

# QSS20: Modern Statistical Computing

## Unit 01: Intro and Setup

# Agenda

- ▶ Course goals
- ▶ Intros
- ▶ Break
- ▶ Nuts and bolts
- ▶ Residual tech setup

# Agenda

- ▶ **Course goals**
- ▶ Intros
- ▶ Break
- ▶ Nuts and bolts
- ▶ Residual tech setup

# Broad goals

- ▶ Build upon your introductory programming course and to equip you with the computing literacy to conduct social science research in the age of “big data.”
- ▶ Two components
  1. **Workflow tools:** Git/GitHub; LaTeX; basic use of command line
  2. **Programming in messy contexts:** applied tasks in Python (data wrangling; basic text analysis); some SQL

# Content of QSS20: Modern Statistical Computing

## Topics:

- ▶ Data wrangling and viz
- ▶ APIs and web-scraping
- ▶ Text as data
- ▶ Supervised machine learning

# Content of QSS20: Modern Statistical Computing

## Topics:

- ▶ Data wrangling and viz
- ▶ APIs and web-scraping
- ▶ Text as data
- ▶ Supervised machine learning

## Workflow tools:

- ▶ Git/GitHub
- ▶ Command line (shell)
- ▶ LaTeX

# Content of QSS20: Modern Statistical Computing

## Topics:

- ▶ Data wrangling and viz
- ▶ APIs and web-scraping
- ▶ Text as data
- ▶ Supervised machine learning

## Workflow tools:

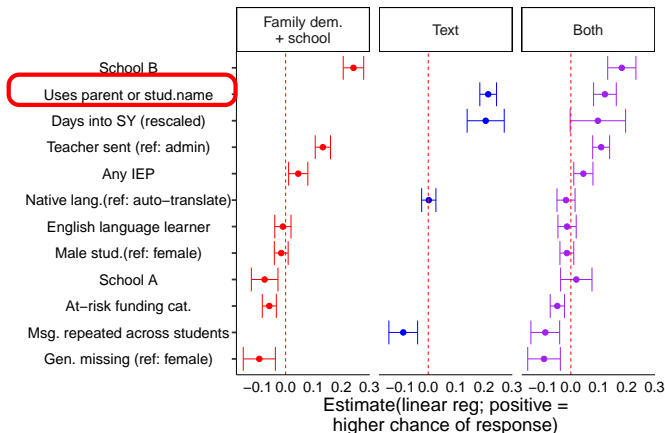
- ▶ Git/GitHub
- ▶ Command line (shell)
- ▶ LaTeX

## Components:

- ▶ DataCamp modules
- ▶ Class activities
- ▶ Problem sets
- ▶ Final project

# A recent example

The **box in red** shows that parents are more likely to respond to text messages from teachers when the teacher uses the parent or their child's name:





# Beyond the statistics, series of workflow and programming tasks before running regression

## 1. Acquire the data:

- ▶ **Ideal:** csv or database
- ▶ **Real:** excel file w/ variable number of tabs and spaces in column names; pdfs containing text; website

## 2. Clean the data:

p_name	s_name	msg_content
Rebecca Johnson	Jennifer John-son	Hi Ms. Johnson! Jenny did great on her math test.
Rebecca Johnson	Jennifer John-son	Hello Rebecca- I'm concerned about Jennifer's grades.

## 3. Reconcile different decisions in data cleaning:



Script #5 Running Thoughts/Comments/Questions #22

vickyme1 opened this issue on Sep 10, 2020 · 1 comment

```
id_orig_nowide_toconcat['PARENTID_constructed_1'] = np.nan
id_orig_nowide_toconcat['PARENTID_constructed_2'] = np.nan

id_orig_nowide_toconcat
```

- ☒ Your comment says 96% match, but I only get 91% match on my end (increase to ~94% after including SY1819). Is 96% still the case for you? Worried I'm using the incorrect file

```
# With local ids only, we get a 96% match rate! up from 75%
osse_merge_1920.found_osse.value_counts(normalize=True)
```

