Handout 6

Data Visualization: Good Vs Bad

Data visualization is a great way to represent huge amounts of data in a simple and intuitive fashion. All data visualizations have the same goal: help viewers easily grasp information to make quick inferences or decisions. However, it is important that visualizations are not overdone and hit the sweet spot where they are catchy, informative, and easy to navigate.

Data visualizations allow people to readily explore and communicate knowledge drawn from data. Visualization methods range from standard scatterplots and line graphs to intricate interactive systems for analysing large data volumes at a glance. But how can we craft visualizations that effectively communicate the right information from our data? What aspects of data and design need to come together to develop accurate insights? The answer lies in the way we see the world: People use their visual and cognitive systems (i.e., our eyes and brain) to extract meaning from visualized data. However, flashy visualizations are not always optimized to help people see what matters. This handout reviews common visualization practices that may inhibit effective analysis, why these designs are problematic, and how to avoid them. The discussion illustrates a need to better understand how visualizations can support flexible and accurate data analysis while mitigating potential sources of bias.

This requires a bit of learning. Putting up a good data visualization is not just a matter of throwing together some data in colorful charts.

A Primer in Visualization: When, Why, and How

Visualizations are powerful tools for discovering and communicating insights in data. However, visualizations are not always necessary—people are not optimized to compute precise statistical quantities from abstract images. Many analysis problems can be solved with direct queries and algorithmic methods. For example, statistical models allow companies to optimize shipping procedures. Purely computational approaches scale further and more accurately estimate precise quantities than people. If you can distill what you need to know about your data into one computable value, you likely do not need a visualization.

However, visualizations often prove robust where statistics fall short. Visualizations take advantage of the universality of visual structure: We can see the shapes these data points make even when we cannot directly enumerate them. Take, for example, Anscombe's Quartet: four datasets with identical means, variance, correlation, and

regressions (Figure). While these datasets appear statistically identical, visualizing them shows substantial qualitative differences in their structure. Our sight detects these high-level structures within 100 milliseconds of looking at a graph, far faster than the blink of an eye.

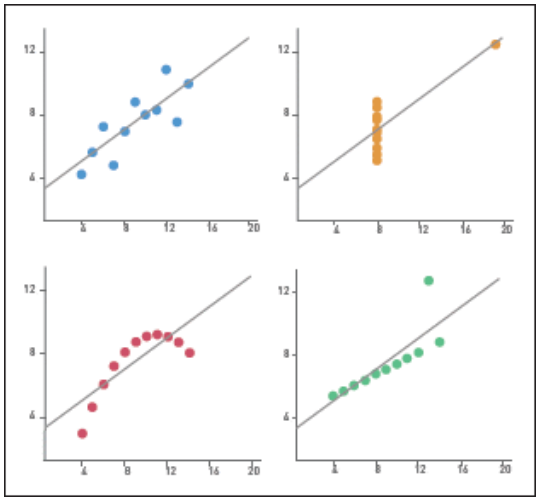


Figure. The four datasets of Anscombe's Quartet share the same basic descriptive statistics, but visualizing these datasets reveals four qualitatively different structures.

How do you decide when to visualize and when to compute? Factors such as uncertainty (how well do statistics represent the data?), transparency (what does the underlying data look like?), context (what additional knowledge could inform analysis and decision making?), scale (how many distinct quantities do we need to evaluate?), exposition (what story must the data tell?), and purpose (do we know what we are looking for?) all help determine when visualizations are valuable. For example, if you cannot readily quantify (or even know) what data properties matter, you can use visualizations to synthesize a diverse set of conclusions. This trade-off between flexibility and precision is often the primary deciding factor for determining when a visualization is necessary: If access to the data underlying a statistic or prediction might change our decisions about that data, we should use a visualization.

Crafting visualizations generally follows a systematic process: clean the data, precompute relevant information, map that information to different visual channels (e.g., position, size, color), and integrate interaction and other details where appropriate. By combining a small number of channels, visualization designers can create intricate interactive systems that reveal patterns in large data collections at a glance. Choosing among these channels, while simple in concept, is where most visualizations go wrong. While many combinations create flashy and engaging graphics, these approaches may inadvertently obscure or even misrepresent data in ways that lead to flawed and biased interpretations. Misleading visualizations appear in our news reports, creating public mistrust in data, in scientific results, leading to incorrect theories, and even in Congress, where policymakers find themselves in

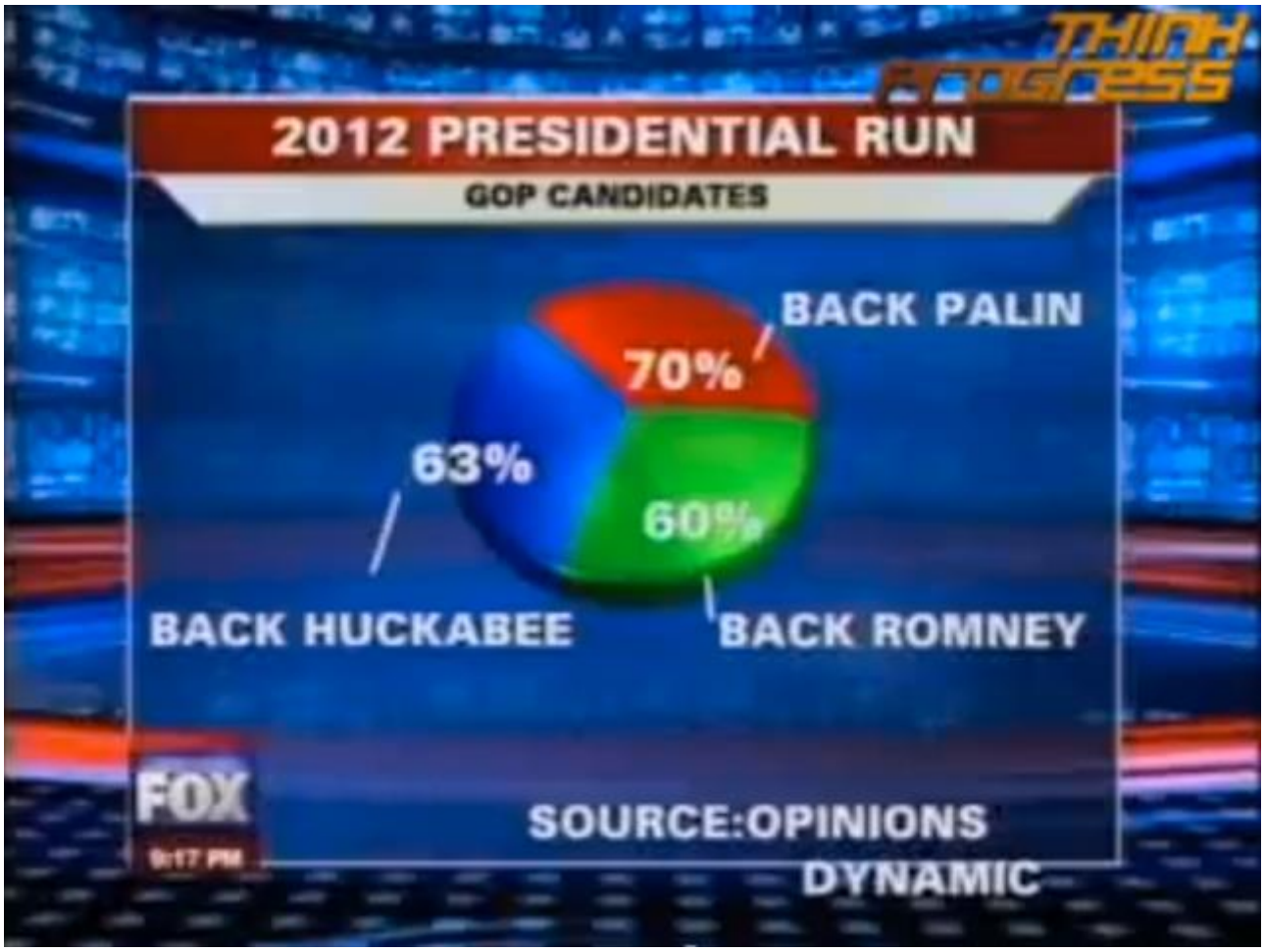
conflict over data. So how do we avoid faulty visualizations? we can start by avoiding well-studied design pitfalls.

