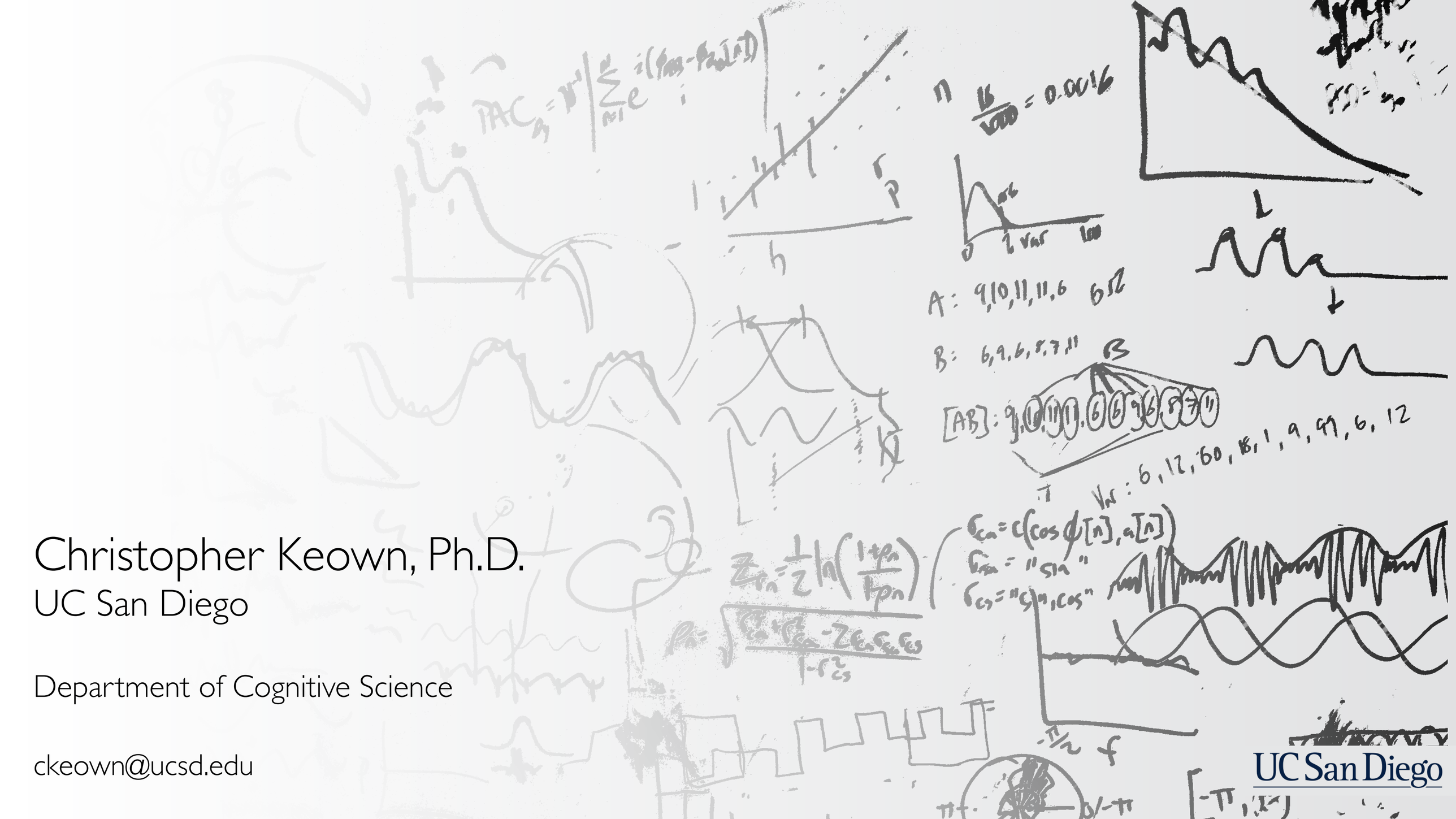


Christopher Keown, Ph.D.

UC San Diego

Department of Cognitive Science

ckeown@ucsd.edu

UC San Diego

COGS 108
Data Science in Practice

```
print("Hello, World!")
```

UC San Diego

Scheduling

Lecture:

TuTh 2:00p-3:20p

Galbraith Hall 242

Professor Office Hours:

Thursdays 4:00-5:00 *and*

by appointment

Prerequisites

- CSE 7 - MATLAB
- CSE 8A or 11 - Java
- COGS 18 - Python

Bottom line: we will assume programming knowledge.

Python will be used for all assignments.

No programming experience?

- *Preferred option* - Take a programming course first. Cogs 18 offered in spring.
- *Can't wait?* - Use online sites like *codecademy.com* or *LearnPython.org*

Waitlist

- Waitlist for this course is huge: >200 people!
- I can't manipulate the list to let anyone in, even if you're a graduating senior, etc.
- Two sections of 108 will be offered in Spring

Section

- Attend the one you registered for if possible
- Attendance is not technically mandatory
- No section during week one

Course links

- Piazza: https://piazza.com/ucsd/winter2019/cogs108/home
- TritonEd
- UCSD Podcast: https://podcast.ucsd.edu/
- GitHub: https://github.com/COGS108/

Setting up your computer

- Tutorial 00
  - Python 3.6 Anaconda distribution
  - Jupyter Notebooks
  - GitHub

General stuff and junk

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

General stuff and junk

- If you email a question, the answer to which is on the syllabus, ***you will not get an email response from us***.

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 24** - Ryan Chesler - SD data hacker and Kaggle guru
- **TBD**
- **TBD**

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 Science in Python, Part II
5. Data gathering, wrangling, and cleaning (How do you find and clean data? (JSON, CSV, XML, SQL, APIs)
6. Jan. 24: Guest lecture – Ryan Chesler to discuss Kaggle
7. Basic data visualization
8. Data privacy, ethics, and HIPAA (anonymization)
9. Data intuition and the "sniff test" (Fermi estimation; distributions and outliers: histograms, CDF, PDFs)
10. Geospatial analysis
11. Linear modeling or non-parametric?
12. OLS (optimization) and multiple linear regression and collinearities
13. Model validation (bootstrapping, resampling, k-fold, leave-p-out, train/test)
14. Dimensionality reduction (PCA); clustering and classification (k-means, knn, SVM)
15. Feature selection
16. NLP and text-mining (bag of words, tf-idf, sentiment analysis)

Final Project!

- Project due on final exam day *in lieu of* an actual exam!
- Deadline: 23:59, Thu, March 21, 2019

General stuff and junk

- First quarter?
-
-
-

UC San Diego

General stuff and junk

- First quarter?
- Data Science majors?
-
-

General stuff and junk

- First quarter?
- Data Science majors?
- You're paying for this.
-

General stuff and junk

- First quarter?
- Data Science majors?
- You're paying for this.
- As an adult, what you do with your time is up to you.

