STATISTICS FOR DATA SCIENCE

Data Visualization and Interpretation

D. Uma

Department of Computer Science and Engineering

umaprabha@pes.edu

STATISTICS FOR DATA SCIENCE

Data Visualization and Good vs. Bad Visualization

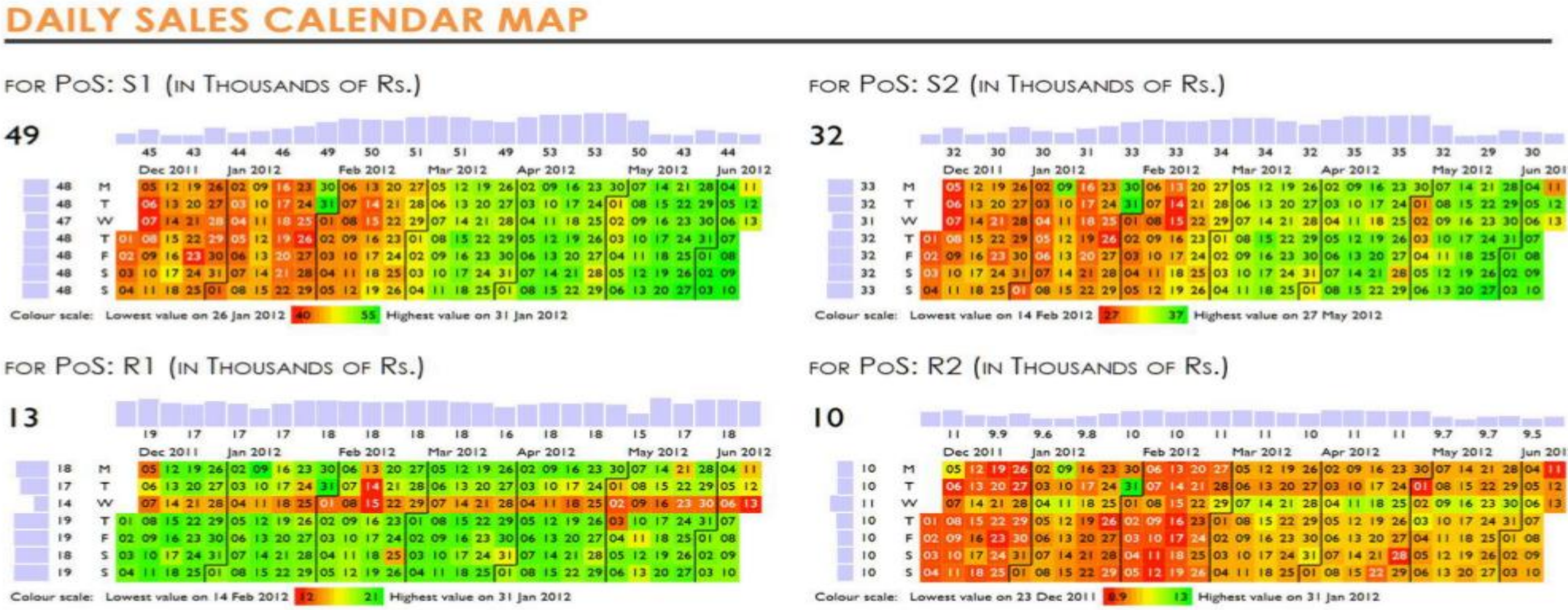
D. Uma

Department of Computer Science and Engineering

Good vs Bad visualization

How can good visualization be used to tell a story?

Restaurant had collected years of data



Source:https://www.slideshare.net/gramener/making-big-data-relevant-importance-of-data-visualization-and-analytics?from_action=save

