ETC1010: Data Modelling and Computing

Lecture 1: Introduction

Di Cook (dicook@monash.edu, @visnut)

7/25/2017

Welcome to Semester 2 2017!

📊 **Lecturers**: Professor Di Cook and Dr David Frazier

📊 **Tutors**: Puwasala Gamakumara, Mitch O'Hara-Wild, Yeasmin Mahbuba

Welcome to Semester 2 2017!

📊 **Lecturers**: Professor Di Cook and Dr David Frazier

📊 **Tutors**: Puwasala Gamakumara, Mitch O'Hara-Wild, Yeasmin Mahbuba

📊 **Unit guide**: Objectives, tentative schedule, grading

Welcome to Semester 2 2017!

- 📊 **Lecturers**: Professor Di Cook and Dr David Frazier

- 📊 **Tutors**: Puwasala Gamakumara, Mitch O'Hara-Wild, Yeasmin Mahbuba

- 📊 **Unit guide**: Objectives, tentative schedule, grading

- 📊 **Textbook**: R for Data Science, Garret Grolemund and Hadley Wickham

Welcome to Semester 2 2017!

- 📊 **Lecturers**: Professor Di Cook and Dr David Frazier

- 📊 **Tutors**: Puwasala Gamakumara, Mitch O'Hara-Wild, Yeasmin Mahbuba

- 📊 **Unit guide**: Objectives, tentative schedule, grading

- 📊 **Textbook**: R for Data Science, Garret Grolemund and Hadley Wickham

- 📊 **Computing**: R and RStudio

Welcome to Semester 2 2017!

- **Lecturers**: Professor Di Cook and Dr David Frazier

- **Tutors**: Puwasala Gamakumara, Mitch O'Hara-Wild, Yeasmin Mahbuba

- **Unit guide**: Objectives, tentative schedule, grading

- **Textbook**: R for Data Science, Garret Grolemund and Hadley Wickham

- **Computing**: R and RStudio

- **Materials**: Moodle for grades, online quizzes; course web site for lecture and lab materials
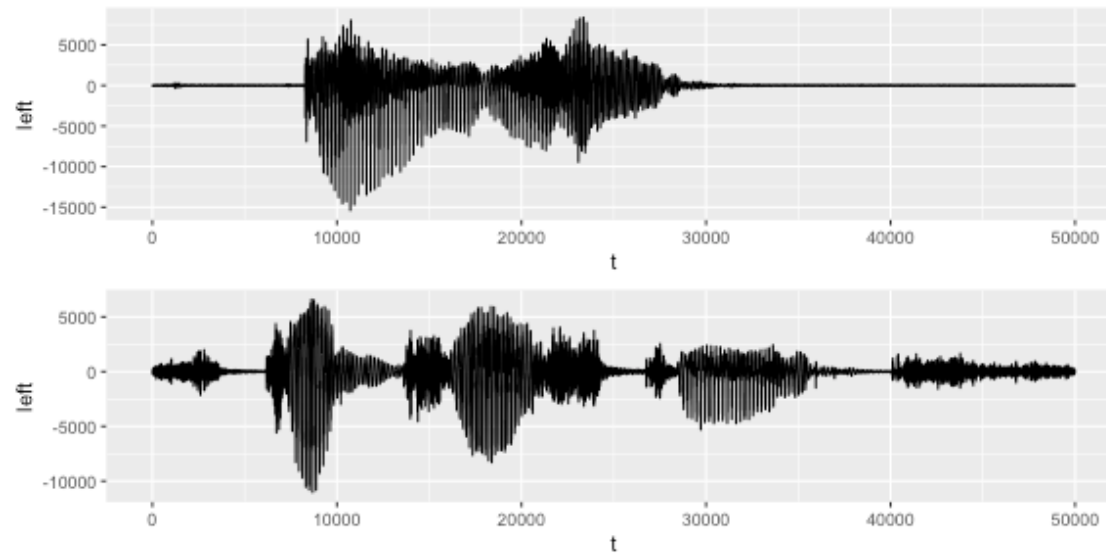
Outline

