UNIVERSITY OF AMSTERDAM

Using users
for online evaluation and learning

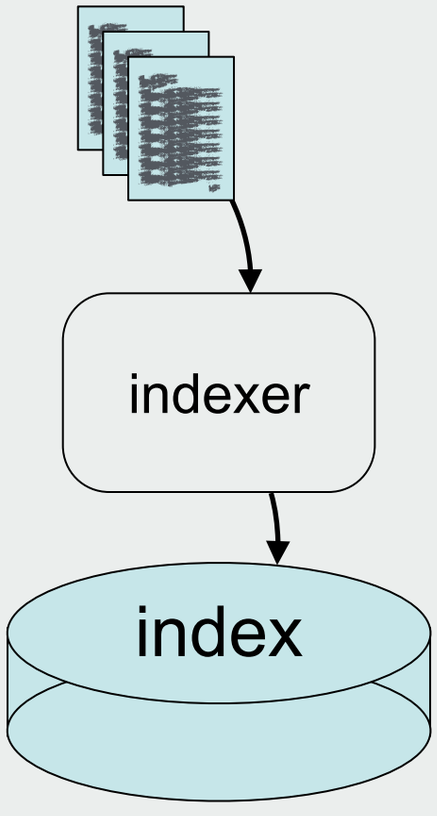
Anne Schuth
anne.schuth@uva.nl
ILPS, University of Amsterdam

