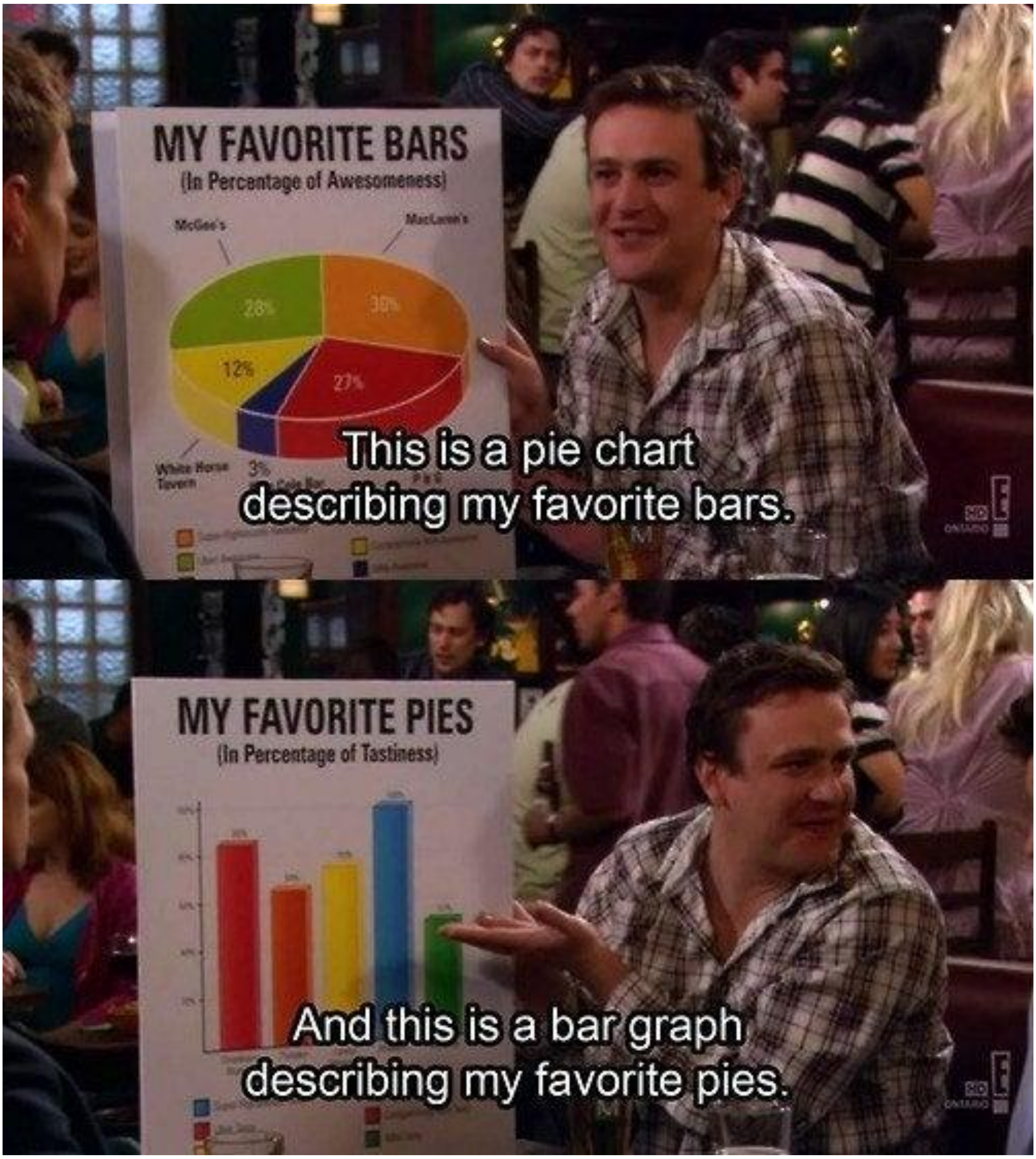
DATA VISUALIZATION



TABEL OF CONTENTS

LECTURE 1: INTRODUCTION

12 FEBRUARY

The return of Dreetje

Definition

Data visualization is an umbrella term that contains both scientific visualization and information visualization. In most of this course we'll be talking about information visualization that is still very abstract. Information visualization is the use of computer-supported, interactive, visual representations of abstract data to amplify cognition. Scientific visualization is more about rendering: What the object would look like in real life. There's also a sub community that focusses on visual analytics.

