THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Introduction to Big Data and Analytics
CSCI 6444
Data Visualization

Prof. Roozbeh Haghnazar

Slides Credit:

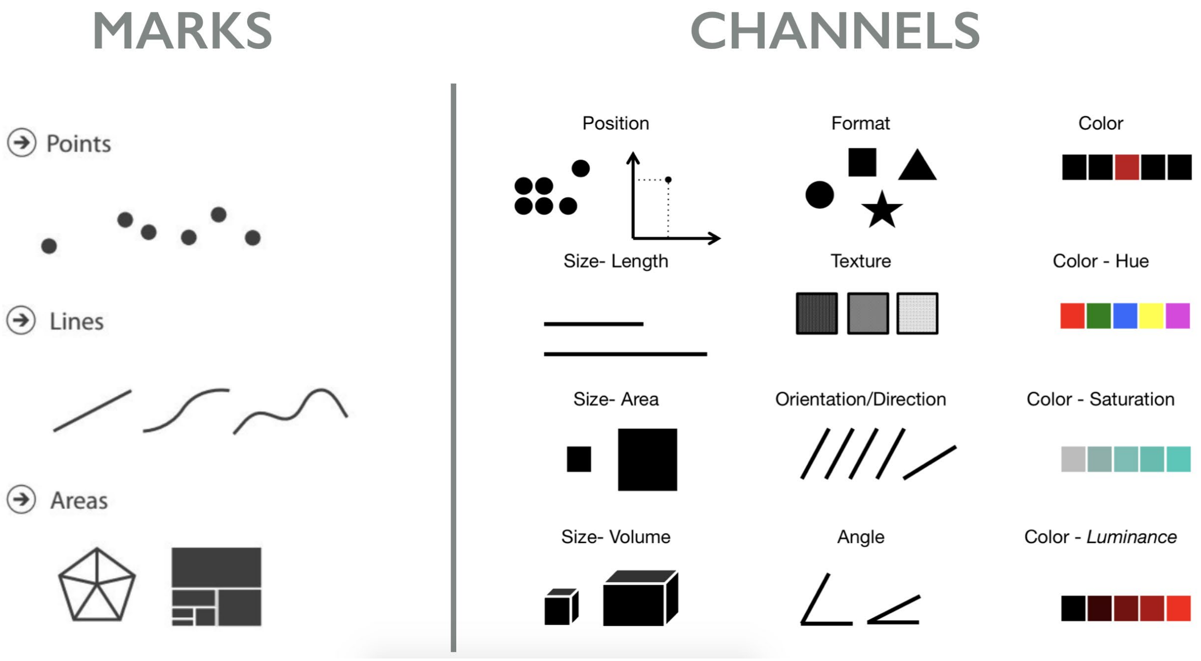
Stephen H. Kaisler

Introduction

- **The human brain processes images 60,000 times faster than text. A study performed by researchers from the University of Minnesota in 1986 found that presentations using visual aids were found to be 43% more persuasive than unaided presentations**
  **** Some researchers have doubt about test method that 3M company used**
- The weapons have changed, the battlefield is different, but the war rages on
- Security professionals have a hard time foreseeing new attacks because of how unique each one is.
- The amount of data to be analyzed has drastically increased. Not only the volume, but the variety of sources, and data types has made analysis difficult.

Using Marks and Channels

- In essence, in data visualization, marks are the geometric shapes we see (like bars, lines, or dots), and channels are the visual aspects of these shapes (like their color, size, or position) that convey specific data attributes. The effective use of marks and channels is crucial for creating clear, comprehensible, and insightful visualizations.
- All channels are not equal.
- the same data attribute encoded with two different visual channels will result in different information content in our heads after it has passed through the perceptual and cognitive processing pathways of the human visual system.
- The use of marks and channels in vis idiom design should be guided by the principles of expressiveness and effectiveness.
- Two principles guide the use of visual channels in visual encoding:
  - Expressiveness
  - Effectiveness

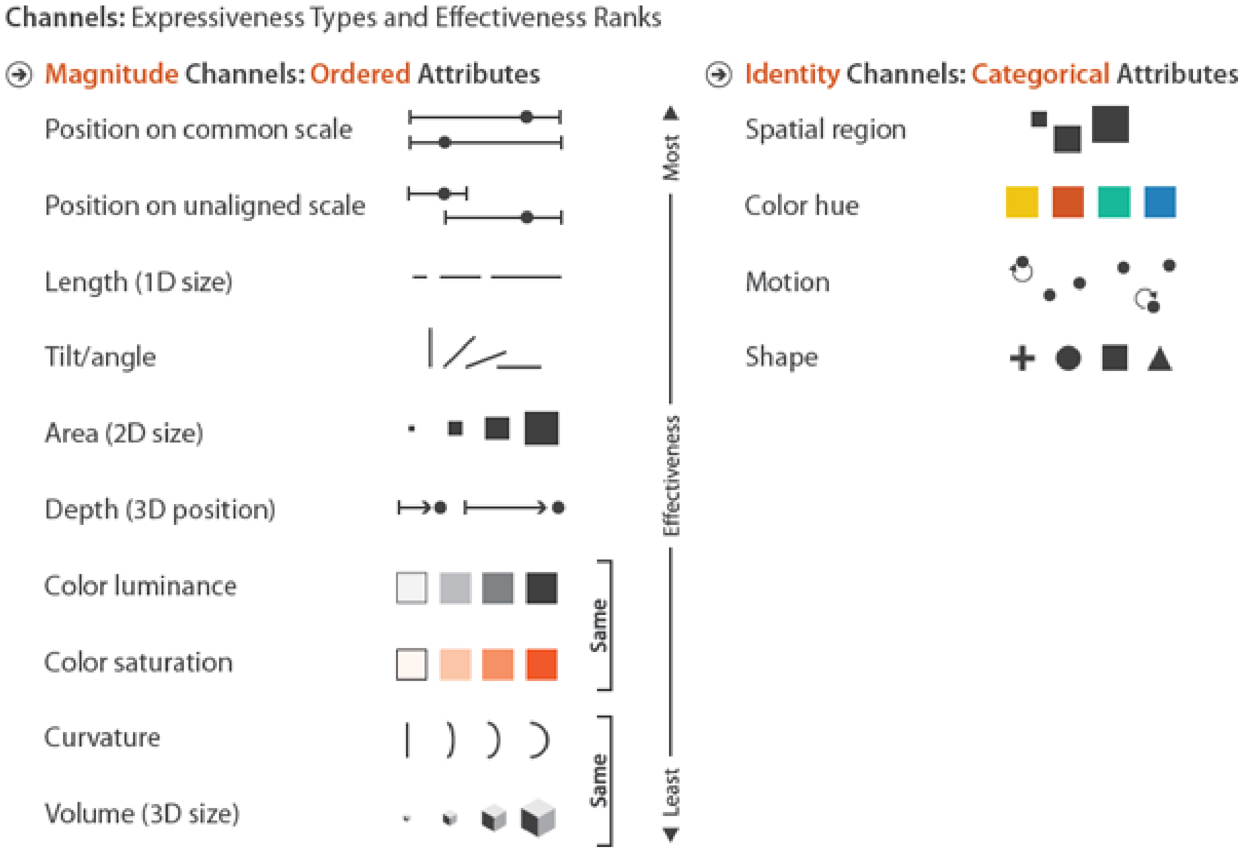


Expressiveness

- The **expressiveness principle** dictates that the visual encoding should express all of, and only, the information in the dataset attributes.
- A visual encoding is expressive if it represents all the distinctions and relationships present in the data without introducing any false implications or misleading interpretations.
- For example, a line graph is expressive for showing trends over time, as it naturally represents continuous data and the relationship between data points.

Expressiveness

- Magnitude channels
  - Purpose: To encode ordered (quantitative) data and show magnitude.
  - Examples: Length, size, area, and color luminance/saturation.
  - How it works: The perceptual system interprets the visual change as a difference in quantity. For instance, a longer bar clearly indicates a larger value.
  - Best practices: Should be used for ordered data. Using a magnitude channel for categorical data can be misleading, as it implies an order that doesn't exist.



Channels: Expressiveness Types and Effectiveness Ranks

Magnitude Channels: Ordered Attributes

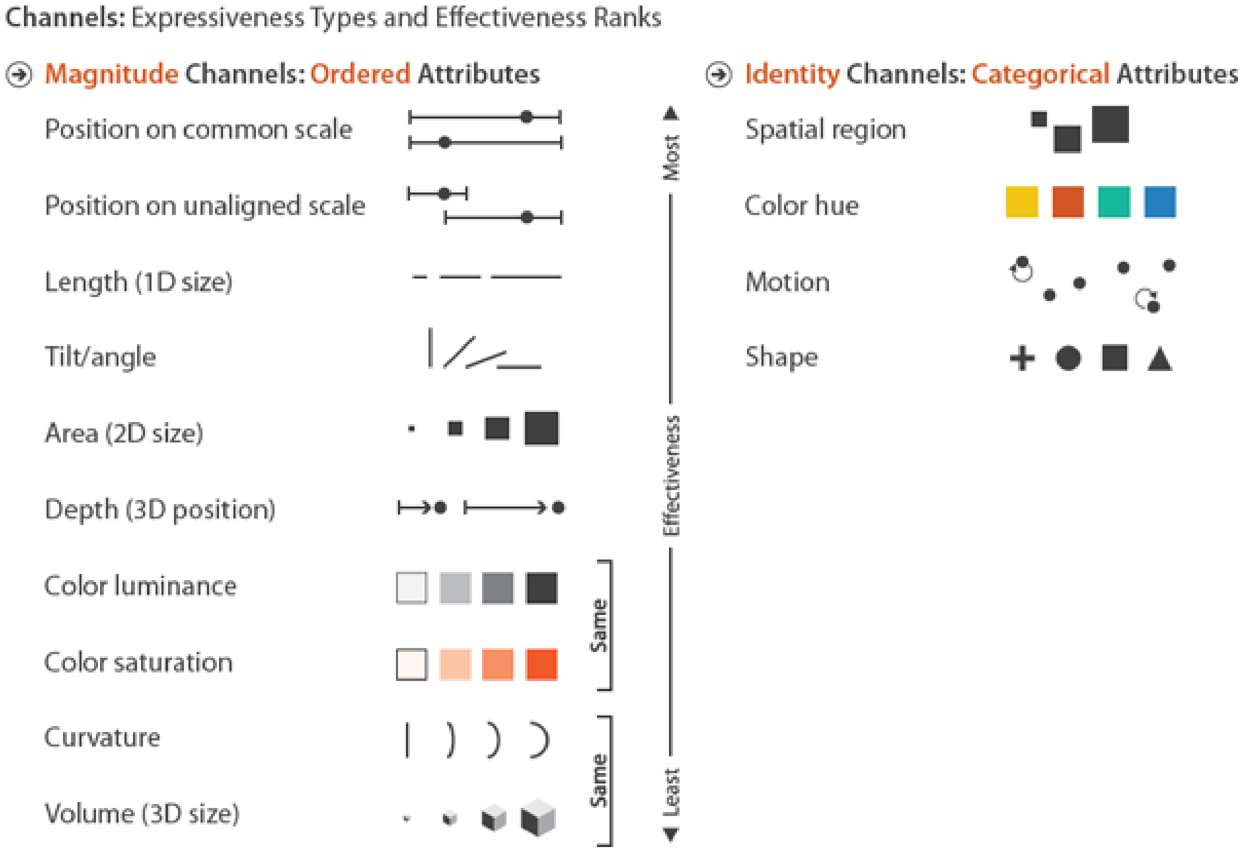
- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)

