Topics in Foundations of Databases

Beijing, July-August 2017
Prof L Libkin (Univ of Edinburgh)
• Course information:

• http://homepages.inf.ed.ac.uk/libkin/teach/beijing2017/

• Easy way to find: Google/Yahoo for me (http://homepages.inf.ed.ac.uk/libkin/)

• Follow “teaching link”

• from there follow link to Beijing 2017
Foundations of big data

- Big data = **data wrangling** (databases) + analytics (mining, machine learning)
  - wrangling: 50% to 80% of the entire effort
  - wrangling = extraction, integration, cleaning, querying, exchanging, etc - everything that is needed to enable analytics
  - Main tools and techniques come from databases
Data scientists’ favorite tools

Tools
Languages, Data Platforms, Analytics

Share of Respondents


Tool: language, data platform, analytics
Future data scientists’ favorite tools

The Most In-Demand Skills for Data Scientists in 2016

- SQL
- Hadoop
- Python
- Java
- R
- Hive
- Mapreduce
- NoSQL
- Pig
- SAS
- C
- Oracle
- Teradata
- SPSS
- Matlab
- Perl
- MySQL
- PostgreSQL
- Ruby
- HTML
- Stata
4Vs of Big Data

- **Volume** (big data is big)
- **Variety** (big data appears in different formats)
- **Veracity** (big data is incomplete)
- **Velocity** (big data changes a lot)
Volume

• Evaluation and optimisation of queries

• Joins/conjunctive queries: ubiquitous in databases
  • review basics: minimization, constraints, chase, acyclic queries

• Approximation of queries

• Queries that scale regardless of database size
  • scale-independence
Variety (1)

• XML data: navigation and query languages
  • how to evaluate efficiently
  • connections with automata and language theory

• Graph databases:
  • navigation, property graphs, languages

• RDF: relational and graph database perspectives
Variety (2)

- Data integration
  - how to answer queries over data pulled from different sources
- Data exchange
  - how to restructure data under a different schema and query it
Veracity

- Incomplete information
  - we still don’t fully understand what SQL does
  - new approximations of correct answers
- Inconsistent data
  - when cleaning is impossible, what can be done?
- Ranking preferences (top-k algorithms)
Suggested projects

• List of papers, divided by topics

• Basic idea: pick a paper, write an essay

• Essay:
  
  • 5-7 pages, explaining the paper in a way that someone who has not read it can understand

  • plus: your own ideas/thoughts

  • examples are: extensions of results, possible special cases that may give better complexity, a quick prototype implementation to see real-life performance of theoretical algorithms etc

  • if you pick a paper among those discussed in class, your original thoughts need to be deeper