Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Halıcıoğlu Data Science Institute
Neurosciences Graduate Program

bvoytek@ucsd.edu
@bradleyvoytek

COGS 108
Data Science in Practice

Hello World!

UC San Diego

Scheduling

Voytek's Office Hours:

Thursdays 11:00a-12:00p *and*

by appointment

General stuff

- Data Science majors?

General stuff

- Data Science majors?
- This isn't high school, you're paying for this.

General stuff

- Data Science majors?
- This isn't high school, you're paying for this.
- As an adult, what you do with your time is up to you.

General stuff

- If something in lecture or a reading is unclear:
  - *ask in class*
  - *ask during section*
  - *email your TAs and IAs* or *cogs108@gmail.com*
  - *post on Piazza*
  - *come to office hours!*

General stuff

- If you email a question, the answer to which is on the syllabus… well, I'm sorry—and I'm sure you're amazing—but ***you will not get an email response from us***.

Course links

- Course email: cogs108@gmail.com
- GitHub: https://github.com/COGS108/
- Piazza: http://piazza.com/ucsd/fall2019/cogs108
- DataHub: https://datahub.ucsd.edu/

GRADING

- Five assignments (10% each)
- Final Project Proposal (10%)
- Final Project (35%)
- Guest Lecture Attendence (5%)

Participation

- Attendance *required* for guest lectures

Data Science at UCSD

Data Science at UCSD

Facebook pioneer donates $75 million to UCSD for data science

Taner Halicioglu is donating $75 million to UC San Diego to make his alma mater a national leader in data science. (Erik Jepsen / UC San Diego)

Past guest lecturers

- Franziska Bell, PhD: Senior Data Science Manager, *Uber*
- Eli Bressert, PhD: Manager, Data Engineering & Analytics, *Netflix*
- Claire Dorman, PhD: Data Scientist, *Pandora*
- Mina Doroud, PhD: Data Scientist, *Twitter* (Senior Data Scientist, *LinkedIn*)
- Carlos Gomez-Uribe, PhD: Director, Core Data Science, *Facebook* (Statistician, *Google*; VP Product Innovation, *Netflix*)
- Hiroki Hiyama, PhD: Senior Data Scientist, *Uber*
- Emi Nomura, PhD: Senior Manager, Data Science, *Pandora*
- Kevin Novak: Head of Data Science, *Uber*
- DJ Patil, PhD: US Chief Data Scientist, *Obama White House* (Head of Data Science, *LinkedIn*; co-coined the term "data science")
- Maksim Pecherskiy: Chief Data Officer, *City of San Diego*
- Sarah Rich, PhD: Data Scientist, *Twitter*
- John Myles White, PhD: Research Scientist, *Facebook* (Author: *Machine Learning for Hackers*)
- Josh Wills, PhD: Head of Data Engineering, *Slack*

Past guest lecturers

Final Project!

- Project due on Final Exam day *in lieu of* an actual Exam!
- Deadline: 23:59, Thu, December 12, 2019

Final Project examples: SD potholes

Out[5]:

We took the longitude and latitude information for each pothole and plotted it using a software called ArcGIS in order to make visualization easier.

As you can see from the open potholes heat map, the most potholes are concentrated in the center (Mira Mesa). However, insead of allocating resources to focus on the area with the most potholes, the repairs seem to neglect Mira Mesa in the closed potholes heat map. This trend can possibly be attributed to the pothole repair schedule that the workers follow. The schedule consists of a set rotational cycle that runs through each San Diego Council District to the next on different days (SanDiego.gov). Although this system gives each Council District an equal opportunity for pothole repair, this system fails to address the most problematic areas for potholes.

Final Project examples: Car accident causes

Data Analysis and Results

Here, the **average injury count is the highest in daylight.** As with weather conditions, this may be just due to there being more drivers or more people in the cars during this time. For the difference in Dark with street lights and Dark without street lights, it might be that less cars travel on dark roads without light. **This being so, this doesn't tell us much yet.**

Final Project examples: Infrastructure

WalkScore

Final Project examples: UCSD grade optimizer

Out[37]:

	course	0	1	2	3	4	5	6	7	8	...	990	991	992	993	994	995	996	997	998	999
0	MATH+20A	2.38	3.21	2.96	2.93	2.62	3.05	2.58	3.02	2.77	...	2.53	3.22	2.91	2.60	2.88	2.62	2.38	2.76	2.80	2.69
1	MATH+20B	2.89	2.21	3.04	3.22	3.34	3.02	2.82	2.53	3.21	...	3.05	2.53	3.39	3.19	2.95	3.04	2.86	3.39	2.77	3.35
2	MATH+20C	2.76	2.69	2.62	2.35	3.21	2.90	2.50	2.76	3.34	...	2.82	2.83	3.33	2.50	3.45	2.62	3.01	2.79	3.02	3.18
3	MATH+20F	3.22	2.63	3.03	3.02	3.31	2.70	3.22	3.17	2.50	...	2.74	2.82	3.33	2.58	2.62	3.17	2.77	2.75	2.95	2.68
4	COGS+1	3.34	2.40	3.19	2.58	3.45	2.56	2.80	3.08	2.76	...	2.66	2.53	2.53	2.71	3.47	3.33	3.52	2.79	2.63	2.90
5	COGS+14A	2.62	3.01	3.25	3.05	3.22	2.93	2.83	2.77	2.81	...	2.83	2.39	2.89	3.08	2.77	3.20	2.35	2.77	3.31	2.21
6	COGS+101A	2.50	2.77	2.90	3.22	2.74	3.19	3.19	2.76	2.95	...	3.21	3.45	3.38	2.52	2.41	2.81	2.53	2.90	2.50	2.80
7	COGS+102A	2.76	2.75	2.89	2.56	2.56	3.15	2.58	2.75	2.70	...	2.52	3.01	3.06	2.80	3.04	2.63	3.45	2.50	3.39	3.16
8	COGS+107A	2.77	3.20	2.76	3.02	2.46	2.38	2.44	2.53	3.33	...	2.80	2.58	2.33	2.87	2.96	2.26	2.83	2.68	3.05	3.02
9	CSE+7	2.90	3.04	2.74	2.58	2.90	2.52	3.39	2.41	3.47	...	3.24	2.62	2.71	3.02	3.22	2.57	2.50	2.95	2.53	2.46

10 rows × 1001 columns

Create a dataframe filled with 1000 random student GPAs from students who have taken a bad professor from 10 different courses.

Final Project examples: DEATH

Source: Owen Shen, Hasan Al-Jamaly, Maximillian Siemers, and Nicole Stone.

UC San Diego

Final Project examples: DEATH

Source: Owen Shen, Hasan Al-Jamaly, Maximillian Siemers, and Nicole Stone.

UC San Diego

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
data scientist

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
cognitive scientist

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
statistician

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
computer scientist

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
neuroscientist

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
ethicist

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
social scientist

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
economist

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
sports analyst

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
political activist

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
data journalist

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
tech guru

Why this course?

You are going to be analyzing lots of data because you're studying to be a:
CEO/small business owner

Why this course?

Because…..

Why this course?

Because…..

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Harvard Business Review

ntist: The Sexiest
e 21st Century

. Patil

Sexiest

tury

What is Data Science?

What is Data Science?

"The best minds of my generation are thinking about how to make people click ads.

"That sucks."

Source: Jeff Hammerbacher, Cofounder (Cloudera) and Assistant Professor (Icahn School of Medicine at Mount Sinai)

BIG DATA

Data science still woefully short on science

KDnuggets™ The Data Science Delusion

“data science isn't science”

Data Science (*cf* Computer Science)

"The first computer science degree program in the United States was formed at Purdue University in 1962."

Source: Wikipedia - Computer Science

Data Science (*cf* Computer Science)

"Since practical computers became available, many applications of computing have become distinct areas of study in their own rights."

Source: Wikipedia - Computer Science

What is Data Science?

Source: NYU Data Science

What is Data Science?

Source: NYU Data Science

What is Data Science?

What is Data Science?

There is much debate among scholars and practitioners about what data science is, and what it isn't. Does it deal only with big data? What constitutes big data? Is data science really that new? How is it different from statistics and analytics?

Source: NYU Data Science

What is Data Science?

What is Data Science?

There is much debate among scholars and practitioners about what data science is, and what it isn't. Does it deal only with big data? What constitutes big data? Is data science really that new? How is it different from statistics and analytics?

????!???

Source: NYU Data Science

UC San Diego

What *isn't* Data Science?

What *isn't* Data Science?

What *isn't* Data Science?

DATA SCIENCE ISN'T MACHINE LEARNING

Defining Data Science?

Trouble defining a field

“The only generally agreed upon definition of mathematics is "Mathematics is what mathematicians do." which is followed by "Mathematicians are people who do mathematics." What is true about defining mathematics is also true about many other fields: there is often no clear, sharp definition of the field.”

Source: Richard Hamming

Data Science (*cf* Computer Science)

"[Computer science] is not really about computers—and it's not about computers in the same sense that physics is not really about particle accelerators, and biology is not about microscopes and Petri dishes…"

Source: Hal Abelson (MIT)

Data Science (*cf* Computer Science)

"…when some field is just getting started and you don't really understand it very well, *it's very easy to confuse the* ***essence*** *of what you're doing with the tools that you use.*"

Source: Hal Abelson (MIT)

What is the *essence* of what we do?

What is our essence?

"Good data scientists understand, in a deep way, that the heavy lifting of cleanup and preparation isn't something that gets in the way of solving the problem: it is the problem."

Source: DJ Patil, Former head of Data Science (LinkedIn) and US Chief Data Scientist under Obama

My (heavily biased) Data Science "essence"

The study of how the quantification of observable phenomena can lead to human understanding of the processes giving rise to those phenomena—or even the ability to predict future outcomes absent human understanding—and why certain phenomena require more or less data to lead to human understanding and/or prediction accuracy.

