

# Big Data Seminar

Lucas Drumond, Josif Grabocka

Information Systems and Machine Learning Lab (ISMLL)  
Institute of Computer Science  
University of Hildesheim, Germany

October 22, 2014

[illegible]

# What is Big Data?

Some definitions:

- ▶ “A collection of data sets so **large and complex** that it becomes difficult to process using on-hand database management tools or traditional data processing applications.”

[http://en.wikipedia.org/wiki/Big\\_data](http://en.wikipedia.org/wiki/Big_data)

# What is Big Data?

Some definitions:

- ▶ “A collection of data sets so **large and complex** that it becomes difficult to process using on-hand database management tools or traditional data processing applications.”  
[http://en.wikipedia.org/wiki/Big\\_data](http://en.wikipedia.org/wiki/Big_data)
- ▶ “Big data is **high-volume, high-velocity and high-variety** information assets that demand cost-effective, innovative forms of information processing for **enhanced insight and decision making.**”  
[www.gartner.com/it-glossary/big-data/](http://www.gartner.com/it-glossary/big-data/)

# What is Big Data?

Big Data is about:

- ▶ Storing and accessing large amounts of (unstructured) data

# What is Big Data?

Big Data is about:

- ▶ Storing and accessing large amounts of (unstructured) data
- ▶ Processing high volume data streams

# What is Big Data?

Big Data is about:

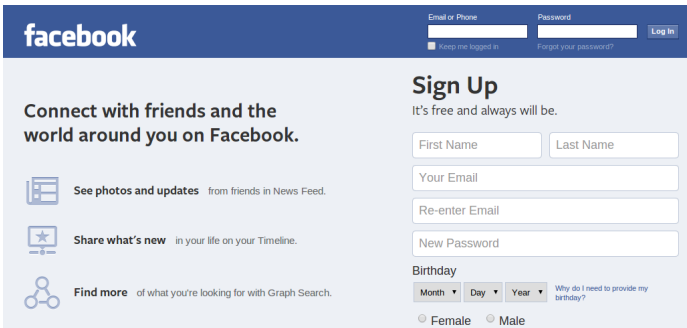
- ▶ Storing and accessing large amounts of (unstructured) data
- ▶ Processing high volume data streams
- ▶ Making sense of the data

# What is Big Data?

Big Data is about:

- ▶ Storing and accessing large amounts of (unstructured) data
- ▶ Processing high volume data streams
- ▶ Making sense of the data
- ▶ Predictive technologies

# Where to find Big Data?



The image shows the Facebook sign-up page. At the top, there's a blue header with the Facebook logo on the left and login fields on the right. The login fields include 'Email or Phone' and 'Password' with a 'Log In' button. Below the login fields are links for 'Keep me logged in' and 'Forgot your password?'. The main content area is divided into two columns. The left column has the heading 'Connect with friends and the world around you on Facebook.' followed by three features: 'See photos and updates' (with a photo icon), 'Share what's new' (with a star icon), and 'Find more' (with a search icon). The right column has the heading 'Sign Up' followed by the text 'It's free and always will be.' Below this are four input fields: 'First Name', 'Last Name', 'Your Email', and 'Re-enter Email'. There is also a 'New Password' field. Below the password fields is a 'Birthday' section with dropdown menus for 'Month', 'Day', and 'Year', and radio buttons for 'Female' and 'Male'. A small link 'Why do I need to provide my birthday?' is next to the birthday fields.

