Advanced Topics in Information Science
CSCI 6720G

Data Management and Big Data Integration

Jarek Szlichta

Data Management Research Group

Agenda for Today

- Big Data
- Course Structure
  - Presentation
  - One pager
  - Class project
  - Mark breakdown
- Course Outline

The Era of Big Data

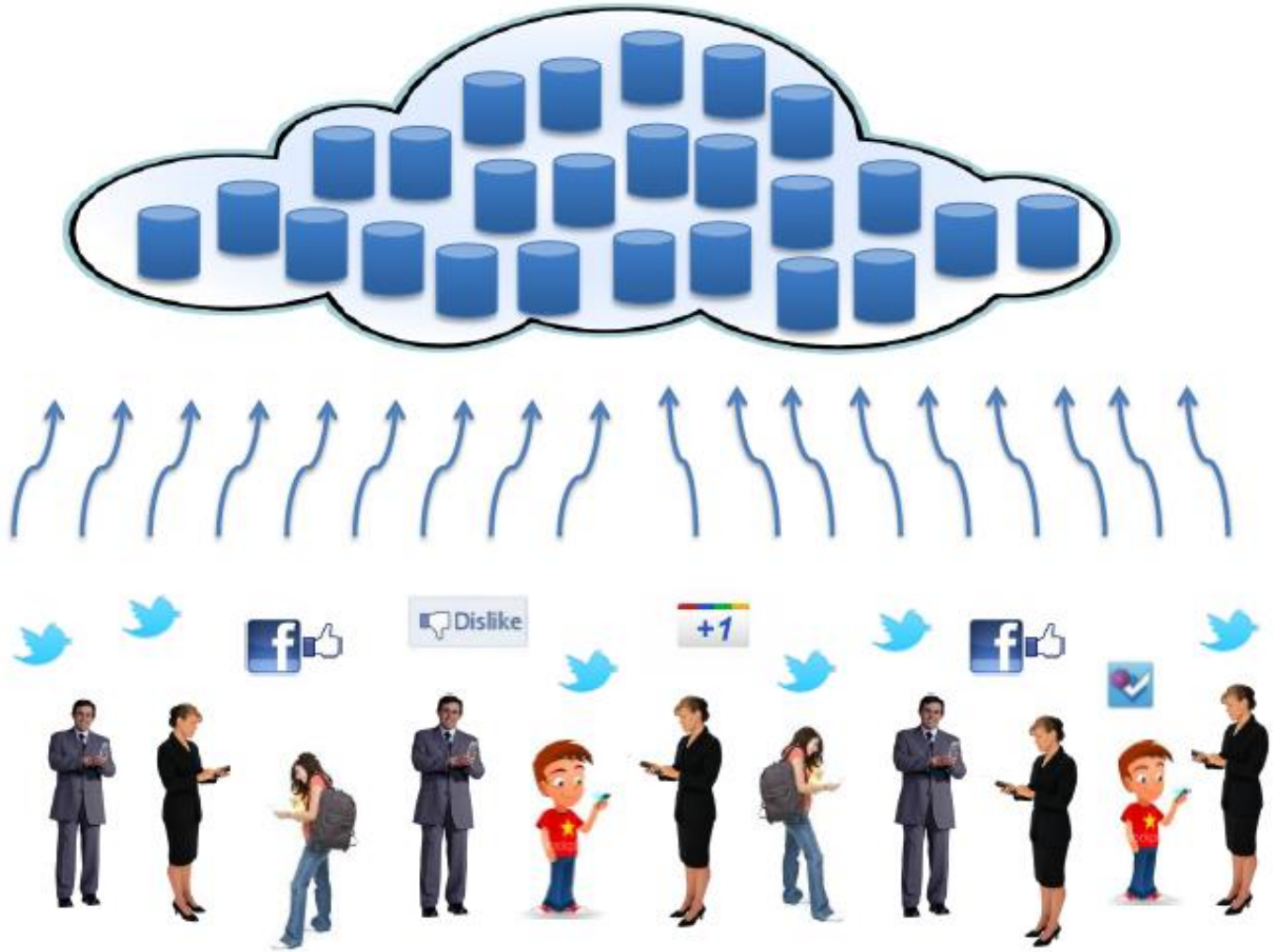
- Unprecedented growth in data being generated and its potential uses/value [TPCTC'12-MC]
  - Tweets, social networks (statuses, check-ins, shared content), blogs, click streams, various logs, …
  - *Facebook: > 845M active users, > 8B messages/day*
  - *Twitter: > 140M active users, > 340M tweets/day*
- Everyone is interested
  - Trade press and popular press: *"Big Data!"*
  - Enterprises, Web companies, online businesses, governments, public health researchers, social scientists, …
  - Untapped value and countless new opportunities to understand, optimize, and/or compete

