Visualization of the information

Carlos Arbonés and Juan P. Zaldivar

GCED, UPC.

Lecture Notes.

Contents

1 Introduction to Visualization

We have to pay attention **users**, in what are their needs, background, work environment, etc. Also focus on the **data**, for example the scale (quantitative vs qualitative), type (1-dimensional, 2D, etc), the number of variables, etc. Finally we have to know **what are the tasks**.

In visualization is crucial accomplishing the following:

1. *Expressiveness:* show exactly **the information in the data**, no more no less.
2. *Effectiveness:* take into account the **cognitive capabilities** of the human visual system, the message has to be easy to get.
3. *Appropriateness:* cost-value ratio that asseses the benefit of the visualization, mainly **time (computation)** and **space (screen-space)** efficiency. There is no need for using the whole screen for plotting extremely simple data for example.

To carry out tasks more **effectively** we need a match between data/task and representation, there are a lot of possible representation and many are ineffective. The chance of finding good solution increase if we understand the full space of possibilities. Representation must be **novel**, enable entirely new kinds of analysis, and **faster**, speed up existing workflows. To validate effectiveness there are many methods and we have to pick the appropriate one for our context.

Some inappropriate practices in data visualization are:

- To have some data set and we **throw it to any chart type**
- Get some **random data and create "visualization"** from it
- Encoding **too much or irrelevant information**
- Using **unsuitable palettes**: we should to able to interpret the visualization without the palette legend
- Using **unreadable text**
- Use **lots of axis**, difficult to easily interpret
- **Correlation does not imply causation**
- Use **space** prudently
- Be careful with **clutter**
- Use **3D** with caution
- Be careful with **scales**
- Use **color** wisely
- Use **standard axis**

1.1 Visualization: The Basics

Computer-based visualization systems provide **visual** representations **of data sets** designed to **help people** carry out tasks more **effectively**. Augment the capabilities of the human rather than replacing it by computation decision making.

Visualization is related to **understanding the underlying data** by helping the user to understand data using their excellent perception capabilities. It **helps the user to carry out tasks more effectively**. If the result is a calculation, we should probably not be using visualization at all. If we know what are we looking for then we do not need it.

We need to be careful when representing datasets. **Summaries lose information** and **details matter**. Famous example: *Anscombe's quartet*. We do not know how the data is distributed, although they might have the same summaries information.

1.1.1 Main applications of visualization

The main applications of visualizations are:

- *Explanatory:* present the results. Visualization is used for **presentation**. To communicate data and ideas, explain and inform providing evidence and influence and persuade. **Commonly only showing a few variables of the data.**
- *Analysis:* Analyse hypothesis. The typical objectives are **showing many variables**, illustrate overview and detail to **facilitate comparison**. Presentation might choose some parts, **analysis will focus on all of them**.
- *Exploratory:* Inspect the data to learn new things, get **insights.**

Visualization is very useful in exploratory data analysis when **we don't know what we're looking for**, **we don't have a priori questions** and **we want to know what questions to ask**.

In conclusion, **exploration** is used for gathering knowledge on the data when nothing is known, **analysis** is used for verification or falsification of hypothesis and **presentation** for communicating the results.

1.2 General Rules

One has to **be honest** with the audience and **check if the data is correct, updated**, etc. Also how it was collected and if it is reliable.

Explain the use of the colors and the legends. If encoding (of colors and shapes) are not standard there has to be a reason. The colors have to take into account the physical conditions of the audience and their background knowledge about the domain.

We need to **make things memorable**, give a certain story when showing the visualizations. The display section is also important, it is not the same to present the visualization in a screen of a phone than a billboard.

1.3 Data, Tasks, Users

1.3.1 Data Types

- **Nominal**: There is no need for a order set of the names/observations.
- **Ordinal**: There has to be an specific order in the observations.
- **Quantitative**: measured data.

We can make nominal data ordered with the use of transformations. We can make certain transformations but this is something artificial. The data can also be analysed by its structure.

1.3.2 Data Structure

Structure	Examples
1-dimensional	Alphabetic lists, source code, texts/documents
2-dimensional	Planar or map data, photos
3-dimensional	Molecules, human body, buildings
Temporal	{start, finish}, e.g., medical records, project management, historical presentations
Multi-dimensional	N attributes -> points in n-dimensional space, e.g., relational databases
Tree	Hierarchies or tree structures, e.g., file directories, business organizations
Network	Connected as graphs, e.g., communications networks, social networks

1.3.3 Tasks

Tasks can be classified as:

- **Overview**: gain a overall knowledge of the data.
- **Zoom**: we can concentrate into a small region in out data. We need to provide this action to people with **interactions**. We can perform zooming in all dimensions. One dimension at a time by moving bar controls or similar functionalities.
- **Filter**: similar to zoom. Focus on some elements, deleting uninteresting elements. What we do is by removing using sliders.
- **Details-on-demand**: we are not going to show all the data we have. At certain points we need to give the users the opportunity to give details, so they click the info they want to see. Select a set of items and get the details when needed.

- **Relate**: interesting depending on the problem to solve. Being able to relate items from one view to another. Or look for similar items in one view. F.e. "Same countries having...".
- **History**: Important to track history. Go back to certain steps of the actions of manipulation of data.
- **Extract**: allow to extract relevant information according to the user.

The first 4 are the most important and every visualization must have it.

1.3.4 Users

- **Limitations**: Background, visualizations limitations, etc.
- **Computational Capacity**: limited resources. We need to update the visualizations due to interactions.
- **Human Capacity**: Memory and attention are finite resources so we are vulnerable to large changes. Also the length of the presentation.
- **Display capacity**: Not enough space for big visualization. People are not going to use infinite space. Scrolling million of pixels is not doable. Maximizing the amount of info in the space available without overwhelming the user. (Ink-ratio)

1.4 Visualization as a Design Process

Problem characterisation: We have some sort of data that can be managed in Ink-ratio. For the same data a lot of possible visualizations. Need to find what the **user wants to solve**. We need to choose some domain experts to help us understand the needs. Sometimes the tasks can not be verbalized, which leads to an **iterative creation of the visualization**, while including more information of the target/task.

Data Abstraction: Transform the data into a more abstract way (a more generic representation). F.e example gender difference among countries, we can transform the data to adjust to explain that gender difference.

Technique and algorithm design: Decide how we let the user select and interact with the visualization.

Validation: we can try to validate if our visualizations are good. If the users can benefit of the visualizations, if the data types are useful to encode the information. We can measure if it is efficient or not. Some of the validation methods can only be applied after the visualization is complete.

2 Good Practices in Visualization

2.1 Effective Visualizations

The main idea is to **communicate a message**. We have a lot of data and we want the reader to get some knowledge from this data. We need that the message is transmitted and the people understand the message. **A visualization is not effective if is too complex or is misleading** (in the sense that the data is not understood).

- **Data density**: not too many elements, this could be confusing. Important that we use the minimum amount of data needed to transmit the message.
- **Visual mappings**: if we have a lot of elements the user will need too much time to process it. As much simple as we can.
- **Amount of information**: keys, labels, etc. helps understanding the data.
- **Color usage**: influences what we can understand and what we can see form a map. If we choose bad colors is very difficult to understand data. We can also use it to guide the user, to mark important things.

To be sure if we are doing a good visualization we need to analyse it and see if the message is understood. We will need to analyse the data and consider if this data contains some important parts that we need to highlight and if the user gets it.

The **general principle** we need to follow is that we need to strive to give our viewer the greatest number of useful ideas in the shortest time with the least ink (simple visualizations).

- **Create visuals when necessary**, when we need to transmit and explain something.
- **Know the message in advanced** to plainly the visualization that satisfies the audience. Need to adapt it to who will be looking at it.
- **Respect visual capacities of humans**.

We will need several attempts to achieve a good visualization. If we want to communicate contents we need a good design, is an **iterative process**. Refine the visualization until we get what we want. We need to consider **legibility**, the idea is that it should be readable in the screen we are going to display it. Our visual ability is not infinite.

2.2 Use of color

We can use color to:

- **Distinguish**: we will use palettes that allows us to represent categorical data, use colors different to the other but not have any other implications.
- **To encode values**: to encode quantitative data. We can use sequential palettes and diverging palettes.
- **To highlight** elements/values. F.e all elements in grey except the element we need to highlight.

