Introduction to Data Science

CSC2621 Data Science

Syllabus Introduction

Who am I?

- Dr. John Bukowy
- Ph.D. in Physiology?!
  - Other degrees in:
    - Biomedical Engineering
    - Electrical Engineering
- Industry Experience
  - 2 years at MERGE Healthcare
- 5 year at MSOE
- When I am not working I…

What Will I Be Graded On?

- 5 Activities will have grades associated with them
  - Labs
  - Quizzes
  - Midterm Project
  - Final Project
  - Final Exam

Item	Percentage
Labs	35%
Quizzes	10%
Midterm Project	15%
Final Project	15%
Final Exam	25%

Letter Grade	Percentage Needed
A	>=93%
AB	>=89%
B	>=85%
BC	>=81%
C	>=77%
CD	>=74%
D	>=70%
F	<70%

What Will I Be Graded On?

- Labs
  - Labs are structured problems where we get to try the things that we have learned. This class will place a lot of emphasis on the labs. I view labs as being very important.
  - Labs will (should) be posted on Monday of the week that they are assigned.
  - Labs are due 1-week from lab day. If your lab day is Wednesday the labs submission is due the next Tuesday at 11:59 pm
  - Grade - %

What Will I Be Graded On?

- Quizzes
  - You should expect to take a short quiz at the beginning of each lab session.
  - The quiz will be hand written and take ~15 minutes.
  - The quizzes will cover lecture material since the last quiz as well as comprehension questions regarding the lab for the week.
  - Grade - %

What Will I Be Graded On?

- Midterm Project
  - The midterm project is a individual project that takes the place of a midterm exam.
  - The expectation is that you will complete this project completely on your own.
  - This will take place during week 8 and class time will be given to work on the project.
  - Grade - %

What Will I Be Graded On?

- Term Project
  - The term project will have you take a data science project from start to finish.
  - You will give the term project presentations in the final week of class (week 15)
  - After we have our final roster, we will decide the groups. I am happy to hear preferences/feedback on how to organize.
  - Think of this project as preparing materials to teach the class. Pay attention to how the example projects are structured in this class – model your projects after these.
  - Grade - %

What Will I Be Graded On?

- Final Exam
  - There will be a cumulative final exam that takes place during week 16.
  - The final exam will be a written exam.
  - More details to come
  - Grade - %

What Will I Be Graded On?

- Questions?

Class Introduction

Data Science

- What is Data Science?
- Why we’re interested in Data Science?
- What makes a good data scientist?

What is Data Science?

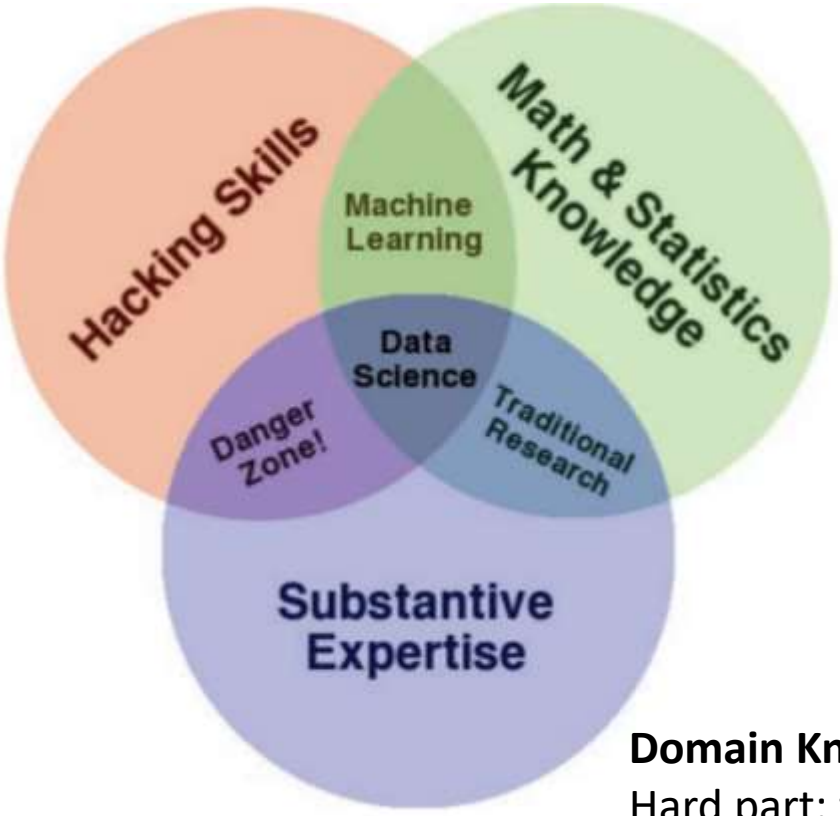
- Data Science is the science and art of **extracting** and **communicating insight** from data, using **computing**, **statistical** and **visualization** tools.

- "The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that's going to be a hugely important skill in the next decades."

Hal Varian, chief economist at Google, source (2009 interview)

Data Science : Combination of Different Skills

Computer Science:
Programming skills



Statistics:
- Tools to explore and model the data
- Scientific thinking/methods & study design

Domain Knowledge:
Hard part: thinking part
- What questions to ask?
- How to interpret the results**?**

Drew Conway, video
[5:45-15:35]

Data Science : Combination of Different Skills

- “Data science, as it’s practiced, is a **blend of Red-Bull-fueled hacking and espresso-inspired statistics**. … Data science is the civil engineering of data. Its acolytes possess **a practical knowledge** of tools and materials, **coupled with a theoretical understanding** of what’s possible.”

  Mike Driscoll, source

- “Data Scientist = statistician + programmer + coach + storyteller + artist”

  Shlomo Aragmon, source

Why Data Science?

Discover Glassdoor's Best Jobs in 2022

Using Glassdoor's unique data on jobs, salaries, and companies, we compiled a list of the 50 Best Jobs in America to help people find jobs they'll love. Each job stands out for its earning potential (median salary), job satisfaction, and job openings. Are you considering a new position? Check out this comprehensive list to see what jobs made the list this year, and view open jobs at companies across the country.

	Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1	Enterprise Architect	$144,997	4.1/5	14,021	View Jobs
#2	Full Stack Engineer	$101,794	4.3/5	11,252	View Jobs
#3	Data Scientist	$120,000	4.1/5	10,071	View Jobs
#4	Devops Engineer	$120,095	4.2/5	8,548	View Jobs
#5	Strategy Manager	$140,000	4.2/5	6,977	View Jobs
#6	Machine Learning Engineer	$130,489	4.3/5	6,801	View Jobs

Why Data Science?

Linkedin article: [Jobs on the Rise in 2021]

Data science specialists

According to the 2020 U.S. Emerging Jobs Report, data roles, specifically data scientist and data engineer roles, are increasing steadily — reflecting about a 35% average annual growth for both roles.

Why Data Science?

- 10 or 15 years ago, educational institutions didn't have programs or courses specialized in teaching data science.

- What created the interest? The age of big data: massive amounts of data + inexpensive computing power
  - Online behavior is datafied: "shopping, communicating, reading news, listening to music, searching for information, expressing our opinions"
  - Companies were interested in analyzing large less-structured datasets.
  - The big data hype encouraged many industries to collect data and to adopt a **data-driven** culture: education, social welfare, government, finance.

Why Data Science?

- 10 years ago, DJ Patil who co-coined the term data scientist, co-authored the famous article: “Data-scientist : The Sexiest Job of the 21st century”.

- 10 years later (July 2022), the authors wrote a new article “Is Data Scientist Still the Sexiest Job of the 21st century?”
  - “A decade later, the job is more in demand than ever with employers and recruiters”
  - They describe how the job has changed (ex: importance consideration of ethics, automation of some aspects of data science, algorithms drifts observed due to COVID-19)

How is data science different than traditional engineering and science?

Mode of inquiries:

- Traditional engineering: method-driven
- Traditional scientific method: hypothesis-driven
- Data science: data-driven

Modes of Inquiry: Method Driven

Traditional engineering approach:

- We focus on understanding a method or technique
- We identify when and how to apply that technique
- We becomes experts at applying the technique

Modes of Inquiry: Hypothesis Driven

Traditional scientific method:

- We form a hypothesis
- We design an experiment (including collecting data) to test the hypothesis
- If the experiment is able to reject the hypothesis, we generate a new hypothesis
- Otherwise, we design another experiment to test the hypothesis

Modes of Inquiry: Data Driven

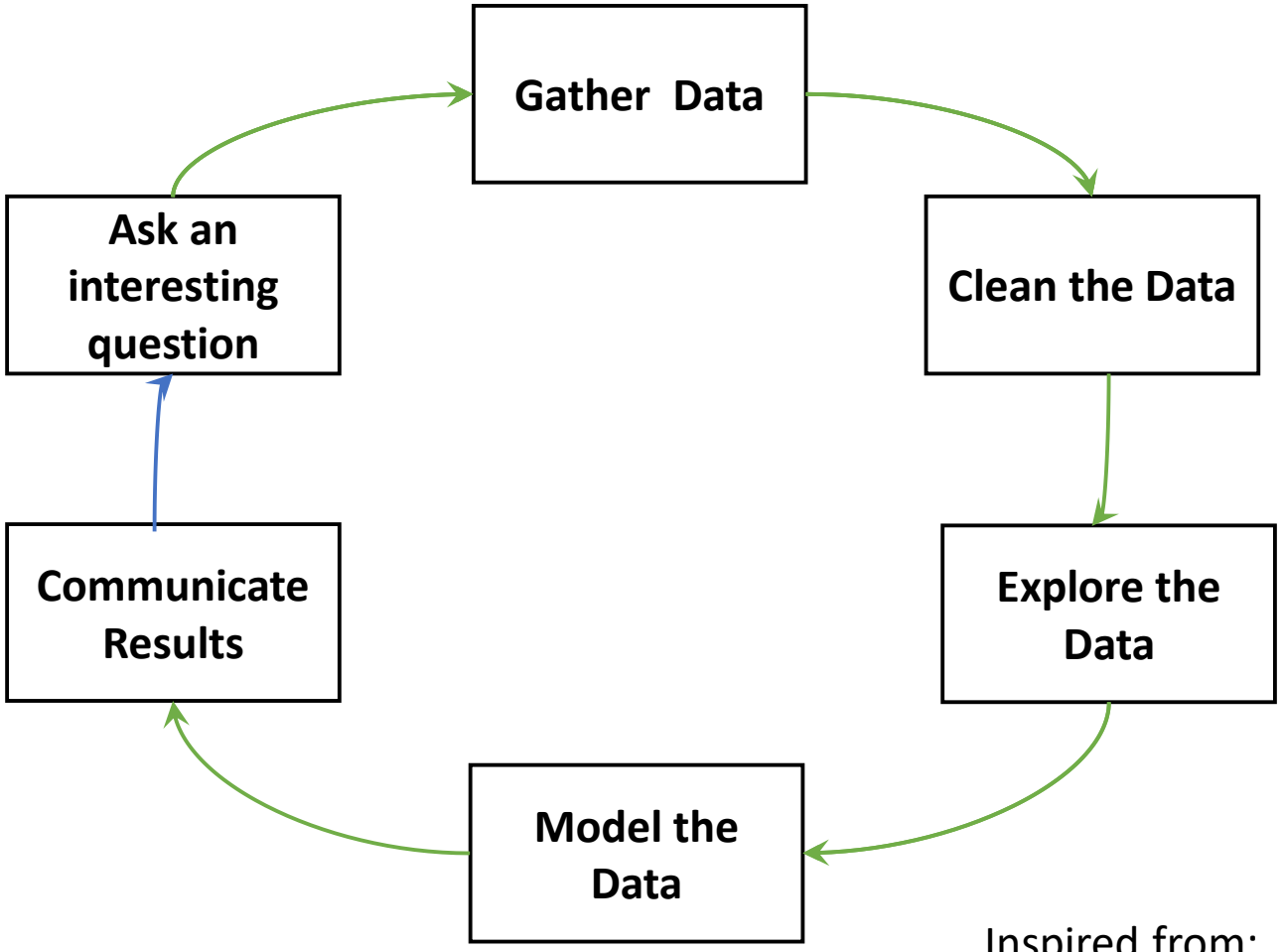
Data Science:

- We start with a data set
- We explore the data set to identify patterns
- From these patterns, we ask questions and form hypotheses
- We may be able to use the data to answer the hypothesis or may need to design a new experiment

This is a new mode of inquiry and what makes Data Science different from traditional science and engineering.

https://learninganalytics.upenn.edu/ryanbaker/GIFTCh12.pdf

Data Science Process

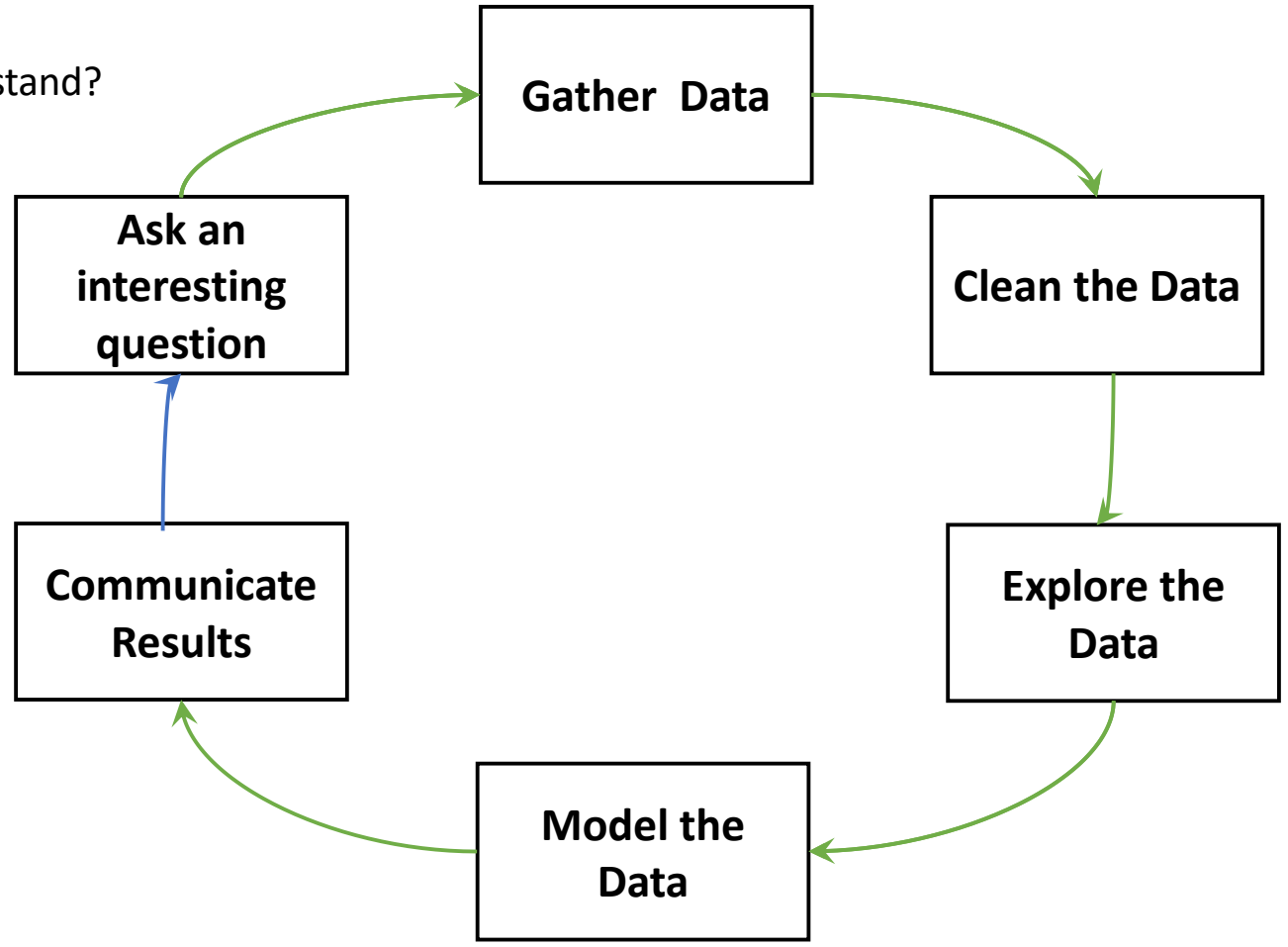


Inspired from:

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

What is the goal?

- Predict? Recommend? Forecast?
- Explain or Understand?
- Make a decision?



Inspired from:

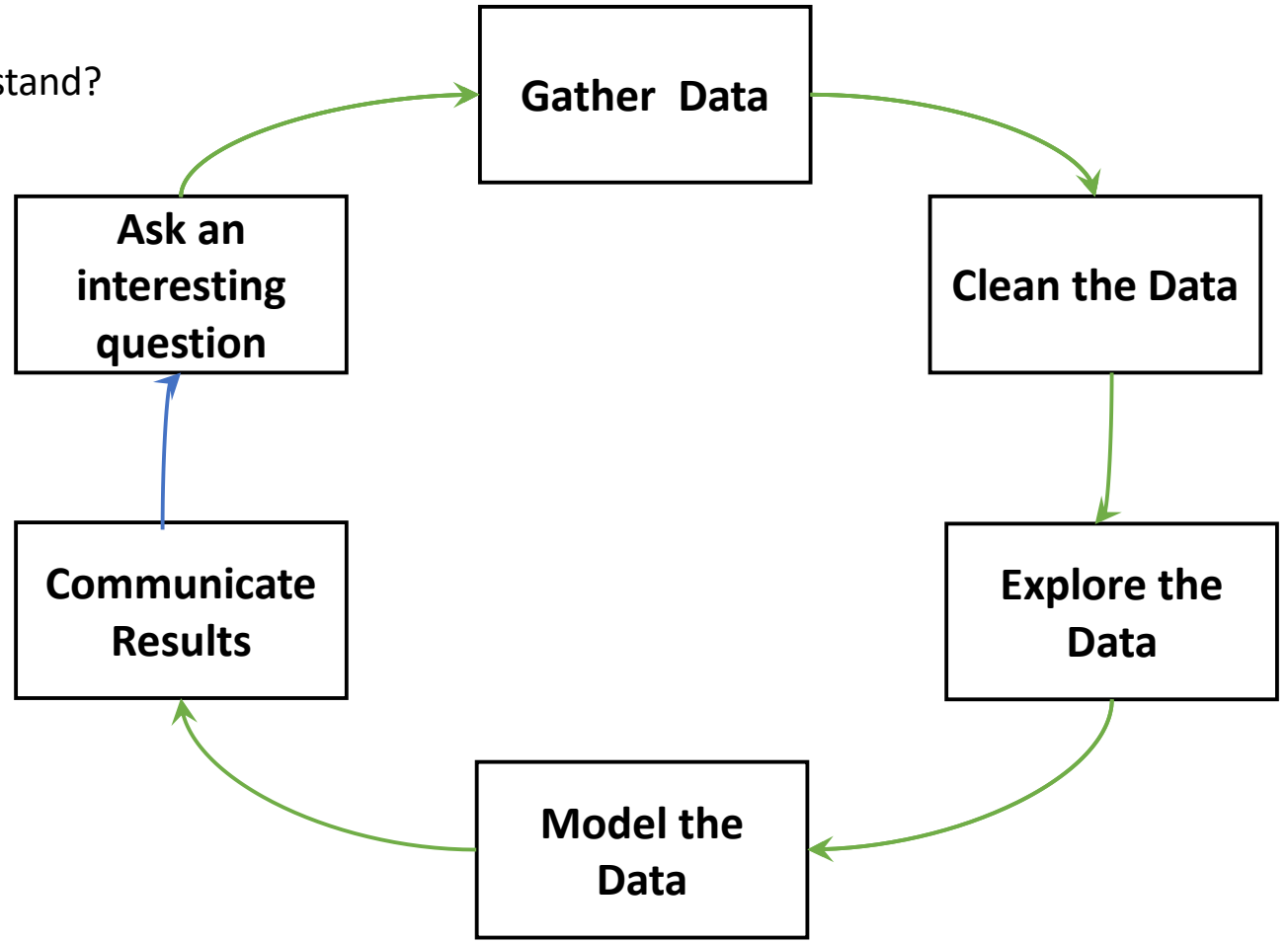
Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

What is the goal?

- How was the data collected? By whom?
- Which data are relevant?
- Any ethical or privacy issues?

What is the goal?

- Predict? Recommend? Forecast?
- Explain or Understand?
- Make a decision?



Inspired from:

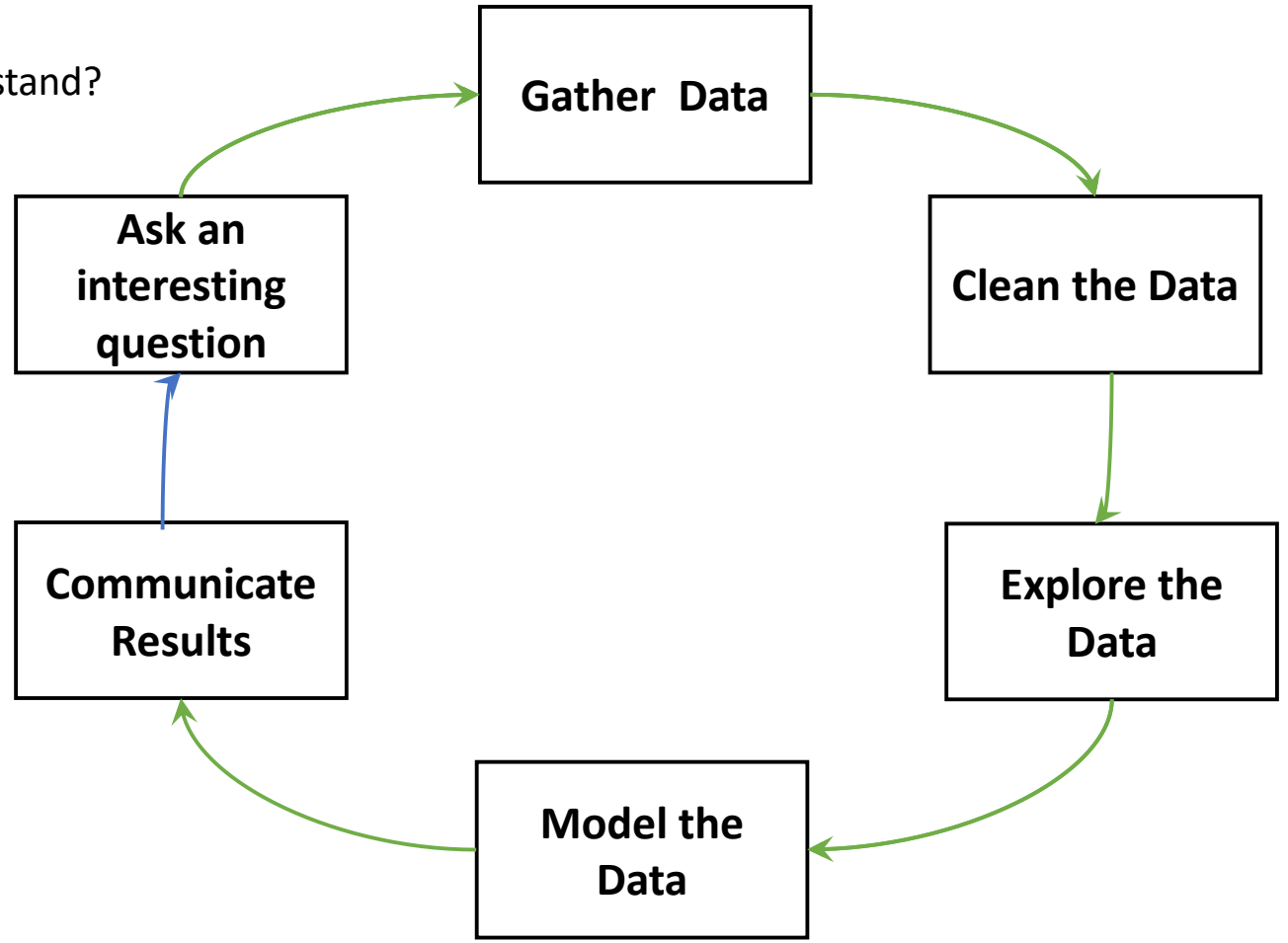
Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

What is the goal?

