

# STOR 390: Introduction to Data Science

Spring 2017

Iain Carmichael ([iain@unc.edu](mailto:iain@unc.edu))



**THE UNIVERSITY**  
*of* **NORTH CAROLINA**  
*at* **CHAPEL HILL**

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

– George Box

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

– George Box

# Model

A simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions - New Oxford American Dictionary

# Model

A simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions - New Oxford American Dictionary

or

“an abstract representation of some process, be it a baseball game, an oil company’s supply chain, a foreign government’s actions or a movie theater’s attendance” - Weapons of Math Destruction

Newton's three laws of motion are  
a simple model of the universe

$$F = ma$$

Newton's three laws of motion are  
a simple model of the universe

$$F = ma$$

Special/General relativity

Newton's three laws of motion are  
a simple model of the universe

$$F = ma$$

Special/General relativity

Vast majority of physics applications use  
Newtonian mechanics



# Some people are introverts, some people are extroverts

Places people into two categories (or  
maybe on a continuum)

[https://www.ted.com/talks/susan\\_cain\\_the\\_power\\_of\\_introverts](https://www.ted.com/talks/susan_cain_the_power_of_introverts)

# Some people are introverts, some people are extroverts

Places people into two categories (or maybe on a continuum)

Fails to capture a lot about you

# Some people are introverts, some people are extroverts

Places people into two categories (or maybe on a continuum)

Fails to capture a lot about you

Helpful for understanding how people operate

# Relationship advice...

“Absence makes the heart grow fonder”

# Relationship advice...

“Absence makes the heart grow fonder”

or is it

“Out of sight, out of mind”

# Relationship advice...

“Absence makes the heart grow fonder”

or is it

“Out of sight, out of mind”



# Self driving cars use a lot of models

Where is the car on the road?

Where are other cars it going?



# Self driving cars use a lot of models



Where is the car on the road?

Where are other cars it going?

What is an object I should avoid?



# Self driving cars use a lot of models



Where is the car on the road?

Where are other cars it going?

What is an object I should avoid?

Is that a stop sign?

