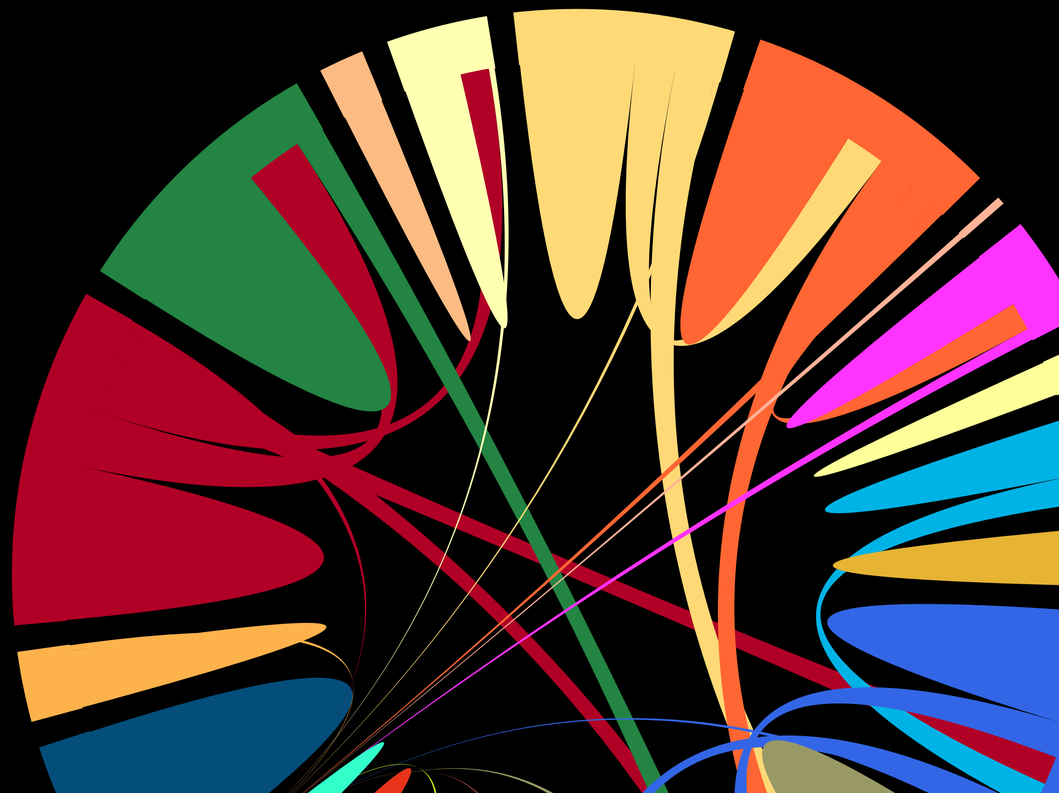
Information Visualization
INFO250

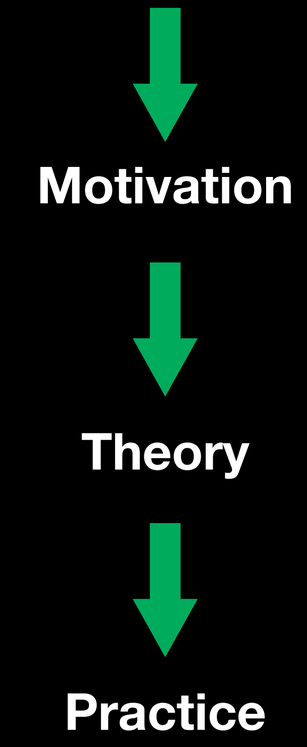
Chapter 3:
Chart Junk
Visualization Best Practices
Tufte's Principles
Practical Guidelines

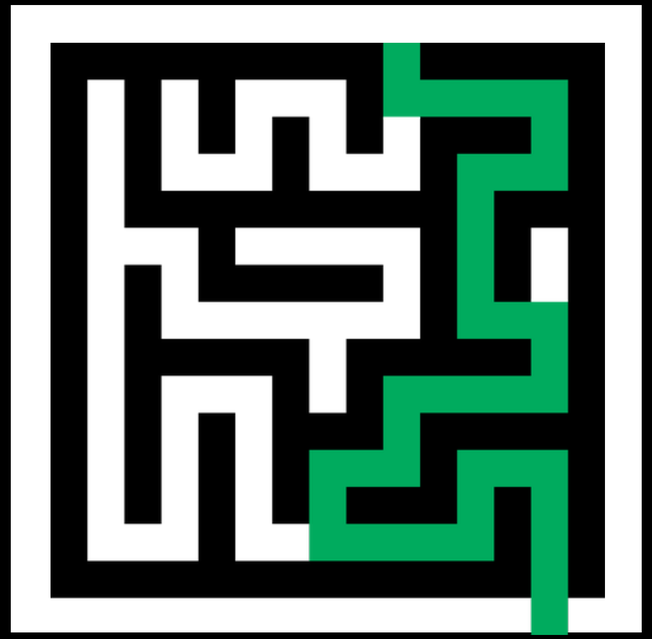


Luís Cruz - l.cruz@tudelft.nl

Outline

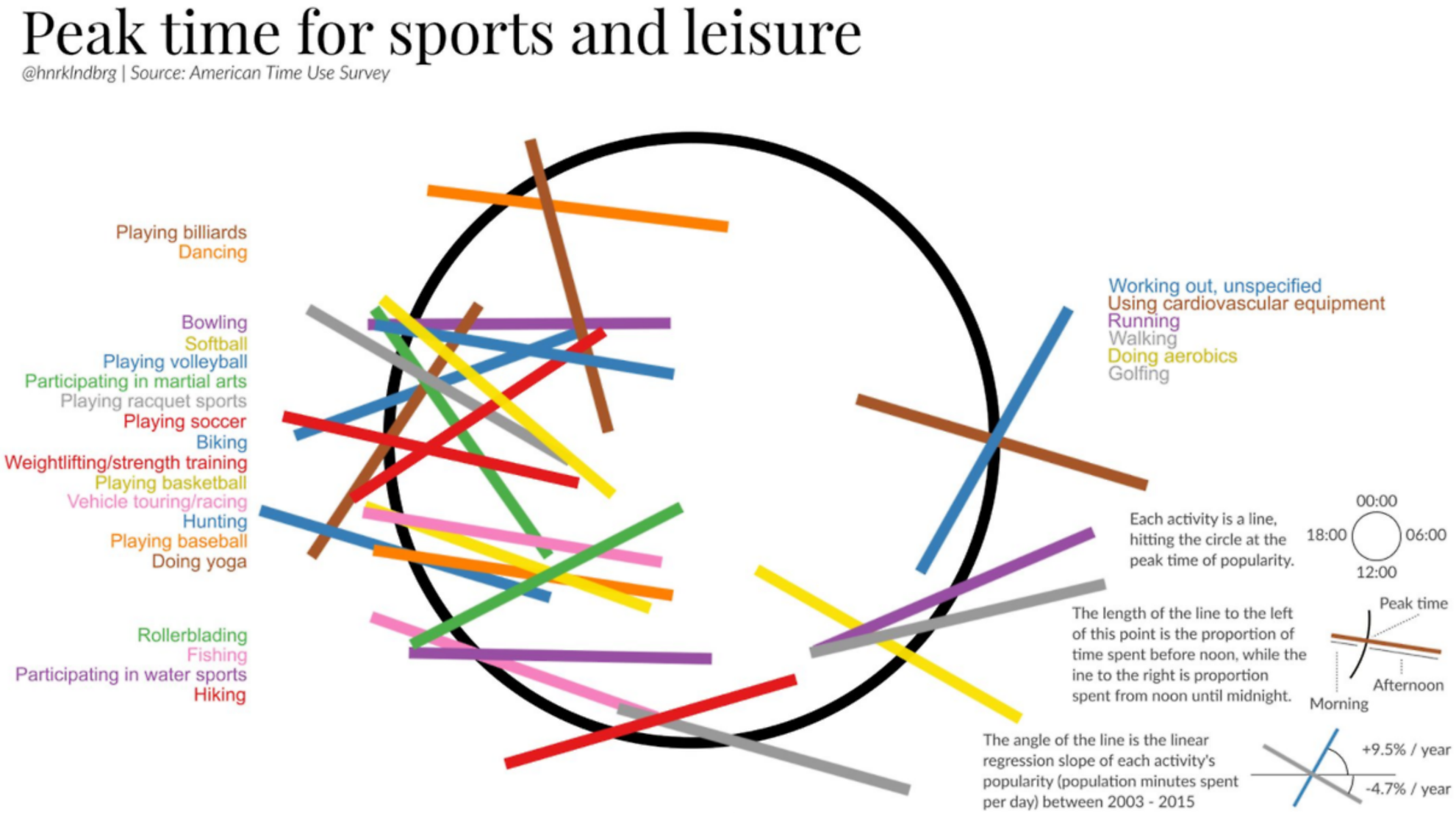
- The importance of good visualizations
- Chartjunk
- Graphical excellence
- Graphical integrity
- Real-world examples
- Practical guidelines





Examples of Visualizations

Examples of Visualizations

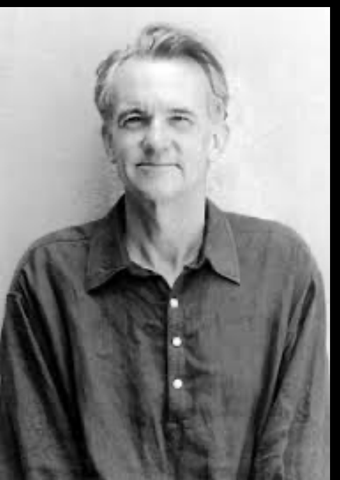


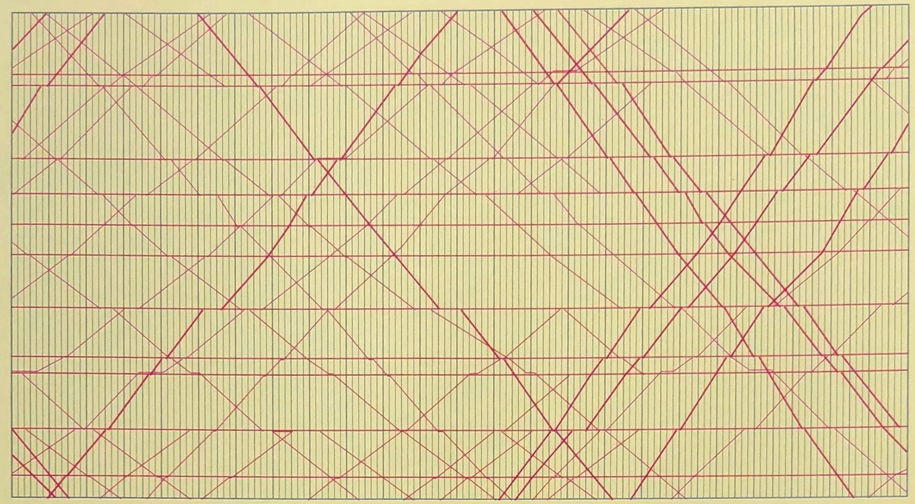
What is the theory behind a good visualization?

The Visual Display of Quantititative Information

Edward Tufte (1984)

- Coined the term *Chartjunk.*
- Graphical **Excellence**.
- Graphical **Integrity**.





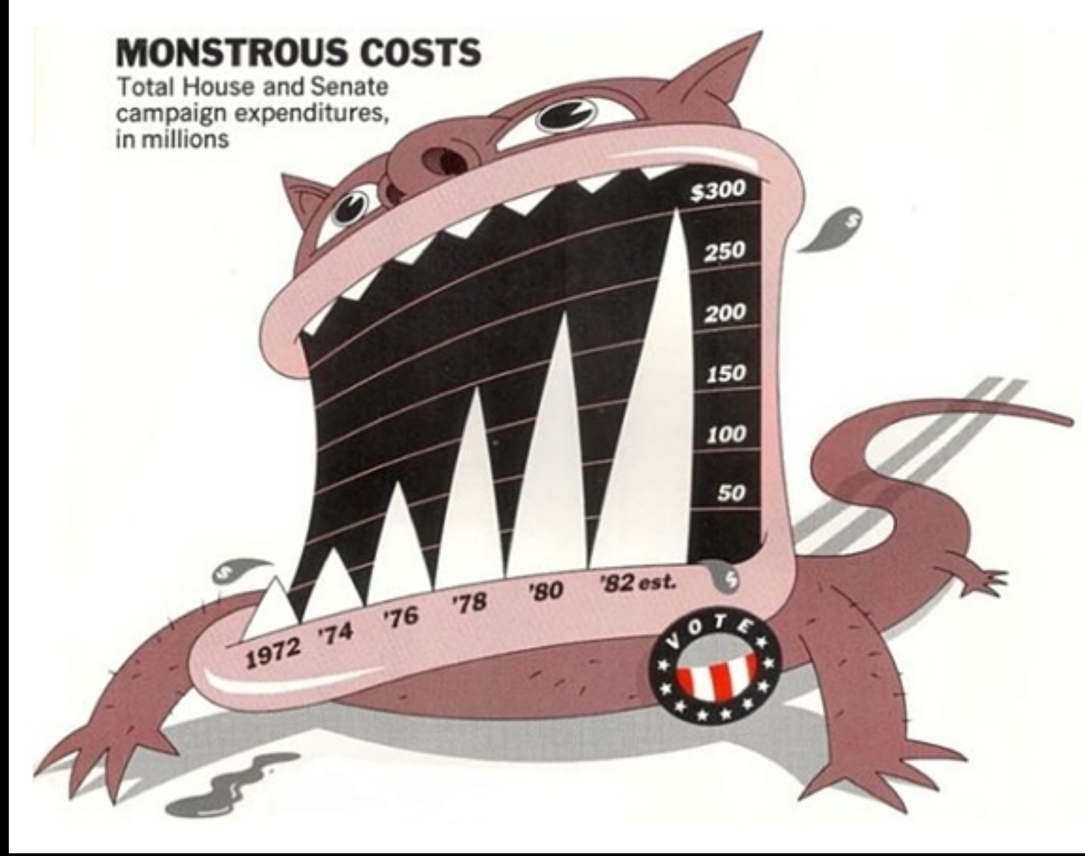
SECOND EDITION

The Visual Display of Quantitative Information

EDWARD R. TUFTE

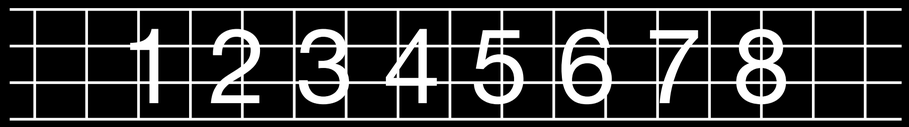
Chartjunk

- **Visual elements** in charts and graphs that are **not necessary** to comprehend the information represented on the graph or that **distract the viewer** from this information.

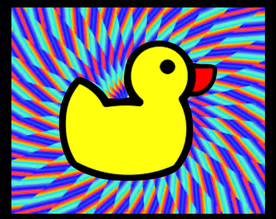


- Examples:
  - Unintentional optical art

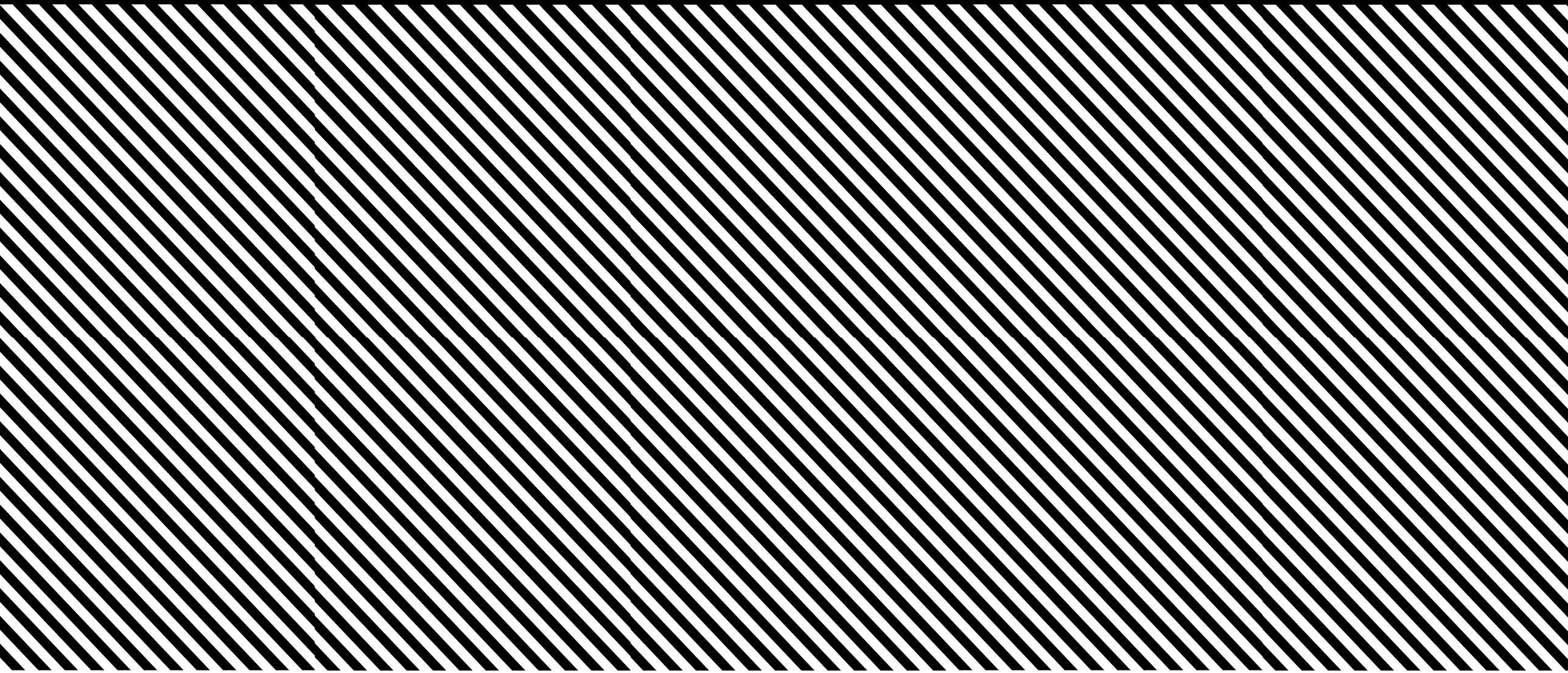
  - Over-busy grid lines and excess ticks



