Living Labs for IR Evaluation

#LL4IR

Overview

Anne Schuth
University of Amsterdam

Krisztian Balog
University of Stavanger

Liadh Kelly
Trinity College Dublin

Living Labs
for IR evaluation

http://living-labs.net
@livinglabsnet

the lab

use cases

conclusions

Living Labs for IR

- **New** lab

Living Labs for IR

- **New** lab
- **Realistic** IR **evaluation**

Living Labs for IR

- **New** lab
- **Realistic** IR **evaluation**
- Exposing experimental systems to **real users**

Living Labs for IR

- **New** lab
- **Realistic** IR **evaluation**
- Exposing experimental systems to **real users**
  - **Unsuspecting** users

Living Labs for IR

- **New** lab
- **Realistic** IR **evaluation**
- Exposing experimental systems to **real users**
  - **Unsuspecting** users
  - Users performing **real tasks**

Living Labs for IR

- **New** lab
- **Realistic** IR **evaluation**
- Exposing experimental systems to **real users**
  - **Unsuspecting** users
  - Users performing **real tasks**
  - Users issuing **real queries**

API

API

Bruxelles – Wikislovník

[illegible]

bruxelles » Encyklopedie seznam.cz

[illegible]

Praha - Bruxelles, [illegible] - vlakem.net

[illegible]

Hotel a Bruxelles, Alloggi ... a Bruxelles, Belgio ...

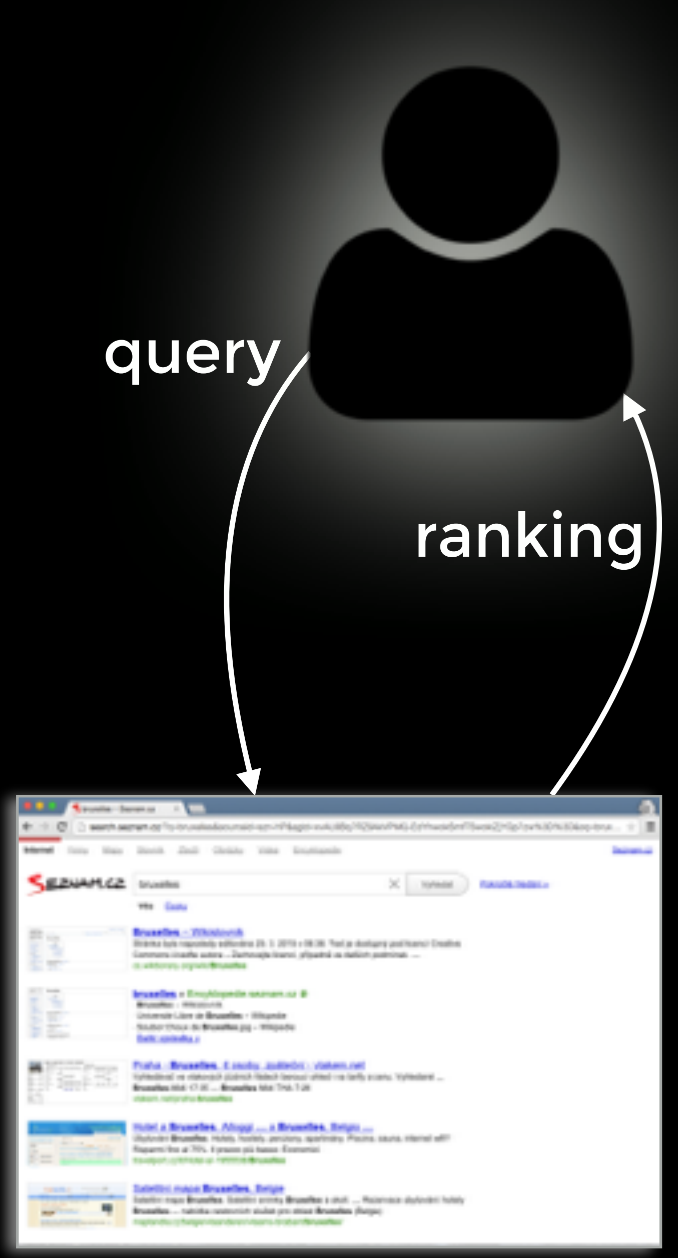
[illegible]

Satelitní mapa Bruxelles, Belgie

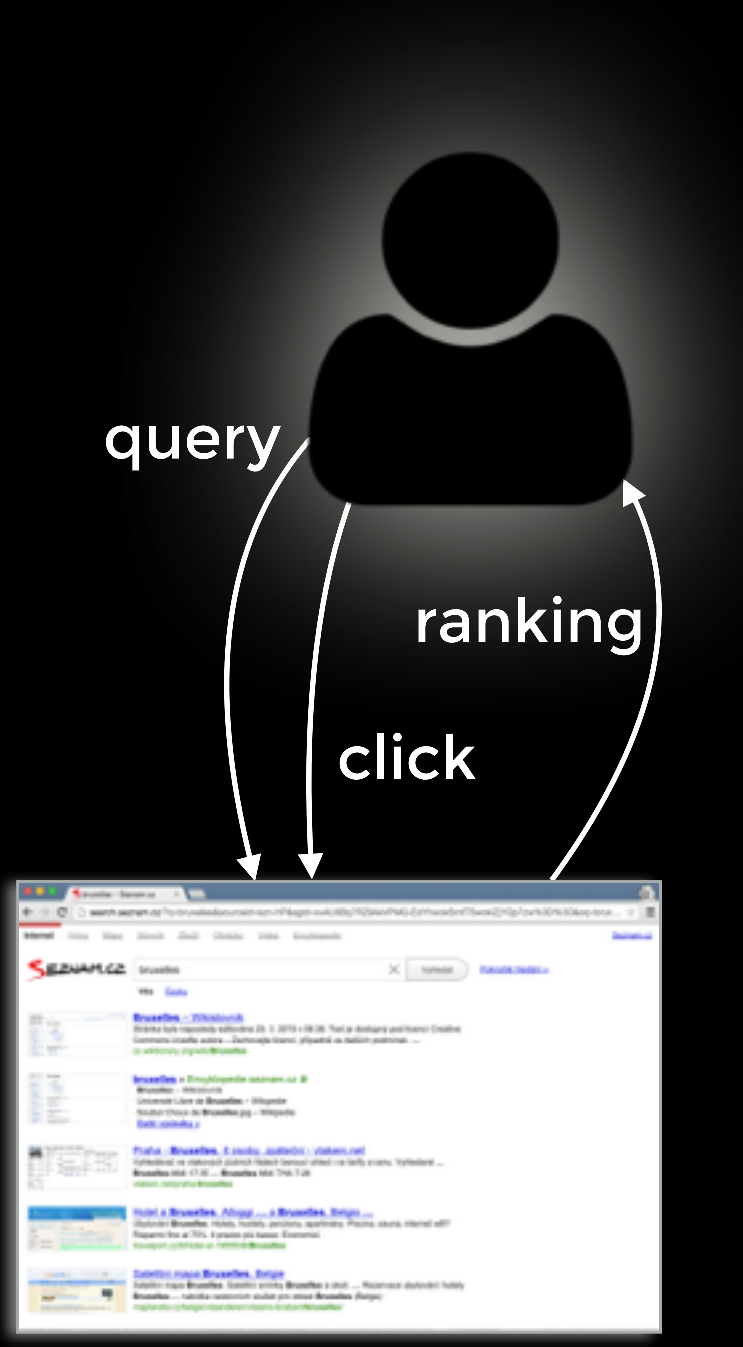
[illegible]

