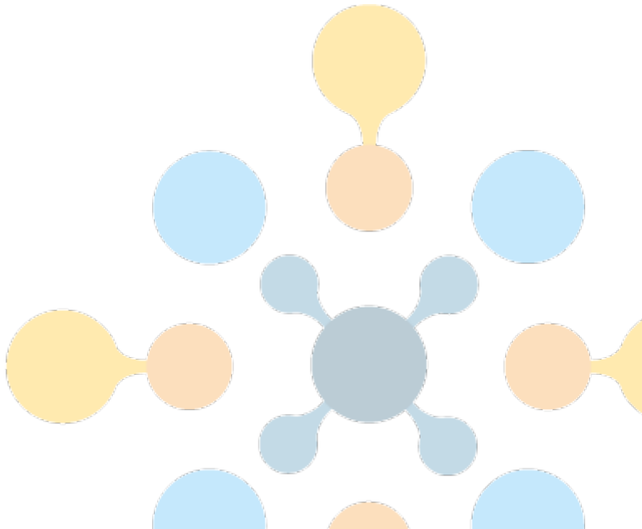
An Empirical Evaluation of the GPT-4 Multimodal Language Model on Visualization Literacy Tasks

Alexander Bendeck
John Stasko

Georgia Tech Visualization Lab



September 25, 2023

ChatGPT can now see, hear, and speak

September 25, 2023

ChatGPT can now see, hear, and speak

Ten Wild Ways People Are Using ChatGPT's New Vision Feature

Published Sep 29, 2023 at 7:09 PM EDT

The New ChatGPT Can 'See' and 'Talk.' Here's What It's Like.

The image-recognition feature could have many uses

But wait…

- **What is the *visualization literacy* of multimodal LLMs?**
  - Are models like GPT-4V actually any good at reading charts?

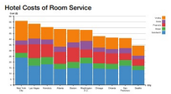
- This is a complicated question
  - There are potentially many ways to measure vis literacy

Vis Literacy of GPT-4V: Our Approach

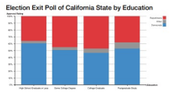
Visualization Literacy Assessment Test (VLAT)

(a) Line Chart

(b) Bar Chart



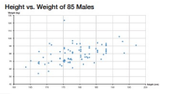
(c) Stacked Bar Chart



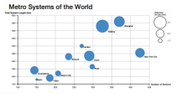
(d) 100% Stacked Bar Chart

(e) Pie Chart

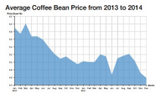
(f) Histogram

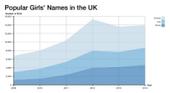


(g) Scatterplot



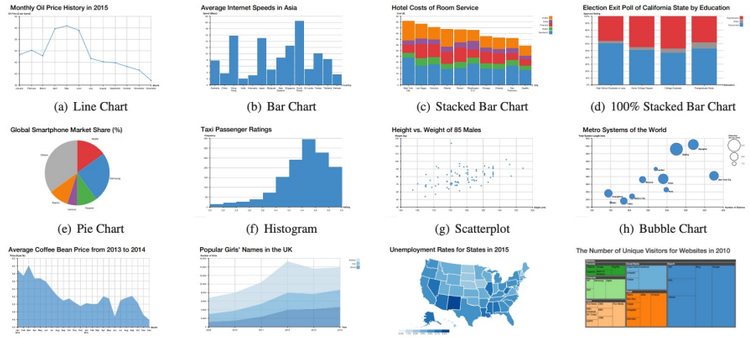
(h) Bubble Chart





Vis Literacy of GPT-4V: Our Approach

Visualization Literacy Assessment Test (VLAT)



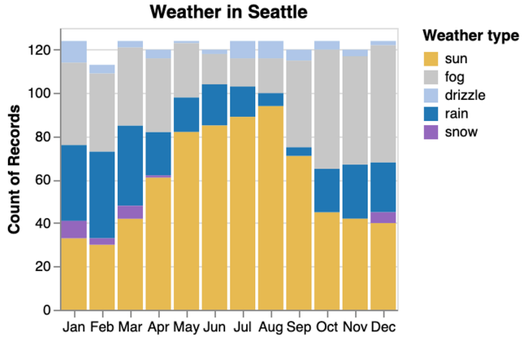
(a) Line Chart (b) Bar Chart (c) Stacked Bar Chart (d) 100% Stacked Bar Chart

(e) Pie Chart (f) Histogram (g) Scatterplot (h) Bubble Chart

Chart Question Answering

CQA Q349:
What month has the most sun?

Based on the provided chart, ***August*** *has the most sun, with the* ***yellow segment*** *(representing sunny days) being the* ***tallest in that month's column****.*



Vis Literacy of GPT-4V: Our Approach

Visualization Literacy Assessment Test (VLAT)

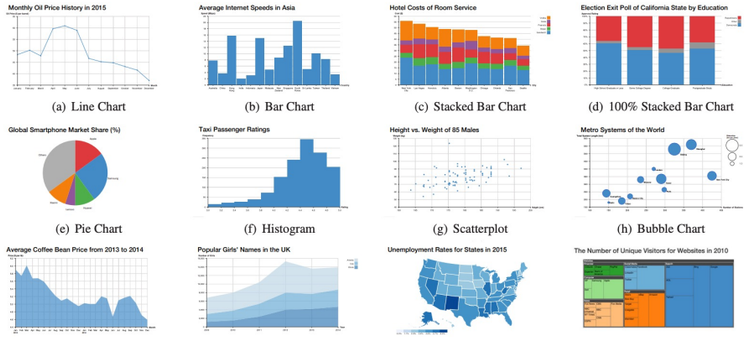
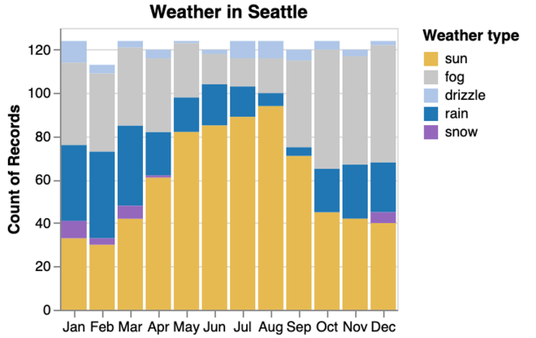


Chart Question Answering

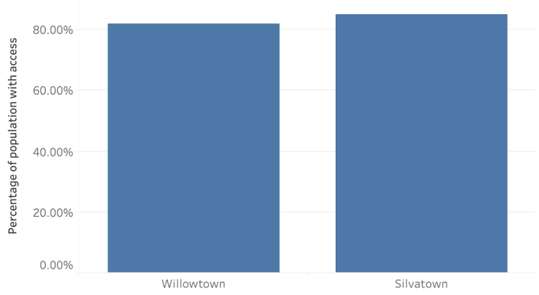
CQA Q349:
What month has the most sun?

Based on the provided chart, ***August*** *has the most sun, with the* ***yellow segment*** *(representing sunny days) being the* ***tallest in that month's column****.*

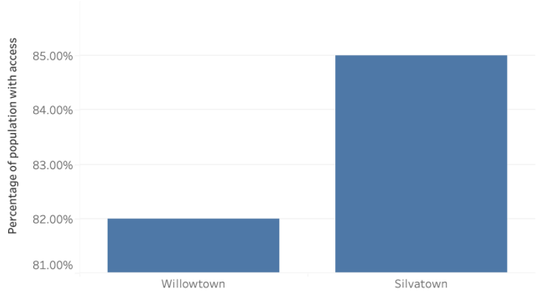


Deceptive Visualization Design Choices

Access to safe drinking water in Willowtown and Silvatown, as of 2010.



Access to safe drinking water in Willowtown and Silvatown, as of 2010.



Vis Literacy of GPT-4V: Our Approach

Visualization Literacy Assessment Test (VLAT)

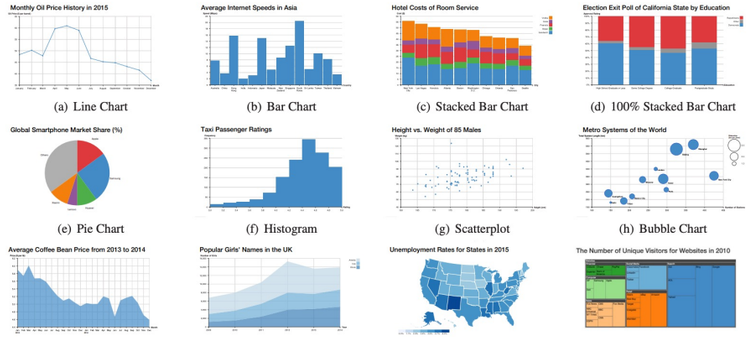
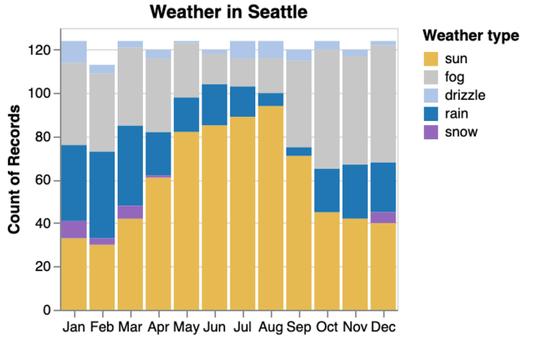


Chart Question Answering

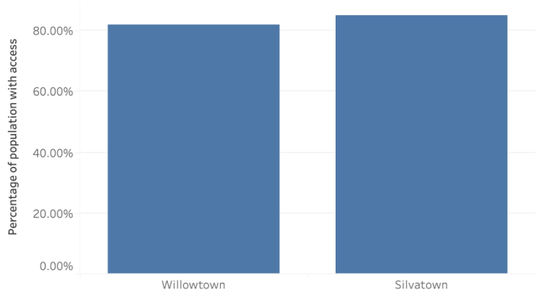
CQA Q349:
What month has the most sun?

Based on the provided chart, ***August*** *has the most sun, with the* ***yellow segment*** *(representing sunny days) being the* ***tallest in that month's column****.*

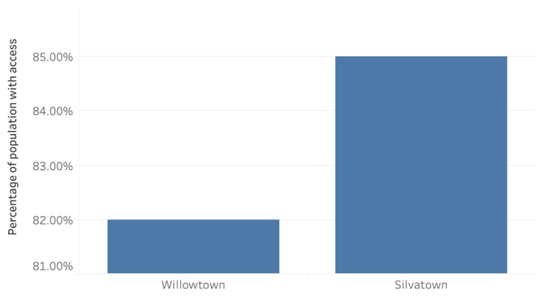


Deceptive Visualization Design Choices

Access to safe drinking water in Willowtown and Silvatown, as of 2010.

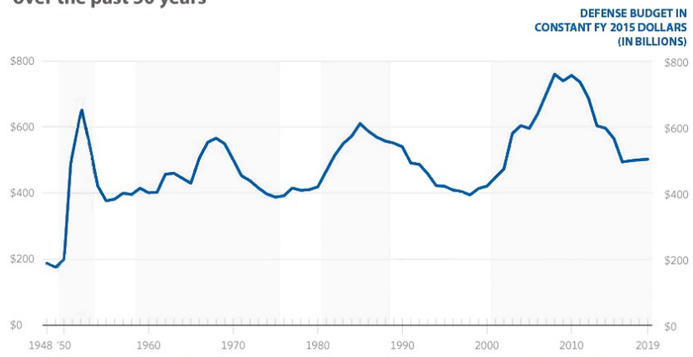


Access to safe drinking water in Willowtown and Silvatown, as of 2010.



