Introduction to Data Science

UNIVERSITY OF
SAN FRANCISCO

Abbie M Popa

BSDS 100 - Intro to Data Science with `R`

Outline

- Course Overview
- What is Data Science?
  - A brief history
  - Applications

Part I: Course Overview

A Little About Me

- B.S. Cognitive Neuroscience (Brown University)
- Ph.D. Neuroscience (UC Davis)
  - Studied how anxiety in teenagers affects electrical signals in the brain related to attention and control of behavior
  - Used data science tools to make sense of large messy data generated by human brains
- Now... USF Data Institute
  - Will continue application of data science to understand signals from the brain, now applying network based analyses
  - Teaching BSDS100!

A Little About Me

- Born in Wisconsin, grew up in Pennsylvania
- I enjoy cooking
- I love playing board games
- I once ran a marathon, very slowly
- I love my cat

Thank You!

Thank you to Paul Intrevado and James Wilson for original course materials!

Course Description and Syllabus

All lecture notes, the syllabus, assignments, and course description are available at this course website:

https://github.com/abbiepopa/BSDS100

Course Expectations

My expectations for you:

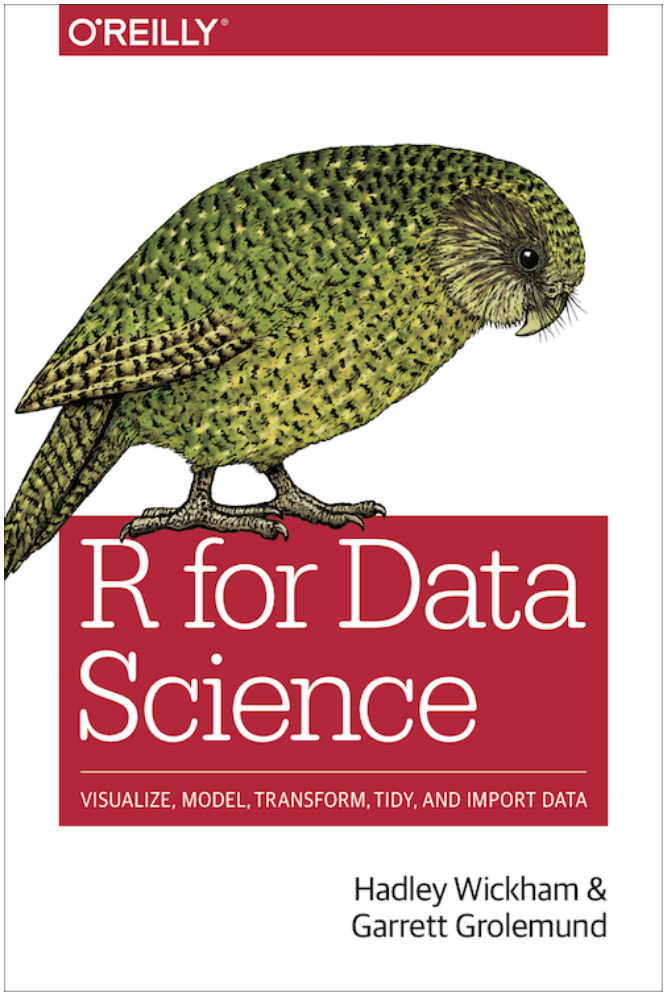
- Brief review of syllabus
- Attempt all activities, stay focused and on task during in class activities
- Respect each other (don't take over someone else's keyboard!)

Course Expectations

You can expect from me:

- I will be available to answer questions by e-mail or in office hours
- I will respect you
- This class is a priority for me, I will be prompt in my responses and uploads of course material

Main Text



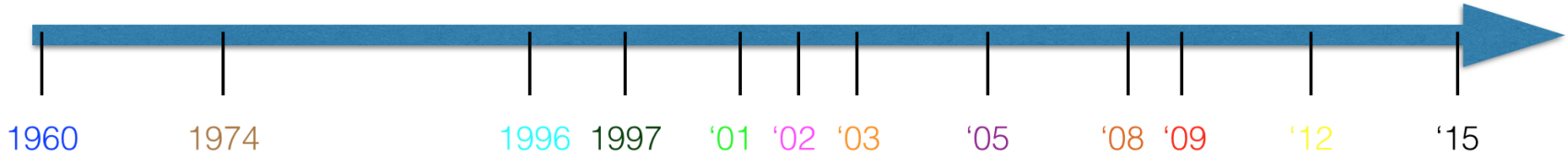
Available online here: http://r4ds.had.co.nz/index.html

Part II: What is Data Science?