Identity Channels: Categorical Attributes

- Spatial region
- Color hue
- Motion
- Shape

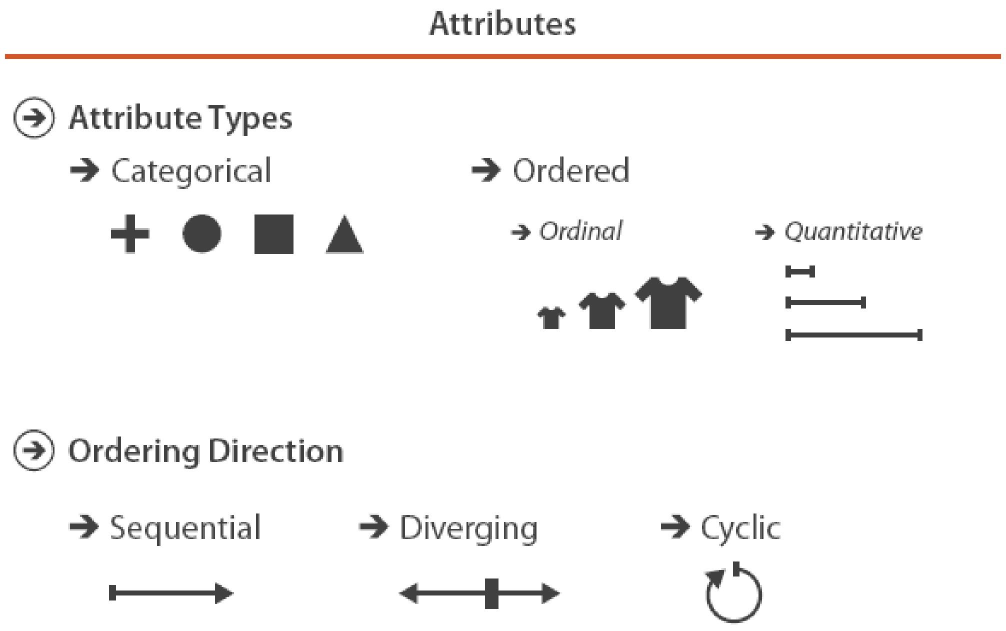
EXPRESSIVENESS

- Identity channels
  - Purpose: To encode categorical (qualitative) data and distinguish between different groups.
  - Examples: Color hue, shape, and spatial region.
  - How it works: The perceptual system automatically groups items with the same visual property (e.g., same color) together, which is useful for showing relationships and categories.
  - Best practices: Should be used for categorical data. Using an identity channel for ordered data will lose the magnitude information.



Attribute Types

- Categorical
- Ordered: Ordinal and Quantitative
  - Sequential versus Diverging
  - Cyclic
- Hierarchical Attributes



Attributes

Attribute Types

Categorical

Ordered

Ordinal

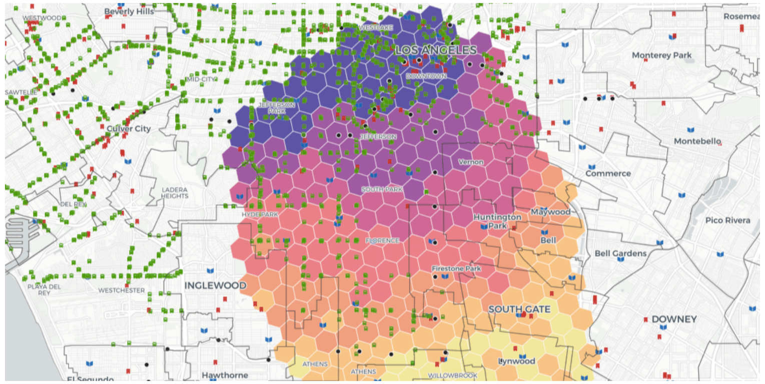
Quantitative

Ordering Direction

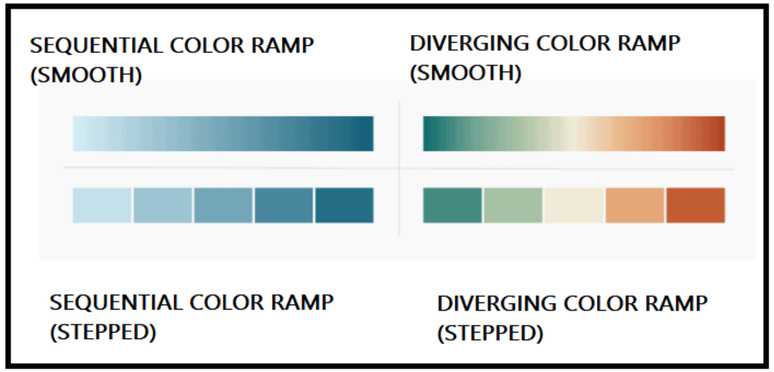
Sequential

Diverging

Cyclic



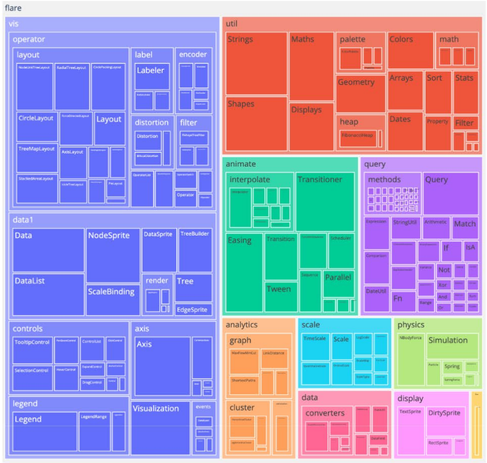
Z Score Visualization… Sequential versus Diverging?

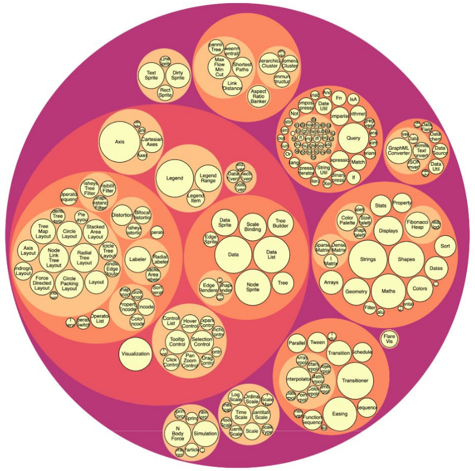

SEQUENTIAL COLOR RAMP (SMOOTH)
DIVERGING COLOR RAMP (SMOOTH)
SEQUENTIAL COLOR RAMP (STEPPED)
DIVERGING COLOR RAMP (STEPPED)

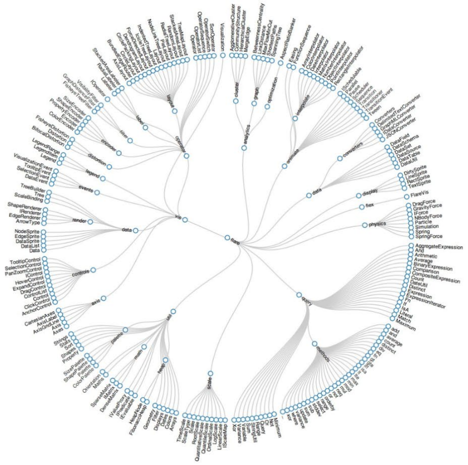
Attribute Types

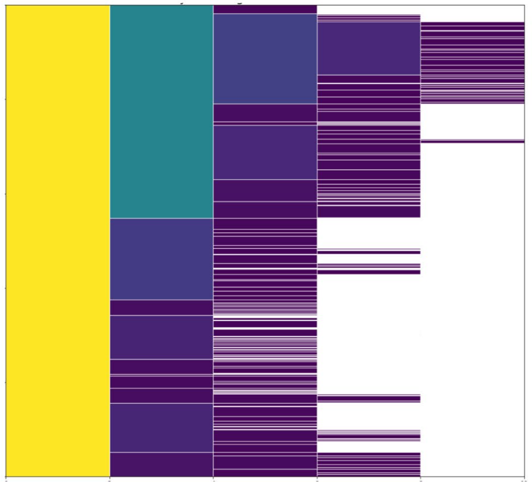
- Categorical
- Ordered: Ordinal and Quantitative
  - Sequential versus Diverging
  - Cyclic
- **Hierarchical Attributes**

A single "hierarchical attribute" can represent the entire hierarchy (e.g., "Time"), allowing users to navigate and filter through all its levels (Year, Quarter, Month, Day) at once. Alternatively, each level can be a separate "normal" attribute, requiring a separate definition for the relationship between them



