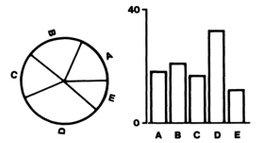
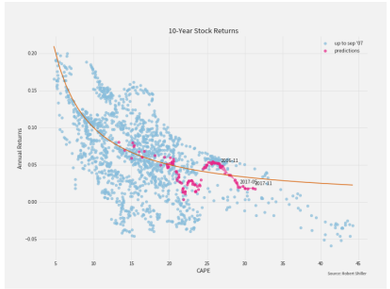
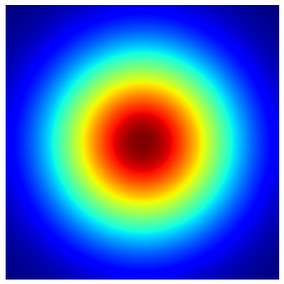
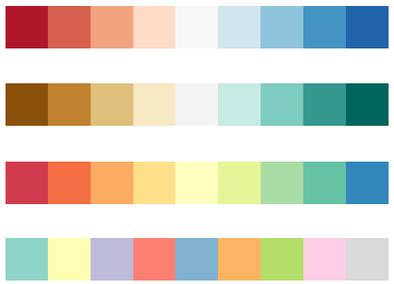








Data Visualization

Data Visualization

A very, very brief introduction (part I)

Chandrasekhar Ramakrishnan

Hi everyone. I would like to present a brief introduction to data visualization this afternoon.

Everyone who works with data needs to visualize it. I know you have made many visualizations in your life, but I think many of you have not been exposed to thinking formally about visualizations. I presented this slide deck in a data-visualization-focused meet-up (which is why there are not in the SDSC theme), and when Fernando and Guillaume were asking for volunteers to present, I offered to go through this topic, since I thought it might be valuable for some to hear this perspective on presenting data.

Goal of Visualization

Let me get started by making this term “Visualization” a bit more precise. Before I start, let me add that you should feel free to interrupt me at any time and ask question if you have any. This forum seems to usually run this way anyway, but I like to be interrupted with questions.

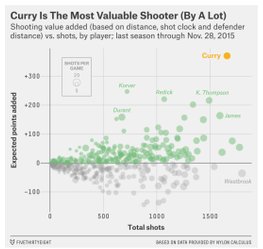
The **goal** of data visualization is to enable **quantitive reasoning** with your **eyes**.

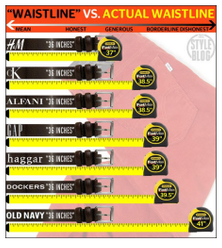
There are many ways to define the goal of visualization. This is one: *"The goal of data visualization is to enable quantitative reasoning with your eyes."*

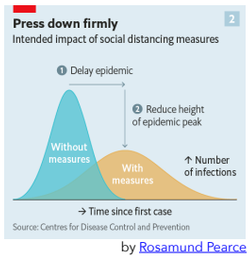
It is not the only approach to the topic, but it is the one we will pursue here. I like this definition because it emphasizes to to key things about data visualization: it has to do with quantitative data, and must respect the properties and characteristics of this kind of data; and it has to do with the eyes, so it must at least be aware of the characteristics, the strengths and the weaknesses of the human visual processing system.

These kinds of visualizations cover an important subset of all visualizations, and formulating the goal of visualizing data this way makes it possible for us to objectively analyze and compare visualizations, which is valuable for building intuition.

Data Visualization vs. Infographics







by Rosamund Pearce

by /u/valeriepieris

Not all **presentations** of data try to achieve this goal. I want to distinguish between between *data visualizations* and *infographics*. The data visualizations we will talk about today present data to enable quantitative reasoning with your eyes. You see some examples on your left. Infographics, like you see on the right, present data in a way that does not enable quantitive reasoning with your eyes. I think the graphics on the right are fantastic illustrations of certain ideas, so they are not bad, but just not the sort of thing we will talk about.

Same Data, Different Goals

From FlowingData https://flowingdata.com/2010/09/30/advertised-vs-actual-waistline/

Consider these two visualizations of the same data set. They are both good, but they look very different because they pursue different goals.

The one on the left is a visualization of the kind we will be talking about: one designed to enable quantitative reasoning. The one on the right has a different goal: its goal is to be entertaining. This is of course fine, and this visualization does a good job of realizing this goal, but it is not the kind of visualization we will be talking about.

TLDR

With the right approach, it is possible to design a visualization that is good for any comparison you want to make

Without any support, color is easy to mess up; but with good tools, color is easy to get right

Transforming data can be key in visualizing it effectively

We are going to explore how to create visualizations to realize our goal. We will develop a framework for thinking about visualizations that will make it possible to design a visualization that is good for any kind of comparison we want to make. This framework may not always give us the best solution, but it will allow us to identify some bad solutions, reducing the search space to let us focus on possibilities that actually have potentital. Along the way, we will devote some extra attention to things are important for making readable visualizations.

Table of Contents

This introduction to data visualization is made up of several sections. To make visualizations, we need to use tools, and we will start with a high-level survey of the tool landscape. Then we will move to the meat of the presentation and develop a framework for thinking about visualizations. Using this framework, we will look at how to make choices to realize effective visualizations. Then we will into greater detail in two topics that are keys to making clear visualizations: the choice of color and the process of transforming data.

References

Edward Tufte

Visual Display of Quantitative Information

Envisioning Information

Visual Explanations

Online

- Maneesh Agrawala
  Stanford University
- Jeffrey Heer
  U. Washington
- Jock Mackinlay
  Tableau

What I am presenting here borrows extensively from others. The classic trilogy by Edward Tufte is full of wonderful insight and inspiration for making powerful visualizations. And from the academic domain, Maneesh Agrawala, Jeff Heer, and Jock Mackinlay are all computer scientists who specialize in data visualization and make some great resources available online. You can use those resources to go into greater depth into the things I will be talking about today.

Follow Along

https://github.com/ciyer/intro-data-viz

I also wanted to mention that you can access my material on github. The contents of the presentation are there as well as the code used to make many of the visualizations I show, so if you are curious how I did something, you will find that helpful.

Tools for Visualization

There are many tools out there for making visualizations. You are certainly already be very familiar with one or more. In this presentation, we will not go into any of them in detail, but you may be interested to know what is out there.

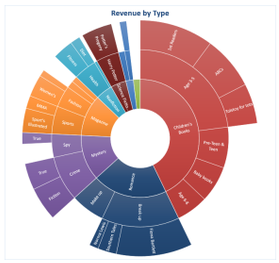
Point and Click

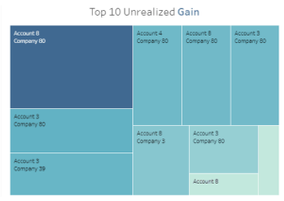
Excel / Numbers / etc.

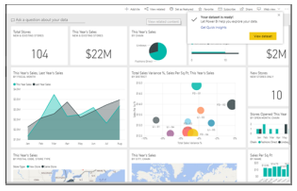
Tableau

Spotfire

Power BI







Some tools for making visualizations are built on graphical user interfaces (GUIs). These include spreadsheet programs like Excel, as well as other tools that are more specialized for data visualization like Tableau, Spotfire, and Power BI. These can be used to make interactive visualizations and dashboards without any programming. Tableau, for example, operationalizes a lot of research about visualization into its software and can produce very nice results, but you are of course limited to what the tool offers.

1. https://www.microsoft.com/en-us/microsoft-365/blog/wp-content/uploads/2016/02/3-ways-to-drive-business-decisions-using-the-new-Excel-2016-charts-2.png
2. https://www.tableau.com/about/blog/2019/9/answer-your-investment-questions-faster-tableau
3. https://docs.microsoft.com/en-us/power-bi/consumer/end-user-experience

Programming

R

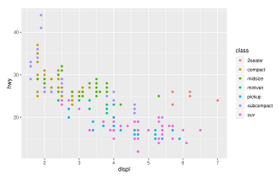
ggplot

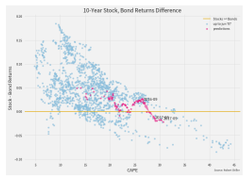
Python

- matplotlib
- seaborn
- bokeh
- altair
- holoviews

JavaScript

- Vega/Vega-Lite
- D3
- C3
- plot.ly





Programming languages also provide frameworks for data visualization, including, of course, widely used languages for data analysis like R and Python. The landscape in R is cleaner: ggplot is the clear choice. In Python there is the classic warhorse matplotlib, but several newer frameworks are actively being developed to provide alternatives to matplotlib that are offer both more modern APIs and can render for presentation on the web and support interactivity.

And JavaScript, though not typically an an environment for data analysis, is the primary programming language for the web and has powerful frameworks for visualizing data.

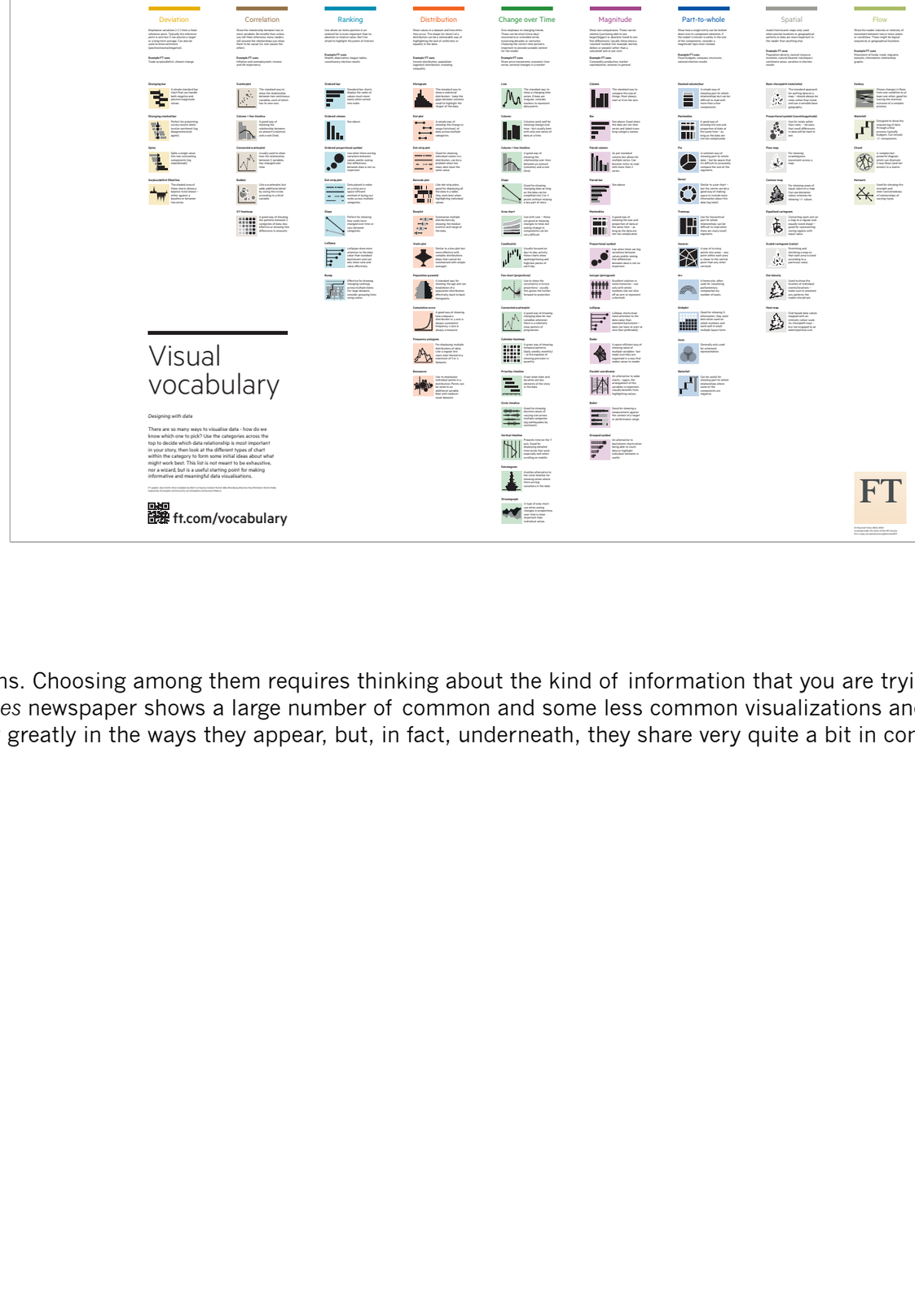
I am not going to go into greater detail about any of these tools, but I do want to highlight Vega and Vega-Lite, which are declarative specifications for describing data visualizations. I like them personally, and through built-in support in environments like Jupyter Notebooks, they are gaining access to a wider audience.

