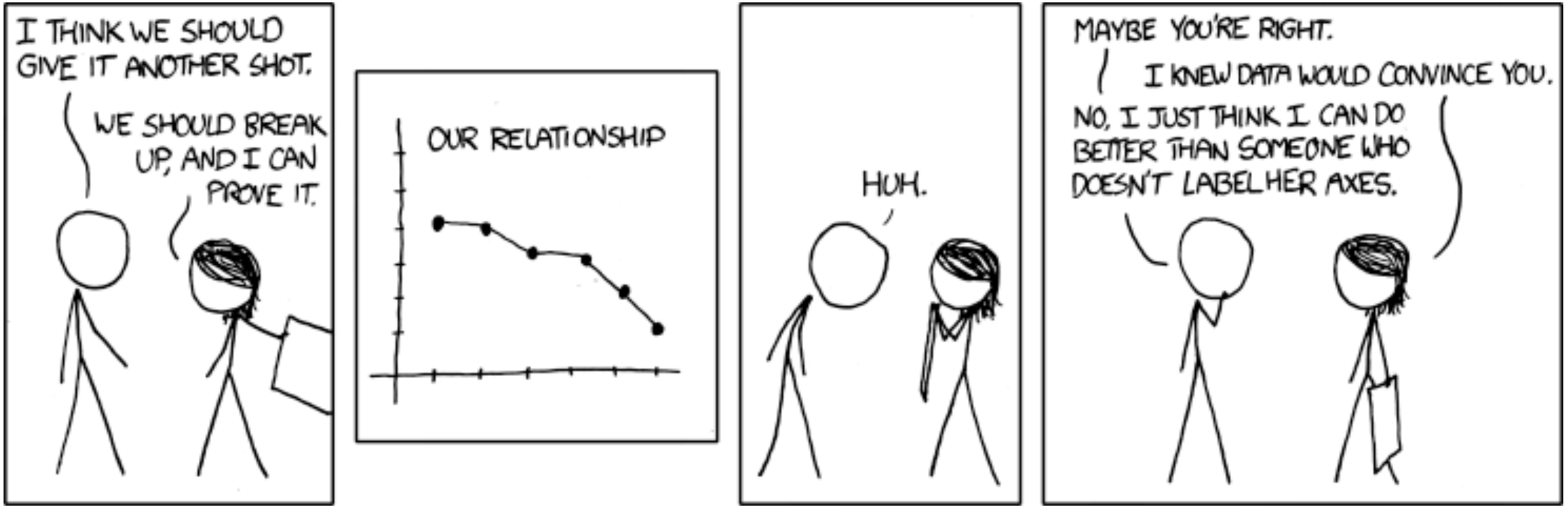
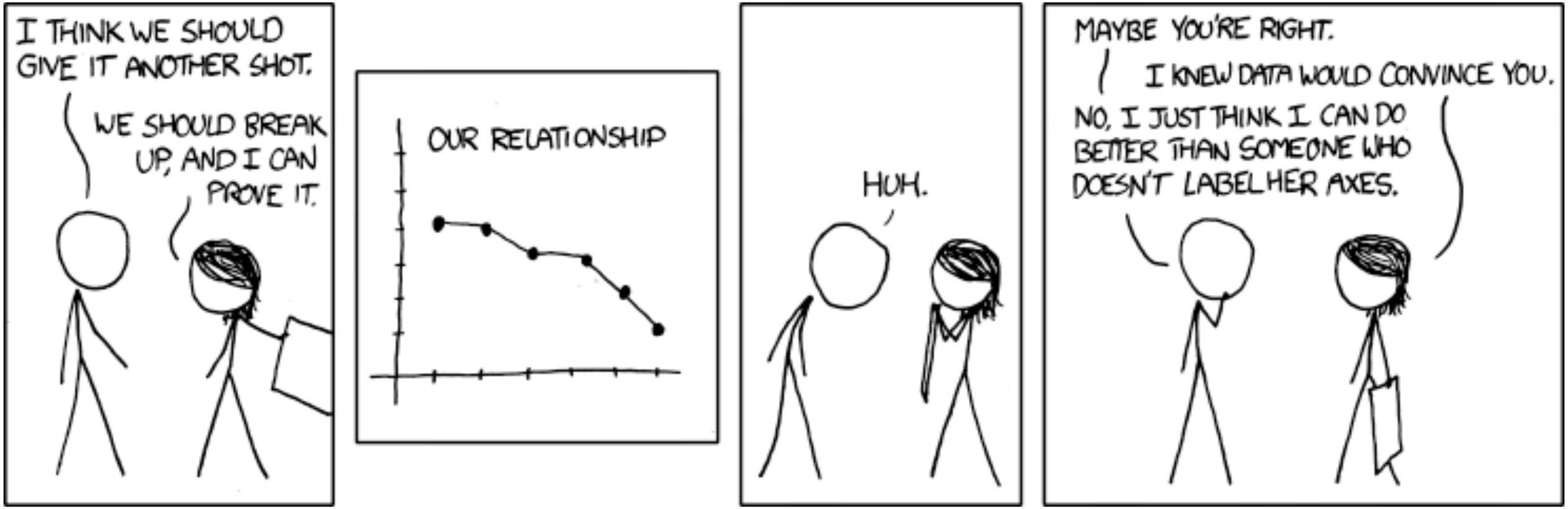


Source: xkcd.com/833



Tooltip: "And if you labeled your axes, I could tell you exactly how MUCH better!"

Source: xkcd.com/833

Exploratory Data Analysis

January 18, 2021

DATA 550 - Lecture 3

20 minutes

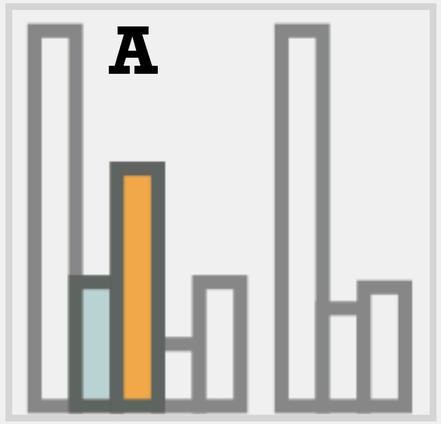
Part 1: Formalizing choice of visualization

