Data Visualization:

A Practical Guide

Danielle Presgraves, Ph.D.

Learning Objectives:

1. Classify the variable type (categorical, ordinal, numeric, etc.)

2. Identify the appropriate visual representation based on variable and question

3. Simplify the visualization elements to reduce cognitive load on your audience

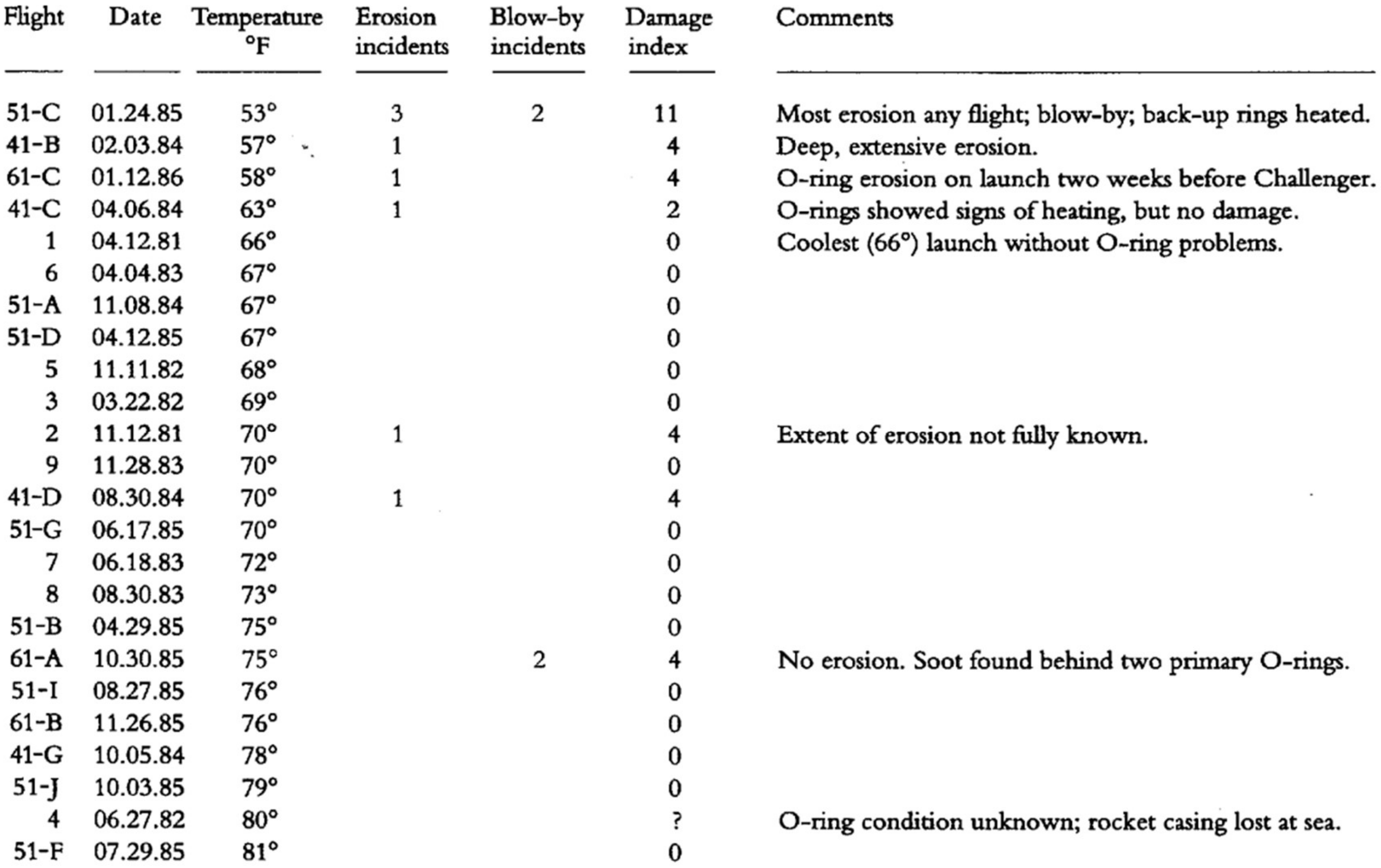
4. Prioritize accuracy, ethics, and honest communication

5. Integrate storytelling to emphasize data understanding, not just presentation

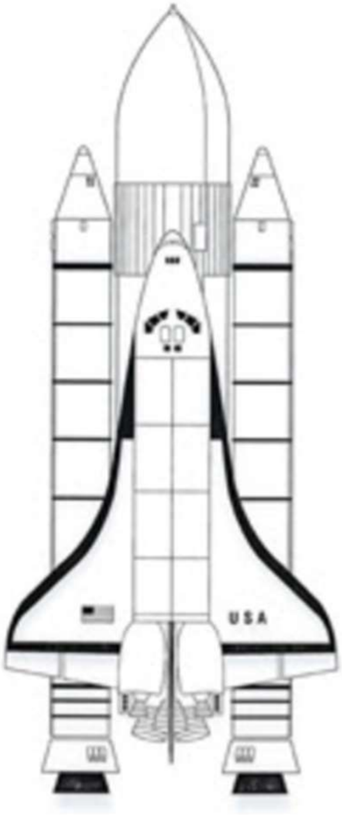
Agenda:

1. A famous data visualization “Missed Opportunity”
2. A tangent about historic visualization visionaries
3. Technical Aspects
4. Artistic & human factors choices
5. Narratives: The Hero’s Journey

The Challenger pre-launch presentation from concerned engineers had 13 charts, including:



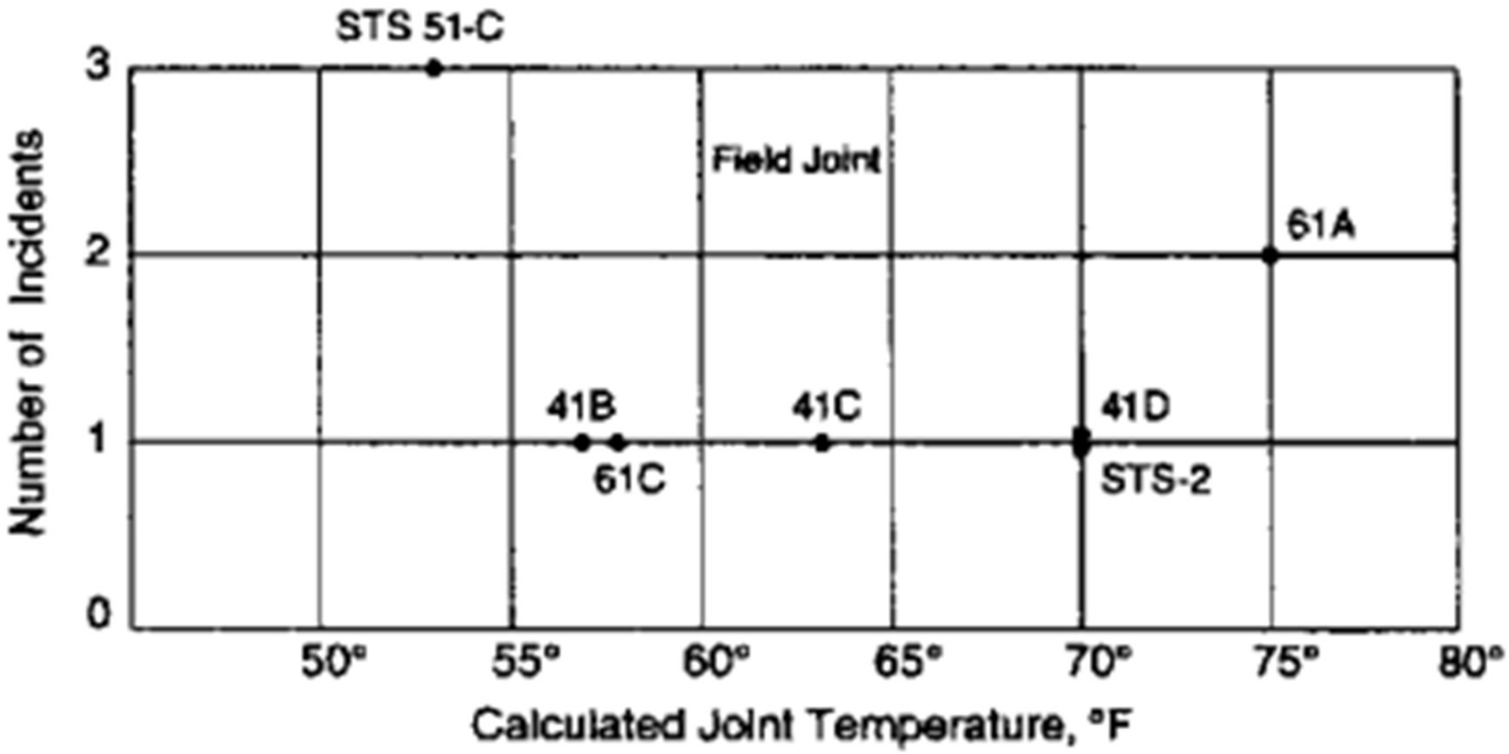
Flight	Date	Temperature °F	Erosion incidents	Blow-by incidents	Damage index	Comments
51-C	01.24.85	53°	3	2	11	Most erosion any flight; blow-by; back-up rings heated.
41-B	02.03.84	57°	1		4	Deep, extensive erosion.
61-C	01.12.86	58°	1		4	O-ring erosion on launch two weeks before Challenger.
41-C	04.06.84	63°	1		2	O-rings showed signs of heating, but no damage.
1	04.12.81	66°			0	Coolest (66°) launch without O-ring problems.
6	04.04.83	67°			0	
51-A	11.08.84	67°			0	
51-D	04.12.85	67°			0	
5	11.11.82	68°			0	
3	03.22.82	69°			0	
2	11.12.81	70°	1		4	Extent of erosion not fully known.
9	11.28.83	70°			0	
41-D	08.30.84	70°	1		4	
51-G	06.17.85	70°			0	
7	06.18.83	72°			0	
8	08.30.83	73°			0	
51-B	04.29.85	75°			0	
61-A	10.30.85	75°		2	4	No erosion. Soot found behind two primary O-rings.
51-I	08.27.85	76°			0	
61-B	11.26.85	76°			0	
41-G	10.05.84	78°			0	
51-J	10.03.85	79°			0	
4	06.27.82	80°			?	O-ring condition unknown; rocket casing lost at sea.
51-F	07.29.85	81°			0	



