Research Group
Databases and Information Systems

informatik institut für universität innsbruck

universität innsbruck

# Evaluation of IR Systems

Günther Specht
Eva Zangerle

Summer Term 2019

# Evaluation of IR Systems

- How to measure how 'good' a retrieval system performs?
- Without any evaluation hardly any improvements are possible
- Need to justify changes

- Today:
  - Evaluation overview
  - Evaluation metrics for (automatic) evaluations
  - Kappa
  - Accuracy
  - Precision@k, Recall@k
  - Mean Average Precision
  - User-based Evaluation
  - Crowdsourcing

# Evaluation of IR Systems

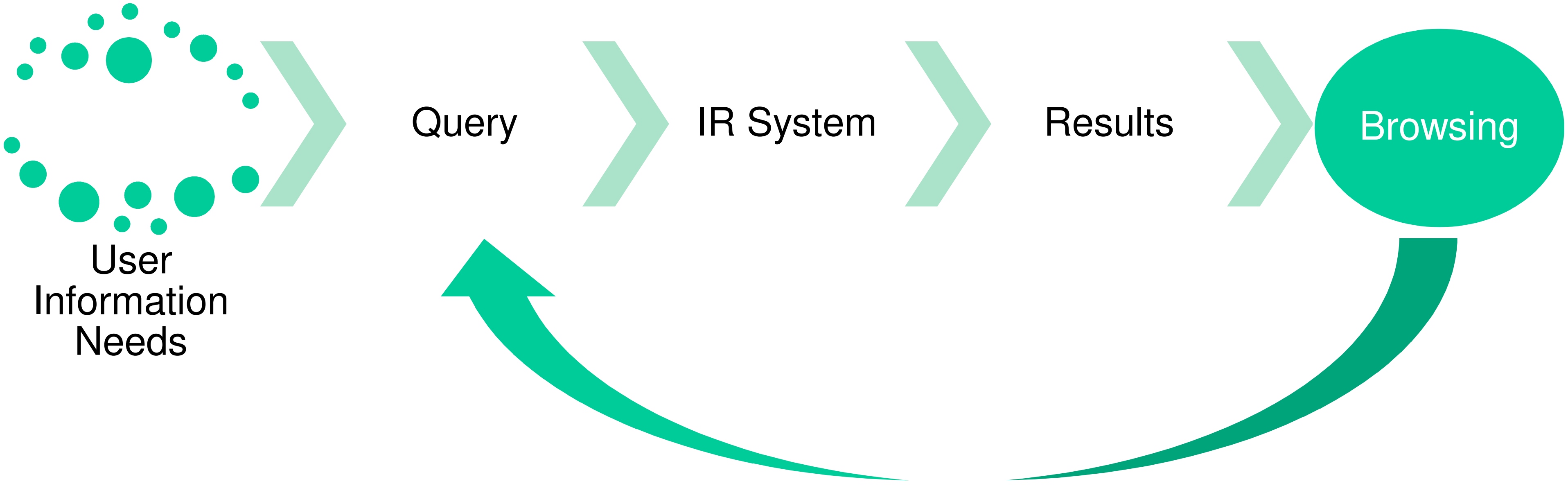
- First attempt
  - Speed of query evaluation
    - Queries per second evaluated
  - Speed of indexing
    - Number of tweets added to the index per second
  - Footprint of index
  - Query expressivity
    - No wildcards
    - Simple keyword search
- Factors above can be measured
- What about the user?

# Evaluation of IR Systems

- Most important measure: user happiness ☺
  - **Relevance of retrieved documents**
  - Speed of response, latency
  - Size of indexed documents

  - Also: user interface design
    - Responsiveness
    - Clarity
    - Layout

# Information Needs



User Information Needs

Query

IR System

Results

Browsing

Reformulation
Query Expansion
Relevance Feedback

# User Happiness

- What are criteria for user happiness?

  in

  - (Web) search engine
  - Online shop
  - Social network

# Evaluation of IR Systems

- Relevance
  - Relevance related to information need (not the query itself)
    - Document containing all search keywords may not be relevant at all in regards to information need
    - Cf. apple vs. Apple
  - Binary (relevant or non-relevant)
- Translation of information need to query:
  - Sometimes difficult. Example:

    Information need: Information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.

    Query: wine and red and white and heart and attack and effective

# Measuring Ad-hoc Effectiveness

- Test collection required
  - Document collection
  - Queries: Test suite of information needs (queries expressing these information needs)
  - Relevancy judgements for query-document pairs
    - Team of human judges
    - At least one judge per document

  - Ground truth

# Test Collections, e.g

- Cranfield Collection
  - First evaluation collection
  - Collection is fully evaluated
  - 1.400 documents (abstracts of aerodynamics journals)
  - 225 queries
  - Imporant, since first one; today: too small.

- TREC
  - Text Retrieval Conference
  - 1.89 mio documents (newswire articles) in total
  - 450 information needs (topics) in total
  - TREC 6-8: 150 information needs for 528,000 documents
  - trec.nist.gov

- PAN Workshops
  - For author ship attribution, plagiarism detection etc.

# TREC Example

```
<top>
<num> 52 </num>
<title> Topic:  Accusations of Cheating by Contractors on U.S. Defense
Projects</title>
<desc> Description:
Document will refer to an alleged illegality committed by any entity
seeking a contract on behalf of the U.S. Military Forces. </desc>
<narr> Narrative:
A relevant document will mention an alleged impropriety or improprieties
by individuals or entities (companies, corporations), either domestic or
foreign, attempting to provide services or products related to the U.S.
military effort.  To be relevant, document will specifically:
(1) identify the wrongdoer, and
(2) describe in at least general terms (e.g., overcharging, bribing,
swindling, illegal gifts, bid-riggings found, insider information) the
nature of the wrongdoing. </narr>
</top>
```

# Assessing Relevance

- Manual task
- User information needs have to be realistic (not randomly generated keywords) → realistic distribution
- Relevance assessment tedious, costly, time-consuming

- Humans for gold standard judgement?
  - Different opinions, backgrounds, expections
  - But: systems are used by the same users
- Multiple judges for relevance assessment

- Measurement of agreement of judges on relevance judgements?

# Kappa Statistic

- Measure for agreement of judges: Kappa statistic

$$kappa = \frac{P(A) - P(E)}{1\ - P(E)}$$

- P(A) proportion of the times the judges agreed
- P(E) proportion of the times they would be expected to agree by chances (next slide)
  - Two-class decision: chance of agreement is 0.5

- kappa = 1 for total agreement
- kappa = 0 agreement per chance (dt. Zufall)

# Kappa Statistic

| | | Judge 2 | | |
|---|---|---|---|---|
| | | yes | no | total |
| Judge 1 | yes | 300 | 20 | 320 |
| | no | 10 | 70 | 80 |
| | total | 310 | 90 | 400 |

- P(A) = (300 + 70) / 400 = 0.925
- P(relevant) = (320 + 310) / (400 + 400) = 0.7878
- P(nonrelevant) = (80 + 90) / (400 + 400) = 0.2125
- P(E) = P(relevant)$^2$ + P(nonrelevant)$^2$ = 0.665
- Kappa = 0.776

# Kappa Statistic

Manning:

- 0 -0.66 = seen as dubious base for evaluation
- 0.67-0.80 = fair agreement
- 0.81-1.00 = good agreement
-

Landis and Koch:

- 0.41-0.60 = „mittelmäßige (moderate) Übereinstimmung“,
  0.61-0.80 = „beachtliche (substantial) Übereinstimmung“,
  0.81-1.00 = „(fast) vollkommene ((almost) perfect) Übereinstimmung

- Within TREC dataset, normally kappa > 0.67, < 0.8

# Unranked Retrieval Evaluation

# Unranked Retrieval Evaluation

User

| Information System | Relevant | Not Relevant |
|---|---|---|
| Retrieved | True Positive | False Positive |
| Not Retrieved | False Negative | True Negative |

# Precision and Recall

- Precision
  - Fraction of relevant documents which are retrieved
  - $precision = \frac{\#\,relevant\ retrieved\ docs}{\#retrieved\ docs} = \frac{\#tp}{\#tp+\#fp}$
- Recall
  - Fraction of documents which are relevant
  - $recall = \frac{\#\,relevant\ retrieved\ docs}{\#relevant\ docs} = \frac{\#tp}{\#tp+\#fn}$

| | **relevant** | **not relevant** |
|---|---|---|
| retrieved | true positive | false positive |
| not retrieved | false negative | true negative |

