Machine Learning

– winter term 2016/17 –

Chapter 01: Introduction

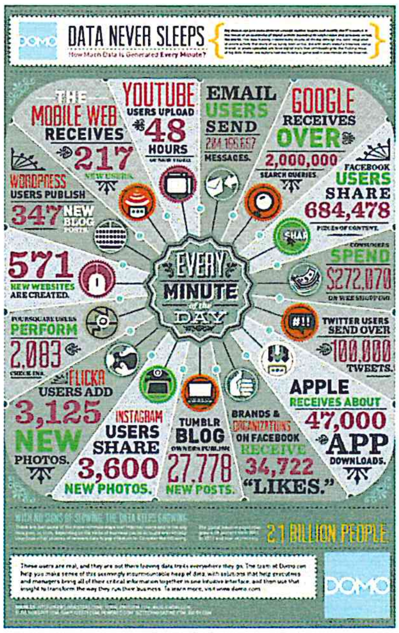
Prof. Adrian Ulges
Masters "Computer Science"
DCSM Department
University of Applied Sciences RheinMain

Outline

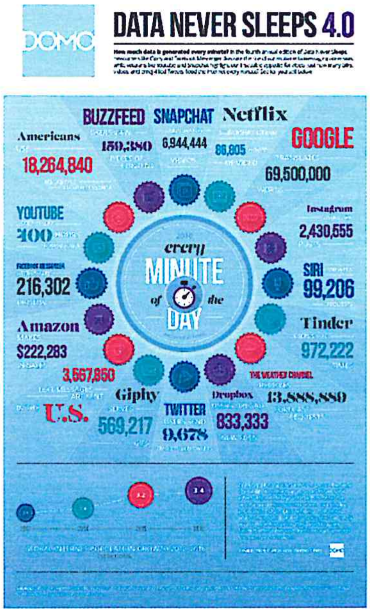
1. Motivation
2. Machine Learning Basics
3. Machine Learning vs X
4. Benchmarking ML Systems

Machine Learning ("ML") and Big Data images from [16]

(June 2012)


DOMO
DATA NEVER SLEEPS
How Much Data Is Generated Every Minute?
THE MOBILE WEB RECEIVES 217 NEW USERS
YOUTUBE USERS UPLOAD 48 HOURS OF NEW VIDEO
EMAIL USERS SEND 204,166,667 MESSAGES.
GOOGLE RECEIVES OVER 2,000,000 SEARCH QUERIES
WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS
FACEBOOK USERS SHARE 684,478 PIECES OF CONTENT.
571 NEW WEBSITES ARE CREATED.
EVERY MINUTE of the DAY
CONSUMERS SPEND $272,070 ON WEB SHOPPING
FOURSQUARE USERS PERFORM 2,083 CHECK-INS.
TWITTER USERS SEND OVER 100,000 TWEETS.
FLICKR USERS ADD 3,125 NEW PHOTOS.
APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS.
INSTAGRAM USERS SHARE 3,600 NEW PHOTOS.
TUMBLR BLOG OWNERS PUBLISH 27,778 NEW POSTS.
BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE 34,722 "LIKES."
2.1 BILLION PEOPLE
DOMO

(June 2016)

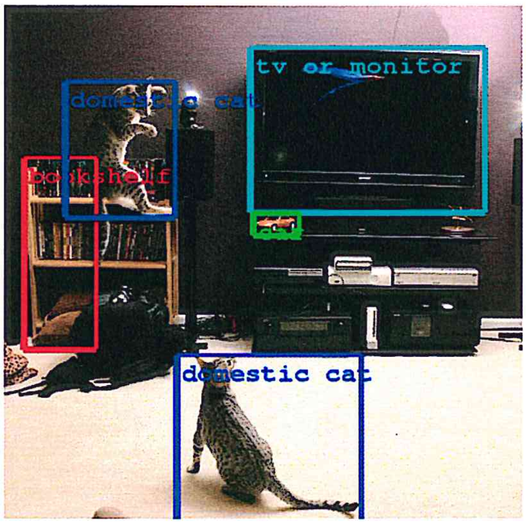


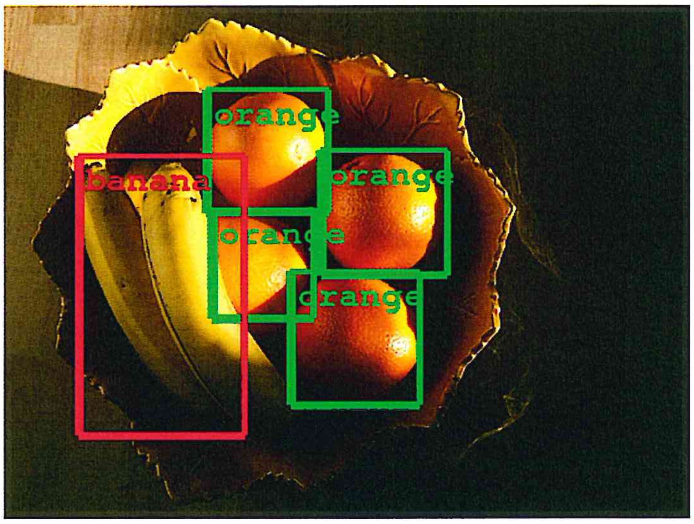
- There is more and more **unstructured** data around *(web 2.0, mobile devices, IoT, self-driving cars, ...)*

Machine Learning: Some recent Achievements image from [17]

"Imagine a historian of science writing about computer vision in the year 2100. They will identify the years 2011 to 2015 (and probably a few years beyond) as a time of **huge breakthroughs***, driven by deep convolutional nets."*

(M. Nielsen, "Neural Networks and Deep Learning")

