



HKUST
VISLAB

COMP 4462

Data Visualization Tutorial

Leo Yu Ho, Lo
Ming Yao

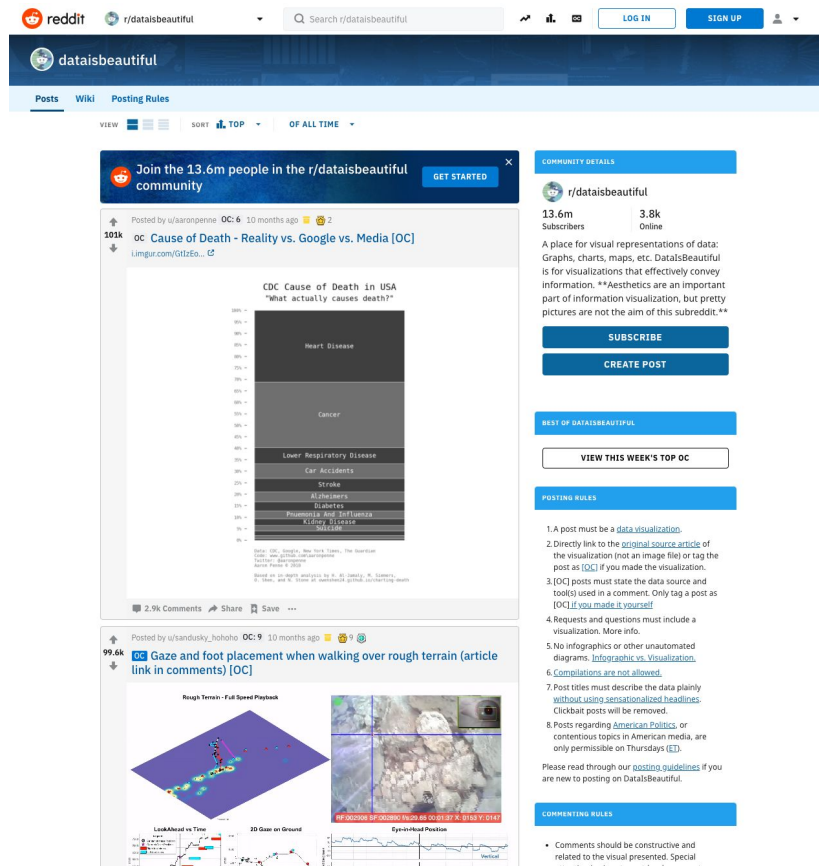
Tuesday 26 February, 2019
<https://bit.ly/vis-t03>

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?

