

DSFA

Spring 2019

Lecture 1

STSCI+ORIE+CS 1380

Introduction

*I would found an institution
where any person can study
data science. - Ezra Cornell*

A course for anyone who wants to study *data visualization, prediction, machine learning, and programming in Python*. We'll analyze real-world data sets on crime, health, transportation, literature, and more!

STSCI + ORIE + CS 1380
Data Science For All
Spring 2019 TR 10:10-11:25am

No experience required – Open to all – Fulfills MQR-AS

Who are we?



- Professor Booth

- Professor Wilson

+ Teaching Assistants

Who are we?



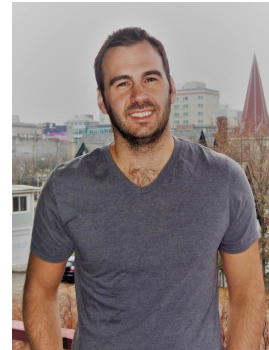
- Polina Kirichenko (pk575)
Operations Research



- Sean Sinclair (srs429)
Operations Research



- Skyler Seto (ss3349)
Statistics



- Antonio Sirianni (ads334)
Sociology
-

Who are we?



- Victoria Bao (yb244)
Statistics



- Daniel Sanky (ds869)
Information Science
-

Who are you?

Take this class if you:

- are **curious** about data
- don't know much/any **CS**
- don't know much/any **Stats**
- don't know much/any **OR**

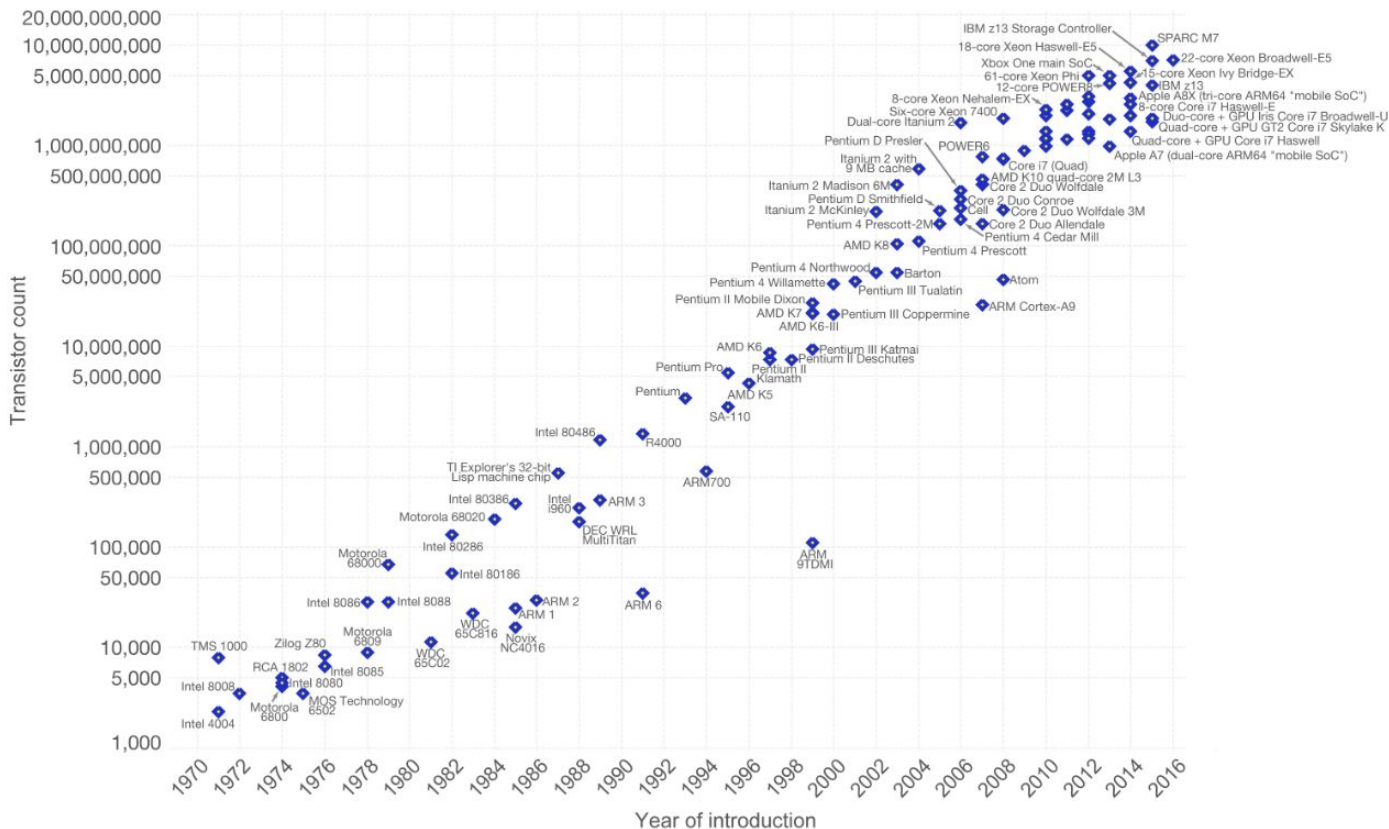
Don't take this class if you:

- have already taken both **CS & Stats** intro classes
(it will be too slow for you)

Why Data Science?

Moore's Law – The number of transistors on integrated circuit chips (1971-2016)

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are strongly linked to Moore's law.

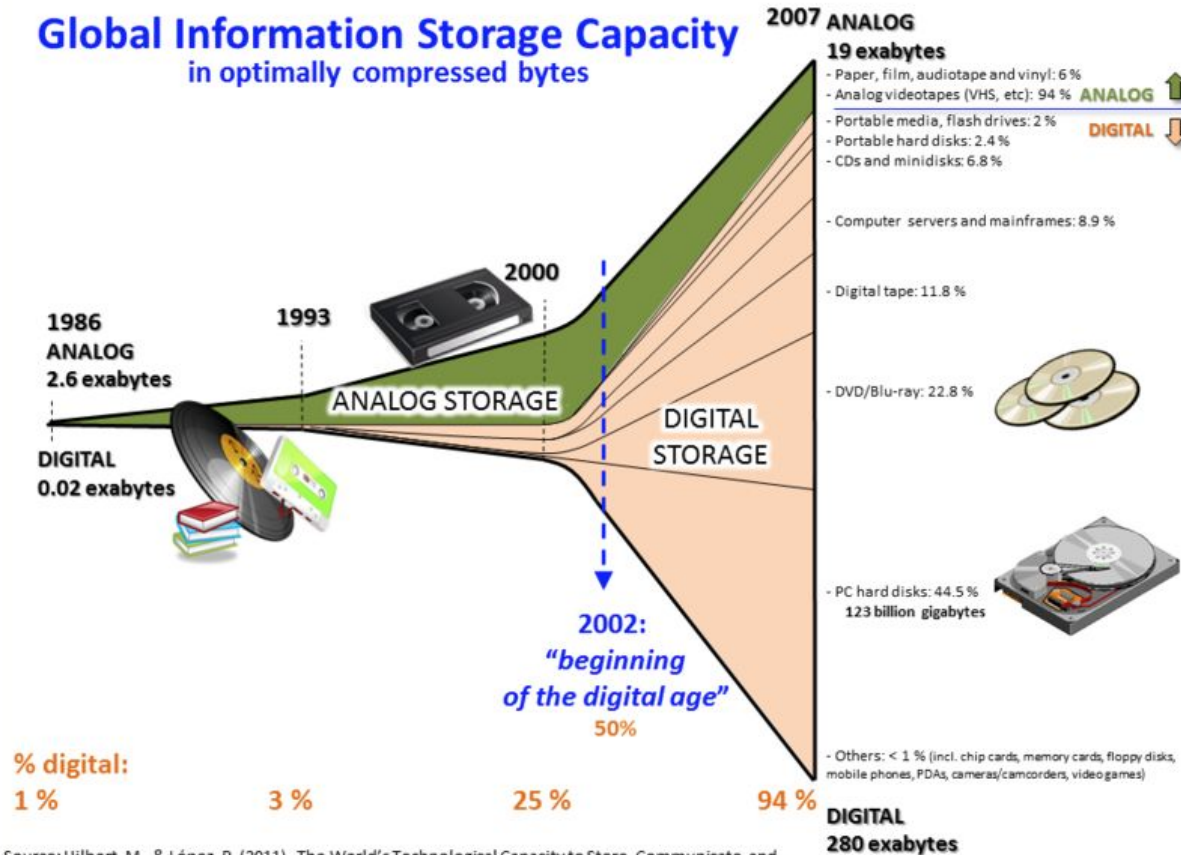


Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)

The data visualization is available at [OurWorldinData.org](https://www.ourworldindata.org). There you find more visualizations and research on this topic.

