

# Course Reminders

- Due this Friday (11:59 PM)
  - **PROJECT GROUP SIGNUP**
  - 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)

Q, All Videos News Images Books ; More

Tools

About 282,000 results (0.44 seconds)

www.youtube.com > watch

### How to Create a Personal Access Token in GitHub - YouTube



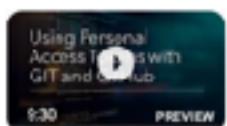
Personal access tokens (PATs) are an alternative to using passwords for authentication to GitHub when using ...

YouTube · CoderDave · Mar 11, 2021

7 key moments in this video

www.youtube.com > watch

### Using Personal Access Tokens with GIT and GitHub - YouTube



A short walk through of how to use Personal Access Tokens to work with GitHub. Written instructions can be ...

YouTube · Ec Goad · Feb 9, 2021

9:30 3 key moments in this video

www.youtube.com > watch

### GitHub password no longer works? GitHub Personal Access ...



Enroll to the 23-hours long Git and GitHub course ... Now for proper interaction with GitHub you need to ...

YouTube · Bogdan Stashchuk · Sep 2, 2021

7 key moments in this video

## Tokens are

- More secure (no dictionary attacks)
- Unique per person or per device
- You can have lots of them, different PATs for different roles in different projects

Our Scott Yang wrote this great HOWTO

<https://docs.google.com/document/d/1Sb6tQwUVBhzcmBGWw4UnhGIYcMDdyUy3gaRKcQzYur4/edit>

# 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 Whatever is MOST popular EVER and whatever is HOTTEST RN on Kaggle

# 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 Group Signup - 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 (9%) - a GitHub repo will be created for your group; ‘submit’ on GitHub (due Fri Week 4)

# Project Proposal (9%)

Full project guidelines are here:

[https://github.com/COGS108/Projects/blob/master/  
FinalProject\\_Guidelines.md](https://github.com/COGS108/Projects/blob/master/FinalProject_Guidelines.md)

# Data tidiness & intuition

Jason G. Fleischer, Ph.D  
UC San Diego

• • •

Department of Cognitive Science  
[jfleischer@ucsd.edu](mailto:jfleischer@ucsd.edu)  
<https://jgfleischer.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

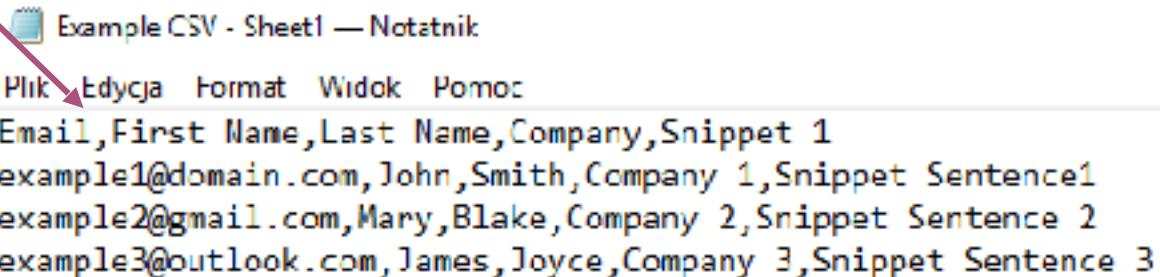
# (Semi-)Structured Data

*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"



Email	First Name	Last Name	Company	Snippet
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

undo redo print preview | 100% | \$ % .0 .00 123 | Arial | 10 | B I S A | field tools

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 Sentence2	
4	example3@outlook.com	James	Joyce	Company 3	Snippet Sentence3	
5						
6	CSV file					
7						
8						

Example CSV - Sheet 1 — 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"}

The diagram illustrates a JSON object consisting of a single key-value pair. The key, 'Name', is highlighted in large black font at the top left. The value, 'Isabela', is highlighted in large black font at the top right. Two pink arrows point from the words 'key' and 'value' at the bottom left and bottom right respectively, towards their corresponding parts in the JSON string.

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  
    },
```

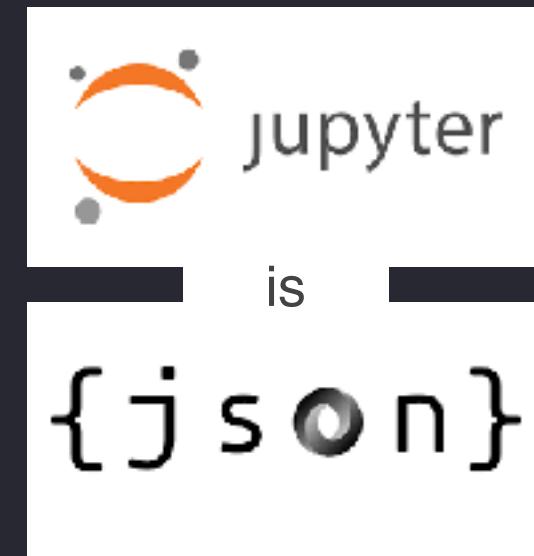
These are all  
nested within  
"Good For"

JSON

emoji\_tone.ipynb

```
{ "cells": [ { "cell_type": "markdown", "metadata": {}, "source": [ "This example represents the output the t-SNE dimensionality reduction algorithm on embeddings computed from Unicode emojis using Keras" ] }, { "cell_type": "code", "execution_count": null, "metadata": {}, "outputs": [], "source": [ "import pandas as pd\n", "import holoviews as hv\n", "hv.extension('bokeh')\n" ] }, { "cell_type": "markdown", "metadata": {}, "source": [ "## Declaring data" ] }, { "cell_type": "code", "execution_count": null, "metadata": {}, "outputs": [] } ] }
```





# Jupyter notebooks suck to version control

<https://nextjournal.com/schmudde/how-to-version-control-jupyter>

