ACE 592 SAE:

Data Science for Applied Economics

Instructor: Jared Hutchins

TA: Rocio Valdebenito

What is Data Science?

What is data science?

Data science is a somewhat catch all term for what is now considered essential skills and practice for data analysis and processing.

or according to Dr. Matthew Brett:

"Data science is an approach to data analysis with a foundation in code and algorithms."

or according to a Berkeley course:

"Data Science is about drawing useful conclusions from large and diverse data sets through exploration, prediction, and inference."

…ok, so what is data science?

As a student that analyzes data, most of it you do already: you process it, you use statistics to describe it, or you visualize it.

In addition to this, **data science** adds some crucial things to our toolkit:

- New **kinds of data**, especially *unstructured data.*
- **Workflow** using open source software and tools.
- New **analytical** (e.g. machine learning) and **visualization** techniques.

Why does it matter to our field?

Here are four reasons:

#1: Getting data

#2: Learning new tools.

#3: Community and collaboration

#4: Presenting your work

Reason #1: There is more data out there than ever before, but it isn't always in .csv format.

To learn about human behavior, the most cutting-edge economics papers are using:

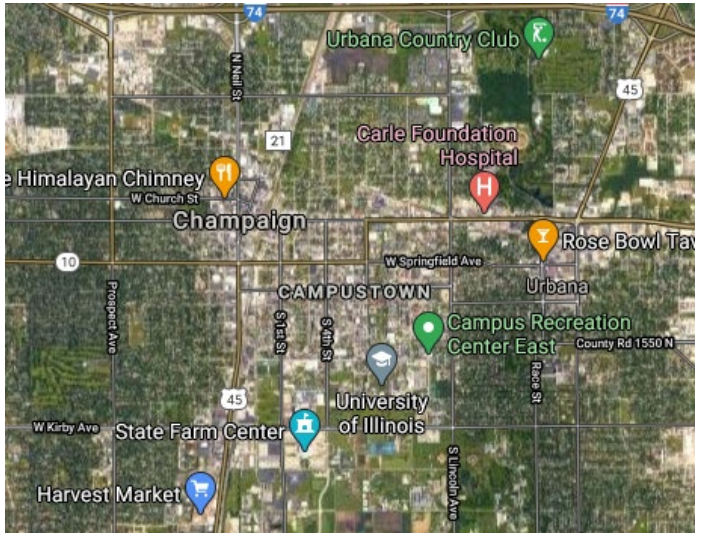
- Text data (government reports, tweets etc.).
- Image data (weather data, soil data, satellite images).
- Information scraped from websites.

This drastically increases the number of things we can research!

Reason #1: There is more data out there than ever before, but it isn't always in .csv format.

This is data

So is this



Reason #2: Coding and algorithms help us do our work better.

Our main coding tools for this course:

Why Python/Anaconda?

- Multi-purpose and appropriate for many tasks.
- A good introduction to object-oriented programming.
- Very good package library.

Reason #3: Coding using git and GitHub = Accessible and transparent research

git

- Becoming standard practice for researchers to use GitHub to make their research accessible and reproducible.
- Don’t reinvent the wheel; learn from what others have already done.
- Easy to track your own work and collaborate.

Reason #4: New ways to present your work.

Jupyter makes it easy to:

- Code interactively.
- Present and comment your code.
- Make presentations.
- Make your work interactive.

Bonus Reason: Employability

JOE Listings (Job Openings for Economists)

August 1, 2020 – January 31, 2021

The Federal Reserve Bank of Philadelphia

Research

[+] Data Scientist/Machine Learning Economist

Air Liquide

Research & Development

Computational & Data Science

[+] Data Scientist / Economist

Uber

[+] Economist (Data Scientist)

The World Bank

Equity Policy Lab

Poverty and Equity Global Practice

[+] Poverty Economist / Data Scientist

Stanford University

Immigration Policy Lab

[+] Postdoctoral Research Fellow (Data Science, Health)

Course Objectives

By the end of the course, you will be able to:

1. Obtain and process text, image, and numeric data using Python.
2. Analyze data using basic data visualization and machine learning in Python.
3. Construct a git repository and collaborate on a research project on GitHub.
4. Document code and communicate results using Jupyter notebooks.

Course Objectives

Objectives 1 and 2

For obtaining, processing, and analyzing data.

Objective 3

For tracking our work and collaborating.

Objective 4

For presenting and making our work transparent.

Lectures and Discussion

Lectures

- Tuesday and Thursday, 4 - 5:20 pm.

Discussion Sections

- Friday, 10 am.

Prerequisites

No specific prereqs, but it helps if you are:

- Interested in or currently doing quantitative research.
- Familiar with coding and scripting.
- Taken a statistics or econometrics course.

I do not assume you know Python, git, or Jupyter.

Our Basic Plan

1. Introduction
2. Text as Data
3. Images as Data
4. Numbers as Data
5. Analysis

Our Basic Plan

1. Introduction (3 weeks)
    a. Python, git, and Jupyter basics.
    b. Pandas, numpy, matplotlib.
    c. APIs, requests, scraping.

This is a good time to catch up if you feel have gaps in your knowledge.

Our Basic Plan

2-4. The Data Types (3 weeks each)

a. How do we obtain it?
b. How do we process/clean it?
c. What tools can we use to analyze it?
d. What economics questions need it?

Our Basic Plan

5. Analysis and Visualization (3 weeks)
    a. How do we decide how to visualize data?
    b. Basics of unsupervised learning.
    c. Basics of supervised learning.

When do we use each of these tools?

Assessments

- 3 Homework Assignments + 1 Bonus Assignment (20 pts each, 60 pts total)
  - Submitted individually as a Jupyter notebook.
  - One for each data type: text, image, and numeric.
- One final, team project (40 pts)
  - Presentation (10 pts)
  - GitHub Repo (10 pts)
  - Write up (20 pts)

Homeworks

- One for each data module.
- Written up individually, but group discussion is encouraged.
- Submit as a Jupyter notebook to your Box folder.
- Due the week after we finish discussing the topic.

Final Project

- Teams must be:
  - At least 3 people.
  - **Formed by February 18th.**

- Projects must consist of:
  - An analysis of an economics question, cleared by me by **March 18th**
  - A GitHub repository of your project.
  - A presentation done at the end of the semester.

Extra Credit Assignments

- **Data Viz Competition**
  - For every homework, you may submit your visualization to receive 2 points extra credit.
  - Submitted visualizations will be voted on by the class to judge who made the best visualization.

Resources to Learn

No textbook, but we have the following resources:

- DataCamp tutorials (you are all signed up for it).
- Various online textbooks (available in the syllabus).
- Stack Overflow (if you are *really* desperate).

Much of learning to code is self-guided, and:

How much you learn $\propto$ How much you work

Things to do right now:

1. **Get a GitHub account.**
2. **Sign up for DataCamp.**
3. **Install Anaconda on your computer.**

Let Rocio and I know if you have any issues here.

Things to start thinking about

1. **Readying your computer for working with our tools.**
2. **Forming your group (must be formed by Feb 18).**
3. **Thinking about a research topic (March 18)**