Alfredo Vellido

Visualization, Visual Analytics and Data Mining

An intro

IDEAI

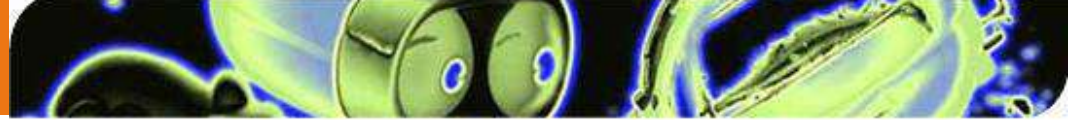
eurecat
Centre Tecnològic de Catalunya

BRAIN @ WORK
INFORMATION COMPETENCE AS BOOSTER FOR PROSPECTIVE SCIENTISTS

UPC

Now, who are you and why did you make it here?

Now, who am I and how did I make it here?



From Leonardo da Vinci to Ramón y Cajal & beyond

SPAIN

PHYSICS (electr. & autom. control) UPV/EHU

ENGLAND

NEURAL NETS (HIV) **LJMU**

NEURAL NETS (+) e-COMMERCE (B2C) **LJMU**

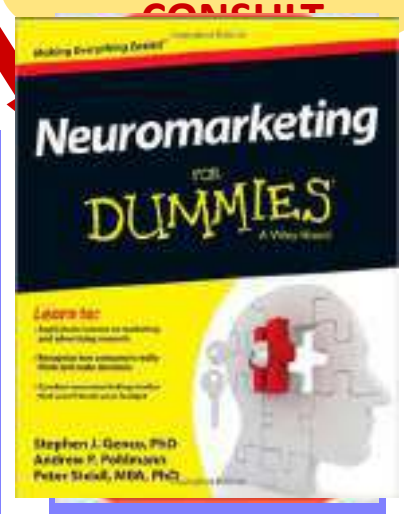
NEURAL NETS (+) **UNILEVER** Product Analysis

Data Mining iSOCO (BCN)

Associate Professor

RAMÓN & CAJAL

Data Mining



PRE-DOC

DOC

POST-DOC

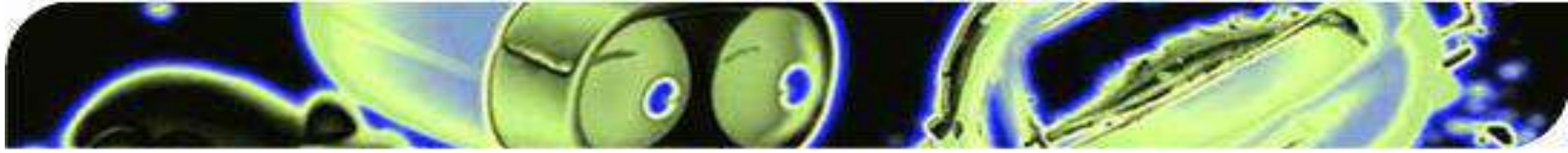
XXth Century

XXIst Century

Visual Data Mining

Contents

- A brief **introduction** to **information visualization**
- Visualization & **history**
- **Perception**: the brain is looking
- Visual **exploratory** DM
- Visual Analytics

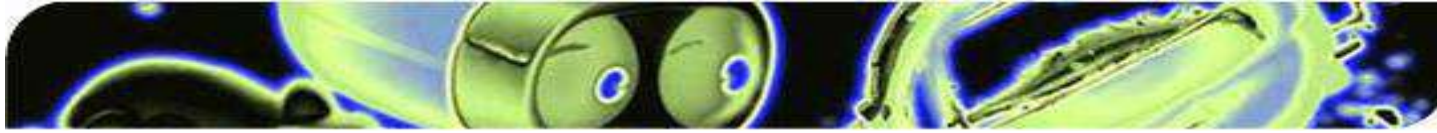


The eye of the beholder

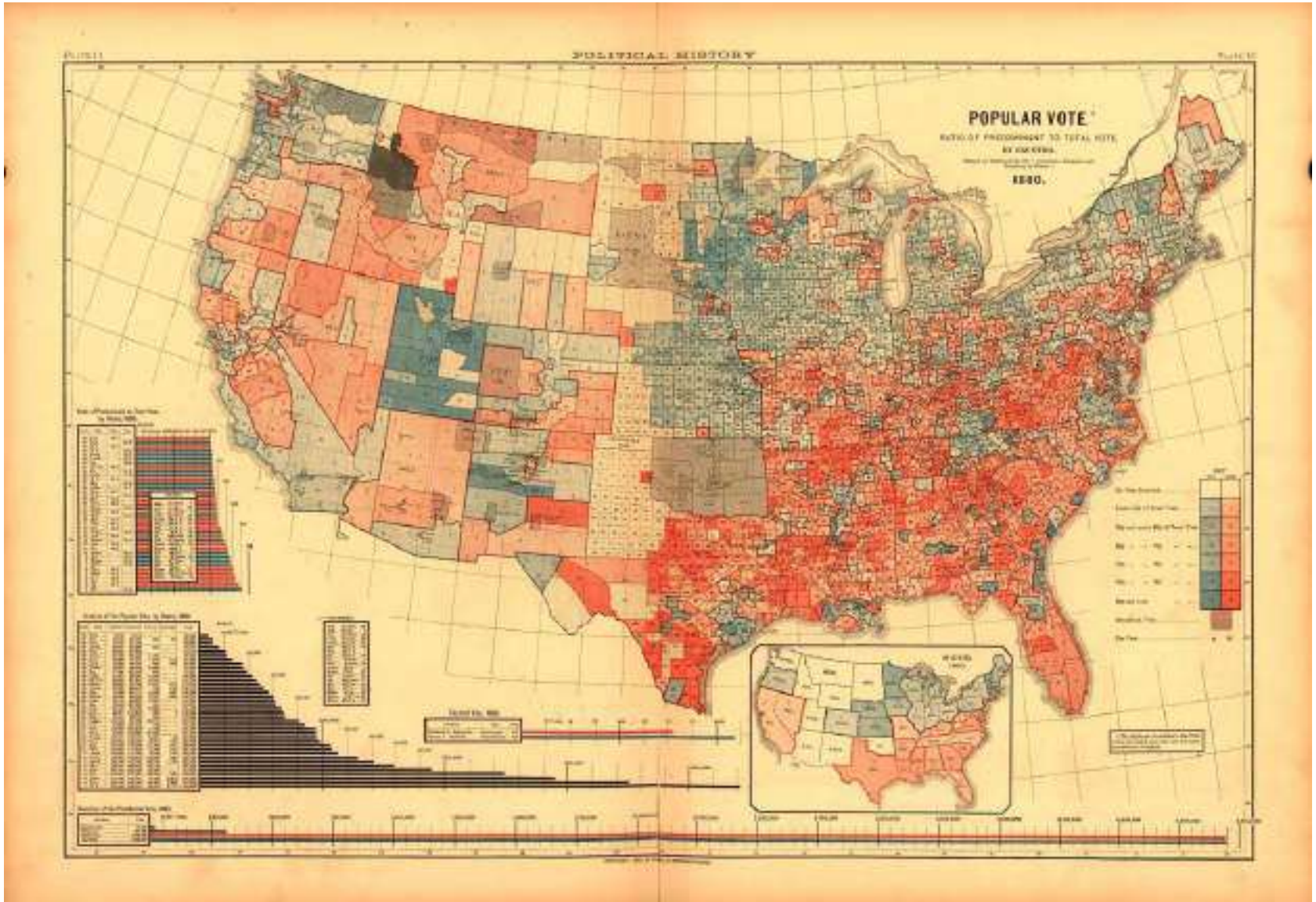
"… visualization is […] such a powerful amplifier of human abilities that it should be illegal, unprofessional, and unethical to do data analysis using only statistical and algorithmic processes"

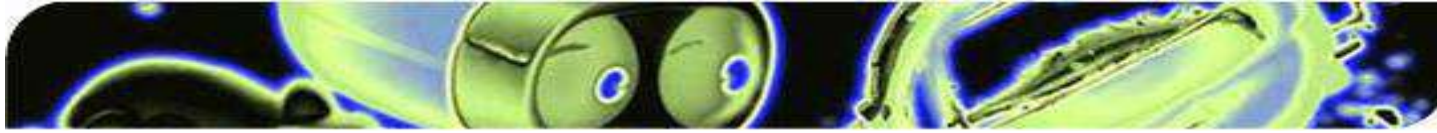
Ben Shneiderman
Information visualization and HCI pioneer

Visual DM



Map of the presidential elections: USA, 1880



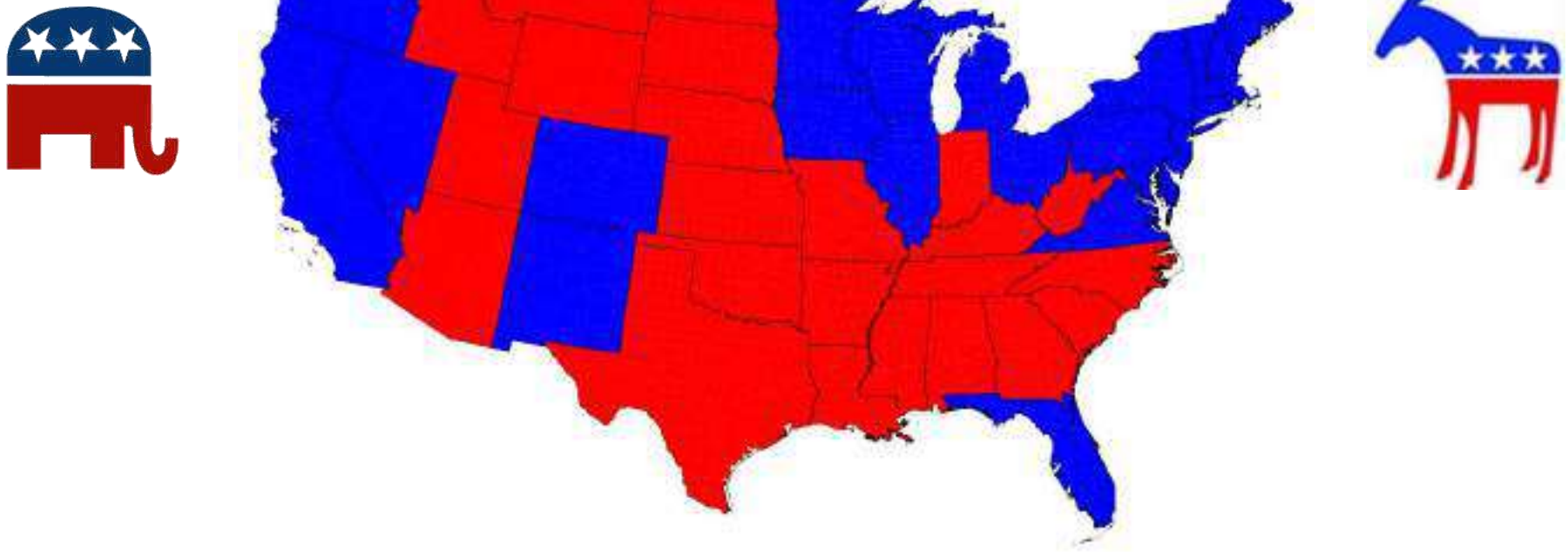


Maps and cartograms of the 2012 US presidential election results*

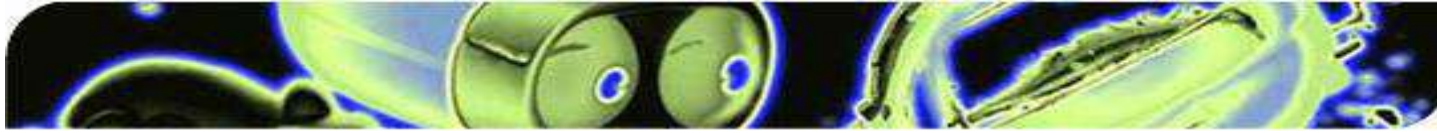
- **The map most press published after USA'12 presidential elections…**

Republicans

Democrats

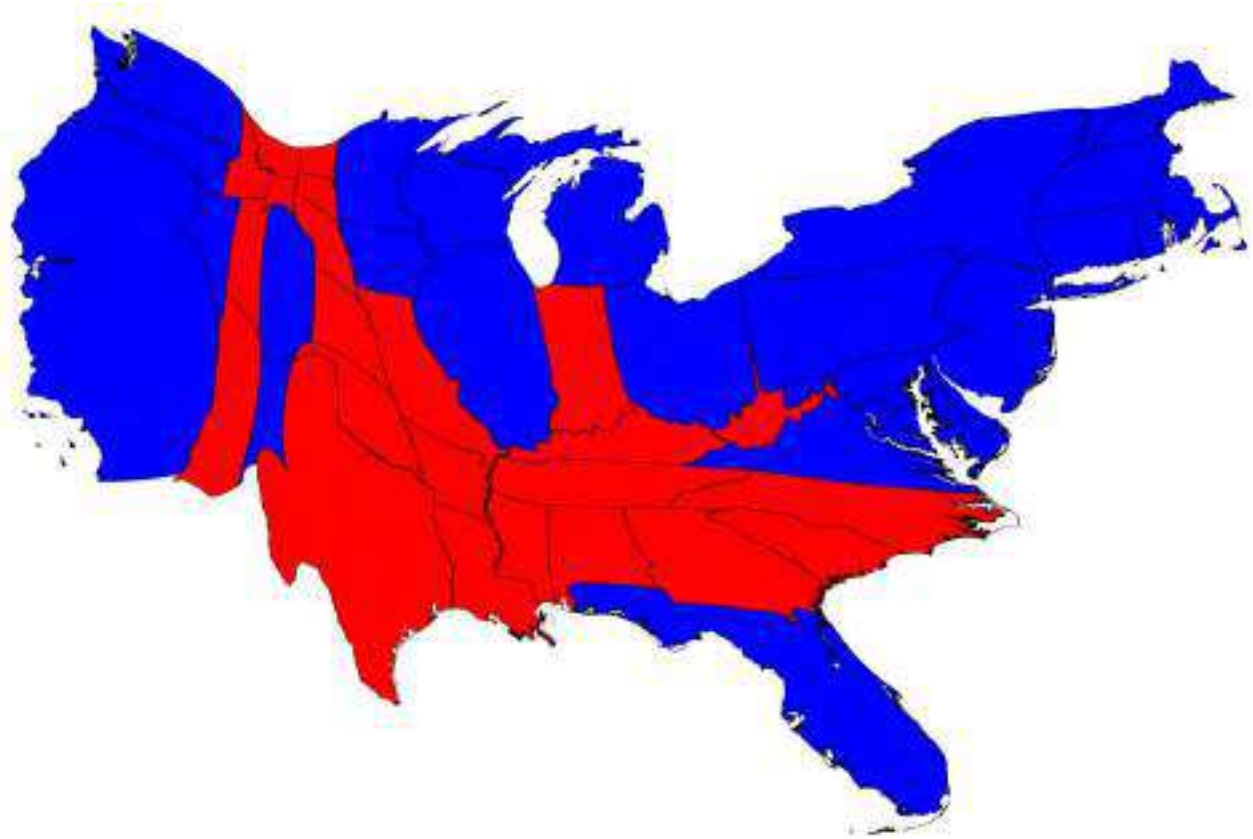


*www-personal.umich.edu/~mejn/election/ Michael Gastner, Cosma Shalizi, and Mark Newman (University of Michigan)



Maps and cartograms of the 2012 US presidential election results (2)

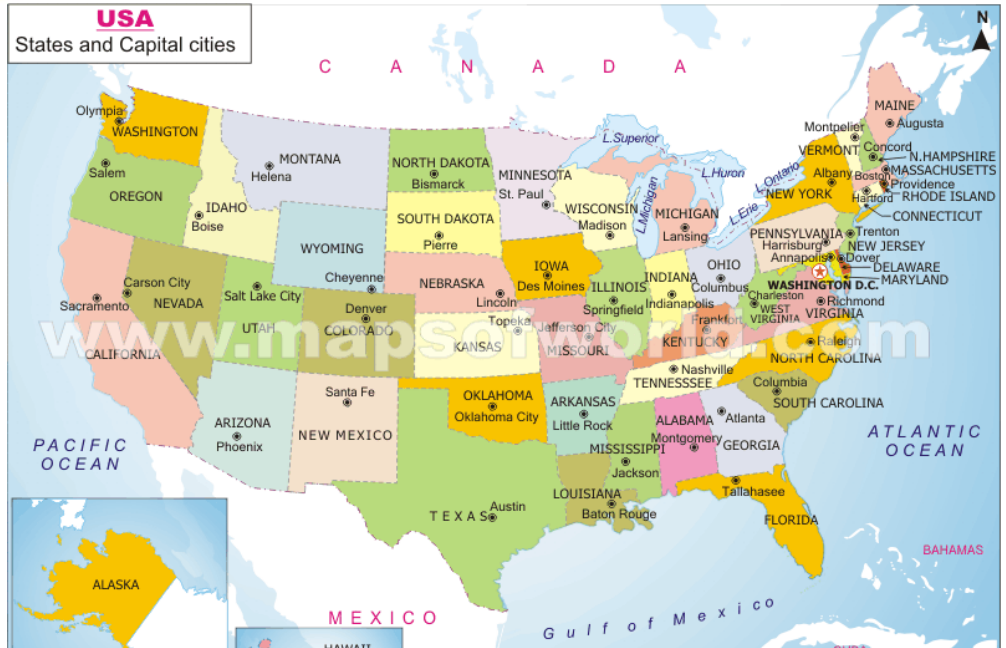
- **…which is not the same as a *"cartogram"*, corrected according to state population…**

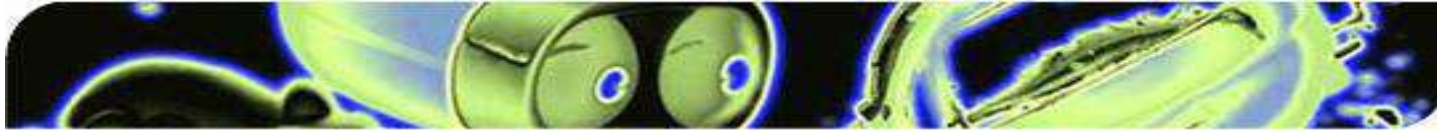


Visual DM

Maps and cartograms of the 2012 US presidential election results (4)