2.2.1 Tips for Color Selection

- Use color **only when needed** to serve a particular communication goal.
- Select **suitable color palettes**.
- Non-data components should be displayed just visibly enough to perform their role (e.g., light grey).
- Avoid using a combination of red and green

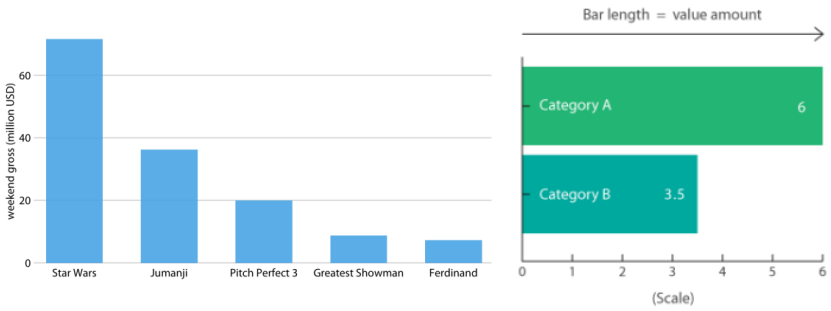
3 Visualization techniques

3.1 Display quantities

3.1.1 Bar charts

Are used to **compare/lookup (really easy)**. Can scale to hundreds of elements.

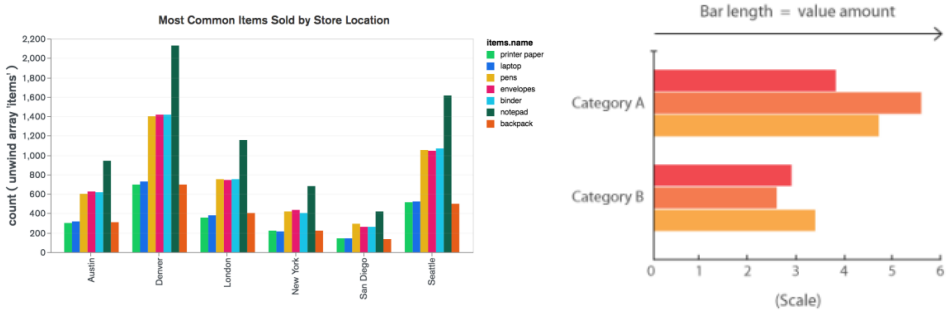
- They always **must start at 0**. If not, the proportion of the quantities is lost, causing misunderstanding to the user.
- **Labels easy to read**. Think of the orientation of labels (horizontal if possible). Labels of the bars should not be too long.
- **Order based on data or labels**. Alphabetical order if we want to make label search easier. By quantity if we want to facilitate value search
- **Neutral colors** better, something related to gray or just gray. Other colors to emphasize.
- **Grid lines** needed if we are looking to have a **precision**.
- If data is **ordered in time**, better use **line chart**.
- **Don't use hundreds of bars**



3.1.2 Paired bar charts

Usualy used to compare. Easy to identify specific data in the same category, but not between different categories. **Length** is used to express quantity. **Color** is used to separate values in each category. **Spatial regions** separate categories.

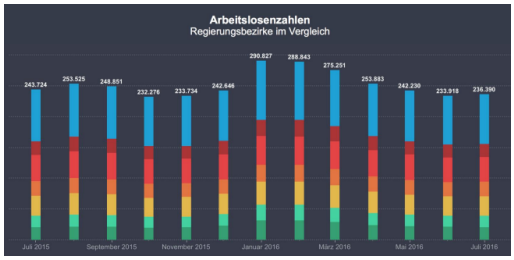
- They always must **start at 0**.
- **Bar chart guidelines** apply.
- **Don't use them** if one category is **time**.

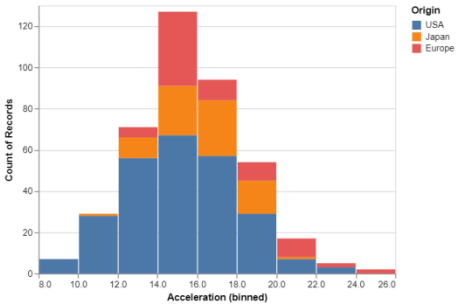


3.1.3 Stacked bar chart

Contains the same information as *Paired bat charts*, but now bars are stacked vertically. More **difficult to compare different groups**, also **difficult** within the same category. The **total quantity of the stacked bar has to make sense**, not only the divisions.

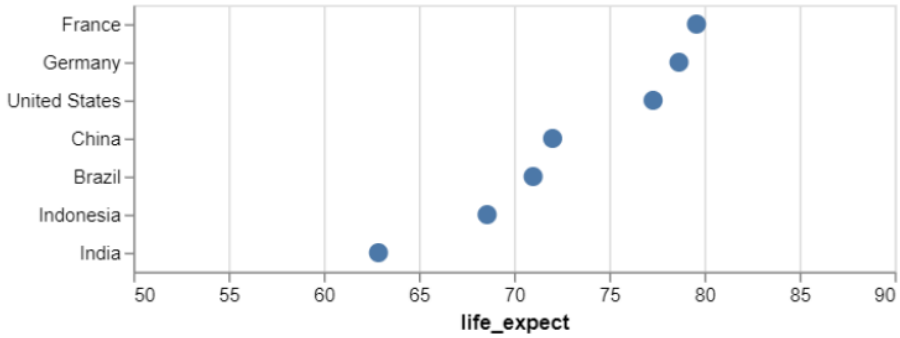
- **Start at 0**.
- Same **guidelines** as **bar charts**.
- Difficult to compare between groups
- Difficult to compare within groups
- Don't use when total quantity does not make sense
- Use few categories





3.1.4 Dot plot

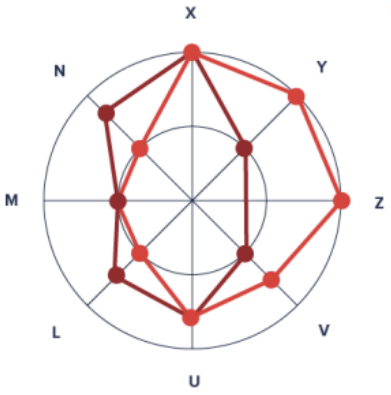
- **Don't need** to start at **0**
- Must be **ordered by quantity**
- Suitable when **small differences** must be shown
- If values are relevant, **label axes suitably**

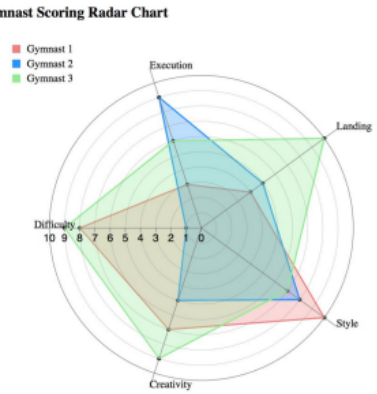


3.1.5 Radar chart

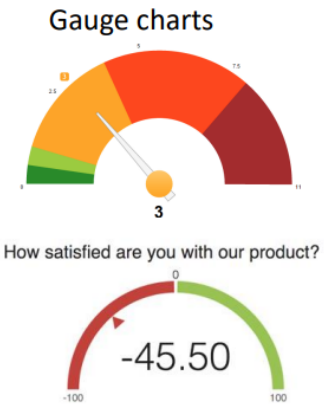
Instead of bars, it shows the data in a radar way. It is **analogous to paired/grouped column charts**. Easy to compare the values in the category.

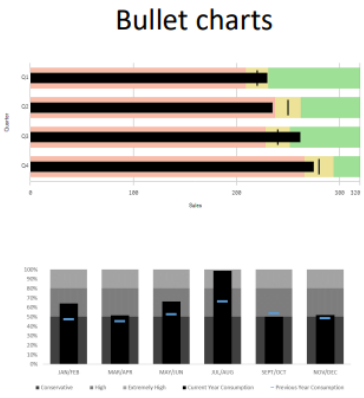
- Multiple dimensions
- Space efficient
- Different designs (points, area ...)
- Do not scale very well
- Can be small





