Prof. Dr. Adrian Ulges

Empolis Workshop “Machine Learning”

Machine Learning 101

Hochschule RheinMain
Department DCSM *(Design, Computer Science, Media)*

Überblick

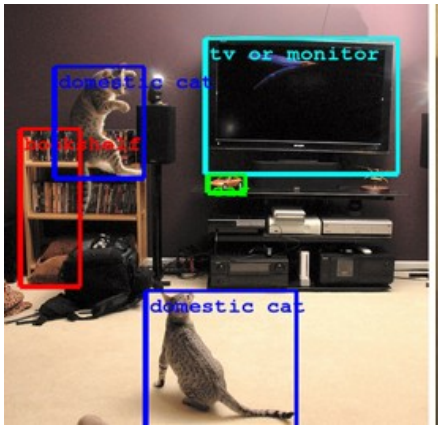
Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

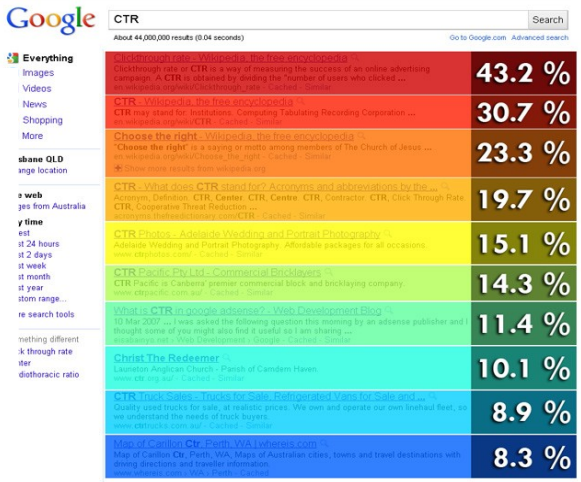
- Machine Learning 101
  - Grundlagen, Begriffsbildung
  - Design-Cycle, Benchmarking, Cross-Validation
  - Python-Beispiel: Titanic

ML: Definition

Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

- **Arthur Samuel (1959)**: „The field of study that gives computers the ability to learn ***without being explicitly programmed***“
- **Tom Mitchell (1998)**: „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***.“





TITLE: Impact of position on clickthrough rate in search results
Segment: Australian Retail/e-Commerce Traffic, Year: 2011
Credits: http://dejanseo.com.au

ML: Beispiele

kaggle™

- **Large-scale hierarchical Text Classification**
  (categorize Wikipedia articles into one of 325,056 categories)
- **Sentiment Analysis on Movie Reviews**
- **Job Recommendation**
  (predict which jobs users will apply to)
- **Axa Driver Telematics challenge**
  (given GPS routes, identify a cars driver)
- **Flight Quest**
  (optimize flight routes based on wheather and traffic)
- **TFI Restaurant Revenue Prediction**
  (predict annual sales of restaurants)
- **Algorithmic Trading** challenge
- **Whale Detection** Challenge
- **Merck molecular Activity** Challenge
- **Psychopathy Prediction** based on Tweets
- Forecasting **ESC Votings**

ML: Eingabedaten

Löw, Joachim, geb. 3.2.1960 in Freiburg, Mittelfeldspieler und Stürmer, 1,81 m, 72 kg, Stammverein Eintracht Freiburg.

- Die ***Eingabedaten*** für Lernverfahren sind üblicher Weise aus **Samples** und **Labels**.
- ***Samples*** $\mathbf{x_1,\ldots,x_n}$: beschreiben Objekte der Welt *(Bilder / Dokumente / Personen / Queries / Situationen / …)*
- Samples bestehen üblicher Weise aus ***Merkmalen***
- Typen von Merkmalen ...
  - **nominal** *(Kategorien ohne Ordnung, z.B. Farben)*
  - **ordinal** *(Kategorien mit Ordnung, z.B. Häufigkeiten)*
  - **Intervallskaliert** *(Zahlen mit Abstandsmaß, z.B. Kalendertage)*
  - **verhältnisskaliert** *(Zahlen mit Nullpunkt, z.B. Gewicht)*
  - **Fehlend!**
- Alternativen (z.B. objekt-orientiert, Graphen): *Heute nicht*
- ***Labels*** $\mathbf{y_1,\ldots,y_n}$: repräsentieren eine Aussage über Objekte
  - Eine **Zuordnung** zu einer Klasse / Gruppe
  - Ein reellwertiges **Attribut** (z.B. ein Preis)

