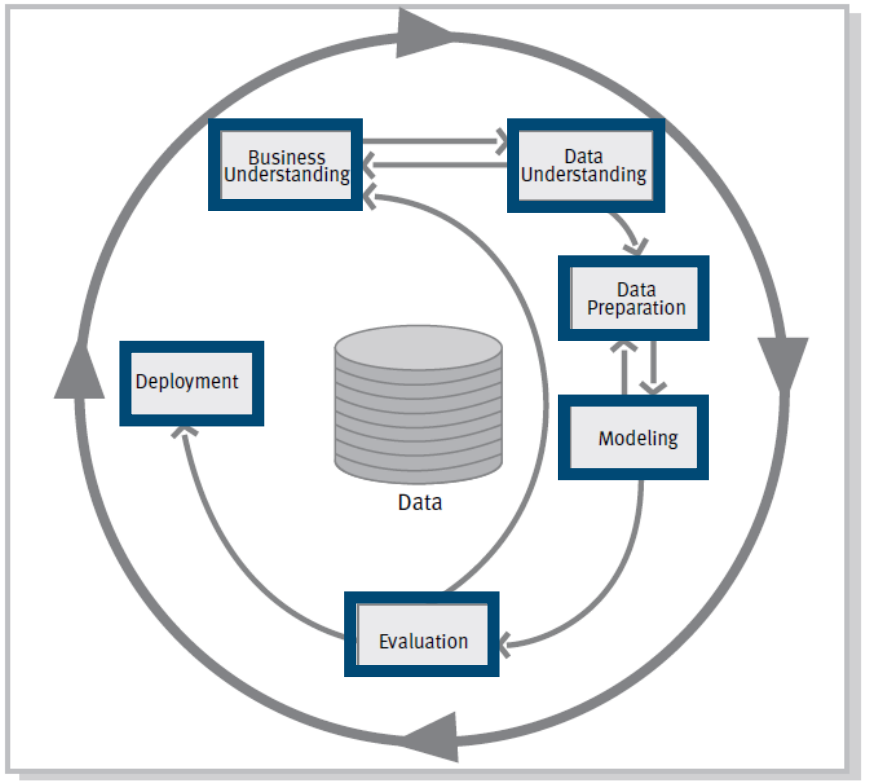


Cross Industry Standard for Data Mining
CRISP-DM

- Initiative launched in late 1996 by three “veterans” of data mining market.
  - Daimler Chrysler (then Daimler-Benz), SPSS (now part of IBM IBM) , NCR
- Developed and refined through series of workshops (from 1997-1999)
- Over 300 organization contributed to the process model
- Published CRISP-DM 1.0 (1999):
  - First Version: *https://www.the-modeling-agency.com/crisp-dm.pdf*
  - With Examples: *ftp://public.dhe.ibm.com/software/analytics/.../**CRISP_DM.pdf***

Industrial Standards for Big Data Analytics
Cross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

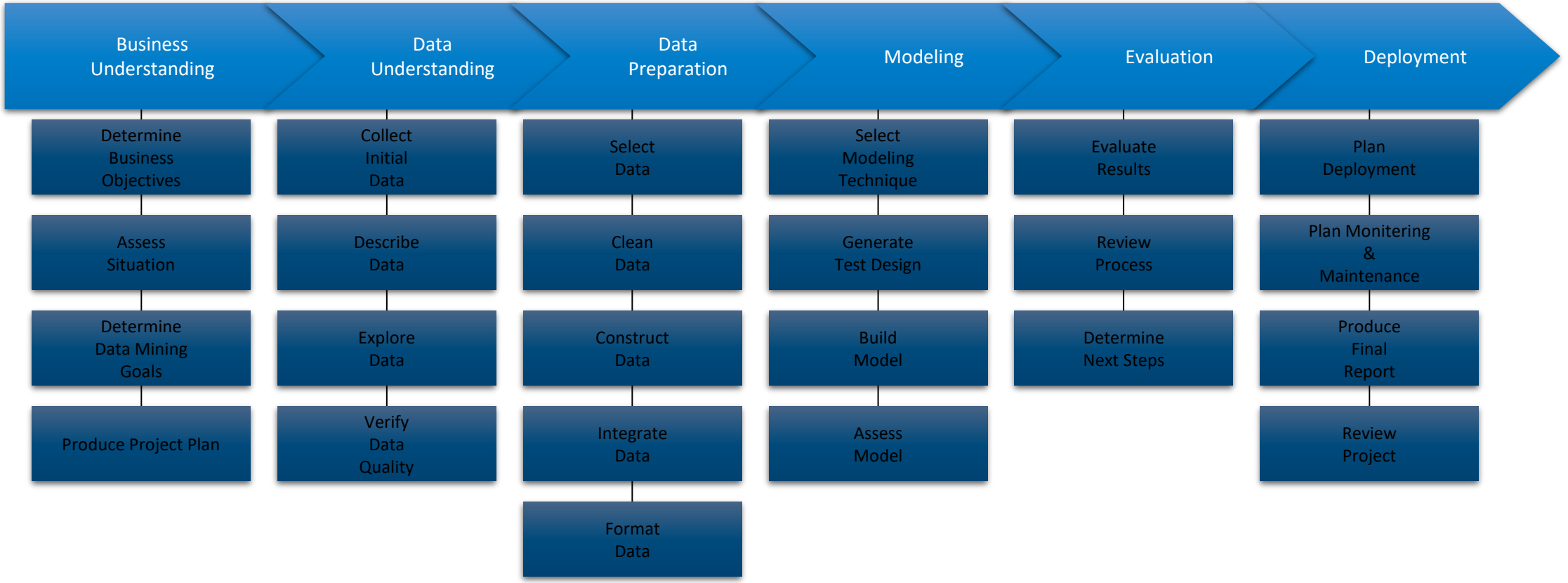
CRISP-DM:



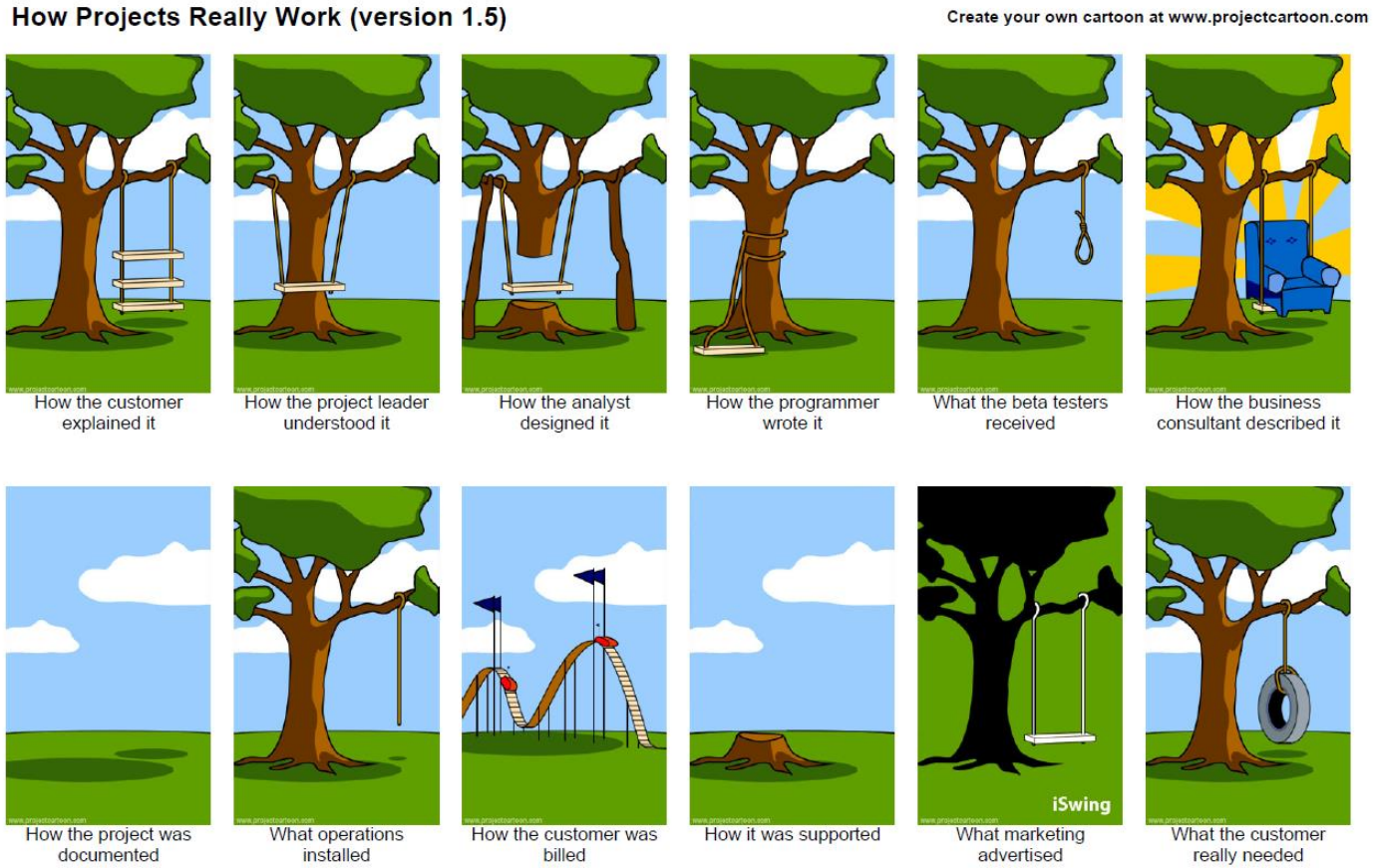
*https://the-modeling-agency.com/**crisp-dm**.pdf*

