CX4242:

Data Cleaning

Mahdi Roozbahani
Lecturer, Computational Science and Engineering, Georgia Tech
Data Cleaning
How dirty is real data?
How dirty is real data?

Examples

• Jan 19, 2016
• January 19, 16
• 1/19/16
• 2006-01-19
• 19/1/16

How dirty is real data?

Discuss with your neighbors (group of 2-3) for 2 minutes.

This topic is in connection with 5+ kinds of “data dirtiness”.
How dirty is real data?

- Non-standardized naming
- Date format
- Human mistake/ typos
- Cultural differences
- Missing data
- Duplicates
- Outliers
- Machine failure
- White spaces/ tab/ indent
How dirty is real data?

- Missing or corrupted (NaN, null)
- Numbers stored as string ("1232")
- Different units
- Spelling/typos
- Different string encodings
- Outliers (due to data recording)
- geocoding, timezone offsets (missing +, -)
- Duplicate data
- Fake data (malicious)
- Sql injection
- Different software version generating slightly different formats
- Cap locks
- Semi-colons
- Structure (json objects)
- Invisible characters
- Different delimiters
- Indentation
Importance of Data Cleaning
“80%” Time Spent on Data Preparation

Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says [Forbes]

What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%
Data Janitor
Writing “Clean Code”

- Be careful with trailing whitespaces
- Indent code (spaces vs tabs) following coding practices in your team/company

https://google.github.io/styleguide/javaguide.html#s4.2-block-indentation

...there’s no way I'm going to be with someone who uses spaces over tabs...


Trailing whitespace is evil. Don't commit evil into your repo.

http://codeimpossible.com/2012/04/02/trailing-whitespace-is-evil-don-t-commit-evil-into-your-repo/
Both available **free** for GT students on

[http://safaribooksonline.com/](http://safaribooksonline.com/)
Data Cleaners

Watch videos

• Data Wrangler (research at Stanford)
• Open Refine (previously Google Refine)

Write down

• Examples of data dirtiness
• Tool’s features demo-ed (or that you like)

Will collectively summarize similarities and differences afterwards

Open Refine: http://openrefine.org
Data Wrangler: http://vis.stanford.edu/wrangler/
Wrangler is an interactive tool for data cleaning and transformation. Spend less time formatting and more time analyzing your data.

UPDATE: The Wrangler research project is complete, and the software is no longer actively supported. The team behind Wrangler has moved on to work on a commercial venture, Trifacta.

Why wrangle?

- Too much time is spent manipulating data just to get analysis and visualization tools to read it. Wrangler is designed to accelerate this process: spend less time fighting with your data and more time learning from it.
- Wrangler allows interactive transformation of messy, real-world data into the data tables analysis tools expect. Export data for use in Excel, R, Tableau, Protovis, ...
- Want to learn more about Wrangler's design? Take a look at our research paper.
- Wrangler is still a work-in-progress. Please share your feedback and feature requests!

TRY IT NOW
Welcome!

OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

OpenRefine is available in English, Chinese, Spanish, French, Russian, Portuguese (Brazil), German, Japanese, Italian, Hungarian, Hebrew, Filipino, Cebuano, Tagalog

OpenRefine is supported by:

Google News Initiative

Introduction to OpenRefine

1. Explore Data

OpenRefine can help you explore large data sets with ease. You can find out more about this functionality by watching the video below and going through these articles.
What can Open Refine and Wrangler do?

O = Open Refine  W = Data wrangler

- [w] well structured formatting at the beginning
- [w,o] redo and undo
- [o] More features like statistical analysis
- [w,o] generating a programming language output
- [w] it will give you suggestions
What can Open Refine and Wrangler do?

- [w,o] undo, redo
- [o,w] history of data
- [o] transform data (e.g., take log)
- [w] data editing/highlighting/interaction may be easier
- [o] clustering
- [w] transpose/pivot
- [w] fill in missing data
- [w] suggestions + preview

O = Open Refine
W = Data wrangler
How do they compare?

• Similarities
  • work directly on data
  • provide visual feedback
  • browser-based
  • can only handle common use cases(?)
  • free!!!
  • undo/redo, history (people make mistakes)
  • input: plain text

G = Google Refine
W = Data wrangler
How do they compare?

• Differences
  • W generates transform code
  • G recognizes clusters
  • W gives natural language suggestions
  • G works offline (your sensitive data stay with you)
  • G has more sophisticated functions?
  • W seems to be able to transform overall data format
  • W supports expression syntax (e.g., log())
  • G more scalable(?)

G = Google Refine
W = Data wrangler
The videos only show *some* of the tools’ features. Try them out.

Open Refine: [http://openrefine.org](http://openrefine.org)
Data Wrangler: [http://vis.stanford.edu/wrangler/](http://vis.stanford.edu/wrangler/)