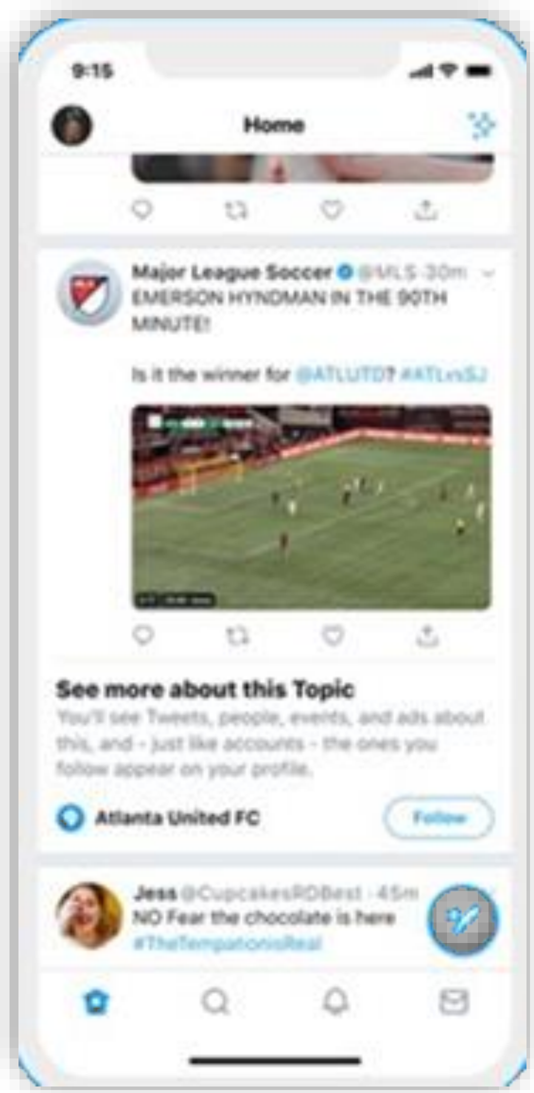
Evaluation

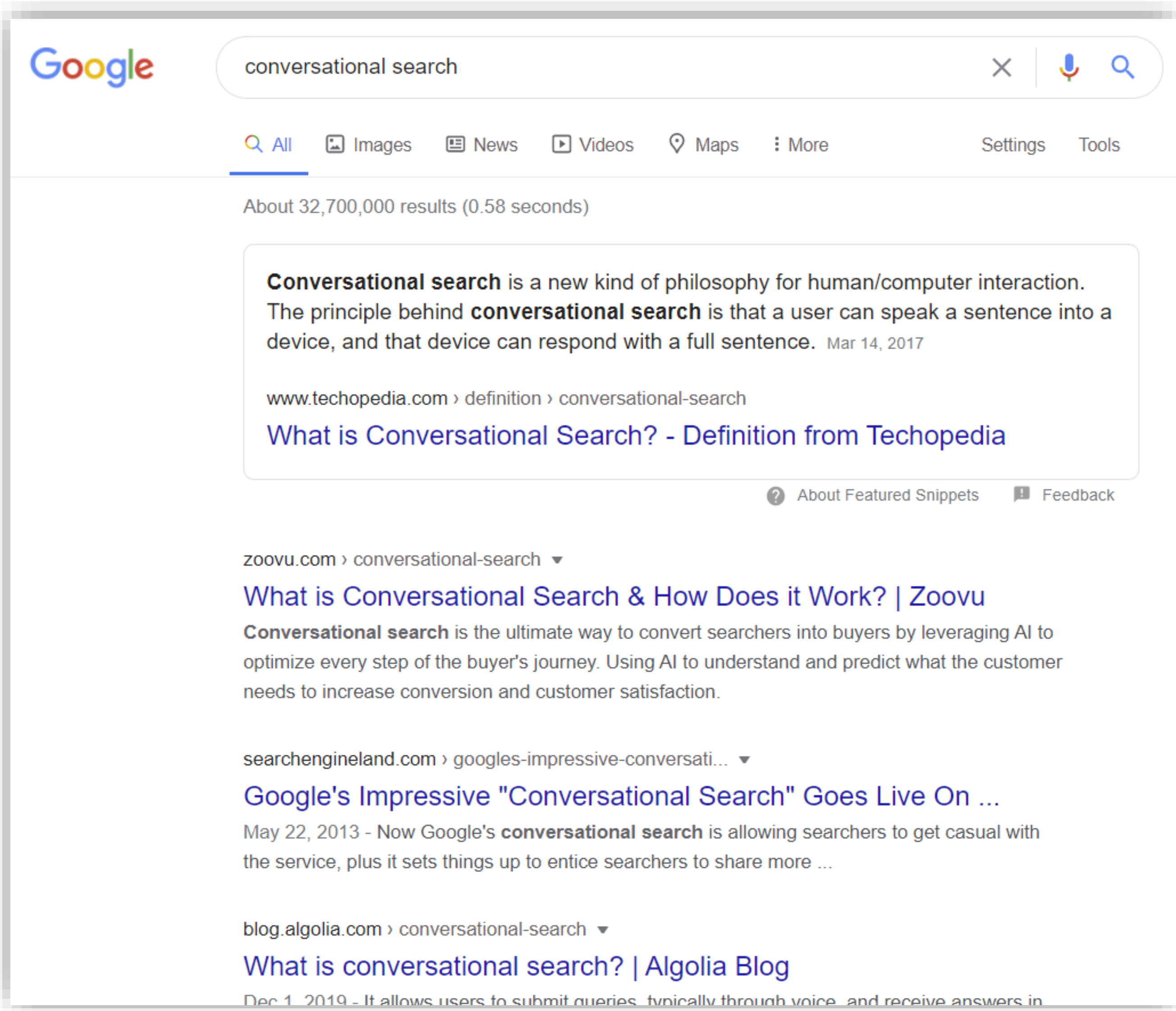
Experimental protocols, datasets, metrics

Information Retrieval

Topic feeds

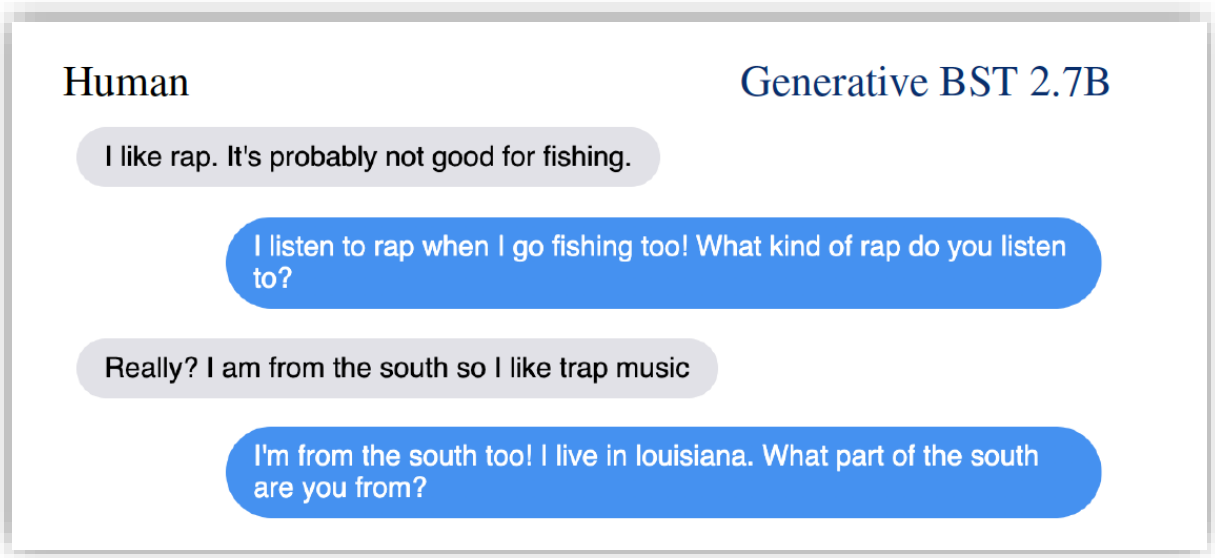

9:15
Home
Major League Soccer @MLS · 30m
EMERSON HYNDMAN IN THE 90TH MINUTE!
Is it the winner for @ATLUTD? #ATLvsSJ
See more about this Topic
You'll see Tweets, people, events, and ads about this, and - just like accounts - the ones you follow appear on your profile.
Atlanta United FC
Follow
Jess
NO Fear the chocolate is here

