http://poloclub.gatech.edu/cse6242

CSE6242 / CX4242: Data & Visual Analytics

Data Cleaning

Duen Horng (Polo) Chau
Associate Professor, College of Computing
Associate Director, MS Analytics
Machine Learning Area Leader, College of Computing
Georgia Tech

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

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

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)

60 seconds

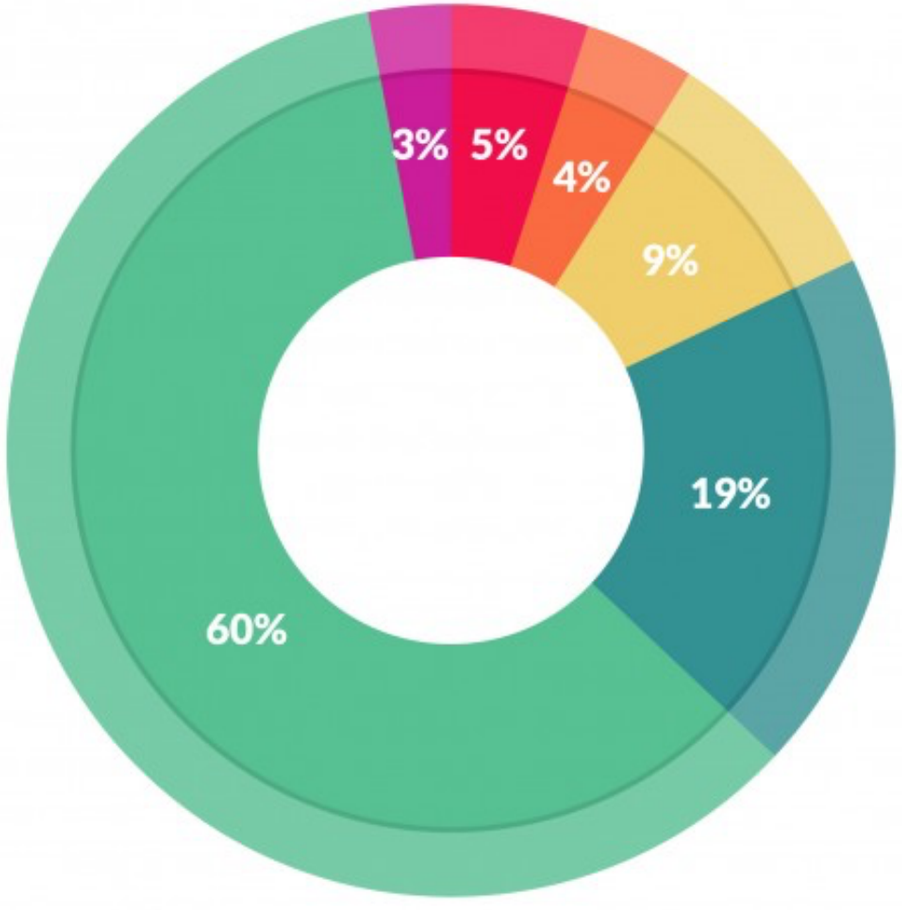
Comes up with **5+ kinds of “data dirtiness”**

How dirty is real data?

Importance of Data Cleaning

“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

Writing “Clean Code”

