

Course Reminders

- Due this Friday (11:59 PM)
 - Q1
 - D1
 - A1
- Projects
 - You will be assigned a GitHub repo this weekend - please accept the invitation (it will expire)
 - You will also be assigned a previous project to review (links will be on Canvas)

COGS 108 Final Projects

The COGS 108 Final Project will give you the chance to explore a topic of your choice and to expand your analytical skills. By working with real data of your choosing you can examine questions of particular interest to you.

- You are encouraged to work on a topic that matters to the world (your family, your neighborhood, a state/province, country, etc).
- Taboo Topics: Movie Predictions/Recommendation System; YouTube Data Analysis, Kickstarter success prediction/analysis,prediction of what makes a song popular on Spotify

Final Project: Objectives

- 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(s).
- 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.

Upcoming Project Components

Project Planning Survey (1%) - 1 submission per group (due Fri Week 2)

Project Review (5%) - Before Mon of week 3, your group will be assigned a previous COGS 108 project to review; A google Form will be released to guide your thinking/discussion about and review of what a previous COGS 108 group did for their project. (due Fri Week 3)

Project Proposal (8%) - a GitHub repo will be created for your group; ‘submit’ on GitHub (due Fri Week 4)

Project Proposal (8%)

Full project guidelines are here:

<https://github.com/COGS108/Projects/blob/master/FinalProject%20Guidelines.md>

Data tidiness & intuition

Jason G. Fleischer, Ph.D
UC San Diego

• • •

Department of Cognitive Science
jfleischer@ucsd.edu
<https://jgfleischer.com>
[@jasongfleischer](https://twitter.com/jasongfleischer)



Data Structures Review

Structured data

- can be stored in database SQL
- tables with rows and columns
- requires a relational key
- 5-10% of all data

Semi-structured data

- doesn't reside in a relational database
- has organizational properties (easier to analyze)
- CSV, XML, JSON

Unstructured

- non-tabular data
- 80% of the world's data
- images, text, audio, videos

(Semi-)Structured Data

Data that is stored in such a way that it is easy to search and work with. These data are stored in a particular format that adheres to organization principles imposed by the file format. These are the data structures data scientists work with most often.

CSVs

Each column separated by a comma

Has the extension ".csv"

Example CSV - Sheet1 — Notatnik				
Plik	Edycja	Format	Widok	Pomoc
Email	First Name	Last Name	Company	Snippet 1
example1@domain.com	John	Smith	Company 1	Snippet Sentence1
example2@gmail.com	Mary	Blake	Company 2	Snippet Sentence 2
example3@outlook.com	James	Joyce	Company 3	Snippet Sentence 3

Each row is separated by a new line



Example CSV



File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

fx

	A	B	C	D	E	F
1	Email	First Name	Last Name	Company	Snippet 1	
2	example1@domain.com	John	Smith	Company 1	Snippet Sentence1	
3	example2@gmail.com	Mary	Blake	Company 2	Snippet Sentence 2	
4	example3@outlook.com	James	Joyce	Company 3	Snippet Sentence 3	

CSV file

Example CSV - Sheet1 — Notatnik

Plik Edycja Format Widok Pomoc

```
Email,First Name,Last Name,Company,Snippet 1
example1@domain.com,John,Smith,Company 1,Snippet Sentence1
example2@gmail.com,Mary,Blake,Company 2,Snippet Sentence 2
example3@outlook.com,James,Joyce,Company 3,Snippet Sentence 3
```

JSON: key-value pairs

nested/hierarchical data

```
{"Name": "Isabela"}
```

key

value

JSON

These are all
nested within
attributes

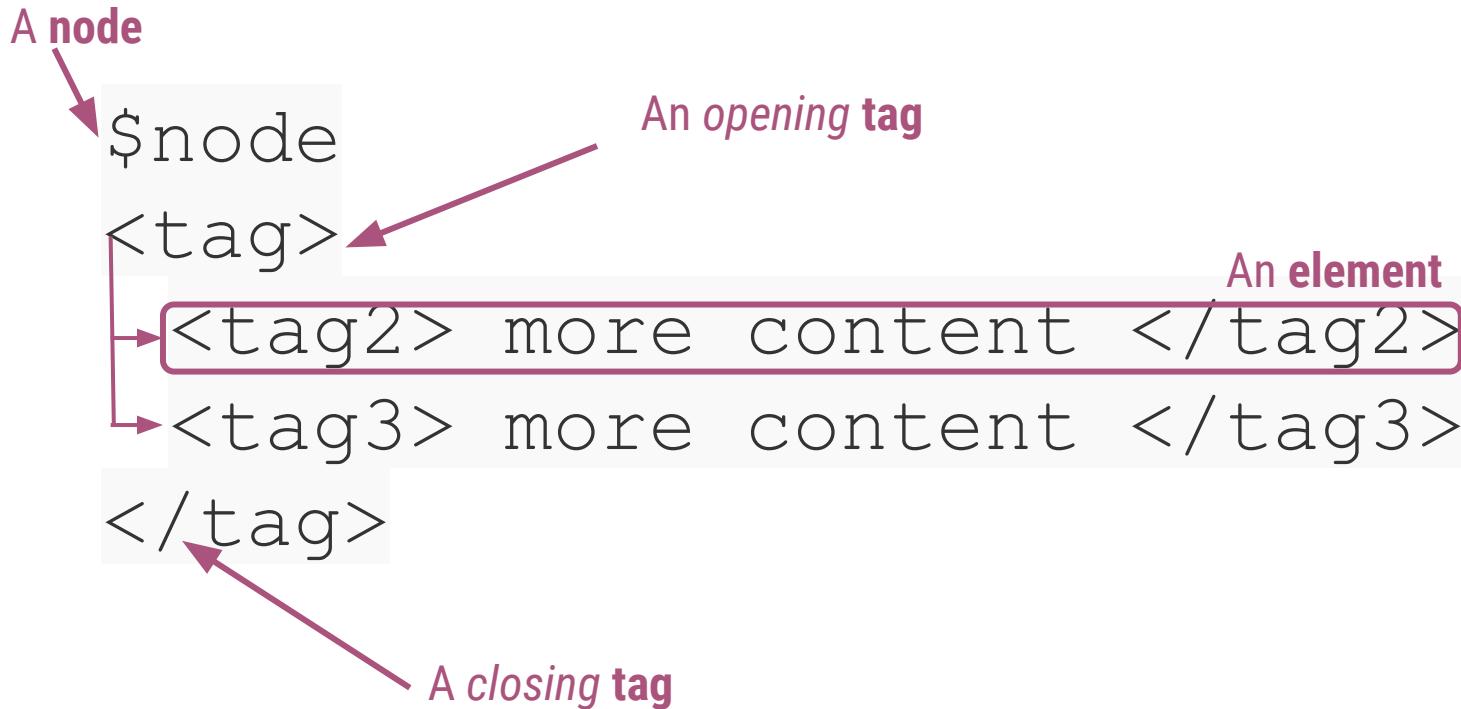
```
"attributes": {  
    "Take-out": true,  
    "Wi-Fi": "free",  
    "Drive-Thru": true,  
    "Good For": {  
        "dessert": false,  
        "latenight": false,  
        "lunch": false,  
        "dinner": false,  
        "breakfast": false,  
        "brunch": false  
    },
```

