



**HKUST**  
VISLAB

HKUST  
HCI Initiative

# **COMP 4462**

# **Data Visualization Tutorial**

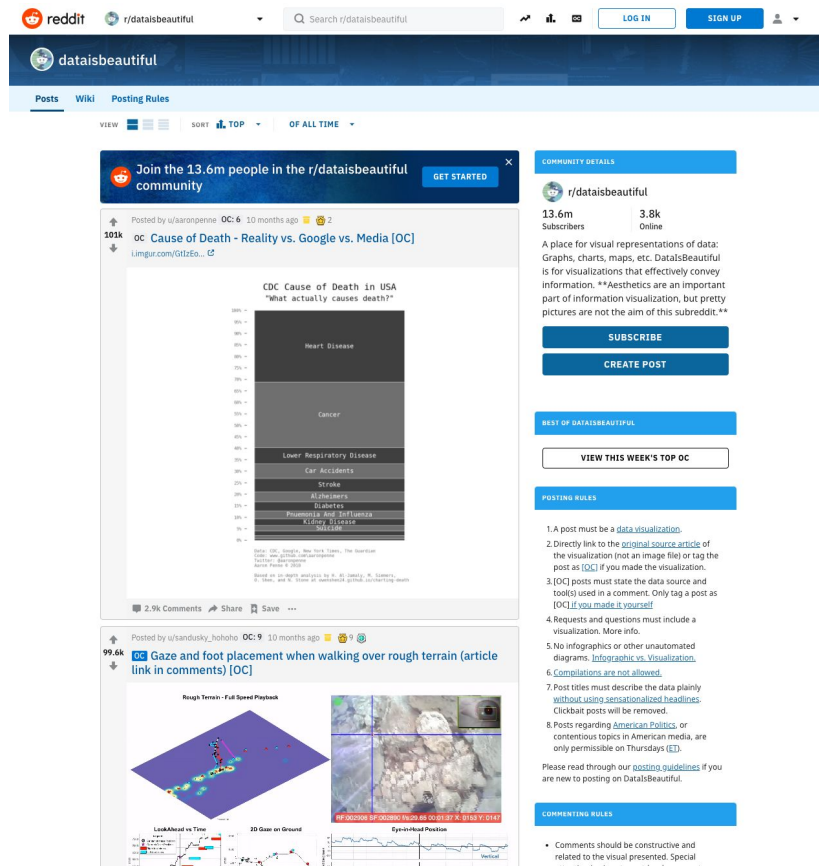
PAN Ziqi  
CHEN Chang

Friday 18 October, 2024

**Where to find:  
visualizations and datasets?**

