Data mining and Machine Learning

Valerio Consorti
mail: valerio.consorti@generali.com

Overview on the course

Goal:

- To be able to *design* and *implement* simple end-to-end data analytics process

To be able to design…

- To understand the various logical steps that constitute a typical data analytics process
- To be able to build the proper set of variable to describe the data
- To choose the proper model to solve the specific problem
- To be able to properly evaluate the performance of the analysis

To be able to implement…

- To become familiar with the python package scikit-learn
- To be able to implement custom transformers and estimators
- To be able to implement complex multi-step analysis

Structure of the lectures

	Wednesday 03/03	Tursday 04/03	Friday 05/03
09:30 - 11:00	*Theory:* Data Analysis & Regression problems	*Theory:* Classification problems	*Theory:* Ensemble models for classification
11:15 - 11:30	break	Break	break
11:30 -13:00	*Technology:* scikit-learn: Transformers & Estimators	*Technology:* scikit-learn: pipelines, and model optimization	*Practice:* Advanced classification
… - 14:00	Lunch break	Lunch break	Lunch break
14:00 - 18:00	*Practice:* Linear regression	*Practice:* Simple classification	*Practice:* Advanced classification

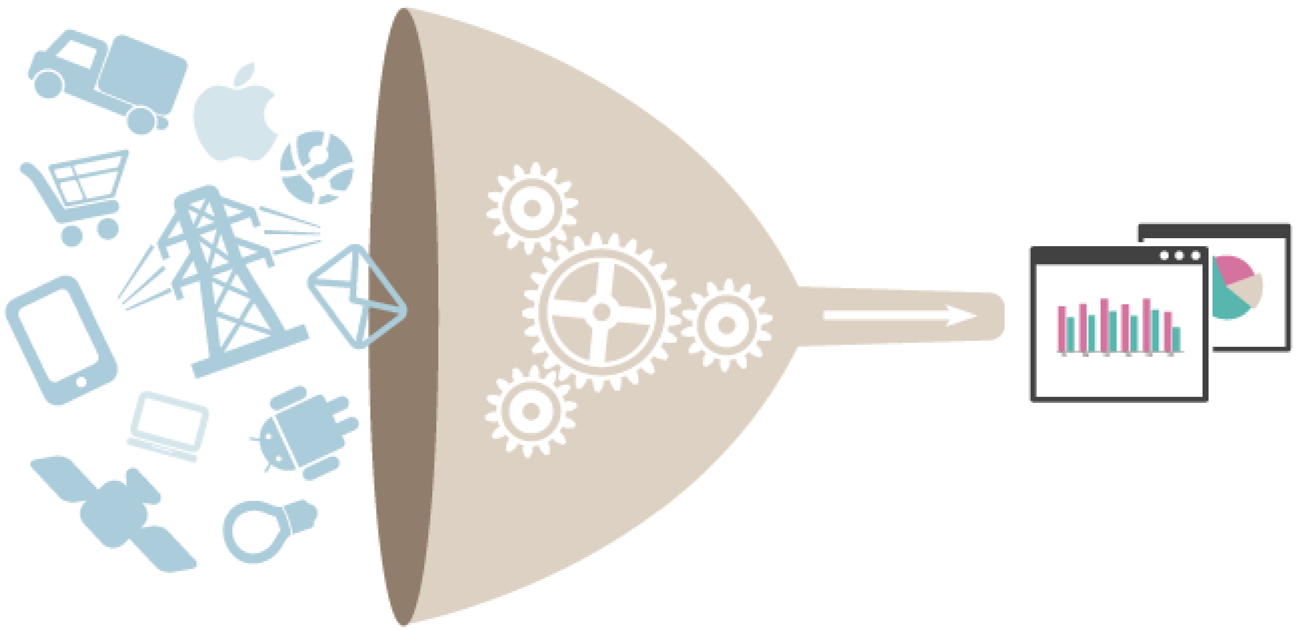
My goal:

- To provide you the minimal theoretical background necessary to pragmatically approach an analytical problem
- To provide you some technology that enable you to implement your own analytical solution
- *To let you exercise facing some real-word problem building an hands-on experience on data analytics*

Day 1: The data analytics process

What people think data analytics look like…

Collect data → The Machine process the data → Get the results



What people think data analytics look like…

Collect data → The Machine process the data → Get the results

What people think data analytics look like…

Collect data

- Which data?
- How do they look like?
- Which are the available variables?
    - How are they distributed?
- Is there any missing value?
- …

The Machine process the data

- Which kind of processing?
- Which model?
- …

Get the results

- Which results do I expect?
- …

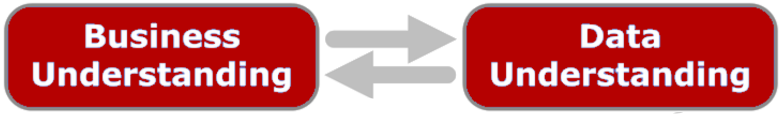
NO WAY

How data analytics actually look like…

Business Understanding

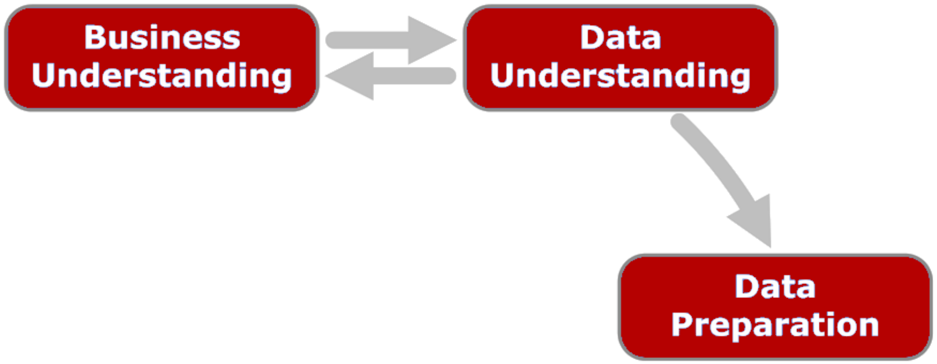
- *Business understanding*:
  - Focus on the business problem in terms of objectives and requirements
  - Translate the business problem into a data-mining problem

How data analytics actually look like…



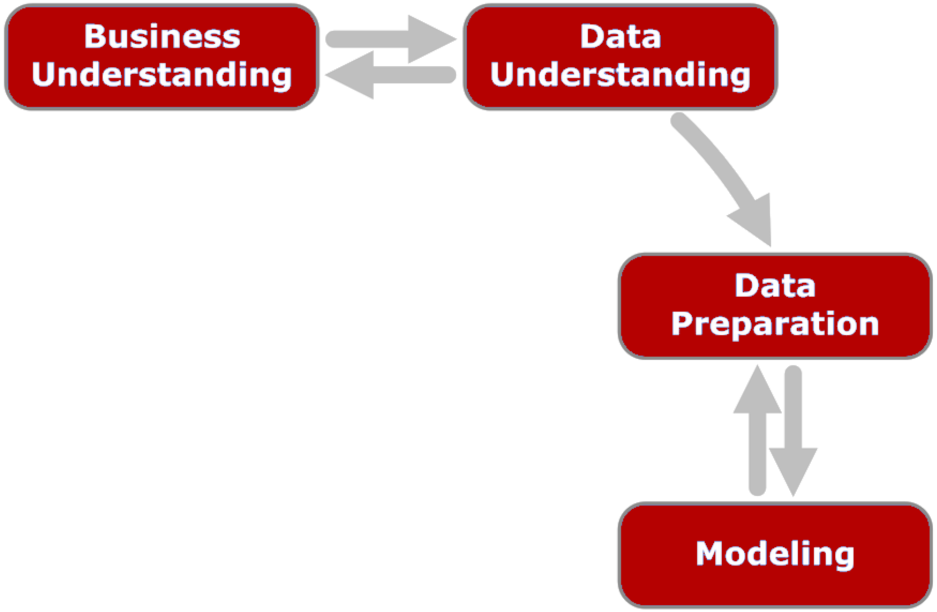
- *Business understanding*:
    - Focus on the business problem in terms of objectives and requirements
    - Translate the business problem into a data-mining problem
- *Data understanding:*
    - Data collection
    - Data exploration: variables, data quality, get first insight into data

How data analytics actually look like…



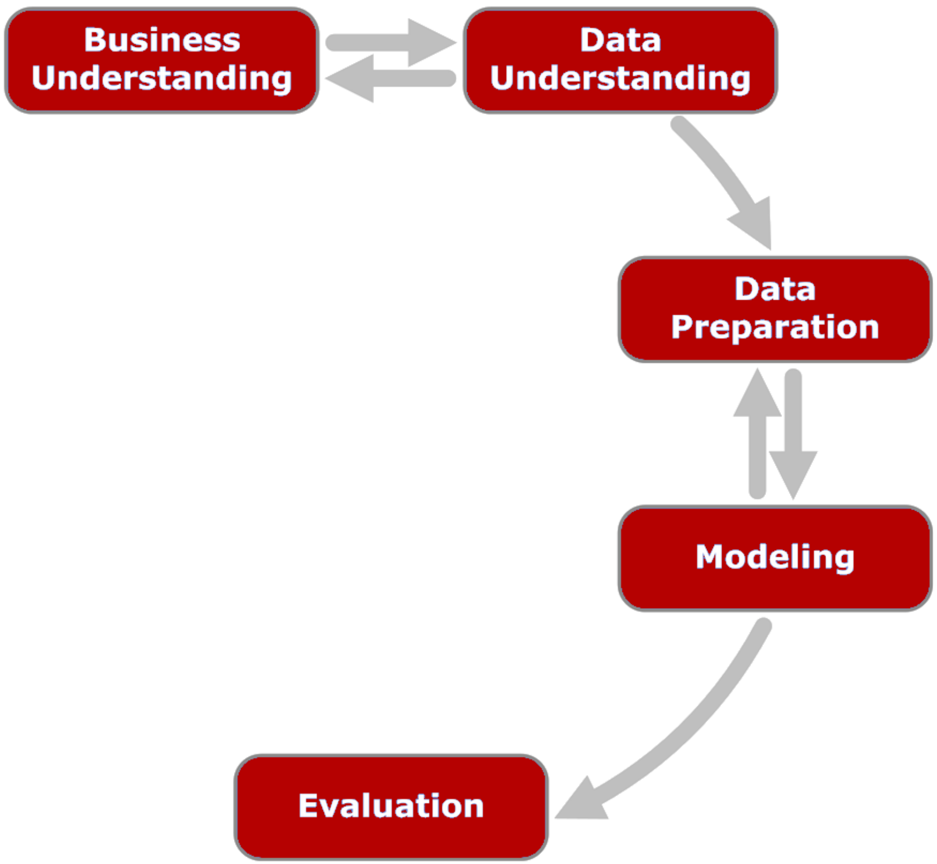
- *Business understanding*:
  - Focus on the business problem in terms of objectives and requirements
  - Translate the business problem into a data-mining problem
- *Data understanding:*
  - Data collection
  - Data exploration: variables, data quality, get first insight into data
- *Data preparation:*
  - Data cleaning, selection, transformation …
  - Definition of the final data-set to be used in the analysis