Summary of 13 charts

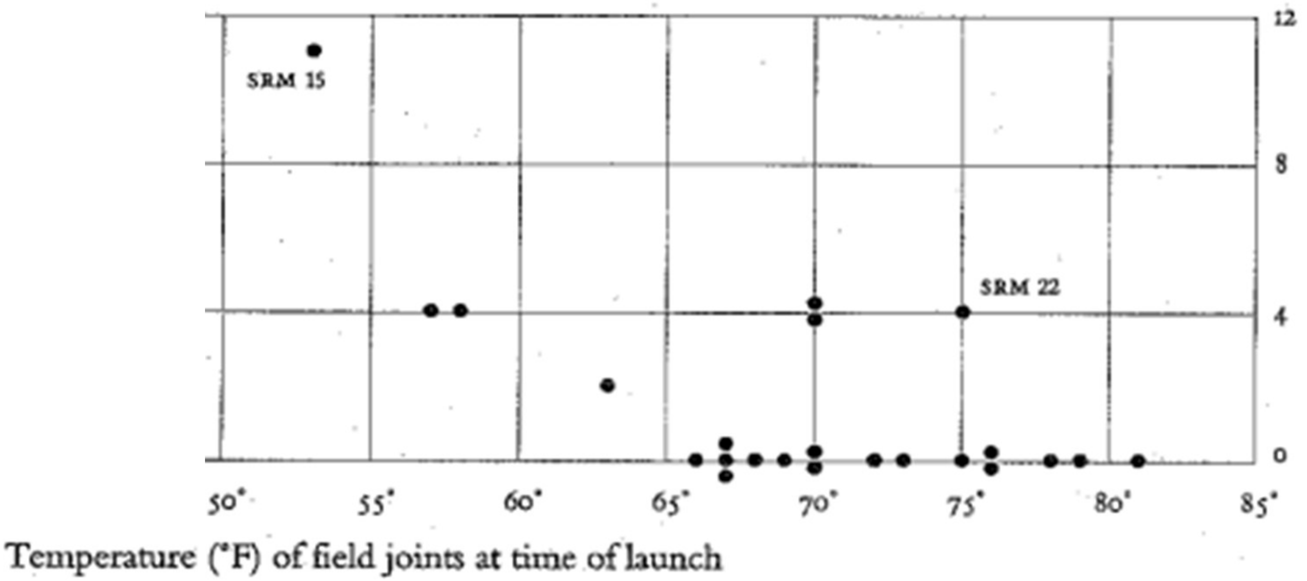
Summary of Tufte's Challenger Argument

Post-launch diagram*



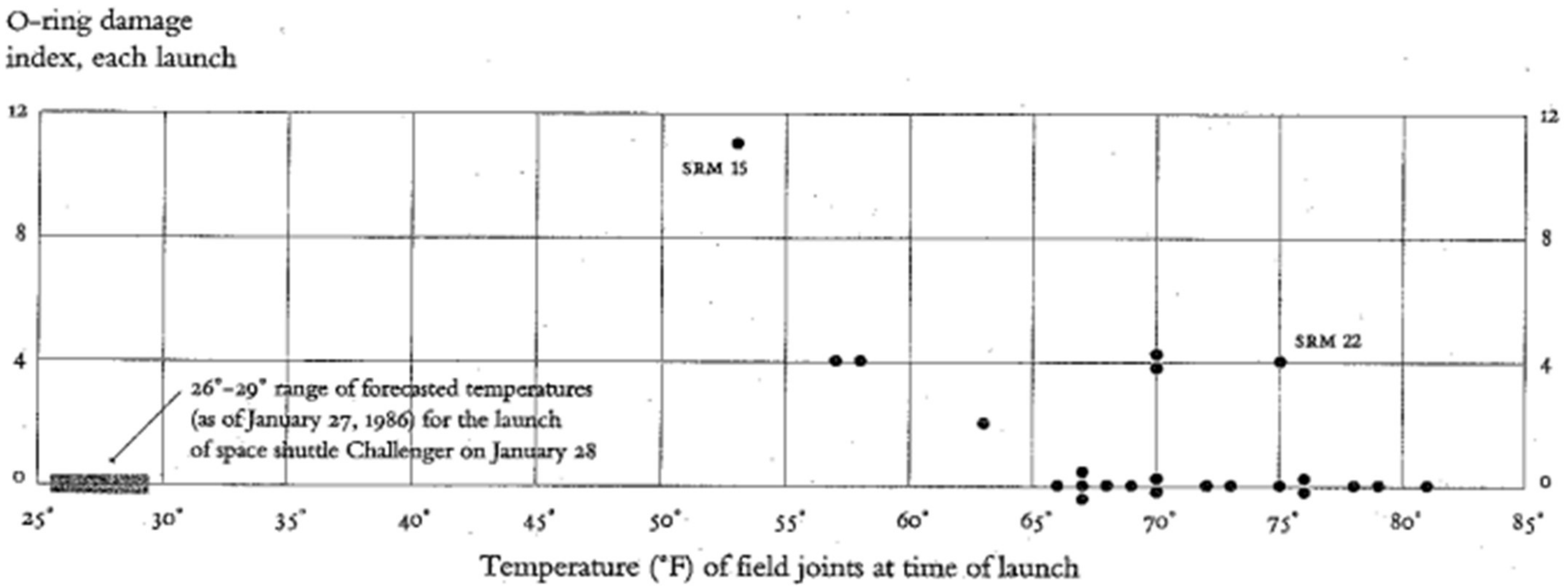
* drawn after incident by a lawyer and executive director

Post-launch Redrawn by Tufte

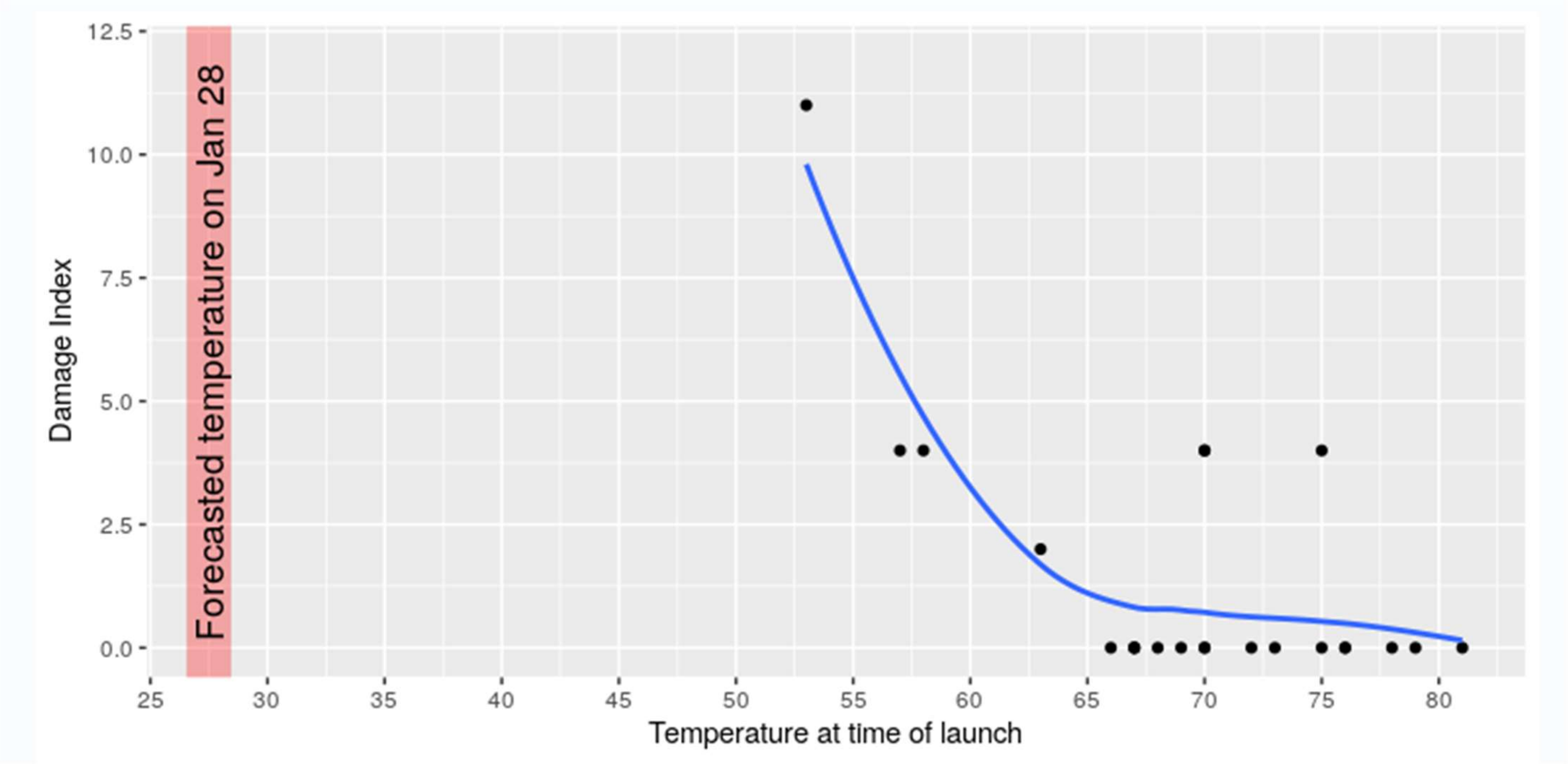


Tufte, Visual and Statistical Thinking (1997)

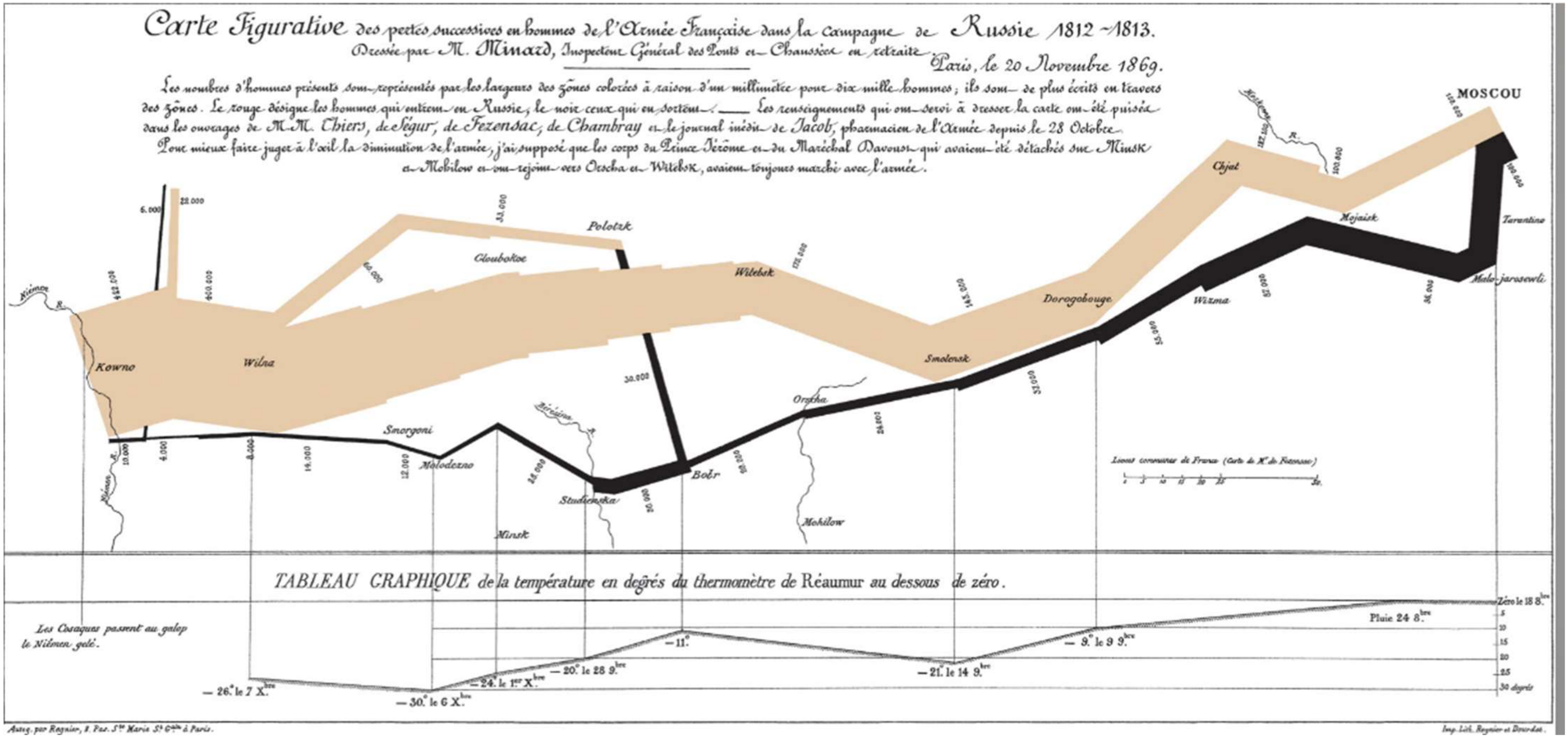
Post-launch Redrawn by Tufte



Tufte, Visual and Statistical Thinking (1997)



R package: DAAG has O-rings Dataset shown here



"probably the best statistical graphic ever drawn" (Tufte, Visual Display of Quantitative Information)

Data Visualization Requires Clear Thinking

1. **Technical Foundations**
    - Tidy data
    - Variable type
    - Question type

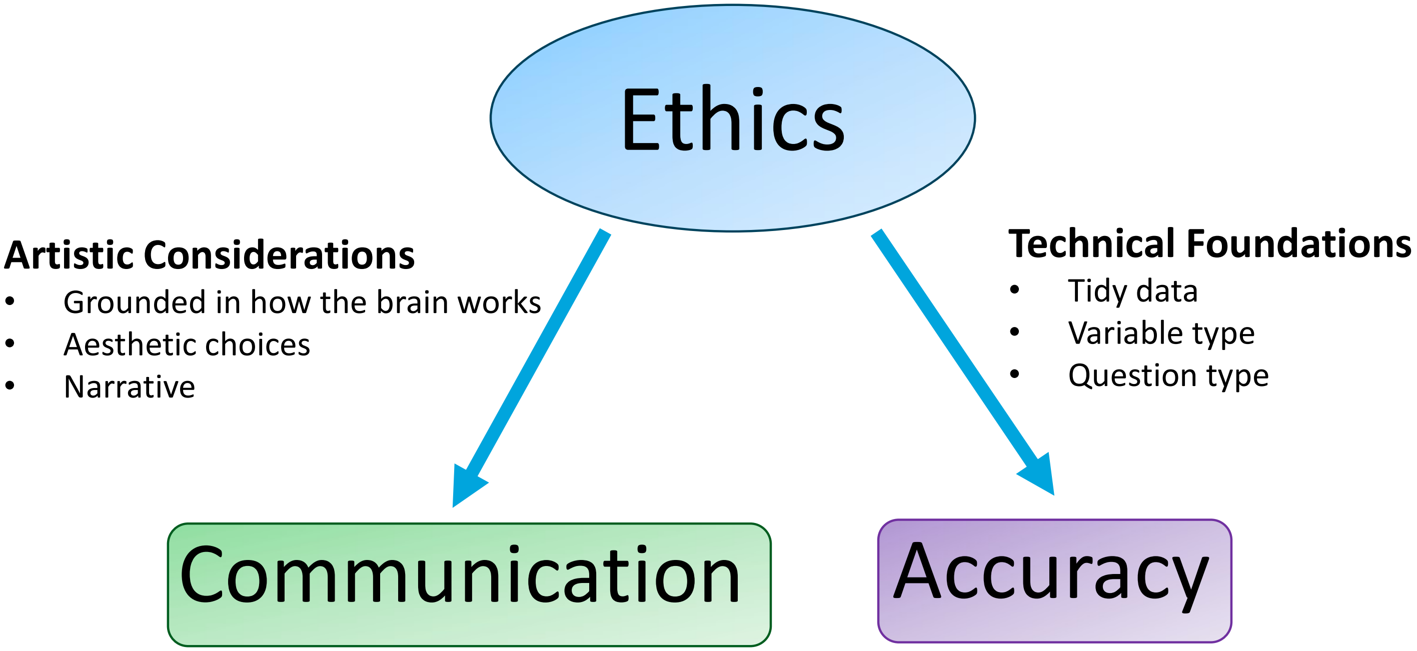
2. **Artistic Considerations**
    - Grounded in how the brain works
    - Aesthetic choices
    - narrative