Here are 5 common mistakes that lead to bad data visualization. Avoid these to get the most out of your data visualizations.

1. Bad Data

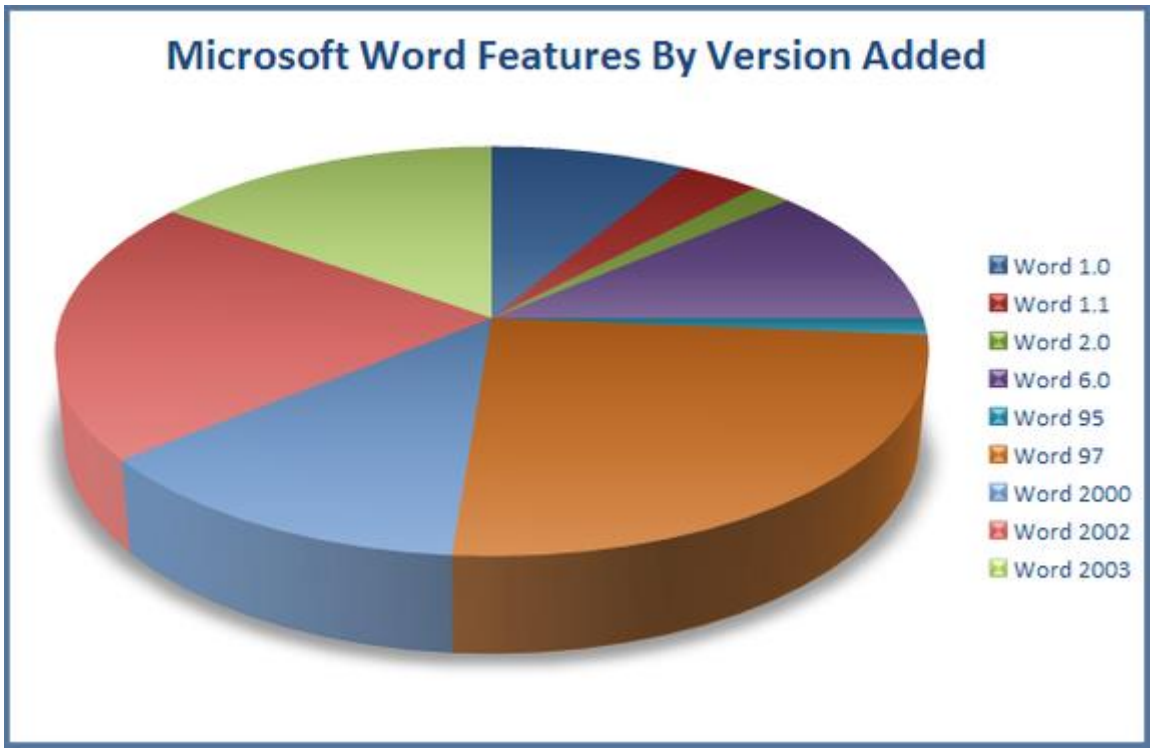
There's an old principle in computer science: "Garbage In, Garbage Out". In the context of data visualization, this means that bad data will lead to bad visualizations. Start with the basics: is your data clean? Use checks at every stage the data goes through collection, sourcing, cleaning, and compiling before it is visualized. Common errors include data duplication, missed data, NA values not marked, and so on.

For instance, in this pie chart, the three sectors of the pie add up to 193%, which makes no sense. Such mistakes in data would render your final visualizations useless.



