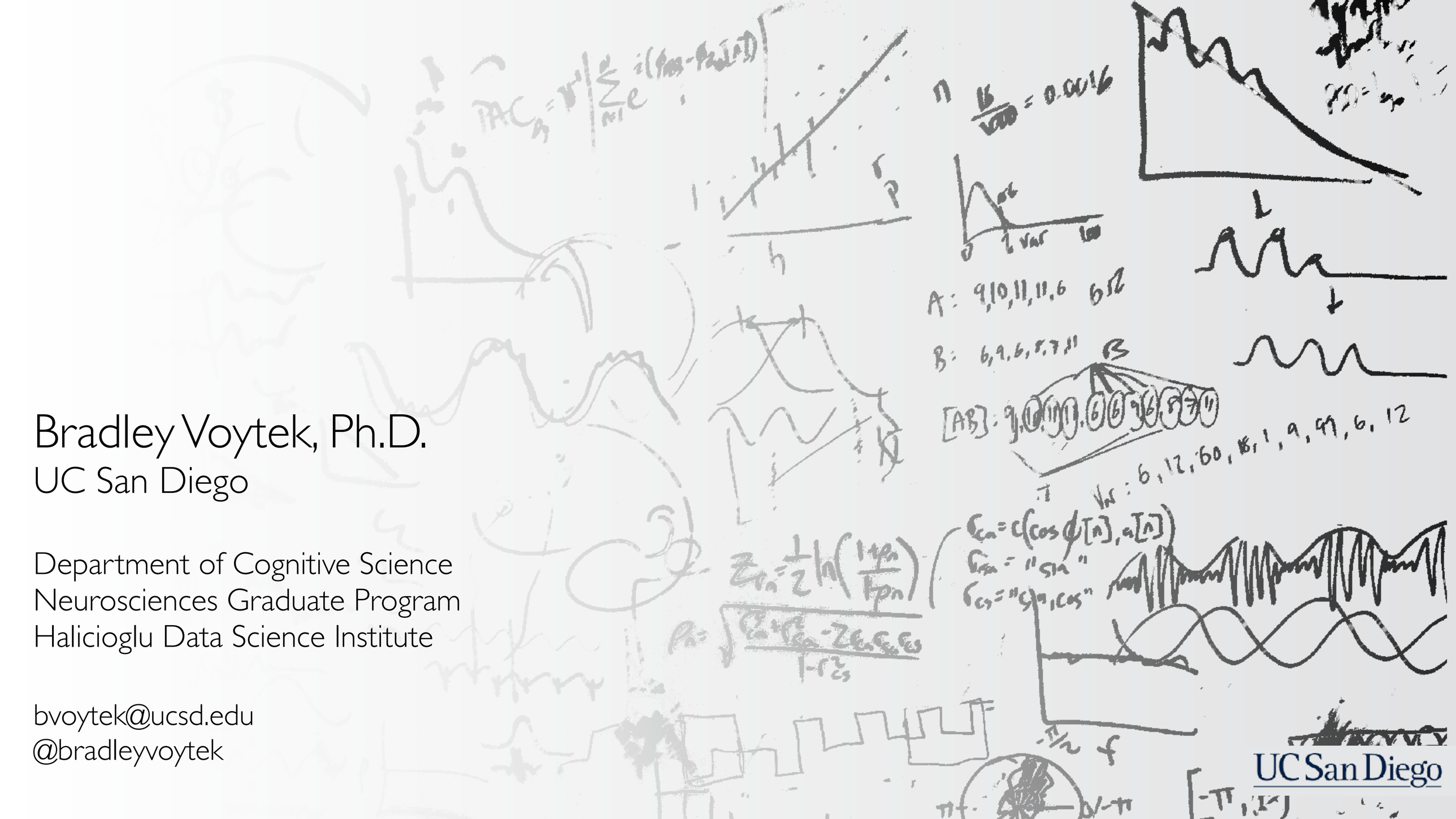


Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Neurosciences Graduate Program
Halicioglu Data Science Institute

bvoytek@ucsd.edu
@bradleyvoytek

UC San Diego

COGS 108
Data Science in Practice

Hello World!

UC San Diego

Scheduling

Lecture:

TuTh 2:00p-3:20p

Galbraith Hall 242

Voytek's Office Hours:

Wednesdays 10:00a-11:00a
and

by appointment

General stuff and junk

- First quarter?
-
-

General stuff and junk

- First quarter?
- Data Science majors?
-

General stuff and junk

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

General stuff and junk

- First quarter?
- 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 and junk

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

General stuff and junk

- 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

- Piazza: https://piazza.com/ucsd/winter2018/cogs108/home
- GitHub: https://github.com/COGS108/
- TritonEd

GRADING

- Five assignments (12% each)
- Final Project (35%)
- Participation (5%)

Participation

- Attendance *required* for guest lectures

Past guest lectures

- **UCSD faculty**

- **Practicing data scientists:**
  - Eli Bressert, PhD: Manager, Data Engineering & Analytics, Netflix
  - Mina Doroud, PhD: Data Scientist, Twitter (Senior Data Scientist, LinkedIn)
  - Hiroki Hiyama, PhD: Senior Data Scientist, Uber
  - Emi Nomura, PhD: Data Scientist, Jawbone (Senior Manager, Data Science, Pandora)
  - Maksim Pecherskiy: Chief Data Officer, City of San Diego
  - Sarah Rich, PhD: Data Scientist, Twitter
  - Claire Dorman, PhD: Data Scientist, Pandora
  - Franziska Bell, PhD: Senior Data Science Manager, Uber
  - John Myles White: Research Scientist, Facebook (Author: *Machine Learning for Hackers*)
  - Carlos Gomez-Uribe, PhD: Director, Core Data Science, Facebook (Statistician, Google; VP Product Innovation, Netflix)

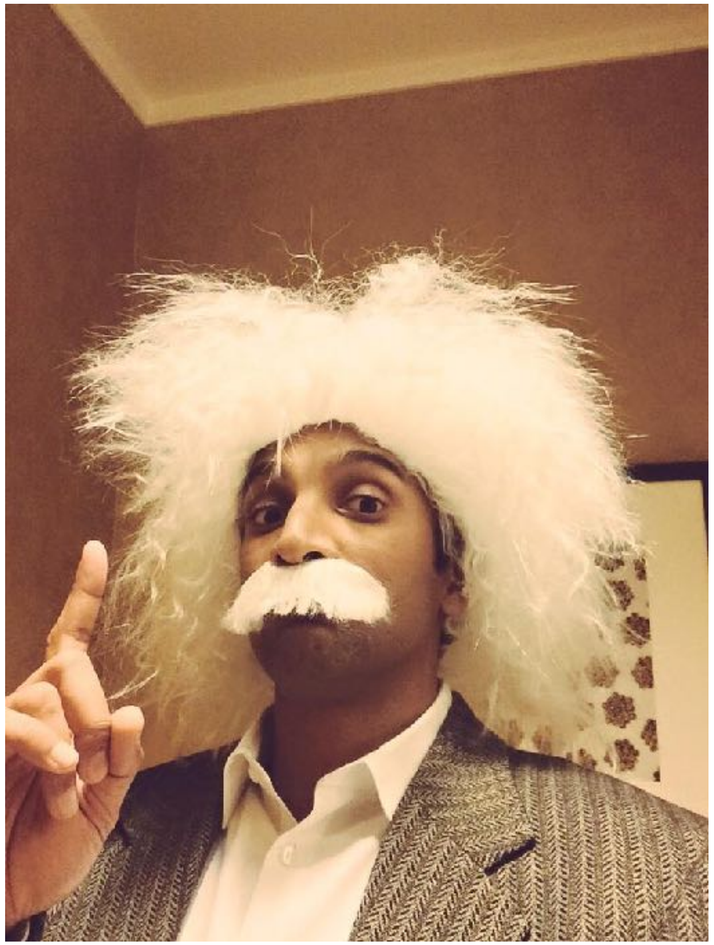
Guest lecturers

- **Jan 25** Kevin Novak: Chief Data Officer, Tala (Formerly: Head of Data & Engineering, *Uber*)
- **Feb 13** Ilkay Altintas, PhD: Chief Data Science Officer, San Diego Supercomputer Center (SDSC)
- **Feb 27** Josh Wills: Director of Data Engineering, Slack (Formerly: Director of Data Science, Cloudera; Analytics, Google)

Proposed course order

1. Introduction: Why data analysis? (prediction and classification)
2. Python!
3. Data Science in Python (jupyter, pandas, numpy, scipy, scikit-learn, etc.)
4. Data gathering, wrangling, and cleaning (How do you find and clean data? (JSON, CSV, XML, SQL, APIs)
5. Data privacy, ethics, and HIPAA (anonymization)
6. **Jan 25** Guest lecture: Kevin Novak: Chief Data Officer, Tala (Formerly: Head of Data & Engineering, *Uber*)
7. Basic data visualization
8. Data intuition and the "sniff test" (Fermi estimation; distributions and outliers: histograms, CDF, PDFs)
9. Non-parametric statistics
10. Linear modeling
11. **Feb 13** Ilkay Altintas, PhD: Chief Data Science Officer, San Diego Supercomputer Center (SDSC)
12. NO CLASS!
13. OLS (optimization)
14. Multiple linear regression and collinearities
15. **Feb 27** Josh Wills: Director of Data Engineering, Slack (Formerly: Director of Data Science, Cloudera; Analytics, Google)
16. Model validation (bootstrapping, resampling, k-fold, leave-p-out, train/test)
17. Dimensionality reduction (PCA); clustering and classification (k-means, knn, SVM)
18. Feature selection
19. NLP and text-mining (bag of words, tf-idf, sentiment analysis)
20. Geospatial analysis

