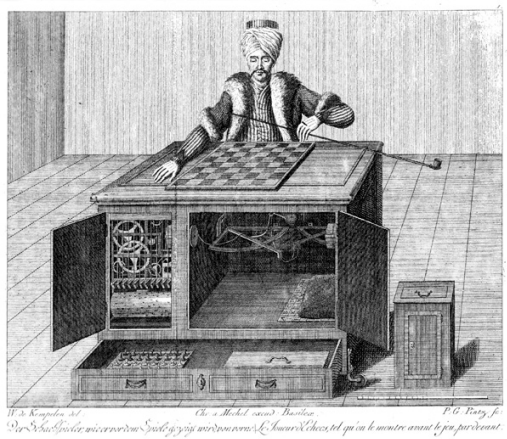
Evaluierung von Machine-Learning-Methoden

Unterschiedliche Aspekte der Evaluierung

Agenda – What You Will Learn…

- Typical Use Cases for Machine Learning
- Basic Concepts
- Evaluating Learning to Rank
- Evaluating Classification Algorithms
- Assess the Validity of Machine Learning Models
- Conclusion

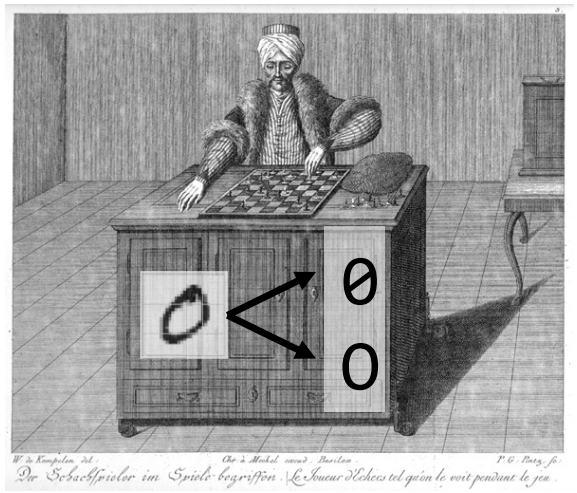


Machine Learning?

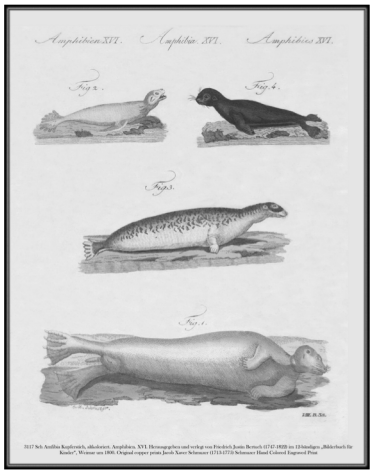
- Lots of data is available
- We expect some kind of pattern hidden in the data
- We don‘t know a logical/mathematical expression to solve our problem

What are typical Machine Learning Problems?

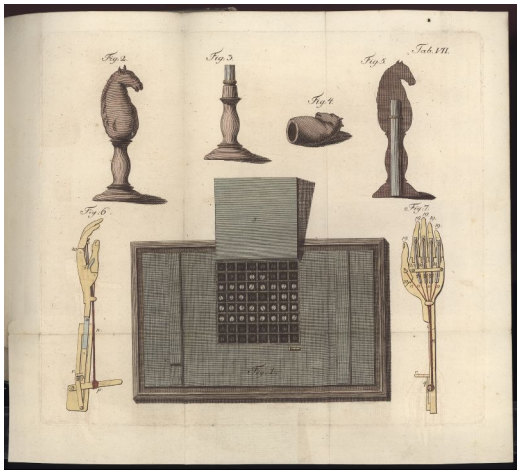
Classification



Clustering



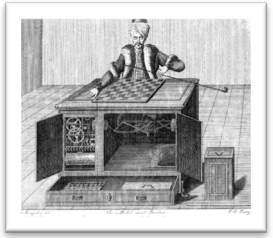
Regression

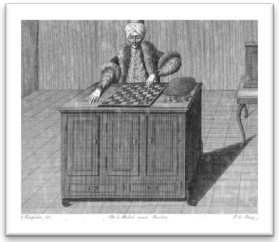


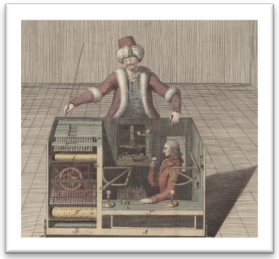
Learning to Rank

Classes of Machine Learning Algorithms

- Reinforcement Learning
  - Learning through reward
- Unsupervised Learning
  - Clustering is derived from data set
- Supervised Learning
  - Based on a ground or some form of expert knowledge



