

poloclub.github.io/#cse6242

CSE6242/CX4242: **Data** & **Visual** Analytics

Data Cleaning

Duen Horng (Polo) Chau

Associate Professor, College of Computing

Associate Director, MS Analytics

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 you neighbors (group of 2-3)

60 seconds

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

How dirty is real data?

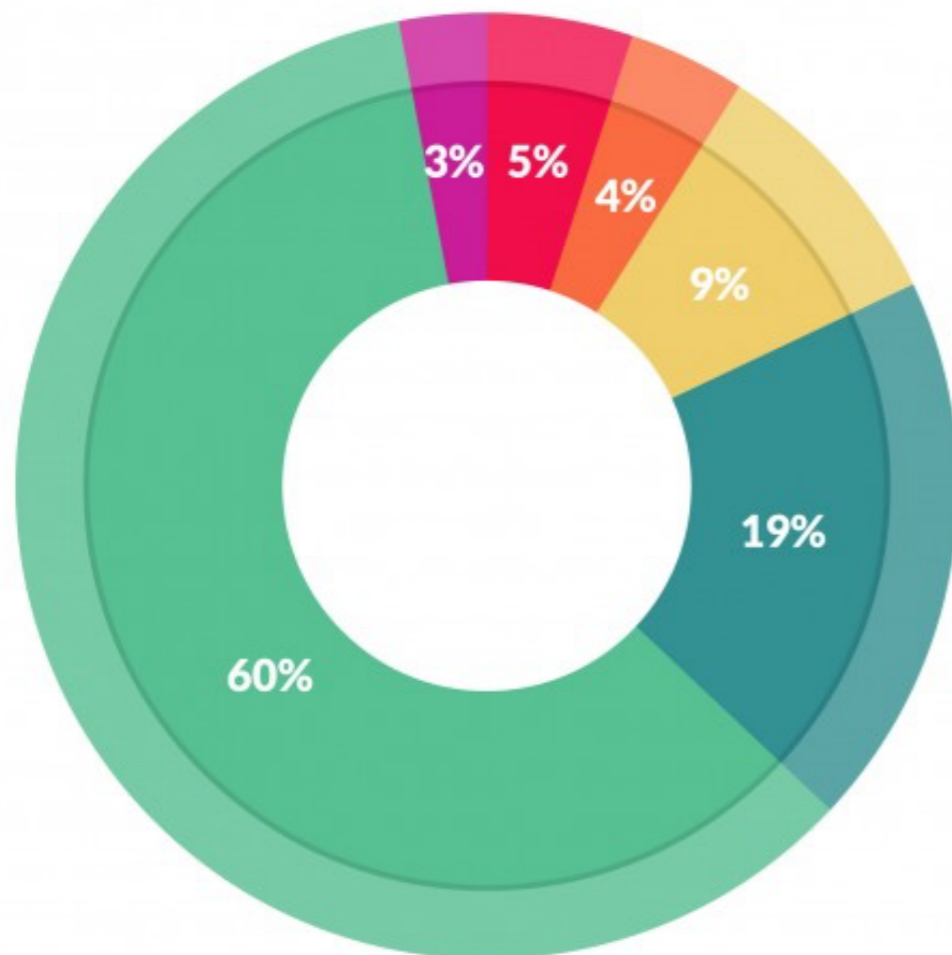
- Typos
- Missing data/fields
- Units (different)
- Data types
- Abbreviations
- Variations of the same thing
- Duplicates
- Encoding
- dashes, parentheses
- Delimiters
- White spaces
-