API

API

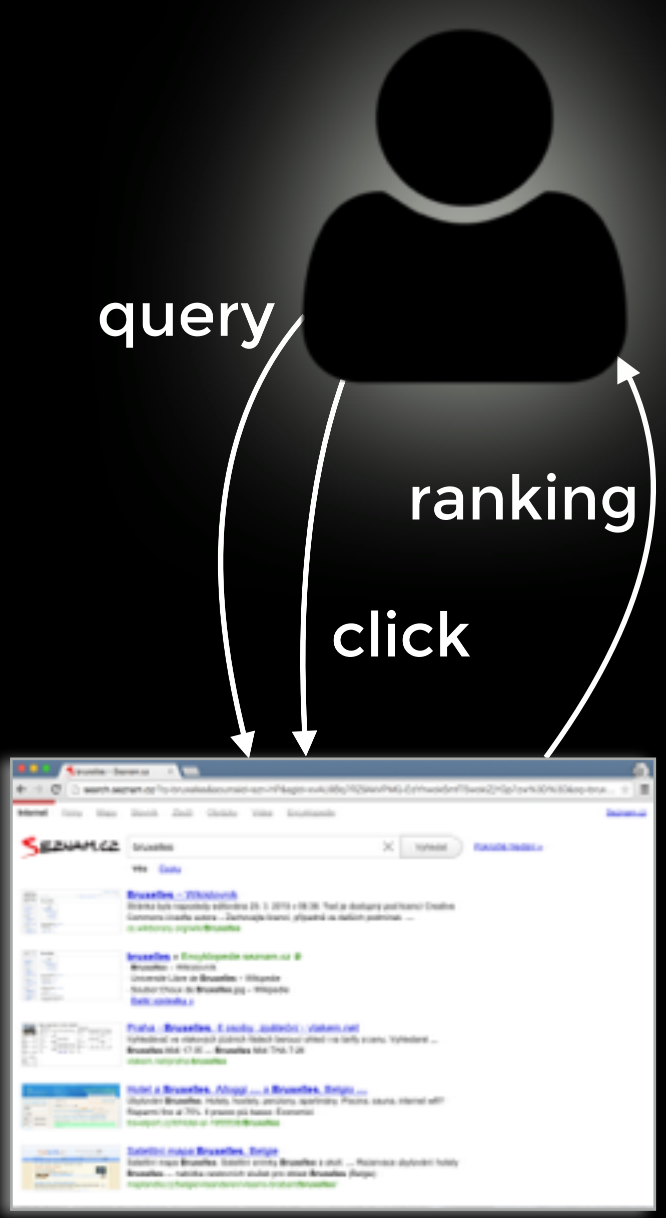


API

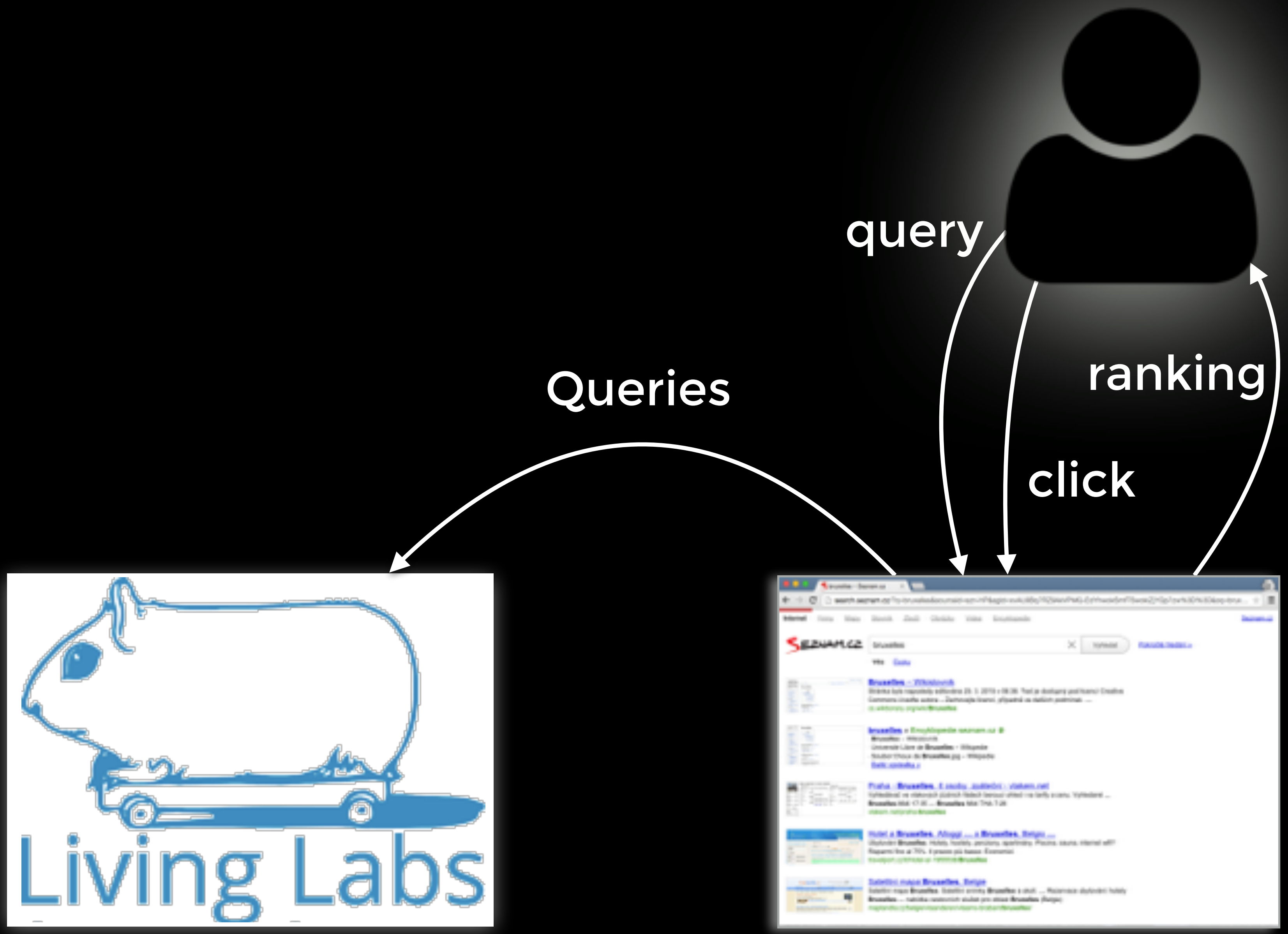


API

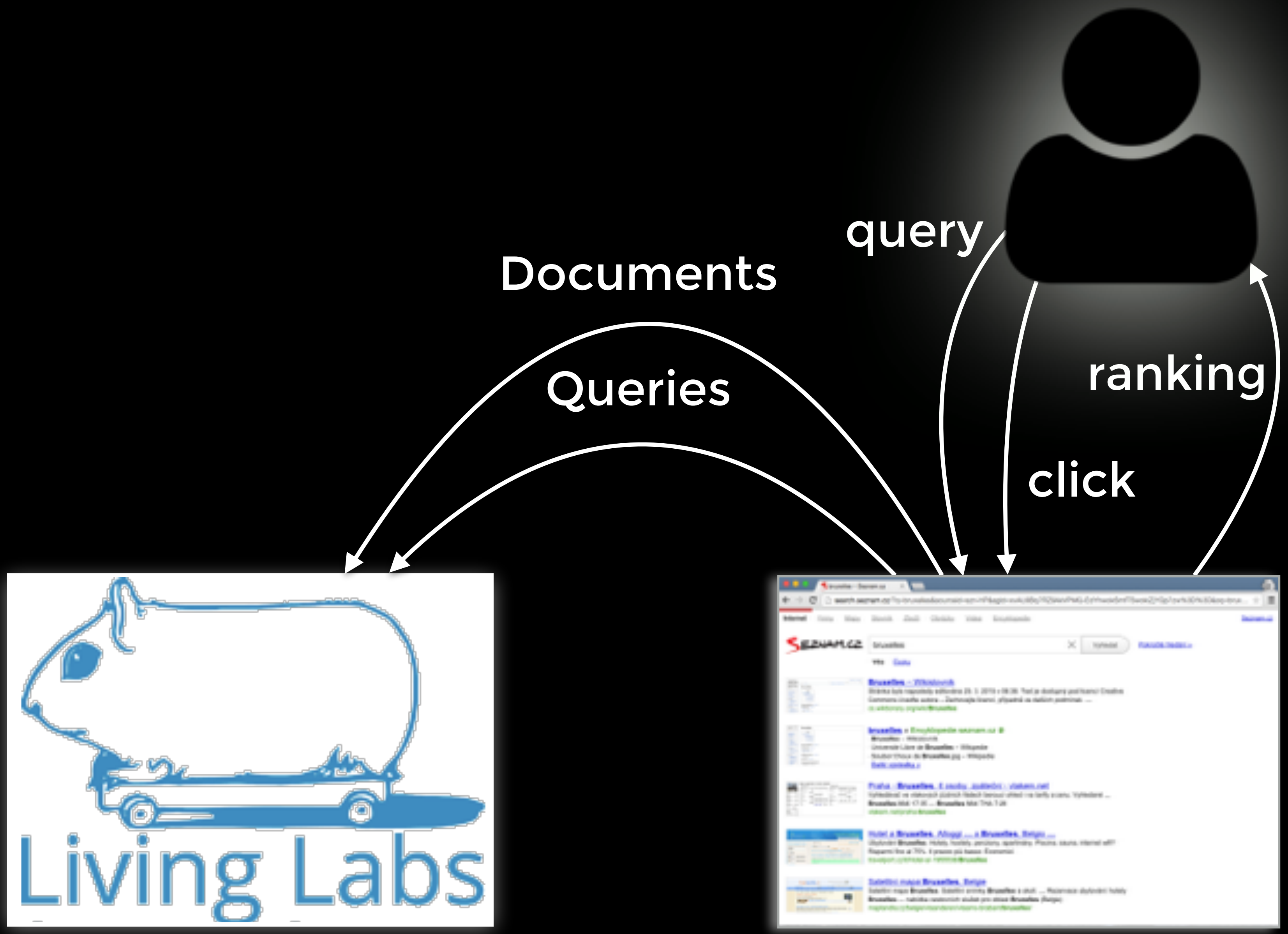
Living Labs



API



API



query

Documents

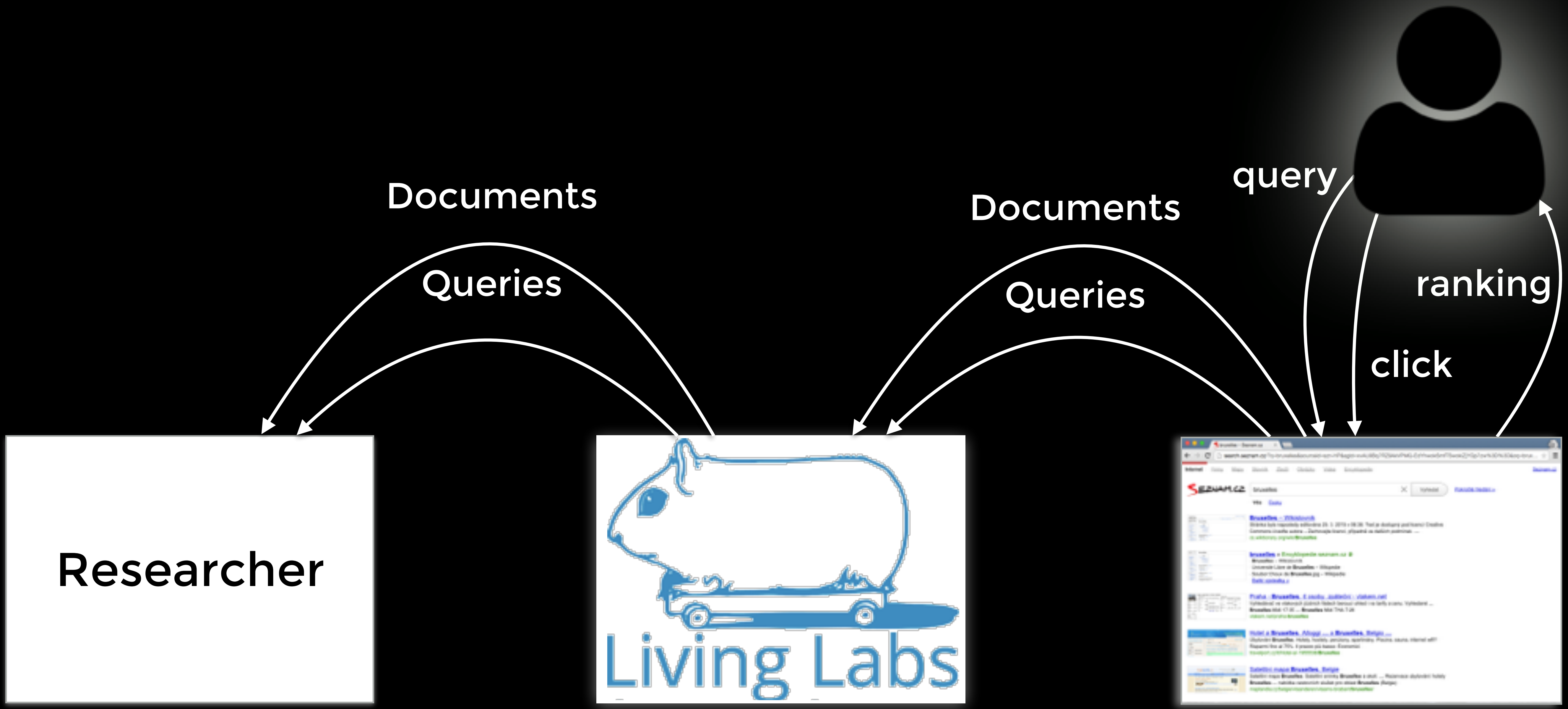
Queries

ranking

click

Living Labs

API



Documents

Queries

Documents

Queries

query

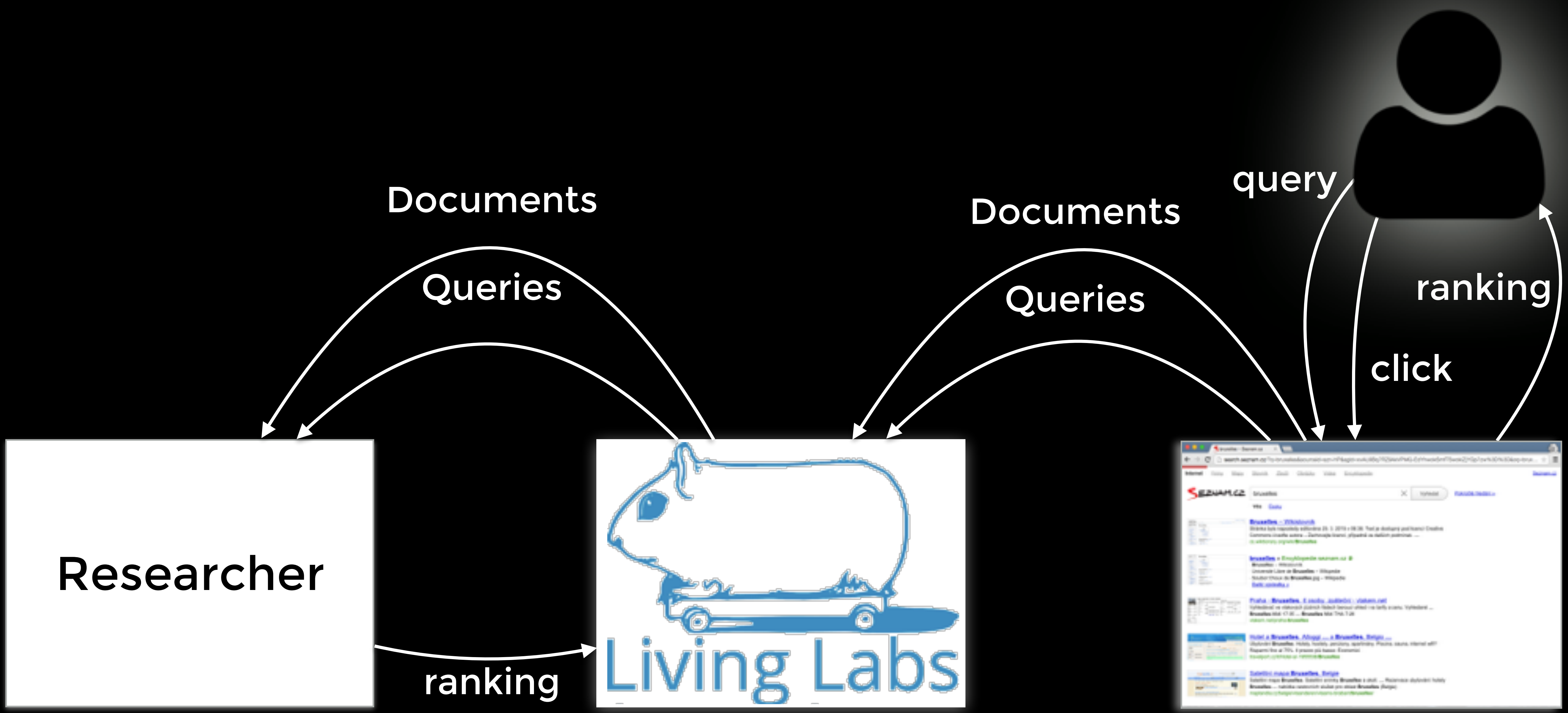
ranking

click

Researcher

Living Labs

API



Documents

Queries

Researcher

ranking

Living Labs

Documents

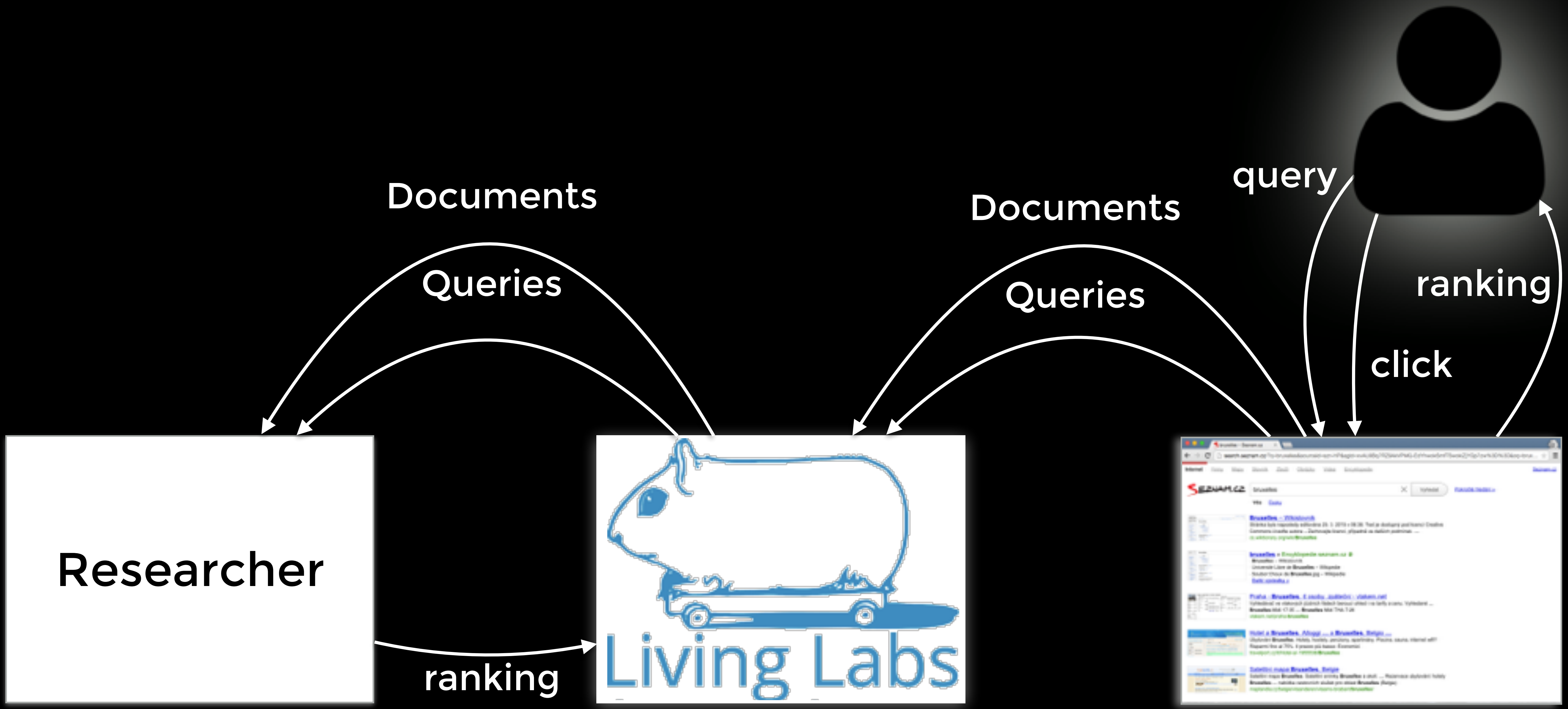
Queries

query

ranking

click

API



Documents

Queries

Researcher

ranking

Living Labs

Documents

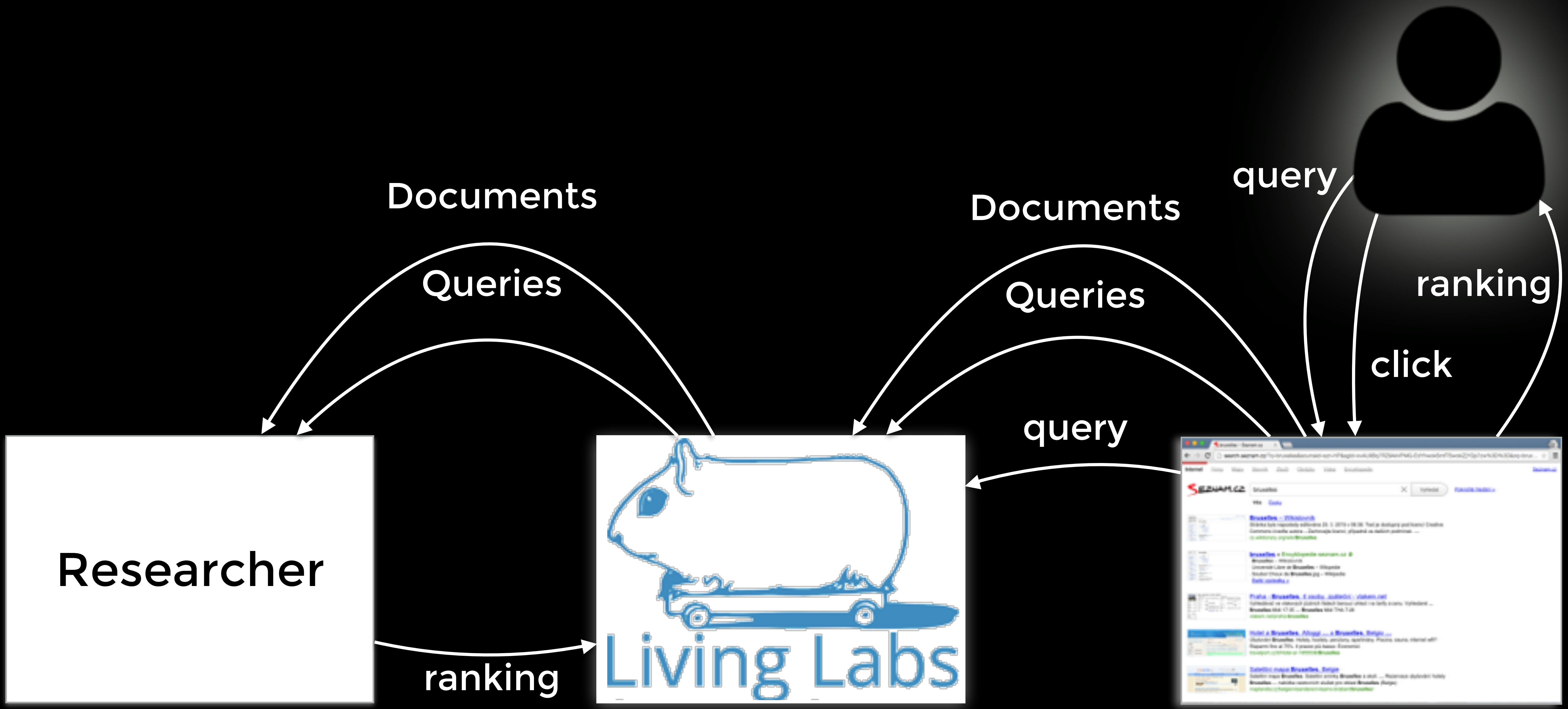
Queries

query

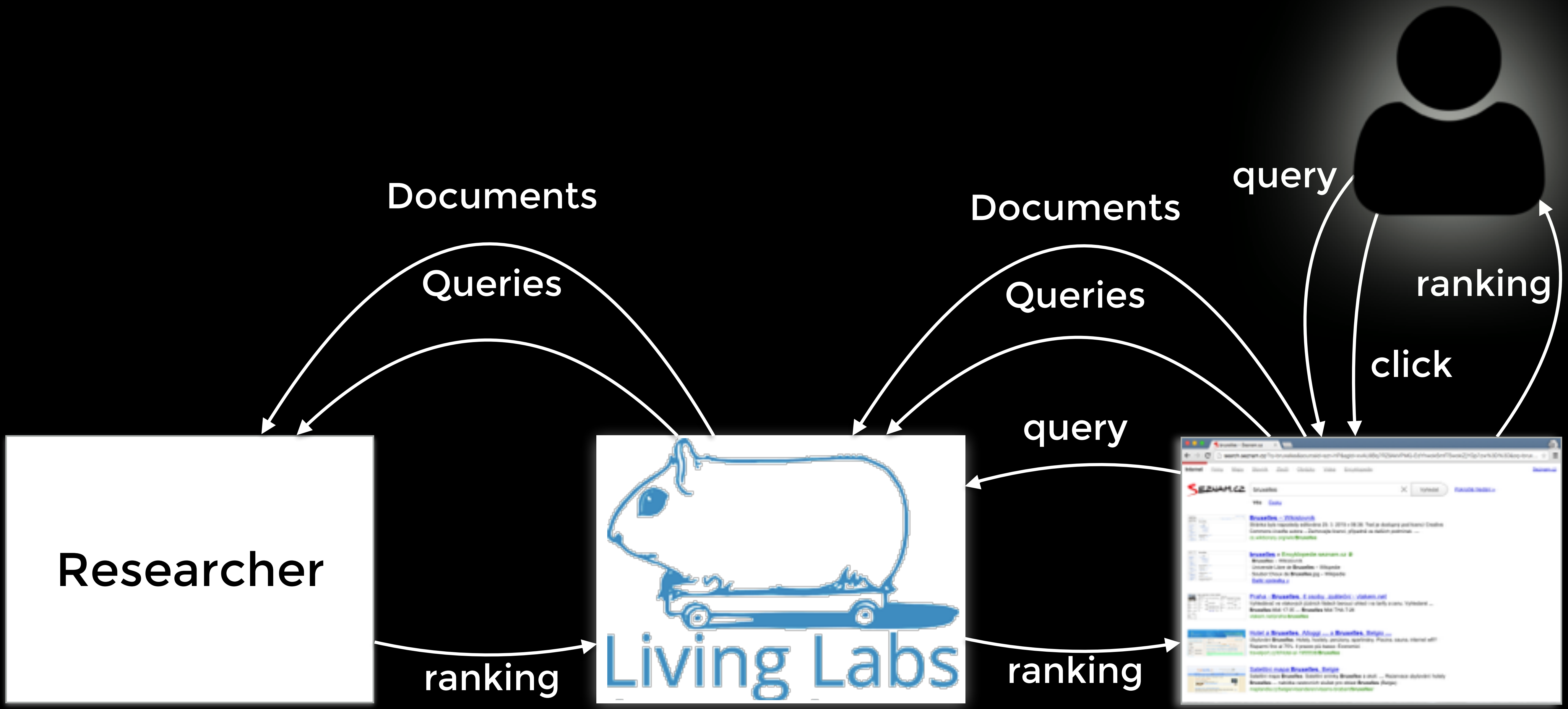
ranking

click

