Introduction to Data Science

Cody Carroll

BSDS 100 - Lecture 1

Outline

- Course Overview & Syllabus
- What is Data Science?

Course Overview & Syllabus

Nice to meet ya!

- Texas → Japan → CA
- Ph.D. in Statistics from UC Davis
  - Statistical research expertise: statistical analysis of curves
  - Other applied DS interests: CA wildlife, conservation, computer vision & deep learning

About Me

Before becoming a professor I was:

- a ESL Teacher
- a Data Scientist
- a DJ

About Me

I teach jointly across Math and MSDS:

- BSDS 100 - Intro to Data Science with R
- MATH 372 - Linear Regression
- MATH 371 - Statistics with Applications
- MSDS 610 - Communication for Analytics
- MSDS 630 - Advanced Machine Learning
- MSDS 699 - Machine Learning Lab

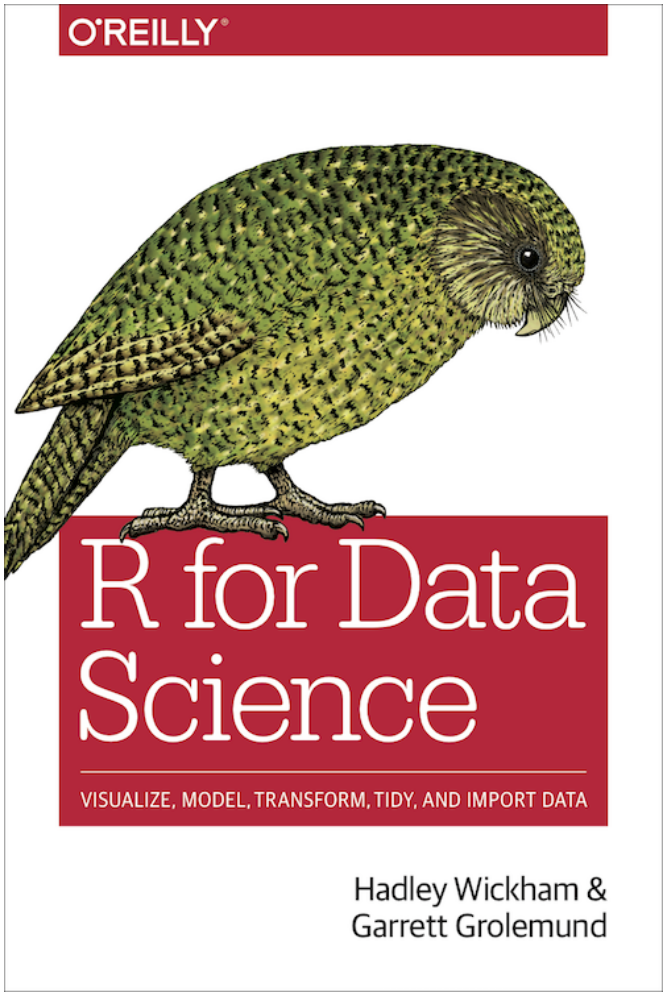
Course Description and Syllabus

The syllabus and course description can be found here:

https://github.com/codycarroll/Intro-DS-F23

All lecture notes and assignments will be posted here over the semester.

Main Text



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

Other Resources

The Best Place for Answers to R Questions?

Part II: What is Data Science?

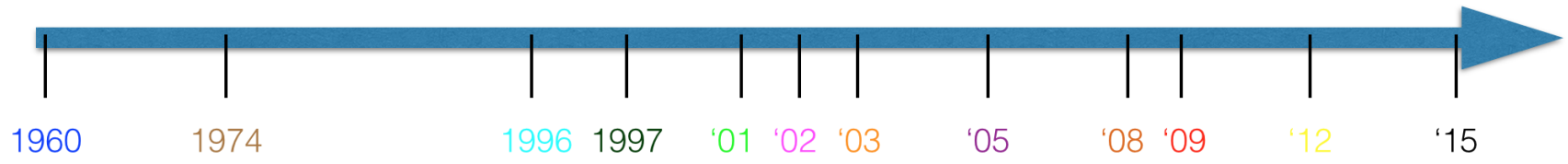
What is Data Science?

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

Where is Data Science?

