The Power of Big Data

Dr Gianluca Demartini
Associate Professor in Data Science
University of Queensland, Australia

gianlucademartini.net
@eglu81

Gianluca Demartini

demartini@acm.org

- BSc MSc in CS at U. of Udine, Italy
- PhD in CS at U. of Hannover, Germany
  - Entity Retrieval
- Worked at U. Sheffield iSchool (UK), the eXascale Infolab (CH), UC Berkeley (on Crowdsourcing), Yahoo! (ES), L3S Research Center (DE)
- Faculty member at the School of ITEE, U. Queensland since 2017
- Tutorials on
  - Entity Search at ECIR 2012 and RuSSIR 2015
  - Crowdsourcing at ESWC 2013, ISWC 2013, ICWSM 2016, WebSci 2016, Facebook

www.gianlucademartini.net

The Distinguished Speakers Program is made possible by

Association for Computing Machinery

Advancing Computing as a Science & Profession

For additional information, please visit http://dsp.acm.org/

About ACM

- ACM, the Association for Computing Machinery (www.acm.org), is the premier global community of computing professionals and students with nearly 100,000 members in more than 170 countries interacting with more than 2 million computing professionals worldwide.

- OUR MISSION: We help computing professionals to be their best and most creative. We connect them to their peers, to what the latest developments, and inspire them to advance the profession and make a positive impact on society.

- OUR VISION: We see a world where computing helps solve tomorrow's problems – where we use our knowledge and skills to advance the computing profession and make a positive social impact throughout the world.

Big Data

- Defined as **V**s
  - **Volume**: Just about *size*, Giga, Tera, Petabytes
  - **Variety**: *Formats*, text, databases, pictures, excel
  - **Velocity**: *Speed*, 10 000 tweets per second, 2 000 pictures on Instagram per second

Data is huge

- Banks, city councils, governments, shops, etc.

- Facebook processes 750TB/day of data
  - 48k iPhones every day
  - 7PB of photo storage / month

- This requires computers (a lot of them!)

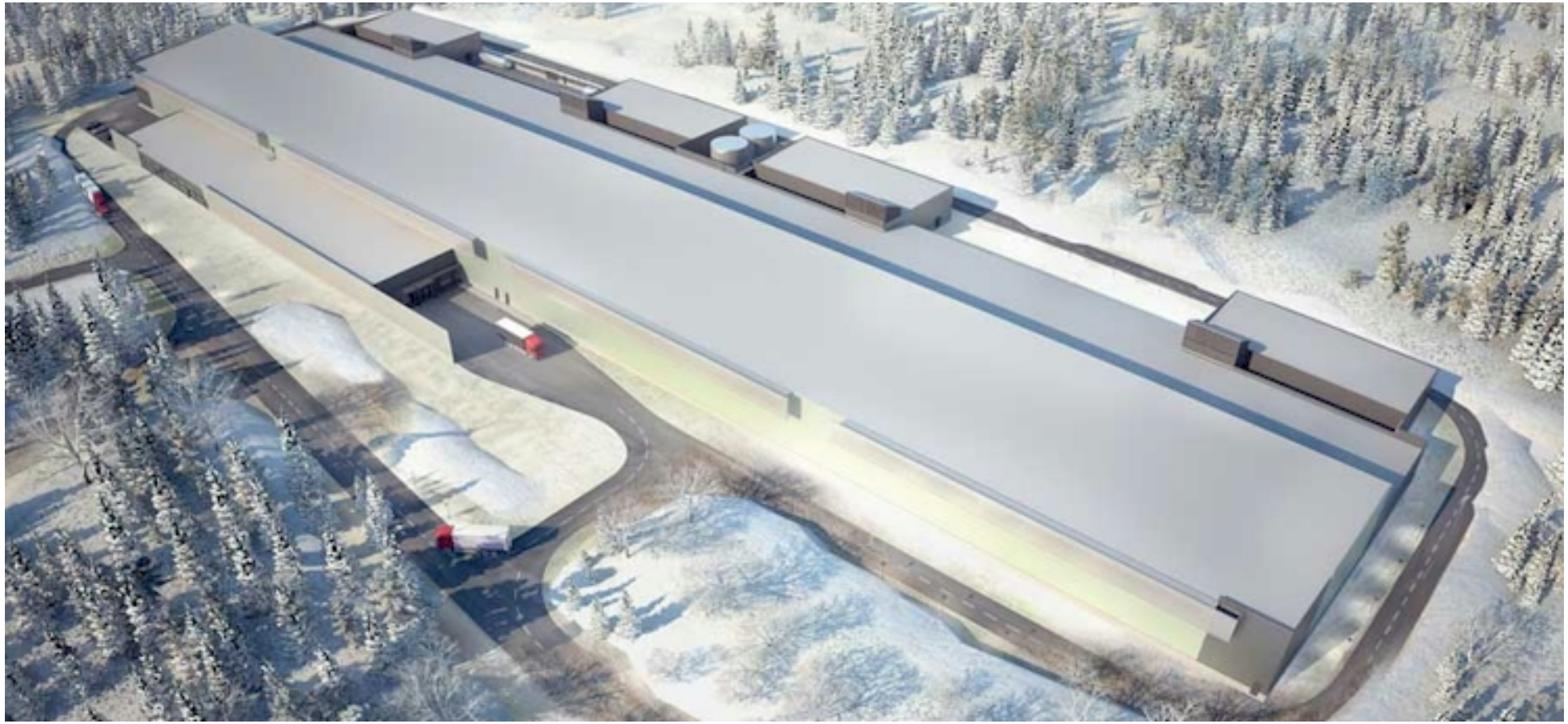
Data is fast (Velocity)

