

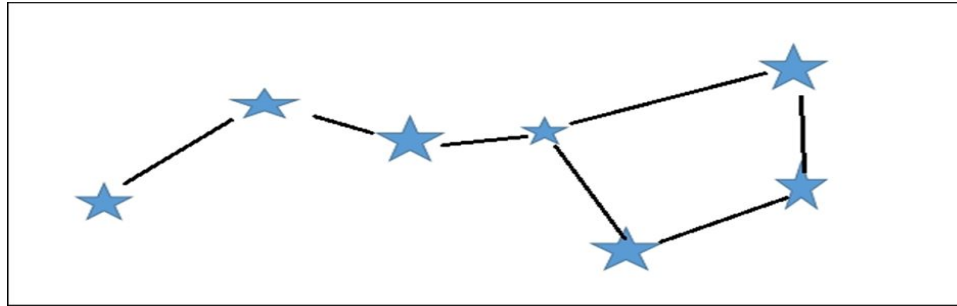
# BIG DATA VISUALIZATION

# Introduction: Data Visualization

Data means information/numbers

Visualization means picturing, or picturing the information

Data visualization is to make something complex appear simple



Connect stars to create a picture to help one visualize the constellation

# Motives for using Data Visualization

- To explain the data or put the data in context
- To solve a specific problem
- To explore the data to reach a better understanding or add clarity
- To highlight or illustrate otherwise invisible data
- To predict, for example, potential sales volumes

## *As per Data Visualization: The future of data visualization, Towler, 2015:*

*"Data visualization is entering a new era. Emerging sources of intelligence, theoretical developments, and advances in multidimensional imaging are reshaping the potential value that analytics and insights can provide, with visualization playing a key role."*

# For visualization one typically needs to consider

- The size and volume of the data to be visualized.
- The data's cardinality and context.
- What is it you are trying to communicate? What is the point that you want to communicate?
- Who is your audience? Who will consume this information?
- What kind or type of visual might best convey your message to your audience?

# The most common visualization methods

- Table
- Histogram
- Scatter plot
- Line, bar, pie, area, flow, and bubble charts
- Data series or a combination of charts
- Time line
- Venn diagrams, data flow diagrams, and **entity relationship (ER)** diagrams

# Challenges of big data visualization

## Big Data:

- A large assemblage of data and datasets that are so large or complex that traditional data processing applications are inadequate.
- In 2001, then Gartner analyst Doug Laney introduced the 3Vs concept.
- The 3Vs, according to Doug Laney, are volume, variety, and velocity.
- The 3Vs make up the dimensionality of big data: volume (or the measurable amount of data), variety (meaning the number of types of data), and velocity (referring to the speed of processing or dealing with that data).

# Challenges of big data visualization

Excel to gauge your Data:

- Excel is not a tool to determine whether your data qualifies as big data





# Big Data: The 3Vs - Volume, Variety and Velocity

## **Volume:**

How much of something there will be?

With every click of a mouse, big data grows to be petabytes (1,024 terabytes) or even Exabyte's (1,024 petabytes) consisting of billions to trillions of records generated from millions of people and machines

# Big Data: The 3Vs - Volume, Variety and Velocity

## **Velocity:**

Velocity is the rate or pace at which something is occurring.

The measured velocity experience can and usually does change over time.

Velocities directly affect outcomes.

With Internet of Things (IoT), this pace will only quicken.

# Big Data: The 3Vs - Volume, Variety and Velocity

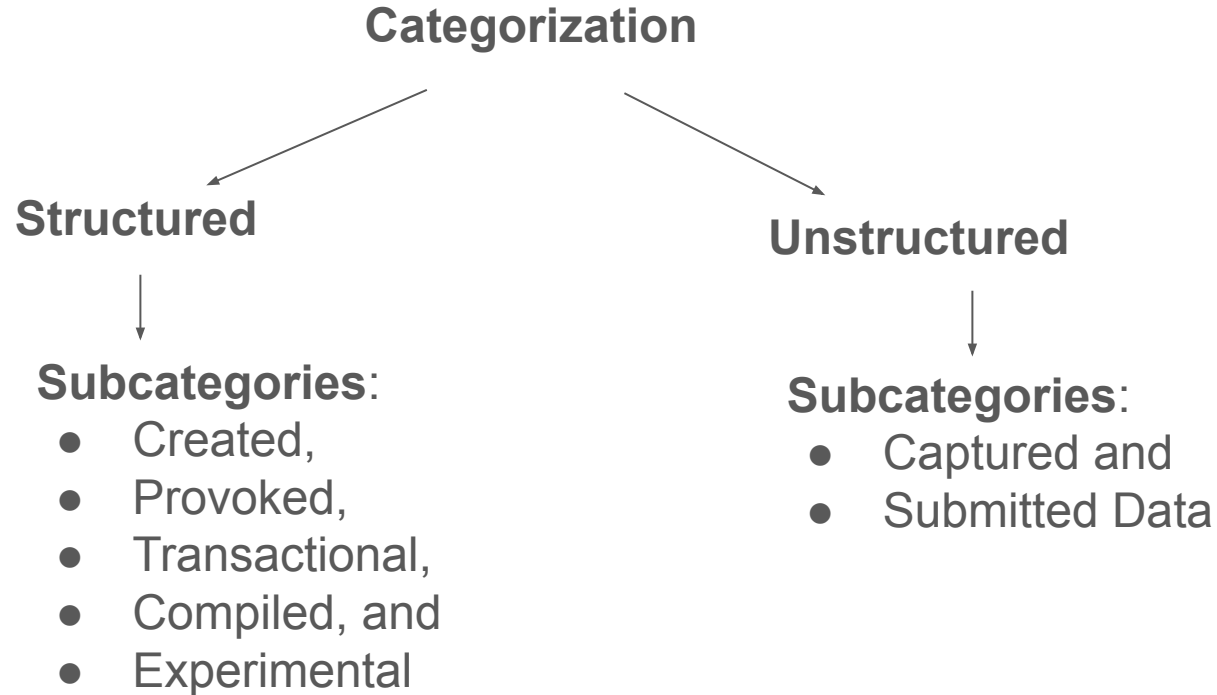
## Variety:

Relational databases are considered to be highly structured (contain text in `VCHAR`, `CLOB`, or `BLOB` fields).

Data today (and especially when we talk about big data) comes from many kinds of data sources, and the level in which that data is structured varies greatly from data source to data source.

Formats that go beyond pure text, photo, audio, video, web, GPS data, sensor data, relational databases, documents, SMS, pdf, flash, and so on.

# Categorization



# Various data formats (Varieties)

1. **Created data:** This is the data being created for a purpose; such as **focus group surveys or asking website users to establish an account on the site** (rather than allowing anonymous access).
2. **Provoked data:** This is described as **data received after some form of provoking**, perhaps such as providing someone with the opportunity to express the **individual's personal view on a topic, such as customers filling out product review forms**.
3. **Transactional data:** This is data that is described as **database transactions**, for example, the record of a sales transaction
4. **Compiled data:** This is data described as **information collected (or compiled) on a particular topic such as credit scores**.

# Various data formats (Varieties)

5. **Experimental data:** Described as when **someone experiments with data and/or sources of data to explore potential new insights**. For example, combining or relating sales transactions to marketing and promotional information to determine a (potential) correlation.

6. **Captured data:** **Data created passively due to a person's behavior** (like when you enter a search term on Google, perhaps the creepiest data of all!).

7. **User-generated data:** **Data generated every second by individuals, such as from Twitter, Facebook, YouTube, and so on** (compared to captured data, this is data you willingly create or put out there).

**Big Data comes with no common or expected format and the time required to impose a structure on the data has proven to be no longer worth it.**

# Challenges in Big Data Visualization for Big Data

Ability to effectively deal with data quality, outliers, and to display results in a meaningful way.



# Data Quality

The value of almost anything and everything is directly proportional to its level of quality and higher quality is equal to higher value.

Data is no different. Data (any data) can only prove to be a valuable instrument if its quality is certain.

The general areas of data quality include:

Accuracy	Consistency (across sources)
Completeness	Reliability
Update status	Appropriateness
Relevance	Accessibility



# Data Quality

The quality of data can be affected by the way it is **entered, stored, and managed**.

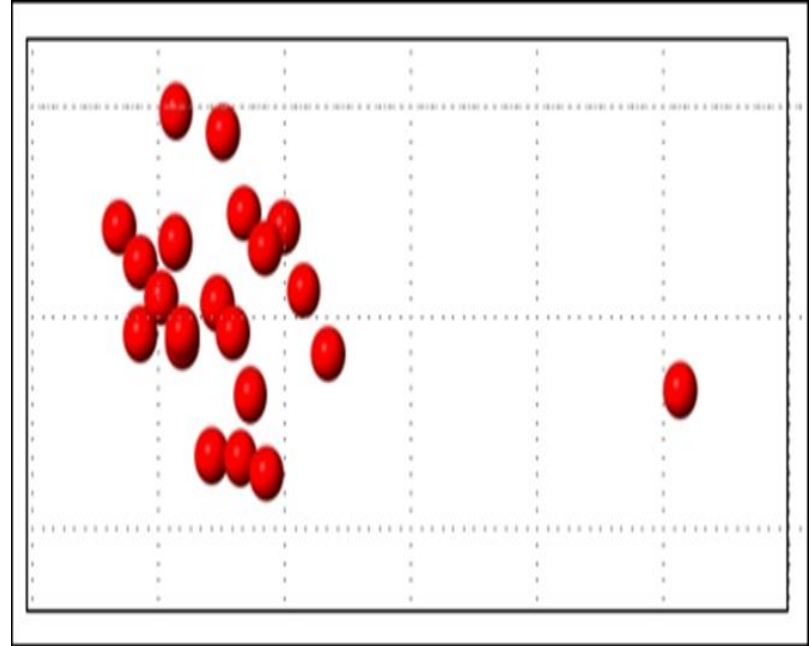
The process of addressing data quality (**Data Quality Assurance - DQA**), requires a routine and regular review and evaluation of the data, and performing ongoing processes termed **profiling and scrubbing**.

**Effective profiling and scrubbing of data** necessitates **the use of flexible, efficient techniques capable of handling complex quality** issues hidden deep in the depths of very large and ever accumulating (big data) datasets.

