ONLINE PREFERENCE LEARNING WITH BANDIT ALGORITHMS

Eyke Hüllermeier
Intelligent Systems Group
Department of Computer Science
University of Paderborn, Germany

eyke@upb.de

TFLM 2015, Theoretical Foundations of Machine Learning, Bedlewo, Poland, 17-FEB-2015

PREFERENCES ARE UBIQUITOUS

INTELLIGENT SYSTEMS

Preferences play a key role in many applications of computer science and modern information technology:

COMPUTATIONAL ADVERTISING	RECOMMENDER SYSTEMS	COMPUTER GAMES
AUTONOMOUS AGENTS	ELECTRONIC COMMERCE	ADAPTIVE USER INTERFACES
PERSONALIZED MEDICINE	ADAPTIVE RETRIEVAL SYSTEMS	SERVICE-ORIENTED COMPUTING

PREFERENCES ARE UBIQUITOUS

INTELLIGENT SYSTEMS

Preferences play a key role in many applications of computer science and modern information technology:

COMPUTATIONAL ADVERTISING	RECOMMENDER SYSTEMS	COMPUTER GAMES
AUTONOMOUS AGENTS	ELECTRONIC COMMERCE	ADAPTIVE USER INTERFACES
PERSONALIZED MEDICINE	ADAPTIVE RETRIEVAL SYSTEMS	SERVICE-ORIENTED COMPUTING

medications or therapies specifically tailored for individual patients

COMMERCIAL INTEREST

INTELLIGENT SYSTEMS

Amazon files patent for "anticipatory" shipping

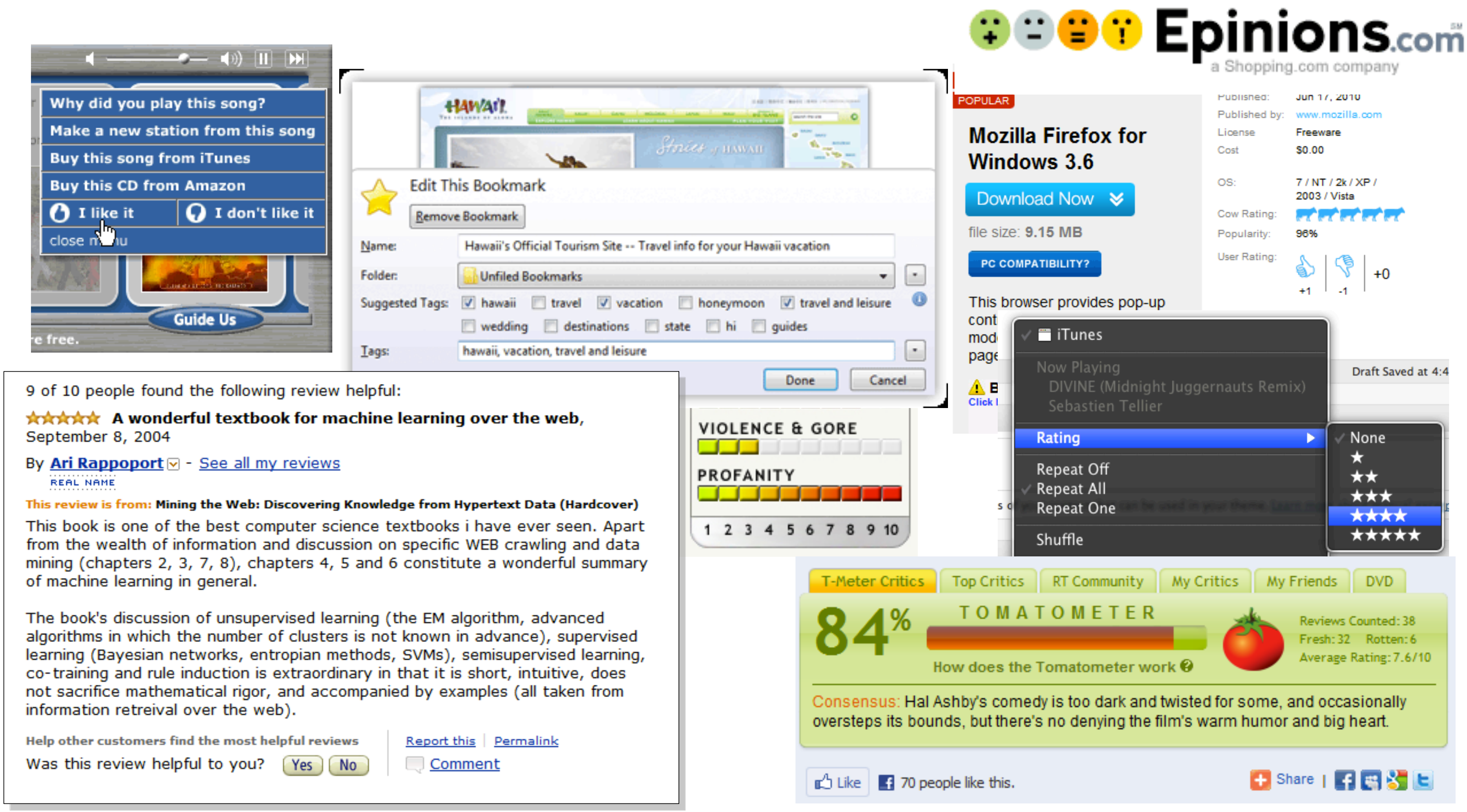
10 Comments / Shares / Tweets / Stumble / Email / More +

Amazon.com has filed for a patent for a shipping system that would anticipate what customers buy to decrease shipping time.

Amazon says the shipping system works by analyzing customer data like, purchasing history, product searches, wish lists and shopping cart contents, the Wall Street Journal reports. According to the patent filing, items would be moved from Amazon's fulfillment center to a shipping hub close to the customer in anticipation of an eventual purchase.

PREFERENCE INFORMATION

INTELLIGENT SYSTEMS



PREFERENCE INFORMATION

INTELLIGENT SYSTEMS

| Offizielle Homepage | **Daniel Baier** |
www.**daniel-baier**.com/
Willkommen auf der offiziellen Homepage von Fussballprofi **Daniel Baier** - TSV 1860 München.

NOT CLICKED ON

Prof. Dr. **Daniel Baier** - Brandenburgische Technische Universität **...**
www.tu-cottbus.de/fakultaet3/de/.../team/.../prof-dr-**daniel-baier**.html
Vökler, Sascha; Krausche, **Daniel**; **Baier**, Daniel: Product Design Optimization Using Ant Colony And Bee Algorithms: A Comparison, erscheint in: Studies in **...**

CLICKED ON

Daniel Baier
www.weltfussball.de/spieler_profil/**daniel-baier**/
Daniel Baier - FC Augsburg, VfL Wolfsburg, VfL Wolfsburg II, TSV 1860 München.

Daniel Baier - aktuelle Themen & Nachrichten - sueddeutsche.de
www.sueddeutsche.de/thema/**Daniel_Baier**
Aktuelle Nachrichten, Informationen und Bilder zum Thema **Daniel Baier** auf sueddeutsche.de.

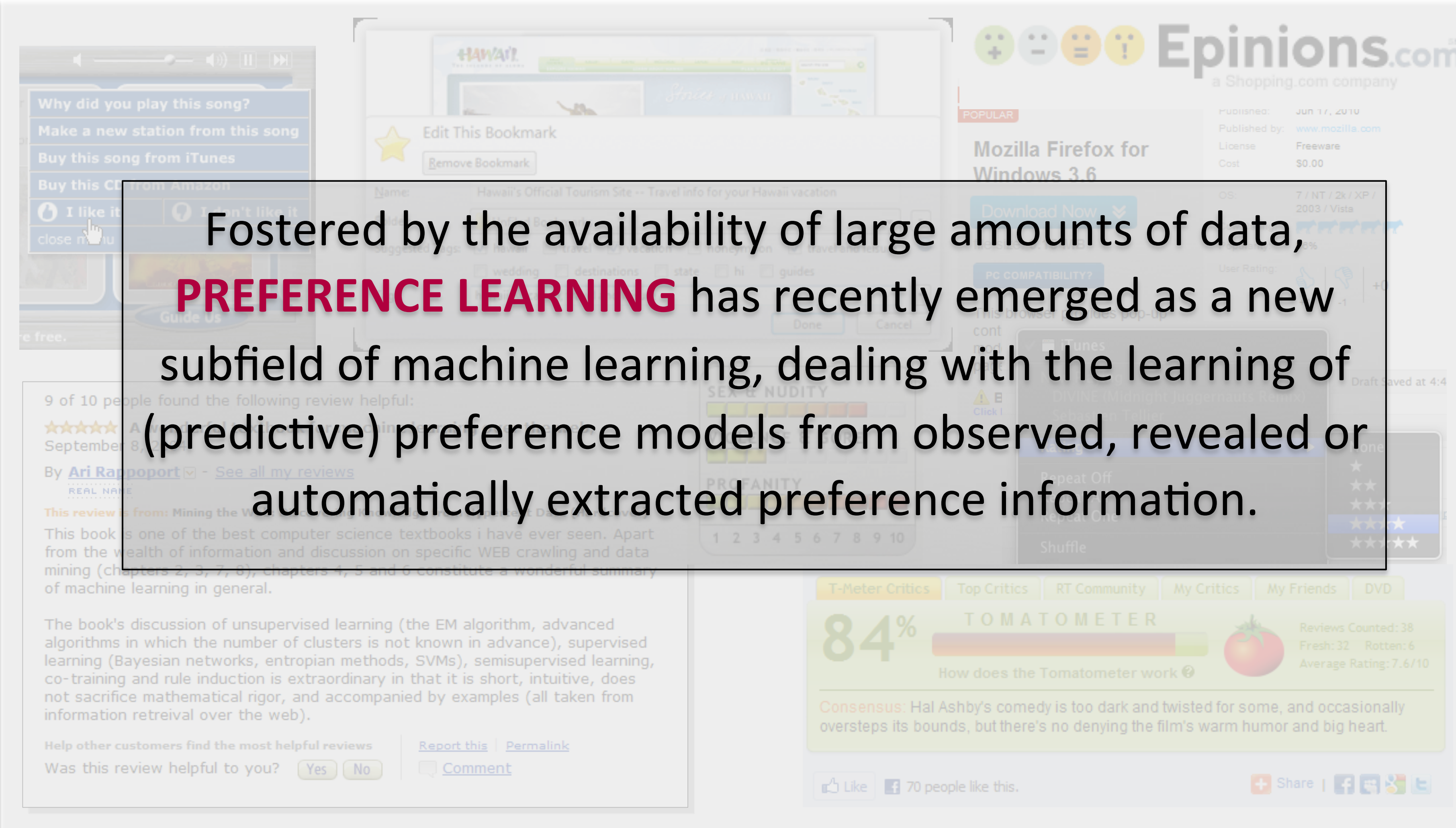
Daniel Baier | Facebook
de-de.facebook.com/**daniel.baier**.589
Tritt Facebook bei, um dich mit **Daniel Baier** und anderen Nutzern, die du kennst, zu vernetzen. Facebook ermöglicht den Menschen das Teilen von Inhalten mit **...**