- What is data?
- What can we do if we have good data handling skills?
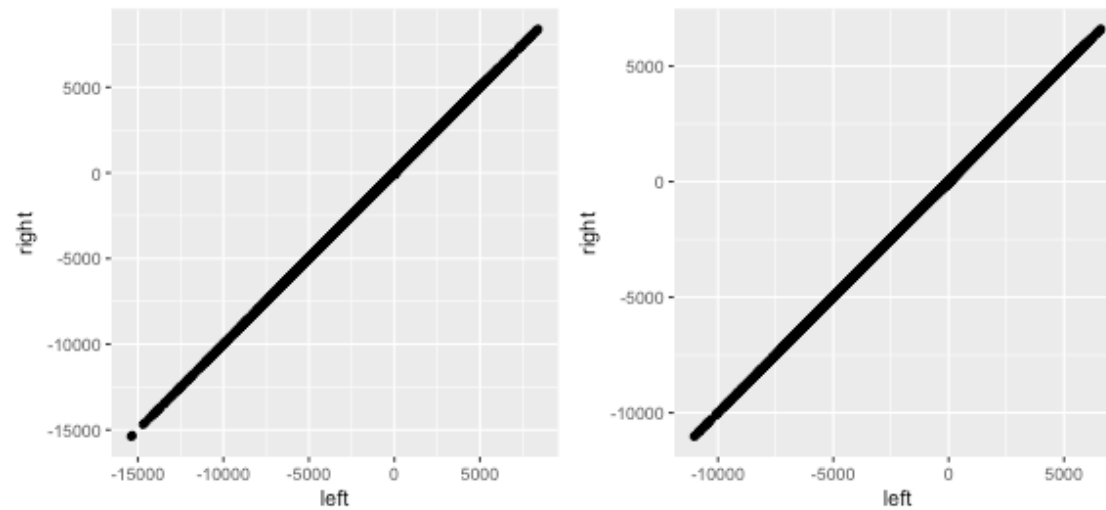 - **Insert lots of examples here**
- Why R?

Music

- 📊 Read in a wave file
- 📊 Digitise it as time and amplitude
- 📊 Plot
- 📊 Compare with other sounds

