Visual Analytics & Information Visualization

Lecture 19

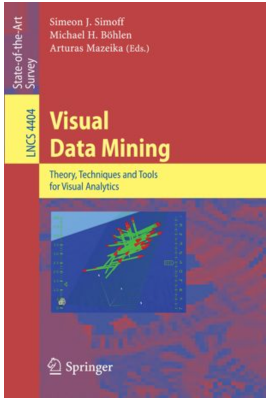
IDS-L19

Chair of Process and Data Science

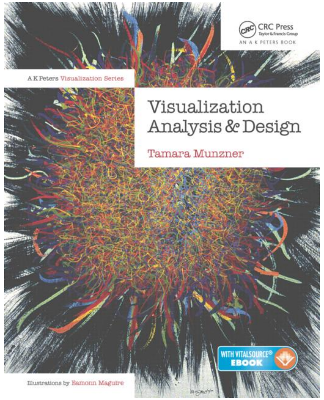
RWTH AACHEN UNIVERSITY

Outline of Today's Lecture

- Visual Analytics
  - Scope and Process
  - Examples
  - Challenges
- Advanced visualization techniques
- User tasks
- Visual encoding principles



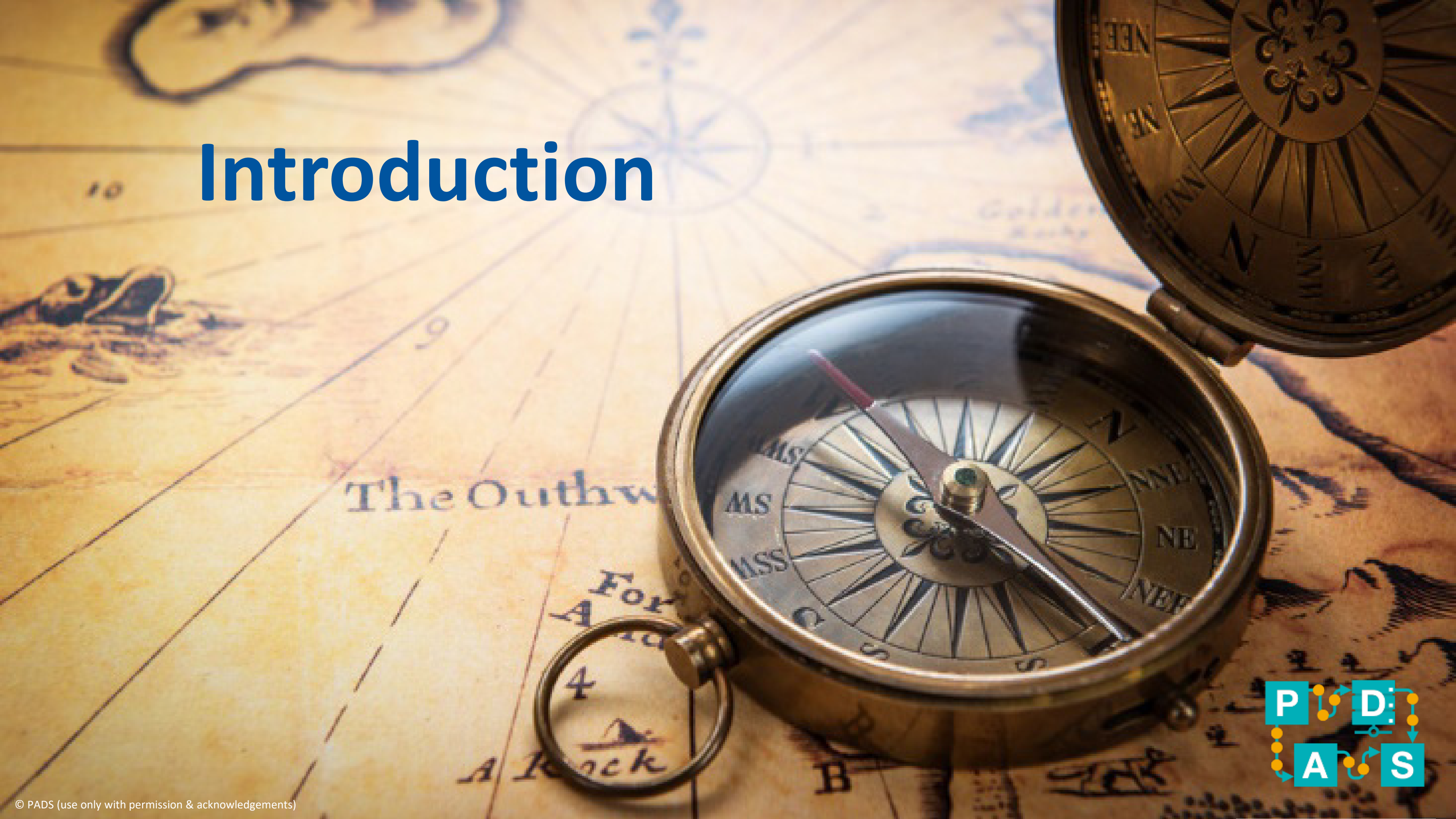
Visual Data Mining Theory, Techniques and Tools for Visual Analytics by Simeon Simoff, Michael Böhlen, Arturas Mazeika.



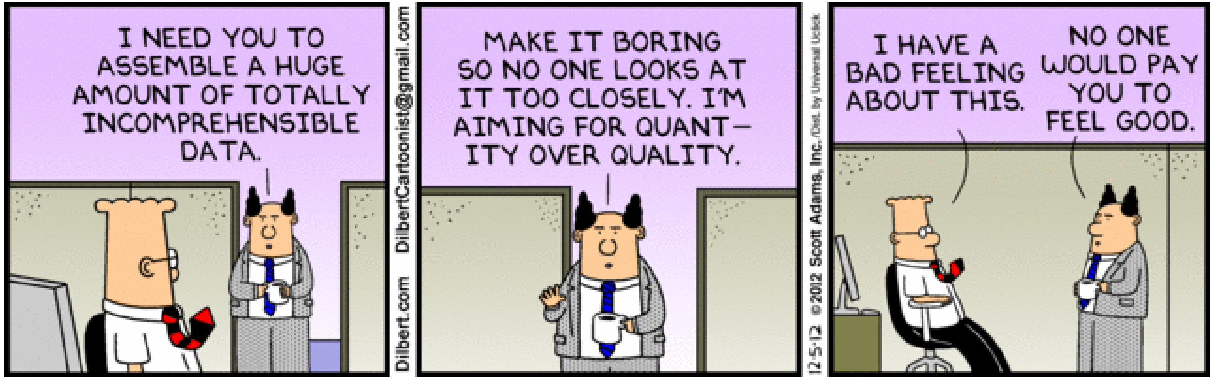
Visualization Analysis and Design: Principles, Techniques, and Practice by Tamara Munzner.

Chair of Process and Data Science

Introduction



Introduction



- **In the last years there was a rapid increase in the amount of data produced by computer systems.**
- **Raw data has no value itself, only extracted information has.**
- **Not just known unknowns, also unknown unknowns.**
- **Exploration is a vital first step (understanding, spotting data quality problems, building thrust, etc.).**

Chair of Process and Data Science

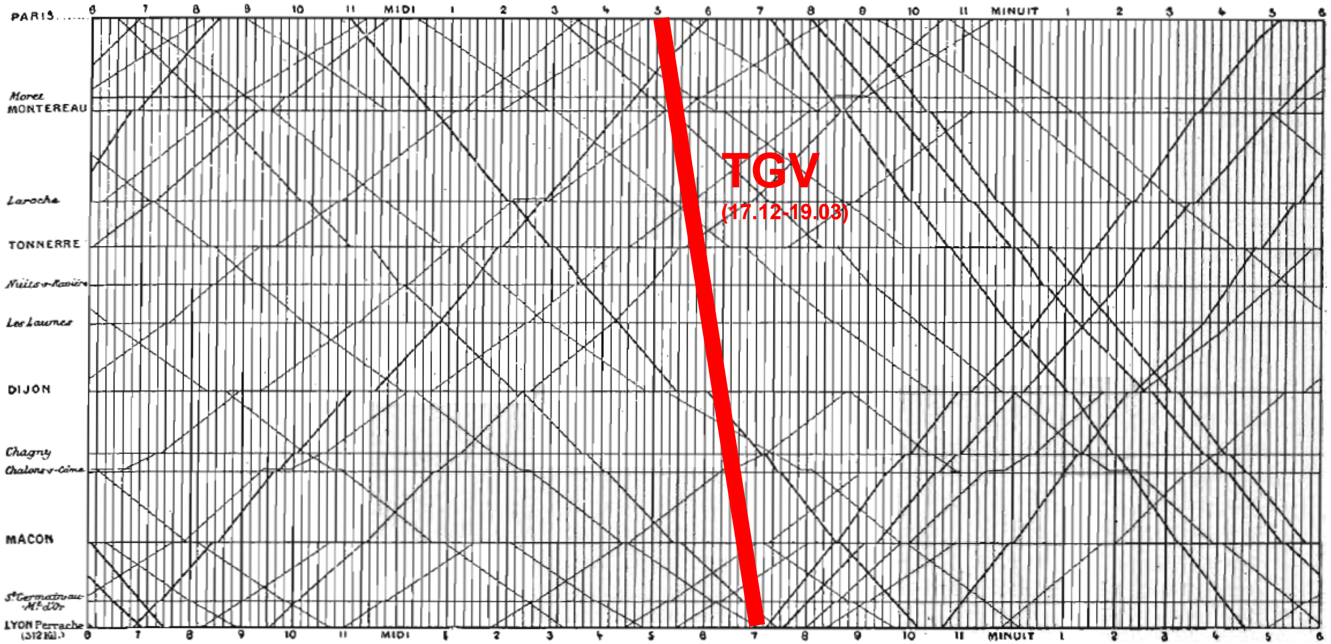
Classical Visualization Examples

John Snow (1813-1858) was an English physician and considered one of the fathers of modern epidemiology. He traces the source of a cholera outbreak in Soho, London, in 1854. His findings inspired fundamental changes in the water and waste systems of London.

John Snow's map of the 1854 cholera outbreak in Soho, London

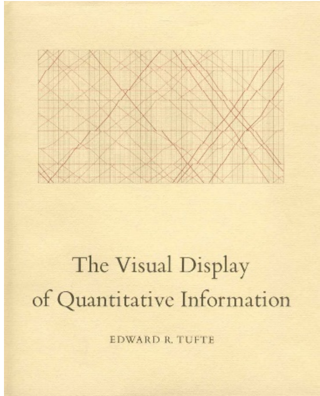
Chair of Process and Data Science

Classical Visualization Examples

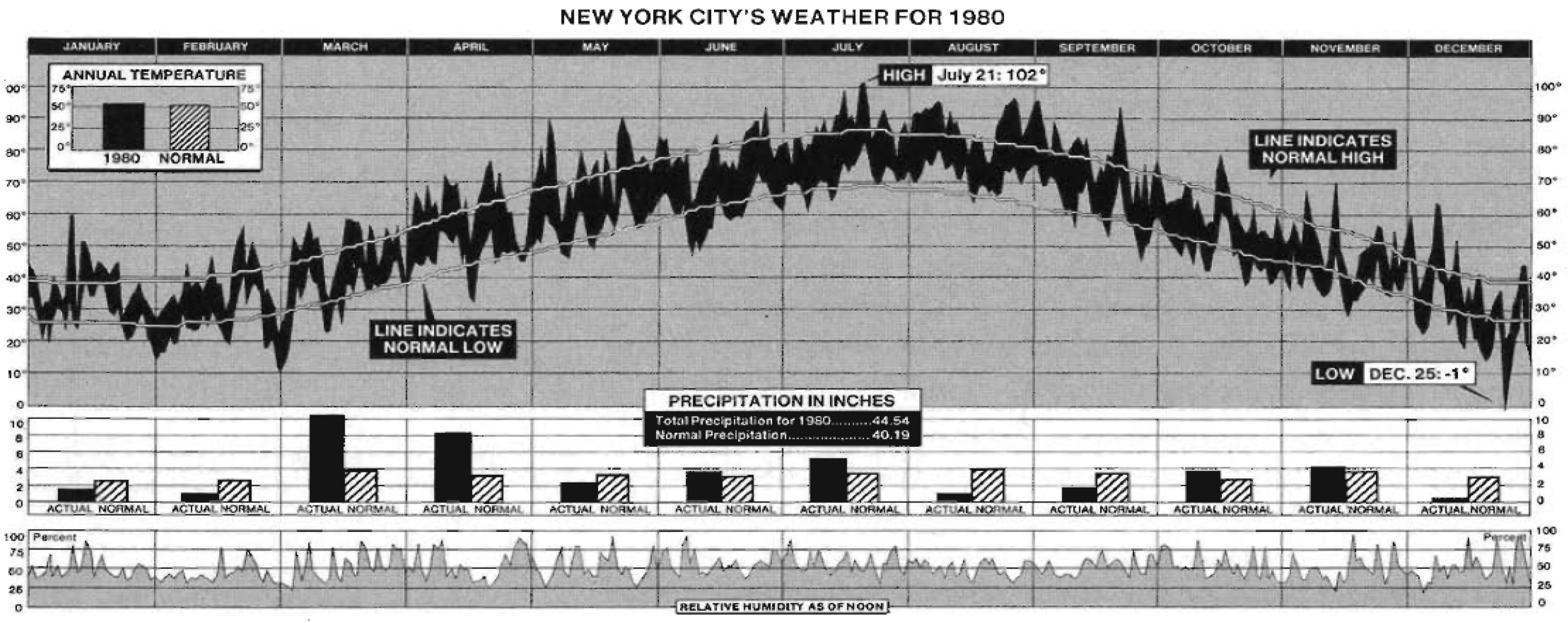


E. J. Marey, *La méthode graphique* (Paris, 1885), p. 20. The method is attributed to the French engineer, Ibry.

Étienne-Jules Marey 1885.