# What is data science?

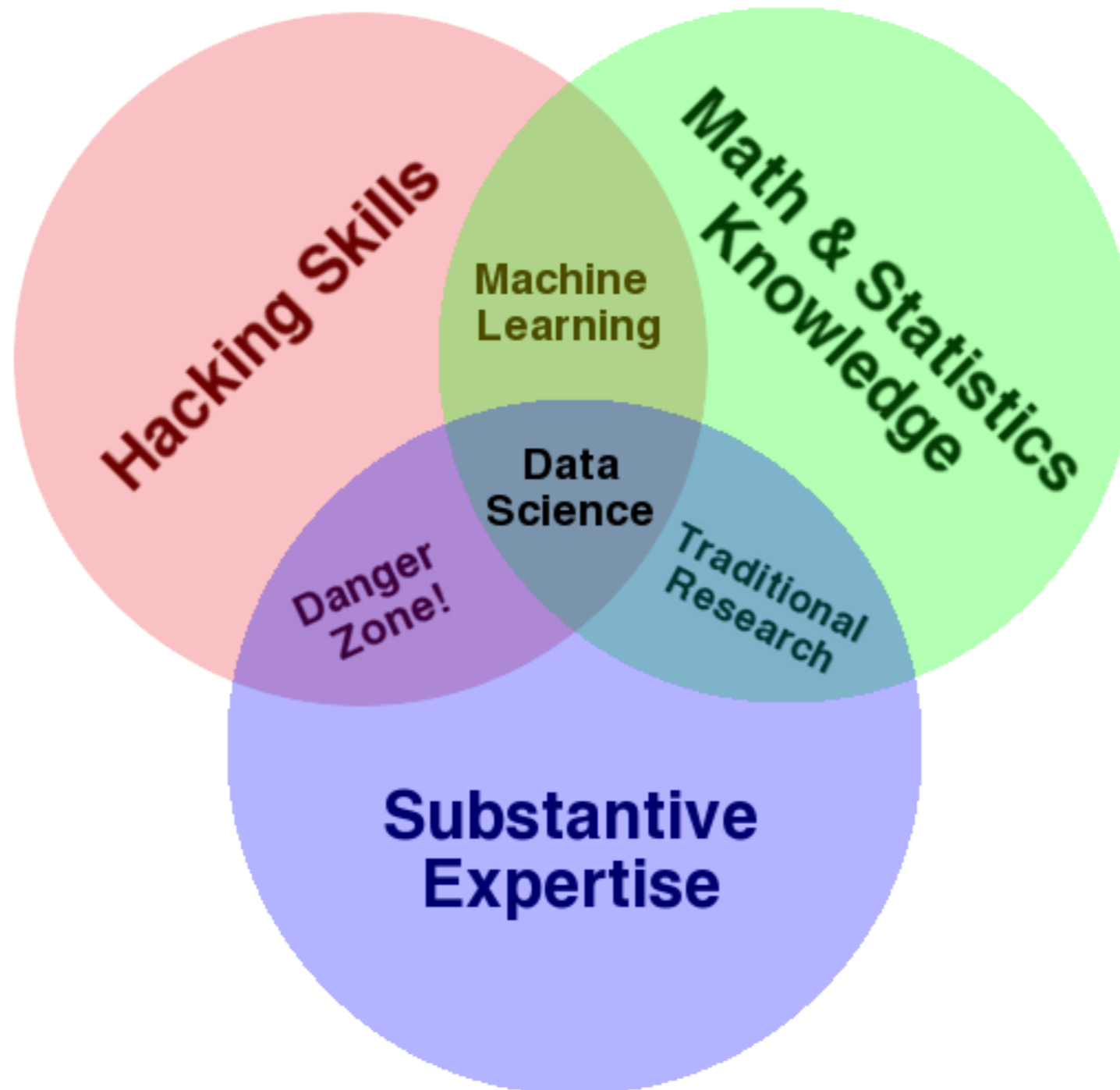
---

Brian Caffo, Jeff Leek, Roger Peng

@jtleek

www.jtleek.com

# The data science Venn Diagram



# Lots of buzz words



**Big Data** can mean a lot of things

Lots of observations

# **Big Data** can mean a lot of things

Lots of observations

Lots of variables

# **Big Data** can mean a lot of things

Lots of observations

Lots of variables

Non-standard data

- Text, images, networks

# **Big Data** can mean a lot of things

Lots of observations

Lots of variables

Non-standard data

- Text, images, networks

or that someone is trying to impress you...



# Big data = data is ubiquitous

Neuroscience

Ecommerce

Cars

Finance

Medicine

Journalism

# Big data = data is ubiquitous

Neuroscience

Ecommerce

Cars

Finance

Medicine

Journalism

Where is data absent?

Is this always a good thing?

# Use data to **understand** something

What customers are interested in my product?

Who will respond to this cancer treatment?

# Use data to **understand** something

What customers are interested in my product?

Who will respond to this cancer treatment?

“Classical” science, now applied to many areas

# Use data to **understand** something

What customers are interested in my product?

Who will respond to this cancer treatment?

“Classical” science, now applied to many  
areas

New and interesting

- problems
- datasets
- algorithms

# Use data to **do** something

Facebook can do facial recognition

Write an algorithm to beat the stock market

Program a computer to beat humans at Go

# Use data to **do** something

Facebook can do facial recognition

Write an algorithm to beat the stock market

Program a computer to beat humans at Go

More like “engineering”

# Use data to **do** something

Facebook can do facial recognition

Write an algorithm to beat the stock market

Program a computer to beat humans at Go

More like “engineering”

Same algorithms, different goals



# Course Information

STOR 390: Introduction to Data Science

- TuTh: 5:00 - 6:15 pm
- Greenlaw 101

Instructor: Iain Carmichael ([iain@unc.edu](mailto:iain@unc.edu))

Teaching Assistants:

- Varun Goel ([varung@live.unc.edu](mailto:varung@live.unc.edu))
- Brendan Brown ([bb@live.unc.edu](mailto:bb@live.unc.edu))

