Evaluation and Replication (the scientist)

Trusted Data Analytics Seminar

TU Delft, 15 June 2018

Julio Gonzalo (UNED, Madrid, Spain)

Evaluation & Replication

- **Reproducibility**: experimental results can be replicated
- **Generalizability**: Results make valid predictions outside the lab
- **Validity**: Results are meaningful, unbiased and based on valid measurements

Data Scientists Anonymous

(Textual) Data Science Evaluation

Prediction is very difficult, especially about the future
(Niels Bohr, physicist)
Replication is very difficult, especially after the first occurrence
(Stefano Mizzaro, data scientist)

- Hard to predict performance
  - for a new problem
  - for a similar problem
  - for the same problem on a different dataset
- Even hard to replicate results

The bridge metaphor in Computer Science

Satellite control systems

NLP

IR

RecSys

New Cola Flavor

objective

subjective

human data

Standard Evaluation Methodology

Task formalization, RQs (← Algorithms)

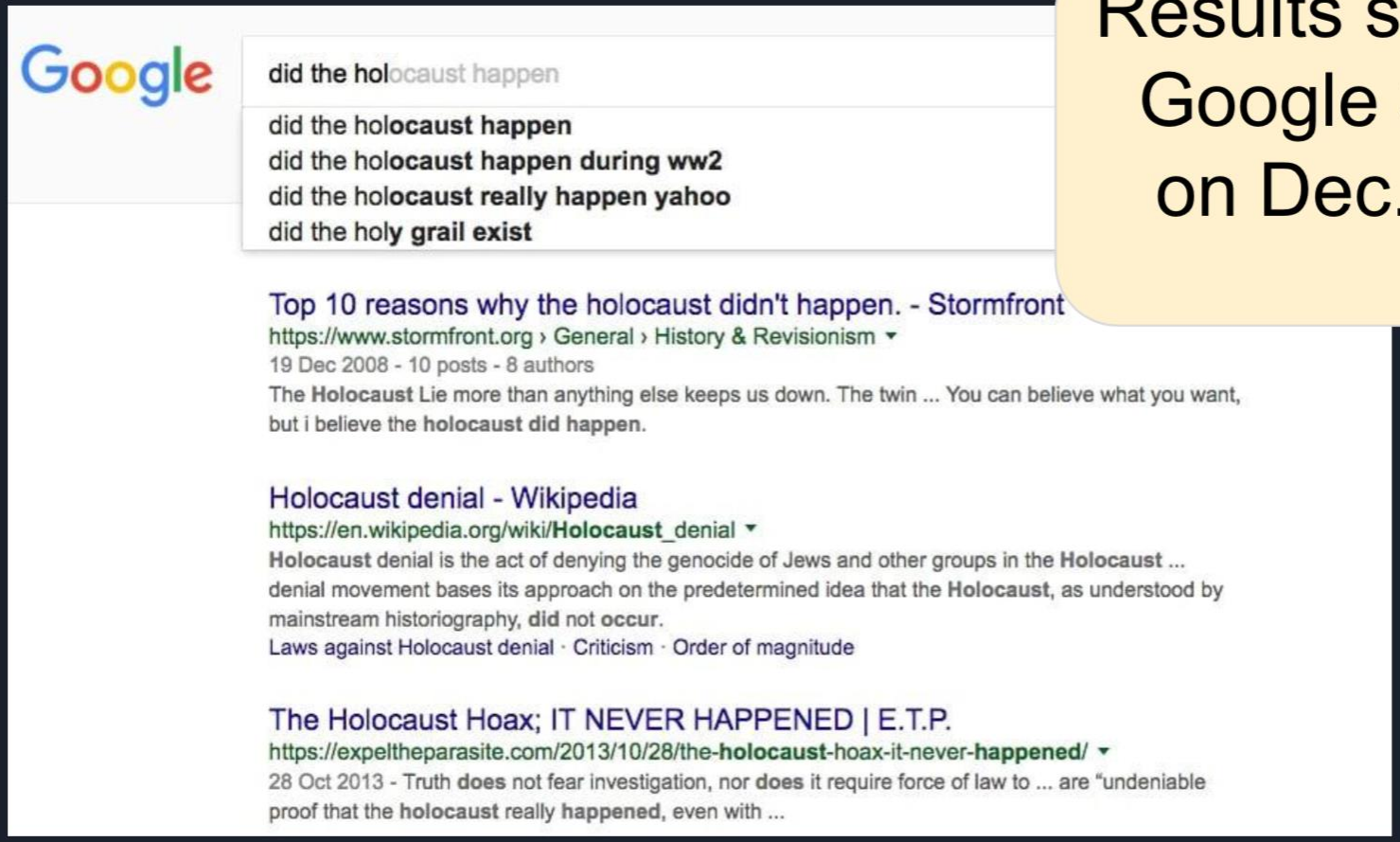
Data Selection / Acquisition / Harvesting

Experimental Setting

Analysis

- Evaluation metrics, statistical significance
- Failure analysis, qualitative analysis
- scope & limitations of results

Right or wrong?