Search

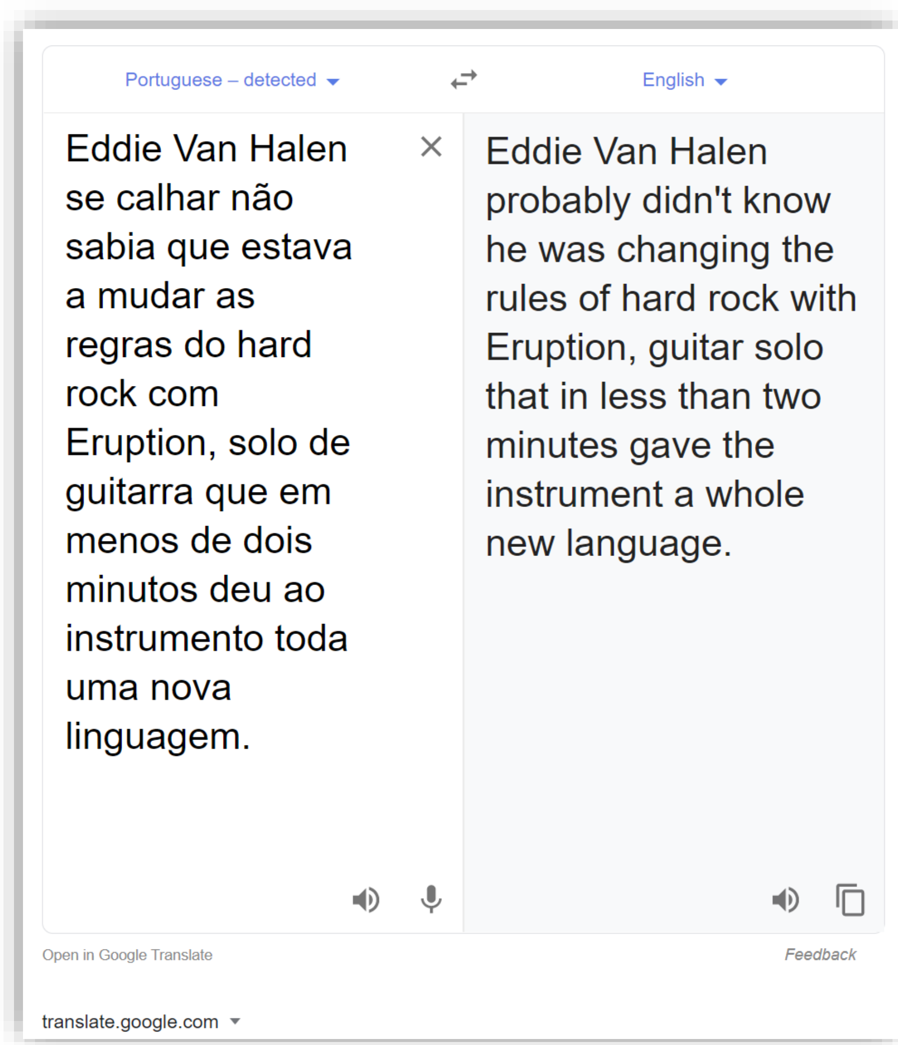


Google
conversational search
All
Images
News
Videos
Maps
More
Settings
Tools
About 32,700,000 results (0.58 seconds)
Conversational search is a new kind of philosophy for human/computer interaction. The principle behind conversational search is that a user can speak a sentence into a device, and that device can respond with a full sentence. Mar 14, 2017
www.techopedia.com › definition › conversational-search
What is Conversational Search? - Definition from Techopedia
About Featured Snippets
Feedback
zoovu.com › conversational-search
What is Conversational Search & How Does it Work? | Zoovu
Conversational search is the ultimate way to convert searchers into buyers by leveraging AI to optimize every step of the buyer's journey. Using AI to understand and predict what the customer needs to increase conversion and customer satisfaction.
searchengineland.com › googles-impressive-conversati...
Google's Impressive "Conversational Search" Goes Live On ...
May 22, 2013 - Now Google's conversational search is allowing searchers to get casual with the service, plus it sets things up to entice searchers to share more ...
blog.algolia.com › conversational-search
What is conversational search? | Algolia Blog

Answer generation



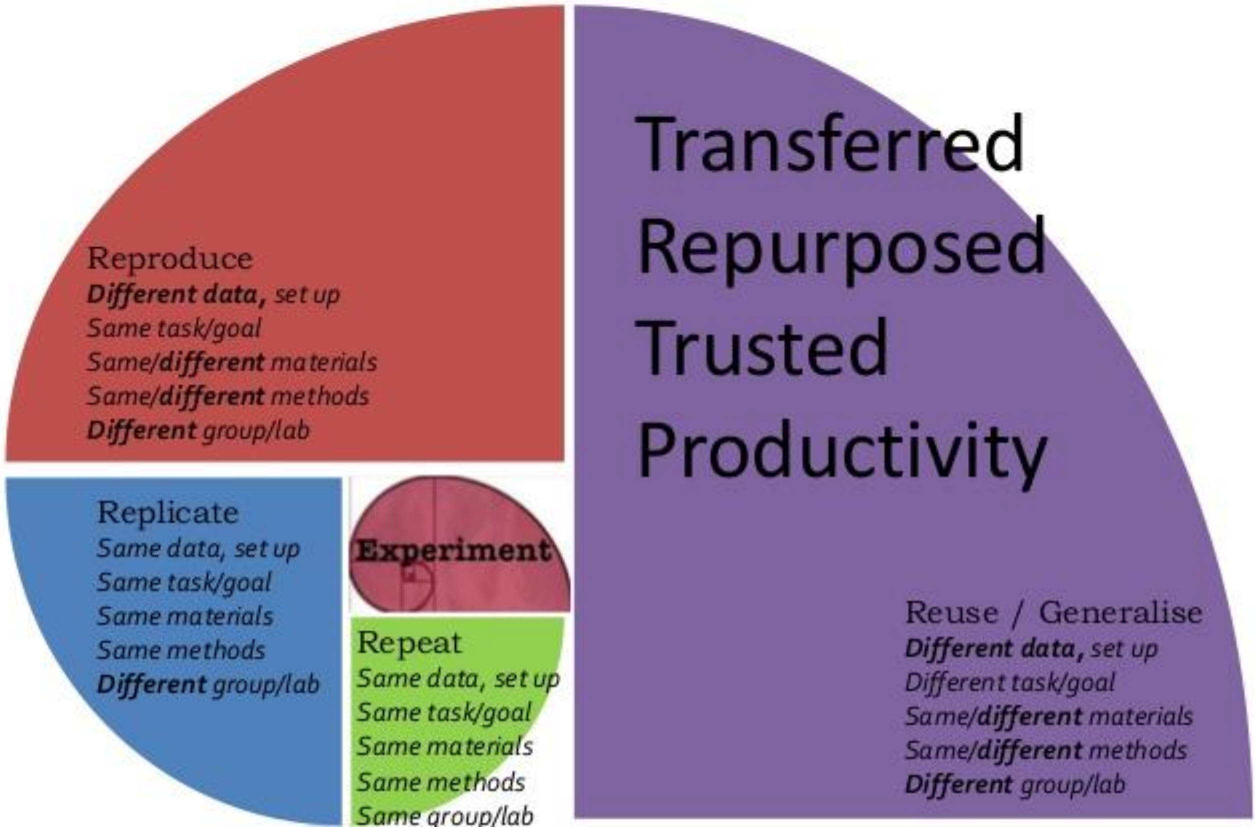
Machine translation



How to benchmark the correctness of natural language processing and information retrieval algorithms?

