Data Visualization: Introduction and Overview

```
$ echo "Data Sciences Institute"
```

Acknowledgement

We wish to acknowledge this land on which the University of Toronto operates. For thousands of years it has been the traditional land of the Huron-Wendat, the Seneca, and most recently, the Mississaugas of the Credit River. Today, this meeting place is still the home to many Indigenous people from across Turtle Island and we are grateful to have the opportunity to work on this land.

Welcome!

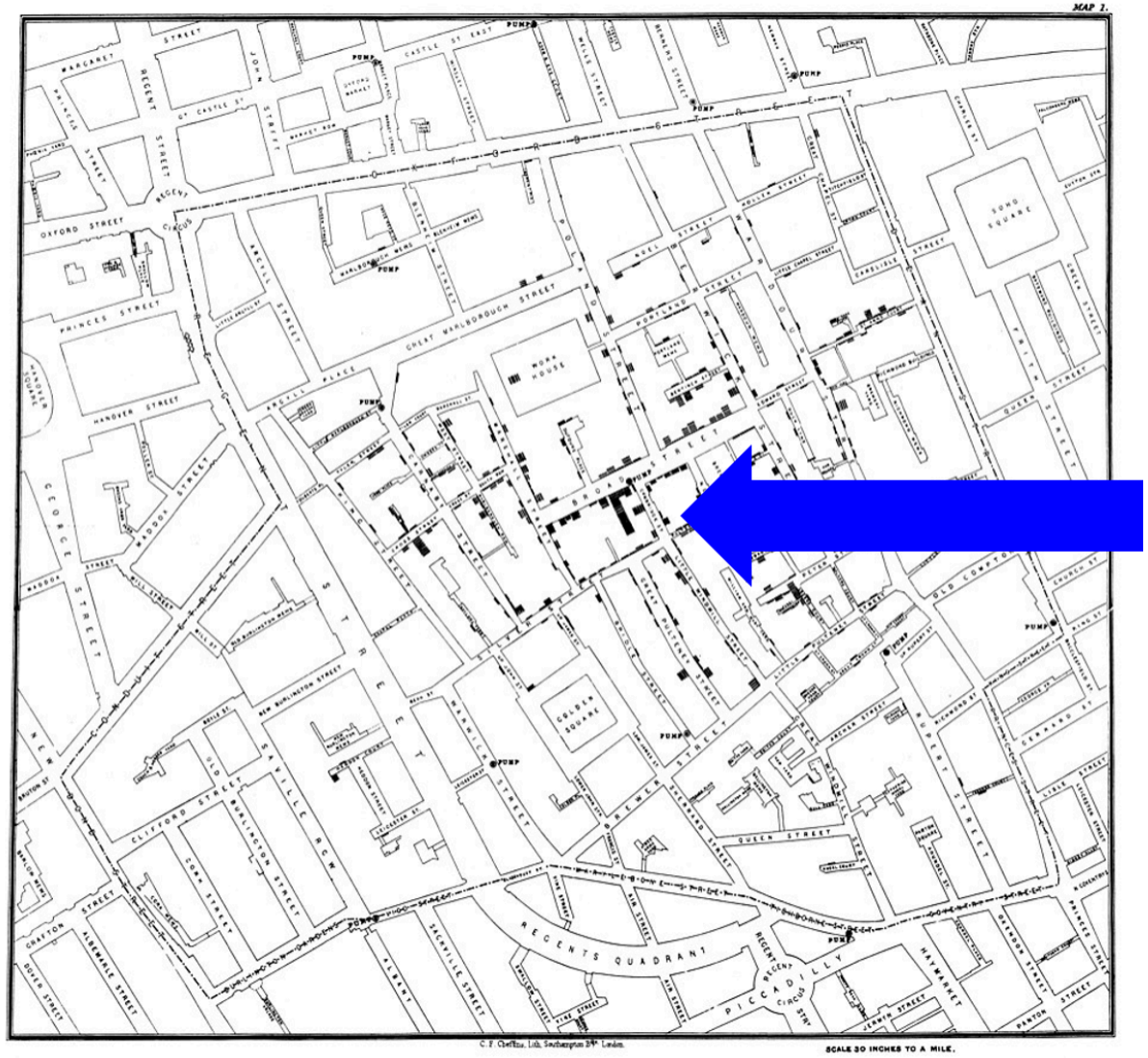
- TF: krystal Wang (she/her) fw2400@cumc.columbia.edu
- LS: Tianyi tianyi21e@gmail.com
- LS: Anjali Deshpande deshpande2013@gmail.com
- LS: Vishakh Patel Vishakh8128@gmail.com

Overview of this slide deck

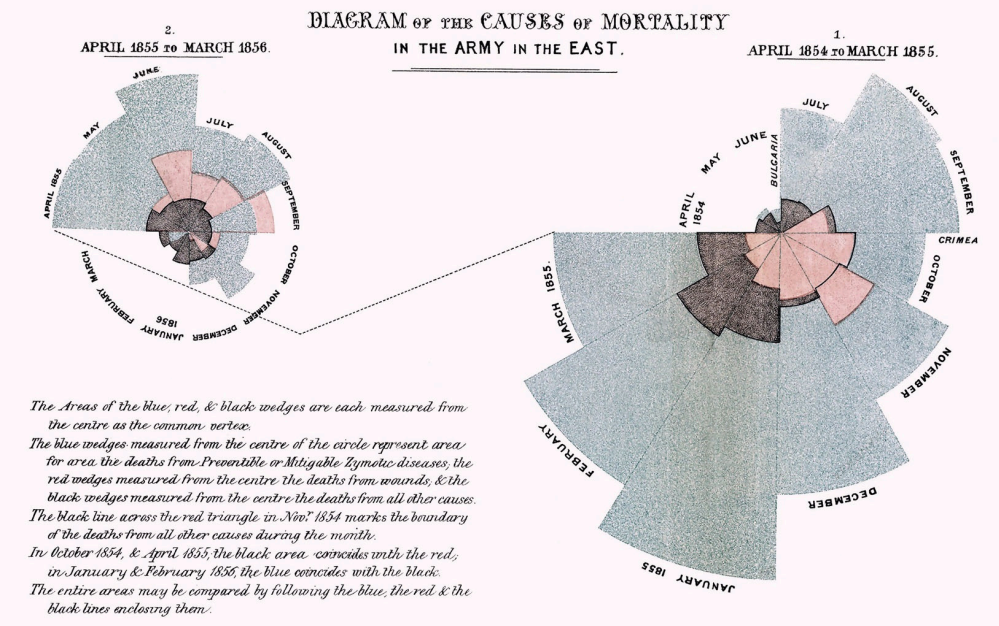
- Explore why we should care about data visualization
- Question what makes ‘good’ data visualization
- Introduce a range of software and tools that are used for data visualization

Case Study: Why should we care about data visualization?

- No matter how good or groundbreaking our data science work is, if we can’t communicate it, its impact will be severely limited.
- Often, the best way to communicate insights from data is in visual format.
- We can see examples of this idea throughout history.

