Gov 50: 1. Introduction

Matthew Blackwell

Harvard University

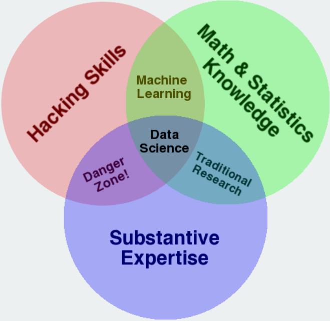
Roadmap

1. Welcome and Motivation

2. Course Details

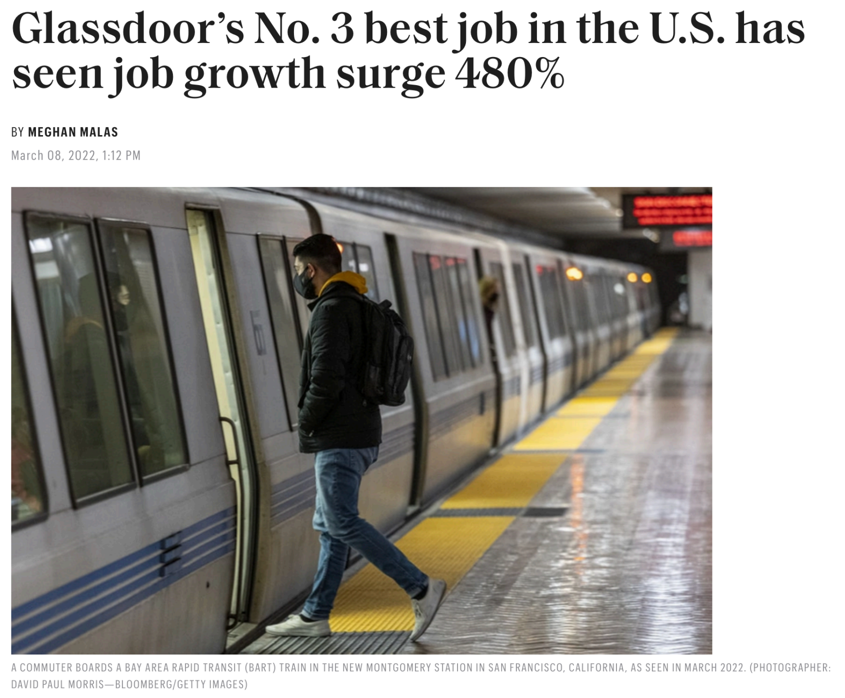
1/ Welcome and Motivation

What is data science?



- **Data science**: wrangling, visualizing, and analyzing data to understand the world
- Who does data science? Tech companies, non-tech companies, nonprofits, governments.

Credit: Drew Conway



Glassdoor’s No. 3 best job in the U.S. has seen job growth surge 480%

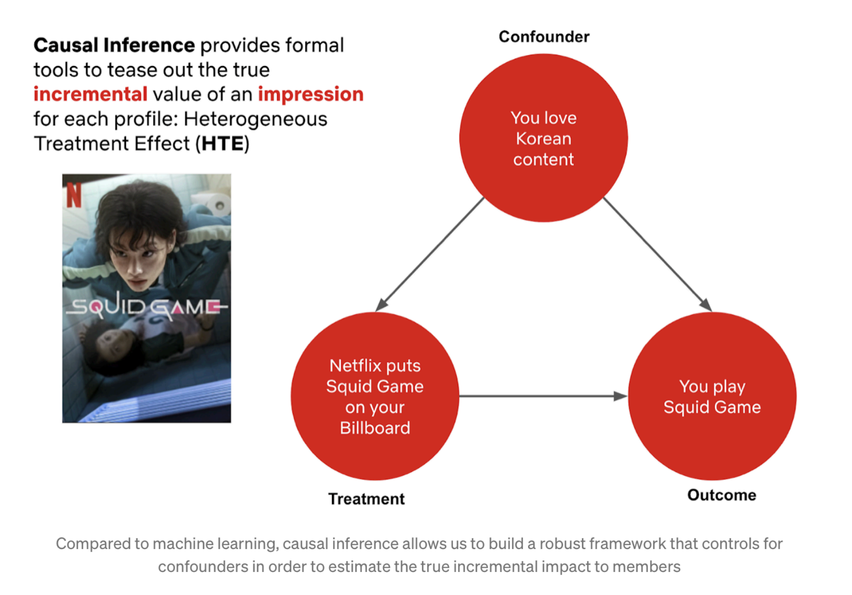
BY **MEGHAN MALAS**

March 08, 2022, 1:12 PM

A COMMUTER BOARDS A BAY AREA RAPID TRANSIT (BART) TRAIN IN THE NEW MONTGOMERY STATION IN SAN FRANCISCO, CALIFORNIA, AS SEEN IN MARCH 2022. (PHOTOGRAPHER: DAVID PAUL MORRIS—BLOOMBERG/GETTY IMAGES)

