Advanced Data Science and Fallacies

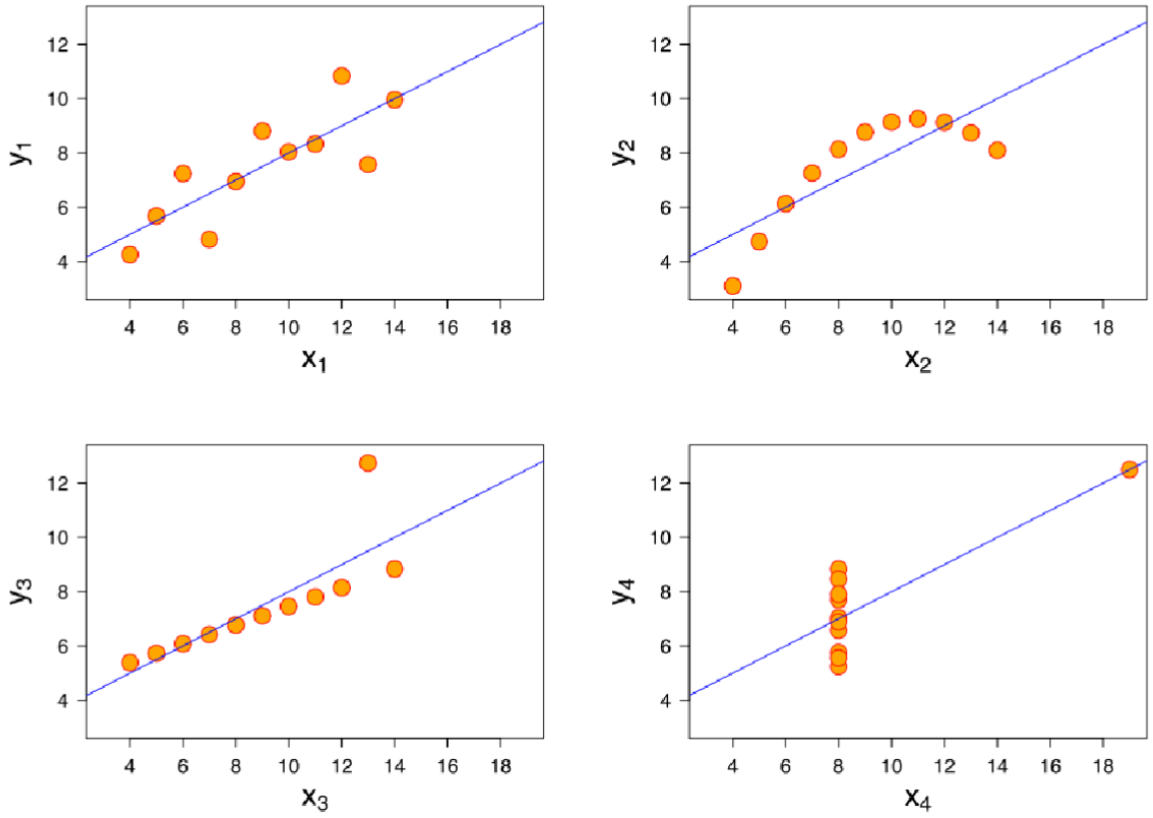
MIE223

Winter 2025

1 Potential Data Interpretation Fallacies

1.1 Anscombe Quartet: Visualize

When we look at mean of x, sample variance of x, mean of y, sample variance of y, correlation between x and y, and linear regression line, the accuracy differs.
If datasets have varying levels of accuracy, the mean, variance, and correlation will be the same. However, the linear regression line will differ.

