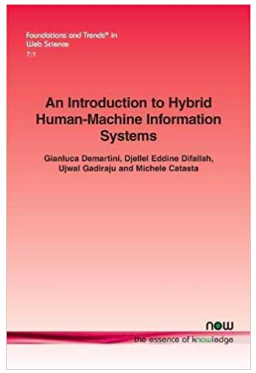
THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Knowledge Graph Quality Management

Gianluca Demartini

demartini@acm.org
@eglu81



www.gianlucademartini.net

Research Interests

- **Entity-centric Information Access** (2005-now)
  - Structured/Unstruct data (SIGIR 12), TRank (ISWC 13, WSemJ 16)
  - Entity Extraction (WWW 14), Gender Bias (SIGIR 18), Entity Cards (SIGIR 19)
  - IR Evaluation (IRJ 2015, ECIR 16 Best Paper, CIKM 17, SIGIR 18, CIKM 19)
- **Human-in-the-loop Information Systems** (2012-now)
  - Entity Linking (WWW 12, VLDBJ), CrowdQ (CIDR 13)
  - Remove noise (WWW 19)
  - Huml systems overview (COMNET 15, FnT 17)
- **Better Crowdsourcing Platforms** (2013-now)
  - Platform Dynamics (WWW 15), Wikidata (CSCWJ 18, ISWC 19)
  - Pick-a-Crowd (WWW 13), Scheduling Tasks (WWW 16)
  - Agreement (ICTIR 17, HCOMP 17), Pricing Tasks (HCOMP 14)
- **Human Factors in Crowdsourcing** (2015-now)
  - Malicious Workers (CHI 15), Attack Schemes (HCOMP 18 Best Paper)
  - Modus Operandi (UBICOMP 17), Power Workers (WSDM 20), Mood (HT 19)
  - Time (HCOMP 16), Complexity (HCOMP 16), Abandonment (WSDM 19)

Thanks to:

European Commission

Australian Government
Australian Research Council

EPSRC
Engineering and Physical Sciences Research Council

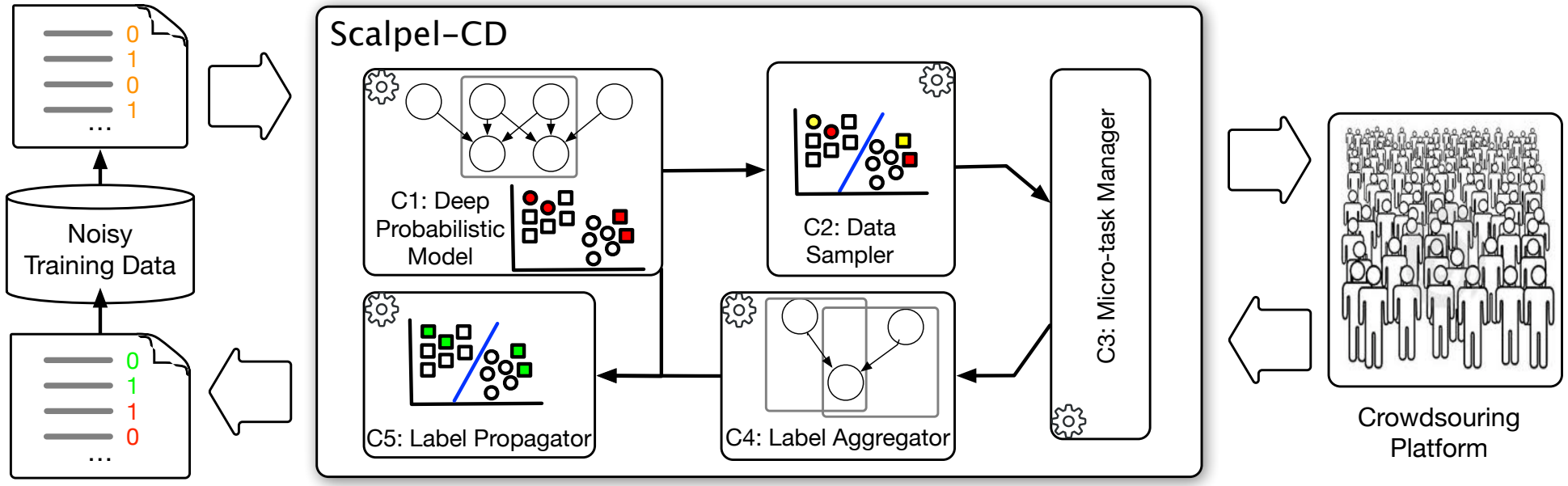
facebook research

Data Quality in Knowledge Graphs (KGs)

- Correctness
  - Values in the KG are correct
- Completeness
  - Schema level
    - All classes are present
  - Instance level
    - All instances for a class
- Inconsistencies (Bias)
  - Difference possible values for an attribute
- KGs for Biomedical Applications
  - Extraction of scientific concepts from scientific literature
  - Health Cards for Consumer Health Search

Correctness - Human-in-the-loop Data Curation

- **Noise detection** in training data
  - Deep learning (detection) + Crowdsourcing (fix) + Label propagation (save cost)



Jie Yang, Alisa Smirnova, Dingqi Yang, Gianluca Demartini, Yuan Lu, and Philippe Cudré-Mauroux. Scalpel-CD: Leveraging Crowdsourcing and Deep Probabilistic Modeling for Debugging Noisy Training Data. In: The Web Conference (**WWW 2019**). San Francisco, CA, USA, May 2019.

KG Class Completeness

- Estimating the cardinality of a class in a KG that evolves over time
  - Knowing the cardinality allows to estimate the completeness level (e.g., 80%)

Michael Luggen, Djellel Difallah, Cristina Sarasua, Gianluca Demartini, and Philippe Cudré-Mauroux. **Non-Parametric Class Completeness Estimators for Collaborative Knowledge Graphs**. In: The International Semantic Web Conference (ISWC 2019 - Research Track). Auckland, New Zealand, October 2019.