Licensed under CC-BY-SA by the author Max Roser.

Global Information Storage Capacity in optimally compressed bytes



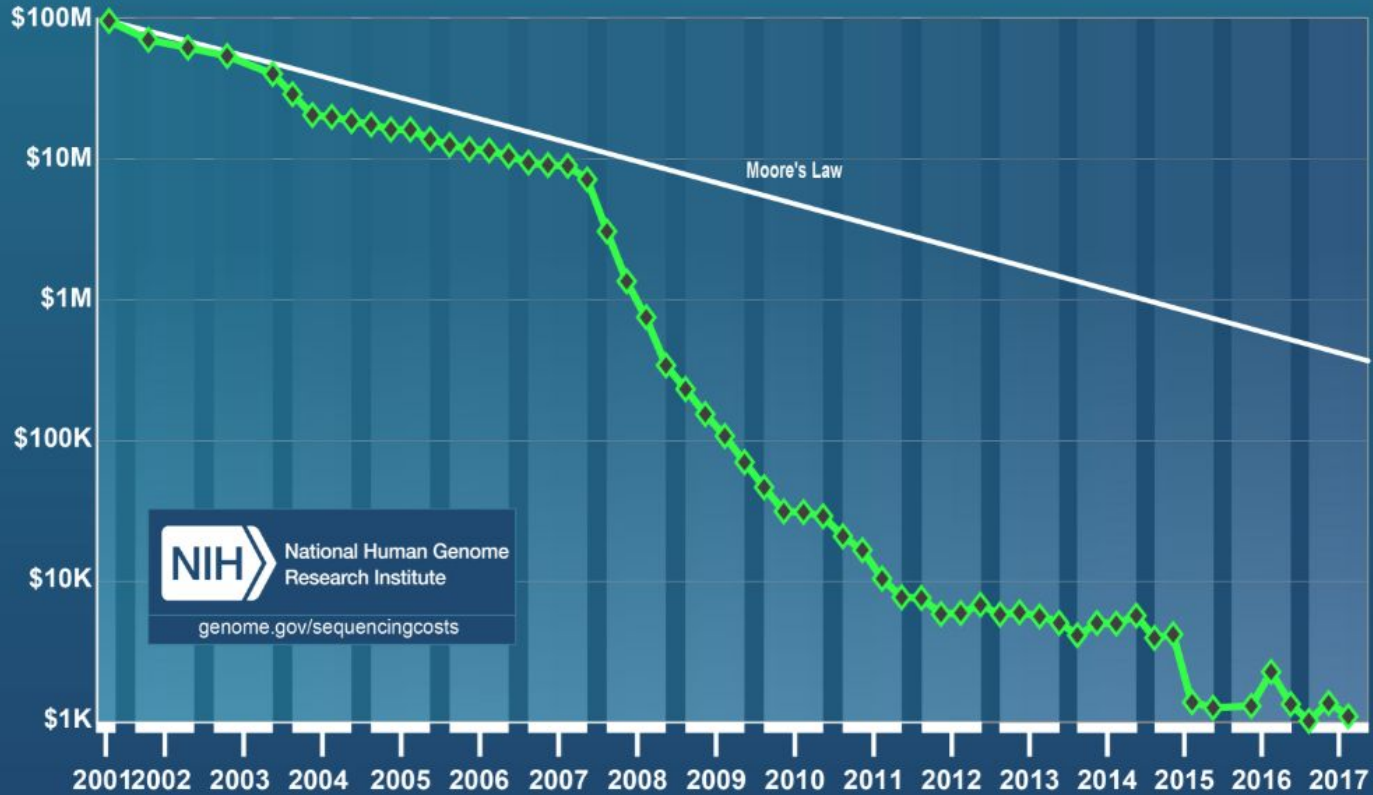
Growth of and digitization of global information-storage capacity^[1]



Digital Data Terminology

- **Bit** - binary unit: 0/1
 - **Byte** - eight bits
 - **Kilobyte** - 2^{10} or 1024 bytes
 - **Megabyte** - 2^{20} bytes or 1024 kilobytes
 - **Gigabyte** - 2^{30} bytes or 1024 megabytes
 - **Terabyte** - 2^{40} bytes or 1024 gigabytes
 - **Petabyte** - 2^{50} bytes or 1024 terabytes
 - **Exabyte** - 2^{60} bytes or 1024 petabytes
-

Cost per Genome



Who needs data science?