3.1.6 Gauge & Bullet chart





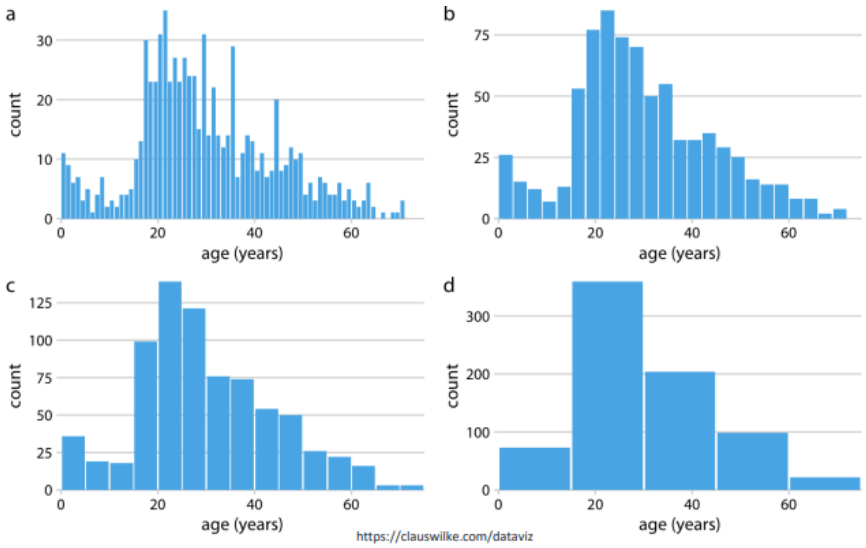
- Adaptation of **real gauges**
- Very common in **business analytics** (used to display KPIs)
- Current value (front) vs reference (background)
- Using **angle** to encode values (less optimal)
- Use too much space
- Commonly include the data in text too

- Version of gauge charts using bars
- using the background of the bar chart to encode reference value(s)
- Space efficient (may encode multiple values in the same space)
- Better for perception (comparing lengths instead of angles)

3.2 Display distributions

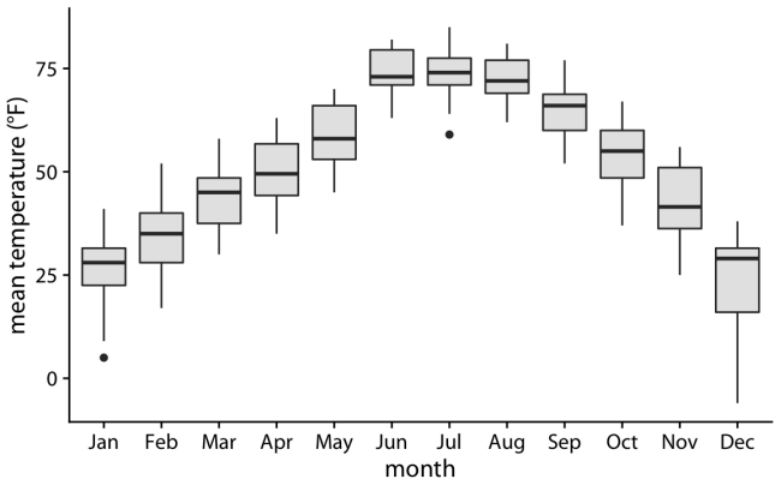
3.2.1 Histograms

Not focused on the values itself but in the **distribution/trend** of the whole set of values. Complicated but important to **choose the number of bars** for the distribution (might complicate the interpretation and the visualization of the distribution).



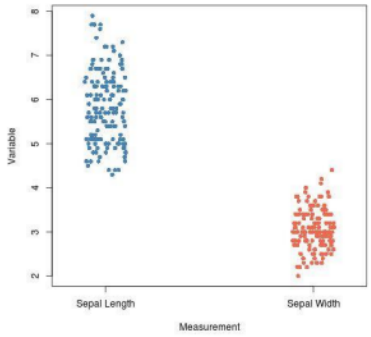
3.2.2 Boxplot

- Useful for **several distributions** at the same time
- Gives insights on data distribution (median, minimum, maximum, outliers...)
- Box plots **hide/abstract** too much data
- Hidden information may be relevant
- Can use **alternative charts (violin, streep...)** to show the internal distribution



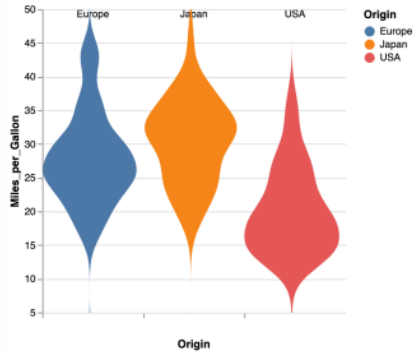
3.2.3 Strip chart

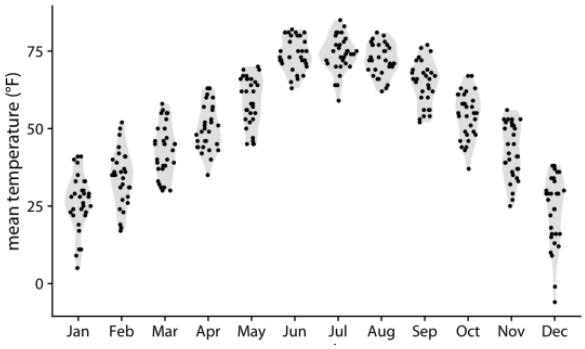
- Shows all the data points (revealing the distribution)
- Difficult to interpret
- Use random positioning in one axis to avoid operlapping.



3.2.4 Violin plot

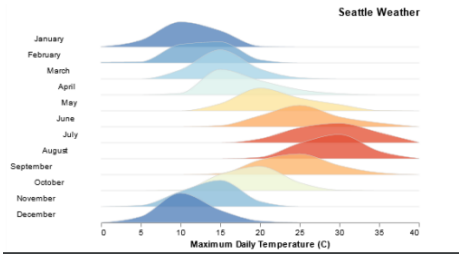
- An accumulation in the horizontal axis to illustrate the distribution (density chart)
- Reflected for aesthetic purposes.
- We lose the statistical properties
- Need to calculate the shape
- May still hide some data





3.2.5 Ridge plot

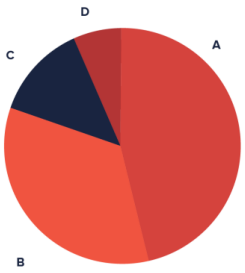
- Like a half violin plot in horitzontal
- Allows more data
- Allows overlapping if done carefully (can be a problem)
- No accurate value estimation possible



3.3 Display proportion

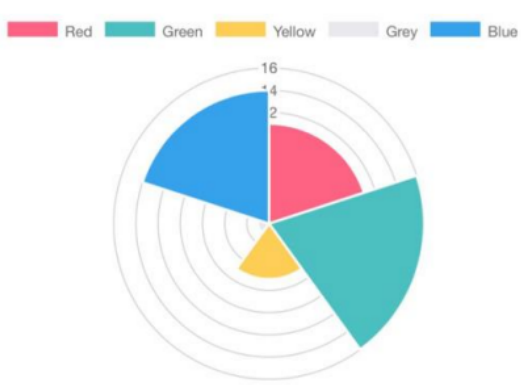
3.3.1 Pie chart

- They are proportions, should add to 100%
- **Angle** is used to display **quantity**
- Use **few categories**
- Start at 12:00 and sort in descending order (clockwise)
- Similar values will be **difficult to appreciate** visually
- Very influenced by the color palette chosen
- Too much space for the low information shown
- Certain key proportions ($1/4^{th}$, half) may be easier to read
- Diffcult to get them well
- The community **hates** them



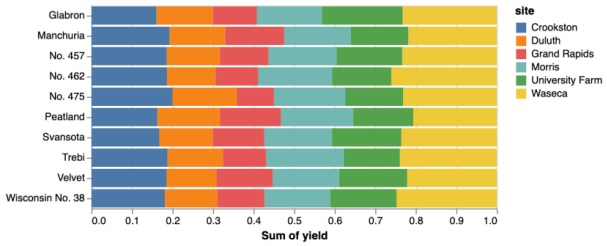
Polar area chart

We encode the value with the area of each of the sector. They still occupy a lot of space but it is more easy to compare/estimate the values and distribution. Allows to stack categories but that makes everything harder.



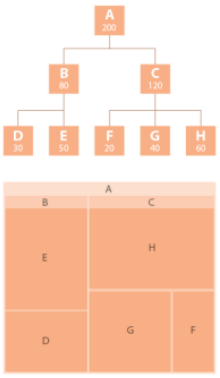
3.3.2 Normalized stack bar

Easy to compare the different categories. When many categories, the ones that are not adjacent are difficult to compare.



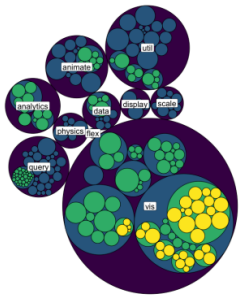
3.3.3 Tree maps (enclosure diagrams)

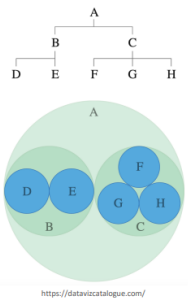
Areas are much more difficult to compare than bars. But we can use them to show hierarchycal data.