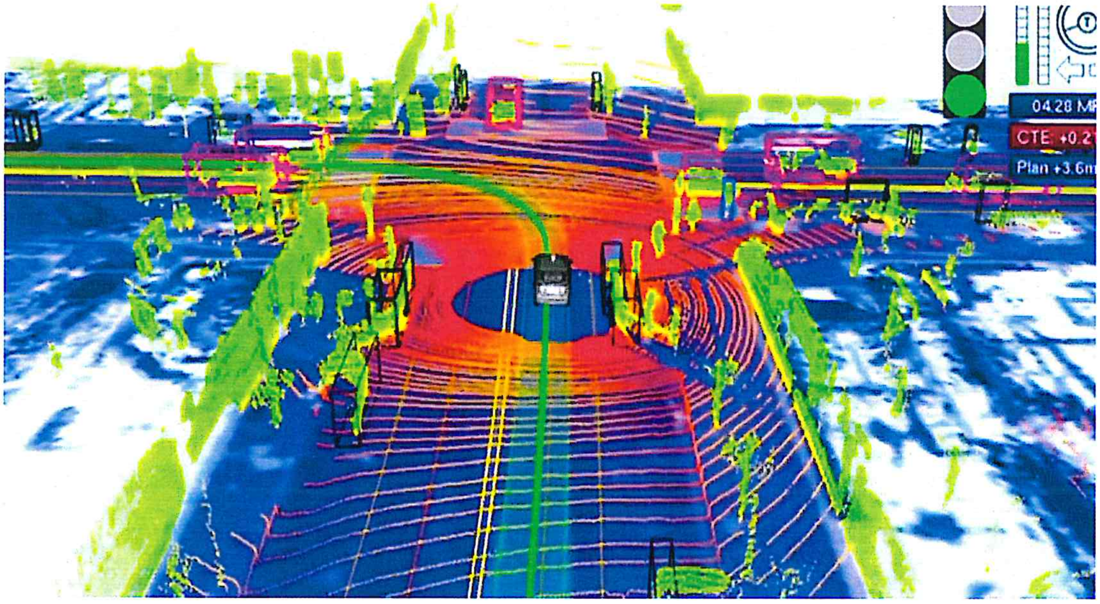




Machine Learning: Some recent Achievements image from [12]

"In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome."

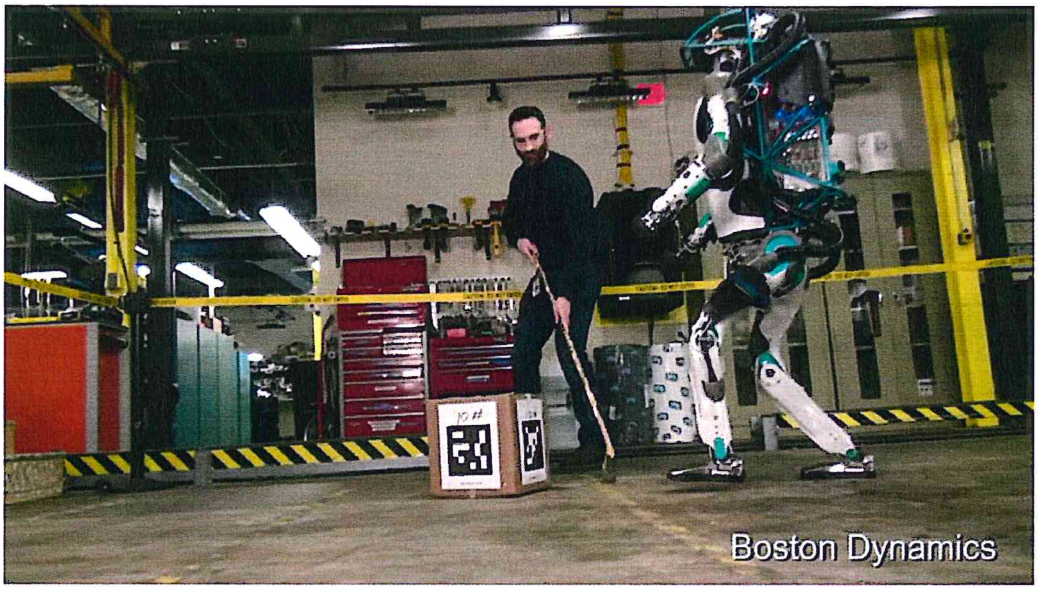
(Andrew Ng, Stanford University)



Machine Learning: Some recent Achievements image from [3]

"A survey carried out among AI experts recently shows they think machines will be as intelligent as humans by the year 2040."

(Nick Bostrom)



Atlas, The Next Generation

Machine Learning: Motivation

> *"If data had mass, the earth would be a black hole."*
>
> (Steven Marsland)

> *"As more data becomes available, more ambitious problems can be tackled. As a result, machine learning is* **widely used in computer science** *and other fields. However, developing successful machine learning applications requires a substantial amount of* **'black art'** *that is hard to find in textbooks."*
>
> (P. Domingos, *A few Useful Things to Know about Machine Learning*)

Example Applications

Have a look on kaggle.com

- Flight Quest

  optimize flight routes based on wheather and traffic
- TFI Restaurant Revenue Prediction

  predict annual sales of restaurants to open
- Job Recommendation Challenge

  predict which jobs users will apply to
- Whale Detection Challenge

  detect whale calls from audio, prevent collision with ship traffic
- Discovering trolling in user comments
- ...

There's also the 'classics'

- OCR, handwriting recognition, object recognition
- search engines, recommender systems, targeted advertising
- natural language processing, spam filtering

Outline

1. Motivation

2. Machine Learning Basics

3. Machine Learning vs X

4. Benchmarking ML Systems

An ML Sample Application images from [4] [13]

- A computer system is to make a non-trivial decision
- Example: **spam filtering**
- Why not **hard-code** the decision logic?

Problems

- high effort to grasp problem's **complexity**
- easy to code *something*, difficult to reach the **optimal** program
- **feasibility checking**: What accuracy can be reached by a decision?
- code is extremely difficult to **maintain**
- keeping track of **data changes** (e.g., when spammers change strategies) is almost impossible
- there is no way to take **user feedback** into account

Machine Learning: Definition

> *"Machine learning is a scientific discipline that explores the construction and study of* **algorithms that can learn from data**. *Such algorithms operate by building a* **model** *from* **example inputs** *and using that to make* **predictions or decisions**, *rather than following strictly static program instructions."*

(en.wikipedia.org)

> *"The field of study that gives computers the ability to learn* **without being explicitly programmed.**"

(Arthur Samuel (1959))

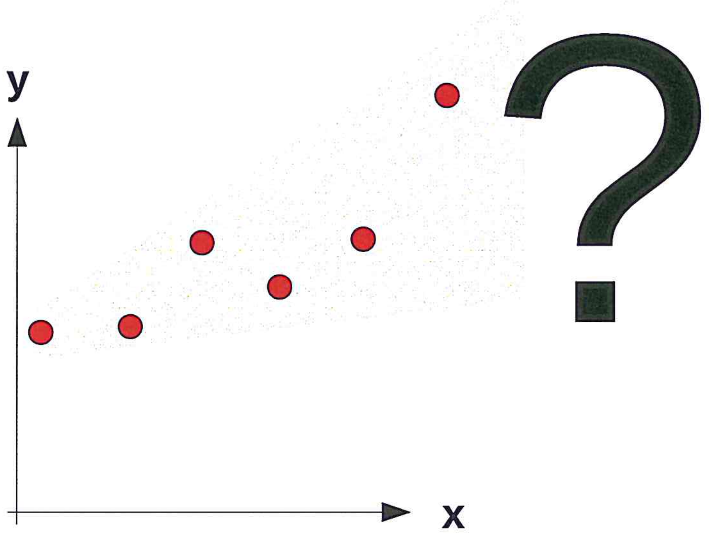
> *"A computer program is said to* **learn** *from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with Experience E."*

(Tom Mitchell (1998))

Remark

These definitions are entirely **non-operational**!