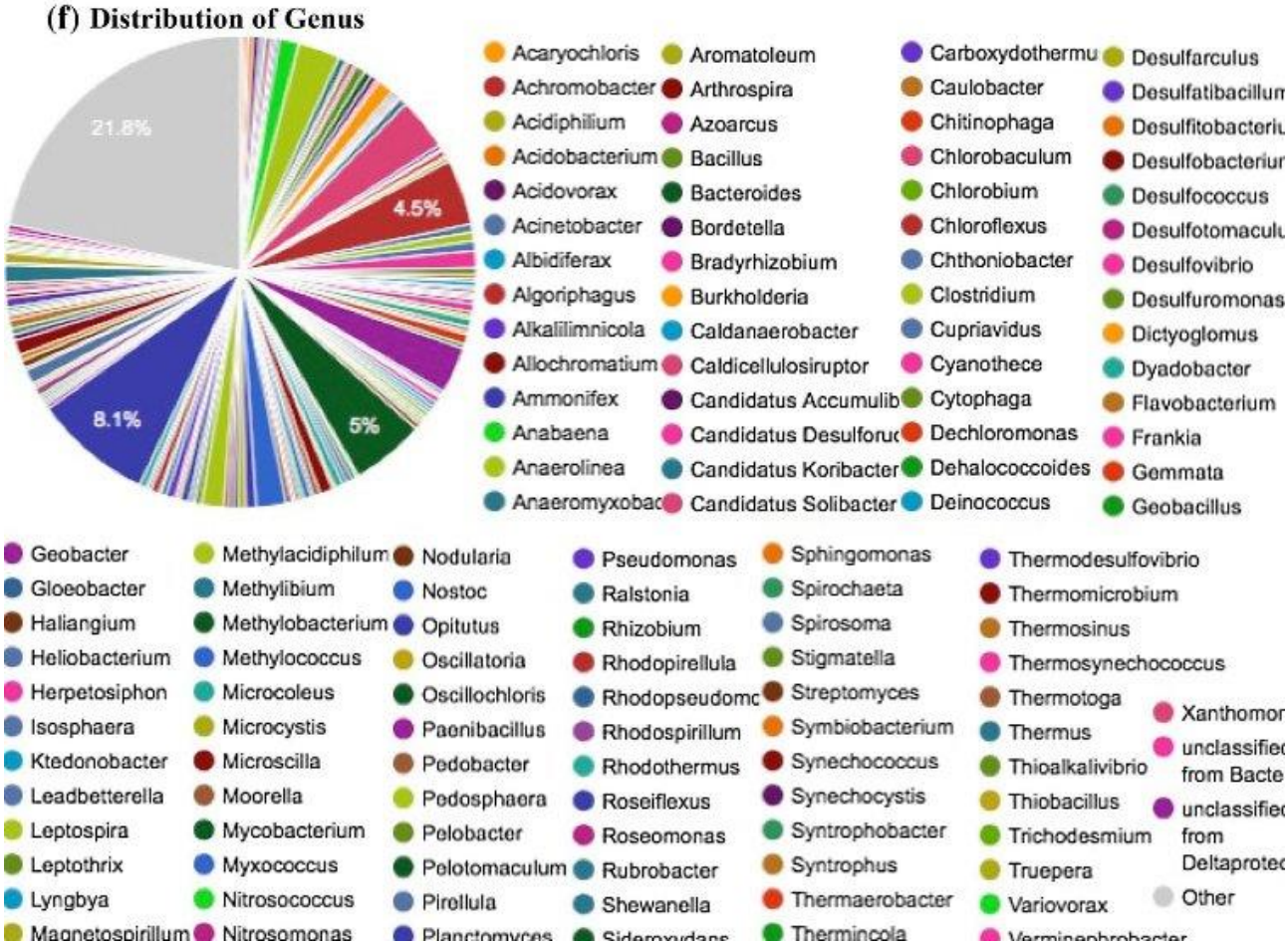
2. Wrong Choice of Data Visualization

Once your data is ready, you should be careful about what type of visualization you use. For instance, in the visualization below, a pie chart was the wrong choice. The intention there was to show how many features a given Microsoft Word version has. The pie chart, on the other hand, shows the proportion of features in a particular version as a percentage of the total features in all versions. A bar chart would be a better choice for this data.



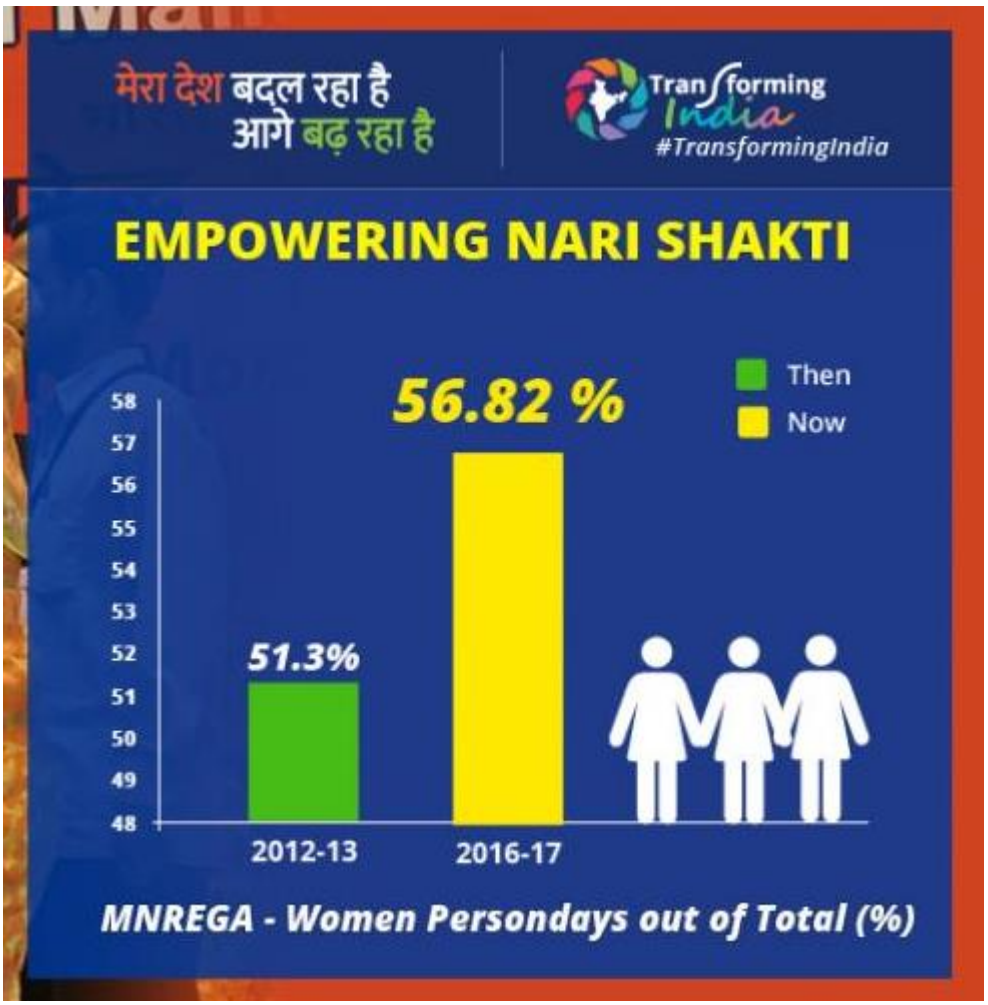
3. Too Much Color or Information

While it has been established that using different colors help people interpret data visualizations quicker, too much color can confuse the viewer. It is important to choose a limited number of colors — 5 is a good upper limit — that are distinct from each other. Thanks to the crazy colors, the visualization below is seriously messy.