  - Graphical duck



Unintentional optical art

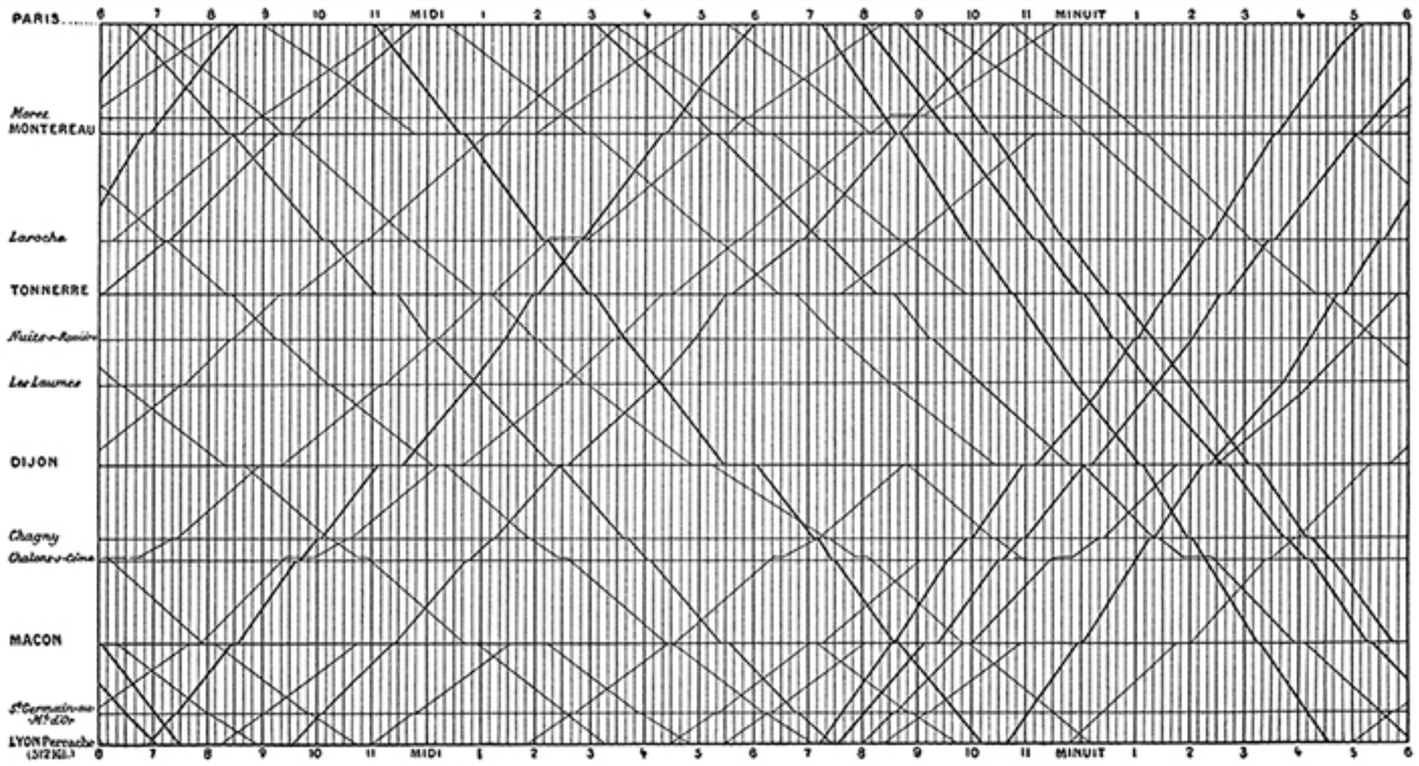


Moiré Effect

Over-busy grid lines and excess ticks

- Grid lines can distract the reader from the data.
- Only lines that **communicate** or **highlight** some **insight** in the data are necessary.

Paris-Lyon 1880 Train Schedule



Graphical Duck

- Visualization that over-emphasizes **style** over (a clear description of the) **data**.
- Analogy to the *Big Duck*—a disfunctional building in New York that only serves the purpose of **decoration**.

- **Duck in iOS**: colors depict the portion of **expenses per category** (e.g., groceries, commute, fashion, etc.)

Visualization Principles
(that work for most visualizations)

1. **Above all else, show the data**. Create the simplest graph that conveys the information you want to present.

2. **Maximize** the **data-ink ratio**. Every bit of ink requires a reason. Nearly always, that reason should be to present new information

3. **Erase non-data-ink**.

4. **Erase redundant** data-ink.

5. **Revise** and **edit**.

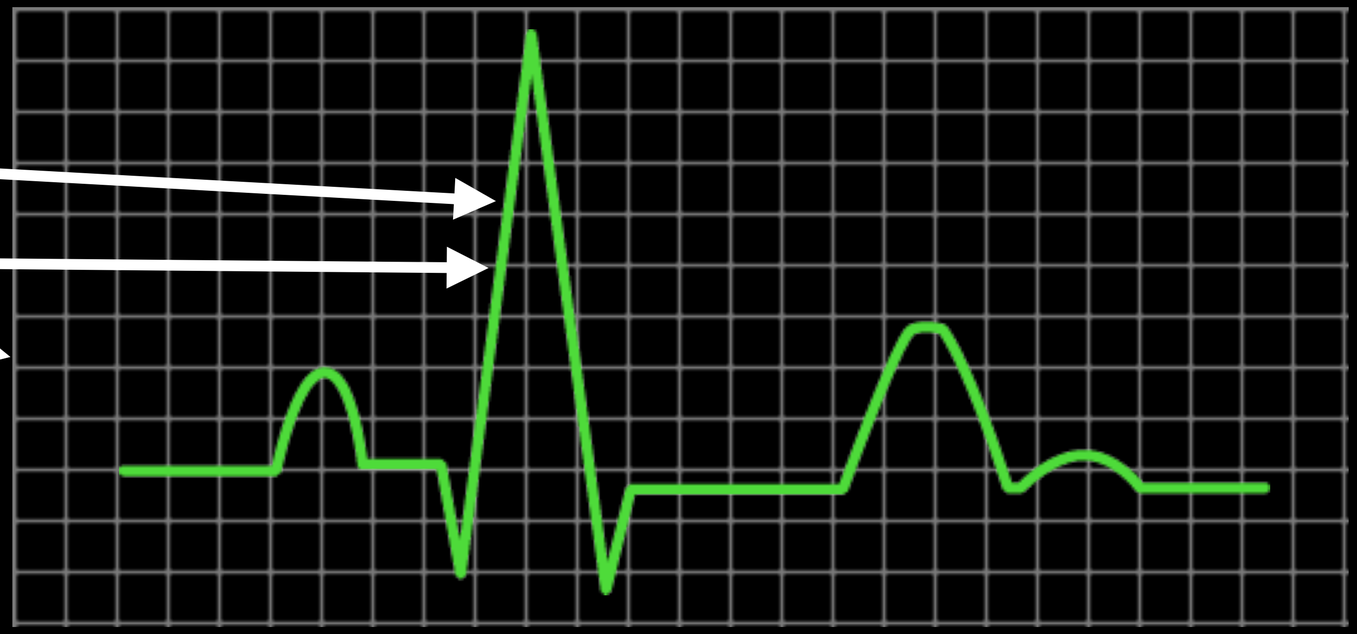
— — What is data-ink ratio? 🤔 — —

Data-Ink Ratio

Visualization Principles

- *Proportion of a graphic's ink devoted to the non-redundant display of data-information.*

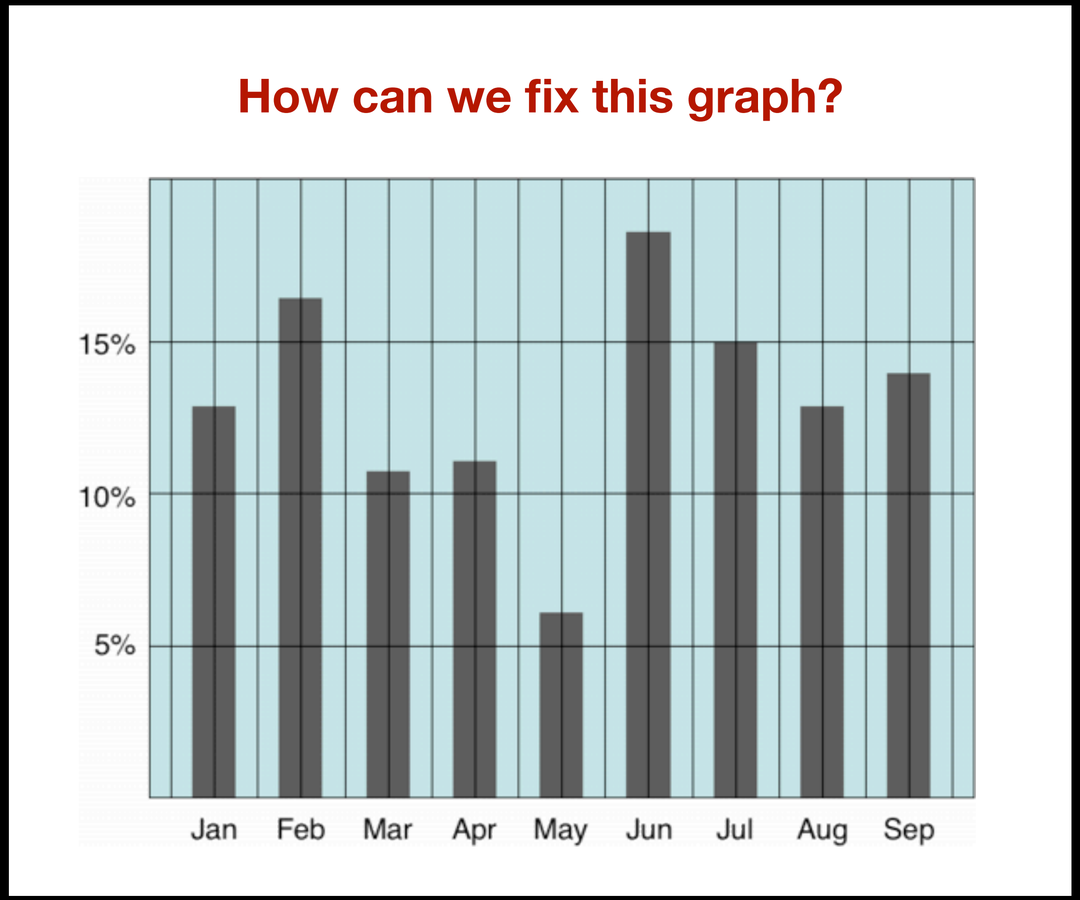
$$\frac{data\ ink}{total\ ink}$$



3. Erase Non-Data-Ink

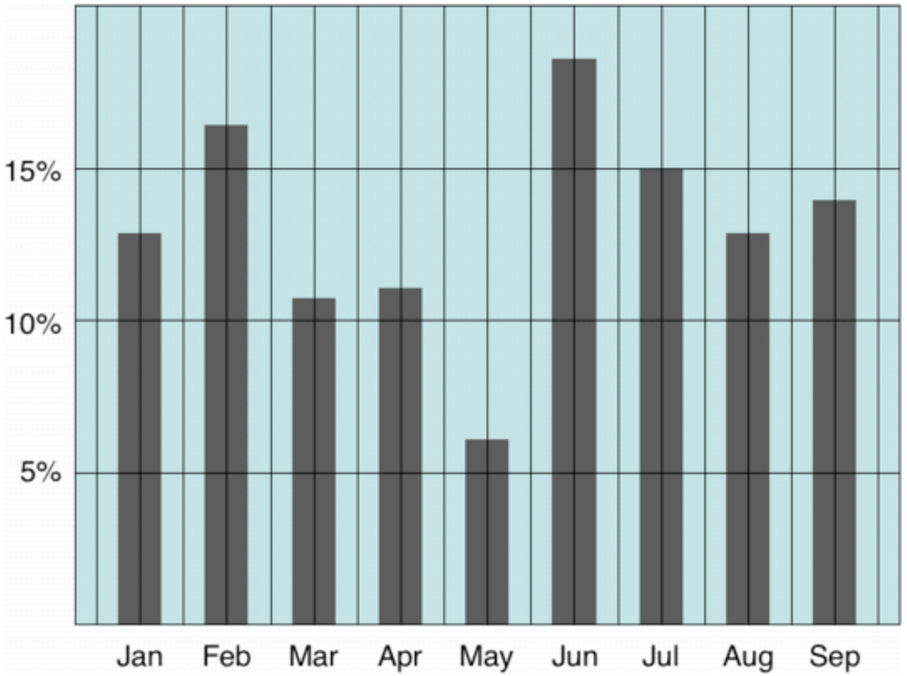
Visualization Principles

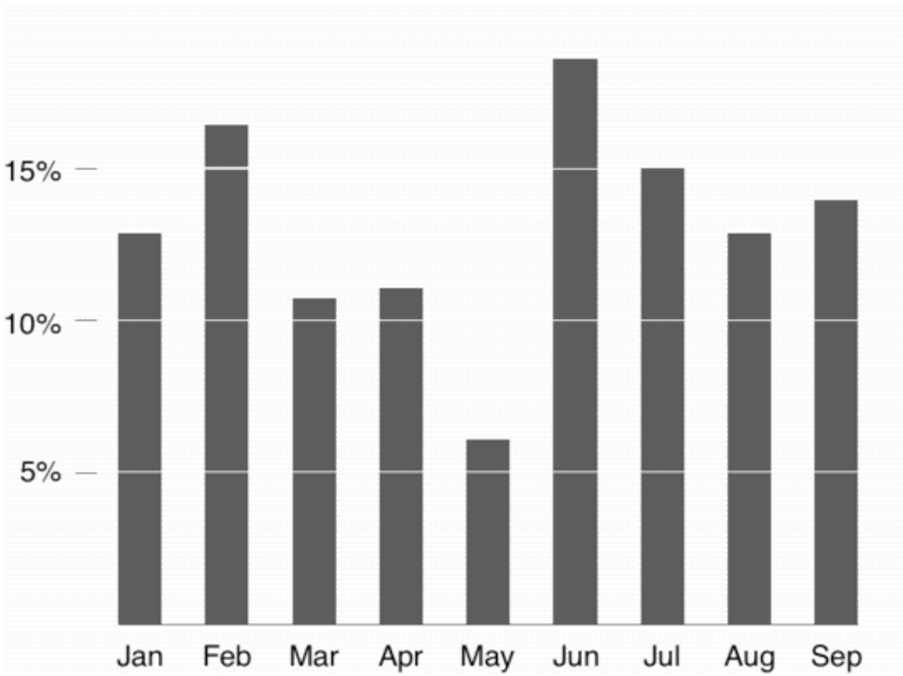
- Examples of non-data-ink:
  - Use of 3D effects
  - Background images
  - Shadow effects
  - Unnecessary borders
  - Unnecessary grid lines.



3. Erase Non-Data-Ink

Visualization Principles

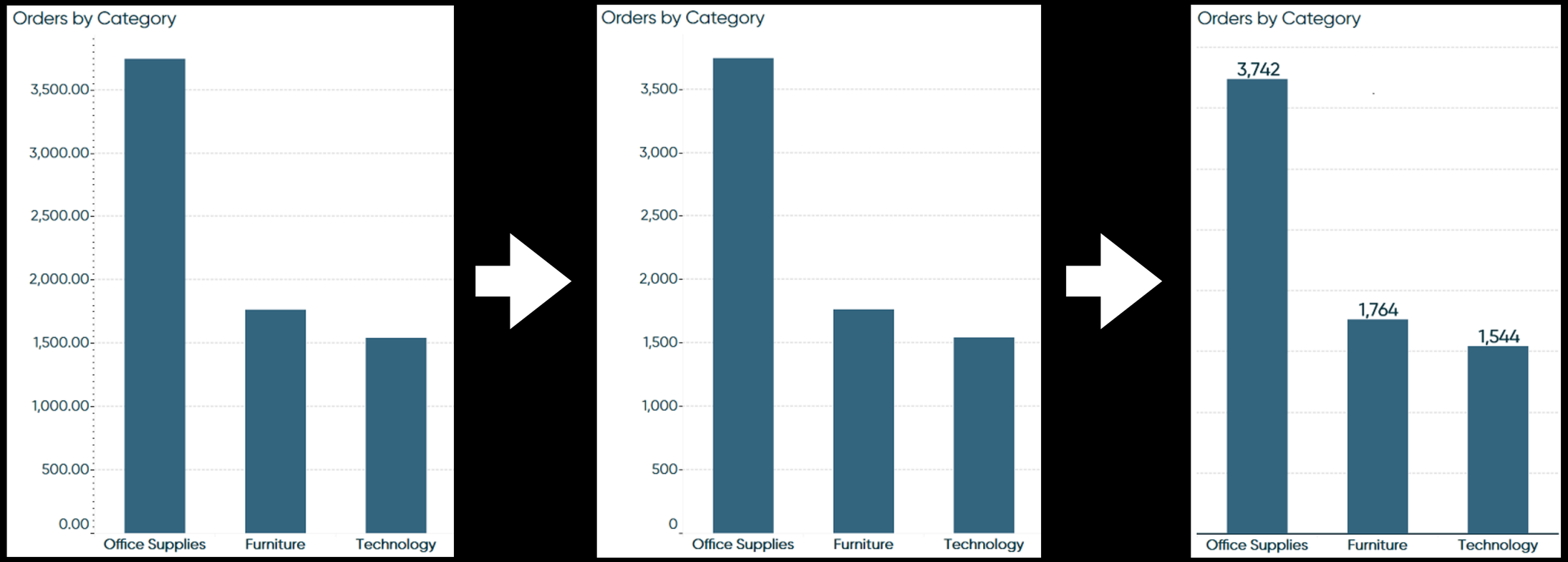




4. Erase Redundant Data-Ink

Visualization Principles

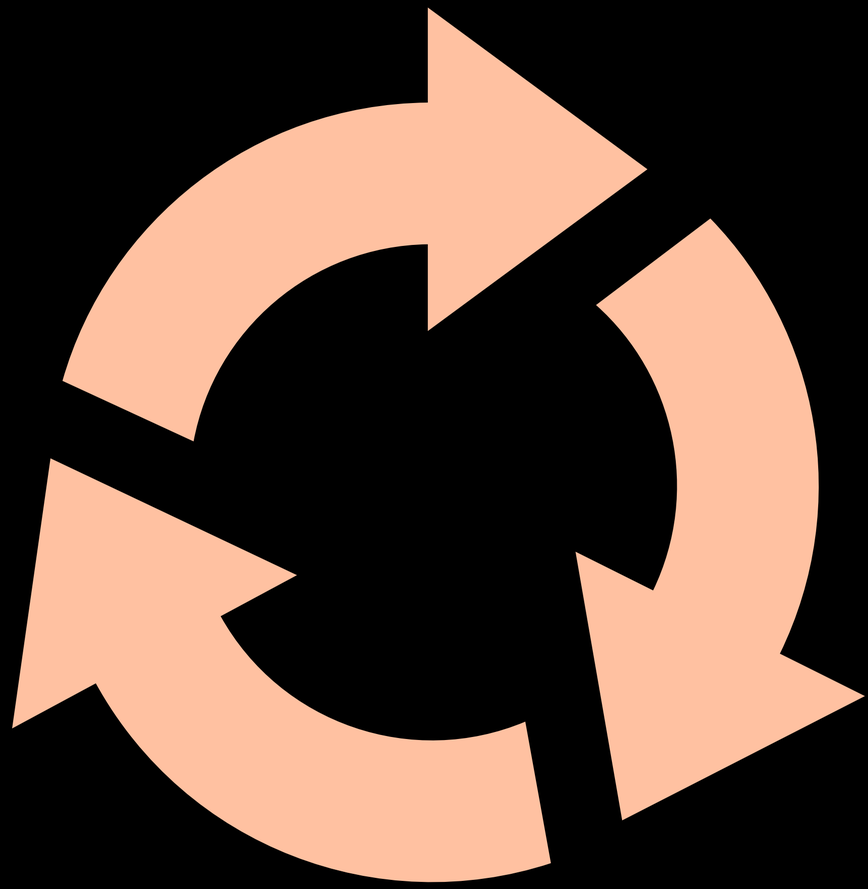
- Unless redundancy has a distinctly worthy purpose.



