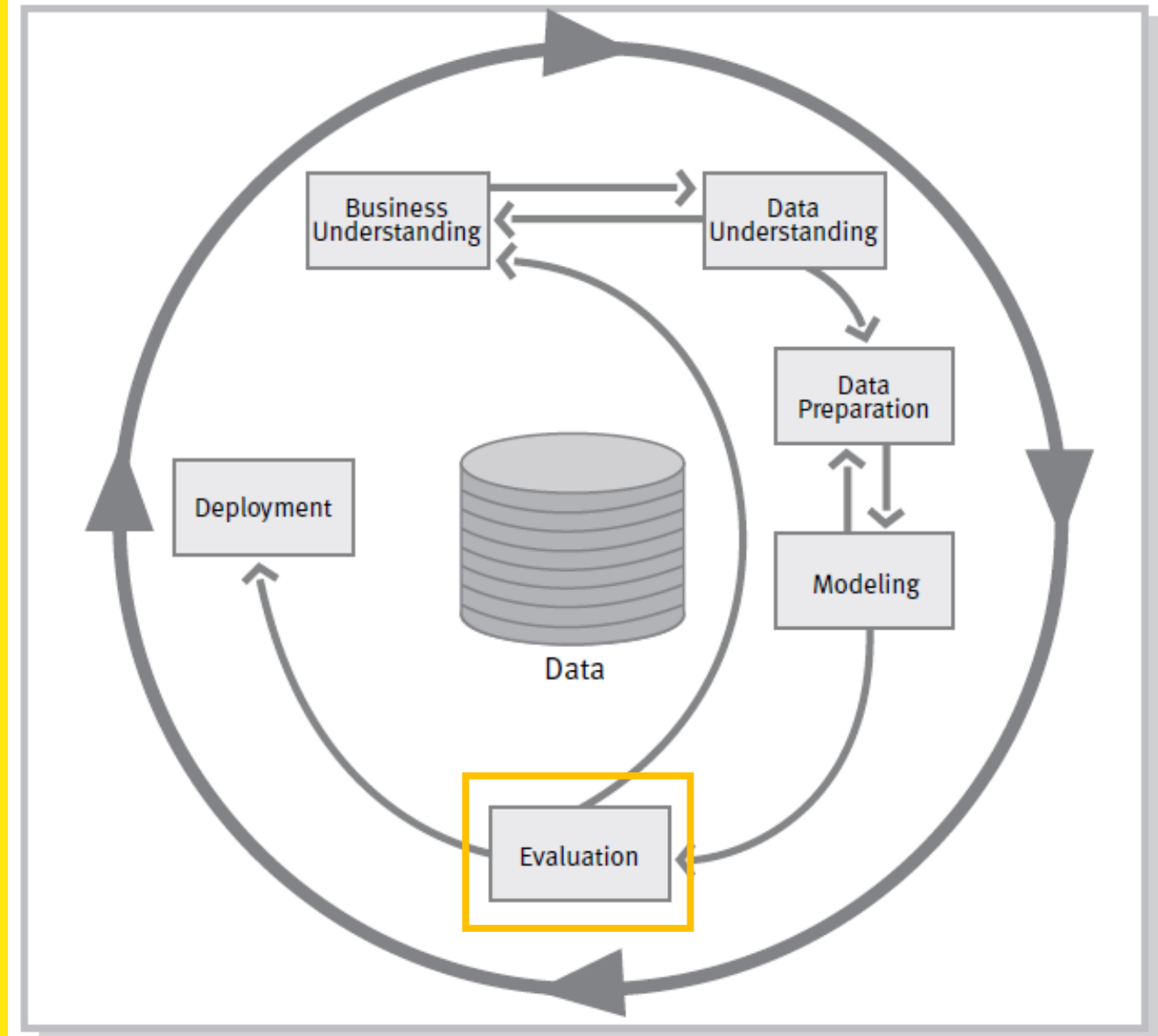


# Maschinelles Lernen – Evaluierung