CRISP-DM is a comprehensive data mining methodology and process model that provides anyone- from novices to data mining Experts- with a complete blueprint for conducting a data mining project. CRISP-DM breaks down the life cycle of a data mining project into six phases

Cross Industry Standard for Data Mining
CRISP-DM



Industrial Standards for Big Data Analytics
Business Understanding?



Source: http://projectcartoon.com/

How Projects Really Work (version 1.5)

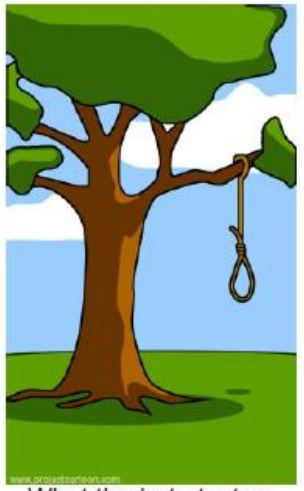
Create your own cartoon at www.projectcartoon.com


How the customer explained it

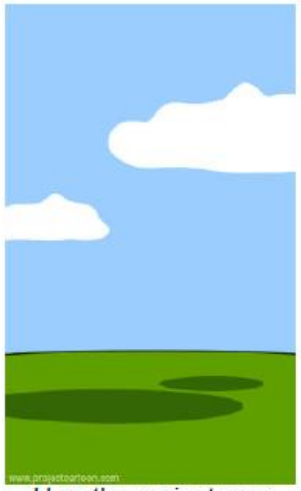

How the project leader understood it

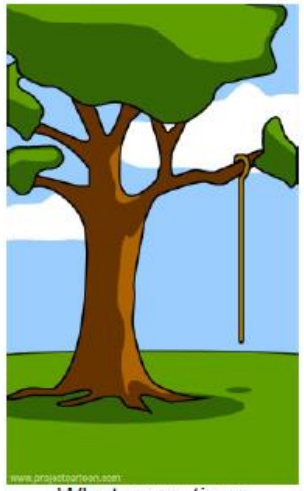

How the analyst designed it


How the programmer wrote it

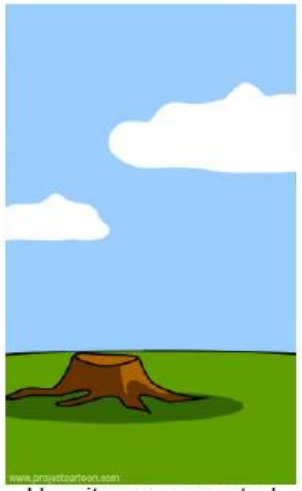

What the beta testers received


How the business consultant described it


How the project was documented


What operations installed


How the customer was billed


How it was supported

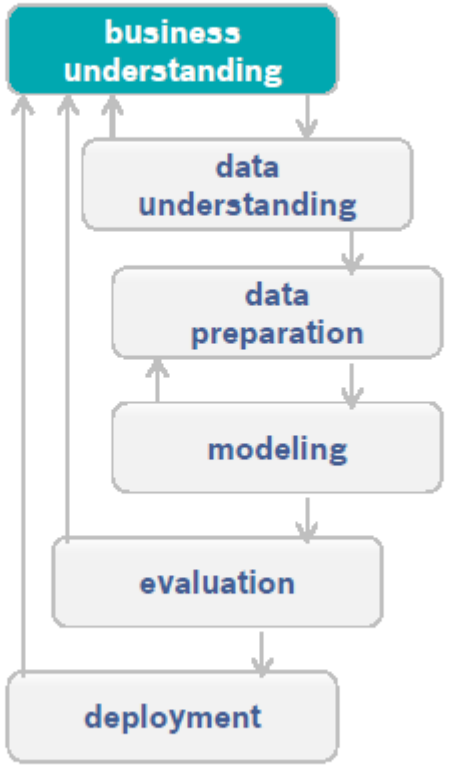
What marketing advertised


What the customer really needed

BOSCH

Introduction – Business Understanding

Data Mining Steps (CRISP-DM)*



What is the **problem** that data mining should solve for my business?

