"Give us your ranking, we'll have it clicked!"

Living Labs for IR Evaluation

LL4IR@CLEF'15

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

News

News

- Funding from EFS ELIAS
    - This meeting
    - Developing the API
- Funding from Microsoft Azure
    - For hosting the API
- Lots of improvements of the API
    - Tracking of errors
- Lots of interest from site that may want to join
    - Several academic search engines?
    - Recipe search?

Introduction

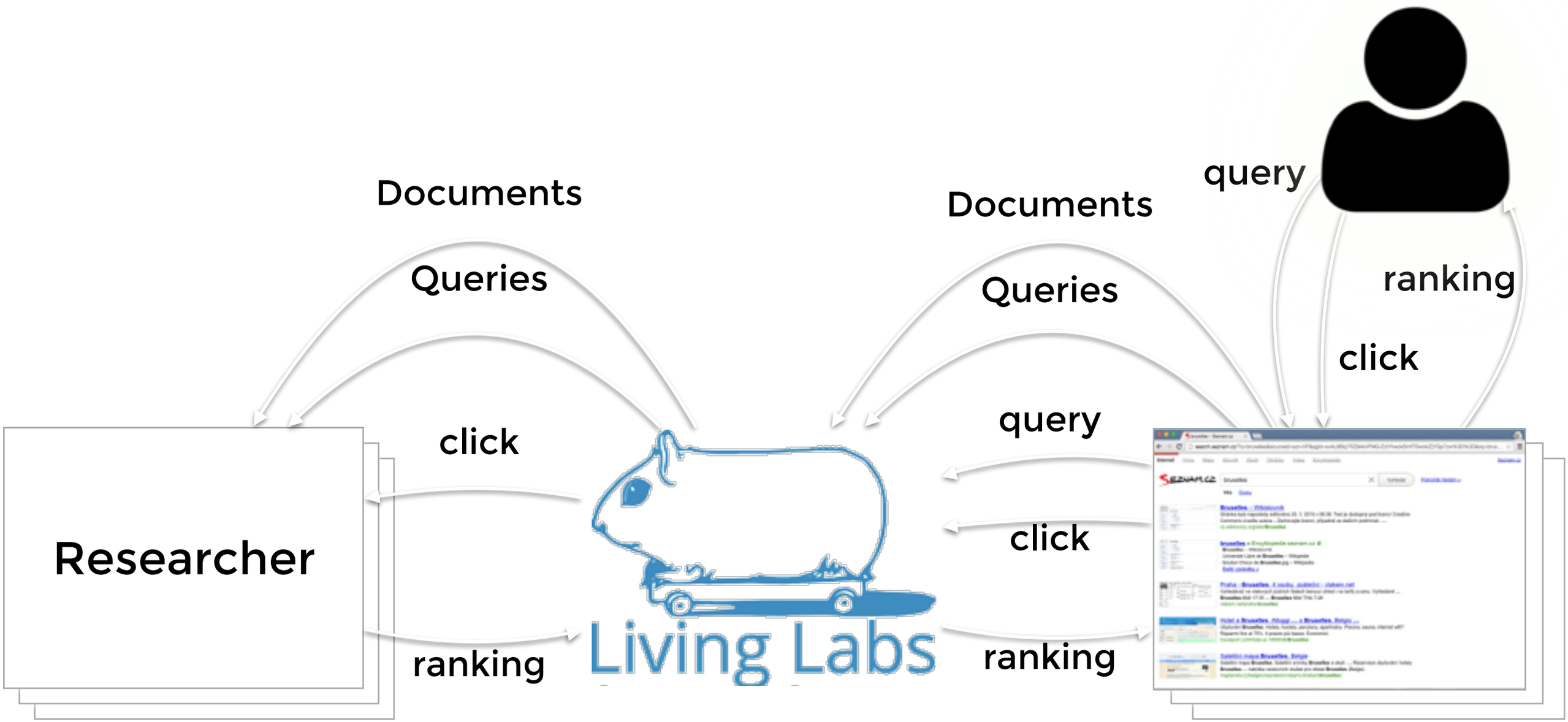
Overview

- **Overall goal**: make information retrieval evaluation more realistic
  - Evaluate retrieval methods in a *live setting* with *real users* in their *natural task environments*
- **Focus**: medium to large sized organizations with fair amount of search volume
  - Typically lack their own R&D department, but would gain much from improved approaches
  - Or, would like to collaborate with academic researchers

Key idea

- Focus on frequent (head) queries
  - Enough traffic on them (both real-time and historical)
  - Ranked result lists can be generated offline
- An API orchestrates all data exchange between live sites and experimental systems
- **Head First: Living Labs for Ad-hoc Search Evaluation**. Balog et al. CIKM'14.

