The Power of Big Data

Dr Gianluca Demartini
Senior Lecturer in Data Science, University of Sheffield

gianlucademartini.net
@eglu81

Gianluca Demartini

g.demartini@sheffield.ac.uk

- B.Sc., M.Sc. at U. of Udine, Italy
- Ph.D. at U. of Hannover, Germany
  - Entity Retrieval
- Worked at the eXascale Infolab U. Fribourg (Switzerland), UC Berkeley (on Crowdsourcing), Yahoo! (Spain), L3S Research Center (Germany)
- Senior Lecturer in Data Science at the iSchool, **U. of Sheffield**
- Tutorials on Entity Search at ECIR 2012 and RuSSIR 2015, on 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!)

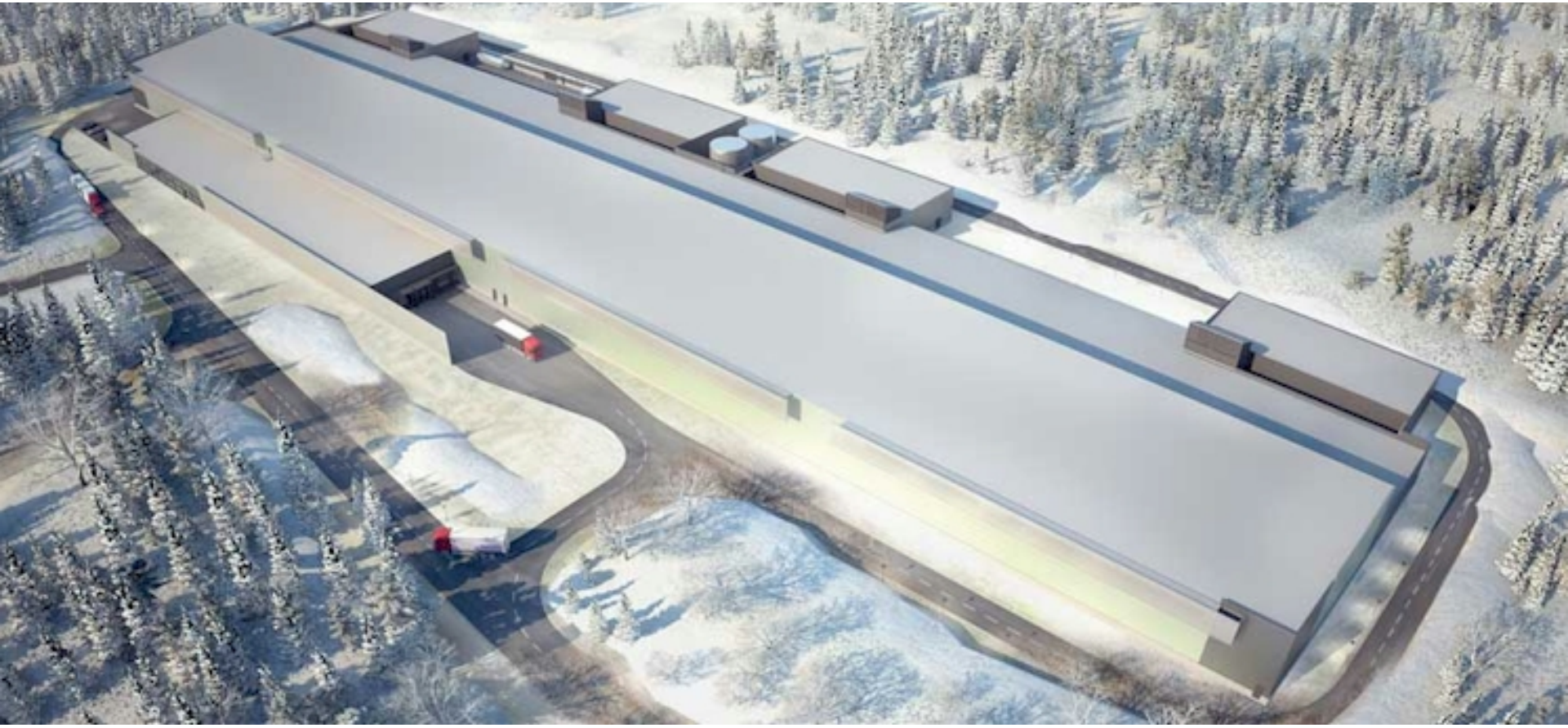
Data is fast (Velocity)

- 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
- **Apache Spark** http://spark.apache.org/
  - Integrated solution

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.