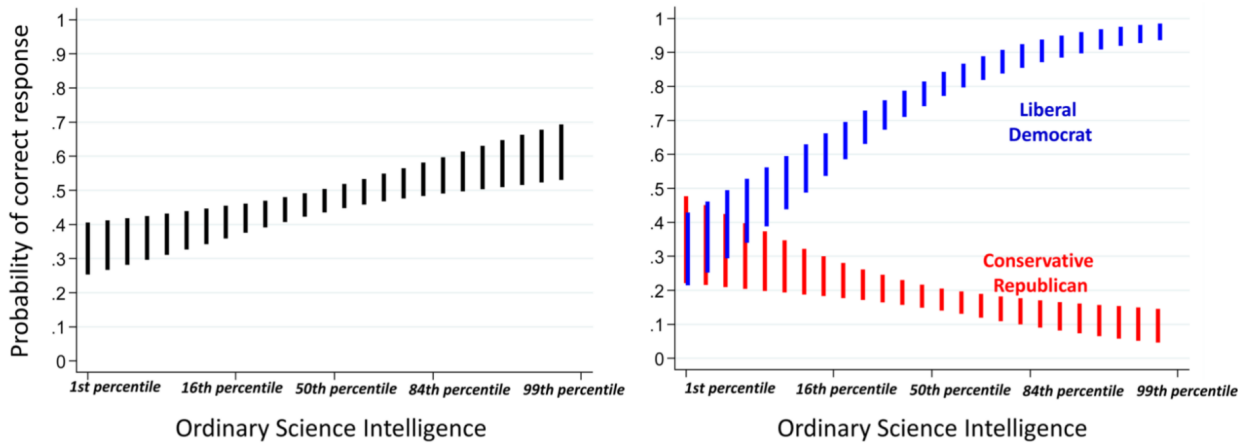


1.2 Spurious Correlation

Ratios induce spurious correlations. Solution: compositional data analysis.
x, y, and z are all statistically independent. The ratios induce high correlation between them.
Normalizing by a common number (such as in ratios) results in spurious correlation.

1.3 Confounding Variables

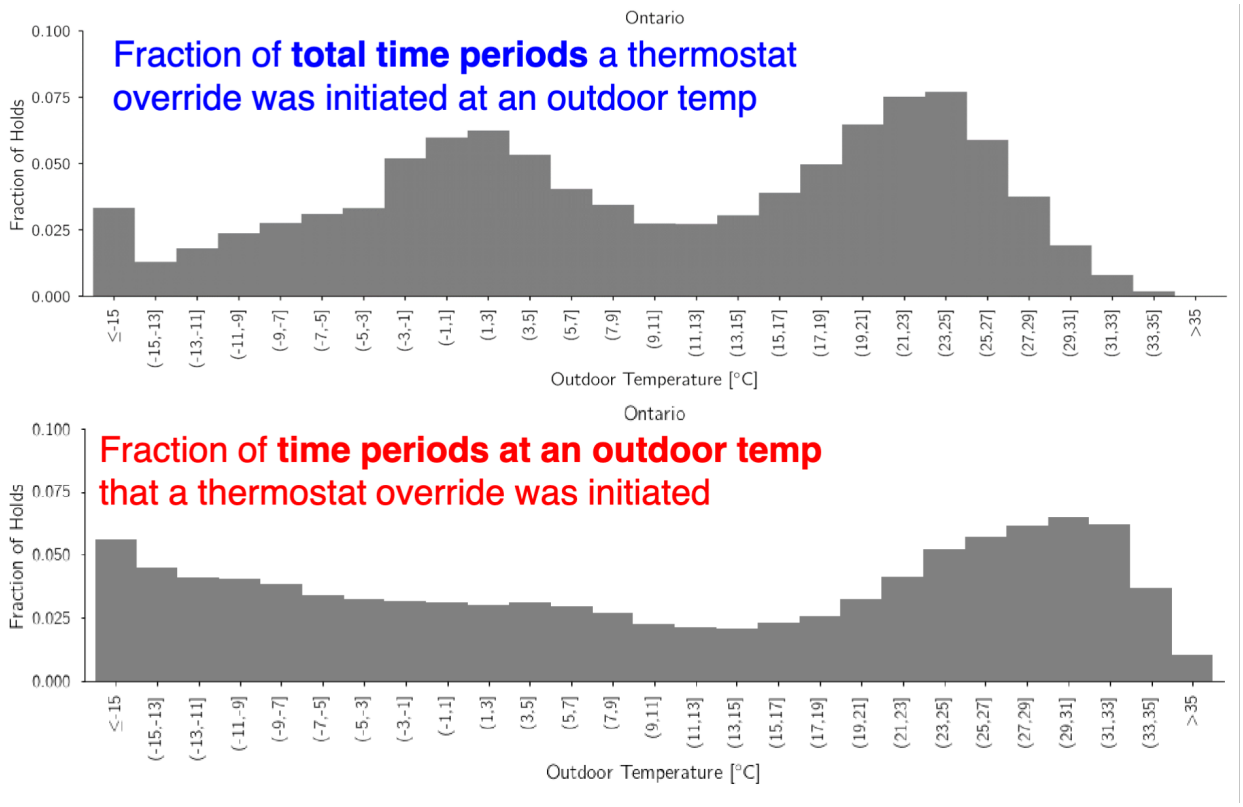
- Beware of (latent) confounding variables
  - Averaging over them can hide important trends