# Dealing with outliers

As per Sham Mustafa, founder and CEO of data scientist marketplace Correlation One:

*“Anyone who is trying to interpret data needs to care about outliers. It doesn't matter if the data is financial, sociological, medical, or even qualitative. Any analysis of that data or information must consider the presence and effect of outliers. Outliers (data that is “distant” from the rest of the data) indicating variabilities or errors – need to be identified and dealt with.”*

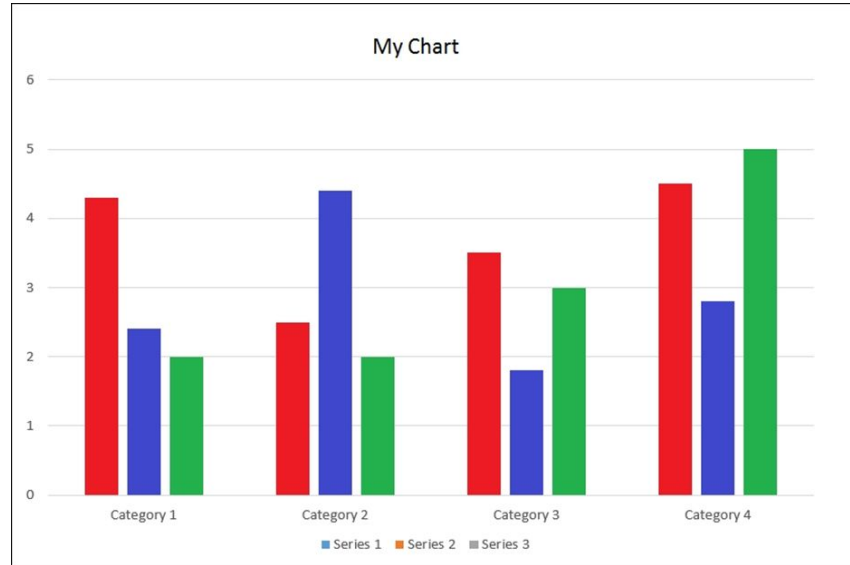


# Dealing with outliers

Methods for dealing with these outliers may be simply moving them to another file or replacing the outliers with other more reasonable or appropriate values

The outliers you identify in your data acts as an **indicator that the data itself is bad or faulty** or **are the outliers' random variations caused by new and interesting points or characteristics within your data?**

# The powers of data visualization: A picture is worth a thousand words and Seeing is believing



As per Millman/Miller Data Visualization: *"The whole point of data visualization is to provide a visual experience."*

# Data Visualization

Data visualization is a key technique permitting individuals to perform analysis, identify key trends or events, and make more confident decisions much more quickly.

In fact, data visualization has been referred to as the visual representation of business intelligence.

# Adding a fourth V?

The idea of establishing and improving the quality levels of big data: **Veracity**

Data that is disparate, large, multi formatted, and quick to accumulate and/or change (big data) causes uncertainty and doubt (can I trust this data?).

The uncertainty that comes with big data may cause the perhaps valuable data to be excluded or overlooked.

A method for dealing with big data veracity is by assigning **a veracity grade or veracity score for specific datasets to evade making decisions** based on analysis of uncertain and imprecise big data.

# Visualization philosophies

- The proper arrangement of related information
- Appropriately using color(s)
- Correctly defining decimal placements
- Limiting the use of 3D effects or ornate gauge designs

# Visualization philosophies

Variety	Velocity	Volume
Without context, data is meaningless and the same applies to visual displays (or visualizations) of that data.	The usability or value of the data will be (at least) reduced if the data is not timely.	Pitfalls: Slow performance, larger wait time, attempting to plot points for analysis with large amount of information.
For example, <b>data sourced from social media may present entirely different insights</b> depending on <b>user demographics</b> (age group, sex, or income bracket), <b>platform</b> (Facebook or Twitter), or <b>audience</b> (who consume the visualizations).	The effort and expense required to source, understand, and visualize data is squandered if the results are stale, obsolete, or potentially invalid by the time the data is available to the intended consumers	Too much information displayed in one place can cause the viewer to have what is referred to as sensory overload and also available viewing space can be detrimental to the value of a visualization trying to depict too many data points or metrics
Acquiring a <b>proper understanding</b> (establishing a context) of the data <b>takes significant domain expertise</b> as well as the ability to properly analyze the data	The challenge of speedily crunching numbers exists within any data analysis, but when considering the varieties and volumes of data involved in big data projects, it becomes even more evident.	Visualizations of data should be used to uncover trends and spot outliers much quicker



# All is not lost

There are various approaches (or strategies) that have come to exist and can be used for preparing effective big data visualizations for big data

- You can change the type of the visualization, for example, switching from a column graph to a line chart can allow you to handle more data points within the visualization.
- You can use higher-level clustering. In other words, you can create larger, broader stroke groupings of the data to be represented in the visualization
- You can remove outliers from the visualization. Outliers can be removed and if appropriate, be presented in a separate data visualization.
- You can consider capping, which means setting a threshold for the data you will allow into your visualization. This cuts down on the range or data making for a smaller, more focused image.

# Approaches to big data visualization

Big data visualization tools:

- Hadoop
- R
- Data Manager
- D3
- Tableau
- Python
- Splunk

These tools are done in an effort to meet the challenges of big data visualization and support better decision making.

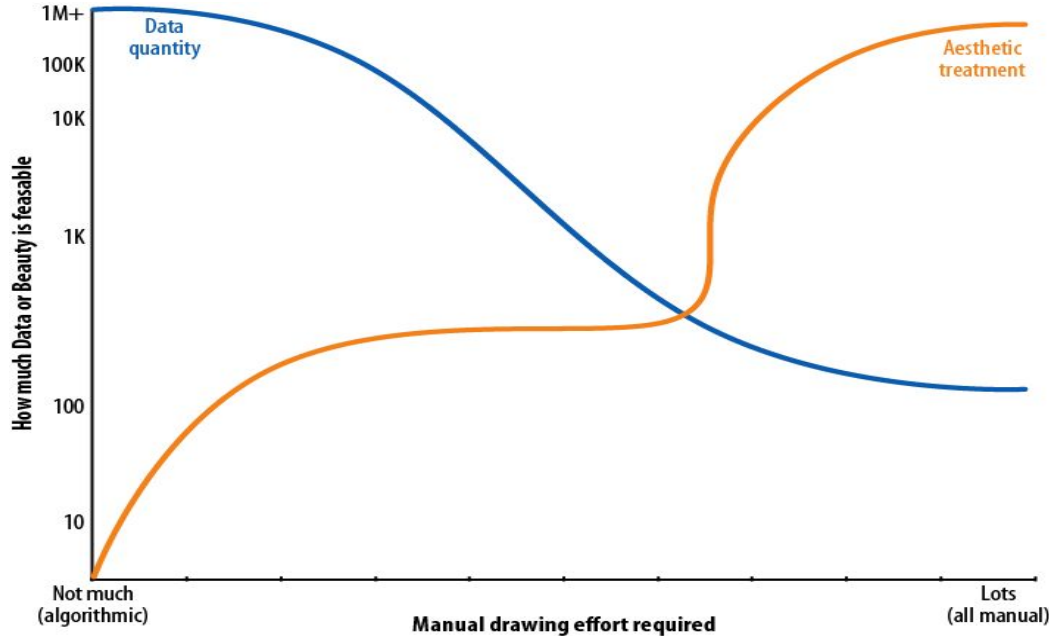
# Infographics versus Data Visualization

The terms infographics and data visualization used in different ways, or interchangeably in different contexts or with positive or negative connotations attached.

People use **infographic** to refer to representations of information perceived as casual, funny, or frivolous, and **visualization** to refer to designs perceived to be more serious, rigorous, or academic.

**The distinction between infographics and data visualizations (or information visualizations) is based on both form and origin.**

# Infographics versus Data Visualization



The difference between infographics and data visualization may be loosely determined by the method of generation, the quantity of data represented, and the degree of aesthetic treatment applied

# Infographics

The term infographic is useful for referring to any visual representation of data that is:

- Manually drawn (and therefore a custom treatment of the information);
- Specific to the data at hand (and therefore nontrivial to recreate with different data);
- Aesthetically rich (strong visual content meant to draw the eye and hold interest); and
- Relatively data-poor (because each piece of information must be manually encoded).

# Infographics

Infographics are illustrations where the data representation is manually laid out or sketched, probably with drawing software such as Adobe Illustrator.

Because of their manually-drawn process of creation, infographics have the option of being aesthetically rich.

Another consequence of their manual origins is they tend to be limited in the amount of data they can convey, simply due to the practical limitations of manipulating many data points.

Similarly, it is difficult to change or update the data in an infographic, as any changes must be implemented manually.

# Data Visualization

The terms data visualization and information visualization are useful for referring to any visual representation of data that is:

- Algorithmically drawn (may have custom touches but is largely rendered with the help of computerized methods);
- Easy to regenerate with different data (the same form may be repurposed to represent different datasets with similar dimensions or characteristics);
- Often aesthetically barren (data is not decorated); and
- Relatively data-rich (large volumes of data are welcome and viable, in contrast to infographics).

# Data Visualization

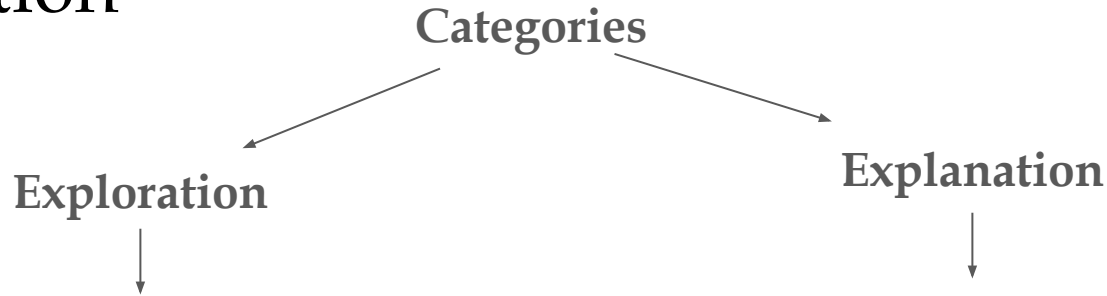
Data visualizations are initially designed by a human, but are then drawn algorithmically with graphing, charting, or diagramming software.

The advantage of this approach is that it is relatively simple to update or regenerate the visualization with more or new data.

While they may show great volumes of data, information visualizations are often less aesthetically rich than infographics.



