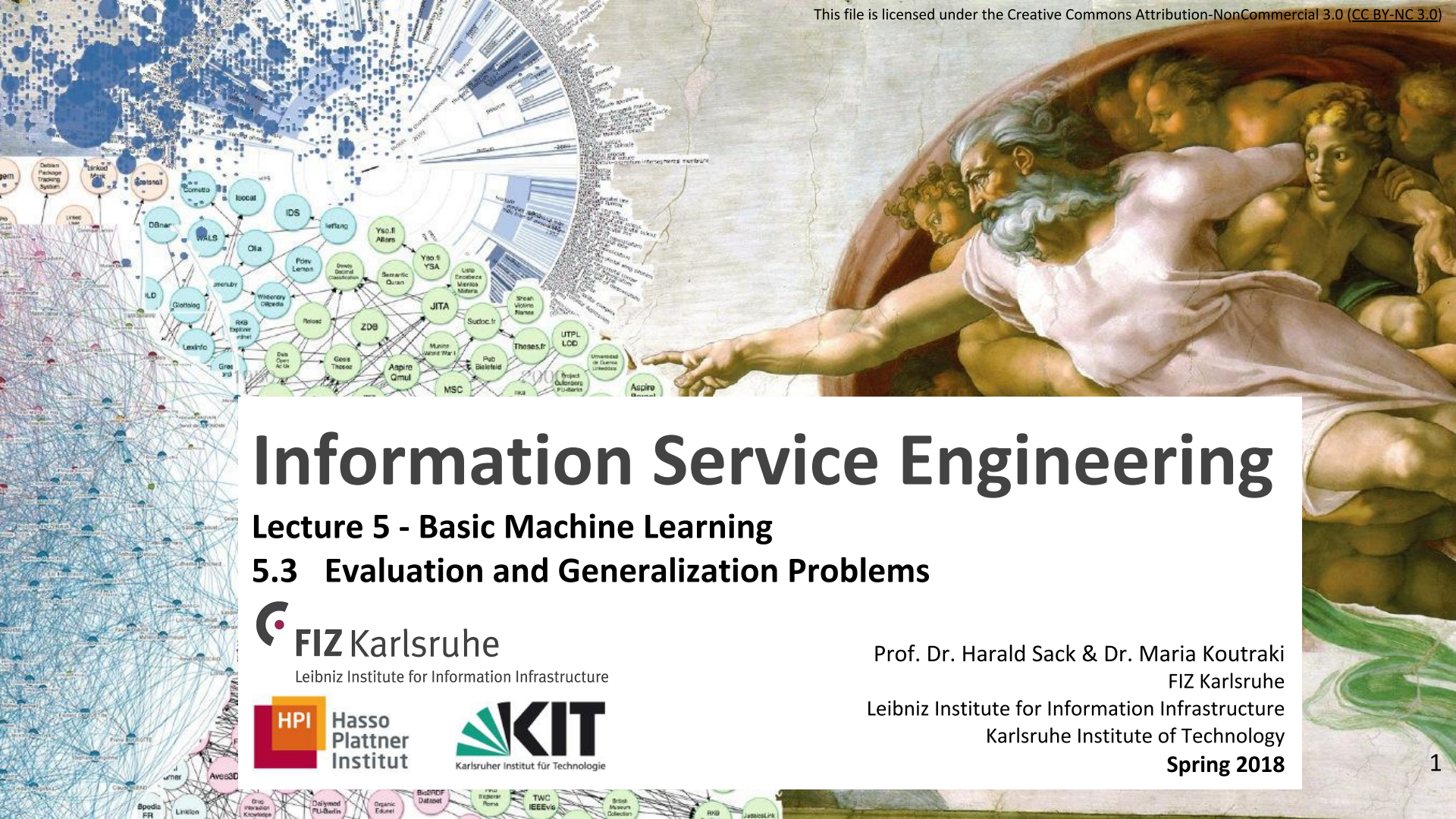




Information Service Engineering

Lecture 5 - Basic Machine Learning

5.3 Evaluation and Generalization Problems

 **FIZ Karlsruhe**

Leibniz Institute for Information Infrastructure

 **Hasso Plattner Institut**

 **KIT**
Karlsruher Institut für Technologie

Prof. Dr. Harald Sack & Dr. Maria Koutraki
FIZ Karlsruhe

Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology
Spring 2018

Information Service Engineering

Lecture 5: Basic Machine Learning

5.1 A Brief History of AI

5.2 Introduction to Machine Learning

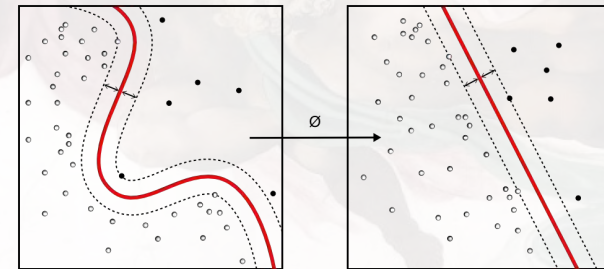
5.3 Evaluation and Generalization Problems