Data Visualization Requires Clear Thinking

1. **Technical Foundations**
   - Tidy data
   - Variable type
   - Question type

✹ No coding in this presentation, but you might notice philosophical similarities to Leland Wilkinson “The grammar of visualization” and Hadley Wickham’s tidyverse/ggplot package

2. **Artistic Considerations**
   - Grounded in how the brain works
   - Aesthetic choices
   - Narrative



Ethics

Artistic Considerations

- Grounded in how the brain works
- Aesthetic choices
- Narrative

Technical Foundations

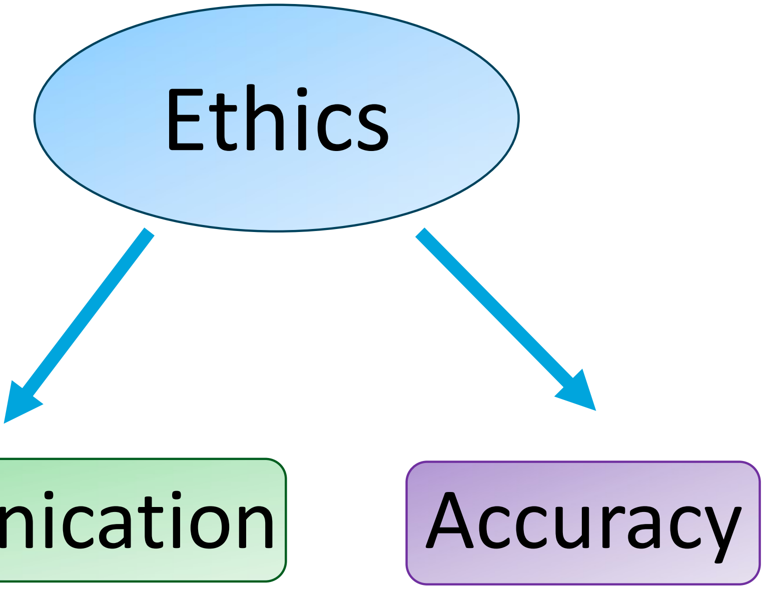
- Tidy data
- Variable type
- Question type

Communication

Accuracy

Tell an honest story

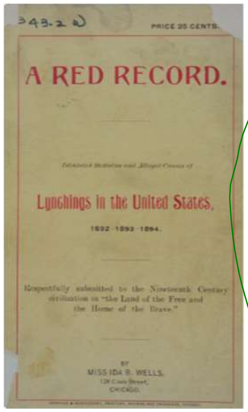
- Support conclusions with transparent methods
- Avoid distorted axes, cherry-picking, and hidden baselines



Ethics

Communication

Accuracy



343.2 W

PRICE 25 CENTS.

A RED RECORD.

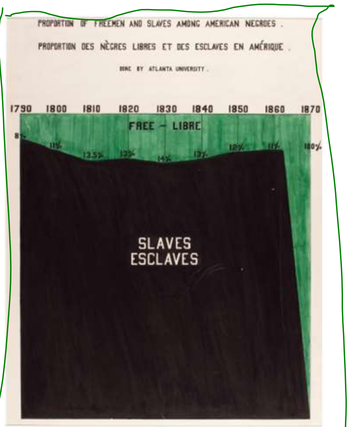
Tabulated Statistics and Alleged Causes of

Lynchings in the United States.

1892-1893-1894.

Respectfully submitted to the Nineteenth Century civilization in "the Land of the Free and the Home of the Brave."

BY
MISS IDA B. WELLS,
128 Clark Street,
CHICAGO.



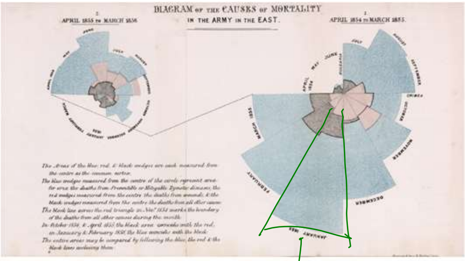


DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.
2
APRIL 1855 TO MARCH 1856.
1
APRIL 1854 TO MARCH 1855.

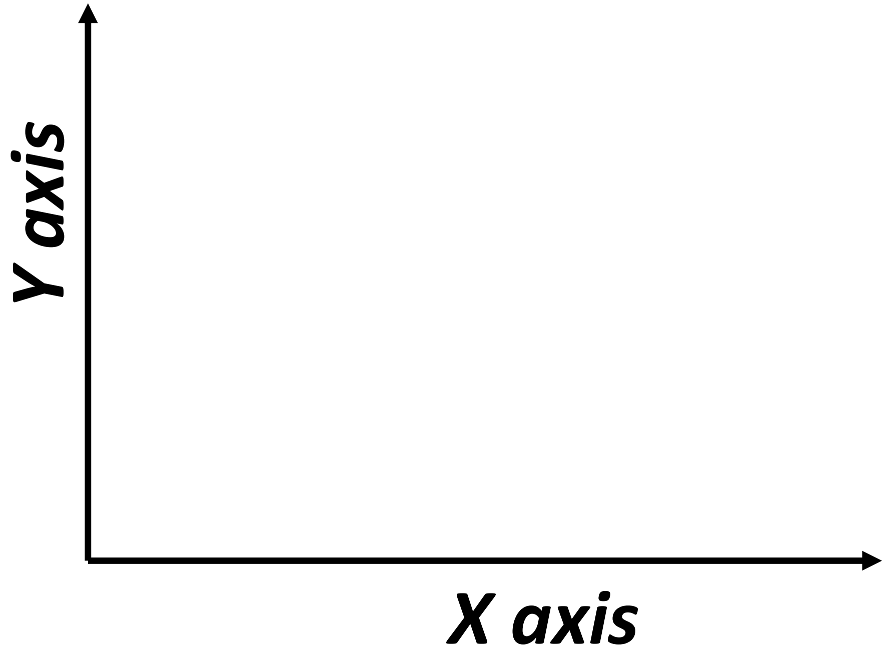
visualizatons of lyncings

WEB Debois
sociologist
worlds Fair

Flu
leading
cause
of death

Florence
Nightengale
standardized
nursing

Anatomy of a plot



- Axes: choose meaningful baselines; avoid dual axes unless essential
- Variable mapping: encode key data in position first

Anatomy of a plot

X axis = horizontal axis

Y axis = vertical axis

Variable type:

Categorical

- Ordinal: can be ordered
  ex.
- Nominal: cannot be meaningfully ordered
  ex.
- Binary
  ex.

Numeric

- Continuous: measure
  ex.
- Discrete: count
  ex.

Anatomy of a plot

X axis = horizontal axis

Y axis = vertical axis

Variable type:

Categorical

- Ordinal: can be ordered
  ex. Age, Socio economics status, education level
- Nominal: cannot be meaningfully ordered
  ex. Religion, blood group, cause of death, treatments, dog breeds
- Binary
  ex. Success and Failure

Numeric

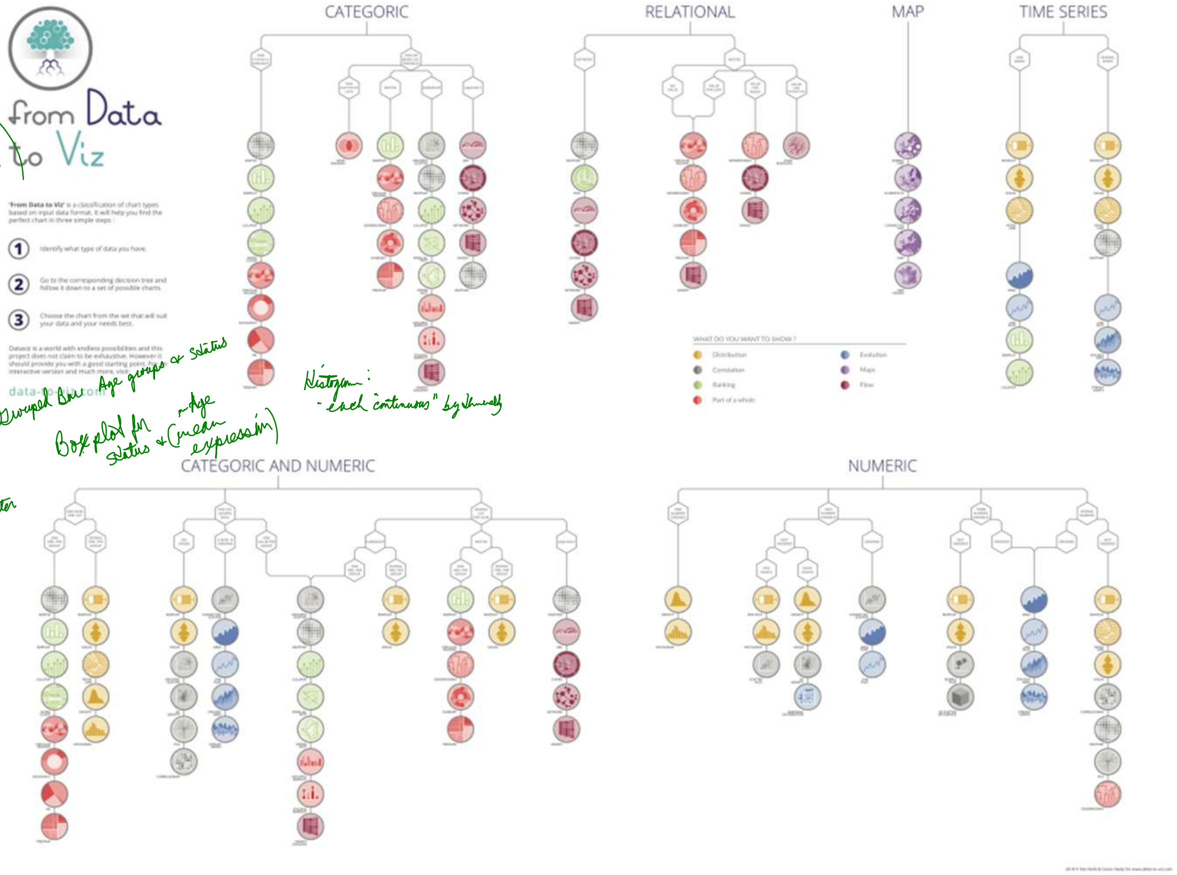
- Continuous: measure
  ex. Age, height, distance
- Discrete: count
  ex. Number of patients, results of rolling a die

Wes Anderson color pallete

Variable type governs plotting possibilities

Boxplot dest

is the skewness (median displacement) statistic



'From Data to Viz' is a classification of chart types based on input data format. It will help you find the perfect chart in three simple steps:

1. Identify what type of data you have.
2. Go to the corresponding decision tree and follow it down to a set of possible charts.
3. Choose the chart from the set that will suit your data and your needs best.