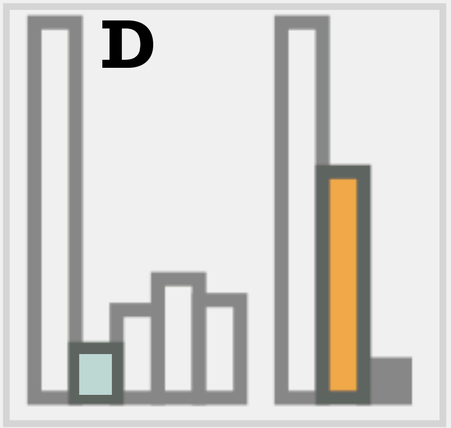
Choosing a Visualization

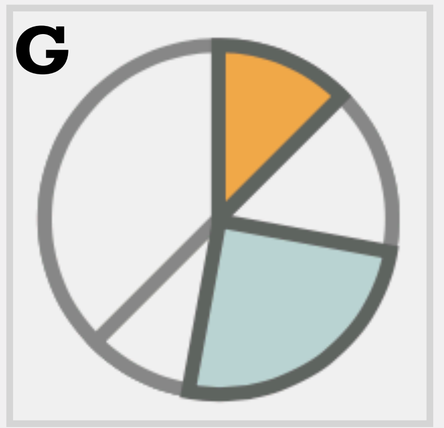
Principle	Definition	Examples
• Visualization choice	Must be informed by the **data** you have, the **research question** being asked and the **audience** that cares	- Summary statistics >> cannot show distributions - Pick the simplest plot that best shows most/all of the data needed to answer the research question - Tailor the visualization to your audience (within reason)

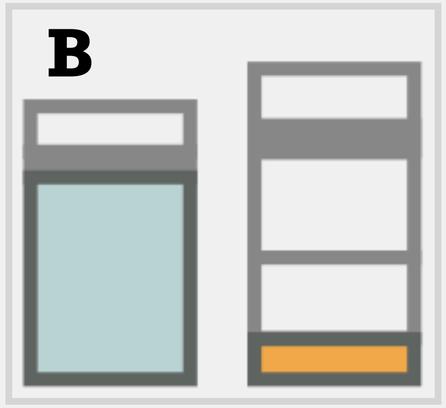
Rank the 9 plots for their effectiveness in answering:

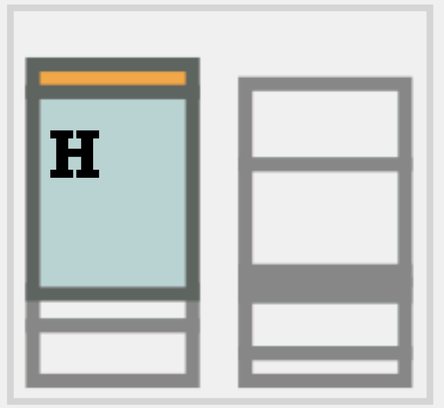
What is the difference between the orange and green quantities?