API


Documents
Queries
Researcher
ranking
Living Labs
Documents
Queries
query
query
ranking
click

API



Documents

Queries

Documents

Queries

query

ranking

click

query

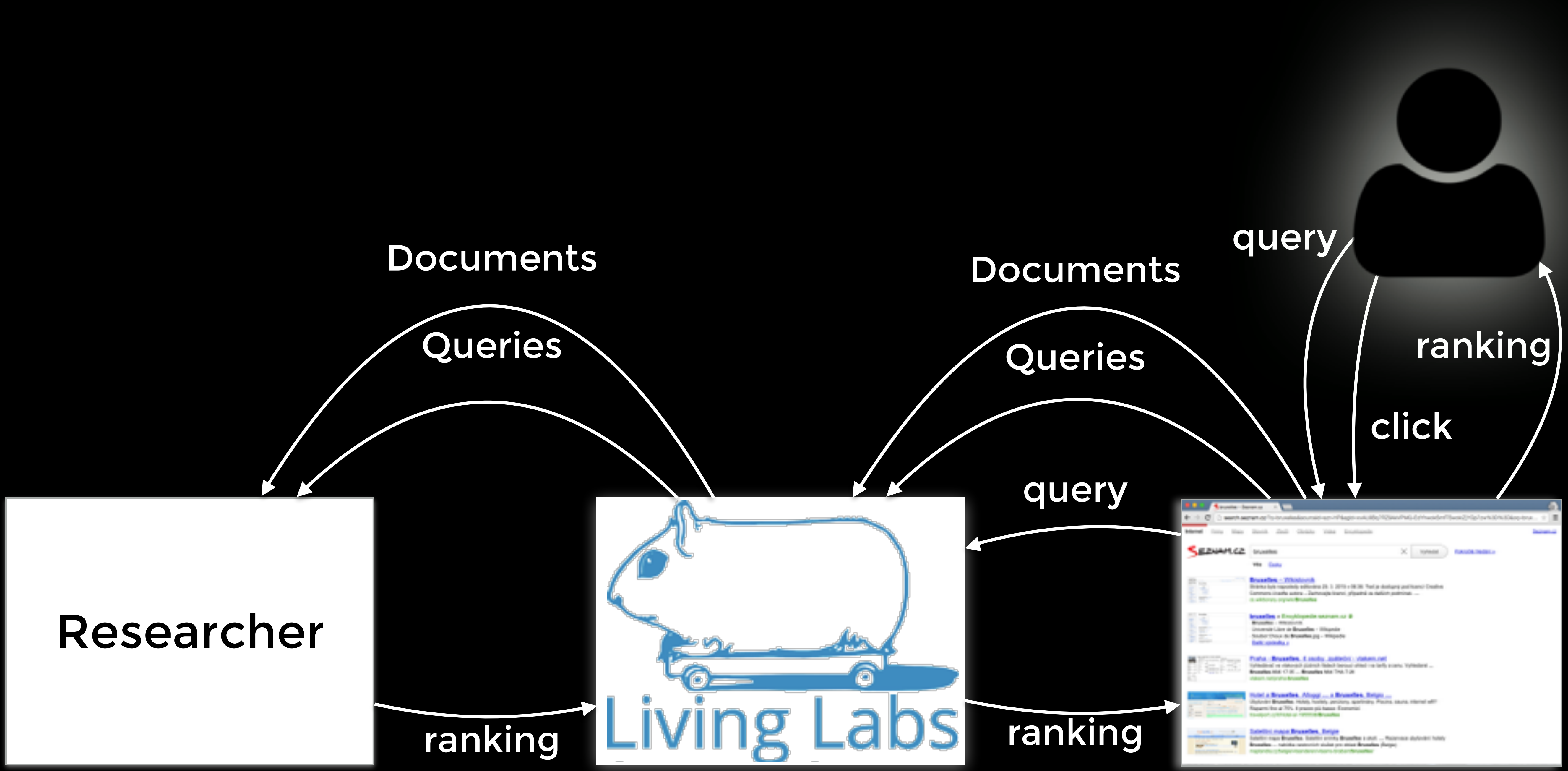
Researcher

ranking

Living Labs

ranking

API



Documents

Queries

Researcher

ranking

Living Labs

Documents

Queries

query

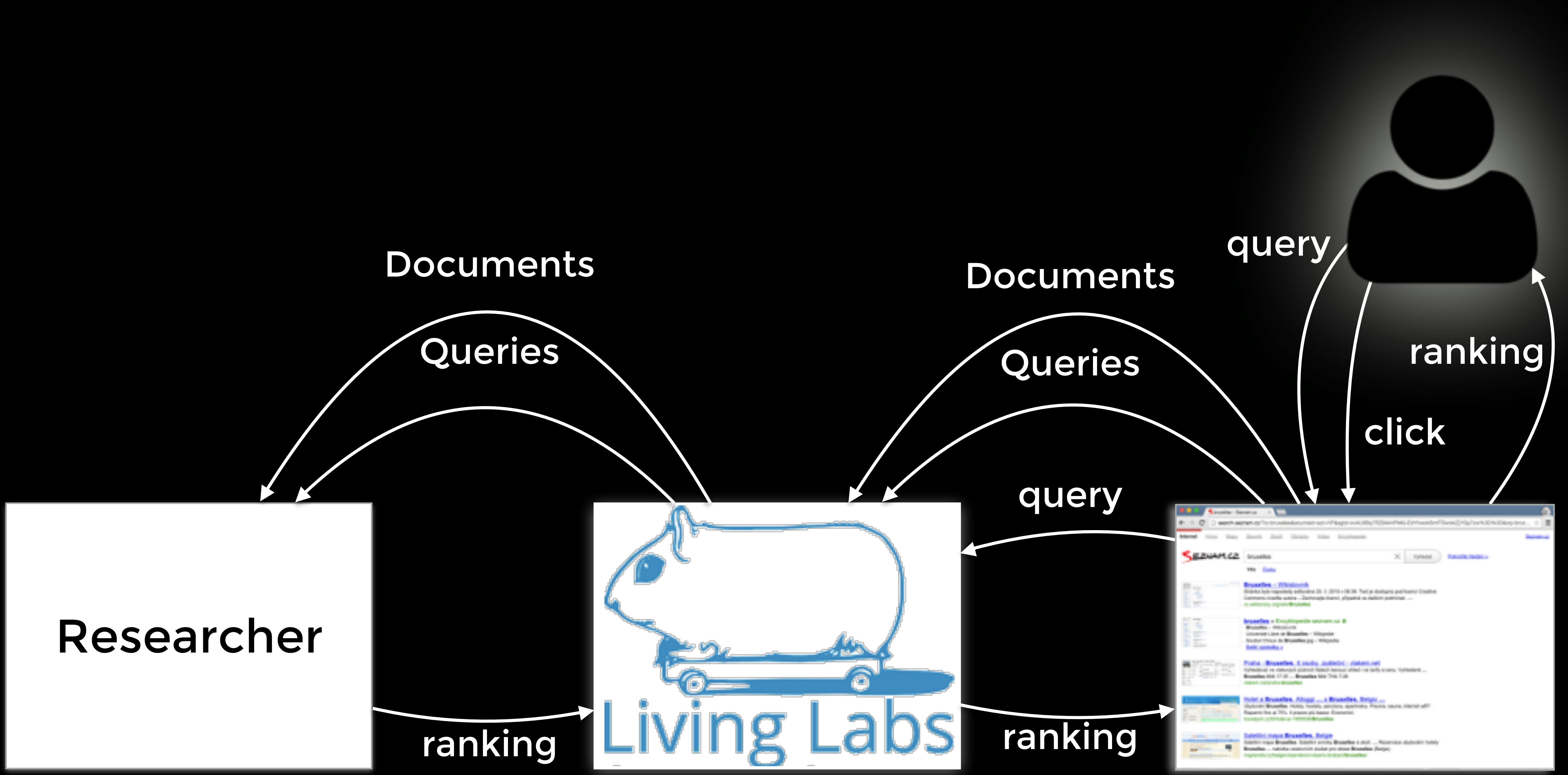
ranking

query

click

ranking

API



Documents

Queries

ranking

Researcher

Living Labs

Documents

Queries

query

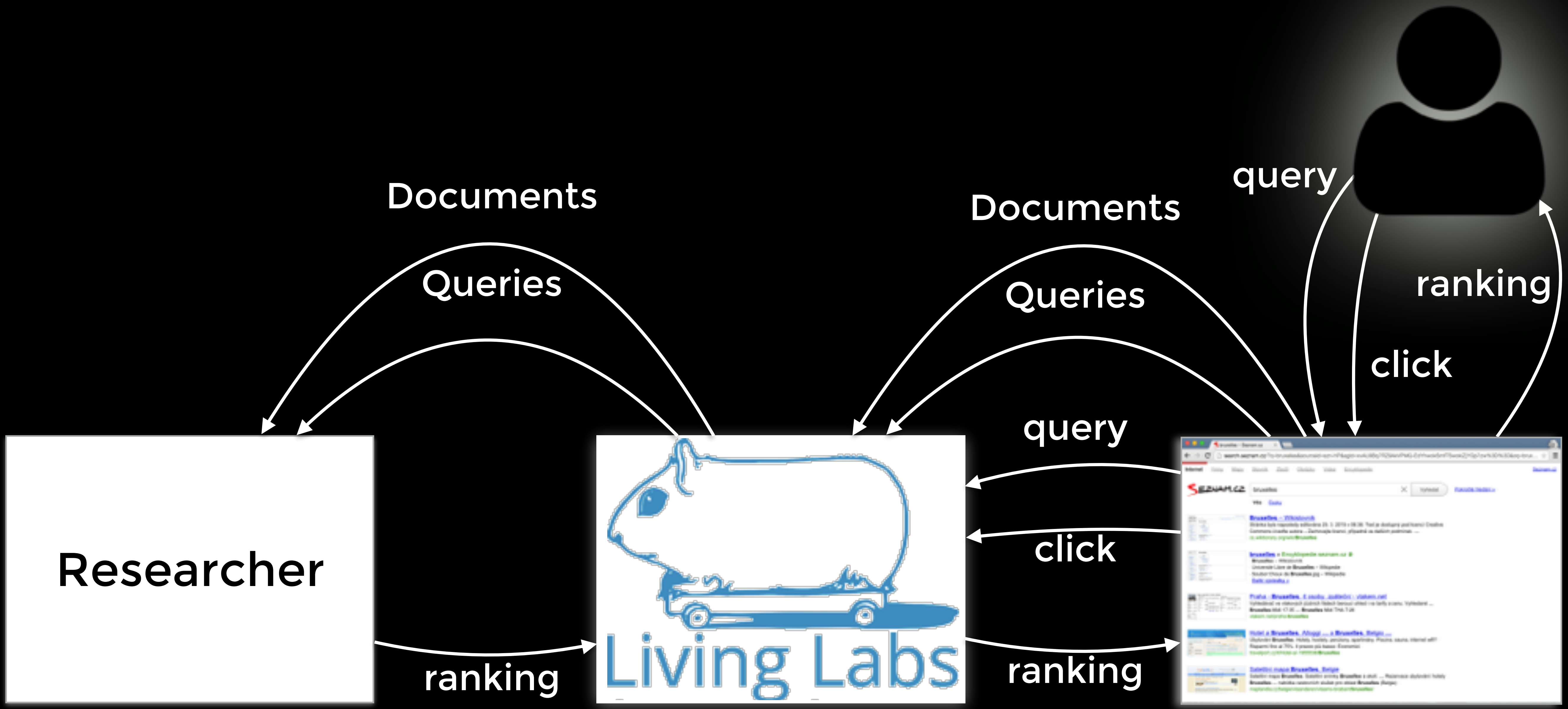
ranking

query

ranking

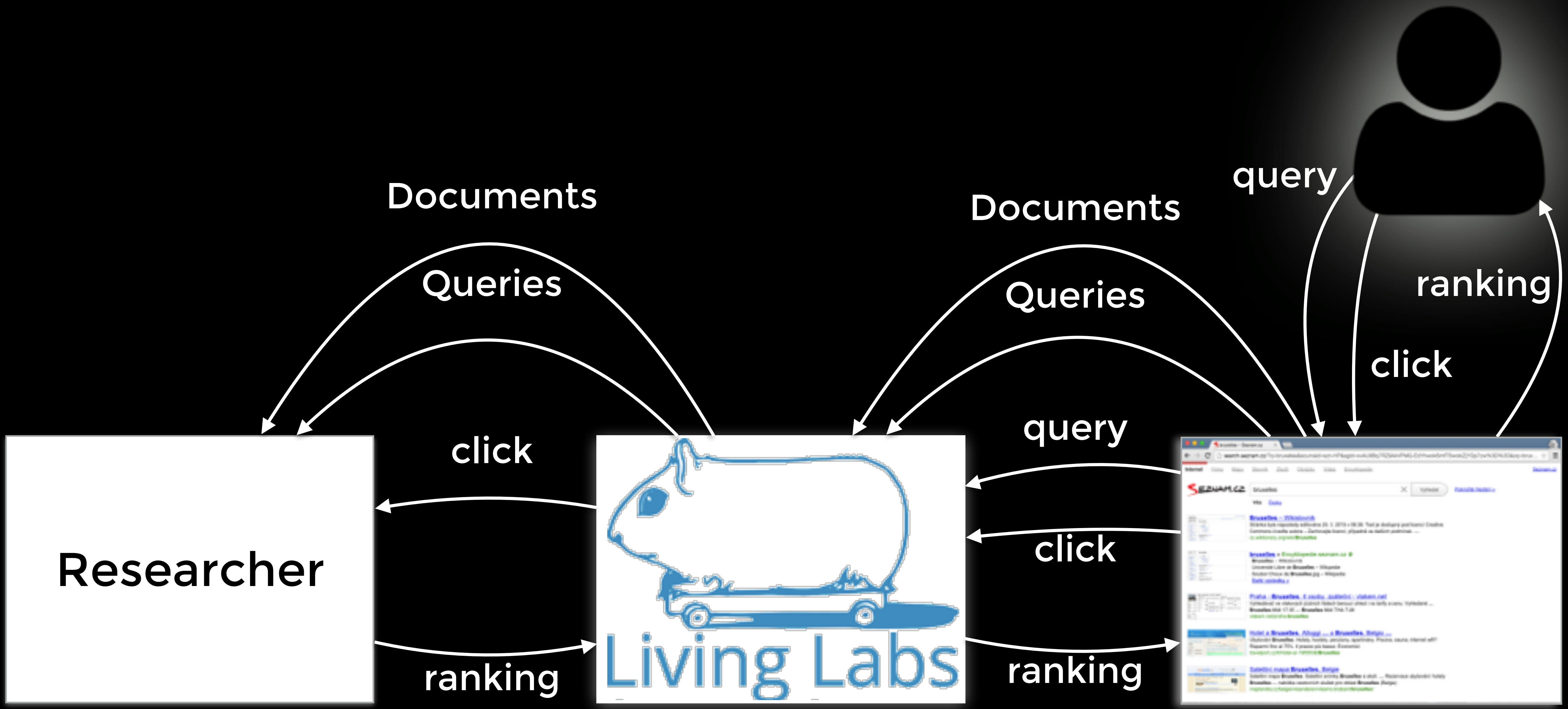
click

API



Documents
Queries
Researcher
ranking
Living Labs
Documents
Queries
query
click
ranking
query
ranking
click

