Multivariate data visualization and the limits of human perception

maxluster Oct 16, 2018 · 9 min read

The human vision system is able to process an incredible amount of data in the blink of an eye, but there are limits to the amount of information we can efficiently extract from a graphic. No matter how well designed a visualization is, our ability to *understand the data*—to connect meaning to shapes, colors, sizes, positions, trends, clusters, outliers — deteriorates as more and more variables are represented.

Our perceptual limitations are at odds with the complex, multivariate systems that surround us. Data visualization expert Edward Tufte puts it best:

> "All the interesting worlds (physical, biological, imaginary, human) that we seek to understand are inevitably and happily multivariate in nature."
>
> Edward Tufte, *Envisioning Information*

Despite our limitations, multivariate systems are critical for us to understand. Thoughtful analysis of complex systems can change the direction of technology, science, public discourse, and policy. As data visualizers, designers, analysts, scientists, it's our role to push against the limits of human perception — to do our best to make complex systems clear.

In this article I'm going to walk through an analysis of one important system, *economic opportunity in the United States*. Using this system as a case study, we can explore the limits of our perception.

To facilitate our exploration, I've created a series of data visualizations. With each subsequent visualization, one additional variable will be depicted. I'll provide analysis but please feel free to follow along, examine the graphics yourself, and take note of what you can see (and what you cannot).

The data comes from a longitudinal study conducted by Opportunity Insights covering nearly the entire U.S. population. With this unique dataset we can examine relationships between:

1. Parent household income rank
2. The average income rank of their children born 1978–1983
3. Race
4. Gender

I'll begin by visualizing a single variable, parent household income rank, which has been organized into percentiles.

One variable

Parent household income rank

Design

To show the range of parental income ranks in the data, I'm going to represent each percentile with a mark — a point, depicted by a small circle. I'll draw one mark for each percentile and set each mark's position along the x-axis in even intervals. Labels will be rendered in light grey to keep attention on the data.

https://beta.observablehq.com/@maxluster/parental-income-ranks

Analysis

This is a single variable data visualization — it doesn't tell us much. However, from the position of the points we can quickly see that parental incomes were organized into 100 even interval percentile ranks.

Two variables

Average child income rate by parent household income rank

Design

To visualize the relationship between parent income rank and our second variable, average child income rank, I'm going to position each mark along the y-axis. I'll also add a reference line to help us see income mobility between generations. Points above the line will indicate upward mobility — where average child income ranks exceeded parent incomes; points below the line will show us downward mobility.

https://beta.observablehq.com/@maxluster/avg-child-household-income-rank-by-parent-household-income-

Analysis

The dots form a curve in our mind. The *slope* and *shape* of the curve can be rapidly perceived and compared with the reference line. These visual signs can be mapped to conceptual meaning:

Slope: The slope of the curve indicates the strength of the relationship between parental income rank and average child income rank. A quick comparison with the

reference line shows us a regression towards the mean — on average, children raised in poor households had higher incomes than their parents, children raised in wealthy households had lower incomes.

Shape: The shape of the curve indicates the change in the strength of the relationship across the economic spectrum. The increasing slope of the tails shows us that children that grew up in households at the extremes of wealth and poverty were more likely to remain in their economic situation compared to their peers from middle class households.

Three variables

Average child income rate by parent household income rank and race

Design

This dataset has the following designations for race.

- Native American
- Asian (non-Hispanic)
- Black (non-Hispanic)
- White (non-Hispanic)
- Hispanic

To encode information about race, I'm going to manipulate the colors of the marks. Because race is a categorical variable, my goal is to select a set of colors where the intensity of the colors is constant, but the hue of the colors are maximally different from one another and can be easily distinguished.

https://beta.observablehq.com/@maxluster/avg-child-household-income-rank-by-race-and-by-parent-househ

Analysis

The marks of each color easily group together and separate from marks of other colors in our mind. They form 5 distinct curves that we can rapidly examine and compare. In addition to our ability to quickly evaluate the slope and shape of each curve, now we can compare their *position* to see the difference in overall outcomes for each racial group.