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%

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>



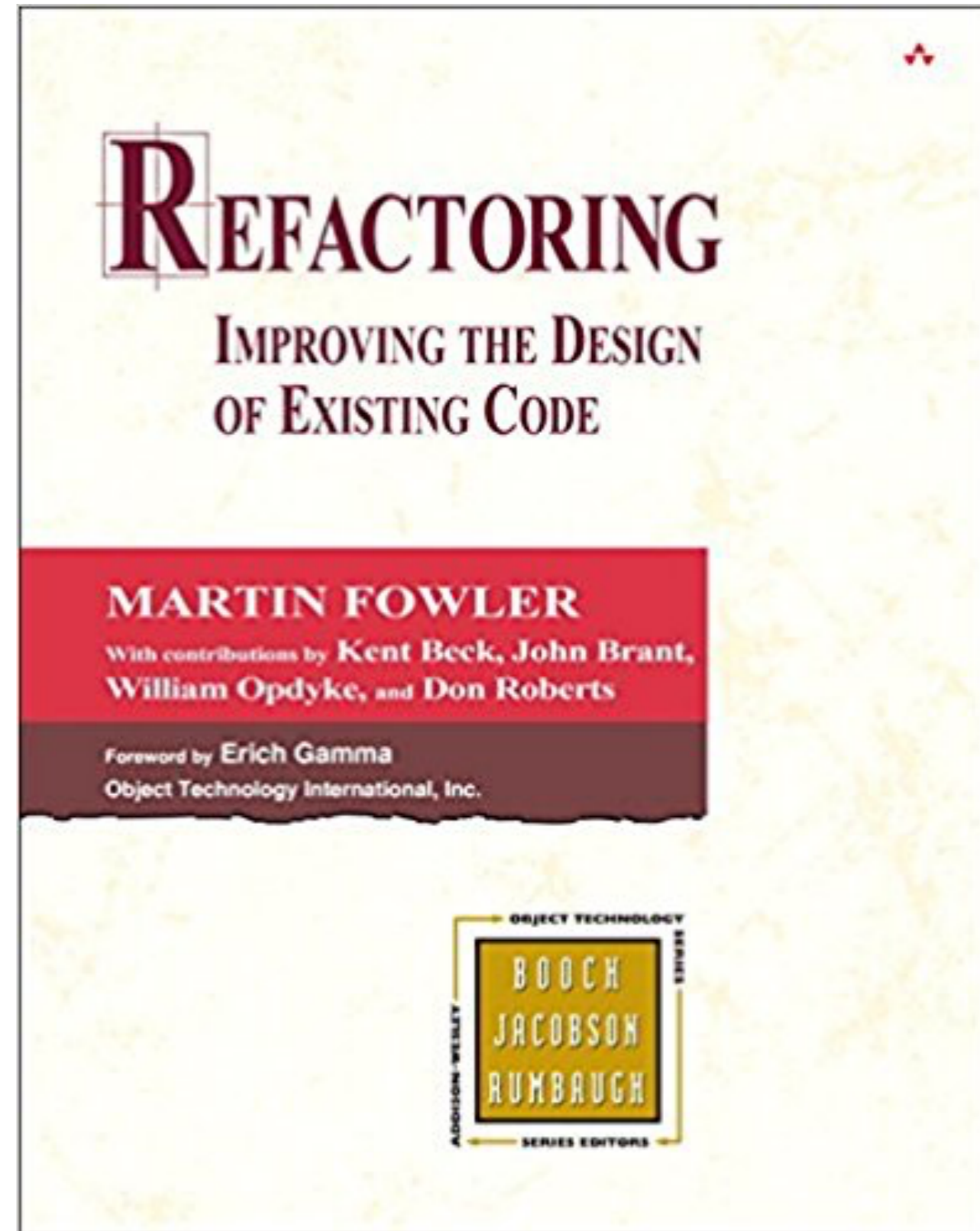
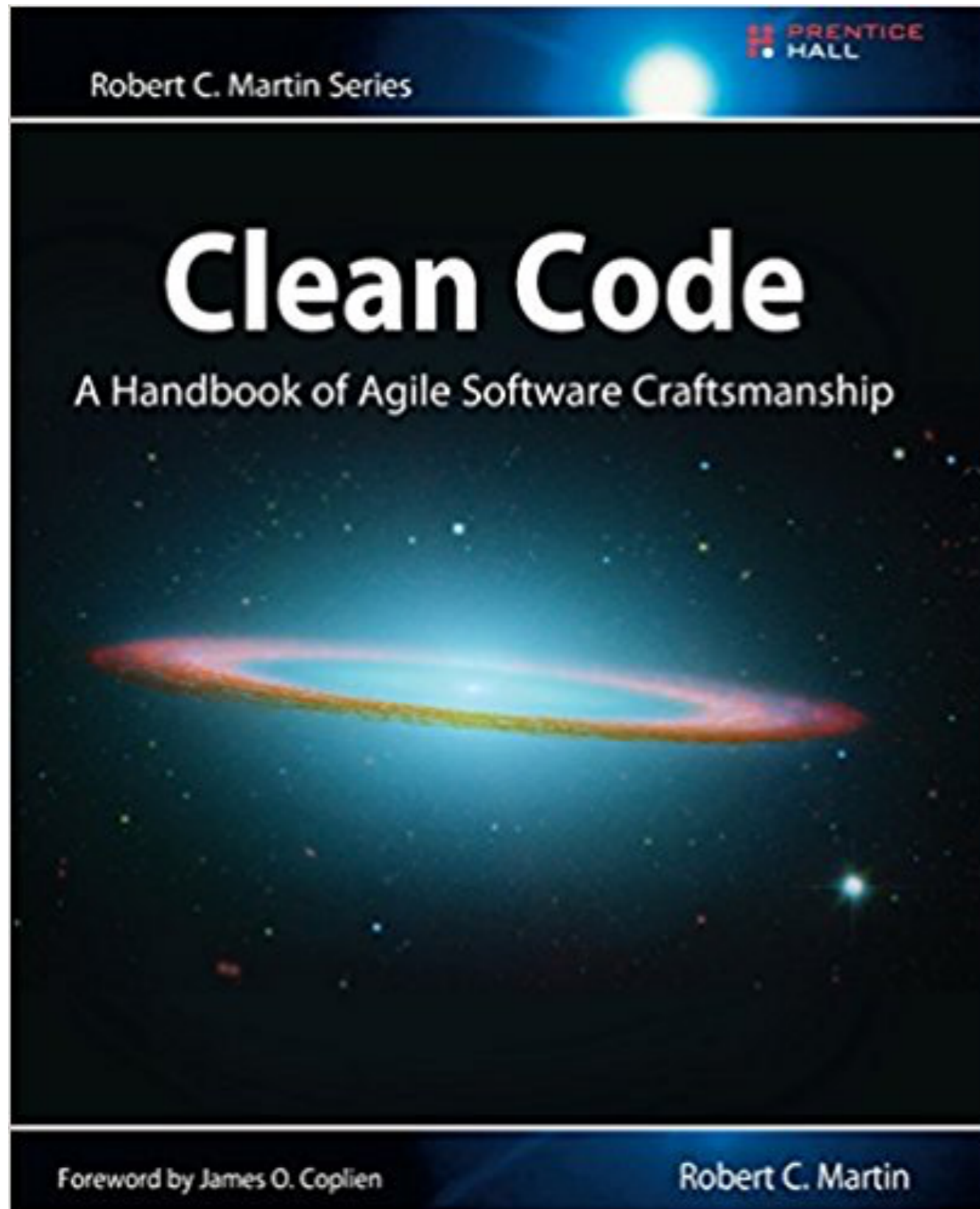
...there's *no way* I'm going to be with someone who uses spaces over tabs...