How data analytics actually look like…



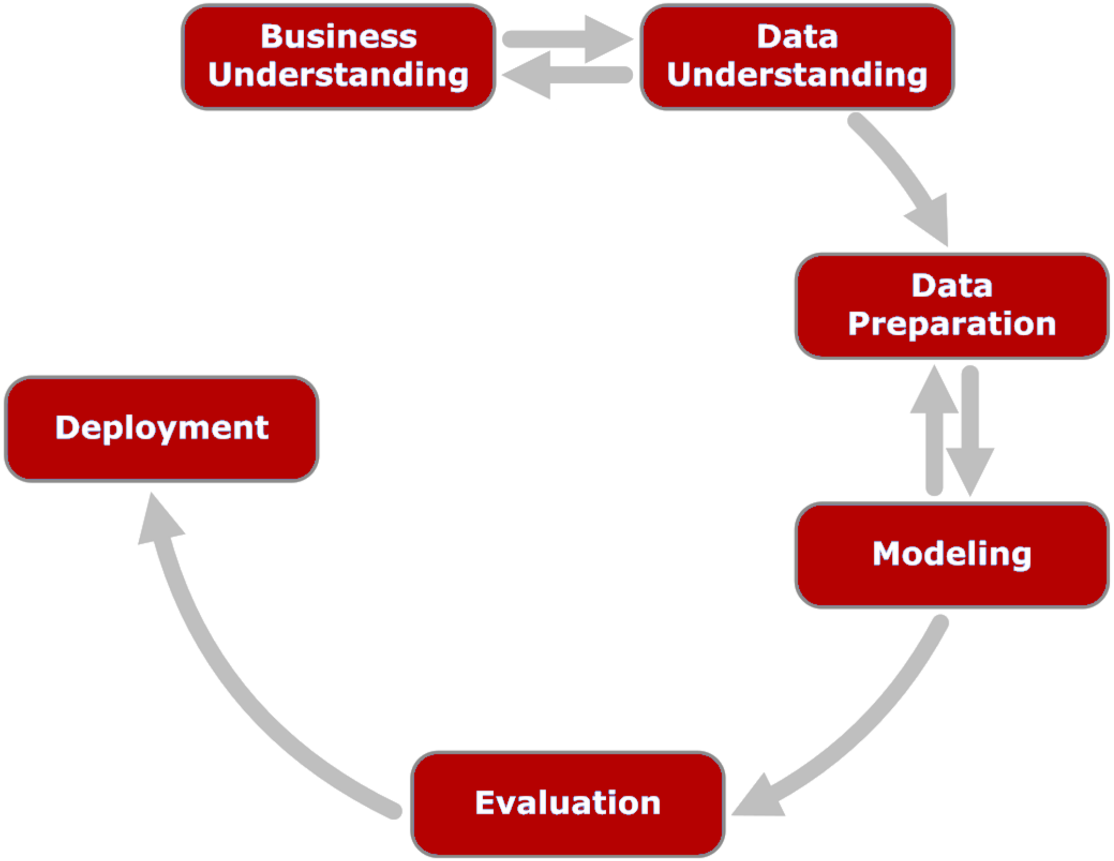
- *<u>Business understanding</u>*:
  - Focus on the business problem in terms of objectives and requirements
  - Translate the business problem into a data-mining problem
- *<u>Data understanding:</u>*
  - Data collection
  - Data exploration: variables, data quality, get first insight into data
- *<u>Data preparation:</u>*
  - Data cleaning, selection, transformation …
  - Definition of the final data-set to be used in the analysis
- *<u>Modelling:</u>*
  - Select and optimize models that can better describe the problem

How data analytics actually look like…



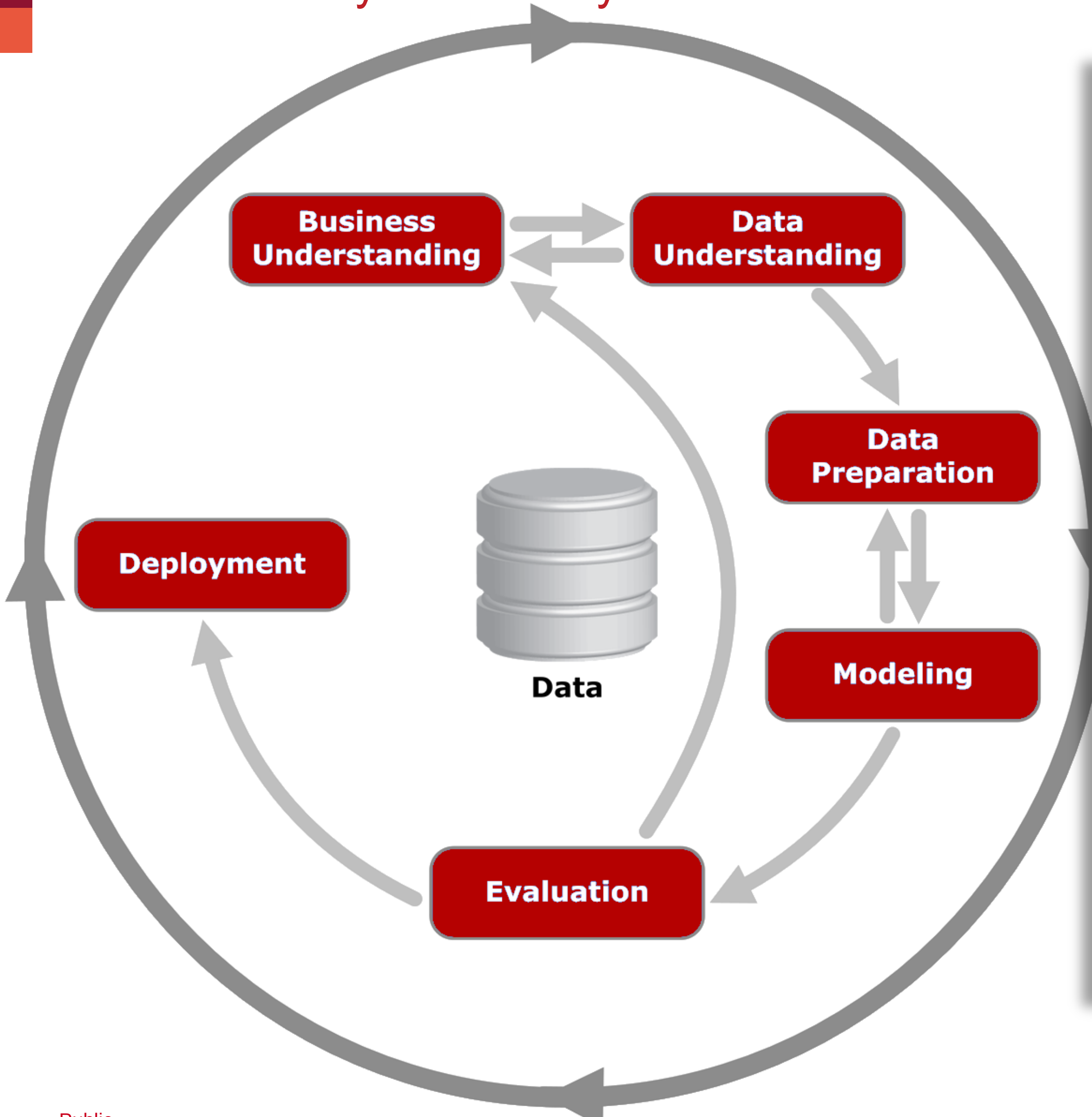
- *<u>Business understanding</u>*:
  - Focus on the business problem in terms of objectives and requirements
  - Translate the business problem into a data-mining problem
- *<u>Data understanding:</u>*
  - Data collection
  - Data exploration: variables, data quality, get first insight into data
- *<u>Data preparation:</u>*
  - Data cleaning, selection, transformation …
  - Definition of the final data-set to be used in the analysis
- *<u>Modelling:</u>*
  - Select and optimize models that can better describe the problem
- *<u>Evaluation:</u>*
  - Evaluate the performance of the analysis process in terms of business requirements

How data analytics actually look like…



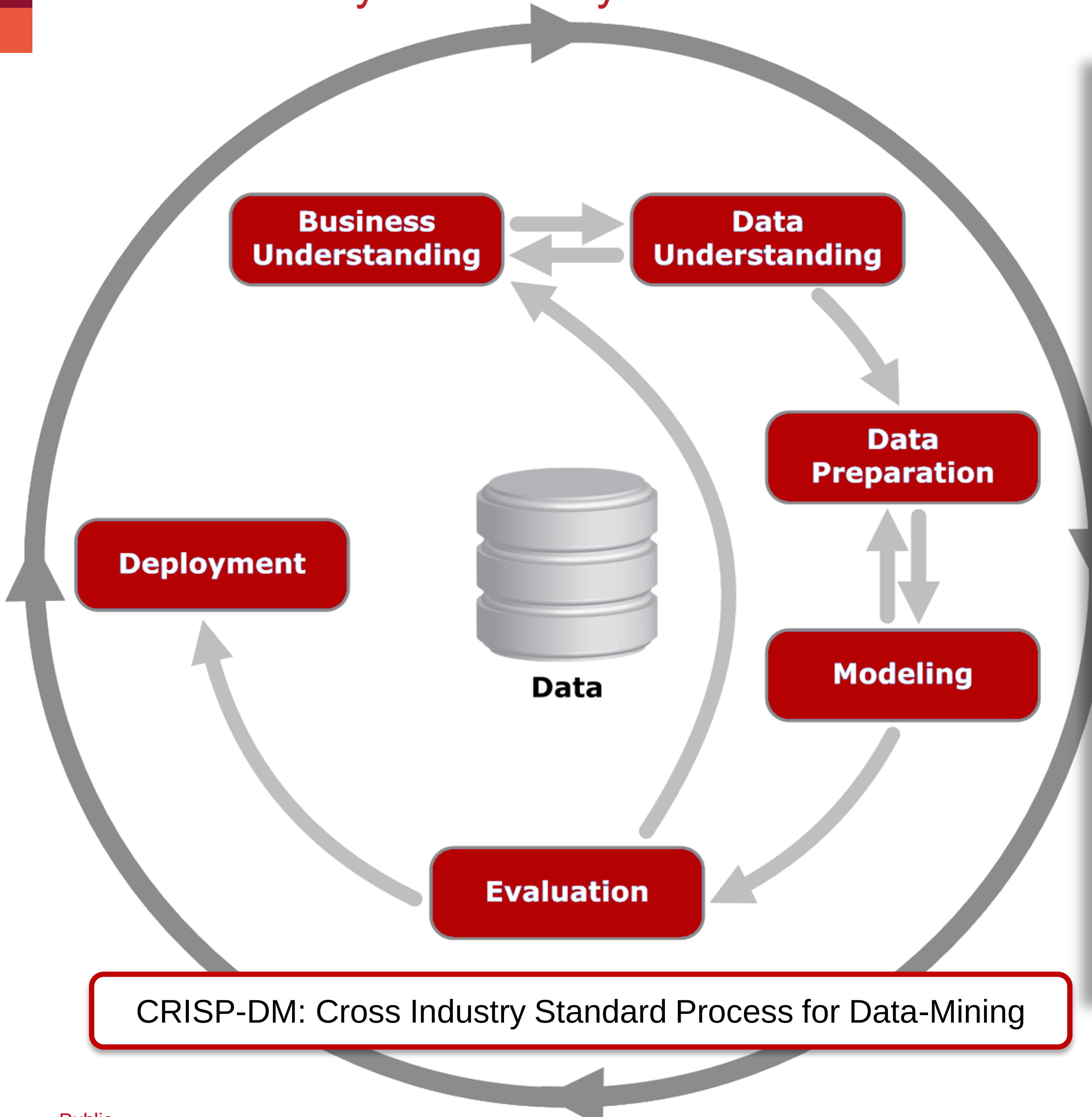
- *Business understanding*:
  - Focus on the business problem in terms of objectives and requirements
  - Translate the business problem into a data-mining problem
- *Data understanding:*
  - Data collection
  - Data exploration: variables, data quality, get first insight into data
- *Data preparation:*
  - Data cleaning, selection, transformation …
  - Definition of the final data-set to be used in the analysis
- *Modelling:*
  - Select and optimize models that can better describe the problem
- *Evaluation:*
  - Evaluate the performance of the analysis process in terms of business requirements
  - If the quality of the results matches the business expectation: deploy