Title-Chart Misalignment

Defense budget on a steady decrease as a percentage of GDP over the past 50 years



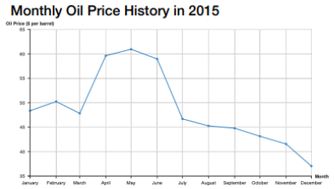
IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS VOL. 23, NO. 1, JANUARY 2017

VLAT

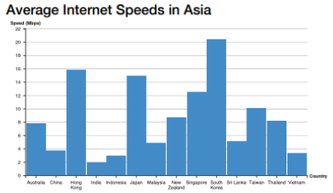
VLAT: Development of a Visualization Literacy Assessment Test

Sukwon Lee, Sung-Hee Kim, and Bum Chul Kwon, *Member, IEEE*

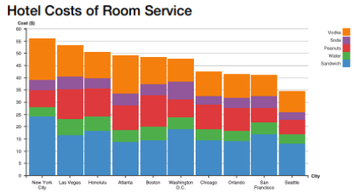
- Visualization literacy assessment test for humans
  - 12 data visualizations and 53 multiple-choice test items



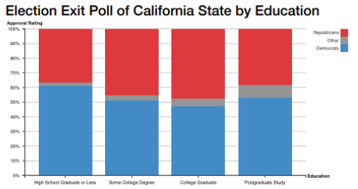
(a) Line Chart



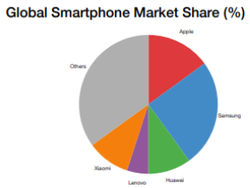
(b) Bar Chart



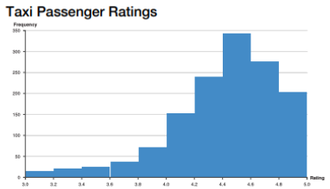
(c) Stacked Bar Chart



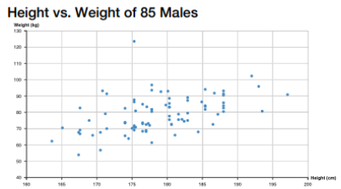
(d) 100% Stacked Bar Chart