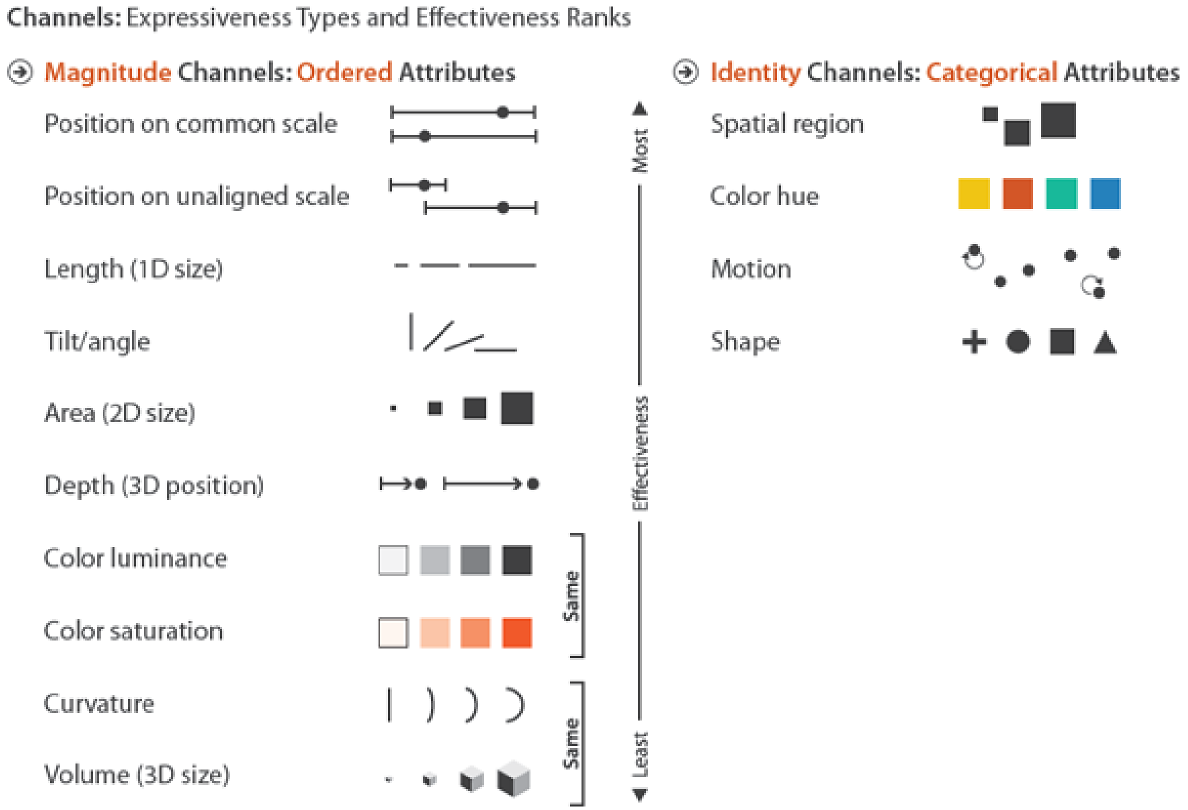






Channel Rankings

- While it is possible in theory to use a magnitude channel for categorical data or a identity channel for ordered data, that choice would be a poor one because the expressiveness principle would be violated.
- The spatial channels are the only ones that appear on both lists; none of the others are effective for both data types.



Channels: Expressiveness Types and Effectiveness Ranks

Magnitude Channels: Ordered Attributes

Position on common scale

Position on unaligned scale

Length (1D size)

Tilt/angle

Area (2D size)

Depth (3D position)

Color luminance

Color saturation

Curvature

Volume (3D size)

Identity Channels: Categorical Attributes

Spatial region

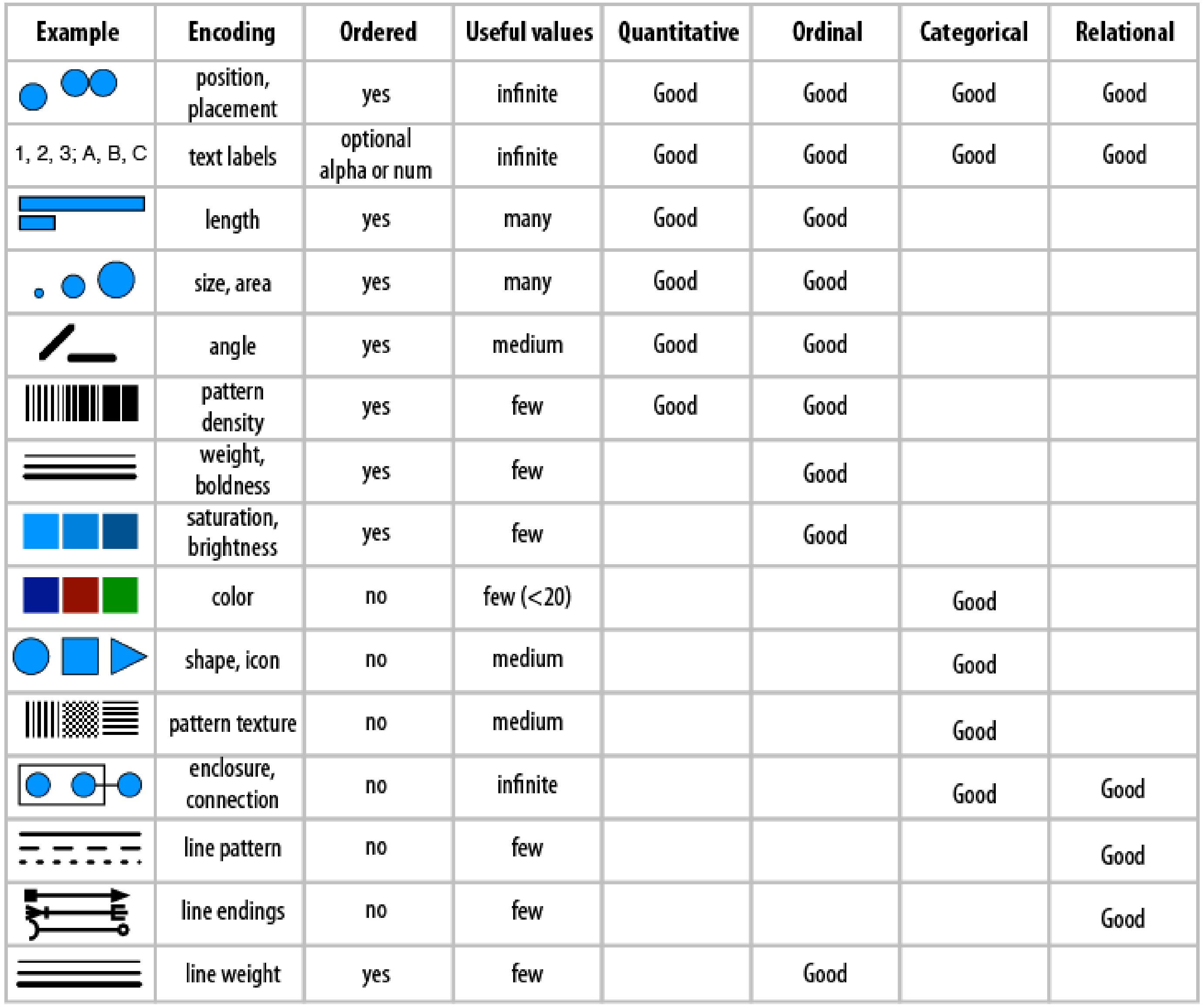
Color hue

Motion

Shape

CHANNEL RANKINGS

- **mental model:** internal mental representation used for thinking and reasoning
- The attributes encoded with position will dominate the user’s **mental model** compared with those encoded with any other visual channel.



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		

EFFECTIVENESS

- The **effectiveness principle** dictates that the importance of the attribute should match the **salience** of the channel
  - How are these rankings justified?
  - Why did the designer decide to use those particular visual channels?
  - How many more visual channels are there?
  - What kinds of information and how much information can each channel encode?
  - Why are some channels better than others?
  - Can all of the channels be used independently or do they interfere with each other?
- In other words, **the most important attributes** should be encoded with the **most effective channels** in order to be most noticeable, and then decreasingly important attributes can be matched with less effective channels

Effectiveness

- **Accuracy :**The obvious way to quantify effectiveness is **accuracy**
  - how close is human perceptual judgement to some objective measurement of the stimulus?
- We perceive different visual channels with different levels of accuracy; they are not all equally distinguishable.

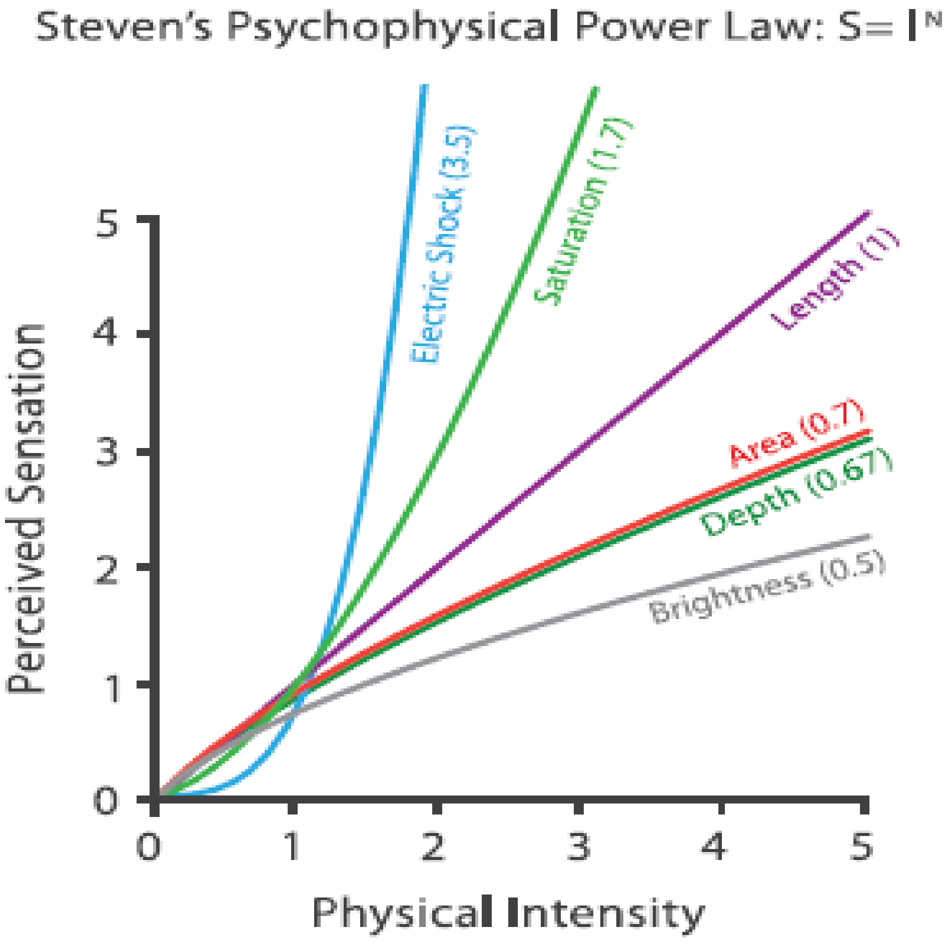
Effectiveness (Accuracy)