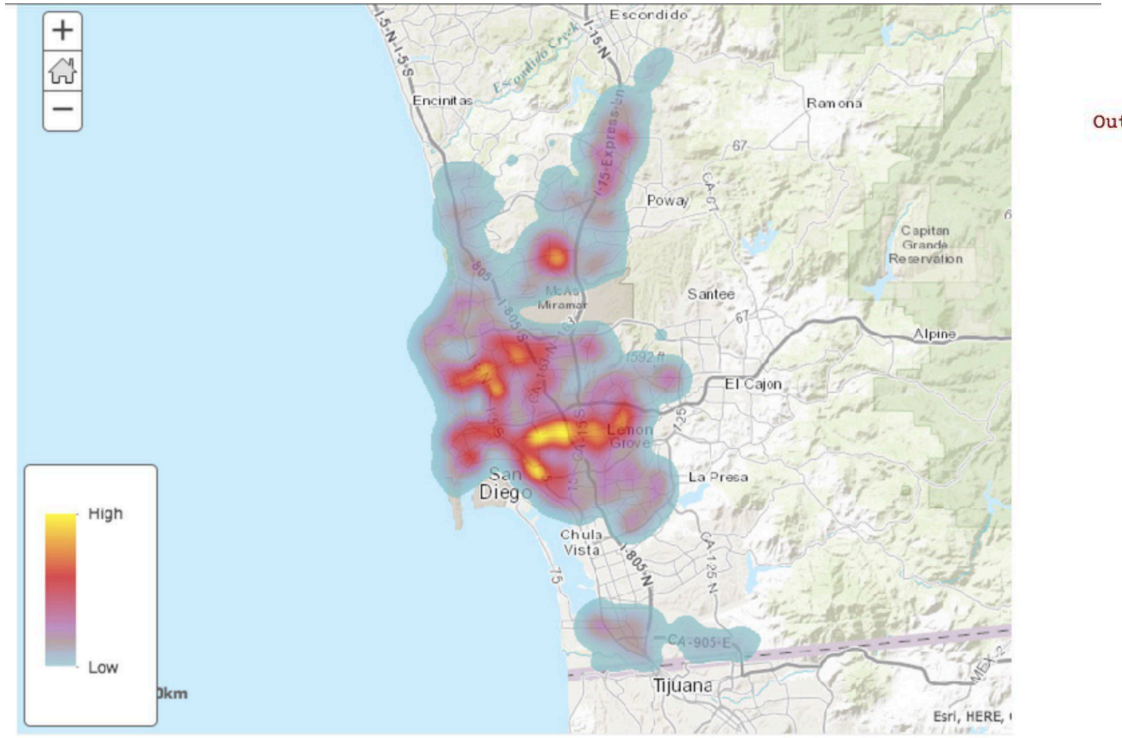
Final Project!



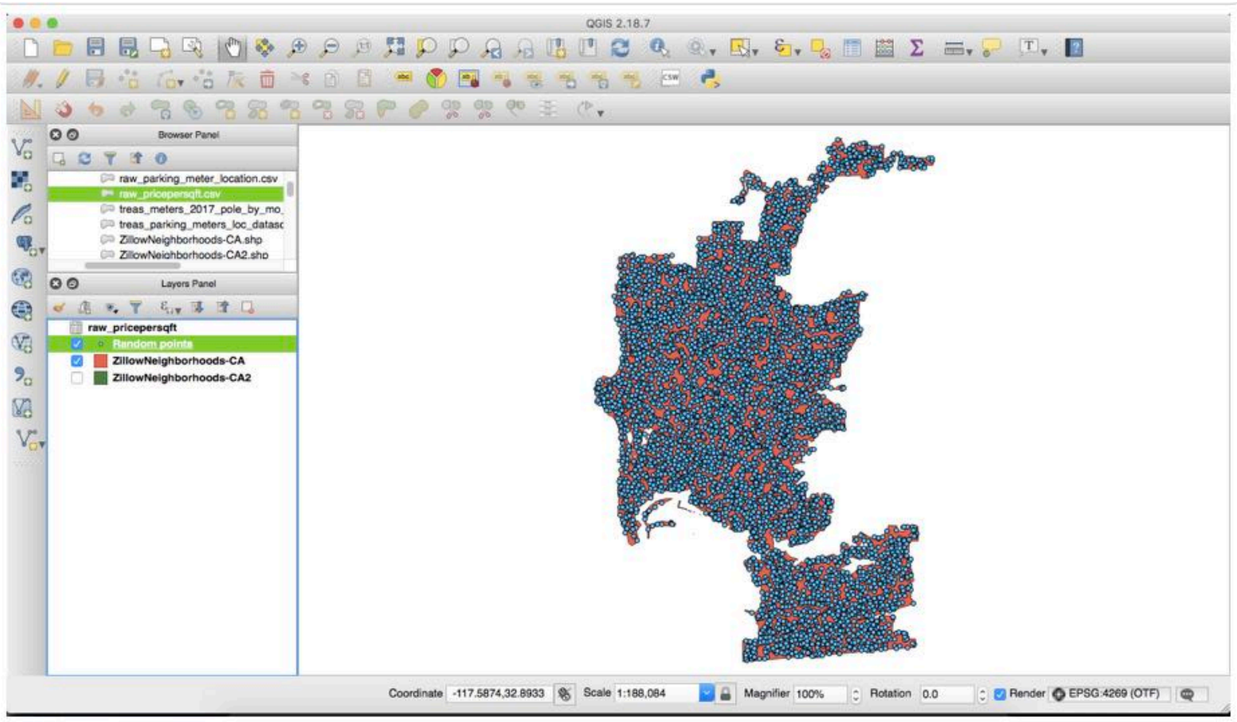
Final Project!

- Project due on Final Exam day *in lieu of* an actual Exam!
- Deadline: 23:59, Thu, March 22, 2018

Final Project examples: SD potholes



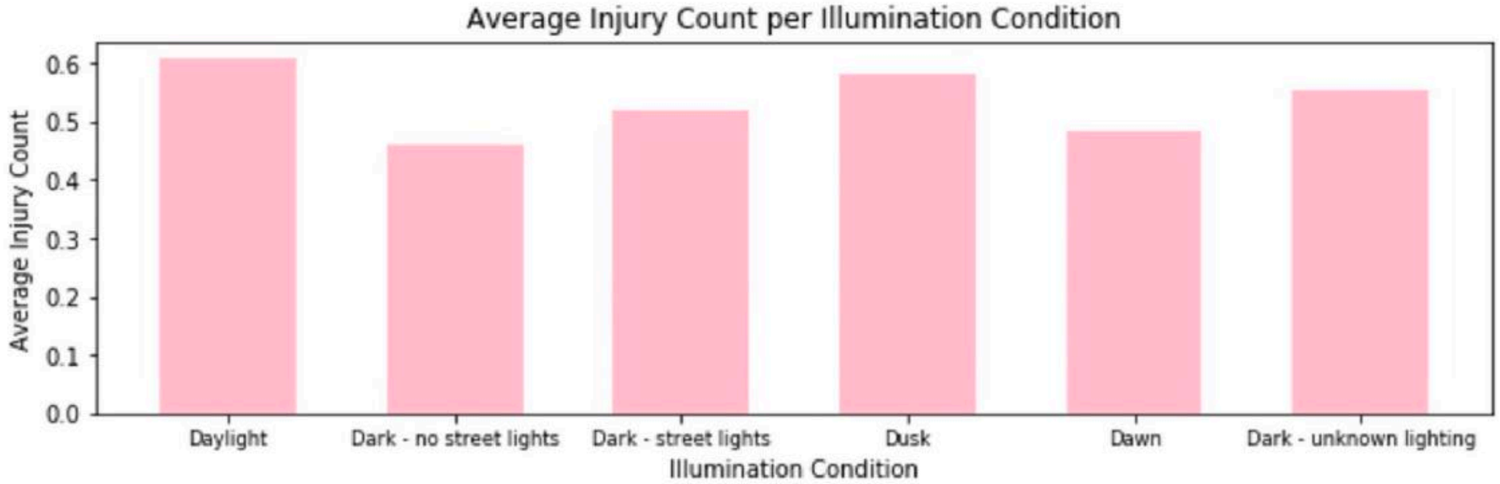
Out[5]:



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, instead 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



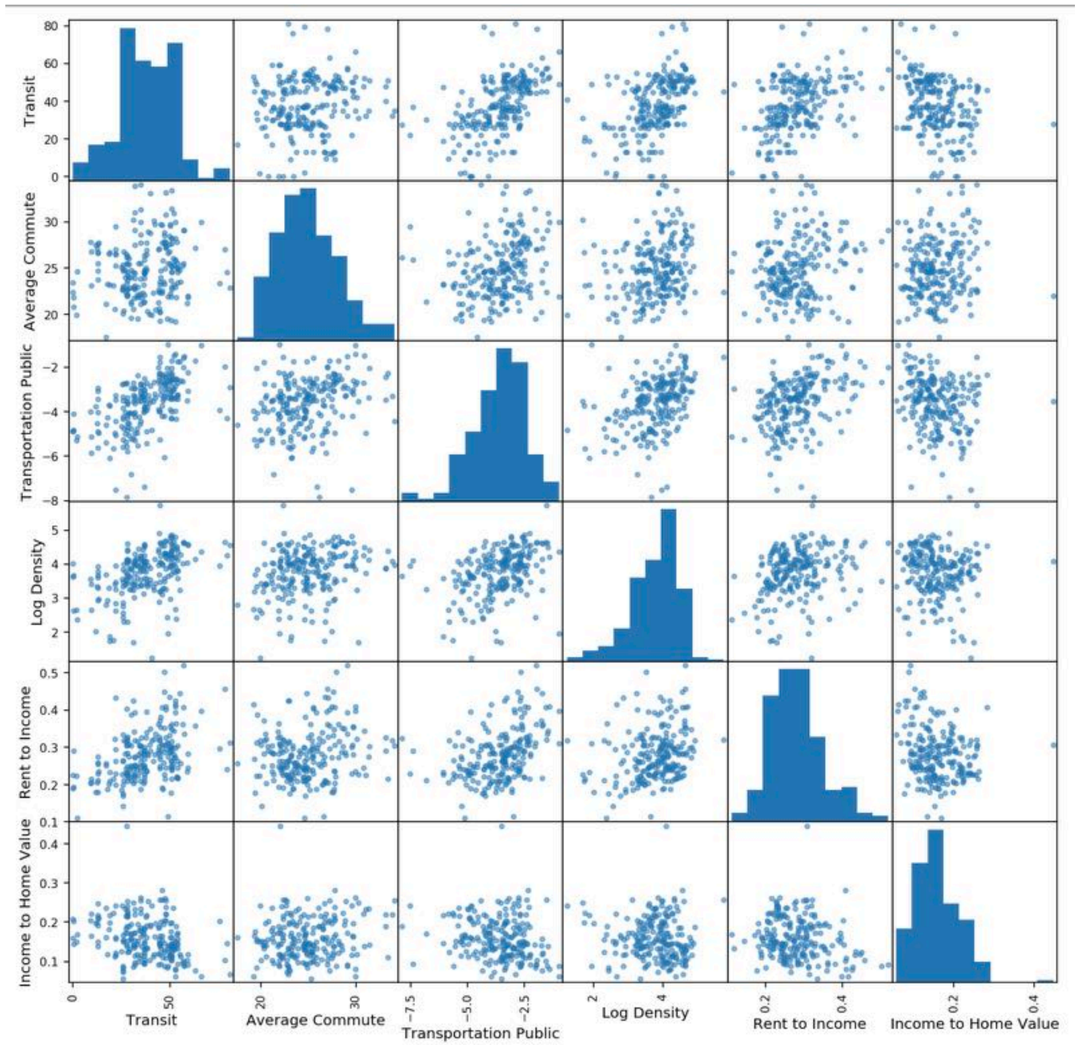
Data Analysis and Results

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

Walk Score®



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.