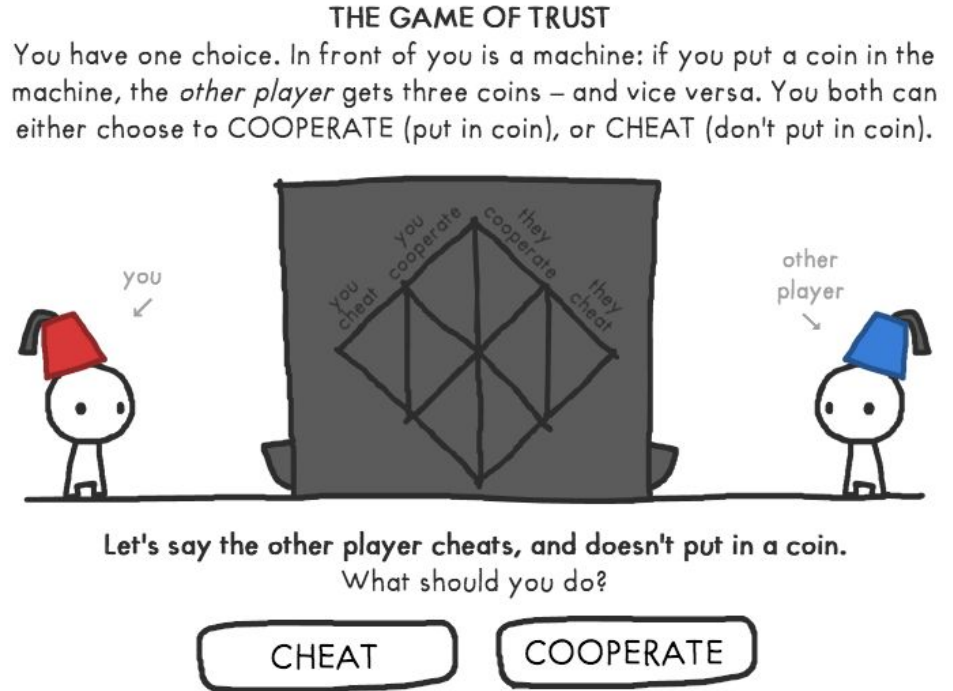
Data is Beautiful

- 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



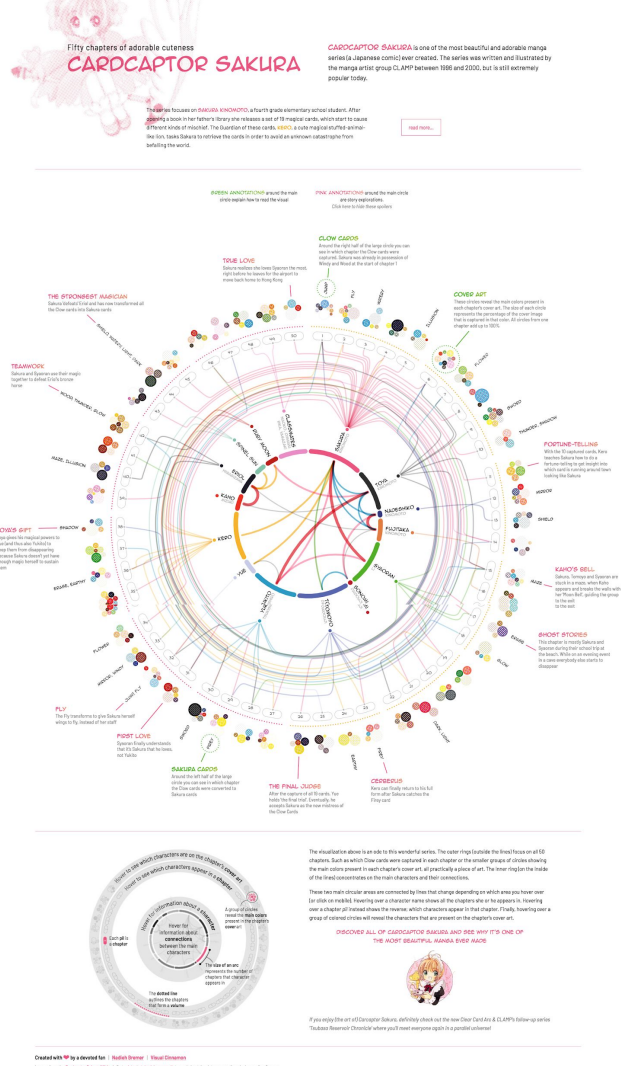
Nick Case

- 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](#)



