Doc. RNDr. Irena Holubová, Ph.D. & PROFINIT

DATA SCIENCE

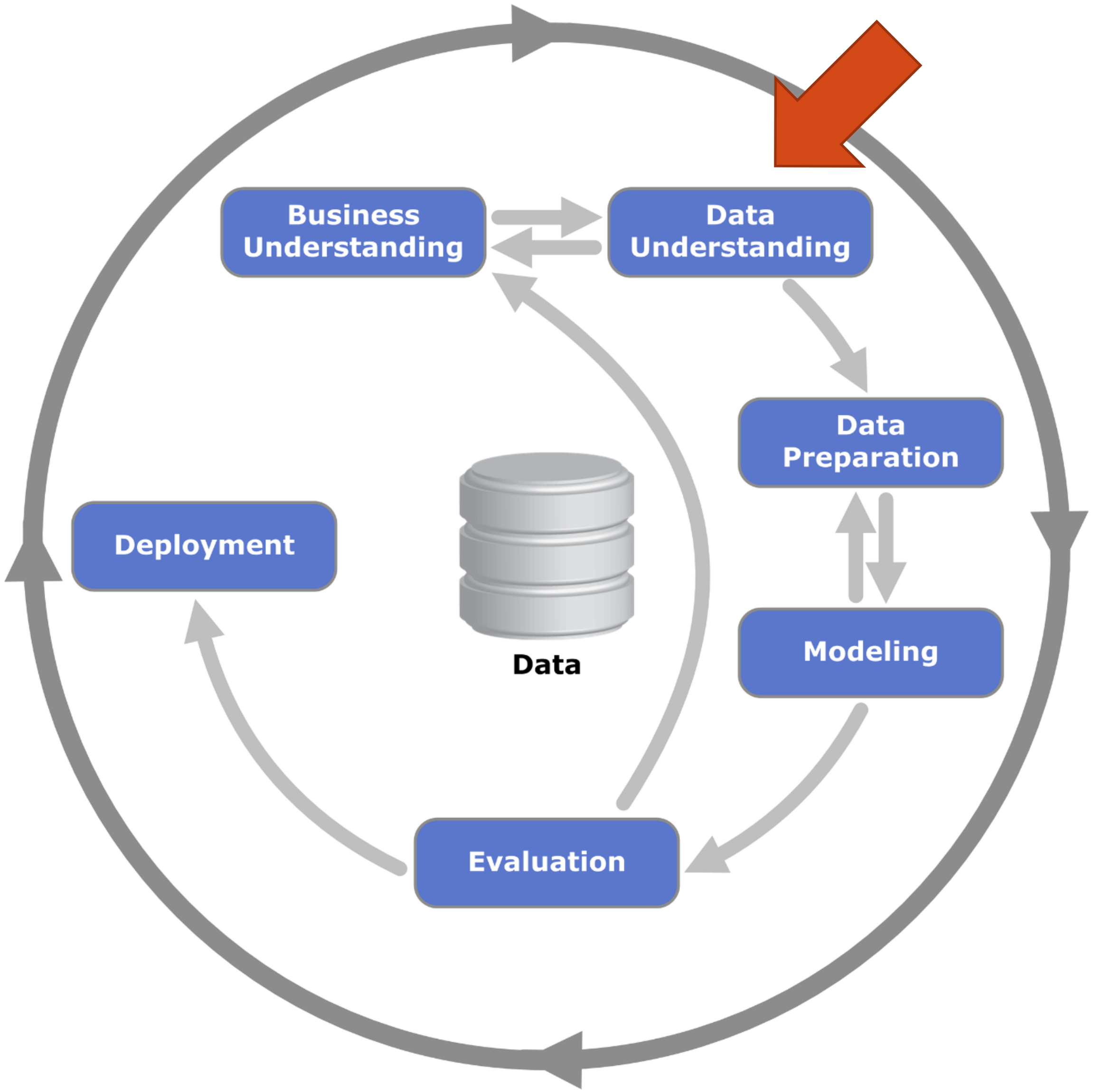
NDBI048

Data Visualization

https://www.ksi.mff.cuni.cz/~holubova/NDBI048/

CRISP-DM PHASES

I. Business Understanding
II. Data Understanding
III. Data Preparation
IV. Modeling
V. Evaluation
VI. Deployment



https://www.datascience-pm.com/crisp-dm-2/

> *"One picture is worth ten thousand words."*
> Chinese proverb

BASIC TECHNIQUES

- Descriptive statistics
  - Numbers that summarize properties of the data
    - Frequency, mean, standard deviation, …
  - Most can be calculated in a single pass through the data
- Visualization
  - Conversion of data into a visual format → characteristics of the data and the relationships among data items or attributes can be analysed / reported
    - Data objects, their attributes, and the relationships among data objects → points, lines, shapes, colours, …
  - One of the most powerful techniques for data exploration
  - Humans have a well developed ability to analyse large amounts of information that is presented visually
    - Can detect general patterns and trends, outliers and unusual patterns

DESCRIPTIVE STATISTICS

DESCRIPTIVE STATISTICS

- Terms
  - **statistical unit**: person, observation, family, town, …
  - **population**: (complete) set of statistical units
  - **variable**: a unit feature for which we have values
  - **distribution**: which values of variable are present and how often

Aims of descriptive statistics:

- to describe observed **distributions** of individual variables
- to describe observed **relationships** between multiple variables

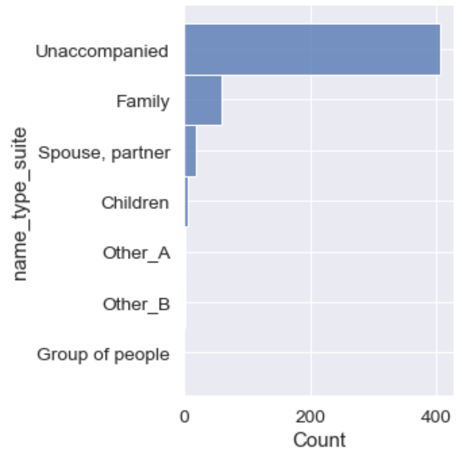
BASIC VARIABLE TYPES

- Categorial
  - nominal (unordered): names, trees, pets
  - ordinal (ordered): education, approval
- Numeric
  - difference-like (zero has no meaning): temperature, rating
  - ratio-like (zero mean „none“): distance, amount, time
  - discrete/continuous

CATEGORIAL VARIABLE DISTRIBUTION

- Frequency tables
- Frequency graphs
  - Stacked (ordinal)

name_type_suite	count	count_rel
Children	7	0.014028
Family	60	0.120240
Group of people	1	0.002004
Other_A	3	0.006012
Other_B	2	0.004008
Spouse, partner	20	0.040080
Unaccompanied	406	0.813627

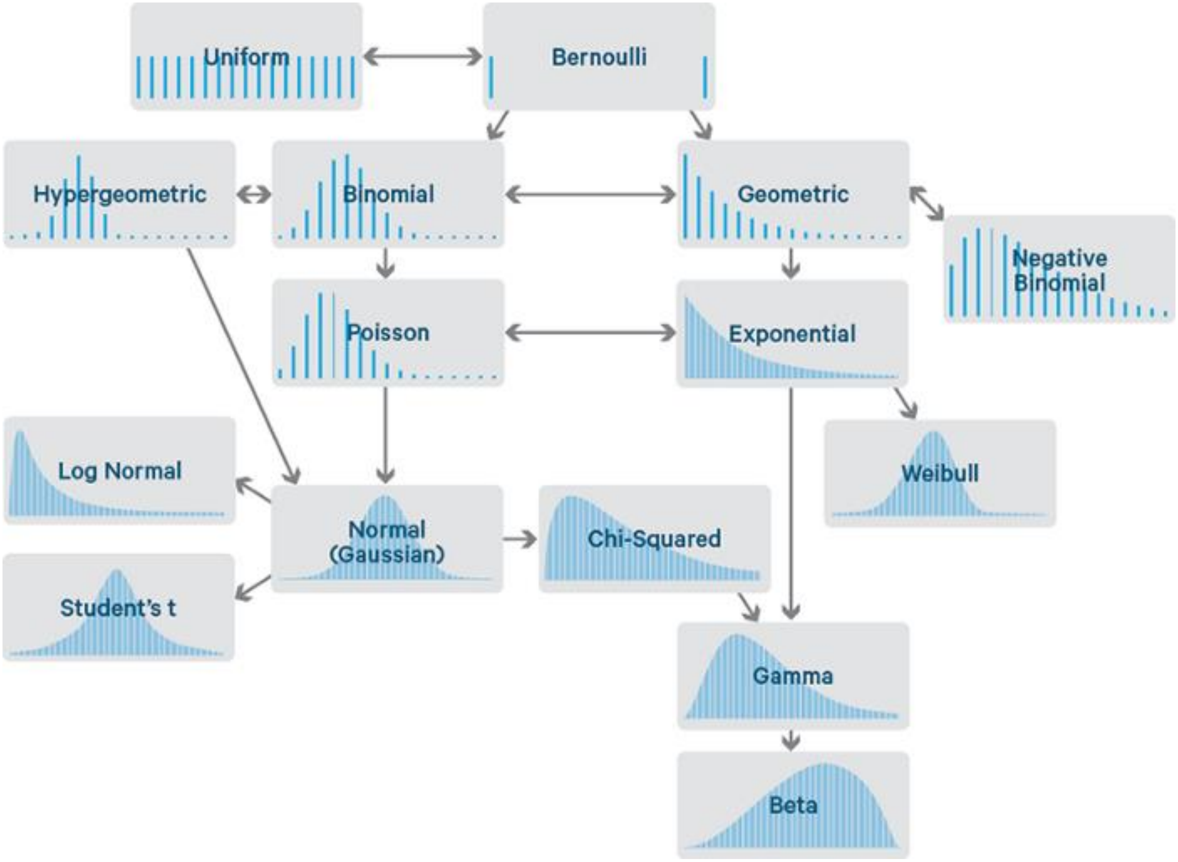


NUMERIC VARIABLE DISTRIBUTION

- A little of unique values → treat as categorial
- A lot of unique values:
  - full information: empirical cumulative distribution function (ECDF), stripplot
  - balanced: histogram, density estimation
  - compressed: boxplot, numerical statistics
    - What range are values in? → min, max
    - Which value is the „center“? → (trimmed) mean/average, median, mode
    - What is the dispersion of values? → standard deviation, interquartile range
    - What other values or thresholds are important? → quantiles, second mode
    - What is the shape of the distribution? → approximating by a standard distribution

PROBABILITY DISTRIBUTIONS

- discrete / continuous
- symmetrical / skewed
- light / heavy tail
- unimodal / bimodal



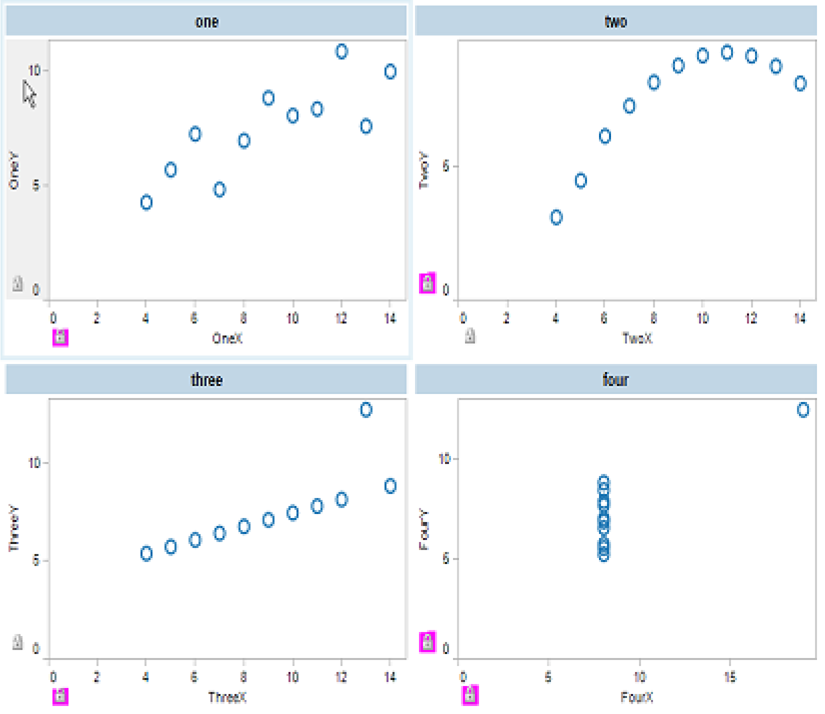
VISUALISATION

MOTIVATION

Similar motivation as for statistics but visualization can reveal / distinguish data/trends/patters, … which statistics can not (easily)