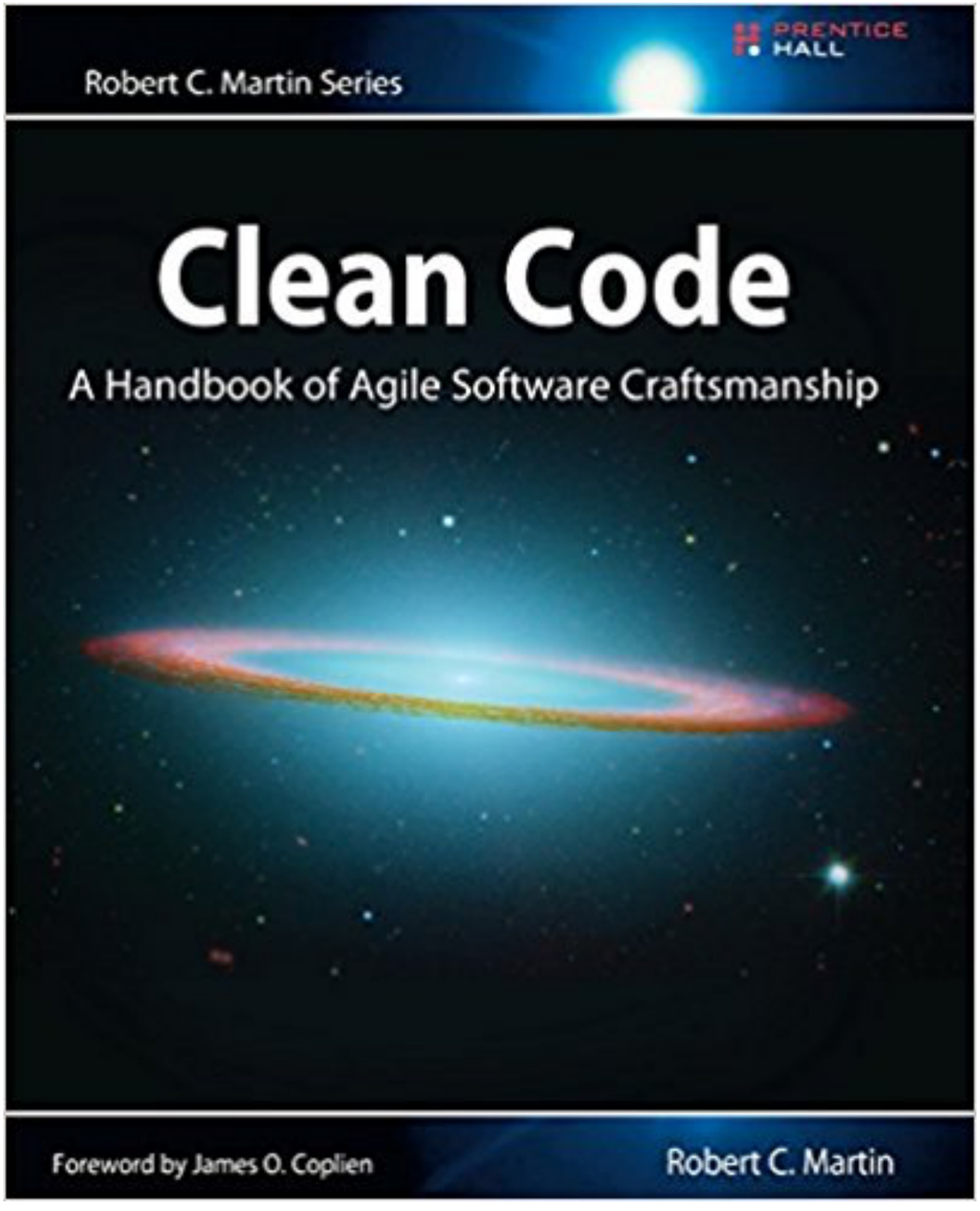
- Be careful with **trailing whitespaces**
- Indent code (**spaces vs tabs**) following coding practices in your team/company
  https://google.github.io/styleguide/javaguide.html#s4.2-block-indentation