Methodology

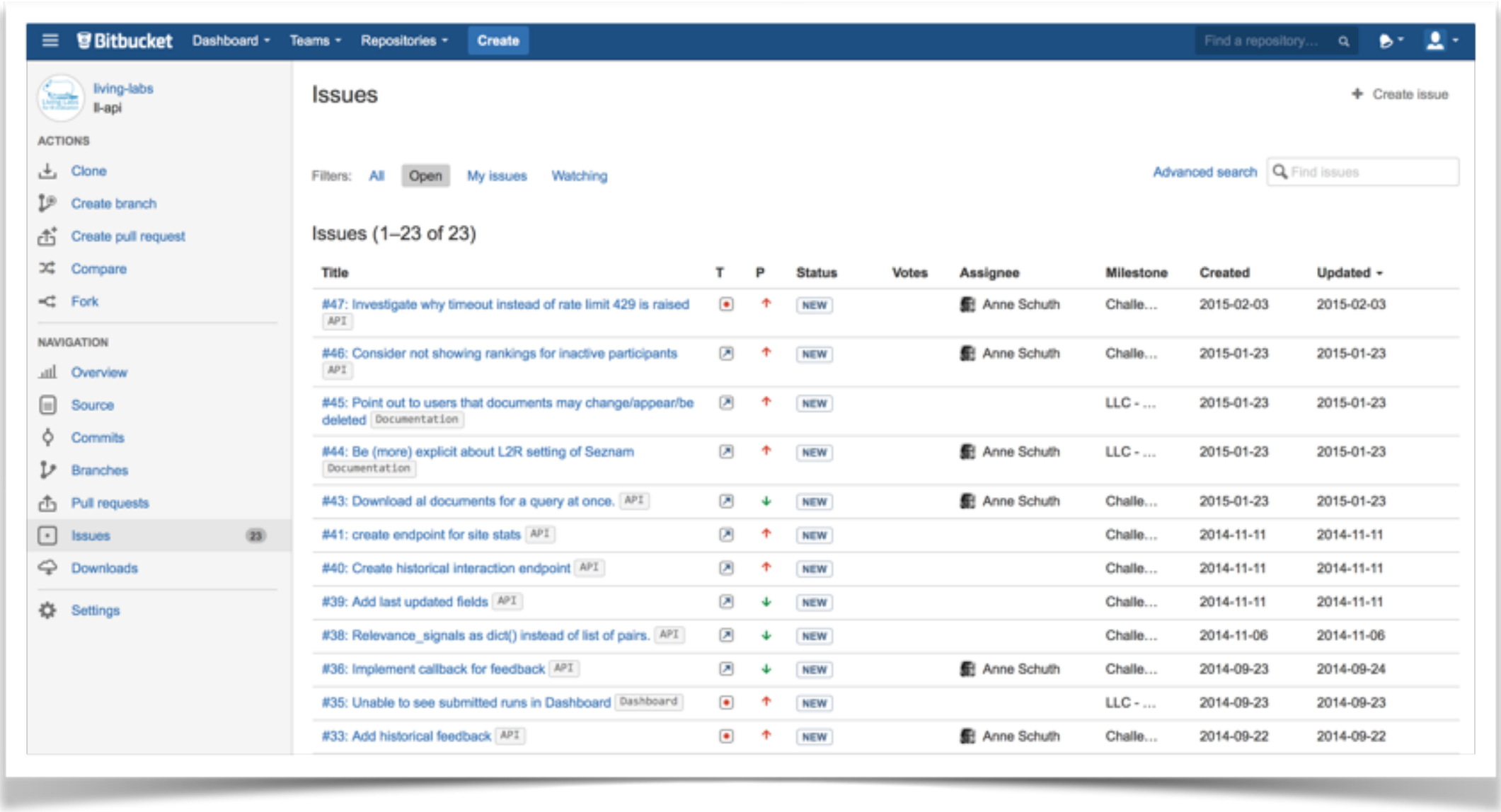


Use cases

- ~~Three~~ Two ad-hoc search tasks

	~~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

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



please report issues here!

