Scalable Data Analysis (CIS 602-02)

Data Integration

Dr. David Koop
Wrangler

- Data cleaning takes a lot of **time** and **human effort**
- "Tedium is the message"
- Repeating this process on multiple data sets is even worse!
- Solution:
  - interactive interface (mixed-initiative)
  - transformation language with natural language "translations"
  - suggestions + "programming by demonstration"
Potter's Wheel: Example

Format

[(.*)], (.*), to \2\1

Split at ' '

2 Merges

[V. Raman and J. Hellerstein, 2001]
Wrangler Interface

• Automated Transformation Suggestions
• Editable Natural Language Explanations

  ▶ Fill Bangladesh by copying values from above
  ▶ Fill Bangladesh by averaging the 5 values from above
  ▶ Fill Bangladesh by averaging the 5 values from above

• Visual Transformation Previews

• Transformation History

[S. Kandel et al., 2011]
Wrangler Evaluation

• Compare with Excel
• Tests:
  - Extract text from a single string entry
  - Fill in missing values with estimates
  - Reshape tables
• Allowed users to ask questions about Excel, not Wrangler
• Found significant effect of tool and users found previews and suggestions helpful
• Complaint: No manual fallback, make implications of user choices more obvious for users
Pandas Demo

Reading Presentations

• Tuesday (Visualization):
  - Chaitanya Chandurkar and Ramya Reddy Mara

• Others: Network down, need to check time submitted

• Remember, you do not have to do a reading response if you are presenting
Reading Responses

• Planning to grade before the next class
• Since you have not received feedback, the scoring will be lenient
• That changes once feedback is provided
• Remember to emphasize your understanding of the material and focus on the critique
Projects

• Options:
  - Data analysis on some existing data: think about the questions you want to try to answer
  - Improve some technique for data analysis

• Data Sources:
  - Search the web for topics you're interested in
  - [https://github.com/caesar0301/awesome-public-datasets](https://github.com/caesar0301/awesome-public-datasets)
  - Local data

• If you are doing a research project in a particular area, let's try to work something out so that the course project relates
Data Integration

• Days of maintaining a closed system with only "your information" are gone
• Want to do analysis that brings in data from multiple sources
• How?
Introduction to Data Integration

A. Doan, A. Halevy, Z. Ives
Principles of Data Integration, 2012

http://research.cs.wisc.edu/dibook/slides/Chapter_1.ppt
Data Integration Summary

• Lots of data sources, how do we answer questions where we need to access data from more than one?

• Problem of heterogeneity

• AI-Complete problem: difficulty is the same as making computers as intelligent as people

• Two techniques:
  - Mediation
  - Data Warehouses
Dataspaces

• Mediation and warehouses both require an agreed-upon schema
• You can't use data unless you know how it maps to some global schema
• What if we just dump data into a "lake" and build methods to answer queries on-demand?
Dataspaces

M. Franklin, A. Halevy, D. Maier, J. Widom

http://db.cs.berkeley.edu/dblunch-fa2005/alon.ppt
Dataspaces

- What does search/query mean in a dataspace?
- Uncertainty
- Reuse human effort whenever possible
iTrails: Pay-as-you-go Information Integration in Dataspaces

M. V. Salles, J. Dittrich, S. Karakashian, O. Girard, L. Blunschi

iTrails

• No schema vs. Schema first vs. Pay-as-you-go
• Graph model representation
• Uncertainty?
• More general processing than trails?
Reminders

- Reading Presentations
- Reading Responses
- Project Ideas