FC Augsburg: Mein Tag in Bad Gögging: **Daniel Baier**
www.fcaugsburg.de/cms/website.php?id=/index/aktuell/news/...
2. Aug. 2012 – **Daniel Baier** berichtet heute, was für die Profis auf dem Programm stand. Hi FCA- Fans,. heute liegen wieder zwei intensive Trainingseinheiten **...**

- *Preferences are not necessarily expressed explicitly, but can be extracted* ***implictly from people's behavior****!*
- *Massive amounts of very* ***noisy data****!*

PREFERENCE LEARNING

INTELLIGENT SYSTEMS



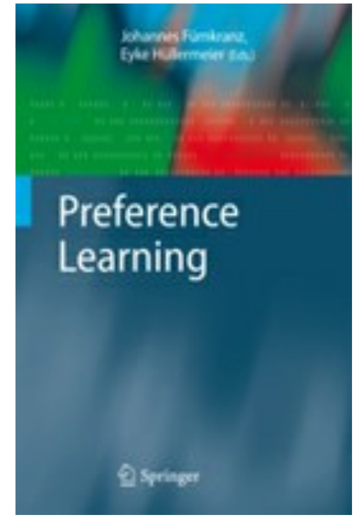
Fostered by the availability of large amounts of data, **PREFERENCE LEARNING** has recently emerged as a new subfield of machine learning, dealing with the learning of (predictive) preference models from observed, revealed or automatically extracted preference information.

PL IS AN ACTIVE FIELD

INTELLIGENT
SYSTEMS

Tutorials:

- European Conf. on Machine Learning, 2010
- Int. Conf. Discovery Science, 2011
- Int. Conf. Algorithmic Decision Theory, 2011
- European Conf. on Artificial Intelligence, 2012
- Int. Conf. Algorithmic Learning Theory, 2014



J. Fürnkranz &
E. Hüllermeier (eds.)
Preference Learning
Springer-Verlag 2011

Special Issue on
Representing,
Processing, and
Learning Preferences:
Theoretical and
Practical Challenges
(2011)

Special Issue on
Preference Learning
Forthcoming

PL IS AN ACTIVE FIELD

INTELLIGENT
SYSTEMS

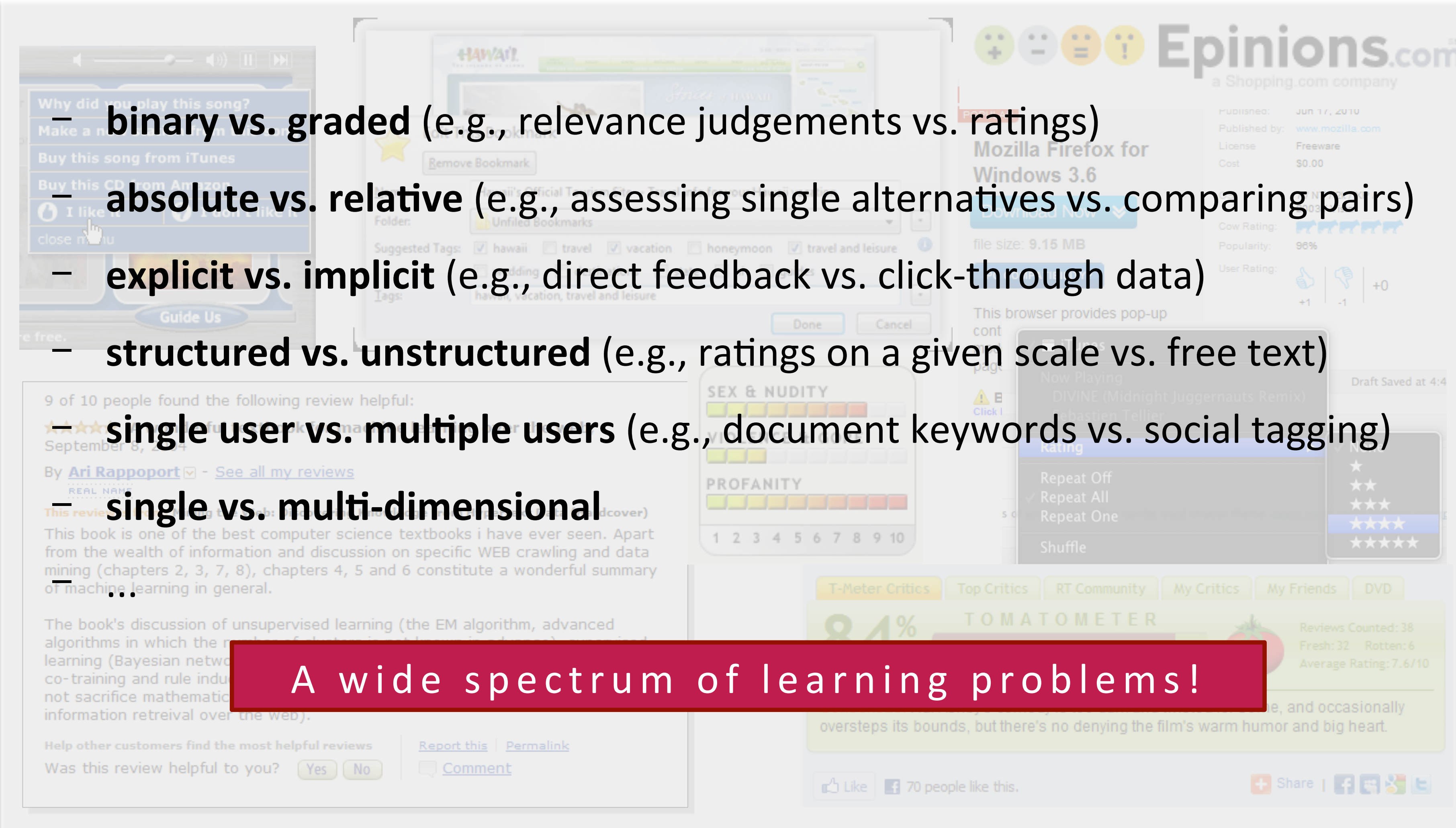
- NIPS 2001: New Methods for Preference Elicitation
- NIPS 2002: Beyond Classification and Regression: Learning Rankings, Preferences, Equality Predicates, and Other Structures
- KI 2003: Preference Learning: Models, Methods, Applications
- NIPS 2004: Learning with Structured Outputs
- NIPS 2005: Workshop on Learning to Rank
- IJCAI 2005: Advances in Preference Handling
- SIGIR 07–10: Workshop on Learning to Rank for Information Retrieval
- ECML/PDKK 08–10: Workshop on Preference Learning
- NIPS 2009: Workshop on Advances in Ranking
- American Institute of Mathematics Workshop in Summer 2010: The Mathematics of Ranking
- NIPS 2011: Workshop on Choice Models and Preference Learning
- EURO 2009-12: Special Track on Preference Learning
- ECAI 2012: Workshop on Preference Learning: Problems and Applications in AI
- DA2PL 2012: From Decision Analysis to Preference Learning
- **Dagstuhl Seminar on Preference Learning (2014)**
- NIPS 2014: Analysis of Rank Data: Confluence of Social Choice, Operations Research, and Machine Learning

CONNECTIONS TO OTHER FIELDS

INTELLIGENT SYSTEMS

- Structured Output Prediction
- Learning Monotone Models
- Classification (ordinal, multilabel, ...)
- Information Retrieval
- Learning with weak supervision
- **Preference Learning**
- Recommender Systems
- Economics & Decison Science
- Statistics
- Social Choice
- Graph theory
- Optimization
- Operations Research
- Multiple Criteria Decision Making

MANY TYPES OF PREFERENCES



- **binary vs. graded** (e.g., relevance judgements vs. ratings)
- **absolute vs. relative** (e.g., assessing single alternatives vs. comparing pairs)
- **explicit vs. implicit** (e.g., direct feedback vs. click-through data)
- **structured vs. unstructured** (e.g., ratings on a given scale vs. free text)
- **single user vs. multiple users** (e.g., document keywords vs. social tagging)
- **single vs. multi-dimensional**
- ...

A wide spectrum of learning problems!

OUTLINE

INTELLIGENT SYSTEMS

PART 1	PART 2	PART 3
Preference learning	Ranking problems	Preference-based bandit algorithms

OBJECT RANKING [Cohen et al., 1999]

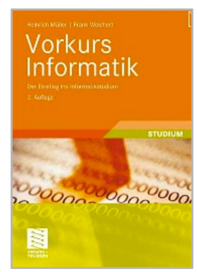
TRAINING

$$
\begin{array}{ccc}
(0.74, 1, 25, 165) & \succ & (0.45, 0, 35, 155) \\
(0.47, 1, 46, 183) & \succ & (0.57, 1, 61, 177) \\
(0.25, 0, 26, 199) & \succ & (0.73, 0, 46, 185)
\end{array}
$$

Pairwise preferences between objects

$\succ$

$\succ$



OBJECT RANKING [Cohen et al., 1999]

PREDICTION (ranking a new set of objects)

$$Q = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9, \boldsymbol{x}_{10}, \boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \boldsymbol{x}_{13}\}$$

$$\boldsymbol{x}_{10} \succ \boldsymbol{x}_4 \succ \boldsymbol{x}_7 \succ \boldsymbol{x}_1 \succ \boldsymbol{x}_{11} \succ \boldsymbol{x}_2 \succ \boldsymbol{x}_8 \succ \boldsymbol{x}_{13} \succ \boldsymbol{x}_9 \succ \boldsymbol{x}_3 \succ \boldsymbol{x}_{12} \succ \boldsymbol{x}_5 \succ \boldsymbol{x}_6$$

$Q =$ 

PREFERENCE LEARNING TASKS

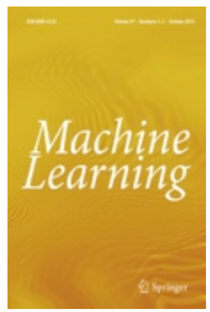
Theoretically challenging, because

- supervision is weak (partial, noisy,...),
- sought predictions are complex/structured,
- performance metrics are hard to optimize,
- ...

LABEL RANKING [EH et al., 2008]

... learning models that map instances to **TOTAL ORDERS** over a fixed set of alternatives/labels:

$\mapsto$



$\succ$

$\succ$

$\succ$

`(1.35,0,35,324)`

... likes more
... reads more
... publishes more in
...

