Economical Evaluation of Systems: Users as a Signal

Kevin Roitero
Research Talk
New York, 7 June 2019

CV: https://kevinroitero.com/resources/Kevin-Roitero-CV.pdf

About Myself

About Myself

November 2016 -Today (1 November 2019)
Ph.D. Candidate at University of Udine, Italy.
Work on Information Retrieval, Crowdsourcing, Artificial Intelligence, Statistical modeling, Transfer Learning, Data mining and Analysis. Supervisor: Stefano Mizzaro.

April 2017 - June 2017
Visiting period at University Of Sheffield, UK.
Work on Crowdsourcing and statistical modeling, with a focus on user interaction and engagement, and agreement metrics.

March 2018 - July 2018
Visiting period at RMIT University, Australia.
Work on Agreement metrics, relevance scales, and Crowdsourcing Evaluation.

September 2018 - December 2018
Research Intern at Spotify London.
Work on Statistical Modeling, Machine and Transfer Learning for BaRT.

Outline

Outline

- Measure Agreement
- Incorporate Agreement in Evaluation
- Effect of Judgment Scales on Agreement
- Bias
    - Bias in (IR) Evaluation
    - Bias in Scholarly Publishing
- Economical Evaluation of Systems

Measure Agreement

Setting

- micro-task crowdsourcing
- many workers do the same task
- agreement among workers can / should be leveraged
- leveraging agreement can be useful for:
  - estimating the reliability of collected data
  - understanding behavior of the workers

Agreement Formalization

$$
\begin{array}{c c}
 & \begin{array}{cccccc} \text{Item}_1 & \text{Item}_2 & \cdots & \text{Item}_j & \cdots & \text{Item}_N \end{array} \\
\begin{array}{c} \text{Assessor}_1 \\ \text{Assessor}_2 \\ \vdots \\ \vdots \\ \text{Assessor}_i \\ \vdots \\ \vdots \\ \text{Assessor}_M \end{array} &
\begin{bmatrix}
r_{1,1} & r_{1,2} & \cdots & r_{1,j} & \cdots & r_{1,N} \\
r_{2,1} & r_{2,2} & \cdots & r_{2,j} & \cdots & r_{2,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r_{i,1} & r_{i,2} & \cdots & r_{i,j} & \cdots & r_{i,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r_{M,1} & r_{M,2} & \cdots & r_{M,j} & \cdots & r_{M,N}
\end{bmatrix}
\end{array}
$$

Agreement Formalization

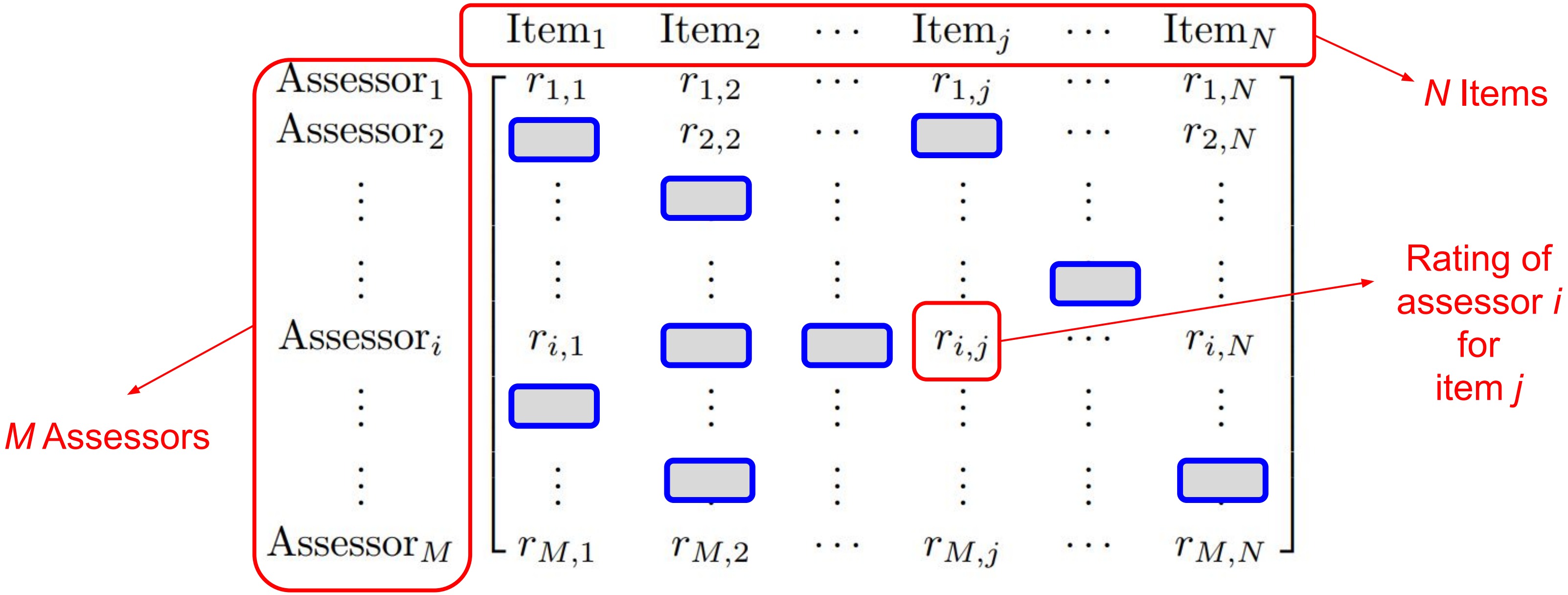
$$\begin{array}{cc}
 & \begin{array}{cccccc} \text{Item}_1 & \text{Item}_2 & \cdots & \text{Item}_j & \cdots & \text{Item}_N \end{array} \\
\begin{array}{c} \text{Assessor}_1 \\ \text{Assessor}_2 \\ \vdots \\ \vdots \\ \text{Assessor}_i \\ \vdots \\ \vdots \\ \text{Assessor}_M \end{array} &
\begin{bmatrix}
r_{1,1} & r_{1,2} & \cdots & r_{1,j} & \cdots & r_{1,N} \\
r_{2,1} & r_{2,2} & \cdots & r_{2,j} & \cdots & r_{2,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r_{i,1} & r_{i,2} & \cdots & \boxed{r_{i,j}} & \cdots & r_{i,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r_{M,1} & r_{M,2} & \cdots & r_{M,j} & \cdots & r_{M,N}
\end{bmatrix}
\end{array}$$

N Items

Rating of assessor *i* for item *j*

M Assessors

Agreement Formalization



This matrix is often **very** sparse in crowdsourcing

There are Several Agreement Measures

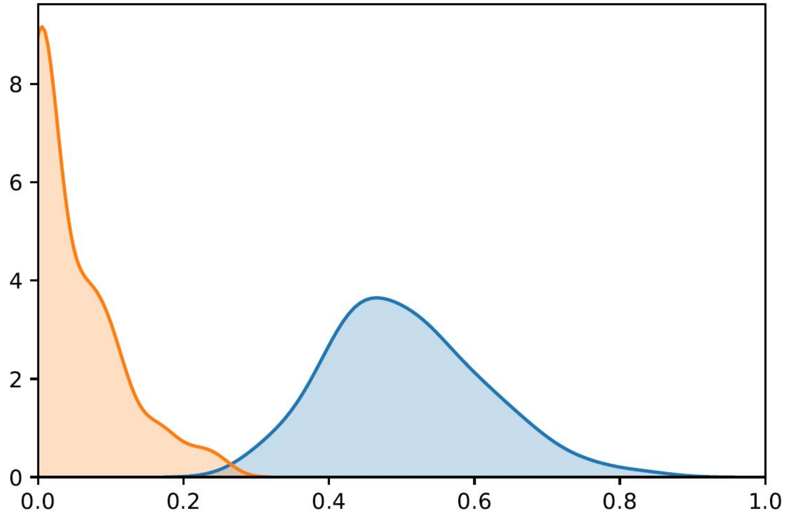
- Percentage Agreement (PA)
- Scott’s π
- Cohen’s κ
- Intraclass Correlation Coefficient (ICC)
- Fleiss κ
- Krippendorff’s Alpha

Current Agreement Measures Are Inadequate

- measures often borrowed from other scenarios with **different assumptions** (which usually do not hold for crowdsourcing):
  - one assessor rates all items
  - all assessors rate all items
  - limited and fixed (= known) number of assessors
- measures are often designed for estimating **data reliability**, not **agreement**
  - **reliability**: the capacity of any measurement tool to differentiate between respondents when measured twice under the same conditions. [Berchtold]
  - **agreement**: the capacity of any other measurement tool applied twice on the same respondents under the same conditions to provide strictly identical results. [Berchtold]
  - reliability can be considered as a necessary but not sufficient condition to demonstrate agreement. [Berchtold]

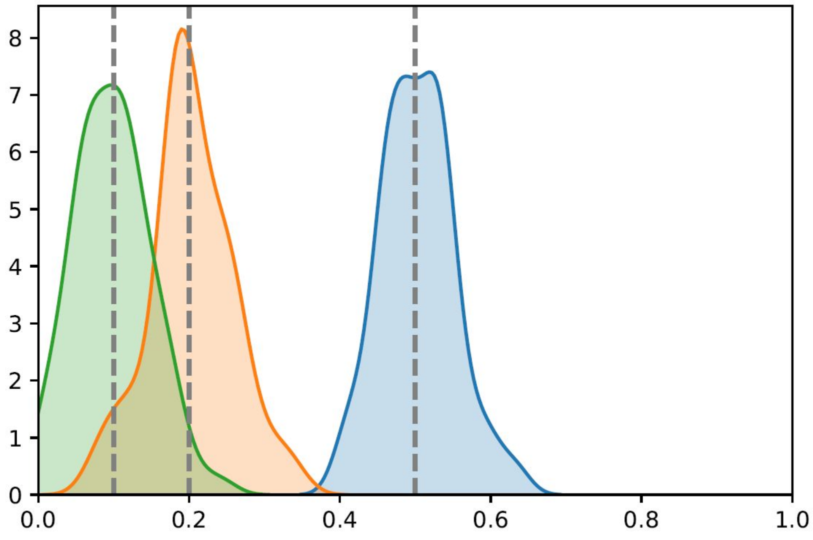
Problems

- there is more variability of judgments in the centre of the scale w.r.t. scale boundaries.
  → can lead to over-estimate agreement close to scale boundaries.



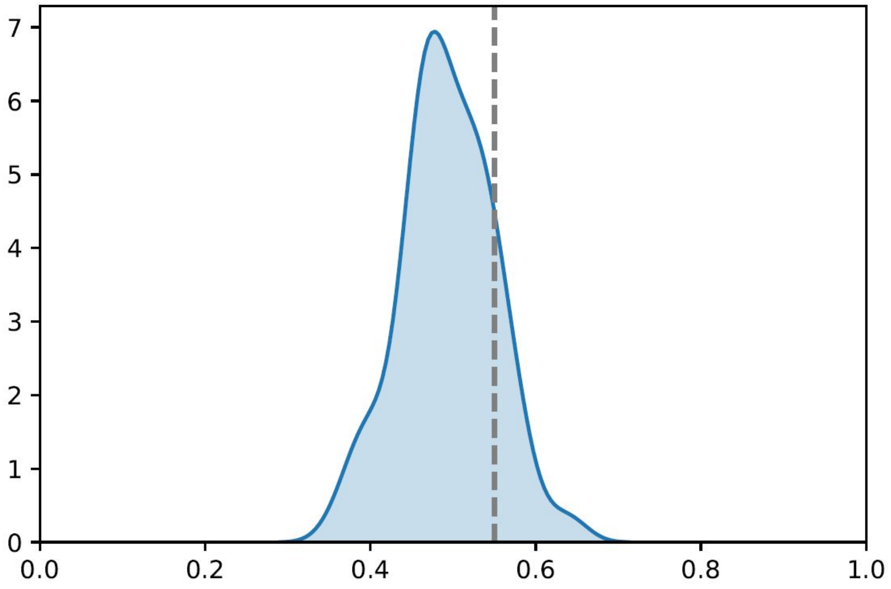
Problems

- the concentration point can be different for different items
  → can lead to over/under-estimate agreement