[TPCTC'12-MC] - Mike Carrey's keynote at TPCTC'12

Social Networks - Facebook

- Daily user activity
  - 2.5B content items shared
  - 2.7B ‘Likes’
  - 350M photos uploaded
- Data volume statistics
  - 100+ PB in a single HDFS cluster
  - 500+ TB of new data generated per day
  - 70K queries per day
  - 105 TB scanned per half-hour (Hive)

Data Integration



Facebook Country

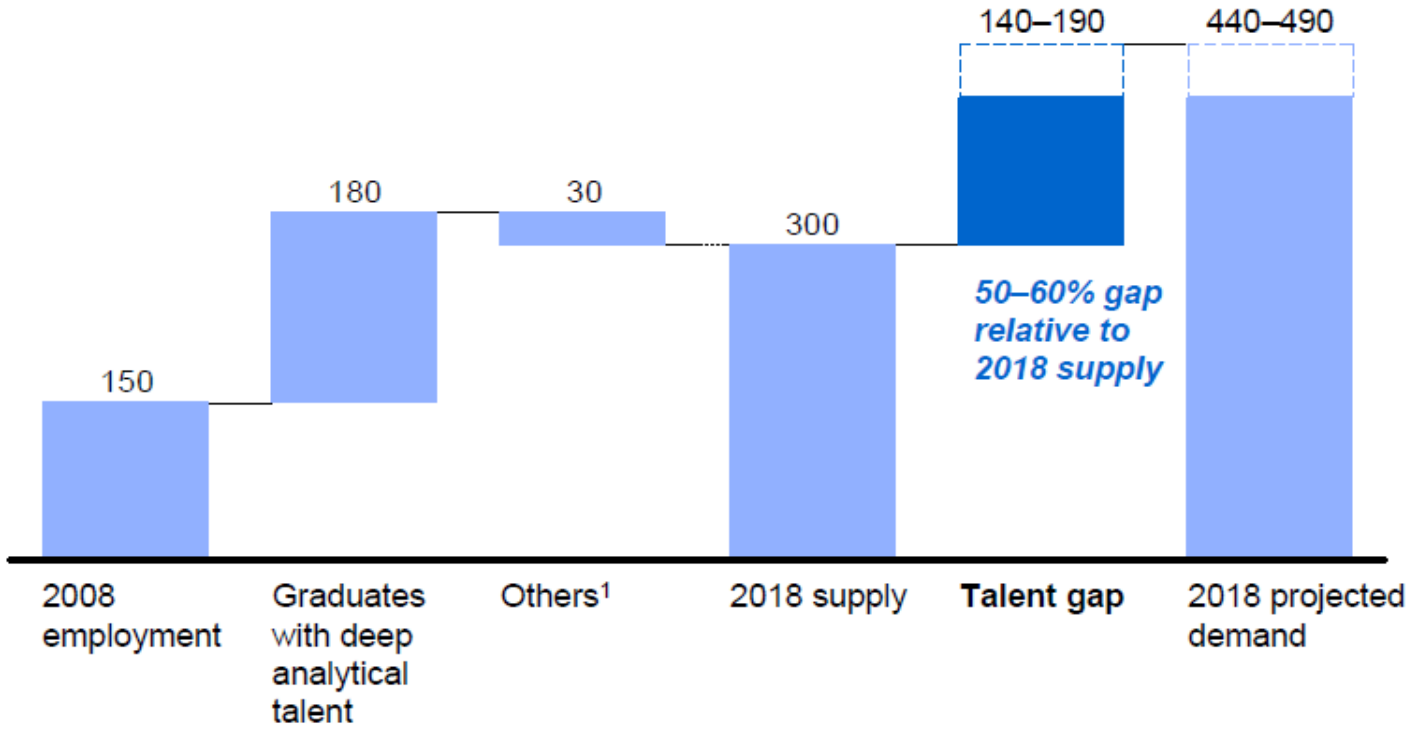
1. China (1.339 billion)
2. India (1.218 billion)
3. Facebook (900 million)

Data contains value and knowledge

Good news: Demand for Big Data

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018
Thousand people



1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).
SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

Course Structure

- Presentation & Discussion
  - Each week 2 presentations
  - Recent research papers on
    - Big data integration (data integration, data analytics, data curation, data cleaning, web search, …)
- Class project
  - 1 person per team..
  - Related to the course topic

