If you walked in here feeling sick,
please turn around and walk back out.
Podcast recording of this class is turned on

In-person illness policy

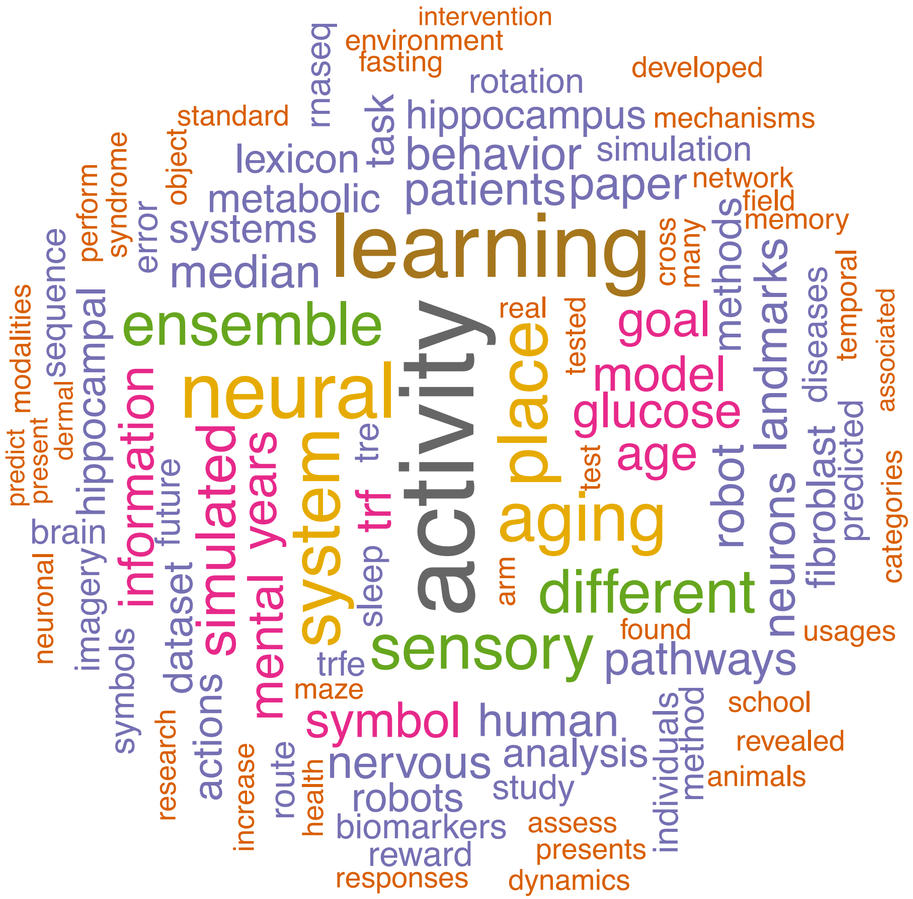
Please do not attend any in-person activity (lecture/section/office hours) if you are feeling ill, especially if you are sneezing/coughing and have a fever. If you feel mildly ill but without sneezing/coughing, or if you have bad allergies, then you may come to in-person events while wearing a well-fitting mask.

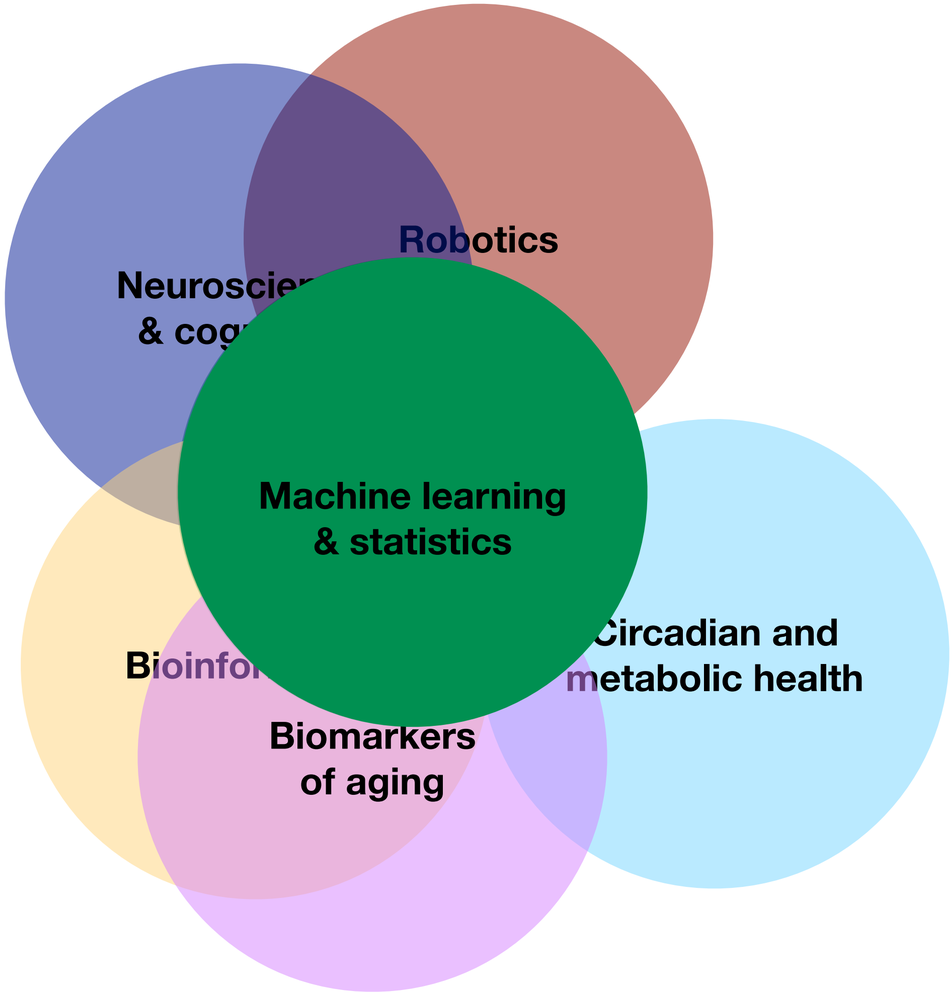
Welcome to COGS 108!

