University of Glasgow

# Catching up with Industry - Online Evaluation of Information Access Systems

**Frank Hopfgartner** **@FTHopf**

University of Glasgow

# We have a problem!

## If SIGIR had an Academic Track, what would be in it?

David Hawking
Microsoft (Bing)
Canberra, Australia
david.hawking@acm.org

**ABSTRACT**

It used to be the case that very little industry research was presented at SIGIR. Now the balance has radically changed – many accepted papers have industry authors and many rely on industry data sets – To the extent that a leading academic member of the SIGIR community has light-heartedly proposed the creation of an Academic Track.

Behind the levity lies the important question of how a researcher can make a meaningful contribution to the field, in the absence of petabyte-scale sets of documents and massive user-interaction logs. Theoretical contributions can revolutionize thinking, but have greatest impact when applicable in practice, and when empirically validated.

In my years at Funnelback and more recently at Microsoft I have been very aware of high-impact but not-well-solved IR problems involving relatively tiny datasets. Many of them are characterized by sparsity of user interaction data and are hence not well-suited to simple machine learning approaches or to large scale A/B testing. My talk will illustrate and attempt to characterize these problems and to suggest fruitful areas for academic research.

If time permits, I will mention some areas in which academic research has contributed to current large-scale industry practice.

**Categories and Subject Descriptors**

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*relevance feedback*

**Keywords**

Open IR problems for academics.

**BIO**

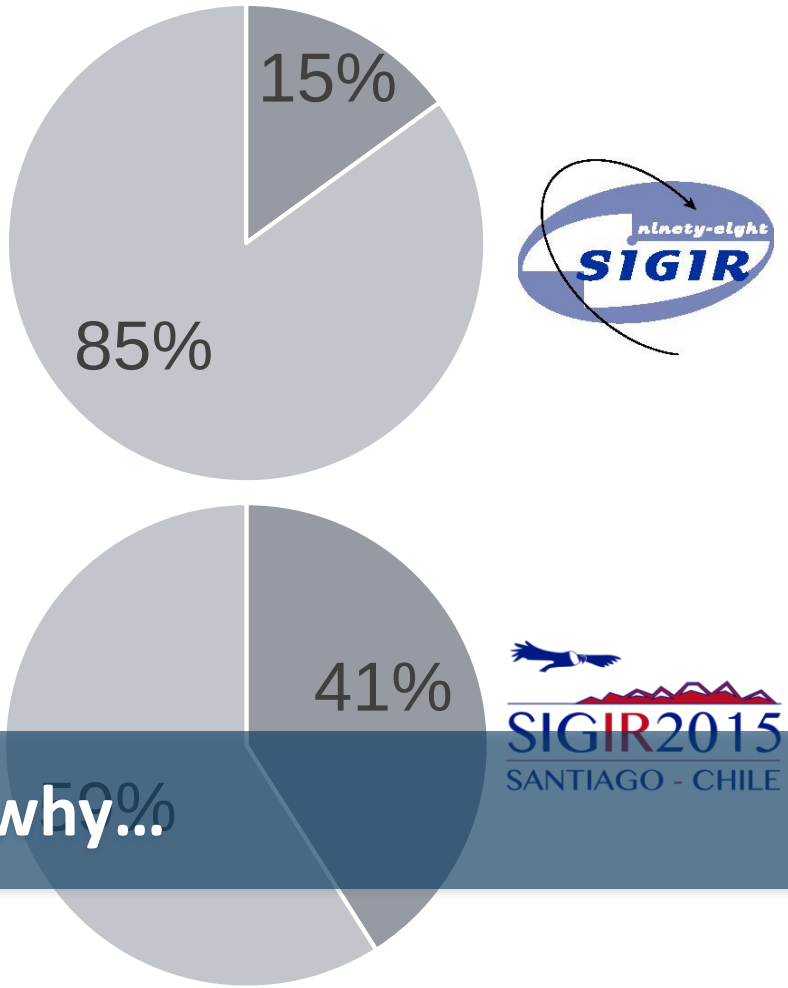
David Hawking is an applied scientist in the Relevance & Intent group of Microsoft. He is based in Canberra and works on Bing.

David has been an IR researcher since 1991 and, among other things, has worked on parallel IR, distributed IR, term proximity models, efficiency, automatic assessment of health webpage quality, evaluation, test collection development, web IR and enterprise search. In 1998 he left the Australian National University to continue his research at Australia's largest research organization CSIRO. CSIRO's industrial focus led to the commercialisation of the team's research and eventually to the spinning off of the Funnelback internet and enterprise search company in late 2005. David joined the spinoff as Chief Scientist in 2008 and worked in that role until joining Microsoft in 2013.

David continues as an adjunct professor at the Australian National University and has graduated nine PhDs. He has authored around a hundred publications in the IR area (see `david-hawking.net/publications.shtml`) and twice served as program chair for SIGIR. He chaired the Industry Track at SIGIR 2014 and was Web Track coordinator at the international Text Retrieval Conference (TREC) from 1997-2004. In this role he was responsible for the creation and distribution of widely used text retrieval test collections. He holds an honorary doctorate from the University of Neuchâtel (Switzerland), the 2005 Chris Wallace award (Australasia) for computer science research and the 2012 Tony Kent Strix award (jointly with Doug Cutting).


*SIGIR'15*, August 09 - 13, 2015, Santiago, Chile.
ACM 978-1-4503-3621-5/15/08
http://dx.doi.org/10.1145/2766462.2776784





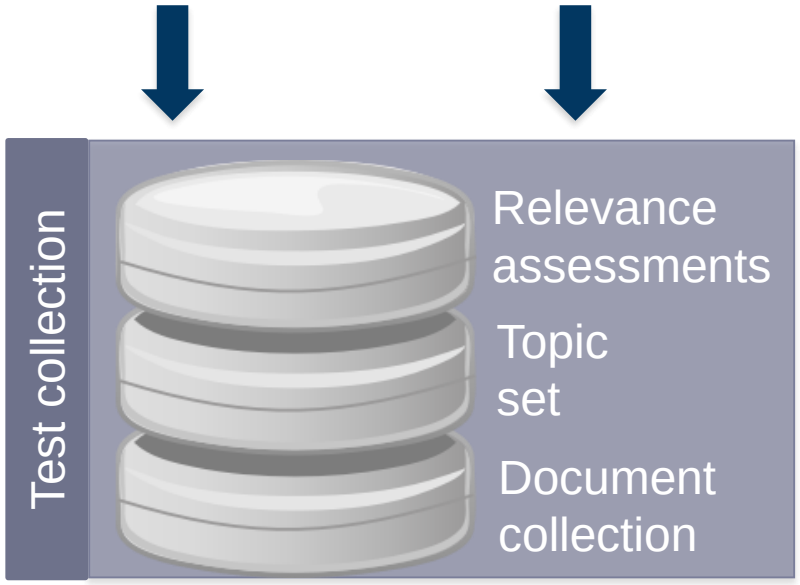

ninety-eight SIGIR

SIGIR2015
SANTIAGO - CHILE

**And here is why…**

University of Glasgow

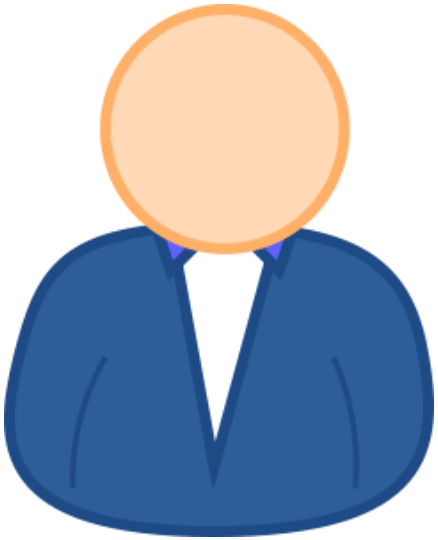
# How do Academics evaluate information access systems?



University of Glasgow

# Academic Evaluation Protocol (Cranfield)

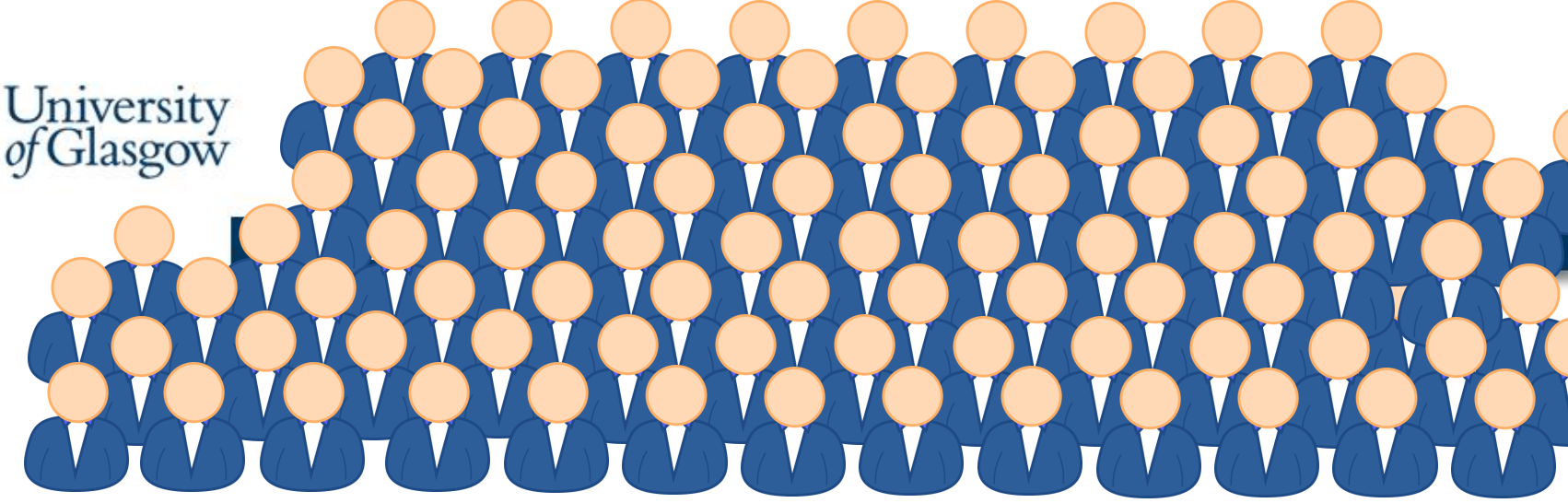
Find as many documents as possible for a given search task!

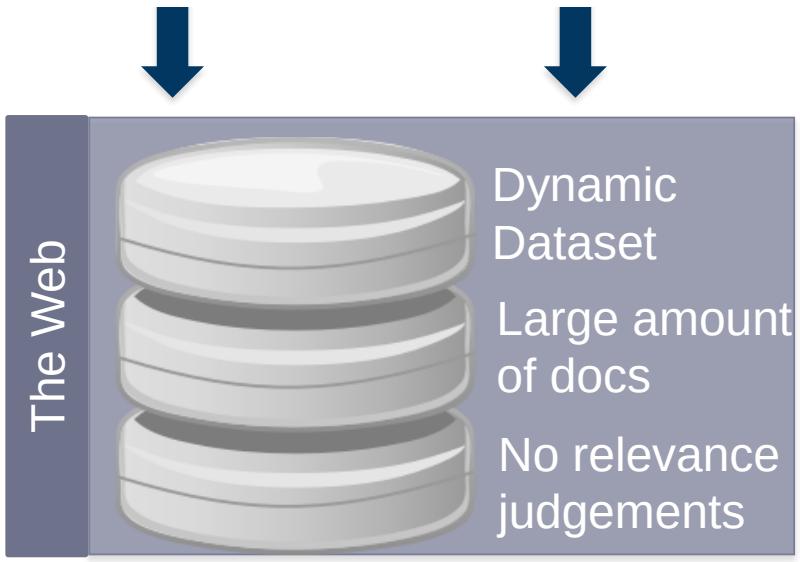


I tell you what is relevant!

Act naturally while I watch everything you are doing!