Any questions about tools before we move on?

1. https://ggplot2.tidyverse.org

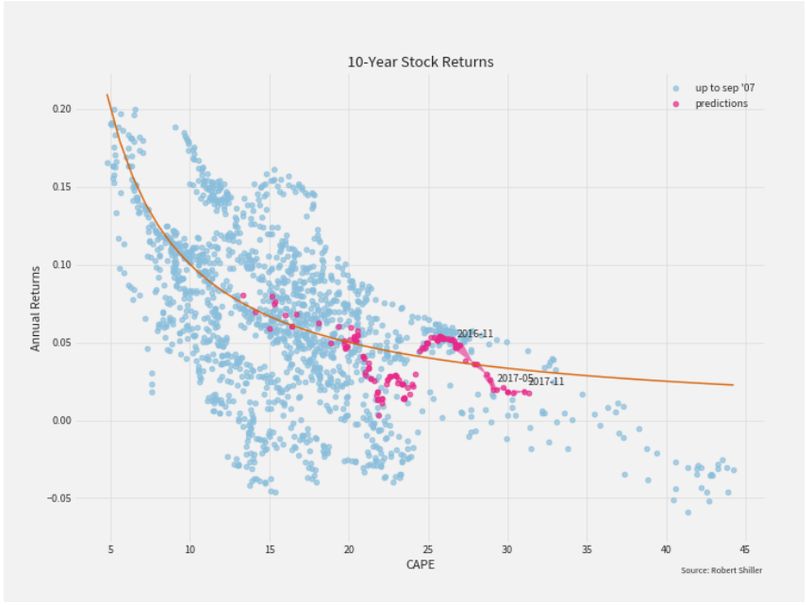
Thinking About Data Viz

Now we come to the core of today's topic. We defined our goal for making visualizations as supporting quantitative reasoning visually. This has implications for how we build visualizations.



There are many different kinds of visualizations. Choosing among them requires thinking about the kind of information that you are trying to communicate and what aspects of the data you want to highlight. This chart from *The Financial Times* newspaper shows a large number of common and some less common visualizations and explains what situations they are good for. Looking at this catalog illustrates that visualizations can vary greatly in the ways they appear, but, in fact, underneath, they share very quite a bit in common.

Parts of a Visualization

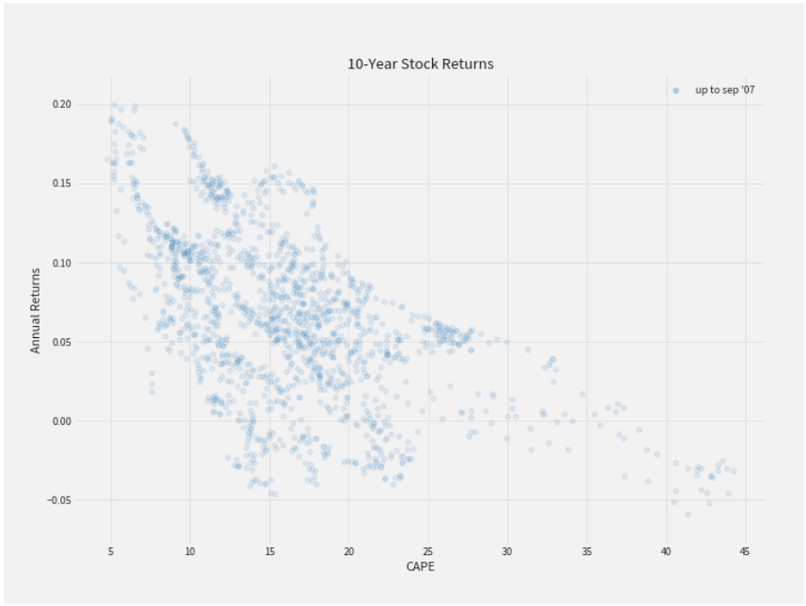


Part of what they share is their structure. Let us dissect a visualization to take a look at its structure.

Marks

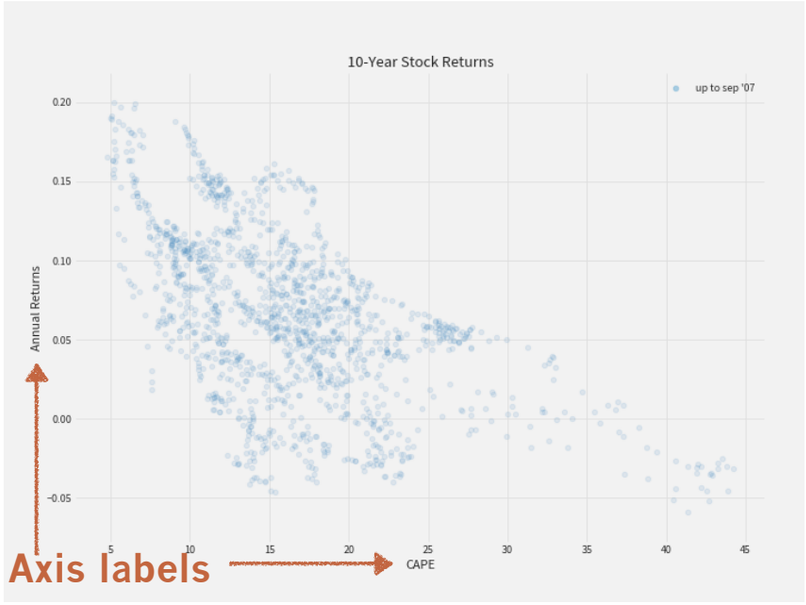
The core of a visualization is made up of the marks that represent data. The marks are the most important part of any visualization, but, alone, they are not sufficient to communicate an good understanding of the data.

Context



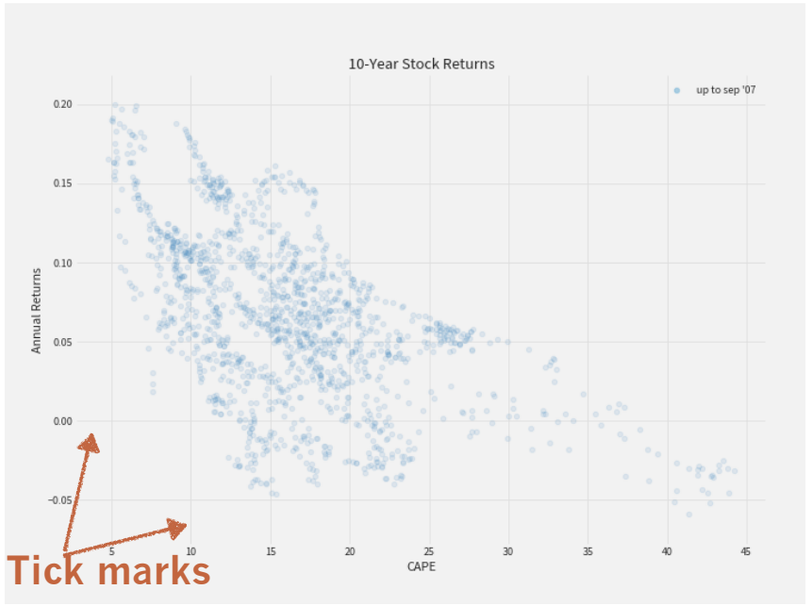
To understand data, context is necessary, and the next layer of a visualization provides context for interpreting the marks. This primary layers of context include... (the axis labels, axis tick marks, title, and a legend).

Context



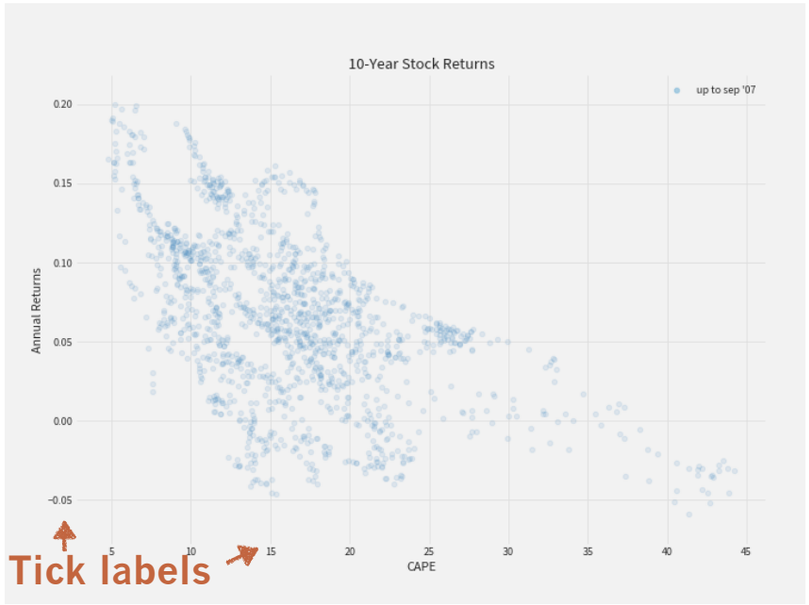
...the axis labels...