– How to find confounders? Search.
– What if latent? Learn mixture models (EM), deep learning, etc.

Political affiliation is a confounder. Belief in a scientific statement is conditioned on political affiliation.

1.4 Correctly Normalizing



blue

- peaks between 0 and 1 and 23 and 25

- Not at certain temperatures with equivalent probability
- Looking at a joint measurement not a conditional measurement

red

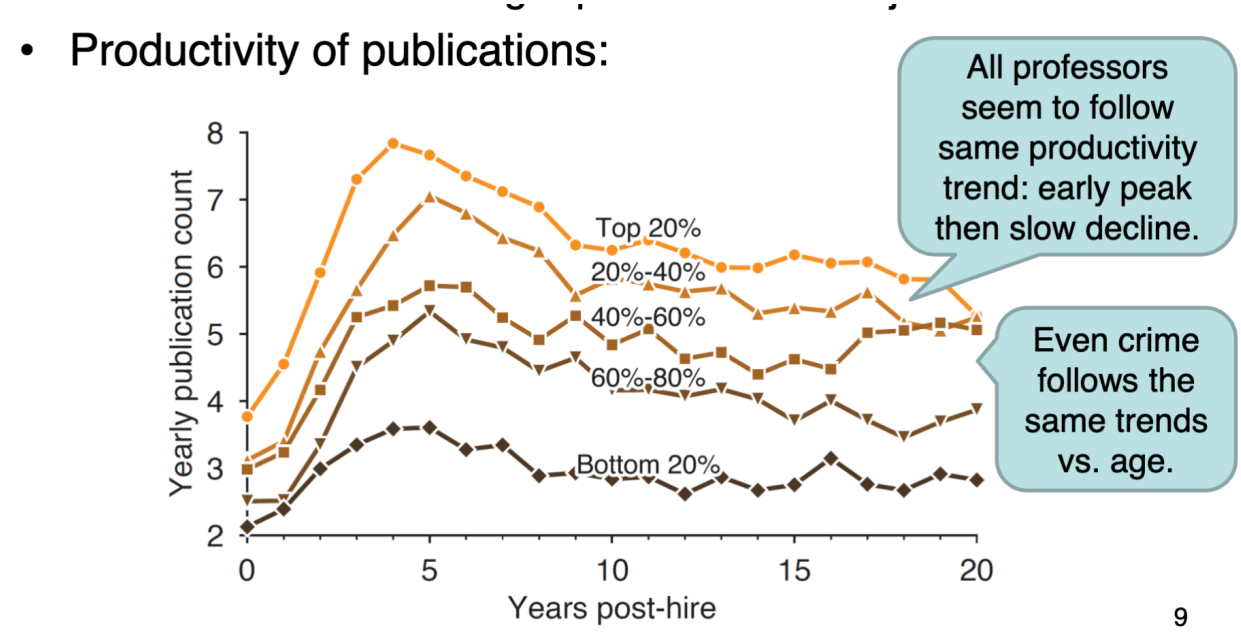
- Normalized by the number of days in each temperature range

