

COMP 4462 Lecture Note

Student HONG, Lanxuan
SID 21035307

1 Introduction to Visualization

Vis allows people to analyze data when they don't know exactly what questions they need to ask in advance. – Tamara Munzner

1.1 First glance

The first question for this course is, what is the importance of visualization and why visualization? A quick answer to it is that visualization helps explore the data. It presents a more intuitive, human friendly way to show data, and further helps decision making. This actually requires the correct and elegant presenting of data, which will be covered in the whole course.

Before go into details about how visualization is designed and how different rules should be applied, let's have a quick glance at how visualization involves a plenty of contents. As shown in Fig.1, there are lots of presenting methods including **distribution, correlation, rankings, part and whole, evolution, map, and flow.**

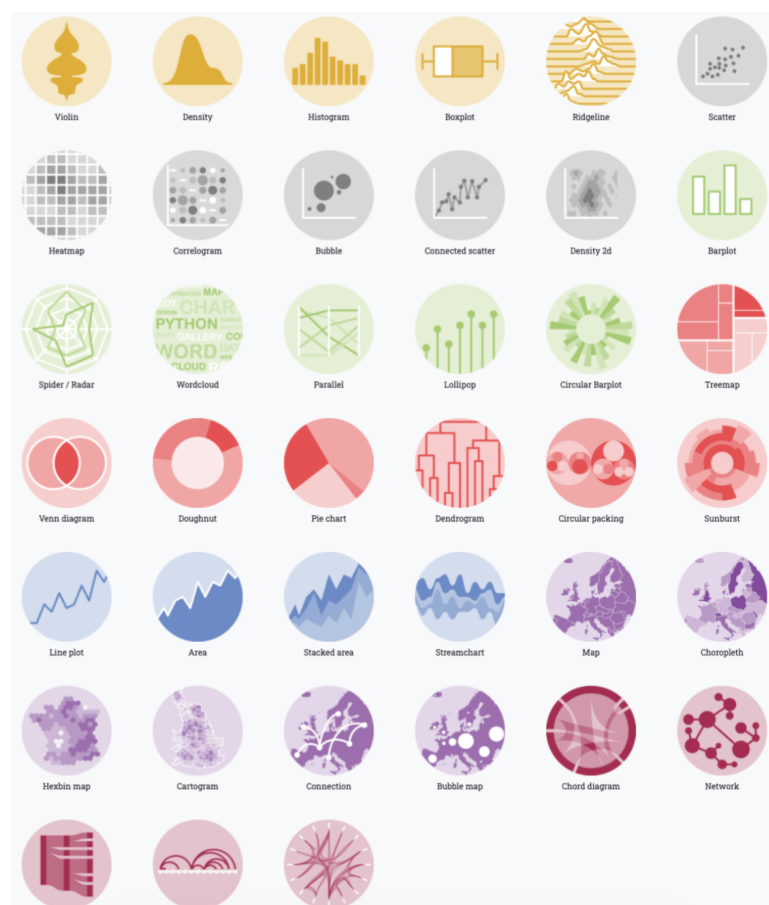


Figure 1: Chart Types

1.2 Data abstraction

The things that we want achieve through data visualization is, to build a model of data to find patterns, show patterns, and analyze patterns. As shown in Fig.2, the four basic dataset types are **tables**, **networks**, **fields**, and **geometry**; other possible collections of items include clusters, sets, and lists. These datasets are made up of different combinations of the five data types: **items**, **attributes**, **links**, **positions**, and **grids**.

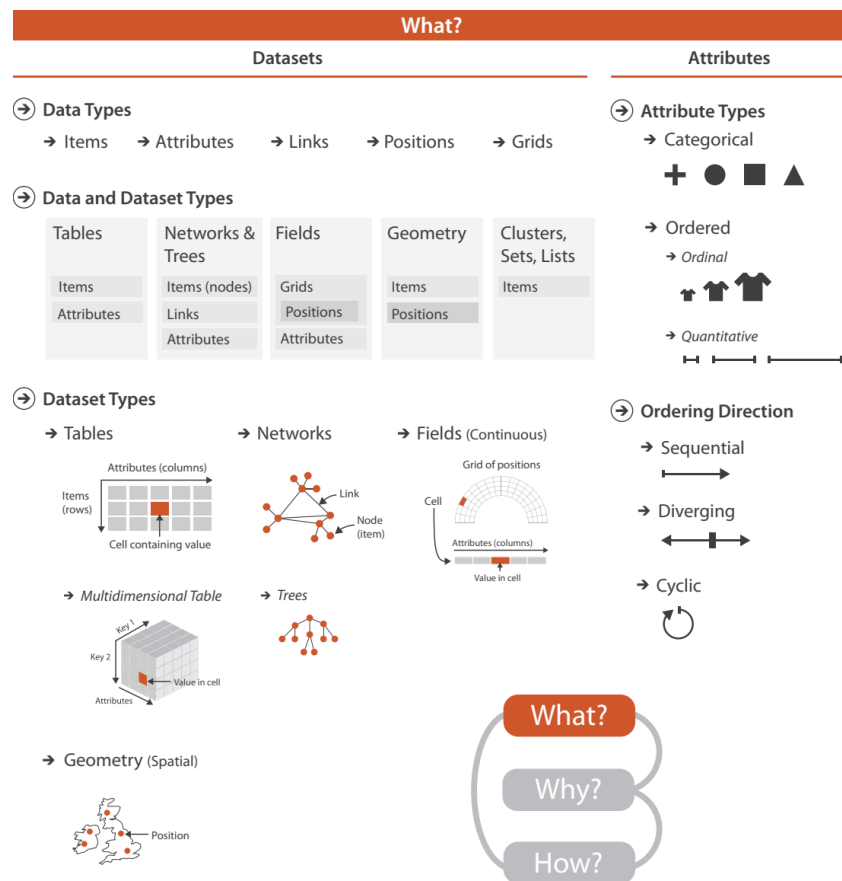


Figure 2: Data Abstraction

Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Figure 3: Data and Datasets

All these methods and figures are to contribute to the presenting of various data types. We can, in terms of how exact the data points are, put the data into four different boxes, namely **nominal**, **ordinal**, **interval** and **ratio**. For example, colors are of nominal type as they can only be categorized¹, rankings of schools are ordinal type, which can be ranked but cannot be evenly spaced, meaning that we cannot quantify the difference between each rankings. IQ scores and temperature of interval type can be evenly spaced, but only matters as a difference, not its absolute value. The ratio type, such as length and Kelvin temperature, has a natural 0 which means nothing.