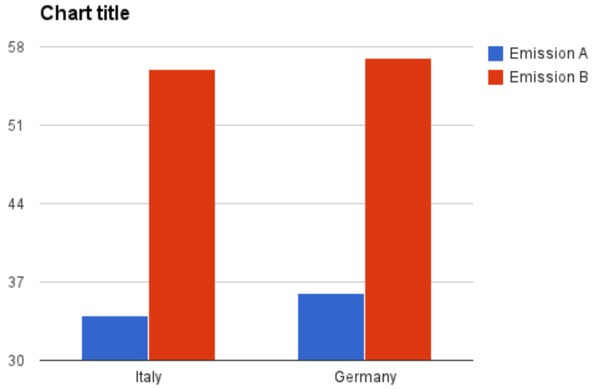
Good vs Bad visualization

Some Do’s and Don'ts

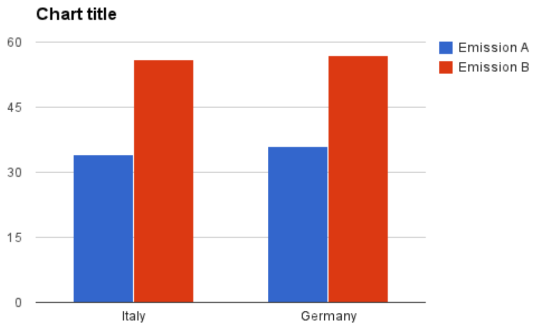
- Use the full axis
- Avoid distortion
- Sort the data for ease of comparison
- Use consistent intervals on any axis or indicate a break
- Use the chart type wisely
- Don’t use colors and effects without reason
- Don’t use 3D

Good vs Bad visualization

For bar charts, the numerical axis (often the y axis) must start at zero.



WRONG



Correct

Source: https://www.eea.europa.eu/data-and-maps/daviz/learn-more/chart-dos-and-donts

Sort your data for easier comparisons

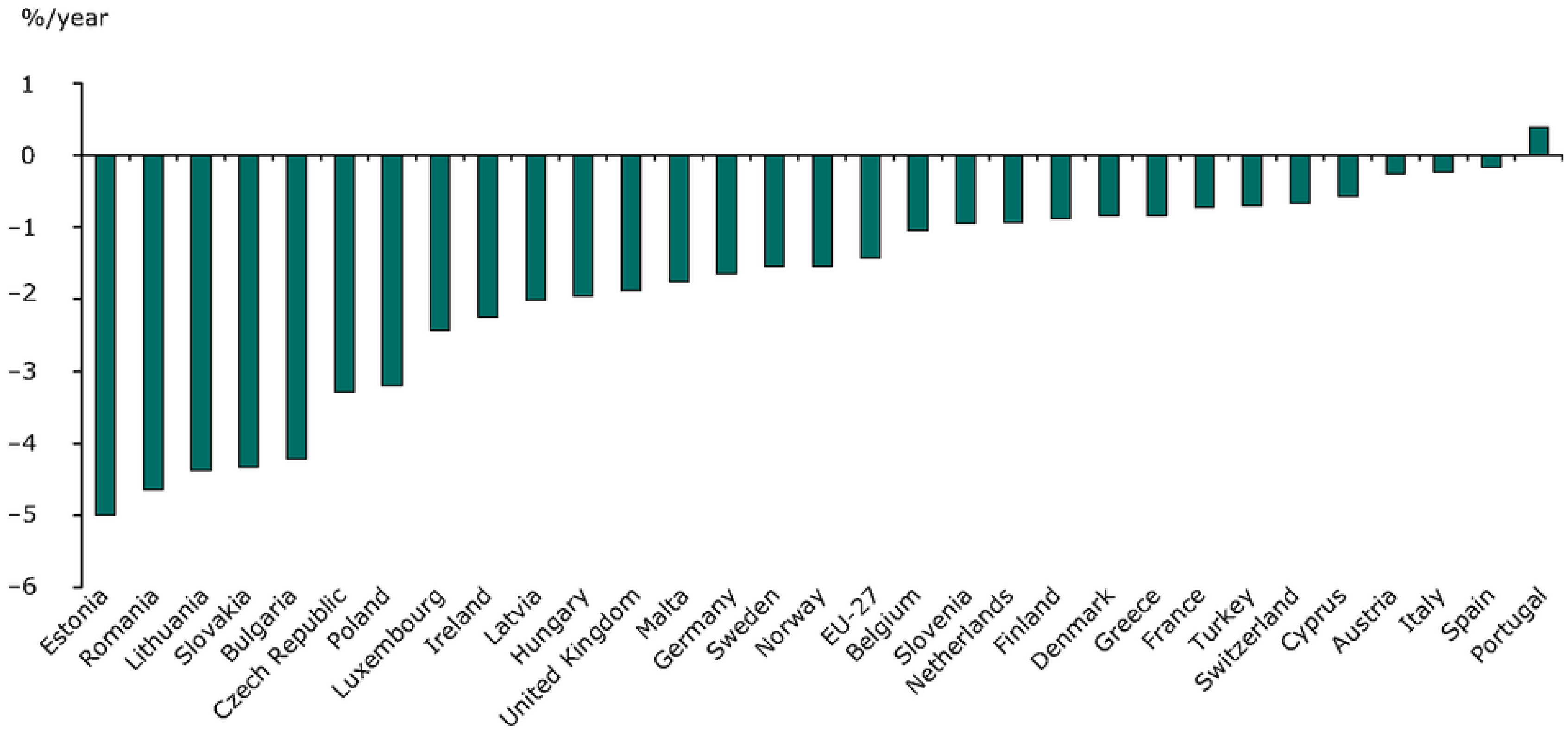
The bar chart is a good example, where the chart x-axis is sorted on the y-values not on the alphabetic order of the country names.

