



HKUST
VISLAB

COMP 4462

Data Visualization Tutorial

Leo Yu Ho, Lo
Ming Yao

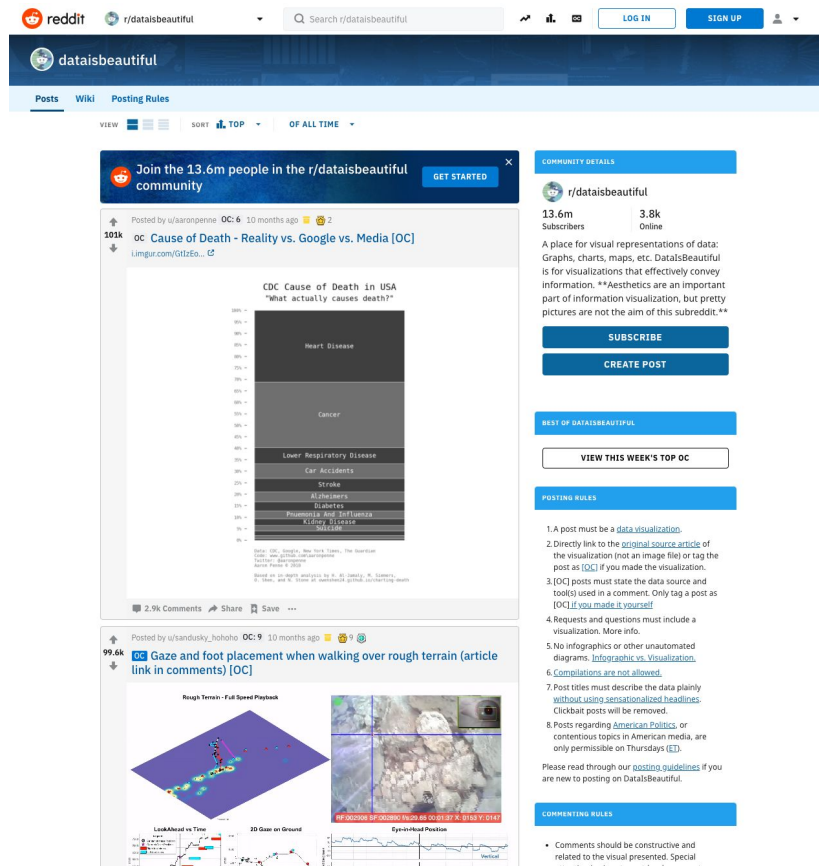
Tuesday 26 February, 2019
<https://bit.ly/vis-t03>

Course Project & Top-Vis Competition

- Project (50%)
 - Grouping: [Sign up sheet](#) (27 Feb)
 - Phase 1: Proposal presentation (27 Mar & 29 Mar)
 - Find a dataset
 - What kind of data? How large is it? Why this dataset?
 - Visualization tasks / data processing / visual encoding
 - Good to have a mock up
 - Phase 2: Project presentation (3 May & 8 May)
 - Make it real! Coding & demo
 - Share stories in the data
- Top-Vis Competition (10%)
 - 2 mins to present 2 visualizations
 - 24 Apr & 26 Apr
 - Write up a short essay
 - Why you chose this visualization?
 - What data are visualized? How are they encoded?