How data analytics actually look like…



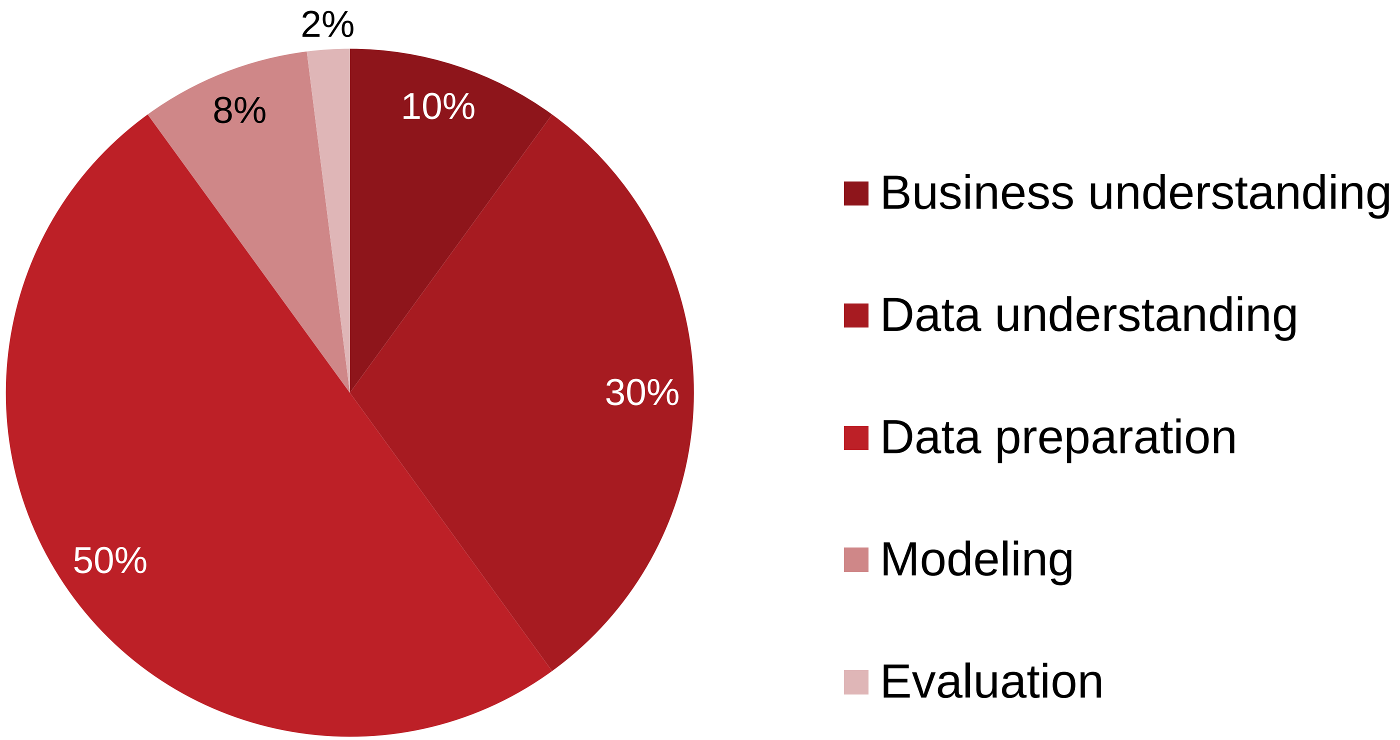
- *Business understanding*:
  - Focus on the business problem in terms of objectives and requirements
  - Translate the business problem into a data-mining problem
- *Data understanding:*
  - Data collection
  - Data exploration: variables, data quality, get first insight into data
- *Data preparation:*
  - Data cleaning, selection, transformation …
  - Definition of the final data-set to be used in the analysis
- *Modelling:*
  - Select and optimize models that can better describe the problem
- *Evaluation:*
  - Evaluate the performance of the analysis process in terms of business requirements
  - If the quality of the results matches the business expectation: deploy
  - Else: identify which are the business needs that are not yet satisfied and upgrade the analysis

How data analytics actually look like…

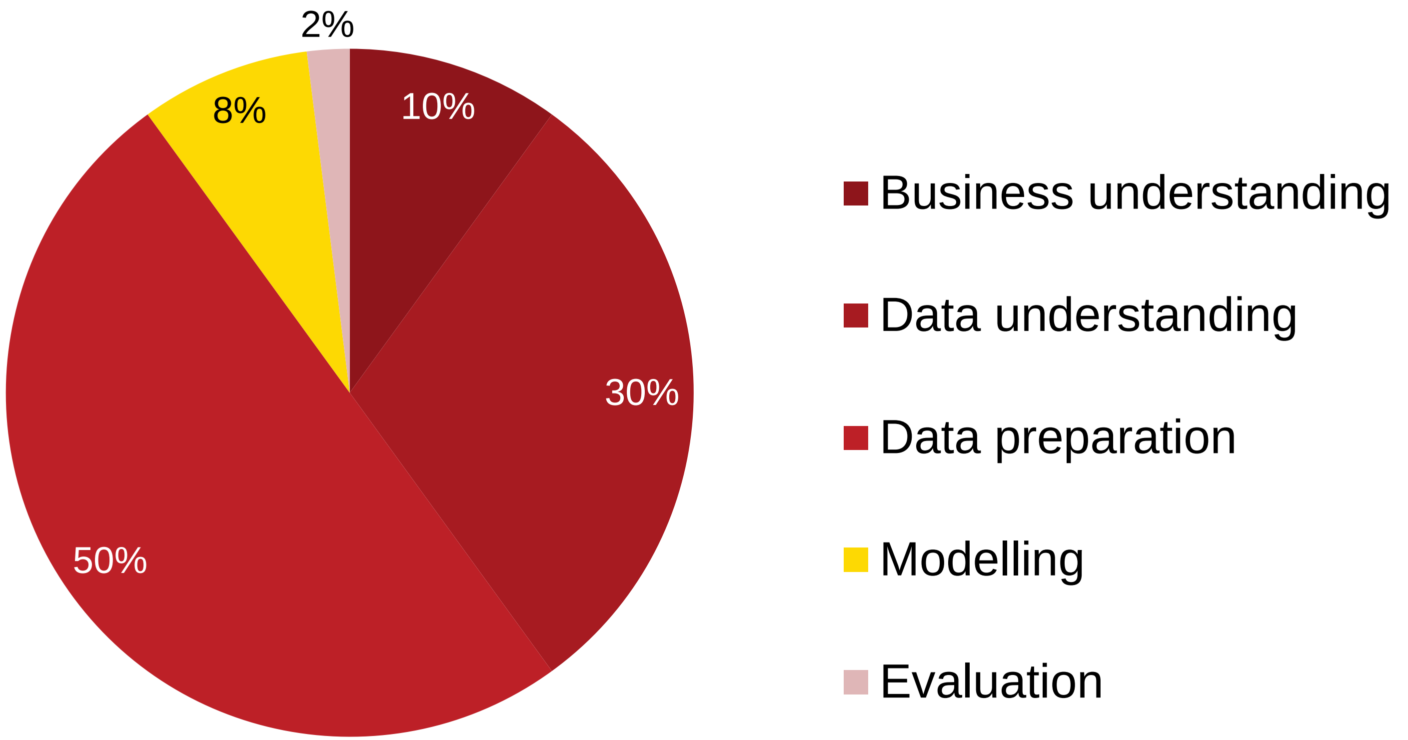


- *Business understanding*:
    - Focus on the business problem in terms of objectives and requirements
    - Translate the business problem into a data-mining problem
- *Data understanding:*
    - Data collection
    - Data exploration: variables, data quality, get first insight into data
- *Data preparation:*
    - Data cleaning, selection, transformation …
    - Definition of the final data-set to be used in the analysis
- *Modelling:*
    - Select and optimize models that can better describe the problem
- *Evaluation:*
    - Evaluate the performance of the analysis process in terms of business requirements
    - If the quality of the results matches the business expectation: deploy
    - Else: identify which are the business needs that are not yet satisfied and upgrade the analysis