- How was the data collected? By whom?
- Which data are relevant?
- Any ethical or privacy issues?

What is the goal?

- Predict? Recommend? Forecast?
- Explain or Understand?
- Make a decision?



Clean & Explore

- Are there any anomalies, missing values or outliers?
- Visualize the data.
- Any patterns

Inspired from:

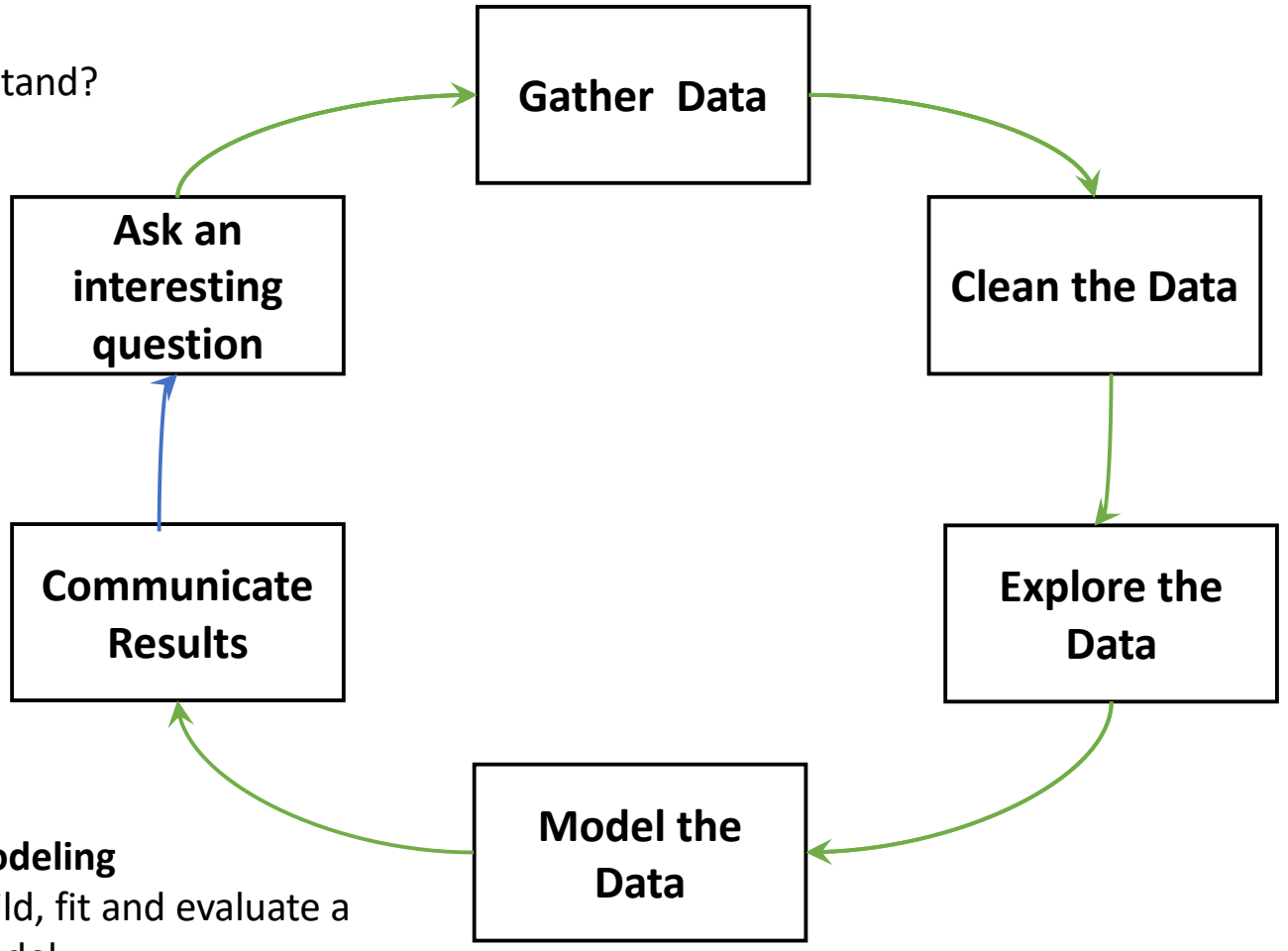
Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

What is the goal?

- How was the data collected? By whom?
- Which data are relevant?
- Any ethical or privacy issues?

What is the goal?

- Predict? Recommend? Forecast?
- Explain or Understand?
- Make a decision?



Clean & Explore

- Are there any anomalies, missing values or outliers?
- Visualize the data.
- Any patterns

Modeling
Build, fit and evaluate a model

Inspired from:

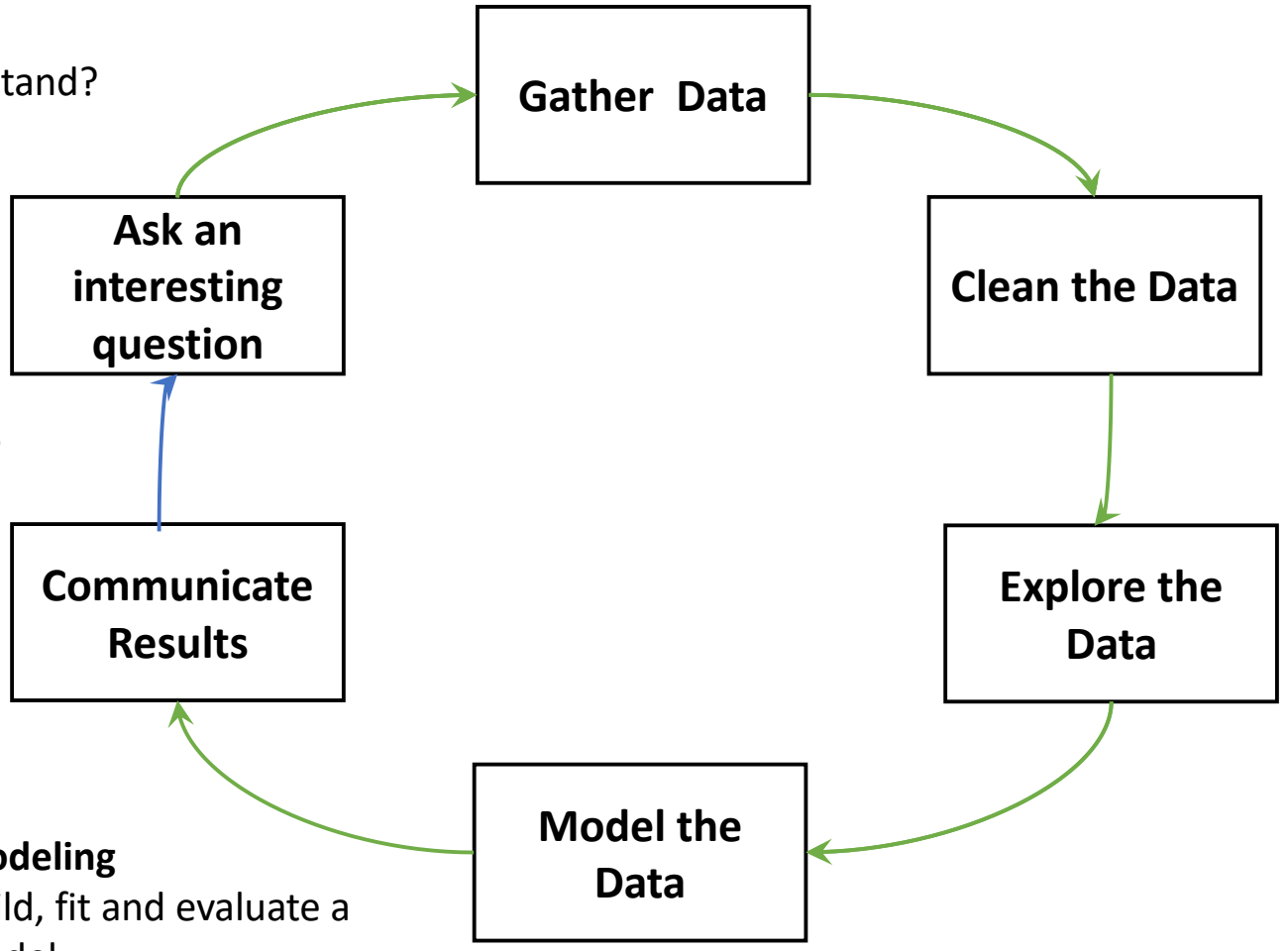
Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

What is the goal?

- How was the data collected? By whom?
- Which data are relevant?
- Any ethical or privacy issues?

What is the goal?

- Predict? Recommend? Forecast?
- Explain or Understand?
- Make a decision?



Clean & Explore

- Are there any anomalies, missing values or outliers?
- Visualize the data.
- Any patterns

Results

- What did we learn?
- What story to tell?
- Do results make sense?

Modeling

Build, fit and evaluate a model

Inspired from:

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

Characteristics of a Data Scientist

- **Curiosity**: having an inquisitive mindset, the desire to dig into the data and to learn about the application domain

- **Communication**: telling a story with the data and effectively communicate the results

And sometimes a lot of patience to clean large messy datasets!

Case Study: Who (used to) goes to the gym?

- During graduate school I found myself on a committee that managed and advocated for our campus's onsite gym.
- Gym was not staffed but student’s and faculty had 24/7 access.
- The gym had a somewhat sordid history
  - Surrounding departments thought some bros were too noisy with their mad-heavy lifts
  - There was a commercial gym across the road (free vs. ~$50/month)
  - Previous discussion about shutting down the campus gym
- Was approached to create macro for generating use gym use statistics based off card reader access

Analysis request

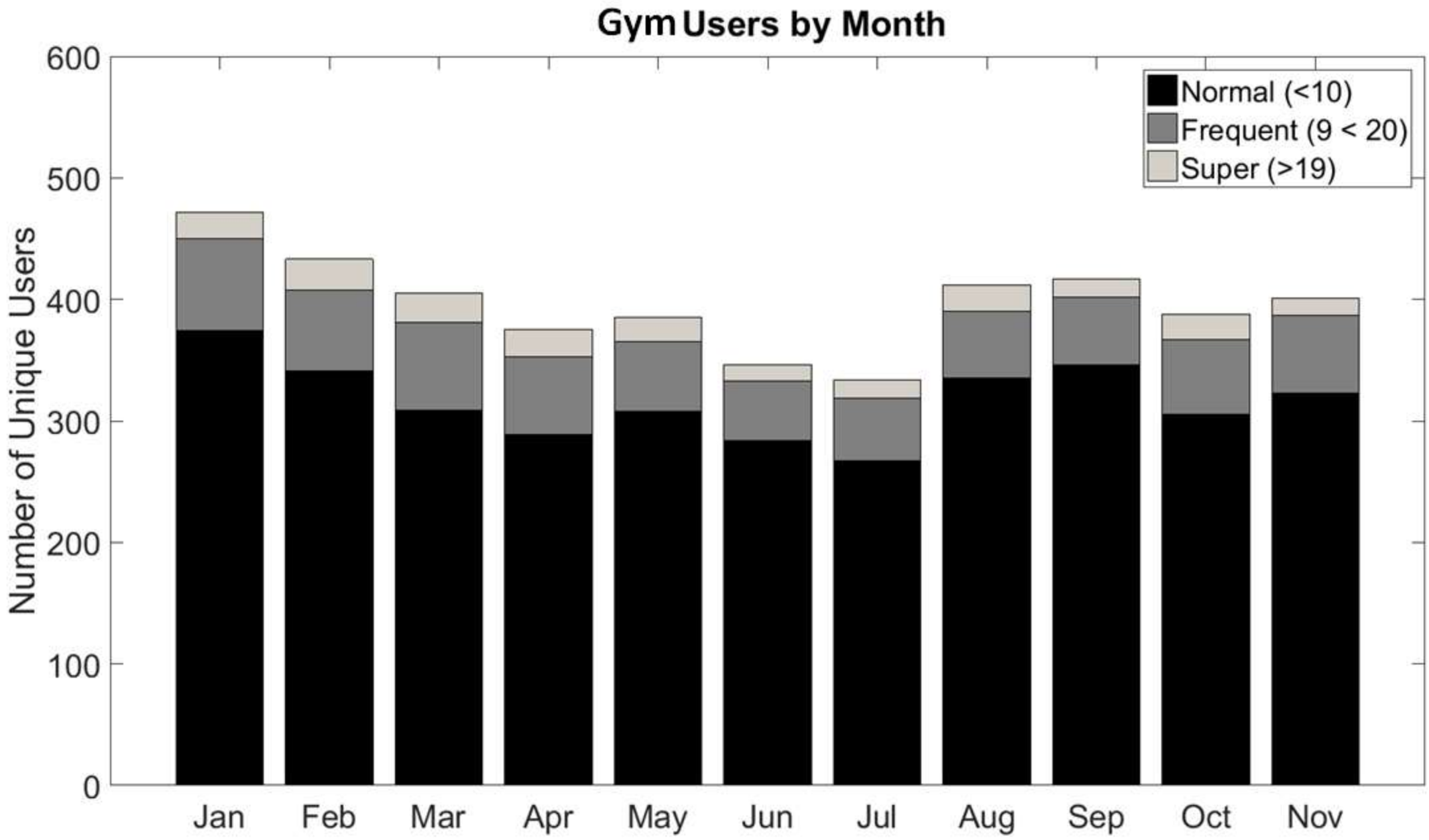
The metrics leadership was asking for included:

- Gym User Groups: what were the behaviors of individuals using the gym? They made arbitrary use definitions of
    1) Normal users (<10 total visits in a month)
    2) Frequent users (9 < 20 total visits in a month)
    3) Super users (> 20 total visits in a month)

- Number of unique users by month

Dataset

Date	ID	Position
1/1/2016 11:16:46 AM	000086	CELL BIOLOGY
1/1/2016 11:34:30 AM	043428	CELL BIOLOGY
1/1/2016 11:46:48 AM	016591	CELL BIOLOGY
1/1/2016 12:02:49 PM	016591	CELL BIOLOGY
1/1/2016 12:08:32 PM	000086	CELL BIOLOGY
1/1/2016 12:23:42 PM	076421	CELL BIOLOGY
1/1/2016 2:09:06 PM	009219	CELL BIOLOGY
1/1/2016 2:42:10 PM	005058	MEDICAL STUDENT
1/1/2016 2:44:18 PM	021453	MEDICAL STUDENT
1/1/2016 3:26:15 PM	021453	MEDICAL STUDENT
1/1/2016 4:52:57 PM	064585	MEDICAL STUDENT

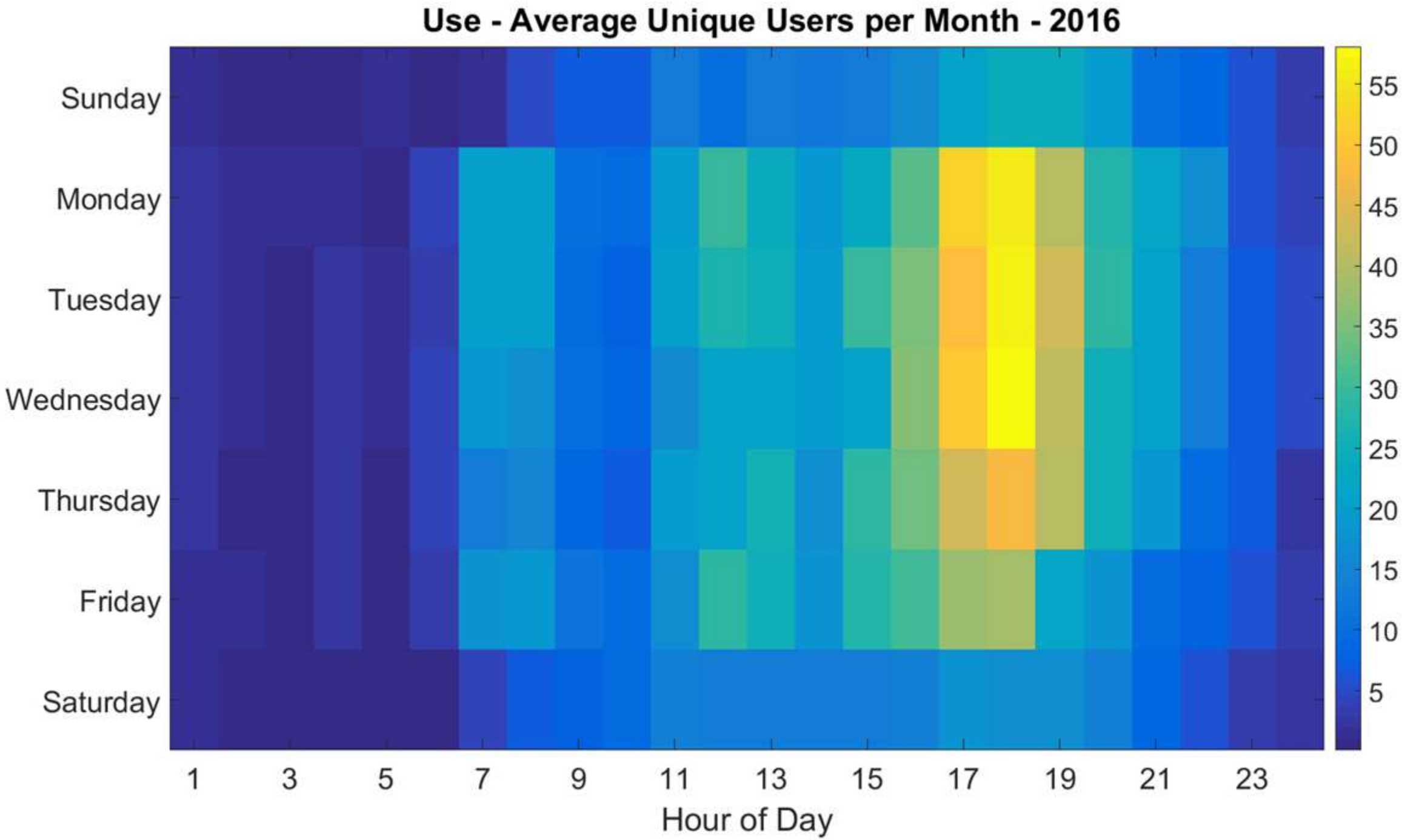

Gym Users by Month
Number of Unique Users
600
500
400
300
200
100
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Normal (<10)
Frequent (9 < 20)
Super (>19)

