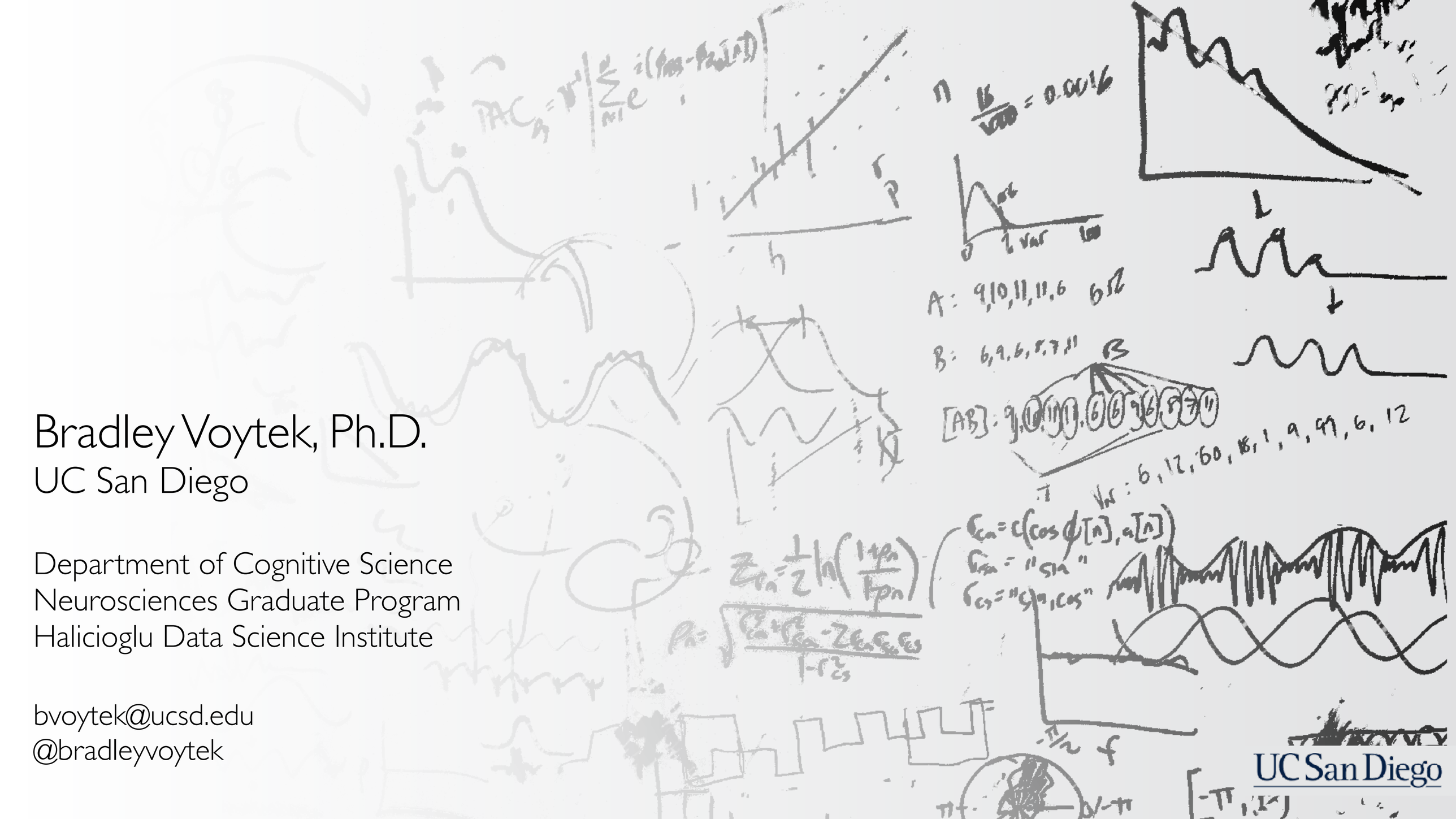


Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Neurosciences Graduate Program
Halicioglu Data Science Institute

bvoytek@ucsd.edu
@bradleyvoytek

UC San Diego

Administrative stuff

- **Research opportunities**
  - Email professors!
  - Do a 99/199 (course credits in a lab)
  - Professors often ignore/don't respond
  - So look at their faculty web pages and email their students/staff!

Administrative stuff

- **REQUIRED LECTURE ANNOUNCEMENTS!**
  - İlkay Altıntaş: Chief Data Science Officer, San Diego Supercomputer Center
  - 2018 February 13 (Tuesday)

  - Special guest!
  - 2018 February 20 (Tuesday)

- **NO CLASS (don't cheer)**
  - 2018 Feb 15 (Thursday)

Administrative stuff

A1: Setting-Up

- Released: Monday, January 15th (W2)
- Due: Sunday, January 21st (W2)

A2: Data Exploration

- Released: Monday, January 22nd (W3)
- Due: Sunday, February 4th @ 11:59 pm (W4)

A3: Data Privacy

- Released: Monday, February 5th (W5)
- Due: Sunday, February 11th @ 11:59 pm (W5)

A4: Data Analysis

- Released: Monday, February 12th (W6)
- Due: Sunday, February 25th @ 11:59 pm (W7)

A5: Name TDB

- Released: Monday, February 26th (W8)
- Due: Sunday, March 4th @ 11:59 pm (W9)

Project Schedule

Intermixed with this schedule is your final project.

Project Proposal

Due Sunday, February 18th @ 11:59 pm (W6)

Project Check In

Due Thursday, March 1st (W8)

Final Project

Due Thursday, March 22nd @ 11:59 pm (Finals Week)

Source: https://github.com/COGS108/Assignments

Administrative stuff

The broad objectives for the project are to:

- Identify the problems and goals of a **real** situation and dataset.
- Choose an appropriate approach for formalizing and testing the problems and goals, and be able to articulate the reasoning for that selection.
- Implement your analysis choices on the dataset.
- Interpret the results of the analyses.
- Contextualize those results within a greater scientific and social context, acknowledging and addressing any potential issues related to privacy and ethics.
- Work effectively to manage a project as part of a team.

The basic project steps:

- Find a real world dataset and problem that you believe can be solved with one or more of the techniques we have learned in class.
- After selecting a dataset and identifying the goal, write out a proposed analysis plan and submit it through TritonEd for review (due Sunday, Feb 18).
- Apply the techniques outlined and come up with a result for the dataset that you proposed.
- Assemble a Jupyter notebook that communicates your hypothesis, methods, and results (this is the final product due Thursday, March 22).

UC San Diego

Source: https://github.com/COGS108/Projects/blob/master/CapstoneProject.pdf

Administrative stuff

Project Proposal - Detailed Description

For the Project Proposal you need to write a report, in the style outlined below, about how you might approach your question of interest. Specifically, every Report must contain seven sections, briefly outlined here, with more specific direction provided in the proposal template notebook:

1) Research Question: What's your question?
2) Hypothesis: What's your prediction?
3) Dataset(s): What data will you use to answer your question? Describe your dataset(s).
4) Background: Why is this question of interest, what background information led you to your hypothesis, and why is this important?
5) Proposed Methods: What methods will you use to analyze your data?
6) Ethics: Acknowledge and address any potential ethics and privacy issues related to your project.
7) Discussion: Discuss the potential impact of your project, as well as trying to anticipate any problems you may encounter.

Source: https://github.com/COGS108/Projects/blob/master/CapstoneProject.pdf

Administrative stuff

UC San Diego Halicioglu Data Science Institute Launch Event

The *special objectives* for the *optional* UC San Diego Halicioglu Data Science Institute launch event are to:

- Communicate your results effectively to both experts and laypersons.
- Use data scientific approaches to address questions *specifically concerning civic utility and social good*.

A panel of local Data Science experts from the university, government, and industry will evaluate 4-8 projects, selected by Prof. Voytek for their potential for addressing critical questions of civic utility and/or social good.

These Projects *need not be the complete and final project you will submit for grading*, however they do need to be relatively thorough and complete to be considered for presentation on the afternoon of the launch event.