Data Science in Practice

Jason G. Fleischer, PhD Dept. of Cognitive Science UC San Diego https://jgfleischer.com

A bit about me






U.S. COAST GUARD

Photo by Evgeni Zotov

Photo by Sigfrid Lundberg

photo by Paul Domenick

Waitlist

- There is almost none this quarter for the first time in years
- For those few who are WL’d just move to another section
- You can go to another section than the one you’re signed up for as long as you’re not taking away a seat for someone who’s supposed to be there

Course Objectives

- Formulate a plan for and complete a data science project from start (question) to finish (communication)
- Explain and carry out descriptive, exploratory, inferential, and predictive analyses in Python
- Communicate results concisely and effectively in reports and presentations
- Identify and explain how to approach an **unfamiliar** data science task

Class info and resources
(e.g. syllabus, PDFs of slides)
are on GitHub
https://github.com/COGS108

No programming experience (or you forgot it all)?

- *Preferred option*
  - Take a programming course first
  - COGS 18 : Introduction to Python
- *Can't wait?*
  - Use online sites like *codecademy.com* or *LearnPython.org*
  - *Python Data Science Handbook*

Discussion Section

Discussion Sections start in Week 1 with a Python review!!

- You don't *have to* go to section, but you **should**. Why go?
  - Short review of important concepts from that week
  - Individual help with anything (labs, assignments, project)
  - Set time to work on labs and assignments will prevent you from being late with your work
- Can I switch sections? Yes, but…
  - Don't take away a seat from someone who's registered in that section
  - Ideally you should stick with whichever section you decide to attend so you can develop a rapport with your TA

Grading

*** indicates group submission**

	number	each	% of grade
Labs	8	2%	16%
Quizzes	8	1%	8%
Assignments	4	6%	24%
In-class exercises	20	0.05%	1%
Project Review*	1	2%	2%
Project Proposal*	1	10%	10%
Project Checkpoints*	2	10%	20%
Final Report*	1	15%	15%
Final Video*	1	4%	4%
Team Evaluation	1	1%	1%

Grading

Extra credit worth up to 3% of the final grade will be awarded for

- Exceptional participation on EdStem discussion board: Roughly the top 3 - 5% of contributors will get 0.5% bonus to their final grade. Starting/participating in good discussions, organizing things, answering questions, etc.
- Being present for >= 2/3 of the in class exercises will net 0.5% bonus to the final grade
- Answering the pre & post course surveys will give 0.5% extra credit (0.25% each for 0.5% total)
- Attending guest lecture(s) in-person is 0.5% extra credit
- Filling out all 7 of the weekly project progress surveys (0.5% of grade, see Project section below)
- If >2/3 of the students fill out SET teaching evaluations at the end of the quarter there will be an extra 0.5% of the final grade for everyone. BTW this EC criteria has not been met since the zoom classes of the pandemic, why won't the current generation of students actually fill out evals?

Weekly Lecture Quizzes

- 8 of the 9 weekly quizzes covering the previous week's lecture topics
- Posted Fridays around 5pm, Due Mondays at 11:59pm
- NO late submisison
- Lowest score dropped
- Goal: to help you keep on top of the material covered in lecture
- How:
  - Taken on Canvas
  - Two attempts
  - ~10-20 Questions, 20 minutes per attempt

8 Discussion Lab exercises

Completed individually and graded mostly programmatically (for correctness) and partly manually (for effort and good thinking).

- These are meant to get you practice programming around the topics covered in class.
- You will have to look some stuff up on your own. This is by design.
- Instructions must be followed perfectly to receive credit.
- You'll have the opportunity to practice in discussion section.

Discussion labs will be due on Fridays by 11:59PM

75% credit if submitted less than 5 days after deadline. No submission after 5 days late.

7 LATE DAYS allowed per person without penalty
to be used for Discussion Labs + Assignments

4 Assignments

Completed individually and graded almost completely programmatically (for correctness).

- These are meant to get you practice programming around the topics covered in class.
- The first two are much simpler/shorter, the last two are harder/longer.
- You will have to look some stuff up on your own. This is by design.
- Instructions must be followed PERFECTLY to receive credit.
- You'll have the opportunity to practice in discussion section.