¹sometimes color can actually be ranked based on transparency, which will be mentioned in the color section

2 Human Perception and Information Processing

When a person looks at a graph, the information is visually decoded by the person's visual system. A graphical method is successful only if the decoding is effective. No matter how clever and how technologically impressive the encoding, it fails if the decoding process fails.

– William Cleveland and Robert McGill

We know that human perceive data, but we are not sure of how we perceive. The study of perception helps a better control of data presentation and harness human perception. There are many different theories of perception. Most define perception as the process of **recognizing, organizing, and interpreting**. In terms of how visual information is processes, the workflow can also be described as from **feature⇒pattern⇒interpretation**.

2.1 The process of visual information

There are two ways (path) to form and process visual information. They are **Bottom-up information process** and **Top-down attentional process**. The bottom-up process drives the pattern building. They become connections in our knowledge bank while the top-down process, start from the previous experience and knowledge, reinforces the relevant information. Humans tend to rely on an initial piece of information (the “anchor”) or exposure to one stimulus (“priming”) to make subsequent judgement or decision.

2.2 Cognition: from pattern to interpretation

The **Gestalt Law** shows how human generate patterns based on knowledge bank.

1. **Proximity**: Things with greater spatial concentration are more likely to be seen as a group.
2. **Similarity**: Things with same color/shape/size are visually grouped to indicate relatedness.
3. **Connectedness**: Connectedness with lines indicates stronger relation.
4. **Enclosure**: When objects are enclosed by lines or placed in a common container, they appear as a collective. **This is the strongest pattern!!!**
5. **Continuity**: Visual entities tend to be smooth and continuous.
6. **Common Fate**: Elements with the same moving direction are perceived as a collective or unit. (in a static presenting, elements don't physically move, but the trend and pattern can also be seen as a moving direction)
7. **Symmetry**: Used to compare time series data.
8. **Closure**: Humans tend to ignore gaps to create familiar shapes and images.
9. **Figure and Ground**: Smaller components tend to be seen as figures against a larger background. Also, human tend to see horizontal distance rather than the vertical distance.

2.3 Illusion: interpreting process

The study of illusion tells us what might cause optical illusions and how do we prevent from delivering wrong messages. There are three main classes of illusions, separated by their cause, namely **physical**, **physiological** and **cognitive**. For each class of illusion, four types of result will be raised. The following table shows some examples of the problems.

Here are some "solutions" or tips in visualization design to prevent from such misunderstanding. Notice that the surrounding context also matters especially in the judgement of shapes and objects, take the letter 'B' and number 13 as example.

2.4 Attention: forming patterns from features

Recall the top-down process and bottom-up process introduced before, some patterns are **pre-attentive**, meaning that they are naturally easier to recognize. These channels include: **shape**, **orientation**, **size**, **color**, **motions(blinking/vibrating)** and **spatial**. Notice that they are not always easy to recognize, especially there are more than one pattern in the presenting. In a rapid visual search, we must make our patterns easy to attend to, with distinct but simple symbols and apply the rules of orthogonality.

Encoding Lessons

- Ordered Variables
 - Prefer encodings that are more easily decodable (accurate)
 - Best to show quantitative variables with position or length
- Favor Separable Encodings
 - Use color (sensible) and other attributes e.g., shape, size, etc.
 - Do not overload symbols (e.g., two at most)
 - Avoid mixing two aspects of color or two aspects of shape
- Small Multiples
 - Reduce the need for multiple encodings
 - But this may make direct comparison more difficult
- Highlighting
 - To draw attention to one group, use a pre-attentive attribute
- Arrangement: put things to compare nearby

Figure 4: Encoding Lessons

2.5 Color

Color serves to highlight, identify, and group elements. It can be **sequential**, **diverging**, **categorical**, **highlight**, or **alert**. The fundamental uses of color in information design is to **label**, **measure**, and to **present reality**. This actually shows that color can be either quantitative or categorical, or just to distinct certain pattern from others.

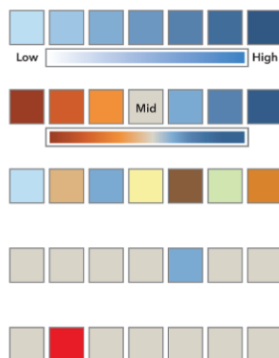


Figure 5: Color

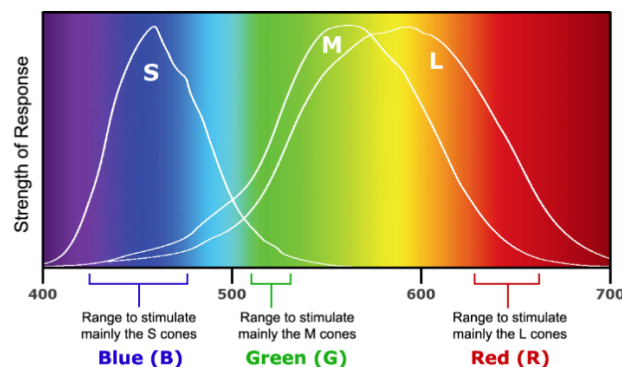


Figure 6: Color Perception

2.5.1 Color Space

Color space is a quantitative model to describe color according to human perception. Three cone types of human can response to **blue**(shortest wave), **green**(medium wave), and **red**(longest wave).(see Fig.6) The most basic **RGB Color System** is describing a color as combination of three primary colors.