5.4 Basic ML Algorithms

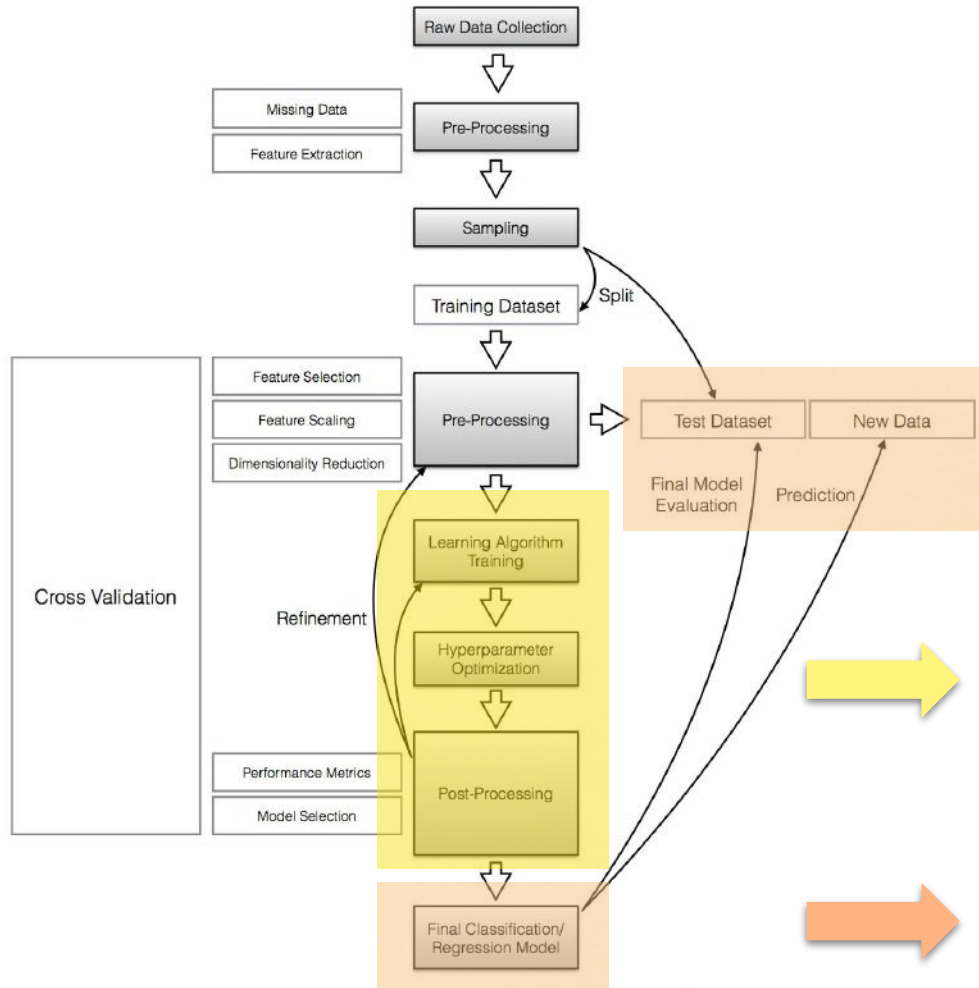
5.5 Basic ML Algorithms II

5.6 Unsupervised Learning

5.7 Deep Learning



Supervised Learning



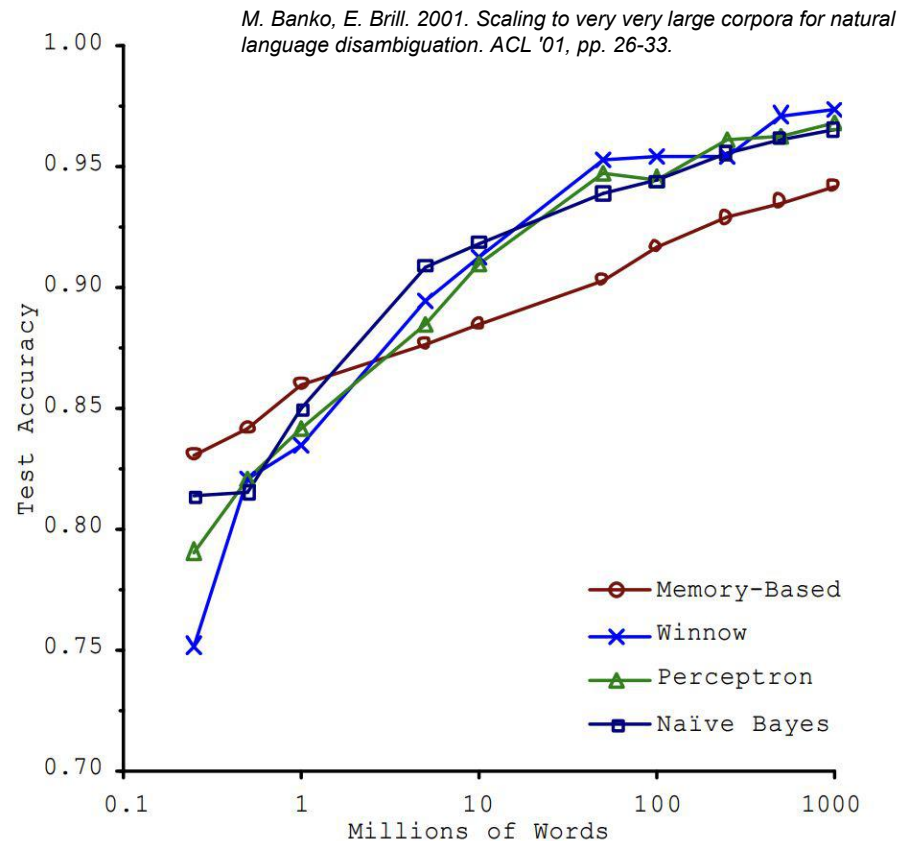
Training/Optimisation Phase

Prediction Phase

Main Challenges of Machine Learning

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

- **Insufficient Quantity of Training Data**



Main Challenges of Machine Learning