It will be otherwise very difficult if not impossible for users to do a proper comparison across the many bars.

It is in any case easy with a quick eye-scan to find your own country in the list.

Good vs Bad visualization

Contd..

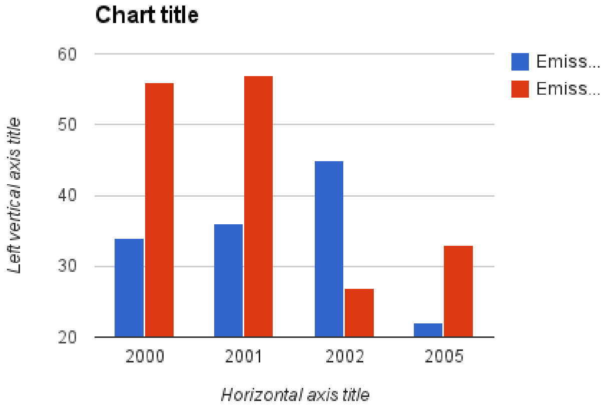


Source: https://www.eea.europa.eu/data-and-maps/indicators/final-energy-consumption-intensity-2/assessment

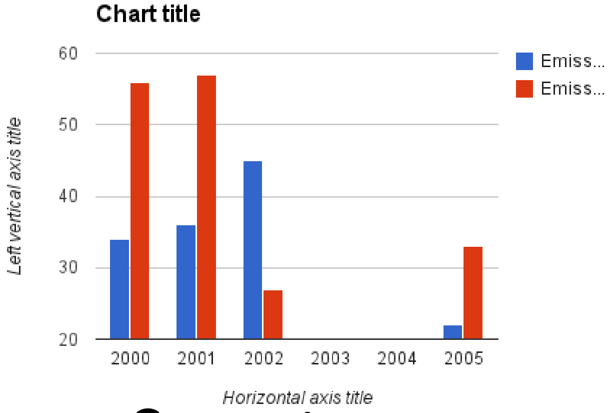
Good vs Bad visualization

Use consistent intervals on axis (be transparent on data gaps)

- Be clear when some data is missing. Explain the reason why is missing. Use the full axis and do not skip values when you have numerical data.
- The x-axis in the "wrong example" below has a time-series with inconsistent intervals (missing years 2003 and 2004) giving a distorted view of data over time.