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

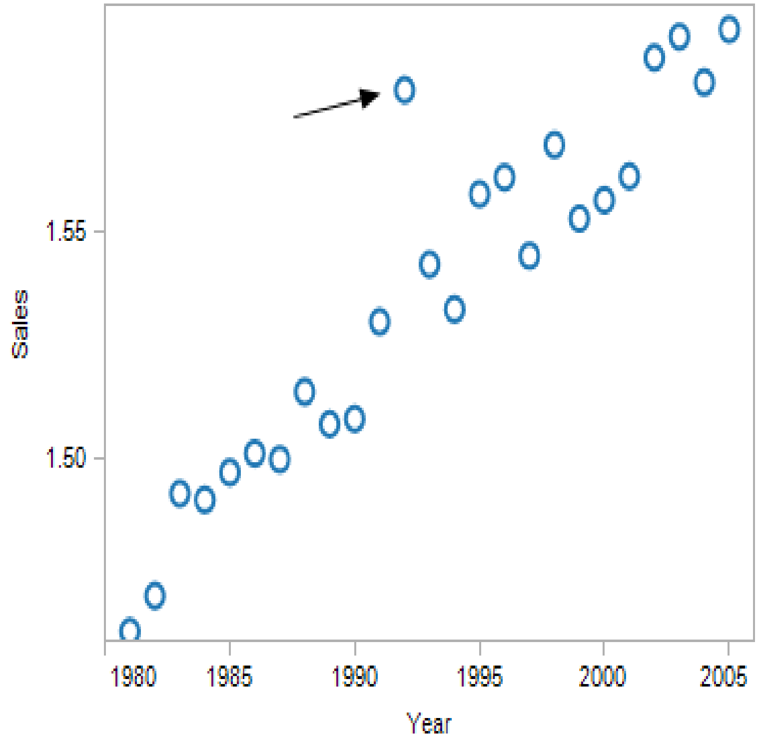
Source: Tufte, Edward R (1983), *The Visual Display of Quantitative Information*, Graphics Press



Four data sets with nearly identical linear model (mean, variance, linear regression line, …)

	A	B
1	Year	Sales
2	1981	1.4622
3	1982	1.47004
4	1983	1.49253
5	1984	1.49118
6	1985	1.49722
7	1986	1.50138
8	1987	1.50008
9	1988	1.51493
10	1989	1.50781
11	1990	1.50899
12	1991	1.53037
13	1992	1.58137
14	1993	1.54299
15	1994	1.53307
16	1995	1.55845
17	1996	1.56213
18	1997	1.54488
19	1998	1.56927
20	1999	1.55305
21	2000	1.5571
22	2001	1.56235
23	2002	1.58847
24	2003	1.59309
25	2004	1.58303
26	2005	1.5947

MOTIVATION



Find an outlier….

DATA VISUALIZATION

- Information visualization has two equally important aspects
  - Structural modeling
    - Detection, extraction and simplification of the underlying information
  - Graphical representation
    - Transform initial representation into a graphical one which provides visualization of the structure
      - Different types of structures require different type of visualization
      - e.g., time series vs. hierarchical information

EXPLORATORY VS. EXPLANATORY VISUALIZATION

Exploratory

- What the data is
- What is hidden in the data
- Enables to look at the data from different angles

Explanatory

- Helping to make sense of the data by choosing the right technique
- Needs to know the context from which the user come and what they need to know
- Strategic placement of elements and choice of attributes to help the users to focus on what is important

BIG DATA VISUALIZATION

- Decision about what technique to use became more difficult with Big Data
  - Visualization is needed to decide which portion of data to explore further
  - Visualization algorithms (i.e., graph drawing) should scale well to billions of entities (nodes)
    - The first application was probably the visualization of web-related data
      - i.e., pages, relations, traffic, …
  - New techniques may be needed
  - Trends might not be clear
  - Noise reduction might be even more necessary

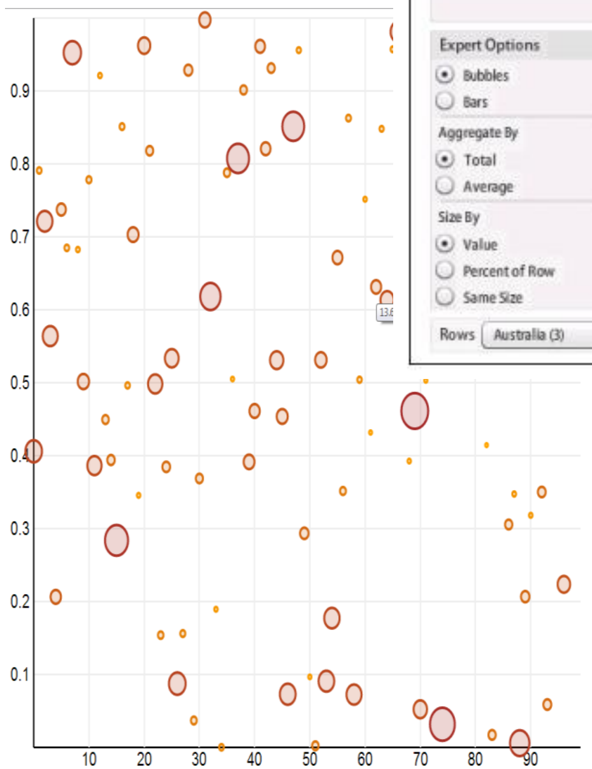
CHALLENGES OF DATA VISUALIZATION

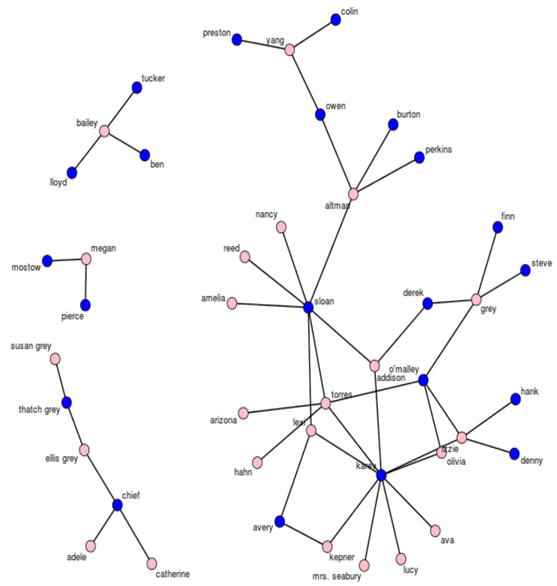
- Determine the medium
  - Table – individual precise values, comparison of individual values, multiple levels of aggregation, …
  - Graph – pattern trends and exceptions, a set of values is seen as whole, …
  - Schema
- Design the components of the medium
  - Which data to emphasize, which colors to choose, …

VISUALIZATION TYPES

DATA RELATIONSHIPS

- Scatter plot
  - Classical statistical diagram that lets us visualize relationships between numeric variables
  - Can carry additional information
    - Color, shape, size, …
- Matrix chart
  - Summarizes a multidimensional data set in a grid
- Network diagram
  - A set of objects (vertices) connected by edges
  - Visualization of the network is optimized to keep strongly related items in close proximity to each other



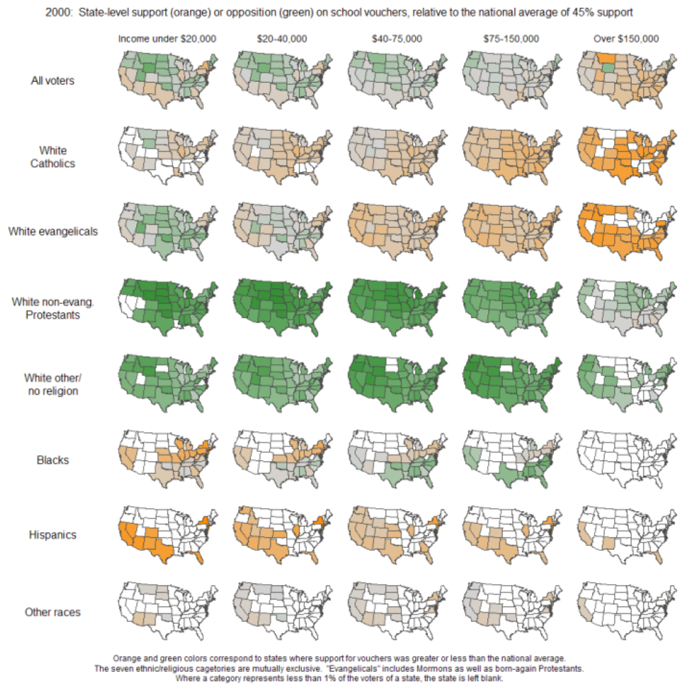


Camping Equipment
Camping Equipnull
Golf Equipment
Expert Options
Bubbles
Bars
Aggregate By
Total
Average
Size By
Value
Percent of Row
Same Size
Australia
2,285
1,029
Austria
3,098
5,035
Belgium
944
4,064
Yr - 2007
Yr - 2006
Rows
Australia (3)
Columns
Yr - 2006 (2)
Color
Camping Equipment (
Size
4,383

