Stat 480 - Syllabus discussion


Welcome to
Stat 480

What is this course about?

Data Acquisition

- data ingestion: flat files, data bases, web sites, other (binary) sources
- ethical issues

Data Exploration

- numerical and graphical summaries
- types of graphics and good visualization practices
- (simple) modeling

What is this course about? (cont'd)

Data Management

- Pipeline for data analysis: filtering, transformation, aggregation
- data (re-)shaping
- normal forms of data

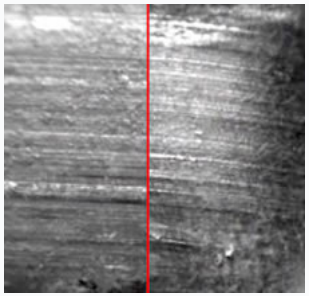
Communicating Findings

- writing reports
- web-based applets

Reproducibility/Repeatability of Findings

Data comes in many formats

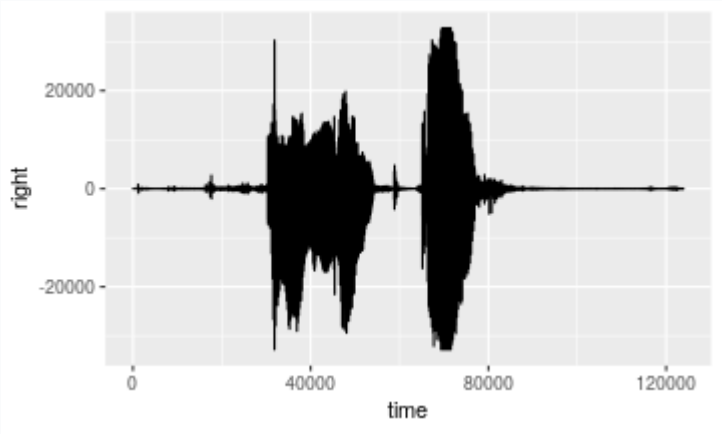
... as sound

... as image 

... in a monitoring device 

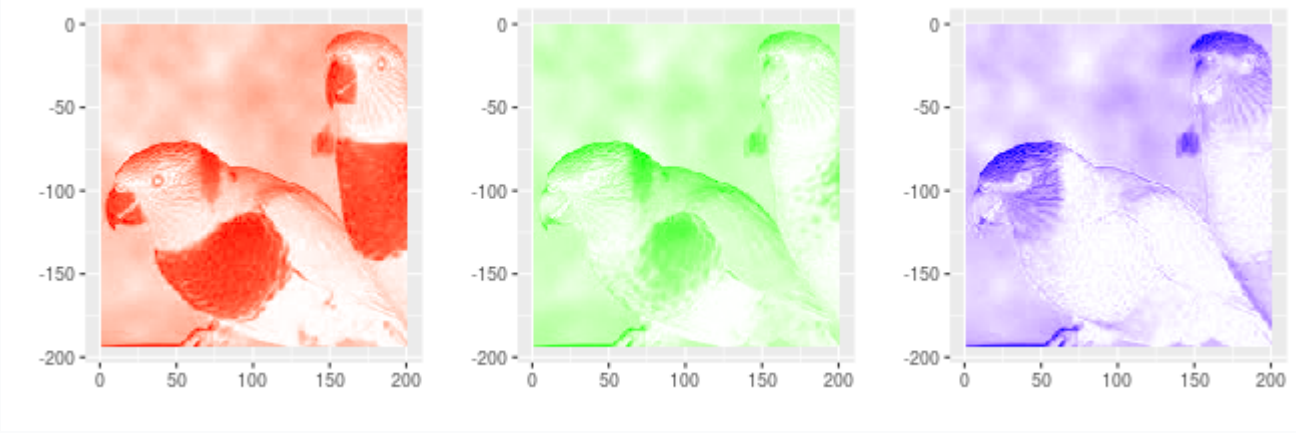
```r
library(tuneR)
ilr_class <- readWave("data/i-like-r.wav")
str(ilr_class)
```

```
## Formal class 'Wave' [package "tuneR"] with 6 slots
##   ..@ left     : int [1:123904] 2 2 2 2 -1 -1 -1 0 1 3 ...
##   ..@ right    : int [1:123904] -2 -2 -2 -1 2 1 0 0 -2 -4 ...
##   ..@ stereo   : logi TRUE
##   ..@ samp.rate: int 44100
##   ..@ bit      : int 16
##   ..@ pcm      : logi TRUE
```



```
library(jpeg)
img <- readJPEG("data/imgres.jpg")
str(img)
```

```
##  num [1:193, 1:200, 1:3] 0.235 0.235 0.239 0.239 0.243 ...
```



... what kind of birds are these?