Principles

On the first example, the X axes starts at 60%, and because of that the differences appear way bigger than they are. Bar values should start at zero. Second example with Bush tax cuts: Same thing. Unemployment under Obama: The labels are not on the same height, which is misleading. 3D can be very misleading as well because some things can look bigger or smaller than they actually are. Next example: Occlusion problem, the bars in the front block the bars in the back. You can't see all the data. Easy to solve by making different charts, especially if you don't have that much data. "Save the pies for desert": They can be hard to read, it's difficult to see which slice is bigger than the other. Don't use too much ink for many decorations.

Main principles: If the purpose is effective communication, it's key to do it with clarity and precision.

Design aesthetics

Show data and maximize the data clarity. Make the data stand out from the background. **Data-ink ratio**: "A large share of ink on a graphic should present data information (…)." Don't use too much ink on the background, such as color or grid lines. Maximize the data-ink ratio within reason. Erase non-data ink, within reason. Erase redundant data-ink.

"Perfection is achieved not when there is nothing more to add, but when there is nothing left to take away."

Workflow

1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact

Tools & IPI

- Spotfire
- Tableau
  Will get exercise sessions on this. They are tools like Excel but have more powerful visualization tools.
- Gephi

Go for the easier options.

COURSE OVERVIEW

Goals: Know the foundations, learn about existing techniques and tools, build.

50% of the grade is a group project. The other 50% is a closed book oral exam. She'll provide example questions. Mostly it's applying the principles and critiquing visualizations. Team project: Build an interactive piece of visualization. Most of development work will happen in class. Limit the number of things and be creative. You can bring your own data.

Milestones:

1. Form teams (1/3)
2. Project proposal (23/3)
3. Intermediate presentation (22/4)
4. Final presentation (20/5)
5. Report (31/5)

There's a discussion board on Toledo with inspiring examples. Don't forget to start thinking about teams.

LECTURE 2: PERCEPTION AND PRINCIPLES

FEBRUARY 19, 2020

Perception is how our brain perceives and interprets visuals. Objects appear smaller when they are included in something larger. That's why you eat more food when it's on a bigger plate.

PRE-ATTENTIVE PROCESSING

How do we make something pop out? That's how you see patterns. Think of 'Where's Waldo'? Pre-attentive takes more time, parallel processing. What kind of features and representations are pre-attentive? Color is already one. Another one is the shape. Enclosure is another one. Orientation. You can combine them but it's not always a good idea. For attentive visualization you need to put in more effort.

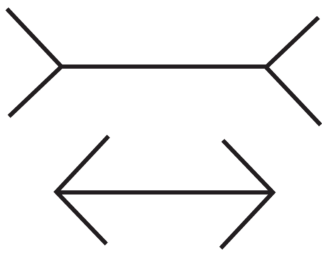


FIGURE 1: MULLER-LYER ILLUSION

ENCODING METHODS

It is easier to estimate how much bigger a bar is than a circle. You are more likely to estimate a lower number when it's a circle. When you try volumes, it even gets worse. Position and length are typically very accurate. Area, volume and color density are even more difficult. Categorical data: sex. Ordinal data: age. Ordinal; there's an order. In quantitative there's not necessarily an order. Sequential: People grow order. Diverging: There's a break point. Cyclic.

Intuitive use of color: People will assume a lighter color will mean a lower saturation of the data whereas a darker color is a higher saturation of the data.