Dataviz is a world with endless possibilities and this project does not claim to be exhaustive. However it should provide you with a good starting point. For an interactive version and much more, visit:

data-to-viz.com

Grouped Bar Age groups or status

Boxplot for status + (mean expression) ~Age

Histogram:
- each "continuous" by themself

. Scatter

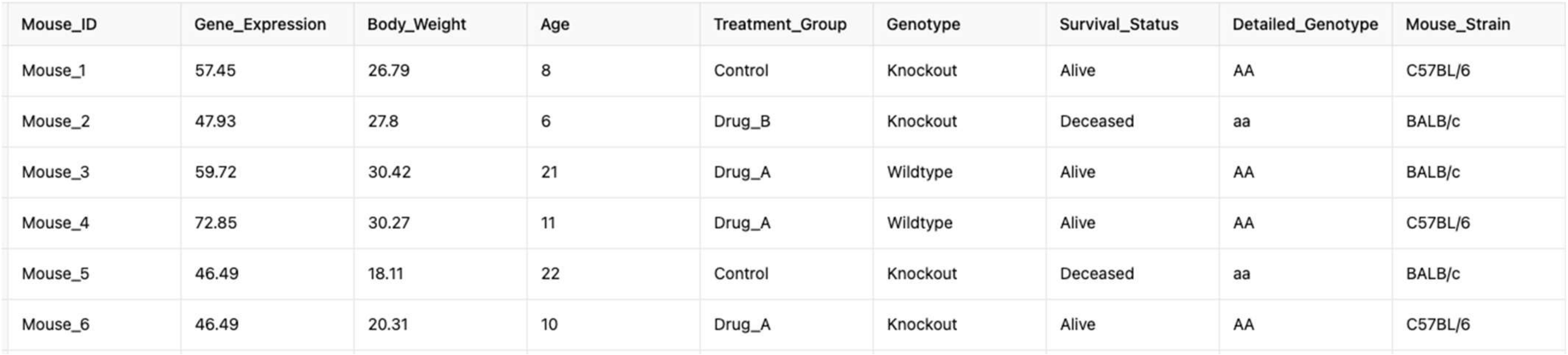
https://www.data-to-viz.com/

Variable type (and number) governs plotting possibilities

# variables	Variable Type	Recommended Plots	Use Case
1 (univariate)	Categorical	Bar Plot~~, Pie Chart~~	Comparing category frequencies
	Numerical	Histogram, Density Plot	Understanding distribution
2 (Bivariate)	Categorical & Categorical	Grouped Bar Chart, Mosaic Plot	Comparing proportions of two groups
	Numerical & Categorical	Boxplot, Violin Plot, Strip Plot	Comparing distributions across categories
	Numerical & Numerical	Scatter Plot, Line Plot, Hexbin Plot	Examining relationships or trends
3+ (Multivariate)	Multiple Categorical	Stacked Bar Chart	Analyzing categorical interactions
	Multiple Numerical	Scatterplot Matrix	Comparing multiple numeric relationships
	Mixed	Faceted Plots, Heatmap, Bubble Chart	Visualizing mixed data relationships

https://www.data-to-viz.com/

A subset from a dataset of 200 mice:



Mouse_ID	Gene_Expression	Body_Weight	Age	Treatment_Group	Genotype	Survival_Status	Detailed_Genotype	Mouse_Strain
Mouse_1	57.45	26.79	8	Control	Knockout	Alive	AA	C57BL/6
Mouse_2	47.93	27.8	6	Drug_B	Knockout	Deceased	aa	BALB/c
Mouse_3	59.72	30.42	21	Drug_A	Wildtype	Alive	AA	BALB/c
Mouse_4	72.85	30.27	11	Drug_A	Wildtype	Alive	AA	C57BL/6
Mouse_5	46.49	18.11	22	Control	Knockout	Deceased	aa	BALB/c
Mouse_6	46.49	20.31	10	Drug_A	Knockout	Alive	AA	C57BL/6

How would we visualize the following?

1. Is there a relationship between **Gene_Expression** and **Body_Weight**?

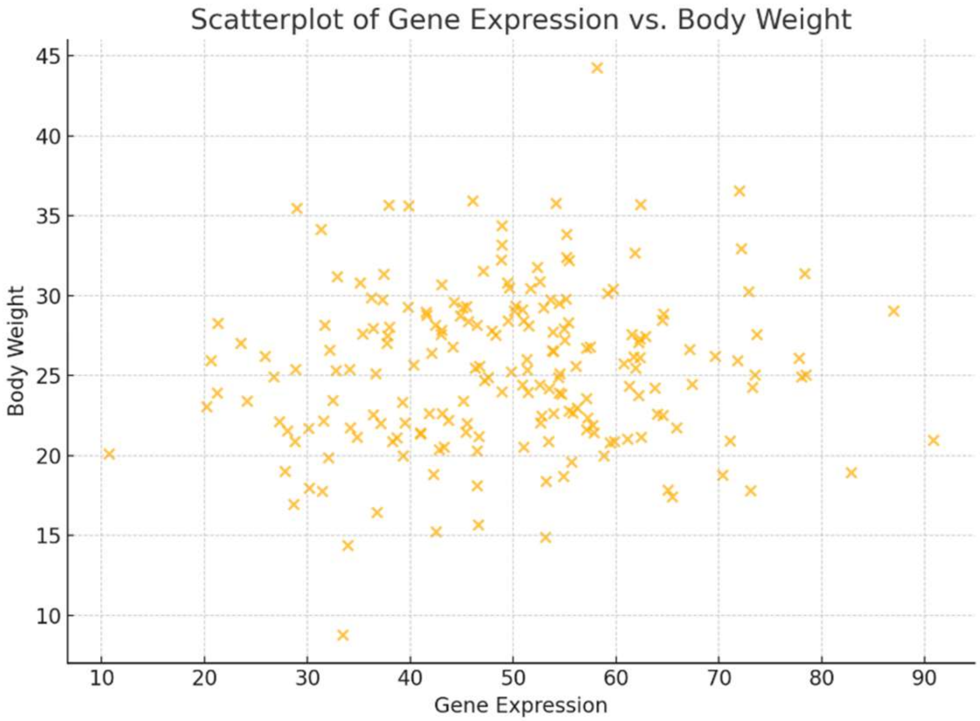
2. How does **Gene_Expression** differ among **Treatment_Group**? Or Genotype? Or Mouse_Strain?

3. Does **Body_Weight** vary between different **Detailed_Genotype** groups? Or Mouse_Strain?

4. Is there an association between **Treatment_Group** and **Survival_Status**?

How would we visualize the following?

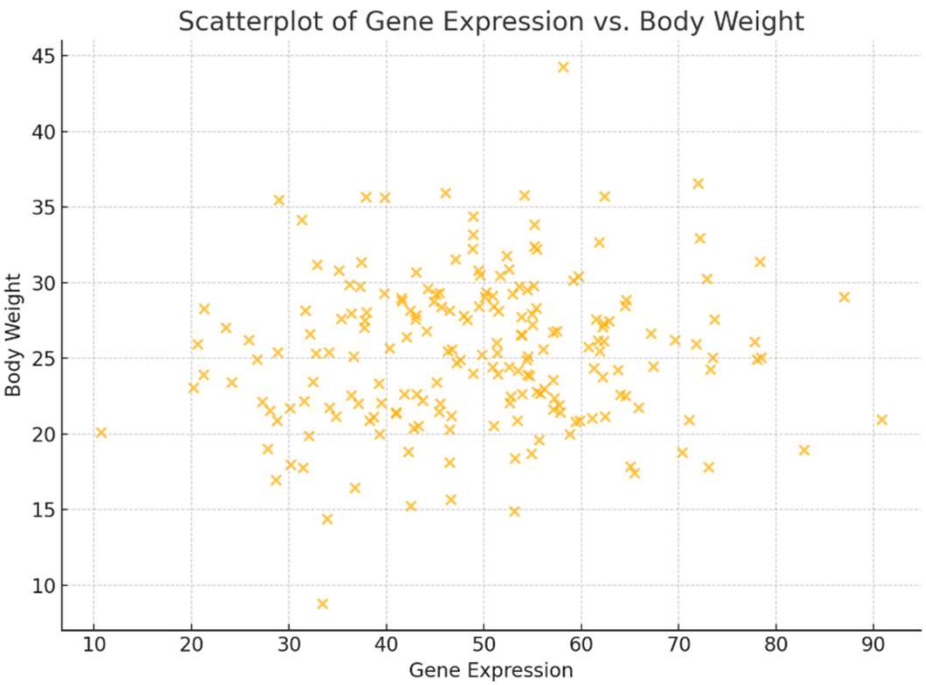
1. Is there a relationship between **Gene_Expression** and **Body_Weight**?

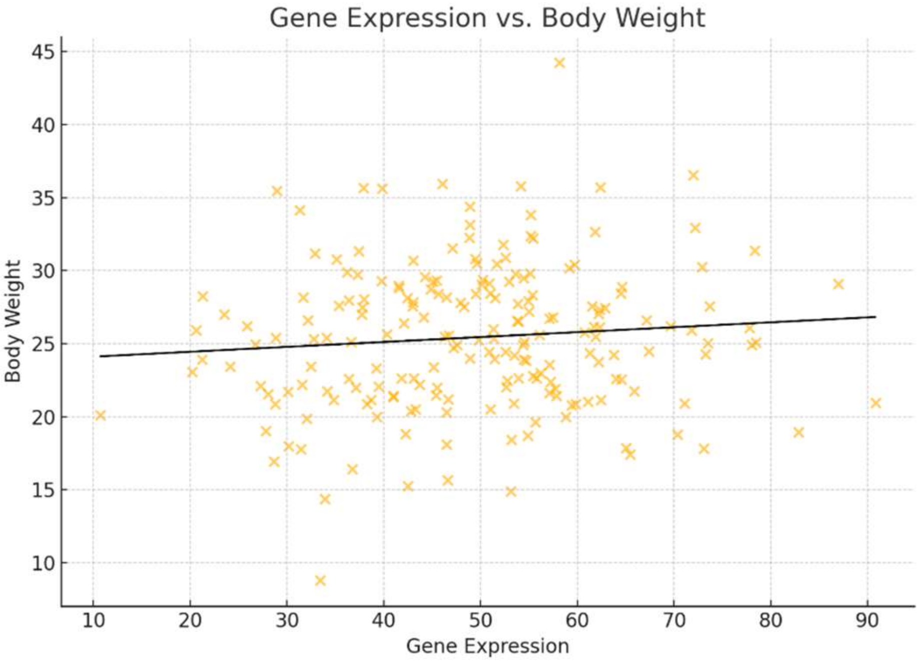


https://www.data-to-viz.com/

How would we visualize the following?

1. Is there a relationship between **Gene_Expression** and **Body_Weight**?

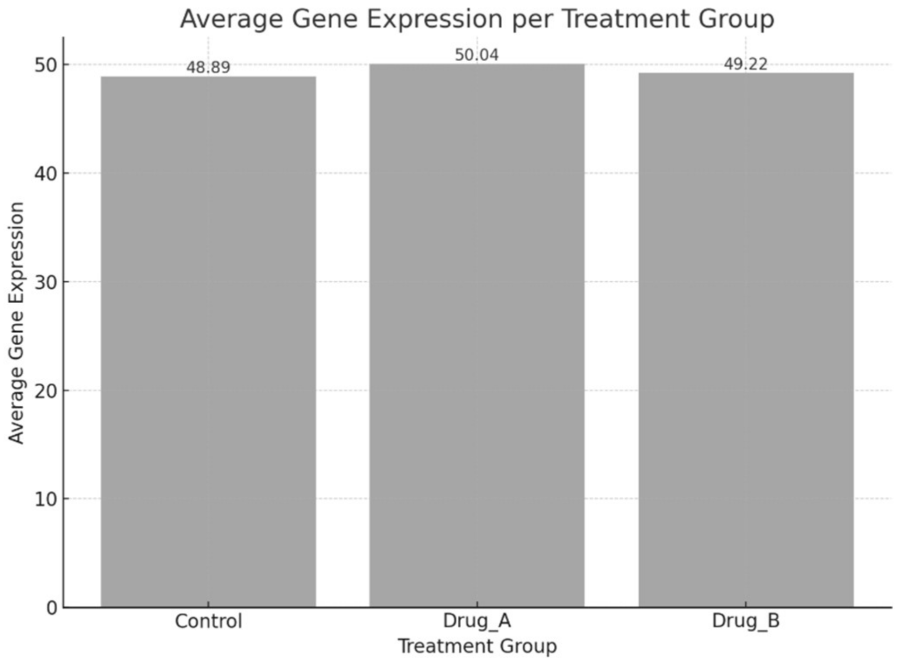


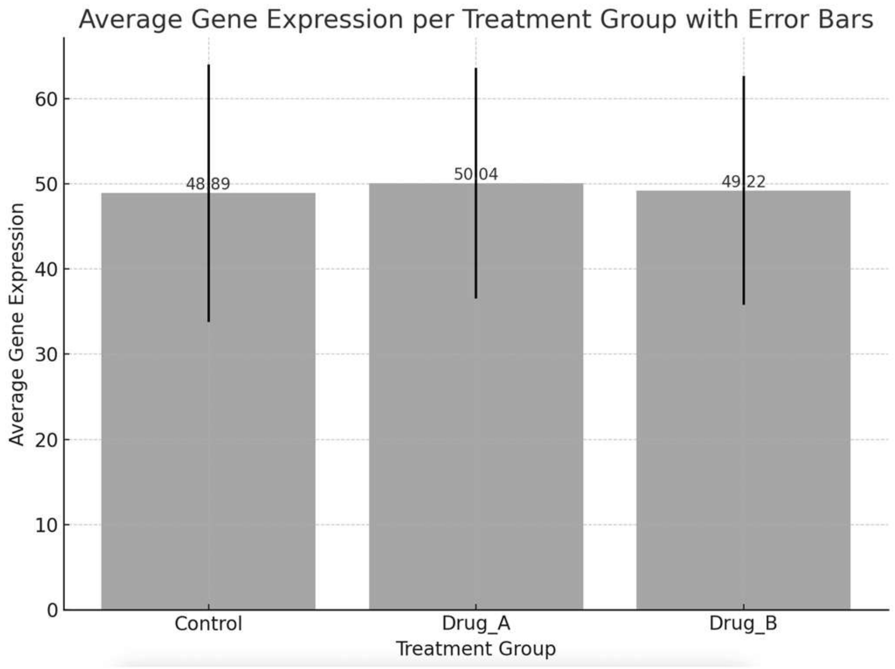


https://www.data-to-viz.com/

How would we visualize the following?