This system might lead to negative intensity, where some colors cannot be reproduced by RGB lights. Then we can do a simple transformation to get **XYZ Color System**, where **Y** represents luminance.

Scientist also goes further from the XYZ system to the new **xy Chromaticity Diagram**, which is a famous diagram representing colors. The property of this diagram is that for any line on this diagram, any color on the line can be represented by a certain ratio of the end point.

CIE also create another famous color space other than the xy standard, which is **Lab** color space and **HSV** Color space. For printings other than digital presenting, there are also **CMYK** color space. Check [this article](#) for a general sum up.

2.5.2 Color for Information Visualization

The key idea of all these systems, is that it give human a way to quantitatively describe and index a color. As is mentioned before, the functions of colors are to label, quantify(measure) and to highlight(present). There are 3 common types of color scheme to represent 3 types of data. Other color selection model: transparency, harmony², readability, deficiency. Refer to the *color brewer*.

1. Qualitative Color Scheme for **categorical** data.
2. Sequential Color Scheme for **sequential data**.
3. Diverging Color scheme for **diverging patterns**.

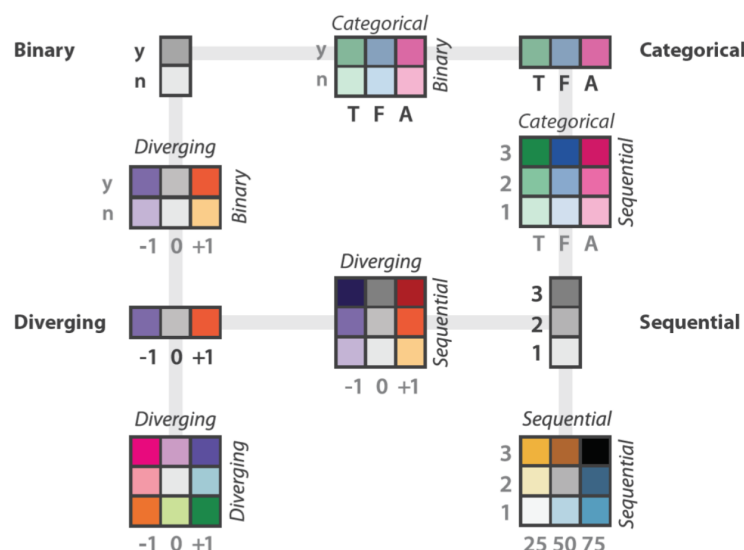


Figure 7: Color Schemes

²Monochromatic, Analogous, Complementary, Split-Complementary, Triadic, Tetradic, etc

3 Visual Design Rules

3.1 Visual Design Principles

Before go into small details and rules, let's look at the big picture of how to achieve graphical excellence. Graphical excellence consists of complex ideas communicated visually with **precision**, **efficiency**, and **clarity**. According to **Tufte** (remember this name!!!), this means:

- Graphical Integrity: Telling the truth
 - lie factor
 - express data in context
 - correct(consistent, standardized, meaningful) scale
- Data-Ink Ratio:
- Chart Junk Avoidance

Lie factor refers to the different ratio of attribute value between visual and data.(see Fig.8) He also suggest not to use unnecessary 2d/3d presenting to avoid lie factor.

$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

$$\text{size of effect} = \frac{|\text{second value} - \text{first value}|}{\text{first value}}$$

Figure 8: Lie Factor

To solve the scale problem, one should not change the scale in the middle/end of the axis or x/y axis. It is possible sometimes to change scale at the beginning to show context or historical data, but it should be very careful for misunderstandings and **DO NOT differ the beginning point**. For some special cases, the axis should be flipped for a more accurate meaning. This also include the **standardization** request, which is, when comparing 2 data with different scale, it's better to normalize them before directly compare the raw data.

Another request over a true visualization is an elegant and intuitive visualization, meaning that it should give the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space. What Tufte suggest, is to *"Maximization of Data-Ink Ratio, with Reasons"*. This can be interpret together with his third principle, which is *avoid chart junk*. But notice that it is not always wise to make things to its simplest version as shown in Fig.9.

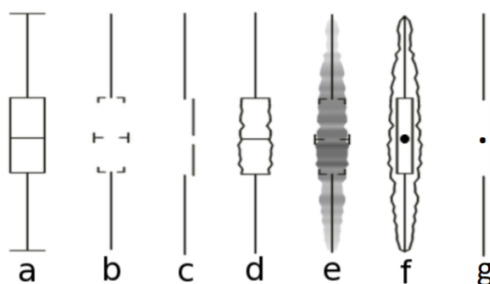


Figure 9: Do not remove everything

3.2 Rules of Thumb

There are 8 rules of thumbs listed as follows:

- No unjustified 3D
- No unjustified 2D
- Eyes beat memory
- Resolution over immersion
- Overview first, zoom and filter, detail on demand
- Responsiveness is required
- Get it right in black and white
- Function first, form next

3.2.1 No Unjustified 3D

3D visualization are effective for spatial data, especially when the task involves shape understanding of inherently 3d structure. In other context, however, while **depth cues** conveyed by occlusion, distortion, shadows and lightening enable us to perceive 2D images as 3D objects, it cost a trade-off in the visual encoding context. Recall that the vis-channel ranking in Fig.12 shows that the **spatial position channels apply only to planar spatial position**, not arbitrary 3D position. It also shows that **depth is a less accurate attribute** than length and area. What's more, the important depth cue **occlusion** will hide information (can be solved at a cost of time and cognitive by interaction of rotation), with the other depth cue **distortion** leading to inaccurate perception. Another drawback of 3D visualization is the dramatic **impaired text legibility**.