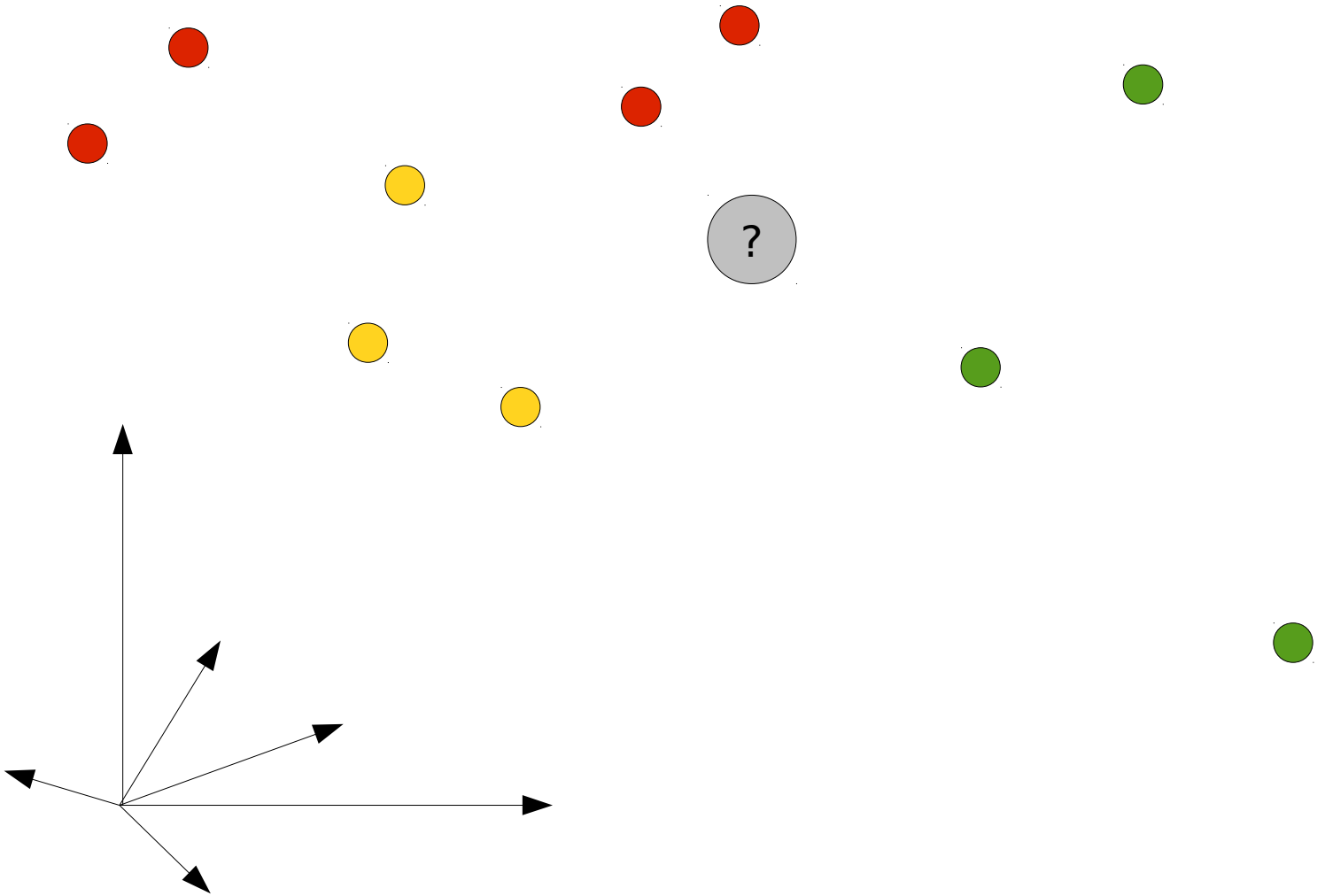
ML: Eingabedaten

Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

	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	male		0	0	244373	13		S
20	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18		S
21	20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7225		C
22	21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26		S
23	22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S
24	23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292		Q
25	24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
26	25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21075		S
27	26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)	female	38	1	5	347077	31.3875		S
28	27	0	3	Emir, Mr. Farred Chehab	male		0	0	2631	7225		C
29	28	0	1	Fortune, Mr. Charles Alexander	male	19	3	2	19950	263	C23 C25 C27	S
30	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female		0	0	330959	7.8792		Q
31	30	0	3	Todoroff, Mr. Lalio	male		0	0	349216	7.8958		S
32	31	0	1	Uruchurtu, Don. Manuel E	male	40	0	0	PC 17601	27.7208		C
33	32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female		1	0	PC 17569	146.5208	B78	C
34	33	1	3	Glynn, Miss. Mary Agatha	female		0	0	335677	7.75		Q
35	34	0	2	Wheadon, Mr. Edward H	male	66	0	0	C.A. 24579	10.5		S
36	35	0	1	Meyer, Mr. Edgar Joseph	male	28	1	0	PC 17604	82.1708		C
37	36	0	1	Holverson, Mr. Alexander Oskar	male	42	1	0	113789	52		S
38	37	1	3	Mamee, Mr. Hanna	male		0	0	2677	7.2292		C
39	38	0	3	Cann, Mr. Ernest Charles	male	21	0	0	A./5. 2152	8.05		S
40	39	0	3	Vander Planke, Miss. Augusta Maria	female	18	2	0	345764	18		S
41	40	1	3	Nicola-Yarred, Miss. Jamila	female	14	1	0	2651	11.2417		C
42	41	0	3	Ahlin, Mrs. Johan (Johanna Persdotter Larsson)	female	40	1	0	7546	9475		S
43	42	0	2	Turpin, Mrs. William John Robert (Dorothy Ann Wonnacott)	female	27	1	0	11668	21		S
44	43	0	3	Kraeff, Mr. Theodor	male		0	0	349253	7.8958		C