The R* Nautilus

with thanks to Nicola Ferro for the visualisation



Organizations dedicated to **replicable** and **reproducible** benchmarks

- There are several organizations dedicated to the definition of reproducible benchmakrs:
  - TREC: http://trec.nist.gov/tracks.html
  - CLEF: http://clef2017.clef-initiative.eu/
  - SemEVAL: http://alt.qcri.org/semeval2017/
  - Visual recognition: http://image-net.org/challenges/LSVRC/
- These experimental setups define:
  - a **protocol**
  - a **dataset** (documents and relevance judgments)
  - a set of **metrics** to evaluate performance.

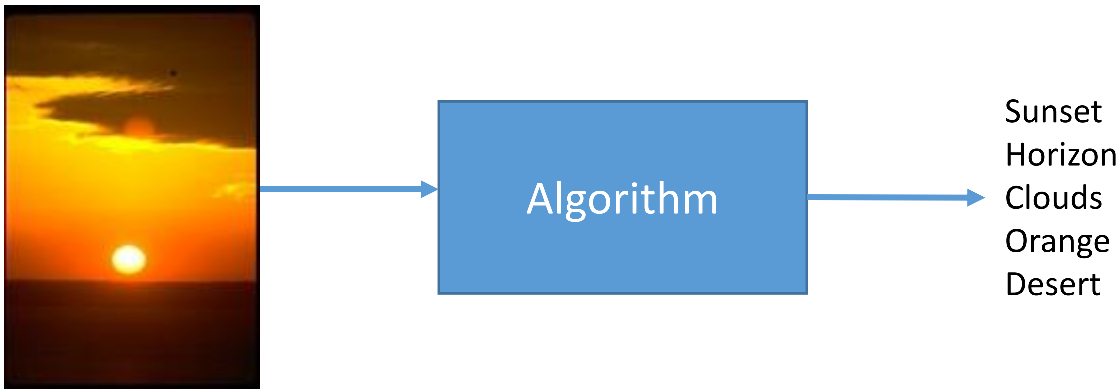
Reproducible experimentation

- Experimental protocol
  - Is the task/problem clear? Is it a standard task?
  - Detailed description of the experimental setup:
    - identify all steps of the experiments.
- Reference dataset
  - Use a well known dataset if possible.
    - If not, how was the data obtained?
  - Clear separation between training and test set.

- Evaluation metrics
  - Prefer the commonly used metrics by the community.
  - Check which statistical test is most adequate.

What is your task?

- Experimental setups are designed to develop a *algorithm* to address a specific task



  - Topic detection; Search by exemple; Ranking annotations; Real-time summarization; Conversational search

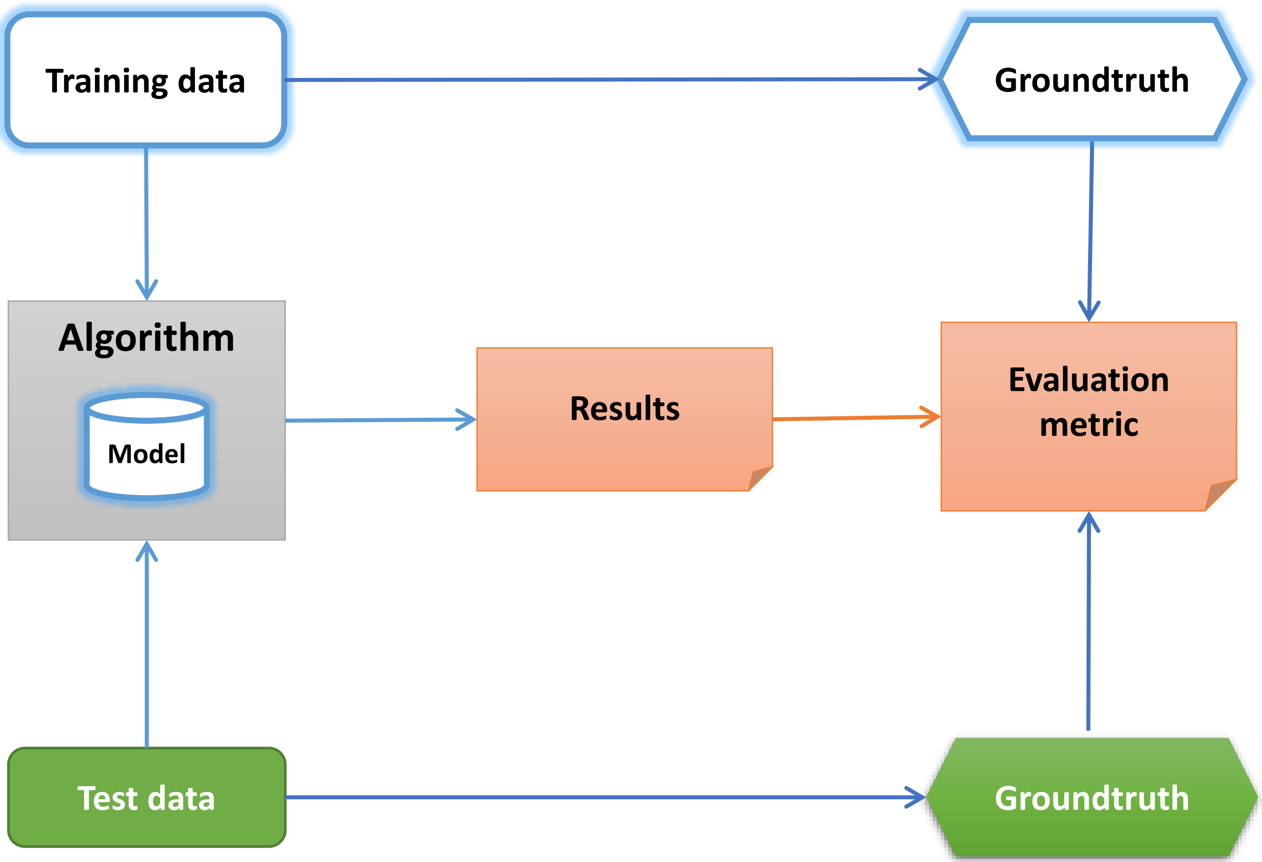
- Benchmarks exist for all the above tasks.