```
{
  "cell_type": "code",
  "execution_count": null,
  "metadata": {},
  "outputs": [],
  "source": [
    "import pandas as pd\n",
    "import holoviews as hv\n",
    "hv.extension('bokeh')"
  ]
},
```

A large orange arrow pointing to the right, containing the word "DETOUR" in black capital letters. This graphic serves as a visual metaphor for the alternative approach being suggested.

DETOUR

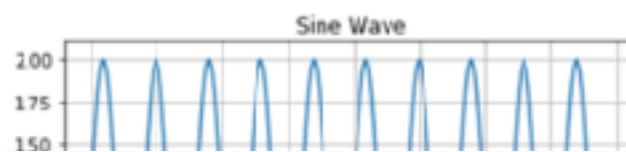
```
In [10]: import numpy as np
import matplotlib.pyplot as plt

# Data for plotting
t = np.arange(0.0, 2.0, 0.01)
s = 1 + np.sin((5 * 2)* np.pi * t)

# Note that using plt.subplots below is equivalent to using
# fig = plt.figure() and then ax = fig.add_subplot(111)
fig, ax = plt.subplots()
ax.plot(t, s)

ax.set(xlabel='time (s)', ylabel='voltage (mV)', title='Sine Wave')
ax.grid()
```

Cut[10]:



"outputs": [

{

  "data": {

    'image/png':

"iVBORw0KGgoAAAANSUhEUgAAAYWAAAECAYAAAB1xKBvAAAABHNCSVQICAgIfAhkiAAAAAlwSFzAAALEgAACxIB0t1+/AAAADl0RVh0U29mdHdcmUAbWF0cGxvdGxpYiB2ZXJzaW9uIDIuMi4yLCBudHRwCi8vbWF0cGxvdGxpYi5vcmlcvhp/UCwAAIABJREFUeJzsvXmcHNd13/s9vc4+2EgABHeQEkkVSXGGRFLembFNSPn7Wyy45i5UXh5ZjvcSy4xcr78WK5bwkzvKSeIlloqaVxZKcOJLN+FHc0dxJEVxxAgQBAiCIdbDP0tPT+80fVdXdmOnl1q17ezBm/T6f+QDdXVXnVtU996z3HFFKESNGjBgxYvRDYrkHECNGjBgxVgZigREjRowYMBQQC4wYMWLEiKGFWGDEiBEjRgwtxAIjRowYMWJoIRYYMWLEiBFDC7HAiBEDEJG/JiKPL/c4YsQ4nxELjBgfGojIXSLyoojMiMg

# Jupyter notebooks suck to version control

<https://nextjournal.com/schmudde/how-to-version-control-jupyter>