To solve the problem, here are some alternative ways or constraints when visualizing 3 channels of data. For example, transform into new data abstraction by cluster hierarchy, or constrained navigation steps through carefully designed view points.

3.2.2 No Unjustified 2D

Compared to 1D list data, 2D data are more intuitive and direct though, also has some drawbacks in certain areas. In terms of showing **maximal amount of information** and **looking up tasks**, list is more efficient then 2D maps and network. If the task is not about spatial identity or connection, then 2D visualization might not be needed.

3.2.3 Eyes Beat Memory

Using our eyes to switch between different views that are visible simultaneously has much lower cognitive load than consulting our memory to compare a current view with what was seen before. In other words, the **internal memory** performs worse than **external cognition**. This imply that for animation design, we need to switch animation in terms if time into small multiples as a overview window so that it can be read off by the perceptual system instead of remembered.

Also, we should be careful of the phenomenon of **change of blindness** that we fail to notice changes if our attention is directed elsewhere when designing animation.

3.2.4 Resolution Beats Immersion

If there are trade-off between resolution and immersion, resolution is usually far more important, except for special cases when 3D data need to be present such as VR and XR.

3.2.5 Visualization Mantra: Overview first, zoom and filter, then details on demand

A vis idiom that provides an **overview** is intended to give the user a broad awareness of the entire information space, i.e. to summarize. Zoom and filter would allow users to go into details into more attributes of data. This is a **top-down** visualization design process applicable for more general data. For storytelling visualization, **bottom-up** method is also useful for presenting local data first and then show the pattern.

3.2.6 Function First, Form Next

Start with focus on functionality instead of start with beauty.

3.2.7 Responsiveness is Required

A less than 1 second response(or **latency**) can be considered a quick and immediate response. A response longer than that might lead to impatience of users. The visualization designer should provide response to the users, for example, confirming that the action has completed by highlight the selection, showing the progress of loading, loading lower resolution or text first, etc.

3.2.8 Get it Right in Black and White

4 Visual Channel

When it comes to visualization design, we should think of how do we combine the data and datasets into a visualization result. The way we encode certain data structures into visual idioms according to data abstraction is through **marks** and **channels**, where marks shows the graphical element, and channel present their appearance based on its attributes. The big picture (Fig.10) shows different kinds of marks and channels.

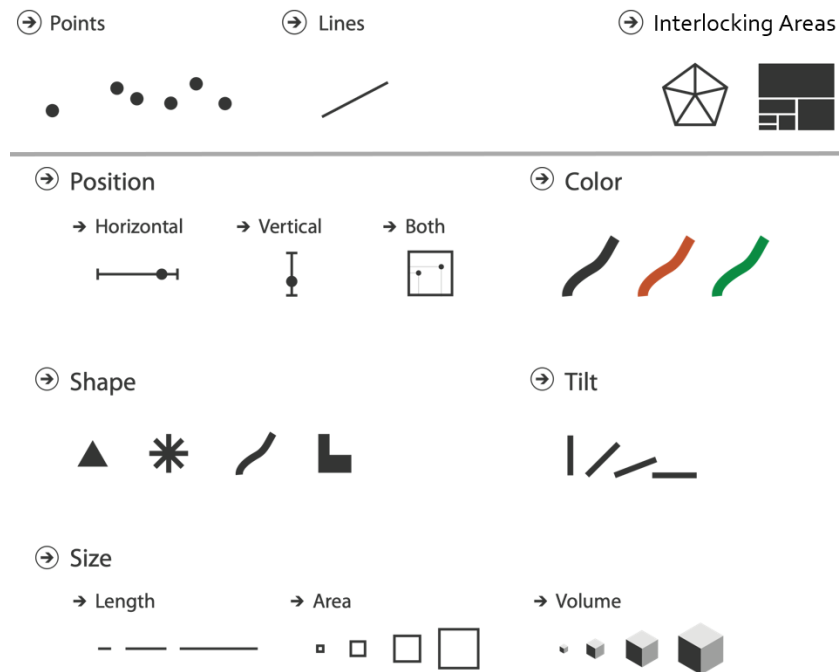


Figure 10: Big Picture of Marks and Channels

4.1 Marks

Marks are geometric primitive objects classified according to the number of spatial dimensions they require.(see Fig.11) **Marks for items** can be 0D(as a dot), 1D(as a line), 2D(as an area), or even 3D(as a volume).

- **point** marks conveys positions. Can be size-, shape-, color-coded.
- **line** marks convey position and length. Can be width-, color-, shape-coded.
- **area** mark are fully constrained. Can be size- and color-coded.

Marks for links can be in form of containment or connection.

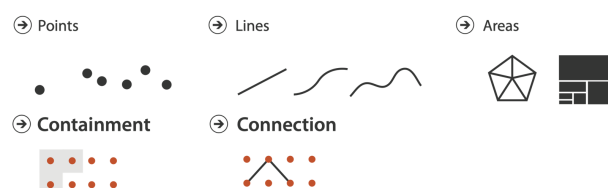


Figure 11: Marks

4.2 Channels

Channels, controlling the appearance of marks, can be interpreted by color, position, shape, angle, size, etc. All these channels can be separate into 2 types based on human perceptual system, naming **identity** channel or **magnitude** channel.