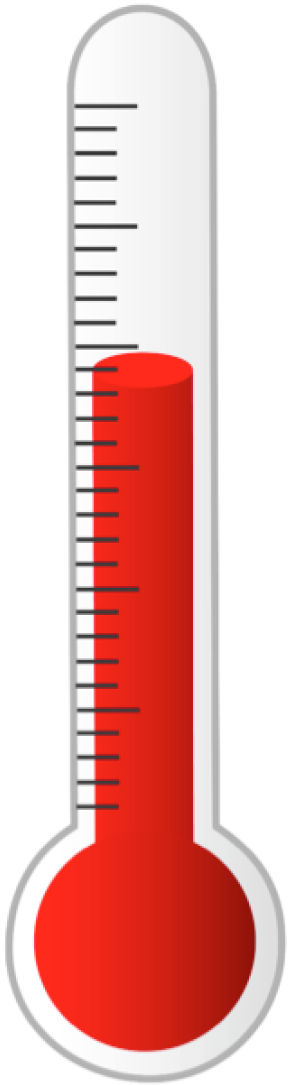
Results shown by Google Search on Dec. 2016.

Source: The Guardian

Task biases

	PUBLIC	PRIVATE
INCENTIVE	Publishability	User satisfaction Attention retention
BIG (?) DATA OUTCOMES	70% publications come from public institutions	70% of research is private
	IRRELEVANT (reproducible?)	IRREPRODUCIBLE (relevant?)

Metrics define the problem



"Temperature is what a thermometer measures"