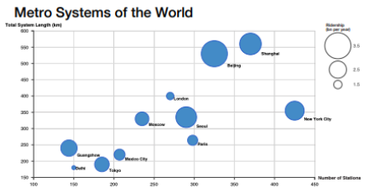
(e) Pie Chart



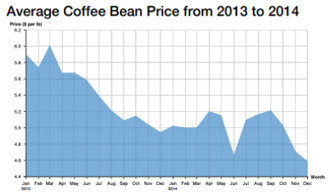
(f) Histogram



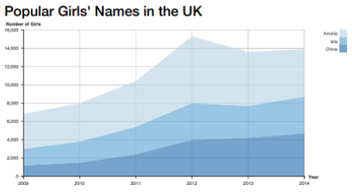
(g) Scatterplot



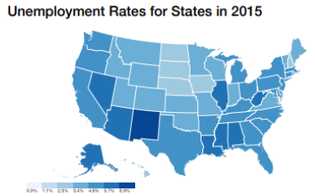
(h) Bubble Chart



(i) Area Chart

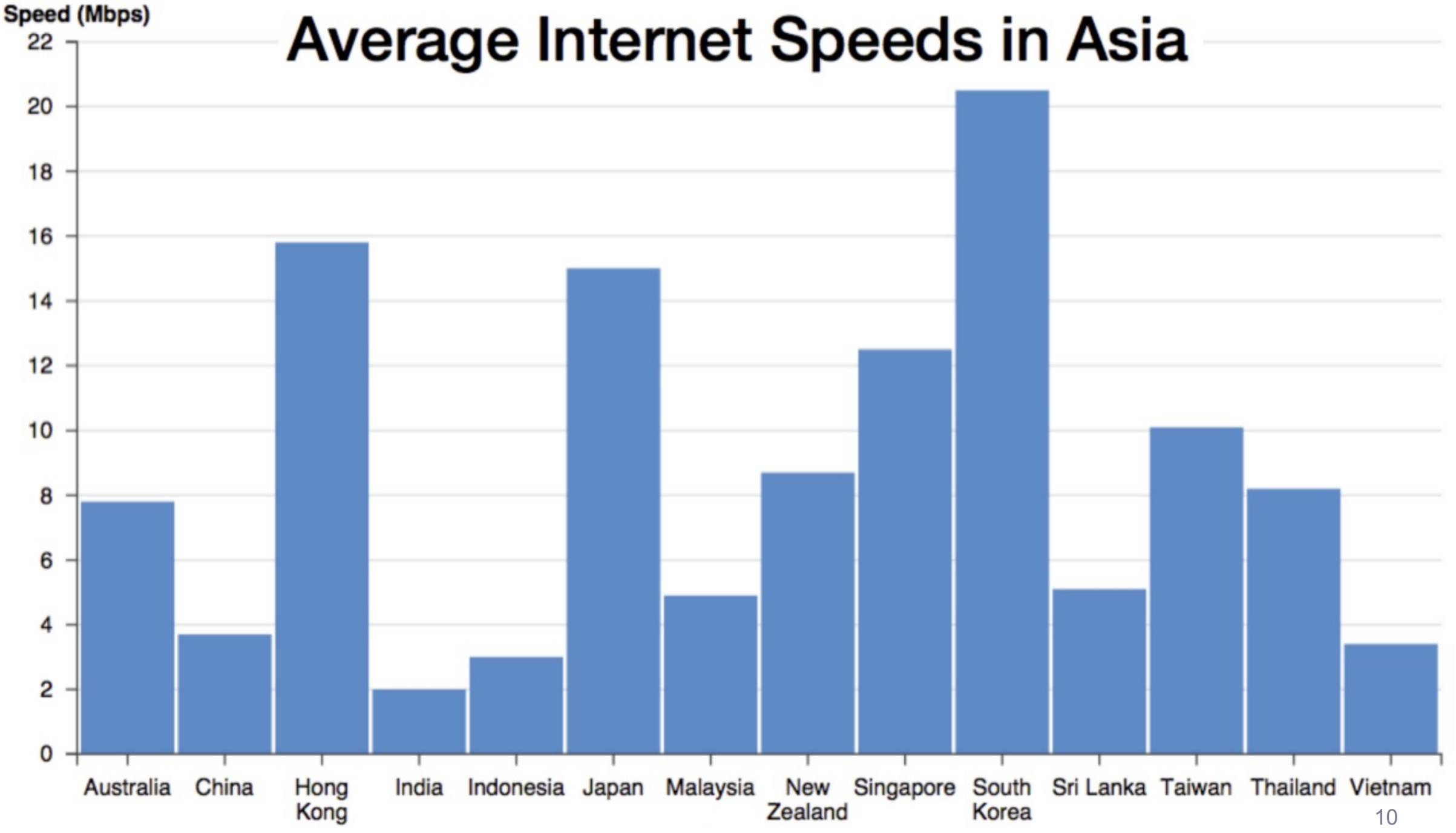


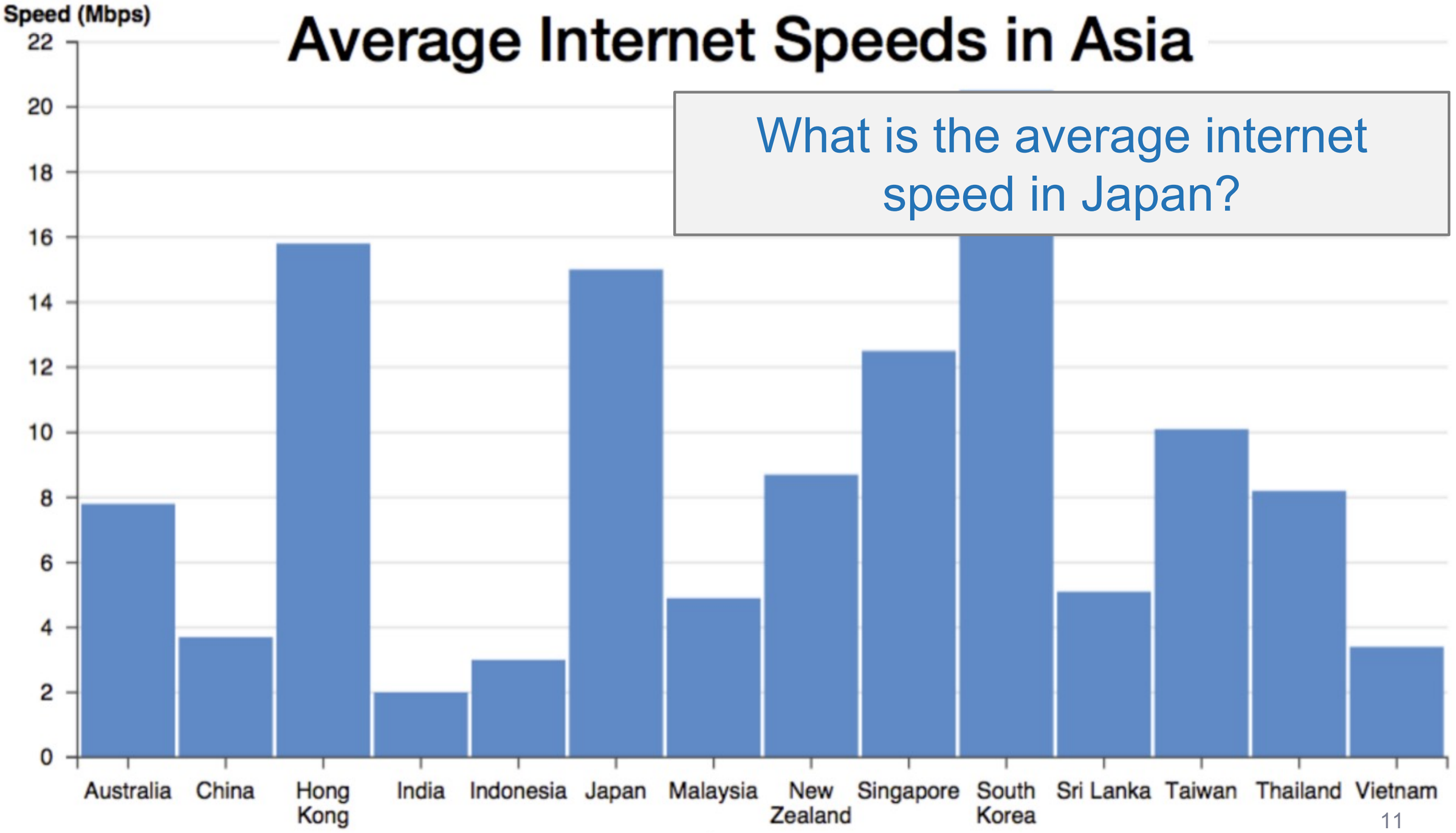
(j) Stacked Area Chart



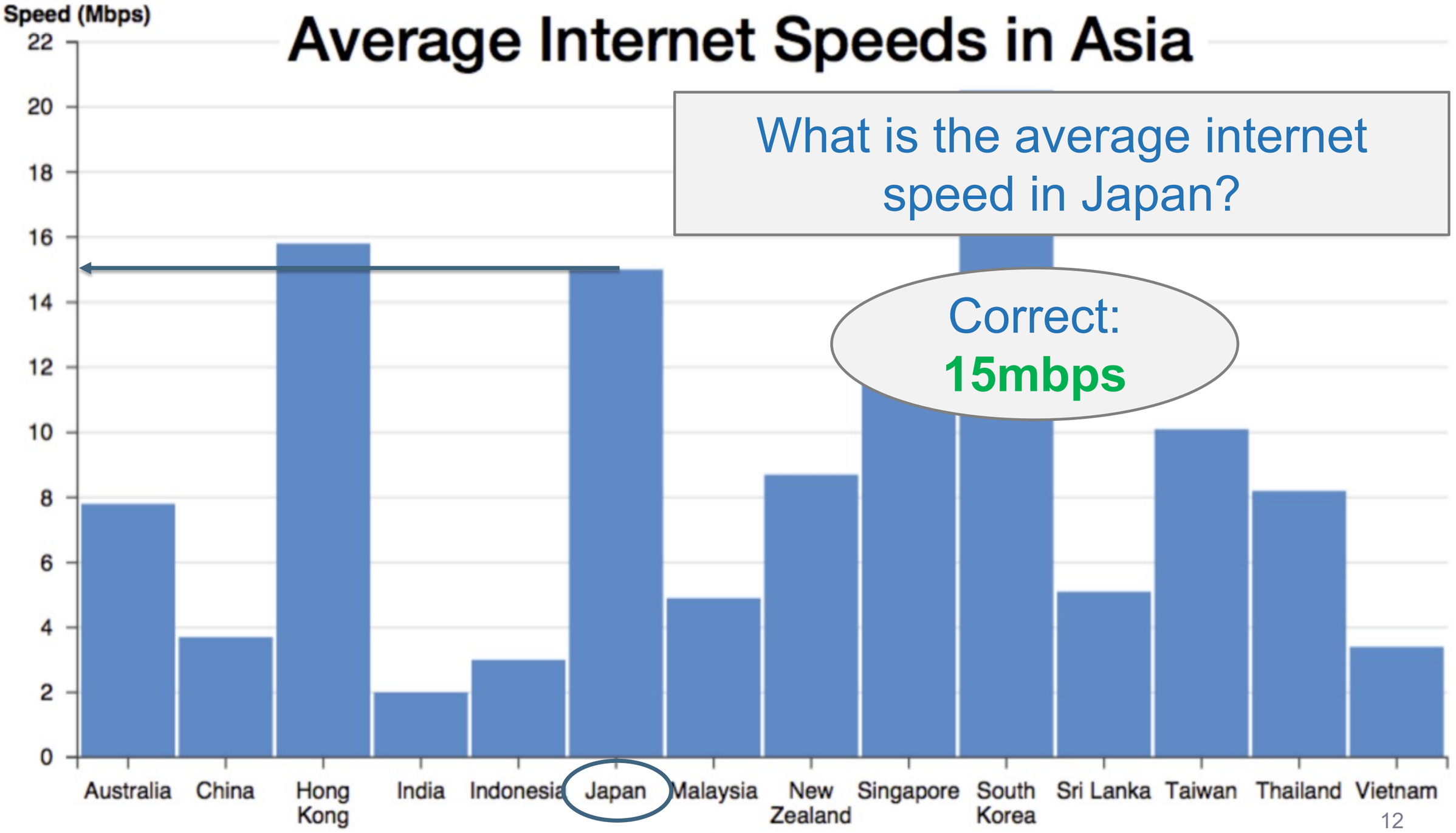
(k) Choropleth Map

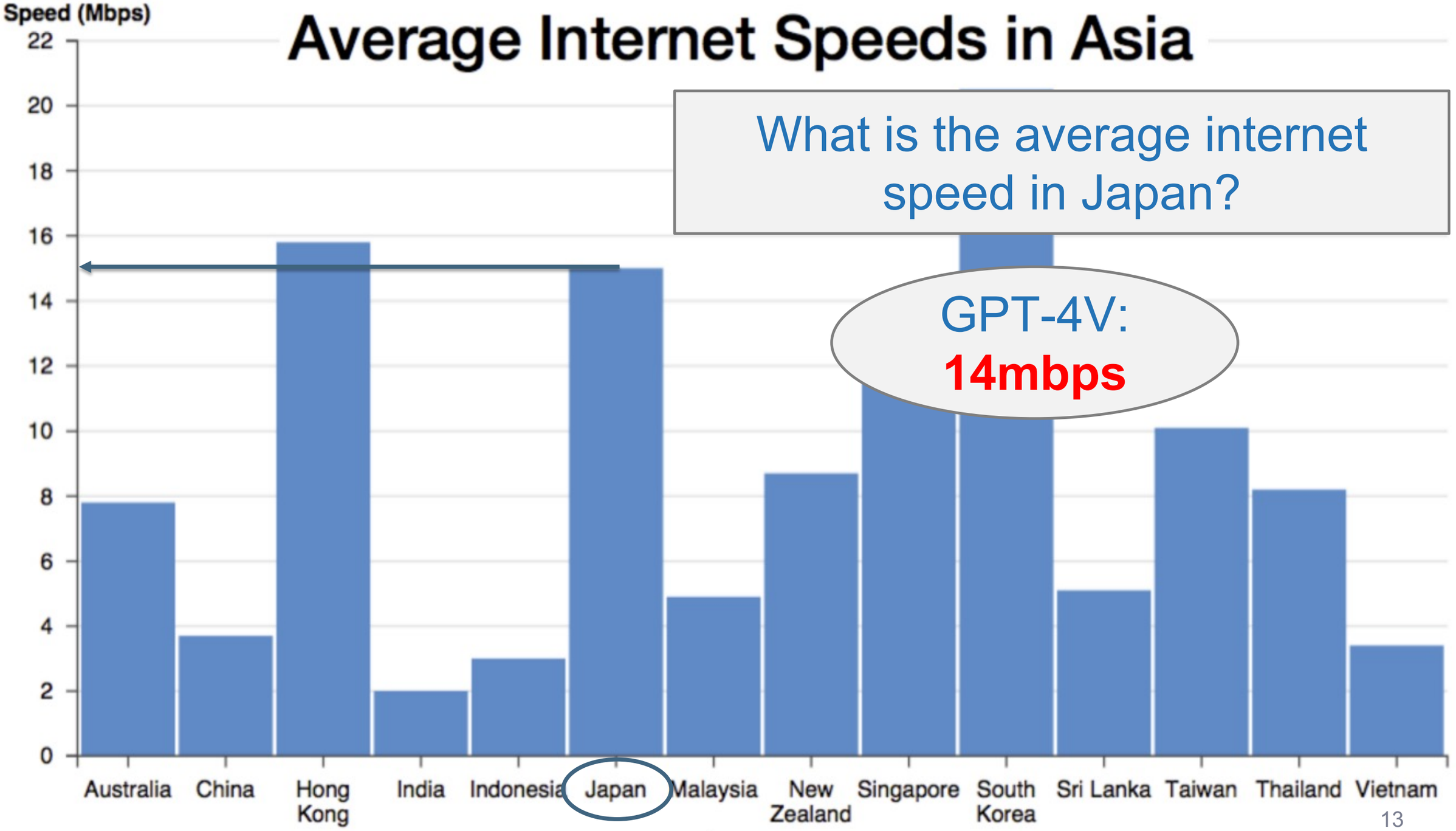
(l) Treemap

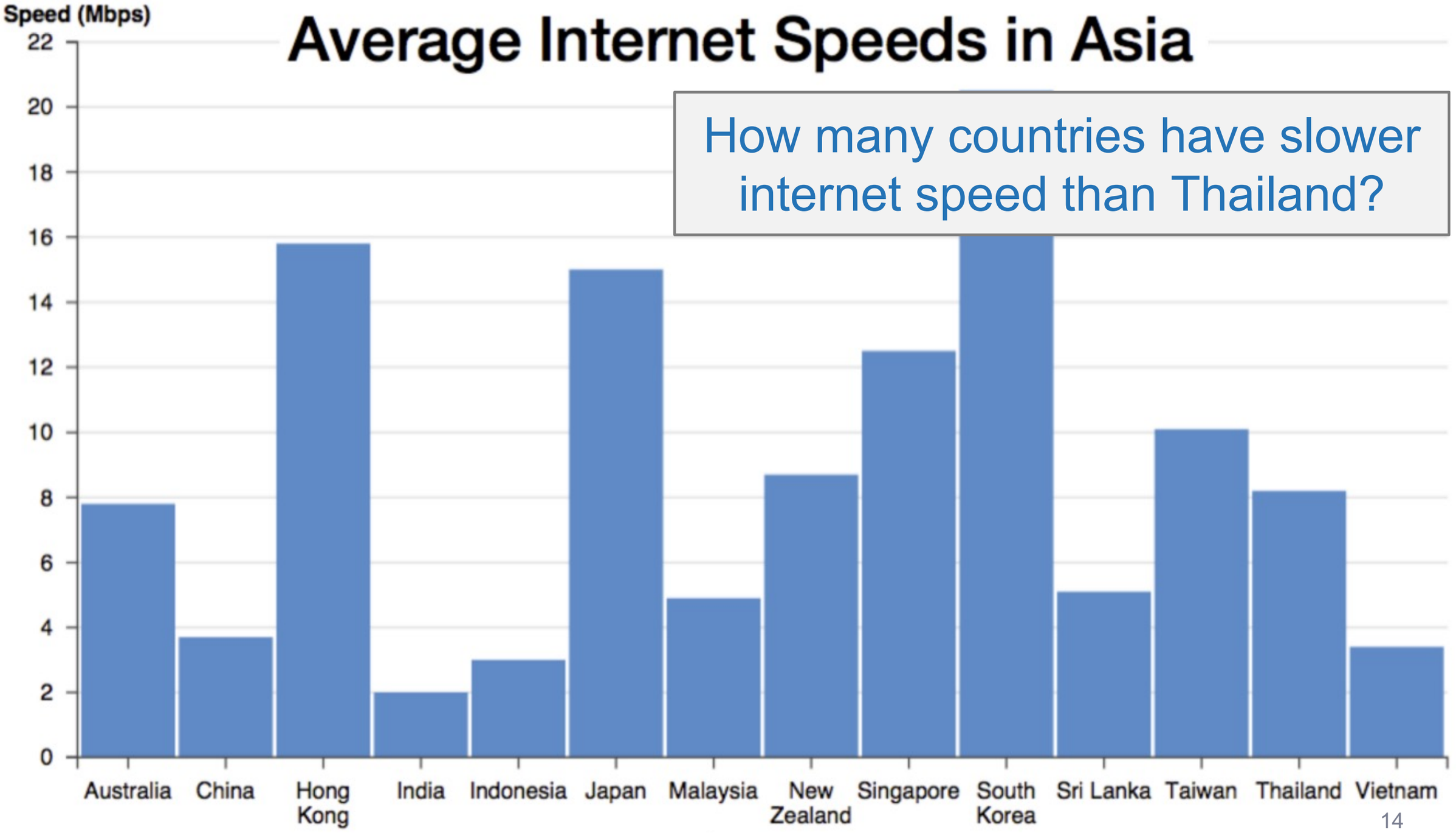

Average Internet Speeds in Asia
Speed (Mbps)
22
20
18
16
14
12
10
8
6
4
2
0
Australia
China
Hong Kong
India
Indonesia
Japan
Malaysia
New Zealand
Singapore
South Korea
Sri Lanka
Taiwan
Thailand
Vietnam



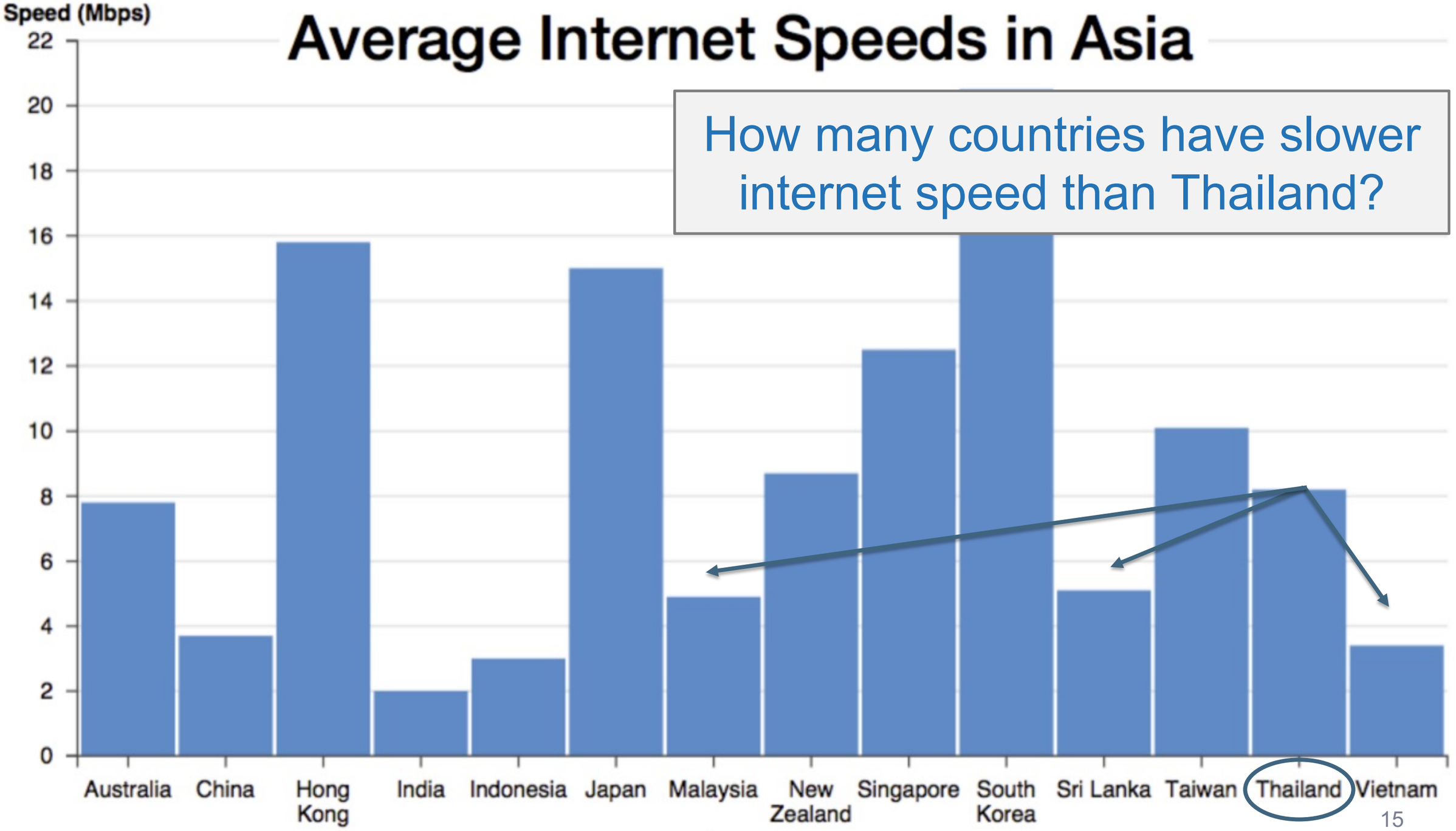
What is the average internet speed in Japan?