ML: Eingabedaten

Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain



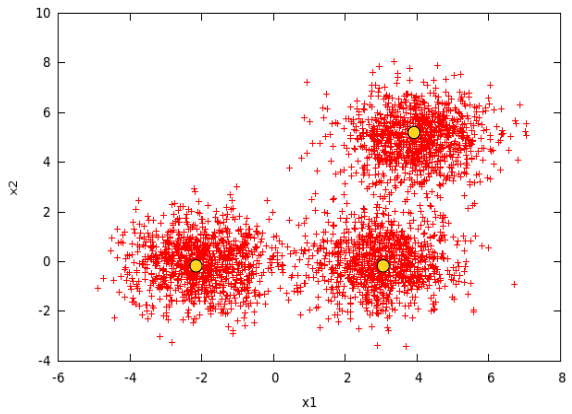
ML: Problemstellungen

Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

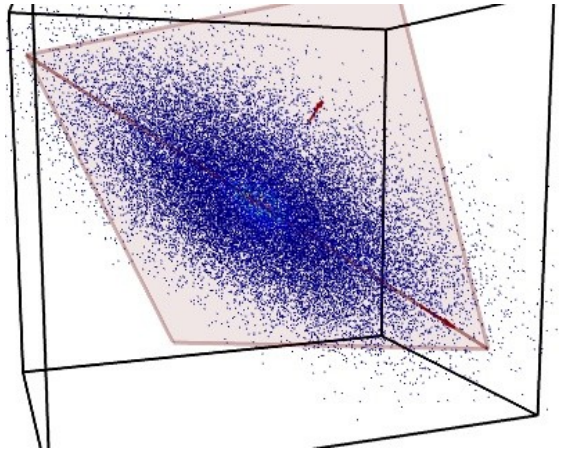
- ***Überwachtes Lernen***
  - **Gegeben**: Trainingssamples $\mathbf{x_1,...,x_n}$ und Labels $\mathbf{y_1,...,y_n}$
  - **Aufgabe**: Ordne Samples **x** die richtige Klasse **y** zu
- ***Unüberwachtes Lernen***
  - **Gegeben**: Trainingssamples $\mathbf{x_1,...,x_n}$ ***ohne Labels***
  - **Wichtig**, denn wir besitzen oft viele Daten aber keine Labels!
  - **Aufgabe**: Inferenz der ***Struktur der Daten***
- ***Halbüberwachtes Lernen (Mischform)***
  - **Gegeben**: Trainingssamples $\boldsymbol{x_1, ..., x_n}$, ***manche*** mit Labels
- **Weitere Unterscheidungen**:
  - Transdutives vs. Induktives Lernen
  - Reinforcement learning, **active learning**

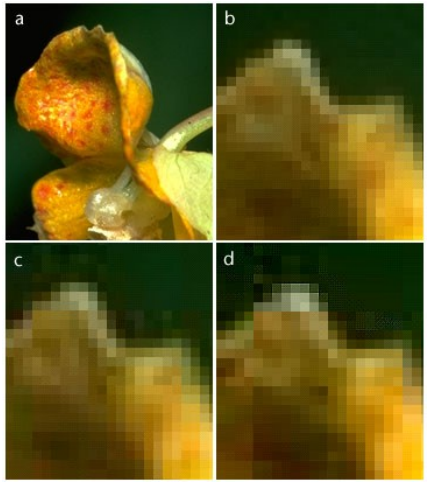
Unüberwachtes Lernen

Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain



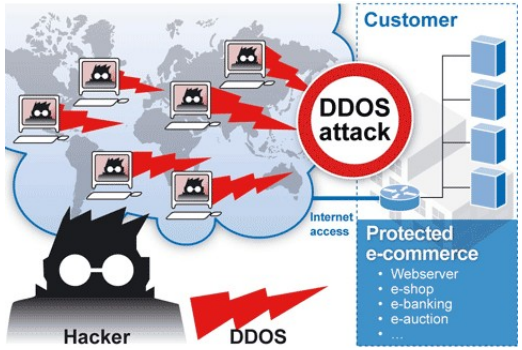
Clustering: Teile die Daten in kohärente Gruppen ein





Dimensionality Reduction: Komprimiere die Daten

Itemset Mining: Finde häufige Substrukturen in den Daten



Anomaly Detection: Finde Outlier / Ausreißer in den Daten

Überwachtes Lernen

Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

- **Zum überwachten Lernen:** Hier unterscheiden wir zwei Fragestellungen: ***Regression*** und ***Klassifikation***
- **Klassifikation**
  - Die Labels $\mathbf{y_1,...,y_n}$ sind ***nominale Variablen*** (d.h., ***Klassen***)
  - Aufgabe: Ordne Sample **x** einer Klasse **y** zu
  - ***Beispiel***: SPAM-Filter, Optical Character Recognition (OCR)
- **Regression**
  - Die Labels $\mathbf{y_1,...,y_n}$ sind ***reellwertige Variablen***
  - Aufgabe: Ordne einem Sample **x** einen Zahlenwert zu
  - ***Beispiel***: Fahrtdauer-Prognose im Routenplaner