Classical Visualization Examples

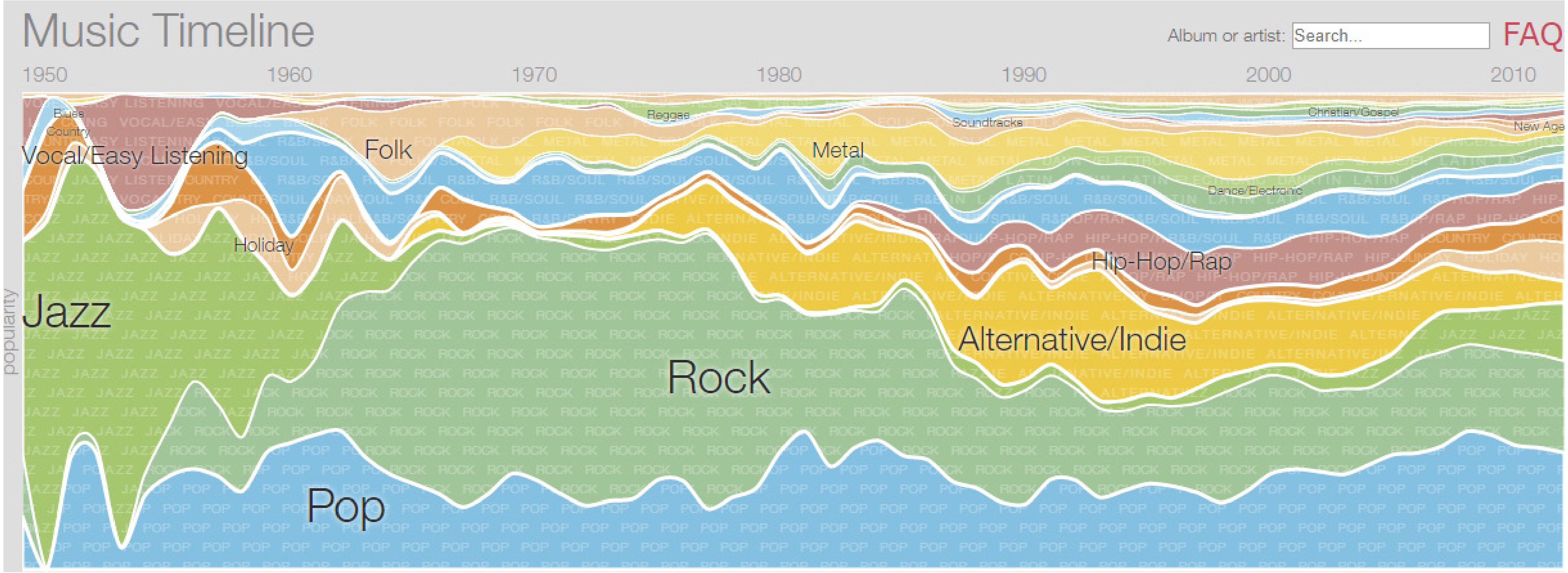


New York Times, January 11, 1981, p. 32.

NYT 1981, showing 1888 numbers!

Chair of Process and Data Science

Classical Visualization Examples

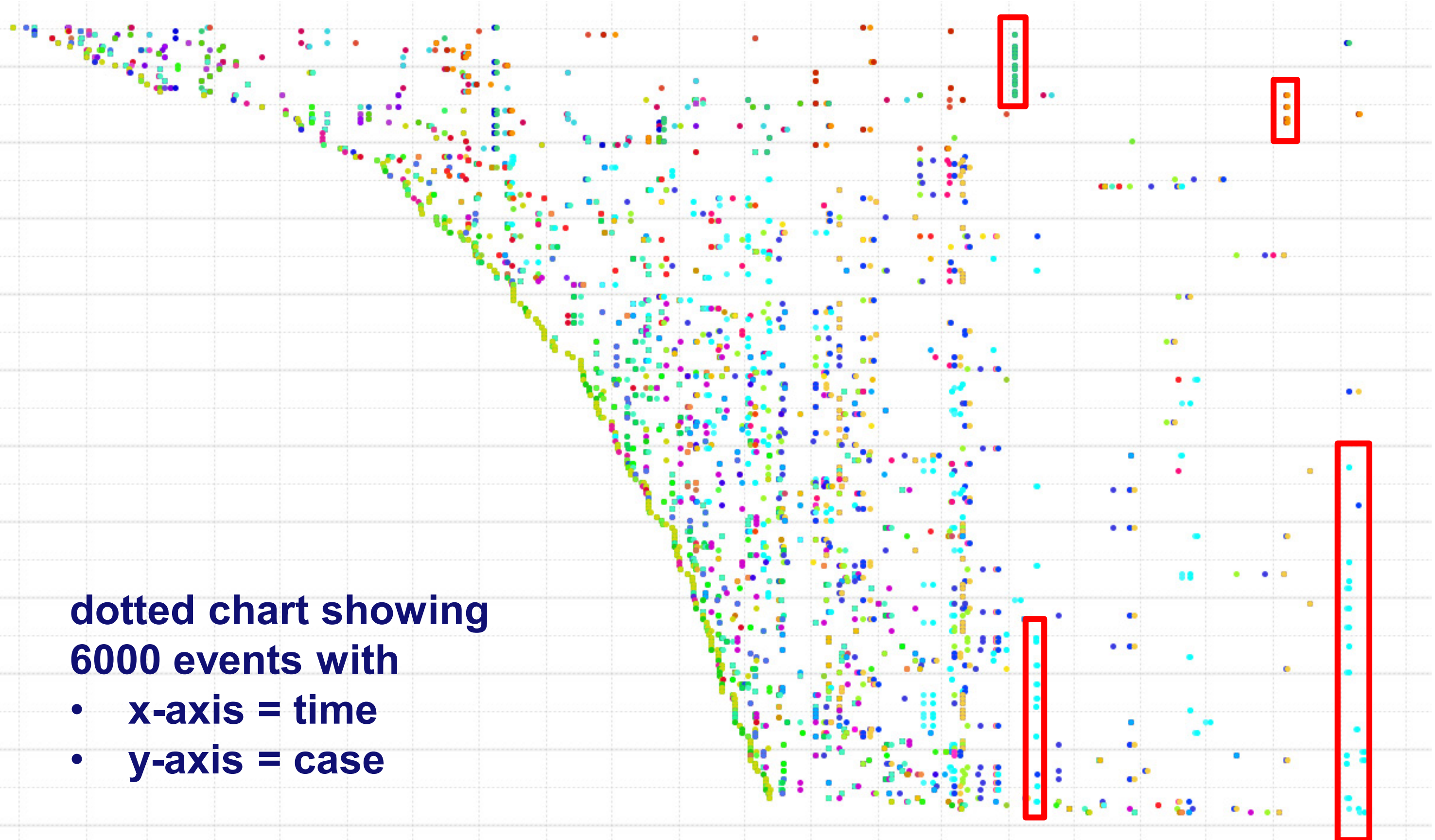


music-timeline.appspot.com

Based on album and artist statistics aggregated from Google Play Music

Chair of Process and Data Science

Classical Visualization Examples

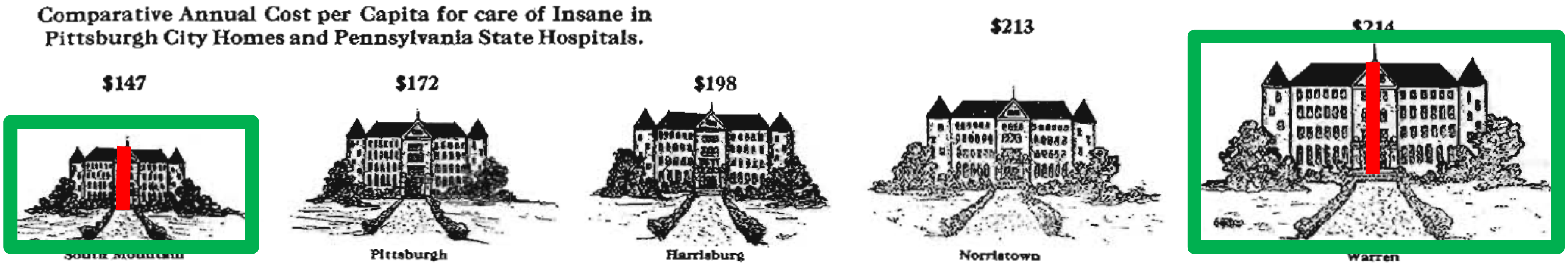


dotted chart showing 6000 events with

- x-axis = time
- y-axis = case

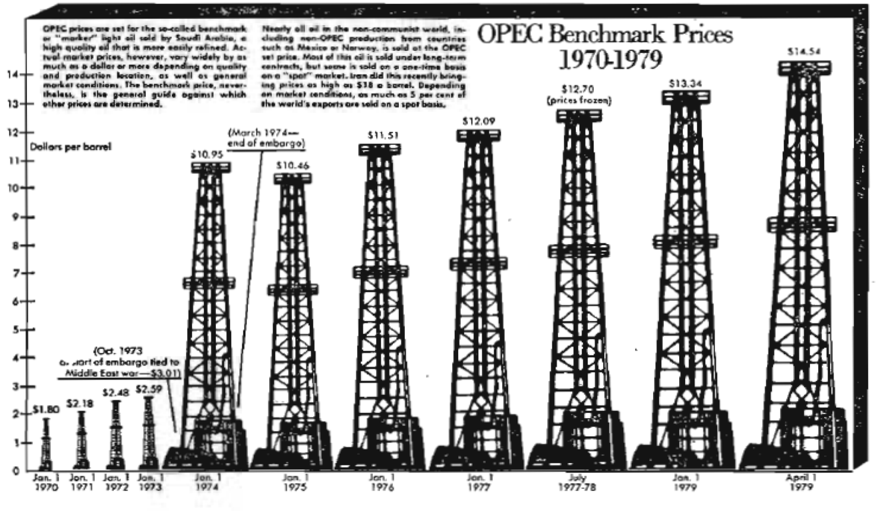
Classical Visualization Examples

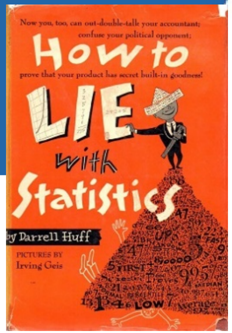
Misleading magnitudes.



Pittsburgh Civic Commission, *Report on Expenditures of the Department of Charities* (Pittsburgh, 1911), p. 7.

