HKUST
VISLAB

HKUST
HCI Initiative

COMP 4462
Data Visualization Tutorial

PAN Ziqi
CHEN Chang

Friday 18 October, 2024

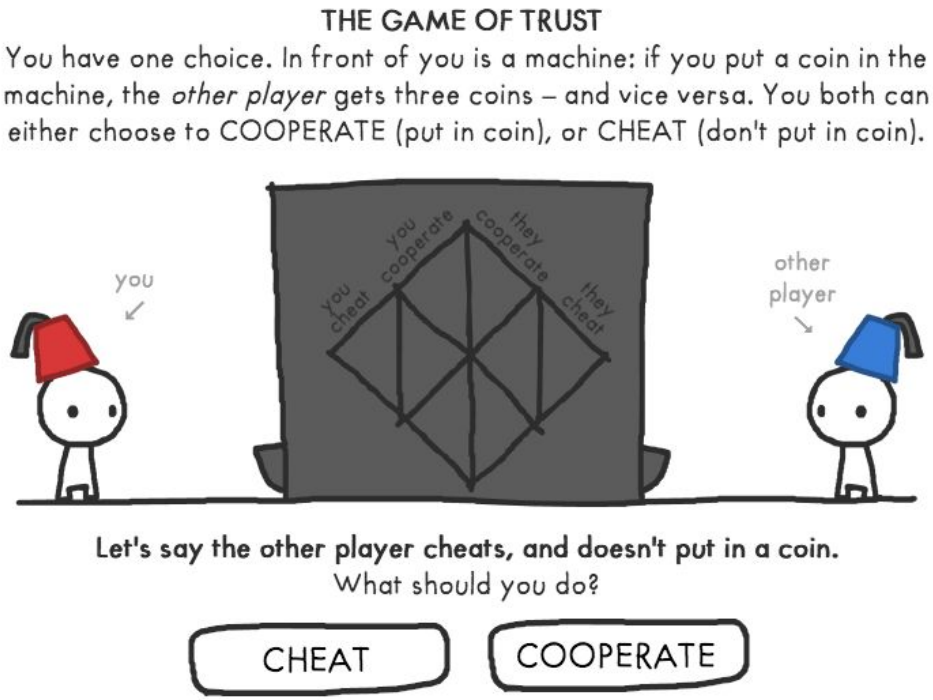
Where to find: visualizations and datasets?

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

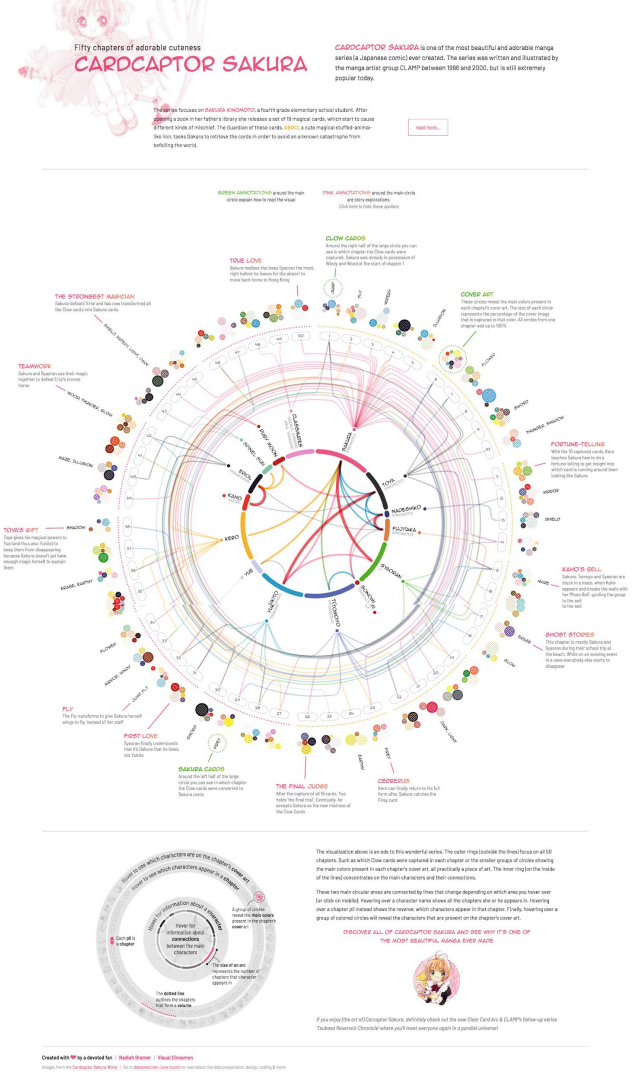
Nick Case

- 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 <=> which word in English?