What are **success criteria** for the data mining activity?

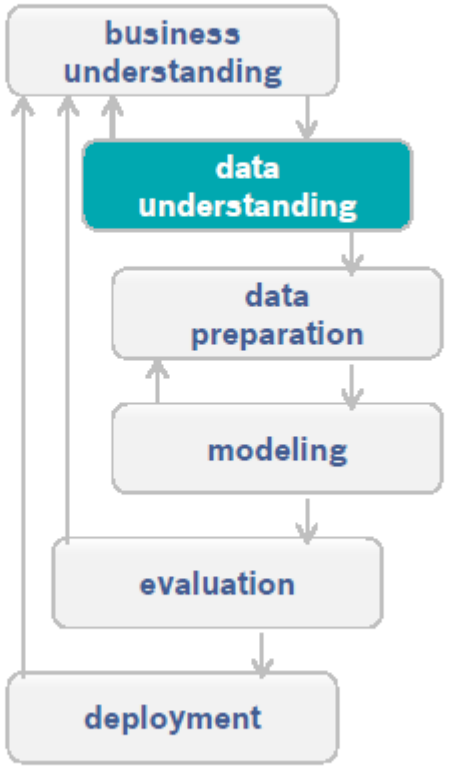
What roles / competencies / collaborations do I need for the project?

Tasks involved: Determine business objectives and requirements, assess situation, determine data mining problem definition and objectives

* Cross-Industry Standard Process for Data Mining

Introduction – Data Understanding

Data Mining Steps (CRISP-DM)*



What **type of data** do you have?

(Structured, text, image, audio, sensor signals, ...)

What is the **update rate** of your data? *(Batch, streaming)*

Are there **multiple** data sources? How can I combine them?

What **errors**, **inconsistencies** or **missing values** are in the data? What are sources for them?

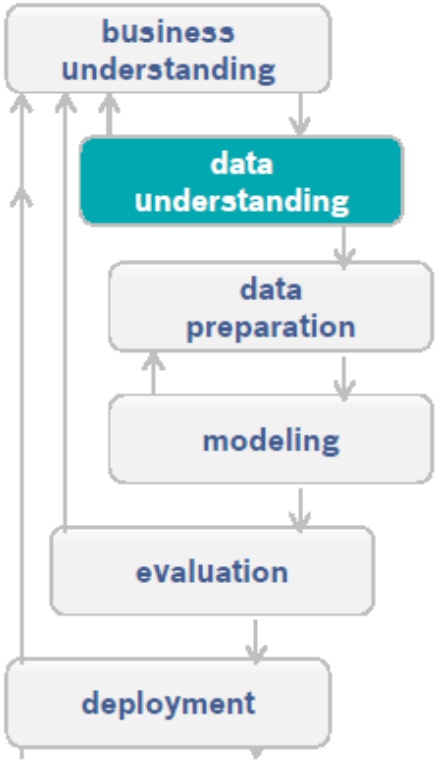
What does your data mean?

Tasks involved: Collect initial data, describe data, explore data, and verify data

* Cross-Industry Standard Process for Data Mining

Introduction – Data Understanding

Data Mining Steps (CRISP-DM)*



Seminar

Student

Attribute	Length	Type	Rules
Name	40	Alpha	At least 2 words
Email Address	50	Mixed	Must contain @
Phone #	10	Numeric	Reject all "555"
Address	30	Mixed	Format - ### alpha
City	20	Alpha	none
State	2	Alpha	Must be a valid state