5. Revise and Edit

Visualization Principles

- **Critical thinking is key**.
- Be **critical about all principles** and design elements.
- Similar to writing or any other **communication** work.
- Ask for **feedback**.

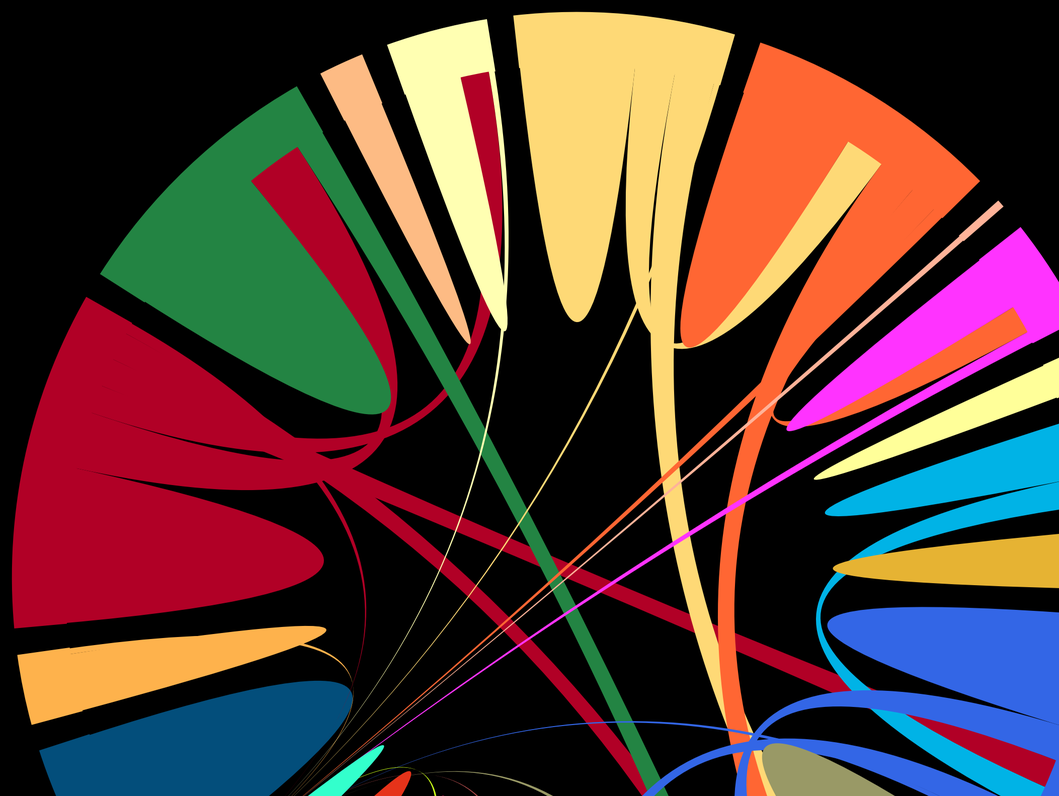


Criticism of Tufte's visualization principles

- Data-ink ratio does not consider visual variations such as **color**.
- People tend to like graphs with **appealing designs**. In some cases, chartjunk may help having a **less boring** graph. (Inbar, 2007)
- Related, recognizable images can help memorability. Redundancy helps recall and understanding. (Borkin, et al., 2015)
- Conclusion: graphs ought to be subject to numerous **revisions** and principles used **wisely**. (vis. principle #5: revise and edit)

Information Visualization
INFO250

Class 11 - Graphical Integrity

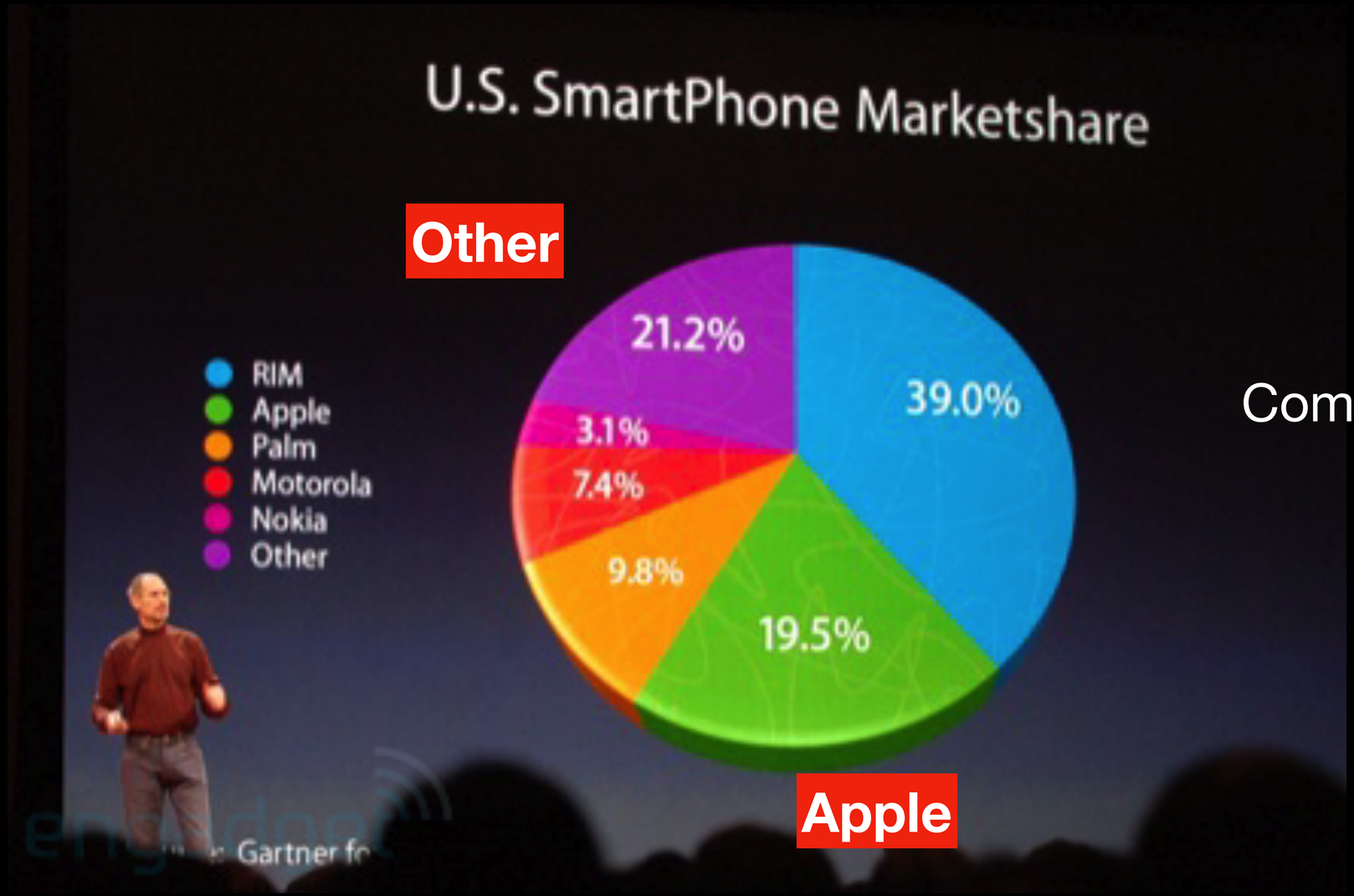


Luís Cruz - l.cruz@tudelft.nl

Graphical Integrity

How accurately visual descriptions represent the actual data being represented?

Graphical Integrity

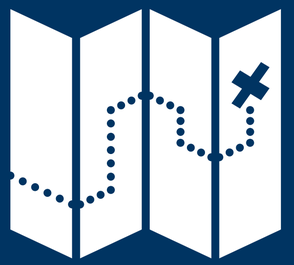


Compare **19.5%** with **21.2%**

by Ryan Womack

Principles for Graphical Integrity

Edward Tufte (1984)



1. The **representation** of numbers, as physically measured on the surface of the graph itself, should be **directly proportional to the numerical quantities represented**

2. Clear, detailed and thorough **labeling should be used to defeat graphical distortion and ambiguity**. Write out explanations of the data on the graph itself. Label important events in the data.

3. Show **data variation**, not design variation.

4. (In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.) **Important but out-of-scope of this class.**

5. The number of information carrying (variable) **dimensions depicted** should not exceed the **number of dimensions in the data**.

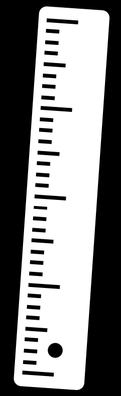
6. Graphics **must not quote data out of context**. Representations should not imply an unintended context.

Lie Factor

Useful for a precise discussion on graphical integrity.

$$\text{Lie Factor} = \frac{\text{Size of the effect shown in the graph}}{\text{Size of the effect shown in the data}}$$

0.95 < Lie Factor < 1.05 ✅

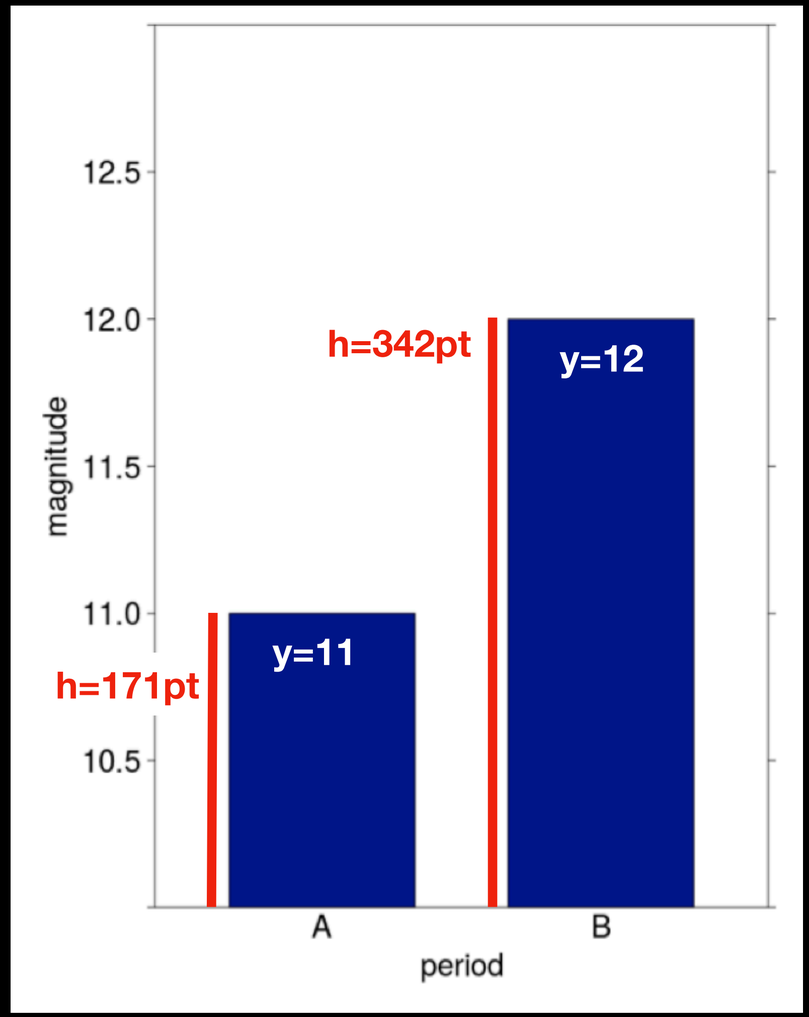


Lie Factor

$$Effect\ Size\ of\ Data = \frac{y_B - y_A}{y_A} = \frac{12 - 11}{11} = 0.09$$

$$Effect\ Size\ of\ Graph = \frac{h_B - h_A}{h_A} = \frac{342 - 171}{171} = 2$$

$$LieFactor = \frac{EffectSize\ of\ Graph}{EffectSize\ of\ Data} = \frac{2}{0.09} = 22.22$$

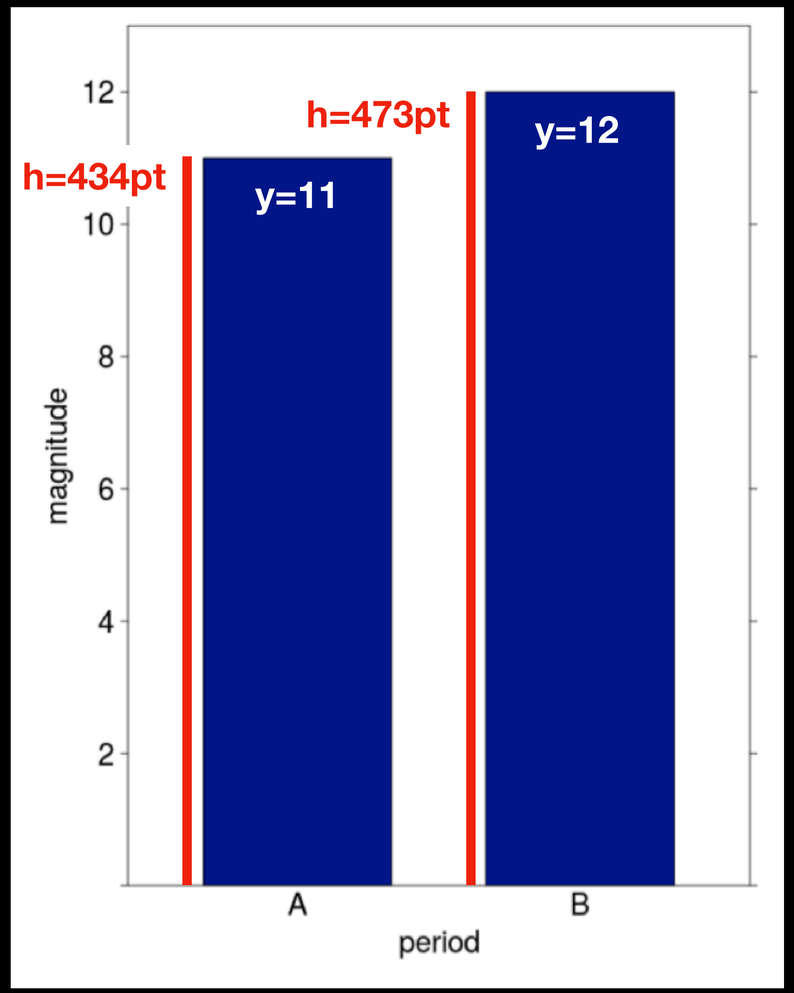


Lie Factor — fixed

$$Effect\ Size\ of\ Data = \frac{y_B - y_A}{y_A} = \frac{12 - 11}{11} = 0.09$$

$$Effect\ Size\ of\ Graph = \frac{h_B - h_A}{h_A} = \frac{473 - 434}{434} = 0.09$$

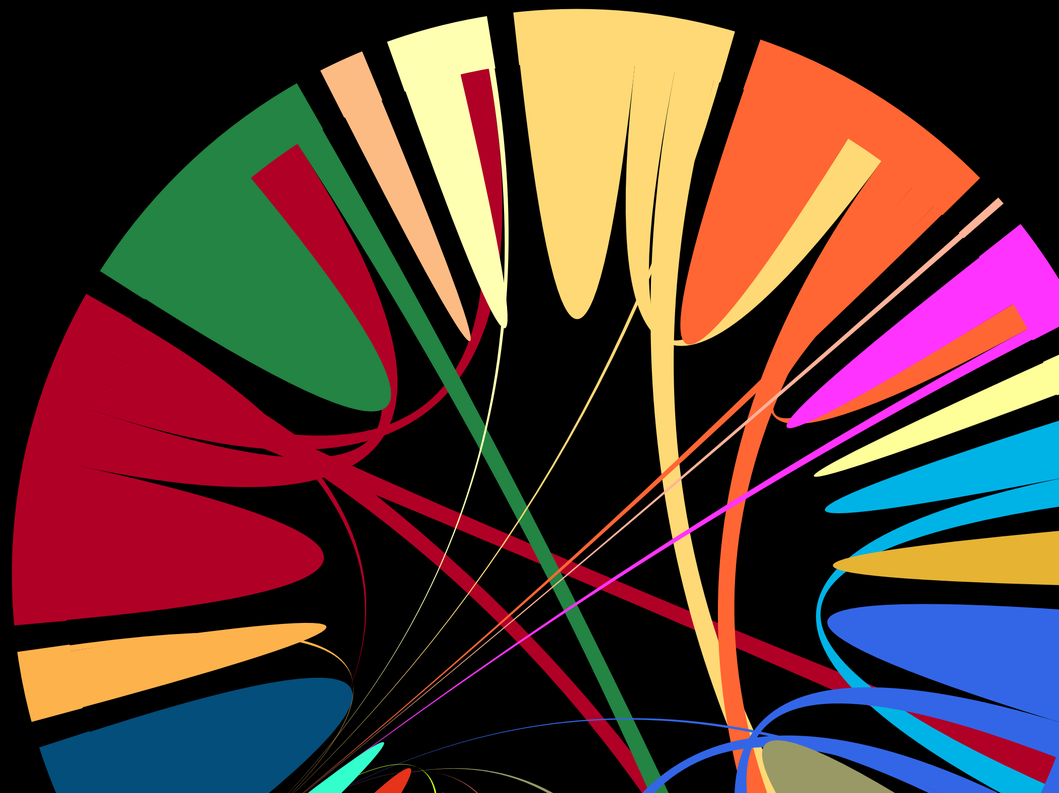
$$LieFactor = \frac{EffectSize\ of\ Graph}{EffectSize\ of\ Data} = \frac{0.09}{0.09} = 1$$



