DLI Accelerated Data Science Teaching Kit

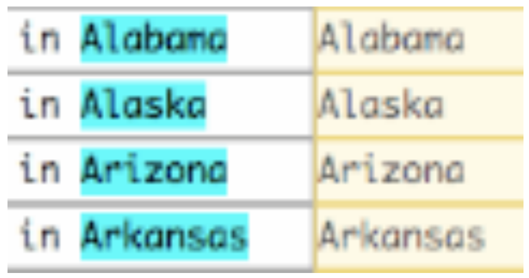
Lecture 3.3 - Data Cleaners: OpenRefine and Wrangler

The Accelerated Data Science Teaching Kit is licensed by NVIDIA, Georgia Institute of Technology, and Prairie View A&M University under the Creative Commons Attribution-NonCommercial 4.0 International License.

Data Cleaners

Watch videos

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



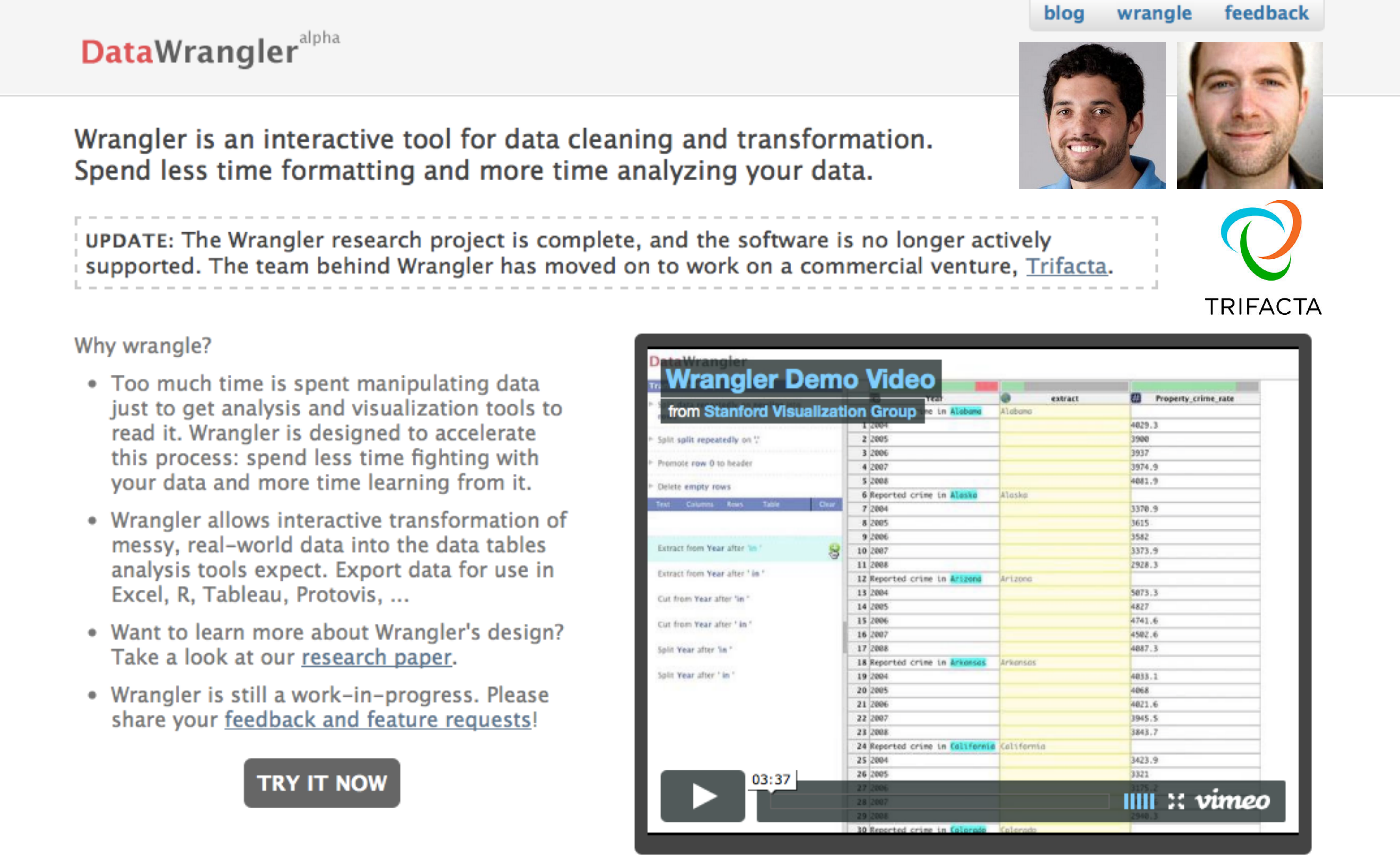
in Alabama	Alabama
in Alaska	Alaska
in Arizona	Arizona
in Arkansas	Arkansas

