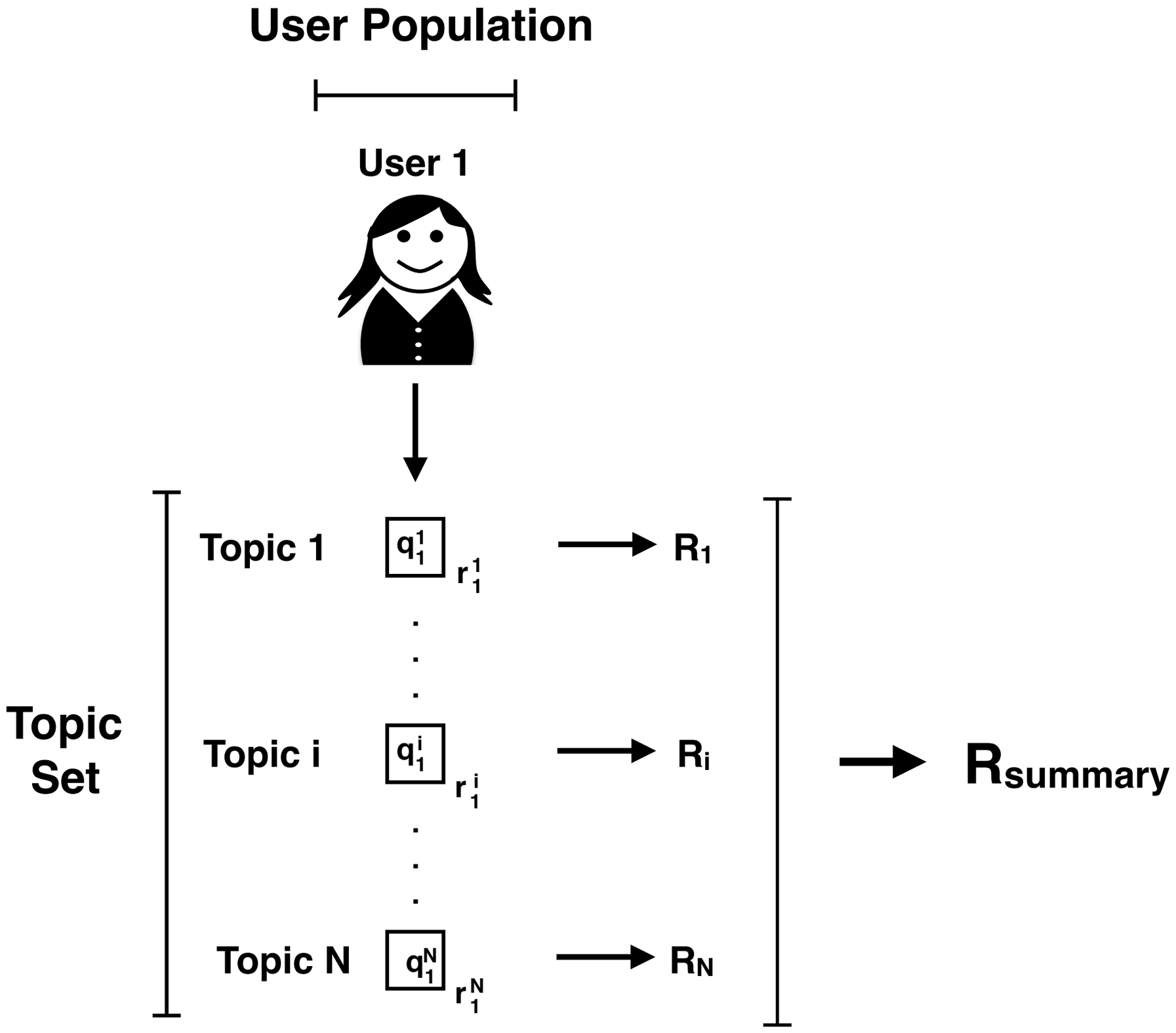
Query Variations and their Effect on Comparing Information Retrieval Systems

Guido Zuccon[1], Joao Palotti[2], Allan Hanbury[2]

[1] Queensland University of Technology, Brisbane, Australia
[2] Vienna Institute of Technology, Vienna, Austria

TREC-style Evaluation



An example of query variations

What would be your query to Google if you have this on your skin?

q: “Crater type bite mark”

Health Consumers

q: “Ring wound below wrinkled eyelid”

Health Professionals

q: “skin lesions”

From the CLEF eHealth 2015 collection

Query variations affect retrieval?

q: “Crater type bite mark”

What Bit Me? Mystery Bug Bites Solved | SafeBee
www.safebee.com › Outdoors ▾
Jun 16, 2015 - What it's **like**: You may feel a sharp **sting** when you're **bitten** or nothing at all. ... The brown recluse has a violin-shaped **mark** on its back that isn't ... six weeks to go away, and the **bite** can leave a large **crater** and scarring.