Wrong



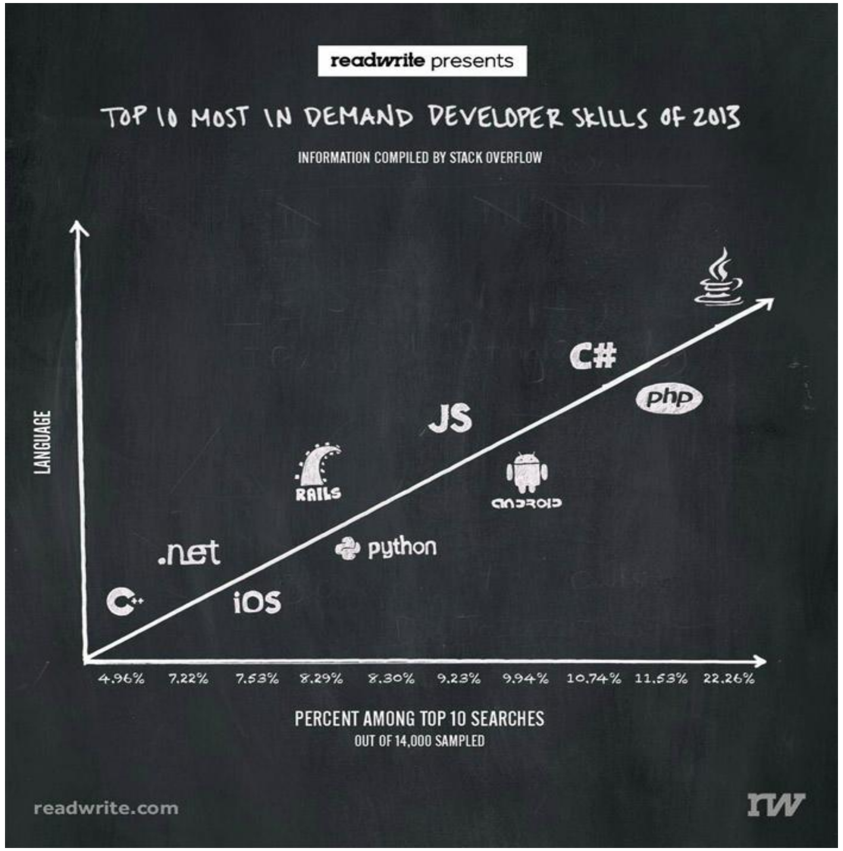
Correct

Source: https://www.eea.europa.eu/data-and-maps/daviz/learn-more/chart-dos-and-donts

Good vs Bad visualization

An axis should have *something* on it

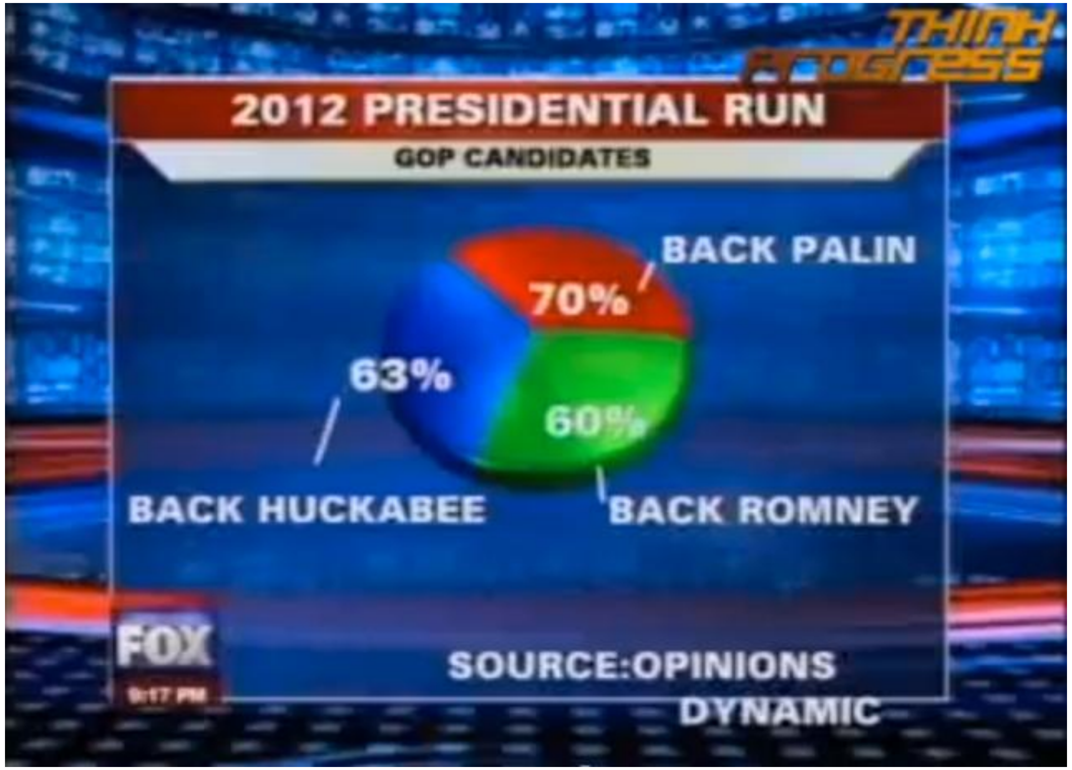
We would have thought this obvious, but a line graph should have *something* numerical on each axis. The graph below does not. Its vertical axis is labeled "Language" but even in the most generous interpretation this is a categorical variable and thus not appropriate for display using a line graph.