ML: Terminologie

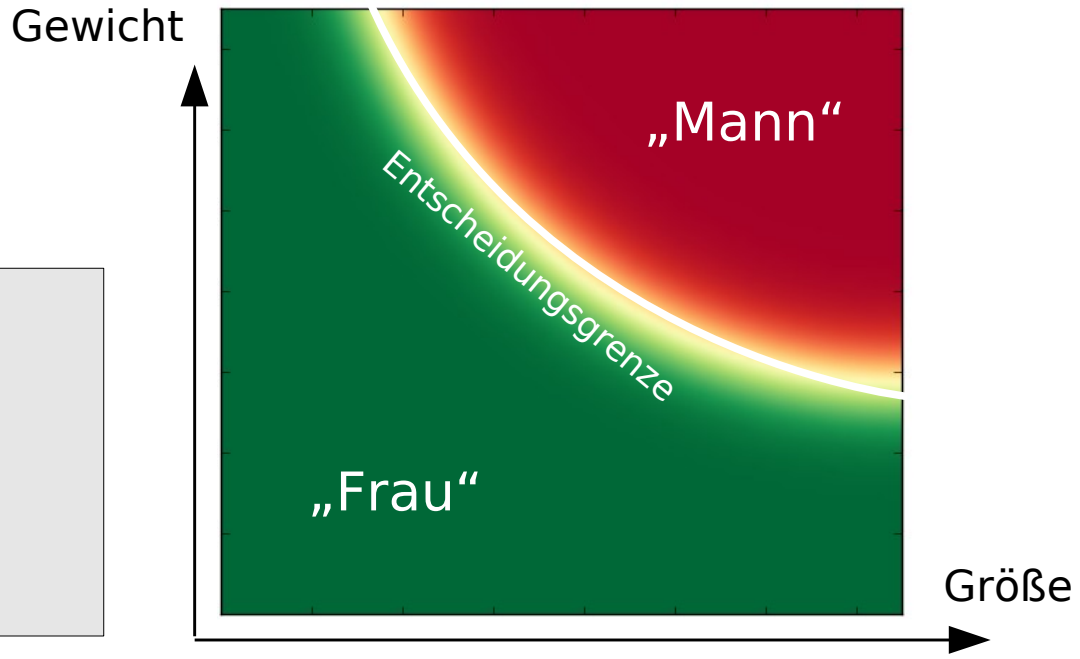
Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

- **Szenario**: ***Klassifikation*** (ordne ein **Sample x** einer **Klasse y** zu)
- **Ziel**: Berechne ***P(y|x)*** für **jede Klasse *y***. Hieraus ergeben sich ***Entscheidungsregionen*** und ***-Grenzen***.

Dummy-Beispiel

x = (Größe,Gewicht)

P(C="Frau"|x) ist niedrig
P(C="Frau"|x) ist hoch



ML: Terminologie

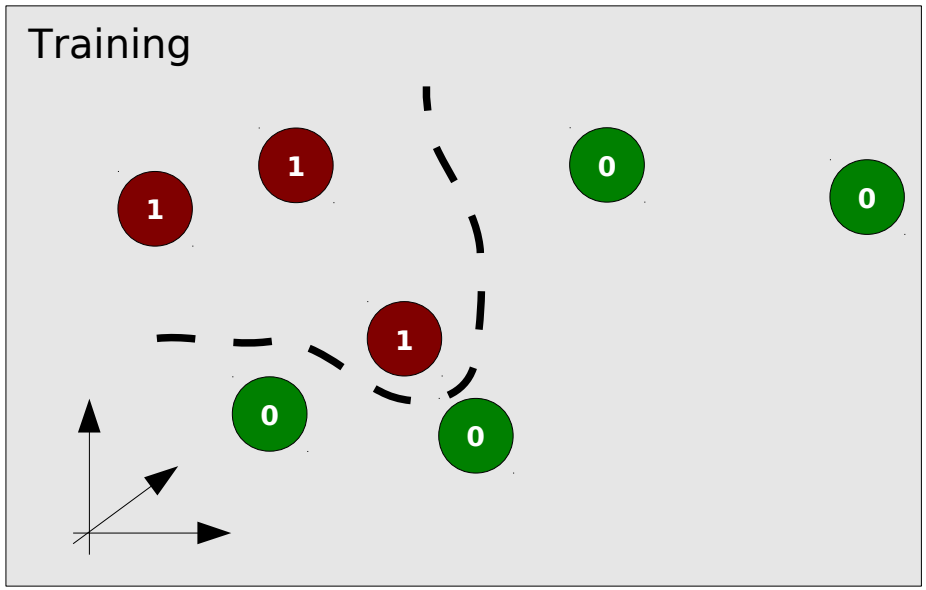
Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