We can view all 500 points as a single image and form a macro understanding of the data — race played a key role in determining the opportunity children had towards economic success no matter how poor or wealthy their childhood home. Additionally, we can visually select any one of the curves and subject it to the same detailed analysis of *slope* and *shape* as in our two-variable visualization. We can easily see differences in:

- *Position:* Income ranks for children of different racial groups were substantially different across the entire spectrum of parental incomes
- *Slope:* The relationship between parental income rank and avg. child income rank was much stronger for some racial groups than others
- *Shape:* The strength of the relationship increased at the extremes for all racial groups, but this effect happened at different rates for each race.

Four variables

Average child income rate by parent household income rank, race, and gender

Design

This dataset reports average child income rank, comparing outcomes for black and white men and women. Data for other groups was not available.

Since we've already used color to represent race, I'm going to differentiate the marks by shape to encode information about gender. My goal is to select shapes that appear to be the same size but are easily differentiated from one another.

https://beta.observablehq.com/@maxluster/avg-child-income-rank-by-race-parent-household-income-rank-a

Analysis

Comparing the position of the curves, we can quickly see that white men tended to have higher incomes than any of our other groups. But what else can we see?

When examining the data further, notice how you need to mentally pause. What question do you want to investigate first? Do you look at the blue marks vs. the orange marks? The circles vs. the triangles? We can't make all possible comparisons in a single instant — we must toggle between comparisons in our mind, building a understanding of each in turn.

At first glance your vision system most likely prioritized grouping the marks by color (race). When a visualization encodes variables by both color and shape, color takes priority for our attention. However, if you focus your attention on shape, it's possible to visually compare the group of *triangles* (gender: female) to the group of *circles* (gender: male). Then, if you visually select a single shape, triangles for example, you can separate the triangles by color to compare *yellow* (race: black) triangles to *blue* (race: white) triangles.

> *When designing multivariate data visualizations we must make decisions about which relationships will be easiest to see, which will require additional mental effort, and which will be out of reach entirely.*

The challenge of envisioning the information would be amplified if the visualization depicted more data. For example, if we wanted to show all five of the racial groups (for a total of 1,000 points) we would probably have major issues with marks occluding other marks. And what if we wanted to depict a fifth variable, how would we visually encode it? Would we use size, texture, color intensity? We *could* use one of these visual variables to create an even more dense multivariate visualization. Alternatively, we could take a good design that depicts two or three variables and multiply it.