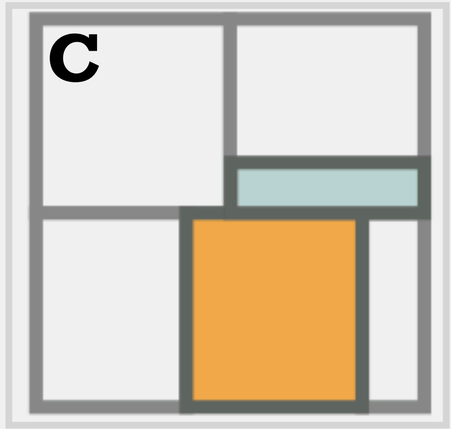


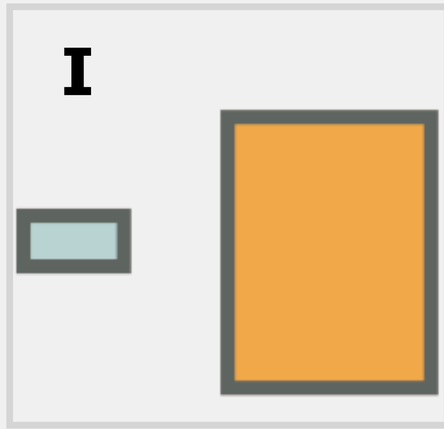




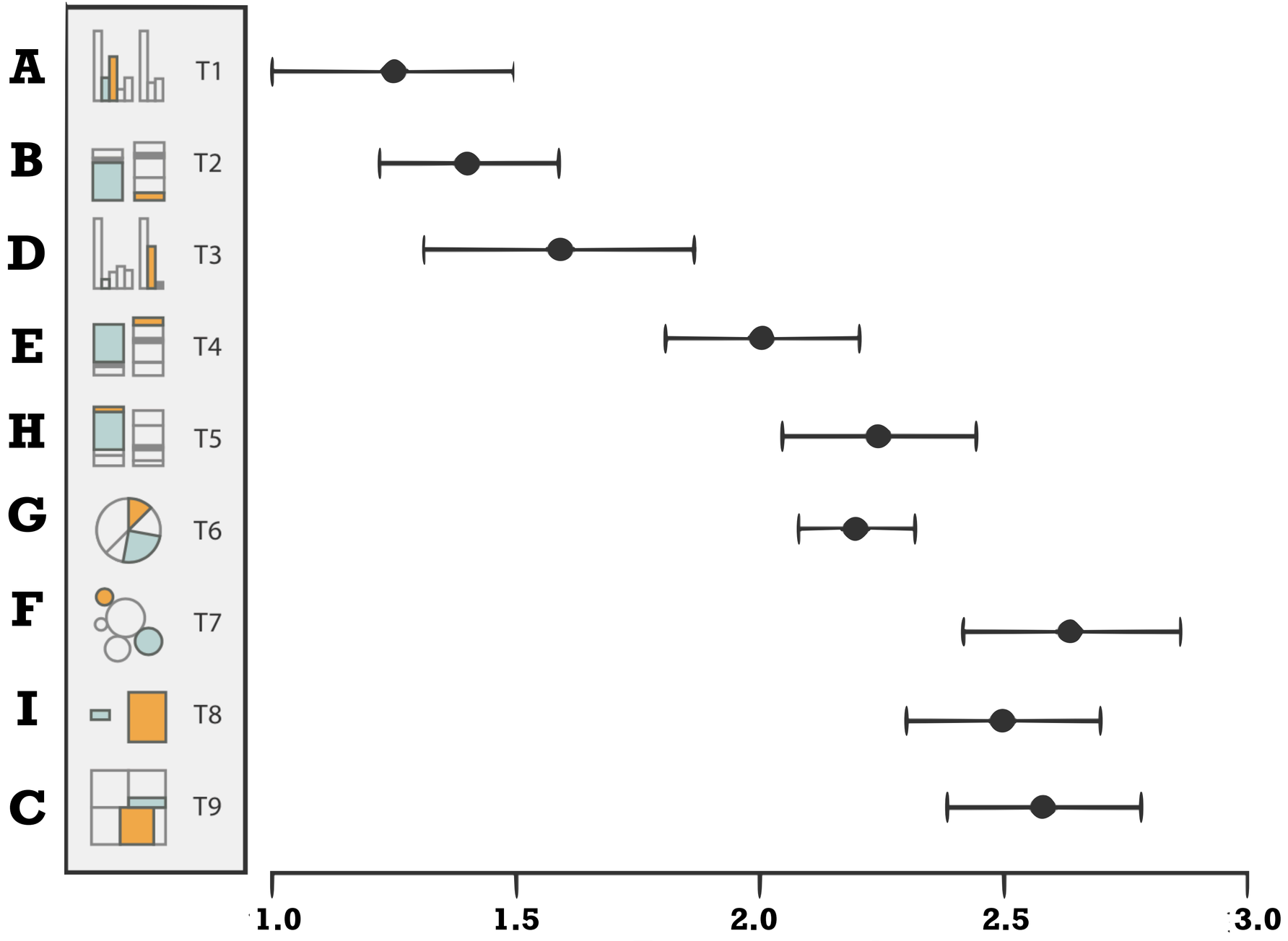




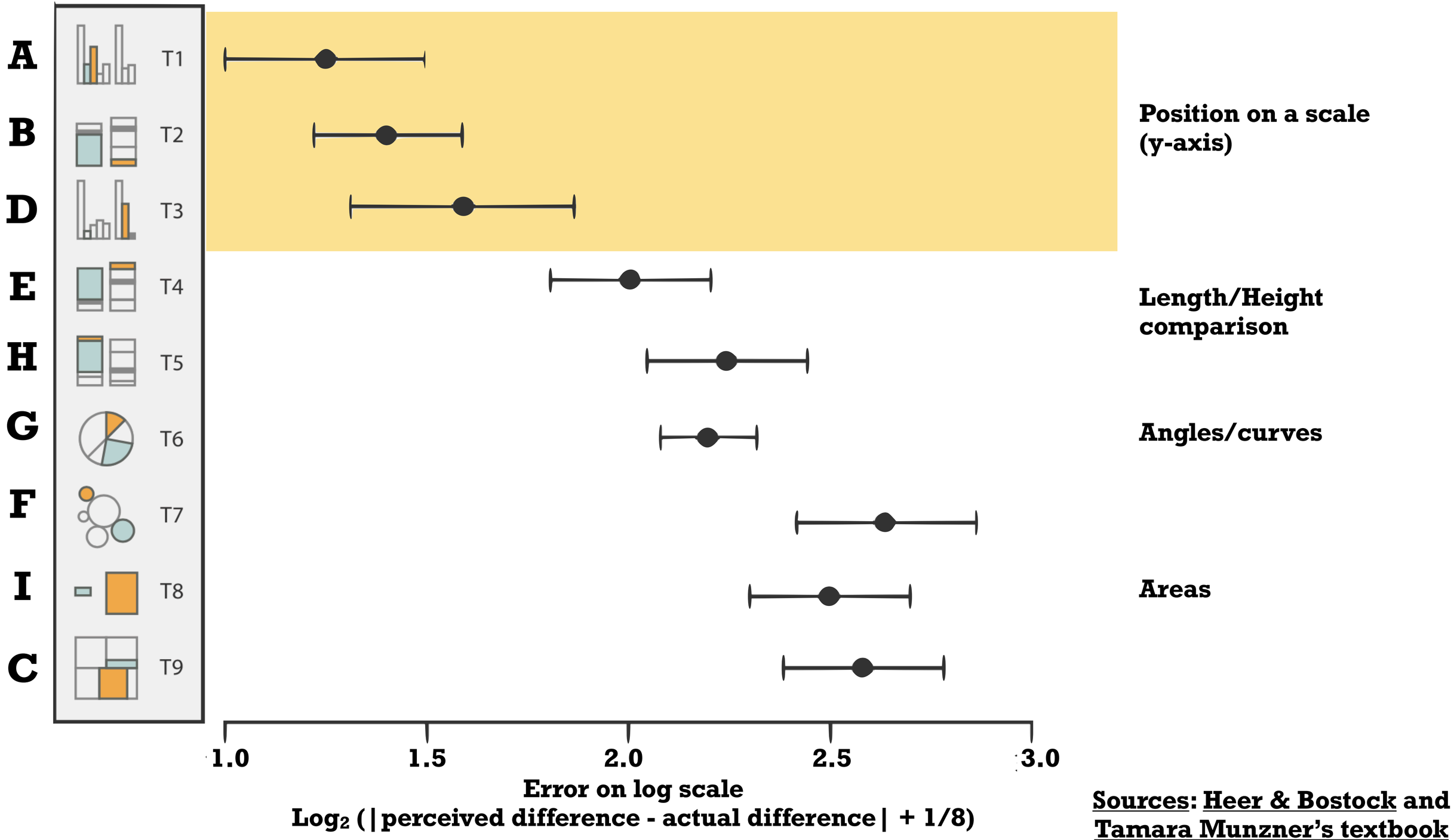




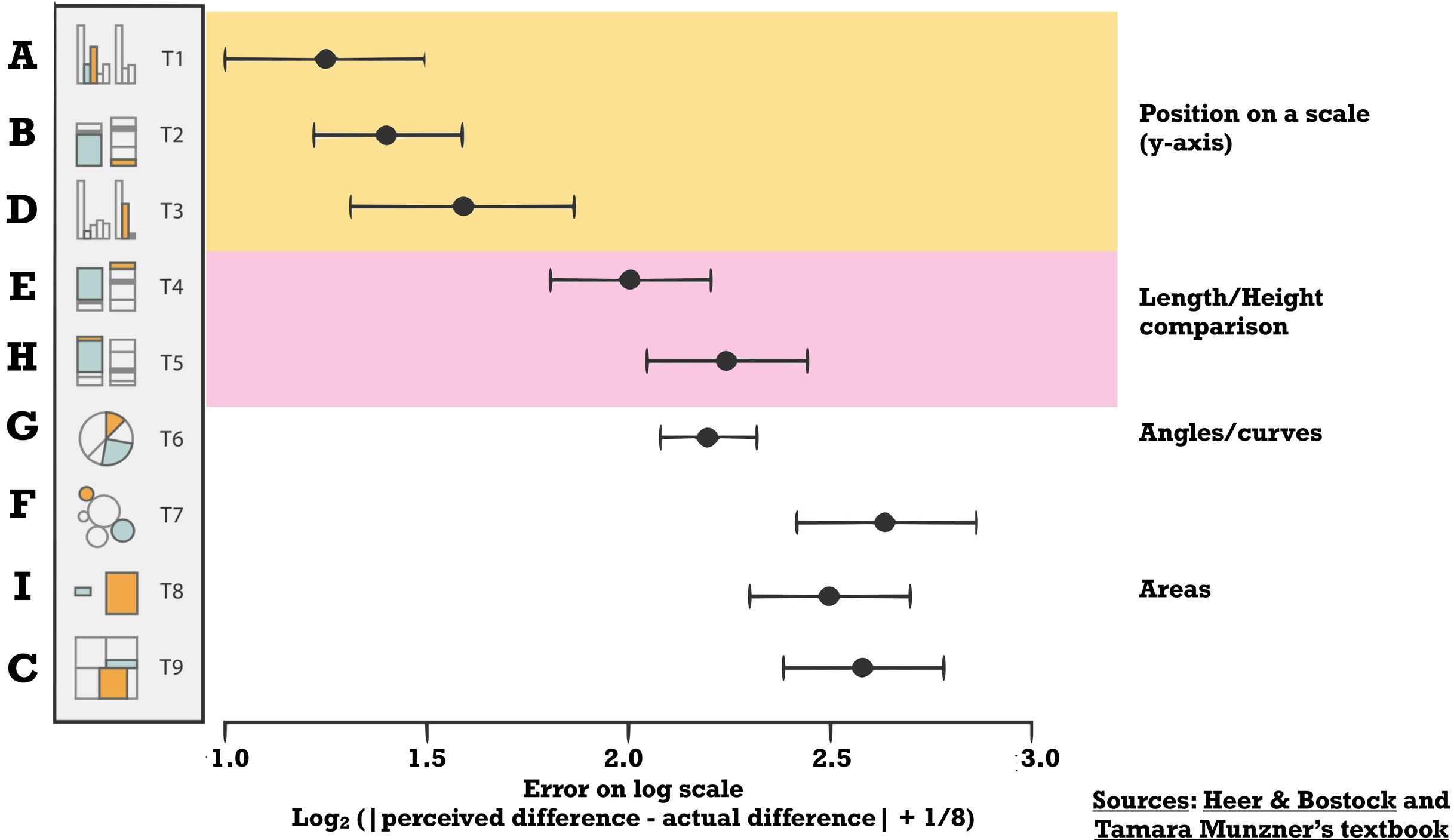
Sources: Heer & Bostock, Cleveland & McGill, and Tamara Munzner's textbook



