From Rules to Reasoning: Building Effective Visualizations for Complex, Real-World Data

Xiaoyi Yang
Khoury College of Computer Sciences
Northeastern University

Nov 14th, 2025

About me

- BS in Statistics and Math at University of Wisconsin-Madison
- Master and PhD in Statistics at Carnegie Mellon University
- Tenure-track Assistant Professor at Math department, Creighton Univeristy
- Teaching-track Assistant Professor at Khoury College of Computer Sciences, Northeastern University
- Thesis: *Learning social networks from text data*
  using covariate information
- Teaching: *Information presentation and visualization*
  Foundations of Data Science
  Machine Learning

Visualization in Statistics and CS

When we talk about the visualization, what will be the first three terms in your mind? Or, try to use a sentence to explain what is a visualization?

Visualization in Statistics and CS

In Statistics

- Visualization as a tool for **exploratory data analysis (EDA)**.
- Focus on data integrity, distributional patterns, outliers.
- Aims to **summarize** uncertainty and statistical inference.
- Grounded in **graphics for reasoning and model diagnostics**.

In Computer Science

- Visualization as an **interactive system** or **interface**.
- Focus on **scalability**, **interactivity**, and system performance.
- Aims to **support human–computer interaction** and sensemaking.
- Integrates with ML, databases, and web technologies.

Getting into deeper understanding of visualization

In the classroom: To compare student's GPA across different majors, which figure you can use?

Getting into deeper understanding of visualization

In the classroom: To compare student's GPA across different majors, which figure you can use?

In the reality: A music producer wants to track the performance of three singers in his company. He hopes to get an overall sense of how the downloads have been performing in the past 12 months. He's particularly curious to see if the singers' popularity overlaps — for example, whether an increase in one singer's downloads corresponds to a drop in another's. He also wants to compare how different streaming platforms contribute to the overall downloads and identify if any song has ever become a "hit of the day" on a particular service. Finally, he wants to know which five songs are the most profitable. The producer is looking for a simple and clean summary that he can easily share with the management team.

Task abstraction

First, understand what the client is saying...

- What data you will need to fulfill the requirement?
- What visualization you can make?

Task abstraction

Text	Analytical Tasks
1. Understand overall performance trends over time	Show 12-month trend of total downloads
2. Compare singer popularity & overlapping popularity	Compare downloads between 3 singers over time; Part-to-whole breakdown by singers
3. Compare streaming platform contribution	Part-to-whole breakdown by platform
4. Detect "hit of the day"	Identify local peaks
5. Find top 5 most profitable songs	Ranking task

Then, it seems that you have a plan

Analytical Tasks	Visual Encoding
1. Show 12-month trend of total downloads	Line chart
2. Compare downloads between 3 singers over time; Part-to-whole breakdown by singers	Multi-line chart; small multiples of stacked bar plots
3. Part-to-whole breakdown by platform	Stacked bar (value or percentage)
4. Identify local peaks	Highlight peaks; annotate spikes
5. Ranking task	Sorted bar chart

Remember, you also need a clean, simple summary show to the people who are not in data science field.

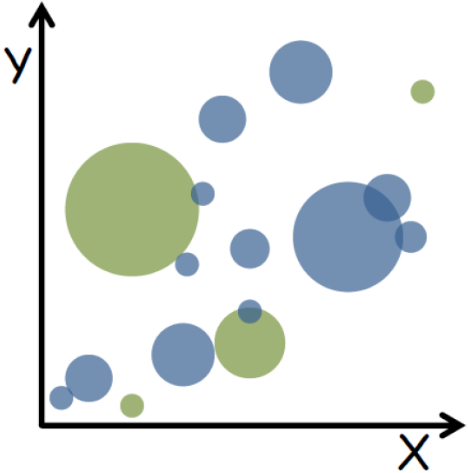
The question is why?

- Why we can match the figures to our task?
- Why and maybe when we can combine some of the figures?
- Why we need to choose from the current visualization choices? Can we design new visualization to better represent our data?