API doc: **doc.living-labs.net**

Living Labs API CLEF Lab Dashboard Documentation

Living Labs Documentation

Guide for CLEF participants

Living Labs API CLEF Lab Dashboard Documentation Search

1. Guide for CLEF Participants

Note
This guide is being updated as it is being used. Please tell us what you think is missing. Our contact details are at the bottom of this page New

This guide is meant to be a practical guide to participating in the CLEF Living Lab. Since we deviate significantly from the typical TREC style evaluation setup that most participants are likely to be familiar with, we will focus primarily on those differences.

Participating in the lab involves following these steps:

1. Read the lab description and *Key Concepts* below. Make sure you're *Getting Help* when needed.
2. Sign up:
    1. Register at CLEF.
    2. Register with the lab. You can do this at any moment until the test phase begins. New
    3. Sign and send the lab the agreement form. You will receive a link to this form.
    4. Sign up for individual sites (use-cases) you want to obtain data for. You will receive a link by email to do so.
3. Implement your method as a client that can talk to the API. Examples are provided. See *Implement a Client* below.
4. Run your client:
    1. The client you implement should probably run continuously over several weeks and can potentially constantly update runs.
    2. When the test phase starts, download test queries and submit your test runs. Again, the test phase will last for several weeks but there is no need (nor the possibility) to update runs.
5. Write up your findings. Publication details will become available.
6. Come to and present your work at CLEF 2015 in Toulouse, France in September 2015.

We hope that all steps but 3. and 4. are self explanatory. Below we detail these two steps in Sections *Implement a Client* and *Running a Client* respectively.

1.1. Schedule

Date	Description
1 Nov, 2014	Training period begins (Note that you can join any time after this date!) New
1-15 Apr, 2015	Uploading test runs
15 Apr, 2015	Testing period begins
15 May, 2015	Testing period ends
17 May, 2015	Results released

Dashboard: **living-labs.net:5001**

Living Labs API | CLEF Lab | Dashboard | Documentation

Dashboard | Home | Participants | Sites | My | UiS

Queries for REGIO Jatek

Qid	Site_qid	Querystring
R-q1	2c36a0df0f6b0d161a38504864a1109df9571543	monster high
R-q2	0d58e9f46857cd209c665fae08f73052371fea0c	magnetiz
R-q3	30ca705660eda1ae8db8f87b77ae07bc69d8ecb5	duplo
R-q4	d0e70782e0c61e9be38754882329503979 3ed0dc	lego friends
R-q5	8c3f47547b5eb2d5ead87215b81e12deed11c21e	geomag
R-q46	dd674cba73acd04b151f80e0df8f5b3e8f31353b	angry
R-q47	0b46210c8e4c32fad3901bd686f22ce32715c592	interaktív
R-q48	ce59527338c60184f609e0127cc523149b1acdc9	minnie
R-q49	6cf635a089c2bbb20bce1874572a8f8754dfee51	busz
R-q50	3c994c32de03650422c00af78bfb14f079144f56	my little p
R-q38	5bc238dd91088541300285533 1f89bb2167f7303	kártya
R-q39	f6e84b5adacb1fce7df0257c34860e5e0e5475bd	találd ki
R-q40	74bc49d2651e35ecf2efe215d5fa90dc16814d29	gyöngy
R-q41	40ff5708d0a6934bbfd5250fc07a77ce0d9dee8e	ugráló
R-q42	28b6ea3446e0f0e997df894f5e6169d16f00e74e	kisvakond