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

Trailing whitespace is evil — leads to “false differences”.
Don't commit evil into your repo.

<https://stackoverflow.com/questions/300489/why-is-it-bad-to-commit-lines-with-trailing-whitespace-into-source-control>

Both available **free** for GT students on
<https://www.oreilly.com/>



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



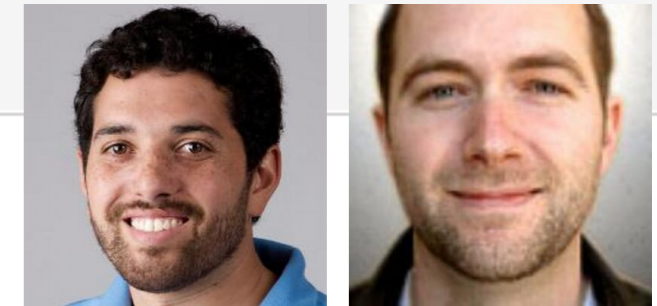
Write down

- 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](#)

The screenshot shows a video player interface for a demo of DataWrangler. The video title is "Wrangler Demo Video from Stanford Visualization Group". The video content displays a data table with the following columns: "Year", "State", "extract", and "Property_crime_rate". The data rows are numbered 1 through 30. The "State" column contains state names like Alabama, Alaska, Arizona, Arkansas, and California. The "Property_crime_rate" column contains numerical values. The video player includes a play button, a progress bar showing 03:37, and the Vimeo logo.

Year	State	Property_crime_rate
2004	Alabama	4029.3
2005	Alabama	3900
2006	Alabama	3937
2007	Alabama	3974.9
2008	Alabama	4081.9
2004	Alaska	3370.9
2005	Alaska	3615
2006	Alaska	3582
2007	Alaska	3373.9
2008	Alaska	2928.3
2004	Arizona	5073.3
2005	Arizona	4827
2006	Arizona	4741.6
2007	Arizona	4502.6
2008	Arizona	4087.3
2004	Arkansas	4033.1
2005	Arkansas	4068
2006	Arkansas	4021.6
2007	Arkansas	3945.5
2008	Arkansas	3843.7
2004	California	3423.9
2005	California	3321
2006	California	3175.2
2007	California	2940.3
2008	California	2940.3
2004	Colorado	3321



OpenRefine

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



Home

Community

Documentation

Download

Contact Us

Blog

Enhanced with Java profiler



JPROFILER

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 13



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

Open Refine: <http://openrefine.org>

Data Wrangler: <http://vis.stanford.edu/wrangler/>