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:
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

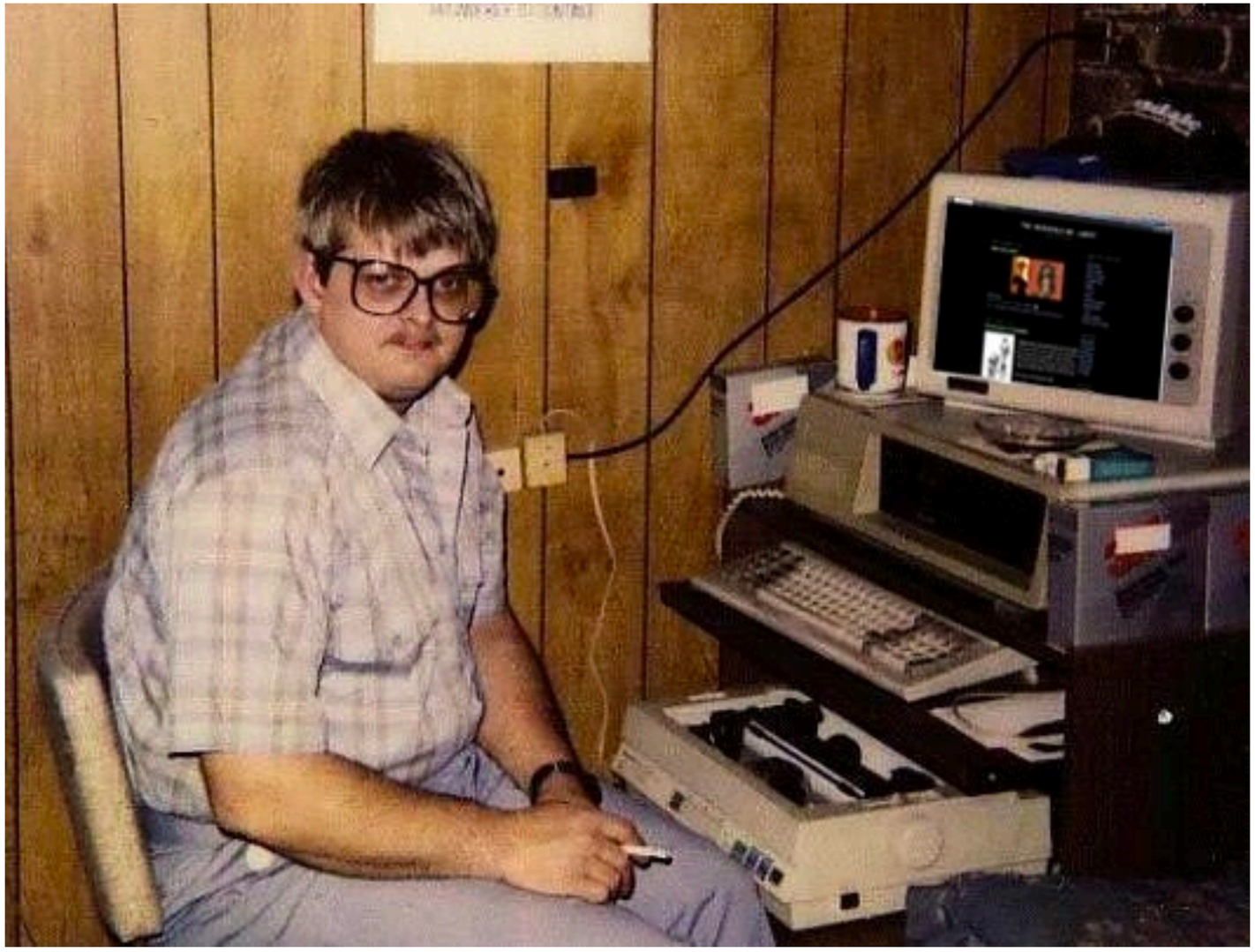
FROM THE OCTOBER 2012 ISSUE

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 still woefully short on science

Lacking rigor

“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

A New Field Emerges

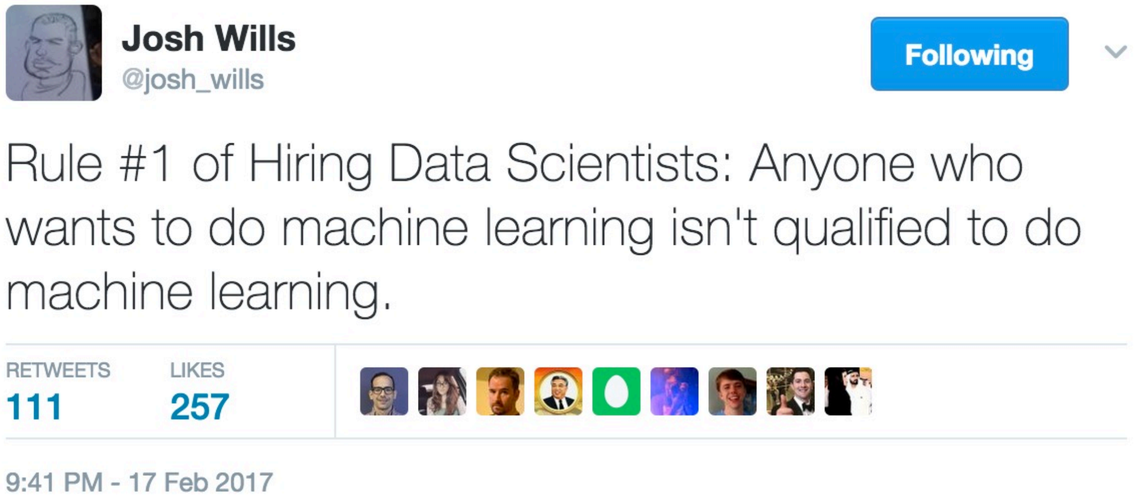
There is significant and growing demand for data-savvy professionals in businesses, public agencies, and nonprofits. The supply of professionals who can work effectively with data at scale is limited, and is reflected by rapidly rising salaries for d , and data analysts.