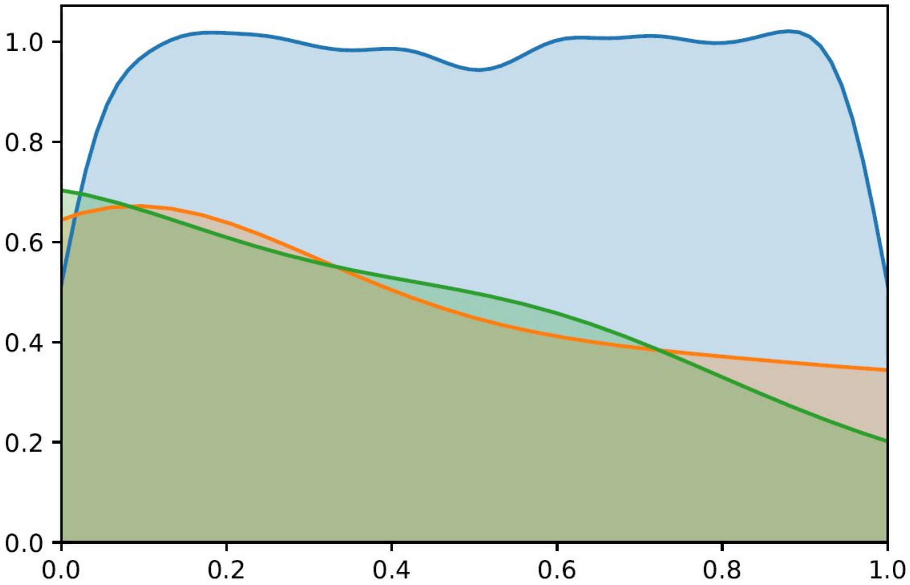
Problems

- additional information is often not considered (e.g., gold questions)



Problems

- different ideas of "**agreement by chance**" definition
- correction by chance assumptions are often violated in crowdsourcing setting



Real Problems with State-of-the-Art Measures

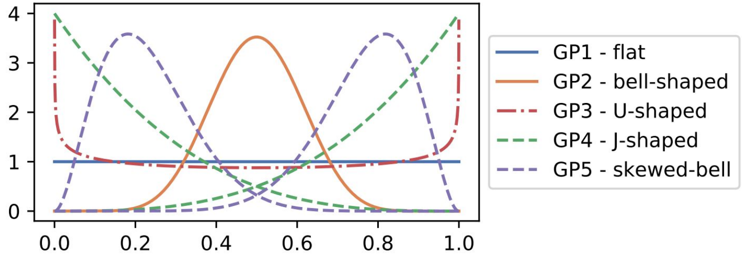
- Percentage Agreement (PA)
  - does not consider agreement by chance
  - works only with nominal data
  - depends on the scale granularity (can not compare different scales)
- Scott’s π and Cohen’s κ
  - work only with two assessors
  - work only with nominal data
- Intraclass Correlation Coefficient (ICC)
  - assessor have same marginal probability of an answer (not true in crowdsourcing)
  - equivalent to weighted Cohen’s κ
- Fleiss κ
  - Generalizes κ to multiple assessors (i.e., shares the same issues)

Real Problems with State-of-the-Art Measures

- Krippendorff’s Alpha: an attempt to generalize previous metrics
    - **Random guessing can have high agreement**
    - **Random guessing may have more agreement than honest coding**
    - High agreement, low reliability
    - Zero change in percentage agreement causing radical drop in reliability.
    - **Eliminating disagreements does not improve agreement**
    - Honest work as bad as coin flipping.
    - Two datasets: same quality, same agreement; but higher reliability in one.
    - punishing larger sample and replicability (i.e., data quantity dependent)
    - **“reverse answer” problem** $([1,0,0,0,1] \neq [0,1,1,1,0])$

Our Measure: Φ

- "*agreement is the amount of concentration around a data value*"
- if not agreement, we have disagreement → negative agreement
- in practice:
  - first, we fit a distribution over the histogram of the ratings
  - then, we measure the dispersion of such distribution
- the fitting distribution has to be general enough to capture:
  - random judgments → flat distribution
  - agreement → bell-shaped distribution
  - agreement around scale boundaries → J-distribution
  - disagreement → U shaped distribution
  - → use a Beta function



- we should have a minimal number of parameters, to avoid overfitting

Our Measure: Φ

- we use a Beta distribution to model our scenario: $B(a, b)$
- we re-parametrize the distribution in terms of the mean value μ and the precision p as $\mu = \frac{a}{a+b}; p = a + b$
- now, we can treat separately mean and dispersion
- we can have a metric that is agnostic of the mean value
- then, we transform to have values in the [-1, +1] range:

$$\Phi = 1 - 2^{\frac{-p \, log2}{2}}$$

Our Measure: Φ

- we use Bayesian inference to compute Φ:

$$\mathrm{P}(\vec{\mu}, \Phi | X) = \prod_{i=1}^{N} \prod_{j=1}^{M} B(X_{i,j} | \mu_i, \Phi)^{O_{ij}}$$

$$\prod_{i=1}^{N} \mathcal{N}(1/2, \sigma_\mu^2 \mathbf{I}) \mathcal{N}(0, \sigma_\Phi^2) C,$$

probability of observing the mean values, with a common dispersion, given the observed data

- Then, we estimate Φ using

$$\hat{\Phi} = \arg\max_{\Phi} \mathrm{P}(\vec{\mu}, \Phi | X).$$

Our Measure: Φ

- we use Bayesian inference to compute Φ:

$$\mathrm{P}(\vec{\mu}, \Phi | X) = \prod_{i=1}^{N} \prod_{j=1}^{M} B(X_{i,j} | \mu_i, \Phi)^{O_{ij}}$$

$$\prod_{i=1}^{N} \mathcal{N}(1/2, \sigma_\mu^2 \mathbf{I}) \mathcal{N}(0, \sigma_\Phi^2) C,$$

probability of observing the mean values, with a common dispersion, given the observed data

- Then, we estimate Φ using

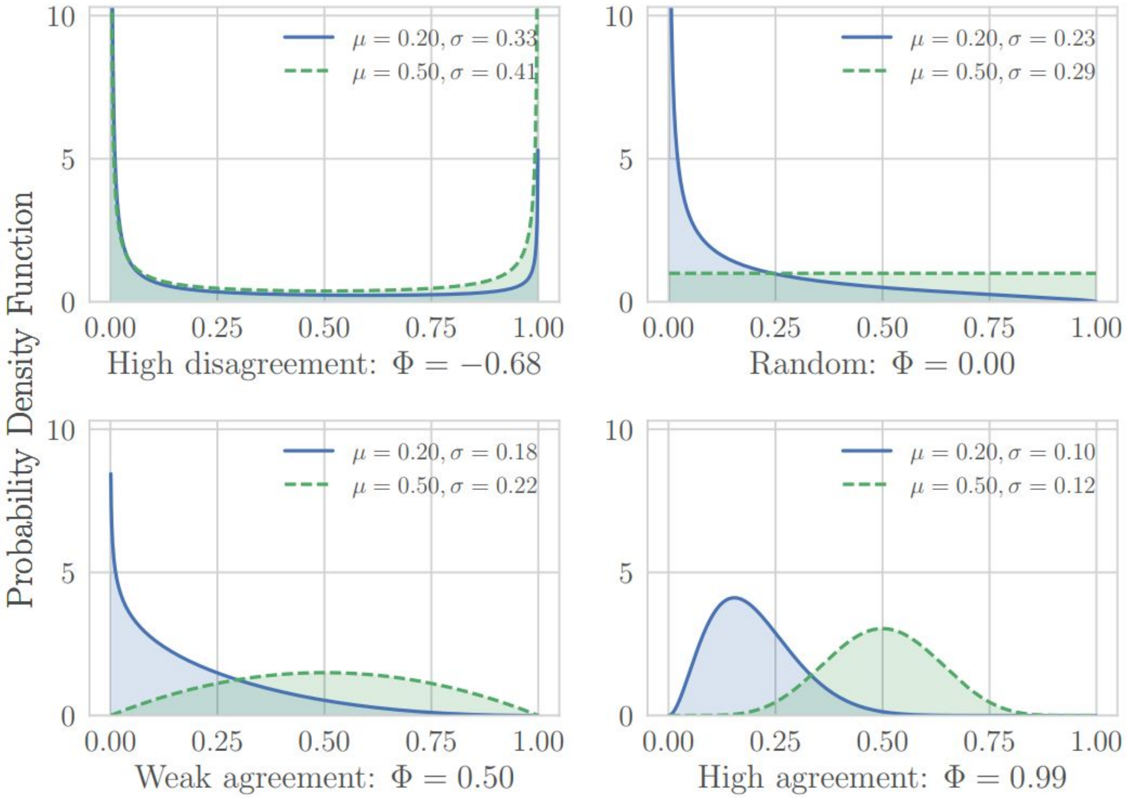
$$\hat{\Phi} = \arg\max_{\Phi} \mathrm{P}(\vec{\mu}, \Phi | X).$$

the formula can change to incorporate custom ground truth

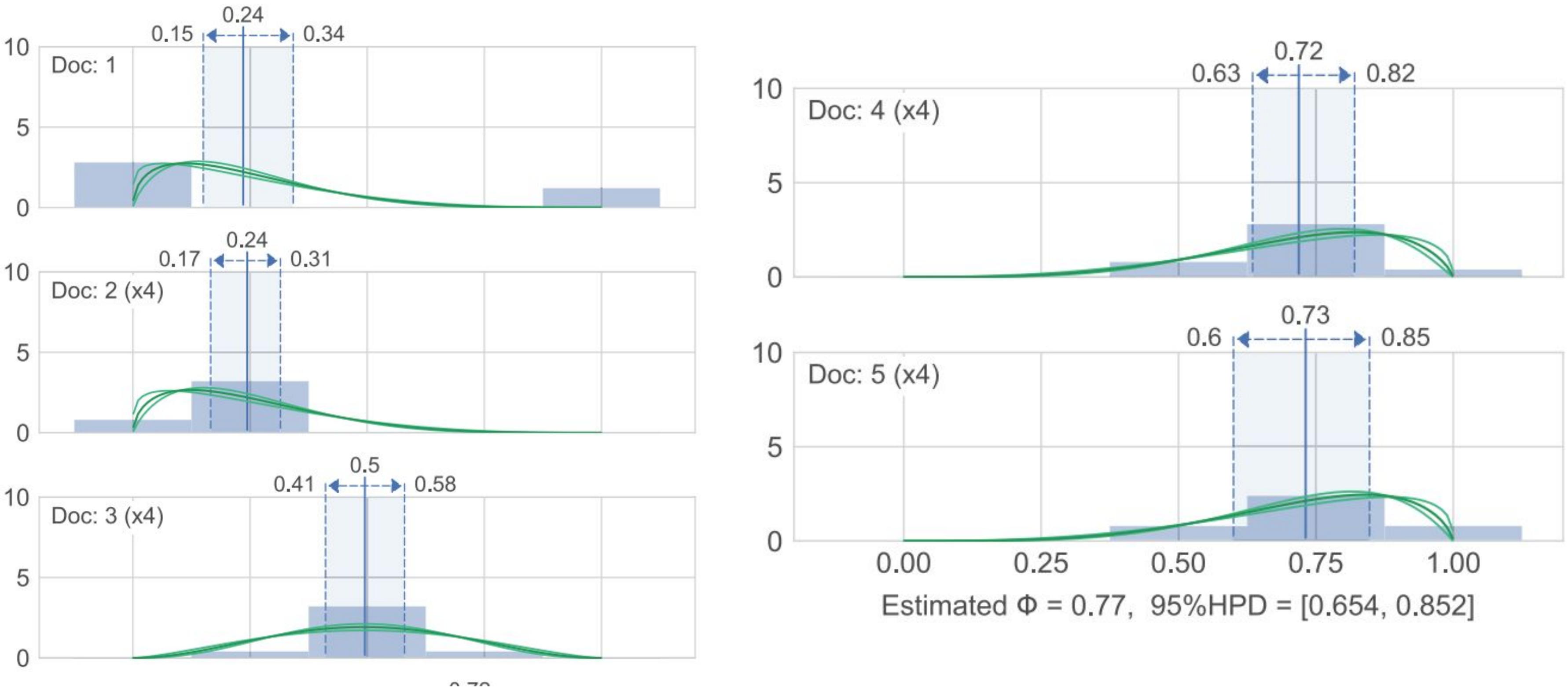
Φ Interpretation

- **High Disagreement**. When **$\Phi < 0$**, there is no central tendency value but rather a tendency to exclude a central area (polarized behavior)
- **Random**. When **$\Phi=0$**, the behavior is equivalent with a unbounded uniform process censored on the scale
- **Weak Agreement**. When **$0 < \Phi \leq 0.5$**, the distribution has no inflection point, but there is a unique central tendency or a dispersion that is smaller than a uniform process
- **High Agreement**. When **$\Phi > 0.5$**, the distribution is bell shaped with two inflection points, more narrow around the mean as Φ grows