API



Documents

Queries

click

ranking

Researcher

Living Labs

Documents

Queries

query

click

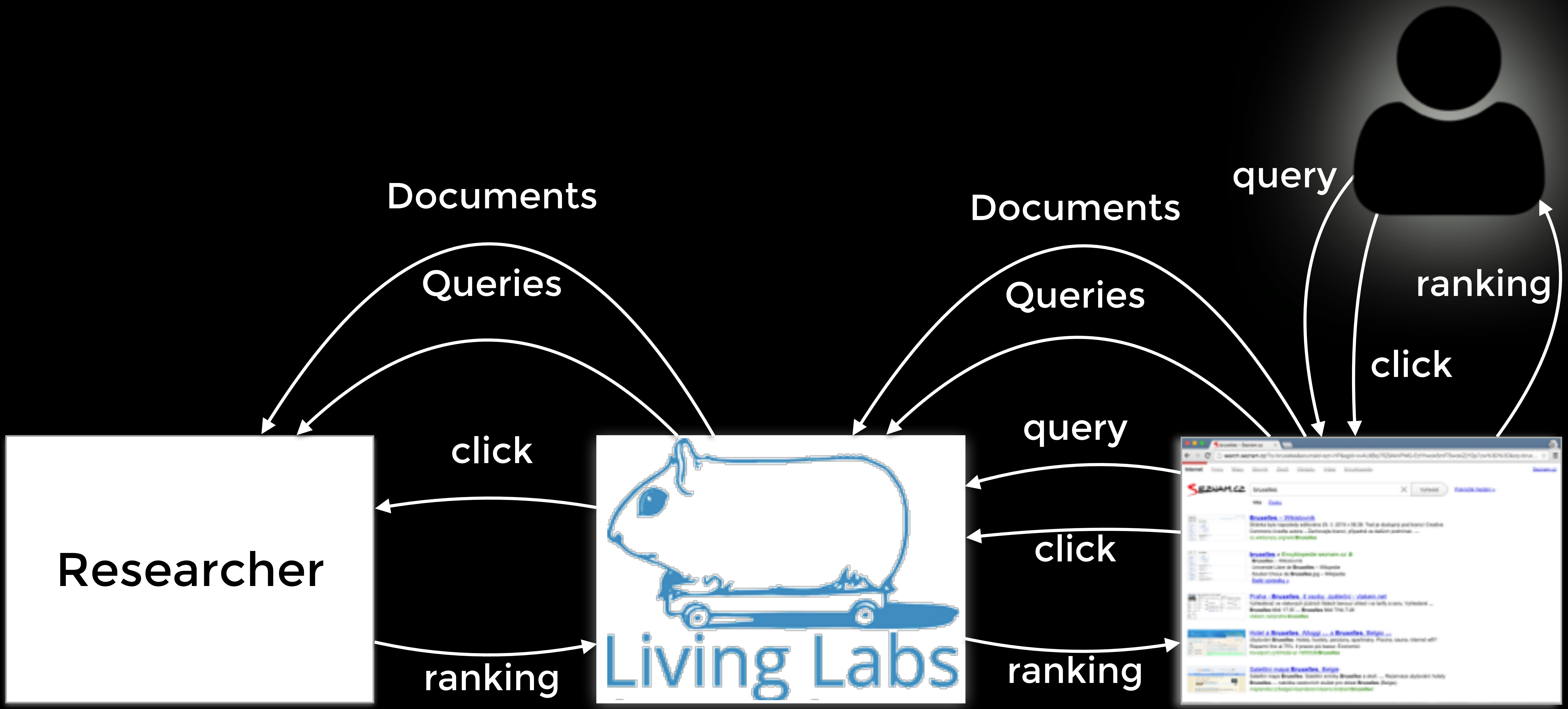
ranking

query

ranking

click

API



Documents

Queries

click

ranking

Researcher

Living Labs

Documents

Queries

query

click

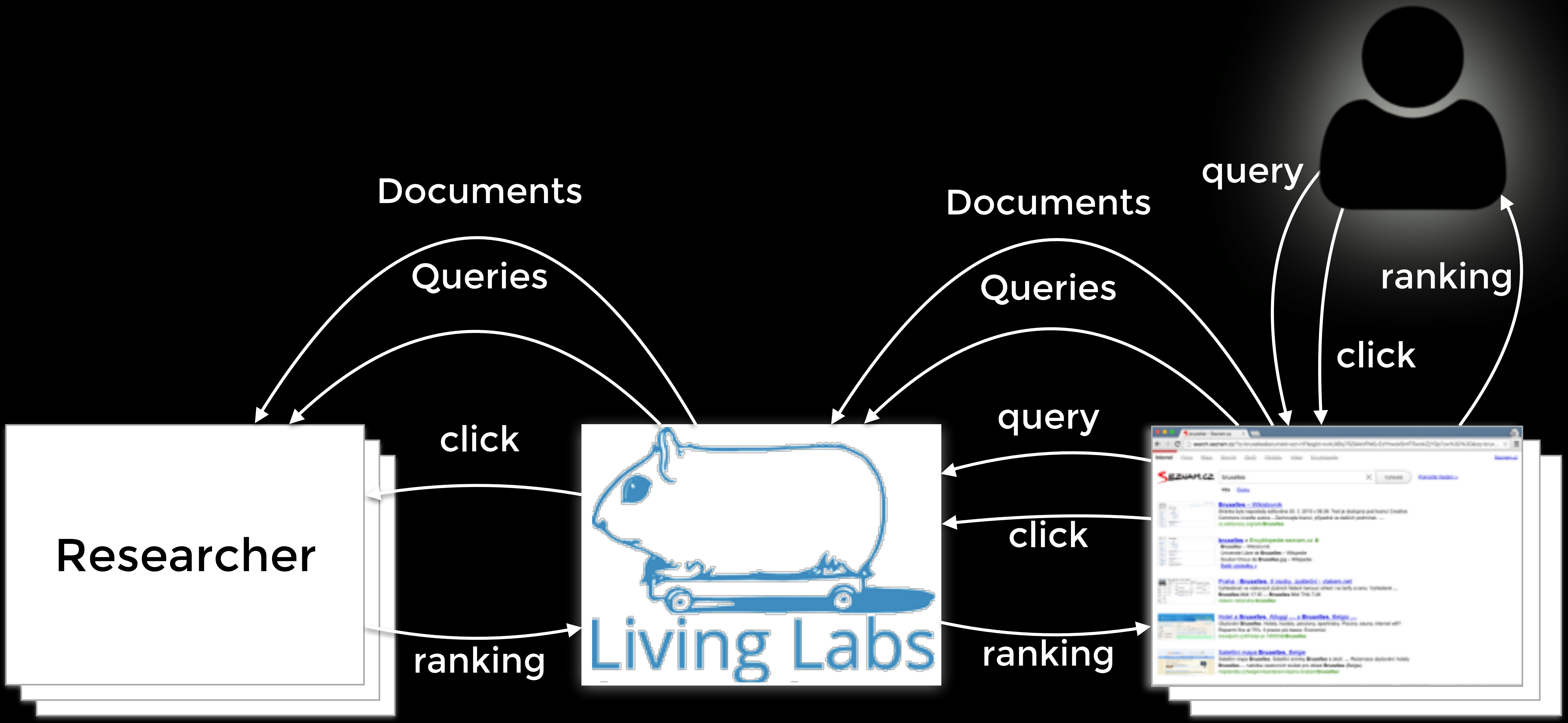
ranking

query

ranking

click

API



Documents

Queries

click

ranking

Researcher

Living Labs

Documents

Queries

query

click

ranking

query

ranking

click

API

- **API** (open source) to communicate

API

- **API** (open source) to communicate
  - Queries, documents, runs, clicks, …

API

- **API** (open source) to communicate
  - Queries, documents, runs, clicks, …
- Both **researchers** and **search engines** use API

API

- **API** (open source) to communicate
  - Queries, documents, runs, clicks, …
- Both **researchers** and **search engines** use API
  - **Easy** to connect new search engines

API

- **API** (open source) to communicate
  - Queries, documents, runs, clicks, …
- Both **researchers** and **search engines** use API
  - **Easy** to connect new search engines
- **Fast** (for crucial requests)

API

- **API** (open source) to communicate
  - Queries, documents, runs, clicks, …
- Both **researchers** and **search engines** use API
  - **Easy** to connect new search engines
- **Fast** (for crucial requests)
- REST-full, JSON

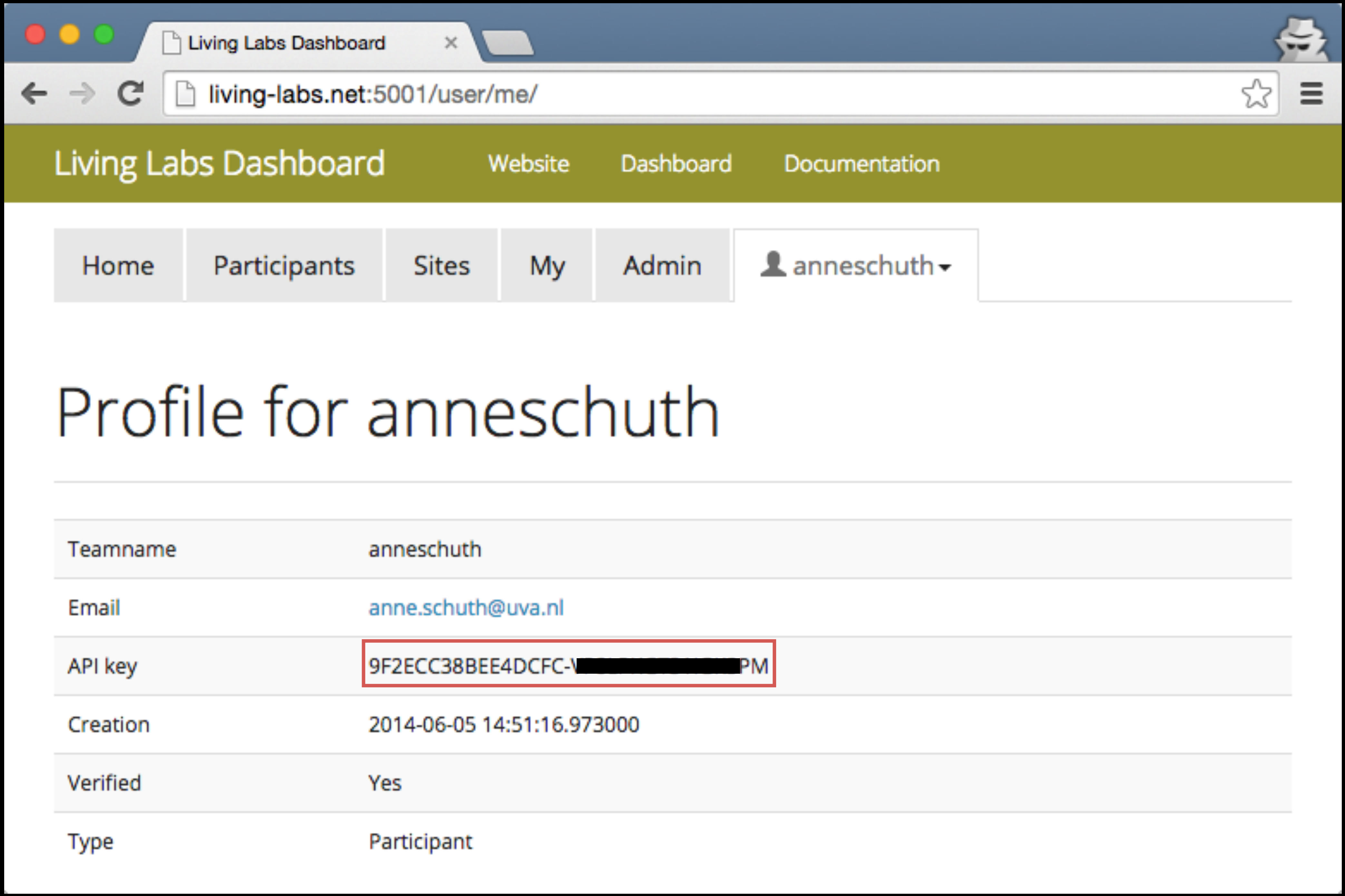
API

- **API** (open source) to communicate
  - Queries, documents, runs, clicks, …
- Both **researchers** and **search engines** use API
  - **Easy** to connect new search engines
- **Fast** (for crucial requests)
- REST-full, JSON
- Example clients

API

- **API** (open source) to communicate
  - Queries, documents, runs, clicks, …
- Both **researchers** and **search engines** use API
  - **Easy** to connect new search engines
- **Fast** (for crucial requests)
- REST-full, JSON
- Example clients
  - **Easy** to get started

Dashboard



Profile for anneschuth

Teamname	anneschuth
Email	anne.schuth@uva.nl
API key	9F2ECC38BEE4DCFC-[redacted]PM
Creation	2014-06-05 14:51:16.973000
Verified	Yes
Type	Participant

API

- **Request**
  GET /api/participant/query/(key)

- **Response**

```
{
"queries": [ {
    "creation_time": "Mon, 10 Nov 2014 17:42:24",
    "qid": "S-q1",
    "qstr": "jaguar",
    "type": "train"
    }, ...]
}
```

Head Queries

- Evaluate systems on the **same set of queries**

Head Queries

- Evaluate systems on the **same set of queries**
- **Stable** volume

Head Queries

- Evaluate systems on the **same set of queries**
- **Stable** volume
- **Historical** click and usage data is available

Head Queries

- Evaluate systems on the **same set of queries**
- **Stable** volume
- **Historical** click and usage data is available
- No (or less) **privacy** issues

Head Queries

- Evaluate systems on the **same set of queries**
- **Stable** volume
- **Historical** click and usage data is available
- No (or less) **privacy** issues
- Researchers can upload rankings **offline**

Evaluation

- **Train** queries

Evaluation

- **Train** queries
    - ‘Immediate’ feedback

Evaluation

- **Train** queries
  - ‘Immediate’ feedback
  - Raw and aggregated feedback

Evaluation

- **Train** queries
  - ‘Immediate’ feedback
  - Raw and aggregated feedback
- **Test** queries

Evaluation

- **Train** queries
  - ‘Immediate’ feedback
  - Raw and aggregated feedback
- **Test** queries
  - **No updates** during test period

Evaluation

- **Train** queries
  - ‘Immediate’ feedback
  - Raw and aggregated feedback
- **Test** queries
  - **No updates** during test period
  - Feedback after test period

Evaluation

- **Train** queries
  - ‘Immediate’ feedback
  - Raw and aggregated feedback
- **Test** queries
  - **No updates** during test period
  - Feedback after test period
  - Only Aggregated feedback

Evaluation

- **Train** queries
  - ‘Immediate’ feedback
  - Raw and aggregated feedback
- **Test** queries
  - **No updates** during test period
  - Feedback after test period
  - Only Aggregated feedback
- **Metric**: Team Draft Interleaving

Evaluation

- **Train** queries
    - ‘Immediate’ feedback
    - Raw and aggregated feedback
- **Test** queries
    - **No updates** during test period
    - Feedback after test period
    - Only Aggregated feedback
- **Metric**: Team Draft Interleaving
    - Fraction of **wins** against production