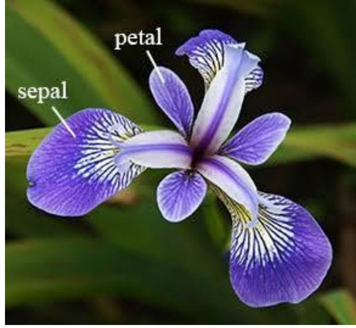
VISUALIZATION TYPES

DATA RELATIONSHIPS

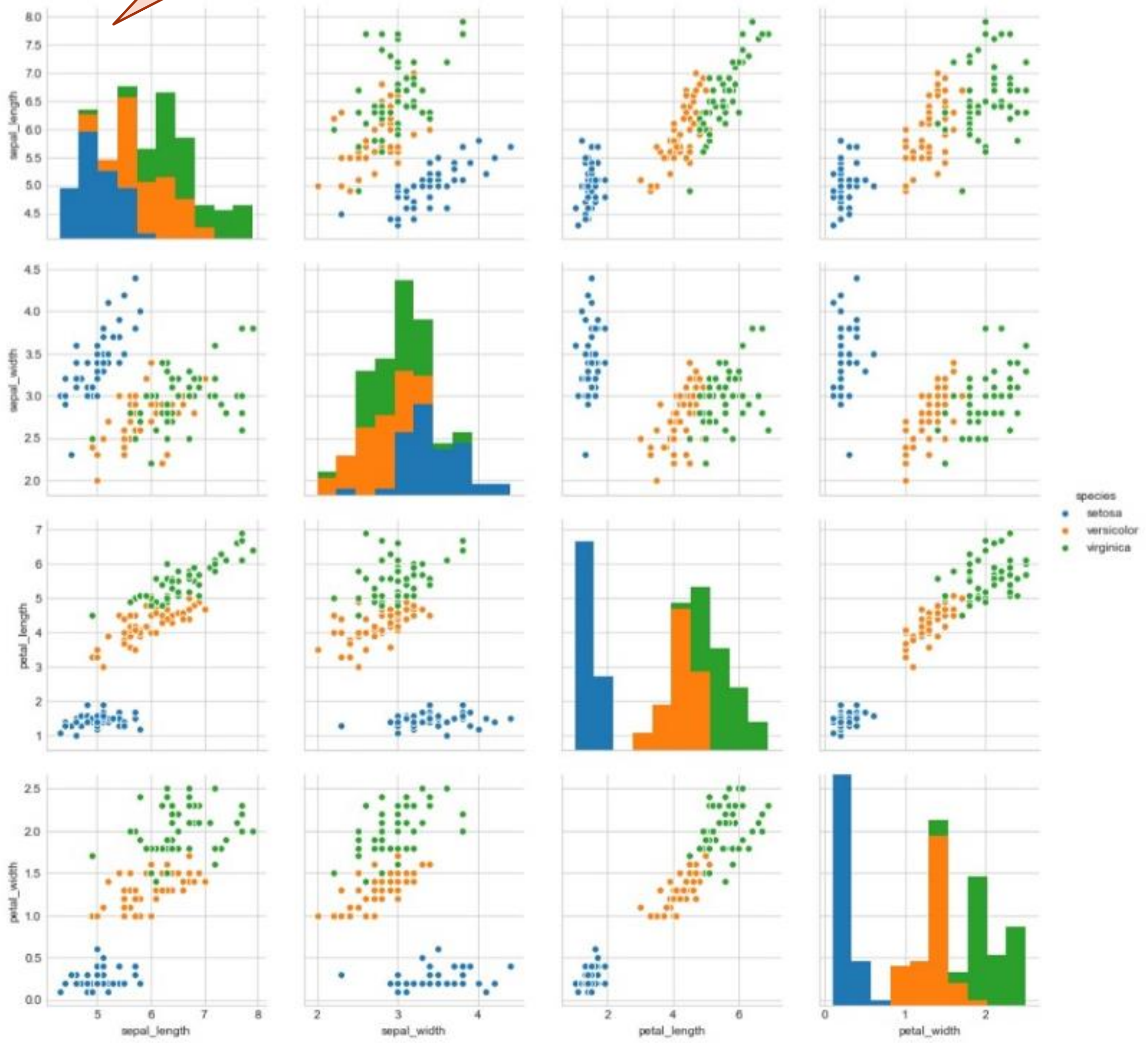
- (Scatter) plot matrix
  - Matrix of scatter (or other) plots
  - Each scatter plot is created between different combinations of variables



distribution of data for the variable in the column



Iris data set



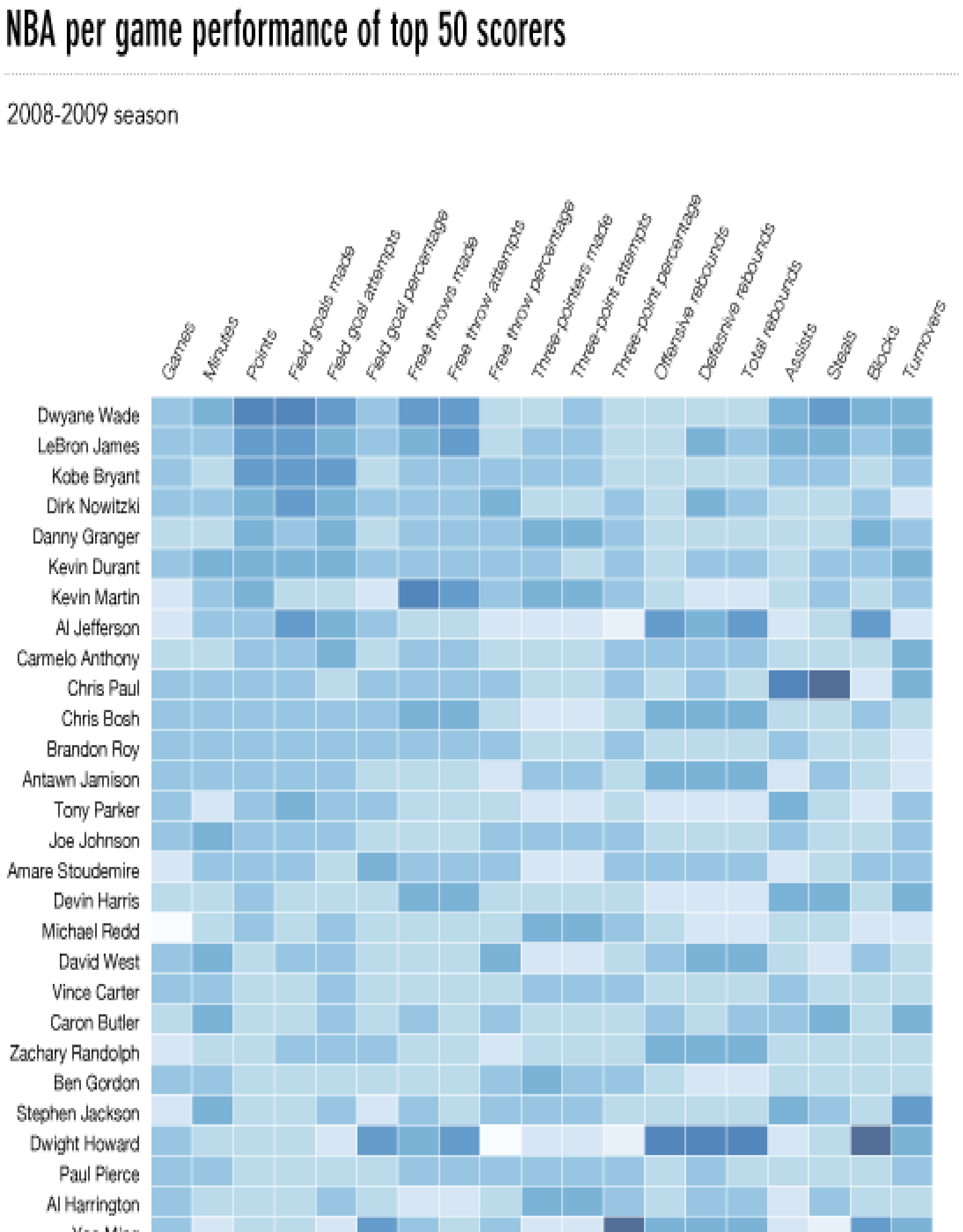
2000: State-level support (orange) or opposition (green) on school vouchers, relative to the national average of 45% support

Orange and green colors correspond to states where support for vouchers was greater or less than the national average.
The seven ethnic/religious cagetories are mutually exclusive. "Evangelicals" includes Mormons as well as born-again Protestants.
Where a category represents less than 1% of the voters of a state, the state is left blank.

VISUALIZATION TYPES

DATA RELATIONSHIPS

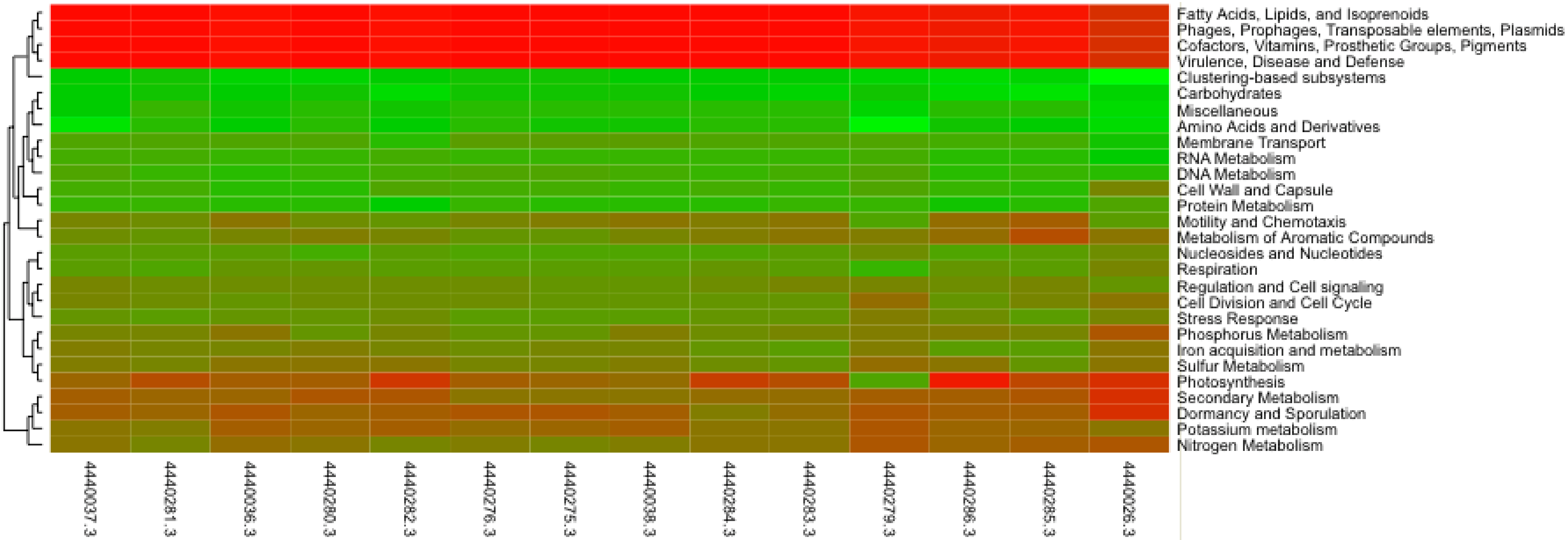
- Correlation matrix (heat map)
  - Combines data to quickly identify which variables are related
  - Shows how strong the relationship is between the variables



VISUALIZATION TYPES

DATA RELATIONSHIPS

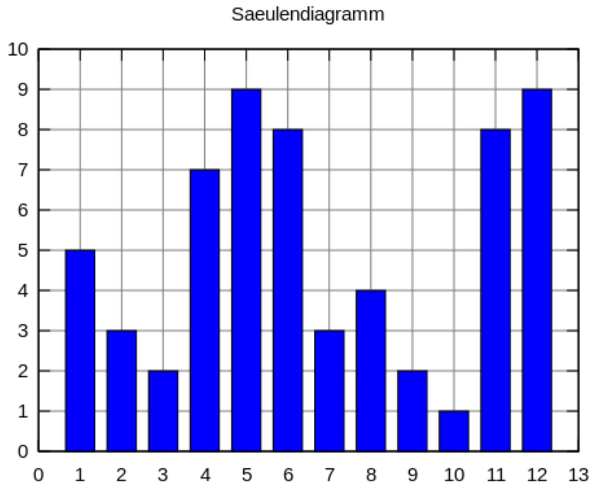
- Heat map is often combined with a dendrogram
  - Aggregates rows or columns based on their overall similarity into a tree structure

