AN INTRODUCTION TO PREFERENCE LEARNING

Eyke Hüllermeier
Department of Mathematics and Computer Science
Marburg University, Germany

Philipps Universität Marburg

Paris Dauphine, 29-JAN-2014

AGENDA

1. Introduction
   - An example: The Choquet integral
   - Preference modeling, elicitation, and learning
   - Preference Learning: A first glimpse
   - Machine Learning
2. Ranking Problems
3. Model-based Preference Learning
4. Summary & Outlook

AGGREGATION OF CRITERIA

	Math	CS	Statistics	English	score
$\boldsymbol{x}_1$	16	17	12	19	0.7
$\boldsymbol{x}_2$	10	12	18	9	0.4
$\boldsymbol{x}_3$	19	10	18	10	0.6
	…	…	…	…	…
$\boldsymbol{x}_n$	8	18	10	18	0.5

ADDITIVE & NON-ADDITIVE MEASURES

Non-additive measure (capacity) $\mu : 2^X \rightarrow [0,1]$: We require ...

- $\mu(\emptyset) = 0$, $\mu(X) = 1$ normalization
- $\mu(A) \leq \mu(B)$ for $A \subset B \subseteq X$ monotonicity
- [$\mu(A \cup B) = \mu(A) + \mu(B)$ for $A \cap B = \emptyset$] not necessarily additivity

ADDITIVE & NON-ADDITIVE MEASURES

Non-additive measures allow for modeling **interaction** between criteria:

- Positive (synergy): $\mu(A \cup B) > \mu(A) + \mu(B)$
- Negative (redundancy): $\mu(A \cup B) < \mu(A) + \mu(B)$

In a machine learning context: criteria = attributes/features

$\mu(A) =$ **joint importance** of the feature subset A

$\neq$ sum of individual importance degrees

DISCRETE CHOQUET INTEGRAL

The **discrete Choquet integral** of $f : X \to \mathbb{R}_+$ with respect to μ is defined as follows:

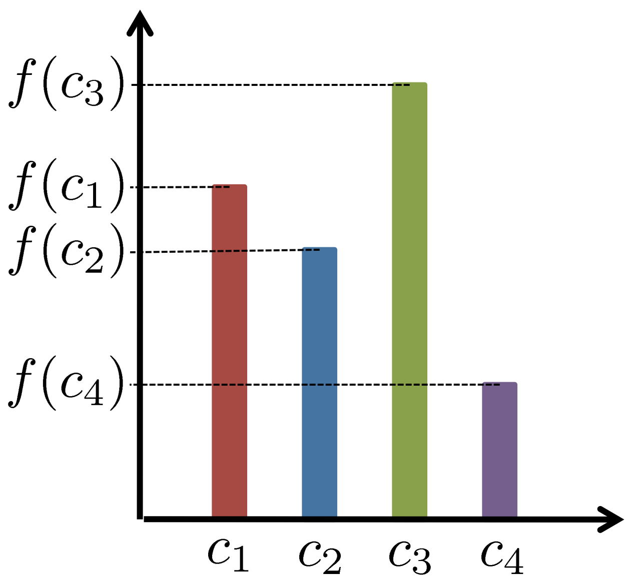
$$\mathcal{C}_\mu(f) = \sum_{i=1}^{m} \left(f(x_{(i)}) - f(x_{(i-1)})\right) \cdot \mu\left(A_{(i)}\right) \ ,$$

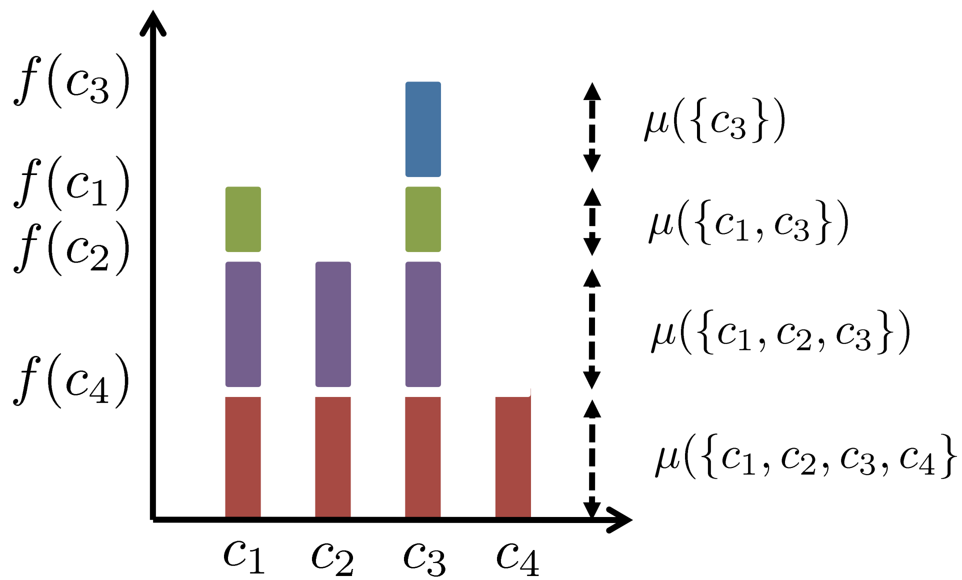
where $(\cdot)$ is a permutation of $\{1, \ldots, m\}$ such that
$0 \leq f(x_{(1)}) \leq f(x_{(2)}) \leq \ldots \leq f(x_{(m)})$, and $A_{(i)} = \{x_{(i)}, \ldots, x_{(m)}\}$.

Special cases:

- min
- max
- weighted mean (additive measure)
- OWA

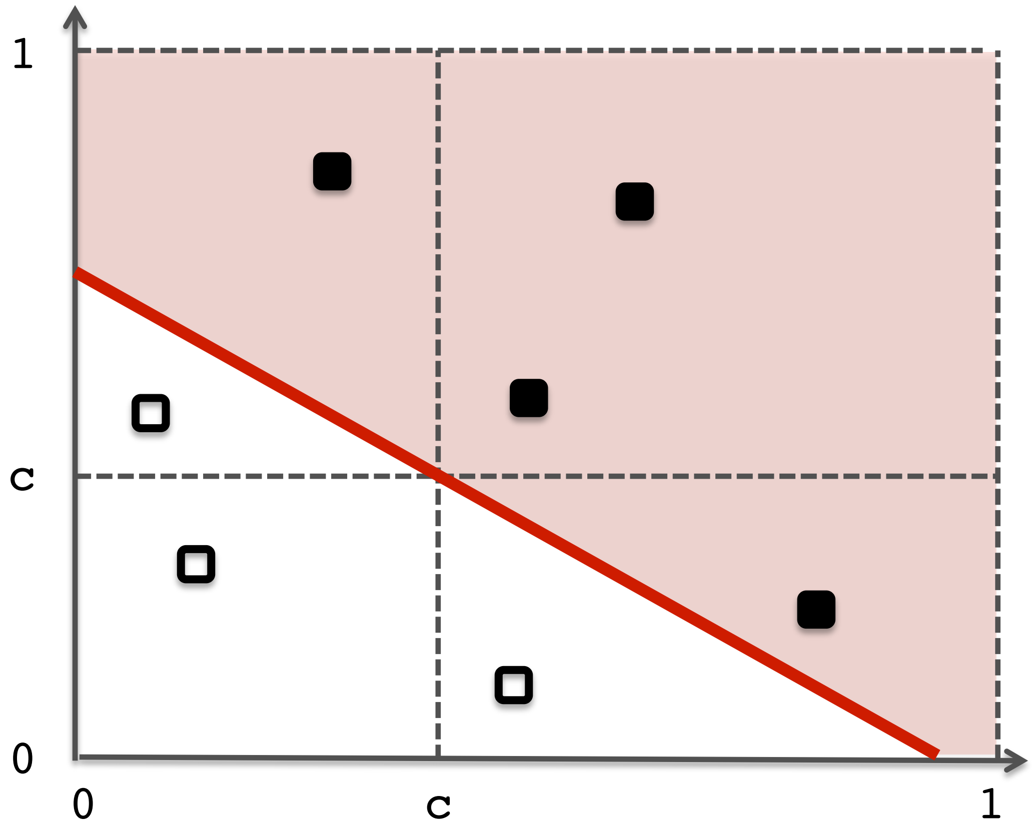
DISCRETE CHOQUET INTEGRAL





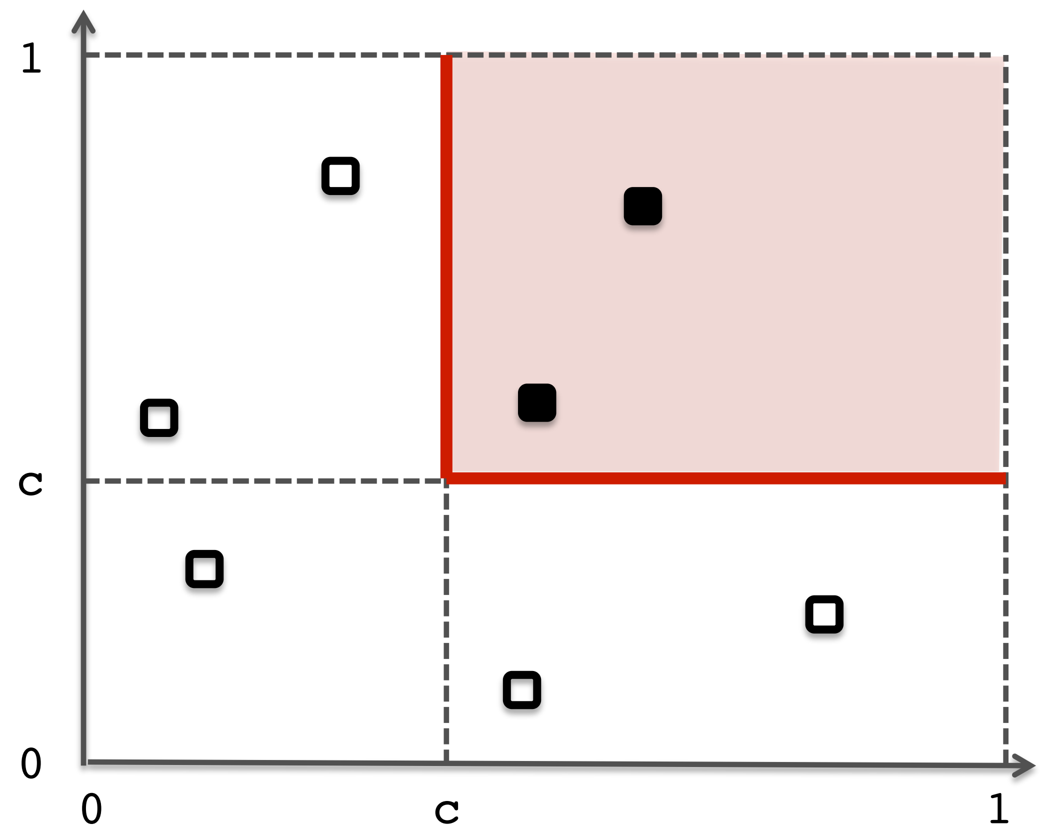
$$\mathcal{C}_\mu(f) = \sum_{i=1}^{4} w_i \cdot f(c_i) = \sum_{i=1}^{4} \mu(\{c_i\}) \cdot f(c_i) \qquad \mathcal{C}_\mu(f) = \sum_{i=1}^{4} \mu\Big(A_{(i)}\Big) \cdot \Big(f(c_{(i)}) - f(c_{(i-1)})\Big)$$

DECISION BOUNDARY IN TWO DIMENSIONS



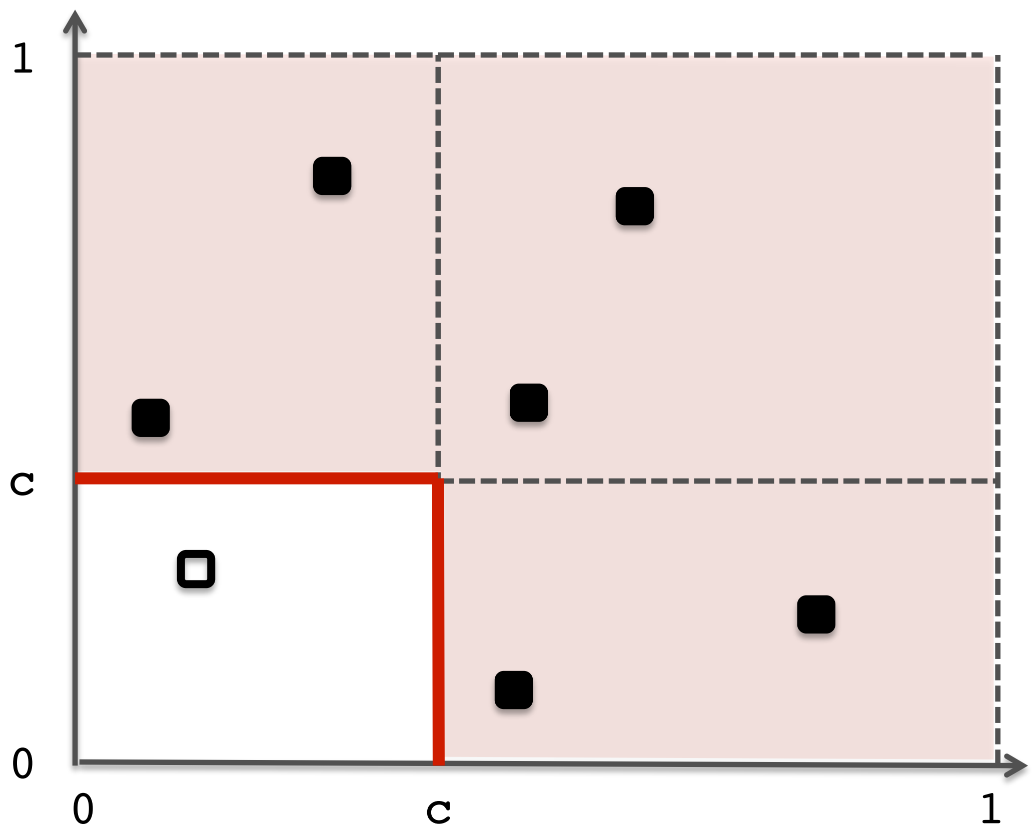
$$(x, y) \mapsto \mathbb{I}\Big(\alpha x + (1-\alpha)y > c\Big)$$

