Getting and Cleaning Data Overview

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Getting and Cleaning Data Content

- Raw vs. tidy data
- Downloading files
- Reading data
  - Excel, XML, JSON, MySQL, HDF5, Web, ...
- Merging data
- Reshaping data
- Summarizing data
- Finding and replacing
- Data resources

Connecting and listing databases

```
ucscDb <- dbConnect(MySQL(), user = "genome", host = "genome-mysql.cse.ucsc.edu")
result <- dbGetQuery(ucscDb, "show databases;")
dbDisconnect(ucscDb)
result
```

Merging data - merge()

```
mergedData2 <- merge(reviews, solutions, by.x = "solution_id", by.y = "id",
    all = TRUE)
head(mergedData2[, 1:6], 3)
reviews[1, 1:6]
```

Raw versus processed data

Raw data

- The original source of the data
- Often hard to use for data analyses
- Data analysis *includes* processing
- Raw data may only need to be processed once

http://en.wikipedia.org/wiki/Raw_data

Processed data

- Data that is ready for analysis
- Processing can include merging, subsetting, transforming, etc.
- There may be standards for processing
- All steps should be recorded

http://en.wikipedia.org/wiki/Computer_data_processing