Compare left and right channels

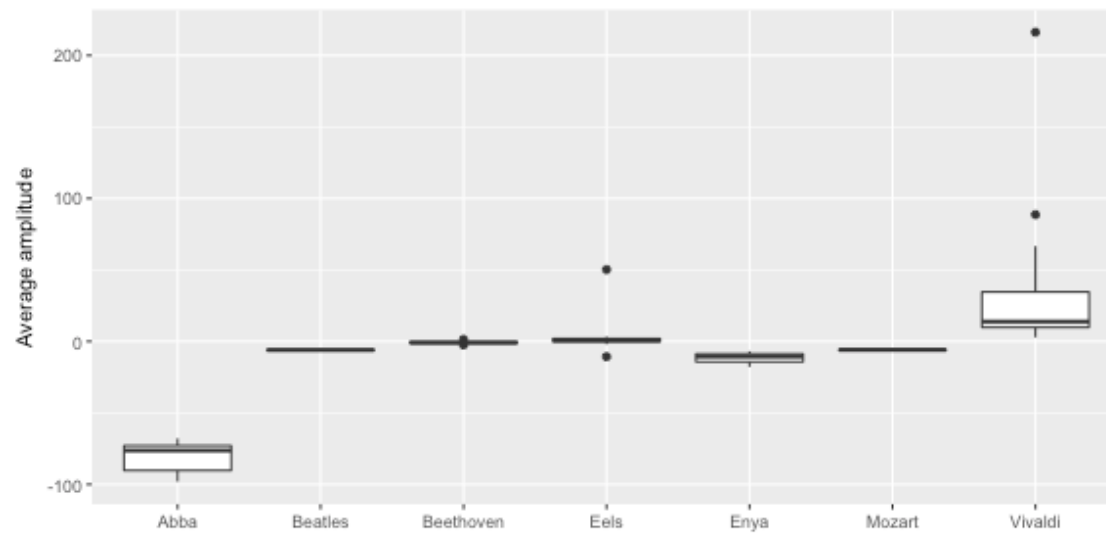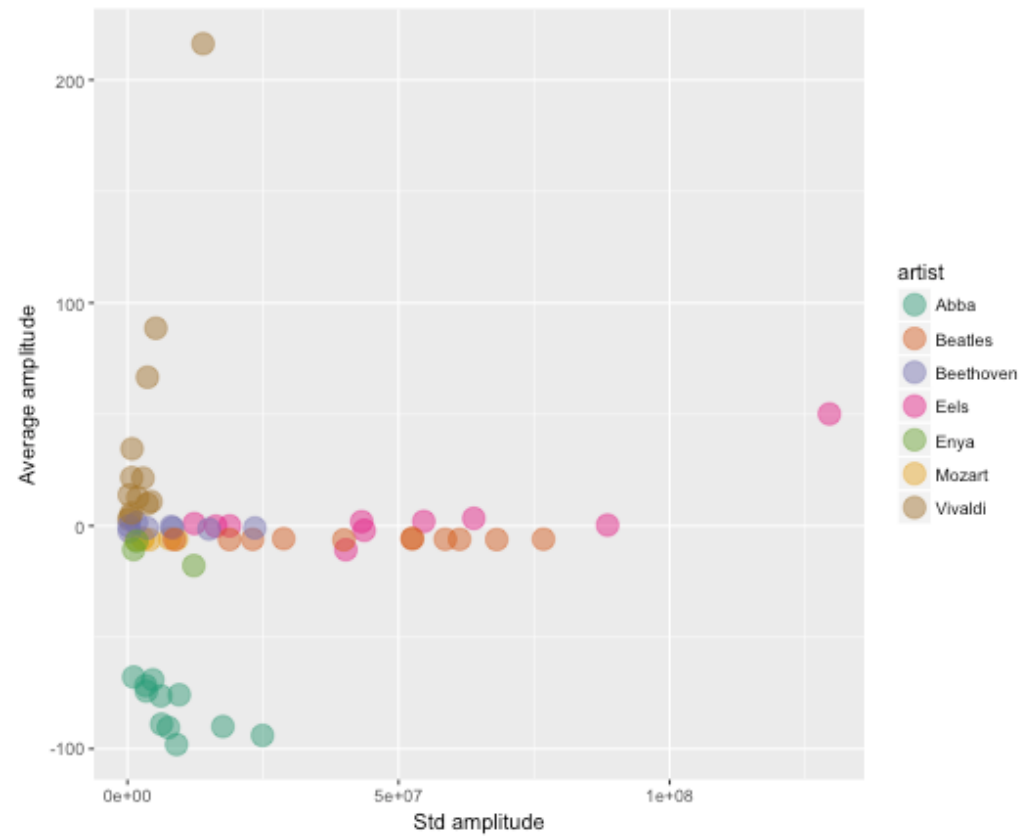Oh, same sound is on both channels! A tad drab.

Compute statistics

```
# A tibble: 2 x 5
        word           m        s    mx       mn
       <chr>       <dbl>    <dbl> <dbl>    <dbl>
1       data 0.004059919 1602.577  8393  -15386
2 statistics 0.009019820 1506.626  6601  -11026
```

My music - don't laugh