# Categories of Data visualization: Exploration versus Explanation



- When you have a whole bunch of data and you're not sure what's in it.
- To know what's inside your data set, translating it into a visual medium can help you quickly identify its features, including interesting curves, lines, trends, or anomalous outliers.
- Exploration is generally best done at a high level of granularity
- There may be a whole lot of noise in your data, but if you oversimplify or strip out too much information, you could end up missing something important.
- This type of visualization is typically part of the data analysis phase, and is used to find the story the data has to tell you

- when you already know what the data has to say and you are trying to tell that story to somebody else
- The story you are trying to tell is known to you at the outset, and therefore you can design to specifically accommodate and highlight that story.
- Can make certain editorial decisions about which information stays in, and which is distracting or irrelevant and should come out
- out. This is a process of selecting focused data that will support the story you are trying to tell
- Exploratory data visualization is part of the data analysis phase, and presentation phase

# Hybrids: Exploratory Explanation

Involves a curated dataset that is nonetheless presented with the intention to allow some exploration on the reader's part.

These visualizations are usually interactive via some kind of graphical interface that lets the reader choose and constrain certain parameters, thereby discovering for herself whatever insights the dataset may have to offer.

So in these hybrid designs there is a certain freedom-of-discovery aspect to the information presented, but it is usually not totally raw; it has been distilled and facilitated to some extent.

# Exploratory Visualization: Informative versus Persuasive versus Visual Art

There are three main categories of explanatory visualizations based on the relationships between the three necessary players:

- The designer,
- The reader, and
- The data.

It discusses designing visualizations of data with known parameters and stories

# The Designer-Reader-Data Trinity

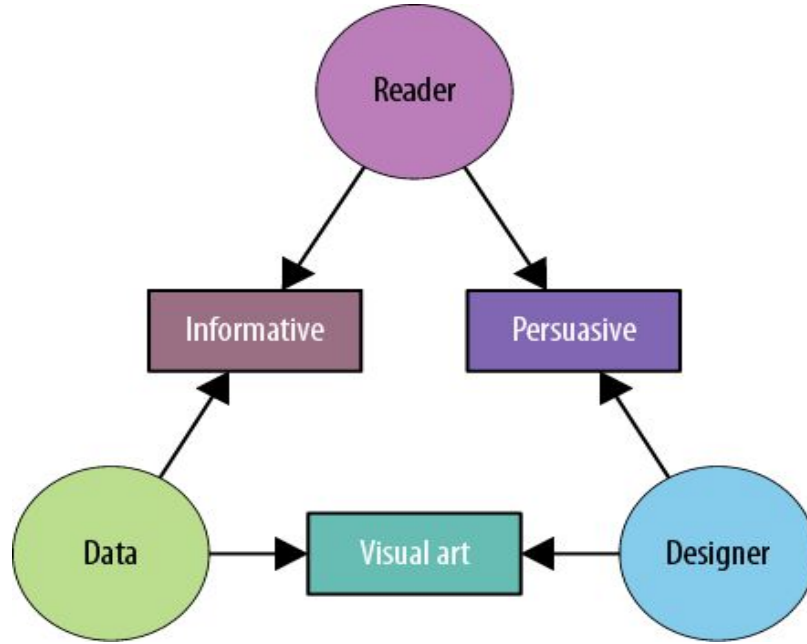
An effective explanatory data visualization as supported by a three-legged stool consisting of the designer, the reader, and the data.

Each of these “legs” exerts a force, or contributes a separate perspective, that must be taken into consideration for a visualization to be stable and successful.

Each of the three legs of the stool has a unique relationship to the other two.

It is necessary to account for the needs and perspective of all three in each visualization project, the dominant relationship will ultimately determine which category of visualization is needed

# The Designer-Reader-Data Trinity



The nature of the visualization depends on which relationship (between two of the three components) is dominant

# Informative

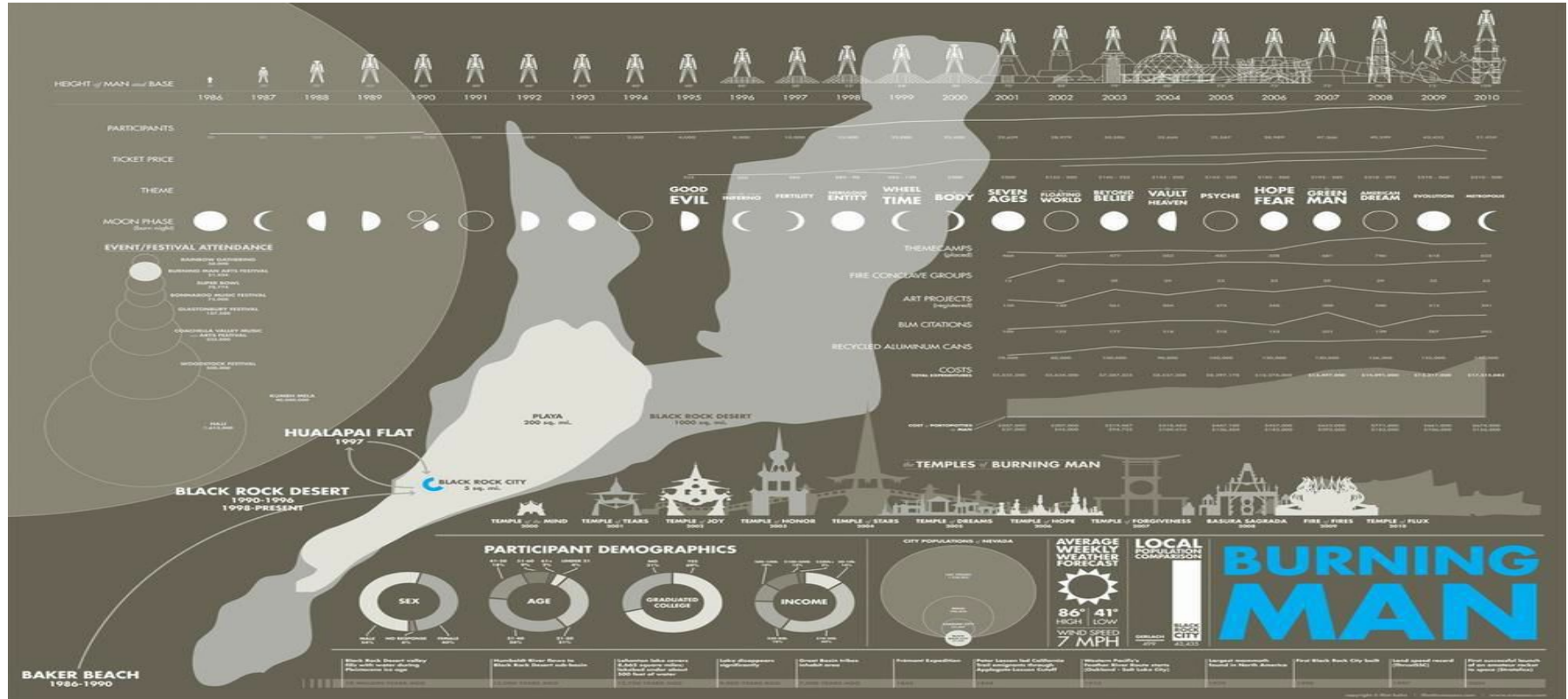
An informative visualization primarily serves the relationship between the reader and the data.

It aims for a neutral presentation of the facts in such a way that will educate the reader

Informative visualizations are often associated with broad data sets, and seek to distill the content into a manageably consumable form.

They form the bulk of visualizations that the average person encounters on a day-to-day basis — whether that's at work, in the newspaper, or on a service-provider's website

# Flint Hahn's Burning Man infographic:



# Persuasive

A persuasive visualization primarily serves the relationship between the designer and the reader.

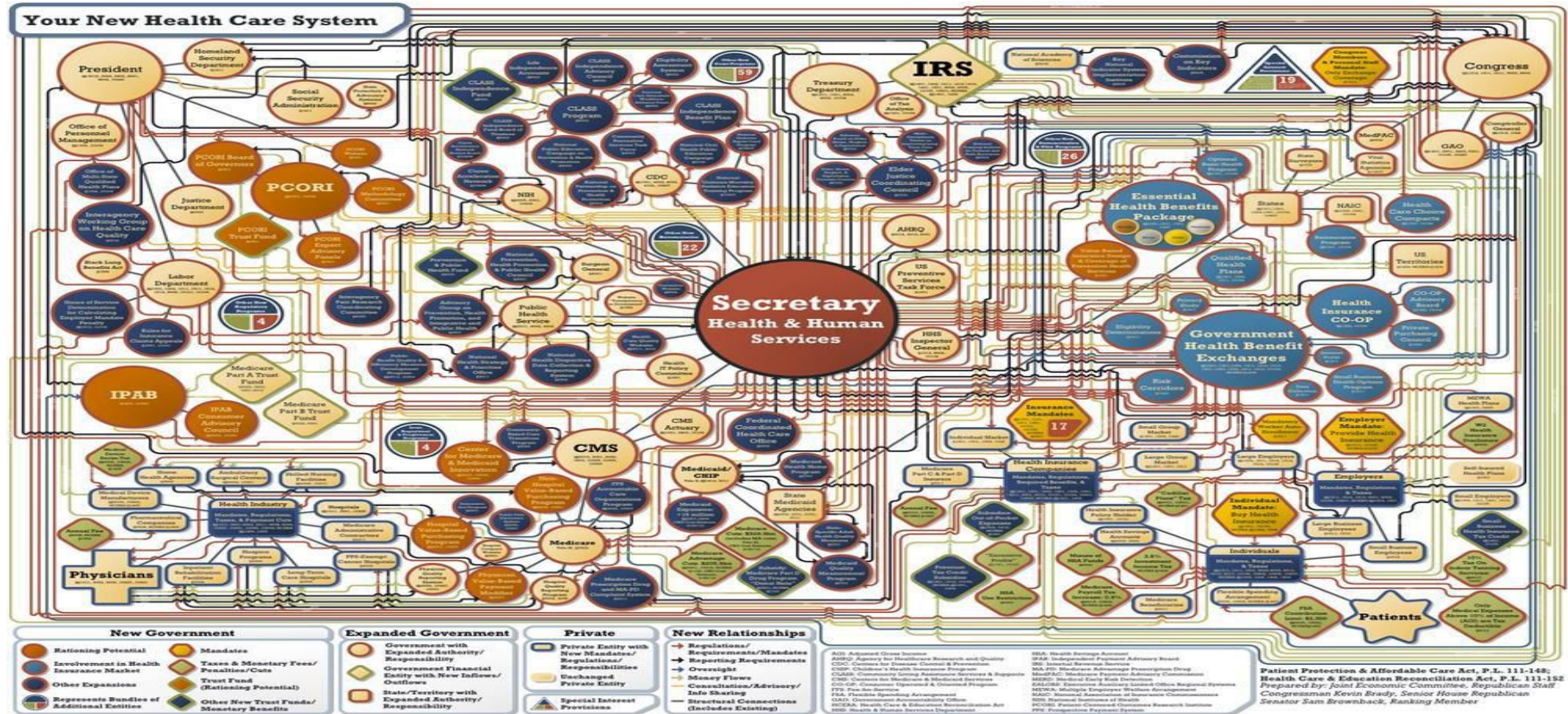
It is useful when the designer wishes to change the reader's mind about something.