Assignments will be due on Wednesdays by 11:59 PM

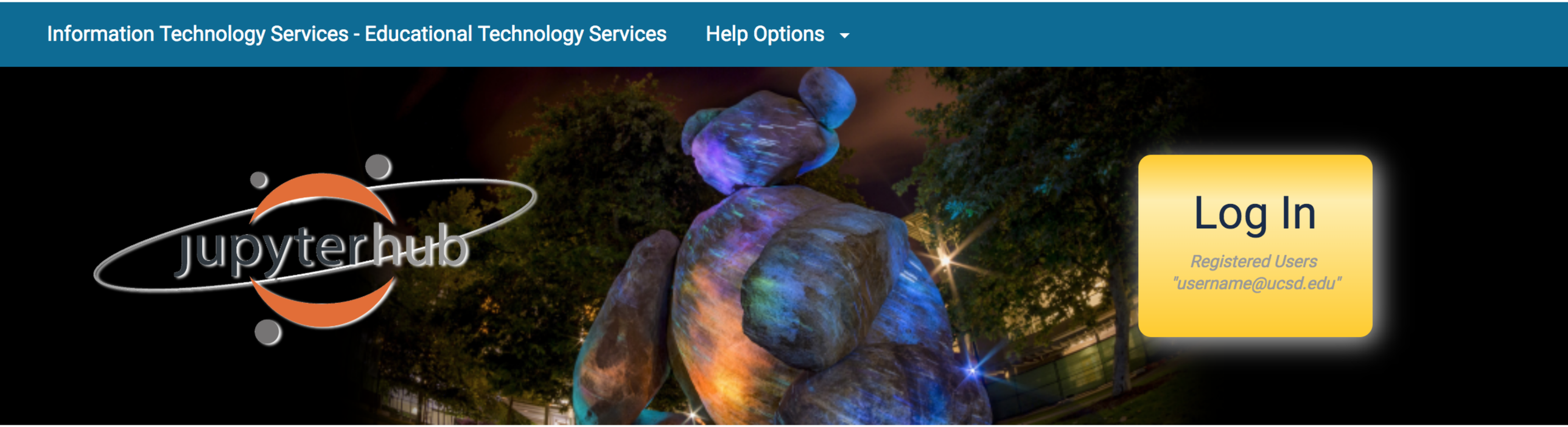
75% credit if submitted less than 5 days after deadline. No submission after 5 days late.

7 LATE DAYS allowed per person without penalty
to be used for Discussion Labs + Assignments

Assignment/lab Submission @ Datahub: https://datahub.ucsd.edu

DATA SCIENCE / MACHINE LEARNING PLATFORM

UC San Diego



UC San Diego Jupyterhub (Data Science) Platform

Before next Mon: log onto datahub & have a working installation of Jupyter on your computer

Group Projects: the main focus of COGS 108

Week 3: find your group of 3-5 people

Week 4: review a previous project,

Week 5: make a project proposal,

Week 6: get feedback and fix things

Weeks 7 - 10: 2x: a checkpoints, get feedback, and fix things

Finals week: Turn in a final written report w/ a short video summary

How to find a group:

1. go to discussion section, talk to people there
2. post on Looking for Teammates on EdStem
3. talk to people you are sitting near after class

Project examples

From https://github.com/COGS108/FinalProjects-Sp23

- Analyzing Predictors of an Anime's Score on MyAnimeList
- Exploring the relationship between gender and dialogue sentiment in MCU films
- Will California meet its goal of all vehicle sales being EVs by 2035?
- Predicting Football Player Transfer Net Worth
- Housing Price Impact on California's Population: Comparative Analysis with Texas

What is your question? How will you get relevant data? What is your hypothesis?

Project Resources

- Information to prevent your group from hating each other (20% have problems!!) https://github.com/COGS108/Projects/blob/master/COGS108_TeamPolicies.md

- https://github.com/COGS108/Projects/blob/master/FinalProject_Guidelines.md

15% of groups have at least one “loafer”

20% of people agreed that
“working with my group was more stressful than anticipated”

Successful groups…

- Set group expectations
- Respect each other and each other’s time
- Communicate often and clearly
- Have a positive culture (important to start right!!)
- Are diverse (skills, background, personality, culture, race/ethnicity, gender)
- Accept and support individual differences
- Help each other learn

Confusion and struggling is expected and ITS OK!!!

This class is a mile wide and an inch deep…. you will need to teach yourself!

If something is unclear:

First try for a while to understand it yourself, or to educate yourself from the internet

Still struggling?

- *ask in class*
- *ask during section*
- *post on EdStem*
- *ask a classmate*
- *come to office hours*

Policy on getting help from other people / AI

MORE HELP OK ——————————— LESS HELP OK

| | |

Projects Labs Assignments

1. You can help someone learn, but be Socratic wherever possible
2. Straight up giving someone answers or code is cheating.
3. If you get help CREDIT / CITE the software package, webpage, publication, person, or AI that produced the code or writing you are reusing.
4. Don’t use stuff you don’t understand! You are responsible for being able to explain your solutions or you may lose points

How to use AI the right way

1. Asking it to straight up write your code based on a text description of the problem is dangerous.
   Works well for a common problem, more errors for a rarer task.
   Works well for good descriptions, more errors for poor ones
2. No matter the problem rarity, it can be helpful to ask it to suggest different approaches to the problem, and to list the pros and cons of each approach.
3. Use it to help you write test code
4. Ask it to explain errors and to help you debug
5. CITE YOUR USE!!!
   # Gemini Prompt: Load data from Google Drive

How these things work together

- Most effort and learning in the project
- Regular small stakes assessments (Quiz, Labs) to practice
- Periodic large stakes assessments (Assignments, Project) that repeat the concepts in the small stakes stuff but longer, harder
- Lots of practice with the core skills:
  - thinking clearly with data
  - learning to learn
  - using common tools like Git and Python