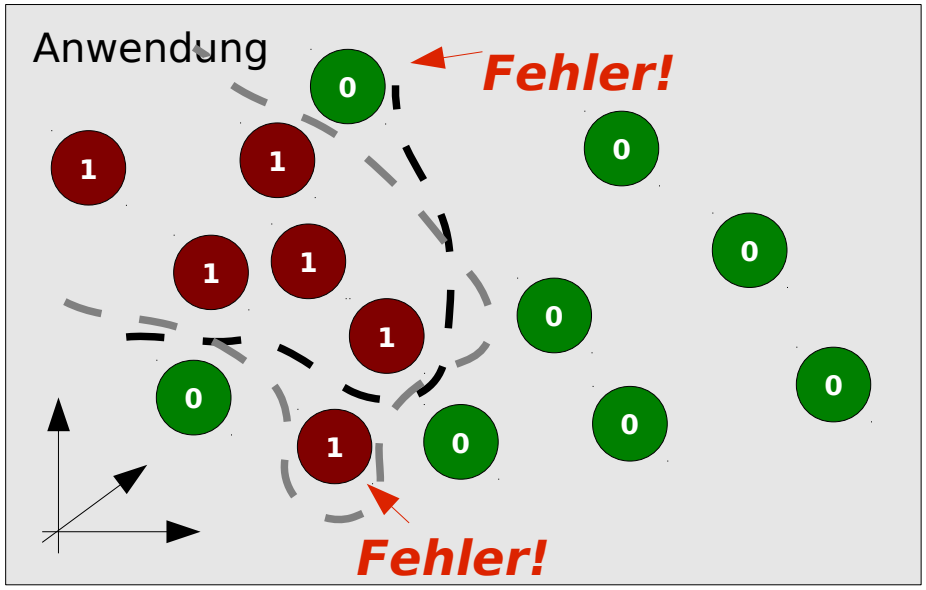
- ***Ansatz***: Wir bestimmen für jede Klasse eine ***Klassifikatorfunktion*** Φ_c^θ**:** $\mathbb{R}^n \rightarrow$ **[0,1]**
- Gegeben ein Objekt **x**, verwenden wir Φ_c^θ**(x)** als ***Approximation*** für **P(c|x)**, um **x** zu klassifizieren: Berechne Φ_c^θ**(x)** für **alle Klassen c** und wähle die Klasse mit ***maximalem Wert***.
- Die genaue Form von Φ_c^θ**(x)** wird von den ***Parametern*** θ bestimmt. Die Parameter werden durch ***Training*** ermittelt
- Hierzu verwenden wir eine **Trainingsmenge**:
  - Beispiele (***Samples***) $\mathbf{x_1, x_2, \ldots, x_n} \in \mathbb{R}^n$
  - Klassenzuordnungen (***Labels***) $\mathbf{y_1, y_2, \ldots, y_n} \in \mathbf{\{1,..,C\}}$

ML: Generalisierung und Overfitting

Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

Die Fähigkeit der eines Klassifikators, von den Trainingsdaten auf (neue) Testdaten zu schließen, bezeichnen wir als ***Generalisierung***.