Fallacies

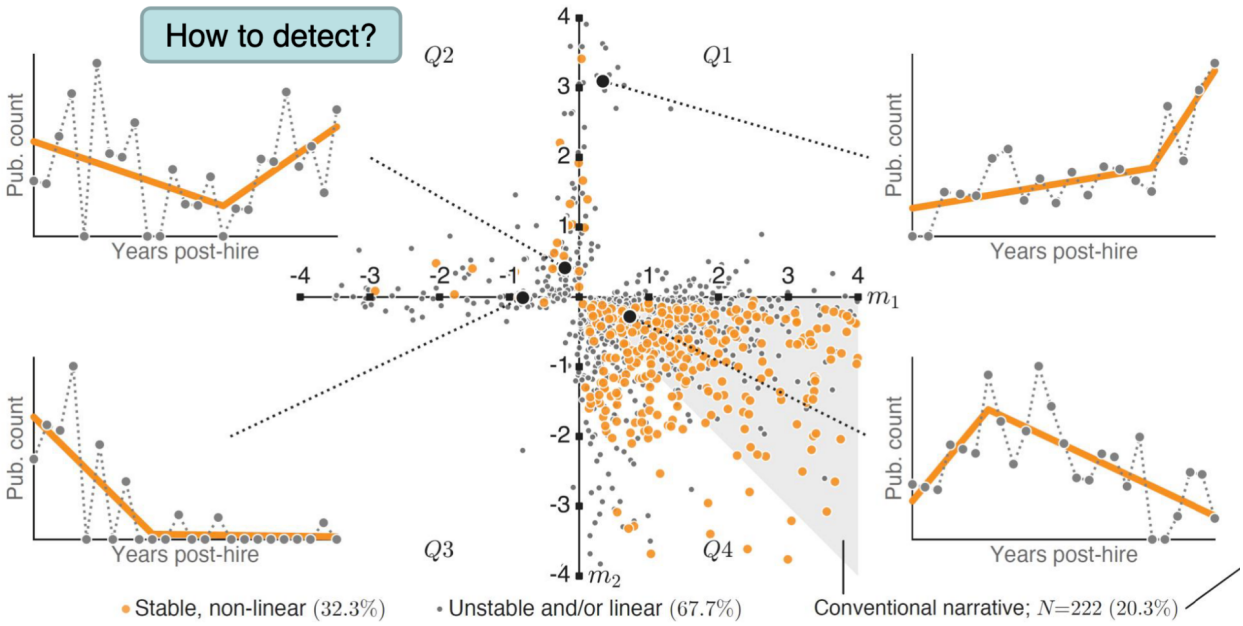
- Don't look at a visualization before you make a hypothesis

1.5 Myth of the Average

- High rate of plane crashes in the 40s
  - Seats built for "average" pilot were non-adjustable!
- Productivity of publications:
- We've averaged out the interesting behaviour of the data



If plot early and later career productivity, a more complex picture of productivity types emerges...



x axis m1, y axis m2. m1 is positive, m2 is negative.
We didn't see this many different behaviours because of averaging the data.
There are smaller subsets of trends that get shadowed by the majority trend

1.6 Simpson's Paradox

Possibly the most problematic issues in data analysis – unobserved variables and/or sample bias can completely change your interpretation and decision!

	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

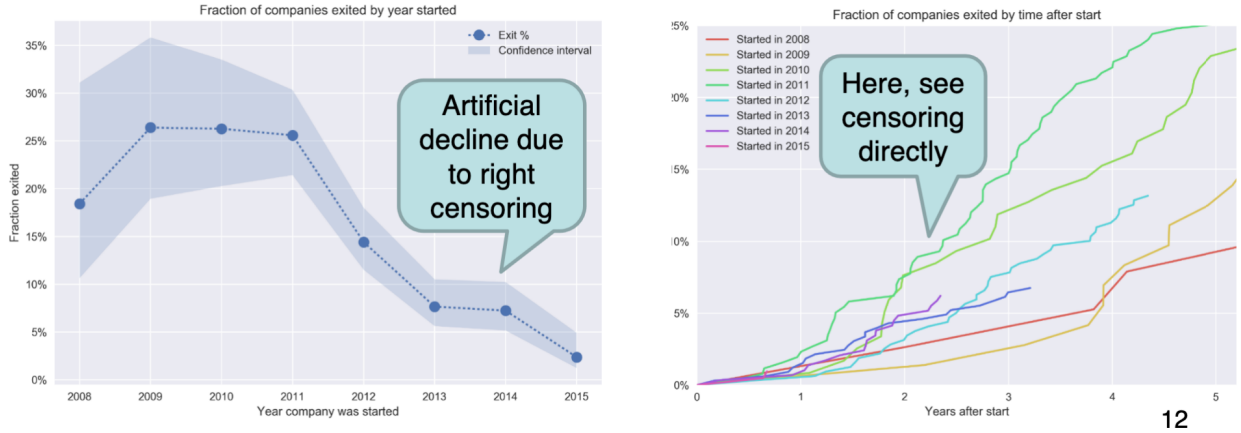
How do resolve this? – Which answer is correct depends on more knowledge
Aggregating the data can lead to a different conclusion than looking at the data in its entirety. There is causal interference in the data. Certain doctors may prescribe differently. Treatment A is more effective but more doctors prescribed treatment B to small stones. Small stones generally have a higher recovery rate compared to large stones in general.
Look at Wikipedia page on Simpson's paradox that has examples that can be on final exam.