Context



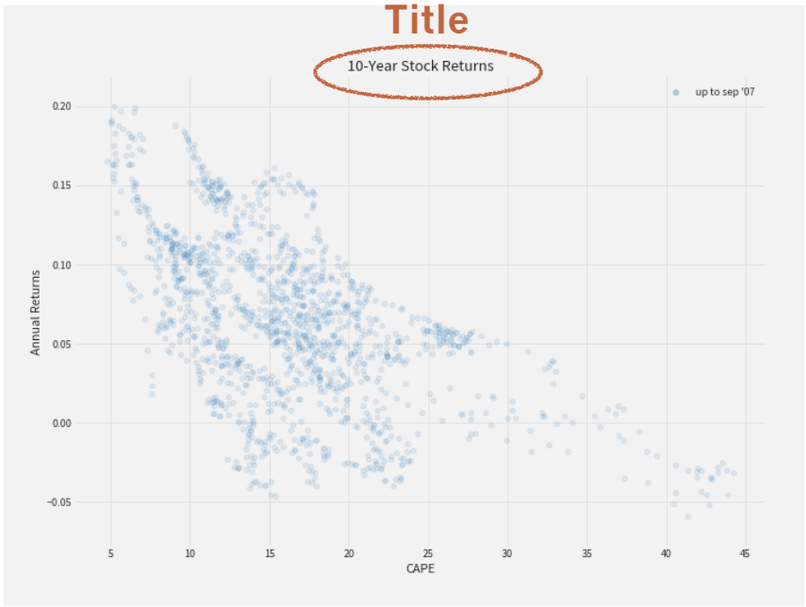
...axis tick marks...

Context



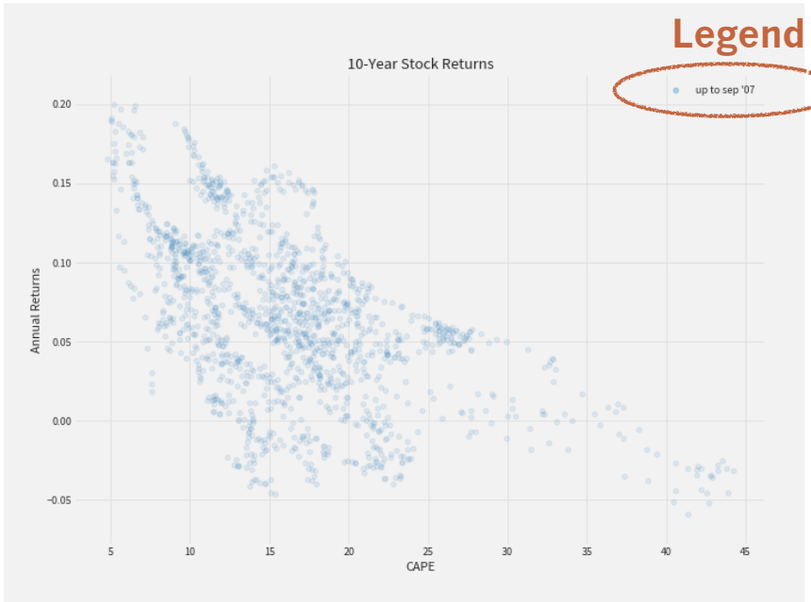
...axis tick labels...

Context



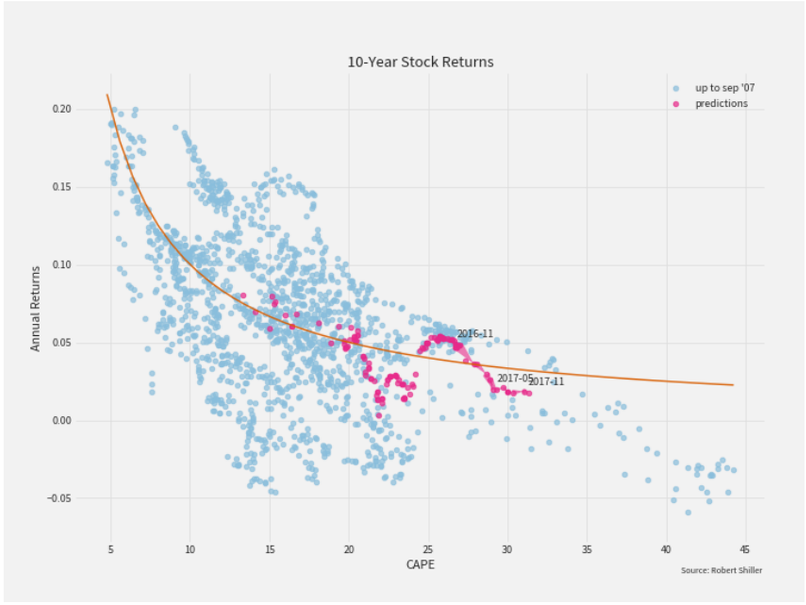
...title...

Context



...and a legend.

More Context

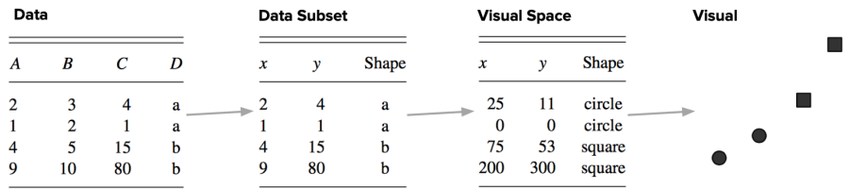


Much of the power of a visualization comes from context. Beyond those basic layers of context, it can be helpful to add more context by layering marks, for example in this visualization that, in addition to the data, that includes a theoretical model of the data in orange, predictions made by a model in pink, labels for certain data points as text in black, and an indication of the source of the data in the lower right-hand corner.

Marks in Detail

Let us focus our attention on the marks. Specifying marks forms the core of the visualizations process. But how are marks specified?

Marks Through Mapping



One structured way of thinking about visualization is that put forward by Leland Wilkinson in his *Grammar of Graphics*. In that perspective, marks are specified by defining maps from data to ink (pixels) in an image.

Put another way: visualizations are realized by mapping data variables to visual variables.

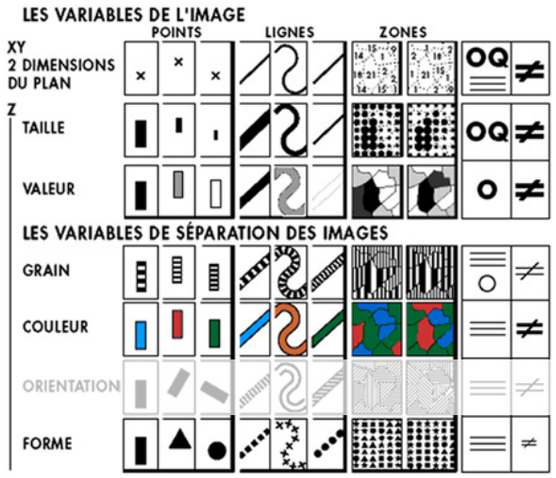
Here we see a schematic depiction of that process. It starts with tabular data and applies first an operation to select a subset of the data we want to visualize, and assigns the data columns to visual variables we want to control – the three variables *(x,y)* position and shape. The next step is to map from the data space to the space of the graphics display. In this example, that means converting from the range of data in the A and C columns into raster coordinates of the graphics system, and mapping the D categorical variable to shapes.

This information is handed to the graphics API to render the result as an image.

A key part of the process is the choosing visual variables we can control with data. Other than position and shape, what other visual variables can we control?

Sémiologie Graphique

Jacques Bertin *Sémiologie Graphique* (1967)



http://pauline-blot.blogspot.ch/2012/02/jacques-bertin.html

Jacques Bertin was a cartographer and semiotician who started thinking systematically about how to represent data in visual form. He identified eight visual variables that can be controlled: *x, y* position, size, brightness, texture, hue (color), orientation, and shape. In practice orientation is not used very frequently, so we will ignore it, leaving us seven variables.

Now we know what variables we *can* control, it is natural to ask if some kinds of data variables are better matches for mapping certain visual variables. E.g., imagine we have a table of fuel efficiency data for automobiles in 1999 and 2008, and we want to see if 2008 cars are more or less fuel efficient in aggregate, what variables should we map fuel efficiency and year to? x, y position? Color? Something else?

Scales of Measurement

There is a systematic way of answering this question, but we need to review some concepts from statistics to formulate the answer.

In statistics, measurements occur on a scale. Different scales admit different kinds of operations. Since the goal of visualization is to permit quantitive reasoning with your eyes, it should not be surprising that the process of mapping from data to visual variables needs to take scales of measurements and their possible operations into account in order to be effective.

Who remembers the four scales of measurement?

Scales of Measurement

Scale
Nominal
Ordinal
Interval
Ratio

Statistics distinguishes between four scales of measurement: nominal, ordinal, interval, and ratio.

What are the examples of each of these scales?

Scales of Measurement

Scale	Examples
Nominal	Gender, color, city, species
Ordinal	Unhappy/indifferent/happy, S/M/L/XL
Interval	Dates, temp (C)
Ratio	Meters, money, temp (K)

An example of a nominal scale would be something like gender. These just have names and you can test equality, but there is no ordering. Ordinal scales are things like clothing sizes. They have names and ordering, but you cannot perform quantitative operations on them. Interval and ratio scales are quantitative numerical scales and admit operations like subtraction.

Scales of Measurement

Scale	Type	Properties
Nominal	Label	Qualitative, no ordering
Ordinal	Label or Number	Ordered, but not otherwise comparable
Interval	Number	Differences can be compared
Ratio	Number	Has a fixed zero, can divide one value by another

As we go down the list, we accumulate valid operations that we can do to the data.

Scales of Measurement

Scale	Operations
Nominal	=
Ordinal	=, <
Interval	=, <, -
Ratio	=, <, -, /

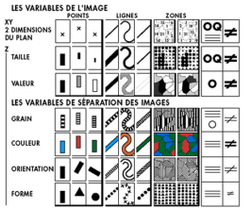
For measurements on a nominal scale, we can determine whether they are equal, but there are no other valid mathematical operations. For ordinal scales, we can check equality and see if values are greater than or less than one another. Interval scales admit these two operations and subtraction as well: we can subtract two values and get a meaningful quantity. Finally, for ratio scales, we can do all of the above and division as well.

Selecting Mappings

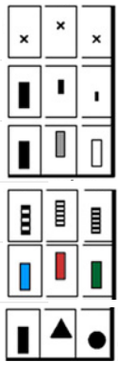
Our goal of visualizing data is to let the eye do quantitative operations on the data. So it makes sense that we need to take the operations allowed on the data based on scale of measurement into account when making a visualization.

Sémiologie Graphique

Jacques Bertin Sémiologie Graphique (1967)



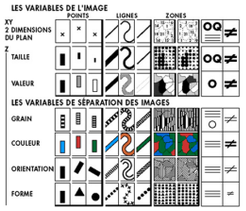
Variable	Operations
Position	=, <, -
Size	=, <, -
Brightness	=, <, -
Texture	=, <
Hue	=
Shape	=



Bertin created a catalogue of visual variables that can be manipulated, which we have seen, and furthermore, he considered the kinds of operations that we can do with those visual variables using our eyes. For shapes, we can just identify them. The same for hues. Brightness, size, and position can be identified and can also be ordered, and we can perceive differences in these quantities as well.