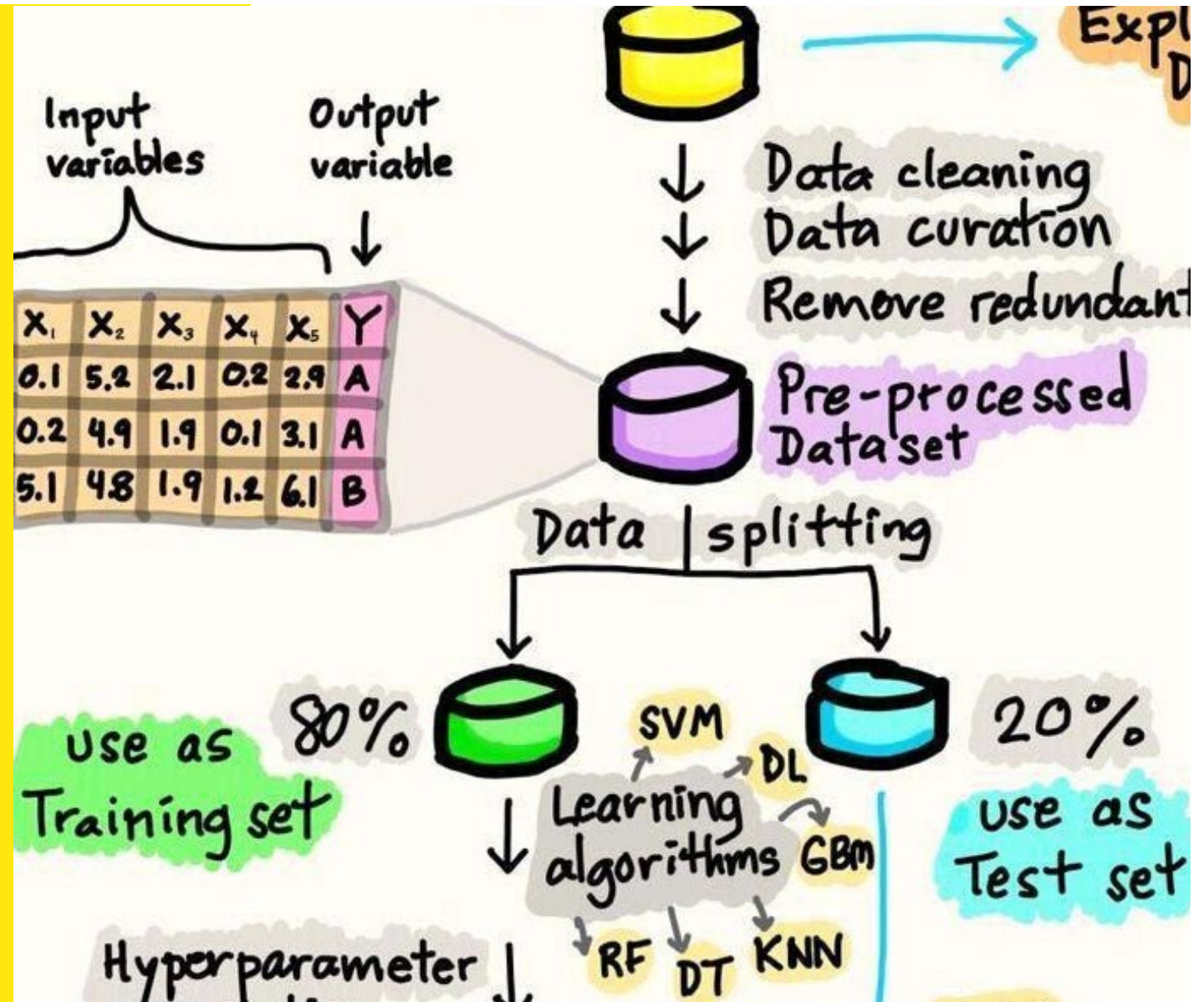
– Hans Friedrich Witschel, Andreas Martin