KG Evolution

- Wikidata graph and edit history as of Oct 2018

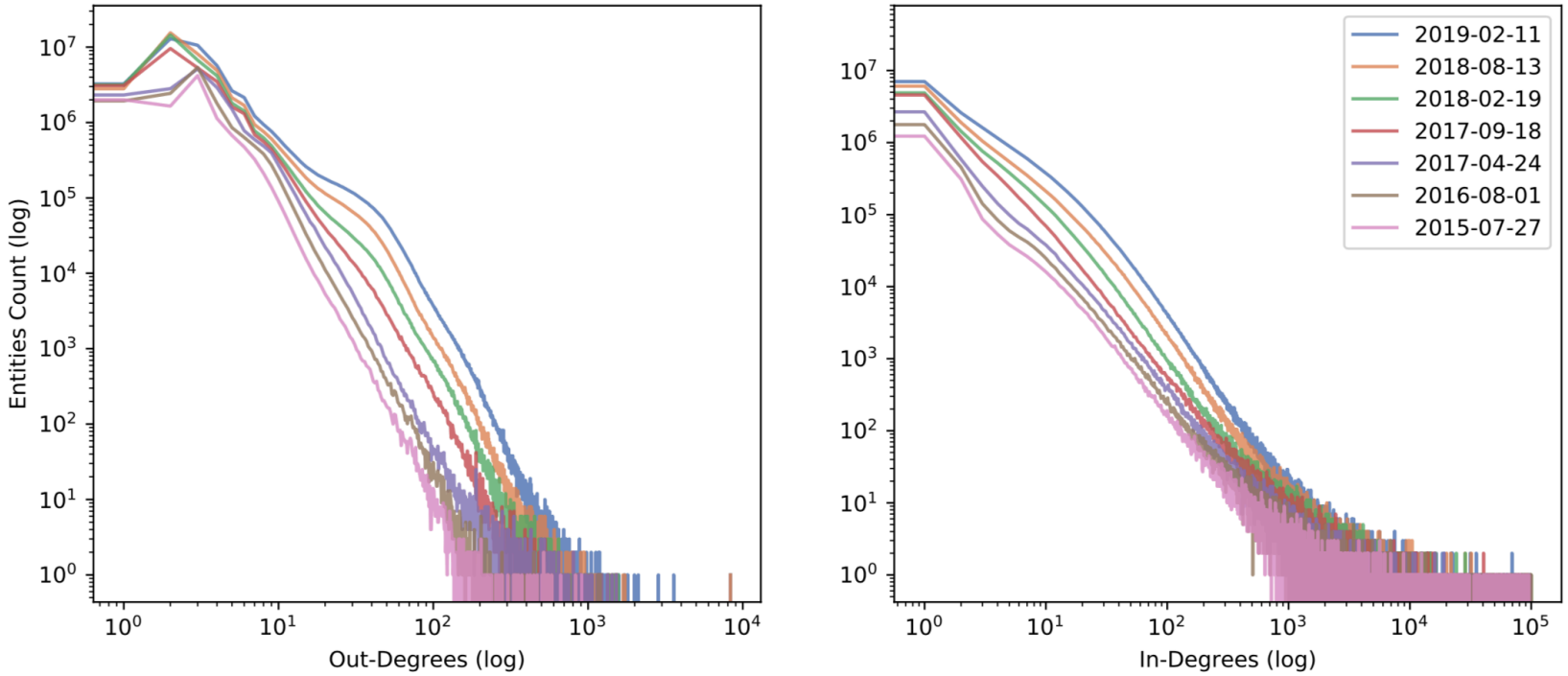


Fig. 2: The growth evolution of Wikidata: a temporal view on how the in and out-degree distributions have evolved since the inception of the project.

Class Completeness - Edits over Time

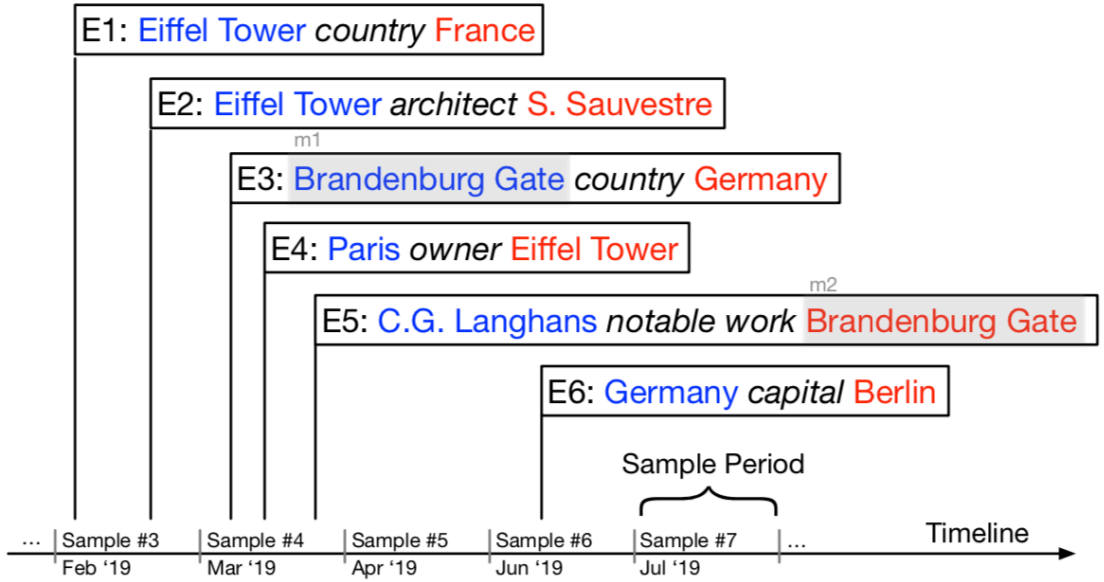
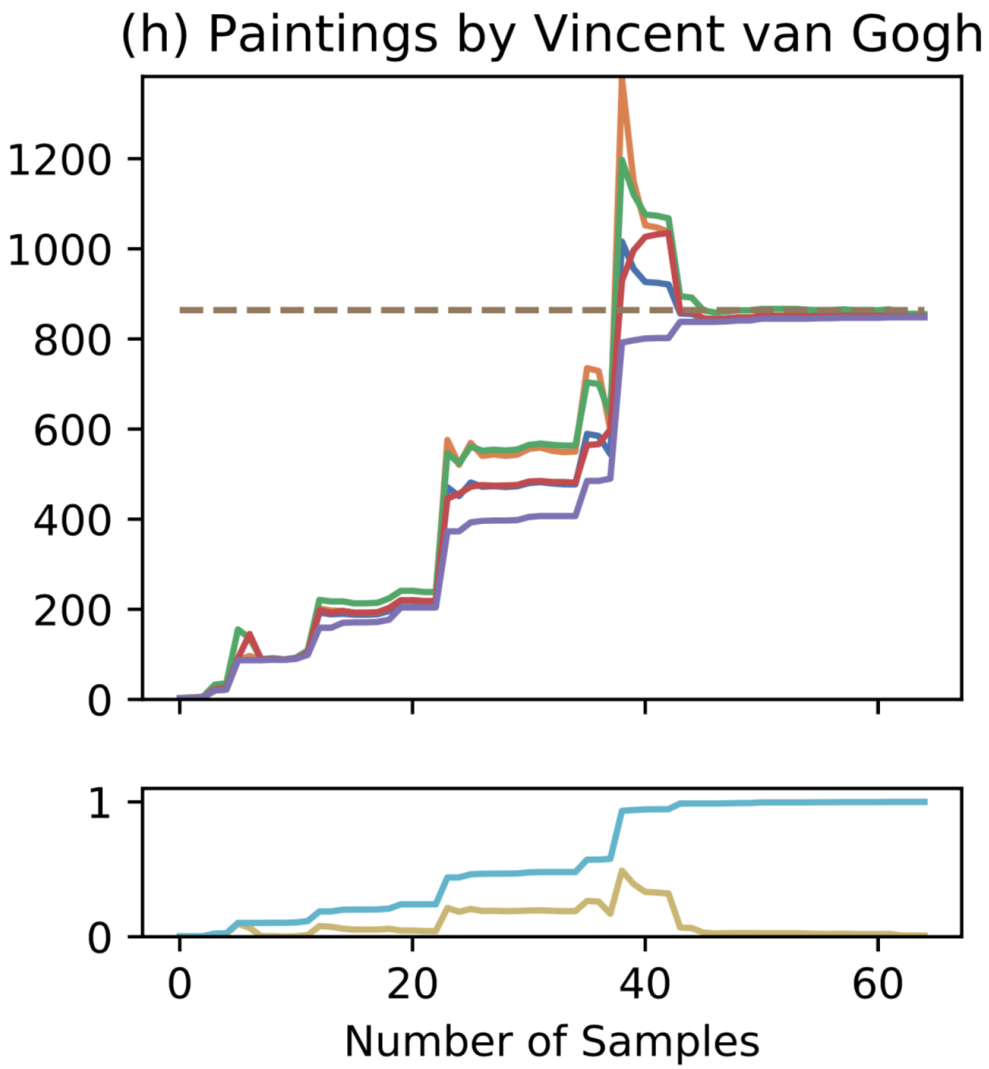


