CS3245

Information Retrieval

9

Lecture 9: IR Evaluation

Live Q&A
https://pollev.com/jin

Last Time

The VSM Reloaded

… optimized for your pleasure!

Improvements to the computation and selection process

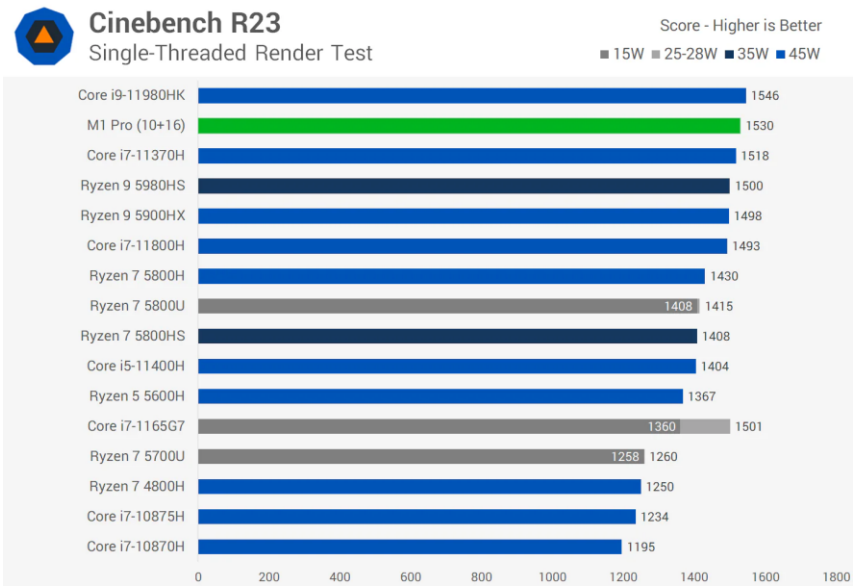
Use of heuristics to avoid unnecessary / time consuming computations

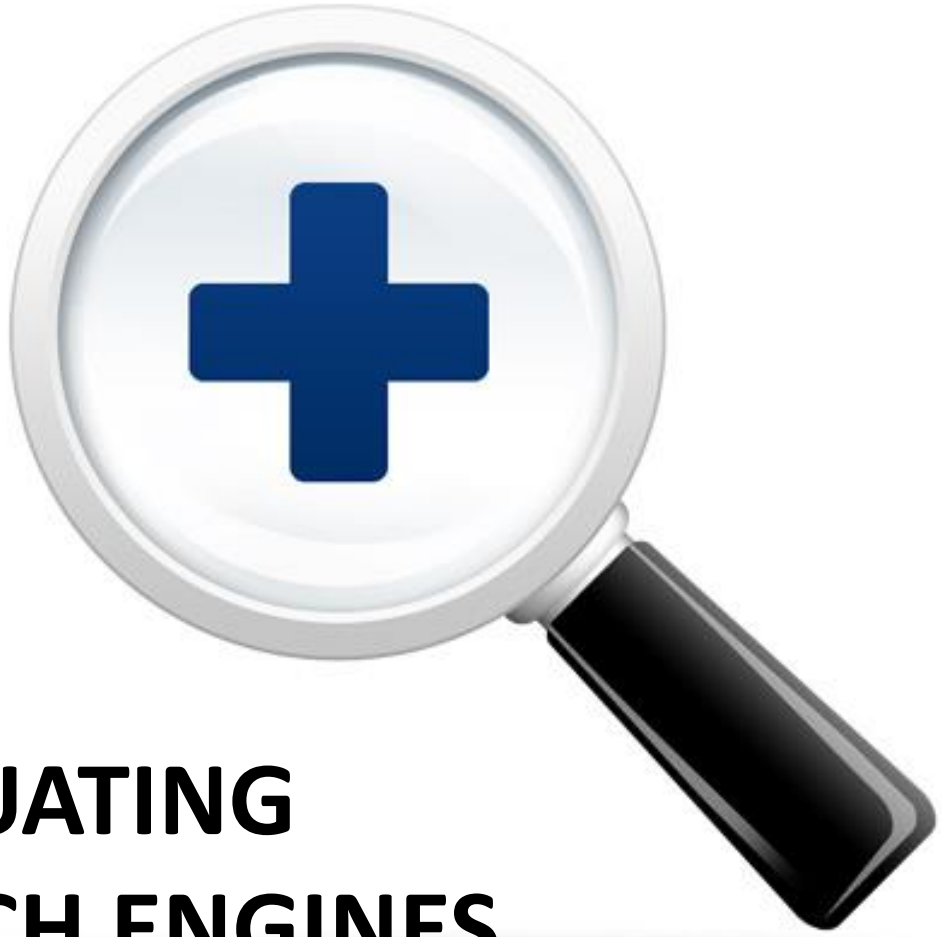
1. Index elimination
2. Tiered lists
3. Early termination
4. Cluster pruning