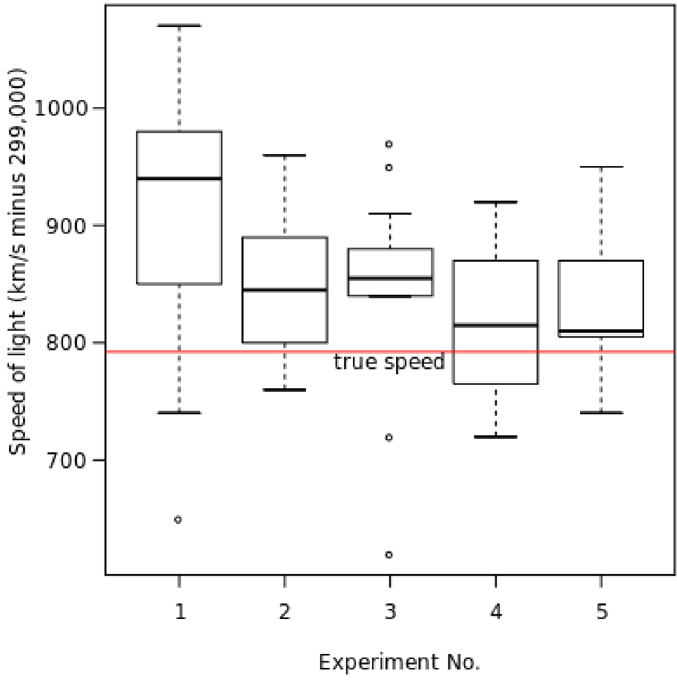


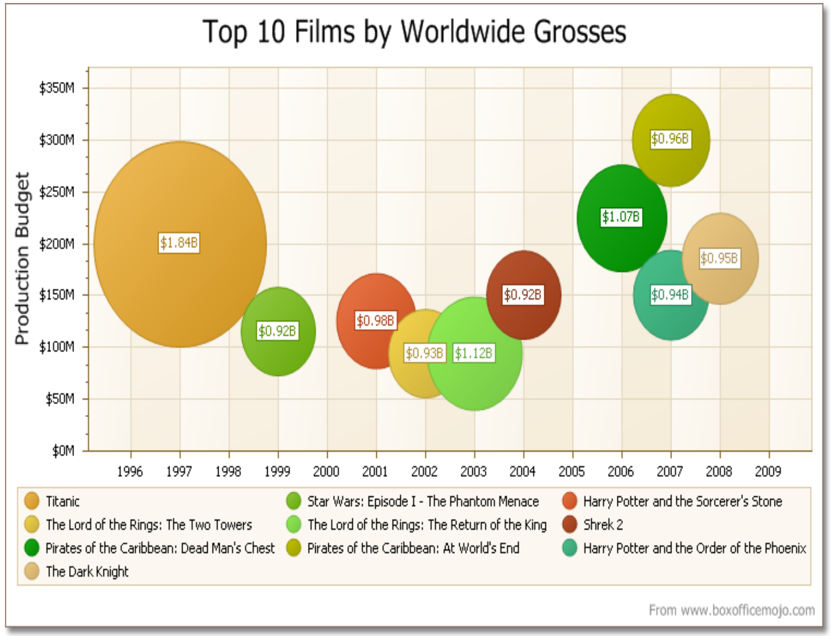
VISUALIZATION TYPES

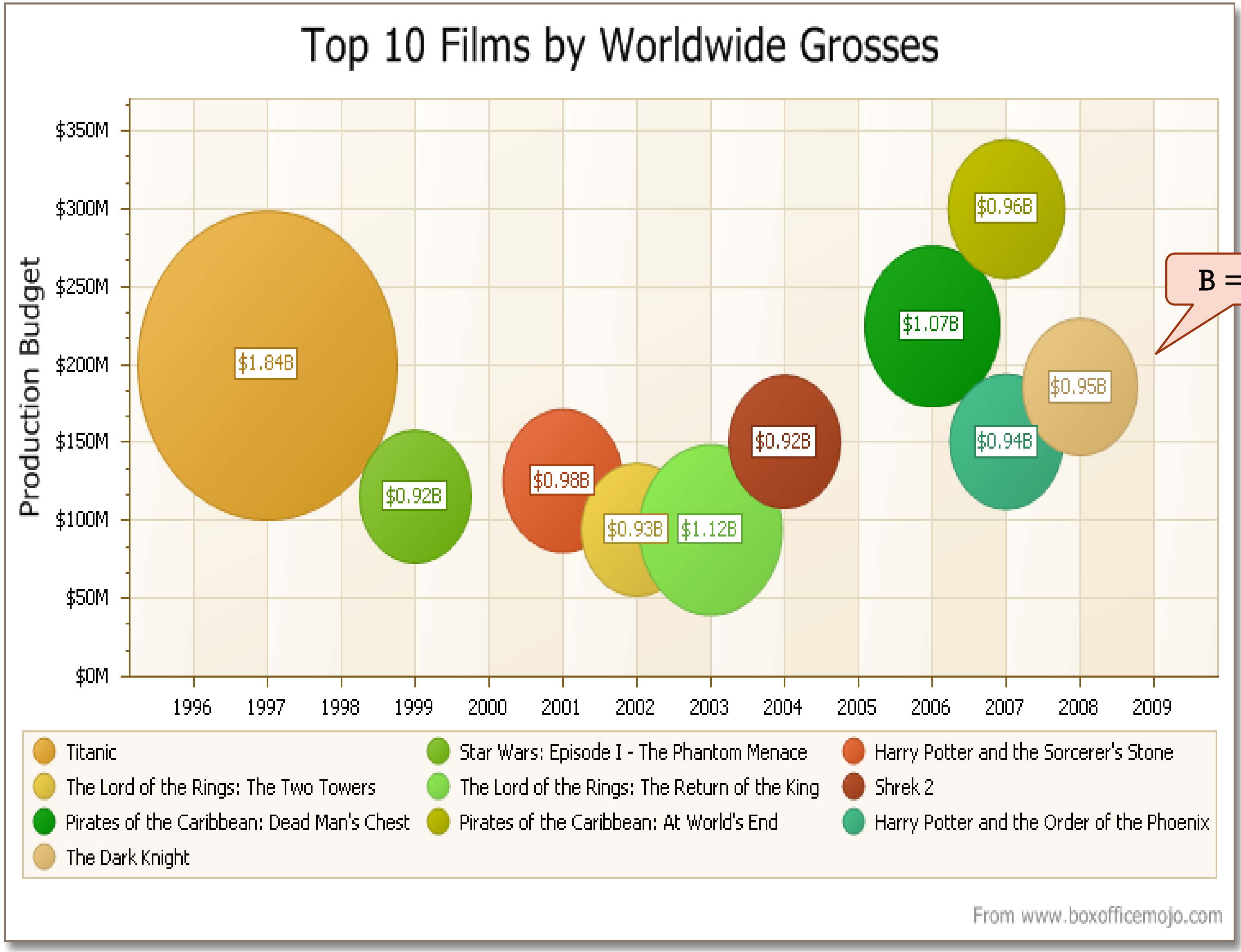
COMPARISON OF A SET OF VALUES

- Bar Chart
  - Classical method for numerical comparisons
  - Histograms
- Box plot (box-and-whisker plots)
  - Five statistics (minimum, lower quartile, median, upper quartile and maximum) summarizing the distribution of a set of data
- Bubble chart
  - Circles in a bubble chart represent different data values
  - Triplet (v_1, v_2, v_3) of data = bubble
    - Two of the v_i values = xy location
    - Third = size







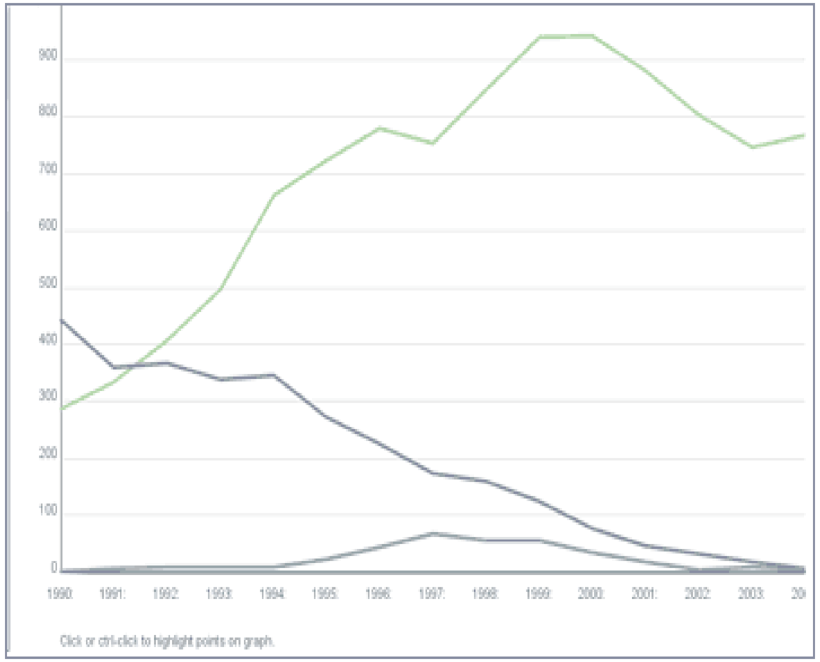


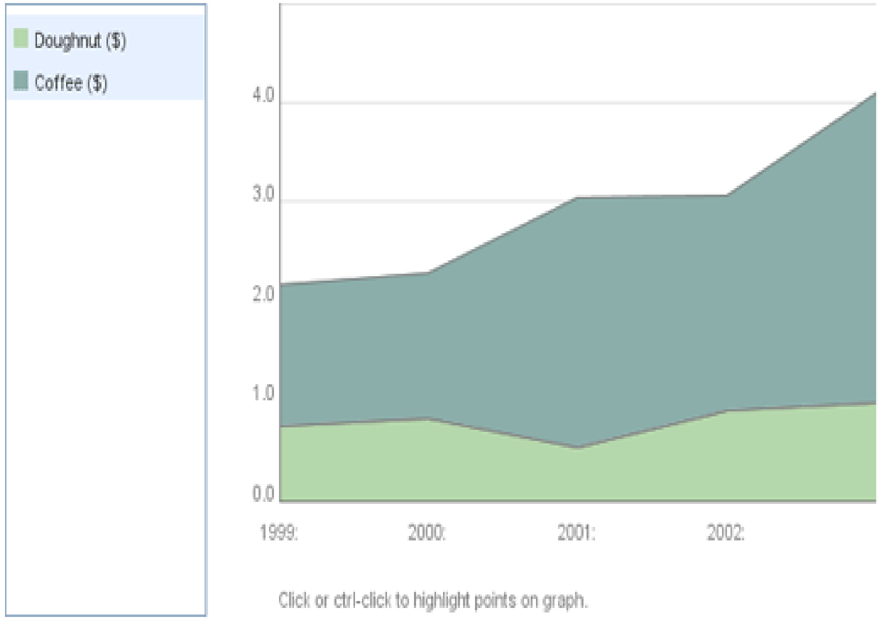
Top 10 Films by Worldwide Grosses
Production Budget
$350M
$300M
$250M
$200M
$150M
$100M
$50M
$0M
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
$1.84B
$0.92B
$0.98B
$0.93B
$1.12B
$0.92B
$1.07B
$0.96B
$0.94B
$0.95B
Titanic
Star Wars: Episode I - The Phantom Menace
Harry Potter and the Sorcerer's Stone
The Lord of the Rings: The Two Towers
The Lord of the Rings: The Return of the King
Shrek 2
Pirates of the Caribbean: Dead Man's Chest
Pirates of the Caribbean: At World's End
Harry Potter and the Order of the Phoenix
The Dark Knight
From www.boxofficemojo.com

B = 10^9

VISUALIZATION TYPES
TRENDS OVER TIME

- Line graph
  - Classical method for visualizing continuous change
- Stack graph
  - Visualizing change in a set of items
  - The sum of the values is as important as the individual items