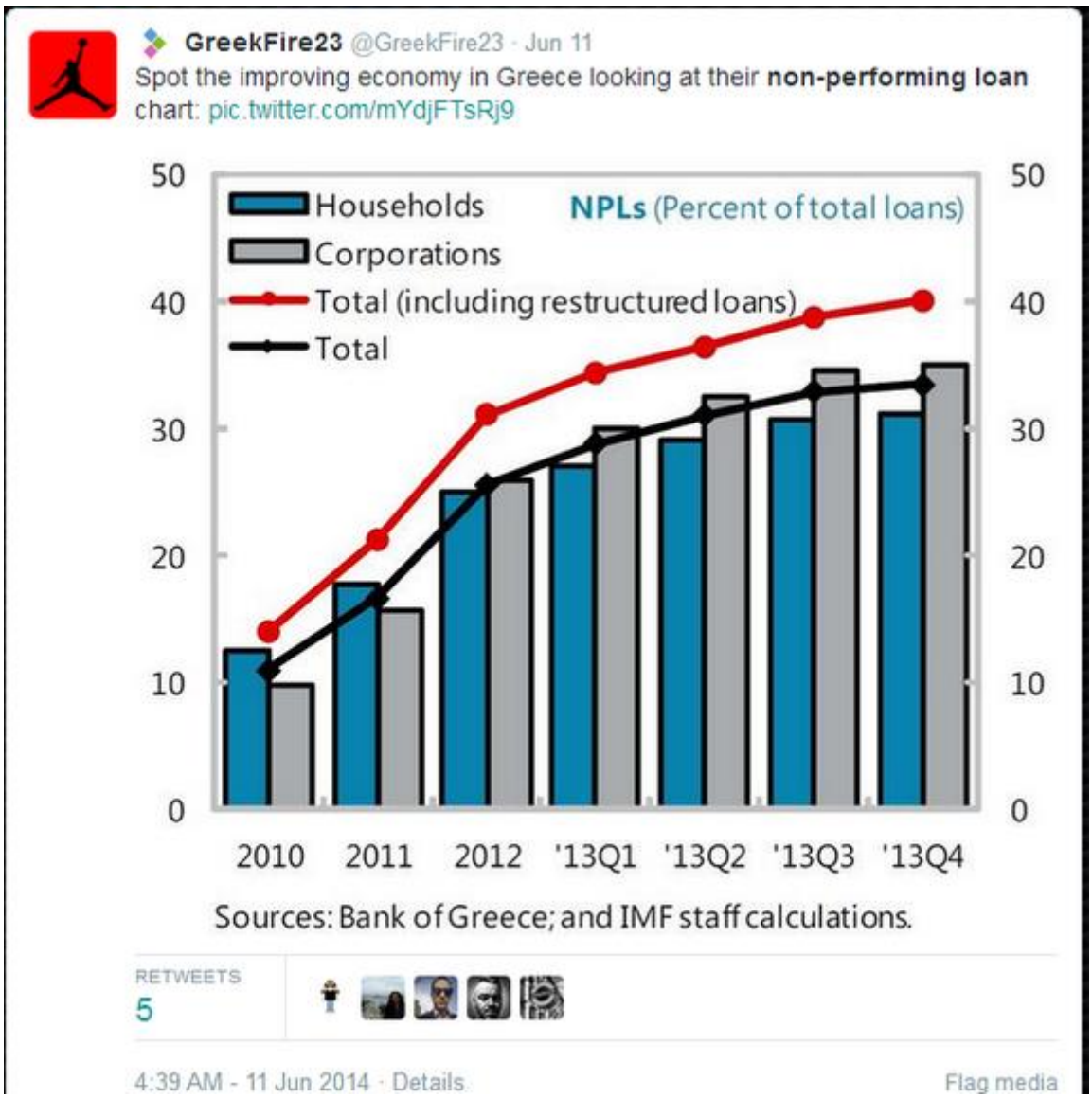
4. Misrepresentation of Data

For instance, this bar chart seems to show that the percentage of women covered under a job guarantee scheme more than doubled from 2012-13 to 2016-17. However, when we look at the y-axis, we see that it begins from 48%, not 0%. This misrepresents the marginal improvement of around 5.5% as a 2x increase.



5. Inconsistent Scales

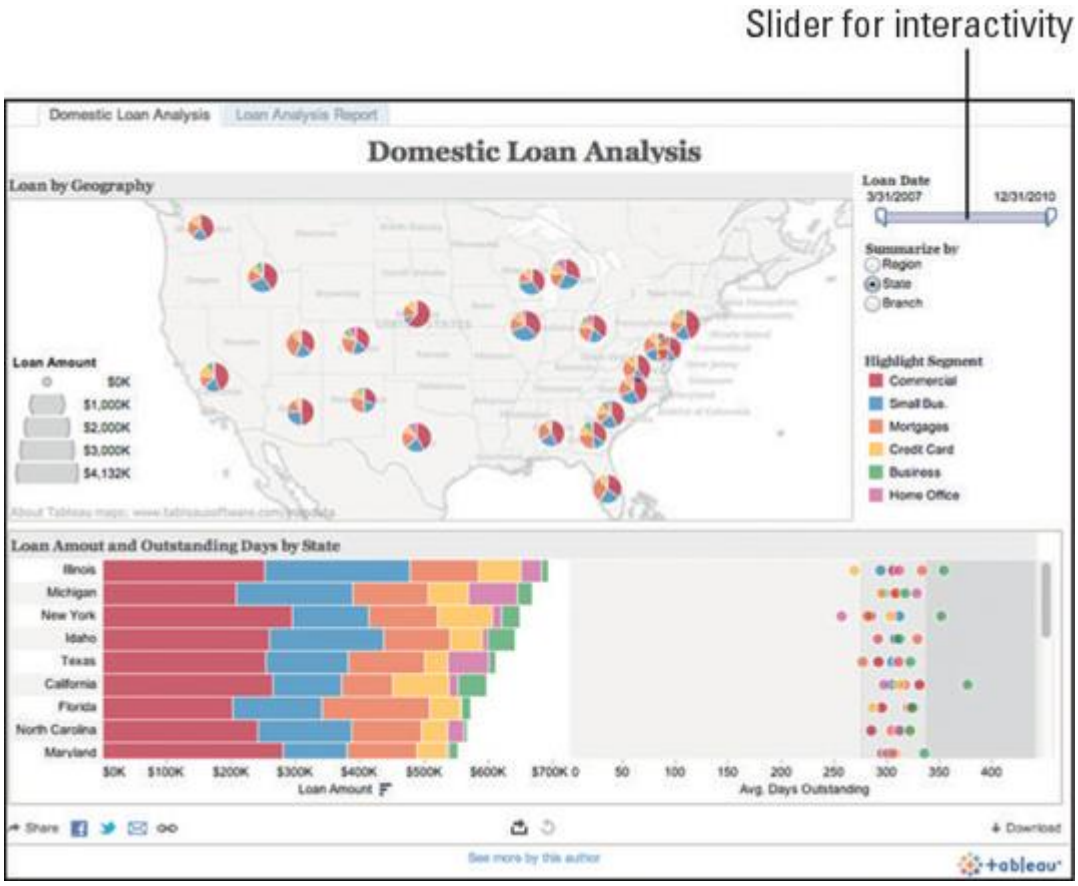
You can, of course, represent multiple variables in a single data visualization. However, it is always a good practice to represent them on a single scale to avoid confusion. For example, the graph below shows bars for the years 2010 to 2012, and then breaks down the year 2013 into bars of four quarters. This could be confusing since the x-axis scale has not been kept standard.



Data Visualization: Examples of the Good and the Bad

Data visualization 1

In this example, the following figure shows a dashboard that analyzes the status of domestic loans in the United States.