- Twitter fire hose
  - In 2011, 1 000 Tweets per second (TPS)
  - In 2014, 20 000 TPS
  - With peaks: 143K TPS
- Services on top
  - DataSift: aggregate, filter and extract insights
- Not only internet companies!
  - Stock exchange, sensors in water network, smart cities, fitness trackers, etc.

Scale-up vs Scale-out

- Scale-up
  - Increasing the power of your computer (i.e, disk, memory, processor)
- Scale-out
  - Use many standard computers and distribute data and computation over them

Facebook Data Center (Sweden)



Machines

- Google has around 900,000 servers (260 million watts == 200K homes)
- Google accounts for roughly 0.013% of the world's energy consumption
- CERN Large Hadron Collider 180MW

Data from July 2013

Fundamental work

- Google File System, 2003
  - access to data using large clusters of commodity machines
- Big Table, 2003-2006
  - data storage system
  - Distributed map Key -> Value
- Map/Reduce, 2004
  - Programming paradigm over a cluster of machines

Open-Source analogous

- HDFS (Hadoop File System)
  - Distributed File System
- Apache Hbase http://hbase.apache.org/
  - Distributed database
- Apache Hadoop http://hadoop.apache.org/
  - Distributed computation

C.L. Philip Chen, Chun-Yang Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences, Volume 275, 10 August 2014, Pages 314-347, ISSN 0020-0255, http://dx.doi.org/10.1016/j.ins.2014.01.015.

(http://www.sciencedirect.com/science/article/pii/S0020025514000346)

Should we care?

- This data is about us!
- **Data**: GMail, Facebook, debit cards, shopping fidelity cards, transport, mobile phones, …
- **Usage**: Mortgage application, health insurance, car insurance

Algorithms rule the world

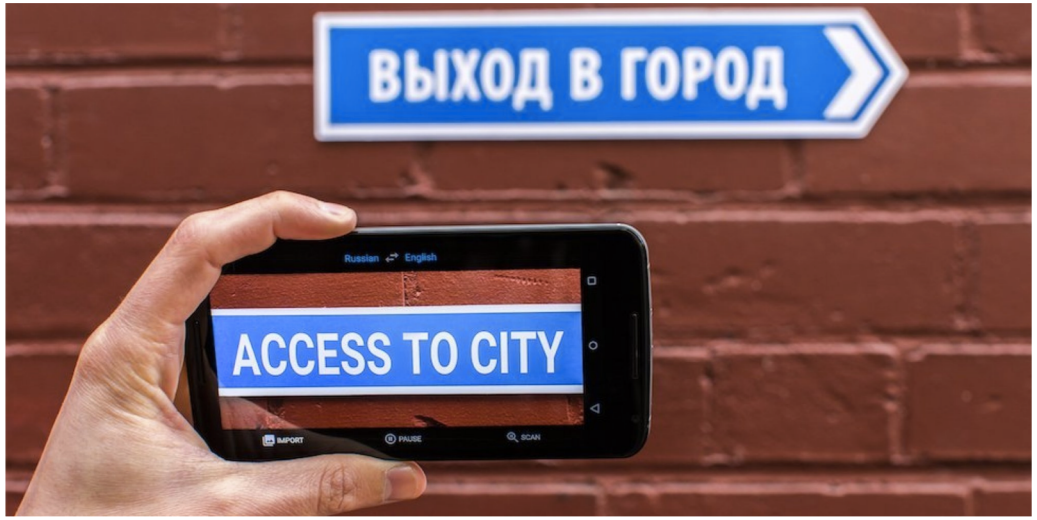
- Some data must not be processed by people!
  - GMail content is processed by computers to decide which advertisement you see on the Web

Algorithms rule the world

- **Uber** prices are decided by a software programs
  - The boss of Uber drivers is a computer
  - It decides how they work and how much money they make
- Computers know a lot about people but not the other way around

Is it all bad?