Sémiologie Graphique

Jacques Bertin Sémiologie Graphique (1967)



Variable	Kind of Data
Position	Nominal, Ordinal, Quantitative (Interval/Ratio)
Size	Nominal, Ordinal, Quantitative (Interval/Ratio)
Brightness	Nominal, Ordinal, Quantitative (Interval/Ratio)
Texture	Nominal, Ordinal
Hue	Nominal
Shape	Nominal

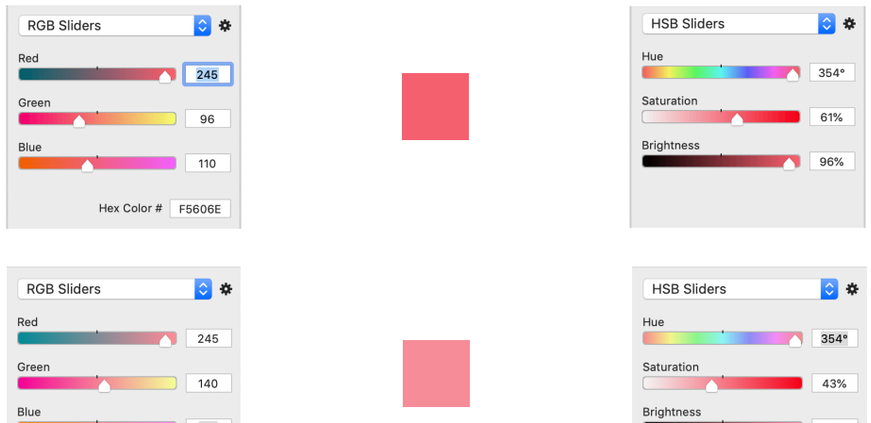
Putting this information together, he created a catalogue of possible mappings of data to visual variables that obey the constraint that the valid operation on the scale of measurement of the data should be possible with the visual variable as well.

Aside: Color Spaces

Name	Expansion	Usage
RGB	Red, Green, Blue	Light / Computer systems
CMYK	Cyan, Magenta, Yellow, Black	Pigment / Printing
HS [B,L,V]	Hue Saturation Brightness/Lightness/Value	Controlling h, s, b independently

Let me say a little more about the distinction between color and hue. In casual language, they are often used interchangeably, but in more precise terms, color is the result of a combination of primary objects. The primary objects depend on the color space used to describe colors. RGB is a common color space. It is convenient for describing colors for machines that work with light. CMYK is another common color space that is used in printing. Both of these are convenient for machines, but not easy for people to reason in. HSB (hue, saturation, brightness) is a much easier color space to think in.
In HSB, the primary objects are Hue, which can be thought of as the “pure” color being represented, Saturation, which ranges between unsaturated (white) and fully saturated, and Brightness (or Lightness or Value), which ranges between dark (black) and fully bright.

Aside: RGB vs. HSB



Here is an example of the same colors specified in RGB and HSB. Looking at the squares, we see that they are both variations on the same color, the bottom one being a pastel version of the top one.
Looking at the RGB values, it it not at all obvious that the two colors are similar. But, in the HSB space, it is clear: they have the same hue and brightness, the second color is is less saturated, which makes it pastel.

Although Bertin didn't distinguish between color, hue, and saturation, later visualization researchers do apply that distinction.

Choosing a Mapping

Year to position or fuel efficiency to position?

Data Variable	Kind	Visual Variable
Cty	Quantitative	X? Y? Size? Brightness?
Hwy	Quantitative	X? Y? Size? Brightness?
Year	Nominal	X? Y? Size? Brightness? Texture? Hue? Shape?

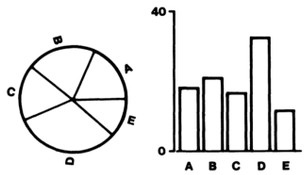
I gave an example of comparing automobile fuel efficiency in 1999 vs. 2008. Let me tell you a bit more about the data: I have three data variables – city fuel efficiency, highway fuel efficiency, and year. The first two are quantitative variables, the third, in this case, is a nominal variable.

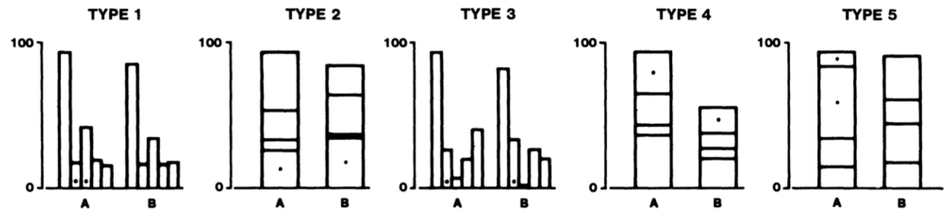
Using what we learned, we can reduce the set of options to consider. We know we cannot map a quantitative variable to shape. So, for city and highway fuel efficiency, we have four possible visual variables to control. For year, we have seven possibilities.

Are all possible combinations equally good, or can we narrow down our options any further?

Graphical Perception

Cleveland & McGill *Graphical Perception* (1984)





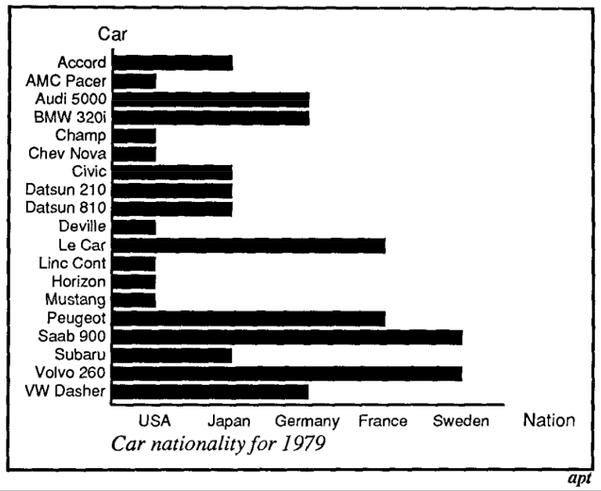
Images from *Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods*, William S. Cleveland and Robert McGill, Journal of the American Statistical Association, Vol. 79, No. 387 (Sep., 1984), pp. 531-554

This is the question that William Cleveland and Robert McGill looked at in the 1980s. Whereas Bertin evaluated mapping strategies analytically, and identified all possible mappings, Cleveland and McGill set out to acquire quantitative data comparing the efficacy of these possible mappings.

They had subjects perform quantitative tasks using the same data presented with different mappings and evaluated the subjects' performance in terms of speed and accuracy. What they found, is that not all valid mappings are perceived equally effectively.

Poor Mapping (nominal -> length)

From Mackinlay *Automating the Design of Graphical Presentations of Relational Information* (1986)

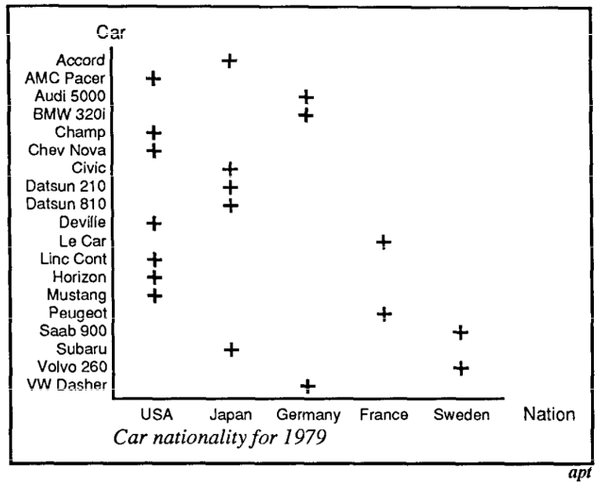


I'm following Jock Mackinlay's exposition of Cleveland & McGill. Mackinlay put together a good example to illustrate this point.

According to Bertin, I can map a nominal variable to length. It is a valid mapping because I can compare equality of length. But this mapping is difficult to read. **How many cars in this list are from Germany?** (3)

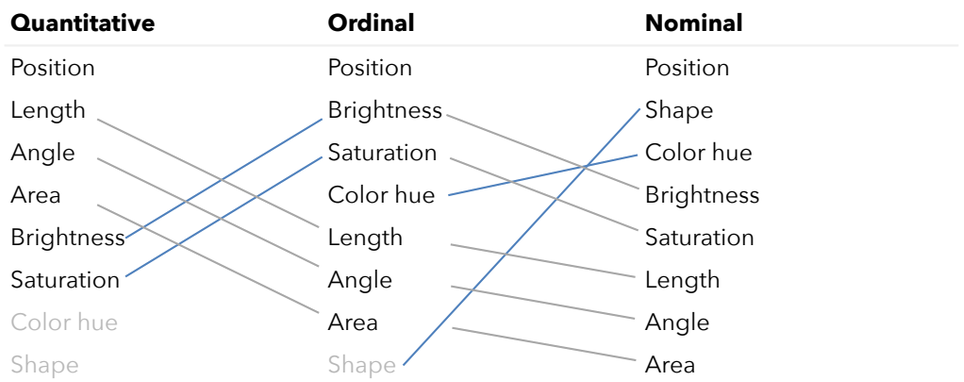
Good Mapping (nominal -> position)

From Mackinlay *Automating the Design of Graphical Presentations of Relational Information* (1986)



A nominal variable can also be mapped to position, and this is clearly a better choice. **How many cars are from Japan?** (5) Was that easier and faster to realize than in the previous visualization?

Ranking Mappings

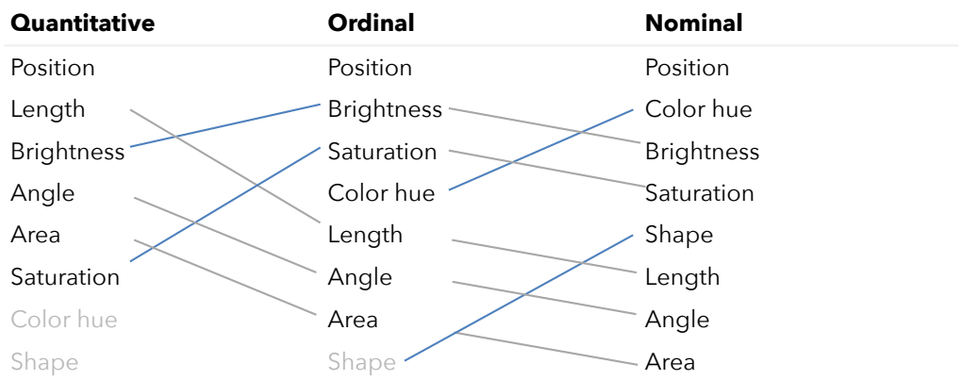


Redrawn from https://www.tableau.com/sites/default/files/whitepapers/designing-great-visual-communications.pdf

Here is Mackinlay's slopeograph summarizing the research ranking mappings.

We see here that position turns out to the be the best mapping for any kind of variable for carrying out quantitative tasks. For the other variables, the ordering shifts around. For example, Mackinlay identifies length as the second best option for quantitative variables, whereas the second best for ordinal variables is brightness, and for nominal variables it is shape.