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

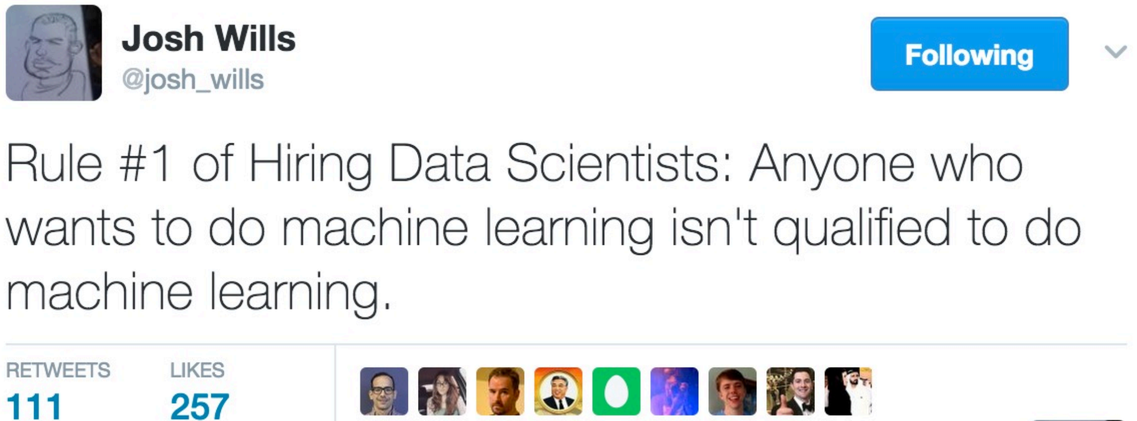
UC San Diego

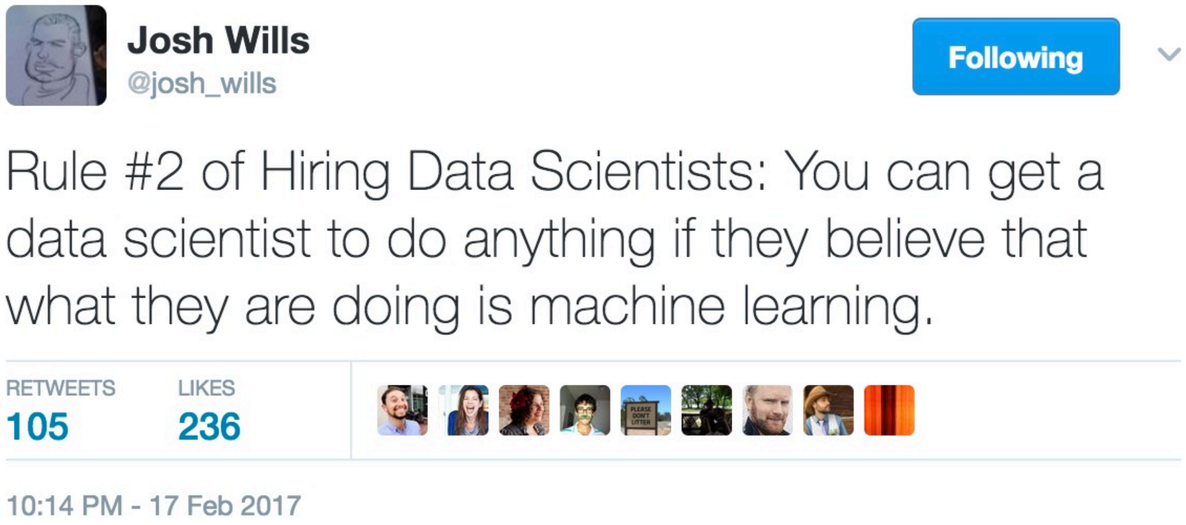
Source: BIDS

What *isn't* Data Science?

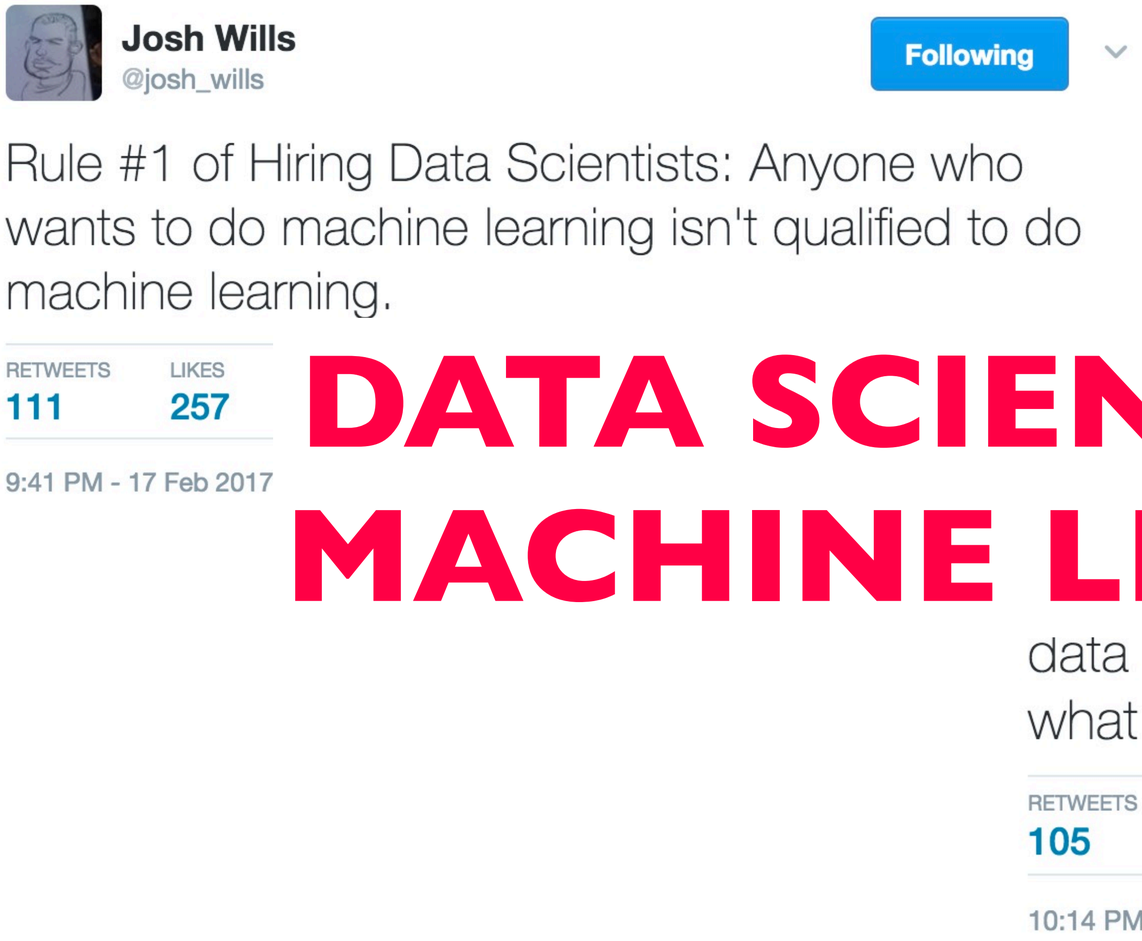


What *isn't* Data Science?

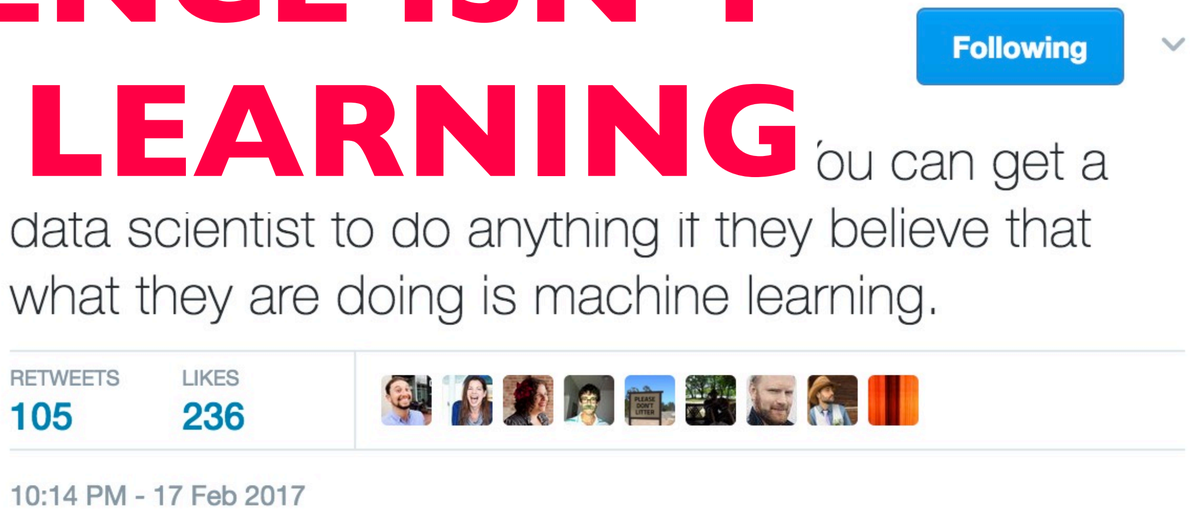




What *isn't* Data Science?



DATA SCIENCE ISN'T MACHINE LEARNING



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.