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:
neuroscientist

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:
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?

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:
journalist

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

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."

UC San Diego

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

BIG DATA

Data science

BIG DATA

Data science still woefully short on science

BIG DATA

Data science still woefully short on science

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."

UC San Diego

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?

UC San Diego

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?

???!!???

UC San Diego

Source: NYU Data Science

What is Data Science?

BERKELEY
Institute for
Data Science

Data Science Studies
Berkeley

This working group designs and conducts research projects across disciplines and methods to understand the challenges posed by data scientists' practices in the academic context. In collaboration with our partner institutions, it also develops innovative quantitative and qualitative metrics to measure the evolution of data science environments.

Source: BIDS

What is Data Science?

A New Field Emerges

There is significant and growing demand for data-savvy professionals in businesses, public agencies, and nonprofits. The supply of professionals who can work effectively with data at scale is limited, and is reflected by rapidly rising salaries for data engineers, data scientists, statisticians, and data analysts.

UC San Diego

Source: BIDS

What is Data Science?

A New Field Emerges

There is significant and growing demand for data-savvy professionals in businesses, public agencies, and nonprofits. The supply of professionals who can work effectively with data at scale is limited, and is reflected by rapidly rising salaries for data engineers, data scientists, statisticians, and data analysts.

The field of data science is emerging at the intersection of the fields of social science and statistics, information and computer science, and design. The UC Berkeley School of Information is ideally positioned to bring these disciplines together and to provide students with the research and professional skills to succeed in leading edge organizations.

Source: BIDS

What is Data Science?

A New Field Emerges

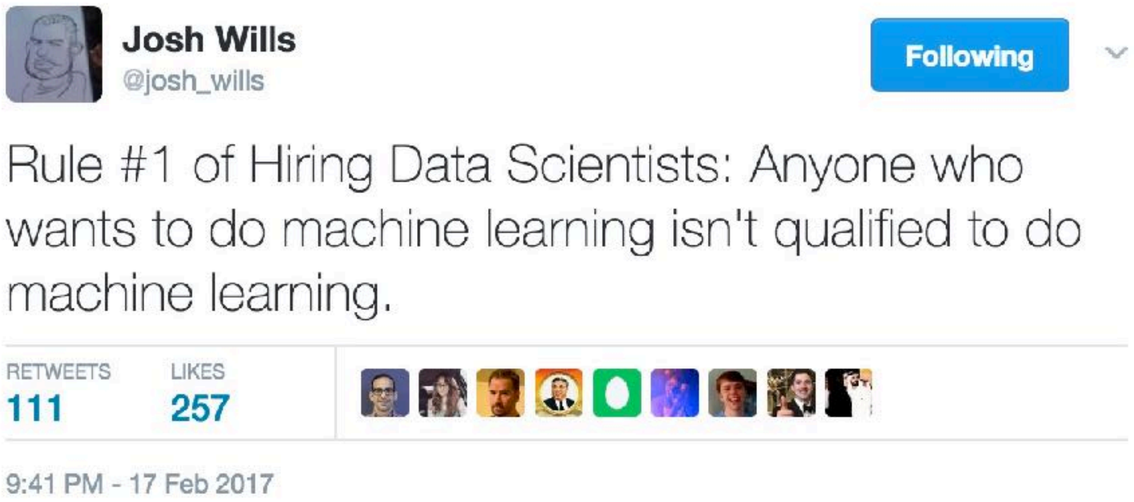
There is significant and growing demand for data-savvy professionals in businesses, public agencies, and nonprofits. The supply of professionals who can work effectively with data at scale is limited, and is reflected by rapidly rising salaries for d , and data analysts.

??!!??

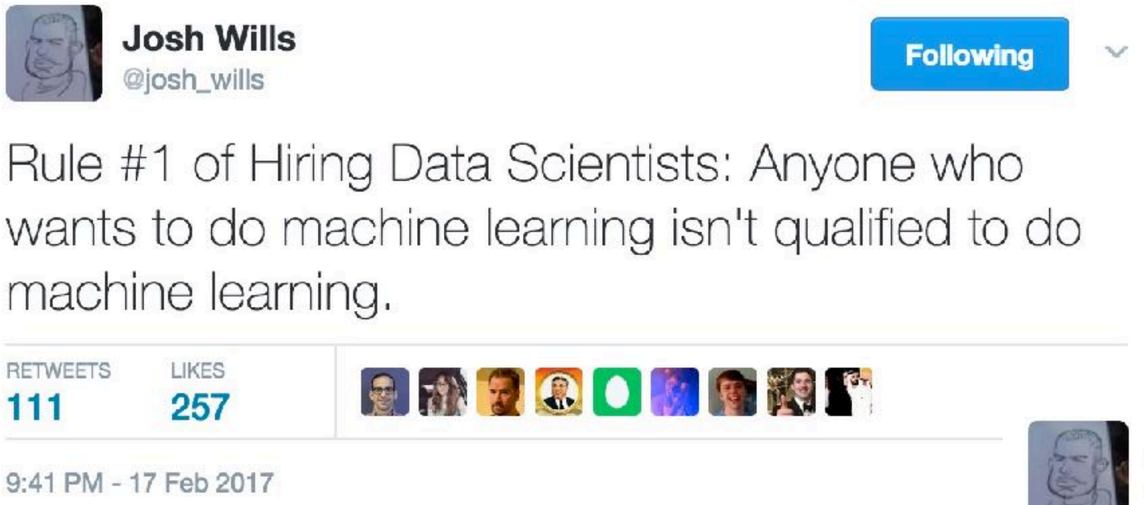
The field of data science is e f social science and statistics, information and computer science, and design. The UC Berkeley School of Information is ideally positioned to bring these disciplines together and to provide students with the research and professional skills to succeed in leading edge organizations.

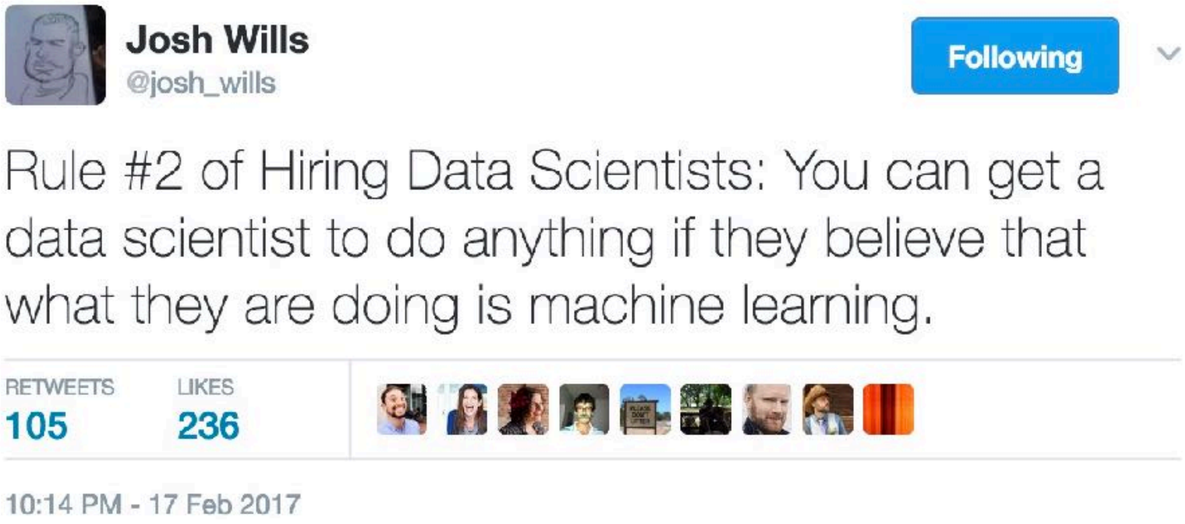
Source: BIDS

What *isn't* Data Science?



What *isn't* Data Science?





What *isn't* Data Science?

What is Data Science?

"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."

UC San Diego

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

Data Science - Defining a field

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…

Data Science - Defining a field

In contrast Data Science is the study of:

1) How and why data can be used that way, what kinds of data are there.

2) What makes "good" versus "bad" quality data for different questions, etc.

Data Science *vs.* Data Engineering

Data Science is the empirical study of data whereas Data Engineering is the application of data science methods and techniques to draw conclusions.

Data Science *vs.* Data Engineering

This is directly comparable to the difference between Computer Science (the scientific study) and Computer Engineering (the application and, commonly, job title).