## Clear Output Manually

The simplest solution is to always clear the output before committing. **Cell → All Output → Clear → Save**. This removes any binary blobs that have been generated by the notebook. There are three main drawbacks:

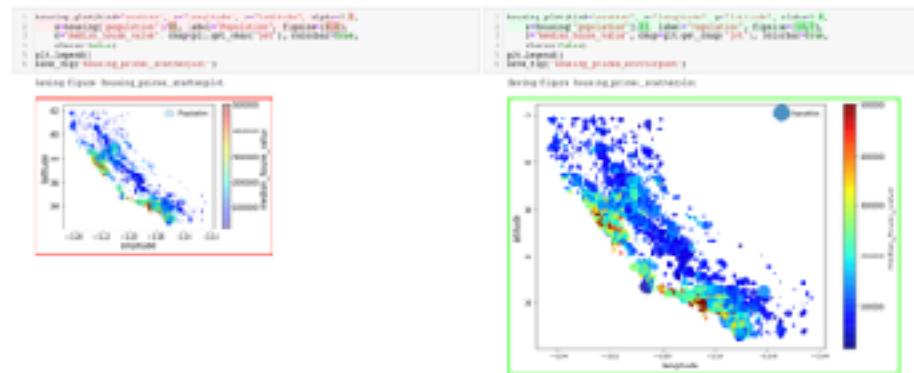
- It is a manual process.
- Collaborators on other machines will need to rerun the notebook to see the output, requiring additional time and setup.

# Jupyter notebooks suck to version control

<https://nextjournal.com/schmudde/how-to-version-control-jupyter>

## ReviewNB

ReviewNB is a GitHub app that also offers visual diffing with an interface that looks similar to the traditional Jupyter IDE. Because the outputs are visualized, problems associated with committing binary blobs disappear.

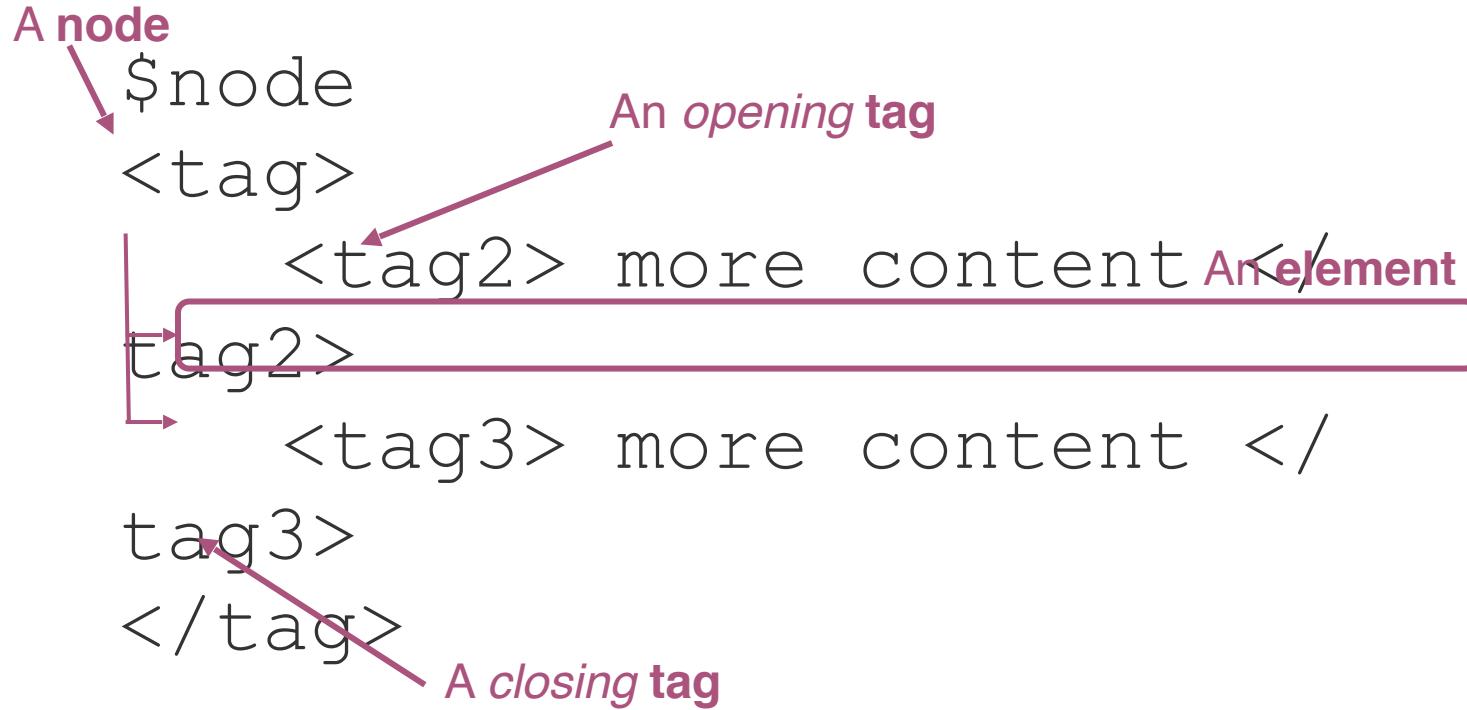


ReviewNB example courtesy of the [ReviewNB website](#)

Back to data formats...

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

*nested/hierarchical data*



XML

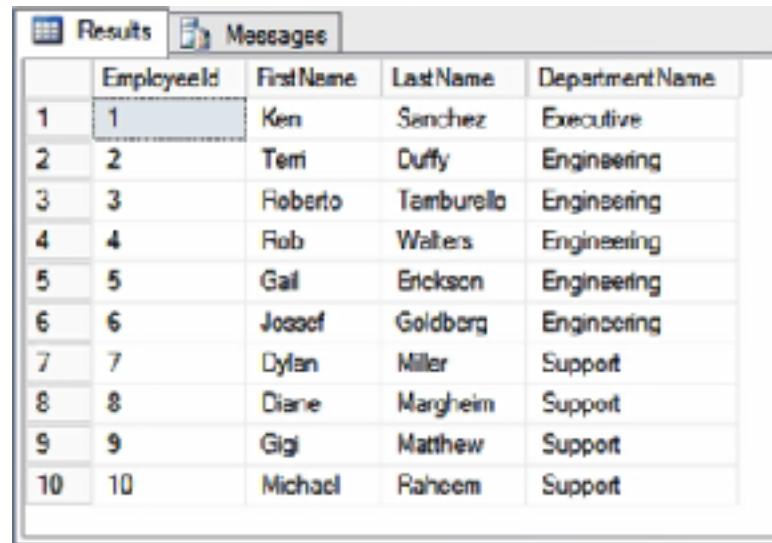
```
<?xml version="1.0" encoding="UTF-8"?>
<customers>
    <customer>
        <customer_id>1</customer_id>
        <first_name>John</first_name>
        <last_name>Doe</last_name>
        <email>john.doe@example.com</email>
    </customer>
    <customer>
        <customer_id>2</customer_id>
        <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 grid displays a table of employee data with the following columns: EmployeeId, FirstName, LastName, and DepartmentName. The data consists of 10 rows, each representing an employee with a unique EmployeeId from 1 to 10, and corresponding FirstName, LastName, and DepartmentName.