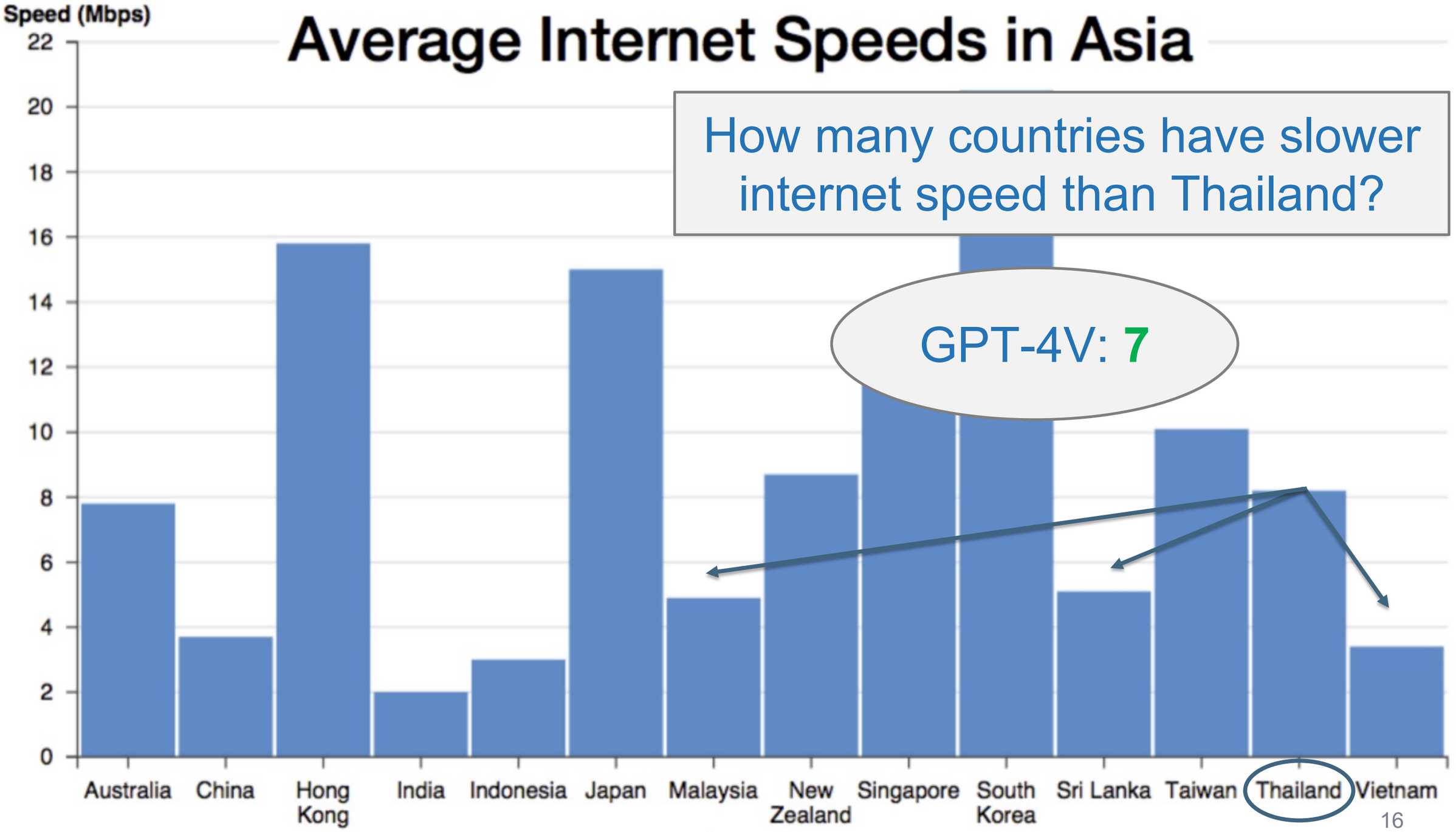

Average Internet Speeds in Asia
Speed (Mbps)
22
20
18
16
14
12
10
8
6
4
2
0
Australia
China
Hong Kong
India
Indonesia
Japan
Malaysia
New Zealand
Singapore
South Korea
Sri Lanka
Taiwan
Thailand
Vietnam
What is the average internet speed in Japan?
Correct:
15mbps


Average Internet Speeds in Asia
Speed (Mbps)
22
20
18
16
14
12
10
8
6
4
2
0
Australia
China
Hong Kong
India
Indonesia
Japan
Malaysia
New Zealand
Singapore
South Korea
Sri Lanka
Taiwan
Thailand
Vietnam
What is the average internet speed in Japan?
GPT-4V:
14mbps



How many countries have slower internet speed than Thailand?


Average Internet Speeds in Asia
Speed (Mbps)
22
20
18
16
14
12
10
8
6
4
2
0
Australia
China
Hong Kong
India
Indonesia
Japan
Malaysia
New Zealand
Singapore
South Korea
Sri Lanka
Taiwan
Thailand
Vietnam
How many countries have slower internet speed than Thailand?


Average Internet Speeds in Asia
Speed (Mbps)
22
20
18
16
14
12
10
8
6
4
2
0
Australia
China
Hong Kong
India
Indonesia
Japan
Malaysia
New Zealand
Singapore
South Korea
Sri Lanka
Taiwan
Thailand
Vietnam
How many countries have slower internet speed than Thailand?
GPT-4V: 7

GPT-4V on the VLAT

- GPT-4V performs around the **16th percentile** of humans, and shows divergent performance on different tasks:

Task type	correct	omit	incorrect	% correct
Identify hierarchy	1	0	0	100.0
Find trends	4	0	1	80.0
Make comparisons	9	1	3	69.2
Find extremum	8	2	2	66.7
Find anomalies	1	0	1	50.0
Determine range	2	0	3	40.0
Find clusters	1	0	2	33.3
Retrieve value	3	6	4	23.1

Chart Question Answering (CQA)

- Original work from CHI 2020

CHI 2020 Paper CHI 2020, April 25–30, 2020, Honolulu, HI, USA

Answering Questions about Charts and Generating Visual Explanations

Dae Hyun Kim
Stanford University
Stanford, CA, USA
dhkim16@cs.stanford.edu

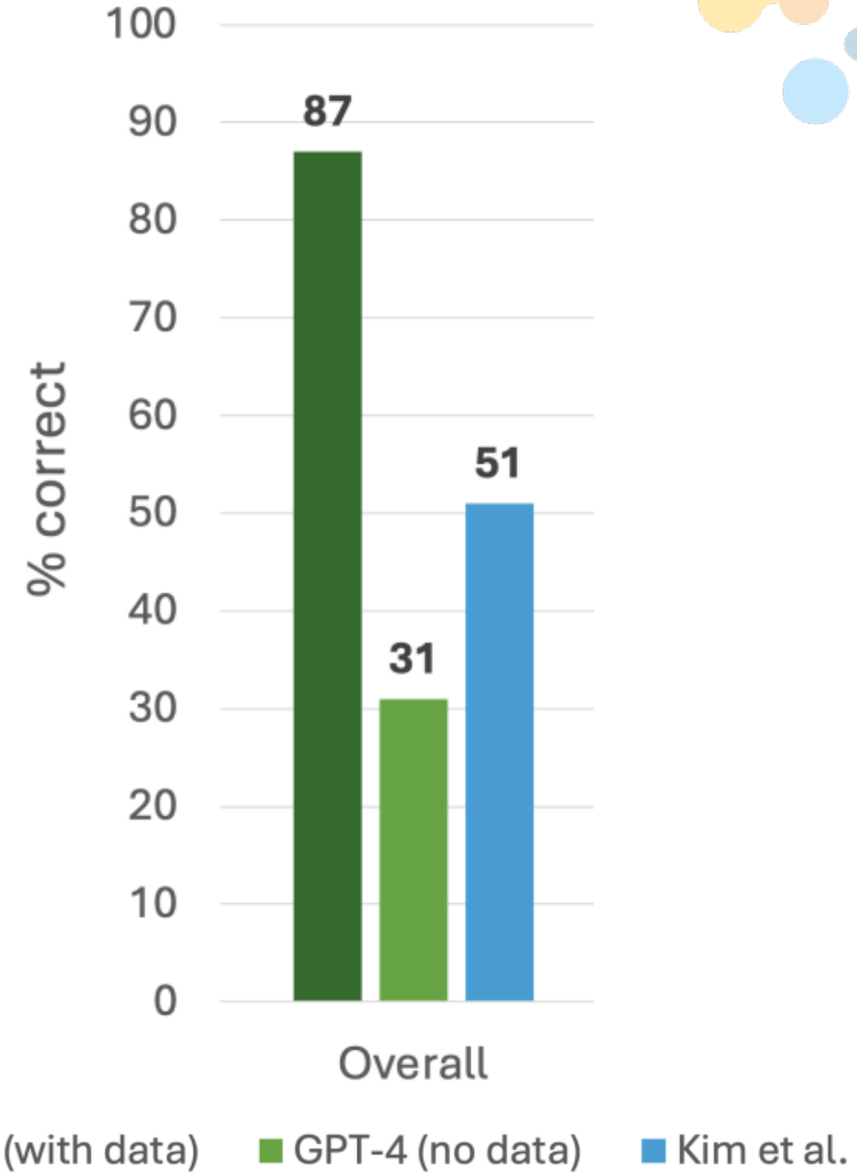
Enamul Hoque
York University
Toronto, ON, Canada
enamulh@yorku.ca

Maneesh Agrawala
Stanford University
Stanford, CA, USA
maneesh@cs.stanford.edu

- Benchmark dataset
  - 629 questions
  - 47 charts (mostly bar & some line)
- We compare GPT-4V's performance with underlying data vs. without
  - Paper's system (our baseline) has data extraction step before QA

CQA Results

- **GPT-4V with data** performs very well
  - Data extraction step from **Baseline system**
- **GPT-4V without data** performs quite poorly



CQA Results

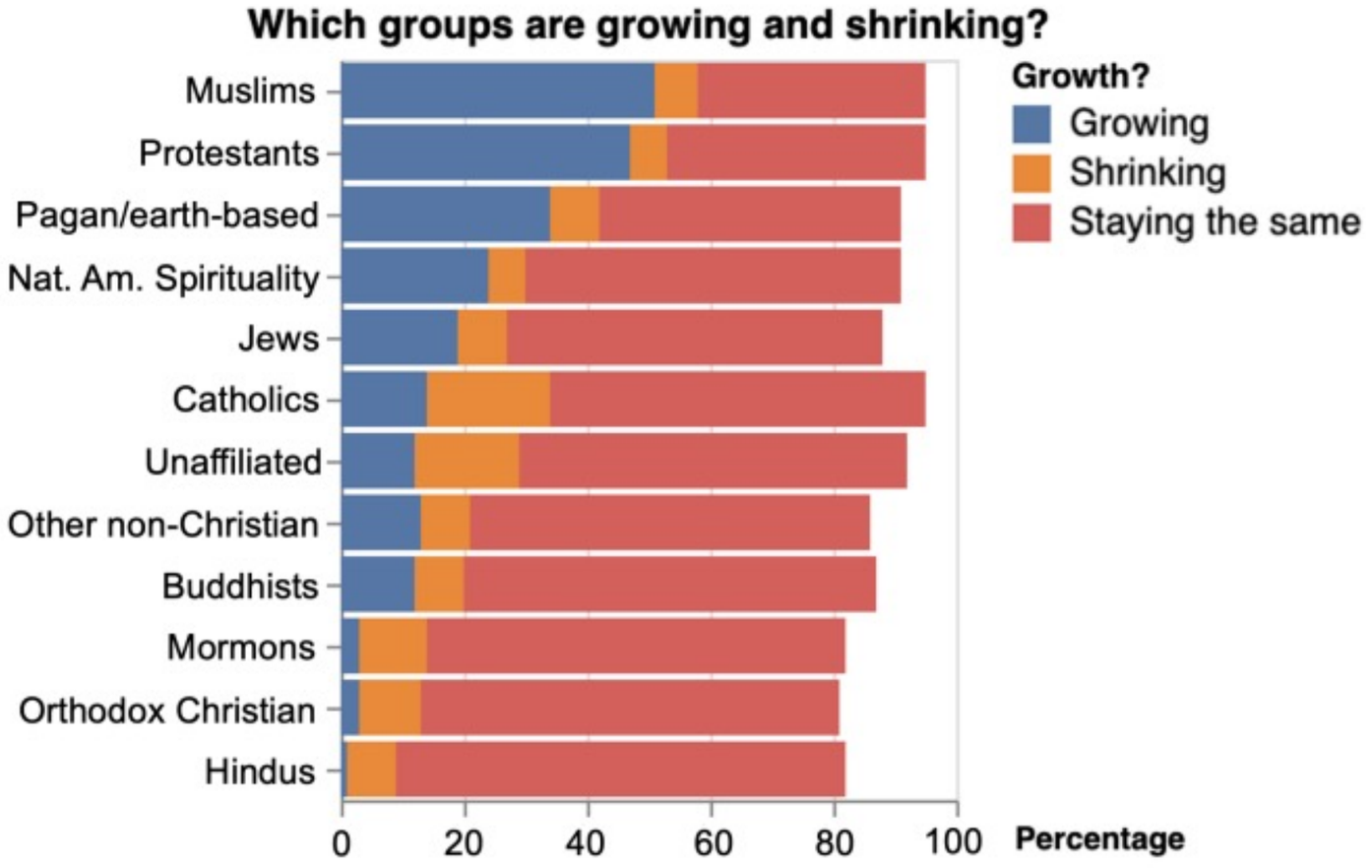
Task	# Questions	Accuracy w/ data	Accuracy w/o data
Compute derived value	125	96%	7%
Lookup	193	93%	23%
Find extrema	267	87%	52%
Make comparisons	25	84%	44%
Multiple	70	69%	37%