# Website

**<https://idc9.github.io/stor390/>**

# Iain Carmichael

BA in Math and Physics from Cornell

PhD candidate in Statistics

Gamalon Machine Intelligence

Research

- networks, probability and high-dimensional statistics
- neuroscience and law

# Brendan Brown

PhD student in statistics

2+ years experience in data science for the UNC system office

- visualization
- presentation
- forecasting, modeling, with large datasets

# Varun Goel

PhD candidate in Geography

Data Scientist at Indian School of Business, Hyderabad - Involved in informing agricultural public policy through data science

Current Research

Spatial Statistics, GIS, Disease ecology,

Population Health

# Waitlist...

The waitlist is very long

Sign up at: <https://stat-or.unc.edu/waitlist/>

I do not control the waitlist

# Course organization

Homework: 35%

- ~ 4 data analyses

Labs: 35%

- Start in class, due the next class

Class participation: 15%

Final project: 15%

Extra Credit: up to 5%

# Group work for homework and final project

The instructor will assign teams

Final grade will be adjusted by peer ratings

As a last resort a team may fire an uncooperative member



# Final Project

Novel data analysis

- get a data set
- Analyze it
- Write a blog post

In a team

# Goals and learning objective

core R programming skills

statistical and programming best practices

communication, problem solving, teamwork

literate programming

- R Markdown

# Goals and learning objective

core R programming skills

statistical and programming best practices

communication, problem solving, teamwork

literate programming

- R Markdown

# Goals and learning objective

core R programming skills

statistical and programming best practices

communication, problem solving, teamwork

literate programming

- R Markdown

# Goals and learning objective

core R programming skills

statistical and programming best practices

communication, problem solving, teamwork

literate programming

- R Markdown

# Topics (see syllabus)

Visualization with ggplot2

Data manipulation dplyr

R Markdown

Programming e.g. functions, loops, if/else, comments

Tidy Data, relational data, data import

Reproducibility

Strings/regular expressions

EDA

Classification, clustering, regression

Web scraping

Text data and Natural Language Processing

# Additional topics (if time permits)

interactive graphics with shiny

effective visualization for communication

date/time data

github

GIS data

data privacy/ethics

# Google is your best friend as a programmer

Lots of resources on the course website

**[https://idc9.github.io/stor390/course\\_info/references.html](https://idc9.github.io/stor390/course_info/references.html)**





# Install R and R Studio

[https://idc9.github.io/stor390/notes/getting\\_started/getting\\_started.html](https://idc9.github.io/stor390/notes/getting_started/getting_started.html)



<http://rhrv.r-forge.r-project.org/>



<https://www.rstudio.com/about/trademark/>

# R vs. Python

Better to be really good an one then mediocre at both

Both and pluses and minuses

# Class survey: how will this information be used against you?

Please fill out the survey

I may use major/year information to make teams

This data will not be released outside the class

# First lab: data.gov



[DATA](#) [TOPICS](#) - [IMPACT](#) [APPLICATIONS](#) [DEVELOPERS](#) [CONTACT](#)

