

# Course Reminders

- Due today (11:59PM)
  - Q1 (covers lectures last week)
- Due this Wednesday (11:59 PM)
  - Pre-course survey
  - Practice Assignment
- Due this Friday (11:59 PM)
  - D1
  - #FinAid quiz on Canvas

# DS3 Fall Kickoff Week



- WED** | Meet the Board 12 - 2 pm, Red Shoe Room ⑪  
10.11 | Fall GBM 6 - 8 pm, PC Ballroom West ⑪
- THU** | Data Science Talent Day 1  
10.12 | 11 - 2 pm, PC Ballroom East
- FRI** | Data Science Talent Day 2  
10.13 | 10 - 12 pm, Virtual  
Cloud Powered Data Science Part 1 ⑪  
3:30 - 4:30 pm, CSE 1202
- SAT** | Bonfire Social ⑪  
10.14 | 4 - 8 pm, La Jolla Shores

⑪ Free Food!!

# 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

# Structured Data

*Databases!*

*What is a DB?? An organized collection of structured information*

*Manage huge datasets*

*Control access to data*

*Allow users to find a subset of data with a query*

*Run an analysis on data inside the DB and return a report*

# Structured Data

## *Examples of relational DB*

- *SQLite*
- *MySQL*
- *Postgres*

*Relational DB work using tables of data with “relationships” established between tables*

## *Examples of non-relational DB*

- *Hadoop*
- *Hive*
- *Apache CouchBase*

*NoSQL DB work with key-value pairs to lookup data, and that’s exactly like JSON slides coming up.*

# (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

Has the  
extension  
“.csv”

Each  
column  
separated  
by a  
comma

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](#)

| 100% | \$ % .0+ .00 123 | Arial | 10 | **B** *I* A

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	
5						
6						
7						
8						

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

```
{
  "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,
```



is

{j s o n}

# 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')"  
  ]  
},
```



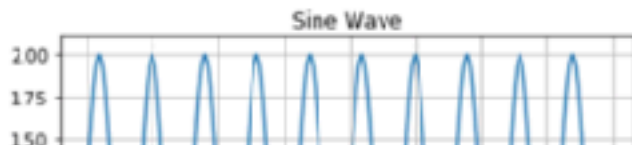
```
In [10]: import numpy as np
import matplotlib.pyplot as plt

# 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":

"iVBORw0KGgoAAAANSUhEUgAAAYwAAAEWCAYAAAB1xKBvAAAAABHNCSVQICAgIfAhkiAAAAAlwSFlzAAALEgAACxIB0t1+/AAAADl  
ORVhOU29mdHdhcmUAAbWF0cGxwdGxpYiB2ZXJzaW9uIDIuMi4yLCBodHRwOiBvbWw0cGxwdGxpYi5vcmlvhp/UCwAAIABJREFUeJz  
svXmcHNd13/s9vc4+2EgABHeQEkVSXGGRFLembFNSPn7Wyy45i5UXh5ZjvcSy4xcr78WK5bkwzvKSeIll0qaVxZKcOJLN-FHc0dx  
JEVxAAgQBAiCiIdbDP0tPT+80fVdXdm0nllq17ezBm/T6f+QDdXVXnVtU996z3HFFKESNGjBgxYvRDYrkHECNGjBgxVgZigREjRow  
YMbQQC4wYMWLEiKGFwGDEiBEjRgwtxAIjRowYMWJoIRYYMWLEiBFDC7HAiBEDEJG/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:

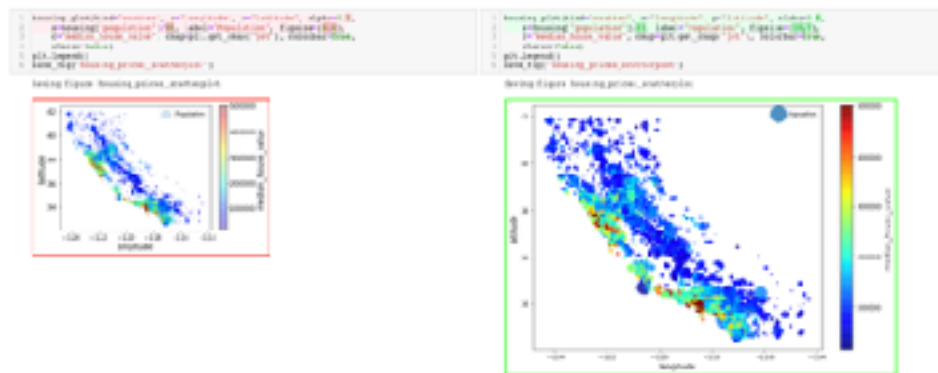
- It is a manual process.
- Collaborators on other machines will need to rerun the notebook to see the output, requiring additional time and setup.