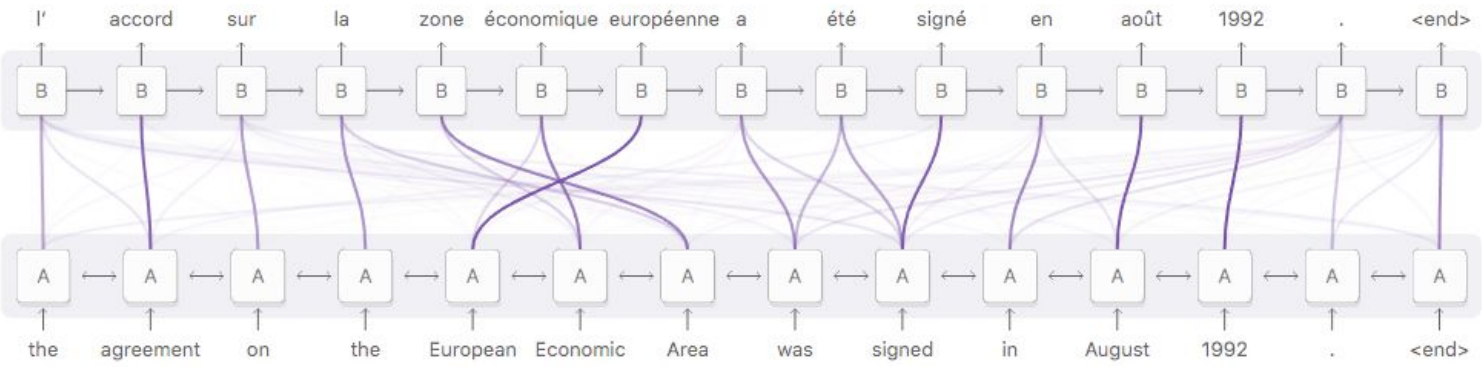
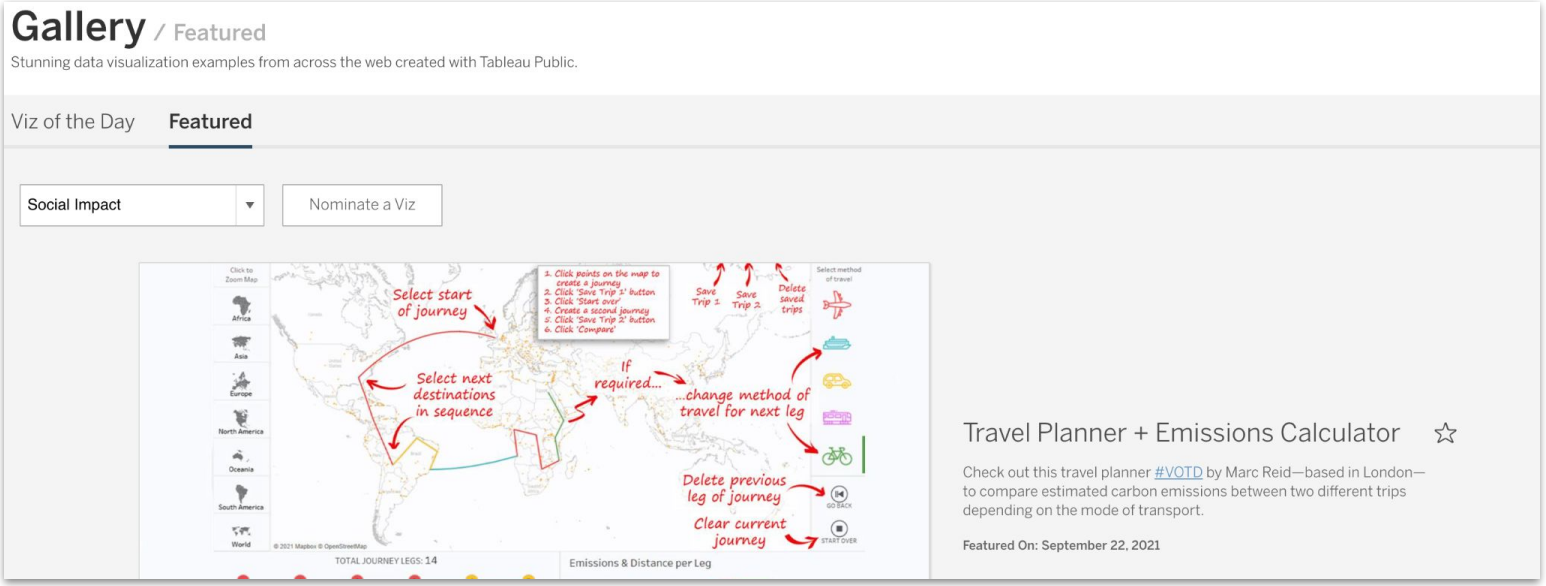