The case of recommendation

NETFLIX CHALLENGE

predict user ratings

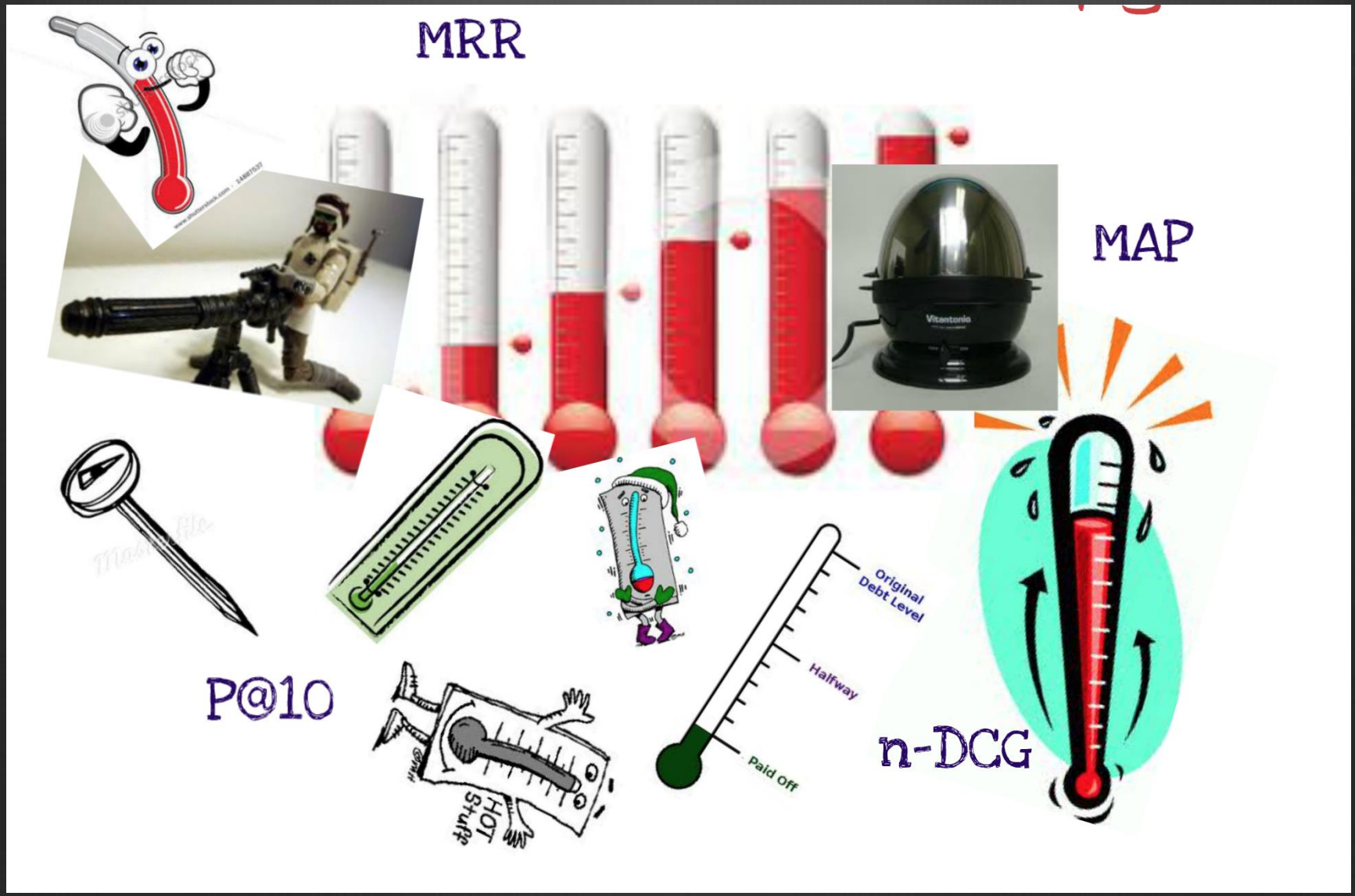
→ **classification metrics**

REAL TASK

suggest something the user likes

→ **ranking metrics**

Let's go ranking



Metric Selection Hall of Fame

Use the most popular

Use the simplest

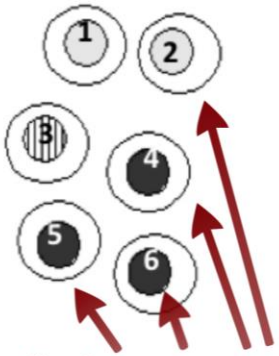
Get creative

Mirror, mirror, who is the prettiest…?

Wait… Do all MAP implementations give the same output?