1.7 Plotting "Right Censored" Time Series Data

when trying to measure an effect, we need to be aware of the horizon effect with time. if you don't measure long enough you may not see that effect.

- Be aware of horizon effects with time

- "right censored": event may not have happened in all samples
  - Or data may be incomplete (people left study, lost track, etc.)
- Especially prominent in "survival/failure analysis"
  - How long an engine lasts, how long a patient survives, etc.
- For plotting: need to make incompleteness in data clear...



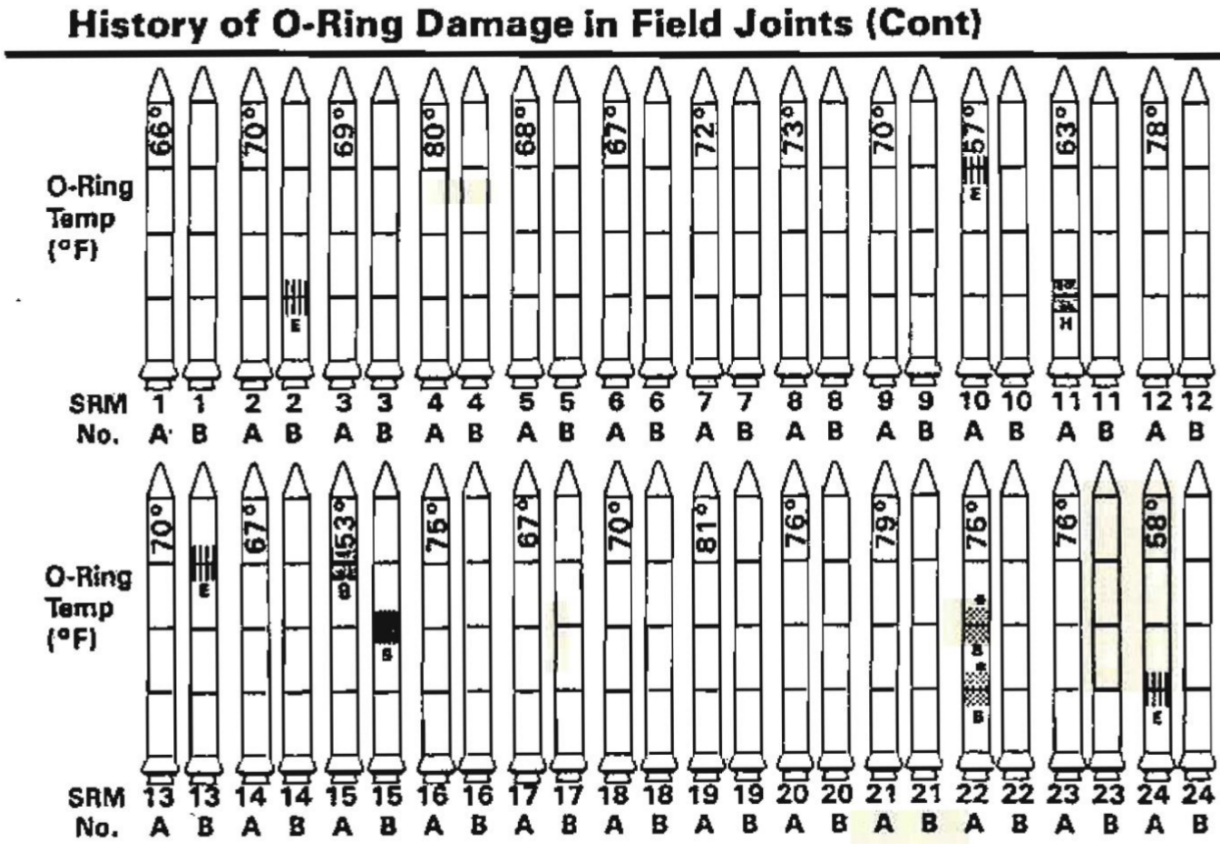
companies that started later have fewer years to exit compared to older startups.