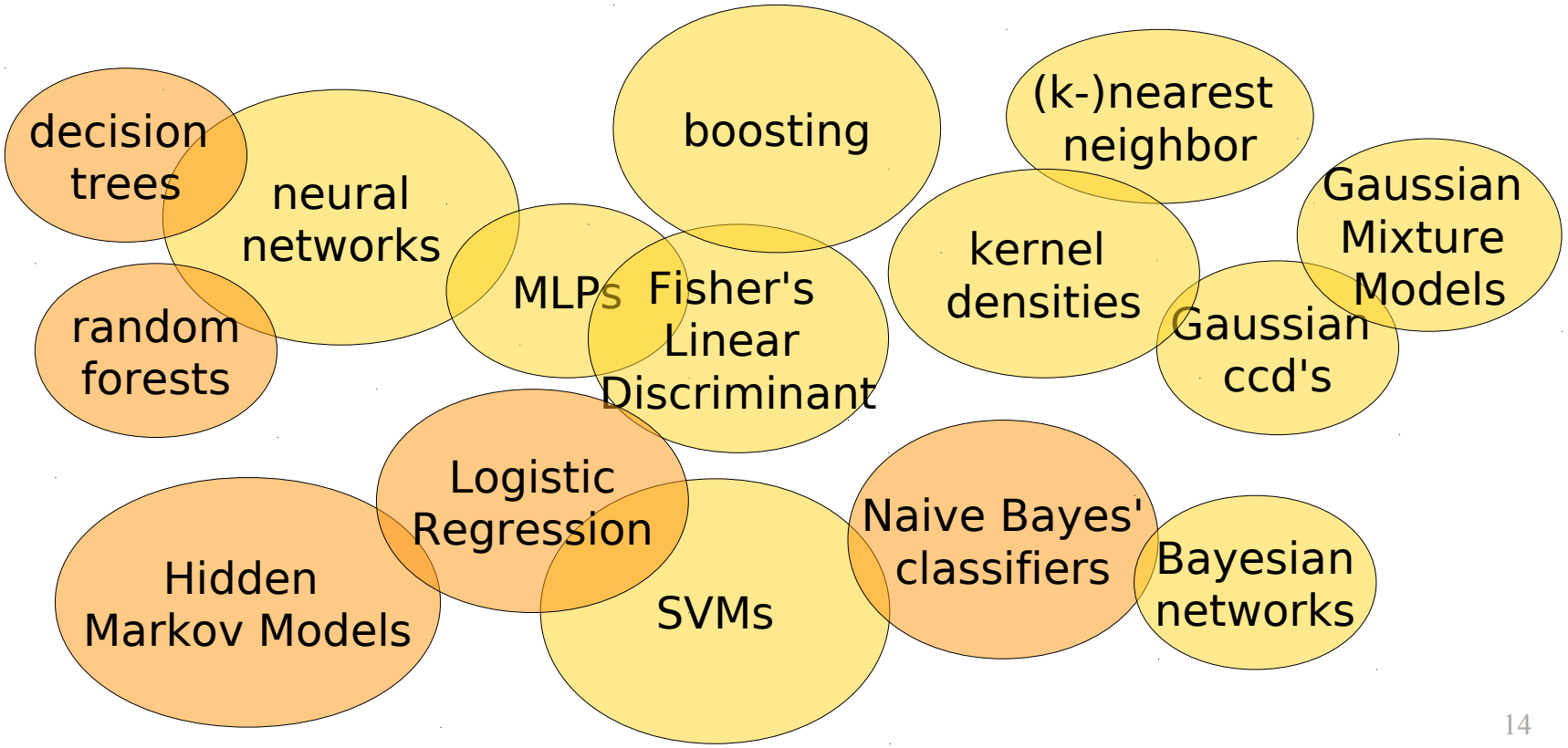




ML: Ansätze

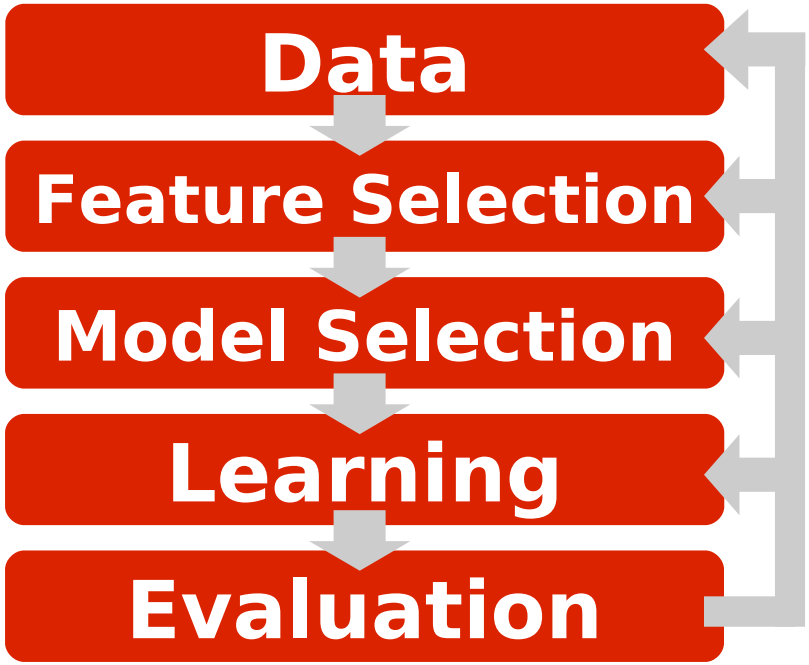
Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

- ***Zahlreiche Methoden*** des maschinellen Lernens beschäftigen sich mit der Approximation von **P(y|x)** ...
- Den ***universell besten Klassifikator*** gibt es nicht
  (→ No-free-lunch-Theorem)



ML: Design Cycle

Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain



- **Design eines ML-Systems** = ***iterative Auswahl*** von Daten / Merkmalen / Modellen / Parametern
- **Schlüsseltreiber**: ***Benchmarking***

ML: Benchmarking

Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

Fehler:
5,3%

- Wir **unterteilen** die Datenmenge in Trainings- und Testdaten
- Wir trainieren auf den ***Trainingsdaten*** und benchmarken auf den ***Testdaten***
- ***Todsünde: „Testing on the Training Data“***

Benchmarking: Gütemaße

Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

		Estimated Emotion							
		Anger	Boredom	Disgust	Fear	Happiness	Sadness	Neutral	*Emotion Recog. Rate*
True Emotion	Anger	**19**	0	2	0	**3**	0	0	79.2%
	Boredom	1	**8**	1	1	0	1	**7**	42.1%
	Disgust	0	1	**6**	0	1	0	3	54.5%
	Fear	1	3	2	**7**	2	0	1	43.8%
	Happiness	**3**	0	3	2	**5**	0	2	33.3%
	Sadness	0	0	0	0	0	**14**	0	100.0%
	Neutral	0	**5**	1	0	0	0	**13**	68.4%
	HMM Recog. Rate	79.2%	47.1%	40.0%	70.0%	45.5%	93.3%	50.0%	

Bsp.: ***Confusion-Matrix*** *(welche Klassen werden besonders häufig verwechselt?)*

