

# Introduction to Data Science

Cody Carroll

BSDS 100 - Lecture 1



# Outline

- Course Overview & Syllabus
- What is Data Science?

# **Course Overview & Syllabus**



# 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.



# 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 the instructor

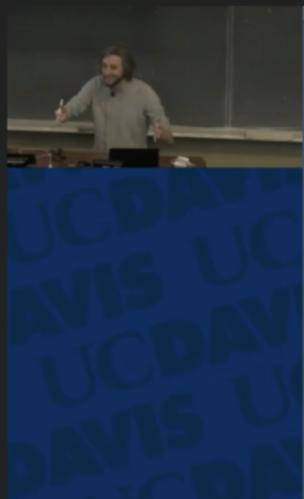
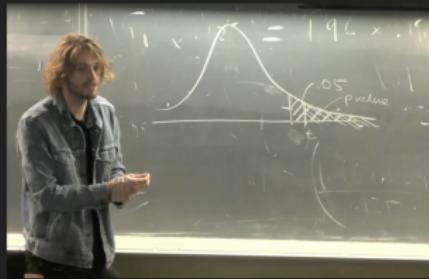
Before coming to USF:





# About the instructor

Before coming to USF:



```
lec2.R | RGuideWeek1&2.R | lec5.R* | Go to file/function | Addins |
survey = read.csv("C:/Users/.../Desktop/STA100/datasets/classsurvey.csv")
str(survey)
ggplot(survey, aes(x=height, y=shoe size)) +
  geom_point()
surveyline = lm(shoe size ~ height, data=survey)
summary(surveyline)
#Predict a new value:
#What shoe size would we expect someone who is 180cm (6') to have
#by hand:
33:1 (Top Level) | R Script |
Console | Terminal | ~Desktop/STA100/Datasets/ |
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.340708 1.627774 -9.424 <2e-16 ***
height 0.140610 0.009786 14.486 <2e-16 ***
---
Signif. codes: 0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 ' ' 1
Residual standard error: 1.15 on 156 degrees of freedom
Multiple R-squared: 0.5736, Adjusted R-squared: 0.5709
F-statistic: 289.9 on 1 and 156 DF, p-value: < 2.2e-16
> |
```



# About the instructor

## iNaturalist



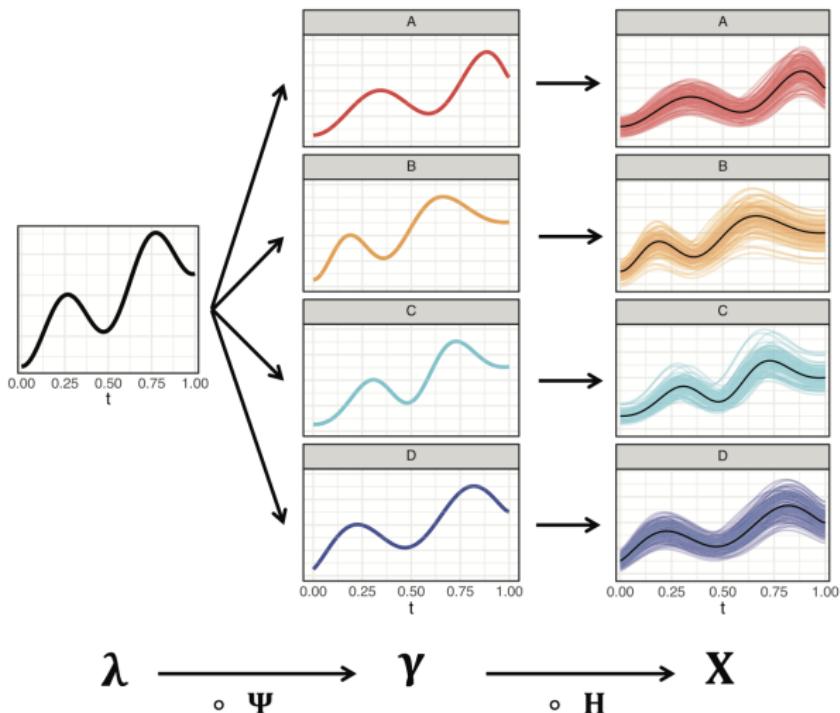
The screenshots illustrate the iNaturalist workflow:

- Screenshot 1:** A close-up photograph of a Lesser Two-spot Octopus resting on a rocky substrate. The photo is timestamped "2/27/22, 3:16 PM".
- Screenshot 2:** The iNaturalist app interface showing the identification results. The species is identified as "Lesser Two-spot Octopus" (*Octopus bimaculoides*). The location is "Mushroom House" in "North Pacific Ocean, CA, US". A map shows the observation point near San Francisco. The "DATA QUALITY" section indicates "Casual Grade" and "Needs ID".
- Screenshot 3:** A detailed view of the species page for "Lesser Two-spot Octopus". It includes a larger image of the octopus, a description: "The California two-spot octopus (*Octopus bimaculoides*), often simply called a "bimac", is an octopus species native to many parts of the Pacific Ocean including the coast of California. One can identify the species by the circular blue eyespots on each side of its head. Due to their friendly temperament and relative hardness, most experts consider them the best pet octopus. Bimacs usually live to be about two years old. They are closely related to Verriell's two-spot... (Source: Wikipedia, California\_two-spot\_octopus, CC BY-SA 3.0)", and a "MAP OF OBSERVATIONS" showing collection points along the West Coast of North America.



# About the instructor

Functional data & time warping:

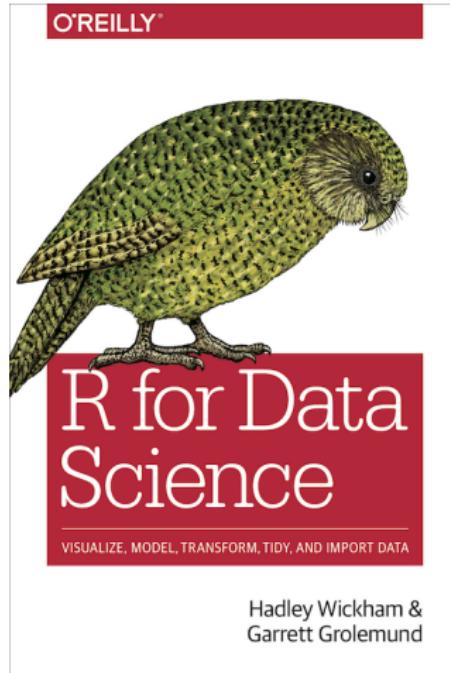




# About the Instructor

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



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



# Other Resources



# The Best Place for Answers to R Questions?





## Friends!

- Find someone to pair up with.
- Self-introduce and trade emails/contact info.
- Major? Year? Interests? Something memorable?
- What makes you curious about data science?
- How would you explain what “data science” is?  
(your perception, not an “official” definition)

## 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...
- Practitioners tend to agree on the *components* of data science:
  - gathering and cleaning data
  - database management
  - exploratory analysis
  - predictive modeling
  - data summary and visualization

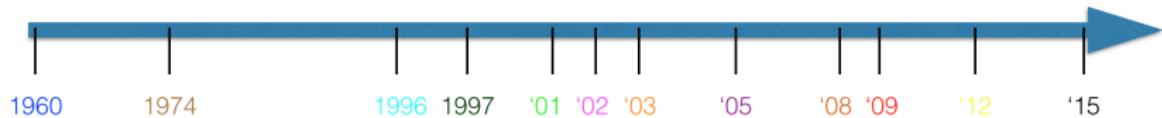


# Where is Data Science?



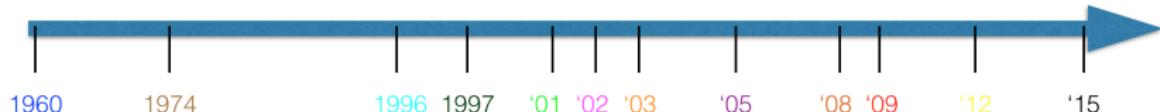
- Twitter

# The Evolution of Data Science



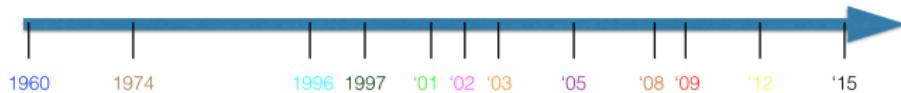
- 1960: Peter Naur (CS Ph.D.) published *Datalogy: the science of data and its place in education*.
- 1974: Peter Naur published *Concise Survey of Computer Methods*.
  - defines data science as “the science of dealing with data, once they have been established.”
  - continues to say that “... the relation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

# The Evolution of Data Science



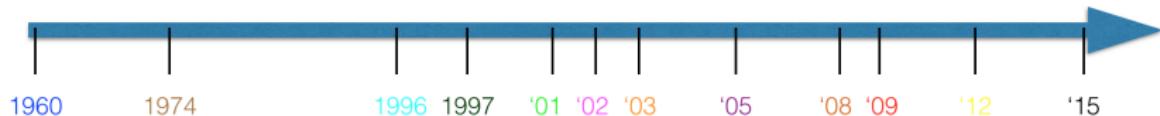
- **1996:** International Federation of Classification Societies meet in Tokyo and for the first time include “data science” in the conference title: “Data science, classification, and related methods.”
- **1997:** C.F. Jeff Wu gave the inaugural lecture “Statistics = Data Science?” for appointment to the H. C. Carver Professorship at the University of Michigan.