# Accuracy

- Accuracy
  - Number of correct classifications
  - $accuracy = \frac{\#tp+\#tn}{\#tp+\#tn+fn+\#fp}$
  - For most queries, 99% of all documents are irrelevant
  - Optimize accuracy: label all documents as non-relevant → accuracy = 99%

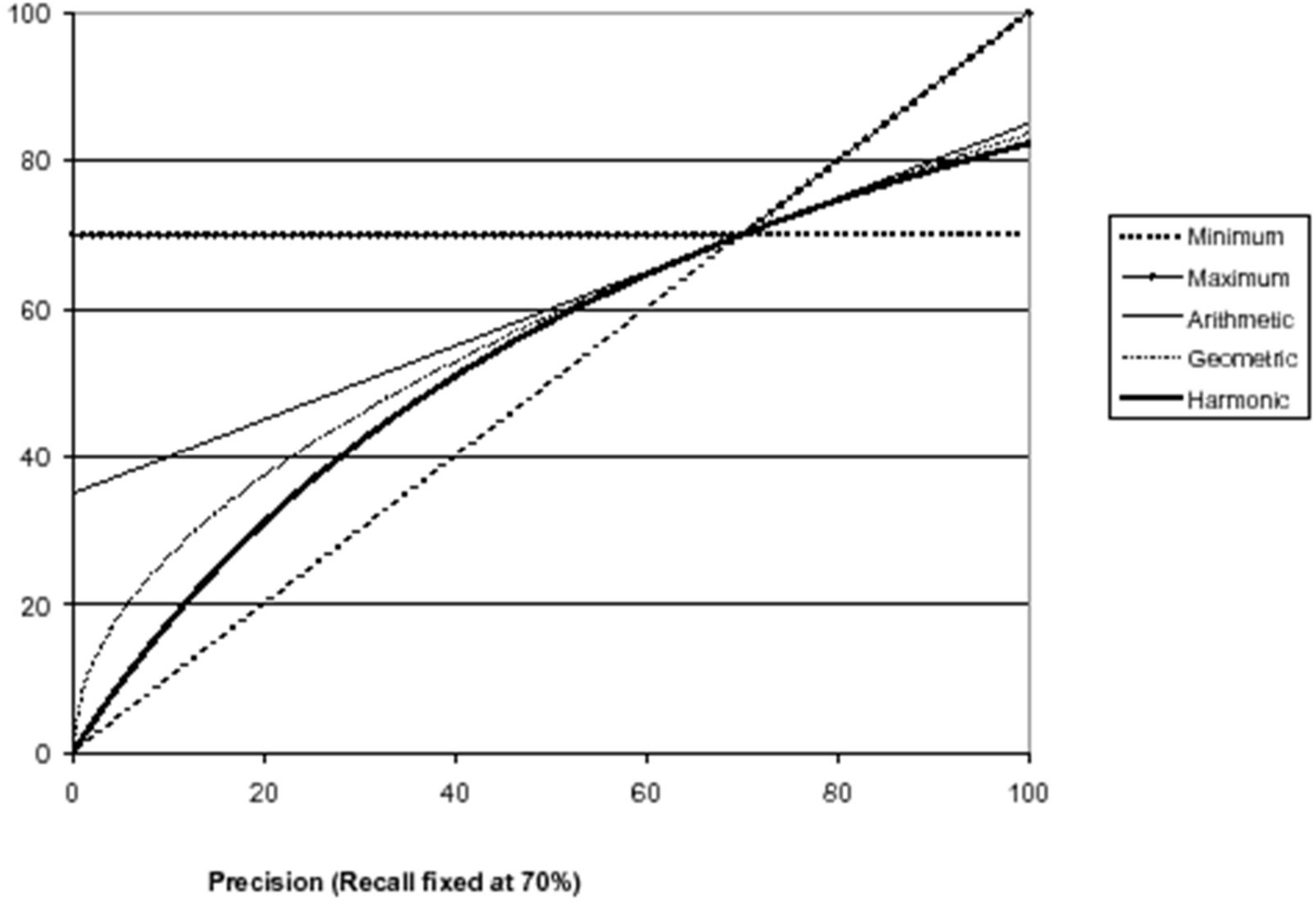
    But: empty resultset not satisfying for user

- Recall vs. Precision depends on aim of information system
  - Consider web search: better optimize for recall or for precision?

# F-Measure

- F-measure (or F1-measure)
  - Combination of recall and precision
  - Harmonic mean of recall and precision
  - $F1 = \frac{2 * P * R}{P + R}$
  - Why harmonic mean and not
    - Arithmetic mean (a+b)/2
    - Geometric mean (a*b)^1/2
    - Harmonic Mean is conservative, not that affected by outliers

# F-Measure


100
80
60
40
20
0
0
20
40
60
80
100
Minimum
Maximum
Arithmetic
Geometric
Harmonic
Precision (Recall fixed at 70%)

# Ranked Retrieval Evaluation

# Ranked Retrieval Evaluation

- Ranked Retrieval
- Precision, recall and F1 also applicable for ranked result sets

- But: ranking of documents not reflected

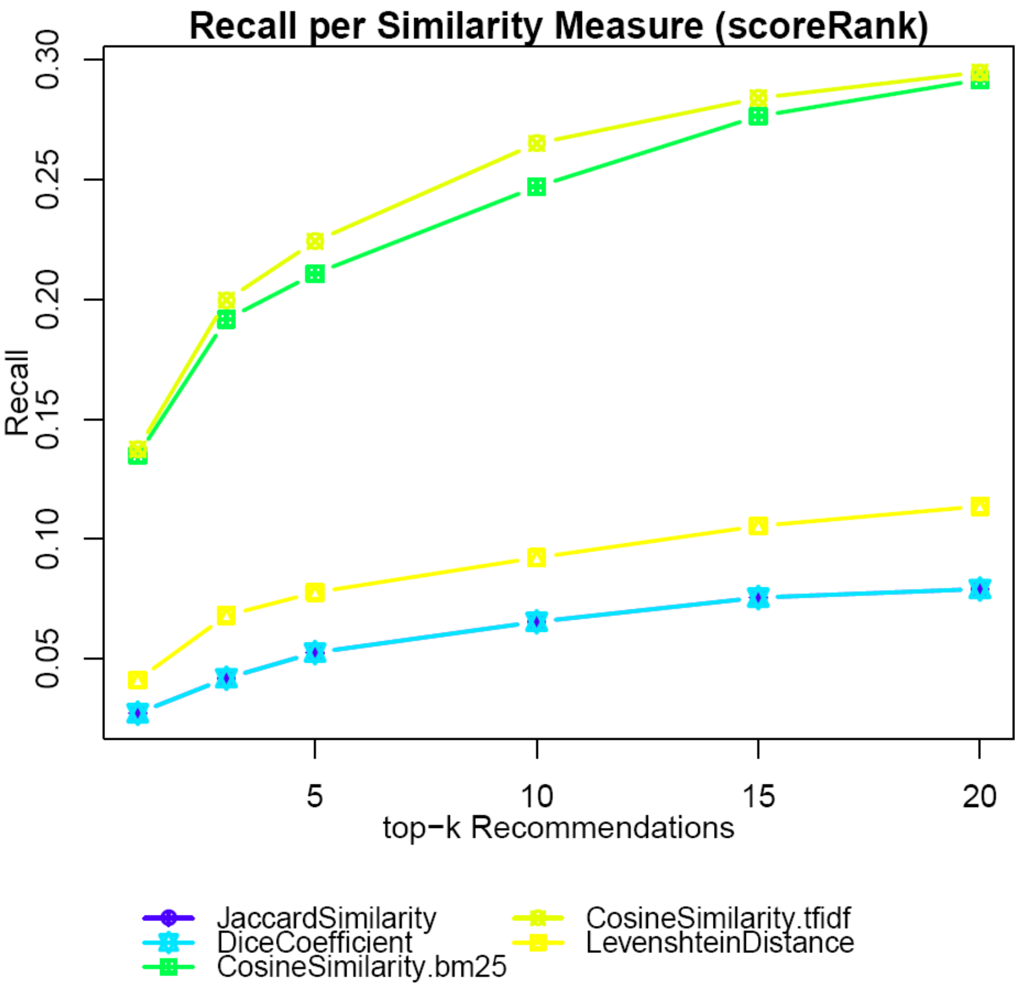
  → Precision, recall and F1 are set-based measures