Deadline: To be considered eligible for presenting at this event, you will need to submit your Project Notebook by Sunday, Feb 25 at 23:59.

This *optional* submission, to be considered for the event, should follow the same outline and rubric as above for the final project notebook. You must have preliminary results, but it can be be a work-in-progress (for example, discussion section and conclusions need not necessarily be fleshed out).

UC San Diego

Source: https://github.com/COGS108/Projects/blob/master/CapstoneProject.pdf

Administrative stuff

- **Available Data**
  - https://data.sandiego.gov/datasets/
  - https://data.sandiego.gov/resources/

- **Project Ideas**
  - ***Bike Friendly Miles:*** Using publicly available data from City of San Diego, Sandag, and Open Street Map, try to estimate the amount of bike-friendly miles in the City and how that has changed over the years.
  - ***Track Increase / Decrease of Bicycle Ridership:*** Using publicly available data sources, track the levels of bicycle ridership, the trends over the years, the neighborhoods where it's the most prevalent, time of day, and why do you think that is. Can you predict where bike lanes should be improved (e.g., from a shared to a protected lane) or added based on use/predicted use?

COGS 108
Data Science in Practice

HOW DO YOU DO STUFF?!

UC San Diego

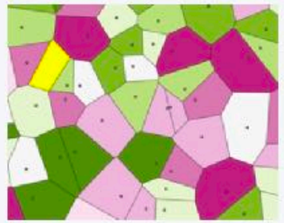
d3.js

Sunburst

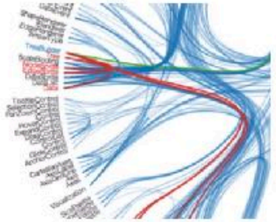
Node-Link Tree

Treemap

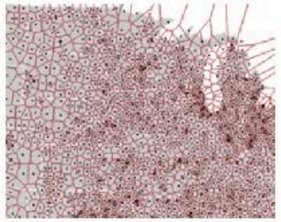
Voronoi Diagram



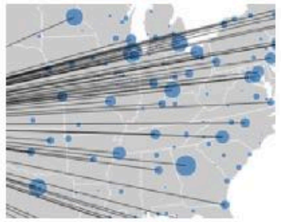
Hierarchical Edge Bundling



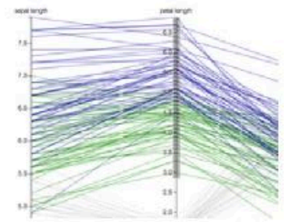
Voronoi Diagram



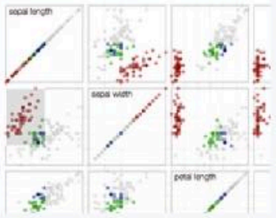
Symbol Map



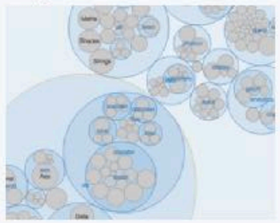
Parallel Coordinates



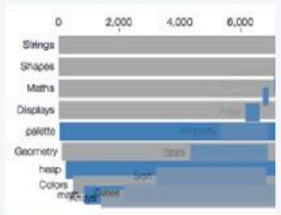
Scatterplot Matrix



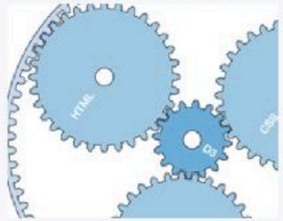
Zoomable Pack Layout



Hierarchical Bars

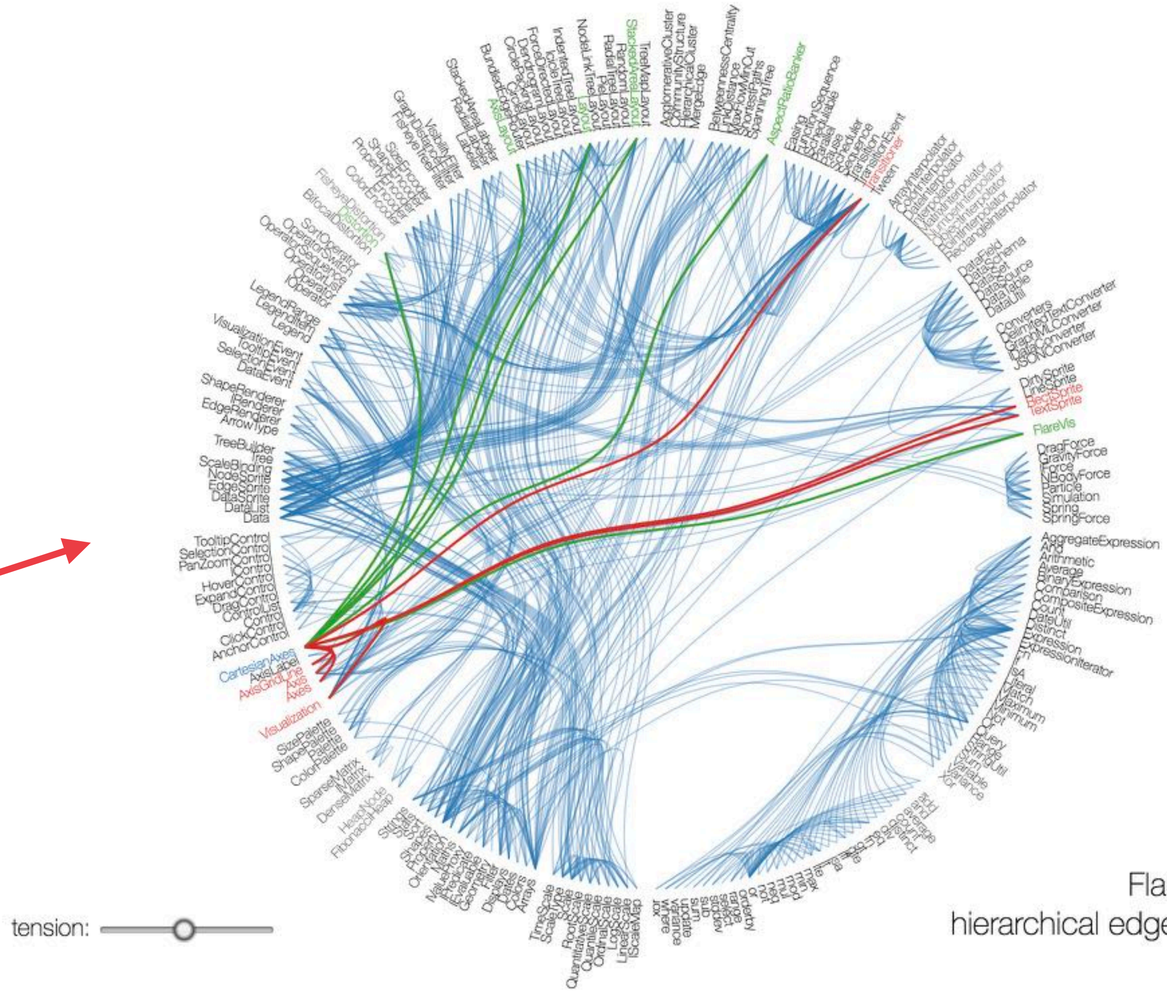


Epicyclical Gears



Source: https://d3js.org/

d3.js

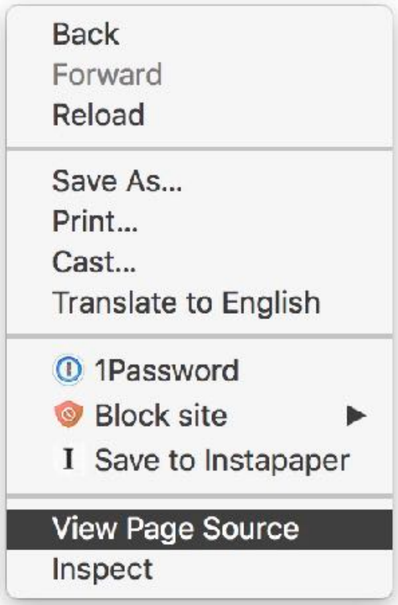


Source: https://d3js.org/

UC San Diego

d3.js

view-source:http://mbostock.github.io/d3/talk/20111116/bundle.html



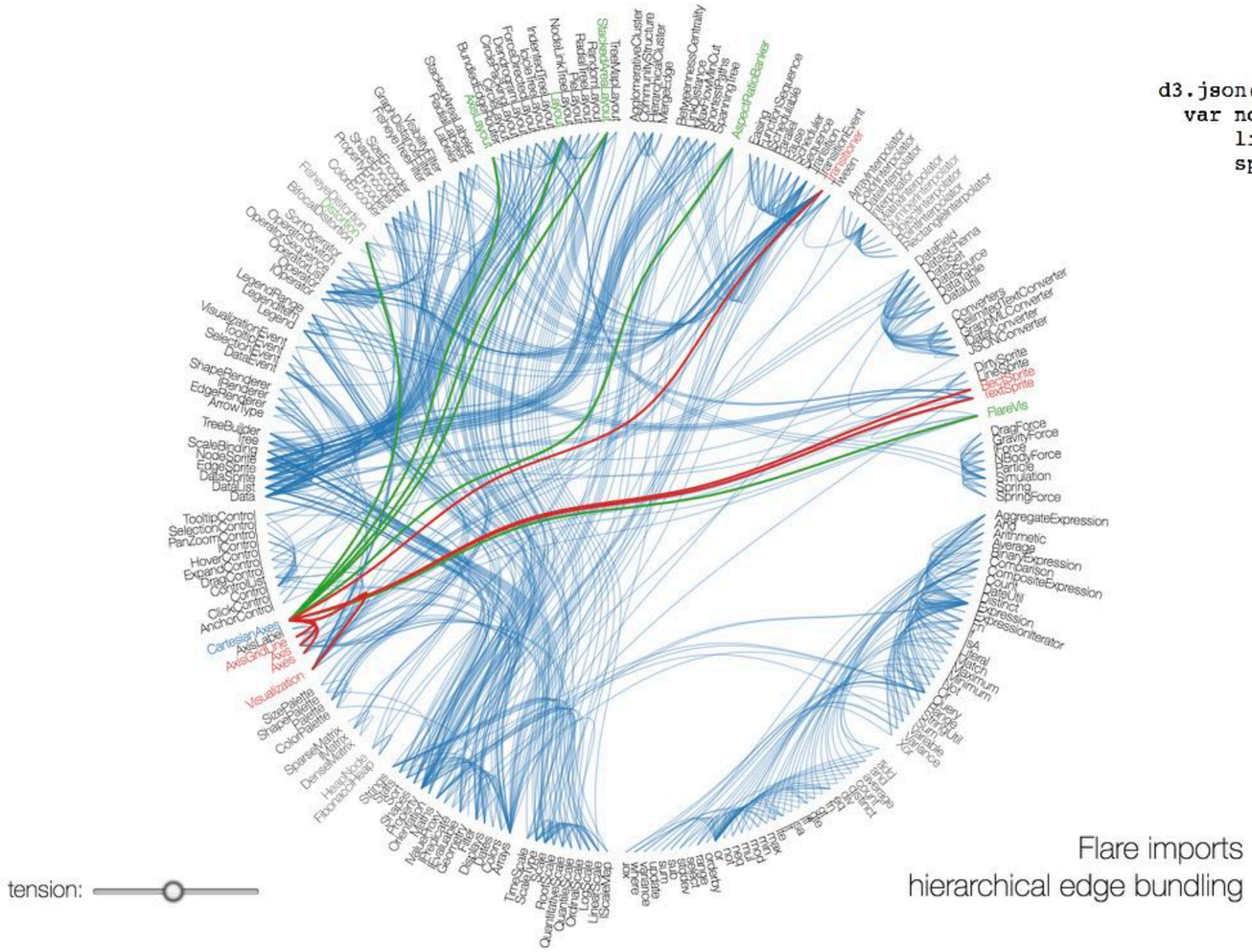

```
  </style>
</head>
<body>
  <h2>
    Flare imports<br>
    hierarchical edge bundling
  </h2>
  <div style="position:absolute;bottom:0;font-size:18px;">tension: <input style="position:relative;top:3px;" type="range" min="0" max="100" value="85"></div>
  <script type="text/javascript" src="d3/d3.js"></script>
  <script type="text/javascript" src="d3/d3.layout.js"></script>
  <script type="text/javascript" src="packages.js"></script>
  <script type="text/javascript">