CQA Results

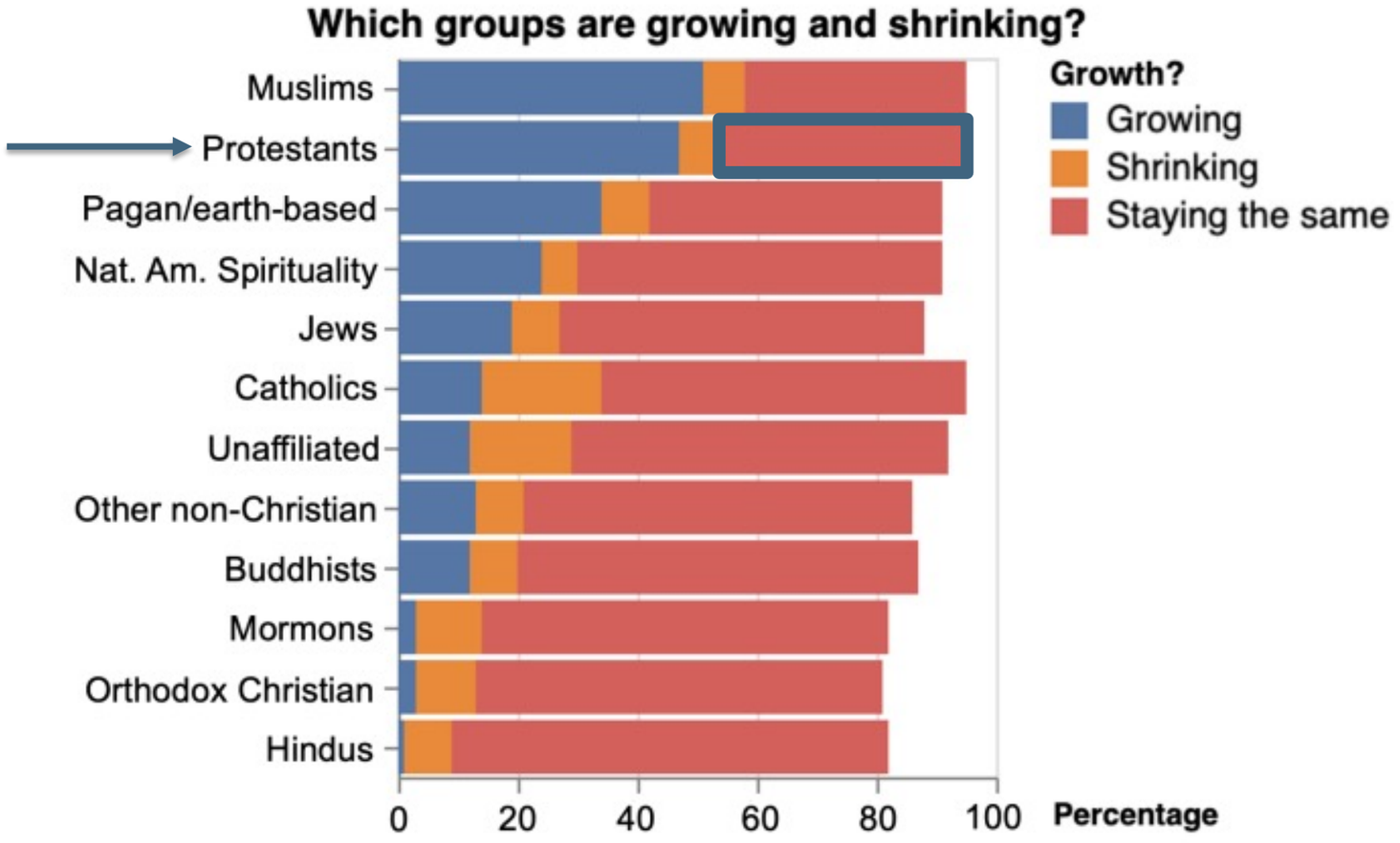
Task	# Questions	Accuracy w/ data	Accuracy w/o data
Compute derived value	125	96%	7%
Lookup	193	93%	23%
Find extrema	267	87%	52%
Make comparisons	25	84%	44%
Multiple	70	69%	37%

CQA Results

Task	# Questions	Accuracy w/ data	Accuracy w/o data
Compute derived value	125	96%	7%
Lookup	193	93%	23%
Find extrema	267	87%	52%
Make comparisons	25	84%	44%
Multiple	70	69%	37%

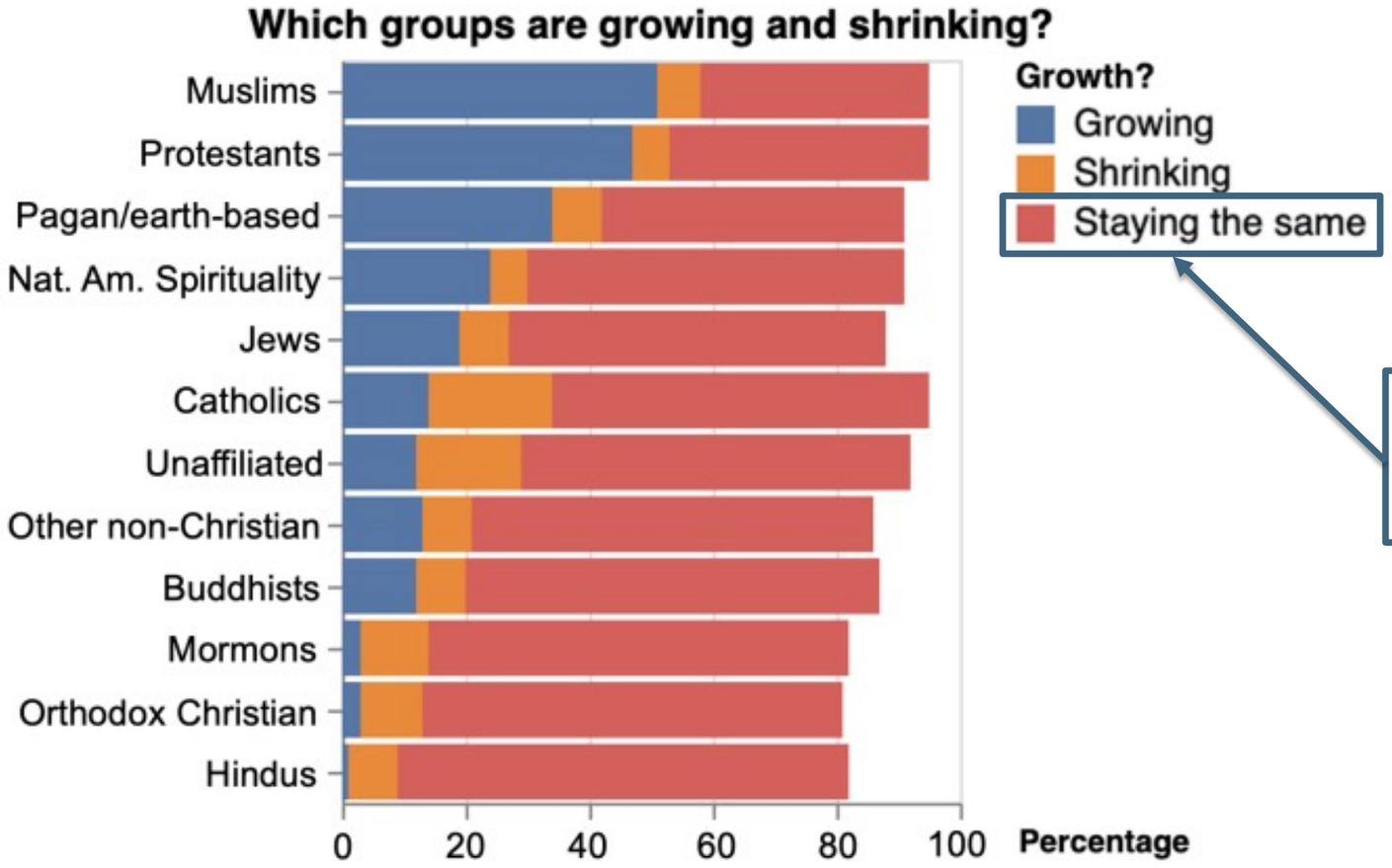


CQA Q85:
What is the percentage of red Protestants?



CQA Q85:
What is the percentage of red Protestants?

Challenge: Color



CQA Q85:
What is the percentage of red Protestants?

*The red portion [...] indicates the percentage that is **shrinking**.*

*According to the provided data, **6%** of Protestants are in the shrinking category.*

Deceptive Vis Designs

- Based on work from CHI 2015
  - Human-subjects study
- How effective are common vis distortions?
  - We study **6 distortions**

How Deceptive are Deceptive Visualizations?: An Empirical Analysis of Common Distortion Techniques

Anshul Vikram Pandey
School of Engineering,
New York University
anshul.pandey@nyu.edu

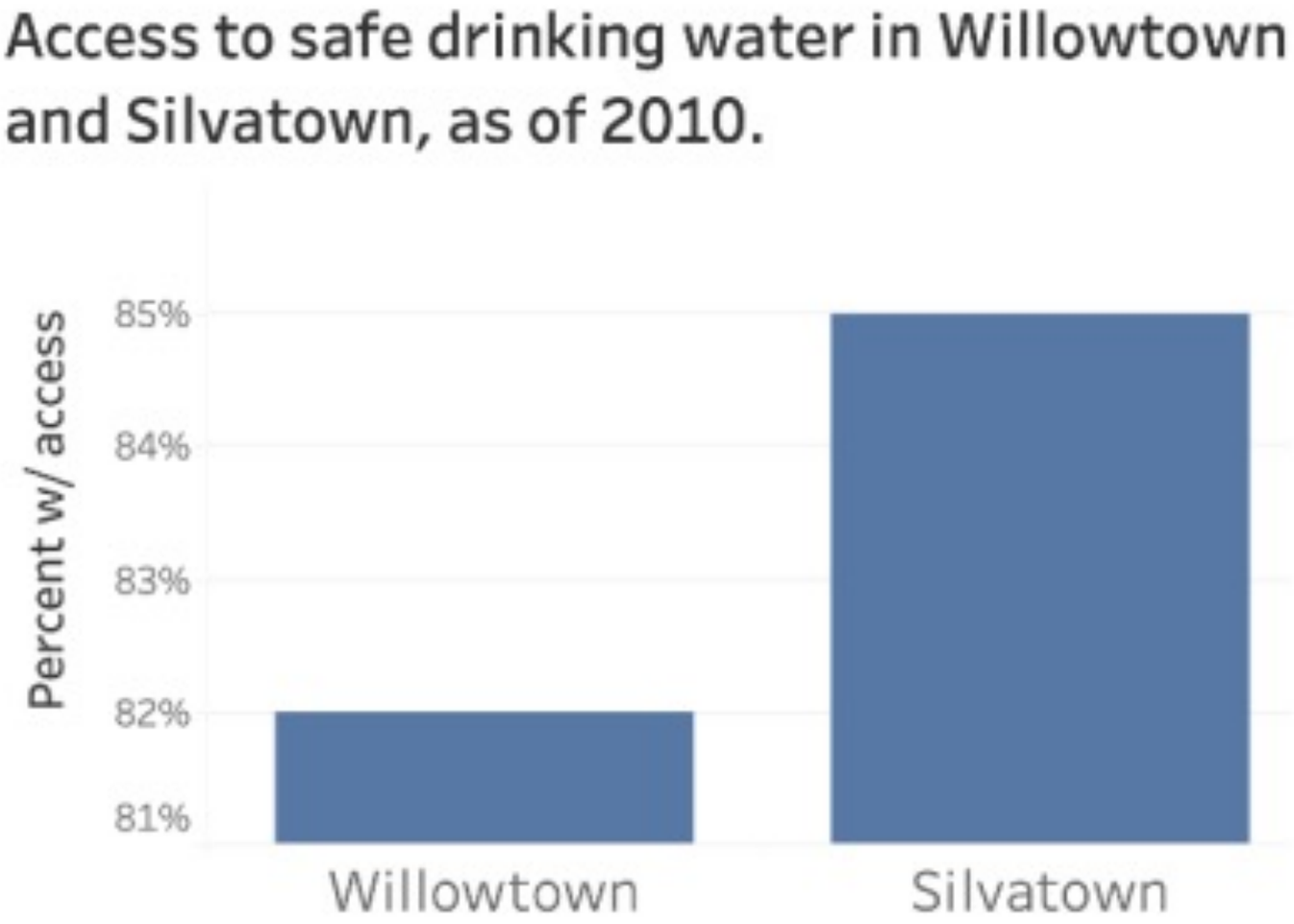
Katharina Rall
School of Law,
New York University
kr1326@nyu.edu