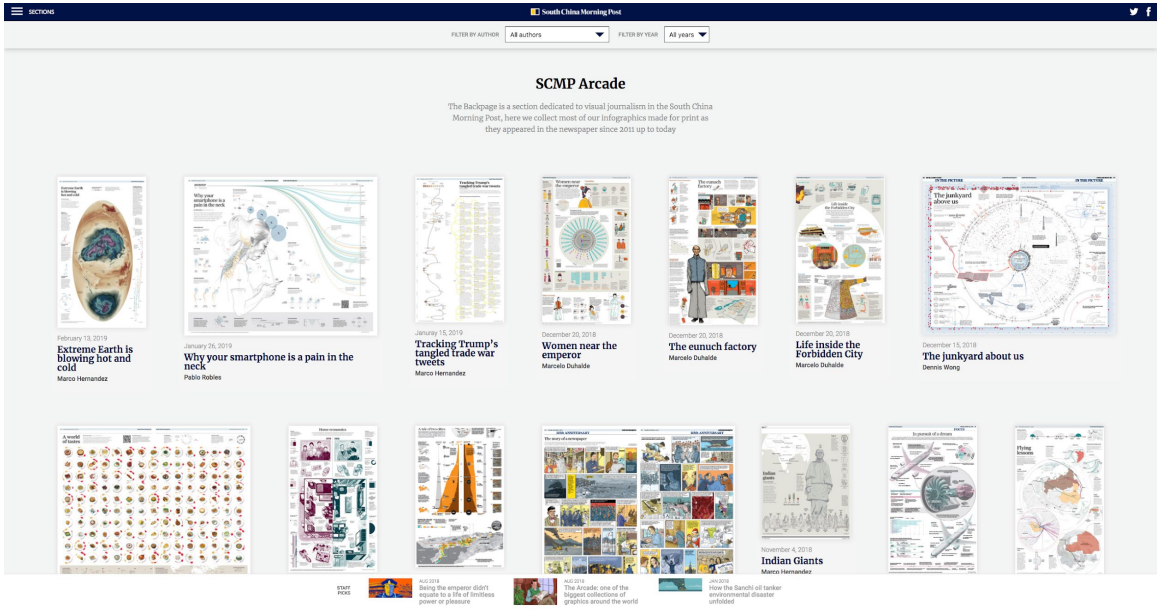
Tableau Public Gallery

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



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

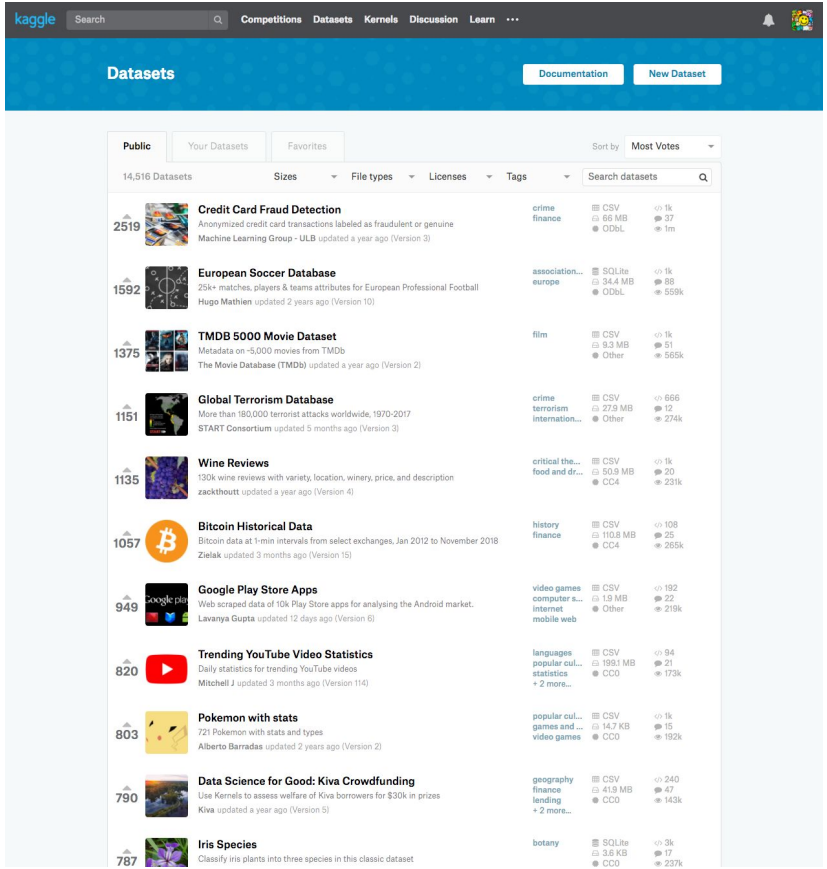


Junk Charts

- A collection of bad visualizations
  - How to lie with visualizations
  - Like Data is Ugly subreddit
  - With explanations
  - Update frequently