## The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and [more](#).

**GET STARTED**

SEARCH OVER **166,943 DATASETS**



### BROWSE TOPICS



Agriculture



Climate



Consumer



Environment



Education



Energy



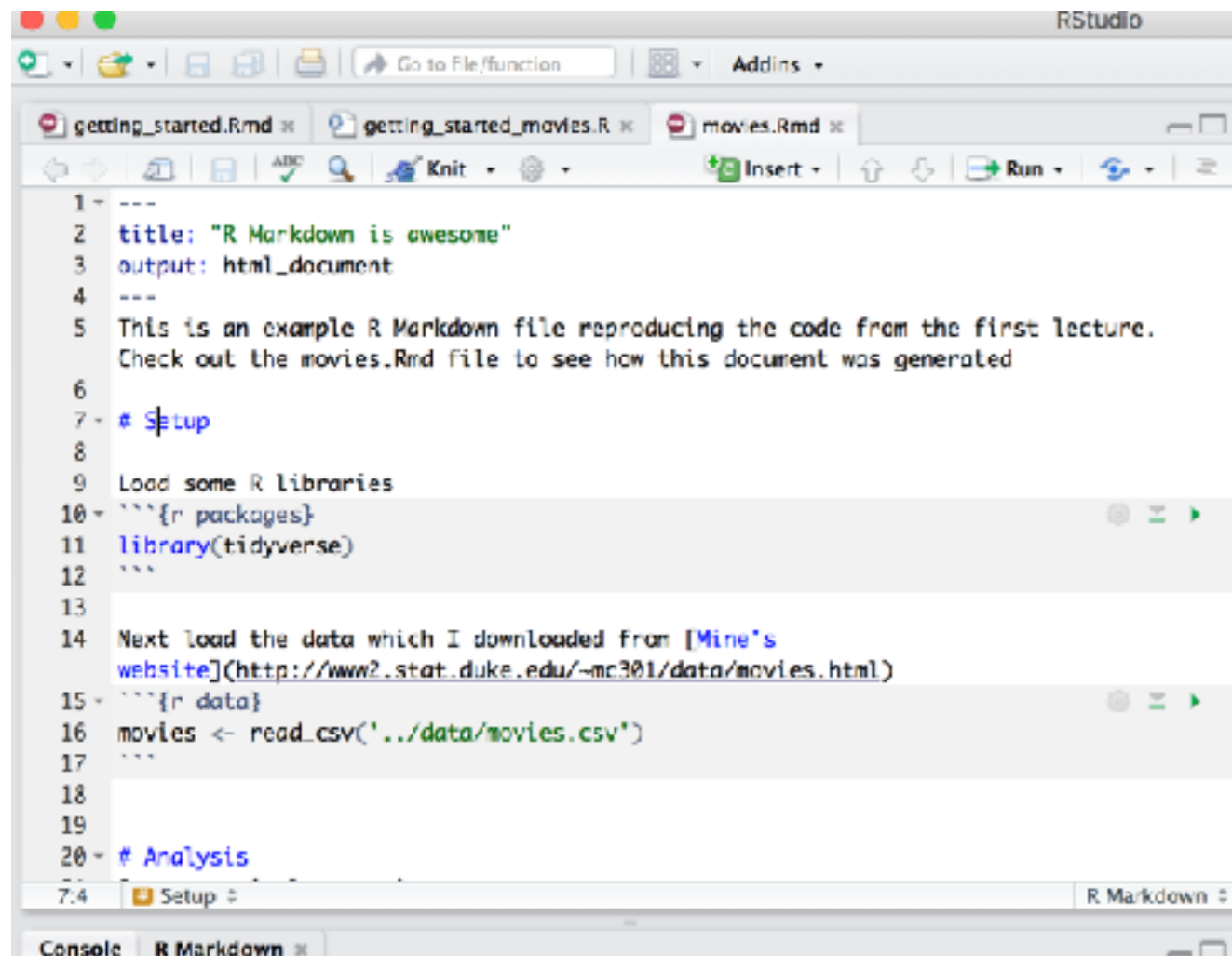
Economic

# Write code for humans, not computers

literate programming

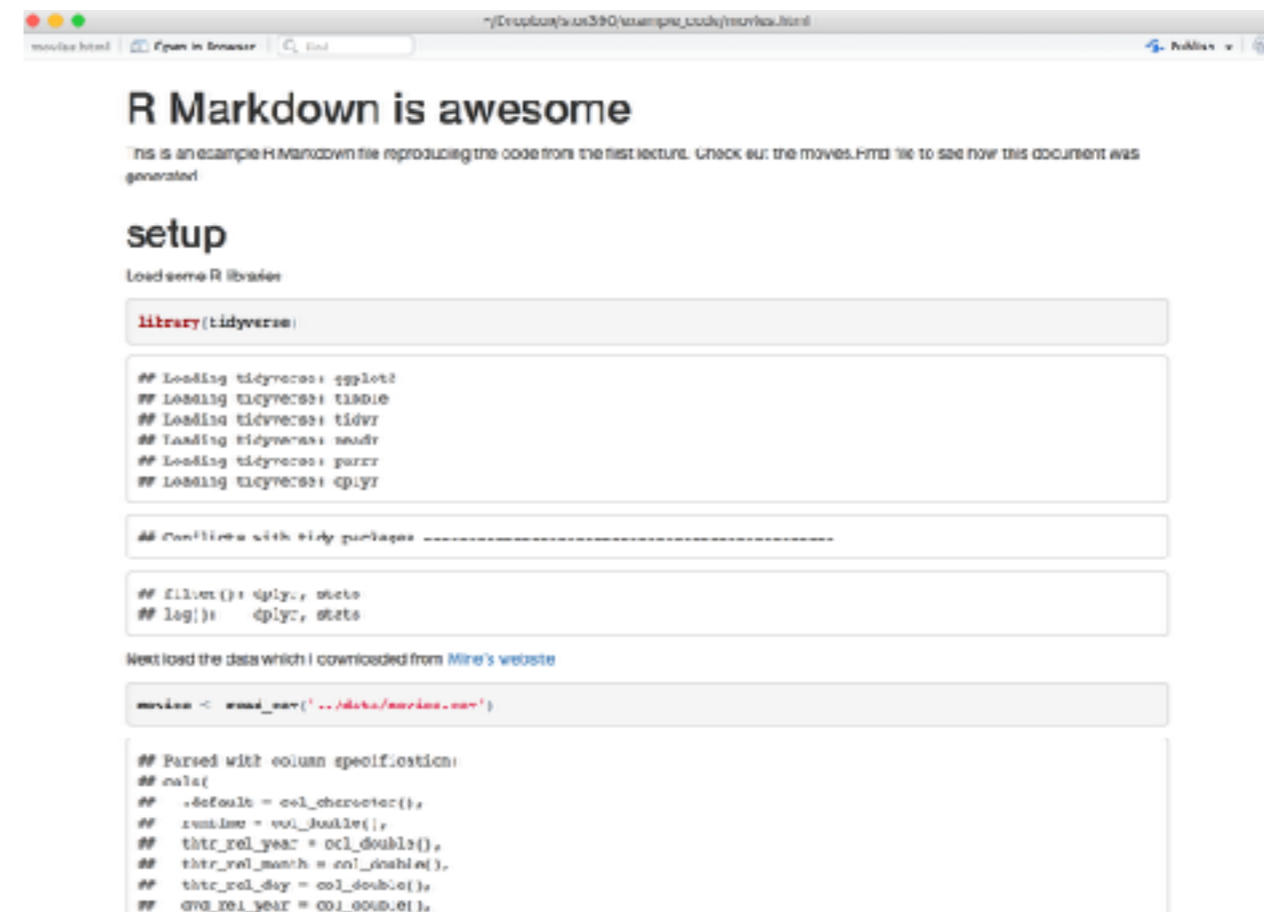
- <http://brandonrose.org/clustering>
- <https://cran.r-project.org/web/packages/tidyttext/vignettes/tidyttext.html>
- [https://github.com/idc9/brain-networks/blob/master/explore\\_igraph/EDA.ipynb](https://github.com/idc9/brain-networks/blob/master/explore_igraph/EDA.ipynb)

# R Markdown is awesome



```
1 ---
2 title: "R Markdown is awesome"
3 output: html_document
4 ---
5 This is an example R Markdown file reproducing the code from the first lecture.
6 Check out the movies.Rmd file to see how this document was generated
7
8 # Setup
9 Load some R libraries
10 ```{r packages}
11 library(tidyverse)
12 ```
13
14 Next load the data which I downloaded from [Mine's
15 website](http://www2.stat.duke.edu/~mc301/data/movies.html)
16 ```{r data}
17 movies <- read_csv("../data/movies.csv")
18 ```
19
20 # Analysis
```

7.4 Setup R Markdown



```
R Markdown is awesome
This is an example R Markdown file reproducing the code from the first lecture. Check out the movies.Rmd file to see how this document was generated
setup
Load some R libraries
library(tidyverse)
# Loading tidyverse: ggplot2
# Loading tidyverse: tidyr
# Loading tidyverse: tibble
# Loading tidyverse: tidier
# Loading tidyverse: readr
# Loading tidyverse: purrr
# Loading tidyverse: dplyr
## Conflicts with tidy packages -----
# filter() dplyr, stats
# log() dplyr, stats
Next load the data which I downloaded from Mine's website
movies <- read_csv("../data/movies.csv")
## Parsed with column specification:
## cols:
##   $default = col_character(),
##   runtime = col_double(),
##   ttr_col_year = col_double(),
##   ttr_col_month = col_double(),
##   ttr_col_day = col_double(),
##   avg_col_year = col_double()
```