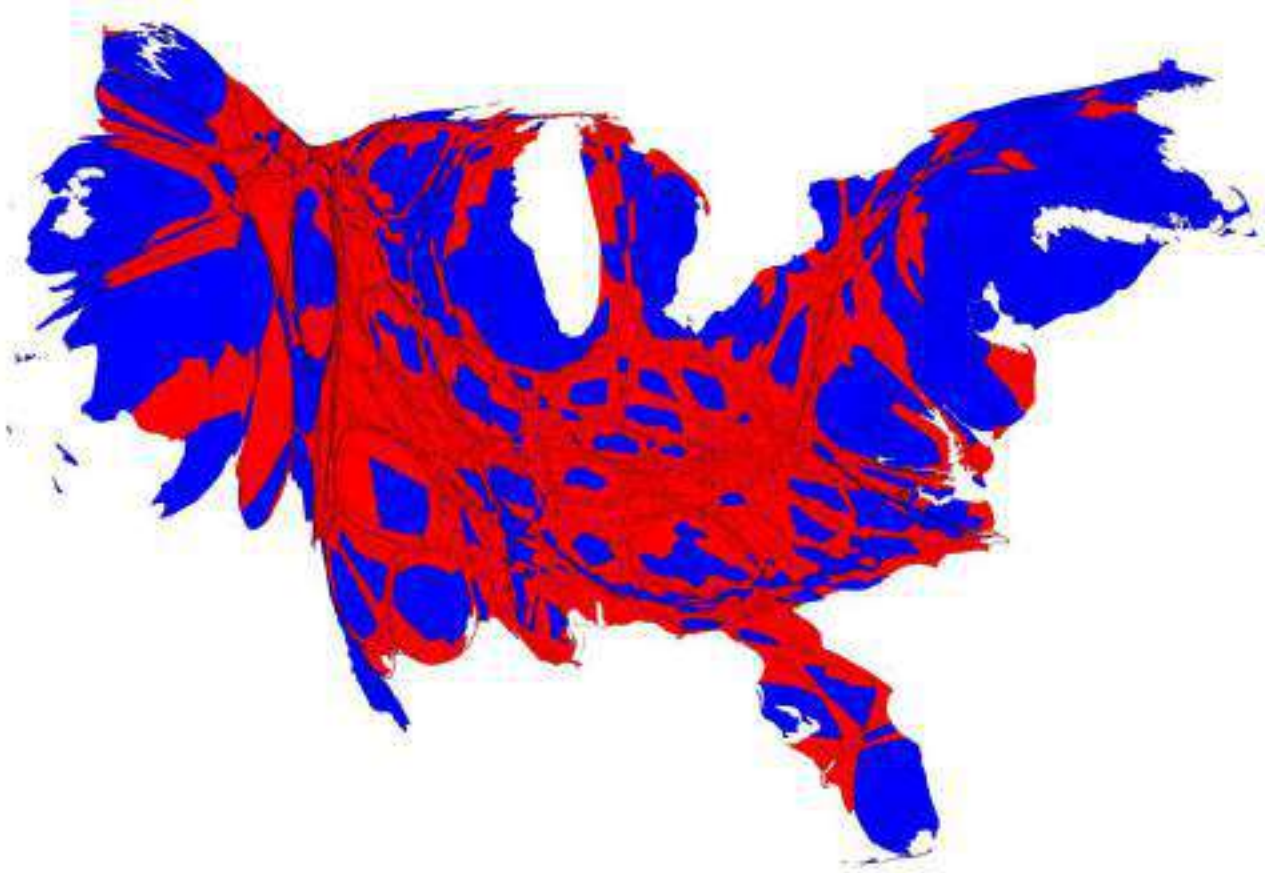
- **…and what about visualizing the results by *county*? (so *USA Today* did it!)**

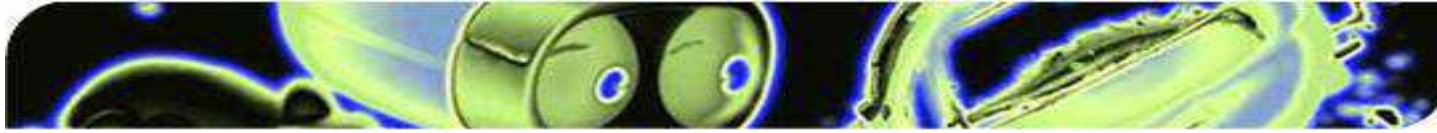




Maps and cartograms of the 2012 US presidential election results (5)

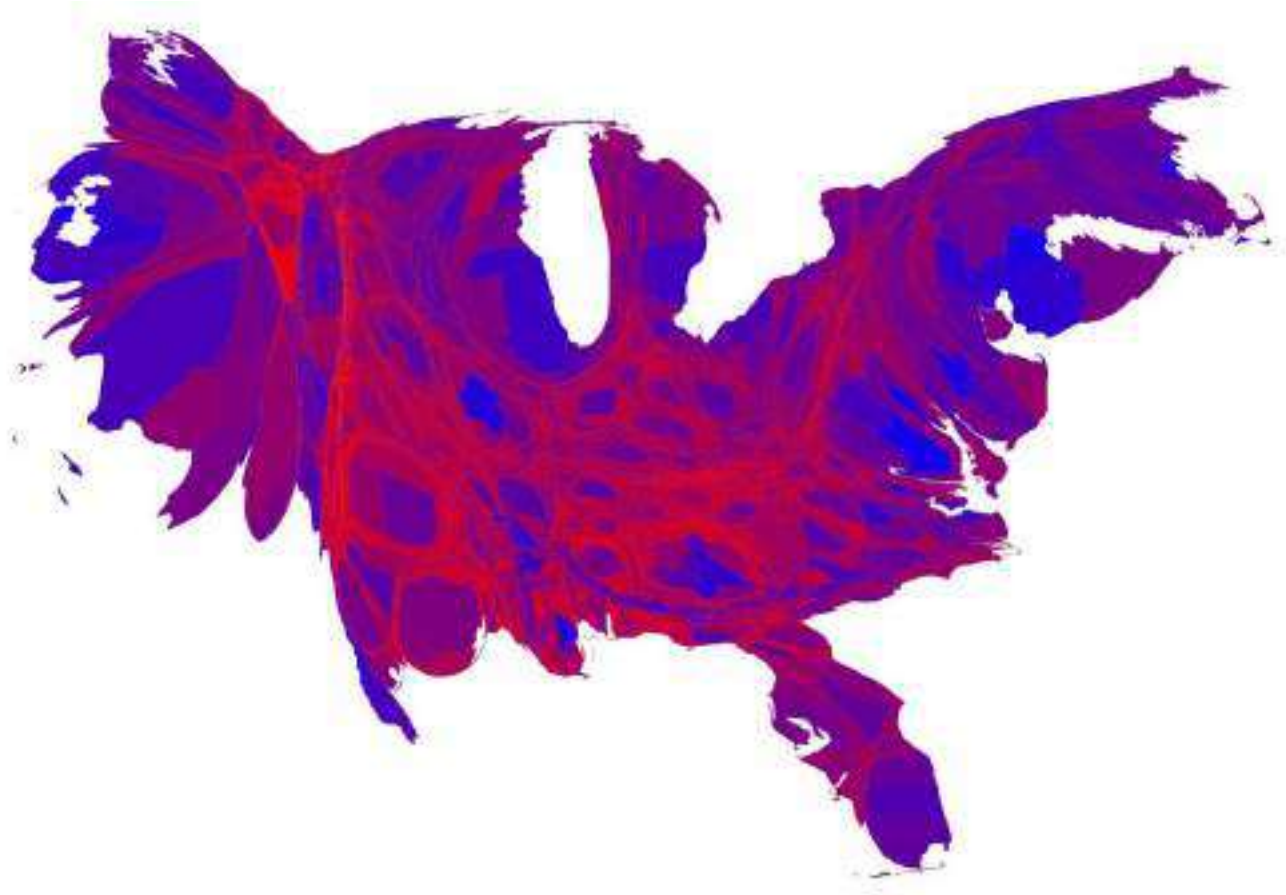
- **…again it does not look the same if we use a *cartogram*…**



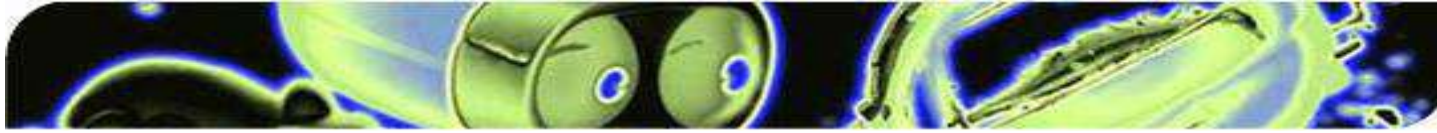


Maps and cartograms of the 2012 US presidential election results (6)

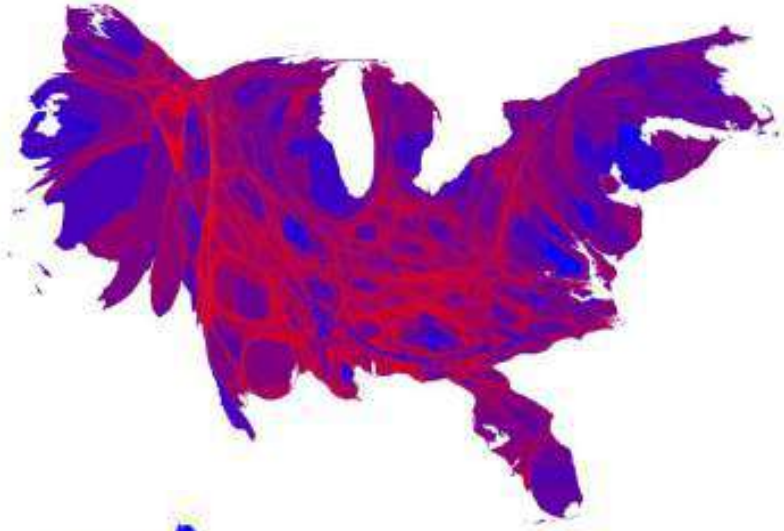
- …even less the same if we used non-linear **blue** and **red combinations** to introduce **voting percentages** (saturated at 70%)…



Visual DM

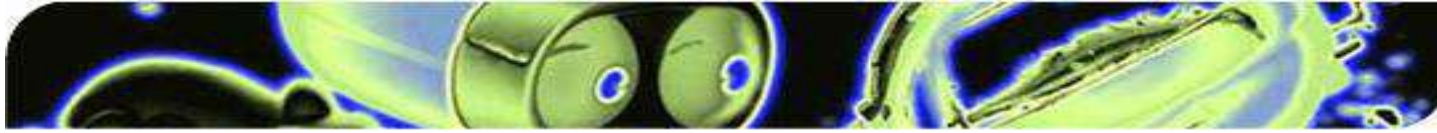


Maps and cartograms of the 2012 US presidential election results (7)

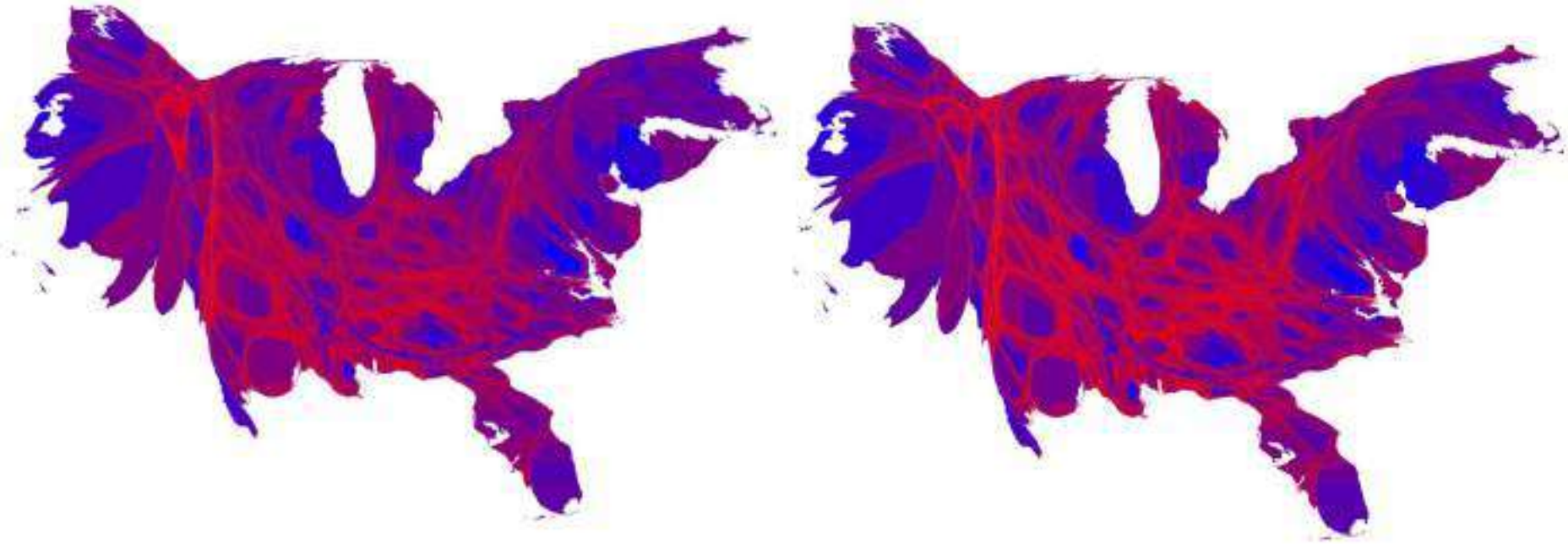


- … **COMPARE!!!**

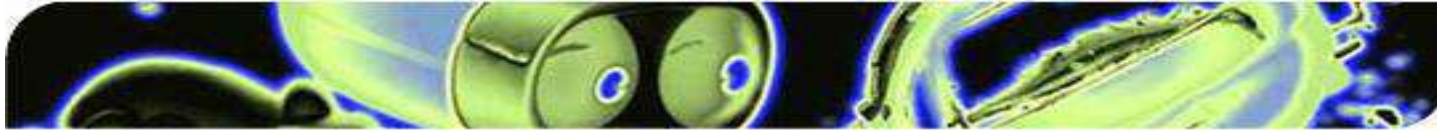
Visual DM



Maps and cartograms of the 2012 vs 2016 US presidential election results (8)

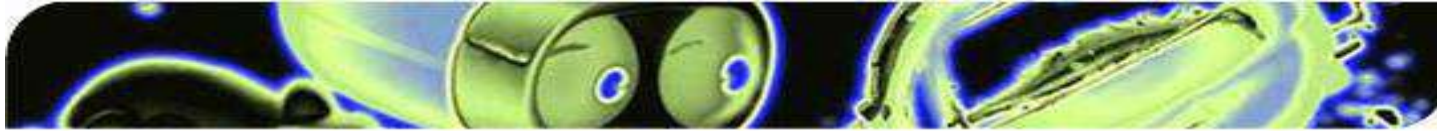


Visual DM

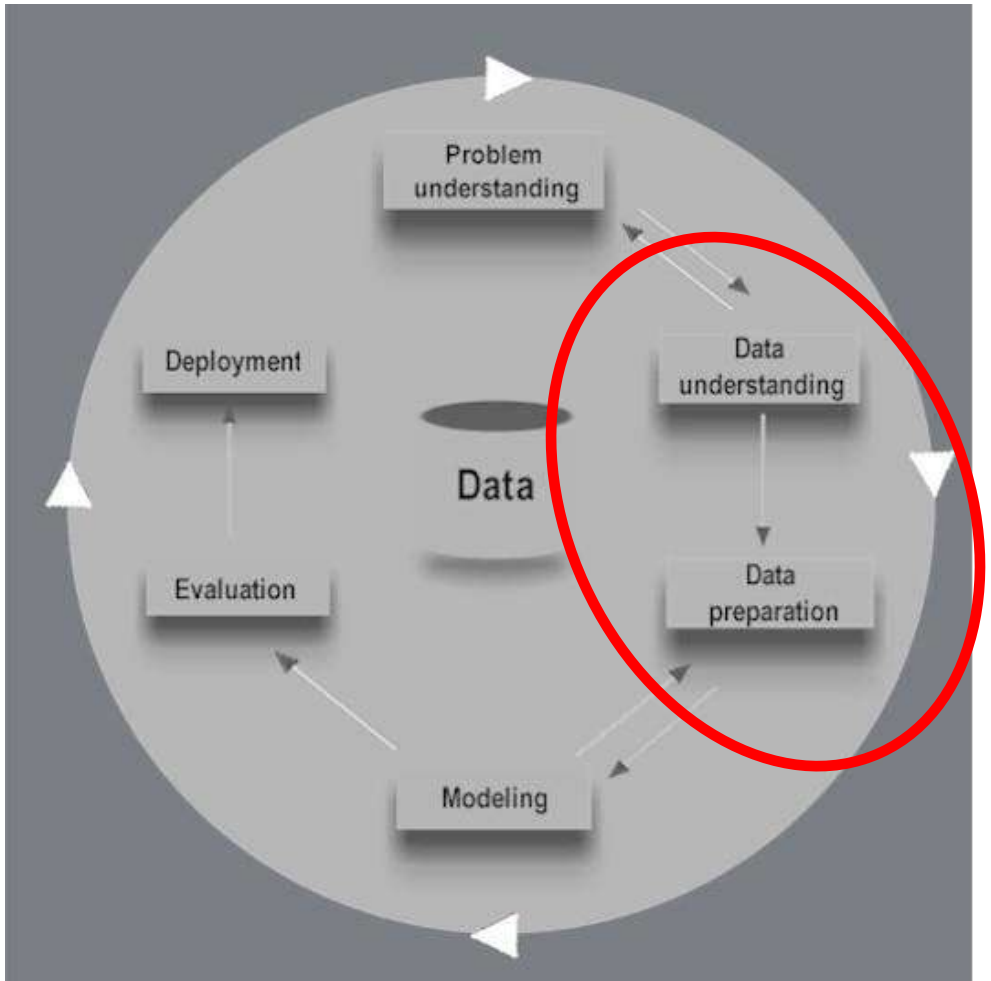


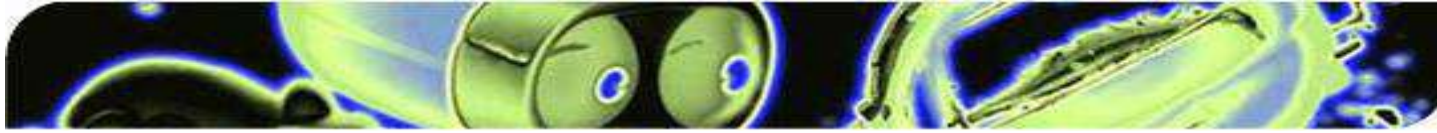
Map oldies: keep'em coming!

Visualization: where in DM?