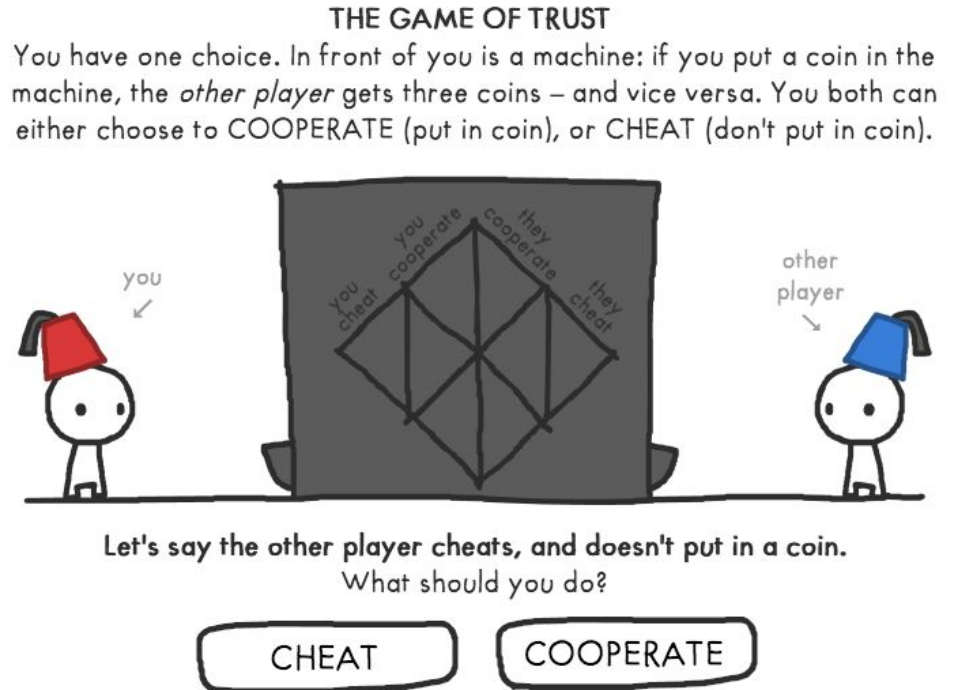
# Data is Beautiful

- 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



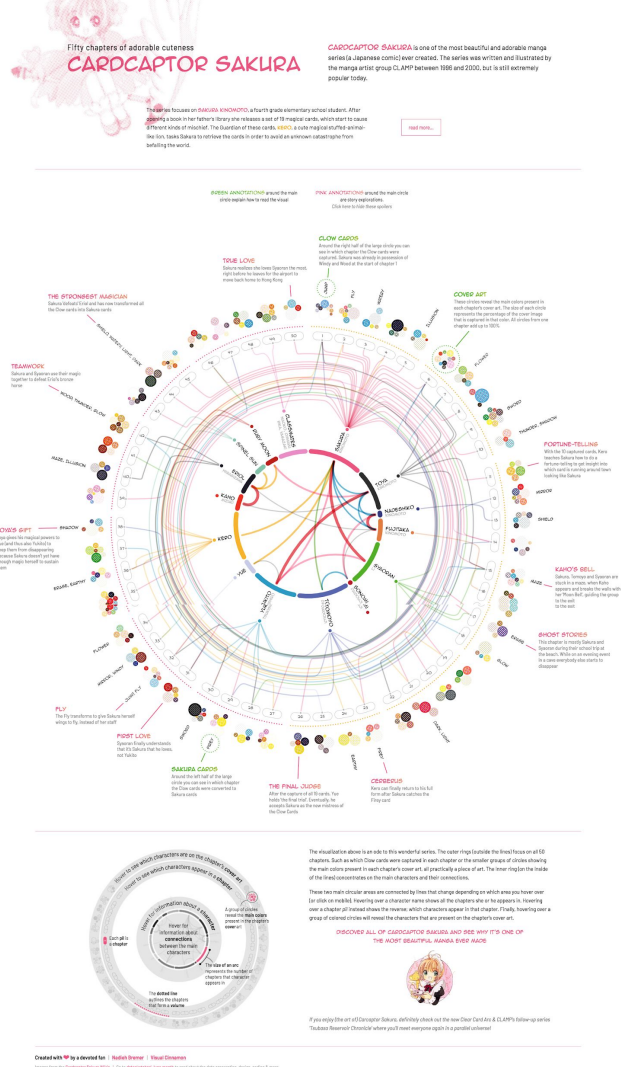
# Nick Case

- 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](#)