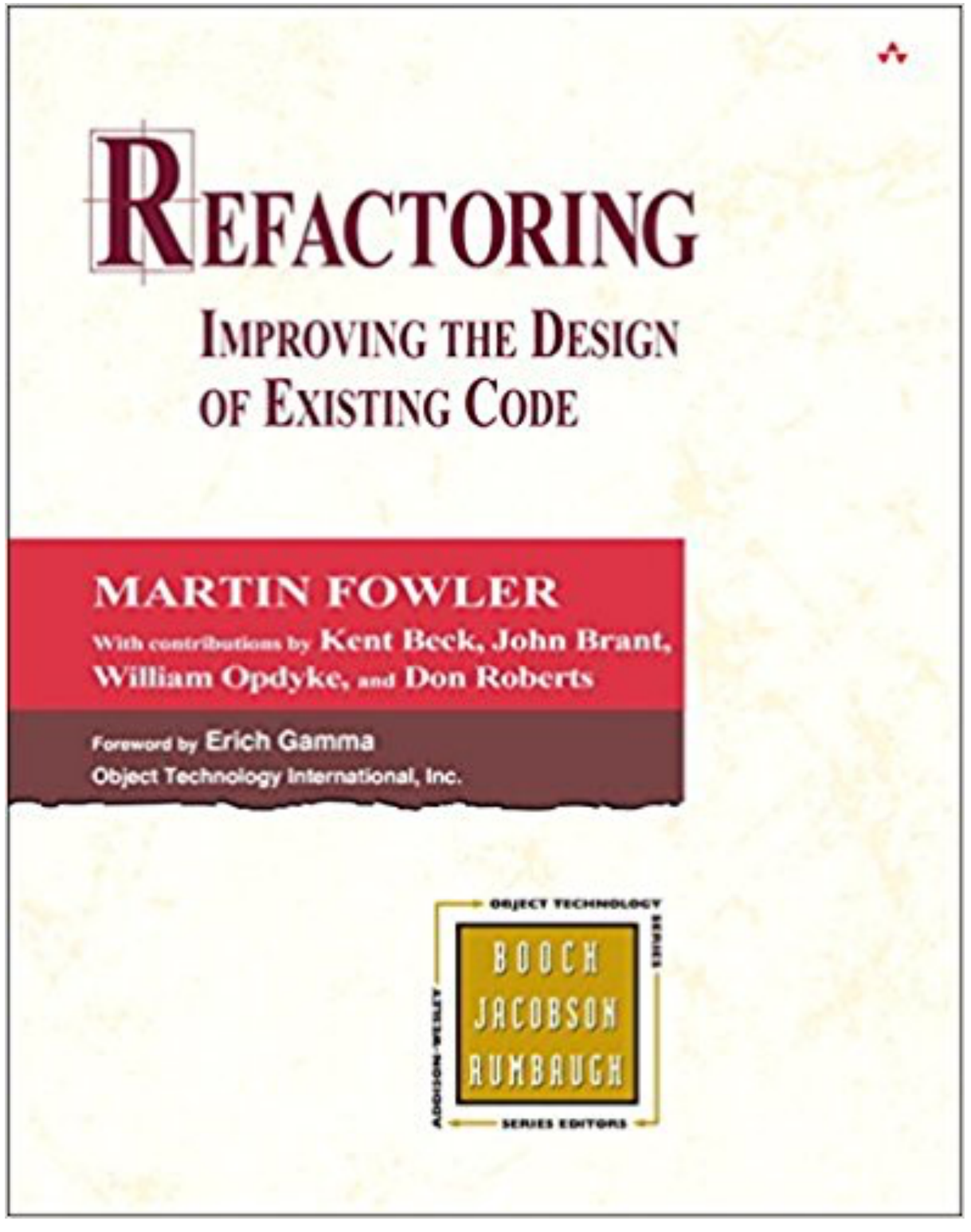
…there’s *no way* I'm going to be with someone who uses spaces over tabs…

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/


Robert C. Martin Series
PRENTICE HALL
Clean Code
A Handbook of Agile Software Craftsmanship
Foreword by James O. Coplien
Robert C. Martin


REFACTORING
IMPROVING THE DESIGN OF EXISTING CODE
MARTIN FOWLER
With contributions by Kent Beck, John Brant, William Opdyke, and Don Roberts
Foreword by Erich Gamma
Object Technology International, Inc.
OBJECT TECHNOLOGY SERIES
BOOCH JACOBSON RUMBAUGH
ADDISON-WESLEY
SERIES EDITORS

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

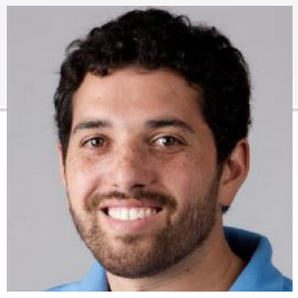
Write down

- 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.