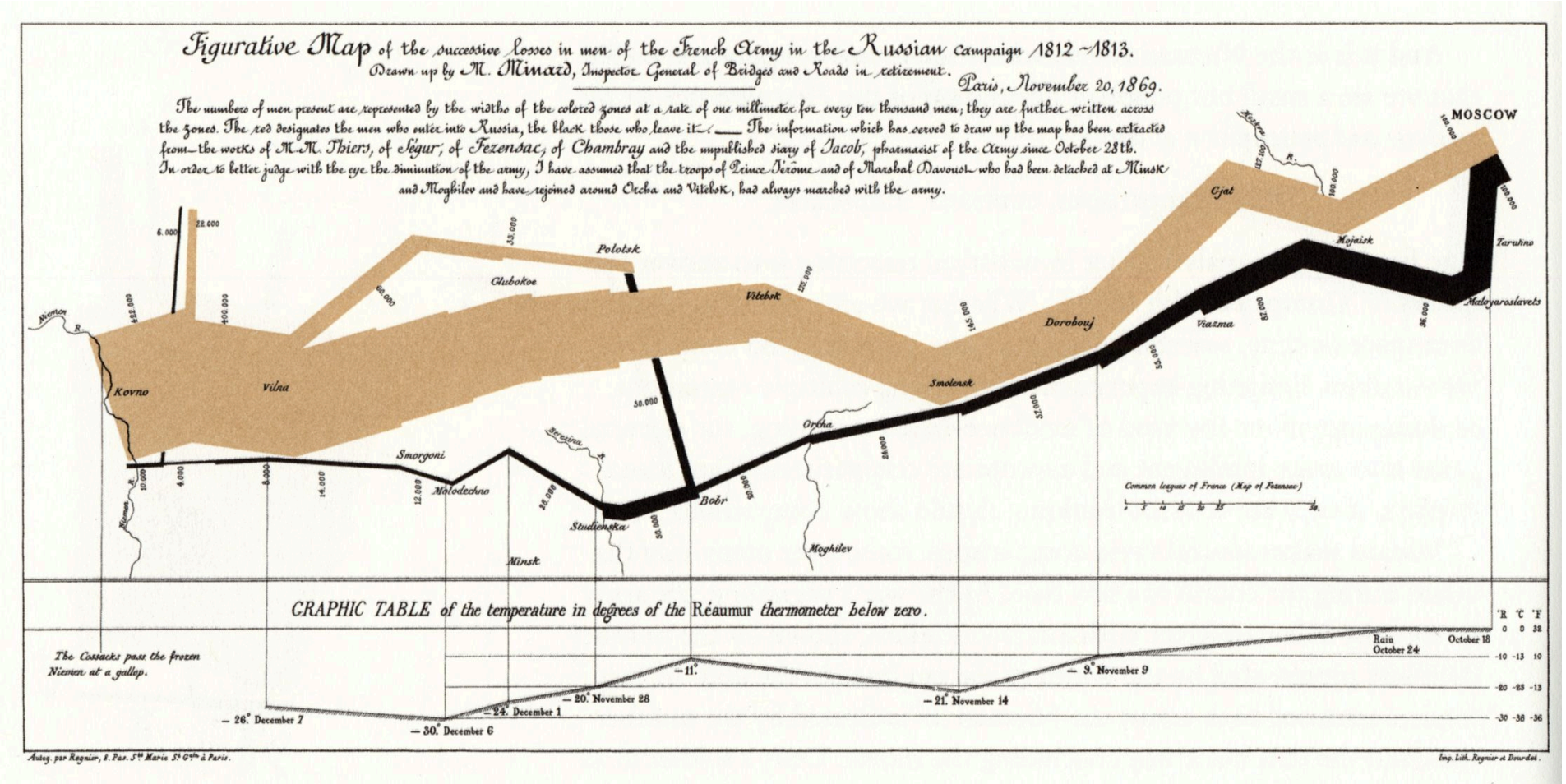


- By plotting black bars at the locations of each cholera death, Dr. John Snow was able to provide evidence tracing an 1854 outbreak to a specific water pump.
- This visualization is often cited as one of the early origins of the field of epidemiology.



- This 1858 visualization by Florence Nightingale shows soldiers' deaths due to wounds in battle (pink), other causes (black), and disease (blue) and was used to advocate for prioritizing nutrition, ventilation, and shelter, revolutionizing army medicine.

- Minard, in 1869, visualized: "Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812–1813."



So why visualize information?

- These historical examples gave us different reasons
  - Cholera outbreak → Allows us to see data in context
  - Nightingale's rose diagram → Presents an argument and advocates
  - Napoleon's Russian Campaign → An example of story telling

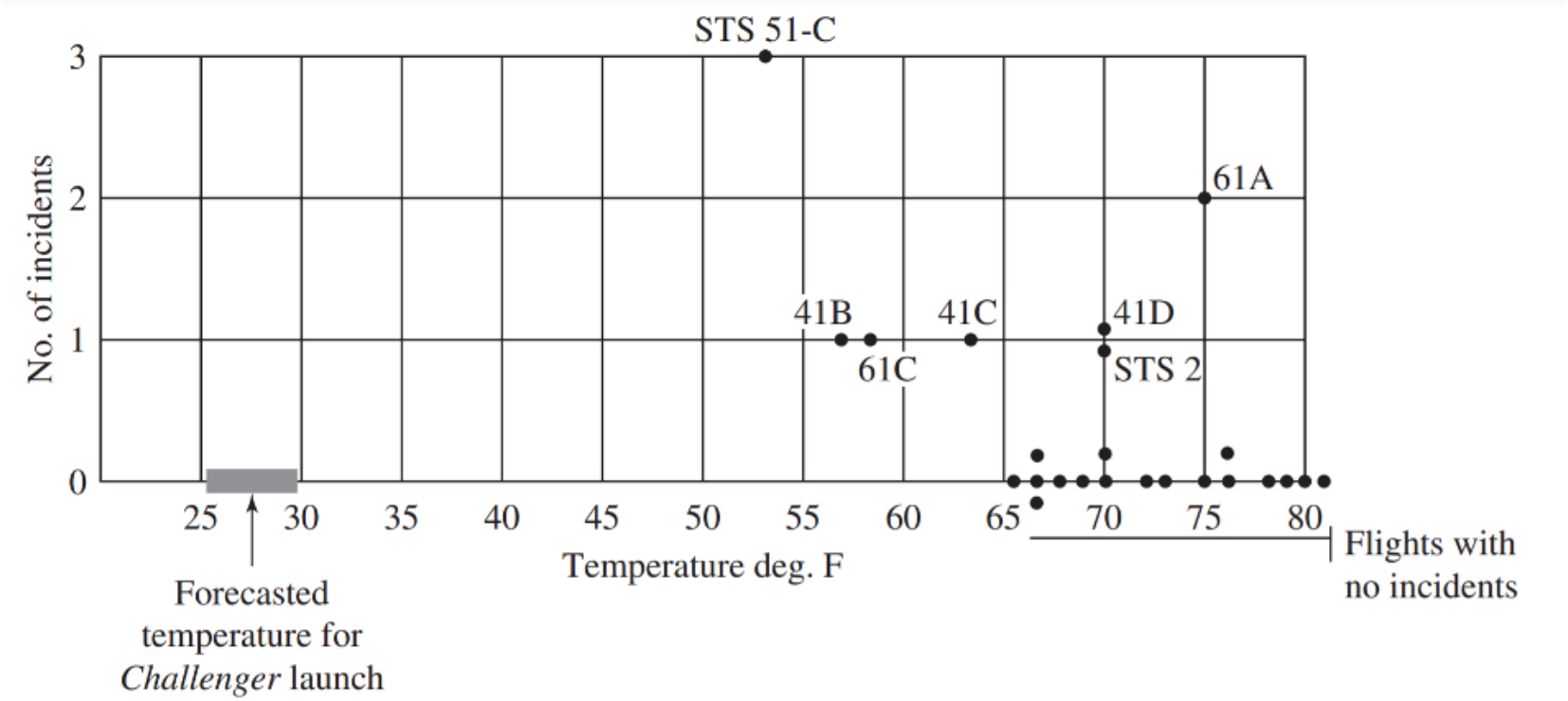
But data visualization is not always so straightforward...