# Testverfahren und -metriken

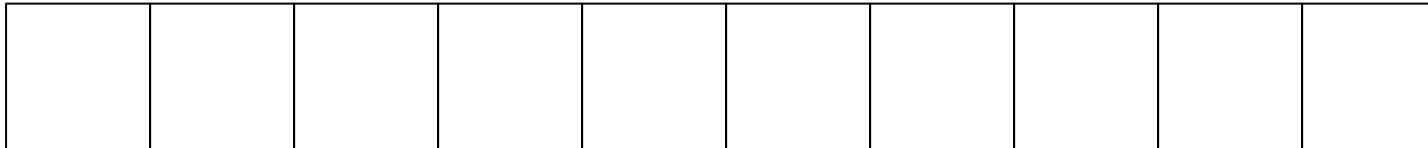
- Verfahren
  - a. Holdout, Kreuzvalidierung
  - b. Lernkurven
  
- Metriken
  - a. Genauigkeit / Fehlerrate
  - b. Precision, Recall, F-measure
  - c. Area Under (ROC) Curve
  - d. Kosten

# Testverfahren



# Holdout und Kreuzvalidierung

- **Grundidee:** Qualität der Vorhersagen auf **unbekannten** Daten evaluieren!
- Typische **Holdout-Evaluierung**: nutze 80% der Daten zum Trainieren, den Rest zum Testen
- **Kreuzvalidierung**, z.B. 10-fach:

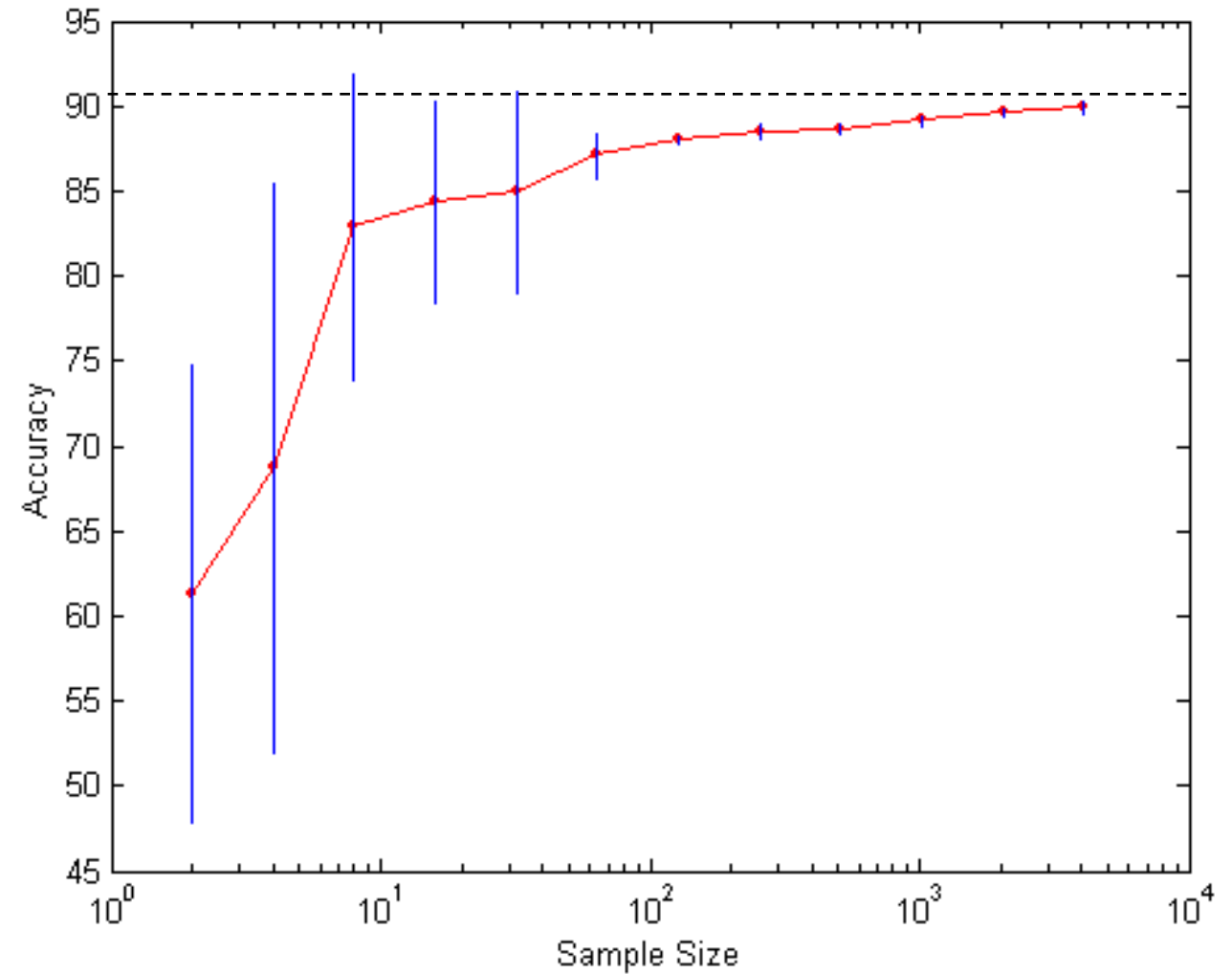


***Training***

***Test***

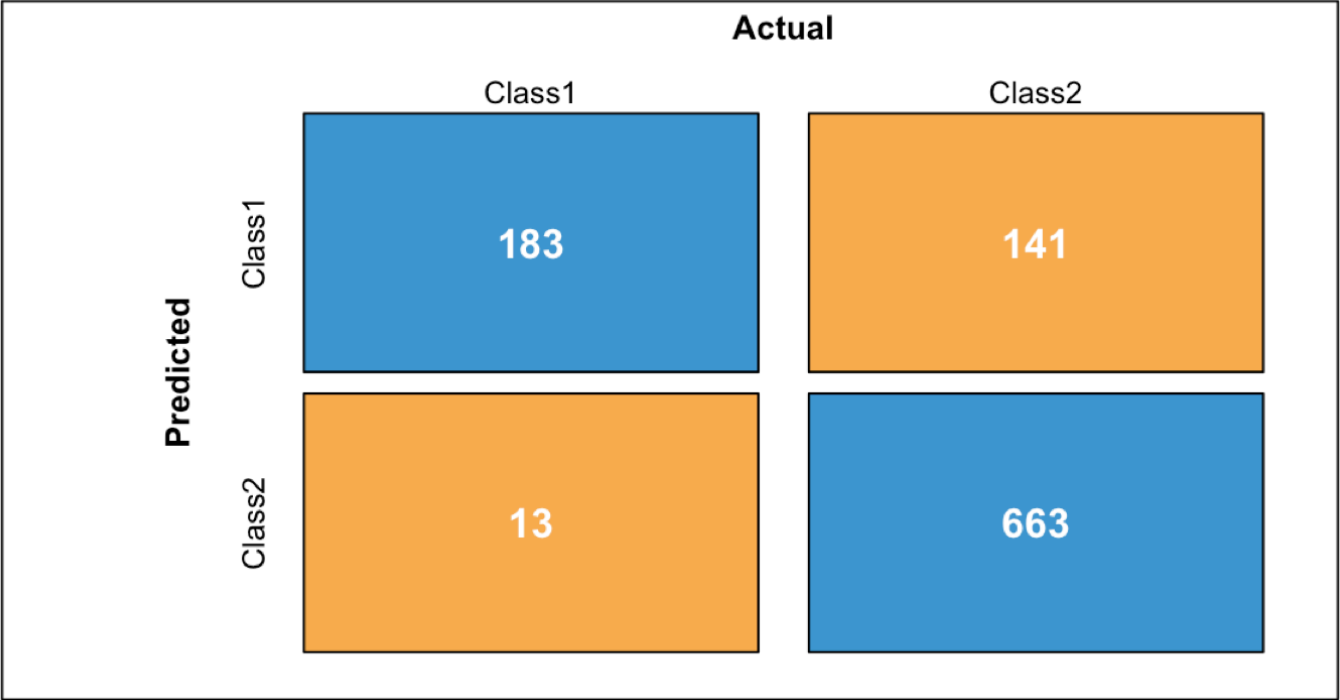
# Lernkurve

- Zeigt, wie die Vorhersagegenauigkeit von der Grösse der Trainingsmenge abhängt
- Wichtig, wenn Trainings- und Testdaten manuell vorklassifiziert werden müssen



# Testmetriken

## CONFUSION MATRIX



### DETAILS

<b>Sensitivity</b> 0.934	<b>Specificity</b> 0.825	<b>Precision</b> 0.565	<b>Recall</b> 0.934	<b>F1</b> 0.704
<b>Accuracy</b> 0.846		<b>Kappa</b> 0.608		