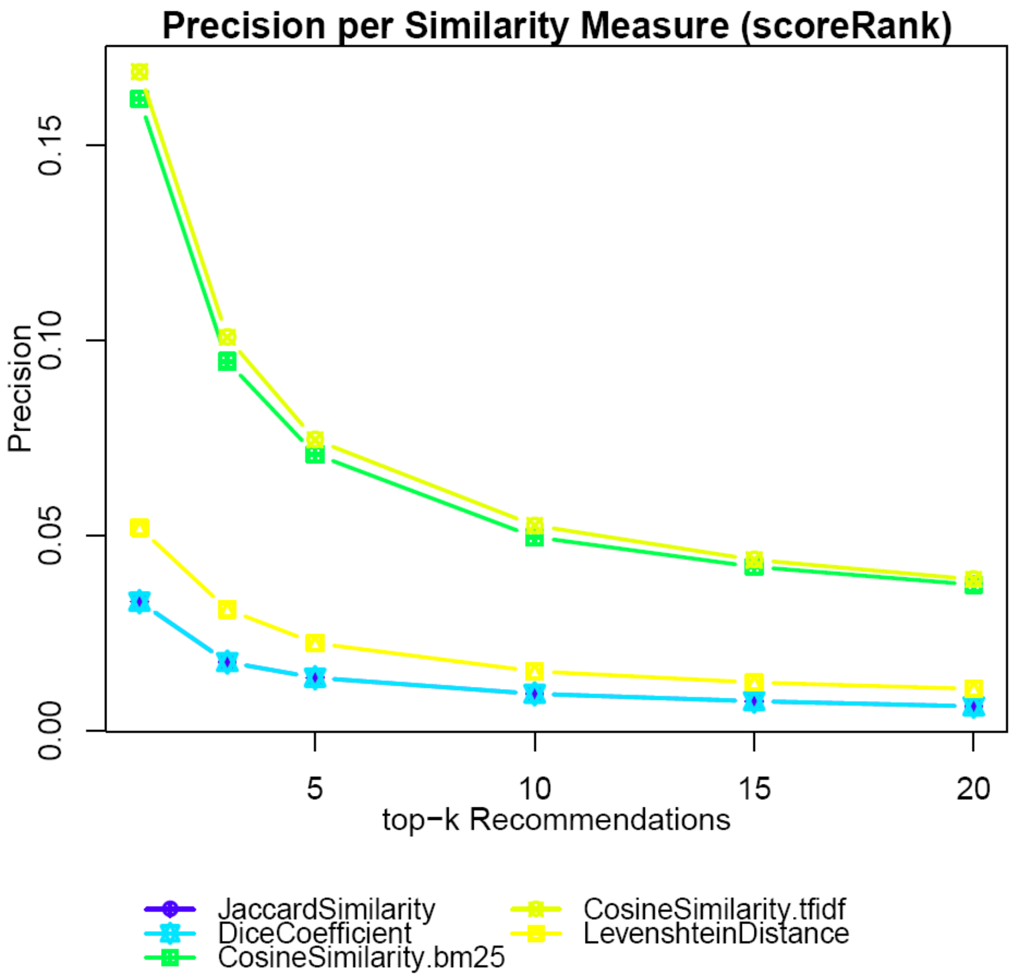
# Single Value Summaries

- Single value summaries required to quickly grasp performance of information retrieval system

- Precision@k, recall@k
  - Precision/recall of top-k results
  - Evaluate precision/recall performance of top-k ranked results
  - Mostly used for web search and recommender systems

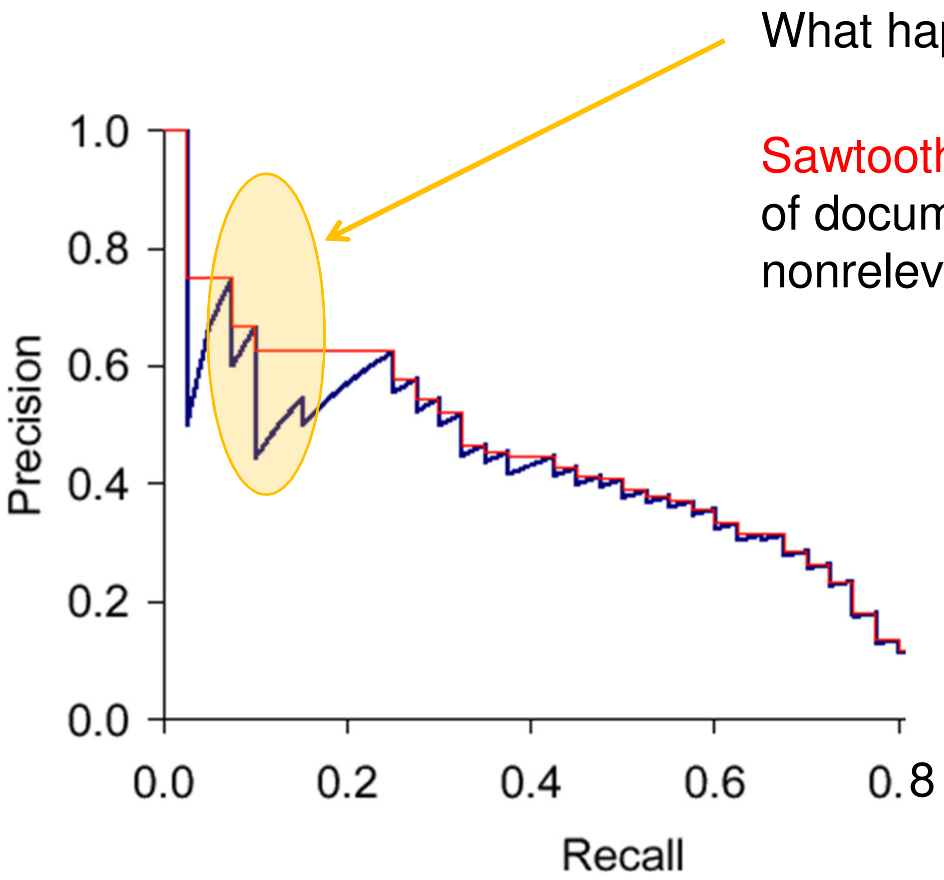
# Precision@k, Recall@k





# Precision-Recall Curve

- Precision-Recall curve for top-k results (blue line) (typically: with sawtooth shape)



What happened here?

Sawtooth shape: consider recall of documents k+1 and k if k+1 is nonrelevant

Figure taken from Manning, Raghavan, Schütze: Introduction to Information Retrieval, Cambridge University Press, 2008.