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

- Insufficient Quantity of Training Data
- Nonrepresentative Training Data

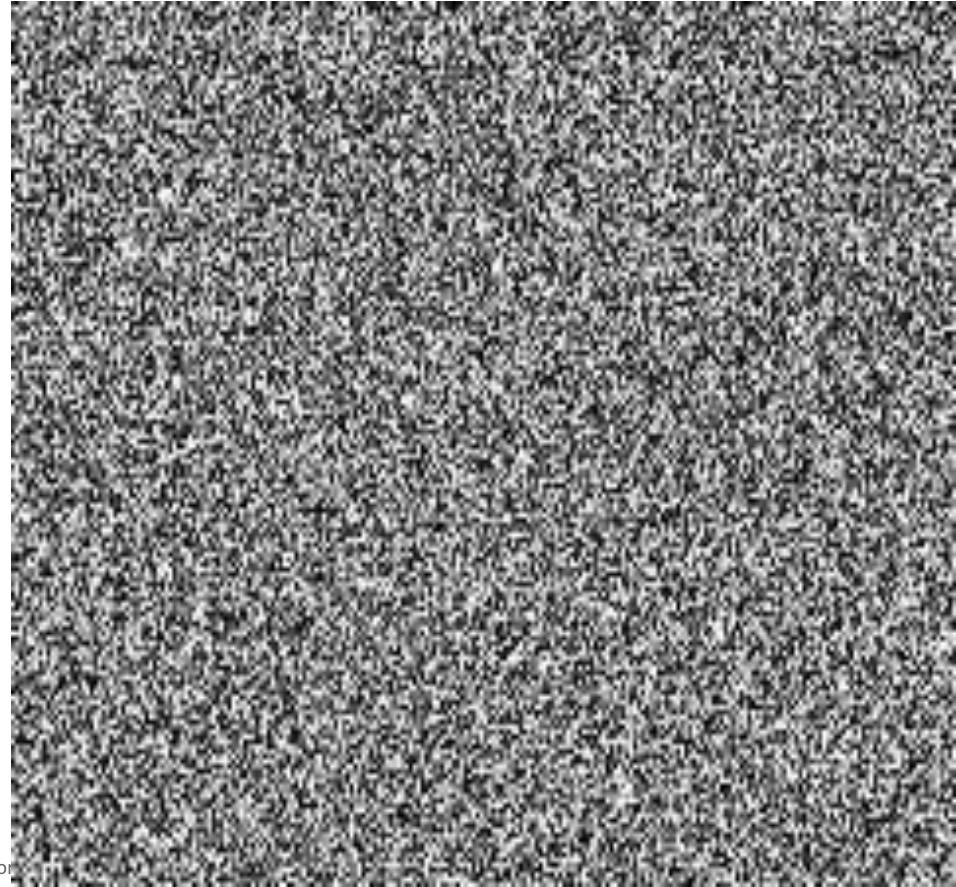


“**Harry S. Truman** has the last laugh as he holds up a copy of the **Chicago Tribune** at Union Station in St. Louis, Missouri, on **November 3, 1948**, after winning the election the previous day. He was so widely expected to lose that the Tribune ran this incorrect headline.”