Alternative Ranking

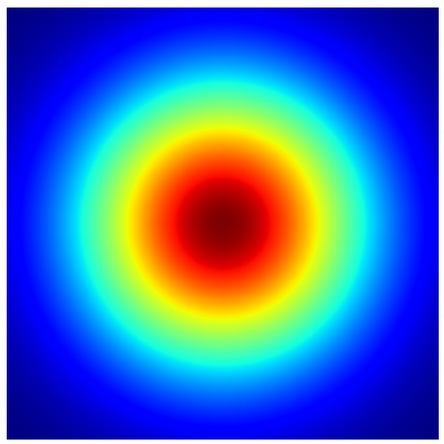


Inspired by https://www.tableau.com/sites/default/files/whitepapers/designing-great-visual-communications.pdf

From my experience, I have my own personal rankings that look slightly different. For example, I think brightness can work very well for quantitative variables, and I prefer color over shape for nominal variables, but for the most part, my experience agrees with Mackinlay's ranking.

Respect the Data

Visual inferences should be valid data inferences.



This is a good moment to provide a reminder to evaluate the validity of what you are seeing. Visualizations should allow you to process data with your eyes, and the things you see should be justified by the data.

Because of the way our eyesight works, we are compelled to perceive visual information in certain ways. To make quantitative reasoning possible through your eyes, it is necessary to choose mappings that ensure that visual inferences are valid data inferences. Look at the above visualization. It is constructed by mapping data to x, y, and color. How do you expect the underlying data to be distributed? Do you see contours and bands of similar values. We will later look back at this and see if the data matches, and see if this visualization does a good job of suggesting inferences that are valid data inferences.

Visual inferences should be valid data inferences (Agrawala).

Example

Let us use these principles to produce a visualization. Visualizations should be motivated by questions, and the question we are trying to answer helps us make choices about which mappings we want to focus on.

Example Data: Automobile Details

manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact
audi	a4	2	2008	4	manual(m6)	f	20	31	p	compact
audi	a4	2	2008	4	auto(av)	f	21	30	p	compact
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact
audi	a4	2.8	1999	6	manual(m5)	f	18	26	p	compact
audi	a4	3.1	2008	6	auto(av)	f	18	27	p	compact
audi	a4 quattro	1.8	1999	4	manual(m5)	4	18	26	p	compact
audi	a4 quattro	1.8	1999	4	auto(l5)	4	16	25	p	compact
audi	a4 quattro	2	2008	4	manual(m6)	4	20	28	p	compact
audi	a4 quattro	2	2008	4	auto(s6)	4	19	27	p	compact
audi	a4 quattro	2.8	1999	6	auto(l5)	4	15	25	p	compact
audi	a4 quattro	2.8	1999	6	manual(m5)	4	17	25	p	compact
dodge	caravan 2wd	3.8	1999	6	auto(l4)	f	15	22	r	minivan

Here is an excerpt from the data I have referred to several times. It contains information about a collection of automobiles from two different years: 1999 and 2008. For each car, the data set includes the manufacturer, model, engine displacement, year of production, number of cylinders, transmission, drivetrain, city and highway fuel efficiency, fuel type, and class of automobile.

Task

Given fuel efficiency data for automobiles in 1999 and 2008, answer the following questions:

Are 1999 cars more or less efficient than 2008 cars?

Did individual models become more or less efficient?

To answer these questions, we need to make a visualization that displays three variables: **city fuel mileage**, **highway fuel mileage**, and **year**.

We will use automobile fuel-efficiency data to investigate the following question: Are 1999 autos more or less efficient than 2008?

To answer these questions, we need to make a visualization that displays three variables: city fuel mileage, highway fuel mileage, and year.

Textual Summary

In *1999*, mean city gas mileage was **17.02 MPG**, compared to **16.70 MPG** in *2008*

In *1999*, mean highway gas mileage was **23.43 MPG**, compared to **23.45 MPG** in *2008*

One way to respond is to compute a mean and just say what the answer is. This is succinct, accurate, and easy to understand. When seen this way, it looks like city fuel efficiency declined a little, and highway fuel efficiency was unchanged (the values are in MPG, which inverted compared to the L/km units which are used in Europe).

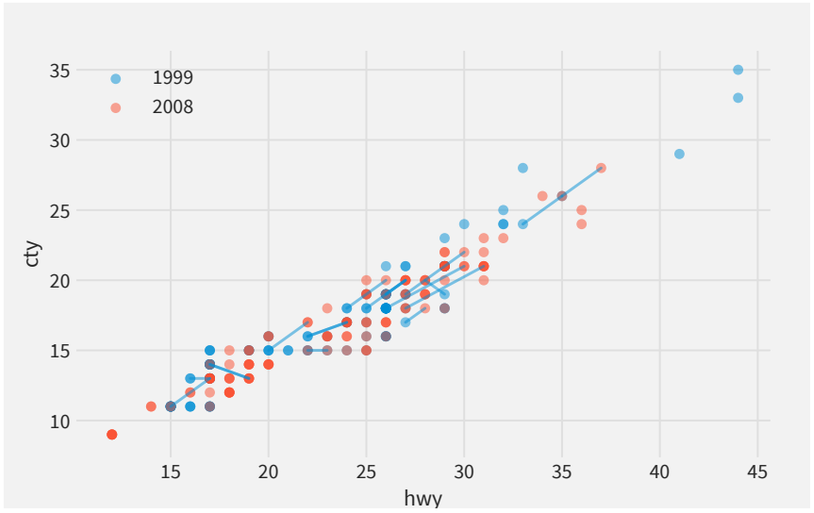
But it is not very convincing. You can either believe it or not, but you cannot question the conclusion or look at it in any other way. It doesn't show how much data was evaluated to come to this conclusion. Did I look at 2 cars, 200?

Table

	1999	2008
Cty (MPG)	17.02	16.70
Hwy (MPG)	23.43	23.45

Alternatively I can state the results as a table. This makes the numerical values easier to spot an compare, but it still has the same weaknesses as the sentence: it does not show us much data. It's not possible to see if there were outliers in the data that may have influenced the mean, for example.

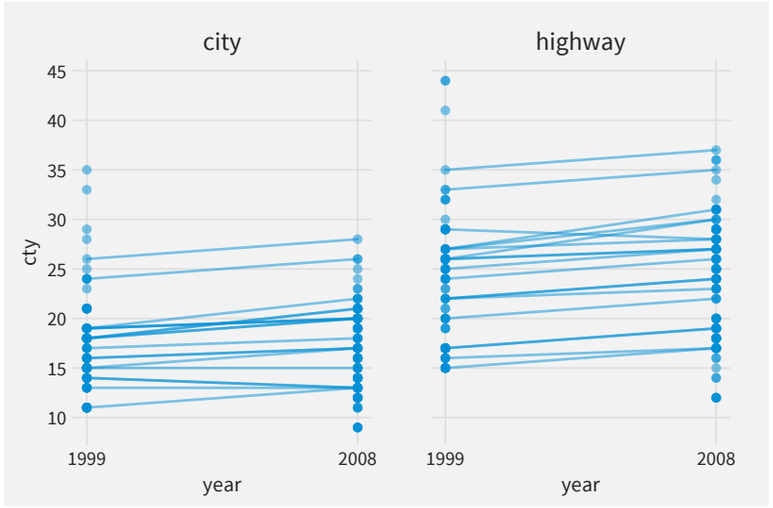
x(cty), y(hwy), color(year)



Another approach is to make a scatterplot with city mileage on the x axis, highway milage on the y axis, and use color for year. I've also drawn lines that connect cars of the same model between the different years.

This shows all the data and makes it possible to see things like outliers, but it's not so easy to see if 1999 cars were more efficient than 2008 cars.

x(year), y(cty/hwy)



We know that the best visual variable for carrying out quantitative tasks with all types of measurements is position. Since year is a key variable for the comparison we want to make, I can give it centrality by mapping year to position. Again, I have drawn lines that connect the same model when it appears in both years.

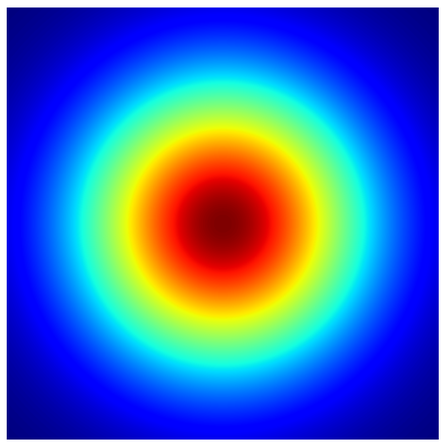
Here we again see all the data, and we can also easily perform quantitative operations. We see, visually, that the mean city fuel efficiency in 2008 was slightly lower, but we also see that there were some outliers that pull up the 1999 means. These cars do not appear in 2008. Why not? We can ask questions like, “Was our result an artifact of sampling bias?”

Color

Let us turn our attention to color.

Respect the Data

Visual inferences should be valid data inferences (Agrawala)

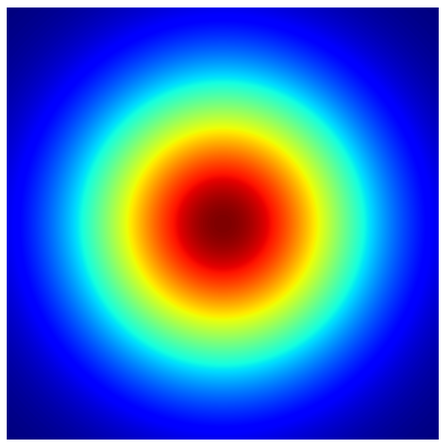


Here is the plot from earlier again. What do you think the data looks like? Are there bands of similar values? Are values in the central dark red region greater or less than those in the yellow ring, and are those greater or less than those in the bright blue ring?

If we were to make a surface from this data, where would it be high and where would it be lower?

Respect the Data

Visual inferences should be valid data inferences (Agrawala)



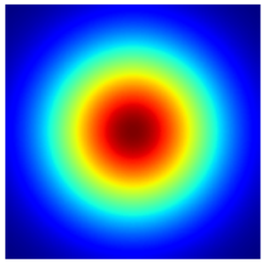
Here is the same data shown using a grayscale colormap on the right. The choice of colormap can greatly influence interpretation. The grayscale image shows that there are no bands of data at all. The variable varies continuously from the center to the edges.

Pure hue is not a good mapping for quantitative variables because of the way we perceive hues. Grayscale is better.

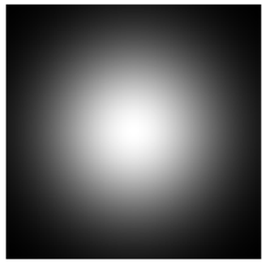
Color Palettes

Visual inferences should be valid data inferences (Agrawala)

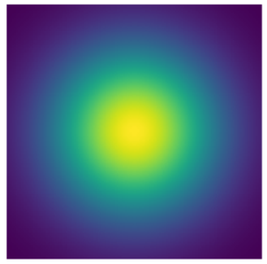
jet



gray



viridis



Jet and rainbow are terrible colormaps for visualization, and grayscale is a safe choice. But we are not doomed to black and white, which might be too visually bland. We can use color if we use a carefully designed palette like viridis. It is more colorful than grayscale, but still designed to be interpretable. And if you look at it, you get a similar visual impression.