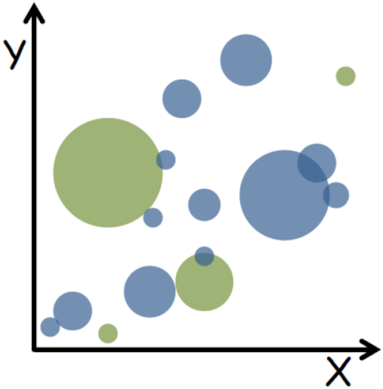
Break down the visualization

In order to answer those questions, we need to break down the visualizations.

What you have seen?



Visualization Building Blocks



Mark = basic graphical element in an image

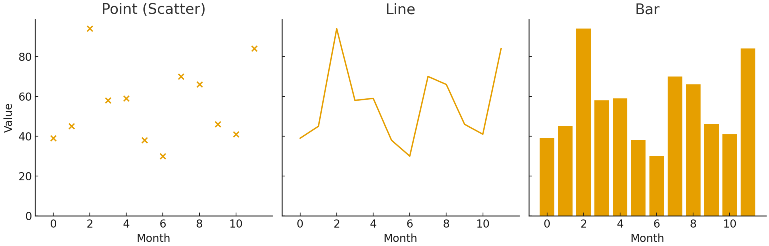
eg: point, line, area/bar, text

Channel = way to control the appearance of independent of the dimensionality of the geometric primitive

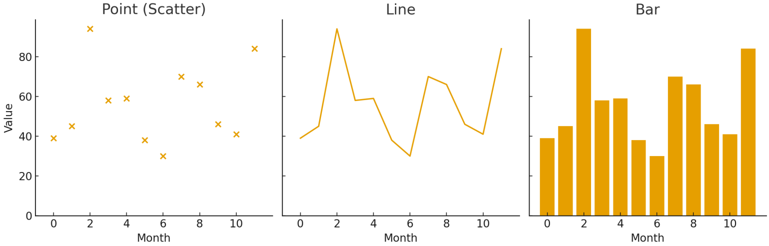
eg: position, length, angle/area, color, size, shape

Previous example: point mark with 4 channels: horizontal position, vertical position, size and color

Marks


Point (Scatter)
Line
Bar
Value
0
20
40
60
80
0
2
4
6
8
10
Month
Month
Month

Marks



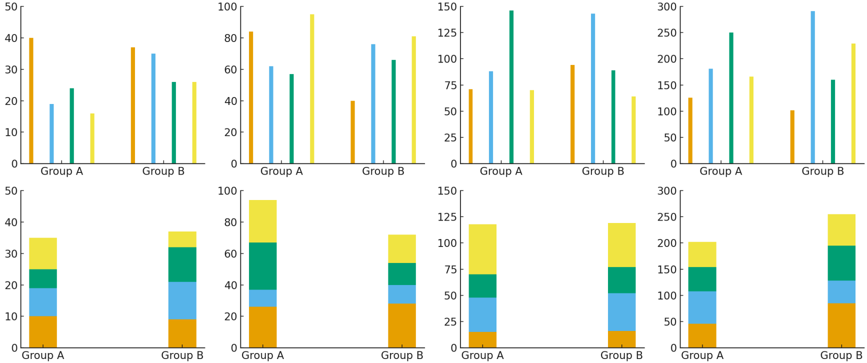
- Mental emphasis = individual values
- Forces precision reading.

- Mental emphasis = trend, direction, continuity
- Values in between are valid, connected and interpolated

- Mental emphasis = magnitude comparison
- Viewers judge length from a baseline

Channels: Cleveland & McGill, 1984

Estimate the values for blue segments.



Result

Here are the answers:

- The first row: 19, 35, 62, 76, 88, 143, 181, 291
- The second row: 9, 12, 11, 12, 33, 36, 62, 43

Calculate your MSE.

In general, which layout is easier to estimate which one is harder?