Main Challenges of Machine Learning

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

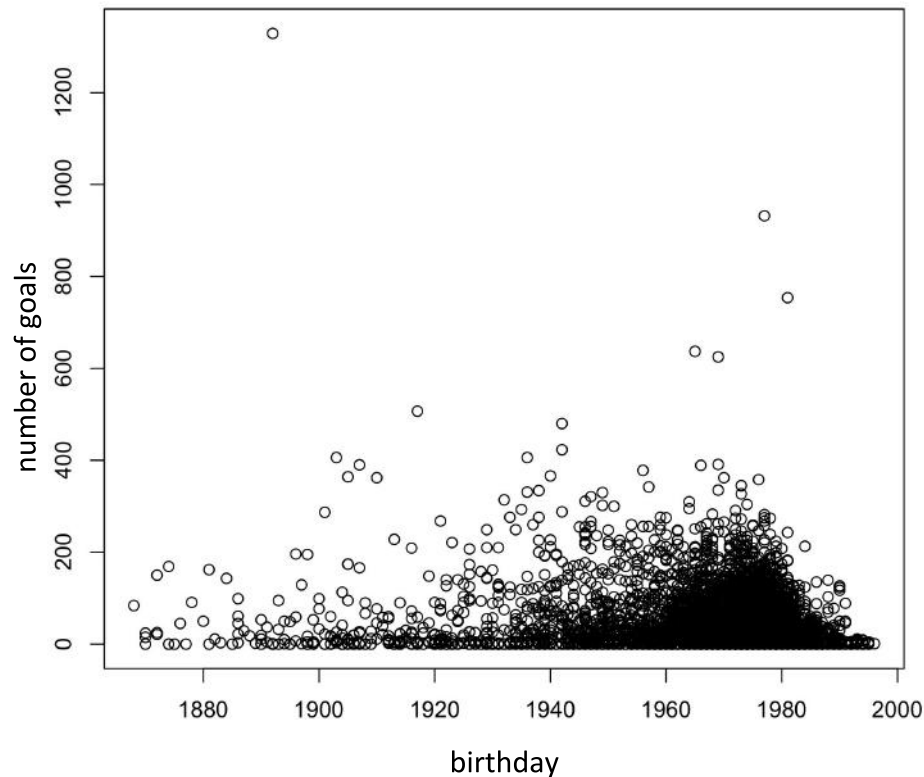
- **Insufficient Quantity of Training Data**
- **Nonrepresentative Training Data**
- **Poor-Quality Data**



Main Challenges of Machine Learning

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

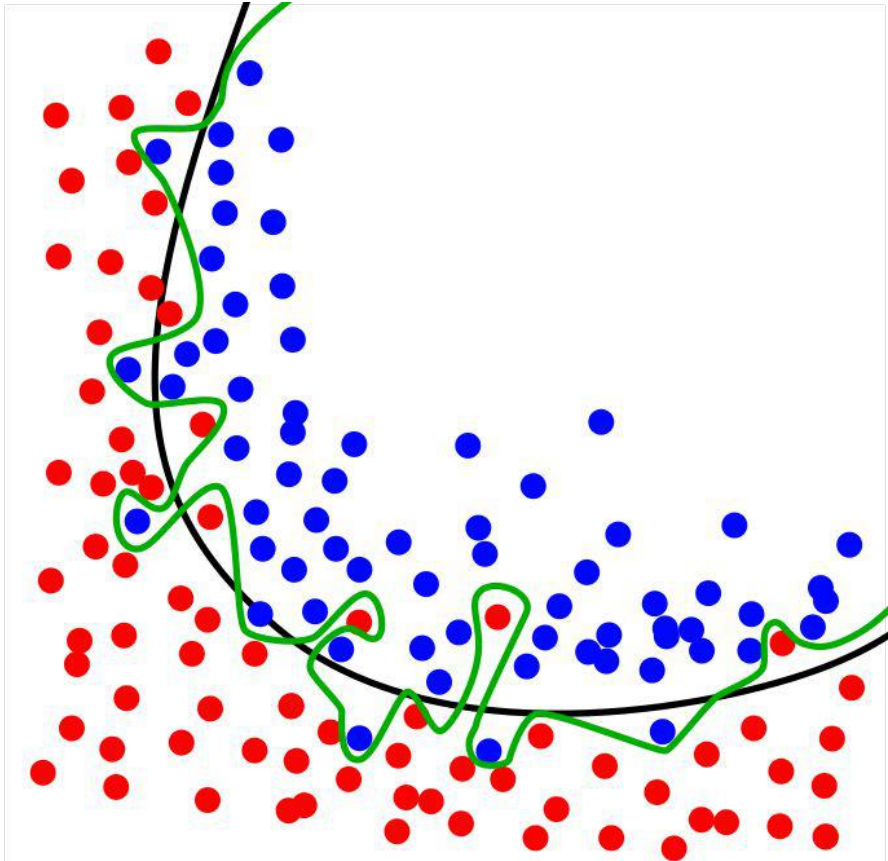
- Insufficient Quantity of Training Data
- Nonrepresentative Training Data
- Poor-Quality Data
- Irrelevant Features