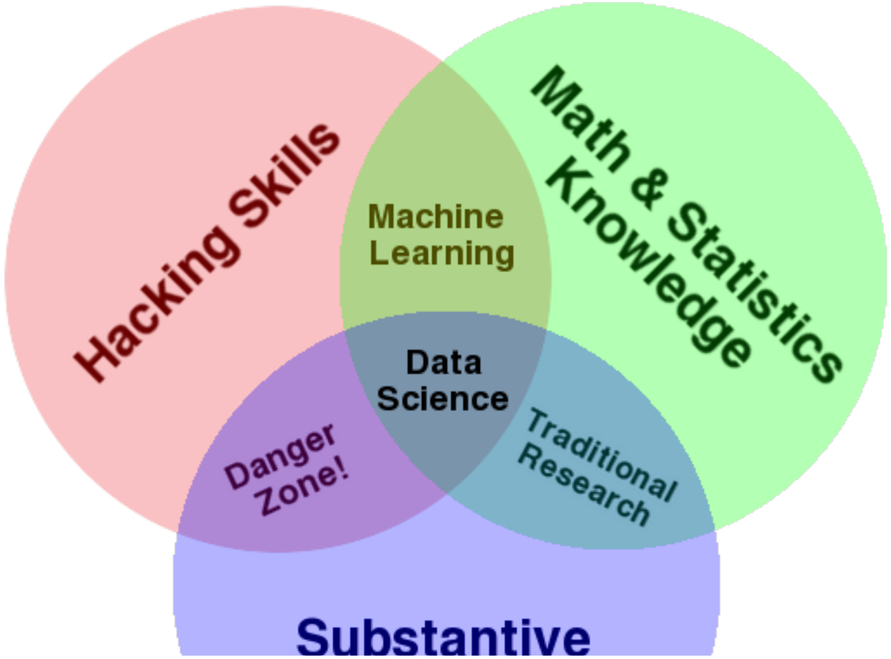
How to get an easy A

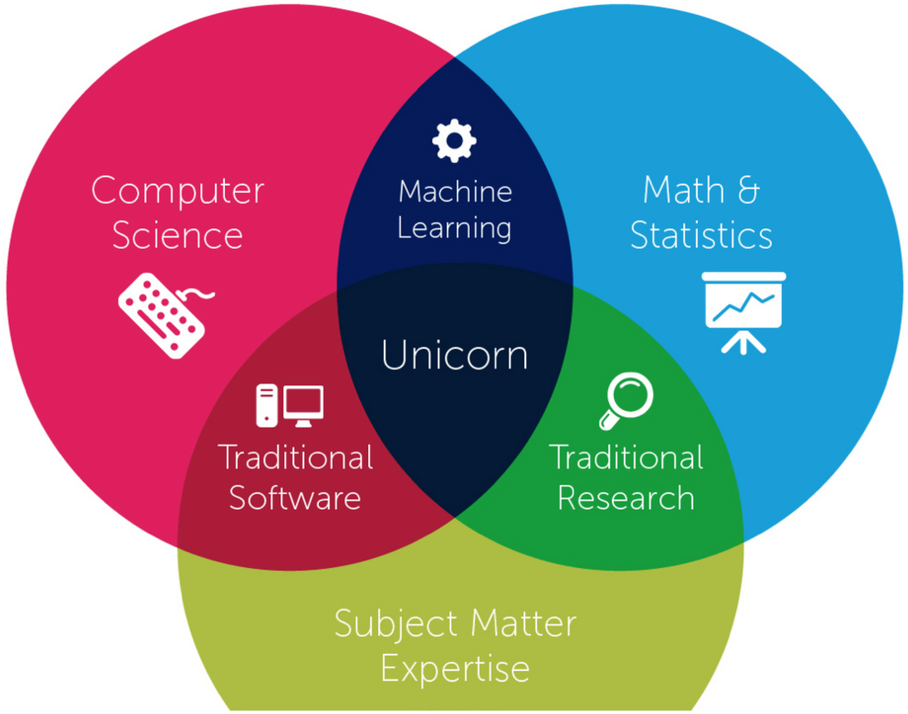
- Pay attention to the lectures. Go to section
- Keep track of time, plan things in advance
- Still new to programming / statistics?
  - Try to do the work yourself until you've struggled for 30 minutes, then ask for help (section, TA, classmates, me, Ed)
  - Limit your use of AI
  - Ask your group mates to teach you how to do something, don't let them do it for you
- Expert at CS / stats?
  - Teach other people, nothing cements knowledge like teaching it
  - Learn to use AI effectively

What COGS 108 logistics questions do you have?

"The scientific process of extracting value from data"

What is data science?