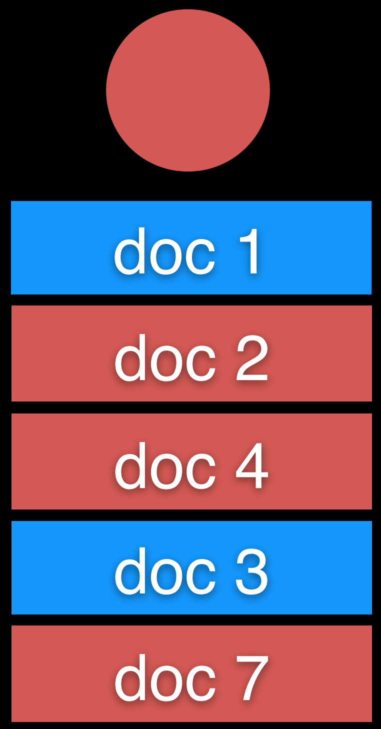
Team Draft Interleaving

Production	Researcher
doc 1	doc 2
doc 2	doc 4
doc 3	doc 7
doc 4	doc 1
doc 5	doc 3

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

Team Draft Interleaving

Production Researcher


doc 1

doc 2

doc 4

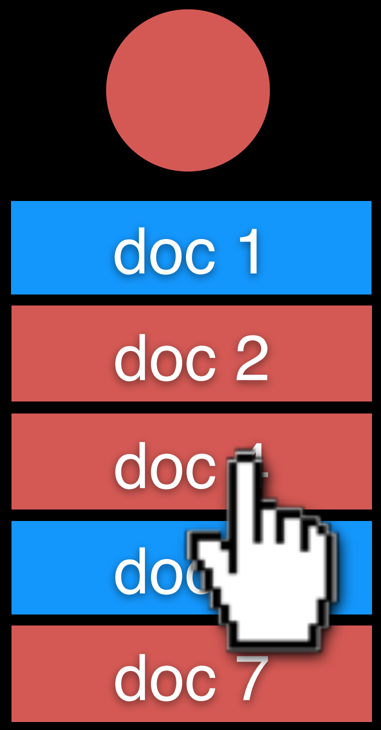
doc 3

doc 7

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

Team Draft Interleaving

Production Researcher

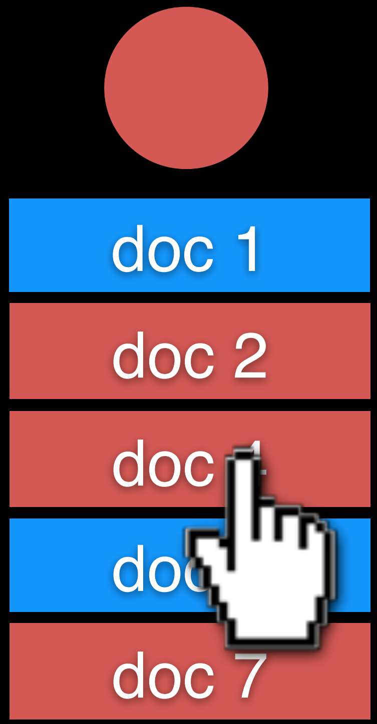


F. Radlinski, M. Kurup, and T. Joachims.
How does clickthrough data reflect retrieval
quality? In CIKM ’08. 2008

Team Draft Interleaving

Production Researcher

Researcher is preferred over Production



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

Participation

- 39 teams signed up

Participation

- 39 teams signed up
  - Industry:
    904labs, Microsoft, Plista, Yahoo

Participation

- 39 teams signed up
    - Industry:
      904labs, Microsoft, Plista, Yahoo
    - Academia:
      au, bw, cz, fr, ie, in, jp, nl, no, uk, us

Participation

- 39 teams signed up
  - Industry:
    904labs, Microsoft, Plista, Yahoo
  - Academia:
    au, bw, cz, fr, ie, in, jp, nl, no, uk, us
- 20 teams signed our agreement

Participation

- 39 teams signed up
  - Industry:
    904labs, Microsoft, Plista, Yahoo
  - Academia:
    au, bw, cz, fr, ie, in, jp, nl, no, uk, us
- 20 teams signed our agreement
- 12 teams submitted runs

Participation

- 39 teams signed up
  - Industry:
    904labs, Microsoft, Plista, Yahoo
  - Academia:
    au, bw, cz, fr, ie, in, jp, nl, no, uk, us
- 20 teams signed our agreement
- 12 teams submitted runs
- 3 teams submitted 5 runs for test queries

the lab

use cases

conclusions

the lab

use cases

conclusions

Use Cases

Provider	
Data	
Site traffic	
Info needs	

Use Cases

	Local domain search
Provider	uva.nl
Data	raw queries and (generally textual) documents
Site traffic	relatively low
Info needs	(mostly) navigational

Use Cases

	Local domain search	**Product search**
Provider	uva.nl	regiojatek.hu
Data	raw queries and (generally textual) documents	raw queries and (highly structured) documents
Site traffic	relatively low	relatively low (~4K sessions/day)
Info needs	(mostly) navigational	(mostly) transactional

Use Cases

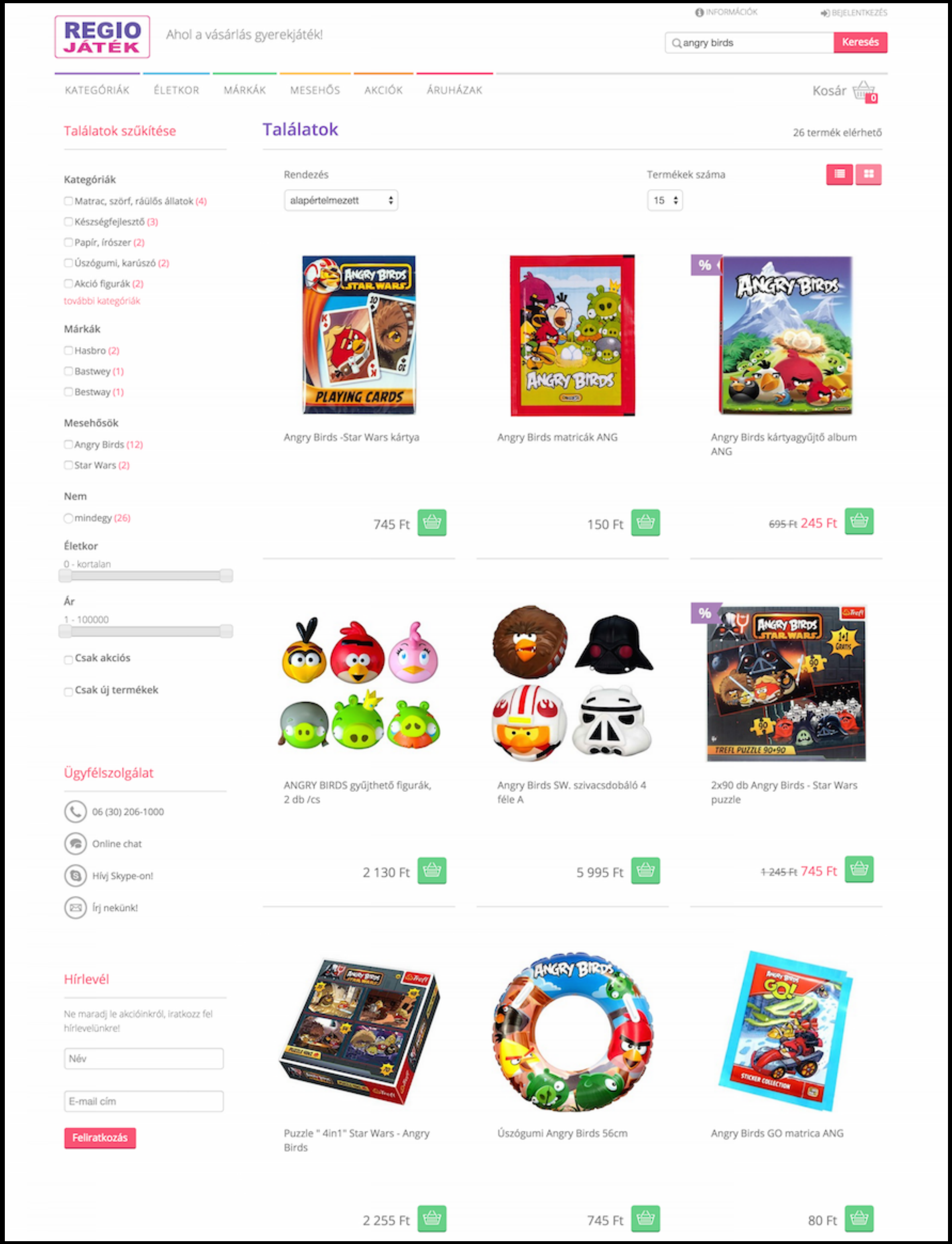
	Local domain search	**Product search**	**Web search**
Provider	uva.nl	regiojatek.hu	seznam.cz
Data	raw queries and (generally textual) documents	raw queries and (highly structured) documents	pre-computed document-query features
Site traffic	relatively low	relatively low (~4K sessions/day)	high
Info needs	(mostly) navigational	(mostly) transactional	vary

Use Cases

Not implemented on time

	Local domain search	**Product search**	**Web search**
Provider	uva.nl	regiojatek.hu	seznam.cz
Data	raw queries and (generally textual) documents	raw queries and (highly structured) documents	pre-computed document-query features
Site traffic	relatively low	relatively low (~4K sessions/day)	high
Info needs	(mostly) navigational	(mostly) transactional	vary

Product Search



REGIO JÁTÉK

Ahol a vásárlás gyerekjáték!

INFORMÁCIÓK

BEJELENTKEZÉS

angry birds

Keresés

KATEGÓRIÁK ÉLETKOR MÁRKÁK MESEHŐS AKCIÓK ÁRUHÁZAK

Kosár 0

Találatok szűkítése

Kategóriák

□ Matrac, szörf, ráülős állatok (4)
□ Készségfejlesztő (3)
□ Papír, írószer (2)
□ Úszógumi, karúszó (2)
□ Akció figurák (2)

további kategóriák

Márkák

□ Hasbro (2)
□ Bastwey (1)
□ Bestway (1)

Mesehősök

□ Angry Birds (12)
□ Star Wars (2)

Nem

○ mindegy (26)

Életkor

0 - kortalan

Ár

1 - 100000

□ Csak akciós

□ Csak új termékek

Ügyfélszolgálat

☎ 06 (30) 206-1000

Online chat

Hívj Skype-on!

✉ Írj nekünk!

Hírlevél

Ne maradj le akcióinkról, iratkozz fel hírlevelünkre!

Név

E-mail cím

Feliratkozás

Találatok

26 termék elérhető

Rendezés: alapértelmezett

Termékek száma: 15

Angry Birds -Star Wars kártya — 745 Ft

Angry Birds matricák ANG — 150 Ft

% Angry Birds kártyagyűjtő album ANG — ~~695 Ft~~ 245 Ft

ANGRY BIRDS gyűjthető figurák, 2 db /cs — 2 130 Ft

Angry Birds SW. szivacsdobáló 4 féle A — 5 995 Ft

% 2x90 db Angry Birds - Star Wars puzzle — ~~1 245 Ft~~ 745 Ft

Puzzle " 4in1" Star Wars - Angry Birds — 2 255 Ft

Úszógumi Angry Birds 56cm — 745 Ft

Angry Birds GO matrica ANG — 80 Ft

Product Search

- Toy store
- Highly structured documents representing products
- Many fields:
  - age_max, age_min, arrived, available, bonus, price, brand, category, category_id, characters, description, etc, …

Product Search - Participation

Product Search - Participation

- 3 teams submitted runs for train queries

Product Search - Participation

- 3 teams submitted runs for train queries
    - UIS

Product Search - Participation

- 3 teams submitted runs for train queries
    - UIS
    - IRIT

Product Search - Participation

- 3 teams submitted runs for train queries
  - UIS
  - IRIT
  - GESIS

Product Search - Participation

- 3 teams submitted runs for train queries
    - UIS
    - IRIT
    - GESIS
- 5 runs submitted for test queries

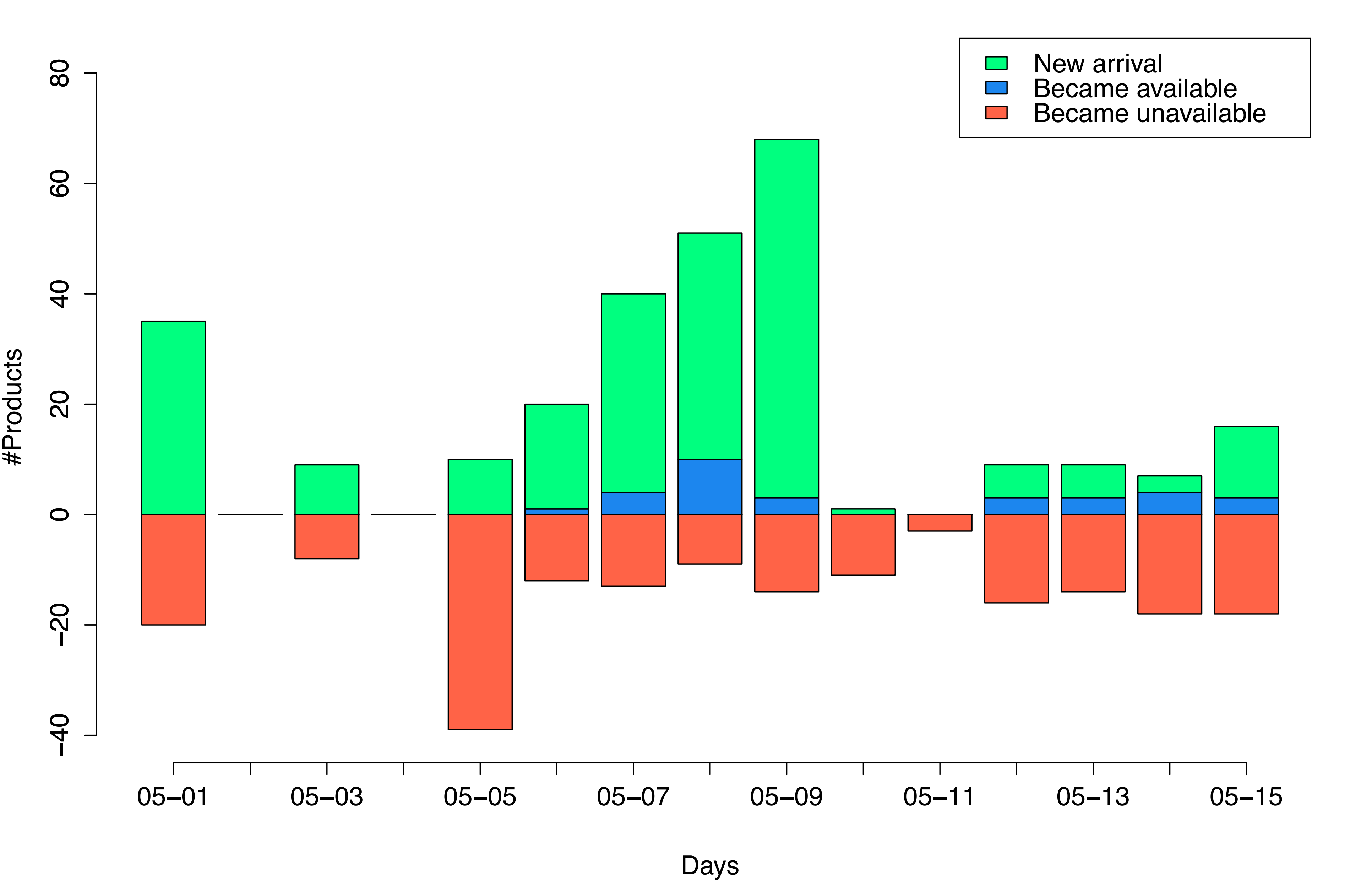
Product Search - Participation

- 3 teams submitted runs for train queries
    - UIS
    - IRIT
    - GESIS
- 5 runs submitted for test queries
- One baseline

Product Search - Participation

- 3 teams submitted runs for train queries
    - UIS
    - IRIT
    - GESIS
- 5 runs submitted for test queries
- One baseline
    - Sorting by historical clicks

Product Search - Inventory



Product Search - Inventory

- Participants **should** update available products

Product Search - Inventory

- Participants **should** update available products
- Rankings **may** contain stale products

Product Search - Inventory

- Participants **should** update available products
- Rankings **may** contain stale products
- These products were removed **after** interleaving