2 Challenger Disaster

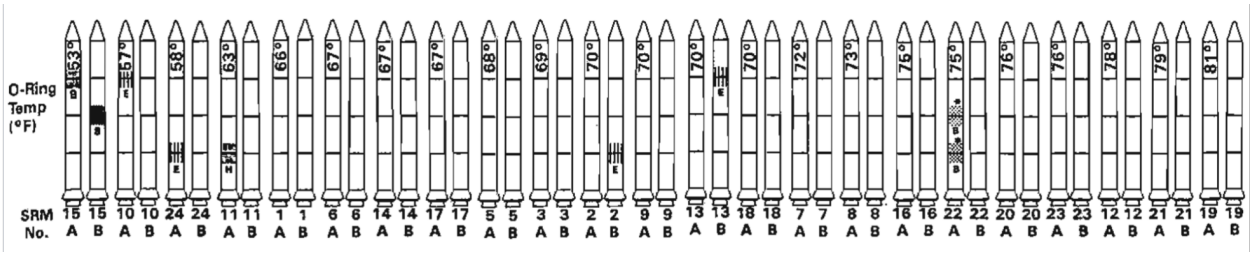
2.1 Tufte's Challenger Example

Should the Space Shuttle Challenger have been launched when the launch temperature was -1C?

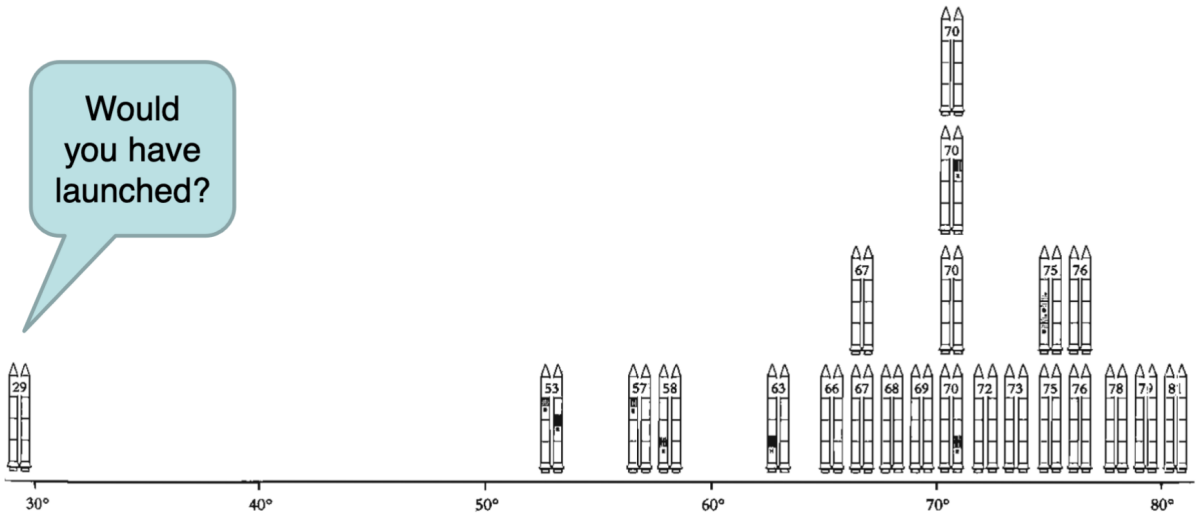
- Hashes show post-launch damage evaluation.
- This chart does not help us answer the question. Why not?