2. How does **Gene_Expression** differ among **Treatment_Group**? Or Mouse_Strain?

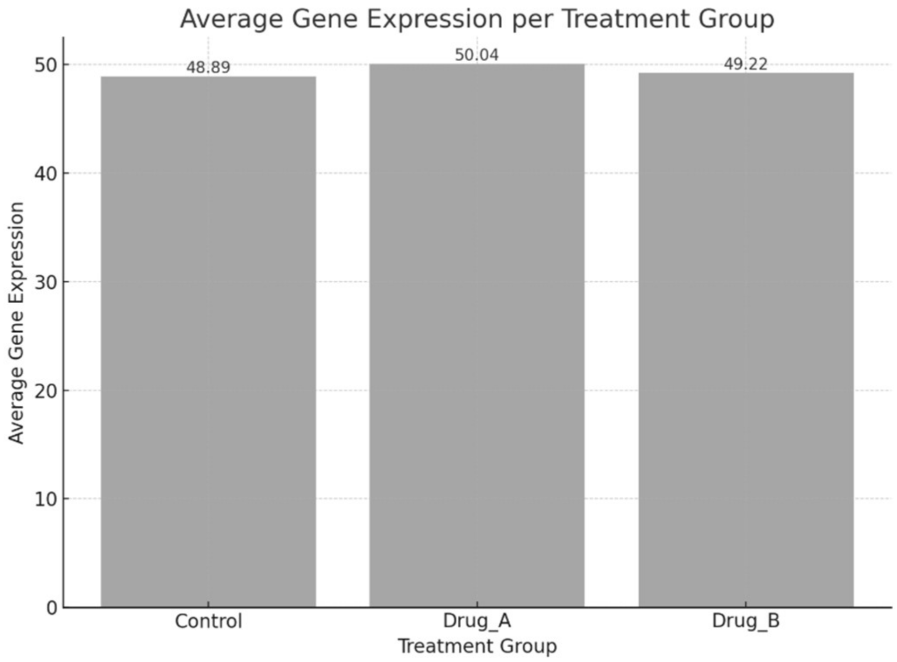


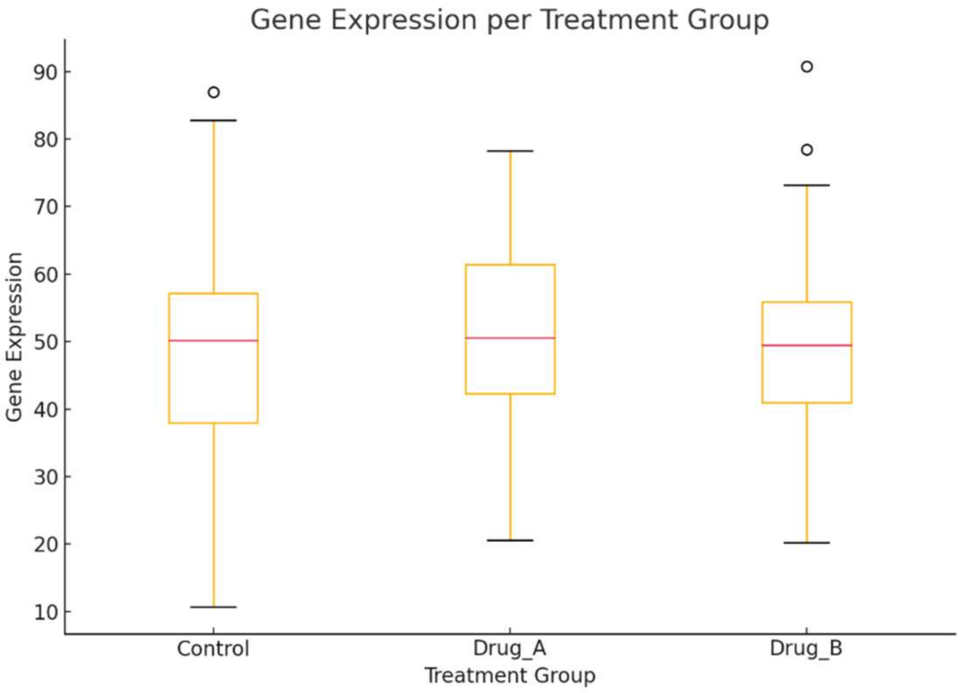


https://www.data-to-viz.com/

How would we visualize the following?

2. How does **Gene_Expression** differ among **Treatment_Group**? Or Mouse_Strain?

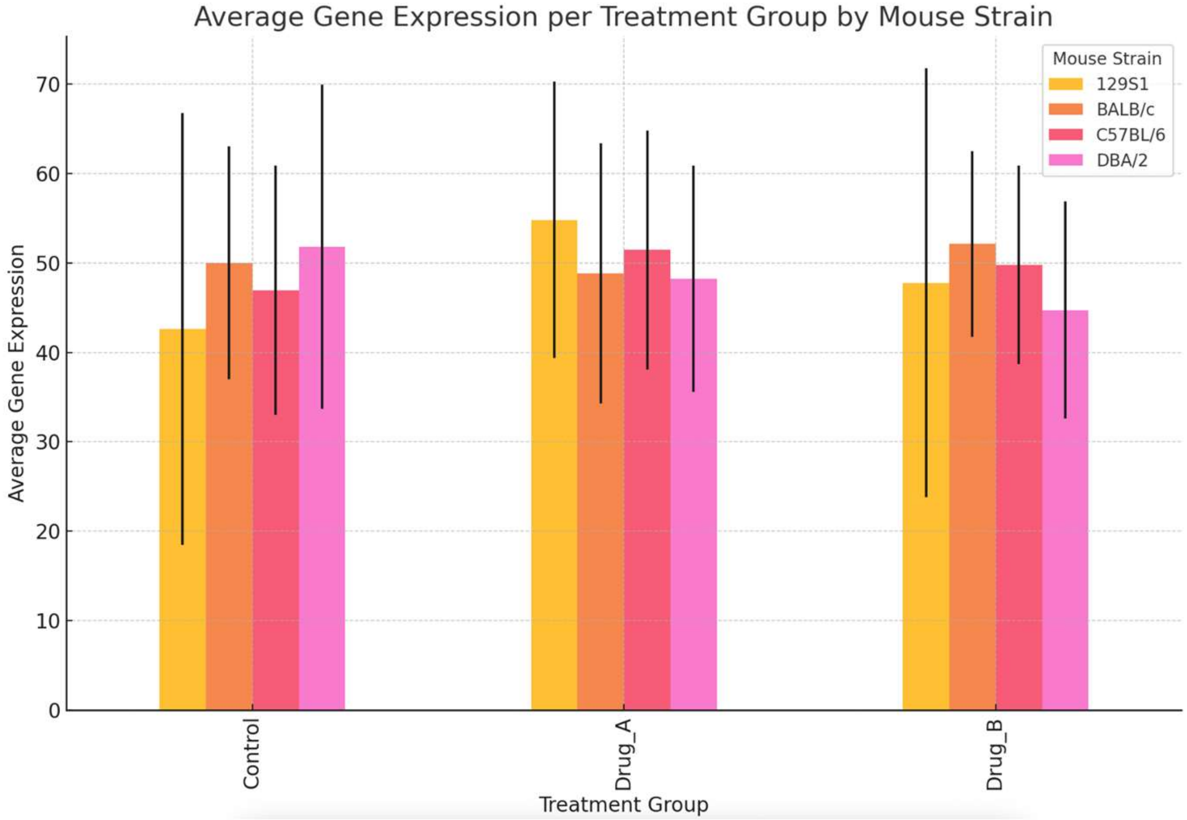




https://www.data-to-viz.com/

How would we visualize the following?

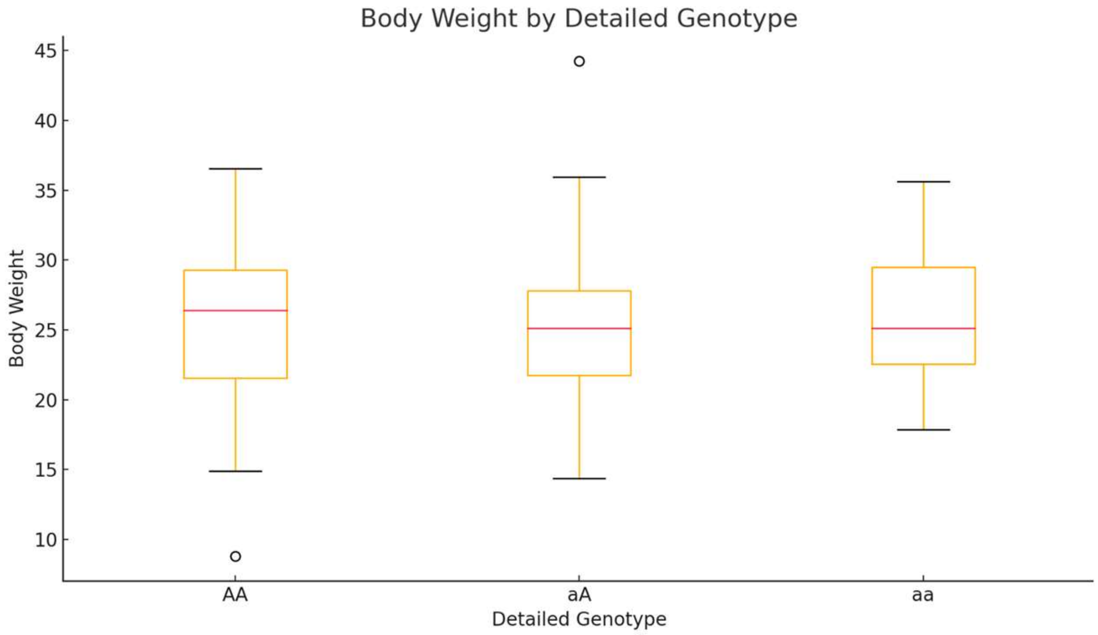
2. How does **Gene_Expression** differ among **Treatment_Group** and **Mouse_Strain**?



https://www.data-to-viz.com/

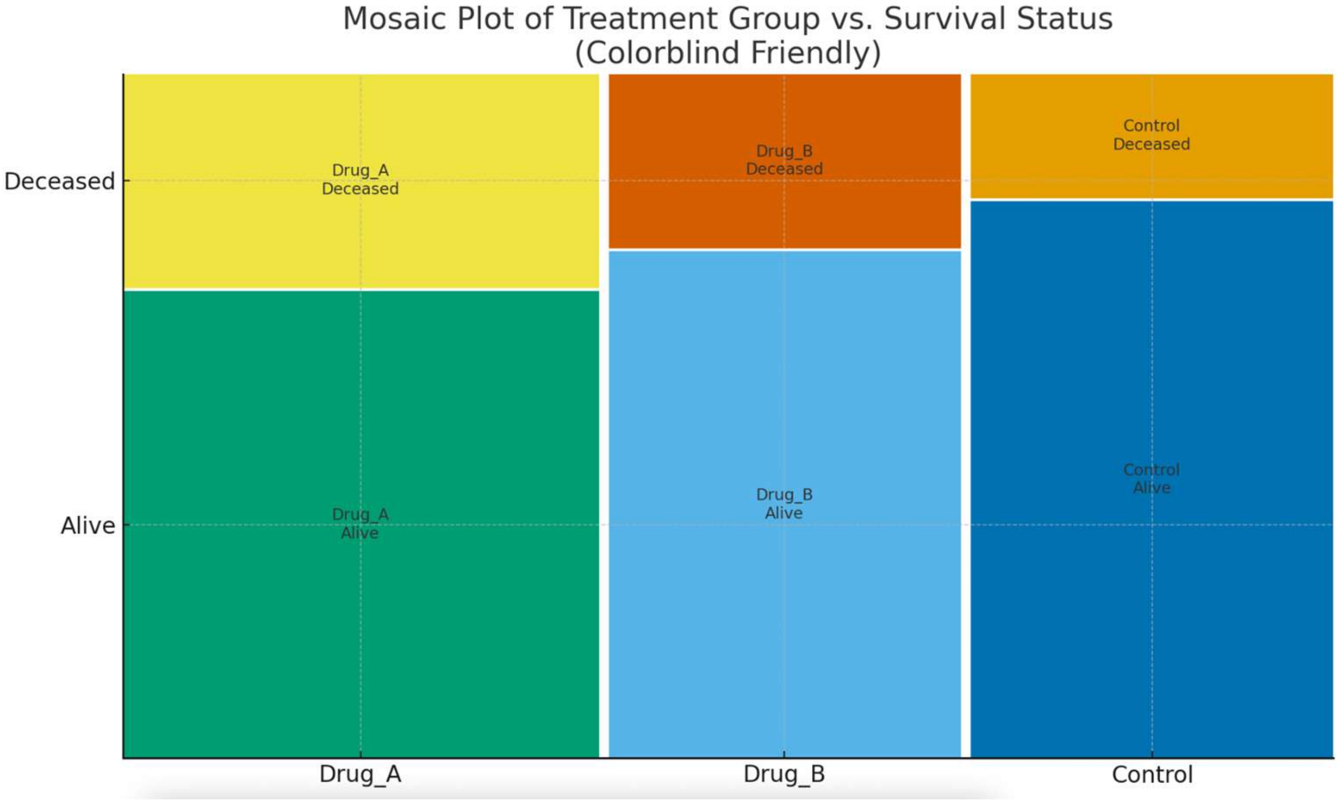
How would we visualize the following?