University of Glasgow

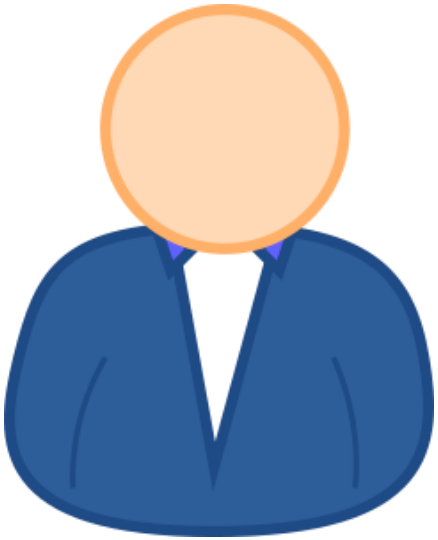




University of Glasgow

# Industry Evaluation Protocol (A/B Testing)

Continue using our information access service



Use it for whatever you want to use it for

Use it whenever and wherever you want

University of Glasgow

## Ron Kohavi's Keynote at KDD'15

# Example 1: MSN Home Page Search Box

- OEC: Clickthrough rate for Search box and popular searches

A


B


Differences: A has taller search box (overall size is the same),
has magnifying glass icon, "popular searches"
B has big search button, provides popular searches without calling them out

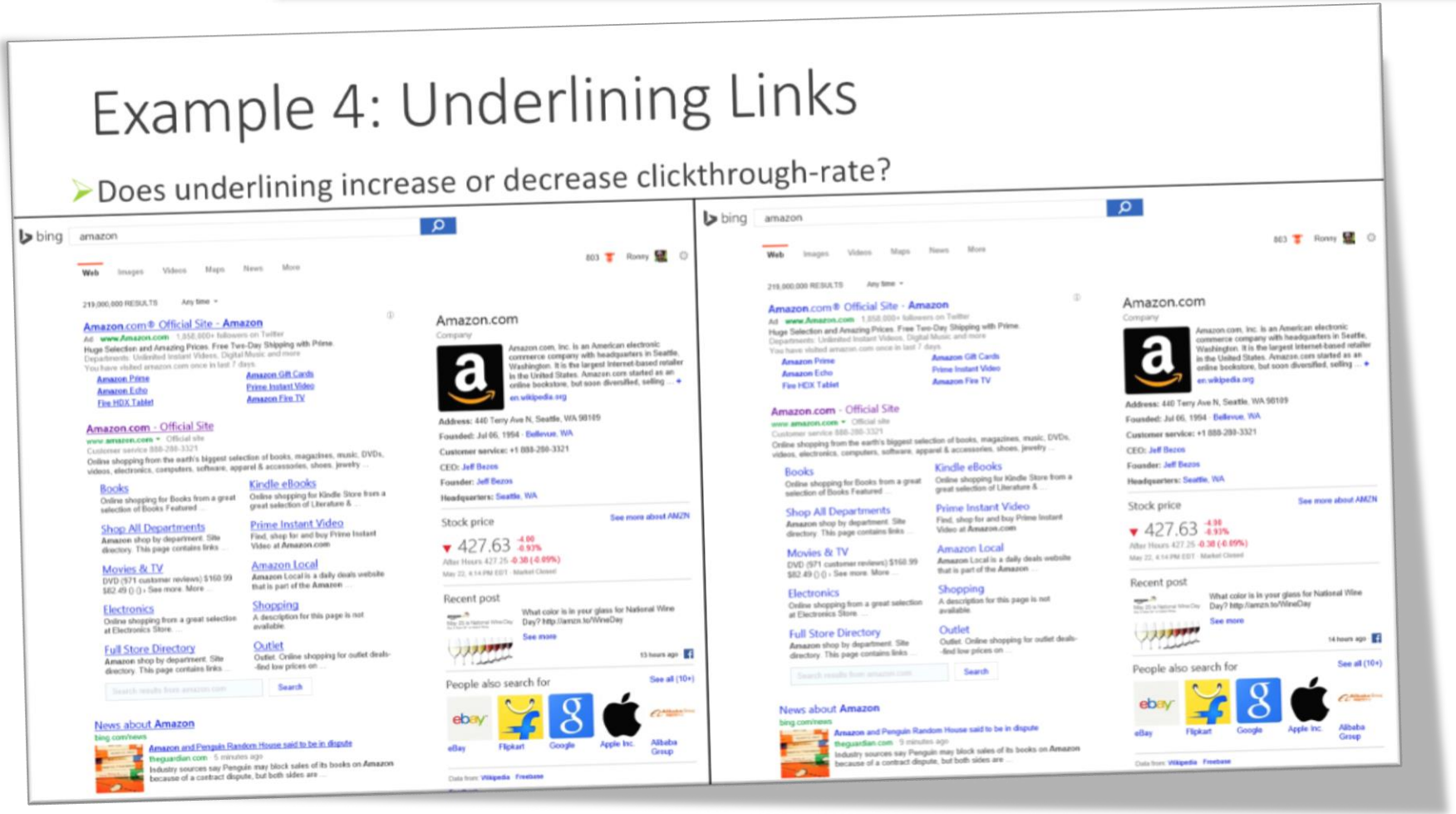
Ronny Kohavi

11

R. Kohavi
*"Online Controlled Experiments: Lessons from Running A/B/n Tests for 12 Years, In Proc. KDD'15, Sydney, Australia*, p. 1, 2015.

University of Glasgow

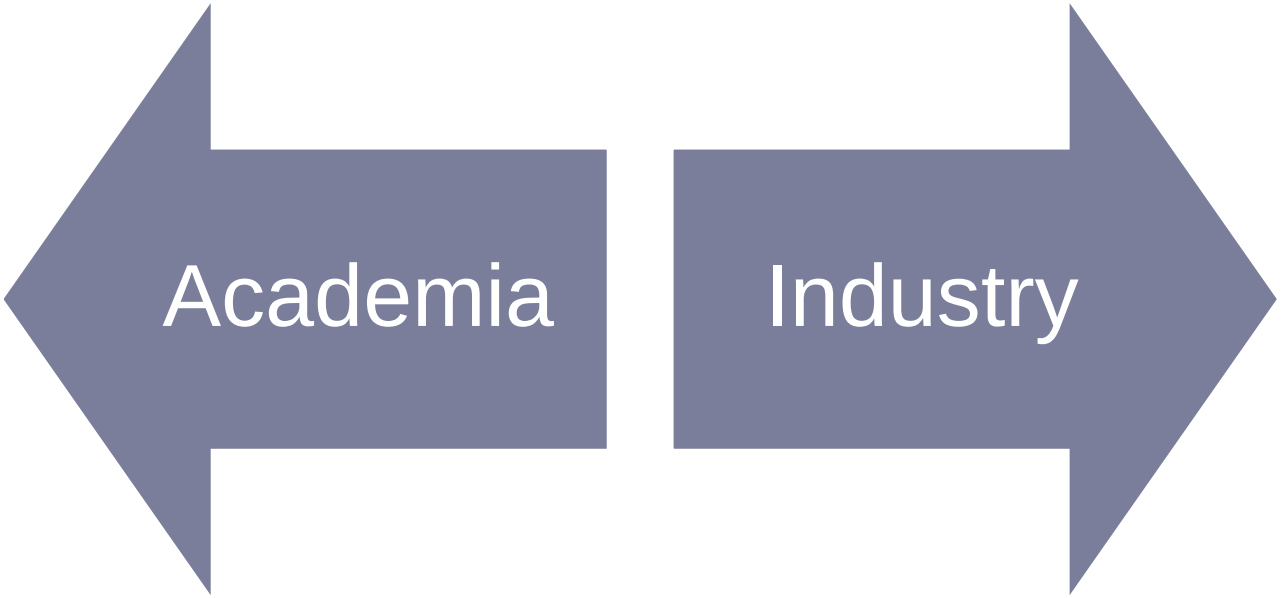
# Ron Kohavi's Keynote at KDD'15




R. Kohavi
*"Online Controlled Experiments: Lessons from Running A/B/n Tests for 12 Years, In Proc. KDD'15, Sydney, Australia,* p. 1, 2015.

University of Glasgow

## Evaluation Gap




- Static, often rather old datasets
- Offline Evaluation
- Atypical search task
- Observer-expectancy effect
- Missing context/background
- Missing incentive to satisfy users' information needs

- Dynamic dataset
- Online A/B testing
- Real users
- Real users!
- Real users!!
- Real users!!!

University of Glasgow

## Narrowing the growing gap between Academia and Industry

Do we want **real users** who interact with a system following their own information need?

Do we want **realistic** settings where users are not restricted by closed laboratory conditions?

Do we want to test **scalability** of our systems and algorithms?

**Yes**

**But how do we get access to users and infrastructure?**

University of Glasgow

## Option 1: Industry-inclined PhD research

"Doctorates in cooperation with industry do have great value, but must not be seen as a panacea to better prepare doctoral graduates for the job market."

League of European Research Universities

Source: http://www.leru.org/index.php/public/news/society-needs-talent-from-universities/

University of Glasgow

## Option 2: Living Labs

"A **living laboratory** on the Web that brings researchers and searchers together is needed to facilitate ISSS (Information-Seeking Support System) evaluation."

D. Kelly, S. Dumais, J. O. Pedersen.
"*Evaluation Challenges and Directions for Information Seeking Support Systems*", Computer, 42(3):60-65, 2009.

University of Glasgow

## Overview

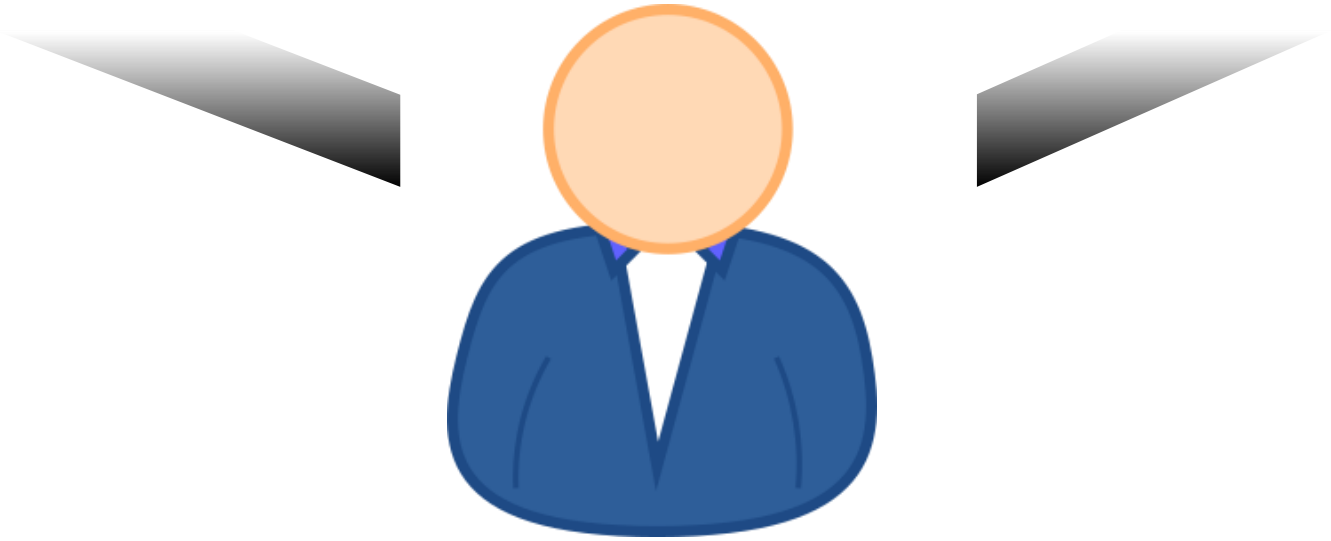
- Introduction
- **Living Labs**
- News Recommendation Evaluation Lab (NewsREEL) Scenario
- NewsREEL 2016
- NewsREEL and You

University of Glasgow

## What are living labs?

Real-life test and experimentation environment to fill the **pre-commercial gap** between fundamental research and innovation.

Rely on feedback from real users to develop convincing demonstrators that **showcase potentials** of an idea or a product.



University of Glasgow

## Questions to be addressed

Who is interested in my product?

Is there a need for my product?

How will people really use the technology?

What is the willingness to pay?

What parameters do I need?

University of Glasgow

## Question

What should a living lab for the promotion of information access research look like?

University of Glasgow

# Search Data is Sensitive!

## What Revealing Search Data Reveals

AOL posted, but later removed, a list of the Web search inquiries of 658,000 unnamed users on a new Web site for academic researchers. An interview with one of those unnamed users, Thelma Arnold, combined with her data reveal what she was searching for, why and on which Web sites.