Abba is just different from everyone else!

Education

- OECD PISA survey is the world's global metric for quality, equity and efficiency in school education.
- Workforce readiness of 15-year old students
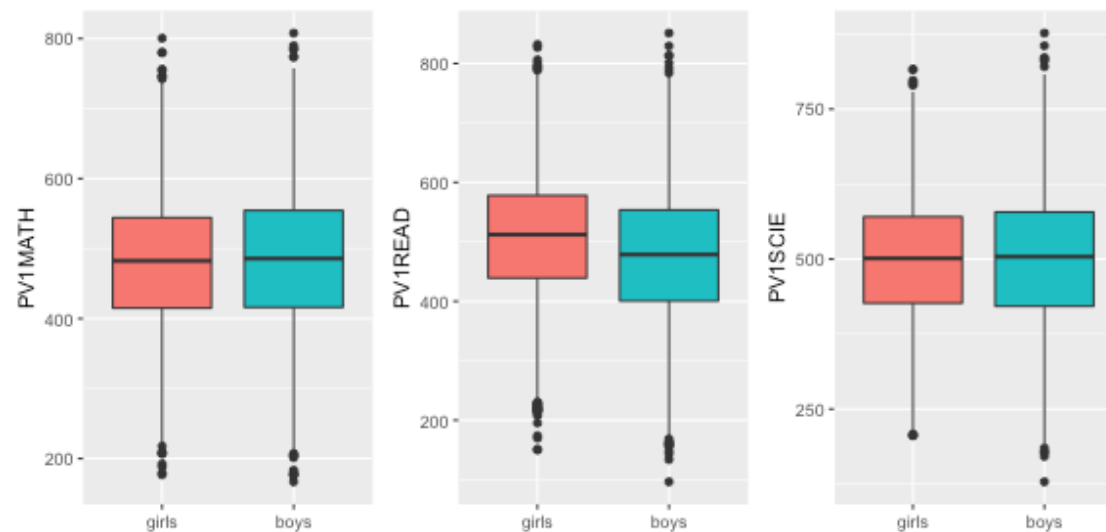- 14530 students tested in Australia in 2015
- How many schools?
- Math, reading and science tests, surveys on school and home environment, 921 variables
- Data available from http://www.oecd.org/pisa/data/