# 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*

**A node**

\$node

<tag>

**An opening tag**

<tag2> more content </

**An element**

tag2>

<tag3> more content </

tag3>

</tag>

**A closing tag**

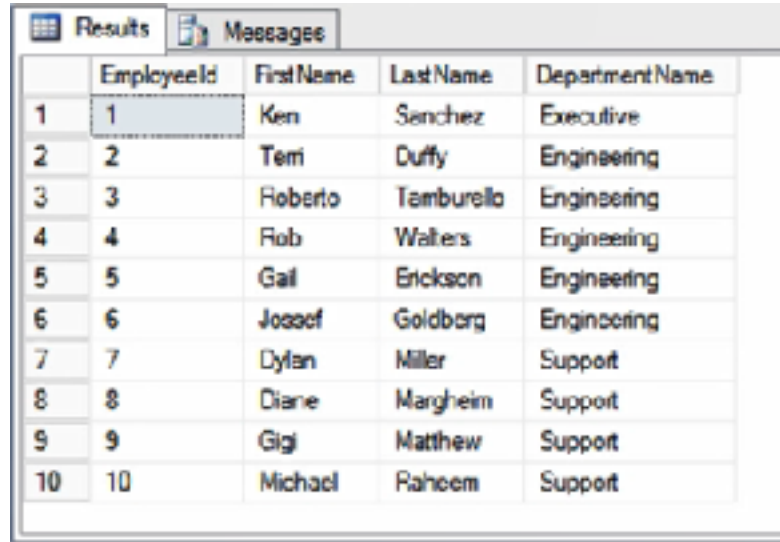
# XML

```
<?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

# Relational Databases: A set of interdependent tables

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

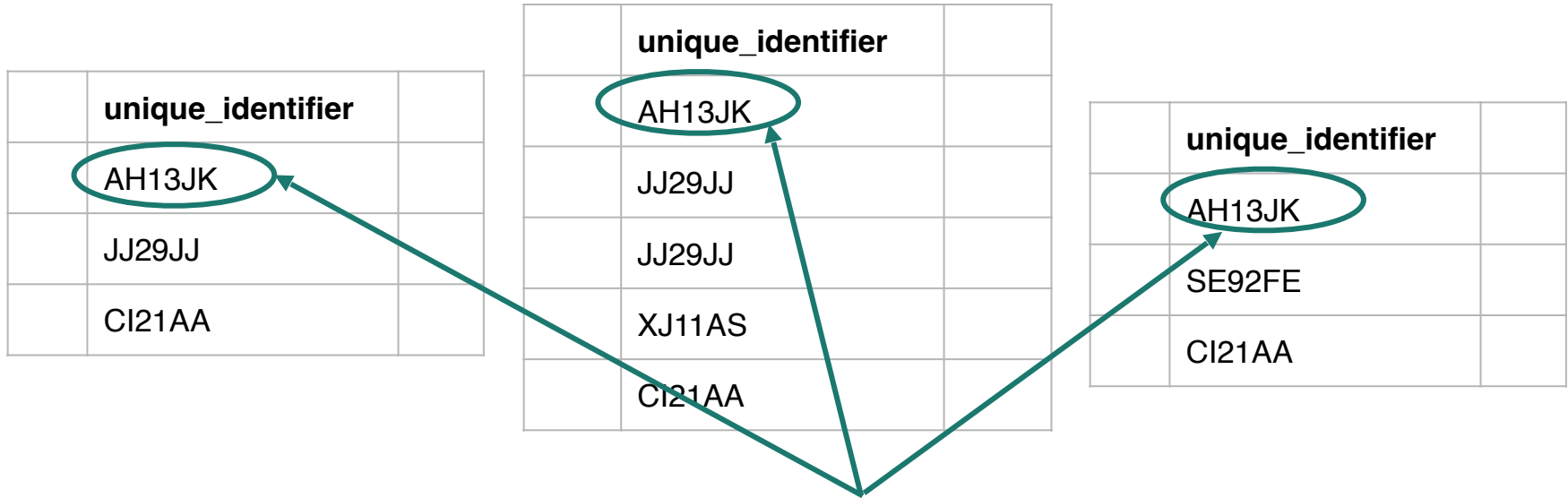


A screenshot of a database application window. At the top, there are two tabs: 'Results' and 'Messages'. The 'Results' tab is active, displaying a table with five columns: 'EmployeeId', 'FirstName', 'LastName', and 'DepartmentName'. The table contains 10 rows of data. The first row is highlighted with a dashed border. The data is as follows:

	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	Josef	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



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

## 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

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

relational database

# Unstructured Data

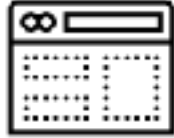
*Some datasets record information about the state of the world, but in a more heterogeneous way. Perhaps it is a large text corpus with images and links like Wikipedia, or the complicated mix of notes and test results appearing in personal medical records.*



# Unstructured Data Types



Text files  
and  
documents



Websites  
and  
applications



Sensor  
data



Image  
files



Audio  
files



Video  
files



Email  
data



Social  
media  
data



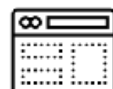
Positive:  
70%

Negative:  
20%

Neutral:  
10%



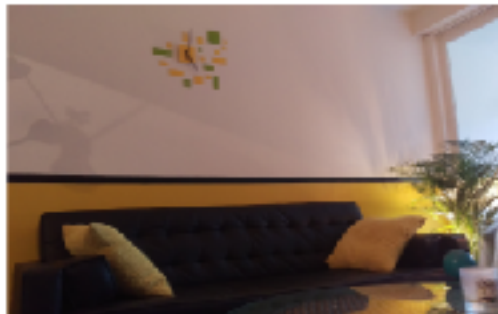
Text:  
Sentiment Analysis



# 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

Table junk

## 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

Year NA		15-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60-69 years	70-79 years	80-89 years	90+ years
Lingard (F)	Total participants	292	1,088	1,888	1,653	1,515	1,308	1,110	1,130	1,314