**A sample of Thelma Arnold's search data released by AOL**

| | | | | | |
|---|---|---|---|---|---|
| 4417749 | swing sets | 2006-04-24 | 15:39:30 | 4 | http://www.byoswingset.com |
| 4417749 | swing sets | 2006-04-24 | 15:39:30 | 9 | http://www.buychoice.com |
| 4417749 | swing sets | 2006-04-24 | 15:39:30 | 10 | http://www.creativeplaythings.com |
| 4417749 | swing sets | 2006-04-24 | 15:39:30 | 5 | http://www.childlife.com |
| 4417749 | swing sets | 2006-04-24 | 15:39:30 | 6 | http://www.planitplay.com |
| 4417749 | that do not shed | 2006-04-28 | 9:05:54 | 2 | http://www.gopetsamerica.com |
| 4417749 | dog who urinate on everything | 2006-04-28 | 13:24:07 | 6 | http://www.dogdaysusa.com |
| 4417749 | walmart | 2006-04-28 | 14:07:32 | 1 | http://www.walmart.com |
| 4417749 | womens underwear | 2006-04-28 | 14:12:28 | 10 | http://www.bizrate.com |
| 4417749 | jcpenny | 2006-04-28 | 14:16:05 | | |
| 4417749 | jcpenny | 2006-04-28 | 14:16:49 | 1 | http://www.jcpenney.com |
| 4417749 | tortus and turtles | 2006-04-29 | 13:12:47 | | |
| 4417749 | manchester terrier | 2006-05-02 | 9:05:31 | 1 | http://www.manchesterterrier.com |
| 4417749 | delta | 2006-05-02 | 11:49:26 | | |
| 4417749 | fingers going numb | 2006-05-02 | 17:35:47 | | |
| 4417749 | dances by laura | 2006-05-02 | 17:59:32 | | |
| 4417749 | dances by lori | 2006-05-02 | 17:59:57 | | |
| 4417749 | single dances | 2006-05-02 | 18:00:18 | 1 | http://solosingles.com |
| 4417749 | single dances in atlanta | 2006-05-02 | 18:01:13 | | |
| 4417749 | single dances in atlanta | 2006-05-02 | 18:01:50 | | |
| 4417749 | dry mouth | 2006-05-06 | 16:49:14 | 2 | http://www.mayoclinic.com |
| 4417749 | dry mouth | 2006-05-06 | 16:49:14 | 8 | http://www.wrongdiagnosis.com |
| 4417749 | tyroid | 2006-05-06 | 16:55:34 | | |
| 4417749 | thyroid | 2006-05-06 | 16:55:44 | | |
| 4417749 | competitive market analysis of homes in lilburn | 2006-05-14 | 12:14:52 | | |
| 4417749 | competitive market analysis of homes in lilburn | 2006-05-14 | 12:16:17 | | |
| 4417749 | competitive market analysis of homes in lilburn | 2006-05-14 | 12:16:43 | | |

**Why the search**

"I was thinking about my grandchildren"

"I was looking for some."

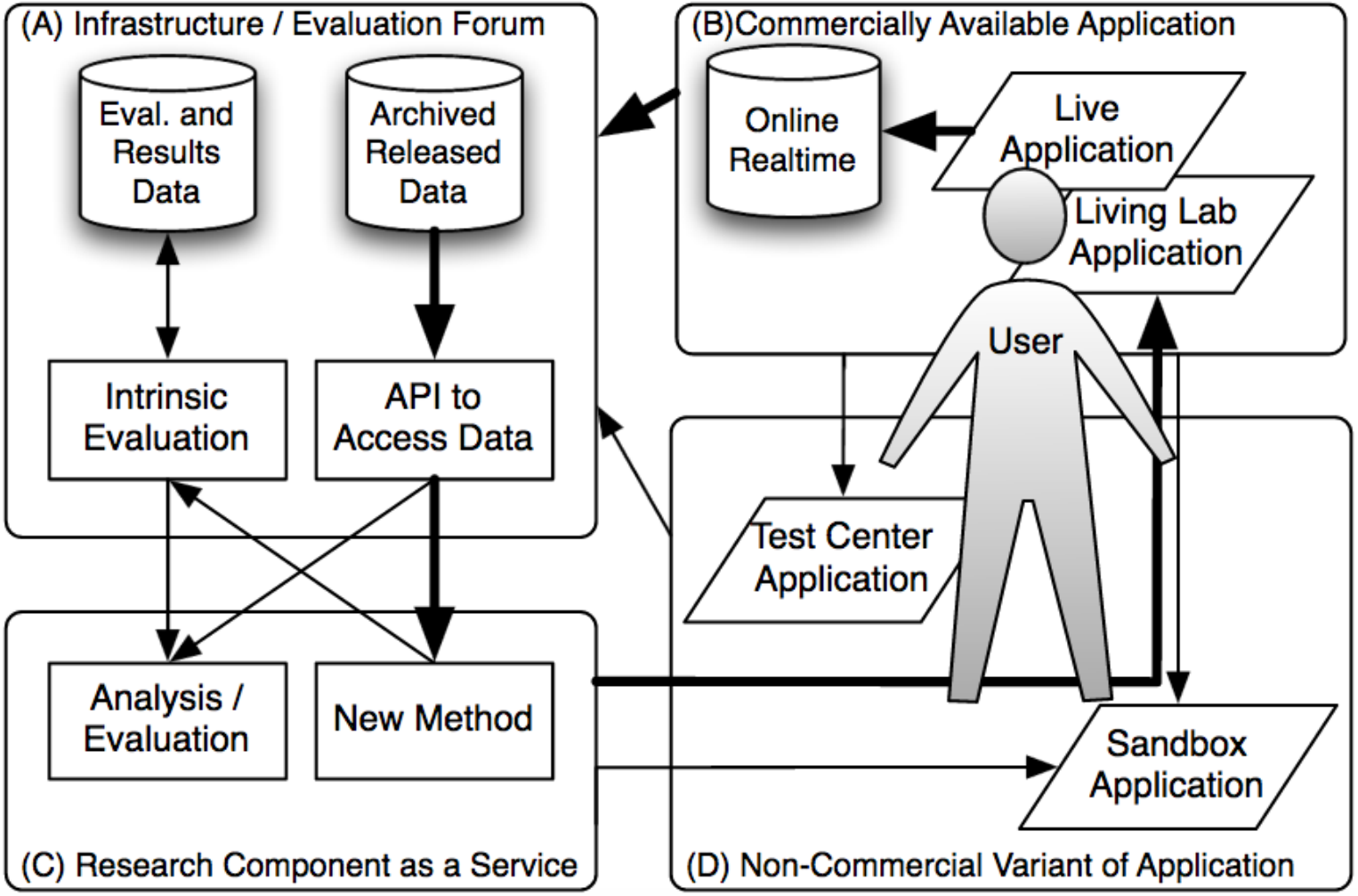
"A woman was in the [public] bathroom crying. She was going though a divorce. I thought there was a place called 'Dances by Lori,' for singles."

"I wanted to find out what my house was worth."

Source: http://www.nytimes.com/imagepages/2006/08/08/business/09aol-graphic.html

University of Glasgow

# Towards a Living Lab for Information Retrieval…




L. Azzopardi and K. Balog.
"*Towards a Living Lab for Information Retrieval Access Research and Development*", In Proc. CLEF'11, pages 26-37, 2011

University of Glasgow

## Living Labs as Academic Research Challenges

- 2013: News Recommender Systems Challenge, organised as workshop and challenge at ACM RecSys'13
- Since 2014: News Recommendation Evaluation Lab (NewsREEL), organised as part of CLEF
- 2016: Living Labs for Information Retrieval (LL4IR), organised as part of CLEF
- Since 2016: OpenSearch, organised as part of TREC
- Since 2016: Open Live Test for Question Retrieval, organised as part of NTCIR

University of Glasgow

# TREC OpenSearch 2016

## KNOWN ISSUES

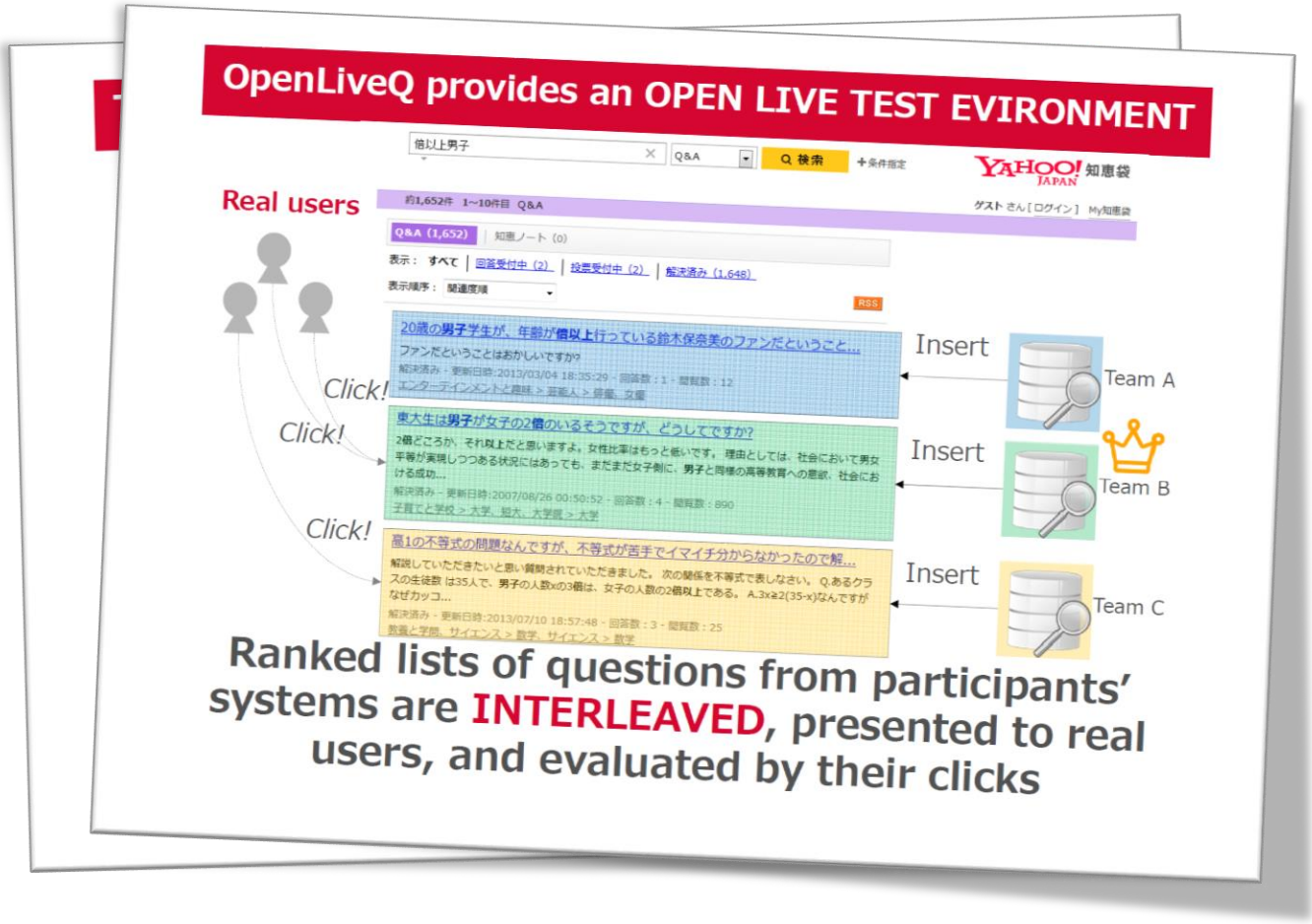
- Head queries only
  - Considerable portion of traffic, but only popular info needs
- Lack of context
  - No knowledge of the searcher's location, previous searches, etc.
- No real-time feedback
  - API provides detailed feedback, but it's not immediate
- Limited control
  - Experimentation is limited to single searches, where results are interleaved with those of the production system; no control over the entire result list
- Ultimate measure of success
  - Search is only a means to an end, it is not the ultimate goal