- Our responses to the sensory experience of magnitude are characterizable by power laws, where the exponent depends on the exact sensory modality:
  - most stimuli are magnified or compressed, with few remaining unchanged.
- **Stevens' Power Law:** The apparent magnitude of all sensory channels follows a power function based on the stimulus intensity: $S = Ki^n$
  - S is the perceived sensation
  - I is the physical intensity
  - K is a constant that varies depending on the sensory channel and the units of measurement.
  - The power law exponent n ranges from the sublinear 0.5 for brightness to the superlinear 3.5 for electric current

Effectiveness (Accuracy)

- The sublinear phenomena are compressed
- The superlinear phenomena are magnified
- length has an exponent of *n* = 1.0, so our perception of length is a very close match to the true value.

- **Brightness (sublinear, *n*≈0.5)**: If you double the physical intensity of a light source (e.g., the number of lumens), the perceived brightness does not double; it only increases by a factor of about the square root of 2 (√2). This is a sublinear relationship because the perception grows more slowly than the intensity.
- **Electric Current (superlinear, *n*≈3.5)**: If you increase the intensity of an electric current, the perceived sensation increases at a much faster rate. For instance, if you double the physical intensity of the electric current, the perceived sensation increases by a factor of $2^{3.5}$ (which is about 11.3). This is a superlinear relationship because the perception grows more quickly than the intensity.

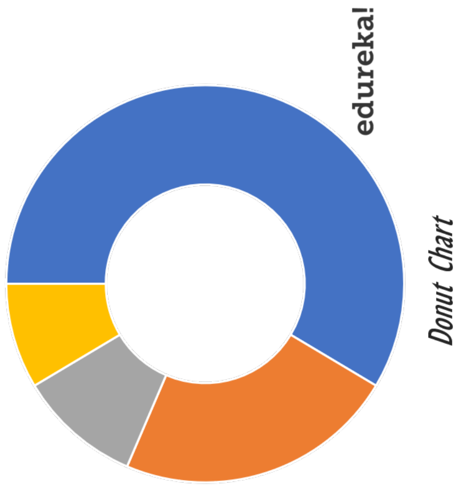


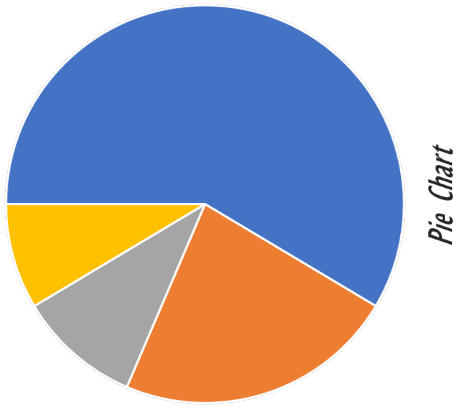
Effectiveness (Accuracy)

- Cleveland and McGill’s experiments on the magnitude channels [Cleveland and McGill 84a] showed that aligned position against a common scale is most accurately perceived, followed by unaligned position against an identical scale, followed by length, followed by angle. Area judgements are notably less accurate than all of these.

Here is a breakdown of the visual channels mentioned in order of accuracy:

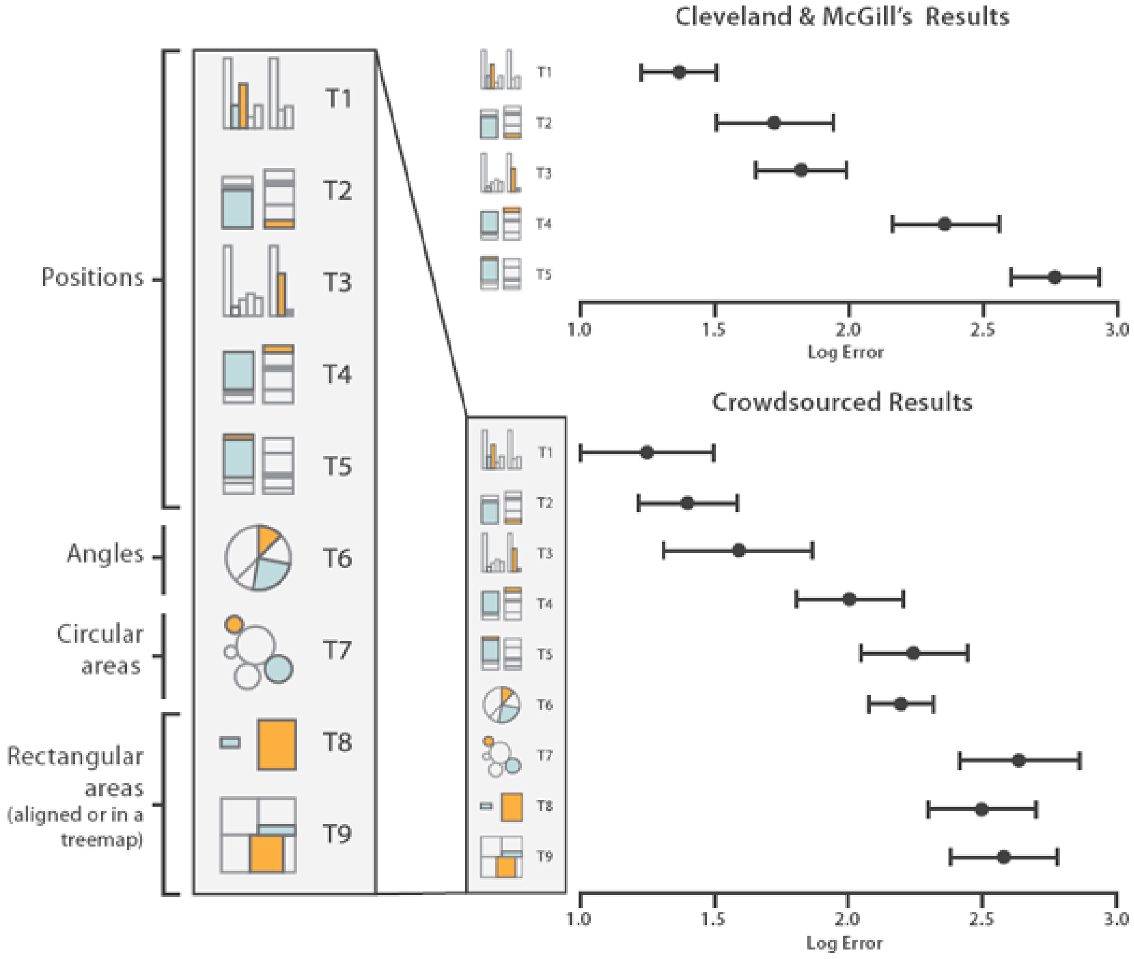
1. **Aligned Position Against a Common Scale**: This is the most accurate type of encoding for our visual system. It refers to the placement of visual elements along a common scale, such as the position of bars in a bar chart or points on a line graph, where all elements start from a common baseline. People can very precisely compare the length or height of these elements relative to each other and the scale.
2. **Unaligned Position Against an Identical Scale**: This is similar to aligned position but without a common baseline for all elements. An example might be the position of marks in a scatter plot. This is less accurate than aligned positions because the lack of alignment makes direct comparison slightly more challenging.
3. **Length**: This refers to the ability to judge the length of an object, such as a bar in a bar chart, without an aligned scale. While still quite accurate, it is not as precise as when position is aligned against a common scale.
4. **Angle**: This is the visual encoding of information through angles, such as in pie charts or the slope of a line in a line chart. Judgments based on angle are less accurate than those based on length or position, as our ability to perceive precise differences in angles is not as developed.
5. **Area**: Area encodings, such as the size of bubbles in a bubble chart, are less accurate still. This is because our ability to compare areas is not as refined as our ability to compare lengths or positions. The perception of area does not scale linearly with the actual area size, which makes it difficult to make accurate judgments of relative size.