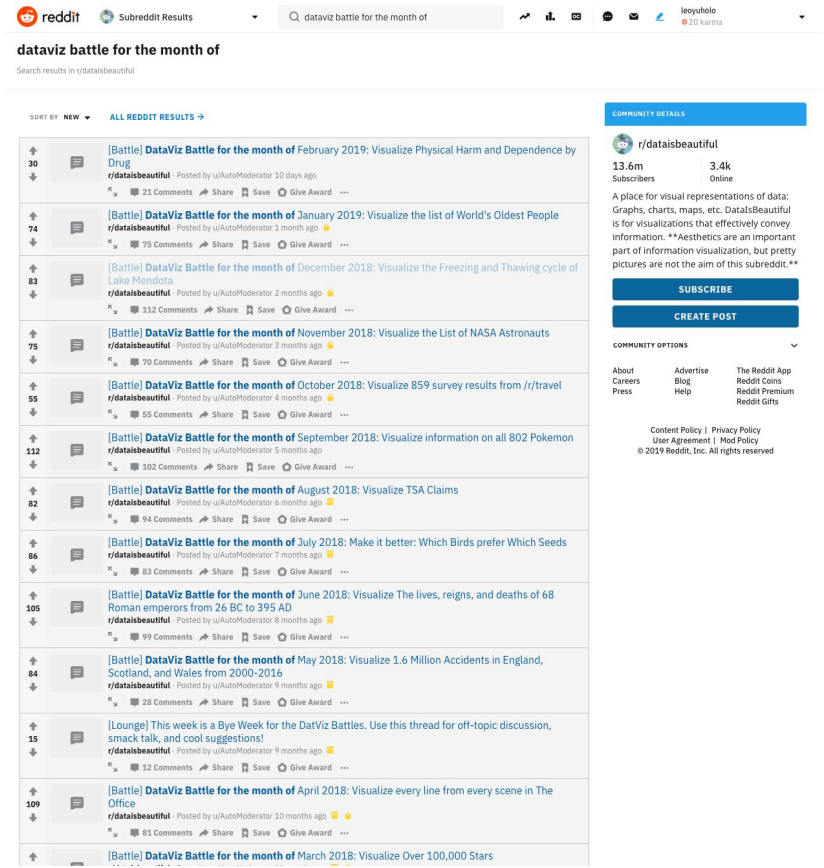
Kaggle Datasets

- 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



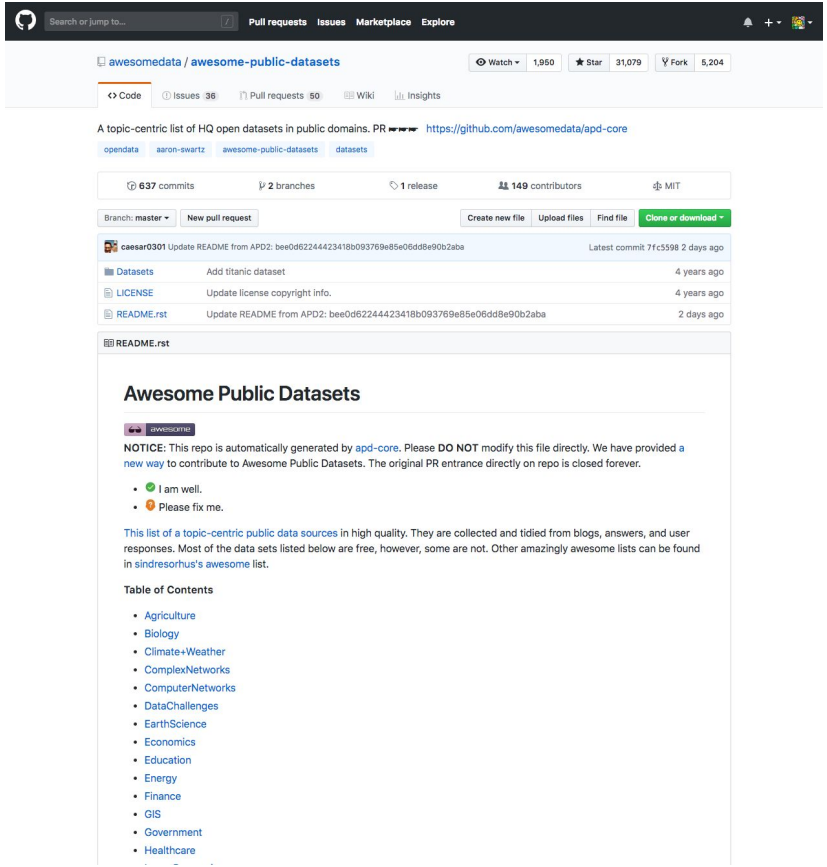
Dataviz Battle on r/dataisbeautiful

- 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