Benchmarking: Gütemaße

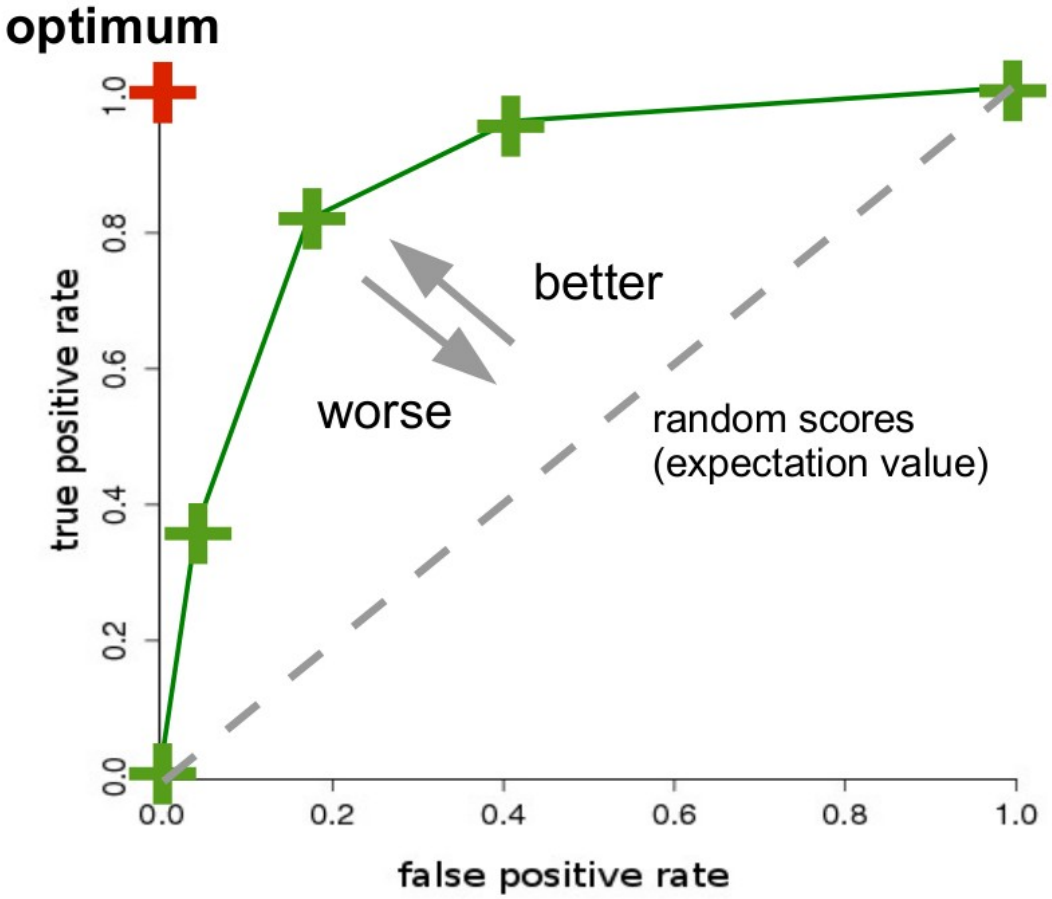
Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

- Oft unterscheiden wir nur zwischen **zwei Klassen** (1 und 0)
- Wir unterscheiden vier Arten von Ergebnissen
  - ***true/false positive***
  - ***true/false negatives***
- Häufig ermitteln Klassifikatoren keine absolute Entscheidung, sondern einen ***Score*** **s(x)**
- Wir wählen eine **Schwelle T** und entscheiden uns für Klasse 1, falls **s(x) >= T**
- **T** bestimmt die ***Sensitivität*** des Klassifikators