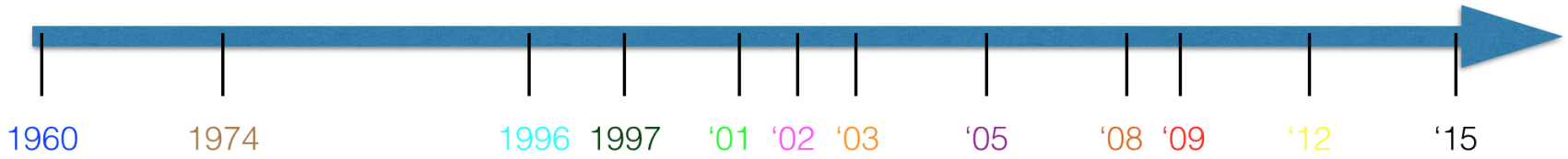
- Twitter

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."

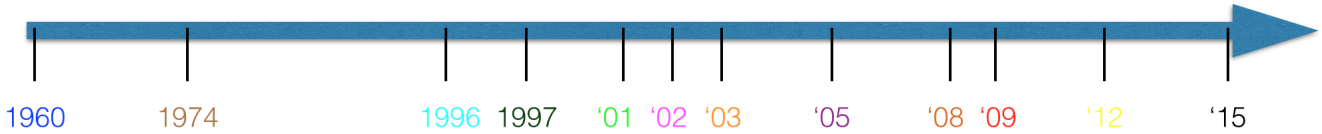
The Evolution of Data Science



- 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."

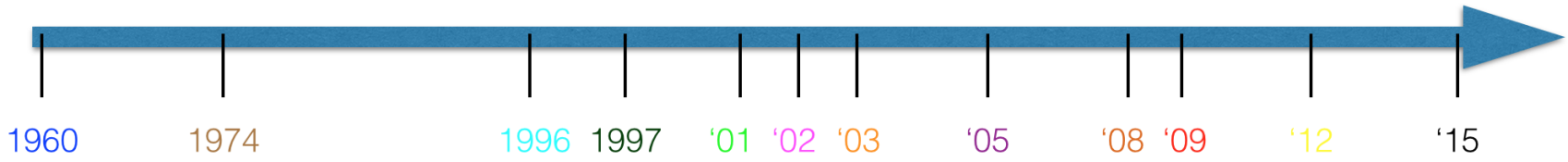
- 1997: C.F. Jeff Wu gave the inaugural lecture “Statistics = Data Science?" for appointment to the H. C. Carver Professorship at the University of Michigan.

The Evolution of Data Science



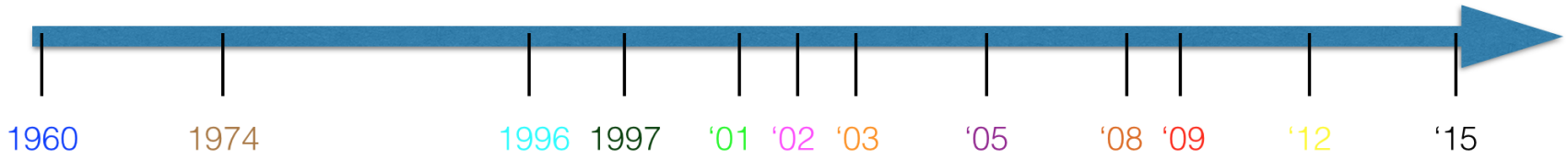
- 2001: William Cleveland (Bell Labs) published *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*.
  - "Faculty members should devote their careers to advances in computing with data and who form partnerships with computer scientists."
- 2002: *Data Science Journal* 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.

The Evolution of Data Science

- 2021: DS jobs starting to shift toward Data Engineering tasks
- 2022: A big correction in tech industry - DS market tightens
- 2023+ : Recovery?