The Challenger Disaster

- In 1986, the space shuttle *Challenger* exploded 73 seconds after launch, killing everyone aboard.
- The explosion occurred because hot propellant gases burned through rubber seals (“O-rings”) on the shuttle’s right solid rocket booster.
- For months, up until the night before the launch, concerns about the O-rings and the safety of the launch had been raised, but launch proceeded anyway

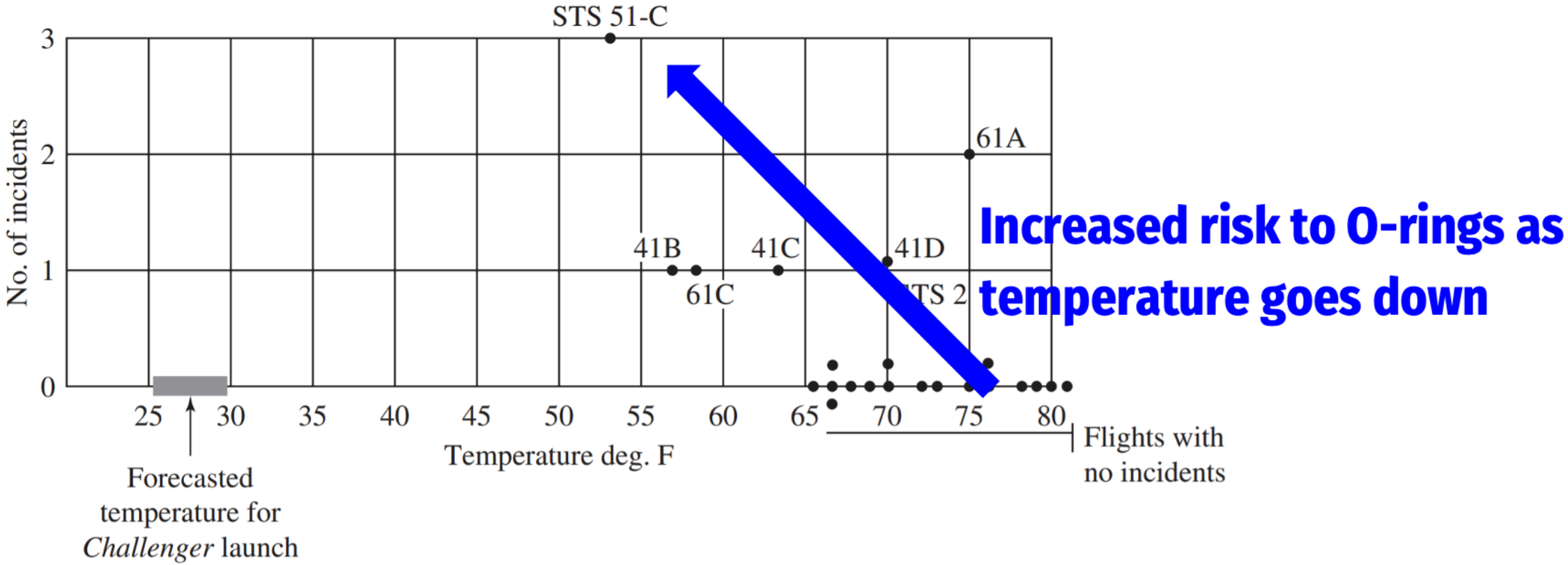
The Challenger Disaster - Tufte's Graph

- In 1997, Edward Tufte famously published a visualization showing the relationship between temperature during launches of test shuttle flights (pre-*Challenger*) and incidents of damage to O-rings (image: Tufte, 1997 👇)



The Challenger Disaster - Tufte's Graph

- Tufte's graph seems to show that as launch temperatures decreased, O-ring incident rate increased; thus, by launching at temperatures lower than those tested, NASA unnecessarily endangered *Challenger* (image: Tufte, 1997 👇)

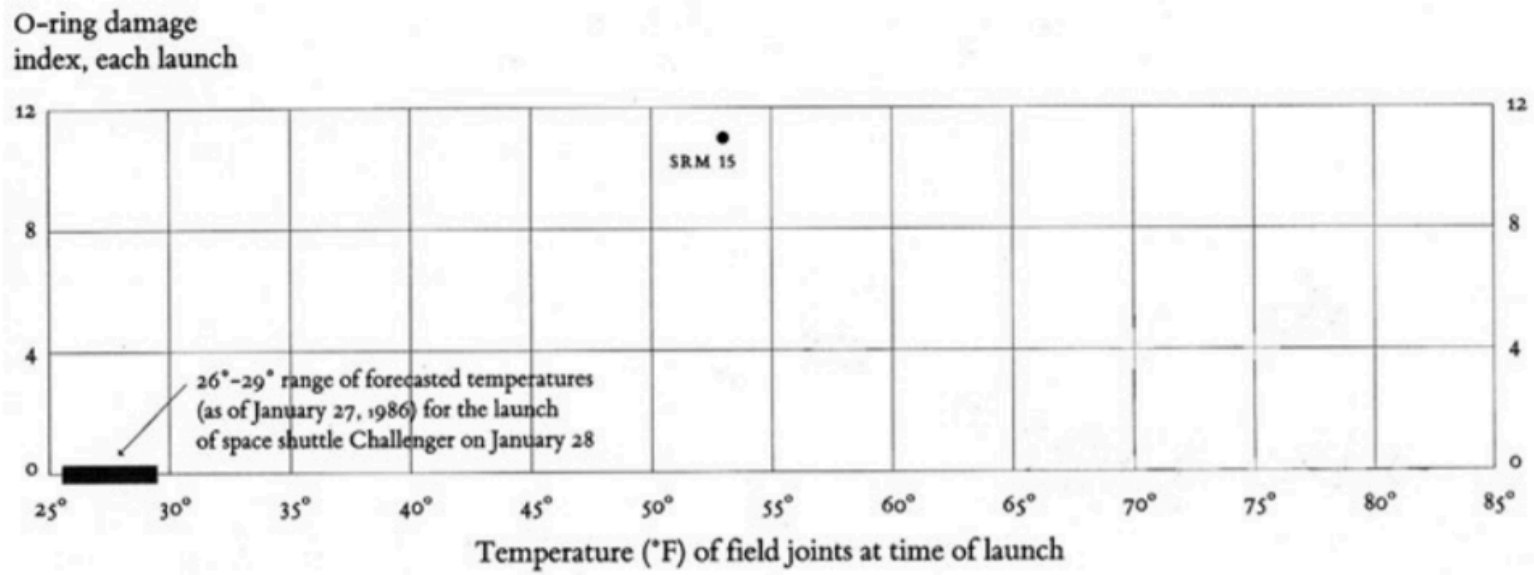


The Challenger Disaster - Tufte's Graph

- Tufte argued that the engineers responsible for communicating the results of the O-ring tests had failed to represent the data that might have saved the crew of *Challenger*.
- Tufte's graph is used as a classic case study demonstrating the importance of data visualization as a way to communicate complex information.
- Another case of data visualization saving lives?

The Challenger Disaster - But...

- The accuracy of Tufte's visualization has been debated, with Robison identifying several errors in Tufte's use of data.
- Robison suggests that only one test launch had actually produced relevant O-ring temperature data, producing this visual instead: (image: Robison, 1997, 2002 👇)



The Challenger Disaster - But...