Machine Learning's Hello World



- **Goal**: predict a person's weight in the future
- **Given**: a set of 2D samples $(x_1, y_1), ..., (x_n, y_n)$ *(x is time, y is person's weight)*
- **Approach (***linear regression***)**: fit a line to the points and use this line for prediction
- Does this qualify as **machine learning**?

Machine Learning's Hello World

Linear Regression

- We define a line as a function

$$\mathcal{M}_\theta(x) = a \cdot x + b$$

- We measure the quality of a particular line $\theta = (a, b)$ using an error function E:

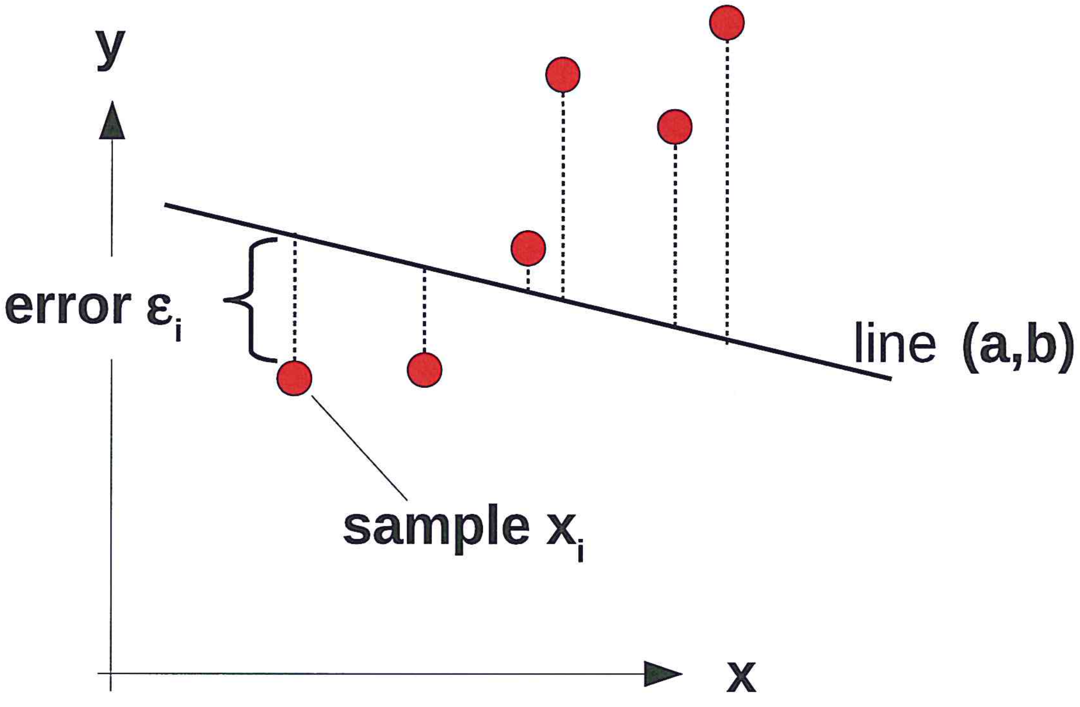
$$E(\theta) = \sum_{i=1}^{n} \overbrace{\Big(\underbrace{a \cdot x_i + b}_{\mathcal{M}(x_i)} - y_i \Big)^2}^{\text{error } \epsilon_i^2}$$

- The **best line** is the one that **minimizes** E:

$$(a^*, b^*) = \underset{\theta \in \mathbb{R}^2}{\arg\min} \; E(\theta)$$

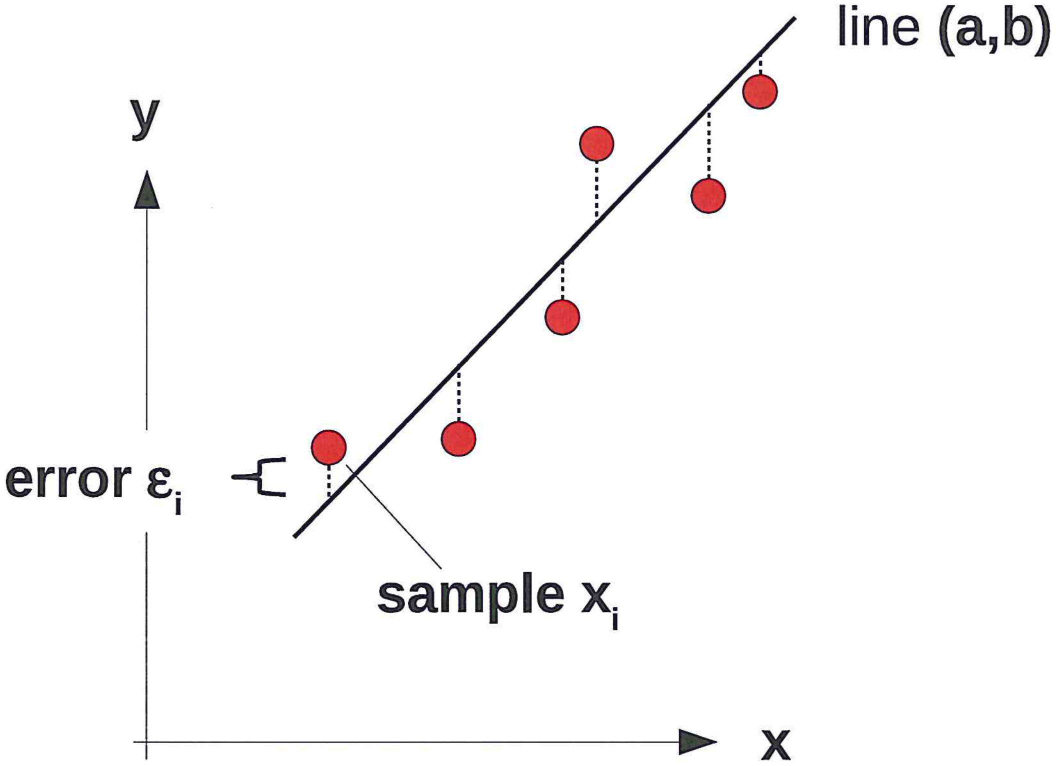
Machine Learning's Hello World

a bad line: The errors $\epsilon_1^2, \ldots, \epsilon_n^2$ are high → E is high



Machine Learning's Hello World

a good line: The errors $\epsilon_1^2, \ldots, \epsilon_n^2$ are low → E is low



Machine Learning's Hello World

We minimize E by settings its gradient to zero, followed by a bit of algebra:

$$\partial E/\partial b = \partial\Big(\sum_i (a \cdot x_i + b - y_i)^2\Big)/\partial b = 0$$

$$2 \cdot \sum_i (a \cdot x_i + b - y_i) = 0$$

$$n \cdot b = \sum_i y_i - a \cdot \sum_i x_i$$

$$b = \bar{y} - a\bar{x}$$

$$\partial E/\partial a = \partial\Big(\sum_i (a \cdot x_i + b - y_i)^2\Big)/\partial a = 0$$

$$2 \cdot \sum_i (a \cdot x_i + b - y_i) \cdot x_i = 0$$

$$\sum_i (a \cdot x_i + (\bar{y} - a\bar{x}) - y_i) \cdot x_i = 0$$

$$a\Big(\sum_i x_i^2 - \bar{x}\sum_i x_i\Big) = \sum_i y_i x_i - \bar{y}\sum_i x_i \qquad // :n$$

$$a = s_{xy}/s_x^2$$

Machine Learning Terminology

- The problem above is a **regression problem**:
  It is about predicting a real-valued variable y.
- We call the points $(x_1, y_1), ..., (x_n, y_n)$ the **training samples**.
- We call our line function $\mathcal{M}(x) = a \cdot x + b$ the **model**.
- We call the process of estimating the model parameters (here, a and b) **training**.
- A typical training strategy is to formulate an error criterion (here, E) and **optimize** this criterion.
- In our example, we were able to pin down the solution by hand (we say: there is an *analytical* solution). In practice, optimization can be much trickier. Not all functions are easy to optimize.
- Therefore, learning is often done by **local search**.

Machine Learning Terminology

- Obviously, picking the **right model** for the right data is tricky.
  It is the core problem in machine learning, actually.
- **Example**: Would those be better models?

$$\mathcal{M}_{a,b,c,d}(x) = a + b \cdot sin(c \cdot x + d)$$

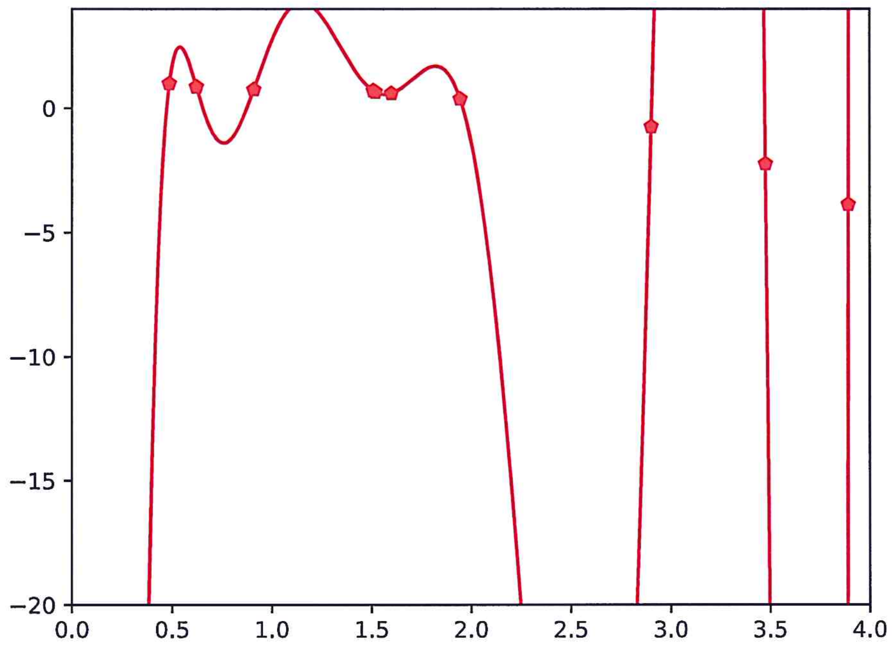
$$\mathcal{M}_{a,b,c,d,e}(x) = a + b \cdot sin(c \cdot x + d) + e \cdot x$$

$$\mathcal{M}_{a_0,a_1,...,a_{100}}(x) = \sum_{i=0}^{100} a_i \cdot x^i$$

Machine Learning Terminology

Example: Runge's Phenomenon

- Fitting an 8-degree polynomial to 9 points



Machine Learning Terminology

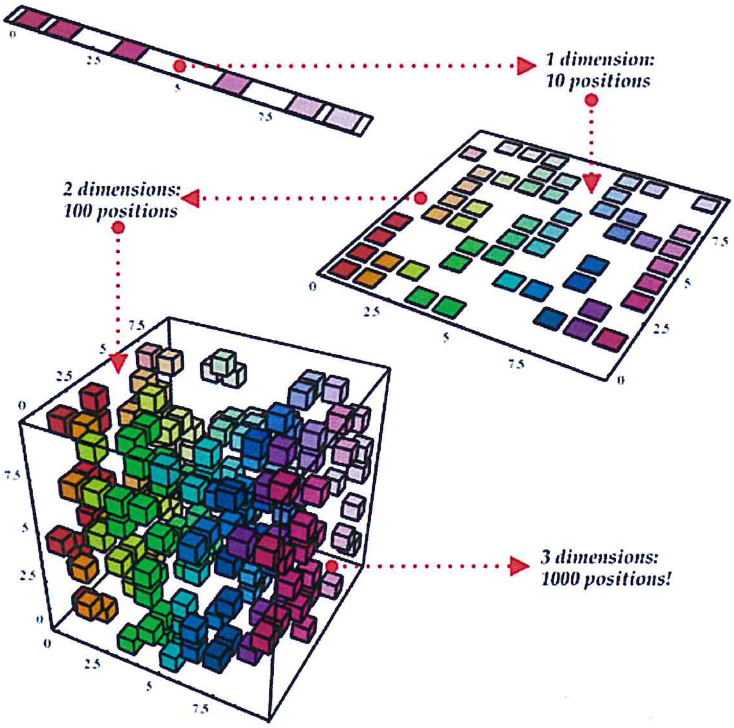
> *"The real value of a scientific explanation lies not in its ability to explain (what one has already seen), but in predicting events that have yet to (be seen)"*
>
> (Blumer et al. 1987)

> *"With four parameters I can fit an elephant and with five I can make him wiggle his trunk."*
>
> (John von Neumann)

- Our polynomial model works well on the training data but **generalizes poorly** to novel data. The model **overfits**.
- We call $\theta = (a, b)$ the model's **parameters**.
  In practical models there may be thousands of parameters.
- Overfitting is usually more severe...
  - ... the more parameters a model has.
  - ... the fewer training samples we have.

Machine Learning: The Curse of Dimensionality image from [10]

- Overfitting is also more severe the **more dimensions** we have.
- **Reason**: As the number of dimensions increases, we require *more and more data* to populate our input space!

