The Power of Big Data

Gianluca Demartini
Lecturer in Data Science
University of Sheffield
gianlucademartini.net

Gianluca Demartini

- B.Sc., M.Sc. in CS at U. of Udine, IT
- Ph.D. in CS on Entity Retrieval at U. of Hannover
- Previously at eXascale Infolab U. Fribourg (CH), UC Berkeley (on Crowdsourcing), Yahoo! (ES), L3S Research Center (DE)
- **Lecturer in Data Science** at the iSchool, U. of Sheffield
  - Deputy programme coordinator for **MSc Data Science** since 2014
  - Module coordinator for Data Mining and Visualisation
- Tutorials on Entity Search at ECIR 2012 and RuSSIR 2015, on Crowdsourcing at ESWC13, ISWC13, and SearchSolutions 2015
- J Web Semantics editorial board, PC member of top conferences in IR, SemWeb

www.gianlucademartini.net

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!)

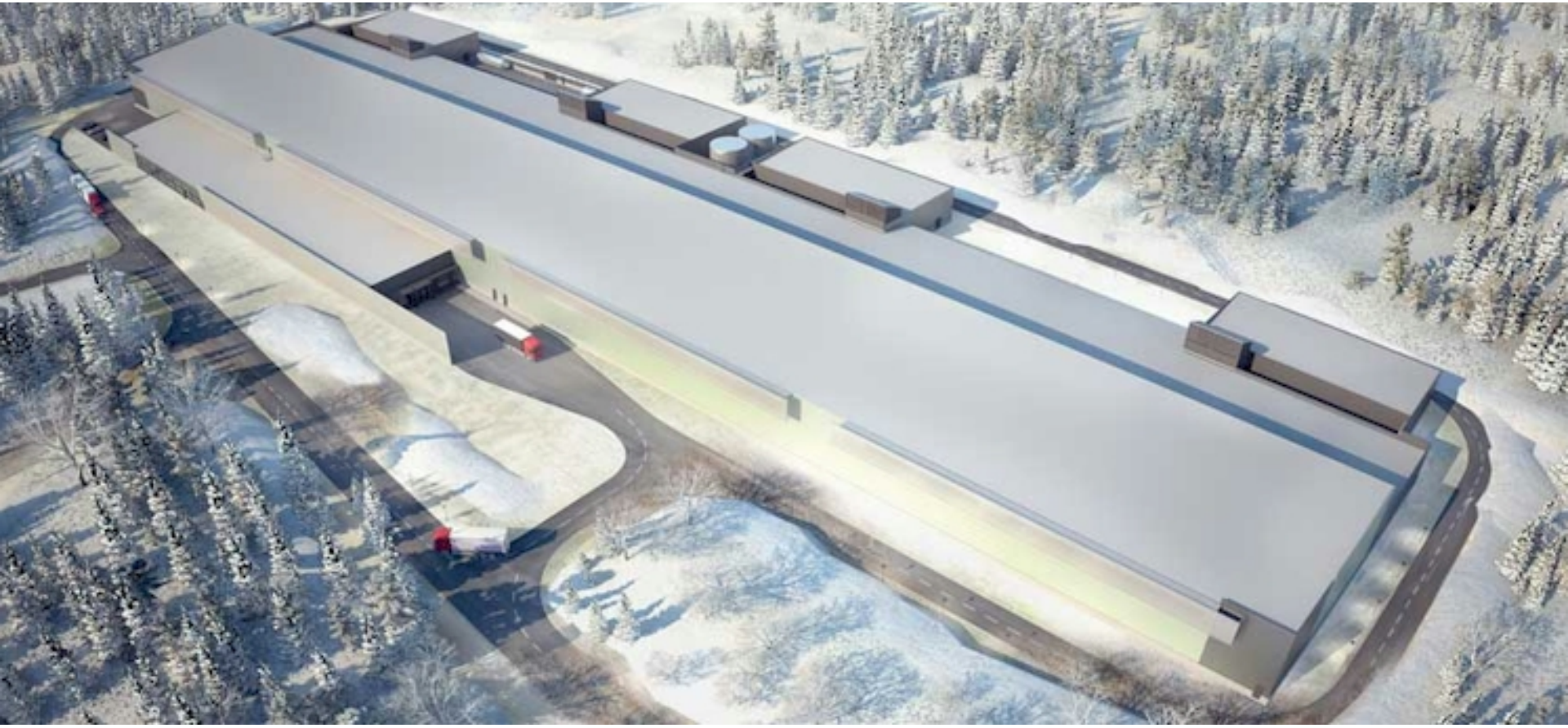
Data is fast (Velocity)