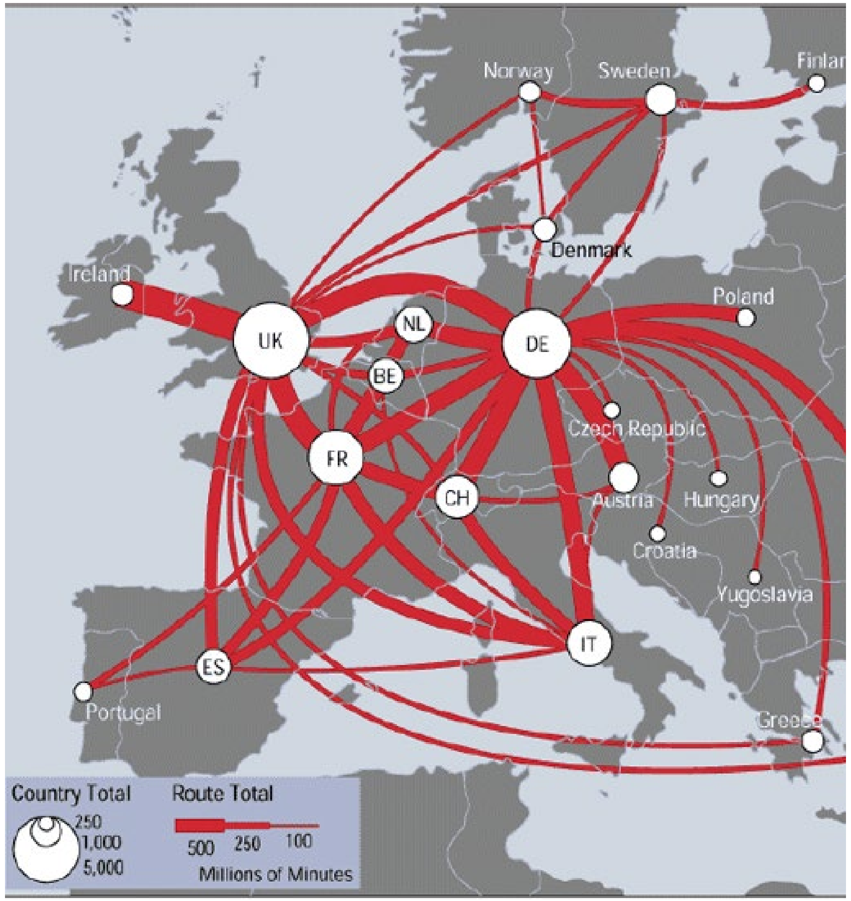
Effectiveness (Accuracy)

- Error rates across visual channels, with recent crowdsourced results replicating and extending seminal work from Cleveland and McGill [Cleveland and McGill 84a]



Effectiveness

- **Discriminability:** if you encode data using a particular visual channel, are the differences between items perceptible to the human as intended?
- The characterization of visual channel thus should quantify the number of bins that are available for use within a visual channel, where each bin is a distinguishable step or level from the other.
- For instance, some channels have a very limited number of bins. Consider line width: changing the line size only works for a fairly small number of steps. Increasing the width past that limit will result in a mark that is perceived as a polygon area rather than a line mark. A small number of bins is not a problem if the number of values to encode is also small. For example, Figure 5.9 shows an example of effective linewidth use. Linewidth can work very well to show three or four different values for a data attribute, but it would be a poor choice for dozens or hundreds of values. The key factor is matching the ranges: the number of different values that need to be shown for the attribute being encoded must not be greater than the number of bins available for the visual channel used to encode it. If these do not match, then the vis designer should either explicitly aggregate the attribute into meaningful bins or use a different visual channel.



EFFECTIVENESS

- **Separability:**You cannot treat all visual channels as completely independent from each other, because some have dependencies and interactions with others.

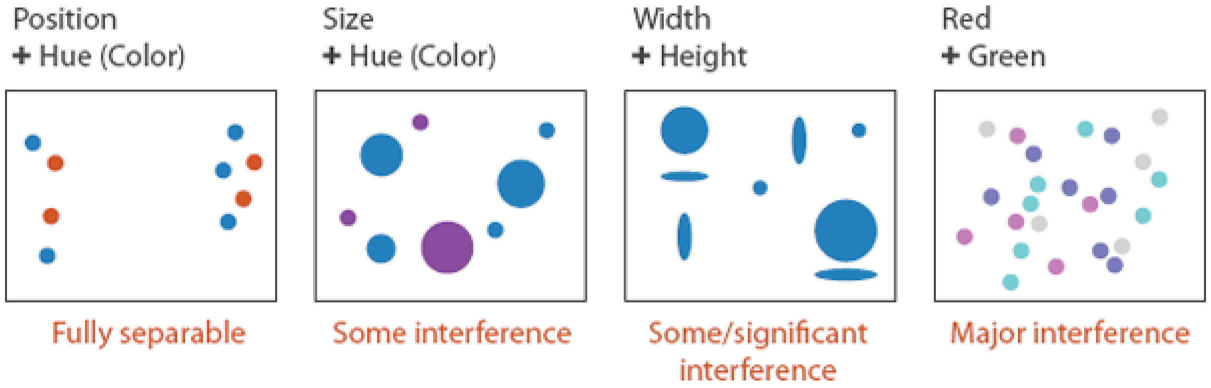
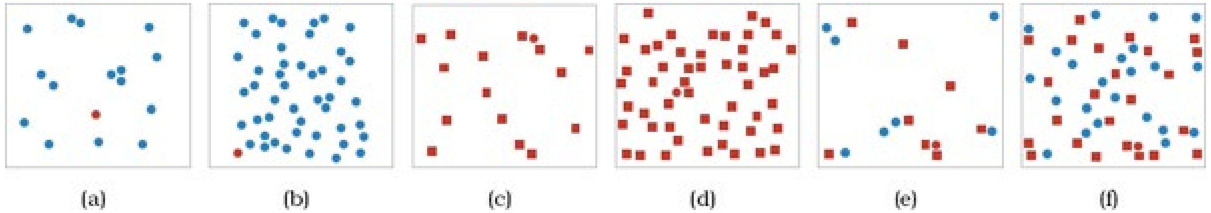
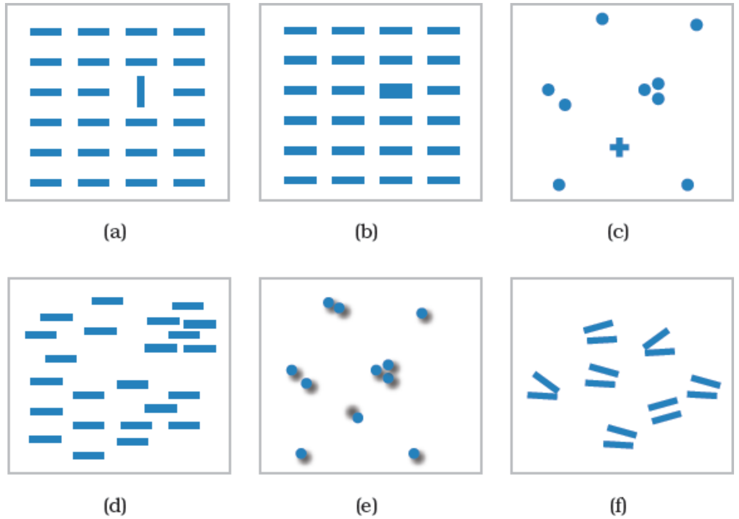


Figure 5.10 shows pairs of visual channels at four points along this continuum. On the left is a pair of channels that are completely separable: position and hue. We can easily see that the points fall into two categories for spatial position, left and right. We can also separately attend to their hue and distinguish the red from the blue. It is easy to see that roughly half the points fall into each of these categories for each of the two channels.
Next is an example of interference between channels, showing that size is not fully separable from color hue. We can easily distinguish the large half from the small half, but within the small half discriminating between the two colors is much more difficult. Size interacts with many visual channels, including shape.The third example shows an integral pair. Encoding one variable with horizontal size and another with vertical size is ineffective because what we directly perceive is the planar size of the circles, namely, their area. We cannot easily distinguish groupings of wide from narrow, and short from tall. Rather, the most obvious perceptual grouping is into three sets: small, medium, and large. The medium category includes the horizontally flattened as well as the vertically flattened.
The far right on Figure 5.10 shows the most inseparable channel pair, where the red and green channels of the RGB color space are used. These channels are not perceived separately, but integrated into a combined perception of color. While we can tell that there are four colors, even with intensive cognitive effort it is very difficult to try to recover the original information about high and low values for each axis. The RGB color system used to specify information to computers is a very different model than the color processing systems of our perceptual system, so the three channels are not perceptually separable.

Effectiveness

- **Popout :**Many visual channels provide visual **popout**, where a distinct item stands out from many others immediately



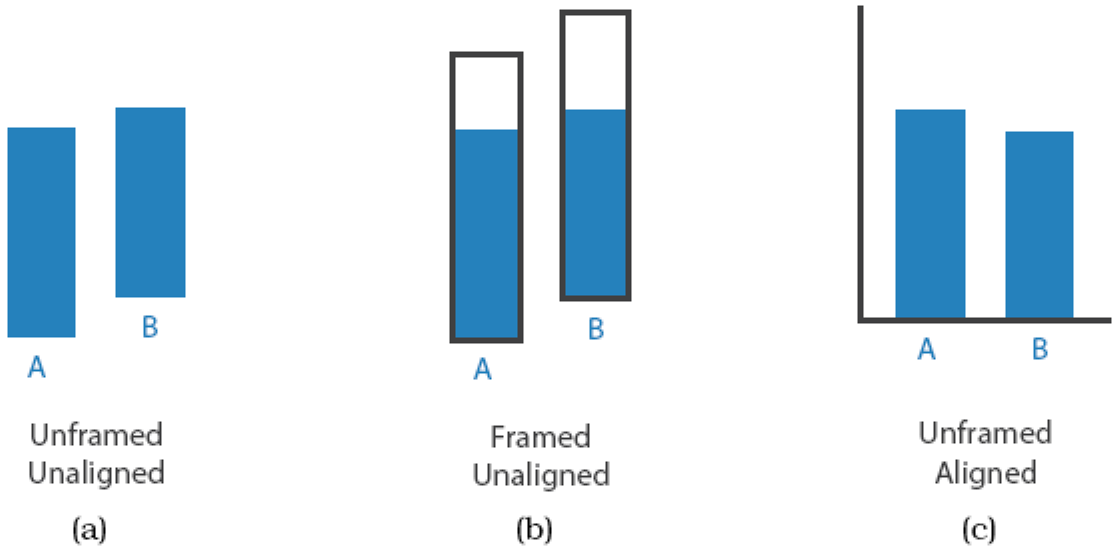


EFFECTIVENESS

- **Grouping :**The effect of perceptual grouping can arises from either the use of link marks or from the use of identity channels to encode categorical attributes
- Containment is the strongest cue for grouping, with connection coming in second.