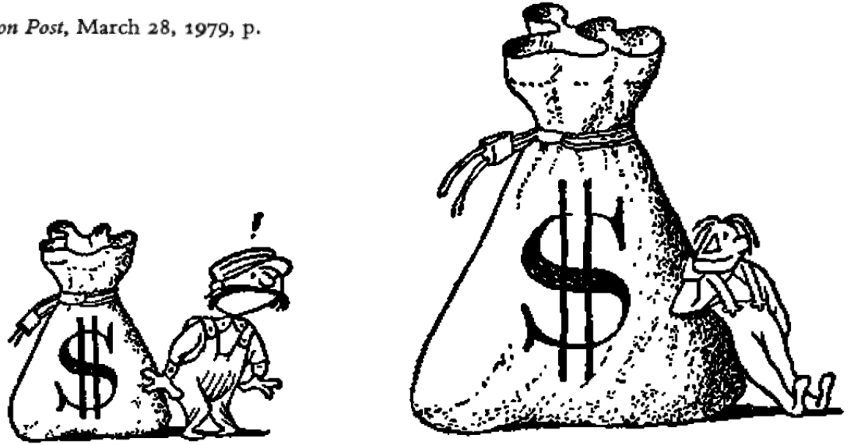
Classical Visualization Examples





Same problems

Washington Post, March 28, 1979, p. A-18.



The trick

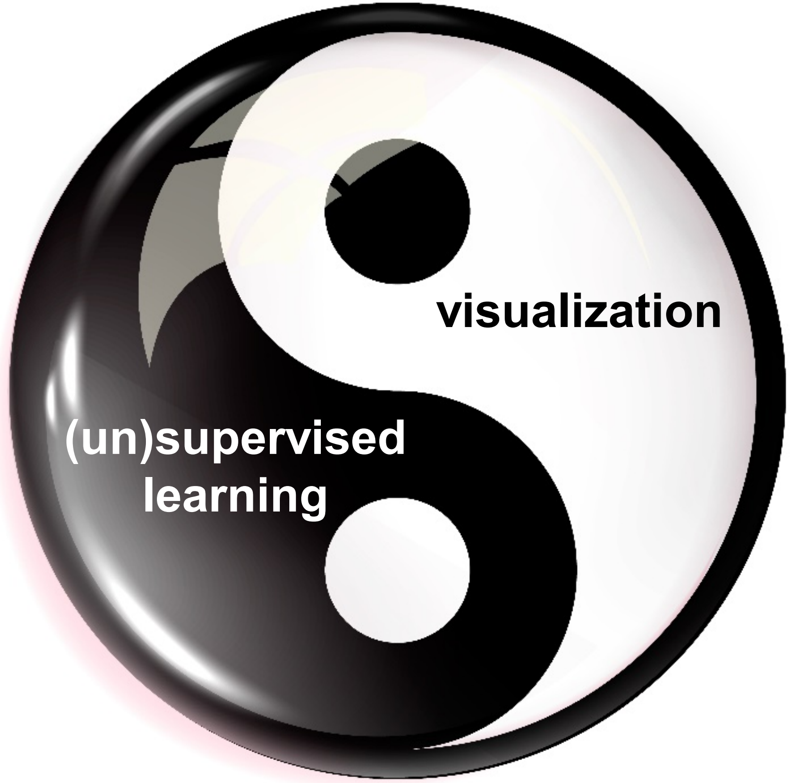
h

$\pi r^2 h$

$2h$

$$\pi(2r)^2 2h = 8\pi r^2 h$$

Complemening



Chair of Process and Data Science

Introduction

- <u>Our goal:</u> Extract relevant information from the flood of data to fulfill a certain purpose (e.g. support decision making processes in companies)
- Challenges:
  - Big, messy, inconsistent data
  - Extract information *quickly*
  - State of the art analysis software might be outdated and cannot handle these big amounts of data
- *Visual analytics* aims to solve these challenges

Chair of Process and Data Science

Visual Analytics is more than Visualization

Visualization

- Usually either handles natural geometric structures (e.g. MRI data) or abstract data structures (e.g. trees and graphs)
- Solely focuses on the representation of the data
- No to little interaction possibilities
- Not supported by learning techniques
- Main goals: presentation, confirmatory analysis or exploratory analysis

Visual Analytics is more than Visualization

Visual Analytics

- Basic idea:
  - Handles heterogeneous data from multiple sources
  - Combining *computational capabilities*, *statistics* and *mathematics* with *human skills* as perceiving relations and drawing conclusions
  - Benefits from information visualization in combination with human interaction
- Advantages:
  - Provides *overview* of the data
  - Provides possibility to explore the data
  - *Reduces complexity* of certain tasks
  - Helps to make decisions in *time-critical* situations

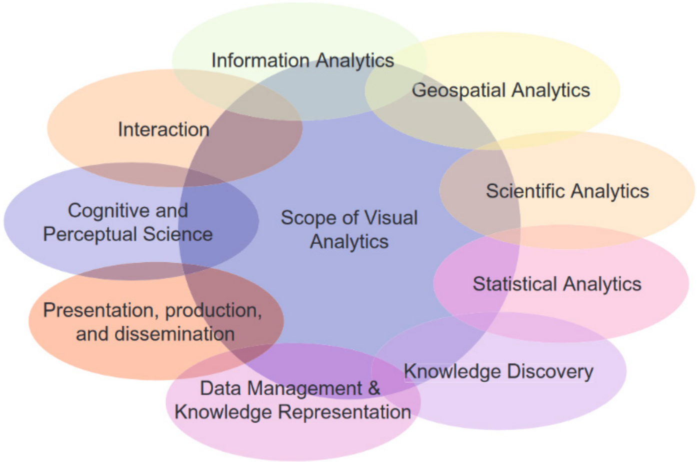
"Turn information overload into an opportunity"

Chair of Process and Data Science

Visual Analytics Scope and Tasks

Visual Analytics - Scope

- **Visual Analytics combines many different fields which can be sorted into four major categories**
  - ***Data representation* and *transformation***
  - ***Visual representation* and *interaction* techniques**
  - ***Production* and *presentation* of results to communicate the information appropriately**
  - ***Analytical reasoning techniques* which help the users to get insights and support decision making etc.**

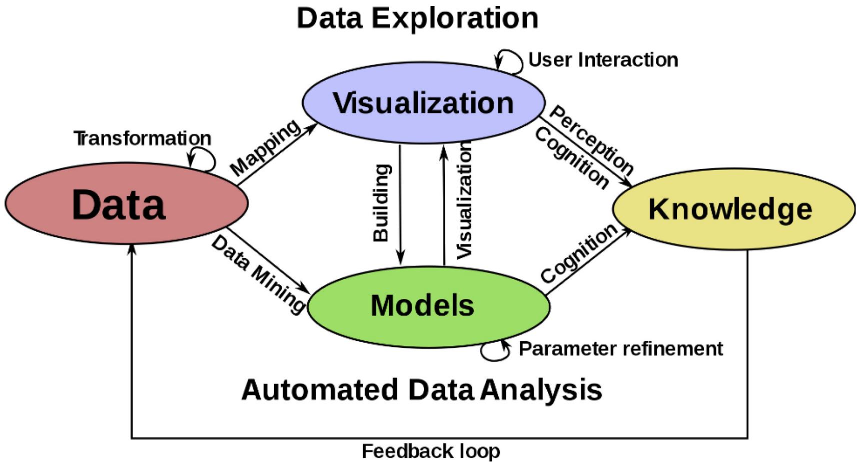


D. Keim et. al, Visual Analytics Scope and Challenges in: Visual Data Mining, Springer, 2008

Visual Analytics - Process

- Overall goal is to obtain *knowledge* from the given data
- The *data* needs to be preprocessed/transformed to be suitable for the analysis (see previous lecture)
- The *models* are obtained by applying data mining techniques to the data which can then be refined based on the visualization
- The *visualization* provides insights and interaction possibilities to the user

Key idea: combine visualization with model learning



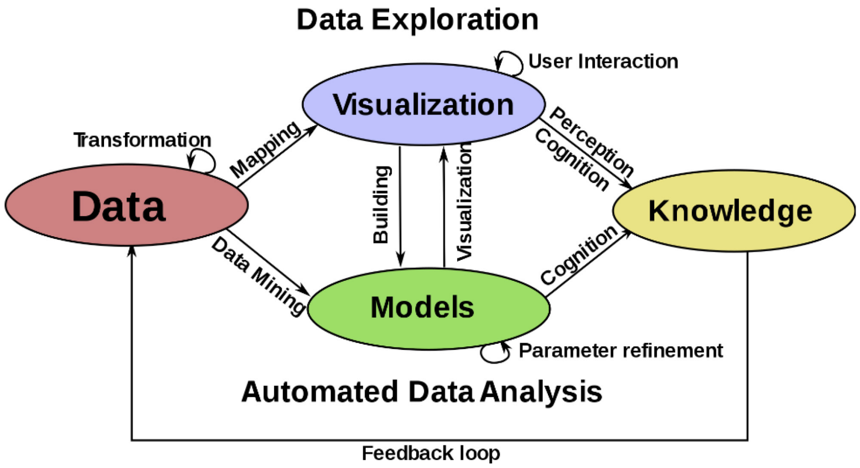
D. Keim et. al, Solving problems with visual analytics, Eurographics Association, 2010

Chair of Process and Data Science

Visual Analytics - Process

To obtain the best results a steady interchange between automated data mining techniques and manual user interactions is required

"Analyze First – Show the Important – Zoom, Filter and Analyze Further – Details on Demand"

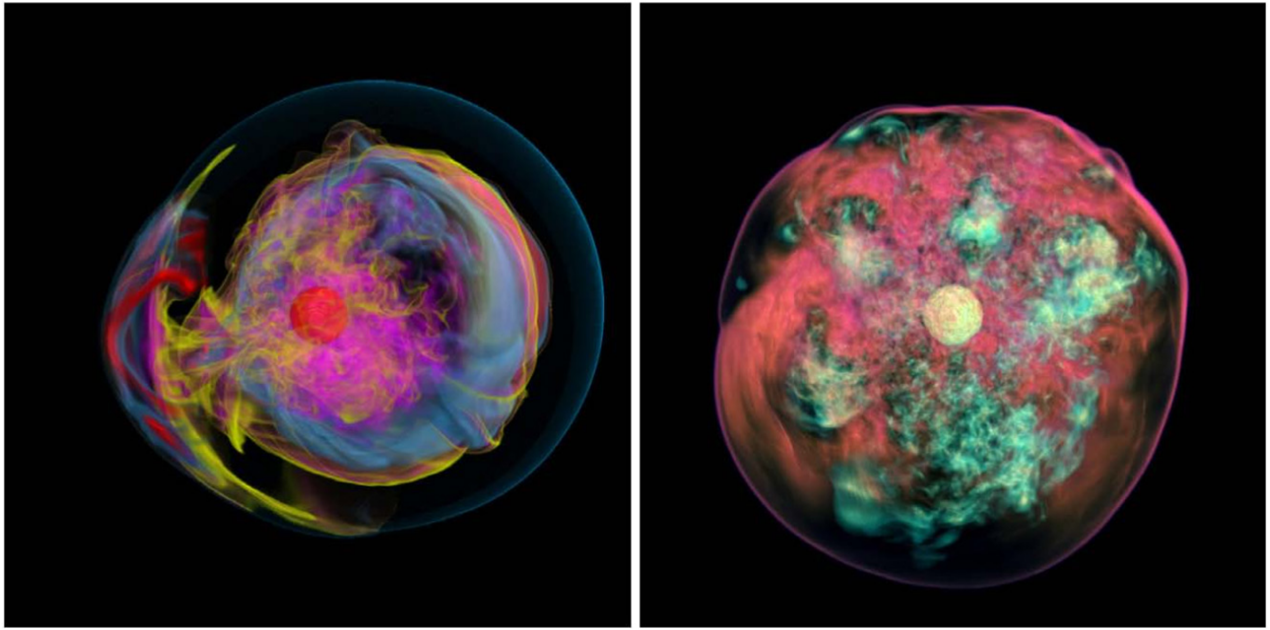


D. Keim et. al, Solving problems with visual analytics, Eurographics Association, 2010

Visual Analytics Examples

Visual Analytics Examples

- The Terascale Supernova Initiative Project runs complex simulations to gain a complete understanding of core collapse supernovas
- Complex relationships inside the supernova (e.g. rotation, radiation, magnetic fields etc.) produce tens of terabytes of data per simulation
- Data is only understandable after visualizing it.