d3.json("flare-imports.json", function(classes) {
  var nodes = cluster.nodes(packages.root(classes)),
      links = packages.imports(nodes),
      splines = bundle(links);
```

Source: https://d3js.org/

d3.js




```
d3.json("flare-imports.json", function(classes) {
  var nodes = cluster.nodes(packages.root(classes)),
      links = packages.imports(nodes),
      splines = bundle(links);
```

view-source:http://mbostock.github.io/d3/talk/20111116/bundle.html

view-source:mbostock.github.io/d3/talk/20111116/flare-imports.json


```
{"name":"flare.vis.axis.CartesianAxes","size":6703,"imports":
["flare.animate.Transitioner","flare.display.RectSprite","flare.vis.axis.Axis","fl
are.display.TextSprite","flare.vis.axis.Axes","flare.vis.Visualization","flare.vis
.axis.AxisGridLine"]},
{"name":"flare.vis.controls.AnchorControl","size":2138,"imports":
```


UC San Diego

Source: https://d3js.org/

Real data crunching

Test No. 8

33. Accrued depreciation is a term used in the real estate appraisal field. Which of the following methods would accrued depreciation have its greatest effect on?

a. Market data approach
b. Cost approach
c. Capitalization of net income approach
d. Sales comparison approach

Ans.(b):

34. What transfers less than an entire leasehold, with the original lessee being primarily liable for the rental agreement?

a. Sublease
b. Sandwich lease
c. Residential lease
d. Assignment

Ans.(a):

- PDF of 10 tests, 200 questions each

Real data crunching

Test No. 8

33. Accrued depreciation is a term used in the real estate appraisal field. Which of the following methods would accrued depreciation have its greatest effect on?

a. Market data approach
b. Cost approach
c. Capitalization of net income approach
d. Sales comparison approach

Ans.(b):

34. What transfers less than an entire leasehold, with the original lessee being primarily liable for the rental agreement?

a. Sublease
b. Sandwich lease
c. Residential lease
d. Assignment

Ans.(a):

- PDF of 10 tests, 200 questions each
- How many unique questions?

Real data crunching

Test No. 8

33. Accrued depreciation is a term used in the real estate appraisal field. Which of the following methods would accrued depreciation have its greatest effect on?

a. Market data approach
b. Cost approach
c. Capitalization of net income approach
d. Sales comparison approach

Ans.(b):

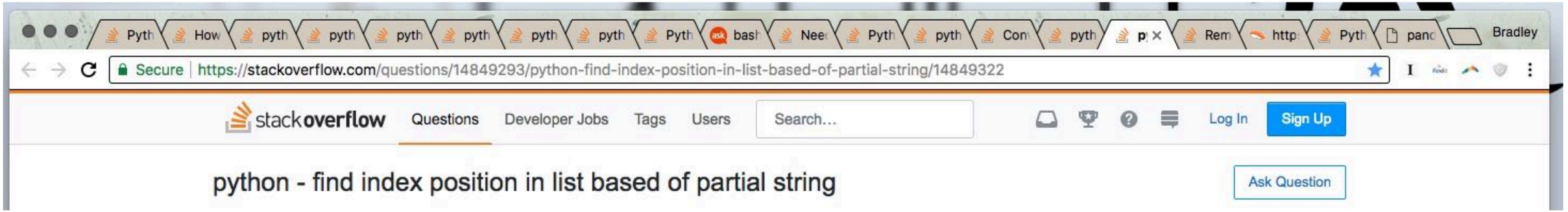
34. What transfers less than an entire leasehold, with the original lessee being primarily liable for the rental agreement?

a. Sublease
b. Sandwich lease
c. Residential lease
d. Assignment

Ans.(a):

- PDF of 10 tests, 200 questions each
- How many unique questions?
- Which ones are duplicated, and how often?