How data analytics actually look like…



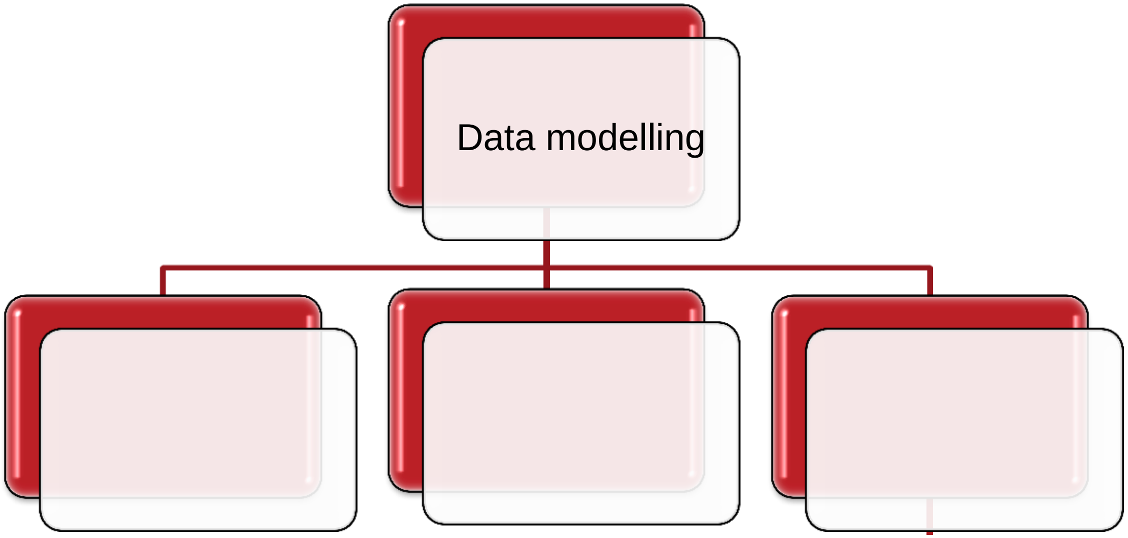
How data analytics actually look like…



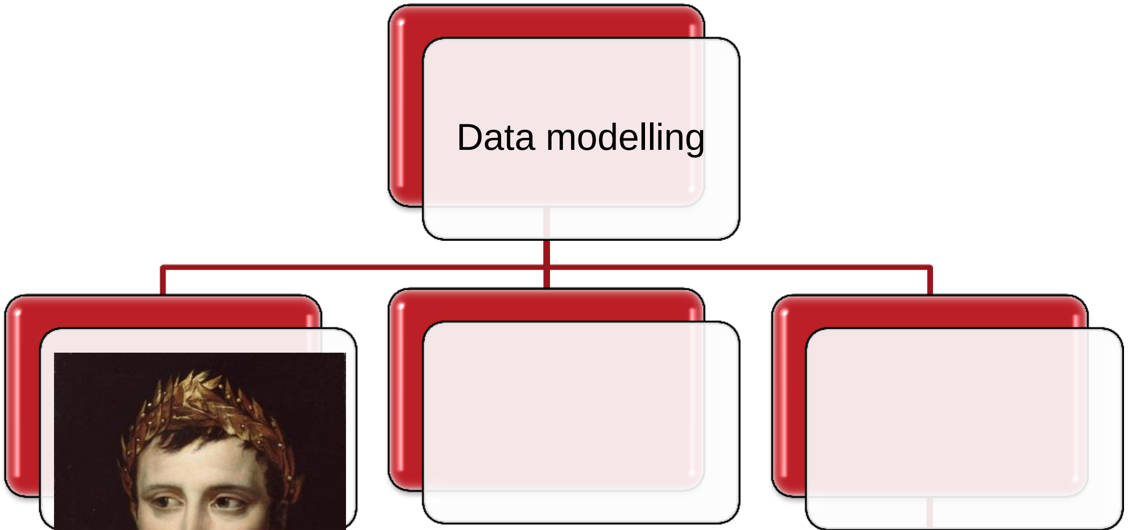
The theory modules of the course will be focused on models, while during exercises you will face the complete data-analytics process

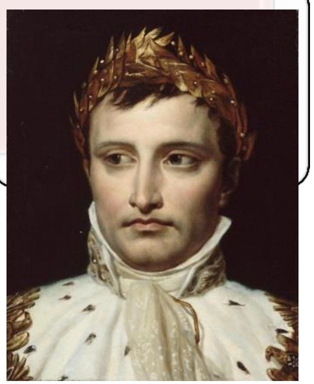
Day 1: Data modelling

Data modelling: the portrait of an analytical problem



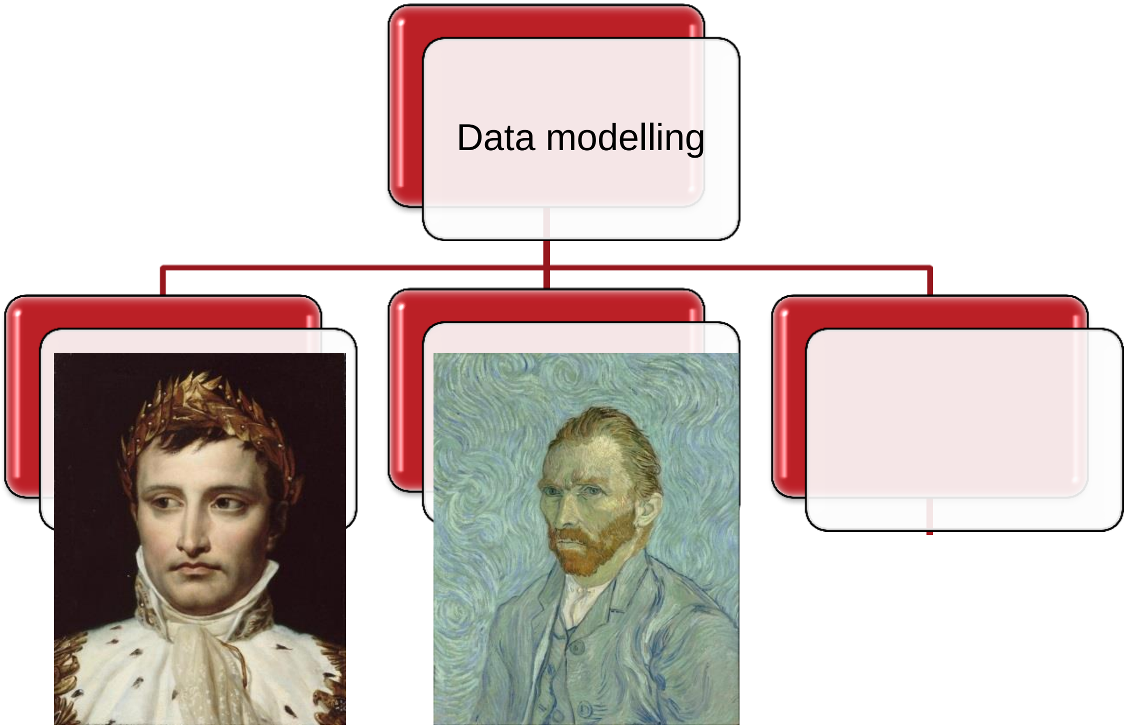
Data modelling: the portrait of an analytical problem





Jacques-Louis David:
Napoléon crowned
(1807)

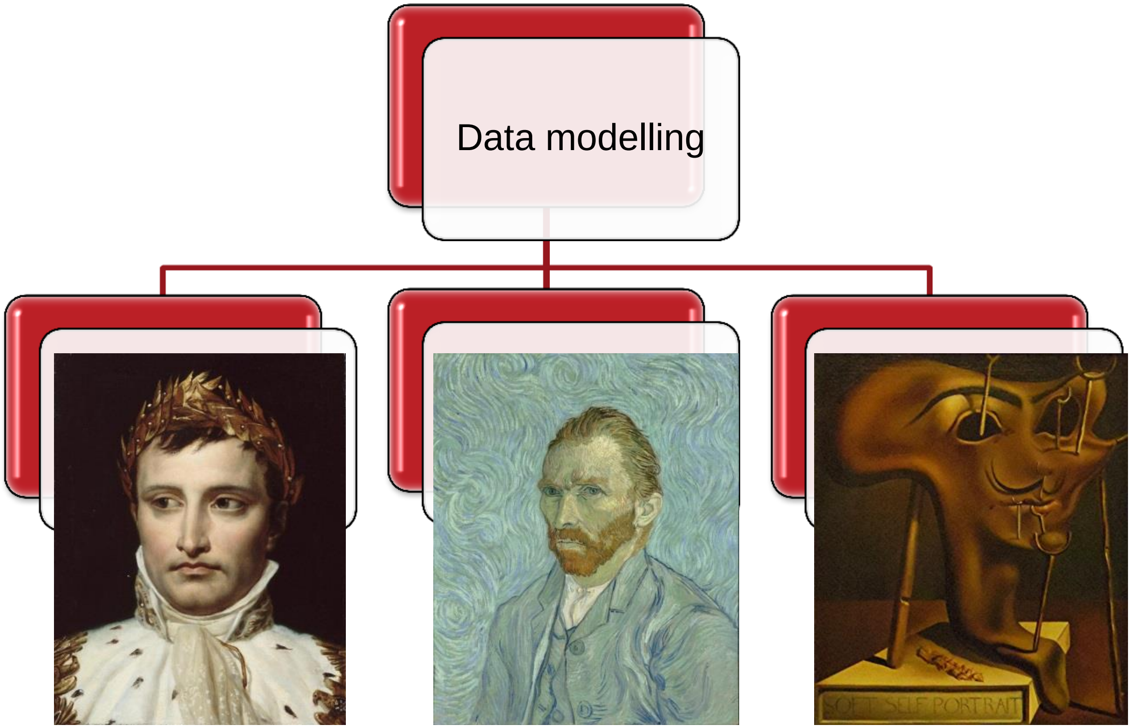
Data modelling: the portrait of an analytical problem



Jacques-Louis David: *Napoléon crowned* (1807)

Vincent Van Gogh: *Self-portrait (1889*)

Data modelling: the portrait of an analytical problem



Jacques-Louis David: *Napoléon crowned* (1807)