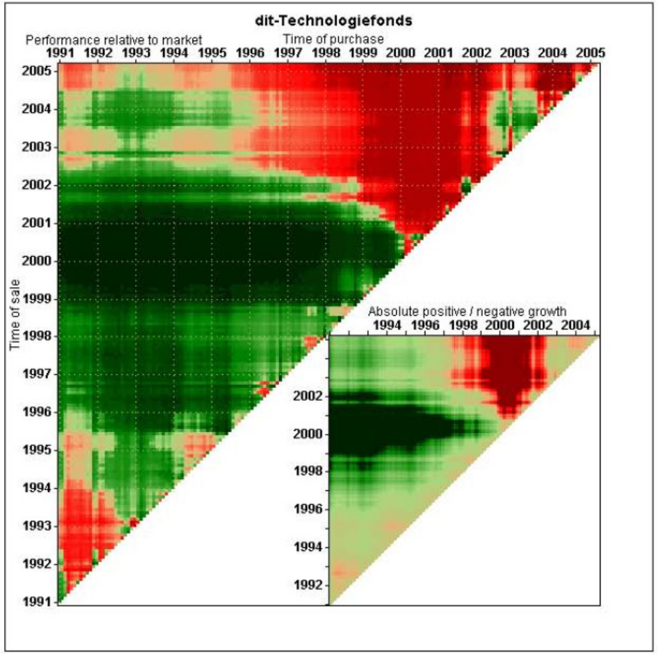


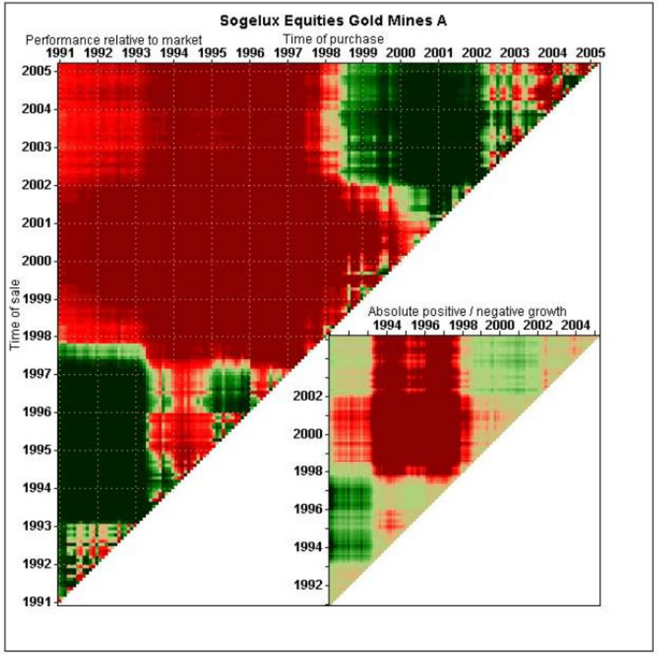
Visual Design and Analysis: Abstractions, Principles and Methods by T. Munzner

Chair of Process and Data Science

Visual Analytics Examples

- **Visual performance analysis of financial data.**
- **Small triangle shows the absolute performance of one stock and the big one compares one stock to the entire market.**
- **Red represents negative growth and green represents positive growth.**



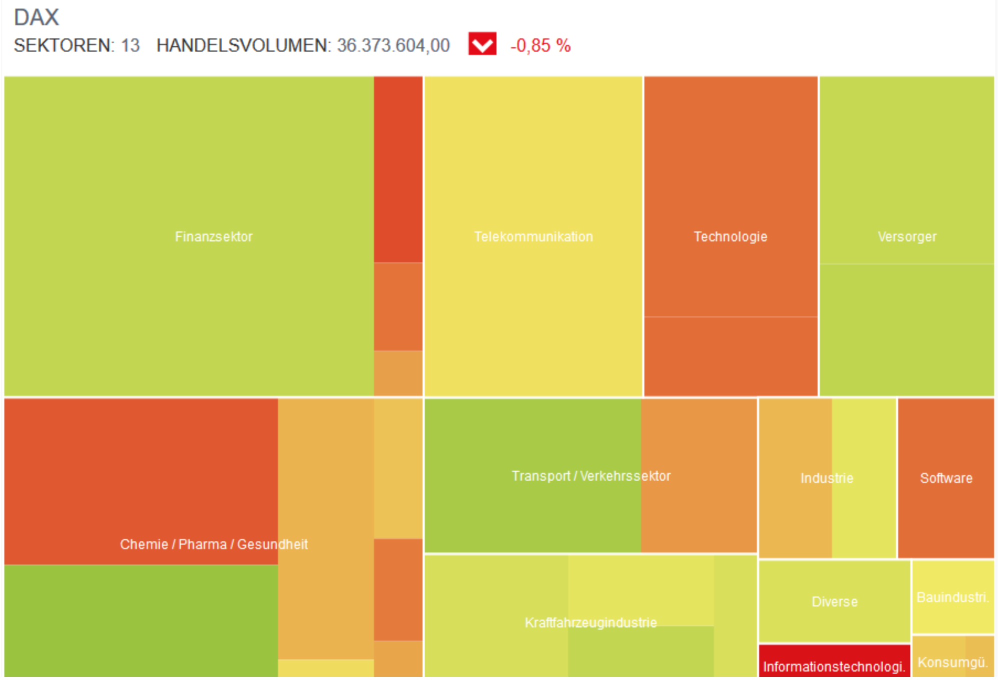


D. Keim et. al, Visual Analytics Scope and Challenges in: Visual Data Mining, Springer, 2008

Chair of Process and Data Science

Visual Analytics Examples

- **Website to visualize the daily development of companies and sectors on the stock market with several interaction options**
- **Red color indicates bad performance and green good performance**
- **In this example: Size of each sector is determined by sales volume and each sector is split further into colored rectangles which correspond to sub segments of the sector**

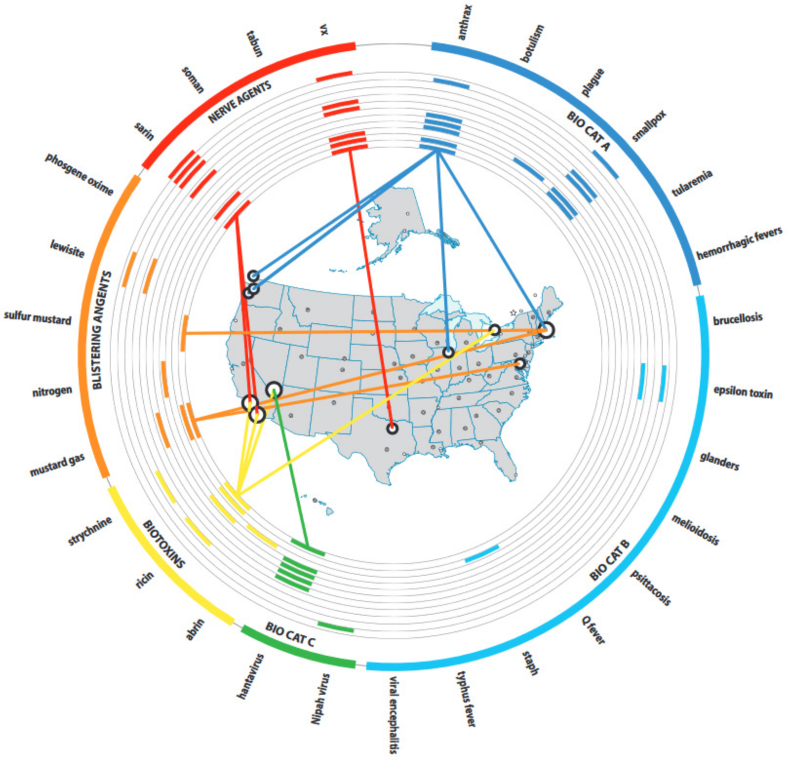


https://www.maxblue.de/maerkte-analysen/boersen-kurse/marketmap.html

Chair of Process and Data Science

Visual Analytics Examples

- VisAware for BioWatch system detects biological and chemical attacks in the US
- Different categories of biological and chemical agents are shown in the outer colored ring
- Concentric rings represent sequential time samples and show the presence of the agent over time with the outer rings being older time samples
- The map shows where sensors are set up
- The width of the connecting line between sensor location and time ring shows the probability of an actual attack



D. Keim et. al, Visual Analytics Scope and Challenges in: Visual Data Mining, Springer, 2008

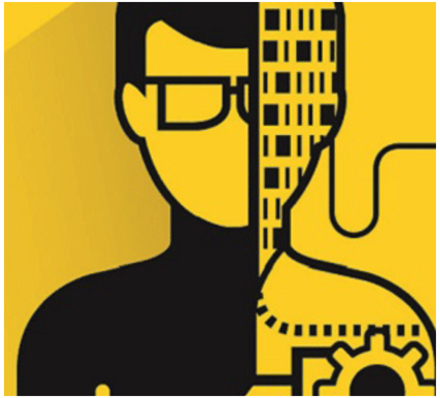
Visual Analytics Challenges

Visual Analytics Challenges

- Problem solving, decision science and human information discourse
- User acceptance
- Synthesis of problems
- Data quality (see previous lecture)
- Scalability
- Integration
- Evaluation

Problem solving, decision science and human information discourse

- **Overall goal: Problem solving, decision making supported by technology**
- **Two main requirements:**
  - **Understanding of the technology**
  - **Capabilities like reasoning and comprehension of logic**
- **To achieve this the visual analytics software should present analytic results through *meaningful* visualization and *clear* representation**
- **But what does meaningful and clear mean?**
- **Might depend greatly on the context**



Chair of Process and Data Science

User acceptance

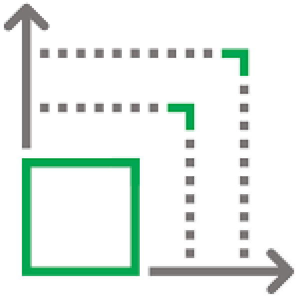
- General problem when introducing new tools/systems to an environment
- People are used to their routines and do not want to change them
- Using a new tool/system means changing the routine
- If people do not use the tool, a lot of time and money was possibly wasted
- Presenting the benefits of new visual analytics tool needs to be an important part of the deployment process

Synthesis of problems

- Real-world problems are often more complex and consist of several sub-problems
- Solving one sub-problem by itself is often already challenging but doable
- Correlation between sub-problems makes it challenging to solve the entire problem

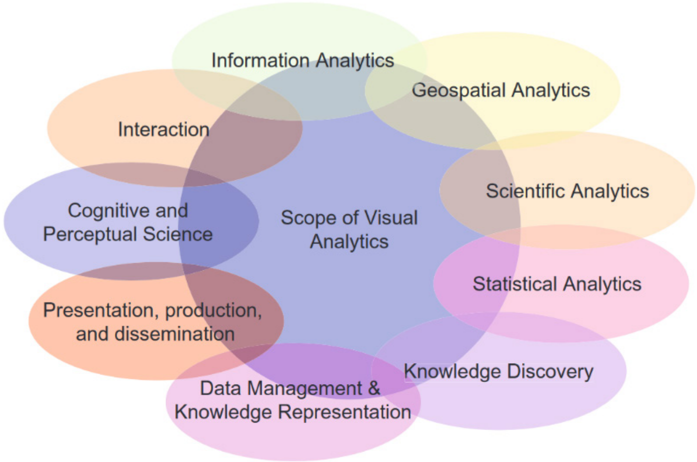
Scalability

- Over time storage capacities have improved
- Only storing more data does not help if the analysis methods can not handle more data
- Visual analytics methods need to be able to handle data sets of increasing size