If temperature is what concerns us, let's reorder by temperature:



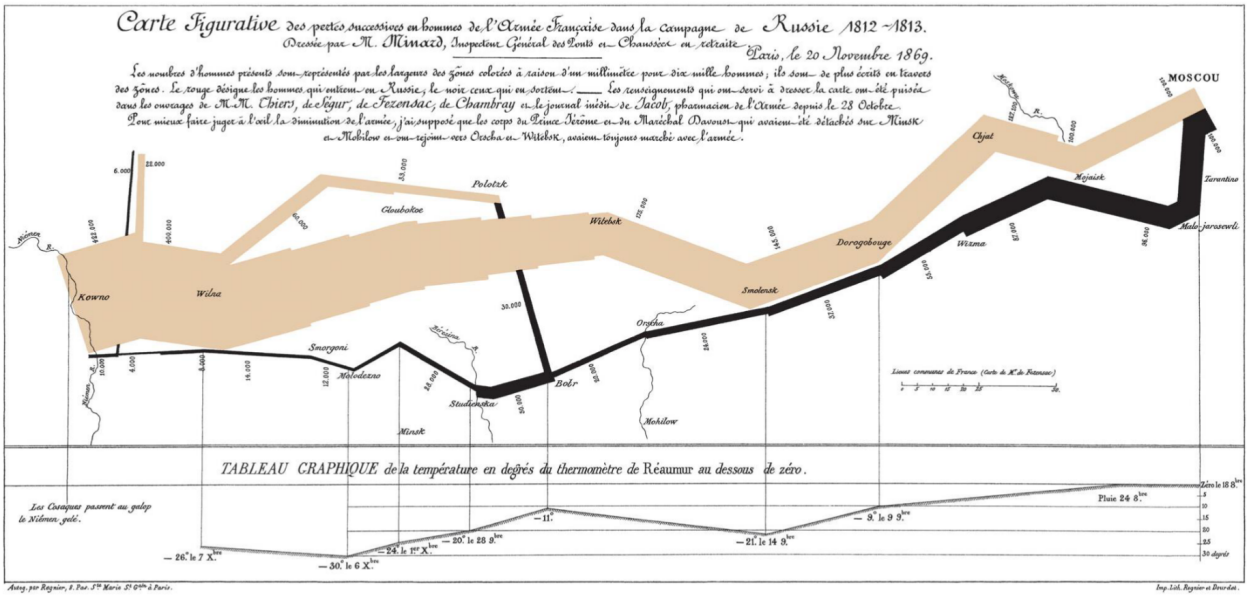
OK, some concentrated damage at lower temperatures, but not highly conclusive.
Let us accurately represent the temperature scale and try one more time:



3 Custom Visualization

3.1 Minard's Map of the 1812 Russian Campaign

Sometimes you have to construct your own visualization...



Modern redrawing of Charles Minard's map of Napoleon's disastrous Russian campaign of 1812. The graphic is notable for its representation in two dimensions of six types of data: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates

4 Custom Interactive Visualization

Hans Rosling's Global Temporal Analysis of Life Expectancy and GDP

You want the area to be proportional to the population of the country to better scale quadratically than radius which is linear.

5 Debugging Data-oriented Code

For machine learning or complex data science analysis

5.1 Synthetic Data for Debugging

- Your code will have bugs!
- How to debug data analysis / machine learning?
  - Make a synthetic dataset
    - E.g., (Features x: Age, Gender; Label y: Snapchat, Facebook)
      - Everyone under 20 uses Snapchat
      - Everyone 20 and over uses Facebook
      - Random selection for gender
      - or use label as a feature
    - Does your analysis uncover the expected trends?
    - What happens as you add noise?
      - 90%, 80%, 70% of people under 20 use Snapchat?

Inability to uncover expected trends indicates a bug in your code or the need to revisit your analysis approach or hypothesis space.