	EmployeeId	FirstName	LastName	DepartmentName
1	1	Ken	Sanchez	Executive
2	2	Terri	Duffy	Engineering
3	3	Roberto	Tamburello	Engineering
4	4	Rob	Walters	Engineering
5	5	Gail	Erickson	Engineering
6	6	José	Goldberg	Engineering
7	7	Dylan	Miller	Support
8	8	Diane	Margheim	Support
9	9	Gigi	Matthew	Support
10	10	Michael	Rahiem	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	<b>JJ29JJ</b>	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
<b>JJ29JJ</b>	2018-03-12	D'eonte	98
<b>JJ29JJ</b>	2018-01-02	Monica	66
XJ11AS	2018-12-16	Mark	43
CI21AA	2018-08-21	Anh	99

## rating

id	stars
AH13JK	4.9
<b>JJ29JJ</b>	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

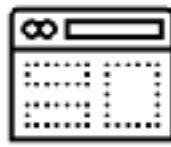
# Unstructured Data

*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



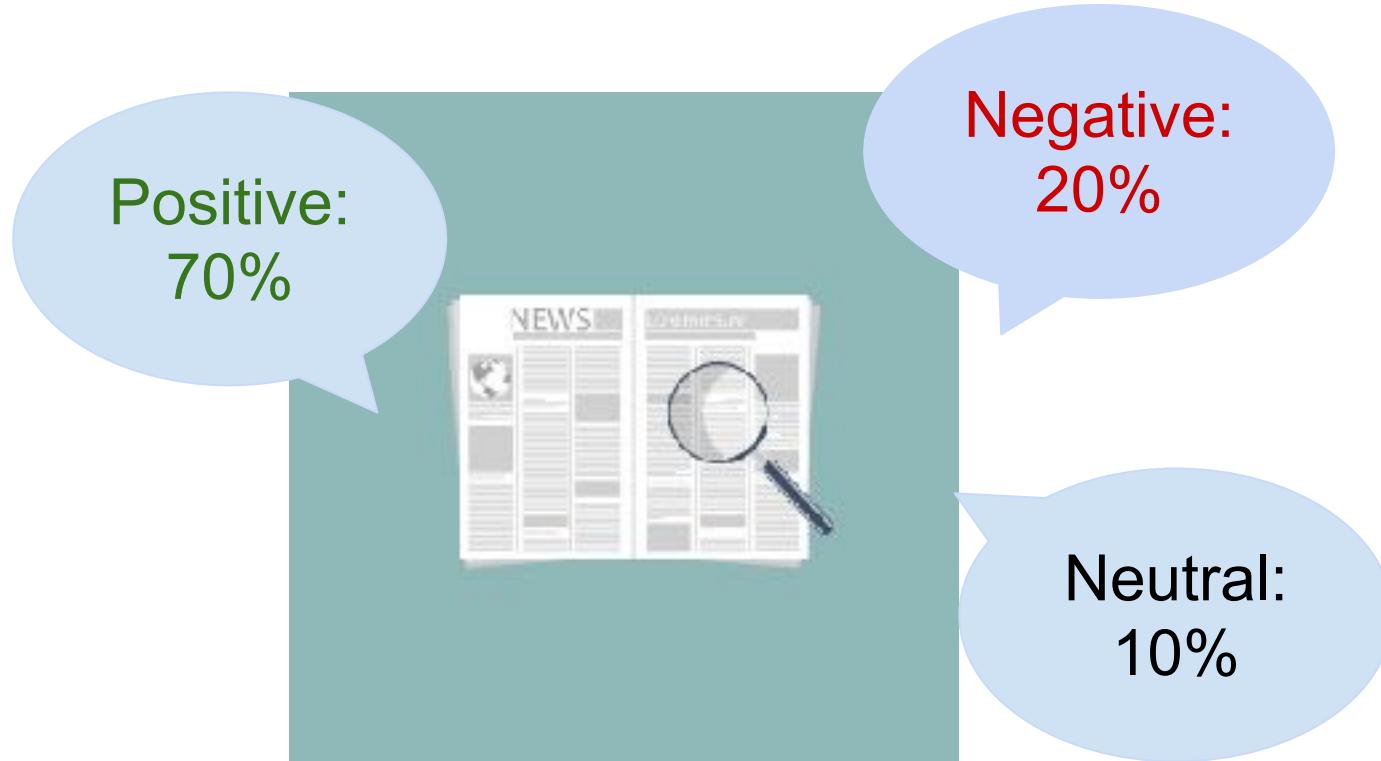
Email  
data

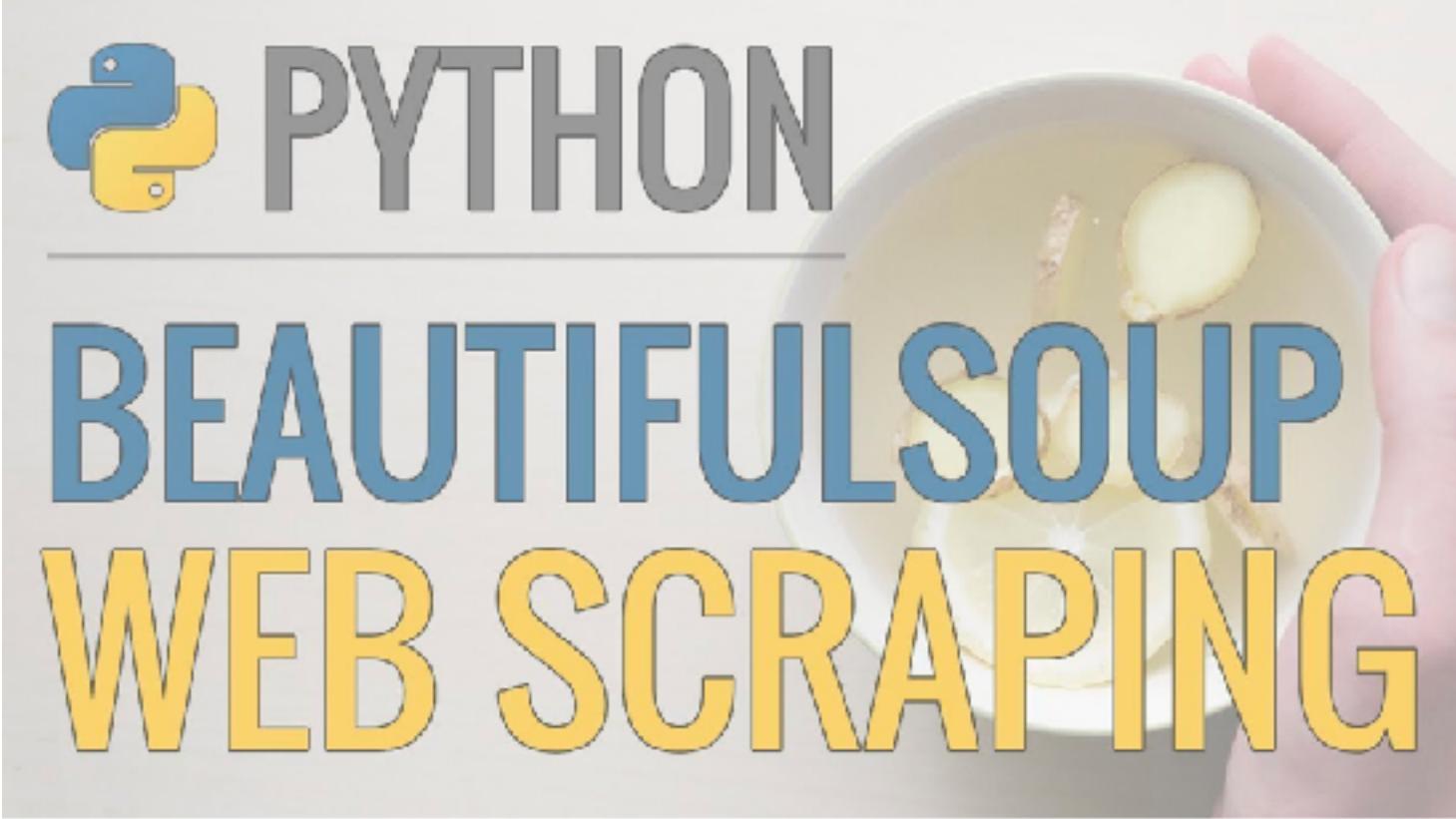


Social  
media  
data



# Text: Sentiment Analysis



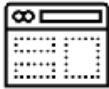


**PYTHON**

---

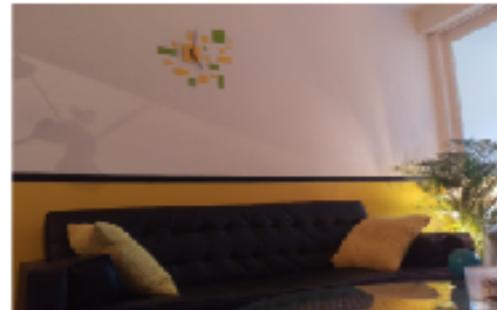
**BEAUTIFULSOUP**

**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



# Australian Bureau of Statistics

## 1800.0 Australian Marriage Law Postal Survey, 2017

Released on 15 November 2017

**Table 5 Participation by Federal Electoral Division(a), Males and Age, Gender apartheid**

Table junk

Yeah NA		15-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60-69 years	70-79 years	80+ years	NA/NA years	NU/NU years	RU/RU years
Lingard(c)	Total participants	292	1,056	1,582	1,682	1,515	1,518	1,710	1,781	1,753	1,753	1,514
Lingard(c)	Eligible participants	572	2,400	3,789	3,996	3,601	3,508	3,645	3,331	2,980	2,458	
Lingard(c)	Participation rate (%)	51.0	36.4	38.7	41.4	42.0	43.2	46.5	51.9	59.2	64.1	