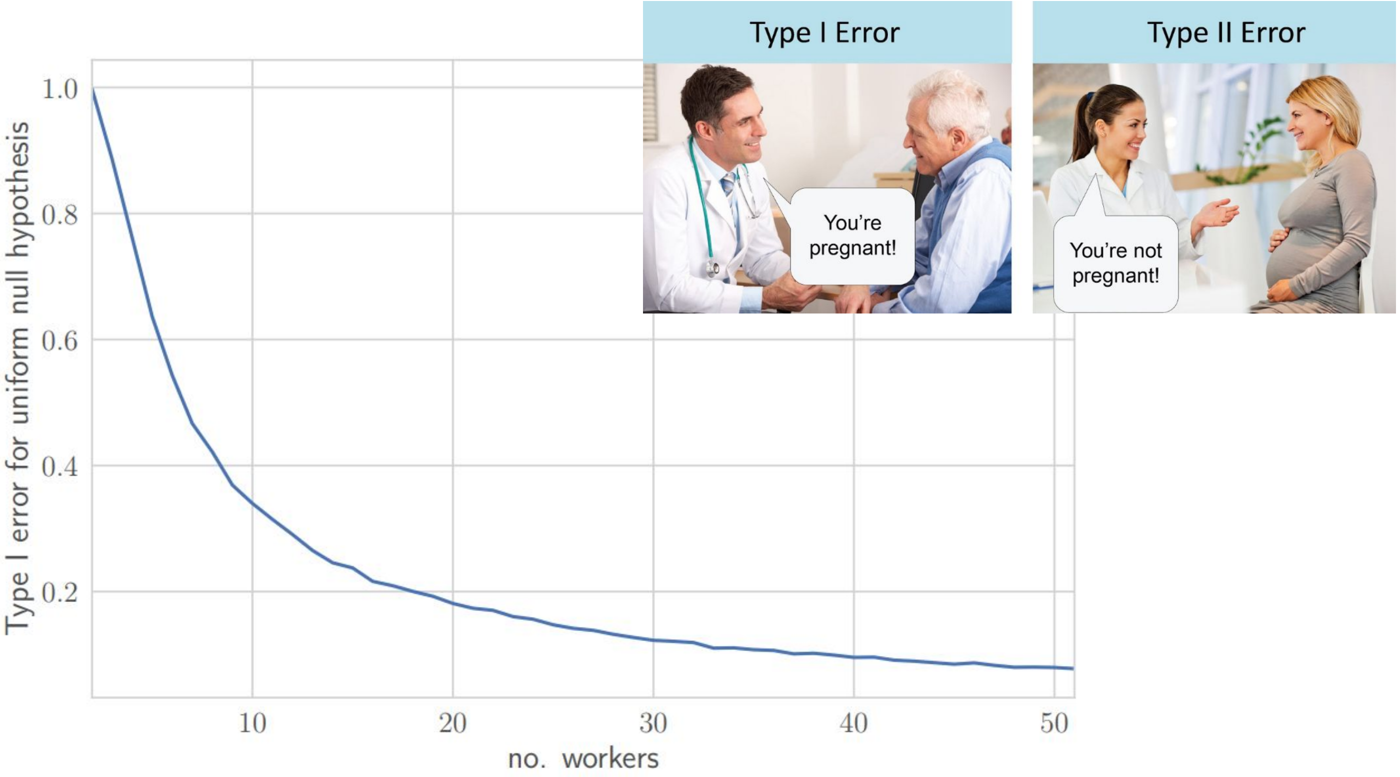
Examples of Φ Shapes



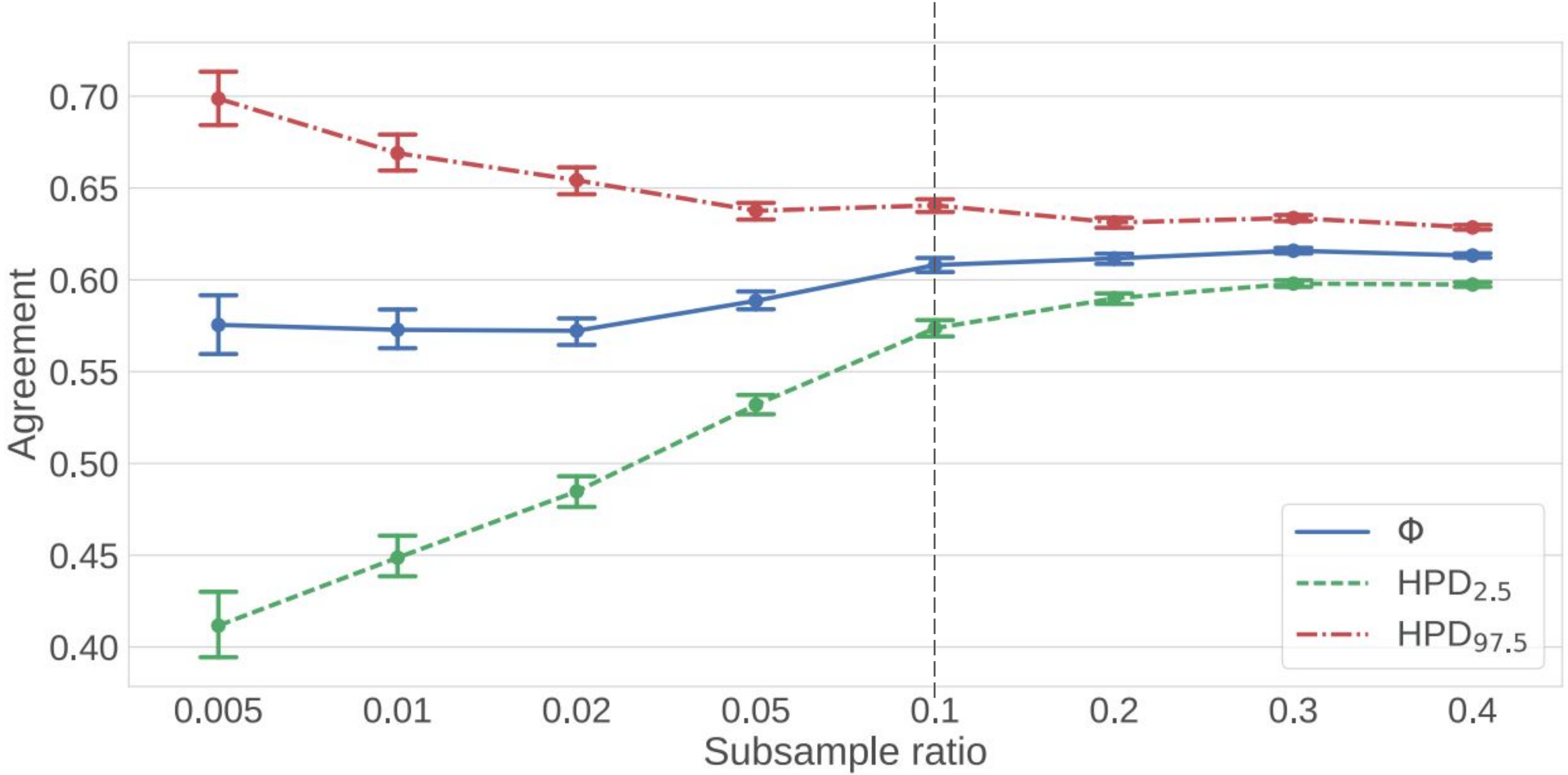
Φ in Action on Real Data



Robustness of Φ



Confidence Interval, Robustness of Φ



Incorporating Agreement in Evaluation

Setting

- test collection evaluation scenario:
  - Document collection
  - Information Needs ≈ Queries (called topics)
  - A set of Information Retrieval systems (called runs)
- each system retrieves a ranked list of documents for each topic
- human made (= from experts) relevance judgments for a subset of documents of each topic
- metrics (such as Precision, Recall, NDCG, etc.) are computed
- systems are then ranked

Setting

- test collection evaluation scenario:
  - Document collection
  - Information Needs ≈ Queries (called topics)
  - A set of Information Retrieval systems (called runs)
- each system retrieves a ranked list of documents for each topic
- human made (= from experts) relevance judgments for a subset of documents of each topic
- metrics (such as Precision, Recall, NDCG, etc.) are computed
- systems are then ranked

Very Expensive (Money + Time). A more economical evaluation process is needed.

Evaluation of IR systems

N Topics / Queries

$$
\begin{array}{cc}
 & \begin{matrix} \text{Topic}_1 & \text{Topic}_2 & \cdots & \text{Topic}_j & \cdots & \text{Topic}_N \end{matrix} \\
\begin{matrix} \text{system}_1 \\ \text{system}_2 \\ \vdots \\ \vdots \\ \text{system}_i \\ \vdots \\ \vdots \\ \text{system}_M \end{matrix} &
\begin{bmatrix}
AP_{1,1} & AP_{1,2} & \cdots & AP_{1,j} & \cdots & AP_{1,N} \\
AP_{2,1} & AP_{2,2} & \cdots & AP_{2,j} & \cdots & AP_{2,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
AP_{i,1} & AP_{i,2} & \cdots & \boxed{AP_{i,j}} & \cdots & AP_{i,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
AP_{M,1} & AP_{M,2} & \cdots & AP_{M,j} & \cdots & AP_{M,N}
\end{bmatrix}
\end{array}
$$

Average Precision of system *i* for topic *j*

M Systems

Test Collection Evaluation Using Crowdsourcing

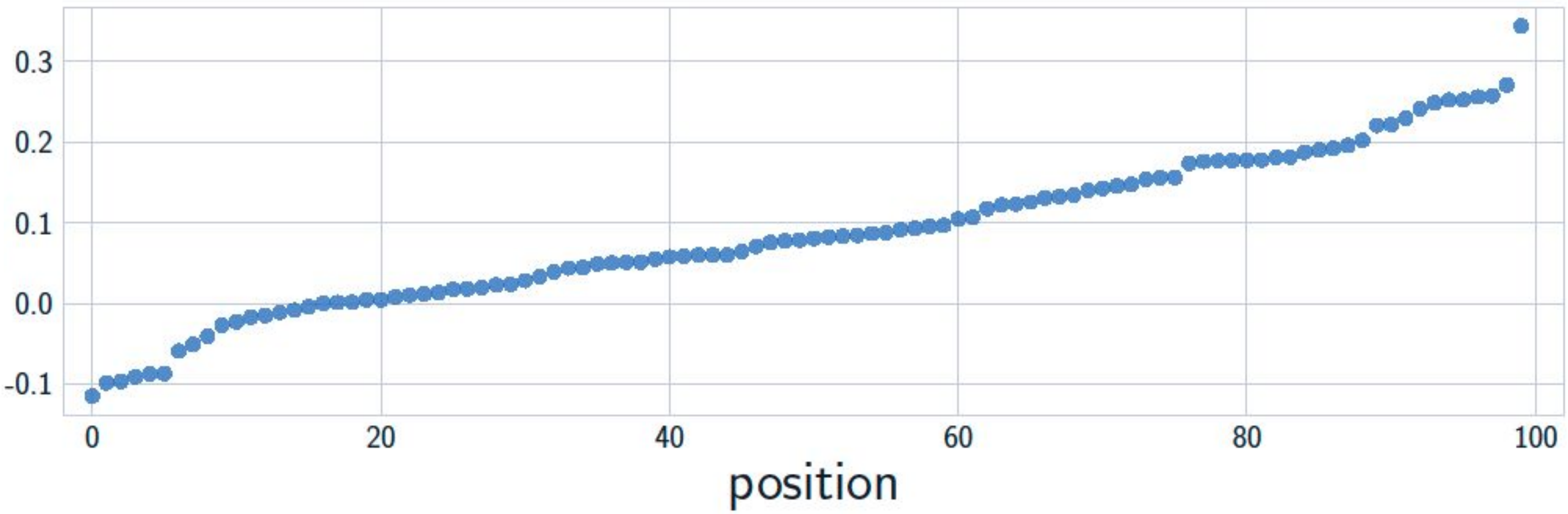
- relevance assessment is a crucial activity in IR effectiveness evaluation.
- relevance is often uncertain:
  - ambiguous query
  - terms with multiple meanings
  - unclear or ambiguous information needs
  - non-ideal work conditions of relevance assessors
  - inadequate or even malicious assessors
  - …
- even worse when crowdsourcing is used!

