Data wrangling, data quality

Leonid Libkin Liat Peterfreund *Pierre Senellart*

7 December 2022

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?
- What kind of (database) model and query to be used?
- How to assess the quality of data?
- How to deal with missing, imprecise, duplicated data?
- How to do all of this efficiently?

Curriculum and provisional schedule

- Classes on Friday morning
- 8 sessions + project defenses (TBC)
- Topics covered:
  - Joins and conjunctive queries
  - Web content acquisition and information extraction
  - Spanners: querying extracted information
  - Data integration and data exchange
  - Data cleaning and data deduplication
  - Incomplete information
  - Advanced querying of graph data
  - Probabilistic databases

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