<b>Primary keynotes</b>		Comma on										
<b>Merged cells</b>		Total participants	442	1,461	2,066	2,357	2,188	2,067	2,224	2,108	2,134	1,772
Suburban		Eligible participants	796	2,391	3,994	4,155	3,634	3,398	3,427	3,066	2,931	2,355
		Participation rate (%)	56.5	48.6	51.7	56.7	60.2	60.5	64.5	63.0	72.8	75.2
Northern Territory (TERR)		Total participants	734	2,519	3,531	4,010	3,703	3,573	3,934	3,828	3,887	2,346
		Eligible participants	1,322	5,361	7,783	8,151	7,343	6,364	7,072	6,367	6,891	4,813
		Participation rate (%)	55.5	42.7	45.4	49.2	51.1	51.8	55.6	60.0	66.0	69.5
<b>Australian Capital Territory Divisions</b>		Summary of data inside data										
Canberra(c)		Total participants	1,764	4,789	4,817	4,973	4,626	4,453	5,074	4,826	5,169	4,394
		Eligible participants	2,260	5,471	6,446	6,509	5,983	5,305	6,302	5,902	6,044	5,057
		Participation rate (%)	76.1	74.0	74.7	76.4	77.3	76.7	80.5	81.8	85.5	89.9
Forster(c)		Total participants	1,472	4,587	5,176	5,786	6,025	5,463	5,193	4,208	3,948	3,465
		Eligible participants	1,904	5,354	7,123	7,322	7,960	7,155	6,480	5,206	4,092	3,945
		Participation rate (%)	77.6	73.8	72.2	74.0	76.1	76.4	80.1	80.8	84.1	87.8
		NA Yeah										
Australian Capital Territory (Total)		Total participants	9,243	9,470	9,955	10,759	10,053	9,910	10,145	9,934	9,117	7,659
		Eligible participants	4,164	12,325	13,566	14,331	13,943	12,360	12,782	11,108	10,736	9,002
		Participation rate (%)	77.8	73.9	73.2	71.1	78.4	78.5	80.3	81.3	84.9	87.3
<b>AUSTRALIA</b>												
Total		Total participants	151,297	430,186	441,558	450,546	452,206	479,360	524,620	517,693	543,449	506,799
		Eligible participants	201,435	635,396	646,916	655,250	656,446	659,341	680,050	659,150	664,720	597,366
		Participation rate (%)	75.1	68.5	68.3	69.2	70.4	72.5	75.6	78.5	81.8	84.8