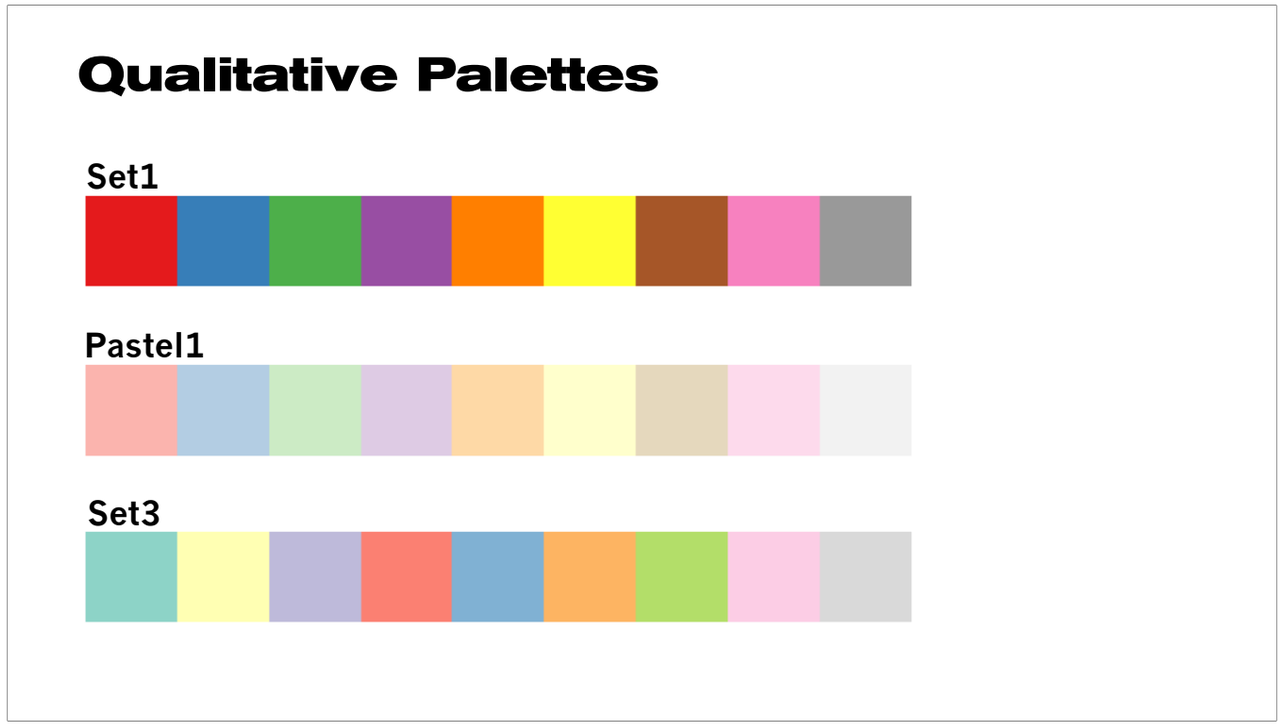
Kinds of Palettes

Qualitative for categorical data

Diverging for numerical data with a clear central point

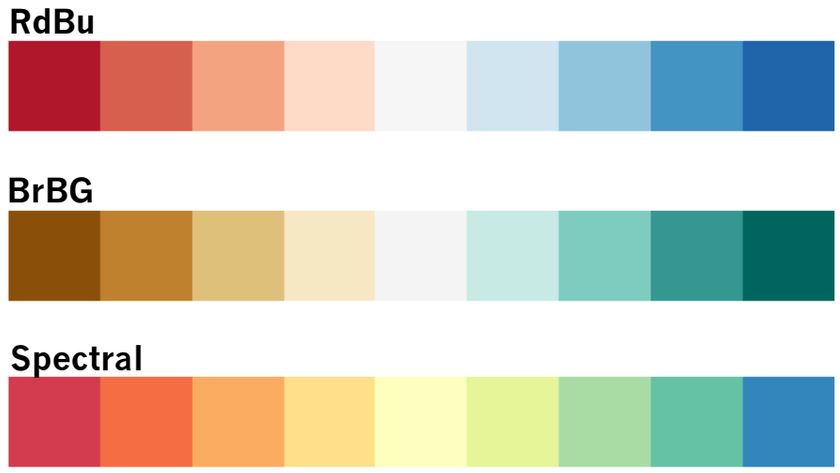
Sequential for numerical data otherwise

Palettes come in three different flavors. Palettes for categorical data are called qualitative. Those for numerical data are either diverging, if the data has a clear central point or sequential otherwise.

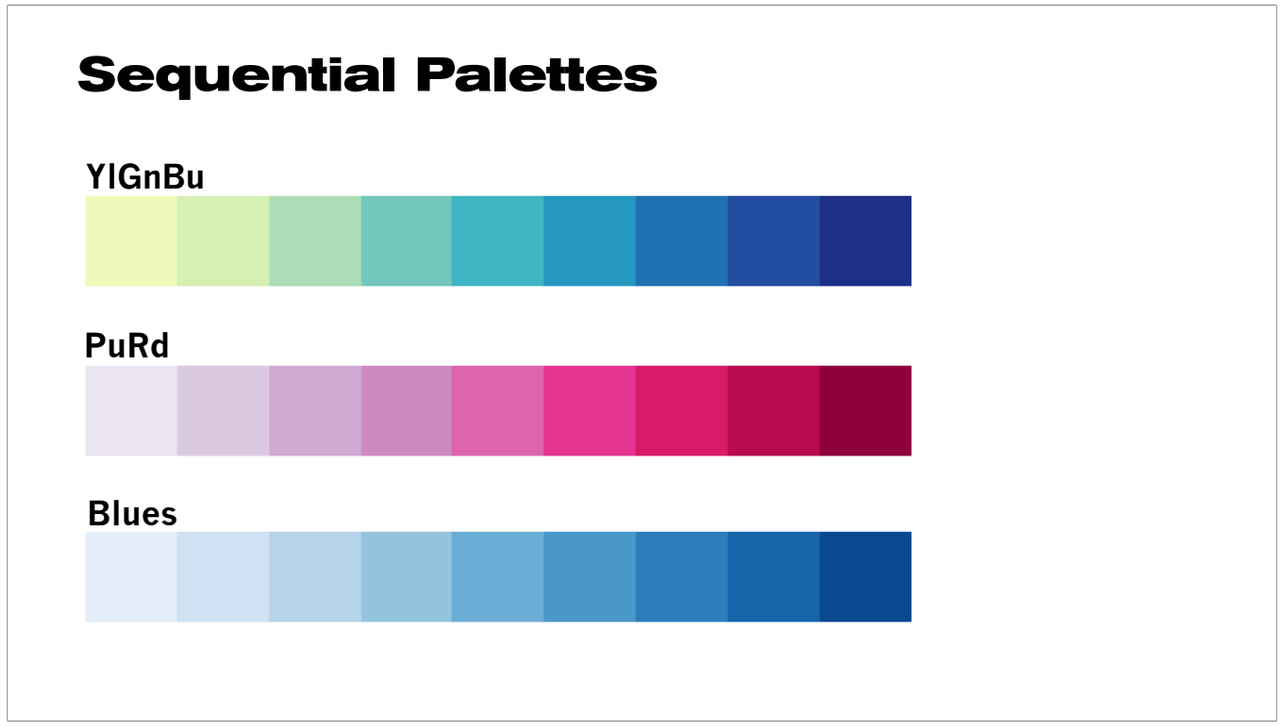


Qualitative palettes look like this. They are designed to make it easy to identify matching values and to avoid implying ordering.

Diverging Palettes



Diverging palettes are designed to make it easy to identify a central point and if the values are above or below that point.



Sequential palettes are designed to emphasize ordering of values.

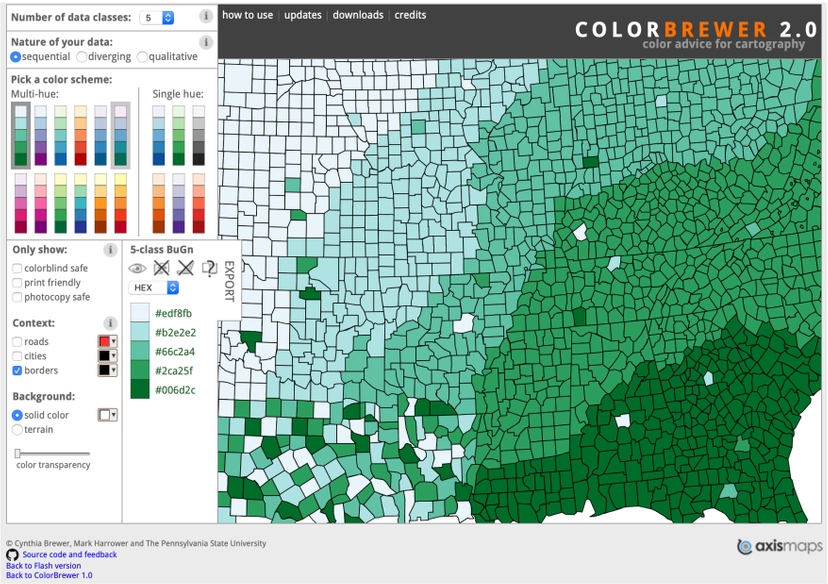
Color Advice

- Can map **quantitative** variables to **brightness**
- Can map **nominal** variables to **hue** (but remember, some people are colorblind)
- Use a **carefully designed** palette
- If unsure, try a **grayscale** or a **monochrome** palette

You can map quantitative variables to brightness, and nominal variables to hue. We saw that with a poor palette choice, it is possible to confuse our ability to do quantitive operations. If you use color, use a carefully designed palette.

If you see something in a colorful visualization and are uncertain as to whether the phenomenon is really in the data, you can use a grayscale or monochrome palette, varying just brightness, to verify.

ColorBrewer



Where do you get a carefully designed palette? My favorite place is using ColorBrewer.

Cynthia Brewer is a cartographer who has designed very good palettes for a variety of situations and made them available as ColorBrewer. The previous examples of color palettes all come from her. On her site, you can find color palettes suitable for a variety of situations, including color-blindness-safe palettes. The ColorBrewer palettes widely available; they are built into matplotlib, ggplot, vega, and bokeh, for example.

Use those or another source of well designed color palettes to ensure that you do not use color in a confusing way.

Example

Let us take a look at an example of working with color.

Example Data: Antibiotics

bacteria	gram	antibiotic	mic	genus
Mycobacterium tuberculosis	negative	penicillin	800.000	Mycobacterium
Mycobacterium tuberculosis	negative	streptomycin	5.000	Mycobacterium
Mycobacterium tuberculosis	negative	neomycin	2.000	Mycobacterium
Salmonella schottmuelleri	negative	penicillin	10.000	Salmonella
Salmonella schottmuelleri	negative	streptomycin	0.800	Salmonella
Salmonella schottmuelleri	negative	neomycin	0.090	Salmonella
Proteus vulgaris	negative	penicillin	3.000	Proteus
Proteus vulgaris	negative	streptomycin	0.100	Proteus
Proteus vulgaris	negative	neomycin	0.100	Proteus
Klebsiella pneumoniae	negative	penicillin	850.000	Klebsiella
Klebsiella pneumoniae	negative	streptomycin	1.200	Klebsiella
Klebsiella pneumoniae	negative	neomycin	1.000	Klebsiella
Brucella abortus	negative	penicillin	1.000	Brucella

Here is data on the *minimum inhibitory concentration* (mic) of three antibiotics, *neomycin*, *streptomycin*, and *penicillin* against various bacteria.