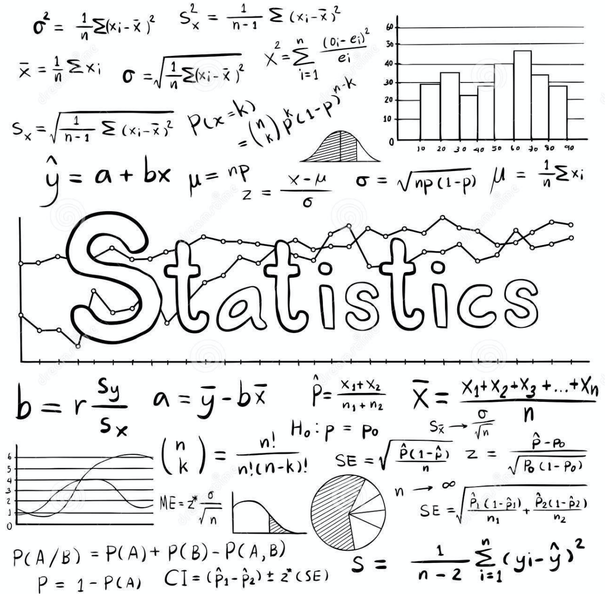


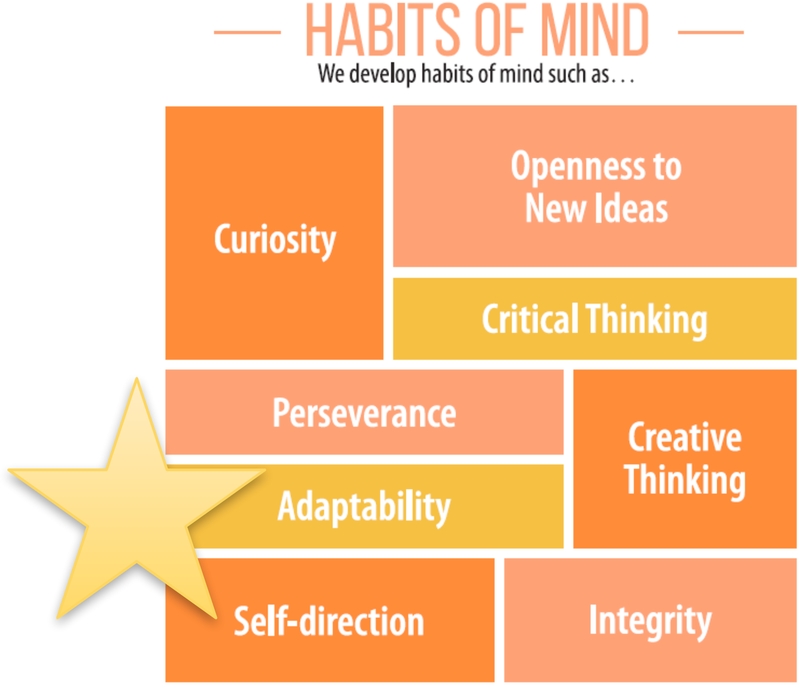


Thinking clearly with help from data

http://c

How this class will train you for DS





Data scientist is actually MANY jobs

https://hbr.org/2018/11/the-kinds-of-data-scientist

A final piece of advice for those hiring data scientists: Look for people who are in love with solving problems, not with specific solutions or methods, and for people who are incredibly collaborative. No matter what kind of data scientist you are hiring, to be successful they need to be able to work alongside a vast variety of other job functions — from engineers to product managers to marketers to executive teams. Finally, look for people who have high integrity. As a society, we have a social responsibility to use data for good, and with respect. Data scientists hold the responsibility for data stewardship inside and outside the organization in which they work.