# Interpolated Precision

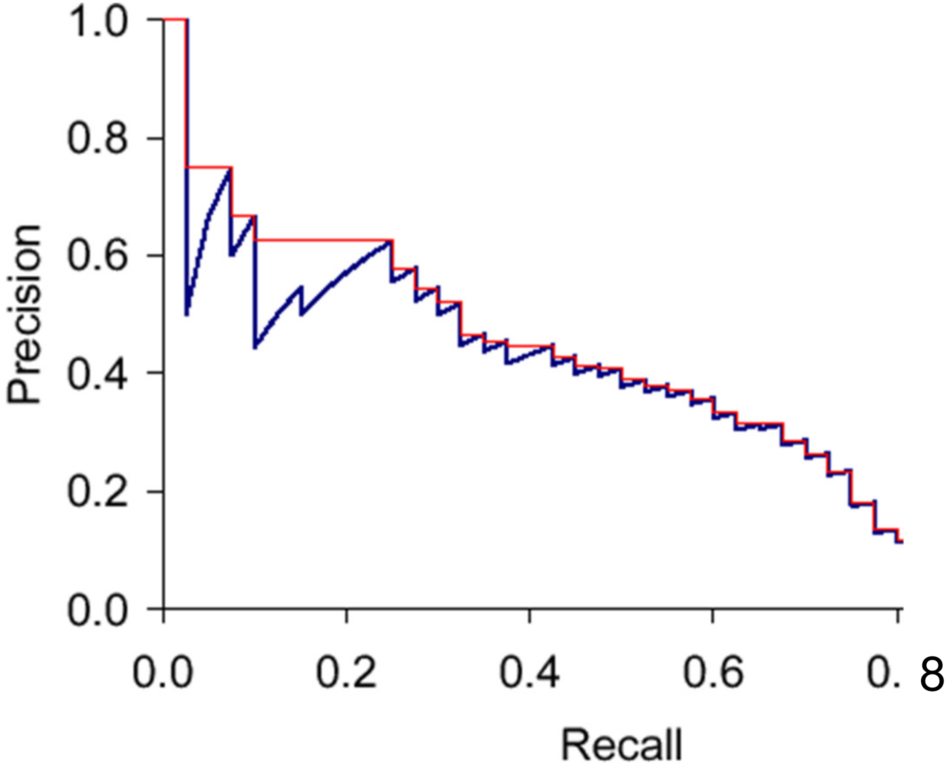
- Interpolated precision
  - No sawtooth shapes, smoothed curve

$$p_{inter} = \max_{r' \geq r} p(r')$$

  - Why?
  - You would look three documents further if you knew that there is a relevant document!

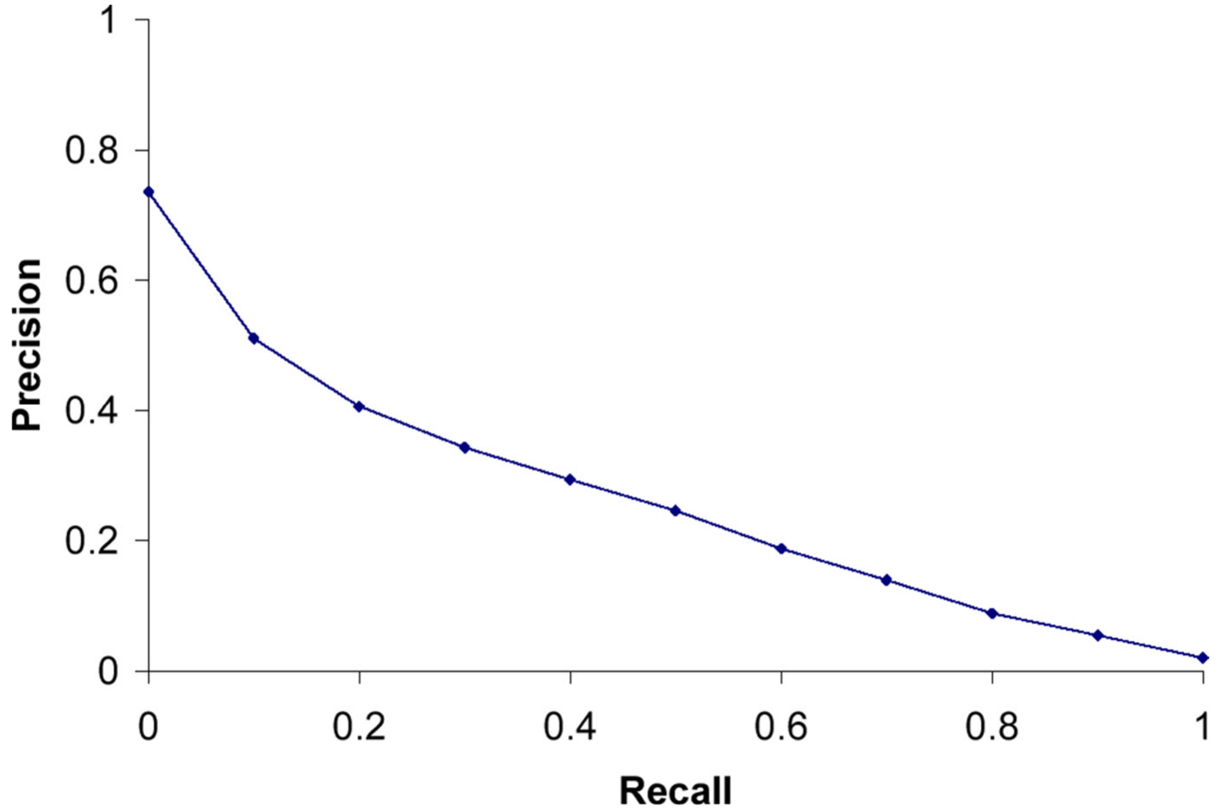
# Interpolated Precision

| recall | interpolated precision (red) |
|---|---|
| 0.0 | 1.00 |
| 0.1 | 0.67 |
| 0.2 | 0.63 |
| 0.3 | 0.55 |
| 0.4 | 0.45 |
| 0.5 | 0.41 |
| 0.6 | 0.36 |
| 0.7 | 0.28 |
| 0.8 | 0.13 |
| 0.9 | 0.10 |
| 1.0 | 0.08 |



# Eleven Point Interpolated Avg. Precision

- Eleven point interpolated average precision
  - Computed for each recall level as in prev. slide
  - For all tests
  - Averaged over whole set of information needs at the eleven levels 0.0, 0.1, 0.2, …0.9, 1.0



# Mean Average Precision

- Alternative:
- MAP – Mean Average Precision
  - Avg. precision = average precision value for top-k documents
  - Average of all information needs which are evaluated
  - Relevant documents for information need $q_j \in Q = \{d_1, \dots d_{m\_j}\}$
  - $R_{jk}$ = set of ranked retrieval results (Q queries, $m\_j$ docs per Query)

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m} \sum_{k=1}^{m_j} precision(R\_jk)$$

  - For single information need: avg. precision approx. area under uninterpolated precision-recall curve
  - Stable measure