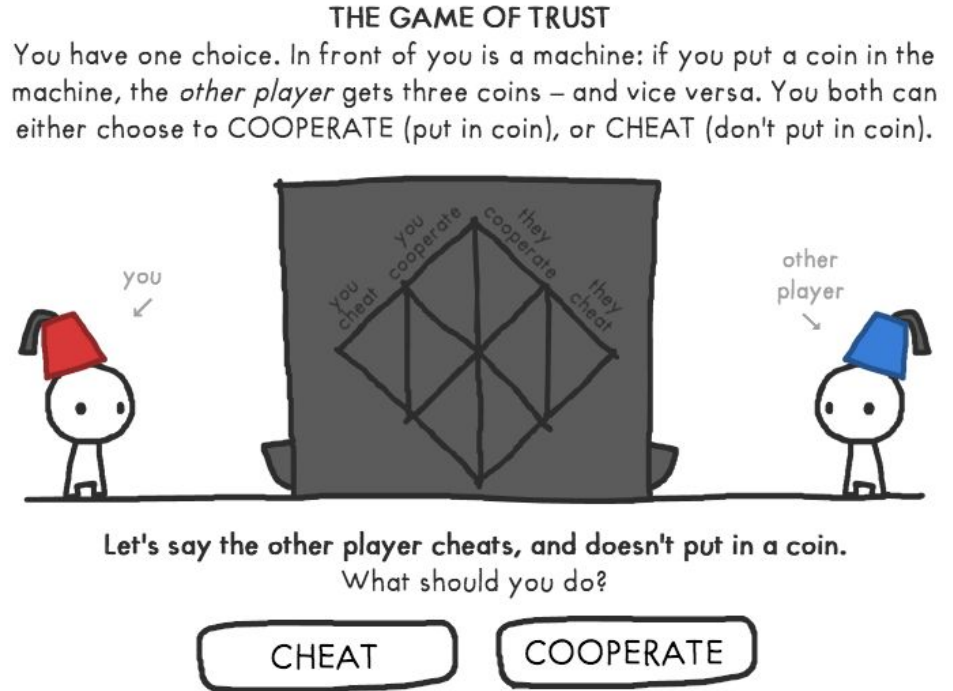
Data is Beautiful

- New visualizations everyday
- Top post of all time
 - Visualization with highest voting of all time
- A lot of remarkable ideas
- Mainstream:
 - Meaning of data > visual effect
 - And some are visually impressive
- Another subreddit: Data is Ugly
 - Lying with charts
 - Deceiving, scam
 - Some are from very authoritative sources
 - Famous news websites
 - Governments
 - Famous companies



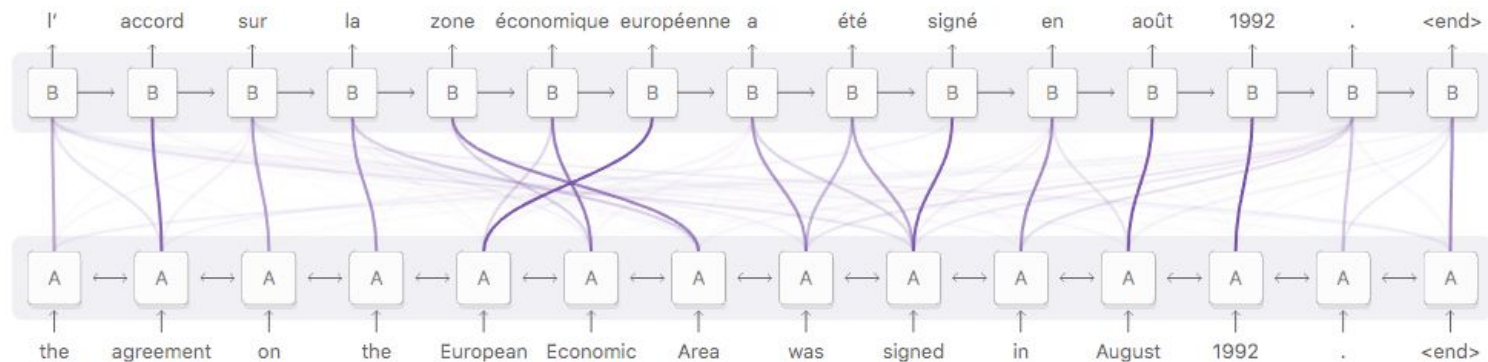
Nick Case

- Narrative visualizations
 - Telling a story with visualizations
- Evolution of Trust
 - Game theory about our society
 - Prisoner dilemma
 - CHEAT?
 - COOPERATE?
 - Interactive
 - Nice graphics and music
 - A sandbox simulator at the end
 - Enjoy!
- More on [Nick Case's webpage](#)