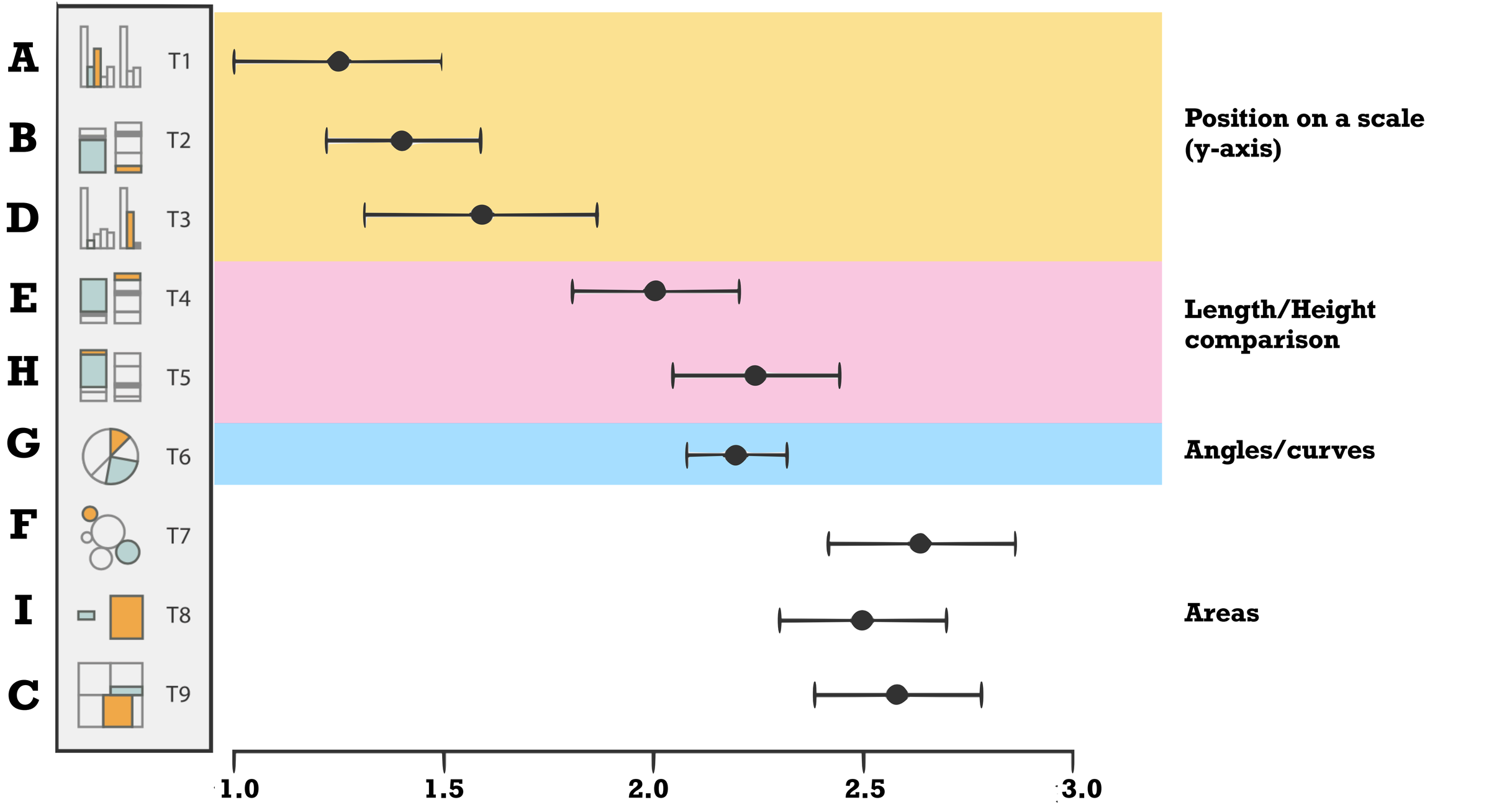
Sources: Heer & Bostock and Tamara Munzner's textbook