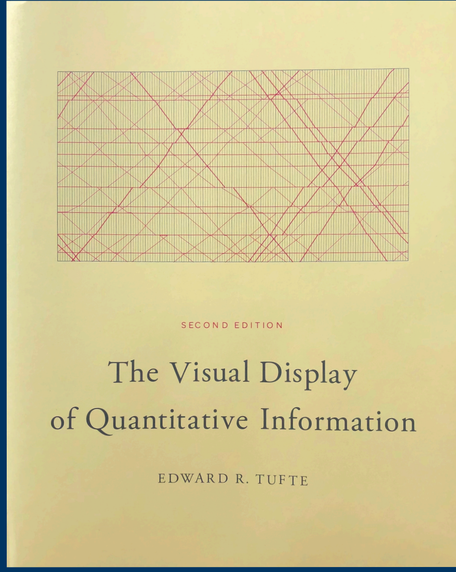
Information Visualization
INFO250

Class 12
Tufte’s principles 36 years later
Color blind
Visualizations in Academic Publications

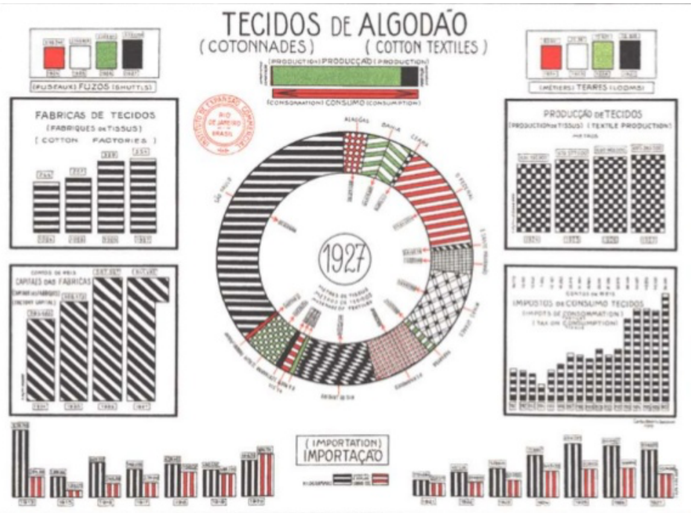
Luís Cruz - l.cruz@tudelft.nl

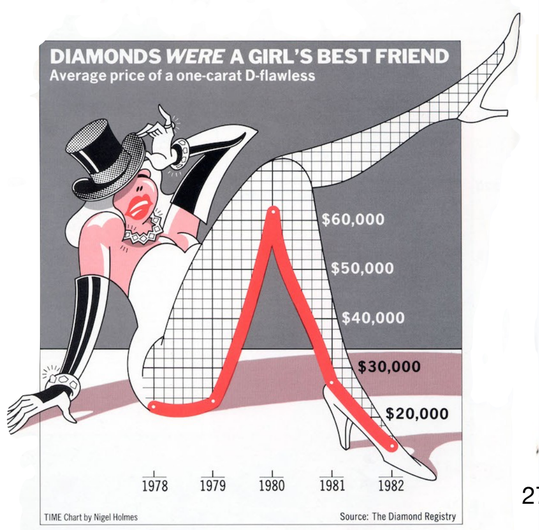


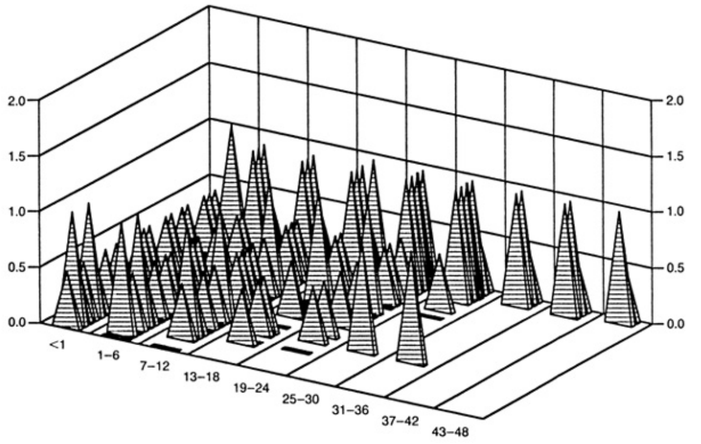
Tufte’s principles 36 years later



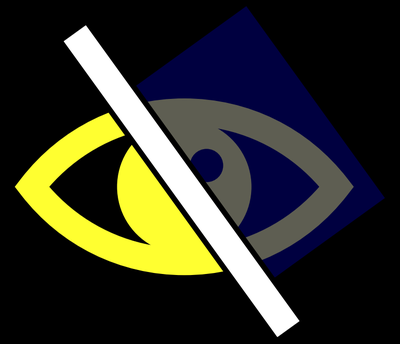
- Tufte’s principles were written in 1984.
- Most examples show really old graphs.
- Since then, there have been **numerous advancements** in the field of information visualization.





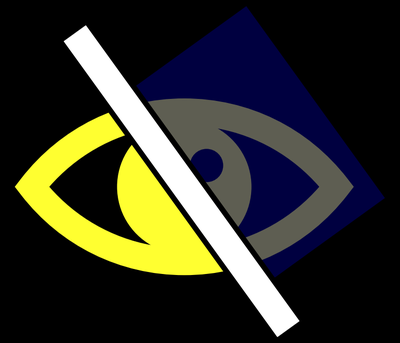


Color blindness



- Decreased ability to see color or differences in color.
- It is estimated to affect **8% of men**, and **0.5% of women**.
- Red–green color blindness is the most common form, followed by blue–yellow color blindness, and total color blindness.
- Information visualisation ought to be colorblind-friendly.

Color blindness - Tips



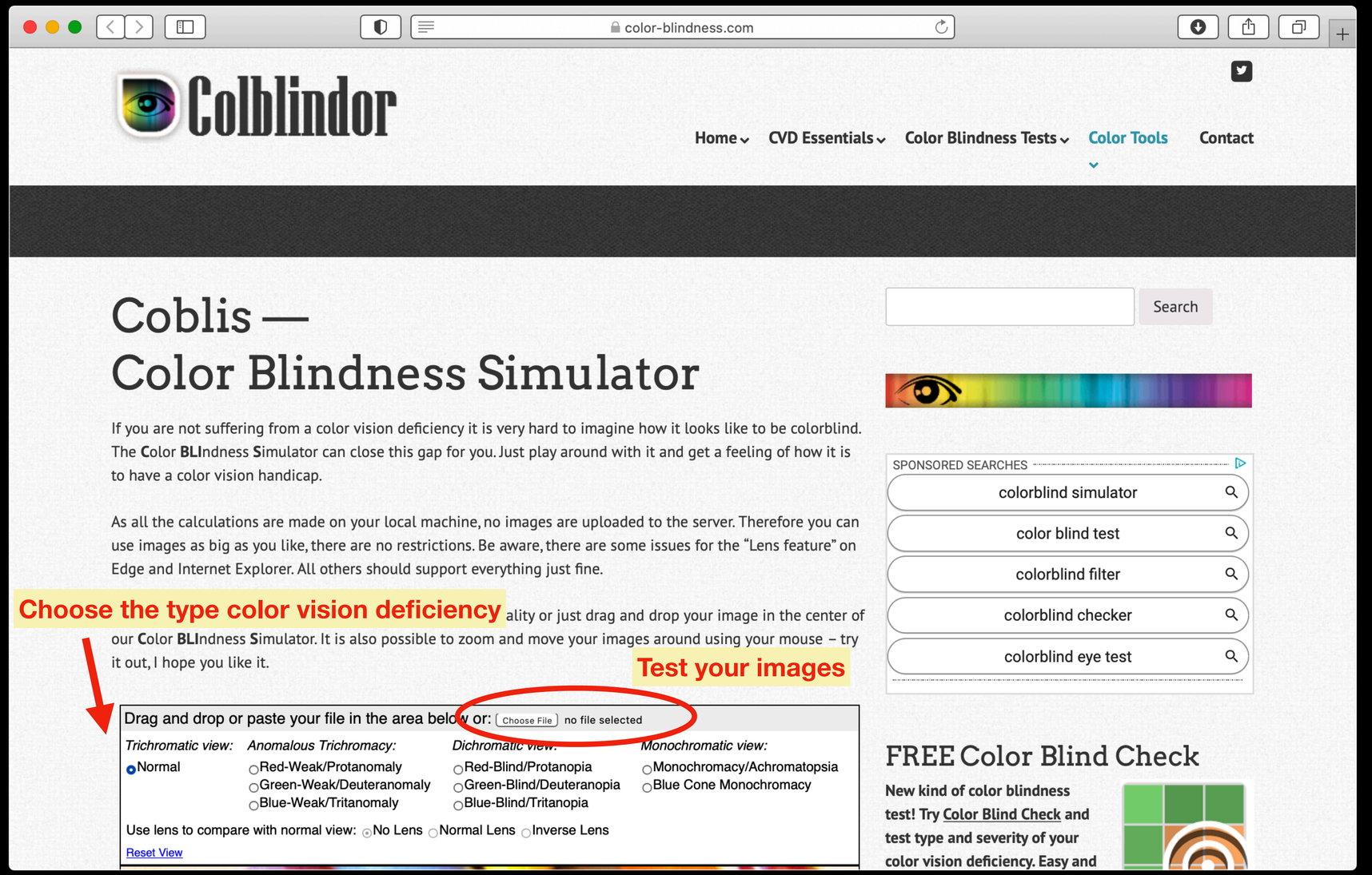
- **Avoid using red and green together.**
  - Sometimes red is used to denote **negative** insights and green to denote **positive**. **It is okay** to use this analogy as long it is redundant with another visual variable.
  - If you have to use red and green, use **different shades** (e.g., light vs dark).
- Use **colorblind-friendly palettes**. Example by Tol (2019) (https://personal.sron.nl/~pault/):

- Test your image with color blindness **simulation tools**. 🔄

Coblis

Color Blindness Simulation

https://www.color-blindness.com/coblis-color-blindness-simulator/

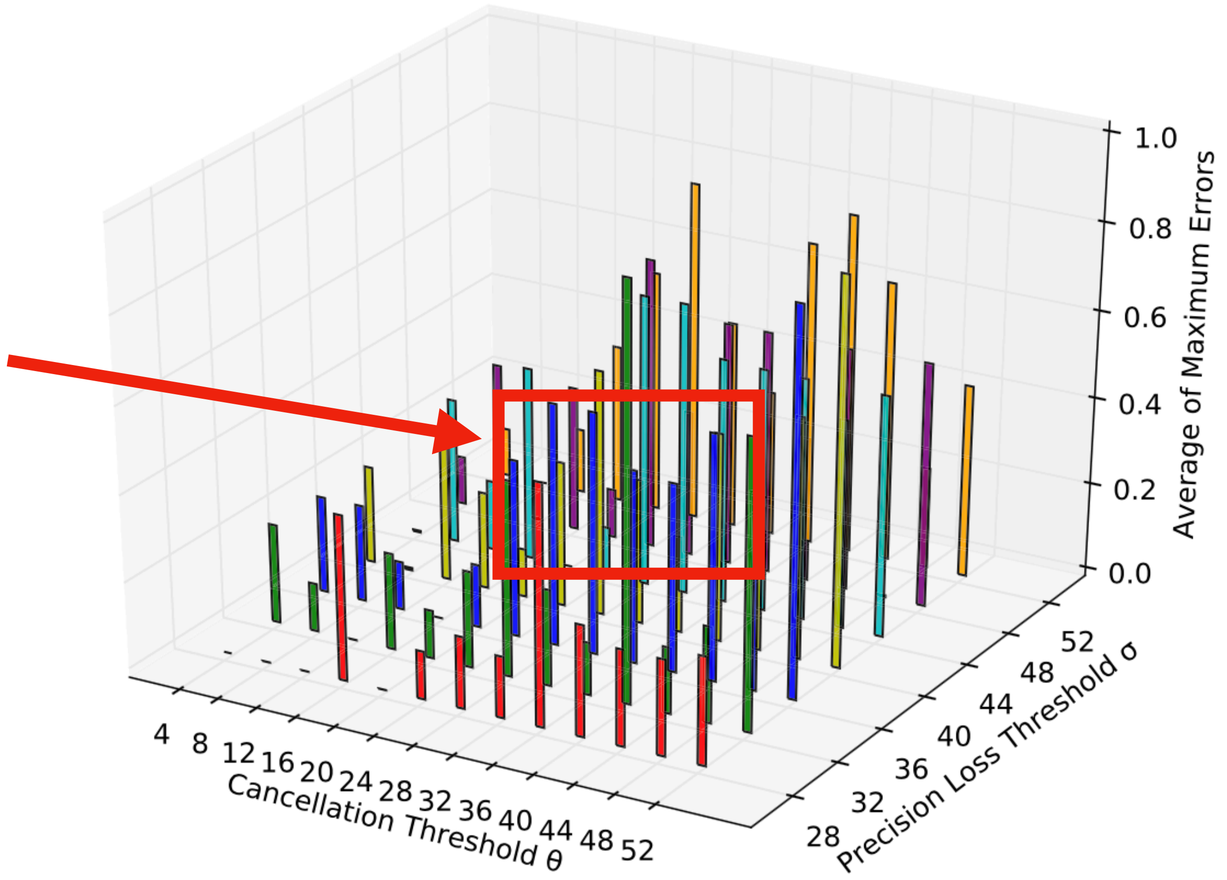


Graphs in the proceedings of ICSE 2020

- The 42th International Conference on Software Engineering (2020).
- One of the **main conferences** in **Software Engineering**.
- Despite many great visualizations, **some of them are confusing and could be improved** with basic visualization principles.
- Some examples in the following slides.

Graphs at ICSE 2020

How to read these bars in the middle?

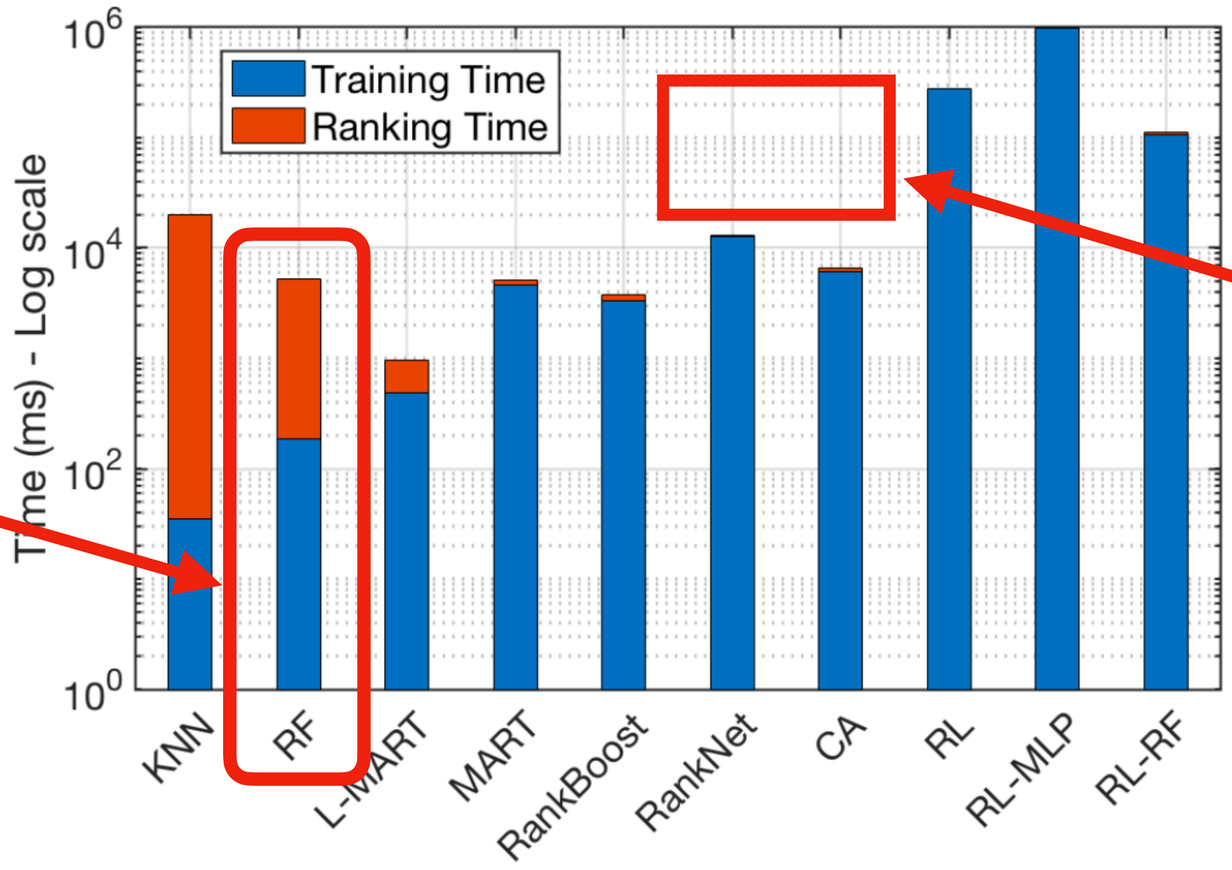


Guo (2020) efficient

Figure 5: Average of maximum errors on 3 summation programs with different inaccuracy thresholds.