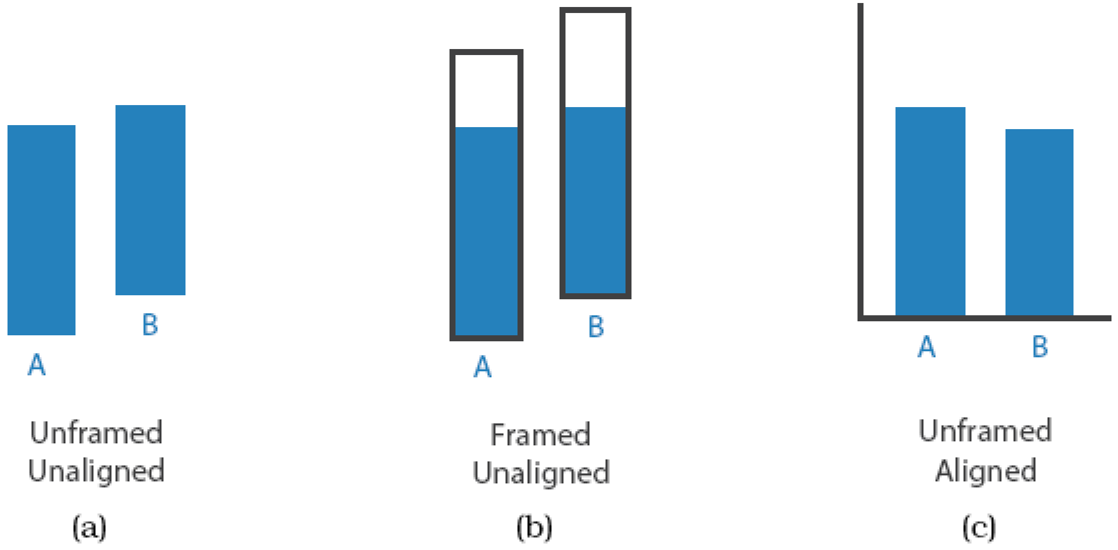
Relative versus Absolute Judgements

- **Weber's Law** is typically stated as the detectable difference in stimulus intensity I as a fixed percentage K of the object magnitude: $\delta I/I = K$.



Relative versus Absolute Judgements

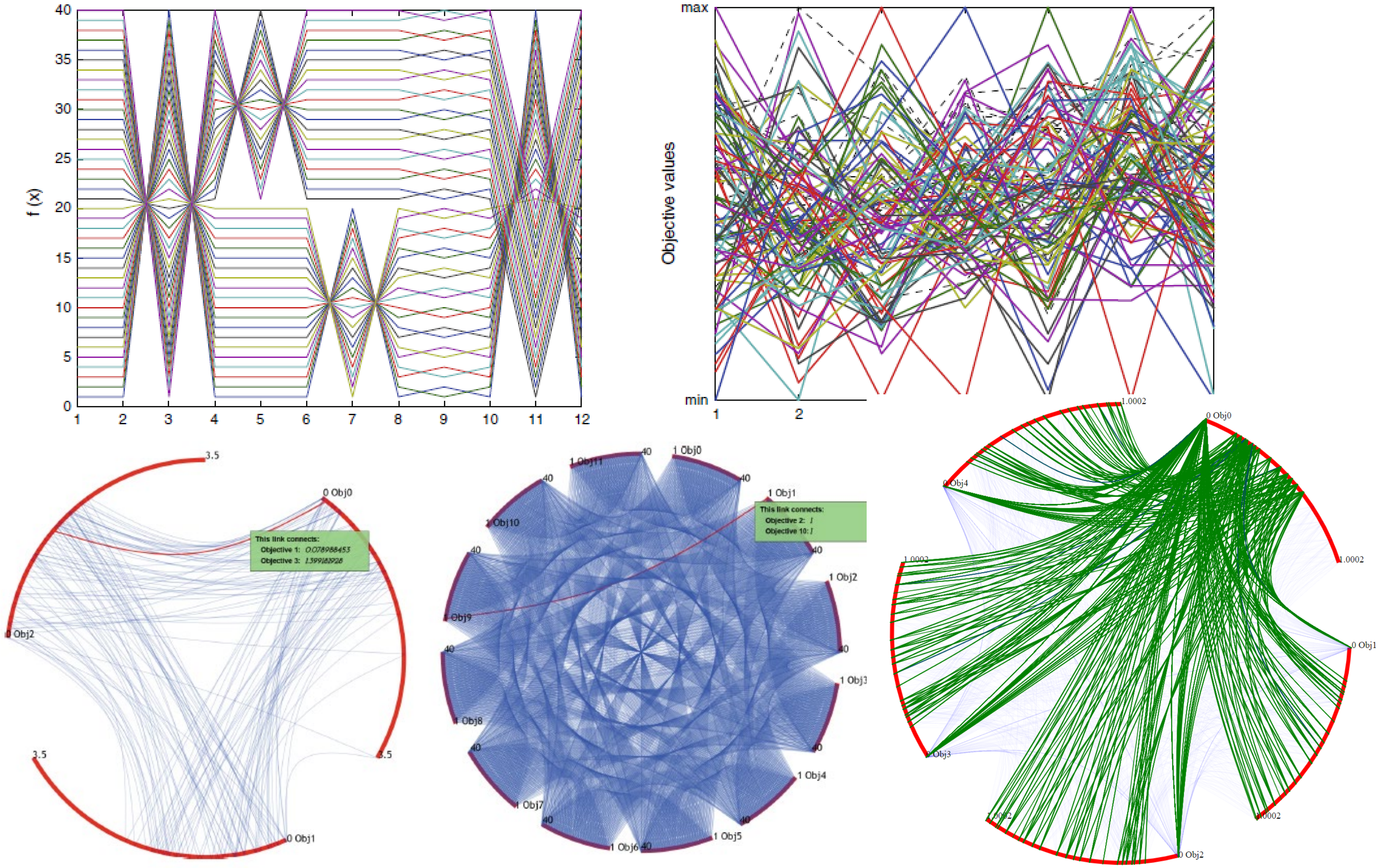
- **Weber's Law** is typically stated as the detectable difference in stimulus intensity I as a fixed percentage K of the object magnitude: δI/I = K.



Effectiveness and Expressiveness

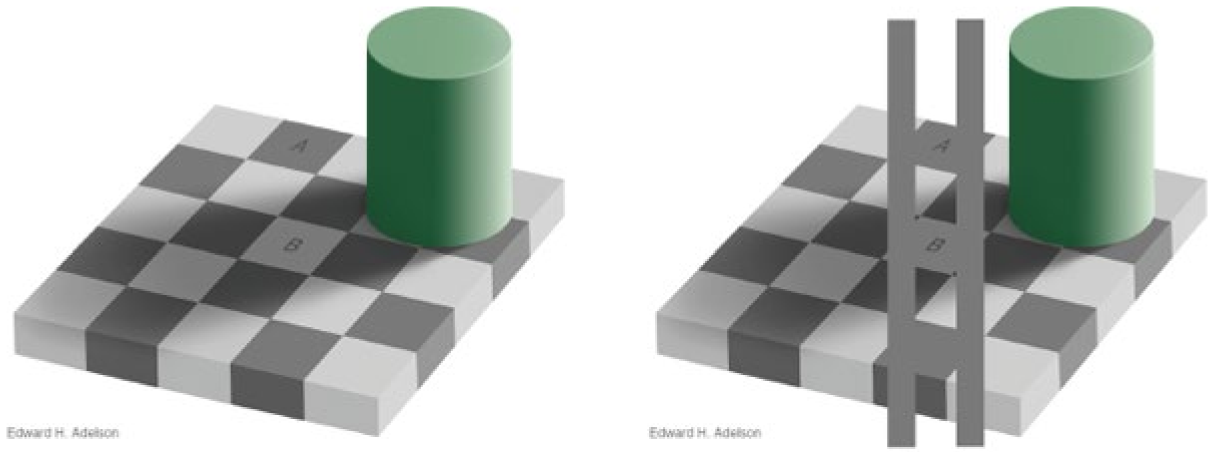
- Effectiveness and expressiveness are two key principles for good data visualization. **Expressiveness** means the visual encoding represents all and only the data attributes, ensuring that ordered data is shown as ordered and unordered data is not given a false order. **Effectiveness** is about how well the visual encoding is perceived, where the most important attributes are matched with the most noticeable visual channels
- A good visualization is both expressive and effective. It is incorrect or misleading if it lacks expressiveness, and it is inefficient or difficult to understand if it lacks effectiveness.
- The expressiveness principle is fundamental, ensuring the chart is a faithful representation of the data, while effectiveness is about optimizing the design for human perception.