Fig. 1: The edits (E_i) of the Knowledge Graph (representing new edges) are leveraged to identify mentions. The source and target of each edge are collected to create a mention from the entity involved. Sample period #4 contains 3 *edits*, in which we identify 6 *mentions*, from which we extract 2 *observations* for class monument (despite the 3 mentions of entities of that class because m_1 and m_2 are only counted once), 1 observation for class country, 1 observation for class city and 1 observation for class person.

Estimators

- N1-UNIF
  - The ratio of the number of instances that have been observed so far
  - If frequency counts are unbalanced it will over/under-estimate
- Chao92
  - Capture/recapture with the concept of sample coverage
  - a high number of singletons might result in significant overestimation
- Jack1 (leave-one-out)
  - n−1 sample periods by removing one sample period from the data at a time
- SOR
  - limit the issue of unpopular entities in N1-UNIF by identifying outliers



Complete and Incomplete Classes in Wikidata

Table 2: Lists of 10 randomly picked examples. Left with a low ρ suggesting a complete class, and right a high ρ suggesting an incomplete class.

	SOR $\rho < 0.001$	Distinct		SOR $\rho > 0.1$	Distinct
municipality of Japan	0.0000	739	urban beach	0.1759	683
Philippine TV series	0.0009	822	hydroelectric power station	0.2975	2,936
Landgemeinde of Austria	0.0000	1,116	aircraft model	0.1800	3,919
district of China	0.0009	975	motorcycle manufacturer	0.1758	690
nuclear isomer	0.0002	1,322	local museum	0.1760	1,150
international border	0.0000	529	waterfall	0.1942	5,322
commune of France	0.0001	34,937	race track	0.2783	946
village of Burkina Faso	0.0005	2,723	film production company	0.2107	2,179
supernova	0.0005	5,906	red telephone box	0.3469	2,716
township of Indiana	0.0002	999	mountain range	0.2390	21,390

More about this paper at ISWC 2019 on **Tue 29 Oct at 15:20** in Session 5D

KG Editing Behaviors

- Wikidata editors
  - few editors with many edits and vice versa
  - few items are edited by many editors and v.v.
- Editor lifespan and contributions
  - Power editors
  - Normal editors
- Edit Sessions, type of edits
  - In Wikidata we find shorter times between edits than in Wikipedia.

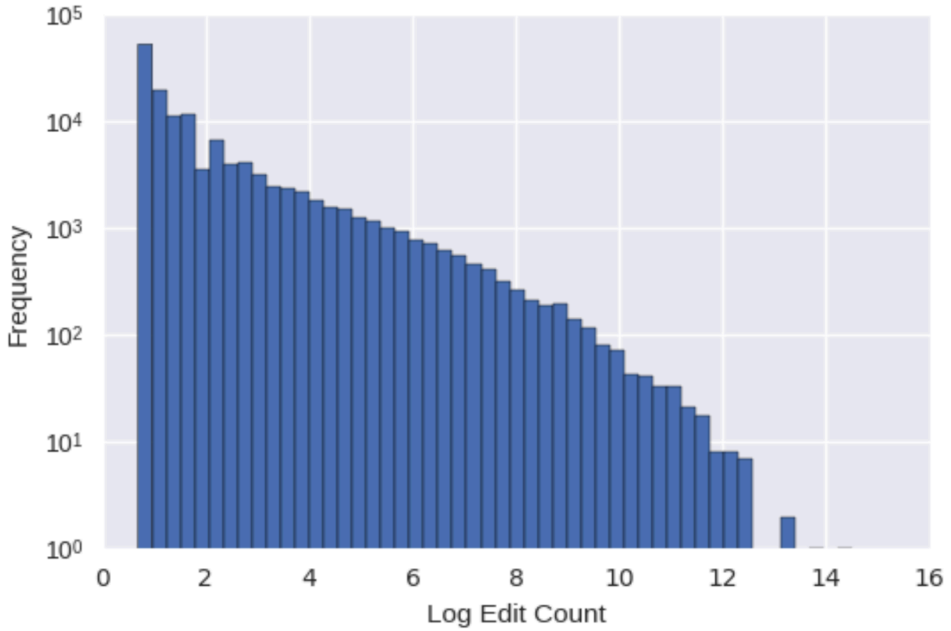


Fig. 1 Total number of edits done by each Wikidata user.

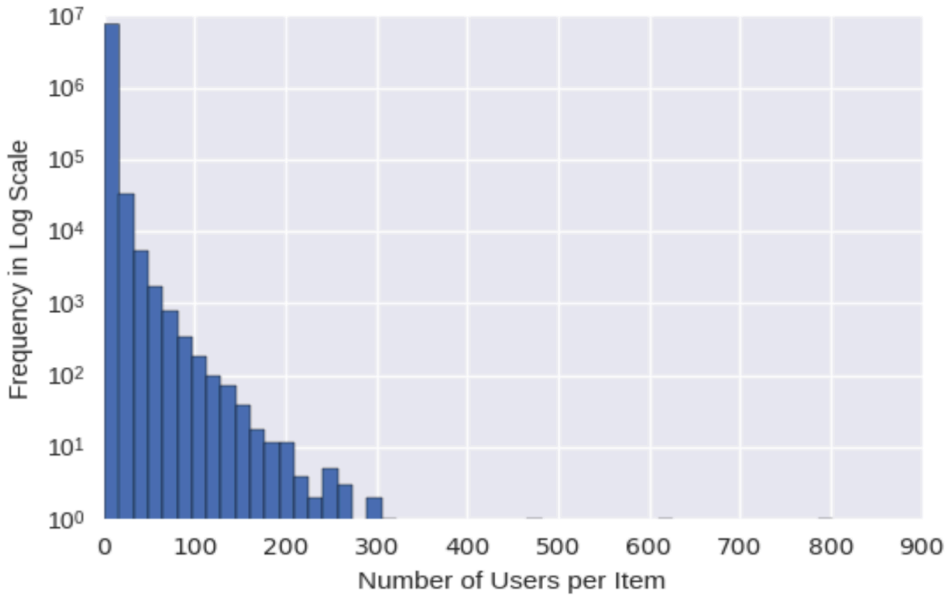


Fig. 2 Histogram of editors per item.