GESTALT GROUPING

Human beings see patterns. We're wired to notice them. That's why we often talk about correlations that don't necessarily make sense: Think of the guy who saw the face of Mary in a lemon slice. Multi-stability: Figure-ground images. Think of the faces and the vase. Closure: Even though an image is not complete, we can see it as such. The next principle is similarity: Rows of circles and triangles because we have a tendency to group things together. With proximity, we do the same. Getting someone into a perceptual set: Sort or prepping someone into a certain kind of thinking, allowing someone to activate a scheme in their brain. It's important that users see your visualization as the figure and not the background. The most important principle is connectedness. If you connect two dots, of course you'll see them as related. Smallness: White/black propeller. You tend to see the smaller object as the figure and not the background.

Surroundedness: If we surround the objects it's easier to see them. Use a combination of closure, common region and layout to ensure that data entities are represented by graphical patterns that will be perceived as figure, not ground.

COLOR

Color helps us break camouflage. Color is a physical property of light itself. We have three color cones: Red, blue and green. They operate at different wavelengths. The one we use for blue is quite distinct, the ones for green and red partially overlap. The blue one is less sensitive. That's why it's hard to read blue text on a black background. We often take color for granted. Colors also have meaning. Green in our culture is positive, red is negative (think traffic lights). That is culturally different. We don't just play with the different colors, but also with saturation and luminance. If you use saturation, an order is implied. If you use hue (different colors), you assume the data is categorical. With luminance, the data is probably ordinal and quantitative. There are diverging color schemes with break points. Colorbrewer2 shows you different color schemes that you could use.

Colors can also interact: Colors can show differently based on the background color. This gets important when you need to read values from a map. Avoid using a grey scale as a method for more than a few (two to four) numerical values. All colors are equal, but they are not perceived as the same, blue and red have higher values than green and yellow. Haloing effect: Enhance the edges, luminance contrast as a highlighting method. Highlighting: Make a small subset clearly distinct from the rest. Use more saturated colors for small symbols, thin lines or small areas. Use less saturated colors for large areas.

Cross-cultural naming: More than 100 languages showed that primary color terms are consistent across cultures. Use the primary colors. Those are hard-wired in our brain as the colors we use a lot. It is necessary to remember a color coding; these colors are the firsts that should be considered. Chromostereopsis: Most people see the red closer than the blue, but some see the opposite effect. Red light is refracted less and is focused towards the ear. Blue light is retracted more and is focused towards the nose. If we use two far pure. Colors in the same image, the eye is not able to focus on both of them. That's why a blue text on a red background is hard to read. Use can be good for highlighting, creating 3D effect, etc. Resolve if united by using colors that are less saturated, surrounding the contrasting colors with a background that moderates the effect of their different wavelengths, separating the contrasting colors.

Color is excellent for labeling and categorization. However, only a small number of colors can be used effectively. To show detail in visualization, always have considerable luminance contrast between background and foreground. Simultaneous contrast with background colors can dramatically alter color appearance, making colors look like another. Beware of the interaction between colors. Small color-coded objects should be given high saturation. Red, green, blue and yellow are hard-wired into the brain as primaries. If it is necessary to remember a color coding, these colors are the first that should be considered. Remember that colors have meanings: use appropriate color palettes for qualitative, quantitative and ordinal data. Respect the color blind.

LECTURE 3: REPRESENTATION OF DATA

FEBRUARY 26, 2020

ENCODING OF VALUES

Anscombe's quartet: If you let some statistics work their magic on the data, you can see something. You can start with statistics or visualization but always do both. They made a movie about Ben Schneiderman, it's on YouTube. The original aircraft altimeter was responsible for many accidents, so it was inconvenient. This is because it worked like a clock with different arms pointing at different values. Change blindness: Picture with the plane engines. It's hard to see small changes. Another example is the one of single numbers. That's why they're doing the beeping noise when they're tracking someone's heart rate – so doctors can immediately know when something changes.