Examples of standard tasks

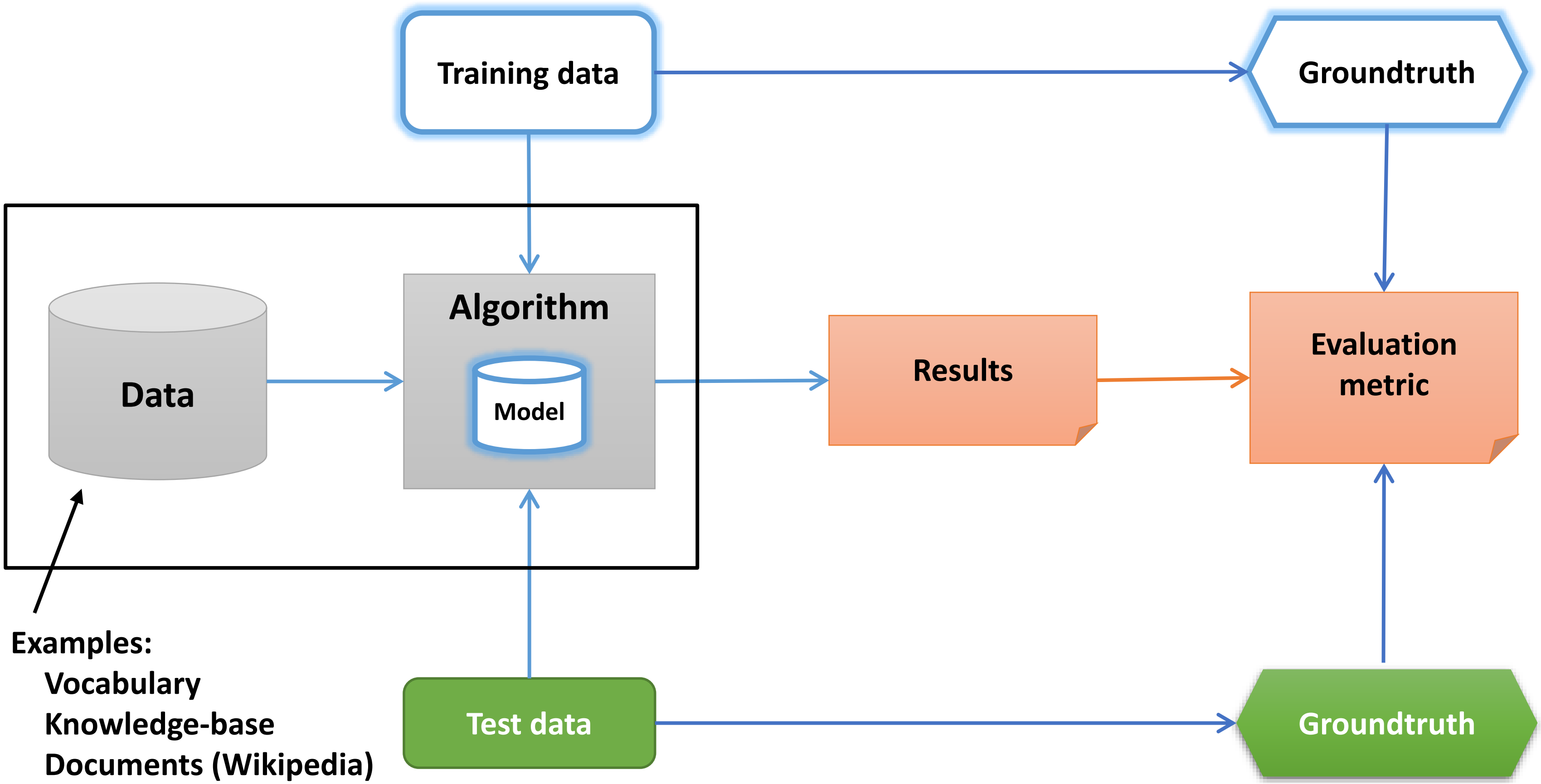
- For example, current “hot” tasks:
  - Conversational recommendation
  - Conversational search: http://www.treccast.ai/
  - Medical Visual QA: https://www.imageclef.org/2019/medical/vqa
  - Health misinformation: https://trec-health-misinfo.github.io/

- Several forums specialize in particular tasks (and region specific challenges):
  - TREC: Blog search, opinion leader, patent search, Web search, document categorization...
  - CLEF: Plagiarism detection, expert search, wikipedia mining, multimodal image tagging, medical image search...
  - Others: Japanese, Russian, Spanish, etc...

Traditional training setup



Extended training setup



Examples:
- **Vocabulary**
- **Knowledge-base**
- **Documents (Wikipedia)**

Essential aspects of a sound evaluation

- Experimental protocol
  - Is the task/problem clear? Is it a standard task?
  - Detailed description of the experimental setup:
    - identify all steps of the experiments.
- Reference dataset
  - Use a well known dataset if possible.
    - If not, how was the data obtained?
  - Clear separation between training and test set.
- Evaluation metrics
  - Prefer the commonly used metrics by the community.
  - Check which statistical test is most adequate.

Reference datasets

- A reference dataset is made of:
  - a collection of documents
  - a set of training data
  - a set of test data
  - the relevance judgments or groundtruth.

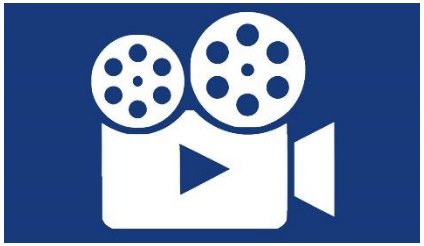
- Reference datasets are as <u>important as metrics</u> for evaluating the proposed method.
  - Many different datasets exist for <u>standard tasks</u>.
  - Reference datasets set the difficulty level of the task.
  - Allow a fair comparison across different methods.

Example of relevance judgments

- Category of a document/image/video

- Query-document pair
  - Q1,D2 TRUE
  - Q1,D7 FALSE
  - Q2,D3 FALSE
  - Q2,D9 TRUE

- Reference translations
  - “Good Morning” -> “Bom dia”



Categories
comedy, cars, explosions

Types of evaluation

- With groundtruth
- A/B testing
- A combination of the two

Ground-truth

- The theoretical "ultimate goal" is to devise a method that produces exactly the same output as the ground-truth.
  - Ground-truth tells the scientist how the algorithm *should* behave.

		Ground-truth		
		True	False	
Method	True	True positive	False positive	Type I error
	False	False negative	True negative	
		Type II error		

- The practical "ultimate goal" can be very different:
  - ground-truth is incomplete, incorrect and only mirrors a small portion of reality.

Obtaining groundtruth/relevance judgments

- Crowdsourcing system
  - DefinedCrowd, Amazon Mechanical Turk, …
  - Limesurvey: https://github.com/LimeSurvey/LimeSurvey
  - Relevation: https://github.com/ielab/relevation
- Quality annotations
  - Redundant annotations
- Cost reduction strategies
  - Convergence
  - Pooling strategies

Annotate these pictures with keywords:




Groundtruth

- Examine the groundtruth:

People
Nepal
Mother
Baby
Colorful dress
Fence

Sunset
Horizon
Clouds
Orange
Desert

Flowers
Yellow
Nature

Beach
Sea
Palm tree
White-sand
Clear sky

- Groundtruth is incomplete
- Not all groundtruth is of equal importance/relevance.

From user annotations to ground-truth

- Judgments can be obtained by **experts** or by **crowdsourcing**
  - Human relevance judgments can be incorrect and inconsistent

- How do we measure the quality of human judgments?