Data Science - Defining a field

Data Science is different from

"using data to come to conclusions *in* science"

Data Science - Defining a field

The scientific method uses data to come to conclusions about physical phenomena and make predictions about it…

Source: NASA

Source: *Vox*

Ptolemaic model

- Ordering: Moon, Mercury, Venus, Sun, Mars, Jupiter, Saturn, Fixed Stars

Source: Wikipedia

Ptolemaic model

- Ordering: Moon, Mercury, Venus, Sun, Mars, Jupiter, Saturn, Fixed Stars
- Ptolemy did not invent or work out this order.

Source: Wikipedia

Ptolemaic model

- Ordering: Moon, Mercury, Venus, Sun, Mars, Jupiter, Saturn, Fixed Stars
- Ptolemy did not invent or work out this order.
- It comes from the ancient "Seven Heavens" religious cosmology common to the major Eurasian religious traditions.

Source: Wikipedia

Ptolemaic model

- That is, Ptolemy begins with observations.

The Scientific Method

Ptolemaic model

- Accurate
- Predictive

Ptolemaic model

- He builds a model given *social and political constraints*.

The Scientific Method

social and political constraints?!?

Source: *Vox*

So what is the point of modeling?

- Prediction

Modeling

2.3 Parsimony

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

2.4 Worrying Selectively

Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.

Source: Box GEP, *J Am Stat Assoc*, 1976

Modeling

- All models are wrong.

Source: Box GEP, *J Am Stat Assoc*, 1976

Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in

tylervigen.com

Modeling

- All models are wrong.
- Some models are useful.

Source: Box GEP, *J Am Stat Assoc*, 1976

So what is the point of modeling?

- Prediction

So what is the point of modeling?

- Prediction
- Classification

So what is the point of modeling?

- Prediction
- Classification
- Knowledge discovery

So what is the point of modeling?

- Prediction
- Classification
- Knowledge discovery
- **DOING USEFUL SHIT**

Who is this guy?

$$Coh_{i,j} = N^{-1}\left|\sum_{n=1}^{N} e^{i(\phi_{\theta i}[n]-\phi_{\theta j}[n])}\right|$$

Source: Voytek *et al.*, *Nature Neurosci* 2015

WARNING!

Postdoctoral Researcher
Natalie Schaworonkow

Postdoctoral Researcher
Stephanie Martin

I'm notoriously bad at my job

A young UCSD scientist vents about how hard it is to obtain grants

UC San Diego neuroscientist Bradley Voytek. (K.C. Alfred / Union-Tribune) (K.C. Alfred / UT San Diego/Zuma Press)

A young UCSD scientist vents about how hard it is to obtain grants

UC San Diego neuroscientist Bradley Voytek. (K.C. Alfred / Union-Tribune) (K.C. Alfred / UT San Diego/Zuma Press)

Observations

Scientists: Advertise Your Failures!

They're a part of every career, and being upfront about them can help put things in perspective

By Elizabeth Landau on September 25, 2017

Most academics are not so vocal about the setbacks they've faced, but Voytek's story has gained traction. Online discussion boards about graduate admissions have invoked Voytek's name to show that it's possible to enter a Ph.D. program—he went to the University of California, Berkeley—with a subpar transcript. One even asks: "Is the neuroscientist Bradley Voytek a real or fictional person?"

My start in Data Science

Knowledge Discovery

Source: Voytek & Voytek, *J Neurosci Methods* 2012

UC San Diego

brainSCANr

Source: Voytek & Voytek, *J Neurosci Methods* 2012

UC San Diego

Source: Voytek & Voytek, *J Neurosci Methods* 2012

Uber

uber
U
Bradley Voytek, Ph.D.
voytek@uber.com
[redacted]
@bradleyvoytek
http://uber.com

Uber NYC

Parametrization?

The previous plot has sparse data: just lat/long and time

Parametrization?

The previous plot has sparse data: just lat/long and time

From that you can get first-order calculated metrics—velocity, acceleration, etc.

Parametrization?

The previous plot has sparse data: just lat/long and time

From that you can get first-order calculated metrics—velocity, acceleration, etc.

But how can you turn such relatively sparse data into actionable business decisions (KPIs)?

Parametrization?

More importantly: how do we *teach students* to learn how to do that?

City dynamics

- Aggregating data into arbitrary geographical units: here neighborhoods, but under the hood you'd use e.g., hexagonal tiling.
- Turns lat/lng pairs and time into temporal profiles of demand between parts of a city.

Source: Voytek, *Uber, Inc.*

UC San Diego

Spatiotemporal dynamics

Source: Voytek, *Uber, Inc.*

Spatiotemporal dynamics

- **Looking at dynamics over time allows you to correlate neighborhoods within and between cities.**

- **Can identify "types" of neighborhoods: those with peak weekend and late night demands are more "party"-like whereas M-F peaks are more "business" regions.**

Source: Voytek, *Uber, Inc.*

UC San Diego