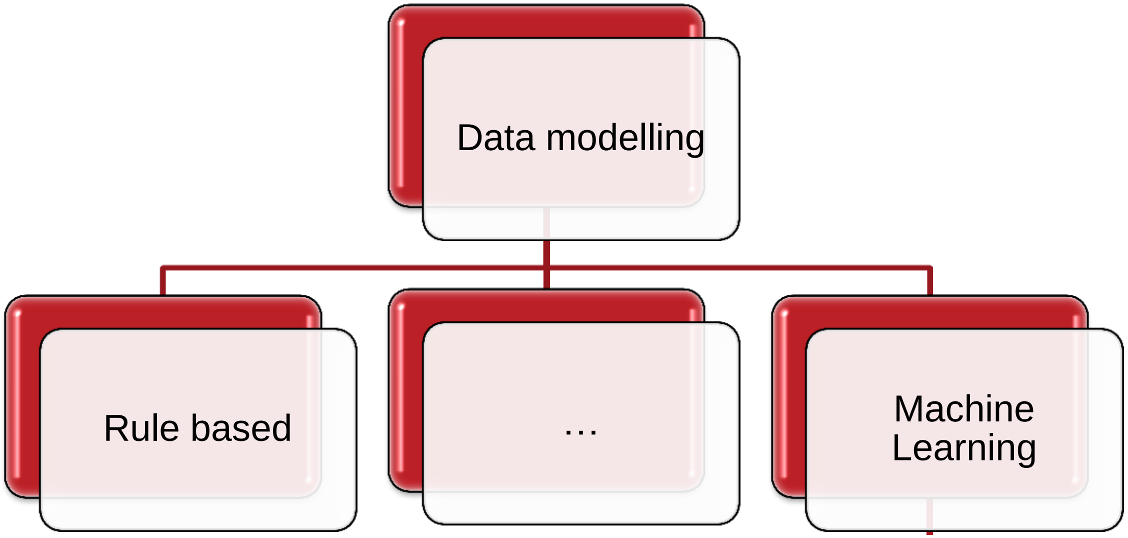
Vincent Van Gogh: *Self-portrait (1889*)

Salvador Dali: *Soft self-portrait with fried bacon (1941*)

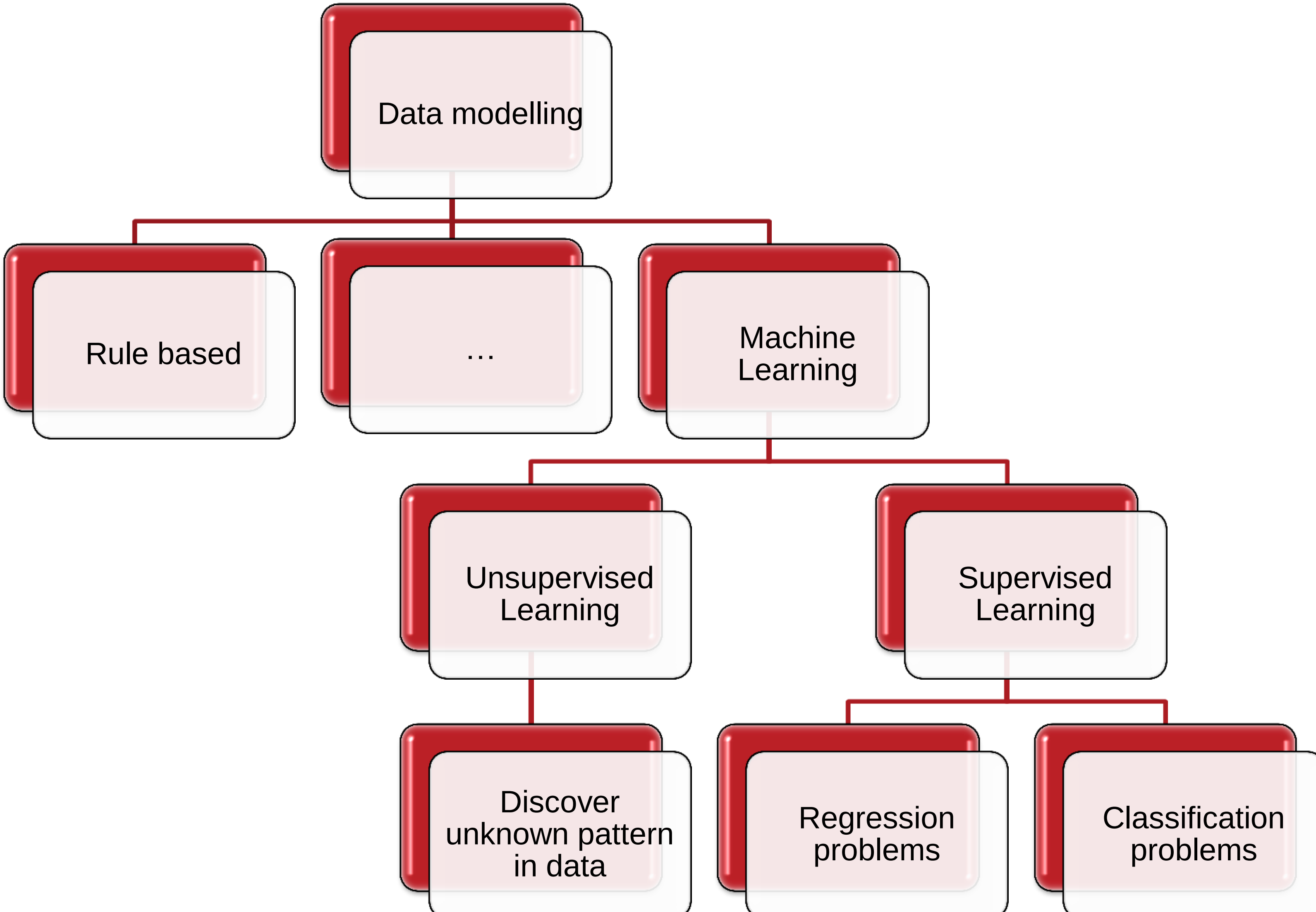
$$Data - Scientist : Model = Painter : \ (Painter's\ brain + feelings\ while\ painting\ a\ portrait)$$

* https://deepart.io/

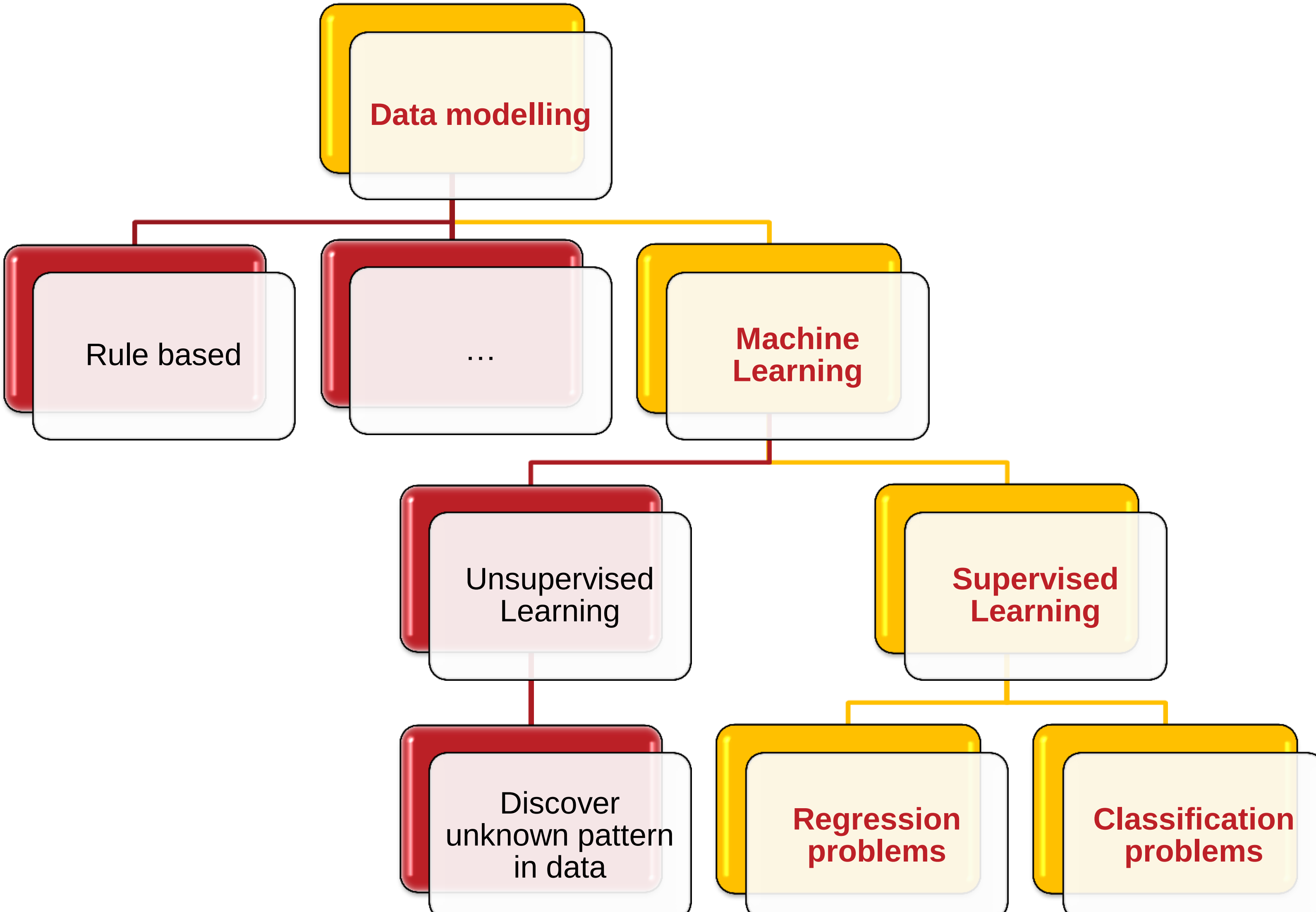
Data modelling: the portrait of an analytical problem



Data modelling: the portrait of an analytical problem



Data modelling: the portrait of an analytical problem



Day 1: Regression models

Regression model: when and why

Main goal: understand whether and how a given phenomenon y (dependent variable) is correlated to a set of independent observations $\vec{x}$

$$y = f\left(\alpha_i, x_j\right), \qquad \begin{cases} i = 1, n \\ j = 1, m \end{cases}$$

Prediction & forecasting

Understanding and modelling the functional relation between observations and a phenomenon means to be able to predict the behaviour of the phenomenon in response to a new set of measurements

Interpretation

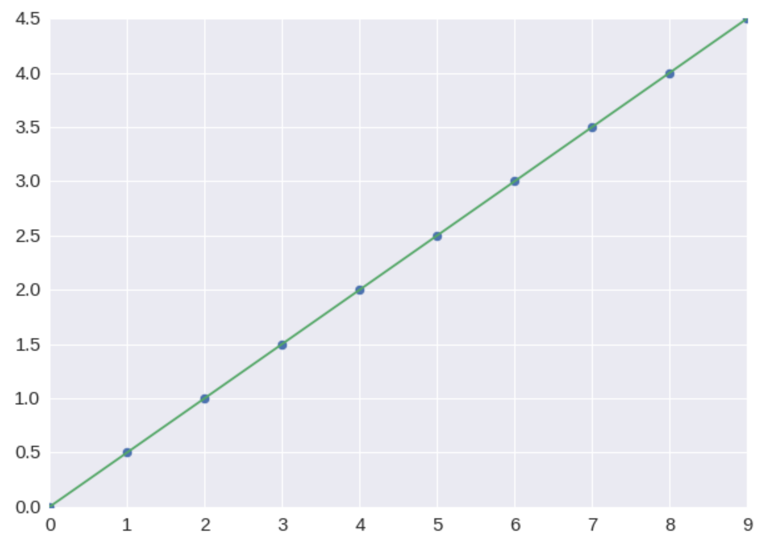
Depending on the functional relation between the phenomenon and the observations it is possible to understand the relevance of each variable in the determination of the phenomenon

