

[poloclub.github.io/#cse6242](https://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?



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

60 seconds

# How dirty is real data?

- ...

# How dirty is real data?

- Coding;
- duplicated data
- Encoding UTF-8| special characters
- Outliers
- Outdated
- Missing data/features; sparse data
- Different names for the same thing; abbreviation
- Data that needs more context
- Feature misalignment
- Different units, types
- Extra works, spelling errors
- Inconsistent formatting
- False positives (irrelevant data)
- Multiple data sources (potentially disagree)
- Not handling delimiters
- Code mixing (diff language)

# 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
-

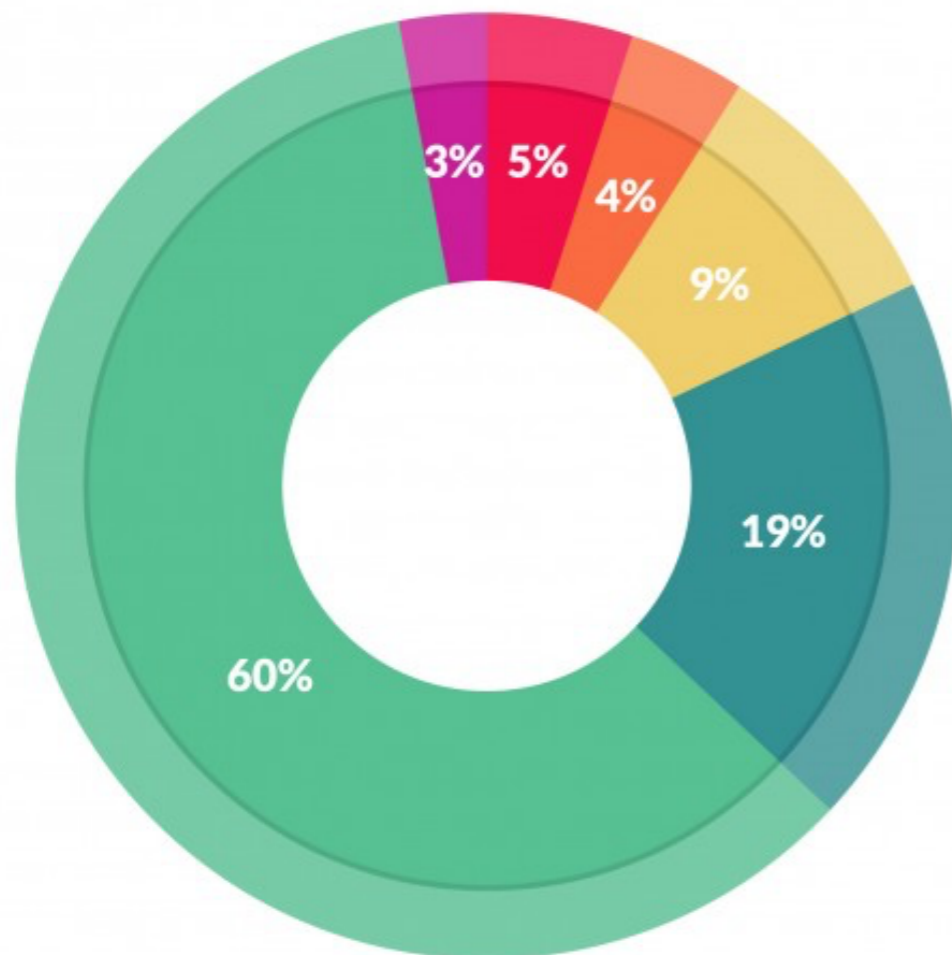
# Importance of Data Cleaning



# “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>



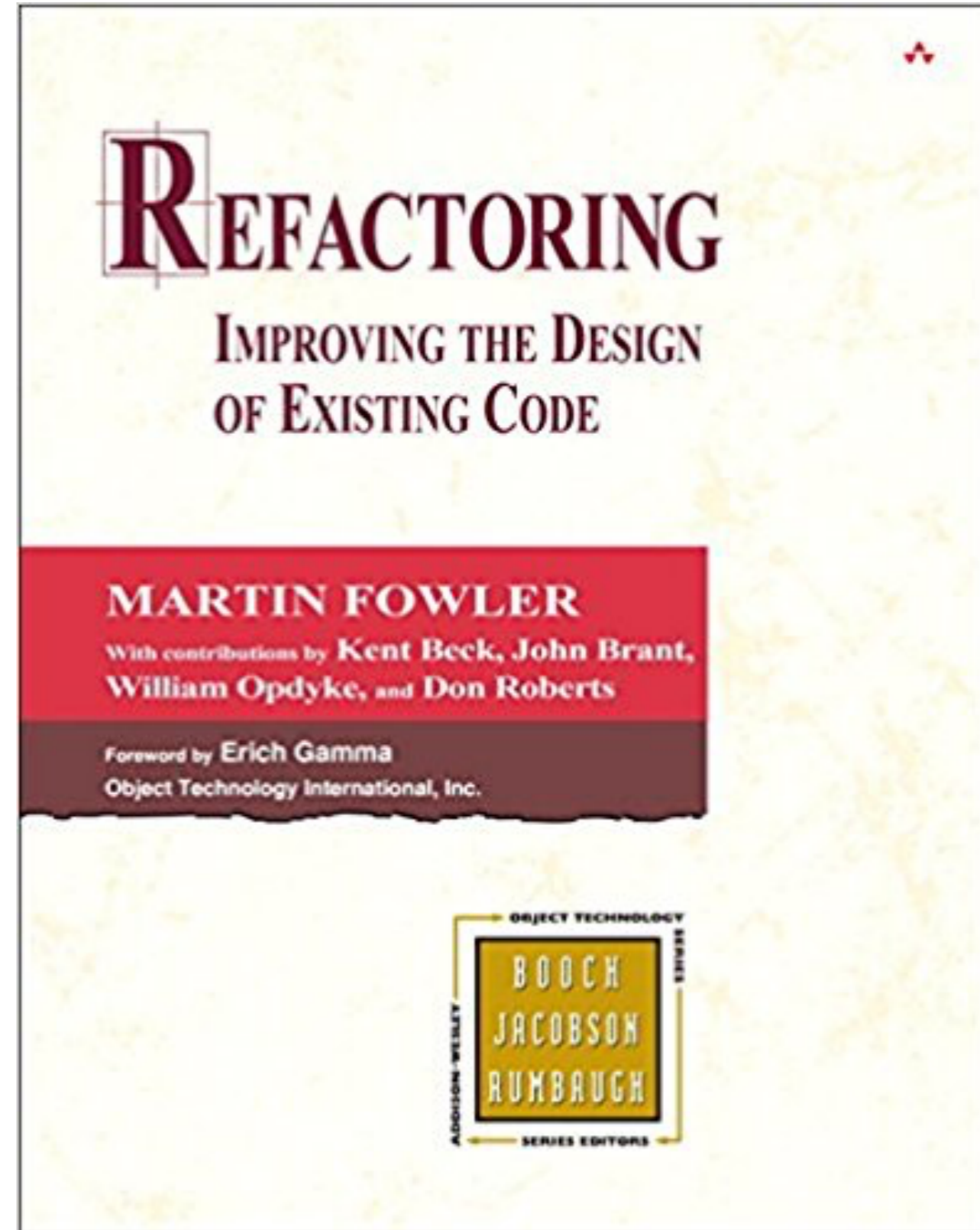
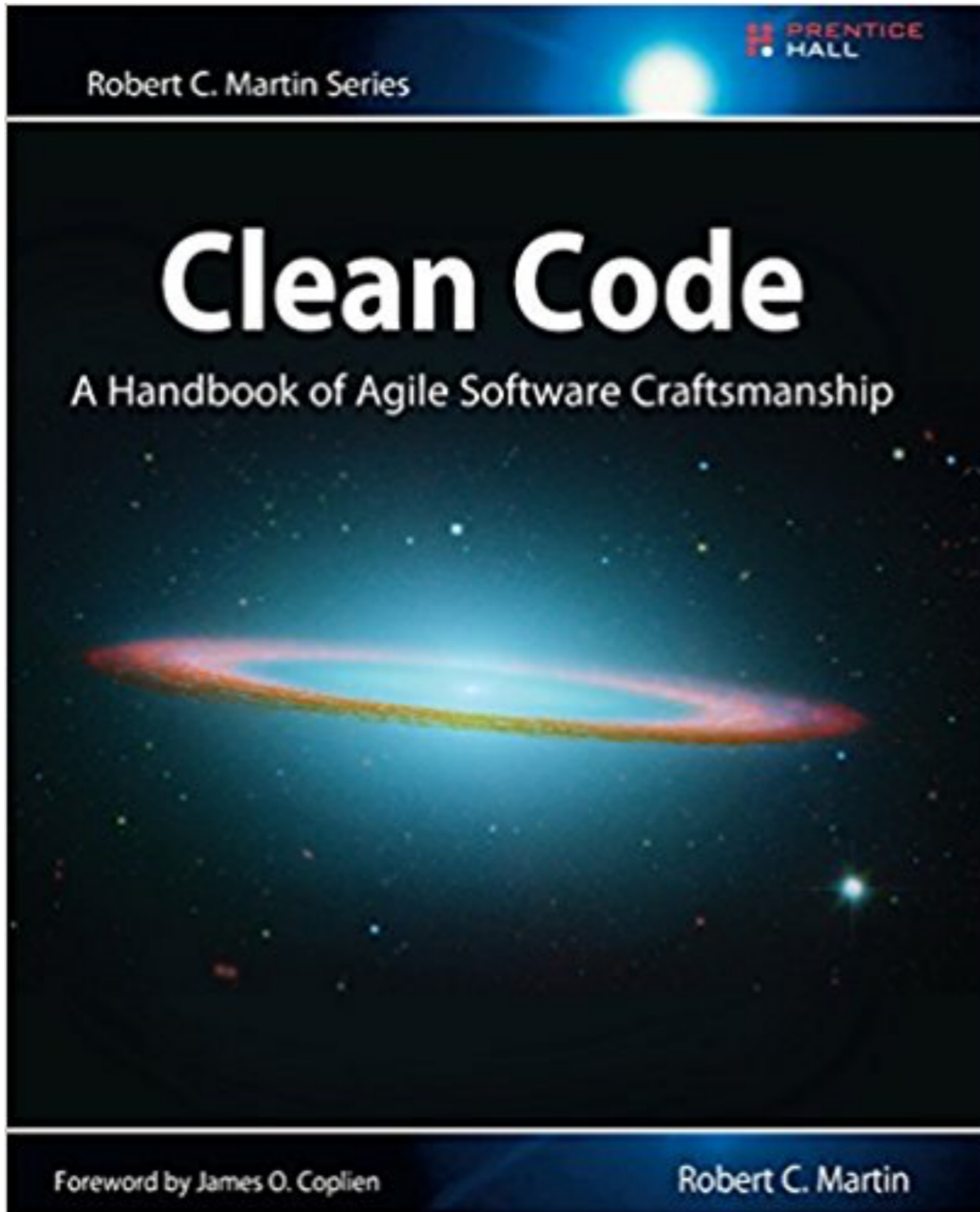
...there's *no way* I'm going to be with someone who uses spaces over tabs...