Cristina Sarasua, Alessandro Checco, Gianluca Demartini, Djellel Difallah, Michael Feldman, and Lydia Pintscher. **The Evolution of Power and Standard Wikidata Editors: Comparing Editing Behavior over Time to Predict Lifespan and Volume of Edits**. In: Computer Supported Cooperative Work (CSCW) Special Issue on Crowd Dynamics: Conflicts, Contradictions, and Cooperation Issues in Crowdsourcing, Springer, 2018.

KG Editing Behaviors

- Power editors tend to increase the diversity of the types of edits
- Editors with long lifespan have a constant contribution over months and tend to increase the diversity of type of their edits
- It is possible to automatically predict the future volume of edits and lifespan duration of an editor based on the available edit history

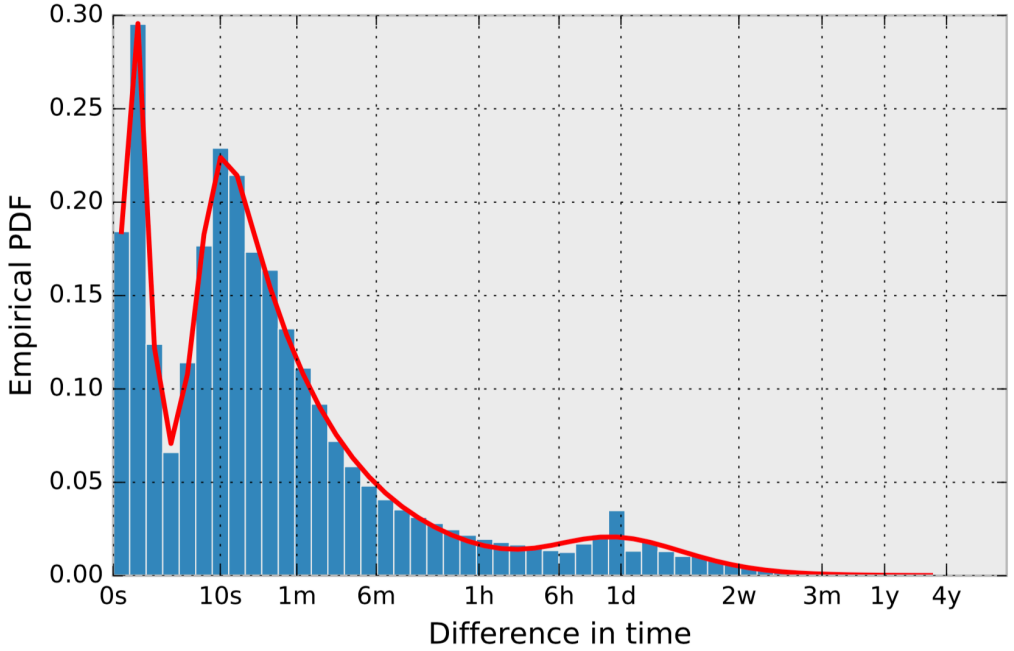
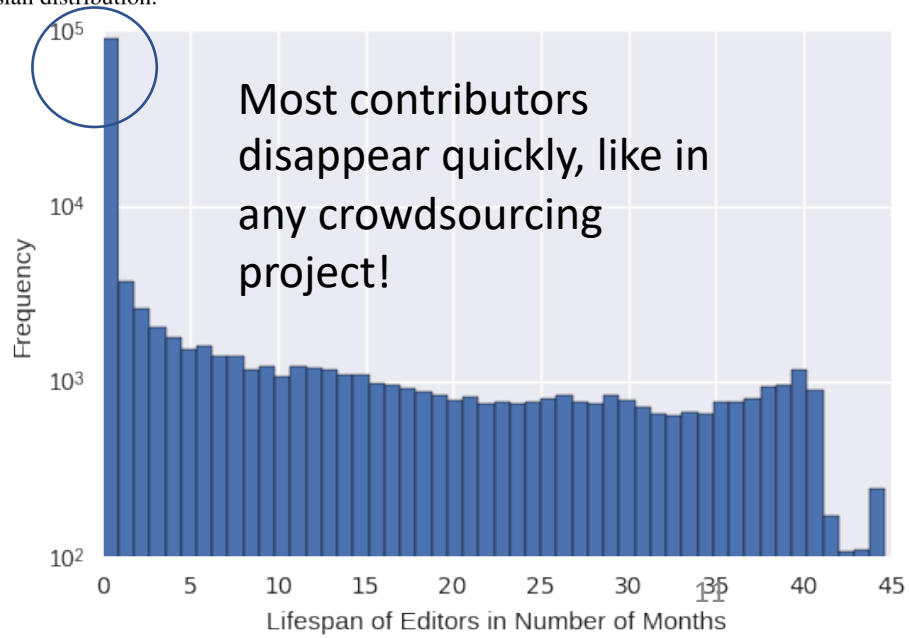


Fig. 8 Distribution of edit differences. The red, continuous line represent the best fit with two log-normal and one expGaussian distribution.



Lessons learned - Bias in KG

- Depending on the contributors, the information stored in KG may differ
- Important to keep track of information provenance
  - Wikidata has *references* for facts
  - It lacks meta-information about how this was identified as a source

- We propose to incorporate provenance metadata about
  - contributors' implicit bias
  - source of evidence
- Reification (statements about statements)
  - Slow, but stores like TripleProv can do it efficiently

Gianluca Demartini. **Implicit Bias in Crowdsourced Knowledge Graphs**. In: The third international workshop on Augmenting Intelligence with Bias-Aware Humans-in-the-Loop co-located with The Web Conference (WWW 2019). San Francisco, CA, USA, May 2019.

Alternative statements in Knowledge Graphs

WIKIDATA

Main page
Community portal
Project chat
Create a new item
Recent changes
Random item
Query Service
Nearby
Help
Donate

Tools

What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Cite this page

Item | Discussion

Catalonia (Q5705)

autonomous community of Spain
Catalunya | Cataluña | Catalonha

In more languages

Statements

instance of	sovereign state (according to 58% of contributors) autonomous community of Spain (according to 42% of contributors) 0 references
	historical nationality 1 reference

Alternative statements in SERPs

Catalonia

Country in Europe (according to 58%)
Autonomous community of Spain (according to 42%)

The Catalonia region, in northeastern Spain, is known for the lively beach resorts of Costa Brava as well as the Pyrenees Mountains. Barcelona, the regional capital, has a historic Gothic Quarter, La Rambla pedestrian mall, museums and several beaches. Antoni Gaudí's distinctive modern art and architecture can be seen at the Sagrada Família Basilica and in the colorful outdoor mosaics of Park Güell.