LABEL RANKING: TRAINING DATA

TRAINING

X_1	X_2	X_3	X_4	preferences
0.34	0	10	174	$A \succ B, C \succ D$
1.45	0	32	277	$B \succ C \succ A$
1.22	1	46	421	$B \succ D, A \succ D, C \succ D, A \succ C$
0.74	1	25	165	$C \succ A \succ D, A \succ B$
0.95	1	72	273	$B \succ D, A \succ D$
1.04	0	33	158	$D \succ A \succ B, C \succ B, A \succ C$

Instances are associated with preferences between labels

... no demand for full rankings!

LABEL RANKING: PREDICTION

INTELLIGENT SYSTEMS

PREDICTION

				A	B	C	D
0.92	1	81	382	?	?	?	?

new instance

ranking ?

LABEL RANKING: PREDICTION

PREDICTION

				A	B	C	D
0.92	1	81	382	4	1	3	2

new instance

$\pi(i) = \text{position of } i\text{-th label}$

A ranking of all labels

LABEL RANKING: PREDICTION

INTELLIGENT SYSTEMS

PREDICTION

0.92	1	81	382	4	1	3	2

LOSS

GROUND TRUTH

0.92	1	81	382	2	1	3	4

A ranking of all labels

LABEL RANKING: PREDICTION

INTELLIGENT SYSTEMS

PREDICTION

0.92	1	81	382	4	1	3	2

A ranking of all labels

LOSS

GROUND TRUTH

0.92	1	81	382	2	1	3	4

KENDALL

$$\mathcal{L}(\pi, \pi^*) = \sum_{1 \leq i < j \leq M} \left[\!\left[(\pi(i) - \pi(j))(\pi^*(i) - \pi^*(j)) < 0 \right]\!\right]$$

LOSS

$$\tau = 1 - \frac{4D(\pi, \pi^*)}{M(M-1)}$$

RANK CORRELATION

METHODS FOR LABEL RANKING

INTELLIGENT SYSTEMS

Reduction to binary classification	Ranking by pairwise comparison [Hüllermeier et al. 08]	Learning pairwise preferences
	Constraint classification [Har-Peled et al. 03]	Learning utility functions
Boosting	Log-linear models for label ranking [Dekel et al. 04]	
Structured output prediction, margin maximization	Structured output prediction [Vembu et al. 09] Local prediction (lazy learning) [Brinker & EH , Cheng et al. 09]	Structured prediction
Statistical inference	Statistical models for label ranking [Cheng et al. 09, Cheng et al. 10]	

PREFERENCE LEARNING TASKS

INTELLIGENT SYSTEMS

task	context (input)	alternative (output)	training information	prediction	ground truth
	representation		type of preference information		
collaborative filtering	ID	ID	absolute ordinal	absolute ordinal	absolute ordinal
dyadic prediction	feature	feature	absolute ordinal	absolute ordinal	absolute ordinal
multilabel classification	feature	ID	absolute binary	absolute binary	absolute binary
multilabel ranking	feature	ID	absolute binary	ranking	absolute binary
label ranking	feature	ID	relative binary	ranking	ranking
object ranking	---	feature	relative binary	ranking	ranking or subset
instance ranking	---	feature	absolute ordinal	ranking	absolute ordinal

... not so much work on ***online preference learning*** *so far.*

PROBABILITIES ON RANKINGS

3 items { A, B, C }

$A \succ B \succ C$	p_1
$A \succ C \succ B$	p_2
$B \succ A \succ C$	p_3
$B \succ C \succ A$	p_4
$C \succ A \succ B$	p_5
$C \succ B \succ A$	p_6

4 items { A, B, C, D }

$A \succ B \succ C \succ D$	p_1
$A \succ B \succ D \succ C$	p_2
$A \succ C \succ B \succ D$	p_3
$A \succ C \succ D \succ B$	p_4
$A \succ D \succ B \succ C$	p_5
$A \succ D \succ C \succ B$	p_6
$B \succ A \succ C \succ D$	p_7
$B \succ A \succ D \succ C$	p_8
$B \succ C \succ A \succ D$	p_9
$B \succ C \succ D \succ A$	p_{10}
$B \succ D \succ A \succ C$	p_{11}
$B \succ D \succ C \succ A$	p_{12}
$C \succ A \succ B \succ D$	p_{13}
$C \succ A \succ D \succ B$	p_{14}
$C \succ B \succ A \succ D$	p_{15}
$C \succ B \succ D \succ A$	p_{16}
$C \succ D \succ A \succ B$	p_{17}
$C \succ D \succ B \succ A$	p_{18}
$D \succ A \succ B \succ C$	p_{19}
$D \succ A \succ C \succ B$	p_{20}
$D \succ B \succ A \succ C$	p_{21}
$D \succ B \succ C \succ A$	p_{22}
$D \succ C \succ A \succ B$	p_{23}
$D \succ C \succ B \succ A$	p_{24}

Need a parameterized family of distributions on the permutation space!

PROBABILITIES ON RANKINGS

INTELLIGENT SYSTEMS

item	A	B	C	D
rank	2	3	4	1

$\longleftrightarrow$ $D \succ A \succ C \succ B$

- Rankings can be represented by permutations $\pi : \{1, \ldots, K\} \to \{1, \ldots, K\}$.
- $\pi(i)$ is the rank of the i-th item.
- The set of all permutations is the symmetric group of order K, denoted $\mathcal{S}_K$.

THE MALLOWS MODEL

... is a **distance-based** probability distribution $\mathbf{P} : \mathcal{S}_K \to [0,1]$, which belongs to the exponential family:

reference ranking

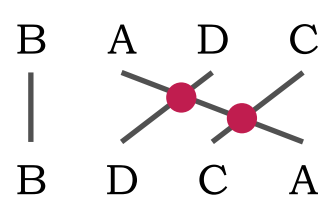
spread parameter

$$\mathbf{P}(\pi \,|\, \pi_0, \theta) = \frac{\exp\big(-\theta \Delta(\pi, \pi_0)\big)}{\phi(\pi_0, \theta)}$$

normalization constant

where Δ is the Kendall distance on permutations (number of item pairs differently ordered):

$$\Delta(\pi, \pi_0) = \#\{1 \leq i < j \leq K \,|\, (\pi(i) - \pi(j))(\pi_0(i) - \pi_0(j)) < 0\}$$



OUTLINE

INTELLIGENT SYSTEMS

PART 1

Preference learning

PART 2

Ranking problems

PART 3

Preference-based bandit algorithms

MULTI-ARMED BANDITS

„pulling an arm“ ⟷ choosing an option

partial information online learning
sequential decision process

MULTI-ARMED BANDITS

INTELLIGENT SYSTEMS

„pulling an arm“ ⟷ putting an advertisement on a website

choice of an option/strategy (arm) yields a **random reward**

partial information online learning
sequential decision process

MULTI-ARMED BANDITS

INTELLIGENT SYSTEMS

$X_1 \sim \mathbf{P}_1$ | $X_2 \sim \mathbf{P}_2$ | $X_3 \sim \mathbf{P}_3$ | $X_4 \sim \mathbf{P}_4$ | $X_5 \sim \mathbf{P}_5$

„pulling an arm“ ⟷ choosing an option

choice of an option/strategy (arm) yields a **random reward**

partial information online learning
sequential decision process

MULTI-ARMED BANDITS

$X_1 \sim \mathbf{P}_1$ $X_2 \sim \mathbf{P}_2$ $X_3 \sim \mathbf{P}_3$ $X_4 \sim \mathbf{P}_4$ $X_5 \sim \mathbf{P}_5$

Immediate reward: 2.5

Cumulative reward: 2.5

MULTI-ARMED BANDITS

$X_1 \sim \mathbf{P}_1$ $X_2 \sim \mathbf{P}_2$ $X_3 \sim \mathbf{P}_3$ $X_4 \sim \mathbf{P}_4$ $X_5 \sim \mathbf{P}_5$

Immediate reward:	`2.5`	`3.1`
Cumulative reward:	`2.5`	`5.6`

MULTI-ARMED BANDITS

$X_1 \sim \mathbf{P}_1$ | $X_2 \sim \mathbf{P}_2$ | $X_3 \sim \mathbf{P}_3$ | $X_4 \sim \mathbf{P}_4$ | $X_5 \sim \mathbf{P}_5$

Immediate reward:	`2.5 3.1 1.7`
Cumulative reward:	`2.5 5.6 7.3`

MULTI-ARMED BANDITS

INTELLIGENT SYSTEMS

$X_1 \sim \mathbf{P}_1$ $X_2 \sim \mathbf{P}_2$ $X_3 \sim \mathbf{P}_3$ $X_4 \sim \mathbf{P}_4$ $X_5 \sim \mathbf{P}_5$

Immediate reward:	2.5	3.1	1.7	3.7	...
Cumulative reward:	2.5	5.6	7.3	11.0	...

maximize cumulative reward → *explore and exploit (tradeoff)*

find best option → *pure exploration (effort vs. certainty)*

MULTI-ARMED BANDITS

INTELLIGENT SYSTEMS

– A **policy** is an algorithm that prescribes an arm to be played in each round, based on the outcomes of the previous rounds.

– Denote by $\mu_i = \mathbf{E}(X_i)$ the expected reward of arm a_i and

$$\mu^* = \max_{1 \leq j \leq K} \mu_j \quad .$$

– Define the **regret** and **cumulative regret**, respectively, as

$$r_t = \mu^* - x_{i(t)}, \quad R^T = \sum_{t=1}^{T} r_t \quad ,$$

where $i(t)$ is the index of the arm played in round t.