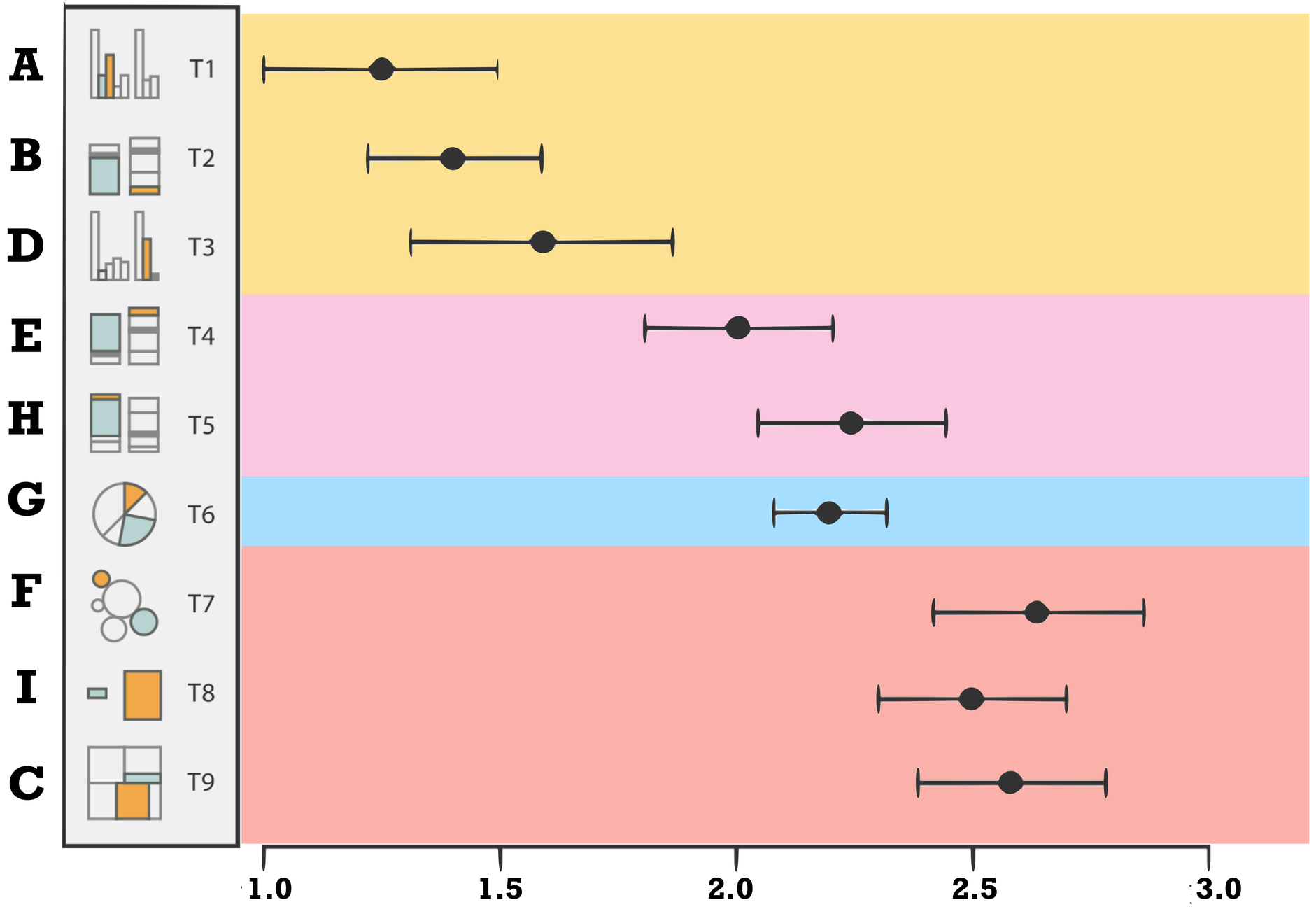
Sources: Heer & Bostock and Tamara Munzner's textbook



Sources: Heer & Bostock and Tamara Munzner's textbook



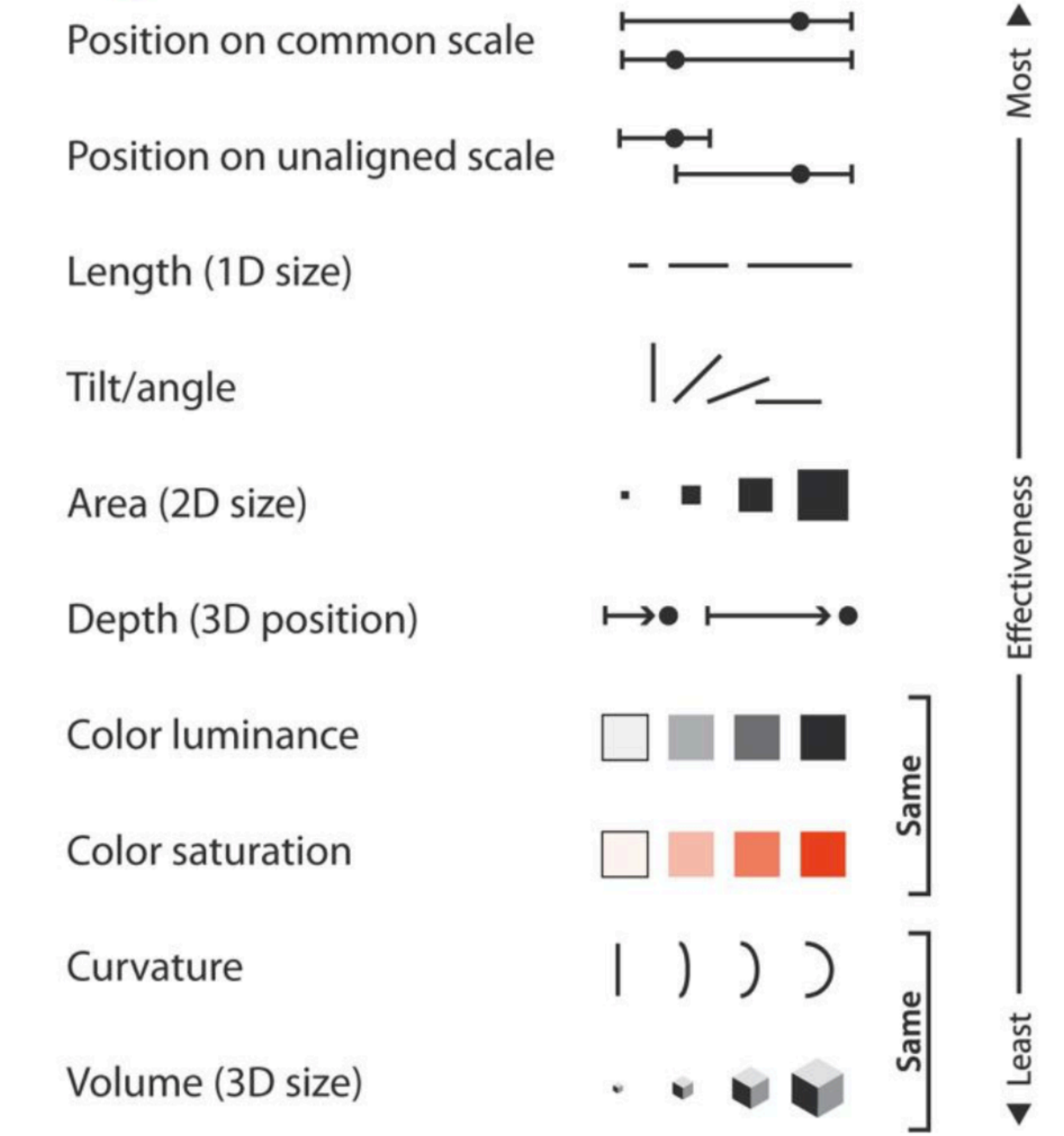
Sources: Heer & Bostock and Tamara Munzner's textbook