Task

Given data on the concentration of antibiotic necessary to inhibit growth of bacteria:

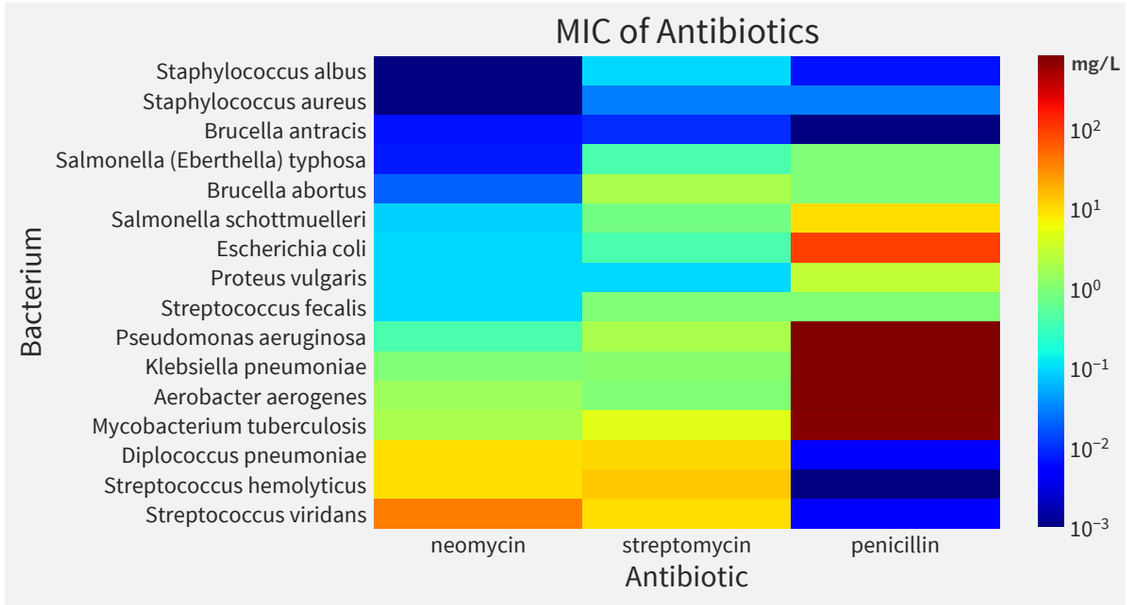
Approximately what concentration of antibiotic is needed against each species of bacteria?

To answer these questions, we need to make a visualization that displays three variables: **species of bacteria**, **antibiotic**, and **minimum inhibitory concentration**.

We will use our antibiotic data to investigate the following question: How effective is each antibiotic against each species of bacteria?

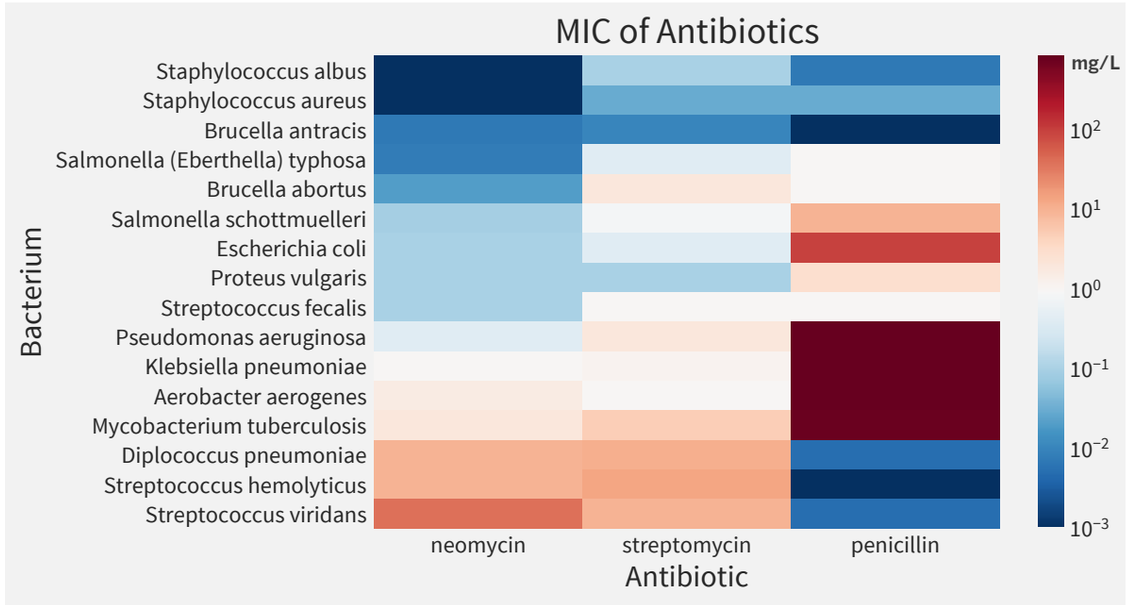
To answer these questions, we need to make a visualization that displays three variables: species of bacteria, antibiotic, and minimum inhibitory concentration.

Jet Colormap



Is the MIC of neomycin against Aerobacter aerogenes more or less than 1 mg/L?

ColorBrewer



Same question: is the MIC of neomycin against Aerobacter aerogenes more or less than 1 mg/L?

Color Advice

Above all do no harm.
Edward Tufte

The hippocratic oath for color in visualizations. Whatever you do, when you use color, make sure you are not damaging your visualizations.

Transformations

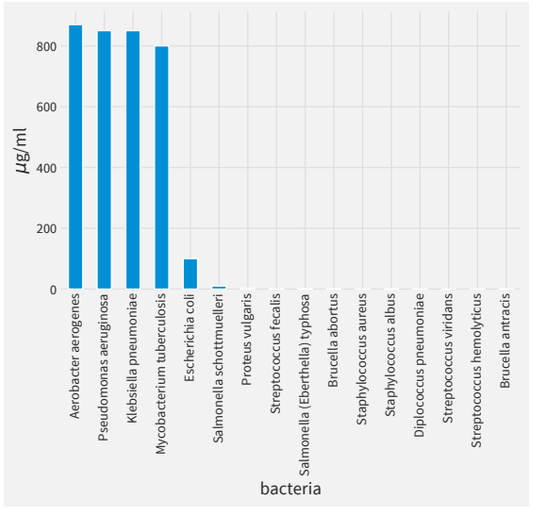
Today, we have been focusing on mapping data to marks. And the last aspect of of this process we will consider today is transformation.

Sometimes it might be necessary to transform data as part of the mapping step because the visualization or the data require it to produce an interpretable result.

Transformations (after Jeff Heer)

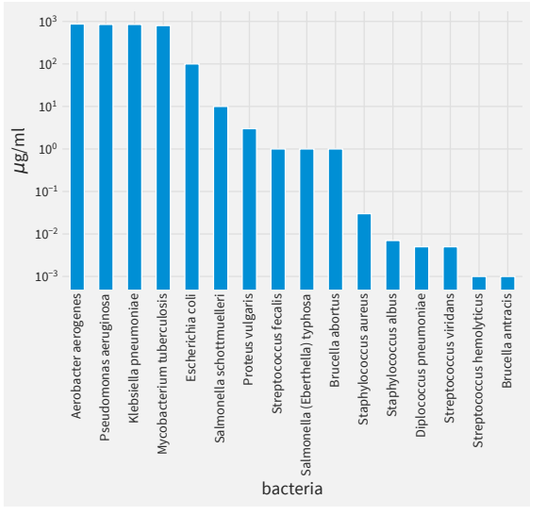
Operation	Implementation
Normalize	e.g., y[i] = y[i] / sum(y), z-score, etc.
Log	log(y)
Deviation from Reference	y = x - reference
Power	pow(y, k)
Box-Cox Transform	pow(y, k) - 1 / k if: k != 0 else: log(y)
Inversion	1/y
Binning	quantiles, histograms
Grouping	e.g., merge categories
Model-space	Transform values to the space of your model of the data

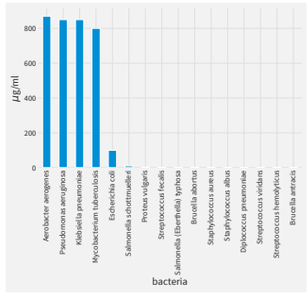
Wide Dynamic Range



One such case is data with a wide dynamic range. For example, consider this bar chart of the minimum inhibitory concentration of penicillin for a variety of species of bacteria. It looks like many bacteria have a mic of 0, but that, of course, makes no sense. By adding a log transform, we can better see the full range of data.

Wide Dynamic Range

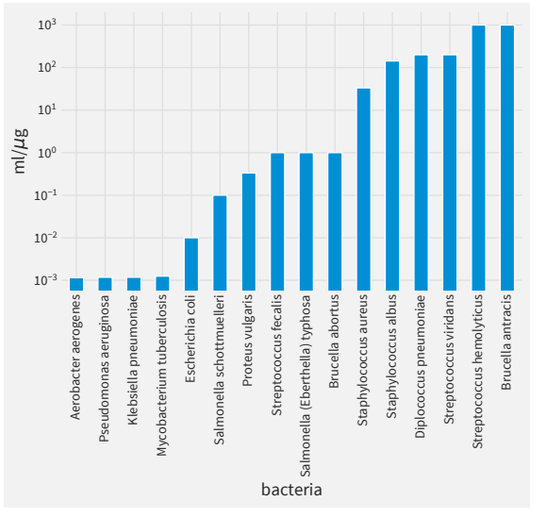


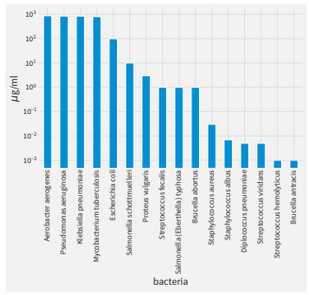


With the addition of a log transform, we see that the MIC is never zero, but sometimes very small. Penicillin is very effective against these bacteria.

We lose some resolution in the high end of the data, but it doesn't matter if the MIC is 800 or 820 $\mu g/ml$. It is more important that we see the differences in the lower end of the range. A dose of 0.001$\mu g/ml$ is quite different from 1$\mu g/ml$.