Hadoop Distributed File System (HDFS)

- Inspired by Google File System
- Scalable, distributed, portable file system written in Java for Hadoop framework
- Primary distributed storage used by Hadoop applications
- HFDS can be part of a Hadoop cluster or can be a stand-alone general purpose **distributed file system**
- Reliability and **fault tolerance** ensured by replicating data across multiple hosts
- Zookeeper for the distributed coordination

MapReduce (MR)

- High-level programming model and implementation for large-scale parallel data processing
- Commodity hardware
- Fault-tolerant
- **Divide & conquer**: partition a large problem into smaller sub-problems
  - **Independent sub-problems** can be executed in parallel by workers
  - Intermediate results from each worker are **combined** to get the final result

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

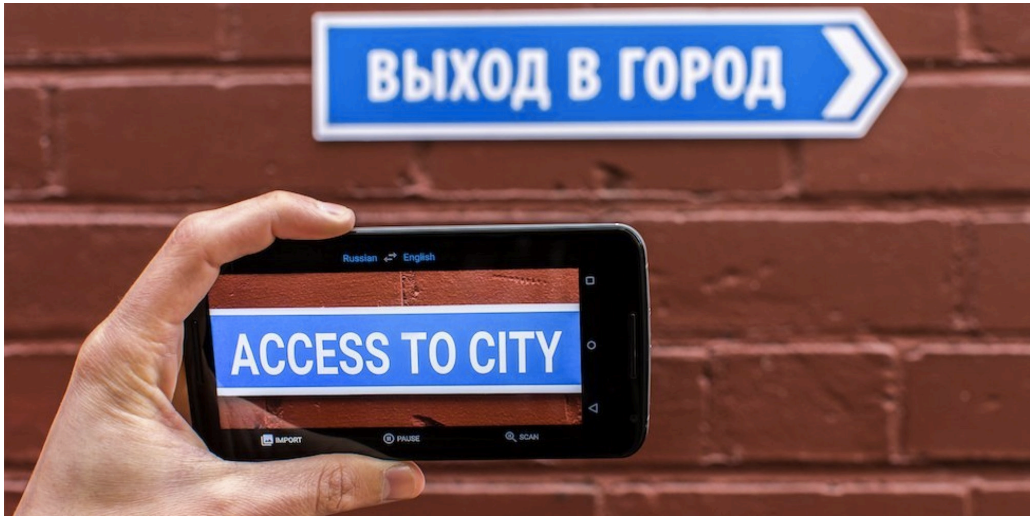
- Some data must not be processed by people!
  - GMail content is processed by computers to decide which advertisement you see on the Web

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
- The Gig economy!

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 in Sheffield go work in:
  - Telecommunication data analysis
  - Cancer research
  - Housing market

Research Interests

- **Entity-centric Information Access** (2005-now)
  - Structured/Unstruct data (SIGIR 12), TRank (ISWC 13, WSemJ 16)
  - NER in Scientific Docs (WWW 14), Prepositions (CIKM 14)
  - IR Evaluation (ECIR 16 Best Paper Award, IRJ 2015)
- **Hybrid Human-Machine Systems** (2012-now)
  - ZenCrowd (WWW 12, VLDBJ), CrowdQ (CIDR 13)
  - Human Memory based Systems (WWW 14, PVLDB)
  - Hybrid systems overview (COMNET, 2015)
- **Better Crowdsourcing Platforms** (2013-now)
  - Platform Dynamics (WWW 15)
  - Pick-a-Crowd (WWW 13), Malicious Workers (CHI 15)
  - Scale-up Crowdsourcing (HCOMP 14), Scheduling (WWW 16)

Supported by: H2020, EPSRC, U Sheffield, Amazon

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*

Matthew Paige "Matt" Damon is an American actor, voice actor, screenwriter, producer, and philanthropist whose career was launched following the success of the drama film Good Will Hunting (1997) from a screenplay... wikipedia.org