It represents a very specific point of view, and advocates a change of opinion or action on the part of the reader.

In this category of visualization, the data represented is specifically chosen for the purpose of supporting the designer's point of view, and is presented carefully so as to convince the reader of same.



# Joint Economic Committee minority's rendition of the proposed Democratic health care plan in 2010



# Visual Art

Visual art, primarily serves the relationship between the designer and the data.

It often entails unidirectional encoding of information, meaning that the reader may not be able to decode the visual presentation to understand the underlying information

The designer may intend only to condense it, translate it into a new medium, or make it beautiful; she may not intend for the reader to be able to extract anything from it other than enjoyment.

This category of visualization is sometimes more easily recognized than others.

Participants address the Fiber Optic Tapestry by tweeting #optictapestry and a primary color—the tapestry displays the colors in algorithmically-determined patterns



# Informative vs Persuasive vs Visual Art

While an informative visualization may not have an intentional point of view in the manner that a persuasive visualization does, all visualizations are going to be biased to some degree, based on the fact that designers are human and have to make choices.

Both informative and persuasive visualizations are meant to be easily decodable—bidirectional in their encoding—whereas visual art merely translates the data into a visual form.

# Ingredients of Successful Visualizations

The three major sources of influence on the design of your data visualization (explanatory visualization).

While exploratory visualization is more about you finding out what's in your data, explanatory visualization is about you as a designer telling the story of the data to your reader.

These three components are your holy trinity when designing data visualizations:

- Designer
- Reader
- Data

# Designer: Why are you here?

As a designer, you have a goal. You may not be aware of it, but you are creating a visualization for some reason.

Being aware of your motivations, goals, and priorities will help you design a successful visualization, rather than merely create an arbitrary visual representation of your data. It is foundation of your process.

Steps in design process:

1. There are different types of visualizations. Knowing which type of visualization you're working with. Example: If you have a story to tell, your visualization is almost certainly informative, persuasive, or visual art.
2. What kind of experience that visualization type should provide to your reader
3. What information should they be able to learn from this visualization?
4. What point or message are you trying to convey?

# Designer

Keep your goal in mind.

To be clear, you should be open to iteration and evolution, serendipity, and the paths that new insight may reveal.

But never lose sight of why you're here in the first place.

Your unique perspective is the value you bring to the table, and it should inform your design choices.



# You Are Creating This for Other People - Reader

Reader - second source of influencer

Reader holds a very special place in the trinity and can be your biggest ally or your biggest hurdle in clear communication—sometimes both.

At all stages of creating your visualization, it is important to put yourself in the shoes of your reader, and to take into consideration the unique viewpoint that he will bring with him

Your success is measured by your reader's success

Explanatory data visualization is a communication medium.

You are selectively encoding specific information in such a way that the reader will be able to decode it and successfully receive your message



# You Are Creating This for Other People - Reader

To be successful, you need to consider the various “distortions” or filters your readers will introduce

To put yourself in the reader’s shoes will force you to simplify your explanations a bit.

It’s merely a process of breaking down the ideas until you can communicate them in clear and transparent terms.

The purpose of the visualization is to take a story that is already known to you and tell it to somebody else, then it stands to reason that the somebody else is exactly **that — not you, but an other**.

# Reader's Context

Considerations to make when designing a data visualization:

- Questions of identity, motivation, and language (specialized knowledge and vocabulary)
- Learned social context
  - What do colors mean?
  - Which direction it the reader used to reading in?
  - Which icons is she familiar with?

# Reader's Context of Use

Different time-frames the reader may be constrained to,

- The factors motivating him to understand your data, and
- The information he needs to gather to meet his own goals or make good decisions

The key questions to ask here are ones like:

- What information does my reader need to be successful?
- How much detail does she need?
- How long does she have to make it effective?

Once you understand the context in which the reader is operating, you can discern which information he needs;

Once you understand the filters he may be using, you can discern how best to present that information to him.

# Data: Third source of Influencer

The best visualizations will reveal what is interesting about the specific data set.

Different data may require different approaches, encodings, or techniques to reveal its interesting aspects.

To start with use default visualization format.

However, the data will yield new knowledge when a different visualization approach or format is used

How do you choose a visualization format that shows your data's best side?

Know your data. Respect your data. Consider the inherent values, relationships, and structures of your data.

# Data: Third source of Influencer

The type of basic questions you will want to ask about your data include:

- Is it a time-series? A hierarchy?
- How many dimensions does it have? Which are the most important ones?
- What sort of relationships do they have (e.g., one-to-one or many-to-many)?
- How variable are they?
- Are the values categorical? Discrete or continuous? Linear or non-linear? How are they bounded?
- How many categories are there?

Each relevant relationship and property of your data needs to be encoded with an appropriate visual property;

The characteristics of each dimension of your data will inform which visual property you choose to use to encode it.

# Choose Appropriate Visual Encodings

Once Data shape is known, we can encode its various dimensions with appropriate visual properties.

For encoding different types of data - visual properties can be varied or may be modified in different ways.

Two key factors are

1. Whether a visual property is naturally ordered, and
2. How many distinct values of this property the reader can easily differentiate.

Natural ordering and Number of distinct values will indicate whether a visual property is best suited to one of the main data types: quantitative, ordinal, categorical, or relational data.

# Natural Ordering - First factor for choosing a visual property

Whether a visual property has a natural ordering?

Human brains detect it automatically, it evaluates relative order independent of language, culture, convention, or other learned factors.

For example,

position has a natural ordering; shape doesn't.

Length has a natural ordering; texture doesn't (but pattern density does).

Line thickness or weight has a natural ordering; line style (solid, dotted, dashed) doesn't.

Depending on the specifics of the visual property, its natural ordering may be well suited to representing quantitative differences (27, 33, 41), or ordinal differences (small, medium, large, enormous).

# Color is not ordered - A tricky one

Color (hue) is not naturally ordered in our brain.

Brightness (lightness or luminance, sometimes called tint) and intensity (saturation) are, but color itself is not.

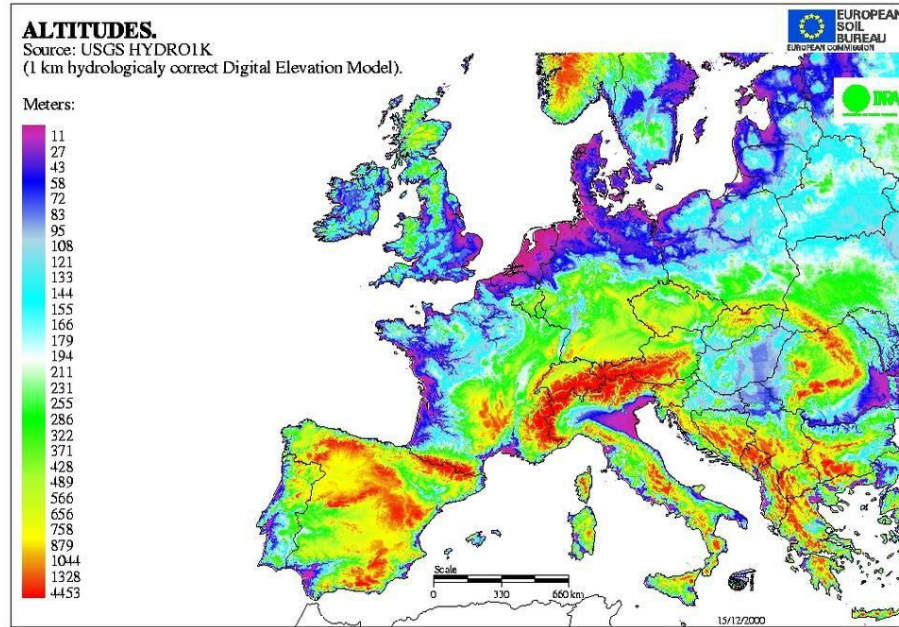
We have strong social conventions about color, and there is an ordering by wavelength in the physical world

In contexts where you're tempted to use "ordered color" (elevation, heat maps, etc.), consider varying brightness along one, or perhaps two, axes.

For example, elevation can be represented by increasing the darkness of browns, rather than cycling through the rainbow



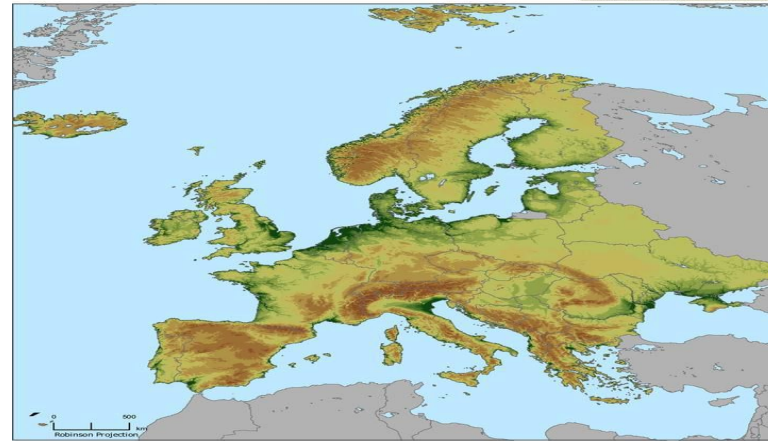
# A rainbow encoding



Does red mean the Alps are hotter than the rest of Europe? Its difficult to understand

# Ordering based on intensity/wavelength

Elevation Zones  
Europe



Digital elevation data (meters above mean sea level) were obtained as a 1 kilometer resolution elevation/bathymetry raster product from ISciences, L.L.C. Elevation zones were created by aggregating ranges of land elevation values into 12 thematic elevation classes. The 2004 ISciences data were resampled from their native 30 arc-second resolution to match GPR's population and land area 2.5 arc-minute spatial footprint. Source: ISciences, L.L.C. 300 N. Fifth Ave., Suite 120., Ann Arbor, MI 48104 <http://www.isciences.com/>.



Copyright 2007, The Trustees of Columbia University in the City of New York, Source: Center for International Earth Science Information Network (CIESIN), Columbia University. Population, Landscape, and Climate Estimates (PLACE).