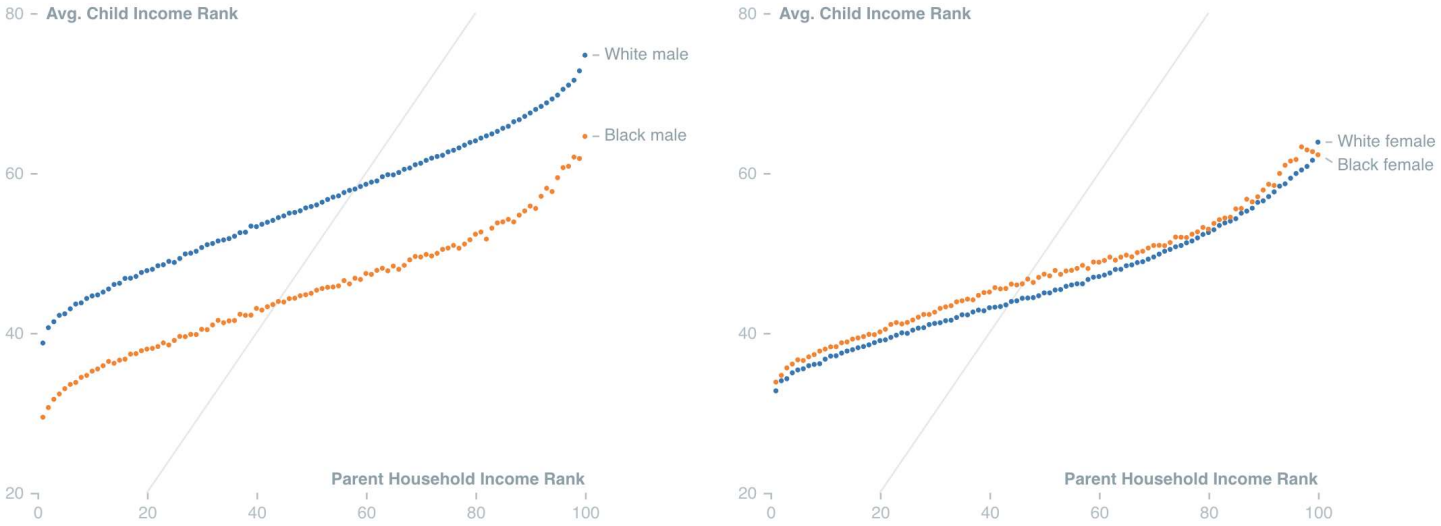
Small multiples

A different approach to visualizing average child income rate by parent household income rank, race, and gender

Design

Lets take a different approach to visualizing all four of our variables.

Rather than rendering a single chart, I'm going to split the data across multiple charts. I'll visualize three variables within each chart — parental income ranks, child income ranks, and race (black/white) — and create multiples of that chart by gender, leaving us with two charts to compare side-by-side. My goal is to have the new charts be small enough to comfortably fit into your field of view, line up on a grid, and share the same data frame (axes).



https://beta.observablehq.com/@maxluster/avg-child-income-rank-by-race-parent-household-income-rank-a/2

Analysis

As with our four variable single chart, we still can't make all possible comparisons in a single instant. However, small multiples give us a way to move efficiently through a sequence of comparisons.

In the chart on the left we can see that no matter how wealthy their parents, on average white men secured significantly higher incomes than black men.

In the chart on the right we can see that for all but those with parents in the top 1% of income, white women secured slightly lower income ranks than black women. And we can see an unusual deviation from the typical shape of the curve in the series for black women — for some reason, income ranks declined for children raised in households with income in the top 3%.

> *Seeing the charts side-by-side enables a more expansive analysis of the data — there was a broad gap in income ranks between white and black people that was explained by income differences in men, not women.*

Why were outcomes for black men so poor relative to white men given the same economic opportunities at childhood? Why was there such a large income gap between black and white men, yet incomes for black and white women were much more similar? These questions are outside the scope of this article, but are pressing questions the analysis compels us to ask.

We can see now that a single multivariate visualization can allow us to distill a complex and abstract idea like '*economic opportunity in the United States*' into tangible ideas that make us feel, think, and act differently.

Pushing against the limits of human perception

We're creatures that see the world in three dimensions, and it's most likely for that reason that well designed visualizations of one, two, or even three variables can be easily digested. Beyond that limit, we've seen that visualization design gets trickier.

However, pushing against the limits of human perception is an essential task. A clear picture of the relationships between *parent income, child income, race,* and *gender* can shape the way we see the world; it can influence policy and public discourse. What other stories can we tell with data? What other systems need to be examined?

Perhaps the fundamentals of human perception will remain forever as they currently are, but our display technologies and data visualization techniques can continue to improve. With better visualizations we can deepen our understanding of the fascinating, multivariate worlds around us.

References

- **Opportunity insights —** A non-partisan research organization with a mission to "develop scalable policy solutions that will empower families throughout the United States to rise out of poverty and achieve better life outcomes." Data from their research and much deeper analysis than provided in this article can be found here.

- **Semiology of Graphics: Diagrams, Networks, Maps**, *Jaques Bertin* — Explains data visualization through semiotics. Examines human perception and our ability to map visual variables (e.g. color, shape, size, position) to variables in the data.

- **Envisioning Information** and **The Visual Display of Quantitative Information**, *Edward Tufte*— Essential guides to visualization design.

Sign up for The UX Collective Newsletter

By UX Collective

A weekly, ad-free newsletter that helps designers stay in the know, be productive, and think more critically about their work. Take a look.

Get this newsletter

You'll need to sign in or create an account to receive this newsletter.

Data Visualization Semiotics Equality Design UX