- Duolingo: Data-driven foreign language learning
  - What is the best way to learn a language depends on your native language
- Language translation



Data Science

- “Data Scientist: The Sexiest Job of the 21st Century”, in Harvard Business Review

- Companies want **data-driven decisions**

- Graduates from the MSc Data Science at UQ go work in:
  - Telecommunication data analysis
  - Cancer research
  - Finance sector
  - …

Research Interests

- **Entity-centric Information Access** (since 2005)
  - Structured/Unstruct data (SIGIR 12), TRank (ISWC 13, WSemJ 16)
  - Entity Extraction (WWW 14), Prepositions (CIKM 14), Entity Cards (SIGIR 19)
  - IR Evaluation (IRJ 2015, ECIR 16 Best Paper, CIKM 17, SIGIR 18, CIKM 19)
- **Human-in-the-loop Information Systems** (since 2012)
  - Entity Linking (WWW 12, VLDBJ), CrowdQ (CIDR 13)
  - Huml systems overview (COMNET 15, FnT 17)
- **Better Crowdsourcing Platforms** (since 2013)
  - Platform Dynamics (WWW 15), Wikidata (CSCWJ 18)
  - Pick-a-Crowd (WWW 13), Scheduling Tasks (WWW 16)
  - Agreement (ICTIR 17, HCOMP 17), Pricing Tasks (HCOMP 14, CSCW 20)
- **Human Factors in Crowdsourcing** (since 2015)
  - Malicious Workers (CHI 15), Attack Schemes (HCOMP 18 Best Paper, JAIR)
  - Modus Operandi (UBICOMP17, HT19, WSDM20), Bias (SIGIR18, ECIR20)
  - Time (HCOMP 16), Complexity (HCOMP 16), Abandonment (WSDM19, TKDE)
- **Better Data** (since 2019)
  - Data Workers (SIGIR 20), Misinfo (SIGIR 20, CIKM 20), Know. Graphs (ISWC 19)
  - Remove noise (WWW 19), Unknown Unknowns (ECAI 20)
  - User Behavior Embeddings (CIKM 20)

Thanks to:

Australian Government
Australian Research Council

European Commission

EPSRC
Engineering and Physical Sciences Research Council

facebook research

Entity-Centric Information Access

tom cruise

Gianluca

All | News | Images | Videos | Shopping | More | Search tools

About 78,300,000 results (0.47 seconds)

Official Tom Cruise: Edge Of Tomorrow, Movies, Bio, News ...
www.tomcruise.com/
OFFICIAL **TOM CRUISE** SITE: View the latest EDGE OF TOMORROW trailer! Watch career movie trailers, videos, and retrospective. Read the **Tom Cruise** ...

Tom Cruise - IMDb
www.imdb.com/name/nm0000129/
Tom Cruise, Actor: Top Gun. If you had told fourteen-year-old Franciscan seminary student Thomas Cruise Mapother IV that one day in the not-too-distant future ...

Tom Cruise - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Tom_Cruise
Tom Cruise is an American actor and filmmaker. Cruise has been nominated for three Academy Awards and has won three Golden Globe Awards. He started his ...
Tom Cruise filmography - Mimi Rogers - Katie Holmes - Nicole Kidman

In the news

Scientologist who worked with Tom Cruise condemned to horrific work camp over lesbian kiss
PinkNews - 2 days ago
A former Scientologist, who worked with celebrities like **Tom Cruise** and John Travolta, has ...

Jerry Bruckheimer confirms Tom Cruise is signed up for Top Gun 2

Tom Cruise

Actor

tomcruise.com

Tom Cruise is an American actor and filmmaker. Cruise has been nominated for three Academy Awards and has won three Golden Globe Awards. He started his career at age 19 in the 1981 film Endless Love. Wikipedia

Born: July 3, 1962 (age 53), Syracuse, New York, United States

Height: 1.7 m

Spouse: Katie Holmes (m. 2006–2012), Nicole Kidman (m. 1990–2001), Mimi Rogers (m. 1987–1990)

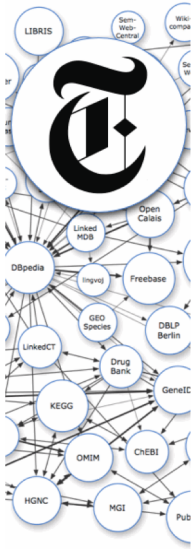
- Entity-seeking queries make up 40-50% of the query volume
  - Jeffrey Pound, Peter Mika, Hugo Zaragoza: Ad-hoc object retrieval in the web of data. WWW 2010: 771-780
  - Thomas Lin, Patrick Pantel, Michael Gamon, Anitha Kannan, Ariel Fuxman: Active objects: actions for entity-centric search. WWW 2012: 589-598
- Show a summary of the most likely information-needs
  - Including related entities for navigation
  - *Roi Blanco, Berkant Barla Cambazoglu, Peter Mika, Nicolas Torzec: Entity Recommendations in Web Search. ISWC 2013*

Web of Data

- Freebase
  - Acquired by Google in July 2010.
  - Knowledge Graph launched in May 2012.
  - Read-only in December 2014 -> WikiData

- Schema.org
  - Driven by major search engine companies
  - Machine-readable annotations of Web pages

- Linked Open Data
  - 31 billion triples, Sept 2011
  - 90 billion triples, Aug 2015 (stats.lod2.eu)