Further information available at: <http://medias.ciesin.columbia.edu/medias>

Elevation

< 5 m	200 - 399.9 m
5 - 9.9 m	400 - 799.9 m
10 - 24.9 m	800 - 1499.9 m
25 - 49.9 m	1500 - 2999.9 m
50 - 99.9 m	3000 - 4999.9 m
100 - 199.9 m	>= 5000 m



This document is licensed under a Creative Commons 3.0 Attribution License <http://creativecommons.org/licenses/by/3.0/>

Publish Date: 03/13/07

The colors diverge from one point, clearly indicating low, medium, and high elevations

# Distinct Value - Second factor for choosing a visual property

How many distinct values it has that your reader will be able to perceive, differentiate, and possibly remember.

For example, there are a lot of colors in the world, but we can't tell them apart if they're too similar.

We can more easily differentiate

- A large number of shapes,
- A huge number of positions, and
- An infinite number of numbers.

When choosing a visual property, select one that has a number of useful differentiable values and an ordering similar to that of your data

Common visual properties to help you select an appropriate encoding for your data type.

Example	Encoding	Ordered	Useful values	Quantitative	Ordinal	Categorical	Relational
	position, placement	yes	infinite	Good	Good	Good	Good
1, 2, 3; A, B, C	text labels	optional alpha or num	infinite	Good	Good	Good	Good
	length	yes	many	Good	Good		
	size, area	yes	many	Good	Good		
	angle	yes	medium	Good	Good		
	pattern density	yes	few	Good	Good		
	weight, boldness	yes	few		Good		
	saturation, brightness	yes	few		Good		
	color	no	few (<20)			Good	
	shape, icon	no	medium			Good	
	pattern texture	no	medium			Good	
	enclosure, connection	no	infinite			Good	Good
	line pattern	no	few				Good
	line endings	no	few				Good
	line weight	yes	few		Good		

# Redundant Encoding

Use of Multiple visual element (use of more than one graphical or visual structure).

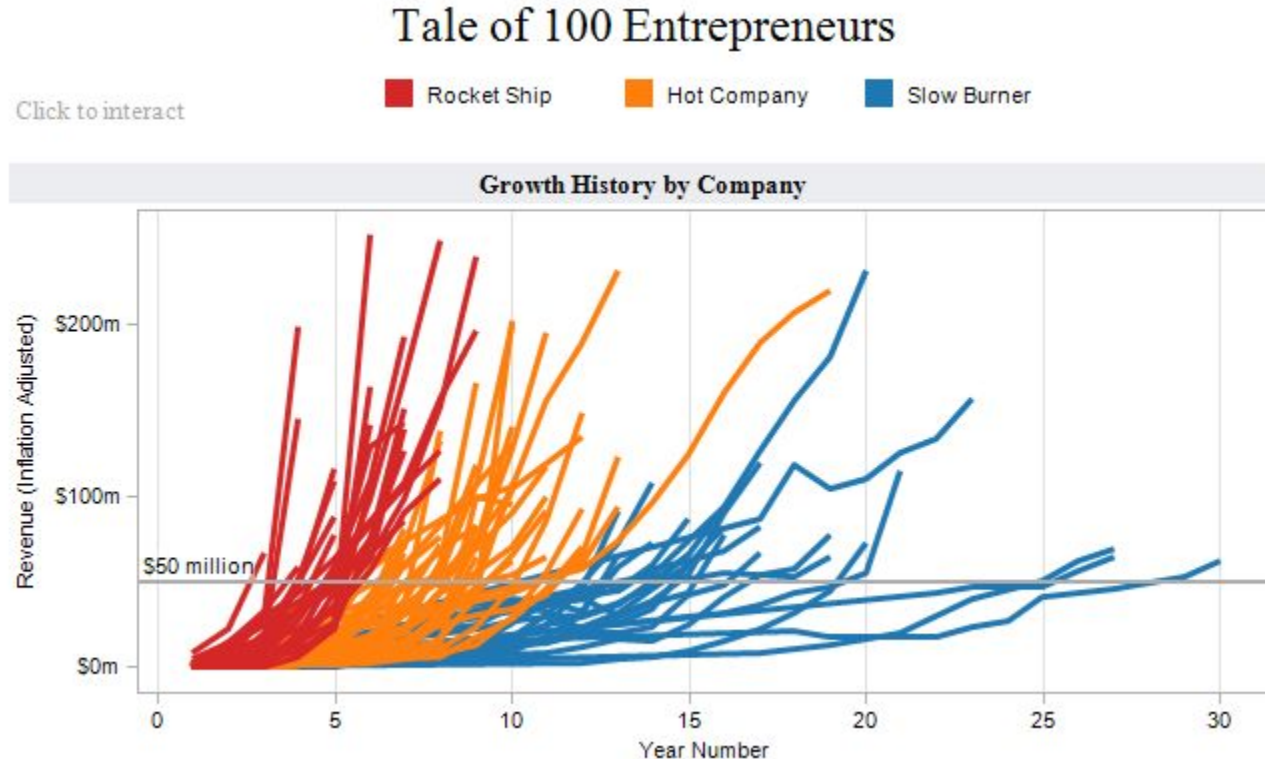
The advantage of redundant encoding is that using more channels to get the same information into your brain can make acquisition of that information faster, easier, and more accurate

For example,

If you've got lines differentiated by ending (arrows, dots, etc.), consider also changing the line style (dotted, dashed, etc.) or color.

If you've got values encoded by placement, consider redundantly encoding the value with brightness, or grouping regions with color

Color redundantly encodes the position of groups of companies in this graph



# Defaults versus Innovative Formats

There are a lot of good default encodings and encoding conventions in the world.

Designing new encoding formats can cost you a lot of time and effort, and may make your reader expend a lot of time and effort to learn.

The choice comes down to a basic cost-benefit analysis.

What is the expense to you and your reader of creating and understanding a new encoding format, versus the value delivered by that format?

1. If you've got a truly superior solution (as evaluated by your reader), then by all means, use it.
2. But if your job can be done (or done well enough) with a default format, save everyone the effort and use a standard solution

# Reader's Context

There are more than one reader, Each different from each other as they are from you.

It may be impossible to take the preconceptions of all readers into consideration at once.

So choose the most important group, think of them as your core group, and design with them in mind.

But, going forward, when we say reader, what we really mean is a representative reader from within your core audience.



# Reader's Context - some facets of the reader's mindset that need to take into account

Titles, tags, and labels

- consider your reader's vocabulary and familiarity with relevant jargon.
- Is the reader from within your industry or outside of it? What about other readers outside of the core audience group?
- Is it worth using an industry term for the sake of precision (knowing that the reader may have to look it up), or would a lay term work just as well?
- Will the reader be able to decipher any unknown terms from context, or will a vocabulary gap obscure the meaning of all or part of the information presented?

# Reader's Context - some facets of the reader's mindset that need to take into account

## Colors

- Color choice depends on readers perception and cultural associations.
- Depending on the culture in question, some colors may be lucky, some unlucky;
- some may carry positive or negative connotations;
- some may be associated with life events like weddings, funerals, or newborn children.

# Reader's Context - some facets of the reader's mindset that need to take into account

## Colors

- Colors take on meaning when paired or grouped with other colors
  - In the United States, red and royal blue to Republicans and Democrats;
  - pink and light blue often refer to boys and girls;
  - red, yellow, and green to traffic signals.
  - The colors red, white, and green may signal Christmas in Canada, but patriotism in Italy.
  - The colors red, white, and blue are patriotic in multiple places: they will make both an American and a Frenchman think of home.

# Reader's Context - some facets of the reader's mindset that need to take into account

## Color blindness

- Color blindness is more properly referred to as color vision deficiency or dyschromatopsia.
- Red-green deficiency is the most common by far, but yellow-blue deficiency also occurs.
- And there are lots of people who have trouble distinguishing between close colors like blue and purple.
- Select color swatches into a group (or enter custom RGB values) and simulate how they are perceived with eight types of dyschromatopsia

# Reader's Context - some facets of the reader's mindset that need to take into account

## Directional Orientation

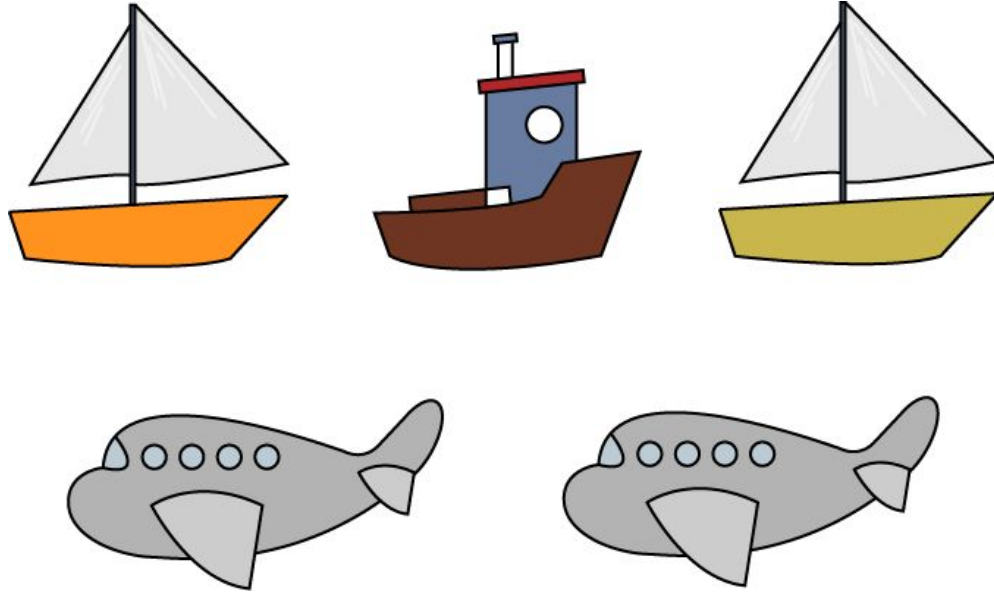
- There exist a culture that readers may reads left-to-right, right-to-left, or top-to-bottom?
- A person's habitual reading patterns will determine their default eye movements over a page, and the order in which they will encounter the various visual elements in your design.

