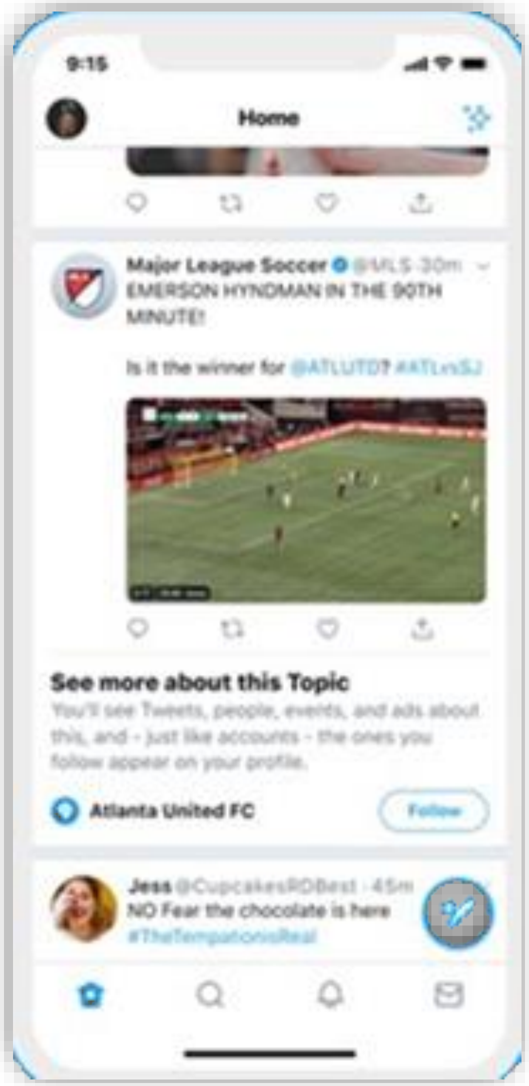
Evaluation

Experimental protocols, datasets, metrics

Information Retrieval

Topic feeds



Major League Soccer @MLS · 30m
EMERSON HYNDMAN IN THE 90TH MINUTE!