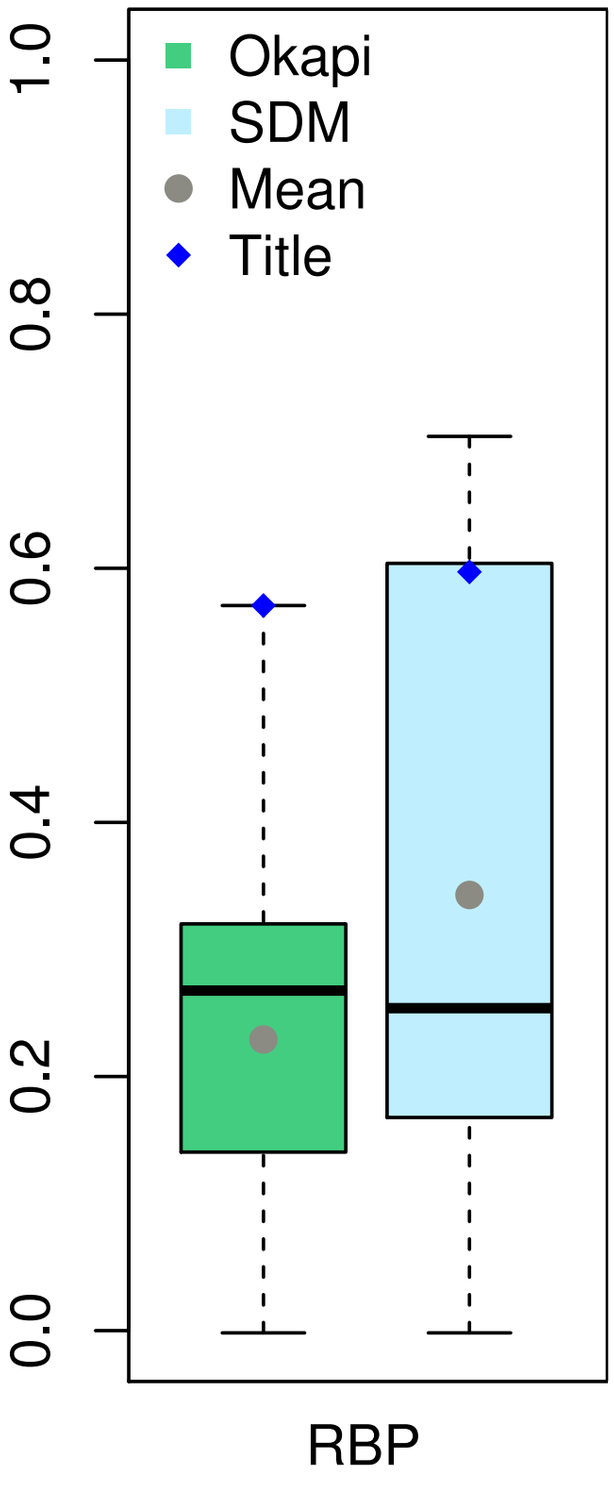
q: “Ring wound below wrinkled eyelid”

Eyelid Lift / Correction | Face - Plastic Surgery - Klinik am Ring
plastic-surgery.klinik-am-**ring**.com/index.php/.../augenlidkorrekturen.ht... ▾
An **eyelid** correction is always performed **under** local anesthesia. ... As long as the **wound** is closed with a hair-thin thread and covered with a thin ... or peeling techniques, the skin and thus the depth of **wrinkles** are sustainably improved.

q: “skin lesions”

Skin Cancer Symptoms: Pictures of Skin Cancer and ...
www.webmd.com/melanoma-**skin**.../slideshow-**skin-lesions**-and-cancer ▾
The Warning Signs of **Skin** Cancer. **Skin** cancers -- including melanoma, basal cell carcinoma, and squamous cell carcinoma -- often start as changes to your **skin**. They can be new growths or precancerous **lesions** -- changes that are not cancer but could become cancer over time.

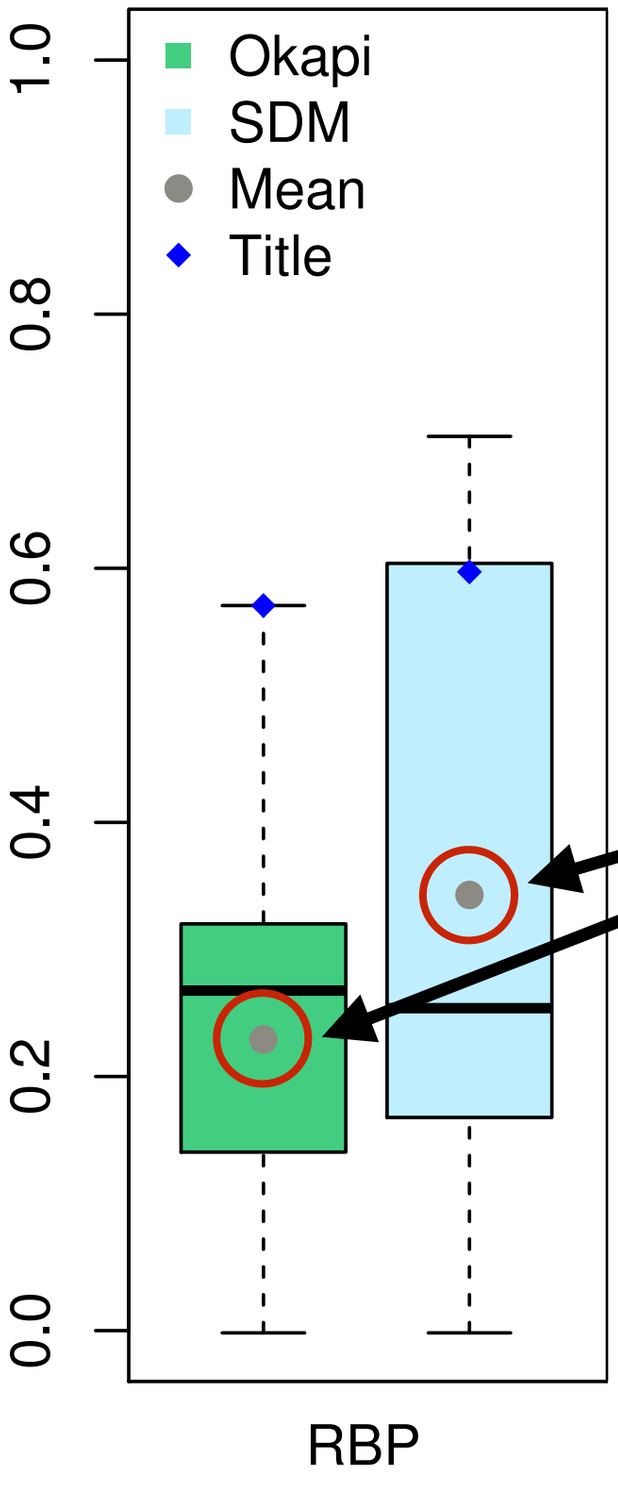
Query variations & evaluation



"the range of scores characterizing effectiveness for a single system arising from query variations is comparable or greater than the range of scores arising from variation among systems using only a single query per topic"