# User-based Evaluation

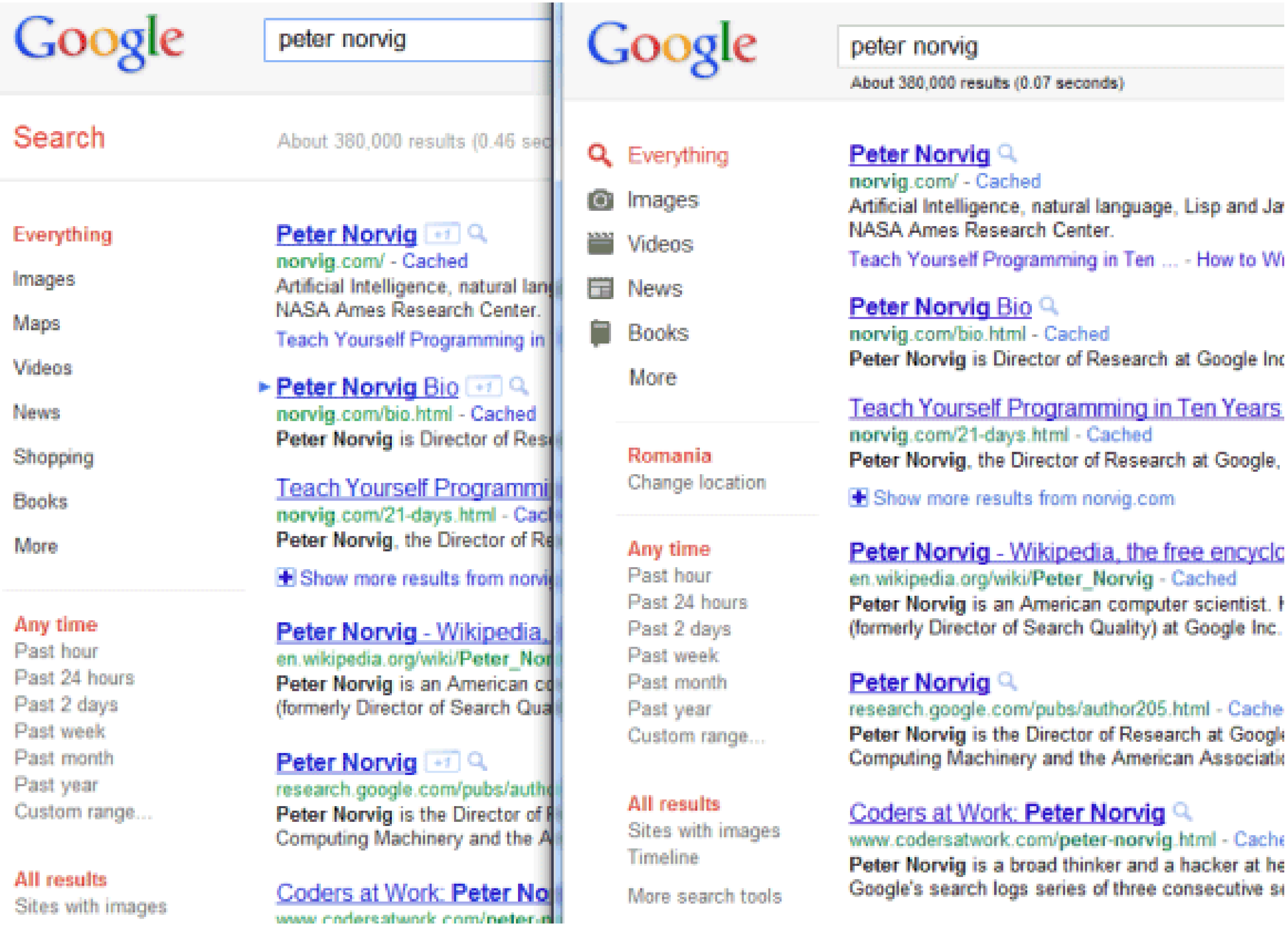
# User-Based Evaluation

- Up until now: user happiness and performance assessed by automated tests and automatically computed metrics.

- What about not directly measurable influence factors?
  - Usability
  - User interface design
  - User interaction with system

- Changes on these factors have to be done carefully
- Pre-tests

# UI Changes



About 380,000 results (0.46 sec

Peter Norvig
norvig.com/ - Cached
Artificial Intelligence, natural lan
NASA Ames Research Center.
Teach Yourself Programming in

Peter Norvig Bio
norvig.com/bio.html - Cached
Peter Norvig is Director of Rese

Teach Yourself Programmi
norvig.com/21-days.html - Cac
Peter Norvig, the Director of Re
Show more results from norvi

Peter Norvig - Wikipedia,
en.wikipedia.org/wiki/Peter_Nor
Peter Norvig is an American co
(formerly Director of Search Qua

Peter Norvig
research.google.com/pubs/auth
Peter Norvig is the Director of R
Computing Machinery and the A

Coders at Work: Peter No
www.codersatwork.com/peter-n

About 380,000 results (0.07 seconds)

Peter Norvig
norvig.com/ - Cached
Artificial Intelligence, natural language, Lisp and Ja
NASA Ames Research Center.
Teach Yourself Programming in Ten ... - How to Wi

Peter Norvig Bio
norvig.com/bio.html - Cached
Peter Norvig is Director of Research at Google Inc

Teach Yourself Programming in Ten Years
norvig.com/21-days.html - Cached
Peter Norvig, the Director of Research at Google,
Show more results from norvig.com

Peter Norvig - Wikipedia, the free encyclo
en.wikipedia.org/wiki/Peter_Norvig - Cached
Peter Norvig is an American computer scientist. H
(formerly Director of Search Quality) at Google Inc.

Peter Norvig
research.google.com/pubs/author205.html - Cache
Peter Norvig is the Director of Research at Google
Computing Machinery and the American Associati

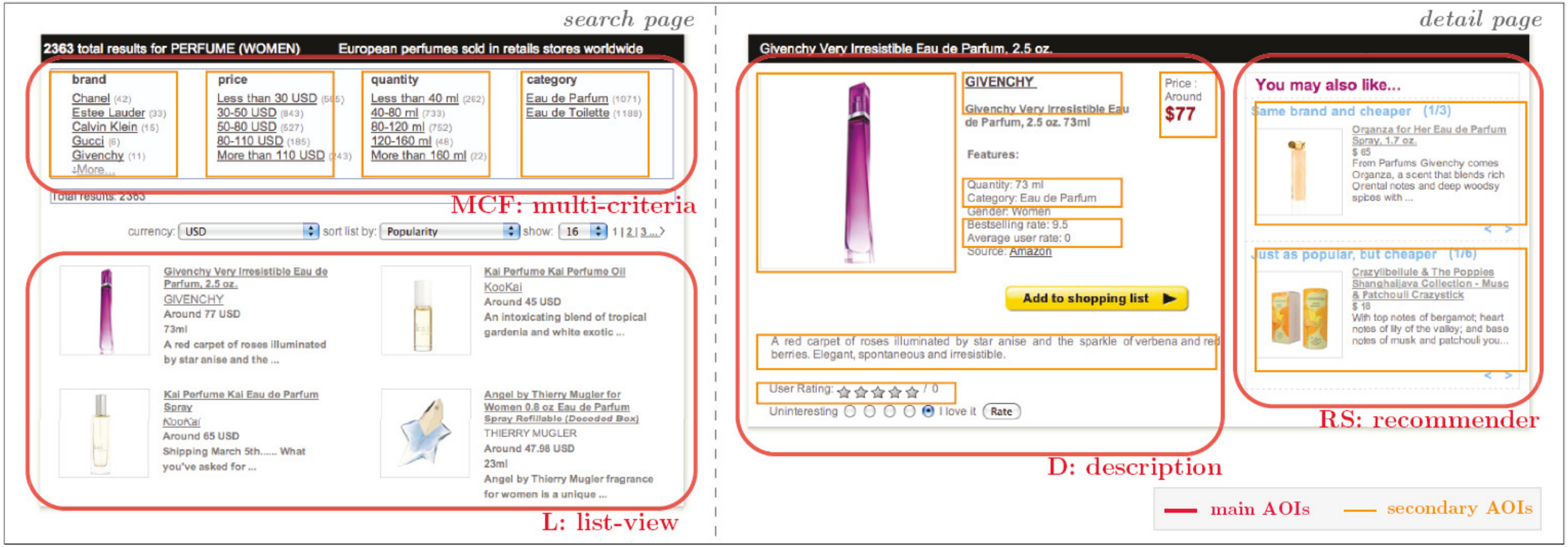
Coders at Work: Peter Norvig
www.codersatwork.com/peter-norvig.html - Cache
Peter Norvig is a broad thinker and a hacker at he
Google's search logs series of three consecutive se

Figure taken from http://googlesystem.blogspot.co.at/

# Human Experimentation

- Small number of test users
- Make small changes to UI, observe human behaviour and assess preferences
  - Video
  - Eyetracking
  - Questionnaire
- Evaluations are run in a lab
- Human users have to be selected carefully, choice of users is crucial for evaluation outcome
- Costly
- Users know that they are observed and their behaviour is evaluated

# Eye Tracking



2363 total results for PERFUME (WOMEN) European perfumes sold in retails stores worldwide

| brand | price | quantity | category |
|---|---|---|---|
| Chanel (42) | Less than 30 USD (565) | Less than 40 ml (262) | Eau de Parfum (1071) |
| Estee Lauder (33) | 30-50 USD (843) | 40-80 ml (733) | Eau de Toilette (1188) |
| Calvin Klein (15) | 50-80 USD (527) | 80-120 ml (752) | |
| Gucci (6) | 80-110 USD (185) | 120-160 ml (48) | |
| Givenchy (11) | More than 110 USD (243) | More than 160 ml (22) | |
| More... | | | |

Total results: 2363

currency: USD sort list by: Popularity show: 16 1 | 2 | 3 ...>

Givenchy Very Irresistible Eau de Parfum, 2.5 oz.
GIVENCHY
Around 77 USD
73ml
A red carpet of roses illuminated by star anise and the ...

Kai Perfume Kai Perfume Oil
KooKai
Around 45 USD
An intoxicating blend of tropical gardenia and white exotic ...

Kai Perfume Kai Eau de Parfum Spray
KooKai
Around 65 USD
Shipping March 5th...... What you've asked for ...

Angel by Thierry Mugler for Women 0.8 oz Eau de Parfum Spray Refillable (Decoded Box)
THIERRY MUGLER
Around 47.98 USD
23ml
Angel by Thierry Mugler fragrance for women is a unique ...