Margaret L. Satterthwaite
School of Law,
New York University
satterth@exchange.law.nyu.edu

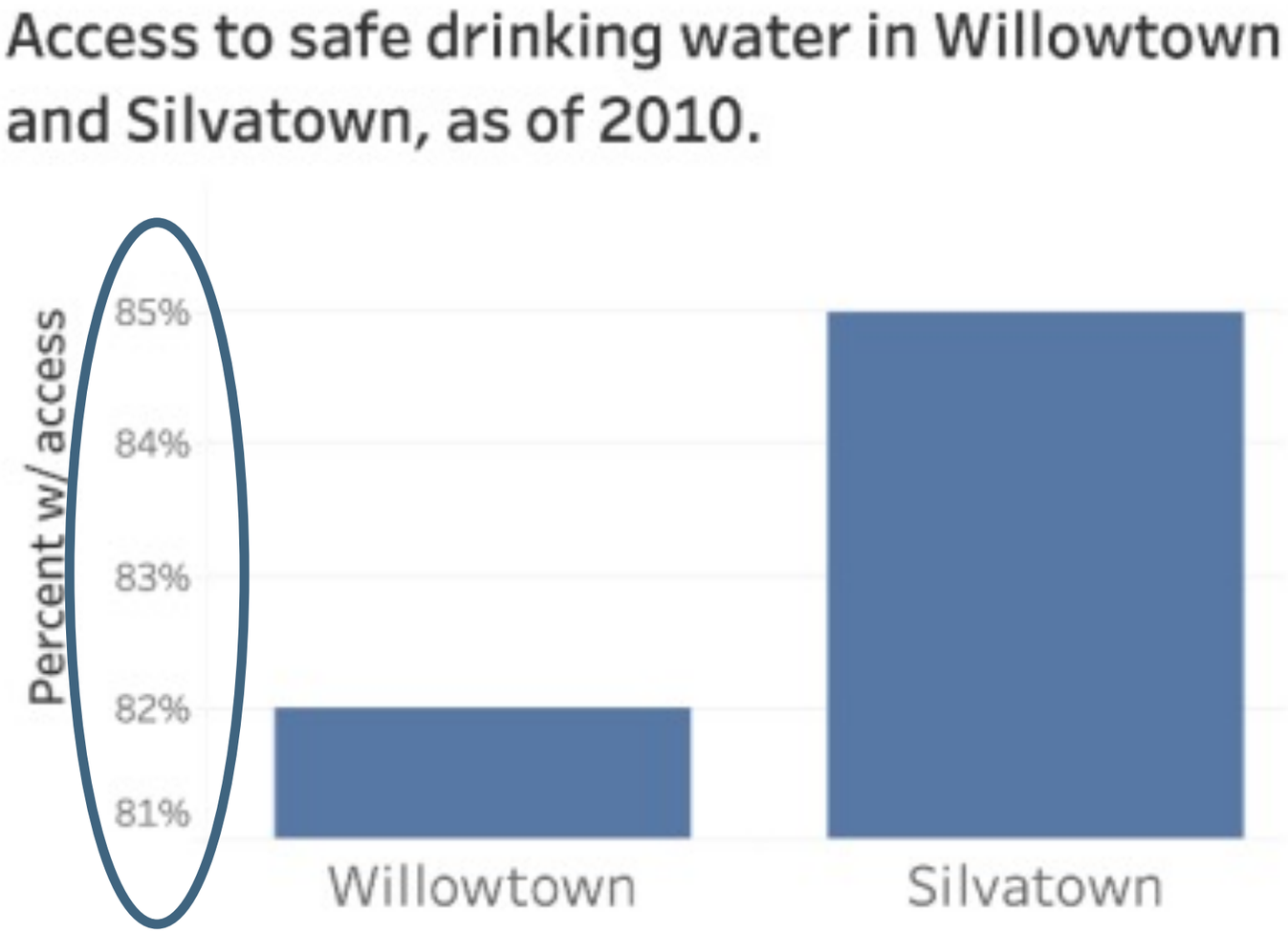
Oded Nov
School of Engineering,
New York University
onov@nyu.edu

Enrico Bertini
School of Engineering,
New York University
enrico.bertini@nyu.edu

Rate 1 to 5:
How much better is Silvatown?

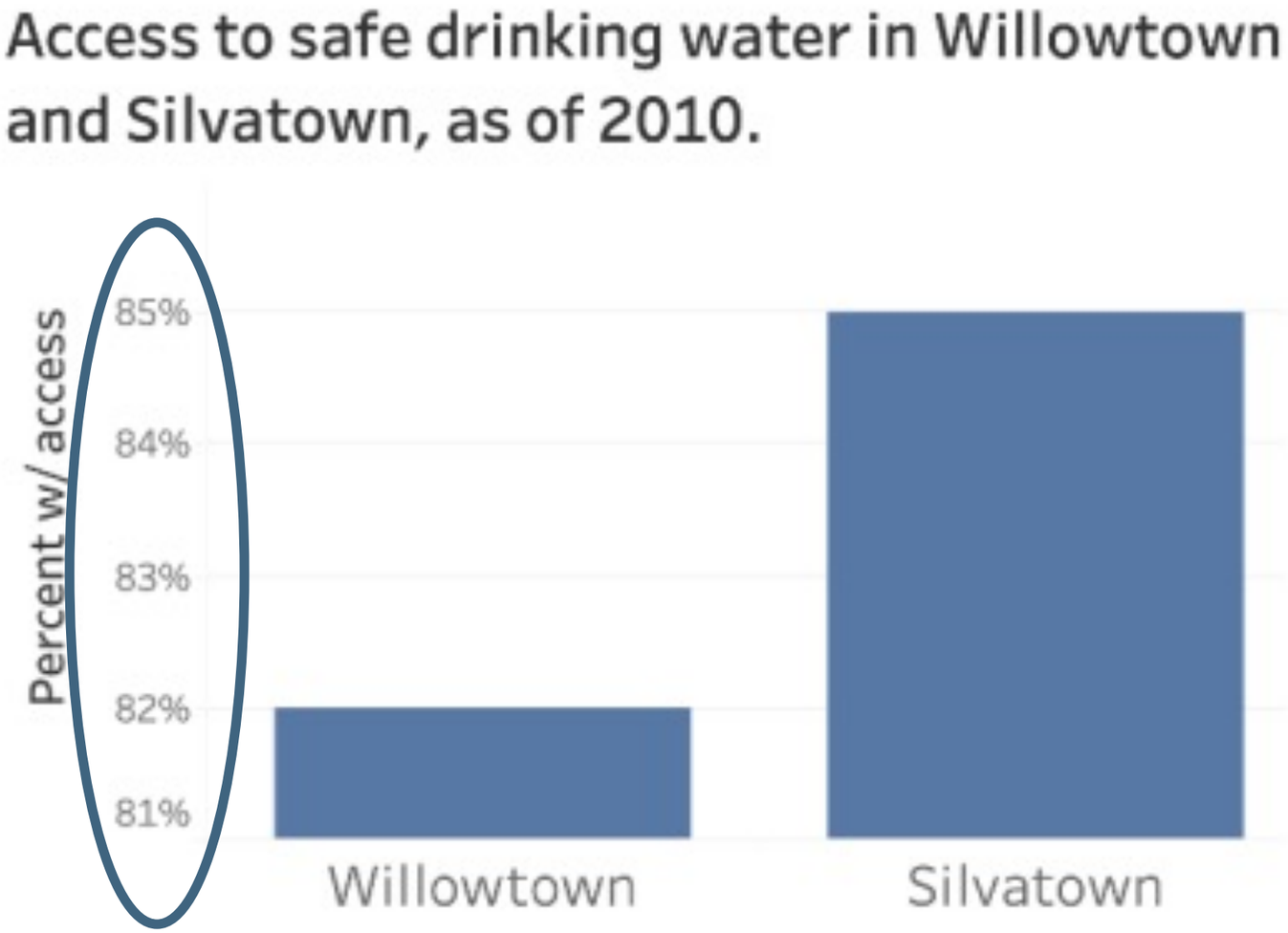


Rate 1 to 5:
How much better is Silvatown?



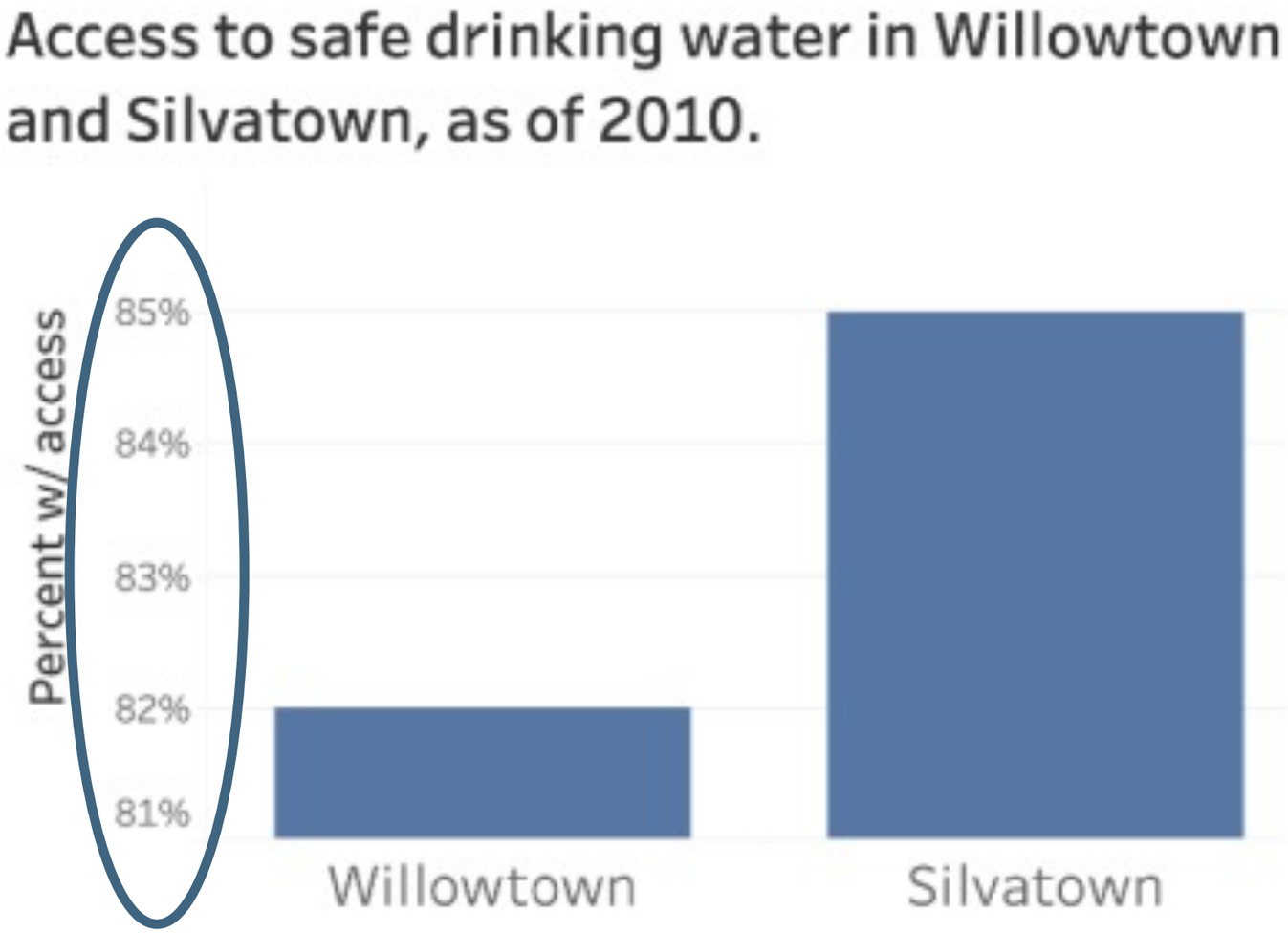
Rate 1 to 5:
How much better is Silvatown?

GPT-4V: **5**



Is there anything deceptive about this visualization?