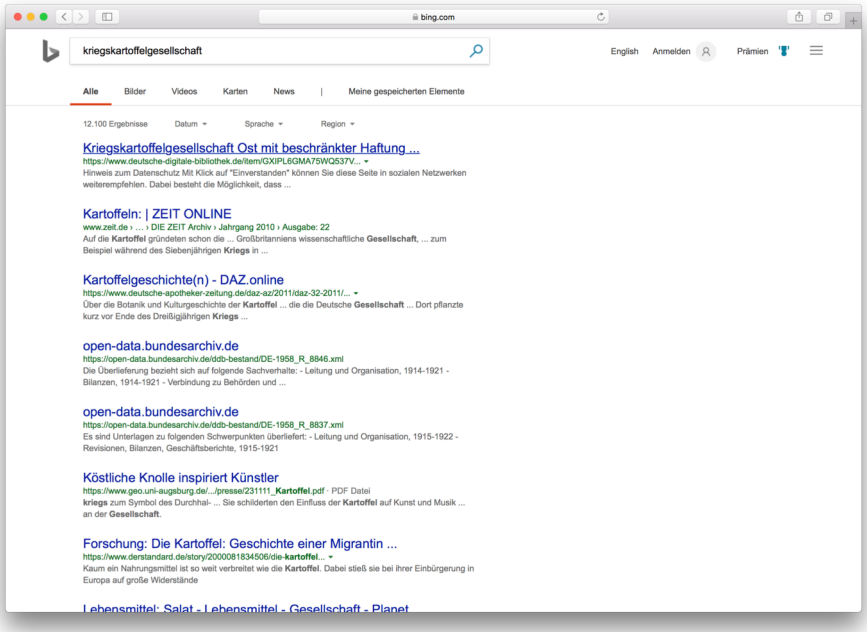


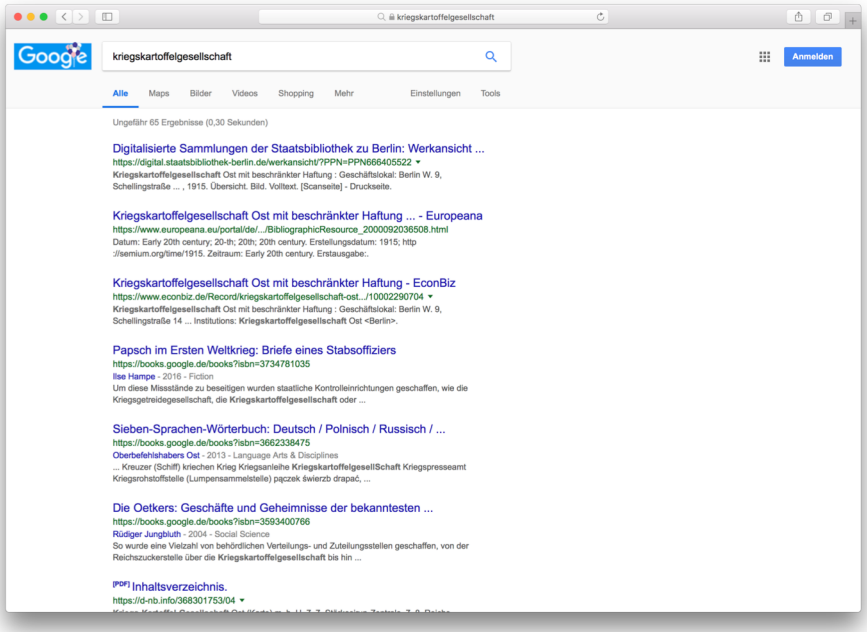


Problems from Daily Work

Evaluation of Two Problem Classes

Which Results are Better?

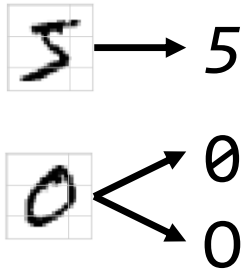




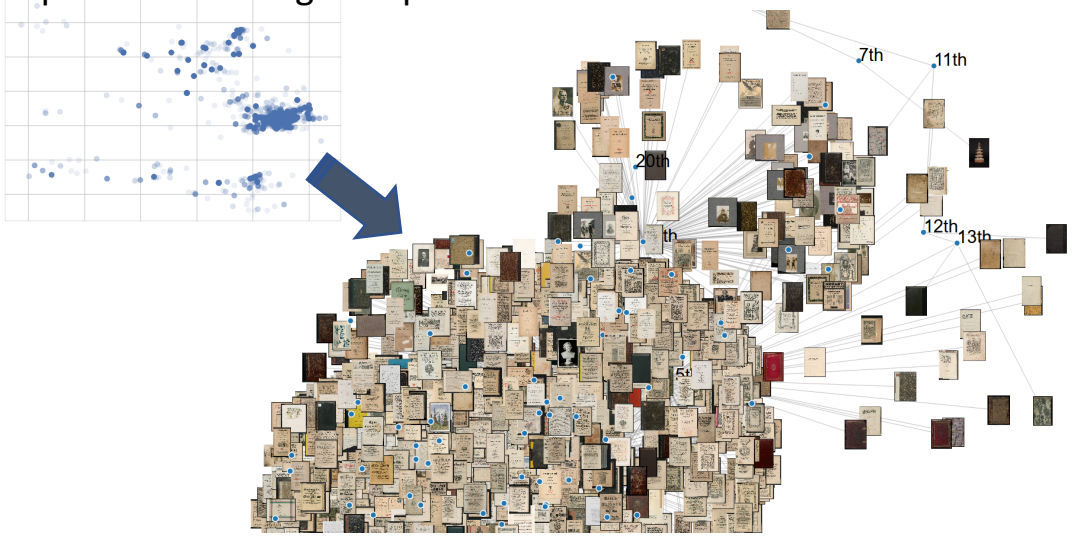
Classification

By Qwertyxp2000 [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], from Wikimedia Commons

Optical Character Recognition (OCR)



Exploration of Large Corpora



D. Zellhöfer: Exploring Large Digital Libraries by Multimodal Criteria; TPDL 2016

Copyright Detection

https://cdn.netzpolitik.org/wp-upload/contentID.png

Basic Concepts

How to Evaluate a System's Answer?