What problems are data scientists working on?

Causality



Causal Inference provides formal tools to tease out the true **incremental** value of an **impression** for each profile: Heterogeneous Treatment Effect (**HTE**)

Confounder

You love Korean content

Treatment

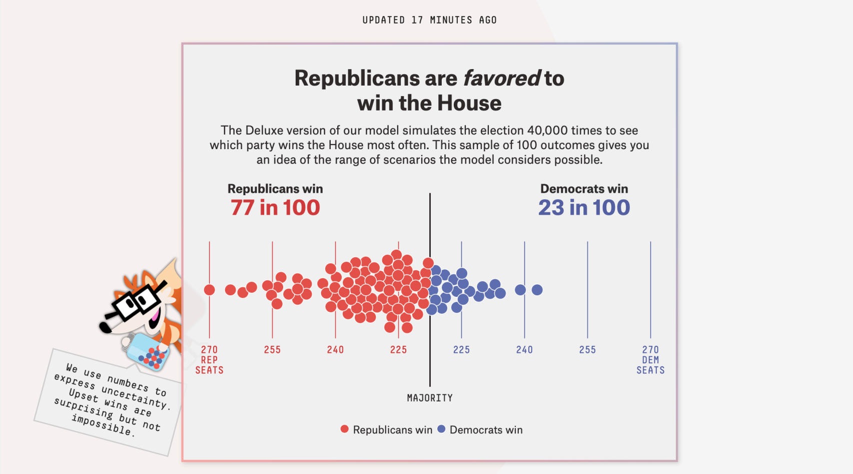
Netflix puts Squid Game on your Billboard

Outcome

You play Squid Game

Compared to machine learning, causal inference allows us to build a robust framework that controls for confounders in order to estimate the true incremental impact to members

Prediction



UPDATED 17 MINUTES AGO

Republicans are *favored* to win the House

The Deluxe version of our model simulates the election 40,000 times to see which party wins the House most often. This sample of 100 outcomes gives you an idea of the range of scenarios the model considers possible.

Republicans win
77 in 100

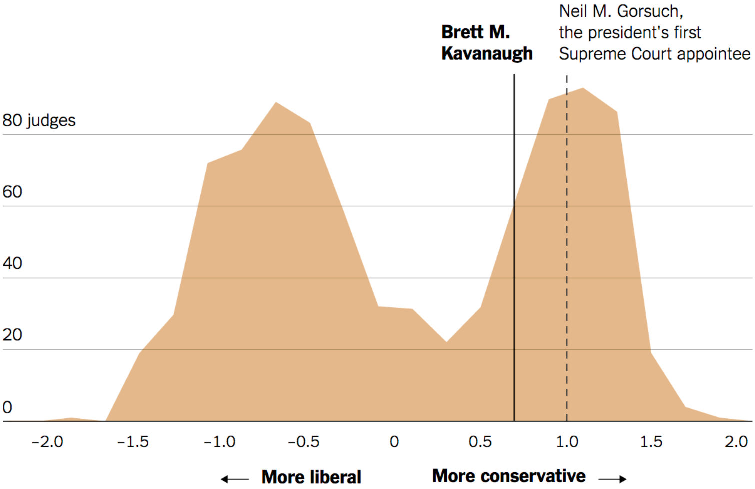
Democrats win
23 in 100

● Republicans win ● Democrats win

We use numbers to express uncertainty. Upset wins are surprising but not impossible.