# Genauigkeit bzw. Fehlerrate

- $Genauigkeit = \frac{\text{Anzahl korrekt klassifizierte Instanzen}}{N}$
- $Fehlerrate = \frac{\text{Anzahl inkorrekt klassifizierte Instanzen}}{N} = 1 - Genauigkeit$

Dabei ist N jeweils die Grösse der Testmenge

# Genauigkeit: immer eine gute Metrik?

- Wir betrachten eine Spendenkampagne (Daten des KDD-Cup 1998). Personen werden durch folgende Attribute beschrieben:
  - a. Hobbys
  - b. Anzahl der Fälle, in denen sie in der Vergangenheit auf bestimmte Arten von Angeboten geantwortet haben
  - c. Geschlecht

-----		
	The following variables indicate the number of known times the donor has responded to other types of mail order offers.	The following variables reflect donor interests, as collected from third-party data sources
MBCRAFT	Buy Craft Hobby	COLLECTABLE (Y/N)
MBGARDEN	Buy Gardening	VETERANS (Y/N)
MBBOOKS	Buy Books	BIBLE READING (Y/N)
MBCOLECT	Buy Collectables	SHOP BY CATALOG (Y/N)
MAGFAML	Buy General Family Mags	WORK FROM HOME (Y/N)
MAGFEM	Buy Female Mags	HOUSEHOLD PETS (Y/N)
MAGMALE	Buy Sports Mags	CD PLAYER OWNERS (Y/N)
MBGARDEN	Gardening Clubs	STEREO/RECORDS/TAPES/CD (Y/N)
	PHOTO	HOME PC OWNERS/USERS
		PHOTOGRAPHY (Y/N)

- Wir sagen erst das Geschlecht voraus, dann das Interesse an Male Magazines
  - a. Warum ist die Genauigkeit so unterschiedlich? Wie gut sind die Klassifikatoren wirklich?



# Confusion Matrix, Precision, Recall und F-Measure

$$\text{Precision (p)} = \frac{a}{a + c}$$

$$\text{Recall (r)} = \frac{a}{a + b}$$

$$\text{F - measure (F)} = \frac{2rp}{r + p} = \frac{2a}{2a + b + c}$$

	VORHERGESAGTE KLASSE		
ECHTE KLASSE		Ja	Nein
	Ja	a	b
	Nein	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

# Precision + Recall: in Worten

- **Precision:** wieviele der als «Ja» vorhergesagten Instanzen sind wirklich welche?  
*z.B. Fraud Detection, Precision = 0.84 → 84% der vorhergesagten Betrugsfälle («Warnungen») sind wirklich betrügerisch*
- **Recall:** wieviele der echten «Ja»-Fälle wurden gefunden?  
*z.B. Fraud Detection, Recall = 0.28 → 28% der echten Betrugsfälle wurden gefunden*

# Area Under (ROC) Curve

- **Grundidee:**

- a. Klassifikatoren geben (fast) immer einen Score (Wahrscheinlichkeit oder Konfidenz) aus, nach dem man die Instanzen sortieren kann
- b. AUC misst, inwieweit die positiven Beispiele nach oben sortiert werden
- c. Formal:  $AUC = \text{Wahrscheinlichkeit, dass ein zufällig gewähltes positives Beispiel höher als ein zufällig gewähltes negatives Beispiel gerankt wird}$

- **Wertebereich:**

- a. Zufällige Sortierung:  $AUC = 0.5$
- b. Optimale Sortierung (erst alle positiven, dann alle negativen):  $AUC = 1$

## Area Under (ROC) Curve: Berechnung

- Lasse den Klassifikator für jede Instanz A die Wahrscheinlichkeit ausgeben, dass A positiv ist (z.B. Interesse am Wintercheck hat)
- Gehe die sortierte Liste durch, nimm jeweils an, dass die aktuelle Position der Schwellenwert für die Klassifizierung ist und berechne
  - a. True Positive Rate / TPR = Recall =  $TP/(TP+FN)$
  - b. False Positive Rate / FPR =  $FP/(FP + TN)$