What is Data Science?

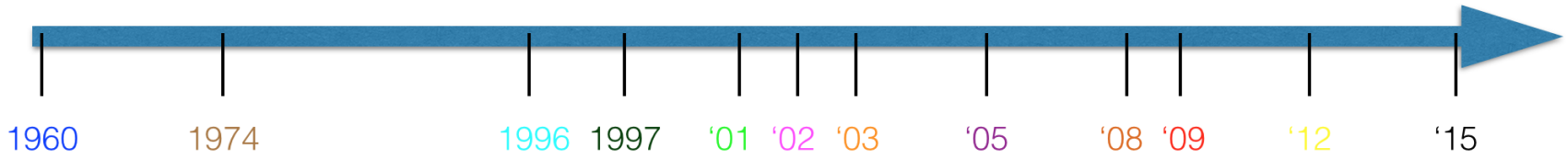
- **Wikipedia**: “the extraction of knowledge from data."
- A precise definition is a bit unclear and has faced much controversy... (we’ll see more on this in a moment)
- Practitioners tend to agree on the *components* of data science:
  - gathering and cleaning data
  - database management
  - exploratory analysis
  - predictive modeling
  - data summary and visualization

The Evolution of Data Science



- 1960: Peter Naur (CS Ph.D.) published *Datalogy: the science of data and its place in education.*
- 1974: Peter Naur published *Concise Survey of Computer Methods*.
  - defines data science as “the science of dealing with data, once they have been established."
  - continues to say that “... the relation of the data to what they represent is delegated to other fields and sciences."

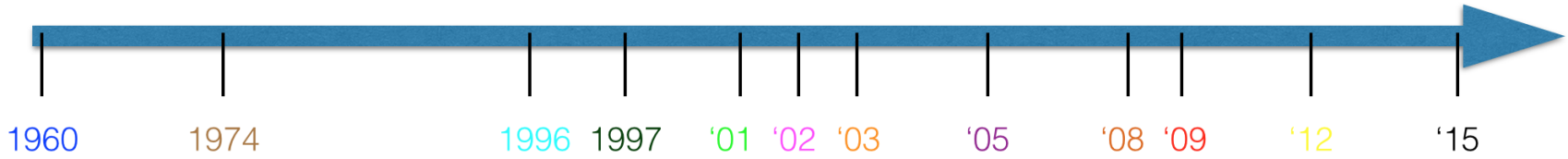
The Evolution of Data Science



- 1996: International Federation of Classification Societies meet in Tokyo and for the first time include “data science" in the conference title: “Data science, classification, and related methods."

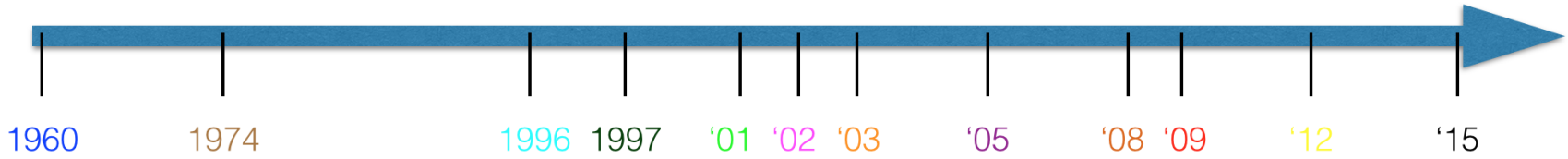
- 1997: C.F. Jeff Wu gave the inaugural lecture “Statistics = Data Science?" for appointment to the H. C. Carver Professorship at the University of Michigan.

The Evolution of Data Science



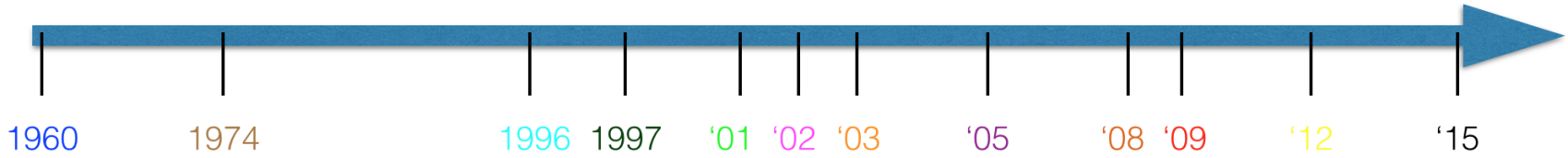
- 2001: William Cleveland (Bell Labs) published *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*.
  - Sets forth 6 areas for a university department involving statistics.
- 2002: *Data Science Journal* is launched
  - Focus on data systems, publications on internet, and applications
- 2003: *Journal of Data Science* is launched
  - Focus on application of statistical and quantitative methods