Data dictionary

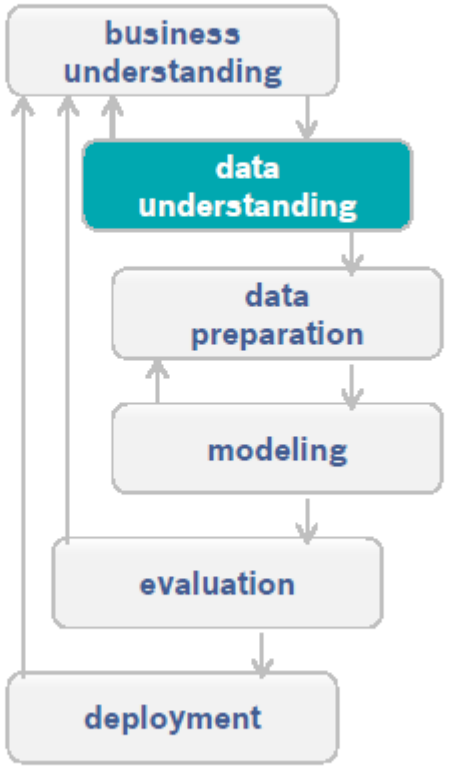
summarizes domain-related and technical information about your data. Documentation of what data you have, what it means and helps to identify inconsistencies and errors.

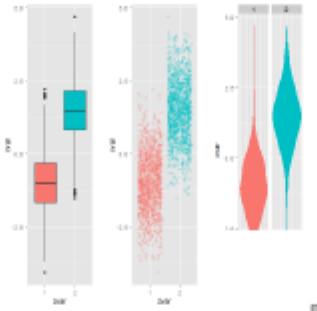
Tasks involved: Collect initial data, describe data, explore data, and verify data

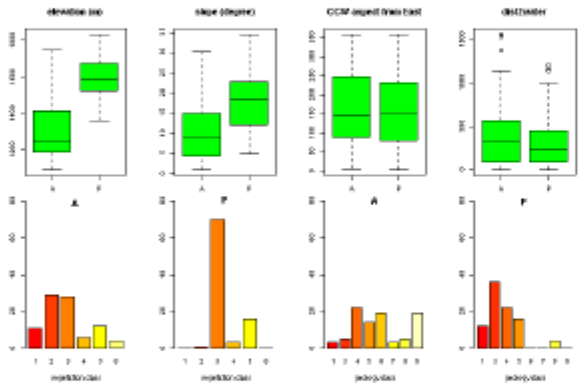
* Cross-Industry Standard Process for Data Mining

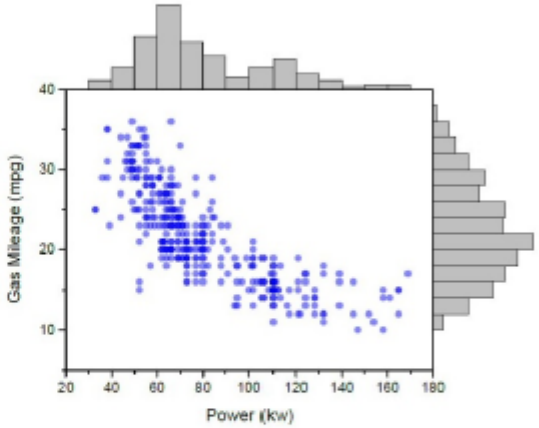
Introduction – Data Understanding

Data Mining Steps (CRISP-DM)*









Descriptive statistics

Explore and describe data using histograms, mean and variance, bar plots, box plots, scatter plots, densities, etc.

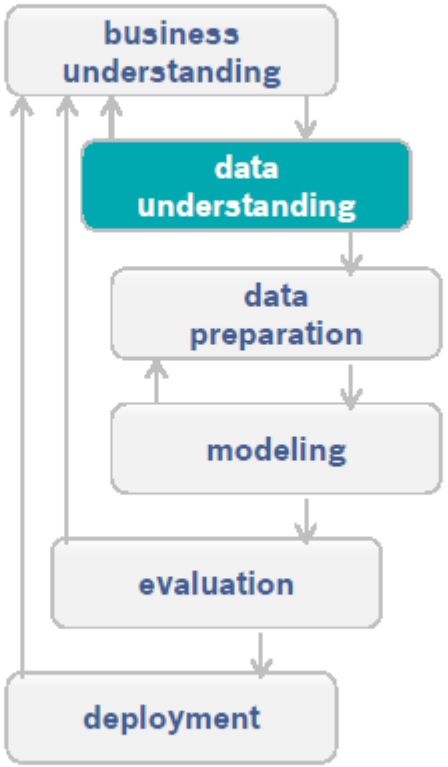
Tasks involved: Collect initial data, describe data, explore data, and verify data