Born: October 8, 1970 (age 43), Cambridge, Massachusetts, USA

Height: 5' 10" (1.78m)

Spouse: Luciana Barroso (m. 2005-present)

Partner: Winona Ryder (1998-2000)

Parents: Kent Damon, Nancy Carlsson-Paige

Children: Isabella Damon, Alexia Barroso, Gia Zavala Damon, Stella Damon

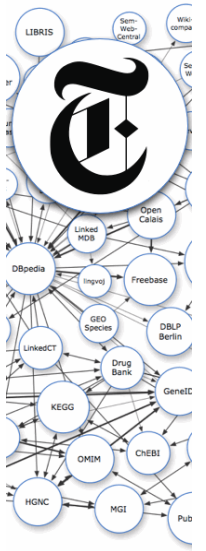
Movies & TV Shows

The Zero Theorem | Elysium | The Departed | We Bought a Zoo | Good Will Hunting

Feedback

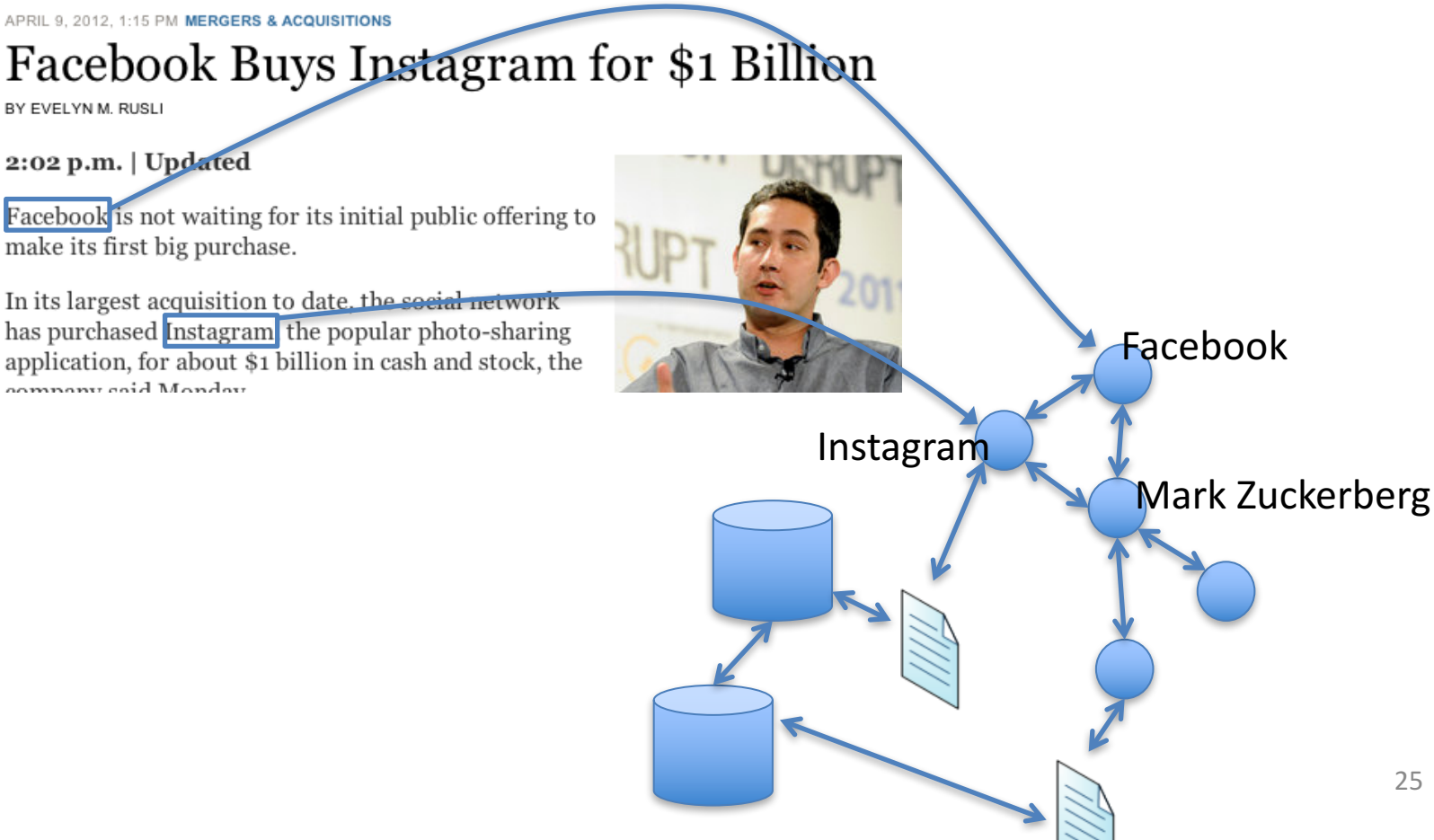
Knowledge Graphs

- 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)