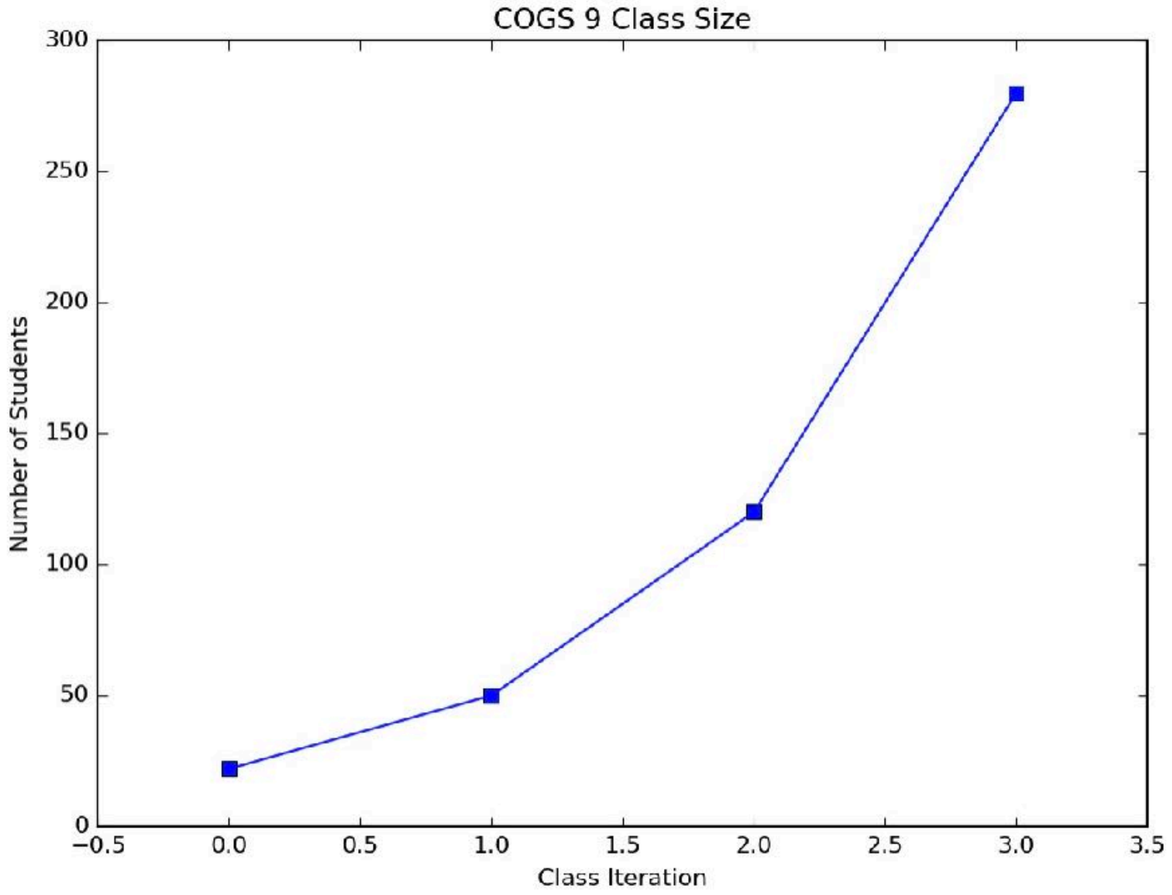
Data Science at UCSD

Data Science at UCSD

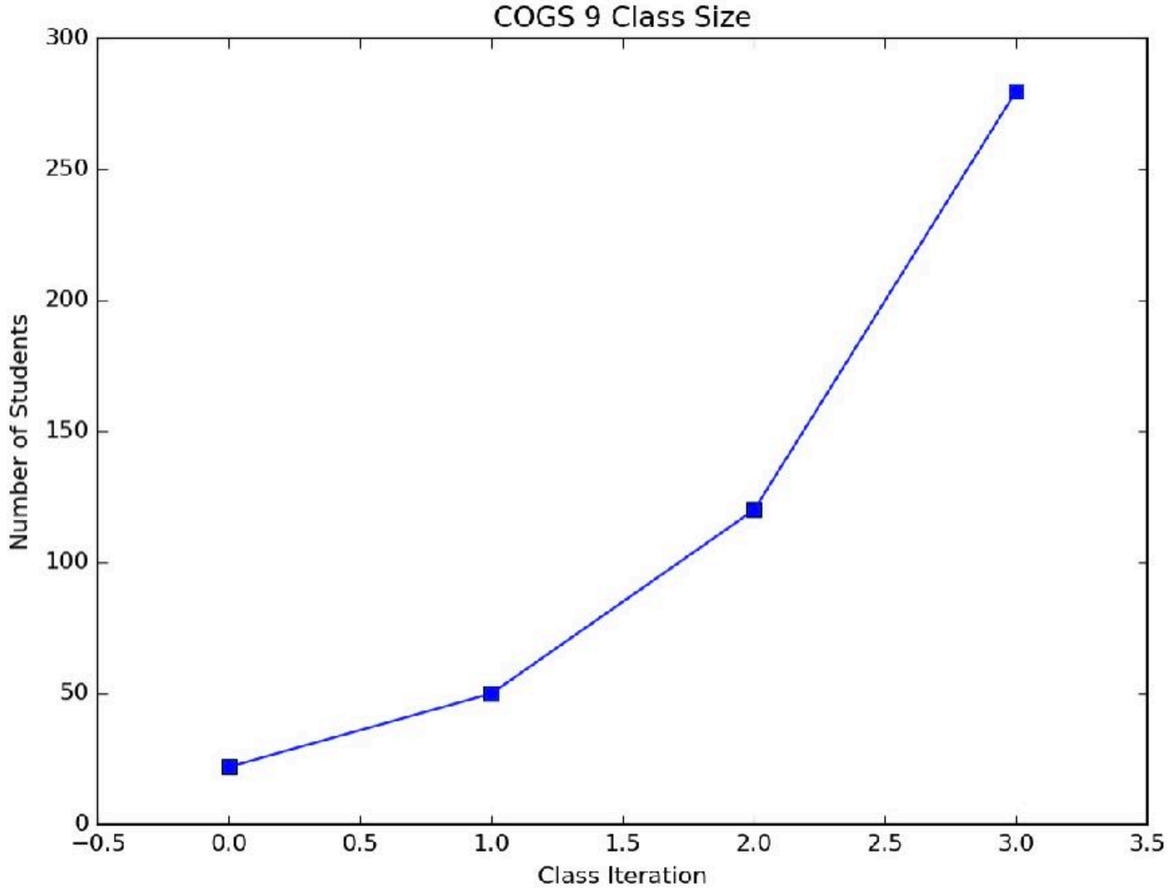
Facebook pioneer donates $75 million to UCSD for data science

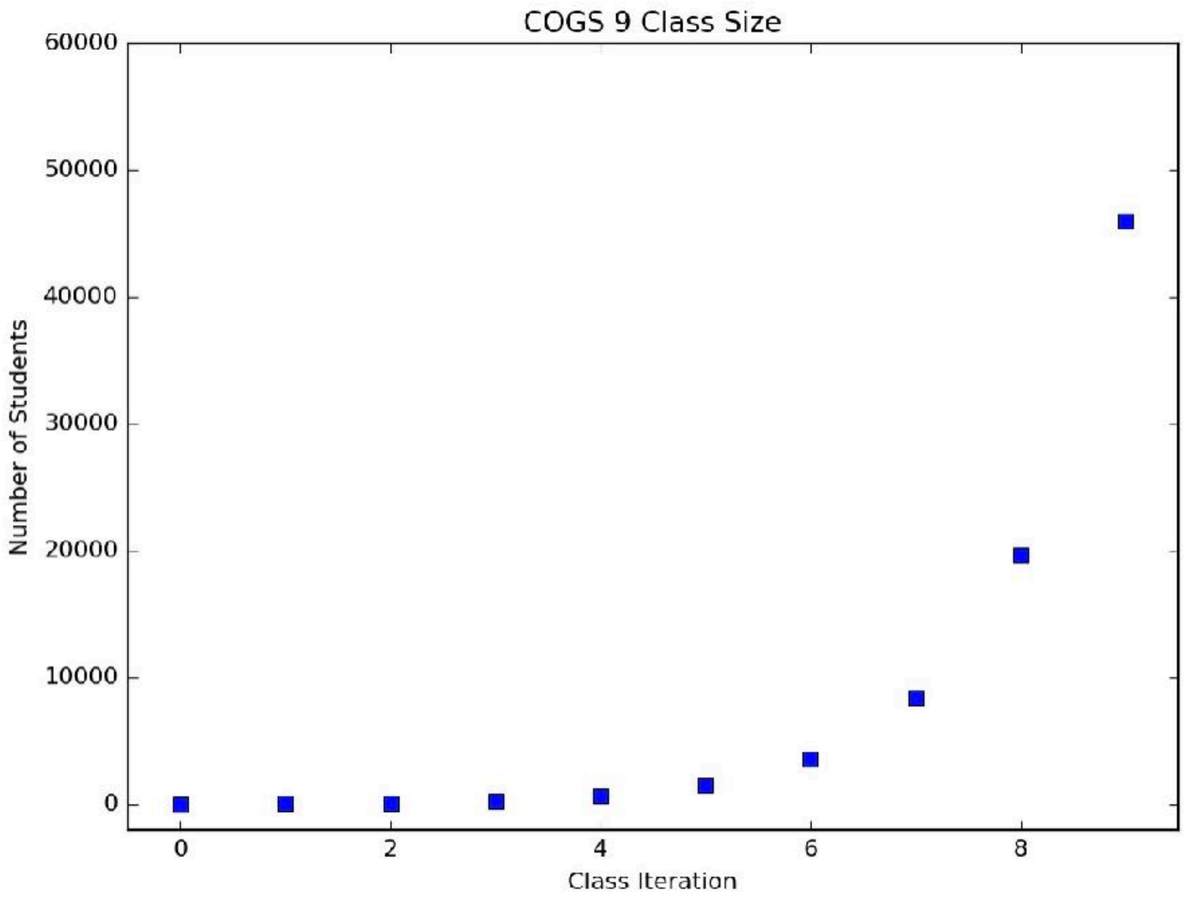
Taner Halicioglu is donating $75 million to UC San Diego to make his alma mater a national leader in data science. (Erik Jepsen / UC San Diego)