Things that work well:

- **Color consistency:** One thing that's evident throughout this visualization is the consistency of the colors in the dashboard. On the right is a single legend — Highlight Segment — that shows the legend for the color, which remains in both the bar chart at the bottom and the pie charts on the map.
- **Simplicity:** The two large charts make digesting the data easy.
- **Interactivity:** The slider in the top-right corner controls the time period displayed on the charts. That interactive feature put users in control of what they view.

Things that don't work:

- **Chart choice:** The small pie charts that are overlaid on the map are of little value. They're hard to view, and without clicking every single one, the user can't determine which of them are worth evaluating. It's also virtually impossible to tell which states or regions the charts pertain to.
- **Color choice:** The abundant use of red, blue, and orange are misleading,

especially in the stacked bar chart at the bottom of the data viz. At a glance, users may think that the colors could mean good versus bad; in fact, they're just associated with a specific segment. This type of color usage harkens back to the recommendation about being careful with the use of RAG colors. An alternative is to use more muted colors, such as a range of grays and blues.

- **Data overload:** There's a lot of data on the screen, but none of it really identifies the most important data or trends that users need to pay attention to. This visualization displays data for viewing instead of adding real value.

Data visualization 2

This figure shows a sales dashboard from a popular blog that displays high-finance charts.

Things that work well:

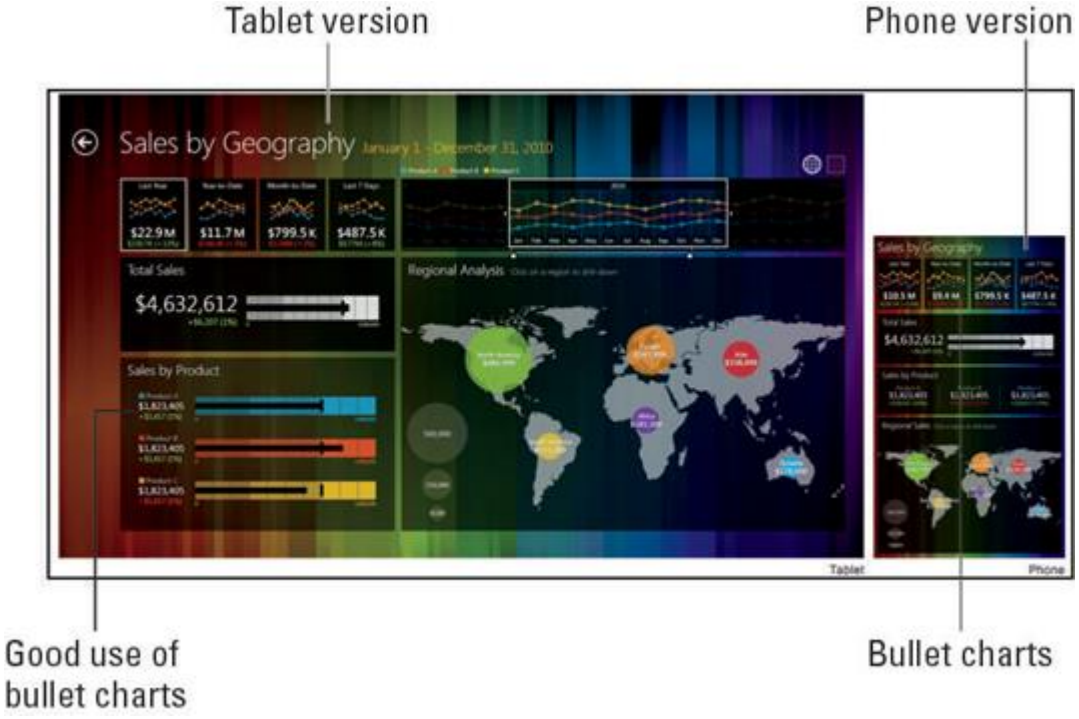
Unfortunately, there's not much in this data viz that works very well.

Things that don't work:

- **Color choice:** Black backgrounds are tricky. Many beginners use a large dark background to make the charts stand out. Over time, however, most users get tired of the dark background, and if they have to print the data visualization, this background is surely a great waste of ink. Avoid using black or very dark backgrounds throughout a data viz unless you can overlay each section with a much lighter color.
- **Chart choice:** The Status pie chart on the left has four almost-equal slices and tells the user absolutely nothing. Using an ascending bar chart would make the small change in the values more evident.
- **Chart choice:** The Line of Business bar charts are tiny and hard to read. Moving the Line of Business column charts to where the line charts are currently located and increasing the height would allow more of the values to be shown. Presently it looks like it just got stuck in the corner.

Data visualization 3

The figure below shows a data visualization that displays sales by geography.



Note that the following things work well and things that don’t work apply to both the tablet and phone versions in the figure.

Things that work well:

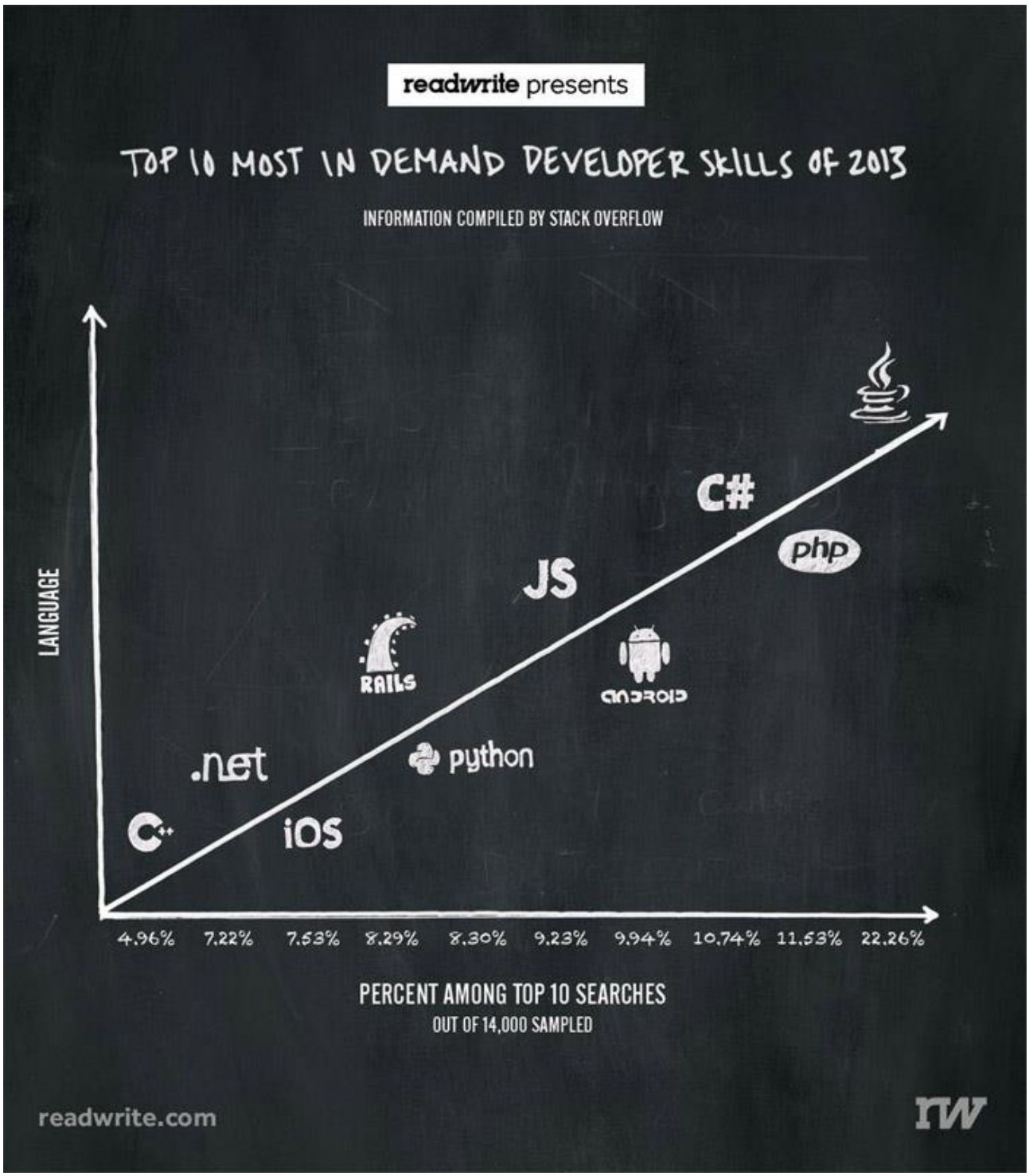
- **Chart choice:** The bullet charts on the left tell a clear story about how the products are selling.
- **Location intelligence:** The use of proportionally sized, brightly colored circles on each country makes this display fairly easy to digest.
- **Mobile viewing:** A great feature of this data viz is that it shows you how it will look on mobile devices. You can see a consistent layout in the two versions.