# How does QSS20 fit with other courses you might have taken/will take?

- ▶ **Data wrangling and visualization with focus on R:**
  - ▶ QSS17 (Data visualization): tidyverse; ggplot
  - ▶ QBS181 (Data wrangling): R and SQL
- ▶ **Deeper dives into the statistics/analysis side:** stats prereqs; some courses in COSC more focused on machine learning
- ▶ **Throughout:** focus on real-world policy applications and ethics of and policy context behind the data

# Overarching goal: transparency and reproducibility



# Why do those matter? Data science in high-stakes contexts

## Misuses of data science:

### Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica  
May 23, 2016

### AUTOMATING INEQUALITY

HOW HIGH-TECH TOOLS PROFILE,  
POLICE, AND PUNISH THE POOR

## Promoting responsible and equitable data science:



DATA FOR  
BLACK LIVES



DATA SCIENCE  
FOR  
SOCIAL GOOD










# Agenda

- ▶ Course goals
- ▶ **Intros**
- ▶ Break
- ▶ Nuts and bolts
- ▶ Residual tech setup

# Meet and introduce your neighbor!

- ▶ Name
- ▶ Favorite class at Dartmouth so far and why?
- ▶ If you could have any data source at your disposal, what would it be and what's a question you would ask?

## A bit about me

Where	What	Languages
	BA in Cultural Anthropology	None
	PhD in Sociology, emphasizing organizations and education	
  <small>GEORGETOWN UNIVERSITY</small> <b>MASSIVE DATA INSTITUTE</b> <small>McCourt School of Public Policy</small>	Organizer and trainer in computational methods	 python™
	Postdoctoral Fellow	 python™

# Course TAs and group tutor

TA: Ramsey Ash

- ▶ Ramsey took QSS20 in winter 2022 and was group tutor last quarter
- ▶ Will be present during most of classes as tech support/coding resource
- ▶ Will help with grading and be a resource via OH & Piazza

TA: You-Chi (Eunice) Liu

- ▶ Eunice took QSS20 in spring 2021
- ▶ Taking a 3A class, so go visit her in OH!
- ▶ In addition to grading, resource via OH & Piazza

Group tutor: Eleanor Sullivan

- ▶ Eleanor took QSS20 in winter 2022
- ▶ Leads group tutoring sessions, may also be a resource via Piazza



# Group tutoring sessions

- ▶ To support all students, especially those with less coding background
- ▶ Thanks to Peer Tutoring Program:  
<https://students.dartmouth.edu/academic-skills/Peer-Tutoring-Program/about-peer-tutoring-program>
- ▶ Tutoring session times **starting next Sunday, 01/15**: Sun 8-9pm; Mon 7-8pm; Thurs 9-10pm (these may change)
- ▶ Sign up for tutoring here (also on course page): <https://grouptutoring.dartmouth.edu/terms/23W/groups/12210>

# Collecting Student Info

Please complete this one brief survey right now:

<https://forms.gle/irAmC6sbBjEwKvL99>

Then take a 5-min stretch break.

# Agenda

- ▶ Course goals
- ▶ Intros
- ▶ **Break**
- ▶ Nuts and bolts
- ▶ Residual tech setup

# Agenda

- ▶ Course goals
- ▶ Intros
- ▶ Break
- ▶ **Nuts and bolts**
- ▶ Residual tech setup

# Course website: most authoritative guide

Please make sure to read the following pages most closely (can click on links in posted slides):

1. **Evaluation and grading:** [https://jhaber-zz.github.io/QSS20\\_site/docs/eval\\_grades\\_py.html](https://jhaber-zz.github.io/QSS20_site/docs/eval_grades_py.html)—covers four late days and exact grade breakdown
2. **Software setup:** [https://jhaber-zz.github.io/QSS20\\_site/docs/software\\_setup.html](https://jhaber-zz.github.io/QSS20_site/docs/software_setup.html)
3. **Course schedule (this may change):** [https://jhaber-zz.github.io/QSS20\\_site/docs/course\\_schedule.html](https://jhaber-zz.github.io/QSS20_site/docs/course_schedule.html)
4. **Final project overview:** [https://jhaber-zz.github.io/QSS20\\_site/docs/finalproject\\_overview.html](https://jhaber-zz.github.io/QSS20_site/docs/finalproject_overview.html)
5. **Options for doing your final project:** [https://jhaber-zz.github.io/QSS20\\_site/docs/finalproject\\_options.html](https://jhaber-zz.github.io/QSS20_site/docs/finalproject_options.html)