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

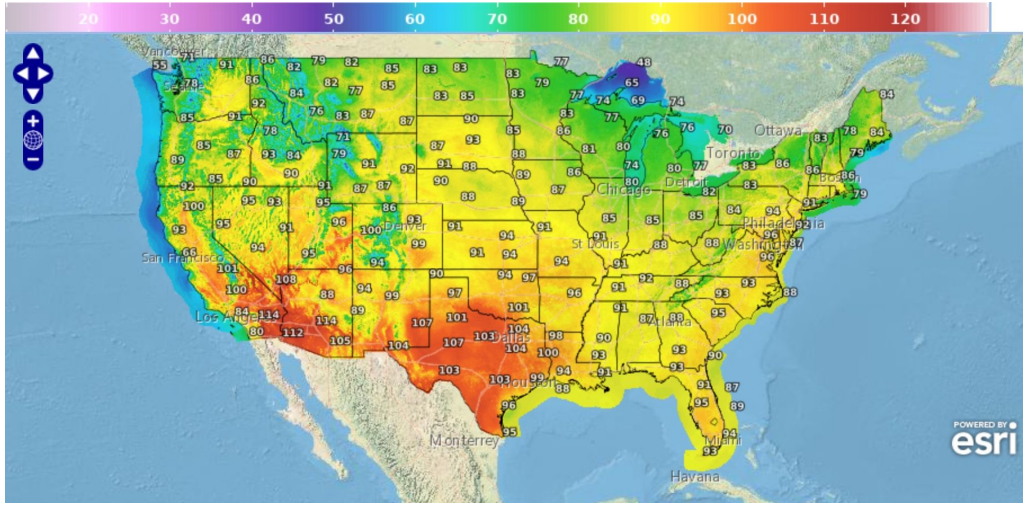
Question 2

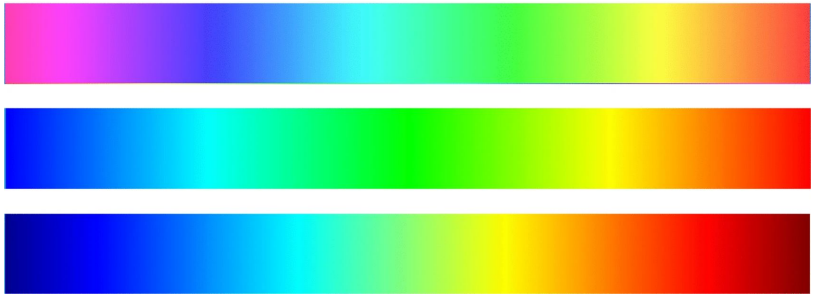
Please:

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).