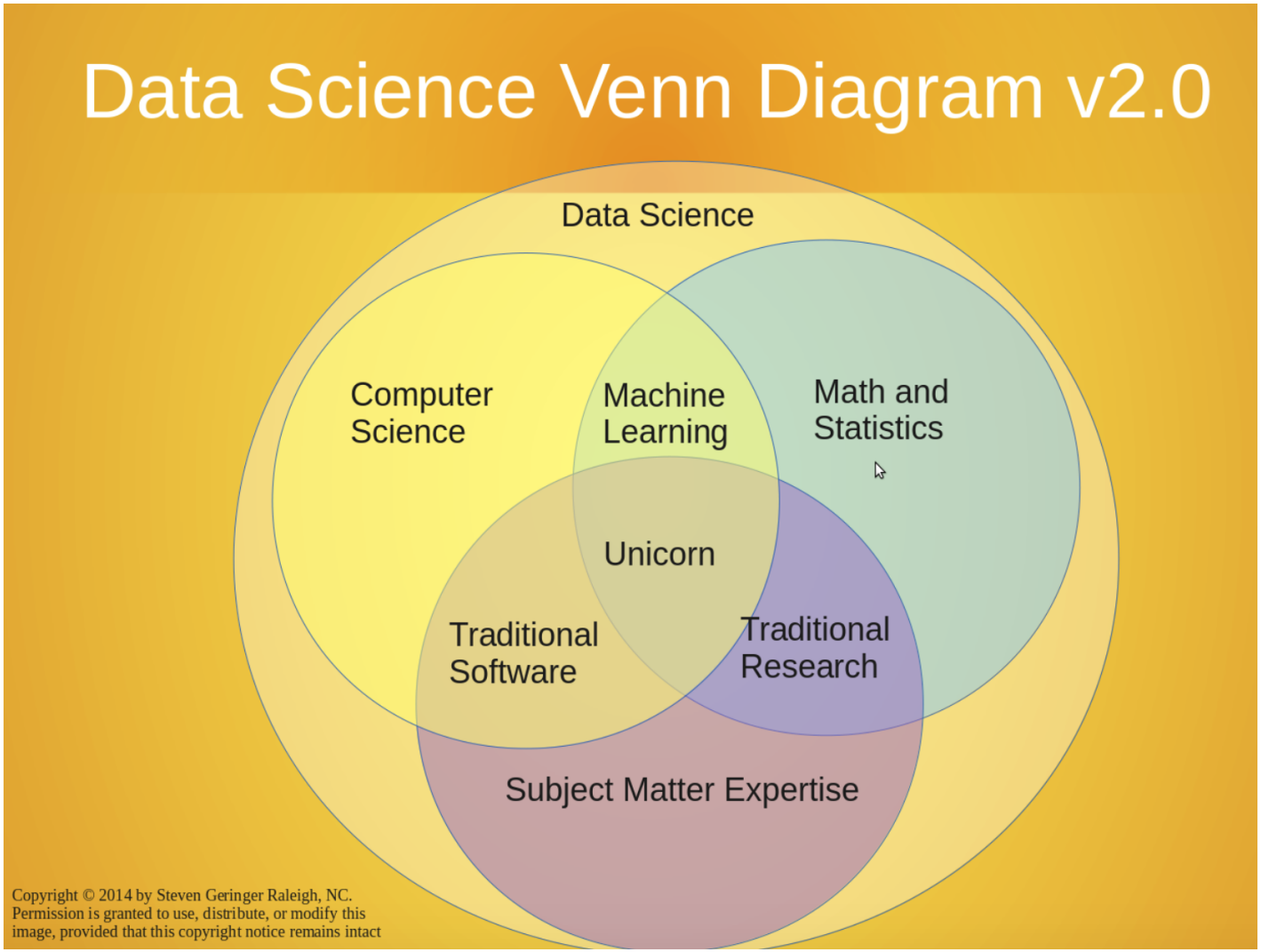
Applications

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

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)