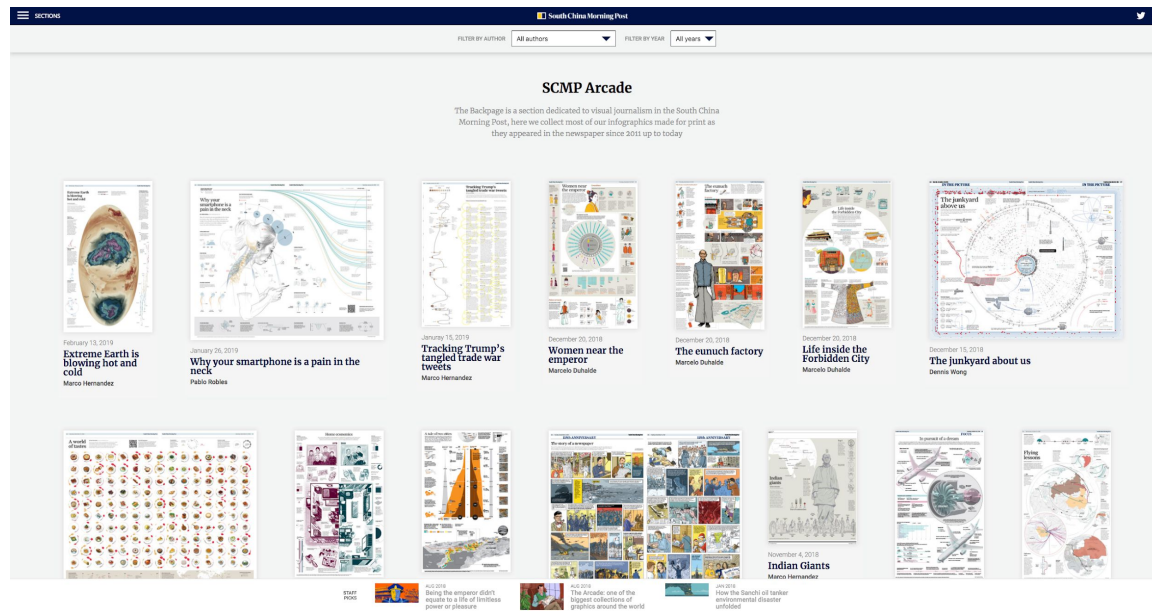
Distill

- Visual Explanation of Machine Learning Algorithms
- Attention and Augmented Recurrent Neural Networks
 - Visualizing a neural translation model
 - Which word in a French sentence \Leftrightarrow which word in English?



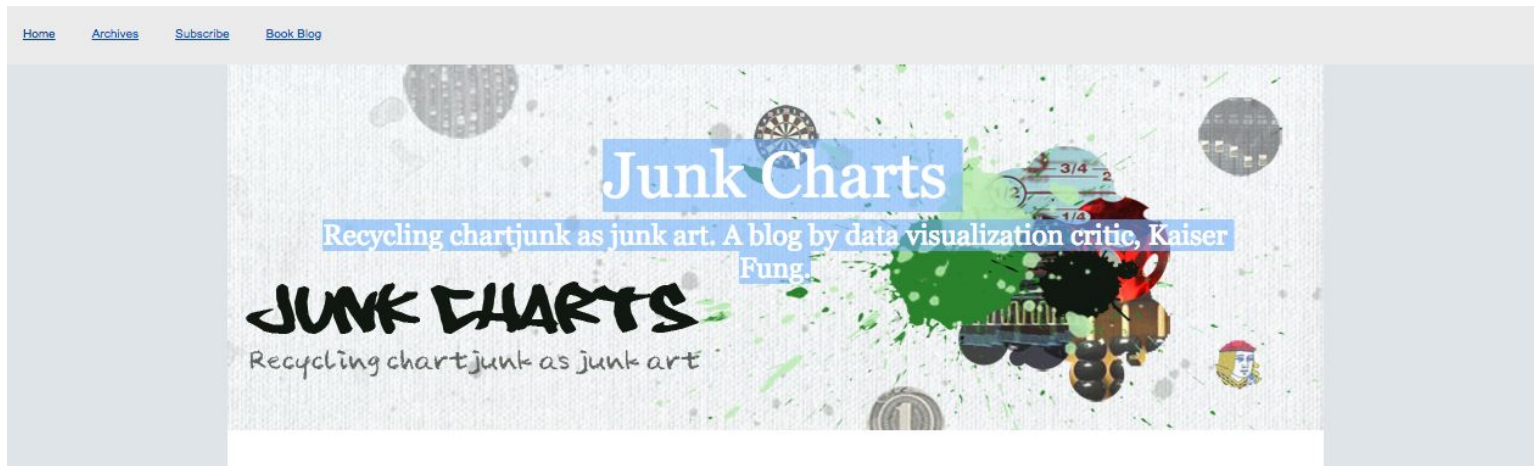
The list of 2018 visualization lists

- 33 lists, each has 10+ visualizations!
- [2018 in visuals: South China Morning Post's infographic highlights](#)
- [SCMP Print Arcade](#)
 - 217 visualizations from 2011 to 2019