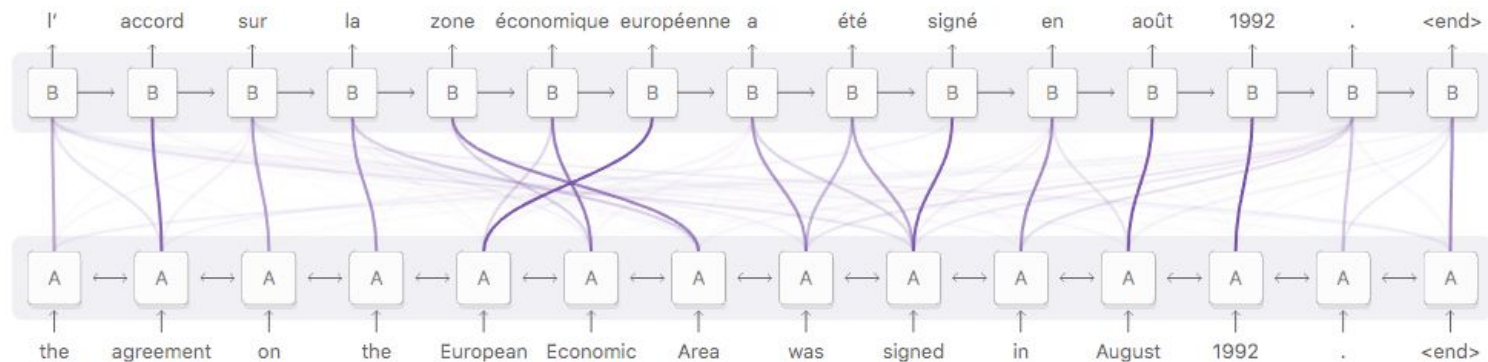
Data Sketches

- Beautiful! Eye pleasing! Fun datasets!
- And they have 24 of them!
- By:
 - [Nadieh Bremer](#)
 - [Susie Lu](#)
- [Cardcaptor Sakura](#)
 - Visualizing 50 chapters of the manga
 - Appeared characters
 - Magic spells
 - Annotations
- Another one on [Dragon Ball Z](#)
- With [explanations](#)!
 - They have journaled the process in details!



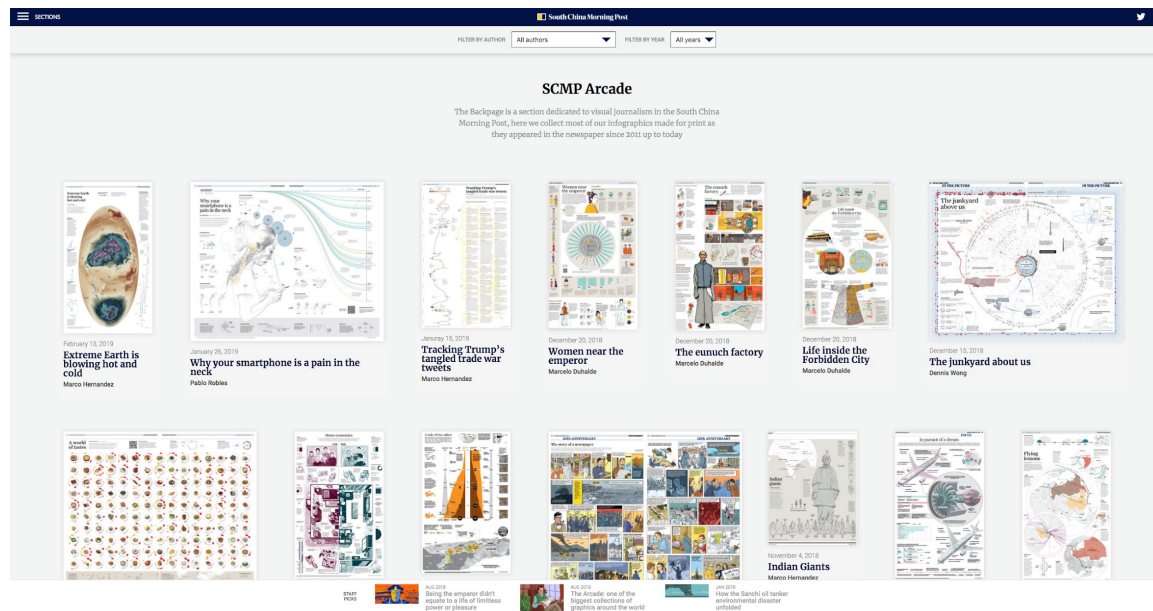
Distill

- 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?