# The Evolution of Data Science



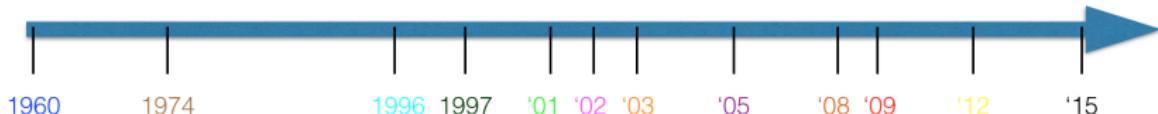
- **2001:** William Cleveland (Bell Labs) published *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics.*
  - "Faculty members should devote their careers to advances in computing with data and who form partnerships with computer scientists."
- **2002:** *Data Science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



# Recent trends

- 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
  - Over 80 trillion gb of data generated in 2022; up from 130 billion in 2005!
  - **Big data** requires innovative techniques for analysis



# Why Data Science?

## DATA SCIENCE US

### SALARY AND DAY RATE BREAKDOWNS

#### PERMANENT – AVERAGE ANNUAL SALARY

Role Type	Entry Level	Mid-Level	Principal/Manager Level	Technical Lead/Director	VP and above
Annual Base Salary - \$USD   East • West • Mid					
Data Science – ML Modelling	\$132k   \$128k   \$100k	\$160k   \$161k   \$118k	\$196k   \$201k   \$168k	\$253k   \$243k   \$213k	\$360k   \$260k   \$240k
Deep Learning & AI	\$132k   \$110k   n/a	\$173k   \$173k   \$119k	\$208k   \$228k   \$160k	\$252k   \$249k   \$202k	\$321k   \$269k   \$250k
ML Engineering – Deployment & Infrastructure	\$138k   \$117k   \$120k	\$180k   \$156k   \$140k	\$228k   \$201k   \$198k	\$287k   \$252k   \$244k	\$340k   \$283k   \$283k
MLOps	\$133k   \$108k   \$120k	\$189k   \$156k   \$140k	\$232k   \$198k   \$178k	n/a   \$250k   \$250k	n/a   \$269k   n/a
Natural Language Processing	\$138k   \$115k   n/a	\$176k   \$155k   \$155k	\$215k   \$195k   \$195k	\$250k   \$225k   \$225k	n/a   \$262k   n/a

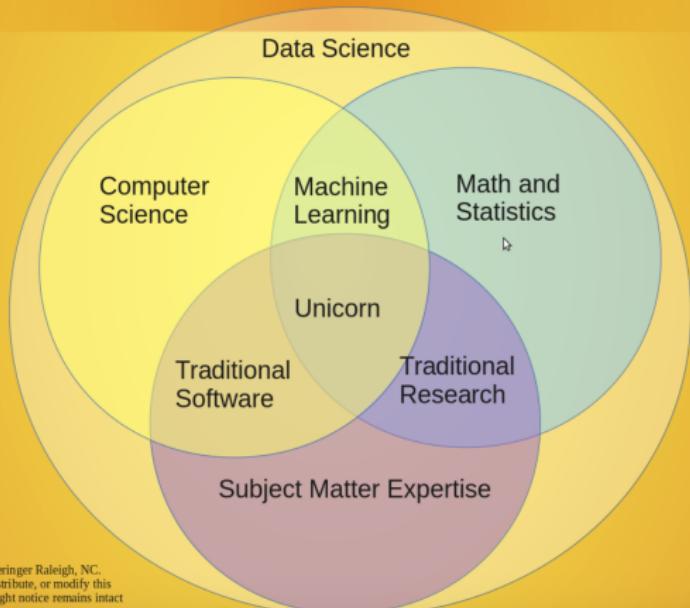
#### CONTRACT – AVERAGE DAY RATES

Role Type	Entry Level	Mid-Level	Principal/Manager Level	Technical Lead/Director	VP and above
Hourly Contract Rate - \$USD   East • West					
Data Science – ML Modelling	\$65   \$65	\$80   \$75	\$95   \$105	\$120   n/a	\$150   n/a
Deep Learning & AI	\$65   \$55	\$85   \$85	\$100   \$115	\$120   n/a	\$150   n/a
ML Engineering – Deployment & Infrastructure	\$70   n/a	\$90   \$80	\$100   \$100	\$130   n/a	\$150   n/a
MLOps	\$65   n/a	\$90   \$85	\$100   \$100	n/a   n/a	n/a   n/a
Natural Language Processing	\$70   n/a	\$85   \$75	\$100   \$100	\$120   n/a	n/a   n/a