Bailey et al, *User Variability and IR System*

Query variations & evaluation

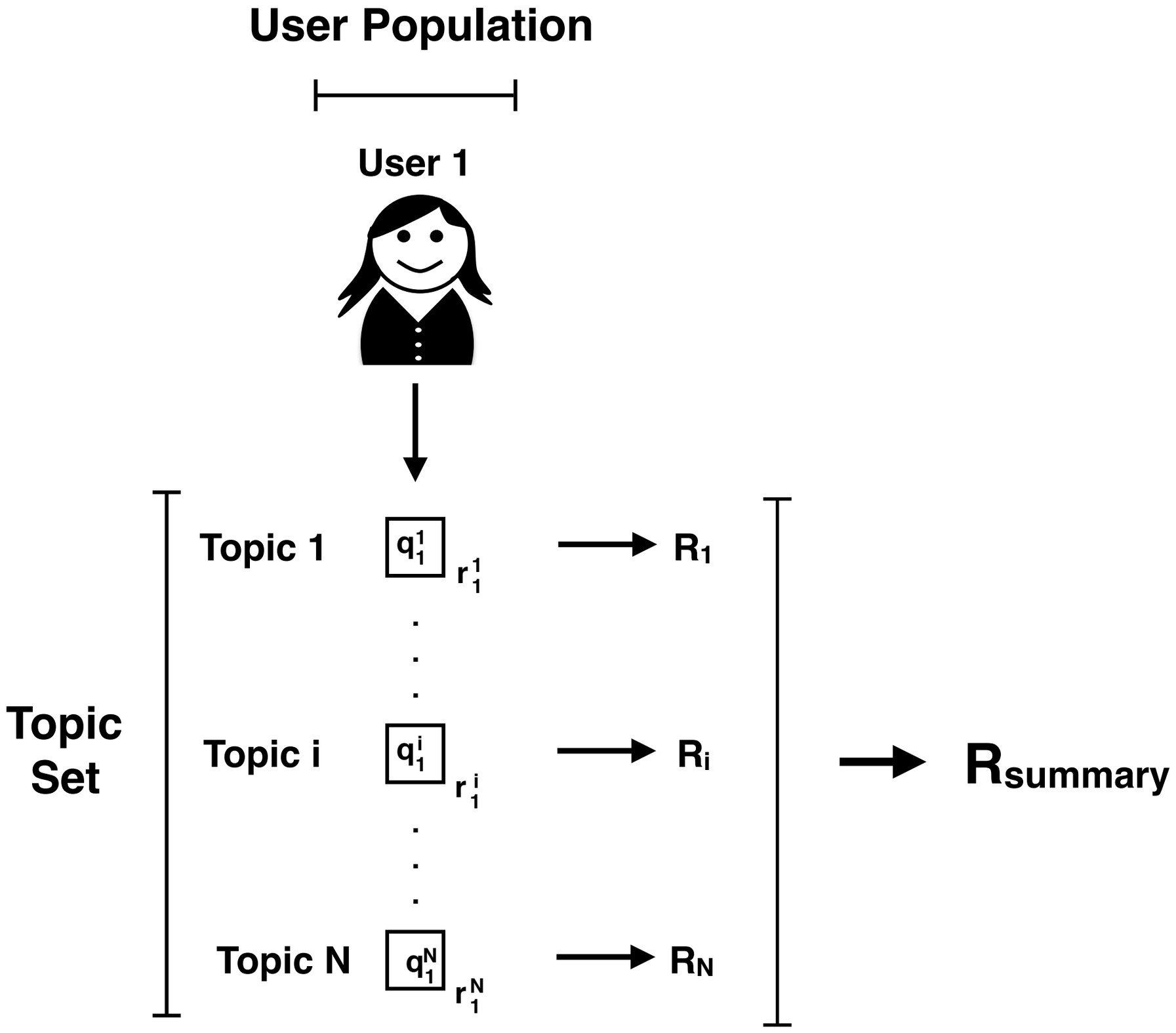


Use the mean effectiveness over query variations?

