

# Converting data

Hadley Wickham

# Outline

- Typical data format + sample data
- Measured vs. id variables
- Using excel for data cleaning

# Common file formats

- Excel
- SPSS
- SAS
  
- I recommend csv

[http://tinyurl.com/](http://tinyurl.com/388y2m)  
388y2m

# Data

- These are pretty typical datasets you find on the web. How can we modify them for use with GGobi?
- We need to put the data in a strict rectangular format. Single row of column headings.

# Normalisation

- Minimise redundancy and inconsistency
- Each “fact” stored only once
- May not be optimal for entry or analysis
- Multiple datasets, each containing information about one “entity”

# Id variables (keys)

- Identify a measurement
- Like indices on a random variable
- Fixed by design of experiment  
(known in advance)
- vs. measurements/measured variables

# Basic structure

- Remove junk (top, bottom, + strange values)
- Give variables short memorable names
- Rearrange
- Save as csv, tab separated
- Load in to GGobi or Mondrian to check



Demo

Your turn - pick another  
dataset and practice

# Homework

- 1-2 page summary of what you've liked and disliked about the class
- What could I do better?

# Next week

- <http://www.geom.uiuc.edu/~banchoff/Flatland/>
- Sections: esp 1, 5, 6, 13, 15-17, 19