The Evolution of Data Science



- 2005: National Science board redefines data scientists:
  - “The information and computer scientists, data and software programmers, disciplinary experts, ... who are crucial to successful management of a digital data collection whose primary activity is to conduct creative inquiry and analysis"
- 2008: DJ Patil (LinkedIn) and Jeff Hammerbacher (Facebook) coined the term "data scientist" to define their jobs

The Evolution of Data Science



- January, 2009: Hal Varian (chief economist at Google) writes that "... the sexy job in the next 10 years will be statisticians."
- October, 2012: Harvard Business Review publishes "Data Scientist: The Sexiest Job of the 21st Century."
- February 5th, 2015: DJ Patil appointed as the first Chief Data Scientist in the White House.

Applications

Marketing analytics, sports analytics, biotechnology, social experiments, e-commerce, government analysis, ...

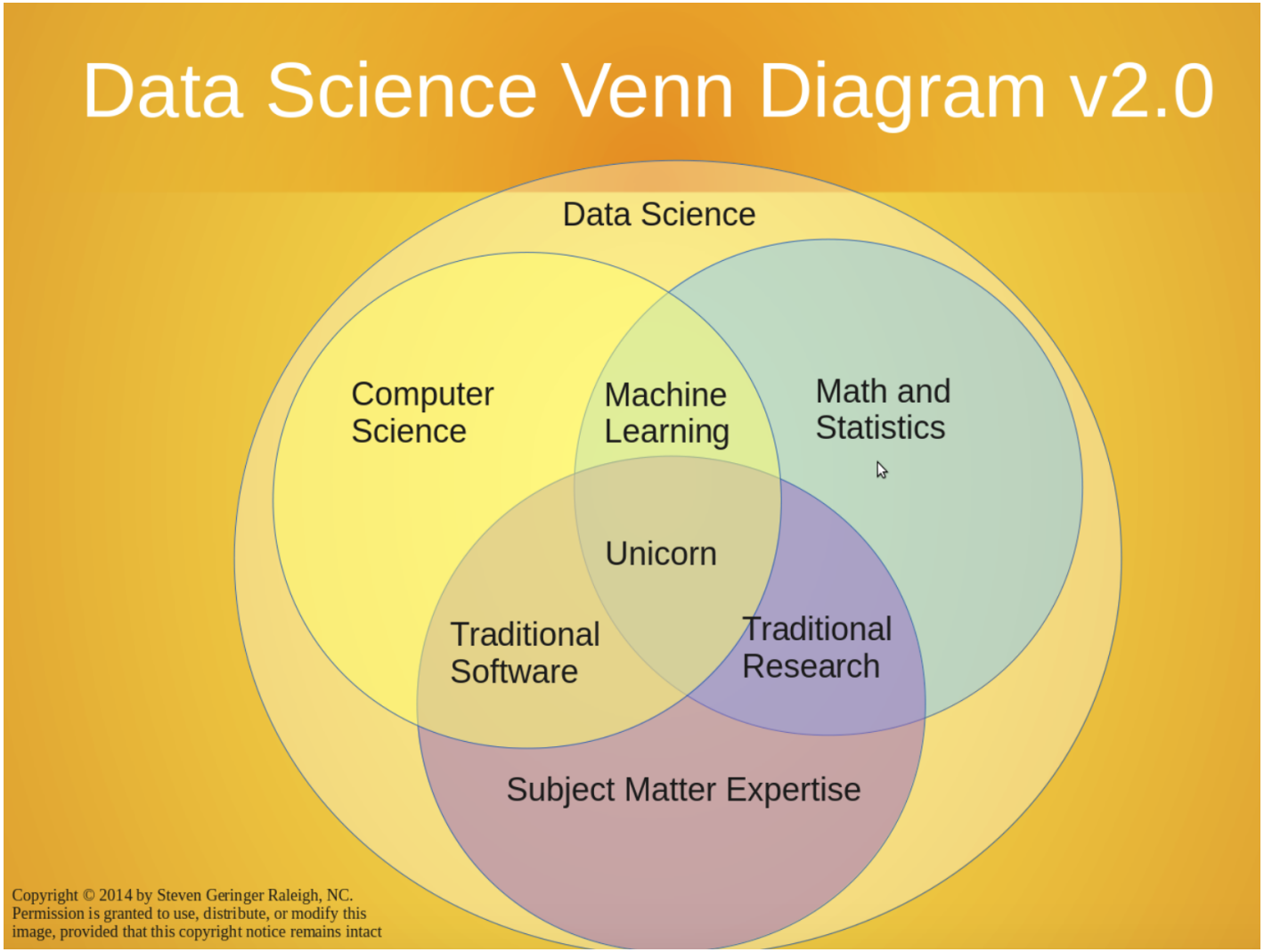
Some Examples

- **In Academia (STEM)** - Clustering teenagers into groups based on results from a wide range of neuropsych surveys
- **In Academia (Humanitites)** - Data mining of medieval texts revealed apothecaries used bioactive ingredients https://www.technologyreview.com/s/611751/data-mining-medieval-text-reveals-medically-bioactive-ingredients/