# Reader's Context - some facets of the reader's mindset that need to take into account

## Compatibility with reality

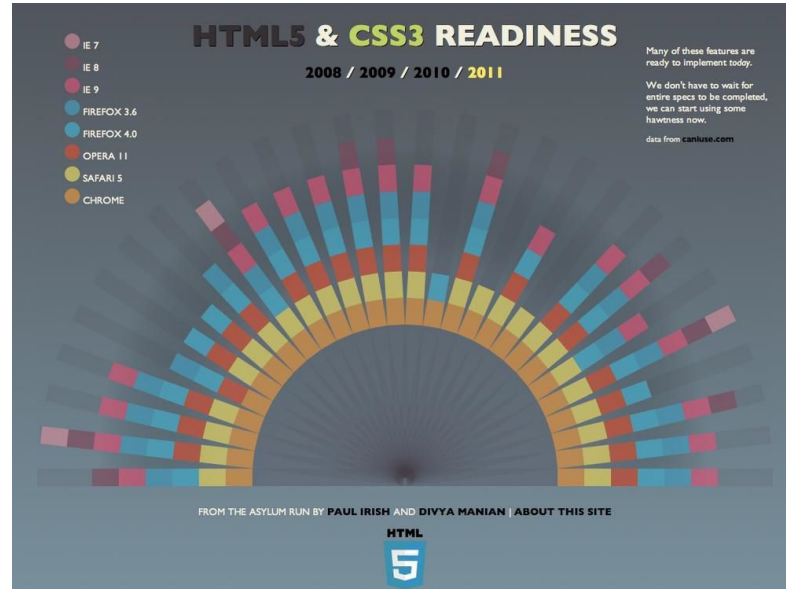
- One way to make decoding easy is to make your encodings of things and relationships as well aligned with the reality (or your reader's reality) of those things and relationships as possible; this alignment is called compatibility.
- This can have many different aspects, including taking cues from the physical world and from cultural conventions.
- Things in the world are full of inherent properties. Use these properties in encoding scheme

# Properties



The visual placement of boats above airplanes is jarring, since they don't appear that way in the physical world.
















# Properties



Representation of browser capabilities



# Properties

Browser	Existing color	Better color
		
		
		
		
		

The representative colors differ greatly from the colors in the browser icons. Other choices would better reflect the icons' colors.

# Reader's Context - some facets of the reader's mindset that need to take into account

## Direction and Reality

- Direction is an interesting property to consider because it has both inherent and learned conventions.

# Patterns and Consistency

- The human brain is amazingly good at identifying patterns in the world.
- We easily recognize similarity in shapes, position, sound, color, rhythm, language, behavior, and physical routine, just to name a few variables.
- This ability to recognize patterns is extremely powerful, as it enables us to identify stimuli that we've encountered before, and predict behavior based on what happened the last time we encountered a similar stimulus pattern.
- Patterns are foundation of language, communication, and all learning.

# Patterns and Consistency

- we also should notice violations of patterns (exceptions).
- Pattern and pattern-violation recognition has two major implications for design.
- The first is that readers will notice patterns and assume they are intentional, whether you planned for the patterns to exist or not.
- The second is that when they perceive patterns, readers will also expect pattern violations to be meaningful.
- As designers, we must be extremely deliberate about the patterns and pattern violations we create.

# Patterns and Consistency

- Don't arbitrarily assign positions or colors or connections to your choices.
- If you change the order or membership of a list of items, either in text or in placement, it will be perceived as meaningful.
- So how should you avoid the potential trap of implying meaning where none is intended? It all comes down to three simple rules.
  - Be consistent in membership, ordering, and other encodings.
  - Things that are the same should look the same.
  - Things that are different should look different.
- Maintaining consistency and intention when encoding will greatly enhance the accessibility and efficiency of your visualization,

# Selecting Structure

The structure of a visualization should reveal something about the underlying data.

One of the most classic data visualization example: the Periodic Table of the Elements.

It takes a complex dataset and makes it simple, organized, and transparent.

The elements are laid out in order by atomic number, and by wrapping the rows at strategic points, the table reveals that elements in various categories occur at regular intervals, or periods.

The table makes it easier to understand the nature of each element—both individually, and in relation to the other elements

# Selecting Structure

# Periodic Table of Elements

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18		
1	<b>H</b> Hydrogen 1.00794	Atomic # Symbol Name Atomic Mass																<b>2</b> <b>He</b> Helium 4.002602		
2	<b>Li</b> Lithium 6.941	<b>4</b> <b>Be</b> Beryllium 9.012182	<div>C Solid</div> <div>Hg Liquid</div> <div>H Gas</div> <div>Rf Unknown</div>		Metals					Transition metals		Nonmetals								
3	<b>11</b> <b>Na</b> Sodium 22.98976928	<b>12</b> <b>Mg</b> Magnesium 24.3050			Alkaline earth metals					Poor metals		Other nonmetals		Noble gases						
4	<b>19</b> <b>K</b> Potassium 39.0983	<b>20</b> <b>Ca</b> Calcium 40.078	<b>21</b> <b>Sc</b> Scandium 44.955912	<b>22</b> <b>Ti</b> Titanium 47.867	<b>23</b> <b>V</b> Vanadium 50.9415	<b>24</b> <b>Cr</b> Chromium 51.9961	<b>25</b> <b>Mn</b> Manganese 54.938045	<b>26</b> <b>Fe</b> Iron 55.845	<b>27</b> <b>Co</b> Cobalt 58.933195	<b>28</b> <b>Ni</b> Nickel 58.6934	<b>29</b> <b>Cu</b> Copper 63.546	<b>30</b> <b>Zn</b> Zinc 65.38	<b>31</b> <b>Ga</b> Gallium 69.723	<b>32</b> <b>Ge</b> Germanium 72.64	<b>33</b> <b>As</b> Arsenic 74.9216	<b>34</b> <b>Se</b> Selenium 78.96	<b>35</b> <b>Br</b> Bromine 79.904	<b>36</b> <b>Kr</b> Krypton 83.56		
5	<b>37</b> <b>Rb</b> Rubidium 85.4678	<b>38</b> <b>Sr</b> Strontium 87.62	<b>39</b> <b>Y</b> Yttrium 88.90585	<b>40</b> <b>Zr</b> Zirconium 91.224	<b>41</b> <b>Nb</b> Niobium 92.90638	<b>42</b> <b>Mo</b> Molybdenum 95.94	<b>43</b> <b>Tc</b> Technetium 97.07	<b>44</b> <b>Ru</b> Ruthenium 101.07	<b>45</b> <b>Rh</b> Rhodium 102.90550	<b>46</b> <b>Pd</b> Palladium 106.42	<b>47</b> <b>Ag</b> Silver 107.8682	<b>48</b> <b>Cd</b> Cadmium 112.4118	<b>49</b> <b>In</b> Indium 114.818	<b>50</b> <b>Sn</b> Tin 118.710	<b>51</b> <b>Sb</b> Antimony 121.757	<b>52</b> <b>Te</b> Tellurium 127.6	<b>53</b> <b>I</b> Iodine 126.90447	<b>54</b> <b>Xe</b> Xenon 131.29		
6	<b>55</b> <b>Cs</b> Cesium 132.9054519	<b>56</b> <b>Ba</b> Barium 137.327	57–71		<b>72</b> <b>Hf</b> Hafnium 178.49	<b>73</b> <b>Ta</b> Tantalum 180.94788	<b>74</b> <b>W</b> Tungsten 183.84	<b>75</b> <b>Re</b> Rhenium 186.207	<b>76</b> <b>Os</b> Osmium 190.23	<b>77</b> <b>Ir</b> Iridium 192.222	<b>78</b> <b>Pt</b> Platinum 195.084	<b>79</b> <b>Au</b> Gold 196.966569	<b>80</b> <b>Hg</b> Mercury 200.59	<b>81</b> <b>Tl</b> Thallium 204.3833	<b>82</b> <b>Pb</b> Lead 207.2	<b>83</b> <b>Bi</b> Bismuth 208.9804	<b>84</b> <b>Po</b> Polonium 209	<b>85</b> <b>At</b> Astatine 222.0176	<b>86</b> <b>Rn</b> Radon 222.0176	
7	<b>87</b> <b>Fr</b> Francium 223.019782	<b>88</b> <b>Ra</b> Radium 226	89–103		<b>104</b> <b>Rf</b> Rutherfordium 261	<b>105</b> <b>Db</b> Dubnium 262	<b>106</b> <b>Sg</b> Seaborgium 266	<b>107</b> <b>Bh</b> Bohrium 264	<b>108</b> <b>Hs</b> Hassium 277	<b>109</b> <b>Mt</b> Meitnerium 268	<b>110</b> <b>Ds</b> Darmstadtium 271	<b>111</b> <b>Rg</b> Roentgenium 272	<b>112</b> <b>Uub</b> Ununbium 285	<b>113</b> <b>Uut</b> Ununtrium 284	<b>114</b> <b>Uuq</b> Ununquadium 289	<b>115</b> <b>Uup</b> Ununpentium 288	<b>116</b> <b>Uuh</b> Ununhexium 292	<b>117</b> <b>Uus</b> Ununseptium 294	<b>118</b> <b>Uuo</b> Ununoctium 294	

For elements with no stable isotopes, the mass number of the isotope with the longest half-life is in parentheses.

Periodic Table Design and Interface Copyright © 1997 Michael Dayah. <http://www.ptable.com/> Last updated: May 27, 2008

**Ptable**  
.com

**Michael Dayah**

57 <b>La</b> Lanthanum 138.90547	58 <b>Ce</b> Cerium 140.12	59 <b>Pr</b> Praseodymium 140.90768	60 <b>Nd</b> Neodymium 144.242	61 <b>Pm</b> Promethium 144.9126	62 <b>Sm</b> Samarium 150.36	63 <b>Eu</b> Europium 151.964	64 <b>Gd</b> Gadolinium 157.25	65 <b>Tb</b> Terbium 158.92535	66 <b>Dy</b> Dysprosium 162.50014	67 <b>Ho</b> Holmium 164.93032	68 <b>Er</b> Erbium 167.259	69 <b>Tm</b> Thulium 168.93271	70 <b>Yb</b> Ytterbium 173.05468	71 <b>Lu</b> Lutetium 174.967
89 <b>Ac</b> Actinium (227)	90 <b>Th</b> Thorium 232.0377	91 <b>Pa</b> Protactinium 231.03688	92 <b>U</b> Uranium 238.02891	93 <b>Np</b> Neptunium 237.04817	94 <b>Pu</b> Plutonium 244.0642	95 <b>Am</b> Americium (243)	96 <b>Cm</b> Curium 247.0704	97 <b>Bk</b> Berkelium 247.0703	98 <b>Cf</b> Californium (251)	99 <b>Es</b> Einsteinium 252.083	100 <b>Fm</b> Fermium (257)	101 <b>Md</b> Mendelevium (258)	102 <b>No</b> Nobelium (259)	103 <b>Lr</b> Lawrencium (260)

For a fully interactive experience, visit [www.ptable.com](http://www.ptable.com).

michael@dayah.com

This rendition of the classic table makes good use of color and line

# Selecting Structure

Designers and satirists are constantly repurposing its familiar rows and columns to showcase collections of everything from typefaces to video game controllers, and, ironically, visualization methods.