Is it the winner for @ATLUTD? #ATLvsSJ

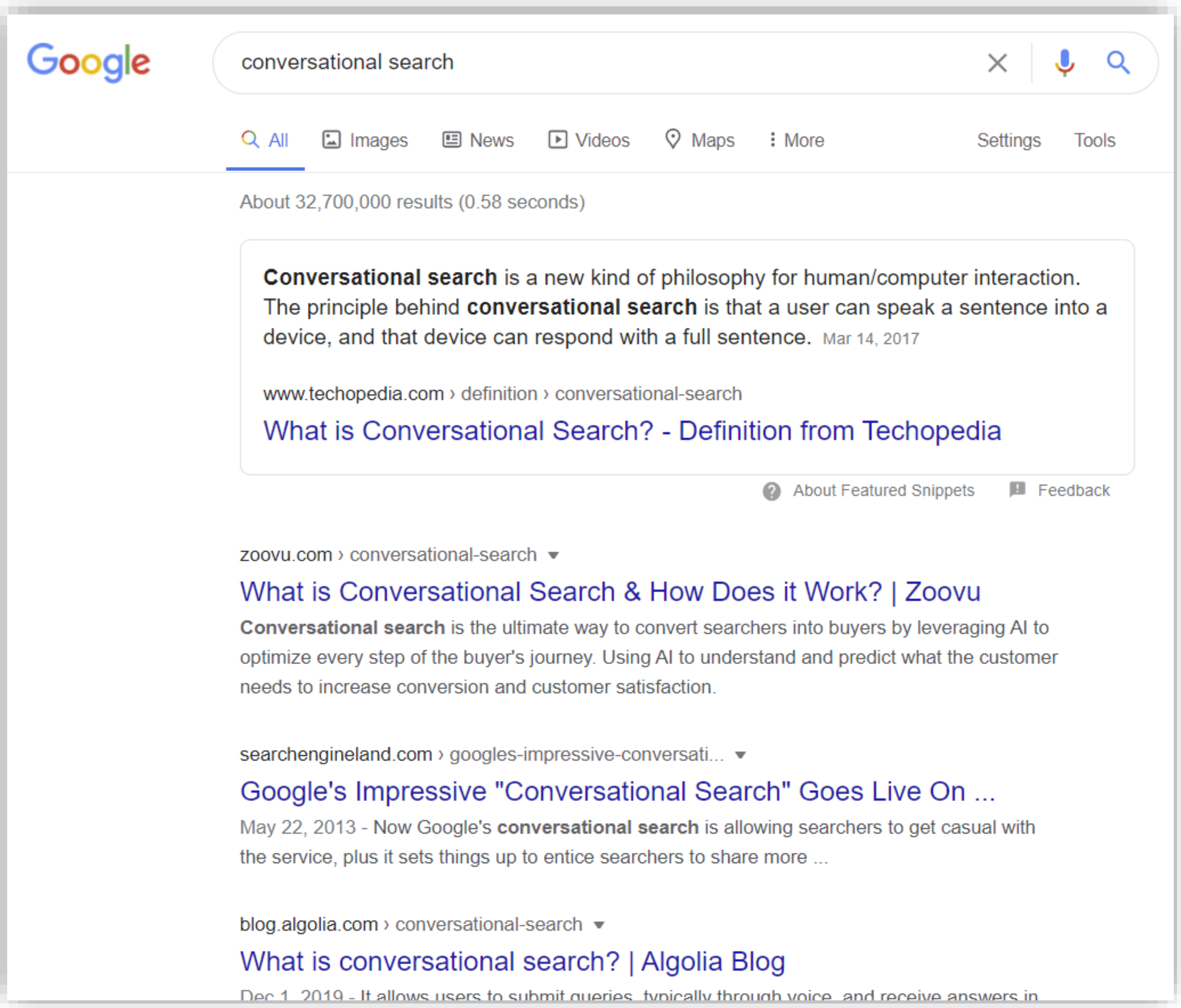
See more about this Topic

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

Jess @CupcakesRDBest · 45m
NO Fear the chocolate is here
#TheTemptationisReal

Search



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

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 ...

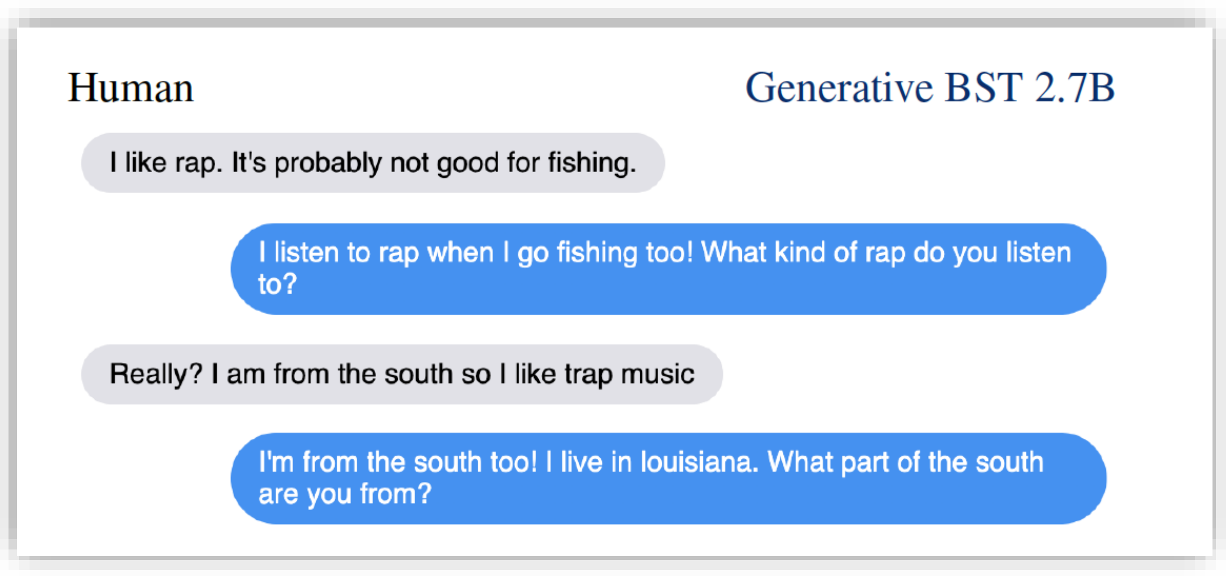
May 22, 2013 - Now Google's **conversational search** is allowing searchers to get casual with the service, plus it sets things up to entice searchers to share more ...

blog.algolia.com › conversational-search ▾

What is conversational search? | Algolia Blog

Dec 1, 2019 - It allows users to submit queries, typically through voice, and receive answers in