Givenchy Very Irresistible Eau de Parfum, 2.5 oz.

GIVENCHY
Givenchy Very Irresistible Eau de Parfum, 2.5 oz. 73ml

Price : Around
$77

Features:

Quantity: 73 ml
Category: Eau de Parfum
Gender: Women
Bestselling rate: 9.5
Average user rate: 0
Source: Amazon

Add to shopping list ▶

A red carpet of roses illuminated by star anise and the sparkle of verbena and red berries. Elegant, spontaneous and irresistible.

User Rating: ☆☆☆☆☆ / 0
Uninteresting ○ ○ ○ ○ ● I love it Rate

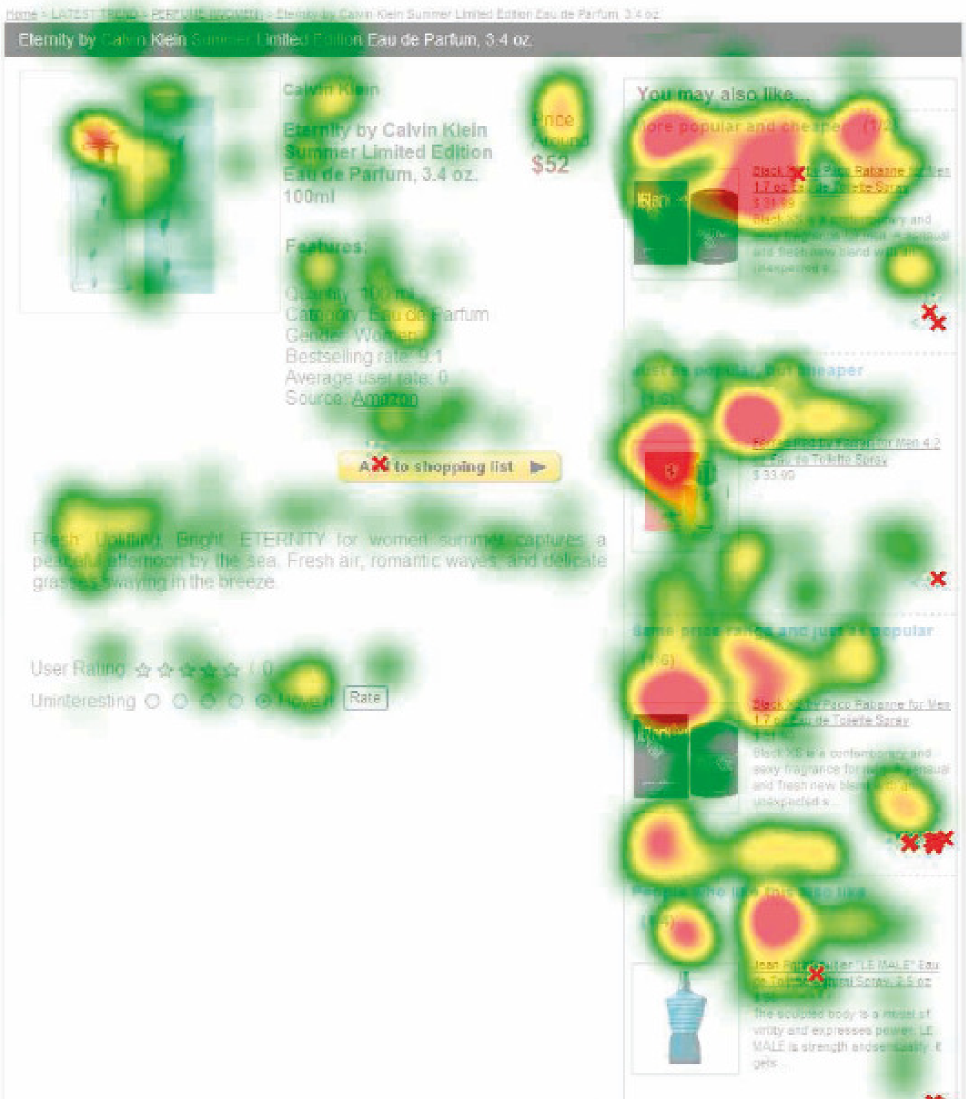
You may also like...

Same brand and cheaper (1/3)
Organza for Her Eau de Parfum Spray, 1.7 oz.
$ 65
From Parfums Givenchy comes Organza, a scent that blends rich Orental notes and deep woodsy spices with ...
< >

Just as popular, but cheaper (1/6)
Crazylibellule & The Poppies Shanghaijava Collection - Musc & Patchouli Crazystick
$ 18
With top notes of bergamot; heart notes of lily of the valley; and base notes of musk and patchouli you...
< >

Figure taken from S. Castagnos, N. Jones, P. Pu: "Eye-tracking product recommenders' usage"

# Eye Tracking



Home > LATEST TREND > PERFUME (WOMEN) > Eternity by Calvin Klein Summer Limited Edition Eau de Parfum, 3.4 oz

Eternity by Calvin Klein Summer Limited Edition Eau de Parfum, 3.4 oz

Calvin Klein

Eternity by Calvin Klein Summer Limited Edition Eau de Parfum, 3.4 oz. 100ml

Price Around $52

Features:

Quantity: 100ml
Category: Eau de Parfum
Gender: Women
Bestselling rate: 9.1
Average user rate: 0
Source: Amazon

Add to shopping list ▶

Fresh. Uplifting. Bright. ETERNITY for women summer captures a peaceful afternoon by the sea. Fresh air, romantic waves, and delicate grasses swaying in the breeze.

User Rating ☆ ☆ ☆ ☆ ☆ / 0

Uninteresting ○ ○ ○ ○ ○ Novel Rate

You may also like...

More popular and cheaper (1/2)

Black XS by Paco Rabanne for Men 1.7 oz Eau de Toilette Spray
$ 31.99
Black XS is a contemporary and sexy fragrance for men. A sensual and fresh new blend with an unexpected s...

Just as popular, but cheaper (1/3)

Ferrari Red by Ferrari for Men 4.2 oz Eau de Toilette Spray
$ 33.99

Same price range and just as popular (1/3)

Black XS by Paco Rabanne for Men 1.7 oz Eau de Toilette Spray
Black XS is a contemporary and sexy fragrance for men. A sensual and fresh new blend with an unexpected s...

People who like this also like (1/5)

Jean Paul Gaultier "LE MALE" Eau de Toilette Natural Spray 2.5 oz
The sculpted body is a model of virility and expresses power. LE MALE is strength and sensuality. It gets...