Things that don‘t work:

- **Color choice:** The striped background colors are distracting. The spectrum of colors attracts users’ attention, but in the wrong direction: away from the data. A solid background would be better.
- **Chart choice:** The four micro line charts on the left, which show some sort of trend, add no value. Can you tell how Sales Last Year is trending by looking at the lines? Removing the lines and just displaying the numbers in large text would be much more effective. Also, the map uses a large amount of real estate compared to the value it adds.

Example:

The infographic, as compiled by tech blog ReadWrite, depicts the Top 10 Most In-Demand Developer Skills of 2013, as compiled by Stack Overflow through keyword searches.



Take a look at the chart. What's good and bad about it?

The Good

The data is properly cited.

The Bad

EVERYTHING ELSE.

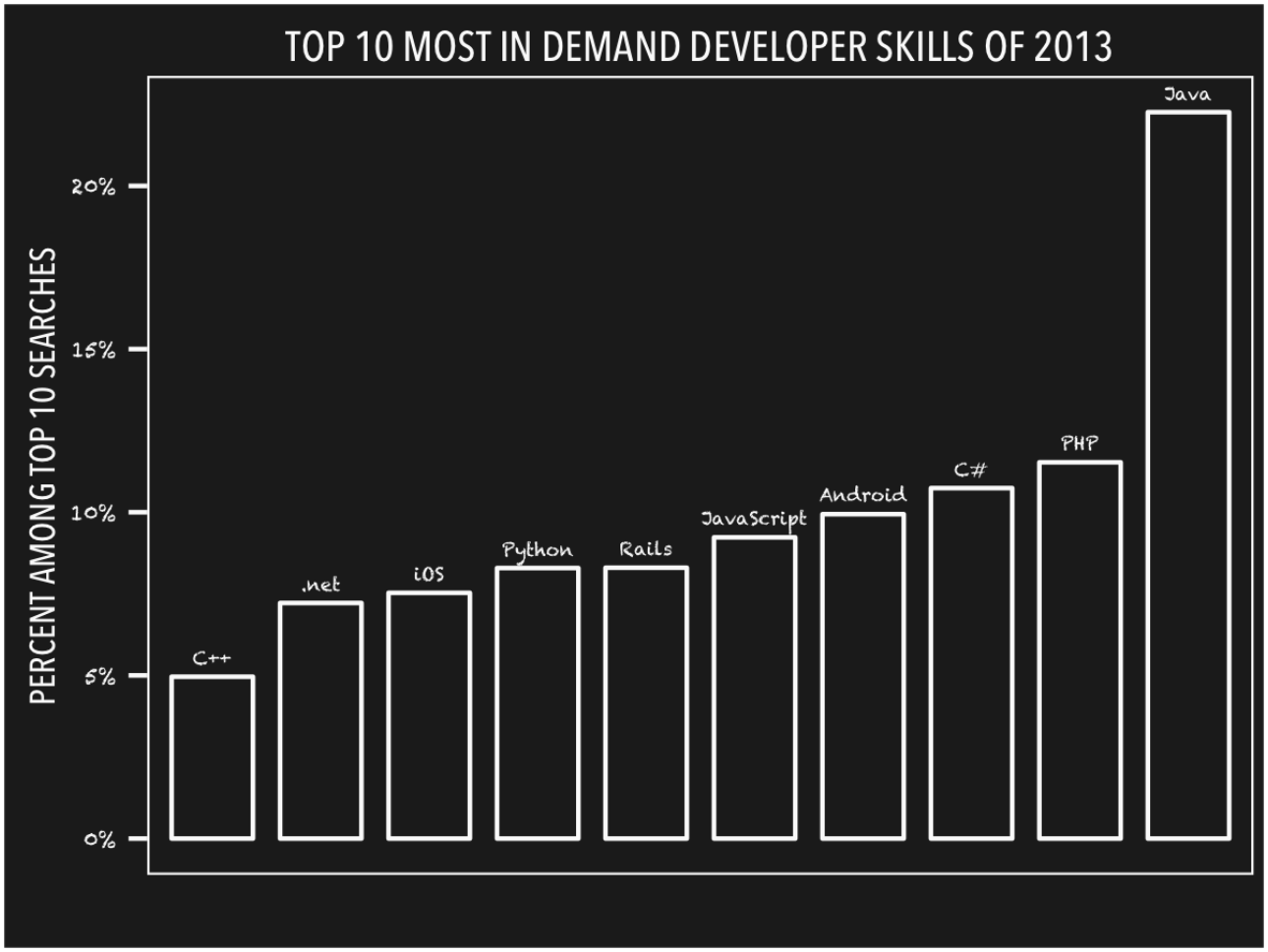
What's terrible about the graph? Let me count the ways:

- Using discrete values in the X-Axis for a continuous measurement (i.e. the percentage). And not only that, discrete values with two significant figures, which make the X-Axis unusually cluttered.
- The Y-Axis is Language. This implies that some programming languages are more language than others. (to be fair, Java is more language than Android)
- Not all entries on the chart are programming languages. (Android, for example, is an operating system.)
- The 45-degree line in the chart implies that the relationship between language and %-of-searches is perfectly linear, where in reality the data has an upward-parabolic shape.
- No relative proportions between the programming languages. We can't accurately see the increase in language Java has relative to Android just by looking at the graph.
- Cannot easily associate a language with the given X-Axis value. The logos representing the programming language oscillate around the line, and it's hard to see at a glance which percentage corresponds to which language.