Recommender Meetup, December 8, 2014

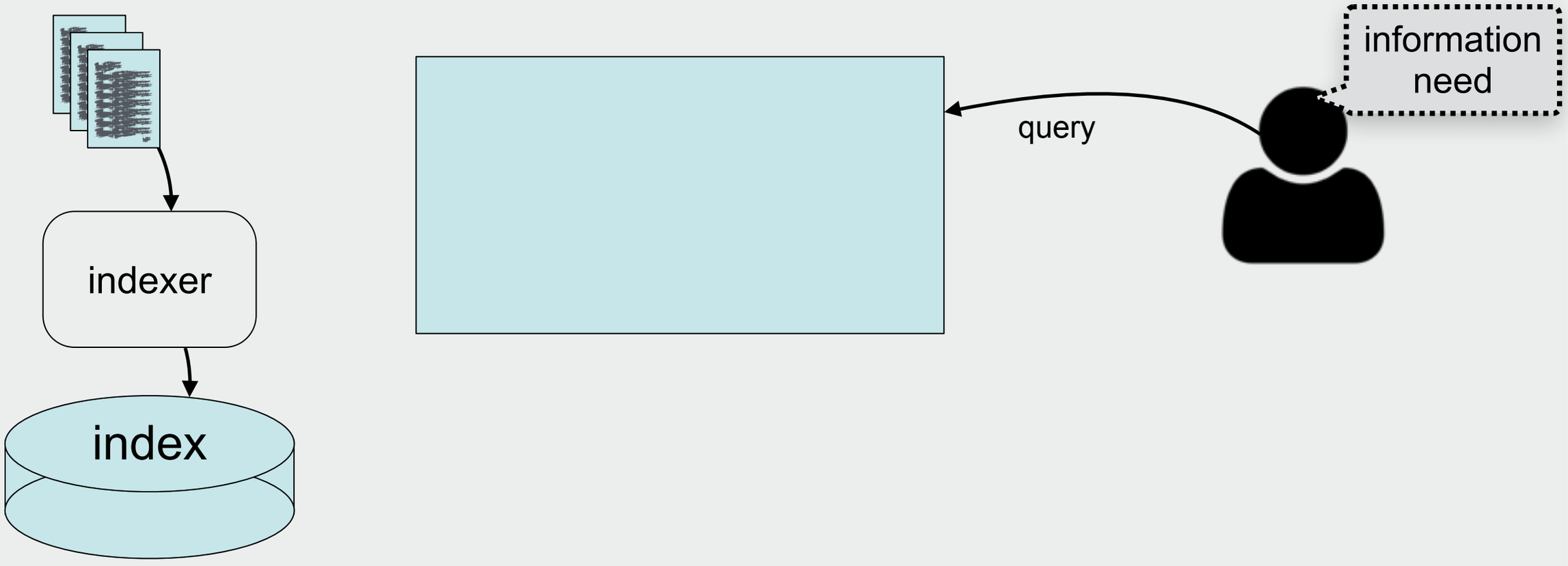
Information Retrieval

Information Retrieval

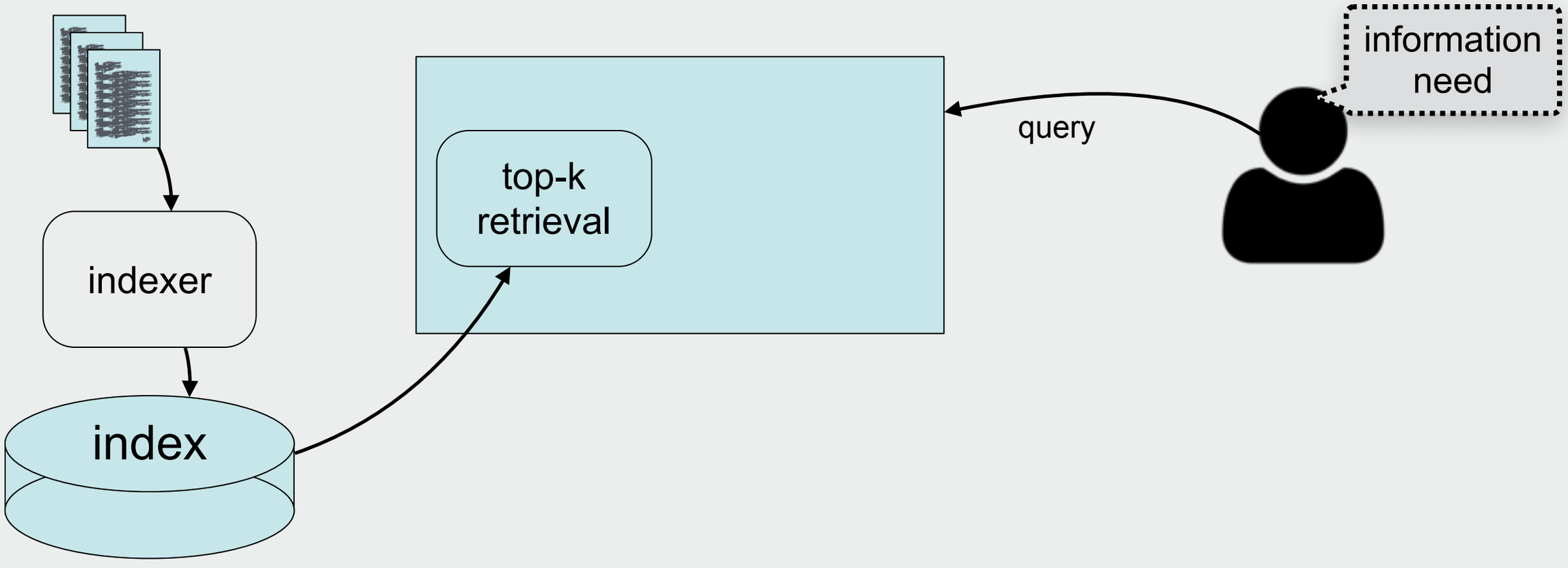
Information Retrieval



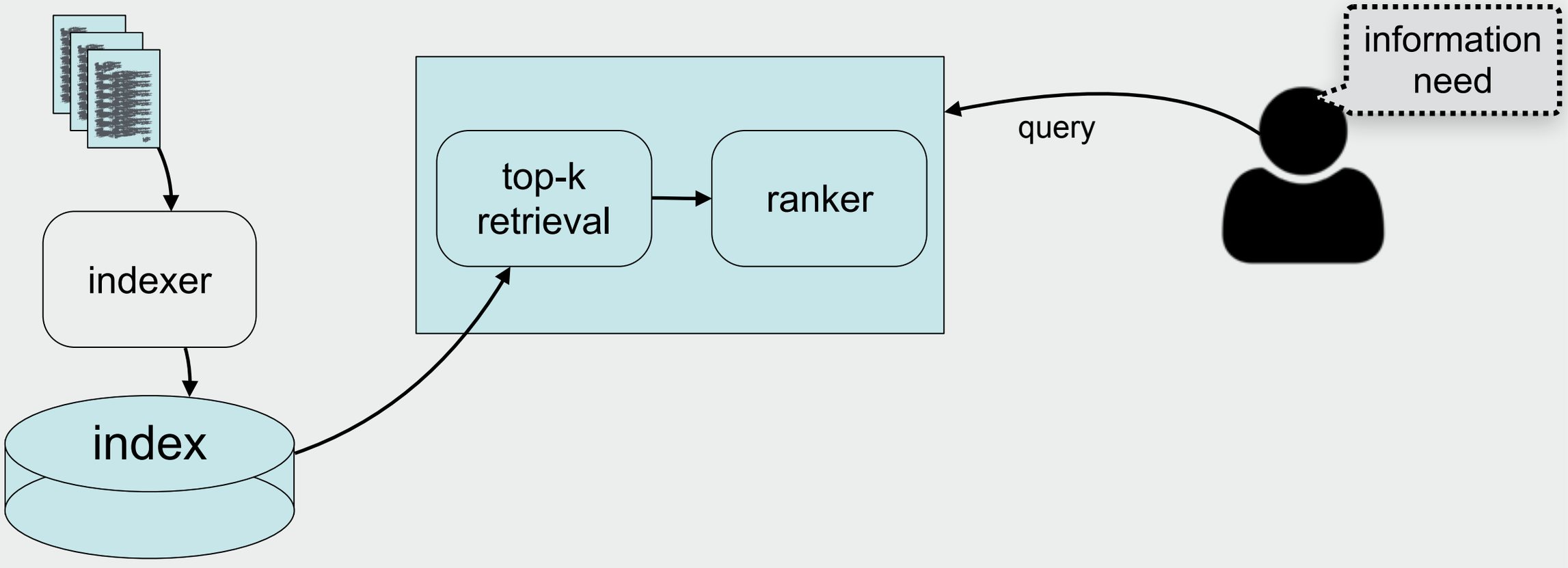
Information Retrieval



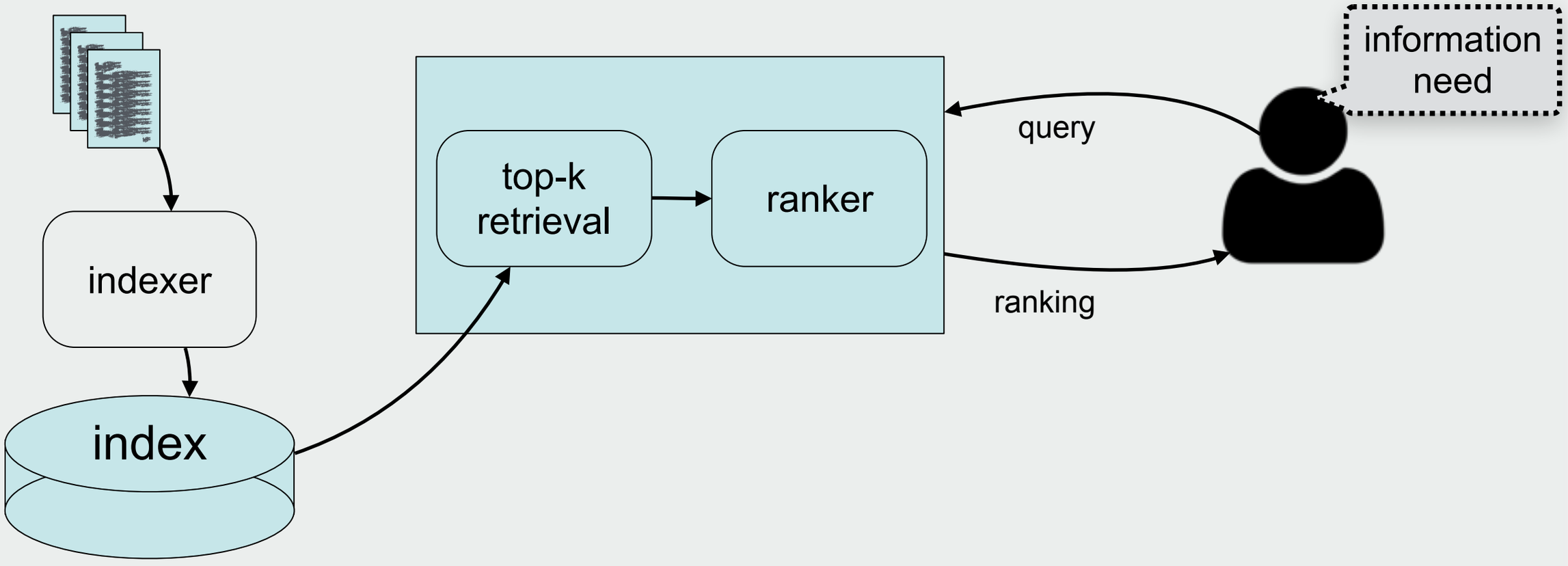
Information Retrieval



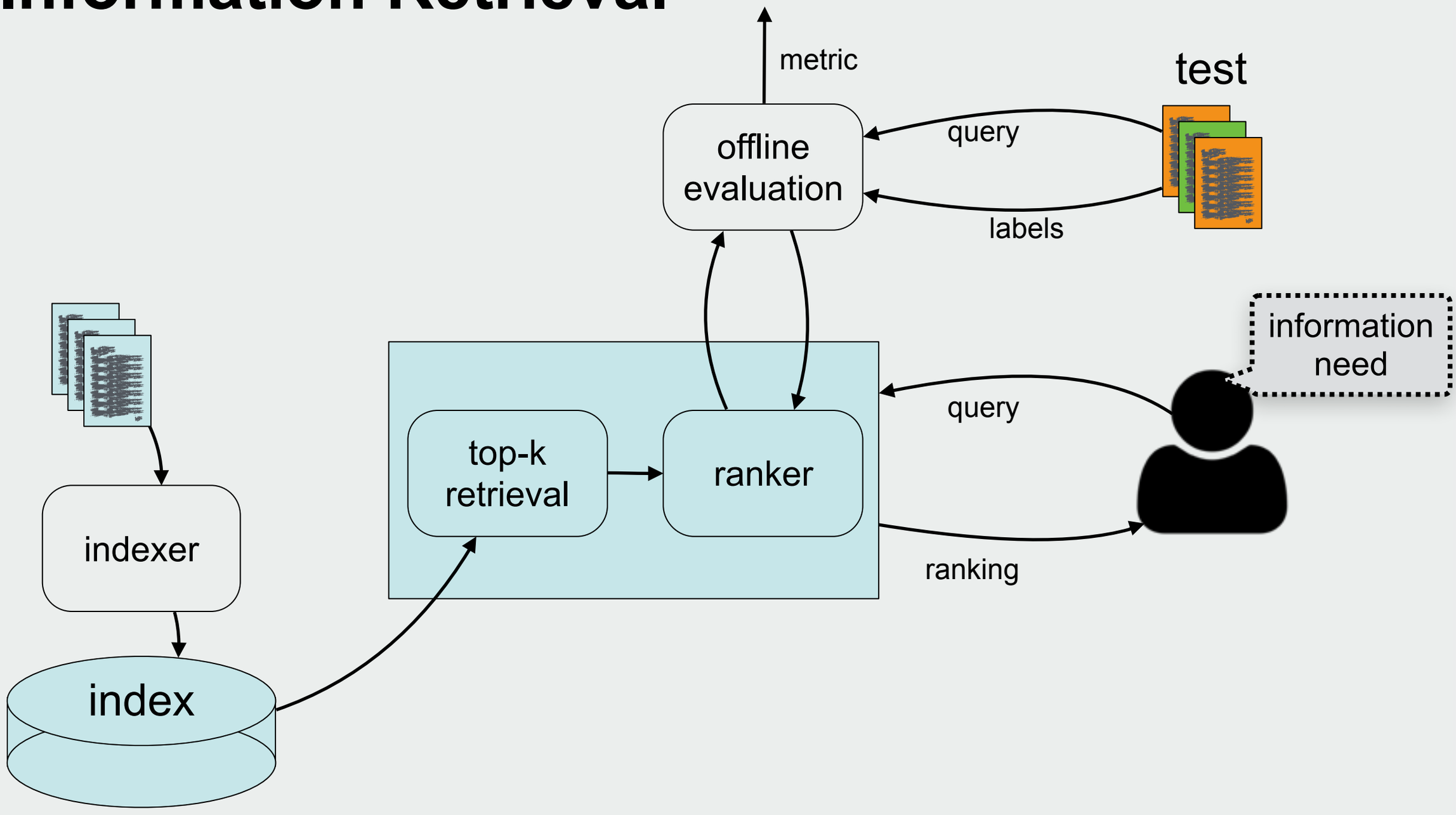
Information Retrieval



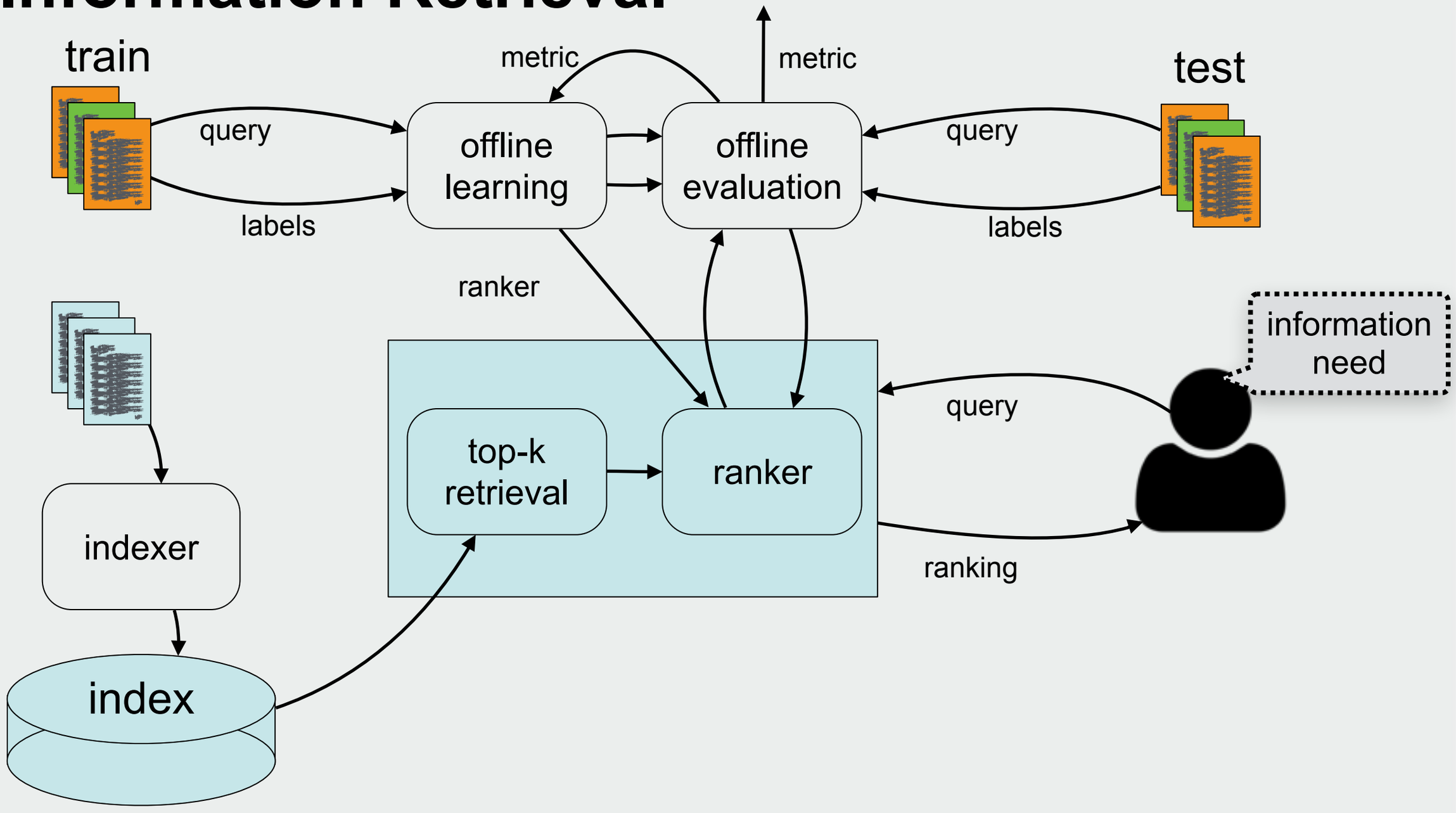
Information Retrieval



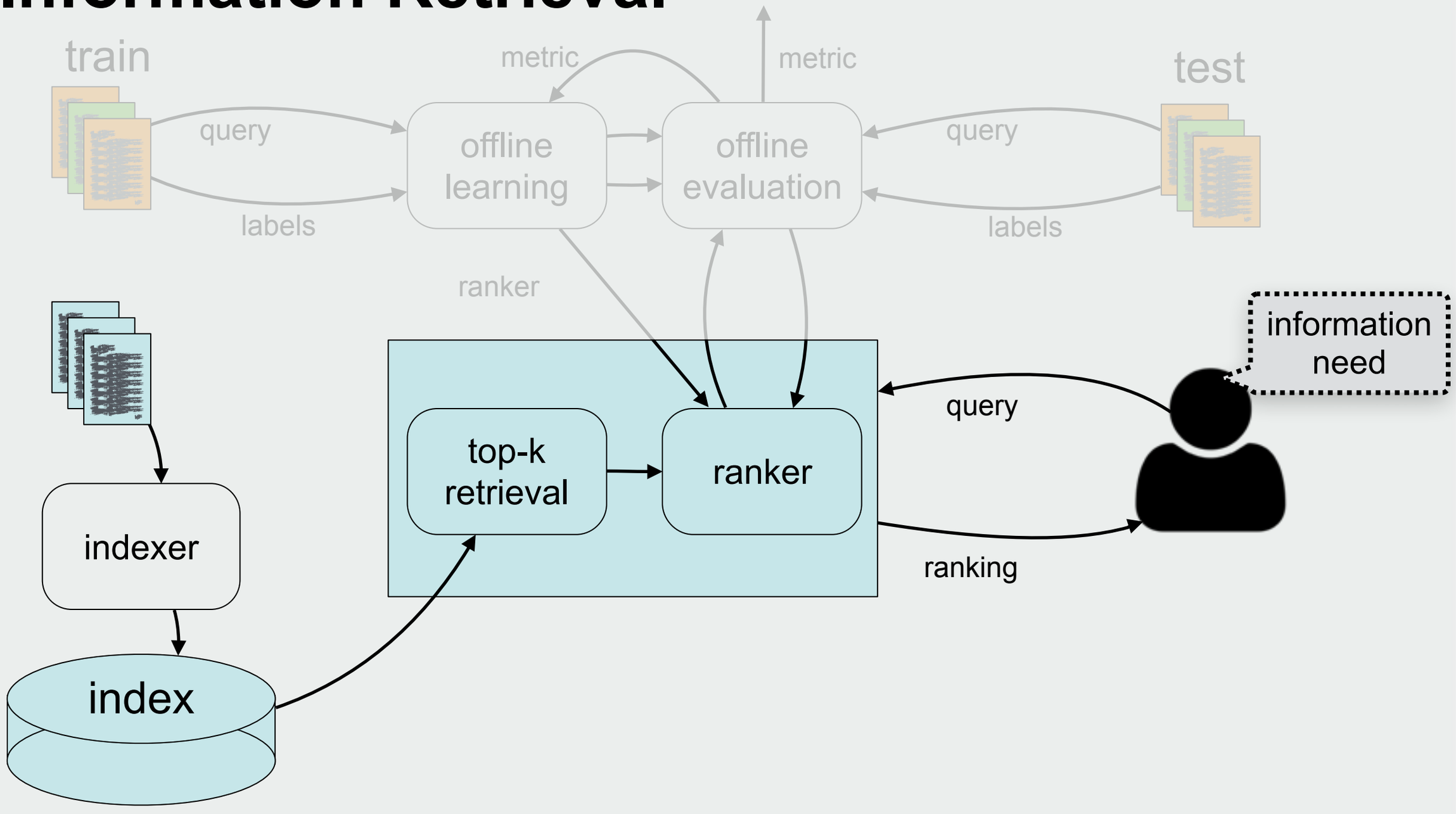
Information Retrieval



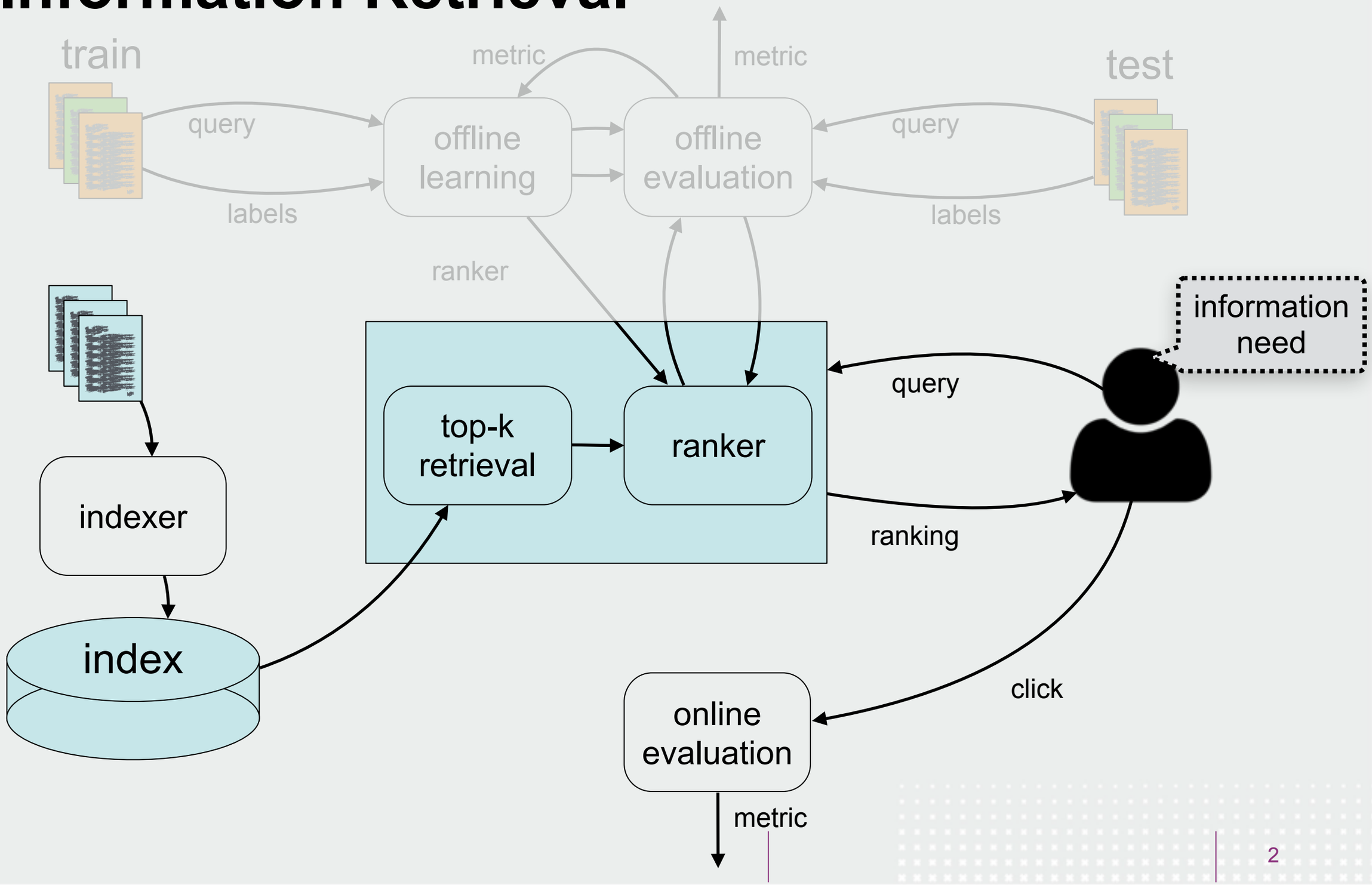
Information Retrieval



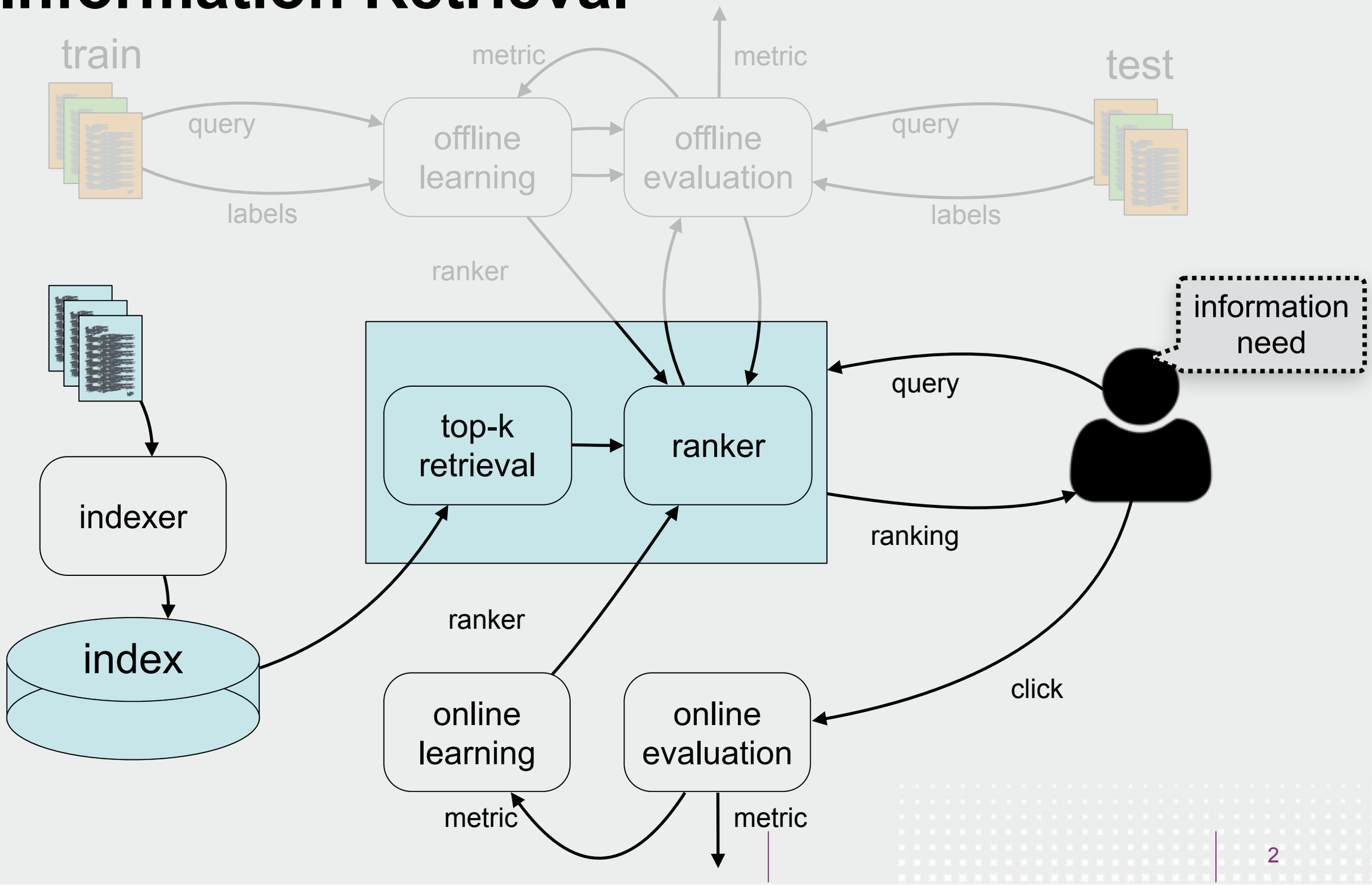
Information Retrieval



Information Retrieval



Information Retrieval



Information Retrieval

Information Retrieval

- **Users** have an **information need**

Information Retrieval

- **Users** have an **information need**
- User come to a **search engine** because they hope (or know) the answer is in the **collection**

Information Retrieval

- **Users** have an **information need**
- User come to a **search engine** because they hope (or know) the answer is in the **collection**
- Information need is translated into a **query**

Information Retrieval

- **Users** have an **information need**
- User come to a **search engine** because they hope (or know) the answer is in the **collection**
- Information need is translated into a **query**
- **Search engine**'s task is to **retrieve** documents **relevant** to information need

Recommender Systems

Information Retrieval

- **Users** have an implicit information need
- User come to a recommender system because they hope (or know) the answer is in the **collection**
- Information need is translated into a **query** (implicit (users are the query))
- Search engine / recommender system's task is to **retrieve** documents **relevant** to information need / the user

Outline

- Information Retrieval
- Online
  - Evaluation
  - Learning to Rank
  - Issues
- Living Labs
- Wrap up

Recommender Systems

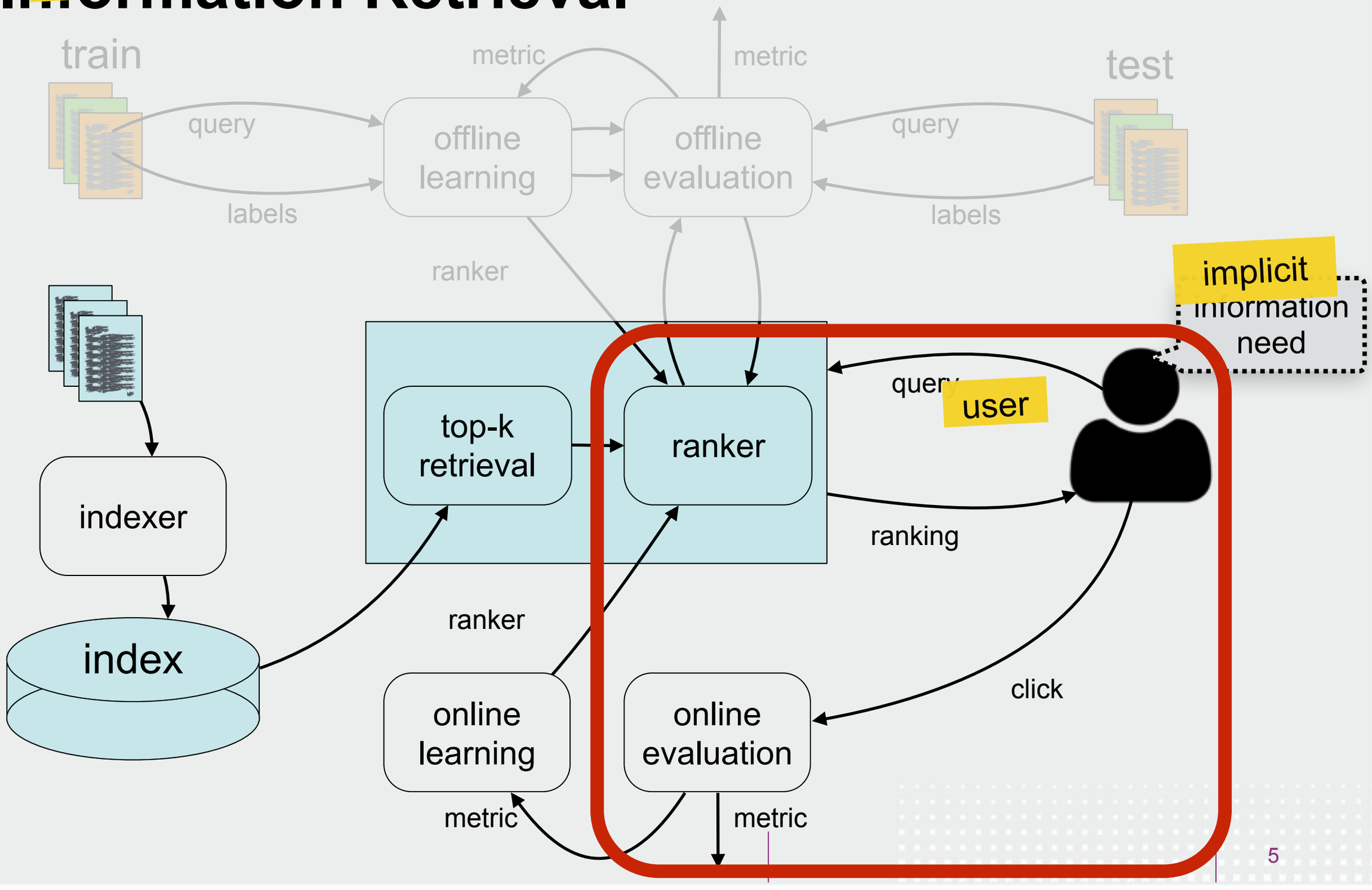
Outline

- Information Retrieval
- Online
  - Evaluation
  - Learning to Rank
  - Issues
- Living Labs
- Wrap up

Recommender Systems

Recommender Systems

Information Retrieval



What about Observational Studies?

Why not compare with historical data?
Here's an example of Kindle Sales over time.
You changed the site, and there was an amazing spike

Kohavi, R. (2013). Online Controlled Experiments. SIGIR '13.

External Events can Dwarf Your Changes

Oprah calls Kindle "her new favorite thing"

- In this example of an A/B test, you'd be better off with version A
- In controlled experiments, both versions are impacted the same way by external events

Kohavi, R. (2013). Online Controlled Experiments. SIGIR '13.

Online Evaluation

- Basic Idea:

  User can tell you what works and what not

- A/B testing
  - “Bucket” users, show A to one bucket, B to the other
  - Between subject design
- Interleaving
  - For ranking evaluation only
  - Avoid buckets, within subject design
  - Reduce variance

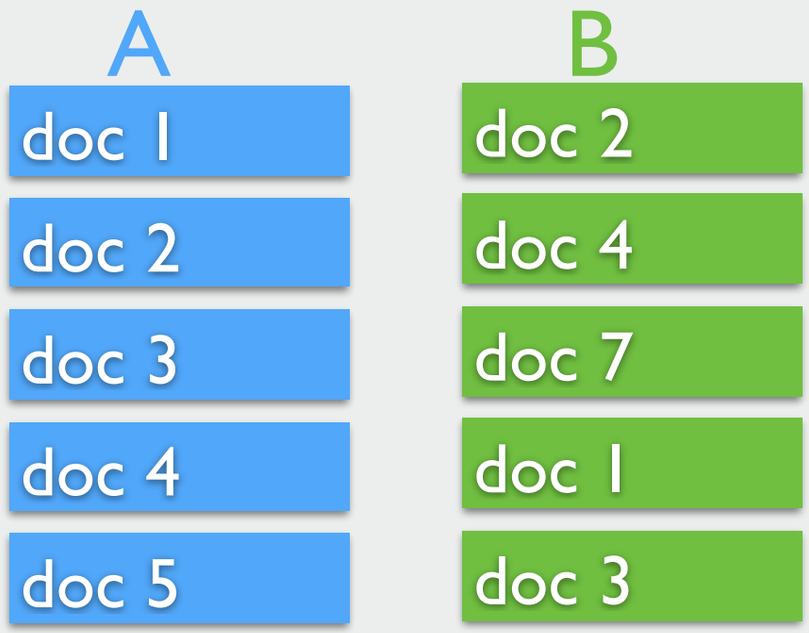
Online Evaluation

- Basic Idea:

  User can tell you what works and what not

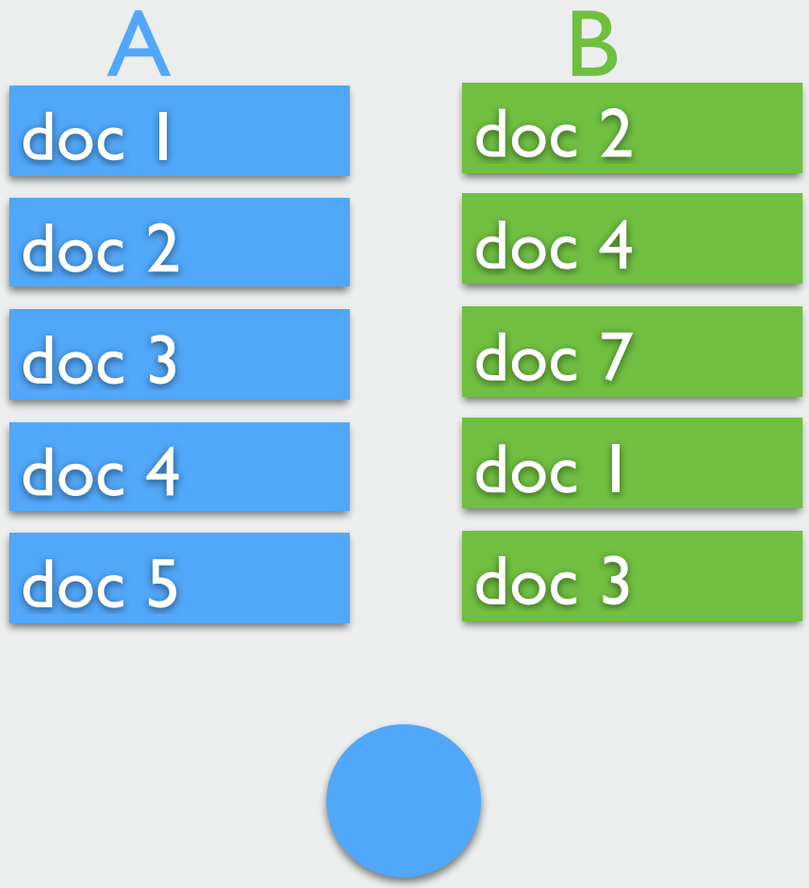
- A/B testing
  - “Bucket” users, show A to one bucket, B to the other
  - Between subject design
- Interleaving
  - For ranking evaluation only (including recommender systems?)
  - Avoid buckets, within subject design
  - Reduce variance

Interleaving



F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM '08. 2008

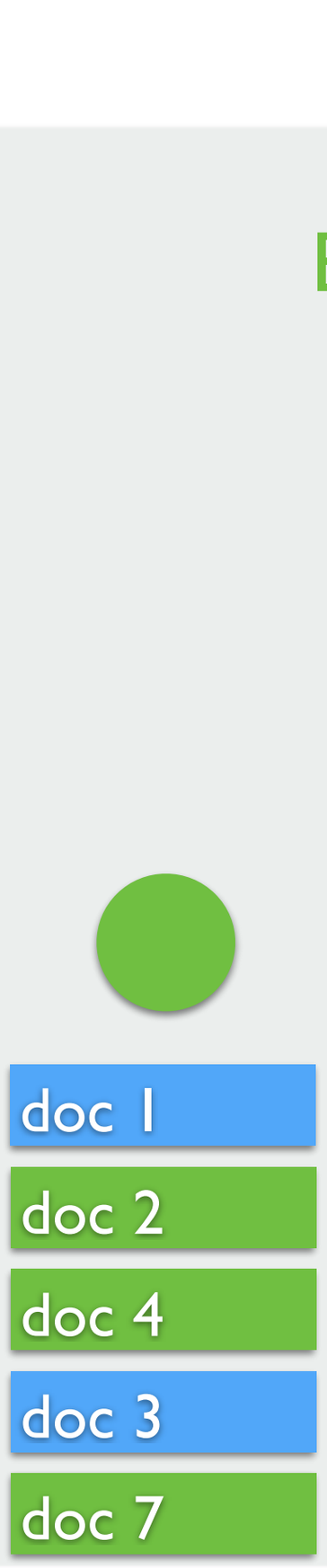
Interleaving



F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM '08. 2008

Interleaving

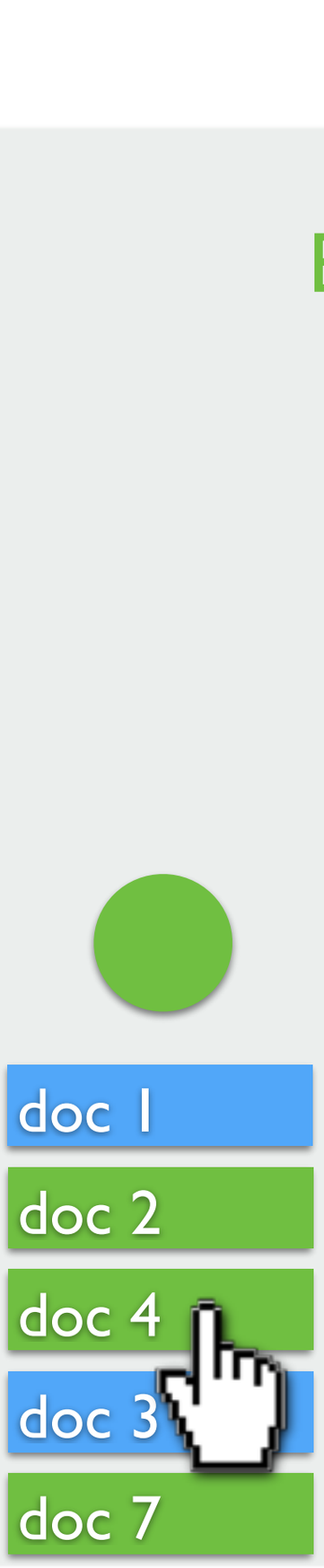
A B



F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM '08. 2008

Interleaving

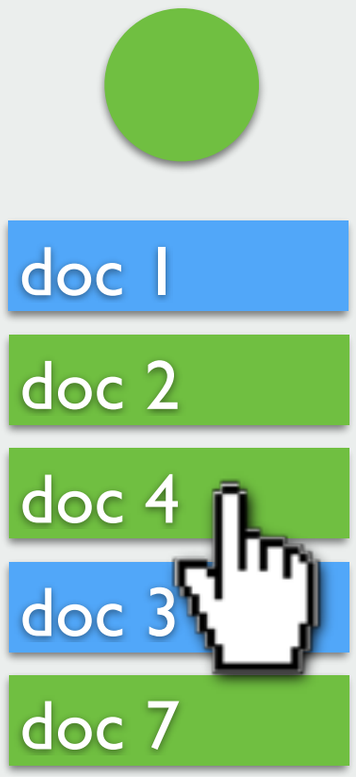
A B



F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM '08. 2008

Interleaving

A B



Inference:

B > A

F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM '08. 2008

Online Evaluation - Clicks

Online Evaluation - Clicks

- Clicks are biased

Online Evaluation - Clicks

- Clicks are biased
  - users won't click on things you didn't show them

Online Evaluation - Clicks

- Clicks are biased
  - users won't click on things you didn't show them
  - user are likely to click on things that appear high

Online Evaluation - Clicks

- Clicks are biased
  - users won't click on things you didn't show them
  - user are likely to click on things that appear high

less so for recommender systems

K. Hofmann, A. Schuth, A. Bellogin, M. de Rijke (2014): Effects of Position Bias on Click-Based Recommender Evaluation. In: ECIR'14, 2014

Online Evaluation - Clicks

- Clicks are biased
  - users won't click on things you didn't show them
  - user are likely to click on things that appear high
  - it matters how you present documents

less so for recommender systems

K. Hofmann, A. Schuth, A. Bellogin, M. de Rijke (2014): Effects of Position Bias on Click-Based Recommender Evaluation. In: ECIR'14, 2014

Online Evaluation - Clicks

- Clicks are biased
  - users won't click on things you didn't show them
  - user are likely to click on things that appear high
    - less so for recommender systems
  - it matters how you present documents
    - snippets, images, colours, font size, grouped with other documents

K. Hofmann, A. Schuth, A. Bellogin, M. de Rijke (2014): Effects of Position Bias on Click-Based Recommender Evaluation. In: ECIR'14, 2014

Online Evaluation - Clicks

- Clicks are biased
  - users won't click on things you didn't show them
  - user are likely to click on things that appear high
    - less so for recommender systems
  - it matters how you present documents
    - snippets, images, colours, font size, grouped with other documents
- Clicks are noisy

K. Hofmann, A. Schuth, A. Bellogin, M. de Rijke (2014): Effects of Position Bias on Click-Based Recommender Evaluation. In: ECIR'14, 2014

Online Evaluation - Clicks

- Clicks are biased
  - users won't click on things you didn't show them
  - user are likely to click on things that appear high
    - less so for recommender systems
  - it matters how you present documents
    - snippets, images, colours, font size, grouped with other documents
- Clicks are noisy
  - they don't always mean what you hope

K. Hofmann, A. Schuth, A. Bellogin, M. de Rijke (2014): Effects of Position Bias on Click-Based Recommender Evaluation. In: ECIR'14, 2014

Why not Just Use Clicks?

greenfield, mn accident

Annandale man dies in car/truck crash in Greenfield

Article by: Star Tribune Updated: January 12, 2010 - 8:59 PM

resize text | print | buy reprints

Recommend 0 Tweet 0 share +

more from west metro

A 21-year-old man from Annandale, Minn., was killed Tuesday afternoon in a car-truck crash on Hwy. 55 in Greenfield, according to the Minnesota State Patrol.