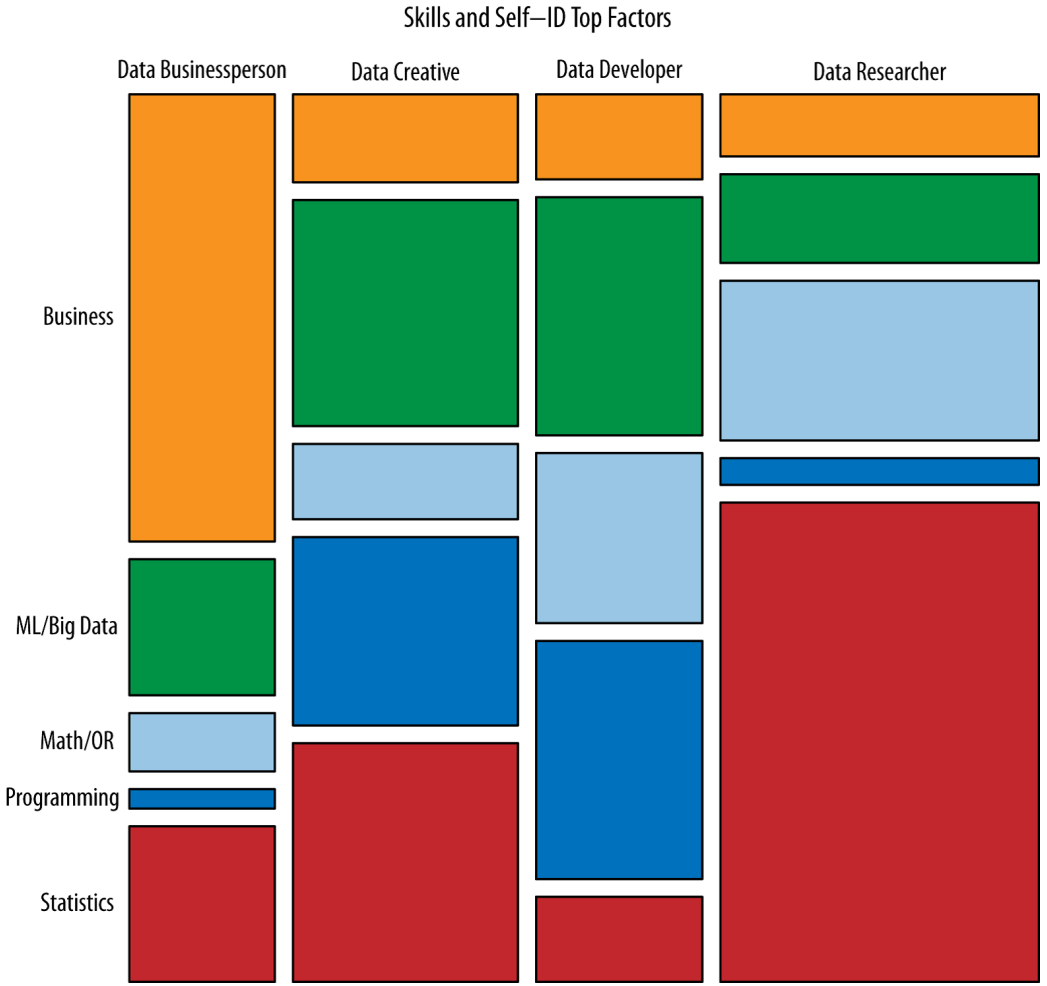
Data Scientists: The unicorn industries want?