Entity Linking

- Looking at data integration across sources

APRIL 9, 2012, 1:15 PM MERGERS & ACQUISITIONS

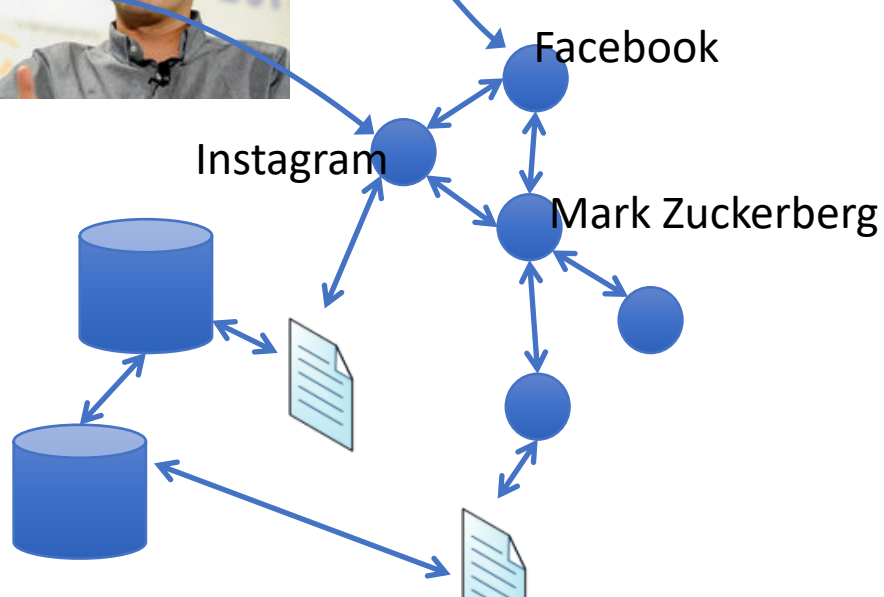
Facebook Buys Instagram for $1 Billion

BY EVELYN M. RUSLI

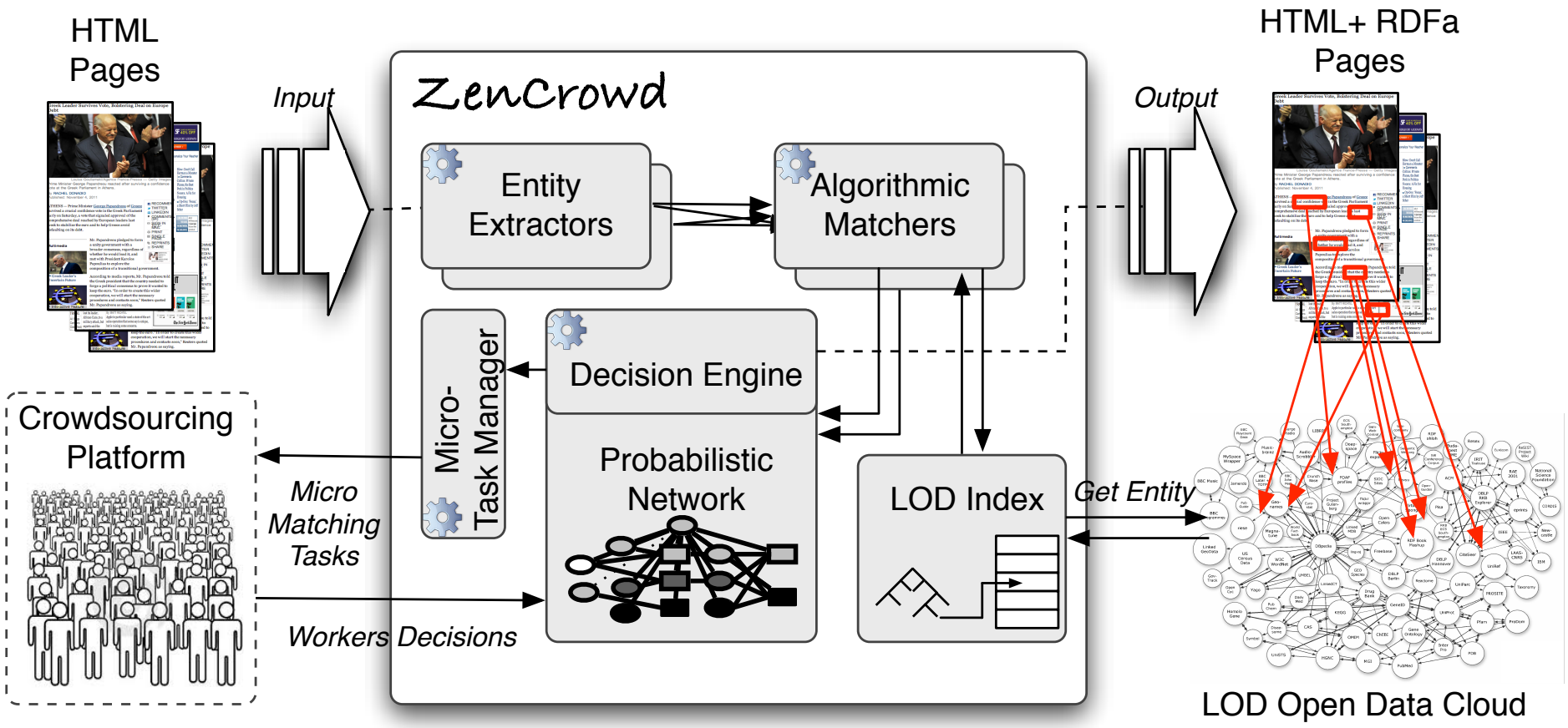
2:02 p.m. | Updated

Facebook is not waiting for its initial public offering to make its first big purchase.