These are all
nested within
"Good For"

JSON

Extensible Markup Language (XML): nodes, tags, and elements

nested/hierarchical data



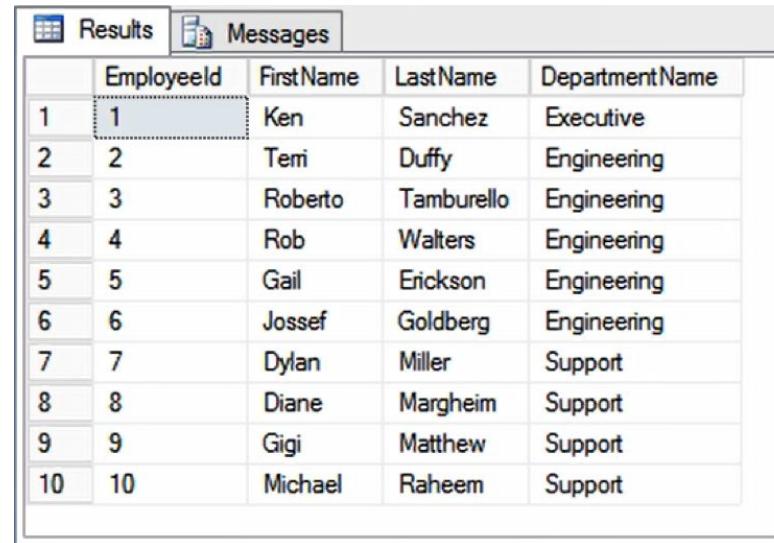
```
<?xml version="1.0" encoding="UTF-8"?>
<customers>
    <customer>
        <customer_id>1</customer_id>
        <first_name>John</first_name>
        <last_name>Doe</last_name>
        <email>john.doe@example.com</email>
    </customer>
    <customer>
        <customer_id>2</customer_id>
        <first_name>Sam</first_name>
        <last_name>Smith</last_name>
        <email>sam.smith@example.com</email>
    </customer>
    <customer>
        <customer_id>3</customer_id>
        <first_name>Jane</first_name>
        <last_name>Doe</last_name>
        <email>jane.doe@example.com</email>
    </customer>
</customers>
```

XML

adapted from Chris Keown

Relational Databases: A set of interdependent tables

1. Efficient Data Storage
2. Avoid Ambiguity
3. Increase Data Privacy



The screenshot shows a Microsoft SQL Server Management Studio (SSMS) interface with the 'Results' tab selected. The results window displays a table of employee data with columns: EmployeeId, FirstName, LastName, and DepartmentName. The data consists of 10 rows, each representing an employee. The first column, EmployeeId, is highlighted with a dotted selection box around the value '1'. The table has a light gray background with white borders between rows and columns.

	EmployeeId	FirstName	LastName	DepartmentName
1	1	Ken	Sanchez	Executive
2	2	Teni	Duffy	Engineering
3	3	Roberto	Tamburello	Engineering
4	4	Rob	Walters	Engineering
5	5	Gail	Erickson	Engineering
6	6	Jossef	Goldberg	Engineering
7	7	Dylan	Miller	Support
8	8	Diane	Margheim	Support
9	9	Gigi	Matthew	Support
10	10	Michael	Raheem	Support

relational database

Information is stored across tables

	unique_identifier	
	AH13JK	
	JJ29JJ	
	CI21AA	

	unique_identifier	
	AH13JK	
	JJ29JJ	
	JJ29JJ	
	XJ11AS	
	CI21AA	

	unique_identifier	
	AH13JK	
	SE92FE	
	CI21AA	

entries are *related* to one another by their unique identifier

relational database

restaurant

name	id	address	type
Taco Stand	AH13JK	1 Main St.	Mexican
Pho Place	JJ29JJ	192 Street Rd.	Vietnamese
Taco Stand	XJ11AS	18 W. East St.	Fusion
Pizza Heaven	CI21AA	711 K Ave.	Italian

health inspections

id	inspection_date	inspector	score
AH13JK	2018-08-21	Sheila	97
JJ29JJ	2018-03-12	D'eonte	98
JJ29JJ	2018-01-02	Monica	66
XJ11AS	2018-12-16	Mark	43
CI21AA	2018-08-21	Anh	99

rating

id	stars
AH13JK	4.9
JJ29JJ	4.8
XJ11AS	4.2
CI21AA	4.7

relational database

restaurant

name	id	address	type
Taco Stand	AH13JK	1 Main St.	Mexican
Pho Place	JJ29JJ	192 Street Rd.	Vietnamese
Taco Stand	XJ11AS	18 W. East St.	Fusion
Pizza Heaven	CI21AA	711 K Ave.	Italian

Two different restaurants with
the same name will have
different unique identifiers

health inspections

id	inspection_date	inspector	score
AH13JK	2018-08-21	Sheila	97
JJ29JJ	2018-03-12	D'eonte	98
JJ29JJ	2018-01-02	Monica	66
XJ11AS	2018-12-16	Mark	43
CI21AA	2018-08-21	Anh	99

rating

id	stars
AH13JK	4.9
JJ29JJ	4.8
XJ11AS	4.2
CI21AA	4.7

relational database

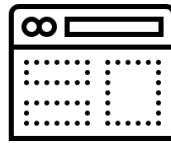
Unstructured Data

Some datasets record information about the state of the world, but in a more heterogeneous way. Perhaps it is a large text corpus with images and links like Wikipedia, or the complicated mix of notes and test results appearing in personal medical records.