Univariate data: Boxplots are often used in scientific papers to visualize some data because they easily show what the results are for the largest part of the population. In a bar gram, all data is visualized horizontally. Robin is not a fan of word clouds, because often a lot of data is lost. It's also hard to read what the difference is between the different words.

Bivariate data: Scatterplot. If the values below the line have a meaning (e.g. stock data) you can use an area filling color to plot the data). For easier tasks, people prefer easier visualizations. For more difficult tasks, people wanted more detailed visualizations. Linked data:

Trivariate data: Almost useless to show the data in a 3D shape – unless the user is able to turn the visualization around and can see it from different angles. A scatterplot matrix is more useful. <video of the guy with the small families and long lives>

Hypervariate data: You can use a simple scatterplot matrix, but you can also turn it around and use every parameter as an axe. Every house would become a line – this is called parallel coordinates, but it takes a while to read. Robin thinks it's a powerful tool to combine a lot of data. He admits you need to get used to the visualization. Star plot (aka radar plot): makes it a little more comprehensible. It's a nice tool to compare data (both to each other but also to average values). Another way is to allow users to filter the data in different scatterplots where all the data is plotted in a different color. If you have a lot of different parameters, you can think about a mosaic plot (think of the example with Titanic survivors). Always keep in mind what the most important story is that you want to tell with your visualization.

Another technique is Chernoff faces. We can easily see faces in data. Faces are difficult.

ENCODING OF RELATIONS

A relation is a logical or natural association of (…). Relations can have a direction. Doesn't always have to be a line, can also be a color (if both things have the same color they're related). If you have a lot of data, it can become more and more complex. Another way is a chord diagram, where you show relations between groups. Maps and diagrams: use a mapping diagram to show which hotels have different facilities such as swimming pools, golf course or restaurants.

Robin did a study on which you can find more information in the slides.

Another way to visualize data is the intersection explorer. Might be an interesting alternative to Venn-diagrams when you have a lot of data. The final part are trees for hierarchical data. Trees get harder when you have more data. Can become quite complex quite quickly. An alternative is cone trees where you use a 3D effect to group data. Another approach is to use a tree map.

If you're creating a tree with nodes and ties it might be useful if you're able to highlight one particular node and move it to the center so you can see what relationships exist. There's a book available on the interwebs. It's understand to not only know what your data is but also who your user is. You also always need context – exercise with visualizing numbers 23 and 7.

PRESENTATION / INTERACTIVITY

PRESENTATION

There's a lot of information available, but limited space to display that information. It's hard to get all information on one display. How can we display all data on one screen?

SPACE LIMITATIONS

Scrolling: You only have a limited part of the information available on the screen, while most information is not visible. **Overview + data**: Detail can get lost when everything is crammed into one screen. A familiar example is the use of maps, so you have one map with the overview (traveling from one city to another) and more detailed views (of the city itself). Nowadays you can zoom in to get more details. **Focus + context**: You have the focus and the context is distorted, in the way that you can see that there is more information available, but you cannot see it straight away. This is called bifocal display features. The information moves smoothly and continuously from context to focus. Display affords for representation: opportunity to use two dimensions (time assigned to horizontal axes, e.g.). **Suppression**: Applies a distance function and a relevance function. Less relevant other items are dropped form the display. A classic example is New York's vision of the world. More relevant information is presented in great detail, the less relevant information presented as an abstraction. Fisheye example: Mac docking station. **Zoom and pan**: Panning is the concept of using a smaller area on a bigger area and move that area around. Most commonly used in Google Maps. Conventional zooming in has no change in data or representation – only filtering. You lose context. It's not distortion whose purpose is to permit focusing rather than filtering. Zooming supports two cognitive tasks: Zooming-in and zooming-out. Geometric zoom is continuous. The other is semantic zoom, which has a discrete transition and additional detail.