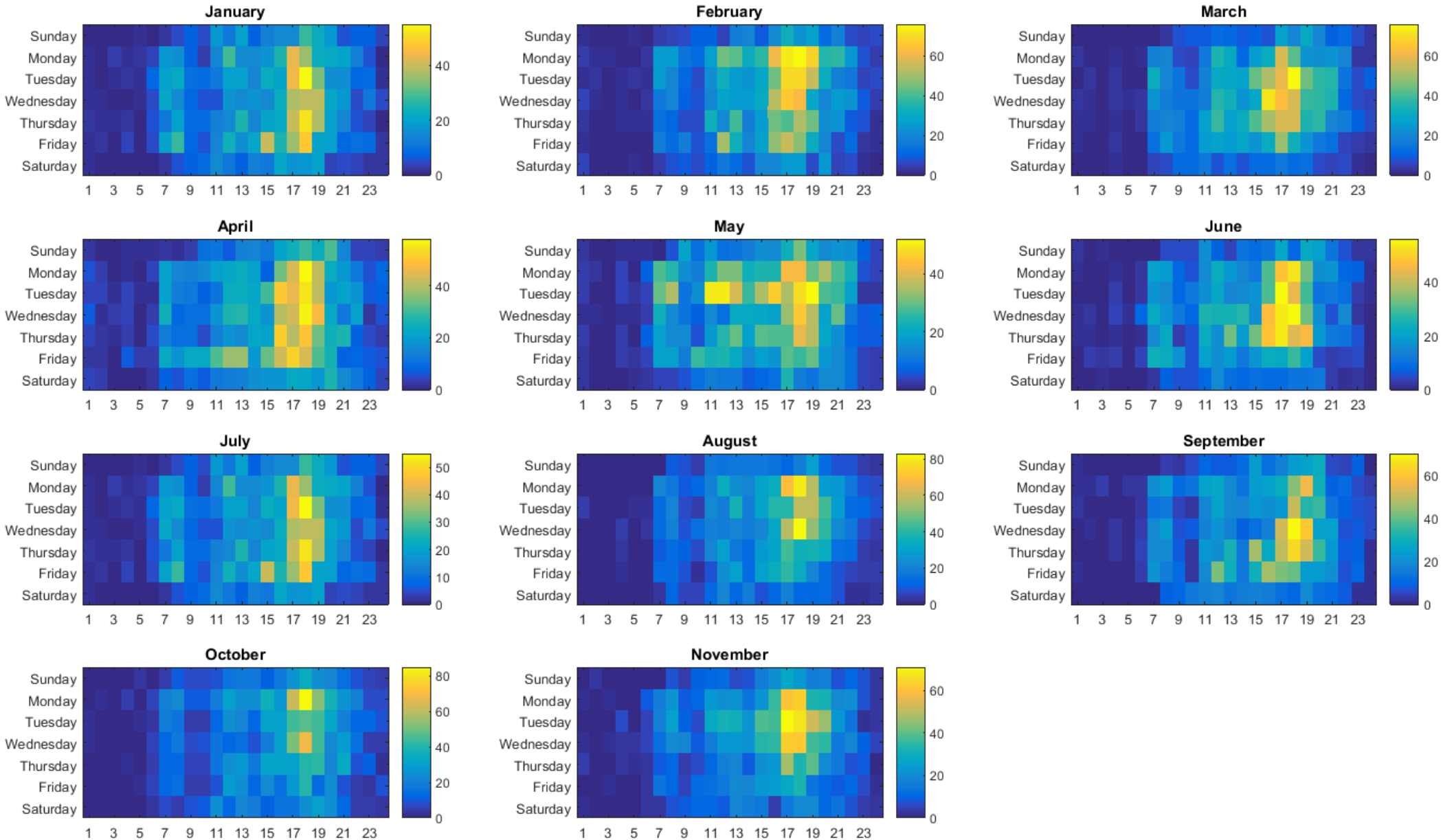
Other Questions?

What else can we do with this data?

- Selfish pursuits?

- Ethical dilemmas?


Use - Average Unique Users per Month - 2016
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
1
3
5
7
9
11
13
15
17
19
21
23
Hour of Day
55
50
45
40
35
30
25
20
15
10
5


January
February
March
April
May
June
July
August
September
October
November
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
1 3 5 7 9 11 13 15 17 19 21 23

Dataset Example - Exercise

Consider the SF Bay Area Bike Share datasets (Kaggle):

- station.csv - Contains data that represents a station where users can pickup or return bikes (id, location, number of docks)
- status.csv - data about the number of bikes and docks available for given station per minute.
- trips.csv - Data about individual bike trips (duration, start time, start station, end time, end station, subscription type: customer or subscriber)
- weather.csv - Data about the weather on a specific day for certain zip codes (the max, min and mean of temperature, dew point, humidity)

Think of some questions that you'd like explore using this dataset.

Sources of Datasets

- Nasdaq data link : financial and economic datasets
- US Government Open Data
- UCI Machine Learning Repository
- Kaggle Datasets
- Mode Analytics
- Google's public datasets directory
- Awesome public datasets

Other Examples of datasets

- **US Wildfire dataset:** contains a spatial database of wildfires that occurred in the United States from 1992 to 2015 (discovery date, final fire size, and a point location at least as precise as Public Land Survey System)
- **Amazon Customer Reviews Datasets:** contains hundred million reviews to Amazon products reported in a period of over two decades since the first review in 1995.
- **Predicting solar energy production**
- **Mental health in tech survey (ongoing survey):** “measure attitudes towards mental health in the tech workplace and examine the frequency of mental health disorders among tech workers.”
- **Details of NFL play-by-play data 2009-2018**

Data Science Skills / What You'll Learn

- **Data munging** – parsing, scraping, formatting, cleaning data
- **Scientific process** – exploring data to observe patterns, stating a hypothesis, and proving or disproving the hypothesis (e.g., using models, statistics, or visualizations)
- **Communication and Visualization** – reports, tables, graphs, interactive data applications, summary statistics
- **Statistics** – traditional analysis
- **Machine learning** – modeling relationships, prediction
- **Domain knowledge** – business, science, etc.