This phenomenon is a particular peeve to your authors precisely because it violates the important principle of selecting an appropriate structure



# Position: Layout and Axes

Axes (and the resulting layout), placement, and position of entities are perhaps the most underutilized visual encodings.

Position:

- Physical position is the easiest to perceive and most powerful visual property but this power is only accessible if you chose to use it.
- It's not difficult to use, and—when used properly—it can convey a huge amount of information.
- Position can encode as many discrete values as you have room to display (sometimes more), and can indicate ordered, temporal, categorical, quantitative, or quantitative values, as well as causal, correlative, sequential, qualitative, or quantitative relationships.
- “These values are different than those values” can be encoded with position.

# Position: Layout and Axes

## Axes:

- Axes contribute value in two ways: both from the bottom up and from the top down.
- From the bottom up, an entity inherits values from an axis or axes without requiring the use of any extra ink or labeling on that entity.
- The reader can easily look at where the entity is placed relative to an axis or axes and understand that the values that correspond to that position on the axis apply to the entity in question.
- The entity doesn't need to be labeled "first" or "Western" or "less valuable"; those values are (appropriately) implied by position alone when an axis is defined.
- If the reader is looking for entities with a specific value or range of values, well-defined axes can help limit the scope of their search to the groups or subsets of entities that have that value. This is the top-down approach.

# The Meaning of Placement and Proximity

## Semantic Distance and Relative Proximity:

Semantic distance: The conceptual relatedness of objects or ideas.

### Example:

A housecat, for instance, is semantically closer to a tiger than to a water buffalo. But it may be closer to a water buffalo than to an orchid, or to Neuschwanstein Castle. Thus we can see that with semantic difference, not only does proximity matter, but relative proximity matters.

# Semantic Distance and Relative Proximity:



Placing the housecat closer to the tiger shows that it is semantically closer to that than to the other items, and forms the category felines

# Semantic Distance and Relative Proximity:



Putting the water buffalo closer to the tiger causes your brain to think of a new category: perhaps, safari animals

# The Meaning of Placement and Proximity

## **Absolute Placement:**

- Absolute placement on the page or screen also carries meaning.
- All else being equal, elements toward the top of the page are generally considered more important than elements on the bottom.
- Most people also read from the top down, so items at the top will be read first

## **Representation of Physical Space:**

- When no other compelling and meaningful ordering exists, consider placing items on the page (or screen) in the same relative order as they'd be found in the physical world.

# The Meaning of Placement and Proximity

## Logical Relationships versus Physical Relationships:

- Though representation and encoding should match reality as closely as possible, there are situations for which this isn't true.
- Sometimes the relationships that need to be communicated are more important than the physical (or understood) relative positions.
- When selecting your layout, consider which of these is the more important consideration.

# The Meaning of Placement and Proximity

## Patterns and Grouped Objects:

- One of the cool things our brains are really good at is automatically collecting and grouping disparate elements into a cohesive set.
- Example: Human don't look at zebra and perceive, it has 43 vertical lines of one arm length.
- Our brains automatically group and assimilate similar features in proximity to each other.
- This is another way to say that our brains are great at picking up on patterns
- It's important to be aware of this function of the visual system, because it may come into play unexpectedly, causing your reader's brain to group or otherwise relate proximate elements within your visualization into a pattern you didn't intend.



# Patterns of Organization

In *Information Anxiety* (Doubleday), Richard Saul Wurman proposed that there are just five ways to organize information, and suggested the mnemonic acronym LATCH:

- Location
- Alphabetical
- Time
- Categorical
- Hierarchical
- Organising Data:

In most cases, you should place the most important (or first) piece of information at the left end or top, and the least important, or last, at the right end or bottom.

# Patterns of Organization

## Importance:

- A hierarchy variant where the most relevant information or the most important entities are at one end (usually the top or left) and the least important are at the opposite end.
- This works for everything from organizational charts to sales figures.
- Importance can also apply to your own priority and the direction you want the reader to consume the information

# Patterns of Organization

## Causality:

- Cause and effect, whether linear or cyclical, is a powerful organizational structure.
- If the organization is cyclical, consider a round structure, probably moving clockwise.
- Otherwise, causality probably flows from cause on the left to effect on the right.

# Patterns of Organization

## Dependence:

- This is a great organizational scheme in which entities can be ranked as relatively more dependent and independent.
- Consider this for hierarchically related entities, such as object classes, libraries, etc.

## Categorical:

- Not hierarchical, but absolutely useful. Consider clumping or grouping otherwise-unordered data into categories.
- If nothing else, it'll help your reader make sense of the volume of data you're presenting.

# Specific Graphs, Layouts, and Axis Styles:

## Quantitative and comparative formats:

Bar graphs:

- Most common and most useful graph types
- Go-to graph for comparing data values within or across categories.
- Bar graphs are good for discrete data; for continuous data, consider using a line graph.
- Bars can contain multiple, stacked data values within the category.
- Stacking values can make comparing the upper values difficult, because they don't share a common baseline with the same data dimension in other bars.
- For this reason, carefully consider which values you put at the bottom of each bar.

# Specific Graphs, Layouts, and Axis Styles:

## Quantitative and comparative formats:

### Histograms:

- Specialized bar graphs designed to show distribution of values across a possible range.
- The total area of the graph represents the sum total of all of the values present.
- The granularity of the categories on the horizontal axis can be as coarse or narrow as is useful.

# Specific Graphs, Layouts, and Axis Styles:

## Quantitative and comparative formats:

### Line graphs:

- Workhorse for continuous data
- Great for showing trends, and—for the right kind of data—can be far less cluttered than bar graphs
- Line graphs have the independent data dimension progressing along the horizontal axis, and the dependent data dimension along the vertical axis. However, there are cases where the data isn't related in such a way.
- When designing filled-area line graphs, be very clear to indicate whether the graph values accumulate vertically, or whether they are layered in front of one another

# Specific Graphs, Layouts, and Axis Styles:

## Quantitative and comparative formats:

### Time series:

- For Data with a time dimension
- Time values typically progressing left to right on the horizontal axis, and another data dimension displayed on the vertical axis

### Pie Graphs:

- Valid for comparing fractions of a whole.
- Best used when there are few relevant fractions, and precision isn't required.



# Specific Graphs, Layouts, and Axis Styles:

## Quantitative and comparative formats:

### Scatter Plots:

- Great for looking for correlations between two quantitative dimensions of data, or for displaying data that varies along two dimensions.
- Three and four dimensions are possible by encoding data points as bubbles, pies, or stacked bar graphs.

# Specific Graphs, Layouts, and Axis Styles:

## Quantitative and comparative formats:

### Tables:

- Tables can be an effective compliment for visualization styles where trends are visible, but precision is harder to perceive.
- Note that table cells can contain data other than text, and the cell itself can be encoded with color or other visual properties, as in the Periodic Table of the Elements.

### Periodic Tables:

- Best choice for periodic data.

# Specific Graphs, Layouts, and Axis Styles:

## Quantitative and comparative formats:

### Treemaps:

- A solution to the problem of representing proportional values and hierarchal relationships at the same time.
- They are excellent for representing many hierarchically nested data values.
- Treemaps are typically used to represent distribution and use of resources, such as budgets or computer storage.

# Specific Graphs, Layouts, and Axis Styles:

## Quantitative and comparative formats:

Heat maps:

- Heat maps are two-dimensional area graphs that use color or brightness to indicate values (or changes in value) of large data sets.
- Color can be used to indicate areas of large changes, out-of-range values, or other interesting characteristics.
- While red-green or red-yellow-green color schemes are common, best practices include muting unremarkable values so that interesting values have a greater contrast against normal values

# Specific Graphs, Layouts, and Axis Styles:

## Quantitative and comparative formats:

Small multiples:

- Are arrays of very small, not very detailed graphs that allow the reader to compare the sense or trend of many values concurrently, or to show how a set of values changes over time.

Marimekko (also known as matrix or mosaic) graphs:

- Marimekko graphs are bar graphs with bars of uniform height that are stacked in two directions to form a larger rectangle.
- The width of each column varies, so that its surface area represents that column's fraction of the whole.
- The sum of each column's stacked sections represents 100% of the value of that column.

# Specific Graphs, Layouts, and Axis Styles: Relational formats:

Data flow diagrams, Entity Relationship Diagrams, etc.:

- Used to document the flow of data through a process, the flow of a user through a software interface or website, the relationship among classes or database tables, etc.
- These all can benefit from an organized layout, sorted by level of abstraction, chronology, importance, hierarchy, or other relevant classification.

Decision maps and flow charts:

- Track the path of a process or decision.
- Nodes represent choices, actions, or other tests or states.
- Node type is usually encoded with shape, and often with color as well.
- Example: UML behavior diagrams

# Specific Graphs, Layouts, and Axis Styles: Relational formats:

## Social network graphs:

- A version of node-edge (or box-and-arrow) graphs, used to display connections among people, companies, etc.
- More interesting social network graphs may include different kinds of nodes and different kinds of relationships or connections.
- Genograms (family trees) are a specific kind of social network graph

# Specific Graphs, Layouts, and Axis Styles: Spatial formats:

## Geographic map:

- Some common practices for representing data on maps include color coding geographic regions; overlaying lines to represent relationships, connections, or movements of things or data; and representing local quantities with bubbles or pies per region.
- Another approach involves distorting the representation of physical space to reflect logical meaning in the data; the resulting map is called a cartogram.



# Specific Graphs, Layouts, and Axis Styles: Spatial formats:

## Non-Geographic map:

- Images can use the metaphorical idea of spatial relations to represent any number of other concepts, from mapping a winning strategy to product maps, to the map of my heart.
- Leveraging a spatial metaphor can be a powerful way to convey intent, relationship, sequence, influence, and process.

# Use of Circles and Circular Layouts

Circles and rounds are very frequently used to represent data.

Circles and circular layouts are also used improperly, in ways that don't allow efficient decoding of the data.