- **In non-profit** - Human rights organizations used data modeling to produce more accurate casualty estimates in Syria
- **In tech sector** - What type of hotels should we advertise to someone browsing our website?

Why Data Science?

- Size, complexity, and amount of data
  - Predicted $\approx$ 40 trillion gigabytes of data in 2020; up from 130 billion in 2005!
  - Big data requires innovative techniques for analysis
- *McKinsey*: "The U.S. faces a shortage of 140K - 190K people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data." (May, 2011)
- *Harvard Business Review*: "Data Scientist: The Sexiest Job of the 21st Century." (October, 2012)

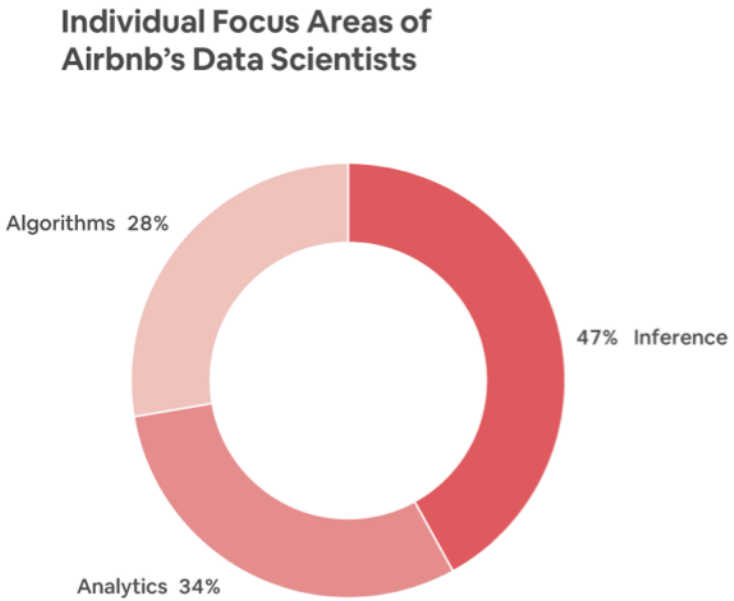
Data Scientists: The unicorn industries want?



Data Scientists: The unicorn industries want?

- The field is inherently interdisciplinary
  - mathematical statistics
  - computer science
  - domain expertise
- The magical Unicorn: having all three skills
  - In 2014, these jobs go unfilled for 6 months or longer on average
- Has lead to the development of data science *teams*
  - hope is to merge skills of analysts