Answer generation



Human

Generative BST 2.7B

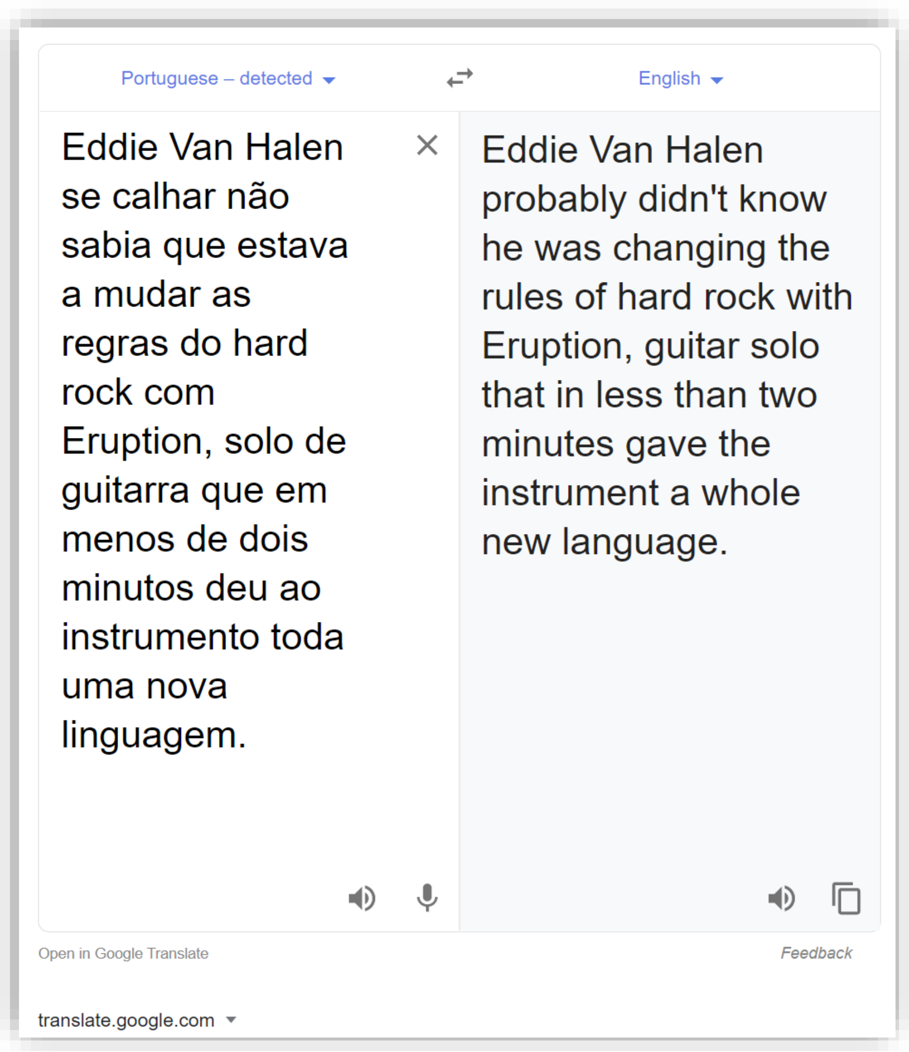
I like rap. It's probably not good for fishing.

I listen to rap when I go fishing too! What kind of rap do you listen to?

Really? I am from the south so I like trap music

I'm from the south too! I live in louisiana. What part of the south are you from?