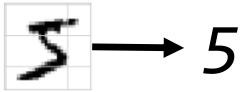
- Relevance – does an object belong to the answer, class etc.?
- Typically a binary scale: irrelevant and relevant
- Driving research since the early 60s [Cleverdon 1962; Lesk & Salton 1968]

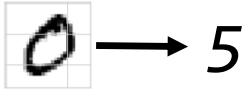
- Commonly interpreted as a probability of relevance [Robertson 1977]

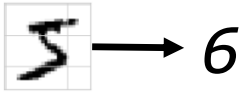
$$P(R|q,d) \approx probability(\text{relevance}|\text{given a query } q \text{ and a document } d)$$

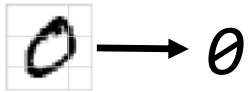
Possible Answers of a System

Is a hand-written digit a five, i.e., does it belong to class 5?

- True positive (tp) 

- False positive (fp) 

- False negative (fn) 

- True negative (tn) 

Precision, Recall, and F-Score

$$Precision = \frac{|tp|}{|tp| + |fp|} \equiv \frac{|\text{relevant elements in result}|}{|\text{elements in result}|}$$

$$Recall = \frac{|tp|}{|tp| + |fn|} \equiv \frac{|\text{relevant elements in result}|}{|\text{relevant elements in collection}|}$$

$$F - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Result Order: 1;2;3;4;
Precision: 0.75
Recall: 0.60
F-score: 0.67

Result Order: 1;2;3;4;6;7;8;
Precision: 0.43
Recall: 0.60
F-score: 0.50

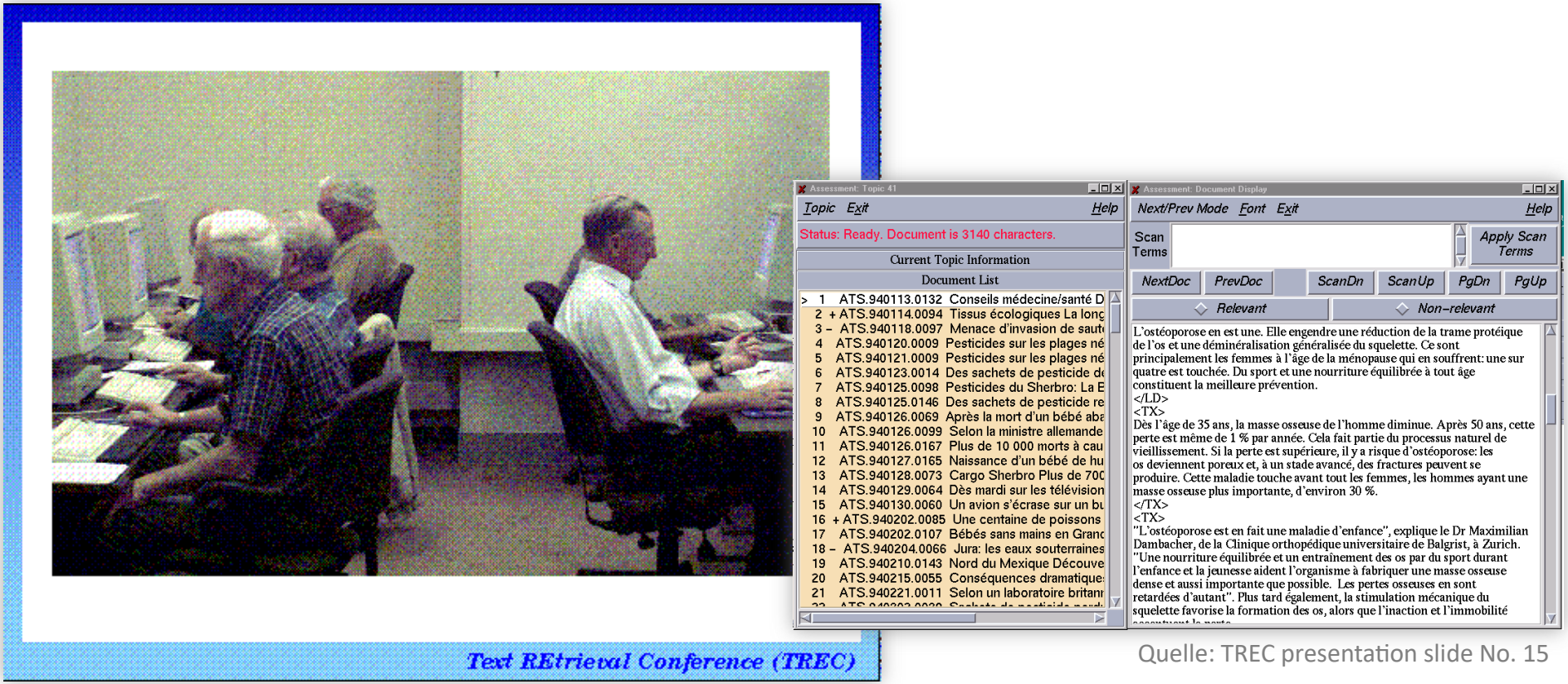
Determine Relevant Elements in a Collection

$$Precision = \frac{|tp|}{|tp| + |fp|} \equiv \frac{|\text{relevant elements in result}|}{|\text{elements in result}|} \qquad Recall = \frac{|tp|}{|tp| + |fn|} \equiv \frac{|\text{relevant elements in result}|}{|\text{relevant elements in collection}|}$$