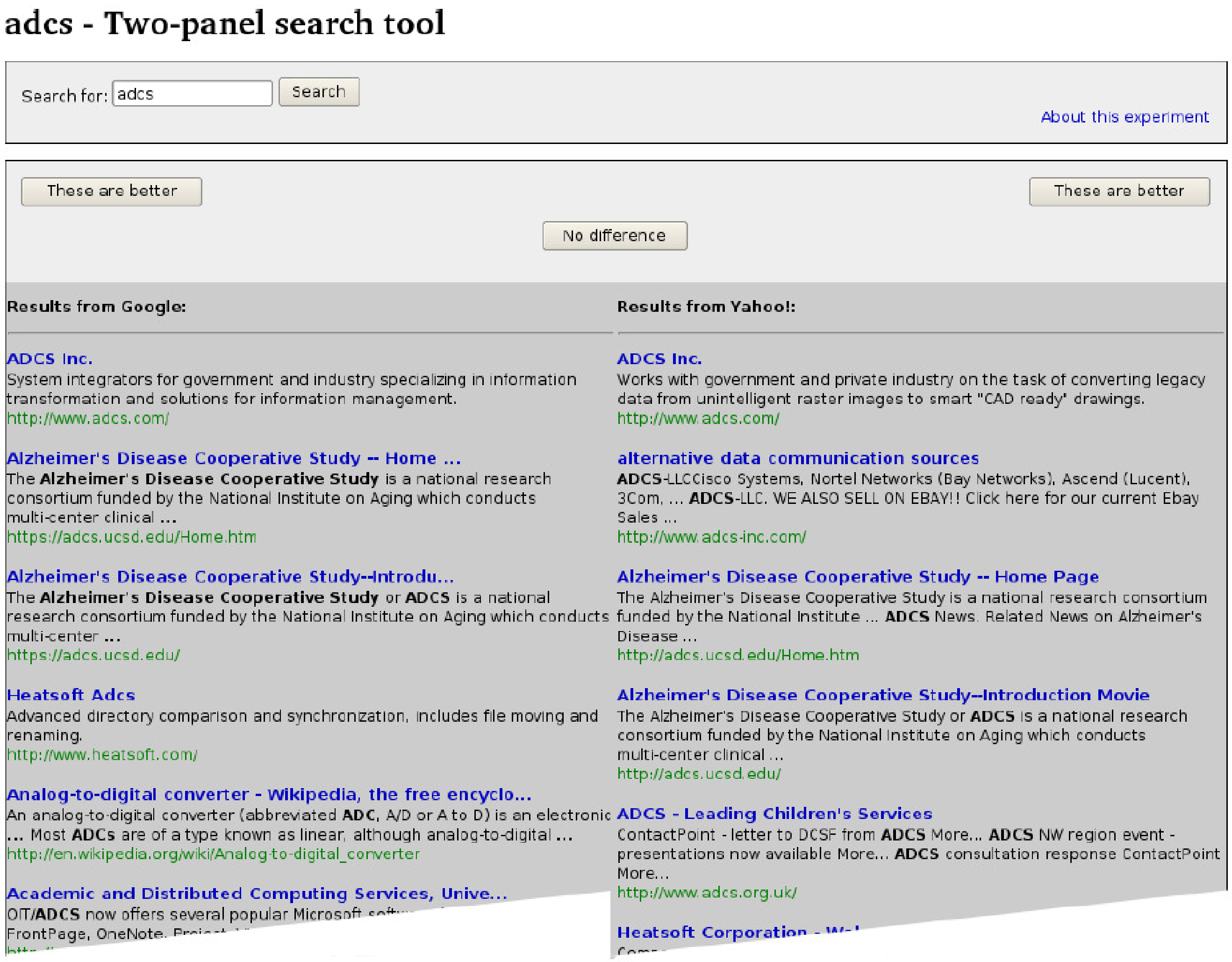
Figure taken from S. Castagnos, N. Jones, P. Pu: "Eye-tracking product recommenders' usage"

# Side-by-Side Panels

- Evaluate results produced by two systems side by side
- Perfect for search engine result lists
- Side by side: system A on left side of screen, system B on right side of screen (e.g. results for same query)

- Users are aware that their behaviour is evaluated
- Evaluations can be done anonymously

# Side-by-Side Panels



adcs - Two-panel search tool

Search for: adcs Search

About this experiment

These are better

These are better

No difference

**Results from Google:**

ADCS Inc.
System integrators for government and industry specializing in information transformation and solutions for information management.
http://www.adcs.com/

Alzheimer's Disease Cooperative Study -- Home ...
The **Alzheimer's Disease Cooperative Study** is a national research consortium funded by the National Institute on Aging which conducts multi-center clinical **...**
https://adcs.ucsd.edu/Home.htm

Alzheimer's Disease Cooperative Study--Introdu...
The **Alzheimer's Disease Cooperative Study** or **ADCS** is a national research consortium funded by the National Institute on Aging which conducts multi-center **...**
https://adcs.ucsd.edu/

Heatsoft Adcs
Advanced directory comparison and synchronization, includes file moving and renaming.
http://www.heatsoft.com/

Analog-to-digital converter - Wikipedia, the free encyclo...
An analog-to-digital converter (abbreviated **ADC**, A/D or A to D) is an electronic **...** Most **ADCs** are of a type known as linear, although analog-to-digital **...**
http://en.wikipedia.org/wiki/Analog-to-digital_converter

Academic and Distributed Computing Services, Unive...
OIT/**ADCS** now offers several popular Microsoft
FrontPage, OneNote,

**Results from Yahoo!:**

ADCS Inc.
Works with government and private industry on the task of converting legacy data from unintelligent raster images to smart "CAD ready" drawings.
http://www.adcs.com/

alternative data communication sources
**ADCS**-LLCCisco Systems, Nortel Networks (Bay Networks), Ascend (Lucent), 3Com, ... **ADCS**-LLC. WE ALSO SELL ON EBAY!! Click here for our current Ebay Sales ...
http://www.adcs-inc.com/

Alzheimer's Disease Cooperative Study -- Home Page
The Alzheimer's Disease Cooperative Study is a national research consortium funded by the National Institute ... **ADCS** News. Related News on Alzheimer's Disease ...
http://adcs.ucsd.edu/Home.htm

Alzheimer's Disease Cooperative Study--Introduction Movie
The Alzheimer's Disease Cooperative Study or **ADCS** is a national research consortium funded by the National Institute on Aging which conducts multi-center clinical ...
http://adcs.ucsd.edu/

ADCS - Leading Children's Services
ContactPoint - letter to DCSF from **ADCS** More... **ADCS** NW region event - presentations now available More... **ADCS** consultation response ContactPoint More...
http://www.adcs.org.uk/

Heatsoft Corporation -

Figure taken from P. Bailey, P. Thomas, D. Hawking: Does brandname influence perceived search result quality? Yahoo!, Google, and WebKumara

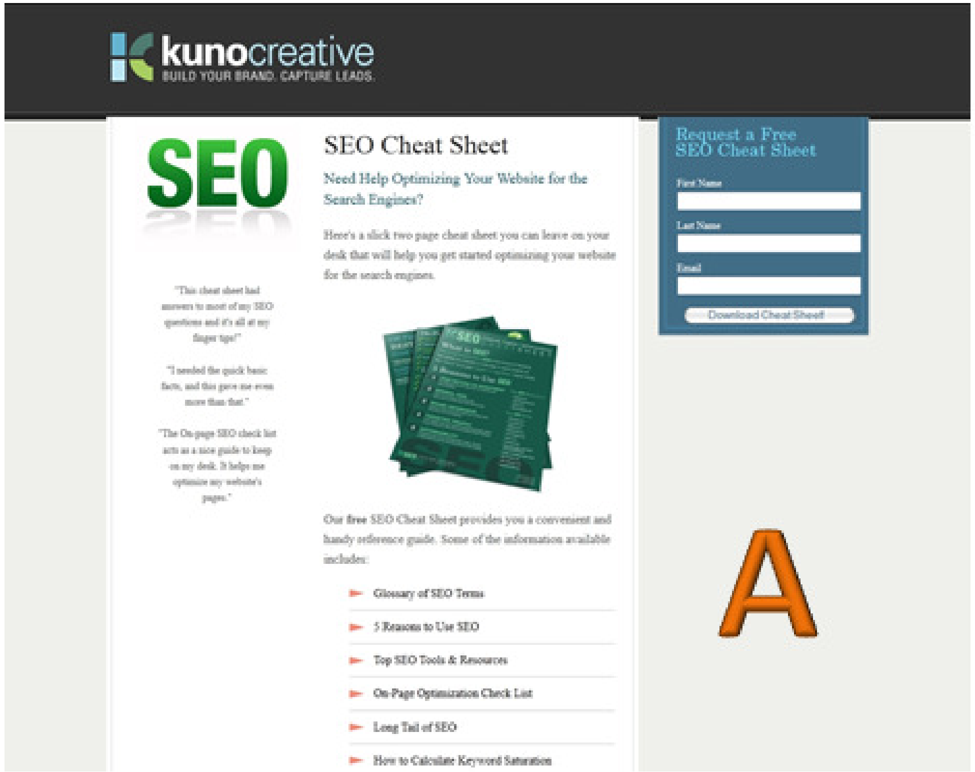
# A/B and Multivariate Testing

Online tests

- Show modified version of (mostly) webpage to pre-selected amount of users and observe reaction
- A/B tests: compare two versions
- Multivariate tests: compare multiple versions
- Huge numbers of participating users possible, preselection based on e.g. demographics

- Users are not aware of being part of an evaluation
- Large number of users
- Many evaluations possible

# A/B and Multivariate Testing



kunocreative
BUILD YOUR BRAND. CAPTURE LEADS.

SEO

## SEO Cheat Sheet

Need Help Optimizing Your Website for the Search Engines?