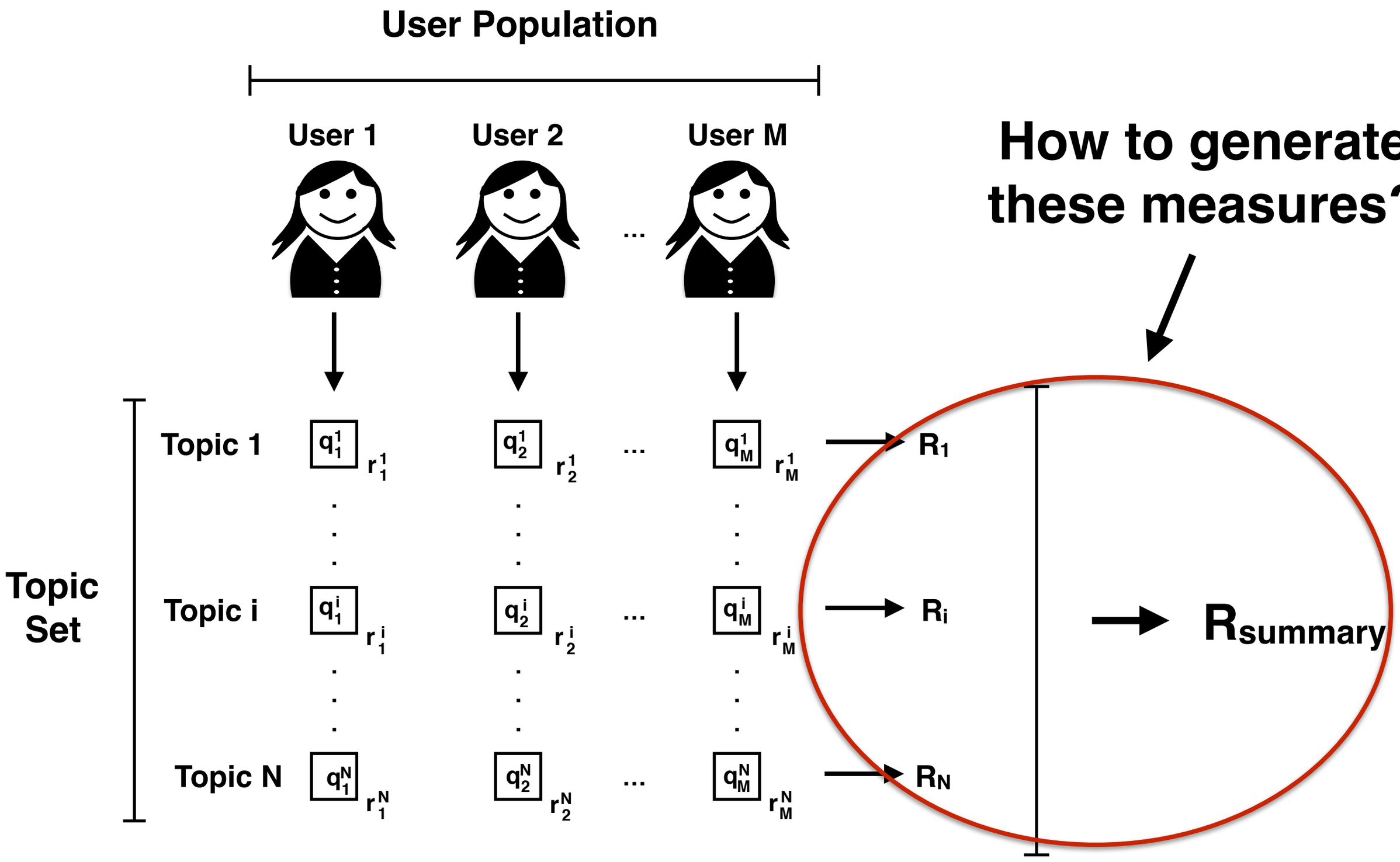
"the range of scores characterizing effectiveness for a single system arising from query variations is comparable or greater than the range of scores arising from variation among systems using only a single query per topic"