- Relevance assessments for all documents (the internet?)
- Relevance assessment on a representative random sample

- A lot of time and effort…

TREC Evaluation

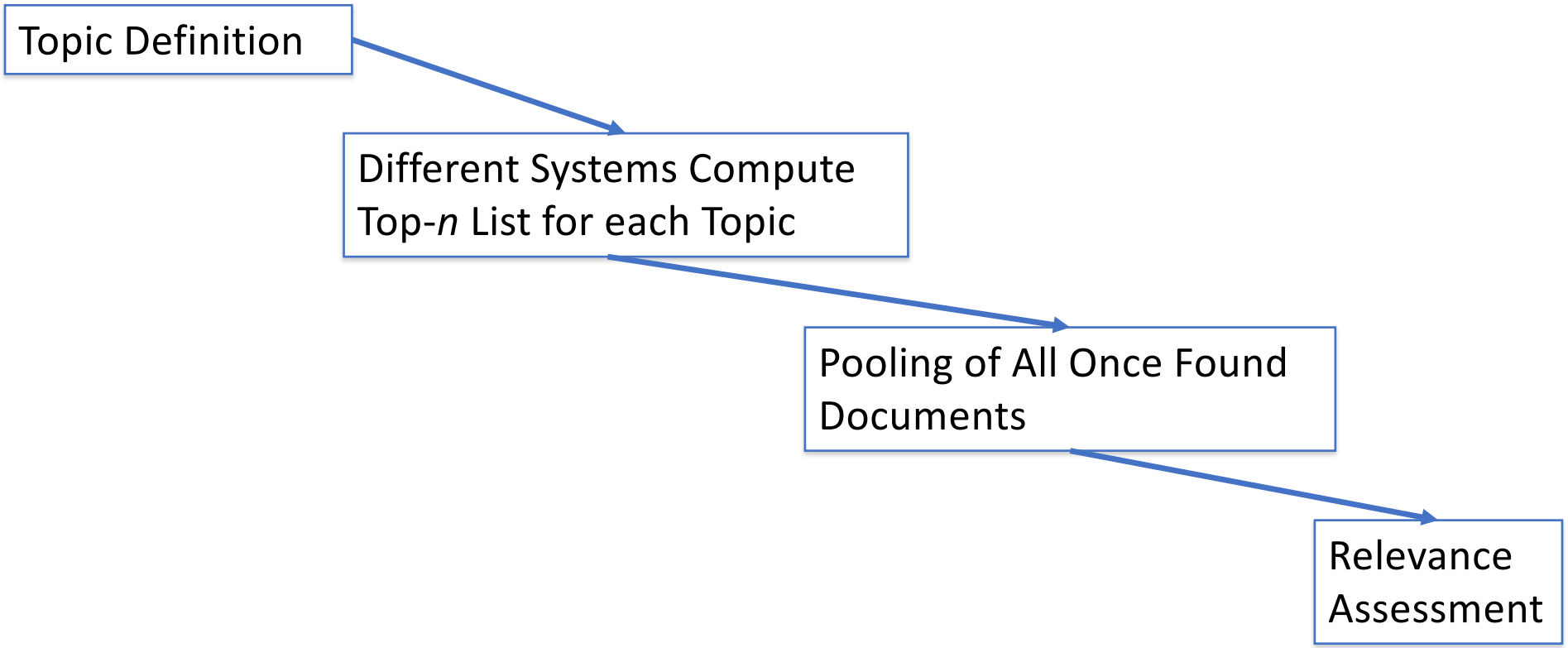


Quelle: TREC presentation slide No. 15

Manual Assessments

- Cost- and time-expensive even with services such as Amazon Mechanical Turk
- Complicated for expert tasks
- But still done by Bing, Google, Staatsbibliothek zu Berlin etc.
  https://www.quora.com/How-does-Google-measure-the-quality-of-their-search-results
- Based on the Cranfield paradigm as represented by conferences such as TREC, CLEF, ImageCLEF…