Mechanism to incorporate different sources of information

Today: Evaluation

- How to assess the IR systems / approaches?
  - Benchmarks
  - A/B Testing





EVALUATING SEARCH ENGINES

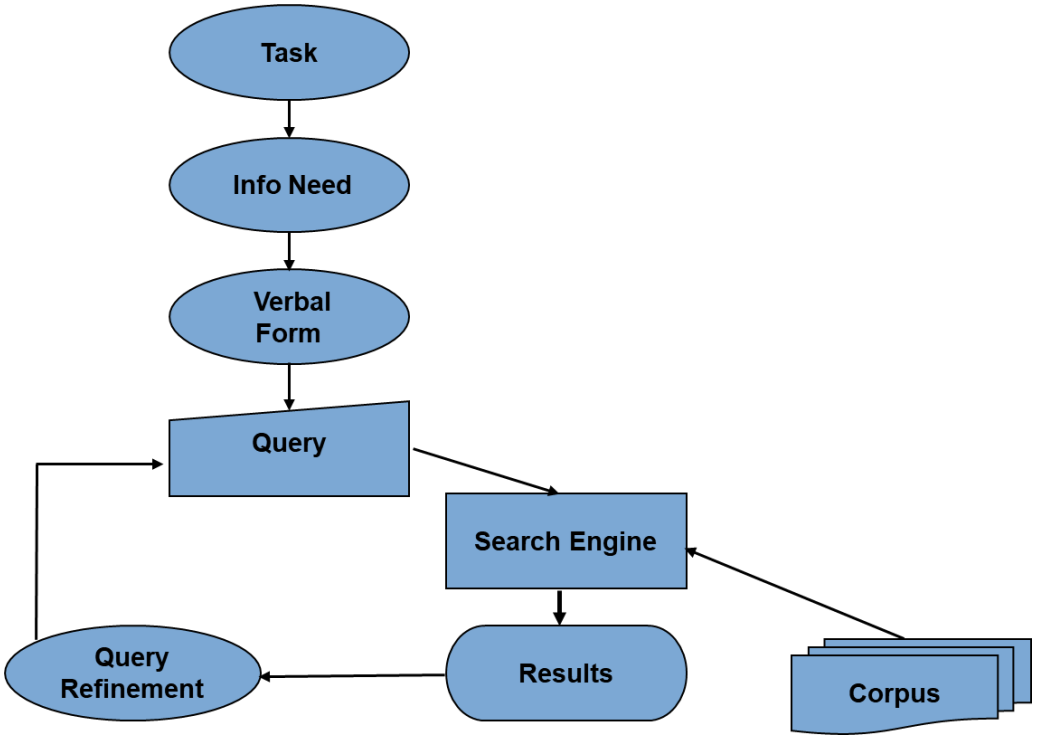
Evaluating an IR system

- How fast does it index?
  - Number of documents/hour
- How fast does it search?
  - Latency as a function of index size
  - Speed on long / complex queries
- Correctness of the implementation?
  - Computation of intersection for AND queries
- Expressiveness of query language?
  - Ability to express complex information needs

Evaluating an IR system

- But most importantly, how **relevant** are results?