Unstructured Data Types



Text files and documents



Websites and applications



Sensor data



Image files



Audio files



Video files



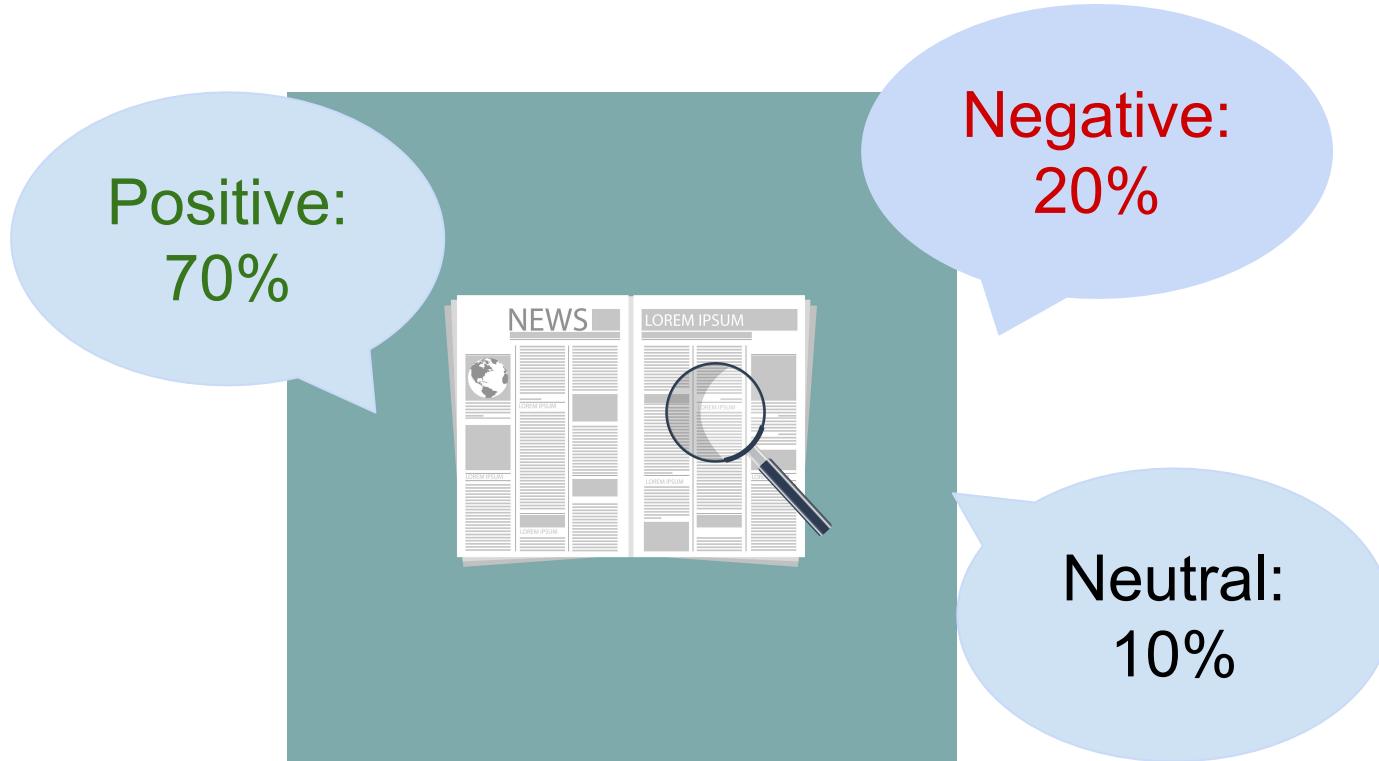
Email data

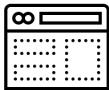


Social media data



Text: Sentiment Analysis

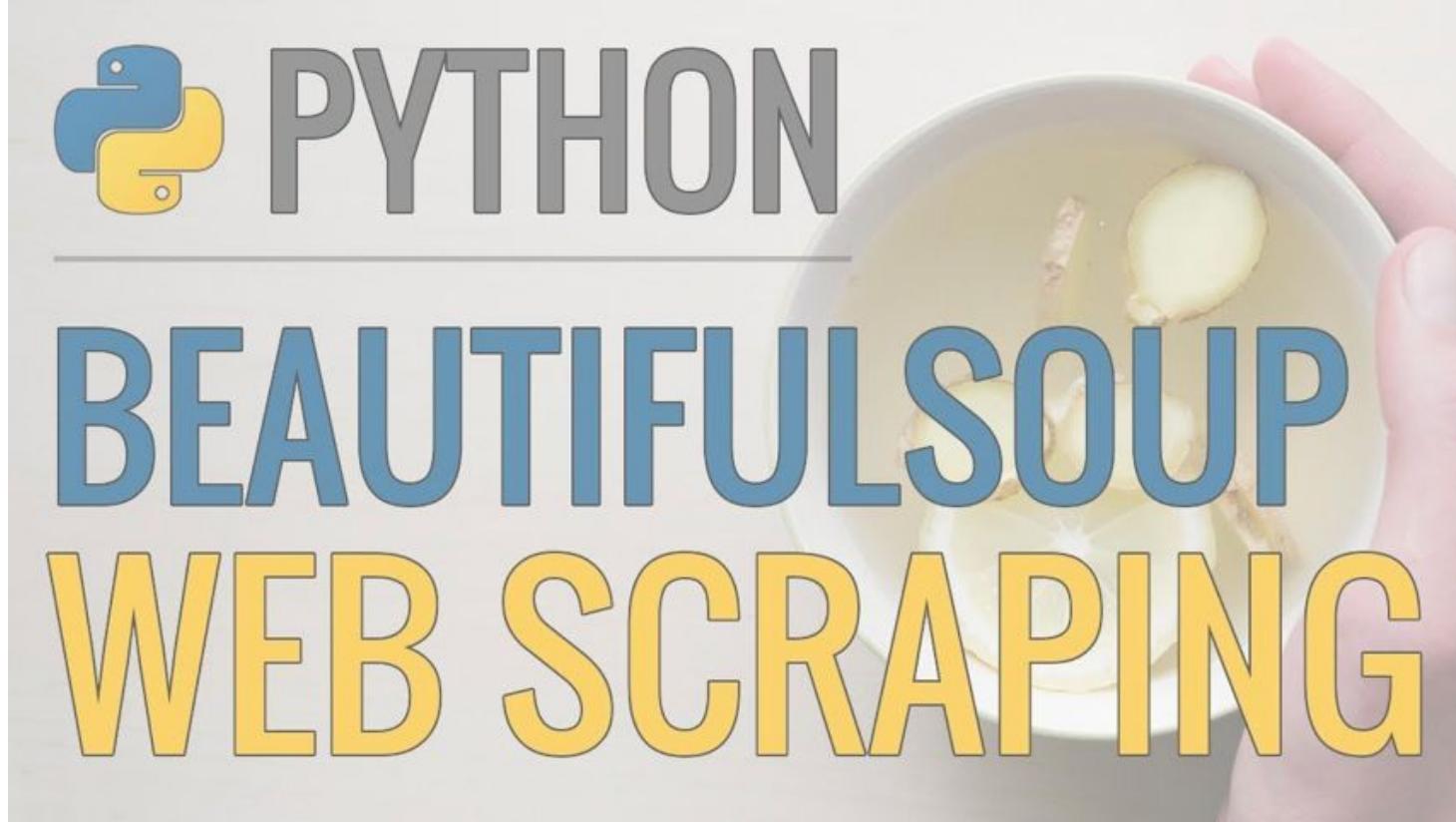




PYTHON

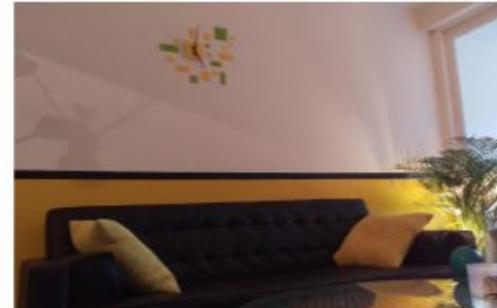
BEAUTIFULSOUP

WEB SCRAPING





Bedroom Or Not?



"The left two photos were correctly predicted as bedrooms; The right two photos were correctly predicted NOT as bedrooms."

Tidy Data

"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." - DJ Patil

untidy data

Australian Bureau of Statistics												
1800.0 Australian Marriage Law Postal Survey, 2017												
Released 15 November 2017												
Table junk												
1	Australian Bureau of Statistics	Yeah NA	18-19 years	20-24 years	25-29 years	30-34 years	35-39 years	40-44 years	45-49 years	50-54 years	55-59 years	60-64 years
2	Lingua(c)	Total participants	292	1,058	1,460	1,653	1,515	1,516	1,710	1,730	1,753	1,574
3	Primary keynotes	Eligible participants	572	2,910	3,040	3,996	3,607	3,506	3,645	3,331	2,960	2,456
4	Merged cells	Participation rate (%)	51.0	36.4	38.7	41.4	42.0	43.2	46.8	51.9	58.2	64.1
5	Solomon	Comma on										
6	Northern Territory	Total participants	734	2,519	3,531	4,010	3,703	3,573	3,934	3,838	3,887	3,346
7	(Total)	Eligible participants	1,322	5,901	7,783	8,151	7,241	6,904	7,072	6,397	5,891	4,811
8		Participation rate (%)	55.5	42.7	45.4	49.2	51.1	51.8	55.6	60.0	66.0	69.5
9	Australian Capital Territory Divisions	Covariate as Subheading	Summary of data inside data									
10	Canberra(d)	Total participants	1,764	4,789	4,817	4,973	4,626	4,453	5,074	4,826	5,169	4,394
11		Eligible participants	2,260	6,471	6,446	6,509	5,983	5,805	6,302	5,902	6,044	5,057
12		Participation rate (%)	78.1	74.0	74.7	76.4	77.3	76.7	80.5	81.8	85.5	86.9
13	Fenner(e)	Total participants	1,477	4,687	5,178	5,786	6,025	5,463	5,191	4,208	3,948	3,465
14		Eligible participants	1,904	6,354	7,121	7,822	7,960	7,155	6,480	5,206	4,692	3,945
15		Participation rate (%)	77.6	73.8	72.7	74.0	75.7	76.4	80.1	80.8	84.1	87.8
16		NA Yeah										
17	Australian Capital Territory (Total)	Total participants	5,242	8,476	8,895	10,735	10,054	9,219	10,205	9,034	9,411	7,659