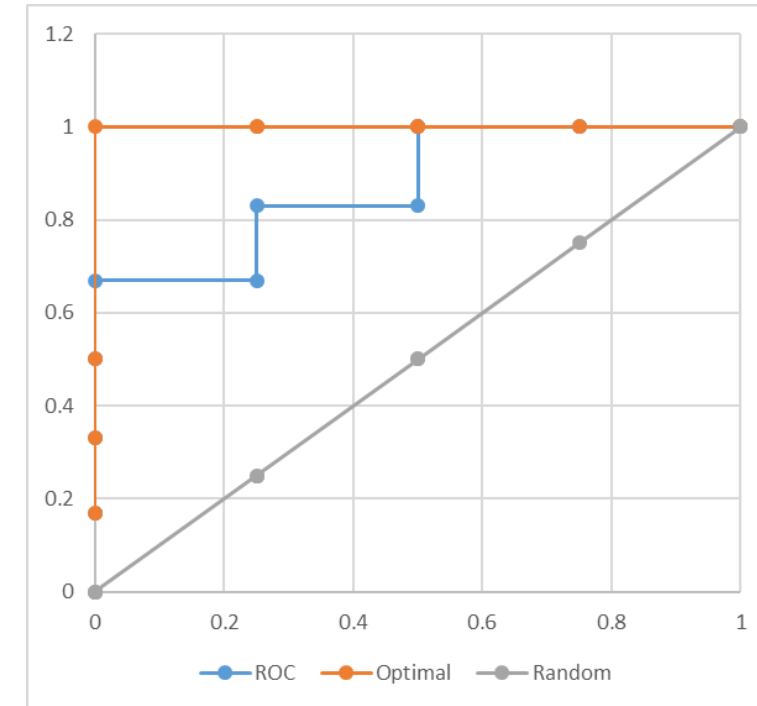
Instanz A	P(+ A)	Echte Klasse
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

# Area Under (ROC) Curve: Beispiel

Instanz	P(+ A)	Echte Klasse	TPR	FPR
1	0.95	+	$1/6 = 0.17$	0
2	0.93	+	$2/6 = 0.33$	0
3	0.87	+	$3/6 = 0.5$	0
4	0.85	+	$4/6 = 0.67$	0
5	0.84	-	$4/6 = 0.67$	$1/4 = 0.25$
6	0.83	+	$5/6 = 0.83$	$1/4 = 0.25$
7	0.76	-	$5/6 = 0.83$	$2/4 = 0.5$
8	0.53	+	1	$2/4 = 0.5$
9	0.43	-	1	$3/4 = 0.75$
10	0.25	-	1	1

Wieviele Interessenten wurden bis  
hierher gefunden?

Wieviele «falsche Alarmer» wurden  
bis hierher generiert?



# Kostenbasierte Evaluation: Idee

- Oft wird ML eingesetzt, um Geld zu verdienen oder zu sparen. Daher resultieren aus Klassifikator-Vorhersagen (unterschiedliche) Gewinne oder Verluste.
  - Beispiel Swiss Bikes:
    - a. Jeder Wintercheck erzielt einen Umsatz von 100 CHF und einen Gewinn von 40 CHF für Swiss Bikes
    - b. Versand eines Werbebriefs kostet 1 CHF
- Kostenmatrix

«Fehlerrate»

	Predicted class	
	Yes	No
“True” class	0	1
	1	0

➔

«Kosten»

	Predicted class	
	Yes	No
“true” class	-39	0
	1	0

# Berechnung der Kosten

- Kostenmatrix

	Predicted class		
“true” class		Yes	No
	Yes	-39	0
	No	1	0

- Confusion matrix auf Testmenge:

	Predicted class		
“true” class		Yes	No
	Yes	12	20
	No	67	101

Welche Kosten ergeben sich aus den Vorhersagen des Klassifikators auf der Testmenge?

# Übung: Kostenmatrizen formulieren

Wie sieht die Kostenmatrix in den folgenden Fällen aus?

- erste Hälfte der Gruppen: Kreditvergabe in einer Bank
- zweite Hälfte der Gruppen: Aufdeckung von Versicherungsbetrug in einer Unfallversicherung

Macht nötige Annahmen über die jeweils relevanten Kosten!