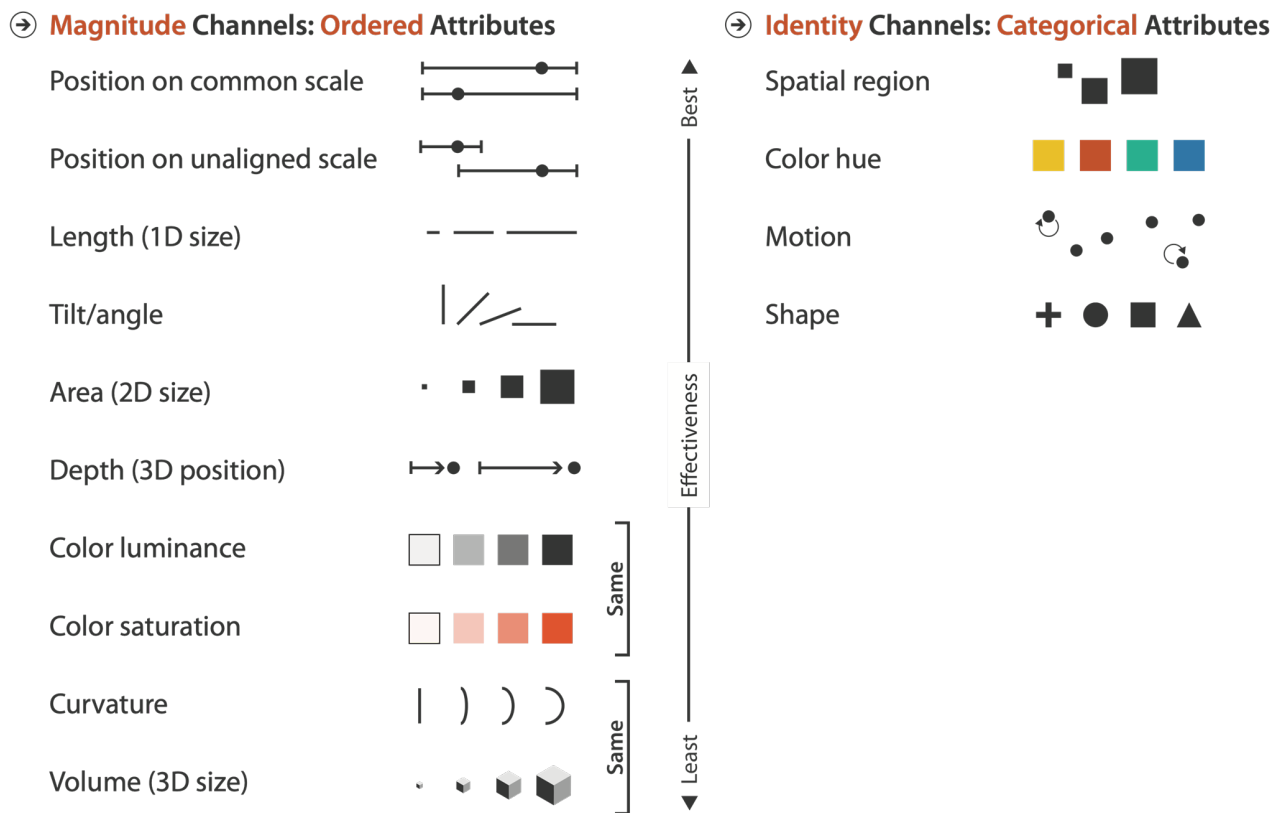


Figure 12: Channels

The following example shows how to add channels on marks. It also shows how different attributes are encoded through single or multi channels. Note that some marks itself have constraints about channels, such as area mark which has size as part of the shape.

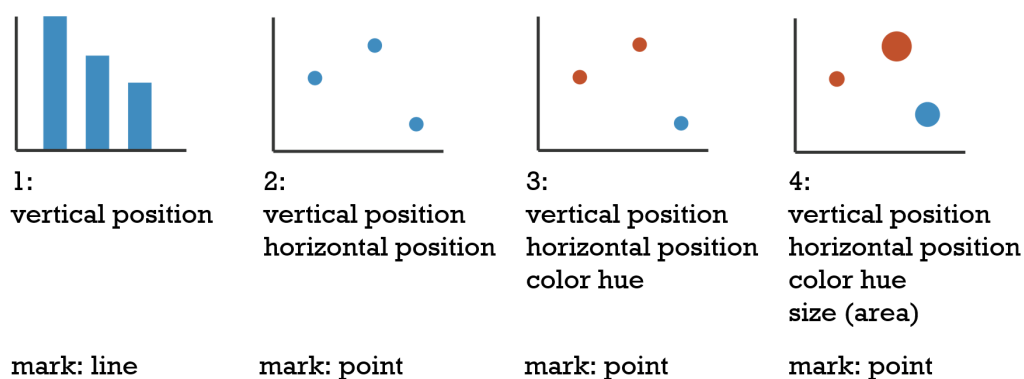


Figure 13: Channel Idiom Analysis Example

The design and analysis of channels should follow the principle of **expressiveness** and **effectiveness**. The expressiveness principle requires that the visual encoding should express in the same way our perceptual system intrinsically senses. Generally identity channels should match for the categorical attributes while magnitude channels should match ordered attributes. The effectiveness principle requires the match of **salience** between attribute and channels.

Channel effectiveness can be analyzed from several aspects. Psycho physicist has shown that (see Fig.14) human present different perception accuracy(and efficiency) to different channels. For example, **length** seem to be the most accurate while brightness is less accurate. This provide ways to assess the accuracy of magnitude estimation for visual encoding.