Test Collection Evaluation Using Crowdsourcing

- when crowdsourcing relevance assessments:
  - each document is redundantly assessed by several crowd workers
  - workers judge the relevance of the document to a specific topic
  - all the relevance judgments for a document by different workers are aggregated to compute a final relevance score
- → only the final aggregated scores are used to compute IR evaluation metrics.
- → agreement information is lost!

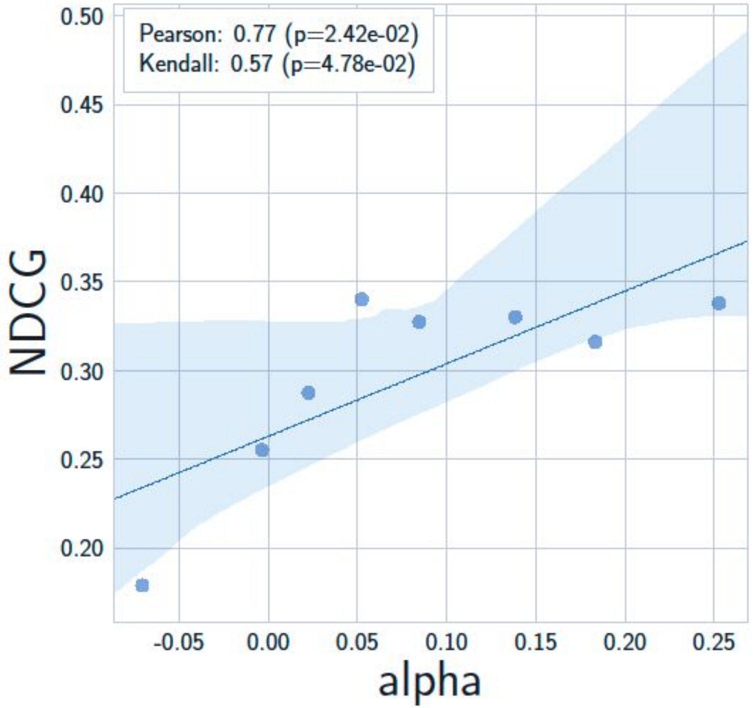
- (log all user behavior, because why not)

Different Topics, Different Agreement



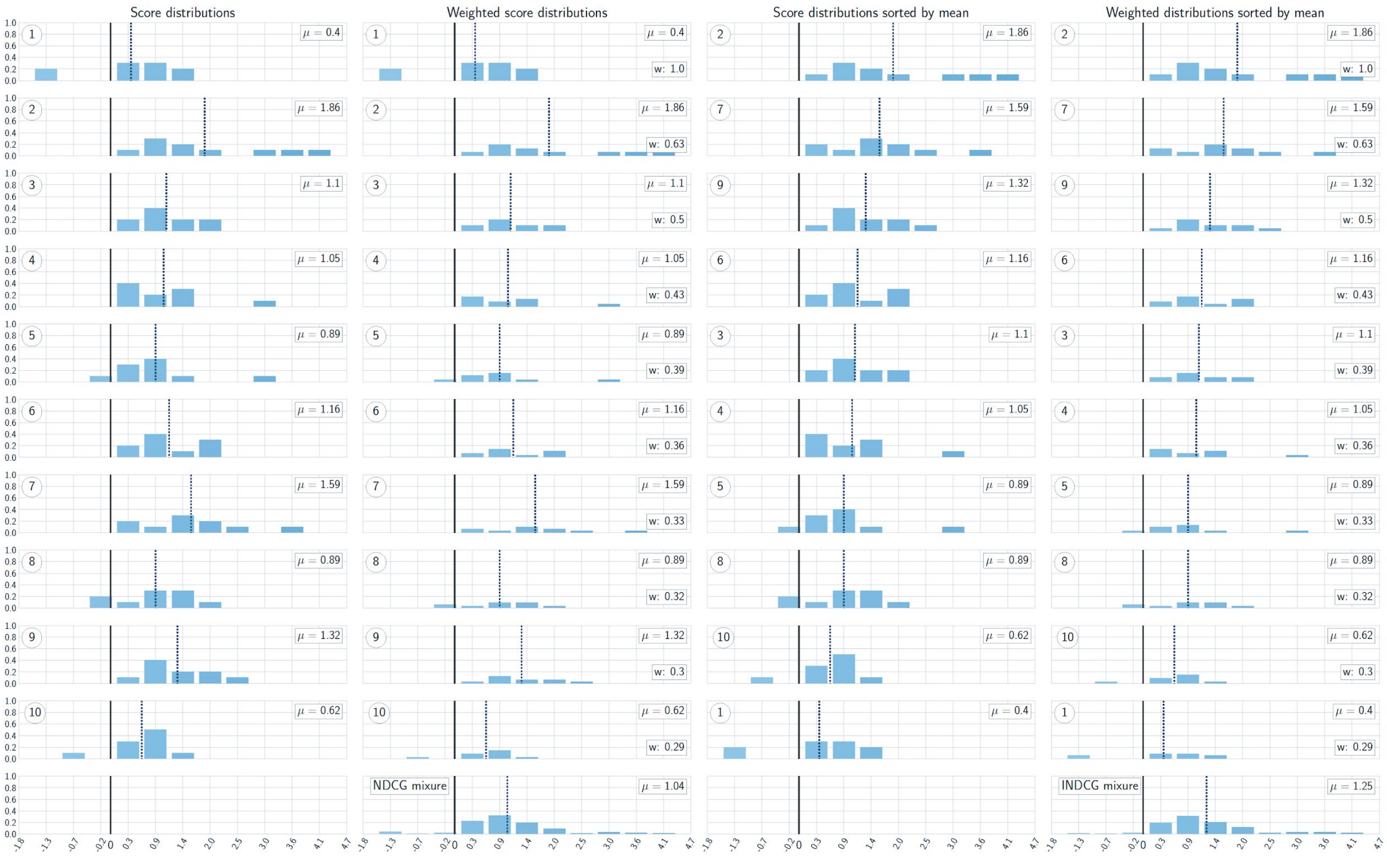
Different topics have a very different agreement

Agreement over Relevance



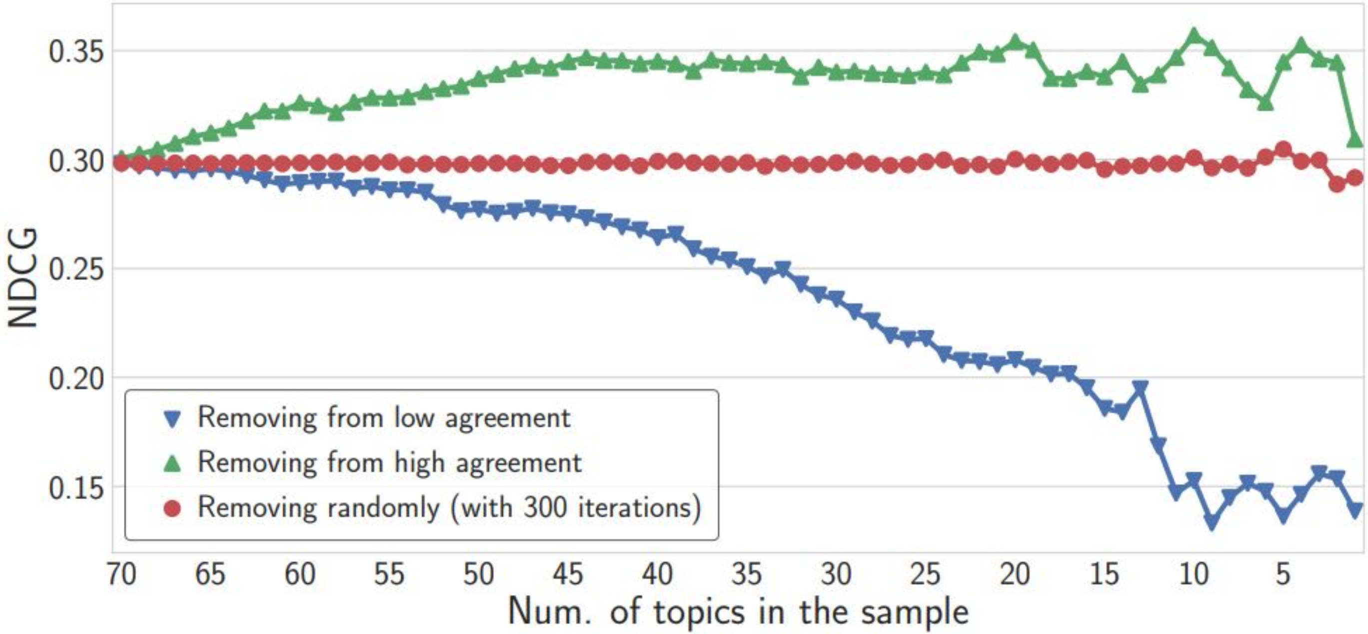
there is a relationship between agreement and effectiveness

Incorporate Agreement in the Evaluation



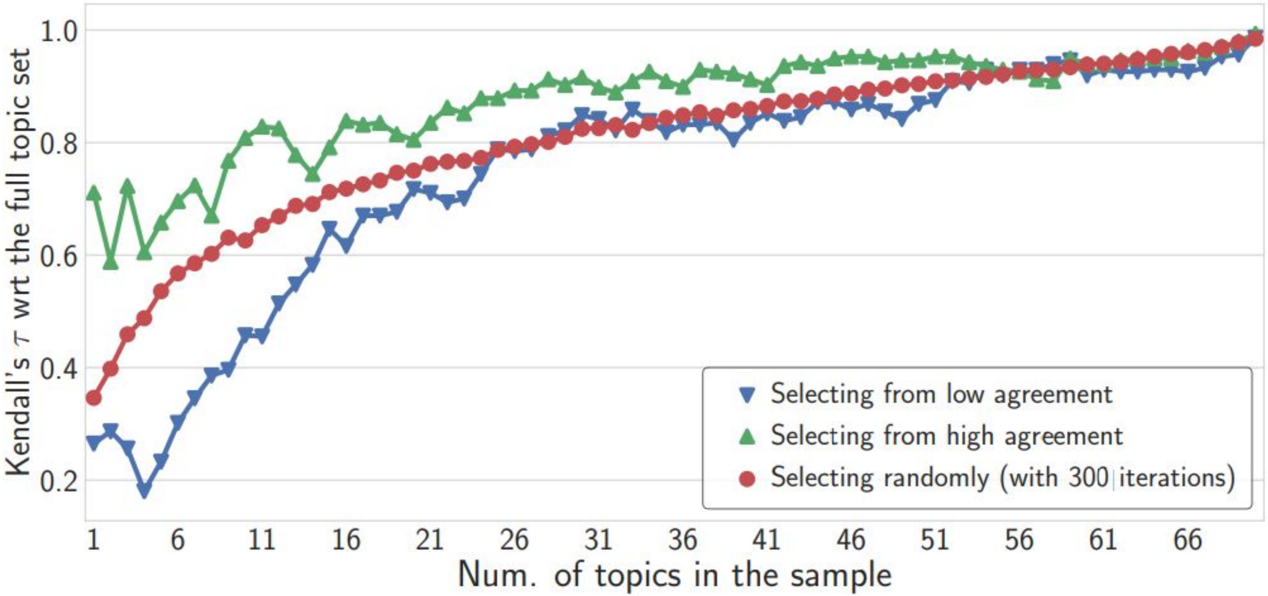
ad-hoc metrics (such as modified NDCG) can naturally incorporate agreement