Data Science Encompasses Many Roles



Data Scientist – Analytics

Defines and monitors metrics, creates data narratives, builds tools

Data Scientist – Algorithms

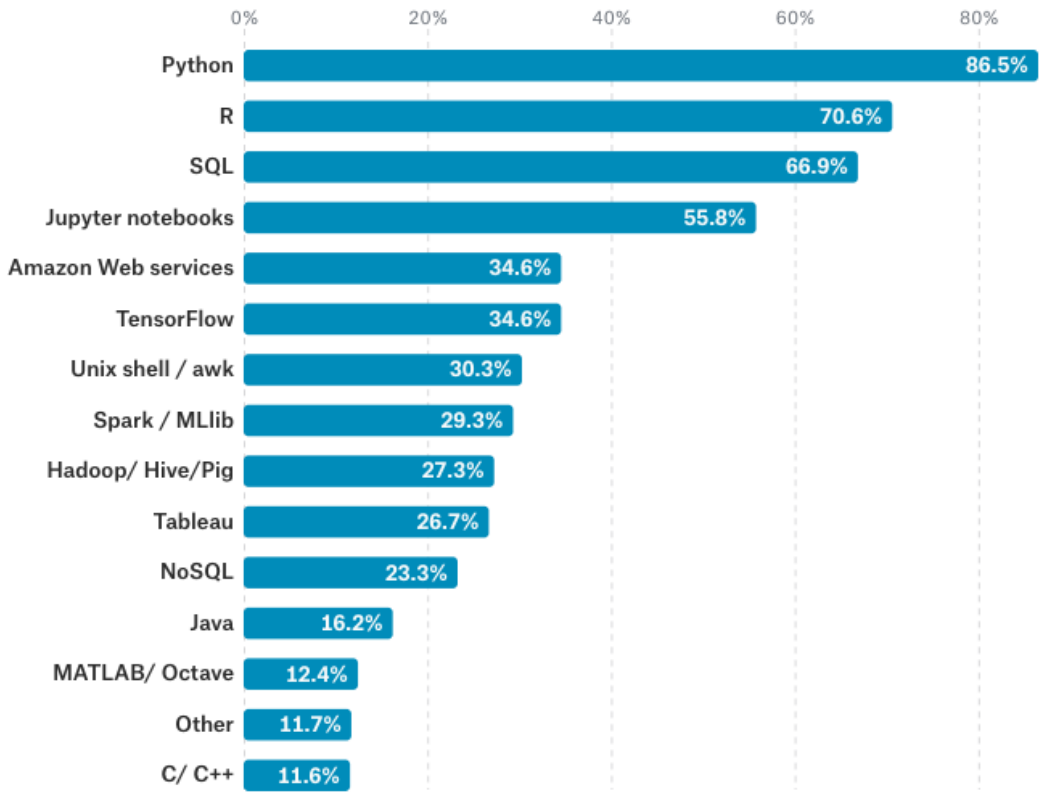
Builds and interprets algorithms that power data products