Tekerős autópálya 1:43 9 m

Méret: pálya hossza 9 m,autók hossza kb. 9 cm,pálya kiterjedése 176 x 182 cm

~~19 995 Ft~~ helyett
12 495 Ft

Kosárba

Cikkszám	11645
Ajánlott életkor	3-6 éves korig


```
{
  "content": {
    "category": "Aut\xf3p\xe1lya, parkol\xf3h\xe1z",
    "main_category": "Aut\xf3k, j\xe1rm\u0171vek",
    "main_category_id": "2",
    "description": "Minden kisfi\xfa v\xe1gyik r\xe1, hogy [...]",
    "price": 19995.0,
    "bonus_price": 12495.0,
    "product_name": "Teker\u0151s aut\xf3p\xe1lya 1:43 9 m",
    "short_description": "M\xe9ret: p\xe1lya hossza 9 m [...]",
    "category_id": "18",
    "brand": ""
  },
  "site_doc_id": "11645",
  "title": "Teker\u0151s aut\xf3p\xe1lya 1:43 9 m"
}
```

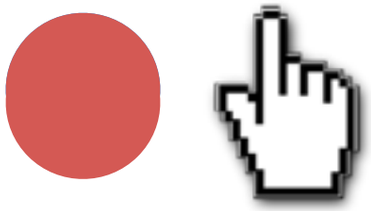
Evaluation

- **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

Researcher is preferred over Production



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

Evaluation

- Test periods
  - Last two weeks of every month
- Same set of queries
- Runs will expire
  - This is new behavior
  - Meant to not waste query impressions