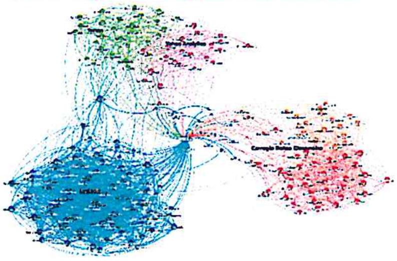


Machine Learning: Challenges images from [14] [9] [2] [1]

Regression

The Hunger Games

Clustering / Segmentation



Recommendation

amazon.com Recommended for You

Amazon.com has new recommendations for you based on items you purchased or told us you own.

The Little Big Things: 163 Ways to Pursue EXCELLENCE

Fascinate: Your 7 Triggers to Persuasion and Captivation

Sherlock Holmes [Blu-ray]

Alice in Wonderland [Blu-ray]

Classification

"cyLikelihood": "VERY_LIKELY"

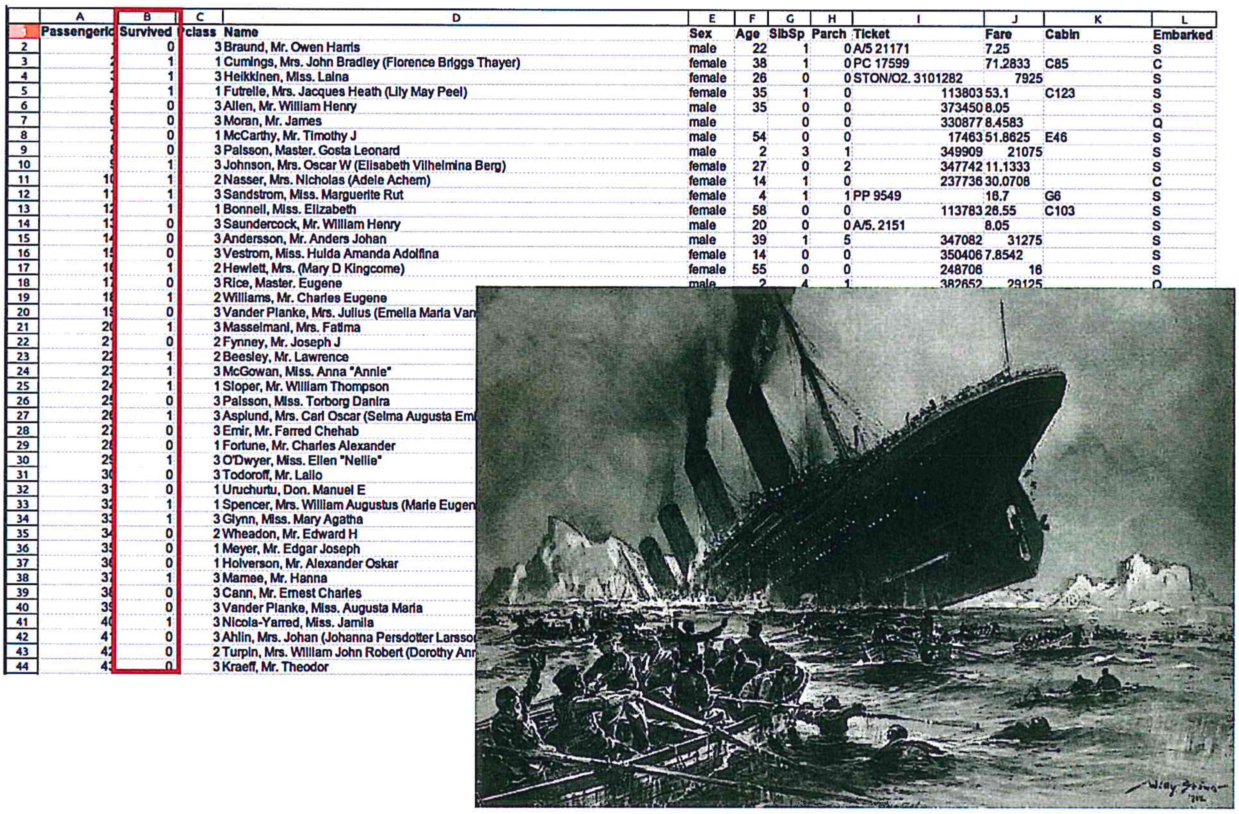
,"description": "ABIERTO\n",
"locale": "es"

Data Reduction

Beer Market
Perceptual Mapping
Regular
Heavy
Popular with Men
Full Bodied
Old Milwaukee
Budweiser
Beck's
Heineken
Meister Brau
Miller
Special Occasions
Blue Collar
Good Value
Dining Out
Premium
Coors
Budget
Stroh's
Premium
Michelob
Miller Lite
Coors Light
Popular with Women
Pale Color
On a Budget
Old Milwaukee Light
Light
Less Filling
Light

Anomaly Detection

Machine Learning: Small Example image from [15]



	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7925		S
5	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21075		S
10	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
14	13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
15	14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31275		S
16	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
17	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
18	17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29125		Q
19	18	1	2	Williams, Mr. Charles Eugene								
20	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Van								
21	20	1	3	Masselmani, Mrs. Fatima								
22	21	0	2	Fynney, Mr. Joseph J								
23	22	1	2	Beesley, Mr. Lawrence								
24	23	1	3	McGowan, Miss. Anna "Annie"								
25	24	1	1	Sloper, Mr. William Thompson								
26	25	0	3	Palsson, Miss. Torborg Danira								
27	26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Em								
28	27	0	3	Emir, Mr. Farred Chehab								
29	28	0	1	Fortune, Mr. Charles Alexander								
30	29	1	3	O'Dwyer, Miss. Ellen "Nellie"								
31	30	0	3	Todoroff, Mr. Lalio								
32	31	0	1	Uruchurtu, Don. Manuel E								
33	32	1	1	Spencer, Mrs. William Augustus (Marie Eugen								
34	33	1	3	Glynn, Miss. Mary Agatha								
35	34	0	2	Wheadon, Mr. Edward H								
36	35	0	1	Meyer, Mr. Edgar Joseph								
37	36	0	1	Holverson, Mr. Alexander Oskar								
38	37	1	3	Mamee, Mr. Hanna								
39	38	0	3	Cann, Mr. Ernest Charles								
40	39	0	3	Vander Planke, Miss. Augusta Maria								
41	40	1	3	Nicola-Yarred, Miss. Jamila								
42	41	0	3	Ahlin, Mrs. Johan (Johanna Persdotter Larsso								
43	42	0	2	Turpin, Mrs. William John Robert (Dorothy Ann								
44	43	0	3	Kraeff, Mr. Theodor								