* Cross-Industry Standard Process for Data Mining

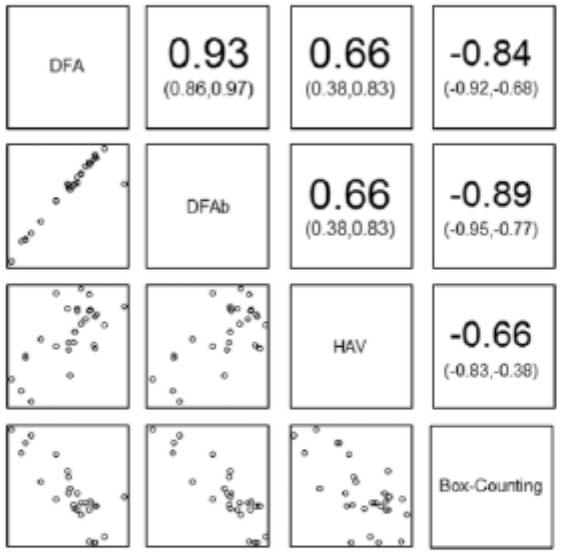
BOSCH

Introduction – Data Understanding

Data Mining Steps (CRISP-DM)*



(b) Fractal-based Dive Parameters



Associations and correlations
between attributes / features can be measured by different correlation coefficients

Pearson's correlation coefficient (linear):

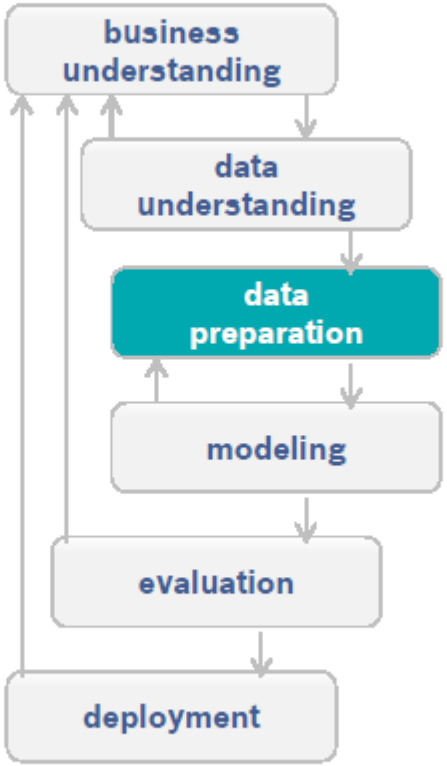
$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

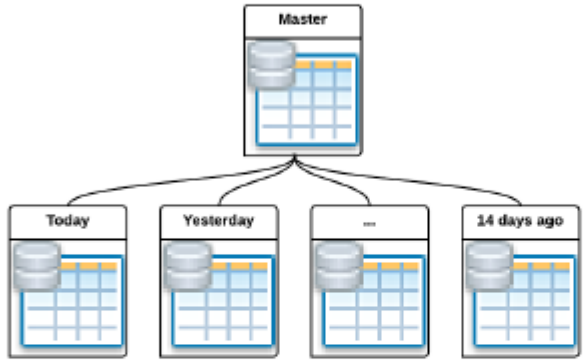
Tasks involved: Collect initial data, describe data, explore data, and verify data

* Cross-Industry Standard Process for Data Mining

Introduction – Data Preparation

Data Mining Steps (CRISP-DM)*





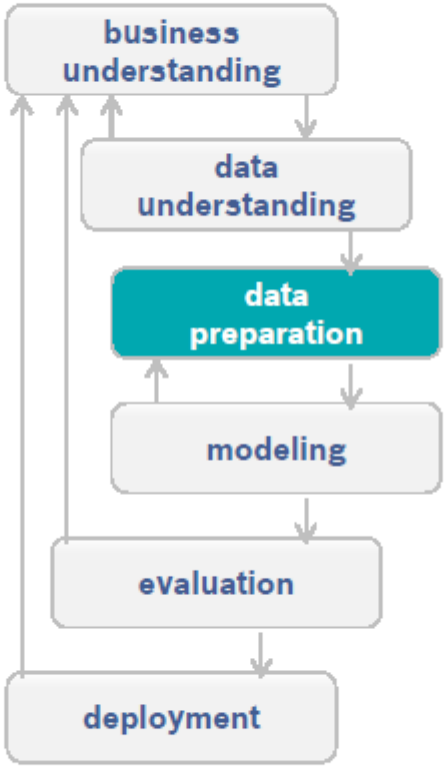
Selecting data I need and gather them in **one master table** – final dataset to feed to modeling