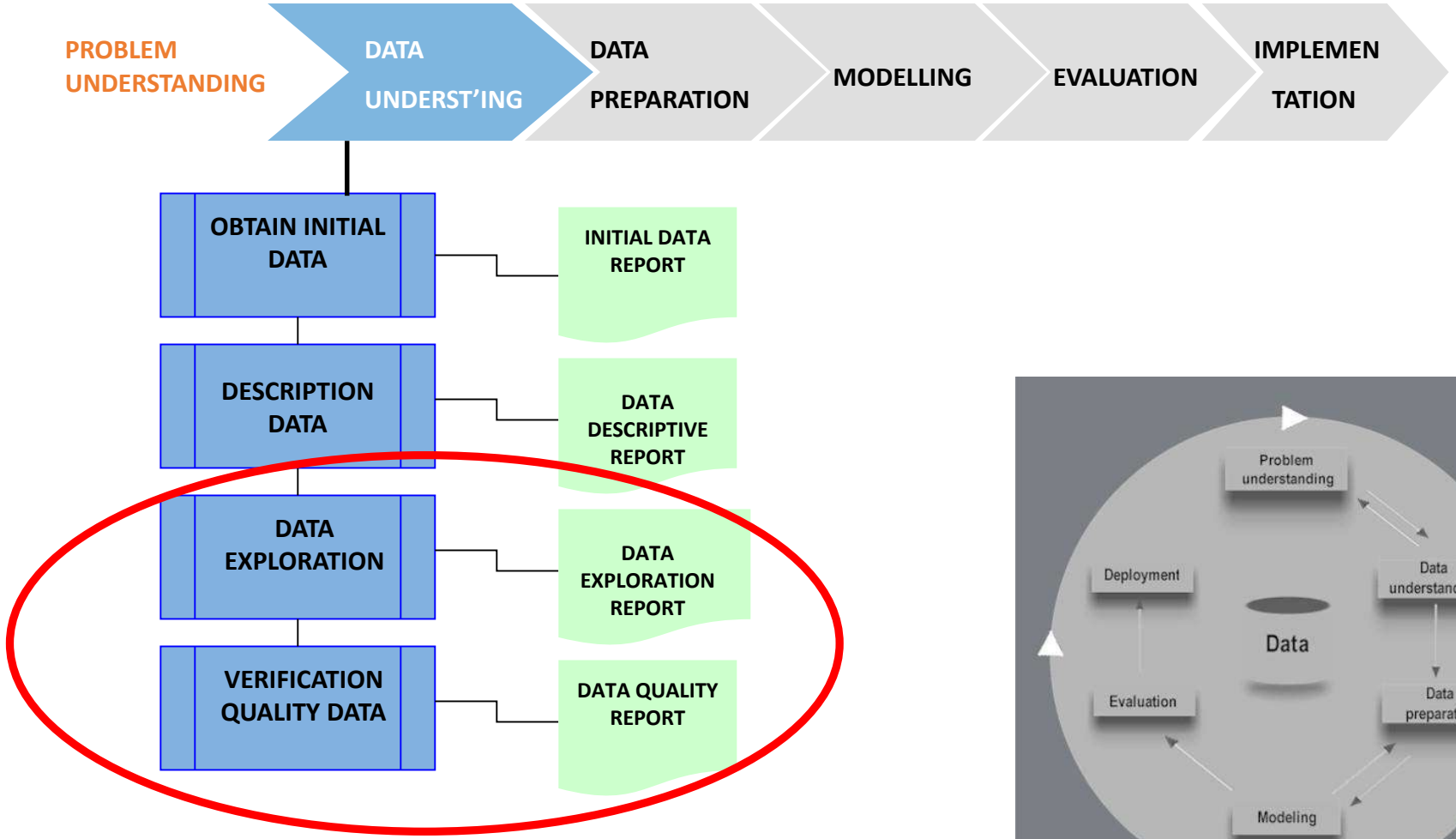


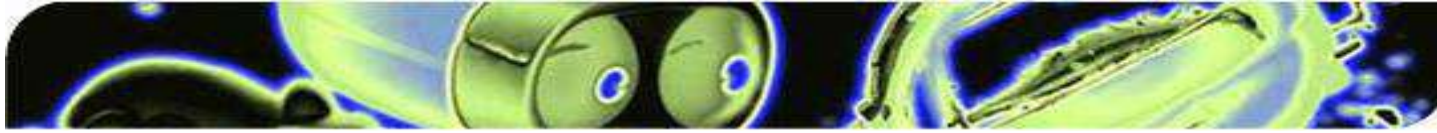
CRISP Data Mining: Methodology phases



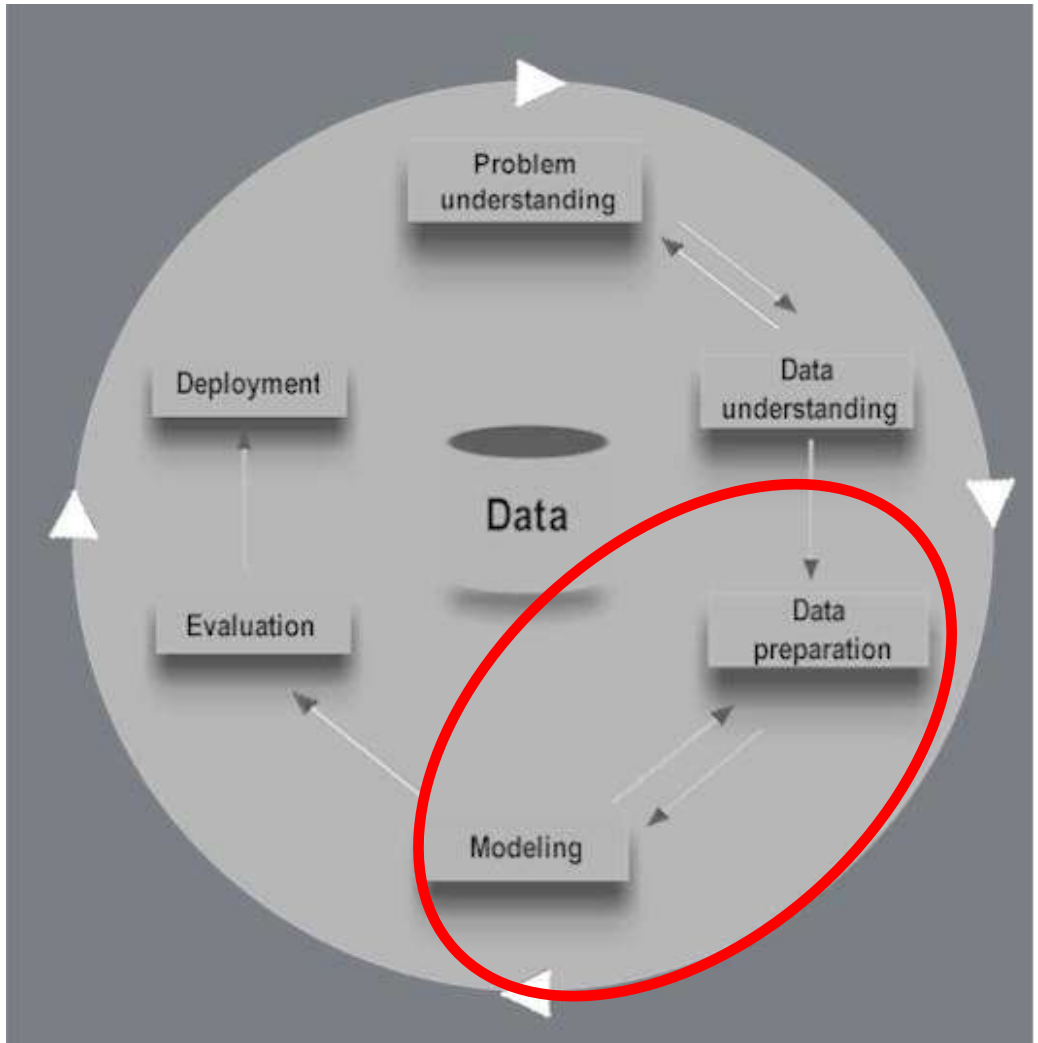


CRISP DM: Phases: Data understanding

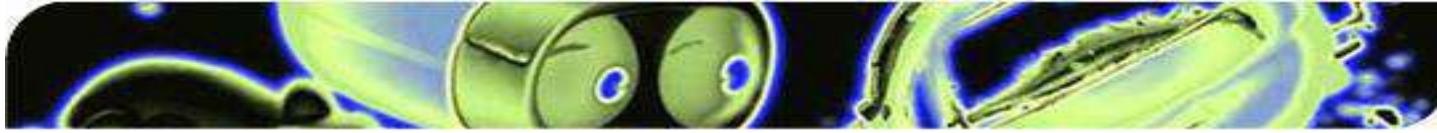




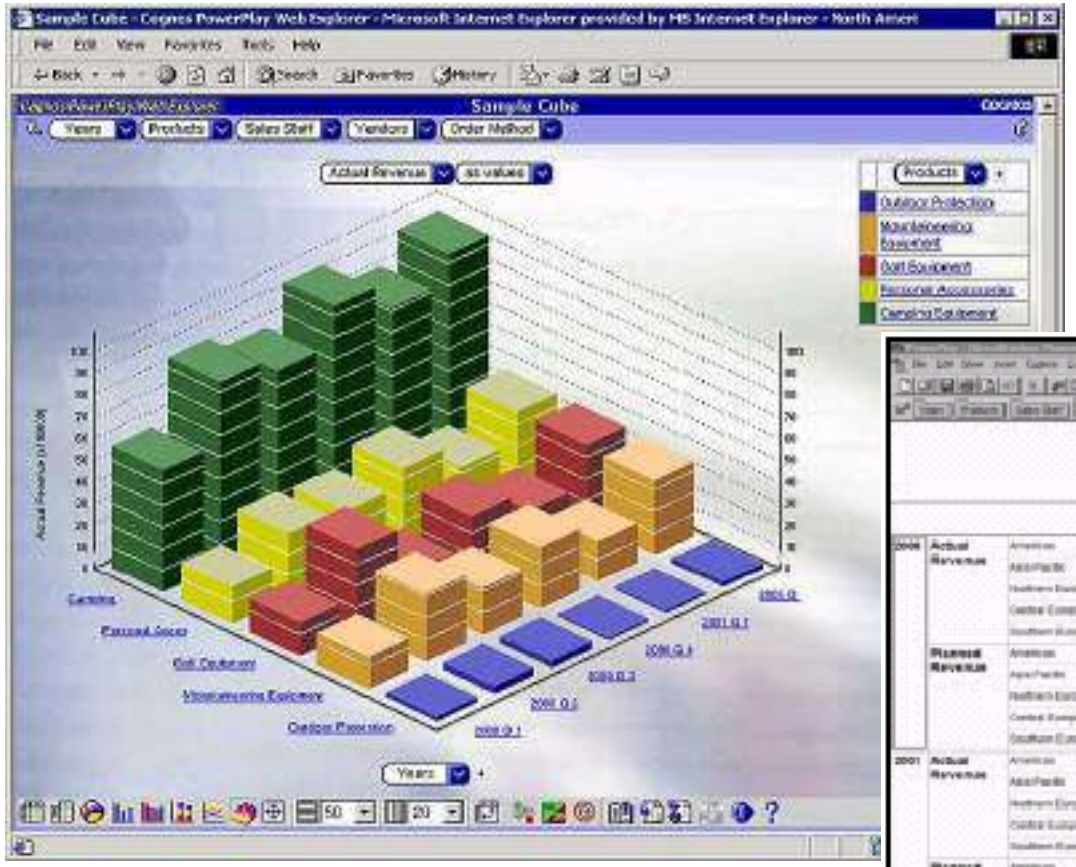
CRISP DM: Methodology phases



Visual DM

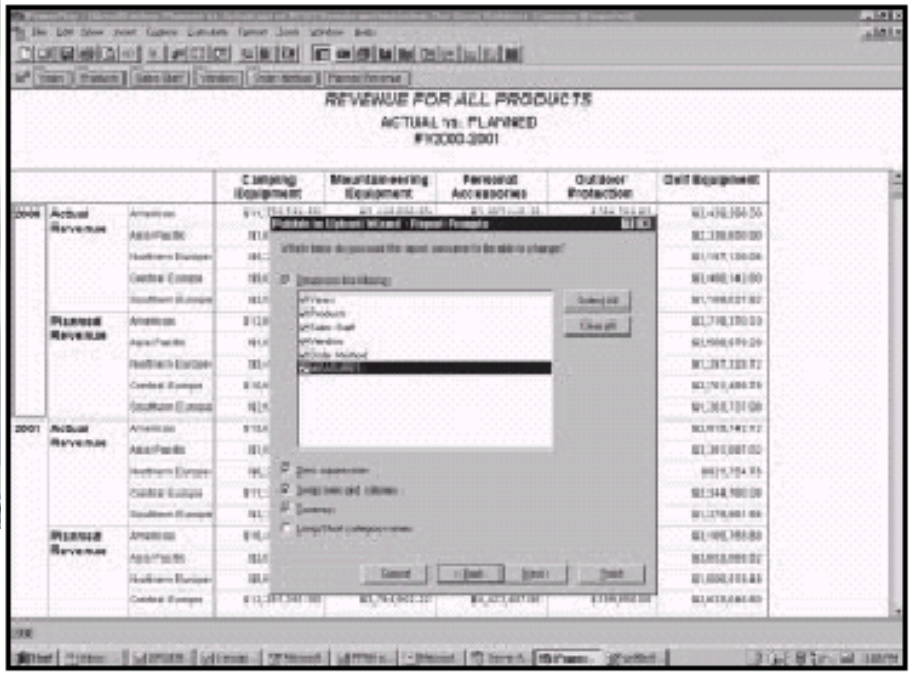


Another take:
CRISP DM / typology of DM problems / DESCRIPTION

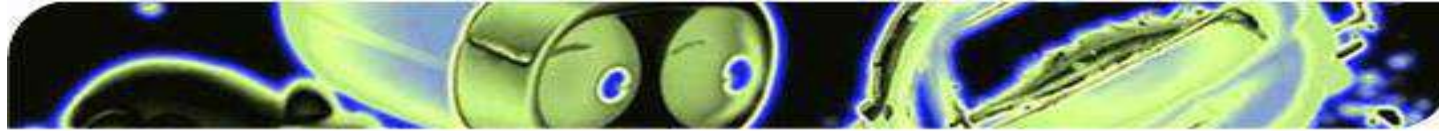


DATA visualization:

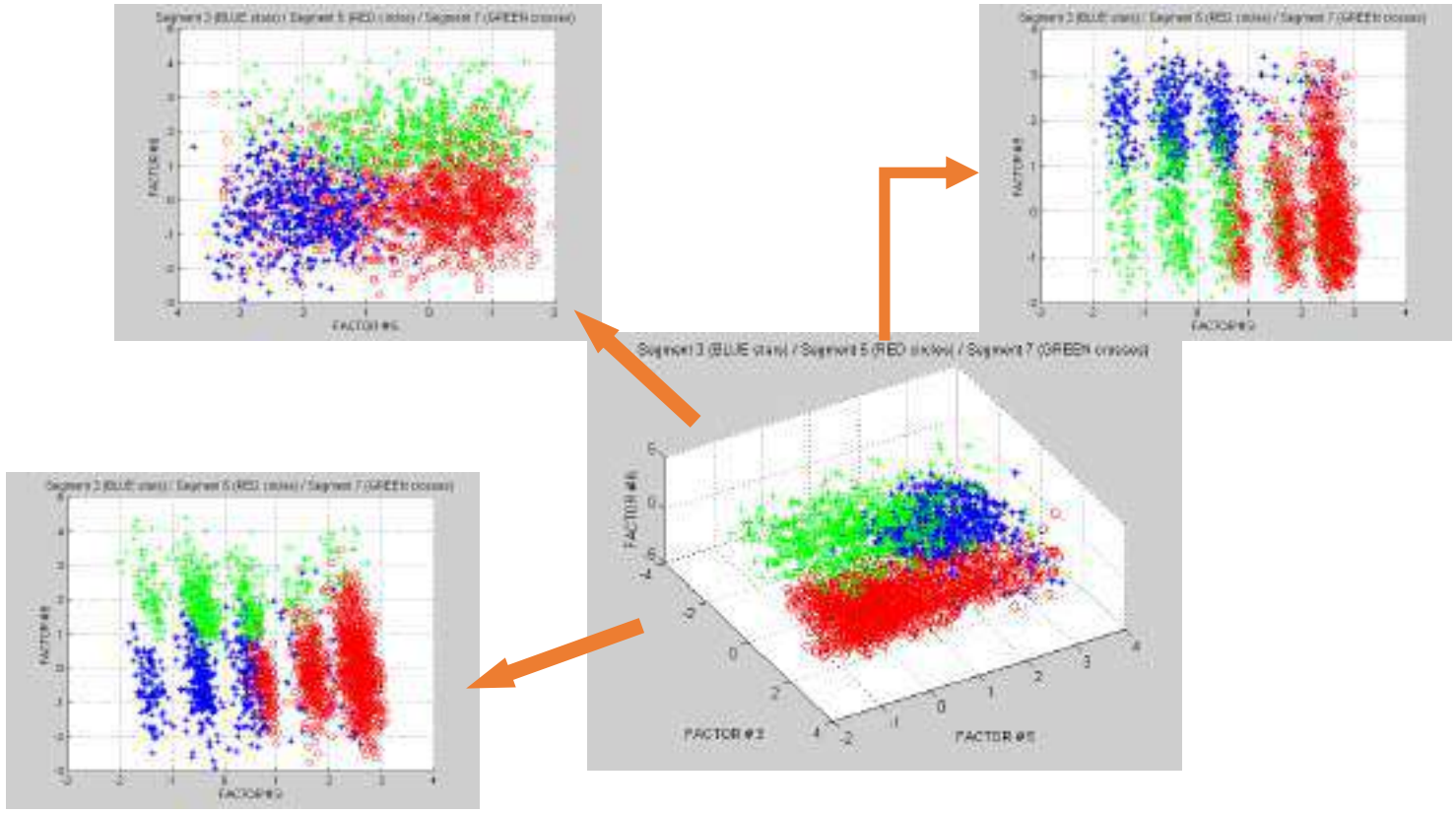
An OLAP example



Visual DM

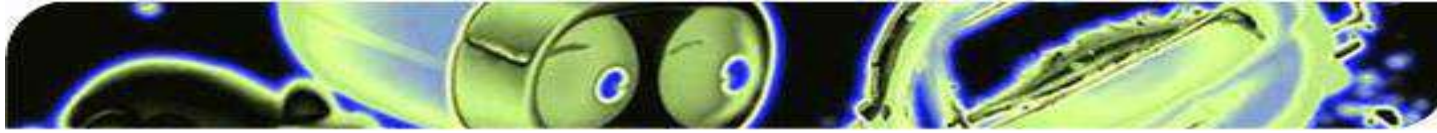


CRISP / typology of DM problems / CLUST./ SEGMENTATION (1)

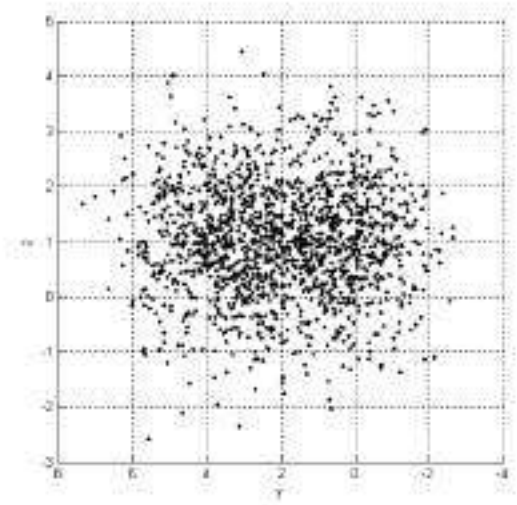


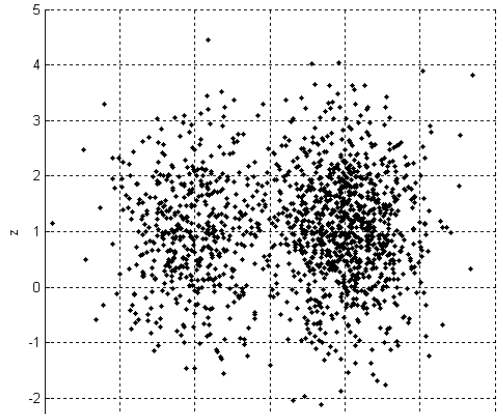
A true clustering example: MODEL visualization

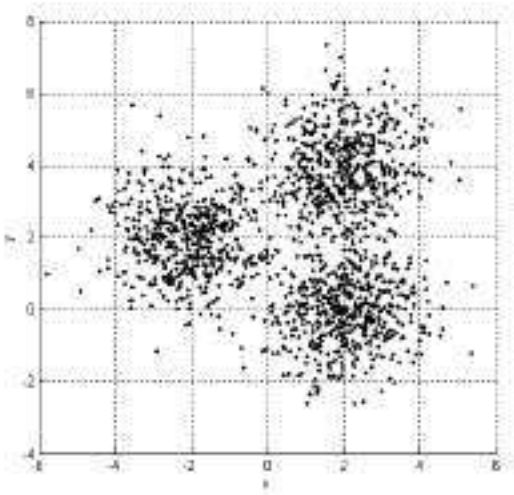
Visual DM

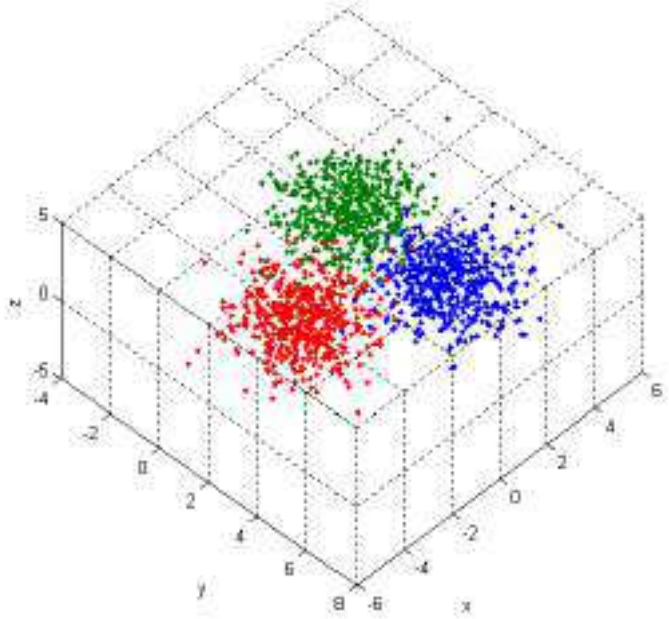


CRISP / typology of DM problems / CLUST./SEGMENTATION (2)