Fixing the Chart

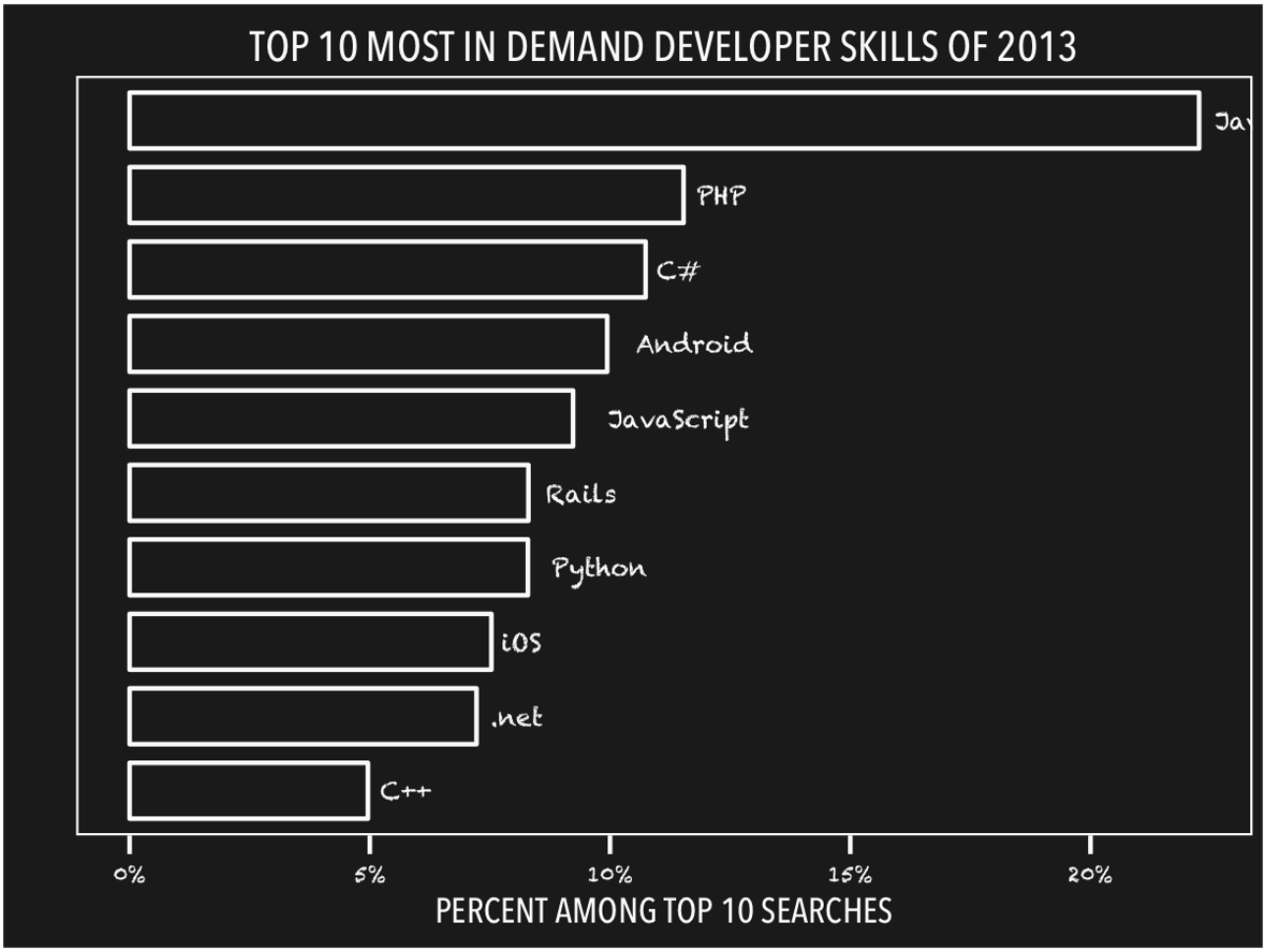
How can we make the chart somewhat logical?

The easiest way to improve the chart is to convert the chart from a line chart to a column chart. Here's a column chart that keeps the intended impression of the original chart:



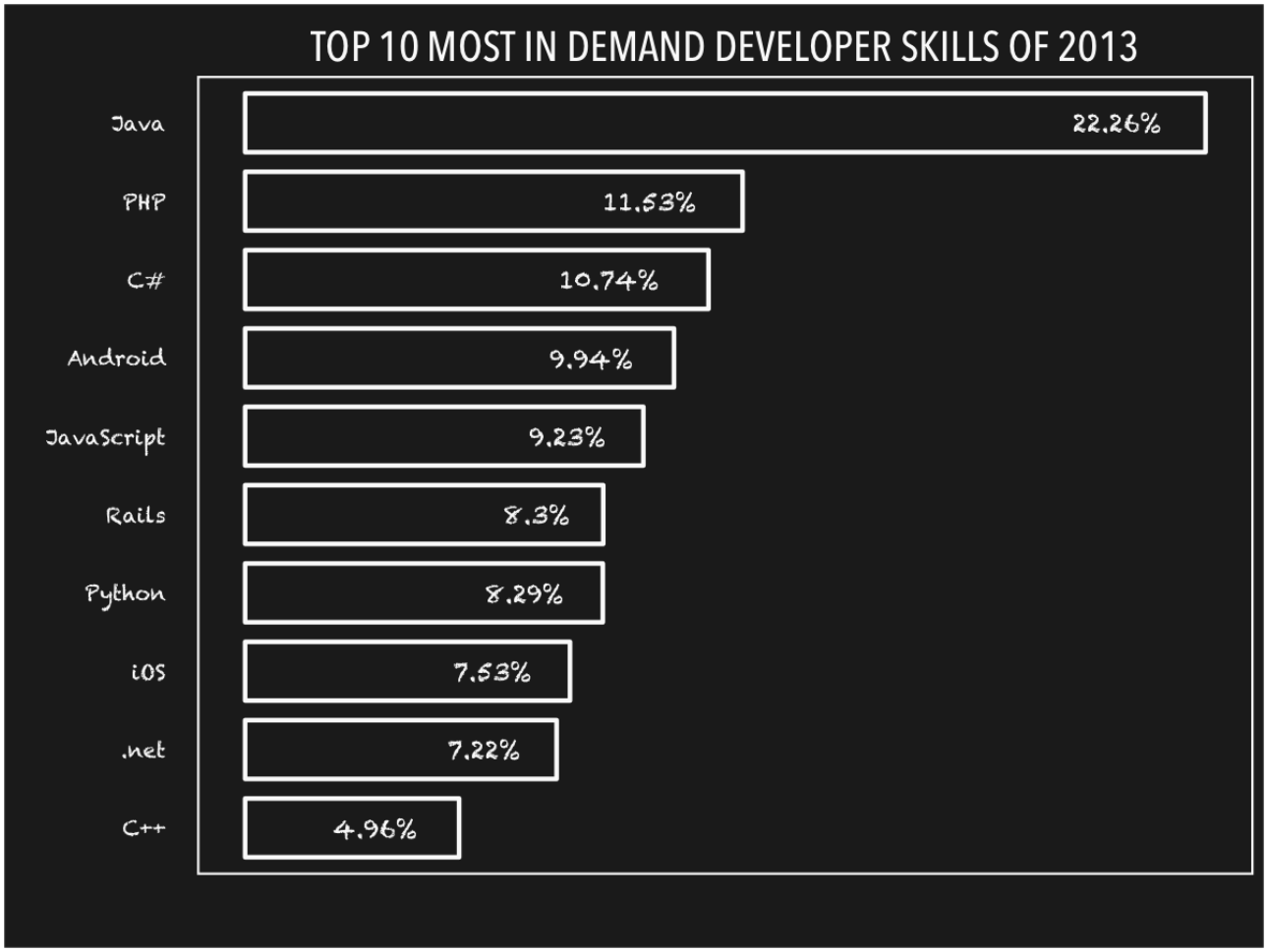
A much bigger improvement, although unfortunately without the cool hand-drawn logos. The axes are no longer illogical and it's easy to determine the relative impact of each language (e.g. Java is clearly, clearly at the top). However, the large amount of text can clutter the bars, and it can be difficult to correlate the raw percentage with the scale at a glance.