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

$p(A)$ -> proportion of times humans agreed

$p(E)$ -> probability of agreeing by chance

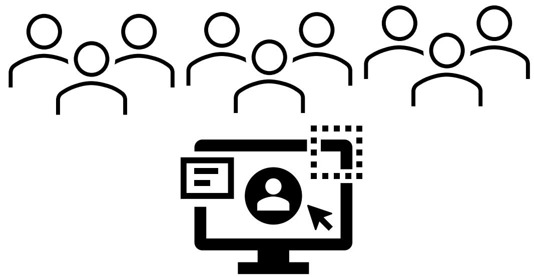
  - Values above 0.8 are considered good
  - Values between 0.67 and 0.8 are considered fair
  - Values below 0.67 are considered dubious

System quality and user utility

- The discussed evaluation procedures only measure the system performance on a given task
    - It can overfit
    - It might be distant from what users expect
- Only real users actually assess the system
    - How expressive is its query language?
    - How large is its collection?
    - How effective are the results?
- A/B testing
    - Make small variation on the system and direct a proportion of users to that system
    - Evaluate frequency with which users prefer one system to the other

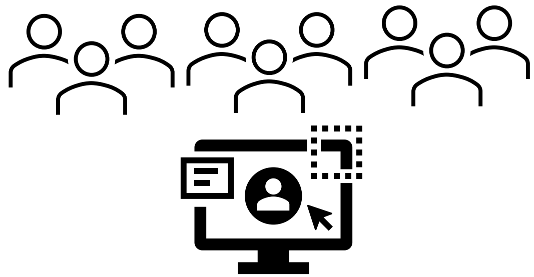
A/B testing

Full
population

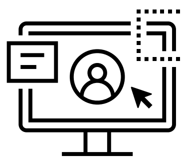


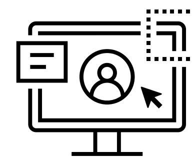
A/B testing

Full population



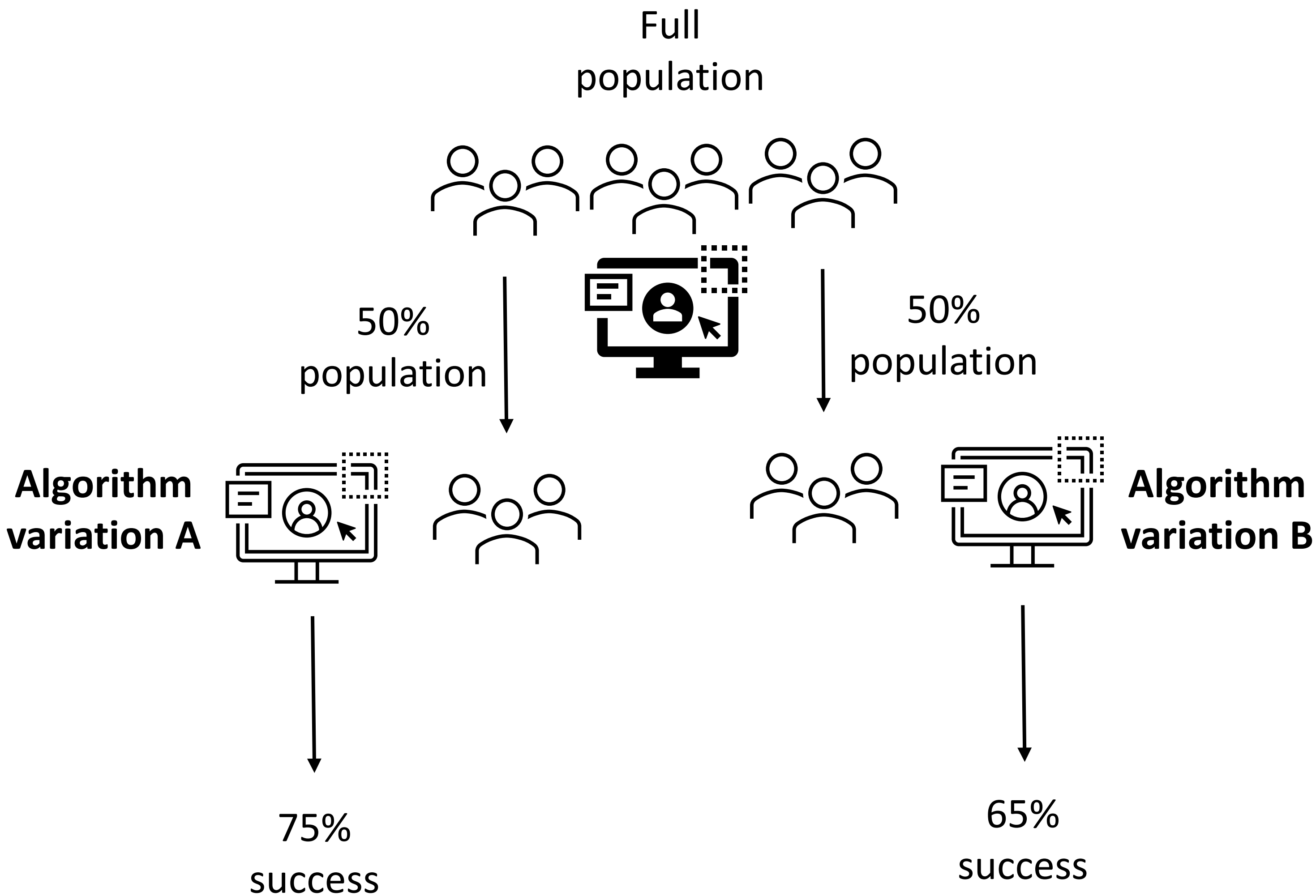
Algorithm variation A





Algorithm variation B

A/B testing


Full
population
50%
population
50%
population
Algorithm
variation A
Algorithm
variation B
75%
success
65%
success

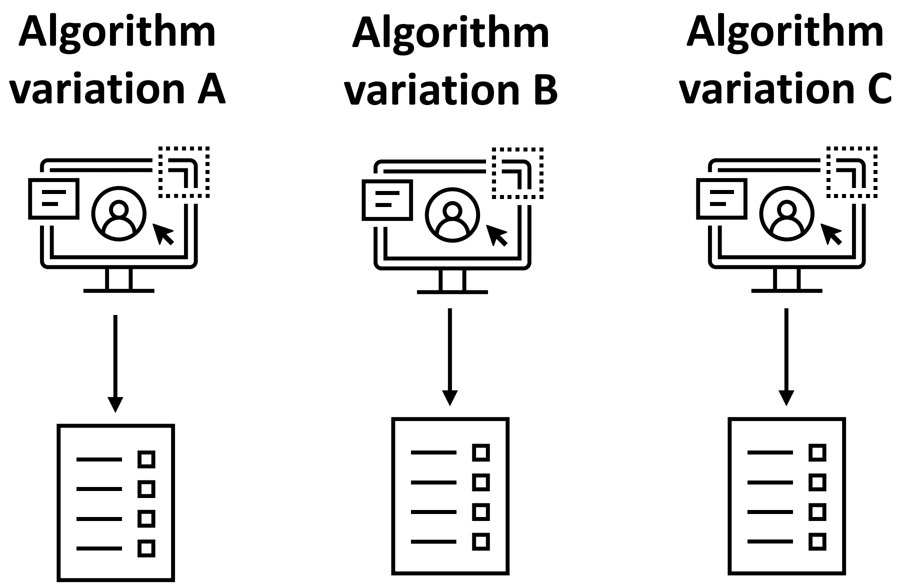
Results pooling

- This technique is used when the dataset is too large to be completely examined.