- A quick recap on the IR process



Evaluating an IR system

- But most importantly, how **relevant** are results?

- **3** key elements for measuring relevance
  1. A set document collection
  2. A set suite of queries
  3. A usually binary assessment of either Relevant or Non-relevant for each query and each document
     - Some work on graded relevance, but not the standard

Measures for a search engine

- But most importantly, how **relevant** are results?

- Relevance is assessed relative to the **information need** *not* the **query**
  - E.g., Information need: *I'm looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.*
  - Query: ***wine red white heart attack effective***
  - i.e., we should find out whether the doc addresses the information need, not whether it contains the terms.

Unranked retrieval evaluation: Precision and Recall

- **Precision (P):** fraction of **retrieved** docs that are **relevant**, i.e., # of relevant doc retrieved / total # of documents retrieved

- **Recall (R):** fraction of **relevant** docs that are **retrieved**, i.e., # of relevant doc retrieved / total # of relevant documents

- Example:
  - For a collection of 5 docs {1, 2, 3, 4, 5} and a query, **3** docs {1, 2, 3} are relevant (and the rest are not). A system returns **2** docs {1, 4}.
  - P = 1 / 2 = 0.5
  - R = 1 / 3 = 0.33

Precision/Recall

- You can get
  - High precision (but low recall) by retrieving only 1 doc and making sure that it is relevant!
  - High recall (but low precision) by retrieving all docs!

- In a good system, precision decreases as either the number of docs retrieved or recall increases
  - This is not a theorem, but a result with strong empirical confirmation

A combined measure: F_1

- Combined measure that assesses precision / recall tradeoff is F_1 measure (**harmonic** mean):

$$\mathrm{F}_1 = \frac{2PR}{P+R}$$

- Harmonic mean is a conservative average
  - Helps to reveal the lower value
- Example: P = 0.8, R = 0.2
  - **Arithmetic** mean = (P + R) / 2 = 0.5
  - **Harmonic** mean = F_1 = 0.32

A combined measure: F_1

- The general form is F measure (**weighted** harmonic mean):

$$F = \frac{(\beta^2+1)PR}{\beta^2 P+R}$$

- β can be used to adjust the relative importance of P and R
  - β = 1, (i.e., F_1) is balanced
  - β < 1, P is more important
  - β > 1, R is more important

Evaluating ranked results

- Relevant documents should be ranked higher than non-relevant documents

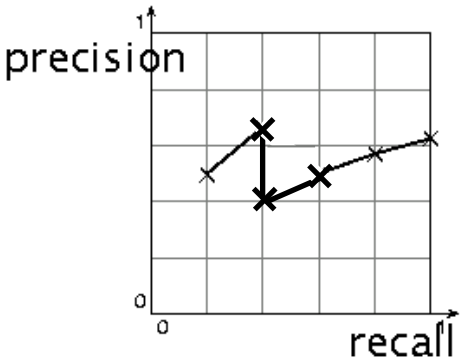
- Example:
  - For a collection of 5 docs {1, 2, 3, 4, 5} and a query, **3** docs {1, 2, 3} are relevant.
  - System A returns 5 docs in the order of {**1**, **2**, **3**, 4, 5}
  - System B returns 5 docs in the order of {**3**, 4, 5, **1**, **2**}
  - Which one is better?

Evaluating ranked results

- A precision-recall curve can be drawn by computing precision at different recall levels (i.e., every time a relevant document is retrieved)
- Example:
  - System B returns 5 docs in the order of {**3**, 4, 5, **1**, **2**}.
  - The key data points in the form of (**R**, **P**) are:
    - (0.33, 1) when doc **3** is retrieved
    - (0.66, 0.5) when doc **1** is retrieved
    - (1, 0.6) when doc **2** is retrieved