	Eligible participants	572	2,400	3,720	3,496	3,603	3,406	3,645	3,333	2,488
	Participation rate (%)	51.0	36.4	38.7	41.4	42.0	43.2	46.9	51.9	64.1
Merged cells		442	1,461	2,066	2,357	2,186	2,057	2,224	2,106	1,712
Suburban	Total participants	442	1,461	2,066	2,357	2,186	2,057	2,224	2,106	1,712
	Eligible participants	750	2,991	3,994	4,155	3,634	3,398	3,427	3,666	2,931
	Participation rate (%)	58.9	48.8	51.7	56.7	60.2	60.5	64.5	68.8	75.2
Northern Territory (TERR)	Total participants	734	2,519	3,531	4,000	3,793	3,573	3,934	3,838	3,346
	Eligible participants	1,332	5,961	7,783	8,151	7,243	6,964	7,072	6,367	4,811
	Participation rate (%)	55.5	42.1	45.4	49.2	51.1	51.8	55.6	60.0	69.5
Covariate as subheading		1,764	4,789	4,817	4,973	4,626	4,453	5,074	4,826	4,394
Canberra (C)	Total participants	1,764	4,789	4,817	4,973	4,626	4,453	5,074	4,826	4,394
	Eligible participants	2,260	5,471	6,448	6,589	5,983	5,585	6,302	5,902	5,057
	Participation rate (%)	78.1	84.0	74.7	76.4	77.3	79.7	80.5	81.8	86.9
Farrer (C)	Total participants	1,472	4,587	5,176	5,786	6,025	5,463	5,191	4,206	3,946
	Eligible participants	1,904	5,354	7,123	7,322	7,960	7,155	6,486	5,206	4,692
	Participation rate (%)	77.6	85.8	72.7	79.0	75.7	76.4	80.1	80.8	84.1
NA Year		3,243	9,476	9,942	10,759	10,693	9,926	10,185	9,634	9,117
Australian Capital Territory (Total)	Total participants	3,243	9,476	9,942	10,759	10,693	9,926	10,185	9,634	9,117
	Eligible participants	4,164	12,325	13,545	14,331	13,943	13,260	12,782	11,108	10,736
	Participation rate (%)	77.8	76.9	73.4	75.1	76.4	75.6	80.3	87.3	84.9
Australia		251,297	433,166	441,556	469,546	452,206	479,360	524,620	517,693	543,449
Total	Total participants	251,297	433,166	441,556	469,546	452,206	479,360	524,620	517,693	543,449
	Eligible participants	201,435	635,965	646,916	695,250	650,446	696,341	680,050	659,150	664,720
	Participation rate (%)	75.1	68.5	68.3	68.2	70.4	72.5	75.6	78.5	81.8

(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 Jervis Bay

Return of the table junk

## untidy data

[illegible]

tidy data

