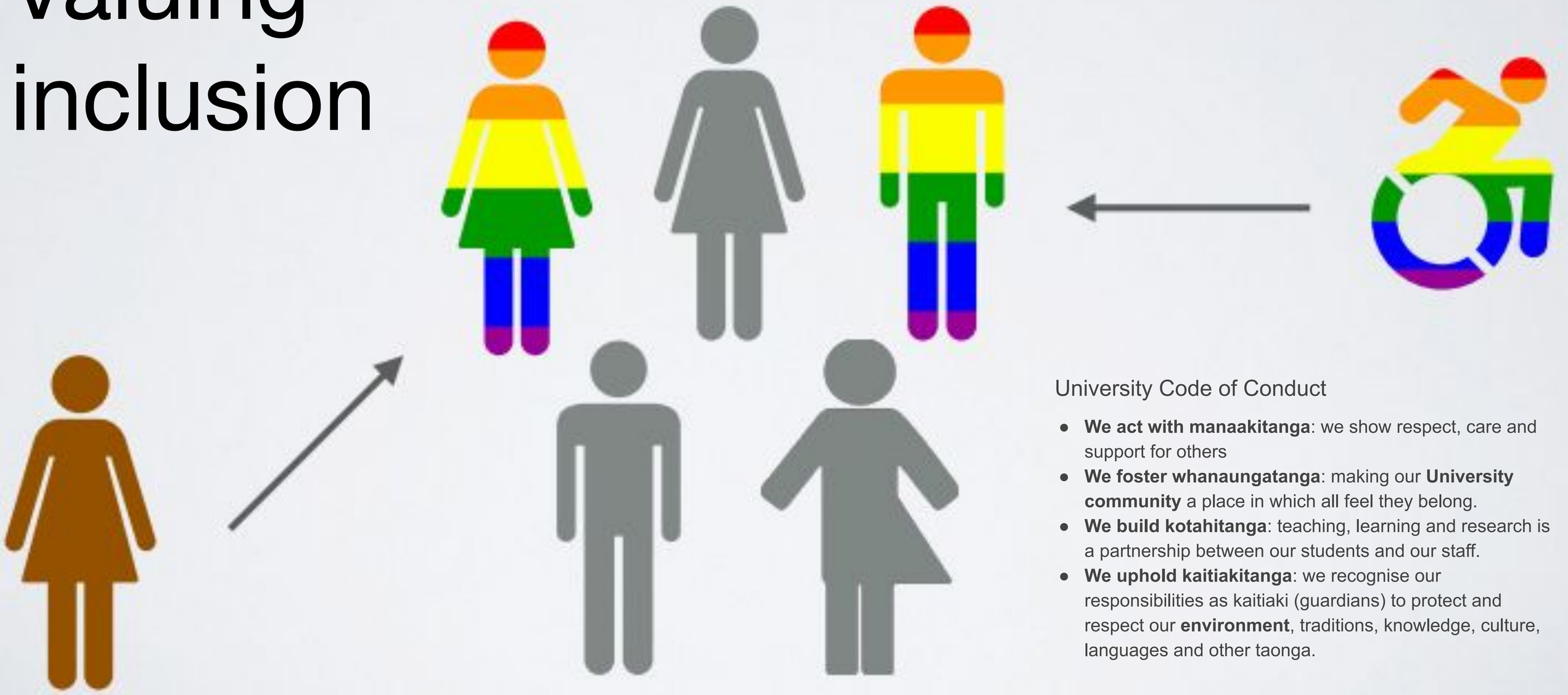
Introduction to cleaning data in OpenRefine

Yvette Wharton
Centre for eResearch

Valuing inclusion



University Code of Conduct

- **We act with manaakitanga**: we show respect, care and support for others
- **We foster whanaungatanga**: making our **University community** a place in which all feel they belong.
- **We build kotahitanga**: teaching, learning and research is a partnership between our students and our staff.
- **We uphold kaitiakitanga**: we recognise our responsibilities as kaitiaki (guardians) to protect and respect our **environment**, traditions, knowledge, culture, languages and other taonga.