Interpolated precision

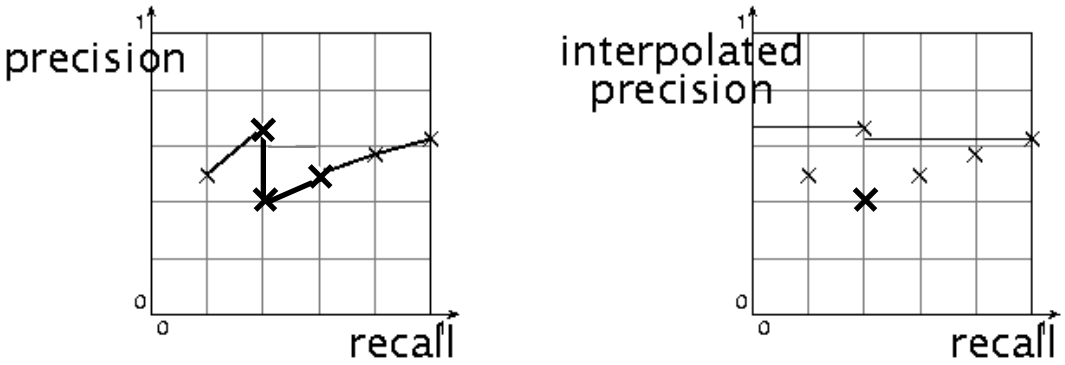
- Sometimes precision does increase with recall locally.



- This should be accounted for since the precision is not as bad as it seems at the low point.