Integration

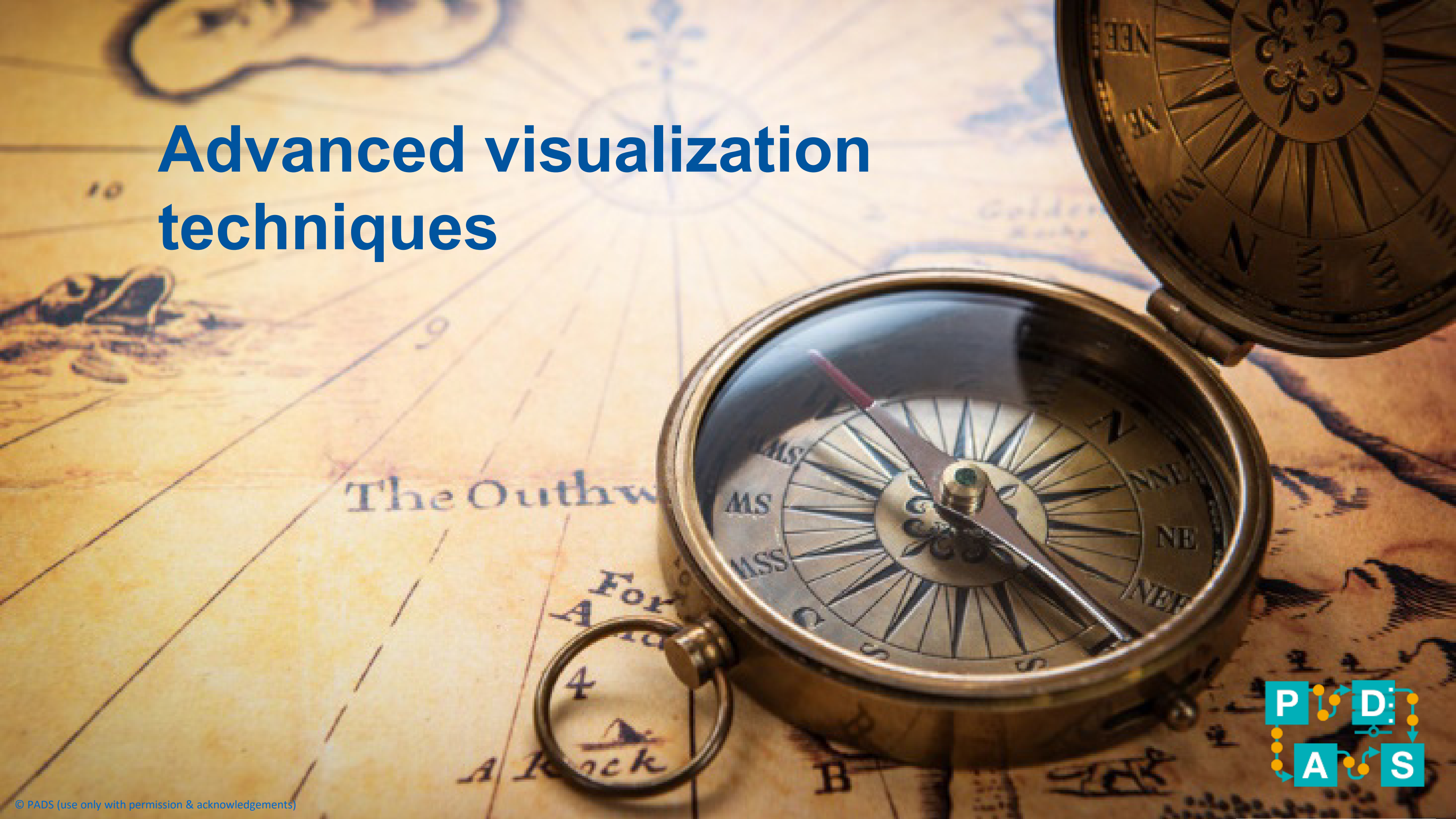
- **Most fields in the scope of visual analytics, when considered by themselves, are represented by a certain system or tool**
- **To achieve the visual analytics goal all these topics need to be integrated into one visual analytics tool/system**
- **Designing and implementing such an integrated environment is challenging**



D. Keim et. al, Visual Analytics Scope and Challenges in: Visual Data Mining, Springer, 2008

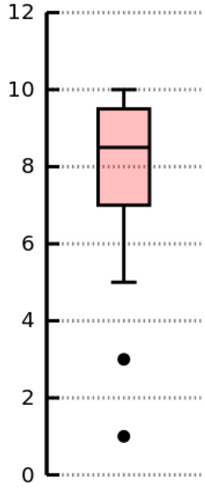
Chair of Process and Data Science

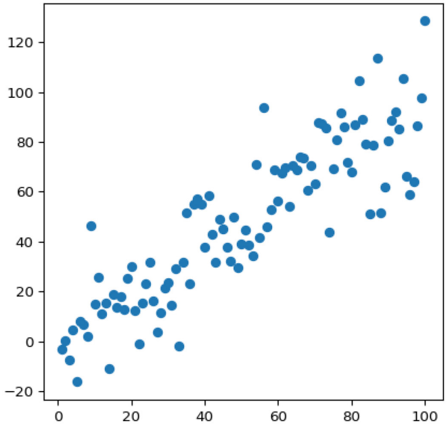
Advanced visualization techniques

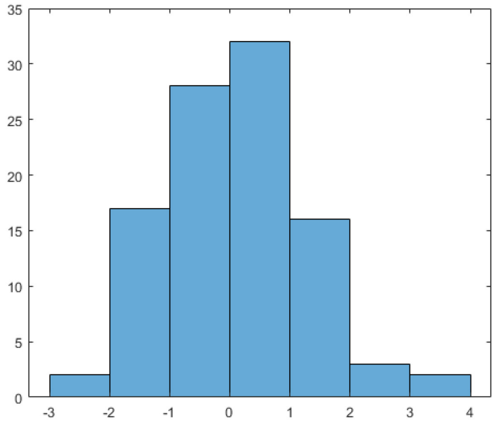


Advanced visualization techniques

In lecture 3 we have already seen basic visualization techniques





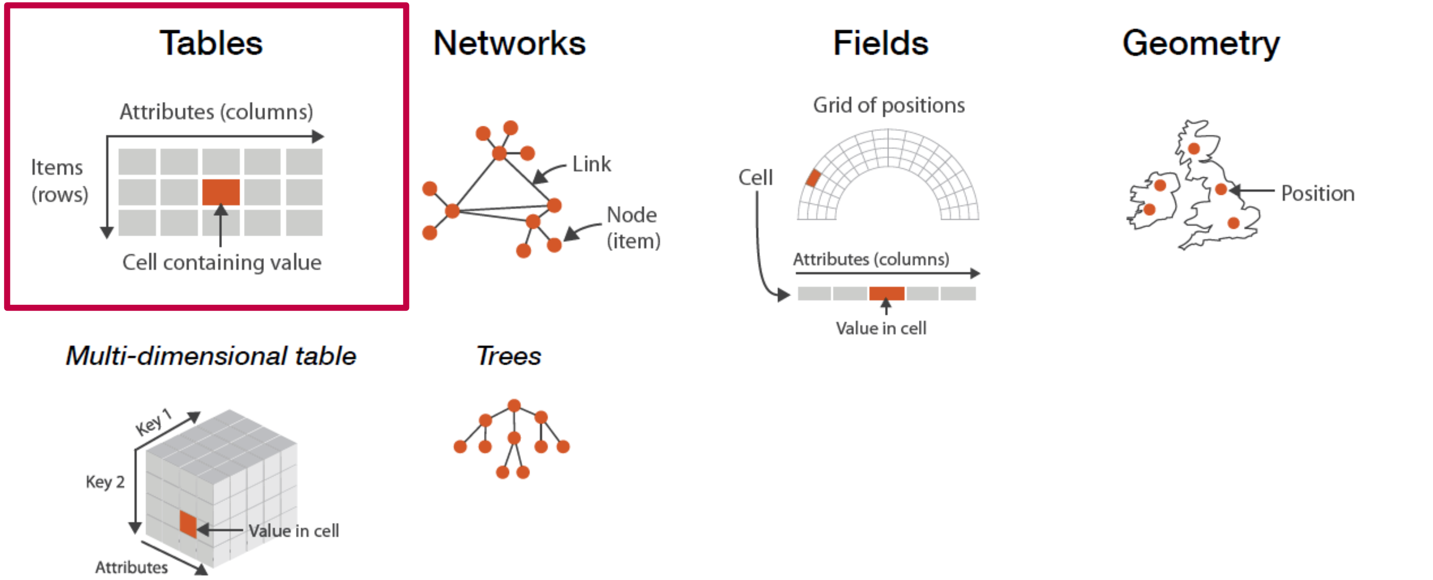


But they might not be enough for advanced analysis.

Chair of Process and Data Science

Advanced visualization techniques

Especially for more complex datasets and data types we need different visualization techniques

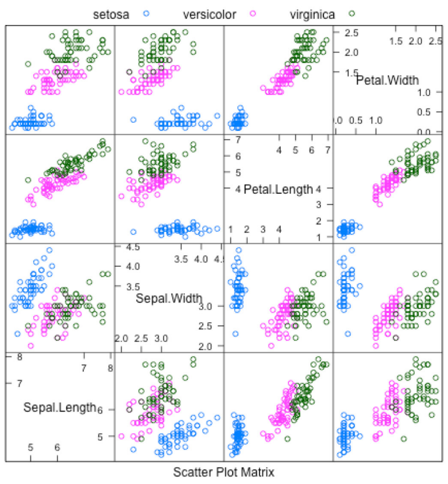


We focus on tabular data in this lecture!

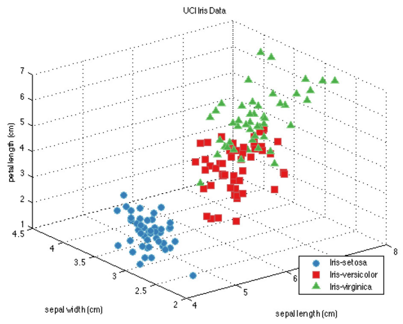
Chair of Process and Data Science

Advanced visualization - Scatter plot matrix

- Shows a scatter plot for each combination of attributes in a dataset with multiple attributes
- Useful to detect correlations, trends and outliers
- Advantage: Easier to interpret and understand than 3-D scatter plot
- Disadvantage: Quickly grows in size as the number of attributes in the data set grows



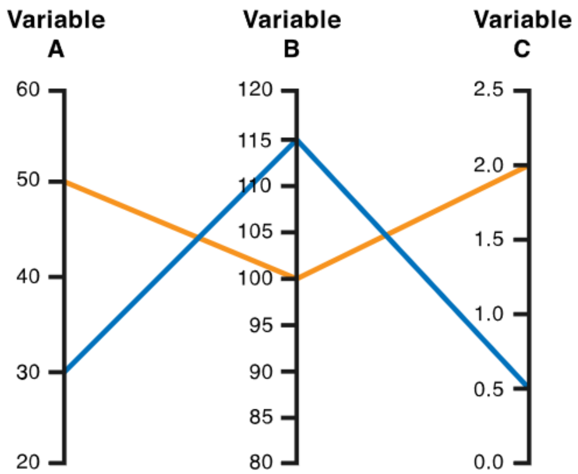
Scatter plot matrix for iris data set using all four attributes



3D Scatter plot matrix for iris data set using three attributes

Advanced visualization - Parallel coordinate plot

- Display each attribute as parallel axes
- Plot the data points for each attribute on the corresponding axis
- Connect all points belonging to one data entry

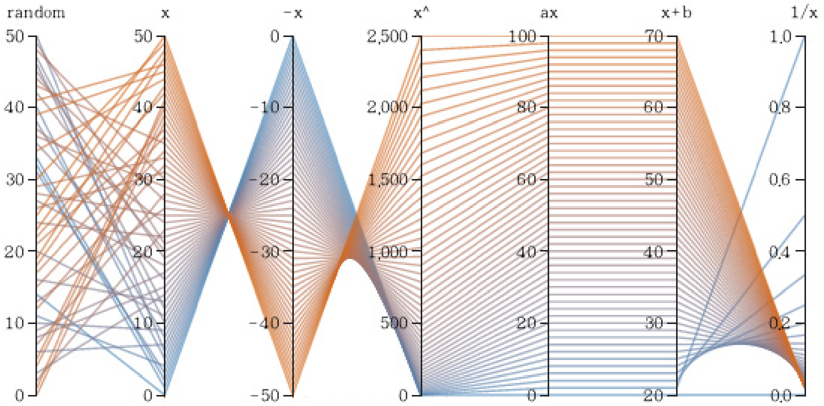


A	B	C
50	100	2.0
30	115	0.5

https://datavizcatalogue.com/methods/parallel_coordinates.html

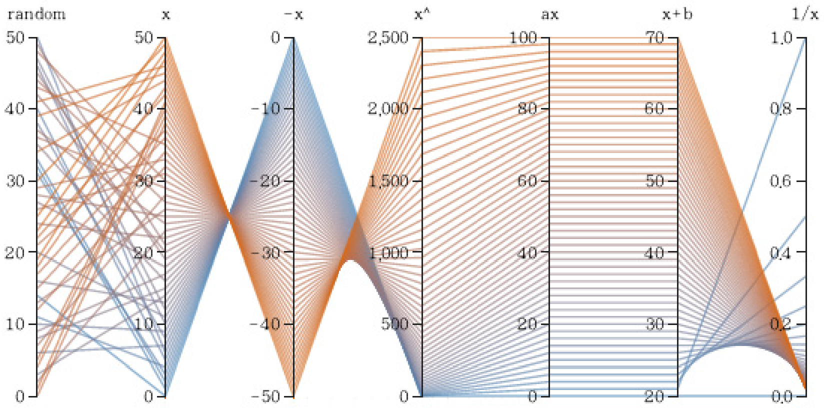
Advanced visualization - Parallel coordinate plot