Machine translation



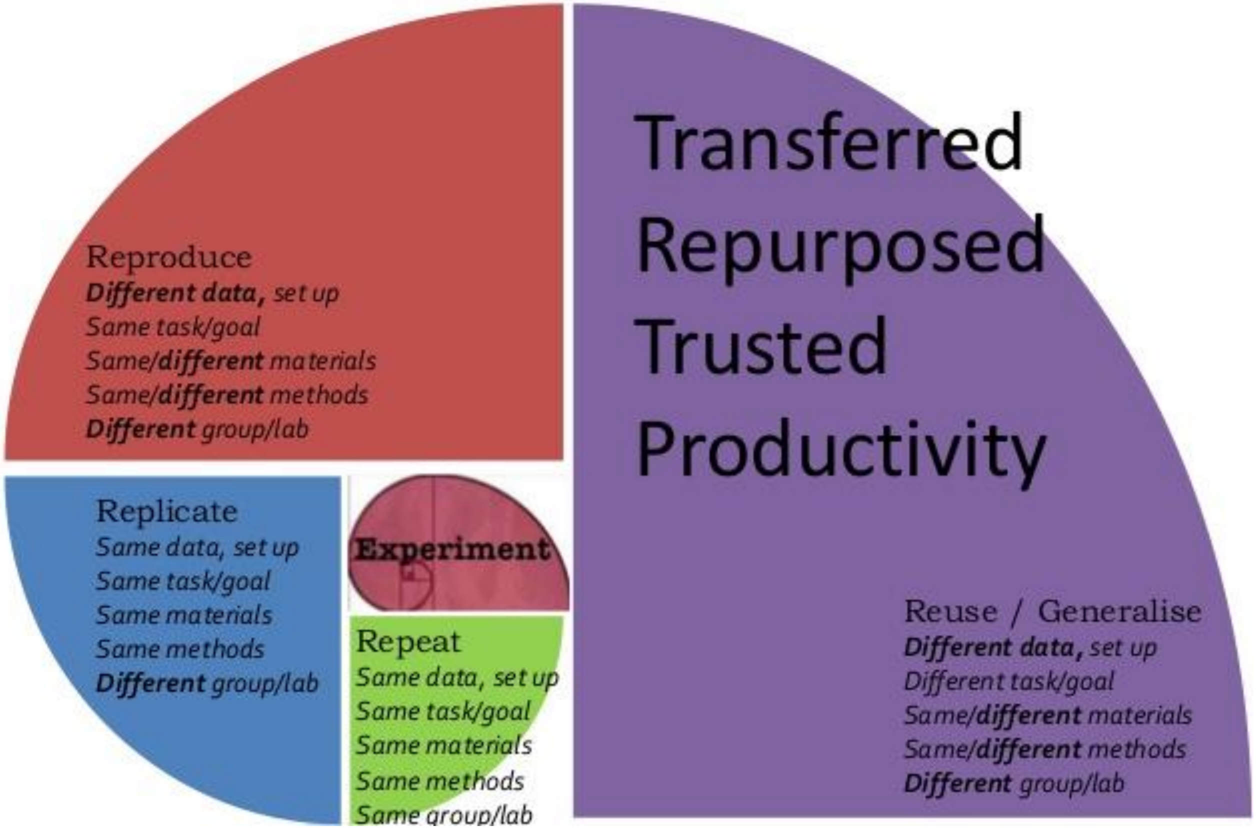
Eddie Van Halen se calhar não sabia que estava a mudar as regras do hard rock com Eruption, solo de guitarra que em menos de dois minutos deu ao instrumento toda uma nova linguagem.

Eddie Van Halen probably didn't know he was changing the rules of hard rock with Eruption, guitar solo that in less than two minutes gave the instrument a whole new language.

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

The R* Nautilus

with thanks to Nicola Ferro for the visualisation



Reproduce
Different data, *set up*
Same task/goal
*Same/****different*** *materials*
*Same/****different*** *methods*
Different *group/lab*

Transferred
Repurposed
Trusted
Productivity

Replicate
Same data, set up
Same task/goal
Same materials
Same methods
Different *group/lab*

Experiment

Repeat
Same data, set up
Same task/goal
Same materials
Same methods
Same group/lab

Reuse / Generalise
Different data, *set up*
Different task/goal
*Same/****different*** *materials*
*Same/****different*** *methods*
Different *group/lab*

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.

Experimental setups

- There are experimental setups made available by different organizations:
  - 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) and suggest a set of metrics to evaluate performance.

What is a standard task?

- Experimental setups are designed to develop a *language processing algorithm* to address a specific task.
  - Topic detection
  - Search by example
  - Ranking annotations
  - Real-time summarization
  - Conversational search