Data Science at UCSD



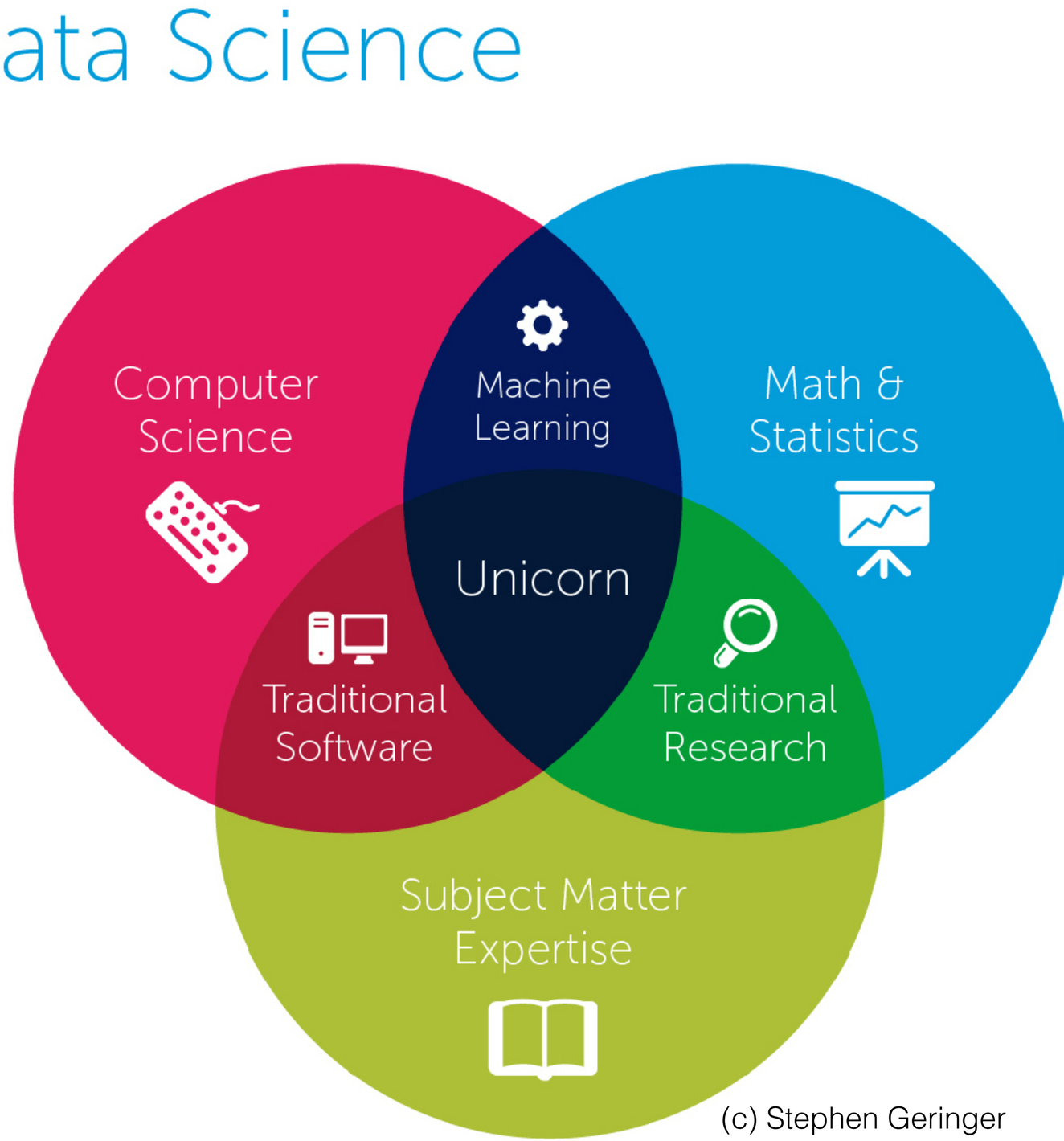
Data Science at UCSD





A New UCSD Major & Minor

- Joint effort between the departments of Computer Science, Math, and Cognitive Science
- Industry & research demand for the "data scientist"



The Human Side of Data Science

or

Why is Cognitive Science a core DS department?

UC San Diego

UCSD Data Science Major

- Joint effort between Computer Science, Math, and Cognitive Science

UCSD Cognitive Science

Machine Learning and Neural Computation

This area of specialization is intended for majors interested in computational and mathematical approaches to modeling cognition or building cognitive systems, theoretical neuroscience, as well as software engineering and data science. Allowed electives include advanced courses in neural networks, artificial intelligence, and computer science.

Who is this guy?

Wait… Who are ***you****?*

bit.ly/COGS108Wi18

Who is this guy?

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

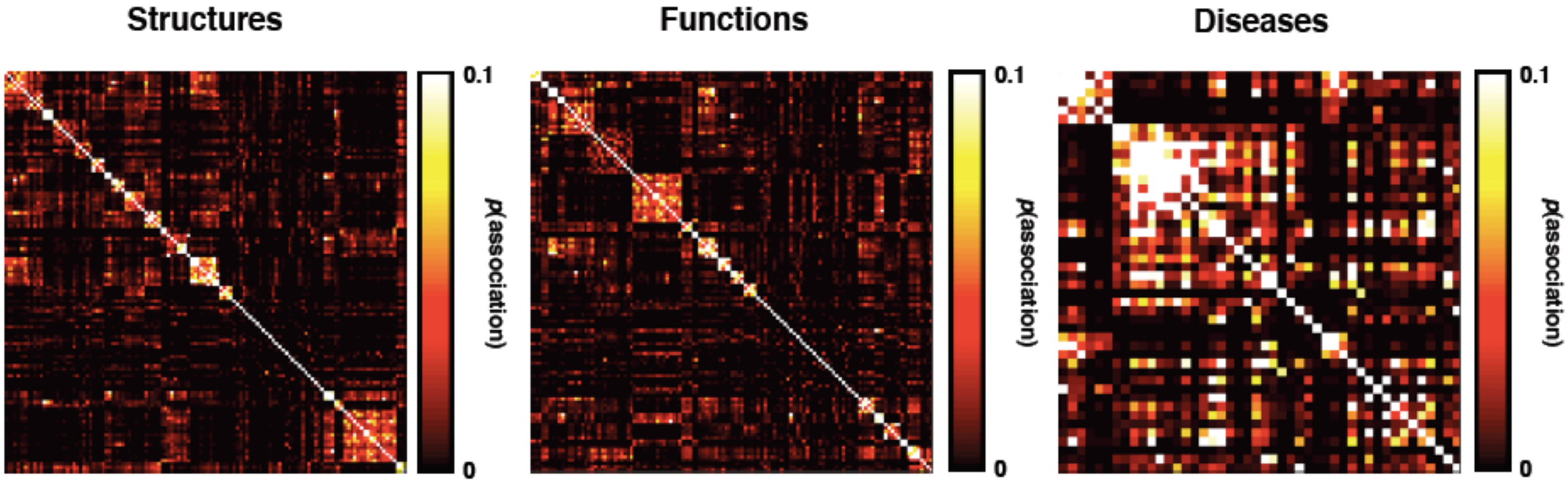
Scientists: Advertise Your Failures!

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

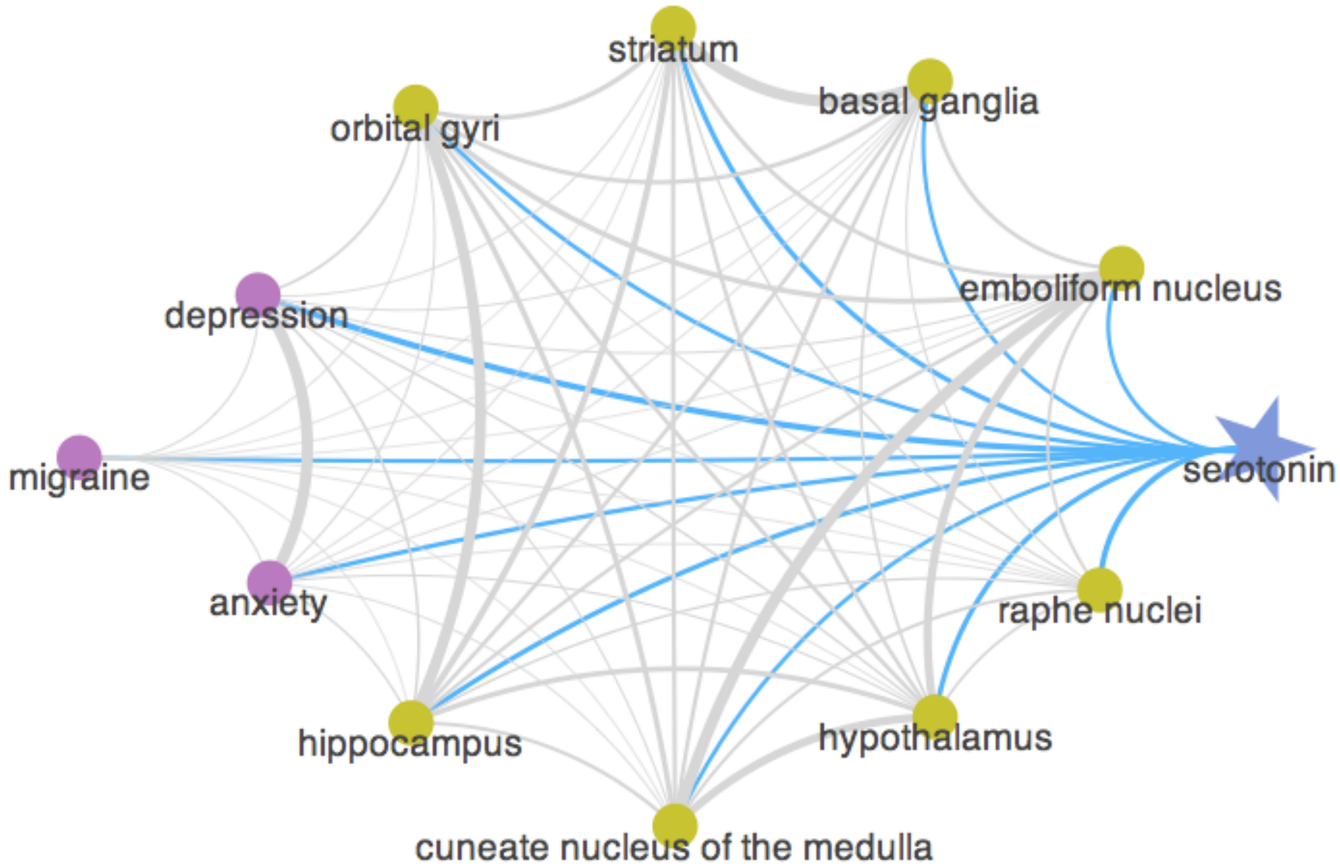
Most academics are not so vocal about the setbacks they've faced, but Voytek's story has gained traction. Online discussion boards about graduate admissions have invoked Voytek's name to show that it's possible to enter a Ph.D. program—he went to the University of California, Berkeley—with a subpar transcript.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