#	area	gender	age	obs	obs (by ind)	Eligible participants	Participated on-site (N)	Excl. participants	Total Participants
1	astoria	female	18-24 years	66	46	66	66	0	66
2	astoria	female	25-34 years	66	76	66	66	0	66
3	astoria	female	35-44 years	66	76	66	66	0	66
4	astoria	female	45-54 years	66	76	66	66	0	66
5	astoria	female	55-64 years	66	76	66	66	0	66
6	astoria	female	65-74 years	66	76	66	66	0	66
7	astoria	female	75-84 years	66	76	66	66	0	66
8	astoria	female	85-94 years	66	76	66	66	0	66
9	astoria	female	95-104 years	66	76	66	66	0	66
10	astoria	female	105-114 years	66	76	66	66	0	66
11	astoria	female	115-124 years	66	76	66	66	0	66
12	astoria	female	125-134 years	66	76	66	66	0	66
13	astoria	female	135-144 years	66	76	66	66	0	66
14	astoria	female	145-154 years	66	76	66	66	0	66
15	astoria	female	155-164 years	66	76	66	66	0	66
16	astoria	female	165-174 years	66	76	66	66	0	66
17	astoria	female	175-184 years	66	76	66	66	0	66
18	astoria	female	185-194 years	66	76	66	66	0	66
19	astoria	female	195-204 years	66	76	66	66	0	66
20	astoria	female	205-214 years	66	76	66	66	0	66
21	astoria	female	215-224 years	66	76	66	66	0	66
22	astoria	female	225-234 years	66	76	66	66	0	66
23	astoria	female	235-244 years	66	76	66	66	0	66
24	astoria	female	245-254 years	66	76	66	66	0	66
25	astoria	female	255-264 years	66	76	66	66	0	66
26	astoria	female	265-274 years	66	76	66	66	0	66
27	astoria	female	275-284 years	66	76	66	66	0	66
28	astoria	female	285-294 years	66	76	66	66	0	66
29	astoria	female	295-304 years	66	76	66	66	0	66
30	astoria	female	305-314 years	66	76	66	66	0	66
31	astoria	female	315-324 years	66	76	66	66	0	66
32	astoria	female	325-334 years	66	76	66	66	0	66
33	astoria	female	335-344 years	66	76	66	66	0	66
34	astoria	female	345-354 years	66	76	66	66	0	66
35	astoria	female	355-364 years	66	76	66	66	0	66
36	astoria	female	365-374 years	66	76	66	66	0	66
37	astoria	female	375-384 years	66	76	66	66	0	66
38	astoria	female	385-394 years	66	76	66	66	0	66
39	astoria	female	395-404 years	66	76	66	66	0	66
40	astoria	female	405-414 years	66	76	66	66	0	66
41	astoria	female	415-424 years	66	76	66	66	0	66
42	astoria	female	425-434 years	66	76	66	66	0	66
43	astoria	female	435-444 years	66	76	66	66	0	66