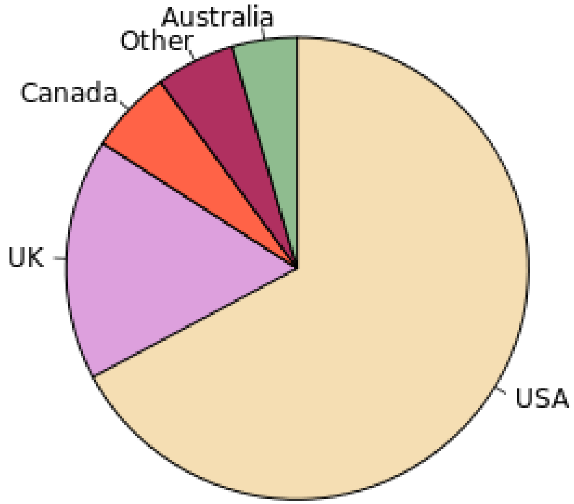


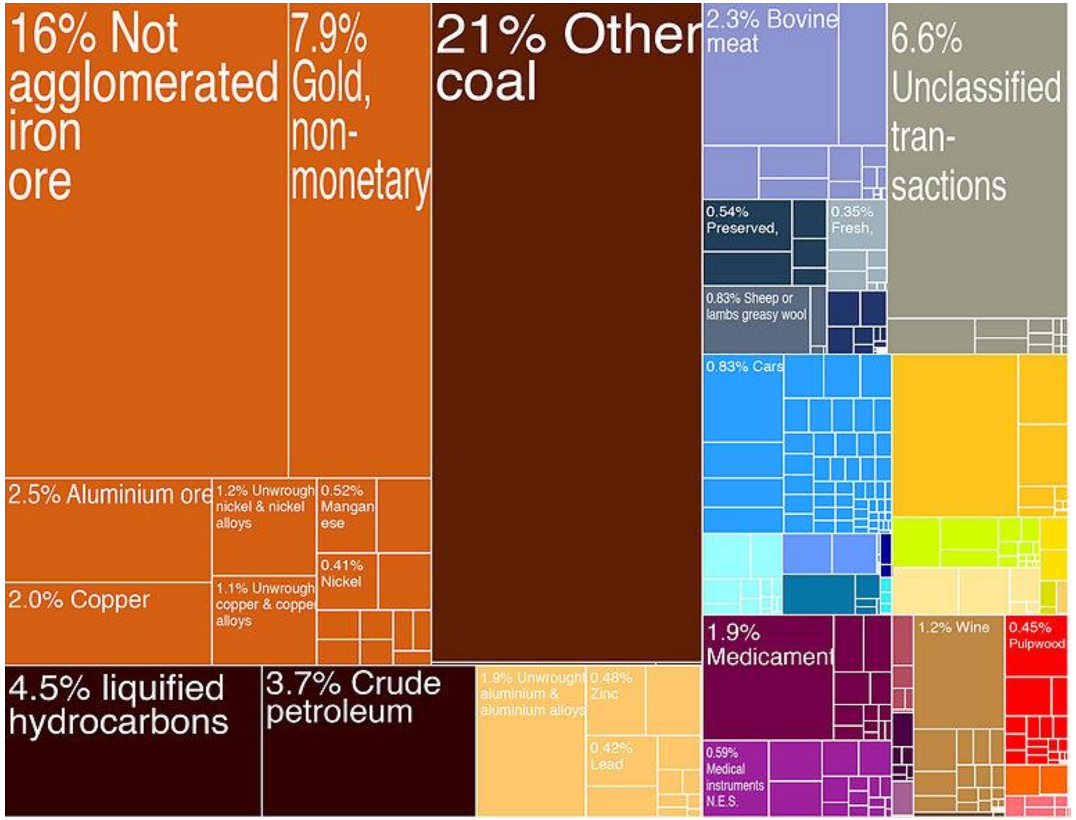


VISUALIZATION TYPES

PARTS OF A WHOLE

- Pie Chart
  - Percentages are encoded as "slices" of a pie, with the area corresponding to the percentage
- Treemap
  - Visualization of hierarchical structures
  - Effective in showing attributes of leaf nodes using size and color coding
  - Enable to compare nodes and sub-trees at varying depth





Economy of Australia

VISUALIZATION TYPES

TEXT ANALYSIS

- Tag cloud
  - Visualization of word frequencies
    - i.e., how frequently words appear in a given text

WHICH VISUALIZATION TECHNIQUE TO USE?

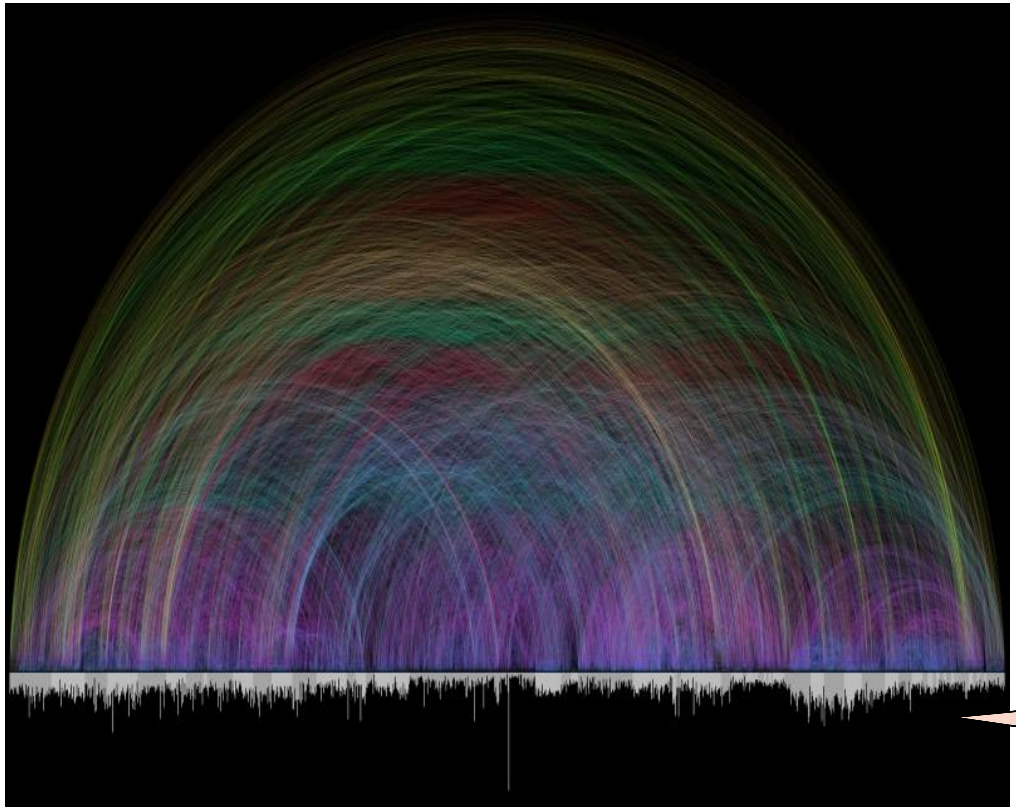
- New visualization software is capable of "guessing" the correct visualization based on the characteristics of the data
  - One-dimensional data $\Rightarrow$ bar chart
  - Two-dimensional data $\Rightarrow$ scatter plot
  - N-dimensional data $\Rightarrow$ multiple scatter plots, matrix chart, …
  - Data with coordinates $\Rightarrow$ map-based charts
- Offers options
- Trend: to simplify the process for common users

BIG DATA VISUALIZATION

- The goal of visualizing Big Data is usually to make sense of a large amount of interlinked information
- In interconnected data the connections between objects are difficult to organize on a linear layout
  - **Circular representations**
  - **Network diagrams**
- Typical "topologies" one can encounter (a bit confusing term based on Manuel Lima's "Visual Complexity" – see references) include arc diagrams, centralized burst, centralized ring, globe, circular ties or radial convergence
  - And many more…

ARC DIAGRAM

- Vertices are placed on a line and edges are drawn as semicircles
- Arcs represent relationships
  - Colors can encode, e.g., distance

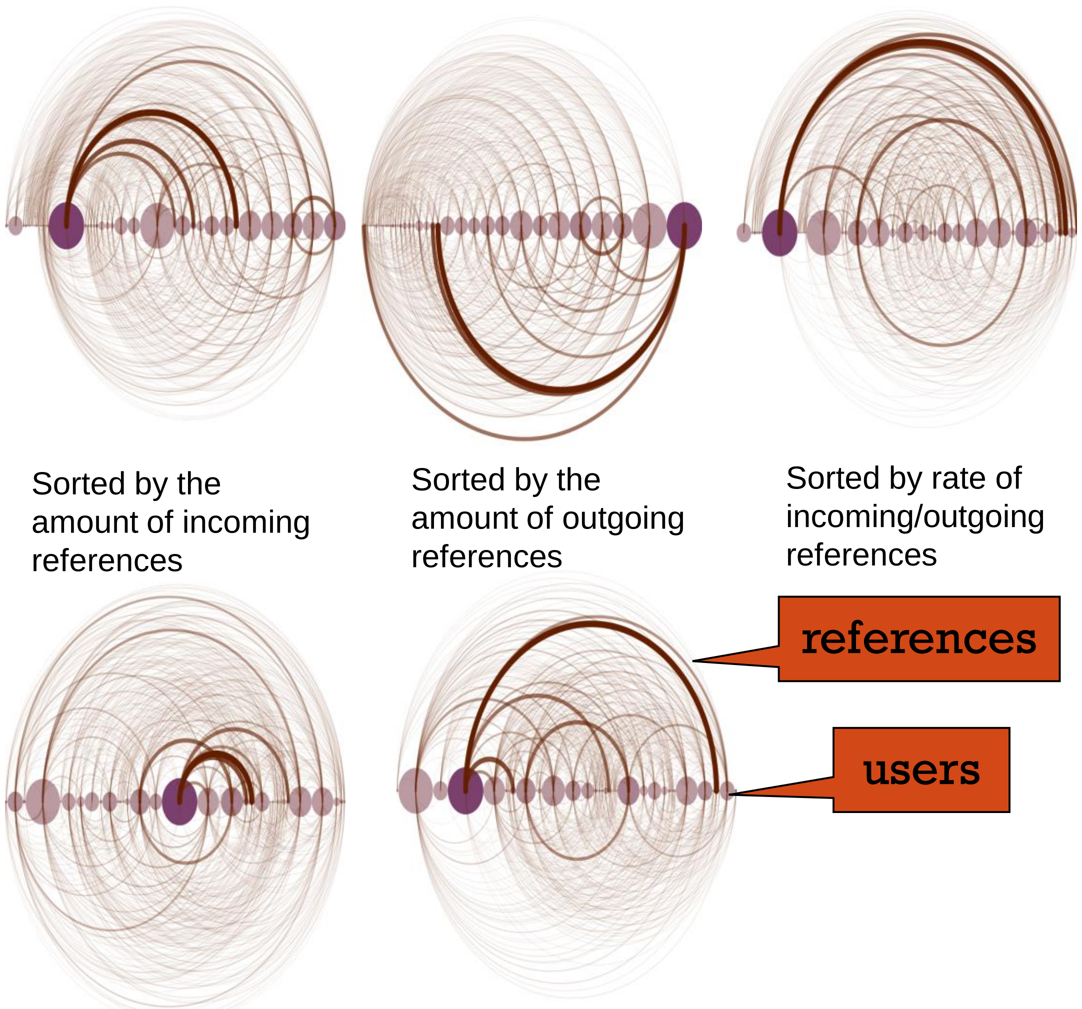


A map of 63,799 cross-references found in the Bible. The bottom bars represent number of verses in the given chapter. Color of arcs represents the distance between the two chapters.