The Pooling Method



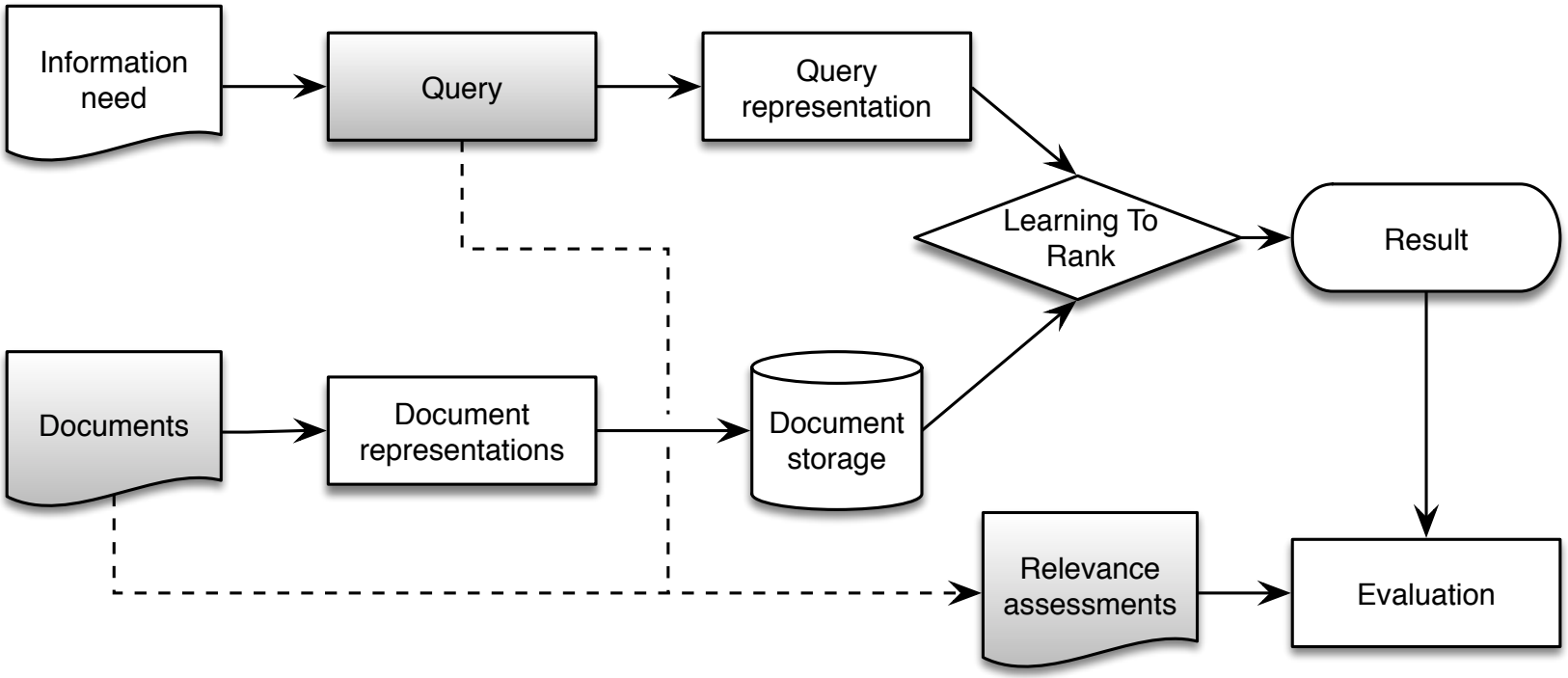
A Brief Critique of Traditional Manual Assessments

- Do not integrate user experience
- Provide few information about assessors, e.g., demographics

- However, there are re-occurring events focusing on user interaction such as TREC interactive track (1997-2002) [Voorhees & Harman 2005], or
- ImageCLEF (2012/13) [Zellhöfer 2013]
  - Provides query and browsed documents derived from different user personas
  - Demographics and characteristics of assessors are published
- Raises cost even more but allows holistic user simulation approach…

Evaluating Learning to Rank

Cranfield-Typed Evaluation of Effectiveness



More User-Oriented Metrics

- Precision and recall are set-based, but learning to rank is ordered
- Precision at cut-off level (top-*k*=5, 10, 20…)

$$Precision\ @\ n = \frac{r^+}{n}$$

$$AP_i = \frac{1}{|R_{q_i}^+|} \sum_{n=1}^{|R_{q_i}^+|} Precision(R[n]) \text{ for one query } q$$

$$MAP = \frac{1}{|Q|} \sum_{i=1}^{|Q|} AP_i \text{ for } q_i \in Q \text{ with } Q \text{ being the set of all considered queries.}$$

- Rank correlation coefficients (Spearman's ρ (rho), Kendall's τ (tau))
  → Vorlesung Statistik

nDCG – A More Advanced Metric

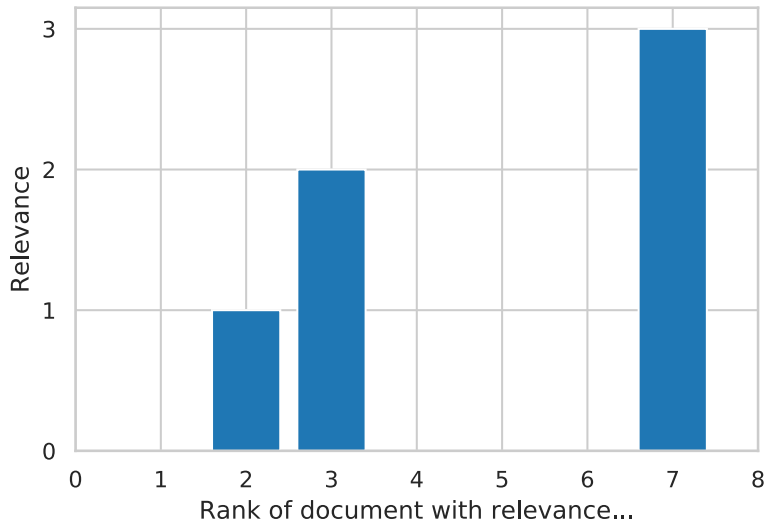
- Normalized discounted cumulative gain [Järvelin & Kekäläinen 2002]
- Relevance is assumed to be graded [0,3]
- Penalizes relevant documents retrieved lately

$$CG[i] = \begin{cases} G[1], & \text{if } i = 1 \\ CG[i-1] + G[i], & \text{otherwise.} \end{cases}$$

$$DCG[i] = \begin{cases} CG[i], & \text{if } i < b \\ DCG[i-1] + \frac{G[i]}{\log_b i}, & \text{if } i \geq b. \end{cases}$$

$$nDCG[i] = \frac{DCG[i]}{iGV[i]} \text{ with } iGV \text{ being the optimal rank (documents ordered by decreasing relevance}$$

nDCG – A More Advanced Metric



CG for rank with 7 results:
6.00

DCG for rank with 7 results:
4.16

nDCG for rank with 7 results:
0.43

ERR – Even More Sophisticated?