44	astoria	female	445-454 years	66	76	66	66	0	66
45	astoria	female	455-464 years	66	76	66	66	0	66
46	astoria	female	465-474 years	66	76	66	66	0	66
47	astoria	female	475-484 years	66	76	66	66	0	66
48	astoria	female	485-494 years	66	76	66	66	0	66
49	astoria	female	495-504 years	66	76	66	66	0	66
50	astoria	female	505-514 years	66	76	66	66	0	66
51	astoria	female	5	66	46	66	66	0	66

data



wrangling



1	area	gender	age	State	Area (sq km)	Eligible participants	Participation rate (%)	Total participants	Total Participants
2	Adelaide	Female	18-19 years	SA	76	1341	83.5	1120	1120
3	Adelaide	Female	20-24 years	SA	76	4620	81.2	3750	3750
4	Adelaide	Female	25-29 years	SA	76	4897	81.8	4004	4004
5	Adelaide	Female	30-34 years	SA	76	4784	79.8	3820	3820
6	Adelaide	Female	35-39 years	SA	76	4319	79	3411	3411
7	Adelaide	Female	40-44 years	SA	76	4310	80.6	3472	3472
8	Adelaide	Female	45-49 years	SA	76	4579	81.4	3728	3728
9	Adelaide	Female	50-54 years	SA	76	4476	84.7	3791	3791
10	Adelaide	Female	55-59 years	SA	76	4622	87.8	4033	4033

# Australian Bureau of Statistics

2000-01 Australian Marriage, Divorce and Family Survey 2017

Released on 15 September 2017

Rate of participation by gender, residence, age group, and age group

TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence, age group, and age group											
TABLE 1											
Rate of participation by gender, residence,											

# 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	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

Doctor's Office Measurements Data

	A	D	E	F	G
1	ID	Height_inches	Weight_lbs	Insulin	Glucose
2	1004	65	190	0.60	163
3	4587	75	215	1.46	150
4	1727	62	124	0.72	177
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	180	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

# Clean Data

Data cleaning is the process of detecting and correcting/  
removing data records that are

- incomplete,
- incorrect,
- inaccurate, or
- irrelevant

and often includes making sure missing data is correctly marked  
and that duplicate records and other errors are removed

Data that is **TIDY** and **CLEAN** is ready for use