http://trec-open-search.org/

Slides: Courtesy of Krisztian Balog

University of Glasgow

# NTCIR-13: Open Live Test for Question Retrieval

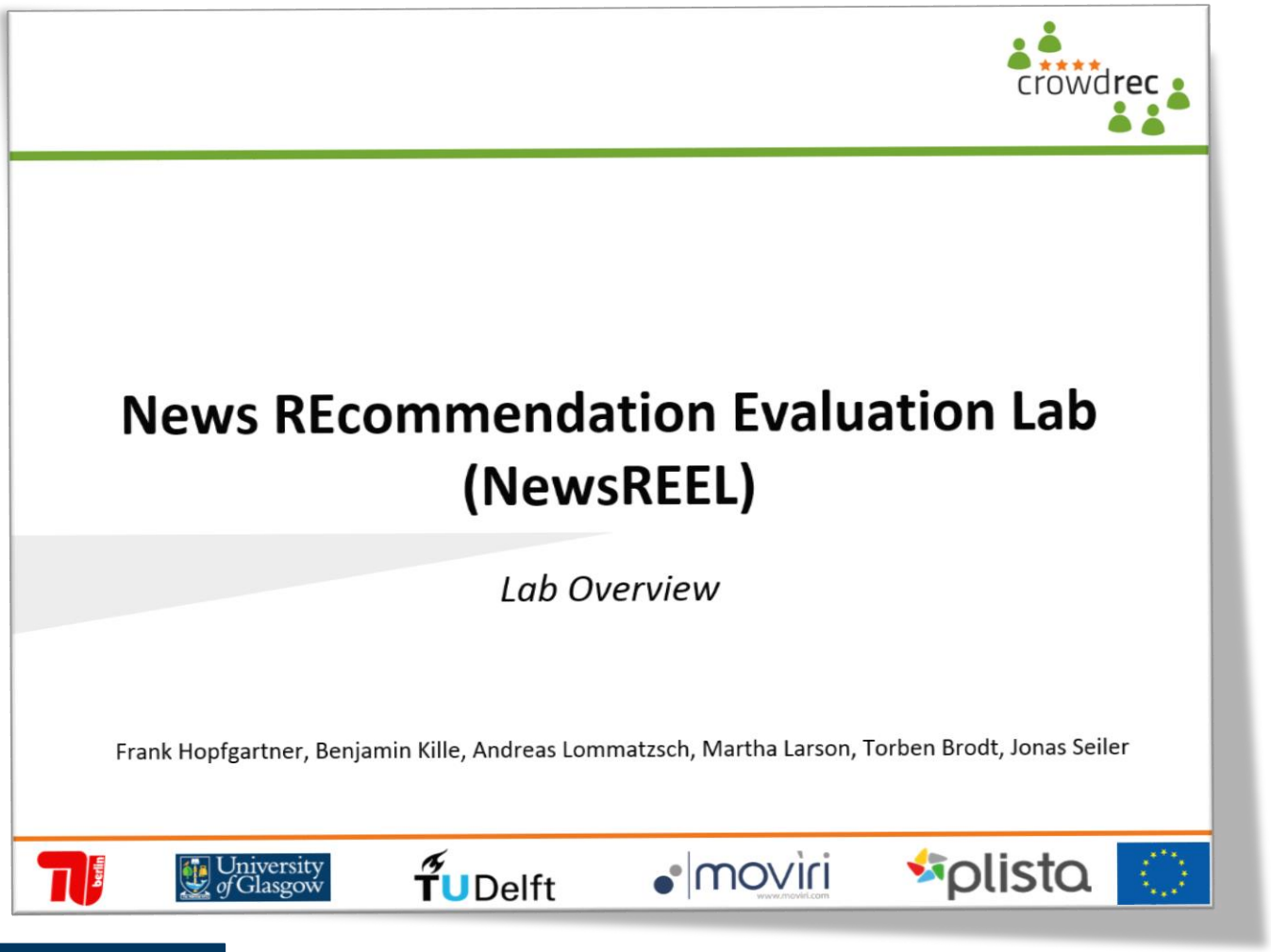



http://www.openliveq.net/

Slides: Courtesy of Makoto P. Kato

University of Glasgow

# News Recommendation Evaluation Lab (CLEF NewsREEL)




http://clef-newsreel.org/

University of Glasgow

## Overview

- Introduction
- Living Labs
- **News Recommendation Evaluation Lab (NewsREEL) Scenario**
- NewsREEL 2016
- NewsREEL and You

University of Glasgow

## CLEF NewsREEL

In CLEF NewsREEL, participants can develop stream-based **news recommendation** algorithms and have them benchmarked (a) online by **millions of users** over the period of a few months, and (b) offline by **simulating a live stream**.

F. Hopfgartner, T. Brodt, J. Seiler, B. Kille, A. Lommatzsch, M. Larson, R. Turrin, A. Sereny.
"*Benchmarking News Recommendations: The CLEF NewsREEL Use Case*", SIGIR Forum, 49(2):57-65, 2015.

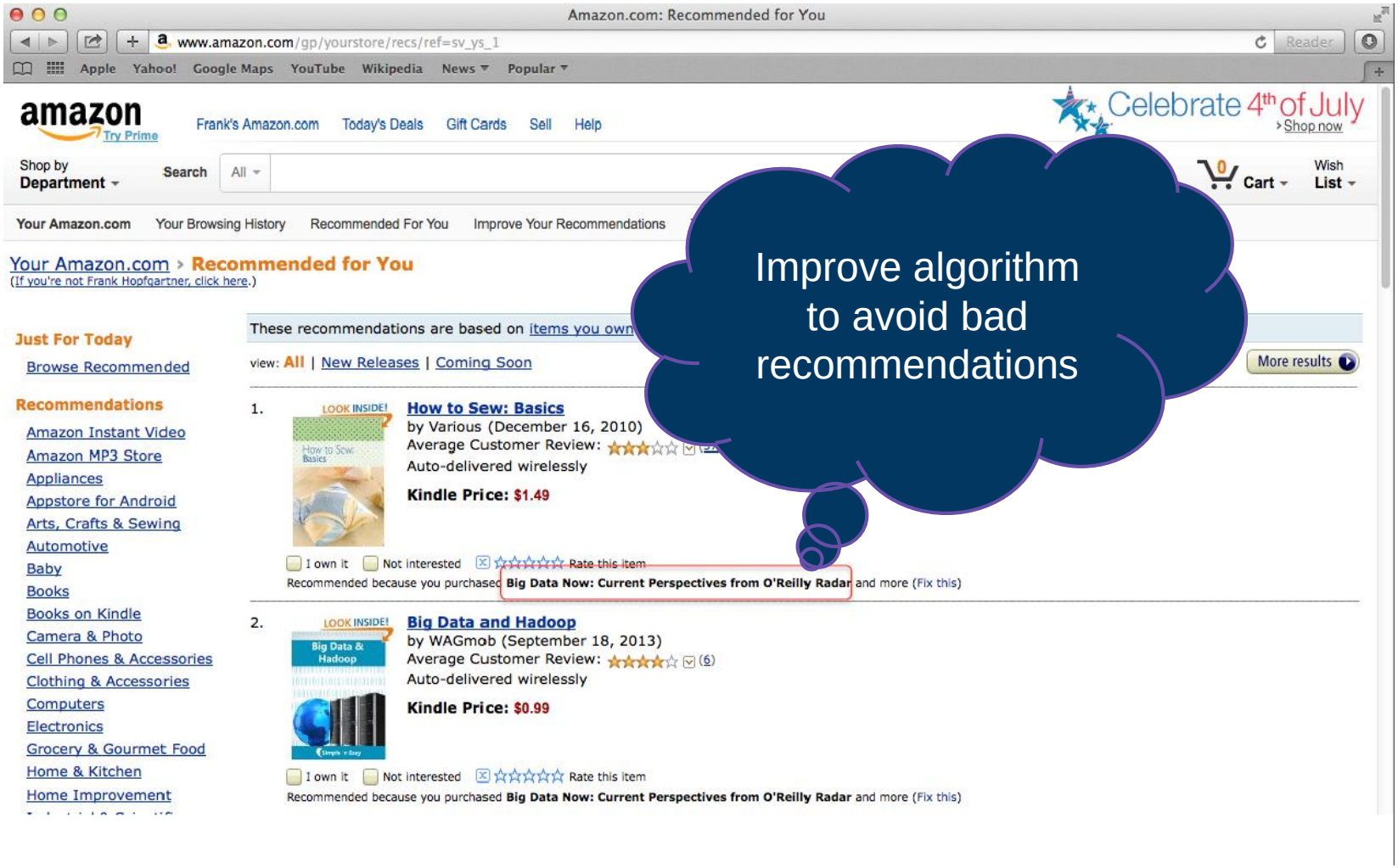
University of Glasgow

## What are recommender systems?

Recommender systems help users to find **items** they were not searching for.