Main Challenges of Machine Learning

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

- Insufficient Quantity of Training Data
- Nonrepresentative Training Data
- Poor-Quality Data
- Irrelevant Features
- **Overfitting the Training Data**

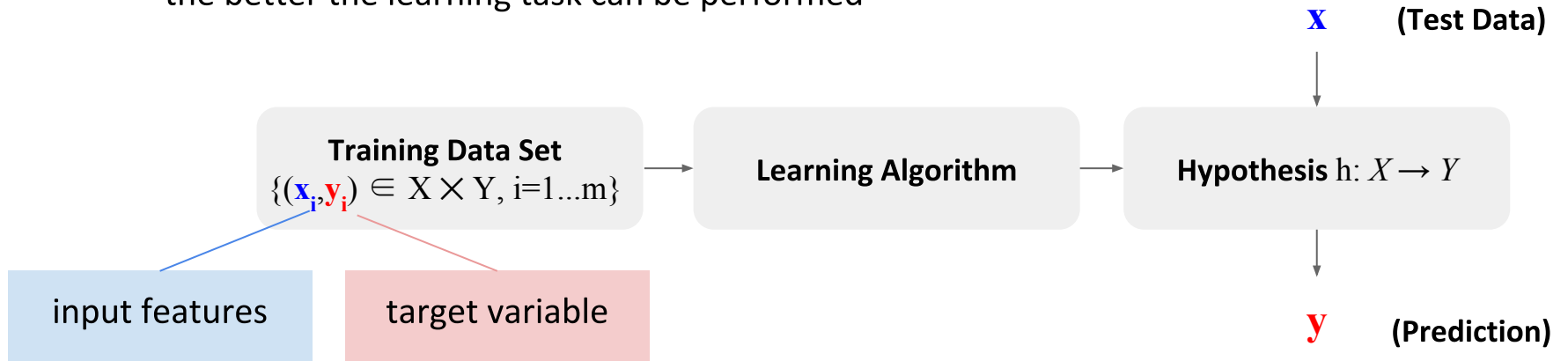


Data Collection

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

- **Raw Data Collection**

- The **larger** and the **more diverse** the collected data, the better the learning task can be performed



Data Preprocessing

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

- **Data Preprocessing**
 - Typically to **create suitable training data**, the collected raw data has to be preprocessed (cleaned up) to remove errors, as e.g., dummy values, absence of data, contradicting data, etc.
- **Data Cleaning Steps**
 - **Parsing**: locates and identifies individual data elements in raw data
 - **Correcting**: corrects parsed individual data components using sophisticated data algorithms
 - **Normalization**: applies conversion routines to transform data into standard formats
 - **Matching**: searching and matching records within and across data based on predefined rules
 - **Consolidating**: merges data into one representation

Training Data and Test Data

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

- The easiest way to obtain training data is to split up the original dataset
 - Training Data (used to train the algorithm)
 - Test Data (used to evaluate the performance of the readily trained algorithm)