3.3.4 Circle packing

Allows to show hierarchycal data with circles rather than rectangles. Same problems as Tree maps.

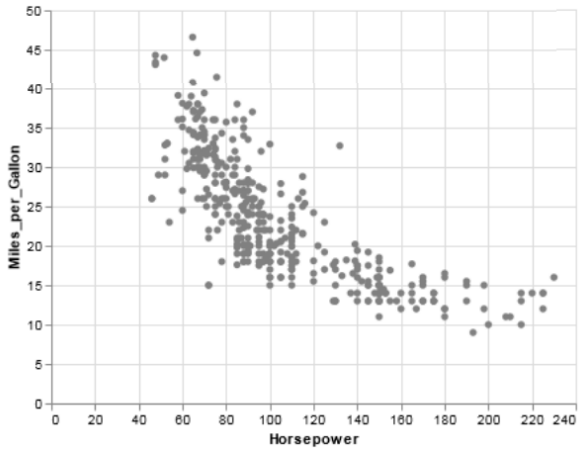




3.4 Display Relationships

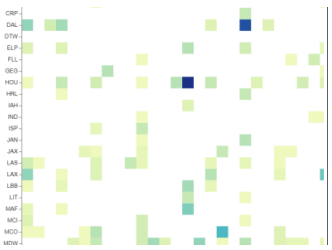
3.4.1 Scatterplots

- Typically represents data without keys. We have two quantitative values represented in points encoded by position.
- Useful to find correlations or outilers. If there are many points, may be difficult to understand when the clutter of points is dense.



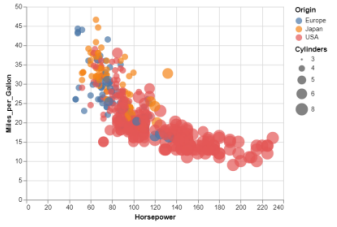
3.4.2 Heat Maps

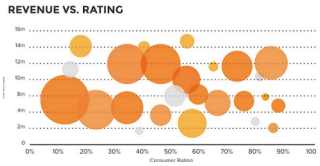
- It is typically represented with an array of rectangles.
- They represent 2 categorical variables and we encode a 3rd quantitative value using color.
- The marks is the area located in a matrix encoded with the 2 categorical attributes.
- For finding outliers and clusters. Categorical variables have no order, so x and y can be ordered as we wish. *Reordered matrix*, with the objective of reordering to find clusters. (Not possible when a category is time).
- Commonly used in bio to encode gen expressions and so on.
- Color palette should not be continuous but discrete to notice the difference between values.



3.4.3 Bubble Charts

- Enhance/Increase the number of variables of scatter plot.
- Cluttering problem will appear much more earlier than a normal scatter plot. I can have overlapping depending on the distribution of the data. Common technique is not defining the points with `opacity = 1`, by making them semi-transparent to see if points overlap.
- Don't use ordered colors when representing categories like in Figure on the right.



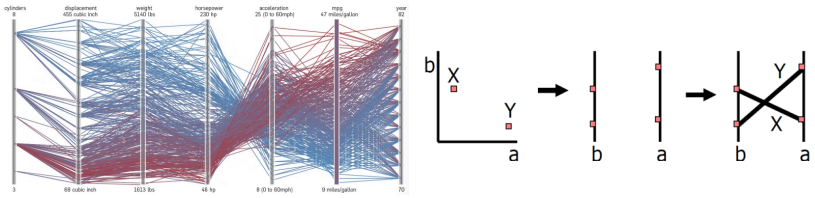


3.4.4 Scatterplot Matrices

- Combine different scatterplots. Provide the tool to analyse pair-wise relationships. Find relationships/correlations.
- Normally the same plot is repeated with a change of axis. Waste of space, waste of half of the matrix.
- Visualization is small due to the dimensions.
- Let the user some interaction tasks; *zoom out, ...*

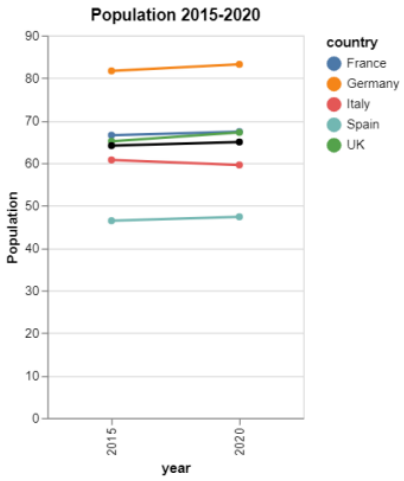
3.4.5 Parallel coordinate plots

- Show multiple variables. Let you see whether variables are correlated.
- Keys can be quantitative or categorical. Axis are scaled to the min/max values.
- Same behaviour for similar correlations. If there is a positive correlation the lines will be parallel and if negative, the lines will cross.
- Not easy to see correlation of not adjacent dimensions. We need to provide interactive tools so users can change the order of the dimensions, highlighting lines, etc.
- Not hundreds of elements. Can represent a large amount of points using special techniques like transparency.
- Scale really well



3.4.6 Slope charts

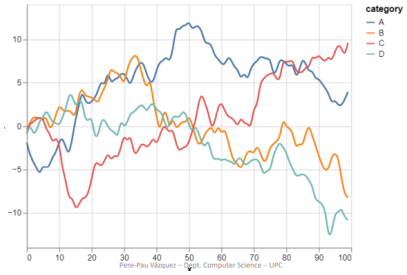
- Often encode two values. Normally of time instances.
- Intended to show increase/decrease of 2 data points along time.
- Really simple but useful. Lets us compare really quick.



3.5 Display Time Series

3.5.1 Line Charts

- The most common representation.
- One key-one value. Data is quantitative.
- Marks are points. Encode the quantity by the position of the points.
- Lines used to show trends.
- Different from bar charts. F.e. lines can go into negative values, meaning that the axis does not have to start at 0.
- Do not scale very well. More than 10-12 lines will have several issues such us color palette, overlapping ...

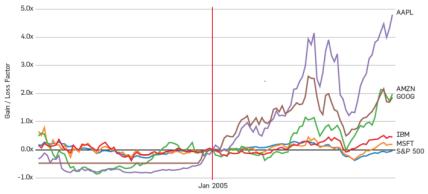


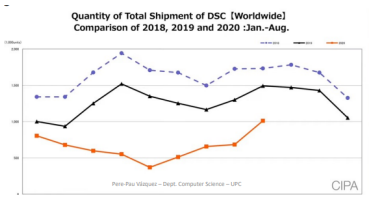
3.5.2 Waterfall Chart

- Used in very specific scenarios. Mostly in business to represent cash flows (money entering and being spent).
- Don't start at 0. Start at the end of the value of the previous one.
- At the end we have another reference bar to encode the final value.
- Usually, use of two colors (green/red). Optionally a third color to the reference bar.

3.5.3 Index Chart

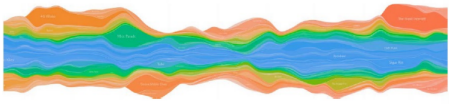
- Solve the problem of time data not being able to be compared evolution-wise. Instead of using as absolute values the values we want to encode, we use a reference.
- The value is the *now value - reference value*. Lets see the evolution. Positive values show an increase and negative decrease.
- Can be indexed in different ways, like in time (see Figure in right). We show different years in the same chart in order to compare easily how the different years have evolved.
- Used when we need to compare values in time fairly by setting starting points.

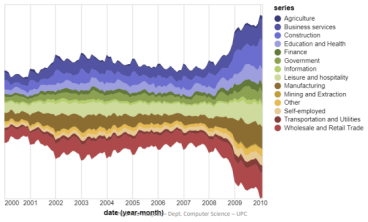




3.5.4 StreamGraph

- Like stacked bars, with a very thin width. Bars can start at both directions of the axis.
- Grasps of trends. Identify the most important aspect and growth.
- Provide zooming, lenses, blah, blah...
- Try to communicate how things have evolved in time. We typically need to recalculate how the geometry is laid out because we want it to be smooth.
- Can have a lot of keys. They scale much better than stacked bars (altough stacked bars do not scale well) .
- Does not support negative values.
- Can not be used with information that can not be added such as temperature.
- Since elements are not aligned is difficult to interpret values and estimate trends.