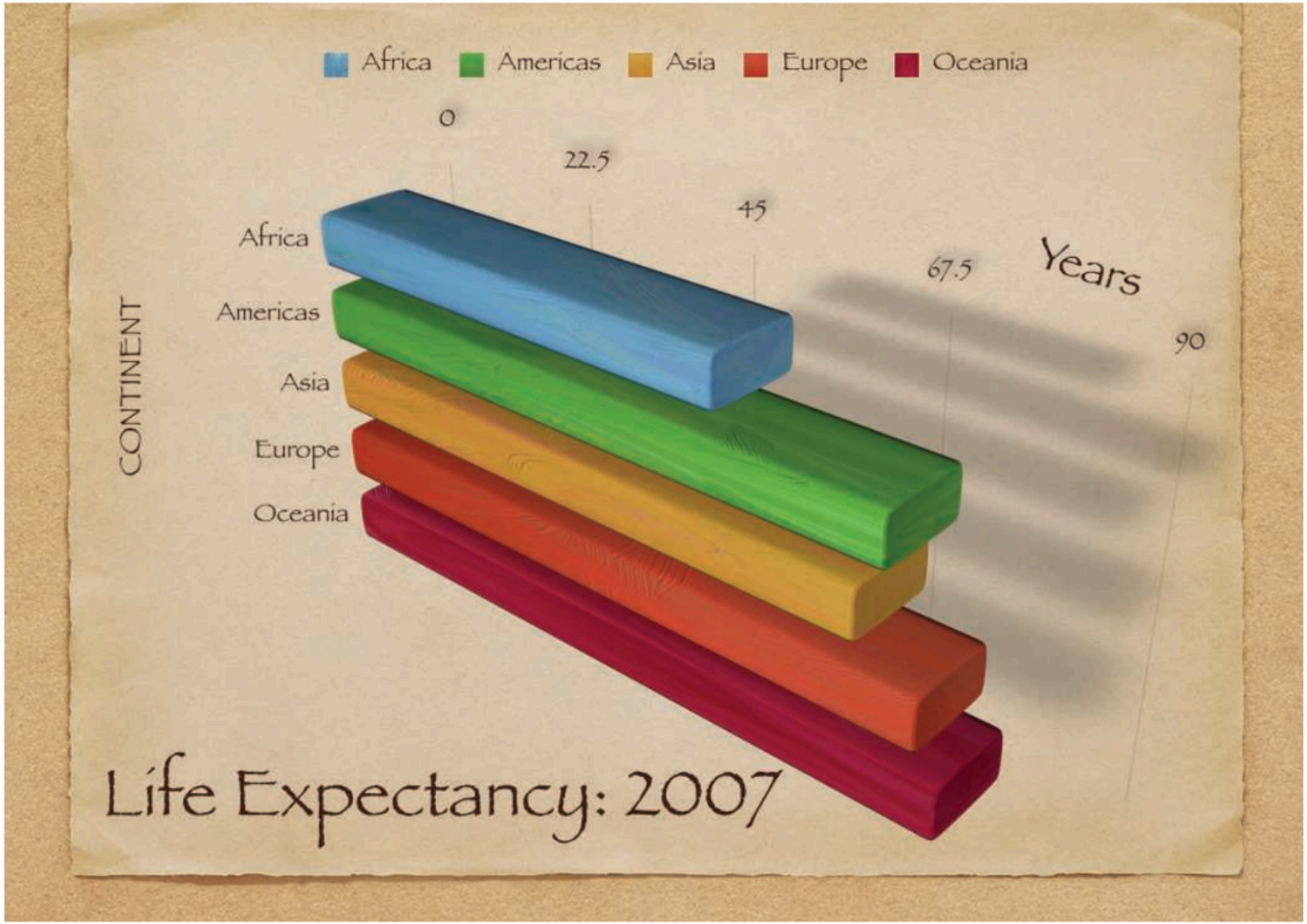
- Per Tufte, the *Challenger* disaster is a case where good data visualization could have facilitated understanding of complex data and saved lives.
- Per Robison, Tufte's work is a case of bad data visualization unfairly placing blame and leading the audience to a faulty conclusion.

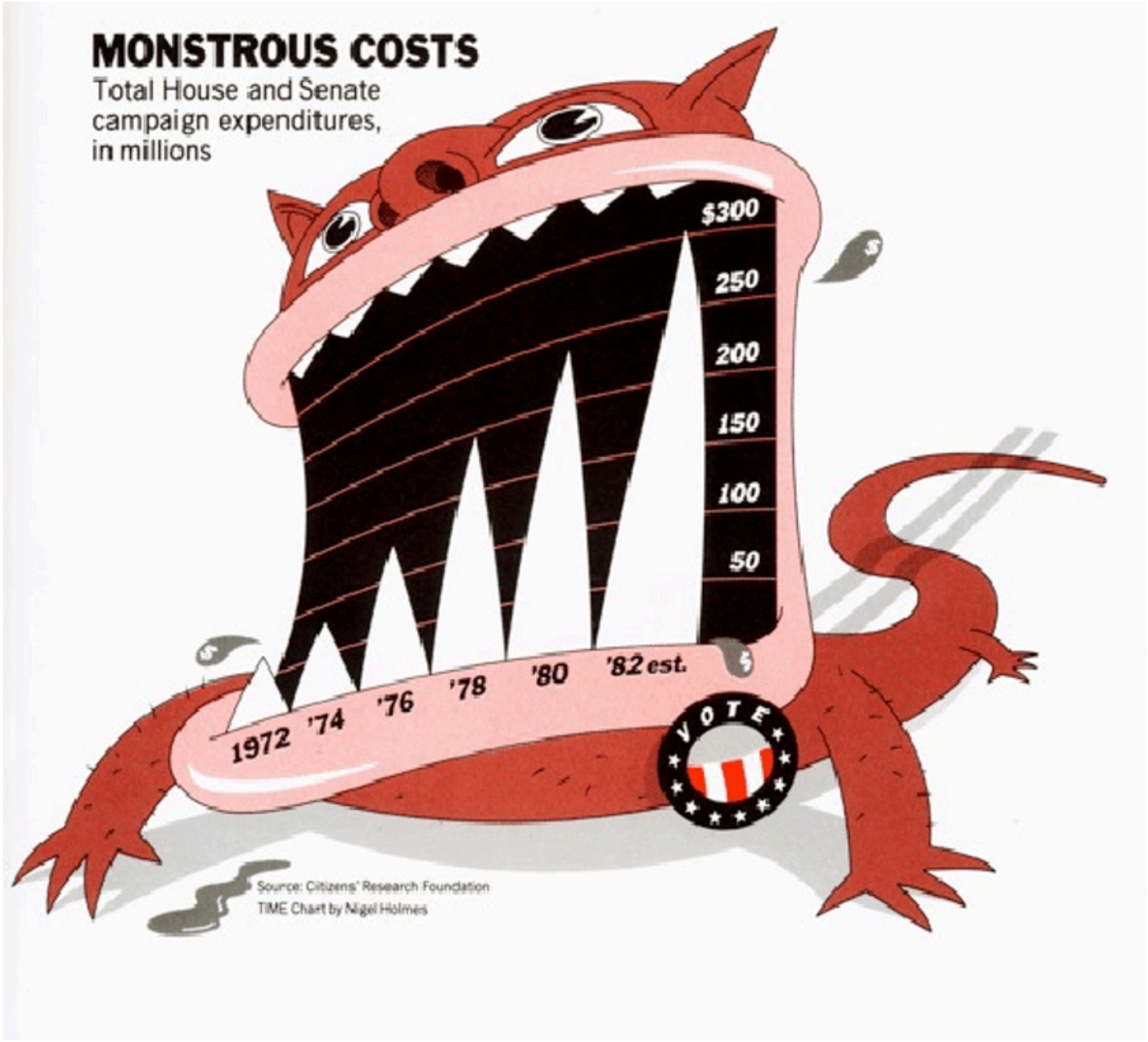
The choices we make about visualizing our data have consequences - so how do we make better ones?

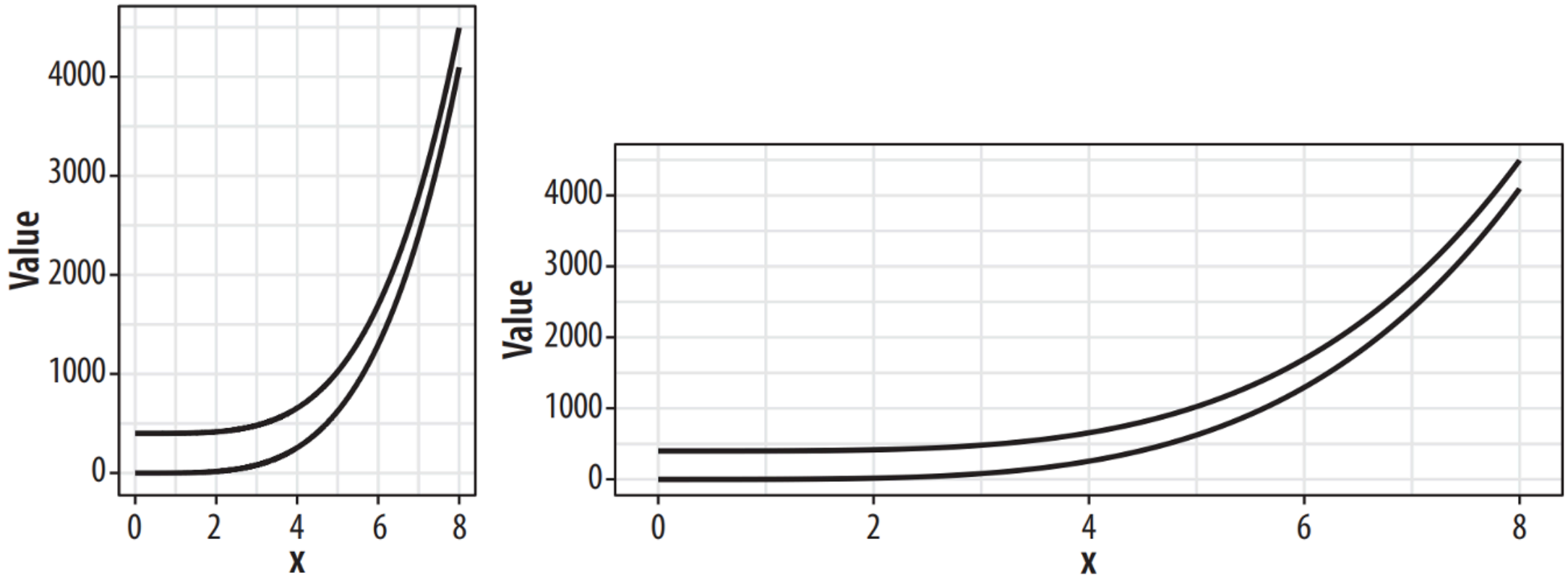
Activity: What is ‘good’ data visualization?