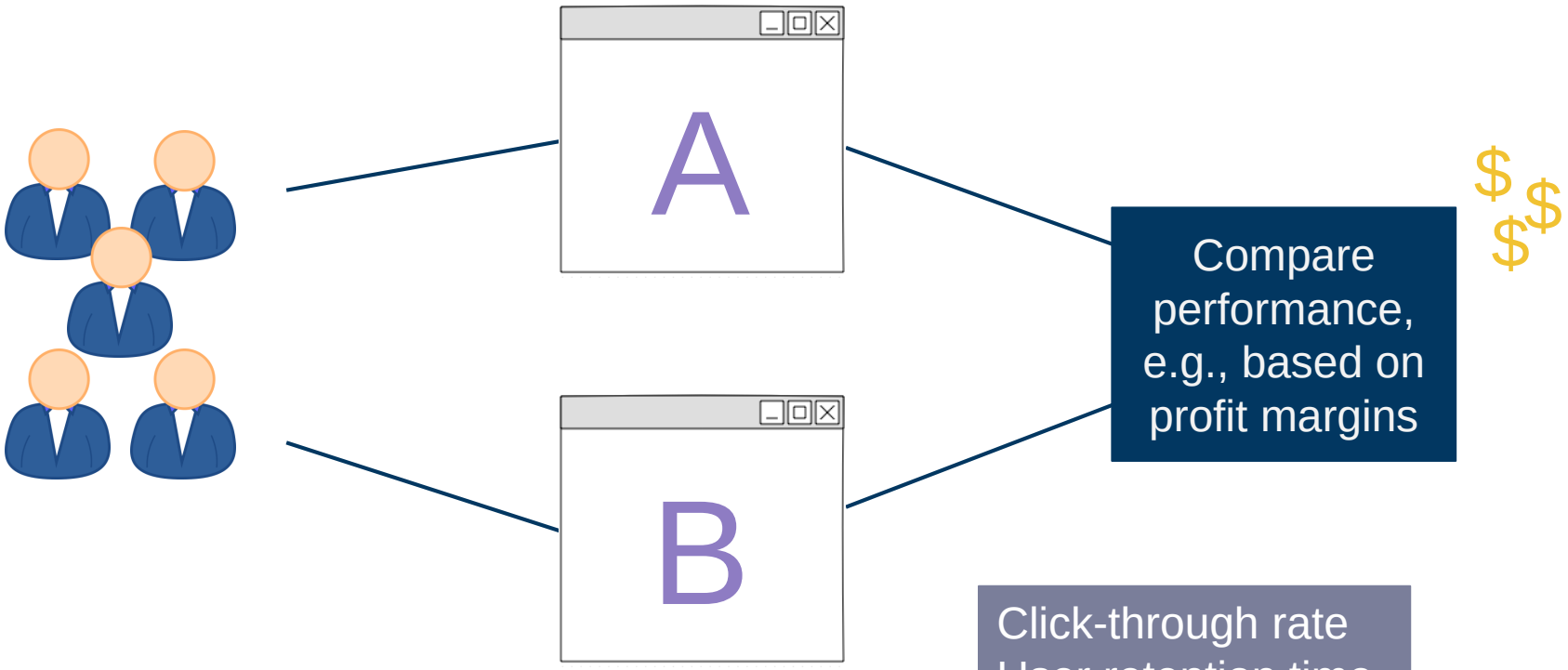
University of Glasgow

# Example: Online Retailer



University of Glasgow

## Online Evaluation (A/B testing)



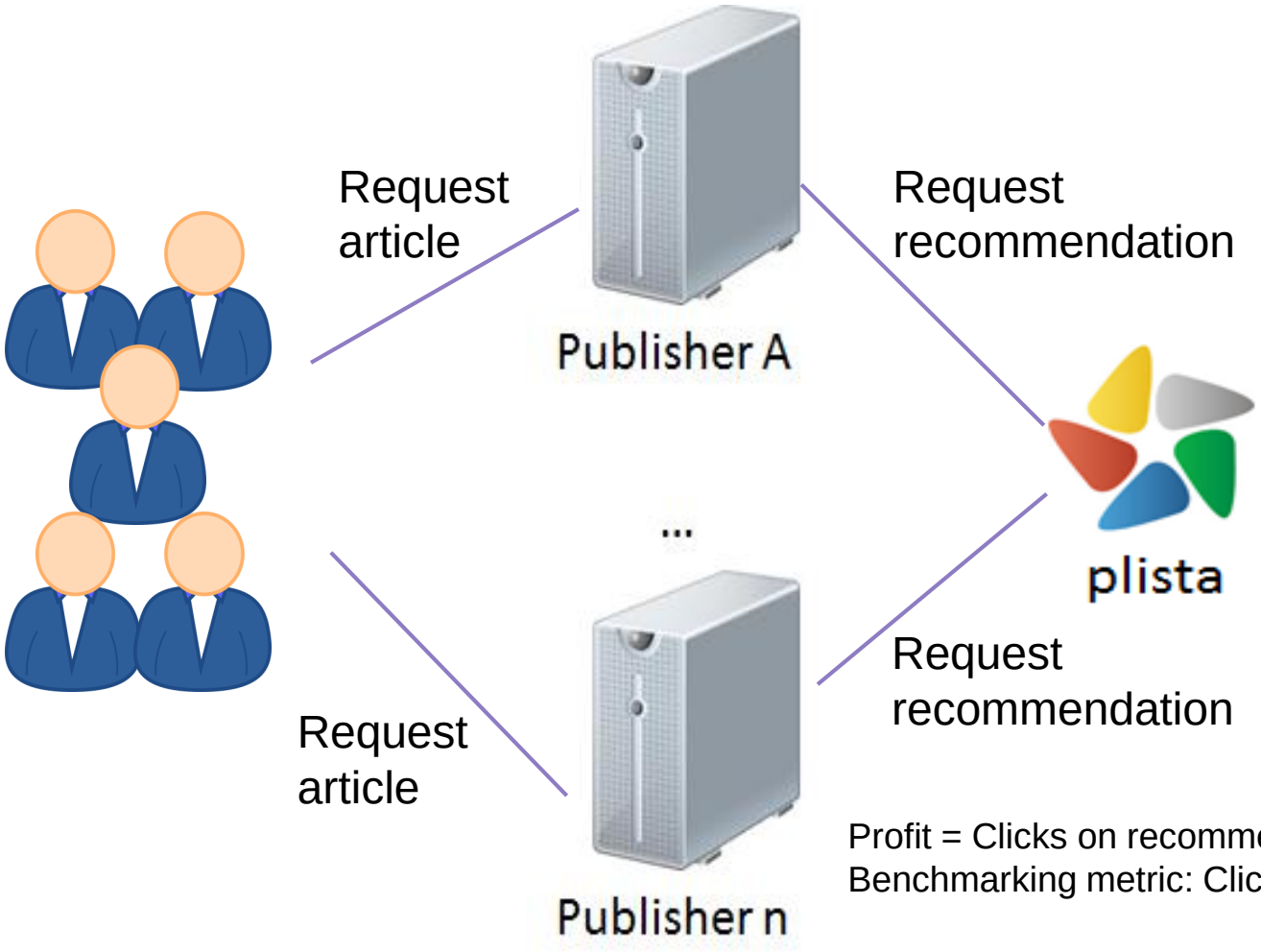
University of Glasgow

# NewsREEL scenario


Image: Courtesy of T. Brodt (plista)

University of Glasgow

## NewsREEL scenario

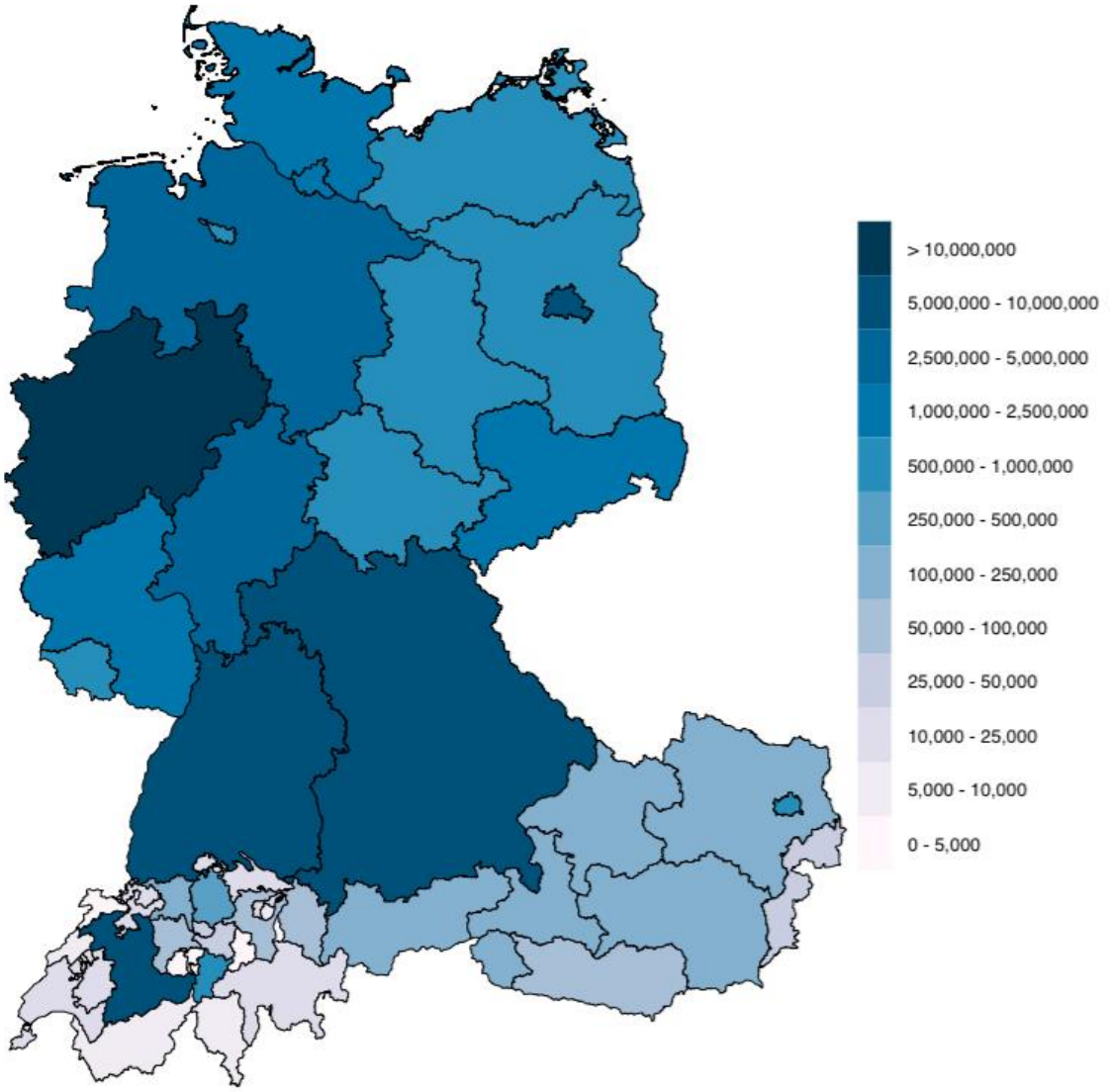



Profit = Clicks on recommendations
Benchmarking metric: Click-Through-Rate

University of Glasgow

## Who are the users?

>80%



University of Glasgow

## Task 1: Online Evaluation

# Recommend news articles in real-time

**Data**

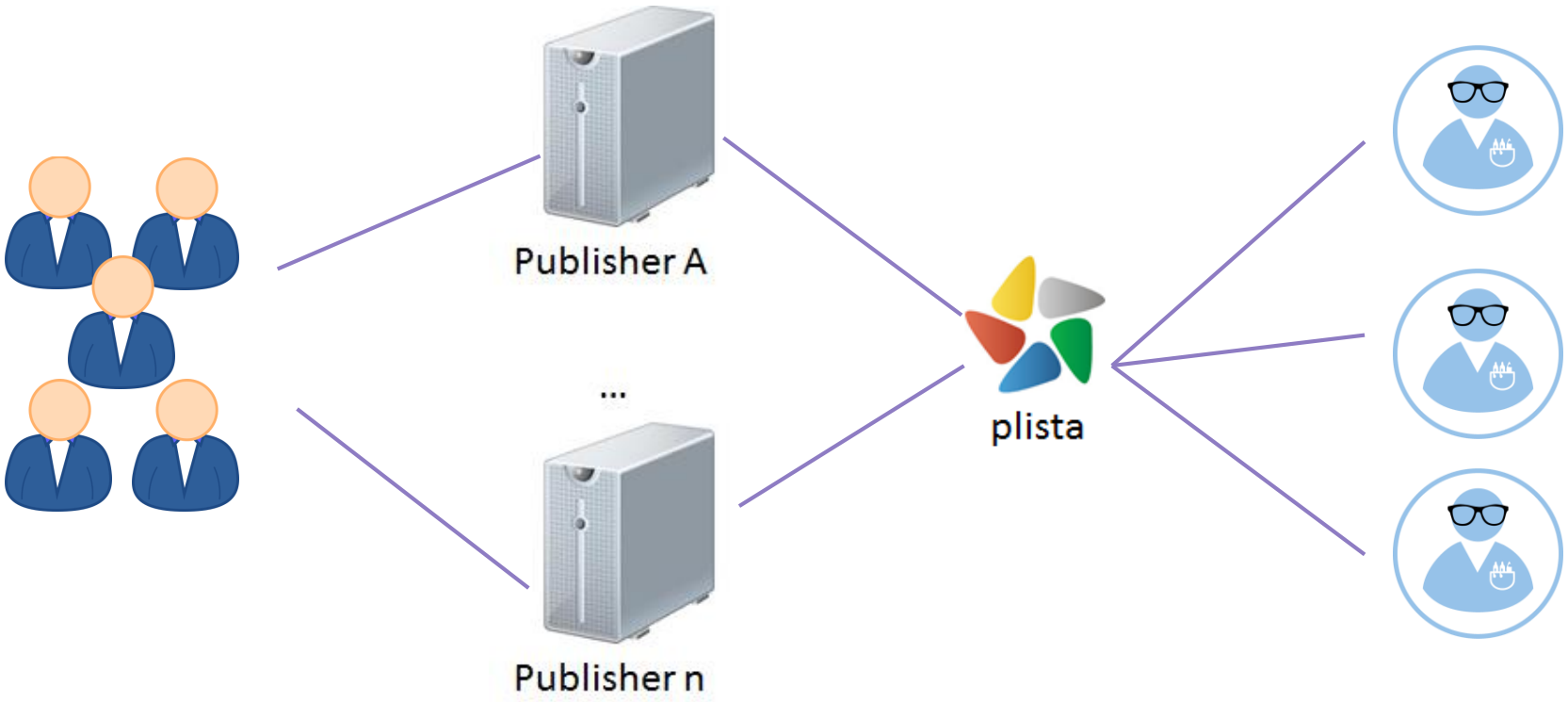
- Provide recommendations for visitors of the news portals of plista's customers
- Ten portals (local news, sports, business, technology)
- Communication via Open Recommendation Platform (ORP)

**Evaluation**

- Benchmark own performance with other participants and baseline algorithms during three pre-defined evaluation windows
- Best algorithms determined in final evaluation period
- Standard evaluation metrics

University of Glasgow

# Real-Time Recommendation




T. Brodt and F. Hopfgartner
"*Shedding Light on a Living Lab: The CLEF NewsREEL Open Recommendation Platform*," In *Proc. of IIiX 2014, Regensburg, Germany*, pp. 223-226, 2014.

University of Glasgow

# NewsREEL

## CLEF-NEWSREEL

### NEWS RECOMMENDATION EVALUATION LAB

HOME | TASKS | HOW TO PARTICIPATE | PUBLICATIONS | ORGANISATION | PREVIOUS CAMPAIGNS | CLEF 2017 | PROGRAM

### Overview

Commercial providers of information access systems (such as Amazon or Google) usually evaluate the performance of their algorithms by observing how large numbers of their customers interact with different instances of their services. Unfortunately, due to the lack of access to large-scale systems, university-based research is struggling to catch up with this large-scale evaluation methodology. NewsREEL, short for News Recommendation Evaluation Lab, aims to bridge this "evaluation gap" between Academia and Industry.

NewsREEL is organised as a campaign-style evaluation lab of CLEF 2017 and addresses the following information access task:

**Whenever a visitor of an online news portal reads a news article on their side, the task is to recommend other news articles that the user might be interested in.**

NewsREEL offers two tasks to study this use case. The first task, **NewsREEL Live**, implements the idea of a "living lab" where the provider of a recommendation service provides access to its infrastructure and user base. The second task, **NewsREEL Replay**, replays a live setting using the recommender system reference framework **Idomaar**.