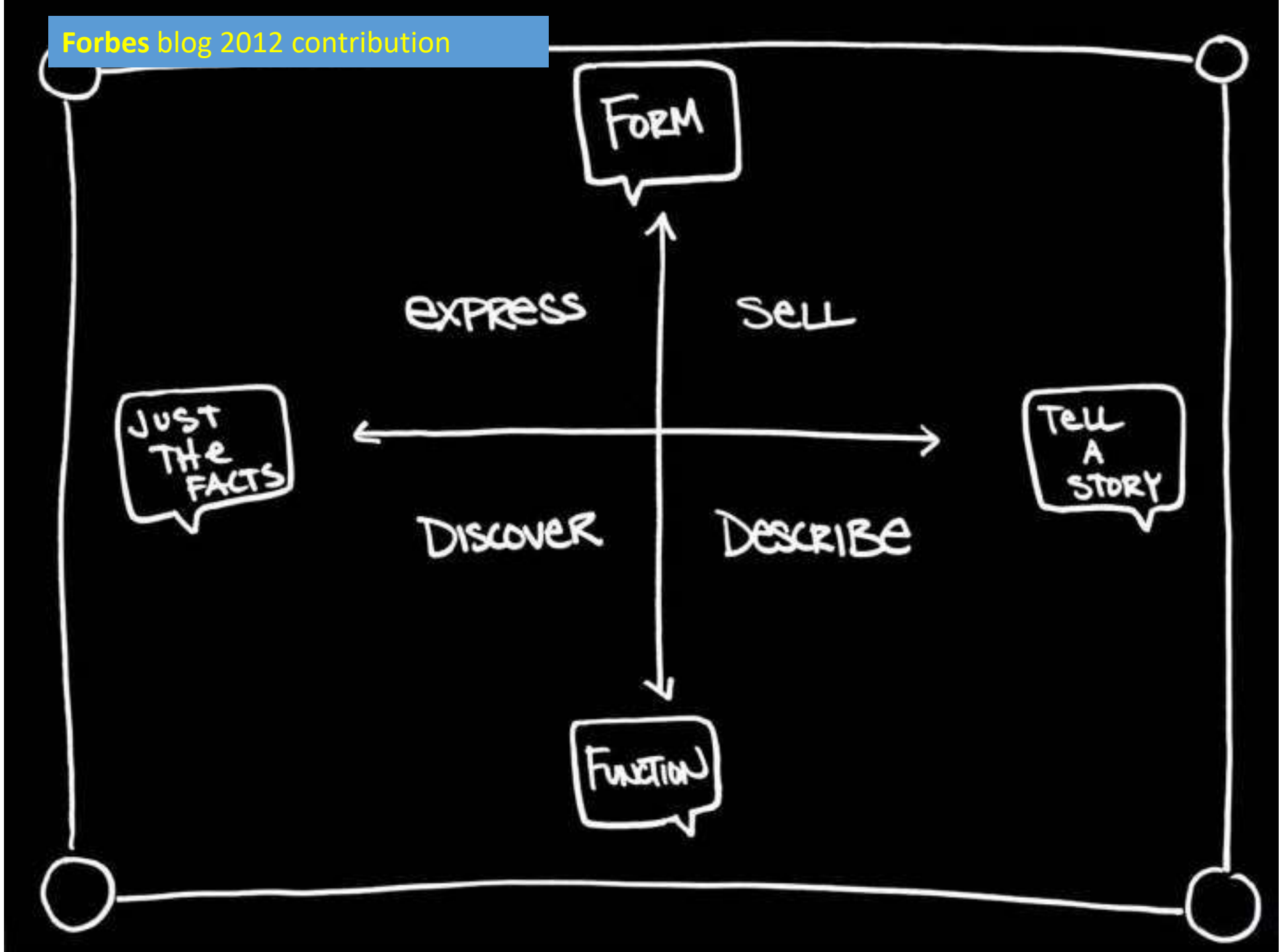






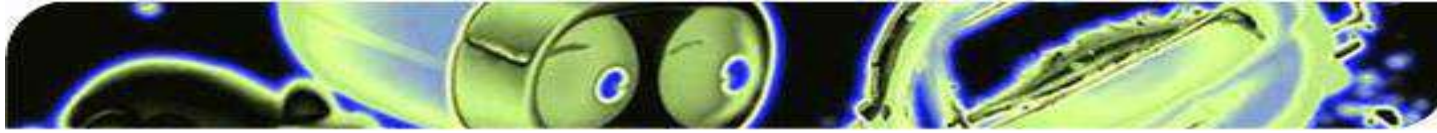


Forbes blog 2012 contribution



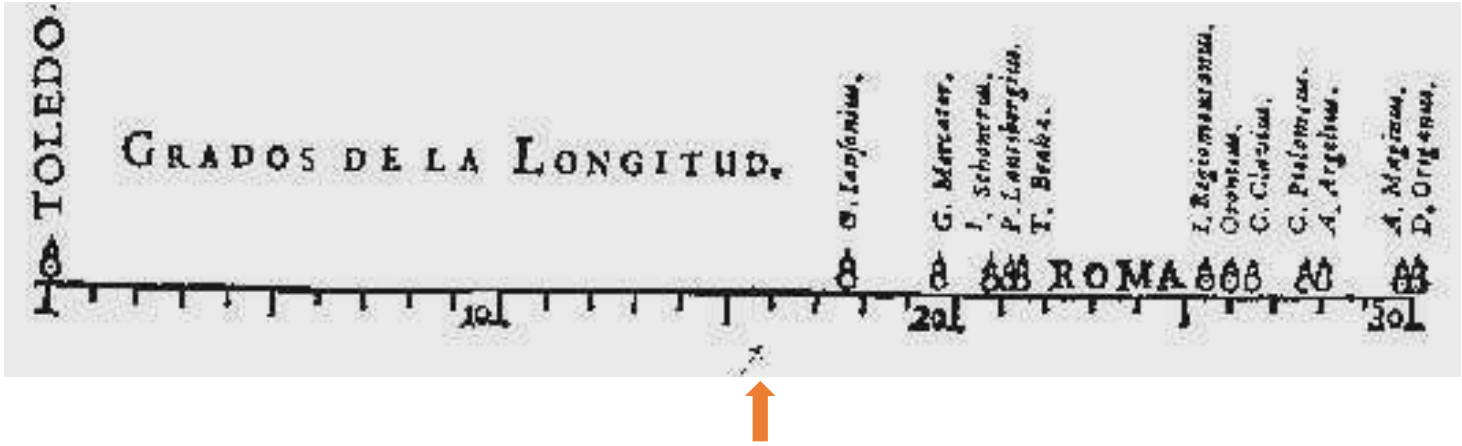
Contents

- A brief **introduction** to **data visualization**
- Visualization & **history**
- **Perception**: seeing with the brain
- Visual **exploratory** analysis

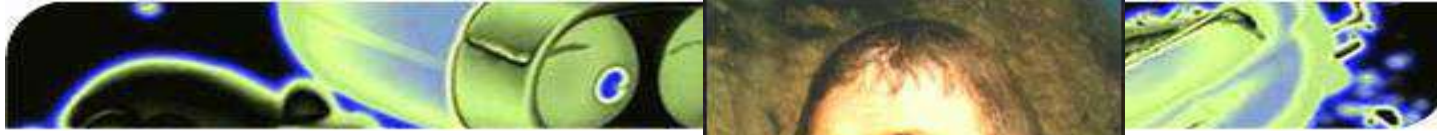


Once upon a time, *circa* 1600…

- **Michael van Langren**, in **1644**, displayed 12 estimations of the longitude **from Toledo to Rome**: This is, possibly, the earliest visualization of statistic data kept on record. A fuzzy arrow indicates the correct longitude (16º30'); All estimations at the time were well off-mark (The word **ROMA** signals Langren's own average estimation).



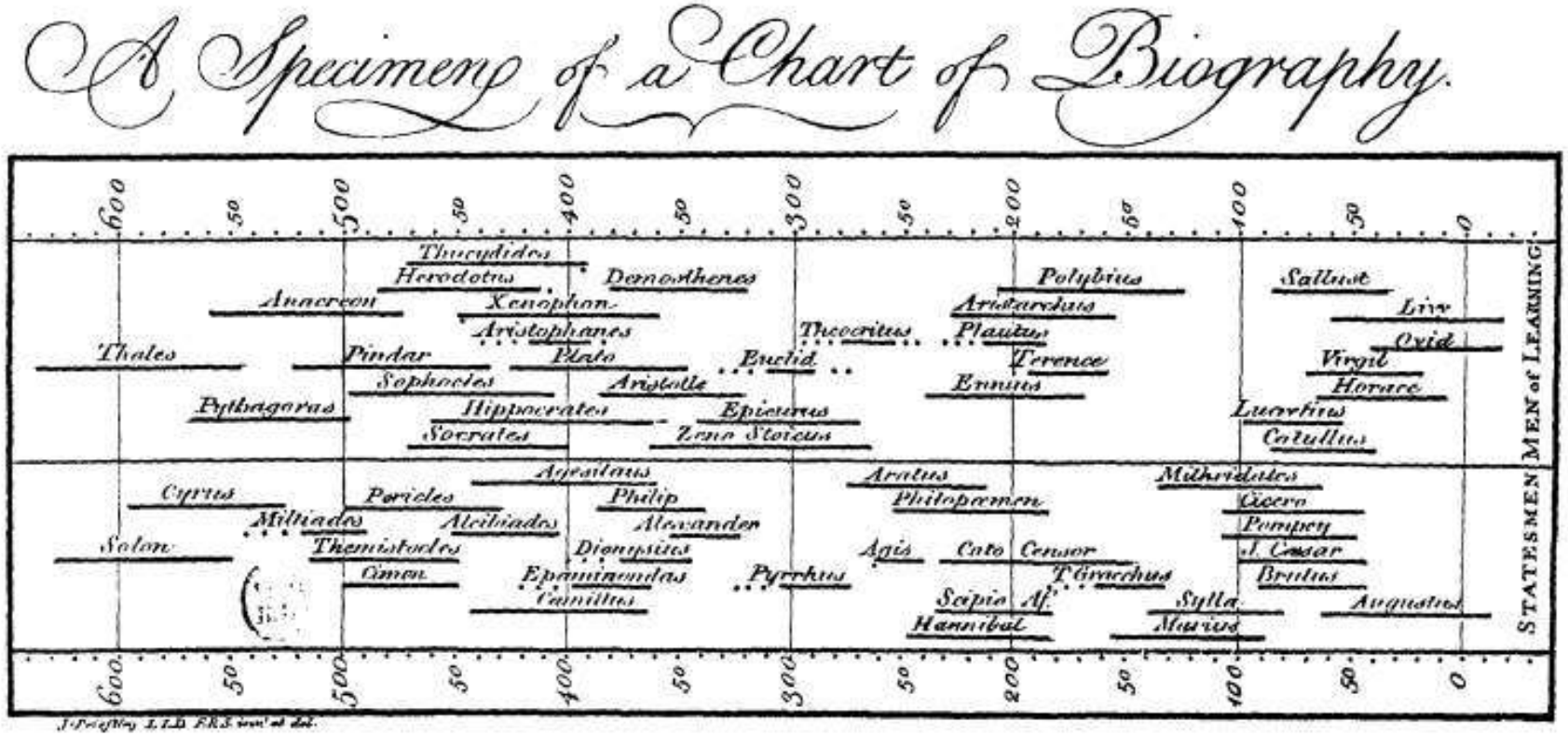
Visual DM



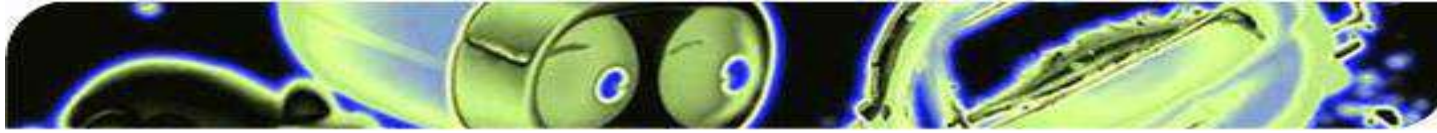
… and reaching 1700…

- **Joseph Priestley** generated this pioneering chart graphic display of v.i.p.'s lives.

(Source: Joseph Priestly, ***A Chart of Biography*, 1765**)

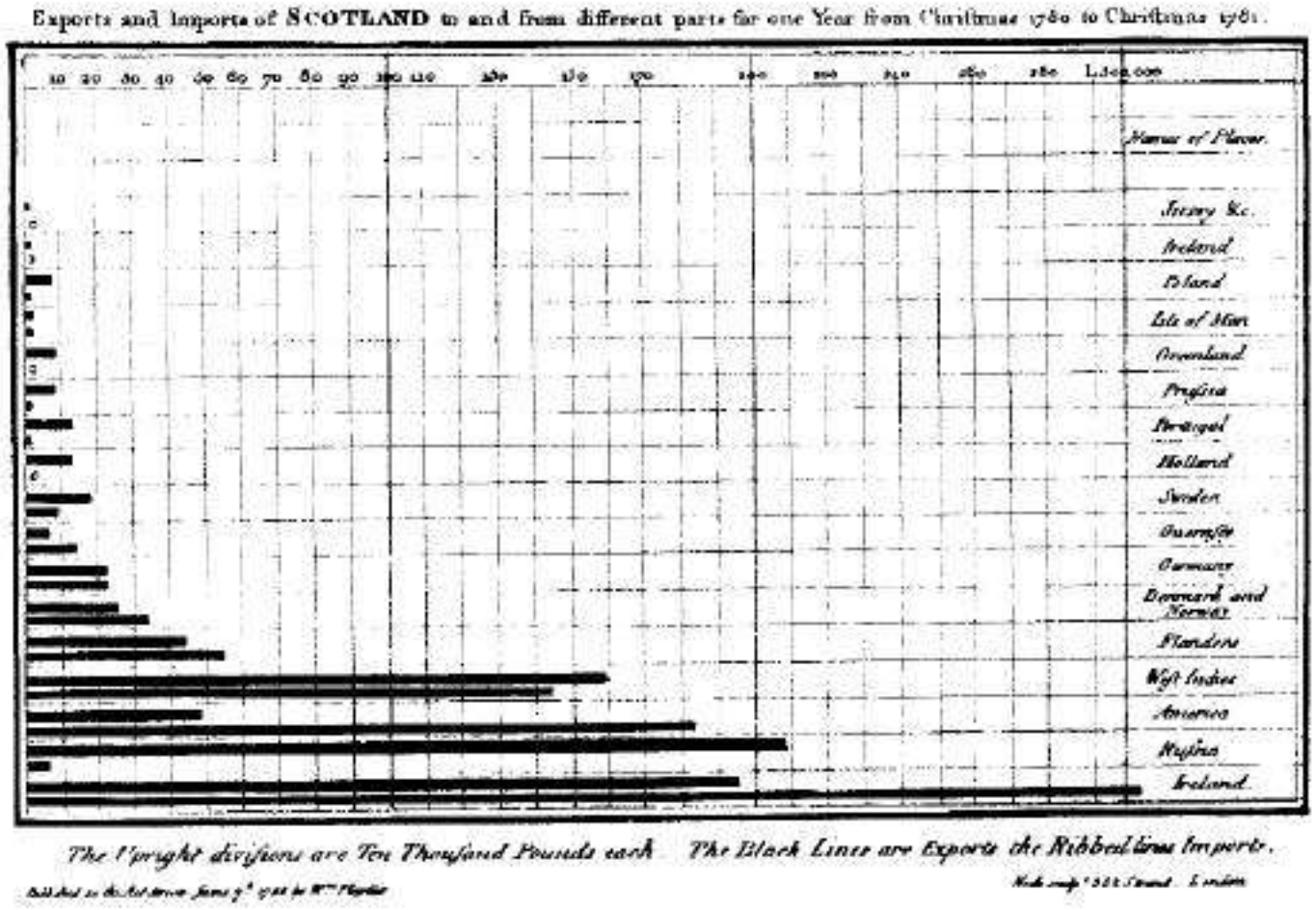


Visual DM

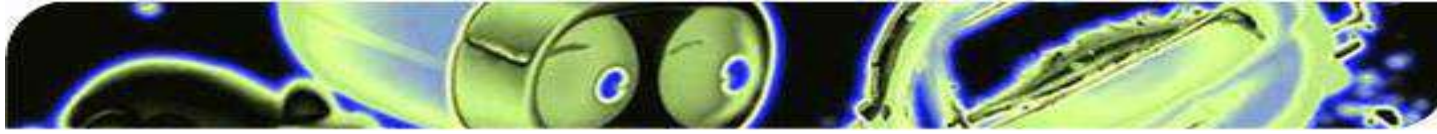


… introducing the industrial revolution …

- ***William Playfair*** (**XVIII-XIX**) explicitly argued that charts communicated better than tables of data. He was credited with inventing the **line, bar, and pie charts**.



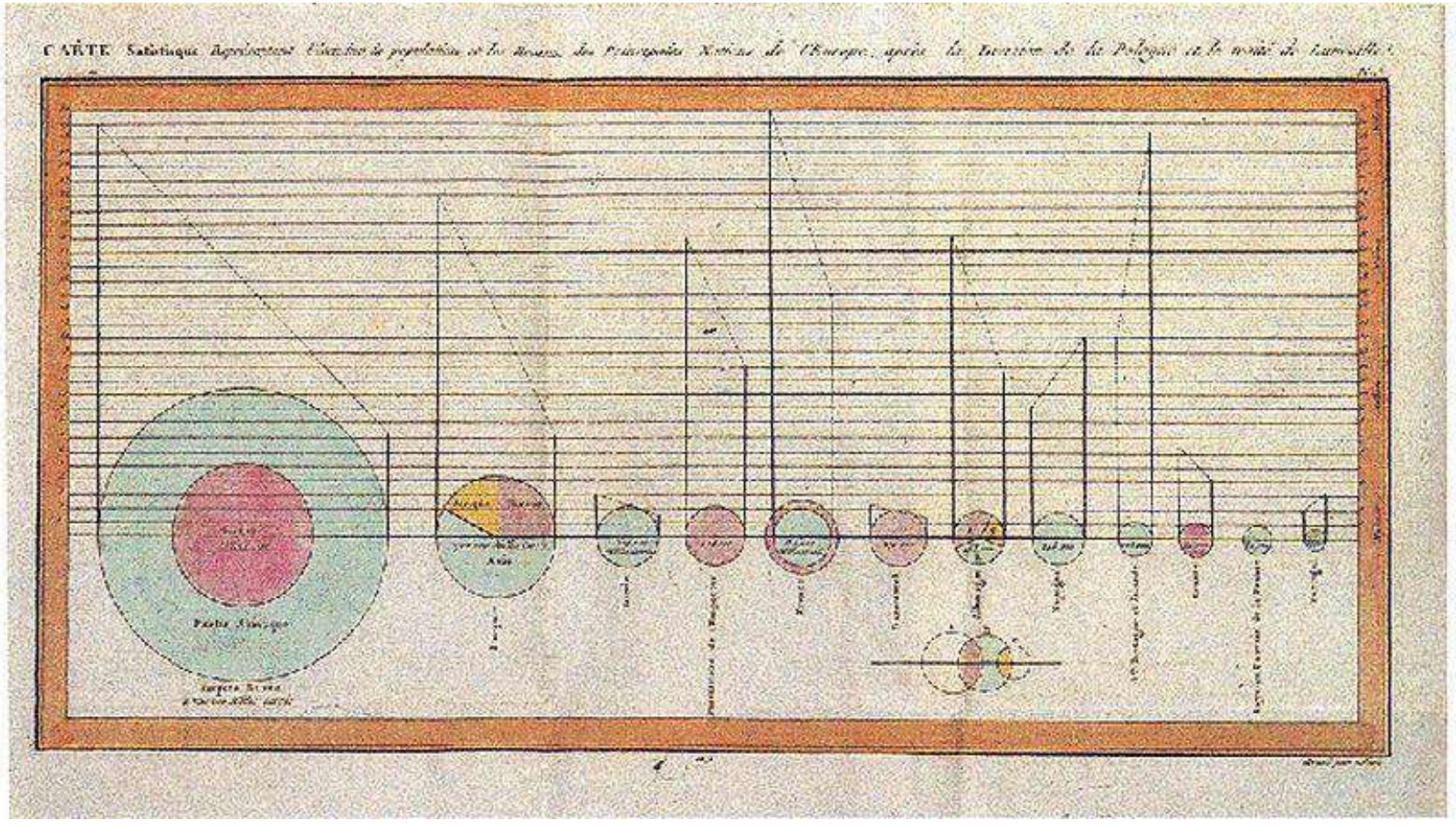
Visual DM



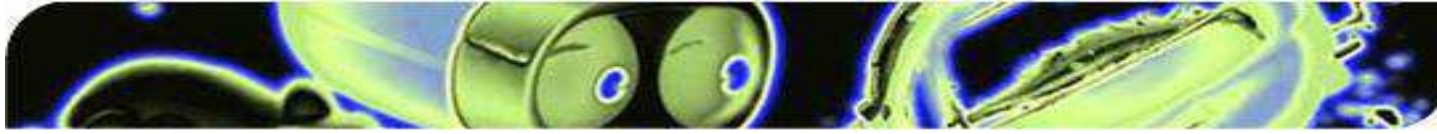
… introducing the industrial revolution …

- ***William Playfair*** : an example of **pie chart**.