*Harnham's US 2023 Data & AI Salary Guide*



## Data Science Venn Diagram v2.0



Copyright © 2014 by Steven Geringer Raleigh, NC.  
Permission is granted to use, distribute, or modify this  
image, provided that this copyright notice remains intact.

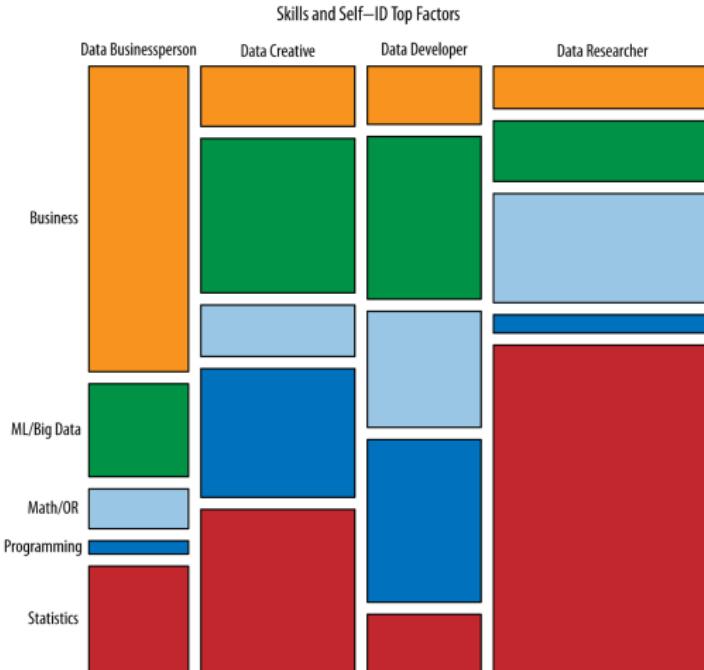
# Data Scientists: The unicorn industries want?



- The field is inherently interdisciplinary
  - mathematical statistics
  - computer science
  - domain expertise
- The elusive “unicorn”: 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."*



# Data Science in Academia

ABOUT USF DESTINATIONS GATEWAYS SEARCH

**USF College of Arts and Sciences**

**ARTS AND SCIENCES**  
Undergrad Programs  
Grad Programs  
Faculty  
Research & Creative Scholarship  
Curricula  
Study Abroad  
Institutes & Centers  
Dean's Welcome  
Dean's Scholar Award  
Lab Safety

**THE MAJOR IN Data Science**

**Welcome**  
Get involved in the emerging Bachelor of Science degree in data University of San Francisco. This interdisciplinary major provides mathematics and quantitative skills problem solving for data-intensive biology, computer science, and m

The core courses in the BBDS major are in mat units distributed among these two departments

**NYU**

**Master of Science in Data Science**

**PROGRAM OVERVIEW**  
Introduction to the MS in Data Science at NYU

**SCHOOL OF INFORMATION STUDIES SYRACUSE UNIVERSITY**

**Future Students**  
Undergraduate  
Graduate  
**Certificate of Advanced Study**  
CAS in Cultural Heritage Preservation  
CAS in Data Science  
Applied Data Science Open Online Course  
CAS in e-Government Management and Leadership

**BERKELEY School of Information**

**UNDERGRADUATE PROGRAMS**

**datascience@berkeley**

**Master of Information and Data Science**  
The UC Berkeley School of Information is about the only program in the country that offers a Master of Information and Data Science online. Answer the questions below to learn more about our program.

