ST599 Statistical Computing and Big Data

Charlotte Wickham & Alix Gitelman

March 31, 2014

What is Big Data?

Some definitions

The four V's: variety, velocity, volume and value (veracity?)

It's big when your usual tools don't work

"When computational time exceeds cognitive time"

A marketing phrase

(More statistical) It's big when your usual inferential approaches don't work

There is always something bigger! We'll try to concentrate on general strategies, but to practice we'll have to learn some specific tools.

1. Getting started

As data gets bigger, basic descriptive and/or exploratory analysis becomes hard:

- the data fits in memory but your usual tools/approaches take too long
- the data doesn't even fit in memory (i.e. you can't get it into R)
- the data won't fit on your computer let alone in memory

But often answering our questions doesn't require *big* data.

Making big data small: subset, summarise, sample

We'll also learn some new tools: dplyr, git, SQL

2. Statistical issues with big data

What happens when we apply our usual statistical tools (t-tests, regression etc.) to big data?

Does it even make sense to apply our usual statistical tools to big data? Do we care about p-values?

Bigger data isn't neccessariy better data

Some things don't change: What's the sample? What's the population?

You might have to think about the implementation of your usual tools

3. Data Mining & Machine Learning

A brief survey of some techniques popular in the data mining and machine learning arenas

What do they do? When are they appropropriate? What don't they do? How do they scale?

Classification and Regression trees, Random Forests, clustering methods, …

The scope of our class

You learn to think about the statistical issues when using "Big" data

You learn some new tools to cope with realistic data

You get some experience doing data analysis in a team, including communicating your results

Syllabus?