(a) The Federal Electoral Divisions are current as at 24 August 2017

(b) Includes those whose age is unknown

(c) Includes Christmas Island and the Cocos (Keeling) Islands

(d) Includes Norfolk Island

(e) Includes Tasmania

Return of the table junk

# untidy data

Australian Bureau of Statistics										
2010-11 Australian Marriage Laws Postal Survey, 2011										
Assessment by sex, age and state/territory										
Table 1a										
Table 1a										
<b>Table 1a</b> Participation by response disclosure, residence, state and age										
Glossary, notes and footnotes										
<b>Definitions</b>										
Total participants										
Digital participants										
Participation rate (%)										
<b>Primary key variables</b>										
Gender										
State/territory										
Age										
Digital participation										
Participation rate (%)										
<b>Geographic cells</b>										
State/territory										
Local government area										
Postcode										
Local government area (LGA)										
<b>Covariates and Subcelling</b>										
Covariate										
Gender(s)										
Centroid(s)										
Postcode										
Area(s)										
Local government area (LGA)										
Postcode										
<b>NA Values</b>										
NA										
Participation rate (%)										
<b>Australian Capital Territory (ACT)</b>										
Total participants										
Digital participants										
Participation rate (%)										
<b>Queensland (QLD)</b>										
Total participants										
Digital participants										
Participation rate (%)										
<b>South Australia (SA)</b>										
Total participants										
Digital participants										
Participation rate (%)										
<b>Tasmania (TAS)</b>										
Total participants										
Digital participants										
Participation rate (%)										
<b>Victoria (VIC)</b>										
Total participants										
Digital participants										
Participation rate (%)										
<b>Western Australia (WA)</b>										
Total participants										
Digital participants										
Participation rate (%)										
<b>Notes</b>										
1. This table includes responses from all participants.										
2. Includes those aged 18 years and over.										
3. Includes those aged 18 years and over, and those aged 17 years who have been granted a dispensation to vote.										
4. Includes those aged 18 years and over.										
5. Includes those aged 18 years and over.										
6. Includes those aged 18 years and over.										
7. Includes those aged 18 years and over.										
8. Includes those aged 18 years and over.										
9. Includes those aged 18 years and over.										
10. Includes those aged 18 years and over.										
11. Includes those aged 18 years and over.										
12. Includes those aged 18 years and over.										
13. Includes those aged 18 years and over.										
14. Includes those aged 18 years and over.										
15. Includes those aged 18 years and over.										
16. Includes those aged 18 years and over.										
17. Includes those aged 18 years and over.										
18. Includes those aged 18 years and over.										
19. Includes those aged 18 years and over.										
20. Includes those aged 18 years and over.										
21. Includes those aged 18 years and over.										
22. Includes those aged 18 years and over.										
23. Includes those aged 18 years and over.										
24. Includes those aged 18 years and over.										
25. Includes those aged 18 years and over.										
26. Includes those aged 18 years and over.										
27. Includes those aged 18 years and over.										
28. Includes those aged 18 years and over.										
29. Includes those aged 18 years and over.										
30. Includes those aged 18 years and over.										
31. Includes those aged 18 years and over.										
32. Includes those aged 18 years and over.										
33. Includes those aged 18 years and over.										
34. Includes those aged 18 years and over.										
35. Includes those aged 18 years and over.										
36. Includes those aged 18 years and over.										
37. Includes those aged 18 years and over.										
38. Includes those aged 18 years and over.										
39. Includes those aged 18 years and over.										
40. Includes those aged 18 years and over.										
41. Includes those aged 18 years and over.										
42. Includes those aged 18 years and over.										
43. Includes those aged 18 years and over.										
44. Includes those aged 18 years and over.										
45. Includes those aged 18 years and over.										
46. Includes those aged 18 years and over.										
47. Includes those aged 18 years and over.										
48. Includes those aged 18 years and over.										
49. Includes those aged 18 years and over.										
50. Includes those aged 18 years and over.										
51. Includes those aged 18 years and over.										
52. Includes those aged 18 years and over.										
53. Includes those aged 18 years and over.										
54. Includes those aged 18 years and over.</td										

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	4820	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	3726	3726
Adelaide	Female	50-54 years	SA	76	4476	84.7	3791	3791
Adelaide	Female	55-59 years	SA	76	4822	87.3	4033	4033
Australian Bureau of Statistics								
10 Australian Bureau of Statistics Postal Survey, 2012 11 December, 2012								
12 Participation by female household members, 18 years and over, Australia, April 2012								
13 Females								
14 Females								
15 Females								
16 Females								
17 Females								
18 Females								
19 Females								
20 Females								
21 Females								
22 Females								
23 Females								
24 Females								
25 Females								
26 Females								
27 Females								
28 Females								
29 Females								
30 Females								
31 Females								
32 Females								
33 Females								
34 Females								
35 Females								
36 Females								
37 Females								
38 Females								
39 Females								
40 Females								
41 Females								
42 Females								
43 Females								
44 Females								
45 Females								
46 Females								
47 Females								
48 Females								
49 Females								
50 Females								
51 Females								
52 Females								
53 Females								
54 Females								
55 Females								
56 Females								
57 Females								
58 Females								
59 Females								
60 Females								
61 Females								
62 Females								
63 Females								
64 Females								
65 Females								
66 Females								
67 Females								
68 Females								
69 Females								
70 Females								
71 Females								
72 Females								
73 Females								
74 Females								
75 Females								
76 Females								
77 Females								
78 Females								
79 Females								
80 Females								
81 Females								
82 Females								
83 Females								
84 Females								
85 Females								
86 Females								
87 Females								
88 Females								
89 Females								
90 Females								
91 Females								
92 Females								
93 Females								
94 Females								
95 Females								
96 Females								
97 Females								
98 Females								
99 Females								
100 Females								
101 Females								
102 Females								
103 Females								
104 Females								
105 Females								
106 Females								
107 Females								
108 Females								
109 Females								
110 Females								
111 Females								
112 Females								
113 Females								
114 Females								
115 Females								
116 Females								
117 Females								
118 Females								
119 Females								
120 Females								
121 Females								
122 Females								
123 Females								
124 Females								
125 Females								
126 Females								
127 Females								
128 Females								
129 Females								
130 Females								
131 Females								
132 Females								
133 Females								
134 Females								
135 Females								
136 Females								
137 Females								
138 Females								
139 Females								