Knowledge Discovery

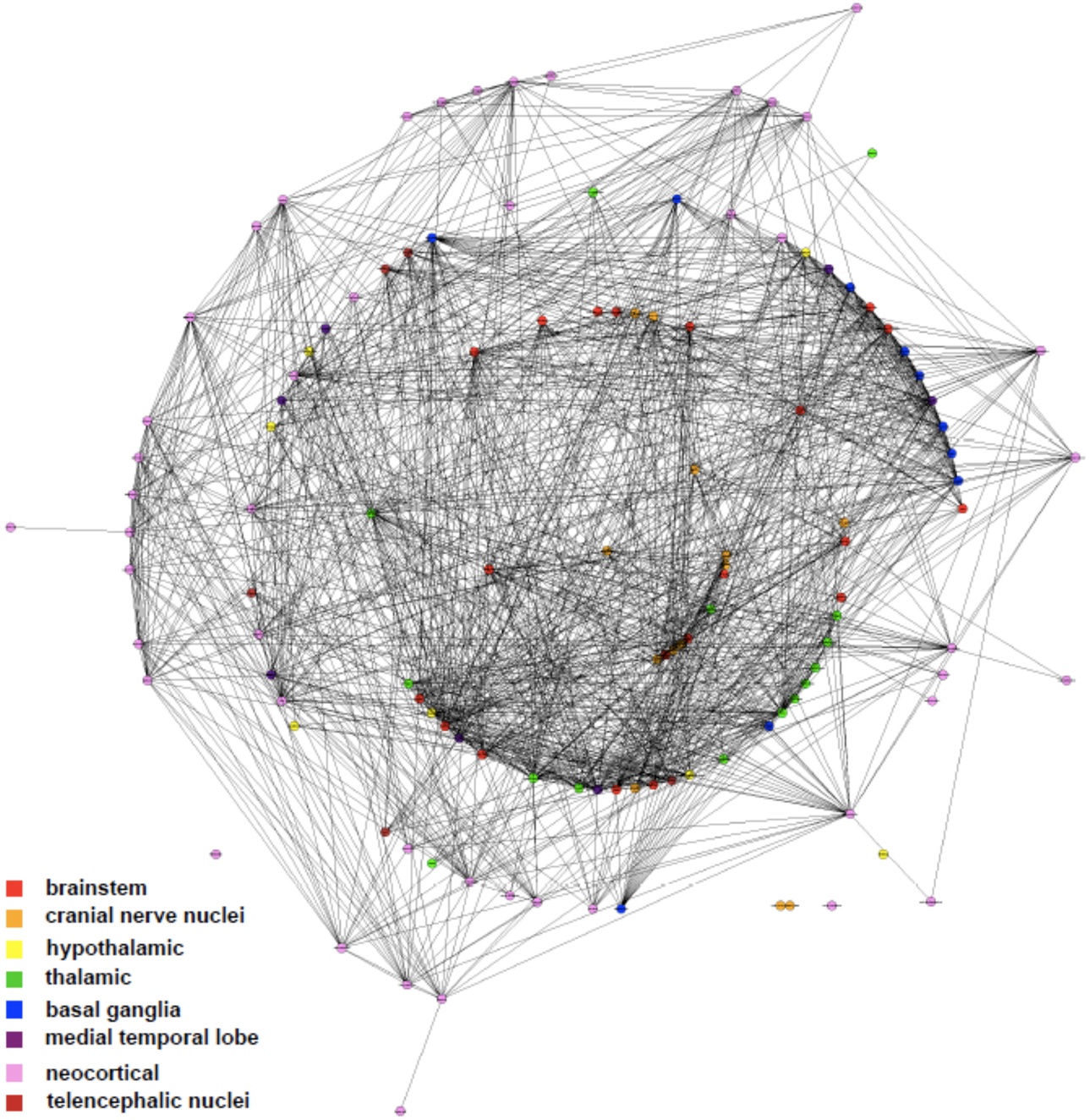


Source: Voytek & Voytek, *J Neurosci Methods* 2012

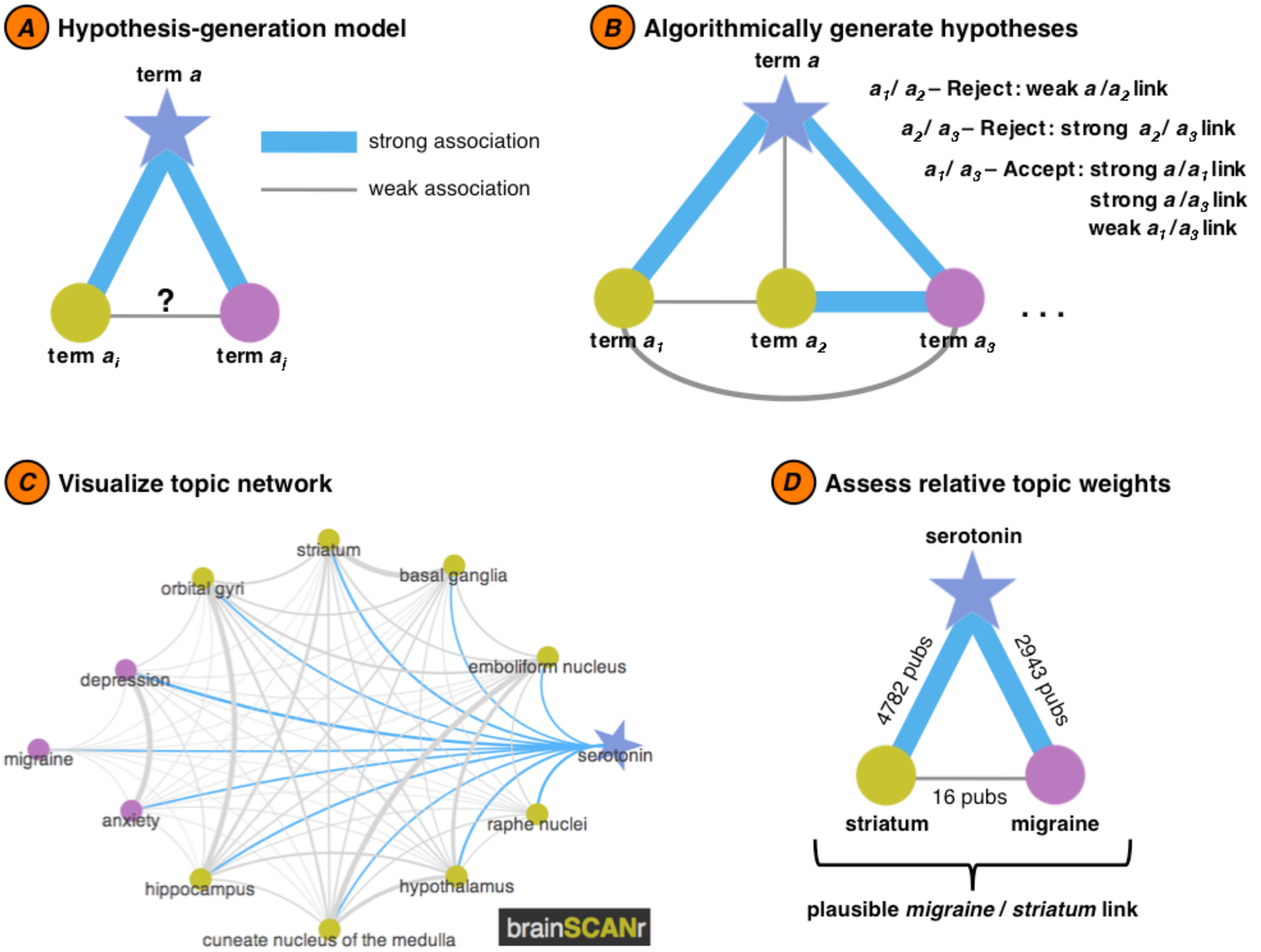
brainSCANr



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

