A Look back on STOR 390

4/27/17

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Where did this course come from?

Data@Carolina grant

Iain Carmichael, Brendan Brown, Varun Goel, Dylan Glotzer, Marshall Markham, Shankar Bhamidi

and many, many more:

https://idc9.github.io/stor390/course_info/acknowledgments.html

Outline

What you learned (and what you didn't)

Why it's important

Broader perspective on data science

What skills you learned

Programming in R

Working with data

Statistical modeling

Effective Communication

You learned how to **program in R**

Loops

If/else

Boolean logic

Data types

vectors, lists, strings, tibbles…

You can use **R Studio**

R, R Markdown, Shiny

Reports, data analysis, dashboards, interactive visualizations, resume, blog post, websites

http://rmarkdown.rstudio.com/gallery.html

https://shiny.rstudio.com/gallery/

You can **work with tidy data**

Visualization

ggplot, shiny

Data munging/manipulation/transformation

dplyr: select, mutate, group_by

joins: filtering, mutating, etc

Loading data

read_csv

You can **work with text data**

Regular expressions

str_match, str_extract

Natural language processing

- tidytext

unnest tokens, document term matrix, tf-idf

You have spent some **time working with data**

data.gov

Biodiversity in North Carolina

MOMA

IMDB

Bike Sharing

iPhone moment tracking

Beauty and the Beast

Harry Potter

Final projects

You know how to **acquire data** for yourself

Web scraping

rvest, SelectorGaget

APIs

geocaching with google maps

Twitter

You have seen **different types of analyses**

Exploratory

Inferential

Predictive

You can do **statistical modeling/machine learning**

Linear regression

Classification

KNN, Nearest Centroid, SVM

Clustering

K-means

Model selection/tuning

cross-validation

Feature engineering

factors, interactions, polynomial terms

You have learned about **effective communication**

General principles/advise

focus on message

adapt to the audience

Effective visual communication

static plots (ggplot), dynamic plots (Shiny)

Literate programming

R Markdown

You have done a full **data analysis**

Ask a question

Acquire data

Analyze some data

Communicate results

Higher level skills

Programming

Ability to acquire data

Identify problems that can be solved with data

Classify data problems

Communication

What you did not learn

More advanced

- programming
- statistics

Lot’s of experience

Be aware you know **enough to be dangerous**

Very easy to make **bad**, but **convincing** data driven arguments

Just because an algorithm says something does **not** imply it is meaningful/correct

Inference is hard

Lot's of great, existing statistics courses teach you inference

Experience

Critical thinking

Why these skills are important

Better understanding of

- data
- science
- technology

See potential opportunities

Empower you to do _____

Understand **strengths and limitations** of data, science and technology

What is easy?

What is hard?

What can go wrong?

Look for **potential opportunities**

Data can get at a lot of problems

Basic understanding can go a long way

The ability to work with data

empowers you to do _____ better

What ever it is you are interested in

medicine, sports, business, law, literature,

“artificial intelligence”

Broader take aways

Teach yourself

Skepticism

Yak-shaving

Problem solving

Trade-offs

Teach yourself

MOOCs

Coursera, edX, Udacidy

Textbooks

Stack exchange

Problem solving

Break up a problem into smaller sub-problems

Details

“Everyone has a plan until they get punched in the mouth.”

–Mike Tyson

Problem solving

Break up a problem into smaller sub-problems

Details

Adapt

Persistence

Be unafraid of **Yak Shaving**

Yak Shaving (noun)

Any apparently useless activity which, by allowing you to overcome intermediate difficulties, allows you to solve a larger problem.

https://en.wiktionary.org/wiki/yak_shaving

“There are three kinds of lies: lies, damned lies, and statistics.”

–Mark Twain

Be skeptical

Where did the data come from?

biases, is it representative?

Does the argument hold merit?

where might it have gone wrong

“There ain’t no such thing as a free lunch.”

–Milton Friedman

There are always **trade-offs**

Time spend writing vs. quality

More rigorous analysis vs. time/resources

The best model depends on the data

Just because you can doesn’t mean you should

Started the course with a quote from George Box

“All models are wrong but some models are useful.”

–George Box

Box quote summarizes data science

Optimism/tenacity

- Maybe we can solve this problem?

Skepticism

- Why should I believe your solution?

Science + engineering

Thanks!

What could we do to make this course better?

Stay in touch!