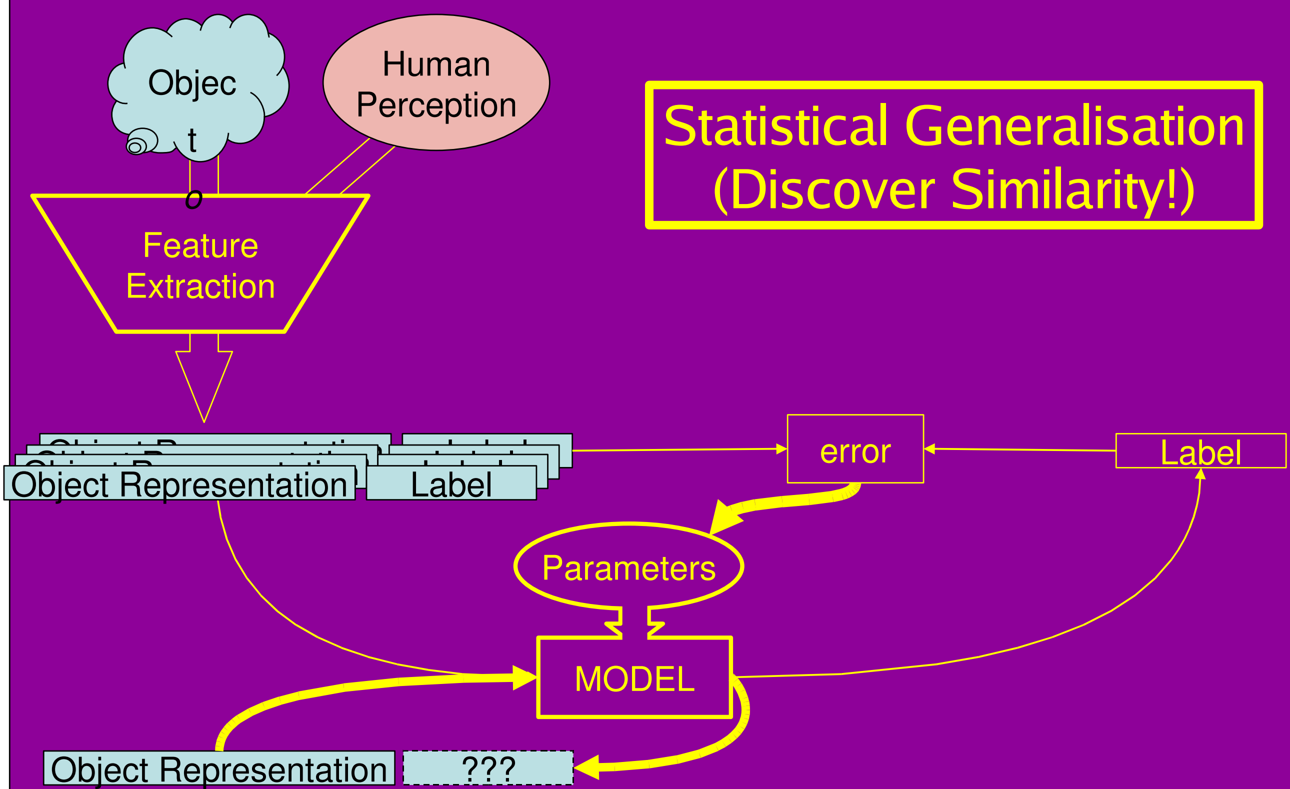
"Tuning":
Error Optimisation in Ad-Hoc Retrieval

Hugo Zaragoza,
Yahoo! Research Barcelona.

(This work was completed at Microsoft Research Cambridge,
in collaboration with:
Ralf Herbrich, Stephen Robertson, Michael Taylor, Nick Craswell & Chris Burges.)

Machine Learning intro.



Machine Learning intro.

- Label type: Task:
  - Binary {A,¬A} — Clasification
  - Discrete {A,B,C} — Multiclass Class.
    (Discrete but unknown) — Clustering
  - Preferences {A > B > C} — Oridinal Regression
  - Continuous {R} — Regression

Machine Learning intro.

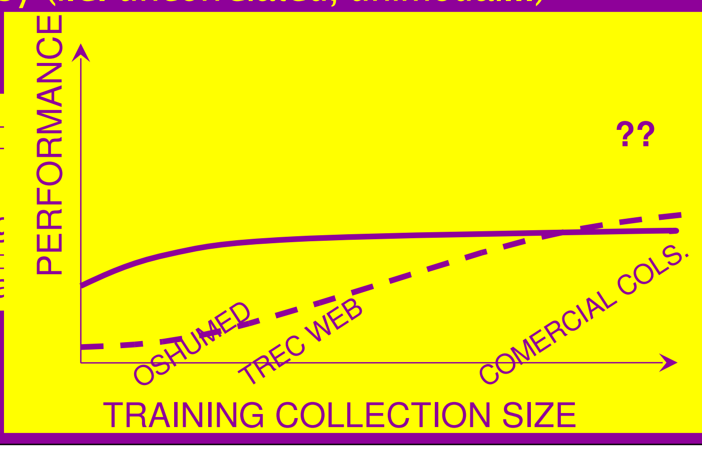
- The promise…
  - Features + labels + ML = Good Model.
  - Concentrate on innovation, tasks, forget the details…
- Until today…
  - *Good* features.
  - *Lots* of labels, good (#labels / #feaure) ratio.
  - *Not too noisy* & nicely noisy (i.e. uncorrelated, unimodal...)
  - *Element-wise* error.
- To

i.e.