Gender differences

```
# A tibble: 2 x 4
  ST004D01T      math      read      sci
     <fctr>     <dbl>     <dbl>    <dbl>
1     girls  491.1820  518.4506  509.0075
2      boys  496.5699  486.9064  511.4427
```


On average reading scores are most different, with girls scoring substantially higher.
There is more variability from individual to individual than from boys to girls.

Your turn

Point out a couple more things that you see in terms of differences between girls and boys?
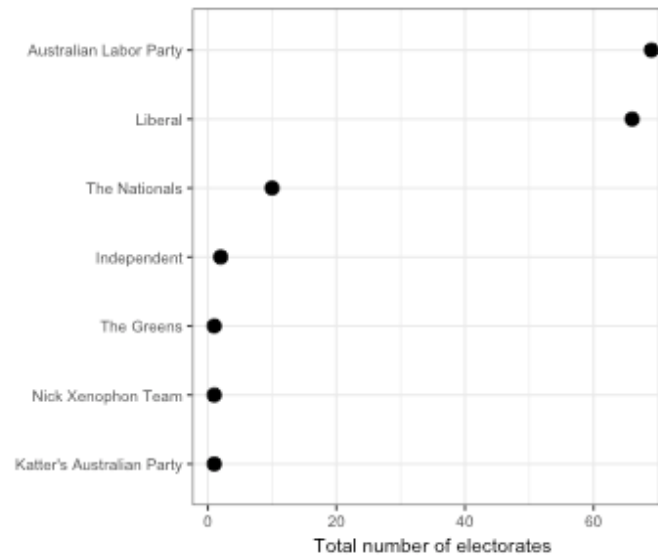
Tennis

Statistics for grand slam and top tournament matches are available through the ATP and WTA web sites. By web scraping these web sites we can pull data together to explore characteristics of different players, surfaces, ... These are compiled into the R package deuce.
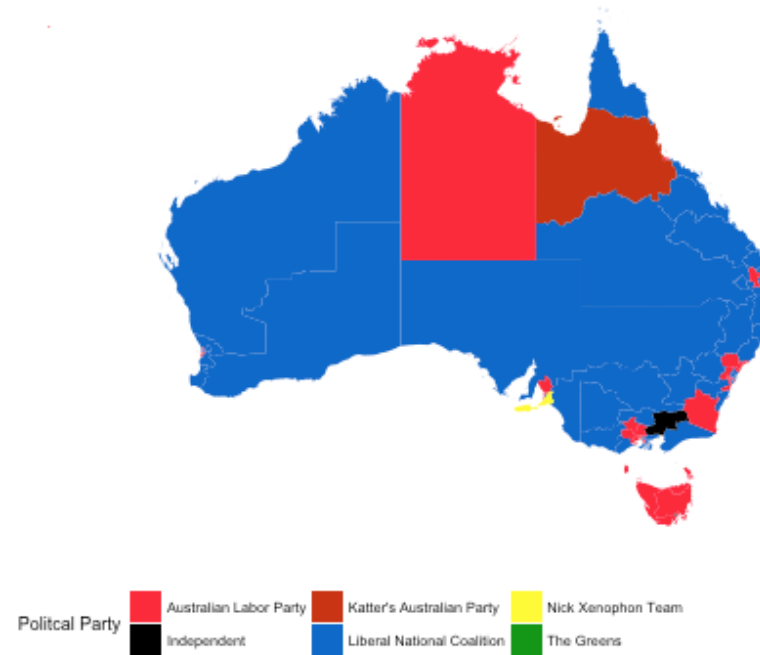
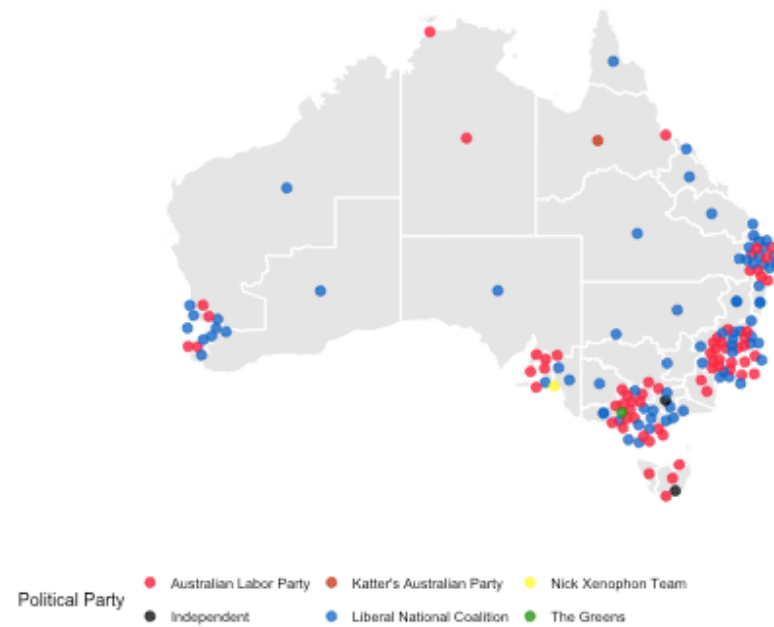Yes, there are 47 year olds playing Grand Slams! See NYTimes, 2004