Tasks involved: Select data, clean data, construct data, integrate data, and format data

* Cross-Industry Standard Process for Data Mining

Introduction – Data Preparation

Data Mining Steps (CRISP-DM)*



	A	B	C	D
1	Main Category	Category	Sub Category	Defects
2	Mechanical	Mechanical	Gear	11
3	Mechanical	Mechanical	Bearing	8
4	Mechanical	Mechanical	Motor	3
5	Electrical	Electrical	Switch	19
6	Electrical	Electrical	Plug	12
7	Electrical	Electrical	Cord	11
8	Electrical	Electrical	Fuse	3
9	Electrical	Electrical	Bulb	2
10	Hydraulic	Hydraulic	Pump	4
11	Hydraulic	Hydraulic	Leak	3
12	Hydraulic	Hydraulic	Seals	1

Pivoting

your master table: turning several rows into one, resulting in a less normalized but more compact table

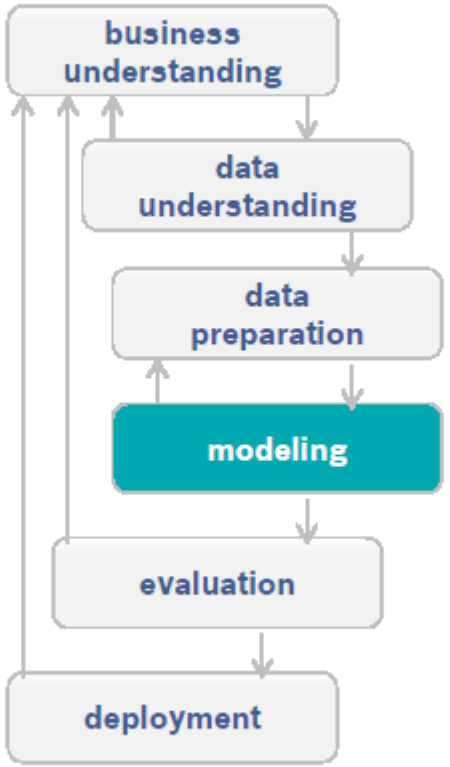
Sum of Defects		Category		
Main Category	Sub Category	Electrical	Mechanical	Hydraulic
Electrical	Switch	19		
	Plug	12		
	Cord	11		
	Fuse	3		
	Bulb	2		
Mechanical	Gear		11	
	Bearing		8	
	Motor		3	
Hydraulic	Pump			4
	Leak			3
	Seals			1

Tasks involved: Select data, clean data, construct data, integrate data, and format data

* Cross-Industry Standard Process for Data Mining

Introduction – Modeling

Data Mining Steps (CRISP-DM)*



... algorithms

• SVM • nearest neighbor • Bayes • logistic regress. • aNN	• linear regression • kernel regression • Gaussian proc.	• boolean rules • decision trees • Apriori alg. • FP-Growth alg.	• k-means alg. • DBSCAN • EM algorithm • GMM	• local outlier factor • k-nearest neighbor • replicator neural networks • one-class SVM
Classification	**Regression**	**Association analysis**	**Cluster analysis**	**Anomaly detection**

Problem Classes...

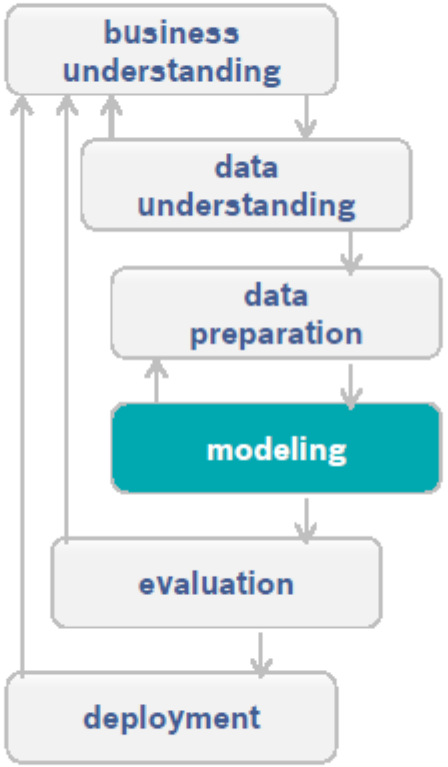
predict (create models)	**relate** (combine inform. that's known to exist)	**explain** (discover new knowledge)