18		Eligible participants	4,164	12,825	15,569	14,331	13,943	12,960	12,782	11,108	10,736	9,002
19		Participation rate (%)	77.8	73.9	73.7	75.1	76.4	76.5	80.3	81.3	84.9	87.3
20	Australia	Total participants	151,297	438,166	441,658	460,548	462,206	479,360	524,620	517,693	543,449	506,799
21		Eligible participants	201,439	635,909	646,916	665,250	656,446	660,841	693,850	659,150	664,720	597,386
22		Participation rate (%)	75.1	68.9	68.3	69.2	70.4	72.5	75.6	78.5	81.8	84.8
23	a) The Federal Electoral Divisions are current as at 24 August 2017	Return of the table junk										
24	b) Includes those whose age is unknown											
25	c) Includes Christmas Island and the Cocos (Keeling) Islands											
26	d) Includes Norfolk Island											
27	e) Includes Jervis Bay											
28		MS Excel or Die										

tidy data

data
wrangling

area	gender	age	State	Area (sq km)	Eligible participants	Participation rate (%)	Total participants	Total Participants
Adelaide	Female	18-19 years	SA	76	1341	83.5	1120	1120
Adelaide	Female	20-24 years	SA	76	4620	81.2	3750	3750
Adelaide	Female	25-29 years	SA	76	4897	81.8	4004	4004
Adelaide	Female	30-34 years	SA	76	4784	79.8	3820	3820
Adelaide	Female	35-39 years	SA	76	4319	79	3411	3411
Adelaide	Female	40-44 years	SA	76	4310	80.6	3472	3472
Adelaide	Female	45-49 years	SA	76	4579	81.4	3728	3728
Adelaide	Female	50-54 years	SA	76	4475	84.7	3791	3791
Adelaide	Female	55-59 years	SA	76	4622	87.3	4033	4033
Adelaide	Female	60-64 years	SA	76	4342	89.3	3879	3879
Adelaide	Female	65-69 years	SA	76	3970	90.7	3602	3602
Adelaide	Female	70-74 years	SA	76	3009	90.3	2716	2716
Adelaide	Female	75-79 years	SA	76	2156	88.5	1908	1908
Adelaide	Female	80-84 years	SA	76	1673	85.1	1423	1423

Tidy Data

1. Each **variable** you measure should be in a single column

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

2. Every **observation** of a variable should be in a different row

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

3. There should be one table for each type of data

Demographic Survey Data

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2		1004	Smith	Jane	female	Frederick	MD
3		4587	Nayef	Mohammed	male	Upper Darby	PA
4		1727	Doe	Janice	female	San Diego	CA
5		6879	Jordan	Alex	male	Birmingham	AL
							Teacher

