http://poloclub.gatech.edu/cse6242

CSE6242 / CX4242: Data & Visual Analytics

Data Cleaning

Duen Horng (Polo) Chau
Assistant Professor
Associate Director, MS Analytics
Georgia Tech

Partly based on materials by
Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos

Data Cleaning

Why data can be dirty?

How dirty is real data?

Examples

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

http://blogs.verdantis.com/wp-content/uploads/2015/02/Data-cleansing.jpg

How dirty is real data?

Discuss with you neighbors (group of 2-3)

90 seconds

Comes up with **5+ kinds of “data dirtiness”**

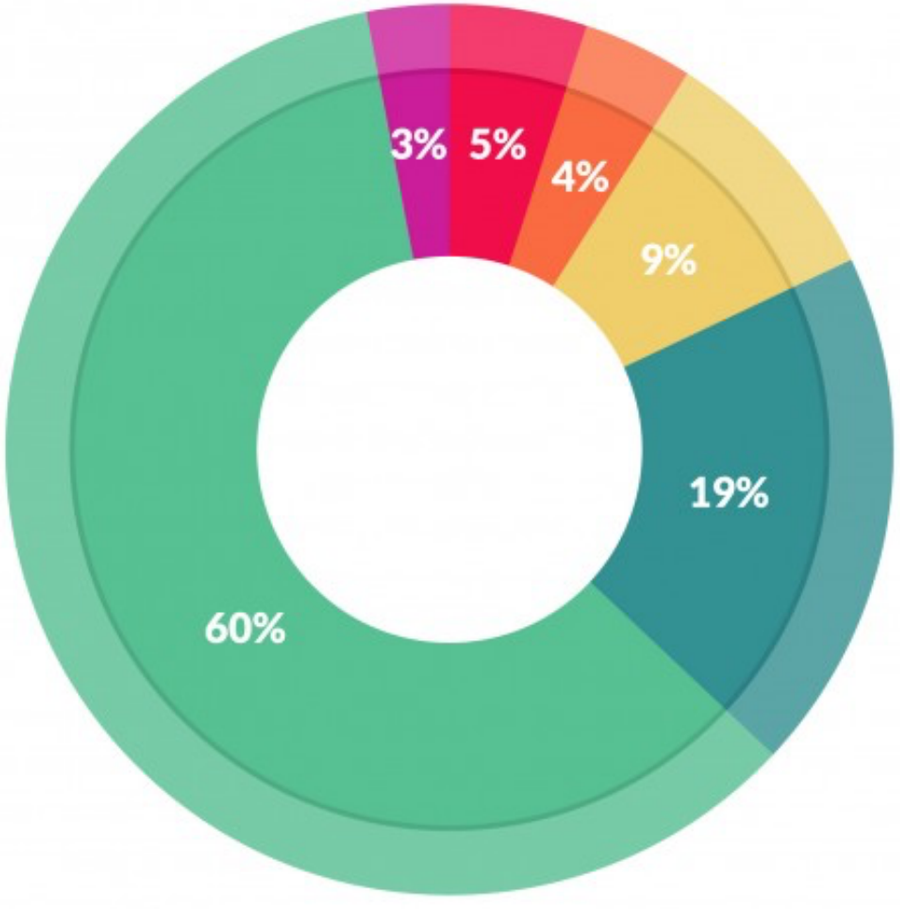
How dirty is real data?

Examples

- unbalanced/“outliers”
- leading zeros…
- different units/measurements (pounds…)
- missing data
- spelling errors
- wrong data types
- cases lower/upper
- data in the wrong place (shifted somehow)
- file format
- inconsistent (last name/first name order exchange)
- encoding errors
- duplication
- different representation for the same thing
- different scales

“80%” Time Spent on Data Preparation

Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says [Forbes]
http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#73bf5b137f75



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

“80%” Time Spent on Data Cleaning

For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights [New York Times]

http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=0

Big Data's Dirty Problem [Fortune]

http://fortune.com/2014/06/30/big-data-dirty-problem/

The Silver Lining

“Painful process of cleaning, parsing, and proofing one’s data”
Is one of the 3 sexy skills of data geeks
(the other two: statistics, visualization)

http://medriscoll.com/post/4740157098/the-three-sexy-skills-of-data-geeks

@BigDataBorat tweeted
“Data Science is 99% preparation, 1% misinterpretation.”

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…

http://www.businessinsider.com/tabs-vs-spaces-from-silicon-valley-2016-5

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/

Data Cleaners

Watch videos

- Data Wrangler (research at Stanford)
- Open Refine (previously **Google Refine**)

in Alabama	Alabama
in Alaska	Alaska
in Arizona	Arizona
in Arkansas	Arkansas

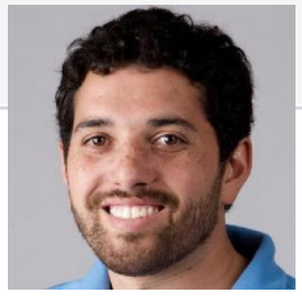
Write down

- 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/

DataWrangler alpha




Wrangler is an interactive tool for data cleaning and transformation.
Spend less time formatting and more time analyzing your data.