DECISION BOUNDARY IN TWO DIMENSIONS



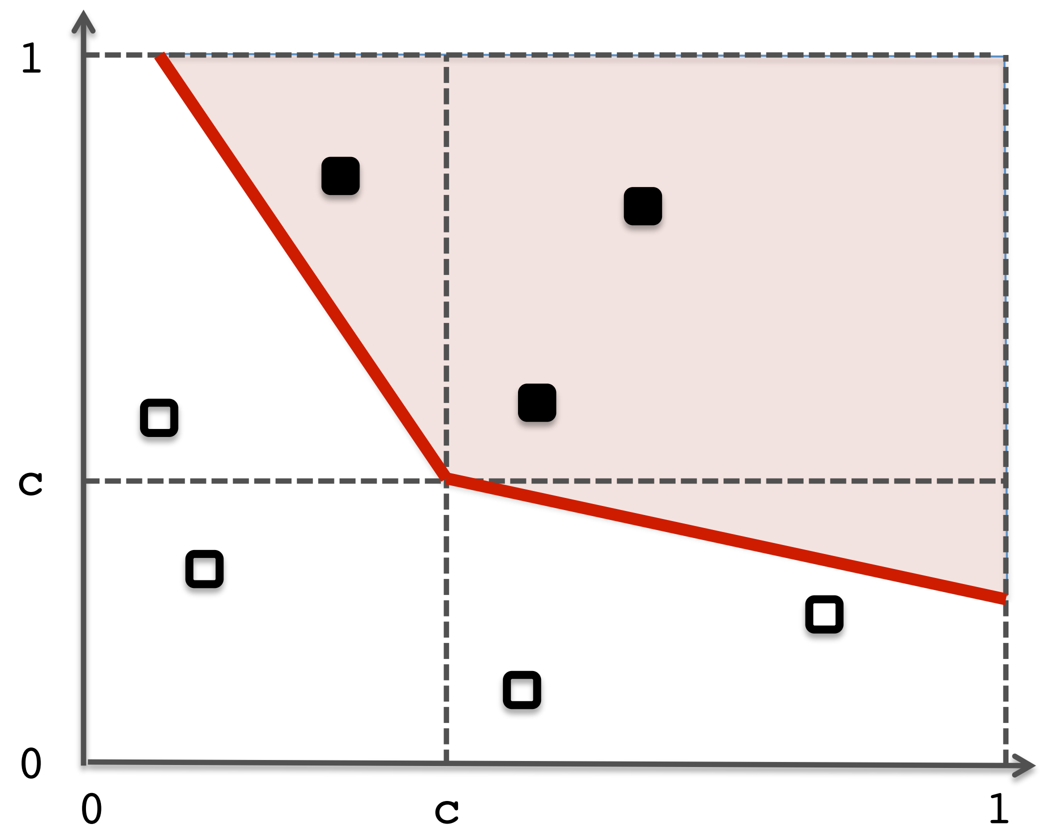
$$(x, y) \mapsto \mathbb{I}\Big(\min(x, y) > c\Big)$$

DECISION BOUNDARY IN TWO DIMENSIONS

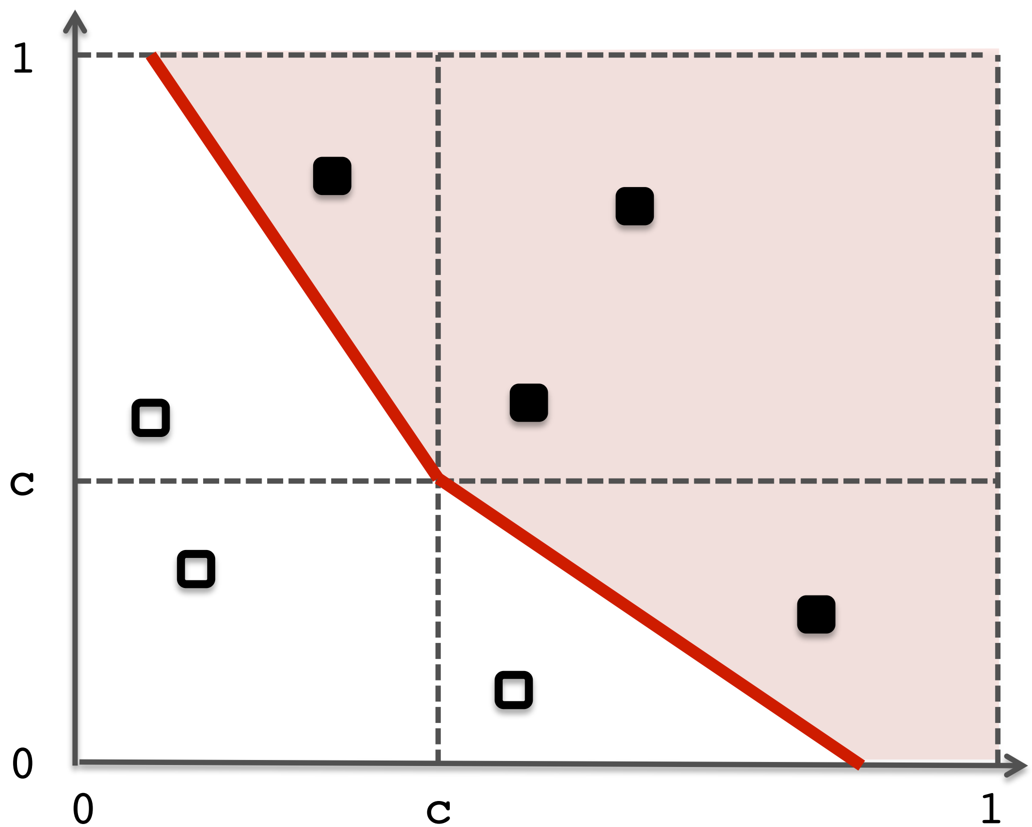


$$(x, y) \mapsto \mathbb{I}\Big(\max(x, y) > c\Big)$$

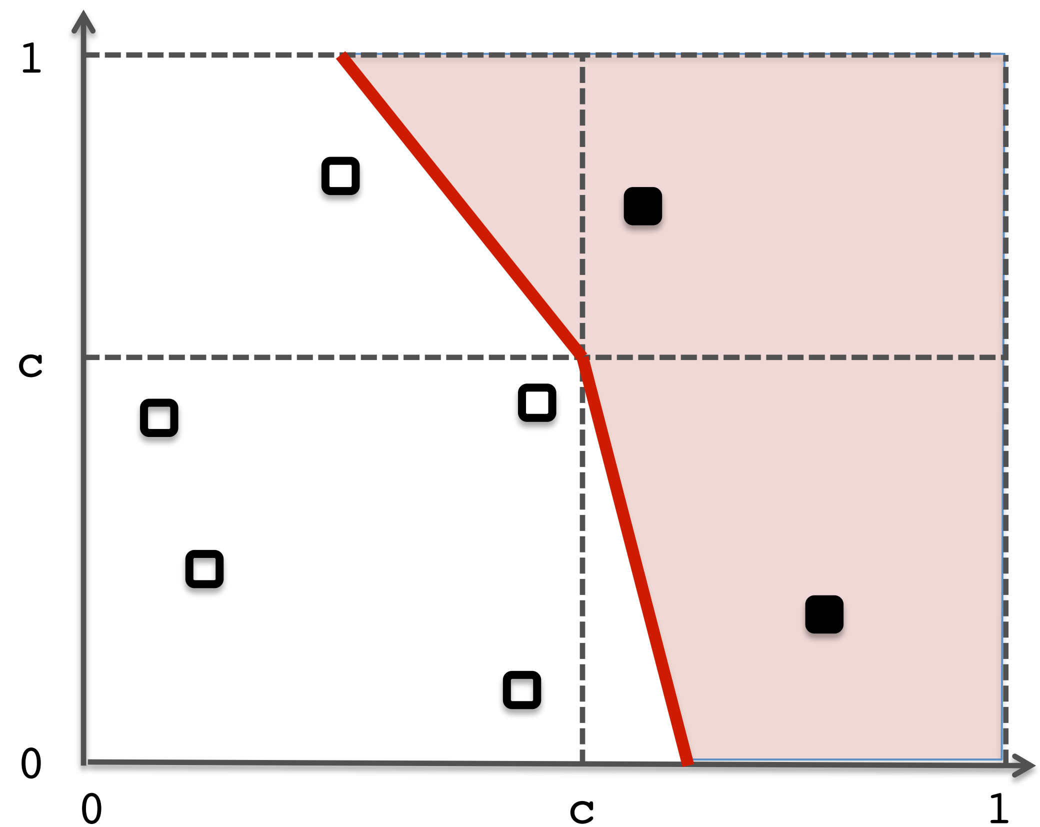
DECISION BOUNDARY IN TWO DIMENSIONS



DECISION BOUNDARY IN TWO DIMENSIONS



DECISION BOUNDARY IN TWO DIMENSIONS

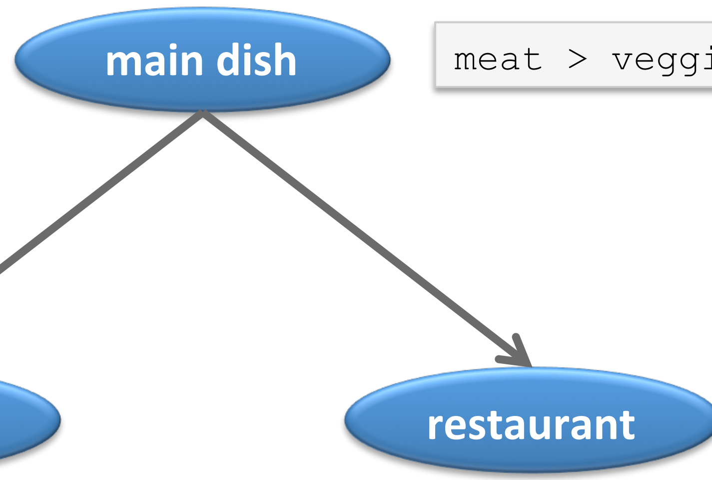


AGENDA

1. Introduction
   - An example: The Choquet integral
   - **Preference modeling, elicitation, and learning**
   - Preference Learning: A first glimpse
   - Machine Learning
2. Ranking Problems
3. Model-based Preference Learning
4. Summary & Outlook

CP NETWORKS

CONDITIONAL PREFERENCE NETWORKS: Compact representation of a partial order relation, exploiting conditional independence of preferences on attribute values.




```
meat > veggie > fish
```

```
meat:    red wine   > white wine
veggie:  red wine   > white wine
fish:    white wine > red wine
```

```
meat:    Italian > Chinese
veggie:  Chinese > Italian
fish:    Chinese > Italian
```

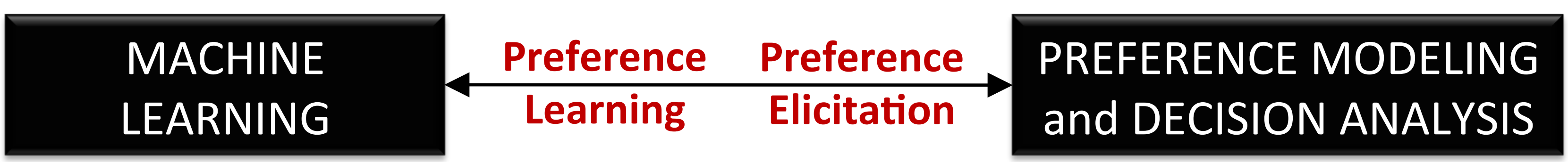
MODELING VS. ELICITATION AND LEARNING

- **Preference modeling:** The model is completely specified by an expert or the decision maker himself.
  - Example Choquet integral: Normalization of features, specification of fuzzy measure, ...
  - Example CP network: Qualitative structure, conditional preferences, ...

- **Preference elicitation/learning:** Parts of the model are specified with the help of external information.

PREFERENCE LEARNING VS. ELICITATION

- typically no user interaction
- holistic judgements
- fixed preferences but noisy data
- regularized models
- weak model assumptions, flexible (instead of axiomatically justified) model classes
- diverse types of training information
- computational aspects: massive data, scalable methods
- focus on predictive accuracy (expected loss)



computer science
artificial intelligence

operations research
social sciences (voting and choice theory)
economics and decision theory

RANKING

TRAINING DATA:

$$
\begin{array}{ccc}
(18, 17, 12, 10) & \succ & (16, 19, 10, 10) \\
(17, 12, 18, 12) & \succ & (15, 14, 16, 11) \\
(12, 19, 18, 18) & \succ & (16, 16, 15, 17) \\
\cdots & \succ & \cdots
\end{array}
$$

The goal might be to find a Choquet integral whose utility degrees tend to agree with the observed pairwise preferences!

ORDINAL CLASSIFICATION / SORTING

TRAINING DATA:

$$\begin{array}{rcl} (12, 17, 11, 8) & \to & ** \\ (19, 15, 17, 16) & \to & *** \\ (9, 12, 14, 10) & \to & * \\ \cdots & \to & \cdots \end{array}$$

The goal might be to find a Choquet integral whose utility degrees tend to agree with the observed classifications!

ORDINAL CLASSIFICATION / SORTING

TRAINING DATA:

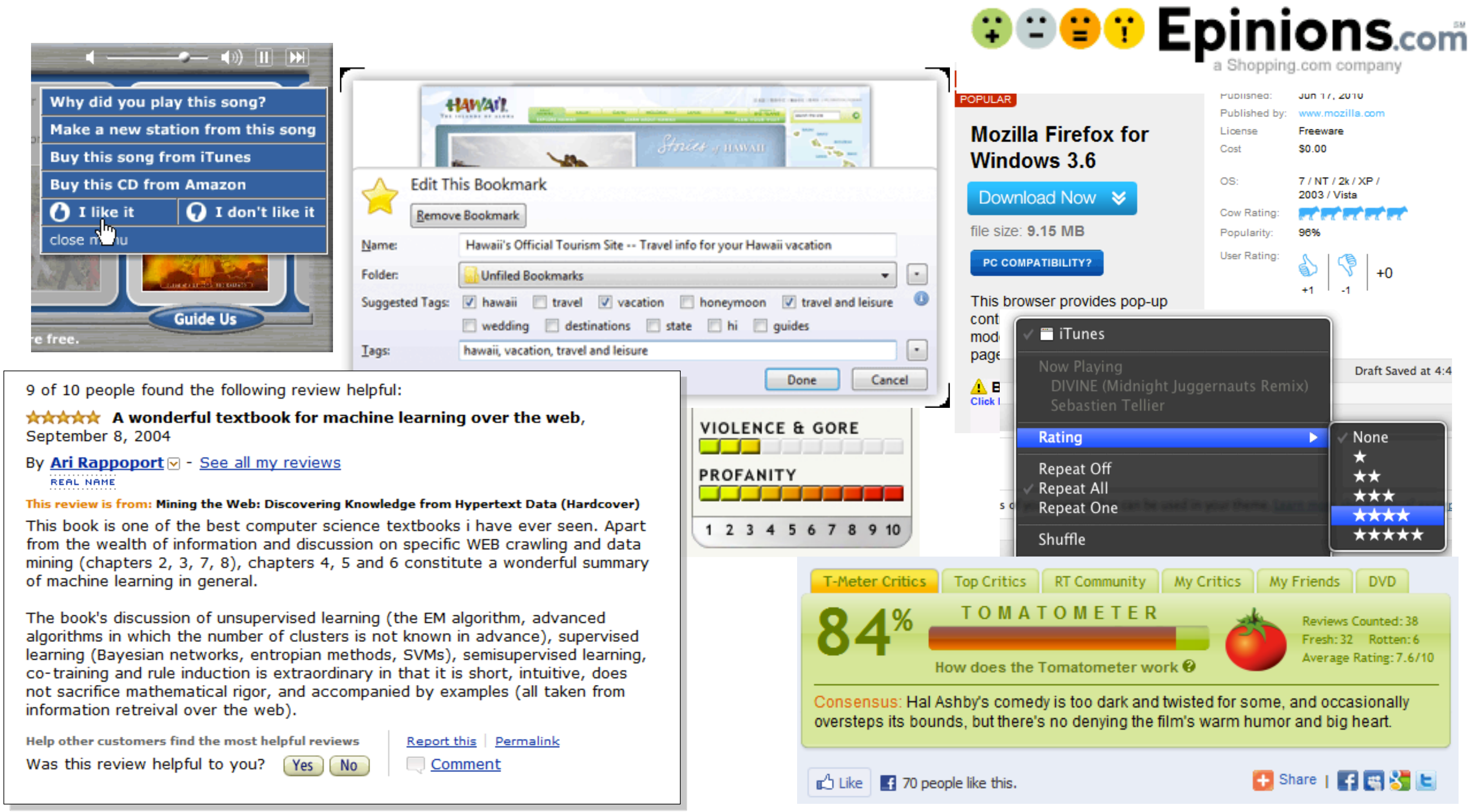
$$\begin{array}{rcl} (12, 17, 11, 8) & \rightarrow & 0.6 \\ (19, 15, 17, 16) & \rightarrow & 0.9 \\ (9, 12, 14, 10) & \rightarrow & 0.2 \\ \cdots & \rightarrow & \cdots \end{array}$$

The goal might be to find a Choquet integral whose utility degrees tend to agree with the observed scores!