- ▶ 1.28 billion users (1.23 billion monthly active in January 2014)
- ▶ Size of user data stored by Facebook: 300 Petabytes
- ▶ Average amount of data that Facebook takes in daily: 600 terabytes
- ▶ Size of Facebook's Graph Search database: 700 Terabytes

Source: [http://allfacebook.com/orcfile\\_b130817](http://allfacebook.com/orcfile_b130817)

# Where to find Big Data?

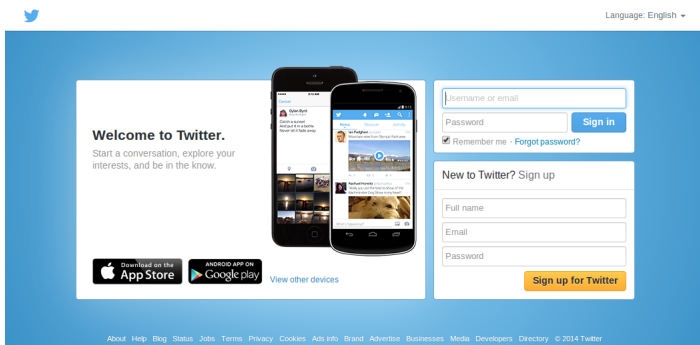


- ▶ 3.3 billion searches per day (on average)<sup>1</sup>
- ▶ 30 trillion unique URLs identified on the Web<sup>1</sup>
- ▶ 20 billion sites crawled a day<sup>1</sup>
- ▶ In 2008 Google processed more than 20 Petabytes of data per day<sup>2</sup>

<sup>1</sup><http://searchengineland.com/google-search-press-129925>

<sup>2</sup>Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), 107-113.

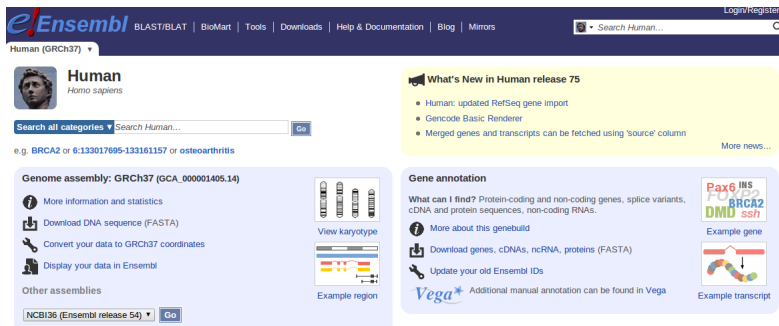
# Where to find Big Data?



- ▶ Average number of tweets per day: 58 million<sup>1</sup>
- ▶ Number of Twitter search engine queries every day: 2.1 billion<sup>1</sup>
- ▶ Total number of active registered Twitter users: 645,750,000<sup>1</sup>

<sup>1</sup><http://www.statisticbrain.com/twitter-statistics/>

# Where to find Big Data?



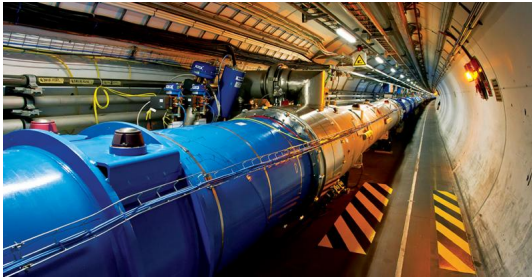
The screenshot shows the Ensembl genome browser interface. At the top, there's a navigation bar with links like BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors. A search bar is present with the text "Search Human...". Below the navigation bar, the main content area is divided into several sections:

- Human (GRCh37)**: A section with a search bar and a "Go" button. Below it, there's a link to "e.g. BRCA2 or 6:133017695-133161157 or osteoarthritis".
- Genome assembly: GRCh37 (GCA\_000001405.14)**: A section with links for "More information and statistics", "Download DNA sequence (FASTA)", "Convert your data to GRCh37 coordinates", and "Display your data in Ensembl". Below this, there's a section for "Other assemblies" with a dropdown menu showing "NCBI36 (Ensembl release 54)" and a "Go" button.
- What's New in Human release 75**: A section with a list of updates: "Human: updated RefSeq gene import", "Gencode Basic Renderer", and "Merged genes and transcripts can be fetched using 'source' column".
- Gene annotation**: A section with a list of features: "What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.", "More about this genebuild", "Download genes, cDNAs, ncRNA, proteins (FASTA)", and "Update your old Ensembl IDs". Below this, there's a link to "Vega" with the text "Additional manual annotation can be found in Vega".
- Example transcript**: A section with a diagram of a transcript and a link to "Example transcript".