Source: *"The Commercial and Political Atlas and Statistical Breviary"*



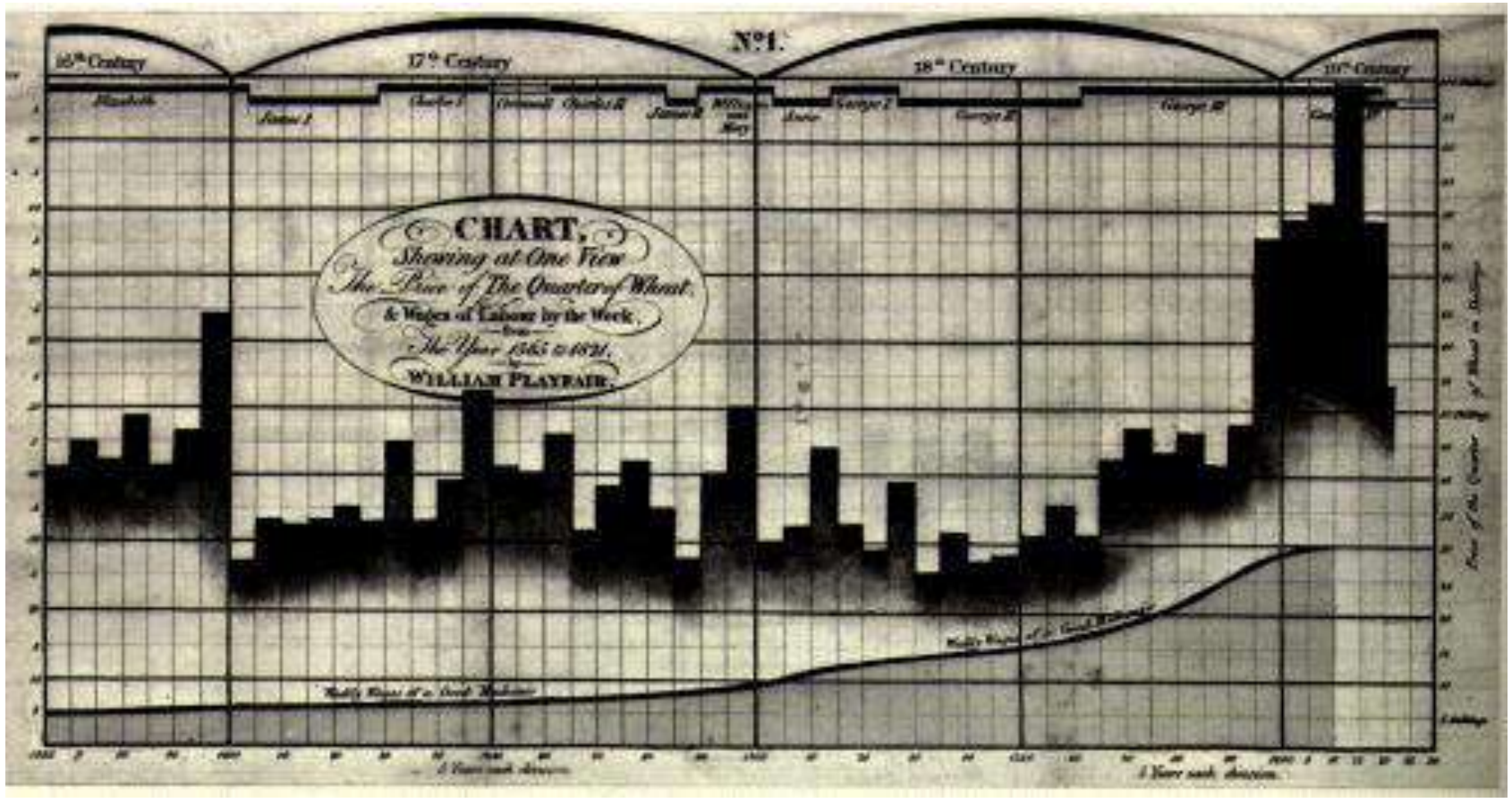
Visual DM



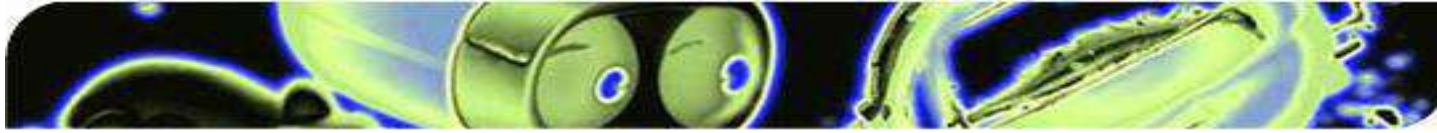
… introducing the industrial revolution …

- ***William Playfair*** created innovative graphics for industrial / economic production: time series and bar charts representing **wheat prices, salaries, and monarchies** along 250+ years

  (**Source**: Playfair, *Letters on our agricultural distresses …*)



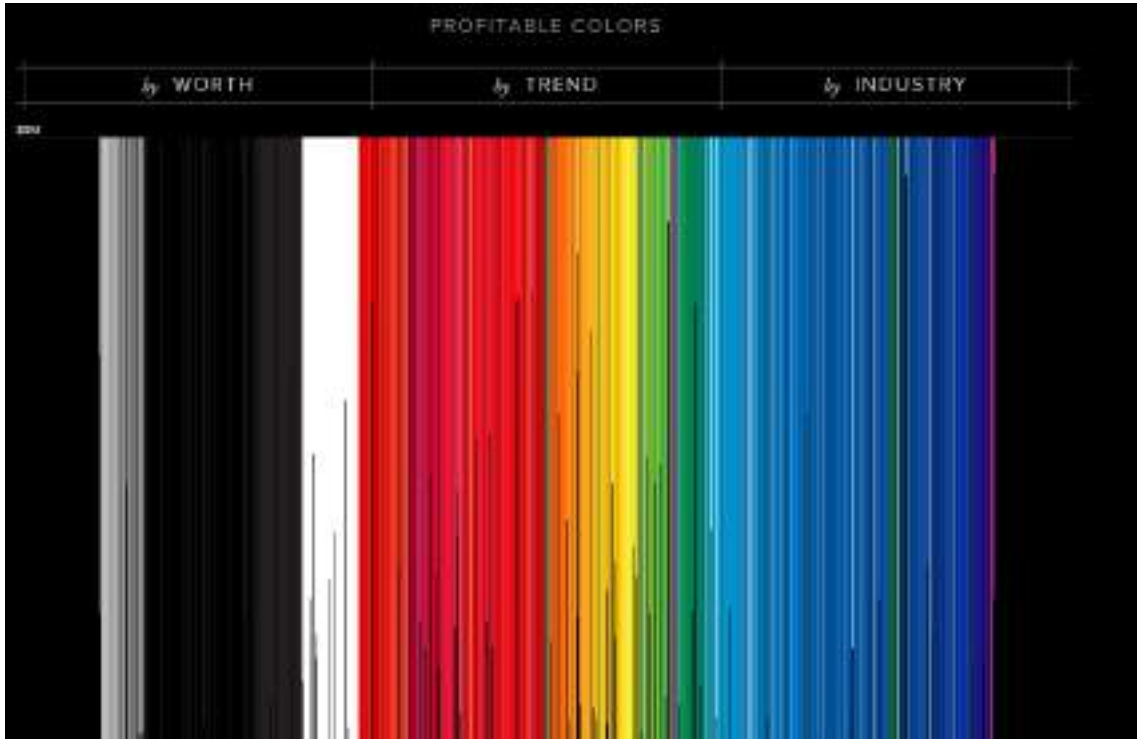
Visual DM



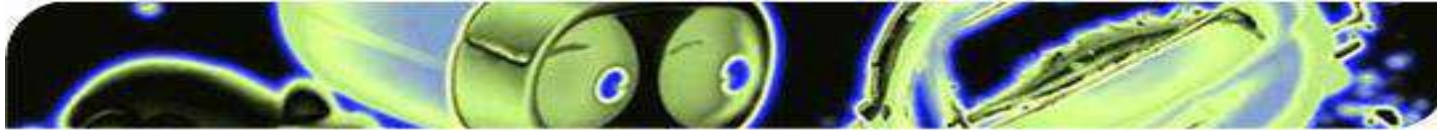
… to computer-based visualization …

- … Although visualization is more than a computer-based task: It is a process of **transforming information into a visual form** enabling the viewer to observe, browse, make sense, and understand the information.
- **These days, it typically employs computers** to process the information and **computer screens** to view it using methods of interactive graphics, imaging, and visual design.
- We must understand, though,

that standardized computer-based information visualization has been around for barely a couple of decades. For this reason, **visualization methods that make use of the possibilities of the computer** are still in their infancy.



Visual DM



… to computer-based visualization …

ieeevis.org/year/2020/welcome

viajes uni freerange eBIB

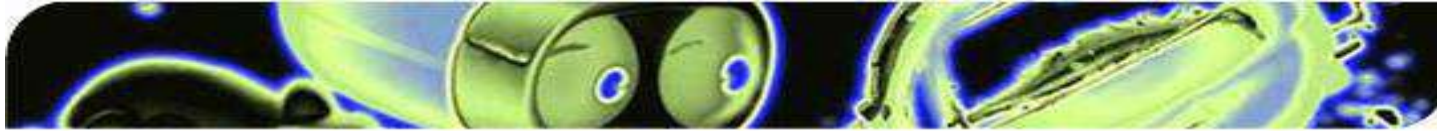
VIS 2020

the premier forum for advances in visualization

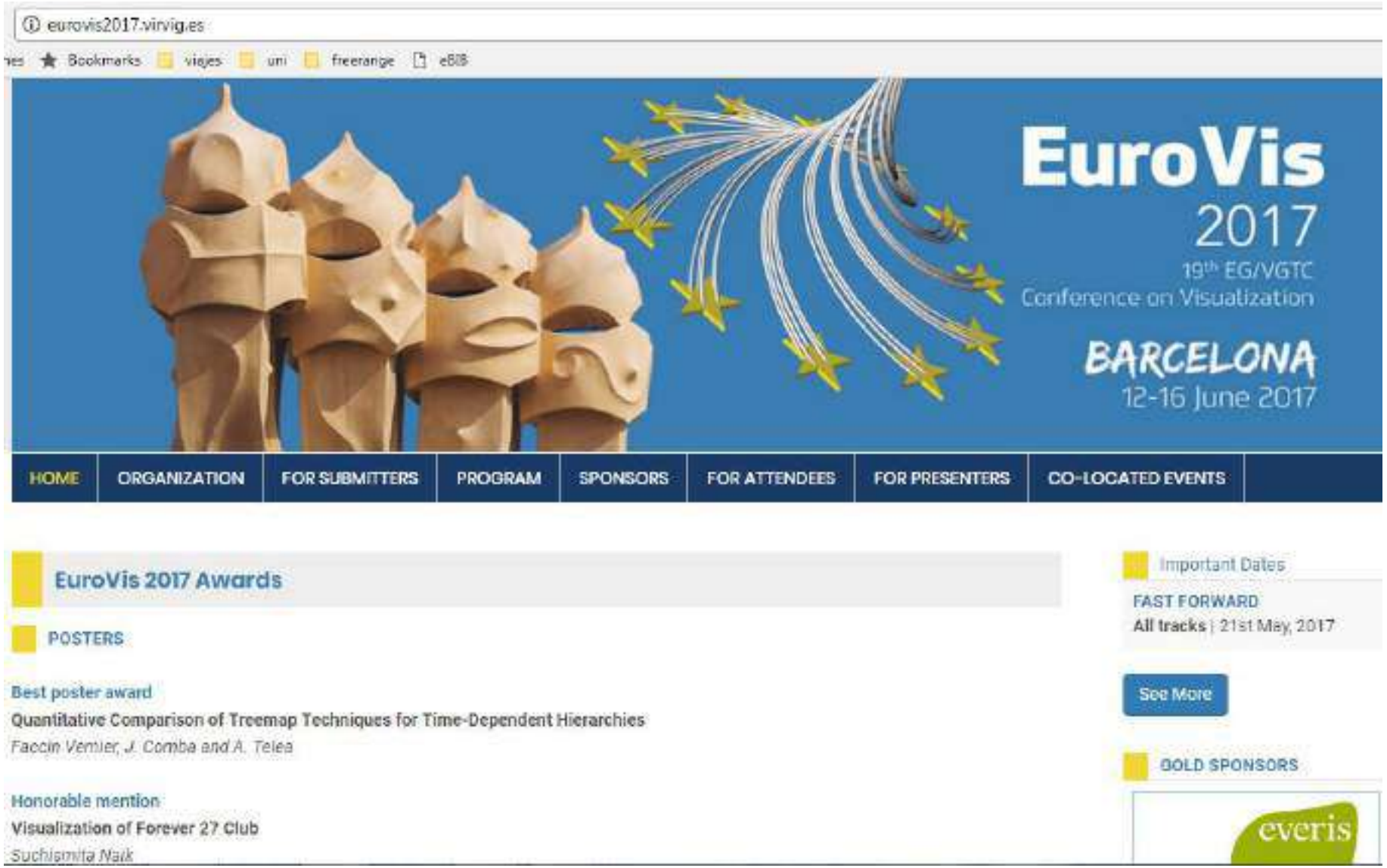
VIS | VAST • INFOVIS • SCIVIS

25-30 October 2020 in Salt Lake City, Utah, USA

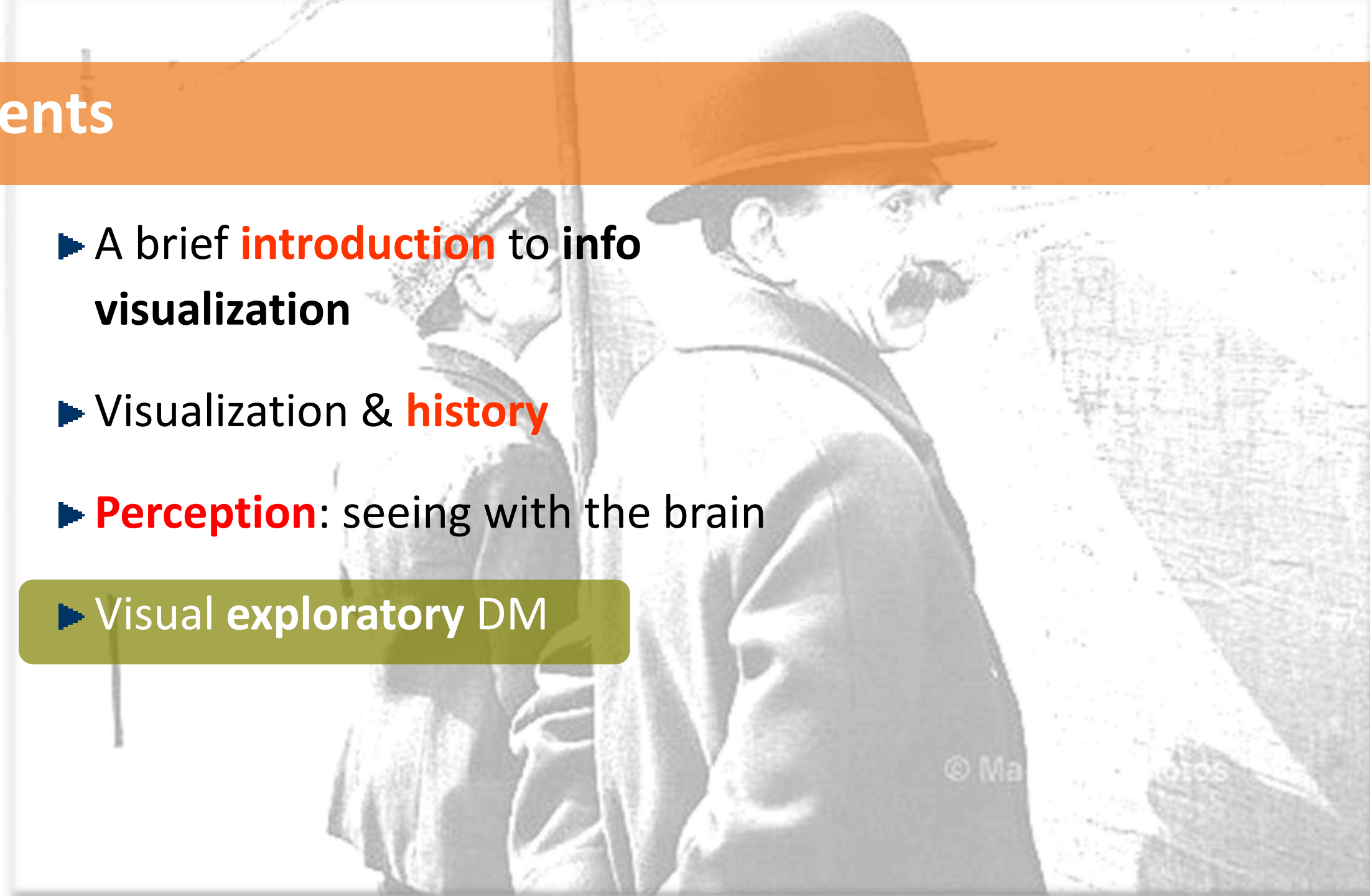
Visual DM



… to computer-based visualization …

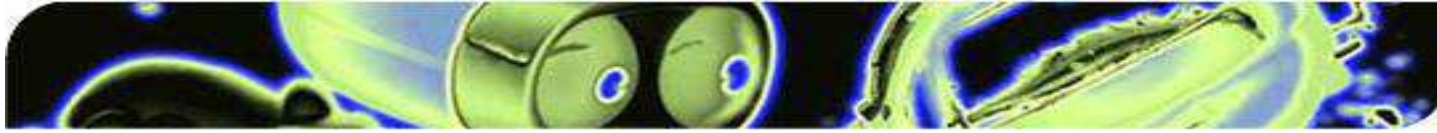


Contents



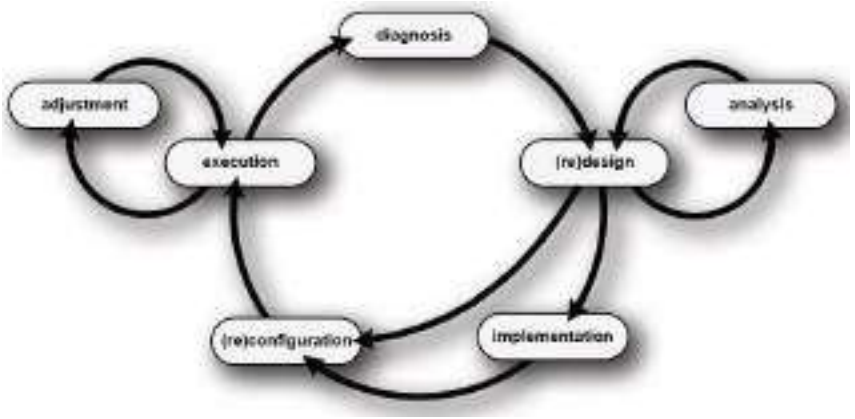
- A brief **introduction** to **info visualization**
- Visualization & **history**
- **Perception**: seeing with the brain
- Visual **exploratory** DM

Visual DM

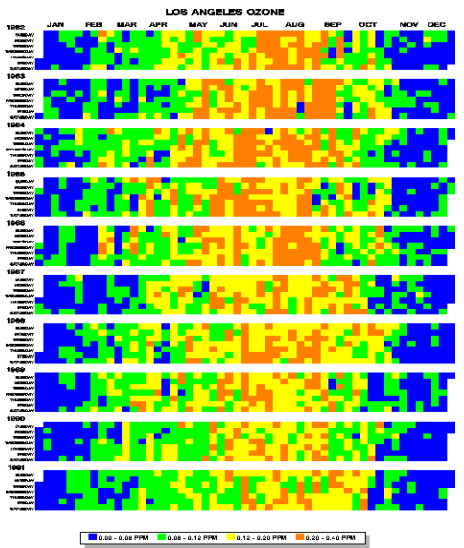


What **type** of visualization are we looking for?