AGENDA

1. Introduction
   - An example: The Choquet integral
   - Preference modeling, elicitation, and learning
   - **Preference Learning: A first glimpse**
   - Machine learning
2. Ranking Problems
3. Model-based Preference Learning
4. Summary & Outlook

PREFERENCES ARE UBIQUITOUS



PREFERENCES ARE UBIQUITOUS

| 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 ...

KNOWLEDGE DISCOVERY AND BIG DATA

Companies (and agencies) are collecting lots of data ...

... hoping to discover something useful in it!

ANTICIPATORY SHIPPING AT AMAZON

Neues Patent: Amazon will schon vor der Bestellung liefern

REUTERS

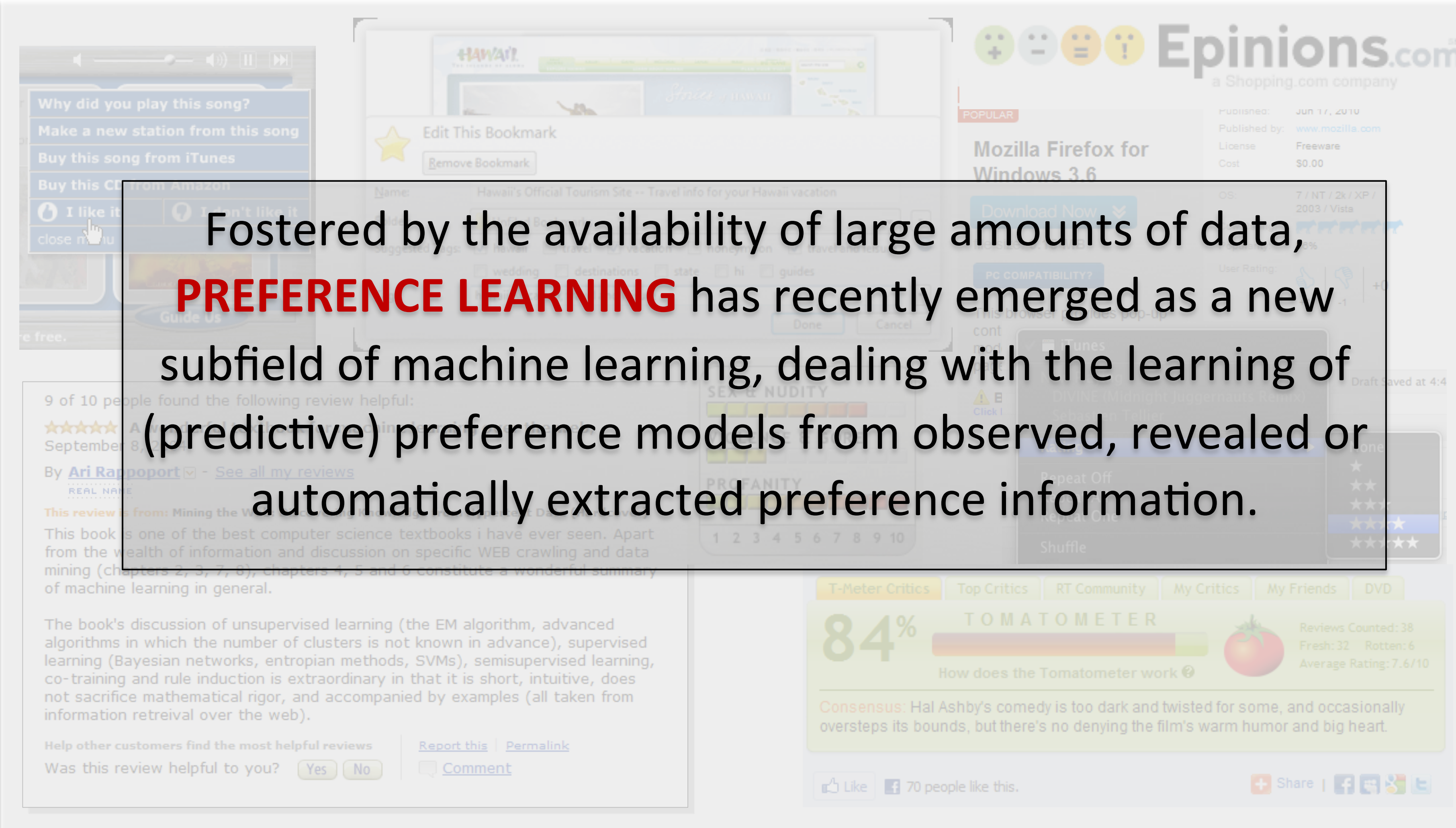
Arbeiterin im Amazon-Versandzentrum bei Berlin: Liefern auf Verdacht

Bevor die Kunden ihre Wahl treffen, soll das passende Produkt schon auf dem Weg in ihre Richtung sein. Amazon wurde ein neues Patent zugesprochen, das ebenso pfiffig wie gruselig klingt.

Noch bevor ein Kunde überhaupt den Button "Kaufen" anklickt, soll die für ihn passende Ware schon auf dem Weg in Richtung seiner Wohnung sein. Dem Versandhändler Amazon wurde ein Patent (PDF) zugesprochen, das einen "vorausschauenden Versand" ("anticipatory shipping") ermöglichen soll. Das heißt: Bestimmte Waren werden schon einmal an ein Versandzentrum geschickt, in dessen Nähe sich ein oder mehrere Kunden höchstwahrscheinlich für das Produkt interessieren. Wird es dann schließlich bestellt, ist es umso schneller beim Empfänger.

Wie Amazon das herausfinden will, erklärt das "Wall Street Journal": Ausgewertet werden könnten demnach frühere Bestellungen, Umtäusche, Wunschzettel bei Amazon, der Inhalt der Einkaufswagen - und sogar, wie lange ein Kunde mit dem Mauszeiger auf einer Produktbeschreibung verweilt.

PREFERENCES ARE UBIQUITOUS



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.

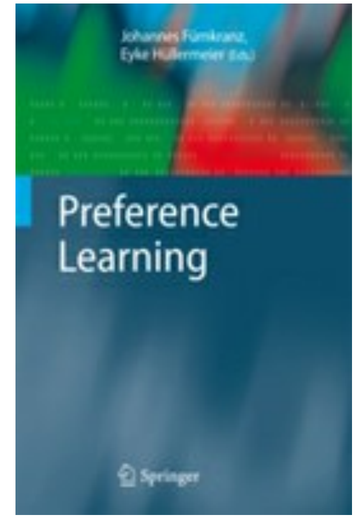
PREFERENCE LEARNING IS AN ACTIVE FIELD

- NIPS–01: New Methods for Preference Elicitation
- NIPS–02: Beyond Classification and Regression: Learning Rankings, Preferences, Equality Predicates, and Other Structures
- KI–03: Preference Learning: Models, Methods, Applications
- NIPS–04: Learning with Structured Outputs
- NIPS–05: Workshop on Learning to Rank
- IJCAI–05: Advances in Preference Handling
- SIGIR 07–10: Workshop on Learning to Rank for Information Retrieval
- **ECML/PDKK 08–10: Workshop on Preference Learning**
- NIPS–09: Workshop on Advances in Ranking
- American Institute of Mathematics Workshop in Summer 2010: The Mathematics of Ranking
- NIPS-11: Workshop on Choice Models and Preference Learning
- EURO-12: Special Track on Preference Learning
- **ECAI-12: Workshop on Preference Learning: Problems and Applications in AI**
- **Forthcoming: Dagstuhl Seminar on Preference Learning (2014)**

PREFERENCE LEARNING IS AN ACTIVE FIELD

Tutorials:

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



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

PREFERENCES IN AI

User preferences play a key role in various fields of application:

RECOMMENDER SYSTEMS

COMPUTATIONAL ADVERTISING

COMPUTER GAMES

ELECTRONIC COMMERCE

AUTONOMOUS AGENTS

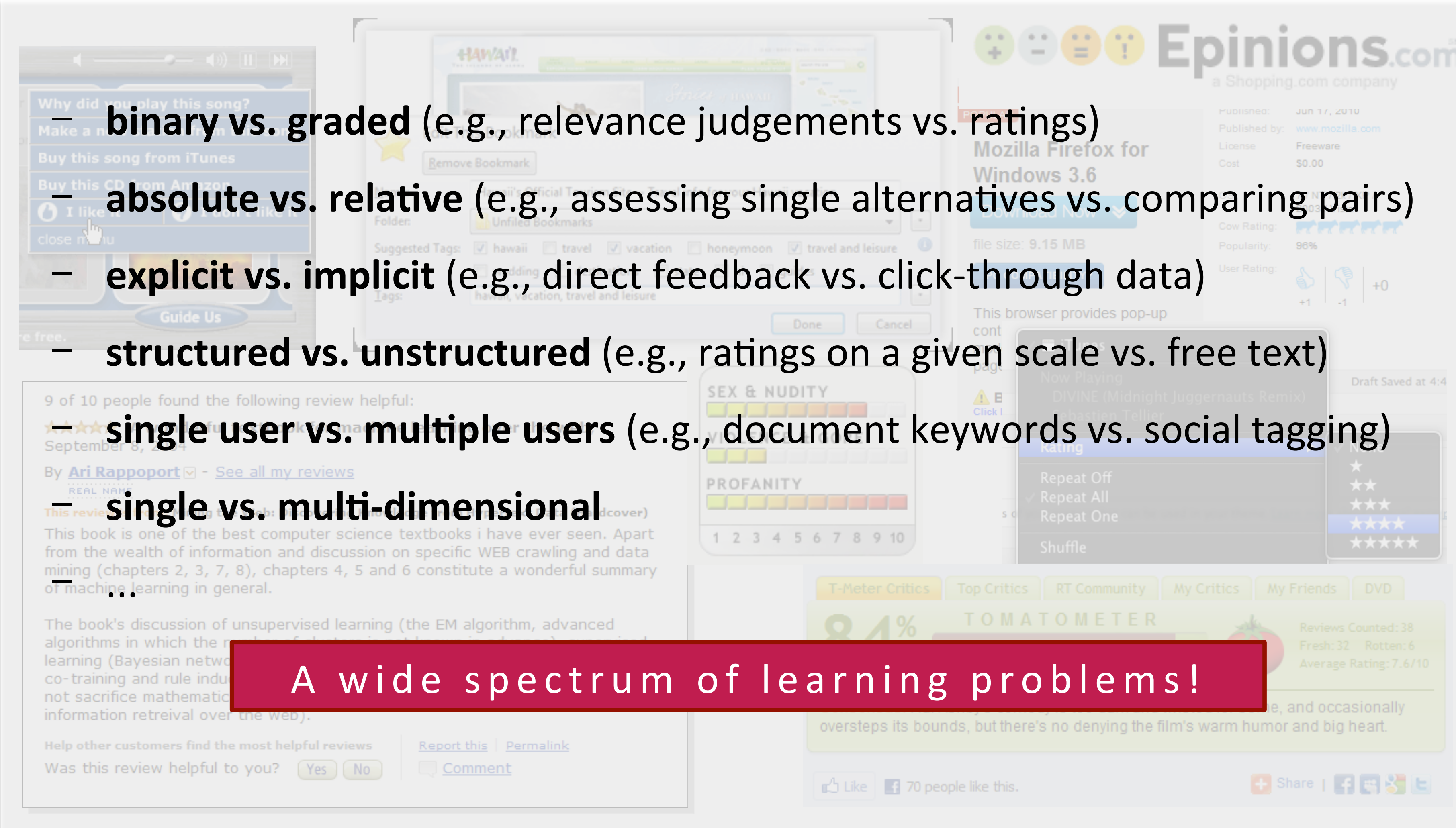
ADAPTIVE USER INTERFACES

ADAPTIVE RETRIEVAL SYSTEMS

PREFERENCES IN AI RESEARCH:

- **preference representation** (CP nets, GAU networks, logical representations, fuzzy constraints, ...)
- **reasoning** with preferences (decision theory, constraint satisfaction, non-monotonic reasoning, ...)
- **preference acquisition** (preference elicitation, **preference learning**, ...)

PREFERENCES LEARNING SETTINGS



- **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**
- ...

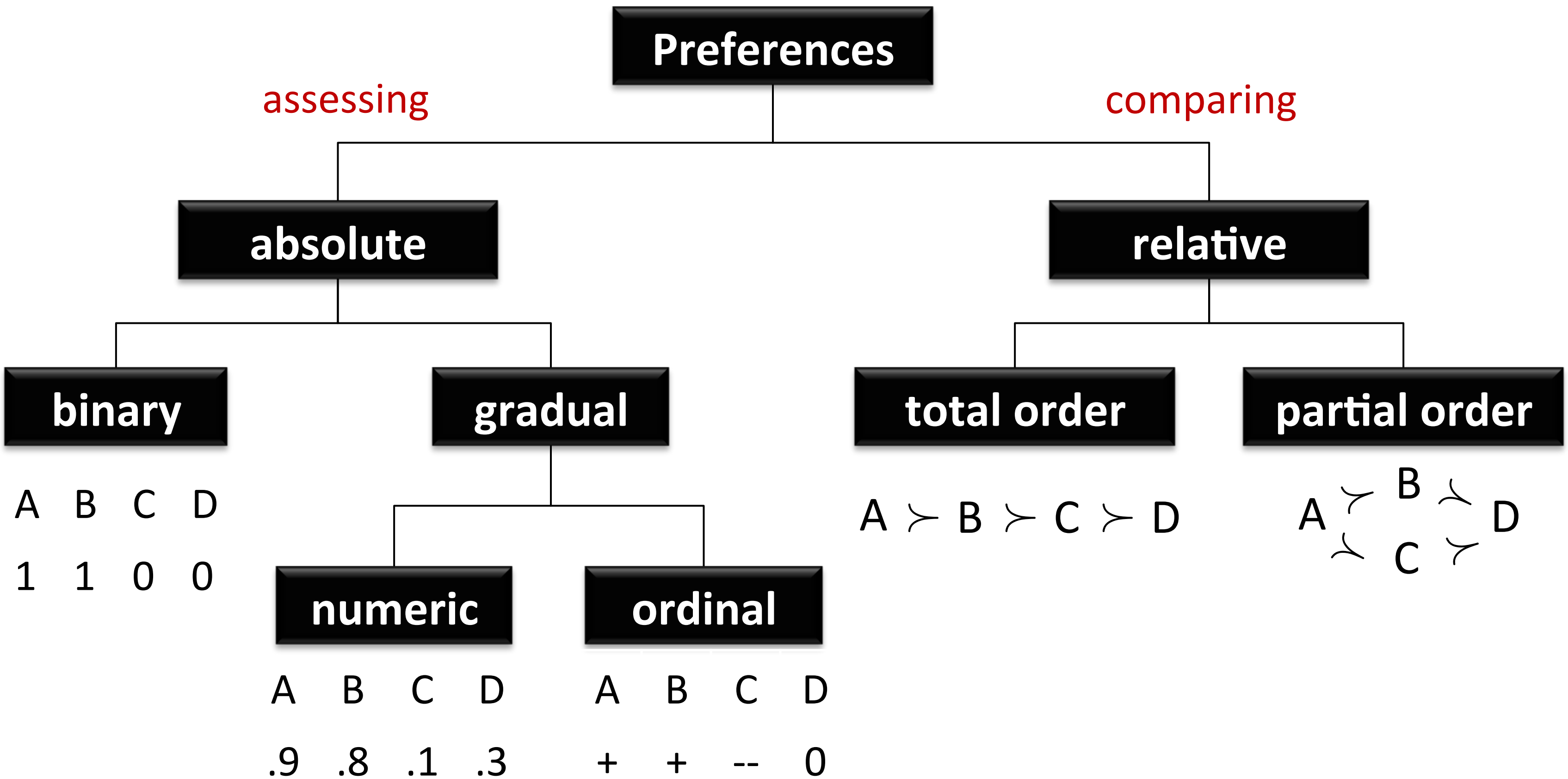
A wide spectrum of learning problems!