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

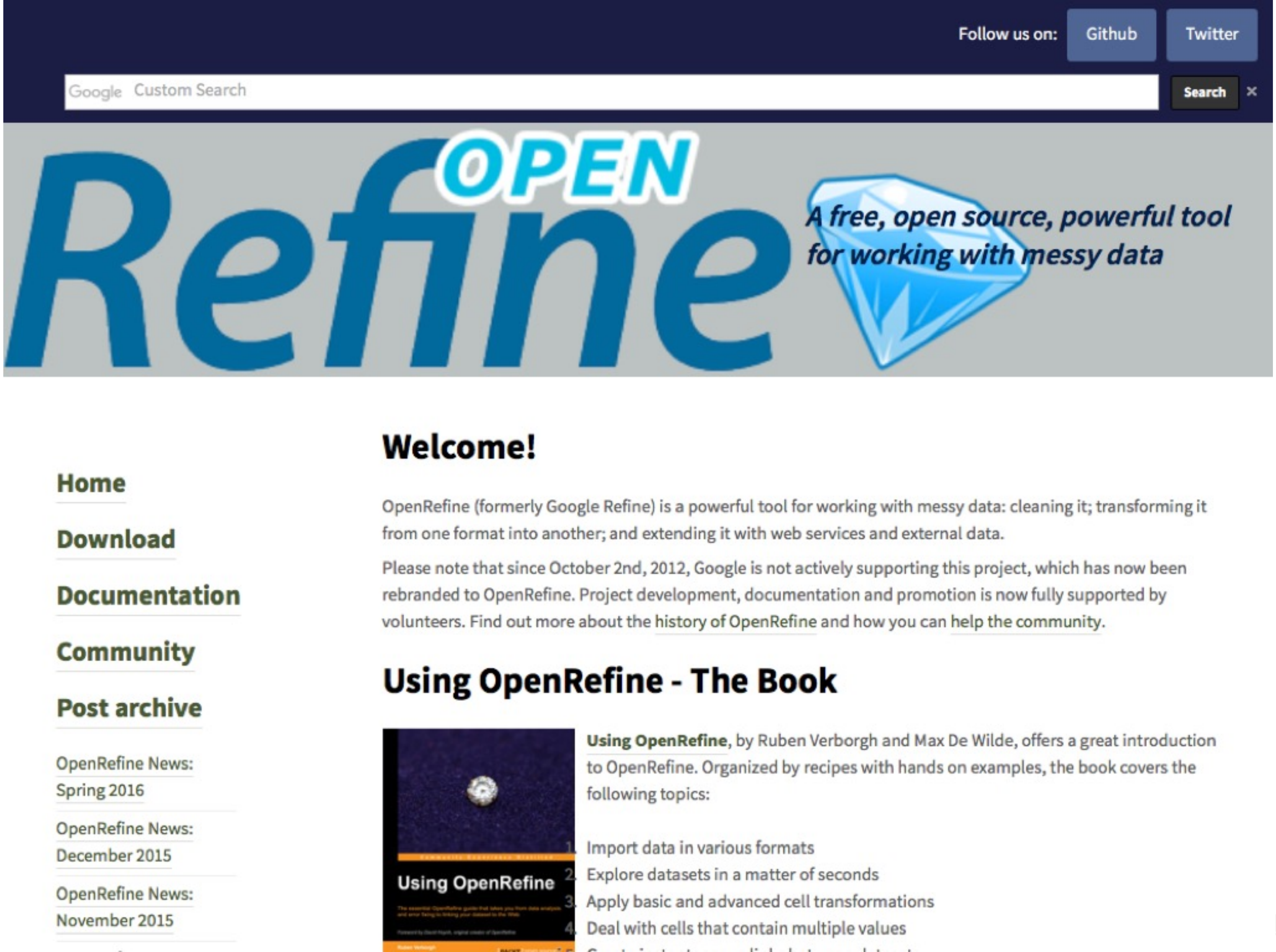
Wrangler is an interactive tool for data cleaning and transformation.
Spend less time formatting and more time analyzing your data.

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.