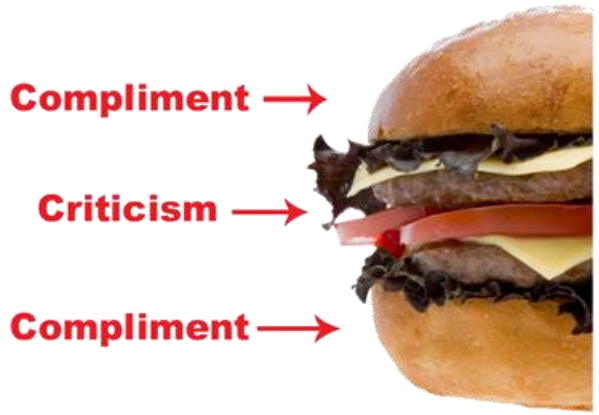
Presentation

- Everyone should read the paper before the lecture!
- Presentation time: 40 minutes
- Make your own slides
- In general your slides should:
  - be easy to read,
  - avoid too much text!
  - use **figures**, **examples** ease understanding!
  - not be too long. Usually you will not need more than one slide per minute. As a rule of thumb you will need 1.5-2 minutes per slide
  - have a clear and easy to follow structure
  - practice your talk before presenting at class

Discussion

- After each presentation
  - Led by the presenter
- Prepare your questions

- Give feedback to the presentation
  - Feedback Burger
    - Basis: Positive details
    - Meat: What could be improved
    - Top: Conclusion / Motivation



One Pager

- **One** page summary of one paper per 2 weeks
  - Typed, 11-12pt font
- It should:
  - summarize the problem(s) addressed/solved by the research paper (1-2 sentences that clearly describe the problem: "The problem is ... ."
  - briefly sketch the main ideas on which the solution of the problem is based
    - briefly describe the research methodology of the paper (1-2 sentences)
  - identify 3 strong points and 3 weak points of the paper
  - summarize any assumptions the solution in the paper is based upon (any restrictions; stated assumptions and non-stated assumptions)
  - raise three non-trivial questions about the paper (including future work)
  - other remarks (if any)
- Is marked

Class Project

- Independent exploration of specific problem
  - within the course topics
  - implementation and performance measurement
  - modeling, design, simulation, and analysis
  - experimentation and evaluation
- Evaluation of project:
  - timeliness, development and presentation of idea (i.e., proposal, progress, and final project report, and final in class presentation).
- No more than 12 pages (LNCS template)
- Project agenda
  - Proposal – submit online
  - Progress report – submit online
  - Final presentation – in-class
  - Final report – conference style paper

Mark Breakdown

- 15% One pagers
- 30% Presentations and discussion leading
- 20% Participation and interactions (discussion, readings, ideas etc.)
- 35% Course project (proposal, progress reports, presentation, final report)

Short Term Agenda

- Next week presentation – some of my research and discussions
- The week after we will start with your presentations
  - First data integration and data curation
- Course Website: data.science.uoit.ca
- Questions
  - jaroslaw.szlichta@uoit.ca

Paper Assignments

- Sigmod, VLDB, ICDE and EDBT proceedings, e.g.,
  - http://www.vldb.org/pvldb/vol9.html
  - http://dl.acm.org/citation.cfm?id=2882903
  - http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7491900
  - http://openproceedings.org/html/pages/2016_edbt.html
- Or individual researchers DBLP, e.g.,
  - http://dblp.uni-trier.de/pers/hd/s/Srivastava:Divesh
  - http://dblp.uni-trier.de/pers/hd/g/Golab:Lukasz

Sample Paper Assignments

- Fei Chiang, Renée J. Miller: A unified model for data and constraint repair. ICDE 2011: 446-457
- George Beskales, Ihab F. Ilyas, Lukasz Golab: Sampling the Repairs of Functional Dependency Violations under Hard Constraints. PVLDB 3(1): 197-207 (2010)
- Mohamed Yakout, Ahmed K. Elmagarmid, Jennifer Neville, Mourad Ouzzani, Ihab F. Ilyas: Guided data repair. PVLDB 4(5): 279-289 (2011)
- Laure Berti-Equille, Tamraparni Dasu, Divesh Srivastava: Discovery of complex glitch patterns: A novel approach to Quantitative Data Cleaning. ICDE 2011: 733-744
- Lukasz Golab, Howard J. Karloff, Flip Korn, Avishek Saha, Divesh Srivastava: Sequential Dependencies. PVLDB 2(1): 574-585 (2009)