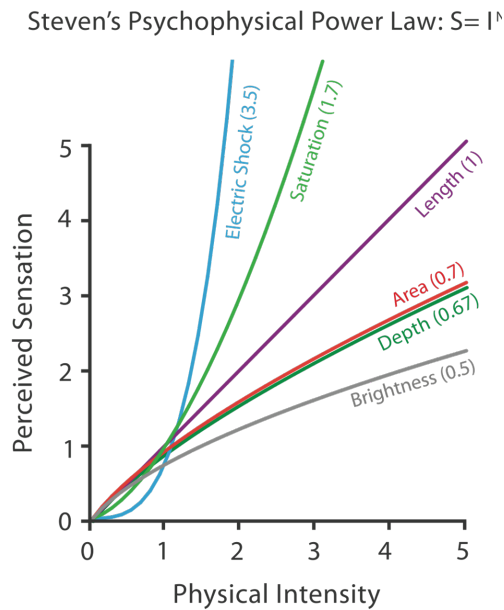


Figure 14: Phychophysical Power Law

Tasks using various graph types can be understood in terms of elementary mental processes.

- anchoring: segment a component to serve as a standard for comparison
- scanning: visual sweep across a distance in a graph
- projection: sent a ray from one point to another
- superimposition: mentally move elements to a new, overlapping location
- detection: detect difference in size of 2 components

5 Visualization Design Idiom

5.1 Task Abstraction

The following Fig.15 shows the big picture of task abstraction. This framework allows designers to consider tasks in abstract form rather than the domain-specific way that users typically think about them. We use a small set to describe why people are using vis, with **verbs** describing **actions** and **nouns** describing **targets**.

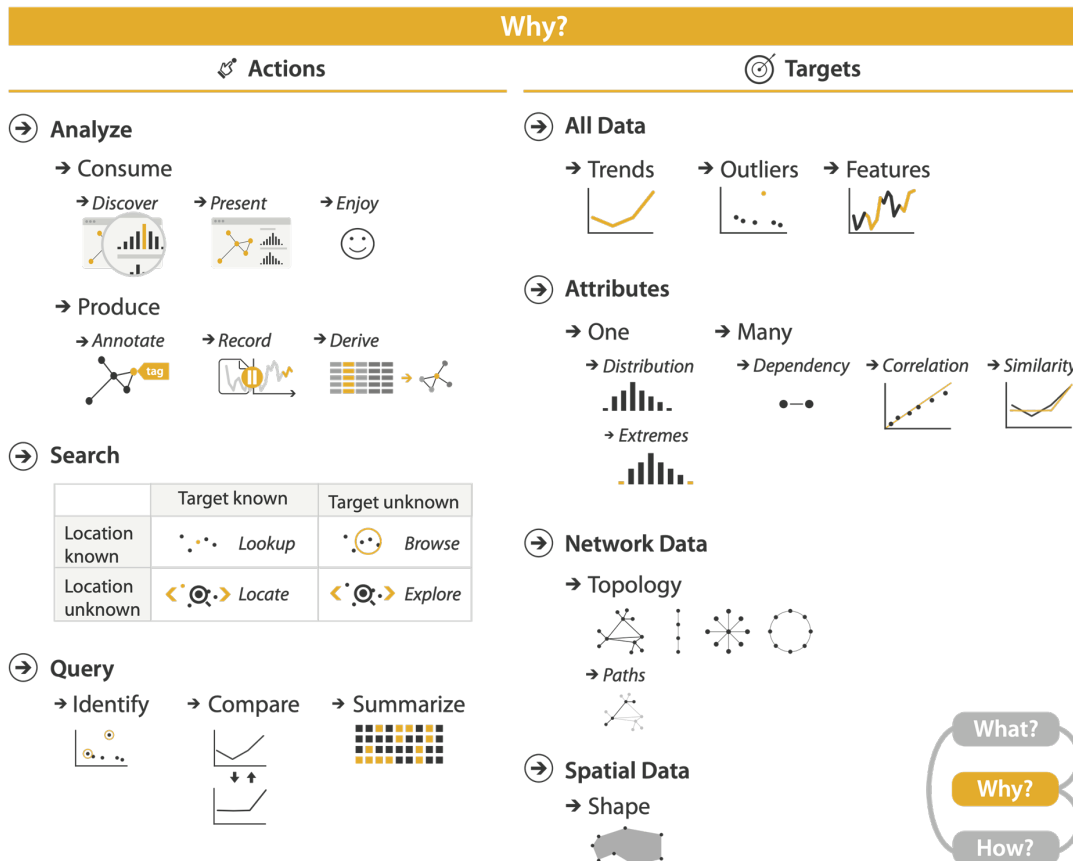


Figure 15: Task Abstraction Big Picture

The highest-level action is to use vis to **analyze**, whether to **consume**(including discover, present and enjoy) or **produce**(including annotate, record and derive) information.

- consume:

1. discover goal is to use vis to find new knowledge that was not previously known.
2. present goal refers to the use of vis for the succinct communication of information for storytelling. It is some information that was understood already but should take place within context of decision making ,planning, forecasting, and instructional processes.
3. enjoy means a casual encounters with vis

- produce:

1. annotate goal means adding annotation to existing visualization as a new attribute.
2. record goal saves vis elements as persistent artifacts.
3. derive is to produce new data elements based on existing data elements

At the middle level, **search**(including lookup, locate, browse and explore) can be classified according to whether the identity and location of targets are known.





	Target known	Target unknown
Location known	 <i>Lookup</i>	 <i>Browse</i>
Location unknown	 <i>Locate</i>	 <i>Explore</i>

Figure 16: Item Action

Once a target for a search has been found, a low level user goal is to **query** these targets through identifying, comparing, or summarizing under different scopes(single, multiple, all).

The highest-level target for all kinds of data are finding **trends, outliers, and features**. For one attribute, the target can be **one** value, focusing on extreme or average, or **many** values, focusing on the dependency, correlation and similarity. The target of network data can be **topology** in general or **paths** in particular, and with spatial data the target can be **shape**.

5.2 Design Idiom

After learning about basic components of task abstraction, it is important to create good visualising idiom design using these abstraction concepts to present user-friendly visualization. There are many choices with mainly 4 classes(see Fig.17).