http://www.chrisharrison.net/index.php/Visualizations/BibleViz

grey/white = book

ARC DIAGRAM



Sorted by the amount of incoming references

Sorted by the amount of outgoing references

Sorted by rate of incoming/outgoing references

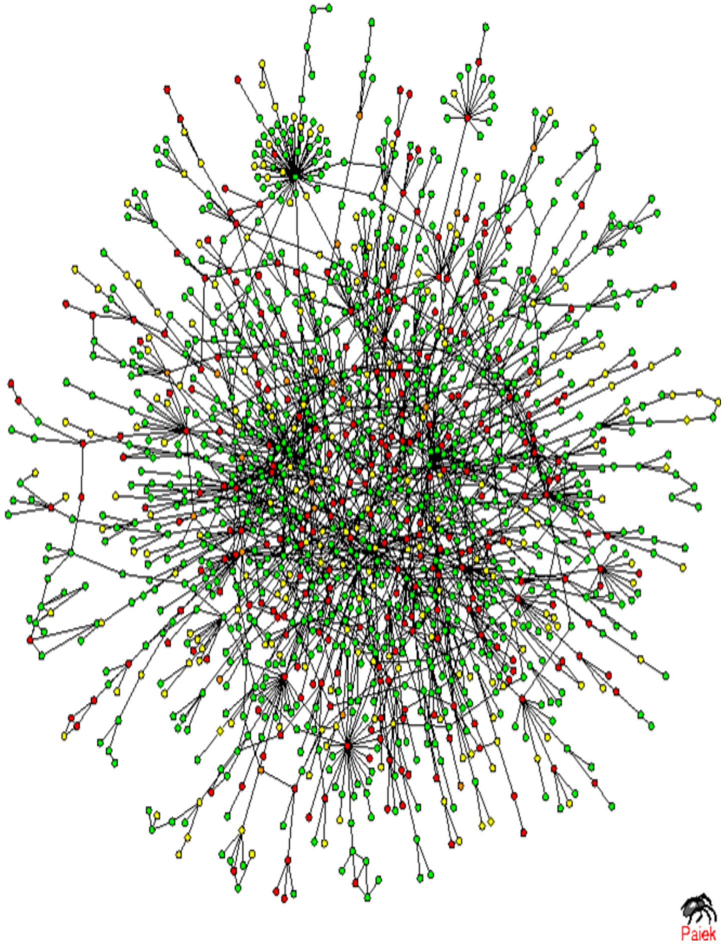
Sorted by user name

Unsorted

- Visualization of IRC communication behavior: Who is talking to whom?
  - Arcs are directional and drawn clockwise:
    - In the upper half of a graph they point from left to right, in the bottom half from right to left
  - Arc strength corresponds to the number of references from the source to the target
  - Circle size = Number of messages
  - Circle color = Average message length
- This visualization favors strong social connections over sociability: Frequent references between the same two users feature more prominently than combined references from several sources to a single target.

http://datavis.dekstop.de/irc_arcs/

CENTRALIZED BURST

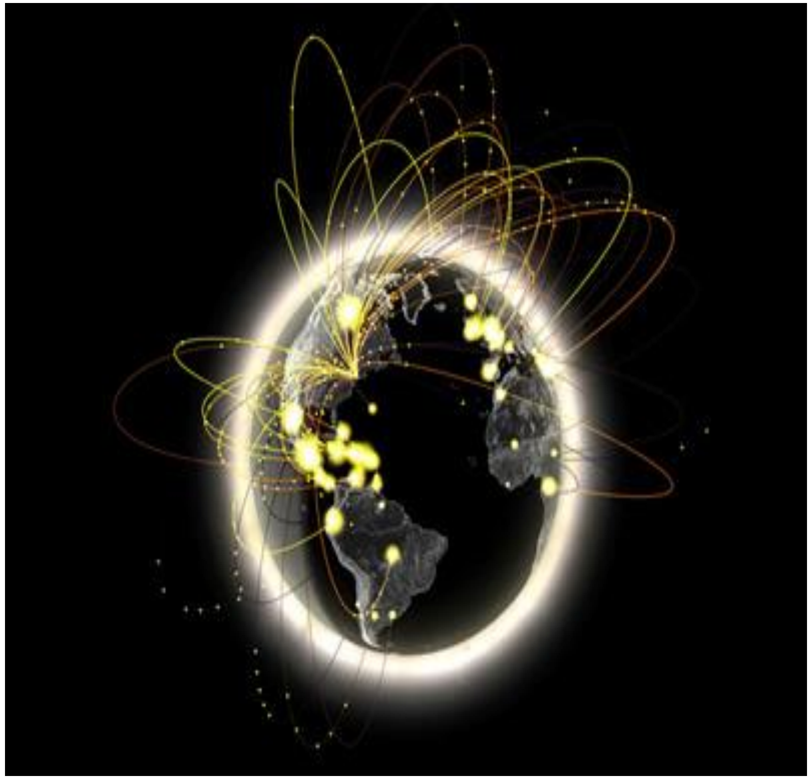


A map of protein-to-protein interactions of a yeast
source: H. Jeong. et al. “Lethality and Centrality in Protein Networks”, Nature, no. 411, 2011: 41-42

- Visualization with strong central tendency
- Can reveal highly connected objects (hubs) which usually correspond to objects with high importance
  - e.g., in a gene network, hubs are interesting points for targeting new drugs
    - Disabling a central gene probably will not allow the organism to adapt

GLOBE

- Globe visualizations are basically projections of other topologies on a globe

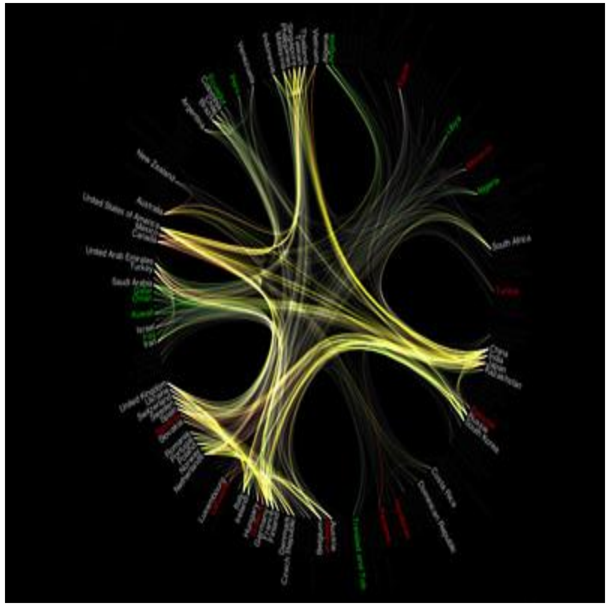


- The global exchange of information in real time by visualizing volumes of long distance telephone and IP data flowing between New York and cities around the world.
- How does the city of New York connect to other cities? With which cities does New York have the strongest ties and how do these relationships shift with time? How does the rest of the world reach into the neighborhoods of New York? The size of the glow on a particular city location corresponds to the amount of IP traffic flowing between that place and New York City. A greater glow implies a greater IP flow.

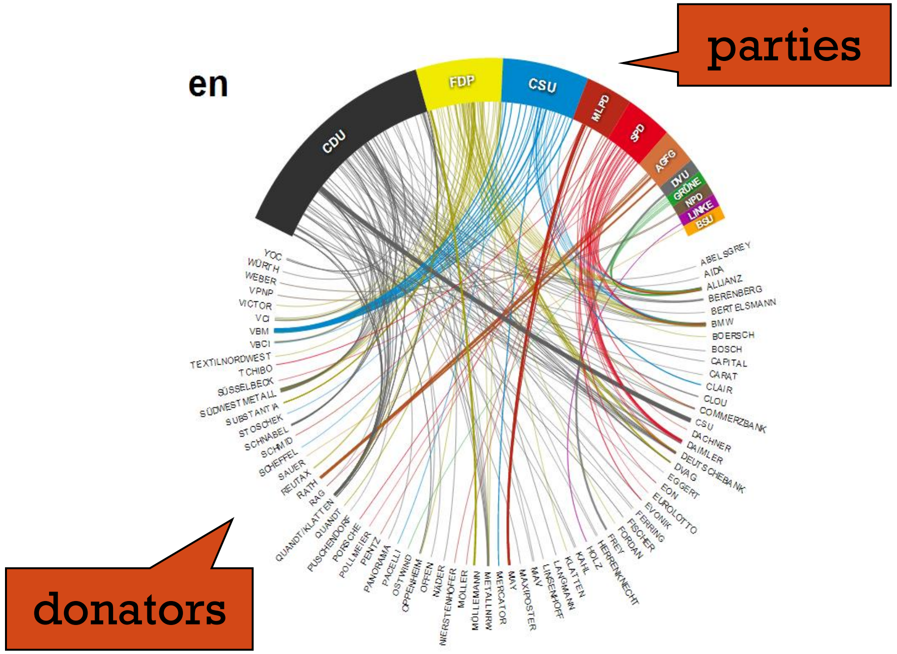
http://www.aaronkoblin.com/work/NYTE/index.html

RADIAL CONVERGENCE

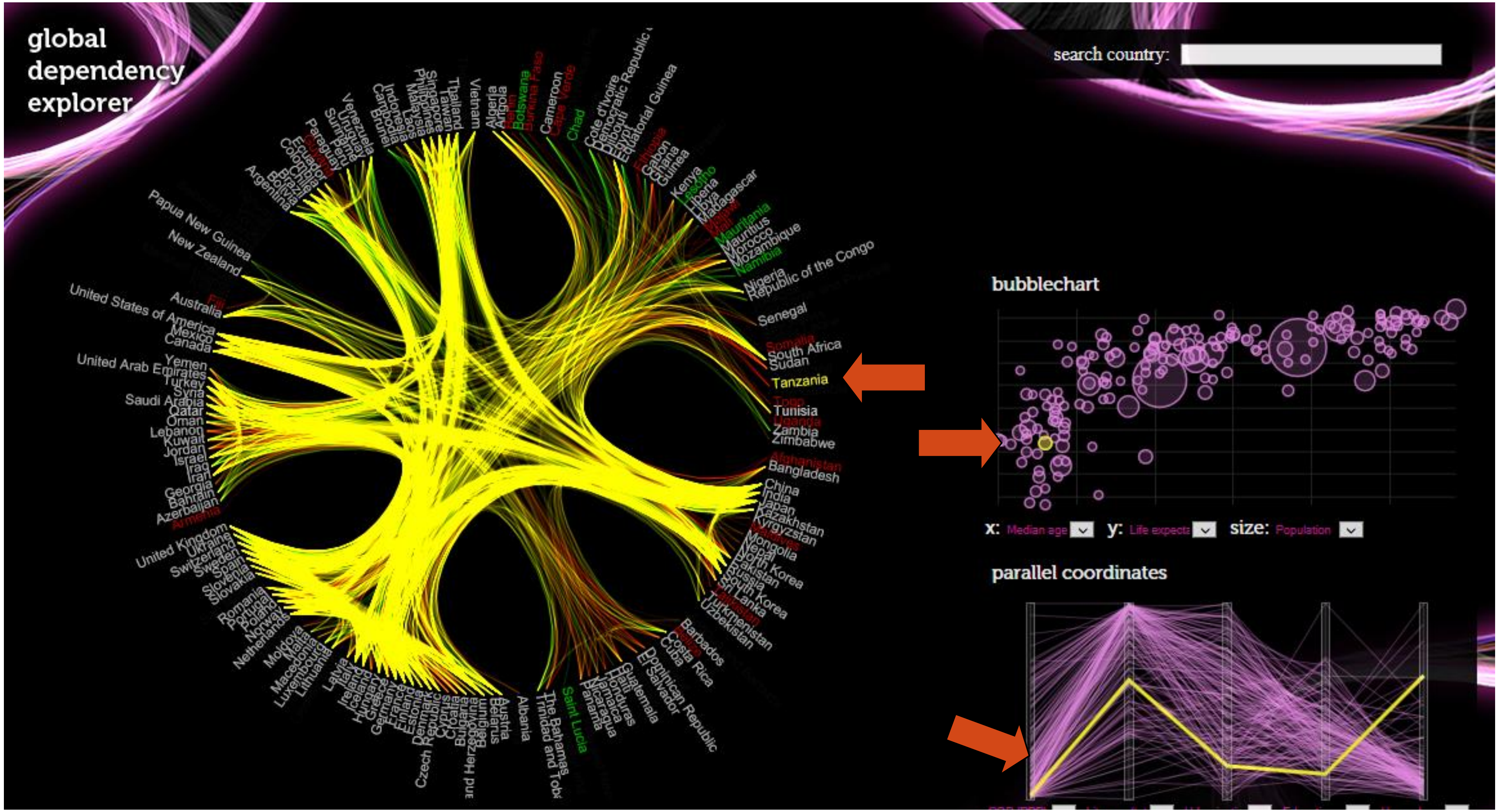
- Also known as radial chart
- Actually a 360 arc diagram



Tracking the commercial ties between most countries across the globe.
http://cephea.de/gde/



Money flow from private donators to parties in the German Bundestag (house of the parliament).
http://labs.vis4.net/parteispenden/


global
dependency
explorer
search country:
bubblechart
x: Median age
y: Life expecta
size: Population
parallel coordinates
Tanzania