THE UCB ALGORITHM

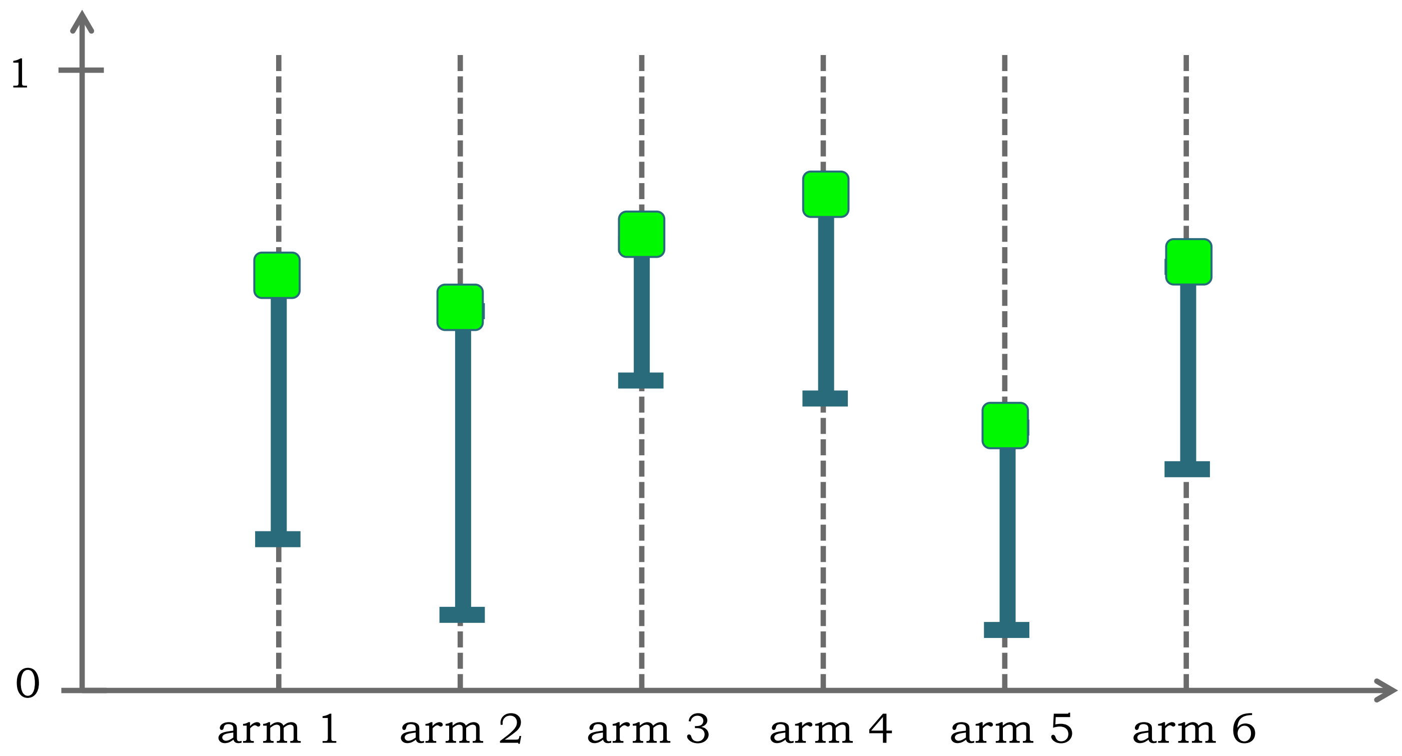
INTELLIGENT SYSTEMS

Algorithm 1 Upper Confidence Bound

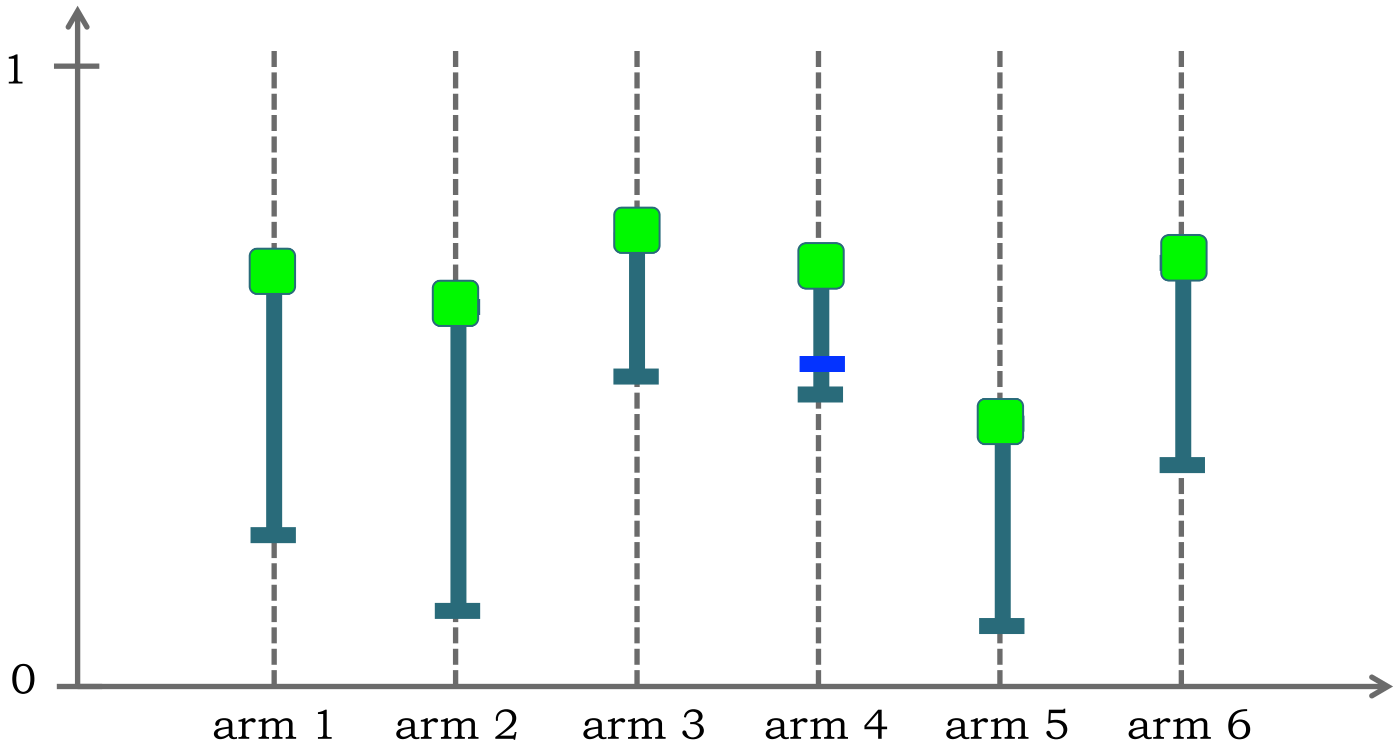
for all $1 \leq i \leq K$ **do**
  $\hat{\mu}_i \leftarrow \infty$ {empirical mean of arm a_i}
  $t_i \leftarrow 0$ {number of times played arm a_i}
end for
$t \leftarrow 1$
while true **do**
  $k \leftarrow \arg\max_i \hat{\mu}_i + \sqrt{\frac{2 \log t}{t_i}}$ {upper confidence bound from Chernoff-Hoeffding}
  play arm a_k, update empirical mean $\hat{\mu}_k$, increment t_k
  $t \leftarrow t + 1$
end while

The UCB algorithm, introduced by Auer et al. (2002), implements the **optimism in the face of uncertainty** principle.

THE UCB ALGORITHM



THE UCB ALGORITHM



BOUND ON EXPECTED REGRET

Theorem: Assume rewards in $[0,1]$ (i.e., distributions $\mathbf{P}_1, \ldots, \mathbf{P}_K$ with support in $[0,1]$). The expected cumulative regret of UCB after any number of rounds T is upper-bounded by

$$\left[8 \sum_{i:\, \mu_i < \mu^*} \left(\frac{\log T}{\Delta_i}\right)\right] + \left(1 + \frac{\pi^2}{3}\right)\left(\sum_{j=1}^{K} \Delta_j\right) \in \mathcal{O}(K \log T) \ ,$$

where $\Delta_i = \mu^* - \mu_i$.

PREFERENCE-BASED BANDITS

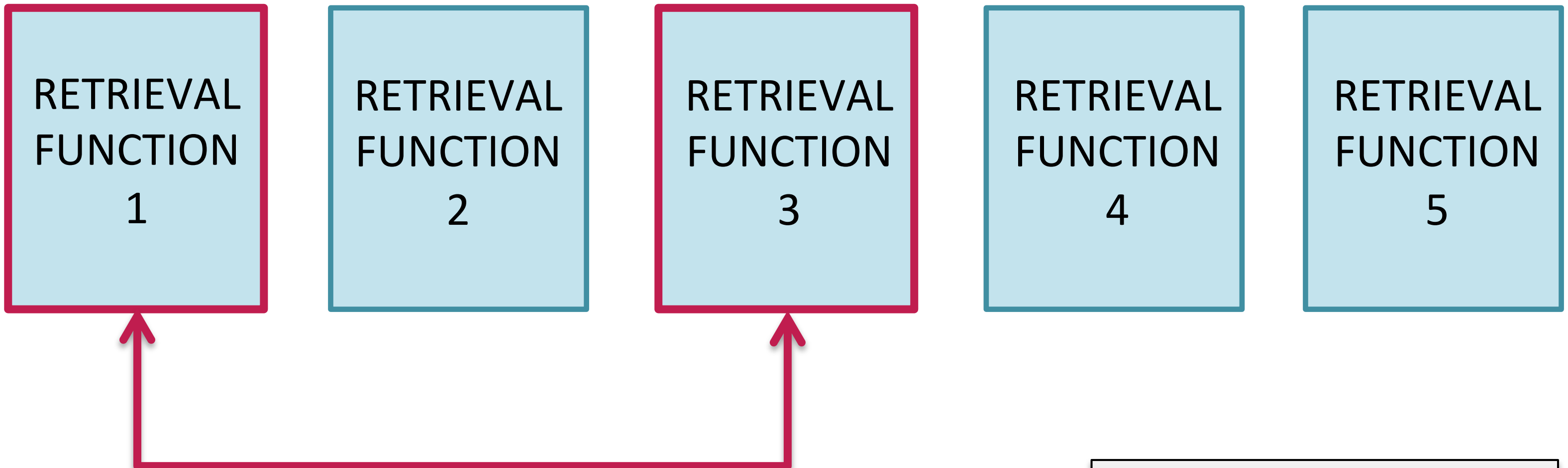
$X_1 \sim \mathbf{P}_1$ $X_2 \sim \mathbf{P}_2$ $X_3 \sim \mathbf{P}_3$ $X_4 \sim \mathbf{P}_4$ $X_5 \sim \mathbf{P}_5$

In many applications,

- the assignment of (numeric) **rewards to single outcomes** (and hence the assessment of individual options on an absolute scale) is difficult,
- while the **qualitative comparison between pairs of outcomes** (arms/ options) is more feasible.