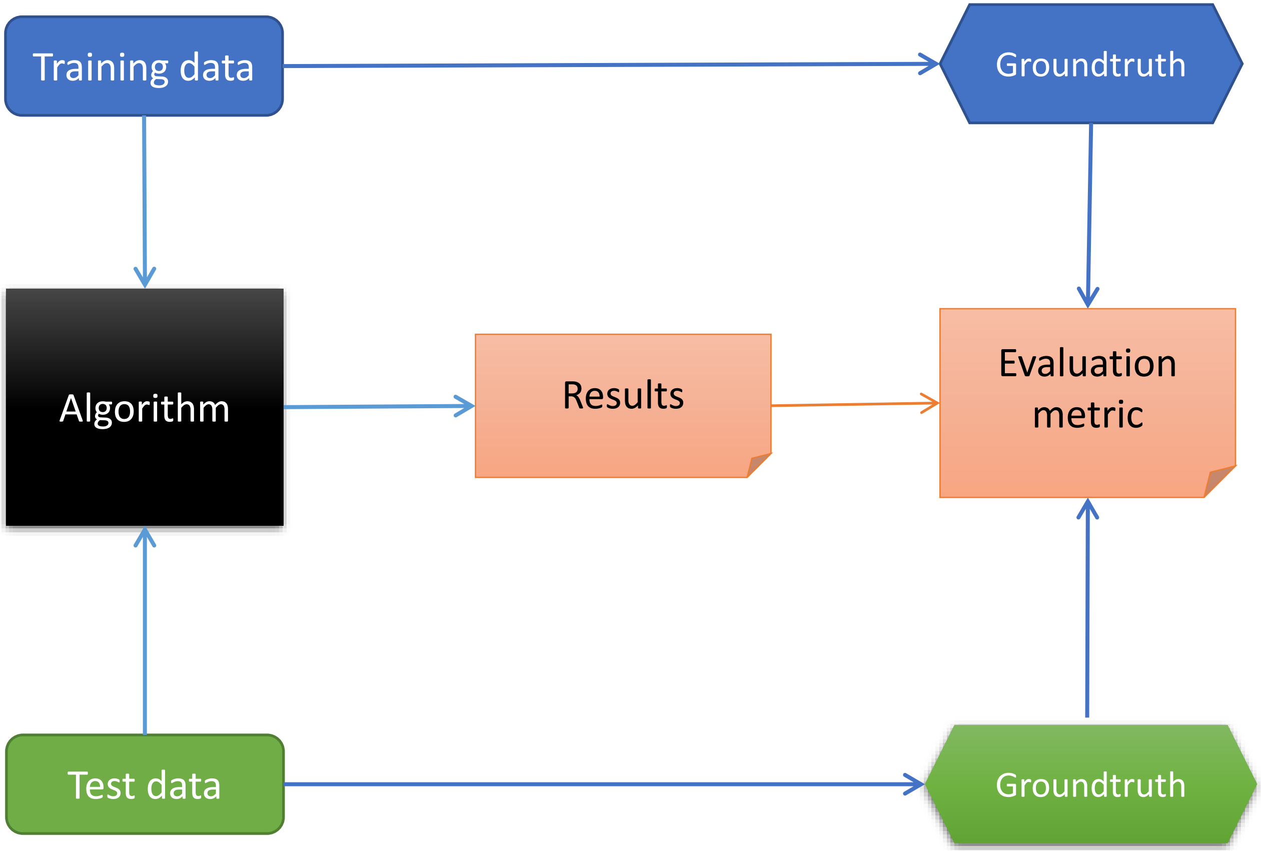
- Datasets exist for all the above tasks.