GPT-4V: **Yes, truncated axis**



Title-Chart Misalignment

- Based on work from CHI 2018 and 2019
  - Human-subjects studies
- Varying degrees of title misalignment
  - 4 levels

Frames and Slants in Titles of Visualizations on Controversial Topics

Ha-Kyung Kong[1], **Zhicheng Liu**[2], **Karrie Karahalios**[1,2]
[1]University of Illinois at Urbana-Champaign, [2]Adobe Research
hkong6@illinois.edu, {leoli, karrie}@adobe.com

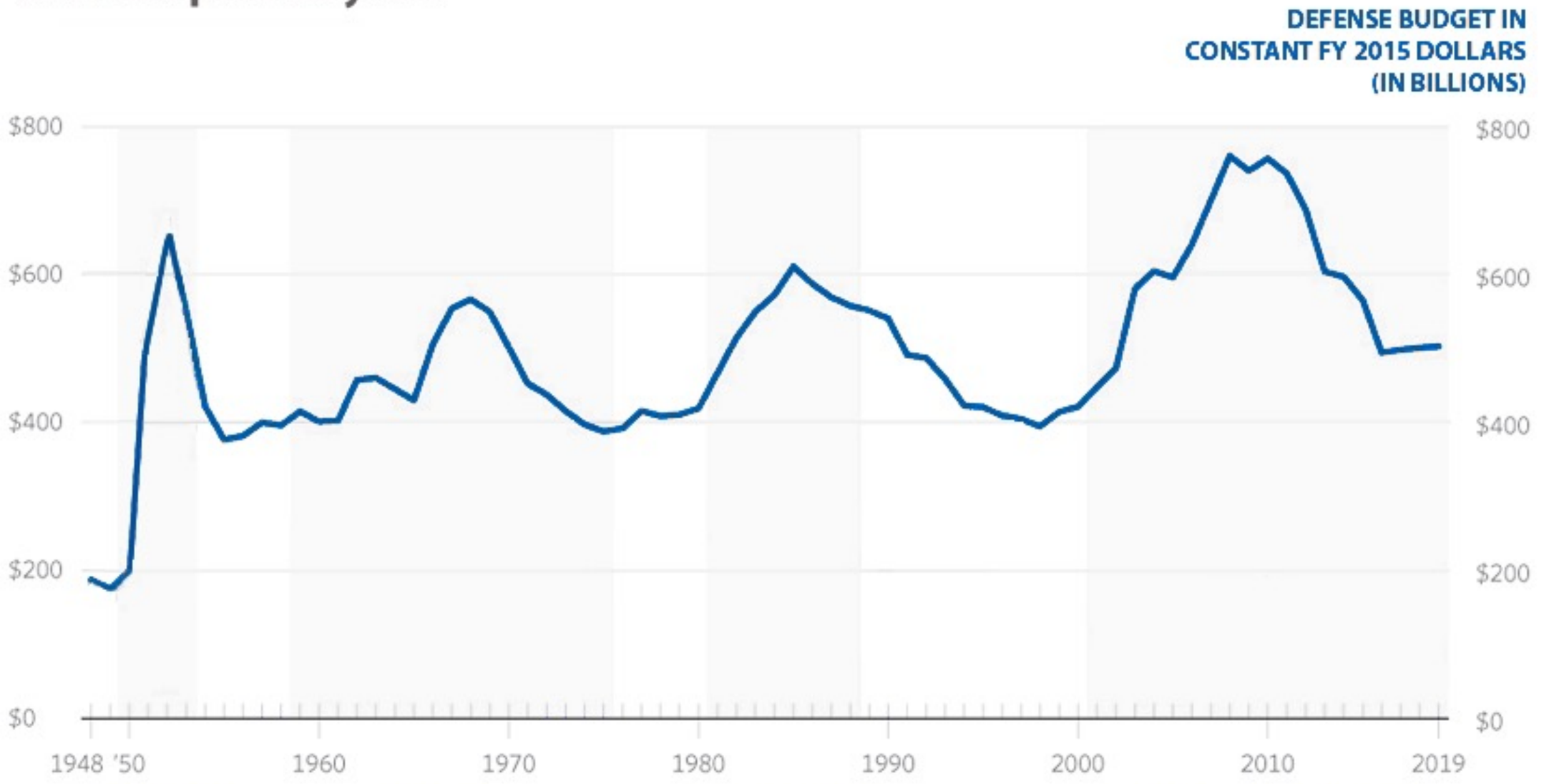
Trust and Recall of Information across Varying Degrees of Title-Visualization Misalignment

Ha-Kyung Kong
UIUC
Urbana, IL
hkong6@illinois.edu

Zhicheng Liu
Adobe Research
Seattle, WA
leoli@adobe.com

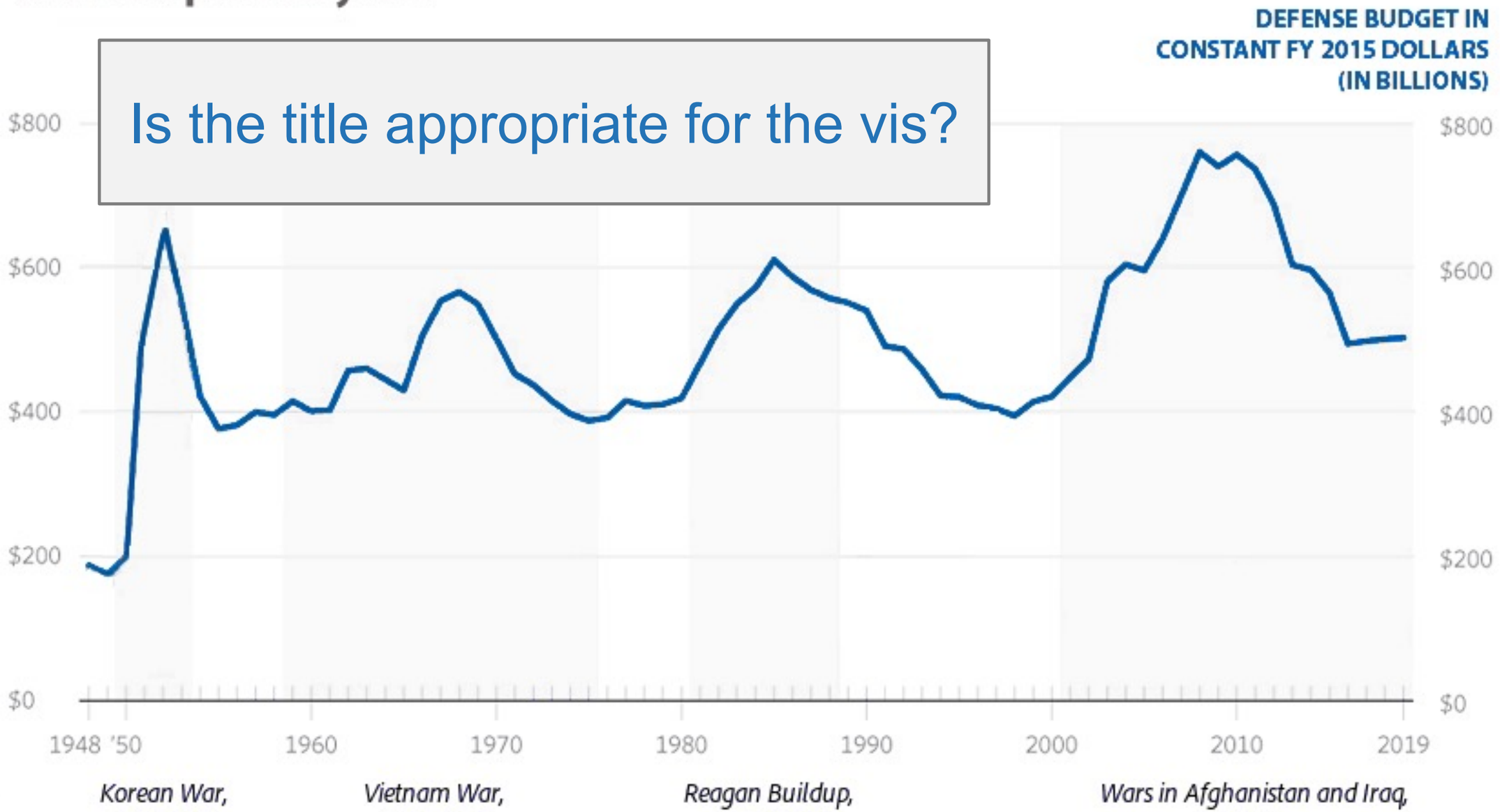
Karrie Karahalios
Adobe Research & UIUC
San Francisco, CA
karrie@adobe.com

Defense budget on a steady decrease as a percentage of GDP over the past 50 years

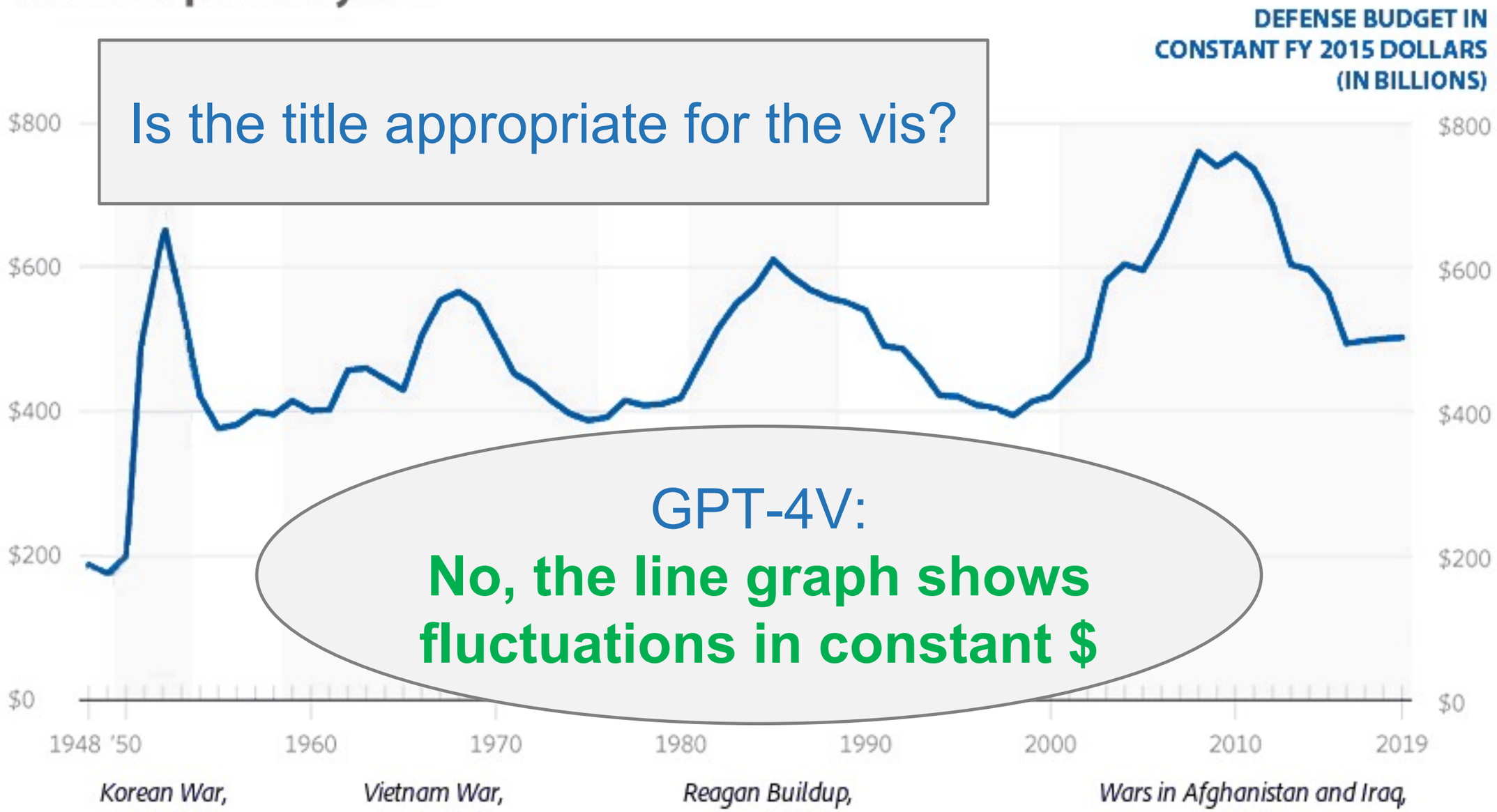


Defense budget on a steady decrease as a percentage of GDP over the past 50 years

Is the title appropriate for the vis?