Source:https://callingbullshit.org/tools/tools_misleading_axes.html

Good vs Bad visualization

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.



Source:https://humansofdata.atlan.com/2016/09/common-mistakes-bad-data-visualization/

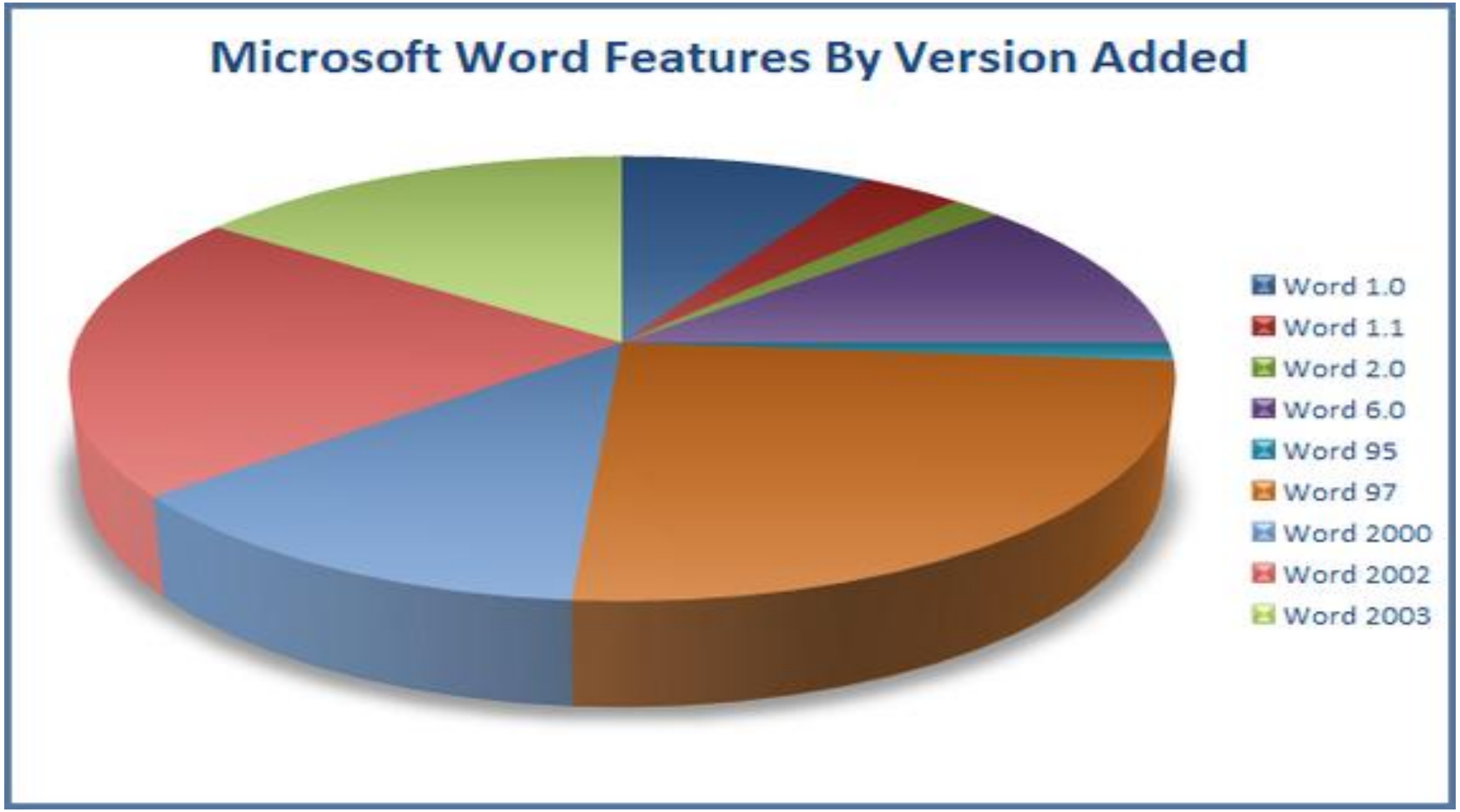
Good vs Bad visualization

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.