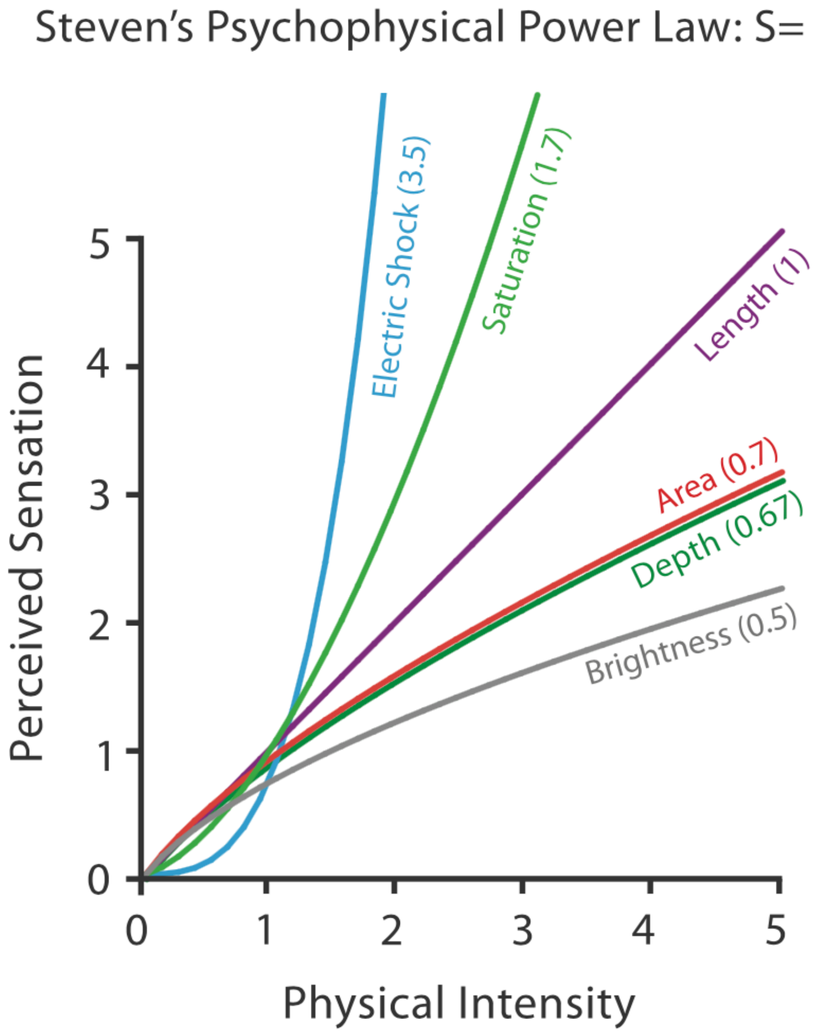
Sources: Heer & Bostock and Tamara Munzner's textbook

Channels: Expressiveness Types and Effectiveness Ranks

Magnitude Channels: Ordered Attributes



Sources: Chapter 5 of Tamara Munzner's textbook



Sources: Chapter 5 of Tamara Munzner's textbook and originally: "On the Psychophysical Law." , Psychological Review 64:3 (1957)

Figure 5.7. Stevens showed that the apparent magnitude of all sensory channels follows a power law $S = I^n$, where some sensations are perceptually magnified compared with their objective intensity (when $n > 1$) and some compressed (when $n < 1$). Length perception is completely accurate, whereas area is compressed and saturation is magnified. Data from Stevens [Stevens 75, p. 15].

30 minutes

Part 2:
Case Study - Planes in WW2

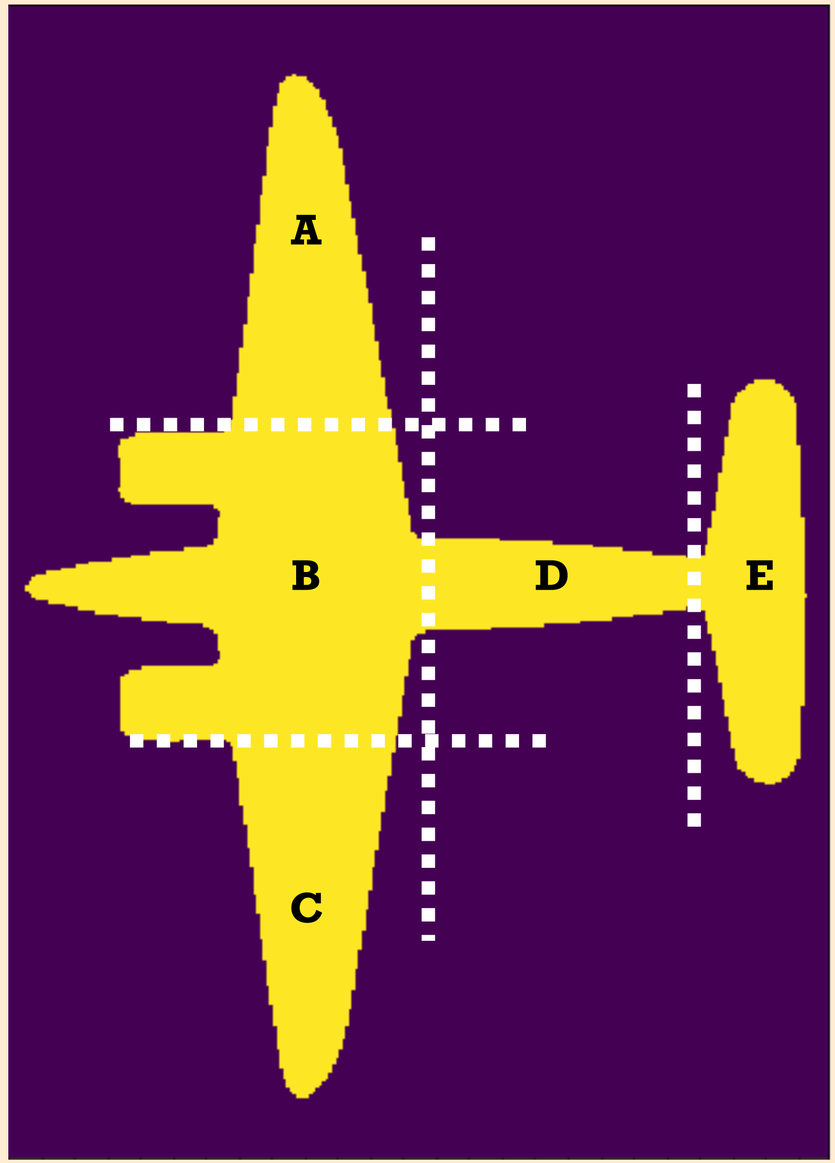
Case Study: Planes in WW2

You have been given a dataset and tasked with trying to solve a problem. In WW2, **expensive fighter planes were going down quite frequently due to bullet fire.** The military decided to conduct an analysis and surveyed all the surviving planes in an effort to catalogue which regions of the plane should be reinforced.