Relative versus Absolute Judgements



Relative versus Absolute Judgements

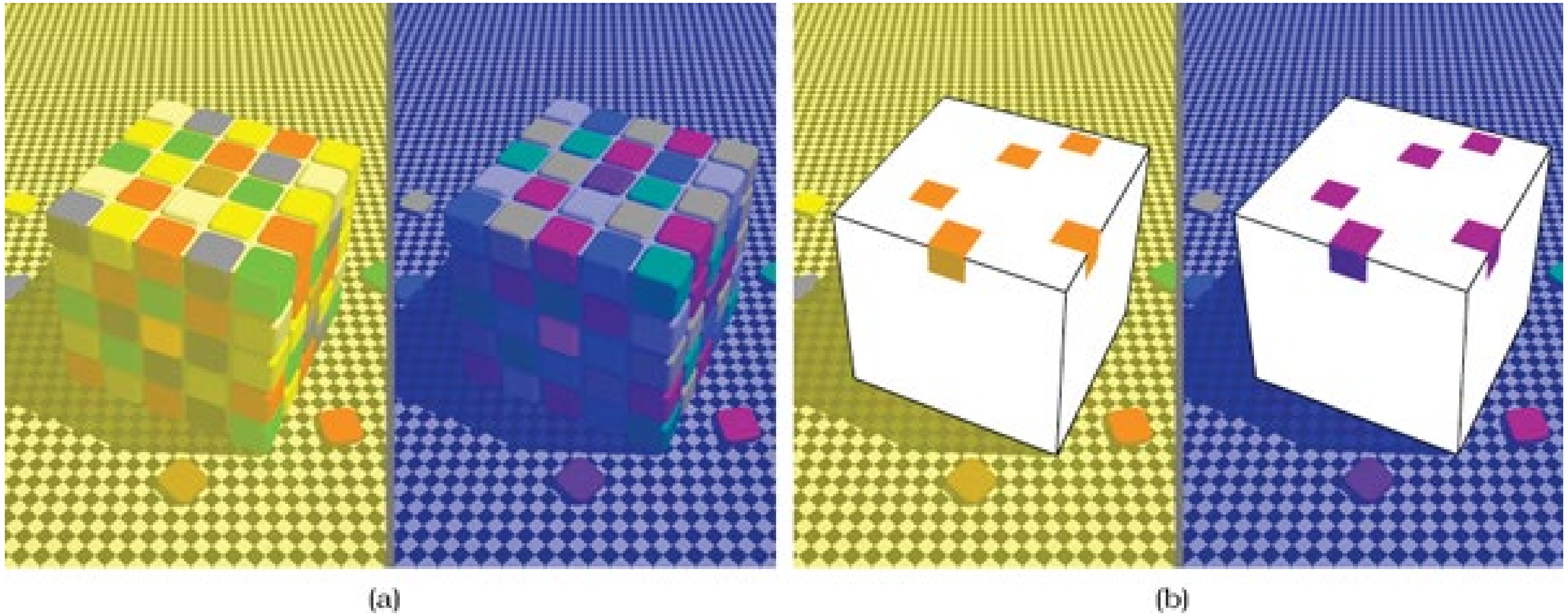
- our perception of color and luminance is completely contextual, based on the contrast with surrounding colors.
- Luminance perception is based on relative, not absolute, judgements. (a) The two squares A and B appear quite different. (b) Superimposing a gray mask on the image shows that they are in fact identical.



(a) (b)

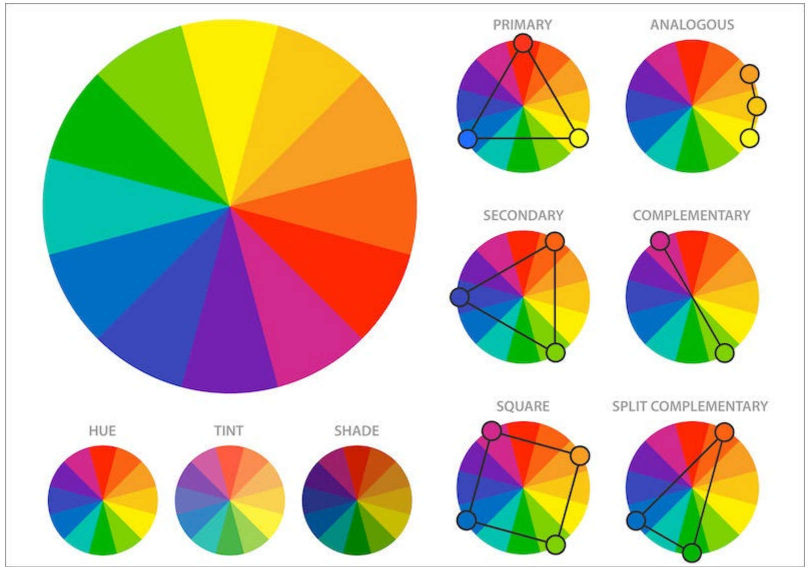
Relative versus Absolute Judgements

- Color perception is also relative to surrounding colors and depends on context. (a) Both cubes have tiles that appear to be red. (b) Masking the intervening context shows that the colors are very different: with yellow apparent lighting, they are orange; with blue apparent lighting, they are purple.



COLOR THEORY

- **Primary Colors**: Red, yellow and blue
- **Secondary Colors**: Green, orange and purple
- **Tertiary Colors:** Yellow-orange, red-orange, red-purple, blue-purple, blue-green & yellow-green

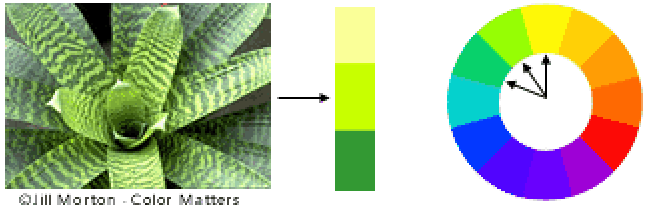


Color Harmony

- In visual experiences, harmony is something that is pleasing to the eye. It engages the viewer and it creates an inner sense of order, a balance in the visual experience.

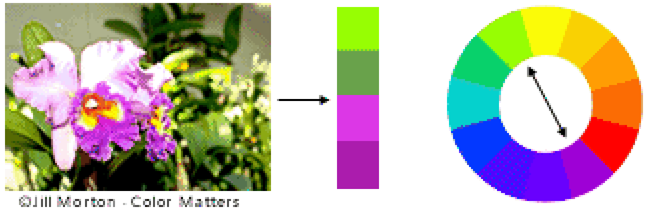
Color Harmony

- A color scheme based on analogous colors



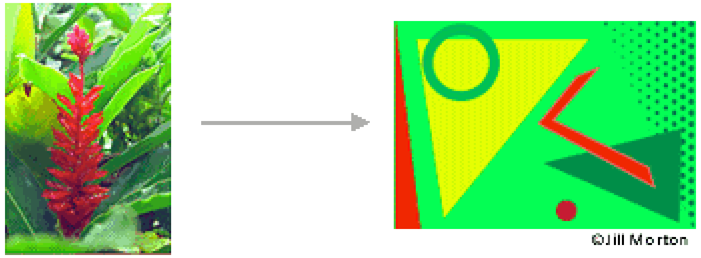
©Jill Morton - Color Matters

- A color scheme based on complementary colors



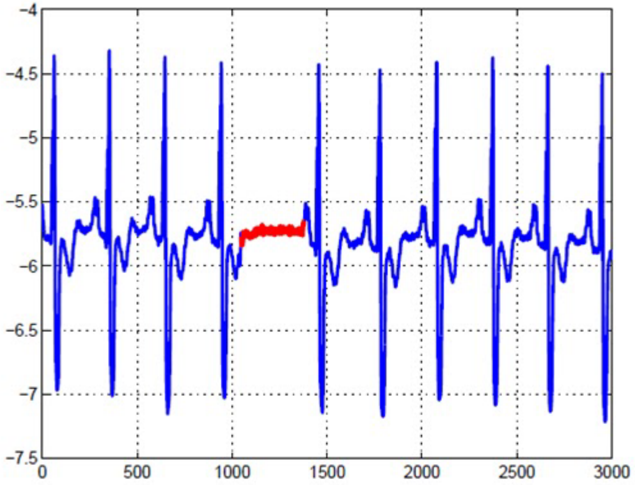
©Jill Morton - Color Matters

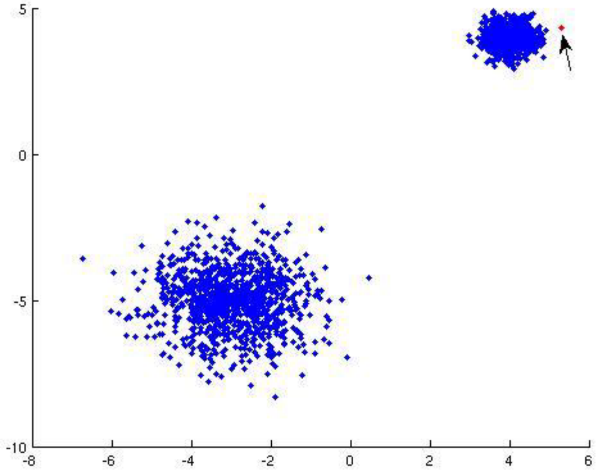
- A color scheme based on nature

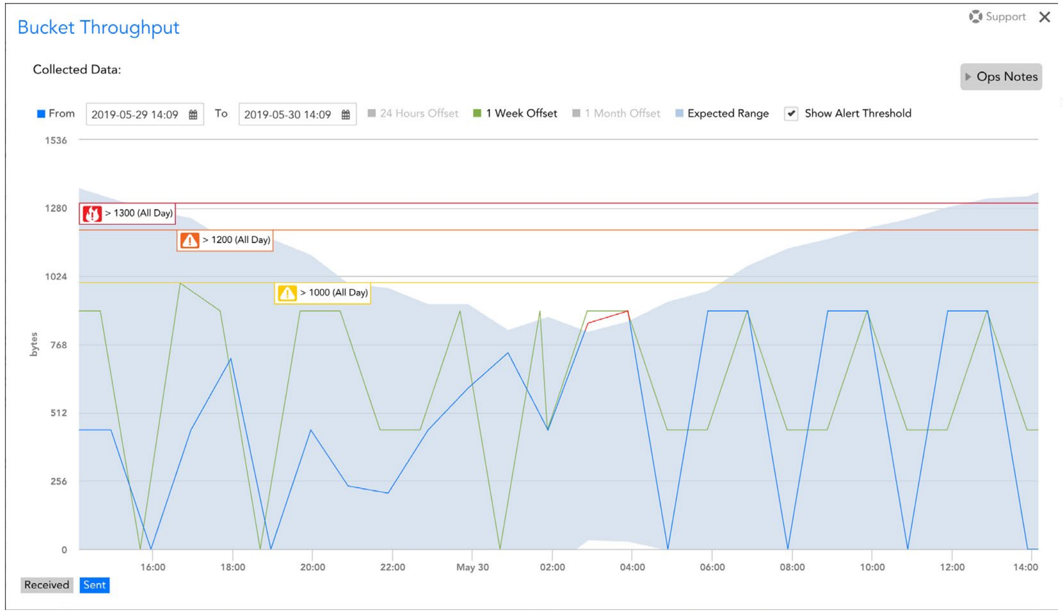


©Jill Morton

Anomaly Detection by Visualization







Bucket Throughput

Collected Data:

🟦 From 2019-05-29 14:09 To 2019-05-30 14:09 ▨ 24 Hours Offset 🟩 1 Week Offset ▨ 1 Month Offset ▨ Expected Range ☑ Show Alert Threshold