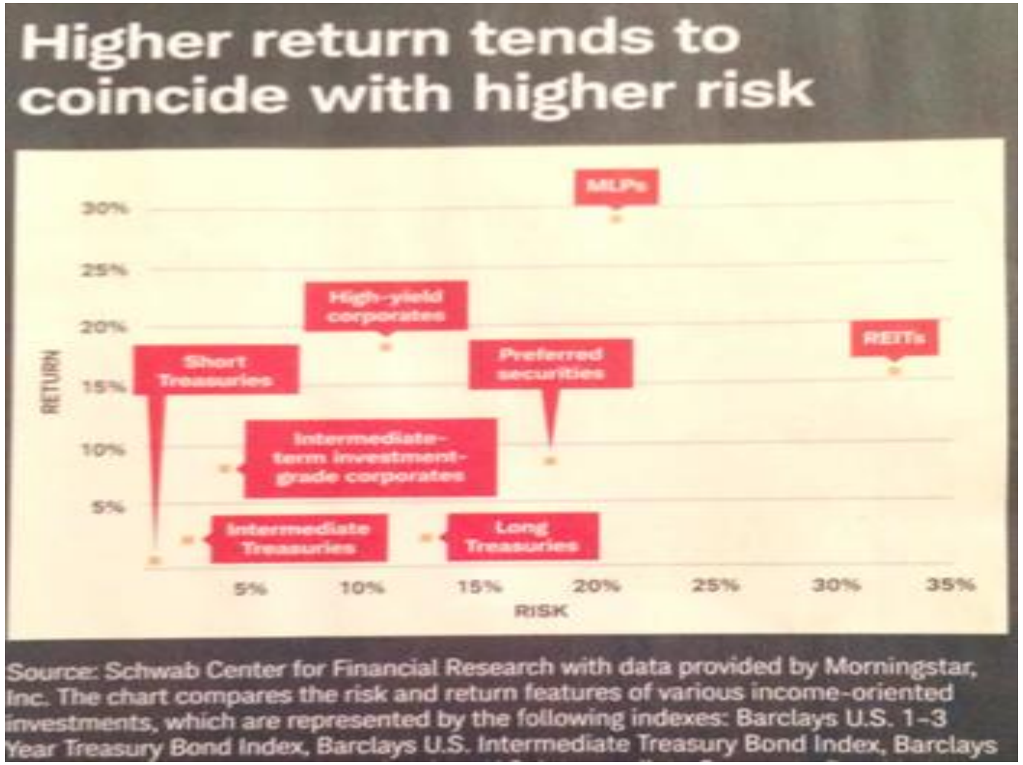
Source:https://humansofdata.atlan.com/2016/09/common-mistakes-bad-data-visualization/

Good vs Bad visualization



Source:https://humansofdata.atlan.com/2016/09/common-mistakes-bad-data-visualization/