Incorporate Document Agreement in the Evaluation



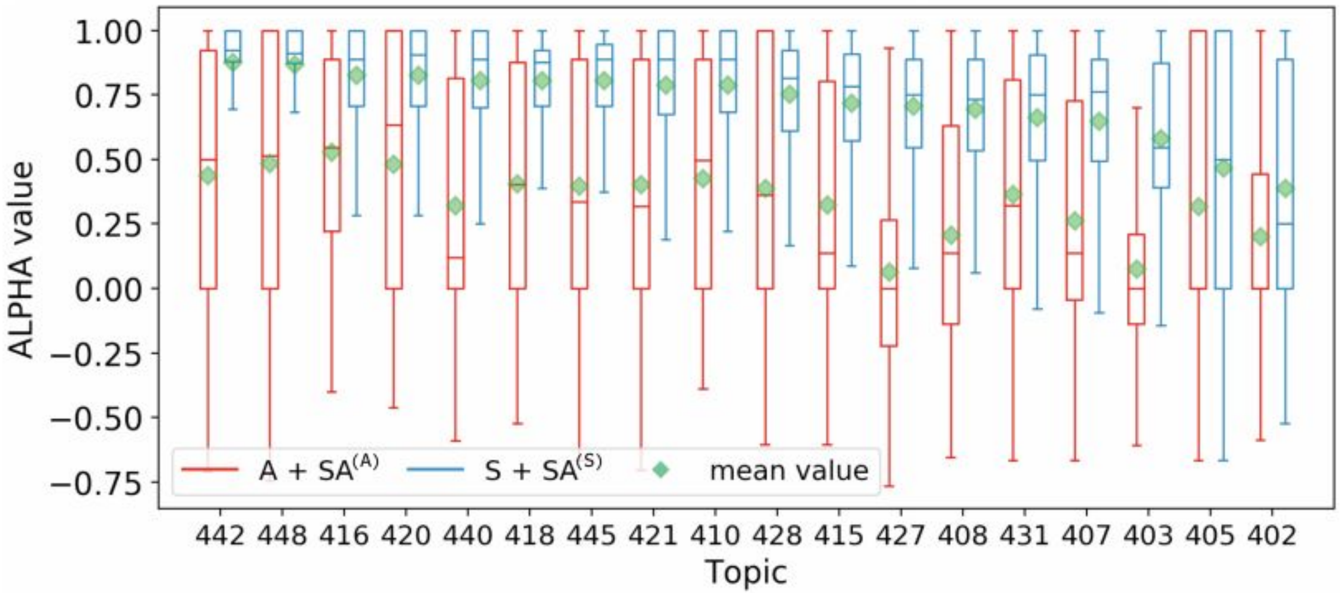
Documents with less/more agreement impact differently in the evaluation

Incorporate Topic Agreement in the Evaluation



leveraging agreement can lead to save resources

Incorporate Crowd Agreement in the Evaluation



different users have a different level of agreement w.r.t. ground truth

Control Crowd Agreement in the Evaluation

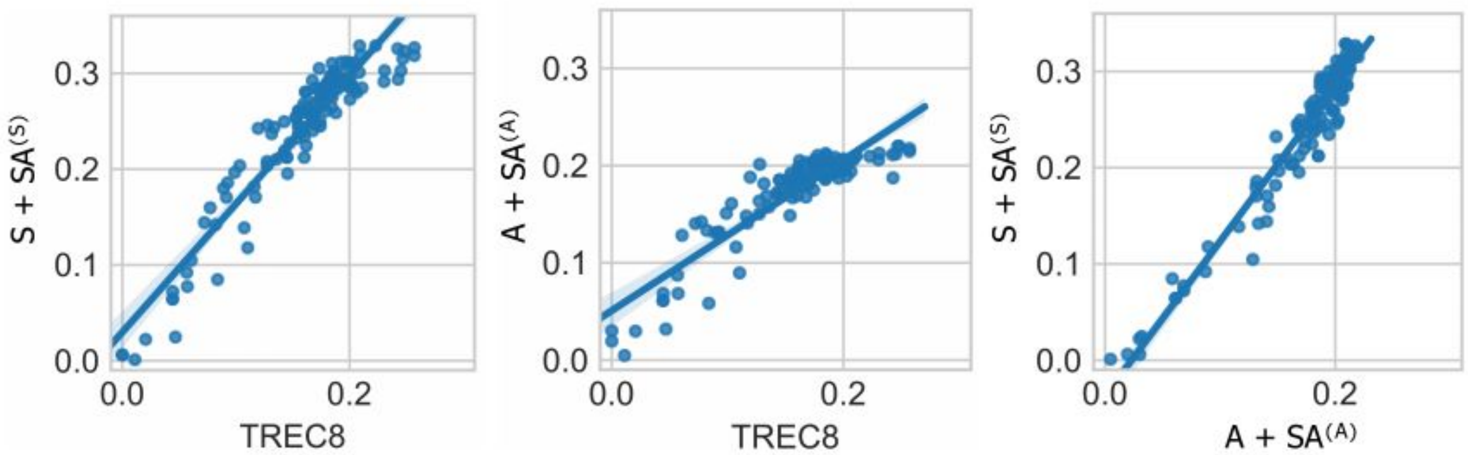
	F	Adj. p-value	ω^2
Two-way ANCOVA			
Number of Sessions			
Reward	76.07	$p < .001$	$\boxed{0.11}$
Task Length	113.01	$p < .001$	$\boxed{0.18}$
Reward:Task Length	43.35	$p < .001$	0.07
Number of Steps			
Reward	48.05	$p < .001$	0.10
Task Length	22.96	$p < .001$	0.05
Reward:Task Length	4.04	$p = .27$	0.01
AVG Time per Session			
Reward	0.08	$p = 1$	-0.01
Task Length	1.38	$p = 1$	0.01
Reward:Task Length	1.01	$p = 1$	0.01
One-way ANOVA			
Number of Sessions			
Quality Control	0.76	$p = 1$	-0.01
Number of Steps			
Quality Control	47.31	$p < .001$	0.09
AVG Time per Sessions			
Quality Control	65.47	$p < .001$	0.12

Short and highly paid tasks tend to maximize agreement.

Length is more important than reward.

External factors can influence User agreement

Incorporate Agreement in the System Evaluation



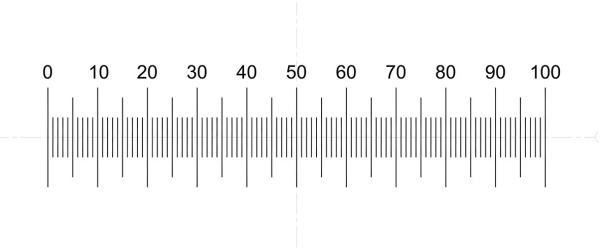
Different agreement levels lead to obtain less / more quality

Effect of Judgment Scale on Agreement

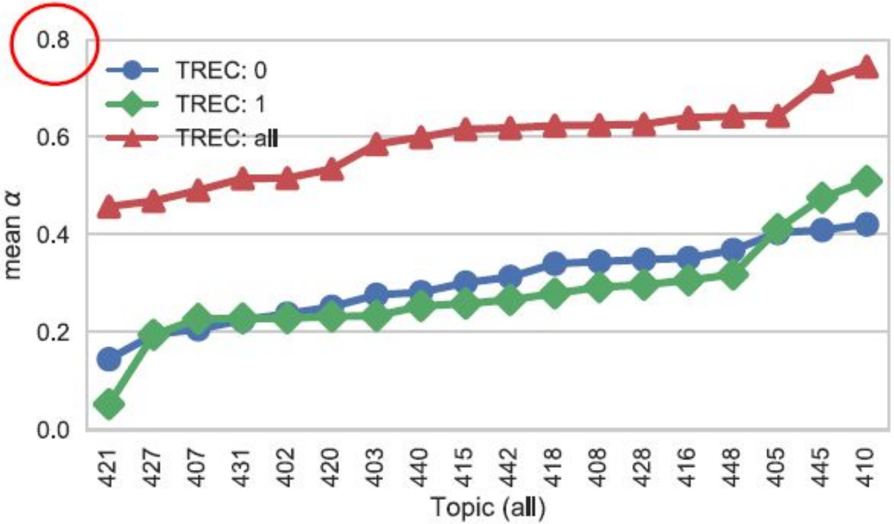
How many Relevance Judgments?

- Historically (until 90s) → two
- NDCG metric (SIGIR 2000) → multi-level judgements
- Sormunen (2002) → four
- Terabyte Track (2004) → three
- Tang et al. (2007) → seven (from 2 to 11)
- Web Track (2014) → six
- Magnitude Estimation (SIGIR 2016) → infinite (0, +∞)
- S100 (SIGIR 2018) → [0,100]

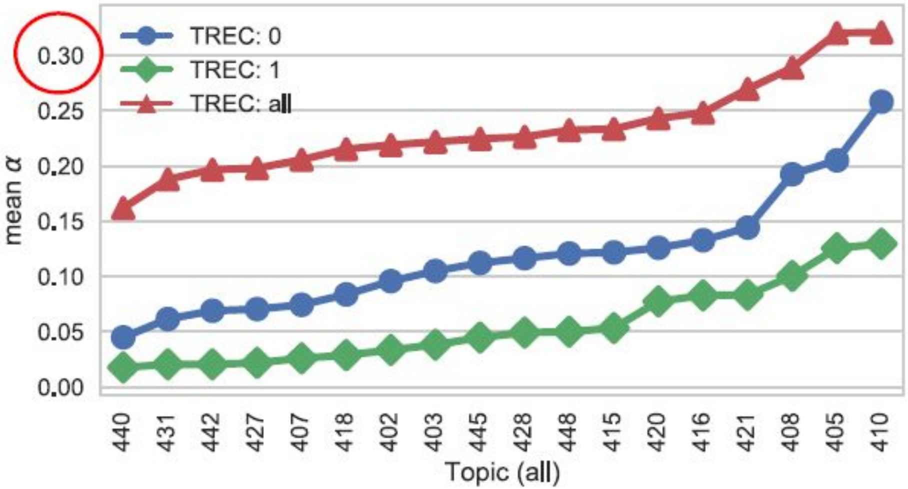
	RELEVANT	NON-RELEVANT	
RETRIEVED	a	b	a + b
NOT RETRIEVED	c	d	c + d
	a + c	b + d	a + b + c + d = N (Total Collection)