Starting from basics: the linear regression

Case 1 (boring): The phenomenon is fully determined by a finite set of independent variables

$$y = f(\alpha_0, \alpha_i, x_i) = \alpha_0 + \sum_i \alpha_i x_i$$

The goal of the training of the model in this case is to analytically solve the problem to find the hyper-plane which pass through all the training points



Questions:

- How many parameters has a linear model in *N* dimensions?
- How many points do I need to fit a linear model in *N* dimensions under the hypothesis that the phenomenon is fully determined by the set on *N* independent variables

Starting from basics: the linear regression

<u>Case 2 (optimistic real life):</u> The phenomenon y is mostly influence by a finite set of independent variables x_i but it depends also on a set of other unknown variables x'_j

$$y = f(\alpha_0, \alpha_i, x_i, \varepsilon) = \left[\alpha_0 + \sum_i \alpha_i x_i\right] + \varepsilon(x'_j)$$

- A geometrical solution of the problem is not anymore possible
- Given the presence of unknown variables that can influence the phenomenon the model that we can build will be ***NOT*** a “true” model
- All what we can do is to find a model that represent the best approximation of the phenomenon
- Assuming that the effect of unknown variables on the phenomenon:
  - is not systematically shifting the phenomenon towards a specific direction
  - has a mean equal to zero
  - has a constant variance independent on the x_i

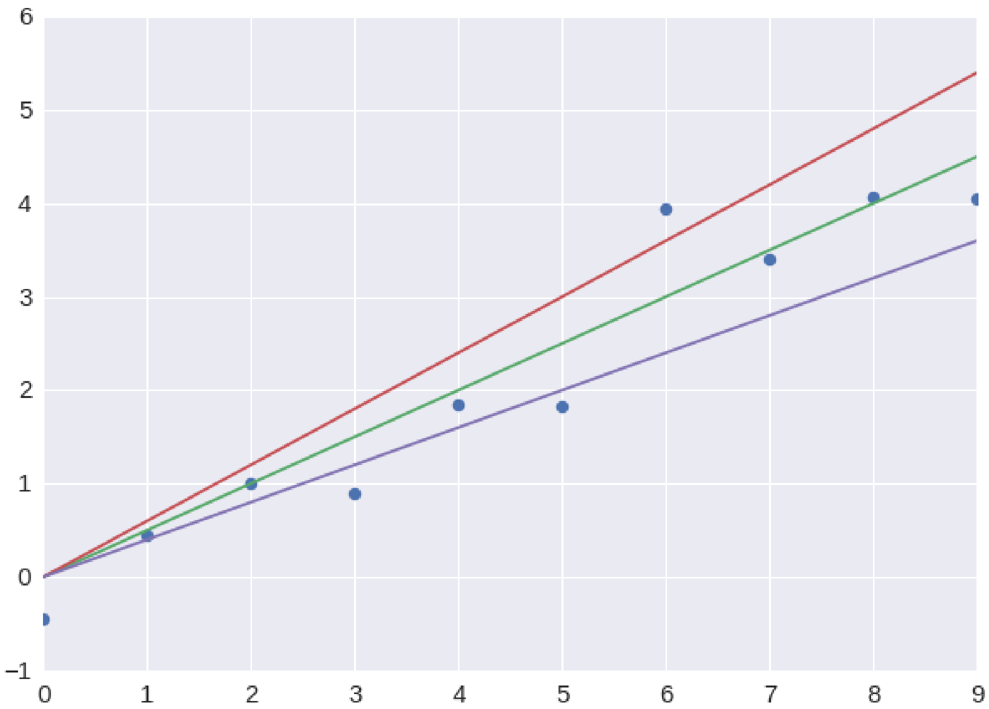
  → the linear model $f(\alpha_0, \alpha_i, x_i)$ will provide the expectation value of the phenomenon y

$$\mu = \langle y \rangle = \langle f(\alpha_0, \alpha_i, x_i, \varepsilon) \rangle = f(\alpha_0, \alpha_i, x_i) + \langle \varepsilon \rangle = f(\alpha_0, \alpha_i, x_i)$$

  → It is possible to interpret the behaviour of the phenomenon in terms of the impact of each one of the known variables according to the value of the corresponding coefficient α_i

How to measure the “level of approximation”

$$\mu = f(\alpha_0, \alpha_i, x_i) = \left[\alpha_0 + \sum\nolimits_i \alpha_i x_i\right]$$



The best model can be chosen as the one that minimise the *sum of squared (SS) residuals*

$$\hat{\alpha} = \min_{\vec{\alpha}} \sum_{i=1}^{N} [y_i - f(\vec{\alpha}, x_i)]^2$$

The process of defining the best parameter set for a given model on a given data-set is called ***training*** of the model

- How good is this model?
- Can the model explain the behaviour of y?

How to measure the “level of approximation”

- We can model the “behaviour” of the y in terms of its variation in the population against its mean value across the whole population $\bar{y}$

$$Total\ SS = SS_{Tot} = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

- The same “behaviour” can be measured also for the prediction of the model $\hat{y} = f(\hat{\alpha}, x)$

$$Regression\ SS = SS_{Reg} = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$

- In linear models the difference between SS_{Tot} the and the SS_{Reg} is fully given by the *residual SS*

$$Residual\ SS = SS_{Res} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

- A good model is the one that is able to reproduce the behaviour of the phenomenon and how well this behaviour is reproduced can be measured with the ratio

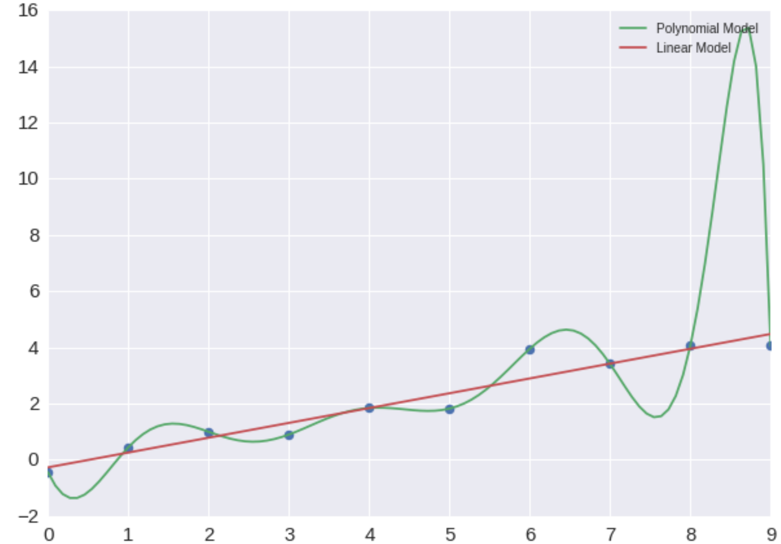
$$R^2 = \frac{SS_{Reg}}{SS_{Tot}} = 1 - \frac{SS_{Res}}{SS_{Tot}}$$

Best models and good models…

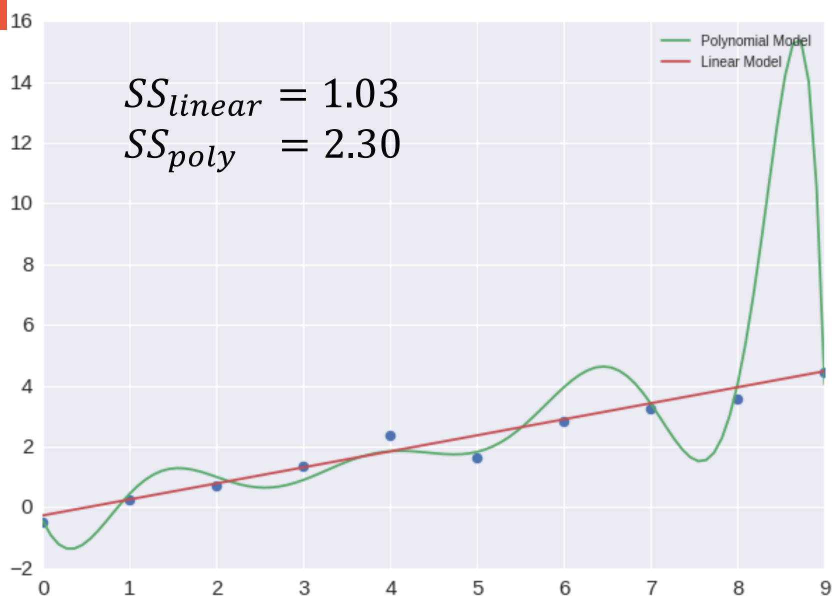
- Which is the best model?
  - $SS_{linear} = 1.89$
  - $SS_{poly} = 1.12 \times 10^{-11}$