Battle over plan to restrict access to metro lakes

Time spent on page: 38 seconds

Microsoft Research

Hassan, A., Shi, X., Craswell, N., & Ramsey, B. (2013). Beyond Clicks: Query Reformulation as a Predictor of Search Satisfaction. In CIKM '13.

Why not Just Use Clicks?

greenfield, mn accident

Annandale man dies in car/truck crash in Greenfield

Article by: Star Tribune Updated: January 12, 2010 - 8:59 PM

Woman dies in a fatal accident in greenfield, minnesota

Session Ends

Microsoft Research

Hassan, A., Shi, X., Craswell, N., & Ramsey, B. (2013). Beyond Clicks: Query Reformulation as a Predictor of Search Satisfaction. In CIKM '13.

Why not Just Use Clicks?

- User performed this search on July 1st
- User was probably looking for

Car-truck crash in Greenfield kills woman, 34

Updated: June 30, 2012 - 9:46 PM

0 comments | resize text | print | buy reprints

Recommend 38 Tweet share +

A 34-year-old woman was killed shortly after 6 p.m. Saturday when the car she was driving collided with a pickup truck at the intersection of County Road 50 and Vernon Street in Greenfield.

more from local

Minn. man sentenced for fake sports apparel

Microsoft Research

Hassan, A., Shi, X., Craswell, N., & Ramsey, B. (2013). Beyond Clicks: Query Reformulation as a Predictor of Search Satisfaction. In CIKM '13.

Why not Just Use Clicks?

- User clicked on a result
- The dwell time is long
- But, user was not satsified

Query **Click** Query

Clicks do not always mean satisfaction

Microsoft Research

Hassan, A., Shi, X., Craswell, N., & Ramsey, B. (2013). Beyond Clicks: Query Reformulation as a Predictor of Search Satisfaction. In CIKM '13.

Online Evaluation - Clicks

- Clicks are biased
  - users won't click on things you didn't show them
  - user are likely to click on things that appear high
    - less so for recommender systems
  - it matters how you present documents
    - snippets, images, colours, font size, grouped with other documents
- Clicks are noisy
  - they don't always mean what you hope

Online Evaluation - Clicks

- Clicks are biased
  - users won't click on things you didn't show them
  - user are likely to click on things that appear high
    - less so for recommender systems
  - it matters how you present documents
    - snippets, images, colours, font size, grouped with other documents
- Clicks are noisy
  - they don't always mean what you hope
  - absence of clicks is not always negative

Why not Just Use Clicks?

Lack of clicks does not always mean dissatisfaction

Microsoft Research

Hassan, A., Shi, X., Craswell, N., & Ramsey, B. (2013). Beyond Clicks: Query Reformulation as a Predictor of Search Satisfaction. In CIKM '13.

Online Evaluation - Clicks

- Clicks are biased
  - understand the bias
  - correct for it

- Clicks are noisy
  - repeat

Outline

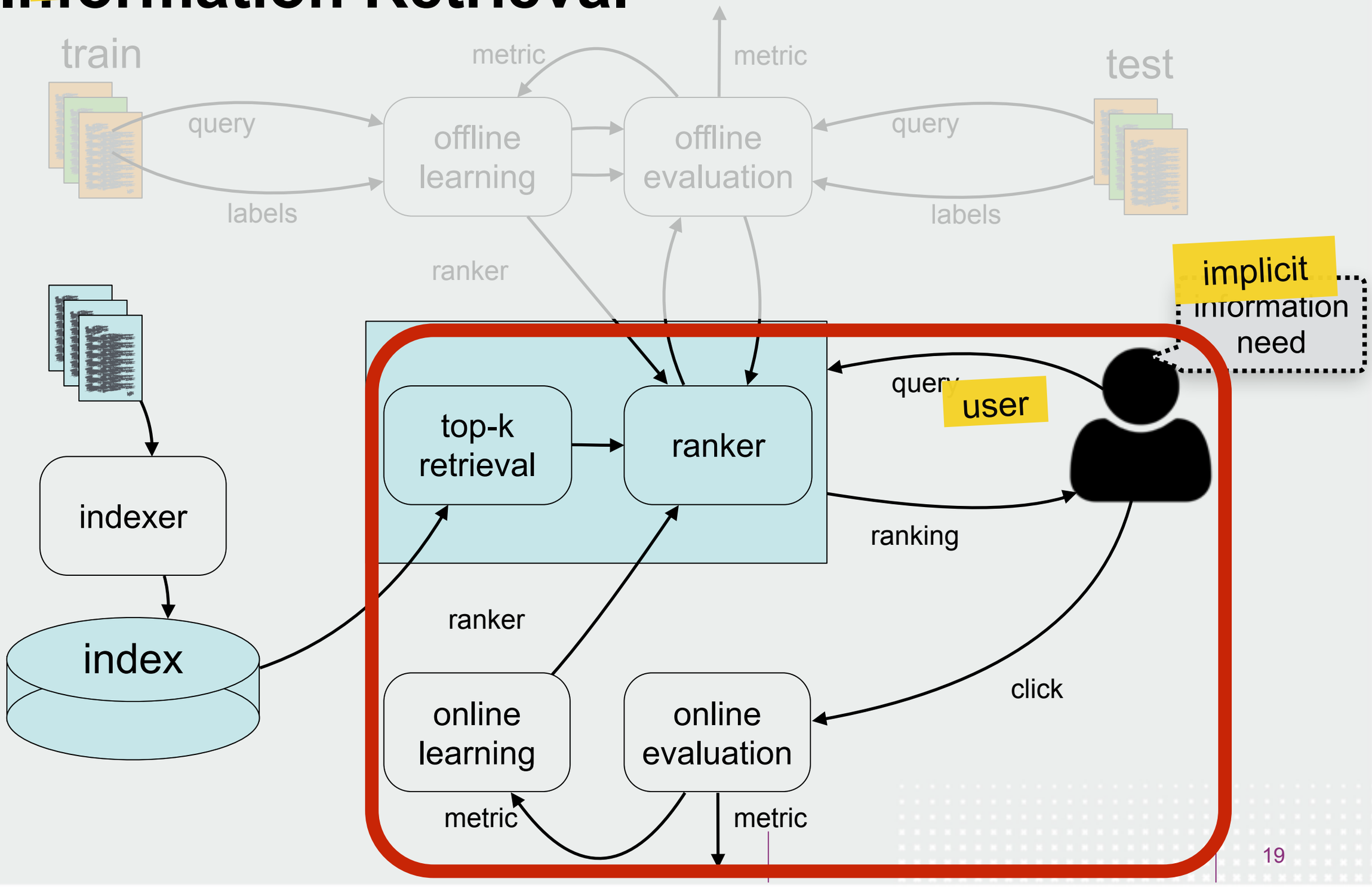
- Information Retrieval
- Online
  - Evaluation
  - Learning to Rank
  - Issues
- Living Labs
- Wrap up

Outline

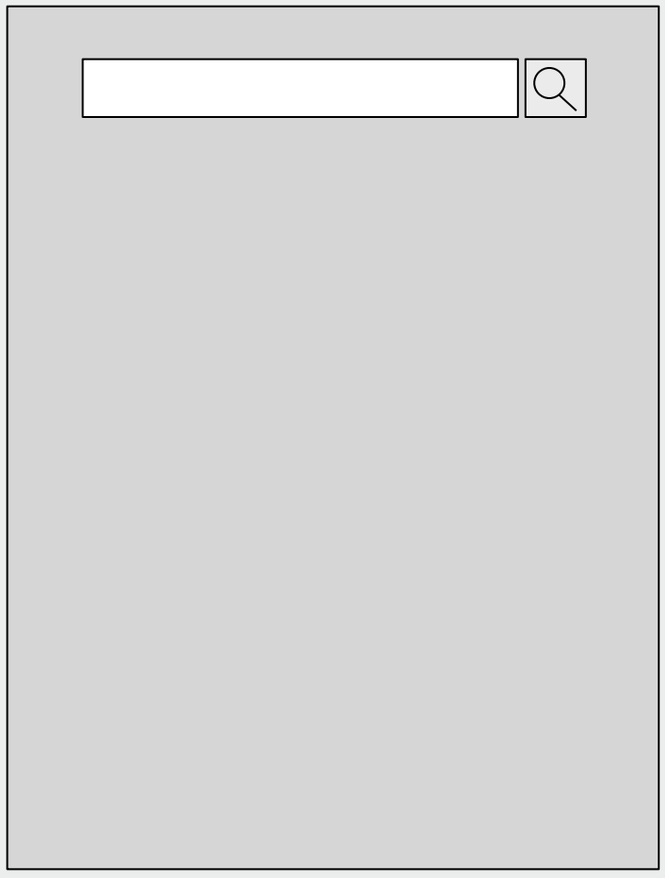
- Information Retrieval
- Online
  - Evaluation
  - Learning to Rank
  - Issues
- Living Labs
- Wrap up

Recommender Systems

Information Retrieval

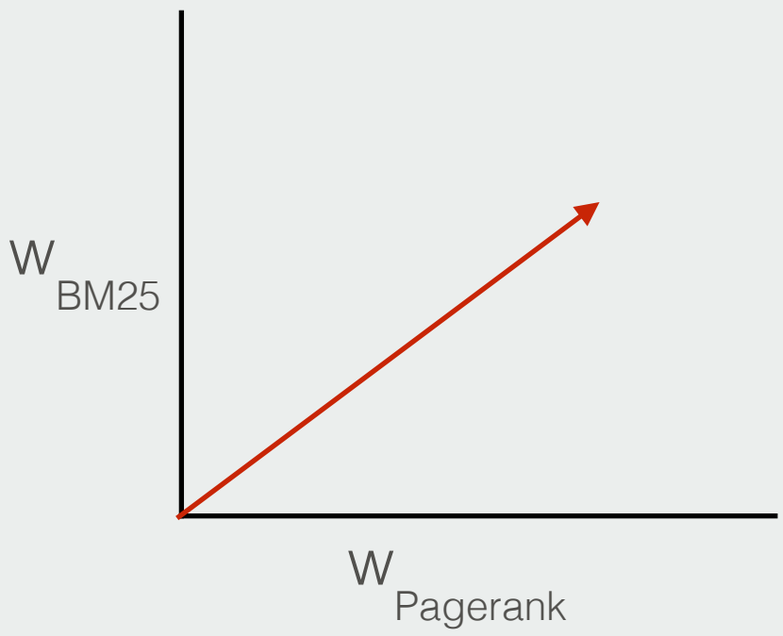


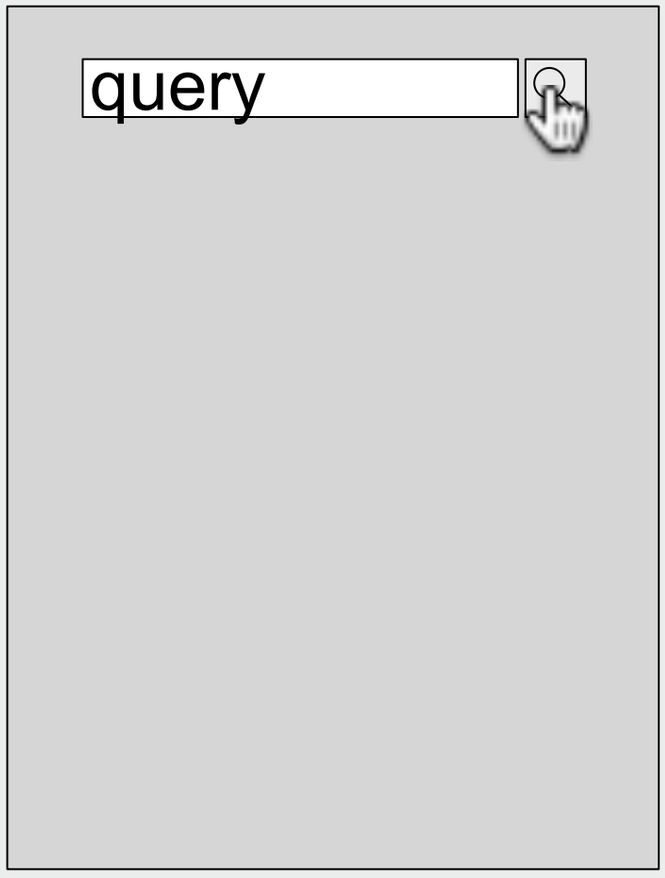
Dueling Bandit Gradient Descent



, 2009; Hofmann et al., 2011]

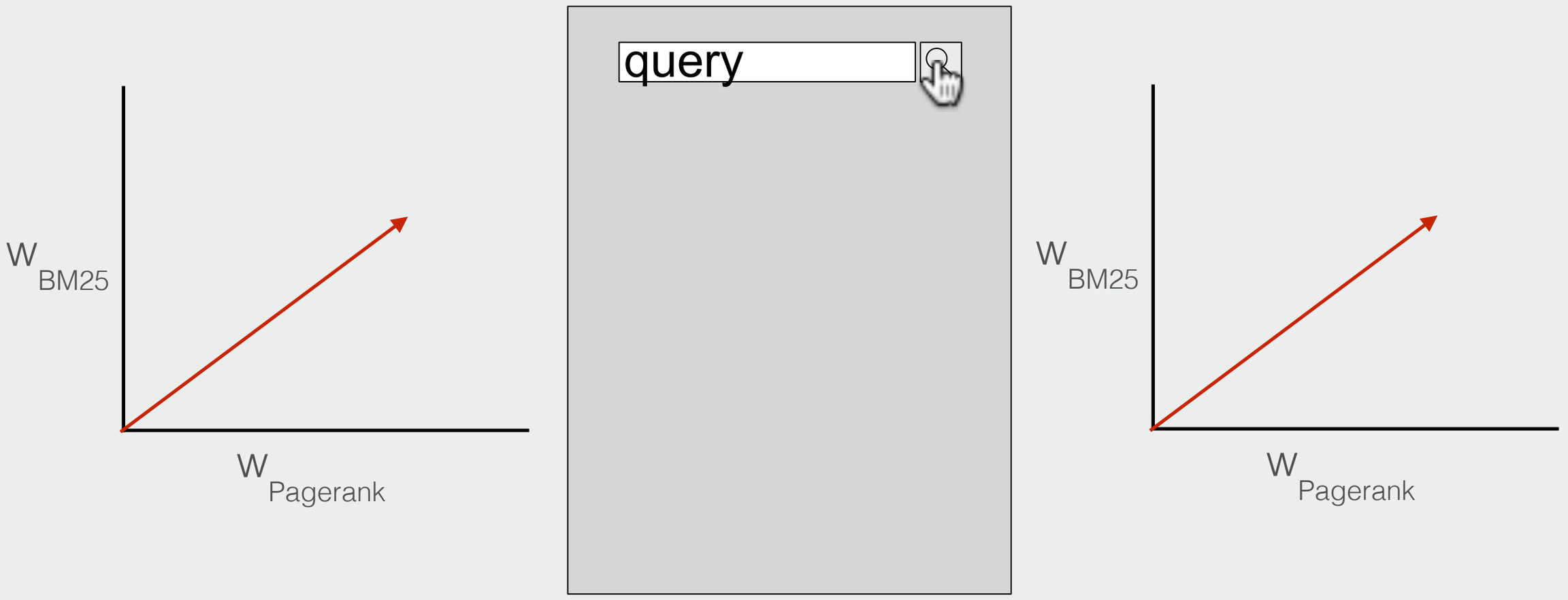
Dueling Bandit Gradient Descent





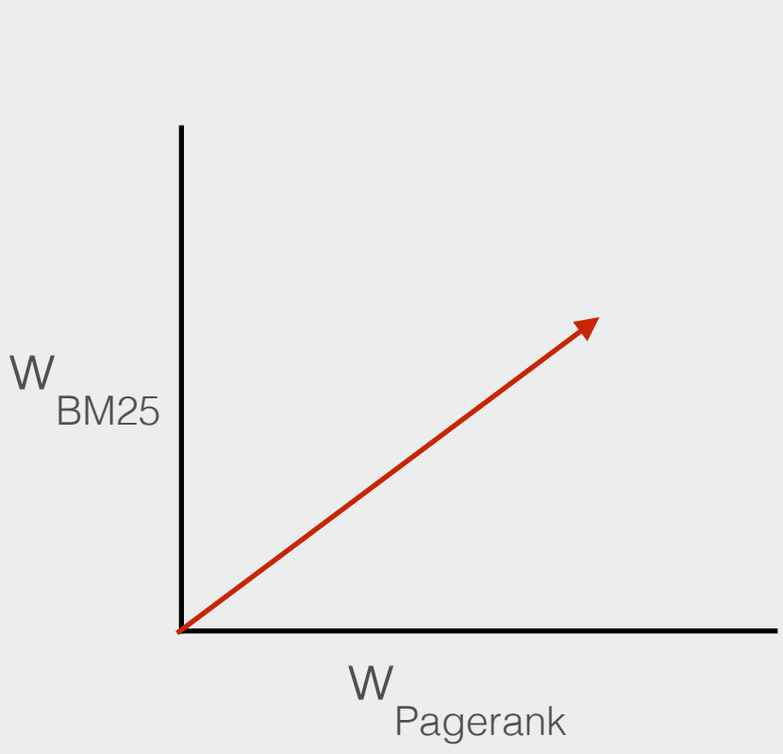
, 2009; Hofmann et al., 2011]

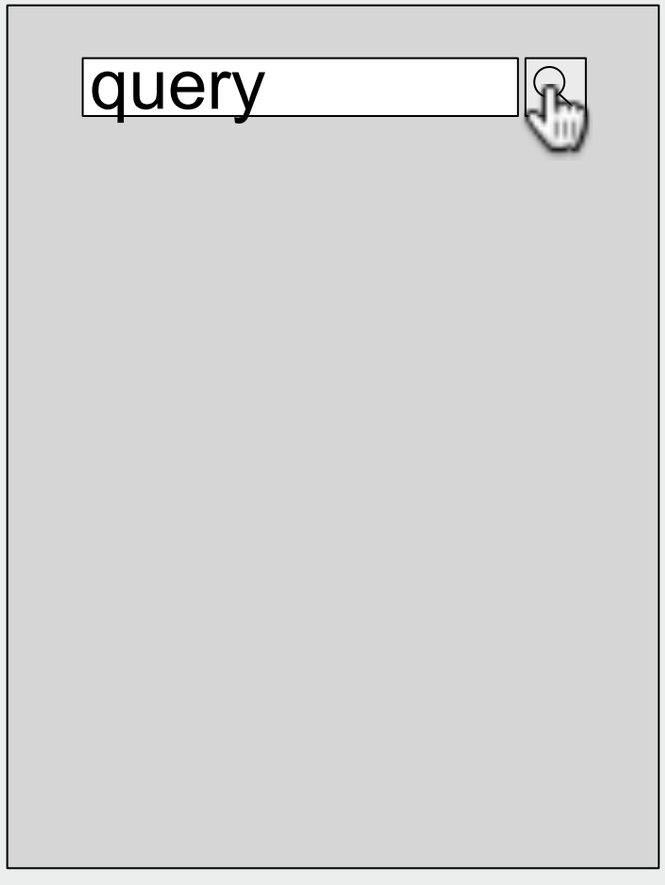
Dueling Bandit Gradient Descent

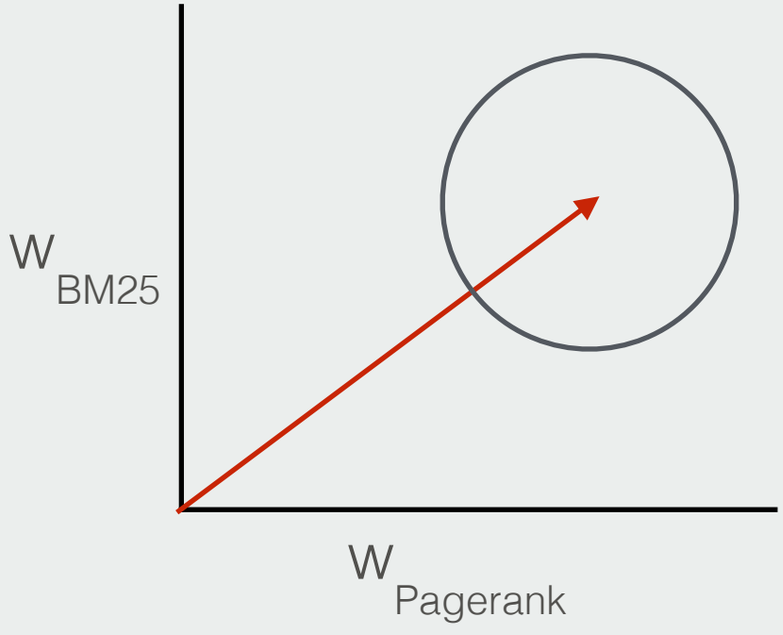


, 2009; Hofmann et al., 2011]

Dueling Bandit Gradient Descent



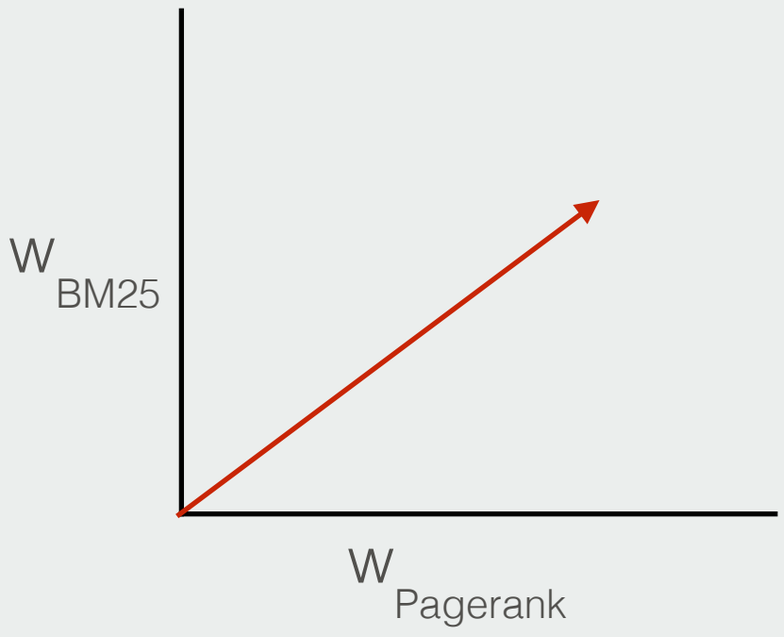


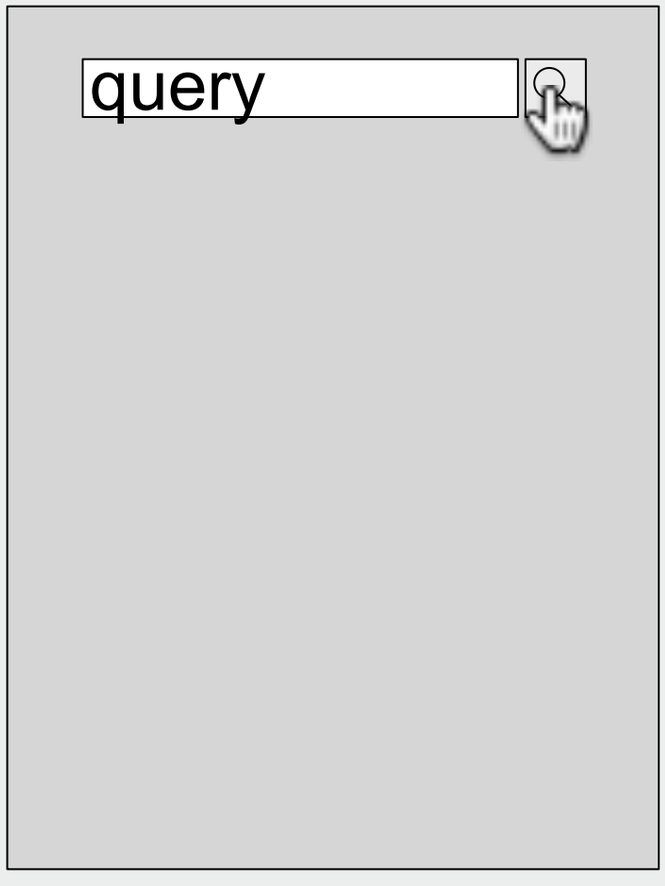


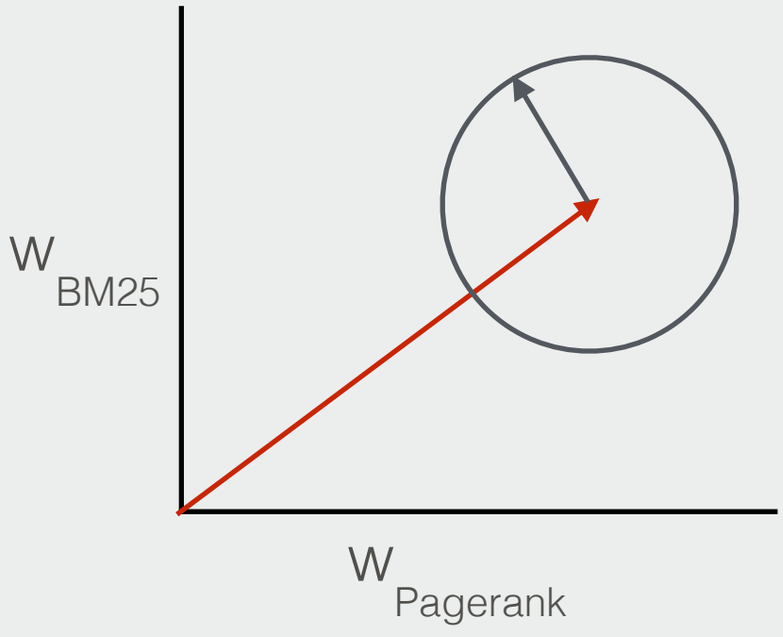
, 2009; Hofmann et al., 2011]