PREFERENCE LEARNING

Preference learning problems can be distinguished along several **problem dimensions**, including

- **representation of preferences, type of preference model**:
  - utility function (ordinal, numeric),
  - preference relation (partial order, ranking, …),
  - logical representation, …
- **description of individuals/users and alternatives/items**:
  - identifier, feature vector, structured object, …
- **type of training input**:
  - direct or indirect feedback,
  - complete or incomplete relations,
  - utilities, …
- …

PREFERENCE LEARNING



→ (ordinal) regression

→ classification/ranking

OBJECT RANKING [Cohen et al., 1999]

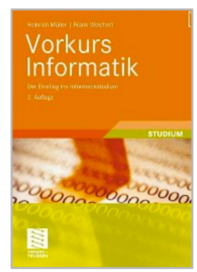
TRAINING

$$(0.74, 1, 25, 165) \succ (0.45, 0, 35, 155)$$
$$(0.47, 1, 46, 183) \succ (0.57, 1, 61, 177)$$
$$(0.25, 0, 26, 199) \succ (0.73, 0, 46, 185)$$

Pairwise preferences between objects

$\succ$

$\succ$



OBJECT RANKING [Cohen et al. 99]

PREDICTION (ranking a new set of objects)

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

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

$\mathcal{Q} =$

COLLABORATIVE FILTERING [Goldberg et al., 1992]

PRODUCTS

USERS

	P1	P2	P3	…	P38	…	P88	P89	P90
U1	★		★★★	…		…		★★★	
U2		★★	★	…		…	★		
…				…		…			
U46	?	★★	?	…	?	…	?	?	★★★
…				…		…			
U98	★★★			…		…	★★★		
U99			★	…		…	★★		

PREFERENCE LEARNING TASKS

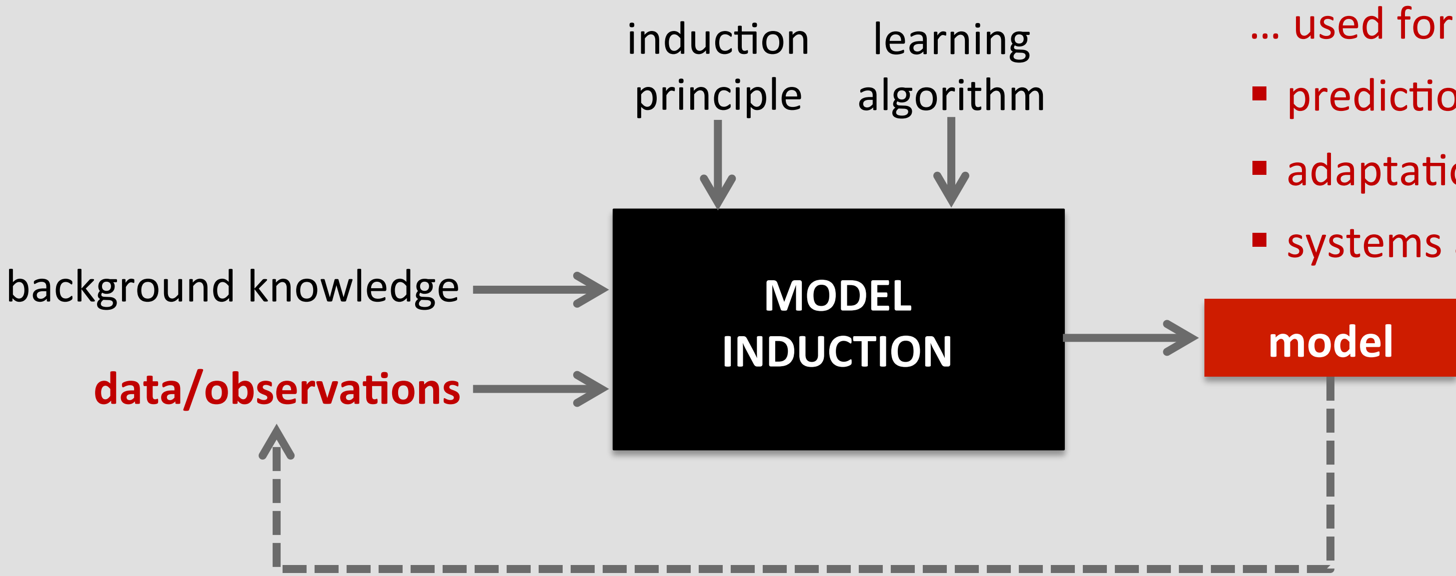
	OBJECT RANKING	COLLABORATIVE FILTERING
product description	features	identifier
preference description	relative	absolute
predictions	ranking	utility degrees
number of users/models	single	many

AGENDA

1. Introduction
   – An example: The Choquet integral
   – Preference modeling, elicitation, and learning
   – Preference Learning: A first glimpse
   – **Machine Learning**
2. Ranking Problems
3. Model-based Preference Learning
4. Summary & Outlook

MACHINE LEARNING FOR PREDICTIVE MODELING

SUPERVISED LEARNING: Algorithms and methods for discovering (alleged) dependencies and regularities in a domain of interest, expressed through appropriate models, from specific observations or examples.



... used for

- prediction, classification
- adaptation, control
- systems analysis

SPECIFICATION OF A MACHINE LEARNING PROBLEM

- What kind of **training data** is offered to the learning algorithm?
- What **type of model** (prediction) is the learner supposed to produce?

$$h : \mathcal{X} \to \mathcal{Y}$$

- What is the nature of the **ground truth**, and how is a model assessed?

LOSS FUNCTION

$$\mathcal{L} : \ \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$$

$$(y, y^*) \mapsto \begin{array}{l}\text{penalty for predicting} \\ y \text{ if the true outcome is } y^*\end{array}$$

SPECIFICATION OF A MACHINE LEARNING PROBLEM

- What kind of **training data** is offered to the learning algorithm?
- What **type of model** (prediction) is the learner supposed to produce?

$$h : \mathcal{X} \rightarrow \mathcal{Y}$$

- What is the nature of the **ground truth**, and how is a model assessed?

$$\mathcal{R}(h) = \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{L}\left(h(\boldsymbol{x}), y\right) \, d\,\mathbf{P}(X,Y)$$

risk ≈ average penalty caused by the model's predictions

unknown data-generating process

- Other criteria, such as complexity ...

MODEL INDUCTION

A simple setting of supervised learning: Given (i.i.d.) training data

$$\mathcal{D} = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \} \subset (\mathcal{X} \times \mathcal{Y})^n$$

and a hypothesis space $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$, induce a model

$$h^* \in \operatorname*{arg\,min}_{h \in \mathcal{H}} \int_{\mathcal{X} \times \mathcal{Y}} \underbrace{\ell\,(h(x), y)}_{\text{loss function}} \; d\, \underbrace{\mathrm{P}(X, Y)}_{\text{unknown data-generating process}}$$

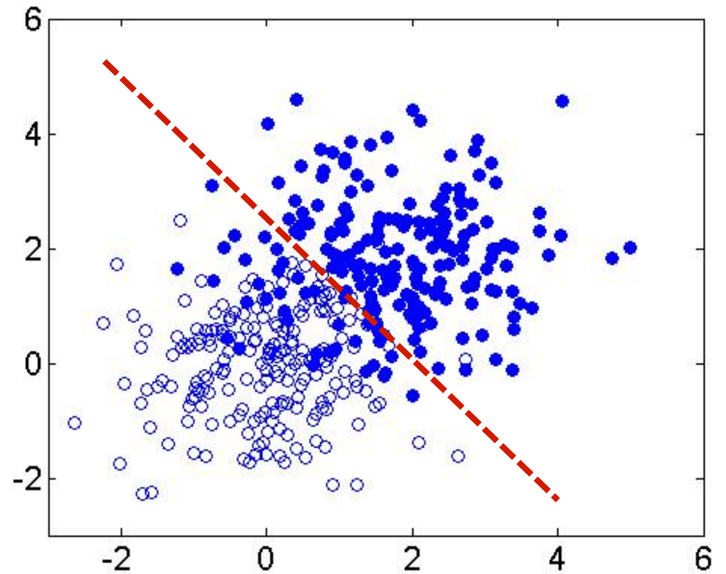
Many other settings exist, such as online learning, semi-supervised learning, active learning, multi-task learning, etc.

EXAMPLE: CLASSIFIER LEARNING

$$\mathcal{X} \subseteq \mathbb{R}^d, \quad \mathcal{Y} = \{y_1, y_2, \ldots, y_k\},$$

$$\ell(\hat{y}, y) = \begin{cases} 0 & \hat{y} = y \\ 1 & \hat{y} \neq y \end{cases}$$

$$\mathcal{H} = \left\{ \boldsymbol{x} \mapsto \mathbb{I}\left(\boldsymbol{x}^\top \alpha + \beta > 0\right) \mid \alpha \in \mathbb{R}^d,\ \beta \in \mathbb{R} \right\}$$



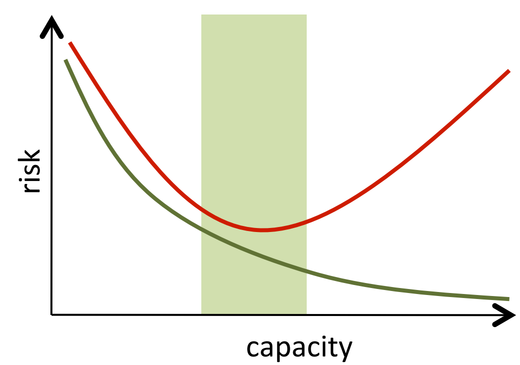
CHOICE OF THE MODEL SPACE

A key to successful learning/generalization is a proper **capacity control**: The model class must be flexible enough (to allow approximation of the pointwise loss-minimizer) but not too flexible (to prevent overfitting the data)

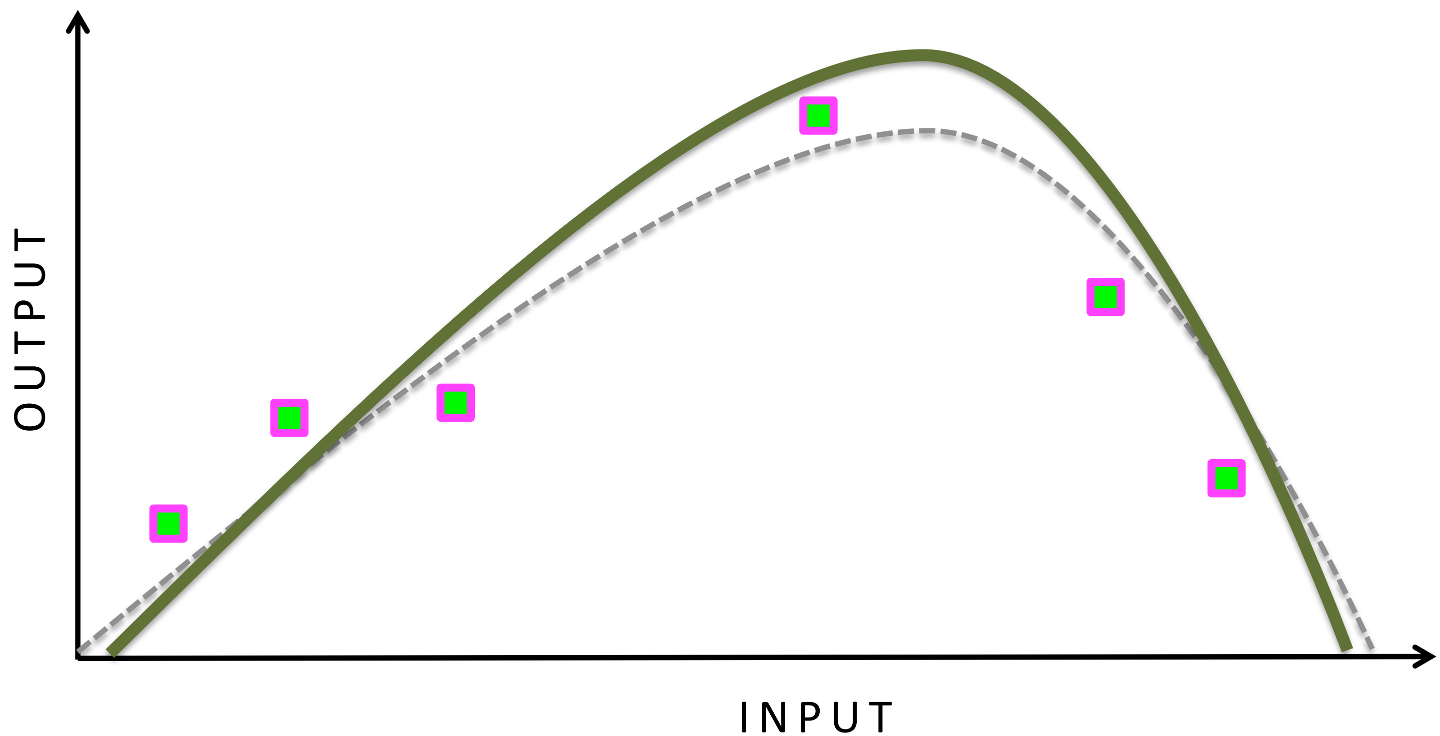
Typical bound on the true risk: With probability $1 - \delta$

$$R(h) \;\leq\; R_{emp}(h) + \sqrt{2\frac{\mathrm{VC}(\mathcal{H})\log\left(\frac{2e|\mathcal{D}|}{\mathrm{VC}(\mathcal{H})}\right) + \log\left(\frac{2}{\delta}\right)}{|\mathcal{D}|}}$$