- 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, 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
- The flux-capacitor needs 1.21 GW (just 5xGoogle)

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

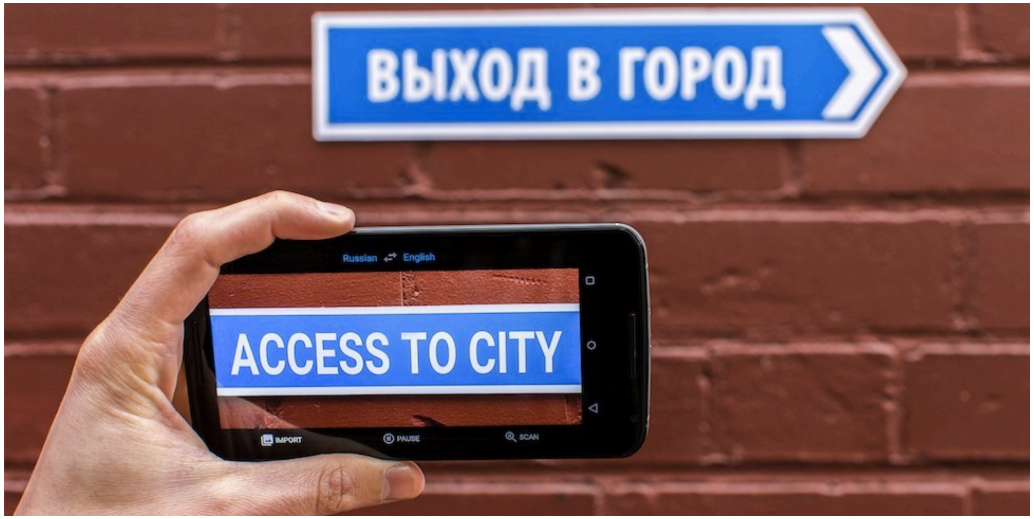
- Some data must not be processed by people!
  - GMail content is processed by computers to decide which advertisement you see on the Web

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



Big Data @ U Sheffield

- MSc Data Science
  - Type II data science who understand technology and data analytics but can also communicate
- How users interact with social media
  - Images
  - Viral content
- Entities as entry access to information
  - Web User support
  - Exploratory search

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

Big Data Research in Sheffield

tom cruise

Web News Images Videos Shopping More Search tools

About 133,000,000 results (0.52 seconds)

In the news

Tom Cruise's nightmare Scientology wedding: Leah Remini opens up about her "reprogramming" after the Cruise ...
Salon - 1 hour ago
Remini details a story of church officials inviting her **Tom Cruise's** home one night to teach ...

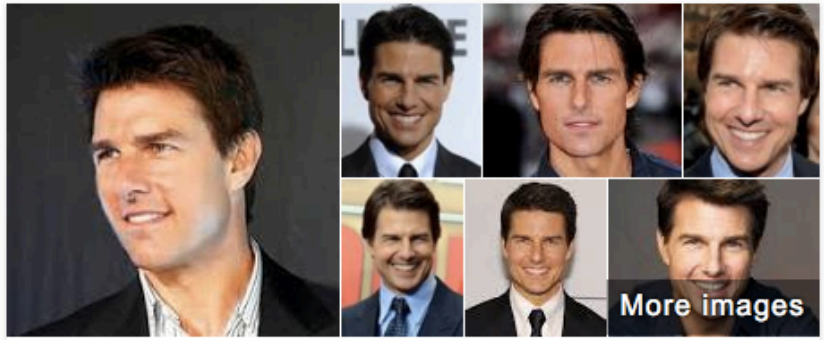
New Details: Leah Remini Makes Explosive Claims About Tom Cruise and His Family in Her Scientology Tell-All
People Magazine - 15 hours ago

Leah Remini reveals 'big kid' Tom Cruise's secret damning feud with John Travolta in explosive Scientology book
Mirror.co.uk - 8 hours ago