By providing this service for millions of users, the recommendation scenario requires solutions to significant research challenges, such as processing information in real-time, handling vast amount of data, and providing suitable


### Important Dates

- Labs registration opens: 4 November 2016
- Registration Closes: 21 April 2017
- Test Period 1: TBA
- Test Period 2: TBA
- Test Period 3: TBA
- Evaluation Period: TBA
- End Evaluation Cycle: 5 May 2017
- Submission of Working Notes: 26 May 2017
- Feedback on Working Notes: 16 June 2017
- Camera Ready Working Notes due: 3 July 2017
- CLEF 2017: 11-14 September 2017

### @clefnewsreel

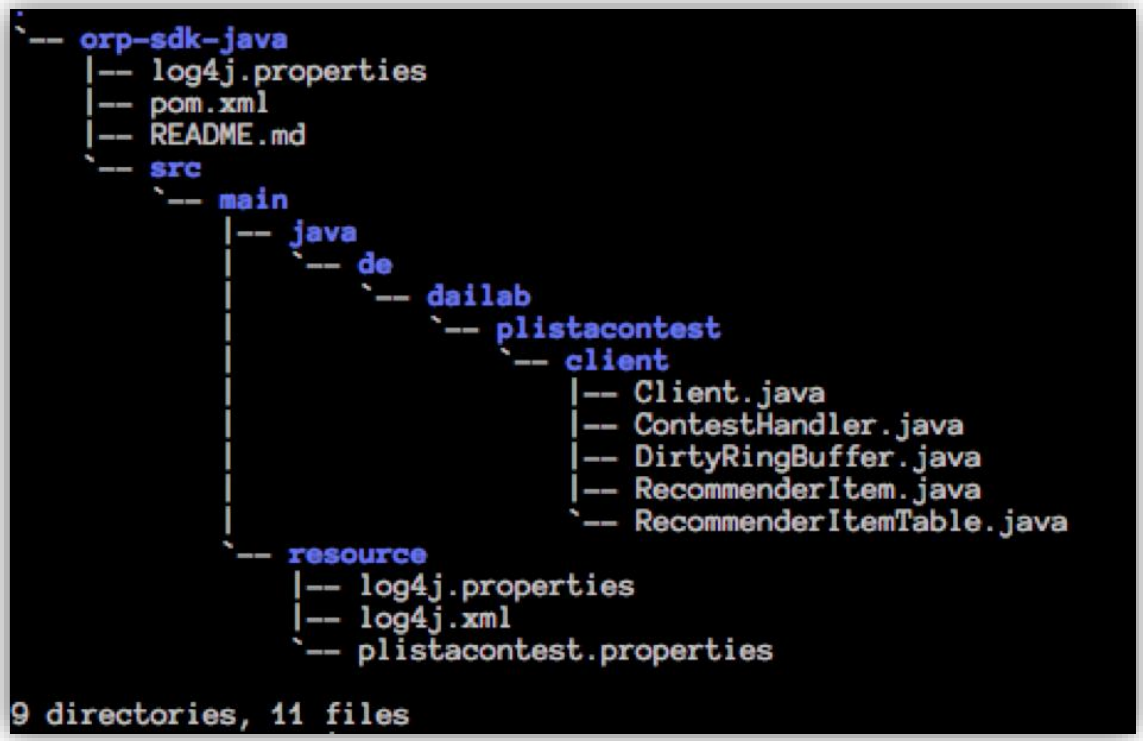
- The poster is on "The Potentials of Recommender Systems Challenges for Student Learning"

http://clef-newsreel.org/

University of Glasgow

## Getting Started: Set up a server

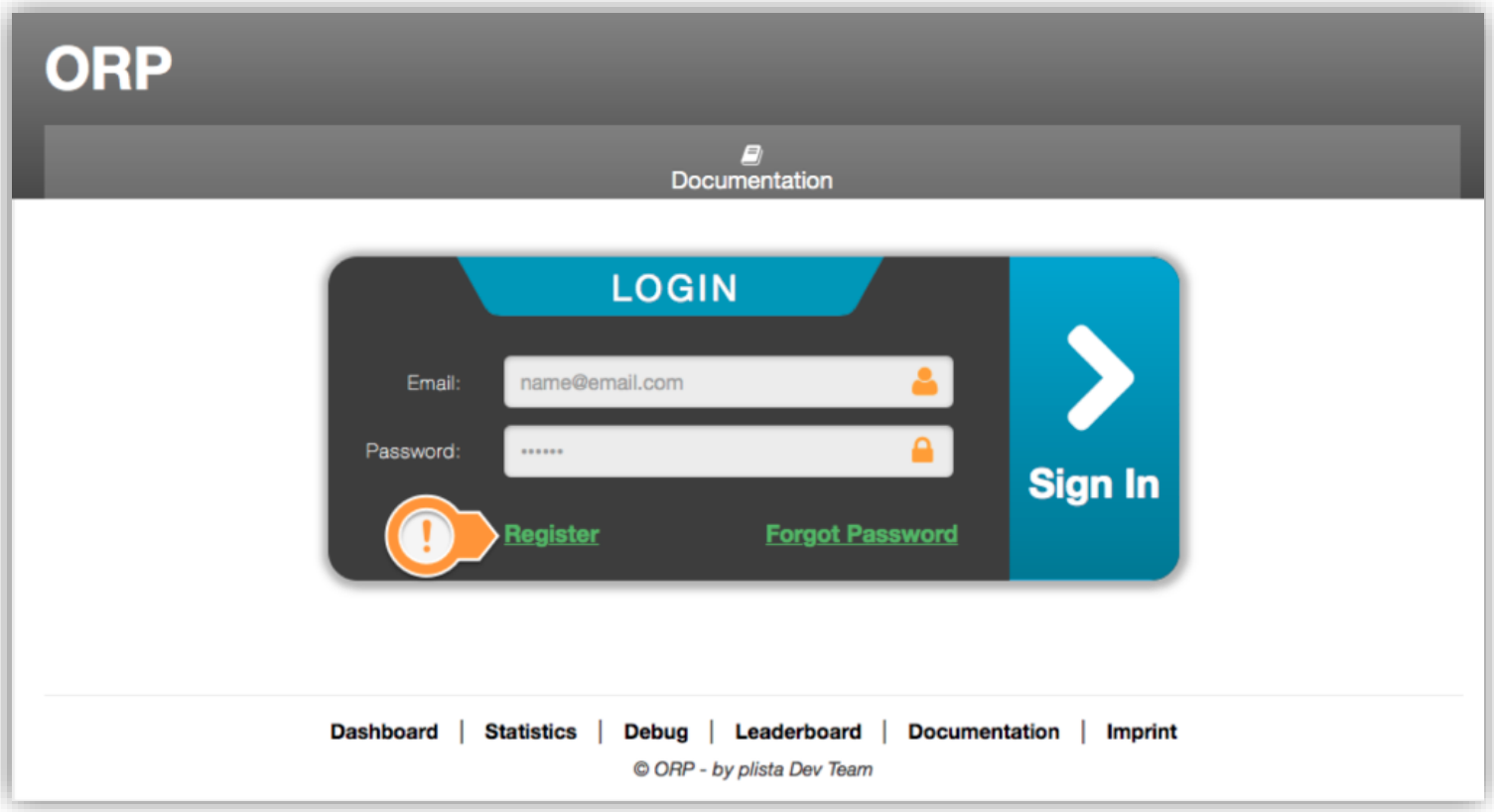
- There are different implementations which handle ORP
  - PHP
  - Java
  - Node.js