Graphs at ICSE 2020

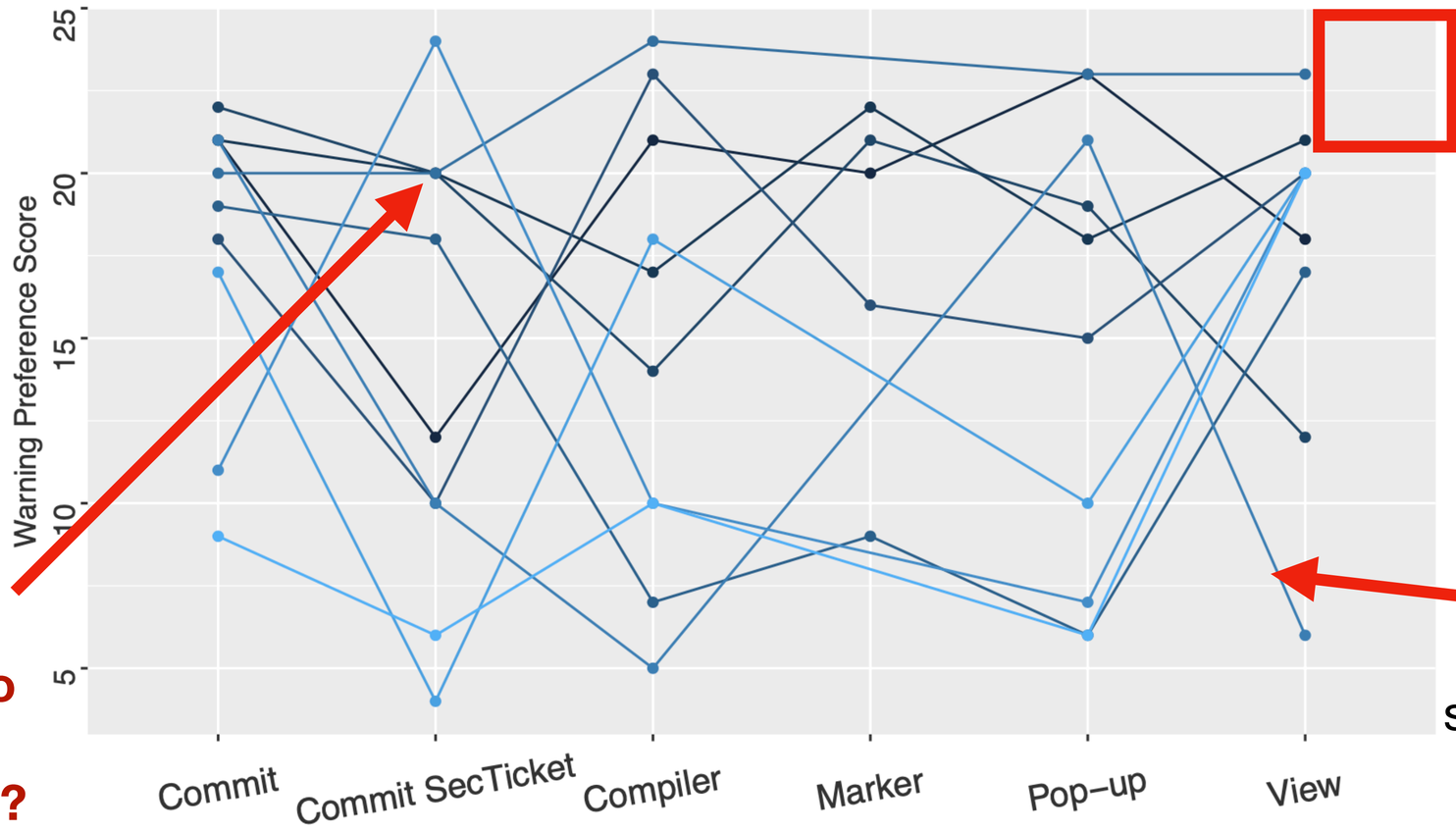
What is bigger:
Ranking time or Training time?



Busy gridlines

Bertolino (2020)

Graphs at ICSE 2020



Unnecessary background

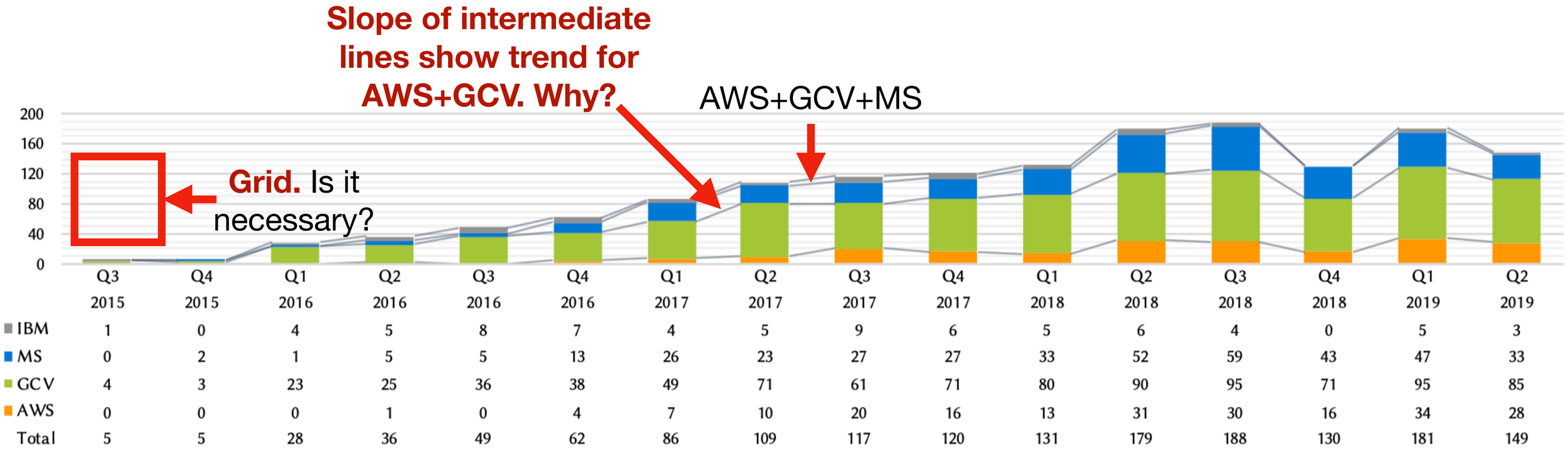
Where did each line go after the intersection?

Lines represent a relationship between two subsequent data points by showing a slope. **What's their meaning here?**

What is the main message of this graph?

Danilova (2020) One

Graphs at ICSE 2020

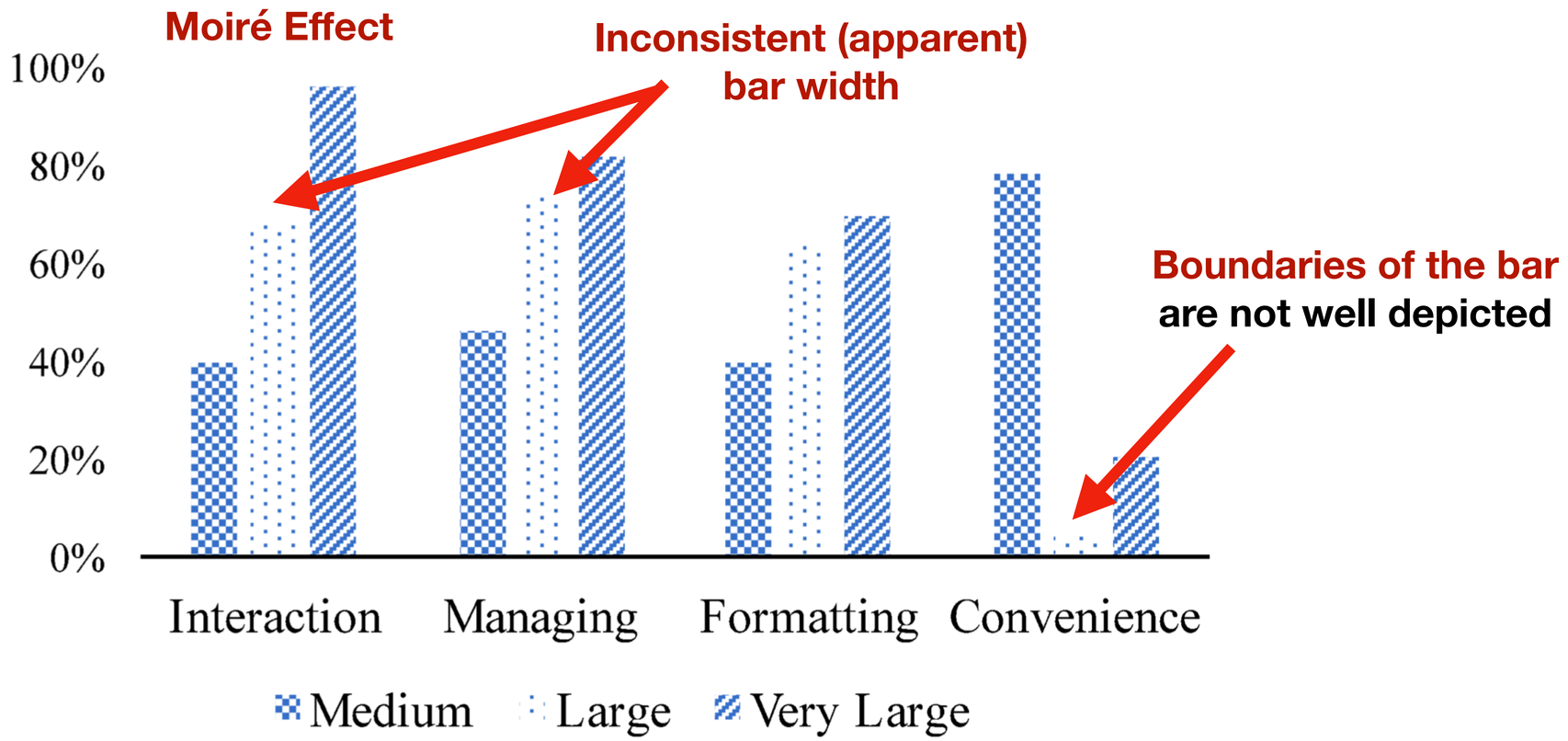


	Q3 2015	Q4 2015	Q1 2016	Q2 2016	Q3 2016	Q4 2016	Q1 2017	Q2 2017	Q3 2017	Q4 2017	Q1 2018	Q2 2018	Q3 2018	Q4 2018	Q1 2019	Q2 2019
IBM	1	0	4	5	8	7	4	5	9	6	5	6	4	0	5	3
MS	0	2	1	5	5	13	26	23	27	27	33	52	59	43	47	33
GCV	4	3	23	25	36	38	49	71	61	71	80	90	95	71	95	85
AWS	0	0	0	1	0	4	7	10	20	16	13	31	30	16	34	28
Total	5	5	28	36	49	62	86	109	117	120	131	179	188	130	181	149

Figure 2: Trend of posts, where IBM = IBM Watson Visual Recognition, MS = Azure Computer Vision, AWS = AWS Rekognition and GCV = Google Cloud Vision. Three MS posts from Q4 2012, Q3 2013 and Q4 2013 have been removed for graph clarity.

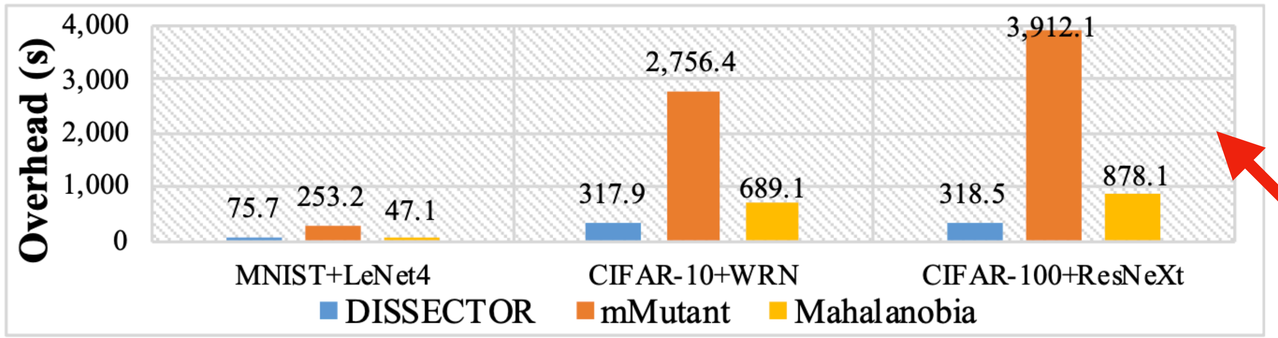
Cummaudo (2020) interpreting

Graphs at ICSE 2020

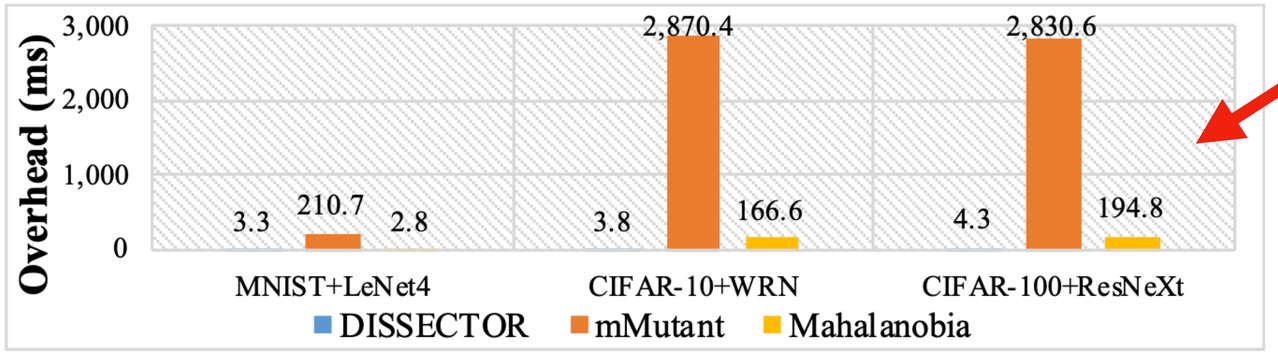


Chen (2020) Studying

Graphs at ICSE 2020



(a) Offline time overhead comparison (in second)



(b) Online time overhead comparison (in millisecond)

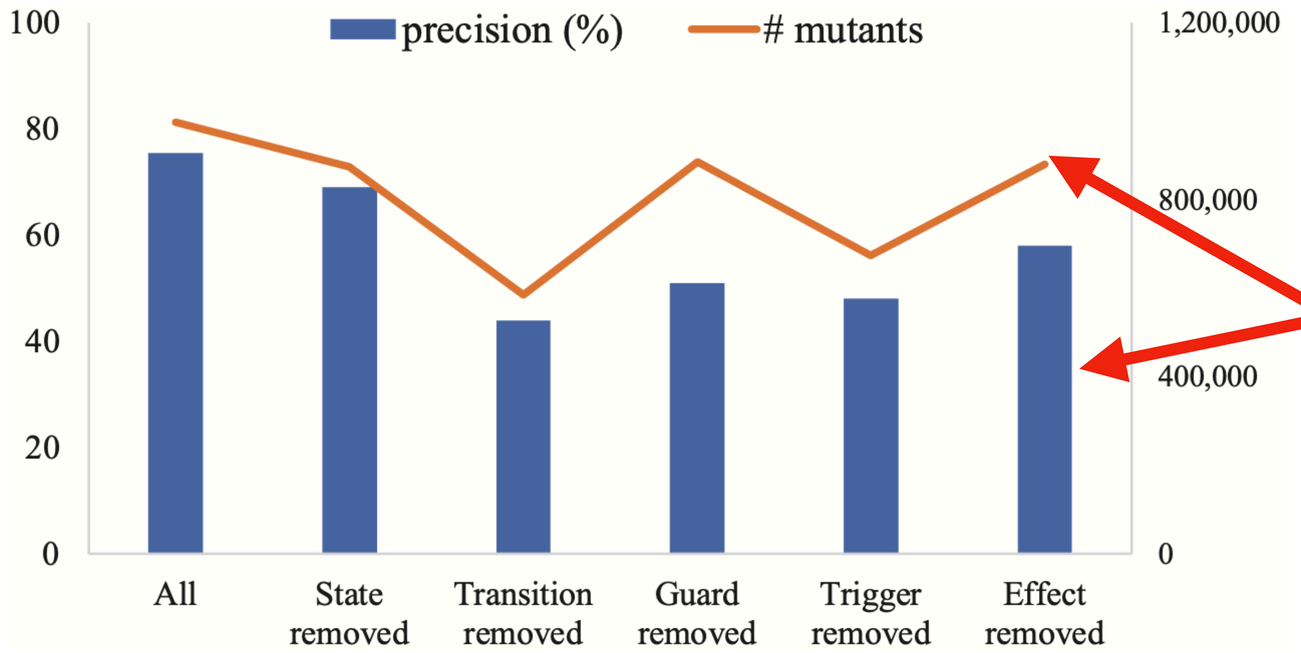
Moiré Effect. Why?

Figure 4: Time overhead comparison

Wang (2020) Dissector

Graphs at ICSE 2020

(this is not a Pareto Chart)



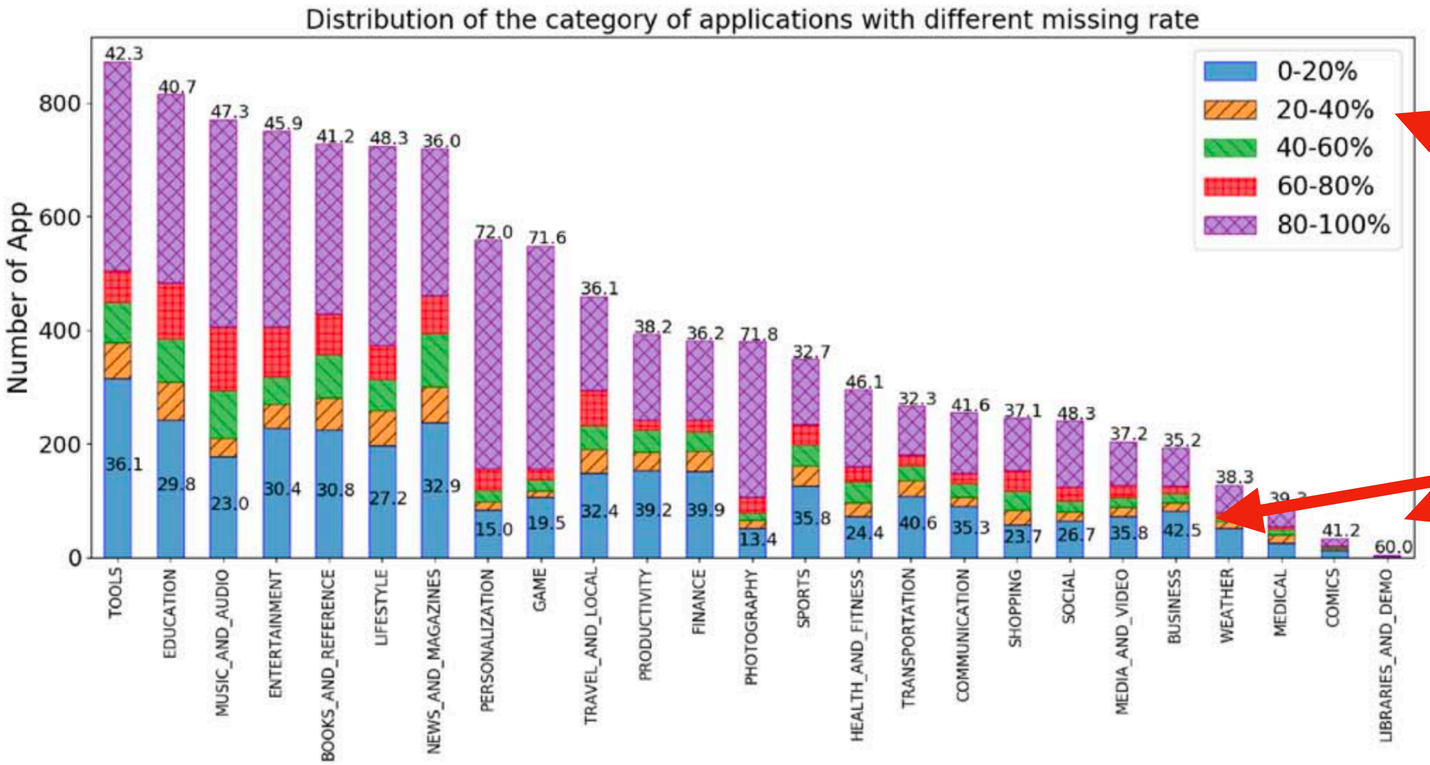
Line and bars compare variables with different units.