**GEORGETOWN UNIVERSITY**

**Graduate Analytics Program**

**Home Page**  
What is Data Science  
Academics  
Admissions  
FAQ – Frequently Asked Questions  
Faculty  
Resources

**data**

**Master of Science in Analytics Concentration in Data Sciences (MS-DS Program)**

**UPCOMING EVENTS**  
No events found

SEARCH THIS SITE

SHARE

APPLY NEWS EVENTS ABOUT CONTACT LOGIN

CAS in Data Science

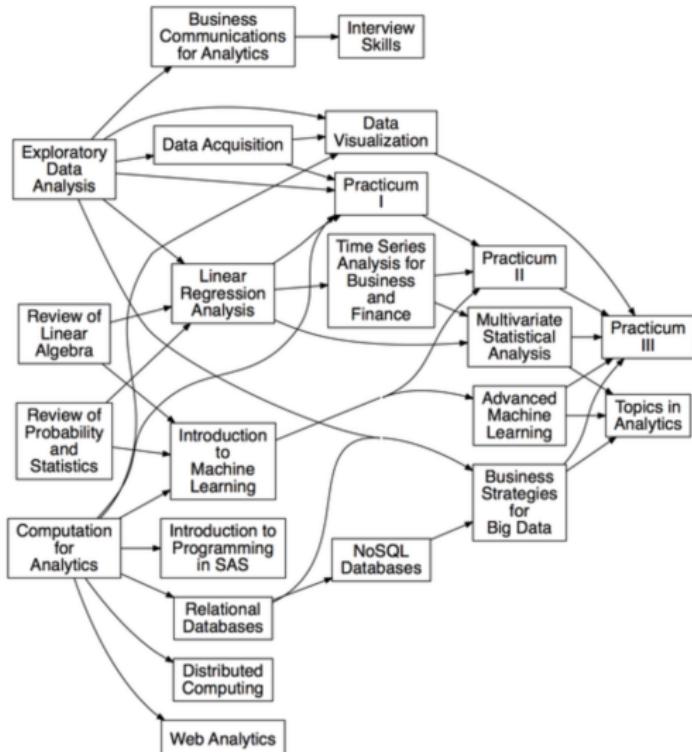
The Certificate of Advanced Study (CAS) in Data Science at the Syracuse University School of Information Studies (SISch) is a 15-credit graduate-level certificate designed for students currently pursuing another graduate degree or one planning to do so. Data Science focuses on teaching you specialized analytical, data mining and management, data visualization and general systems management, but the curriculum can be tailored to fit your education or career goals.

Apply Now Application Checklist Available On-campus | Online | Full-time | Part-time

Data scientists are crucial to solving big data problems in areas as diverse as clinical research, defense intelligence, customer behavior, medical diagnosis, and risk management. The CAS in Data Science at the SISch was the first New York State-approved certificate of its kind, and gives a competitive edge to students and professionals alike by equipping you with a mixture of the technical and theoretical skills. As the field grows, Data Science graduates are shaping the first wave of data science practices and

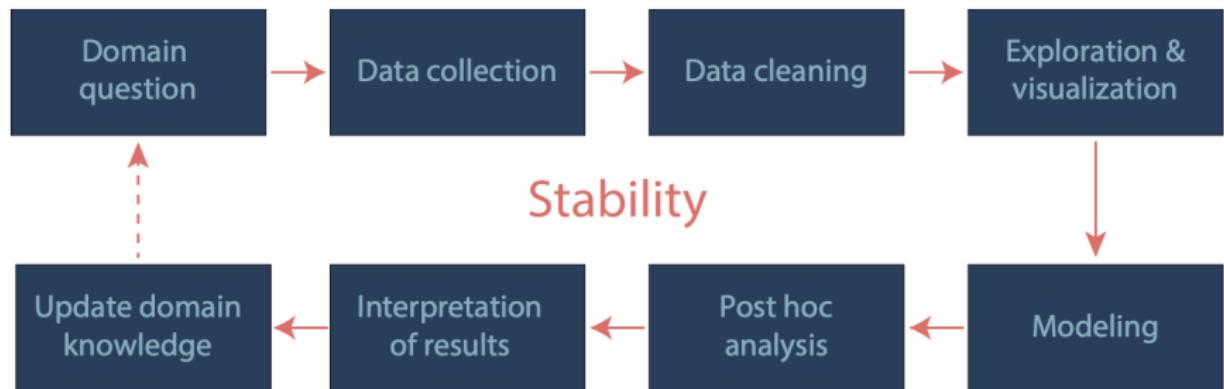


# Academic Programs





# The Data Science Life Cycle



Yu and Kumbier 2020



- ➊ **Data wrangling:** gather, clean, and sample data
- ➋ **Database management:** access big data quickly and reliably
- ➌ **Data exploration:** summarize, visualize, and contextualize data to make a hypothesis
- ➍ **Make predictions:** statistical and machine learning methods
- ➎ **Communicate the results:** more visualization, presentations, summaries



# Resources for Exploring DS

- [Flowingdata.com](#)
  - Contemporary visualization and data manipulation techniques
- [Kaggle.com](#)
  - Kaggle competitions: can 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."