Different Scales, Different Agreement With Experts



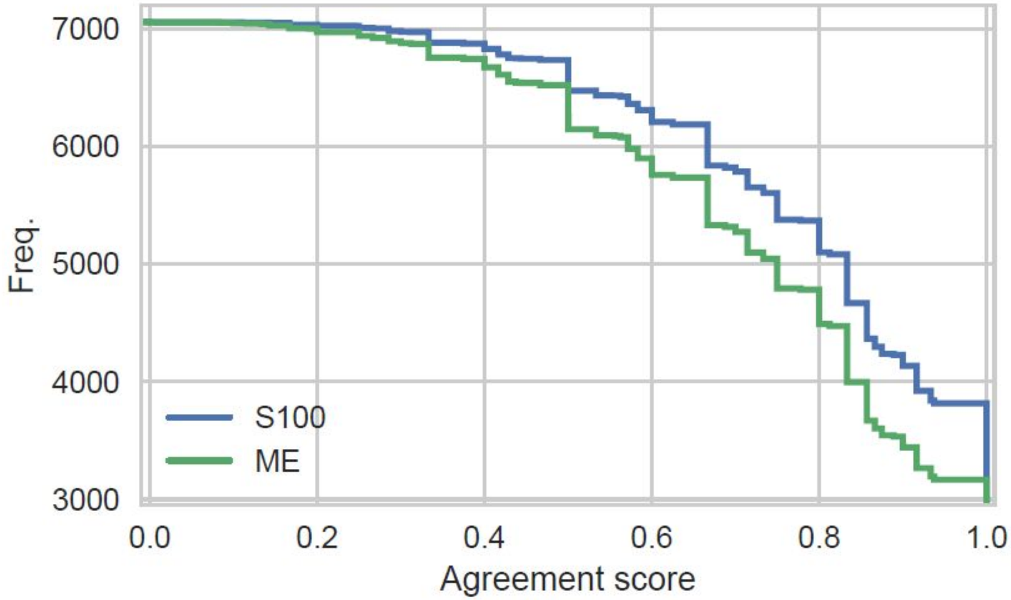
S100



ME

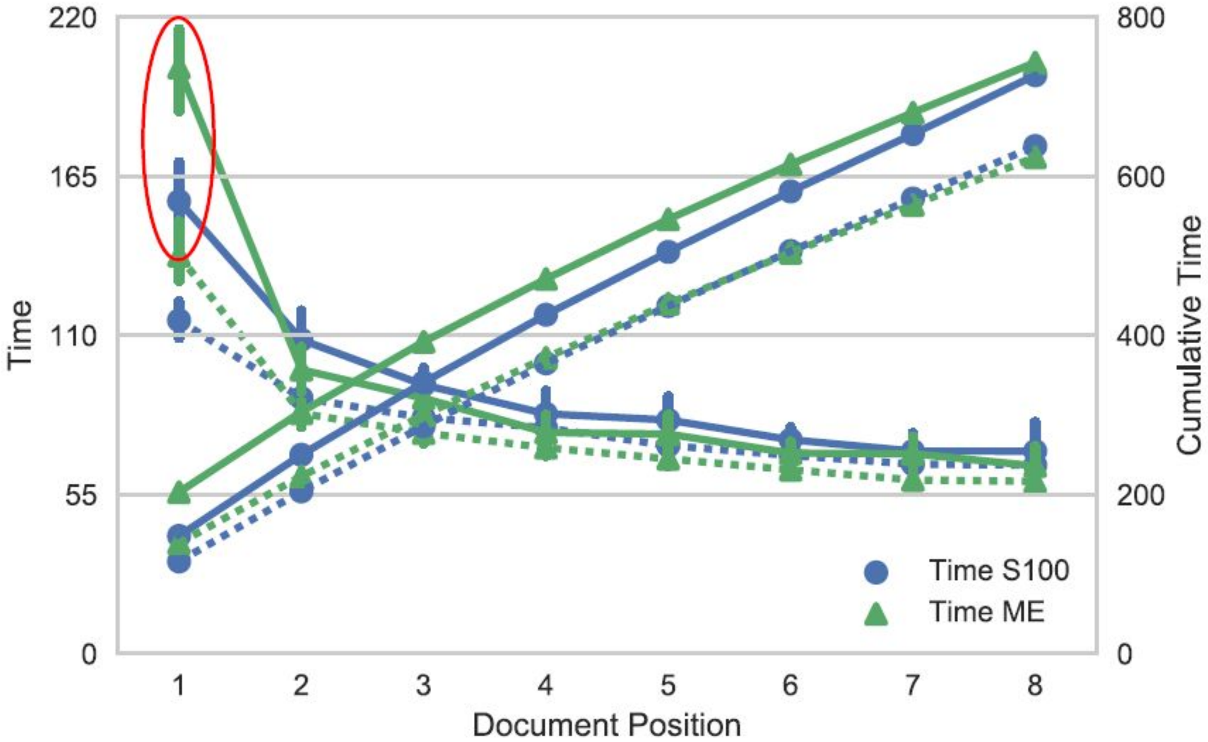
behavior is similar, but agreement is not!

Different Scales, Different Agreement With Experts



even for individual workers different scales have different agreement levels

Different Scales, Different Human Effort



some scales are more intuitive / familiar / easy to use for workers → and produce also higher agreement levels

Conclusions / Ongoing

- different scales behave very differently for evaluation
- different scales are perceived very differently from users

- (ongoing) predict churning of crowd platforms users using their agreement level compute in real time using a (deep) neural network

Relevance to Spotify Research

- use (our) metric to compute agreement between users of spotify
  (on listening behavior, genres, etc.)
- embed agreement into Machine Learning / Retrieval / moderation / etc. evaluation measures
- predict user churning / unsubscribing from premium / etc.
  leveraging user agreement
- use a subset of data with high\less agreement to train ML systems with less data, and thus use more demanding system configurations.
- use the appropriate relevance scale to compute metrics in Spotify ecosystem
- leverage agreement for automatic content moderation for songs / playlist or when crowdsourcing tasks

Resources

- HCOMP: https://aaai.org/ocs/index.php/HCOMP/HCOMP17/paper/viewFile/15927/15258.
- ICTIR: https://dl.acm.org/citation.cfm?id=3121060
- SIGIR - S100: https://dl.acm.org/citation.cfm?id=3210052
- SIGIR - DC: https://dl.acm.org/citation.cfm?id=3210229
- WWW: https://dl.acm.org/citation.cfm?id=3291035

- (submitted) - CIKM - transformation of relevance scales
- (submitted) - TKDE - role of abandonment (and agreement) in Crowdsourcing
- (finalizing) - TOIS - the effect of relevance scales for (crowdsourcing) IR evaluation

Bias

Bias
in (IR) Evaluation

Title

N Topics / Queries

$$
\begin{array}{cc}
 & \begin{array}{cccccc} \text{Topic}_1 & \text{Topic}_2 & \cdots & \text{Topic}_j & \cdots & \text{Topic}_N \end{array} \\
\begin{array}{c} \text{system}_1 \\ \text{system}_2 \\ \vdots \\ \vdots \\ \text{system}_i \\ \vdots \\ \vdots \\ \text{system}_M \end{array} &
\begin{bmatrix}
AP_{1,1} & AP_{1,2} & \cdots & AP_{1,j} & \cdots & AP_{1,N} \\
AP_{2,1} & AP_{2,2} & \cdots & AP_{2,j} & \cdots & AP_{2,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
AP_{i,1} & AP_{i,2} & \cdots & \boxed{AP_{i,j}} & \cdots & AP_{i,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
AP_{M,1} & AP_{M,2} & \cdots & AP_{M,j} & \cdots & AP_{M,N}
\end{bmatrix}
\end{array}
$$

Average Precision of system *i* for topic *j*

M Systems

Title

N Topics / Queries

$$
\begin{array}{ccc}
 & \begin{array}{cccccc} \text{Topic}_1 & \text{Topic}_2 & \cdots & \text{Topic}_j & \cdots & \text{Topic}_N \end{array} & \text{MAP} \\
\begin{array}{c} \text{system}_1 \\ \text{system}_2 \\ \vdots \\ \vdots \\ \text{system}_i \\ \vdots \\ \vdots \\ \text{system}_M \end{array} &
\begin{bmatrix}
AP_{1,1} & AP_{1,2} & \cdots & AP_{1,j} & \cdots & AP_{1,N} \\
AP_{2,1} & AP_{2,2} & \cdots & AP_{2,j} & \cdots & AP_{2,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
AP_{i,1} & AP_{i,2} & \cdots & \boxed{AP_{i,j}} & \cdots & AP_{i,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
AP_{M,1} & AP_{M,2} & \cdots & AP_{M,j} & \cdots & AP_{M,N}
\end{bmatrix} &
\begin{bmatrix} \text{MAP}_1 \\ \text{MAP}_2 \\ \vdots \\ \vdots \\ \text{MAP}_i \\ \vdots \\ \vdots \\ \text{MAP}_M \end{bmatrix} \\
\text{AAP} & \begin{bmatrix} AP_{1,1} & AP_{1,2} & \cdots & AP_{1,j} & \cdots & AP_{1,N} \end{bmatrix} &
\end{array}
$$

M Systems

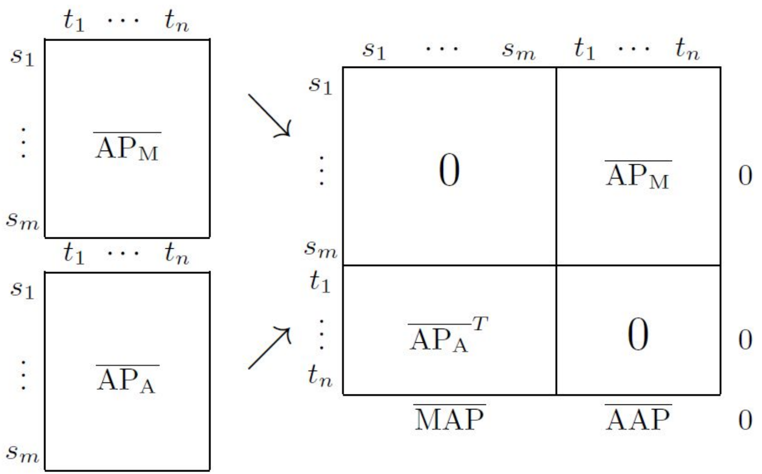
System Effectiveness

Topic Difficulty

Building a Graph

how much a system thinks a topic is difficult

how much a topic thinks a system is effective

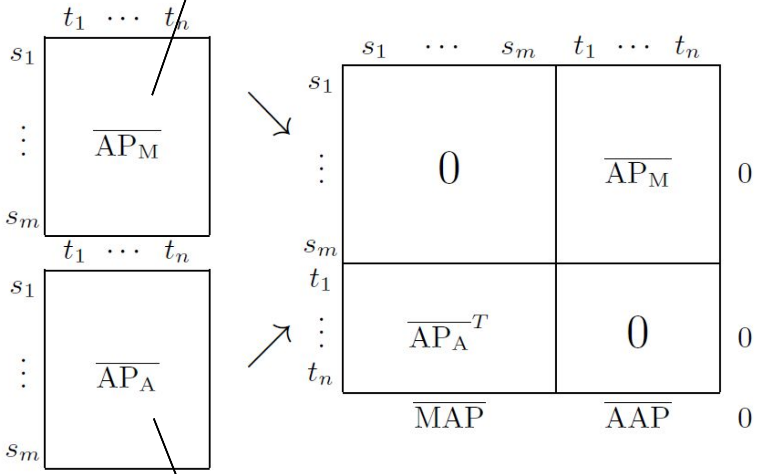


Building a Graph

how much a system thinks a topic is difficult

how much a topic thinks a system is effective

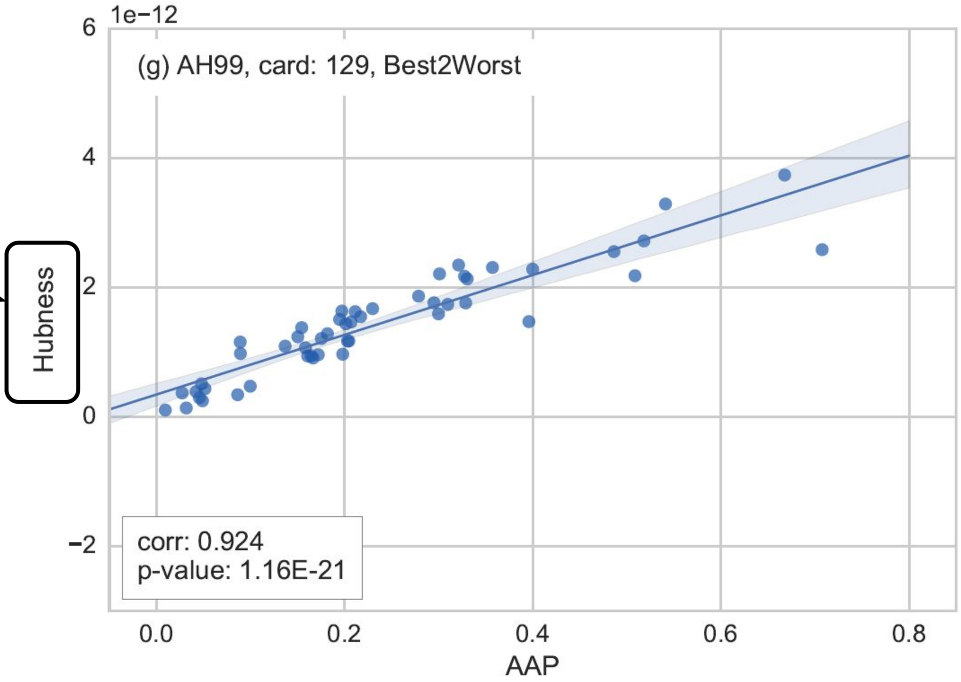
all systems are equally effective. removed system effectiveness bias