Doctor's Office Measurements Data

	A	D	E	F	G
1	ID	Height_inches	Weight_lbs	Insulin	Glucose
2		1004	65	180	0.60
3		4587	75	215	1.46
4		1727	62	124	0.72
5		6879	77	160	1.23
					205

4. If you have multiple tables, they should include a column in each *with the same column label* that allows them to be joined or merged

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Sex	City	State	Occupation
2	1004	Smith	Jane	female	Frederick	MD	Welder
3	4587	Nayef	Mohammed	male	Upper Darby	PA	Nurse
4	1727	Doe	Janice	female	San Diego	CA	Doctor
5	6879	Jordan	Alex	male	Birmingham	AL	Teacher

	A	D	E	F	G
1	ID	Height_inches	Weight_lbs	Insulin	Glucose
2	1004	65	180	0.60	163
3	4587	75	215	1.46	150
4	1727	62	124	0.72	177
5	6879	77	160	1.23	205

Tidy data == rectangular data

A

	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

Tidy Data Benefits

1. consistent data structure
2. foster tool development
3. require only a small set of tools to be learned
4. allow for datasets to be combined

Tabular Data Time

<https://forms.gle/6EKypPu1odgBdbiJ8>



A

ID	Last	First	height_m	height_f
1004	Smith	Jane	NA	65
4587	Nayef	Mohammed	72	NA
1727	Doe	Janice	NA	60
6879	Jordan	Alex	55	NA

B

ID	Last	First	height_m	height_f
1004	Smith	Jane		65
4587	Nayef	Mohammed	72	
1727	Doe	Janice		60
6879	Jordan	Alex	55	

C

ID	Last	First	sex	height
1004	Smith	Jane	female	65
4587	Nayef	Mohammed	male	72
1727	Doe	Janice	fem	60
6879	Jordan	Alex	male	55

D

ID	Last	First	sex	height
1004	Smith	Jane	F	65
4587	Nayef	Mohammed	M	72
1727	Doe	Janice	F	60
6879	Jordan	Alex	M	55

Which of these tables stores data best?



A



B



C



D





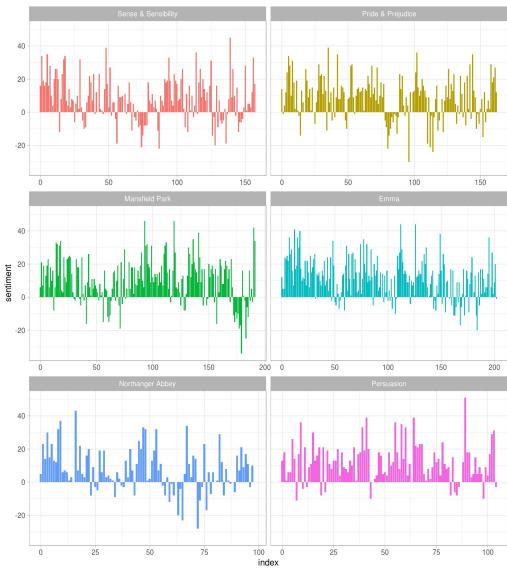
text

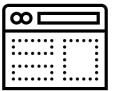
tidy dataset

Word	Novel	Frequency
good	Emma	359
young	Emma	192
friend	Emma	166



results





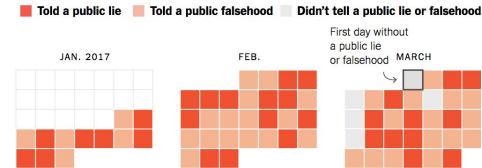
website

A black arrow pointing to the right, indicating the direction of the next page or section.

tidy dataset

date	lie	explanation	url
0 Jan 21, 2017	I wasn't a fan of Iraq, I didn't want to go in...	He was for an invasion before he was against it.	https://www.buzzfeed.com/andrewiakaczynski/lin-20...
1 Jan 21, 2017	A reporter for Time magazine — and I have been...	Trump was on the cover 11 times and Nixon appeared...	http://nation.time.com/2013/11/06/10-things-yo...
2 Jan 23, 2017	Between 3 million and 5 million illegal votes ...	There's no evidence of illegal voting.	https://www.nytimes.com/2017/01/23/us/politics...
3 Jan 25, 2017	Now, the audience was the biggest ever. But th...	Official serial photos show Obama's 2009 inaug...	https://www.nytimes.com/2017/01/21/us/politics...
4 Jan 25, 2017	Take a look at the Pew reports (which show vot...	The report never mentioned voter fraud.	https://www.nytimes.com/2017/01/24/us/politics...

results



text (lyrics)

The Pudding



"I'll be analyzing the repetitiveness of a dataset of 15,000 songs that charted on the Billboard Hot 100 between 1958 and 2017."

AN EXERCISE IN LANGUAGE COMPRESSION

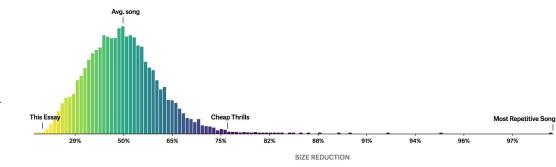
Are Pop Lyrics Getting More Repetitive?

By Colin Morris



song	Artist	Released	Reduction
Cheap Thrills	Sia	2016	76
Around The World	Daft Punk	1997	98
Everybody Dies	J. Cole	2018	27

tidy dataset



results

What are these uber-repetitive outliers? *Around The World* by Daft Punk gets reduced a whopping 98%. It goes from 2,610 characters to 61. Small enough to fit in a tweet - twice!