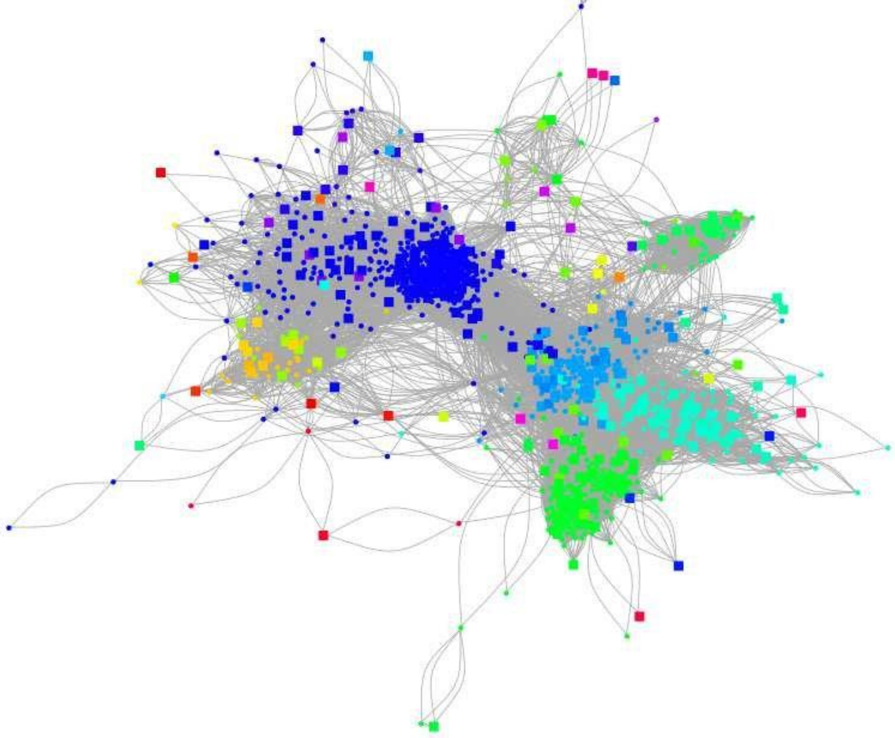
GRAPH DRAWING

- Spatial layout and graph drawing play a key aspect in information visualization
- Good layout needs to express the key features of a complex structure
- Graph drawing algorithms first agree on a criterion of what makes a good graph (and what should be avoided) and then run an algorithm driven by these criteria
- Generally, the primary goal is to optimize the arrangement of nodes so that strongly connected nodes appear close to each other
  - Most widely known graph drawing algorithms combine **force-directed** graph drawing and **spring-embedder** algorithms
  - The strength of a connection needs to be defined

FORCE-DIRECTED PLACEMENT

- Can be traced back to VLSI (very large scale integration) design = creating integrated circuits
  - Aim: optimize the layout of a circuit to a obtain as few number of crossings as possible
- Generally agreed on aesthetics criteria
  - **Symmetry**
  - **Even distribution of nodes**
  - **Uniform edge lengths**
  - **Minimization of edge crossings**
- Some of the criteria can be mutually exclusive
  - e.g., symmetric graph may require crossings which might be avoided

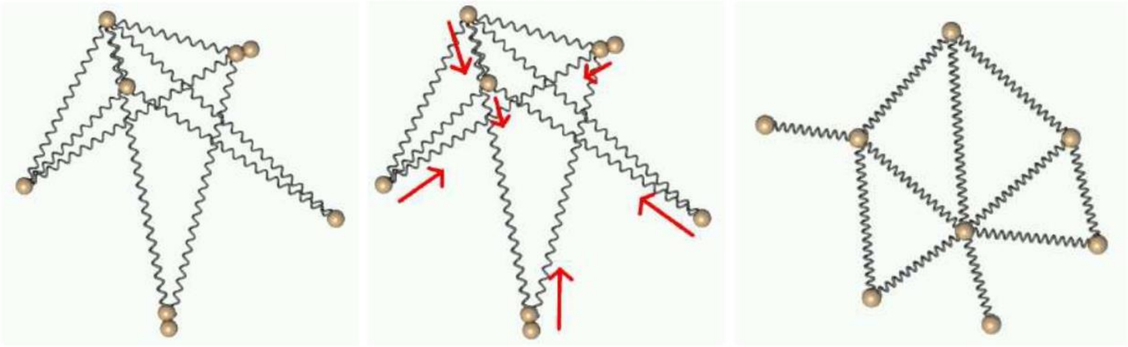
The facebook c network represented with a force-directed placement algorithm [22]. Colors represent the clusters on the map and selected nodes used to train the map are represented by squares (instead of circles)



https://www.researchgate.net/figure/The-facebook-c-network-represented-with-a-force-directed-placement-algorithm-22-Colors_fig2_281597675

FORCE-DIRECTED PLACEMENT

- Replaces vertices in a graph by steel rings and edges by springs
  - Attractive force is applied to a pair of connected nodes
    - Spring-like forces (Hook’s law)
  - Repulsive force is applied to a pair of disconnected nodes
    - Forces of electrically charged particles (Coulomb's law)
- Equilibrium state for the system of forces:
  - Edges tend to have uniform length (spring forces)
  - Nodes that are not connected by an edge tend to be drawn further apart (electrical repulsion)

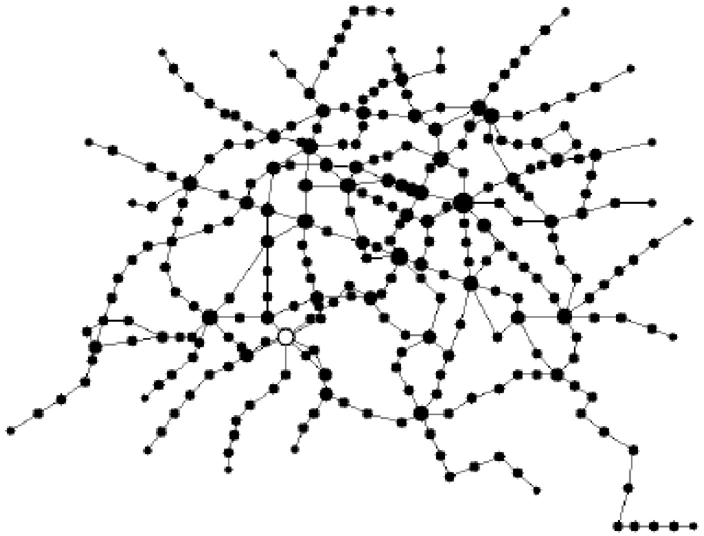


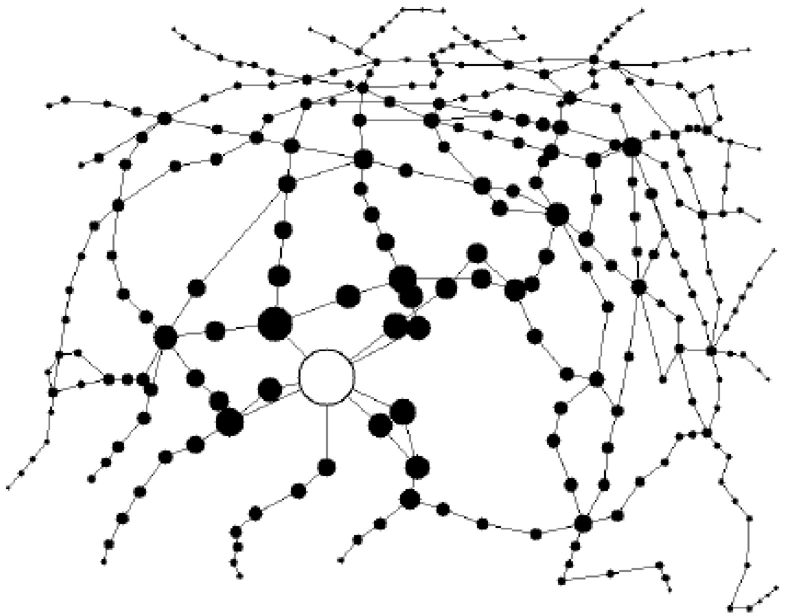
GRAPH DRAWING CHALLENGES

- Current algorithms are inefficient in incremental updating of the layout
  - i.e., one needs to redraw the whole layout when adding/removing a single node
- Networks with heterogeneous link types or node types cannot be efficiently handled
  - e.g., having users of a social network sharing various type of content (images, posts, video) and forming various types of relations (liking, tagging)
- Majority of algorithms focus on strong ties (heavyweight links)
  - Weak ties can be surprisingly valuable because they are more likely to be the source of novel information
    - e.g., hearing about a new job offering is an example of weak link with great social impact

GRAPH DRAWING CHALLENGES

- Scalability
  - Big Data related challenge
  - Problematic scaling as the size and density of the network increases
    - i.e., Big Data are also difficult to visualize
- Limited screen resolution
  - Big Data related challenge
  - Sometimes we simply do not have enough pixels to visualize a complex large-scale network
    - Zoomable interfaces, fish-eye views, …
    - In general solvable by interactivity





GRAPH DRAWING CHALLENGES

SCALABILITY

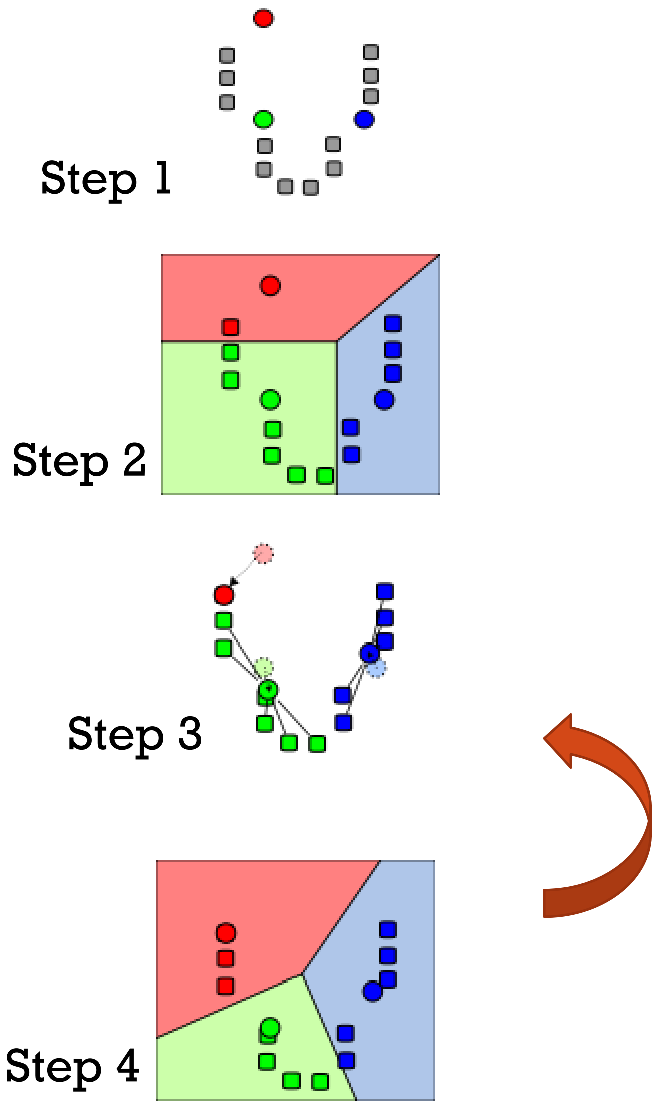
- Solution with dense or large-scale networks can be partially solved by **reducing the complexity of the information** to be visualized
- Link reduction techniques
  - Pruning the original network
- Clustering
  - Dividing the network into smaller components and treat them individually
    - Inefficient if the graph contains large components
- Dimension reduction