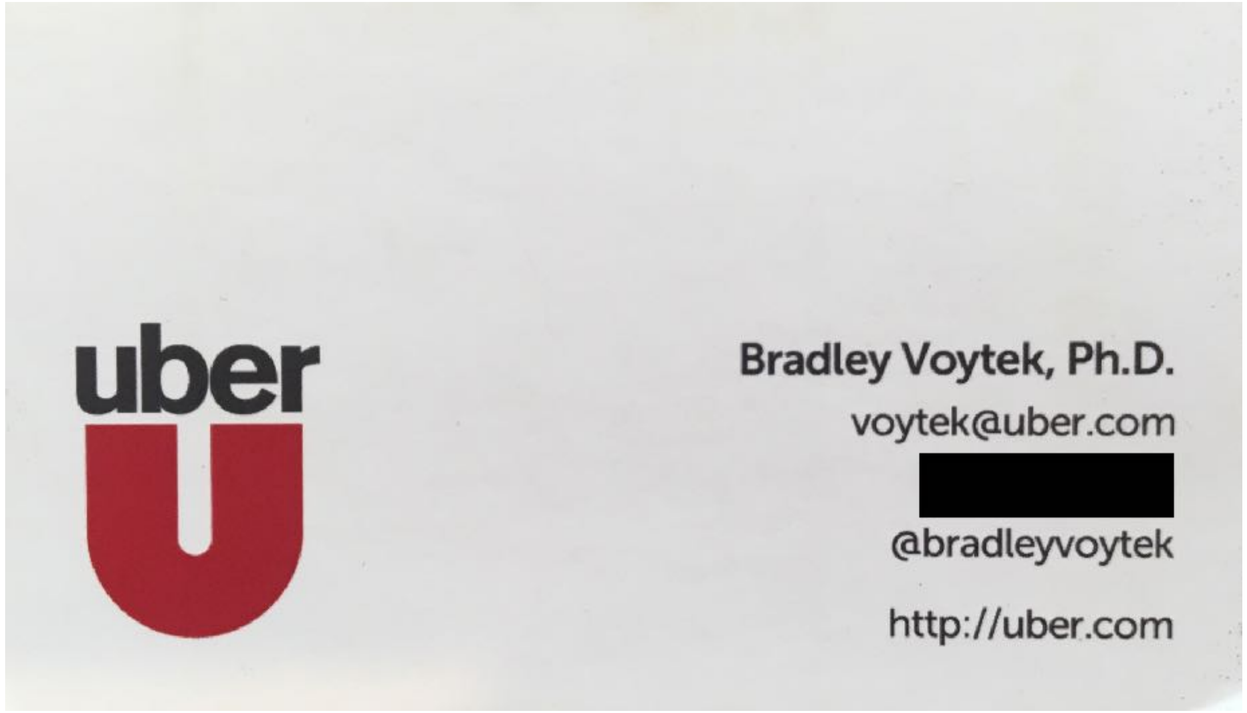


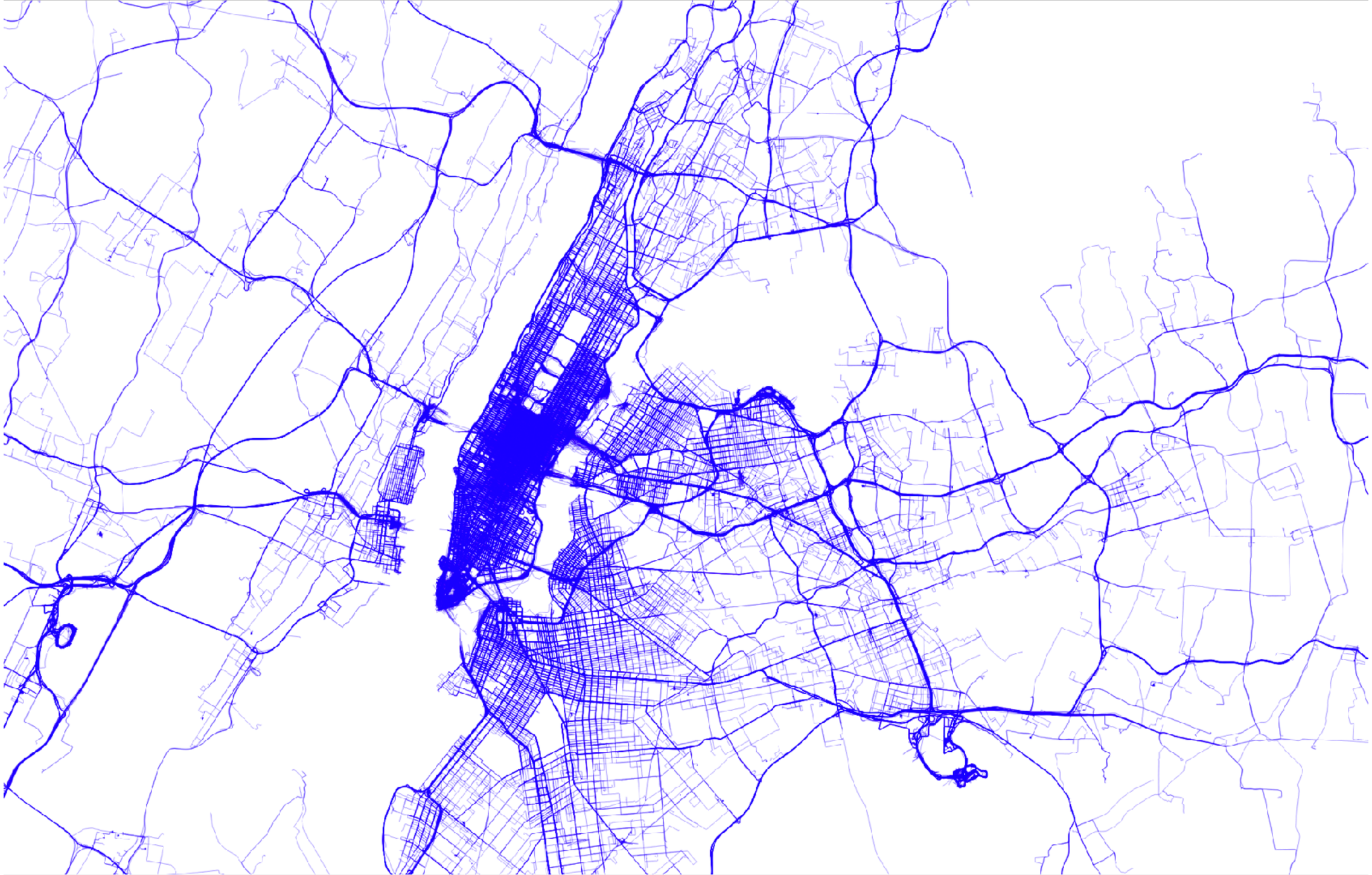
Source: Voytek & Voytek, *J Neurosci Methods* 2012



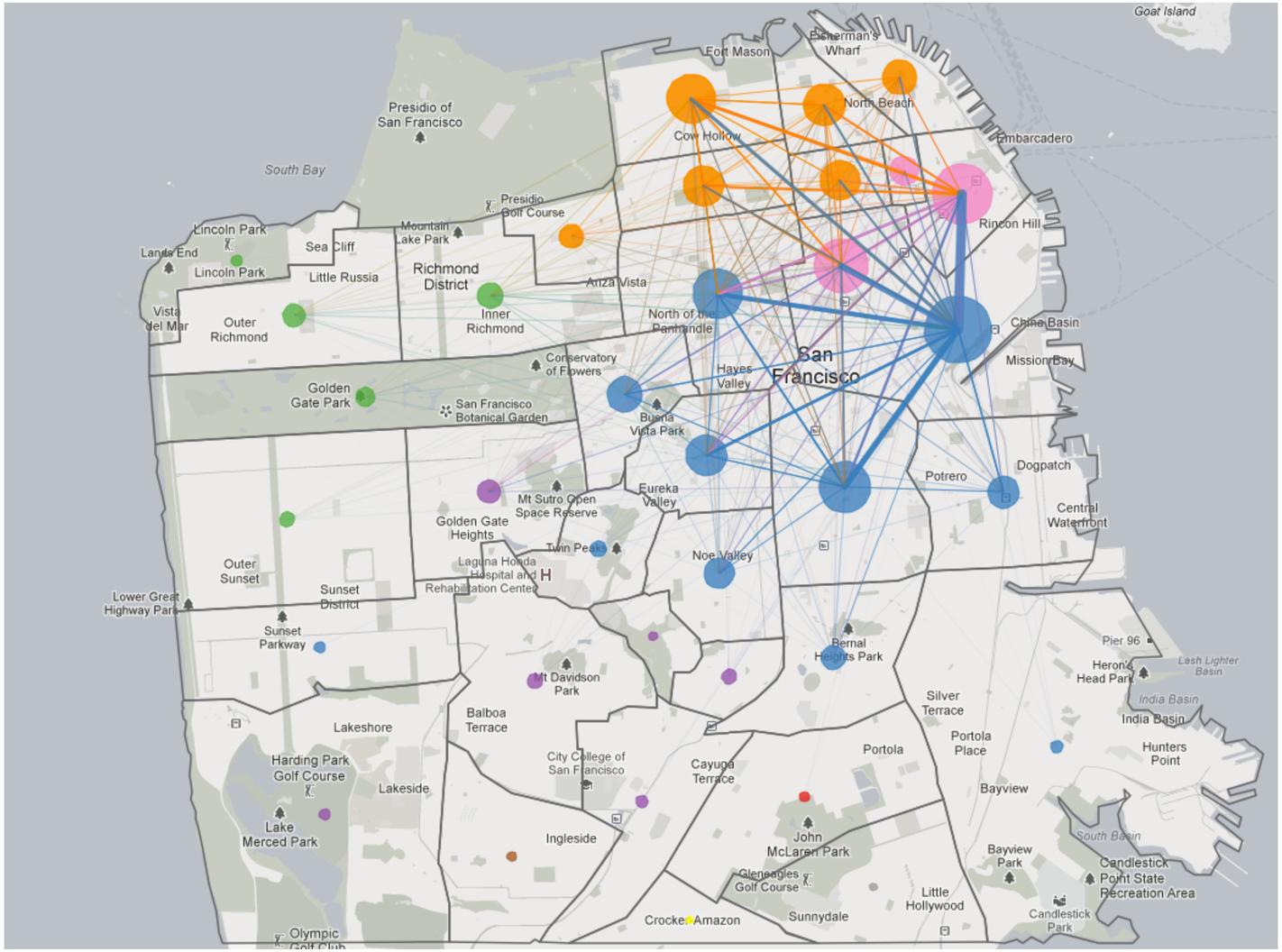
Source: Voytek & Voytek, *J Neurosci Methods* 2012