http://www.clef-newsreel.org/tutorials/task1/

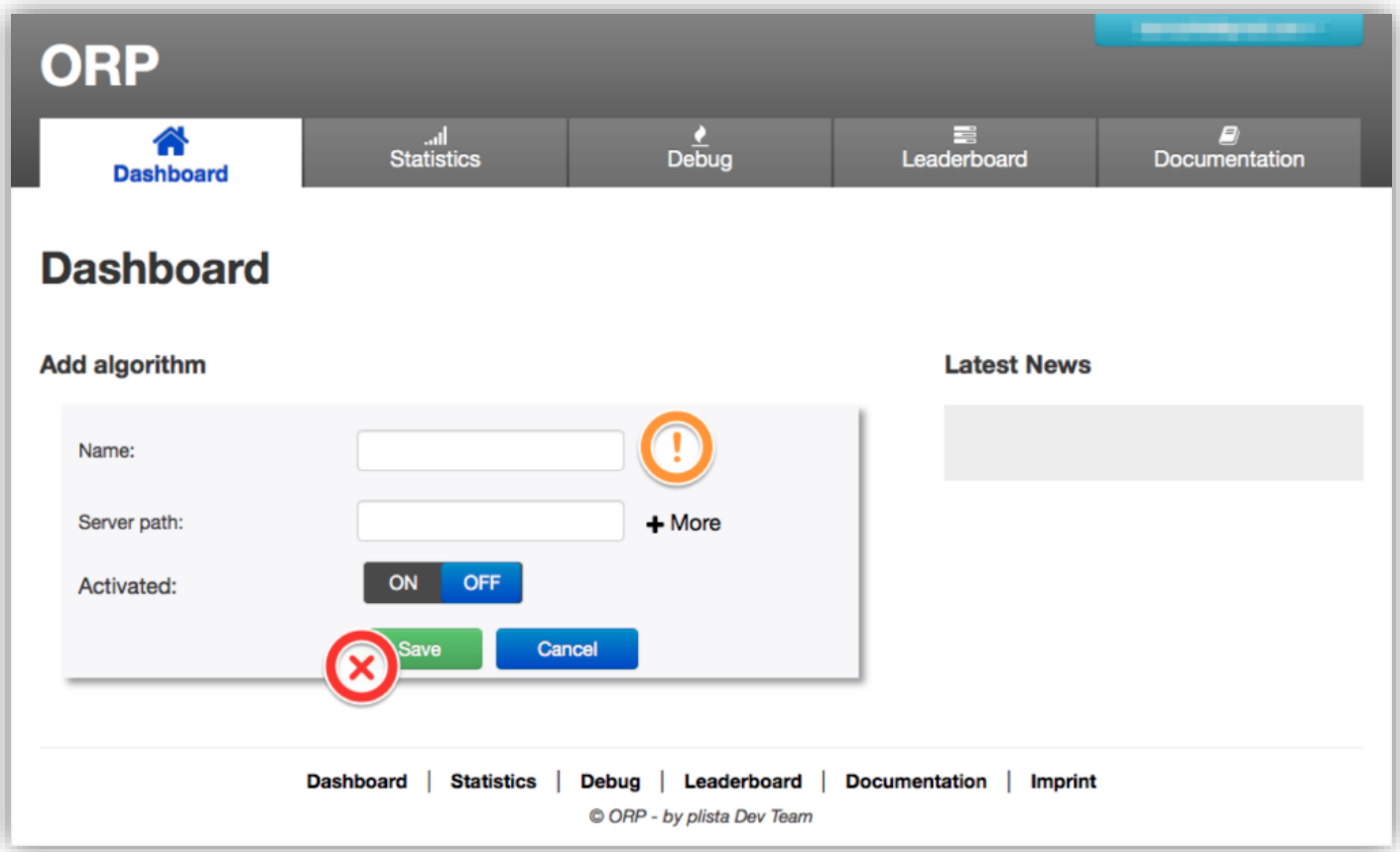
University of Glasgow

# Getting Started: Register (AFTER registering via CLEF)

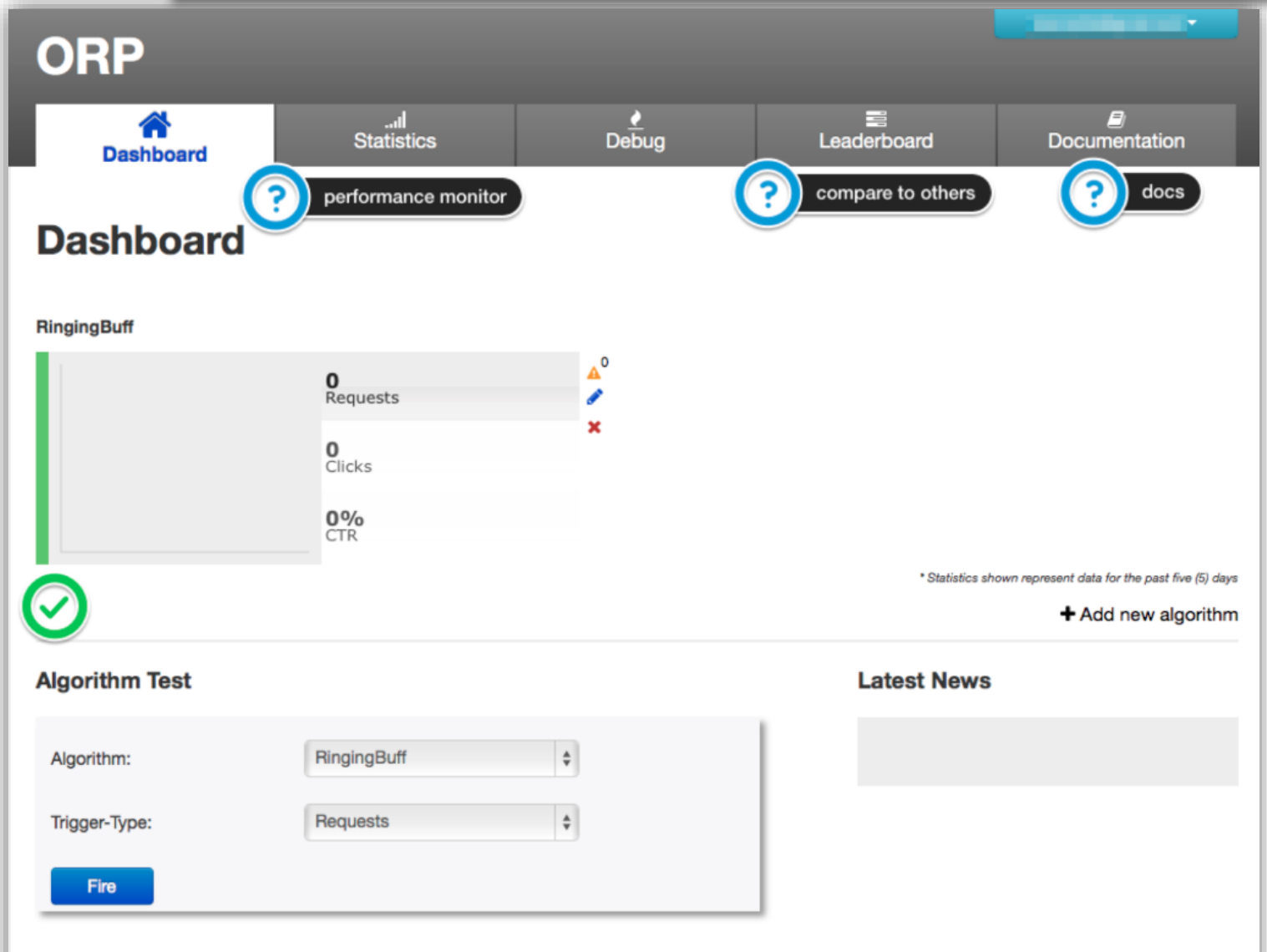


University of Glasgow

# Getting Started: Log into Open Recommender Platform

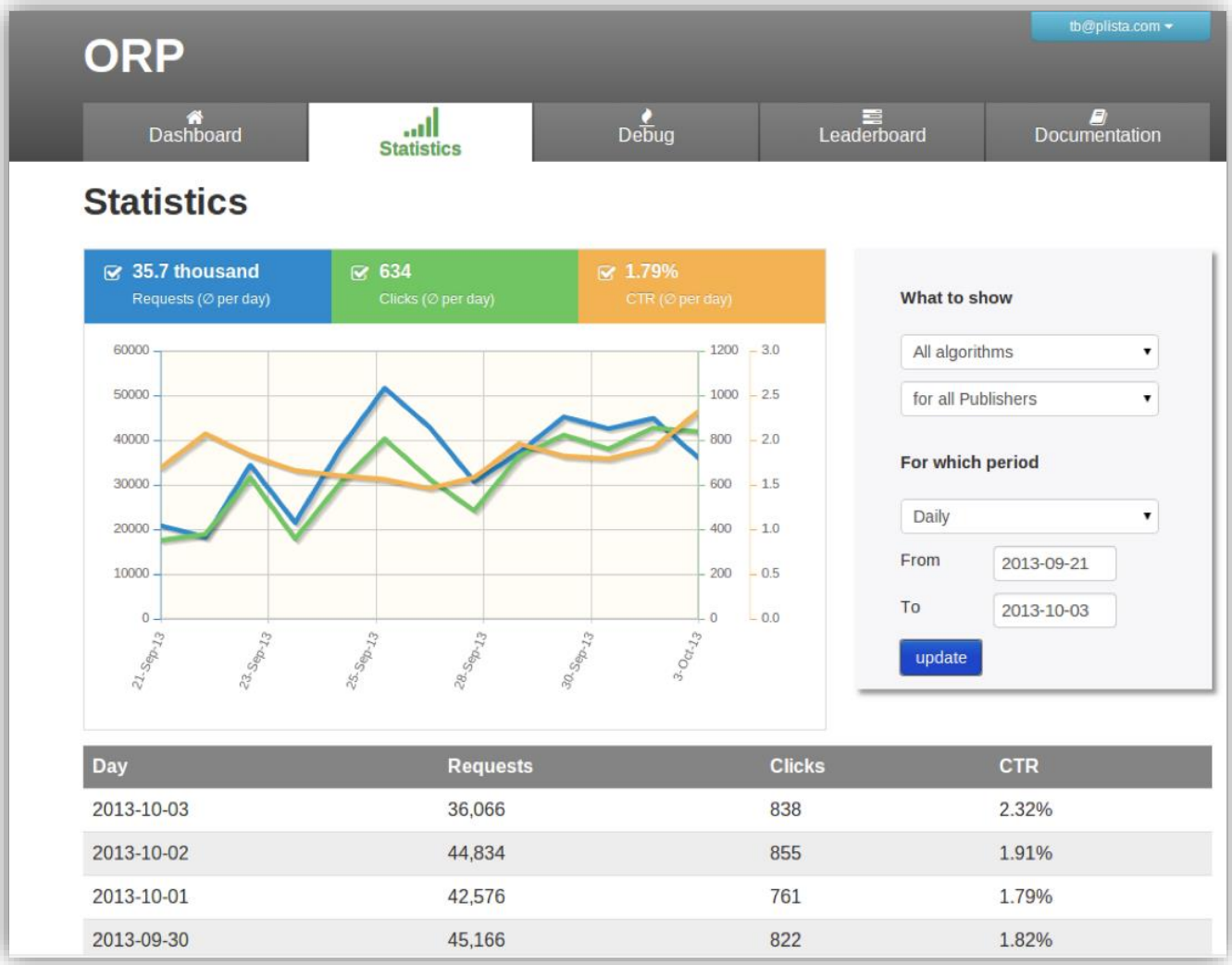


University of Glasgow



University of Glasgow

# Getting Started: Monitor performance




| Day | Requests | Clicks | CTR |
|---|---|---|---|
| 2013-10-03 | 36,066 | 838 | 2.32% |
| 2013-10-02 | 44,834 | 855 | 1.91% |
| 2013-10-01 | 42,576 | 761 | 1.79% |
| 2013-09-30 | 45,166 | 822 | 1.82% |

University of Glasgow

## Overview

- Introduction
- Living Labs
- News Recommendation Evaluation Lab (NewsREEL) Scenario
- **NewsREEL 2016**
- NewsREEL and You

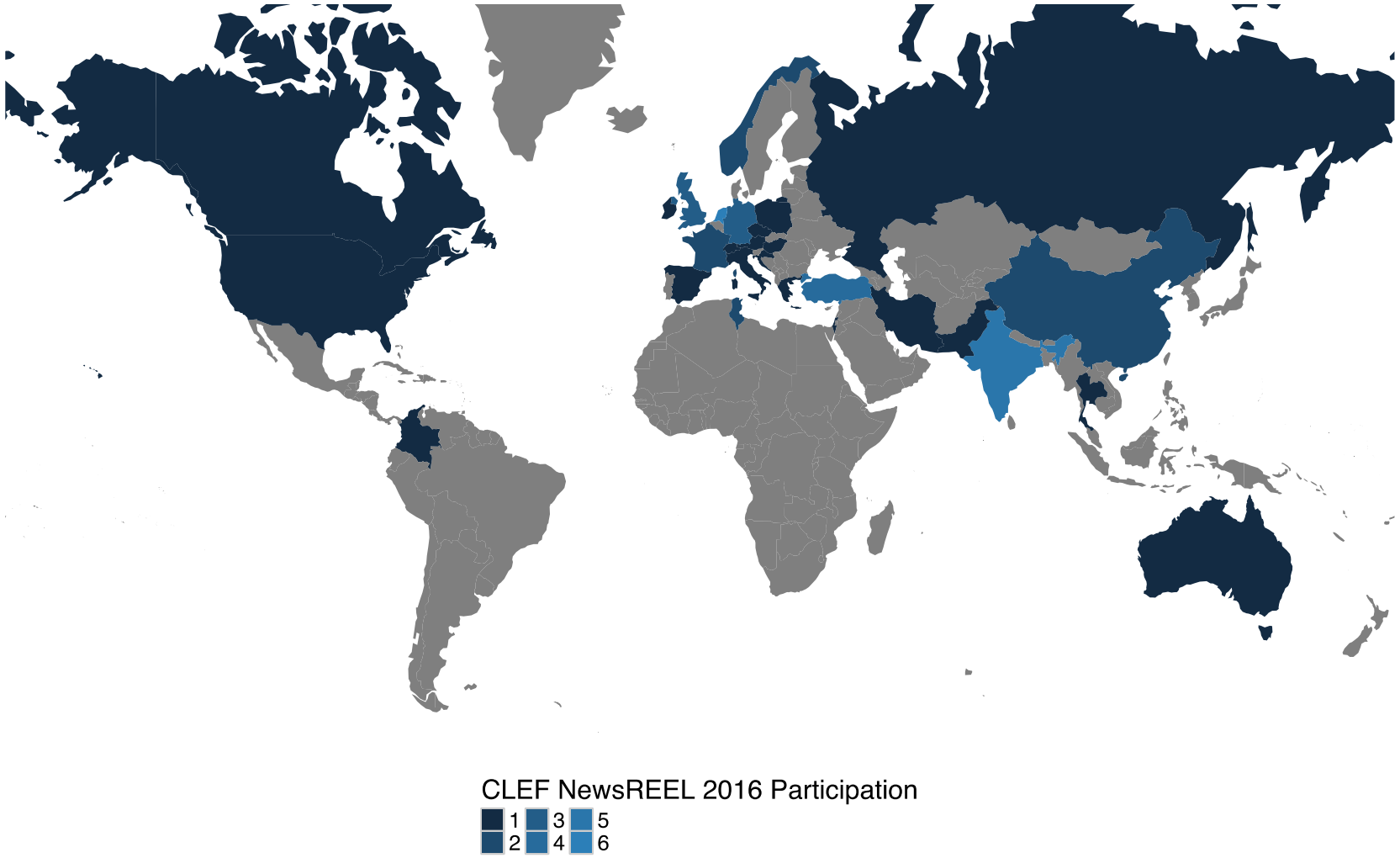
University of Glasgow

## Promotion activities

- Call for participations
- Social media and website
- Tutorials at ACM RecSys'15, ECIR'15, workshop presentations
- Invited talks at workshops, institutes, meetup groups
- Demo at ACM RecSys'16
- …

University of Glasgow

# Registrations for NewsREEL 2016



University of Glasgow

# Evaluation Schedule (Task 1)

## February 2016

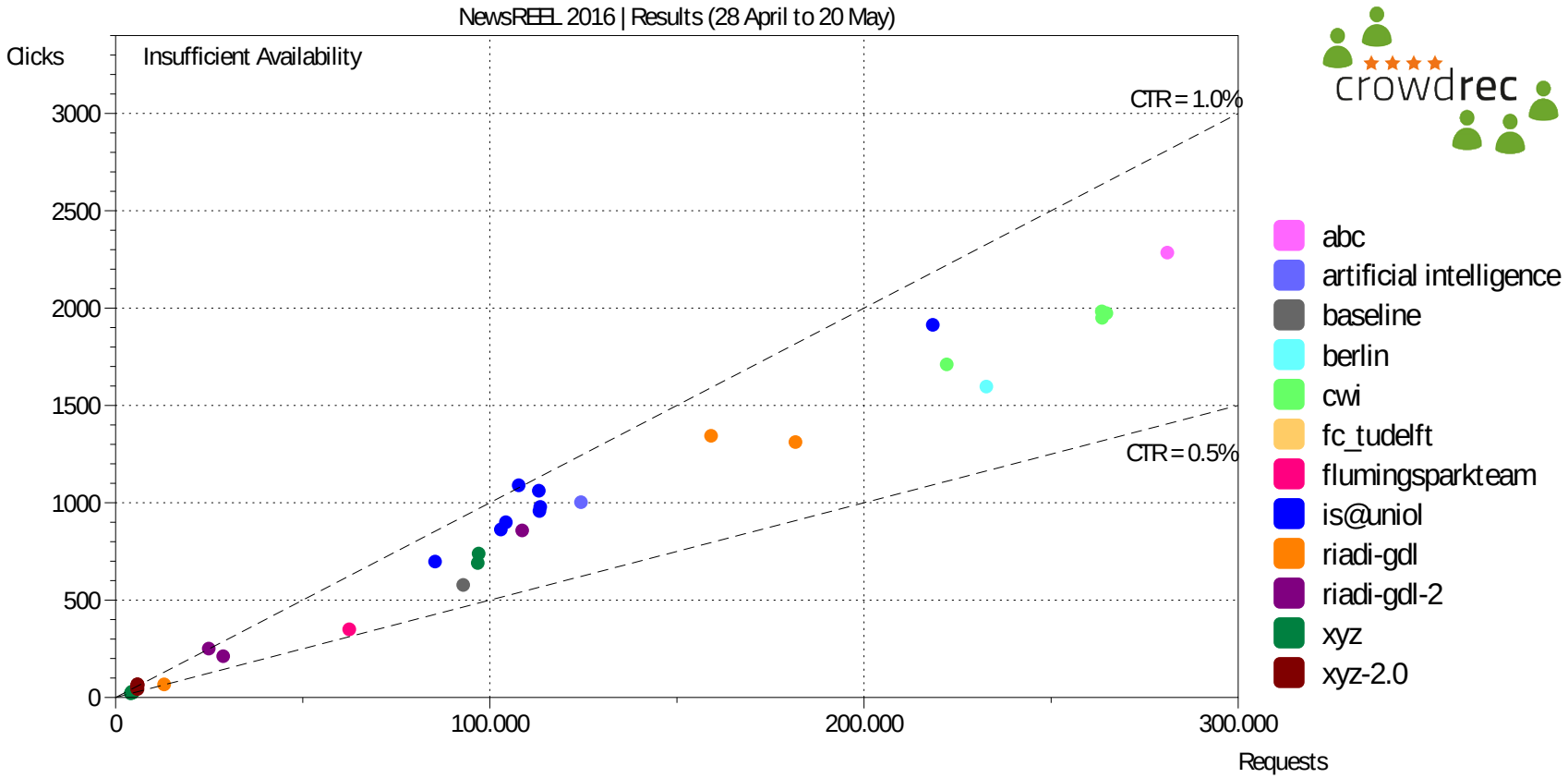
| Sunday | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| 28 | 29 | | | | | |