Figure 9: Ablation results of removing one and only one category of mutation operators.

Alenazi (2020) novel

Graphs at ICSE 2020

Color palette?
Patterns?



Weird way of showing the shape of distribution. Why not using box plots?

Difficult to process labels

Chen (2020) Unblind

Figure 4: The distribution of the category of applications with different rate of image-based buttons missing content description

Graphs at ICSE 2020

Moire Effect…
Inconsistent Patterns…

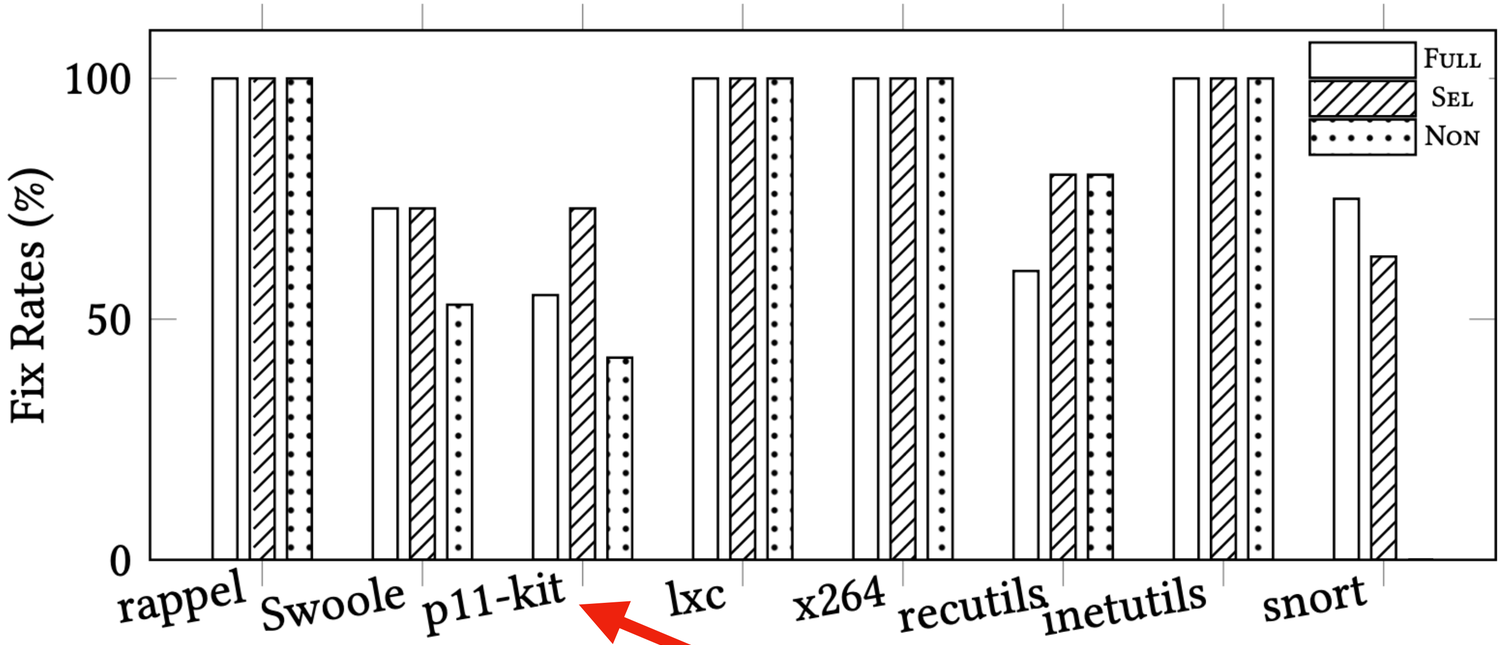


Figure 7: Fix rates with different path-sensitivities

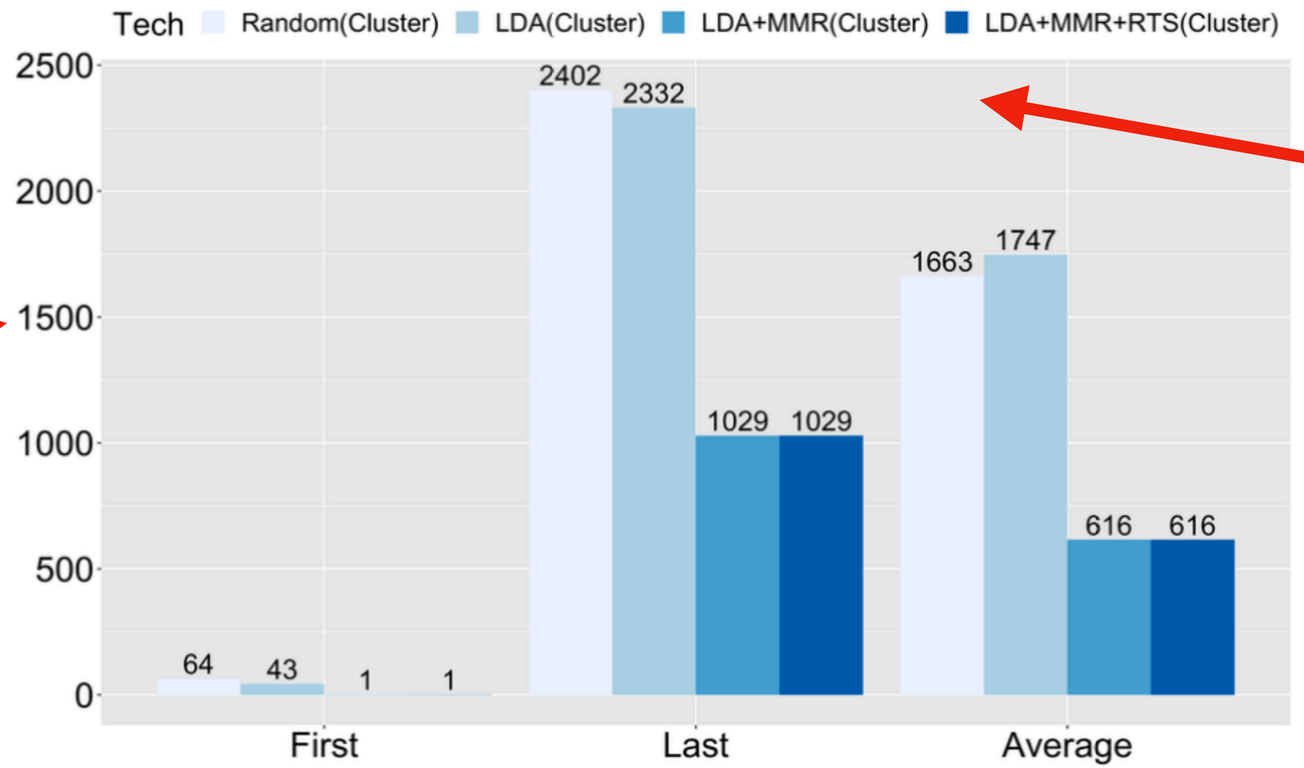
Hong (2020) Saver

Misplaced labels

Graphs at ICSE 2020

Moire Effect…
Inconsistent Patterns…

Why do we have labels here?

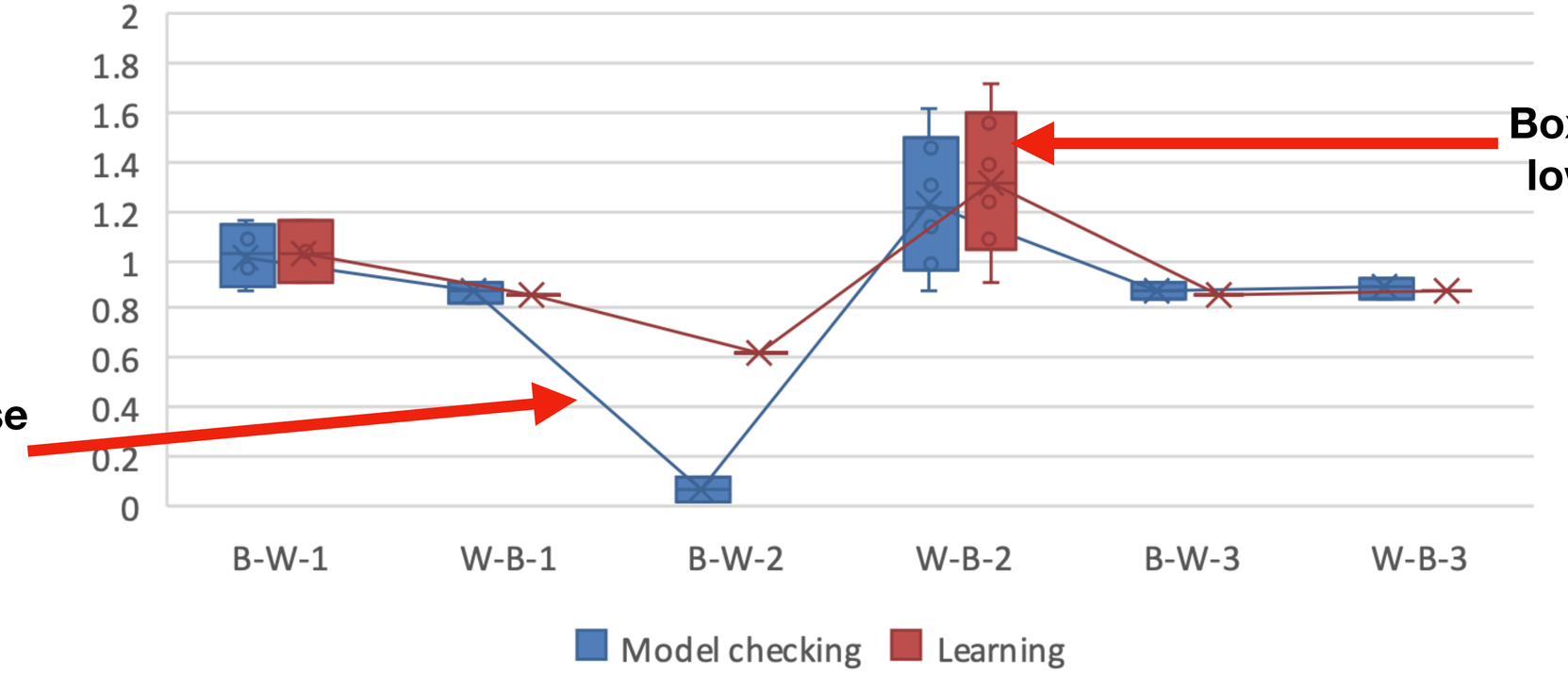


Background

Figure 4: Client project execution

Chen 2020 Taming

Graphs at ICSE 2020



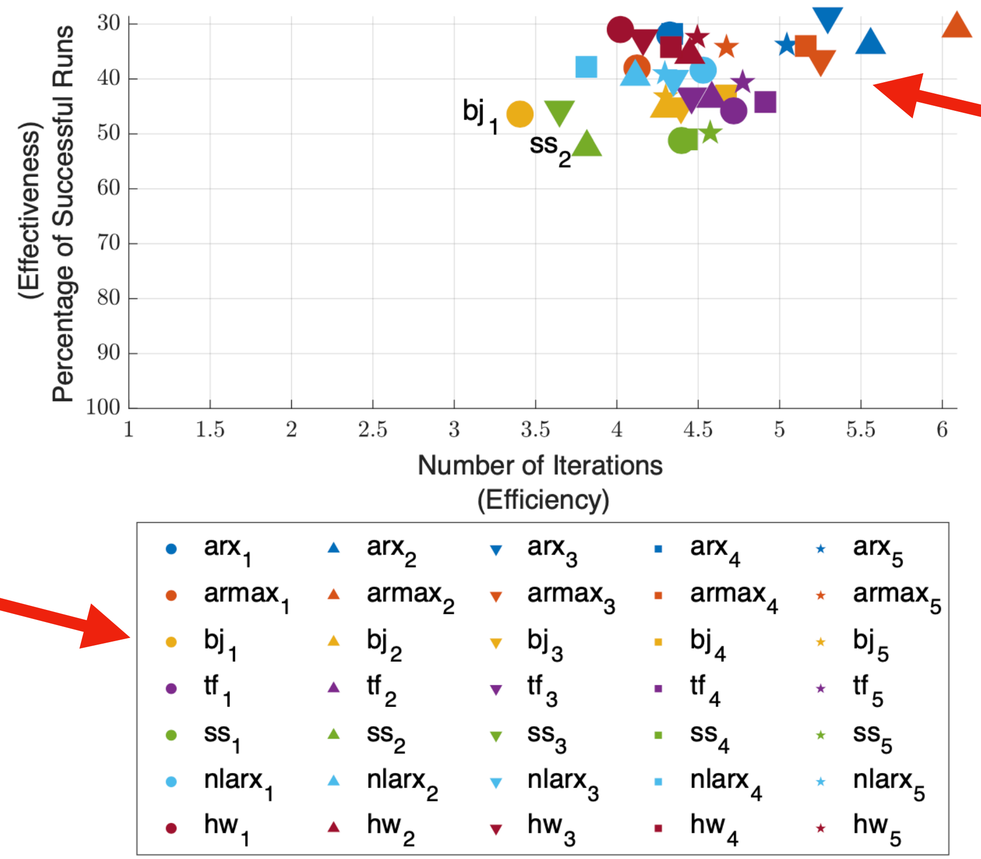
Box plots with low contrast

Why do these lines represent?

Figure 4: Performance of RESPIRE in interactive mode.

Alrajeh2020adapting

Graphs at ICSE 2020



Very Difficult to distinguish

Need to go back and forward to process the legend

Figure 2: Effectiveness and efficiency of different configurations across our non-CI-CPS subject models.

Menghi 2020 Approximation-Refinement

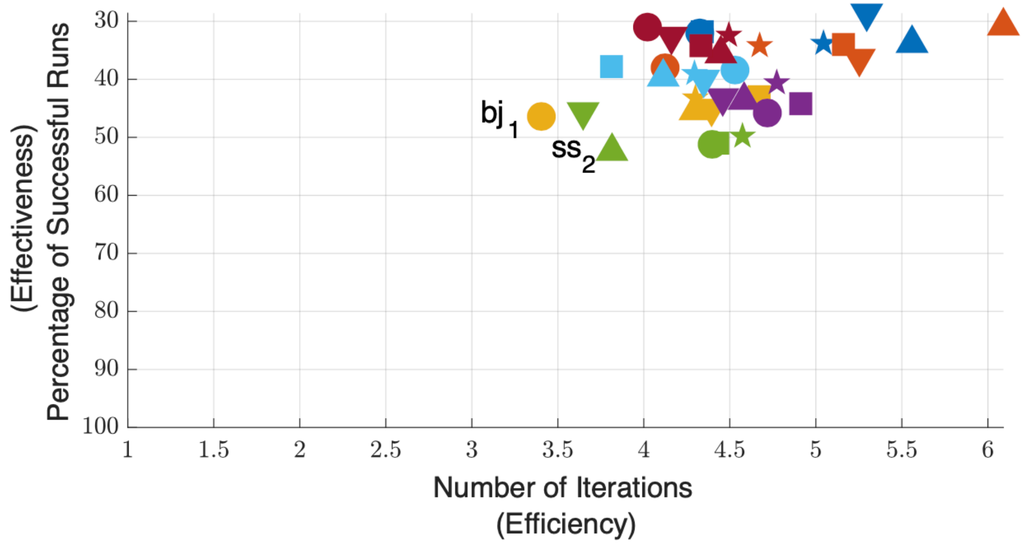


Figure 2: Effectiveness and efficiency of different configurations across our non-CI-CPS subject models.