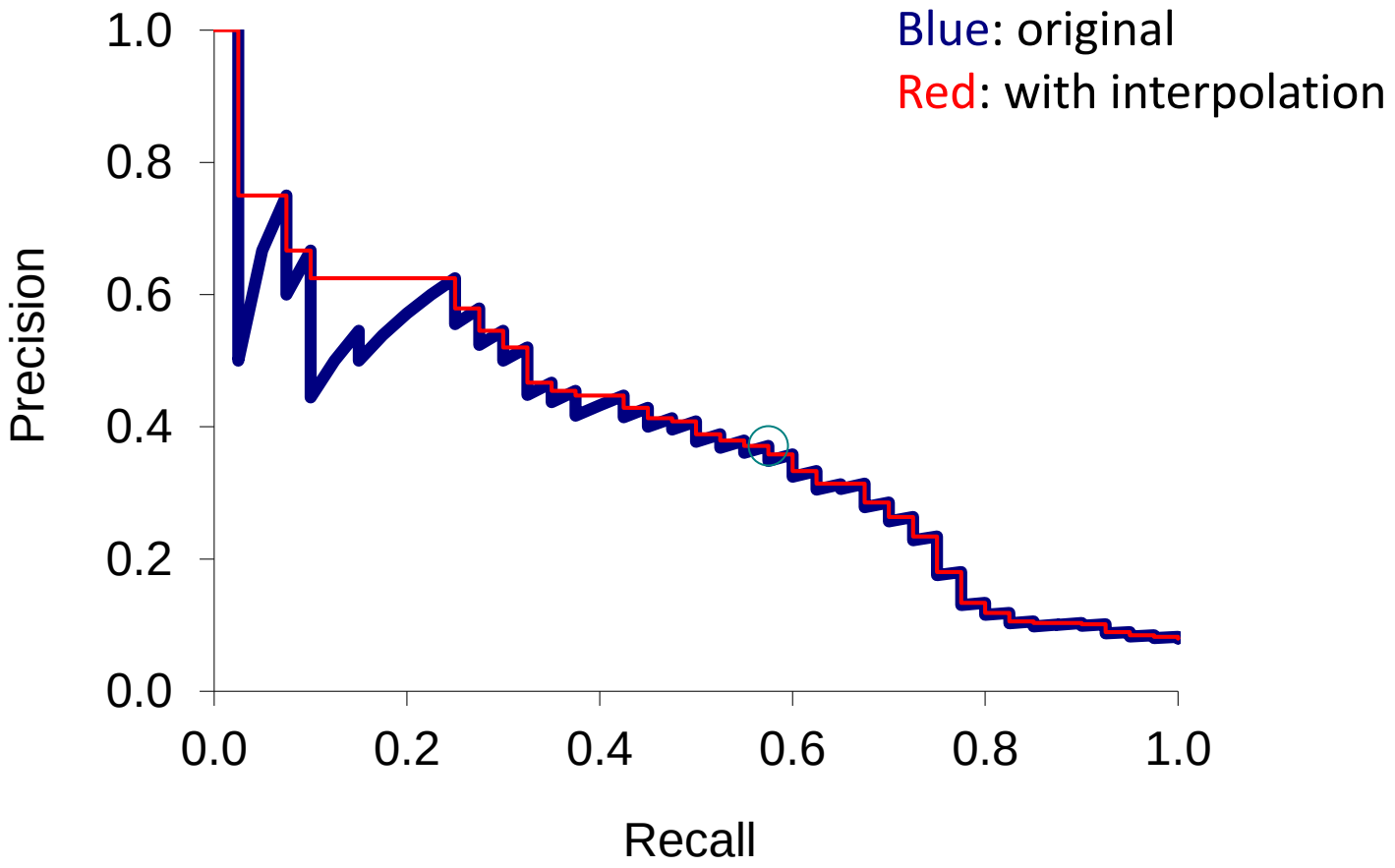
Interpolated precision

- So we take the **maximum** precision to the right of the value as the **interpolated** precision.



- Example:
  - Original data points: (0.33, 1), (0.66, 0.5) and (1, 0,6)
  - Interpolated data points: (0.33, 1), (0.66, **0.6**) and (1, 0.6)

A precision-recall curve



Evaluation

- Graphs are good, but often we want a summary measure!
  - Precision-at-k: Precision of top *k* results
    - Perhaps appropriate for most of web search: all people want are good matches on the first one or two result pages
    - But: averages badly and has an arbitrary parameters of k
  - 11-point interpolated average precision

    The standard measure in the early TREC competitions: you take the precision at 11 levels of recall varying from 0 to 1 by tenths of the documents, using interpolation (the value for 0 is always interpolated!), and average them
    - Evaluates performance at all recall levels

Yet more evaluation measures…

- Mean average precision (MAP)
  - Average of the precision value obtained for the top *k* documents, each time a relevant doc is retrieved
  - Avoids interpolation, use of fixed recall levels
  - MAP for query collection is arithmetic ave.
    - Macro-averaging: each query counts equally
- R-precision
  - If have known (though perhaps incomplete) set of relevant documents of size *Rel,* then calculate precision of top *Rel* docs returned
  - Perfect system could score 1.0.

Variance

- For a test collection, it is usual that a system does poorly on some information needs (e.g., MAP = 0.1) and excellent on others (e.g., MAP = 0.7)
- Indeed, it is usually the case that the variance in performance of the same system across queries is much greater than the variance of different systems on the same query.

- That is, there are easy information needs and hard ones!

CREATING TEST COLLECTIONS FOR EVALUATION

Test Collections

TABLE 4.3 Common Test Corpora

Collection	NDocs	NQrys	Size (MB)	Term/Doc	Q-D RelAss
ADI	82	35			
AIT	2109	14	2	400	>10,000
CACM	3204	64	2	24.5	
CISI	1460	112	2	46.5	
Cranfield	1400	225	2	53.1	
LISA	5872	35	3		
Medline	1033	30	1		
NPL	11,429	93	3		
OSHMED	34,8566	106	400	250	16,140
Reuters	21,578	672	28	131	
TREC	740,000	200	2000	89-3543	» 100,000

Scientific papers

Scientific papers

Medical

Medical

News

News

From document collections to test collections

Still need the other **2** things

1. Test queries
   - Must be relevant to docs available
   - Best designed by domain experts
   - Random query terms generally not a good idea
2. Relevance assessments
   - Human judges, time-consuming
   - Are human panels perfect?