Sunset
Horizon
Coulds
Orange
Desert

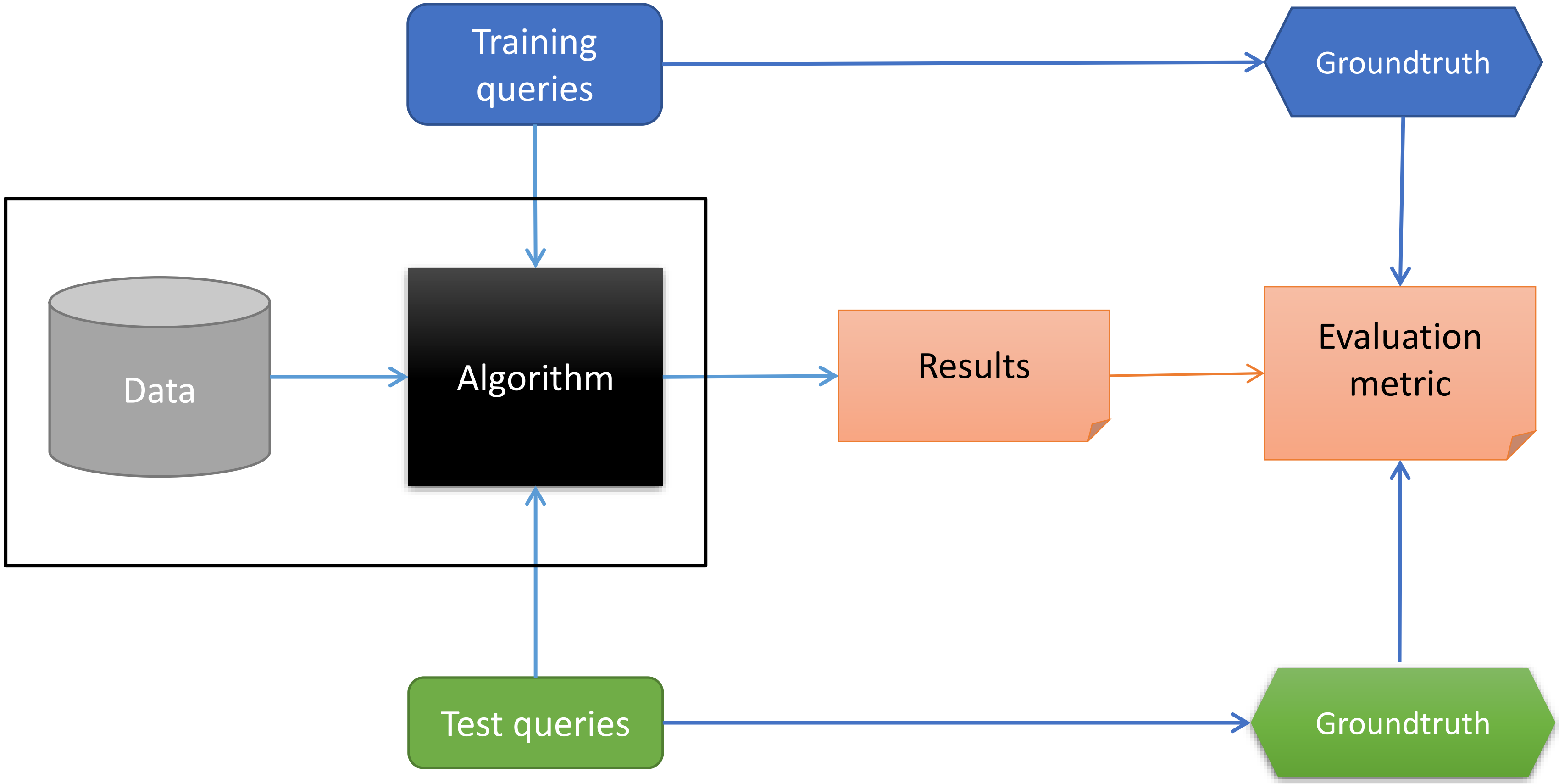
Examples of standard tasks

- 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 exist with different tasks:
  - 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...

A classification evaluation protocol



A retrieval evaluation protocol



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 queries
    - a set of test queries
    - the relevance judgments of the pairs query-document.

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

Ground-truth

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

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

Type I error

Type II error

Annotate these pictures with keywords:

Groundtruth

People
Nepal
Mother
Baby
Colorful dress
Fence

Sunset
Horizon
Coulds
Orange
Desert

Flowers
Yellow
Nature

Beach
Sea
Palm tree
White-sand
Clear sky

Groundtruth can be incomplete, not all groundtruth is of equal importance/relevance.

Relevance judgments -> Groundtruth

- 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

Example of relevance judgments

- Category of a document/image/video
- Query-document pair
- Reference translations

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

- Complete relevance judgments
  - Ranked relevance judgments
  - Binary relevance judgments

- Incomplete relevance judgments (Web scale eval.)
  - Binary relevance judgments
  - Multi-level relevance judgments

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

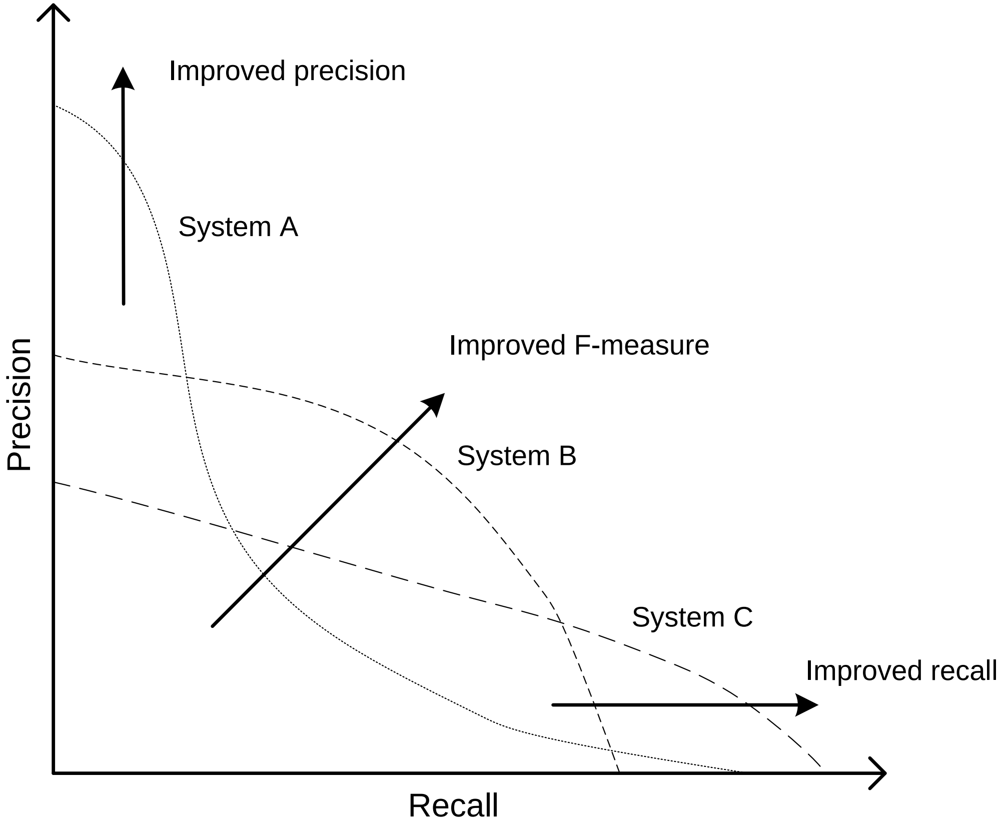
Em PT: exatidão, precisão e abragência.