Real data crunching



Real data crunching

```
In [2]: filename = 'exam.pdf'
        pdf = PyPDF2.PdfFileReader(open(filename, 'rb'))
        total_pages = pdf.getNumPages()
        total_pages

Out[2]: 500
```

Real data crunching

```
In [3]: text_data = ''
        for i in range(0, total_pages):
            pageObj = pdf.getPage(i)
            text_data = text_data + pageObj.extractText()
        text_data

Out[3]: '     Test No. 1, page \n1       1. Any \ncomplaint as to the violat\nion of the United States Civil Rights\n Act of\
        n 1968 should be filed within how many days of its occurrence?\n   a. 180 days\n  b. 365 days\n  c. 30 days\n  d. 90
        days\n  Ans.(b): \nAn aggrieved person may file a\n complaint directly to a U.S. District\n Court within one year of
        the al\nleged discriminatory practice, whether or not a\n verified complaint has been filed with the Secretary of HUD
        .\n    2. A fee \nsimple e\nstate\n is mo\nst likely to be a:\n   a. life\n estate.\n  b. leasehold.\n  c. less\n-tha
```

- How do I parse this into "questions"?

Real data crunching

```
In [4]: new_data = text_data.replace('\n','')
        new_data = new_data.replace('->',' ')
        new_data

Out[4]: '     Test No. 1, page 1       1. Any complaint as to the violation of the United States Civil Rights Act of 1968 sho
        uld be filed within how many days of its occurrence?   a. 180 days  b. 365 days  c. 30 days  d. 90 days  Ans.(b): An
        aggrieved person may file a complaint directly to a U.S. District Court within one year of the alleged discriminatory
        practice, whether or not a verified complaint has been filed with the Secretary of HUD.    2. A fee simple estate is
        most likely to be a:   a. life estate.  b. leasehold.  c. less-than-freehold estate.  d. maximum interest obtainable.
```

- How do I parse this into "questions"?
- Clean it up first…

Real data crunching

In [5]:
```
new_data = text_data.replace('\n','')
new_data = new_data.replace('->',' ')

questions = 200
tests = 10
total_questions = questions * tests

sentence = []

i = 1
j = 0

while len(sentence) < total_questions:
    my_regex = '\ \ ' + str(i) + '\.\ (.*?)a\.\ '
    result = re.search(my_regex, new_data)
    result.group()

    sentence.append(new_data[(result.span()[0]+len(str(i))+4):(result.span()[1]-len(str(i))-2)])
    sentence[j] = sentence[j].strip()

    new_data = new_data[(result.span()[1]-len(str(i))-4):]

    i += 1
    if i>questions:
        i = 1
    j += 1
```

- Extract sentences

Real data crunching

```
In [8]: len(sentence), sentence

Out[8]: (2000,
         ['Any complaint as to the violation of the United States Civil Rights Act of 1968 should be filed within how many da
        ys of its occurrence?',
          'A fee simple estate is most likely to be a:',
          'Mr. Jones, an owner of a packaging firm, purchased a new machine in 2008 and paid $5,500. It was estimated at the
        time of the purchase to have a total economic life of 10 years and a salvage value of $550. Using the straight line m
        ethod of depreciation, the book value at the end of 7 years would be:',
```

- How do I parse this into "questions"?
- Clean it up first…
- Did it work?

Real data crunching

In [9]:
```
import pandas as pd
import numpy as np

df = pd.DataFrame({'id': np.arange(len(sentence))+1, 'questions': sentence})

df['similarity'] = ''
df['indices'] = ''

similarity = []
indices = []


for i in range(0, len(sentence)):
    similarity.append([])
    indices.append([])

    for j in range(0, len(sentence)):
        l_val = Levenshtein.ratio(sentence[i], sentence[j])

        if i != j:
            if l_val >= 0.75:
                similarity[i].append(l_val)
                indices[i].append(j)

    df.at[i, 'similarity'] = similarity[i]
    df.at[i, 'indices'] = indices[i]
```

- Calculate similarity

Real data crunching

```
In [53]: i = 0
         similarity[i], indices[i], sentence[i], sentence[indices[i][0]], sentence[indices[i][1]], sentence[indices[i][2]]

Out[53]: ([1.0, 1.0, 1.0],
          [847, 1304, 1748],
          'Any complaint as to the violation of the United States Civil Rights Act of 1968 should be filed within how many day
         s of its occurrence?',
          'Any complaint as to the violation of the United States Civil Rights Act of 1968 should be filed within how many day
         s of its occurrence?',
          'Any complaint as to the violation of the United States Civil Rights Act of 1968 should be filed within how many day
         s of its occurrence?',
          'Any complaint as to the violation of the United States Civil Rights Act of 1968 should be filed within how many day
         s of its occurrence?')
```

- How do I parse this into "questions"?
- Clean it up first…
- Did it work?
- What is a "unique" question?

Real data crunching

```
In [15]: i = 3
similarity[i], indices[i], sentence[i], sentence[indices[i][0]], sentence[indices[i][1]], sentence[indices[i][2]]

Out[15]: ([0.8872727272727273, 0.9390681003584229, 1.0],
 [643, 1190, 1893],
 'Fred pays $16,240 for a home. If it costs 10% to sell his home before he could sell it at a profit, how much would
it have to appreciate?',
 'Mr. Robert pays $15,240 for a lot. It costs 12% to sell his lot. Before he could sell it at a profit how much would
it have to appreciate?',
 'Mr. Bill pays $152,400 for a home. If it costs 12% to sell his home before he could sell it at a profit, how much w
ould it have to appreciate?',
 'Fred pays $16,240 for a home. If it costs 10% to sell his home before he could sell it at a profit, how much would
it have to appreciate?')
```

- How do I parse this into "questions"?
- Clean it up first…
- Did it work?
- What is a "unique" question?

Real data crunching

	A	B	C	D	E
1	id	questions	similarity	indices	
2	1	Any complaint as to the violation of the United States Civil Rights Act of 1968 should be filed within ho	[1.0, 1.0, 1.0]	[847, 1304, 1748]	
3	2	A fee simple estate is most likely to be a:	[1.0]	[1104]	
4	3	Mr. Jones, an owner of a packaging firm, purchased a new machine in 2008 and paid $5,500. It was est	[]	[]	
5	4	Fred pays $16,240 for a home. If it costs 10% to sell his home before he could sell it at a profit, how mu	[0.8872727272727273, (	[643, 1190, 1893]	

COGS 108
Data Science in Practice

Data intuition and the Sniff Test (Fermi estimation)

UC San Diego

Data Intuition

In today's pattern recognition class my professor talked about PCA, eigenvectors & eigenvalues.

I got the mathematics of it. If I'm asked to find eigenvalues etc. I'll do it correctly like a machine. But I didn't **understand** it. I didn't get the purpose of it. I didn't get the feel of it. I strongly believe in

> you do not really understand something unless you can explain it to your grandmother -- Albert Einstein

Well, I can't explain these concepts to a layman or grandma.

1. Why PCA, eigenvectors & eigenvalues? What was the *need* for these concepts?
2. How would you explain these to a layman?

Data Intuition

In today's pattern recognition class my professor talked about PCA, eigenvectors & eigenvalues.

I got the mathematics of it. If I'm asked to find eigenvalues etc. I'll do it correctly like a machine. But I didn't **understand** it. I didn't get the purpose of it. I didn't get the feel of it. I strongly believe in

> you do not really understand something unless you can explain it to your grandmother -- Albert Einstein

Well, I can't explain these concepts to a layman or grandma.