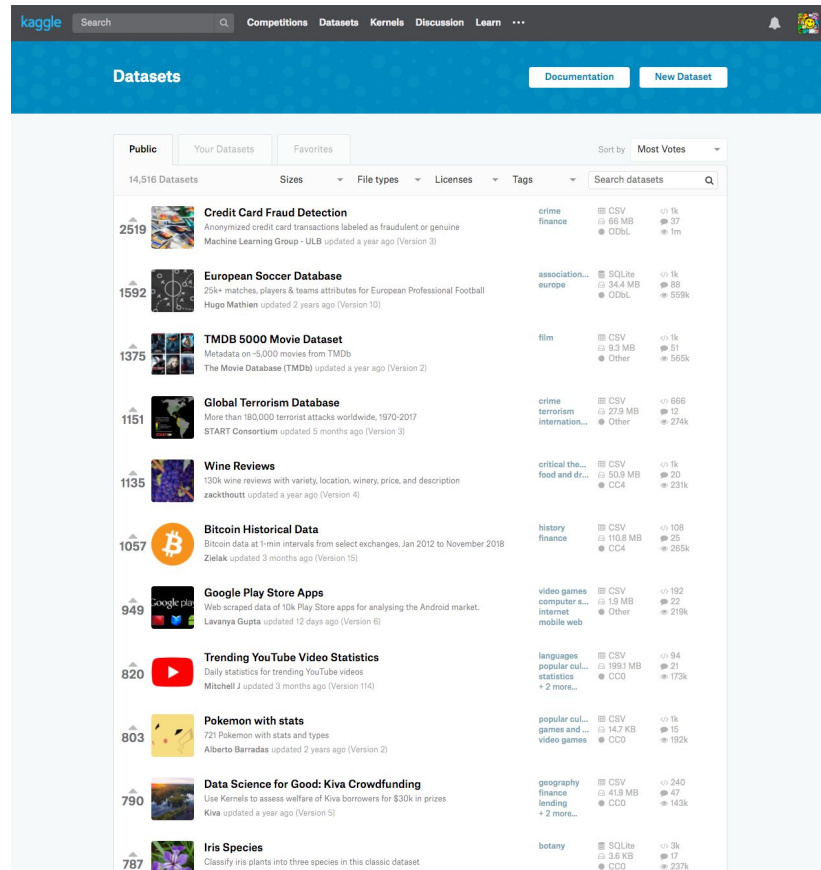
Junk Charts

- A collection of bad visualizations
 - How to lie with visualizations
 - Like [Data is Ugly](#) subreddit
 - With explanations
 - Update frequently



Kaggle Datasets

- No.1 source of datasets
- A lot of datasets
- Data are clean (relatively)
- A lot of kernels (jupyter notebooks)
 - See what the others do with the datasets
- Can seek help very easily
 - Can also raise questions to the authors



The screenshot shows the Kaggle Datasets page. At the top, there's a navigation bar with 'kaggle' logo, a search bar, and links for 'Competitions', 'Datasets', 'Kernels', 'Discussion', and 'Learn'. Below this, a blue header bar contains the word 'Datasets' and two buttons: 'Documentation' and 'New Dataset'. The main content area is a list of public datasets, sorted by 'Most Votes'. The list includes datasets like 'Credit Card Fraud Detection', 'European Soccer Database', 'TMDB 5000 Movie Dataset', 'Global Terrorism Database', 'Wine Reviews', 'Bitcoin Historical Data', 'Google Play Store Apps', 'Trending YouTube Video Statistics', 'Pokemon with stats', 'Data Science for Good: Kiwa Crowdfunding', and 'Iris Species'. Each dataset entry shows its rank, a thumbnail, title, description, file type, size, and number of votes.