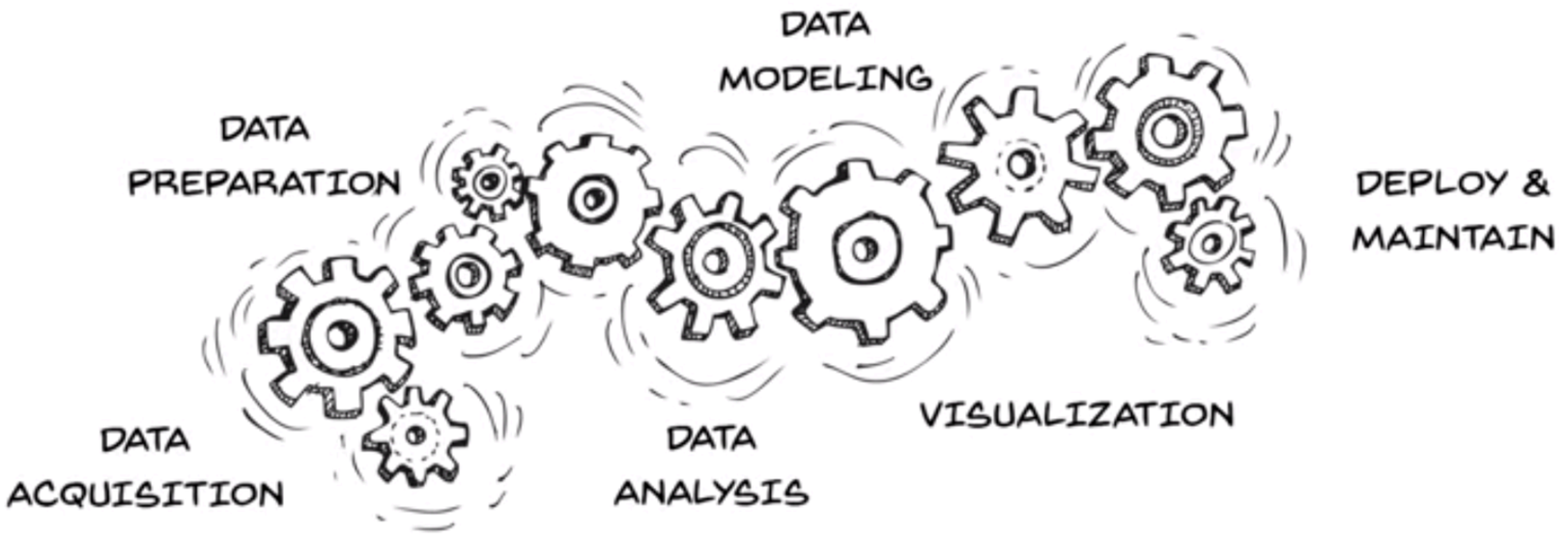
2) What makes "good" versus "bad" quality data for different questions, etc.

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).



WE WILL TELL YOU

HOW DOES IT REALLY WORK UNDER THE HOOD!

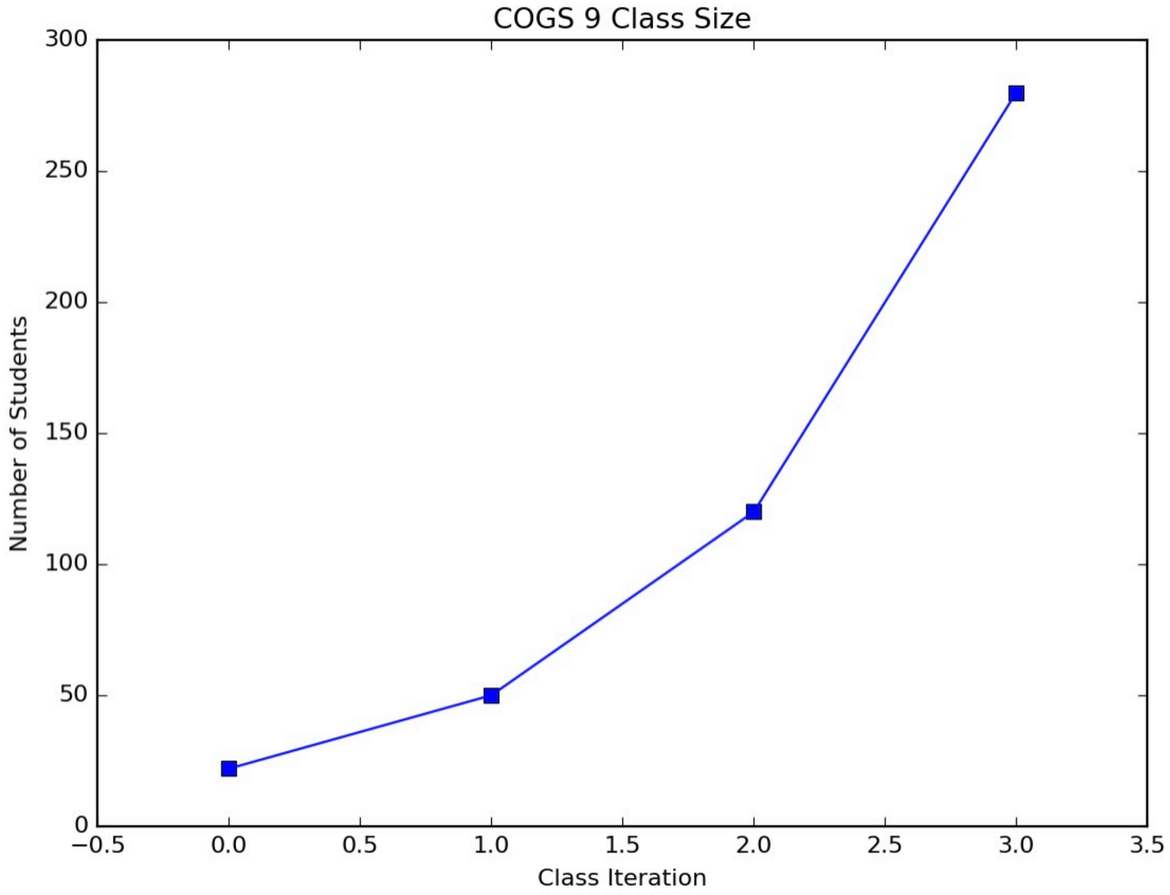
Data Science at UCSD

Data Science at UCSD

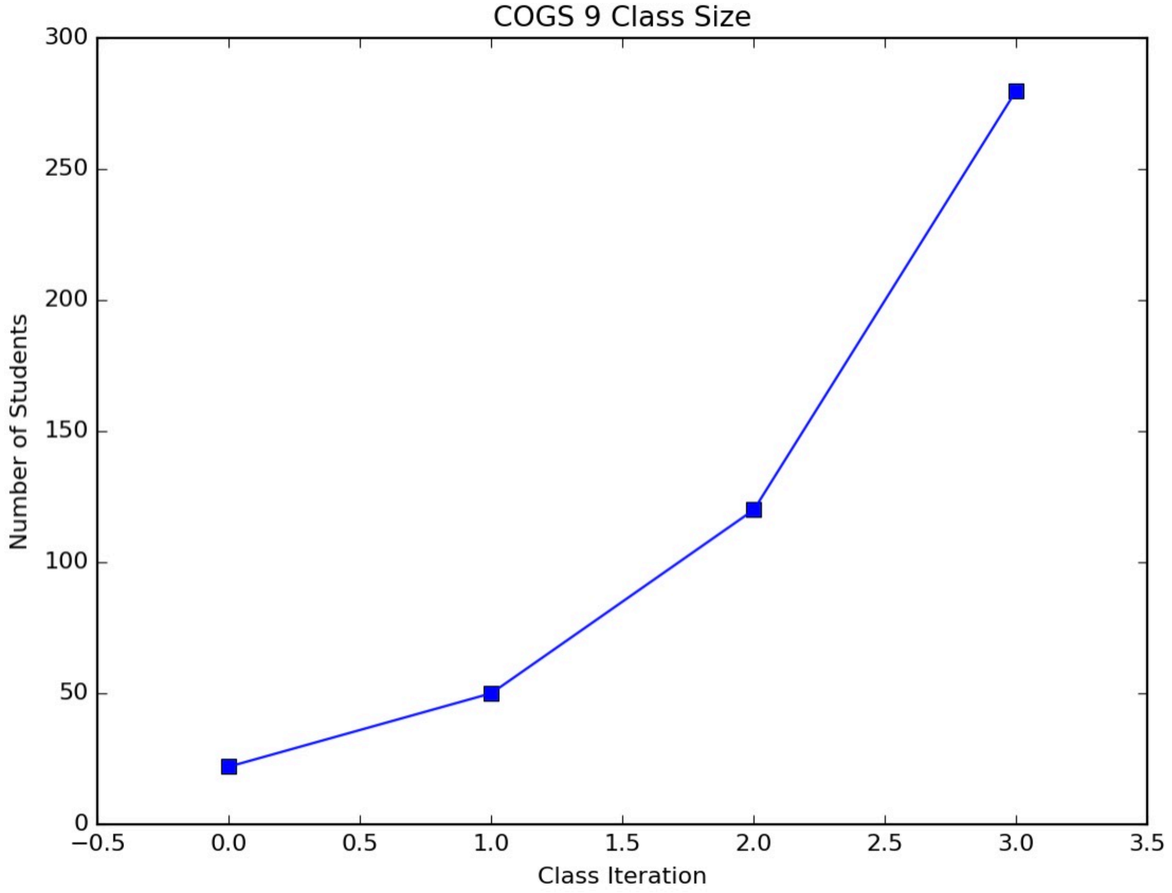
Facebook pioneer donates $75 million to UCSD for data science