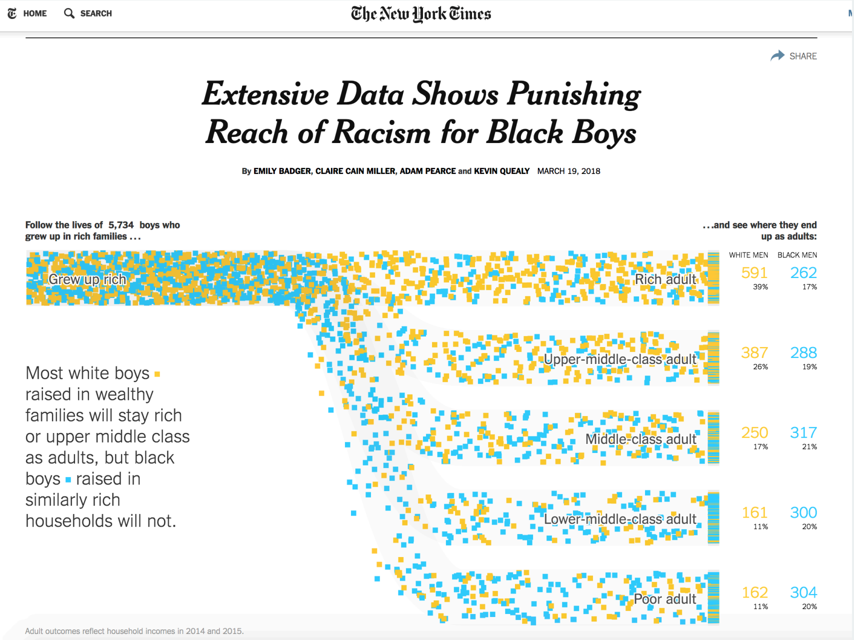
Measurement

How Kavanaugh's Ideology Compares With Other Federal Judges



Based on the campaign finance scores of all current and former federal district and court of appeals judges nominated since 1980. | Source: Database on Ideology, Money in Politics, and Elections; Adam Bonica, Stanford University Department of Political Science; Maya Sen, Harvard University, Kennedy School of Government; Adam Chilton and Kyle Rozema, University of Chicago Law School.

Understanding the socioeconomic world



Extensive Data Shows Punishing Reach of Racism for Black Boys

By **EMILY BADGER, CLAIRE CAIN MILLER, ADAM PEARCE** and **KEVIN QUEALY** MARCH 19, 2018

Follow the lives of 5,734 boys who grew up in rich families . . .

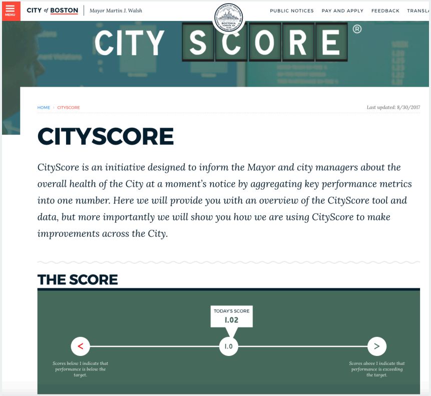
. . .and see where they end up as adults:

Grew up rich	WHITE MEN		BLACK MEN	
Rich adult	591	39%	262	17%
Upper-middle-class adult	387	26%	288	19%
Middle-class adult	250	17%	317	21%
Lower-middle-class adult	161	11%	300	20%
Poor adult	162	11%	304	20%

Most white boys ■ raised in wealthy families will stay rich or upper middle class as adults, but black boys ■ raised in similarly rich households will not.

Adult outcomes reflect household incomes in 2014 and 2015.

Making government work better



HOME › CITYSCORE

Last updated: 8/30/2017

CITYSCORE

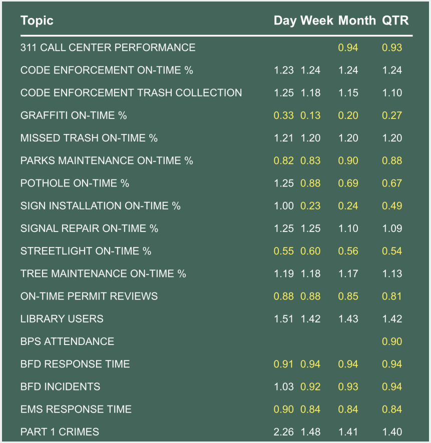
CityScore is an initiative designed to inform the Mayor and city managers about the overall health of the City at a moment's notice by aggregating key performance metrics into one number. Here we will provide you with an overview of the CityScore tool and data, but more importantly we will show you how we are using CityScore to make improvements across the City.