all topics have the same difficulty. removed topics difficulty bias

Results

ability of a topic to identify effective systems

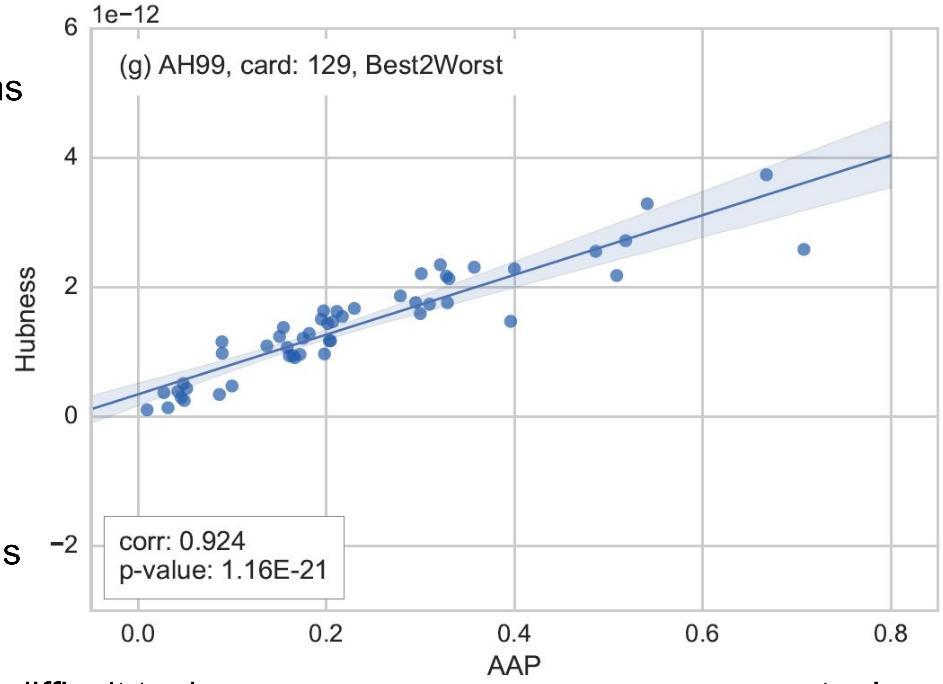


Results

•

most effective systems

least effective systems



difficult topics

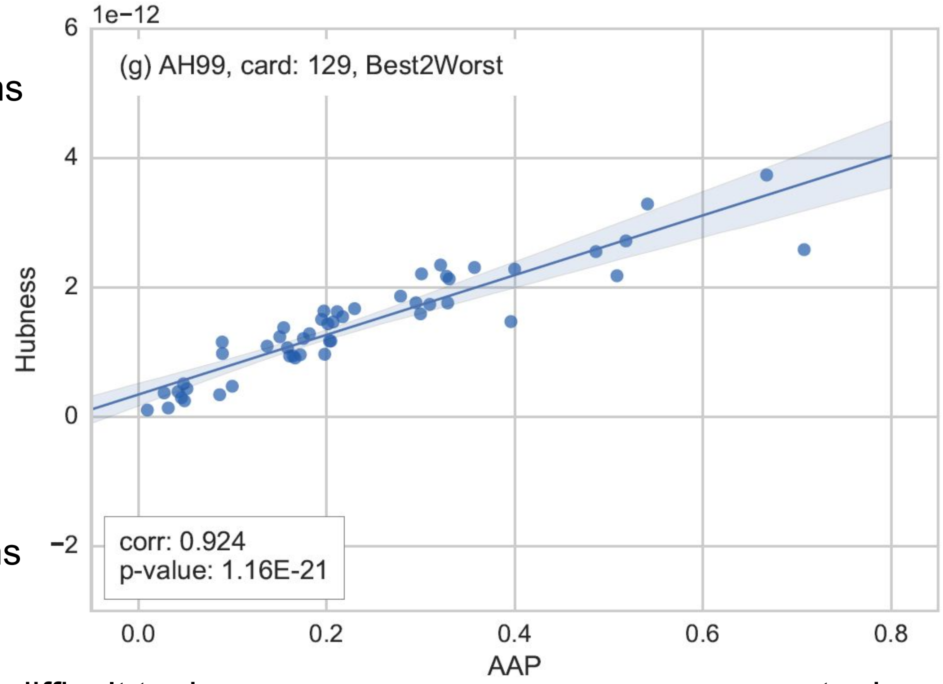
easy topics

Results

-

most effective systems

least effective systems



difficult topics ⟷ easy topics

easy topics are better in identify the final rank of retrieval systems.
→ issue in the evaluation process, → it can be leveraged to find more informative topics, → bias identification

Bias
in Scholarly publishing

Scholarly publishing / Readersourcing

- Scholarly publishing:
  - hard work
  - write + submit
  - Peer Review
  - accept, or goto(hard work)
- not the only possible model
  - open access
  - wikipedia like approaches
- peer review issues
  - slow, biased, subjective
  - unable to detect fraud
  - Page Rank being rejected from SIGIR1998...
  - …
- readersourcing : Crowdsourcing peer review
  - readers read papers and express opinions
  - a formalized model

Scholarly publishing / Readersourcing

N Papers

$$\begin{array}{c c}
 & \begin{array}{cccccc} \text{Item}_1 & \text{Item}_2 & \cdots & \text{Item}_j & \cdots & \text{Item}_N \end{array} \\
\begin{array}{c} \text{Assessor}_1 \\ \text{Assessor}_2 \\ \vdots \\ \vdots \\ \text{Assessor}_i \\ \vdots \\ \vdots \\ \text{Assessor}_M \end{array} &
\begin{bmatrix}
r_{1,1} & r_{1,2} & \cdots & r_{1,j} & \cdots & r_{1,N} \\
r_{2,1} & r_{2,2} & \cdots & r_{2,j} & \cdots & r_{2,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r_{i,1} & r_{i,2} & \cdots & \boxed{r_{i,j}} & \cdots & r_{i,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
r_{M,1} & r_{M,2} & \cdots & r_{M,j} & \cdots & r_{M,N}
\end{bmatrix}
\end{array}$$

$$\begin{array}{c c}
 & \begin{array}{cccccc} \text{Item}_1 & \text{Item}_2 & \cdots & \text{Item}_j & \cdots & \text{Item}_N \end{array} \\
\begin{array}{c} \text{Assessor}_1 \\ \text{Assessor}_2 \\ \vdots \\ \vdots \\ \text{Assessor}_i \\ \vdots \\ \vdots \\ \text{Assessor}_M \end{array} &
\begin{bmatrix}
G_{1,1} & G_{1,2} & \cdots & G_{1,j} & \cdots & G_{1,N} \\
G_{2,1} & G_{2,2} & \cdots & G_{2,j} & \cdots & G_{2,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
G_{i,1} & G_{i,2} & \cdots & \boxed{G_{i,j}} & \cdots & G_{i,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
G_{M,1} & G_{M,2} & \cdots & G_{M,j} & \cdots & G_{M,N}
\end{bmatrix}
\end{array}$$

M Readers

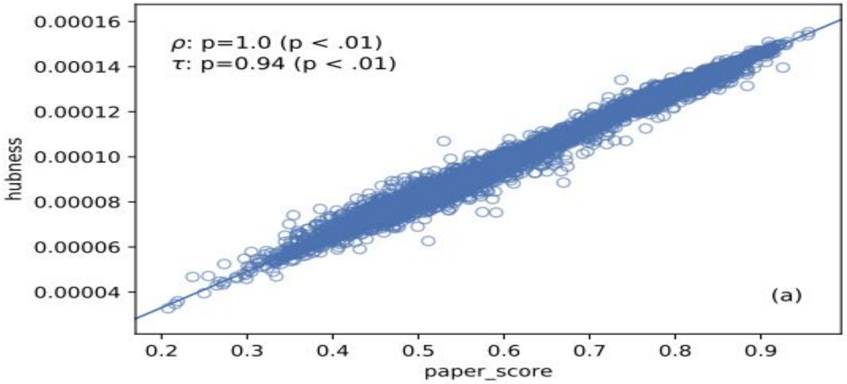
Rating

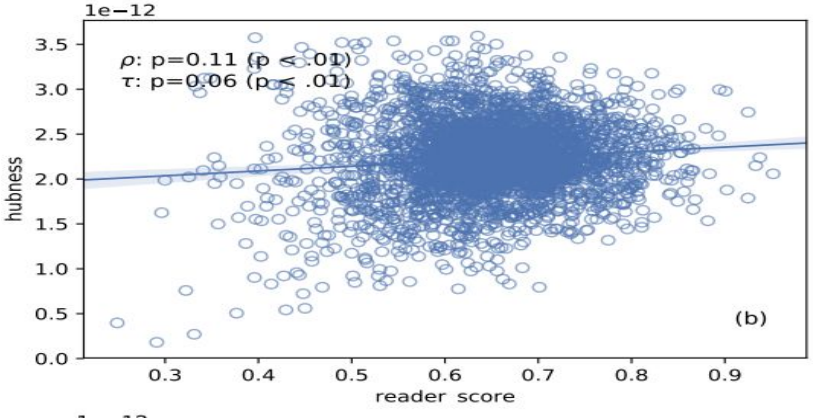
Goodness
of the Rating

Removing Bias

- reader bias:
  - all readers vote with the same median
  - all reader vote with the same judgment scale
  - all readers vote with the same quality
  - (...)
- paper bias:
  - all papers have the same average judgment
  - all papers are judged using the same perceived scale
  - all papers have the same quality
  - (...)
- mixed approaches are possible

Results / Potential





- see if the model has some kind of bias
  - high quality papers are judged by high quality readers
  - readers that overestimate true judgment have an high quality
  - …
- remove bias from the model
  - are the best papers still best papers after removing the bias?
  - ...

Relevance to Spotify

- remove / identify various user biases
  - listen lot of songs
  - free / premium
  - skips lots of songs
  - ...
- see if recommender system evaluation is biased for some users and correct the bias
- identify playlists that are biased in the recommendation phase:
  - playlists that are always recommended at lunch time
- force / embed bias in playlists commendations
  - If I like to have dinner listening to piano songs then ...

Resources

- ECIR: http://dx.doi.org/10.1007/978-3-319-56608-5_55.

- (Submitted) BIRNDL: bias in scholarly publishing models
- (finalizing) Experiments on bias in scholarly publishing for IIR
- (in progress) Experiments on bias in scholarly publishing for a Journal paper

Economical Evaluation of (IR) Systems

Economical Evaluation of IR systems

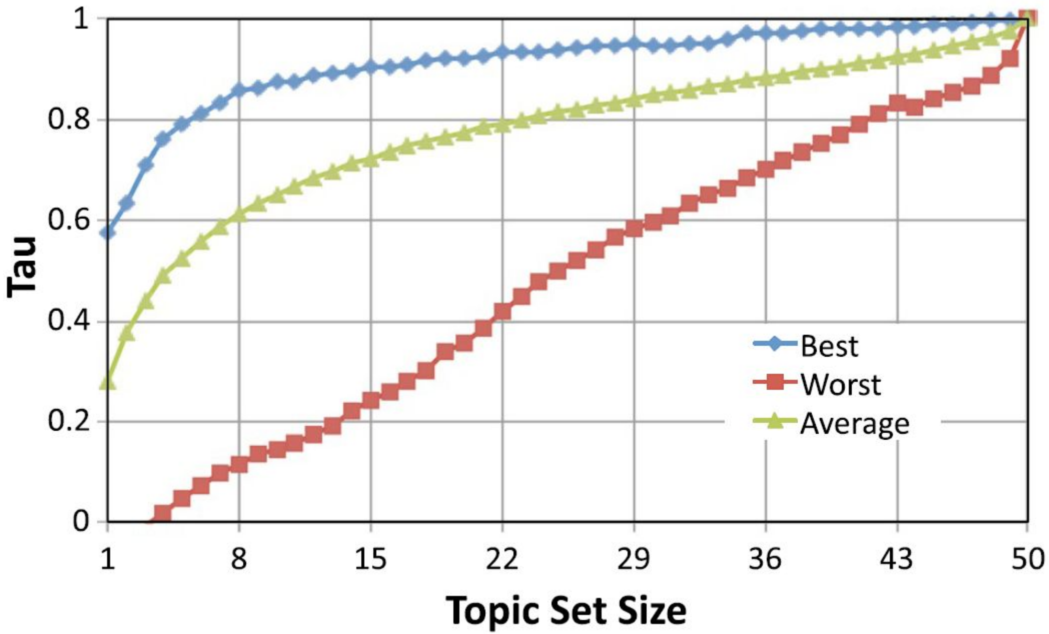
N Topics / Queries

$$
\begin{array}{c}
\\
\text{system}_1 \\
\text{system}_2 \\
\vdots \\
\vdots \\
\text{system}_i \\
\vdots \\
\vdots \\
\text{system}_M
\end{array}
\begin{array}{c}
\begin{array}{ccccccc}
\text{Topic}_1 & \text{Topic}_2 & \cdots & \text{Topic}_j & \cdots & \text{Topic}_N
\end{array} \\
\begin{bmatrix}
AP_{1,1} & AP_{1,2} & \cdots & AP_{1,j} & \cdots & AP_{1,N} \\
AP_{2,1} & AP_{2,2} & \cdots & AP_{2,j} & \cdots & AP_{2,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
AP_{i,1} & AP_{i,2} & \cdots & \boxed{AP_{i,j}} & \cdots & AP_{i,N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
AP_{M,1} & AP_{M,2} & \cdots & AP_{M,j} & \cdots & AP_{M,N}
\end{bmatrix}
\end{array}
$$