Data Scientist – Inference

Establishes causal relationships with statistics

Elena Grewal
Head of Data Science at AirBNB

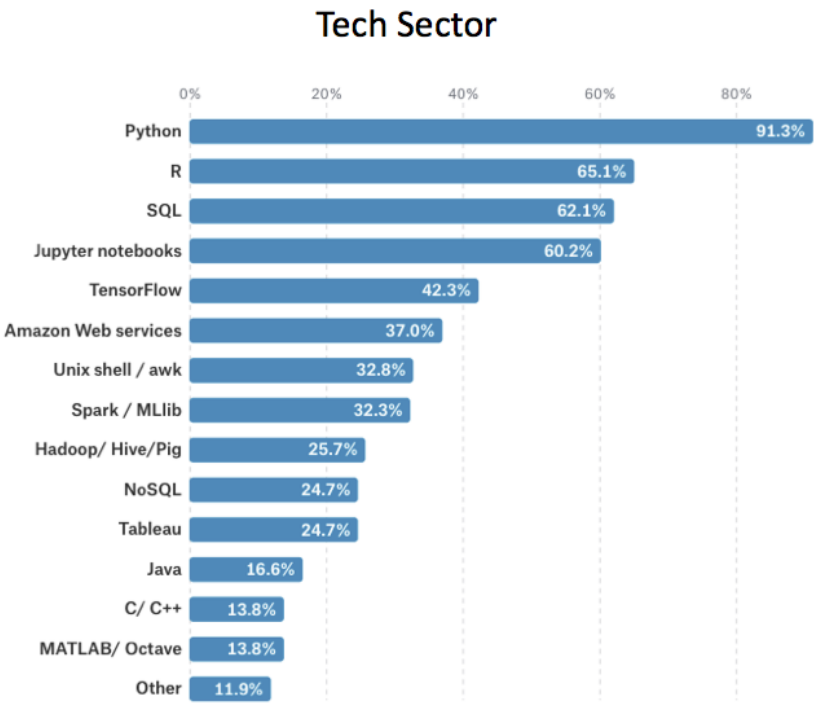
Software: R, Python, and SQL

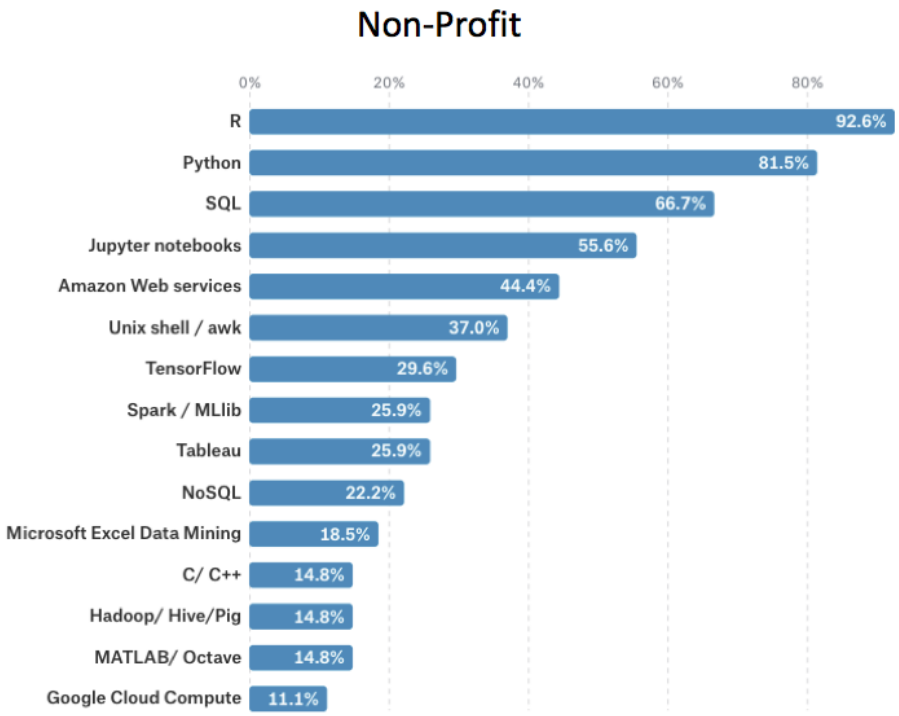


Most data scientists use a mix of Python and R

-https://www.kaggle.com/surveys/2017

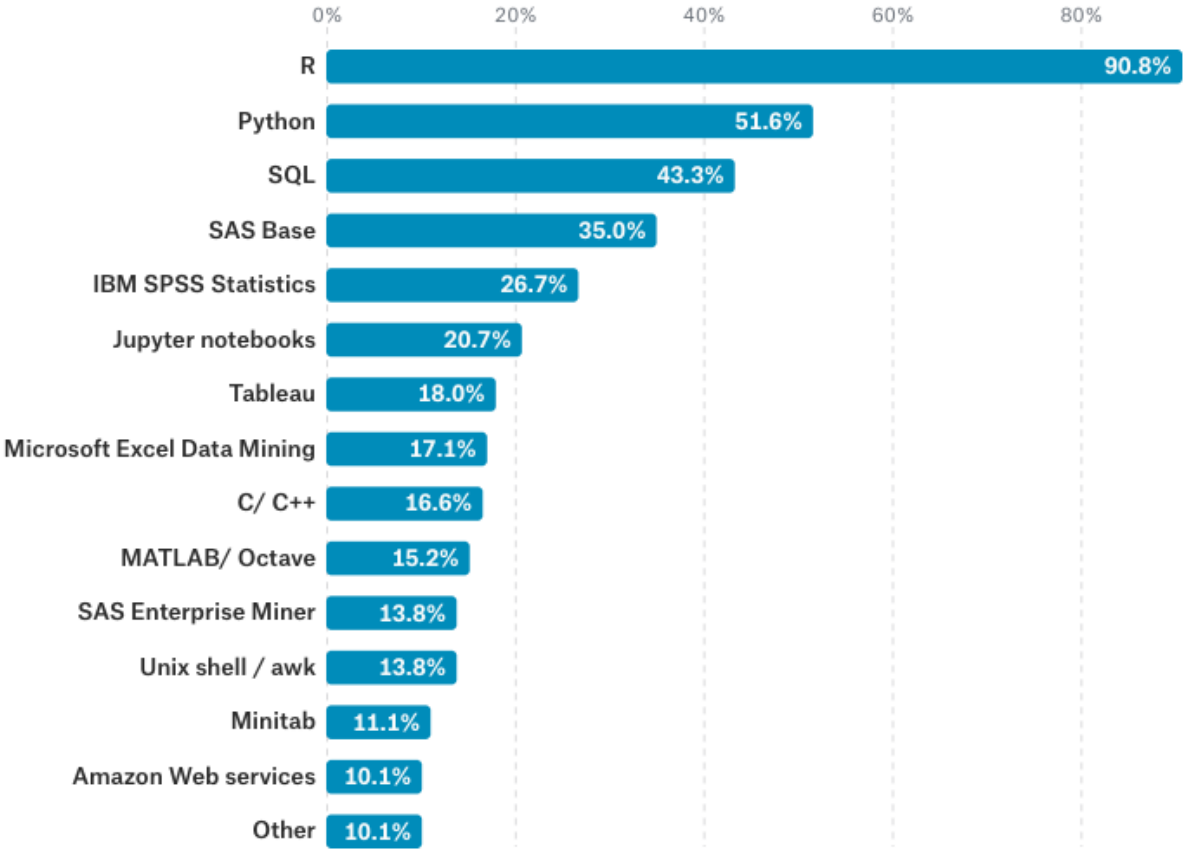
Software: R and Python





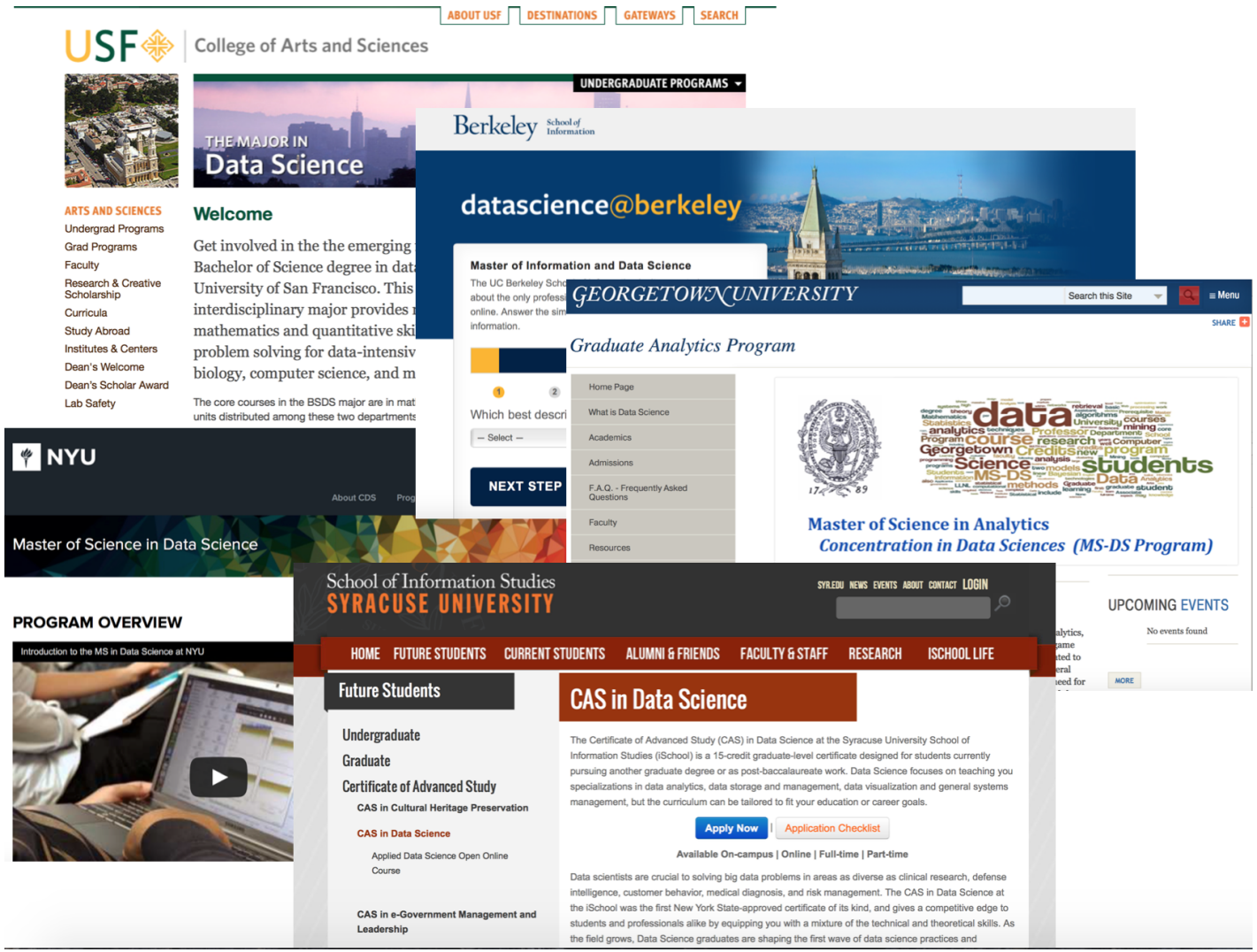
Though preference varies by field

Software: R



And R wins at statistics

Data Science in Academia

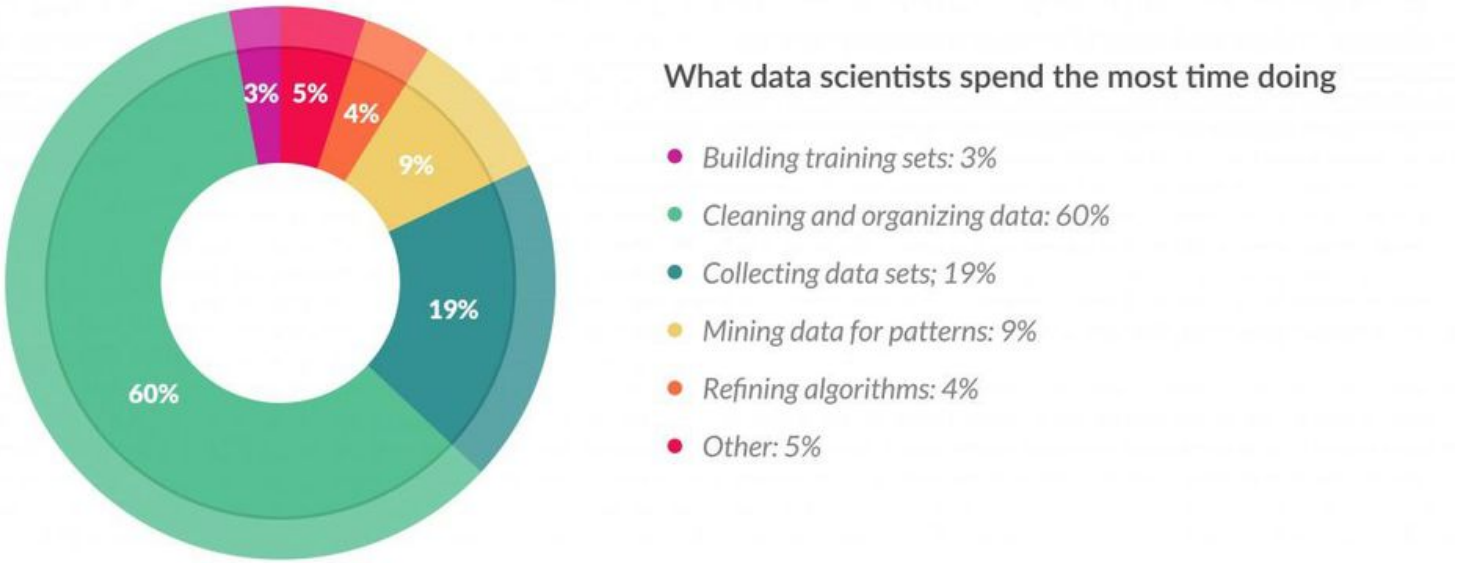


A Data Scientist's Toolkit

Harvard's data science toolkit:

1. **Wrangle the data**: gather, clean, and sample data
2. **Manage the data**: access big data quickly and reliably
3. **Explore the data**: to make a hypothesis
4. **Make predictions**: statistical methods
5. **Communicate the results**: visualization, presentations, summaries

Most Time Spent Data-Munging



From Forbes.com

A Coding Workflow

At each of these steps (wrangling, managing, exploring, etc.) we script!

- Pseudo-code
- Fill in MOST basic components
- Build up to address broader array of cases
- 80% of time will be spent debugging!

Always remember, google and stack overflow are your friends!

Great Resources

- Flowingdata.com
  - Contemporary visualization and data manipulation techniques
- dataelixir.com
  - Gathers data science stories from around the internet
- KDNuggets.com
  - Blog posts on a wide range of data science topics
- pudding.cool
  - Visual data-driven story-telling
- varianceexplained.org
  - Blog for R, statistics, and data science
- Coursera.org
  - Free online courses in data science and machine learning
  - Recent notable course: "The Data Scientist's Toolbox."

Get Involved!

- Kaggle.com
  - Kaggle competitions: win money for solving problems!
- drivendata.org
  - Competitions for non-profit or social good related problems (also often offer cash prizes)