Data Intuition



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

1011

I understood 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 the following quote:



1375



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”

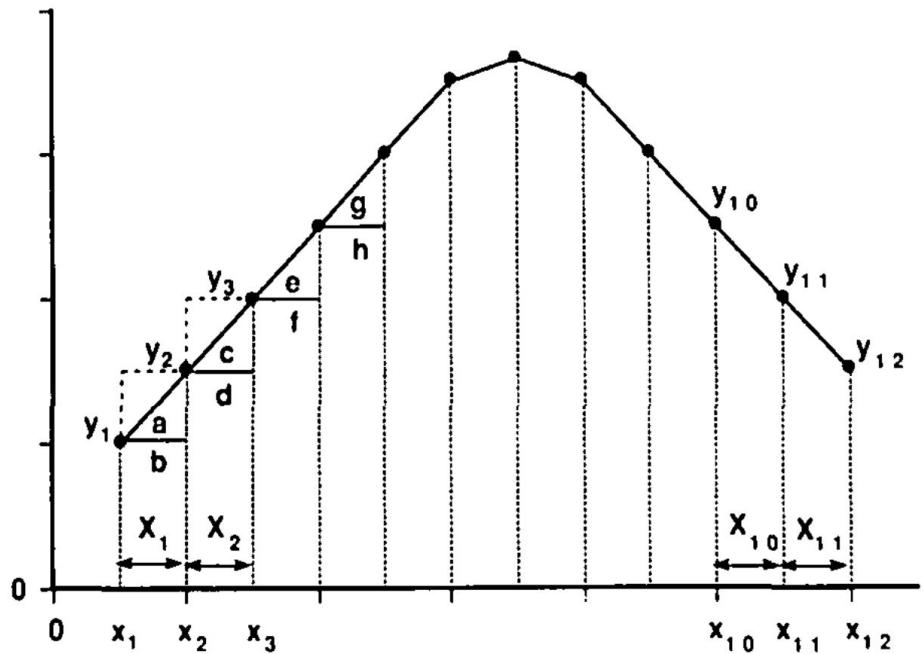


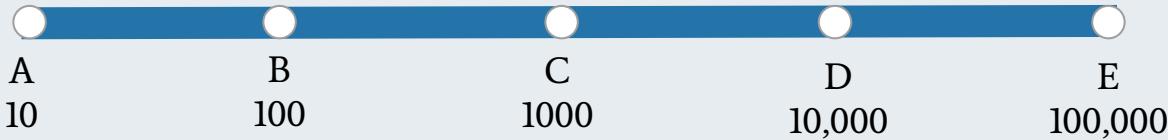
Figure 1—Total area under the curve is the sum of individual areas of triangles *a*, *c*, *e*, and *g* and rectangles *b*, *d*, *f*, and *h*.



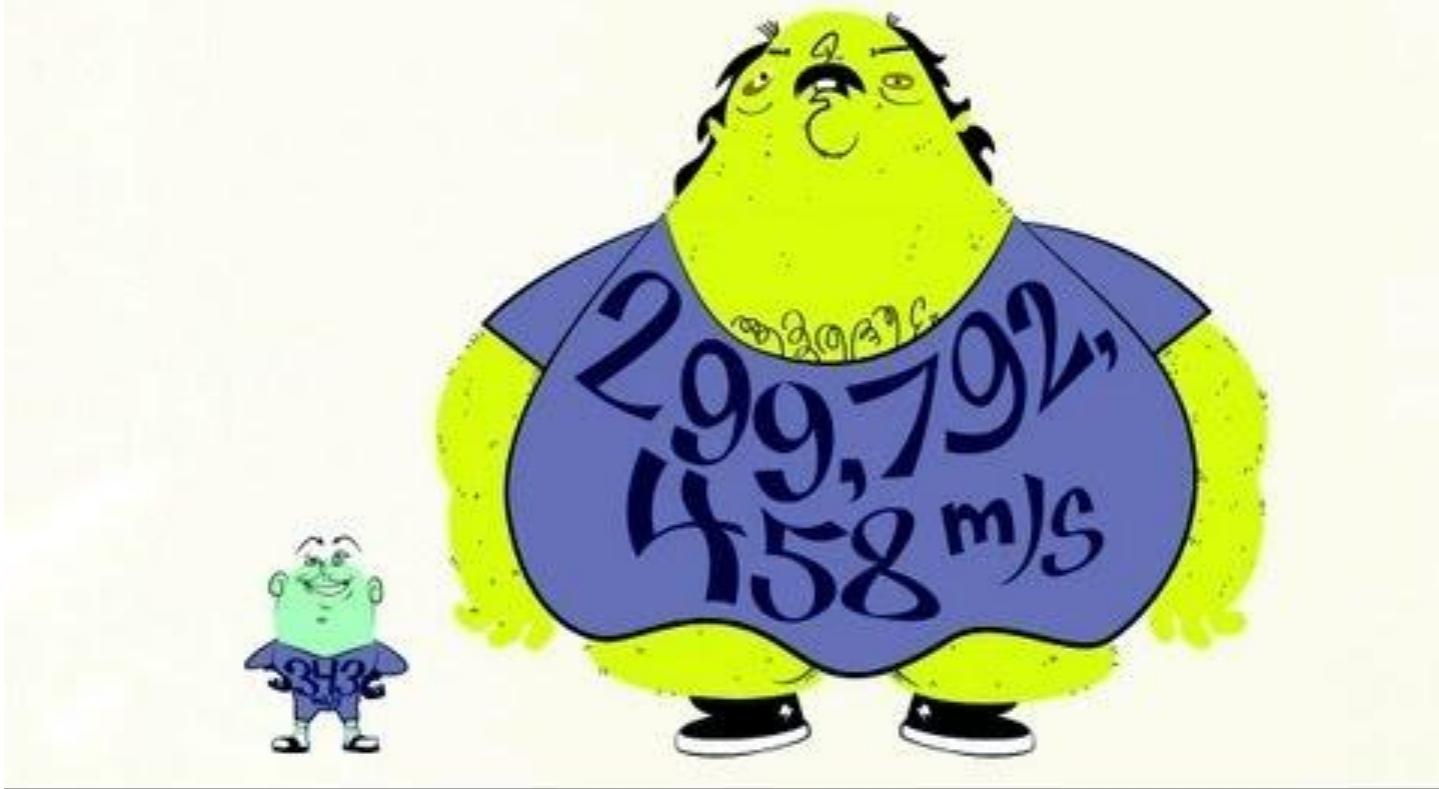
Fermi Estimation

<https://forms.gle/C982naWtU9RvHqAb7>

Approximately how many piano tuners do you think there are in the city of Chicago?







<https://www.youtube.com/watch?v=OYzvupOX8ls>

**Has humanity produced enough paint to
cover the entire land area of the Earth?
—Josh (Bolton, MA)**

Fermi Estimation

<https://forms.gle/shS84W1tai4SDrVF9>



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

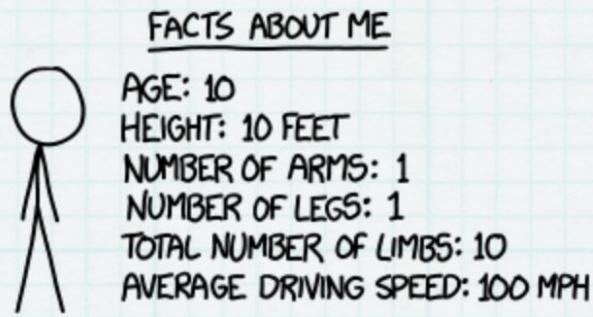




This answer is pretty straightforward. We can look up the size of the world's paint industry, extrapolate backward to figure out the total amount of paint produced. We'd also need to make some assumptions about how we're painting the ground. Note: When we get to the Sahara desert, I recommend not using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



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:



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.