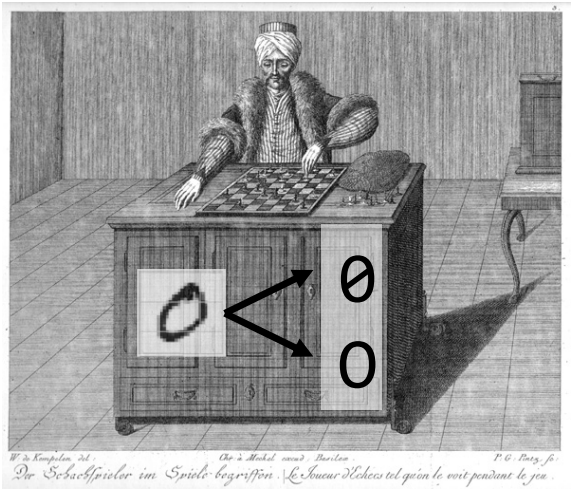
- *Expected Reciprocal Rank* for Graded Relevance as suggested by Yahoo and Google engineers http://olivier.chapelle.cc/pub/err.pdf; Yandex suggested an alternative
- Deals with the problem of nDCG that it assumes a user assesses relevance independently from its position in the list, no matter which documents appeared before (*rank-based user model*)
  → assumption has been proven invalid [Craswell et al. 2008; Chapelle & Zhang 2009]

- Core idea: a user is more likely to stop the inspection of search results after a highly relevant document has been found than after a less relevant document has been seen (*cascade user model*)

- ERR has been shown to correlate better with click metrics

Evaluating Classification Algorithms

The Classification Problem



$\mu_{C_i} : o \rightarrow [0,1] \quad | o \in Objects;\ C_i \in Classes$

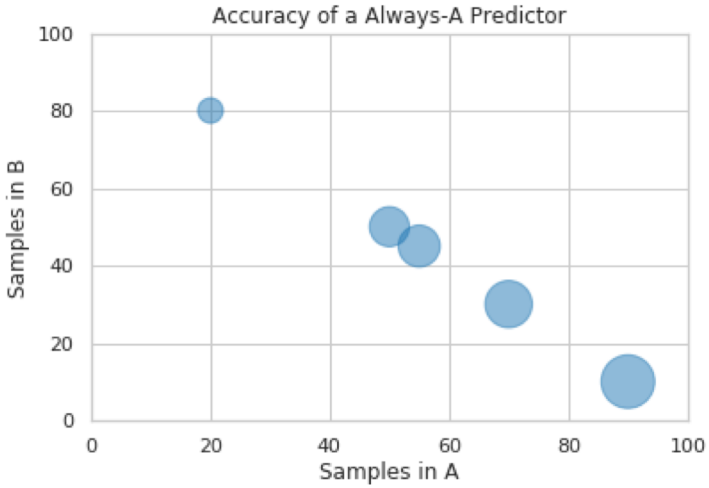
- Membership function μ_{C_i} will be derived inductively from the training data
- Membership value is interpreted a a probability
- Requirement: membership function will generalize for any new type of input data

Accuracy and Error Rate

$$Accuracy = \frac{|tp + tn|}{|tp| + |tn| + |fp| + |fn|} \equiv \frac{|\text{correct predictions}|}{|\text{predictions}|}$$

$$Error\ Rate = \frac{|fp + fn|}{|tp| + |tn| + |fp| + |fn|} \equiv \frac{|\text{wrong predictions}|}{|\text{predictions}|}$$

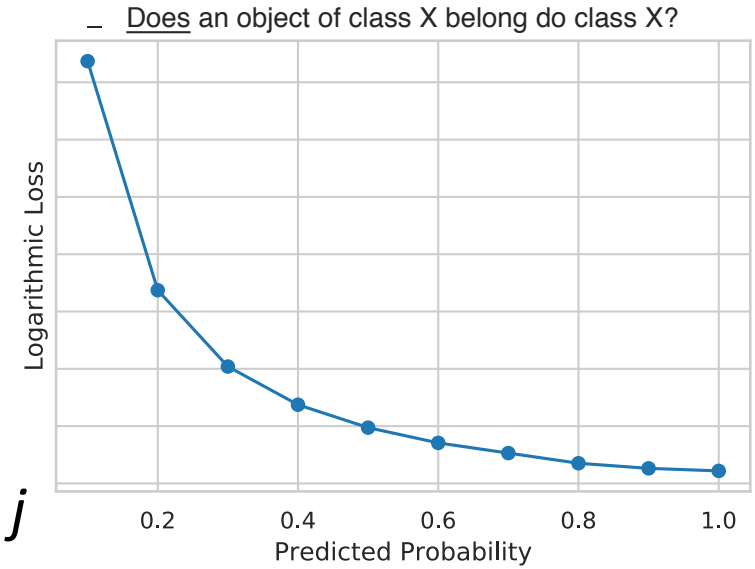
- Assumes equal number of objects per class
- False friend in case of uneven distributions