Bailey et al, *User Variability and IR System*

TREC-style Evaluation



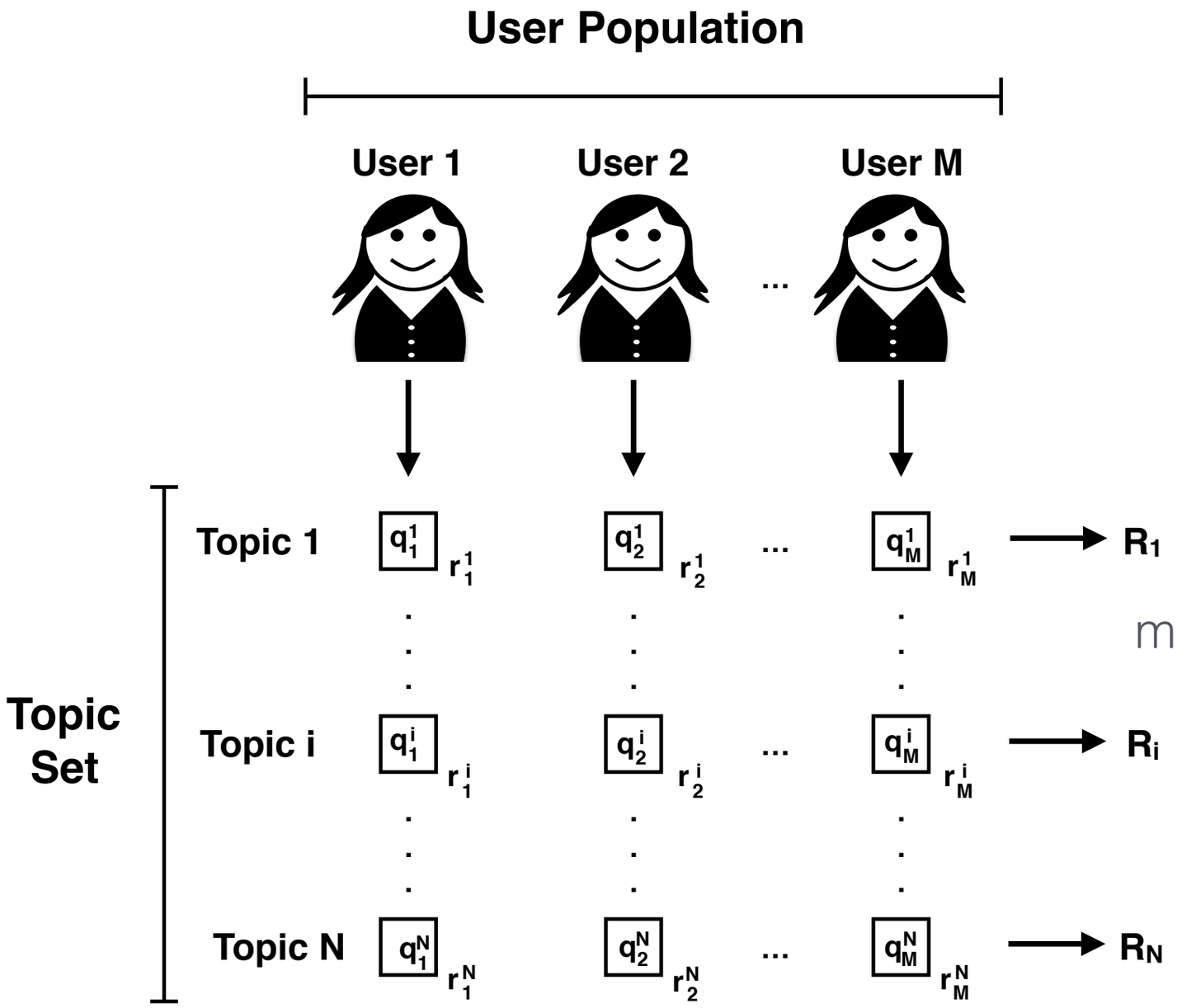
The situation in practice



Modelling query variations

- Consider **queries for a topic as dependent**, i.e. provide a topic-based evaluation, rather than query-based
- For a topic, consider **not just the mean effectiveness** over the queries, but also effectiveness **variance** across queries
- Allows to answers questions like:
  - is the system good on all queries for a topic?
  - are there specific topics where the system is good?
  - are there specific queries for which the system is good, but the system is not good for other queries for that same topic?
  - is a system more “stable” than another system?

Modelling query variations



The effectiveness of a system is:

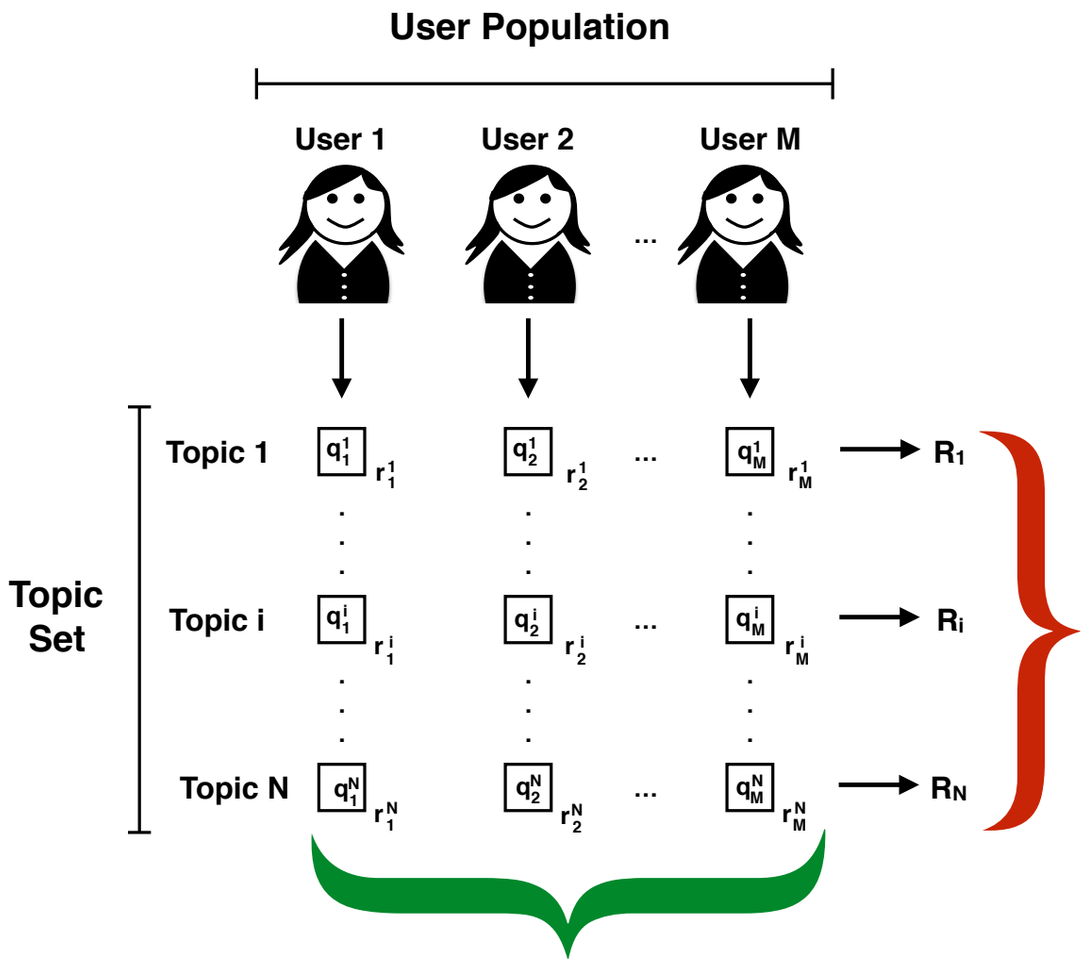
$$\mu - \alpha\sigma^2$$

mean effectiveness over topics

effectiveness variance over topics

risk preference parameter

Modelling query variations

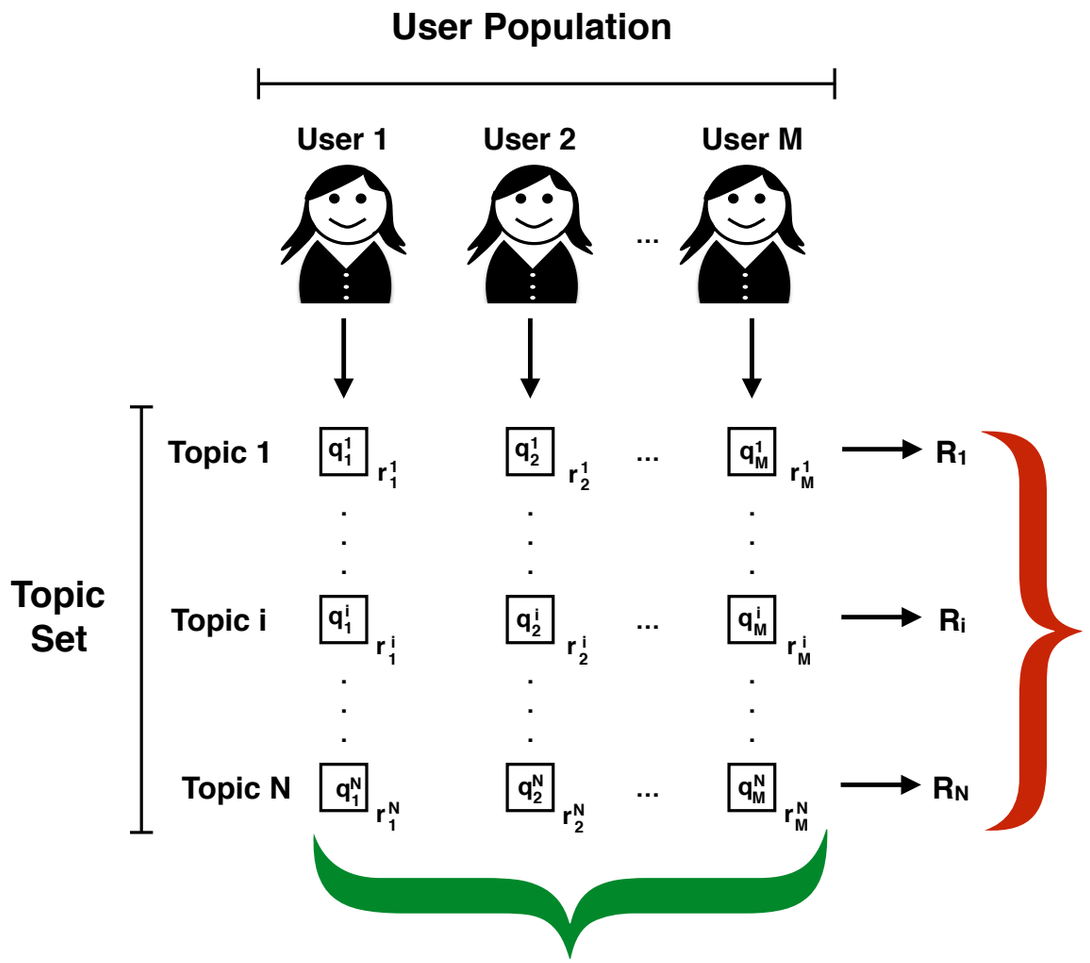


$$\mu - \alpha\sigma^2$$

$$\mu = \sum_{i=1}^{N} E[w_i R_i] = \sum_{i=1}^{N} w_i \cdot \left(\sum_{k=1}^{M} r_i^k p(r_i^k) \right)$$

w_i: importance of a topic for our evaluation (usually w_i=1 for all topics)

Modelling query variations



$$\mu - \alpha\sigma^2$$

$$\sigma^2 = \sum_{i=j=1}^{N} w_i^2 var(R_i) + \sum_{i \neq j}^{N} w_i w_j cov(R_i, R_j)$$

variance within a topic

covariance across topics

Comparing Systems

System A better than system B iff.

$$\mu_A - \alpha\sigma_A^2 > \mu_B - \alpha\sigma_B^2$$

Comparing Systems

System A better than system B iff.

$$\mu_A - \alpha\sigma_A^2 > \mu_B - \alpha\sigma_B^2$$

- **same mean:** *variance* of A *lower* than that of B, i.e. A more stable

- **mean of A lower than that of B**: A may still be better if *difference between variances* is more than the difference between means, i.e. A much more stable than B. Alpha controls the influence of this