Activity

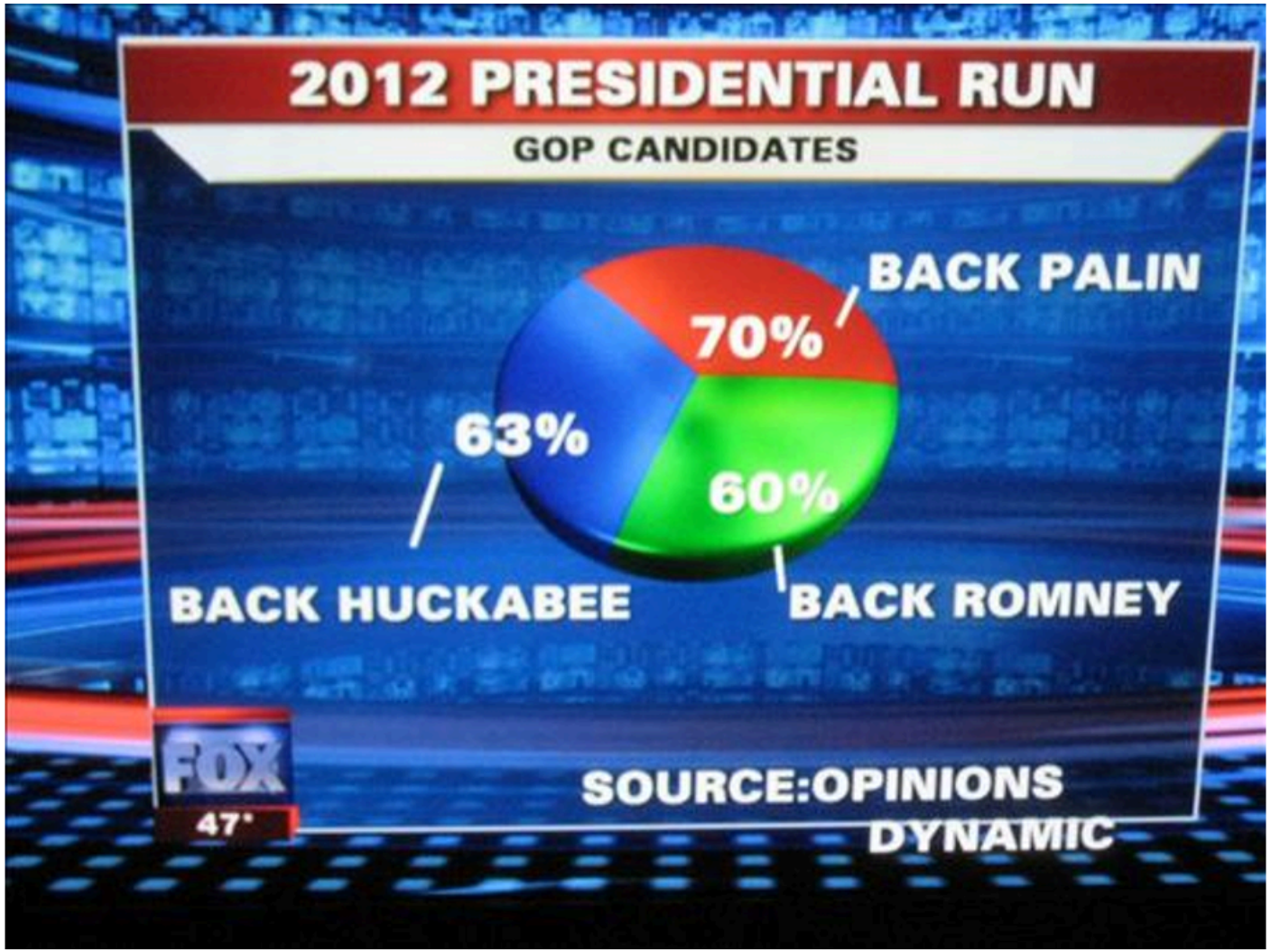
- Look at the following examples of data visualizations. For each, consider:
  - Is the visualization pleasing to look at?
  - Does the visualization accurately represent data?
  - Can we understand what message the maker of the visualization is attempting to convey?
  - Consider factors such as colour, size, use of images. Is this a ‘good’ data visualization? Why or why not?


Africa
Americas
Asia
Europe
Oceania
0
22.5
45
67.5
90
Years
CONTINENT
Africa
Americas
Asia
Europe
Oceania
Life Expectancy: 2007