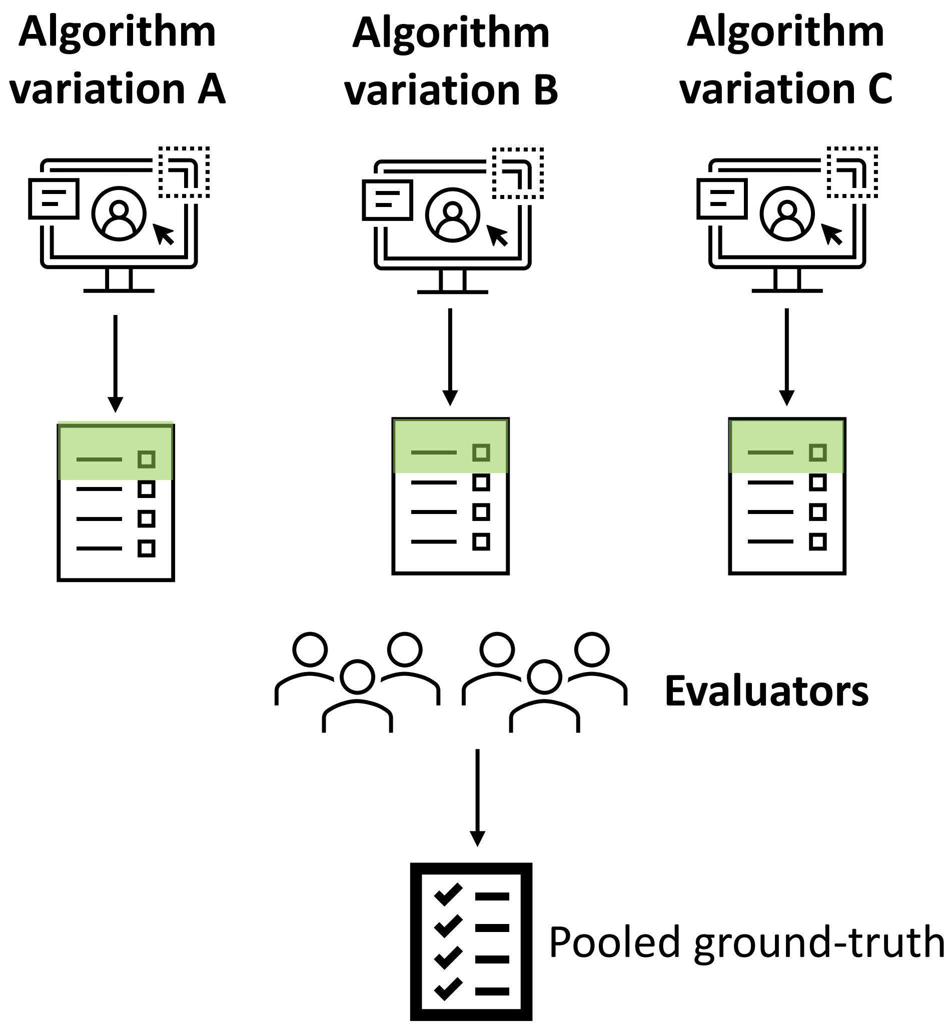
- Considering the results of 10 systems:
  - Examine the top 100 results of each system
  - Label all documents according to its relevance
  - Use the labeled results as ground-truth to evaluate all systems.

- <u>**Drawback: can’t compute recall, AP and MAP**</u>

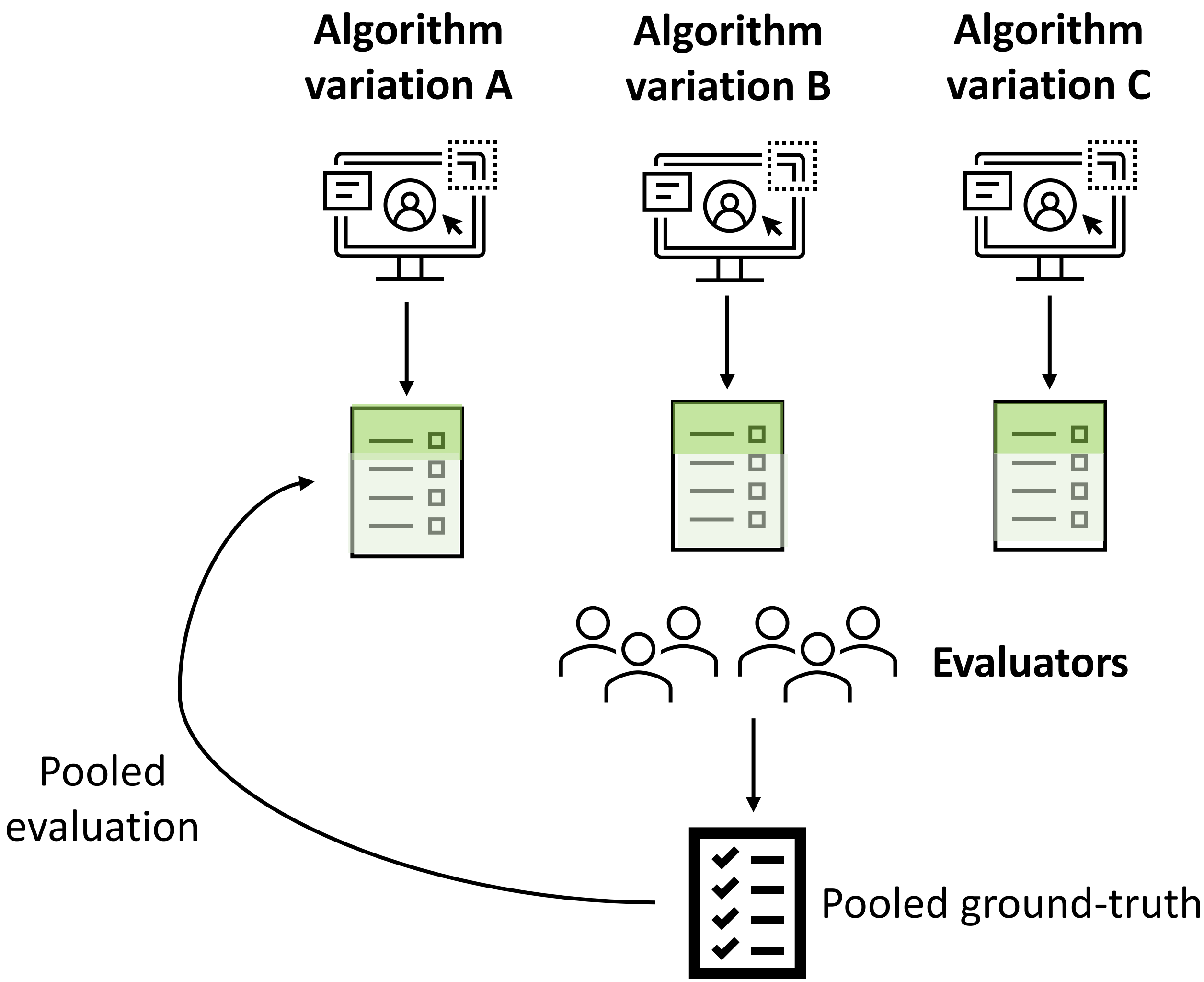
Results pooling



Results pooling



Results pooling



Essential aspects of a sound evaluation

- Experimental protocol
  - Is the task/problem clear? Is it a standard task?
  - Detailed description of the experimental setup:
    - identify all steps of the experiments.

- Reference dataset
  - Use a well known dataset if possible.
    - If not, how was the data obtained?
  - Clear separation between training and test set.

- Evaluation metrics
  - Prefer the commonly used metrics by the community.
  - Check which statistical test is most adequate.