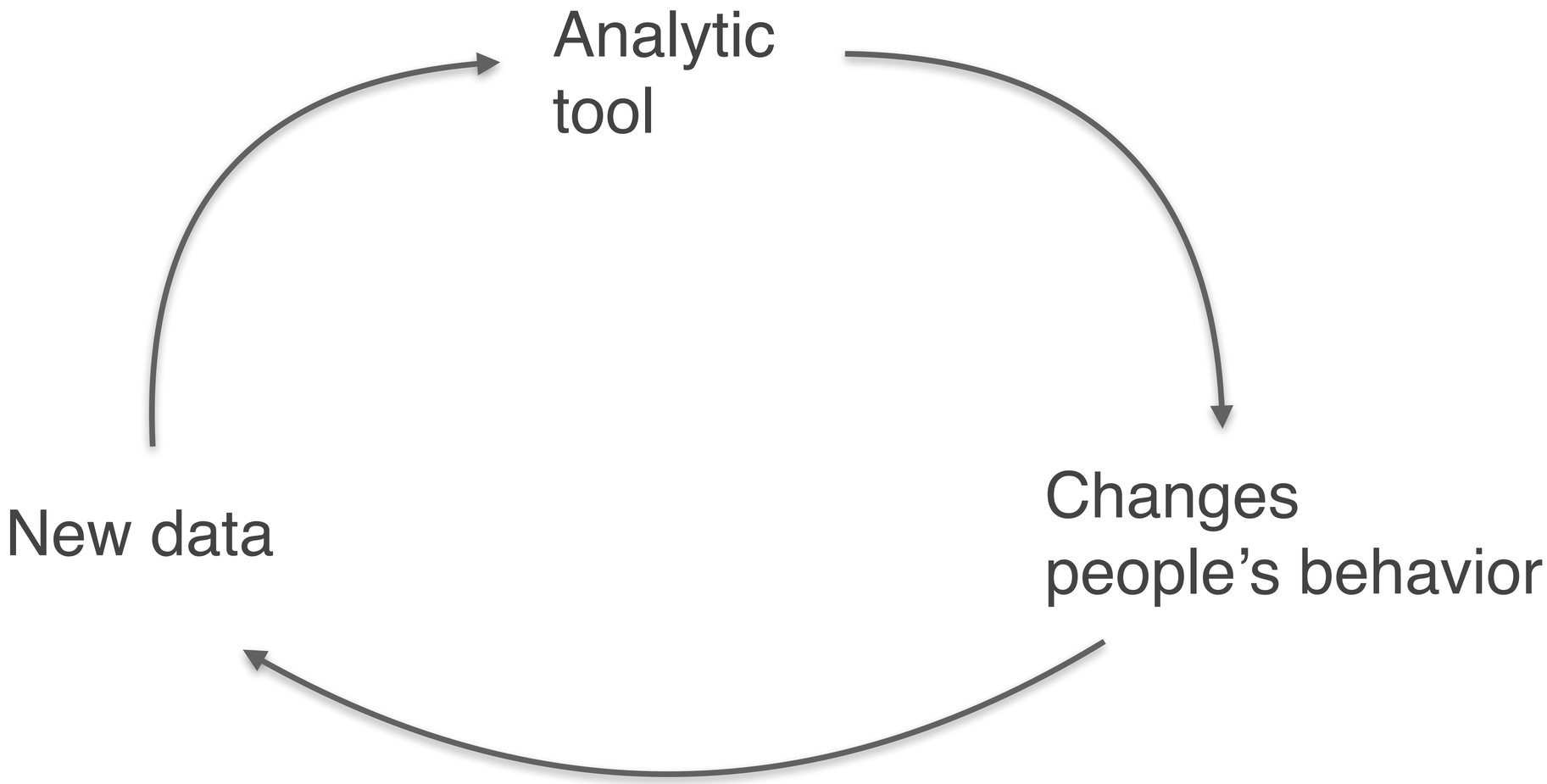
Data science for humans

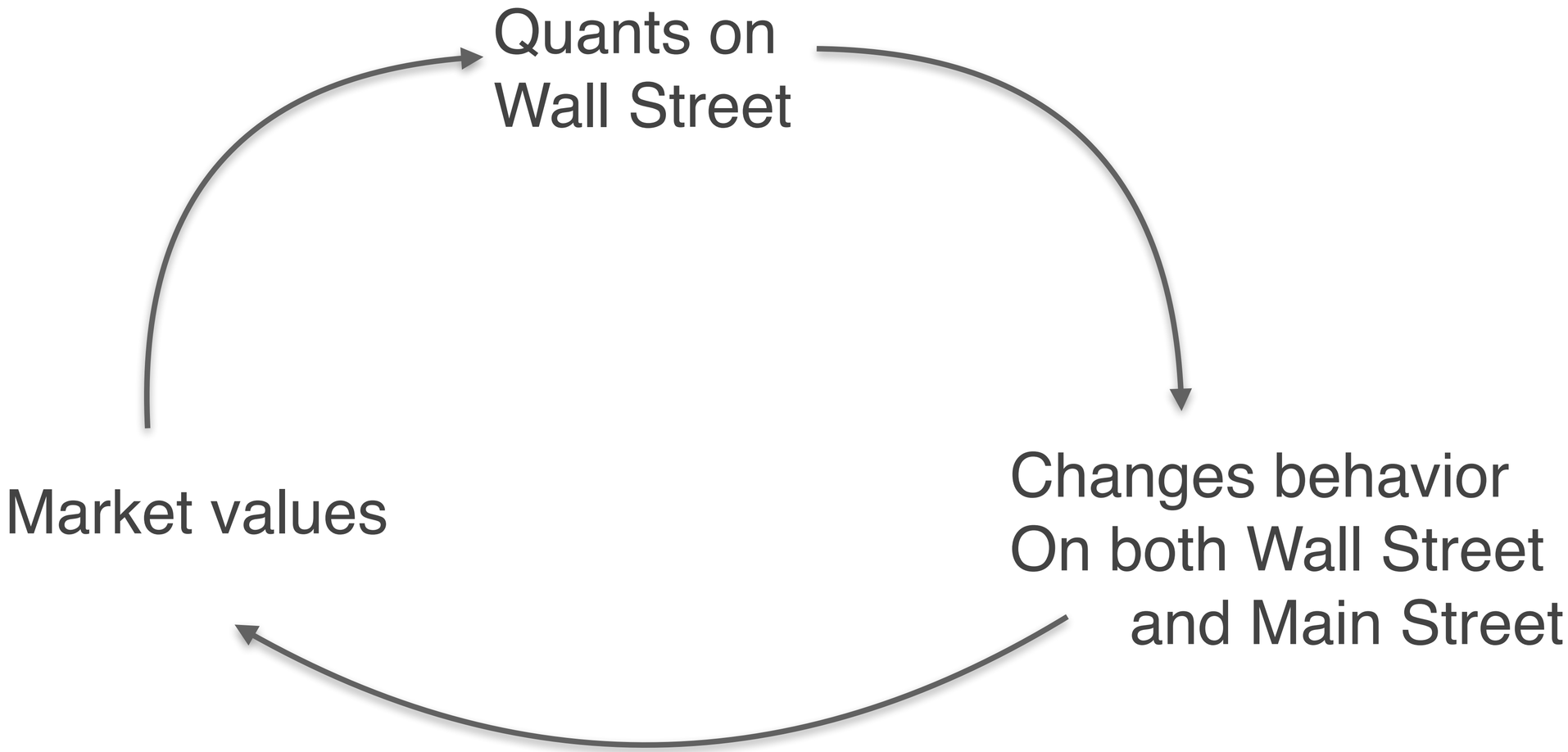
Data science for computers

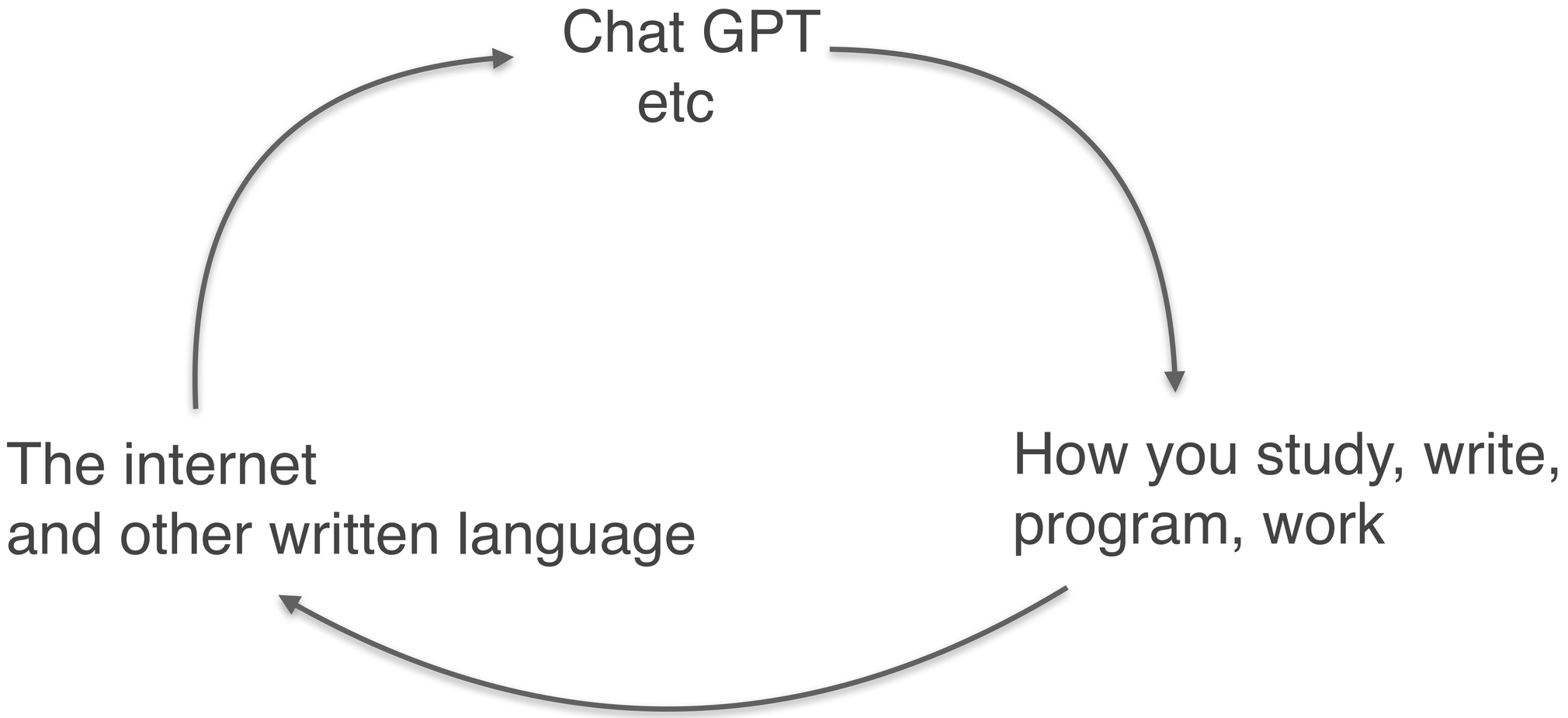
Data science is inherently social, cultural, and political

When do we care about DS?

Only if the model, analysis, tools, etc are useful to people.


Analytic
tool
Changes
people’s behavior
New data


Quants on
Wall Street
Changes behavior
On both Wall Street
and Main Street
Market values


Chat GPT
etc
How you study, write,
program, work
The internet
and other written language

The New York Times

The Hustlers Who Make $6,000 a Month by Gaming Citi Bikes

The bike-sharing program rewards users who help redistribute bikes around New York City. A few riders have figured out how to turn that into profit.

A full dock of Citi Bikes on 34th Street means there's money to be made by moving them elsewhere. Yuvraj Khanna for The New York Times

It was the perfect New York hustle, a scam of subtle perfection. And for three years, it helped Mark Epperson pay his rent.

The hustle, in its simplest form: Borrow a Citi Bike. Ride it one block. Wait 15 minutes, then ride it back.

Earn $6,000 a month (under ideal conditions, and with lots of work).

Occasionally, though, a ride to work ends with the rider’s discovery that the docking station nearest the office is full. A dash to brunch is foiled by an empty dock, with no bikes available.

Both situations are annoying, especially for Citi Bike subscribers, who now pay $220 a year. To fix the imbalance, Citi Bike uses various tactics to move bikes to in-demand stations.

[One] is a program called Bike Angels, in which Citi Bike users move bikes in exchange for points that could be cashed in for swag like water bottles and backpacks, membership discounts and gift cards.

“We imagined people would do it as a recreational fitness kind of thing,” said David B. Shmoys, a data scientist at Cornell University whose research team created Citi Bike’s first rebalancing algorithm in 2014. “We never imagined anyone getting really obsessed.”