Good vs Bad visualization



Higher return tends to coincide with higher risk

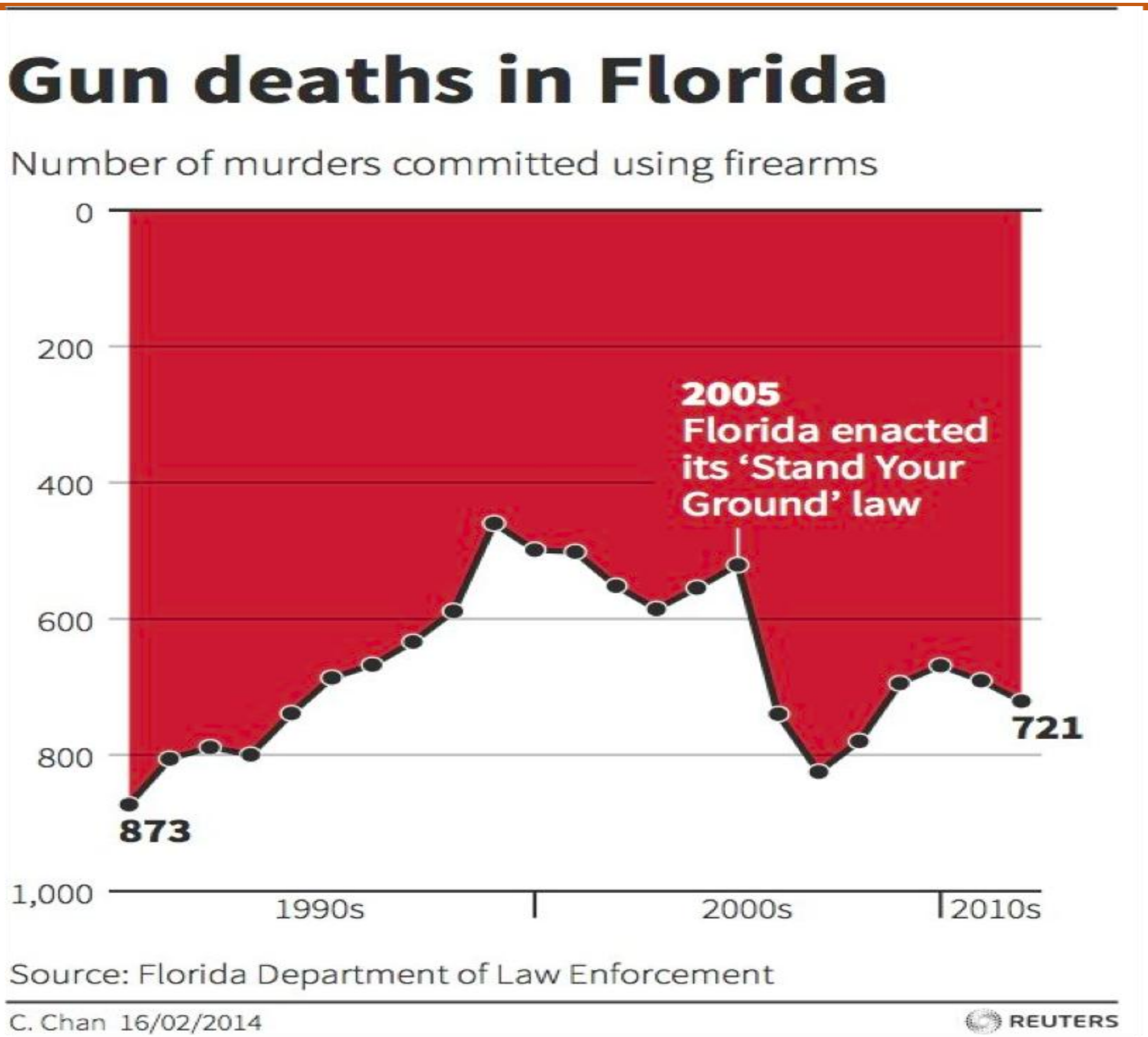
Source: Schwab Center for Financial Research with data provided by Morningstar, Inc. The chart compares the risk and return features of various income-oriented investments, which are represented by the following indexes: Barclays U.S. 1-3 Year Treasury Bond Index, Barclays U.S. Intermediate Treasury Bond Index, Barclays

When there are two variables, and their correlation is of interest, a scatter plot is usually recommended. But not here!

The text labels completely dominate this chart and the designer tried very hard to place them but a careful look reveals that some boxes are placed above the dots while others are placed to their right and the dot for "Short Treasuries" holds refuge quite a while away from the dot. This means the locations of the text boxes do not substitute for the dots.

Source:https:https://junkcharts.typepad.com/junk_charts/2014/12/where-a-scatter-plot-fails.html

Good vs Bad visualization



At first glance, it looks like gun deaths are on the decline in Florida. But a closer look shows that the y-axis is *upside-down*, with zero at the top and the maximum value at the bottom. As gun deaths increase, the line slopes downward, violating a well established convention that y-values increase as we move up the page.

Source:https://callingbullshit.org/tools/tools_misleading_axes.html

Good vs Bad visualization

In summary, data visualizations tell stories.

Relatively subtle choices, such as the range of the axes in a bar chart or line graph, can have a big impact on the story that a figure tells.

When you look at data graphics, you want to ask yourself whether the graph has been designed to tell a story that accurately reflects the underlying data, or whether it has been designed to tell a story more closely aligned with what the designer would like you to believe.

THANK YOU

D. Uma

Department of Computer Science and Engineering

umaprabha@pes.edu