Popular: Purity & Inverse Purity

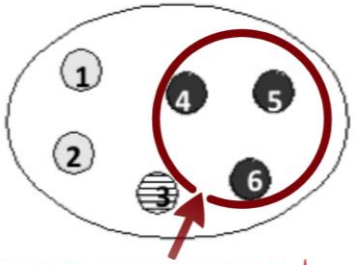
Scattered



All clusters are pure

P: 1,00
IP: 0,48
$F_{0,5}$: 0,65

Joined

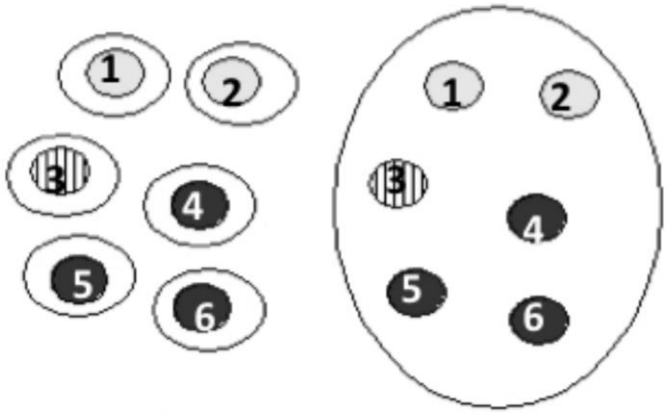


All classes are grouped

P: 0,50
IP: 1,00
$F_{0,5}$: 0,67

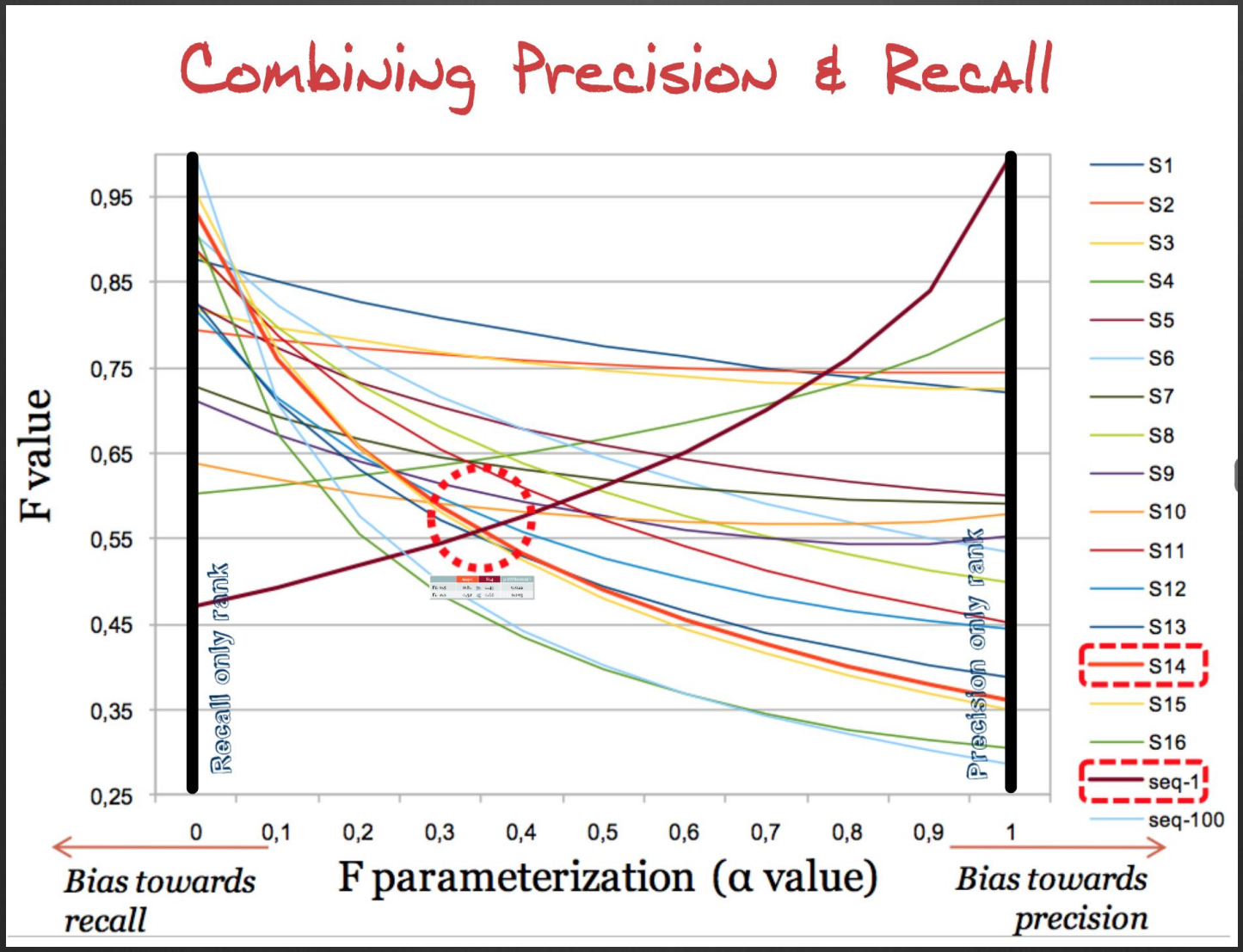
purity: is the cluster clean?
inverse purity: is the class grouped?

Cheat system



P: 0,75
IP: 1,00
$F_{0,5}$: 0,86

What about multiple quality dimensions?



Data Harvesting

- Wisdom of the crowds… only if there is diversity of opinion, independence and decentralization!
- Biases everywhere in our online society:
  - E.g. Twitter population is not a random sample of citizens
  - Cognitive biases and techniques to exploit them (we share what we find outrageous) [economic incentives!]
- "Big data, small data, right data" (Ricardo Baeza)

Ground Turth harvesting: RecSys example

- Random sampling of user/item scores
- Training/Validation/Test split
- Strong bias for popular items!

Data annotation: the entailment case

Crowdworkers find trick to produce test cases quickly

- President Rajoy deposition was, according to the judge, unbelievable ENTAILS Rajoy was lying
- President Rajoy deposition was, according to the judge, unbelievable **NOT ENTAILS** Rajoy was **not** lying

What the algorithms learn

A negation in the consequent is highly correlated with an invalid entailment

Consequence: entailment resolution systems have been highly overrated for years

A/B testing problems

- Weak baselines (“Improvements that don’t add up”)
- Difficult to compare with state of the art
  - Previous approaches difficult to reproduce
  - System outputs are not usually available
- Lab experimentation usually ignore real-world scenarios
- When there are benchmarks, there is overfitting
- Importance of evaluation campaigns

Analysis problems

- sensitivity rather than reliability
- finding rather than explaining differences
- Averages that hide behavior: across metrics, across classes, across test cases. Example: classification efficacy measured as arithmetic mean of harmonic means per class

Analysis of ML outcomes

- Knowledge-based systems vs ignorance-based systems.
- Overfitting: what do ML algorithms actually learn?
- ML elevates correlation to causality:

input X correlates with output Y → ML outputs Y ***because of*** X

Great hockey player ←→ born in January

- Bias & second-order bias:
  - Google holocaust biased crowds, unpredictable results.
  - Second order bias: tag recommendation in Flickr. Without human input, the algorithm cannot improve, is doing a harakiri (Baeza-Yates).

Evaluation and Replication (the scientist)

Trusted Data Analytics Seminar

TU Delft, 15 June 2018

Julio Gonzalo (UNED, Madrid, Spain)


EvALL
Sign In
Evaluate using an existing benchmark
Publish a new system output
Evaluate with your own benchmark
Publish a new benchmark
Menu
Learn more
Browse repository
What is EvALL?