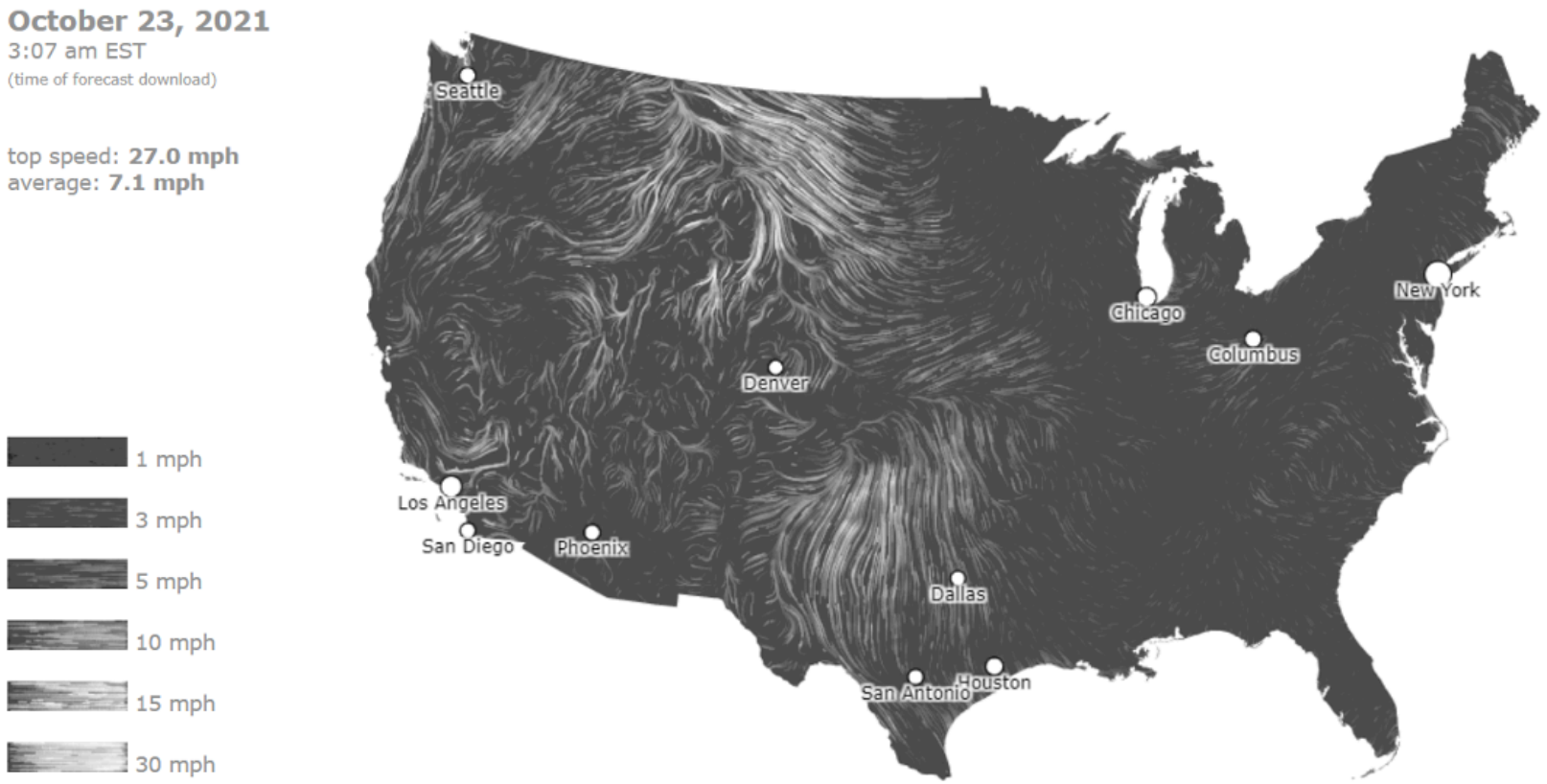

MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions
$300
250
200
150
100
50
1972
'74
'76
'78
'80
'82 est.
VOTE
Source: Citizens' Research Foundation
TIME Chart by Nigel Holmes



(The same data presented with two different aspect ratios)


2012 PRESIDENTIAL RUN
GOP CANDIDATES
BACK PALIN
70%
63%
60%
BACK HUCKABEE
BACK ROMNEY
FOX
47'
SOURCE:OPINIONS
DYNAMIC

wind map



(Click image to view interactive visualization)

Activity

- Each of the questions corresponds to an important quality of data visualizations:
  - Is the visualization pleasing to look at? → **Aesthetic**
  - Does the visualization accurately and honestly present data? → **Substantive**
  - Can we understand what message the maker of the visualization is attempting to convey? → **Perceptual**
- We need to consider all of these qualities when evaluating and designing ‘good’ data visualizations

What data visualization IS

- Dependent on:
  - **Context** → ***where and how*** will our visualization be used? (eg. academic journal, poster, infographic)
  - **Audience** → ***who*** is intended to use our visualization? (eg. subject experts, general public)
  - **Data structure** → ***what*** information do our data capture? (eg. quantities, relationships)

What data visualization is NOT

- Hard and fast rules for every situation → **visualizing data means making decisions**

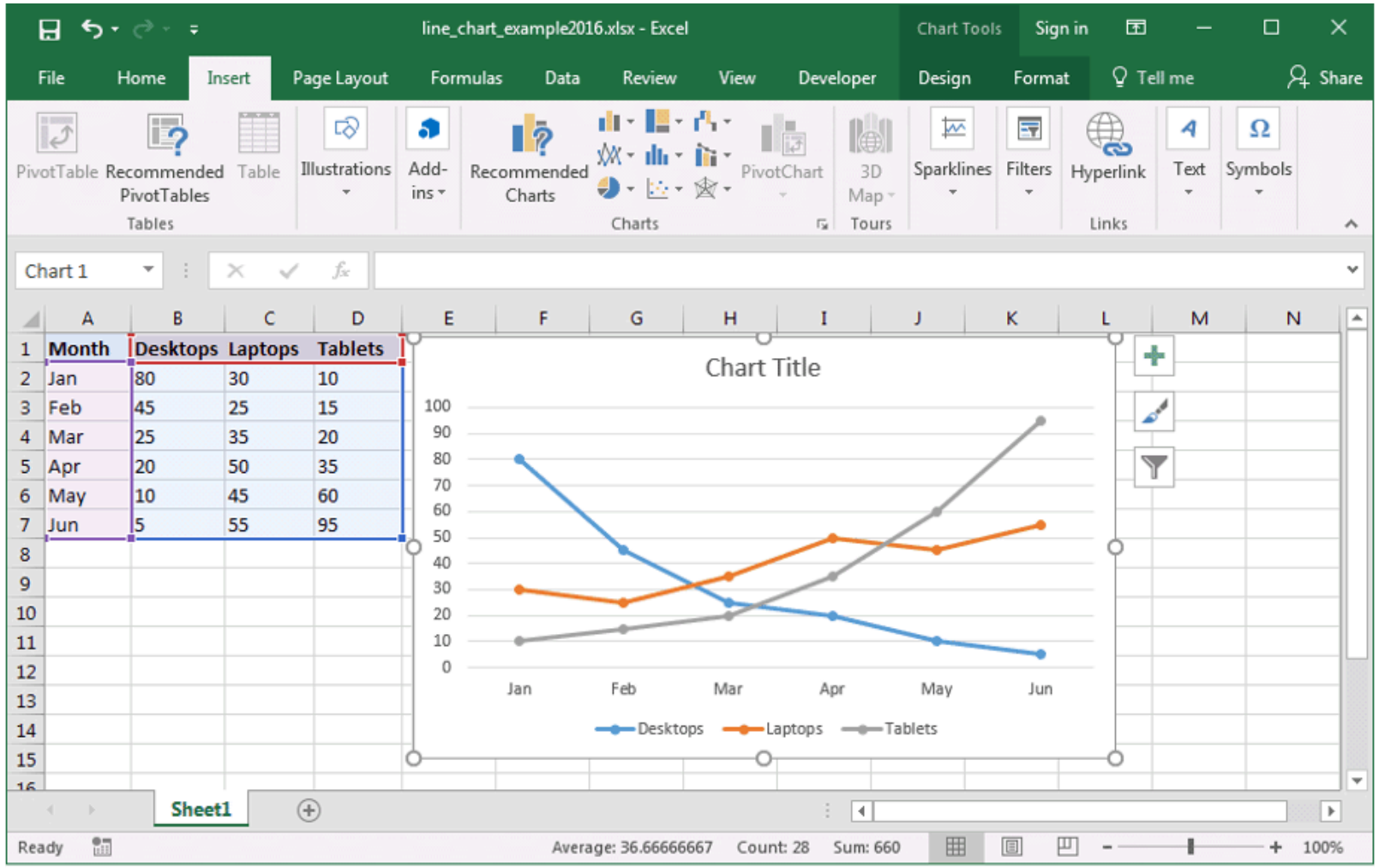
What tools are used for data visualization?