---





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	

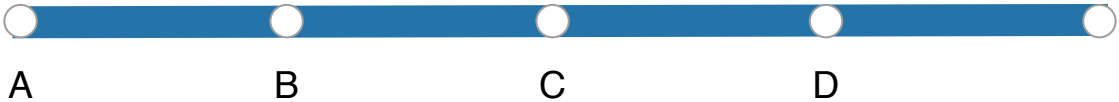
C

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?





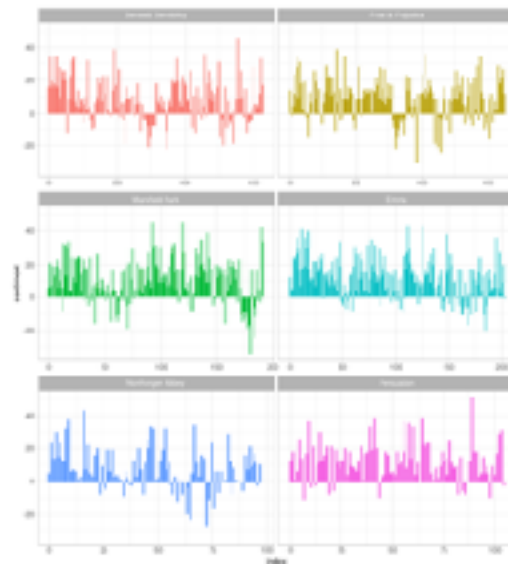
 text

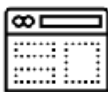


tidy dataset

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

results





website

tidy dataset

date	lie	explanation	url
Jan 21 2017	I want a lion of his I didn't want to get it...	He said that's a lion before he was exposed.	His alternative description of a lion is a... 30...
Jan 21 2017	Answer for Time magazine... and I	Trump was on the cover of Time and News week...	His alternative description of a lion is a... 30...
Jan 21 2017	Between London and a million other cities	TRUMP TO KNOWS OF legal setting	TRUMP TO KNOWS OF legal setting
Jan 21 2017	Now, the president who has helped our	Official press photos show Obama's 2009 inauguration...	Official press photos show Obama's 2009 inauguration...
Jan 21 2017	She is just at the other end of the world	The next time she's mentioned her name...	https://www.nytimes.com/2017/01/23/us/politics/...

results

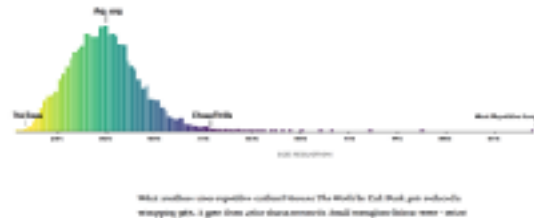


## text (lyrics)

## 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





1011



1375



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

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:

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

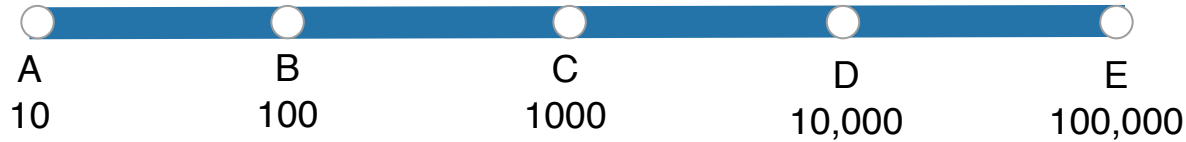
---

# Fermi Estimation

<https://forms.gle/C982naWtU9RvHqAb7>



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









# 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 <sup>[1]</sup> 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. [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,<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 <sup>[5]</sup> 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>[2]</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>

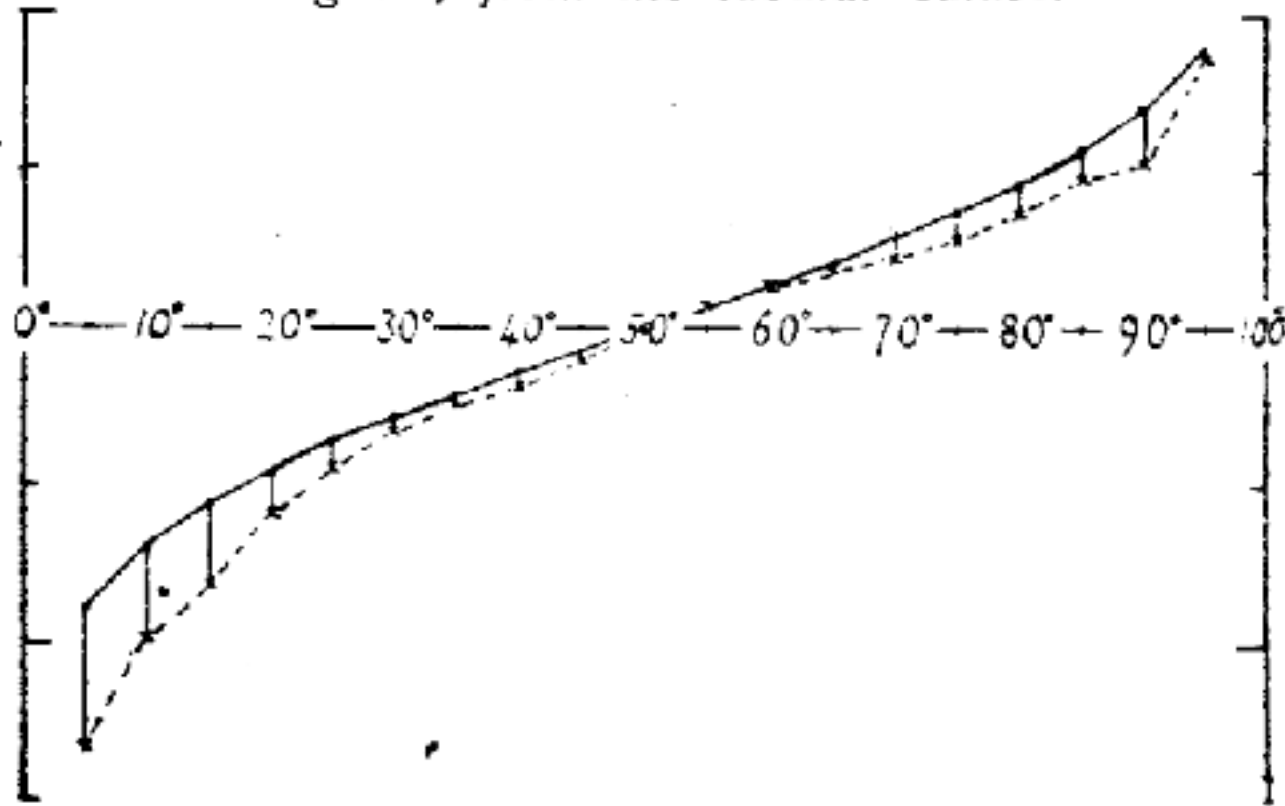
Other kinds of  
guessing and  
intuitions

---



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