- Group A vs Group B
- First row vs. Second row

Channel Rank

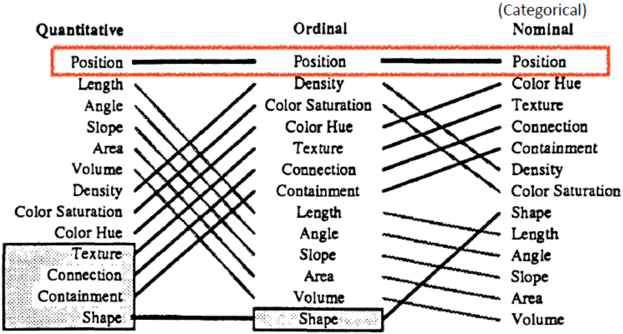


Figure 15: Ranking of Perceptual Tasks. *The tasks shown in the gray boxes are not relevant to that type of data.*

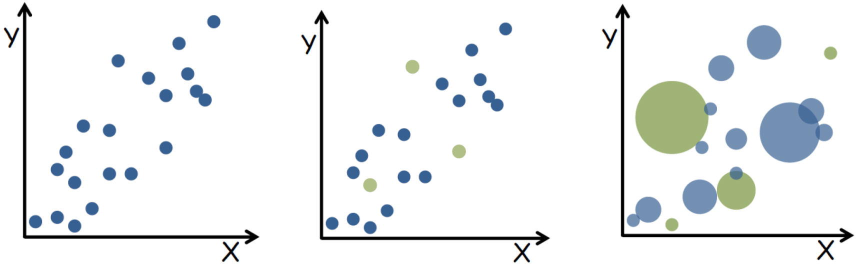
Mackinlay (1986)

When we say, do not use pie chart, it is because...

Combine and create visualizations

- Once you choose a mark, you can add additional dimensions using channels as long as you do it with care and maintain perceptual clarity.
- You can combine views effectively when they share the same primary channel encoding, especially the position channel.

Add additional dimensions



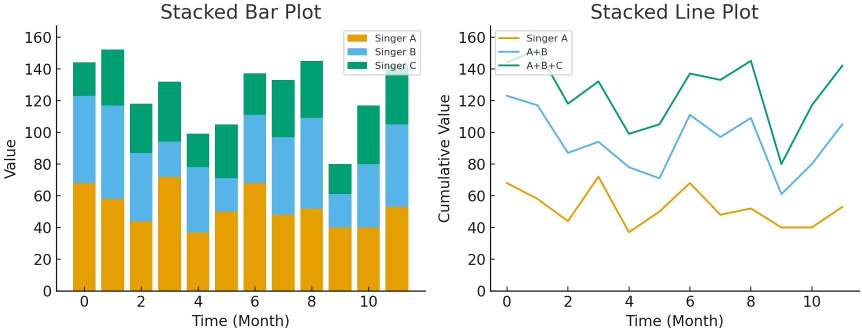
Combine and create visualizations

Recall the second task in our example: Compare downloads between 3 singers over time and part-to-whole breakdown by singers.

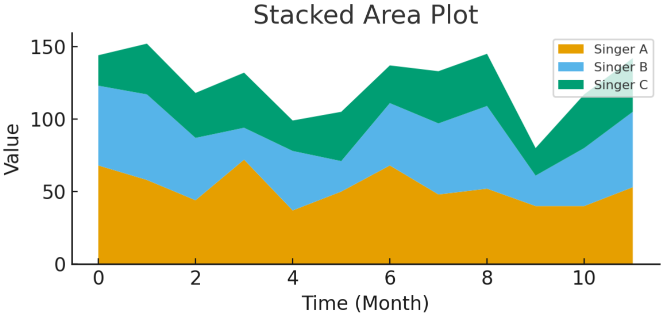
Consider you do have these information:

- download number
- singer
- date

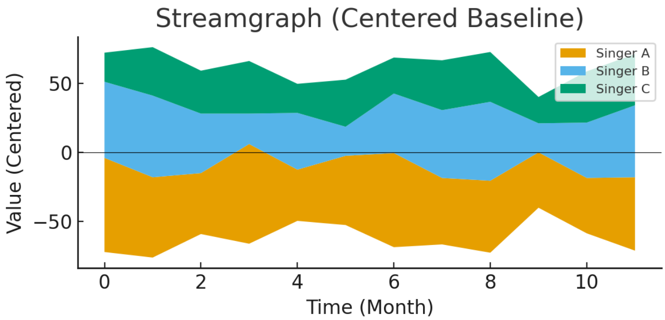
First proposal



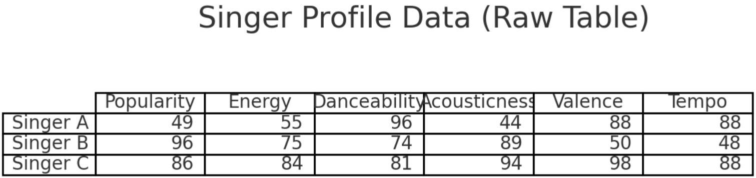
Make some changes



Final version



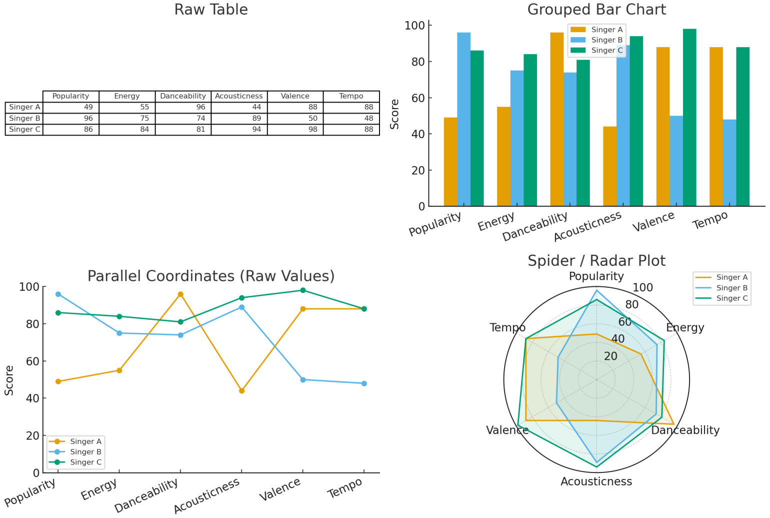
Some other examples



Singer Profile Data (Raw Table)

	Popularity	Energy	Danceability	Acousticness	Valence	Tempo
Singer A	49	55	96	44	88	88
Singer B	96	75	74	89	50	48
Singer C	86	84	81	94	98	88

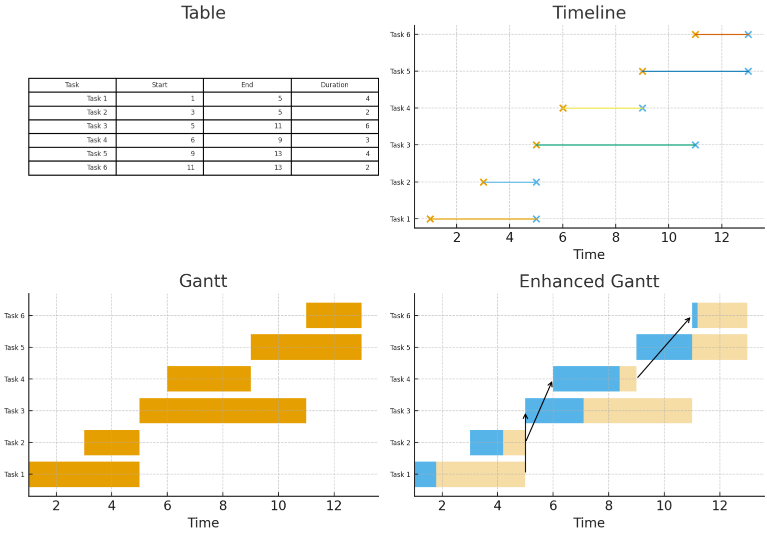
Some other examples: Spider



Raw Table

	Popularity	Energy	Danceability	Acousticness	Valence	Tempo
Singer A	49	55	96	44	88	88
Singer B	96	75	74	89	50	48
Singer C	86	84	81	94	98	88