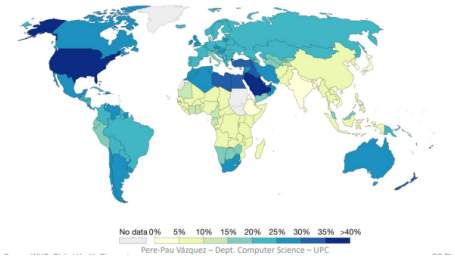


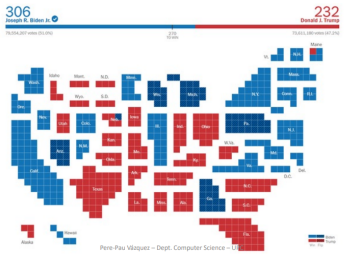


3.6 Display Geospatial Data

3.6.1 Choropleth Maps

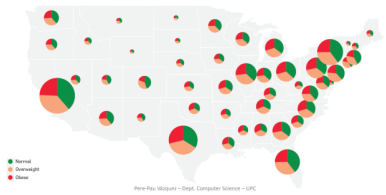
- We use maps when the geographic information is relevant.
- Problems:
  - The bigger the size of the area, the more it attracts our attention.
  - The position (things on top are thought to be more important than things on the bottom.
  - Ignoring the density, we usually have to normalize.
- Sequential discrete color palette.





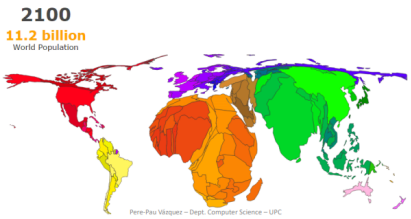
3.6.2 Graduated Symbol Maps

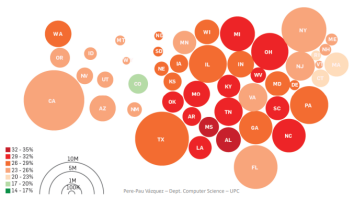
- Use a pie chart that encodes 4 variables. Size is population and the 3 sectors represent the type/proportion of obesity.
- Avoid the problem of the geography interfering with the interpretation of the data.



3.6.3 Cartograms

- Modifying geometry is complex. Depending on the distortion, the countries can loose its shape and identification.



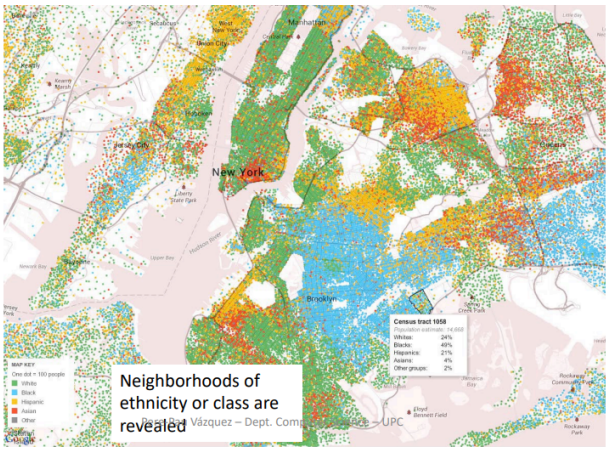


3.6.4 Dot maps

To represent data from a census, dot maps are a common way to visualize geospatial data, where each point on the map represents an observation or a set of data related to a specific geographical location.

Issues:

- If the **size of the symbol is used to represent a quantitative** parameter, **scaling may present perception issues**. Variations in symbol size can make it difficult to make precise comparisons between different locations on the map.
- Perception of size also depends on the local environment of the points on the map. **Points close to each other may give the impression of being a group or a single entity**, which can be misleading in terms of the actual data distribution.
- **If color is used to represent a quantitative parameter, issues related to color perception may also arise**. People's ability to distinguish and comprehend differences in colors varies, which can lead to incorrect interpretations of the data.
- **When working with large datasets, there may be issues of overlap or overplotting of points on the map.** This occurs especially in densely populated areas, where multiple points overlap and make precise visualization challenging.
- **Areas with low population may appear virtually empty on the map**, which can result in a biased perception of data distribution, as individual points in scattered areas may not be clearly visible.

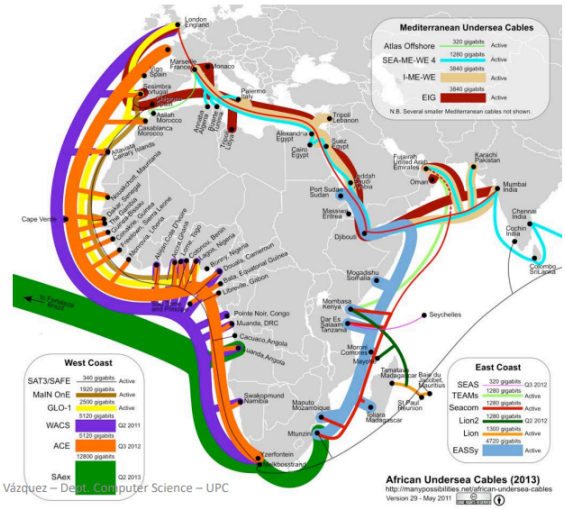


3.6.5 Pixel maps

- Repositions pixels that would otherwise overlap.
- Does not aggregate the data.
- Avoids overlap in the two-dimensional display.
- Provides quite an intuitive result.
- The main idea of the repositioning is to recursively partition the dataset into four subsets containing the data points in four equally-sized subregions.

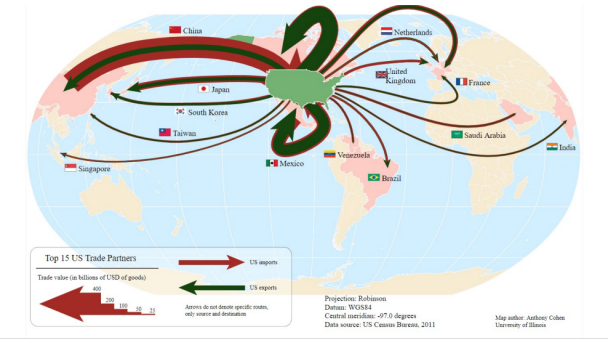
3.6.6 Lines in Geospatial maps

- Map attributes in the map to lines
- Limited application opportunities.



3.6.7 Flow maps

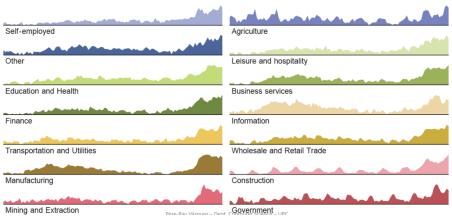
- Depicts movement of a quantity in space.
- Implicitly represents time.
- Can encode a large amount of multivariate information, including path points, direction, line thickness, color, and more.
- May require subtle distortion of the map.



3.7 Other maps

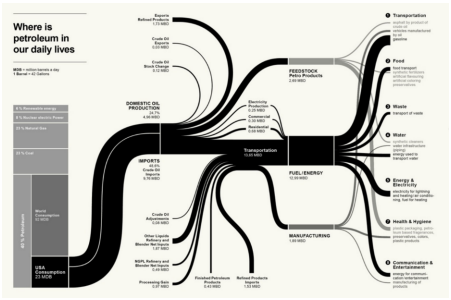
3.7.1 Multiple variables. Small multiples

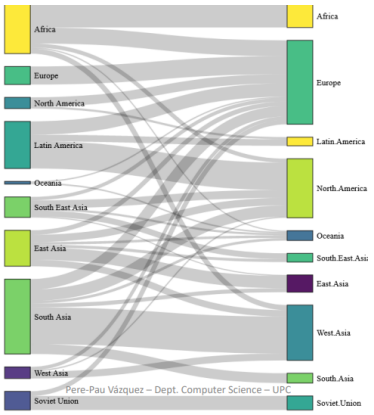
- Grid with axes of smaller charts: This approach is used to facilitate comparison. By arranging smaller charts on a grid with shared axes, it becomes easier to compare different data sets or visualizations side by side.
- The same can be done for time series: Instead of overlapping multiple time series on a single plot, creating small multiples is a useful technique. Small multiples involve creating separate, smaller charts for each time series, making it visually easier to identify trends, seasonal patterns, and other insights in the data.



3.7.2 Sankey Diagrams

- It is a specific type of flow diagram.
- The width of the arrows is proportional to the flow quantity.
- Emphasis is placed on the major flows in the system, allowing for the identification of dominant contributions to the general flow.
- To improve clarity, the diagram minimizes arrow crossings. This may involve omitting or downplaying small or weak flows.
- The relative positioning of nodes is also crucial, as the diagram may become unreadable if there are too many nodes.