UNIVERSITY OF AMSTERDAM

Dueling Bandit Gradient Descent

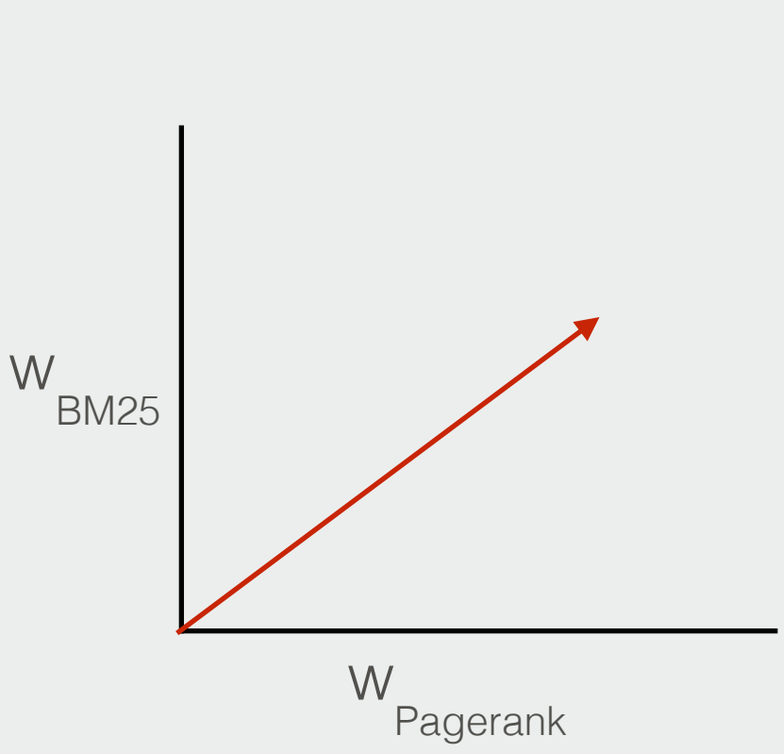


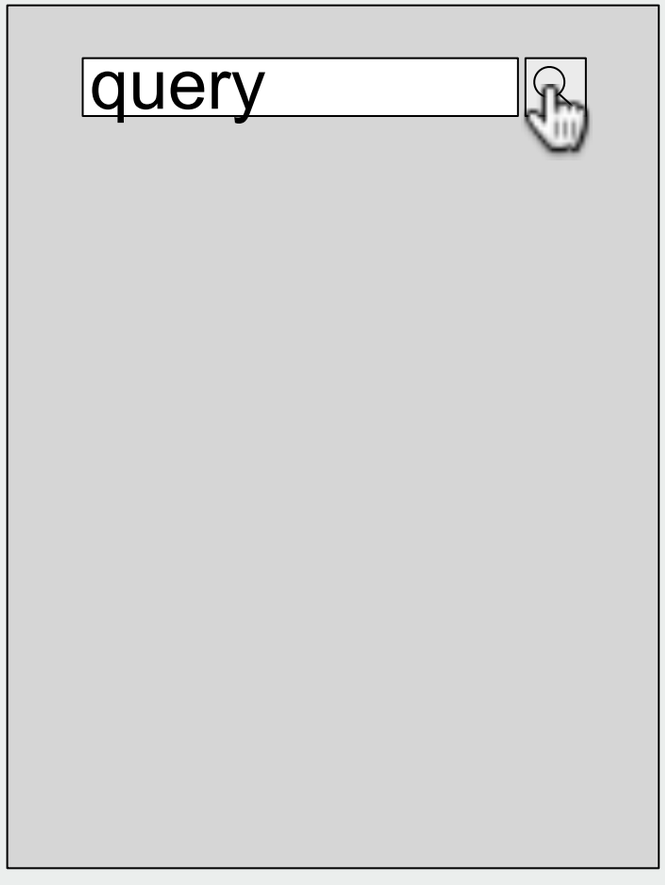


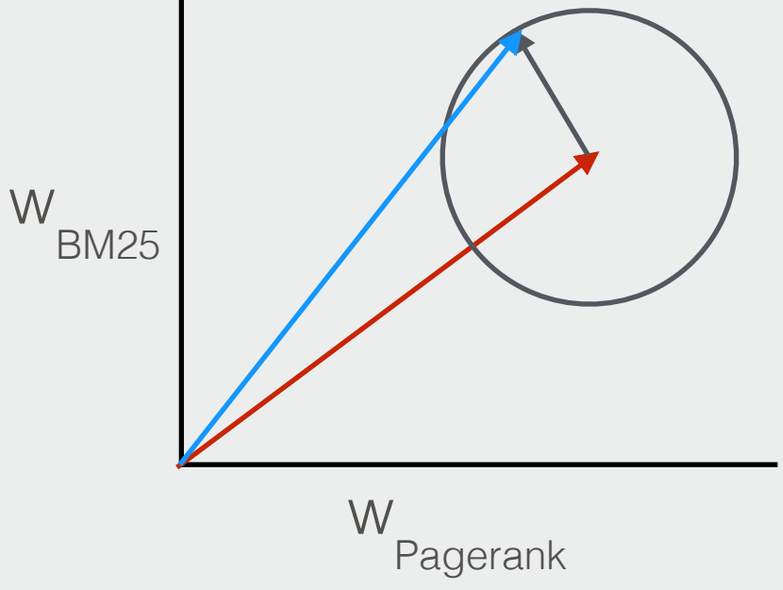


, 2009; Hofmann et al., 2011]

Dueling Bandit Gradient Descent



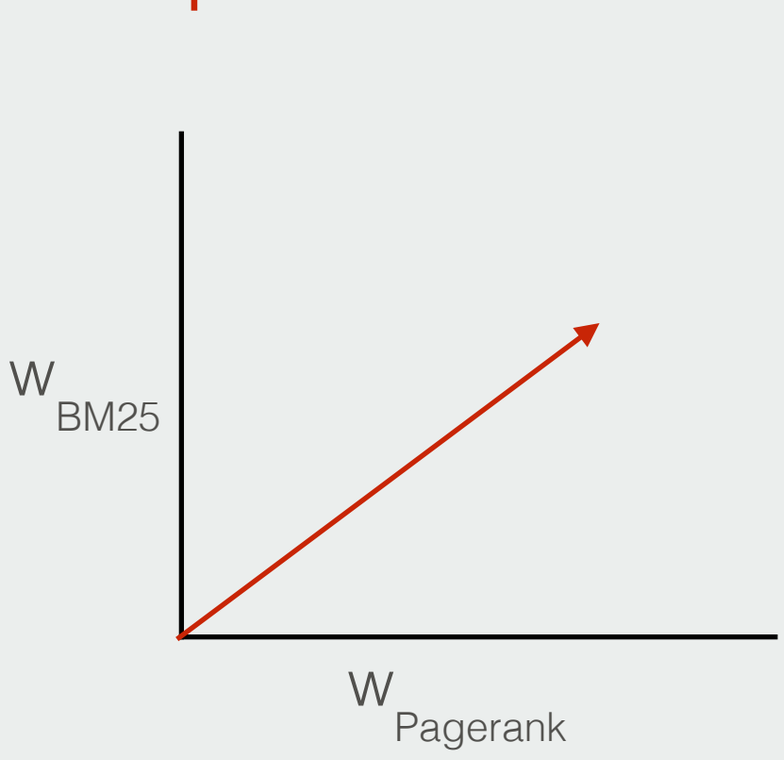


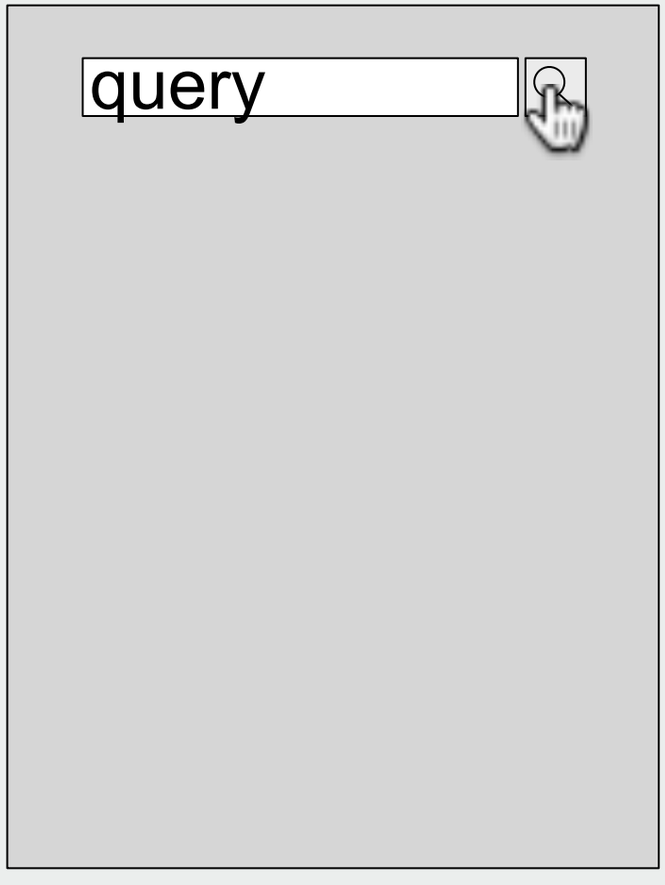


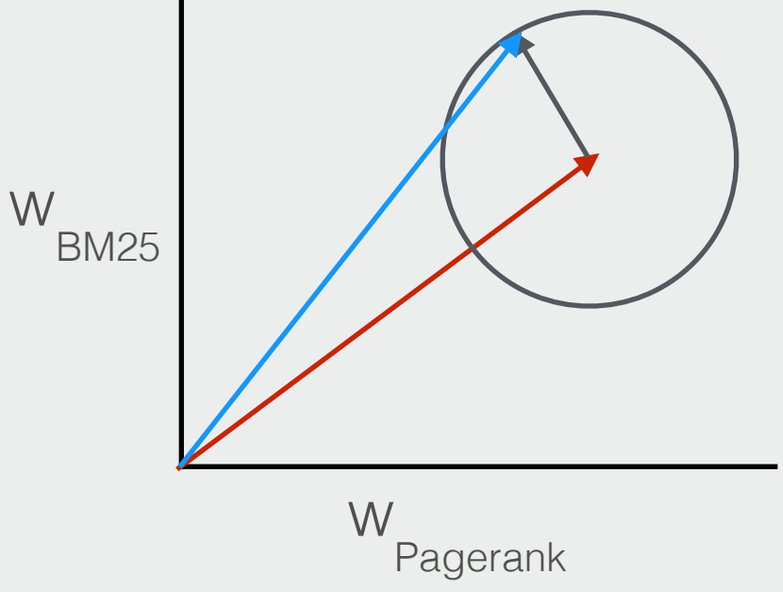
, 2009; Hofmann et al., 2011]

Dueling Bandit Gradient Descent

Exploitative Ranker



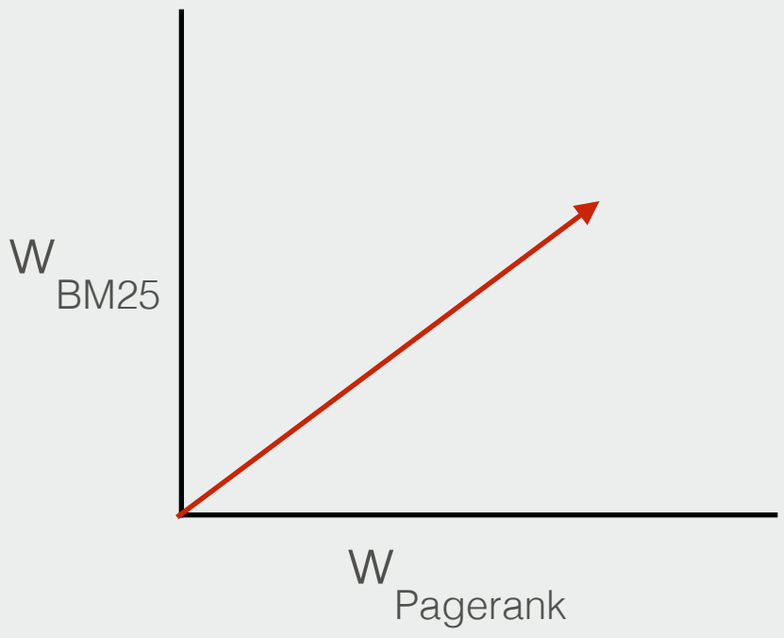


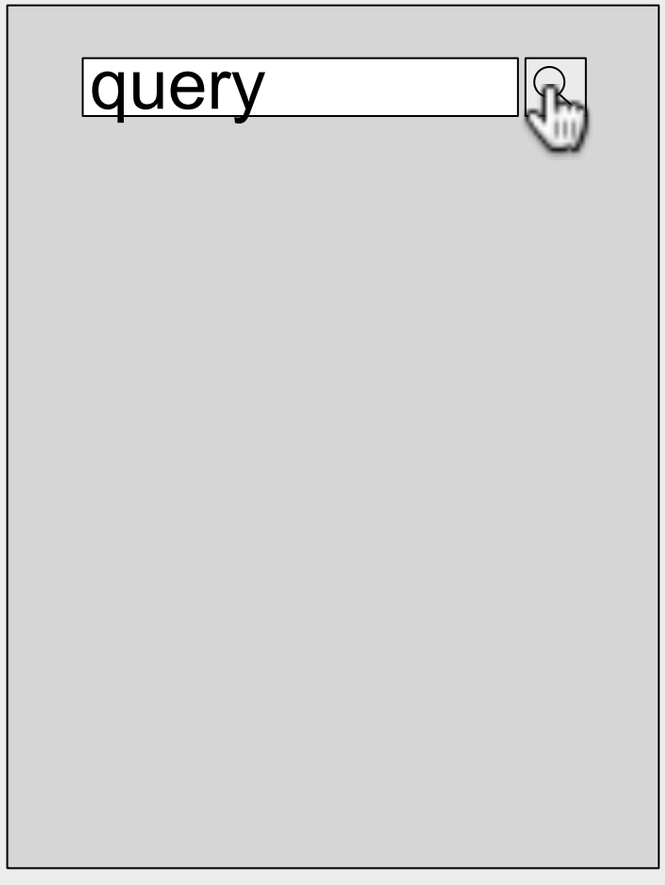


, 2009; Hofmann et al., 2011]

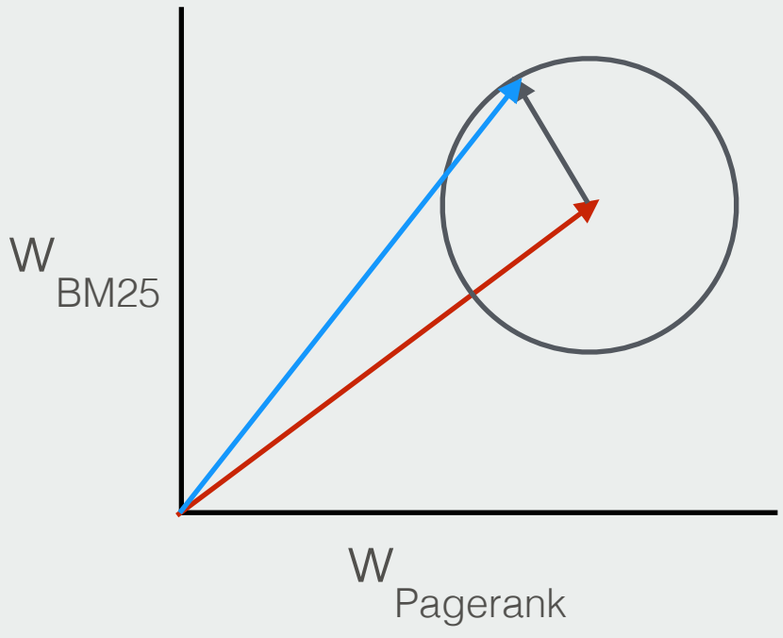
Dueling Bandit Gradient Descent

Exploitative Ranker



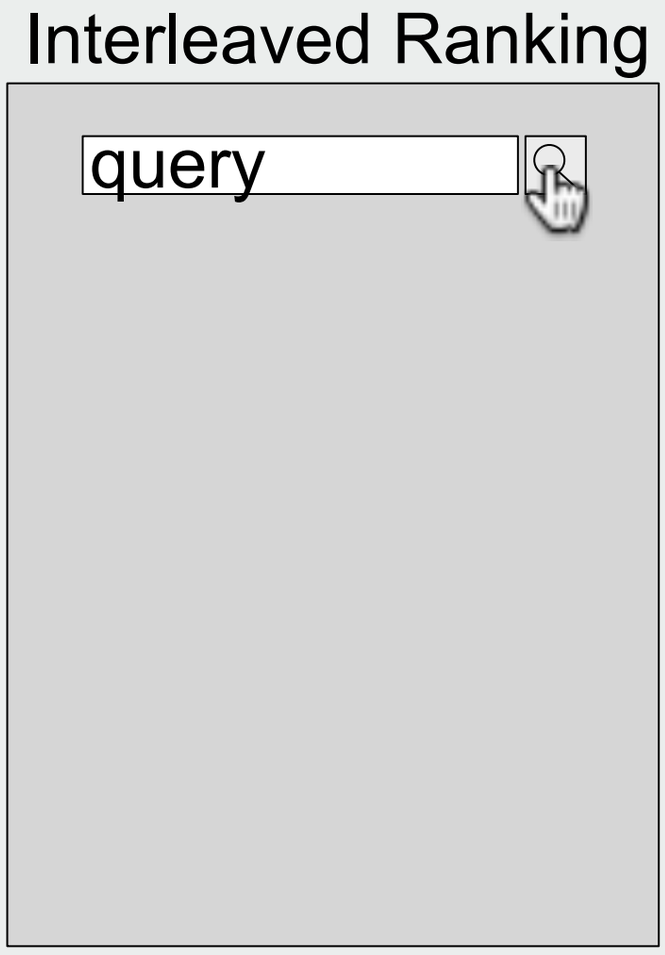


Explorative Ranker



, 2009; Hofmann et al., 2011]

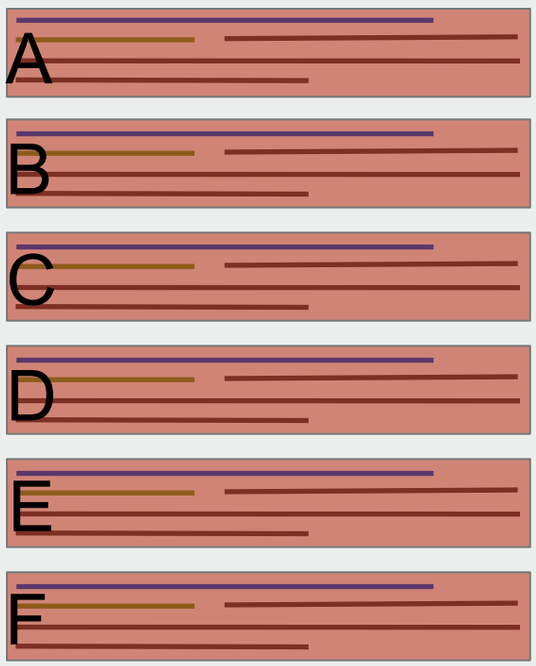
TeamDraft Interleave



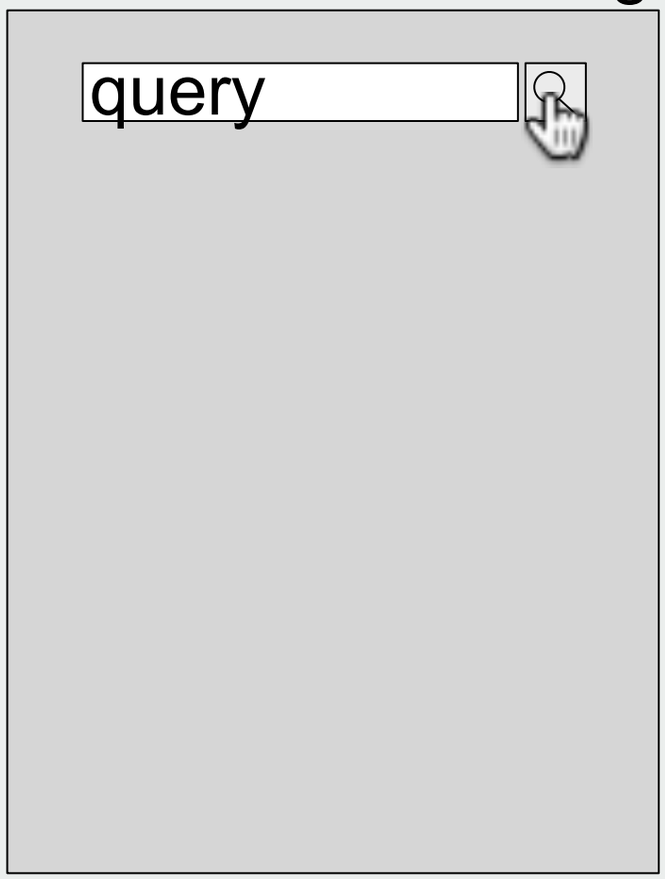
Radlinski, F., Kurup, M., & Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In CIKM '08.