<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 started at Stanford)
- **Open Refine** (previously **Google Refine**)

|             |          |
|-------------|----------|
| in Alabana  | Alabama  |
| in Alaska   | Alaska   |
| in Arizona  | Arizona  |
| in Arkansas | Arkansas |



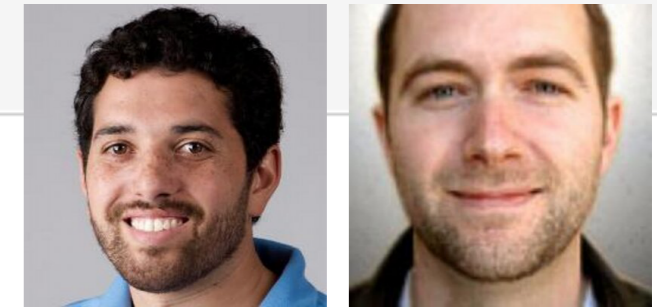
## Write down

- 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\\_video#1](http://openrefine.org_video#1)

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

03:37

# OpenRefine

OpenRefine is a powerful free, open source tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.



[Download](#)

## Main features



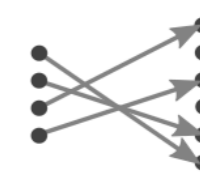
### Faceting

Drill through large datasets using facets and apply operations on filtered views of your dataset.



### Clustering

Fix inconsistencies by merging similar values thanks to powerful heuristics.



### Reconciliation

Match your dataset to external databases via reconciliation services.



### Infinite undo/redo

Rewind to any previous state of your dataset and replay your operation history on a new version of it.



### Privacy

Your data is cleaned on your machine, not in some dubious data laundering cloud.



### Wikibase

Contribute to Wikidata, the free knowledge base anyone can edit, and other Wikibase instances.

# What can Open Refine and Wrangler do?

**O** = Open Refine  
**W** = Data wrangler 15

# What can Open Refine and Wrangler do?

- [O] cluster
- [O] show data distribution
- [O] outlier detection
- [W>O] usability, seems more simple in a good way
- [W] suggestions
- [W] visualize proportion of missing data
- [O,W] History
- [W] relational ops, unfold, pivot, transform
- [O,W] select & edit multiple rows
- [W] export script as javascript
- [O>W] scale to larger data?
- [O] data privacy
- 

**O** = Open Refine

**W** = Data wrangler 16





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

**Open Refine:** <https://github.com/OpenRefine/OpenRefine/wiki/Screencasts>  
**Data Wrangler:** <http://vis.stanford.edu/wrangler/>