# 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

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

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

**TIDY** data is **NOT** the same as **CLEAN** data



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



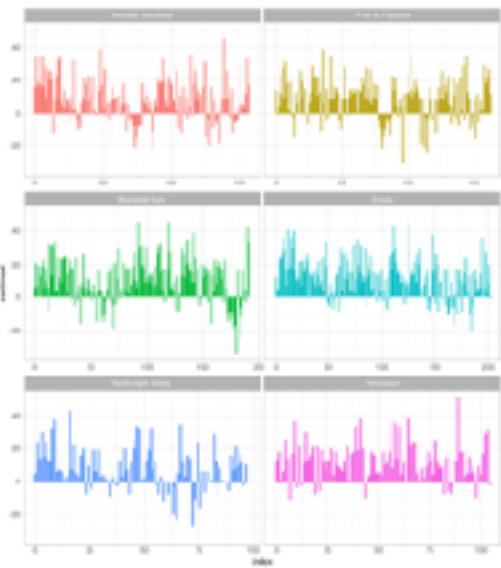
D

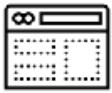


results

tidy dataset

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





# website

1. In the following 100-150 words, discuss the following:  
a) The role of culture and culture shock in understanding the world. How does this relate to the following quote from John Maynard Keynes? "The world has made great progress as regards its material welfare; but it has made very little progress indeed as regards its spiritual welfare."  
b) The relationship between culture and politics.  
c) The relationship between culture and the economy.  
d) The relationship between culture and society.  
e) The relationship between culture and technology.  
f) The relationship between culture and environment.  
g) The relationship between culture and health.  
h) The relationship between culture and education.  
i) The relationship between culture and sports.  
j) The relationship between culture and arts.

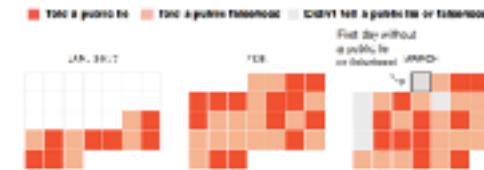


# tidy dataset

Self	It	Explanation	Self
8 Jan 21 2017	I wanted a lot of likes, I didn't want to do it...	He made 12 Instagram posts and 10 posts against it.	How persuasive others considered as cynical X.
1 Jan 21 2017	A woman for Time magazine and I spoke to her...	Trump was on the cover 11 times and Biden 9 times.	https://medium.com/@UWQUOTEDThings-at-
3 Jan 21 2017	Biden was 1 billion and 1 million steps older	TRUMP IS 100 MILLION OF steps longer.	https://www.nytimes.com/2017/01/07/science/bidens-1-
4 Jan 21 2017	"...and the president was the biggest ever IDOL..."	Official presidential inauguration counts from Obama's 2009 inauguration.	https://www.nytimes.com/2017/01/07/science/bidens-1-
5 Jan 21 2017	"...lets a lot of the other experts weigh in about..."	The most recent presidential inauguration.	https://www.nytimes.com/2017/01/07/science/bidens-1-

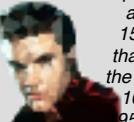


## results



# text (lyrics)

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

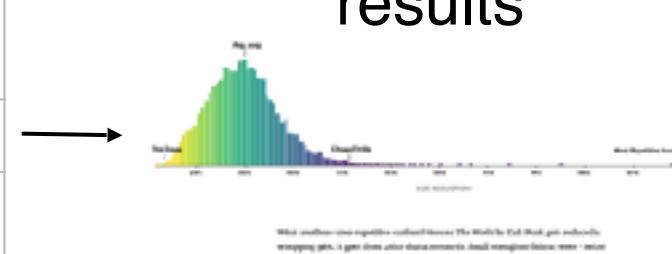


Are Pop Lyrics Getting More Repetitive?

## tidy dataset

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

## results



# 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:



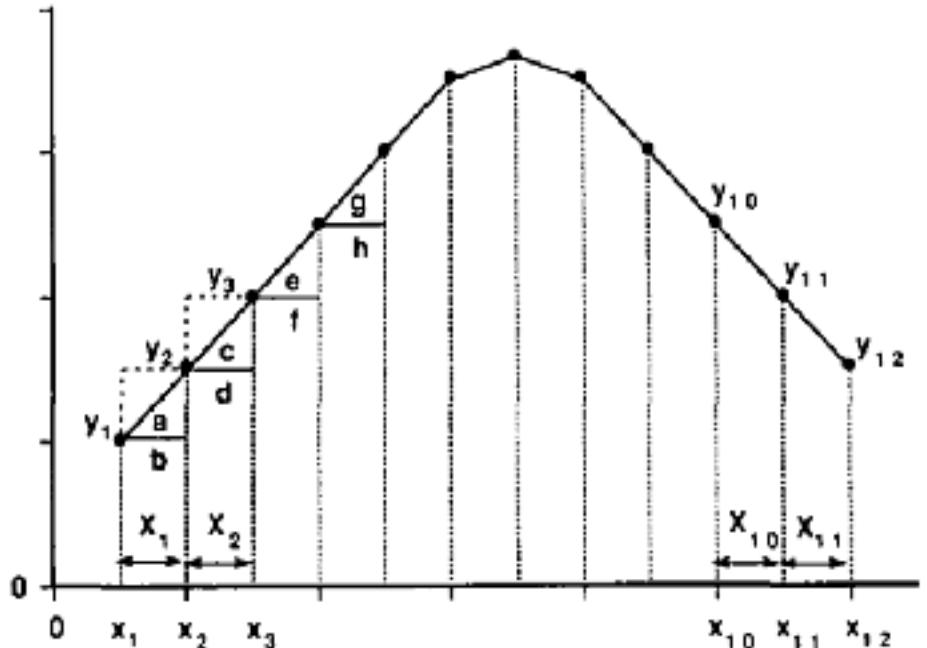
1375

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”

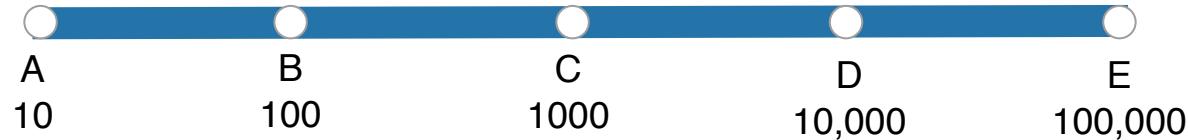
**Figure 1—**Total area under the curve is the sum of individual areas of triangles  $a, c, e$ , and  $g$  and rectangles  $b, d, f$ , and  $h$ .