With limited resources, the military could only reinforce a maximum of two zones. Your task is to look at the bullet data for the planes and help determine which areas of the plane should be reinforced.

You're given a schematic of the plane, and told that the workers added a grid to the schematic, divided it up into regions A,B,C,D,E and recorded a value of 1 wherever there was a bullet hole across all the planes that returned. Areas without bullet holes are marked as 0.

They gave you a csv file with this information called `bullet_data.csv`. Yes, these WW2 workers are very sophisticated and had access to a computer :-).



Case Study: Planes in WW2

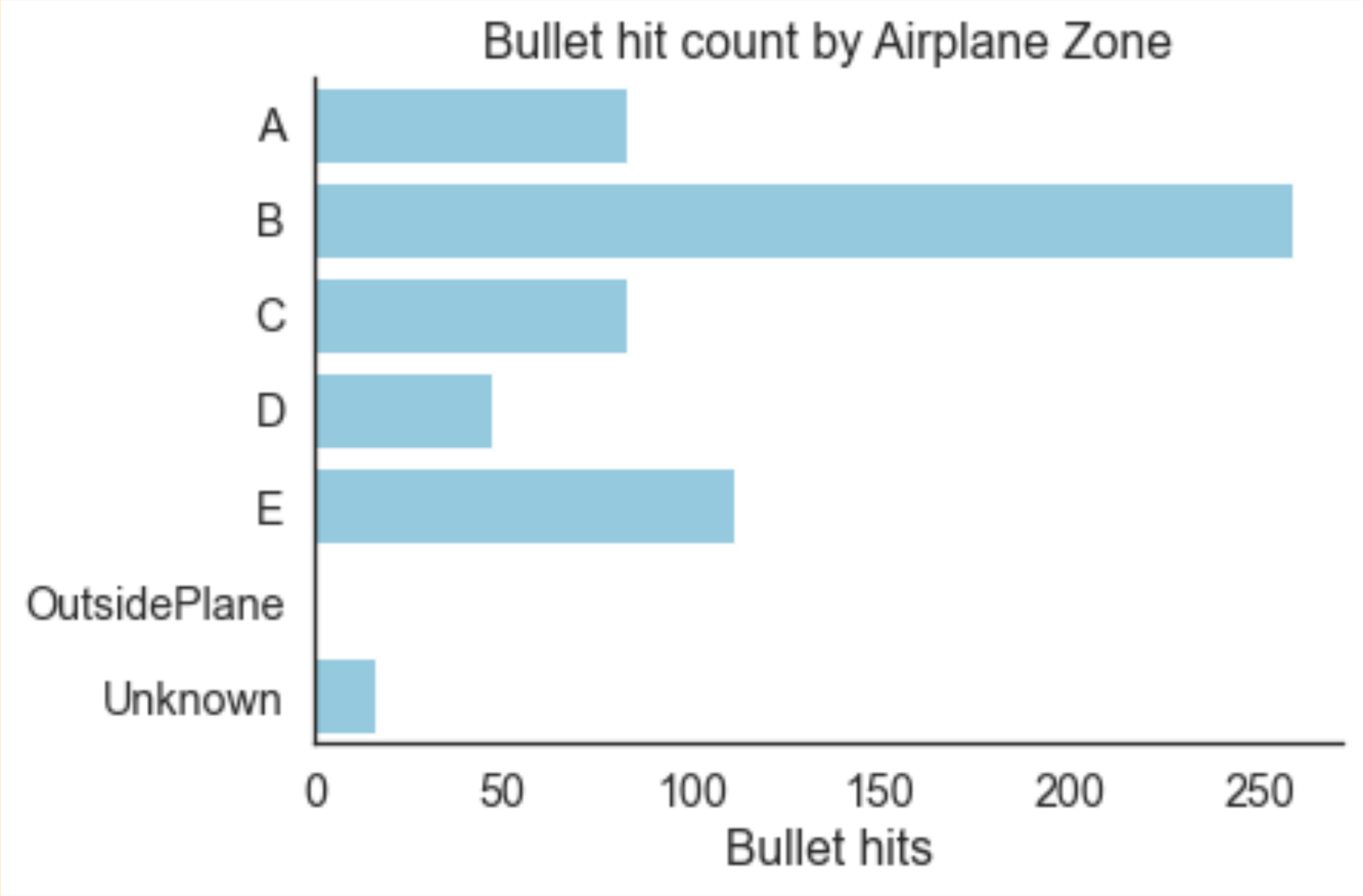
You have been given a dataset and tasked with trying to solve a problem. In WW2, **expensive fighter planes were going down quite frequently due to bullet fire.** The military decided to conduct an analysis and surveyed all the surviving planes in an effort to catalogue which regions of the plane should be reinforced.

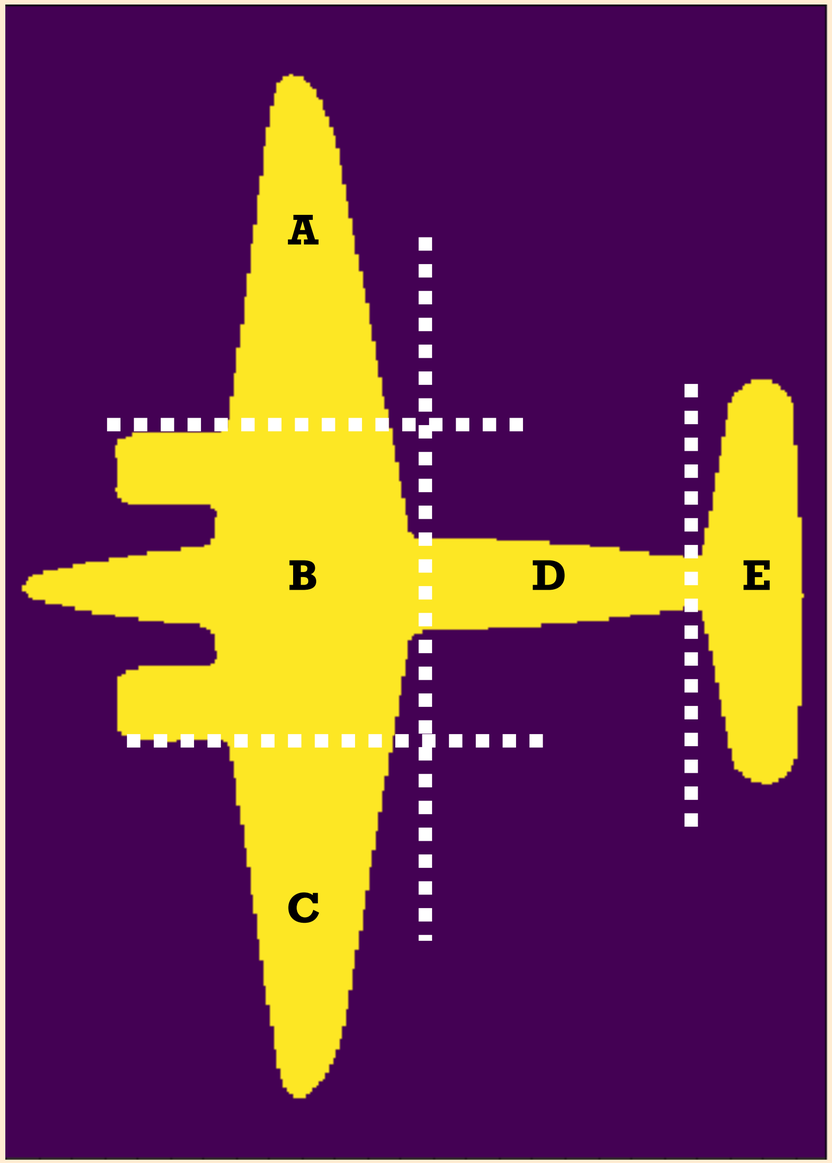
With limited resources, the military could only reinforce a maximum of two zones. Your task is to look at the bullet data for the planes and help determine which areas of the plane should be reinforced.