Over the years, a few users found ways to maximize the program's financial benefits. [Which works like congestion pricing]

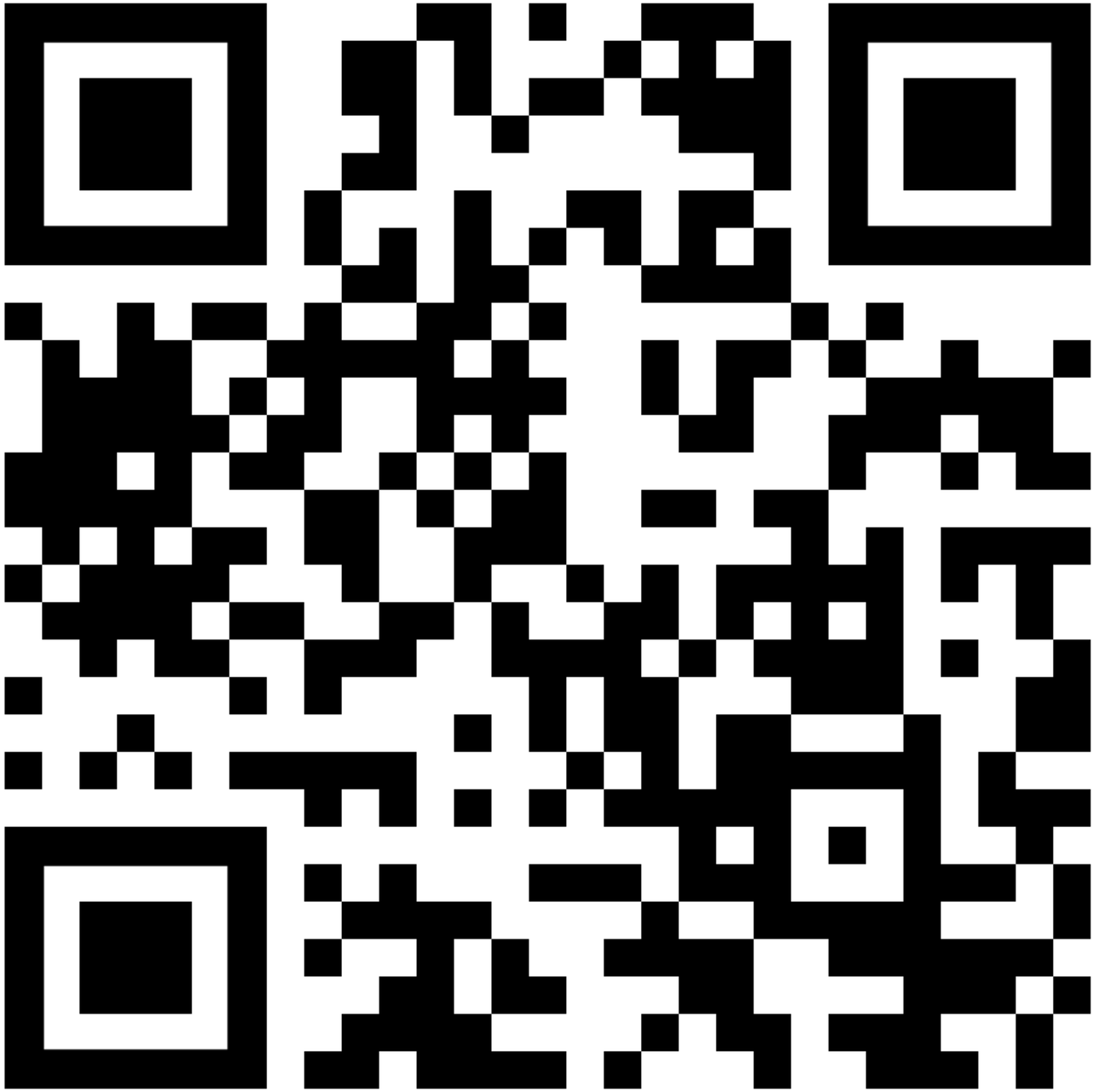
But a few riders realized that by working as a team, and quickly, they could exploit the algorithm. For Mr. Epperson and his fellow hustlers, it "created an opportunity to make a lot of money," he said.

At 10 a.m. seven Bike Angels descended on the docking station at Broadway and 53rd Street, across from the Ed Sullivan Theater. Each rider [..] unlocked a bike [and] rode it one block east, to Seventh Avenue […] docked, ran back to Broadway, unlocked another bike and made the trip again.

By 10:14, the crew had created an algorithmically perfect situation: One station 100 percent full, a short block from another station 100 percent empty. The timing was crucial, because every 15 minutes, Lyft's algorithm resets, assigning new point values to every bike move.

The clock struck 10:15. The algorithm, mistaking this manufactured setup for a true emergency, offered the maximum incentive: $4.80 for every bike returned to the Ed Sullivan Theater. The men switched direction, running east and pedaling west.

https://forms.gle/4UG7nugVvQrS4zQr7



Yes, some data scientists at Citi Bike had to detect this and consider what — if anything — to do about it

Data science is inherently social, cultural, and political

A good data scientist …

- Is curious and always learning
- Is technically proficient
- Is a careful, conscientious, critical thinker
- Communicates results clearly
- Understands that their work is inherently social, cultural, and political

I’m excited to have
you all in COGS 108!