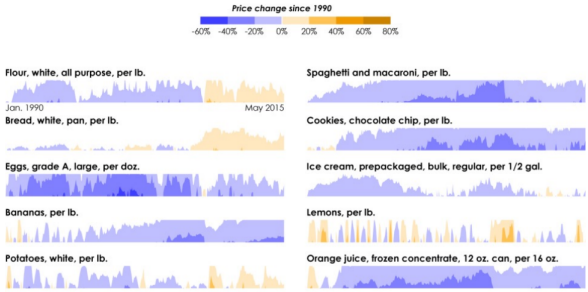




3.7.3 Horizon graphs

Increase data density by overlapping, while keeping the resolution of the graph.

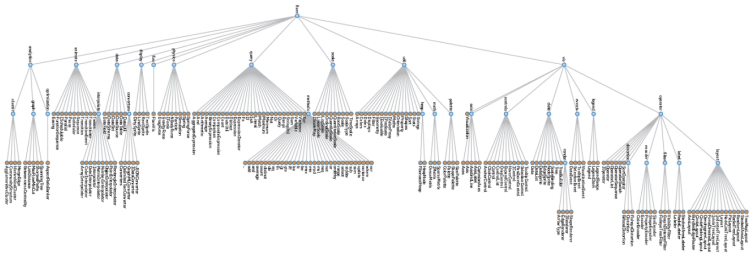
- Start with an area chart
- Mirror negatives to the positive side
- Divide the chart into bands, and mirror again
- Divide the chart into bands, and mirror again
- Result 25% less of vertical space with same resolution



3.8 Hierarchy: Node-link diagram

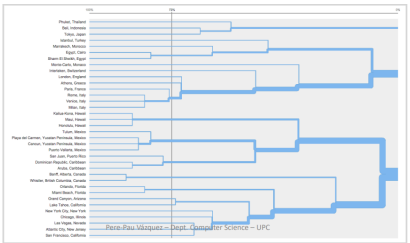
Visualization of hierarchycal relationships to find relationships, groups.

- Line crossing can be a problem
- Might cause a waste of space



3.8.1 Hierarchy: Dendograms

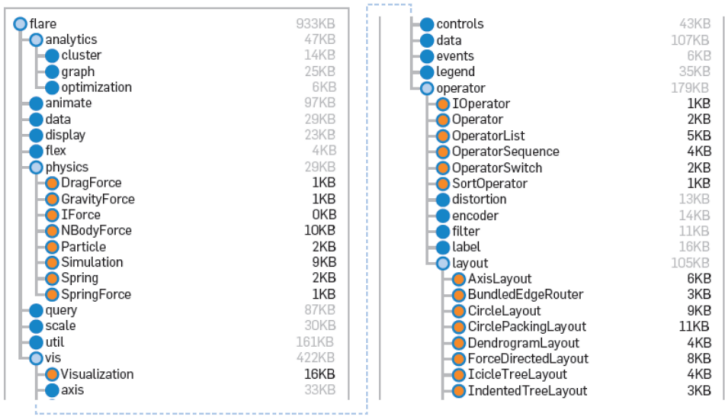
All the leaves are at the same level.



3.8.2 Hierarchy: Indented trees

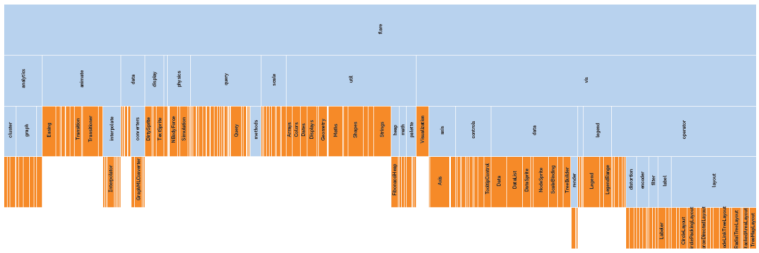
Use in OS to depict the directories.

- Requires a large amount of vertical space.
- Efficient interactive exploration of the tree (to find a specific node).
- Multivariate data shown adjacently (size of file, date, etc.).



3.8.3 Hierarchy: Adjacency diagram

- Space-filling variant of the node-link diagram.
- Nodes are drawn as solid areas.
- The placement of nodes relative to adjacent nodes illustrates their position in the hierarchy.
- Length can be used to encode an additional dimension of information.



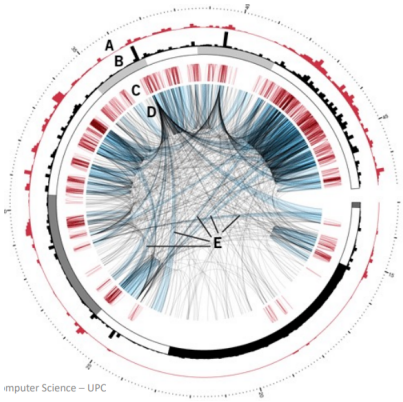
3.8.4 Networks

Contain information about relationships. Information such as: who is connected to, who is a central player (connected to many nodes), groups, cliques. Must place related nodes close and unrelated far away. Reduce crossings may facilitate legibility.

Note: Node-link diagram is an example of a network. Its circle representation is better but implies a lot of crossing

- Drawing is highly complex
- Collinear edges for a large number of nodes
- Very long edges and "meaningless" edge length
- Strong regularity can obscure inherent structures

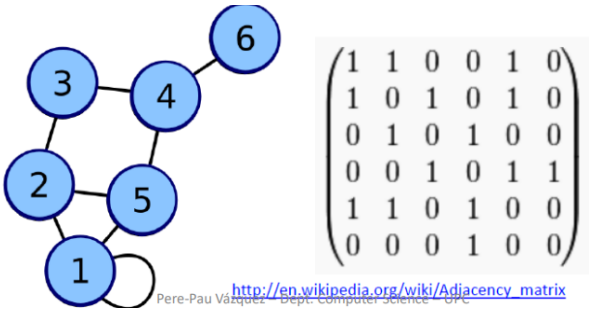
- Very dense drawings for complex graphs
- As a circular layout, the extrenal space can be used for more data.
- Highly regularized and tidy visualization
- Ordering of nodes possible
- Edges or nodes never overlay other nodes
- Easy to visually proceed along edges



3.8.5 Networks: Adjacency matrix

Uses adjacency matrix of the graph. *nxn* adjacency matrix of graph G with n nodes. Use of color might facilitate the interpretation of the links.

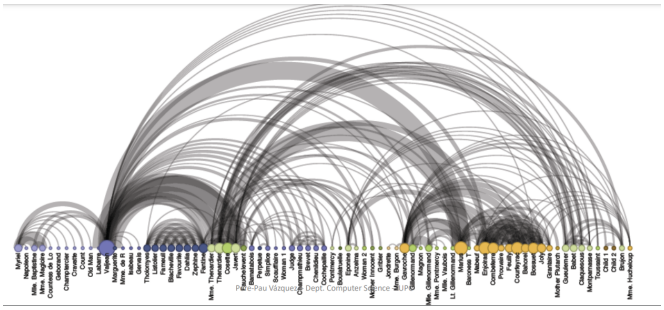
- Difficult to follow "paths" of connections.
- Reordering is expensive
- Crossings are impossible
- Ordering can reveal clusters and bridges (might be done interactively)
- Useful for dense graphs
- Visually scale-able



3.8.6 Networks: Arc diagram

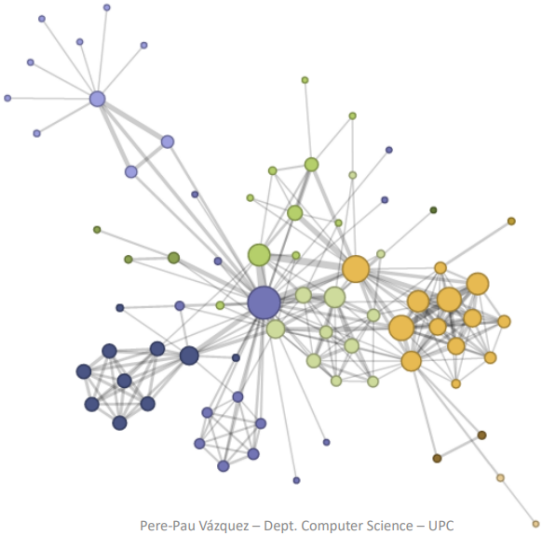
- Lays the nodes in one dimension.
- Circular arcs represent links.
- Good ordering of nodes helps in identifying cliques and bridges.
- Multivariate data can be displayed alongside nodes.
- Problems may arise with the sorting of the data, known as seriation.