Area: 32,108 km²

Population: 7.523 million (2016) Instituto Nacional de Estadística

Provinces: Barcelona, Girona, Lleida, Tarragona

Capital: Barcelona

Points of interest: Sagrada Família, Park Güell, Casa Milà, MORE

KGs for Biomedical Applications

Extracting Scientific Concepts from Literature

1. INTRODUCTION

Nowadays, accessing information on the Internet through search engines has become a fundamental life activity. Current web search engines usually provide a ranked list of URLs to answer a query. This type of information access does a good job for dealing with simple navigational queries by leading users to specific websites. However, it is becoming increasingly insufficient for queries with vague or complex information need. Many queries serve just as the start of an exploration of related information space. Users may want to know about a topic from multiple aspects. Organizing the web content relevant to a query according to user intents would benefit user exploration. In addition, a list of URLs couldn't directly satisfy user information need. Users have

- search engine
- web search engine
- navigational query
- user intent
- information need
- web content
- …

Entity type: scientific concept

Roman Prokofyev, Gianluca Demartini, and Philippe Cudré-Mauroux. **Effective Named Entity Recognition for Idiosyncratic Web Collections**. In: 23rd International Conference on World Wide Web (WWW 2014), Research Track. Seoul, South Korea, April 2014.

Traditional NER

Types:

- Maximum Entropy (Mallet, NLTK)
- Conditional Random Fields (Stanford NER, Mallet)

Properties:

- Require extensive training
- Usually domain-specific, different collections require training on the same domain
- Very good at detecting such types as Location, Person, Organization

Proposed Approach

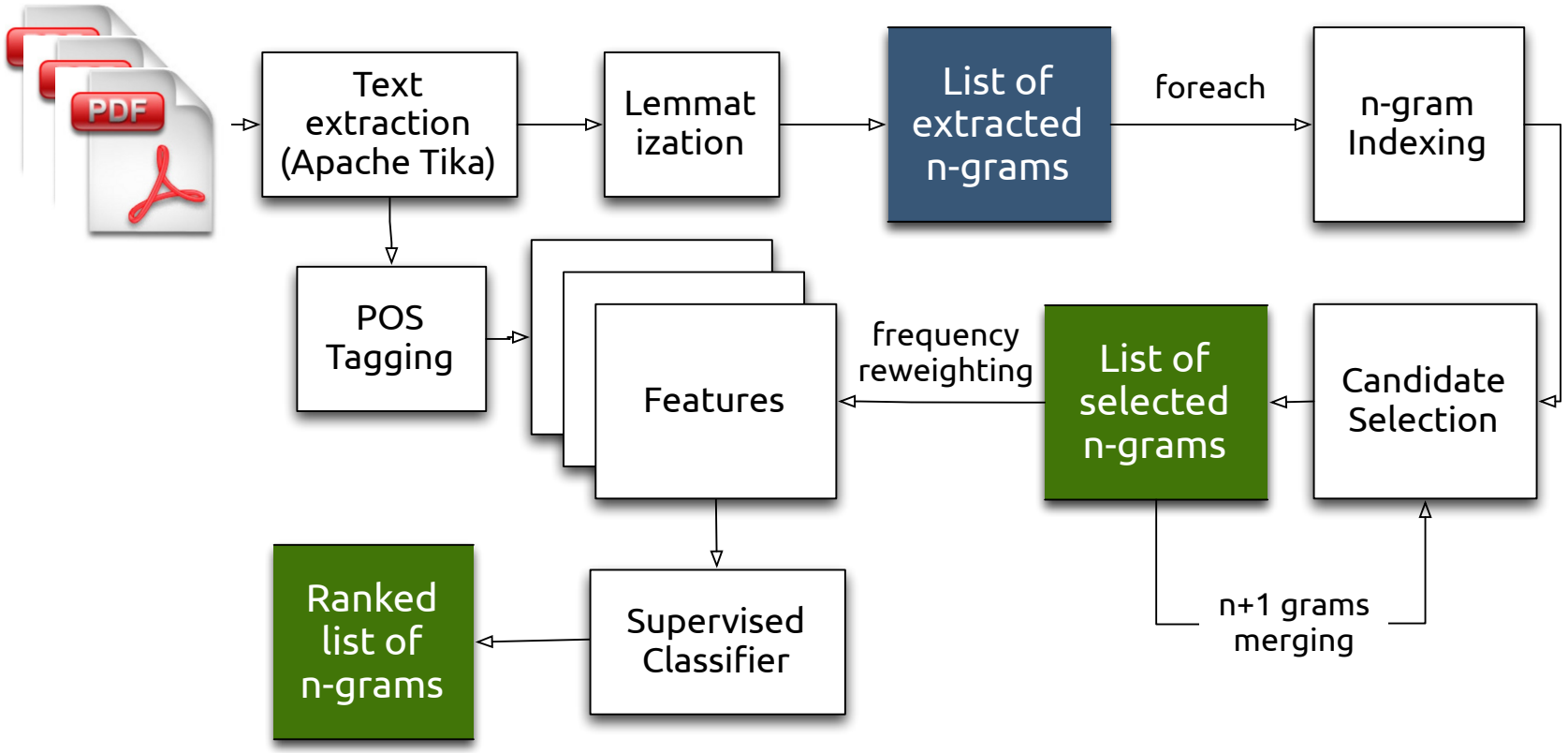
Our problem is defined as a classification task.

Two-step classification:

- Extract candidate named entities using n-gram frequency.
- Classify candidate named entities using supervised classifier.