Results

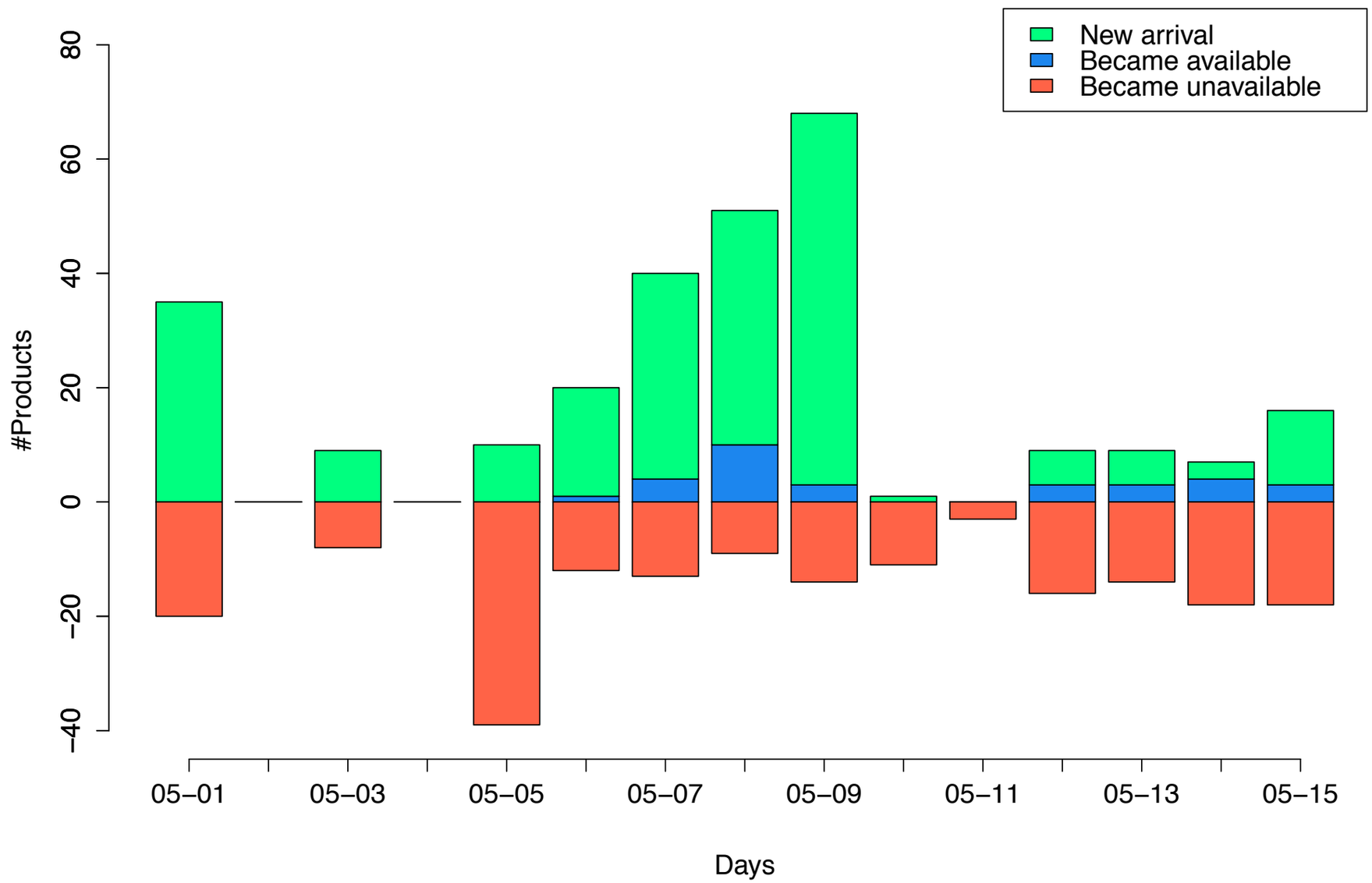
Participants

- 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

Results

Product Search

Product Search - Inventory



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

    (we *estimated* it became 0.28)

Results - Product Search

Round 1 – **Official CLEF Round**

Teamname	Outcome	#Wins	#Losses	#Ties	#Impressions
Baseline	0.4691	91	103	467	661
UiS-Mira	0.3413	71	137	517	725
UiS-Jern	0.3277	58	119	488	665
UiS-UiS	0.2827	54	137	508	699