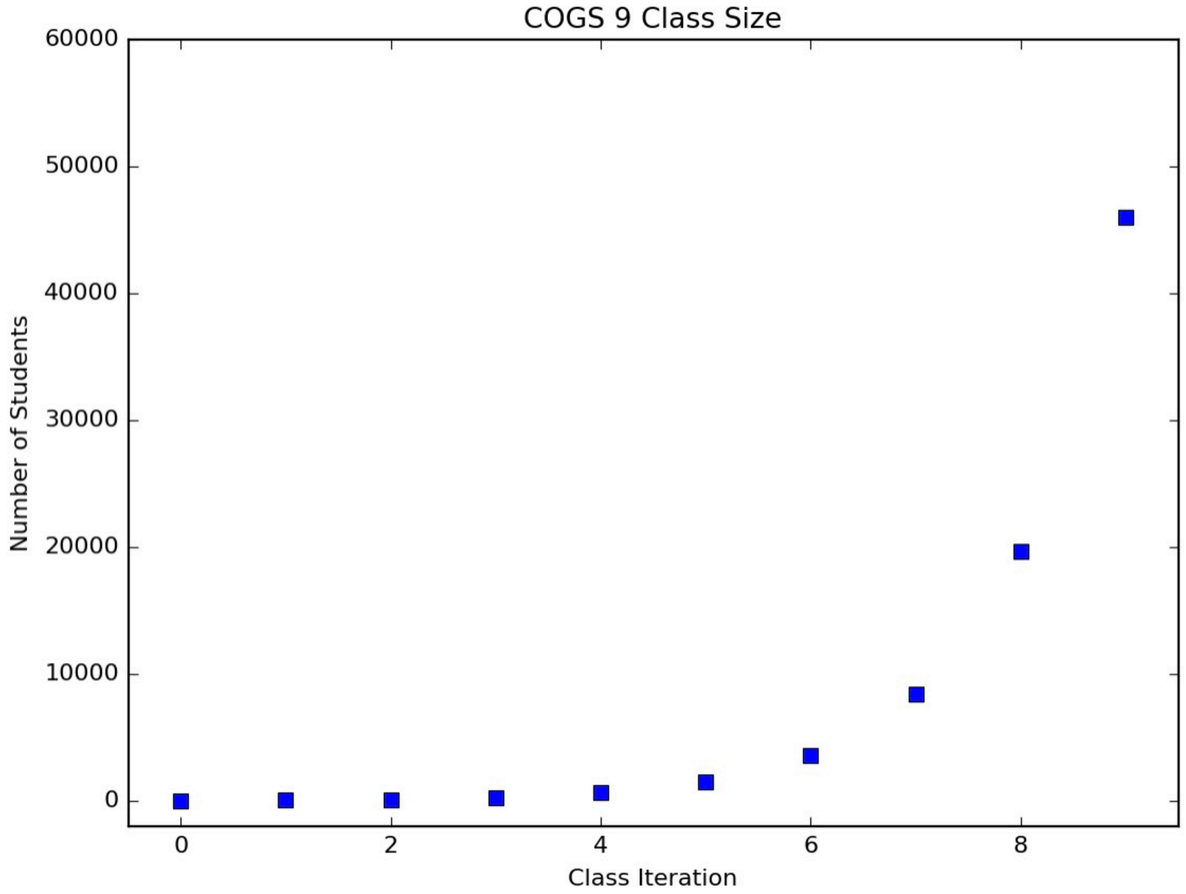
Taner Halicioglu is donating $75 million to UC San Diego to make his alma mater a national leader in data science. (Erik Jepsen / UC San Diego)