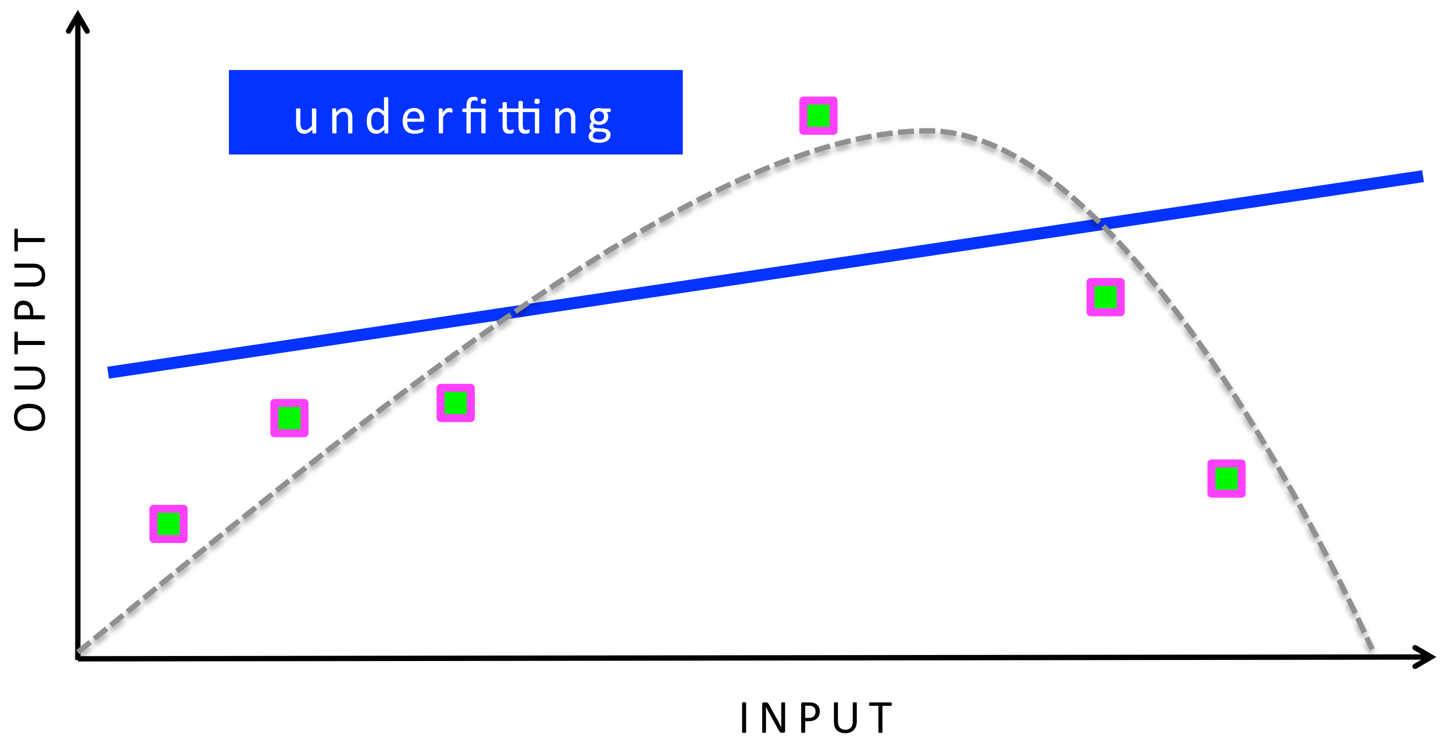
true risk — empirical risk — correction depending on capacity and sample size



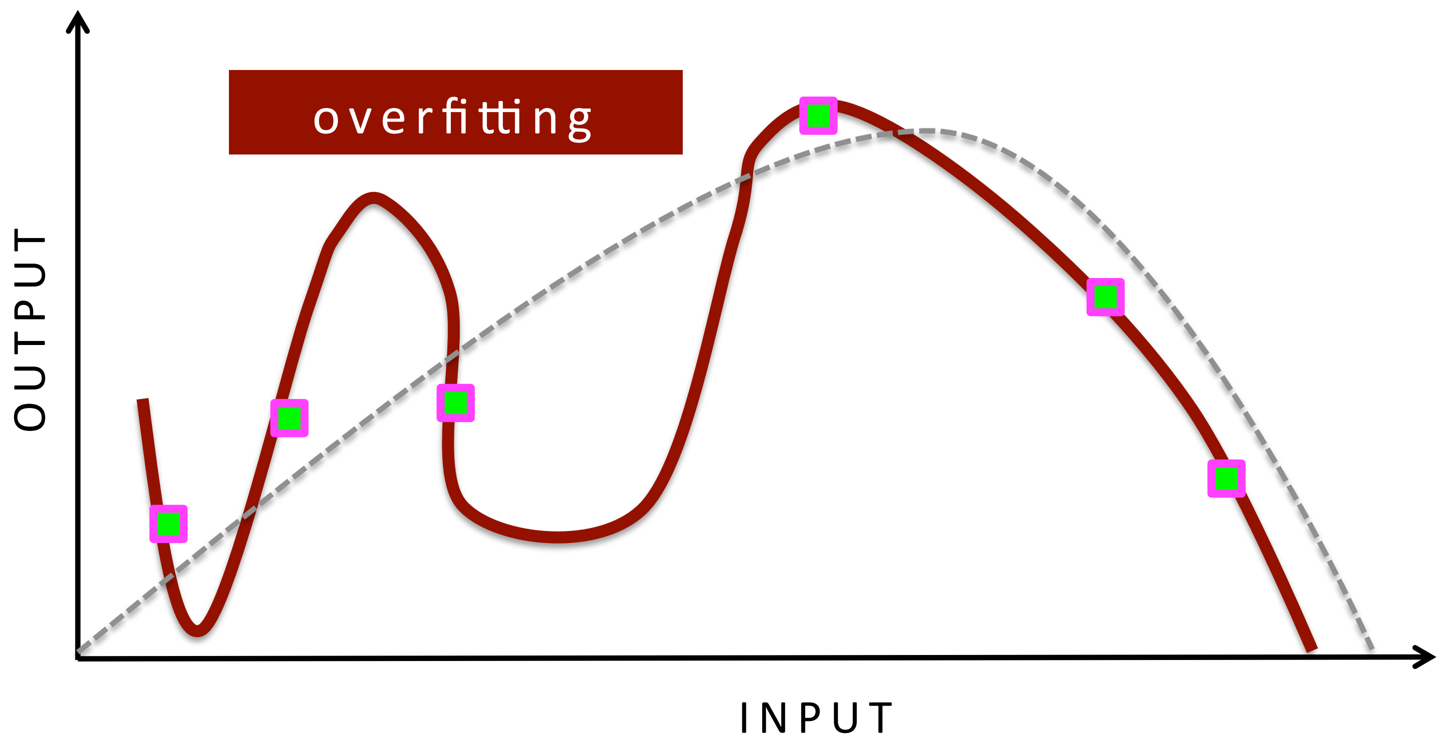
CHOICE OF THE MODEL SPACE



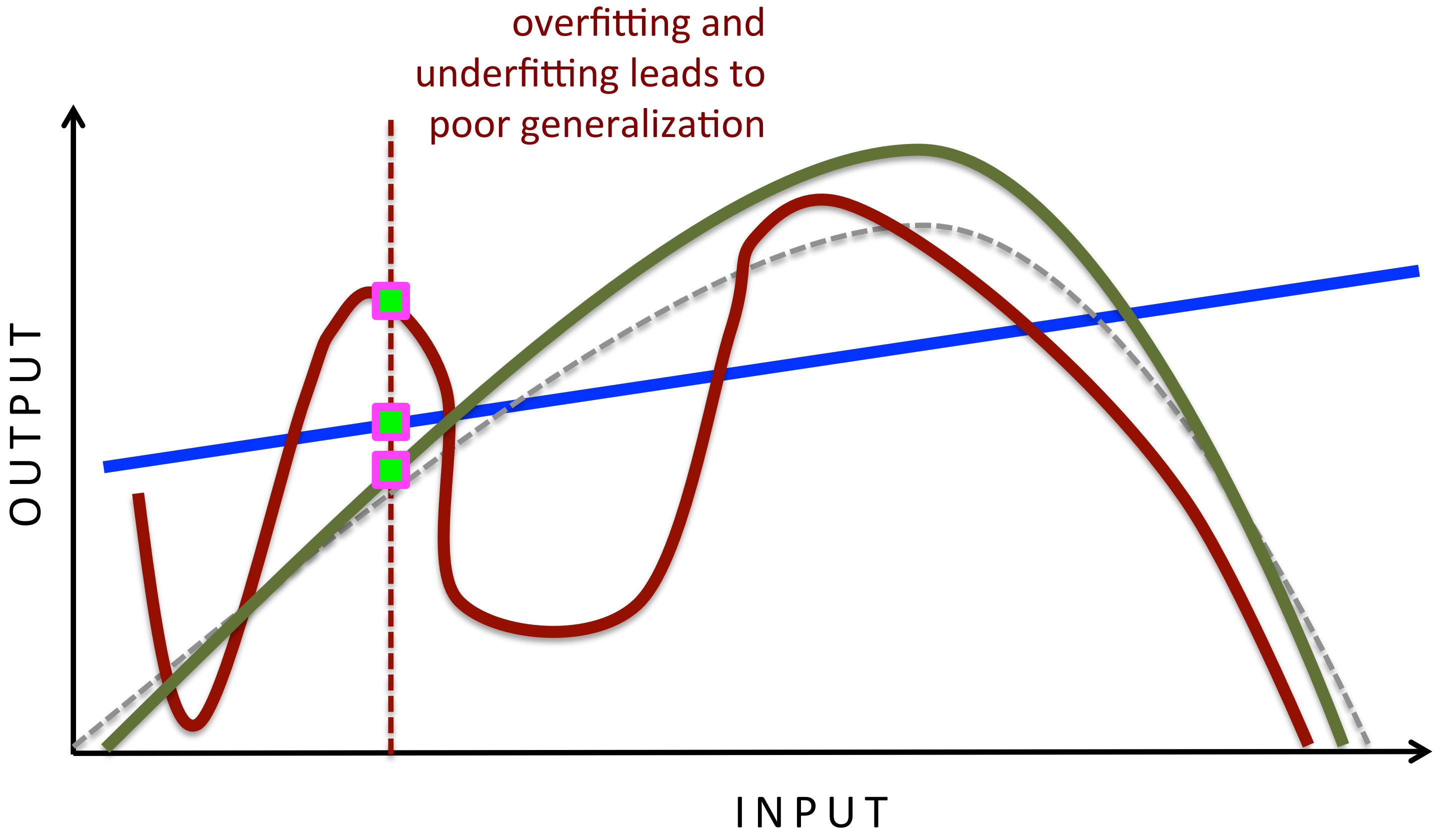
CHOICE OF THE MODEL SPACE



CHOICE OF THE MODEL SPACE



CHOICE OF THE MODEL SPACE



LOGISTIC REGRESSION

- **Logistic regression** modifies linear regression for the purpose of predicting (probabilities of) **a binary class label** instead of real-valued responses.

- The basic model:

log-odds ratio

$$\log\left(\frac{\mathbf{P}(y=1\,|\,\boldsymbol{x})}{\mathbf{P}(y=0\,|\,\boldsymbol{x})}\right) = w_0 + \sum_{i=1}^{m} w_i \cdot x_i$$
$$= w_0 + \boldsymbol{w}^\top \boldsymbol{x}\ ,$$

linear function of predictor variables

where

– $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)^\top \in \mathbb{R}^m$ is an instance to be classified,

– $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)^\top \in \mathbb{R}^m$ is a vector of regression coefficients,

– $w_0 \in \mathbb{R}$ is a constant bias (the intercept).

LOGISTIC REGRESSION: CLASS PREDICTION

- Equivalently, this can be expressed in terms of **posterior probabilities**:

$$\mathbf{P}(y = 1 \,|\, \boldsymbol{x}) = \Big(1 + \exp(-w_0 - \boldsymbol{w}^\top \boldsymbol{x})\Big)^{-1}$$

$$\mathbf{P}(y = 0 \,|\, \boldsymbol{x}) = 1 - \mathbf{P}(y = 1 \,|\, \boldsymbol{x})$$

- **Predictions** are typically made using the following decision rule:

$$\hat{y} \;=\; \begin{cases} 0 & \text{if } \mathbf{P}(y = 1 \,|\, \boldsymbol{x}) < 1/2 \\ 1 & \text{if } \mathbf{P}(y = 1 \,|\, \boldsymbol{x}) \geq 1/2 \end{cases}$$

LOGISTIC REGRESSION: PARAMETER ESTIMATION

- The parameters of the model (bias, regression coefficients) can be obtained through **Maximum Likelihood (ML)** estimation.
- Given a sample of i.i.d. data

$$\mathcal{D} = \left\{(\boldsymbol{x}^{(i)}, y^{(i)})\right\}_{i=1}^{n} \subset (\mathbb{R}^m \times \{0,1\})^n \quad ,$$

the likelihood function is given by

$$\prod_{i=1}^{n} \mathbf{P}\left(y = y^{(i)} \mid \boldsymbol{x}^{(i)}\right) \quad ,$$

and the **ML estimate** is the maximizer of (the log of) this function:

$$(\hat{w}_0, \hat{\boldsymbol{w}}) = \arg \max_{(w_0, \boldsymbol{w})} \sum_{i=1}^{n} y^{(i)} \log \theta^{(i)}(w_0, \boldsymbol{w}) + (1 - y^{(i)}) \log\left(1 - \theta^{(i)}(w_0, \boldsymbol{w})\right)$$

with

$$\theta^{(i)}(w_0, \boldsymbol{w}) = \left(1 + \exp(-w_0 - \boldsymbol{w}^\top \boldsymbol{x}^{(i)})\right)^{-1}$$

LOGISTIC REGRESSION: IMPORTANT FEATURES

- Logistic regression is very popular and widely used in practice.
- It is **comprehesible** and easy to **interpret**, especially since the influence of each variable can easily be captured from the model:

$$\log\left(\frac{\mathbf{P}(y=1\,|\,\boldsymbol{x})}{\mathbf{P}(y=0\,|\,\boldsymbol{x})}\right) = w_0 + \boxed{w_1} \cdot x_1 + w_2 \cdot x_2 + \ldots + w_m \cdot x_m$$

direction and strength of influence of
the first variable on the log-odds ratio
(probability of positive class)

AGENDA

1. Introduction
2. **Ranking Problems**
   - Label ranking
   - Object ranking
   - Instance ranking
3. Model-based Preference Learning
4. Summary & Outlook

LABEL RANKING

... mapping instances to **TOTAL ORDERS** over a fixed set of alternatives/labels:

`(28, 0, 187, 0.4)` $\longmapsto$

$\succ$

$\succ$

$\underbrace{\hspace{4em}}$ instance $\boldsymbol{x} \in \mathcal{X}$
(e.g., features of a person)

$\underbrace{\hspace{4em}}$ ranking of labels/alternatives

$$\mathcal{Y} = \{y_1, y_2, \ldots, y_k\}$$

$$\mathcal{Y} = \{A, B, C, \ldots\}$$

LABEL RANKING

... mapping instances to **TOTAL ORDERS** over a fixed set of alternatives/labels:

$$\underbrace{\texttt{(28, 0, 187, 0.4)}}_{\substack{\text{instance } \boldsymbol{x} \in \mathcal{X} \\ \text{(e.g., features of a person)}}} \longmapsto \underbrace{\boxed{\text{ROME}} \succ \boxed{\text{PARIS}} \succ \boxed{\text{LONDON}}}_{\text{ranking of labels/alternatives}}$$

$$\mathcal{Y} = \{y_1, y_2, \ldots, y_k\}$$

$$\mathcal{Y} = \{A, B, C, \ldots\}$$

THE SUSHI DATA

Rankings of 10 types of sushi by 5000 customers.
Each customer is characterized by 11 features.
Collected by Kamishima et al., reprocessed by Grbovic.