- Problems with the sorting of the data: seriation.
- Multivariate data can be displayed alongside nodes (for example: size, colour, etc)



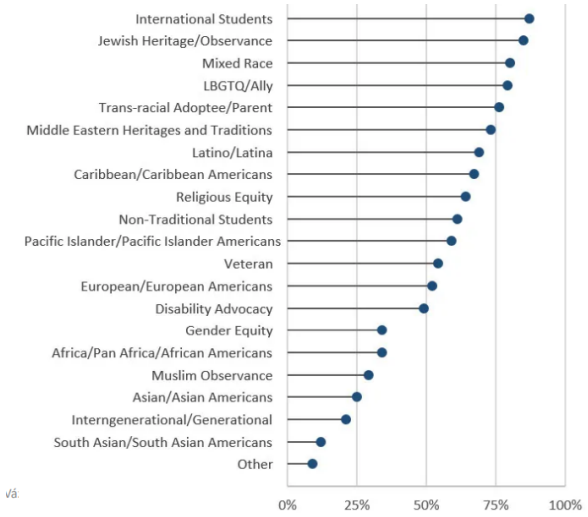
3.8.7 Force-directed layout

- Nodes are represented as charged particles that repel each other.
- Links are modeled as springs that pull related nodes together.
- The layout algorithm uses physical simulation of the forces between nodes to determine their positions.
- Interaction can be added to the visualization to disambiguate links and provide a more dynamic experience.

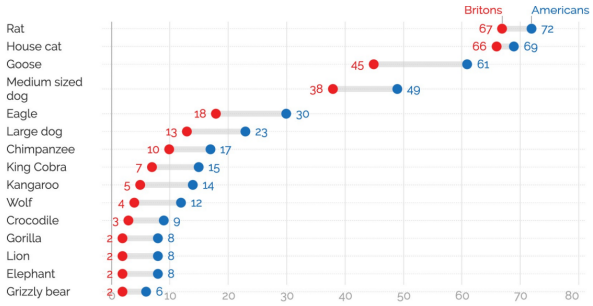


3.8.8 Lollipop

Might not start at 0.



3.8.9 Dot plot with two values



3.8.10 Intersection of sets

Generalization of Venn diagrams. Using Venn (aka Euler) diagrams for more than 3-4 sets is a bad idea. MANY possible intersections.

3.9 Uncertainty

Data is (very commonly) uncertain to a certain degree. We need to communicate this uncertainty. Regular users are not used to uncertainty visualization, because it may lead to the readers considering your data is flawed.

Some common uncertainty methods are violin plots and line ranges.

The only uncertainty communication method that has gone to public is hurricane visualization:

- Cone contains probable path
- Uncertainty grows with time

- Forecasting models exhibit different behaviors (paths)

Problems with cone-based hurricane forecasting vis: Cone size != hurricane size, hurricane impact outside cone

Problems with spaghetti plots forecasting vis: Some are not models, some are old (e.g., 12 hours old), some are statistical models useless for tracking...

4 Perception

We need to understand which elements (distribution of data points, colors, geometric shapes) are involved in how the user perceives the data. The way we design the visualization determines our understanding of the data. In the context of the task we need to solve, we must provide visual tools and affordances.

Affordances: Being able to identify elements that assist you in interacting with the visualization.

- **Simple case scenario**: Small data points and minor attributes. Nearly any visualization will work, but it may lack perceptual properties. With many data points, we might not be able to represent all the data in the visualization.
- **Worst case scenario**: A high number of dimensions that we cannot all represent, and we have to select the most important dimensions. Many visualization techniques may not work properly, even with the best selection of our dimensions. Perhaps some interaction techniques could work.
  - With three or more dimensions, things start to get a bit more complicated (overlapings, ...)
  - With a higher number of dimensions, the position in the matrix and colors can differentiate the dimensions, according to our visualization.

  The perception properties vary for each visualization.

4.1 Preattentive Processing

Create something that can be understood as quickly as possible with minimal cognitive effort. Understand how human visual perception and information processing work. Very related to psychology.

Simplified 3-stage model:

1. The brain can process in parallel to extract low-level properties. Features are processed simultaneously.
2. Extract some structures and create patterns.
3. Perform a sequential search to find the desired information.

Certain variables can be easily identified (fast identification) before beginning to scan the entire visualization. This is important because the user can find what they were looking for and avoid distraction. Identify which features can be perceived rapidly and in what context to use that to your advantage in the visualization.

You have already seen the information before you start searching for it and we know this occurs before your consciousness does by measuring the response time. If it is below a certain threshold, we know that the visualization aid was useful.

The conjunction of several variables is typically not preattentive (e.g., a mix of squares and circles with different colors when looking for the red circle). It requires a more time-consuming serial search for each dimension.

There might be distractors. F.e. only add color based on necessity to ensure that we don't have distractors.

4.2 Perception Laws

Depending on the visual organization of the data, we can perceive some information and not others. Once identified, it is difficult to stop seeing that information.

In general, our brains try to create meaningful things from what we see, even if it is not real. The brain tries to extract some information. The structure of the visualization helps the brain create those ideas to be true.

4.2.1 Pragnänz Law

We tend to perceive simpler shapes.

4.2.2 Law of Closure

The division of complex elements enables us to detect simpler elements/shapes. The mind sees the complex element as a combination of simpler shapes.

4.2.3 Grouping by Spatial Proximity

Even small differences lead to a different perception of the distribution of the elements. Things close together form a group. Increasing the space between them makes the user think that the two groups are not related.

- Grouping by similarity is another variant.

4.2.4 Law of Continuity

There is a tendency to make trends continue. The shape of the edge could create the illusion that two elements are connected.

4.2.5 Law of Common Fate

There is a tendency to group elements that move in the same direction.

4.2.6 Principle of Parallelism

Similar to the law of common fate, but here the groups are based on the orientation of the elements.

4.2.7 Principle of Connectedness

This is a strong principle that can override the other principles. Visual connections can make us reconsider the initial grouping. The line of connection does not have to be strictly connected to the elements (there can be space between the element and the line).

- For example, circuit designs are mainly understood by following the connecting lines.

4.2.8 Law of Symmetry

Elements are grouped based on their symmetry.

4.2.9 Principle of Common Regions

It's based on the boundary lines that contain the elements, even though the elements inside the grouping can have evident differentiation.

4.2.10 Principle of Previous Experience

Our past influences how we perceive the elements. Take this into account to avoid misleading the user and save time by making the visualization more intuitive (e.g., red for negative and positive values).

- There is a tendency to distinguish and detect which elements are in the foreground and which are in the background. We don't want to create elements that are wrong or confusing and do not distinguish between the foreground and background.

4.2.11 Principle of Focal Point

Emphasize the viewer's attention on a specific element of the visualization that serves as an entry point into the visualization. Sometimes, we can enhance it with captivating techniques.

4.2.12 1 + 1 = 3 Effect

A perceptual phenomenon that occurs when elements that are not present appear in our visual system because the organization of the visualization allows it. For example, the small light gray squares in the corners of the big squares.

- We don't want visualizations that allow this principle. For instance, in maps with labels, the labels can be enclosed inside rectangles.

4.3 Application of Perception

4.3.1 Feature Hierarchy

Decide on the hierarchy in which you want to communicate the features, as visual search is hierarchical. Clearly separate them to solve the problem.

- The first scanning process gives an initial understanding of the main elements.
- Eye movements are used to look for new feature candidates, with a limited amount of storage in memory.
- Test whether the extracted features are the ones you were looking for.

4.3.2 Visual variables

...

4.3.3 Texture

F.e. orientation, grain size, blur, shapes, contrast, etc. They might add a lot of noise to the visualization. We can combine methods, orientation and colors, directions, etc.

4.3.4 Glyphs

: Visual representation that encodes data in a specific shape. Encodes attributes of data that can change, so f.e. we can encode two variables with the width and the length of a rectangle.

Challenges of encoding can appear when we cannot distinguish/concentrate on only one dimension (separate dimensions visually). In the rectangle example we can not distinguish/separate both dimensions (length, width).

4.3.5 Direction and orientation

: It is not something easy to do, cause the intuitive thought of arrows can occupy a lot of space. Another methods can be used, like: opacity, width of the direction line.
Things to consider:

- Critical points have to be well represented.
- Depending on the task adjust the visualization.

4.3.6 Transparency

For reducing clutter we can use it sometimes. Not only to solve problems but to encode information. be aware if the colors are the adequates and the contorns are not lost by the interceptions of elements.