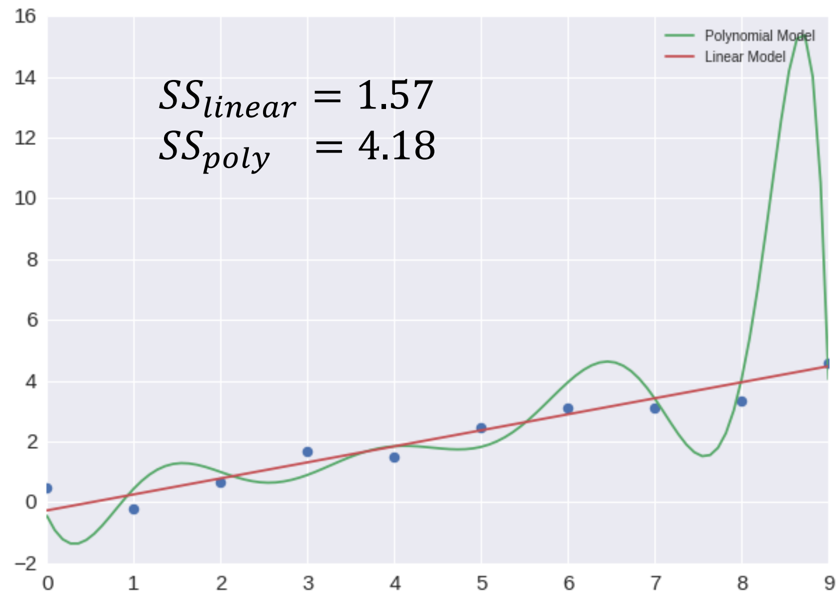
Best models and good models…

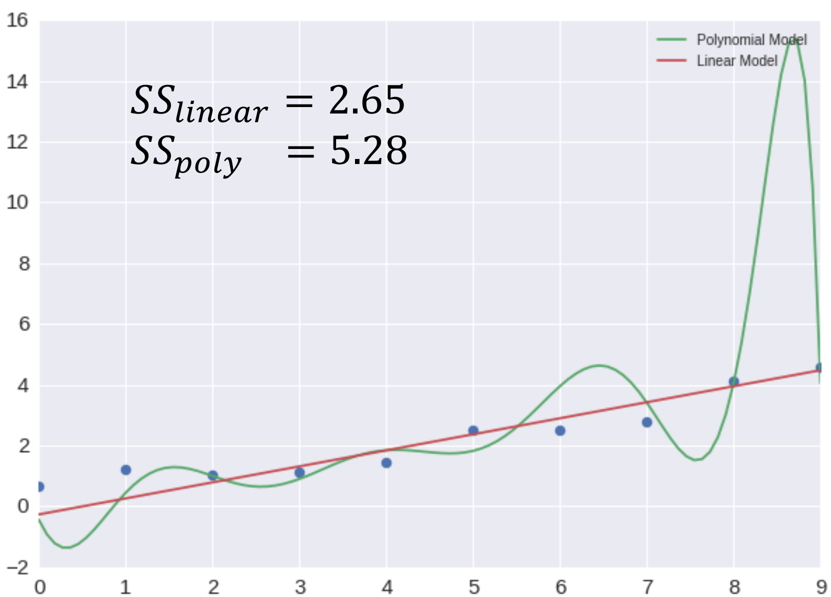
- Which is the best model?
  - $SS_{linear} = 1.89$
  - $SS_{poly} = 1.12 \times 10^{-11}$
- What if we compare the prediction of these 2 models with another set of observation of the same phenomenon?
  - $SS_{linear} = 1.03$
  - $SS_{poly} = 2.30$
- To think about:
  - The performance of the linear model didn't change much
  - The performance of the polynomial changed by 11 order of magnitude!
- What is happened?

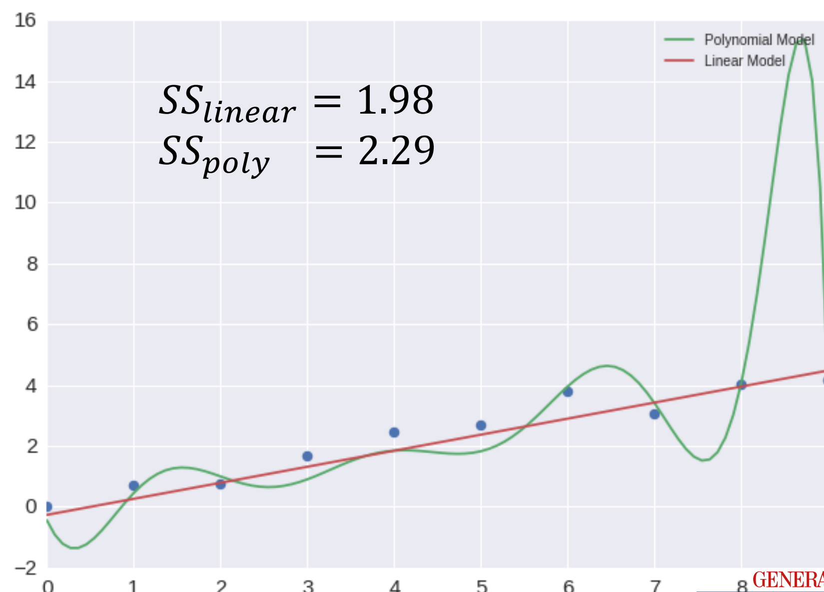


Matter of bad luck?









Is with pleasure that I introduce you: the OVERFITTING

- Overfitting occurs when a model learn “too well” how to reproduce the training set
- The reason of overfitting is that the model tries to describe the stochastic component of the phenomenon as a function of the known observables
- The result is that the model try to reproduce the noise present in the training set as a deterministic component of the signal
- A over-fitted model looses any prediction power

Overfitting: diagnostics and analysis

- Generate two datasets following the function:

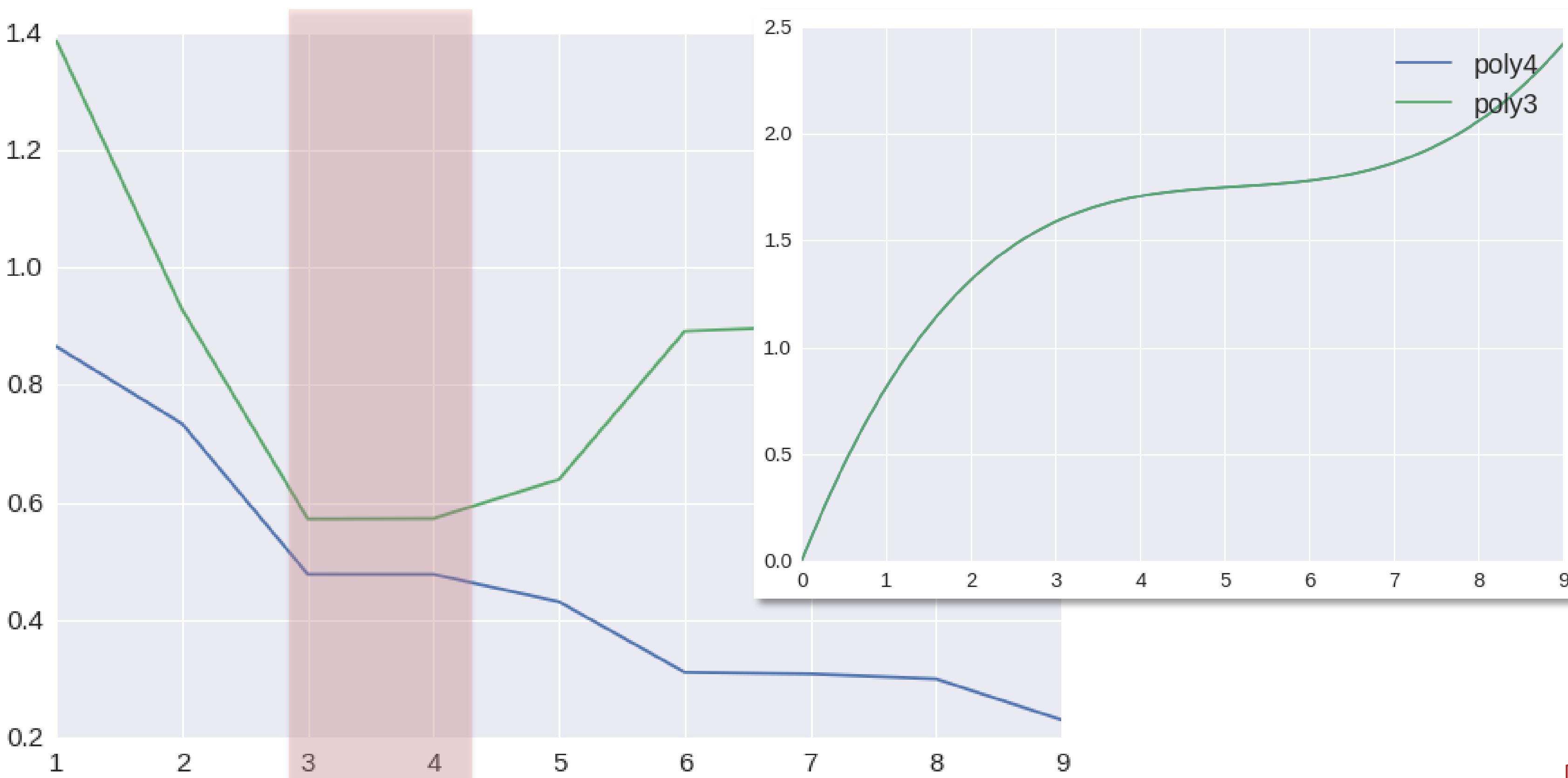
$$y = x - 0.2 \times x^2 + 0.015 \times x^3 - 0.0002 \times x^4 + Gauss(0,0.2)$$

Overfitting: diagnostics and analysis

- Fit polynomial functions with degree from 1 to 9
- Compute the sum of squares for each fitted polynomial
  1. for the test set
  2. for the training set