1. Why PCA, eigenvectors & eigenvalues? What was the *need* for these concepts?
2. How would you explain these to a layman?

Theory *vs* Practice

"Tai's Model"

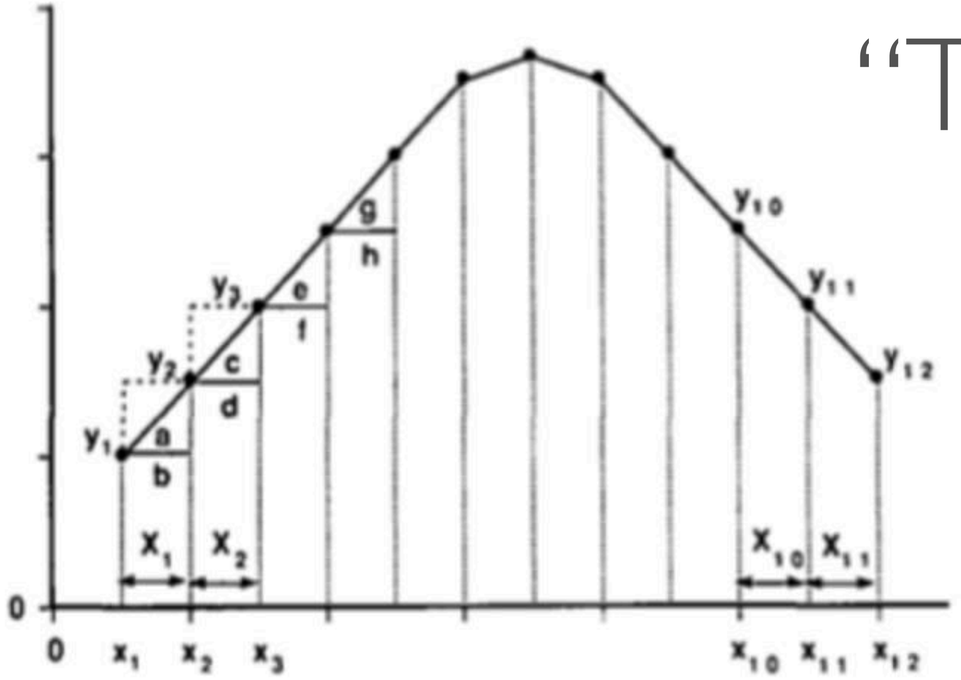


Figure 1—*Total area under the curve is the sum of individual areas of triangles a, c, e, and g and rectangles b, d, f, and h.*

Source: Tai, *Diabetes Care* 1994

Theory *vs* Practice

"In Tai's Model, the total area under a curve is computed by dividing the area under the curve between two designated values on the X-axis (abscissas) into small segments (rectangles and triangles) whose areas can be accurately calculated from their respective geometrical formulas. The total sum of these individual areas thus represents the total area under the curve."

UC San Diego

Source: Tai, *Diabetes Care* 1994

Fermi Estimation

Source: *TED-Ed*

Fermi Estimation

Paint the Earth

Has humanity produced enough paint to cover the entire land area of the Earth?

—Josh (Bolton, MA)

Source: *XKCD*, What If?

Fermi Estimation

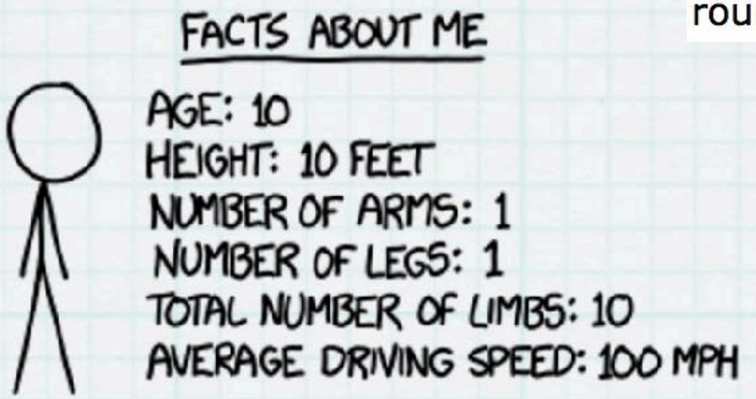
This answer is pretty straightforward. We can look up the size of the world's paint industry, extrapolate backward to figure out the total amount of paint produced. We'd also need to make some assumptions about how we're painting the ground. Note: When we get to the Sahara desert, I recommend not using a brush.

Source: *XKCD*, What If?

Fermi Estimation

But first, let's think about different ways we might come up with a guess for what the answer will be. In this kind of thinking—often called **Fermi estimation**—all that matters is getting in the right ballpark; that is, the answer should have about the right number of digits. In Fermi estimation, you can round[1] all your answers to the nearest order of magnitude:

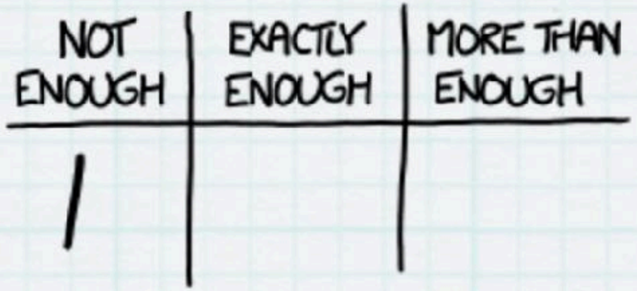
Using the formula $\text{Fermi}(x) = 10^{\text{round}(log_{10}x)}$, meaning that 3 rounds to 1 and 4 rounds to 10.



Source: *XKCD*, What If?

Fermi Estimation

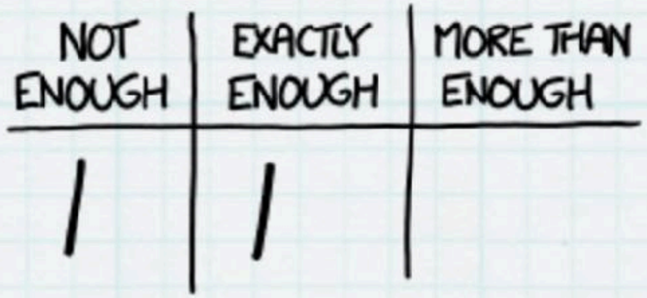
Let's suppose that, on average, everyone in the world is responsible for the existence of two rooms, and they're both painted. My living room has about 50 square meters of paintable area, and two of those would be 100 square meters. 7.15 billion people times 100 square meters per person is a little under a trillion square meters—an area smaller than Egypt.



Source: *XKCD*, What If?

Fermi Estimation

Let's make a wild guess that, on average, one person out of every thousand spends their working life painting things. If I assume it would take me three hours to paint the room I'm in, [2] and 100 billion people have ever lived, and each of them spent 30 years painting things for 8 hours a day, we come up with 150 trillion square meters ... just about exactly the land area of the Earth.



Source: *XKCD*, What If?

Fermi Estimation

How much paint does it take to paint a house? I'm not enough of an adult to have any idea, so let's take another Fermi guess.

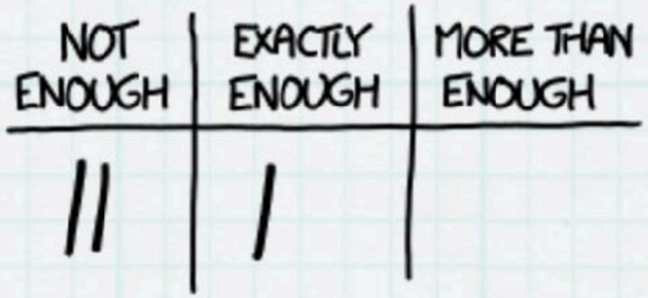
Based on my impressions from walking down the aisles, home improvement stores stock about as many light bulbs as cans of paint. A normal house might have about 20 light bulbs, so let's assume a house needs about 20 gallons of paint.[3] Sure, that sounds about right.

UC San Diego

Source: *XKCD*, What If?