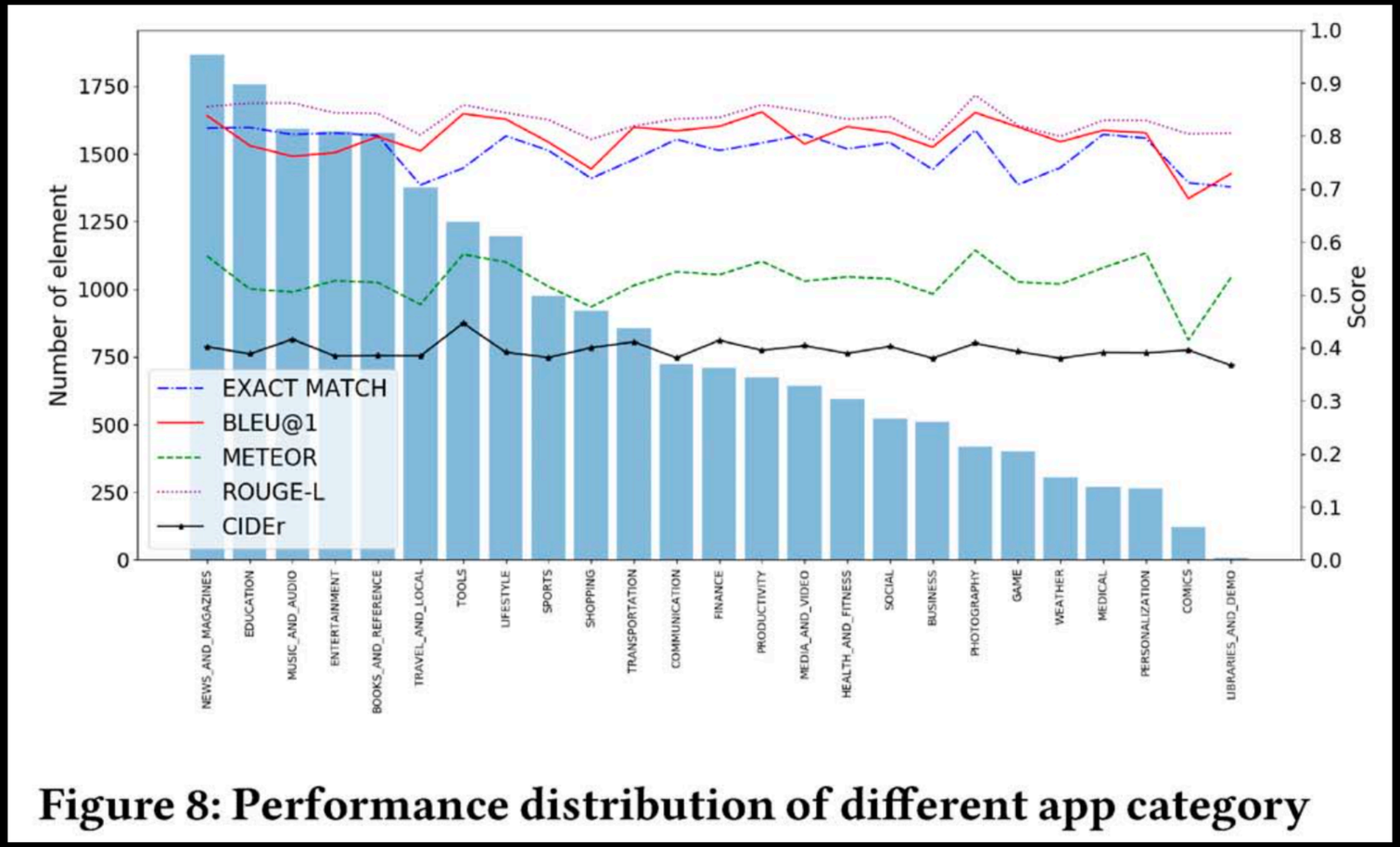


Figure 8: Performance distribution of different app category

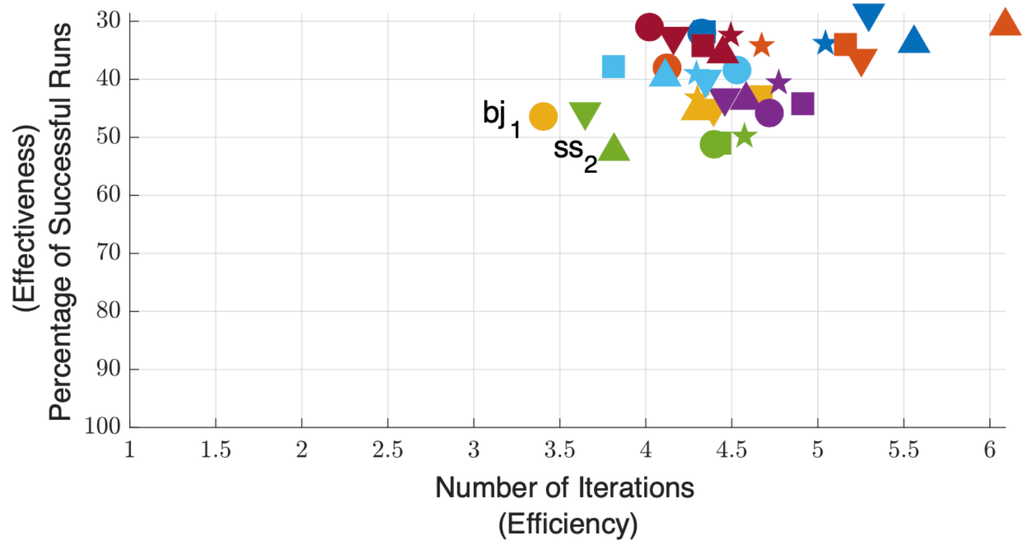
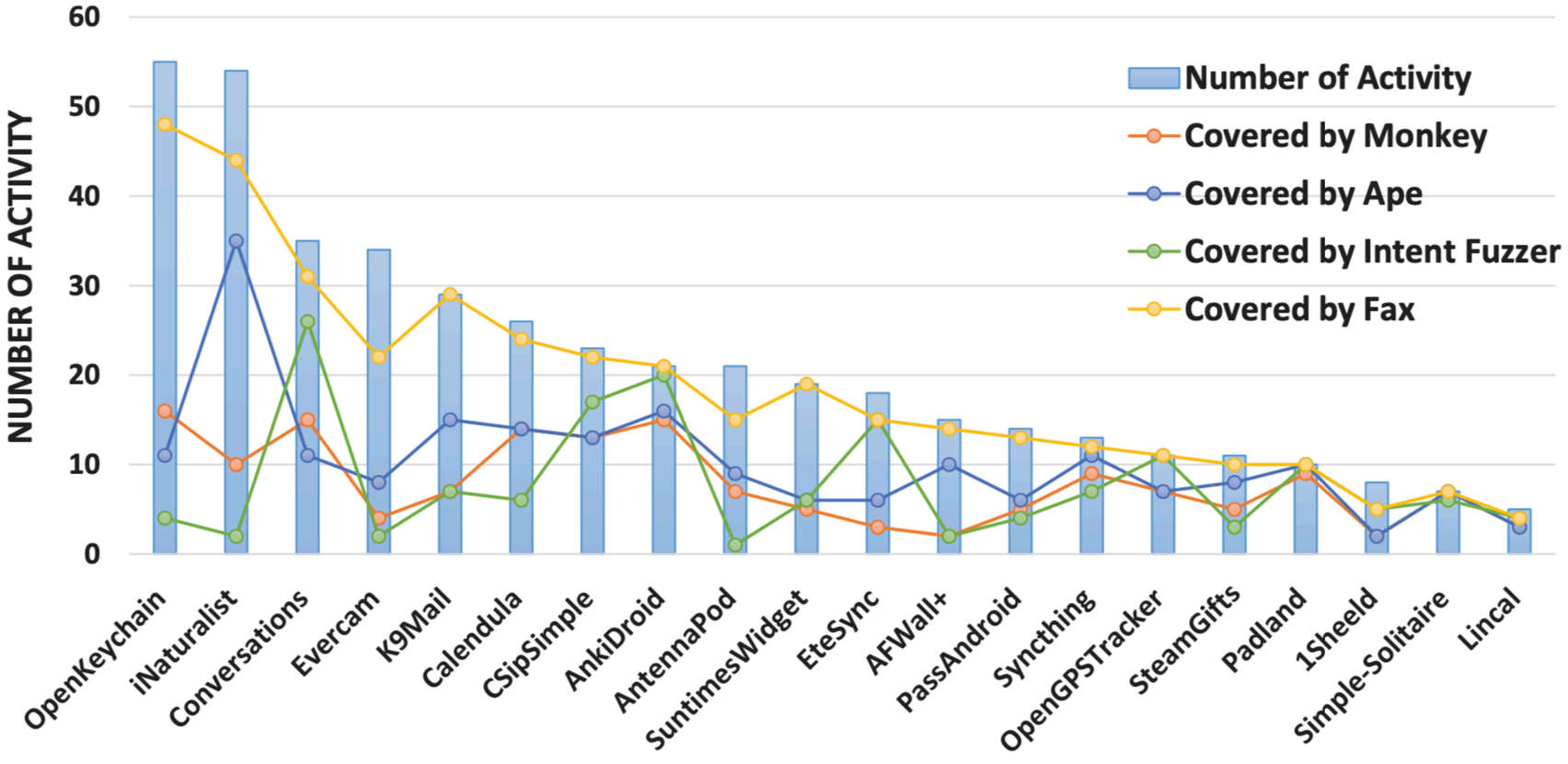
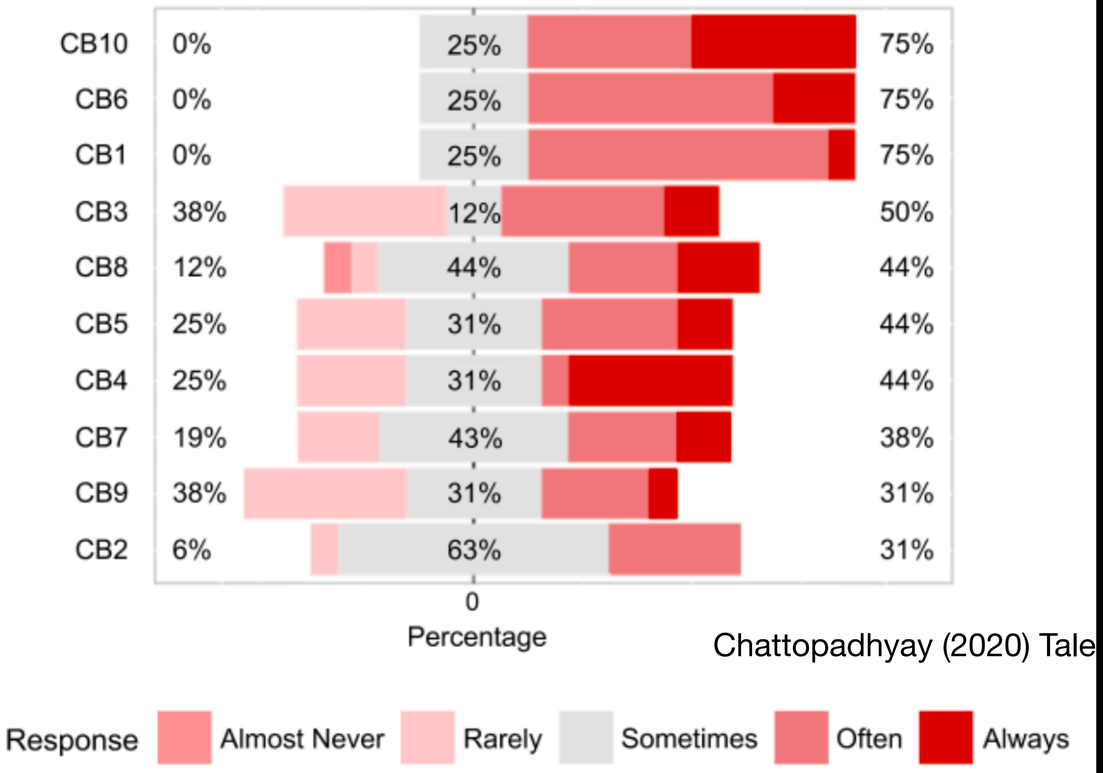


Figure 2: Effectiveness and efficiency of different configurations across our non-CI-CPS subject models.



Yan (2020) multiple

Figure 6: Activity Launching Comparison



Chattopadhyay (2020) Tale

Figure 3: Perceived Frequency of Biases from Interviews ranging from 'Almost Never' to 'Always'. Bias categories are ordered in descending order from most to least frequent (shown as percentages).

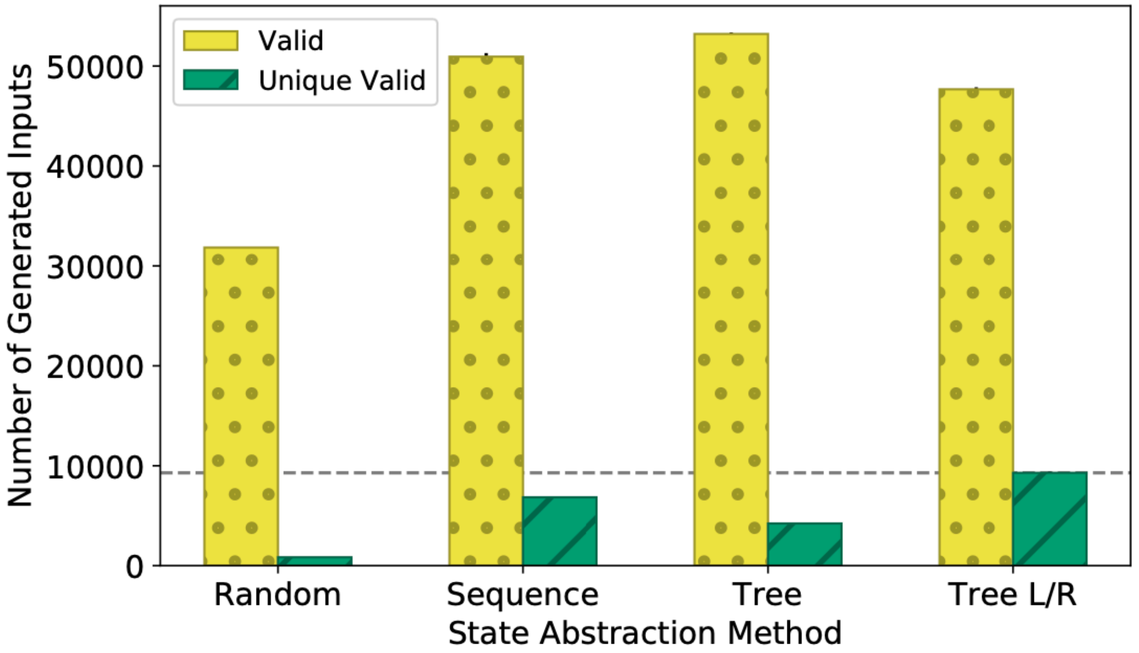


Figure 4: Number of (unique) valid inputs generated, by state abstraction. "Random" is a no-RL baseline.