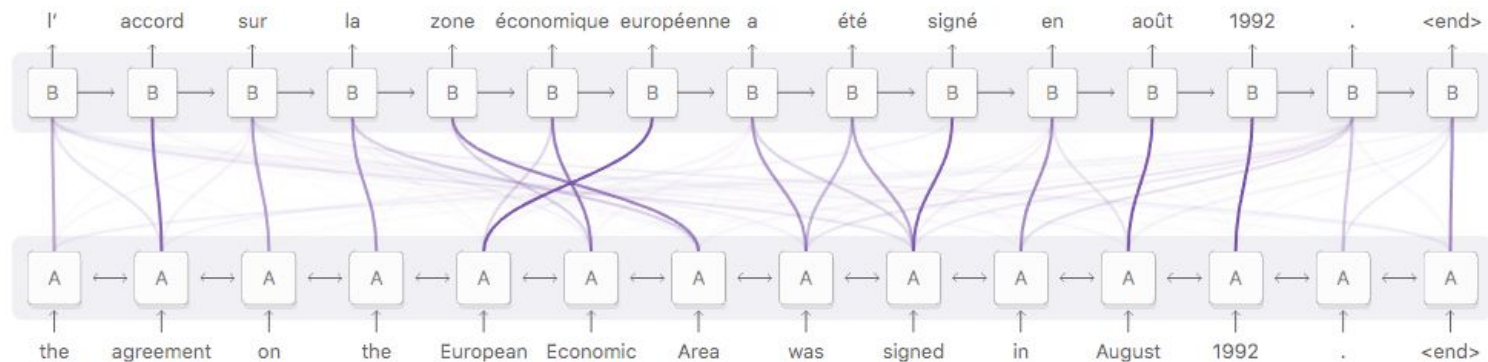
# Data Sketches

- 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?



# Tableau Public Gallery

- A lot of visualizations built with Tableau
  - With tableau worksheets and data
- Other galleries: [plotly](#), [Observable](#), [DataWrapper](#)

## Gallery / Featured

Stunning data visualization examples from across the web created with Tableau Public.