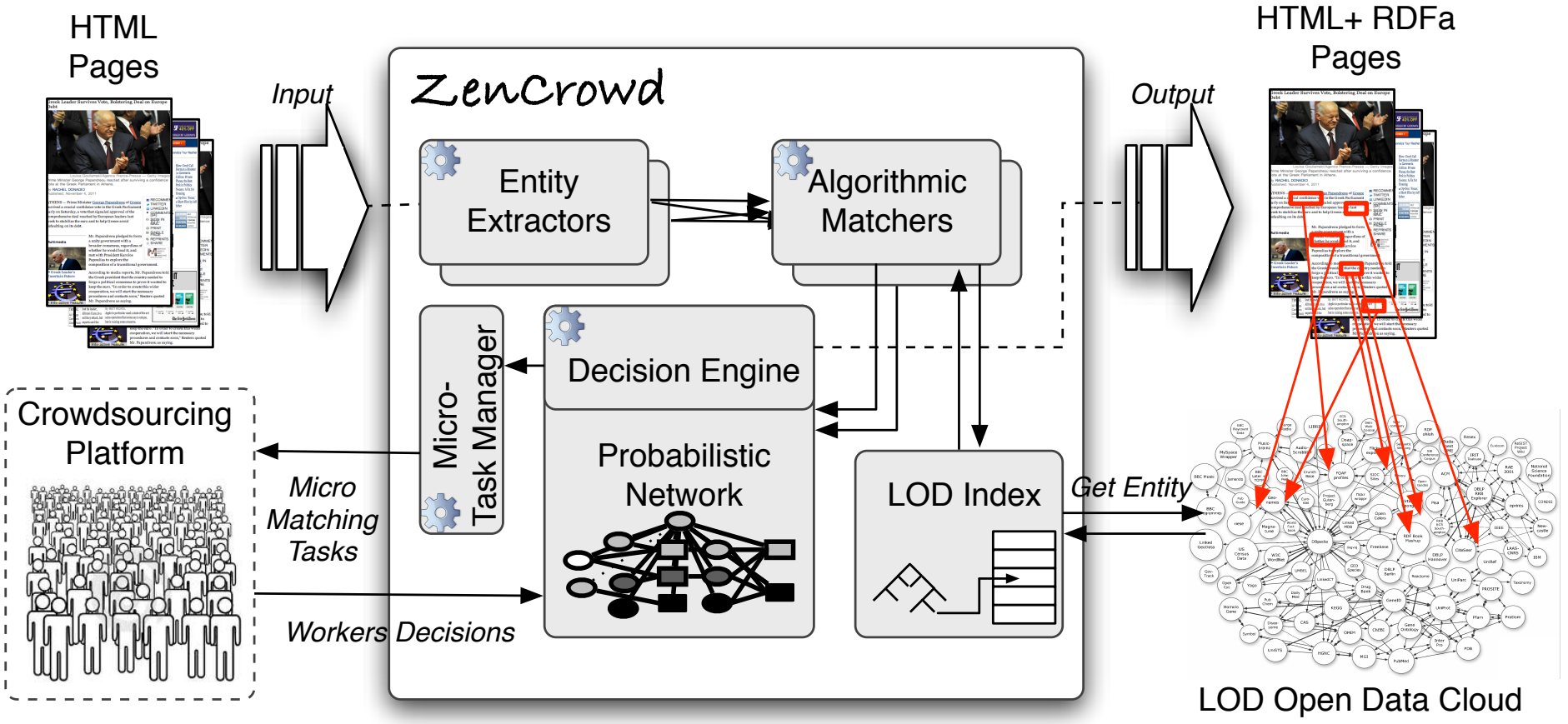


Entity Linking

- Looking at data integration across sources



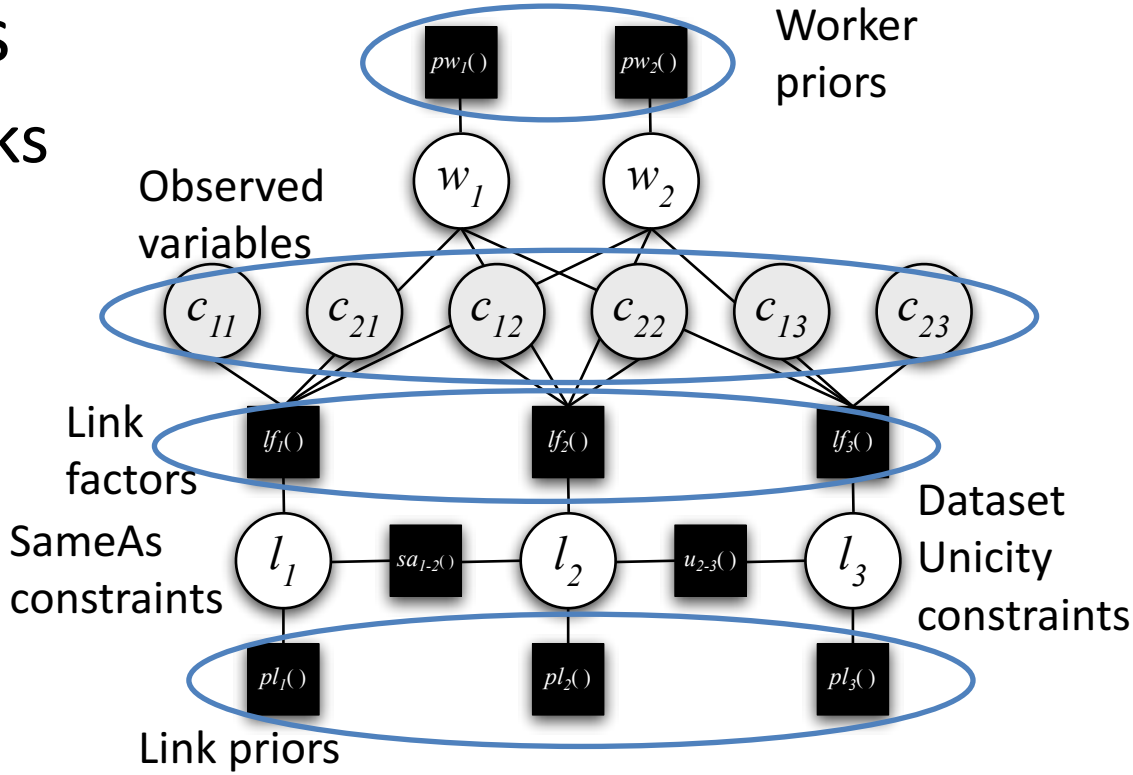
ZenCrowd Architecture