http://www.kamishima.net/sushi/

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

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

A ranking of all labels

new instance

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

LABEL RANKING: PREDICTION

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

SPEARMAN

$$\mathcal{L}(\pi, \pi^*) = \sum_{i=1}^{k} \left(\pi(i) - \pi^*(i)\right)^2$$

LOSS

$$\rho = 1 - \frac{6D(\pi, \pi^*)}{k(k^2 - 1)}$$

RANK CORRELATION

LABEL RANKING: PREDICTION

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

$$\mathcal{L}(\pi, \pi^*) = \sum_{1 \le i < j \le k} [\![(\pi(i) - \pi(j))(\pi^*(i) - \pi^*(j)) < 0]\!]$$

LOSS

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

RANK CORRELATION

LEARNING TECHNIQUES

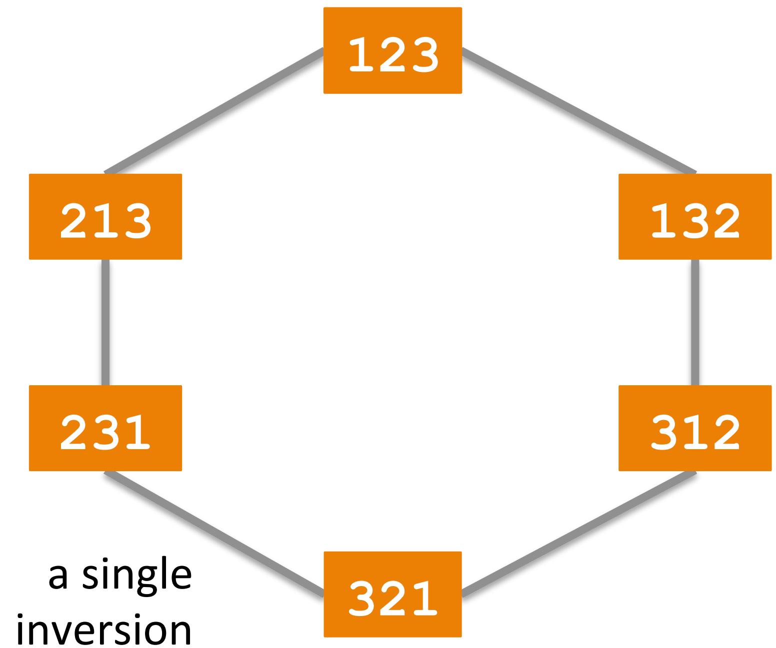
How to learn a label ranker $h : \mathcal{X} \rightarrow \mathcal{S}_k$?

The output space is complex ...

THE PERMUTATION SPACE

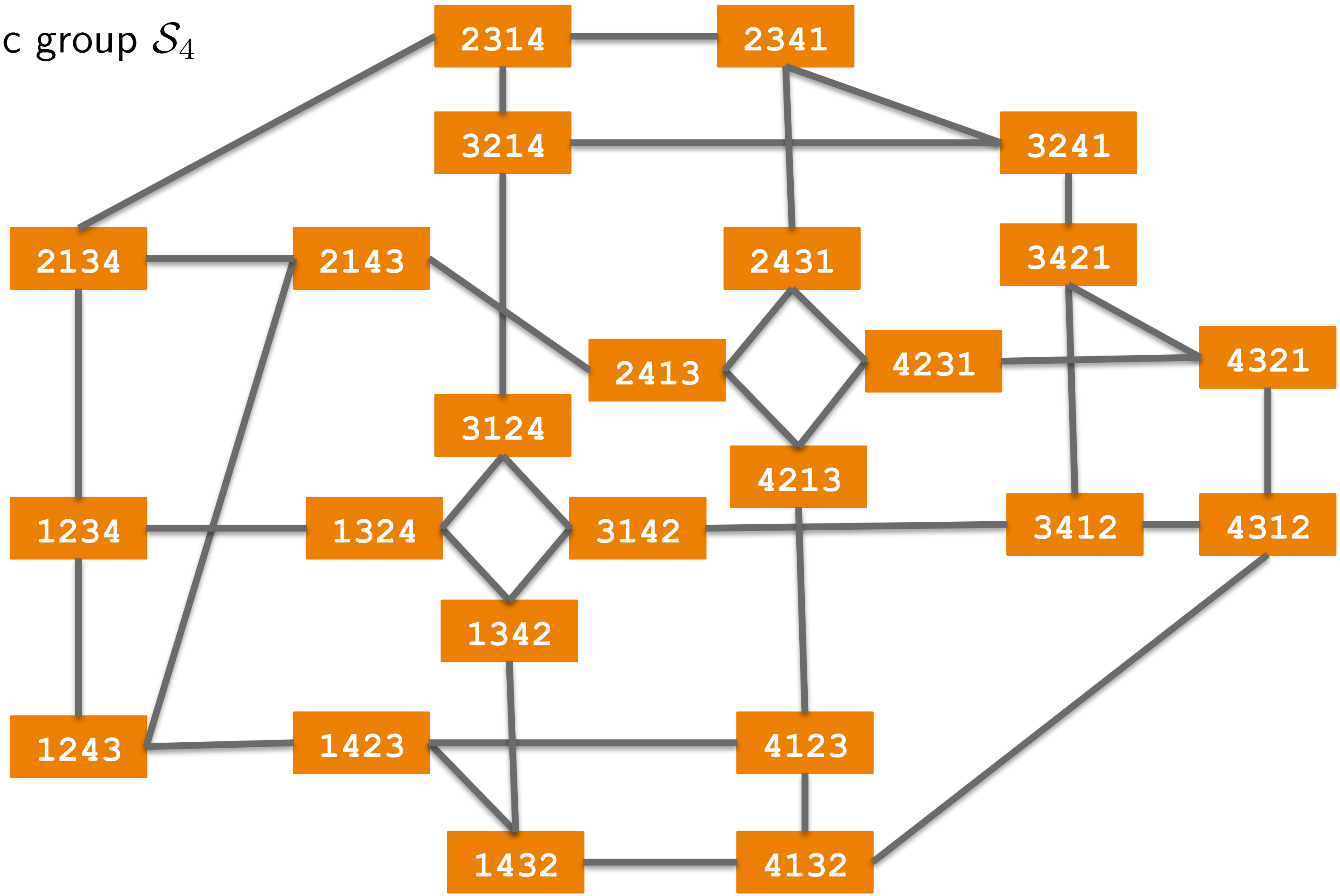
Our output space is the class of permutations (symmetric group):

Symmetric group $\mathcal{S}_3$

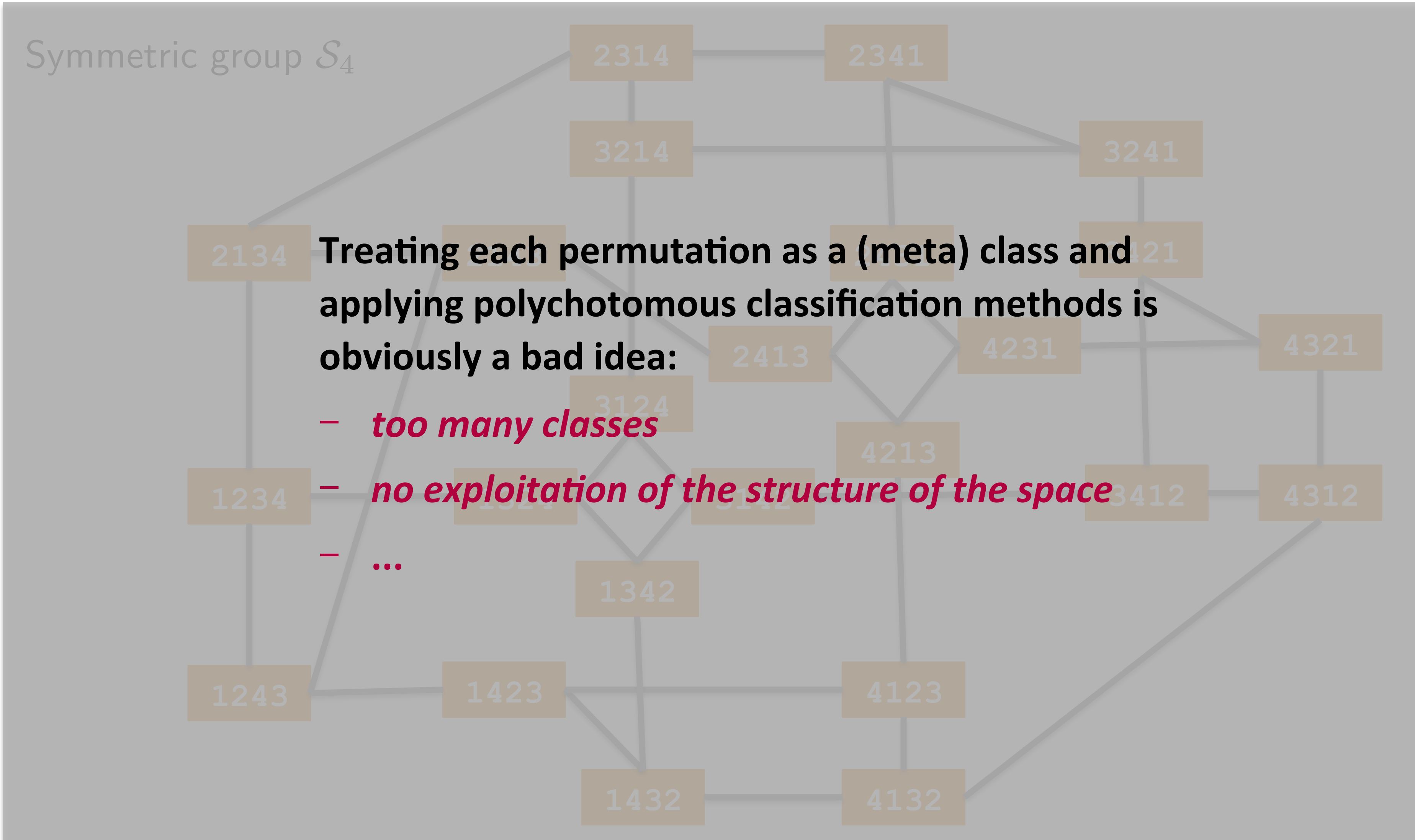


THE PERMUTATION SPACE

Symmetric group $\mathcal{S}_4$



THE PERMUTATION SPACE



Treating each permutation as a (meta) class and applying polychotomous classification methods is obviously a bad idea:

- ***too many classes***
- ***no exploitation of the structure of the space***
- ***...***

LEARNING TECHNIQUES

How to learn a label ranker $h : \mathcal{X} \rightarrow \mathcal{S}_k$?

DIFFERENT APPROACHES:

- Reduction to simpler problems (binary classification)

  Transform the problem, so as to make it amenable to standard ML algorithms.

- Extension of (classification) algorithms

  Generalize standard ML algorithms, so as to make them applicable to label ranking data.

- Probabilistic modeling and statistical inference

  Make use of statistical models for rank data and parameter estimation methods.

RANKING BY PAIRWISE COMPARISON

Ranking by Pairwise Comparison (RPC) trains models

$$\mathcal{M}_{i,j} : \mathcal{X} \to [0,1] \qquad\qquad (1 \leq i < j \leq k)$$

Given a query instance $\boldsymbol{x}$, $\mathcal{M}_{i,j}$ is supposed to predict the probability that $y_i \succ y_j$:

$$\begin{aligned}\mathcal{M}_{i,j}(\boldsymbol{x}) &= \mathbf{P}(y_i \succ y_j) \\ &= 1 - \mathbf{P}(y_j \succ y_i)\end{aligned}$$

$\longrightarrow$ decomposition into $k(k-1)/2$ **binary classification problems**

Training data (for the label pair A and B):

X1	X2	X3	X4	preferences	class
0.34	0	10			1
1.45	0	32			---
1.22	1	46			0
0.74	1	25			1
0.95	1	72			---
1.04	0	33	158	D $\succ$ A, **A $\succ$ B**, C $\succ$ B, A $\succ$ C	1

X1	X2	X3	X4	class
0.34	0	10	174	1
1.22	1	46	421	0
0.74	1	25	165	1
1.04	0	33	158	1

RANKING BY PAIRWISE COMPARISON

At prediction time, a query instance is submitted to all models, and the predictions are combined into a binary preference relation

$$\mathcal{P}(i,j) = \begin{cases} \mathcal{M}_{i,j}(\boldsymbol{x}) \,, & i < j \\ 1 - \mathcal{M}_{i,j}(\boldsymbol{x}) \,, & i > j \end{cases} \quad .$$

predictions $\mathcal{M}_{i,j}(\boldsymbol{x})$ →

	A	B	C	D
A		0.3	0.8	0.4
B	0.7		0.7	0.9
C	0.2	0.3		0.3
D	0.6	0.1	0.7	

How to produce a ranking on the basis of this preference relation?

LOSS DECOMPOSITION

Recall our original goal

$$\mathcal{R}(h) = \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{L}\left(h(x), y\right) \, d\,\mathbf{P}(X,Y) \quad \longrightarrow \quad \min$$

and our representation:

$$h = \mathrm{AGG}\left(\mathcal{M}_{1,2}, \mathcal{M}_{1,3}, \ldots, \mathcal{M}_{k-1,k}\right)$$

Loss decomposition problem: Is it possible to find a suitable loss $\mathcal{L}_p$, to be minimized (in expectation) by the pairwise learners, and an aggregation function AGG, such that $h = \mathrm{AGG}(\mathcal{M}_{1,2}, \ldots, \mathcal{M}_{k-1,k})$ minimizes $\mathcal{L}$ (in expectation)?