3. Does **Body_Weight** vary between different **Detailed_Genotype** group?



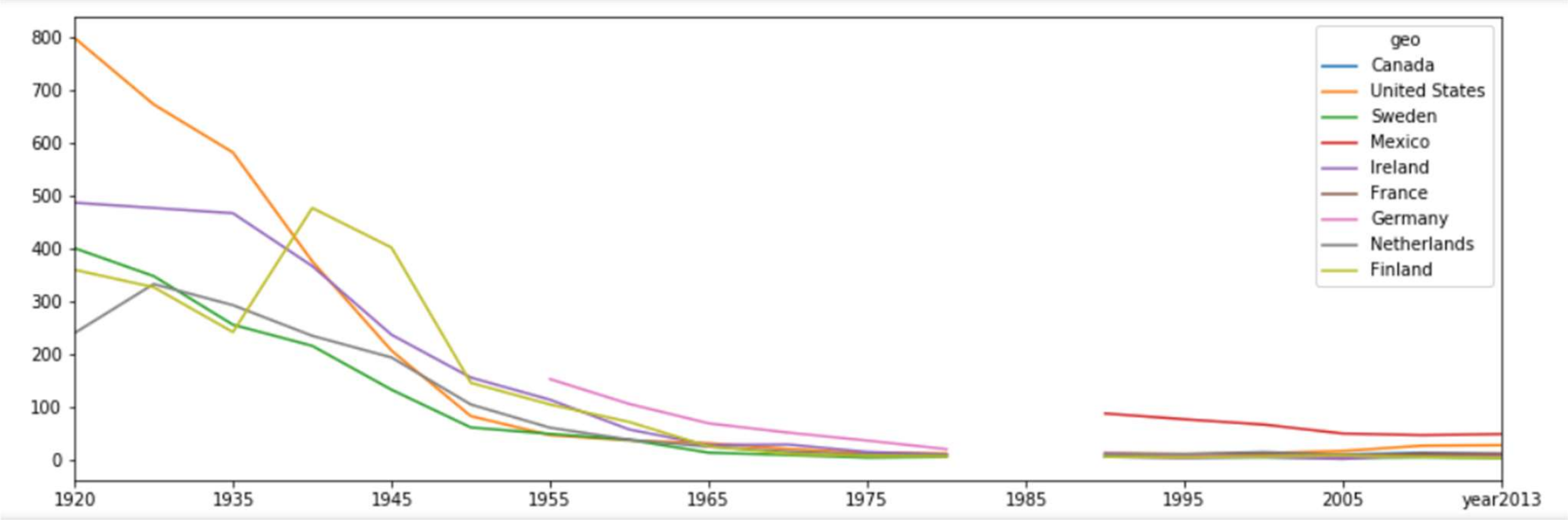
How would we visualize the following?

4. Is there an association between **Treatment_Group** and **Survival_Status**?



https://www.data-to-viz.com/

What is the following graph telling us?



Data is from gapminder.org

Once you have your variables and question, there are still decisions to be made….

How Our Eyes Read Charts

Pre-attentive attributes (Ware, 2004; Illinsky & Steele, 2011):

- Our brains notice and process these instantly
- Color, form, movement, spatial positioning

Gestalt principles:

- proximity, similarity, continuity to signal relationships

Avoid clutter: too much visual noise reduces understanding.

Data Integrity (Tufte)

Decisions still need to be made

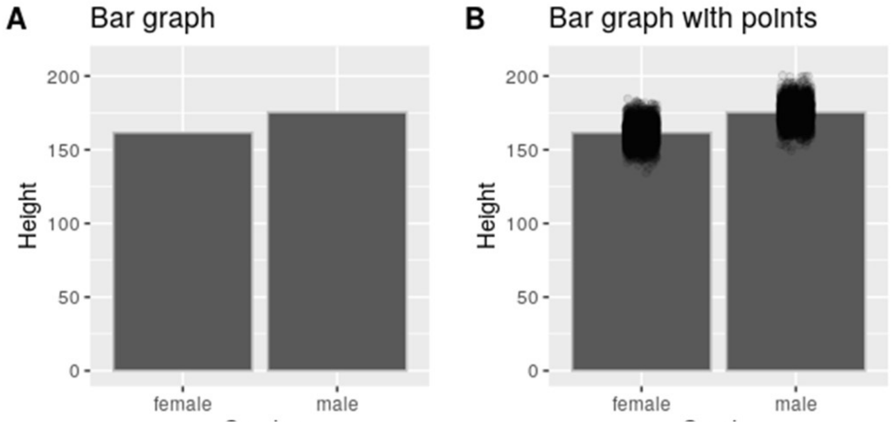


Figure 6.3: **Four** different ways of plotting the difference in height between men and women in the NHANES dataset. Panel A plots the means of the two groups, which gives no way to assess the relative overlap of the two distributions. Panel B shows the same bars, but also overlays the data points, jittering them so that we can see their overall distribution.

Decisions still need to be made

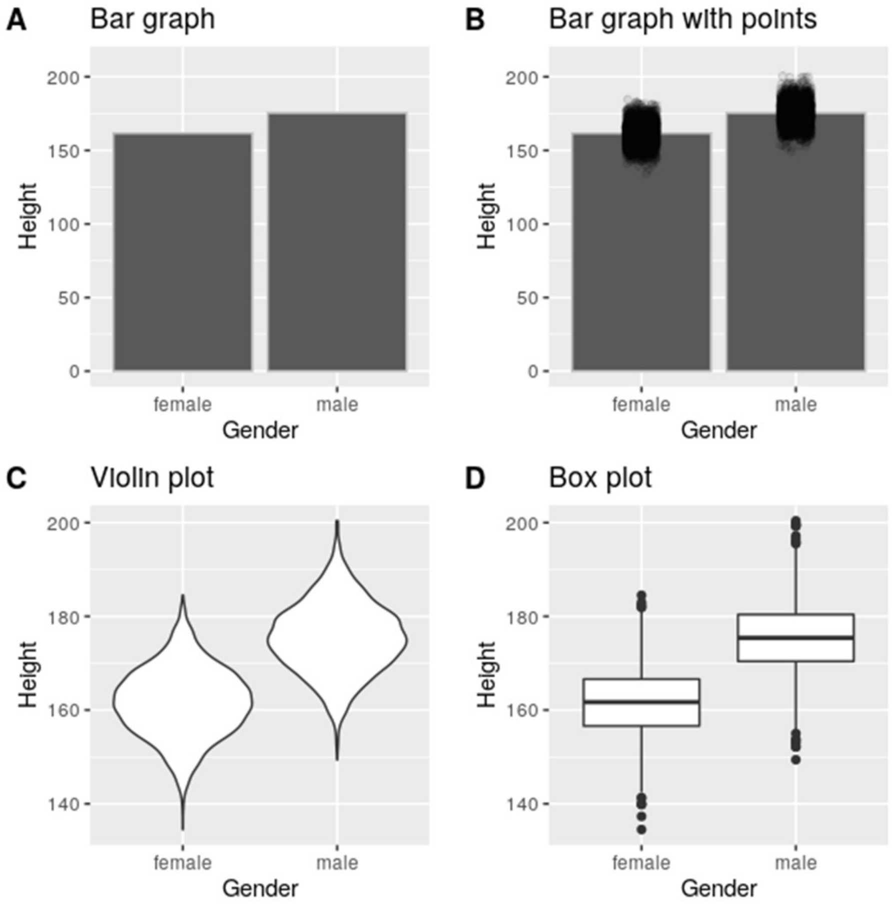


Figure 6.3: **Four** different ways of plotting the difference in height between men and women in the NHANES dataset. Panel A plots the means of the two groups, which gives no way to assess the relative overlap of the two distributions. Panel B shows the same bars, but also overlays the data points, jittering them so that we can see their overall distribution. Panel C shows a violin plot, which shows the distribution of the datasets for each group. Panel D shows a box plot, which highlights the spread of the distribution along with any outliers (which are shown as individual points).

[*Stastics for the 21st Century Viz Example*]()

Principles of good visualization*

Simplify	Remove Chartjunk & Maximize Data-Ink Ratio	• Keep it clean; Avoid unnecessary icons • Let the data shine; Use the least ink possible while ensuring clarity
Be Honest	Avoid distorting data & Misleading Scales	• Start axes at 0 • Maintain proportionality; don't exaggerate differences
Make Comparisons easy	Use multiple charts instead of overcrowding	• Maintain consistency across visualizations
Label Thoughtfully & Integrate Explanations	Place labels directly on data (instead of a separable legend)	• Combine text, graphics, and numbers for seamless storytelling
Show Context & Meaning	Highlight important trends while avoiding unnecessary clutter	• Include reference points, baselines, and annotations
Choose the Right Visualization for the Data	Clarity >>> novelty	• Use the correct graph type

- (Mostly) Edward Tufte
- "How to Lie with Statistics"- Darrell Huff
- "How Charts Lie"- Alberto Cairo
- " Calling Bullshit" – Carl T. Bergstrom & Jevin D. West

Data-Ink Ratio Checklist

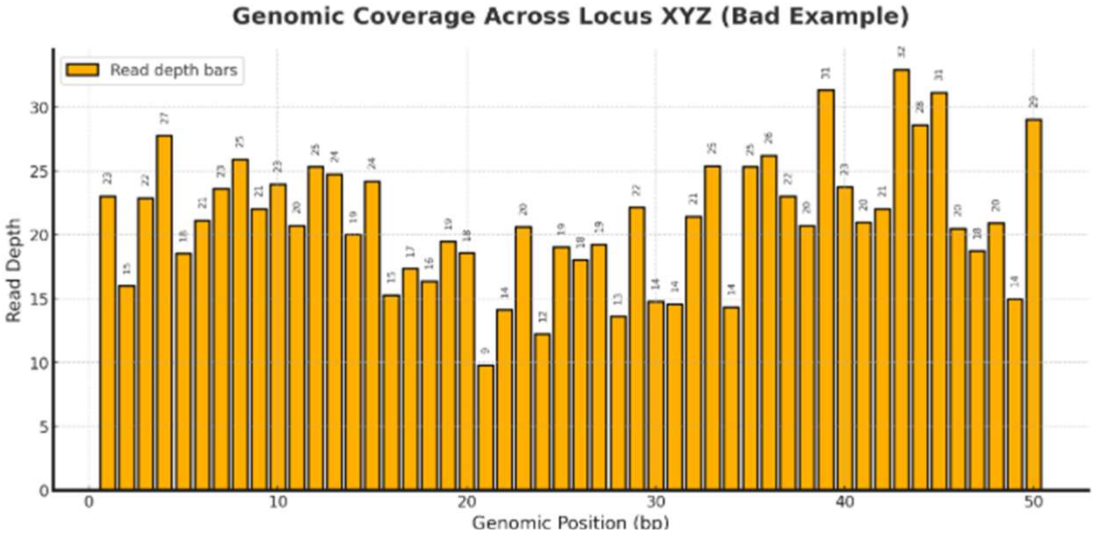
- SHOW THE DATA
- Minimize gridlines, borders, and background decorations
- Avoid 3D effects
- Directly label data when possible
- Reduce Cognitive Load