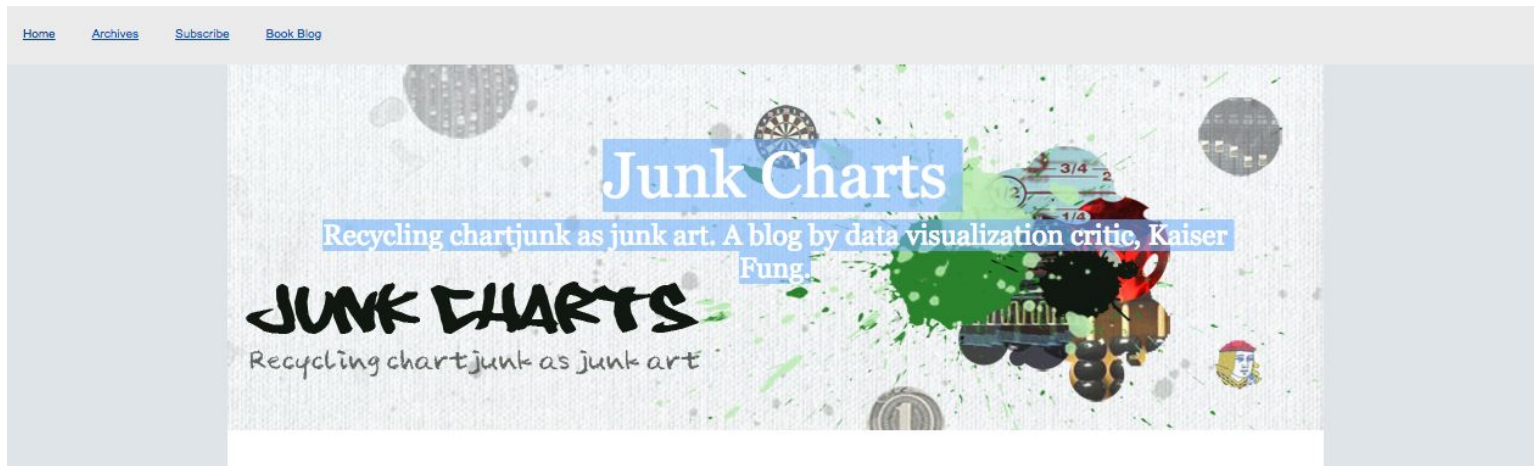
The list of 2018 visualization lists

- 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



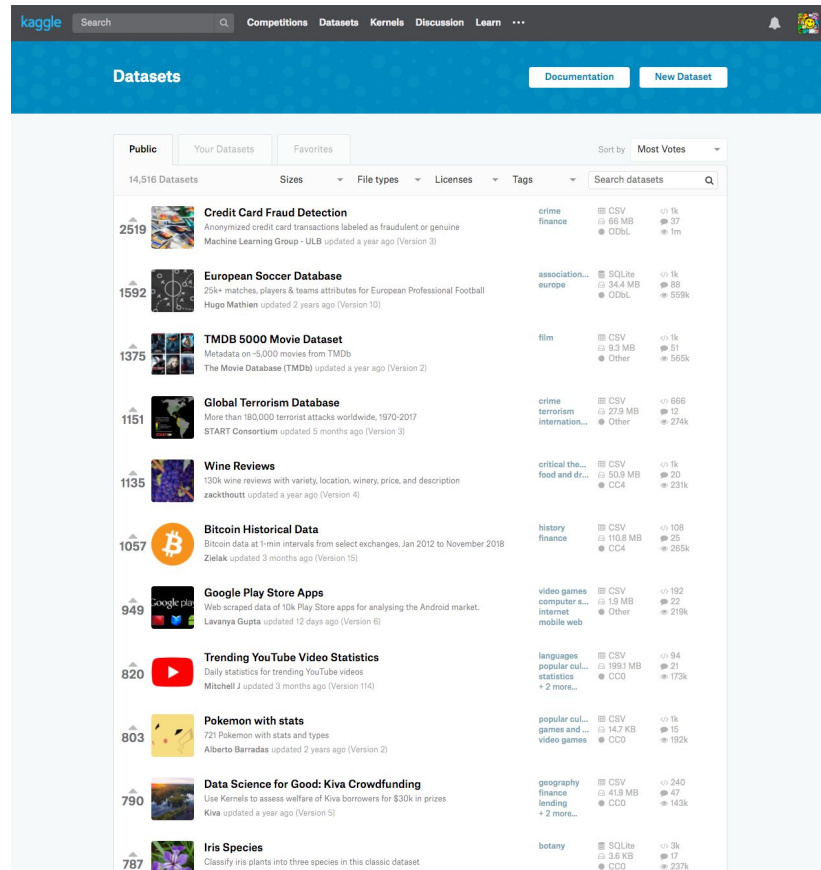
Junk Charts

- 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 contains the word 'Datasets' and buttons for 'Documentation' and 'New Dataset'. The main content area displays 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 MB	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 online users (3.4k). There are also buttons for "SUBSCRIBE" and "CREATE POST". At the bottom of the sidebar, there is a link to the "COMMUNITY OPTIONS" menu.

reddit Subreddit Results dataviz battle for the month of

Search results in r/dataisbeautiful

SORT BY: NEW ALL REDDIT RESULTS

[Battle] DataViz Battle for the month of February 2019: Visualize Physical Harm and Dependence by Drug
r/dataisbeautiful · Posted by u/AutoModerator 10 days ago
21 Comments · Share · Save · Give Award

[Battle] DataViz Battle for the month of January 2019: Visualize the list of World's Oldest People
r/dataisbeautiful · Posted by u/AutoModerator 1 month ago
75 Comments · Share · Save · Give Award

[Battle] DataViz Battle for the month of December 2018: Visualize the Freezing and Thawing cycle of Lake Mendota
r/dataisbeautiful · Posted by u/AutoModerator 2 months ago
112 Comments · Share · Save · Give Award

[Battle] DataViz Battle for the month of November 2018: Visualize the List of NASA Astronauts
r/dataisbeautiful · Posted by u/AutoModerator 3 months ago
70 Comments · Share · Save · Give Award

[Battle] DataViz Battle for the month of October 2018: Visualize 859 survey results from r/travel