MOST POPULAR TOOLS

1. **Tableau**: A powerful business intelligence and data visualization tool that allows users to create interactive and shareable dashboards. It's known for its ability to handle vast amounts of data and perform complex data blending and analytics.
2. **Microsoft Power BI**: A business analytics service provided by Microsoft that provides non-technical business users with tools for aggregating, analyzing, visualizing, and sharing data. Its user interface is fairly intuitive and integrates well with other Microsoft products.
3. **QlikView/Qlik Sense**: Qlik's data visualization and data discovery application with intuitive dashboard and reporting capabilities. It offers self-service data visualization, personalized reports, and dynamic dashboards.
4. **Google Data Studio**: A free tool from Google that turns your data into informative dashboards and reports that are easy to read, easy to share, and fully customizable. It integrates well with other Google services like Google Analytics and Google Sheets.
5. **R with ggplot2**: R is a statistical programming language that is highly extensible and powerful for data analysis. Ggplot2 is a plotting system for R, based on the grammar of graphics, which provides a powerful model for building data visualizations layer by layer.
6. **Python with libraries like Matplotlib, Seaborn, Plotly**: Python is another programming language that is widely used for data analysis and machine learning. Libraries such as Matplotlib and Seaborn are used for static and customizable visualizations, while Plotly is used for interactive plots.
7. **SAS Visual Analytics**: A software suite for data science, advanced analytics, and data visualization. It is particularly strong in the area of statistical analysis.
8. **Looker**: A data-exploration app that helps you make sense of your data. It is web-based and provides a platform for data discovery that makes it easy to find, explore, and understand your data.

JS based libraries and Frameworks

1.Deck.gl: This is probably the most well-known library in the vis.gl suite. Deck.gl is a WebGL-powered framework optimized for creating complex, large-scale geospatial visualizations. It's designed to be highly performant and capable of rendering millions of data points seamlessly.

2.Luma.gl: This is a foundational library in the vis.gl suite that provides a set of classes and utilities to work with WebGL in a more simplified way. It's the underlying rendering engine that powers Deck.gl.

2.React: While React itself is not a data visualization library, it is a JavaScript library for building user interfaces, often used as a framework in web development. React can be used in conjunction with data visualization libraries to create interactive visualizations. Libraries like react-vis, react-chartjs, and react-d3-components provide React components for easier integration of charts and visualizations in React applications.

3.D3.js: One of the most well-known data visualization libraries, D3.js allows you to bind arbitrary data to a Document Object Model (DOM), and then apply data-driven transformations to the document. It enables the creation of highly flexible and interactive data visualizations.

4.Three.js: This is a 3D graphics library that makes it possible to create complex 3D visualizations and animations directly in the browser using WebGL.

5.Chart.js: A simple yet flexible charting library that provides various chart types like line, bar, radar, doughnut, and pie charts. It's easy to use and can be integrated with React and other frameworks.

6.Highcharts: A charting library that offers a wide variety of chart types and integrates easily with your existing website or web application. It's often praised for its ease of use and the professional appearance of its charts.

7.Leaflet: An open-source JavaScript library for mobile-friendly interactive maps. It's designed to be lightweight and simple to use while offering a wide range of map features.

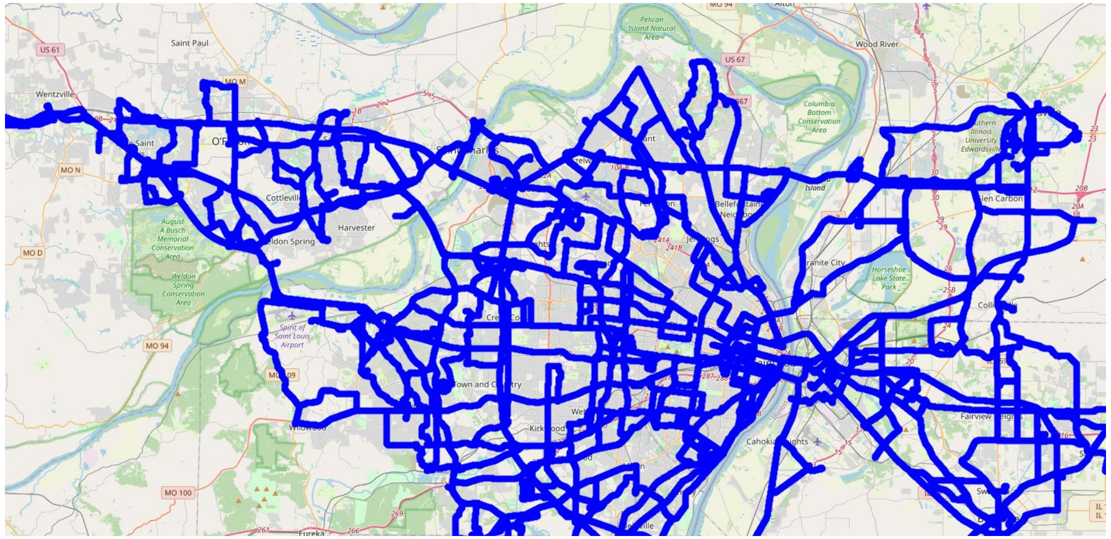
8.P5.js: This library is a JavaScript interpretation of the processing language. It is aimed at artists and designers and is used for creating graphic and interactive experiences, including data visualization, on the web.

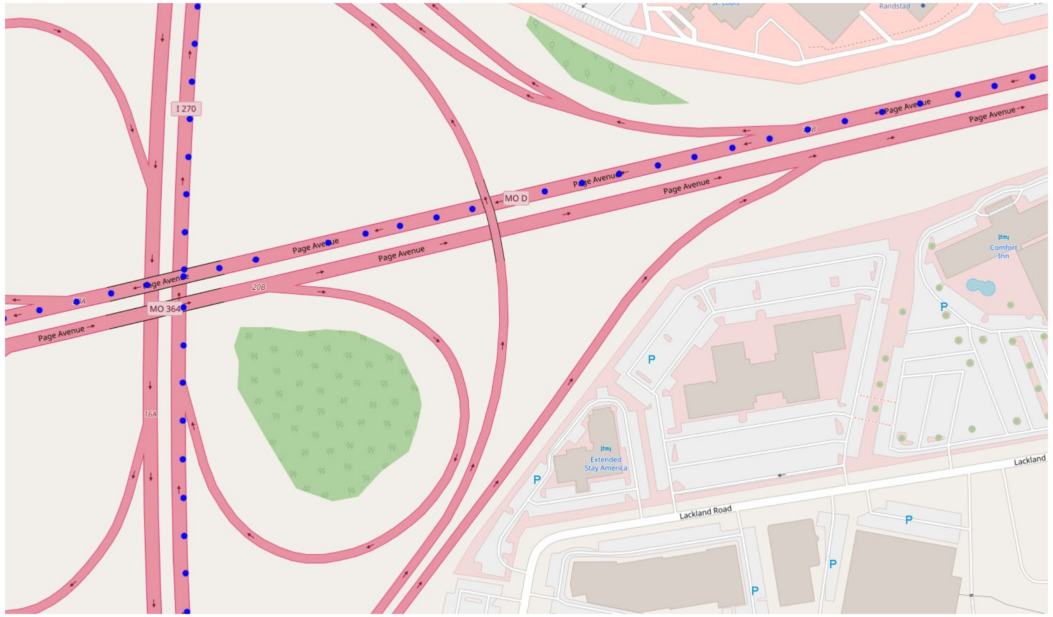
9.Plotly.js: Based on D3.js and stack.gl, Plotly.js is a high-level, declarative charting library that ships with over 30 chart types, including 3D charts, statistical graphs, and SVG maps.

10.Vega and Vega-Lite: Vega is a visualization grammar, a declarative format for creating and saving interactive visualization designs. Vega-Lite is a high-level grammar that provides a simpler and more concise syntax for rapidly creating common types of visualizations.

Assignment (Design your path analysis)

1. You’re given GPS point data from a city drive-test; each point has 4–5 KPIs (e.g., RSRP, SINR, throughput, drop rate). Your goal: visualize the data so a network engineer can instantly spot coverage gaps, congestion, and handoff issues.
2. Design an effective map-based view (e.g., hex/bin tiles, segments by road, or small multiples) with proper encodings (position, diverging/ sequential palettes, legends, bins). Avoid raw scatter overplot; justify every encoding choice.
3. Deliver:
   - Figures,
   - Derived “Quality Index” (formula + reasoning) from the KPIs,
   - 3 concrete insights with recommended actions (≤150 words).


Saint Paul
Wentzville
Wood River
Pelican Island Natural Area
Columbia Bottom Conservation Area
US 61
US 67
MO N
MO D
MO 94
MO 100
Cottleville
Harvester
August A Busch Memorial Conservation Area
Weldon Spring Conservation Area
Spirit of Saint Louis Airport
Granite City
Horseshoe Lake State Park
Glen Carbon
Fairview Heights
Cahokia Heights
Eureka



WEB BASED EXAMPLES

- http://www.cpdee.ufmg.br/~roozbeh.haghnazar/Mohamed_Distribution/
- http://www.cpdee.ufmg.br/~roozbeh.haghnazar/ResearchAnalyzing/
- http://www.cpdee.ufmg.br/~roozbeh.haghnazar/Mohamed_Harmony/

References

- https://www.recordedfuture.com/cyber-intelligence-visualization/
- https://www.sans.org/reading-room/whitepapers/metrics/security-data-visualization-36387
- https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=000AIr
- https://d3js.org/
- https://conferences.oreilly.com/strata/strata2013/public/schedule/detail/27169
- https://learning.oreilly.com/
- https://www.researchgate.net/profile/Roozbeh_Haghnazar_Koochaksaraei/publication/320365138_A_New_Visualization_Method_in_Many-Objective_Optimization_with_Chord_Diagram_and_Angular_Mapping/links/5b620a13458515c4b2591bdf/A-New-Visualization-Method-in-Many-Objective-Optimization-with-Chord-Diagram-and-Angular-Mapping.pdf
- https://www.colormatters.com/color-and-design/basic-color-theory
- https://www.oreilly.com/library/view/designing-data-visualizations/9781449314774/ch04.html