The dream

```
# (ideal) data analysis process
raw_data = GET(data)
proc_data = PROCESS(raw_data)
SUMMARY(proc_data)
PLOT(proc_data)
model = FIT_MODEL(proc_data)
prediction = PREDICT(model)
PRINT(prediction)
> "Woo-hoo! validated model =)"
```

The reality

```
# (real) data analysis process
raw_data = GET(data)
clean_data = CLEAN(data)
proc_data = PROCESS(clean_data)
while (QUALITY(proc_data) != "good") {
   clean_data = CLEAN(proc_data)
   proc_data = PROCESS(clean_data)
   # while loop may run indefinitely
}
SUMMARY(proc_data)
PLOT(proc_data)
model = FIT_MODEL(proc_data)
prediction = PREDICT(model)
PRINT(prediction)
> "Ooops! model sucks =("
```

Images from http://www.gastonsanchez.com/visually-enforced/opinion/2012/09/19/Data-analysis-process/

Tidy Data Principles

- always keep a copy of the raw data
- have a separate copy which is your tidy dataset
- keep metadata record (codebook, readme.txt)
- keep a record of your ‘recipe’ (exact steps taken) to get from raw to tidy data

Reading:: Hadley Wickham, *Tidy Data*, Vol. 59, Issue 10, Sep 2014, Journal of Statistical Software. http://www.jstatsoft.org/v59/i10.

Useful Resources

http://openrefine.org - great introductory videos

Google Group -good for beginner questions and problems.

OpenRefine Google Plus community - help

OpenRefine ecology data - Data Carpentry OpenRefine tutorials

OpenRefine for Social Science data

Typical workflow

- Open OpenRefine (if it doesn't start automatically in your browser go to http://127.0.0.1:3333/)
- Import the dataset – CSV, tab – file, URL
- Explore data using 'facets' and 'filters'
- Use cluster analysis to make consistent
- Rearrange, split, sort
- Repeat
- Export dataset

Installing OpenRefine

1. OpenRefine website: http://openrefine.org/
2. Download section.
3. Choose the appropriate download for your operating system (Windows, Mac or Linux).
4. Follow the installation procedures

Tips

- I need more memory https://github.com/OpenRefine/OpenRefine/wiki/FAQ-Allocate-More-Memory
- backup openrefine data https://github.com/OpenRefine/OpenRefine/wiki/Back-Up-OpenRefine-Data

Faceting

- Grouping the dataset based on one or more parameters, properties, fields, columns
- Like tagging
- You can then explore just those records at the intersection of the facets
- A “species symbol” facet, for instance, would group all records that have the same species name
- Possible to facet on text, number ranges, pairs of numbers, etc.

More information: https://github.com/OpenRefine/OpenRefine/wiki/Faceting

Clustering

- Often faceting will reveal inconsistencies in the data
- Cluster analysis attempts to form clusters of data based on certain algorithms
- OpenRefine allows you try a variety of clustering methods
- These are quite good at revealing inconsistencies, e.g.:**Kriesler** vs **Chrysler**

More information:
https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth

Download data from a URL (API)

- fetch JSON from any web service (databases, registries, mapping services, etc.) based on values in a OpenRefine project
- Open Refine makes it relatively straightforward to call into an API, receive a response, and supplement your dataset with a portion of it.
- two step process - get data then parse data

More information:
https://github.com/OpenRefine/OpenRefine/wiki/Fetching-URLs-From-Web-Services

Regular expressions (regex) using GREL

- regex = codes used for matching patterns. Often there is more than one way to compose a regex pattern-match.
- GREL General Refine Expression Language
- much of Refine's extensible and advanced power comes from regular expressions.
- Useful handout on regex and cheat-sheet.

FAQs for problems

- Sometimes using a browser other than Firefox, OpenRefine does not automatically open.
  - Point your browser at http://127.0.0.1:3333/ or http://localhost:3333 to launch the program.
- Issues getting OpenRefine to run on Windows.
  - install Java (JDK + JRE) and add “JAVA_HOME” and “JDK_HOME” to the environment. This thread includes steps to diagnose possible issues, and links on how to set up environment variables.
- Mac users with the newest operating system MAY have to allow this to run by “allowing everything” to run. They can change the setting back after the exercise OR right click and select open.

More resources

- OpenRefine web site
- OpenRefine Documentation for Users (Wiki site)
- Using OpenRefine book by Ruben Verborgh, Max De Wilde and Aniket Sawant
- Grateful Data is a fun site with many resources devoted to OpenRefine, including a nice tutorial.
- Margaret Heller shows how she uses OpenRefine for Measuring and Counting Impact in Repositories.
- Intersect Course Resources has Jared Berghold's **Cleaning & Exploring your data with Open Refine**.