Comparing Systems

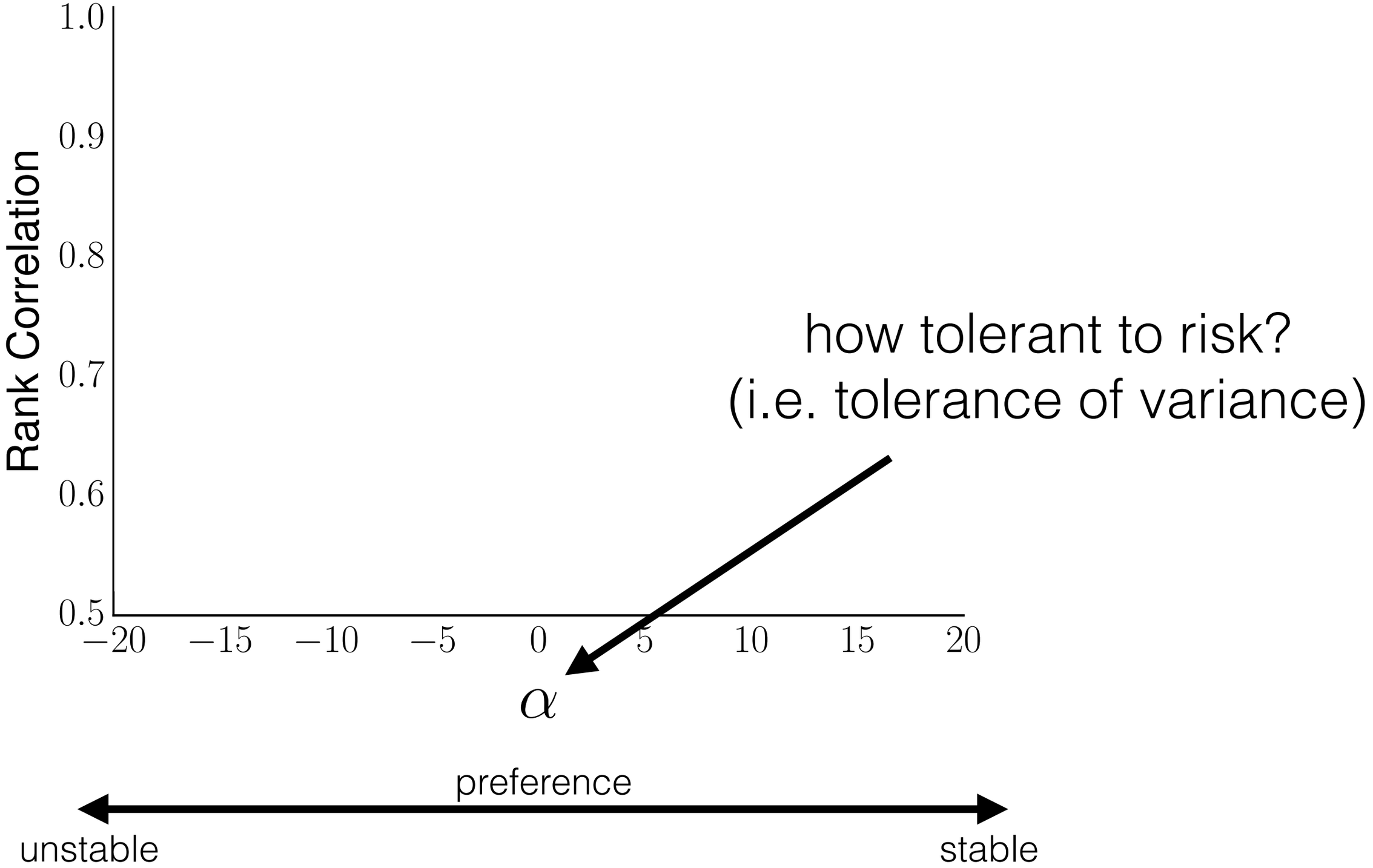
System A better than system B iff.

$$\mu_A - \alpha\sigma_A^2 > \mu_B - \alpha\sigma_B^2$$

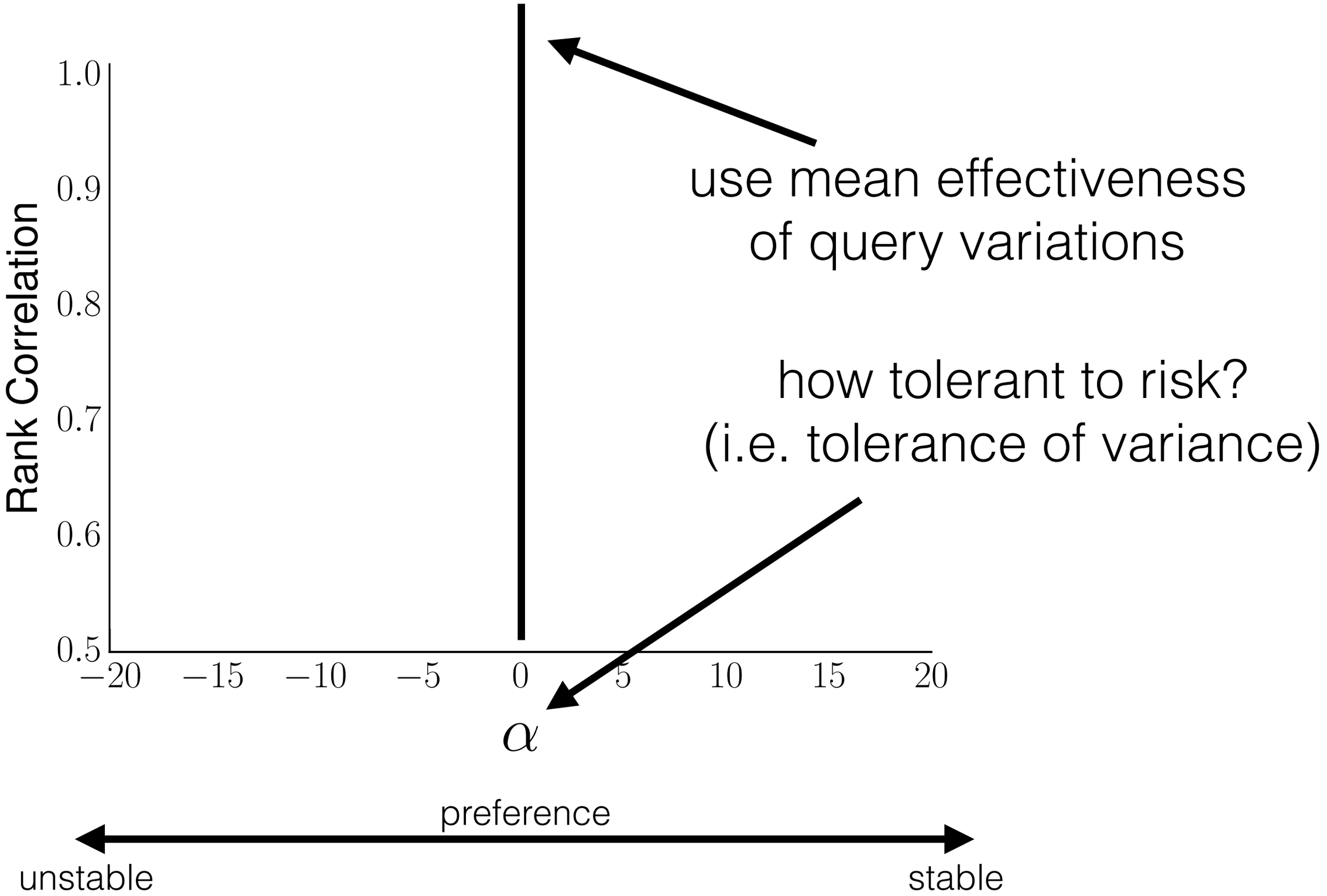
- **same mean, same variance across topics:** A has *lower covariance* across topics than B