PREFERENCE-BASED BANDITS

INTELLIGENT SYSTEMS

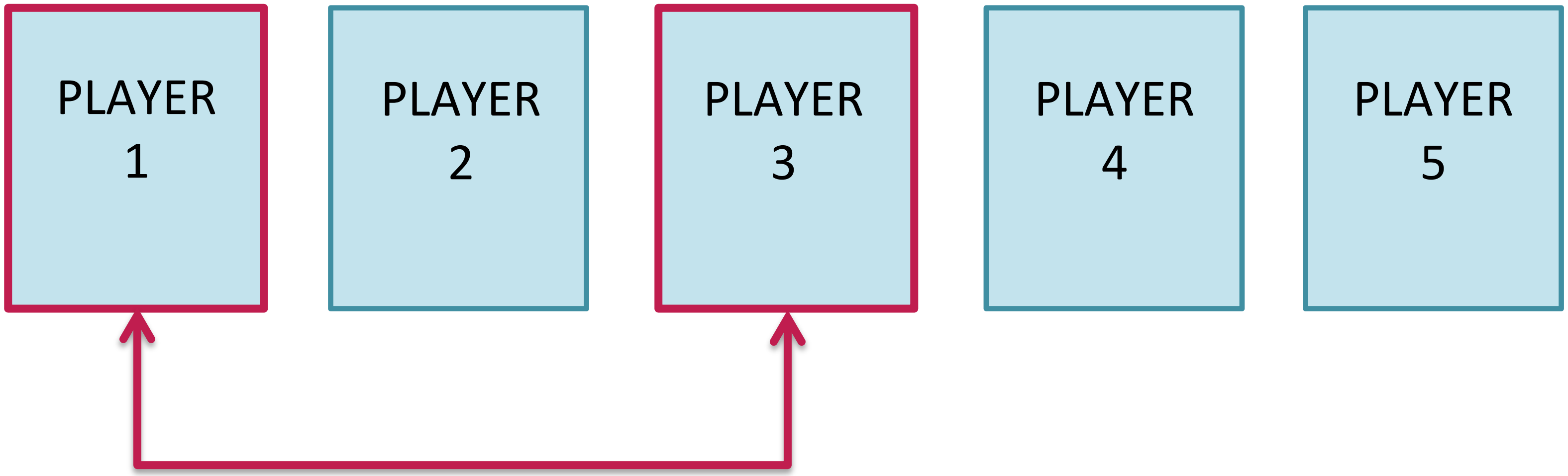


$$X_3 \succ X_1$$

The result returned by the third retrieval function, for a given query, is preferred to the result returned by the first search engine.

Noisy preference can be inferred from how a user clicks through an **interleaved** list of documents [Radlinski et al., 2008].

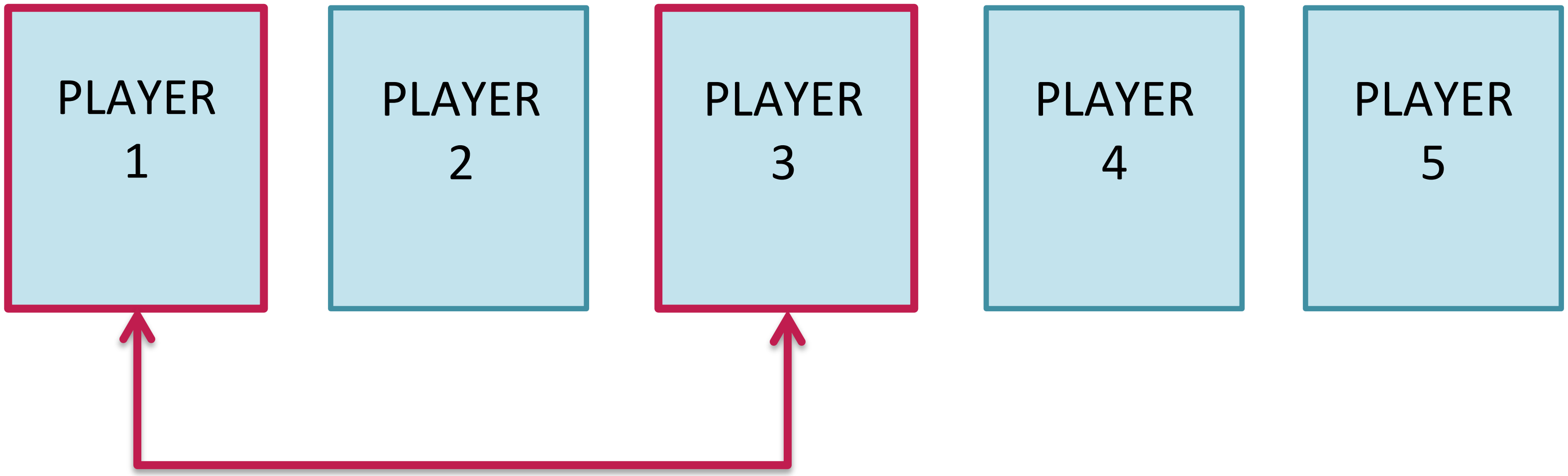
PREFERENCE-BASED BANDITS



$$X_3 \succ X_1$$

Third player has beaten first player in a match.

PREFERENCE-BASED BANDITS



$$X_3 \succ X_1$$

- *This setting has first been introduced as the **dueling bandits problem** (Yue and Joachims, 2009).*
- *More generally, we speak of **preference-based multi-armed bandits** (PB-MAB).*

FORMAL SETTING

- fixed set of arms (options) $\mathcal{A} = \{a_1, \ldots, a_K\}$
- **action space** of the learner (agent) $= \{\, (i,j) \,|\, 1 \leq i \leq j \leq K \,\}$ (compairing pairs of arms a_i and a_j)
- feedback generated by an (unknown, time-stationary) probabilistic process characterized by a **preference relation**

$$\mathbf{Q} = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,K} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ q_{K,1} & q_{K,2} & \cdots & q_{K,K} \end{bmatrix} ,$$

where

$$q_{i,j} = \mathbf{P}\,(a_i \succ a_j)$$

- typically, $\mathbf{Q}$ is reciprocal $(q_{i,j} = 1 - q_{j,i})$

THE PREFERENCE RELATION

– We say arm a_i beats arm a_j if $q_{i,j} > 1/2$.

– The degrees of **distinguishability**

$$\Delta_{i,j} = q_{i,j} - \frac{1}{2}$$

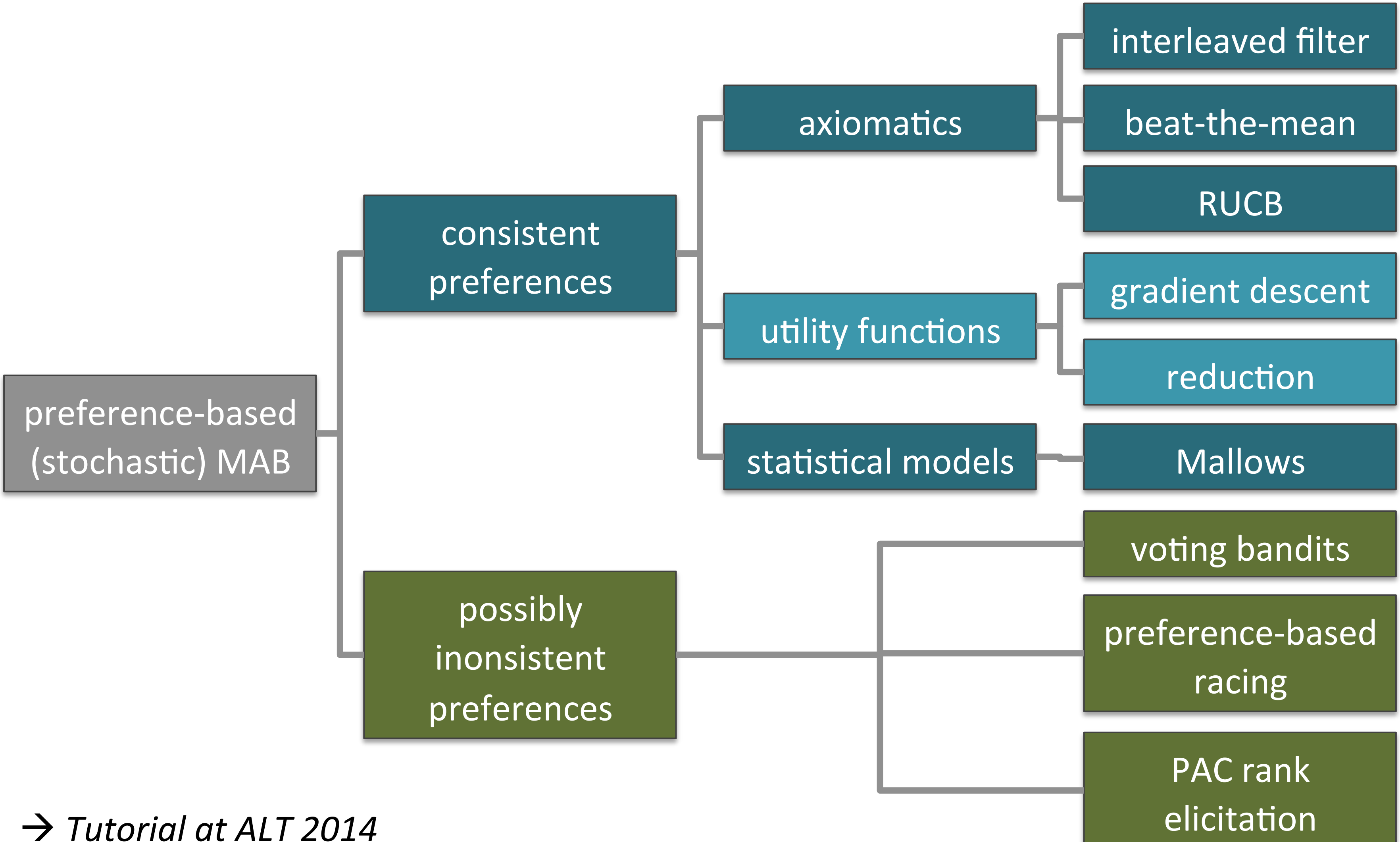
quantify the hardness of a PB-MAB task.

– Definition of **regret** is not straightforward.

– Assumptions on properties of $\mathbf{Q}$ are crucial for learning.