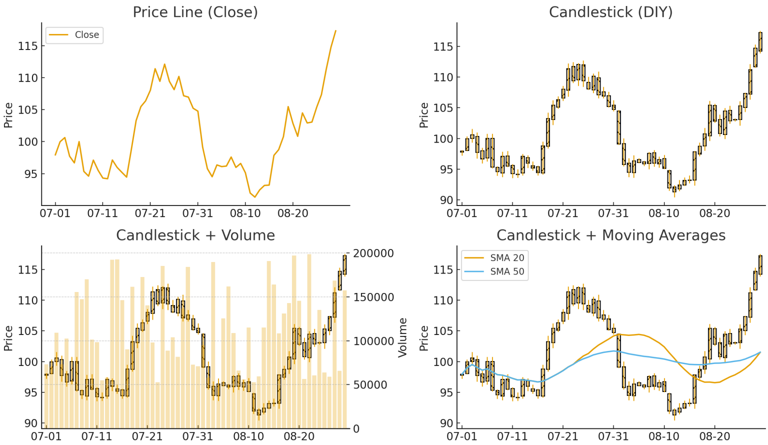
Some other examples: Gantt



Table

Task	Start	End	Duration
Task 1	1	5	4
Task 2	3	5	2
Task 3	5	11	6
Task 4	6	9	3
Task 5	9	13	4
Task 6	11	13	2

Some other examples: Finance



Price Line (Close)

Candlestick (DIY)

Candlestick + Volume

Candlestick + Moving Averages

Take away

- Visualization is not decoration — It's a reasoning tool
- Choose visualization by task, not by habit. Think beyond the chart
- Visualization is a conversation with the data. Embrace design evolution
- Don't fear approximation. Develop intuition vs. Estimate precision.

When the scale increases

1 million Spotify data: https://www.kaggle.com/datasets/amitanshjoshi/spotify-1million-tracks

- Over 60k singers in 24 years, with 82 different genre
- No single strong correlation between any numerical features
- Tech limit

General strategies

Reduce before you render

- Aggregate: bins (2D heat/hex), rolling summaries, top-k
- Filter: categories, time windows, ROI
- Downsample: for previews; keep aggregated totals authoritative

Interaction for scale

- Brush & link: select in one view, highlight in another
- Zoom & pan: preserve overview + focus (mini-map)
- Progressive disclosure: details on demand; pinned comparisons

General strategies

Reduce before you render

- Aggregate: bins (2D heat/hex), rolling summaries, top-*k*
- Filter: categories, time windows, ROI
- Downsample: for previews; keep aggregated totals authoritative

Two other options...

Interaction for scale

- Brush & link: select in one view, highlight in another
- Zoom & pan: preserve overview + focus (mini-map)
- Progressive disclosure: details on demand; pinned comparisons

Story-line

- The Classic Three-Act Structure: Setup, Conflict/Discovery, Resolution
- The Mystery/Detective Story
- The Journey/Quest Structure
- The Comparison/Duality Structure
- The Spiral/Zoom Structure
- The Time Loop/Circular Structure

Or, “interactive” plot

How people interact with such a large dataset is also interesting...

https://xiaoyiyang83.github.io/dynamic-demo/

Take away

- Accept that “Complete Accuracy” isn’t the goal. What scale of truth is still meaningful for the audience?
- Make visualization a two-Way process. The visualization can be interactive and the interaction itself can be the data.
- Visualization is not an endpoint, it’s an instrument: for storytelling, for exploration, for human-in-the-loop data collection.

Questions

The slides and interactive codes will be in the GitHub:
https://github.com/xiaoyiyang83/dynamic-demo

Email: xiaoy.yang@northeastern.edu