Rank	Dataset Name	Description	File Type	Size	Votes
2519	Credit Card Fraud Detection	Anonymized credit card transactions labeled as fraudulent or genuine Machine Learning Group - ULB updated a year ago (Version 3)	CSV	69 MB	1k
1592	European Soccer Database	25k+ matches, players & teams attributes for European Professional Football Hugo Mathien updated 2 years ago (Version 10)	SQLite	34.4 MB	37
1375	TMDB 5000 Movie Dataset	Metadata on ~5,000 movies from TMDB The Movie Database (TMDB) updated a year ago (Version 2)	CSV	9.3 MB	1k
1151	Global Terrorism Database	More than 180,000 terrorist attacks worldwide, 1970-2017 START Consortium updated 5 months ago (Version 3)	CSV	27.8 MB	656
1135	Wine Reviews	130k wine reviews with variety, location, winery, price, and description zackthoutt updated a year ago (Version 4)	CSV	50.8 MB	20
1057	Bitcoin Historical Data	Bitcoin data at 1-min intervals from select exchanges, Jan 2012 to November 2018 Zielak updated 3 months ago (Version 15)	CSV	110.8 MB	108
949	Google Play Store Apps	Web scraped data of 10k Play Store apps for analysing the Android market. Lavanya Gupta updated 12 days ago (Version 6)	CSV	1.9 MB	192
820	Trending YouTube Video Statistics	Daily statistics for trending YouTube videos Mitchell J updated 3 months ago (Version 114)	CSV	1995 MB	94
803	Pokemon with stats	721 Pokemon with stats and types Alberto Barradas updated 2 years ago (Version 2)	CSV	14.7 KB	1k
790	Data Science for Good: Kiwa Crowdfunding	Use Kernels to assess welfare of Kiwa borrowers for \$30k in prizes Kiwa updated a year ago (Version 5)	CSV	41.9 KB	47
787	Iris Species	Classify Iris plants into three species in this classic dataset	SQLite	3.6 KB	3k

Dataviz Battle on r/dataisbeautiful

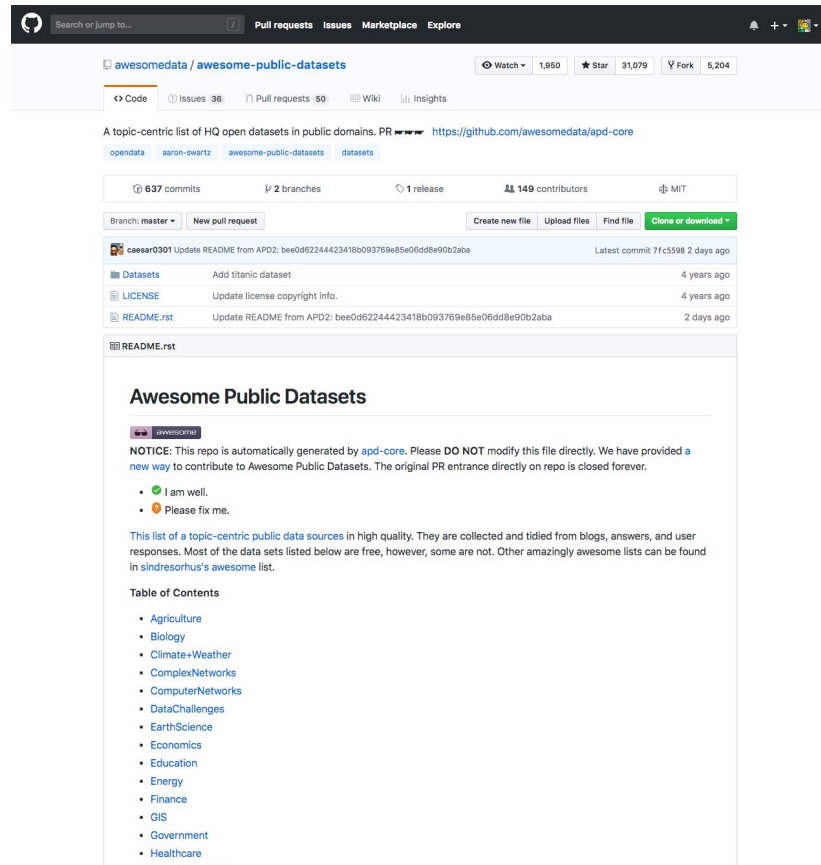
- Monthly competition on r/dataisbeautiful
- A lot of submissions for references
- September 2018: Visualize information on all 802 Pokemon
 - Winners are announced in the Dataviz Battle thread of next month
 - For example, October 2018 announced the winners of visualizing Pokemon