Data Science at UCSD



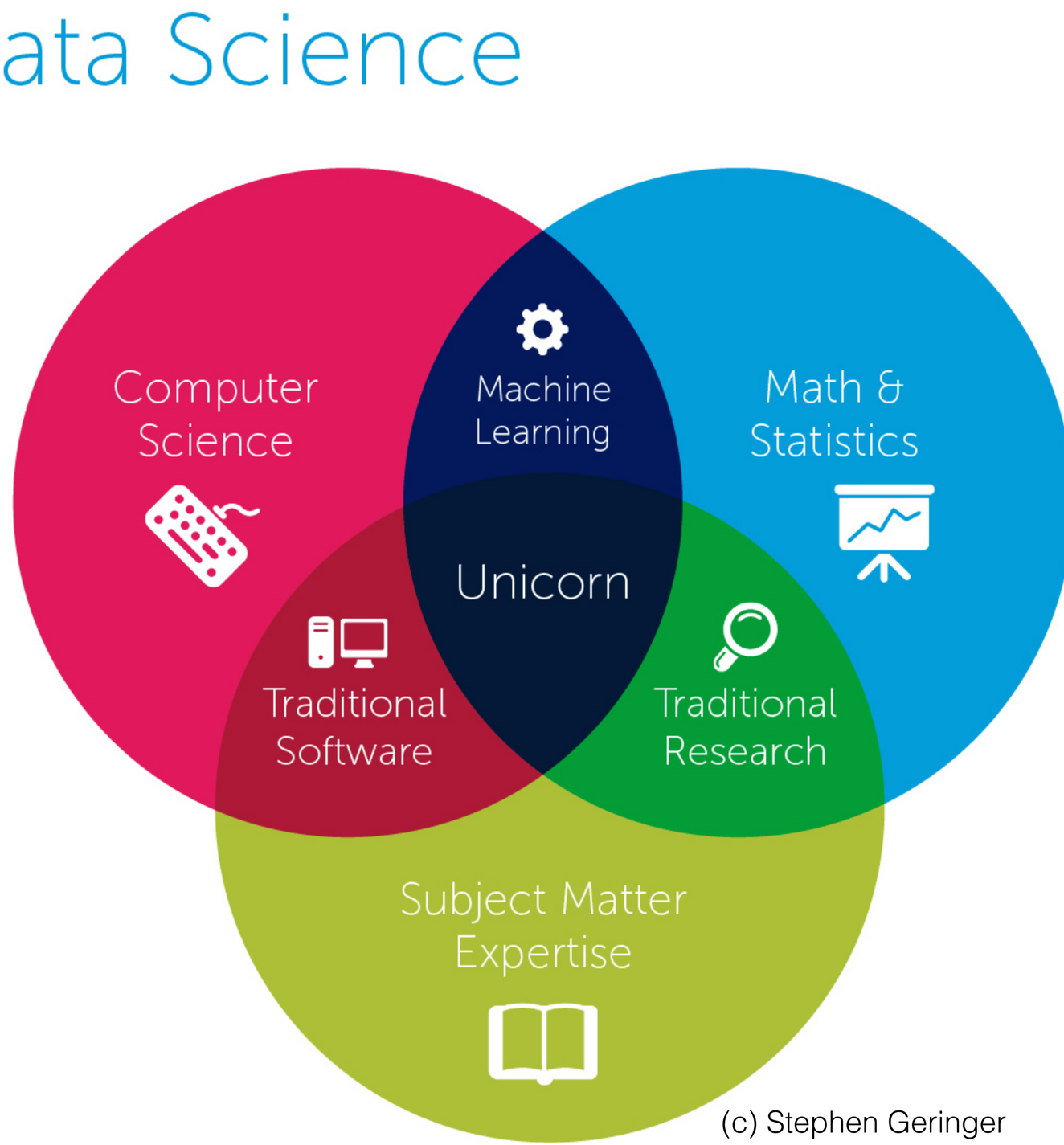
Data Science at UCSD





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
- How humans interact with data and how information and prediction is extracted from data.

UC San Diego

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?

COGS 108 - Winter 2019 Intro Questionnaire

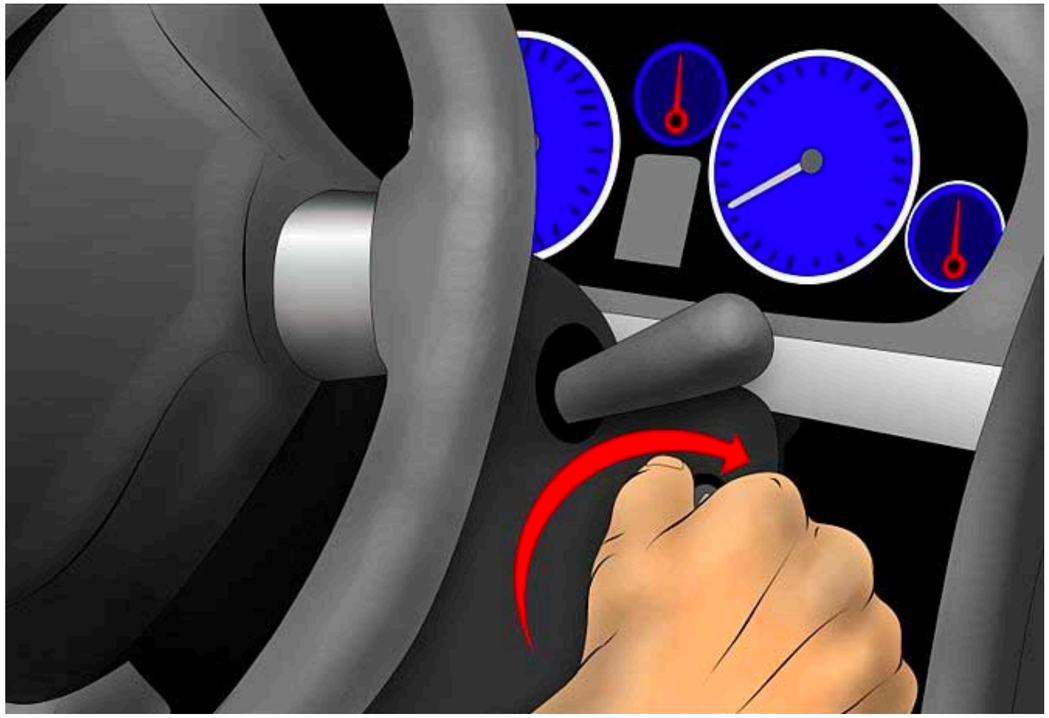
Wait… Who are ***you****?*

https://bit.ly/2Tormjj

Who is this guy?

How to say my last name

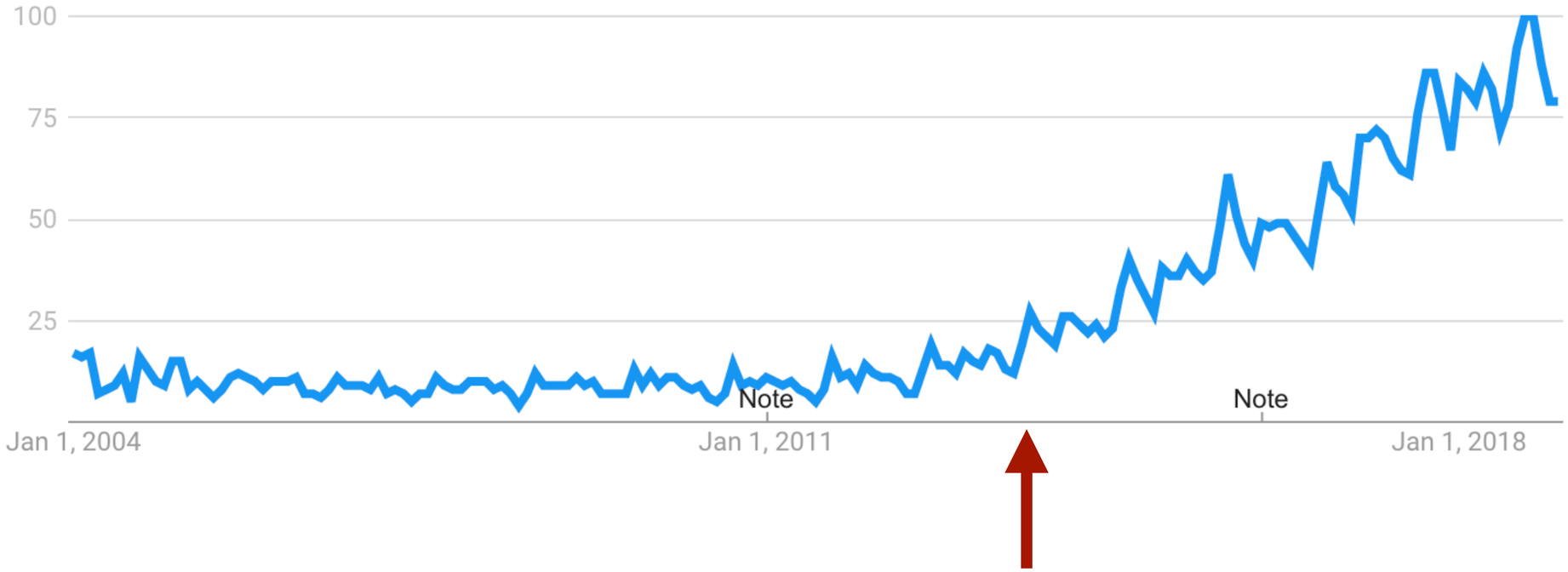
Keown - like *in turn the KEY ON*



Background

- Undergraduate in computer science
- Masters in computational science
- PhD in cognitive science