More news for tom cruise

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 - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Tom_Cruise**
Tom Cruise (born Thomas Cruise Mapother IV; July 3, 1962) is an American actor and filmmaker. Cruise has been nominated for three Academy Awards and ...
Tom Cruise filmography - Mimi Rogers - Katie Holmes - Nicole Kidman



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: 1.70 m

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

Children: Suri Cruise, Connor Antony Cruise, Isabella Jane Cruise

Parents: Thomas Mapother III, Mary Lee Pfeiffer

Movies

View 45+ more

Big Data Research in Sheffield

- Looking at data integration across sources

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

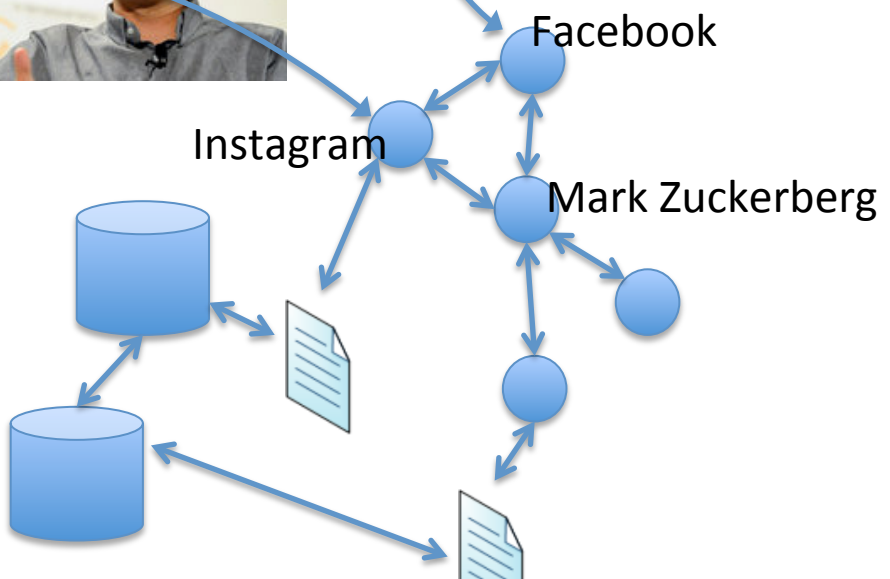
Facebook Buys Instagram for $1 Billion

BY EVELYN M. RUSLI

2:02 p.m. | Updated

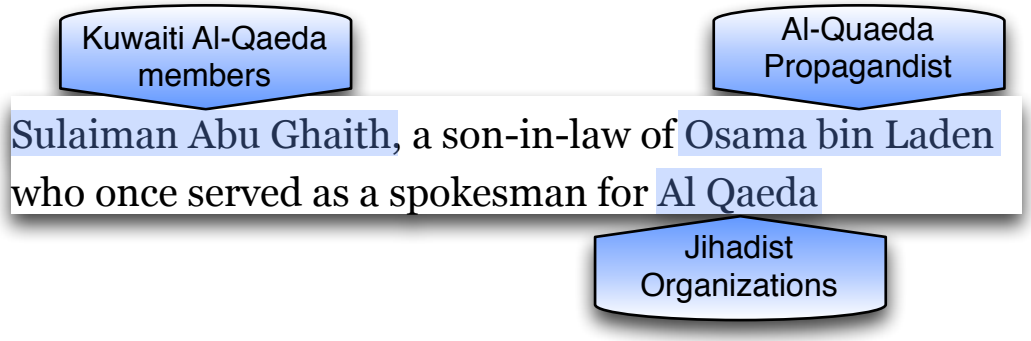
Facebook is not waiting for its initial public offering to make its first big purchase.

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.



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)

Contextual entity types in Web pages



Alberto Tonon, Michele Catasta, Gianluca Demartini, Philippe Cudré-Mauroux, and Karl Aberer. TRank: Ranking Entity Types Using the Web of Data. In: The 12th International Semantic Web Conference (ISWC 2013). Sydney, Australia, October 2013.

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.

Exploratory Search with Entities

- Project with The National Archive (2015-2018)
- Let users explore the UK Gov Web Archive
- Understanding user needs
- Providing novel user interfaces

Is your data ready?

- Decisions about you and your life will be driven by data!
- Are you in control of which data is being collected?
- Are you aware of how the data is stored and processed?
- Do you know how decisions are taken? Do you want to know?