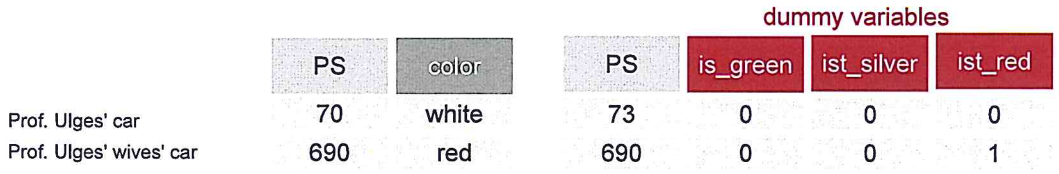
Some more ML Terminology

- ML's goal is to make predictions about real-world **objects**
  - **SPAM filtering**: objects = e-mails
  - **route planning**: objects = routes to drive
- Our goal is to make a prediction regarding an object.
  We call this prediction a **label** or **target**
  - **SPAM filtering**: label $\in \{spam/ham\}$
  - **route planning**: label $\in \mathbb{R}_0^+$ (= *time to destination*)
- We describe each object by a number of **features**
  - **Titanic**: features = gender, ticket price, passenger class, ...
- The collection of the features describing an object
  is called a **feature vector** (and usually denoted with $\mathbf{x}$).
- Features can be **categorial** or **numerical**.
- In case of numerical features, the object can be interpreted as a **point** in a (high-dimensional) space!

ML: Feature Engineering / Feature Extraction

Often, we **prepare** the input data before applying ML

1. Features may be **missing**, i.e. $\mathbf{x}$ is *incomplete*.
   Approach: estimate missing values (*imputation*)
2. **Categorical** features may have to be transformed into numerical ones, typically by introducing **dummy variables** *(one-hot encoding)*

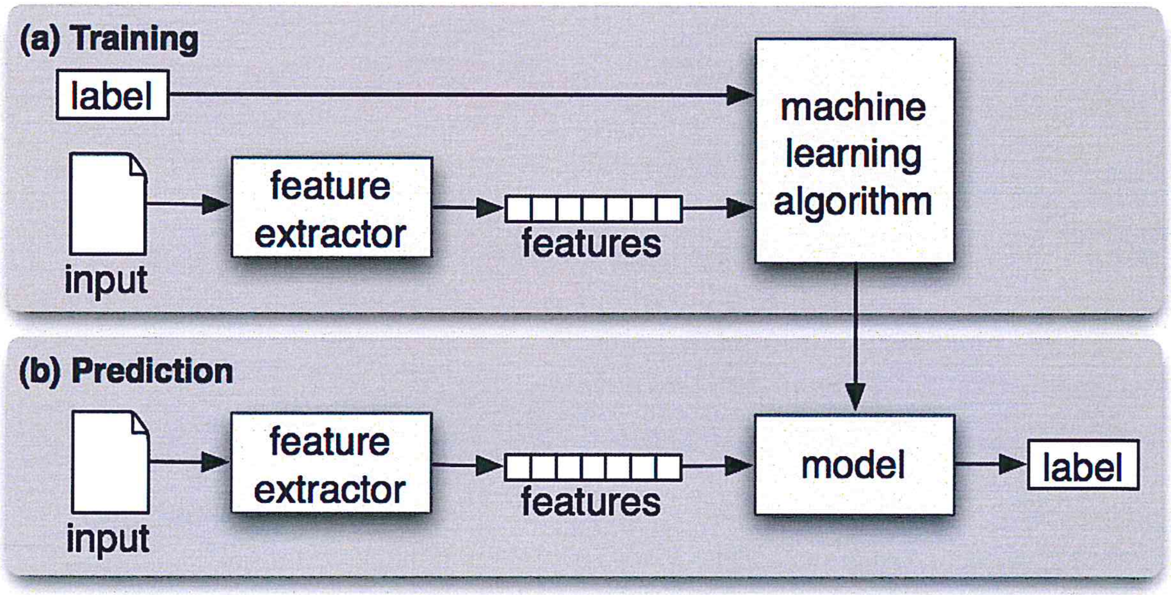


				dummy variables		
	PS	color	PS	is_green	ist_silver	ist_red
Prof. Ulges' car	70	white	73	0	0	0
Prof. Ulges' wives' car	690	red	690	0	0	1

3. Often, we discard **outliers**
4. Often, we try to pre-select **'important'** features.

A Basic ML System Pipeline[1]

1. We **train** the system in an off-line phase, obtaining a **model**
2. We **apply** the system in an on-line phase



[1] image source: 99designs.com

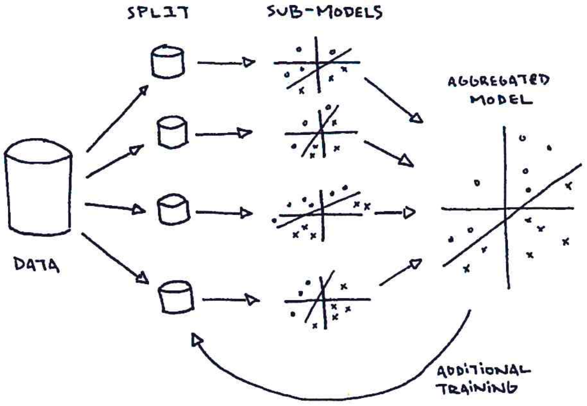
Types of Machine Learning

- **classification vs. regression**
  in classification, the labels $y_1, ..., y_n$ are categorical. In regression, they are real-valued.
- **supervised learning**
  learning from samples $\mathbf{x}_1, ..., \mathbf{x}_n$ with labels $y_1, ..., y_n$
- **unsupervised learning**
  learning only from samples $\mathbf{x}_1, ..., \mathbf{x}_n$, no labels
- **semi-supervised learning**
  learning from samples $\mathbf{x}_1, ..., \mathbf{x}_n$, some with labels
- **active learning**
  the system can pick which samples to label
- **ensemble learning**
  ... is about combining learners for a more robust decision
- **reinforcement learning**
  ... is about learning from feedback instead of labels
- ...