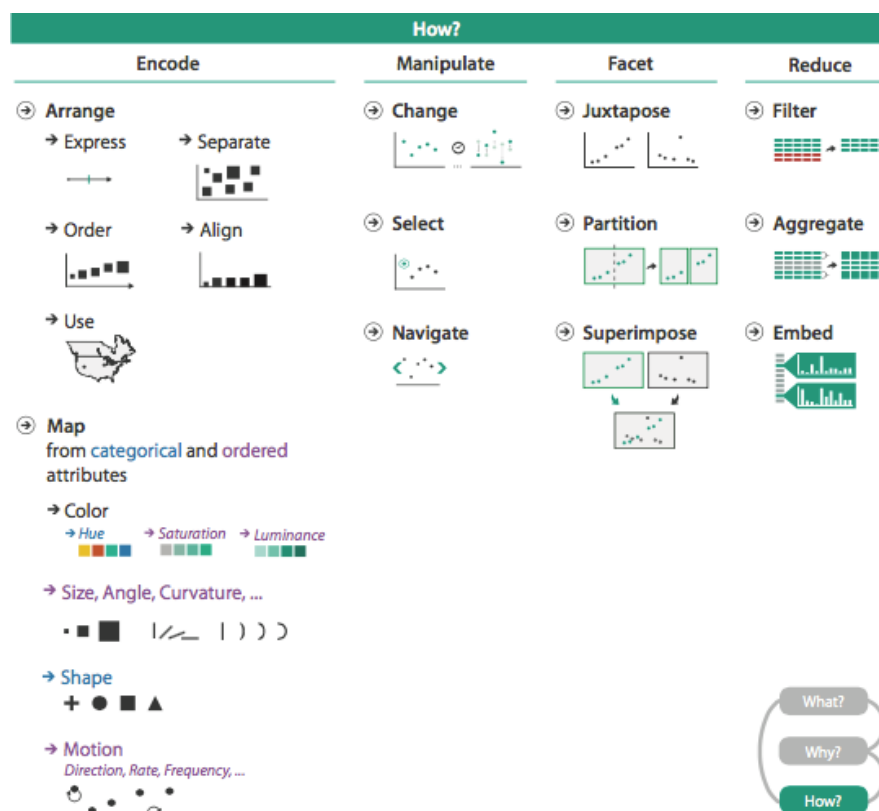


Figure 3.7. How to design vis idioms: encode, manipulate, facet, and reduce.

Figure 17: How to Design Vis Idiom

5.3 Time Series Data Visualization

5.3.1 Abstraction of Time Series Data

Time series data is a set of ordered data values observed at successive points in **time**. Usually we want to simply look into the data, see its trends and features, observe outliers, see correlation, and dependency.

Properties: Granularity, meaning that it can take different unit such as day or hour. It can also be discrete or continuous data according to sampling. Sequential, meaning that it can be uni or bi directional, or even cyclic.

Data Abstraction: with timestamps, time series data should be in **temporal datasets** with quantitative values or categorical types.

Task Abstraction:

- For one attribute in time series: Trends, outliers, and features.
- For multi-attributes at the same time stamp: dependency.
- For multi-attributes in time series: correlation, similarity.

5.3.2 Line Graph

The basic structure of line graph is to have time data encoded by x-axis and important attributes encoded by y-axis(categorical or numerical).

- Marks: points(visible or hidden) and lines.
- Channels: colors, position, and shape.
- Analytical Tasks
 1. find trend: fitting line
 2. find features: emphasize repeated patterns
 3. find extremes: locate peak and valley
- Issues
 - Aspect Ratio
 - Dual-Axis: not always commensurate to use the same unit for different attributes. Also be careful about the mapping of dual axis and each attribute. One smart solution for this problem is by **connected scatter plots**.
- Variations
 - Multiple Line Graph: shared x-y axis and aligned anchor. We can use qualitative color coding and visual cluttering by selecting and highlighting to prevent chaos.
 - Slope graph: use line to connect each attribute(marks) between 2 time stamps to emphasize changes and determine correlation.
 - Connected Scatter Plots: use only attributes in 2d presentation while time data hidden in text channels(see Fig.18). This shows correlation better.

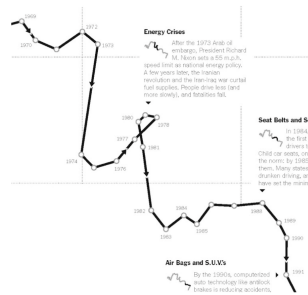


Figure 18: Connected Scatter plots

5.3.3 Area Graph

Similar to line graph with area under line. Channels include width of layer to encode time span, height to encode quantitative attributes and colors. It's easier than the line graph to show the height and relative height, with more chart junks if applied improperly.

- Issues
 - Occlusion: solved by transparency or **Stacked Area Graph**.
 - Chaotic problem in multi area/line charts: solved by partition and juxtapose but with more space. Can also be solved by **Ridgeline Plot**.
- Variation
 - Stack Area Graph: Only lowest layers aligns at the bottom to identify trend in total.
 - Stream Graph: an advanced version of stack area graph, without alignment at the bottom.
 - Ridgeline Plot: partially overlapping and transparency.
 - Horizon Graph: a more advanced version of ridgeline plot.

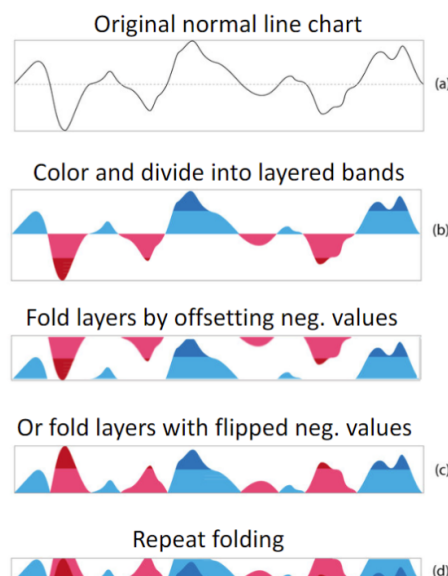


Figure 19: Horizon Graph

5.3.4 Heatmap

The basic structure of heatmap is to have one ordered key attribute, which is time, and some categorical and quantitative attribute which are items at y axis and related cell value at (x,y).