Expected	*0.28*				
GESIS	0.2685	40	109	374	523

Results - Product Search

Round 2 – June 2015

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

Results - Product Search

Round 3 – July 2015

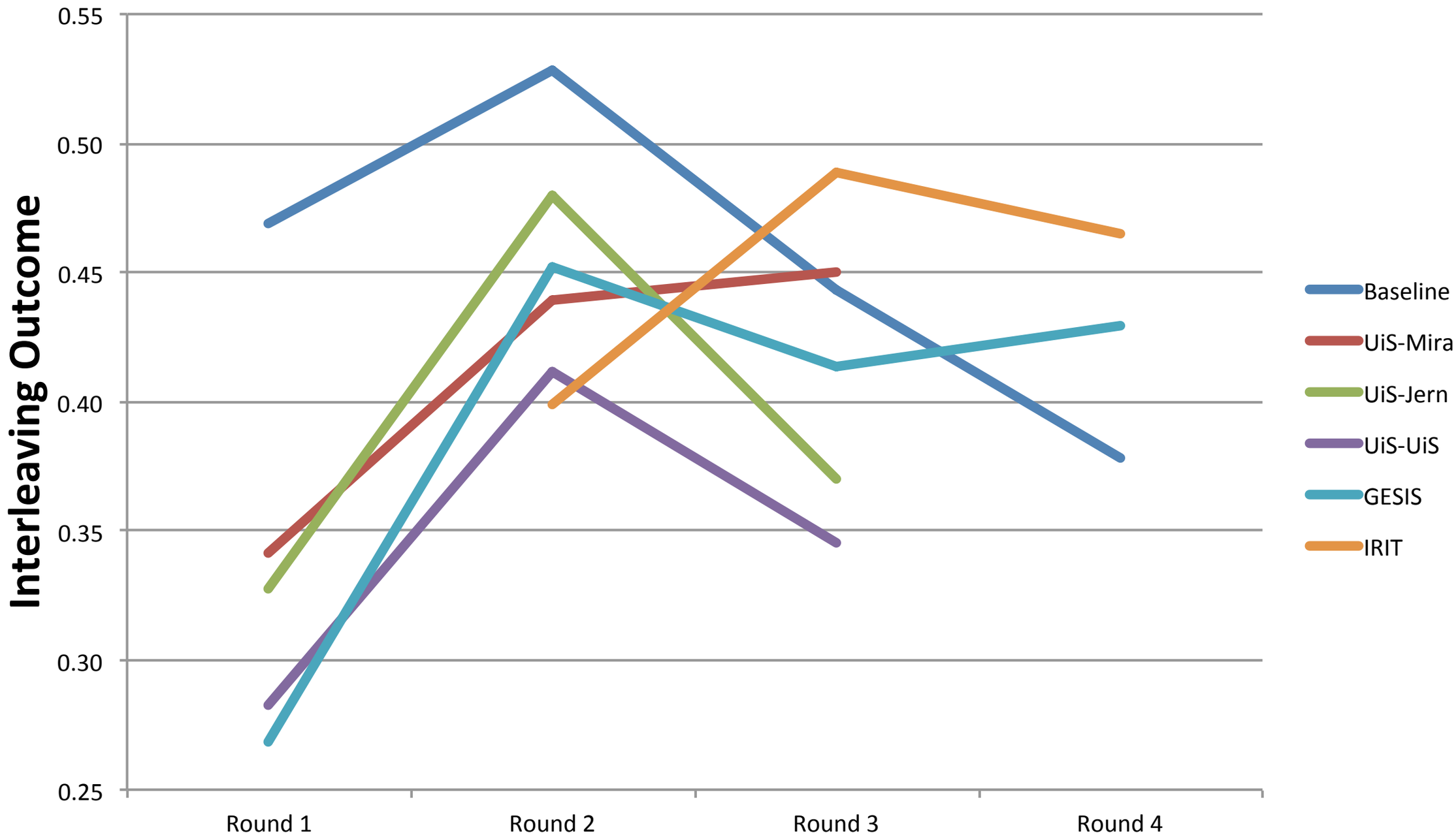
Teamname	Outcome	#Wins	#Losses	#Ties	#Impressions
Expected	0.5				
IRIT	0.4890	89	93	533	715
UiS-Mira	0.4507	64	78	527	669
Baseline	0.4430	66	83	498	647
GESIS	0.4134	74	105	513	692
UiS-Jern	0.3702	67	114	511	692
UiS-UiS	0.3459	55	104	521	680