TeamDraft Interleave

Exploitative Ranking



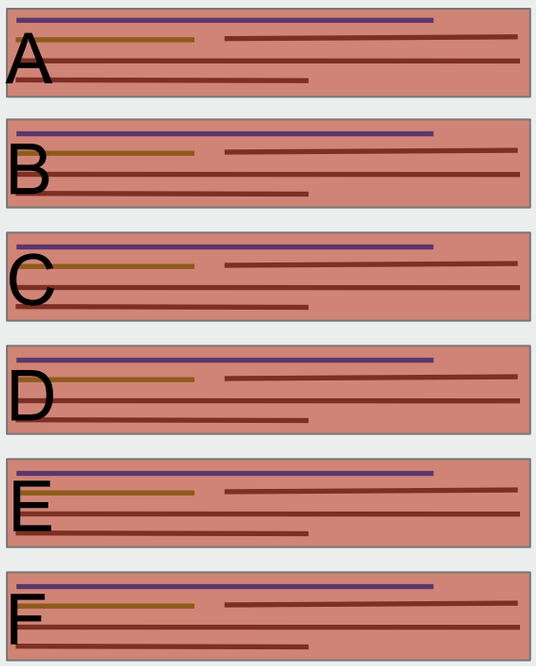
Interleaved Ranking



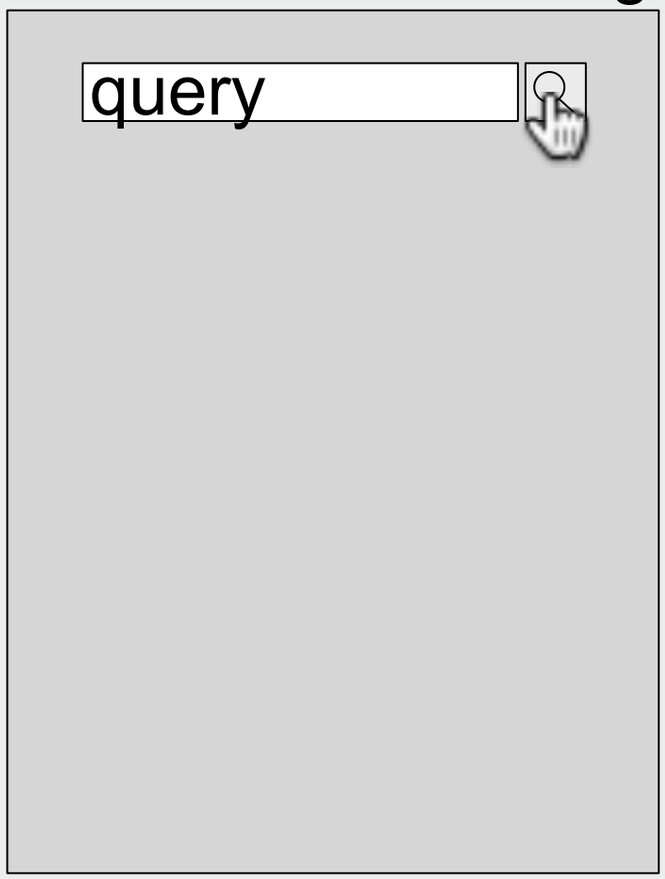
Radlinski, F., Kurup, M., & Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In CIKM '08.

TeamDraft Interleave

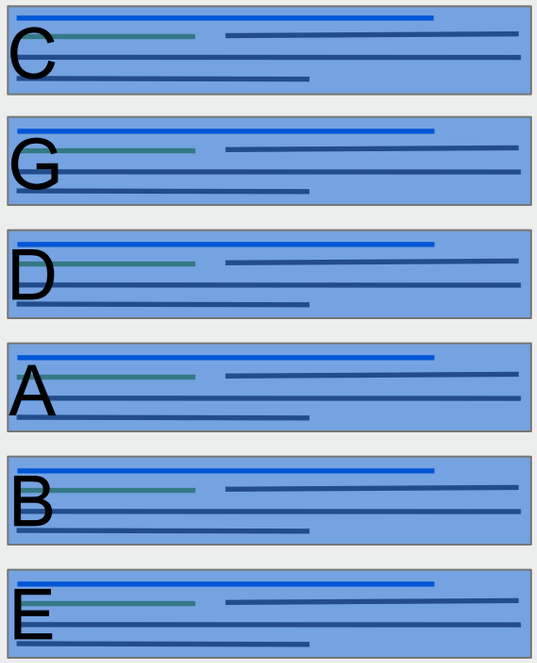
Exploitative Ranking



Interleaved Ranking



Explorative Ranking

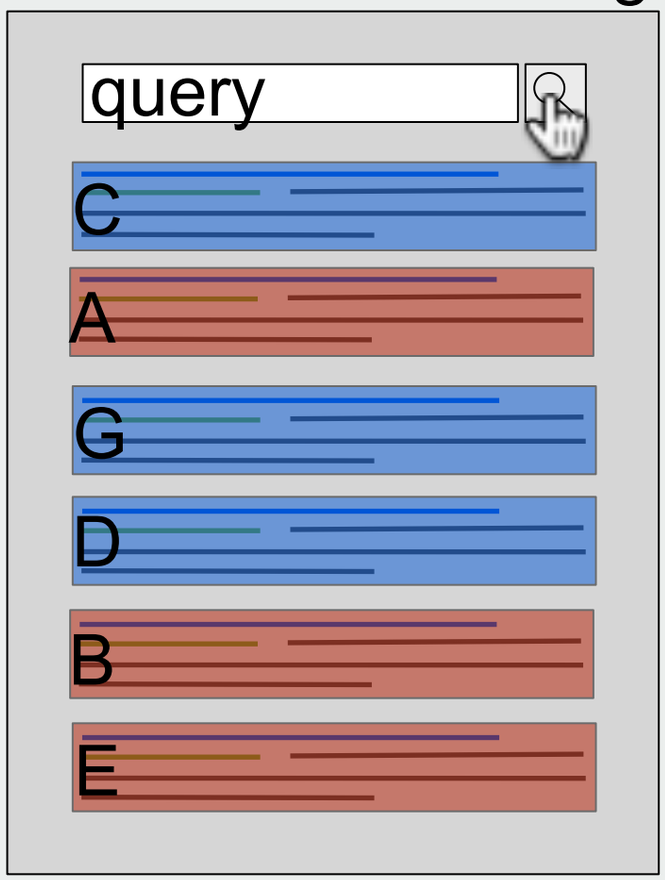


Radlinski, F., Kurup, M., & Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In CIKM '08.

TeamDraft Interleave

Exploitative Ranking

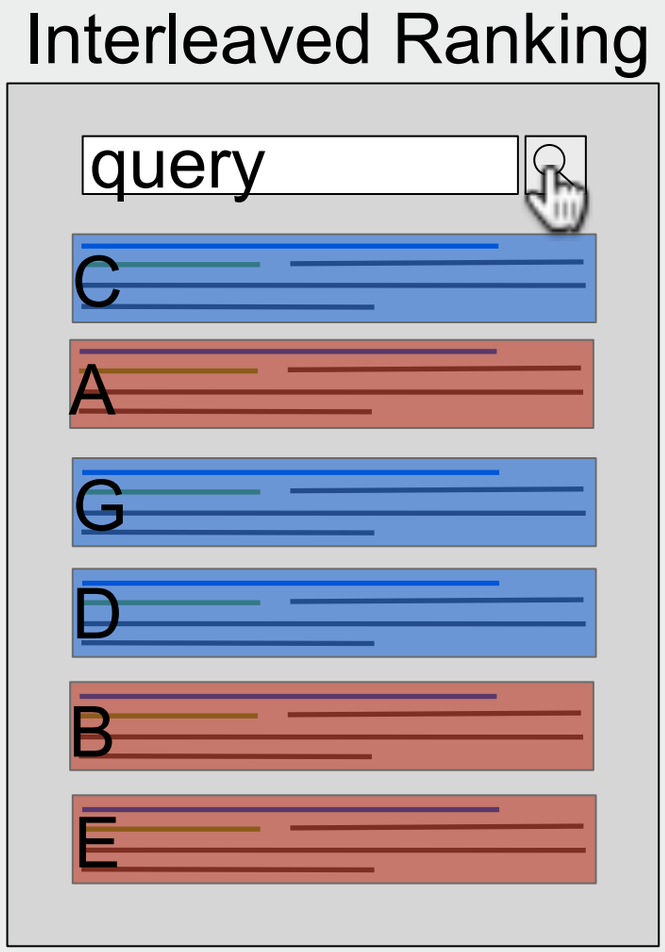
Interleaved Ranking



Explorative Ranking

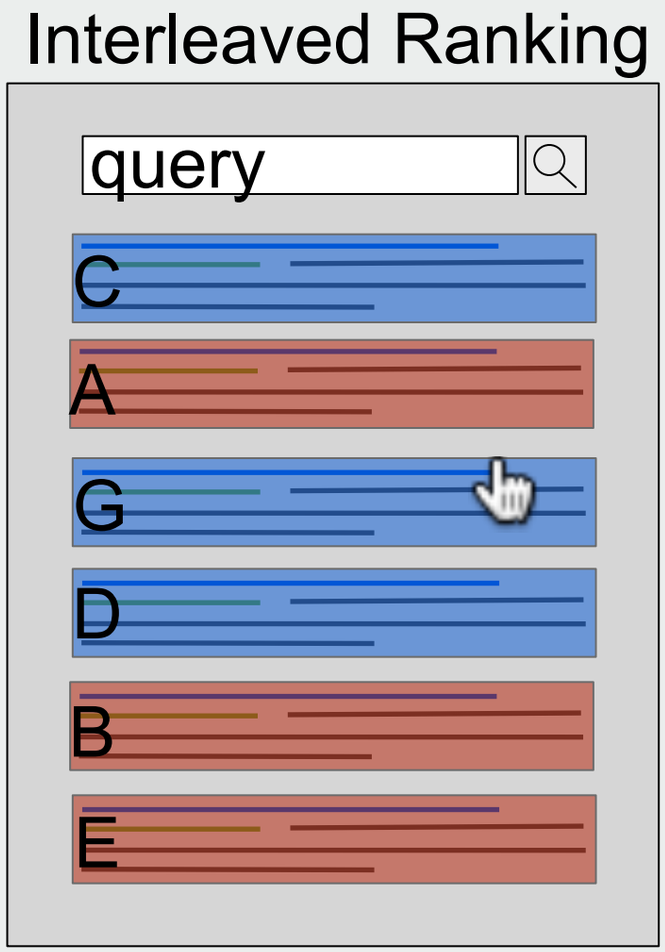
Radlinski, F., Kurup, M., & Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In CIKM '08.

TeamDraft Interleave



Radlinski, F., Kurup, M., & Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In CIKM '08.

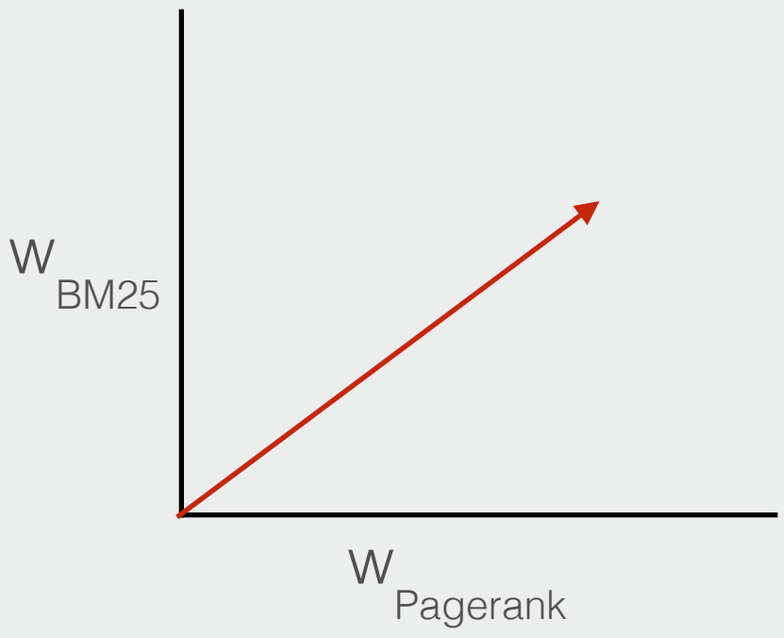
TeamDraft Interleave

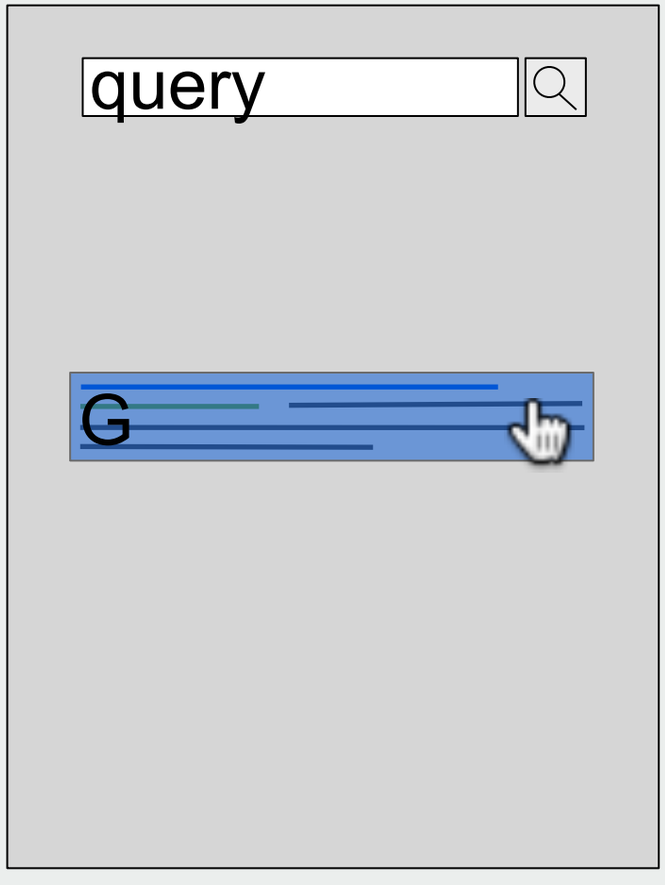


Radlinski, F., Kurup, M., & Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In CIKM '08.

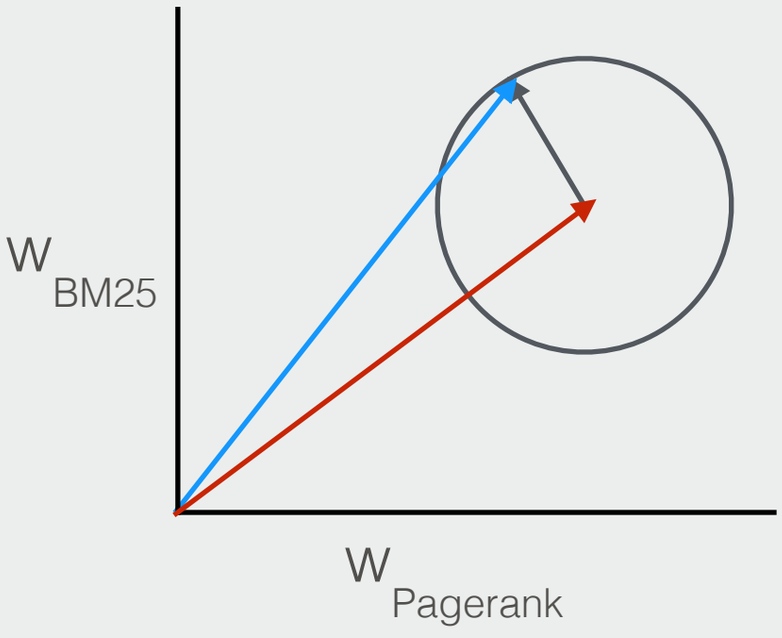
Dueling Bandit Gradient Descent

Exploitative Ranker





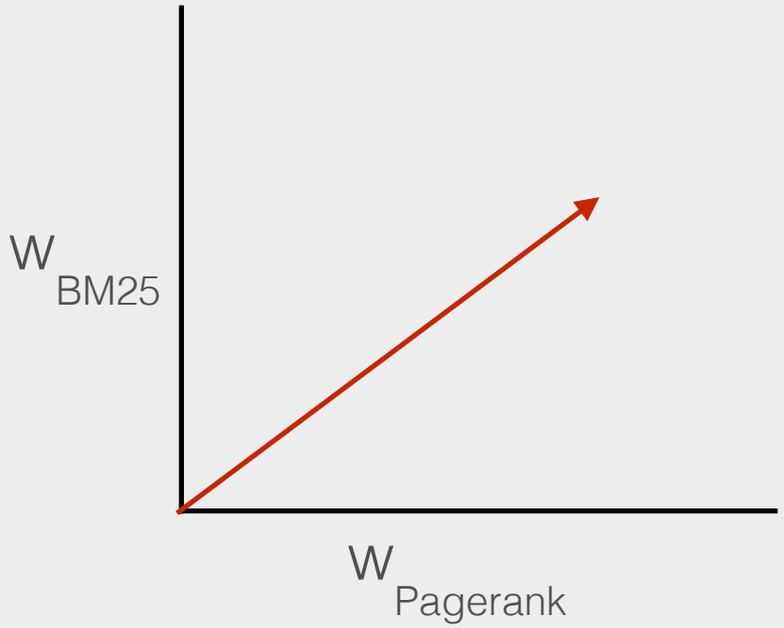
Explorative Ranker

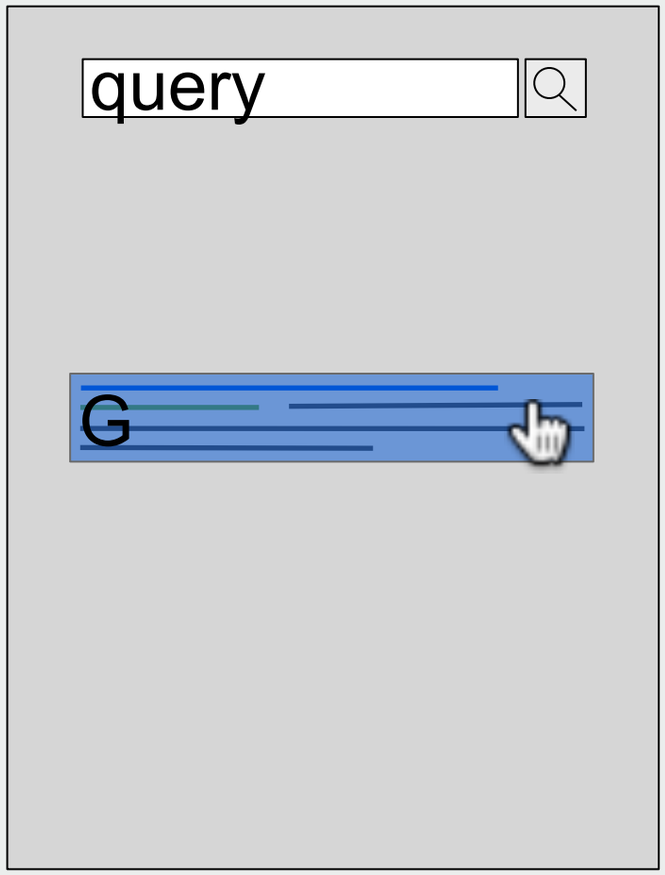


[Yue et al, 2009; Hofmann et al., 2011]

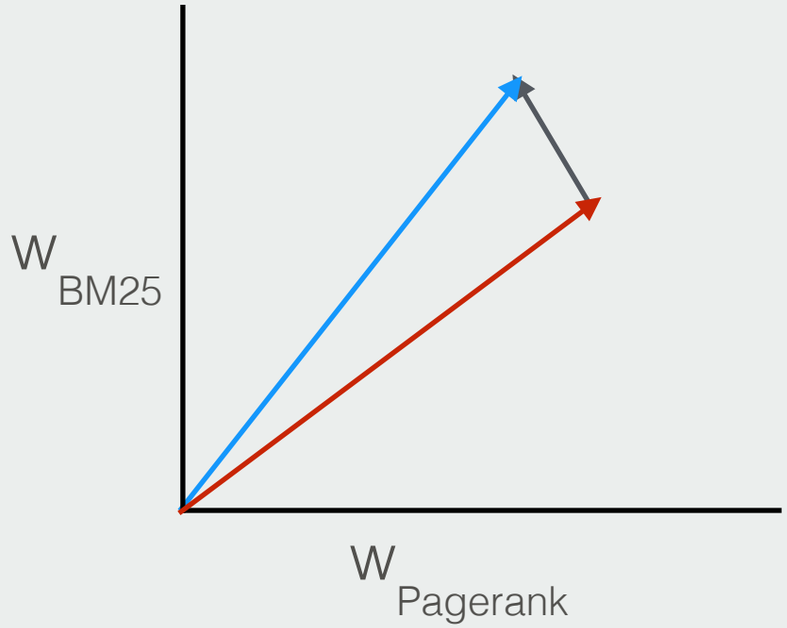
Dueling Bandit Gradient Descent

Exploitative Ranker





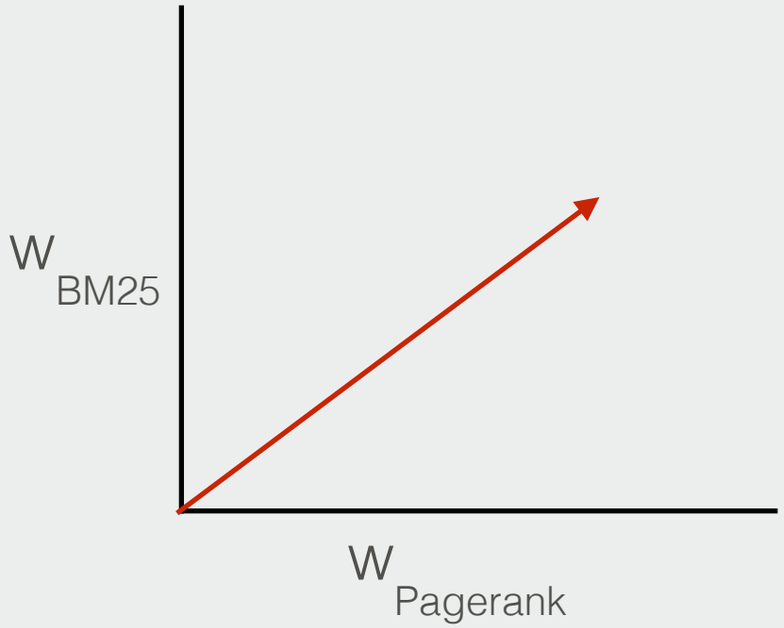
Explorative Ranker

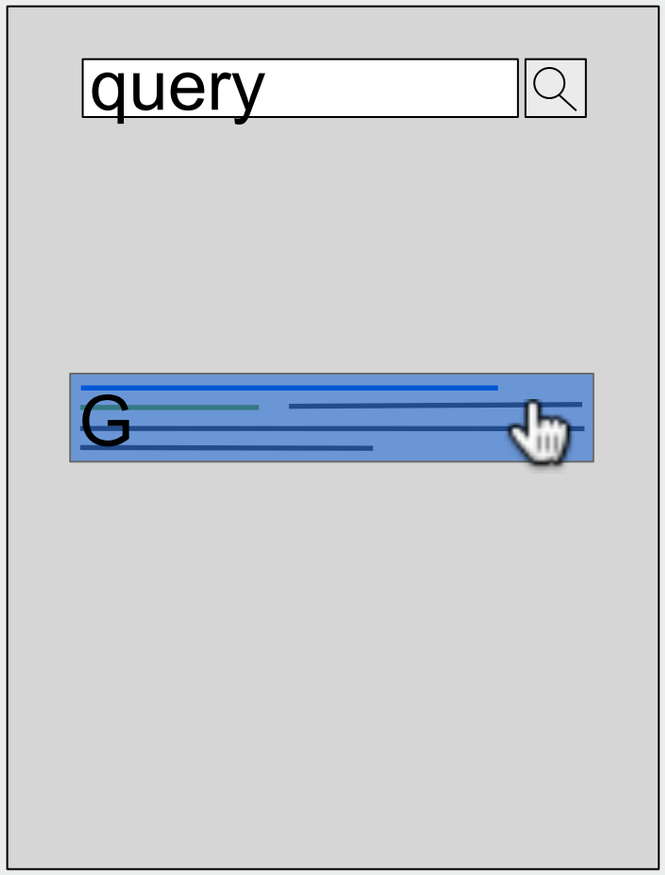


[Yue et al, 2009; Hofmann et al., 2011]