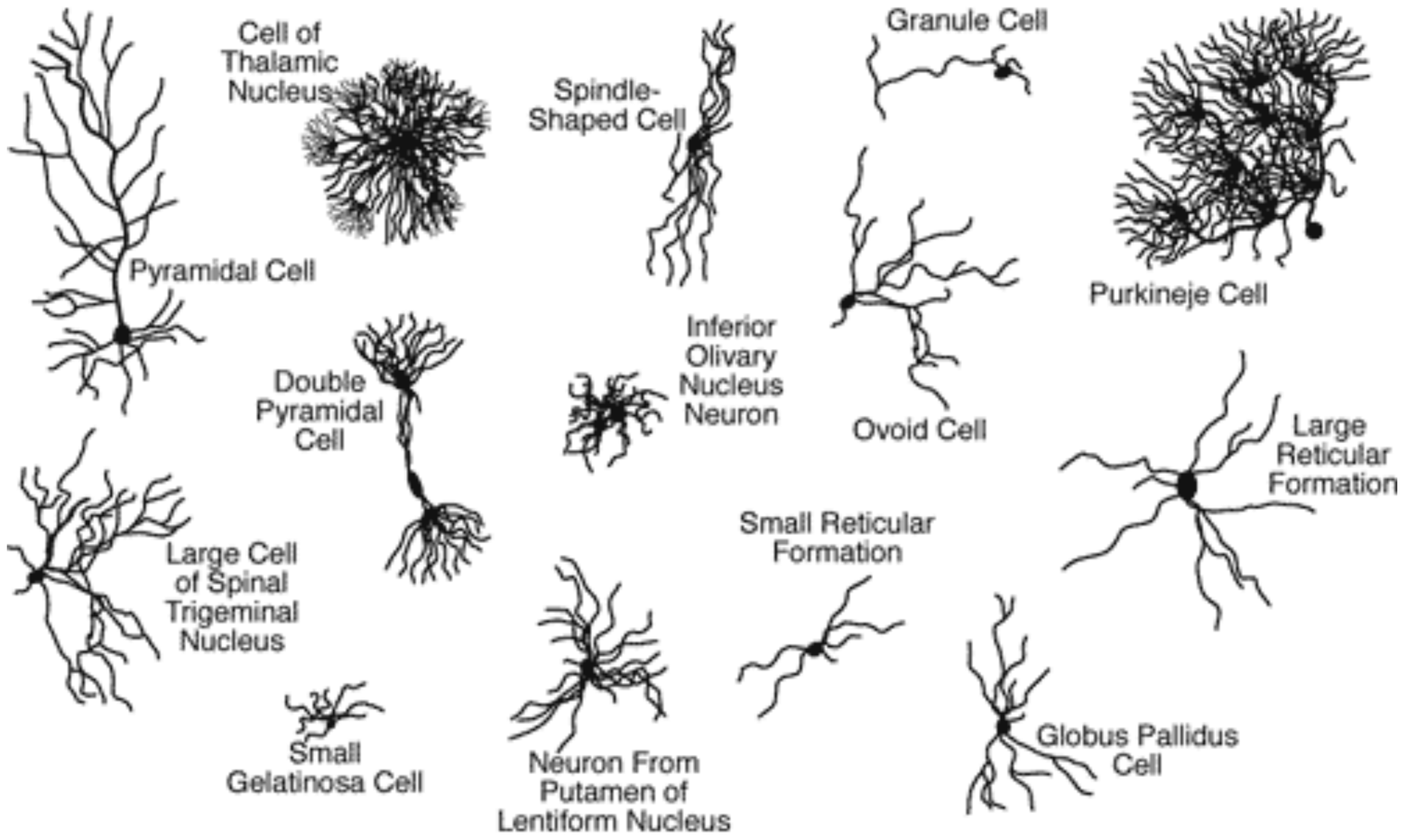
Interest in Data Science - Google Trends



Started my PhD

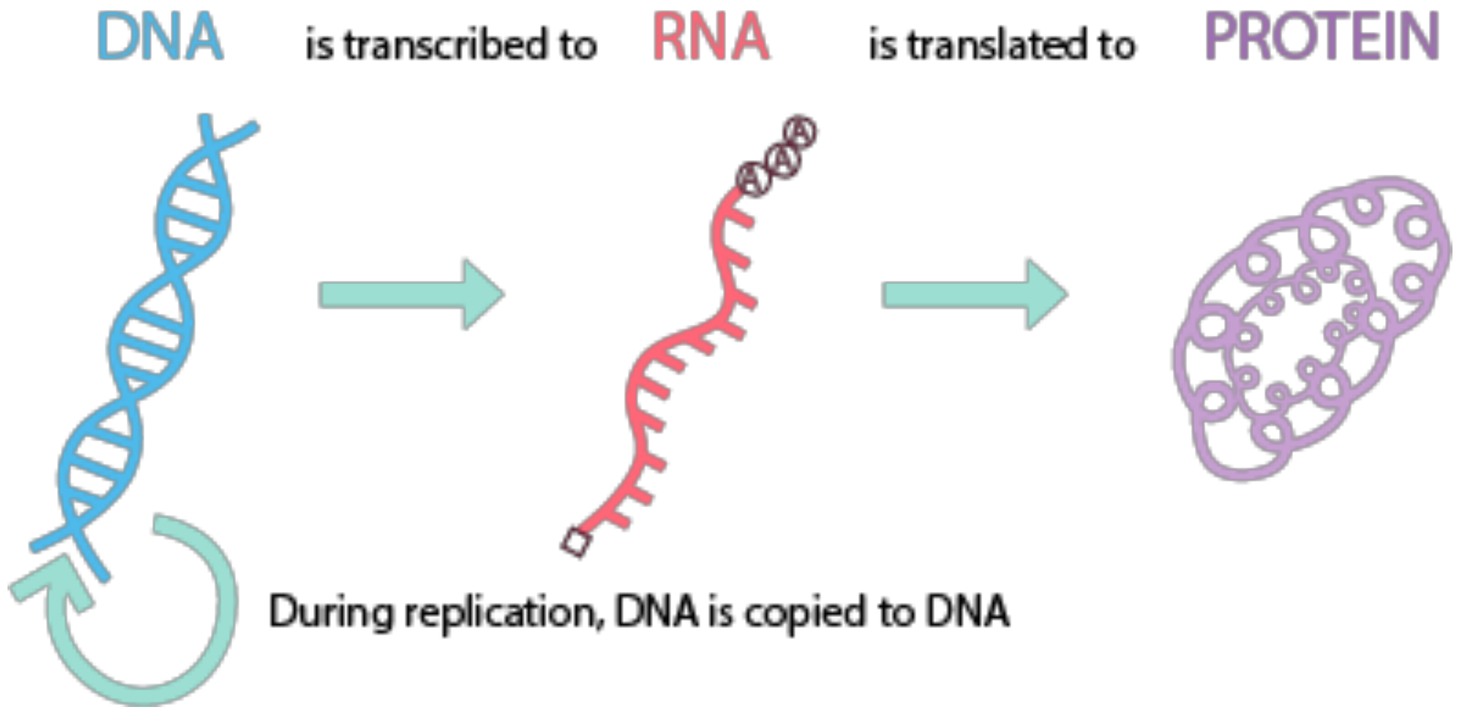
PhD in epigenetics from UCSD Cog Sci

WTH is Epigenetics?



But all cells have the same genome!