- Data scientists
- OR, CS, Stats majors
- Lawyers
- Doctors
- Citizens
- Readers of the news

...ALL

National Challenge

In the United States, it is reported that in 2018 there will be more than 490,000 data science positions available, but only 200,000 qualified people to fill the roles. The **average size of a graduate class of data science students is 23 students**. With approximately only 110 universities offering data science studies, the growing market will continue to pressure the supply in the US.

January 22, 2016

Data Scientists: The Myth and the Reality

Seamus Breslin

OCT. 17, 2017 AT 6:00 AM

The Supreme Court Is Allergic To Math



The Supreme Court does not compute. Or at least would rather not. The justices, the most powerful jurists in the land, seem to have a reluctance — even an allergy — to taking math and statistics seriously.

For decades, the court has struggled with quantitative evidence of all kinds in a wide variety of cases. Sometimes justices ignore this evidence. Sometimes they misinterpret it. And sometimes they cast it aside in order to hold on to more traditional legal arguments. (And, yes, sometimes they also listen to the numbers.) Yet the world itself is becoming more computationally driven, and some of those computations will need to be adjudicated before long. Some major artificial intelligence case will likely come across the court's desk in the next decade, for example. By voicing an unwillingness to engage with data-driven empiricism, justices — and thus the court — are at risk of making decisions without fully grappling with the evidence.

quantify partisan gerrymandering: “It may be simply my educational background, but I can only describe it as sociological gobbledygook.” This was



NEW YORK TIMES BESTSELLER



WEAPONS OF MATH DESTRUCTION



HOW BIG DATA INCREASES INEQUALITY
AND THREATENS DEMOCRACY

CATHY O'NEIL

A NEW YORK TIMES NOTABLE BOOK

Standing is good for you, but wait, N=50!!! Why would Psychological Science or The Economist publish a study with such sample size?



Standing is good for your mind as well as your body

It seems to promote cognitive performance

ECONOMIST.COM

Higher coffee consumption associated with lower risk of early death

Date: August 27, 2017

Source: European Society of Cardiology

Summary: Higher coffee consumption is associated with a lower risk of early death, according to new research. The observational study in nearly 20 000 participants suggests that coffee can be part of a healthy diet in healthy people.

What is Data Science?

Answering questions from data using computation

- **Exploration**
 - Identifying patterns in information
 - Uses visualizations
 - **Inference**
 - Quantifying whether those patterns are reliable
 - Uses randomization
 - **Prediction**
 - Making informed guesses
 - Uses machine learning
-

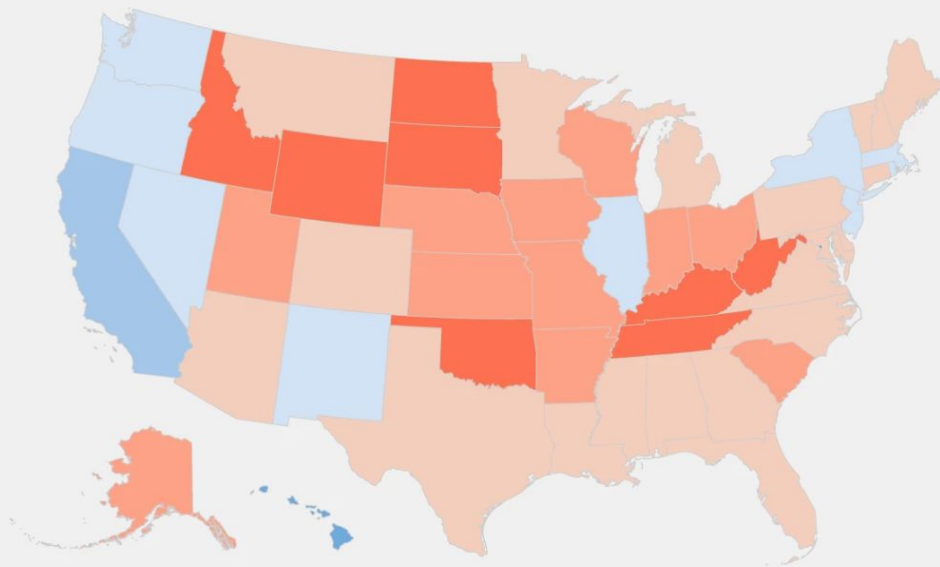
Data Science Stories

- **Agriculture**
 - When will the harvest be ready?
 - How large will the harvest be?
 - **Political Campaigns**
 - How to summarize information from different polls?
 - What is the chance of winning each state or district?
 - Who might be willing to donate if I asked? How to ask?
 - **Medicine**
 - Which patients are at risk of some disease?
 - Which patients would benefit from surgery?
-