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



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

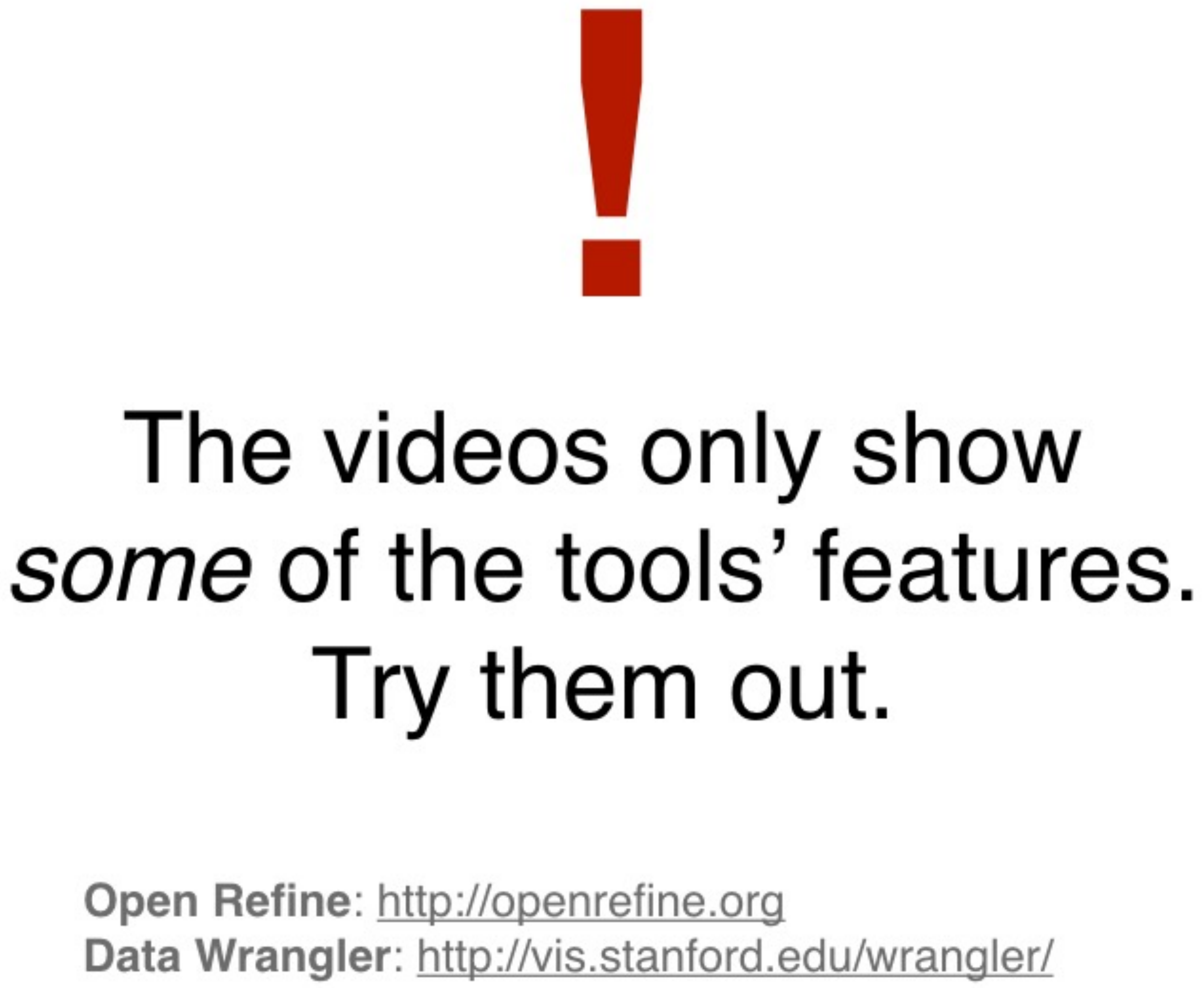
Students watch

- 1st OpenRefine “Explore Data” video at http://openrefine.org
- Wrangler video ar http://vis.stanford.edu/wrangler

What can Open Refine and Wrangler do?

- [O] clustering
- [O] show the “impact” of changes (#rows changed)
- [O] highlight outliers
- [W] preview
- [W] extract part of word, then generalize
- [W] histogram/vis for each column (highlight missing values)
- [W] suggest operations
- [W] “pivoting”
- [O, W] history
- [O, W] offline processing
- [O, W] transformation

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/

DEEP LEARNING INSTITUTE

DLI Accelerated Data Science Teaching Kit

Thank You