r/dataisbeautiful · Posted by u/AutoModerator 4 months ago
55 Comments · Share · Save · Give Award

[Battle] DataViz Battle for the month of September 2018: Visualize information on all 802 Pokemon
r/dataisbeautiful · Posted by u/AutoModerator 5 months ago
102 Comments · Share · Save · Give Award

[Battle] DataViz Battle for the month of August 2018: Visualize TSA Claims
r/dataisbeautiful · Posted by u/AutoModerator 6 months ago
94 Comments · Share · Save · Give Award

[Battle] DataViz Battle for the month of July 2018: Make it better: Which Birds prefer Which Seeds
r/dataisbeautiful · Posted by u/AutoModerator 7 months ago
83 Comments · Share · Save · Give Award

[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
r/dataisbeautiful · Posted by u/AutoModerator 8 months ago
99 Comments · Share · Save · Give Award

[Battle] DataViz Battle for the month of May 2018: Visualize 1.6 Million Accidents in England, Scotland, and Wales from 2000-2016
r/dataisbeautiful · Posted by u/AutoModerator 9 months ago
28 Comments · Share · Save · Give Award

[Lounge] This week is a Bye Week for the DatViz Battles. Use this thread for off-topic discussion, smack talk, and cool suggestions!
r/dataisbeautiful · Posted by u/AutoModerator 9 months ago
12 Comments · Share · Save · Give Award

[Battle] DataViz Battle for the month of April 2018: Visualize every line from every scene in The Office
r/dataisbeautiful · Posted by u/AutoModerator 10 months ago
81 Comments · Share · Save · Give Award

[Battle] DataViz Battle for the month of March 2018: Visualize Over 100,000 Stars