THE SCORE

TODAY'S SCORE 1.02

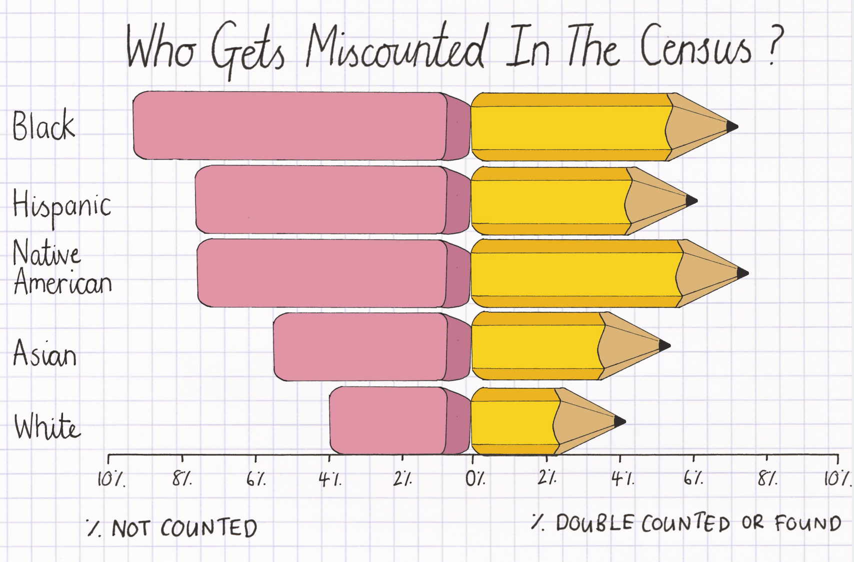
Scores below 1 indicate that performance is below the target.

Scores above 1 indicate that performance is exceeding the target.

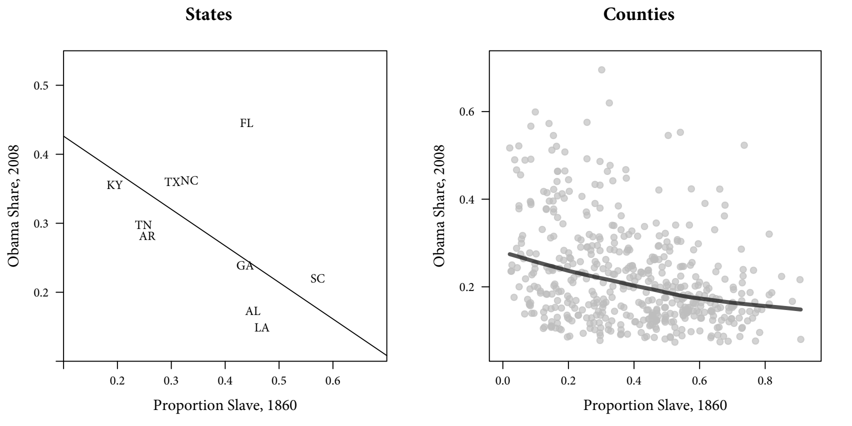


Topic	Day	Week	Month	QTR
311 CALL CENTER PERFORMANCE			0.94	0.93
CODE ENFORCEMENT ON-TIME %	1.23	1.24	1.24	1.24
CODE ENFORCEMENT TRASH COLLECTION	1.25	1.18	1.15	1.10
GRAFFITI ON-TIME %	0.33	0.13	0.20	0.27
MISSED TRASH ON-TIME %	1.21	1.20	1.20	1.20
PARKS MAINTENANCE ON-TIME %	0.82	0.83	0.90	0.88
POTHOLE ON-TIME %	1.25	0.88	0.69	0.67
SIGN INSTALLATION ON-TIME %	1.00	0.23	0.24	0.49
SIGNAL REPAIR ON-TIME %	1.25	1.25	1.10	1.09
STREETLIGHT ON-TIME %	0.55	0.60	0.56	0.54
TREE MAINTENANCE ON-TIME %	1.19	1.18	1.17	1.13
ON-TIME PERMIT REVIEWS	0.88	0.88	0.85	0.81
LIBRARY USERS	1.51	1.42	1.43	1.42
BPS ATTENDANCE				0.90
BFD RESPONSE TIME	0.91	0.94	0.94	0.94
BFD INCIDENTS	1.03	0.92	0.93	0.94
EMS RESPONSE TIME	0.90	0.84	0.84	0.84
PART 1 CRIMES	2.26	1.48	1.41	1.40