Tools for Data Visualization

Microsoft Excel (LibreOffice Calc, Google Sheets, etc)

What is it	Spreadsheet software with ability to generate static data visualizations
Access	- Excel is paid (part of MS Office Suite) - Free alternatives such as Google Sheets and LibreOffice Calc
Reproducible visualizations	- no
Ease of use	- Point and click to select from pre-made visualizations - Can serve as frontend for databases using Power Query
Use cases	- “First line tool” for data analysis and visualization - pretty much everywhere

Microsoft Excel (LibreOffice Calc, Google Sheets, etc)



TechOnTheNet.com

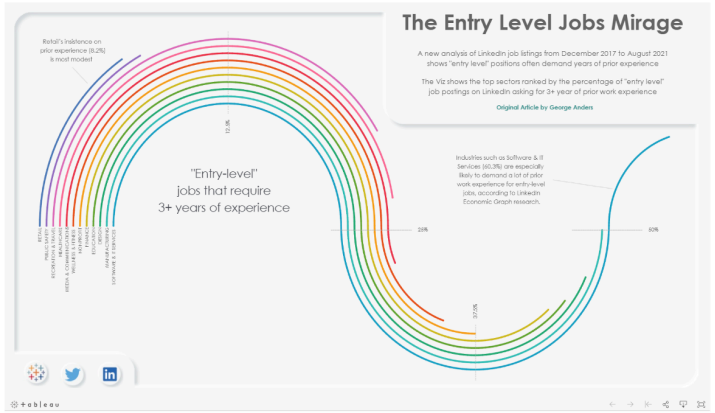
Tableau, Tableau Public

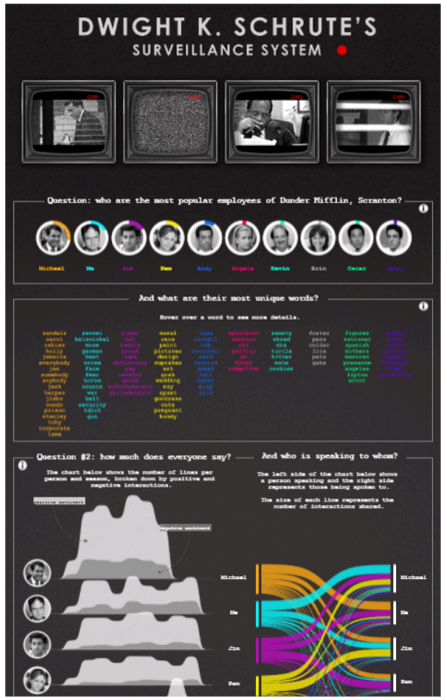
What is it	Combines data from different sources (databases, spreadsheets) into interactive, dynamic visualizations on the web
Access	- Tableau Server and Desktop are paid - Tableau Public is free **but** all visualizations are public and not saved locally
Reproducible visualizations	- no
Ease of use	- Point and click to select from pre-made visualizations
Use cases	- Industry (designed for business intelligence), infographics for media

Tableau, Tableau Public

Click each link to view and interact with public visualizations chosen as Tableau's 'Viz of the Day':

🔗 left, 🔗 middle, 🔗 right

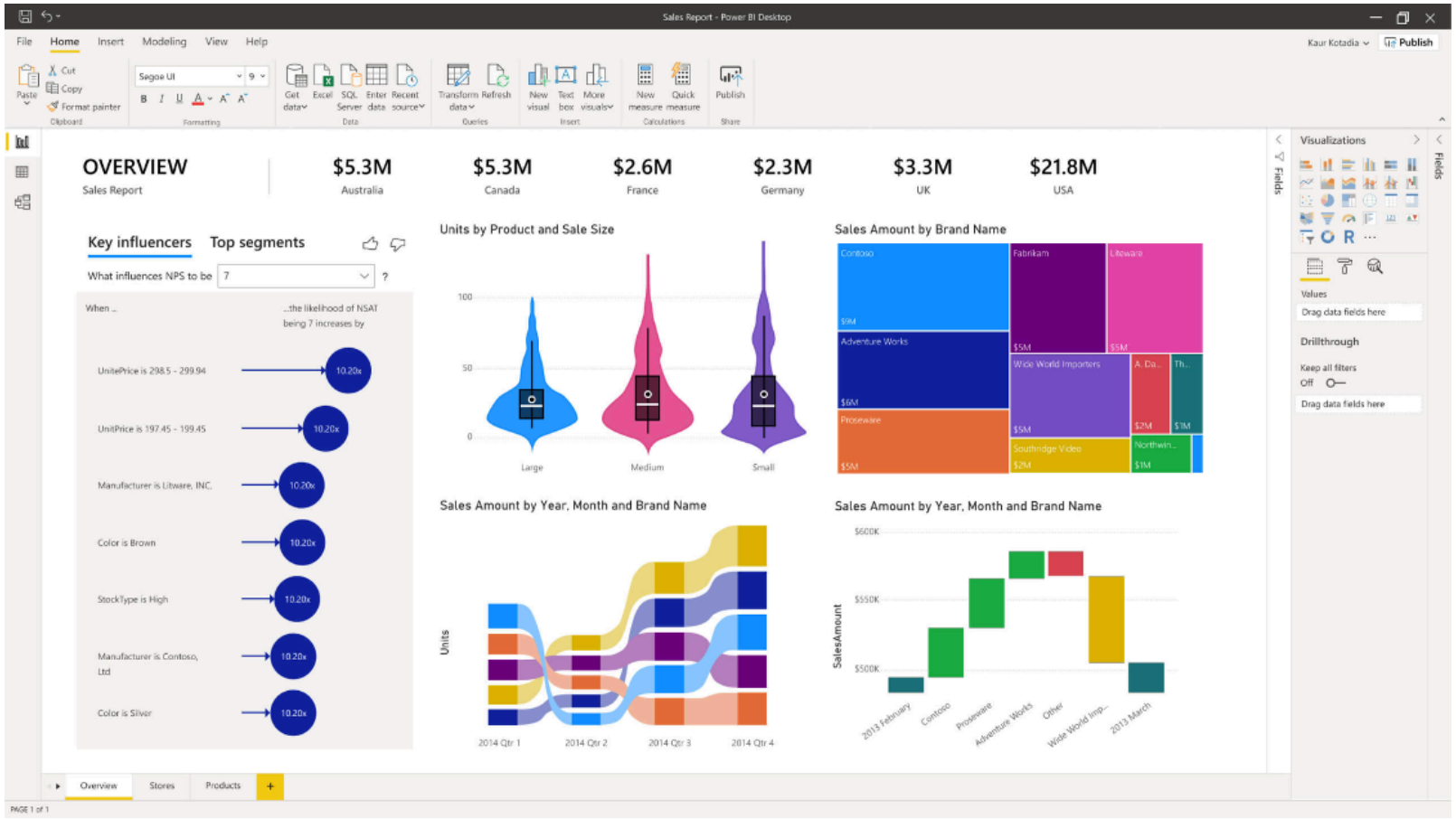




Microsoft Power BI

What is it	Combines data from different sources (databases, spreadsheets) into interactive, dynamic visualizations on the web
Access	Paid (part of MS Office Suite)
Reproducible visualizations	no
Ease of use	Drag and drop to select from pre-made visualizations Can use DAX (Data Analysis Expressions) functions to perform operations on data
Use cases	Industry, government (designed for business intelligence)