Polls underestimated Trump in red states, Clinton in blue states

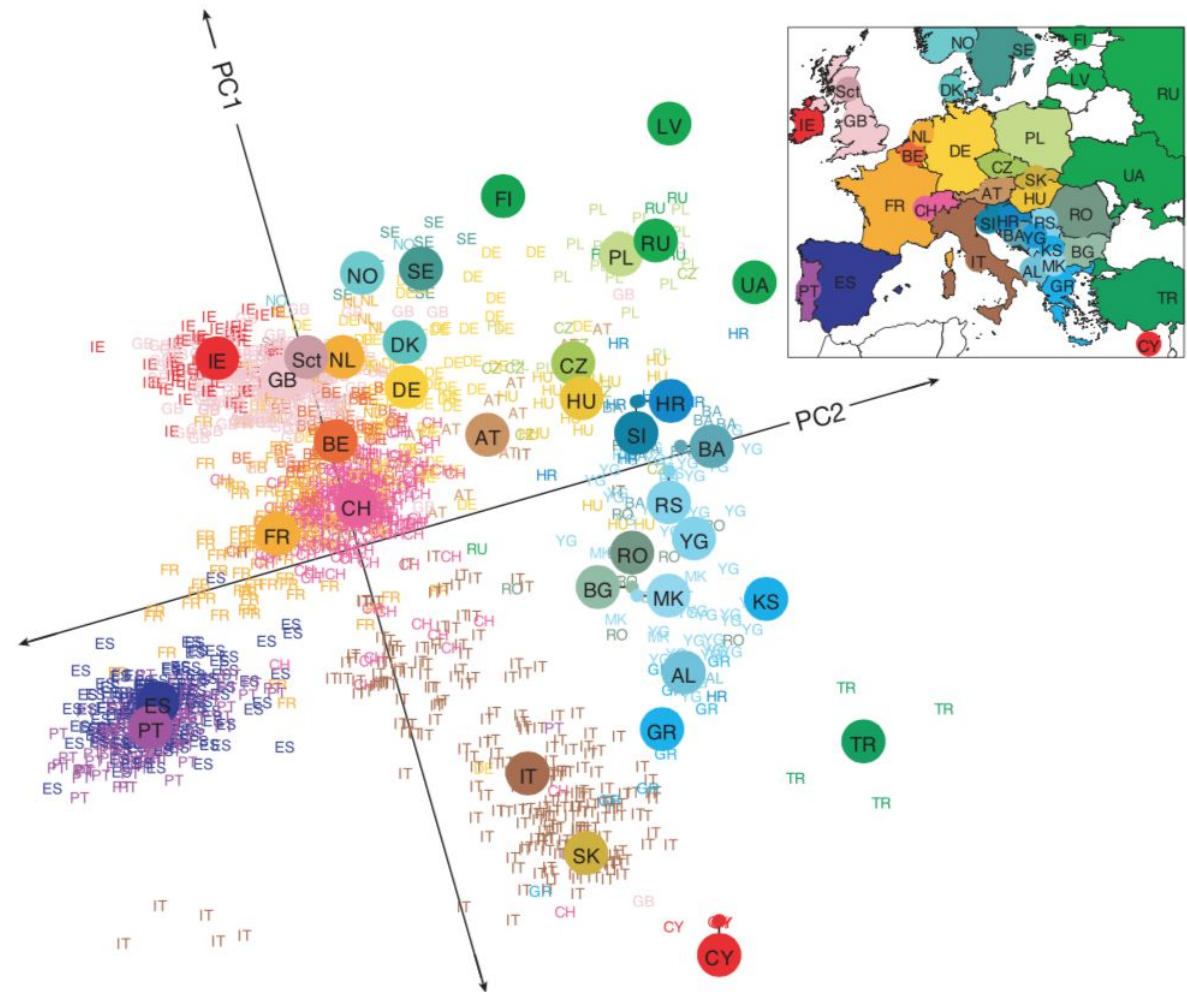
2016 election results vs. FiveThirtyEight's adjusted polling average by state