- Interpretation
  - Positive Correlation: Cluster of similar lines
  - Negative Correlation: No cluster but similar crossing point of axis
  - Outliers: Lines isolated from clusters and lines with different paths



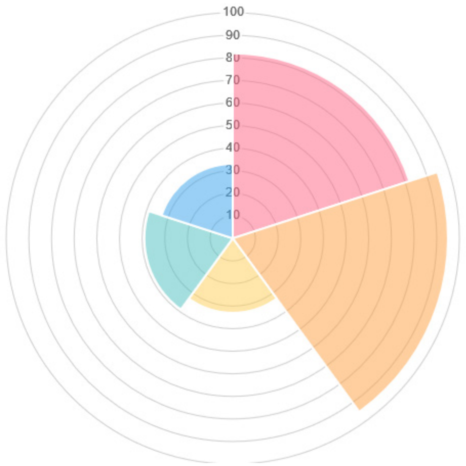
Advanced visualization - Parallel coordinate plot

- Can display a large number of data entries
- Grows rapidly in size as the number of attributes in the data set grows
- How to arrange the variables next to each other? Does the order make a difference for the analysis?



Advanced visualization - Polar area chart

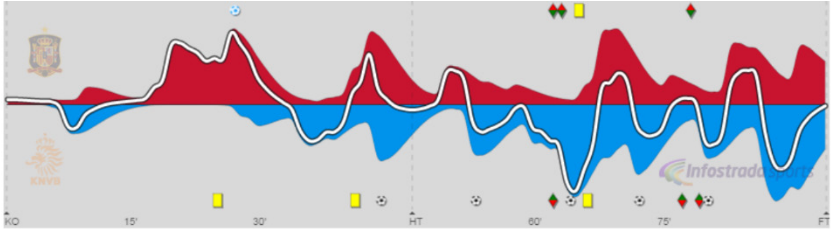
- **Variation of the classic pie chart**
- **Different to pie charts each segment has the same angle**
- **The radius of the segment displays the value of the attribute**



Red	Orange	Yellow	Green	Blue
82	95	33	39	33

Advanced visualization – Stream graphs

- Displays the change in the data over time of different categories
- Plots one ordered key attribute on the x-axis
  - Usually time
- Each segment of the stream represents one category
- Width of stream at point x is determined by the value of the category at time x
- Additional discrete events can be shown (card, goal, free kick, etc.)



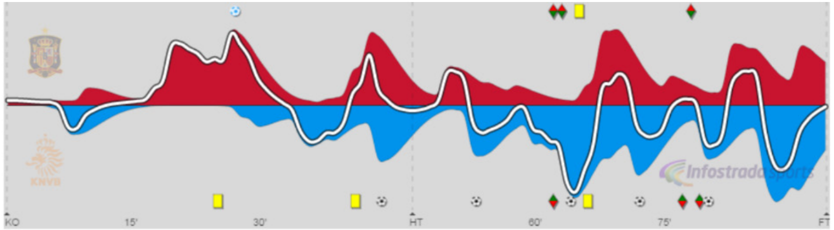
Stream graph visualizing the balance of play for the world cup match between Spain and the Netherlands.

Red area: The strength of the Spanish team to.
Blue area: The strength of the Dutch team to.
White line: Balance of the game between both teams.

M. Grootjans, Visualization of spatio-temporal soccer data, 2014, Master thesis at Eindhoven University of Technology

Advanced visualization – Stream graphs

- **Advantages:**
  - Good for displaying big datasets
  - Good to discover trends and patterns over time (seasonal changes etc.)
- **Disadvantages:**
  - Might get very cluttered with large data sets
  - Categories with small values are difficult to spot
  - Not possible to obtain precise values for the categories from the stream



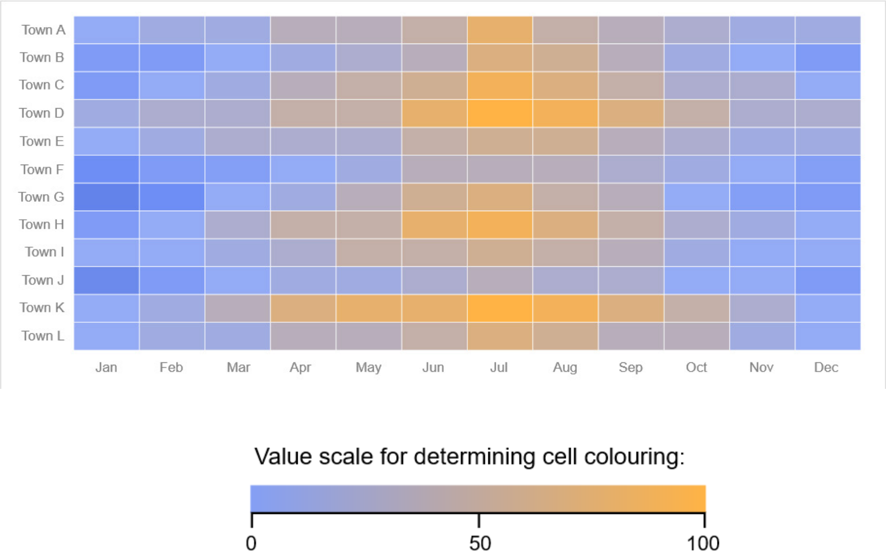
Stream graph visualizing the balance of play for the world cup match between Spain and the Netherlands.

Red area: The strength of the Spanish team to.
Blue area: The strength of the Dutch team to.
White line: Balance of the game between both teams.

M. Grootjans, Visualization of spatio-temporal soccer data, 2014, Master thesis at Eindhoven University of Technology

Advanced visualization - Heatmap

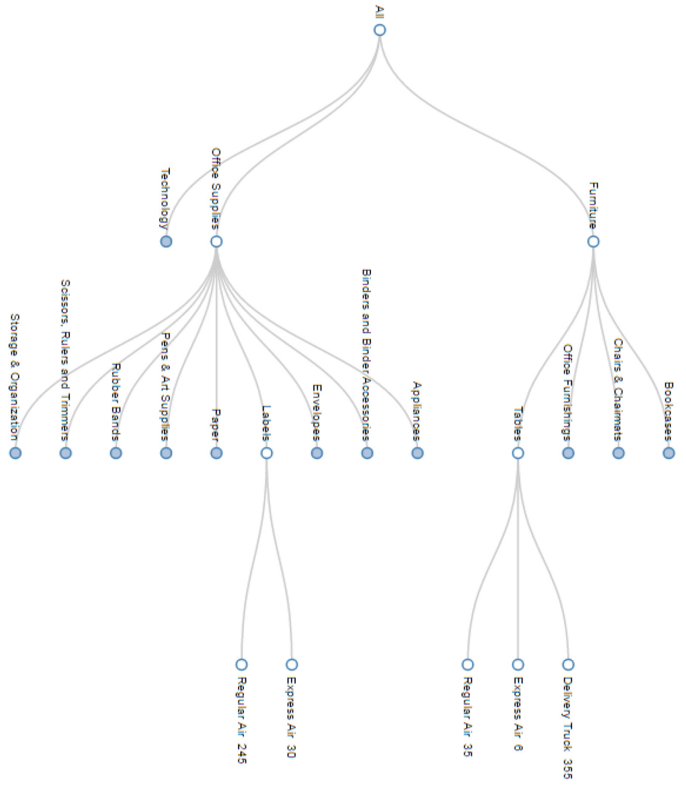
- **Displays the data set by plotting two categorical attributes on the axis and coloring the segments according to a quantitative attribute**
- **Used to find clusters and outliers**
- **Is there a meaningful order for the categorical values on the x- and y-axis?**



https://datavizcatalogue.com/methods/heatmap.html

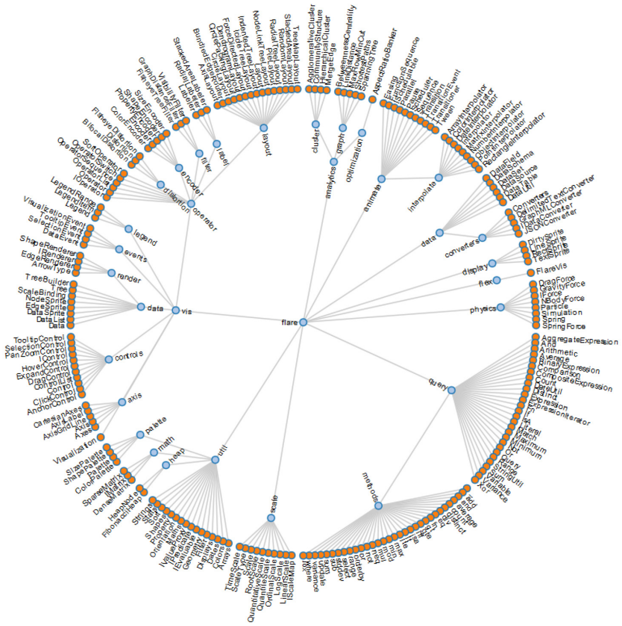
Advanced visualization - Node-link tree

- **Used to represent hierarchical data**
  - **Data points are shown as nodes**
  - **Hierarchical relations are shown as edges**
  - **Root on top and all branches below**
- **Very effective in visualizing tree structures**
- **Tree can get very wide quickly**
- **Low level data points can create a lot of unused white space**



Advanced visualization – Radial Node-link tree

- **Radial variation of the node-link tree**
- **Root in the middle and the hierarchy spreads around it in a radial layout**



http://bl.ocks.org/mbostock/4063550

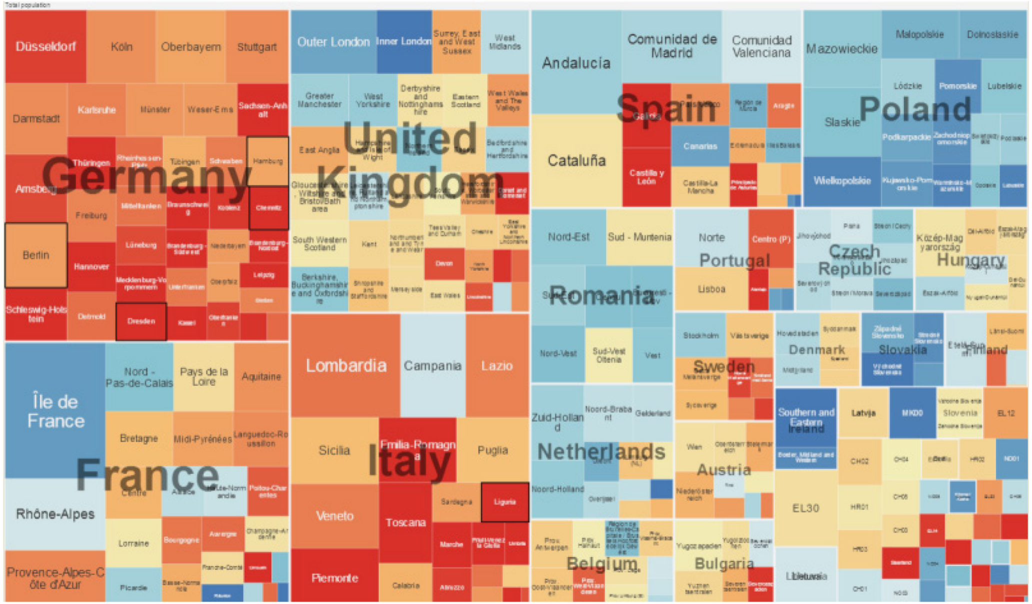
Advanced visualization – Sunburst

- Similar principle to the radial node tree
  - Each ring is a different level of hierarchy with the center being the root
  - Each segment of a ring belongs to one categorical value
  - The size of a segment is either divided equally under the parent node or proportional to a value

https://datavizcatalogue.com/methods/sunburst_diagram.html

Advanced visualization – Tree map

- **Displays a tree as nested rectangles**
  - **Each parent node of the tree is presented as a rectangle**
  - **Each rectangle is divided into smaller rectangle according to the parent's child nodes**
  - **The size of the rectangles is determined by a chosen attribute of the data set**