Combining art and data to inform



Understanding how the past matters



2/ Course Details

About me

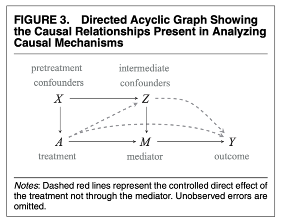
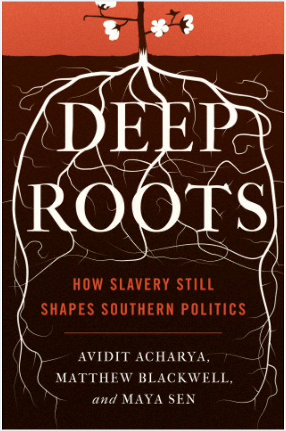


FIGURE 3. Directed Acyclic Graph Showing the Causal Relationships Present in Analyzing Causal Mechanisms

Notes: Dashed red lines represent the controlled direct effect of the treatment not through the mediator. Unobserved errors are omitted.



DEEP ROOTS
HOW SLAVERY STILL SHAPES SOUTHERN POLITICS
AVIDIT ACHARYA, MATTHEW BLACKWELL, and MAYA SEN

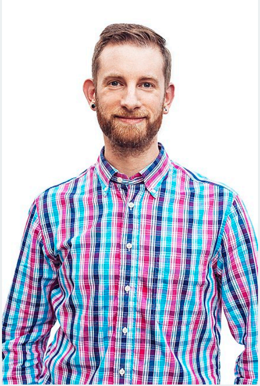
What will you learn in this class?

- Summarize and visualize data
- Wrangle messy data into tidy forms
- Evaluate claims about causality
- Be able to use linear regression to analyze data
- Understand uncertainty in data analysis and how to quantify it
- Use professional tools like R, RStudio, git, and GitHub

Teaching philosophy

- Deliberate pacing and tons of support.
- Emphasize intuition and computational approaches over mathematical equations.
- Practice, practice, practice.

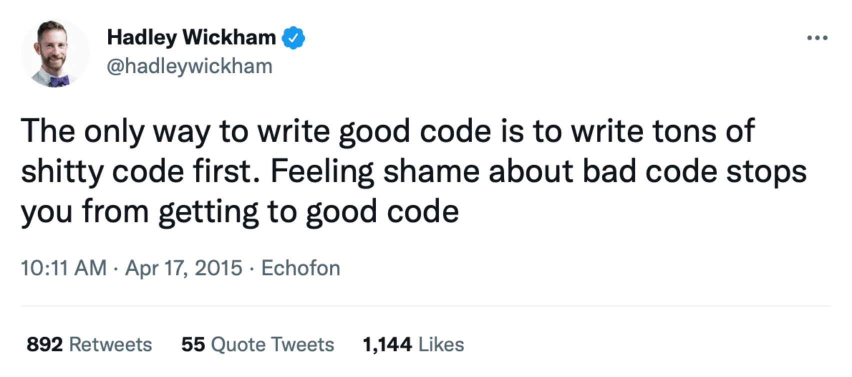
Pep talk, part I



Hadley Wickham (chief data scientist at RStudio)

It's easy when you start out programming to get really frustrated and think, "Oh it's me, I'm really stupid," or, "I'm not made out to program." But, that is absolutely not the case. Everyone gets frustrated. I still get frustrated occasionally when writing R code. It's just a natural part of programming. So, it happens to everyone and gets less and less over time. Don't blame yourself. Just take a break, do something fun, and then come back and try again later.

Pep talk, part II



Hadley Wickham
@hadleywickham

The only way to write good code is to write tons of shitty code first. Feeling shame about bad code stops you from getting to good code

10:11 AM · Apr 17, 2015 · Echofon

892 Retweets **55** Quote Tweets **1,144** Likes

Should I take this course?

- Prerequisites: **NONE** (no prior coding, statistics, data science)
- Gov 50 fulfills Gov methods requirement, data science track, and QRD
- Material useful to students interested in political science, sociology, economics, public policy, health policy, and many other fields in the social sciences.