Here's a slick two page cheat sheet you can leave on your desk that will help you get started optimizing your website for the search engines.

"This cheat sheet had answers to most of my SEO questions and it's all at my finger tips!"

"I needed the quick basic facts, and this gave me even more than that."

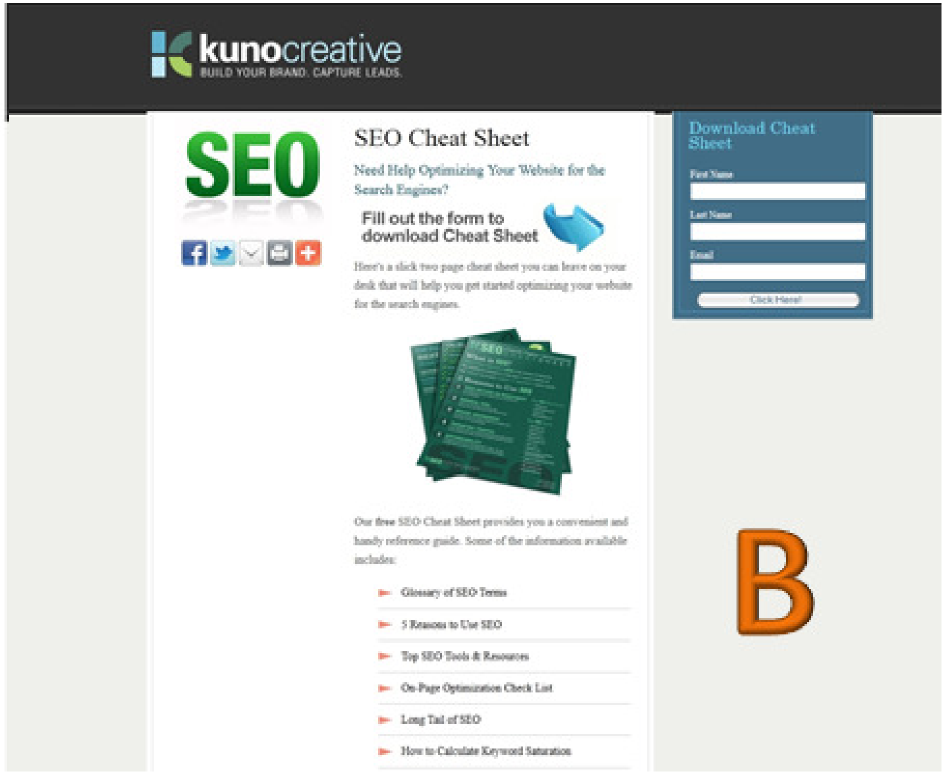
"The On-page SEO check list acts as a nice guide to keep on my desk. It helps me optimize my website's pages."

Our free SEO Cheat Sheet provides you a convenient and handy reference guide. Some of the information available includes:

- Glossary of SEO Terms
- 5 Reasons to Use SEO
- Top SEO Tools & Resources
- On-Page Optimization Check List
- Long Tail of SEO
- How to Calculate Keyword Saturation

Request a Free SEO Cheat Sheet

First Name
Last Name
Email



kunocreative
BUILD YOUR BRAND. CAPTURE LEADS.

SEO

## SEO Cheat Sheet

Need Help Optimizing Your Website for the Search Engines?

Fill out the form to download Cheat Sheet

Here's a slick two page cheat sheet you can leave on your desk that will help you get started optimizing your website for the search engines.

Our free SEO Cheat Sheet provides you a convenient and handy reference guide. Some of the information available includes:

- Glossary of SEO Terms
- 5 Reasons to Use SEO
- Top SEO Tools & Resources
- On-Page Optimization Check List
- Long Tail of SEO
- How to Calculate Keyword Saturation

Download Cheat Sheet

First Name
Last Name
Email

Figure taken from http://www.kunocreative.com/blog/bid/71196/Debunking-Inbound-Marketing-Groupthink-with-Multivariate-Testing

# A/B and Multivariate Testing

- Netflix A/B tests for PlayStation 3

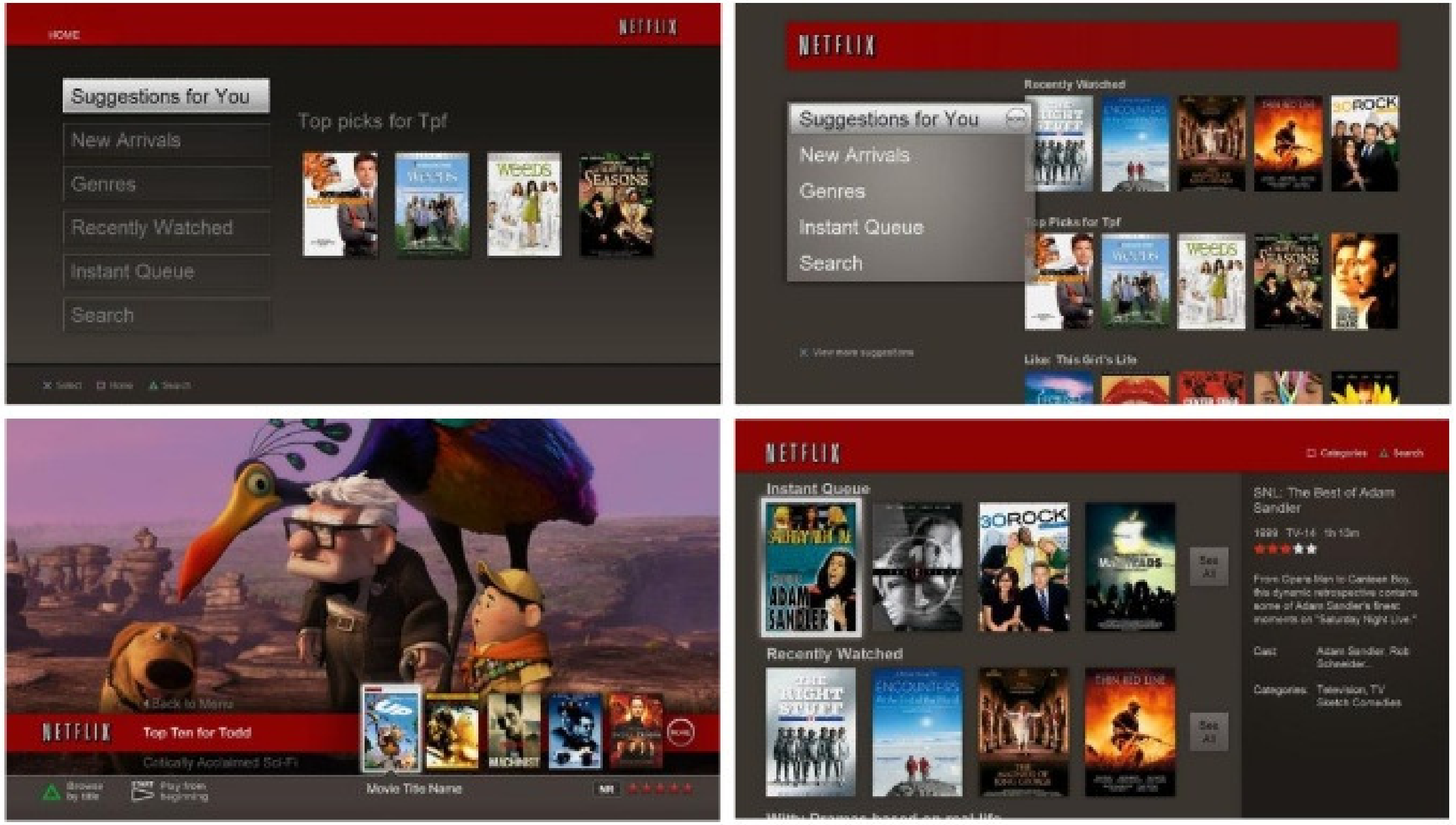


Figure taken from http://gigaom.com/video/netflix-ui-innovation/

# Implicit Feedback / Evaluation

- Implicit metrics which can be collected without direct interaction with the user:
  - Time spent on page
  - Links clicked
  - Rank of item clicked
  - Maximum rank of clicked item
  - Link traversal
  - Items on wishlist
  - Items within basket
  - etc.

- Very useful for A/B testing