Fermi Estimation

The average US home costs about $200,000. Assuming each gallon of paint covers about 300 square feet, that's a square meter of paint per $300 of real estate. I vaguely remember that the world's real estate has a combined value of something like $100 trillion,[4] which suggests there's about 300 billion square meters of paint on the world's real estate. That's about one New Mexico.



Source: *XKCD*, What If?

Fermi Estimation

Of course, both of the building-related guesses could be overestimates (lots of buildings are not painted) or underestimates (lots of things that are not buildings[5] are painted) But from these wild Fermi estimates, my guess would be that there probably isn't enough paint to cover all the land.

So, how did Fermi do?

Source: *XKCD*, What If?

Fermi Estimation

According to the report **The State of the Global Coatings Industry**, the world produced 34 billion liters of paints and coatings in 2012.

There's a neat trick that can help us here. If some quantity—say, the world economy—has been growing for a while at an annual rate of **n**—say, 3% (0.03)—then the most recent year's share of the whole total so far is $1 - \frac{1}{1+n}$, and the whole total so far is the most recent year's amount times $1 + \frac{1}{n}$.

UC San Diego

Source: *XKCD*, What If?

Fermi Estimation

If we assume paint production has, in recent decades, followed the economy and grown at about 3% per year, that means the total amount of paint produced equals the current yearly production times 34.[6] That comes out to a little over a trillion liters of paint. At 30 square meters per gallon,[7] that's enough to cover 9 trillion square meters—about the area of the United States.

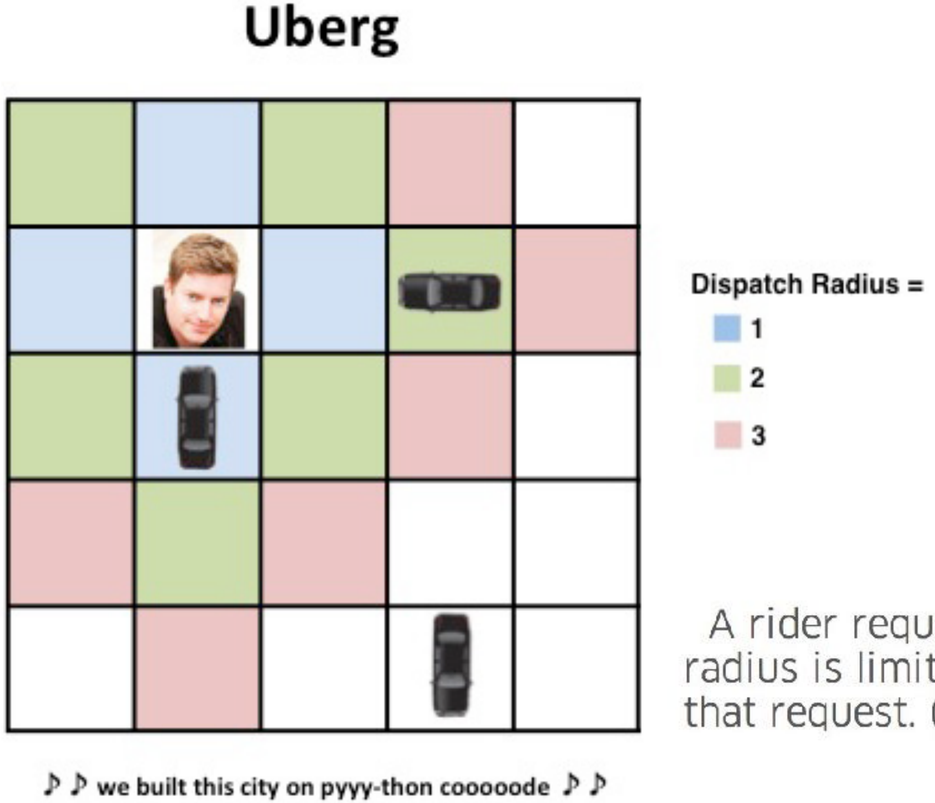
So the answer is no; there's not enough paint to cover the Earth's land, and—at this rate—probably won't be enough until the year 2100.

Source: *XKCD*, What If?