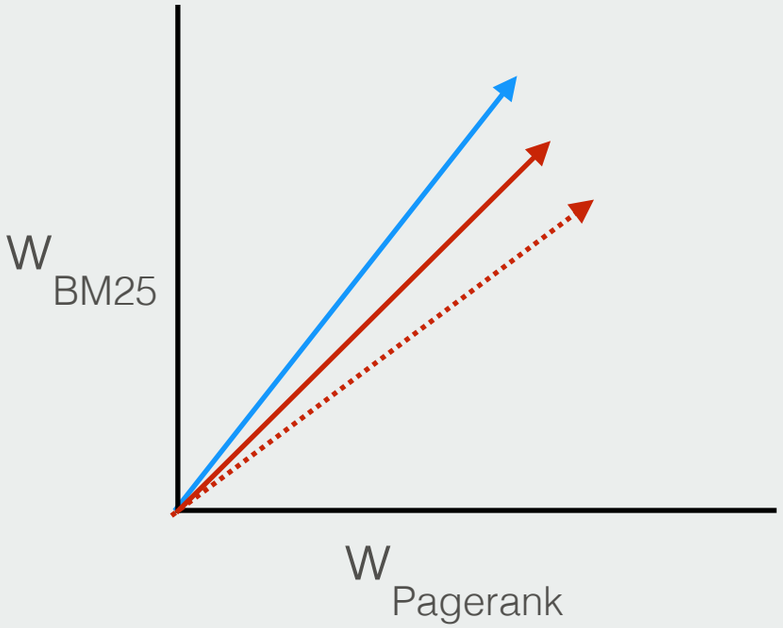
Dueling Bandit Gradient Descent

Exploitative Ranker





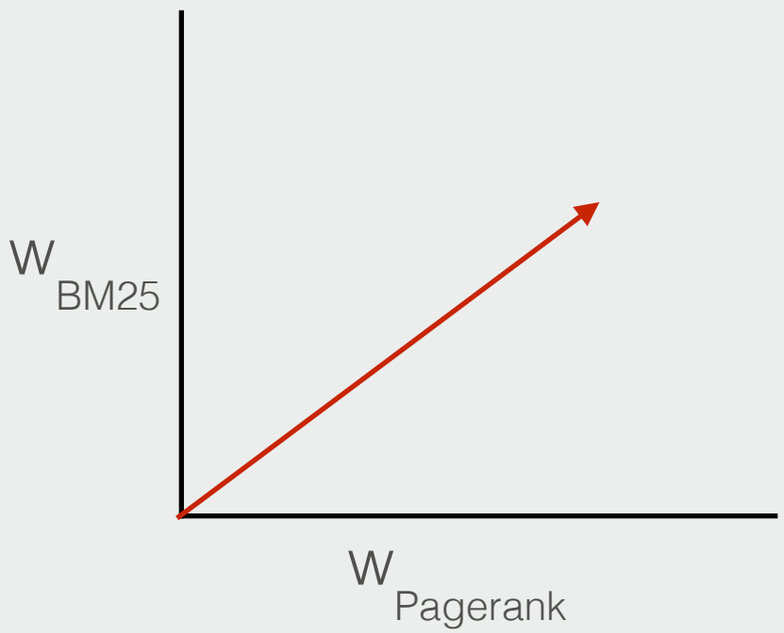
Explorative Ranker

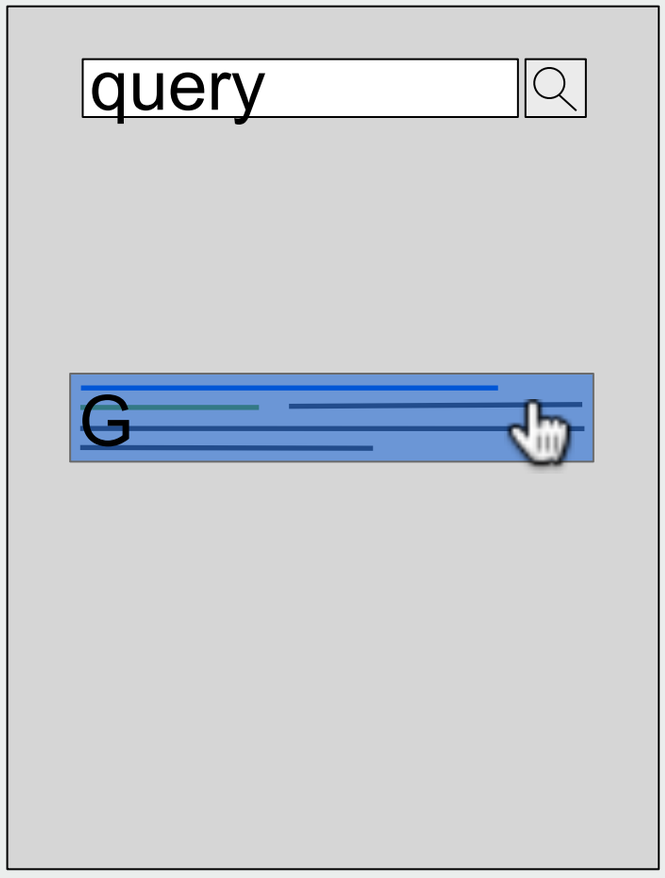


[Yue et al, 2009; Hofmann et al., 2011]

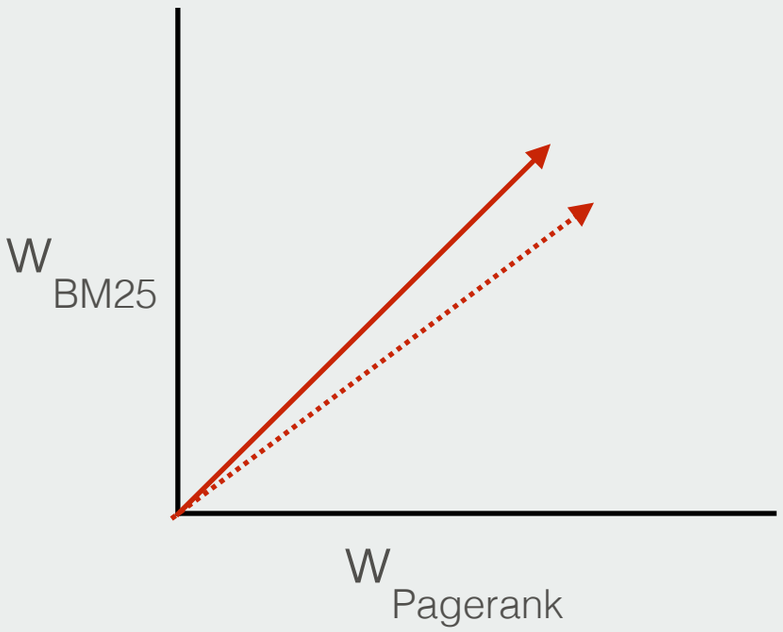
Dueling Bandit Gradient Descent

Exploitative Ranker





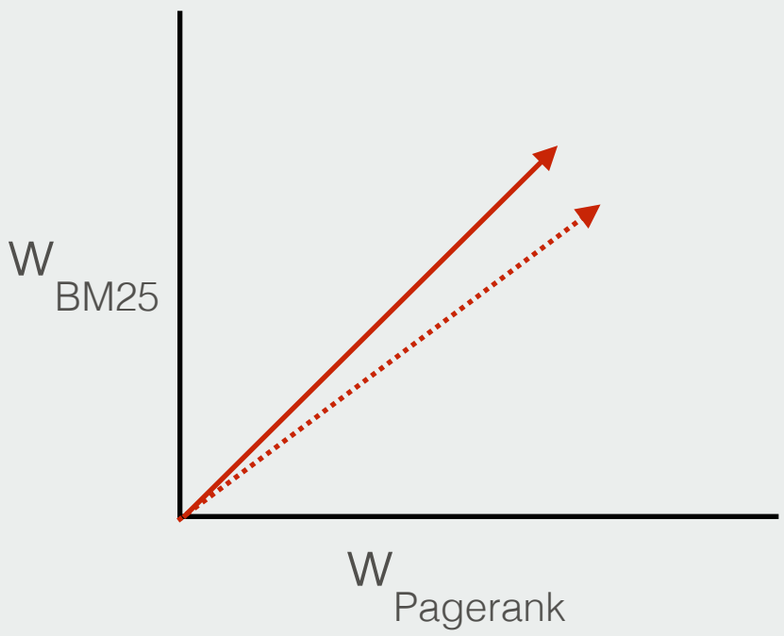
Explorative Ranker

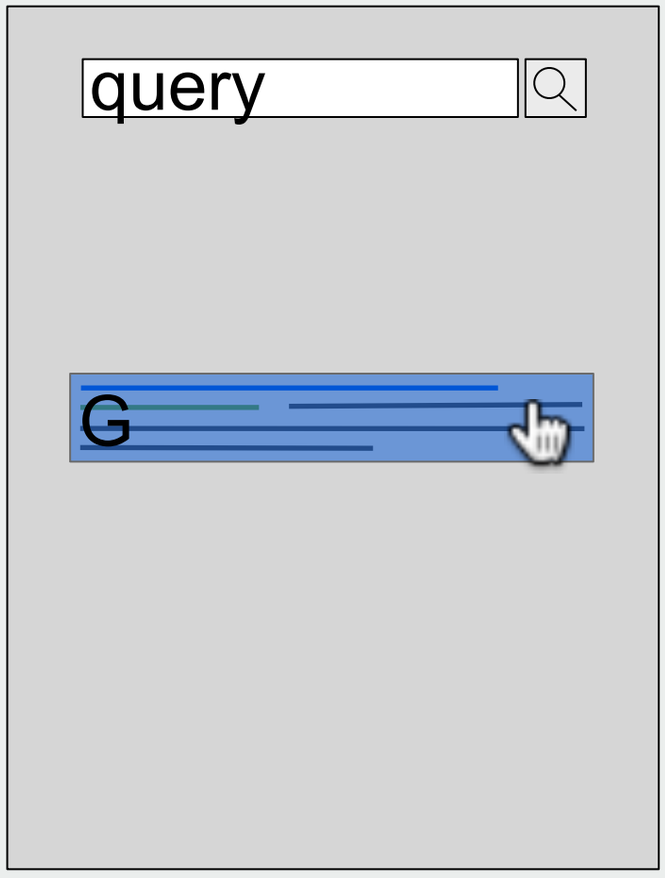


[Yue et al, 2009; Hofmann et al., 2011]

Dueling Bandit Gradient Descent

Exploitative Ranker





[Yue et al, 2009; Hofmann et al., 2011]

Outline

- Information Retrieval
- Online
  - Evaluation
  - Learning to Rank
  - Issues
- Living Labs
- Wrap up

Outline

- Information Retrieval
- Online
  - Evaluation
  - Learning to Rank
  - Issues
- Living Labs
- Wrap up

Issues with Online Evaluation / Learning

Issues with Online Evaluation / Learning

- Overhead of experimental framework

Issues with Online Evaluation / Learning

- Overhead of experimental framework
- Performance metrics are not absolute

Issues with Online Evaluation / Learning

- Overhead of experimental framework
- Performance metrics are not absolute
- Model limitations
  - incrementally updatable (from relative feedback)

Issues with Online Evaluation / Learning

- Overhead of experimental framework
- Performance metrics are not absolute
- Model limitations
  - incrementally updatable (from relative feedback)
- Learning limitations
  - hill climbing (from relative feedback)

Issues with Online Evaluation / Learning

- Overhead of experimental framework
- Performance metrics are not absolute
- Model limitations
  - incrementally updatable (from relative feedback)
- Learning limitations
  - hill climbing (from relative feedback)
- You may hurt a user

Issues with Online Evaluation / Learning

- Overhead of experimental framework
- Performance metrics are not absolute
- Model limitations
  - incrementally updatable (from relative feedback)
- Learning limitations
  - hill climbing (from relative feedback)
- You may hurt a user
- **Users…**
  - **always limiting factor**
  - **no access for researchers**

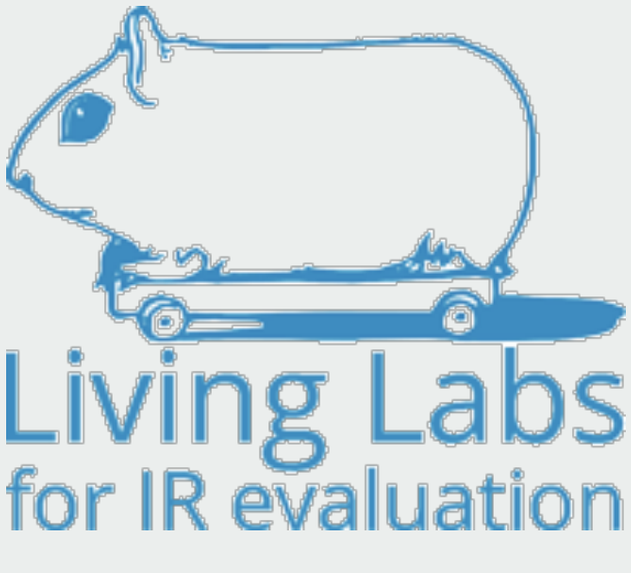
Outline

- Information Retrieval
- Online
  - Evaluation
  - Learning to Rank
  - Issues
- Living Labs
- Wrap up

Outline

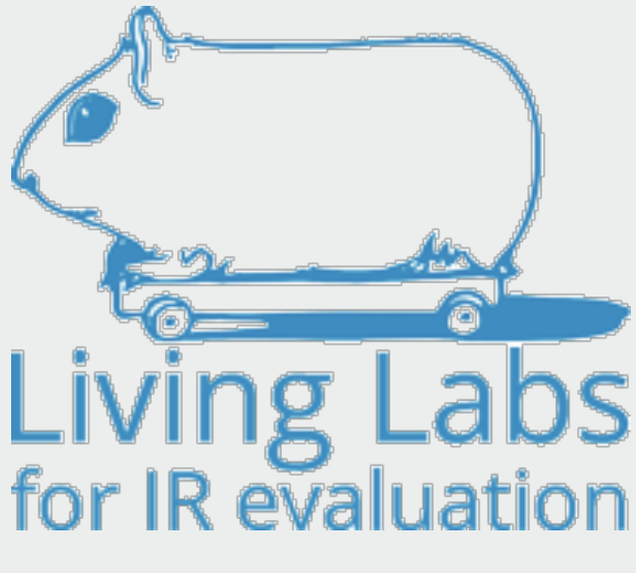
- Information Retrieval
- Online
  - Evaluation
  - Learning to Rank
  - Issues
- Living Labs
- Wrap up

Living Lab for IR Evaluation



- If you are a researcher:
  - “Give us your ranking, we’ll have it clicked!”

Living Lab for IR Evaluation



- If you are a researcher:
  - "Give us your ranking, we'll have it clicked!"

- If you own a search engine:
  - "Give us your clicks, we'll solve your ranking problem!"

Living Lab for IR Evaluation

Living Lab for IR Evaluation

- Sites (real search engines) provide queries
  - sample of ~100 queries
  - per query ~100-1000 candidate documents

Living Lab for IR Evaluation

- Sites (real search engines) provide queries
  - sample of ~100 queries
  - per query ~100-1000 candidate documents
- Participants (researchers) apply their method
  - produce rankings for each query

Living Lab for IR Evaluation

- Sites (real search engines) provide queries
  - sample of ~100 queries
  - per query ~100-1000 candidate documents
- Participants (researchers) apply their method
  - produce rankings for each query
- Site donate traffic
  - rankings from participants are shown to users
  - clicks are fed back

Living Lab for IR Evaluation

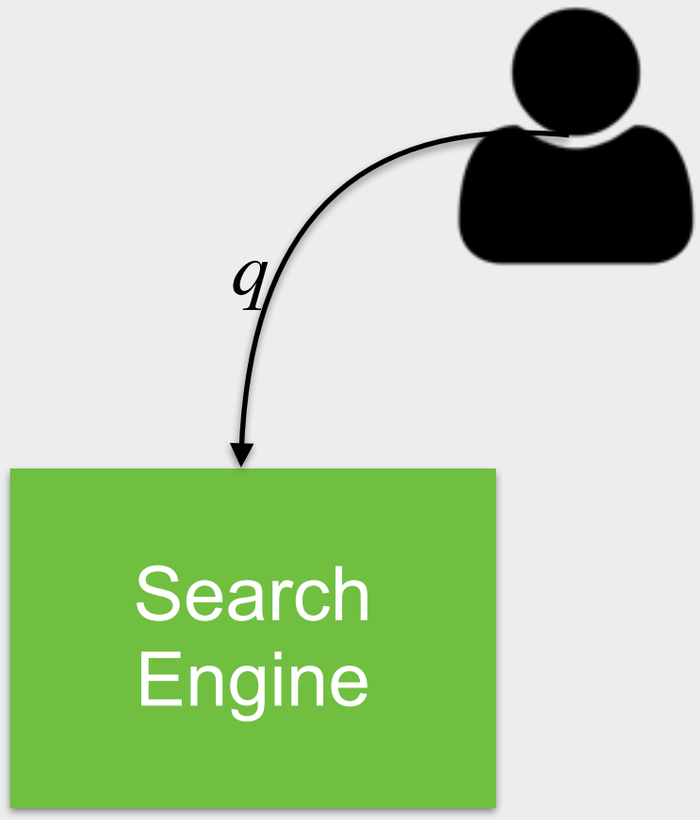
- Sites (real search engines) provide queries
  - sample of ~100 queries
  - per query ~100-1000 candidate documents
- Participants (researchers) apply their method
  - produce rankings for each query
- Site donate traffic
  - rankings from participants are shown to users
  - clicks are fed back
- Participants use clicks on their own rankings
  - to evaluate
  - to improve rankings