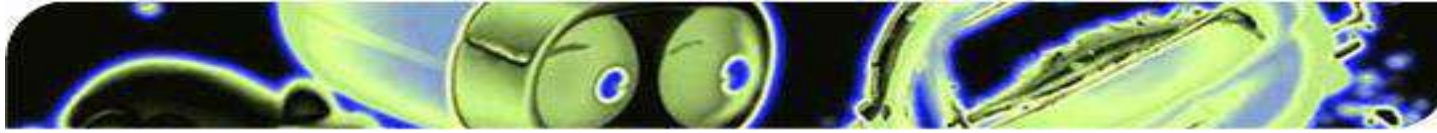
- **Descriptive?** …explicit



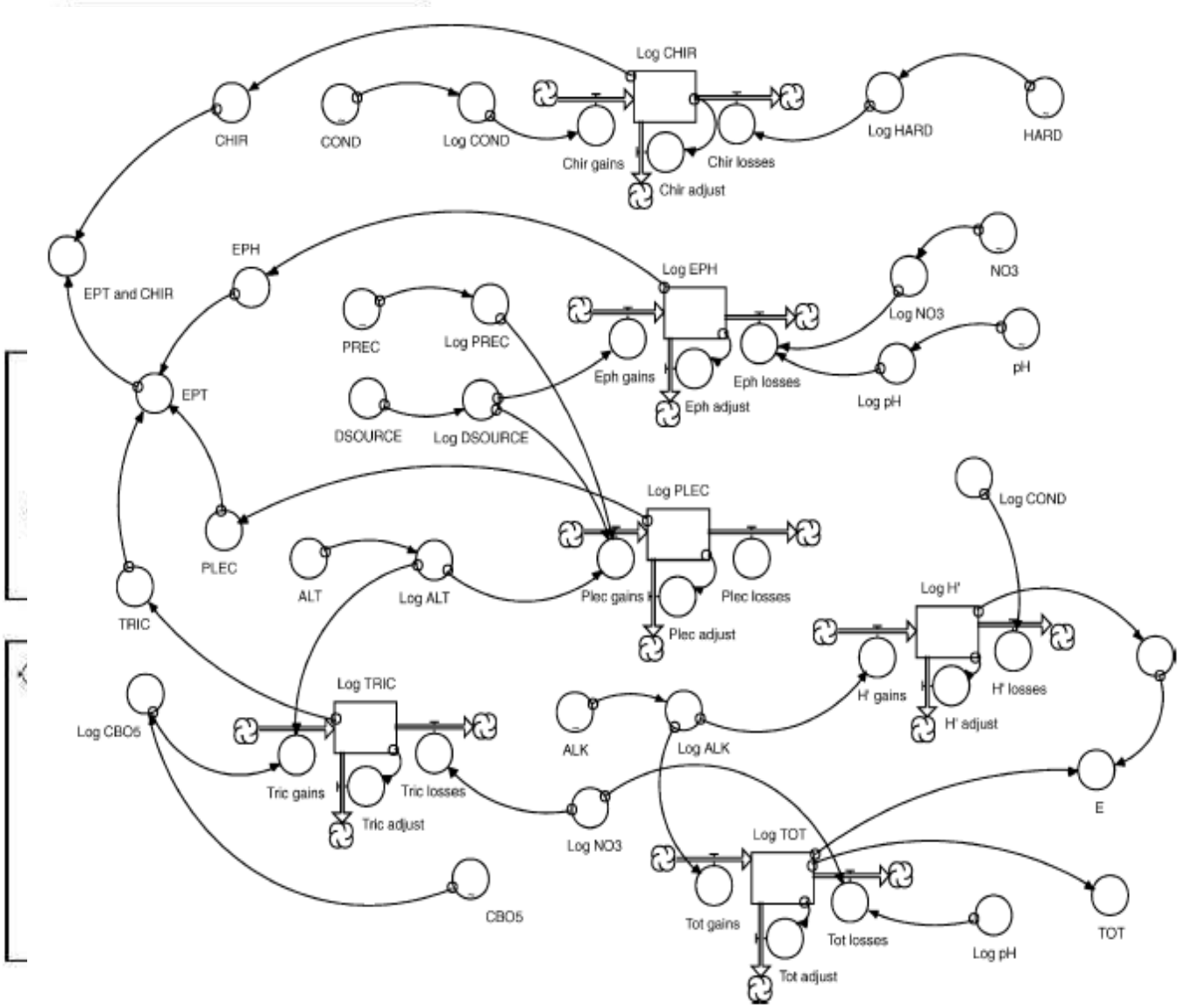
- **Exploratory?** …implicit



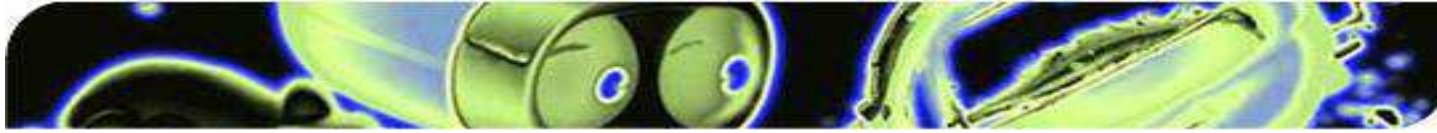
Visual DM

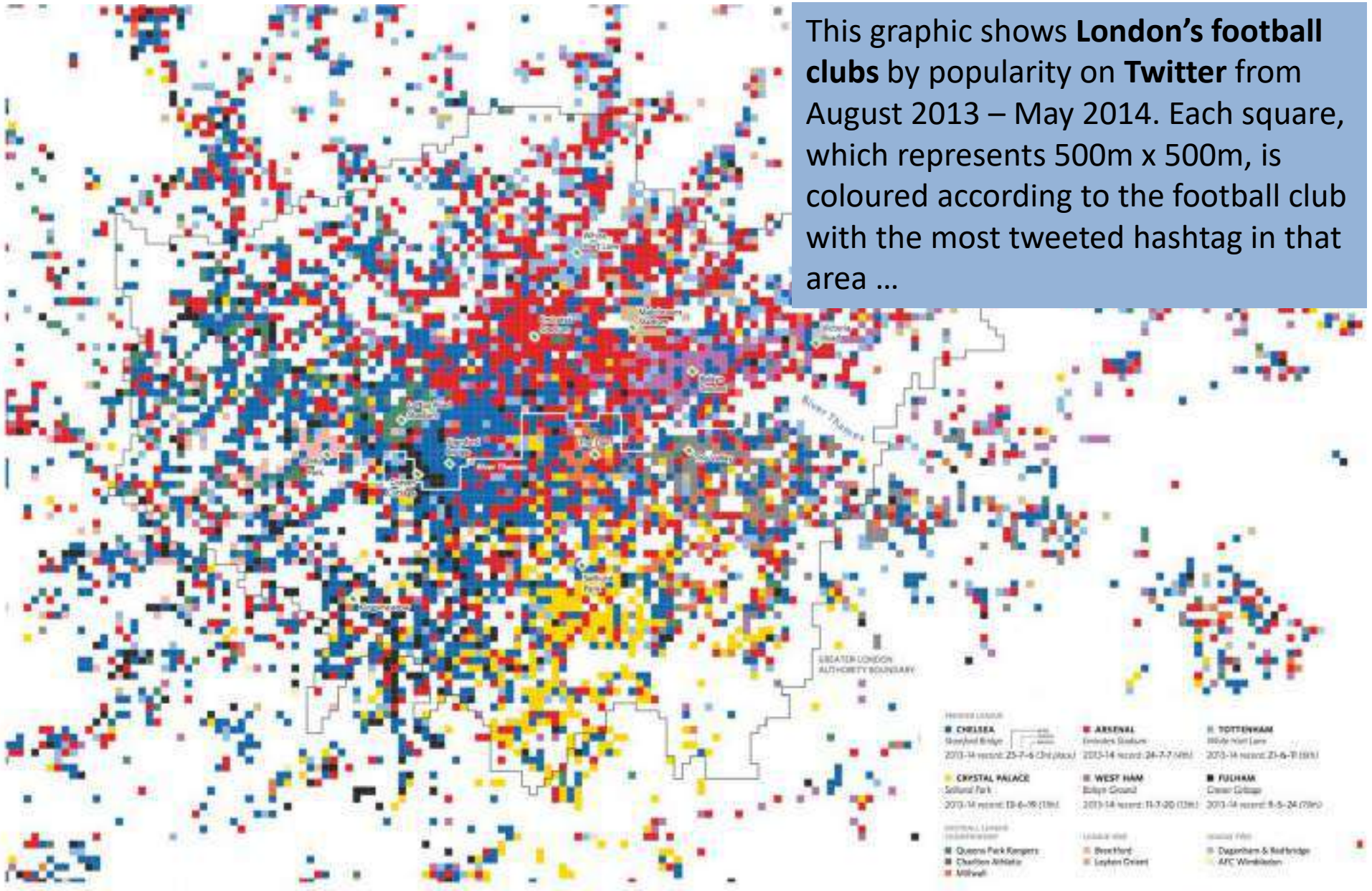


Type DESCRIPTIVE

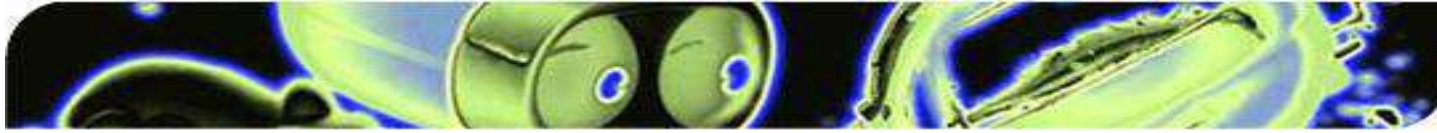


Visual DM





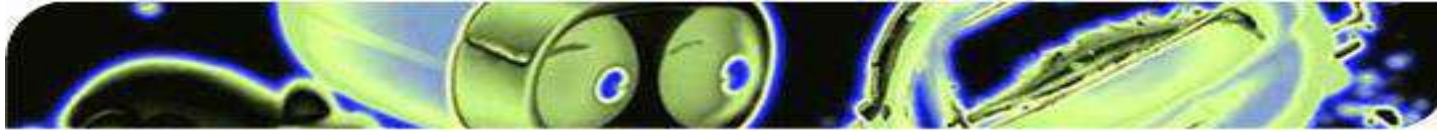
This graphic shows **London's football clubs** by popularity on **Twitter** from August 2013 – May 2014. Each square, which represents 500m x 500m, is coloured according to the football club with the most tweeted hashtag in that area …



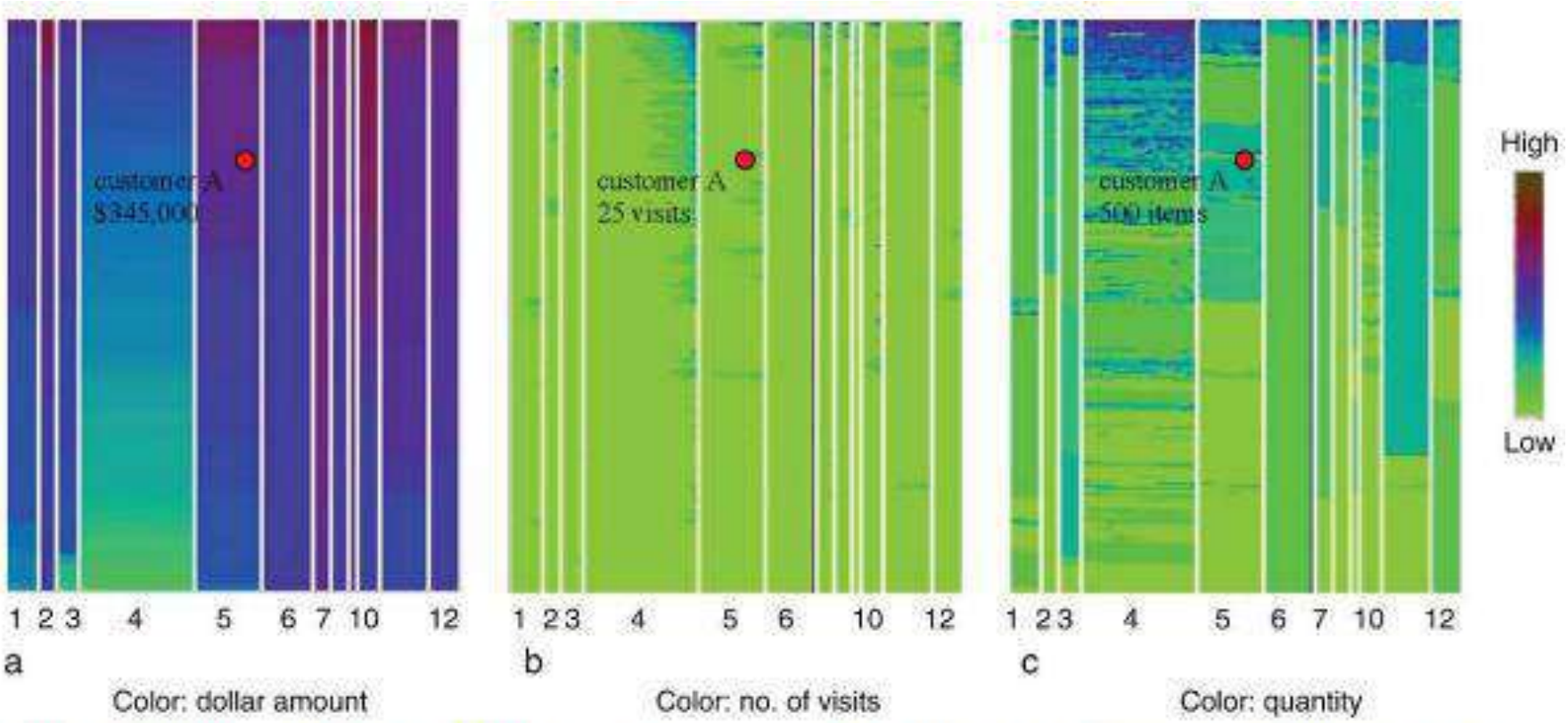
Principles of exploratory visualization:

A good exploratory visualization should ...

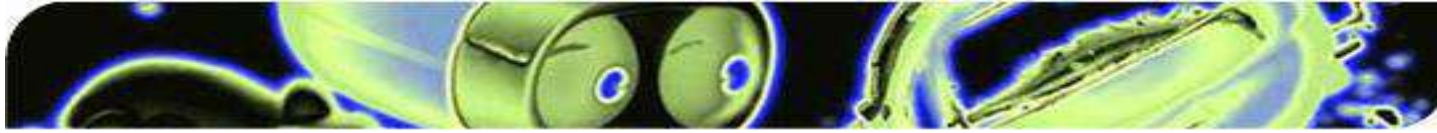
- **Show data and/or results** ...
  - ... at **different levels of detail**, from the overall landscape to the fine detail.
  - ... in a **coherent** manner, even if we are dealing with large collections.
  - ... **avoiding**, as much as possible, **distortion** in their representation.
- **Focus attention in the most relevantes features** ...
  - ... minimizing the impact of **uninformative** and **misleading** data.
  - ... **integrating** statistical results and linguistic descriptions (if possible and relevant**: multimodality).**



Data Exploration:
Some *dimensions …*



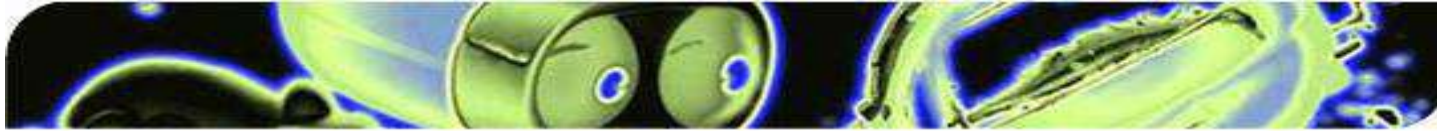
Dense Pixel Displays. Figure 4. Illustrates an example of a multi-pixel bar chart of 405,000 multi-attribute web sales transactions. The dividing attribute is product type; the ordering attributes are number of visits and dollar amount. The colors in the different bar charts represent the attributes dollar amount, number of visits, and quantity (adopted from[7]).