Epigenetics



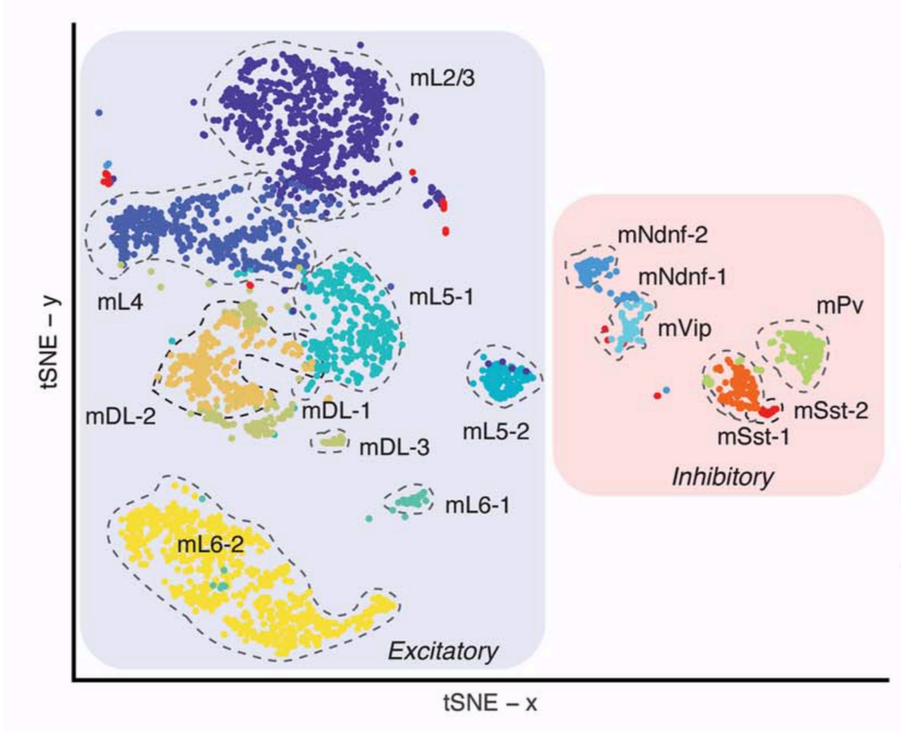
Layer "above" the genome that allows for genes to be turned on and off in individual cells, thereby creating cell type heterogeneity.

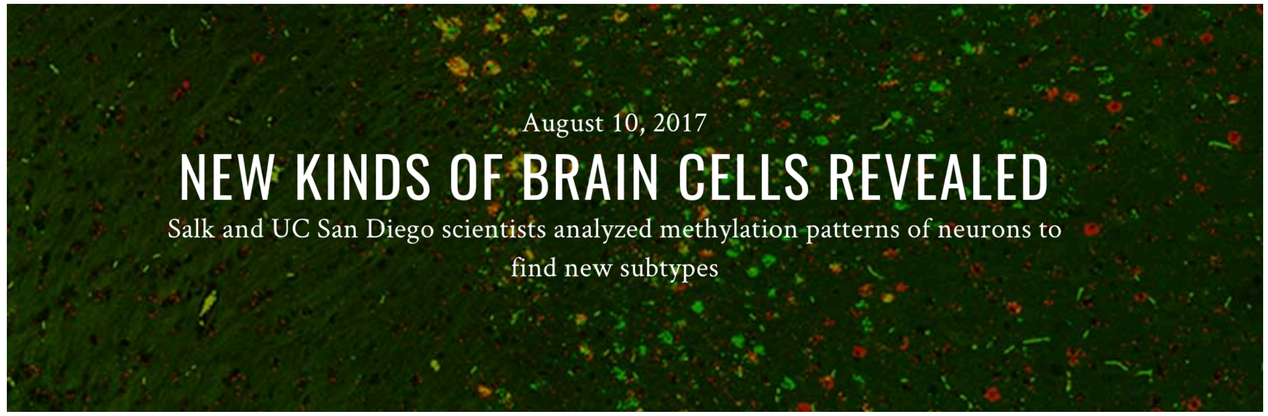
Data Science

~3 billion bases in genome—per cell!
Difficult to process, clean, visualize and identify patterns

How many neuron types are there?

Measure epigenome in many individual cells and cluster them





Entrepreneur

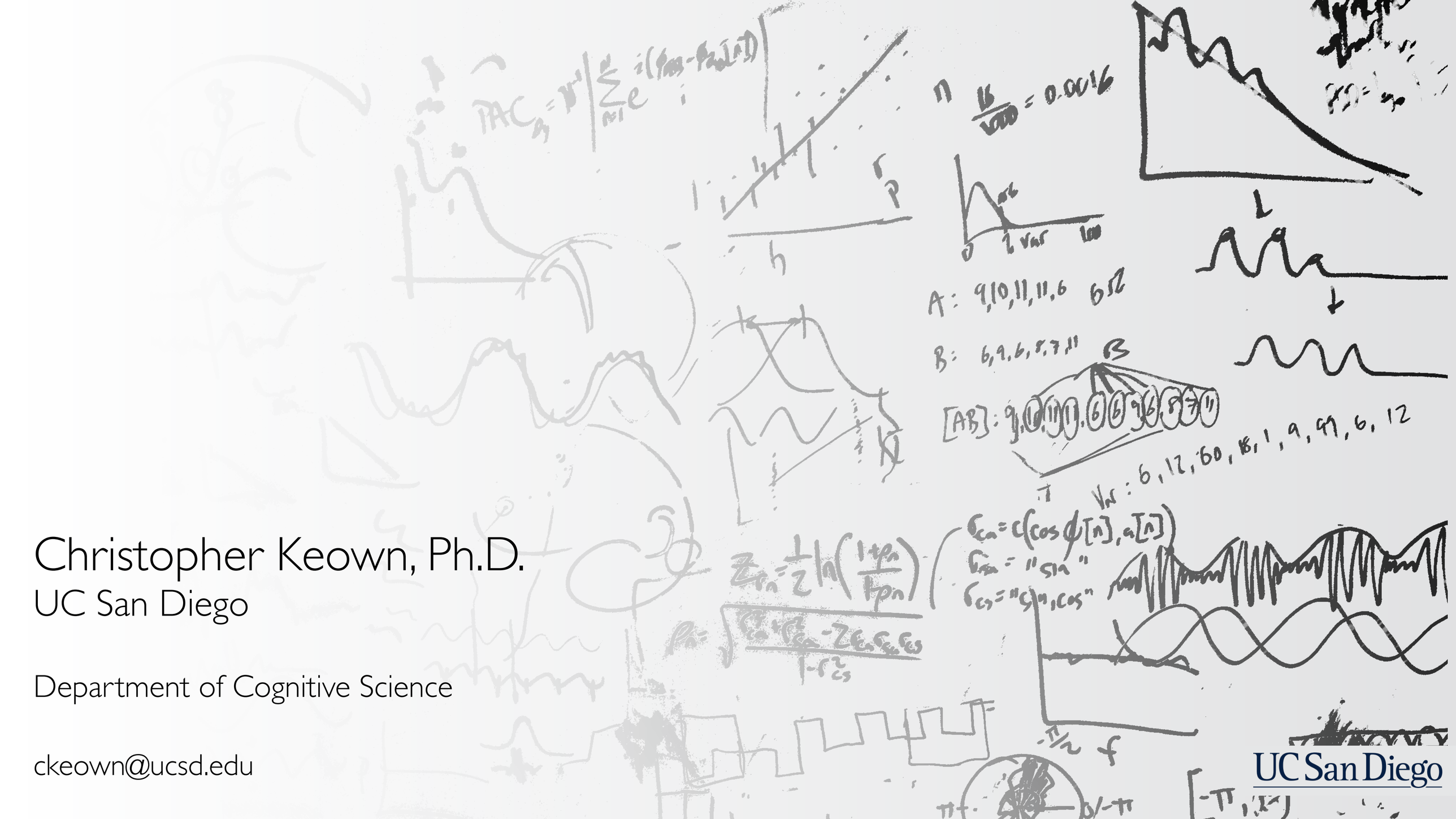
How can scientific research be more effective?

Build tools to make
scientific data more accessible
and more useful

Rose Hendricks

Nick Peterson

San Diego Machine Learning



Christopher Keown, Ph.D.

UC San Diego

Department of Cognitive Science

ckeown@ucsd.edu

UC San Diego