In its largest acquisition to date, the social network has purchased Instagram, the popular photo-sharing application, for about $1 billion in cash and stock, the company said Monday.



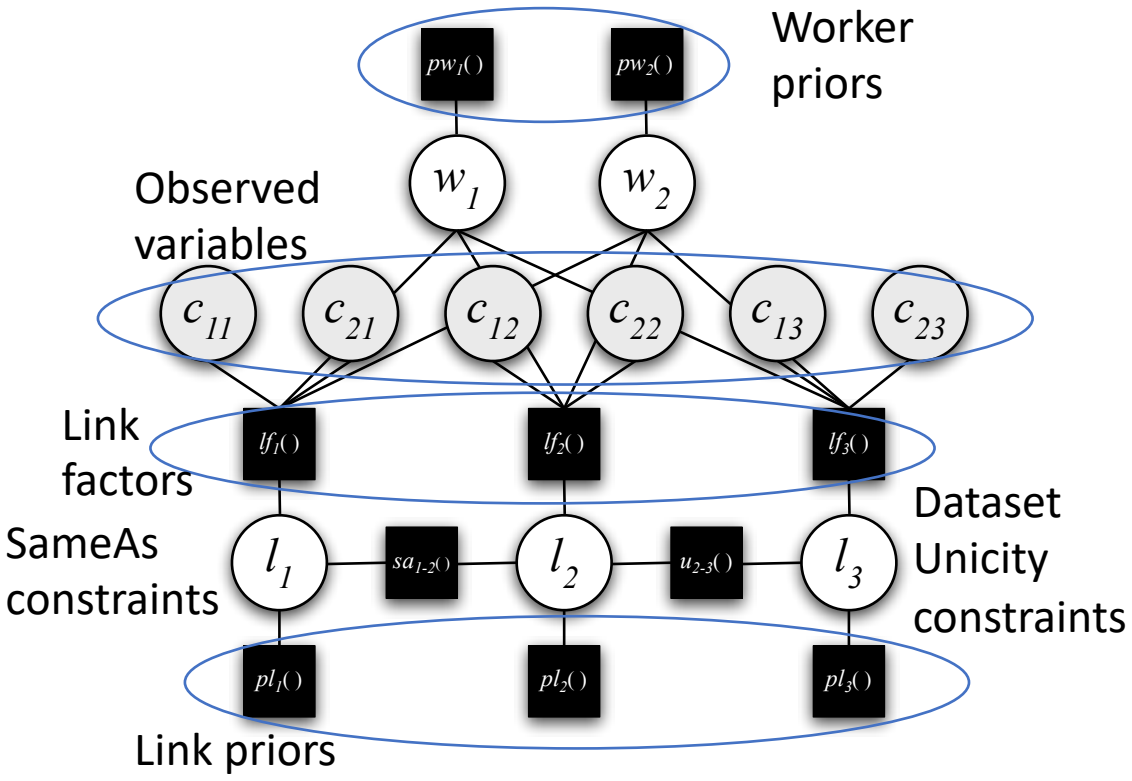
ZenCrowd Architecture



Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking. In: 21st International Conference on World Wide Web (**WWW 2012**).

Entity Factor Graphs

- Graph components
  - Workers, links, clicks
  - Prior probabilities
  - Link Factors
  - Constraints
- Probabilistic Inference
  - Select all links with posterior prob >τ



2 workers, 6 clicks, 3 candidate links

ZenCrowd Summary

- ZenCrowd: Probabilistic reasoning over automatic and crowdsourcing methods for entity linking
- Standard crowdsourcing improves 6% over automatic
- 4% - 35% improvement over standard crowdsourcing
- 14% average improvement over automatic approaches

- Follow up-work (VLDBJ, 2013):
  - Also used for **instance matching** across datasets
  - 3-way blocking with the crowd