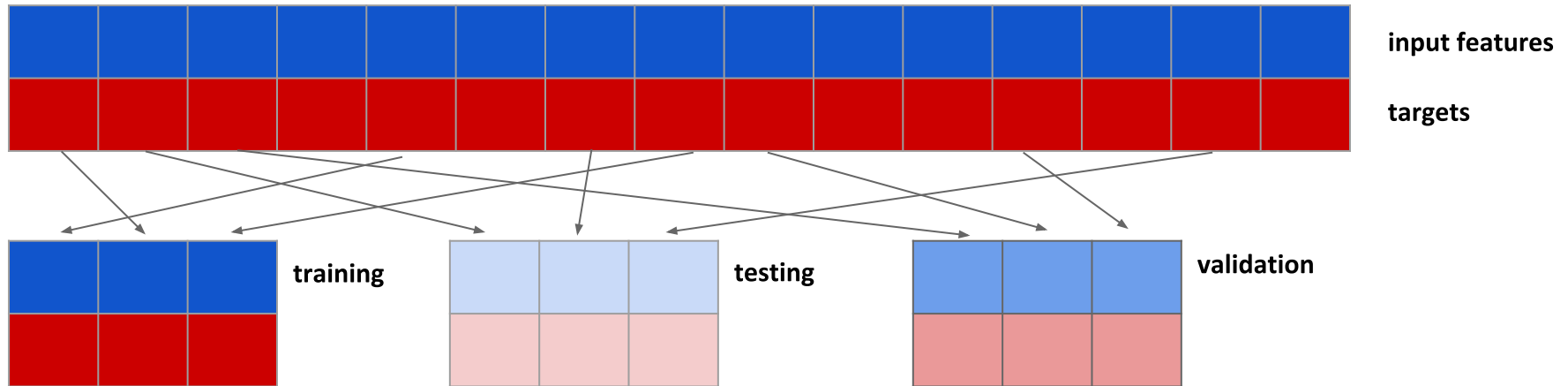


- BUT:
 - Evaluations obtained tend to reflect the particular way the data are divided up.
- SOLUTION:
 - **Statistical sampling** to get a more accurate measurements

Sampling

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

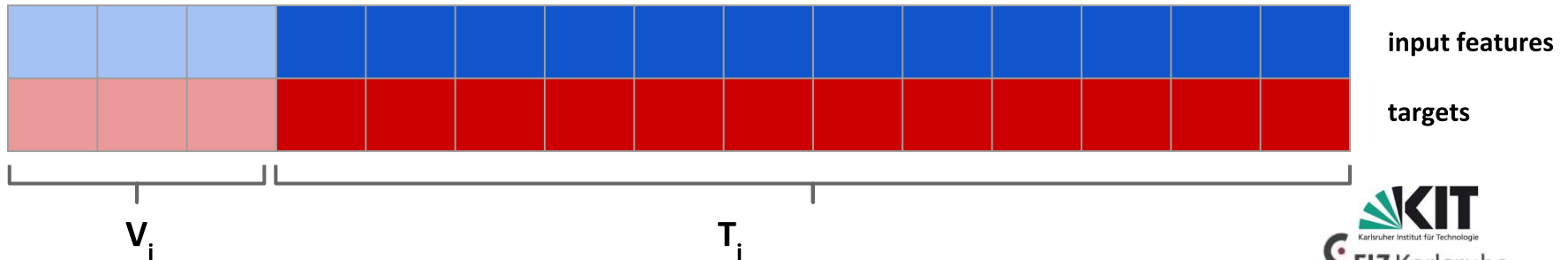
- The collected data should be divided into
 - Training Data (used to train the algorithm)
 - Validation Data (to keep track about the performance of the algorithm while it learns)
 - Test Data (used to evaluate the performance of the readily trained algorithm)



K-fold Cross Validation

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

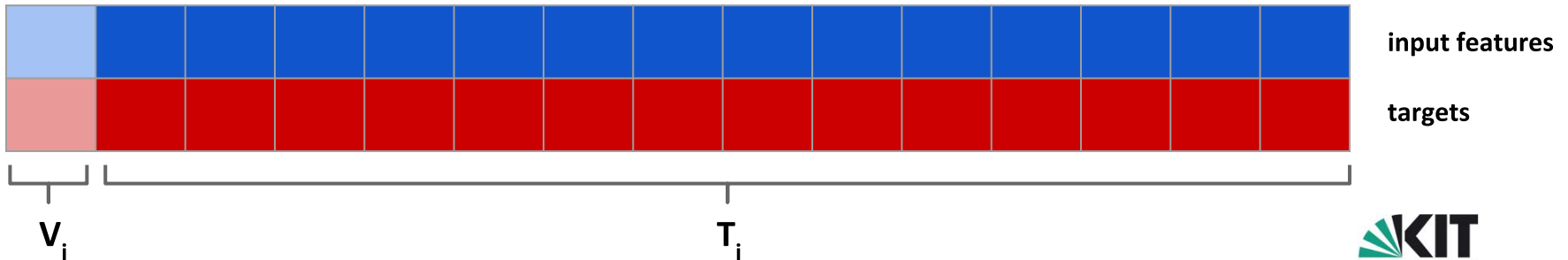
- The aim of **cross-validation** is to ensure that every example from the original dataset has the same chance of appearing in the training and testing set.
- In **K-fold cross-validation**, the dataset X is divided randomly into K equal- sized parts, $X_i, i = 1, \dots, K$.
- To generate each pair,
 - we keep **one of the K** parts out as the **validation set** $V_i = X_i$ and
 - combine the **remaining $K - 1$** parts to form the training set
$$T_i = X_1 \cup \dots \cup X_{i-1} \cup X_{i+1} \cup \dots \cup X_K$$