LINK REDUCTION TECHNIQUES

- **Removing low weight links**
  - Imposing a link weight threshold $\Rightarrow$ only link weights above the threshold are considered
  - Does not take into account the structure of the network
- **Minimum spanning tree**
  - Link reduction to $\mathbf{N - 1}$ edges (on network of $\mathbf{N}$ nodes)
- **Network scaling algorithms**
  - e.g., pathfinder network scaling
    - Extracts paths of length at most $\mathbf{Q}$

CLUSTERING TECHNIQUES

- Goal: to divide a large data set into a number of sub-sets according to some given similarity measures
- Basic methodologies:
  - The choice is a trade-off between quality and speed
  - **Graph-theoretical**
    - Relies on a pre-computed distance matrix
    - Based on how objects are separated
    - e.g., single link (similarity of their *most similar* members), complete link (similarity of their *most dissimilar* members), …
  - **Iterative**
    - Iterative optimization of the clustering structure according to a heuristic function – k-means clustering
      - Repeating re-computation of centroids (step 3 + 4)



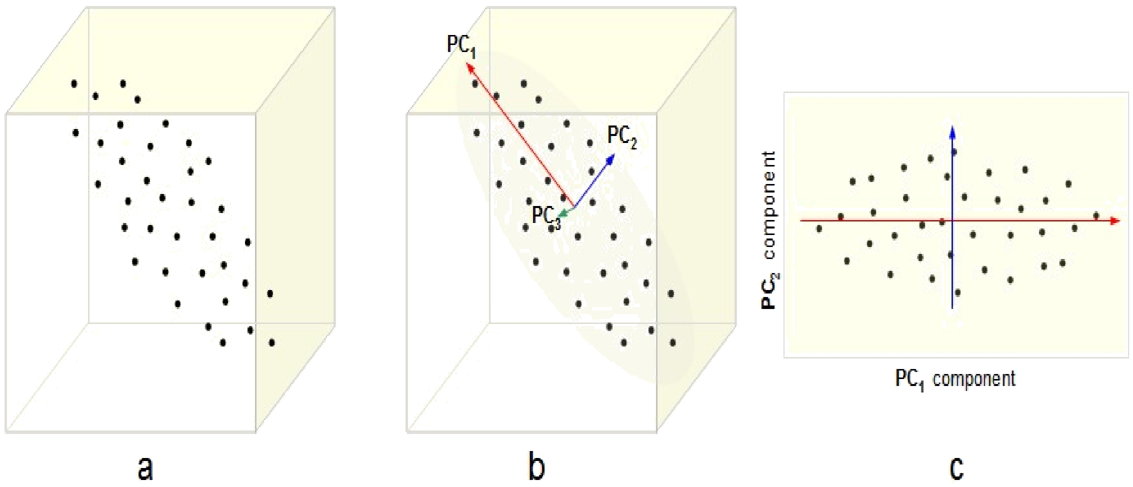
DIMENSION REDUCTION (DR)

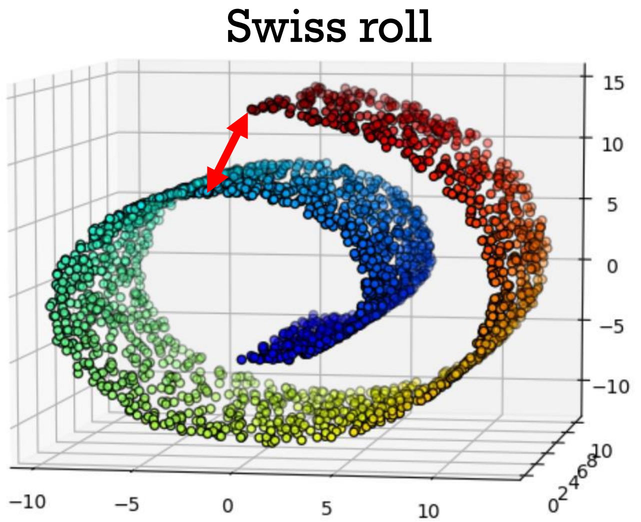
- When dealing with data which have relations expressed as a distance matrix only
  - Either we can use graph drawing methods
  - Or specialized dimension reduction techniques
- Idea: each **data point** consists of **multiple attributes** and the goal is to visualize **similar data points near to each other** in 2D space = **projection** from a multidimensional space **into a 2D space**
  - Generally difficult problems since in general distance space is not metric
    - Unlike Euclidian space

DIMENSION REDUCTION

PRINCIPAL COMPONENT ANALYSIS (PCA)

- Finding a linear transformation which tries to keep as much variability in the data as possible
- Identifies new basis vectors which maximize the amount of information kept after transformation onto the new basis
- New basis vectors correspond to the eigenvectors of the covariance matrix
  - The order of an eigenvalue/eigenvector specifies its informativeness $\Rightarrow$ two first eigenvectors define a projection into 2D space keeping most of the information present in the data



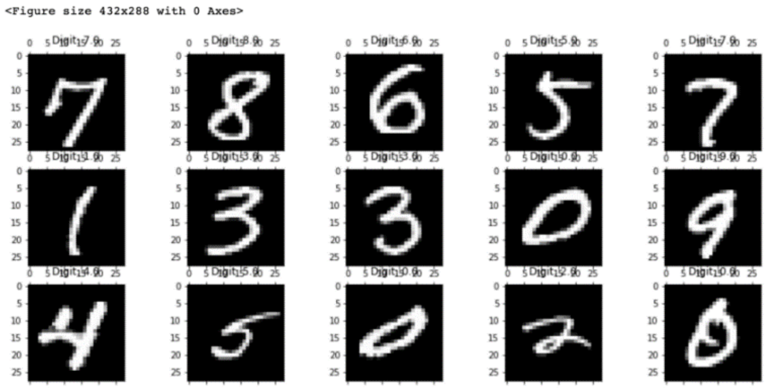


DIMENSION REDUCTION

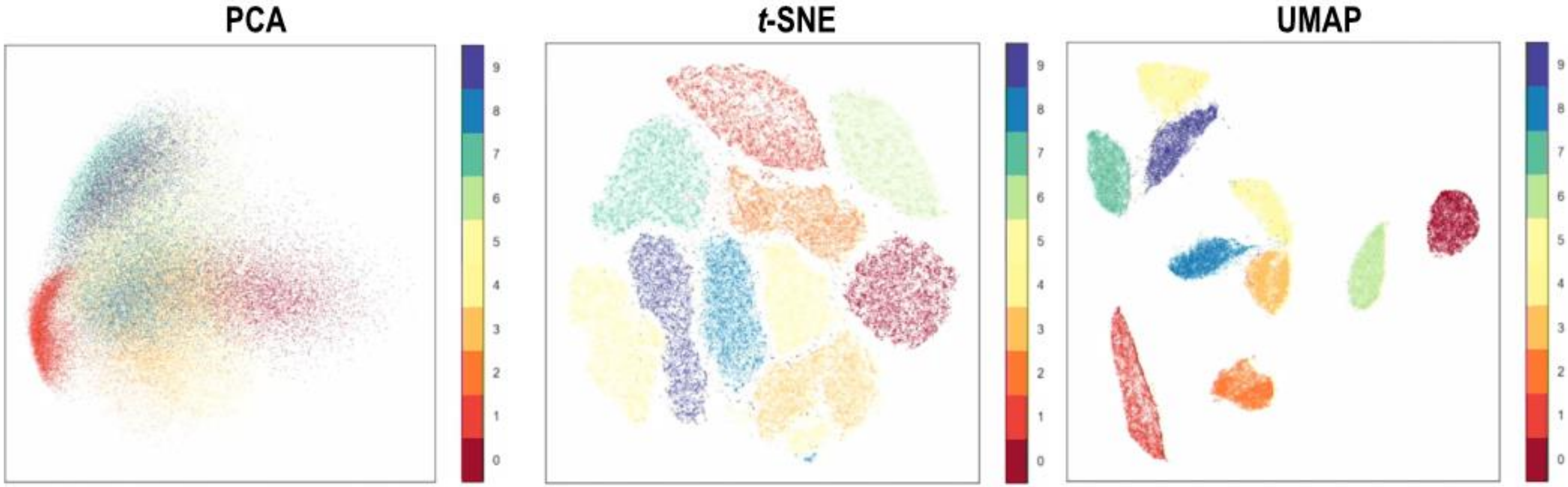
- PCA aims at preserving large pairwise distances
  - Adds most to the variance
- Data forming non-linear manifolds: points close to each other (Euclidean distance) can be in fact far apart
- Non-linear dimensionality reduction:
  - t-SNE (t-distributed stochastic neighbor embedding)
    - Models each high-dimensional object by a two- or three-dimensional point
    - Similar objects become nearby points (with high probability)
  - UMAP (Uniform manifold approximation and projection)
    - Assumption: data lie on a manifold embedded in a high-dim space which we want to project to low-dim space
    - Better preserve distances between clusters

https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a
https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668

DIMENSION REDUCTION



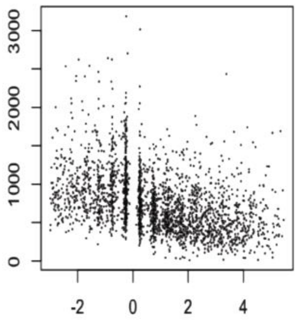
- Example: MNIST dataset
  - handwritten digits: training set = 60,000, test set = 10,000
  - 28x28 pixels = 784 dimensions



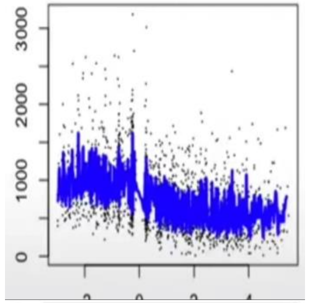
https://jlmelville.github.io/uwot/umap-examples.html#mnist
https://meta.caspershire.net/umap/

SMOOTHING

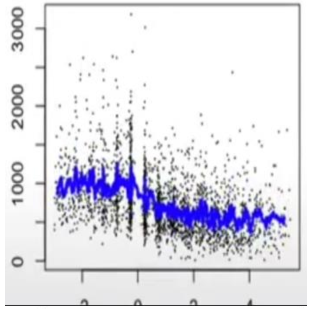
- Revealing patterns in large data
  - The patterns can be partially visible but not evident
- Techniques
  - **Moving average**
    - Representing trend using local averages
    - Sliding window and averaging values over the values
  - Locally weighted scatter plot smoothing (**LOWESS**)
    - Weights for the data points decline with their distance from center point according to a weight function
  - …



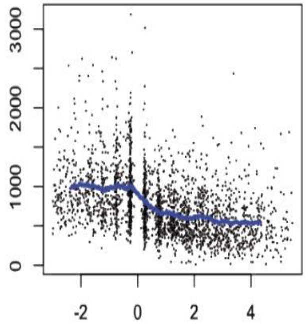
4 points window



20 points window



200 points window



DATA VISUALIZATION TOOLS

- Analytics and visualization tools
  - Standard statistical packages
    - R, Matlab
      - Customizability according to specific needs
  - Libraries for data visualization
    - Python - Matplotlib, Pandas, Seaborn,…
  - Specialized data analytics/visualization solutions
    - SAS, IBM Cognos
      - Limited by the design
      - Ready-to-use solutions on top of a data warehouse
- Visualization tools
  - Tableau, Many Eyes (IBM), Circos, Visual.ly
- Trend: to bring the visualization and analysis to common users (not only data scientists)
  - Easy-to-use software
  - Web interfaces allowing instant sharing of visualizations
  - Drag and drop interfaces

MORE INFORMATION

- Data Visualization Techniques - NDBI042
  - doc. RNDr. David Hoksza, Ph.D.
  - Summer semester

Helped to create this presentation

REFERENCES

- Edward R. Tufte: The Visual Display of Quantitative Information
- Edward R. Tufte: Envisioning Information
- Chaomei Chen: Information Visualization: Beyond the Horizon
- Manuel Lima: Visual Complexity: Mapping Patterns of Information