Results - Product Search

Round 4 – August 2015

Teamname	Outcome	#Wins	#Losses	#Ties	#Impressions
Expected	0.5				
IRIT	0.4654	101	116	767	984
GESIS	0.4292	103	137	804	1044
Baseline	0.3783	87	143	781	1011

Results - Product Search



Results

Web Search

Results – Web Search

Round 1 – Official CLEF Round

Teamname	Outcome	#Wins	#Losses	#Ties	#Impressions
Exploitative Baseline	0.5527	3030	2452	19055	24537
Expected	*0.5*				
Uniform Baseline	0.2161	430	1560	1346	3336

Results – Web Search

Round 2 – June 2015

Teamname	Outcome	#Wins	#Losses	#Ties	#Impressions
Exploitative Baseline	0.6035	3128	2055	18055	23238
Expected	*0.5*				
Uniform Baseline	0.2547	435	1273	1053	2761

Results – Web Search

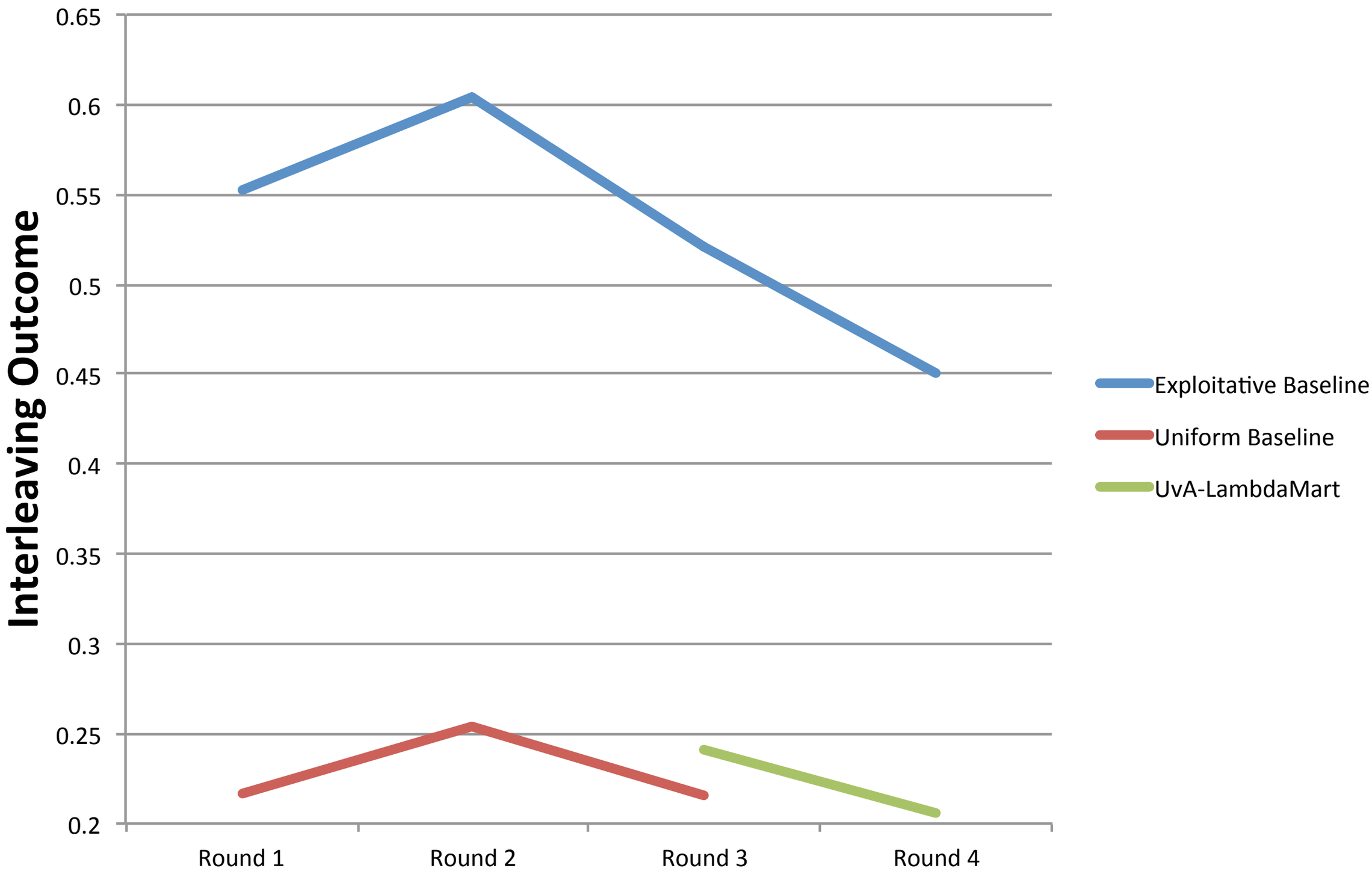
Round 3 – July 2015

Teamname	Outcome	#Wins	#Losses	#Ties	#Impressions
Exploitative Baseline	0.5203	2161	1992	13206	17359
Expected	0.5				
UvA-LambdaMart	0.2405	2264	7148	7863	17275
Uniform Baseline	0.2157	313	1138	922	2373*

Results – Web Search

Round 4 – August 2015

Teamname	Outcome	#Wins	#Losses	#Ties	#Impressions
Expected	0.5				
Exploitative Baseline	0.4500	18	22	134	174
UvA-LambdaMart	0.2059	21	81	89	191


Interleaving Outcome
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
Round 1
Round 2
Round 3
Round 4
Exploitative Baseline
Uniform Baseline
UvA-LambdaMart

Goals

Goals of this Meeting

- Share findings
- Identify obstacles / problems / confusion
- Establish future directions

Future

Our Future

- We will continue
    - Next year at CLEF?
- New Use Cases
    - Academic Search
    - Recipe Search
- New Task?
- Non-head queries?
- Other metrics?
- Relation between online and offline
    - Write your SIGIR paper

Today

Today’s programme

~~16:00-16:10 Introduction to the lab~~

16:10-16:25 Regio use case presentation

16:25-16:40 Seznam use case presentations

16:40-17:25 Lab participants presentations

GESIS, IRIT, UIS (10min each)

17:25-17:35 Questions to partipants

17:35-18:00 Discussion session and wrap-up