

<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

A large-scale photograph of a landfill. In the background, a blue bulldozer is visible, pushing a massive pile of garbage. The foreground and middle ground are filled with a dense, chaotic mass of trash, including plastic bags, crumpled paper, and other debris. In the bottom left corner, a large, worn black tire is prominent. The sky is bright blue, and hundreds of birds, likely seagulls, are seen in flight throughout the scene, suggesting a scavenger colony. The overall image conveys a sense of overwhelming waste and environmental neglect.

# 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

# How dirty is real data?

## Examples

- duplicates
- empty rows
- abbreviations (different kinds)
- difference in scales / inconsistency in description/ sometimes include units
- typos
- missing values
- trailing spaces
- incomplete cells
- synonyms of the same thing
- skewed distribution (outliers)
- bad formatting / not in relational format (in a format not expected)

# **“80%” Time Spent on Data Cleaning**

## **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>

## **Big Data's Dirty Problem [Fortune]**

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

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

# Data Janitor



# 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 |



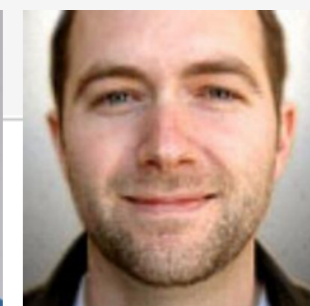
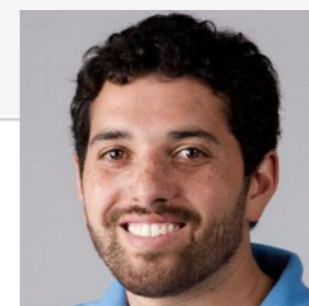
Write down

- 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/>



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

**Wrangler Demo Video**  
from Stanford Visualization Group

| Year                            | extract                   | Property_crime_rate |
|---------------------------------|---------------------------|---------------------|
| 1 2004                          | Reported crime in Alabama | 4029.3              |
| 2 2005                          |                           | 3900                |
| 3 2006                          |                           | 3937                |
| 4 2007                          |                           | 3974.9              |
| 5 2008                          |                           | 4081.9              |
| 6 Reported crime in Alaska      | Alaska                    |                     |
| 7 2004                          |                           | 3370.9              |
| 8 2005                          |                           | 3615                |
| 9 2006                          |                           | 3582                |
| 10 2007                         |                           | 3373.9              |
| 11 2008                         |                           | 2928.3              |
| 12 Reported crime in Arizona    | Arizona                   |                     |
| 13 2004                         |                           | 5073.3              |
| 14 2005                         |                           | 4827                |
| 15 2006                         |                           | 4741.6              |
| 16 2007                         |                           | 4502.6              |
| 17 2008                         |                           | 4087.3              |
| 18 Reported crime in Arkansas   | Arkansas                  |                     |
| 19 2004                         |                           | 4033.1              |
| 20 2005                         |                           | 4068                |
| 21 2006                         |                           | 4021.6              |
| 22 2007                         |                           | 3945.5              |
| 23 2008                         |                           | 3843.7              |
| 24 Reported crime in California | California                |                     |
| 25 2004                         |                           | 3423.9              |
| 26 2005                         |                           | 3321                |
| 27 2006                         |                           | 3175.2              |
| 28 2007                         |                           |                     |
| 29 2008                         |                           | 2940.3              |
| 30 Reported crime in Colorado   | Colorado                  |                     |

03:37

vimeo

# Refine

## OPEN



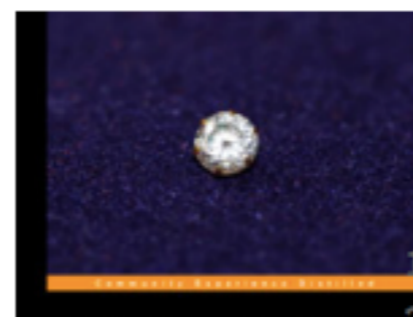
*A free, open source, powerful tool  
for working with messy data*

## Welcome!

OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; extending it with web services; and linking it to databases like [Freebase](#).

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

### Home

### Download

### Documentation

### Community

### Post archive

[A Governance Model for OpenRefine](#)

[Using OpenRefine: a manual](#)

# What can the tools do?

- [W, G] output transformation as scripts
- [G] clustering
- [G] trim data (due with typos, etc.)
- [G, W] undo
- [W] visualization, usability features
- [W] suggestions — “predictive interaction”
- [G] data in more than one format

**G** = Google Refine  
**W** = Data wrangler



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

**Google Refine:** <http://code.google.com/p/google-refine/>  
**Data Wrangler:** <http://vis.stanford.edu/wrangler/>