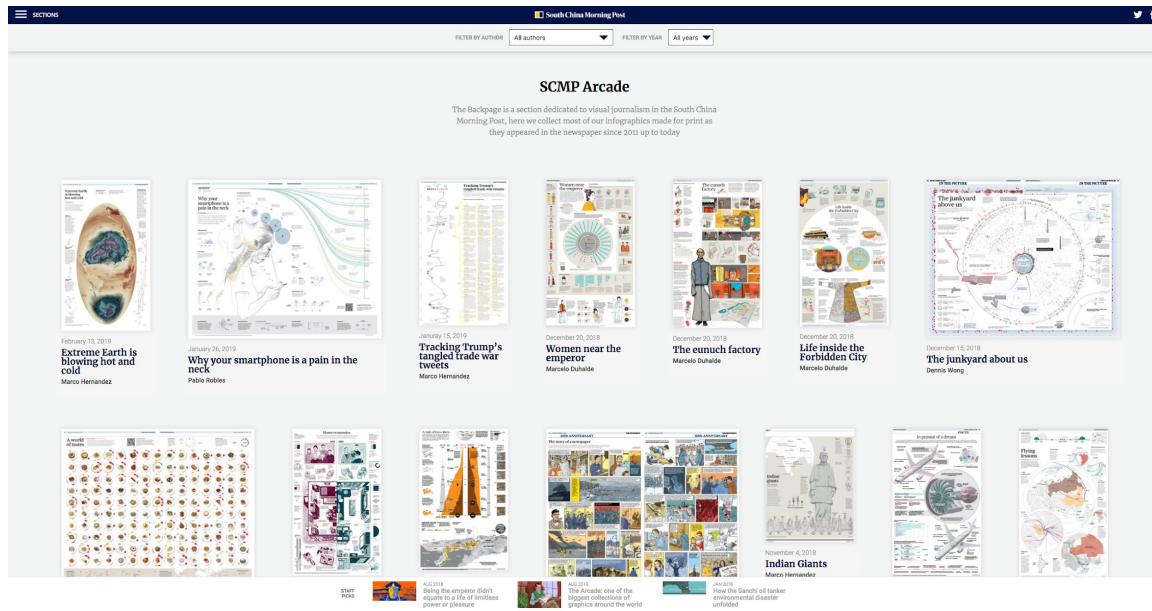
Viz of the Day **Featured**

Social Impact ▼

Nominate a Viz

# The list of 2022 visualization lists

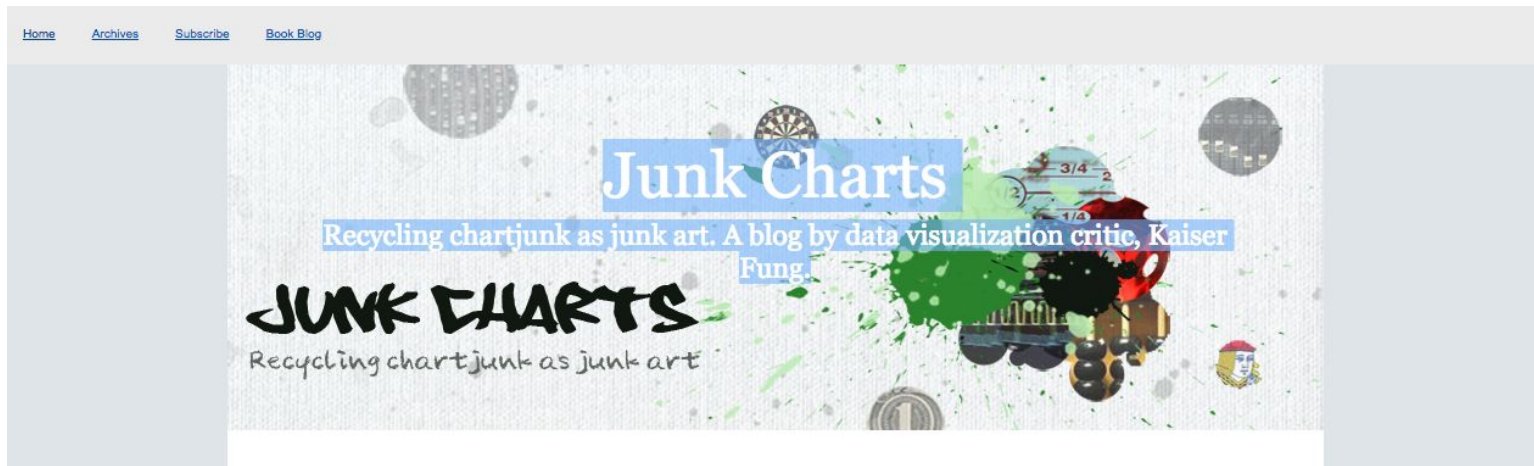
- 38 lists, each has 10+ visualizations! (sadly the author quit being a freelancer and probably the lists will not update...)
- [The list of 2021 visualization lists](#)
- [2020](#), [2019](#), [2018](#), [2017](#), [2016](#), [2015](#)
- [SCMP Print Arcade](#)