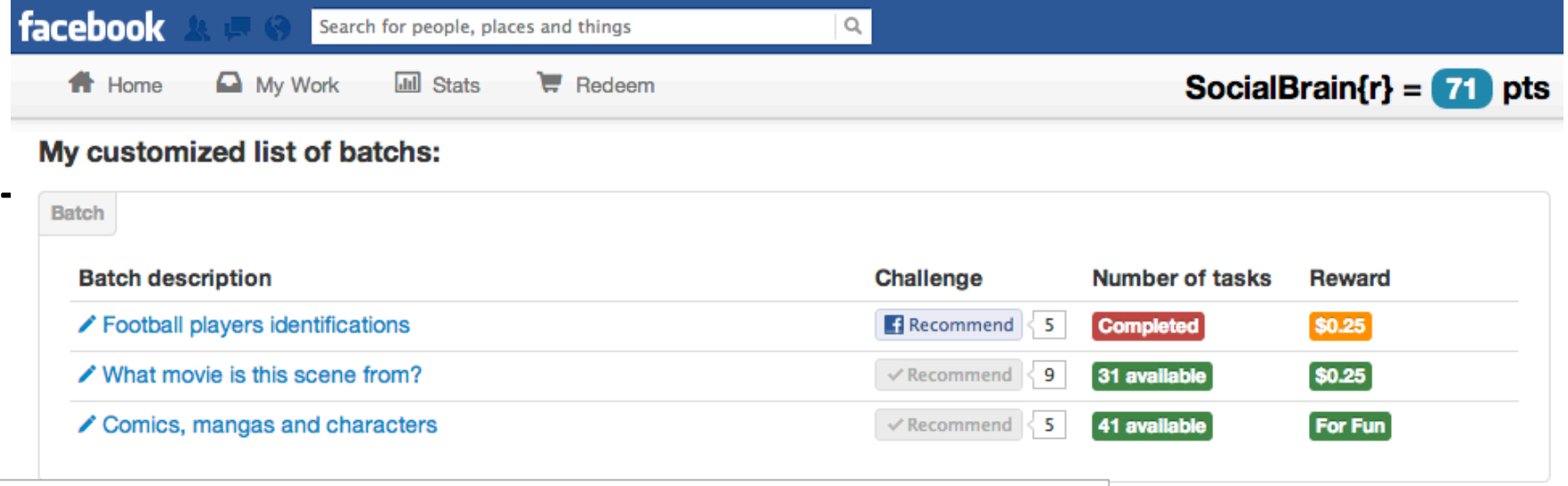
Hybrid Human-Machine Systems

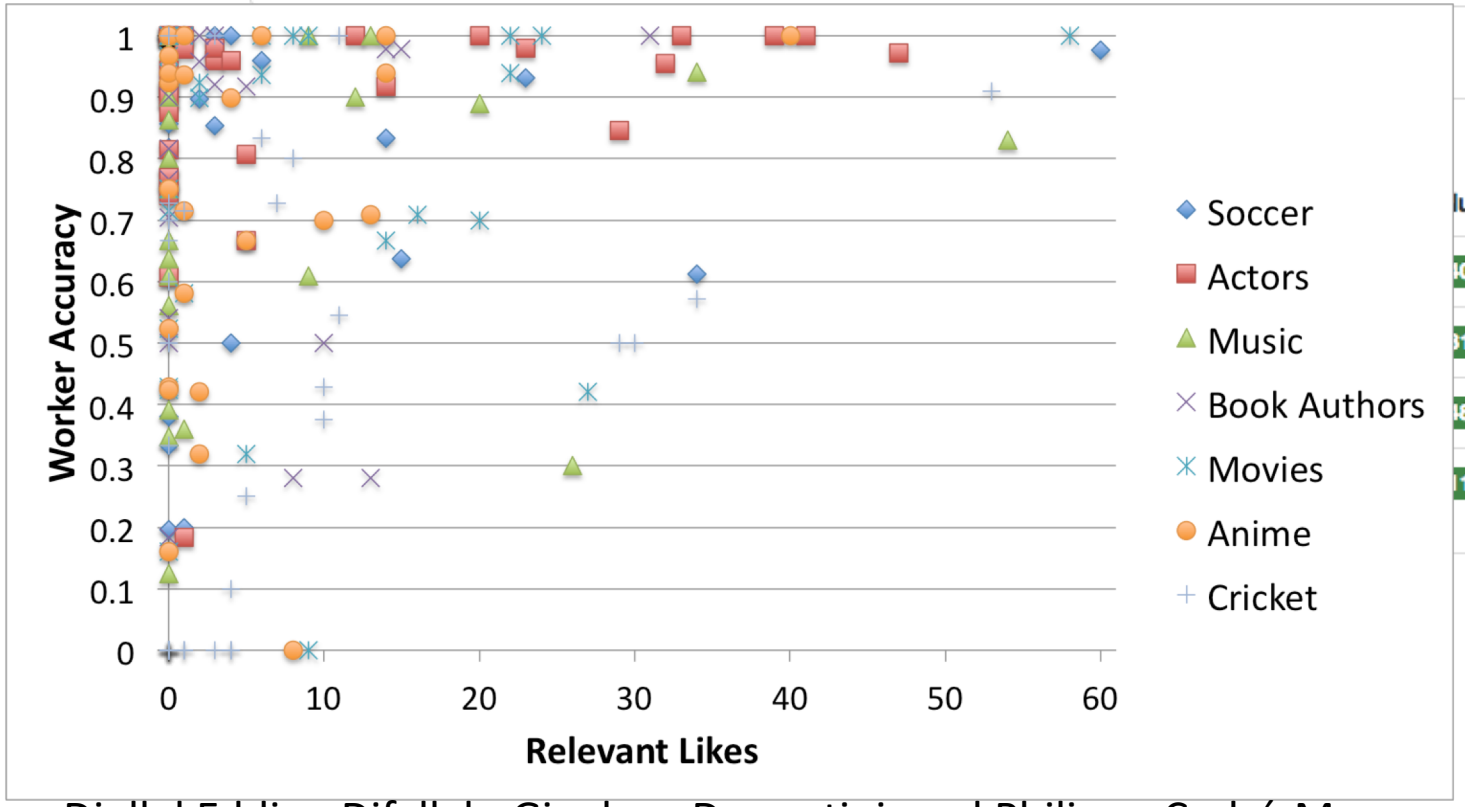
- Use Machines to scale over large amounts of data
- Keep humans in the loop
  - By means of Crowdsourcing
  - To make sure the quality of the data processing is good
- Crowd for Pre-processing vs Post-processing