REPUBLICAN VOTE MARGIN RELATIVE TO POLLS



“Genes mirror
geography
within Europe”

Novembre et al.
Nature 2008



Data Science in Action

Course Structure

How DSFA works

- Lecture Tuesday and Thursday
 - Participation counts for grade
 - Section every week on W or Th
 - Including this week!
 - Attend the one you signed up for
 - Project partner must be enrolled in same section
 - Assignments:
 - Labs (about 10 total)
 - Homework (about 8 total)
 - Projects (3 total)
 - Exams:
 - Two prelims + final exam
-

(Tentative) Section Schedule

Section	Time	Room	TA
DIS201	W 12:20-2:15pm	Thurston Hall 202	Antonio/Polina
DIS202	W 2:30-4:25pm	Upson Hall 202	Daniel
DIS203	W 7:30-9:25pm	Hollister Hall 362	Sean/Skyler
DIS205	R 12:20-2:15	Thurston Hall 202	Victoria

More info:

(Tentative) Office Hours Schedule

TA	Time	Room
Antonio	Wed 4:30-6:30pm	TBD
Daniel	Sat/Sun 2:30-3:30	TBD
Polina	Fri 2-4pm	TBD
Sean	Mon/Wed 1-2pm	TBD
Skyler	Tues/Thurs 9:30-10:30	TBD
Victoria	Mon 12:30-2:30	TBD

Policies, Grading, Etc.

- Details will be posted on the course website
- Blackboard
- Online textbook

Getting help

Questions about material:

- Ask a friend
- Ask on piazza
- Go to section
- Go to office hours

Logistical questions:

- Ask your section TAs
-

Academic Integrity

- Labs:
 - Work together as much as you'd like
- Homework and projects:
 - All work you submit must be your own
 - Share ideas (eg, in English) not solutions (eg, code)

In particular:

- Don't post code on Piazza
 - Cite your sources (including other students)
-

Now what?


- (Now) If you're not enrolled yet sign up
 - (Tomorrow or Thursday) Go to section
 - (By Thursday) Read [Chapter 1](#) (and 2) of the textbook
 - (Constantly) Tell your friends about this class
 - Everyone should take this class
 - There's still space
 - And it's not too late
 - (Next week) Buy an iClicker at the Cornell Bookstore
 - (By the add deadline) [Purchase access](#) to [Vocareum](#)
-

Reef Polling

- Answer in-class quiz questions using a smartphone
 - Create an account from the [login page](#)
 - Free 14 day trial. Six month subscription for \$15
 - Use your Cornell email and NetID to sign in

 - If you use an iClicker it must be registered on Blackboard (and you don't need a Reef account)
-

iClicker Reef login

 iClicker Reef

Email

Password

Remember Me [Forgot Password?](#)

[Sign In](#)

Don't have an account? [Sign Up!](#)

Need to sign in through your campus portal?
If you don't see your university listed, sign in above.

[Go →](#)

Install Anaconda?



Home

Environments

Learning

Community







Documentation

Developer Blog

Feedback

Twitter YouTube GitHub

Applications on Channels

 <p>jupyterlab</p> <p>0.32.1</p> <p>An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.</p> <p>Launch</p>	 <p>jupyter notebook</p> <p>5.5.0</p> <p>Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.</p> <p>Launch</p>	 <p>qtconsole</p> <p>4.3.1</p> <p>PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.</p> <p>Launch</p>
 <p>spyder</p> <p>3.2.8</p> <p>Scientific PYTHON Development EnviRnment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features</p>	 <p>vscode</p> <p>1.26.1</p> <p>Streamlined code editor with support for development operations like debugging, task running and version control.</p>	 <p>glueviz</p> <p>0.13.3</p> <p>Multidimensional data visualization across files. Explore relationships within and among related datasets.</p>

Acknowledgement

This course is based on [Data 8](#), a course taught by Ani Adhikari and John DeNero at the University of California, Berkeley. They and their teaching assistants have developed many of the materials we are using in our own course. We are using those materials with their permission, which we gratefully acknowledge.