Tree map showing Elderly population in Europe per country and its regions. The size of the rectangle represents the total population and the colour shows the amount of people above 65 years.
https://ncva.itn.liu.se/education-geovisual-analytics/treemap?l=en

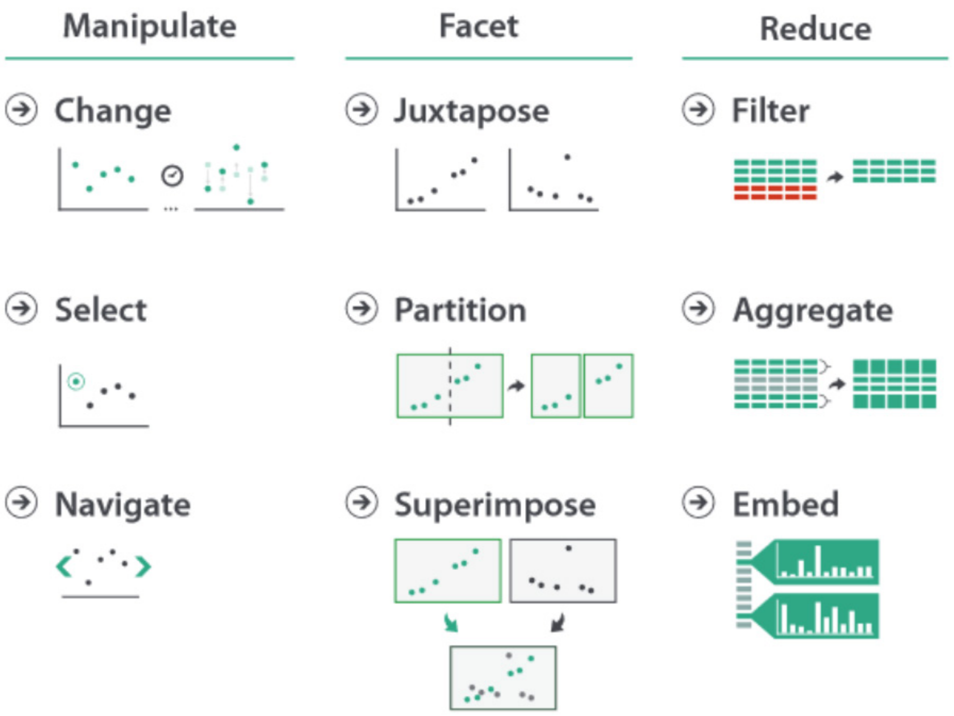
Play with it …

http://mitweb.itn.liu.se/GAV/dashboard/#story=data/nuts-regions-ageing-population-in-europe-2010.xml&layout=[map,treemap]

User interaction

User interaction

- As mentioned before, the user should be able to interact with the presented data
- Three categories of user interaction
  - Manipulate
  - Facet
  - Reduce



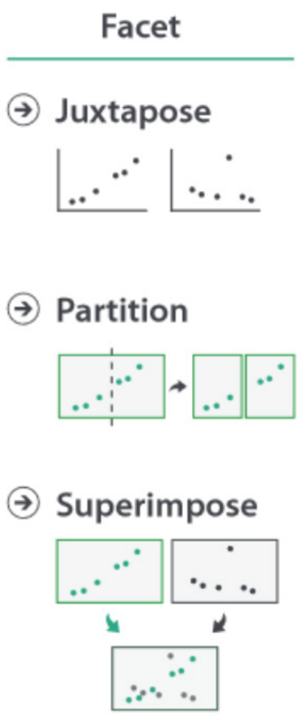
T. Munzner, Visual Design and Analysis: Abstractions, Principles and Methods

User interaction - Manipulate

- Manipulate the visualization to gain meaningful and relevant insights
  - Change the data
  - Select certain parts of the data
  - Navigate through the data
- In comparison to the “reduce” category these interactions just manipulate what is presented (the view) instead of the data behind it

User interaction - Facet

- **Obtain a different view of the data to be able to present/compare relevant parts of the data**
  - **Juxtapose: Putting two views next to each other**
  - **Partition: Split the view**
  - **Superimpose: Overlay two views**



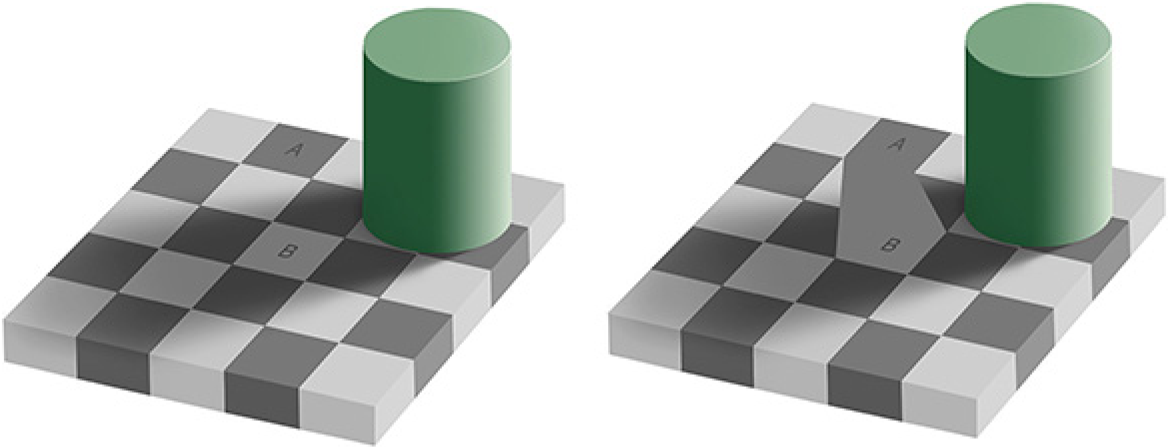
User interaction - Reduce

- **Reduce the data and consequently alter its visualization to only display the relevant part and the correct level of detail**
  - Filter
  - Aggregate
  - Embed

Visual encoding principles

Wrong visualization choices can influence what the user perceives

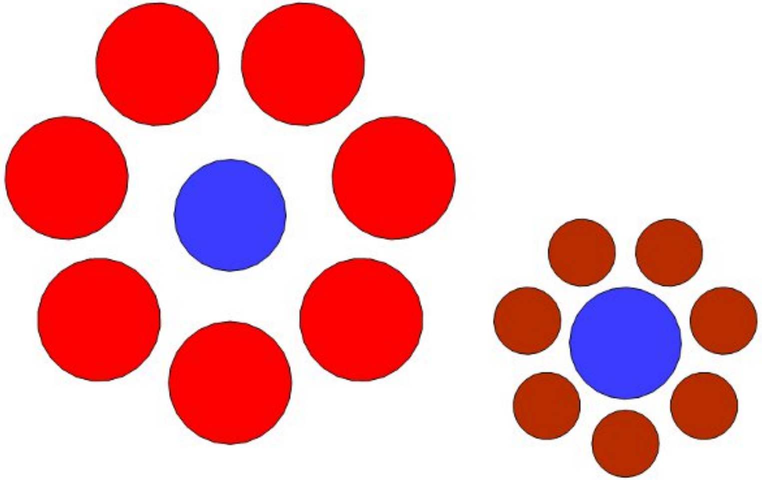
Tile A and B have the same colour



Checker shadow illusion by Edward H. Adelson (1995)

Wrong visualization choices can influence what the user perceives

Both blue dots have the same size

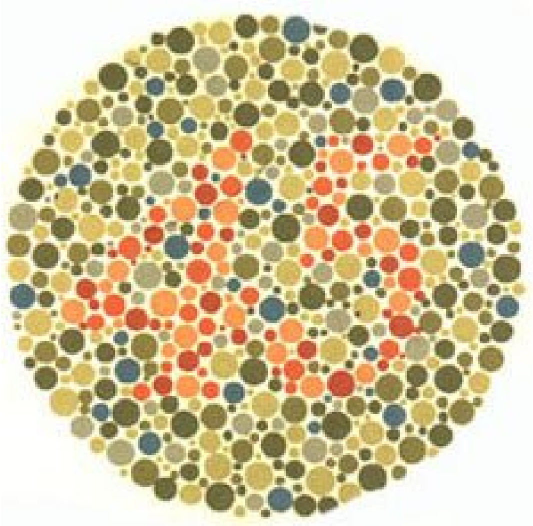


Ebbingshaus illusion by Hermann Ebbinghaus (1901)

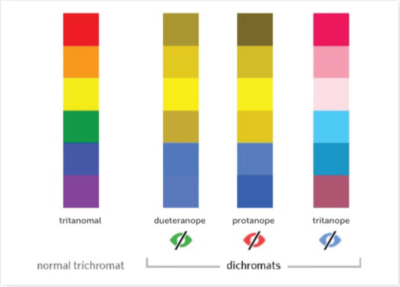
Chair of Process and Data Science

Wrong visualization choices can influence what the user perceives

Colour blind people might not be able to read the displayed number

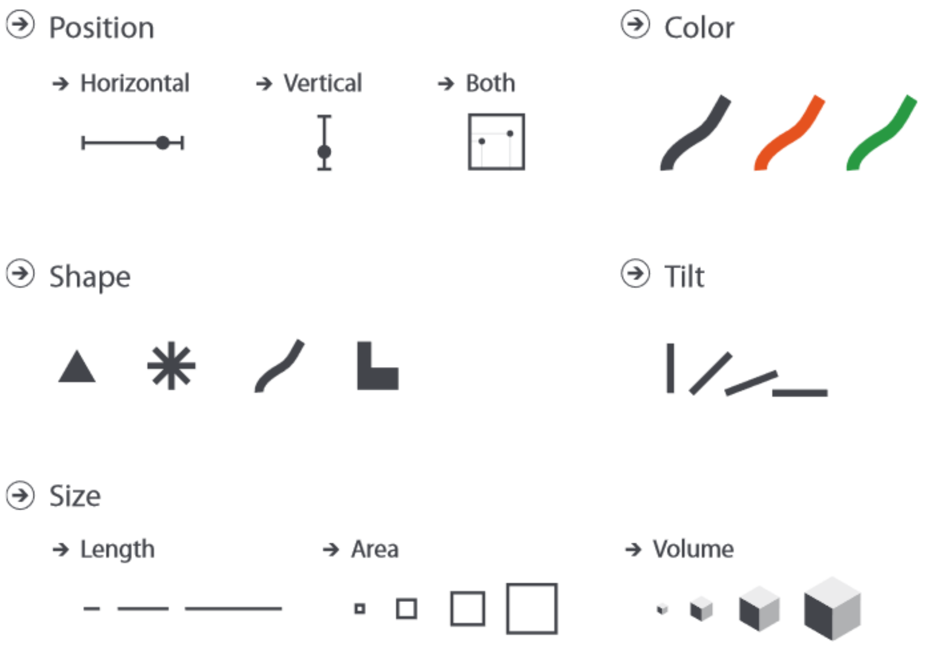


Ishihara color test by Shinobu Ishihara (1917)



Visual encoding principles

- **There are different channels we can use to display our data points and differences between them**
  - **Position**
  - **Shape**
  - **Color**
  - **Tilt**
  - **Size**