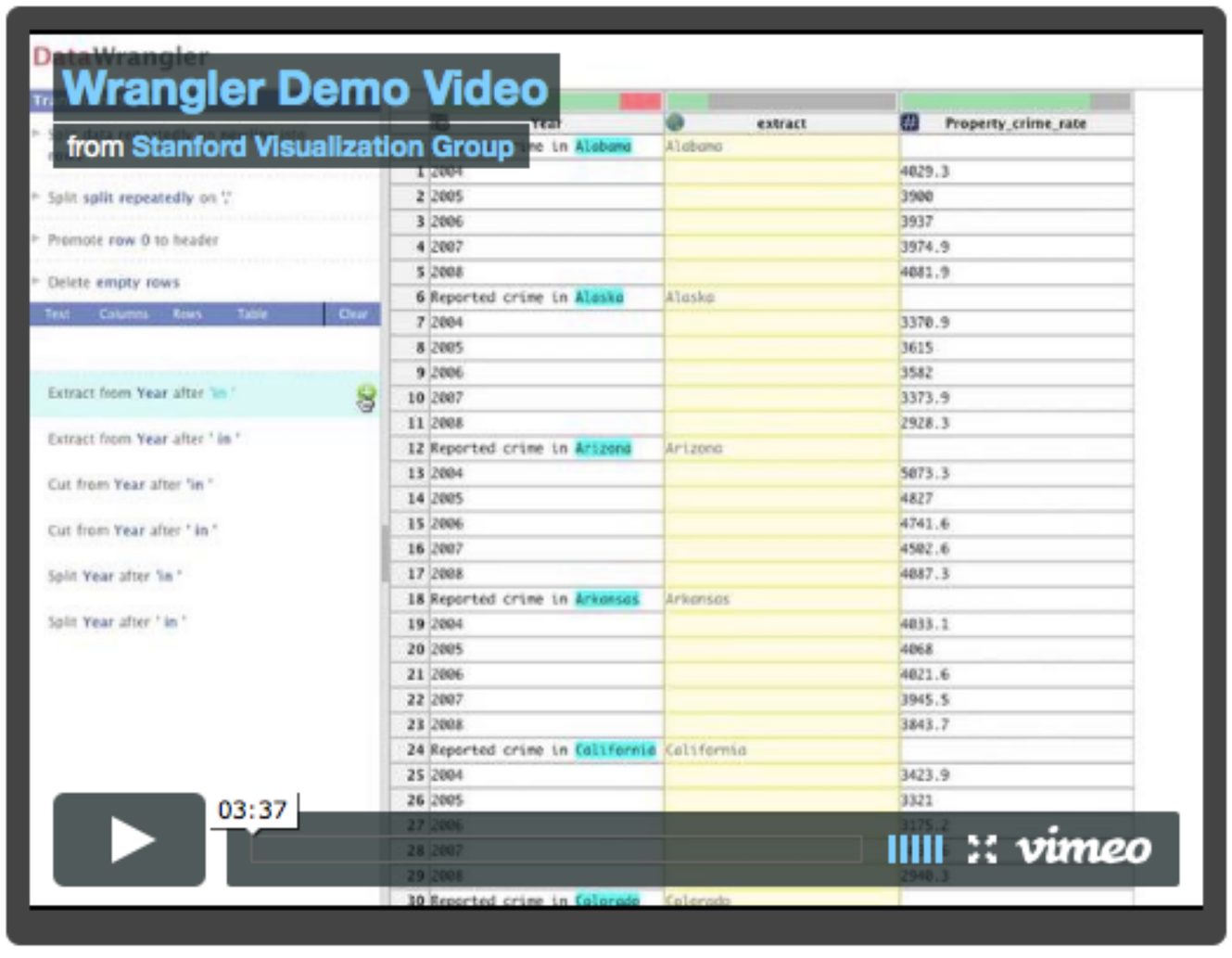
UPDATE: The Wrangler research project is complete, and the software is no longer actively supported. The team behind Wrangler has moved on to work on a commercial venture, Trifacta.

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



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

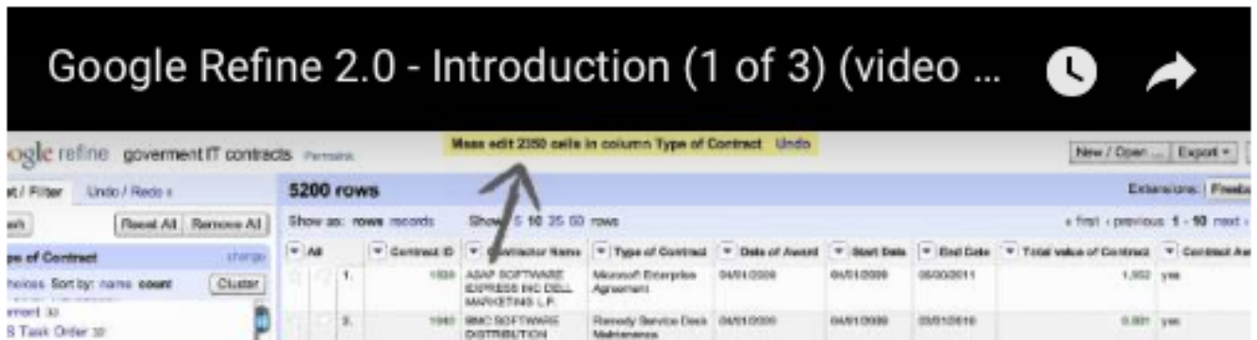
OpenRefine is supported by:

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

!

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

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