MINIMIZING SPEARMAN LOSS

predictions
$\mathcal{M}_{i,j}(\boldsymbol{x})$ →

	A	B	C	D	
A		0.3	0.8	0.4	1.5
B	0.7		0.7	0.9	2.3
C	0.2	0.3		0.3	0.8
D	0.6	0.1	0.7		1.4

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

MINIMIZING SPEARMAN LOSS

Theorem: Suppose the pairwise learners $\mathcal{M}_{i,j}$ yield unbiased probability estimates and let π be a ranking such that

$$\left(\sum_q \mathcal{P}(i,q) > \sum_q \mathcal{P}(j,q) \right) \Rightarrow (\pi(i) < \pi(j)) \ .$$

Then π minimizes risk w.r.t. to the Spearman loss

$$\mathcal{L}(\pi, \pi^*) = \sum_{i=1}^{k} \left(\pi(i) - \pi^*(i) \right)^2 \ .$$

MINIMIZING KENDALL LOSS

Theorem: Risk w.r.t. Kendall loss

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

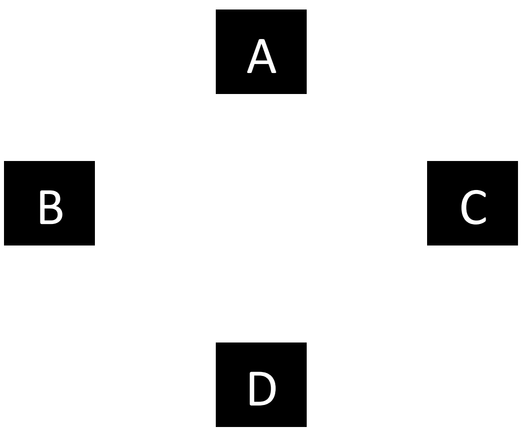
is minimized by

$$\pi^* = \arg\min_{\pi} \sum_{1 \le i < j \le k} \mathcal{P}\big(\pi^{-1}(j), \pi^{-1}(i)\big) \ .$$

$\longrightarrow$ linear ordering problem for weighted tournaments

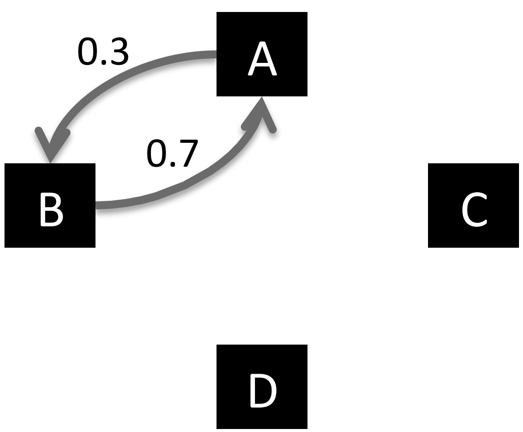
MINIMIZING KENDALL LOSS

	A	B	C	D
A		0.3	0.8	0.4
B	0.7		0.7	0.9
C	0.2	0.3		0.3
D	0.6	0.1	0.7	



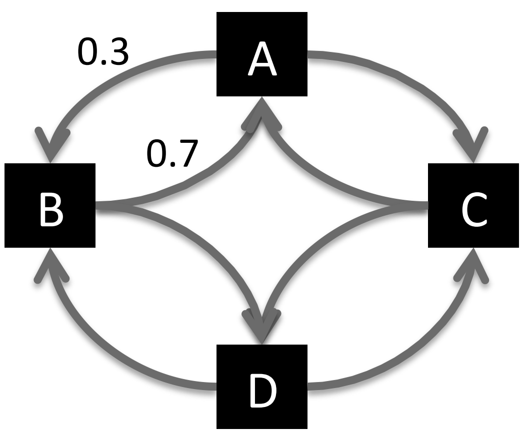
MINIMIZING KENDALL LOSS

	A	B	C	D
A		0.3	0.8	0.4
B	0.7		0.7	0.9
C	0.2	0.3		0.3
D	0.6	0.1	0.7	



MINIMIZING KENDALL LOSS

	A	B	C	D
A		0.3	0.8	0.4
B	0.7		0.7	0.9
C	0.2	0.3		0.3
D	0.6	0.1	0.7	



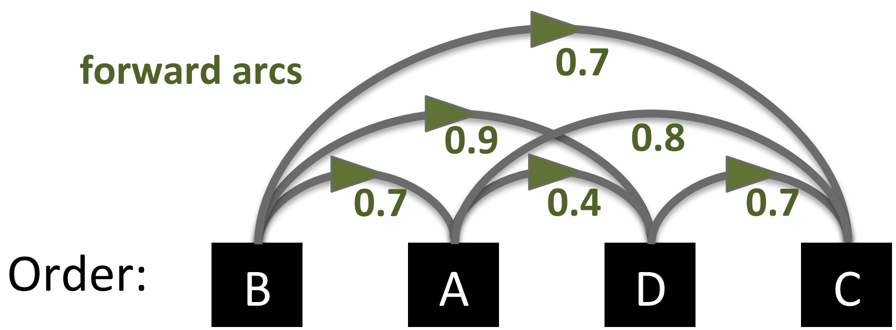
MINIMIZING KENDALL LOSS

	A	B	C	D
A		0.3	0.8	0.4
B	0.7		0.7	0.9
C	0.2	0.3		0.3
D	0.6	0.1	0.7	

Order: B A D C

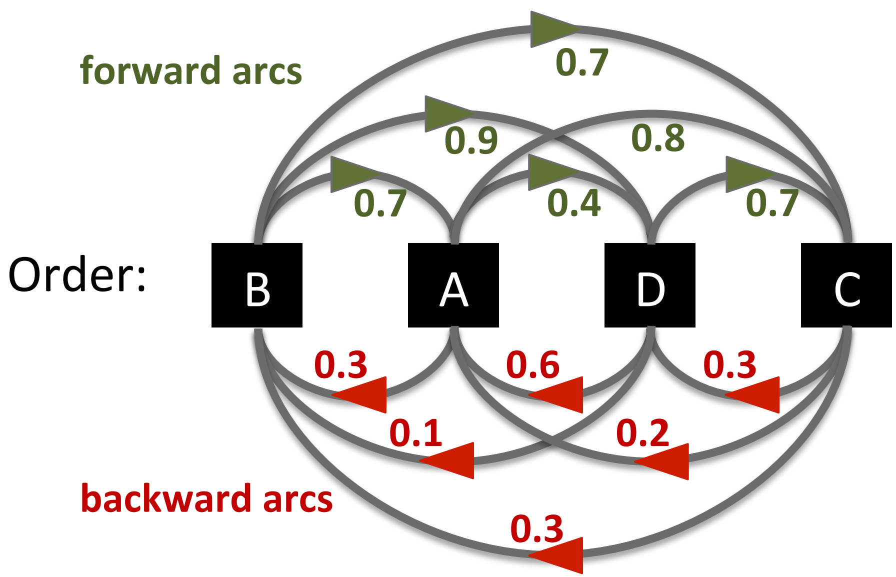
MINIMIZING KENDALL LOSS

	A	B	C	D
A		0.3	0.8	0.4
B	0.7		0.7	0.9
C	0.2	0.3		0.3
D	0.6	0.1	0.7	



MINIMIZING KENDALL LOSS

	A	B	C	D
A		0.3	0.8	0.4
B	0.7		0.7	0.9
C	0.2	0.3		0.3
D	0.6	0.1	0.7	



COST:

0.3 + 0.2 + 0.3 +

0.3 + 0.6 + 0.1 = 1.8

LIMITATIONS OF RPC

Proposition: For the following losses, RPC can not guarantee a risk minimizing prediction:

- 0/1 loss

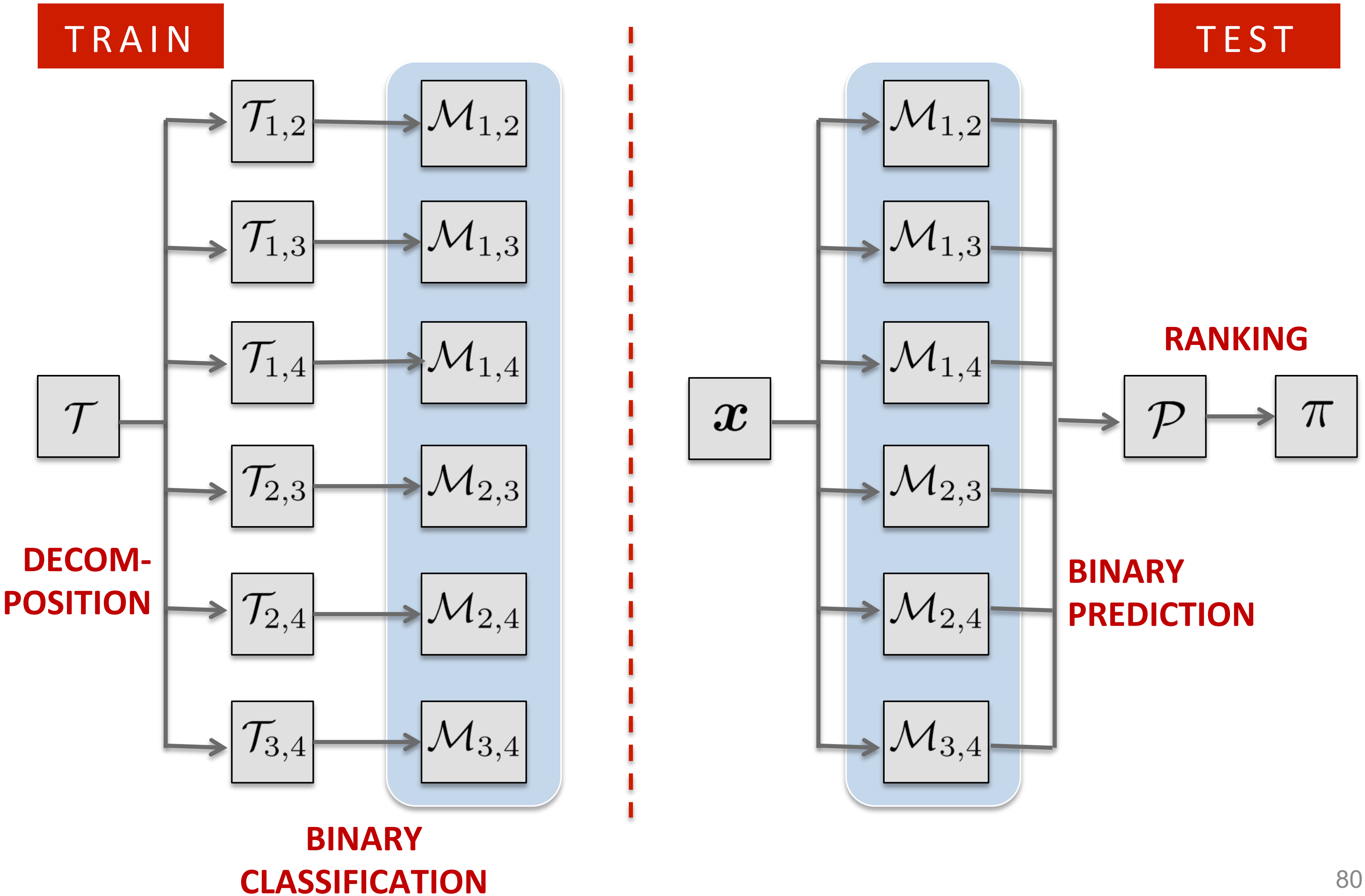
$$\mathcal{L}(\pi, \pi^*) = [\![\pi \neq \pi^*]\!]$$

- Hamming distance

$$\mathcal{L}(\pi, \pi^*) = \sum_{i=1}^{k} [\![\pi(i) \neq \pi^*(i)]\!]$$

- Cayley distance (minimal number of transpositions of any pair of labels needed to turn the first ranking into the second one)
- Ulam distance (minimal number of position changes of labels needed to turn the first ranking into the second one)

RANKING BY PAIRWISE COMARISON [E.H. et al., 2008]



LEARNING TECHNIQUES

How to learn a label ranker $h : \mathcal{X} \rightarrow \mathcal{S}_k$?

DIFFERENT APPROACHES:

- Reduction to simpler problems (binary classification)

  Transform the problem, so as to make it amenable to standard ML algorithms.

- Extension of (classification) algorithms

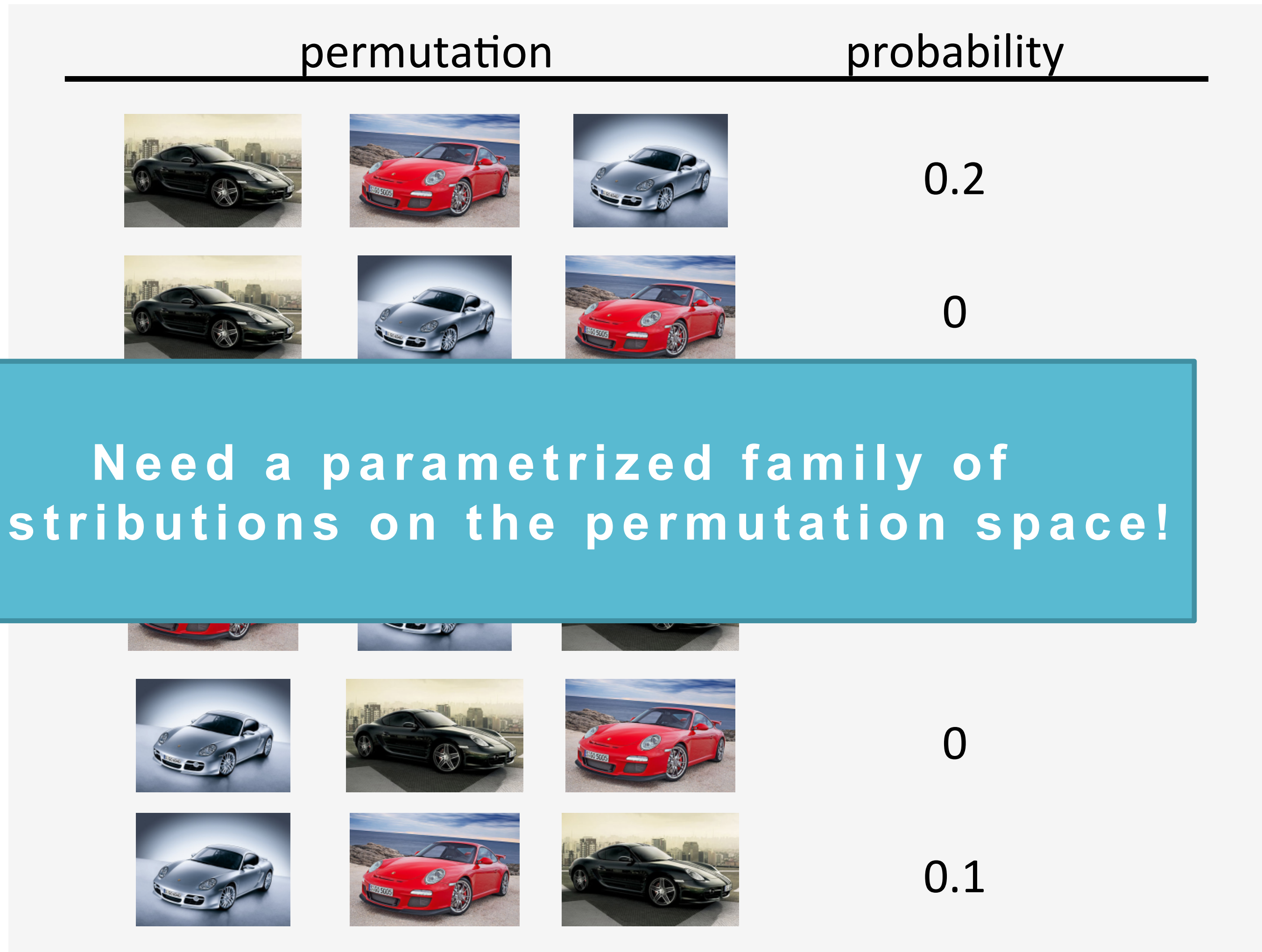
  Generalize standard ML algorithms, so as to make them applicable to label ranking data.

- **Probabilistic modeling and statistical inference**

  Make use of statistical models for rank data and parameter estimation methods.

PROBABILISTIC LABEL RANKER

input $x \mapsto$



Need a parametrized family of distributions on the permutation space!

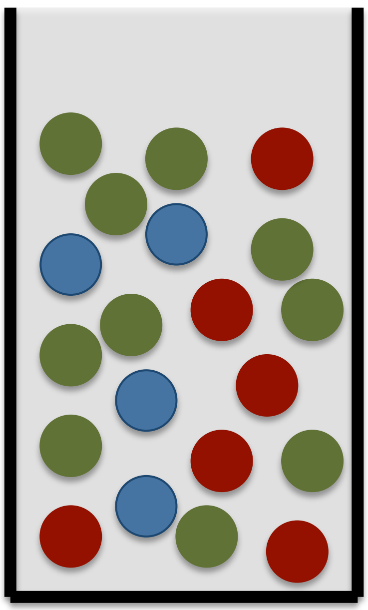
THE PLACKETT-LUCE MODEL

… is a **stagewise** model specified by a vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_k) \in \mathbb{R}_+^k$:

$$\mathbf{P}(\pi \,|\, \boldsymbol{v}) = \prod_{i=1}^{k} \frac{v_{\pi^{-1}(i)}}{v_{\pi^{-1}(i)} + v_{\pi^{-1}(i+1)} + \ldots + v_{\pi^{-1}(k)}}$$

A ranking is produced by choosing labels one by one, with a probability proportional to their respective „skills“.