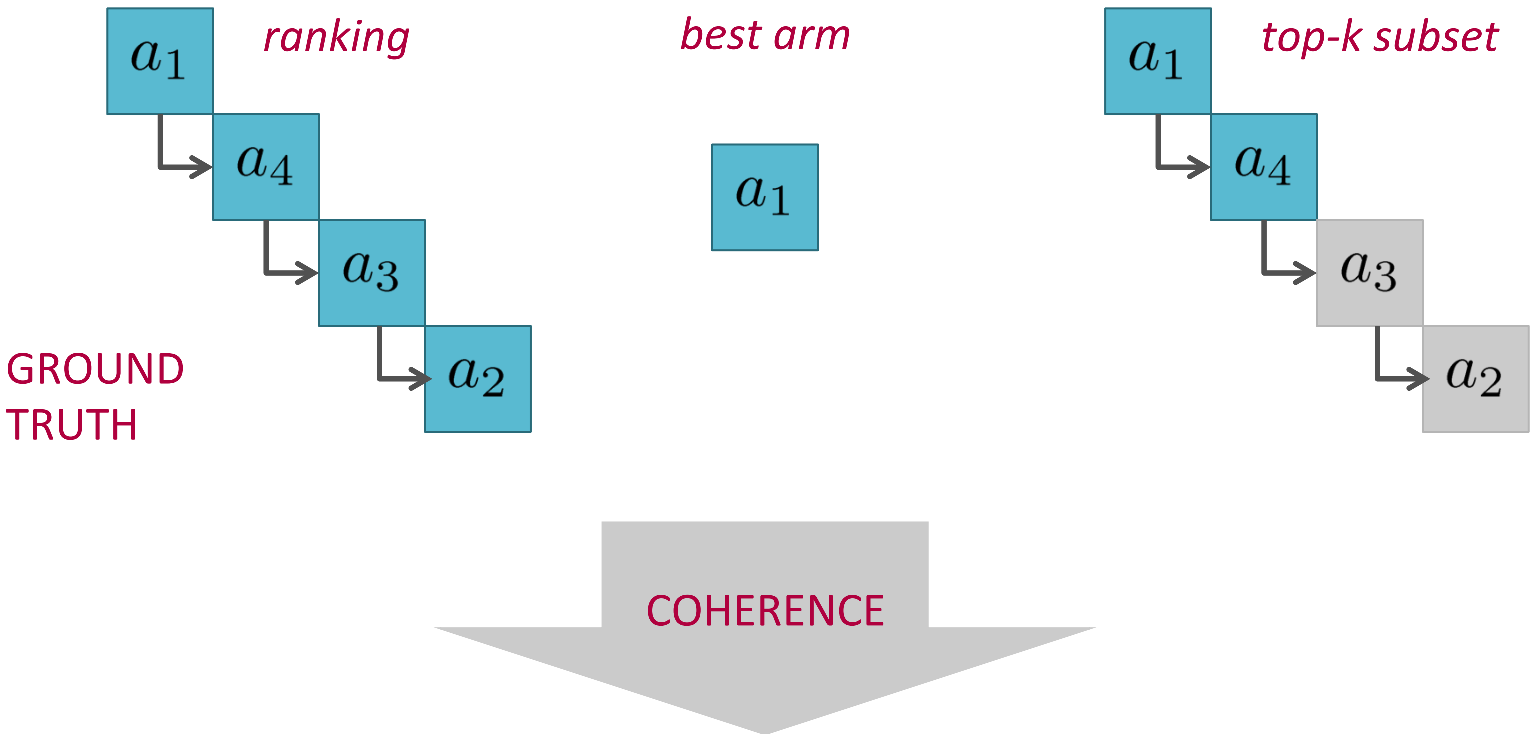
– **Coherence:** The pairwise comparisons need to provide hints (even if "noisy" ones) on the target.

OVERVIEW OF METHODS



→ *Tutorial at ALT 2014*

PROPERTIES OF PREFERENCE RELATION



$$\mathbf{Q} = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,K} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ q_{K,1} & q_{K,2} & \cdots & q_{K,K} \end{bmatrix}$$

... the preference relation is derived from, or at least strongly restricted by the target!

INTELLIGENT SYSTEMS

– Yue and Joachims (2009) proposed Interleaved Filtering, which assumes (i) existence of a total order over arms, (ii) strong stochastic transitivity, (iii) stochastic triangle inequality.

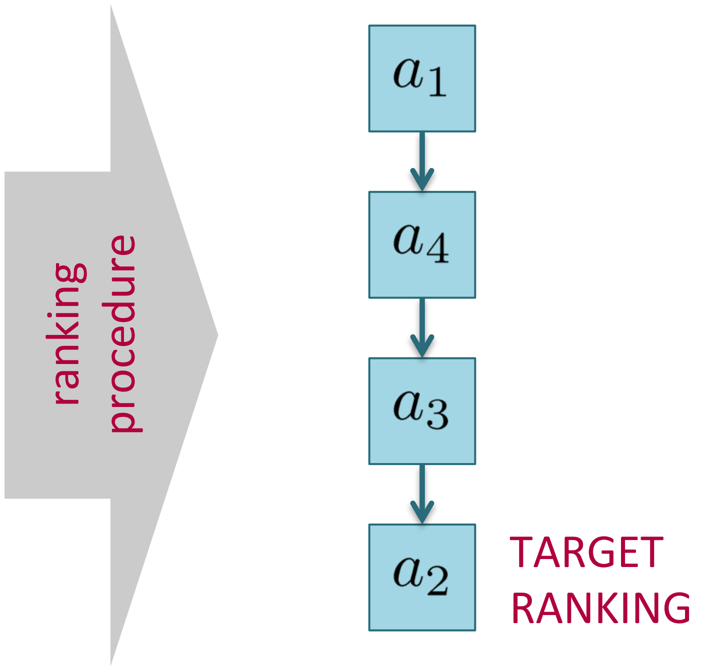
– Zoghi et al. (2014) proposed Relative UCB, which only assumes the existence of a Condorcet winner.

	a_1	a_2	a_3	a_4
a_1	--	0.6	0.6	0.6
a_2	0.4	--	0.8	0.9
a_3	0.4	0.2	--	0.6
a_4	0.4	0.1	0.4	--

PREFERENCE-BASED RANK ELICITATION

Take any preference relation as a point of departure ...

$$\mathbf{Q} = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,K} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ q_{K,1} & q_{K,2} & \cdots & q_{K,K} \end{bmatrix}$$



PROPERTIES OF PREFERENCE RELATION

Borda (weighted voting, sum of expectations):

	a_1	a_2	a_3	a_4	
a_1	--	0.6	0.6	0.6	**1.8**
a_2	0.4	--	0.8	0.9	**2.1**
a_3	0.4	0.2	--	0.6	**1.2**
a_4	0.4	0.1	0.4	--	**0.9**

$$a_2 \succ a_1 \succ a_3 \succ a_4$$

Easy reduction for the case of Borda:

	a_1	a_2	a_3	a_4	
a_1	--	0.6	0.6	0.6	**1.8**
a_2	0.4	--	0.8	0.9	**2.1**
a_3	0.4	0.2	--	0.6	**1.2**
a_4	0.4	0.1	0.4	--	**0.9**

Choosing an arm = pairing it with a randomly chosen alternative:

	a_1	a_2	a_3	a_4
reward 0	0.4	0.3	0.6	0.7
reward 1	0.6	0.7	0.4	0.3

PROPERTIES OF PREFERENCE RELATION

INTELLIGENT SYSTEMS

Statistical approach

Coherence through statistical assumptions on the data generating process, e.g., pairwise probabilities as marginals of a Mallows model:

$$q_{i,j} = \mathbf{P}(\, a_i \succ a_j \,) = \sum_{\pi:\, \pi(i)<\pi(j)} \mathbf{P}(\pi \,|\, \pi_0, \theta)$$

$$= \frac{1}{\phi(\pi_0, \theta)} \sum_{\pi:\, \pi(i)<\pi(j)} \exp\big(-\theta \Delta(\pi, \pi_0)\big)$$

$\longrightarrow$ reference ranking π_0 is the natural target!

PROBABILITY ESTIMATION

- In each iteration $t \in \mathbb{T}$, the learner selects $(i(t), j(t))$ and observes

$$\begin{cases} a_{i(t)} \succ a_{j(t)} & \text{with probability } q_{i(t),j(t)} \\ a_{j(t)} \succ a_{i(t)} & \text{with probability } q_{j(t),i(t)} \end{cases}$$

- Probability $q_{i,j}$ can be estimated by the proportion of wins of a_i against a_j up to iteration t:

$$\widehat{q}_{i,j}^{\,t} = \frac{w_{i,j}^t}{n_{i,j}^t} = \frac{w_{i,j}^t}{w_{i,j}^t + w_{j,i}^t}$$

- As samples are i.i.d., this is a plausible estimate; yet, it might be biased, since $n_{i,j}^t$ depends on the choice of the learner and hence on the data ($n_{i,j}^t$ is a random quantity).

PROBABILITY ESTIMATION

INTELLIGENT SYSTEMS

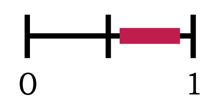
- A high probability confidence interval of the form

$$\left[\,\widehat{q}^{\,t}_{i,j} - c^{t}_{i,j},\ \widehat{q}^{\,t}_{i,j} + c^{t}_{i,j}\,\right]$$

  can be obtained based on concentration inequalities like Hoeffding.

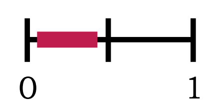
- Option a_i beats a_j with high probability if

$$\widehat{q}^{\,t}_{i,j} - c^{t}_{i,j} > 1/2\ .$$



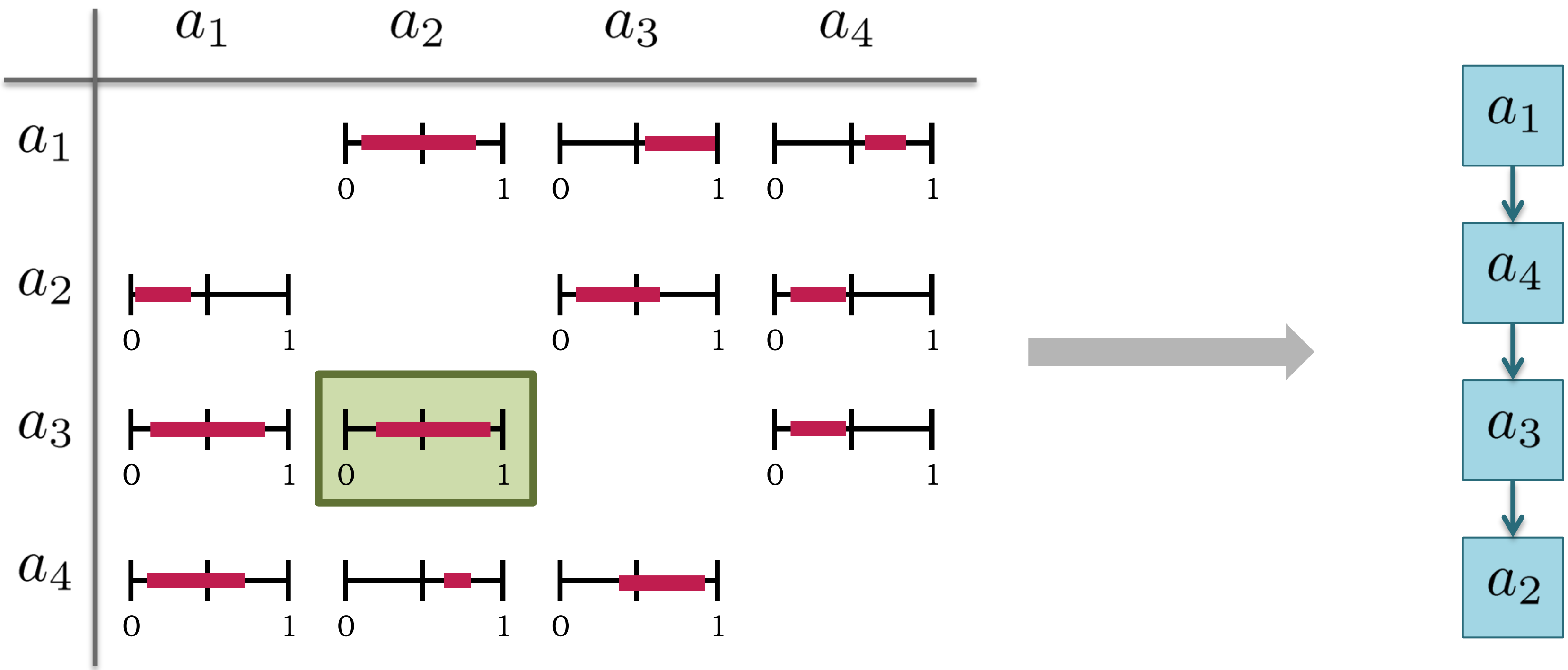
- Option a_j beats a_i with high probability if

