

Data wrangling, data quality

Leonid Libkin Pierre Senellart



7 December 2021

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 **Monday morning** (in conflict with the course on *Computational social choice* 😊)
- 8 sessions + project defenses
- Topics covered:
 - Information extraction
 - Data cleaning, data deduplication
 - Data integration, view-based query answering
 - Provenance management
 - Data exchange
 - Probabilistic databases
 - Incomplete information in databases
 - Approximate query answering

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