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

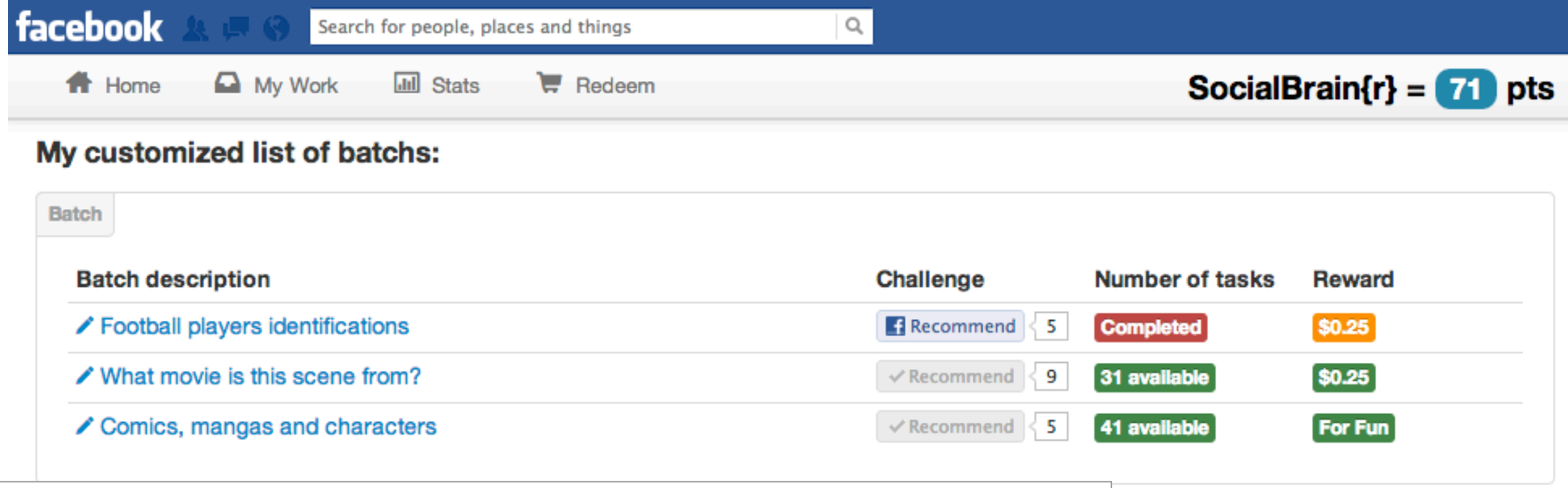
- Follow up-work (VLDBJ, 2013):
  - Also used for **instance matching** across datasets
  - 3-way blocking with the crowd