Types of Machine Learning (cont'd) image from [5]

Online Learning

- Perspective so far: **batch learning**
  (train off-line, apply on-line)
- What if data is too large for memory? Does the model allow a **sharding** of training data? *example: linear regression*



- What if data come in dynamic streams?
  (→ **data stream mining** / online learning)

Outline

1. Motivation

2. Machine Learning Basics

3. Machine Learning vs X

4. Benchmarking ML Systems

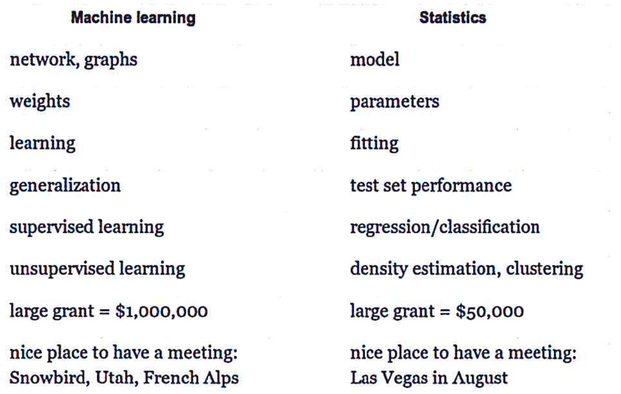
ML vs. Statistics image from [11]

> *"Statistics is the science of learning from data. Machine Learning is the science of learning from data. These fields are identical in intent, although they differ in their history, conventions, emphasis and culture."*

("The rise of the machines" – Larry Wasserman, CMU)

ML vs. Statistics

- The term "machine learning" has a negativ overtone in statistics *("Why do Statisticians hate us?")*
- Both fields address the same problems with similar methods



Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant = $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

ML: Relation to other Fields (cont'd)

ML vs. Artificial Intelligence (AI)

- Machine Learning is a subfield of **subsymbolic** AI
- The other part of AI – **symbolic** AI – is about logic, discrete search and rule-based inference.

ML vs. Optimization

- Machine learning techniques usually employ optimization
- "Machine learning = optimization + **generalization**"

ML vs. Data Mining

- Data mining focuses on **exploratory data analysis** *(by a human expert)*, machine learning on **automatic inference**.
- Data mining experts use machine learning, and machine learning experts use data mining.

Outline

1. Motivation

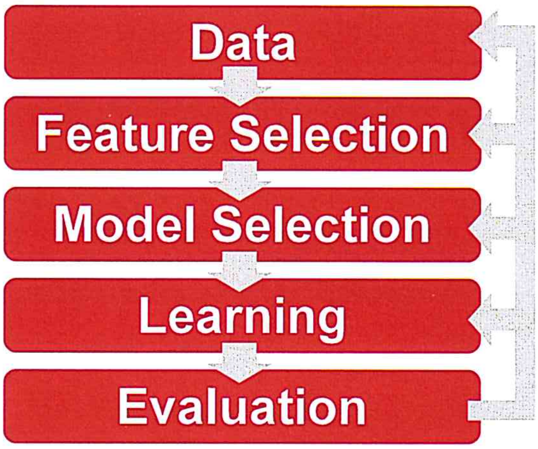
2. Machine Learning Basics

3. Machine Learning vs X

4. Benchmarking ML Systems

Why is Benchmarking so important?

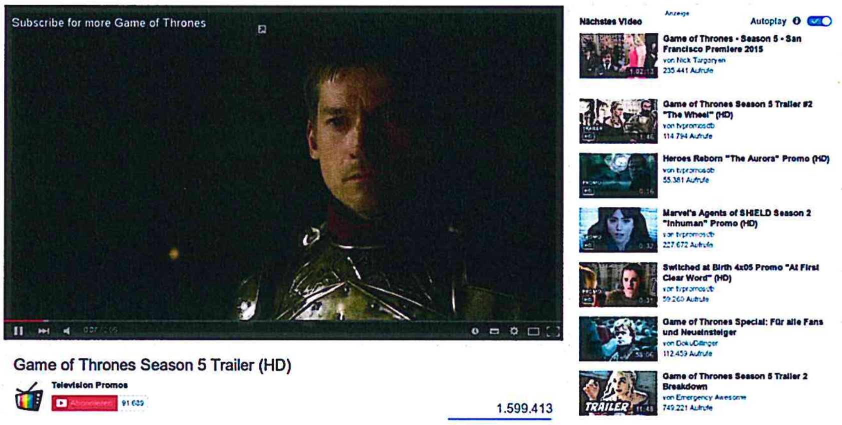
- **Machine Learning** = iterative re-design of...
  - data
  - features
  - models
  - parameters
- **Key Driver**: **Evaluation / Benchmarking**



Benchmarking ML Systems: Motivation image from [7]

Example: YouTube recommender

- We add a **"user history" feature** to the recommender.
- For example, the feature could merge some videos from the user's history into the recommendations.
- YouTube redirects some volume of **traffic** to the new system.
- YouTube benchmarks the new system against the old one.



Game of Thrones Season 5 Trailer (HD)
Television Promos
91 639
1.599.413

Nächstes Video

Game of Thrones - Season 5 - San Francisco Premiere 2015
von Nick Targaryen
235.441 Aufrufe

Game of Thrones Season 5 Trailer #2 "The Wheel" (HD)
von tvpromosdb
114.794 Aufrufe

Heroes Reborn "The Aurora" Promo (HD)
von tvpromosdb
55.381 Aufrufe

Marvel's Agents of SHIELD Season 2 "Inhuman" Promo (HD)
von tvpromosdb
227.672 Aufrufe

Switched at Birth 4x05 Promo "At First Clear Word" (HD)
von tvpromosdb
59.260 Aufrufe

Game of Thrones Special: Für alle Fans und Neueinsteiger
von DokuDillinger
112.459 Aufrufe

Game of Thrones Season 5 Trailer 2 Breakdown
von Emergency Awesome
749.221 Aufrufe

Benchmarking: Ground Rule image from [6]

> *"The most common mistake among machine learning beginners is to test on the training data and have the illusion of success."*
>
> (P. Domingos, *A few Useful Things to Know about Machine Learning*)