Class meetings

- Lectures:
    - Broad coverage of the course material.
    - Coding demonstrations (follow along with your laptop!)
    - Slides/videos will be posted to Canvas shortly before lecture.
- Section:
    - Guided practice through problems and concepts led by our amazing TFs.
    - Material in section will closely mirror assignments.
- Optional speaker series with industry data scientists, TBA!

Teaching fellows

Angelo Dagonel

Dorothy Manevich

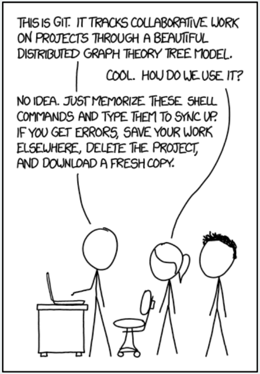
Sooahn Shin

Dominic Valentino

Computing

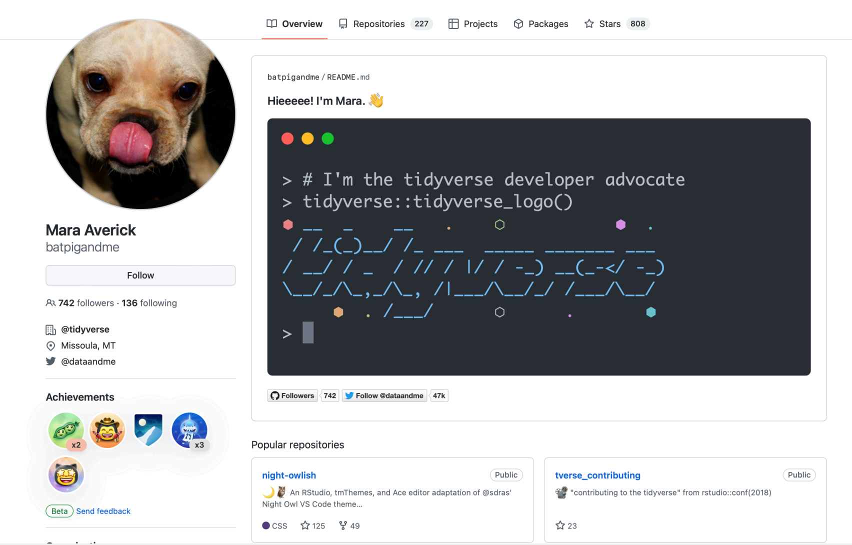
- We’ll use the R statistical environment to analyze data
  - It’s free
  - Extremely popular for data analysis
  - Academics, 538, NYT, Facebook, Google, Twitter, nonprofits, governments all use R.
  - Huge benefit to your resume to have R skills.
- Interface with R via a program called RStudio
- Problem Set 0 on the website helps get everything installed.
- Lots of help in section, study halls, office hours.

git and GitHub


THIS IS GIT. IT TRACKS COLLABORATIVE WORK ON PROJECTS THROUGH A BEAUTIFUL DISTRIBUTED GRAPH THEORY TREE MODEL.
COOL. HOW DO WE USE IT?
NO IDEA. JUST MEMORIZE THESE SHELL COMMANDS AND TYPE THEM TO SYNC UP. IF YOU GET ERRORS, SAVE YOUR WORK ELSEWHERE, DELETE THE PROJECT, AND DOWNLOAD A FRESH COPY.

- Other core tools: git and GitHub
    - **Version control system**: an archive of project versions.
    - Allows you to revert back to old versions easily
    - Makes collaboration much more mangeable.
- Will feel very odd at first, but you git used to it
- Why learn this now?
    - Knowing git/GitHub is a huge plus for data jobs.
    - Your GitHub profile can showcase your amazing new skills with data!