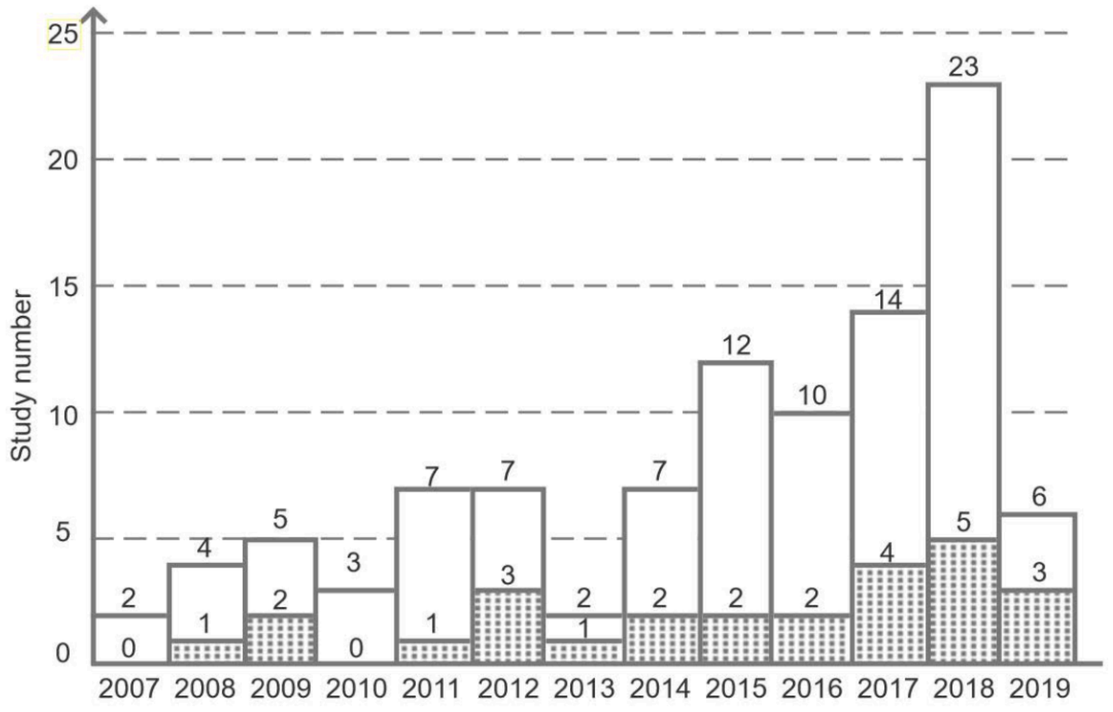


Figure 4: Distribution of reviews with GL over years

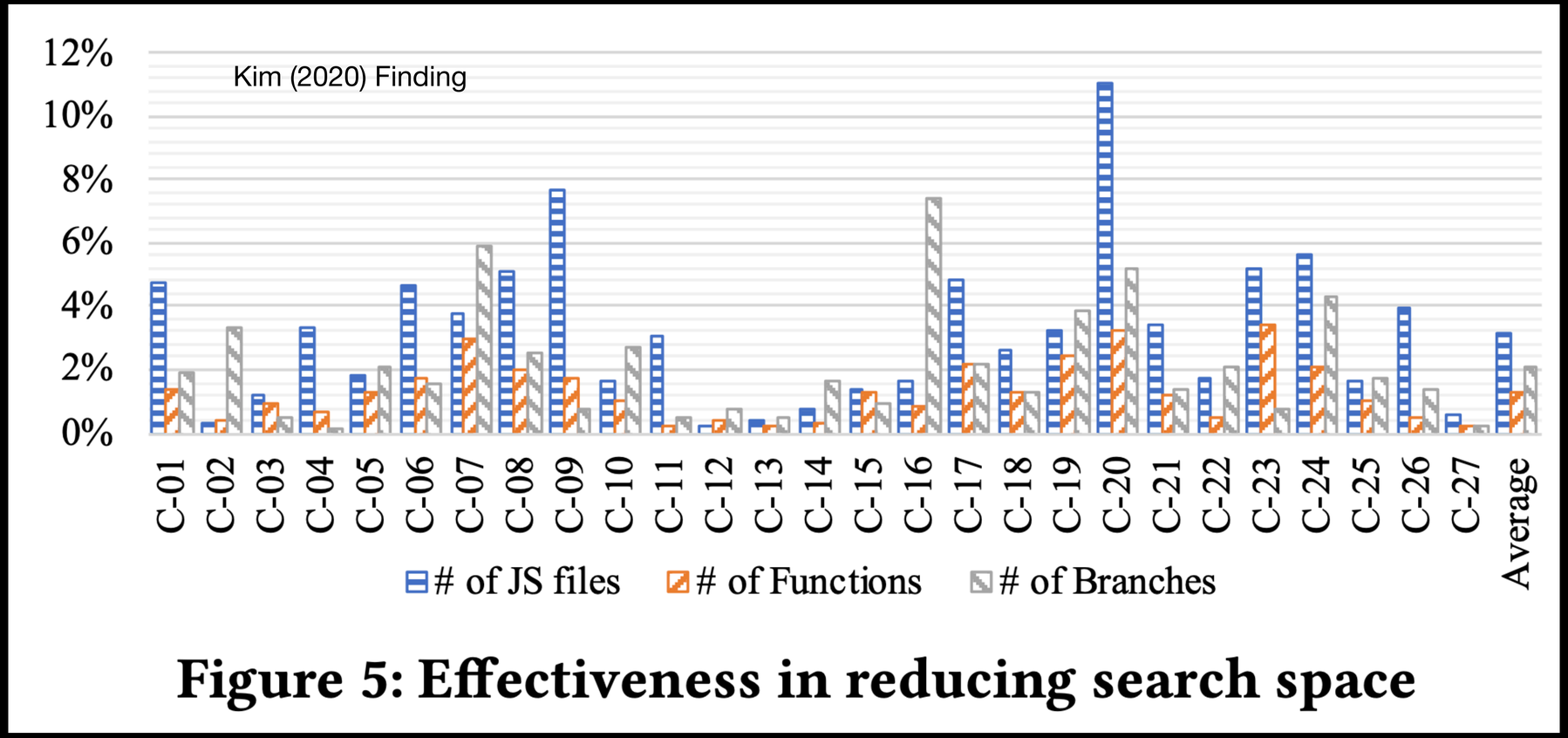
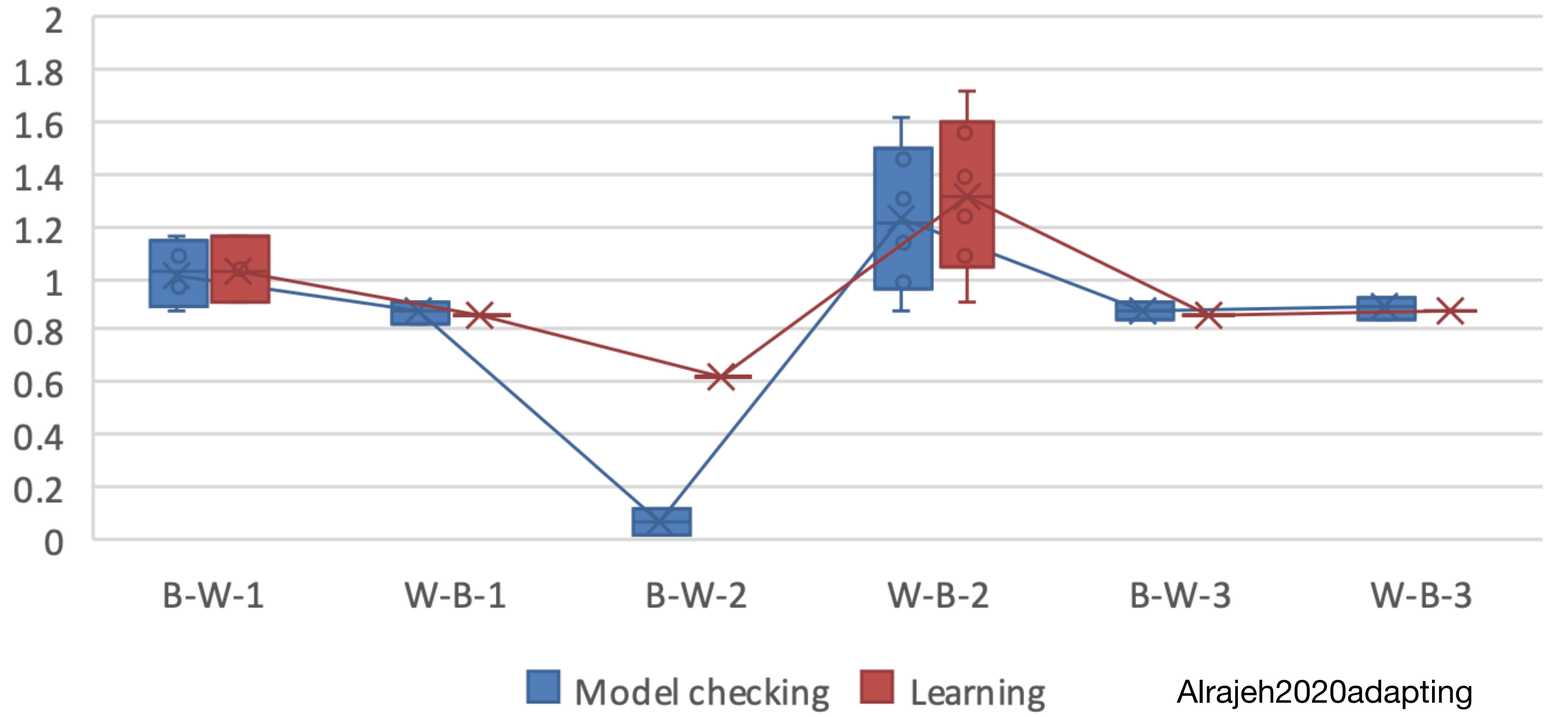


Figure 5: Effectiveness in reducing search space



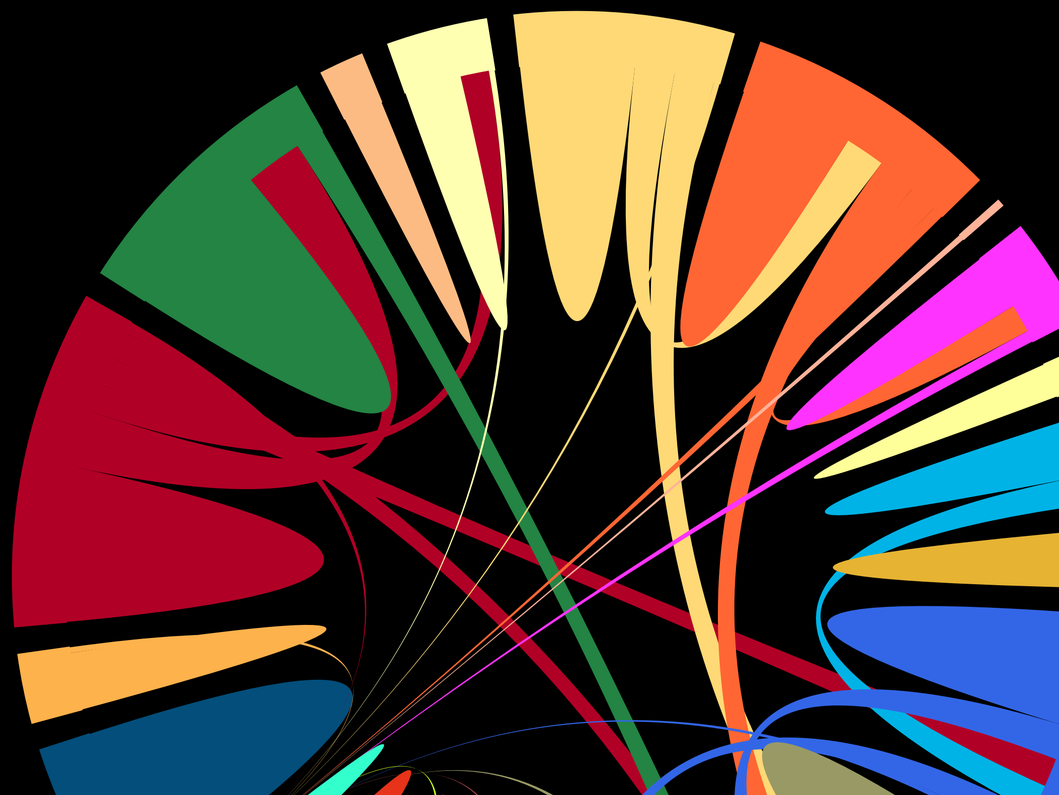
Alrajeh2020adapting

Figure 4: Performance of RESPIRE in interactive mode.

Information Visualization INFO250

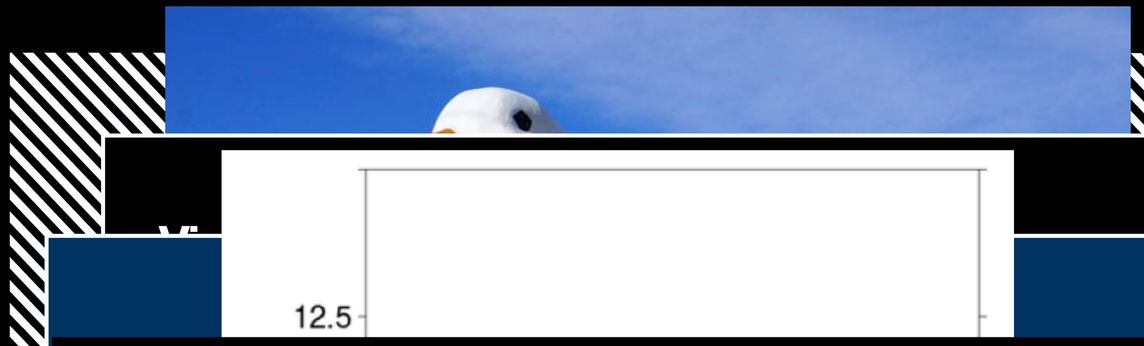
Class 13
Guidelines for Effective Visualizations

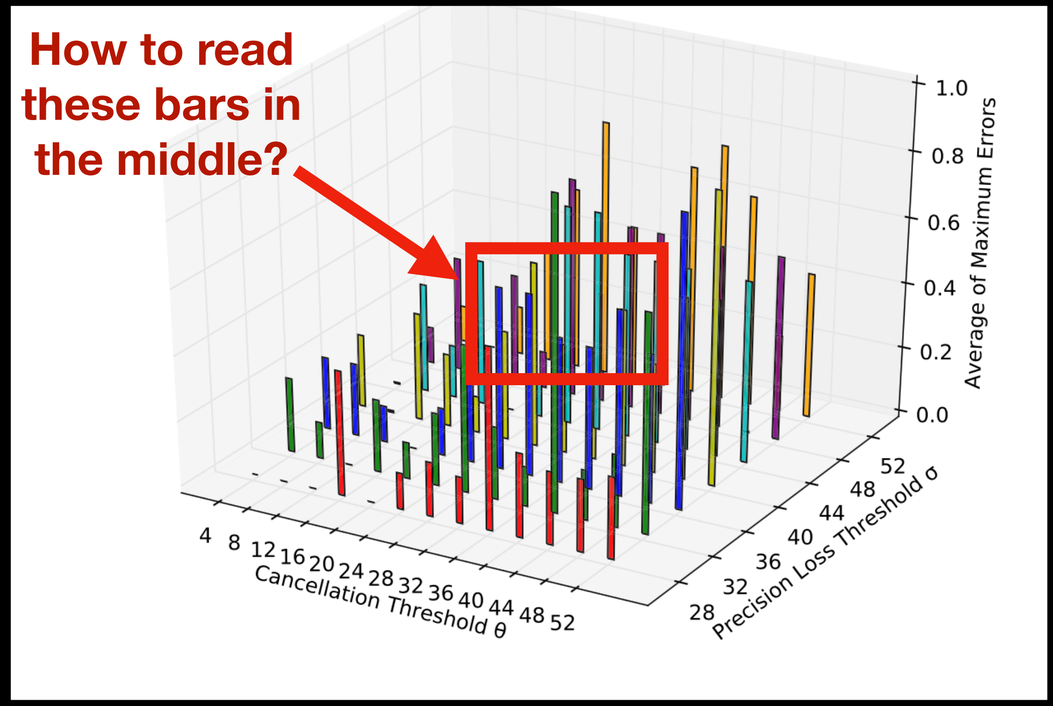
Luís Cruz - l.cruz@tudelft.nl



Recap

- Chartjunk
- Graphical Duck
- Principles for Effective Communication
- Integrity
- Lie Factor
- Graphs at ICSE





Practical Guidelines for Effective Visualizations

(that work most of the time)

Practical Guidelines for Effective Visualizations

Part 1

- Choose a graph type that fits best your case.
  Chart type selection:
  http://experception.net/
  Franconeri_ExperCeptionDotNet_DataVisQuickRef.pdf
- Graphs should have a clear, **self-explanatory** title or caption.
- State the **units of measurement**.
- All axes should be carefully **labelled**.
- Use **two-dimensional** designs.

Practical Guidelines for Effective Visualizations

Part 2

- **Turn off the box around the figure.**
  Same for boxes that are collinear with the axis.
- Only have one x- and one y-axis.
- Use visual variables (color, shape, shade) only for data variation.
- Use diverging color schemes for data that progresses outward from a middle value. E.g., positive and negative values.
- Care for colorblindness. Test your images with tools like Coblis: https://www.color-blindness.com/coblis-color-blindness-simulator/.

Practical Guidelines for Effective Visualizations

Part 3

- Axis must start at a meaningful baseline. ➡ **Bar charts must start at zero.**
- **Never use different colors to represent the same kind of data.**
- Label elements directly, avoiding indirect look-up. E.g., use a legend only when space is tight.
- **Text label should never be rotated** (nor vertical). E.g., rotate bar chart when category names are too long.
- Use bold type/lines only to emphasize something.

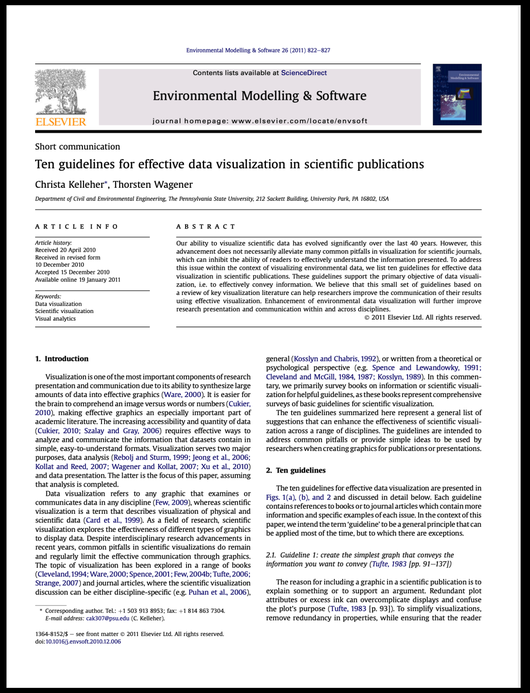
Practical Guidelines for Effective Visualizations

Part 4

- Avoid pairing **green and red together** and opt for color blind-safe paletes.
- Highlight what’s important, tell one story.
- Sort data for easier comparisons. E.g., in a pie chart or bar chart.
- Don’t use 3D effects.
- **Avoid pie charts** (and donut charts). It is difficult to compare many slices in a pie chart. (Very simple charts are exception)

More guidelines

- Material Design Docs. https://material.io/design/communication/data-visualization.html
- Christa Kelleher, Thorsten Wagener (2010). Ten guidelines for effective data visualization in scientific publications. https://doi.org/10.1016/j.envsoft.2010.12.006



Environmental Modelling & Software 26 (2011) 822–827

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Short communication

Ten guidelines for effective data visualization in scientific publications

Christa Kelleher*, Thorsten Wagener

Department of Civil and Environmental Engineering, The Pennsylvania State University, 212 Sackett Building, University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 20 April 2010
Received in revised form
10 December 2010
Accepted 15 December 2010
Available online 19 January 2011

Keywords:
Data visualization
Scientific visualization
Visual analytics

ABSTRACT

Our ability to visualize scientific data has evolved significantly over the last 40 years. However, this advancement does not necessarily alleviate many common pitfalls in visualization for scientific journals, which can inhibit the ability of readers to effectively understand the information presented. To address this issue within the context of visualizing environmental data, we list ten guidelines for effective data visualization in scientific publications. These guidelines support the primary objective of data visualization, i.e. to effectively convey information. We believe that this small set of guidelines based on a review of key visualization literature can help researchers improve the communication of their results using effective visualization. Enhancement of environmental data visualization will further improve research presentation and communication within and across disciplines.

1. Introduction

Visualization is one of the most important components of research presentation and communication due to its ability to synthesize large amounts of data into effective graphics (Ware, 2000). It is easier for the brain to comprehend an image versus words or numbers (Cukier, 2010), making effective graphics an especially important part of academic literature. The increasing accessibility and quantity of data (Cukier, 2010; Szalay and Gray, 2006) requires effective ways to analyze and communicate the information that datasets contain in simple, easy-to-understand formats. Visualization serves two major purposes, data analysis (Reboj and Sturm, 1999; Jeong et al., 2006; Kollat and Reed, 2007; Wagener and Kollat, 2007; Xu et al., 2010) and data presentation. The latter is the focus of this paper, assuming that analysis is completed.

Data visualization refers to any graphic that examines or communicates data in any discipline (Few, 2009), whereas scientific visualization is a term that describes visualization of physical and scientific data (Card et al., 1999). As a field of research, scientific visualization explores the effectiveness of different types of graphics to display data. Despite interdisciplinary research advancements in recent years, common pitfalls in scientific visualizations do remain and regularly limit the effective communication through graphics. The topic of visualization has been explored in a range of books (Cleveland, 1994; Ware, 2000; Spence, 2001; Few, 2004b; Tufte, 2006; Strange, 2007) and journal articles, where the scientific visualization discussion can be either discipline-specific (e.g. Puhan et al., 2006), general (Kosslyn and Chabris, 1992), or written from a theoretical or psychological perspective (e.g. Spence and Lewandowky, 1991; Cleveland and McGill, 1984, 1987; Kosslyn, 1989). In this commentary, we primarily survey books on information or scientific visualization for helpful guidelines, as these books represent comprehensive surveys of basic guidelines for scientific visualization.

The ten guidelines summarized here represent a general list of suggestions that can enhance the effectiveness of scientific visualization across a range of disciplines. The guidelines are intended to address common pitfalls or provide simple ideas to be used by researchers when creating graphics for publications or presentations.

2. Ten guidelines

The ten guidelines for effective data visualization are presented in Figs. 1(a), (b), and 2 and discussed in detail below. Each guideline contains references to books or to journal articles which contain more information and specific examples of each issue. In the context of this paper, we intend the term 'guideline' to be a general principle that can be applied most of the time, but to which there are exceptions.

2.1. Guideline 1: create the simplest graph that conveys the information you want to convey (Tufte, 1983 [pp. 91–137])

The reason for including a graphic in a scientific publication is to explain something or to support an argument. Redundant plot attributes or excess ink can overcomplicate displays and confuse the plot's purpose (Tufte, 1983 [p. 93]). To simplify visualizations, remove redundancy in properties, while ensuring that the reader

* Corresponding author. Tel.: +1 503 913 8953; fax: +1 814 863 7304.
E-mail address: cak307@psu.edu (C. Kelleher).

1364-8152/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2010.12.006

Project 1

- Deadline is in the end of today (Nov 10, 23:59 GMT+8).
- So far the submitted projects tend to be technically good.
- There are still a few things that need improvement.

Points you need to consider

Project 1

- **Creative plot.**
  - The creative plot needs to have a **purpose — convey information**. Some of the creative plots I have seen have added visual variables that are merely **decorative**.
  - Be careful when enforcing the usage of colors. Users may want to print the creative plots or use them with different background colors.
  - You should avoid doing things against the principles studied in the class.
  - Some groups are perceiving creativity as chart junk. Please avoid creating chart junk.

Points you need to consider

Project 1

- **Report.**
  - The report should be **self-contained**. It should explain the whole work for someone that does not know the assignment.
  - It should have a **narrative**.
    Tip: imagine that you are writing a blog post about your project.
  - Use figures to help the communication, but explain what has been done there.
  - **Important!** Whenever possible, you should justify your decisions in the project with the theory given in the class.
  - If you had to come up with a solution for a complex problem, tell that story.
  - (This is not mandatory). It is a good practice to write a **summary** in the beginning of the report. It should briefly describe what your project is about. It should consist of one paragraph with less than 120 words.