Data Exploration:
The CURSE of dimensionality …

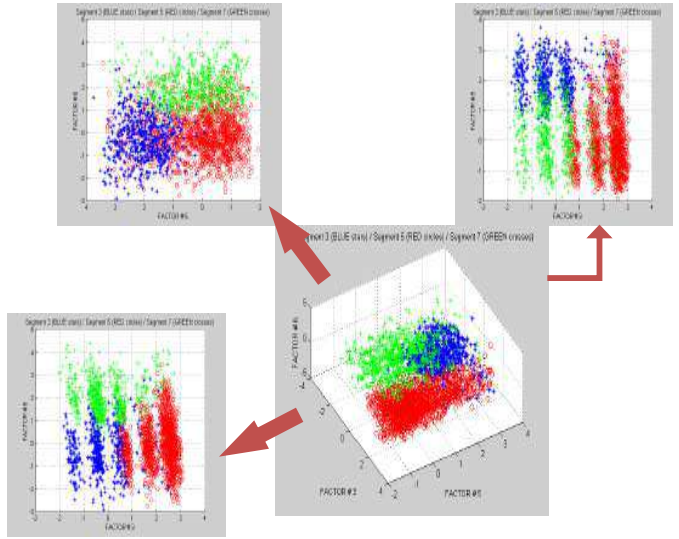
- **Most data available** to us are **stored in** different kinds of **databases and in numeric format, mostly** organized **in table structures.** An extension of these are the **data cubes** generated by ***OLAP*** **processes.**

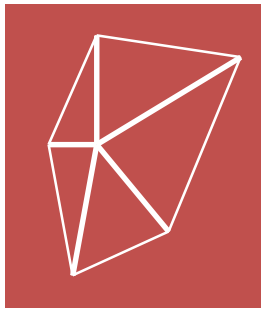
- **How to display multiple dimensions in a visually intuitive manner?**
  A simplified taxonomy of cases:
  - **Low** dimensionality (1-3D)
  - **Moderate** dimensionality (4-10D)
  - **High** dimensionality (>10D)



Data Exploration : low-moderate dim < 10D

- **Spatial coordinates**
  - 3D requires interactivity
- **Further pre-cognitive visual elements allow us to “add” extra dimensions:**
  - color, movement, shape, …
- **Exotic solutions**
  - Glyph*: Chernoff faces, stick-figures, *whiskers...*

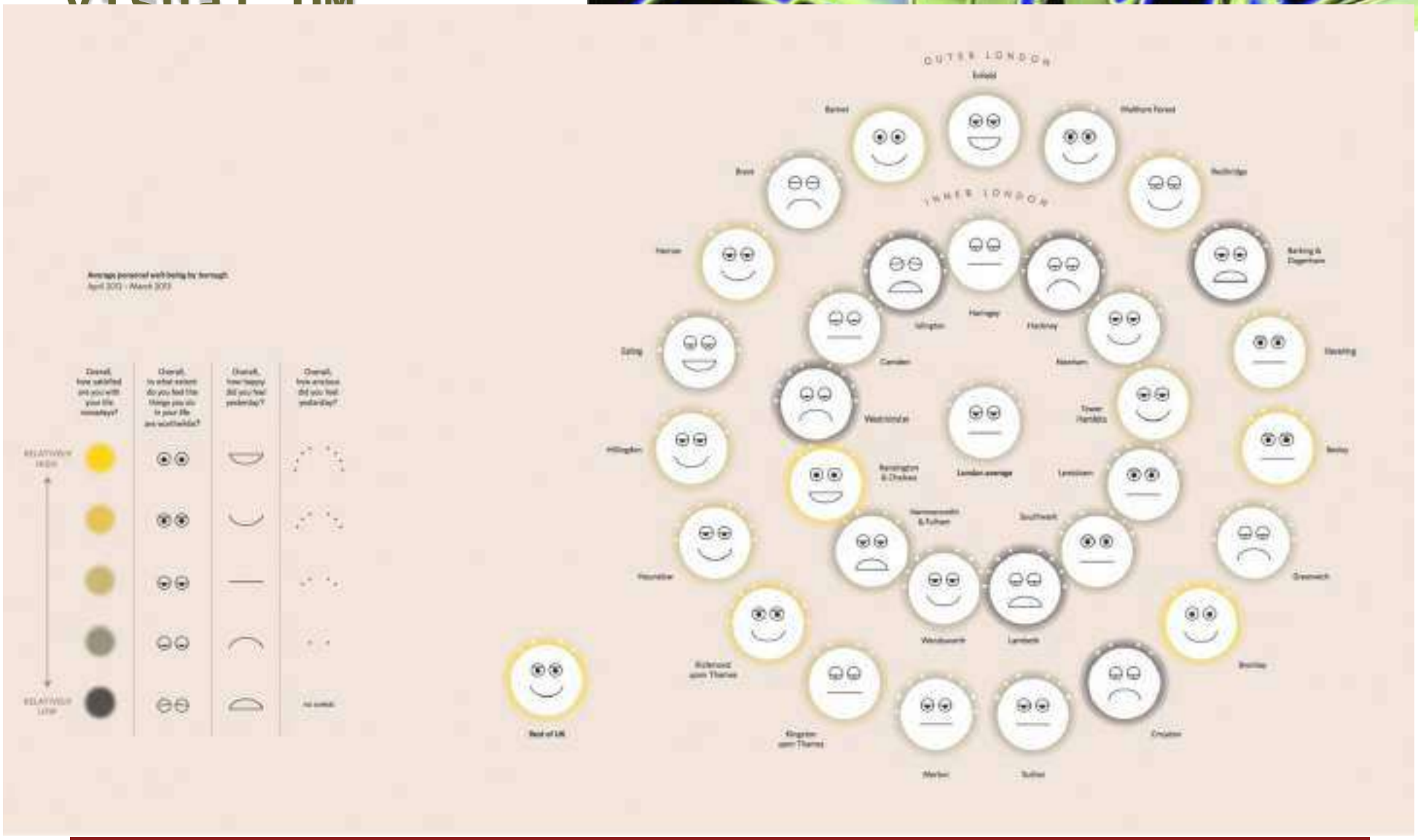




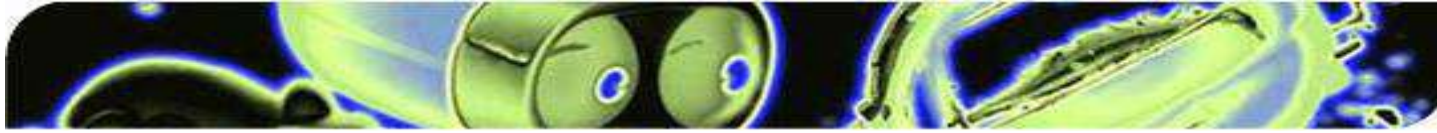
* A **glyph** is a graphical representation of one or more characters, or of part of a character. A character is a textual entity whereas a glyph is a graphical entity.

ideogram, pictogram …

Visual DM



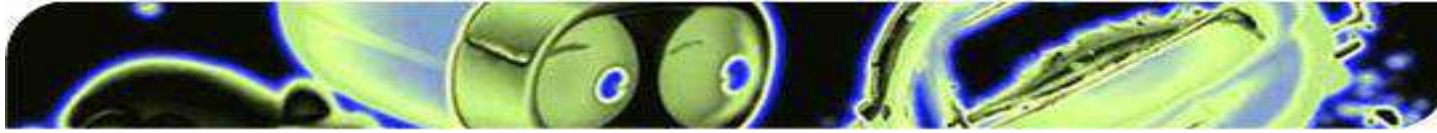
Herman Chernoff (1973). "Using faces to represent points in k-dimensional space graphically". Journal of the American Statistical Association 68 (342): 361–368.



Data Exploration : Data of high dimensionality

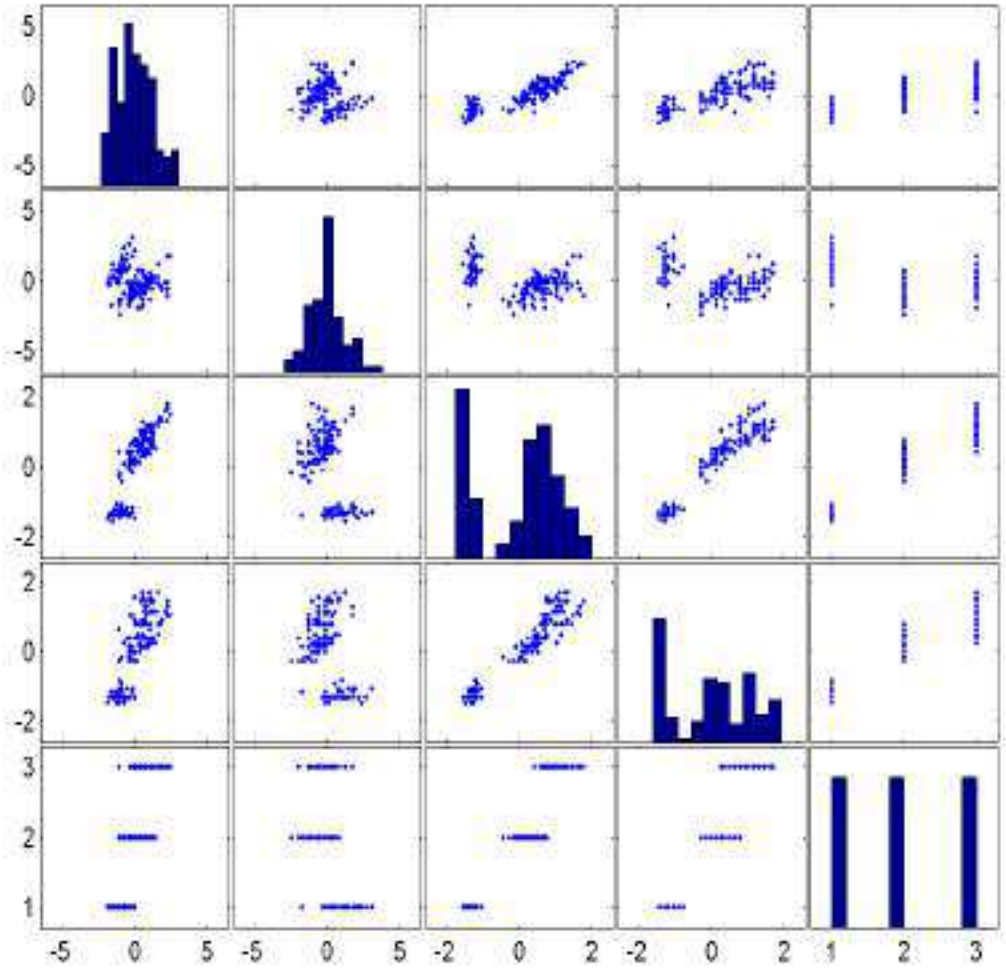
- **How do we visualize data of high (or even very high) dimensionality?**
  - Some of the alternatives are rather straightforward… some others are not…

- **Eliminate dimensions** (data variables): those which are redundant and / or uninformative (at least you manage to alleviate part of the problem…) → Feature selection
- **Divide & conquer**: a classic: create multiple visualizations of low dimensionality.
- **Latent and projection models**

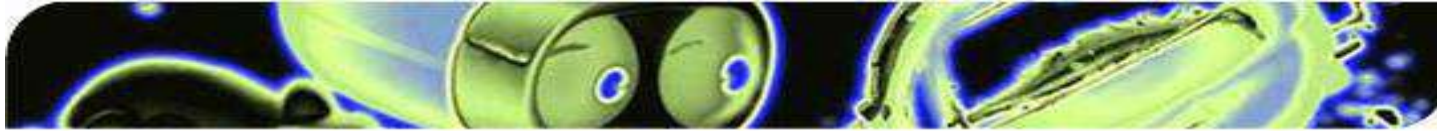


Data Exploration :

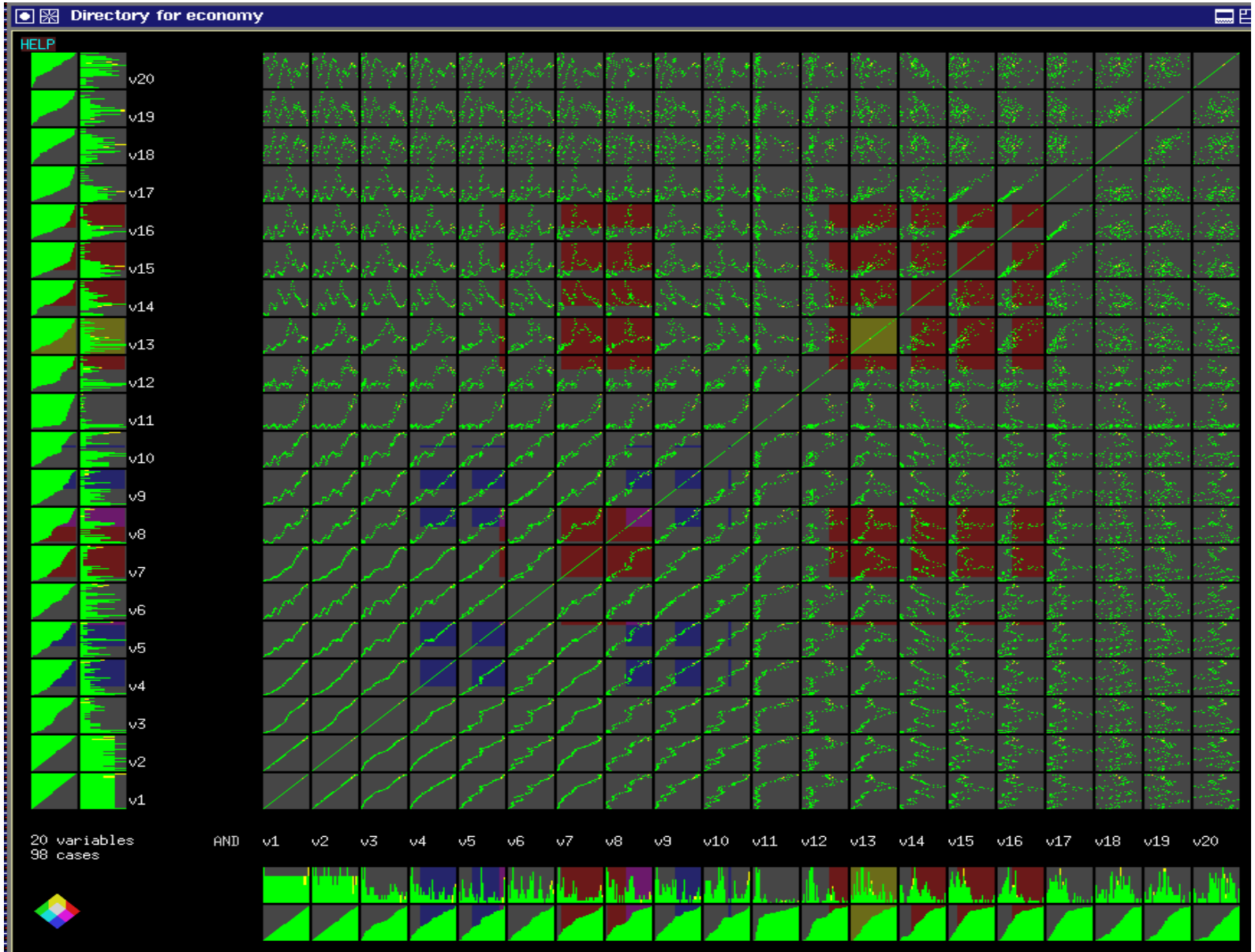
The *Grand Tour*: multiple visualization of *Iris data*

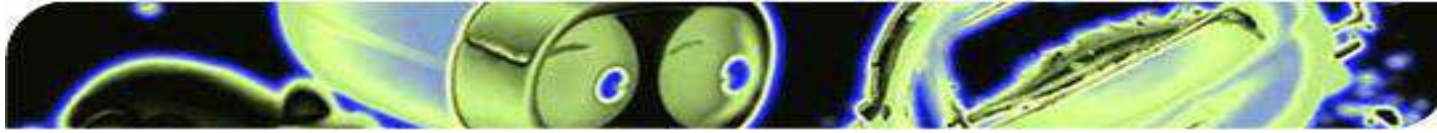


Visual DM



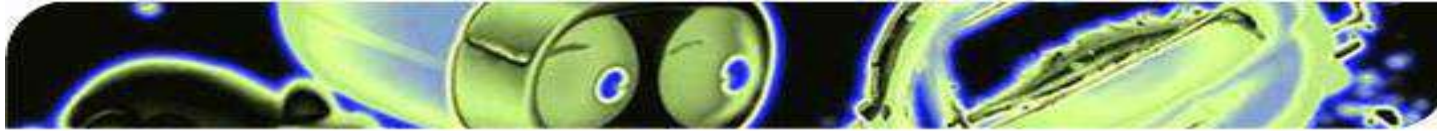
Data Exploration: Too *Grand a Tour*?





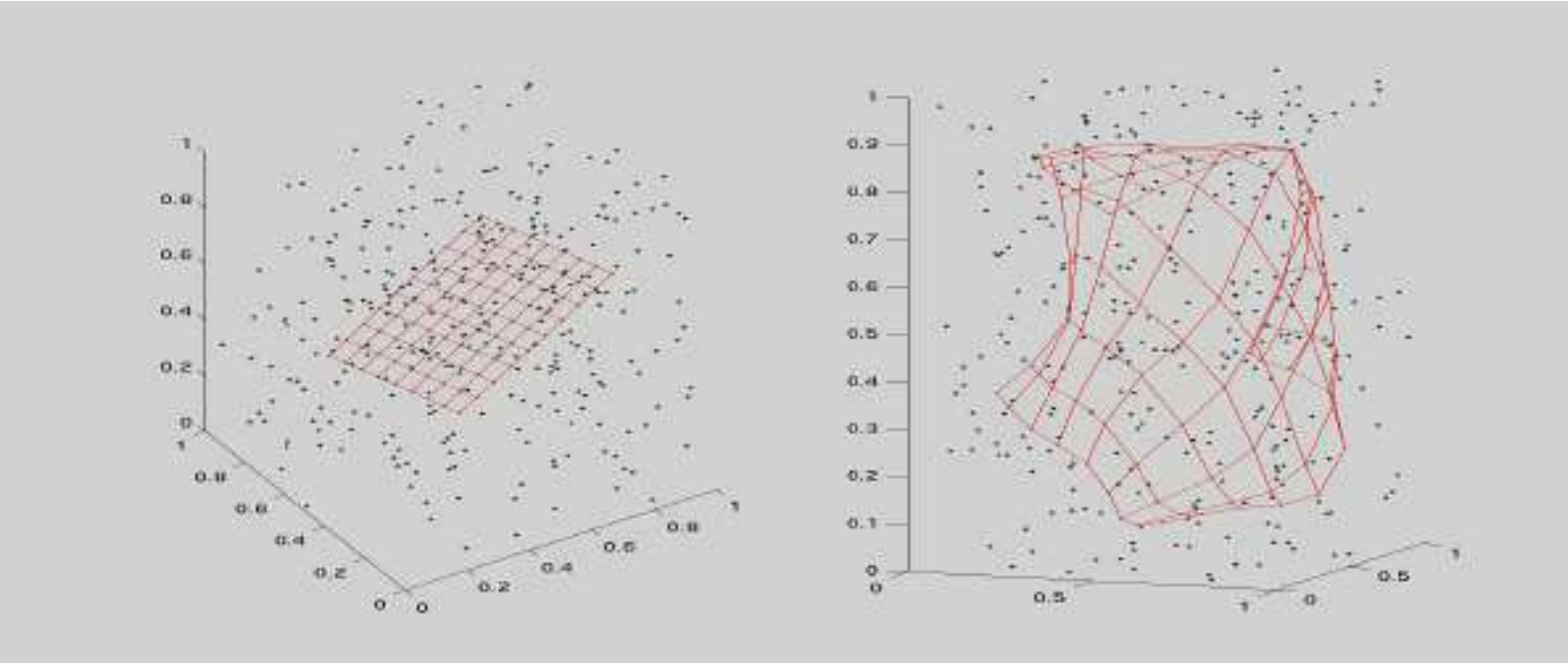
TECHNIQUES: Latency and projection: elements

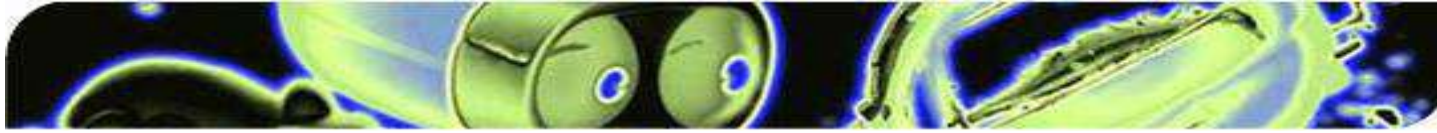
- **Projection**
  - Dimensionality compression (reduction)
  - **Similarity** information **coding**
- **Grouping / Clustering**
  - Finding grouping **structure** in data
  - **Similarity** information **coding**
- ***Vector Quantization & Manifold Learning***
  - Examples of combined **latent representation** and clustering



TECHNIQUES: projection

- **Representation in <4-D, so that the distance-neighborhood relations between multi-dimensional points are faithfully preserved**
  - **It is impossible** to preserve information integrally
  - Some scale normalization is often required
- ***Linear vs. non-linear*** **projections**