The screenshot shows the Reddit interface for the subreddit r/dataisbeautiful. The search bar at the top contains the text "dataviz battle for the month of". Below the search bar, the results are sorted by "NEW" and show a list of posts. The posts are titled "[Battle] DataViz Battle for the month of [Month] [Year]: [Topic]" and include details such as the number of comments, shares, and awards. The posts are arranged in a list, with the most recent at the top. On the right side of the page, there is a sidebar with community details for r/dataisbeautiful, including the number of subscribers (13.6m) and a description of the subreddit's purpose: "A place for visual representations of data: Graphs, charts, maps, etc. DataIsBeautiful is for visualizations that effectively convey information. **Aesthetics are an important part of information visualization, but pretty pictures are not the aim of this subreddit. **". There are also buttons for "SUBSCRIBE" and "CREATE POST".

Rank	Score	Title	Author	Comments	Shares	Awards
30	21	[Battle] DataViz Battle for the month of February 2019: Visualize Physical Harm and Dependence by Drug	u/AutoModerator	10 days ago	21	0
74	75	[Battle] DataViz Battle for the month of January 2019: Visualize the list of World's Oldest People	u/AutoModerator	1 month ago	75	0
83	112	[Battle] DataViz Battle for the month of December 2018: Visualize the Freezing and Thawing cycle of Lake Mendota	u/AutoModerator	2 months ago	112	0
75	70	[Battle] DataViz Battle for the month of November 2018: Visualize the List of NASA Astronauts	u/AutoModerator	3 months ago	70	0
55	55	[Battle] DataViz Battle for the month of October 2018: Visualize 859 survey results from r/travel	u/AutoModerator	4 months ago	55	0
112	102	[Battle] DataViz Battle for the month of September 2018: Visualize information on all 802 Pokemon	u/AutoModerator	5 months ago	102	0
82	94	[Battle] DataViz Battle for the month of August 2018: Visualize TSA Claims	u/AutoModerator	6 months ago	94	0
86	83	[Battle] DataViz Battle for the month of July 2018: Make it better: Which Birds prefer Which Seeds	u/AutoModerator	7 months ago	83	0
105	99	[Battle] DataViz Battle for the month of June 2018: Visualize The lives, reigns, and deaths of 68 Roman emperors from 26 BC to 395 AD	u/AutoModerator	8 months ago	99	0
84	28	[Battle] DataViz Battle for the month of May 2018: Visualize 1.6 Million Accidents in England, Scotland, and Wales from 2000-2016	u/AutoModerator	9 months ago	28	0
11	12	[Lounge] This week is a Bye Week for the DatViz Battles. Use this thread for off-topic discussion, smack talk, and cool suggestions!	u/AutoModerator	9 months ago	12	0
109	81	[Battle] DataViz Battle for the month of April 2018: Visualize every line from every scene in The Office	u/AutoModerator	10 months ago	81	0
		[Battle] DataViz Battle for the month of March 2018: Visualize Over 100,000 Stars				

awesome-public-datasets

- A very thorough list
- With active update



The screenshot shows the GitHub repository page for `awesomedata / awesome-public-datasets`. The repository has 1,950 watches, 31,079 stars, and 5,204 forks. It contains 637 commits, 2 branches, 1 release, and 149 contributors. The latest commit is from `caesar0301` updating the README from APD2: bee0d62244423418b093769e85e06dd8e90b2aba, 2 days ago.

The repository description is: "A topic-centric list of HQ open datasets in public domains. PR <https://github.com/awesomedata/apd-core>".