NOT ENOUGH	EXACTLY ENOUGH	MORE THAN ENOUGH
/		

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.

NOT ENOUGH	EXACTLY ENOUGH	MORE THAN ENOUGH
/	/	

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.

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.

NOT ENOUGH	EXACTLY ENOUGH	MORE THAN ENOUGH
//	/	

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.

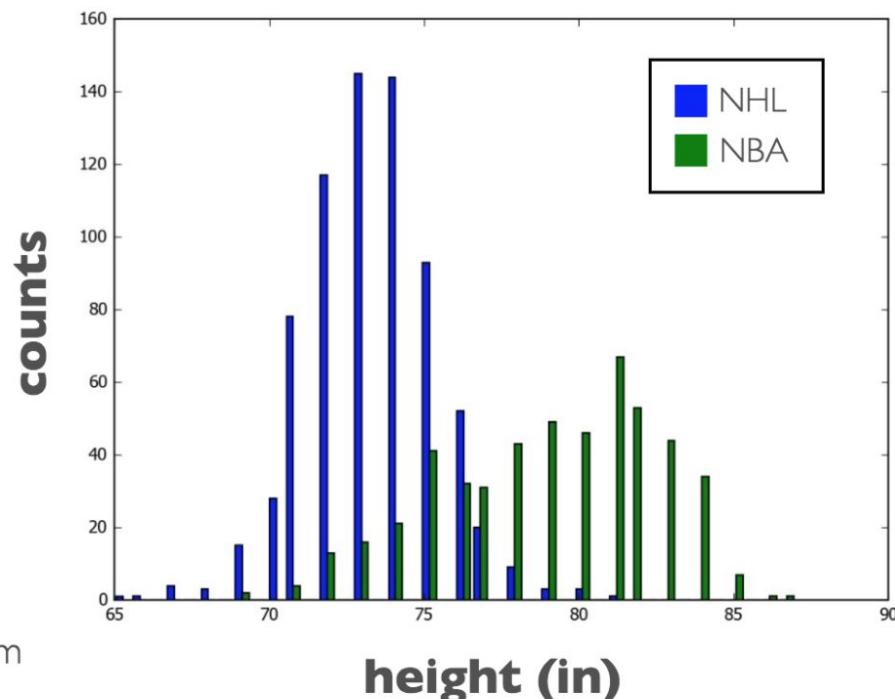
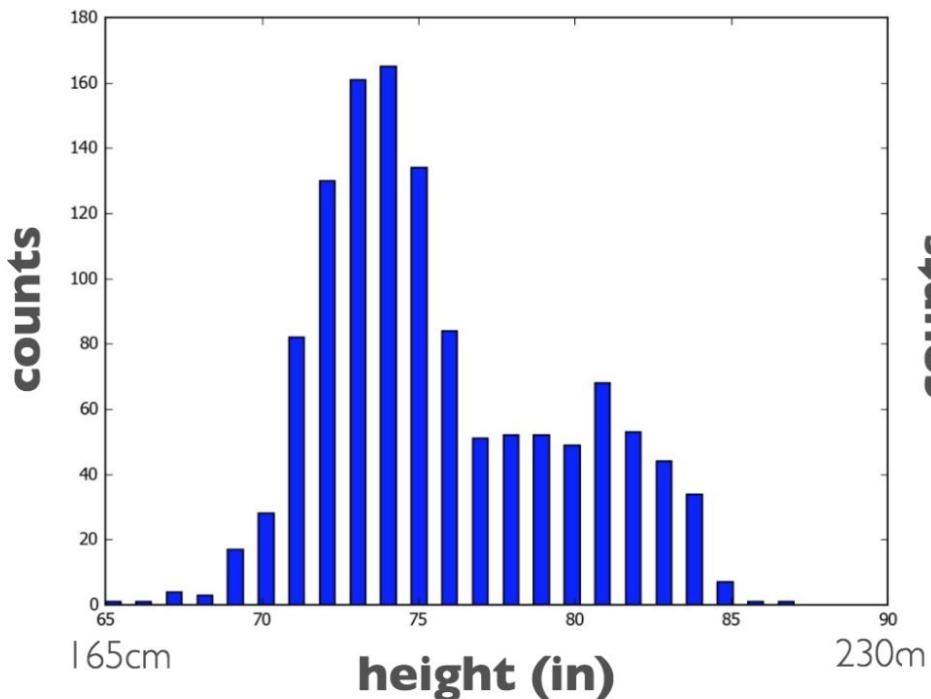
So, how did Fermi do?