Politics

The Australian Electoral Commission provides Federal election results, and the electoral map.

Map it!

Cartogram it!

Political Party
- Australian Labor Party
- Katter's Australian Party
- Nick Xenophon Team
- Independent
- Liberal National Coalition
- The Greens

Combine with Census data

Exploring the Australian Electorate with the R package, eechida

from **Di Cook**

03:31

Video link: https://vimeo.com/167367369
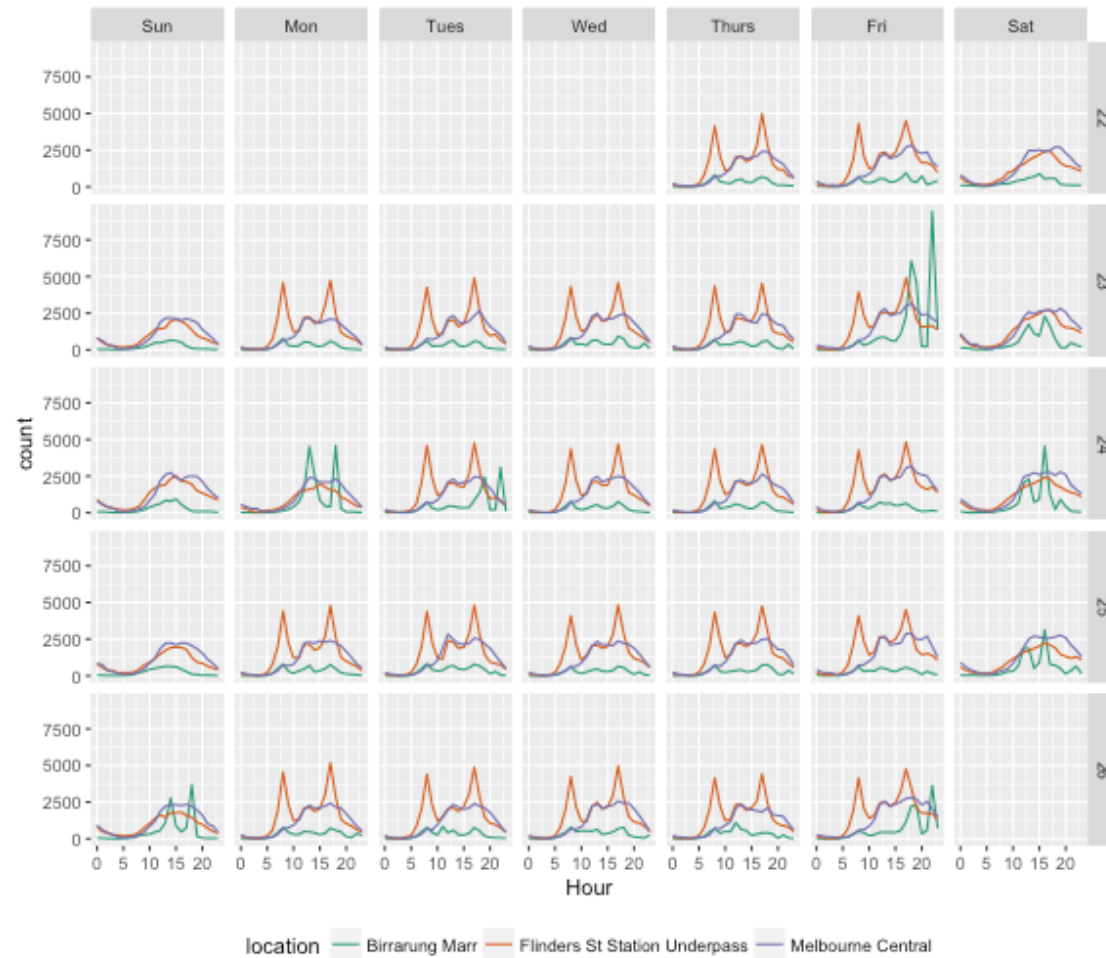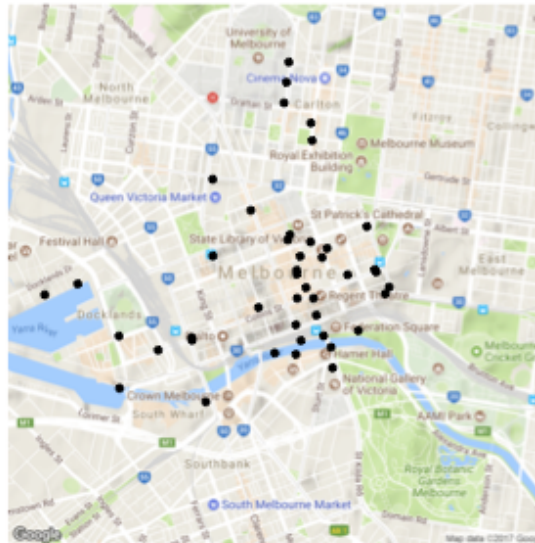
Pedestrian sensors

Each day of June

Google maps

Pull a google map and see where the sensors are located.

Airline traffic

The web site https://www.transtats.bts.gov keeps records for all commercial flights in the USA. The ontime database is interesting. You can download all records since record collection started, about 20Gb, or select months. I've pulled records for May 2017, 20Mb zip'd data.

Variables

```
"YEAR","MONTH","DAY_OF_MONTH","DAY_OF_WEEK",
"FL_DATE","UNIQUE_CARRIER","AIRLINE_ID","CARRIER",
"TAIL_NUM","FL_NUM","ORIGIN_AIRPORT_ID",
"ORIGIN_AIRPORT_SEQ_ID","ORIGIN_CITY_MARKET_ID",
"DEST_AIRPORT_ID","DEST_AIRPORT_SEQ_ID",
"DEST_CITY_MARKET_ID","CRS_DEP_TIME","DEP_TIME",
"DEP_DELAY","DEP_DELAY_NEW","DEP_DEL15",
"DEP_DELAY_GROUP","TAXI_OUT","WHEELS_OFF","WHEELS_ON",
"TAXI_IN","CRS_ARR_TIME","ARR_TIME","ARR_DELAY",
"ARR_DELAY_NEW","ARR_DEL15","CANCELLED",
"CANCELLATION_CODE","DIVERTED","CRS_ELAPSED_TIME",
"ACTUAL_ELAPSED_TIME","AIR_TIME","DISTANCE",
"CARRIER_DELAY","WEATHER_DELAY","NAS_DELAY",
"SECURITY_DELAY","LATE_AIRCRAFT_DELAY"
```

Where did my plane fly?

and in the rest of the month ...