# Course components

1. **Most important - class sessions:** lab-based rather than lecture-based; hands-on practice with more advanced applications / work on problem sets
2. **Piazza-access via Canvas or link in online syllabus**
3. **Office hours** (must attend at least one OH or peer tutoring session)
4. DataCamp for review/basic syntax
5. Five problem sets
6. Social Impact Practicum/final project

# What to expect in an average class

<b>Time window</b>	<b>What</b>
3:30-4:00	Slides; DataCamp questions
4:00-4:10	Break into small groups
4:10-5:00	Work with assigned group on class activity; Ramsey and I circulate around
5:00-5:15	Reconvene as a group and go over questions; outline any prep for next class
5:15-5:20	Anonymous sticky notes with lesson & question

Might deviate if we have visitors/guest speakers, e.g., related to final project/SIP

# Piazza: course communication

- ▶ Access via Canvas or online course page/syllabus
- ▶ Please add an image to your profile by next week's class
- ▶ **Expectations:**
  - ▶ **Order:** first tag the TA(s); they can defer to me if they have problems answering
  - ▶ **Direct emails to me: only** for family emergencies and other personal issues
  - ▶ I will respond within 24 hours on weekdays; by Monday AM on weekends; before a problem set is due, will respond to all questions posted before **3 pm** on due date but not questions between 3 pm and midnight when due



# Office hours

## ► My office hours

- **Two slots:** Mondays 2:15-3:15 pm; Wednesdays 2:15-3:15 pm
- **Location:** Silsby 103 (main floor)
- In-person drop-ins and groups are welcome!
- But if you want to meet privately and/or via Zoom, you can sign up in advance (**no later than the midnight before**) via Calendly (I'll attach to online course schedule):  
<https://calendly.com/jaren-haber-qss>
- When signing up, please let me know if you want privacy and indicate your preferred format:
  1. In person (at my office)
  2. Zoom
- If my door is closed during office hours, that means I am in a private meeting. Please wait in the hall.
- Check Piazza for TA office hours





# Course components

1. Most important - class sessions: lab-based rather than lecture-based; hands-on practice with more advanced applications / work on problem sets
2. Piazza
3. Office hours
4. **DataCamp for review/basic syntax**
5. **Five problem sets**
6. Final research project

# DataCamp: make sure to join via our specific course page so assignments show up

See bottom of course syllabus main page for join link.

Post to [datacamp](#) on Piazza if you need access via a different email.

TITLE ↕	ASSIGNEES ↕	STATUS	DUE BY ▼
 Regular Expressions in Python Regular Expressions for Pattern Matching Chapter	Organization	Past Due	Oct 5, 15:30 EDT
 Joining Data with pandas Data Merging Basics Chapter	Organization	Past Due	Sep 28, 15:30 EDT
 Joining Data with pandas Merging Tables With Different Join Types Chapter	Organization	Past Due	Sep 28, 15:30 EDT
 Data Manipulation with pandas Course	Organization	Past Due	Sep 14, 15:30 EDT

Meant as auxiliary tool/playing a minor role so that you're prepared for in-class activities and so we don't need to review basic syntax. So graded on completion-only basis and only 5% of grade, but if you'd prefer to skip, can reapportion the 5% to the first problem set

# Five problem sets

- ▶ Problem set one will be on GitHub in next few days and is due Sunday, 01.15
- ▶ **Others:** see online course schedule
- ▶ **For each:**
  - ▶ Start well in advance (at least 3-4 days) and space out the parts (Pset 1 should be largely review from COSC 1 and the initial DataCamp modules)
  - ▶ May devote some class time pre deadline to work on the pset/answering questions

# Course components

1. Most important - class sessions: lab-based rather than lecture-based; hands-on practice with more advanced applications / work on problem sets
2. Piazza
3. Office hours
4. DataCamp for review/basic syntax
5. Five problem sets
6. **Final research project**