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

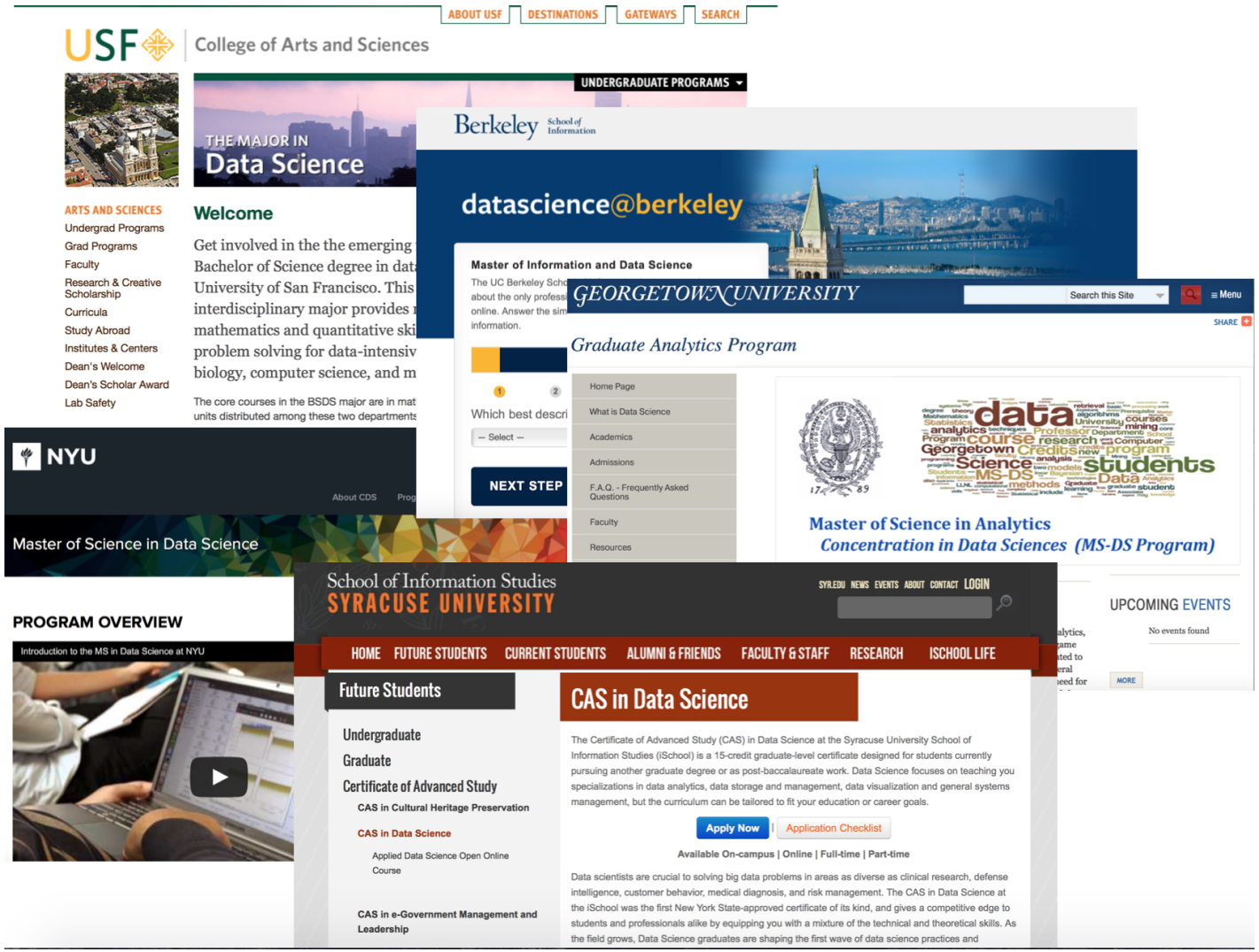
The Analysts of Data Science



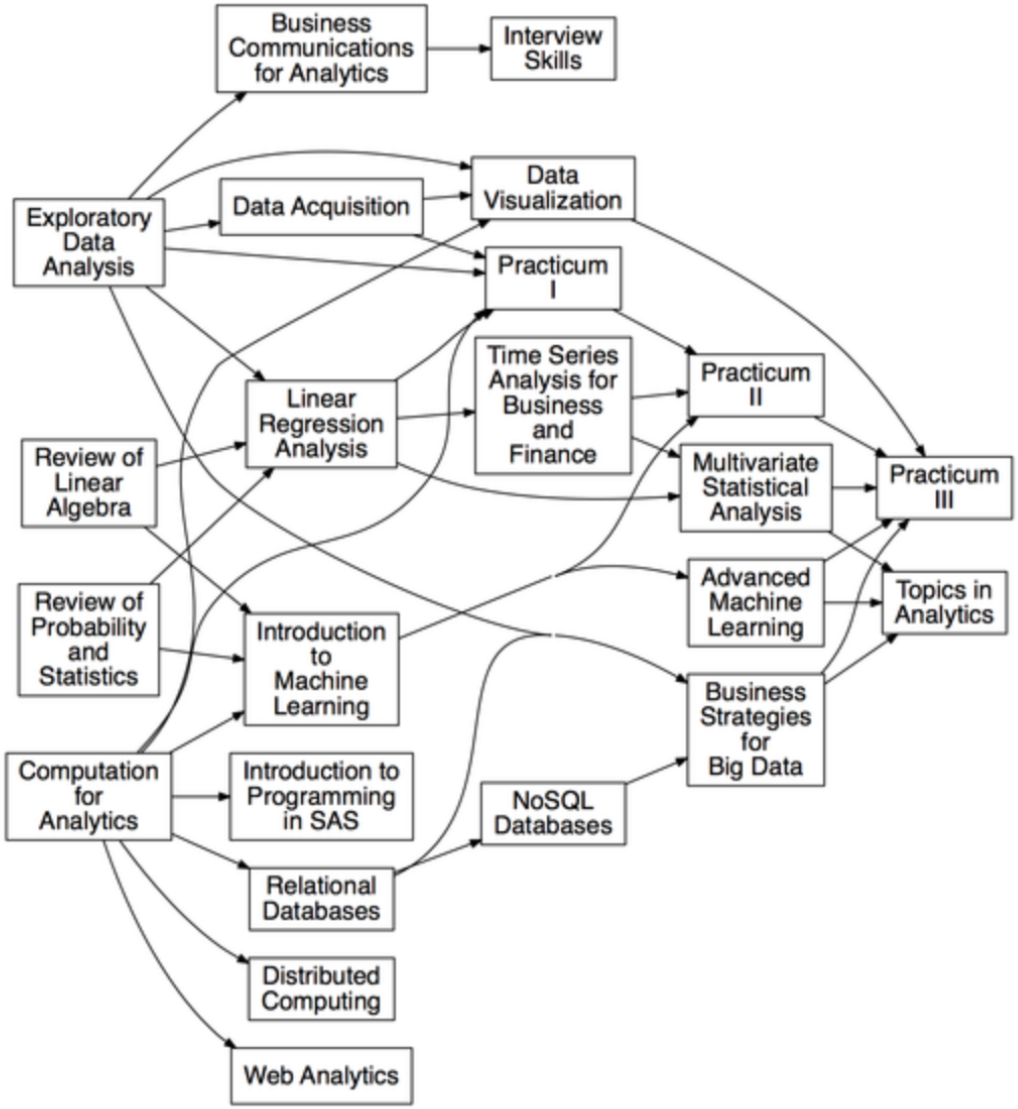
"Analyzing the Analyzers (2013) by Harry, Murphy, and Vaisman."