$$\widehat{q}^{\,t}_{i,j} + c^{t}_{i,j} < 1/2\ .$$



PAIRWISE SAMPLING

INTELLIGENT SYSTEMS



uncertainty about pairwise preferences

translates → into

uncertainty about ranking

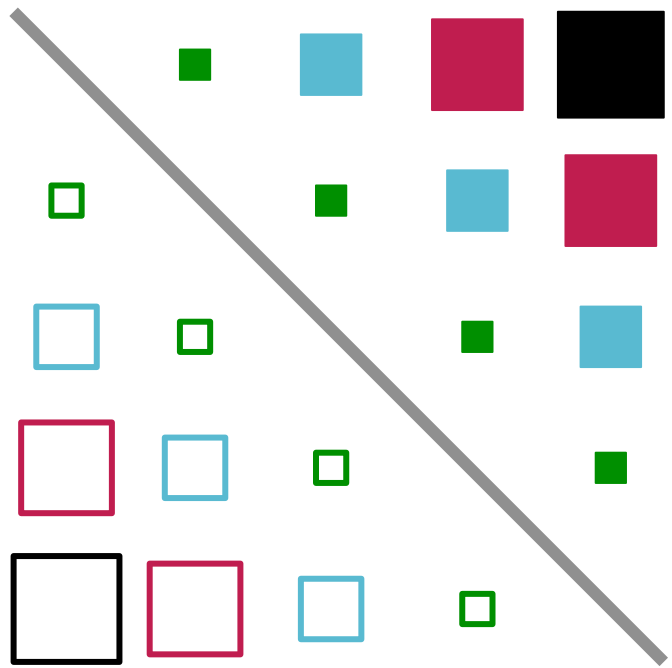
For example, RUCB selects a_c from the set of potential Condorcet winners and compares it to the arm a_d supposed to yield the smallest regret.

THE MALLOWS MODEL

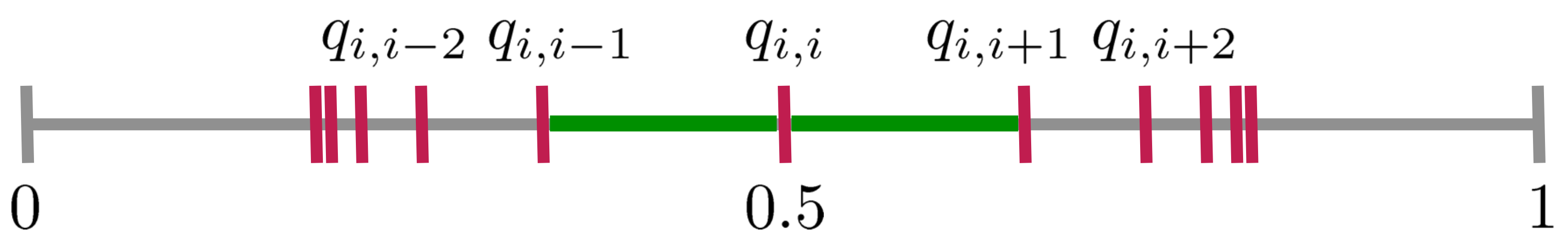
Important observation: With π_0 the identity, the matrix $\mathbf{Q} = (q_{i,j})$ induced by Mallows has a Toeplitz structure:

$$q_{i,j} = h(j-i+1,\theta) - h(j-i,\theta) \ ,$$

with $h(k,\theta) = k/(1-\exp(-k\theta))$.



THE MALLOWS MODEL



- Compared to weaker model assumptions, Mallows induces a **highly regular structure** on the pairwise marginals.
- These are coherent with the target ranking in the sense that $\pi_0(i) < \pi_0(j)$ implies $q_{i,j} > 1/2$ and $\pi_0(i) < \pi_0(j) < \pi_0(k)$ implies $q_{i,j} < q_{i,k}$. (Yet, stochastic triangle inequality does not hold.)
- Most importantly, Mallows assures a **minimum separation** ρ between neighbored options, which depends on θ.
- This allows for establishing a connection to (noisy) **sorting**.

- Busa-Fekete et al. (2014) propose a sampling strategy called **MallowsMPR**, which is based on the **merge sort** algorithm for selecting the arms to be compared.
- However, two arms a_i and a_j are not only compared once but until

$$1/2 \notin \left[\, \widehat{q}_{i,j} - c_{i,j}, \widehat{q}_{i,j} + c_{i,j} \,\right] \ .$$

- **Theorem:** For any $0 < \delta < 1$, MallowsMPR outputs the reference ranking π_0 with probability at least $1 - \delta$, and the number of pairwise comparisons taken by the algorithm is

$$\mathcal{O}\left(\frac{K \log_2 K}{\rho^2} \log \frac{K \log_2 K}{\delta \rho} \right) \ ,$$

where $\rho = \frac{1-\phi}{1+\phi}$, $\phi = \exp(-\theta)$.

PREFERENCE-BASED RANK ELICITATION

Algorithm MallowsMPR(δ)

for $i = 1$ to K **do**
    $r_i \leftarrow i$
    $r'_i \leftarrow 0$
end for
$(\boldsymbol{r}, \boldsymbol{r}') \leftarrow \text{MMRec}(\boldsymbol{r}, \boldsymbol{r}', \delta, 1, K)$
for $i = 1$ to K **do**
    $r_{r'_i} \leftarrow i$
end for
return $\boldsymbol{r}$

Algorithm MMRec($\boldsymbol{r}, \boldsymbol{r}', \delta, i, j$)

if $i < j$ **then**
    $k \leftarrow \lceil (i + j)/2 \rceil$
    $(\boldsymbol{r}, \boldsymbol{r}') \leftarrow \text{MMRec}(\boldsymbol{r}, \boldsymbol{r}', \delta, i, k - 1)$
    $(\boldsymbol{r}, \boldsymbol{r}') \leftarrow \text{MMRec}(\boldsymbol{r}, \boldsymbol{r}', \delta, k, j)$
    for $\ell = i$ to j **do**
        $r_\ell \leftarrow r'_\ell$
    end for
end if

PREFERENCE-BASED RANK ELICITATION

Algorithm MallowsMerge($\boldsymbol{r}, \boldsymbol{r}', \delta, i, j, k$)

$\ell \leftarrow i$, $\ell' \leftarrow k$
for $q = i$ to j **do**
  if $\ell < k$ and $\ell' \leq j$ **then**
    repeat
      observe $o = \mathbb{I}(a_\ell \succ a_{\ell'})$
      $\hat{p}_{\ell,\ell'} \leftarrow \hat{p}_{\ell,\ell'} + o$, $\hat{n}_{\ell,\ell'} \leftarrow \hat{n}_{\ell,\ell'} + 1$
      $c_{\ell,\ell'} \leftarrow \left(\frac{1}{2 n_{\ell,\ell'}} \log \left(\frac{4 n_{\ell,\ell'} C_K}{\delta} \right) \right)^{-1/2}$
    until $1/2 \notin [\hat{p}_{\ell,\ell'} \pm c_{\ell,\ell'}]$
    if $1/2 < \hat{p}_{\ell,\ell'} - c_{\ell,\ell'}$ **then**
      $r'_q \leftarrow r_\ell$, $\ell \leftarrow \ell + 1$
    else
      $r'_q \leftarrow r_{\ell'}$, $\ell' \leftarrow \ell' + 1$
    end if
  else
    if $\ell < k$ **then**
      $r'_q \leftarrow r_\ell$, $\ell \leftarrow \ell + 1$
    else
      $r'_q \leftarrow r_{\ell'}$, $\ell' \leftarrow \ell' + 1$
    end if
  end if
end for
return $\boldsymbol{r}$

PREFERENCE-BASED RANK ELICITATION

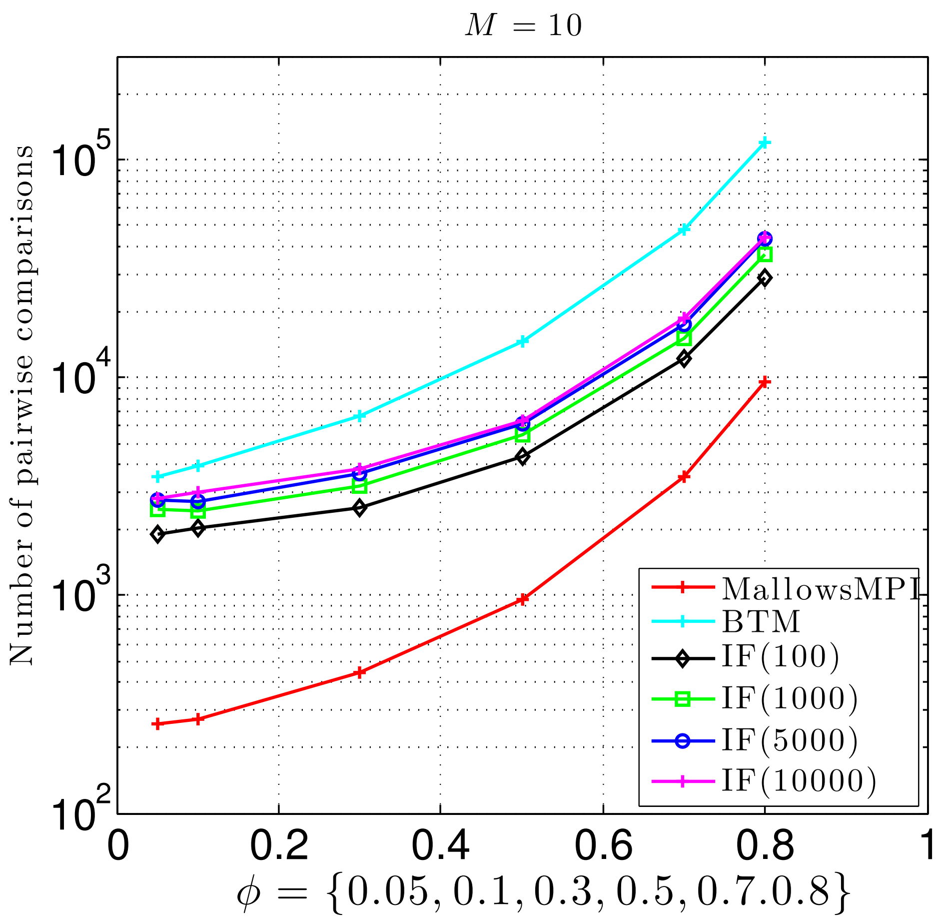
INTELLIGENT SYSTEMS

- For the problem of **finding the best arm**, Busa-Fekete et al. (2014) devise an algorithm that is similar to the one used for finding the largest element in an array.
- Again, two arms a_i and a_j are compared until significance is achieved.
- **Theorem:** MallowsMPI finds the most preferred arm with probability at least $1 - \delta$ for a sample complexity that is of the form