TIME LIMITATIONS

Rapid serial visual representation: People are able to analyze 10 unrelated images per second. Eye-gaze: Think of the different modes in which pictures were shown – some were easier to comprehend; others were more challenging.

Interaction

Select/focus Mark something interesting, for example in a scatterplot. Change the set of data items presented based on some condition. If your main task is to explore the data, you can use different techniques than when you're explaining the data. Keep in mind what the main task is when you're visualizing.

Breadcrumbs are graphs that show the users graphs of all the different steps that they've taken so far.

The return of the Dre: Communication

March 11, 2020

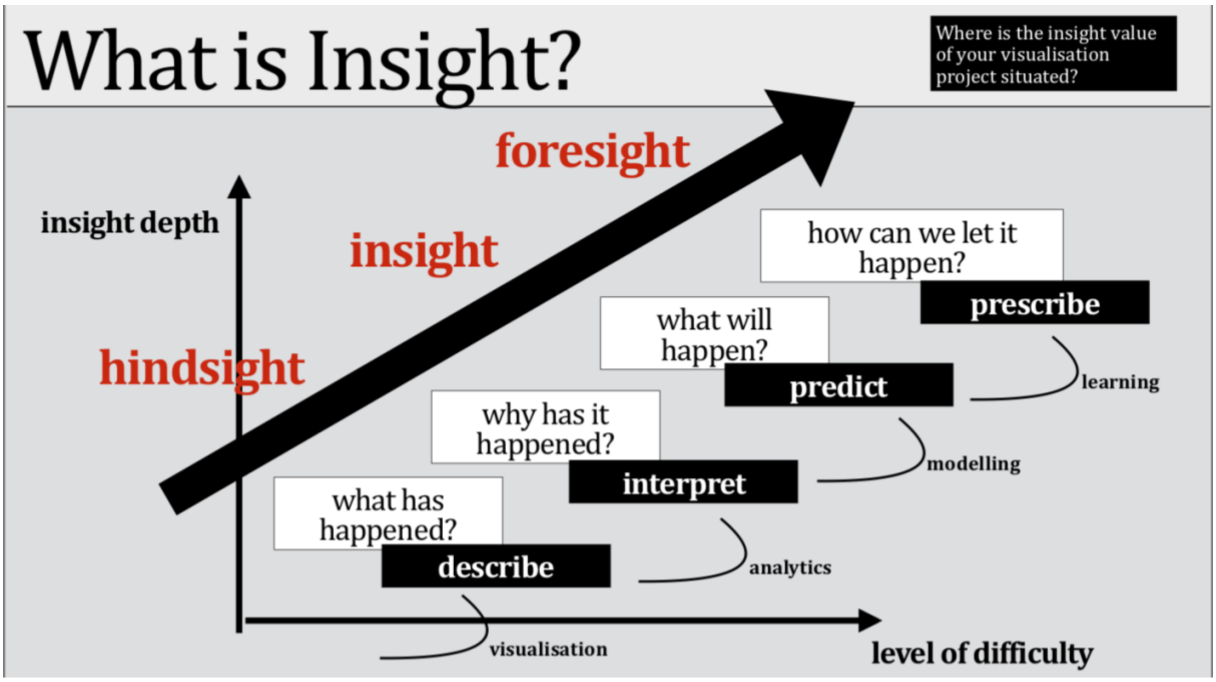
Dre had a blog: Visual aesthetics. Insight is a piece of knowledge that's not necessarily obvious at first glance. It's new, unexpected and useful. An insight has a 'depth' (factual, interpretative, reflective). In a 'good' visualization, an insight manifests itself. Insights can describe, interpret, predict and prescribe. Reflect in your project what your goal is. Consider data proxies, not data itself. Visualize data interpretations, not data itself.

Follow a design approach, avoid visualization pitfalls, follow good practice.