What airports are these

AIRPORT DISPLAY_AIRPORT_NAME

AMA Rick Husband Amarillo International
AUS Austin - Bergstrom International
BWI Baltimore/Washington International Thurgood Marshall CHS Charleston
AFB/International
CLE Cleveland-Hopkins International
CMH Port Columbus International
COS City of Colorado Springs Municipal
CVG Cincinnati/Northern Kentucky International
DAY James M Cox/Dayton International
DFW Dallas/Fort Worth International
DSM Des Moines International
DTW Detroit Metro Wayne County
ELP El Paso International
IAD Washington Dulles International
IAH George Bush Intercontinental/Houston
ICT Wichita Dwight D Eisenhower National
IND Indianapolis International
JAX Jacksonville International
LBB Lubbock Preston Smith International

Cross-rates relative to USD

Pull historical exchange rates from http://openexchangerates.org/api/historical/.

Analysing tweets

Is it possible to distinguish tweets coming from Donald Trump's phone vs his staff's phone? With a twitter api you can collect all tweets between certain times, from different people, with different hashtags, ... David Robinson wrote a post during last year's US election cycle doing just this. Here's a re-creation of his analysis.

Tweets from @realDonaldTrump were collected and passed through a sentiment analysis.

Common words between devices

Sentiment analysis

Poisson test of the differences between whether it is mode likely to emerge from the Android.

Climate change

- Data is collected at a number of locations world wide.

- See Scripps Inst. of Oceanography

- Let's pull the data from the web and take a look ...

-

- Recordings from South Pole (SPO), Kermadec Islands (KER), Mauna Loa Hawaii (MLF), La Jolla Pier, California (LJO), Point Barrow, Alaska (PTB).

📊 CO2 is increasing, and it looks like it is exponential increase.

📊 The same trend is seen at every location - REALLY? Need some physics to understand this.

📊 Some stations show seasonal pattern - actually the more north the more seasonality - WHY?

R is ...

📊 Most commonly used data science software kdnuggets

📊 **Free** to use, **open source** so you can see what code is doing to your data

📊 **Extensible**: Over 11000 user contributed add-on packages currently on CRAN! Bioconductor has more than 1300 packages, and many researchers provide packages through github.

📊 **Powerful**: With the right tools, get more work done, faster, better.

📊 **Flexible**: Not a question of *can*, but *how*.

RStudio is ...

From Julie Lowndes:

If R were an airplane, RStudio would be the airport, providing many, many supporting services that make it easier for you, the pilot, to take off and go to awesome places. Sure, you can fly an airplane without an airport, but having those runways and supporting infrastructure is a game-changer.

The RStudio IDE

📊 Source editor: Docking station for multiple files, Useful shortcuts ("Knit"), Highlighting/Tab-completion, Code-checking (R, HTML, JS), Debugging features

📊 Console window: Highlighting/Tab-completion, Search recent commands

📊 Other tabs/panes: Graphics, R documentation, Environment pane, File system navigation/access, Tools for package development, git, etc

Installing packages

From CRAN

```
install.packages("learnr")
```

From bioconductor

```
source("https://bioconductor.org/biocLite.R")
biocLite("ggbio")
```

From github repos

```
devtools::install_github("earowang/sugrrants")
devtools::install_github("haleyjeppson/ggmosaic")
```

What is R Markdown?

From the R Markdown home page:

📊 R Markdown is an authoring format that enables easy creation of dynamic documents, presentations, and reports from R.

📊 It combines the core syntax of **markdown** (an easy-to-write plain text format) **with embedded R code chunks** that are run so their output can be included in the final document.

📊 R Markdown documents are fully reproducible (they can be automatically regenerated whenever underlying R code or data changes).

Open data, open source

📊 Data is available everywhere today, publicly, free

📊 Software, very powerful software, for analysis of data is available publicly, free

📊 Combined with a knowledge of mathematics and statistics empowers each of us to contribute to understand and improve our world

This course

Data preparation accounts for about 80% of the work of data scientists

Gil Press, Forbes, 2016

This is one of the least taught parts of data science, and business analytics, and yet it is what data scientists spend most of their time on. By the end of this semester, we hope that you will have the tools to be more efficient and effective in this area, so that you have more time to spend on your mining and modeling.

Follow along during class

📊 **These slides** are made in rmarkdown, using styling from the xaringan package.

📊 If you download the Rmd file of lecture notes ahead of, or during class, you can run the code chunks as I talk and check my calculations or models.

Share and share alike