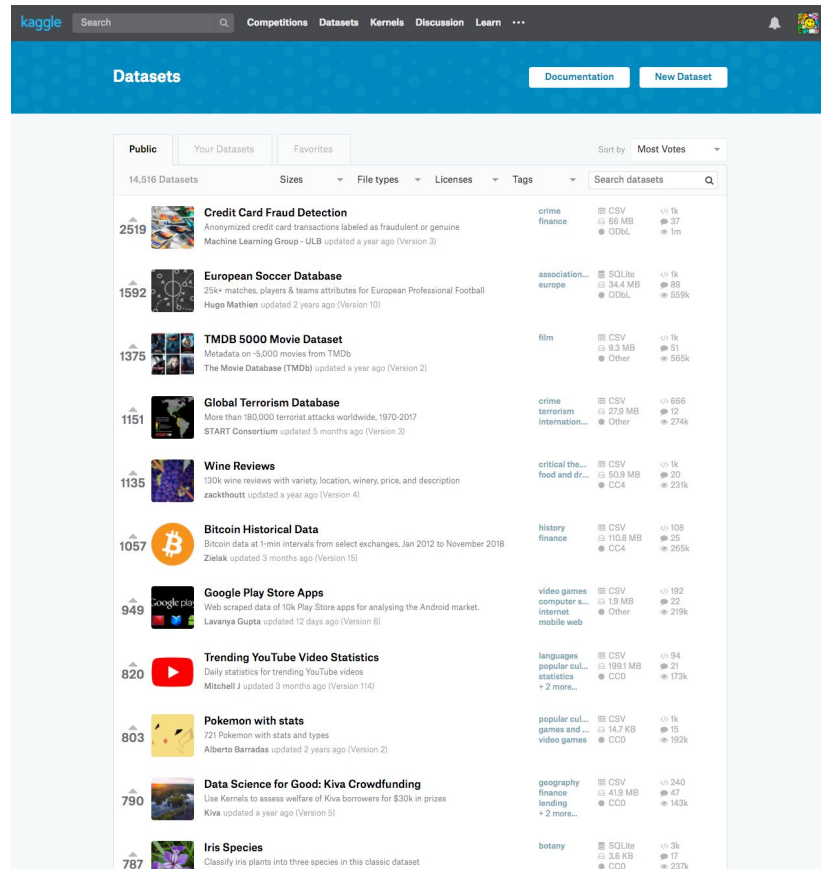
# Junk Charts

- 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 displays a list of public datasets. The list is sorted by 'Most Votes' and shows 14,516 datasets. Each dataset entry includes a rank, a thumbnail icon, the dataset name, a brief description, the file type, size, and the number of votes. The datasets listed are:

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	68 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	199.5 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



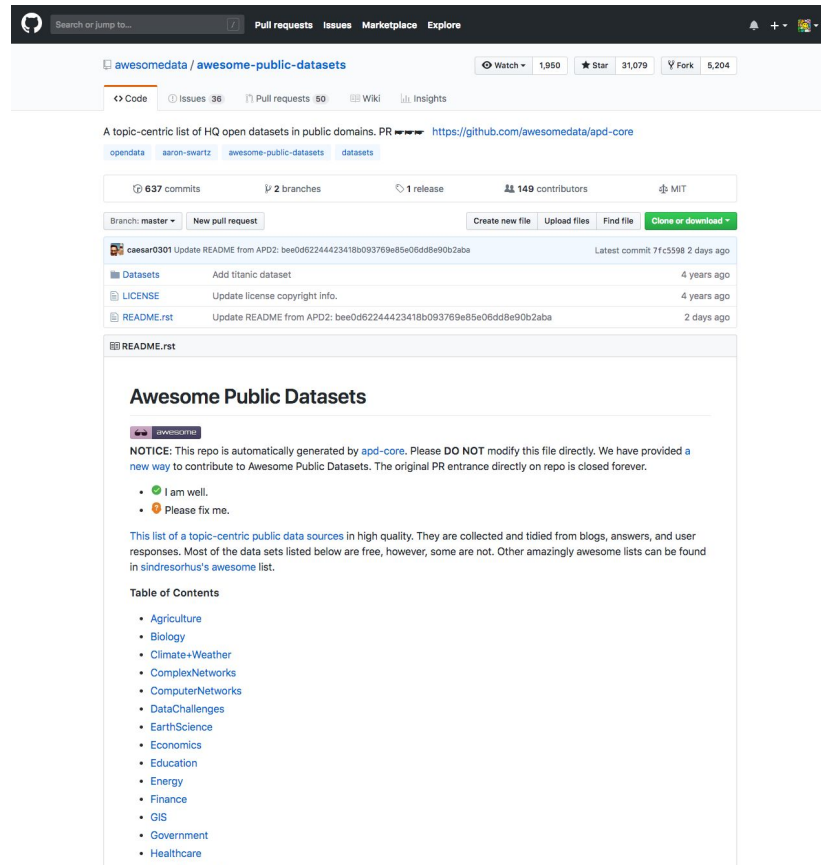
The screenshot shows the Reddit interface for the subreddit r/dataisbeautiful. The search bar at the top contains the text "dataviz battle for the month of". Below the search bar, the title "dataviz battle for the month of" is displayed. The main content area shows a list of threads, each titled "[Battle] Dataviz Battle for the month of [Month Year]: [Topic]". The threads are sorted by "NEW". The threads listed are:

- [Battle] Dataviz Battle for the month of February 2019: Visualize Physical Harm and Dependence by Drug
- [Battle] Dataviz Battle for the month of January 2019: Visualize the list of World's Oldest People
- [Battle] Dataviz Battle for the month of December 2018: Visualize the Freezing and Thawing cycle of Lake Mendota
- [Battle] Dataviz Battle for the month of November 2018: Visualize the List of NASA Astronauts
- [Battle] Dataviz Battle for the month of October 2018: Visualize 859 survey results from r/travel
- [Battle] Dataviz Battle for the month of September 2018: Visualize information on all 802 Pokemon
- [Battle] Dataviz Battle for the month of August 2018: Visualize TSA Claims
- [Battle] Dataviz Battle for the month of July 2018: Make it better: Which Birds prefer Which Seeds
- [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
- [Battle] Dataviz Battle for the month of May 2018: Visualize 1.6 Million Accidents in England, Scotland, and Wales from 2000-2016
- [Lounge] This week is a Bye Week for the DatViz Battles. Use this thread for off-topic discussion, smack talk, and cool suggestions!
- [Battle] Dataviz Battle for the month of April 2018: Visualize every line from every scene in The Office
- [Battle] Dataviz Battle for the month of March 2018: Visualize Over 100,000 Stars

On the right side of the page, there is a "COMMUNITY DETAILS" section for r/dataisbeautiful, showing 13.6m subscribers and 3.4k online users. Below this is a "COMMUNITY OPTIONS" section with links for About, Careers, Press, Advertise, Blog, Help, The Reddit App, Reddit Coins, Reddit Premium, and Reddit Gifts. At the bottom right, there is a "Content Policy" section with links for Privacy Policy, User Agreement, and Mod Policy, and a copyright notice for 2019 Reddit, Inc.

# 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



# Tasks

- Get the whole list of [“Where to find visualizations and datasets” on GitHub](#)
- Project Topics
  - Talk to your group mates for project topics
  - Find a dataset to work on
  - Talk about what interesting insight can be found in the dataset
  - Make amazing visualizations!
- Enjoy the beauty of visualizations! Have fun making your own visualizations!

# **In-class exercise 1**

# Question 1 and answers

Data can be collected about any situation, object, or activity. Take a look at the photograph below and identify at least one variable of each scale type in the scene:



a. Nominal: (names or categories with apparent order) e.g., countries of athletes, gender, etc.

b. Ordinal: (have a clear order but the adjacent items may not be evenly spaced) e.g., order of the ships

c. Interval: (ordered, evenly spaced, no true zero) e.g., temperature in C or F

d. Ratio: (ordered, evenly spaced, with true zero -- 0 means absence) e.g., athlete number, height, etc.

# Question 1 grading scheme and takeaways

Grading scheme: each type of data worth 0.5 points, 2 points in total.

Takeaways:

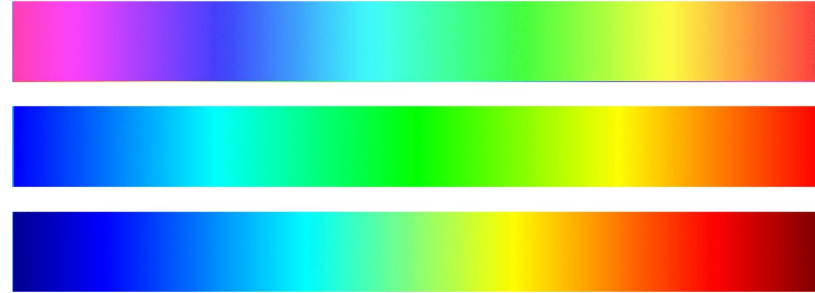
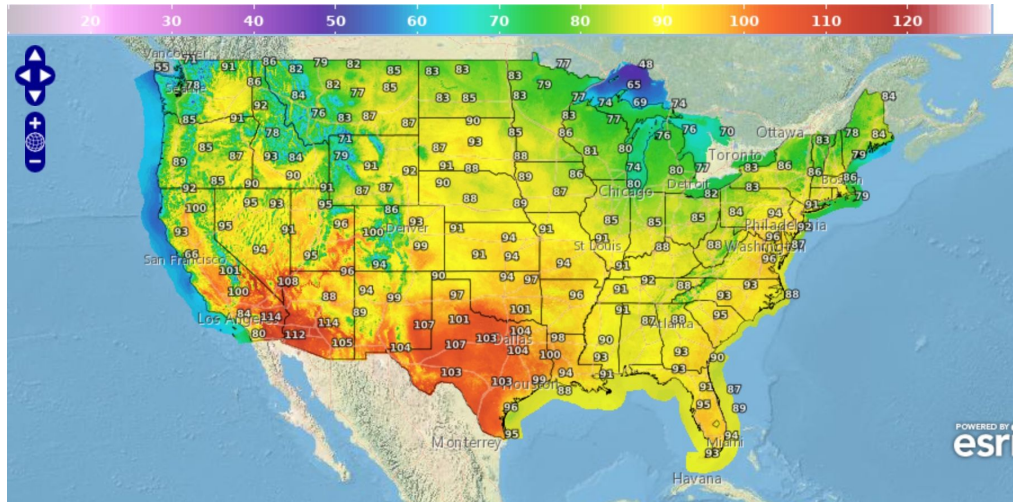
1. For interval or ratio data, it is tricky to be clear about the true zero. You may need to clarify what is the “interval”, and why there is/isn’t true zero.
2. For any question, respect the question itself. Always bring up your ideas according to the materials that have been provided



# Question 2

Please:

- 1) name the color scheme used in the following chart (Figure 2),
- 2) explain why it and its variations (Figure 3) inevitably perform poorly on any test of the perceived metric distance between displayed quantities, and
- 3) propose a better color scheme design (color palette + textual justification).



## Question 2 answers and grading scheme

1) (0.5 points) Rainbow scheme (cannot be sequential/divergent, but can be describing the rainbow scheme without mentioning the word rainbow)

2) (1 point)

Rainbow colormap has two peaks in brightness (near 20 and 90). (point out the problem with the color scheme, 0.5 points)

Data as shown in the legend is sequential and thus rainbow scheme is not suitable for presenting one-directional sequential data. (explain with the characteristic of the data why the color scheme is not suitable, 0.5 points)

3) (0.5 points) Any sequential color scheme mentioned in class (with a clear one-directional change in color saturation, brightness, etc.)

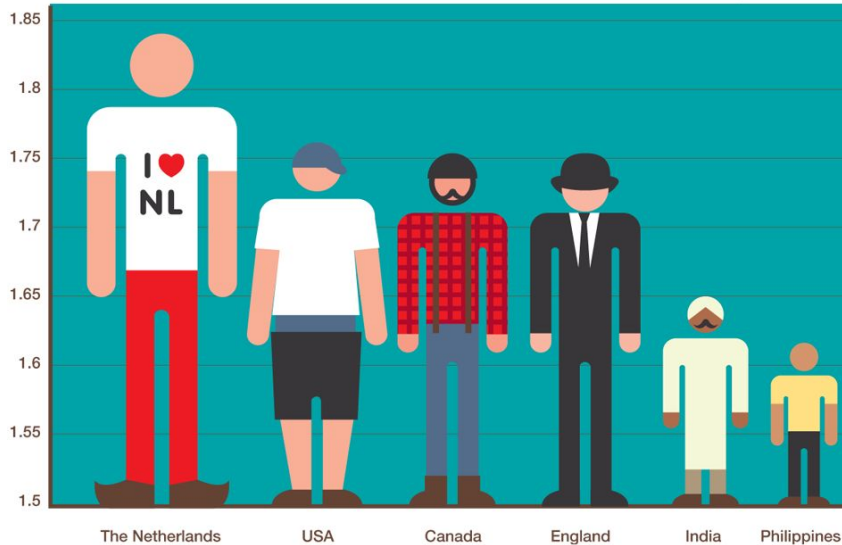
Takeaways: when asked to explain why A (e.g. rainbow scheme) is not good for B (e.g. perceived metric distance), you have to clarify: 1. What is special about A; 2. What is special about B; 3. What's wrong with the relation of A and B