Logarithmic Loss

$$Logarithmic\ Loss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}) \rightarrow [0, \infty)$$

- *N*: number of objects
- *M*: number of classes
- y_{ij}: object *i* belongs to class *j*
- p_{ij}: predicted probability that *i* belongs to *j*

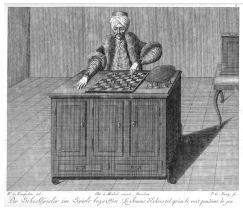


- Penalizes wrong predictions, esp. the ones with high confidence
- Values towards 0 indicate higher general accuracy

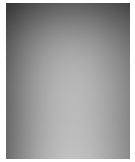
Validity of a Machine Learning Model

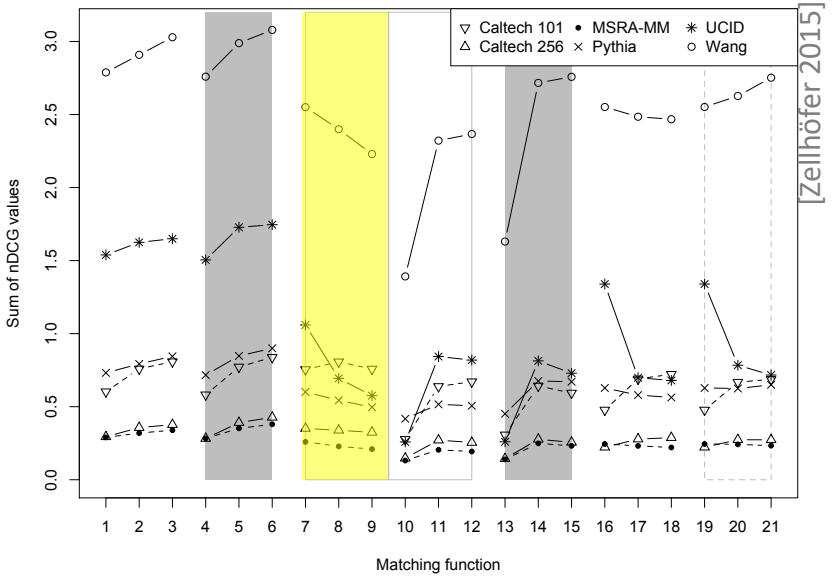
Underfitting

- The model does not work neither with training nor test data
- Easily detectable by using appropriate metrics



?



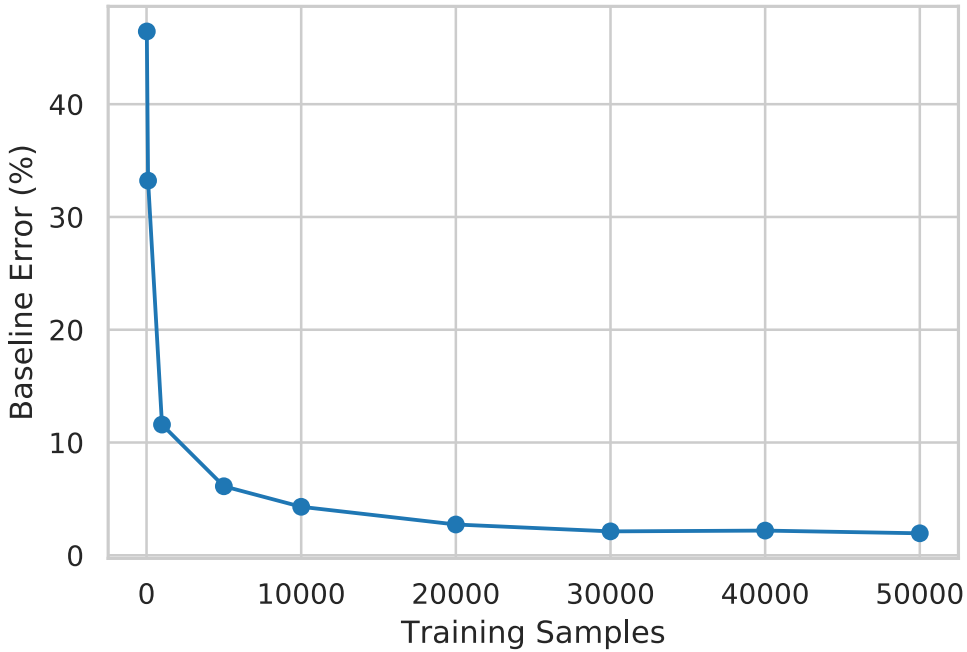


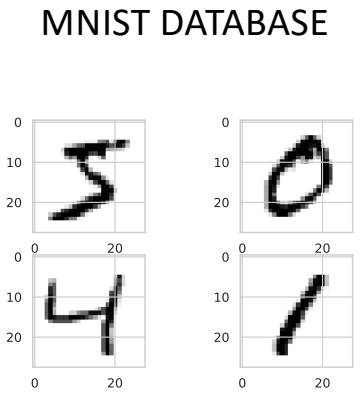
[Zellhöfer 2015]

$$\bigwedge_{\theta_i}(Color\ Layout, Dominant\ Color, Edge\ Histogram, Tamura)$$