The commit history shows:

- `caesar0301` Update README from APD2: bee0d62244423418b093769e85e06dd8e90b2aba (Latest commit 7fc5598 2 days ago)
- `Datasets` Add titanic dataset (4 years ago)
- `LICENSE` Update license copyright info. (4 years ago)
- `README.rst` Update README from APD2: bee0d62244423418b093769e85e06dd8e90b2aba (2 days ago)

The README.rst file content is as follows:

Awesome Public Datasets

NOTICE: This repo is automatically generated by `apd-core`. Please **DO NOT** modify this file directly. We have provided a [new way](#) to contribute to Awesome Public Datasets. The original PR entrance directly on repo is closed forever.

- 🟢 I am well.
- 🟠 Please fix me.


This list of a [topic-centric public data sources](#) in high quality. They are collected and tidied from blogs, answers, and user responses. Most of the data sets listed below are free, however, some are not. Other amazingly awesome lists can be found in [sindresorhus's awesome list](#).

Table of Contents


- [Agriculture](#)
- [Biology](#)
- [Climate+Weather](#)
- [ComplexNetworks](#)
- [ComputerNetworks](#)
- [DataChallenges](#)
- [EarthScience](#)
- [Economics](#)
- [Education](#)
- [Energy](#)
- [Finance](#)
- [GIS](#)
- [Government](#)
- [Healthcare](#)

1001 Datasets and data repositories

- Another list of lists




Sign InSign Up


**Ryan Anderson**

38 COLLEAGUES109 FOLLOWERS108 FOLLOWING


Ryan Anderson's Library > Ryan Anderson's documents

ABOUT R.
R'S KINDREDS
COLLABORATE W/ R.
COLLEAGUES.
R'S LIBRARY >
HOME

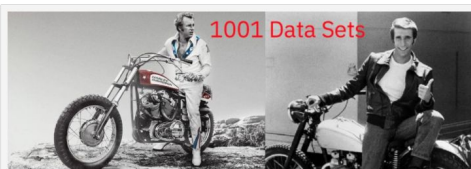


**Document**

2 FOLLOWERS10 RECOMMENDS

1001 Datasets and Data repositories (List of lists of lists)
by  Ryan Anderson

1001 Datasets and Data repositories (List of lists of lists)



1001 Data Sets

This is a LIST of.... "lists of lists". Messy presentation to pull together **Raw Datasets** for my hacks. Suggestions to add? Message me or post comment..

CTRL-F to "FIND" is your best bet - e.g. CTRL-F "food" or "population"

100% of the links below are from external sources (not mine)

Follow me on Twitter <https://twitter.com/ryan77anderson> Need code or pattern once you find the data? Try here: https://dreamtolearn.com/ryan77_journey_to_watson73


QuickDraw

<https://quickdraw.withgoogle.com/data> (source: <https://quickdraw.withgoogle.com/#>)


Source: Medium: The 50 Best Public Datasets for Machine Learning

*What are some open datasets for machine learning? After scrapping the web for hours after hours, we have created a great cheat sheet for high quality and diverse machine learning datasets.

FOLLOWED BY



RECOMMENDED BY



ABOUT THIS DOCUMENT

1001 Datasets and Data repositories (List of lists of lists) - rough list to compile - a rough lists of lists

Created: December 26, 2013

YOU MIGHT ALSO LIKE


Cognitive Wingman
Your Cognitive Companion We are currently passing through the era of Siri, Cortana and Alexa- and evolving to the next generation of CognitiveAI powered assistants. These more advanced assistants will serve their human companions to help them navigate all aspects of personal and professional life. The Cognitive Wingman is a more sophisticated, emotionally intelligent

Tasks

- Get the whole list of [“Where to find visualizations and datasets” on GitHub](#)
- If you don't have a group yet:
 - a. Talk to your classmates about the visualizations you like
 - b. Form a group if interest matches
 - i. Signup on [Google Docs](#)
- Go to your group mates
 - a. Find some visualizations
 - b. Talk to your group mates
 - c. Find a dataset to work on
 - d. Talk about what interesting insight can be found in the dataset
- Make amazing visualizations!

Next tutorial

Python, Jupyter and
Pandas

- Register [Google Colab](#) beforehand
 - Jupyter notebook environment
 - Free!
 - No setup
- Alternatively, you can use jupyter notebook on your computer, but that is cumbersome