Living Lab for IR Evaluation

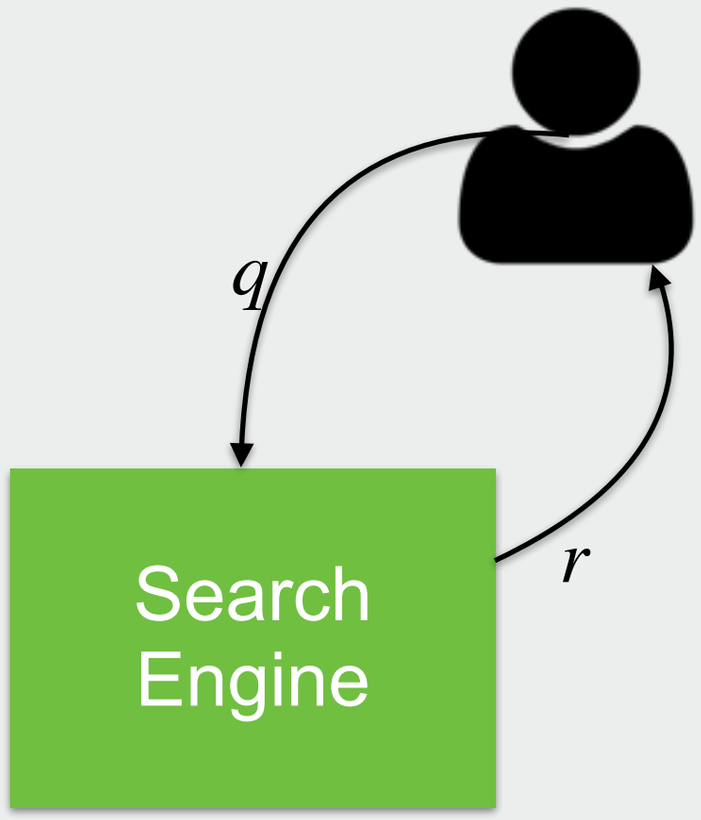
Living Lab for IR Evaluation

Search Engine

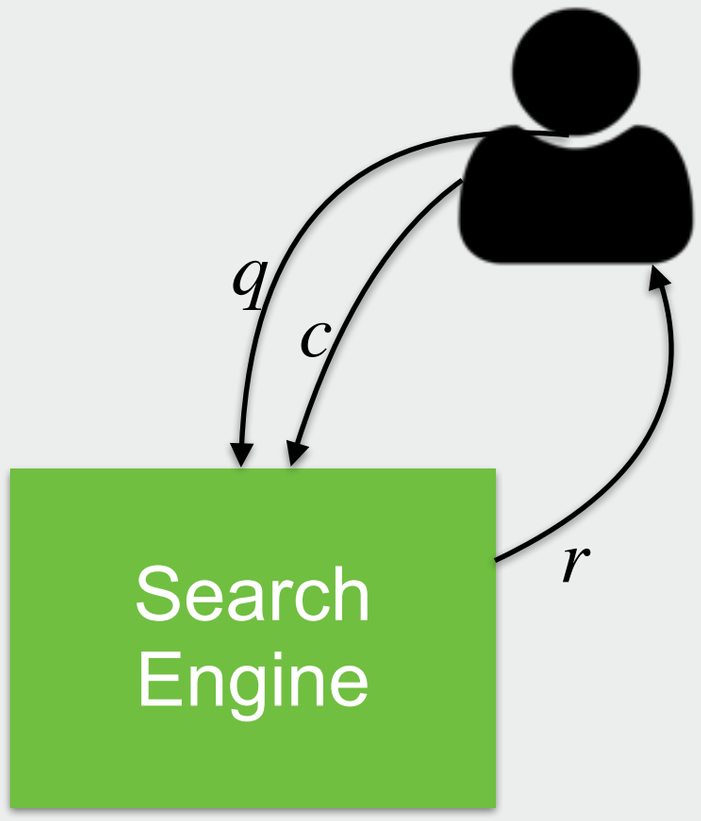
Living Lab for IR Evaluation



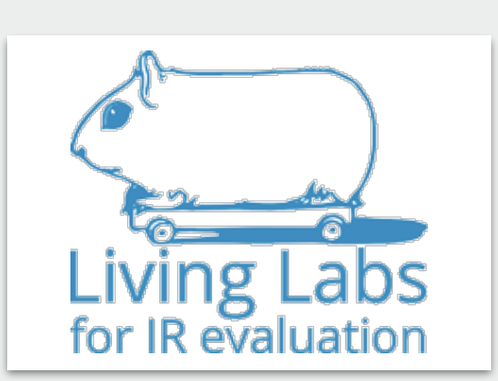
Living Lab for IR Evaluation

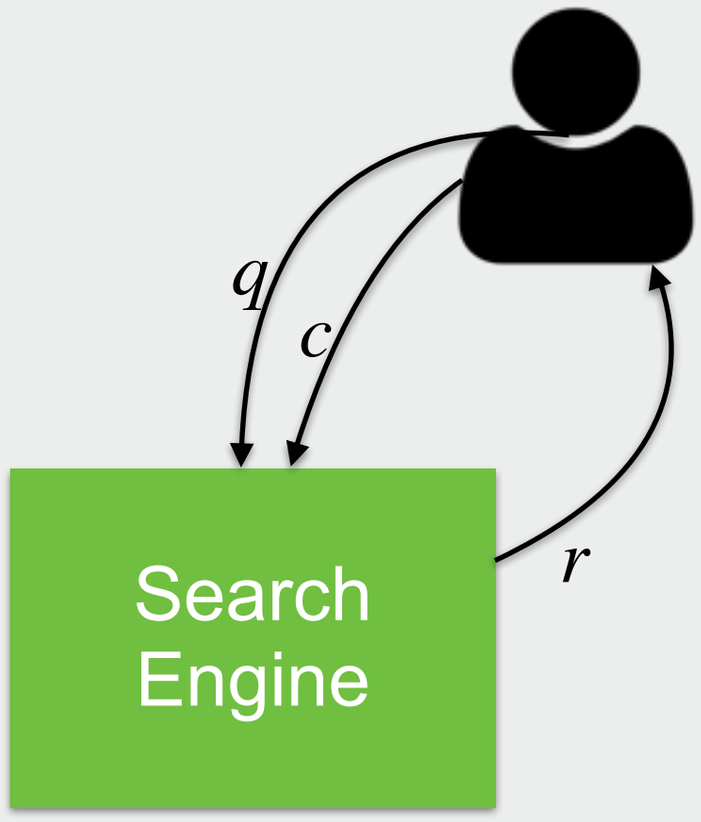


Living Lab for IR Evaluation

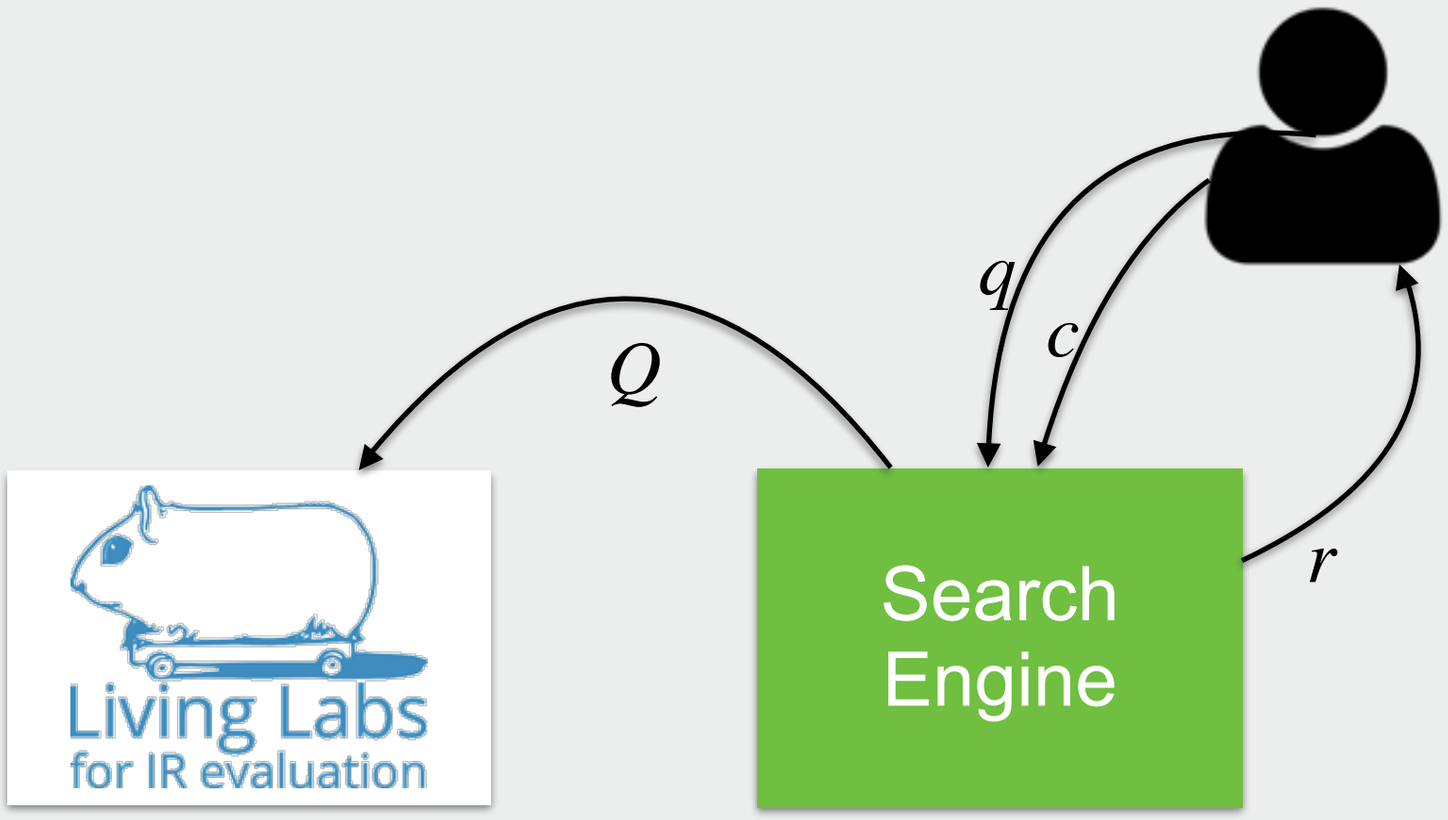


Living Lab for IR Evaluation

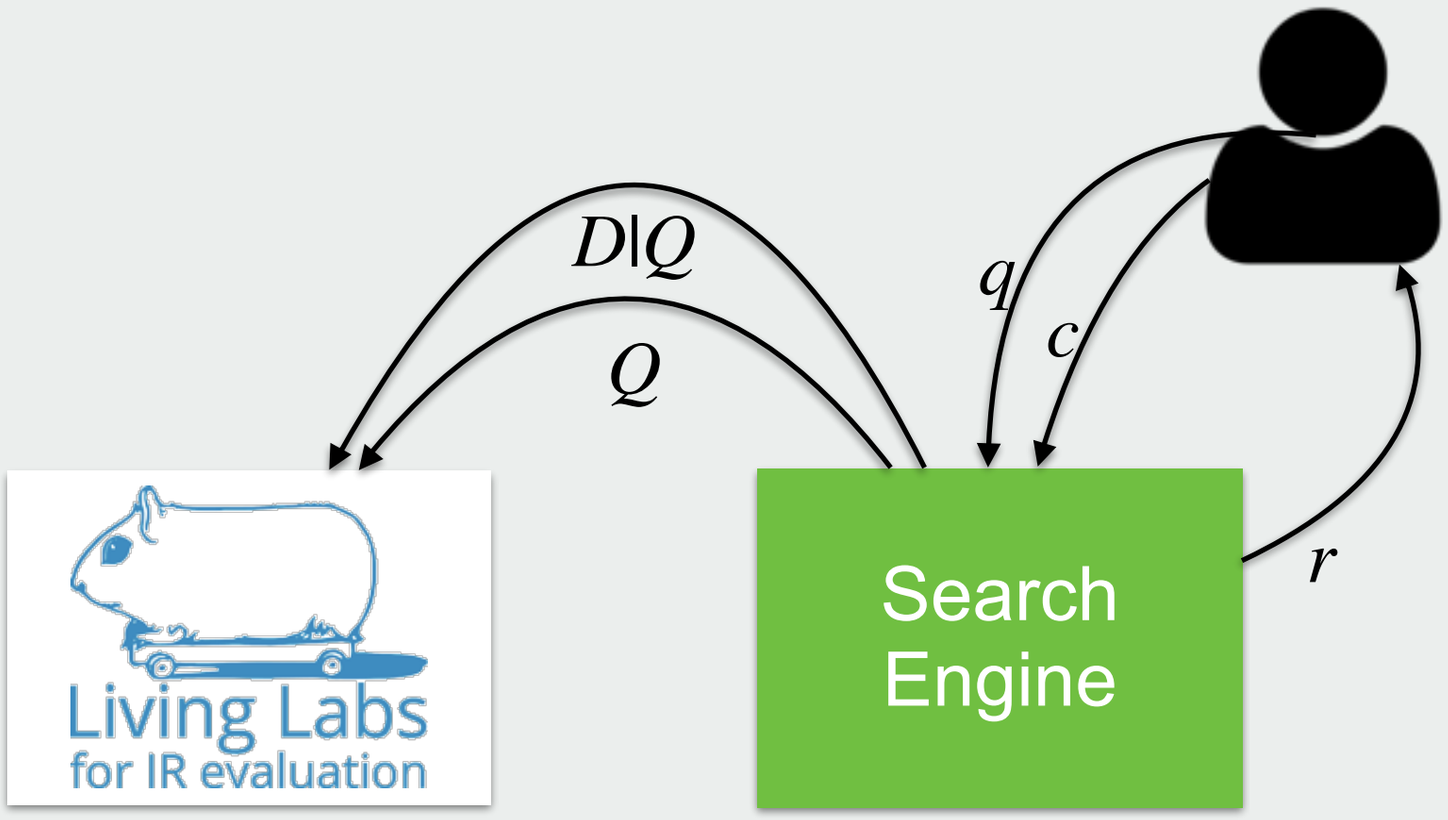




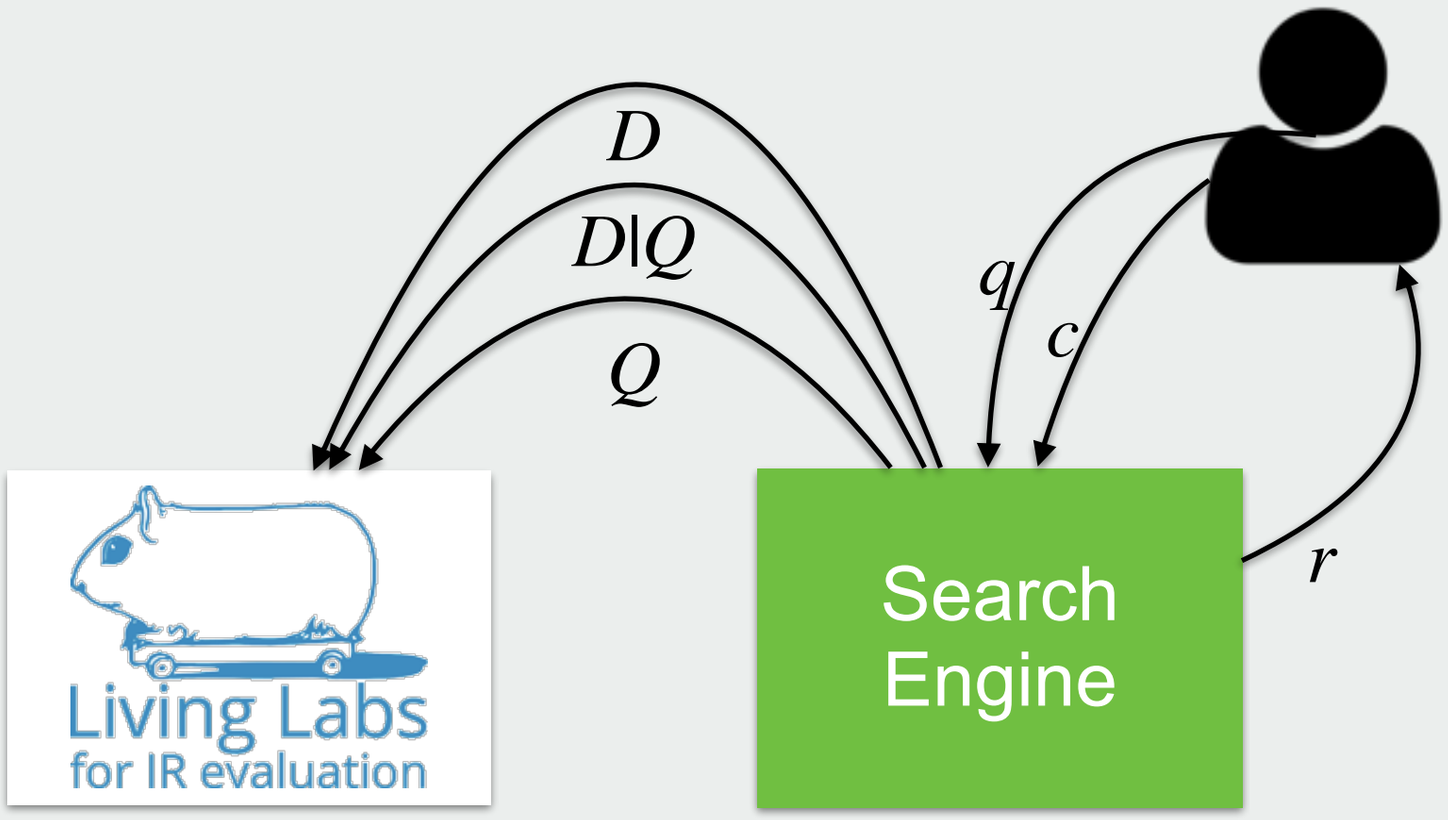
Living Lab for IR Evaluation



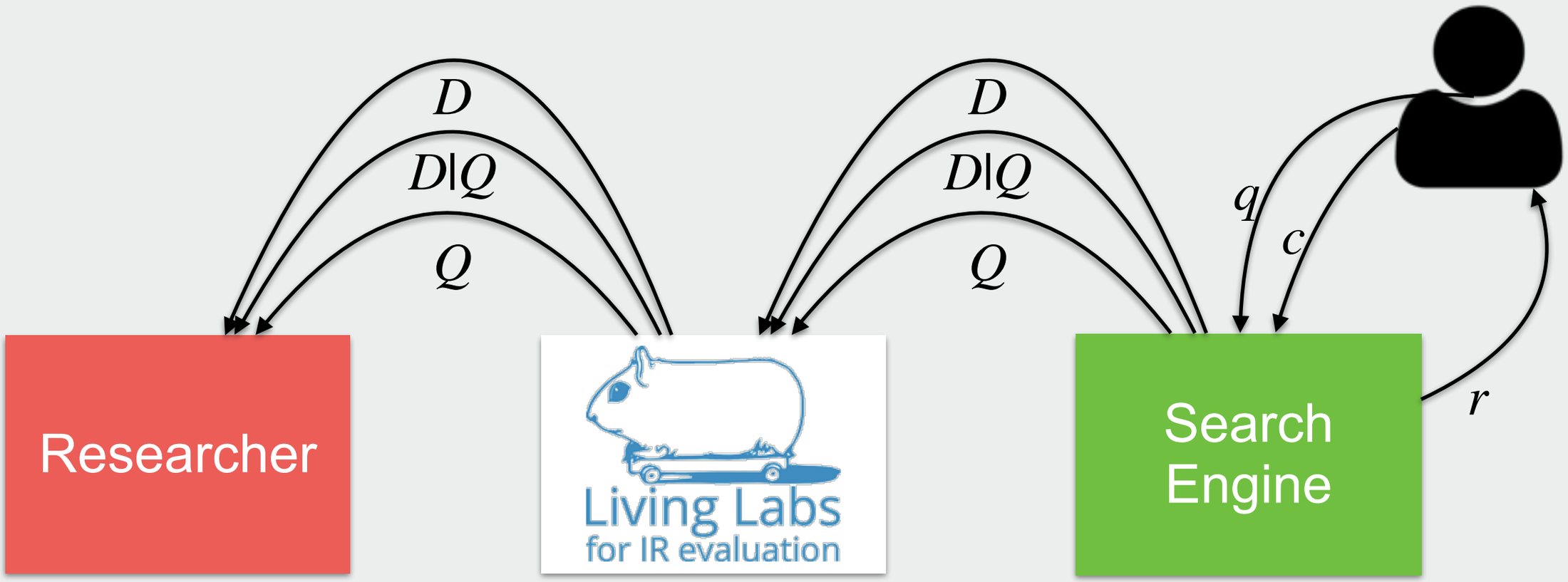
Living Lab for IR Evaluation



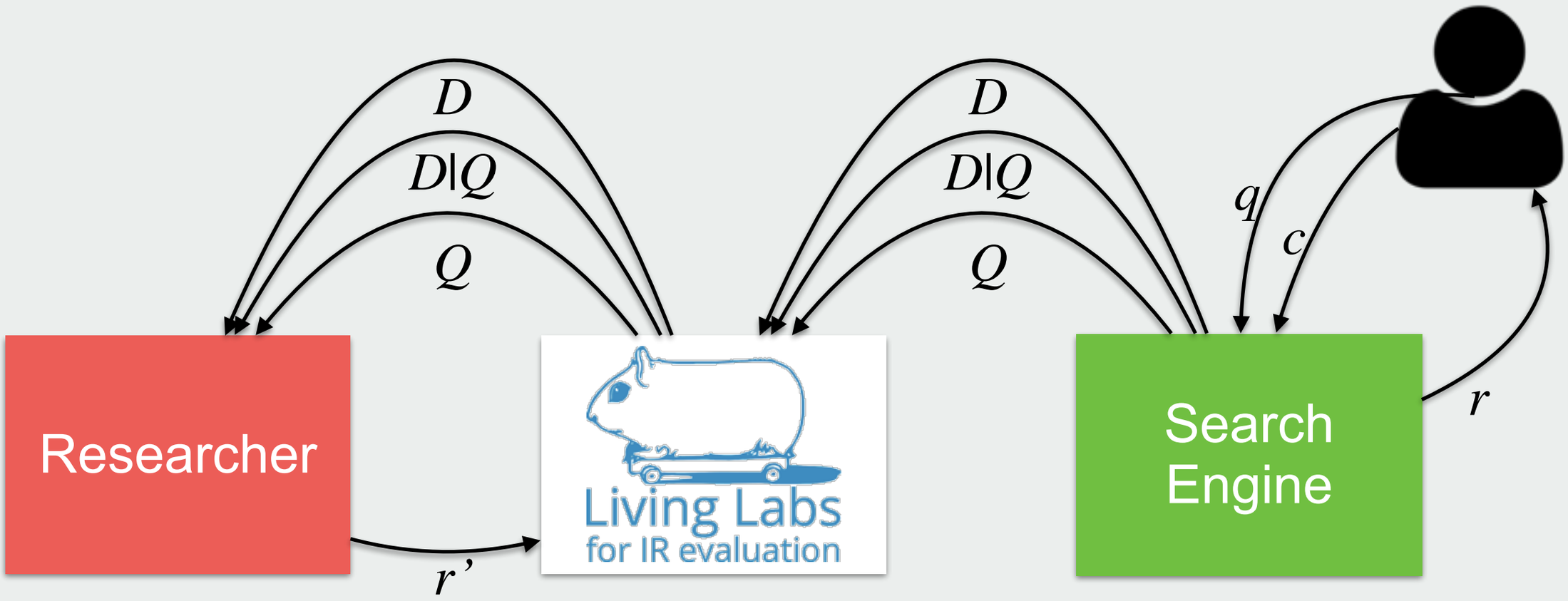
Living Lab for IR Evaluation



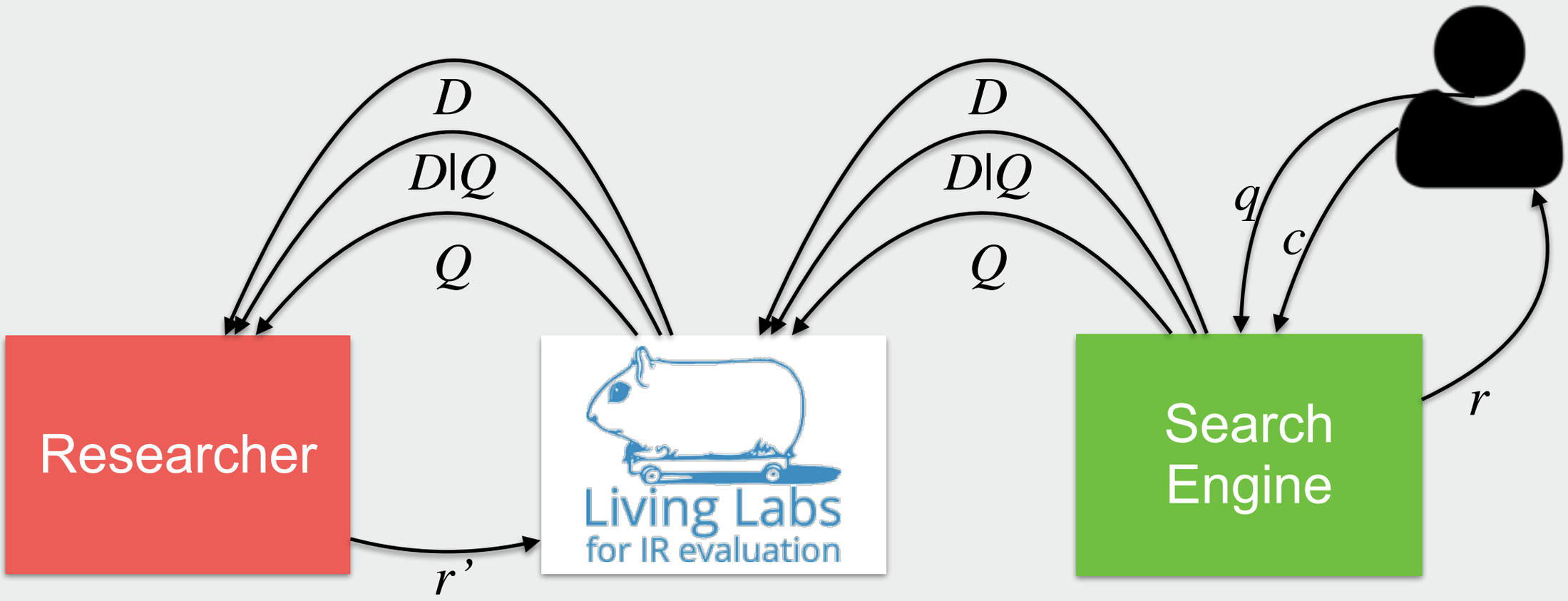
Living Lab for IR Evaluation



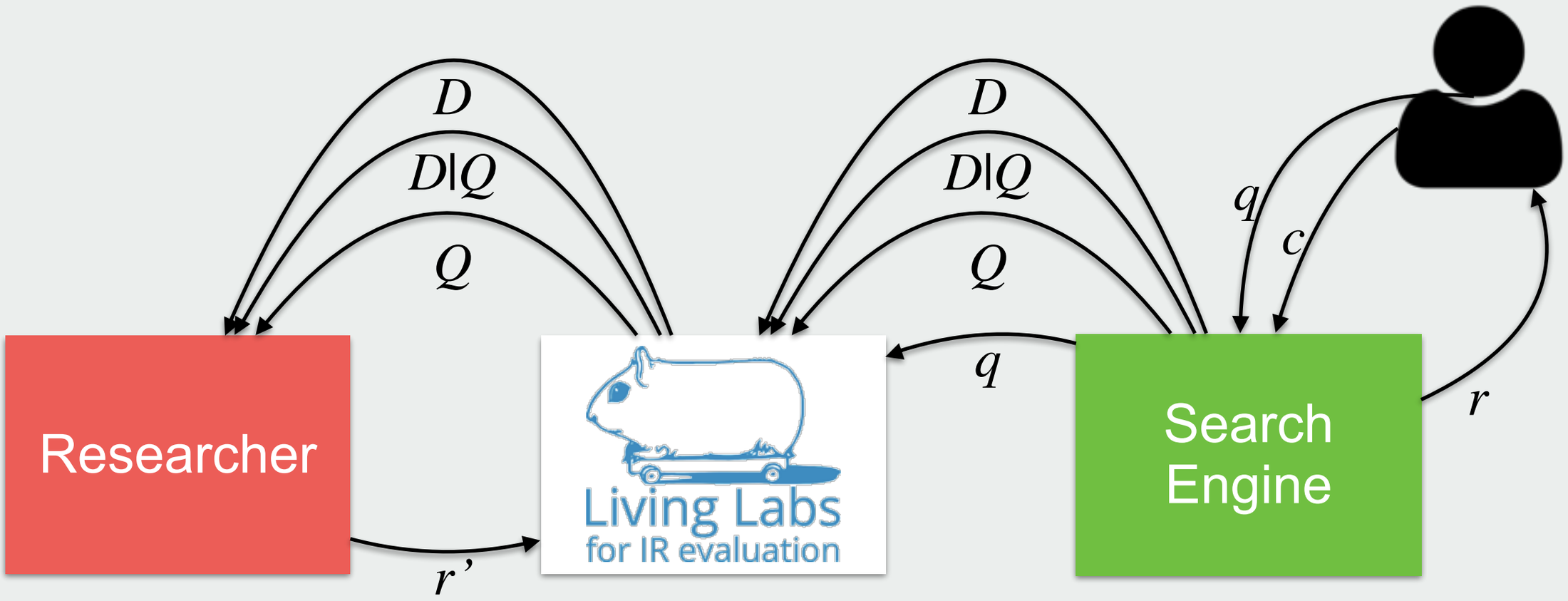
Living Lab for IR Evaluation



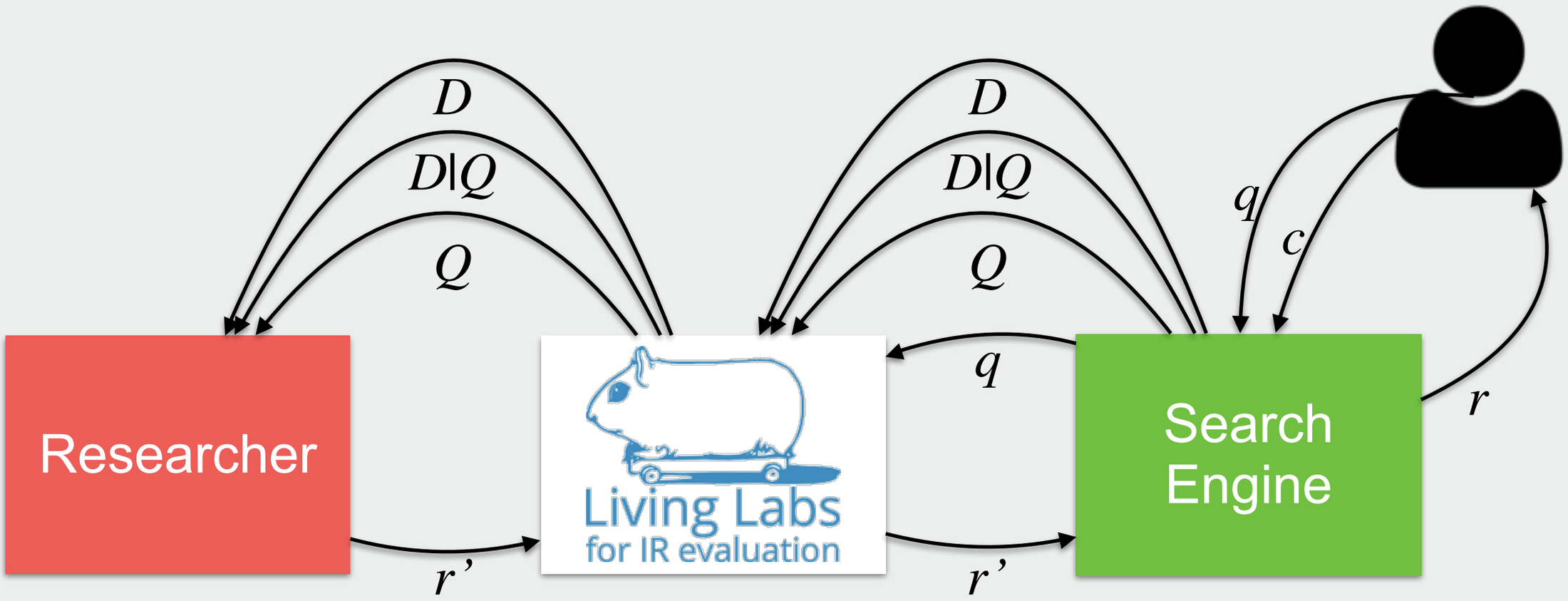
Living Lab for IR Evaluation



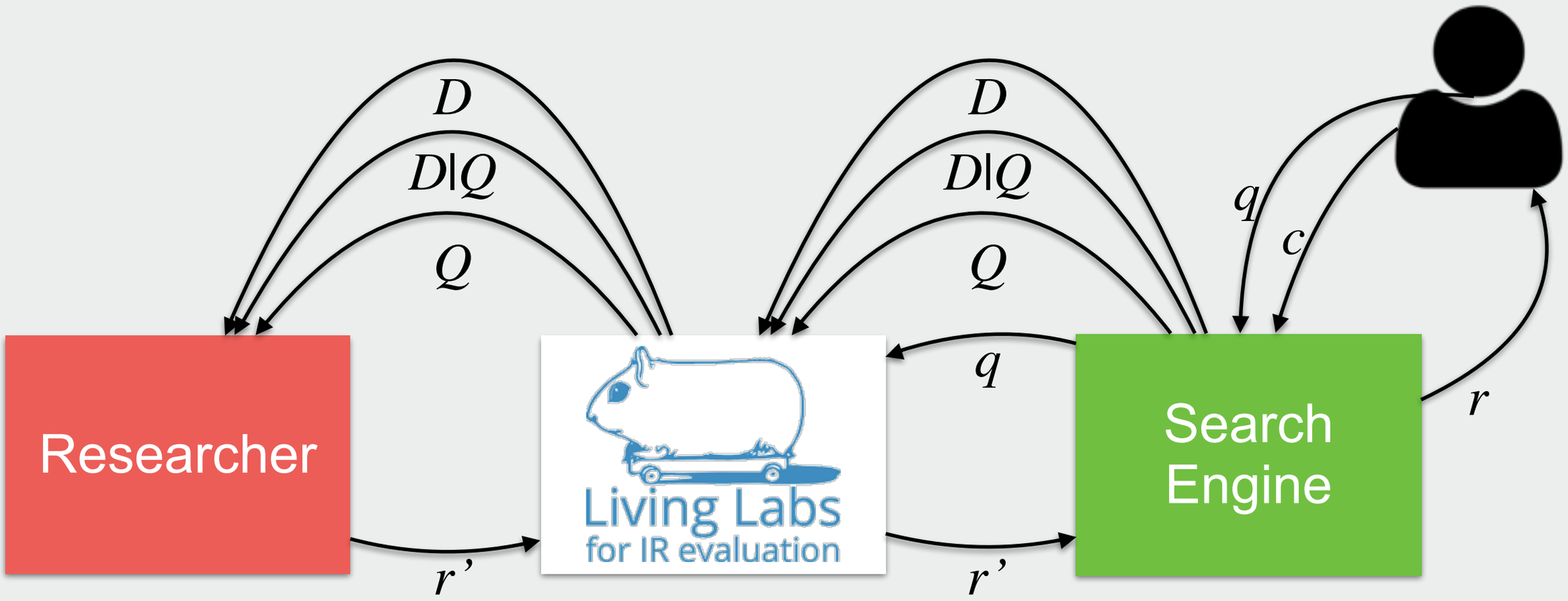
Living Lab for IR Evaluation



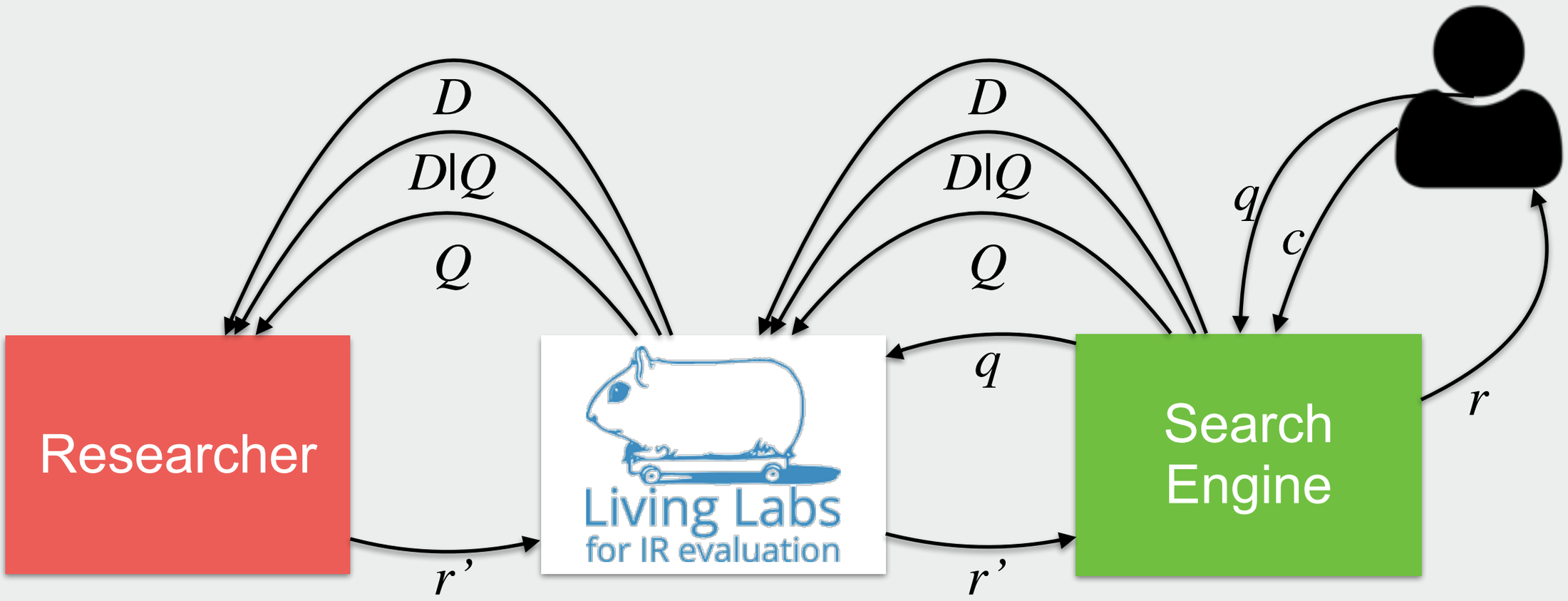
Living Lab for IR Evaluation



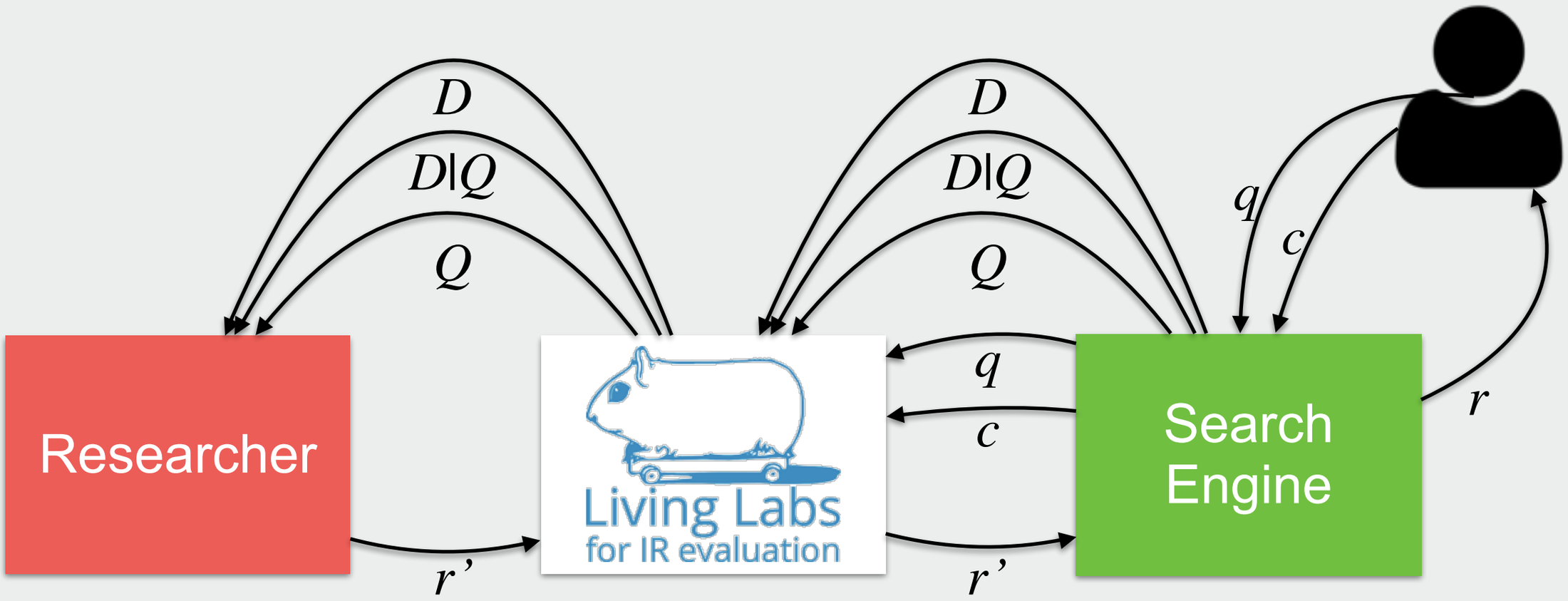
Living Lab for IR Evaluation



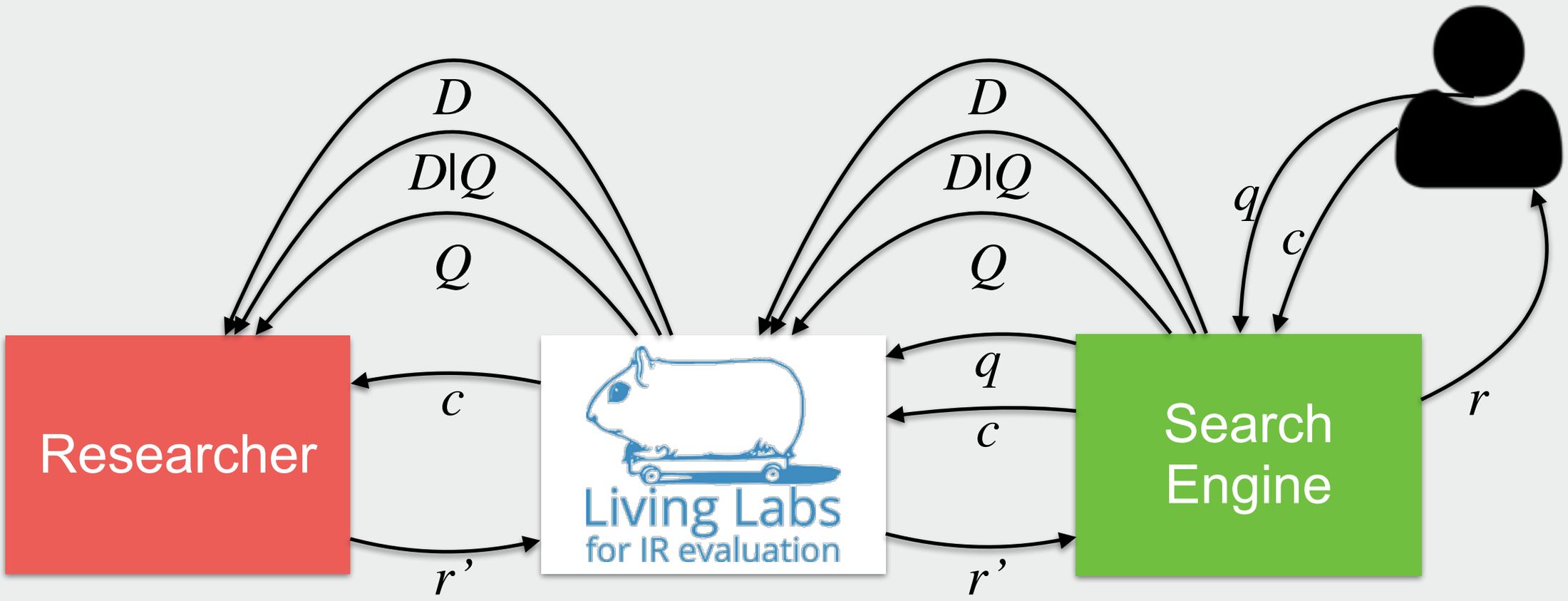
Living Lab for IR Evaluation



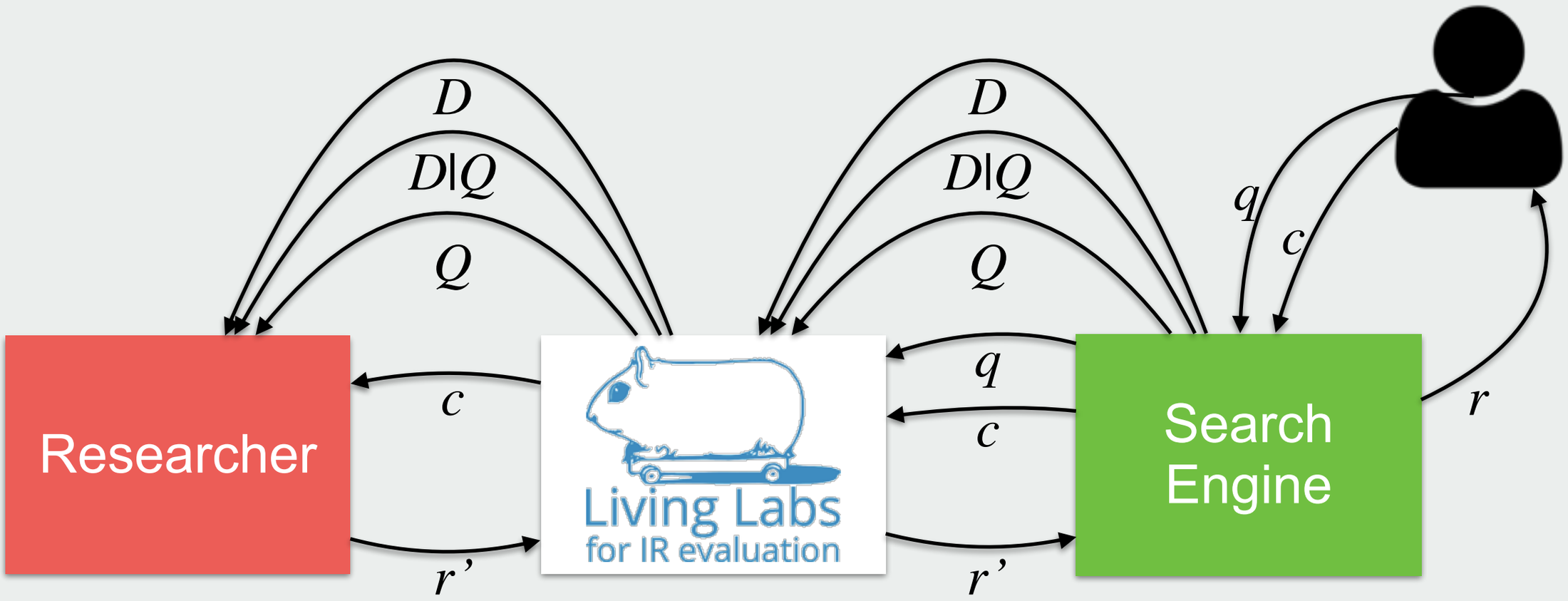
Living Lab for IR Evaluation



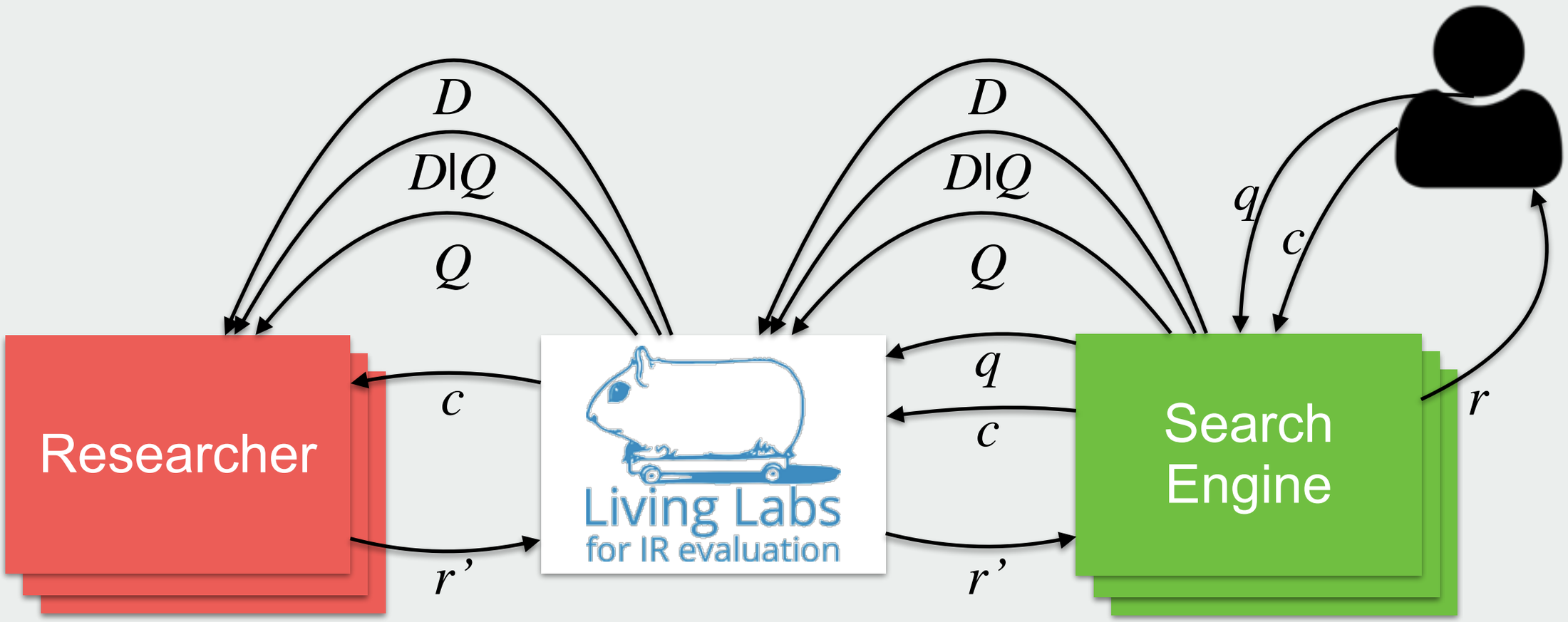
Living Lab for IR Evaluation



Living Lab for IR Evaluation



Living Lab for IR Evaluation



Living Lab for IR Evaluation

Living Lab for IR Evaluation

- Overcome lack of users for researchers

Living Lab for IR Evaluation

- Overcome lack of users for researchers
- Provide search engines with researchers

Living Lab for IR Evaluation

- Overcome lack of users for researchers
- Provide search engines with researchers
- Implementation
  - REST API

Living Lab for IR Evaluation

- Overcome lack of users for researchers
- Provide search engines with researchers
- Implementation
  - REST API
- Collaborations with:
  - Seznam (http://www.seznam.cz)
  - UvA (http://www.uva.nl)
  - Webshop

Living Lab for IR Evaluation

- Overcome lack of users for researchers
- Provide search engines with researchers
- Implementation
  - REST API
- Collaborations with:
  - Seznam (http://www.seznam.cz)
  - UvA (http://www.uva.nl)
  - Webshop
- Runs as a CLEF Lab
  - http://living-labs.net/clef-lab/

Outline

- Information Retrieval
- Online
  - Evaluation
  - Learning to Rank
  - Issues
- Living Labs
- Wrap up

Outline

- Information Retrieval
- Online
  - Evaluation