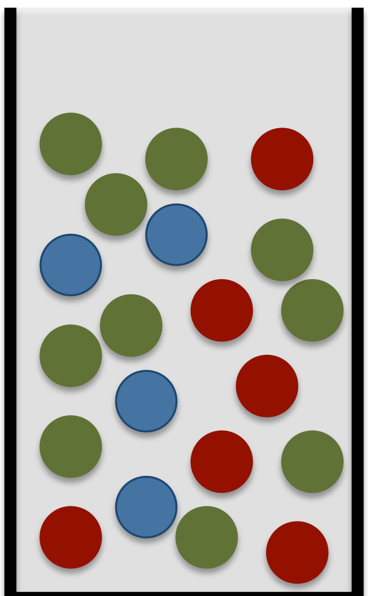
THE PLACKETT-LUCE MODEL

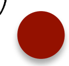
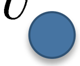


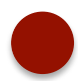
$$v_{\bullet} = 10, \quad v_{\bullet} = 6, \quad v_{\bullet} = 4$$

$$\mathbf{P}(\bullet, \bullet, \bullet) =$$

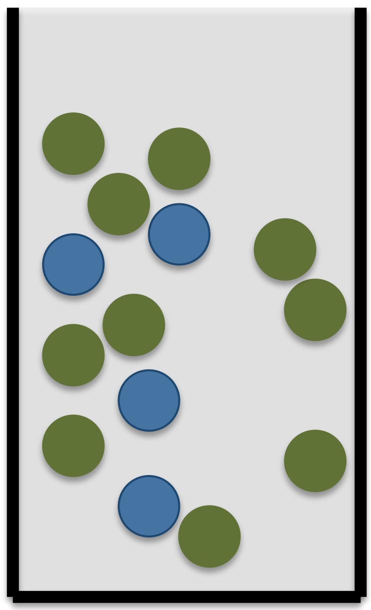
THE PLACKETT-LUCE MODEL



v  $= 10,$ v  $= 6,$ v  $= 4$

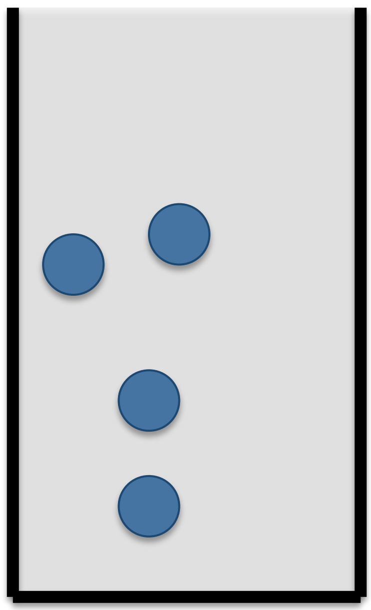
$\mathbf{P}($  $, \quad , \quad) = \frac{6}{20} \times$

THE PLACKETT-LUCE MODEL



$$v = 10, \quad v = 6, \quad v = 4$$

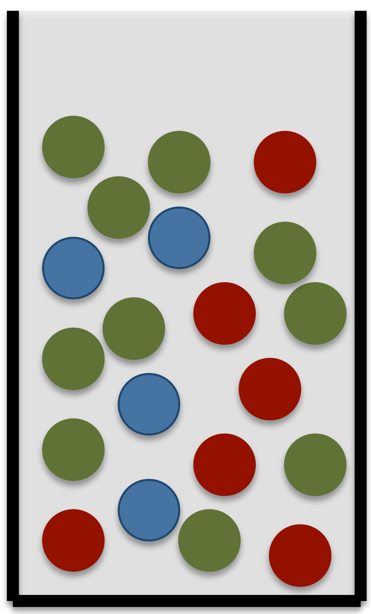
$$\mathbf{P}(\quad , \quad , \quad) = \frac{6}{20} \times \frac{10}{14} \times$$

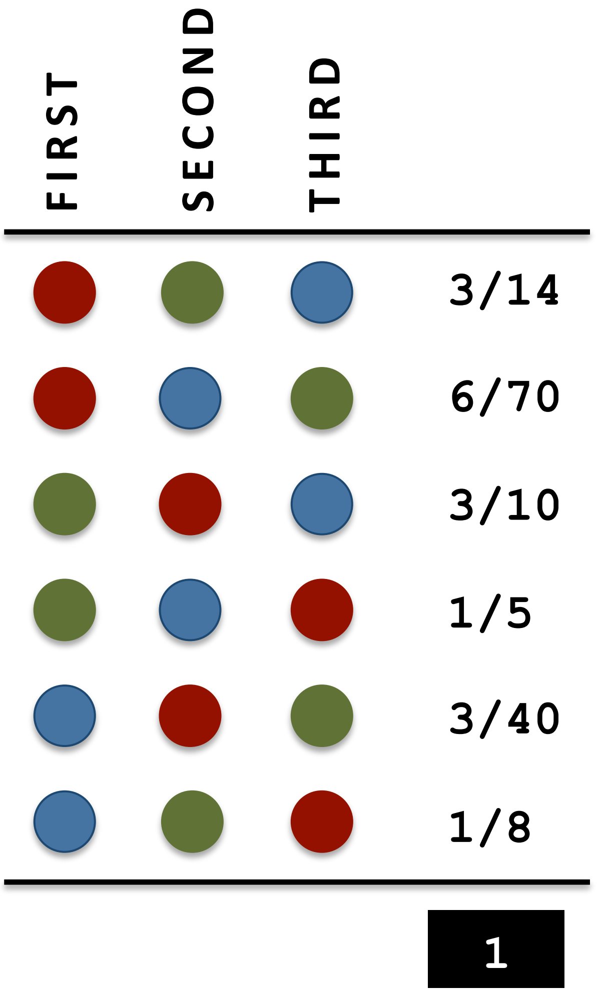


$$v_{\textcolor{green}{\bullet}} = 10, \quad v_{\textcolor{red}{\bullet}} = 6, \quad v_{\textcolor{blue}{\bullet}} = 4$$

$$\mathbf{P}(\textcolor{red}{\bullet}, \textcolor{green}{\bullet}, \textcolor{blue}{\bullet}) = \frac{6}{20} \times \frac{10}{14} \times \frac{4}{4} = \frac{3}{14}$$

THE PLACKETT-LUCE MODEL





PROBABILITY OF INCOMPLETE RANKINGS

Observations are not complete rankings such as

$$\pi : \text{B} \succ \text{C} \succ \text{A} \succ \text{D}$$

but **pairwise preferences** like

$$\sigma : \text{D} \succ \text{C}$$

or **incomplete rankings** like

$$\sigma : \text{B} \succ \text{D} \succ \text{A} \ .$$

PROBABILITY OF INCOMPLETE RANKINGS

Given a probability $\mathbf{P}(\cdot)$ on $\mathcal{S}_k$, the probability of an **incomplete ranking** σ is given by the probability of its linear extensions:

$$\mathbf{P}(\sigma) = \mathbf{P}(E(\sigma)) = \sum_{\pi \in E(\sigma)} \mathbf{P}(\pi)$$

linear extensions

PROBABILITY OF INCOMPLETE RANKINGS

A	B	C	D	0.14
A	B	D	C	0.00
A	C	B	D	0.08
A	C	D	B	0.00
A	D	B	C	0.10
A	D	C	B	0.00
B	A	C	D	0.00
B	A	D	C	0.05
B	C	A	D	0.00
B	C	D	A	0.00
B	D	A	C	0.15
B	D	C	A	0.00
C	A	B	D	0.00
C	A	D	B	0.03
C	B	A	D	0.00
C	B	D	A	0.16
C	D	A	B	0.00
C	D	B	A	0.00
D	A	B	C	0.00
D	A	C	B	0.02
D	B	A	C	0.00
D	B	C	A	0.17
D	C	A	B	0.00
D	C	B	A	0.09

$$\mathbf{P}(\mathsf{A} \succ \mathsf{C}) =$$

PROBABILITY OF INCOMPLETE RANKINGS

A	B	C	D	0.14
A	B	D	C	0.00
A	C	B	D	0.08
A	C	D	B	0.00
A	D	B	C	0.10
A	D	C	B	0.00
B	A	C	D	0.00
B	A	D	C	0.05
B	C	A	D	0.00
B	C	D	A	0.00
B	D	A	C	0.15
B	D	C	A	0.00
C	A	B	D	0.00
C	A	D	B	0.03
C	B	A	D	0.00
C	B	D	A	0.16
C	D	A	B	0.00
C	D	B	A	0.00
D	A	B	C	0.00
D	A	C	B	0.02
D	B	A	C	0.00
D	B	C	A	0.17
D	C	A	B	0.00
D	C	B	A	0.09

$$\mathbf{P}(\mathsf{A} \succ \mathsf{C}) = 0.54$$

LABEL RANKING MODELS

The probability to observe a set of (incomplete) rankings $\mathcal{D} = \{\sigma_n\}_{n=1}^N$, assuming independence, is

$$\mathbf{P}(\mathcal{D}) = \prod_{n=1}^{N} \mathbf{P}\big(E(\sigma_n)\big)$$

The **likelihood** of a set of parameter values is the probability of the data under these values:

$$L(\boldsymbol{v}) = \mathbf{P}(\mathcal{D} \,|\, \boldsymbol{v}) = \prod_{n=1}^{N} \mathbf{P}_{\boldsymbol{v}}\big(E(\sigma_n)\big)$$

→ Maximum Likelihood (ML) Inference

MODELING DEPENDENCY ON INSTANCES

Since rankings π are "contextualized" by instances $\boldsymbol{x}$, we need to model $\mathbf{P}(\pi \,|\, \boldsymbol{x})$ instead of $\mathbf{P}(\pi)$.

Assuming the PL model, this can be done by expressing $\boldsymbol{v} = (v_1, \ldots, v_k)$ as function of $\boldsymbol{x}$:

$$v_i \;=\; f_i(\boldsymbol{x}) = f_i(x_1, \ldots, x_m)$$

MODELING DEPENDENCY ON INSTANCES

Assuming $\boldsymbol{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$, the v_i can be expressed as log-linear functions:

$$v_i = f_i(\boldsymbol{x}) = \exp\left(\sum_{j=1}^{m} \alpha_j^{(i)} \cdot x_j\right)$$

$\longrightarrow$ estimation of parameter set $\left\{ \alpha_j^{(i)} \,|\, 1 \leq i \leq k,\, 1 \leq j \leq m \right\}$

$\longrightarrow$ label ranking model defined by $k \cdot m$ real parameters

MAXIMUM LIKELIHOOD INFERENCE

Given training data $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(n)}, \pi^{(n)}\right)\right\}_{n=1}^{N}$ with $\boldsymbol{x}^{(n)} = \left(x_1^{(n)}, \ldots, x_m^{(n)}\right)$, the log-likelihood is given by

$$\ell = \sum_{n=1}^{N} \left[\sum_{j=1}^{K_n} \log\left(v(\pi^{(n)}(j), n)\right) - \log \sum_{k=j}^{K_n} v(\pi^{(n)}(k), n) \right] ,$$

where K_n is the number of labels in the ranking $\pi^{(n)}$, and

$$v(j, n) = \exp\left(\sum_{i=1}^{m} \alpha_m^{(j)} \cdot x_m^{(n)} \right) \quad .$$

Algorithm based on MM (minorization and maximization) construction principle [Hunter 2004].

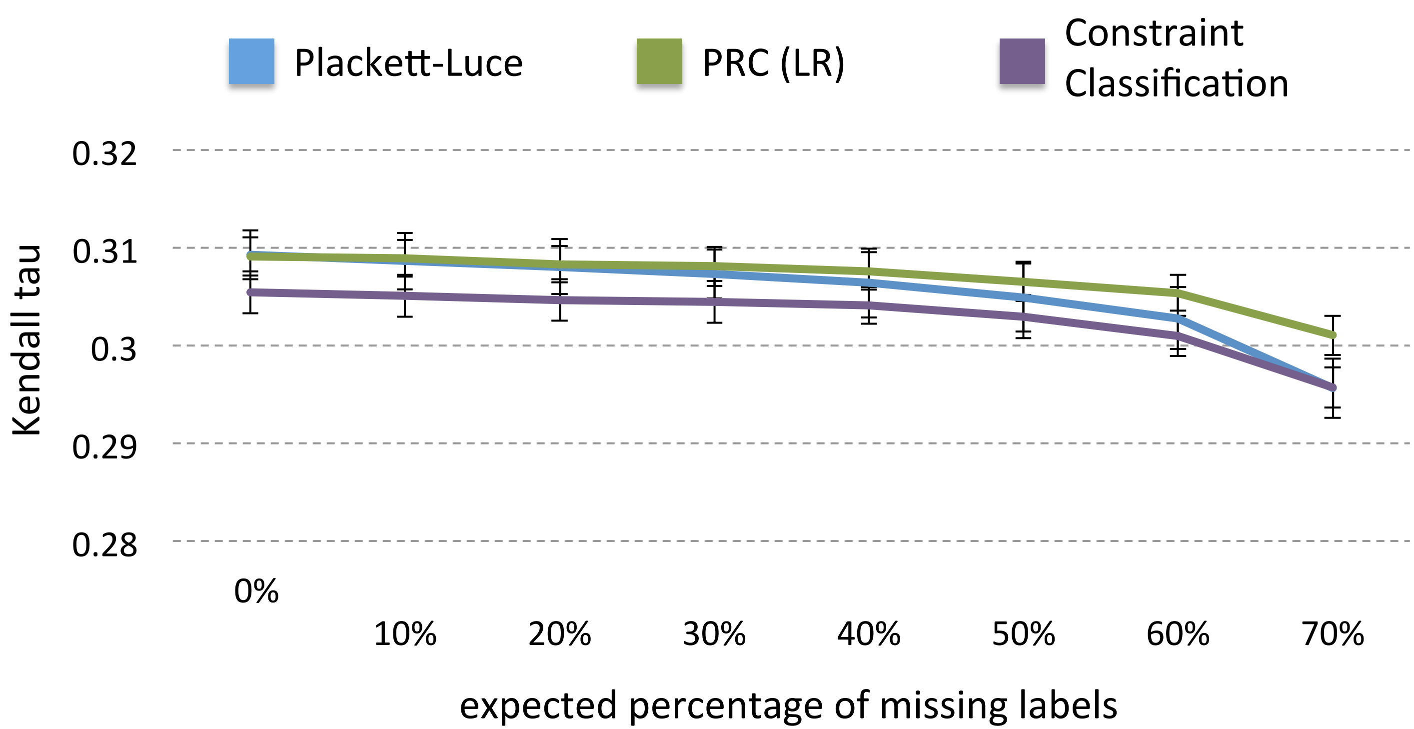
THE SUSHI DATA

Rankings of 10 types of sushi by 5000 customers.
Each customer is characterized by 11 features.

Collected by Kamishima et al., reprocessed by Grbovic.

http://www.kamishima.net/sushi/

EXPERIMENTAL STUDIES



SELECTED LITERATURE

- 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.
- 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.