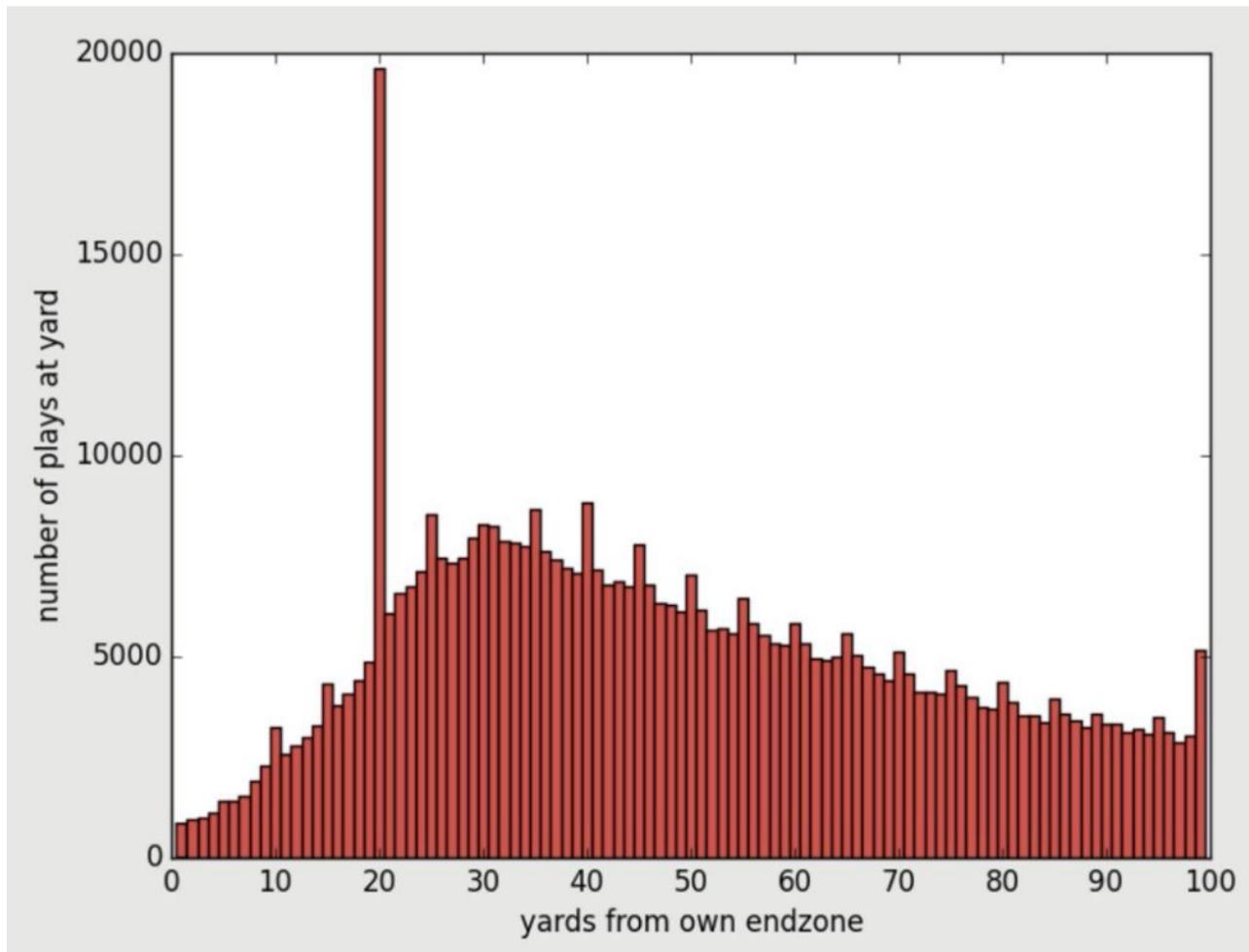
According to the report [**The State of the Global Coatings Industry**](#), the world produced 34 billion liters of paints and coatings in 2012.

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 $1 + \frac{1}{n}$.

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.



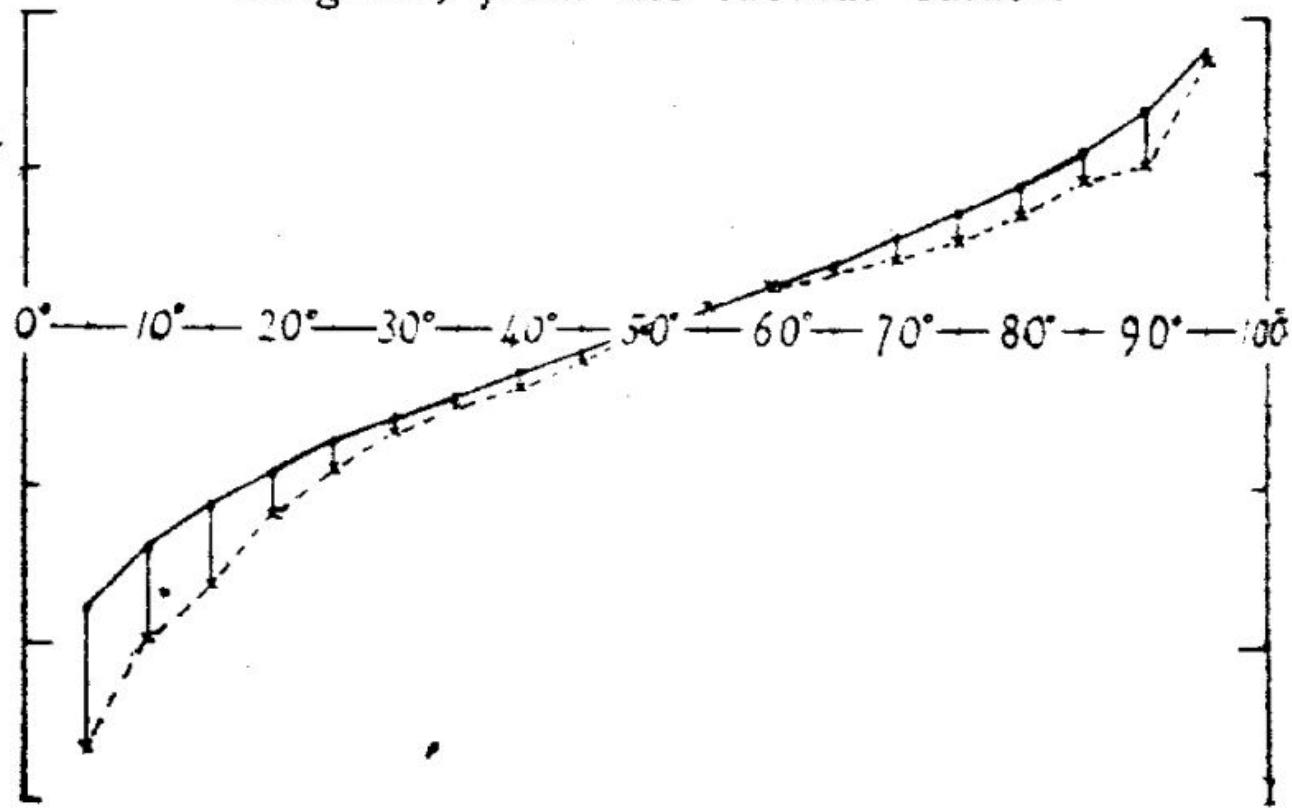


Data Intuition

1. Think about your question and your expectations
2. Do some Fermi calculations (back of the envelope calculations)
3. Write code & look at outputs <- think about those outputs
4. Use your gut instinct / background knowledge to guide you
5. Review code & fix bugs

Diagram, from the tabular values.

Vox Populi



The Wisdom of the Crowds

- Diversity of opinion: Each person should have private information....even if it's just an eccentric interpretation of the known facts
- Independence: People's opinions aren't determined by the opinions of those around them
- Decentralization: People are able to specialize and draw on local knowledge
- Aggregation: Some mechanism exists for turning private judgements into a collective decision

On your own (meaning w/o Googling), please fill out quickly:

<https://forms.gle/CREcpMkYDL>

YTUp2s6