Fermi Estimation

bit.ly/cogs108wi18_fermi

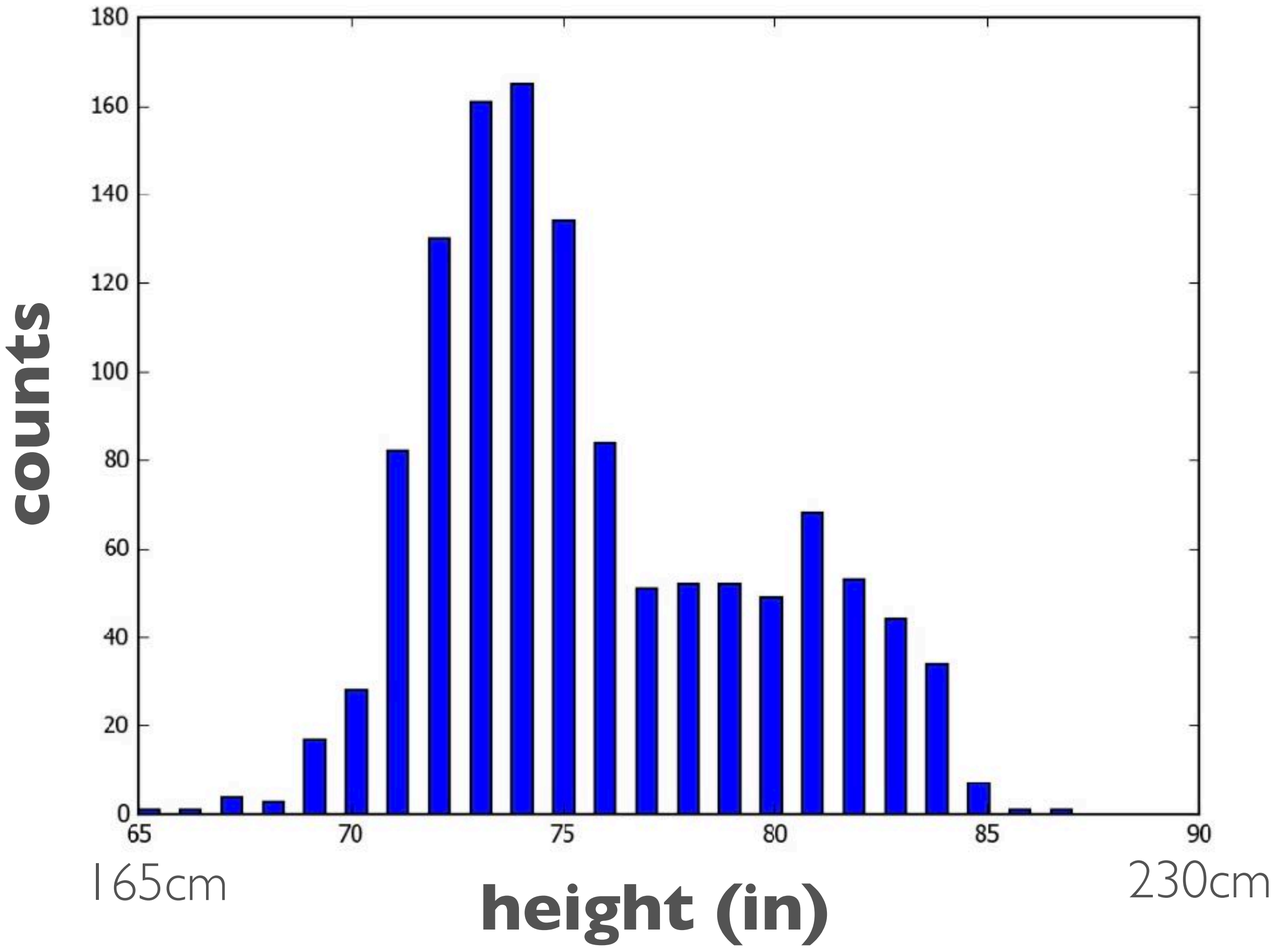
Fermi Estimation in data science



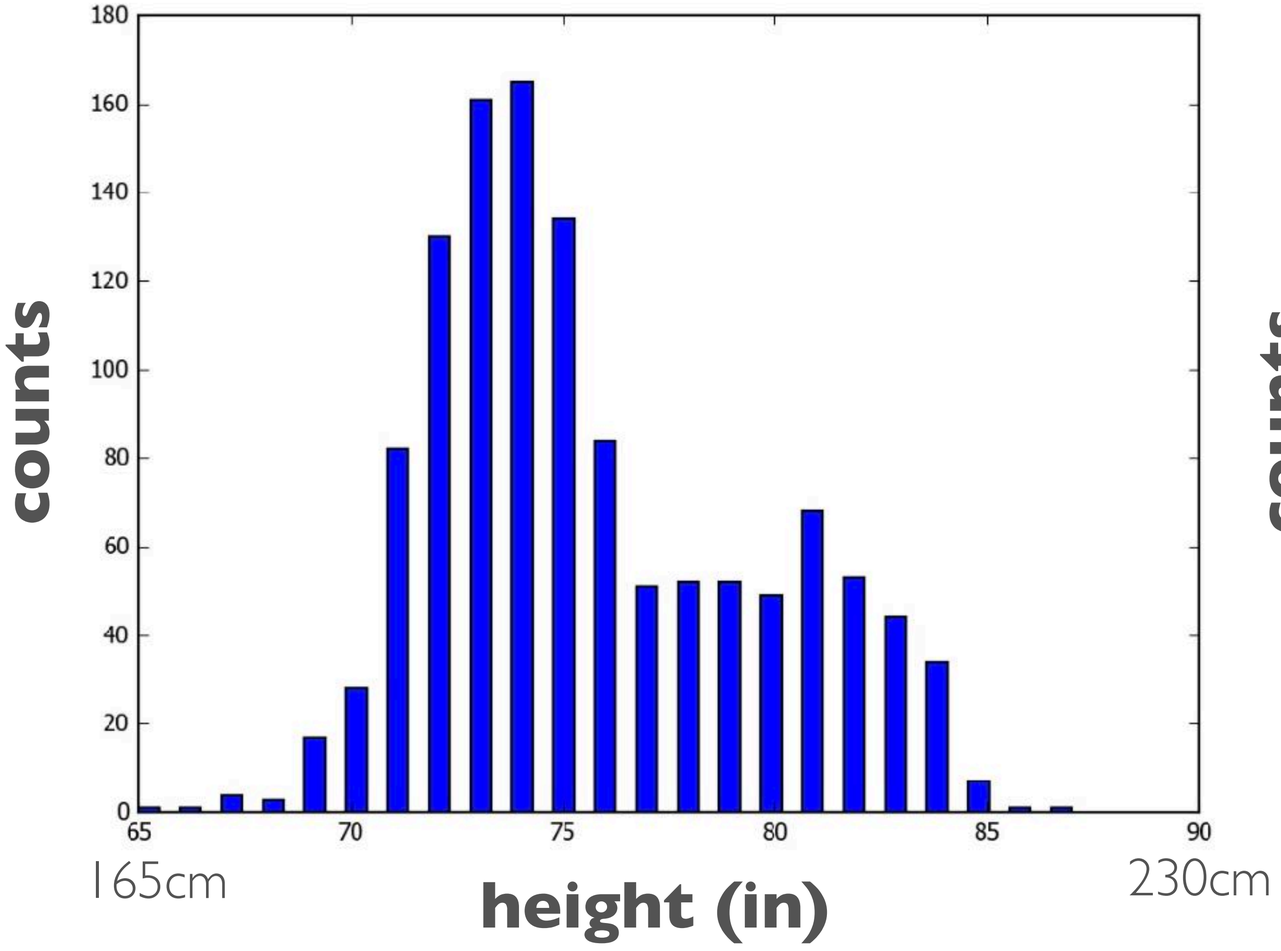
A rider requests a car in the simulated city of Uberg. If the dispatch radius is limited to 1 unit, then only cars within the blue zone will see that request. (Similar to how taxi street hails work: taxis can only pick up the passengers they see.)