Lecture 6: Communication

10 March 2020

Visualization guidelines



Choose a data strategy

a. **Visualize insights, not data itself**

What is insight? Wisdom < knowledge < information < data. An insight is original (new, unexpected, useful), an insight has depth (factual, interpretative, reflective) and drives further ‘sense making’, drives additional insights (intermediate result of iterative process. Initial hypothesis leads to more hypothesis forming).

b. **Consider data proxies, not data itself**

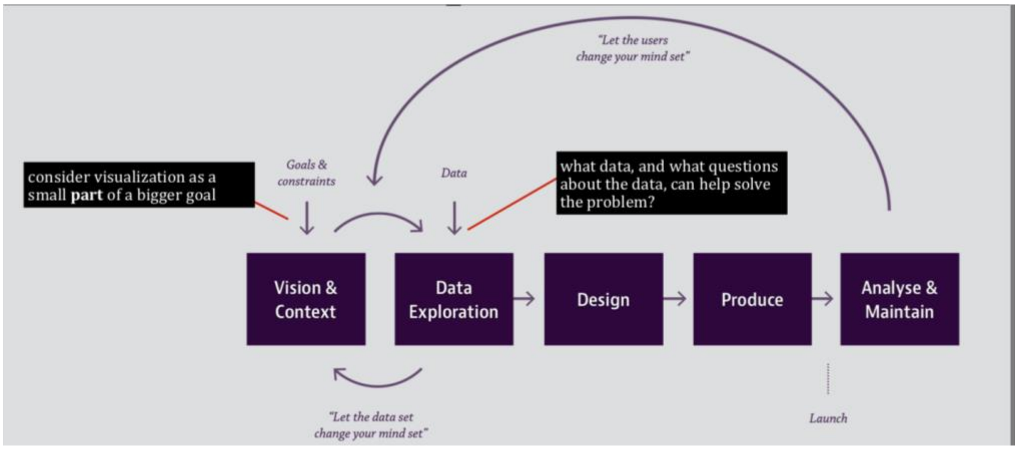
Think of the map with the ‘lights’ of the places where people take pictures. The red is the ones that are uploaded on Flickr, the blue the ones on Twitter. Proxy data are preserved characteristics of specific phenomena that can ‘stand in’ for (often hard to get) direct data measurements. E.g.: Metadata of social images seem to have more insight value than the content of the images themselves.

c. **Visualize data interpretations, not data itself**

Think of the website with all the selfies on it.

CHOOSE A VISUALIZATION STRATEGY

d. **Follow a design approach**



Long story short: You have to go back and forth between the different design septs. Always reevaluate what you’ve created so far and see if you can improve anything.

e. **Avoid visualization pitfalls**

Think of the data item (point, line, area, volume) and the visual attribute (position, size, length, are, volume, orientation, angle, slope, color, etc.). Make sure it’s computable, invertible and comprehensible. People have a hard time to estimate how big a surface is. No visualization or ‘map’ is objective, even the scaling of an axis implies a meaning. There’s a general consensus that the ‘average line slope in a line chart should be around 45°. People like organic shapes (and colors), people like shapes that resemble fractals: Complexity with a certain level of uncertainty. Be critical of maps that don’t’ consider population density – normalize the data first. Be critical of insights. Also be critical of data classification, acknowledge they carry meaning.

f. **Follow good practice**

Show comparisons, contrasts and differences. Individual numbers carry little to no meaning. Consider contrasting numbers with other numbers (evolution over time, different sources, etc.). Show causality: Show or suggest the ‘why’, which is more important than the ‘what’ is inside the data when engaging lay people. Show multivariable data (more than 2 dimensions). Describing phenomena can often occur with 2 dimensions, and a 3rd or 4th dimension often suggests. Integrate all words, diagrams, numbers, images in the visualization itself. Describe the context or other evidence of the insight in or around the visualization itself. Show only the data (remove all graphical elements which can be removed without changing the meaning of the visualization).