  - Learning to Rank
  - Issues
- Living Labs
- Wrap up

Wrap up

Wrap up

- Online evaluation
  - Clicks are biased and noisy
  - But clicks contain relative preferences
  - Interleaving provides reliable relative feedback

Wrap up

- Online evaluation
  - Clicks are biased and noisy
  - But clicks contain relative preferences
  - Interleaving provides reliable relative feedback

- Online Learning
  - Learning from interleaving
  - With the user in the loop

Wrap up

- Online evaluation
  - Clicks are biased and noisy
  - But clicks contain relative preferences
  - Interleaving provides reliable relative feedback

- Online Learning
  - Learning from interleaving
  - With the user in the loop

- Living Labs
  - Users for researchers
  - Researchers for search engines

Thanks

Living Labs for IR evaluation

living-labs.net/clef-lab

Lerot

bitbucket.org/ilps/lerot

anneschuth.nl
anne.schuth@uva.nl
@anneschuth

References

- Buscher, G. (2013). IR Evaluation: Perspectives From Within a Living Lab.
- Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008). An experimental comparison of click position-bias models. In WSDM'08.
- Hassan, A., Shi, X., Craswell, N., & Ramsey, B. (2013). Beyond Clicks: Query Reformulation as a Predictor of Search Satisfaction. In CIKM'13.
- Hofmann, K. (2013). Fast and Reliably Online Learning to Rank for Information Retrieval.
- Hofmann, K., **Schuth, A**., Whiteson, S., & de Rijke, M. (2013). Reusing Historical Interaction Data for Faster Online Learning to Rank for IR. In WSDM'13.
- Hofmann, K., **Schuth, A.**, Bellogin, A., & de Rijke, M (2014): Effects of Position Bias on Click-Based Recommender Evaluation. In ECIR'14.
- Hofmann, K., Whiteson, S., & de Rijke, M. (2011). Balancing Exploration and Exploitation in Learning to Rank Online. In ECIR'11.
- Joachims, T. (2002). Optimizing search engines using clickthrough data. In KDD'02.
- Kohavi, R. (2013). Online Controlled Experiments. SIGIR'13.
- Radlinski, F., Kurup, M., & Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In CIKM'08.
- **Schuth, A.**, Hofmann, K., Whiteson, S., & de Rijke, M. (2013). Lerot: an Online Learning to Rank Framework. In LivingLab'13.
- **Schuth, A.**, Sietsma, F., Whiteson, S., Lefortier, S., & de Rijke, M (2014): Multileaved Comparisons for Fast Online Evaluation. In CIKM'14.
- Yue, Y., & Joachims, T. (2009). Interactively optimizing information retrieval systems as a dueling bandits problem. In ICML '09.