You're given a schematic of the plane, and told that the workers added a grid to the schematic, divided it up into regions A,B,C,D,E and recorded a value of 1 wherever there was a bullet hole across all the planes that returned. Areas without bullet holes are marked as 0.

They gave you a csv file with this information called `bullet_data.csv`. Yes, these WW2 workers are very sophisticated and had access to a computer :-).

Case Study: Planes in WW2





Debrief - EDA is important!

- Look at your data.
- Talk to someone about your data.
- Look at your data another way.
- Think about your data and what it means!

5 min Break

20 minutes

Part 3:
Exploratory Data Analysis

7 Exploratory Data Analysis

7.1 Introduction

This chapter will show you how to use visualisation and transformation to explore your data in a systematic way, a task that statisticians call exploratory data analysis, or EDA for short. EDA is an iterative cycle. You:

1. Generate questions about your data.

2. Search for answers by visualising, transforming, and modelling your data.

3. Use what you learn to refine your questions and/or generate new questions.

EDA is not a formal process with a strict set of rules. More than anything, EDA is a state of mind. During the initial phases of EDA you should feel free to investigate every idea that occurs to you. Some of these ideas will pan out, and some will be dead ends. As your exploration continues, you will home in on a few particularly productive areas that you'll eventually write up and communicate to others.

Types of Research Questions

1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

Research Questions

1. Descriptive

- one that seeks to summarize a characteristic of a set of data
- no interpretation of the result itself as the result is a fact, an attribute of the data set you are working with

- e.g., What is the frequency of viral illnesses in a set of data collected from a group of individuals?
- e.g., How many people live in each US state?

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

2. Exploratory

- one in which you analyze the data to see if there are patterns, trends, or relationships between variables
- looking for patterns that would support proposing a hypothesis to test in a future study

- e.g., Do diets rich in certain foods have differing frequencies of viral illnesses in a set of data collected from a group of individuals?
- e.g., Does air pollution correlate with life expectancy in a set of data collected from groups of individuals from several regions in the United States?

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

3. Inferential

- **one in which you analyze the data to see if there are patterns, trends, or relationships between variables in a representative sample**
- **want to quantify how much the patterns, trends, or relationships between variables is applicable to all individuals units in the population**

- **e.g., Is eating at least 5 servings a day of fresh fruit and vegetables is associated with fewer viral illnesses per year?**
- **e.g., Does air pollution correlate with life expectancy in the United States?**

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

4. Predictive

- **one where you are trying to predict measurements or labels for individuals (people or things)**
- **less interested in what causes the predicted outcome, just what predicts it**

- **e.g., How many viral illnesses will someone have next year?**
- **e.g., What political party will someone vote for in the next US election?**

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

5. Causal

- asks about whether changing one factor will change another factor, on average, in a population.
- Sometimes the underlying design of the data collection, by default, allows for the question that you ask to be causal (e.g., randomized experiment or trial)

- e.g., Does eating at least 5 servings a day of fresh fruit and vegetables cause fewer viral illnesses per year?
- e.g., Does smoking cause cancer?

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

6. Mechanistic

- **one that tries to explain the underlying mechanism of the observed patterns, trends, or relationship (how does it happen?)**

- **e.g., How do changes in diet lead to a reduction in the number of viral illnesses?**
- **e.g., How does airplane wing design changes air flow over a wing, leading to decreased drag?**

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Types of Research Questions

Exploratory Data Analysis

1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

Steps of EDA

1. Describe your dataset

rubric:{reasoning:4}

Task: Describe your dataset. Consider the following questions to guide you in your exploration

- Who: Which company/agency/organization provided this data?
- What: What is in your data?
- When: When was your data collected (for example, for which years)?
- Why: What is the purpose of your dataset? Is it for transparency/accountability, public interest, fun, learning, etc...
- How: How was your data collected? Was it a human collecting the data? Historical records digitized? Server logs?

Steps of EDA

2. Load the dataset

rubric:{correctness:1}

Task: Load your dataset from a file, or URL. This needs to be a pandas dataframe so you can use it with Altair. Remember that others may be running your jupyter notebook so it's important that the data is accessible to them. If your dataset isn't accessible as a URL, make sure to commit it into your repo.

Steps of EDA

3. Explore your dataset

rubric:{correctness:5}

Task: Explore the columns in your dataset. Which ones are interesting/relevant? You can use the same scheme I outlined in Lab 3, Exercise 2 or come up with your own system. By now, you should also know about df.describe() so you can use that as an aid if you think it's useful and appropriate.

Steps of EDA

4. Initial thoughts

rubric={correctness:1}

Task: Use this a place to record any observations you come up with, anything jump out at you as surprising or particularly interesting? Where do you think you'll go with exploring this dataset? Feel free to take notes in this section and use it as a scratch pad. Any content in this area will not be marked.

Steps of EDA

5. Wrangling

rubric={correctness:1}

Task: You can do any wrangling you need to do here. If you prefer to wrangle in R, that's fine - go ahead and wrangle the data in a different notebook, then export the data as a CSV and then load it in again as a new pandas dataframe. Describe what you're doing (or did) using comments within your code.

Steps of EDA

6. Research questions

rubric={reasoning:5}

Task: come up with at least two research questions about your dataset that will require data visualizations to help answer. Recall that for this purpose, you should only aim for "Descriptive" or "Exploratory" research questions.

Steps of EDA

7. Data Analysis & Visualizations

rubric={viz:40, reasoning:10}

Task: Create data visualizations (and justify your choices) using Altair that will help you answer your research questions.

Steps of EDA

8. Summary and conclusions

rubric={reasoning:10}

Task: Summarize your findings and describe any conclusions and insight you were able to draw from your visualizations.