UBER


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



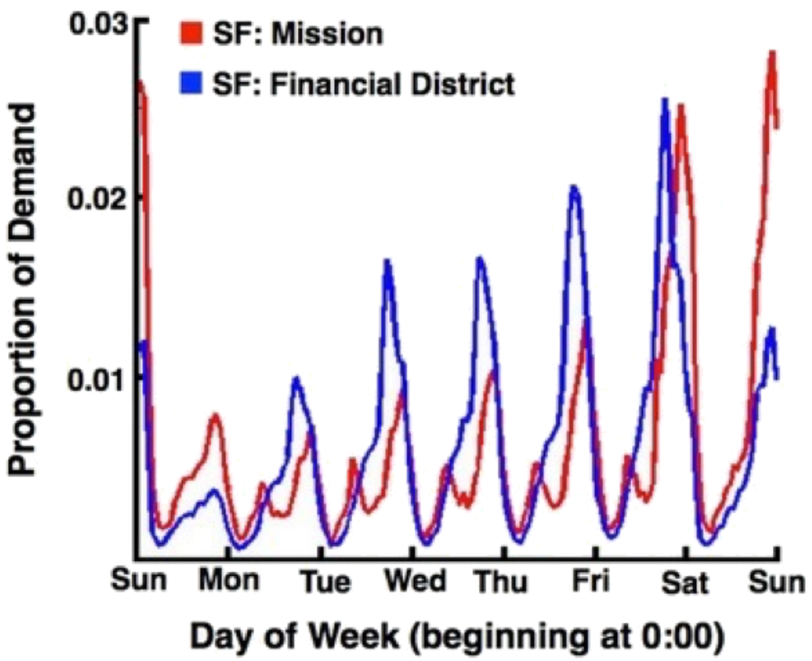
City dynamics

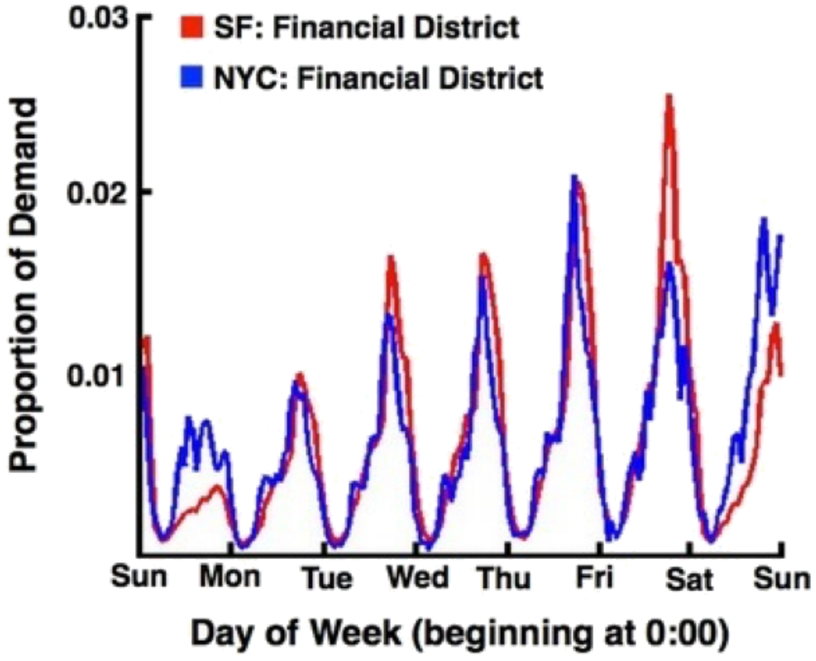


Source: Voytek, *Uber, Inc.*

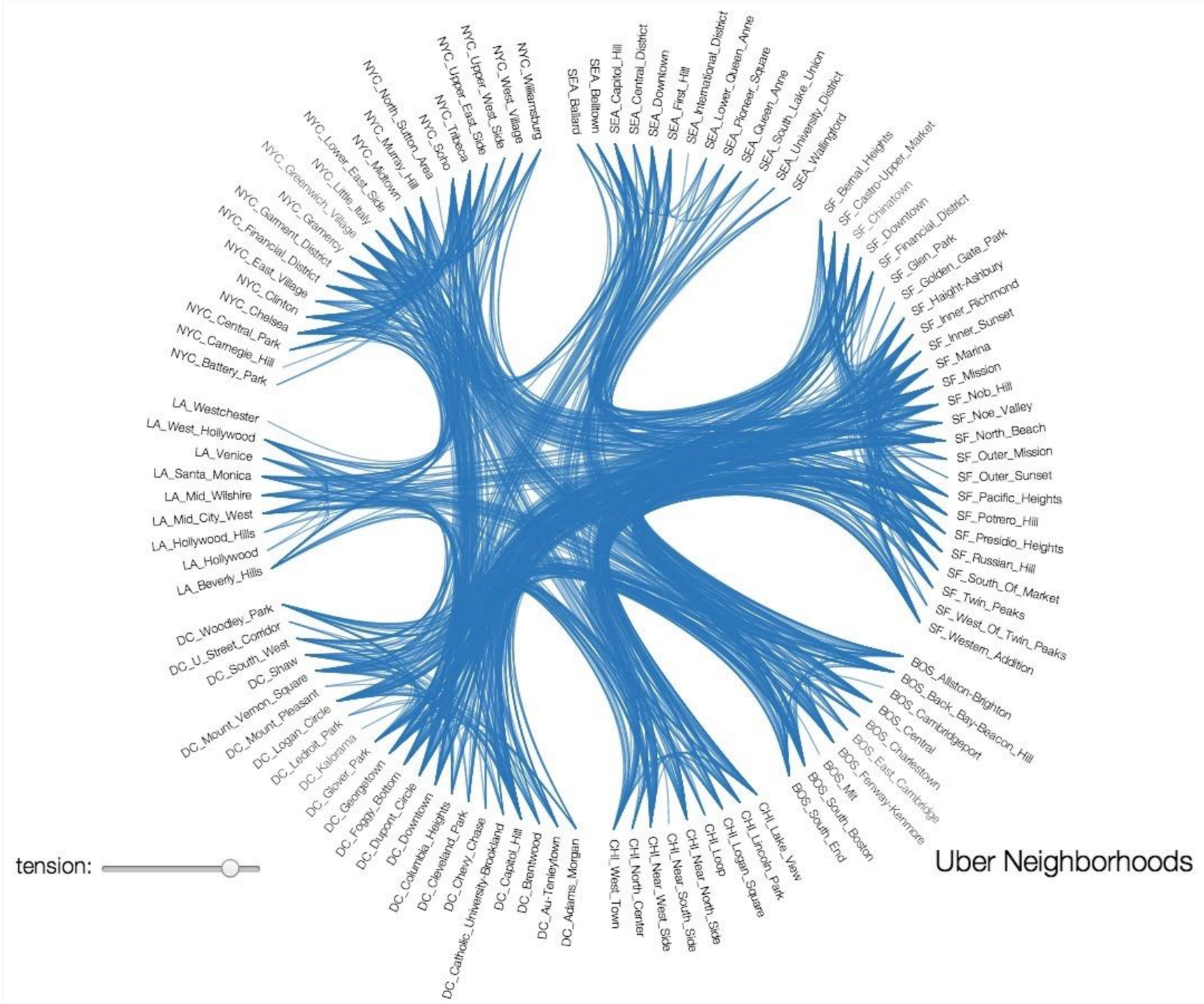
UC San Diego

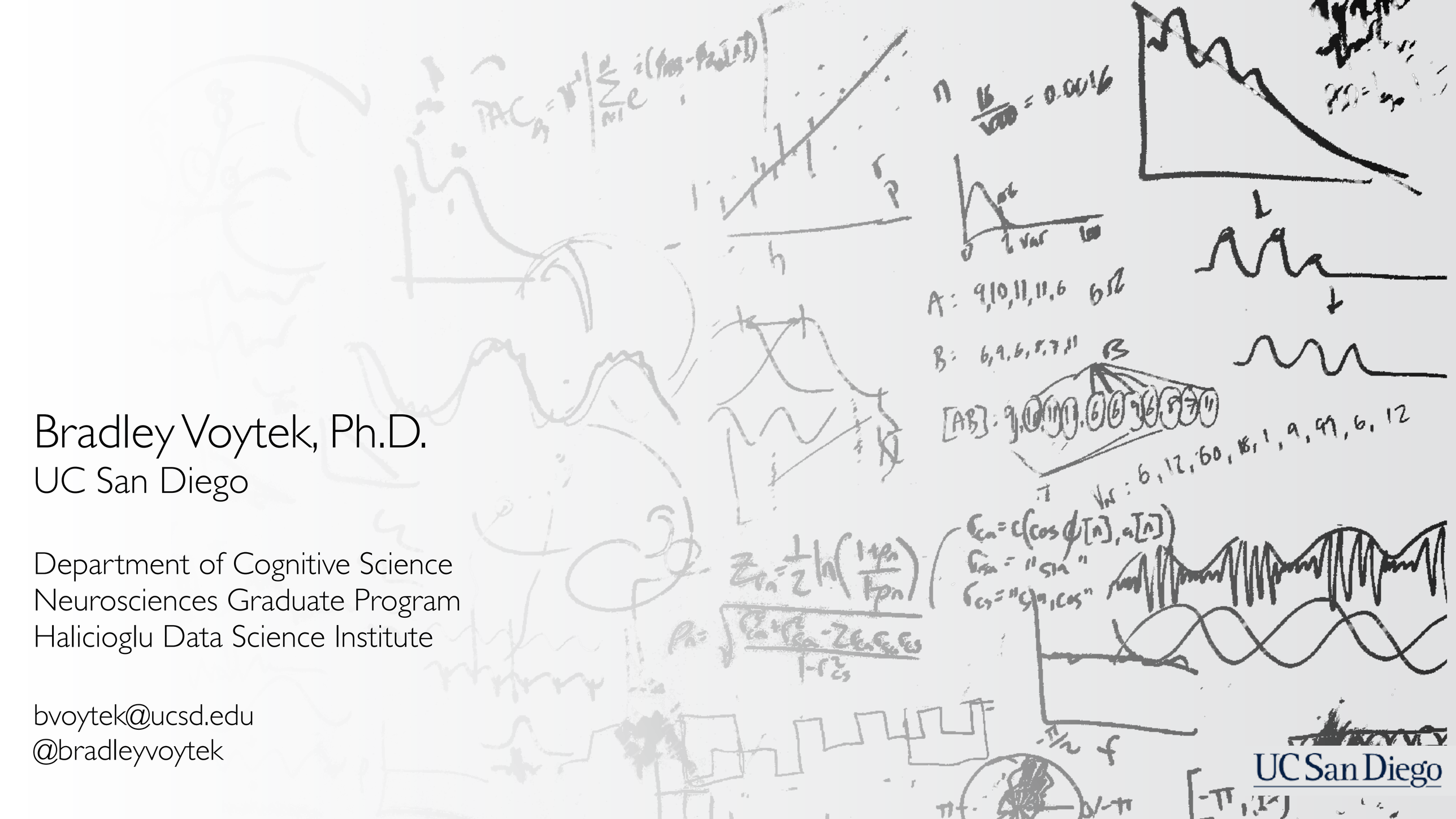
Spatiotemporal dynamics





Source: Voytek, *Uber, Inc.*


NYC_Williamsburg
NYC_West_Village
NYC_Upper_West_Side
NYC_Upper_East_Side
NYC_Tribeca
NYC_Soho
NYC_North_Sutton_Area
NYC_Murray_Hill
NYC_Midtown
NYC_Lower_East_Side
NYC_Little_Italy
NYC_Greenwich_Village
NYC_Gramercy
NYC_Garment_District
NYC_Financial_District
NYC_East_Village
NYC_Clinton
NYC_Chelsea
NYC_Central_Park
NYC_Carnegie_Hill
NYC_Battery_Park
SEA_Ballard
SEA_Belltown
SEA_Capitol_Hill
SEA_Central_District
SEA_Downtown
SEA_First_Hill
SEA_International_District
SEA_Lower_Queen_Anne
SEA_Pioneer_Square
SEA_Queen_Anne
SEA_South_Lake_Union
SEA_University_District
SEA_Wallingford
SF_Bernal_Heights
SF_Castro-Upper_Market
SF_Chinatown
SF_Downtown
SF_Financial_District
SF_Glen_Park
SF_Golden_Gate_Park
SF_Haight-Ashbury
SF_Inner_Richmond
SF_Inner_Sunset
SF_Marina
SF_Mission
SF_Nob_Hill
SF_Noe_Valley
SF_North_Beach
SF_Outer_Mission
SF_Outer_Sunset
SF_Pacific_Heights
SF_Potrero_Hill
SF_Presidio_Heights
SF_Russian_Hill
SF_South_Of_Market
SF_Twin_Peaks
SF_West_Of_Twin_Peaks
SF_Western_Addition
LA_Westchester
LA_West_Hollywood
LA_Venice
LA_Santa_Monica
LA_Mid_Wilshire
LA_Mid_City_West
LA_Hollywood_Hills
LA_Hollywood
LA_Beverly_Hills
DC_Woodley_Park
DC_U_Street_Corridor
DC_South_West
DC_Shaw
DC_Mount_Vernon_Square
DC_Mount_Pleasant
DC_Logan_Circle
DC_Ledroit_Park
DC_Kalorama
DC_Glover_Park
DC_Georgetown
DC_Foggy_Bottom
DC_Dupont_Circle
DC_Downtown
DC_Columbia_Heights
DC_Cleveland_Park
DC_Chevy_Chase
DC_Catholic_University-Brookland
DC_Capitol_Hill
DC_Brentwood
DC_Au-Tenleytown
DC_Adams_Morgan
CHI_West_Town
CHI_North_Center
CHI_Near_West_Side
CHI_Near_South_Side
CHI_Near_North_Side
CHI_Loop
CHI_Logan_Square
CHI_Lincoln_Park
CHI_Lake_View
BOS_South_End
BOS_South_Boston
BOS_Mit
BOS_Fenway-Kenmore
BOS_East_Cambridge
BOS_Charlestown
BOS_Central
BOS_Cambridgeport
BOS_Back_Bay-Beacon_Hill
BOS_Allston-Brighton
tension:
Uber Neighborhoods



Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Neurosciences Graduate Program
Halicioglu Data Science Institute

bvoytek@ucsd.edu
@bradleyvoytek

UC San Diego