Evaluation metrics

- <u>Utility metrics</u> are focused in evaluating the results that are presented to the user
  - Usually, this is done with relevance judgments on the top results
  - Common metrics for binary relevance judgments: Top Precision and Recall
  - Common metrics for binary relevance judgments : NDCG

- Stability metrics are focused in evaluating the robustness of the system results.
  - Usually, this is done with binary relevance judgments across a wide range of data
  - Common metrics: MAP, AP, Precision-Recall curves

Binary relevance judgments

$$Accuracy = \frac{truePos + trueNeg}{truePos + falsePos + trueNeg + falseNeg}$$

$$Precision = \frac{truePos}{truePos + falsePos}$$

$$Recall = \frac{truePos}{truePos + falseNeg}$$

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

		Ground-truth	
		True	False
Method	True	True positive	False positive
	False	False negative	True negative

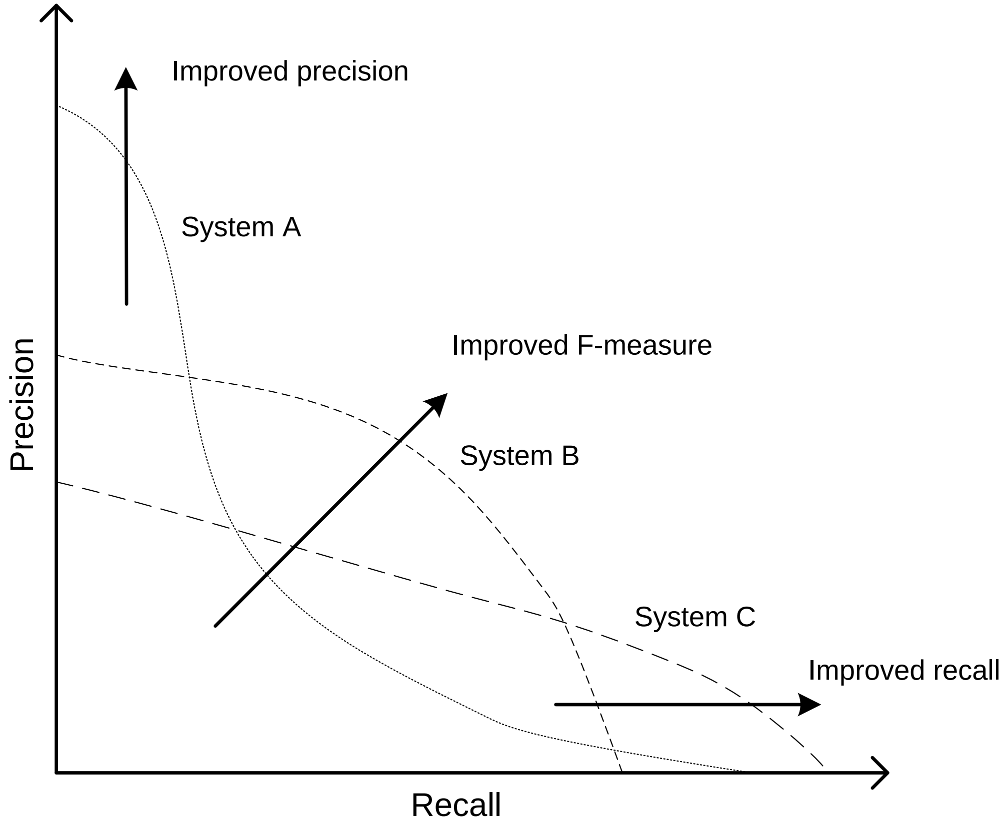
Em PT: exatidão, precisão e abragência.

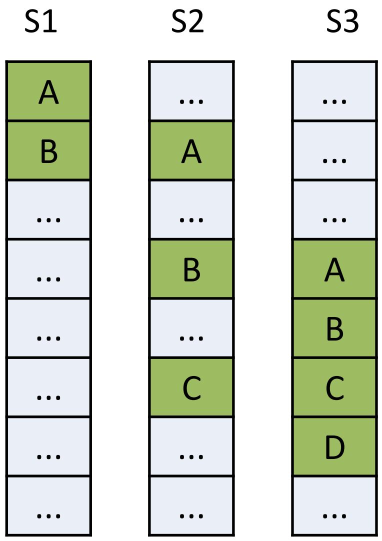
Why not accuracy?

You easily get 99.999999% by not retrieving non-relevant results!!!

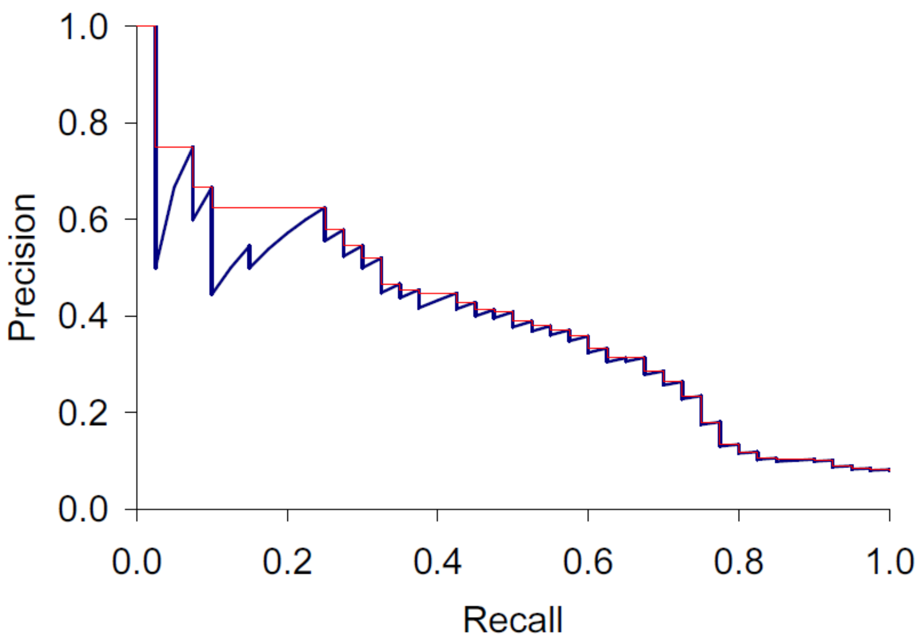
$$Accuracy = \frac{truePos + trueNeg}{truePos + falsePos + trueNeg + falseNeg}$$

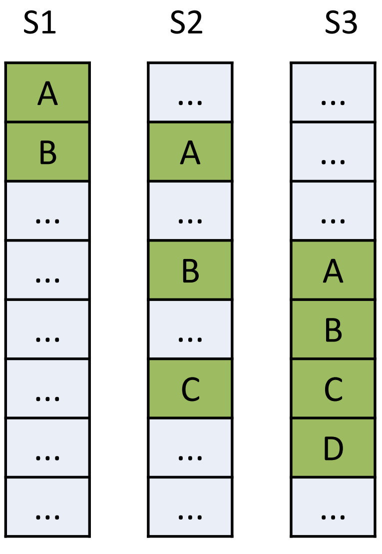
Precision-recall graphs for ranked results





Interpolated precision-recall graphs





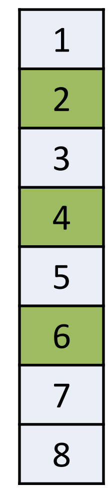
Average Precision

- Web systems favor high-precision methods (P@20)

- Other more robust metric is AP:

$$AP = \frac{1}{\#relevant} \cdot \sum_{k \in \{set\ of\ positions\ of\ the\ relevant\ docs\}} p@k$$