# What's a Social Impact Practicum?

- ▶ Sponsored by Dartmouth Center for Social Impact

- ▶ From them:

*A Social Impact Practicum (SIP) is a project-based experiential learning opportunity connecting undergraduate courses at Dartmouth with community needs identified by nonprofit organizations throughout the Upper Valley.*

*In other words, a SIP is a real-world project with real-world impact.*

- ▶ Can find a database of other SIPs here: <https://students.dartmouth.edu/social-impact/programs-initiatives/students/social-impact-practicums-sips/social-impact-practicum-sip-course>
- ▶ Ashley Doolittle (SIP director) will visit us next week and is happy to meet with students

Partner organization: UNH Center for Start Services

<https://iod.unh.edu/projects/center-start-services>



**THE CENTER FOR  
START SERVICES**

*Institute on Disability / UCED  
University of New Hampshire*

# Overview

- ▶ **Option 1:** How do medical professionals interact with patients with Intellectual and Developmental Disabilities (IDD), and how can medical student training improve this?
- ▶ **Option 2:** How have those with disabilities fared during COVID-19, and what racial inequalities do we see in its impact?



# Project examples from a previous class

- ▶ Geo-visualization of locations of job sites relative to Census tract attributes (e.g., migration rates; unemployment): [https://github.com/rebeccajohnson88/qss20\\_s21\\_proj/blob/main/memos/final\\_papers/dol\\_geocoding\\_writeup.pdf](https://github.com/rebeccajohnson88/qss20_s21_proj/blob/main/memos/final_papers/dol_geocoding_writeup.pdf)
- ▶ Causal analysis of relationship between inspection capacity and findings of legal issues: [https://github.com/rebeccajohnson88/qss20\\_s21\\_proj/blob/main/memos/final\\_papers/dol\\_opmstaffing\\_writeup.pdf](https://github.com/rebeccajohnson88/qss20_s21_proj/blob/main/memos/final_papers/dol_opmstaffing_writeup.pdf)
- ▶ Natural language processing of job contracts: [https://github.com/rebeccajohnson88/qss20\\_s21\\_proj/blob/main/memos/final\\_papers/dol\\_textasdata\\_writeup.pdf](https://github.com/rebeccajohnson88/qss20_s21_proj/blob/main/memos/final_papers/dol_textasdata_writeup.pdf)
- ▶ Supervised machine learning predicting investigations/violations: [https://github.com/rebeccajohnson88/qss20\\_s21\\_proj/blob/main/memos/final\\_papers/dol\\_predictviol\\_writeup.pdf](https://github.com/rebeccajohnson88/qss20_s21_proj/blob/main/memos/final_papers/dol_predictviol_writeup.pdf)

# Structure of project

- ▶ **Milestone 1:** memo or plan for what question you'll ask and analyses you'll run
- ▶ **Milestone 2:** set up your repository and start coding
- ▶ **Final outputs (see course website for more details):**
  - ▶ Final presentation (done in Beamer; LaTeX-based powerpoint software)
  - ▶ Short 10-page report (done in LaTeX)
  - ▶ Github repo and readme with all code to reproduce analyses

# Agenda

- ▶ Course goals
- ▶ Intros
- ▶ Break
- ▶ Nuts and bolts
- ▶ **Residual tech setup**

# Checklist before next class

1. Have you reviewed [the online course syllabus](#)?
2. Are you set up on DataCamp?
  - ▶ Ping in 'datacamp' on Piazza if you need help getting added
3. Have you completed the DataCamp modules (or are you already feel comfortable with basics in Python)?
4. Are you on Piazza?
5. Have you completed your software setup? This means going through: [https://jhaber-zz.github.io/QSS20\\_site/docs/software\\_setup.html](https://jhaber-zz.github.io/QSS20_site/docs/software_setup.html)
  - 5.1 Create a GitHub account and [personal access token](#)
  - 5.2 Create an Overleaf account
  - 5.3 Local Python installation
  - 5.4 Terminal: native on Mac and Linux; WSL or terminal emulator for Windows