Kappa measure for inter-judge (dis)agreement

- Kappa measure
  - Agreement measure among judges
  - Designed for categorical judgments
  - Corrects for chance agreement

- *Kappa (K) = [P(A)-P(E)]/[1-P(E)]*
- *P(A)* – proportion of time judges agree
- *P(E)* – what agreement would be by chance

- Gives 0 for chance agreement, 1 for total agreement

Kappa Measure: Example

# of docs	Judge 1	Judge 2	
300	Relevant	Relevant	**Agree**
70	Non-relevant	Non-relevant	
20	Relevant	Non-relevant	**Disagree**
10	Non-relevant	Relevant	

$P(A) = (300+70) / 400 = 0.925$

$P(\text{non-relevant}) = (70+10+70+20) / (400+400) = 0.2125$

- The chance of a document being assessed as non-relevant

$P(\text{relevant}) = (300+20+300+10) / (400+400) = 0.7875$

- The chance of a document being assessed as relevant

Kappa Measure: Example

# of docs	Judge 1	Judge 2	
300	Relevant	Relevant	**Agree**
70	Non-relevant	Non-relevant	
20	Relevant	Non-relevant	**Disagree**
10	Non-relevant	Relevant	

$P(E) = P(\text{non-relevant})^2$ ← The chance of a document being assessed as non-relevant twice

$+ P(\text{relevant})^2$ ← The chance of a document being assessed as relevant twice

$= 0.2125^2 + 0.7875^2 = 0.665$ ← This should be 0.6653125 with more accurate computation.

Kappa Measure: Example

Kappa = K = (0.925-0.665)/(1-0.665) = 0.776

- Kappa > 0.8 → Good agreement
- 0.67 < Kappa < 0.8 → Tentative conclusions

- Depend on purpose of study
- For >2 judges: average pairwise kappas (or ANOVA)

TREC

- TREC's Ad Hoc task from first 8 TRECs was the standard IR task
  - 50 detailed information needs a year
  - Human evaluation of **pooled** results returned
  - More recently other related things: Web, Hard, QA, interactive track

- A query from TREC 5 (1996)

```
<top>
<num>225</num>
<desc>What is the main function of the Federal
  Emergency Management Agency (FEMA) and the
  funding level provided to meet emergencies?
  Also, what resources are available to FEMA such
  as people, equipment, facilities?</desc>
</top>
```

Interjudge Agreement: TREC 3

information need	number of docs judged	disagreements	NR	R
51	211	6	4	2
62	400	157	149	8
67	400	68	37	31
95	400	110	108	2
127	400	106	12	94

Shows that there are queries that are easier than others

A/B Testing

A/B testing

Purpose: Test a single innovation (i.e., change)
Prerequisite: You have a large search engine up and running.

- Have most users use old system, but divert a small proportion of traffic (e.g., 1%) to the new system with the innovation.
- Evaluate with an "automatic" **Overall Evaluation Criterion (OEC)** like clickthrough on first result
- Now we can directly see if the innovation works.

Slide courtesy Microsoft Inc.

The HiPPO

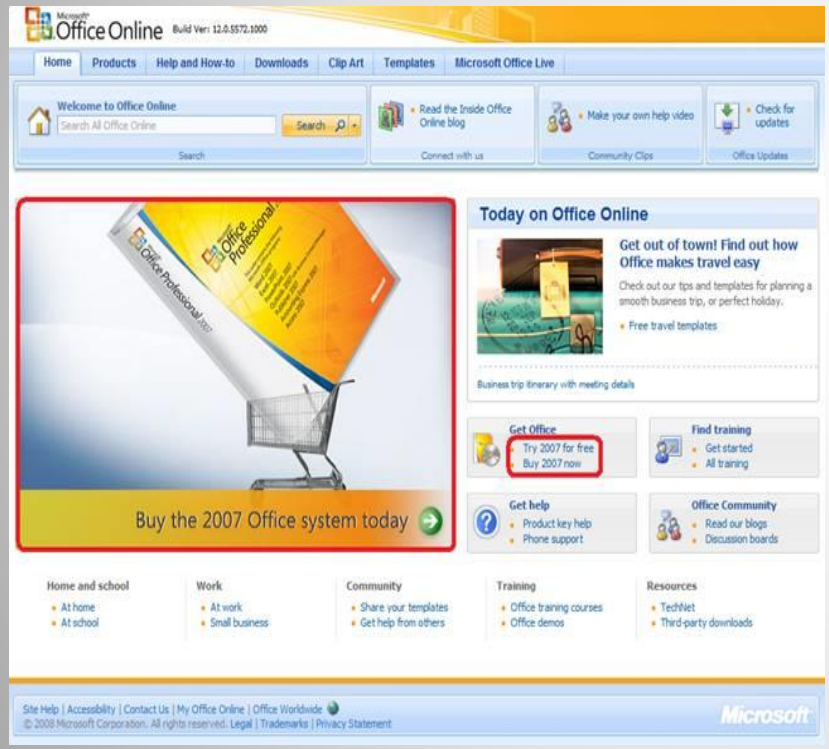
The less data, the stronger the opinions