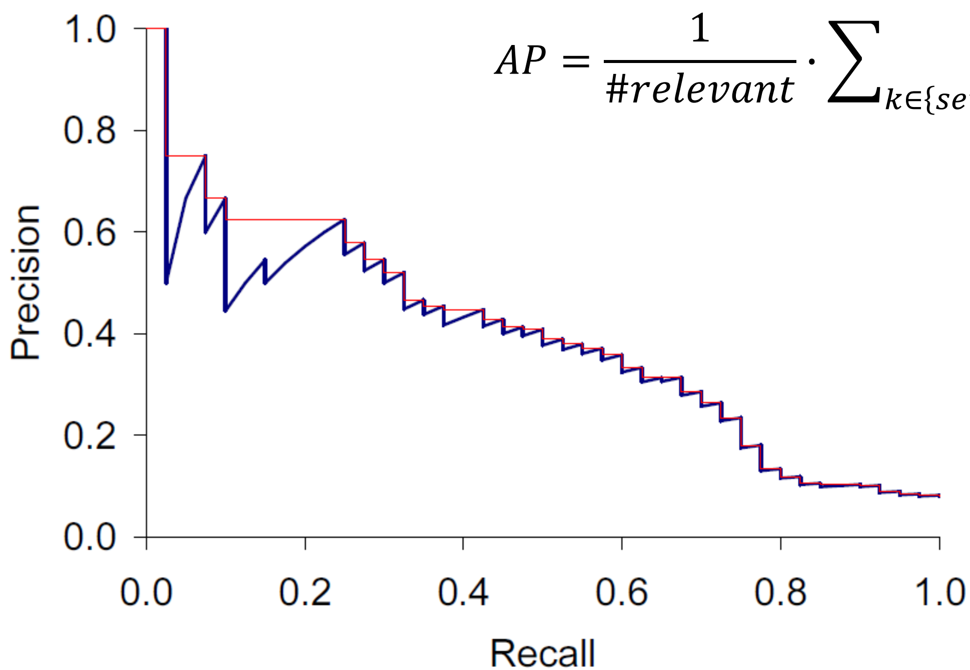
$$AP = \frac{1}{4} \cdot \left(\frac{1}{2} + \frac{2}{4} + \frac{3}{6}\right) = 0.375$$



Average Precision

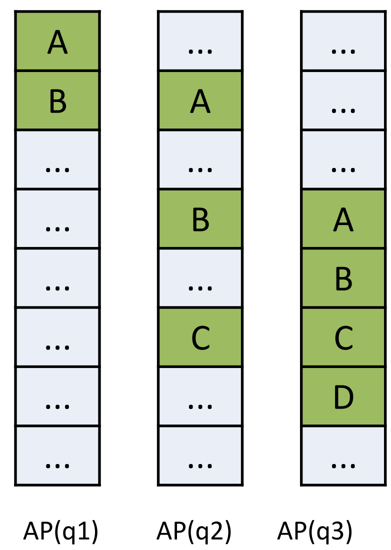
- Average precision is the area under the P-R curve

$$AP = \frac{1}{\#relevant} \cdot \sum_{k \in \{set\ of\ positions\ of\ the\ relevant\ docs\}} p@k$$



Mean Average Precision (MAP)

- MAP evaluates the system for a given range of queries.
- It summarizes the global system performance in one single value.
- It is the mean of the average precision of a set of n queries:



$$MAP = \frac{AP(q_1) + AP(q_2) + AP(q_3) + \ldots + AP(q_n)}{n}$$

Web scale evaluation

- It is impossible to know all relevant documents.
  - It is too expensive or time-consuming.

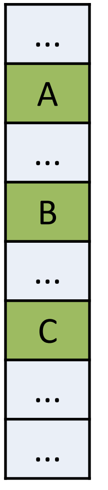
- **nDCG**, **BPref** and **Inferred AP** are three measures to evaluate a system with incomplete ground-truth.

- These metrics use the concept of **pooled results**

E. Yilmaz and J. A. Aslam, Estimating average precision with incomplete and imperfect judgments, ACM CIKM *2006.*
C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. ACM SIGIR 2004.

Relevance

- Some documents are more relevant than others.
    - Documents have different levels of relevance.
- The position of a document in the rank is also important to the user.
    - Relevant documents ranked top count more.



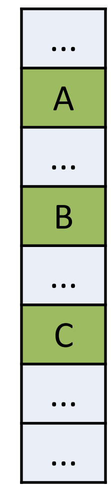
DCG: Incomplete multi-level relevance

- The Discounted Cumulative Gain measure, considers the notion of multi-level relevance:

$$DCG_m \propto 2^{rel_i} - 1 \qquad rel_i = \{0,1,2,3,\dots\}$$

- The DCG measure, also considers the position where the document is on the rank:

$$DCG_m = \sum_{i=1}^{m} \frac{2^{rel_i} - 1}{\log_2(1+i)} \qquad rel_i = \{0,1,2,3,\dots\}$$



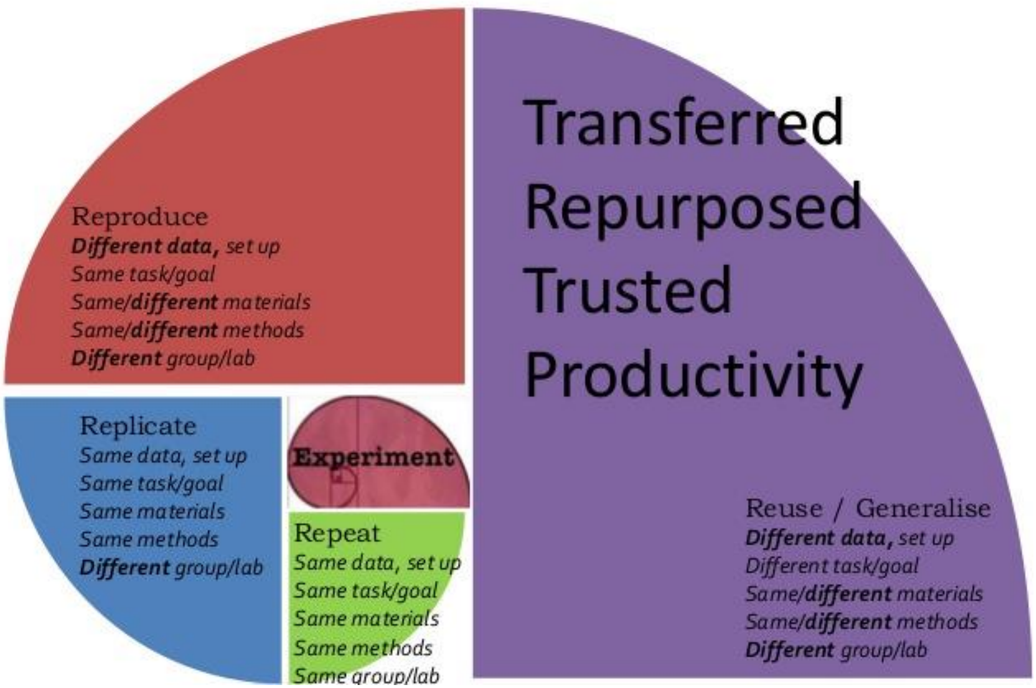
- The normalized metric measures the deviation from the optimal sort order:

$$nDCG_m = \frac{DCG_m}{bestDCG_m}$$

K. Jarvelin, J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," ACM Transactions on Information Systems 20(4), 422–446 (2002).

The R* Nautilus

with thanks to Nicola Ferro for the visualisation



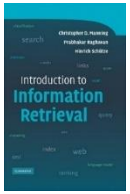
Experimental protocol

Reference dataset

Evaluation metrics

Summary

- Experimental benchmarks: Protocol, Dataset, Metrics



Chapter 8

- The quality of groundtruth
  - Incomplete, incorrect, ambiguous
  - Kappa statistics

- Measuring success
  - Metrics (for replicable+repeatable experiments)
  - A/B testing is the best way of measuring success, but is by far the most expensive
  - Results pooling is a balanced strategy

- Evaluation collections / resources
  - See TRECVID and ImageCLEF for multimedia datasets.
  - See TREC and CLEF forums for large-scale datasets
  - Others exist for Biomedical domain, Geographic IR, Plagiarism,…