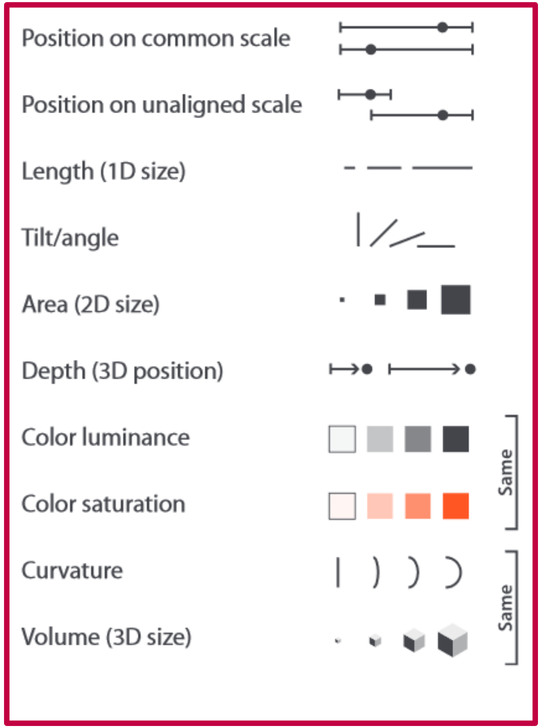


T. Munzner, Visual Design and Analysis: Abstractions, Principles and Methods

Visual encoding principles

- The definition of the channels can be refined and grouped into two groups
  - Magnitude channels
  - Identity channels
- Magnitude channels should be used for *ordered* attributes
- Identity channels should be used for *categorical* attributes

Magnitude Channels

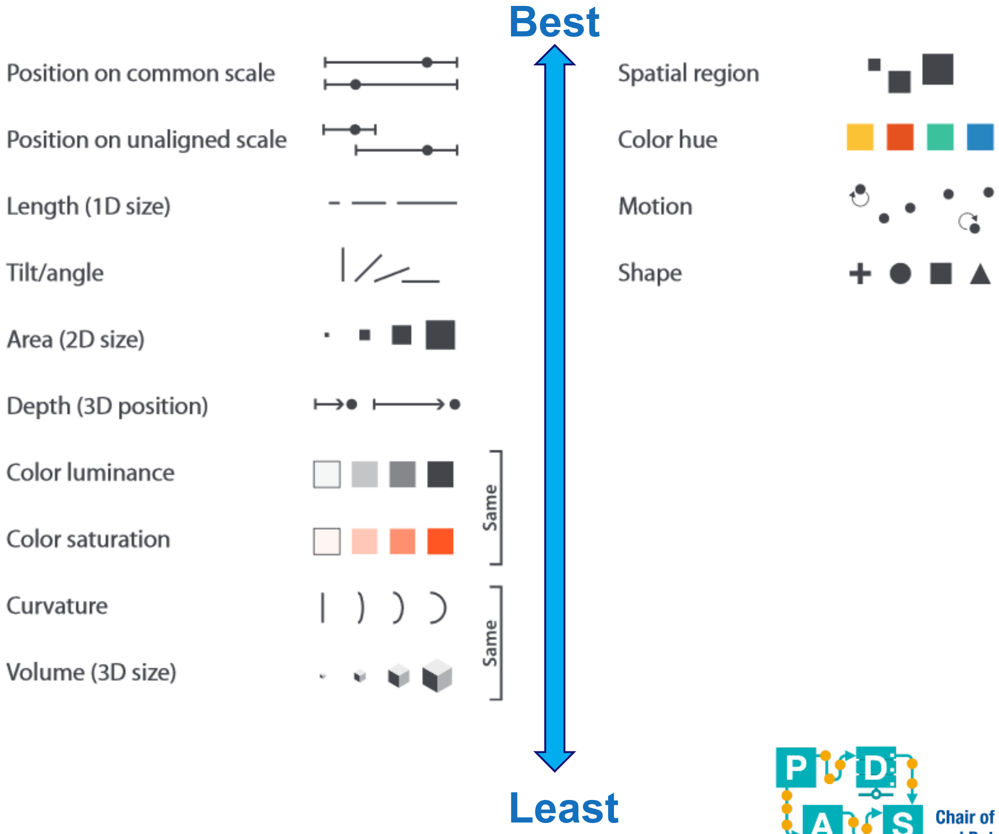


Identity Channels

Chair of Process and Data Science

Visual encoding principles

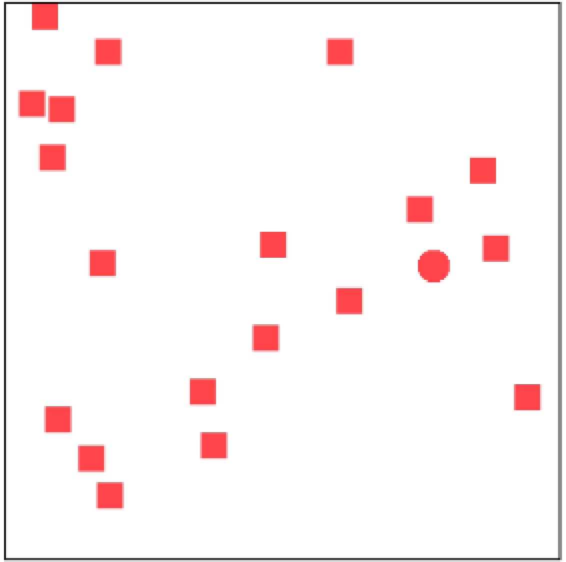
- The channels can be ranked in their effectiveness of expressing differences
- Rule of thumb: encode most important attributes with highest ranked channel

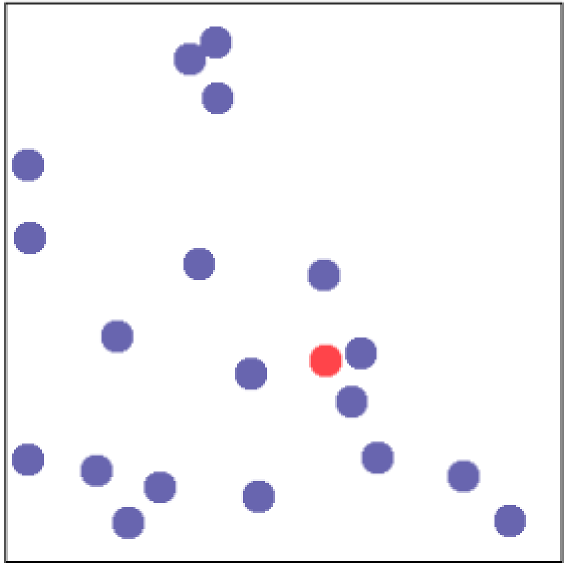


Chair of Process and Data Science

Visual encoding principles

- **Which plot visualizes the difference between its data points better?**

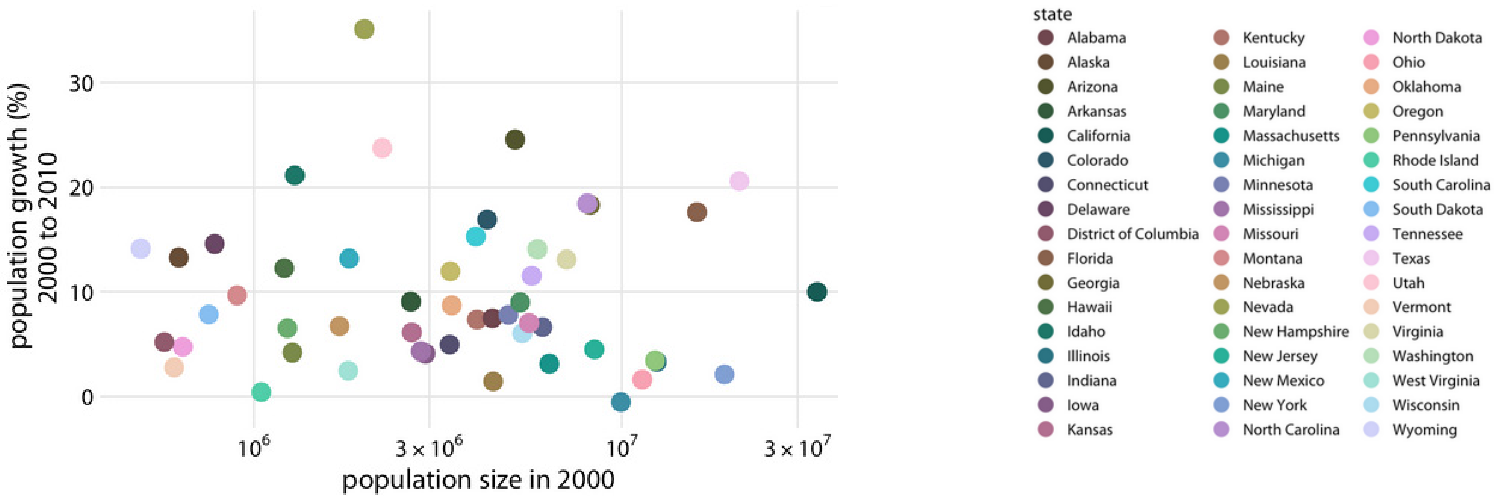




T. Munzner, Visual Design and Analysis: Abstractions, Principles and Methods

Visual encoding principles

- **Which colored dot belongs to which state? The colors are not distinguishable.**



Chair of Process and Data Science

Conclusion

Short summery of lecture

- Visualization is a good way to display relations in the data
- Combined with manual interaction and human perception skills insights into the data can be gained quicker
- Visual analytics aims to support the problem solving and decision making process
- But it's not as easy as it seems
  - What to display?
  - How to display it? Which visualization technique is appropriate?
  - Does the visualization influence the way results are interpreted?

Relevant Literature

- Visual Analytics Scope and Challenges by D. Keim et. al in: Visual Data Mining
- Mastering the Information Age – Solving Problems with Visual Analytics by D. Keim et. al
- Visual Design and Analysis: Abstractions, Principles and Methods by T. Munzner

#	Lecture	date	day
Lecture 1	Introduction	10/10/2018	Wednesday
Lecture 2	Crash Course in Python	11/10/2018	Thursday
Instruction 1	*Python*	*12/10/2018*	*Friday*
Lecture 3	Basic data visualisation/exploration	17/10/2018	Wednesday
Lecture 4	Decision trees	18/10/2018	Thursday
Instruction 2			
Instruction 3			
Instruction 4			
Instruction 5			
Instruction 6			
Lecture 14	Process mining (unsupervised)	28/11/2018	Wednesday
Lecture 15	Process mining (supervised)	29/11/2018	Thursday
Instruction 7	*Process mining and sequence mining*	*30/11/2018*	*Friday*
Lecture 16	Text mining (1/2)	05/12/2018	Wednesday
Instruction 8	*Text mining and process mining*	*06/12/2018*	*Thursday !!*
Lecture 17	Text mining (2/2)	12/12/2018	Wednesday
Lecture 18	Data preprocessing, data quality, binning, etc.	13/12/2018	Thursday
Lecture 19	Visual analytics & information visualization	19/12/2018	Wednesday
backup		20/12/2018	Thursday
Instruction 9	*Text mining, preprocessing and visualization*	*21/12/2018*	*Friday*
Lecture 20	Responsible data science (1/2)	09/01/2019	Wednesday
Lecture 21	Responsible data science (2/2)	10/01/2019	Thursday
Instruction 10	*Responsible data science*	*11/01/2019*	*Friday*
Lecture 22	Big data (1/2)	16/01/2019	Wednesday
Lecture 23	Big data (2/2)	17/01/2019	Thursday
Instruction 11	*Big data*	*18/01/2019*	*Friday*
Lecture 24	Closing	23/01/2019	Wednesday
backup		24/01/2019	Thursday
Instruction 12	*Example exam questions*	*25/01/2018*	*Friday*
backup		30/01/2019	Wednesday
backup		31/01/2019	Thursday
extra	*Question hour*	*01/02/2019*	*Friday*

Lecture 18	Data preprocessing, data quality, binning, etc.	13/12/2018	Thursday
Lecture 19	Visual analytics & information visualization	19/12/2018	Wednesday
backup		20/12/2018	Thursday
Instruction 9	*Text mining, preprocessing and visualization*	*21/12/2018*	*Friday*
Lecture 20	Responsible data science (1/2)	09/01/2019	Wednesday
Lecture 21	Responsible data science (2/2)	10/01/2019	Thursday
Instruction 10	*Responsible data science*	*11/01/2019*	*Friday*

- **Register for the exam!**
- **Exchange students.**
- **Assignments.**


Merry Christmas!
And happy new year