In general there is a lot of interference, we have to select good color palette. The mixture of texture and color can be better to represent multiple intersections of elements.

4.4 Pattern learning

Some patterns are already known. We have to make us of those instead of creating new ones. When we are creating complex charts we may use examples, teaching the people to read our charts. If we do not teach the people we can not afford difficult examples. We can reinforce the visualization with new elements like legends.

4.4.1 Complex surfaces

Some methods can help the user to perceive shapes in a visualization. In general we should not abuse the use of shades.

If we want to mimic reality when using visualization we will need a lot of shapes and more elements, not recommended since it can hide the real data. Generating shadows can be dangerous and using textures is ok but not really promising.

4.4.2 Relative judgements

The detection of difference in the bars is not visually instant when the bars are large. **Weber's law**: a notable difference (JND) is proportional to the intensity of the original stimulus.

$$JND(k) = \Delta I / I$$

The perception of this difference can be highlighted with the encoding of the difference visually with an enclosed region or similar method. This means that for comparison we might interested in adding reference to make the comparison easier. This reference affect the way of understanding the visualization, because they compete with the data for our visual attention.

We have to know how accurate we are when estimating values. When estimating values we are more accurate when estimating lengths than areas (and volumes). The higher the dimensions more difficult to estimate. **Steven's psycho-physical power law**: states that the length is very efficient at transmitting information, since estimation is practically imminent.

4.4.3 Tell truth about data

Try not to distort the perception of the user of the visualization.

The lie is the distortion of the graph according to the actual data.

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

We need to be aware that the visualization must match the actual data. the impression of the visualization stays longer in mind than the actual data.

4.4.4 Innovative charts

We need to ensure people get it. If we create some visualization that uses not standard technique to represent some concept we have to ask somebody what they interpret.

The are 2 ways of analyzing these charts. The first thing is try to guess what the chart is meaning without the labels. Then reading the labels try to see if our interpretation changes. The text content of the visualization can give a previous expectation on what the visualization is about.

4.5 Comparison

To facilitate comparison thought visual design:

- **Identify the elements to compare**: target can be explicitly/implicitly defined.
- **Identify the elements that make the comparison difficult**: When the number of items increases, the complexity increases and this will have to be taken into account for the design of comparison. The complexity of the items itself can be a problem itself, f.e. the shapes of two countries, atom molecules, etc.
- **Proper comparative strategy**:We can select a subset of the whole data.
- **Craft a design that facilitate comparison**: When we know the elements we want to compare we have to make something where the user can estimate quantities, difference. The are multiple ways.

5 Analysis of visualizations

Corn yield response (98 - Introduction to vis) Use of 3D for no reason, the information is so simple there is no need for complex plots. The **axis are not clear**, where is the 0? There is **redundant information**, the quantities are displayed on the axis, with numbers and also the numbers next to the plot is just so much information. The **color palette** is not the best one. Also the **title is not informative**, if we want to emphasize that *Trivapro* is the best treatment the title should be something like *Tivapro outperforms other treatments comparison* however, the title only says *Trivapro corn yield response*, when other treatments are also displayed in the plot.

Same data different charts (111 - Introduction to vis) Number C has three different scales, what it beats completely the purpose of visualization, so we cannot quickly understand what are this values. We need to read the labels and make a cognitive effort to understand and compare. Number B has very light colors, it is very saturated. The thing here is that colors are unnecessary because the bar charts are already separated in space. Unless we need to say something to the user with this colors this is unnecessary. The black grid lines are also so visible we need to concentrate on the data not on other things. Number D, what probably happens is that they cut the y-axis, what is completely wrong because bar charts encode quantity by the length. Therefore if we clip the, we are not encoding. This is the most dangerous.

Math Grades (112 - Introduction to vis) The bad thing here is that the 31% is associated with 2013 because is close to that but 2013 is also orange so it could also refer to the bars. The line chart and the bar chart does not coincide on the layers for the axis. The bar chart has two names and the line chart has years. It is not even clear if both charts are together or meant to be separated.

We need to know what to compare, here is complicated because we might be interested in compare NY city (then this should not be a bar chart it should be better a line chart with 2 points), if we want to compare NY city with Economic Disadvantaged is dubious. If we want to compare something and we have a reference we can use a bullet chart. WE do not know if it is a only plot or three separated plots.

Gene Comparison (22 - Good Practices) At the left there is no message. Is beautiful but we cannot understand anything. To improve it we can use that as context for out visualization, we use it at the background and we highlight some connections. And if we only want to explain the message because the context is not important we can make a simpler diagram that is much better to explain the message. Depending if we only want to transmit the message or also give context (inexpert public) will be different. For scientific purposes the last will be better.

Percentage rokers in Neder. (62 - Vis techniques) Labels of the years are not in the x-axis. Colors are not necessary because we are only showing one category and besides the color palette is "horrible". It shouldn't be a bar plot, but a line plot showing the continuous path with the years.

Arthur 51 (63 - Vis techniques) The proportion of the bars is not proportional to the values, whcih suggest that the scale does not start at 0.

Unemployment rate (65 - Vis techniques) The y-axis does not start at 0 (since it is not a bar chart is not that important). The last values does not correspond to the

percentages.

Groceries (66 - Vis techniques) It is difficult to compare since they are areas. The colors does not help and they dont directly correspond to the values.

One barrel oil (68 - Vis techniques) The length of the bars dont correspond proportionaly. The width of the bars are different but iut doesnt specify what for.

Exercise (205 - Vis techniques)

- **a.** You know what you have is real information. If there are a lot of missing rows if we delete we may lose information not only in data but in context may be useful. Since we don't know the reason of the value missing we cannot calculate other variables like total count, frequency, date. So that 1 row has 1 missing key does not mean that it does not contain important data in it.
- **b.** We keep the information of the other variables, that may be relevant. Disadvantages are that if we use this values to calculate things we may end up with wrong conclusions.

Representation Exercise (206 - Vis techniques) We cannot know because we don't know the population. The population should not be factored out.

Representation Exercise (207 - Vis techniques) Dual axis is intended to show 2 variables that are different like *energy consumption (watts)* and *price of electricity (euros)*. If we use a dual axis plot we first need to know that the variables are correlated and select correct ranges. We can use this plot to analyse the correlation.

Representation Exercise (211? - Vis techniques) Time data is shown in different layers of the stack bars (would be better if time was on the x-axis). Comparing elements that are not aligned is difficult. The time period intervals are not equally distributed. *Thunder bay* is in **Ontairo** and *Ontairo* is in *Canada*. We don't know the population density of the three regions so it can be misleading. Maybe the colors are not adequate because some colors are similar.

Representation Exercise (213? - Vis techniques) The differentiation of *race* and *age* is not evident. We could tend to group the two charts and we should make clear that they are separate. The proportion of the intervals are not the same. Mis-match between the death quantities with the length of the bars, they both indicate different things but since they are closer together make ine think that they are related. We dont know the population of each *race* or *age* and that can lead to some wrong conclusion. Is the rate of birth the same for each *race* or *age*? The emphasis of the color of the bar is not necessary since the *negrita* of the text is enough.

6 Homework

Watch *https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_se*en, Hans R and answer those questions:

- **How many variables (and which) are displayed in the following charts:**
  1. **The one shown at 2:26**
     The are **5** variables displayed. **Life expectancy** on the y-axis, **fertility rate** at the x-axis, **total population** that is the bubble size, **continent or region** that is the color of the bubbles and **time** (year).
  2. **The one shown at 13:07**
     There **5** are variables. **Child survival percentage** on the y-axis, **GDP per capita** in the x-axis, **total population of the country** that is the bubble size, **continent or region** which the country belongs that is the color of the bubbles, **time** (year). The are several data points for each country, in which we do not know the order. This makes certain comparison difficult. To emphasize the countries that we are interested he uses the color.

- **Find an example of visual representation that does not effectively communicate the message (minute and second, and reason why)**

06:04 The visualizations that uses density distributions. This amounts are difficult to estimate when we see density charts that are not overlapping but are on top of the other. Also the values in the x-axis are not completely visible. No y-axis is shown either.

- **What happens in terms of variables when he "splits South Africa"?**

We see the how the data is distributed. It is also highlighted with respect to other data points. When we split something and the rest is not split we compare different things, this may lead to make comparisons that are not good.

- **Bonus: Who is the "ghost"?**

The ghost is China. Hans R refers to China as a *ghost* when explaining that China's income distribution is overlapping the USA one and their economy is growing so fast compared to USA.