First steps of a data science career

Dr. Mine Dogucu

bit.ly/dogucu-talks
mdogucu@uci.edu
@MineDogucu
MineDogucu
• How many of you are first years/sophomores/juniors/seniors?

• How many of you have taken a statistics class?

• How many of you want to be a data scientist?
Coursework
Introduction to Data Science
Course Goals:

• explore data using descriptive statistics and visualizations;
• read, write, and tidy up datasets;
• make predictions and conclusions using models;
• consider impact of decisions related to data on humans, other livings, and the planet;
• write human- and machine-readable code using R.
explore data using descriptive statistics and visualizations;
Artwork by @allison_horst
read, write, and tidy up datasets;
If you'd like to know more, check out this list of changes to the Best 100 Games of All Time list -- and why we made those changes.

Pokémon Go

RELEASED
2016

JUSTIN VACHON → Pokémon GO in 2019 is a game I shouldn't care about. When it launched in 2016 it was in a lot of ways a mediocre experience. Outside of catching the original 151 Pokémon the game itself relied heavily on the nostalgia of the Pokémon franchise and augmented reality gimmick of having them show up in the real world. If you didn't care about the IP, the game itself was very lacking. In 2019, the game is flooded with a multitude of tasks, activities, and events that can involve anyone from yourself to a large group of people. These additions create an experience that incentivizes users to be more dedicated to daily play without feeling like a grind. Friendship has been introduced and allows users to now exchange gifts, trade or even battle each other. Quests (research tasks as they are referred to in-

DID YOU KNOW?

There are currently five (of eight) generations of Pokémon in the game, with more being added each year.
make predictions and conclusions using models;
logistic regression example:

*Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*
Inference using bootstrap sampling and randomization
consider impact of decisions related to data on humans, other livings, and the planet;

Finding data online does not grant permission to use

Human subjects  Bot access  Commercial use
write human- and machine-readable code using R.
The Museum of Modern Art (MoMA) Collection

The Museum of Modern Art (MoMA) acquired its first artworks in 1929, the year it was established. Today, the Museum’s evolving collection contains almost 200,000 works from around the world spanning the last 150 years. The collection includes an ever-expanding range of visual expression, including painting, sculpture, printmaking, drawing, photography, architecture, design, film, and media and performance art.

MoMA is committed to helping everyone understand, enjoy, and use our collection. The Museum’s [website](https://www.moma.org)
rmarkdown

The rmarkdown package helps you create dynamic analysis documents that combine code, rendered output (such as figures), and prose. You bring your data, code, and ideas, and R Markdown renders your content into a polished document that can be used to:

- Do data science interactively within the RStudio IDE,
- Reproduce your analyses,
- Collaborate and share code with others, and
- Communicate your results with others.

R Markdown documents can be rendered to many output formats including HTML documents, PDFs, Word files, slideshows, and more, allowing you to focus on the content while R Markdown takes care of your presentation.

Books

R Markdown
The Definitive Guide

R Markdown Cookbook
Final Projects

Windmills in California Over Time

Windmills Constructed In California Each Year

Turbine Distribution in California

4078 turbines in California.
What’s next?

Exploratory Data Analysis -> Data Visualization
Datasets -> Databases
Communication -> Writing + Speaking
Statistical Modeling -> Probability + Statistics
Ethics -> Philosophy
On Campus Resources
- Attend webinars/seminars related to data science.
- Utilize services offered by career office.
- Use writing center.
Networking
Welcome to CampusGroups!
The official campus community platform at UCI
Consider Using Social Media for Professional Reasons

- Twitter
- LinkedIn

Resource:
Twitter for R programmers
Portfolio
TidyTuesday
A weekly data project in R from the R4DS online learning community

variables observations values

https://github.com/rfordatascience/tidytuesday
Create a website

- distill
- blogdown
- Wordpress

Deploy your website

- GitHub pages
- Netlify
Contests
Undergraduate Statistics Class Project Competition

Introductory Statistics
Projects from courses with no statistics/data science prerequisite course
Submission: A short report (up to 3 pages)

Intermediate Statistics
Projects from non-intro and non-capstone courses
Submission: A short report (up to 3 pages)

Undergraduate Statistics Research Project Competition

Research projects from activities like summer research or a capstone course
Submission: A report (up to 20 pages)

Next Deadline: Dec 22nd

Details:
www.causeweb.org/usproc
CAUSEweb Cartoon Caption Contest
March Randomness

MARCH RANDOMNESS
A contest of probability, intuition, and random chance
Career Preparation Resources
Electronic Undergraduate Statistics Research Conference

FRIDAY, NOVEMBER 5TH, 2021

**Watch**
Undergraduate Statistics Project Competition (USPROC) award winners will present their work.

**Learn**
Info session on graduate school and panel on careers in statistics & data science.

**Be Inspired**
Keynote address by Sydeaka Watson, Senior Research Scientist (Data Scientist) at Eli Lilly and Company.

**Get Involved**
To share your own statistical work and register for the conference, submit an abstract by Friday, October 15th. For more information, go here! Prizes will be given for the Best Video Presentations!

**Register**
To attend this FREE conference, click here. All undergraduate students and faculty are invited.
Build a Career in Data Science

Jacqueline Nolis and Emily Robinson

FEB 11, 2021

Chapter 11: Data Science in Production

Putting data science into production can mean a ton of things: from customer-facing models run millions of times a day to continuously live dashboards for stakeholders. But writing code for production and getting it to work can be intimidating for many data scientists, and lots of us have...
NSF Research Experiences for Undergraduates (REU) search
ASA Internships and Fellowships
Find your next job in data science in the Federal Government

https://usajobs.github.io/microsite-data-science
Keywords:
Statistics
Data
Biostatistics
Bioinformatics
Psychometrics
Econometrics
A First-Gen's Guide to Grad School: How to Get in, Survive, and Thrive

Helping first-gen students along their PhD journey!

https://first-gen-guide.com/resources
Applying to the NSF Graduate Research Fellowship (GRFP)

https://stattrak.amstat.org/2021/07/01/grad-school-in-biostatistics/
Southern California Data Science

New courses: Introduction to Data Science, Introduction to Bayesian Data Analysis, Machine Learning

For fellows: seminar, bootcamp, summer research project

HDR DSC awards:
#2123366 #2123380 #2123384
Last but not least
You might hear things like

- You are not a data scientist because you are not strong in math.
- What you do is not data science because it does not involve advanced computing.
- You cannot be a data scientist because you studied X and Y but not Z.
THIS IS
THE SIGN
YOU’VE BEEN
LOOKING FOR
Questions?

mdogucu@uci.edu
@MineDogucu
MineDogucu