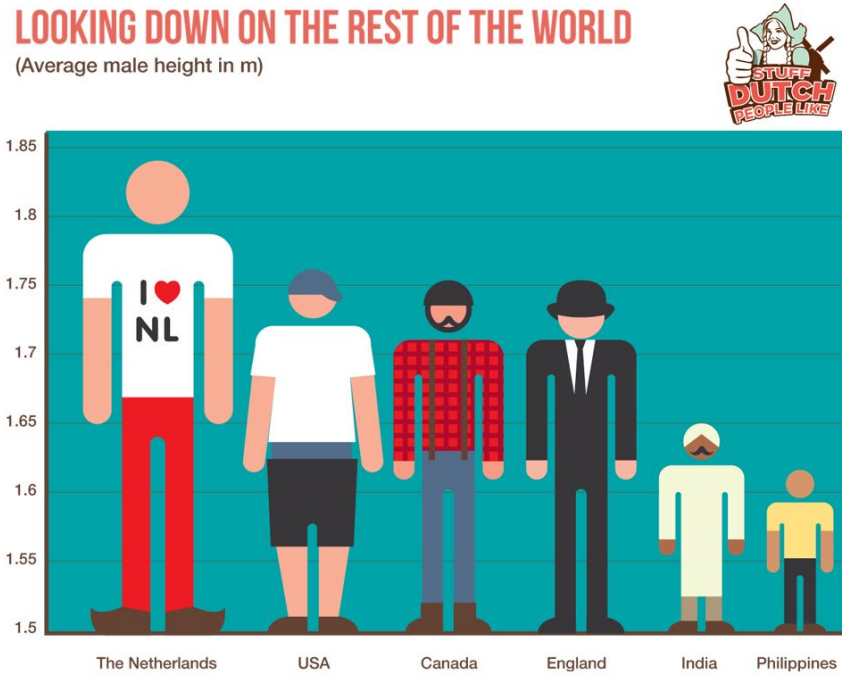
Question 2 answers and grading scheme

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).



- 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

(⅓: 1 point, ⅔: 2 point, all: 2.5 point)

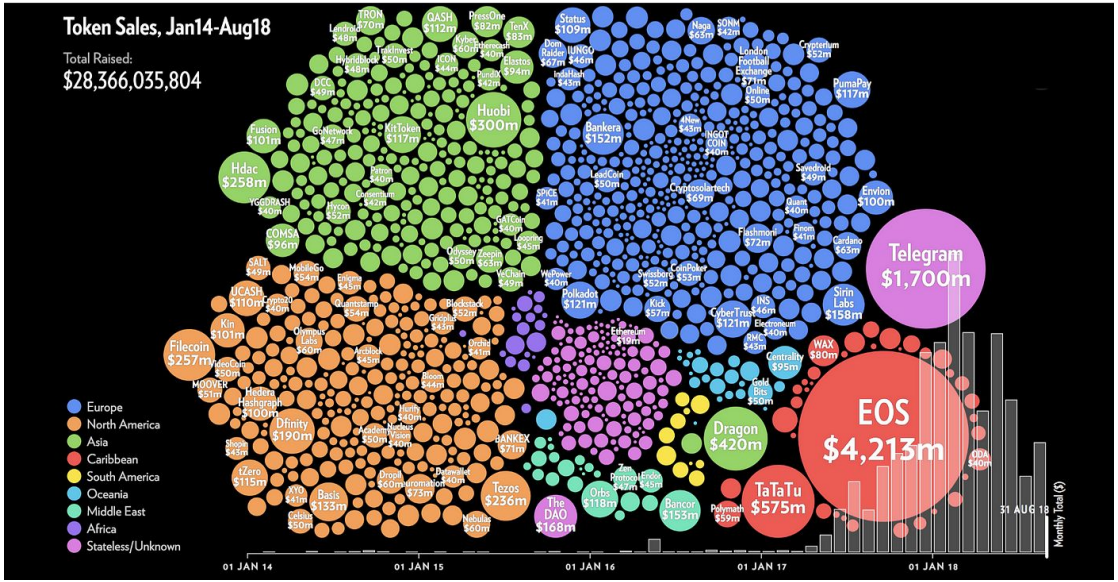
Takeaways: do not repeat on one violation of principles; Observe the data: x,y -> x*y -> others

Question 4

Please describe:

1) what marks and visual channels are used to encode what types of data in this visualization and

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).