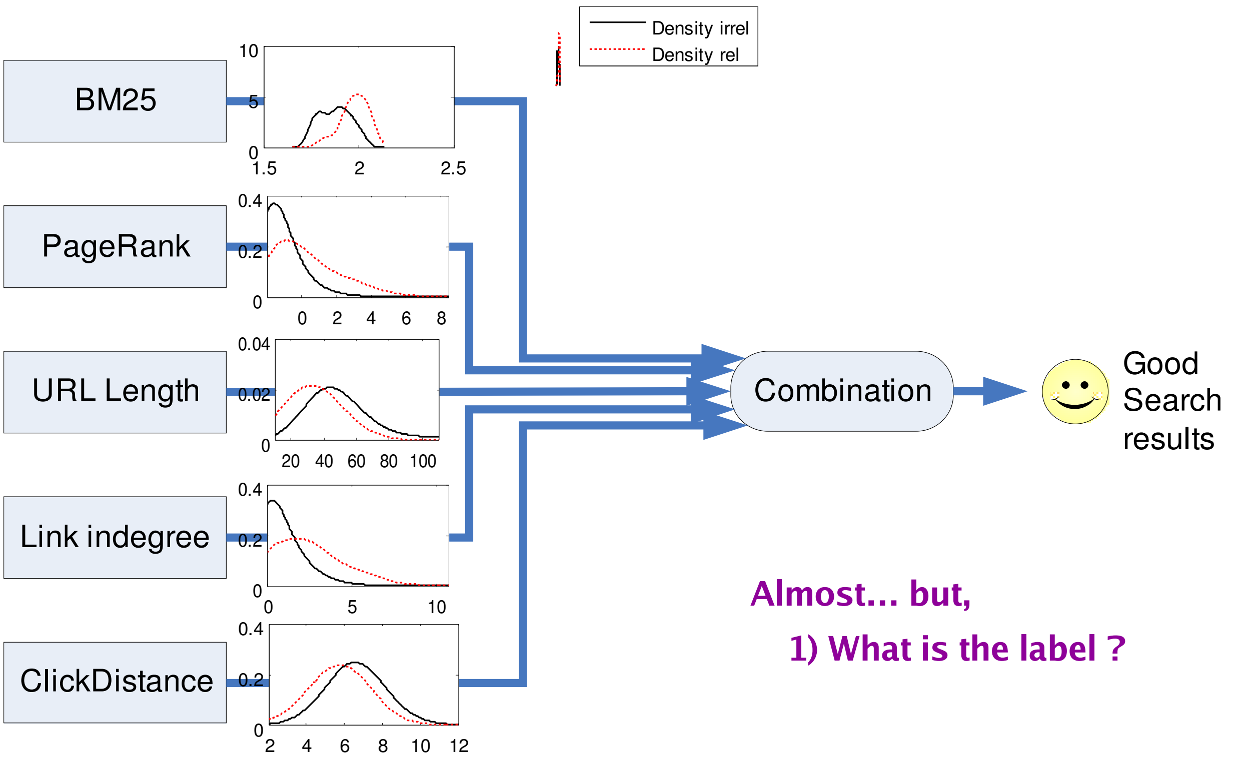
i.e.

- *tf* , *df* → *tf* · log

- HARD: not enought la



Learning Web Ranking Functions



"Ranking" issues

- Rank-dependant evaluation

The *goodness* of one score depends on all the others.

score	*rank*	*relevant?*
17.9	*1*	☒
17.6	*2*	☑
12.5	*3*	☒
8.5	*4*	☒
5.3	*5*	☑
…	…	…

Precision@5 = 2/5

- Query-dependant scale

The relevance scale changes across queries.

query1 = *"Britney"*
doc. A (☑) score = 5
doc. B (☒) score = 0

query2 = *"Britney Spears"*
doc. A (☑) score = 15
doc. B (☒) score = 7

“Ranking” problem

1. We are only interested in **top ranked** objects.

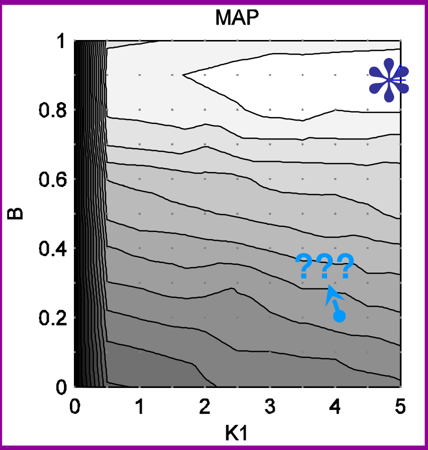
2. Labels are only **relative** to other labels.
   - i.e. *Doc324 is best, Doc311 second best, etc.*