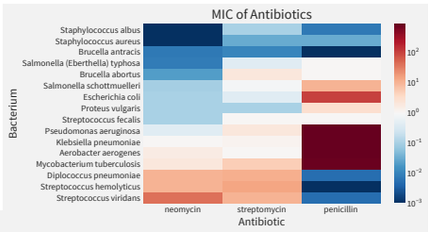
Invert

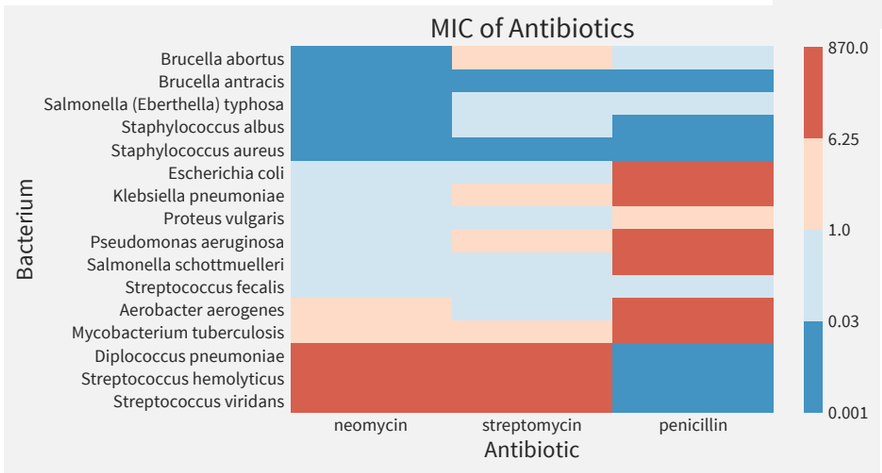




In the graphs above, our eye is drawn to the larger bars. But these are the bacteria for which penicillin performs the **worst**. By inverting, we can draw our eye to the bacteria for which penicillin performs **best**. By inverting, you can avoid having to add a *smaller is better / larger is better. Just invert the units and keep things consistent and intuitive.

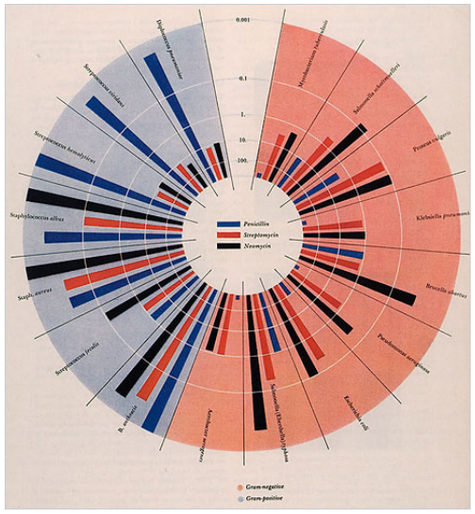
Quantize





Quantization can help answer certain questions by reducing the number of possibilities. If you are just interested in knowing is an antibiotic good or bad. Applying quantization, we can easily identify which is a reasonable (though maybe not the best) choice of antibiotic against different bacteria.

X-Form Tour de Force



http://graphics8.nytimes.com/images/2008/06/01/books/heller-1.jpg

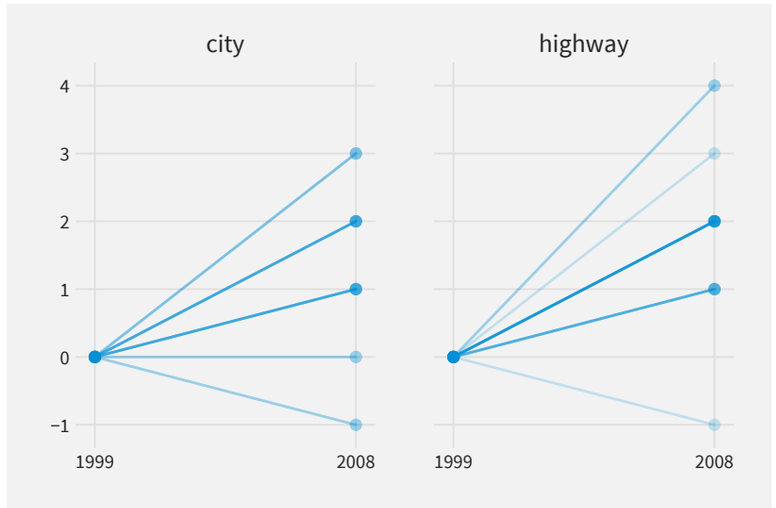
In 1951, Will Burtin published this famous visualization which applies several transformations to make an engaging visualization of antibiotic efficacy. In addition to a log transformation of the MIC, Burtin inverts the scale, so that more effective antibiotics are more prominent, and wraps the bar chart around a circle.

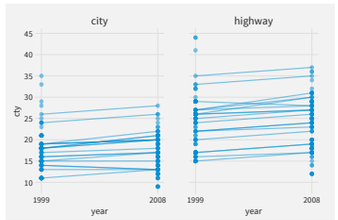
Example

Deviation from Reference

Another useful transformation is to compute a deviation from a reference value. We will apply this transformation to the fuel efficiency data for cars of the same model. We will set the reference to the fuel efficiency in 1999. By transforming this way, we can see changes more clearly.

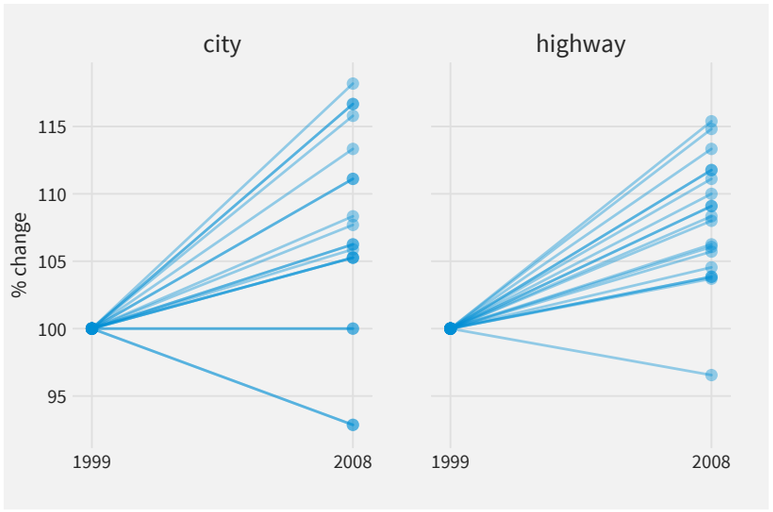
x(year), y(cty/hwy)

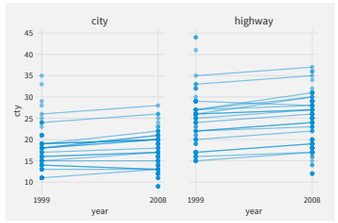


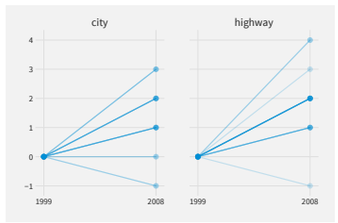


Earlier, we looked at this visualization of automobile fuel efficiency in 1999 and 2008. If we want to emphasize the changes in that occurred, we can, for each model that appears in 1999 and 2008, we can set the 1999 value to 0 and look at only the change between the years. Here we see clearly, that almost all cars improved their fuel efficiency in this time period.

x(year), y(cty/hwy)



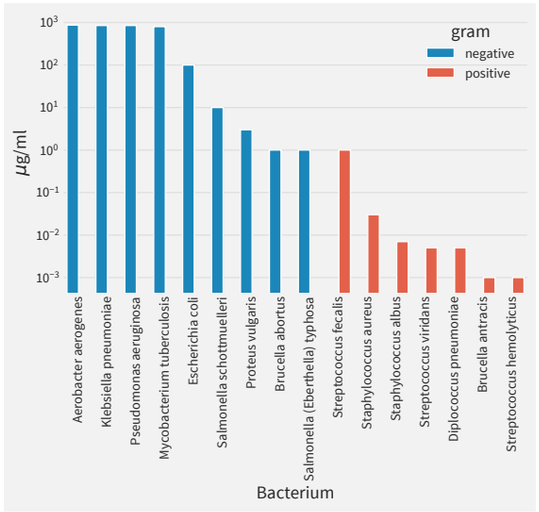


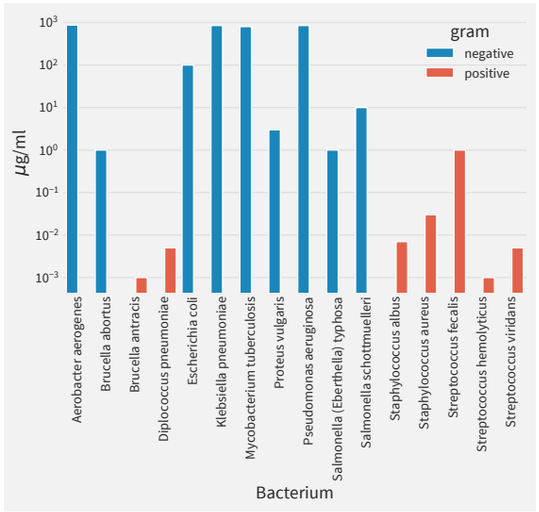


And, if we want a finer-grained understanding of evolution of fuel efficiency, we can also apply another transformation and convert differences to percentages. A 1 mpg improvement going from 10 to 11 mpg is much more substantial than a 1 mpg improvement going from 27 to 28 mpg.

Protip: Sorting

Sort Data





Sorting data according to some meaningful variable makes visualizations easier to read understand. Compare these two visualizations of the MIC of penicillin. In the first one, the bacteria are sorted according to MIC, in the second, it they are sorted alphabetically, which is not a particularly relevant property of the data.

TLDR

- Thinking about **visualizations as mapping** makes it easier to evaluate options
- Using an **appropriate color palette** (ColorBrewer, etc) will improve clarity
- **Employ transformations** to improve perception
- **Sort and structure the data** according to meaningful properties

We are approaching the end of time for today. What did we do over the last 90 minutes? We developed a conceptual framework that makes it possible to design a visualization and evaluate visualizations. We gave some extra attention to color because it is an important aspect of data visualization that is easy to get wrong if you are not careful, but also easy to get right if you draw on some support. Finally, we looked at applying transformations to data to make visualizations easier and more accurate to interpret.

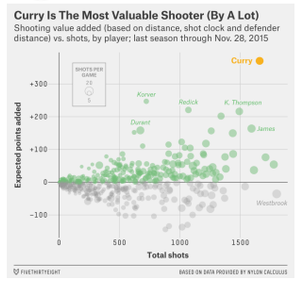
At the heart of quantitative reasoning is a single question: **Compared to what?**

[Good visualizations allow] visually enforcing comparisons of changes, of the differences among objects, of the scope of alternatives.

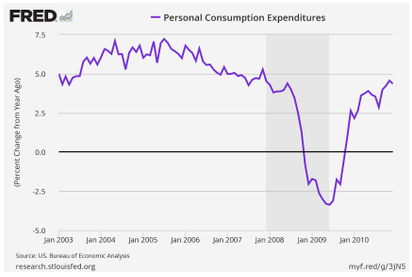
Edward Tufte, *Envisioning Information*, p. 67

Part II — Layering, Faceting & Context

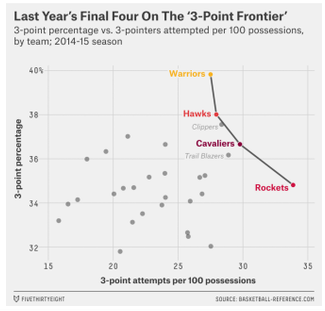
https://medium.com/@cramakrishnan/the-delong-shiller-redux-dc9dd21eefd1



http://fivethirtyeight.com/features/stephen-curry-is-the-revolution/



http://myf.red/g/3jN5



Thank You!