Leave-One-Out Cross Validation

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

- Special case of **cross-validation**, where given a dataset $\mathbf{X}$ of N instances, $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$
 - only **one instance** $\mathbf{x}_i \in \mathbf{X}$ is left out as the **validation set** (instance) $\mathbf{V}_i = \mathbf{x}_i$
 - and **training** uses the $N - 1$ remaining instances $\mathbf{T}_i = \{\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_N\}$

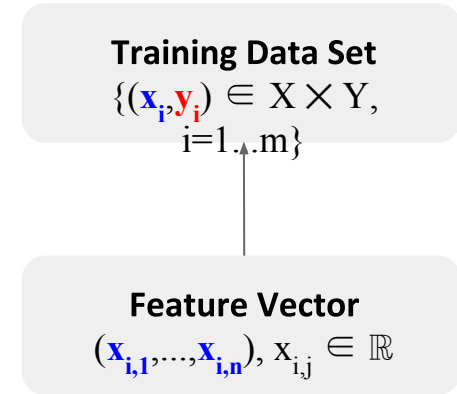


Feature Selection

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

- Select **attributes** of/from the available data that are **relevant to determine the projected outcome**
- Simple Example: **SPAM Detection**
 - Input: emails x
 - Feature Vector:

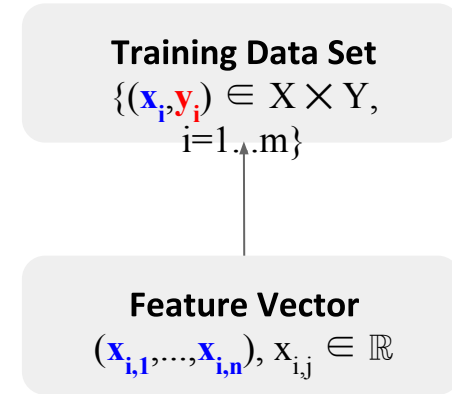
$$f(x) = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \dots \\ f(x_n) \end{bmatrix}, \text{ e.g., } f(x_i) = \begin{cases} 1 & \text{if the email contains "viagra"} \\ 0 & \text{otherwise} \end{cases}$$



Feature Selection

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

- Select **attributes** of/from the available data that are **relevant to determine the projected outcome**
- **Potential Difficulties:**
 - Irrelevant Attributes
 - Missing Attributes
 - Missing Attribute Values
 - Redundant Attributes
 - Attribute Value Noise



Evaluation

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

- To evaluate the performance of a ML model, the following **Accuracy Metrics** can be applied:

$$Accuracy = \frac{\#TP + \#TN}{\#TP + \#FP + \#TN + \#FN}$$

$$Recall = \frac{\#TP}{\#TP + \#FN}$$

$$Precision = \frac{\#TP}{\#TP + \#FP}$$

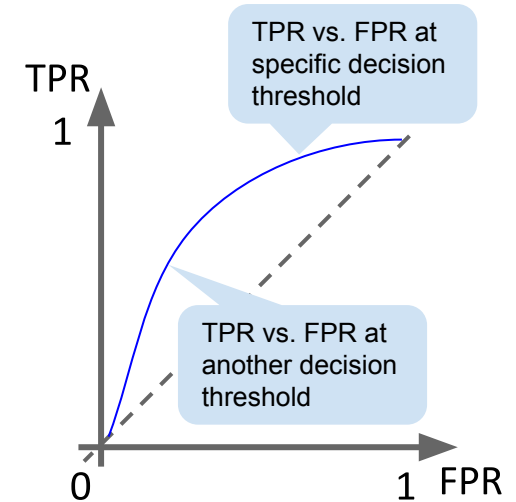
$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

		experiment	
		true	false
ground truth	true	true positive	false negative
	false	false positive	true negative

Evaluation

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

- How do we compare the performance of different models or models using different parameters?
- **ROC Curve (Receiver-Operator Characteristic)**
 - Y-axis: True Positive Rate $TPR = \frac{\#TP}{\#TP + \#FN}$
 - X-axis: False Positive Rate $FPR = \frac{\#FP}{\#FP + \#TN}$
 - An ROC curve plots **TPR vs. FPR at different classification thresholds**. Lowering the classification threshold classifies more items as positive, thus increasing both FP and TP.
 - **AUC (Area under the ROC curve)** measures the entire two-dimensional area underneath the entire ROC curve