- ▶ Ensembl database contains the genome of humans and 50 other species
- ▶ “only” 250 GB<sup>1</sup>

<sup>1</sup><http://www.ensembl.org/>

# Where to find Big Data?



- ▶ Large Hadron Collider has collected data from over 300 trillion proton-proton collisions
- ▶ Approx. 25 Petabytes per year

# Overview

*Part III***Machine Learning Algorithms***Part II***Large Scale Computational Models***Part I***Distributed Database****Distributed File System**

# The rules of selecting a paper:

- 1: Students visit the course website and select a paper under the Section literature (deadline: 29.10).
- 2: The selected paper is notified to [ldrumond@ismll.de](mailto:ldrumond@ismll.de) and [josif@ismll.de](mailto:josif@ismll.de)
  - ▶ Deadline: 29.10
  - ▶ First come, first served
  - ▶ Send three preferred papers to avoid allocation crashes
- 3: The instructors create a schedule for the talks and notify the students. The first talk is scheduled for 12.11.

# Papers list: Part I

Author	Title	Year
Ahmed, N.K. et al.	Graph Sample and Hold: A Framework for Big-graph Analytics	2014
Dean, T. et al.	Fast, Accurate Detection of 100,000 Object Classes on a Single Machine	2013
Dong, X. et al.	Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion	2014
Gonzalez, J.E. et al.	PowerGraph: Distributed Graph-parallel Computation on Natural Graphs	2012
Han, W.-S. et al.	TurboGraph: A Fast Parallel Graph Engine Handling Billion-scale Graphs in a Single PC	2013
Liu, C. et al.	Distributed Nonnegative Matrix Factorization for Web-scale Dyadic Data Analysis on MapReduce	2010

[http://www.ismll.uni-hildesheim.de/lehre/semBI-14w/index\\_en.html](http://www.ismll.uni-hildesheim.de/lehre/semBI-14w/index_en.html)

# Papers list: Part II

Author	Title	Year
Ottaviano, G., Venturini, R.	Partitioned Elias-Fano Indexes	2014
Rakthanmanon, T. et al.	Searching and Mining Trillions of Time Series Subsequences Under Dynamic Time Warping	2012
Recht, B. et al.	Hogwild: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent	2011
Yu, H.-F. et al.	Scalable Coordinate Descent Approaches to Parallel Matrix Factorization for Recommender Systems	2012

[http://www.ismll.uni-hildesheim.de/lehre/semBI-14w/index\\_en.html](http://www.ismll.uni-hildesheim.de/lehre/semBI-14w/index_en.html)

# Regulations of the presentations:

- ▶ Depending on the number of students, there will be one or two seminar presentations per lecture schedule.
- ▶ Each seminar lasts for 50 minutes, including 10 minutes of questions and discussions.
- ▶ All the students should participate in the talks of others.

# Advice on the presentation

- ▶ Understand and describe the underlying theoretic foundation of the methodologies (learning algorithms, equations)
- ▶ Describe the methods in your own formulation and avoid reading out the content of the paper
- ▶ Think analytically and describe the advantages and disadvantages of the paper
- ▶ If applicable, propose ideas and improvements in the end

# Seminar Report

- ▶ Every presenter should prepare a report on the paper he presented.
- ▶ The report should include a description of the method, its strengths and weaknesses
- ▶ The overall tone of the report should be analytic of the work and not a repetition of the paper
- ▶ Additional ideas, experiments or illustrations will be rewarded

# Structure of the Seminar Report

- ▶ Content should not exceed 30 pages
- ▶ Submission deadline, 2 weeks before the term break (28.01.2015).
- ▶ To be submitted (to Lucas Drumond C36Spl):
  - ▶ 3 printed and bound copies
  - ▶ 1 CD with the report, source code and all relevant materials

Any Questions?