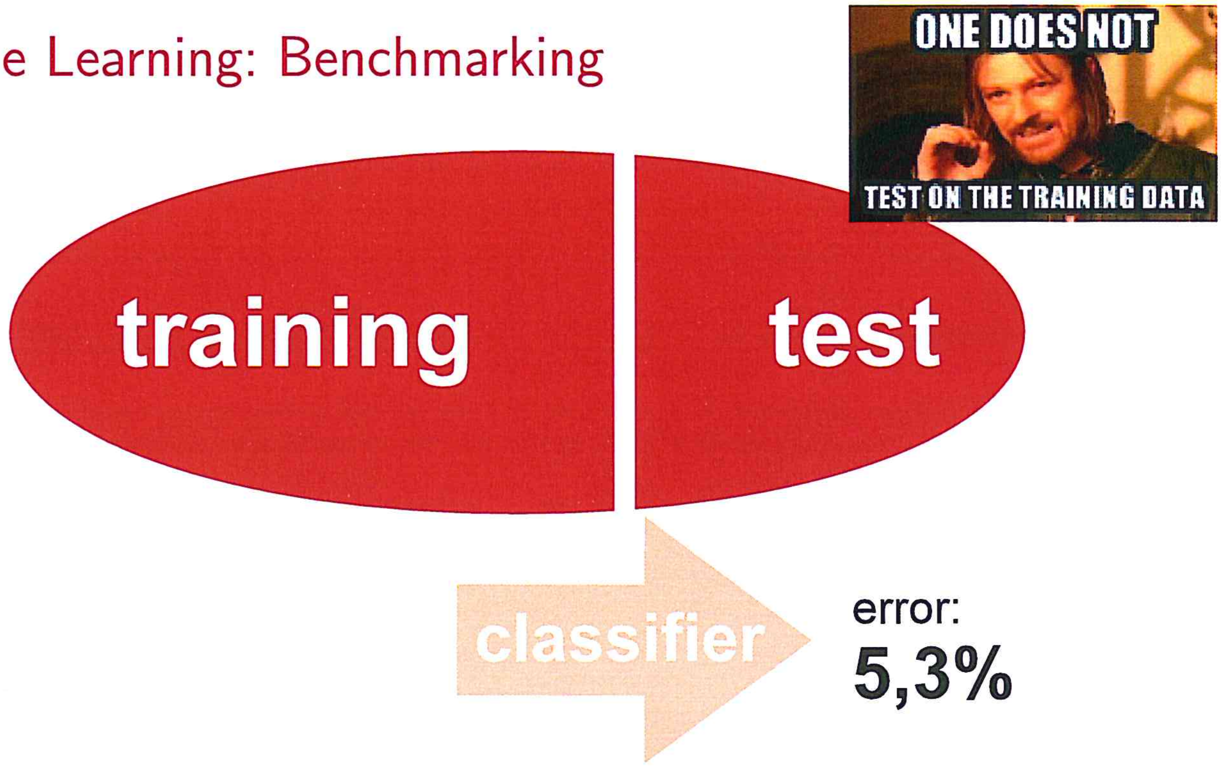
Why?

- All classifiers are prone to **overfitting**
- Achieving perfect accuracy on the training samples is quite simple *(by simply memorizing them)*.
- It is the generalization to **new data** that matters!
- When building classifiers, always set some test data aside!

Adhering to the Ground Rule is trickier than you think!

- **Example**: neural network training
- **Example**: standard datasets in research

Machine Learning: Benchmarking



training

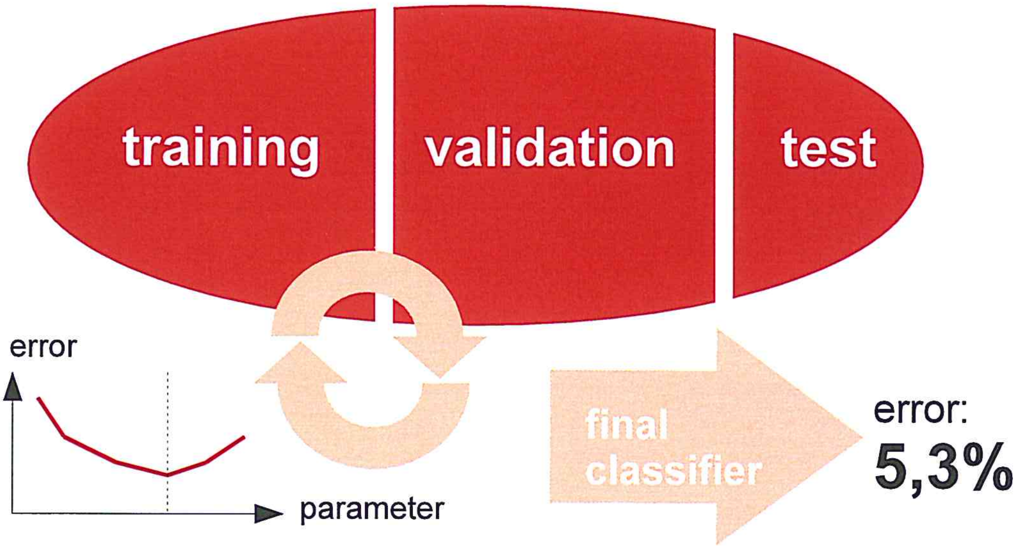
test

classifier

error:
5,3%

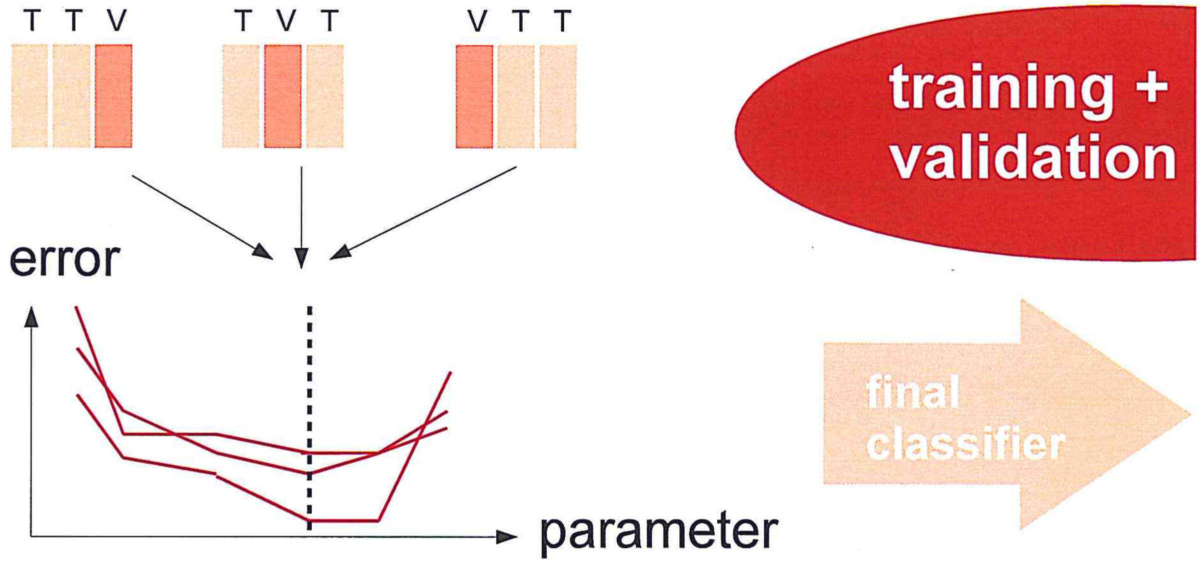
- We separate our dataset into **training and testing data**
- **Tip 1**: choose 'just enough' test data, use the rest for training
- **Tip 2**: use a roughly balanced class distribution

Machine Learning: Benchmarking



- Some of a classifiers' parameters are usually **learned**, other **(free)** parameters are set **manually**
- Typical approach: **grid search**
- **Example**: decision tree (later today) → *parameter* = *depth*
- **Approach**: train → **validate** → test

Machine Learning: Benchmarking



- If there is **small training data**, we apply **cross-validation**: Split the data into subsets ("*folds*"), train/validate multiple times, and average the results
- Extreme case: **leave-one-out validation**