COMMUNITY DETAILS

r/dataisbeautiful

13.6m Subscribers 3.4k Online

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.**

SUBSCRIBE

CREATE POST

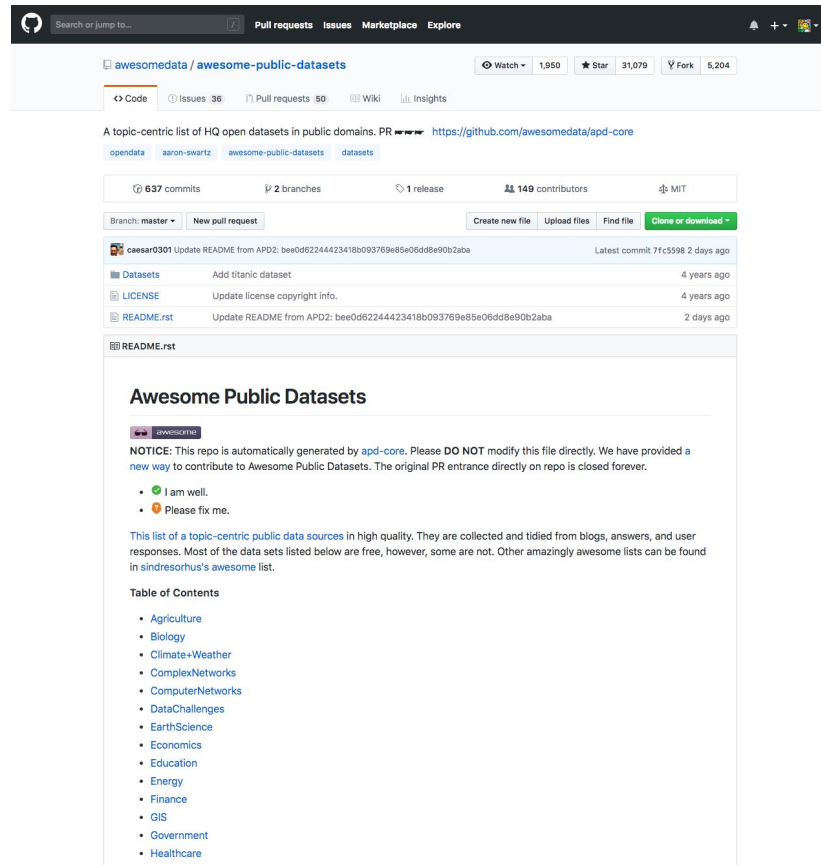
COMMUNITY OPTIONS

About Careers Press Advertise Blog Help The Reddit App Reddit Coins Reddit Premium Reddit Gifts

Content Policy | Privacy Policy User Agreement | Mod Policy © 2019 Reddit, Inc. All rights reserved

awesome-public-datasets

- A very thorough list
- With active update
- Search Engine subsection
 - Websites that have “search for datasets”
- Data Challenge subsection
 - More Kaggle like websites
- Complementary Collection subsection
 - More dataset lists



1001 Datasets and data repositories

- Another list of lists
- A long long list
- Compiled from many sources
 - GitHub
 - Government
 - Blogs
 - Big Companies
 - Quora
 - Reddit
- In the blog.visual.ly subsection
 - Specifically suitable for visualization

The screenshot shows a user profile for Ryan Anderson on the Dreamtolearn platform. The profile has 38 colleagues, 109 followers, and 108 following. The main content area displays a document titled "1001 Datasets and Data repositories (List of lists of lists)" by Ryan Anderson. The document features a header with a pencil icon and the text "dreamtolearn Document". Below the header, there is a section titled "1001 Datasets and Data repositories (List of lists of lists)" with a thumbnail image showing two people on motorcycles. The text of the document describes a list of lists of lists of datasets, mentioning "Raw Datasets" and "CTRL-F to FIND". It also includes a section for "Follow me on Twitter" with a link to Ryan Anderson's Twitter profile. The document is categorized under "KINDRED" and has 2 followers and 10 recommendations. A sidebar on the right contains sections for "FOLLOWED BY", "RECOMMENDED BY", "ABOUT THIS DOCUMENT", and "YOU MIGHT ALSO LIKE".

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

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