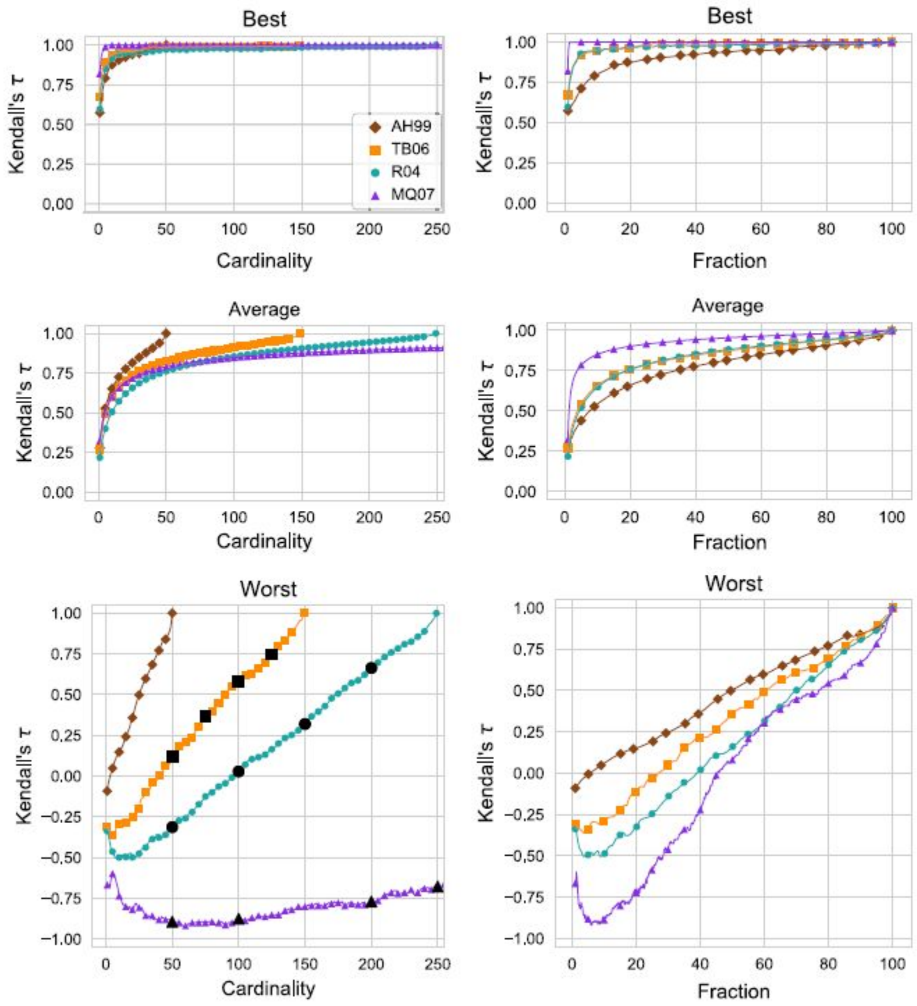
M Systems

Average Precision of system *i* for topic *j*

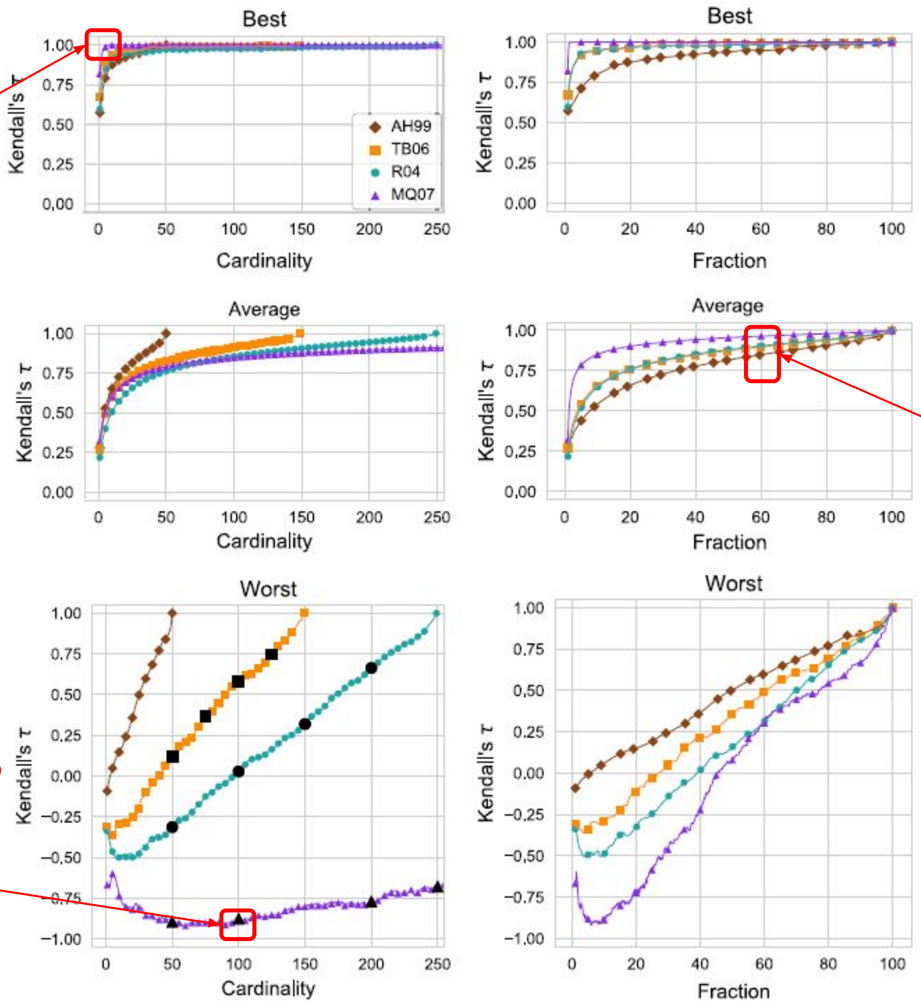
What About using Few Topics?



Few Topics


Best
Average
Worst
Kendall's τ
Cardinality
Fraction
AH99
TB06
R04
MQ07

Few Topics



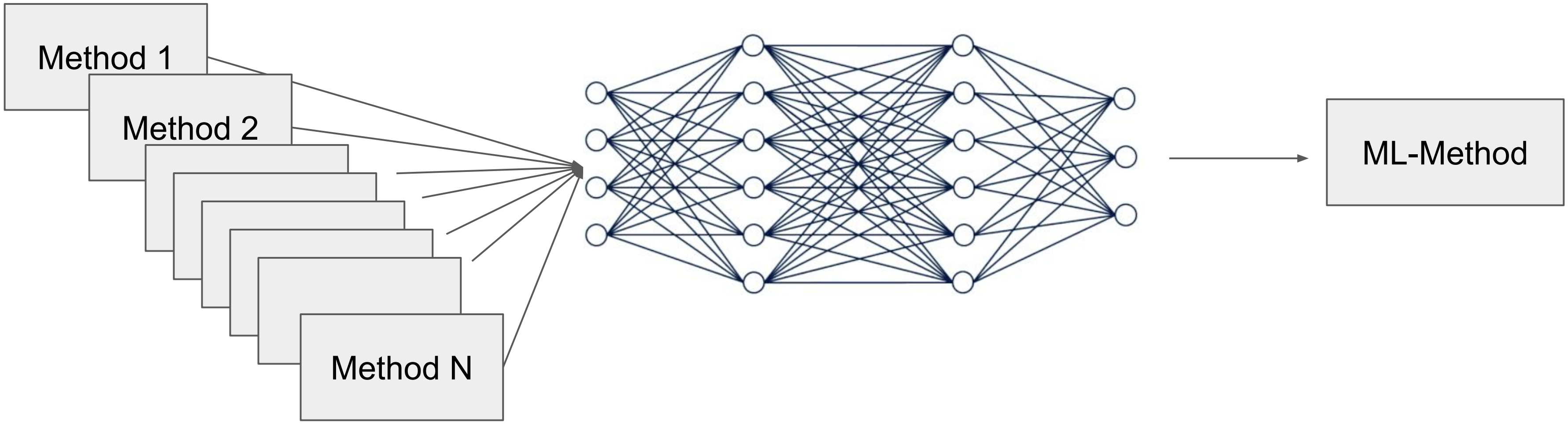
correlation ≈ 1 with 10% of the dataset

random guessing is not a good strategy

correlation ≈ -1 with 10-20% of the dataset

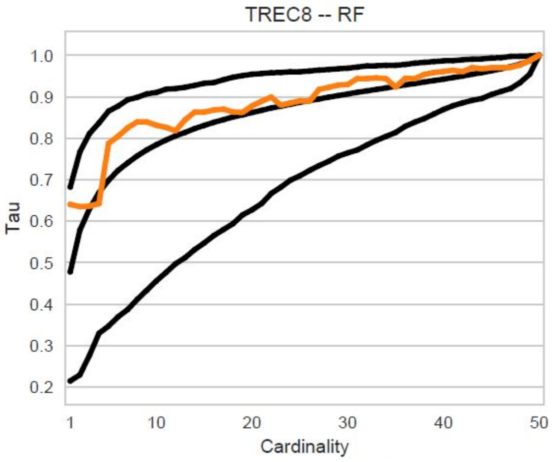
A practical approach to Few Topics

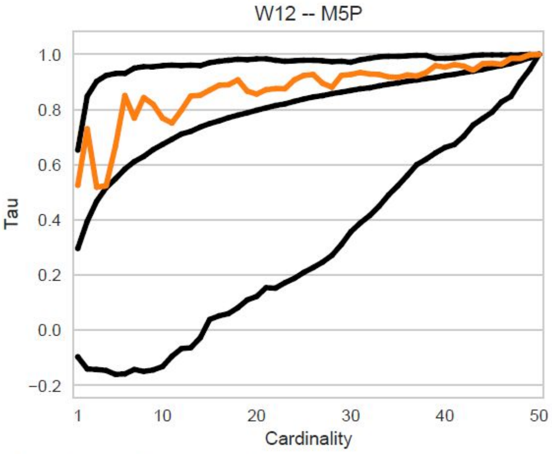
- Evaluation Without Relevance Judgments (EEWRJ)
  - evaluate retrieval systems based on their ranked lists of documents
  - no human effort ≈ no cost!
- strategy based on individual EEWRJ methods + ML combinations of them



A practical approach to Few Topics

- Evaluation Without Relevance Judgments (EEWRJ)
  - evaluate retrieval systems based on their ranked lists of documents
  - no human effort ≈ no cost!
- strategy based on individual EEWRJ methods + ML combinations of them





Relevance to Spotify

- train Spotify recommender system using (much) less data
- identify the subset of good (= more informative) users / playlists / etc.
- use fusion methods and ML combinations of weak EEWRJ-like signals
    - Spotify Search
    - ensembles / cascade of Recommender systems
    - …
- identify (using the outcome of the first part of the talk) the subset of users / playlists / etc. with less / more bias / agreement / both ...

Resources

- IRJ: https://rdcu.be/bFuZq
- SIGIR2018: https://dl.acm.org/citation.cfm?id=3210108
- JDIQ : https://dl.acm.org/citation.cfm?id=3239573
- JDIQ : https://dl.acm.org/citation.cfm?id=3241064

- (rebuttal) - IPM - EEWRJ overview and combination of approaches
- many ongoing experiments:
  - Neural Networks with custom loss to build a good model
  - GANs and Genetic Algorithms to generate fake data for data augmentation
  - Crowdsourcing query variations for retrieval fusion and for data augmentation
  - Multi Armed Bandits to select topics / Documents using “no relevance judgments” signals
  - ...

Thank you!