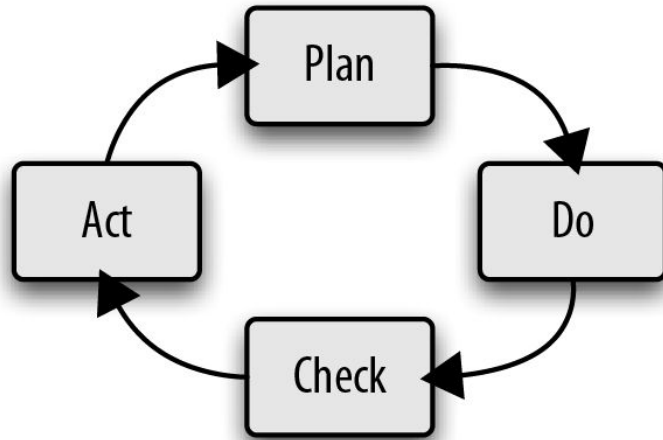
The fundamental problem with using circles: not very good at estimating and comparing circular areas and radial measures, such as arc lengths and pie wedge sizes.

# Use of Circles and Circular Layouts

## Cyclical relationships

- Annual seasons, 24-hour days, biological processes such as the Krebs Cycle, or any other repeating phenomenon can be well-represented by a circular layout.
- Please be careful not to use circular layouts to represent “cycles” for which the process may repeat, but the starting phase doesn’t (or shouldn’t) follow from the previous ending phase

# Use of Circles and Circular Layouts



Good use of circular layout: Krebs Cycle (top) and Deming Cycle (bottom).



FEMA  
causality

misunderstands

cyclical

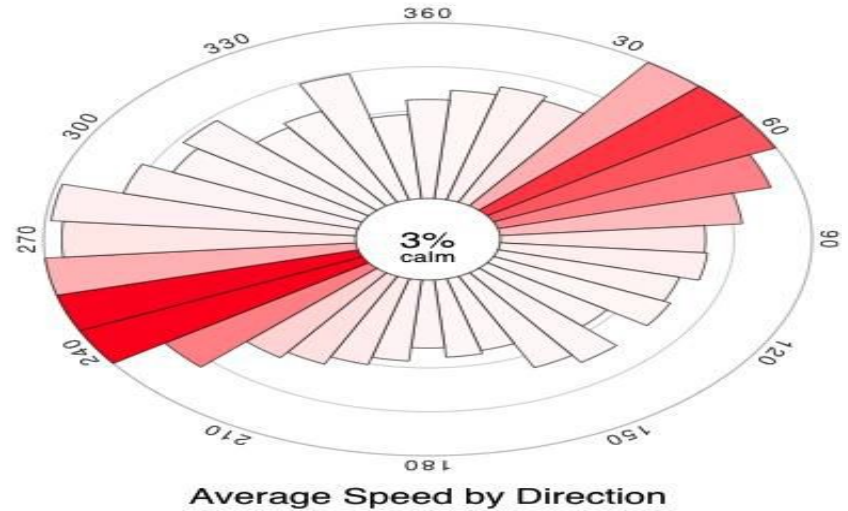
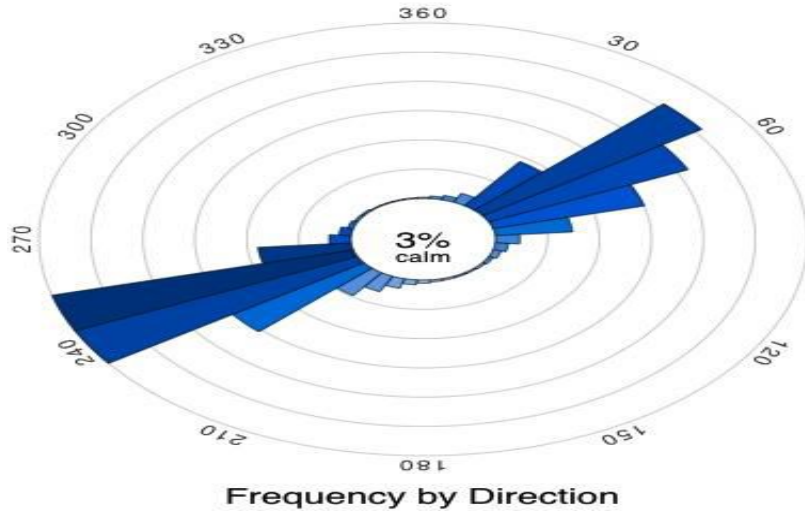
# Use of Circles and Circular Layouts

## Direction

- When dealing with directional data, round layouts that represent direction are entirely appropriate and useful, as they directly indicate the physical reality of direction.
- The wind roses show wind frequency and speed by direction.
- This presentation is more easily understandable than a bar graph or table of data.

# Use of Circles and Circular Layouts

## KGDP: Guadalupe Pass



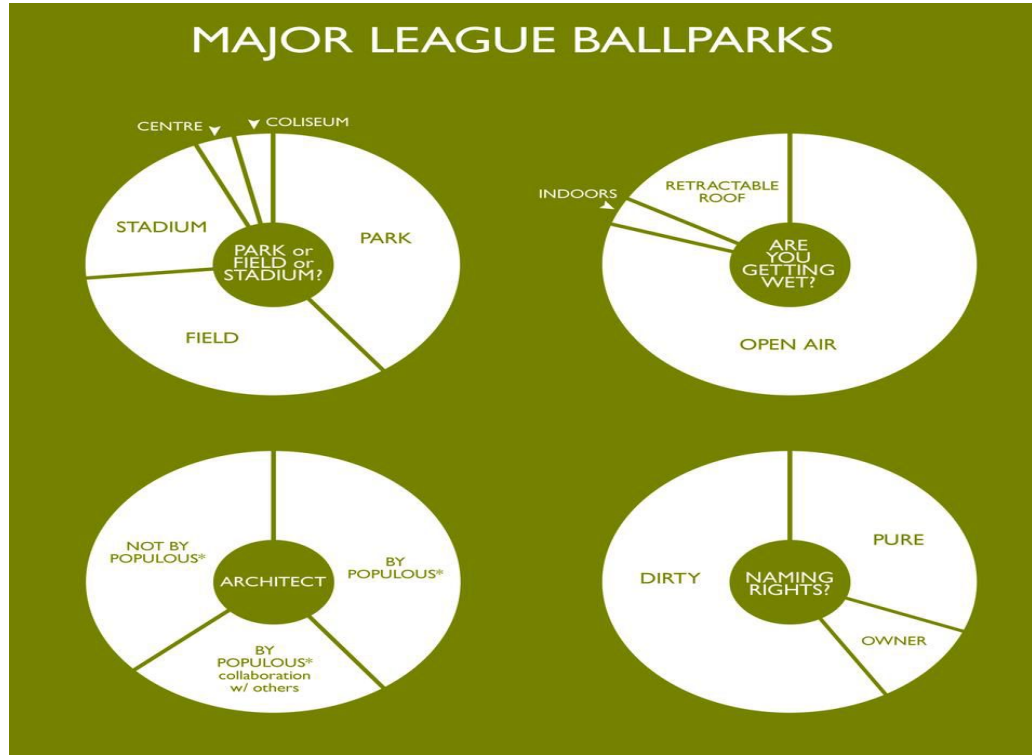
Wind roses show frequency and speed of wind by direction.

# Use of Circles and Circular Layouts

## Pie graphs

- Some data visualization professionals say to use pie chart, some say never ever use pie charts. (The truth lies somewhere in between)
- There are a few relevant caveats.
  - We're better at comparing rectangular areas than circular areas, so comparison of bar length is easier than comparison of circle sizes.
  - We're better at comparing length than angles, so comparison of bar length is easier than comparison of pie wedges.
  - We're better at comparing shapes that have a common baseline than shapes that don't, so comparing wedges that are rotated relative to each other is more challenging than comparing angles that start at the same place.

# Use of Circles and Circular Layouts



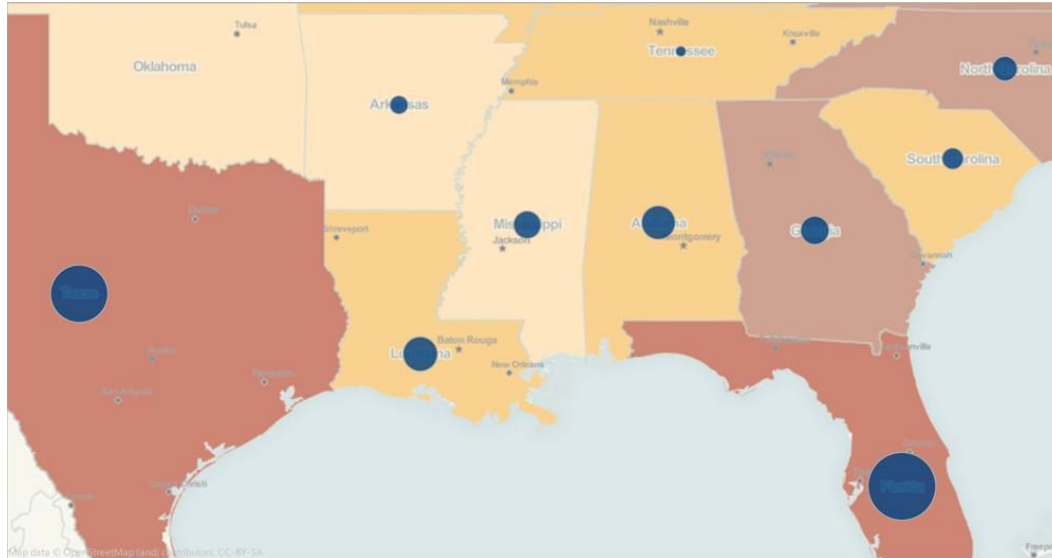
Good pie graphs, with few wedges and low need for precision



# Use of Circles and Circular Layouts

## Area on bubble graphs

- When scatter plots are encoded with different-sized circular marks, they are commonly called bubble graphs.



Circular areas used to represent and compare quantities per region on a map.

# Bad use of Circles and Circular Layouts

## Circular length and circular bar graphs

- The use of circular length is occasionally permissible, as in showing periods of time around a circular clock face.
- The use of circular length a huge “no-no,” as are circular bar graphs: as they create significant distortion of the relative length of the bars.
- Extracting meaningful comparisons from the data is very difficult under those circumstances.
- The better approach, as you may have guessed, is to use a plain old, boring, functional, efficient, bar graph.

# Bad use of Circles and Circular Layouts

## Pies with many or similarly-sized wedges

- If you need precision, or have many fractions/wedges that matter, don't use a pie graph. In those cases, bar graphs with precise labels are a better bet.

## Pies for values other than fractions of a whole

- Pie graphs are only for comparing fractions of a whole measure. If you are comparing other quantities—time data, or anything else—no pie for you!

END