Tasks involved: Select model techniques, generate test design, build model, and asses model

* Cross-Industry Standard Process for Data Mining

Introduction – Modeling

Data Mining Steps (CRISP-DM)*



Assess a model's **classification performance** using a confusion matrix

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Sensitivity = TP / (TP+FN)
Specificity = TN / (TN+FP)

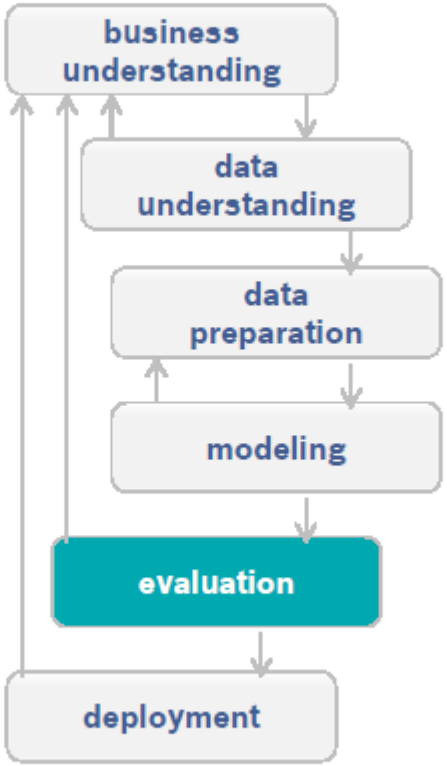
	Spam (Predicted)	Non-Spam (Predicted)	Accuracy
Spam (Actual)	27	6	81.81
Non-Spam (Actual)	10	57	85.07
Overall Accuracy			83.44

Tasks involved: Select model techniques, generate test design, build model, and asses model

* Cross-Industry Standard Process for Data Mining

Introduction – Evaluation

Data Mining Steps (CRISP-DM)*



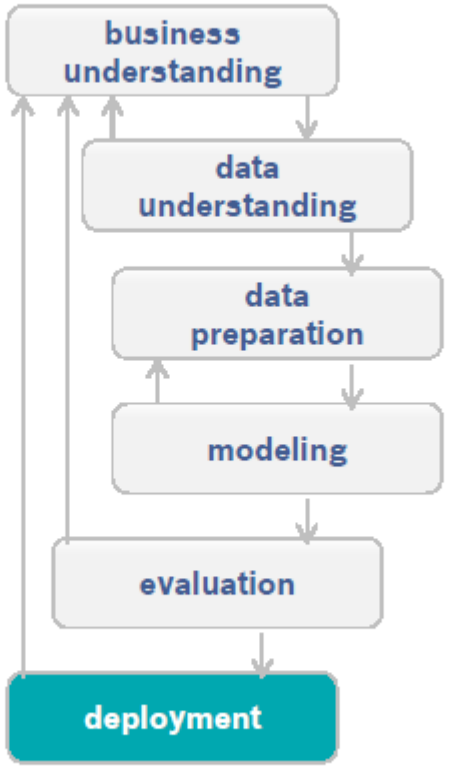
- How to **interpret** the results in terms of the application?
- Do the results fulfill the **data mining goal**? Do they contribute to your business objectives?
- What are findings besides the model? Are these findings related to the **original business goal**?

Tasks involved: Evaluate results, review process, and determine next steps

* Cross-Industry Standard Process for Data Mining

Introduction – Deployment

Data Mining Steps (CRISP-DM)*



- What are **deployable results**?
- How will **information propagate** to users and decision makers? What is the interface?
- What is the **execution model** and the **necessary IT infrastructure**? How will it be maintained?
- How will the use of the model be monitored? How to get **user feedback**?

Tasks involved: Plan deployment, plan monitoring and maintenance, produce final report, and review project

* Cross-Industry Standard Process for Data Mining

Industrial Standards for Big Data Analytics
CRISP-DM References

- Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz, (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler) “CRISP-DM 1.0 - Step-by-step data mining guide”

- “The CRISP-DM Model: The New Blueprint for DataMining”, Colin Shearer, JOURNAL of Data Warehousing, Volume 5, Number 4, p. 13-22, 2000

- https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
- http://www.kdnuggets.com/2016/03/data-science-process-rediscovered.html/2