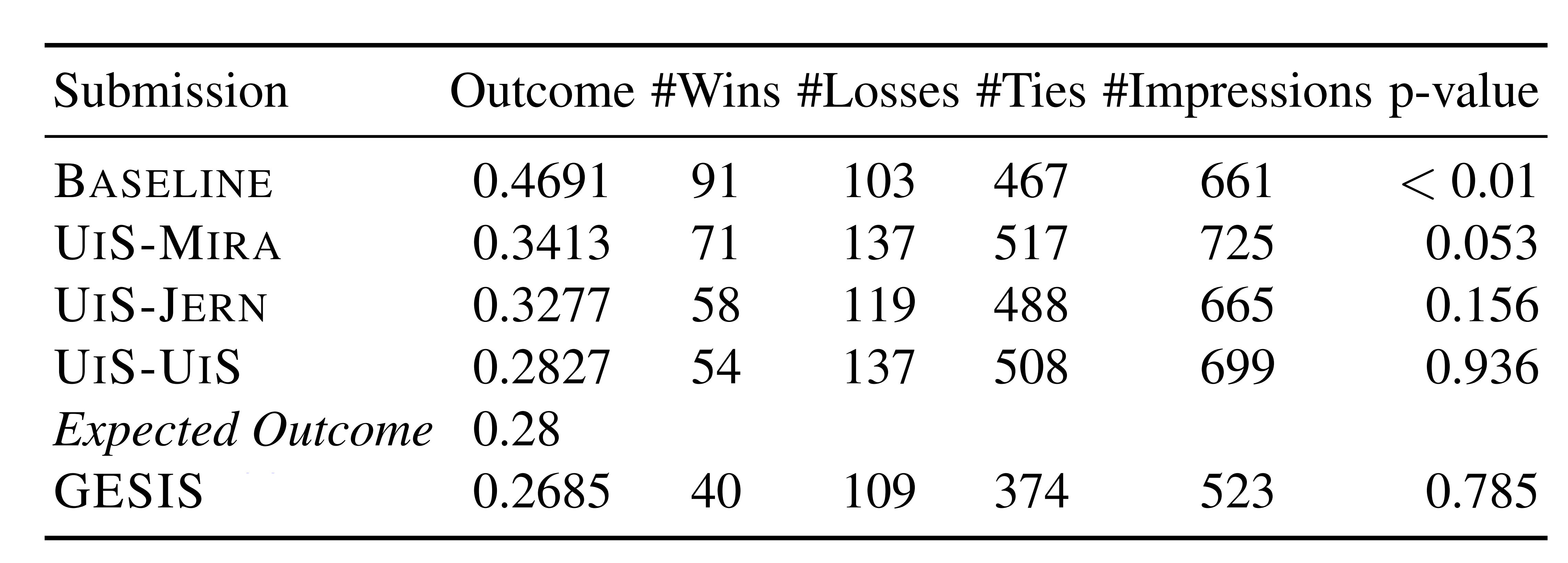
Product Search - Inventory

- Participants **should** update available products
- Rankings **may** contain stale products
- These products were removed **after** interleaving
  - Biasing in favor of production (which never has stale products)

Product Search - Inventory

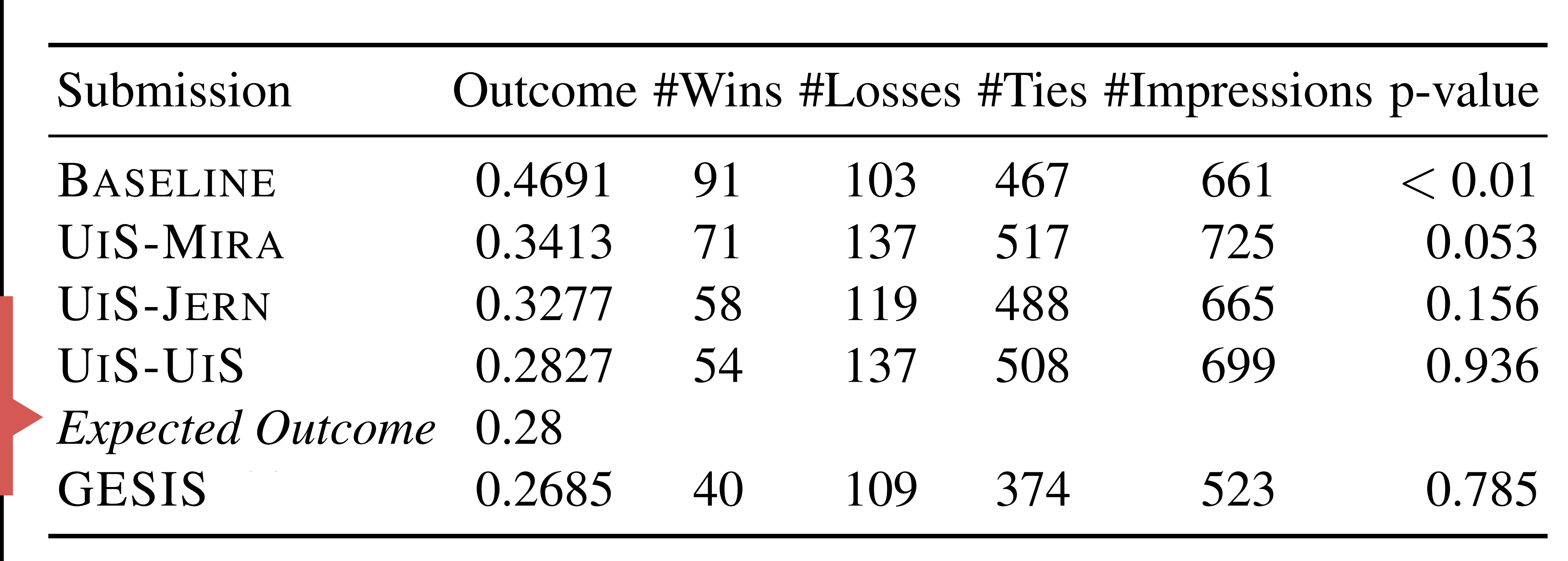
- Participants **should** update available products
- Rankings **may** contain stale products
- These products were removed **after** interleaving
  - Biasing in favor of production (which never has stale products)
  - Expected interleaving outcome is no longer 0.5

Product Search - Results - #1



Submission	Outcome	#Wins	#Losses	#Ties	#Impressions	p-value
Baseline	0.4691	91	103	467	661	< 0.01
UiS-Mira	0.3413	71	137	517	725	0.053
UiS-Jern	0.3277	58	119	488	665	0.156
UiS-UiS	0.2827	54	137	508	699	0.936
Expected Outcome	0.28					
GESIS	0.2685	40	109	374	523	0.785

Product Search - Results - #1

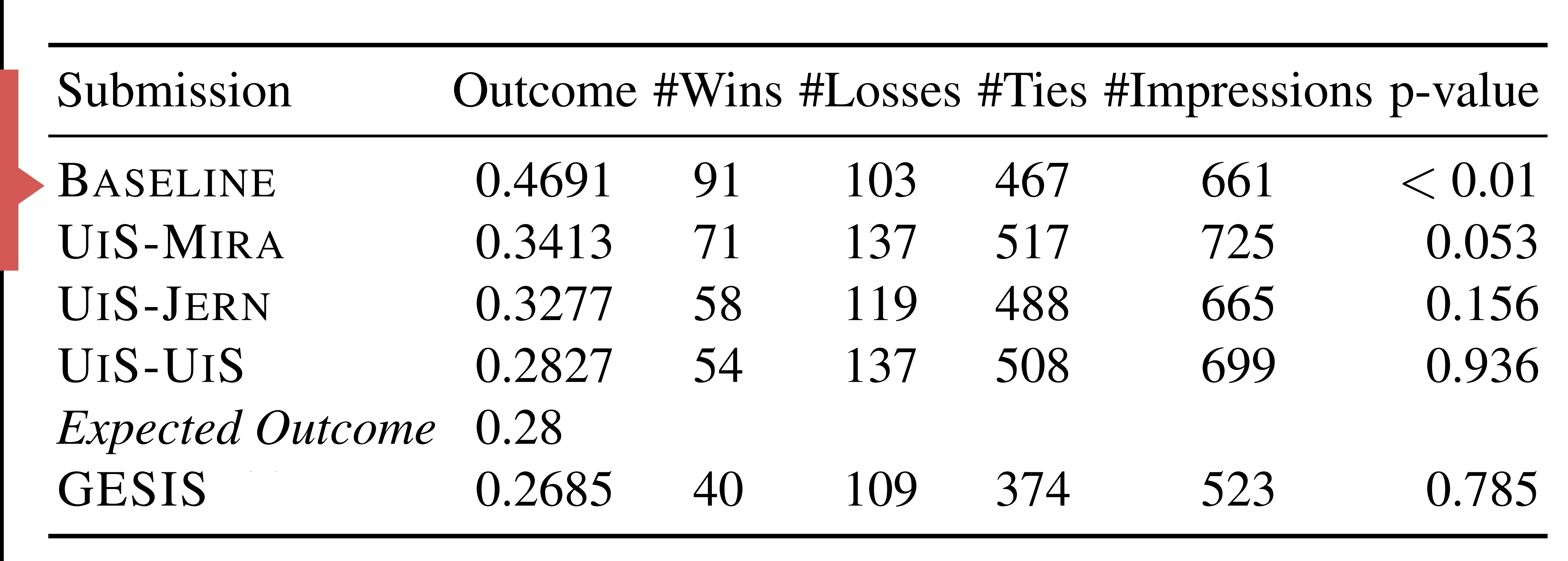


Submission	Outcome	#Wins	#Losses	#Ties	#Impressions	p-value
Baseline	0.4691	91	103	467	661	< 0.01
UiS-Mira	0.3413	71	137	517	725	0.053
UiS-Jern	0.3277	58	119	488	665	0.156
UiS-UiS	0.2827	54	137	508	699	0.936
Expected Outcome	0.28					
GESIS	0.2685	40	109	374	523	0.785

due to inventory changes

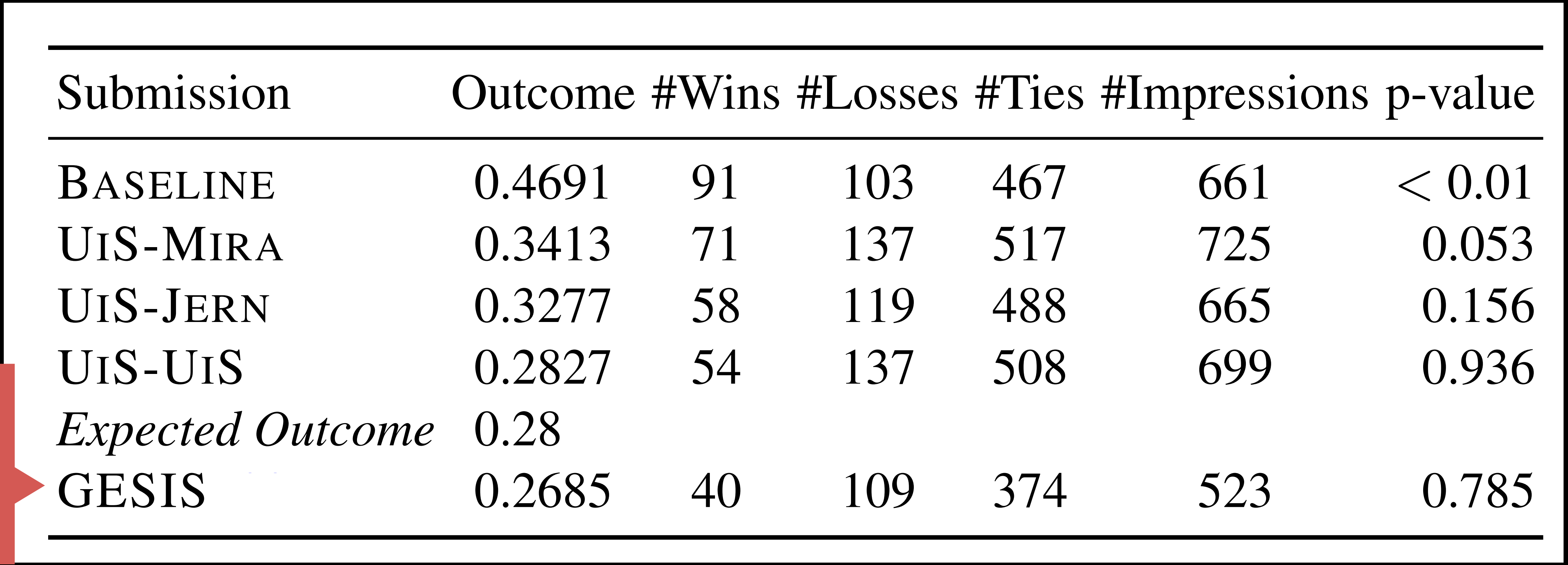
Product Search - Results - #1

ordered by historical clicks



Submission	Outcome	#Wins	#Losses	#Ties	#Impressions	p-value
Baseline	0.4691	91	103	467	661	< 0.01
UiS-Mira	0.3413	71	137	517	725	0.053
UiS-Jern	0.3277	58	119	488	665	0.156
UiS-UiS	0.2827	54	137	508	699	0.936
Expected Outcome	0.28					
GESIS	0.2685	40	109	374	523	0.785

Product Search - Results - #1

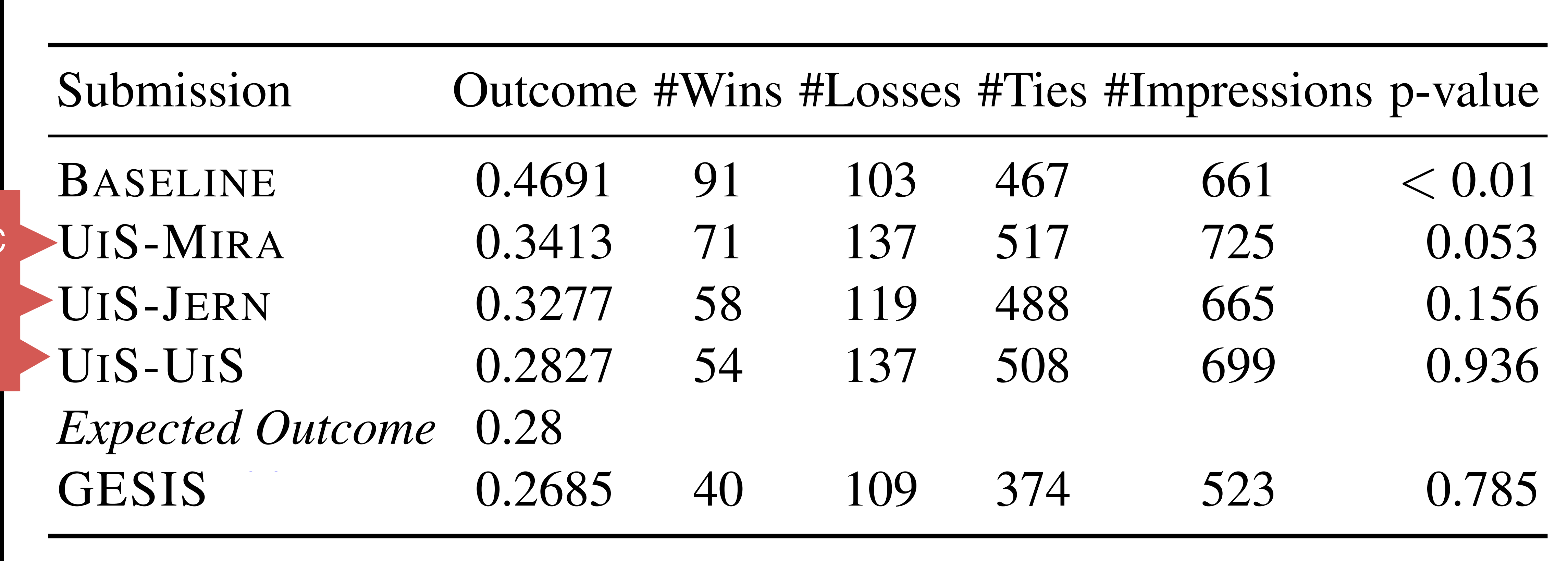


Submission	Outcome	#Wins	#Losses	#Ties	#Impressions	p-value
Baseline	0.4691	91	103	467	661	< 0.01
UiS-Mira	0.3413	71	137	517	725	0.053
UiS-Jern	0.3277	58	119	488	665	0.156
UiS-UiS	0.2827	54	137	508	699	0.936
Expected Outcome	0.28					
GESIS	0.2685	40	109	374	523	0.785