- Our opinions are often wrong – get the data
- HiPPO stands for the Highest Paid Person's Opinion
- Hippos kill more humans than any other (non-human) mammal (really)
- Don't let HiPPOs in your org kill innovative ideas. ExPeriment!
- We give out these toy HiPPOs at Microsoft

Slide courtesy Microsoft Inc.

Office Online

Test new design for Office Online homepage



A

B

Which one is better? (OEC: Clicks on revenue generating links)

Slide courtesy Microsoft Inc.

Office Online

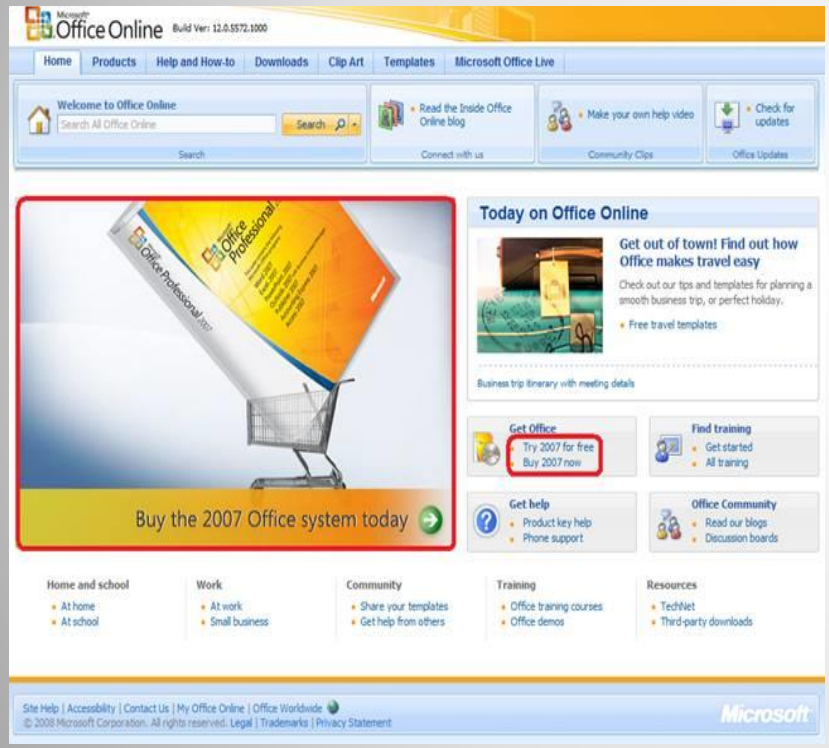
- B was 64% worse
- The Office Online team wrote

  A/B testing is a fundamental and critical Web services… consistent use of A/B testing could save the company millions of dollars

Slide courtesy Microsoft Inc.

Pitfall: Wrong OEC

Remember this example?



A

B

Slide courtesy Microsoft Inc.

Pitfall: Wrong OEC

- B had a drop in the OEC of 64%
- Were sales correspondingly less also?
- No. The experiment is valid if the conversion from a click to purchase is similar
- The price was shown only in B, sending more qualified purchasers to the pipeline
- Lesson: measure what you really need to measure, even if it's difficult!

Summary: Evaluation

Different schemes for lab versus in-the-wild testing

- Benchmark testing
- A/B testing

Resources:

- IIR 8, MIR Chapter 3, MG 4.5