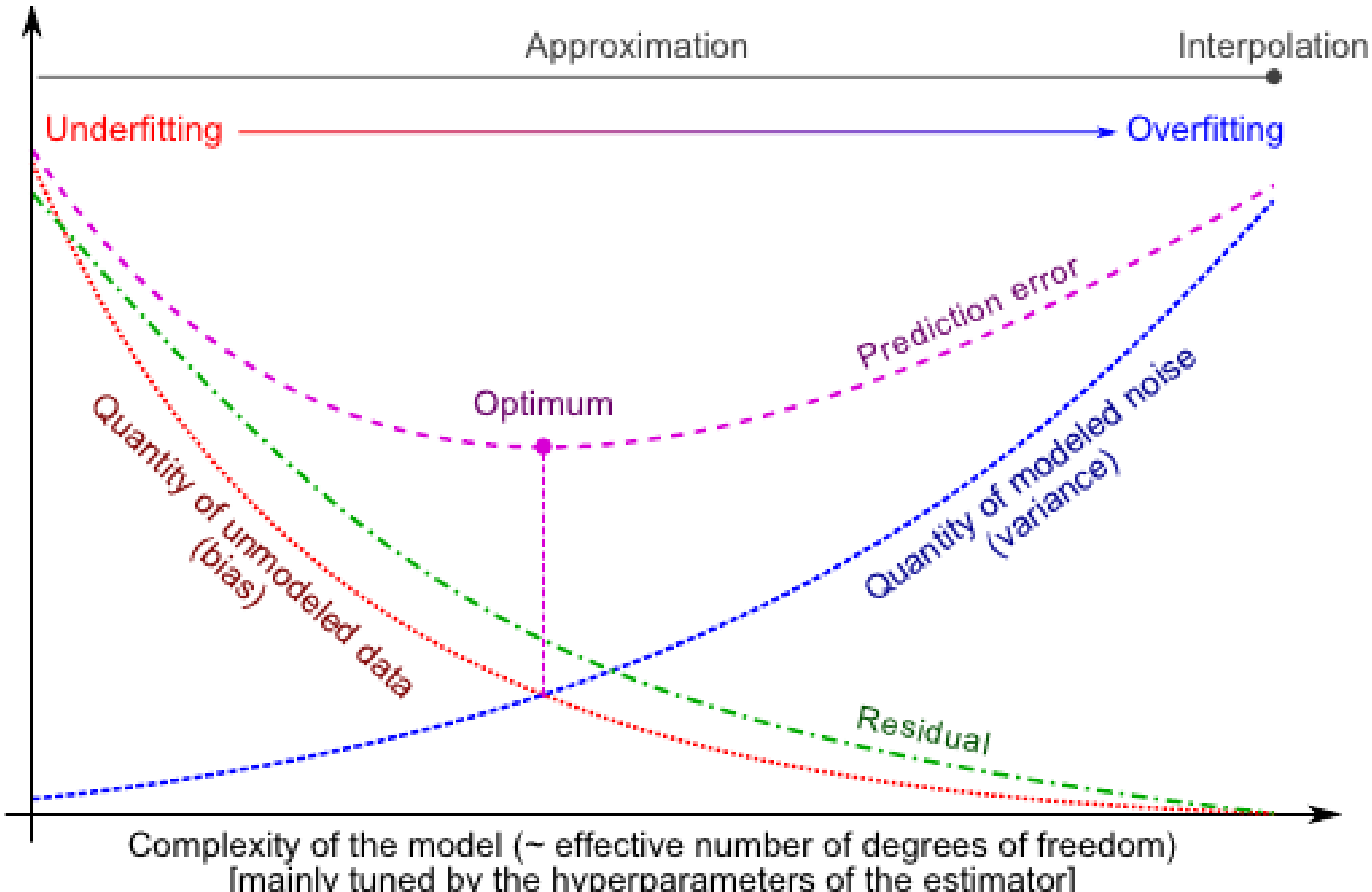
Overfitting: diagnostics and analysis

- Fit polynomial functions with degree from 1 to 9
- Compute the sum of squares for each fitted polynomial
    1. for the test set
    2. for the training set



Pragmatic approach to the model training: metrics

- During training the metric to look at is the out-of-sample error
- $Out - of - sample = Bias + Variance$

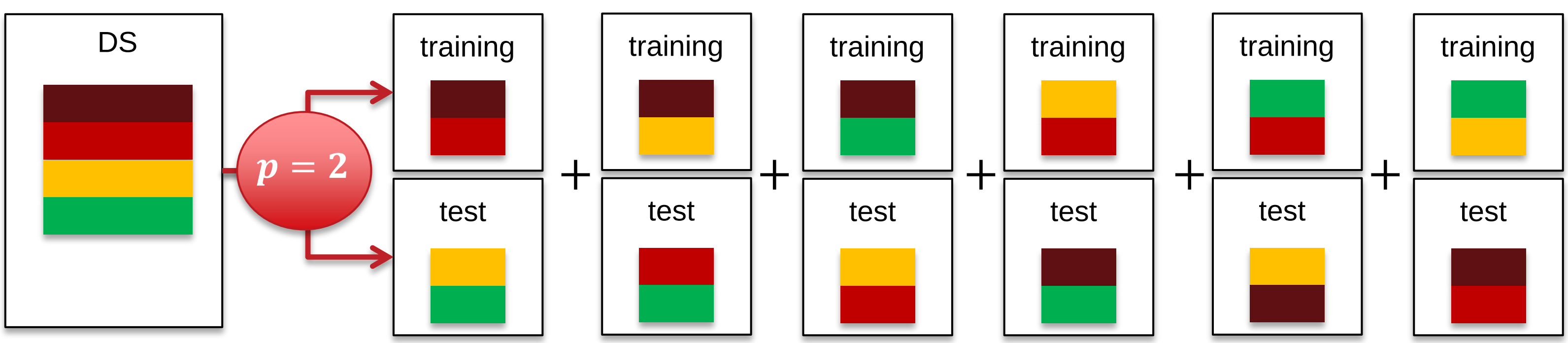


Pragmatic approach to the model training: cross-validation

Often the DS are not large enough to allow a partitioning without without losing significant modelling or testing capability

Strategy 1: ***Exhaustive cross-validation***

- *<u>Leave-p-out cross-validation</u>*:
  - Given a dataset containing n events
  - Use p events out of the n for the validation and $n-p$ for the training
  - Repeat the training and test for all possible combination of the p-events
    - How many combination can we create if n=100 and p=30?
  - Average the results of each train-test to obtain the overall result

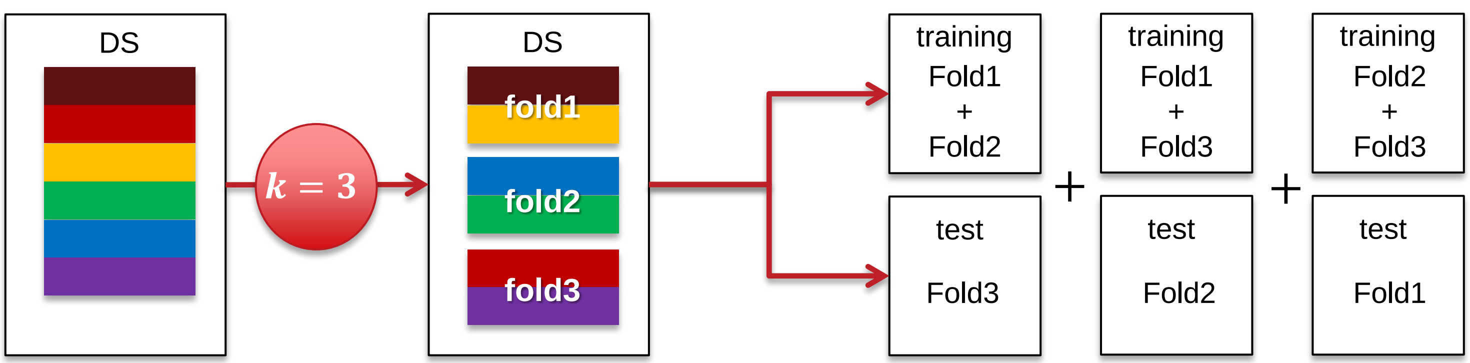


Pragmatic approach to the model training: cross-validation

Often the DS are not large enough to allow a partitioning without without losing significant modelling or testing capability

Strategy 2: ***Non-exhaustive cross-validation***

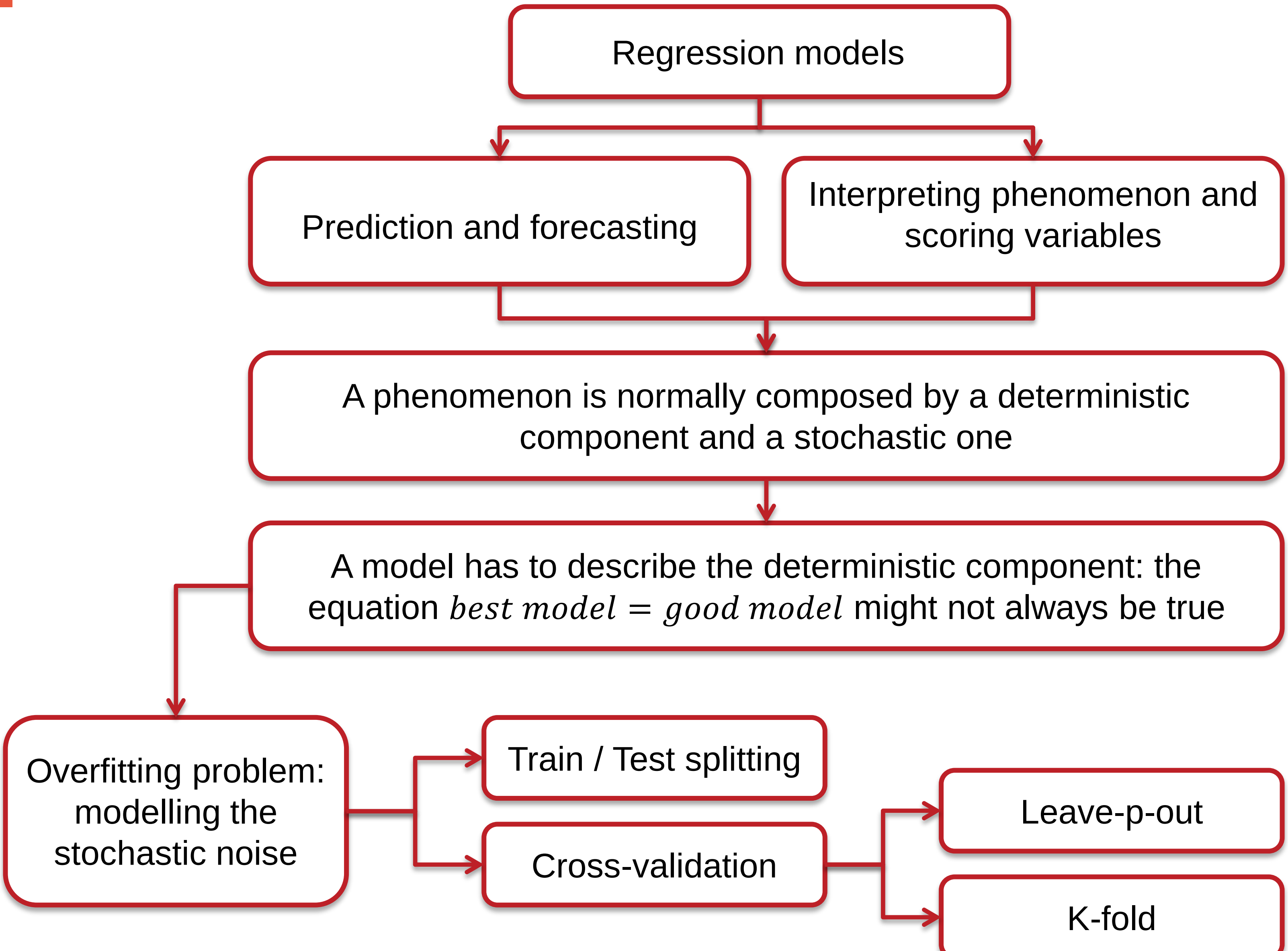
- *k-fold cross-validation*:
  - Split the DS into k random subsample
  - Use $k-1$ samples for the training and the remaining 1 for the test
  - Repeat the process k times and average the results
  - If $k=n$ we come back to the leave-p-out strategy with $p=1$
  - A standard value of k is 10



Day 1: concepts map

- Data Analytics Process
  - It is a complex process build in several recursive steps (CRISP-DM)
  - The solution *<u>MUST</u>* start from the understanding of the business/scientific issue that has to be solved
    - Understanding the business and understanding the data are the most relevant parts of the analytical process
    - The modelling phase is pure art: there are rules to be followed but there are not deterministic prescriptions

Day 1: concepts map



Day 1: scikit-Learn

scikit-learn

- Reference:
  - Link: http://scikit-learn.org/stable/
  - Notebook: notebooks/Lectures/scikit-learn_1