SOLR +
click rerank

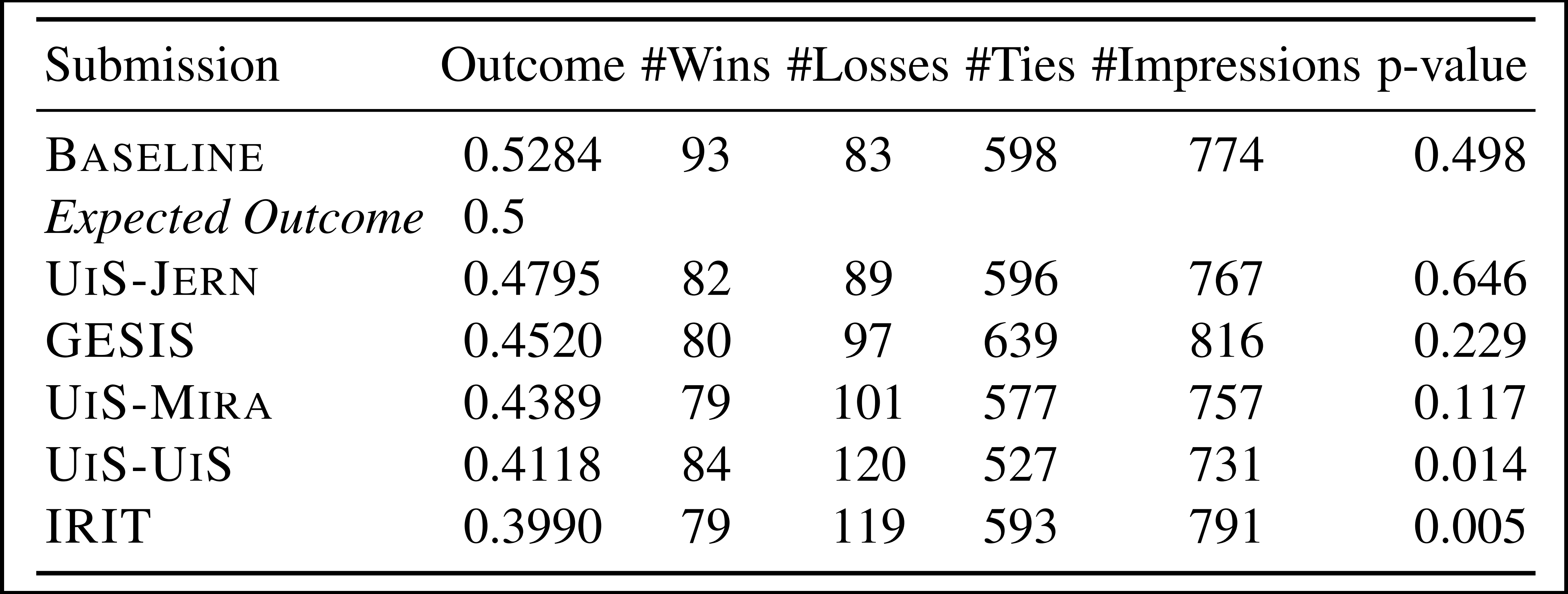
Product Search - Results - #1

Probabilistic Retrieval Models



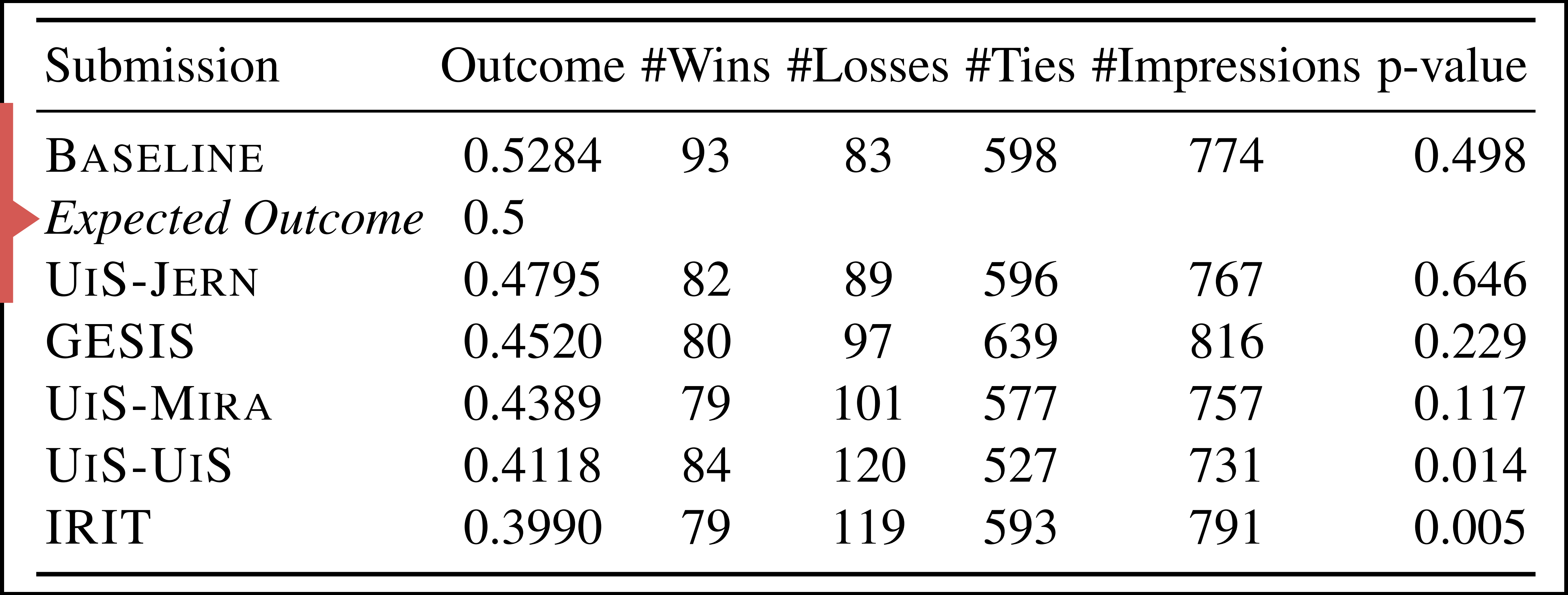
Submission	Outcome	#Wins	#Losses	#Ties	#Impressions	p-value
Baseline	0.4691	91	103	467	661	< 0.01
UiS-Mira	0.3413	71	137	517	725	0.053
UiS-Jern	0.3277	58	119	488	665	0.156
UiS-UiS	0.2827	54	137	508	699	0.936
Expected Outcome	0.28					
GESIS	0.2685	40	109	374	523	0.785

Product Search - Results - #2



Submission	Outcome	#Wins	#Losses	#Ties	#Impressions	p-value
Baseline	0.5284	93	83	598	774	0.498
Expected Outcome	0.5					
UiS-Jern	0.4795	82	89	596	767	0.646
GESIS	0.4520	80	97	639	816	0.229
UiS-Mira	0.4389	79	101	577	757	0.117
UiS-UiS	0.4118	84	120	527	731	0.014
IRIT	0.3990	79	119	593	791	0.005

Product Search - Results - #2

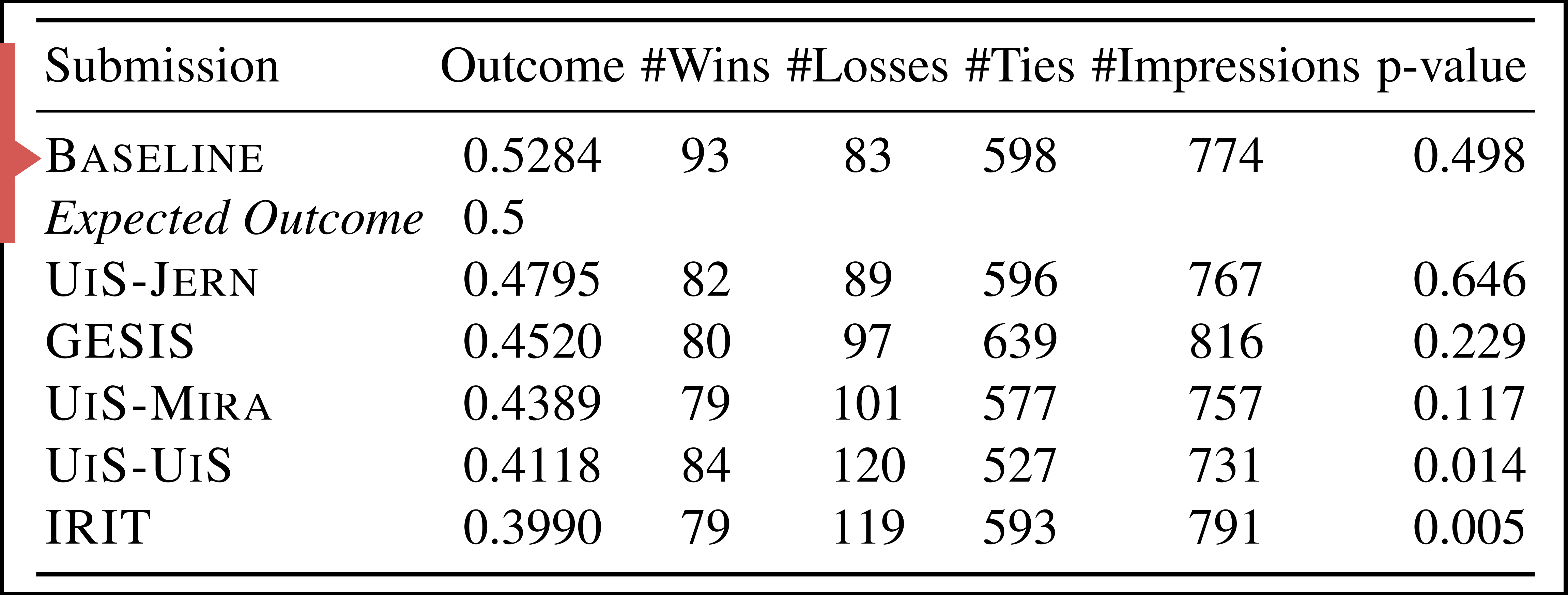


Submission	Outcome	#Wins	#Losses	#Ties	#Impressions	p-value
Baseline	0.5284	93	83	598	774	0.498
Expected Outcome	0.5					
UiS-Jern	0.4795	82	89	596	767	0.646
GESIS	0.4520	80	97	639	816	0.229
UiS-Mira	0.4389	79	101	577	757	0.117
UiS-UiS	0.4118	84	120	527	731	0.014
IRIT	0.3990	79	119	593	791	0.005

no effect inventory changes

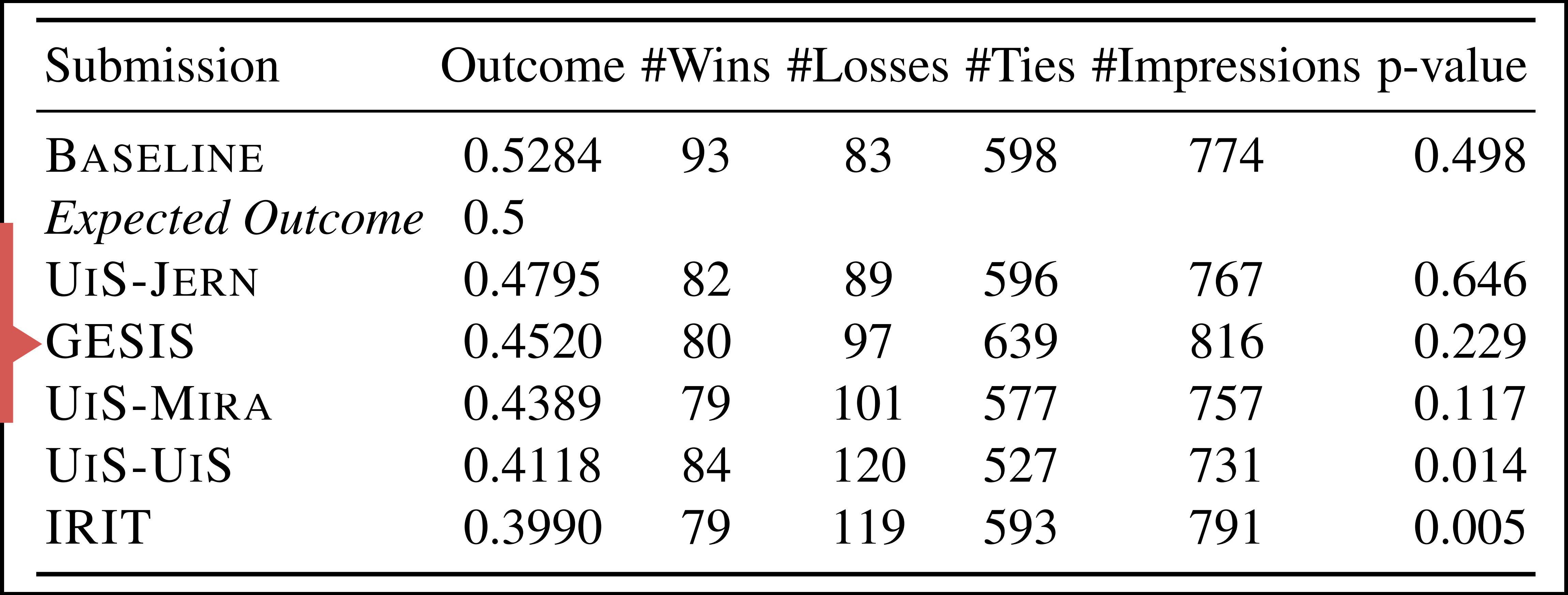
Product Search - Results - #2

ordered by historical clicks



Submission	Outcome	#Wins	#Losses	#Ties	#Impressions	p-value
Baseline	0.5284	93	83	598	774	0.498
Expected Outcome	0.5					
UiS-Jern	0.4795	82	89	596	767	0.646
GESIS	0.4520	80	97	639	816	0.229
UiS-Mira	0.4389	79	101	577	757	0.117
UiS-UiS	0.4118	84	120	527	731	0.014
IRIT	0.3990	79	119	593	791	0.005

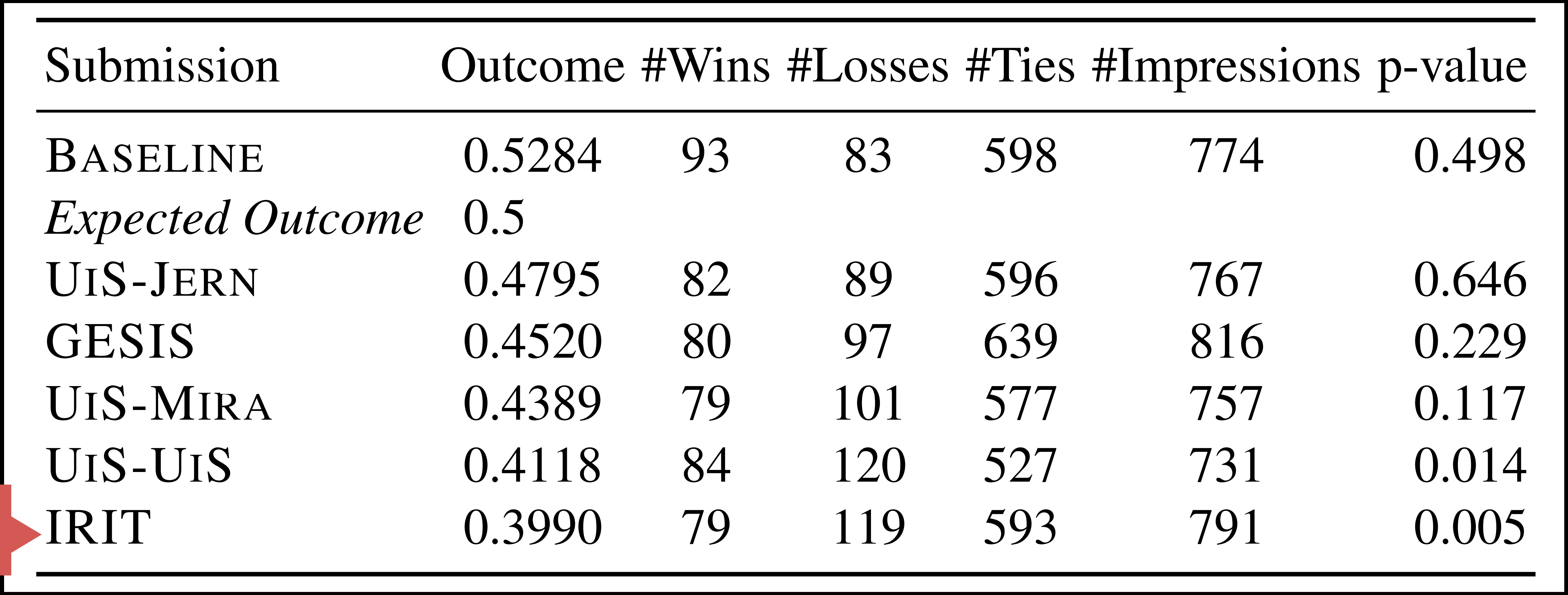
Product Search - Results - #2



Submission	Outcome	#Wins	#Losses	#Ties	#Impressions	p-value
Baseline	0.5284	93	83	598	774	0.498
Expected Outcome	0.5					
UiS-Jern	0.4795	82	89	596	767	0.646
GESIS	0.4520	80	97	639	816	0.229
UiS-Mira	0.4389	79	101	577	757	0.117
UiS-UiS	0.4118	84	120	527	731	0.014
IRIT	0.3990	79	119	593	791	0.005