Software: A Data Scientist's first weapon

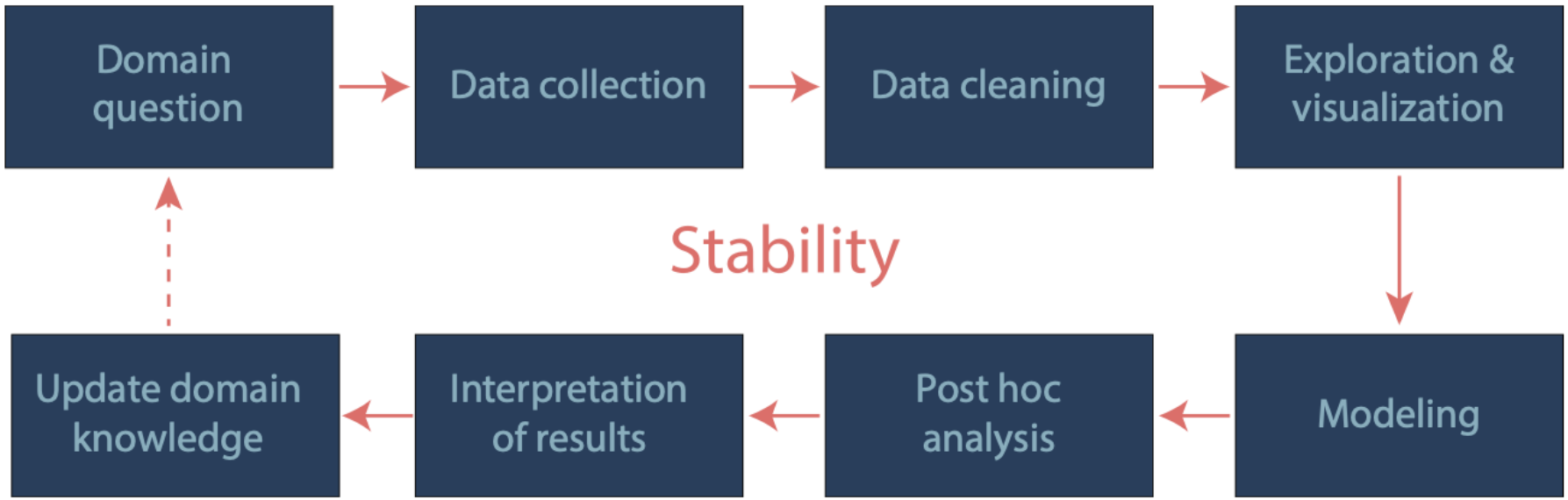
Data Science in Academia



Academic Programs



The Data Science Life Cycle



Yu and Kumbier 2020

A Data Scientist's Toolkit

Data science toolkit:

1. **Data Wrangling**: gather, clean, and sample data
2. **Database Management**: access big data quickly and reliably
3. **Data Exploration**: summarize, visualize, and contextualize data to make a hypothesis
4. **Make predictions**: statistical and machine learning methods
5. **Communicate the results**: more visualization, presentations, summaries

Get Involved! Great Resources

- Flowingdata.com
  - Contemporary visualization and data manipulation techniques
- Kaggle.com
  - Kaggle competitions: win money for solving problems!
- Coursera.org
  - Free online courses in data science and machine learning
  - 972 courses. Great resource for coding, data analysis, etc.
  - Recent notable course: "The Data Scientist's Toolbox."