The Curse of Dimensionality

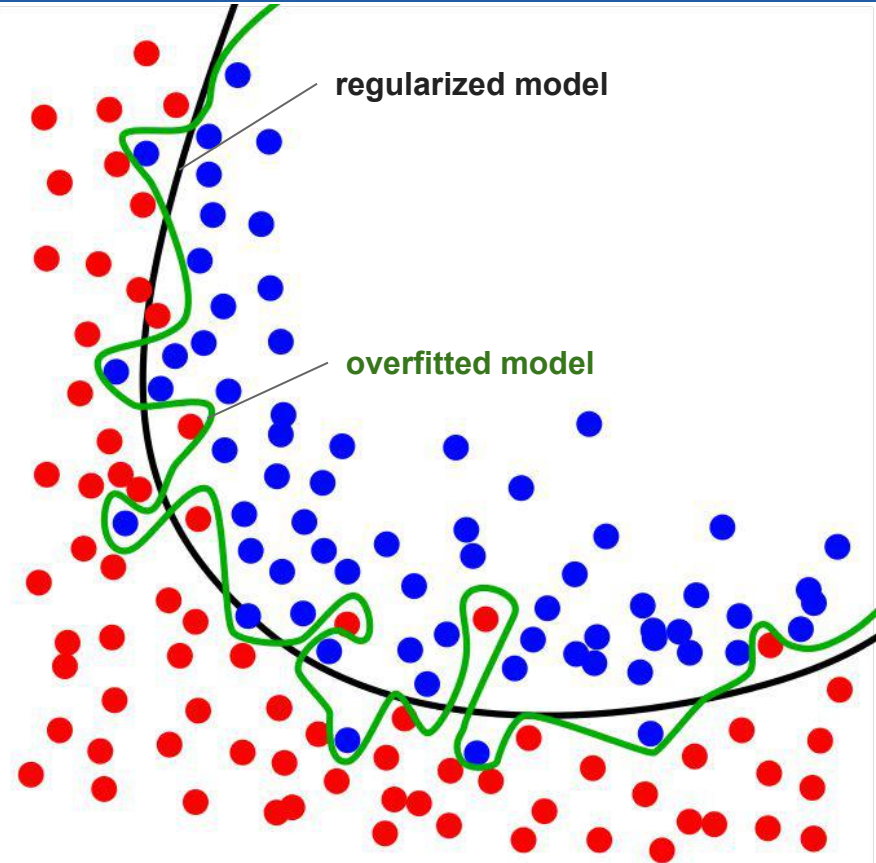
Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

- As **humans**, we are experts in 1D, 2D, and 3D space
 - a bit less in 4D (time)
 - and even less in higher dimensional spaces
- In ML, we are often dealing with much **higher dimensional spaces**, which for us often have **counter-intuitive characteristics**
- **Example:**
 - 1D space with k bins with n samples in each: $k \cdot n$
 - 2D space: $k^2 \cdot n$
 - 3D space: $k^3 \cdot n$
 - ...
- When the **dimensionality increases**, the volume of the space increases so fast that the available data become **sparse**.
- **Sparsity** is problematic for any method that requires statistical significance

Overfitting

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.3 Evaluation and Generalization Problems

- The **overfitted model** follows exactly the training data
 - Too high dependency on potential noise
 - Lack of generalisation
- The **overfitted model** is likely to have a higher error rate on new unseen data, compared to the **regularized model**
 - **Better generalisation** to the underlying classification function
- **Consequence:**
Stop training the model before the algorithm overfits



海川の大河に年々
大章魚と云ふ大魚
是れ捕らるる者
中より空しくして
くく足と云ふ
と目目と云ふ
知れしと云ふ
生死一瞬の間に
市店に於て賣れ
地にもあると云
の食はるるものと

Basic Machine Learning Algorithms

Picture References:

- P. 4: M. Banko, E. Brill. 2001. Scaling to very very large corpora for natural language disambiguation. ACL '01, pp. 26-33.
- P.5: Associated Press photo, <https://en.wikipedia.org/wiki/File:Deweytruman12.jpg>
- P.6: White Noise, <https://commons.wikimedia.org/wiki/File:White-noise-mv255-240x180.png>
- P.15: Hiroshige III, Hunting the Giant Octopus of Namekawa in Etchu Province (1877)
https://commons.wikimedia.org/wiki/File:Hunting_the_Giant_Octopus_of_Namekawa_in_Etchu_Province.jpg