- **no risk preference** (alpha=0): only compare mean effectiveness

Accounting for variations

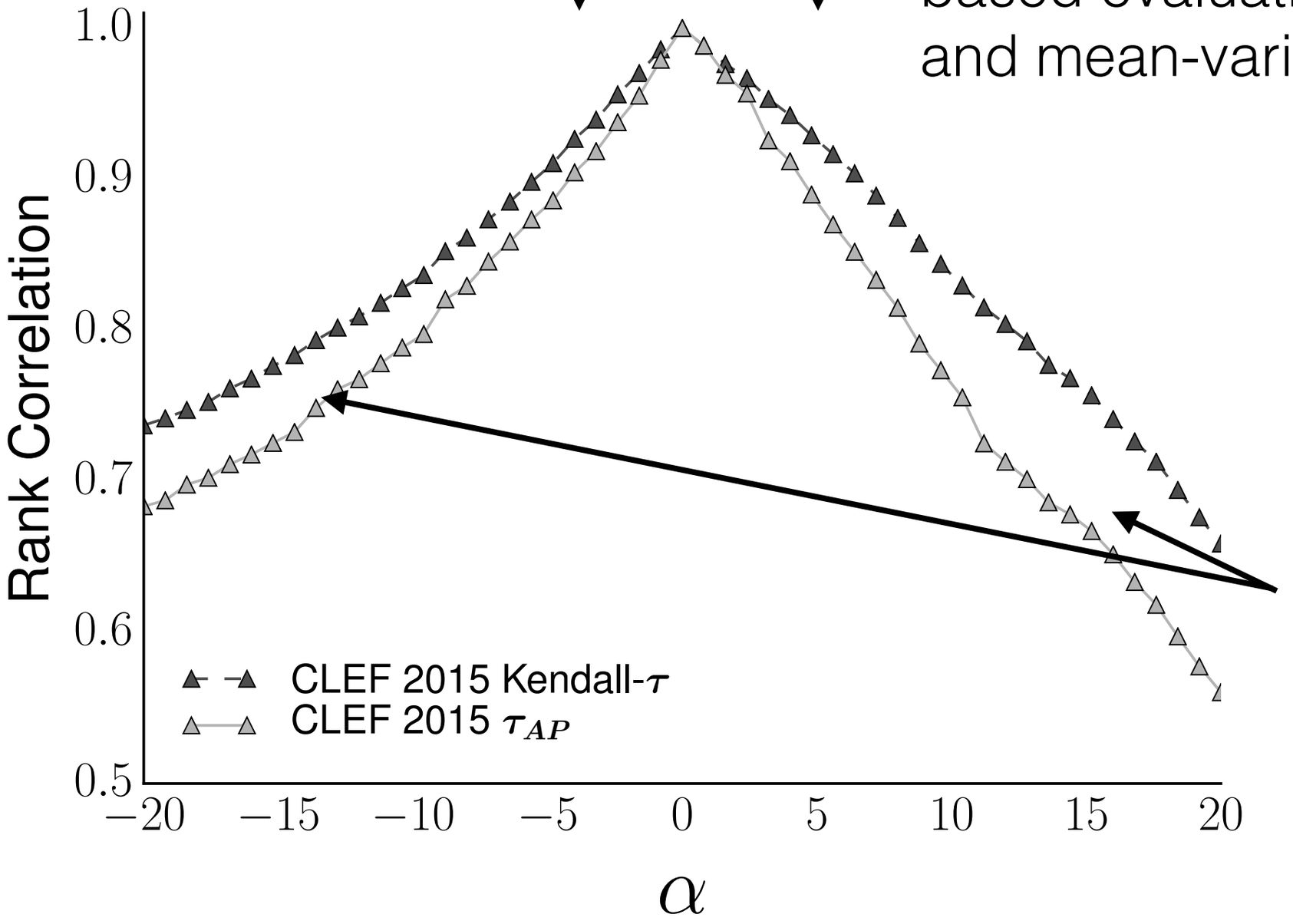


Accounting for variations



Accounting for variations

less correlation between mean-based evaluation (TREC-style) and mean-variance evaluation



variance aversion less correlated with mean-based evaluation (TREC-style)

preference

unstable

stable

Also in the paper…

- **specialisations** of the framework for
    - ***intra-topic evaluation***: query variations, but for a single topic (variance generated by multiple queries)
    - ***inter-topic evaluation***: one query only, for multiple topics (variance generated by multiple topics)
- more **simulations & experiments**, also for specialisations, including more collections
- pilot **user experiment** showing alpha estimates for users (risk preference)
- extended **discussion of key framework components** and their influence on system comparison

Summary

- Users queries for a same topic vary greatly, also in terms of effectiveness. We currently ignore this in TREC-style evaluations

- Unclear **how query variations should be accounted for in the evaluation**

- **Mean effectiveness** over variations hide important properties of systems, e.g. stability over query variations

- We contributed an **evaluation framework for query variations** based on mean-variance analysis

- New evaluation framework leads to **different observations** wrt system effectiveness, but

    - how much variation in effectiveness (risk) users are willing to take?

    - do users prefer high yielding unstable systems, or lower yielding, conservative systems? Valid estimates of alpha?

Advertisement

CLEF 2017 eHealth IR Task

https://sites.google.com/site/clefehealth2017/task-3

- **Query variations**
- **Topicality**, **understandability** and **reliability** assessments
- User **interactions**
- based on **Clueweb12**

Advertisement

Joao is looking for an internship

joaopalotti@gmail.com

- domain specific search
- learning to rank
- evaluation