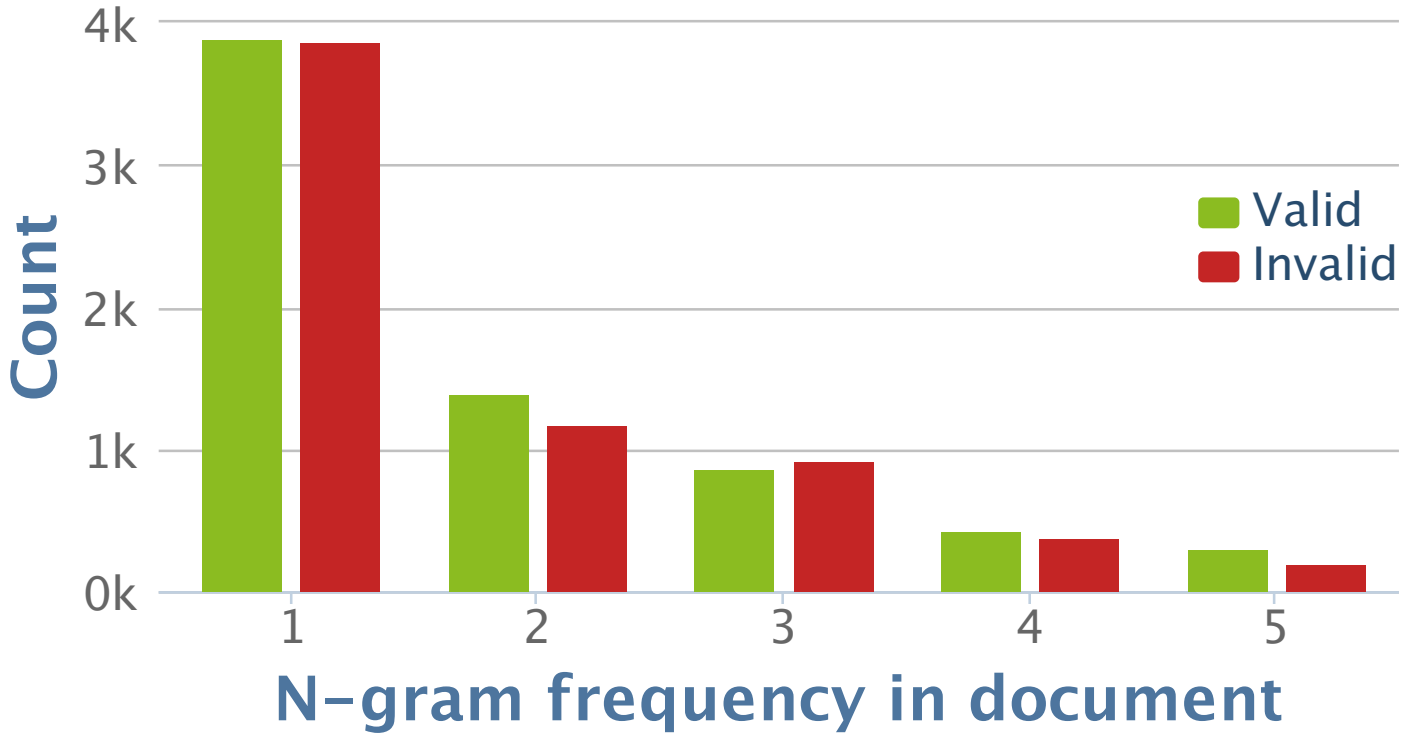
Candidate selection should allow us to greatly reduce the number of n-grams to classify, possibly without significant loss in Recall.

Pipeline


PDF
Text extraction (Apache Tika)
Lemmat ization
List of extracted n-grams
foreach
n-gram Indexing
POS Tagging
Features
frequency reweighting
List of selected n-grams
Candidate Selection
n+1 grams merging
Supervised Classifier
Ranked list of n-grams

Candidate Selection: Discussion

Possible to extract n-grams (n>2) with frequency ≤k



Classifier: Overview

Machine Learning algorithm:

Decision Trees from scikit-learn package.

Feature types:

- POS Tags and their derivatives
- External Knowledge graphs (DBLP, DBPedia)
- DBPedia relation graph
- Syntactic features

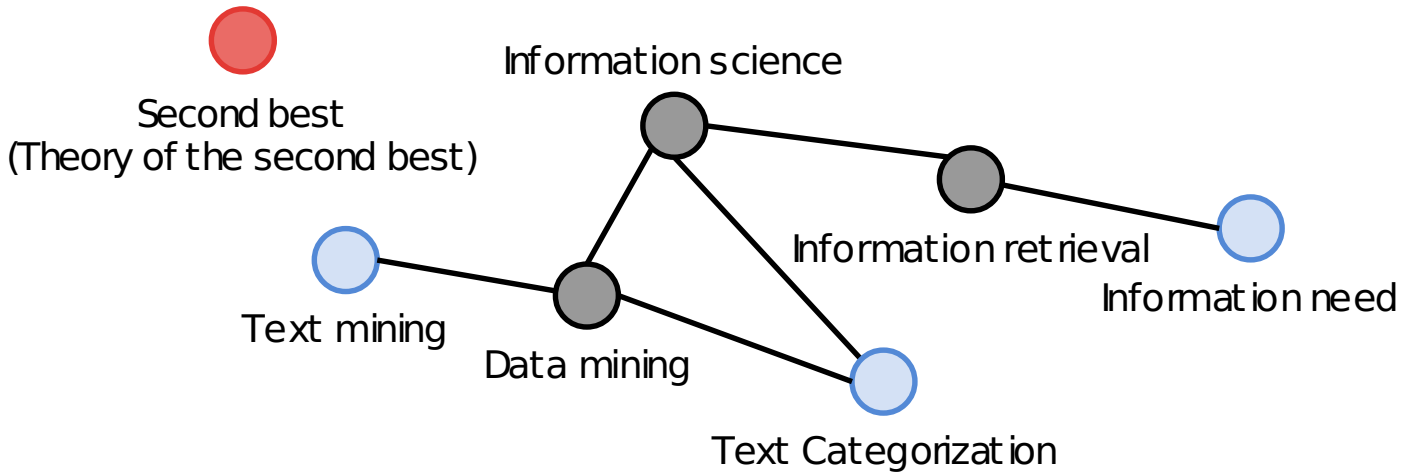
Features: External Knowledge Graphs

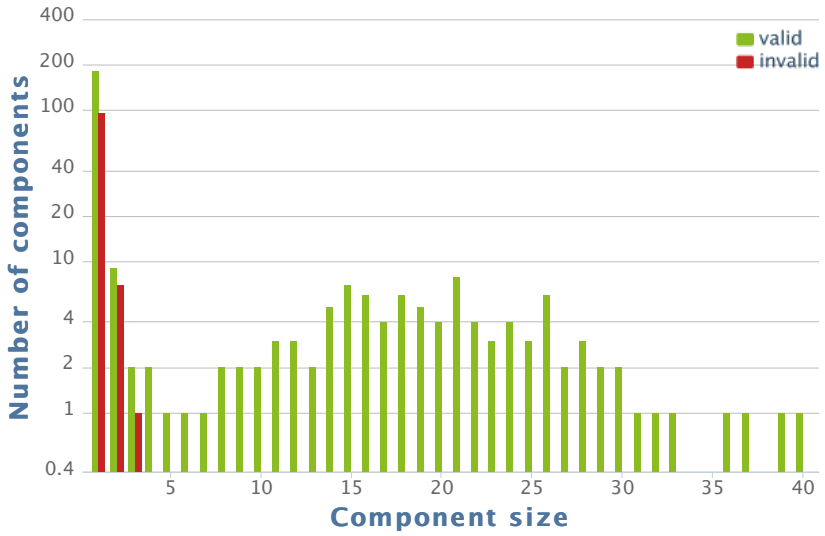
Domain-specific knowledge graphs:

- DBLP (Computer Science): contains author-assigned keywords to the papers
- ScienceWISE: high-quality scientific concepts (mostly for Physics domain) http://sciencewise.info

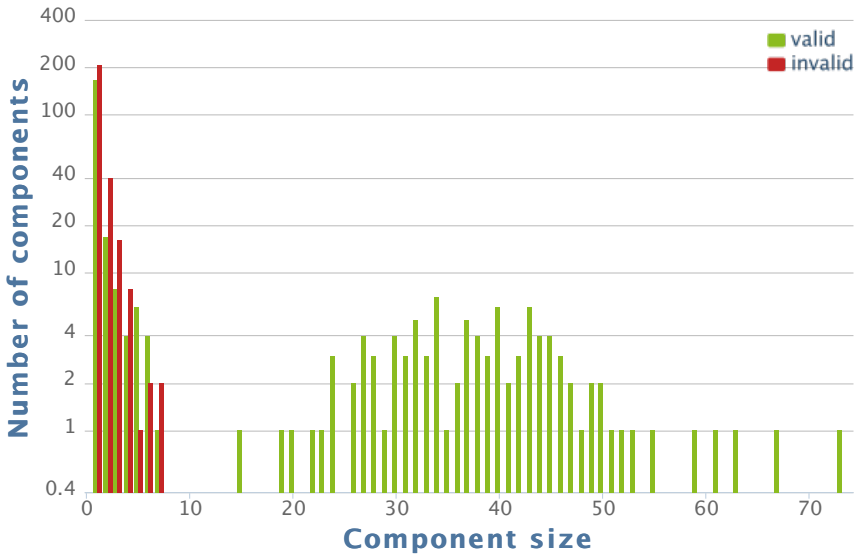
We perform exact string matching with these KGs.

Features: Knowledge Graph





Without redirects



With redirects

Experiments: Feature Importance

	Importance
NN STARTS	0.3091
DBLP	0.1442
Components + DBLP	0.1125
Components	0.0789
VB ENDS	0.0386
NN ENDS	0.0380
JJ STARTS	0.0364

CS Collection, 7 features

	Importance
ScienceWISE	0.2870
Component + ScienceWISE	0.1948
Wikipedia redirect	0.1104
Components	0.1093
Wikilinks	0.0439
Participation count	0.0370

Physics Collection, 6 features

Lessons Learned

- Classic NER approaches are not good enough for scientific literature
- Leveraging the graph of scientific concepts is a key feature
- Domain specific KGs and POS patterns work well
- Experimental results show up to 85% accuracy over different scientific document collections

Health Cards for Consumer Health Search

- A common practice where **people search for health advice online**.
  - 59% of U.S. adults has searched online for health information (Fox & Duggan, 2013)
- Search results strongly bias people's health decisions (Pogacar, 2017)
- People **struggle to understand** health search results (Alpay, 2009).
- 59% of self-diagnosers decided **NOT to confirm their condition with a health professional** (Fox & Duggan, 2013).

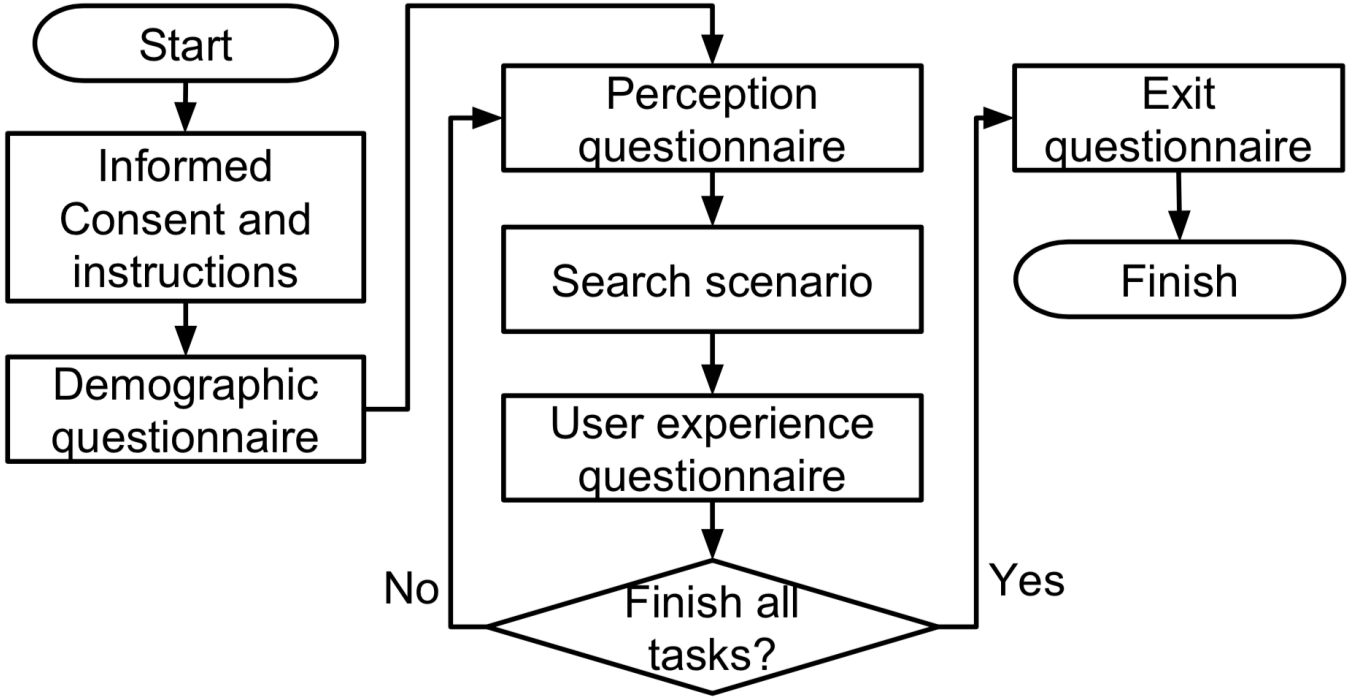
Jimmy, Guido Zuccon, Bevan Koopman, and Gianluca Demartini. **Health Cards for Consumer Health Search**. In: The 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019). Paris, France, July 2019.

KGs for Health Cards in SERP

- Health cards have been used by commercial search engines to present **coherent, easy to understand and trustworthy** health information.
- The appearance of a health card is triggered by a set of queries **related to a specific health condition**.
- Powered by a manually curated background KG
- Our study investigated **the benefits of health cards for broader search tasks** than only to know more about a health condition.