- Marks: area.
- Channels: location(x,y position) and color(quantitative attribute).

5.3.5 Bar Graph

The basic structure of bar graph is to show discrete time data with multiple categories.

- Marks: Line.
- Channels: Horizontal position, Height, Color.
- Analytical Tasks: compare individual values.
- Variation
 - Stacked Bar Chart
 - Box Plot
 - Gantt Chart: use the length of the bar to show duration of time.

5.3.6 Metaphorical in Design

Talk about “When” – Perception of Time

- Basic Timeline
 - Progress bar
- Symbolic Representation of Time
 - Clock
 - Calendar
- Metaphorical Representation of Time
 - Linear: river, travel (map), tree growth, etc.
 - Radial or spiral: tree rings, seashell, hail, galaxy, etc.
- Dynamic Representation of Time
 - Production line
 - Walking

Figure 20: Metaphorical

5.4 Cartography and Geo Spatial Data Visualization

5.4.1 Cartography

cartography is the science of making maps. There are generally 2 types of maps: reference map for natural features and thematic map for other data mapping to locations. Vis designer cares about the design of **thematic map**.

Analytical tasks: spatial distribution, interplay of multiple factors and spatial changes over time.

There are different projection methods with different goals to whether preserving the shape, distance, direction or area.

- Azimuthal: preserve direction and distance from center
- Equal-Area: preserve area
- Conformal: preserve shape and local angle

Different scales are applied with whether details or general view to be presented.

5.4.2 Choropleth Map

Use the reference map as basic, area as marks, with colors of different area showing categorical and numerical attributes.

There are many variations from the basic choropleth map.

- Application of color hue: bi-variate and multi-dimension choropleth map.

One advantage of choropleth map is that it preserve geographical accuracy. It has a trade-off with less focus on the actual influence(importance) from the attributes. Symbol map solve this problem.

5.4.3 Symbol Map

Instead of using area to encode data, which might lead to misunderstanding that the size of area represent the quantity of data, symbol map use a different symbol to present attributes. Here are some basic symbols commonly used.

- Dot Map: position for spatial information, color for category, density for quantity(choose a proper unit to avoid visual occlusion).
- Bubble Map: position for spatial information, color hue for category, sequential colormap or size for quantity.
- Spatial heatmap: an advanced version of bubble map to see the concentration.
- Isoline Map: an advanced version of spatial heatmap, turning sequential data into several categories to simplify color design.
- Other symbol maps: add more channels to the basic map.

5.4.4 Cartogram

Data no longer bounded by the original shape of the map, only position remains for spatial information.

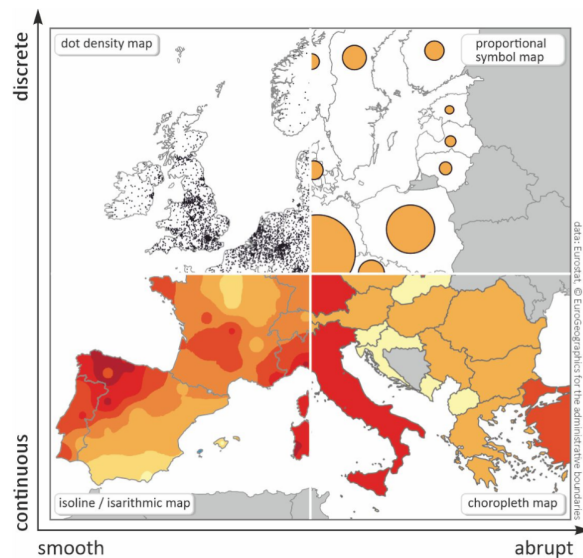


Figure 21: Symbol Maps

6 Visualization of Multivariate Data

Multivariate data to be visualized are of high dimensionality. We use different **dimensions** to present independent data and key attributes. The following are some techniques to visualize multidimensional multivariate data.

6.1 Geometric Projection

Geometric Projection is to map attributes to axes or arbitrary space into spatial data. Geometric projection can be **rectilinear, parallel, or radial**. It is easy for users to detect outliers and correlations. The downside is that all attributes might be considered similarly while not in real world.

1. Scatter Plot for a pair of attributes

- Mark: point + (regression line/area)
- Channel: x-y axis position for qualitative key attributes, color/size for clusters
- Tasks: Identify the trend, outliers, distribution, correlation, clusters.
- Problem: consider using isoline map to reduce visual cluttering but will lose individual information.

2. Scatter Plot Matrix for multi-pairs

3. Radical Coordinates(usually for navigation or time)

4. Star Coordinates

5. Parallel Coordinates

- Mark: line(jagged line between parallel axis)
- Channel: position, color, size
- Task: summarize, explore, clustering, correlation

- Problem: Axis ordering and visual clutter problem. Several solutions: hierarchical parallel coordinate.

6. Circular Parallel Coordinates

7. Radar Chart

6.2 Layout Density

It's the idea to maximize data-int ratio.

1. Space-filling Layout

- (Cluster)Heatmap
- Table Lens: Alignment

2. Dense Layout

6.3 Hierarchical Display

7 Text Data Visualization

8 Validation and Evaluation

There are basically 4 types of validation problem: wrong problem, wrong abstraction, wrong idiom, and wrong algorithm. For each problem, there are some validation approaches.