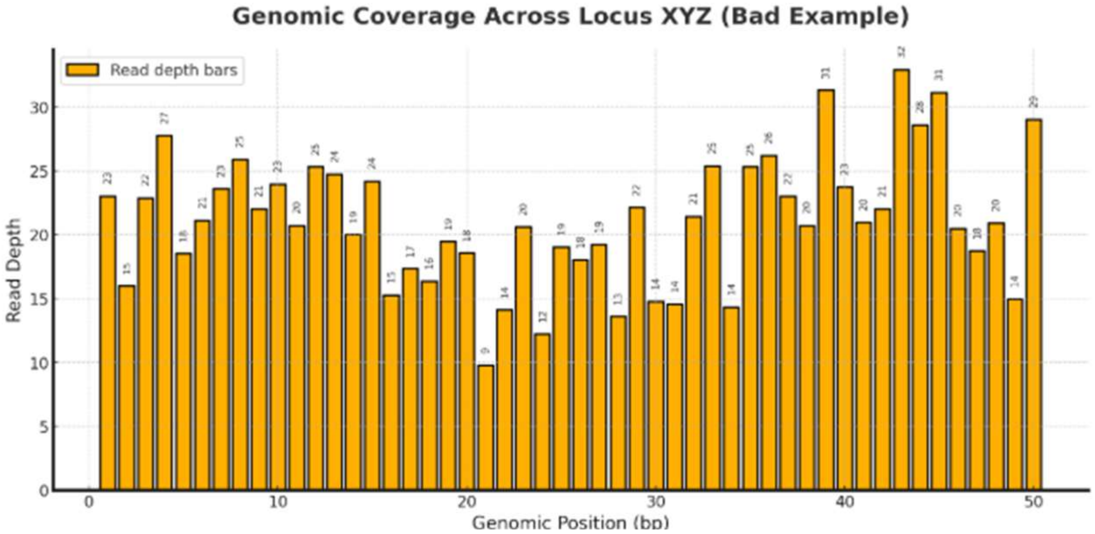
Color Is Not Decoration

- Use color intentionally: categorical vs. sequential
- Avoid red-green combinations for accessibility; use colorblind-safe palettes
- Don't use too many colors: **clarity > variety**

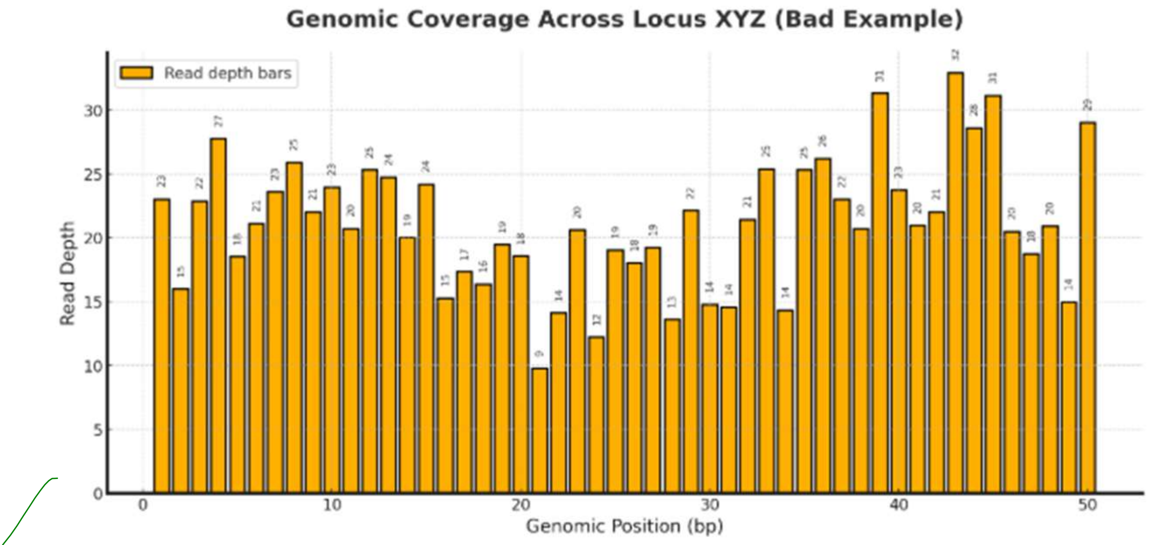
“Graphical excellence is what gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.”

The Visual Display of Quantitative Information


Genomic Coverage Across Locus XYZ (Bad Example)
Read depth bars
Read Depth
0
5
10
15
20
25
30
23
15
22
27
18
21
23
25
21
23
20
25
24
19
24
15
17
16
19
18
9
14
20
12
19
18
19
13
22
14
14
21
25
14
25
26
22
20
31
23
20
21
32
28
31
20
18
20
14
29
0
10
20
30
40
50
Genomic Position (bp)

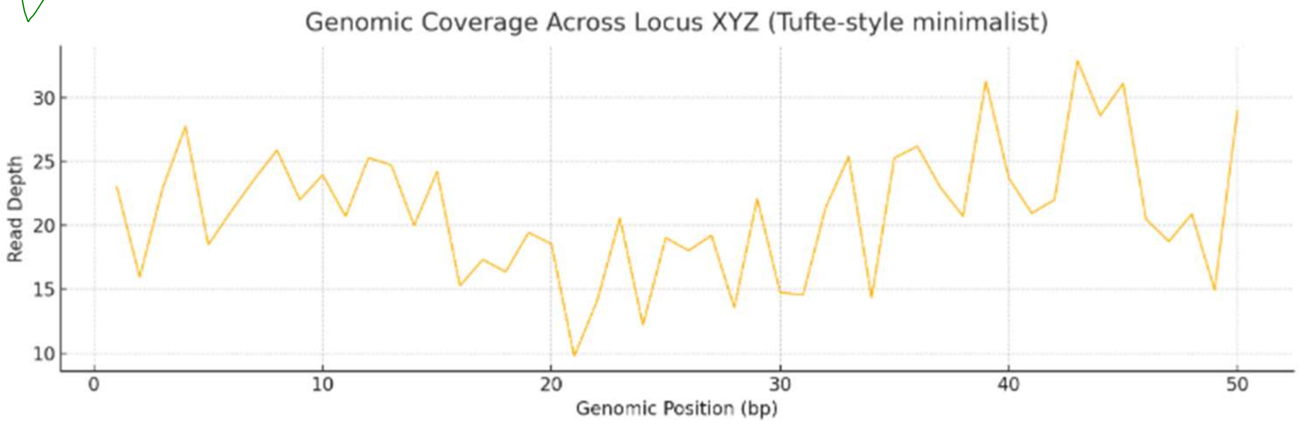


- **Low data-ink ratio**: most ink goes to decoration.
- **Chartjunk**: grid, 3-D-ish bars, bold title—all distract from the pattern.
- **Over-annotation**: numbers on every bar repeat the vertical axis.



- **Low data-ink ratio**: most ink goes to decoration.
- **Chartjunk**: grid, 3-D-ish bars, bold title—all distract from the pattern.
- **Over-annotation**: numbers on every bar repeat the vertical axis.

✓



- **High data-ink ratio**: nearly all ink encodes data.
- **No chartjunk**: viewer sees the trend immediately.
- **Economy of annotation**: axis labels alone are enough; numbers can be read off the scale.

Beyond Technical considerations…..

Data Visualization requires **clear thinking**

→ It valuable to create pc's from samples of the same data for linear modelling

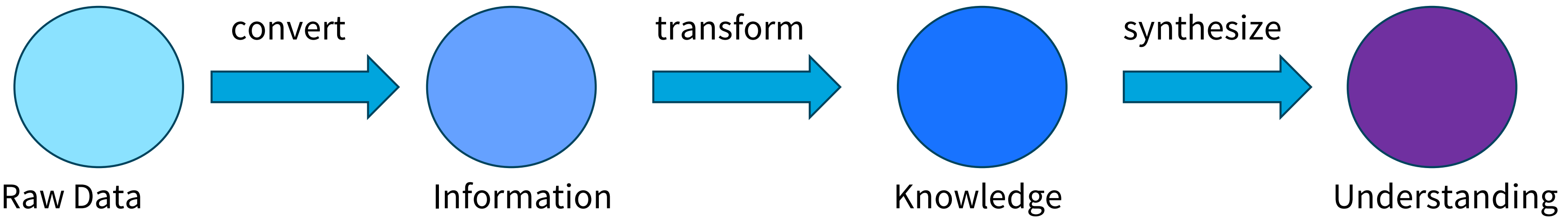
Narrative:

"A story is a set of observations, facts, or events, true or invented, that are presented in a specific order such that they create an emotional reaction in the audience. The emotional reaction is created through the build-up of tension at the beginning of the story followed by some type of resolution towards the end of the story."

"**Most data visualization is done for the purpose of communication**. We have an insight about a dataset, and we have a potential audience, and we would like to convey our insight to our audience. To communicate our insight successfully, we will have to present the audience with a clear and exciting story. ***The need for a story may seem disturbing to scientists and engineers, who may equate it with making things up, putting a spin on things, or overselling results"***

- Claus Wilke

"The role of scientists is to collect data and **transform** them into understanding. **Their role as authors is to present that understanding**."



"The data are supporting actors in the story you tell. The lead actors are the questions and the larger issues you are addressing. **The story grows from the data, but the data are not the story**."

"Only by exploring the boundaries and limits of your data can you find the important story."

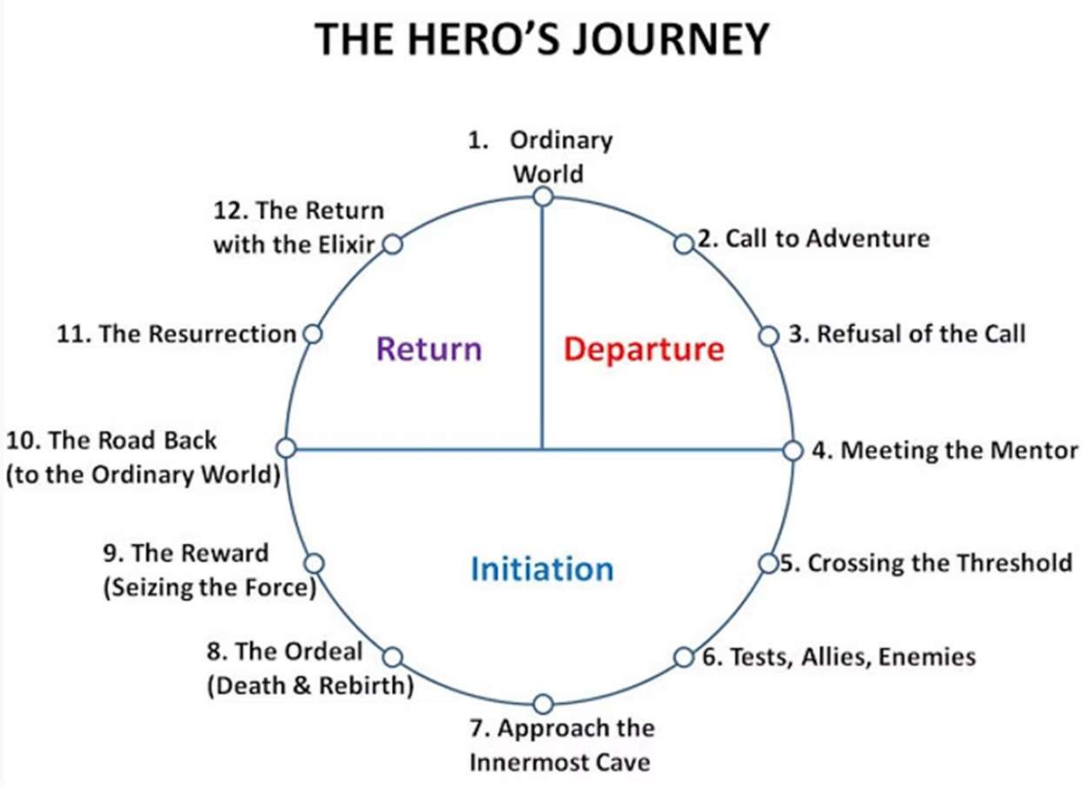

THE HERO'S JOURNEY
1. Ordinary World
2. Call to Adventure
3. Refusal of the Call
4. Meeting the Mentor
5. Crossing the Threshold
6. Tests, Allies, Enemies
7. Approach the Innermost Cave
8. The Ordeal (Death & Rebirth)
9. The Reward (Seizing the Force)
10. The Road Back (to the Ordinary World)
11. The Resurrection
12. The Return with the Elixir
Return
Departure
Initiation