fixed SOLR + click rerank

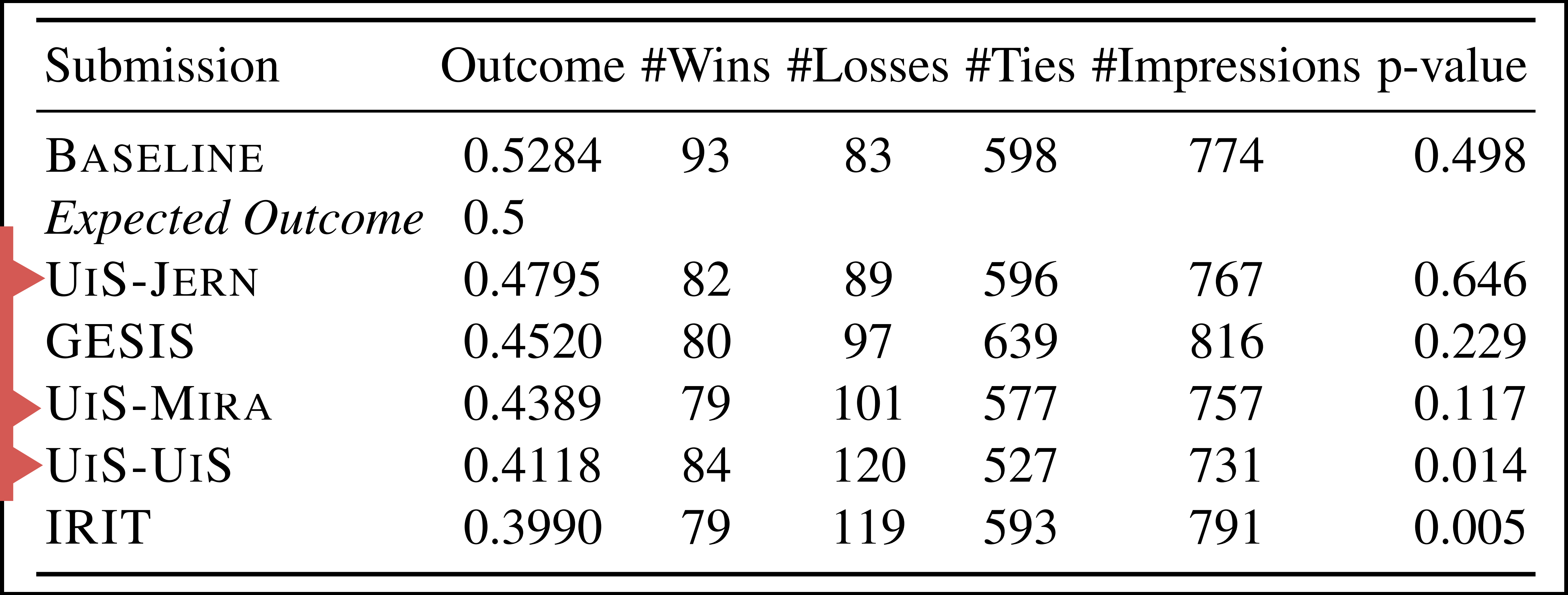
Product Search - Results - #2



Submission	Outcome	#Wins	#Losses	#Ties	#Impressions	p-value
Baseline	0.5284	93	83	598	774	0.498
Expected Outcome	0.5					
UiS-Jern	0.4795	82	89	596	767	0.646
GESIS	0.4520	80	97	639	816	0.229
UiS-Mira	0.4389	79	101	577	757	0.117
UiS-UiS	0.4118	84	120	527	731	0.014
IRIT	0.3990	79	119	593	791	0.005

BM25F

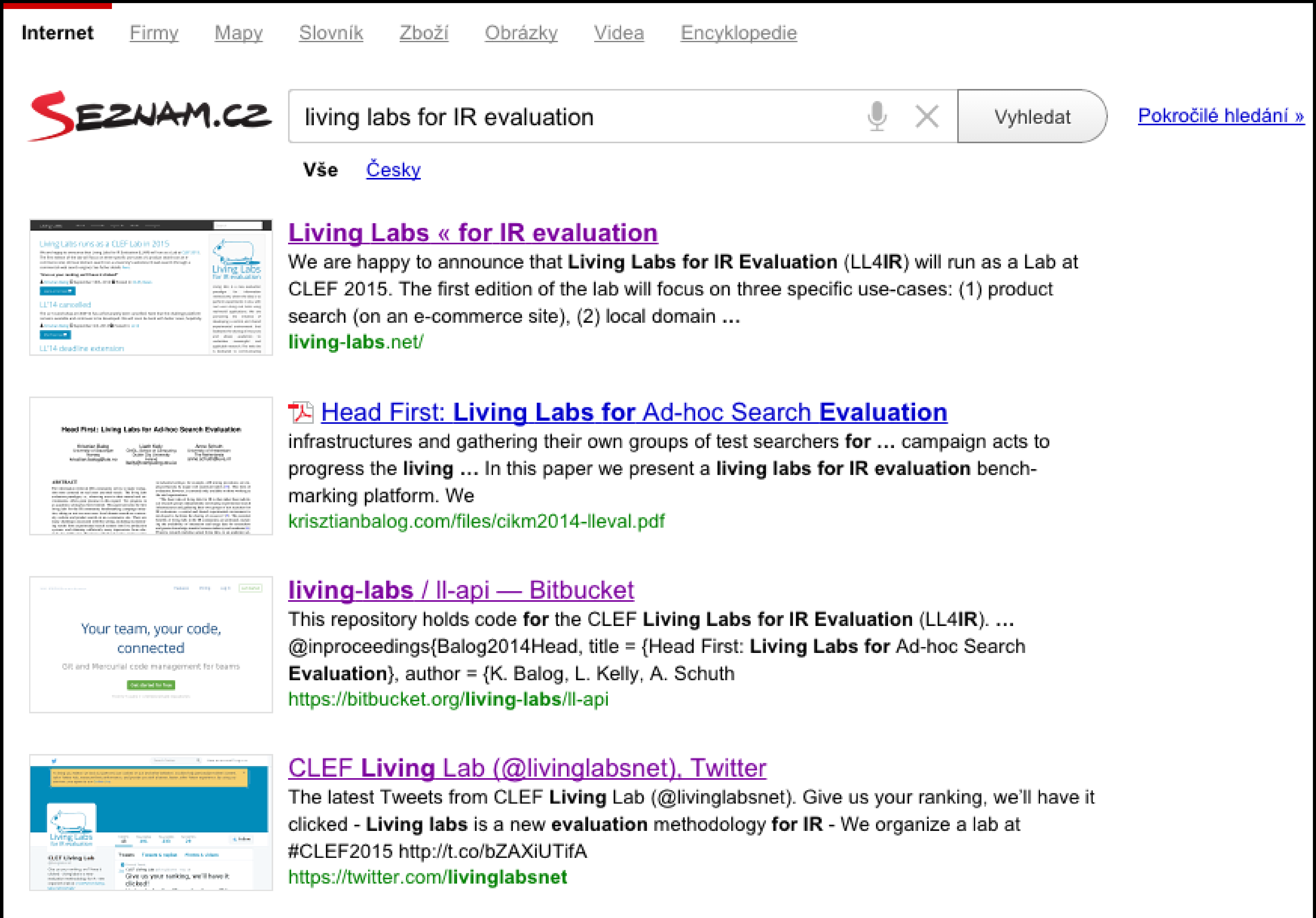
Product Search - Results - #2



Submission	Outcome	#Wins	#Losses	#Ties	#Impressions	p-value
Baseline	0.5284	93	83	598	774	0.498
Expected Outcome	0.5					
UiS-Jern	0.4795	82	89	596	767	0.646
GESIS	0.4520	80	97	639	816	0.229
UiS-Mira	0.4389	79	101	577	757	0.117
UiS-UiS	0.4118	84	120	527	731	0.014
IRIT	0.3990	79	119	593	791	0.005

Probabilistic Retrieval Models

Web Search



living labs for IR evaluation

Living Labs « for IR evaluation

We are happy to announce that **Living Labs for IR Evaluation** (LL4**IR**) will run as a Lab at CLEF 2015. The first edition of the lab will focus on three specific use-cases: (1) product search (on an e-commerce site), (2) local domain ...

living-labs.net/

Head First: **Living Labs for** Ad-hoc Search **Evaluation**

infrastructures and gathering their own groups of test searchers **for** ... campaign acts to progress the **living** ... In this paper we present a **living labs for IR evaluation** benchmarking platform. We

krisztianbalog.com/files/cikm2014-lleval.pdf

living-labs / ll-api — Bitbucket

This repository holds code **for** the CLEF **Living Labs for IR Evaluation** (LL4**IR**). ... @inproceedings{Balog2014Head, title = {Head First: **Living Labs for** Ad-hoc Search **Evaluation**}, author = {K. Balog, L. Kelly, A. Schuth

https://bitbucket.org/**living-labs**/ll-api

CLEF **Living** Lab (@livinglabsnet). Twitter

The latest Tweets from CLEF **Living** Lab (@livinglabsnet). Give us your ranking, we'll have it clicked - **Living labs** is a new **evaluation** methodology **for IR** - We organize a lab at #CLEF2015 http://t.co/bZAXiUTifA

https://twitter.com/**livinglabsnet**

Web Search

- Learning to Rank setting

Web Search

- Learning to Rank setting
  - 557 features

Web Search

- Learning to Rank setting
    - 557 features
- >35K documents

Web Search

- Learning to Rank setting
    - 557 features
- >35K documents
- >0.5M impressions

Web Search - Participation

Web Search - Participation

- 6 teams submitted runs for train queries

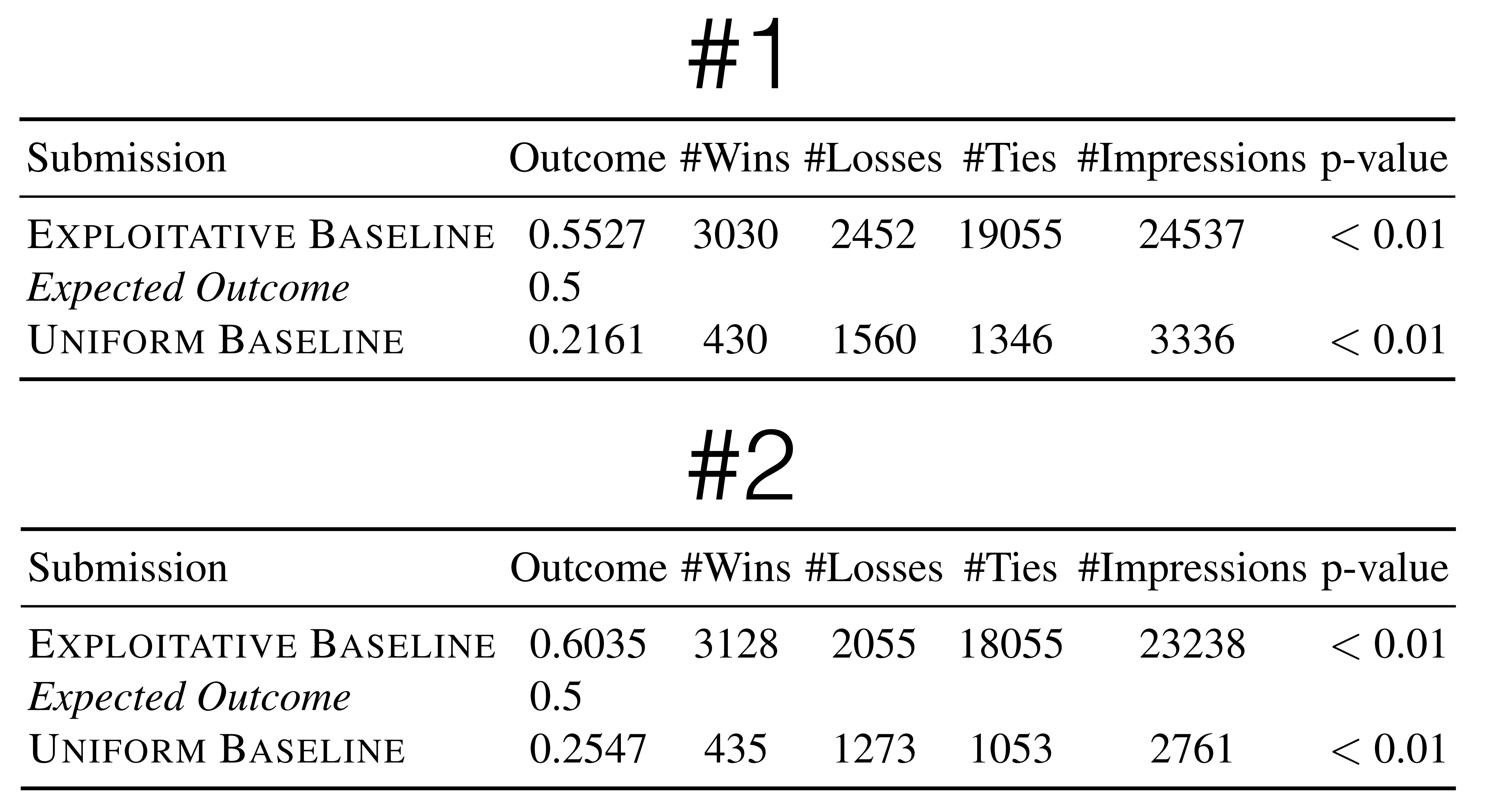
Web Search - Participation

- 6 teams submitted runs for train queries
- 0 teams submitted runs for test queries

Web Search - Participation

- 6 teams submitted runs for train queries
- 0 teams submitted runs for test queries
    - We report only baselines

Web Search - Results



#1

Submission	Outcome	#Wins	#Losses	#Ties	#Impressions	p-value
Exploitative Baseline	0.5527	3030	2452	19055	24537	< 0.01
Expected Outcome	0.5					
Uniform Baseline	0.2161	430	1560	1346	3336	< 0.01

#2

Submission	Outcome	#Wins	#Losses	#Ties	#Impressions	p-value
Exploitative Baseline	0.6035	3128	2055	18055	23238	< 0.01
Expected Outcome	0.5					
Uniform Baseline	0.2547	435	1273	1053	2761	< 0.01

the lab

use cases

conclusions

the lab
use cases
conclusions

Conclusions

- The first evaluation setup of its kind

Conclusions

- The first evaluation setup of its kind
    - With real users, real task, real interactions

Conclusions

- The first evaluation setup of its kind
    - With real users, real task, real interactions
- Two implemented use cases so far

Conclusions

- The first evaluation setup of its kind
    - With real users, real task, real interactions
- Two implemented use cases so far
    - Web search and Product search

Conclusions

- The first evaluation setup of its kind
    - With real users, real task, real interactions
- Two implemented use cases so far
    - Web search and Product search
- Developed an API (code publicly available)

Conclusions

- The first evaluation setup of its kind
    - With real users, real task, real interactions
- Two implemented use cases so far
    - Web search and Product search
- Developed an API (code publicly available)
- Interest from many teams

Conclusions

- The first evaluation setup of its kind
    - With real users, real task, real interactions
- Two implemented use cases so far
    - Web search and Product search
- Developed an API (code publicly available)
- Interest from many teams
    - Participation from some

Conclusions

- The first evaluation setup of its kind
    - With real users, real task, real interactions
- Two implemented use cases so far
    - Web search and Product search
- Developed an API (code publicly available)
- Interest from many teams
    - Participation from some
- No baselines were beaten, yet

Call

- If you …
    - … own
    - … work at
    - … collaborate with
- … a search engine, please consider joining LL4IR!

Call

- If you do IR research: participate

Call

- If you do IR research: participate
    - it is easy (example code runs out of the box)

Call

- If you do IR research: participate
    - it is easy (example code runs out of the box)
    - we run evaluation periods every 2 weeks

Call

- If you do IR research: participate
    - it is easy (example code runs out of the box)
    - we run evaluation periods every 2 weeks
    - next period in less then a week (plenty of time!)

Call

- If you do IR research: participate
  - it is easy (example code runs out of the box)
  - we run evaluation periods every 2 weeks
  - next period in less then a week (plenty of time!)
- Come to our Lab session Thursday Afternoon

Thank You

#LL4IR

Anne Schuth
University of Amsterdam

Krisztian Balog
University of Stavanger

Liadh Kelly
Trinity College Dublin

Living Labs
for IR evaluation

Thanks to:
- CLEF
- ESF ELIAS
- COMMIT
- REGIO Jatek
- Seznam

http://living-labs.net
@livinglabsnet