… we will be using R for that!

R is ...

- **Free** to use, **open source** so you can see what code is doing to your data
- **Extensible**: Over 10000 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.
- **Flexible**
  - Not a question of *can*, but *how*.
- (with python) the most commonly used data science language (see kdnuggets survey)

at the end of the course you will ...

- be able to deal with complex, messy, real data
- have gained familiarity with basic data collection, storage and manipulation
- fluently reshape data into the most convenient form for analysis or reporting
- automate cleaning and analysis in R
- use graphics to explore and understand data
- communicate your findings in a reproducible form
- program a shiny web-app

Syllabus

Full syllabus is available from the course website.

Textbook (optional)

- Garrett Grolemund and Hadley Wickham: *R for Data Science*
- Hadley Wickham: *Advanced R*
- Yihui Xie: *Dynamic Documents with R and knitr*
- additional readings

Course website:

- Materials, assignments, code: https://stat480-at-isu.github.io/
- Canvas (for grades)

Grades

Component	Weight
Homework	30%
Midterm	
Exam	25%
Project	15%
Final	
Report	22.5%
Presentation	7.5%

Homework

- weekly homework assignments.
- homework assignments revise what we covered, plus synthesize some new information.
- plan to spend about 3-4h on each assignment.

Midterm

- in-class programming exam.
- open book, open note, open internet (no direct help from anyone else).
- tentatively scheduled for ??.
- sample exams will be posted as we get closer to the date.

Midterm project

- team-based project (4-5 members).
- data exploration of a given data set.
- your part: lay out data exploration, identify additional data and write up report of findings.
- scope: 2-3 weeks of homework.

Final project

- no final exam.
- team-based project (4-5 members).
- several stages:
    - identify topic and data set
    - identify line of inquiry
    - report findings in report or shiny app
    - present your project in front of the class

Attendance

I expect you to attend class: there will be a substantial amount of time devoted to 'hands-on' examples on the computers. Make use of that time!

If you have to miss class, please

(a) let me know ahead of time.

(b) make sure to catch up with the material (e.g. have a designated note taker, talk to one of your team members, ...)

Disability and Sickness

- Make sure to let me know (if you can, in advance, mini text is fine)
- If you have the flu - !!! Stay at home and take care of yourself !!!
- Keep on top of the weekly homework or you will get swamped!

Lectures

- Electronic copy of the slides are available on the website
- But you'll need to take your own notes!
- If you really want complete notes, organize a roster with others in the class
- Don't goof off on the computers!
- If you're bored, complain!

Getting Help

There's lots of ways to get help in case you are stuck:

(1) Google is your friend! in particular, stackoverflow and R help are usually great resources,

(2) ask a team member,

(3) write email to the instructor with your question

Asking a good question

... is a learned and valuable skill!

Have a look at:

- stackoverflow's Asking a good question
- R's Posting guidelines

What do you know already?

- excel?
- a programming language?
- SAS? R? R markdown?
- database theory? what is a third normal form?

What is this class about?

- very data centric
- I'd like to know what you are interested in

Sports e.g. Baseball salaries and performance

Crime data (incl. type, time, place, demographics etc.)

Health e.g. fitness statistics, or disease rates, or health care costs

Movies e.g. ratings/box office revenues from IMDB

Climate/Weather Data

Travel data, e.g. US flights

Environmental Data: Pollution, Fuel Economy, CO2 Emissions (Carbon Footprint)

Global Data: World Economy, Social indicators, ...

Anything else you can think of?

Now vote

Go to Wufoo Site to let me know your favorites and make suggestions:
https://stat480.wufoo.com/forms/topics-of-interest/