Defense budget on a steady decrease as a percentage of GDP over the past 50 years

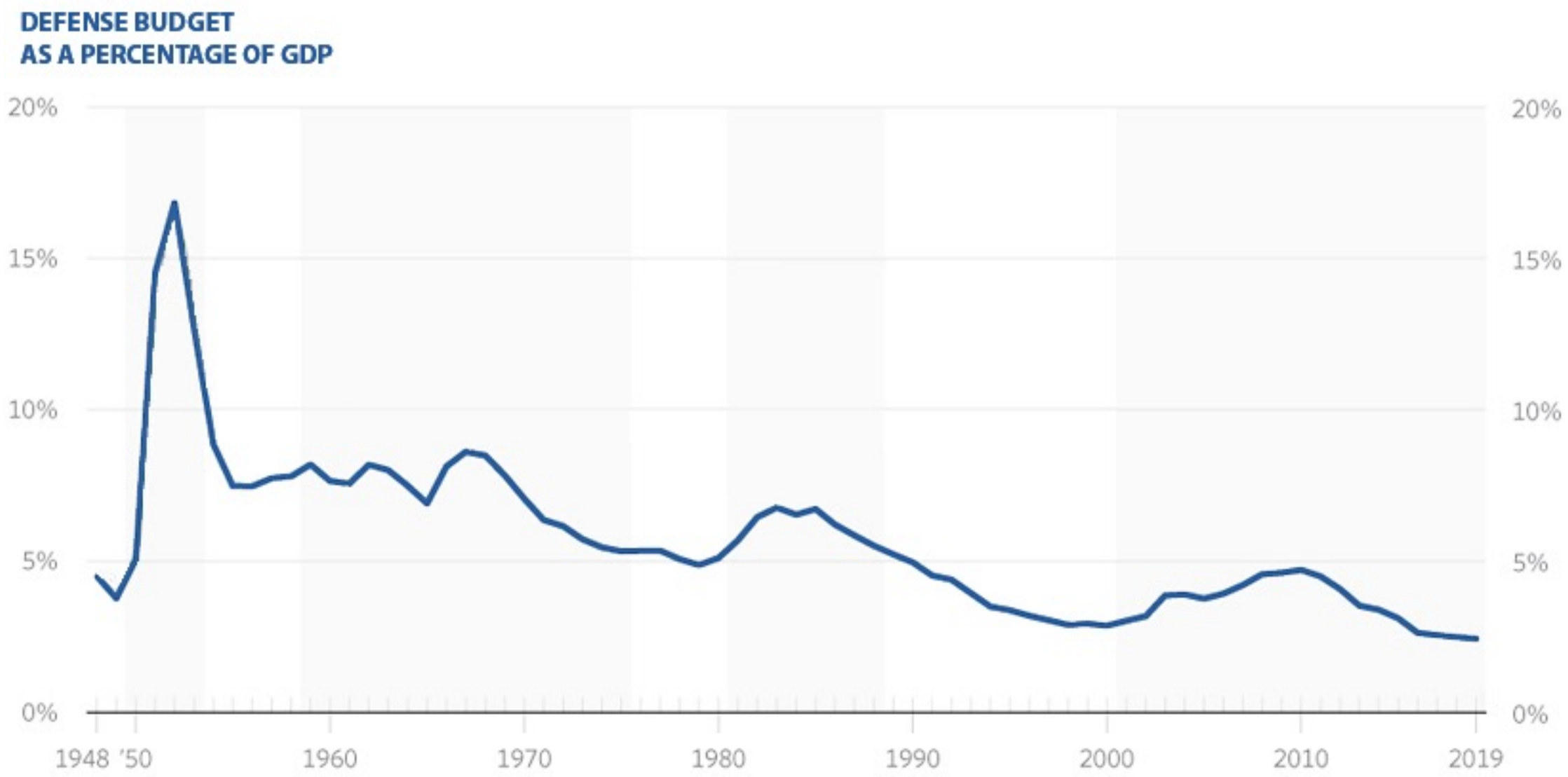


Is the title appropriate for the vis?

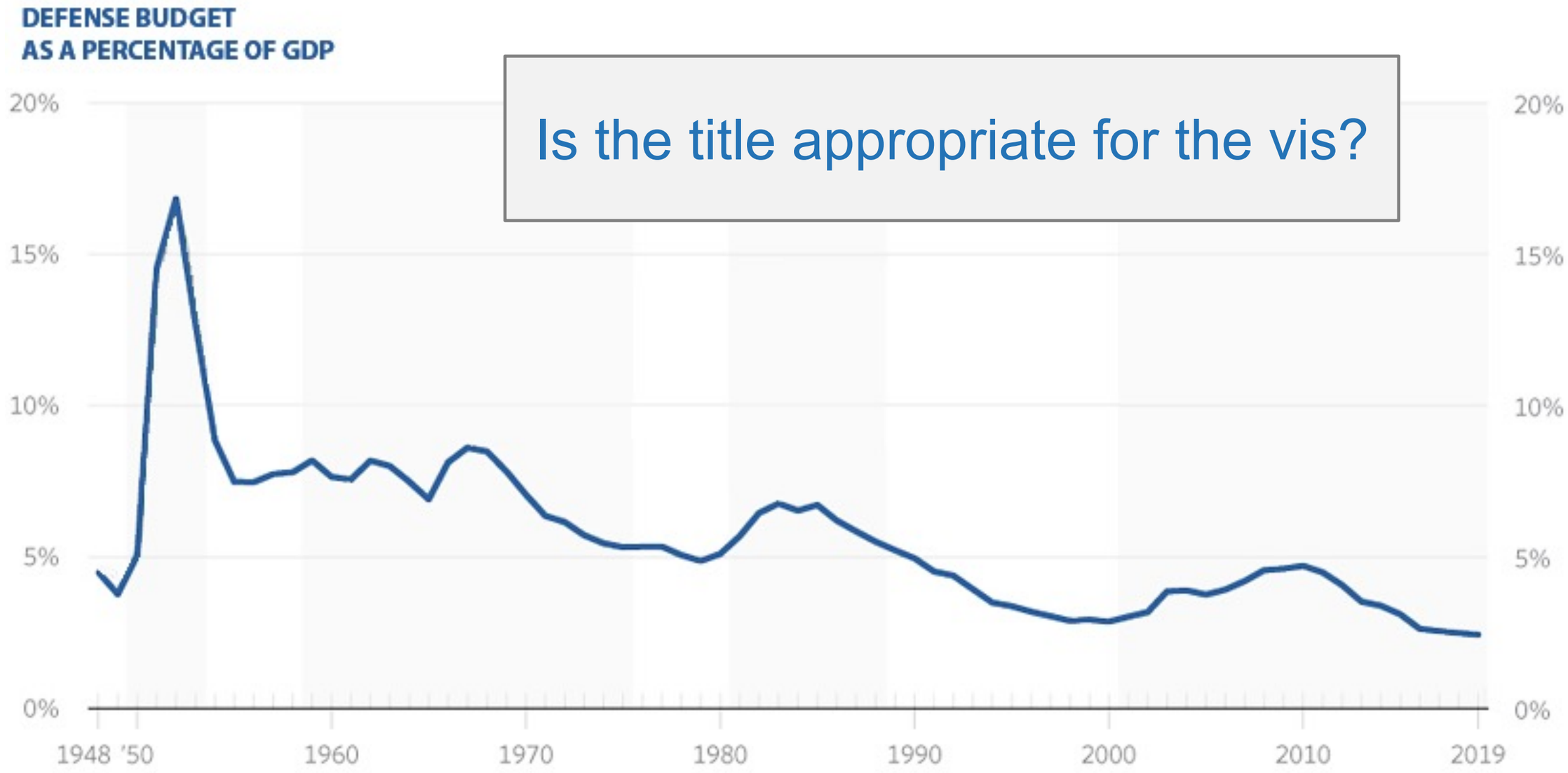
GPT-4V:

No, the line graph shows fluctuations in constant $

Defense budget on a steady decrease as a percentage of GDP over the past 50 years

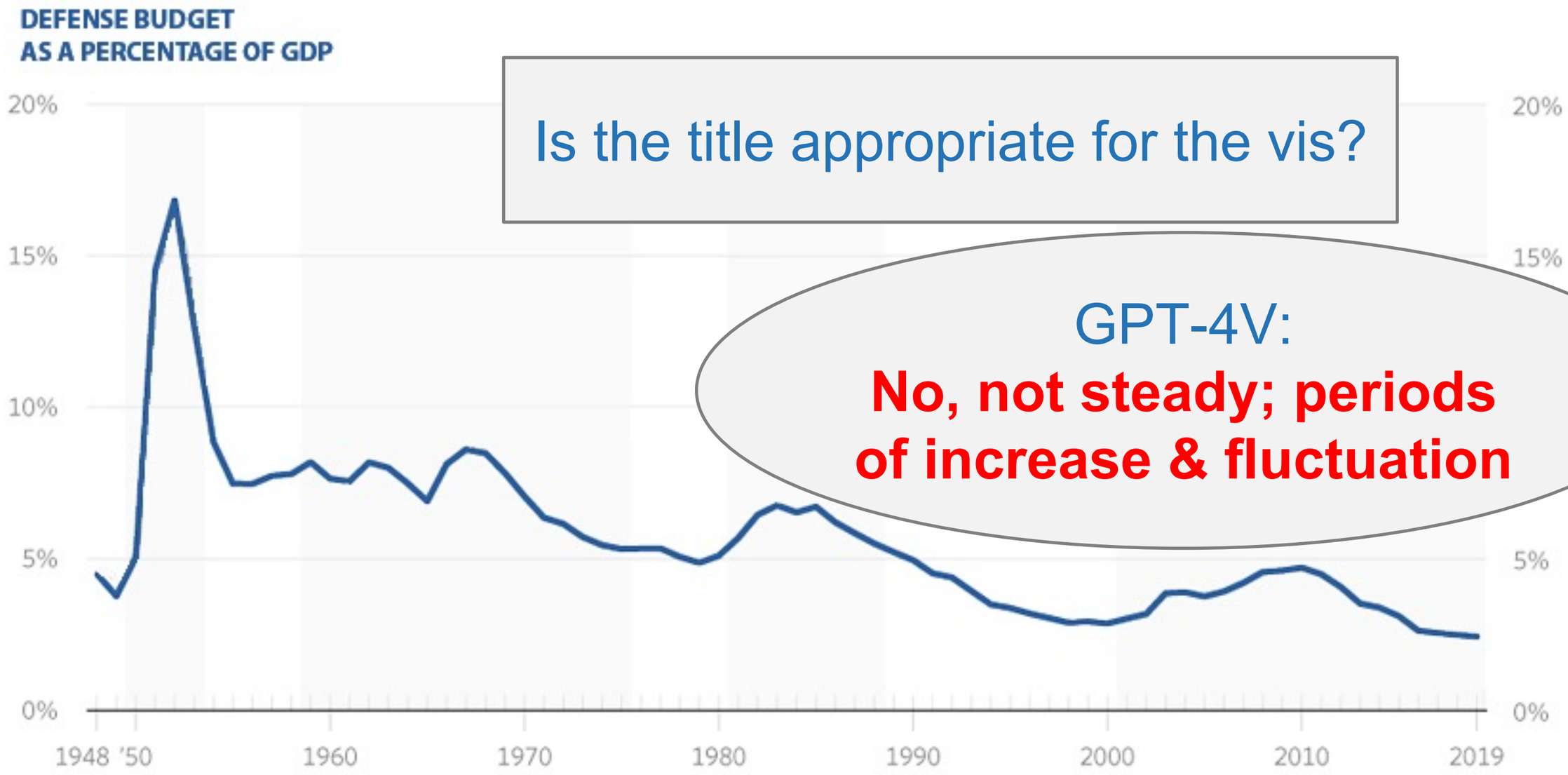


Defense budget on a steady decrease as a percentage of GDP over the past 50 years



Is the title appropriate for the vis?

Defense budget on a steady decrease as a percentage of GDP over the past 50 years



Is the title appropriate for the vis?

GPT-4V:

No, not steady; periods of increase & fluctuation

Recap: GPT-4V's Visualization Literacy

Strengths

- Recognizing trends
- Finding extrema
- Making comparisons

- Some knowledge of vis design best-practices
- Nuanced assessments of titles

Weaknesses

- Retrieving values (without data)
- Reading colors
- Hallucination

- Sometimes fooled by common deception techniques
- Focuses on nitpicky aspects of title wording

Reflections: What Next?

- Future evaluations
  - Why does GPT-4V behave like this? Hard to say
  - Evaluating open-source models may be helpful

- Work for vis folks (sooner or later)
  - Education aids & visualization design helpers
  - Browser extensions for consuming charts online
  - **But: When will these models be “ready”?**

Final Thoughts

An Empirical Evaluation of the GPT-4 Multimodal Language Model on Visualization Literacy Tasks

Alexander Bendeck and John Stasko

- Results reported in much more detail in paper
  - All code, stimuli, & prompts released as supplement
- Sensitivity Analysis
  - Prompt engineering & GPT's extensive knowledge
- Of course, LLMs are a moving target
  - Useful: GPT-4V "snapshot" & eval approach
- **Thank you!**

Supplemental material

Thank you!