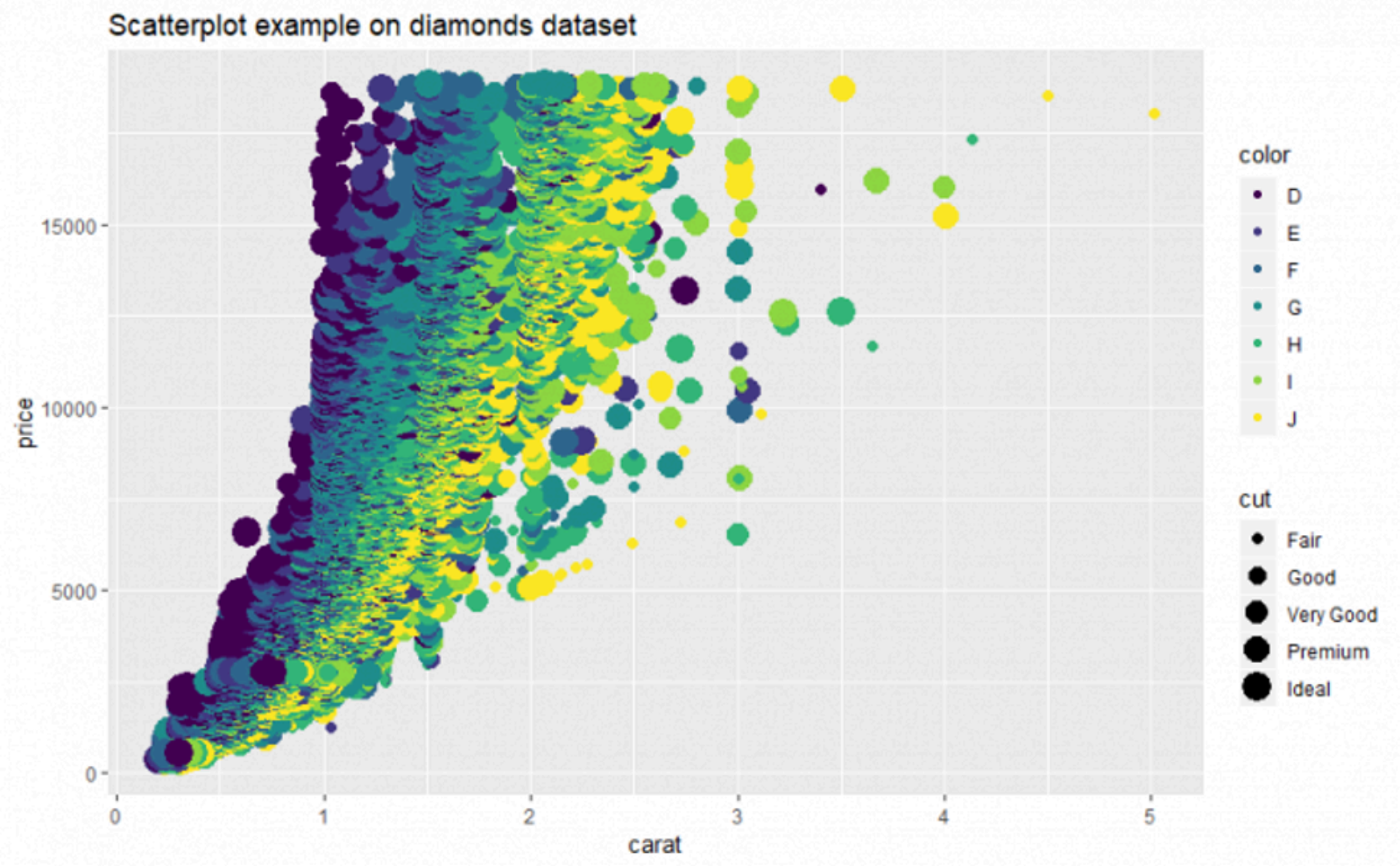
Microsoft Power BI



R

What is it	Programming language with libraries for data visualization (e.g., ggplot2, Plotly, RColorBrewer)
Access	Free and open source (https://www.r-project.org/COPYING)
Reproducible visualizations	yes
Ease of use	Programming language; requires some coding knowledge
Use cases	Academia, industry, government; research and data science contexts

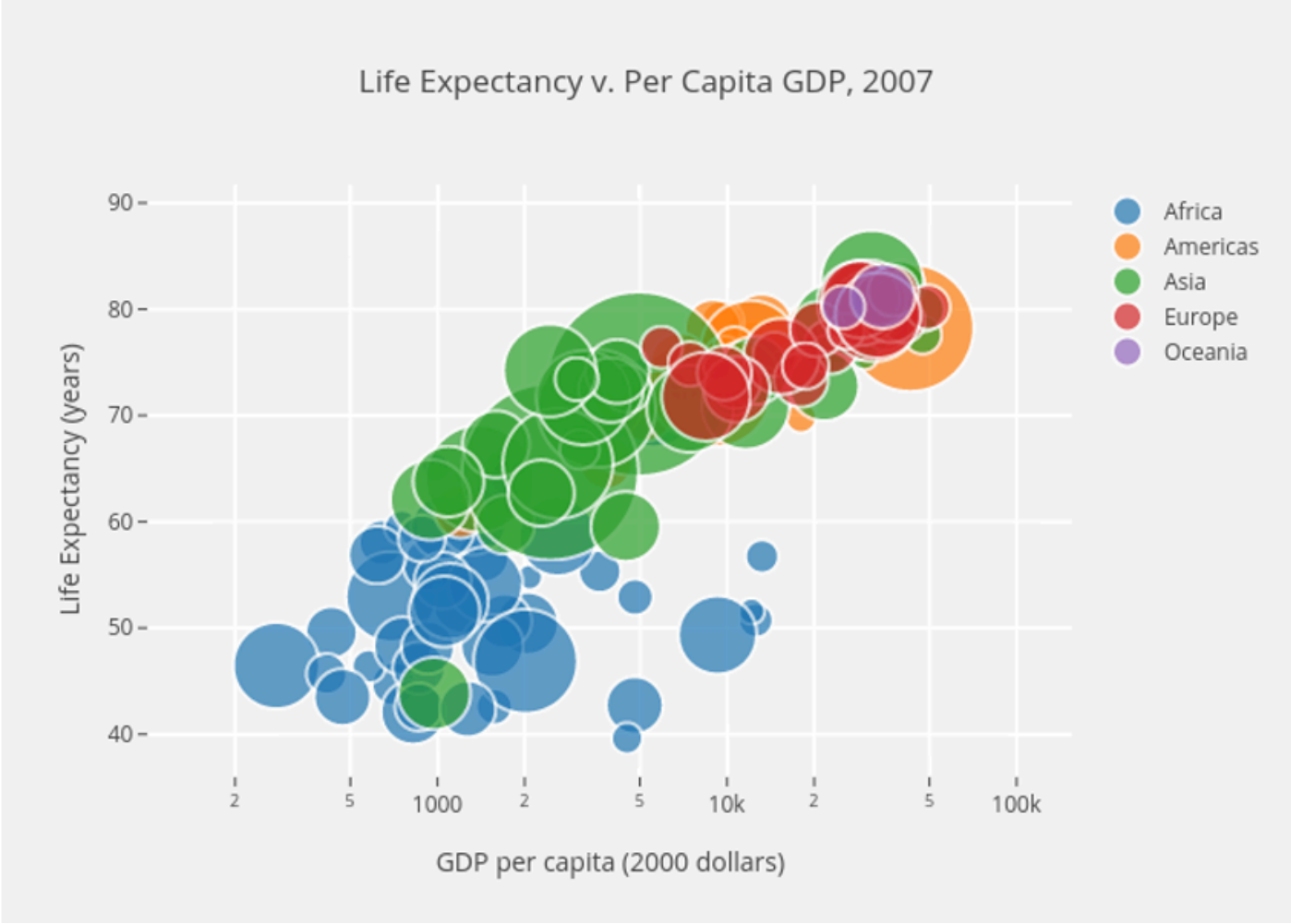
R


Scatterplot example on diamonds dataset
price
15000
10000
5000
0
0
1
2
3
4
5
carat
color
D
E
F
G
H
I
J
cut
Fair
Good
Very Good
Premium
Ideal

Python

What is it	Programming language with libraries for data visualization (e.g., Matplotlib, Plotly)
Access	Free and open source Python Software Foundation License
Reproducible visualizations	yes
Ease of use	Programming language; requires some coding knowledge
Use cases	Government, industry, academia; data science and programming contexts

Python


Life Expectancy v. Per Capita GDP, 2007
Life Expectancy (years)
90
80
70
60
50
40
2
5
1000
2
5
10k
2
5
100k
GDP per capita (2000 dollars)
Africa
Americas
Asia
Europe
Oceania

For our purposes

- Step-by-step walkthroughs are focused on Python (MatPlotLib)
  - Commonly used tool
  - Free and open source
  - Reproducible
  - LOTS of available resources online

BUT...

- General design principles will apply to creating data visualizations in whichever software you decide to use

Coming Up in this Course - Topics

- **First Steps**
  - Get started
  - Make a plot (Matplotlib)
  - Thinking about reproducibility
- **Graphing Our Data**
  - Customize our plot appearance
  - Choosing the right visualization (perceptual qualities)

Coming Up in this Course - Topics

- **Visualization with Purpose**
  - Subplots and combining visualizations
  - Colour theory and accessible design
  - Data visualization as advocacy
- **Getting Fancy**
  - Beyond matplotlib (Seaborn)
  - Interactive data visualizations (Plotly)
  - Qualitative data visualization

Learning Outcomes of this Course

1. Create and customize data visualizations start to finish in Python
2. Use general design principles for creating accessible and equitable data visualizations in Python and other software
3. Use data visualization to purposely tell a story