Choose a user engagement strategy

2. **Visualize a purpose**
   Consider the purpose. Visualization is based on data. Data is a (simple) view of the world. The visualization is shown to the world. The world might change because of your visualization (yeah no but ok). Visualization can have a functional purpose – analyze, explain, inspire. You data can also engage through interaction, where people can influence the way the data is visualized based on what they're interested in.
3. **Integrate narrative and storytelling principles**

   By doing this, you can get 'direct' human attention. Structure a story: how, when where, who wat why? When in doubt, leave things out. Cut from the bottom. Order prevents causally. Put a 'you' in the headline, make an appeal. Use comparisons and contrasts to make your point. Illustrate to make the data easily comprehensible. Simulate: Make the processes behind the data more easily comprehensible. Create a message: Relate, juxtapose (show several realities next to each other to highlight the difference, or the lack thereof, and the causal effect in between), repeat, emphasize, familiarize, illustrate, show the 'human' behind the data, make statement, choose, vote, suggest, …

Be critical

4. **Visualization is subjective**
   Is the data truly representative of certain phenomena? Almost all data has noise, although we 'see' patterns, they might not be there. Can everything be captured by data? What is the meaning of a statistic versus its individual or collective experience of people that are captured in this data? Can everything be compared? Data fallacy: What if all decisions are made purely on the basis of data? Or its visualization?

Lecture 7: Information dashboards

A dashboard is a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen so the information can be monitored at a glance. Think of the example of a car dashboard. Dashboards are visual displays that display information that's needed to achieve specific objectives. They're used to monitor information at a glance.

Variable	Values
Role	Strategic Operational Analytical
Type of data	Quantitative Non-quantitative
Data domain	Sales Finance Marketing Manufacturing Human resources Learning …
Type of measures	Balanced score cards Six sigma Non-performance

Variable	Values
Span of data	Enterprise wide Departmental Individual
Update frequency	Monthly Weekly Daily Hourly Real-time
Interactivity	Static display Interactive display
Mechanisms of display	Primarily graphical Primarily text Integration of graphics and text
Portal functionality	Conduit to additional data No portal functionality

They focus on KPI's. They can convey more meaning than text alone. Dashboards should have clear, intuitive displays. 13 common mistakes:

1. Exceed the boundaries of a single screen (information on dashboards is often fragmented in 1 or 2 ways: Separated into discrete screens wo which one must navigate or separated into different instances of a single screen that are accessed through some form of interaction / require scrolling)
2. Supplying inadequate context for the data
3. Displaying excessive detail or precision
4. Choosing a deficient measure (use of measures that fail to directly express the intended message)
5. Choosing inappropriate display media (pie charts are difficult to interpret accurately; horizontal bar graphs are better. Other types of graphs can be equally ineffective)
6. Introducing meaningless variety
7. Using poorly designed display media (don't put a legend next to a pie chart because this way your eyes have to go back and forth between the graph and the legend; order of the slices shouldn't be random; bright colors produce sensory overkill)
8. Encoding quantitative data inaccurately
9. Arranging the data poorly (most important data should be prominent; data that should be compared ought to be arranged and visually designed to encourage comparisons)
10. Highlighting important data ineffectively or not at all
11. Cluttering the display with useless decoration
12. Misusing or overusing color (too much color undermines its power)
13. Designing an unattractive visual display (

Strategies to create more effective dashboards

- Reducing the non-data pixels
- Enhancing the data pixels
- Designing dashboards for usability
- Exceptionally well organized
- Condensed (primarily in the form of summaries and exceptions)
    - Summarize the data as a single number
    - Show critical values exceptions (show the one or two data points that require attention)
- Specific to and customized for the audience and objectives
- Eliminate and enhance data pixels
- Displayed using concise and often small media that communicate the data and its message in the clearest and most direct way possible
- Support meaningful comparisons
    - Combine items into a single table or graph
    - Place items close to one another
    - Linking items in different groups using a common color
    - Include comparative values (ratios, percentages, etc)

Tldr; Pie charts suck

FIGUUR 1: MOOD.