## March 2016

| Sunday | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| 20 | 21 | 22 | 23 | 24 | 25 | 26 |
| 27 | 28 | 29 | 30 | 31 | | |

University of Glasgow

# CTR (Task 1)




crowdrec

University of Glasgow

# NewsREEL Session at CLEF 2016 in Evora


Winning algorithms were developed by students…

University of Glasgow

## Overview

- Introduction
- Living Labs
- News Recommendation Evaluation Lab (NewsREEL) Scenario
- NewsREEL 2016
- **NewsREEL and You**

University of Glasgow

# NewsREEL at CLEF 2017

## NewsREEL Live

- Provide real-time recommendations for visitors of news portals
- Benchmark own performance with other participants and baseline algorithms during **three pre-defined** evaluation windows
- Best algorithms determined in **final evaluation period**.

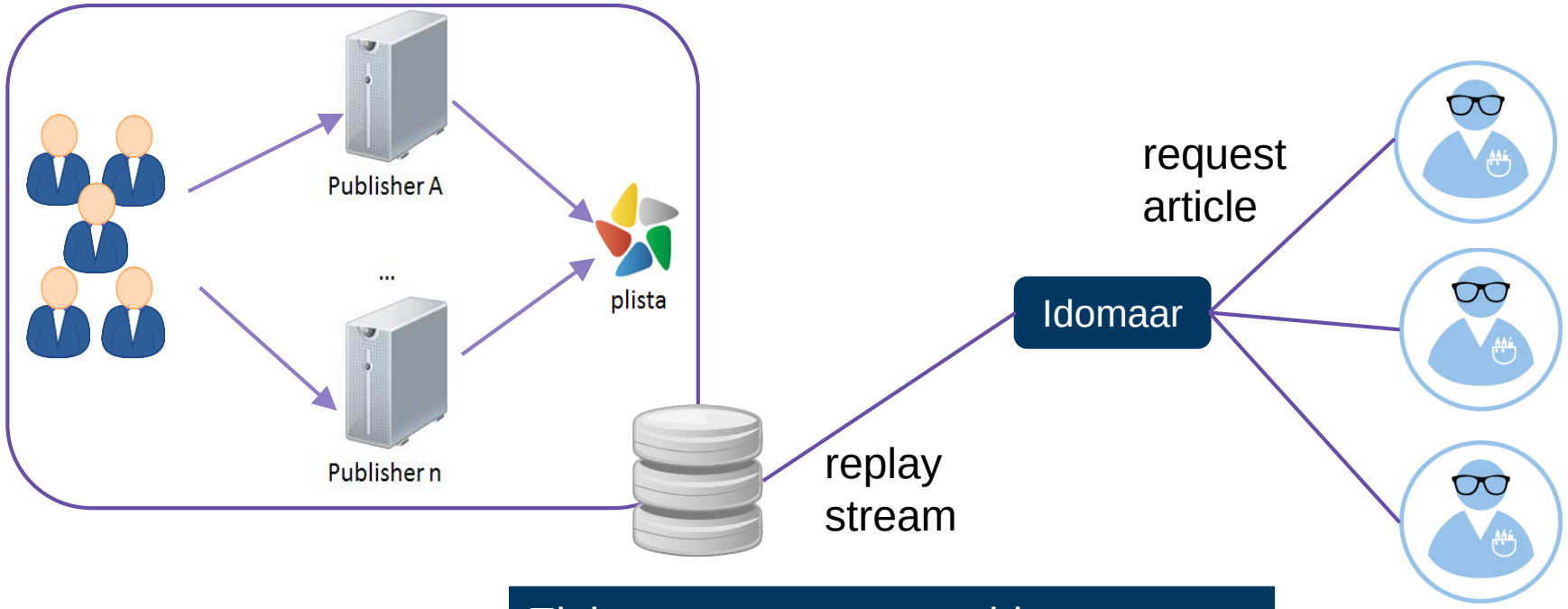
## NewsREEL Replay

- Traffic and content updates of nine **news content provider** websites
- **Simulation** of data stream using Idomaar framework
- Participants have to **predict interactions** with data stream
- Metric: ratio of successful predictions by the total number of predictions

http://clef2017.clef-initiative.eu/

University of Glasgow

# NewsREEL Replay




Eight news content providers
Two million notifications
58,000 item updates
168 million recommendation requests

University of Glasgow

## NewsREEL for Learning & Teaching

"NewsREEL provided an excellent opportunity for my students to gain experience in using web services, processing streams, and developing information access algorithms."

(Senior Academic)

University of Glasgow

## Potentials of NewsREEL for Student Learning

### Learning-by-doing

- Practical lab assignments are part of STEM curriculum to support self-centred student learning and student engagement.

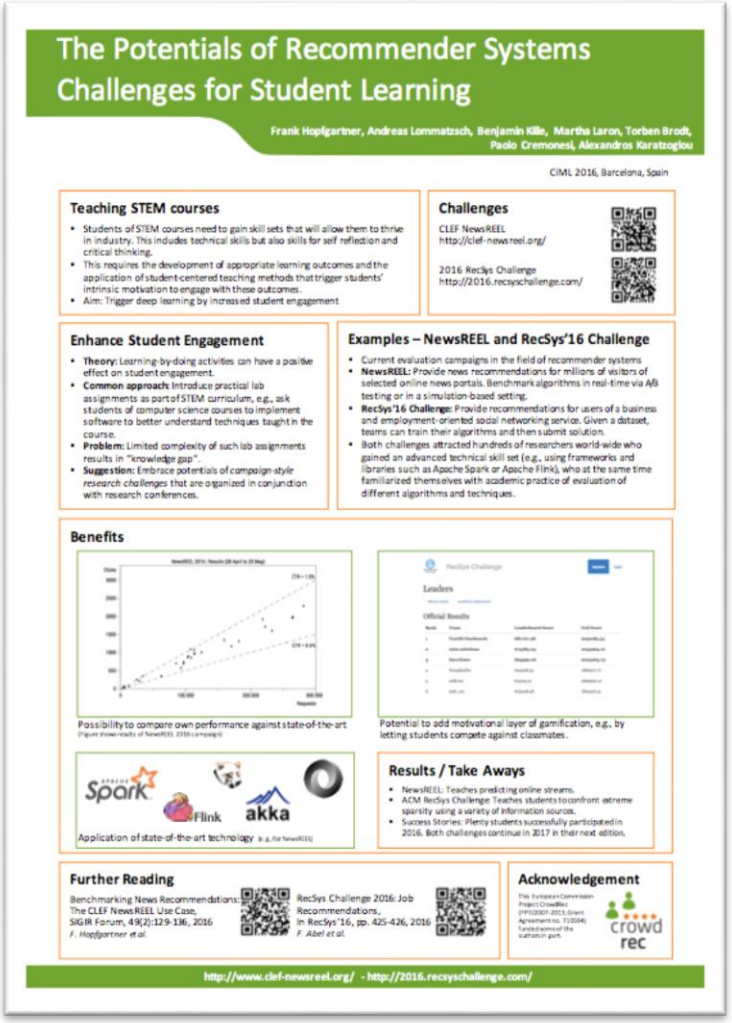
### Limitation

- Limited complexity of curricular assignments results in "knowledge gap" between skillsets taught at HEIs and required by industry.

**Embed NewsREEL in your teaching**

University of Glasgow

## After participating in NewsREEL, students are able to…



The Potentials of Recommender Systems Challenges for Student Learning

Frank Hopfgartner, Andreas Lommatzsch, Benjamin Kille, Martha Larson, Torben Brodt, Paolo Cremonesi, Alexandros Karatzoglou

CIML 2016, Barcelona, Spain

**Teaching STEM courses**

- Students of STEM courses need to gain skill sets that will allow them to thrive in industry. This includes technical skills but also skills for self reflection and critical thinking.
- This requires the development of appropriate learning outcomes and the application of student-centered teaching methods that trigger students' intrinsic motivation to engage with these outcomes.
- Aim: Trigger deep learning by increased student engagement

**Challenges**

CLEF NewsREEL
http://clef-newsreel.org/

2016 RecSys Challenge
http://2016.recsyschallenge.com/

**Enhance Student Engagement**

- **Theory:** Learning-by-doing activities can have a positive effect on student engagement.
- **Common approach:** Introduce practical lab assignments as part of STEM curriculum, e.g., ask students of computer science courses to implement software to better understand techniques taught in the course.
- **Problem:** Limited complexity of such lab assignments results in "knowledge gap".
- **Suggestion:** Embrace potentials of *campaign-style research challenges* that are organized in conjunction with research conferences.

**Examples – NewsREEL and RecSys'16 Challenge**

- Current evaluation campaigns in the field of recommender systems
- **NewsREEL:** Provide news recommendations for millions of visitors of selected online news portals. Benchmark algorithms in real-time via A/B testing or in a simulation-based setting.
- **RecSys'16 Challenge:** Provide recommendations for users of a business and employment-oriented social networking service. Given a dataset, teams can train their algorithms and then submit solution.
- Both challenges attracted hundreds of researchers world-wide who gained an advanced technical skill set (e.g., using frameworks and libraries such as Apache Spark or Apache Flink), who at the same time familiarized themselves with academic practice of evaluation of different algorithms and techniques.

**Benefits**

Possibility to compare own performance against state-of-the-art

Potential to add motivational layer of gamification, e.g., by letting students compete against classmates.

Application of state-of-the-art technology

**Results / Take Aways**

- NewsREEL: Teaches predicting online streams.
- ACM RecSys Challenge Teaches students to confront extreme sparsity using a variety of information sources.
- Success Stories: Plenty students successfully participated in 2016. Both challenges continue in 2017 in their next edition.

**Further Reading**

Benchmarking News Recommendations: The CLEF NewsREEL Use Case, SIGIR Forum, 49(2):129-136, 2016
F. Hopfgartner et al.

RecSys Challenge 2016: Job Recommendations, In RecSys'16, pp. 425-426, 2016
F. Abel et al.

**Acknowledgement**

http://www.clef-newsreel.org/ - http://2016.recsyschallenge.com/

- Demonstrate technical skills (e.g., how to handle data streams, recommendation techniques, use of Web protocols) needed to thrive in industry.
- Value the importance of critical evaluation and self reflection.
- Develop their own ideas and present them to an Academic audience.

F. Hopfgartner, A. Lommatzsch, B. Kille, M. Larson, P. Cremonesi, A. Karatzoglou
"*The Potentials of Recommender Systems for Student Learning*," In *Proc. CiML'16, Barcelona, Spain,* 2016.

University of Glasgow

# Acknowledgements


## People involved in NewsREEL (current and past)

**CO-ORGANISERS**

- Torben Brodt
- Andreas Lommatzsch
- Benjamin Kille
- Jonas Seiler
- Tobias Heintz
- Martha Larson
- Roberto Turrin

**STEERING COMMITTEE**

- Pablo Castells
- Paolo Cremonesi
- Arjen de Vries
- Michael Ekstrand
- Hideo Hoho
- Joemon M. Jose
- Noriko Kando
- Udo Kruschwitz
- Mounia Lalmas
- Jimmy Lin
- Vivien Petras
- Till Plumbaum
- Domonkos Tikk

CLEF ASSOCIATION

University of Glasgow

## Thank you

# Questions?

Frank Hopfgartner
Frank.Hopfgartner@glasgow.ac.uk
@FTHopf
www.hopfgartner.co.uk