# Question 3

Please identify all the problems with the following data visualization (Figure 4). Explain what Principle(s) from Tufte it violates and how to correct the problem(s).

## LOOKING DOWN ON THE REST OF THE WORLD

(Average male height in m)

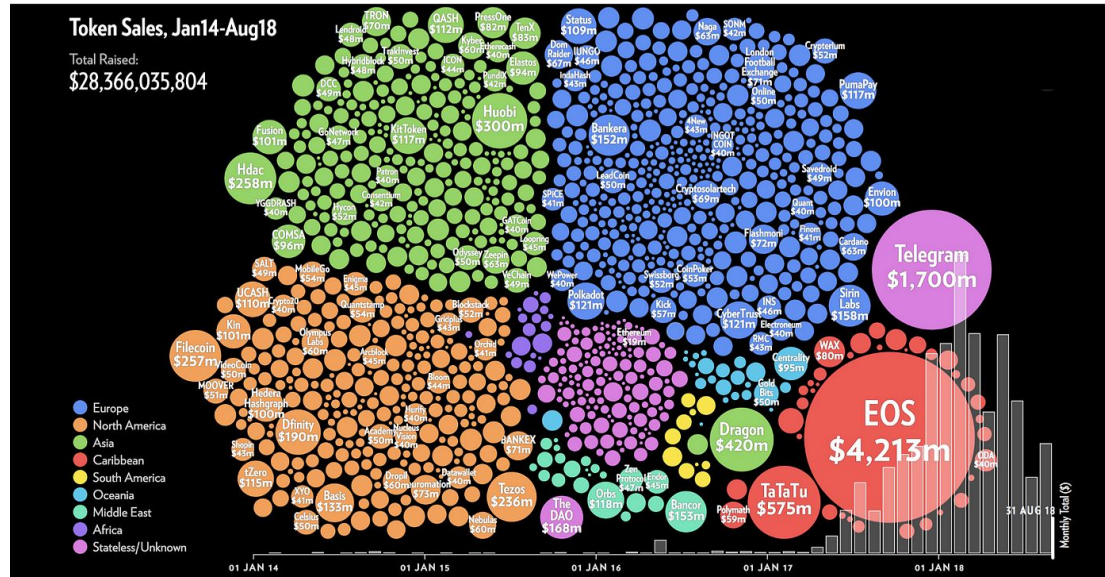


- 2D area encoding for 1D data --> keep the width the same and only vary the height
  - High lie factor due to truncated Y axis --> start Y from 0
  - Low data ink ratio with chart junks --> remove chart junks
- ( $\frac{1}{3}$ : 1 point,  $\frac{2}{3}$ : 2 point, all: 2.5 point)

Takeaways: do not repeat on one violation of principles; Observe the data:  $x, y \rightarrow x * y \rightarrow$  others

Please describe:

2) what tasks this visualization can support.



# Question 4

1) (2 points, 1 point for each mark) Point mark with color hue encoding continent, size encoding amount, and x position encoding time; line mark with x position encoding time and y position (or height/length) encoding monthly total amount.

2) (1.5 points, 1 for the idea, 0.5 for the correct use of language) Identify distribution (bar chart monthly total), identify trend (bubble chart), identify extremes (bar), identify outliers (bubble), compare similarity (bubble groups).