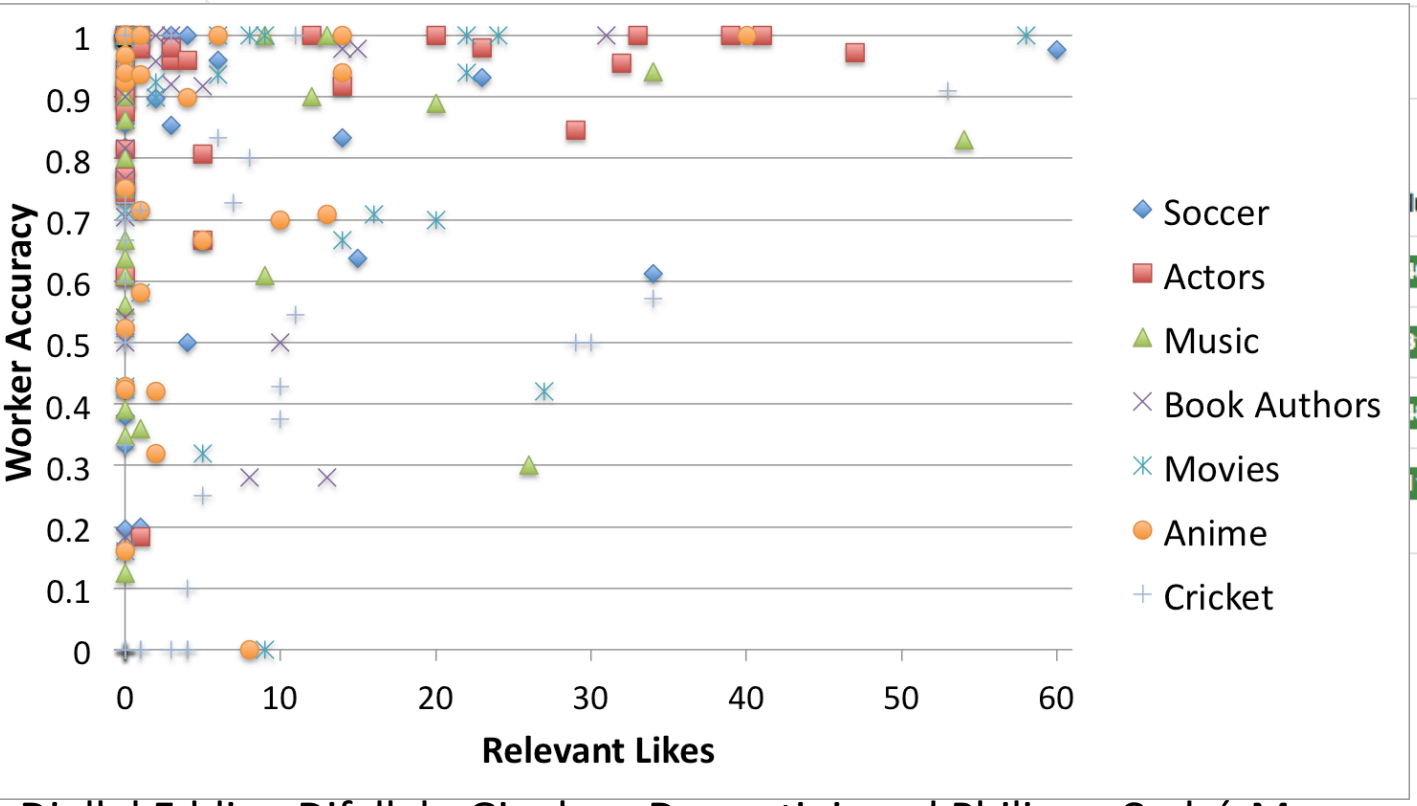
Hybrid Human-Machine Systems

- 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

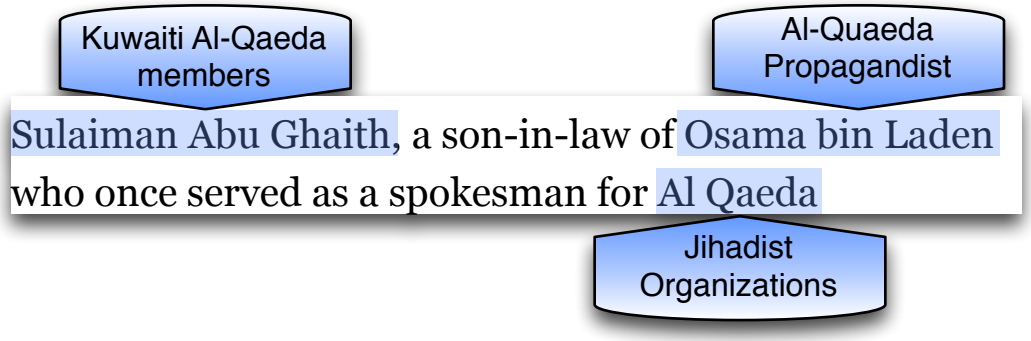
- Crowdsourcing + Prob reasoning works!
- But
  - Different worker communities perform differently
  - Many low quality workers
  - Completion time may vary (based on reward)
- Need to **find the right workers** for your task (see WWW2013 and CHI2015 papers)
- Need to make sure **high priority tasks**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





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

Contextual entity types in Web pages



Alberto Tonon, Michele Catasta, Roman Prokofyev, Gianluca Demartini, Karl Aberer, and Philippe Cudré-Mauroux. **Contextualized Ranking of Entity Types based on Knowledge Graphs**. In: Journal of Web Semantics, Volumes 37-38, Pages 170-183, Elsevier. March 2016.

Search into your browsing history

Michele Catasta, Alberto Tonon, Gianluca Demartini, Jean-Eudes Ranvier, Karl Aberer, and Philippe Cudré-Mauroux. B-hist: **Entity-Centric Search over Personal Web Browsing History**. In: Journal of Web Semantics, Elsevier. July 2014.

Summary

- **Hybrid human-machine systems** can
  - Scale over large amounts of data
  - Reach high accuracy by keeping humans in the loop
- Entities are the new entry point to Web content
  - "Things not string"
  - Google Knowledge Vault (but also Bing, Yahoo!, Yandex)
- Users can benefit from **entity-centric search**, browsing, and exploration of the Web

Gianluca Demartini. **Hybrid Human-Machine Information Systems: Challenges and Opportunities**. In: Computer Networks, vol 90, 5-13, Elsevier, 2015.

Conclusions

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

- There is need for **scalable** and **effective** data management infrastructures
  - Entity-centric approaches
  - Hybrid Human-Machine systems

www.gianlucademartini.net