Another option is to rotate the chart and use bars instead of columns:


TOP 10 MOST IN DEMAND DEVELOPER SKILLS OF 2013
Ja
PHP
C#
Android
JavaScript
Rails
Python
iOS
.net
C++
0%
5%
10%
15%
20%
PERCENT AMONG TOP 10 SEARCHES

This fixes the text issue by giving more room for text with most of the factors, but the Java text clips outside the chart. The correlation issue between language and percent value persists.

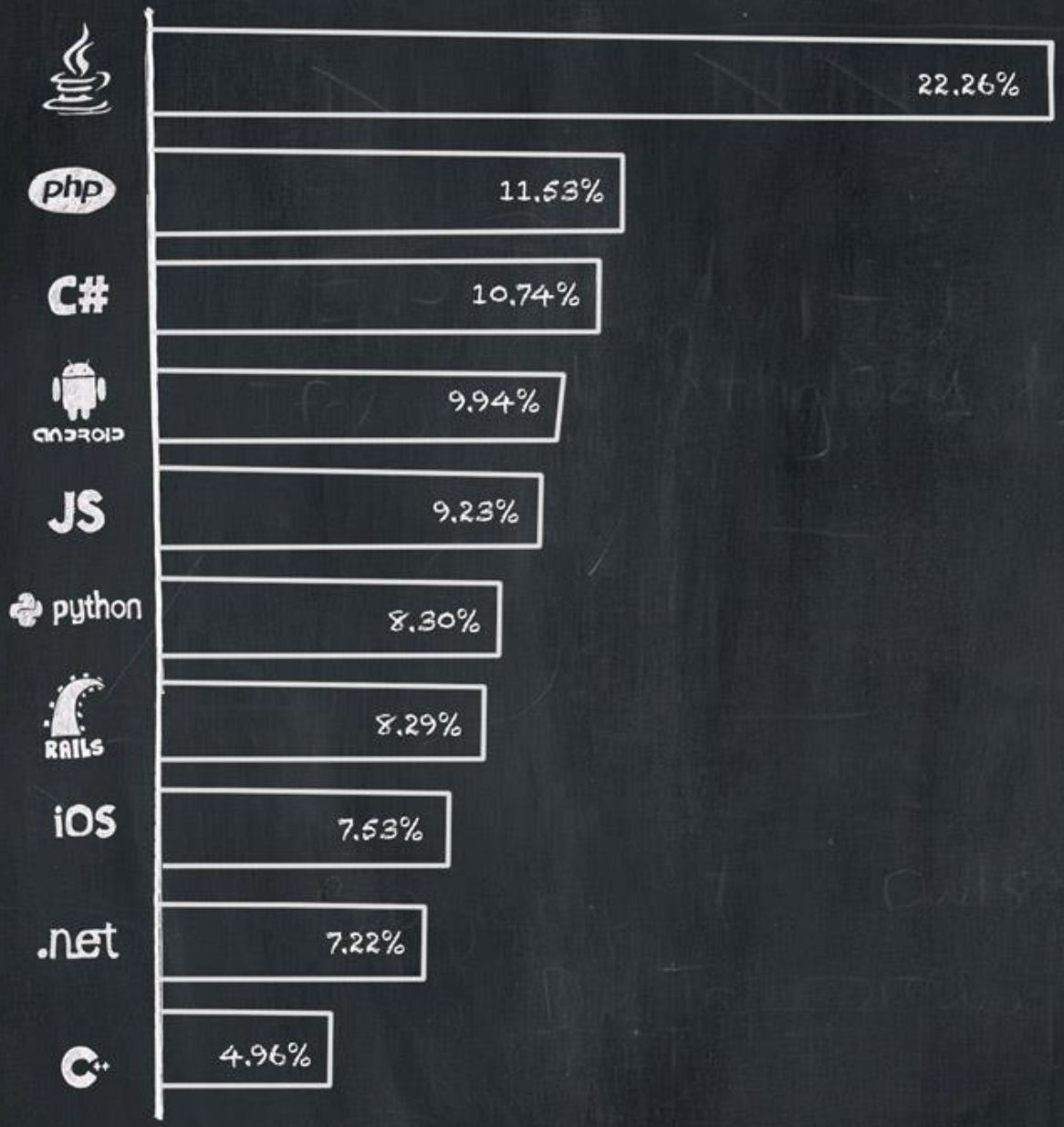
A best-of-both-worlds approach is to display the language on the Y-Axis and the raw percent value on the corresponding bar itself, allowing us to forgo the percent metric axis entirely.



readwrite presents

TOP 10 MOST IN DEMAND DEVELOPER SKILLS OF 2013

COMPILED BY STACK OVERFLOW



OUT OF 14,000 QUERIES ON THE CAREERS 2.0 SEARCH ENGINE

References:

https://www.dummies.com/

https://interactions.acm.org/

https://www.kdnuggets.com/

https://kaijiezhou.wordpress.

https://humansofdata.atlan.

https://minimaxir.com/2014/01/more-language-more-problems/