- *e.g.*
  - *Find the parameters that minimise L2 distance on the highest 20 target values*
  - *Precision at 5*
  - *Average Precision, NDCG, Reciprocal mean, …*

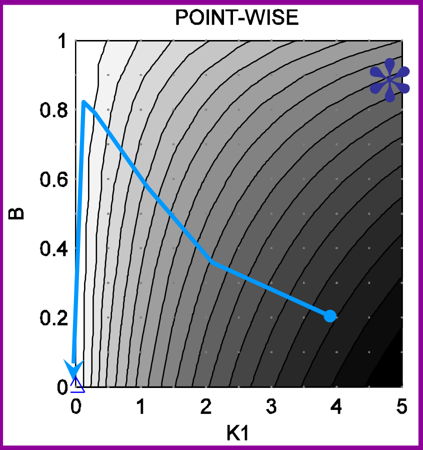
- *Difficult, but ML is starting to propose practical solutions to this problem!*

Example (Topic Distillation, TREC’04)

***Avg. Precision* Surface**



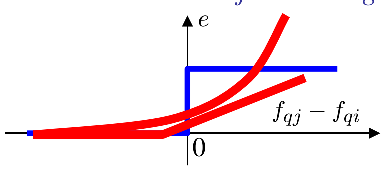
***Point-Wise* (Cross-Entropy)**



Pair-wise error measures

Possible relevance *pairs*	
$(i\ ,\ j)$	error
☒ ☒	$: 0$
☑ ☑	$: 0$
☑ ☒	$:\ e(f_{qi}, f_{qj})$
☒ ☑	$:\ 1 - e(f_{qi}, f_{qj})$

j scores lower ← → *j scores higher*



$$e(f_{qi}, f_{qj}) := \cancel{\mathrm{sign}}(f_{qj} - f_{qi})$$

Smooth-Step,
Hinge Lose,
Exponential…
+ standard
ML machinery

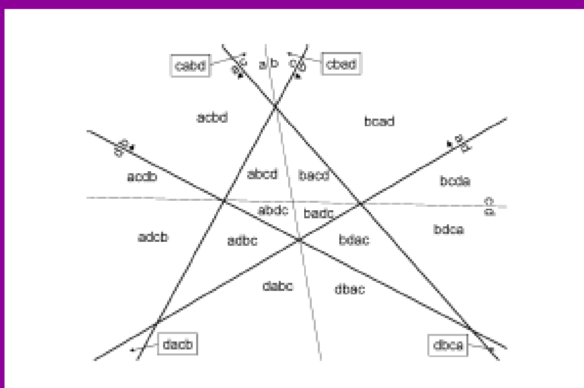
Number of irrelevant above relevant doc *i*:

$$mistakes(i) = \sum_{j \in \mathcal{T}_q} e(f_{qi}, f_{qj})$$

Precision at 5:

$$Precision@K = \frac{1}{Z} \sum_q \overset{\sum}{\cancel{\max}}_{j|rank(j) \leq K} \{\ mistakes(j)\ \}$$

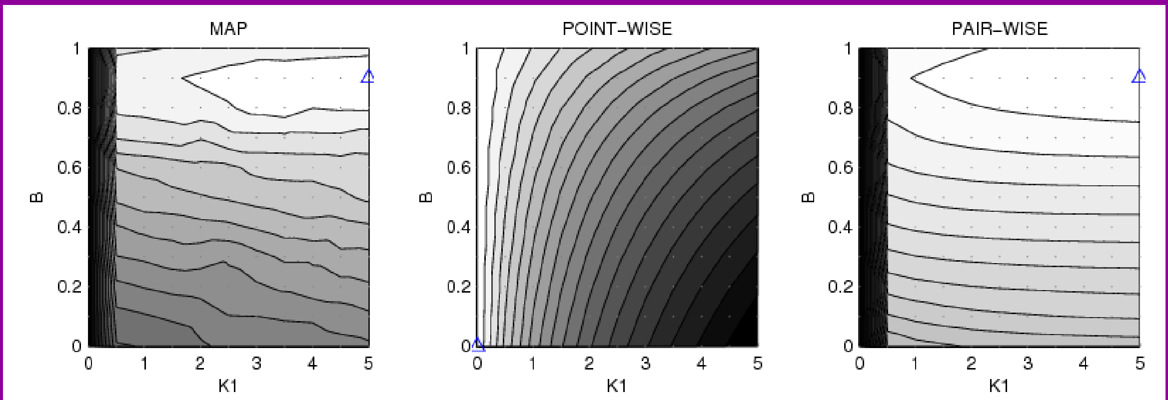
Pair-wise for NDCG: should it work?



[Robertson S., Zaragoza H., MSR-TR-2006-61]

- At a maxima, we cannot decrease pair-wise errors without decreasing NDCG or Avg. Prec.
- Therefore, these measures must share local maxima!

Pair-wise for NDCG: does it work?



average precision

cross-entropy

hinge loss

Pair-wise for NDCG: does it work?