Overfitting

- The model does not generalize from training data





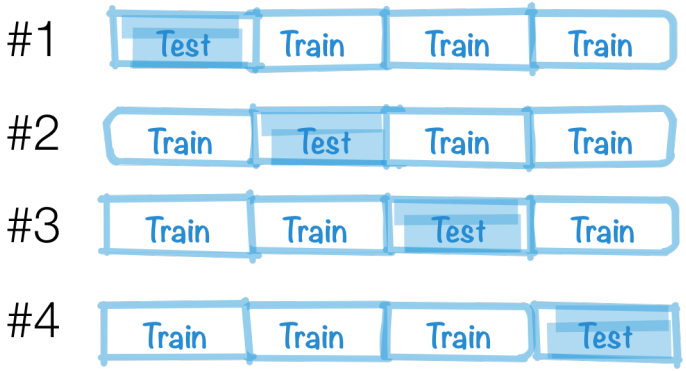
Limiting Overfitting

- Rule of Thumb: 70% for training, 30% for testing*
- Keeping back data for validation limits training data
- Particularly bad if only a small amount of data is available

Limiting Overfitting

- Rule of Thumb: 70% for training, 30% for testing*
- Keeping back data for validation limits training data
- Particularly bad if only a small amount of data is available

- Cross-validation tries to compensate this problem with different strategies and serves as an estimator of the model's validity



Conclusion

Conclusion

- Metrics evaluate only the effectiveness of algorithms

- When obtaining a new data set, lock up validation data elsewhere to avoid bias (overfitting)
- Cross-validation is a good means against overfitting

- Consider financial/practical implications on how to obtain test data
- Consider ethical and governance implications regarding the usage of services such as Mechanical Turk because of low wages, data security etc.

Conclusion

- Using machine learning can also be based on a cost-benefit analysis… in other words: efficiency

Table 1: Comparison of runtime and retrieval effectiveness of the two winning groups at the ImageCLEF 2013 Personal Photo Retrieval subtask (Percentages are given with respect to the first placed ISI_1 run.)

System Characteristic	**ISI_1**	**DBIS_run3**
Type of MIR system	L2R, RF-supported	Logic-based, RF-supported
Number of features	10 (4 visual, 6 Exif)	18 (15 visual, 3 Exif)
CPU cores and model	24 cores, 4 x Intel Xeon X5675 3.07 GHz	8 cores, 2 x Intel Xeon E5520 2.26 GHz
CPU launch date	Q1'2011	Q1'2009
Runtime per topic	ca. 10-15 min.	ca. 1 min.
Metric	**ISI_1**	**DBIS_run3**
MAP, cut-off @ 100	0.5028	0.3954 (78,64%)
nDCG, cut-off @ 20	*0.7425*	*0.6798 (91,56%)*
nDCG, cut-off @ 30	*0.7288*	*0.6546 (89,82%)*
nDCG, cut-off @ 100	0.6878	0.6084 (88,46%)

[Zellhöfer 2015]

Further Reading

- Significance Testing because you will have to determine if the change of a metric's value is significant

- Participate at Kaggle's competitions or simply learn from others
  https://www.kaggle.com/

- https://medium.com/penn-engineering/machine-learning-researchers-try-to-improve-working-conditions-for-amazons-mechanical-turk-workers-f265a5f26e9e

Vielen Dank für Ihre Aufmerksamkeit.

Bibliography

- [Cleverdon 1962] CLEVERDON, Cyril W. ; ASLIB CRANFIELD RESEARCH PROJECT (Ed.): Aslib Cranfield Research Project: Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, USA, 1962
- [Lesk & Salton 1968] LESK, M. E. ; SALTON, Gerard: Relevance Assessments and Re- trieval System Evaluation. In: Information Storage and Retrieval 4 (1968), Nr. 4, 343–359
- [Robertson 1977] ROBERTSON, Stephen E.: The Probability Ranking Principle in IR. In: Journal of Documentation Bd. 4. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1977, 281–286
- [Järvelin & Kekäläinen 2002] JÄRVELIN, Kalervo ; KEKÄLÄINEN, Jaana: Cumulated Gain-based Evaluation of IR Techniques. In: ACM Trans. Inf. Syst. 20 (2002), Nr. 4, 422–446
- [Voorhees & Harman 2005] VOORHEES, Ellen M. ; HARMAN, Donna K.: TREC: Experiment and Evaluation in Information Retrieval. Cambridge, Mass. : MIT Press, 2005 (Digital Libraries and Electronic Publishing)
- [Craswell 2008] CRASWELL, N.; ZOETER, O.; TAYLOR, M.; RAMSEY, B.: An Experimental Comparison of Click Position-Bias Models. In: Proc. 1st Intl. Conf. on Web Search and Data Mining (2008), 87-94
- [Chapelle & Zhang 2009] CHAPELLE, O.; ZHANG, Y.: A Dynamic Bayesian Network Click Model for Web Search Ranking. In: 18th Intl. Conf. on World Wide Web (2009), 1-10
- [Zellhöfer 2013] ZELLHÖFER, David: Overview of the ImageCLEF 2013 Personal Photo Retrieval Subtask. In: CLEF 2013 Labs and Workshop, Notebook Papers, 23-26 September 2013, Valencia, Spain. 2013
- [Zellhöfer 2015] ZELLHÖFER, David: A Preference-based Relevance Feedback Approach for Polyrepresentative Multimedia Retrieval. Brandenburg University of Technology, Cottbus - Senftenberg, 2015