# 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=0YzvupOX8ls>

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

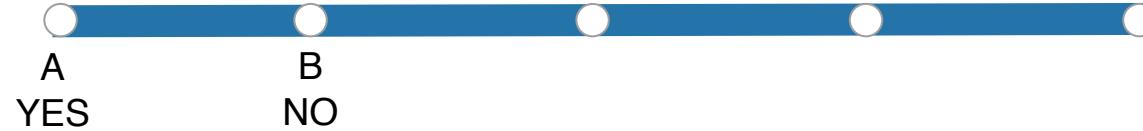
**—Josh (Bolton, MA)**

# Fermi Estimation

<https://forms.gle/shS84W1tai4SDrVF9>



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

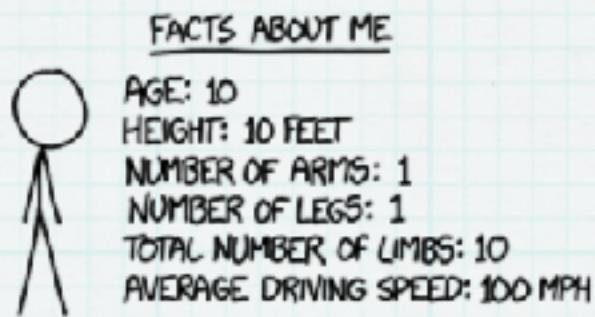




This answer is pretty straightforward. We can look up the size of the world's paint industry, extrapolate backward to figure out the total amount of paint produced. We'd also need to make some assumptions about how we're painting the ground. Note: When we get to the Sahara desert, I recommend not using 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,<sup>[2]</sup> 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.<sup>[3]</sup> 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,<sup>[4]</sup> 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.<sup>[6]</sup> That comes out to a little over a trillion liters of paint. At 30 square meters per gallon,<sup>[7]</sup> 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

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

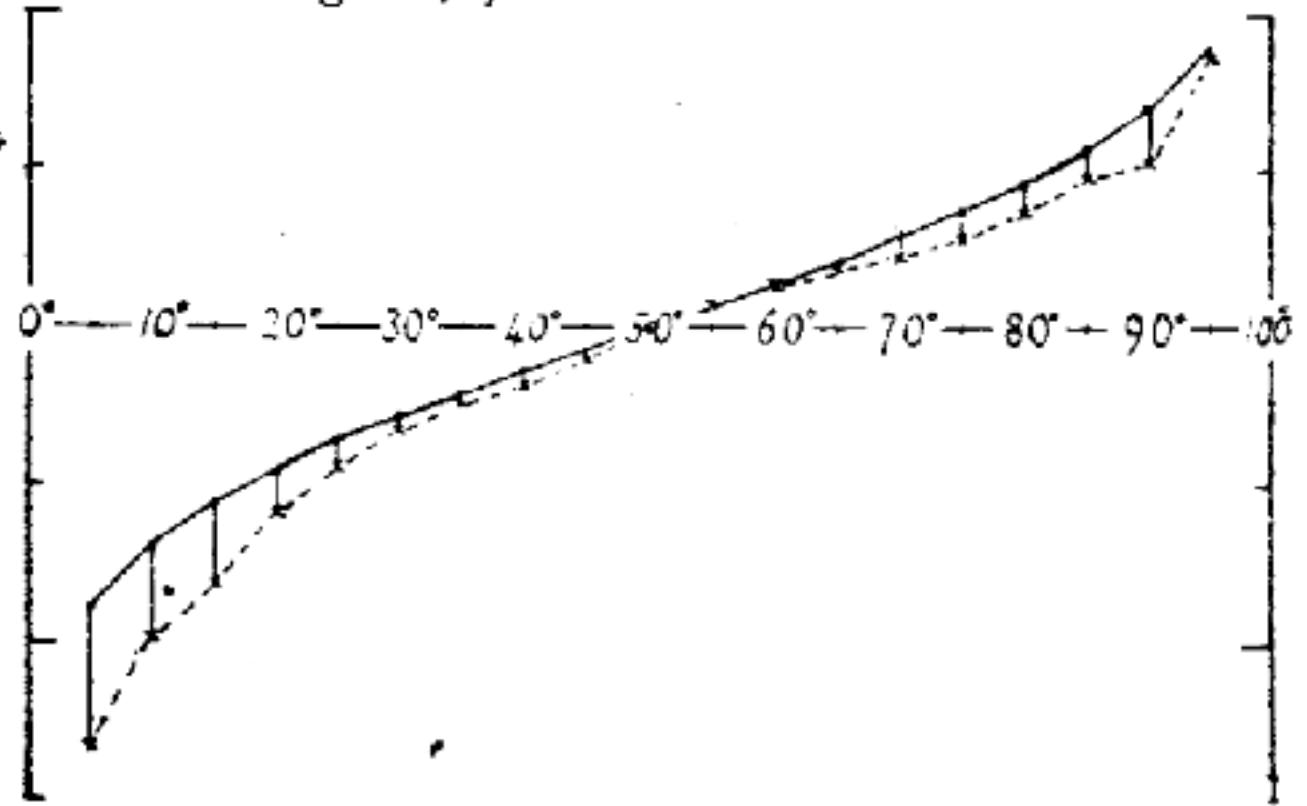
<https://forms.gle/CREcpMkYDLYTUp2s6>



Other kinds of  
guessing and  
intuitions

*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