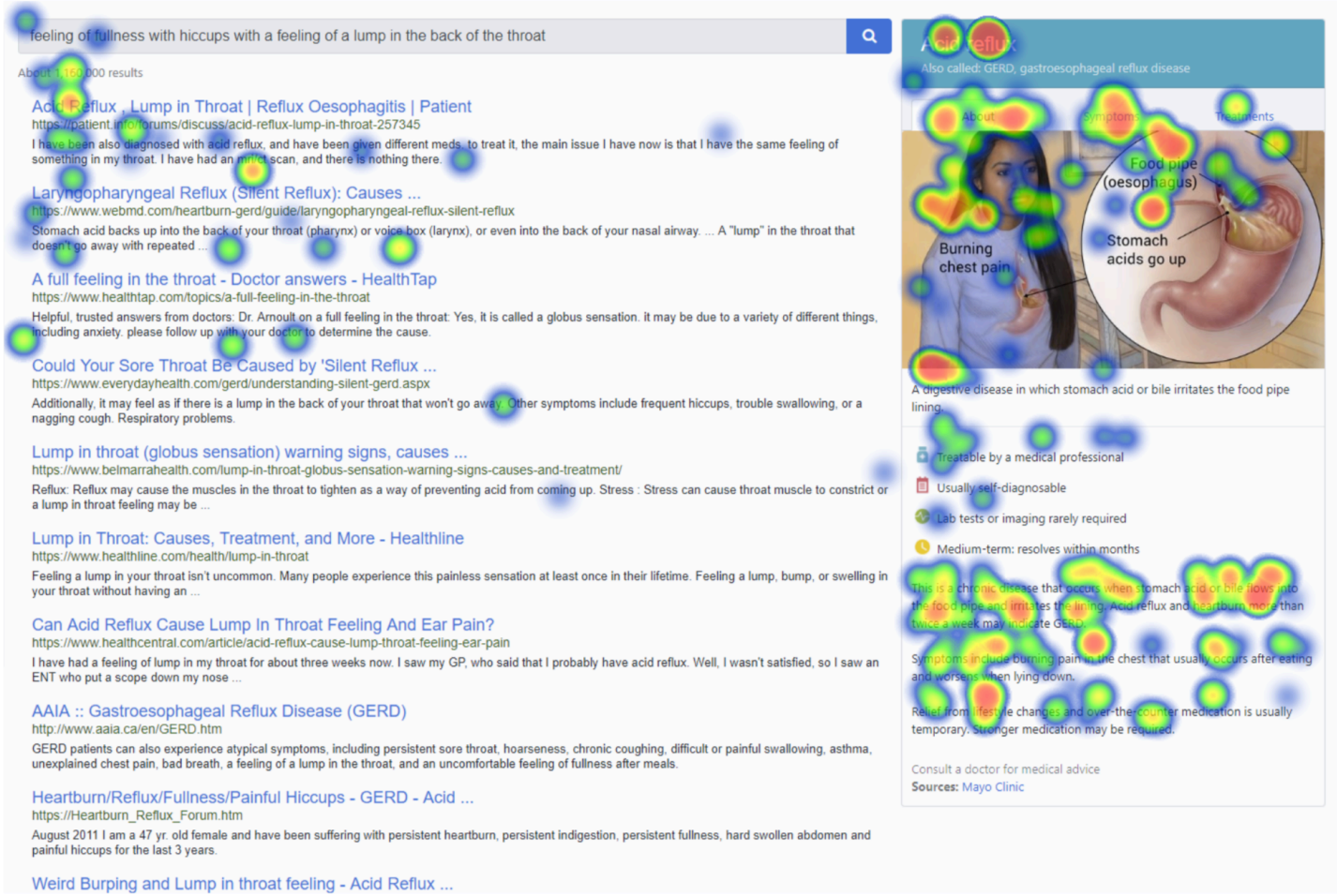
User Study – in lab

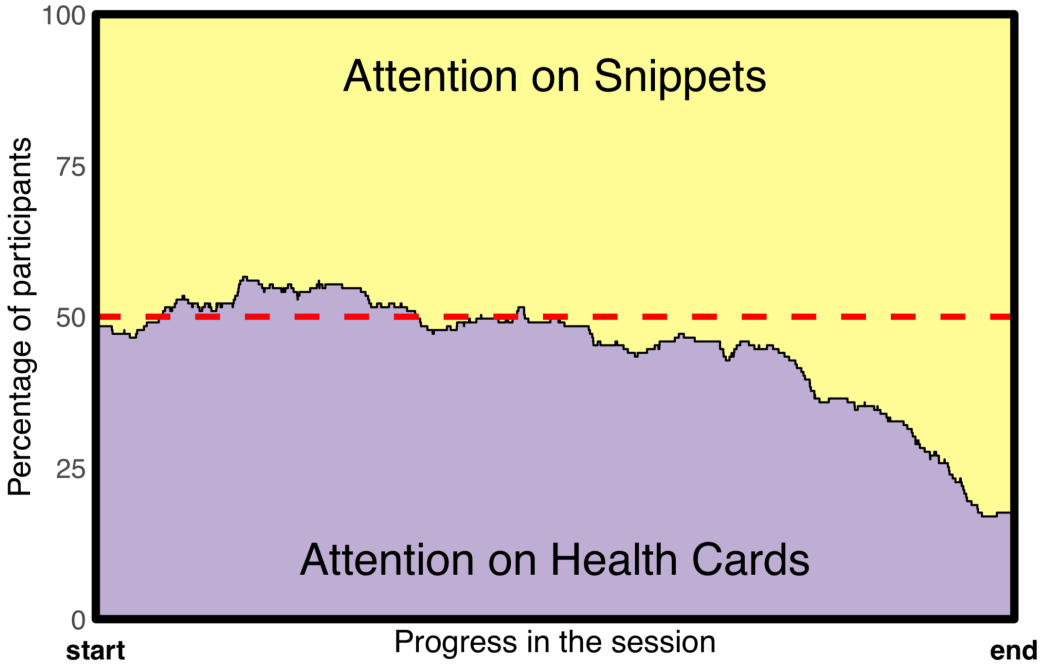
- 48 Participants x 8 health search tasks= 384 Data Points.
- Participants worked on web search tasks using 2 types of UI:
  - With health cards
  - Without health cards



Results - Eye-tracking

- Participants spent 55.40% of their time to observe the health cards





Other Observations

- Participants completed the search tasks **faster and more accurately** when they selected information from health cards
- Overall, presenting health cards reduced the **effort** spent and improved the user's **satisfaction**.
- Health cards **helped the less knowledgeable** to perform as effective as the knowledgeable (in term of correctness).
- Health cards were significantly beneficial for well-defined health search tasks (Factual).
- In contrast, health cards provided no significant benefits for "exploratory" health search tasks (Intellectual).

www.gianlucademartini.net
@eglu81
demartini@acm.org

Conclusions

- Data quality in KG
  - Removing noise (deep learning + crowdsourcing)
- Wikidata
  - Class completeness estimation
  - Editor behaviors / predicting lifespan of editors / editor bias in the KG
- Applications of KG
  - Entity extraction in scientific documents
  - Health Cards in SERP to support Consumer Health Search tasks