Data wrangling, data quality

Leonid Libkin *Pierre Senellart*

3 December 2019

Ideal vs actual world

Ideal world for a data scientist:

- A single dataset, with a fixed, simple structure (e.g., one table with features and label)
- Structured data
- Exact, complete information
- Precise values, certain data

Ideal vs actual world

Ideal world for a data scientist:

- A single dataset, with a fixed, simple structure (e.g., one table with features and label)
- Structured data
- Exact, complete information
- Precise values, certain data

Actual world:

- Many datasets to be combined, with different structures and schemas
- Plain text, semi-structured data
- Duplicated information, missing information
- Imprecise values, uncertain data

Data wrangling, data quality

- How to wrangle real-world data and turn it into a nice structured form?
- How to assess the quality of data?
- How to deal with missing, imprecise, duplicated data?
- How to keep track of where data comes from?
- How to do all of this efficiently?

Curriculum and provisional schedule

- Classes on Wednesdays afternoon (in conflict with the course on *Ethics and AI* ☹)
- 8 sessions + project defenses

  22/01 Information extraction (Pierre)
  29/01 Data cleaning, data deduplication (Pierre)
  05/02 Data integration, view-based query answering (Leonid)
  12/02 Provenance management (Pierre)
  26/02 Data exchange (Leonid)
  04/03 Probabilistic databases (Pierre)
  11/03 Incomplete information in databases (Leonid)
  18/03 Approximate query answering (Leonid)
  25/03 Project defenses

Evaluation

One homework (40% of the total grade): take one research paper, summarize it, explain it in your own words, comment on its strengths and limitations

One project (60% of the total grade): take one (or more) research paper, build something cool from it (implement it, improve the algorithm, test it on some interesting dataset, etc.), present it in a defense