Image Source : mikeduran.com

1. Ordinary World: Set the baseline
2. Call to Adventure: An anomaly or pattern
3. Threshold: From data to implications
4. Climax: Insight or discovery
5. Resolution: Application or next step

You can also use a three-act structure

https://www.storyboardthat.com/articles/e/heroic-journey

Hero's Journey Wikipedia

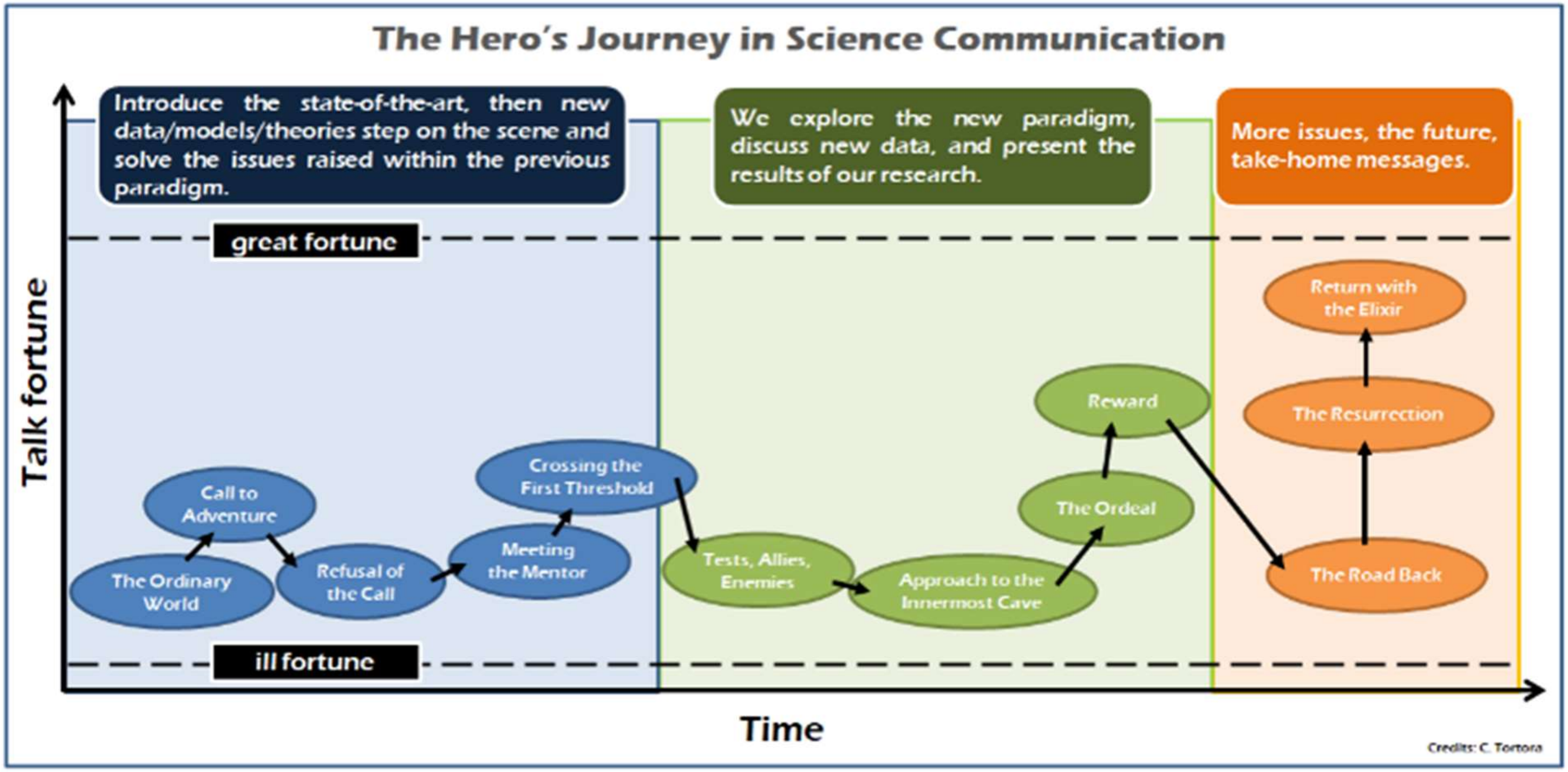

The Hero's Journey in Science Communication
Introduce the state-of-the-art, then new data/models/theories step on the scene and solve the issues raised within the previous paradigm.
We explore the new paradigm, discuss new data, and present the results of our research.
More issues, the future, take-home messages.
great fortune
ill fortune
Talk fortune
Time
The Ordinary World
Call to Adventure
Refusal of the Call
Meeting the Mentor
Crossing the First Threshold
Tests, Allies, Enemies
Approach to the Innermost Cave
The Ordeal
Reward
The Road Back
The Resurrection
Return with the Elixir
Credits: C. Tortora

Figure 4. The Hero's Journey in Science communication in a nutshell. The talk emotional arc is presented on the Vonnegut's 'beginning-end' and 'ill fortune-great fortune' axes. The talk fortune is shown as a function of talk time, as a progression through lows and highs, divided in the 12 stages and the three acts: Departure or Separation (blue), Initiation (green) and Return (orange). Stages in the lowest part of the diagram correspond to issues/problems/failures, while on the top to discoveries and positive results. Note that the precise vertical position of each stage in the diagram is flexible. Credits: C. Tortora

ChatGPT: break down an example of scientific discovery into the Hero’s Journey for mouse genomics.

The Ordinary World

Headline: The Baseline: Conserved Genes Across Mouse Strains

Start with a **basic genomic overview**: A chart showing conserved genes across common mouse strains. Everything looks stable and expected. Gene expression is consistent. No red flags

Visual: Clean bar chart or heatmap showing stable gene expression across several strains (e.g., C57BL/6, BALB/c, DBA/2)

Notes: Emphasize expected patterns, genetic stability, and no major surprises. Use calm tones (blues/greys).

Call to Adventure

Headline: An Anomaly Appears: A Mutational Hotspot in Gene X

Introduce a surprising **mutation hotspot** discovered in a subgroup—perhaps affecting a gene known to regulate synaptic function.

Visual: Genome browser snapshot highlighting mutation in a red box; zoomed-in region on chromosome 7, for example

Notes: This mutation was unexpected and potentially impactful—cue the intrigue. This isn’t what we expected….

Crossing the Threshold

Headline: From Genes to Behavior: A Phenotypic Clue

Show how this mutation correlates with early onset of motor deficits in a mouse model. You’re now linking **genotype to phenotype**.

Visual: Split panel: (left) genotype heatmap showing mutation, (right) bar chart of reduced maze performance or locomotor activity

Notes: Mutation in Gene X is correlated with motor deficits—potential link to neurodegeneration. This gene isn’t just different…it is doing something.

Trials and Revelations

Headline: Exploring the Wider Genomic Landscape

Expand to other strains or cross-species comparisons:

•Is this mutation conserved in rats? In humans?

•Use PCA plots or UMAPs to show clustering by expression profile or epigenetic markers.

Visual: UMAP or PCA plot showing gene expression clusters by mouse strain; additional heatmap showing epigenetic markers

Notes: Mutation not isolated—appears in functionally related genes across networks. Introduce microscopy images of damaged neurons. This gene is part of a bigger story

The Climax / The Ordeal

Headline: The Pathway Revealed: A Network of Neurodegeneration

Reveal a **network visualization** of the gene’s pathway. It connects to several neurodegeneration-linked genes.

This could be a **biomarker**. Or a **target**.

Visual: Gene interaction network with Gene X as a central hub; bold red links to known Alzheimer's/Parkinson’s genes

Notes: This gene isn't just an outlier—it's a potential master regulator. High tension moment. This is the point of no return—we’ve found a key node.

Return with the Elixir

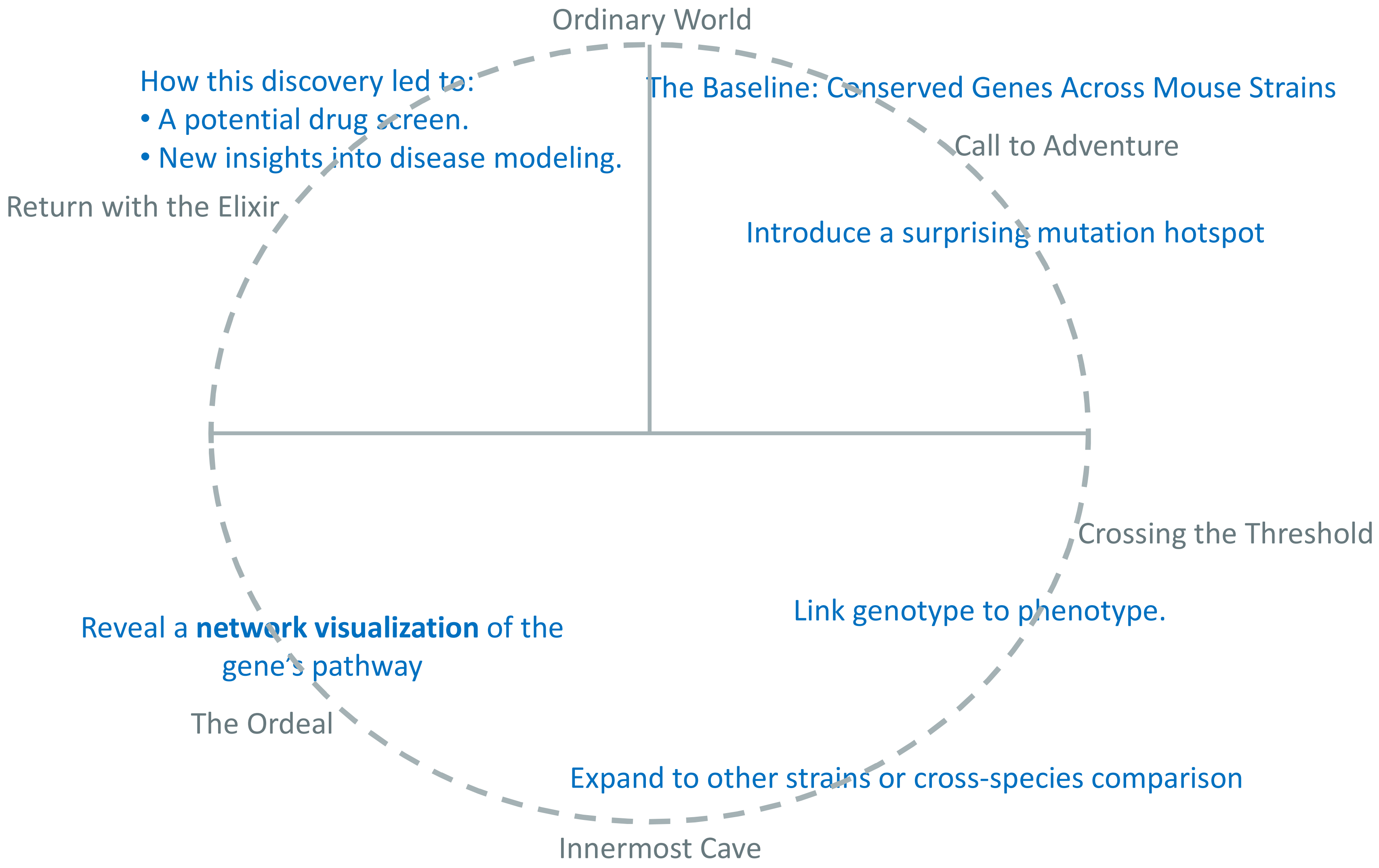
Headline: Breakthroughs and Hope: From Discovery to Intervention

Wrap up with how this discovery led to:

- A potential drug screen.
- New insights into disease modeling.
- A CRISPR-based experiment that reversed symptoms in a pilot study.

Visual: Before-and-after mouse performance chart; image of treated vs. untreated brain slices

Notes: Preliminary results from CRISPR experiments show symptom reversal. Ending on hope. Knowledge is power—and this gene could change lives



Ordinary World

The Baseline: Conserved Genes Across Mouse Strains

Call to Adventure

Introduce a surprising mutation hotspot

Crossing the Threshold

Link genotype to phenotype.

Expand to other strains or cross-species comparison

Innermost Cave

The Ordeal

Reveal a **network visualization** of the gene’s pathway

Return with the Elixir

How this discovery led to:

- A potential drug screen.
- New insights into disease modeling.

Summary

1. Variable type & question-driven design leads to effective visuals

2. Cognitive principles guide audience attention and comprehension
   - Especially limit cognitive load
   - Emphasize relationships

3. Ethics and clarity are non-negotiable in data communication

4. Use visualizations as part of a larger narrative
   - If you don’t give your audience a ’story’, they will find one for you

https://forms.office.com/r/dceAZNUVc1