Benchmarking: Gütemaße

Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

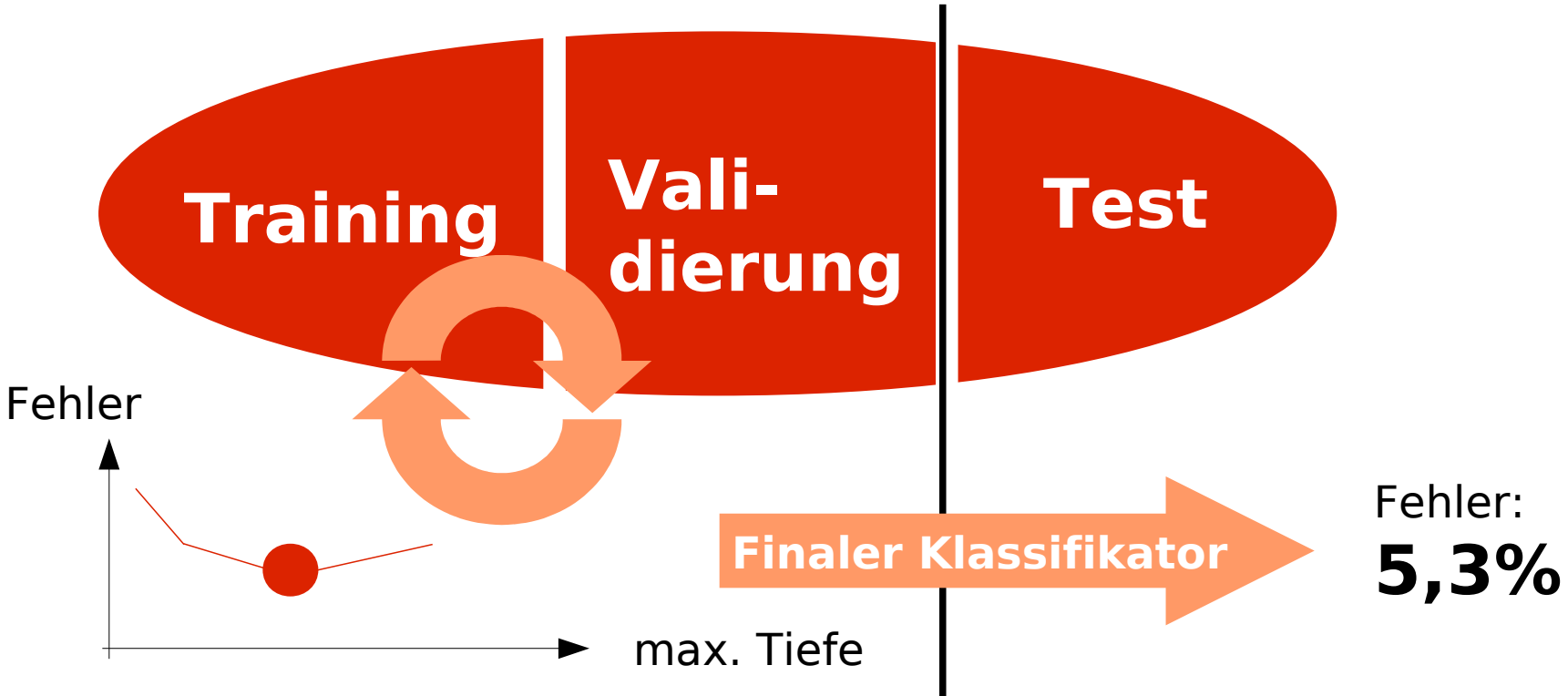
- Wir **variieren T** und erhalten eine Kurve, die sogenannte ***ROC-Kurve***



Validierungsdaten

Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

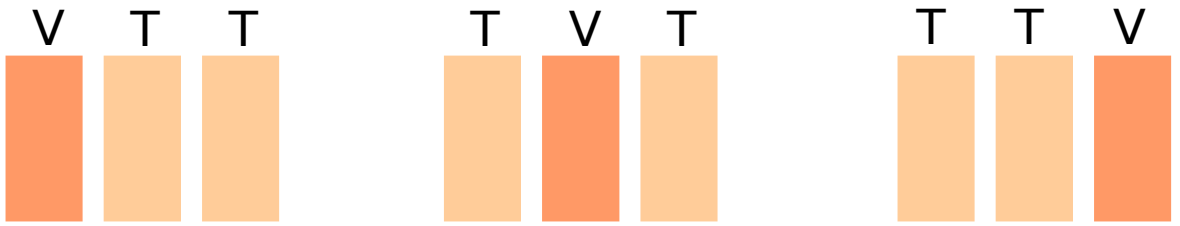
- Einige **Parameter** des Klassifikators werden ***gelernt***, andere (freie) Parameter werden ***manuell*** („trial-and-error“ / grid search) gesetzt
- **Beispiel Entscheidungsbaum**: maximale Tiefe
- **Notwendig**: Training → Validieren → Testen



Cross-Validation

Prof. Adrian Ulges
Fachbereich DCSM / Informatik
Hochschule RheinMain

- Sind die Trainingsdaten klein, teilt man sie in Teile (oder „***Folds***“) und trainiert/validiert ***mehrfach***
- Die Ergebnisse werden über die Folds **gemittelt**

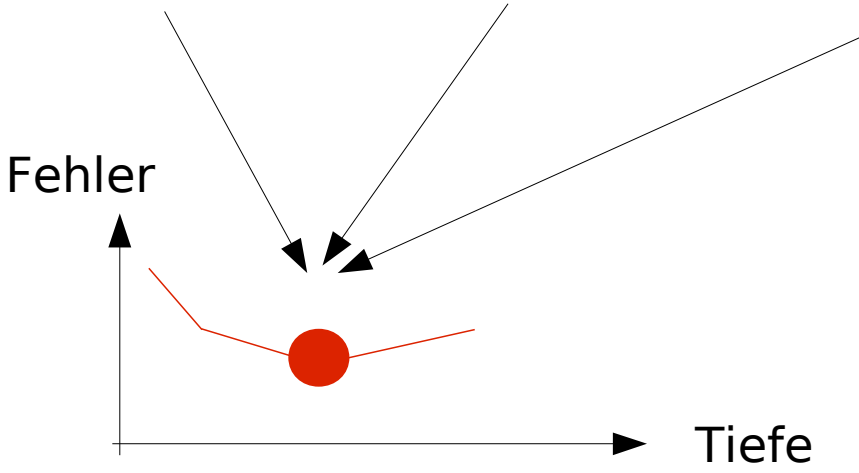


Training + Validierung

Fehler: **5,3%**

Fehler: **6,2%**

Fehler: **4,8%**



Finaler Klassifikator