Sample GitHub profile



Mara Averick
batpigandme

742 followers · 136 following

@tidyverse
Missoula, MT
@dataandme

Achievements

batpigandme / README.md

Hieeeee! I'm Mara. 👋

```
> # I'm the tidyverse developer advocate
> tidyverse::tidyverse_logo()
⬢ __  _    __   .    ⬡           ⬢  .
 / /_(_)__/ /_ ___  ____ _______ ___
/ __/ / _  / // / |/ / -_) __(_-</ -_)
\__/_/\_,_/\_, /|___/\__/_/ /___/\__/
     ⬢  .  /___/      ⬡      .       ⬢
> 
```

Popular repositories

night-owlish Public
🌙🦉 An RStudio, tmThemes, and Ace editor adaptation of @sdras' Night Owl VS Code theme...
● CSS ☆ 125 49

tverse_contributing Public
"contributing to the tidyverse" from rstudio::conf(2018)
☆ 23

GovCodes workshops

- Gov department providing supplemental GovCodes workshops to provide additional computing practice.
- First meeting: tomorrow! Be on the lookout for a sign-up email.
  - Topic: getting everything installed and working on your computer!
  - Good to attend if Problem Set 0 is giving you trouble.

Textbook

- 3 primary textbooks (links on syllabus):
  - Modern Dive (free online)
  - "Quantitative Social Science: An Introduction in tidyverse" by Kosuke Imai (not free)
  - Introduction to Modern Statistics (free online)
- We'll move back and forth.
- Sometimes same material in two/three different books. Choose which helps most!

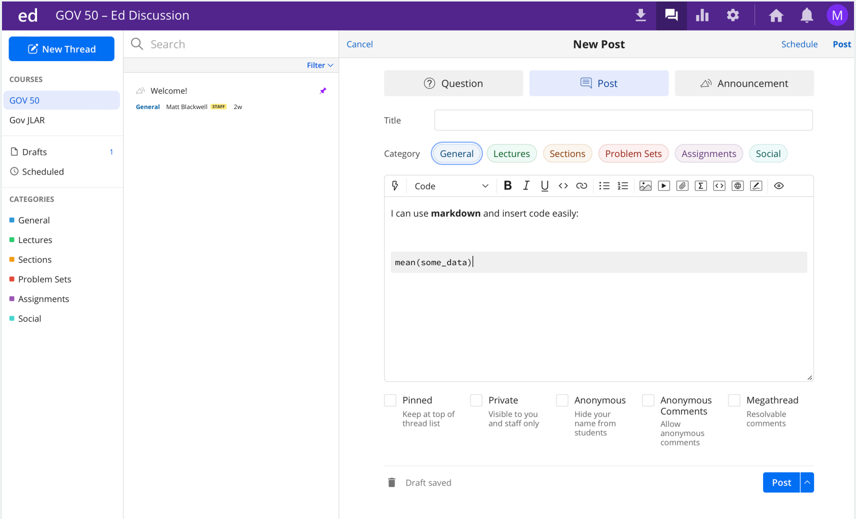
Assignments

- Roughly weekly homeworks throughout semester
    - Posted on Thursday morning, due following Wednesday.
    - Dates on syllabus
    - Lowest score dropped.
- Two take-home “exams” which are just HWs done by yourself.
- Final project: a data essay
    - Find data, pose a research question, answer it using data.
    - Submitted as a public GitHub repository and website
    - First item in your public data portfolio

Tutorials

- Getting practice with R can be overwhelming, so we'll introduce new skills through online tutorials.
- Guided practice on R, helping to introduce new concepts.
    - Low stakes/stress: graded simply on completion.
    - Due on Monday nights
- Lecture/HW won't be the first time you're trying some code!

Ed discussion board



Welcome!

General Matt Blackwell STAFF 2w

I can use **markdown** and insert code easily:

```
mean(some_data)
```

Grades

- Grade breakdown as follows:
  - R tutorials (10% of final grade)
  - Homeworks (40% of final grade)
  - Exams (30% of final grade)
  - Final project (20% of final grade)
- Final grade is curved
- **Bump-up**: we bump up grades of students close to the cutoff who make valuable contributions to the course.

Study Halls

- Study Halls: a place to work on Gov 50 and get help.
    - Will happen weekly, exact number of hours will depend on enrollment.
    - Peer tutors with experience in statistics and R will be on hand to help you if you get stuck or have question.
    - Best to come in groups and work together, grab a tutor when stuck.
- Bottom line: **we want you to succeed in this class!**

What should you do today?

- Try to get everything set up on your computer (Problem Set 0)
- Start Tutorial 1 on basics of R and data visualization
    - Can be done on the web before installing R on your computer.
- Respond to sign-up requests for GovCodes and section times.
- **Tell your friends**: data science is more fun with friends along for the journey.