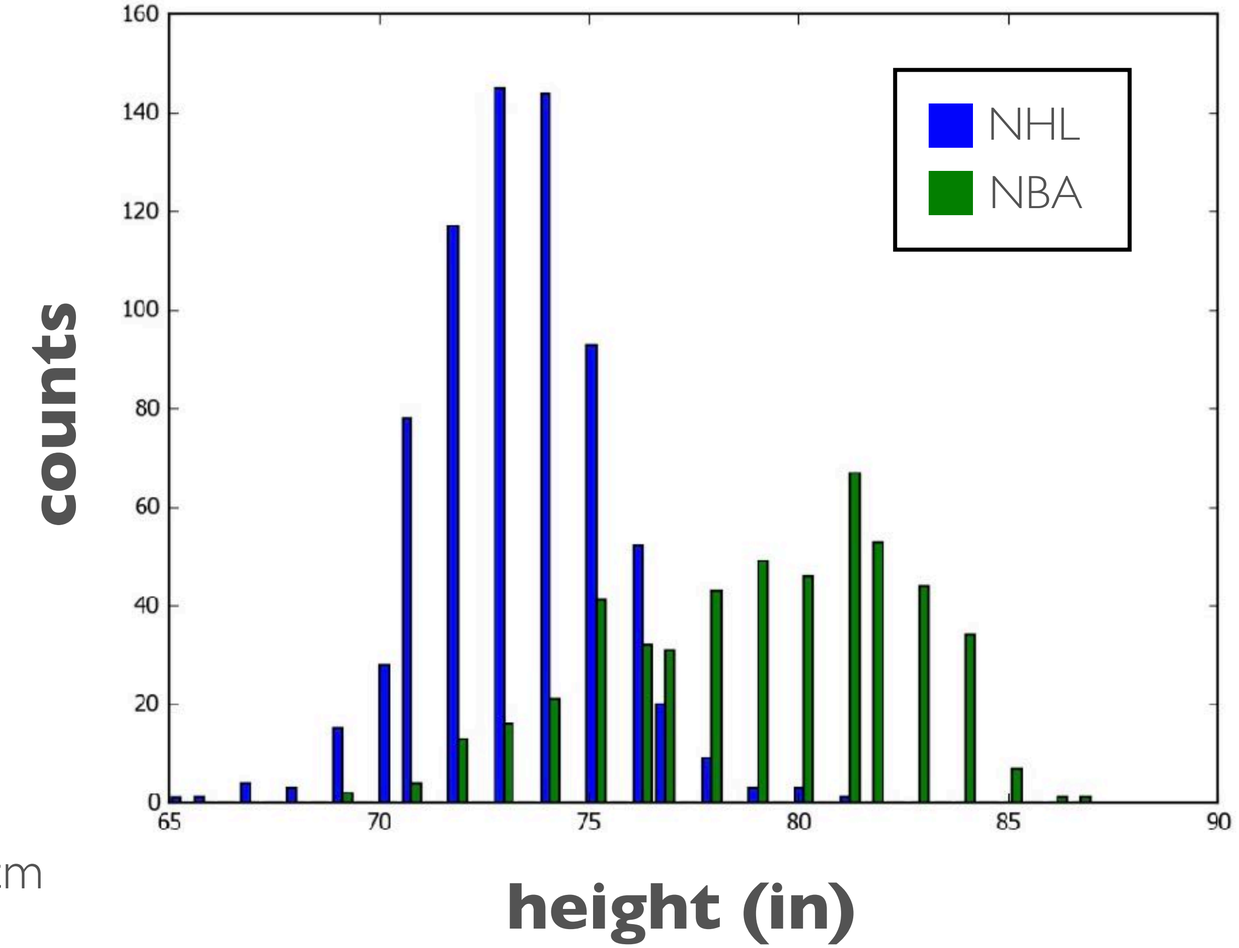
Source: Voytek, *Uber, Inc.*

Data intuition

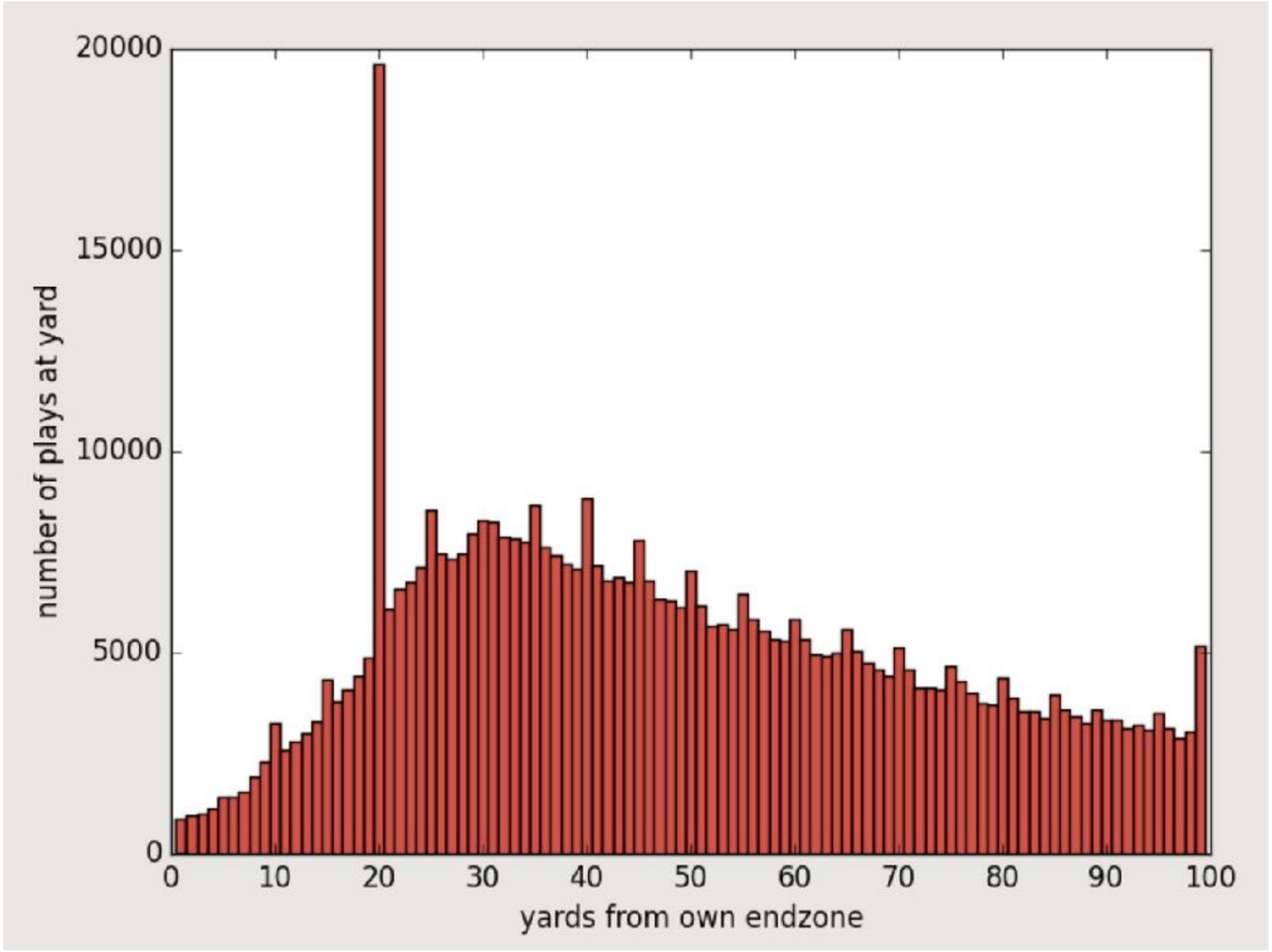


Data intuition

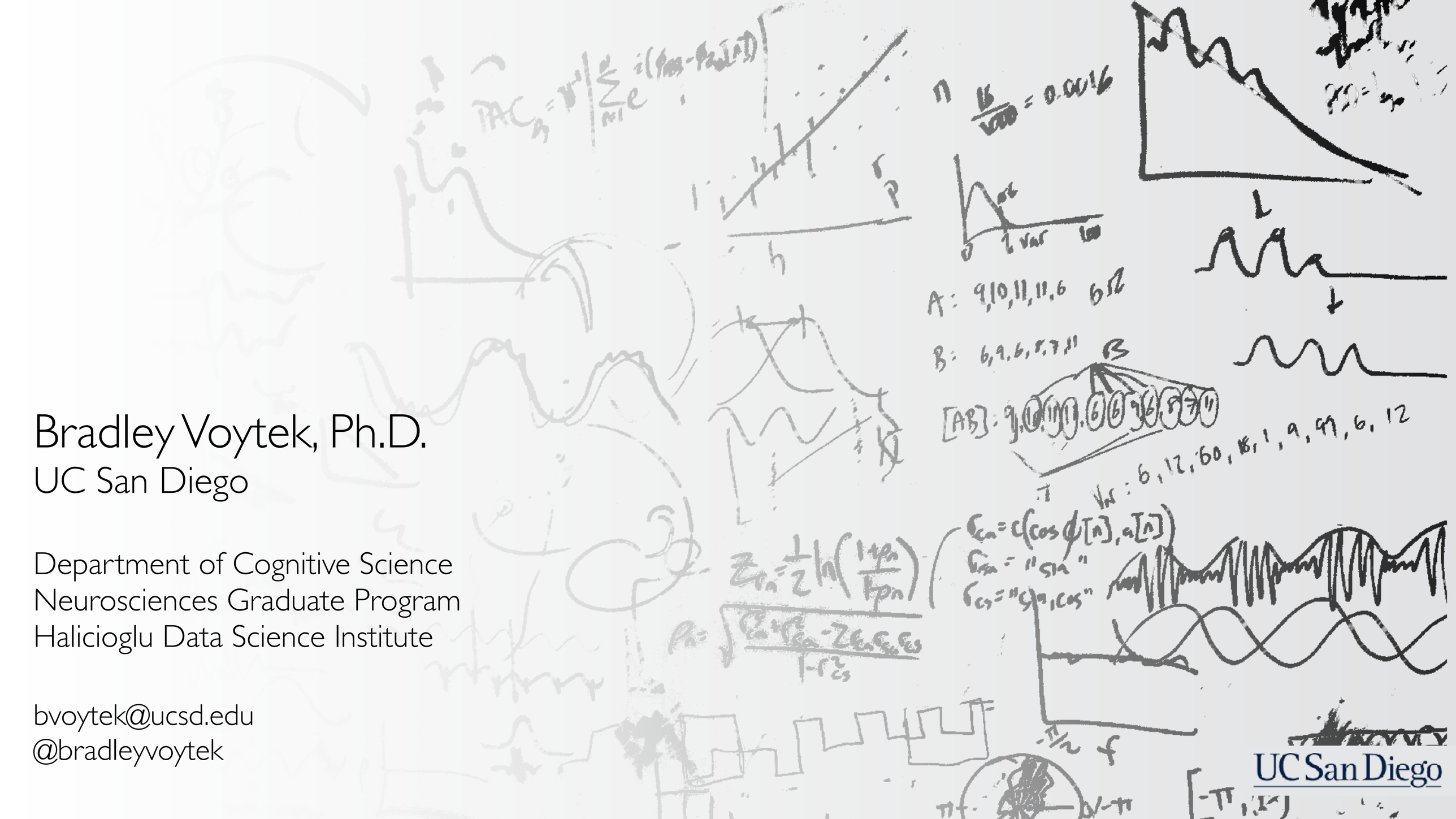




Data intuition



source: https://gutterstats.wordpress.com/2015/11/03/are-nfl-officials-biased-with-their-ball-placement/



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