$$\mathcal{O}\left(\frac{K}{\rho^2}\log\frac{K}{\delta\rho}\right) \quad ,$$

where $\rho = \frac{1-\phi}{1+\phi}$.

EMPIRICAL VALIDATION



Sample complexity for K=10, δ = 0.05 and different values of ϕ.

Theorem: Assume that the ranking distribution is Mallows. Then, for any $\epsilon > 0$ and $0 < \delta < 1$, MallowsKLD returns parameter estimates $\widehat{\pi}_0$ and $\widehat{\theta}$ for which

$$\mathrm{KL}\left(\mathbf{P}(\cdot \,|\, \pi_0, \theta), \mathbf{P}\left(\cdot \,|\, \widehat{\pi}_0, \widehat{\theta}\right) \right) =$$

$$= \sum_{\pi \in \mathcal{S}_M} \mathbf{P}(\pi \,|\, \pi_0, \theta) \log \frac{\mathbf{P}(\pi \,|\, \pi_0, \theta)}{\mathbf{P}\left(\pi \,|\, \widehat{\pi}_0, \widehat{\theta}\right)} < \epsilon \ ,$$

and the number of pairwise comparisons requested by the algorithm is

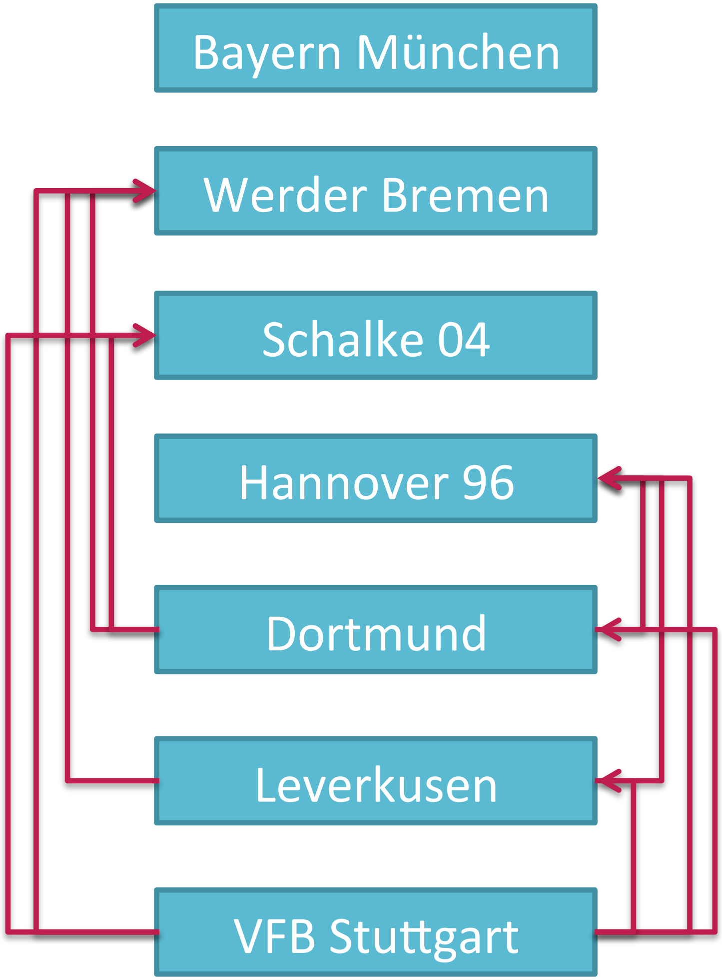
$$\mathcal{O}\left(\frac{M \log_2 M}{\rho^2} \log \frac{M \log_2 M}{\delta \rho} + \frac{1}{D(\epsilon)^2} \log \frac{1}{\delta D(\epsilon)} \right) \ ,$$

where $\rho = \frac{1-\phi}{1+\phi}$, $\phi = \exp(-\theta)$ and

$$D(\epsilon) = \frac{\phi}{6(\phi+1)^2} \left(1 - \frac{2}{\exp\left(\frac{\epsilon}{M(M-1)}\right) + 1} \right) \ .$$

EMPIRICAL VALIDATION

- In general, the approach performs quite well compared to baselines.
- However, it may fail if the underlying data is not enough „Mallowsian“ ...



SUMMARY & CONCLUSION

- Growing interest in **preference learning**
- **Online preference learning** not yet strongly developed
- Preference-based online learning with multi-armed bandits (PB-MAB):
  - **emerging** research topic,
  - no complete and **coherent framework** so far,
  - many **open questions and problems** (e.g., necessary assumptions on preference relation to guarantee certain bounds on regret or sample complexity, lower bounds, statistical tests for verifying model assumptions, generalizations to large (structured) set of arms, contextual bandits, adversarial setting, practical applications, etc., ...)

SELECTED LITERATURE (PB-MAB)

INTELLIGENT SYSTEMS

- N. Ailon, K. Hatano, and E. Takimoto. Bandit online optimization over the permutahedron. CoRR, abs/1312.1530, 2014.
- N. Ailon, Z. Karnin, and T. Joachims. Reducing dueling bandits to cardinal bandits. ICML 2014.
- P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47:235-256, 2002.
- **R. Busa-Fekete and E. Hüllermeier. A Survey of Preference-based Online Learning with Bandit Algorithms. Proc. ALT-2014, Int. Conf. Algorithmic Learning Theory, Bled, 2014.**
- **R. Busa-Fekete, E. Hüllermeier, and B. Szorenyi. Preference-based rank elicitation using statistical models: The case of Mallows. ICML 2014.**
- R. Busa-Fekete, B. Szorenyi, and E. Hüllermeier. PAC rank elicitation through adaptive sampling of stochastic pairwise preferences. AAAI 2014.
- R. Busa-Fekete, B. Szorenyi, P. Weng, W. Cheng, and E. Hüllermeier. Top-k selection based on adaptive sampling of noisy preferences. ICML 2013.
- W.W. Cohen, R.E. Schapire and Y. Singer. Learning to order things. J. of Artif. Intelligence Res., 10:243–270, 1999.
- J. Duchi, L. Mackey, and M. Jordan. On the consistency of ranking algorithms. ICML 2010.
- J. Fürnkranz and E. Hüllermeier, editors. Preference Learning. Springer-Verlag, 2011.
- E. Hüllermeier, J. Fürnkranz, W. Cheng, K. Brinker. Label ranking by learning pairwise preferences. Artif. Intell., 172, 2008.
- F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? CIKM 2008.
- T. Urvoy, F. Clerot, R. Feraud, and S. Naamane. Generic exploration and k-armed voting bandits. ICML 2013.
- Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The K-armed dueling bandits problem. Journal of Computer and System Sciences, 78(5):1538-1556, 2012.
- Y. Yue and T. Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. ICML 2009.
- Y. Yue and T. Joachims. Beat the mean bandit. ICML 2011.
- M. Zoghi, S. Whiteson, R. Munos, and M. de Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. ICML 2014.

SELECTED LITERATURE (PL)

INTELLIGENT SYSTEMS

- E. Hüllermeier, J. Fürnkranz, W. Cheng and K. Brinker. Label ranking by learning pairwise preferences. Artificial Intelligence, 172, 2008.
- W. Cheng, J. Hühn and E. Hüllermeier. Decision tree and instance-based learning for label ranking, ICML-09, Montreal, 2009.
- W. Cheng, K. Dembczynski and E. Hüllermeier. Label ranking using the Plackett-Luce model. ICML-10, Haifa, Israel, 2010.
- W. Cheng, W. Waegeman, V. Welker and E. Hüllermeier. Label ranking with partial abstention based on thresholded probabilistic models. NIPS 2012.
- J. Fürnkranz, E. Hüllermeier, W. Cheng, S.H. Park. Preference-Based Reinforcement Learning: A Formal Framework and a Policy Iteration Algorithm. Machine Learning, 89, 2012.
- E. Hüllermeier and J. Fürnkranz. On predictive accuracy and risk minimization in pairwise label ranking. J. Computer and System Sciences , 76, 2010.
- E. Hüllermeier and P. Schlegel. Preference-based CBR: First steps toward a methodological framework. ICCBR-11, London, 2011.
- R. Akrour, M. Schoenauer, M. Sebag. Preference-Based Policy Learning, ECML 2011.
- W.W. Cohen, R.E. Schapire and Y. Singer. Learning to order things. Journal of Artificial Intelligence Research, 10:243–270, 1999.
- O. Dekel, C.D. Manning, Y. Singer. Log-Linear Models for Label Ranking. NIPS-2003.
- D. Goldberg, D. Nichols, B.M. Oki and D. Terry. Using collaborative filtering to weave and information tapestry. Communications of the ACM, 35(12):61–70, 1992.
- S. Har-Peled, D. Roth and D. Zimak. *Constraint classification: A new approach to multiclass classification*. Proc. ALT-2002.
- D.R. Hunter. MM algorithms for generalized Bradley-Terry models. The Annals of Statistics , 32(1):384–406, 2004.
- S. Vembu and T. Gärtner. Label ranking: a survey. In: Preference Learning. J. Fürnkranz and E. Hüllermeier (eds.), Springer-Verlag, 2011.