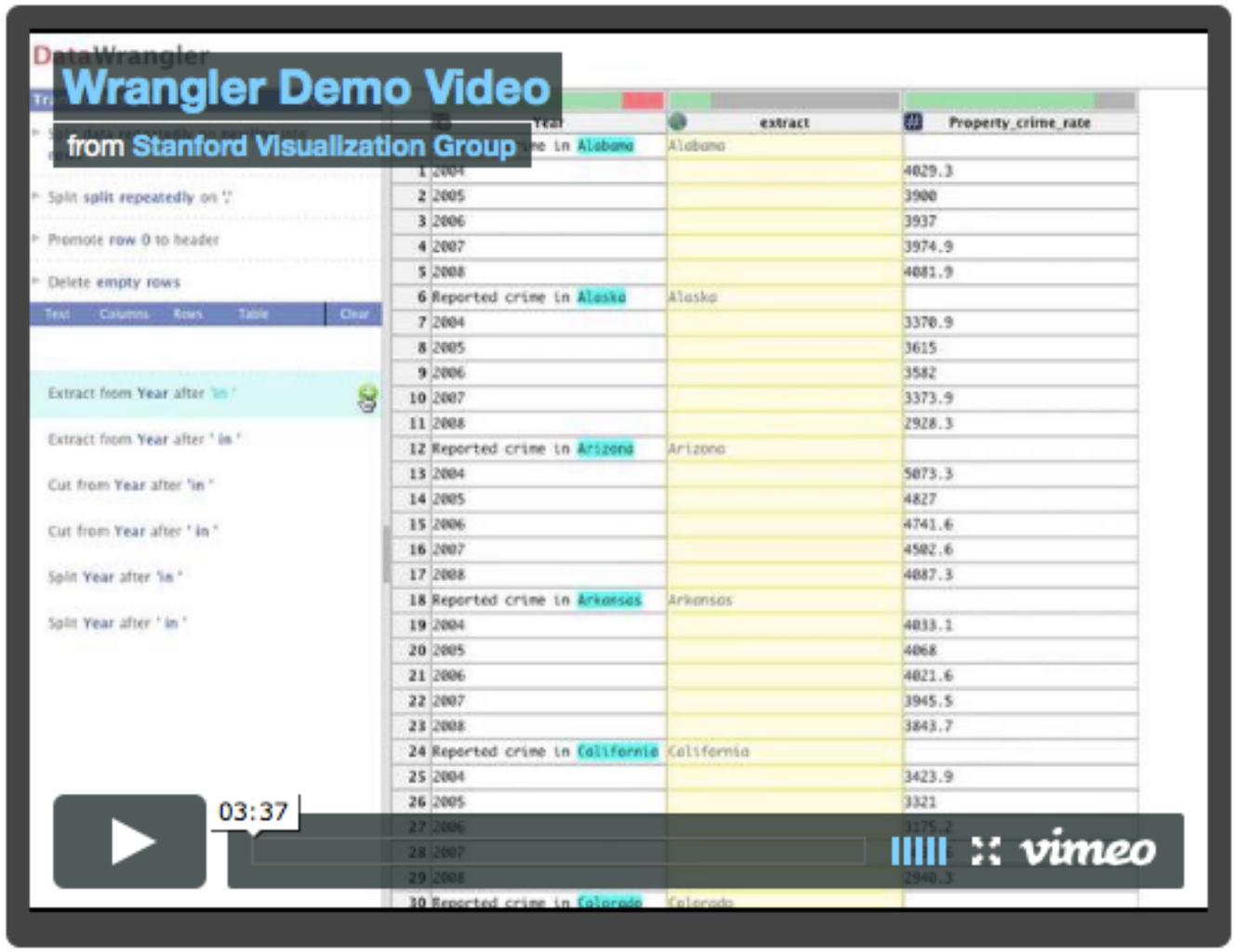
UPDATE: The Wrangler research project is complete, and the software is no longer actively supported. The team behind Wrangler has moved on to work on a commercial venture, Trifacta.

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



OpenRefine

A free, open source, powerful tool for working with messy data

Home

Download

Documentation

Community

Post archive

OpenRefine News: Spring 2016

OpenRefine News: December 2015

OpenRefine News: November 2015

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.

Please note that since October 2nd, 2012, Google is not actively supporting this project, which has now been rebranded to OpenRefine. Project development, documentation and promotion is now fully supported by volunteers. Find out more about the history of OpenRefine and how you can help the community.

Using OpenRefine - The Book

Using OpenRefine, by Ruben Verborgh and Max De Wilde, offers a great introduction to OpenRefine. Organized by recipes with hands on examples, the book covers the following topics:

1. Import data in various formats
2. Explore datasets in a matter of seconds
3. Apply basic and advanced cell transformations
4. Deal with cells that contain multiple values

What can the tools do?

- [O] identify similar words
- [O, W] history (in W: script)
- [O] clustering
- [O] show the “impact” of changes (#rows changed)
- [O, W] transformation
- [O, W] offline processing
- [O] highlight outliers
- [W] “pivoting”
- [W] previewing
- [W] extract part of word, do some generalization
- [W] histogram/vis for each column (highlight missing values)
- [W] suggested operations

O = Open Refine
W = Data wrangler

!

The videos only show *some* of the tools' features. Try them out.

Open Refine: http://openrefine.org
Data Wrangler: http://vis.stanford.edu/wrangler/