Lessons Learnt

- Crowdsourcing + Prob reasoning works!
- But
  - Different worker communities perform differently
  - Many low quality workers
  - Completion time may vary (based on reward)
- Need to **find the right workers** for your task (see WWW2013 and CHI2015 papers)
- Need to make sure **high priority tasks** are completed fast (see WWW2016 paper)

Pick-A-





Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Pick-A-Crowd: Tell Me What You Like, and I'll Tell You What to Do. In: **WWW2013**

How people interact with information

- We ask people to label data to then train AI models
- That is, we ask people to look at data and make decisions

- How do people label fake news? (SIGIR 2020, CIKM 2020)
- How do data scientists prepare data? (SIGIR 2020)

Crowdsourcing Truthfulness Judgements

- ~600 MTurk US workers
- To assess truthfulness of
  - US political statements (Politifact)
  - non-US political statements (ABC)
- 3 scales (3, 6, and 100 levels)

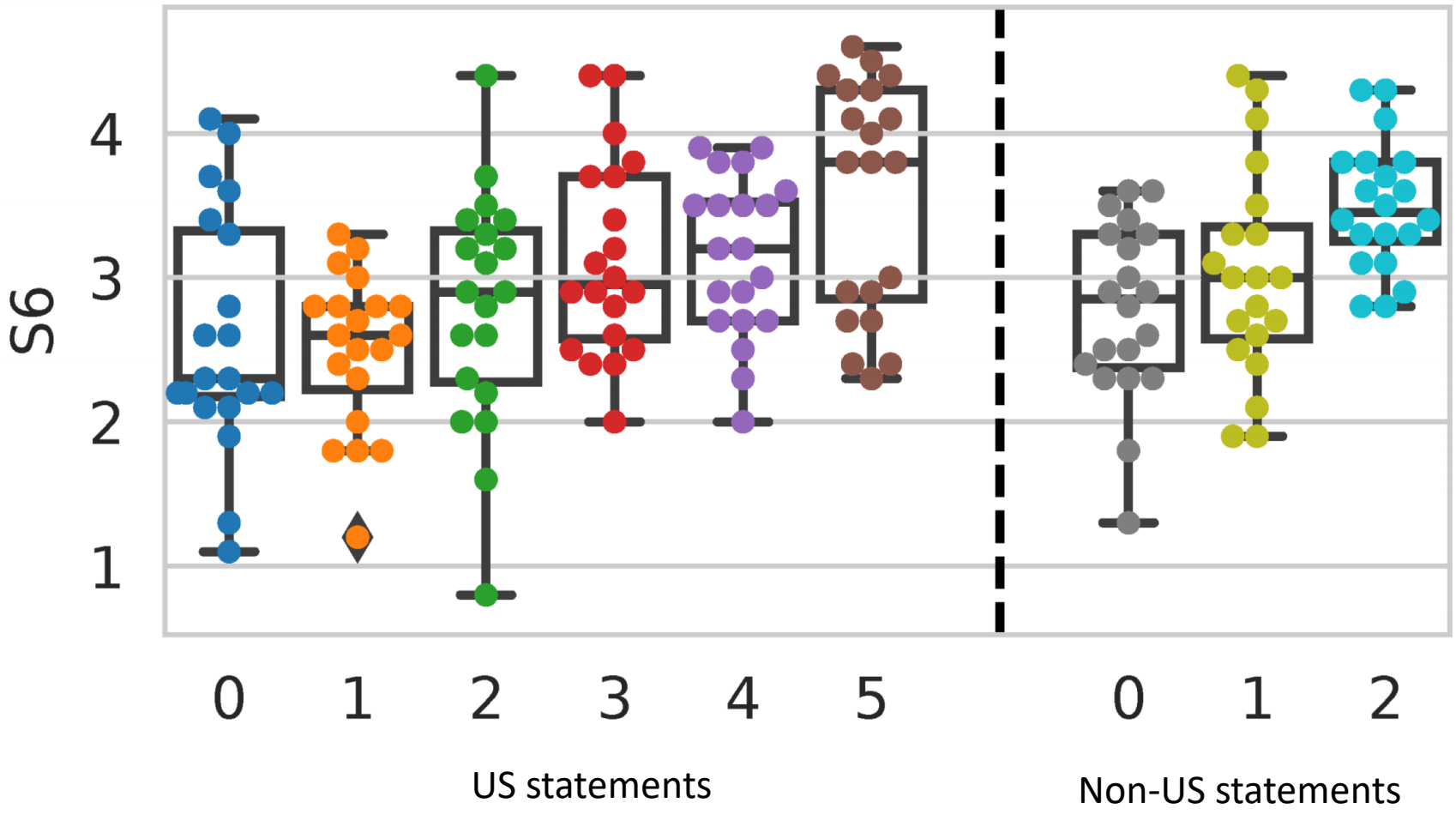
Table 1: Example of statements in the PolitiFact and ABC datasets.