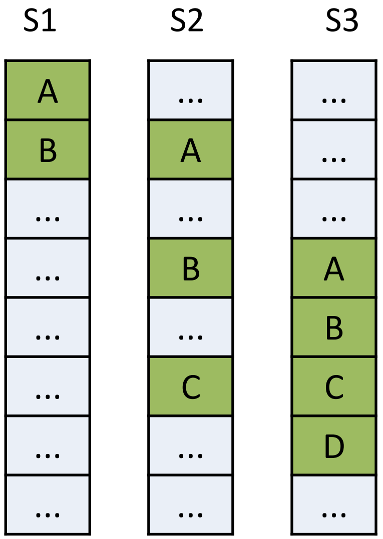
Why not accuracy?

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

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

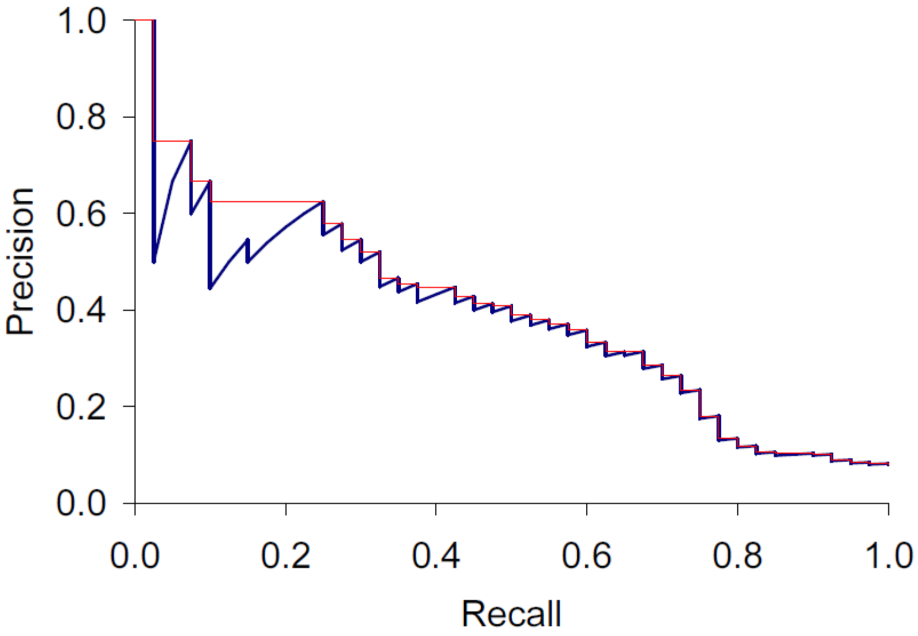
Precision-recall graphs for ranked results





S1	S2	S3
A	…	…
B	A	…
…	…	…
…	B	A
…	…	B
…	C	C
…	…	D
…	…	…

Interpolated precision-recall graphs



S1	S2	S3
A	…	…
B	A	…
…	…	…
…	B	A
…	…	B
…	C	C
…	…	D
…	…	…

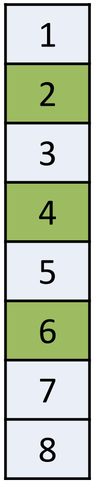
Average Precision

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

- Other more robust metric is AP:

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

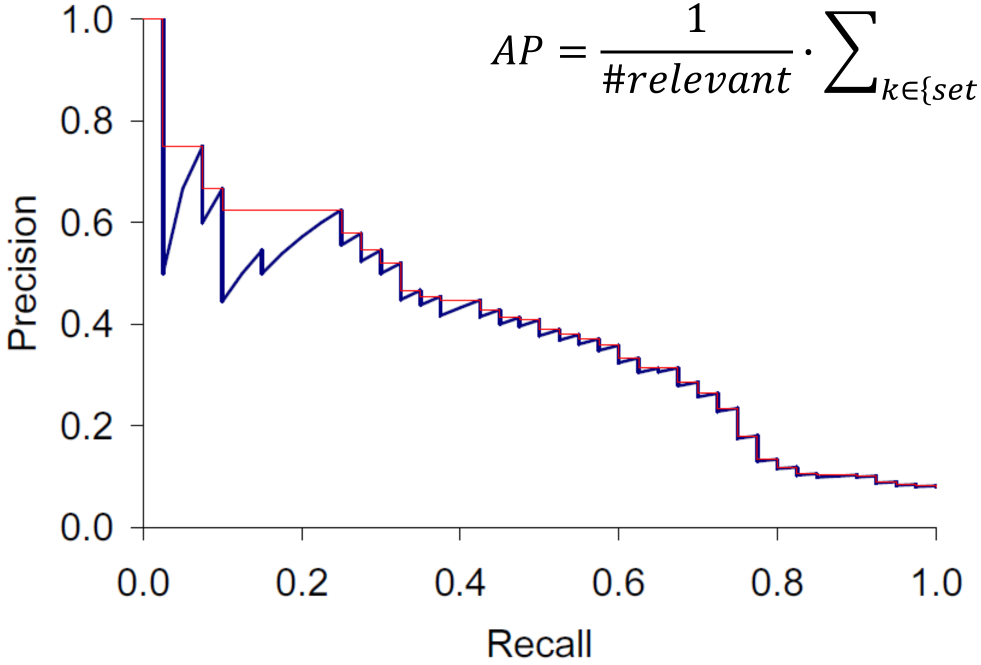
$$AP = \frac{1}{4} \cdot \left(\frac{1}{2} + \frac{2}{4} + \frac{3}{6}\right) = 0.375$$



Average Precision

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

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

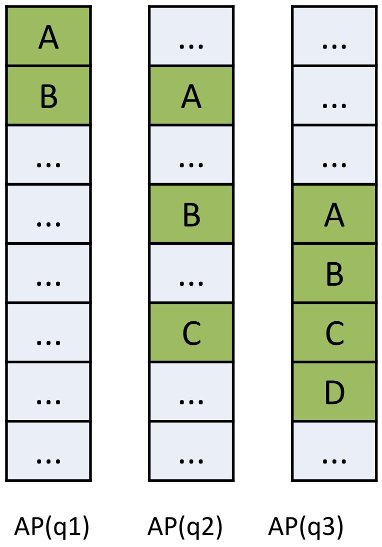


Mean Average Precision (MAP)

- 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.

Results pooling

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

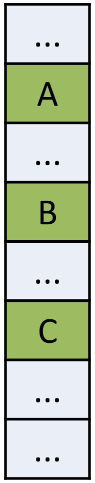
- Considering the results of 10 systems:
  - Examine the top 100 results of each system
  - Label all documents according to its relevance
  - Use the labeled results as ground-truth to evaluate all systems.

- **Drawback: can’t compute recall, AP and MAP**

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.

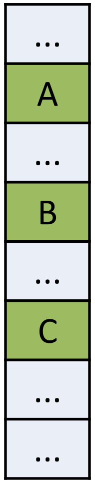


…
A
…
B
…
C
…
…

DCG: Incomplete multi-level relevance

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

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



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

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

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

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

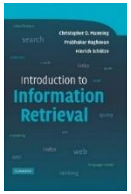
K. Jarvelin, J. Kekalainen, “Cumulated gain-based evaluation of IR techniques,” ACM Transactions on Information Systems 20(4), 422–446 (2002).

Efficiency metrics

Metric name	Description
Elapsed indexing time	Measures the amount of time necessary to build a document index on a particular system.
Indexing processor time	Measures the CPU seconds used in building a document index. This is similar to elapsed time, but does not count time waiting for I/O or speed gains from parallelism.
Query throughput	Number of queries processed per second.
Query latency	The amount of time a user must wait after issuing a query before receiving a response, measured in milliseconds. This can be measured using the mean, but is often more instructive when used with the median or a percentile bound.
Indexing temporary space	Amount of temporary disk space used while creating an index.
Index size	Amount of storage necessary to store the index files.

Summary

- Metrics for complete relevance judgments
  - Binary: Precision, Recall, F-measure, Average Precision, Mean AP
  - Ranked: Spearman, Kendal-tau



Chapter 8

- Metrics for incomplete relevance judgments
  - Binary: Bpref, InfMAP
  - Multi-valued: Normalized DCG

- Evaluation collections / resources
  - See TRECVID and ImageCLEF for multimedia datasets.
  - See TREC and CLEF forums for Web and large-scale datasets
    - User search interaction, Geographic IR, Expert finding, Blog search, Plagiarism,…
  - Use **trec_eval** application to evaluate your system