	Statement	Speaker, Year
PolitiFact Label: mostly-true	"Florida ranks first in the nation for access to free prekindergarten."	Rick Scott, 2014
ABC Label: in-between	"Scrapping the carbon tax means every household will be $550 a year better off."	Tony Abbott, 2014

- All data:
- https://github.com/kevinRoitero/crowdsourcingTruthfulness

Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro and Gianluca Demartini. **Can The Crowd Identify Misinformation Objectively? The Effects of Judgments Scale and Assessor's Bias**. In: The 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)

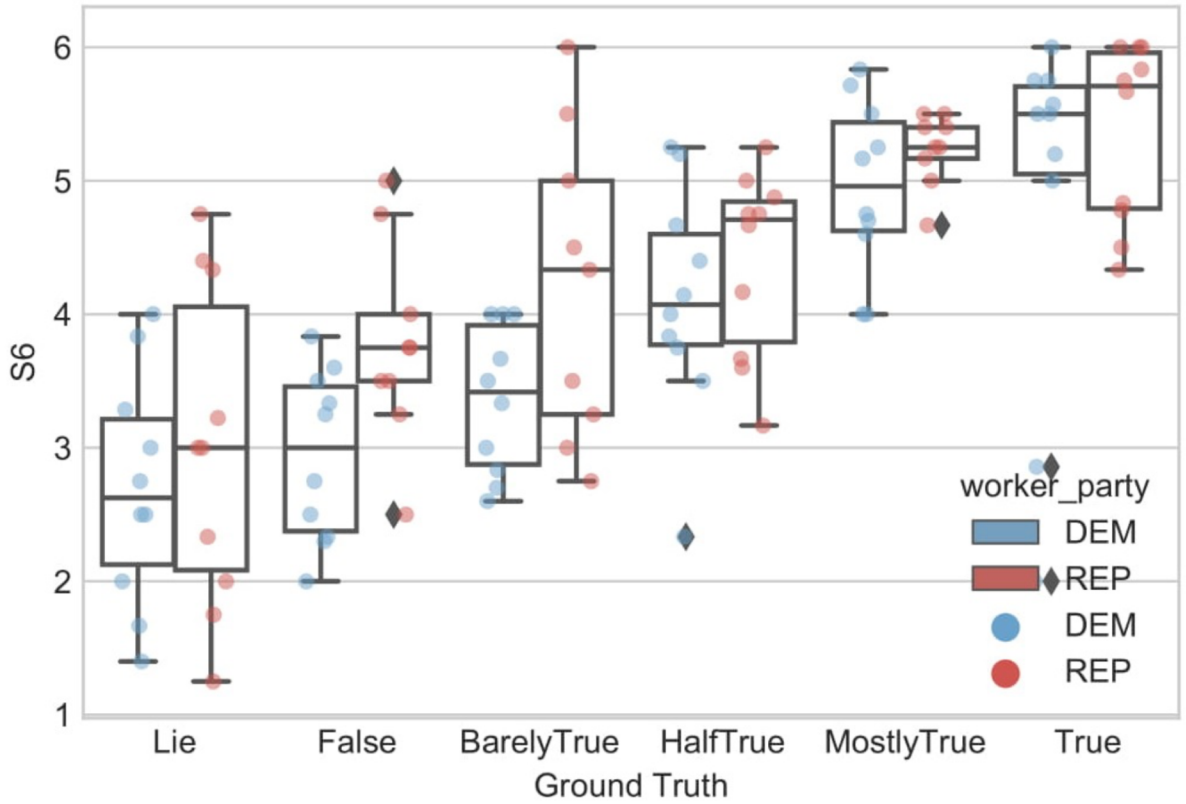
Crowd Performance VS Expert Ground Truth



Fake News labelling - Political bias

- Fact checkers are expert journalists verifying sources and validating news

- Can we (non-experts) do the same?
- Non-expert people who vote REP are more likely to believe to statements by REP politicians



David La Barbera, Kevin Roitero, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. **Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias**. In: The 42nd European Conference on Information Retrieval (ECIR 2020). Lisbon, Portugal, April 2020.

Summary

- **Human-in-the-loop AI** systems can solve complex tasks at scale by combining
  - The ability of machines to scale over **very large amounts of data**
  - The quality of human intelligence and **manual content curation**
- Humans come with challenges
  - Data-driven (activity logging and log analysis) **behavior understanding**
  - System optimization (improving **efficiency and effectiveness**)
- Ongoing research
  - Better AI with humans to *pre-process* or *post-process* data
  - Means to deal with **implicit bias** to **improve the quality of data** with humans in the loop

Gianluca Demartini, Djellel Eddine Difallah, Ujwal Gadiraju, and Michele Catasta. **An Introduction to Hybrid Human-Machine Information Systems**. In: Foundation and Trends in Web Science Vol. 7: No. 1, pp 1-87. 2017.

Conclusions

gianlucademartini.net
demartini@acm.org
@eglu81

- Data is ubiquitous
- It is used to make decisions and influences businesses, jobs, and leisure time

- There is need for scalable data management infrastructures
  - Entity-centric approaches
  - Hybrid Human-Machine Information Systems