TECHNIQUES: projection: methods

- **Examples of methods based on inter-point distances, where:**

  dx = distance in the original space

  dy = distance in the projection space

  h = neighborhood function

$E = \Sigma\ (dx - dy)^2$	**MDS, PCA**
$E = \Sigma\ (dx - dy)^2 \,/\, dx$	**Sammon's projection**
$E = \Sigma\ (dx - dy)^2\ e^{-dy}$	**CCA**
$E = \Sigma\ dx^2\ h(dy)$	**SOM**

... and in which we aim to minimize an inherent projection distorsion (E)

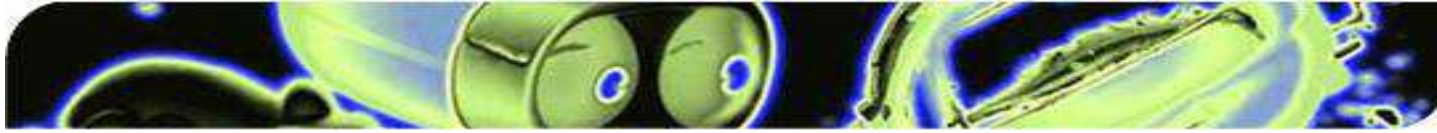
Visual DM

Some dimensionality reduction algorithms

They can be divided into 3 major groups:

➡	PCA/LDA	linear	Matrix Factorization		
	ICA	linear	Matrix Factorization		
	MDS	non-linear	Matrix Factorization		
	Sparce NNMF	non-linear	Matrix Factorization	2010	https://pdfs.semanticscholar.org/664d/40258f12ad28ed0b7d4c272935ad72a150db.pdf
	cPCA	non-linear	Matrix Factorization	2018	https://doi.org/10.1038/s41467-018-04608-8
	ZIFA	non-linear	Matrix Factorization	2015	https://doi.org/10.1186/s13059-015-0805-z
	ZINB-WaVE	non-linear	Matrix Factorization	2018	https://doi.org/10.1038/s41467-017-02554-5
	Diffusion maps	non-linear	graph-based	2005	https://doi.org/10.1073/pnas.0500334102
	Isomap	non-linear	graph-based	2000	10.1126/science.290.5500.2319
➡	t-SNE	non-linear	graph-based	2008	https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf
	- BH t-SNE	non-linear	graph-based	2014	https://lvdmaaten.github.io/publications/papers/JMLR_2014.pdf
	- Flt-SNE	non-linear	graph-based	2017	arXiv:1712.09005
	LargeVis	non-linear	graph-based	2018	arXiv:1602.00370
➡	UMAP	non-linear	graph-based	2018	arXiv:1802.03426
	PHATE	non-linear	graph-based	2017	https://www.biorxiv.org/content/biorxiv/early/2018/06/28/120378.full.pdf
	scvis	non-linear	Autoencoder (MF)	2018	https://doi.org/10.1038/s41467-018-04368-5
	VASC	non-linear	Autoencoder (MF)	2018	https://doi.org/10.1016/j.gpb.2018.08.003

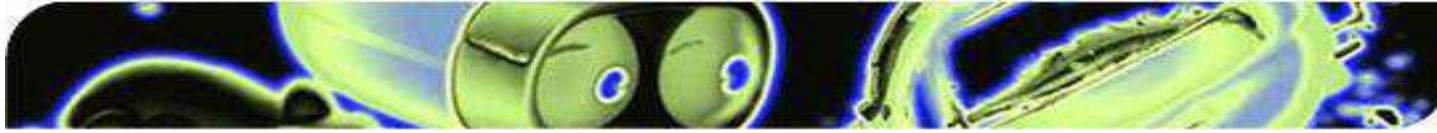
... and many more



TECHNIQUES:

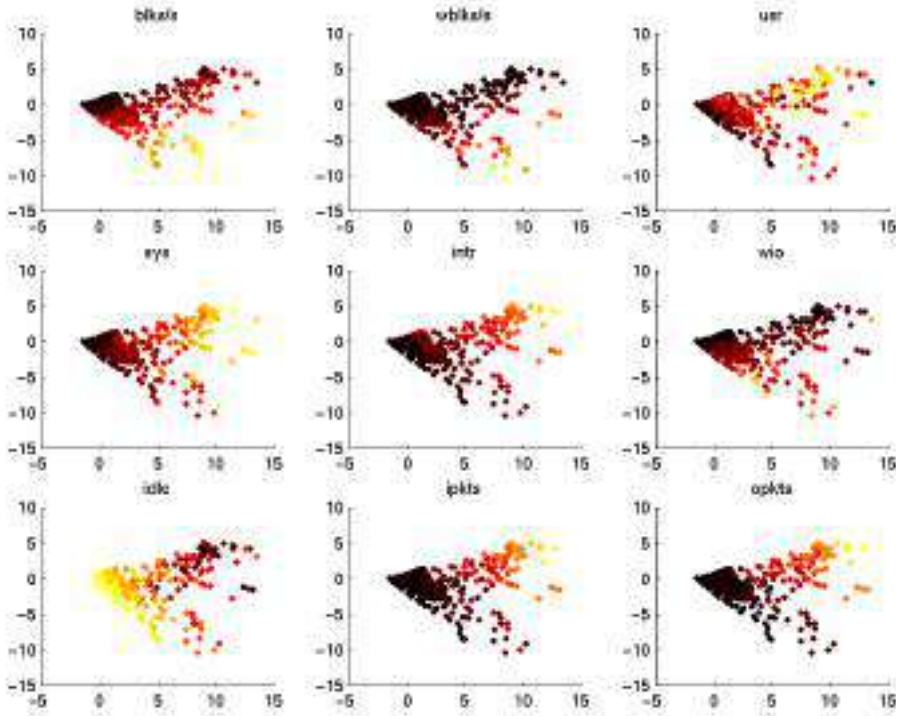
Projection: discussion, pros & cons

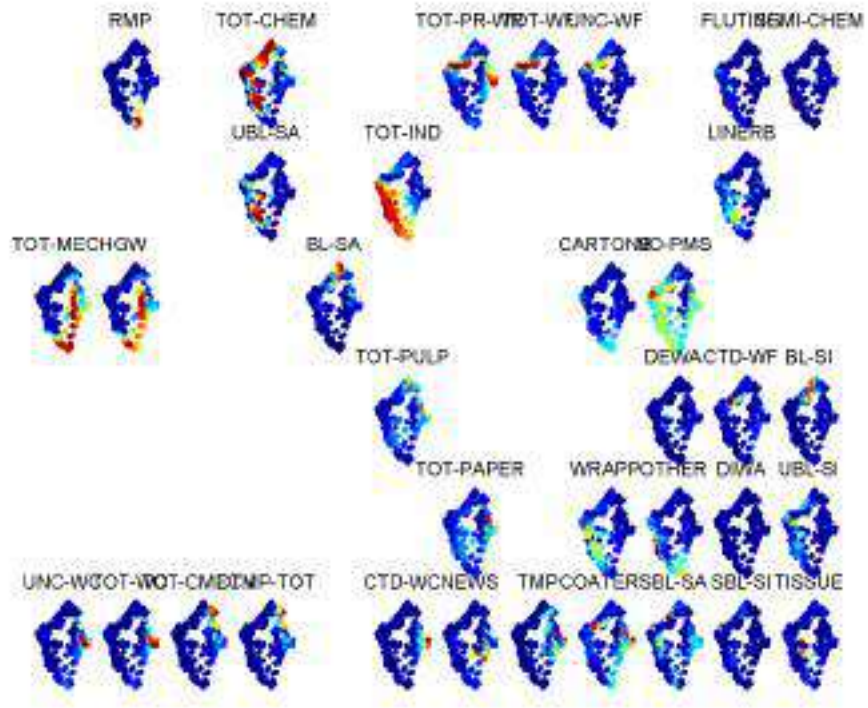
- **Projection techniques code proximity / similarity information in spacial coordinates (sometimes, with extra precognitive elements such as colour ...)**
  - **They allow…**
    - … **Finding "natural" data structure** (groups, *clusters*) on the basis of some sort of similarity
    - … Finding the "**shapes**" of these groupings
  - **But ...**
    - Projection is always limited by **error** and **information loss.**
    - New projection coordinates are **not always readily interpretable** (latency by definition), given that the original relations between data dimensions are lost **(interpretability!)**.
    - Quite often, the **computacional effort** is to be taken into account, as most of these methods are based on distances between multivariate points **(scalability!)**.



TECHNIQUES: multiple visualizations

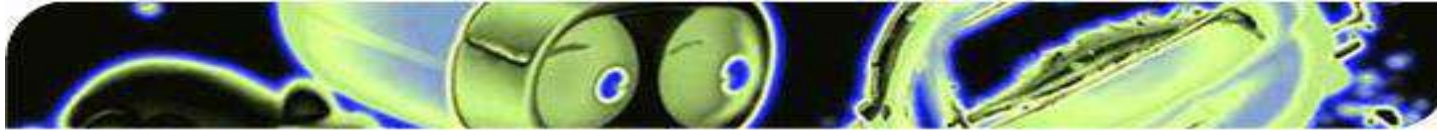
- **How to get some of the info conveyed by observable variables back into the projections?** One possibility: Using multiple visualizations.
  - Parallel coordinates and pre-cognitive stimuli (colour, position...)

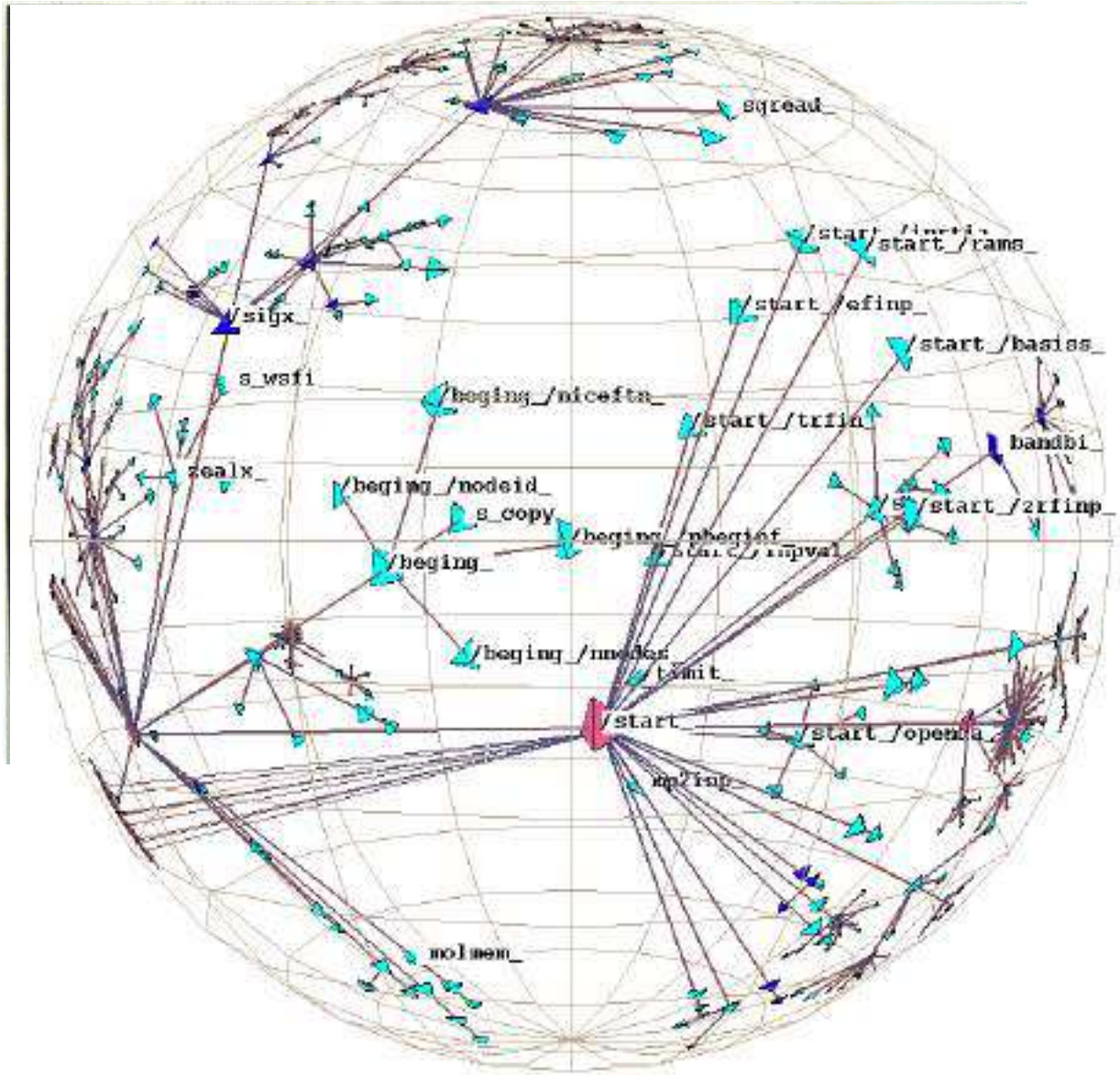




Beyond projection
Visualization gone wide:
text, hierarchies, graphs and other exotisms

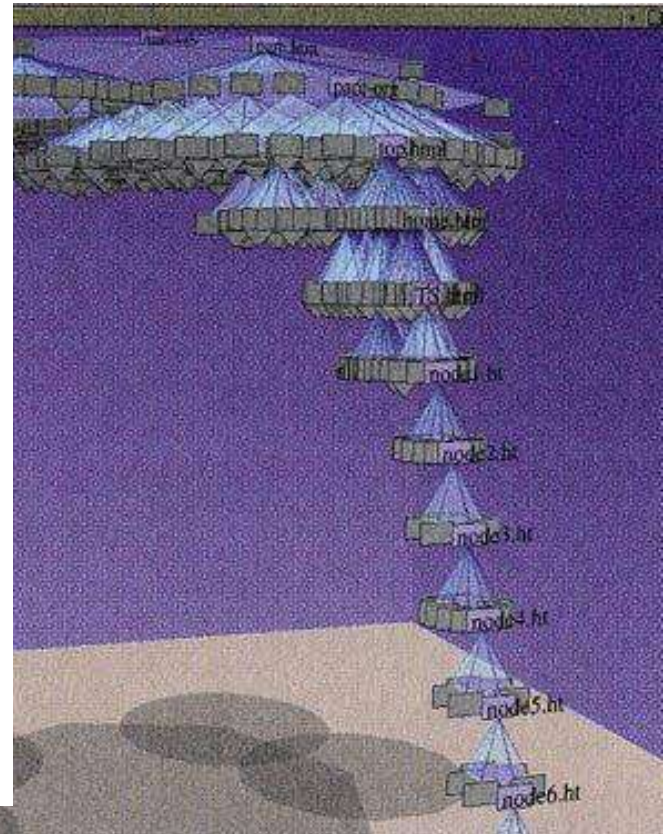
Visual DM



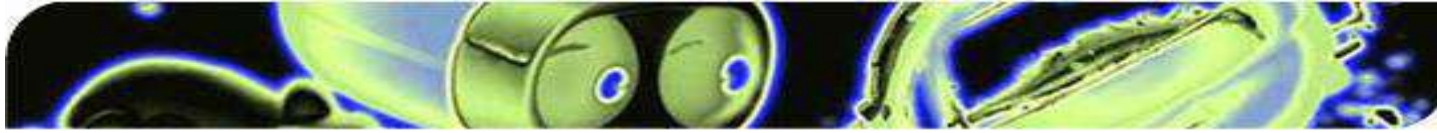


hierarchies:

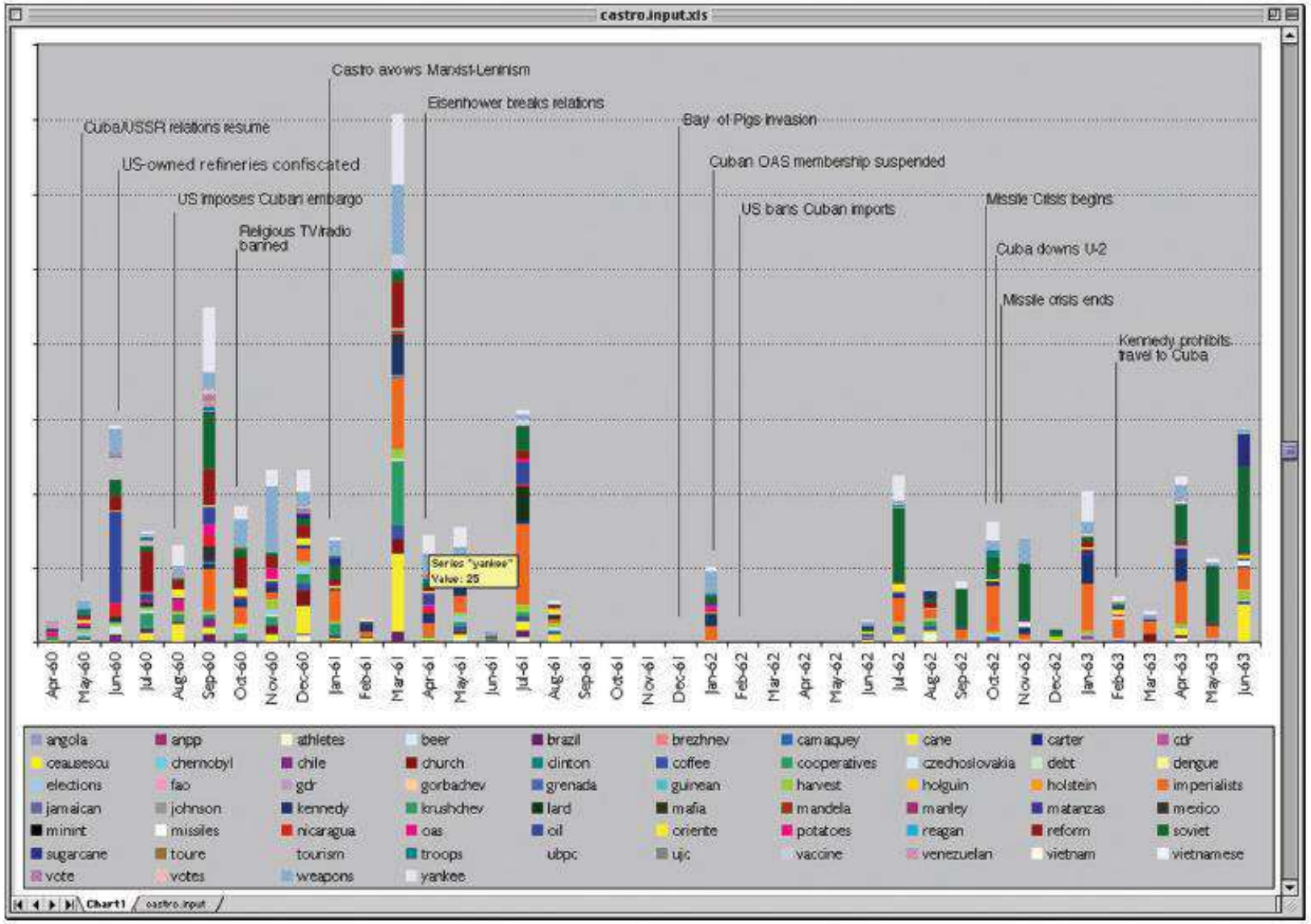
Conic trees



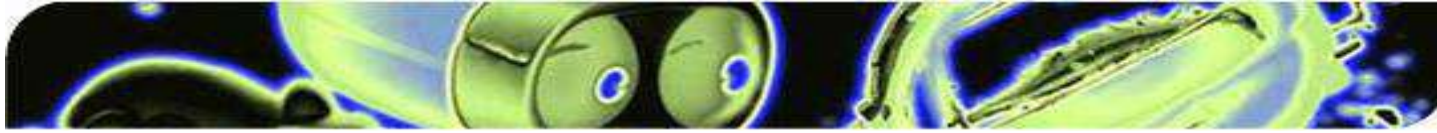
Visual DM



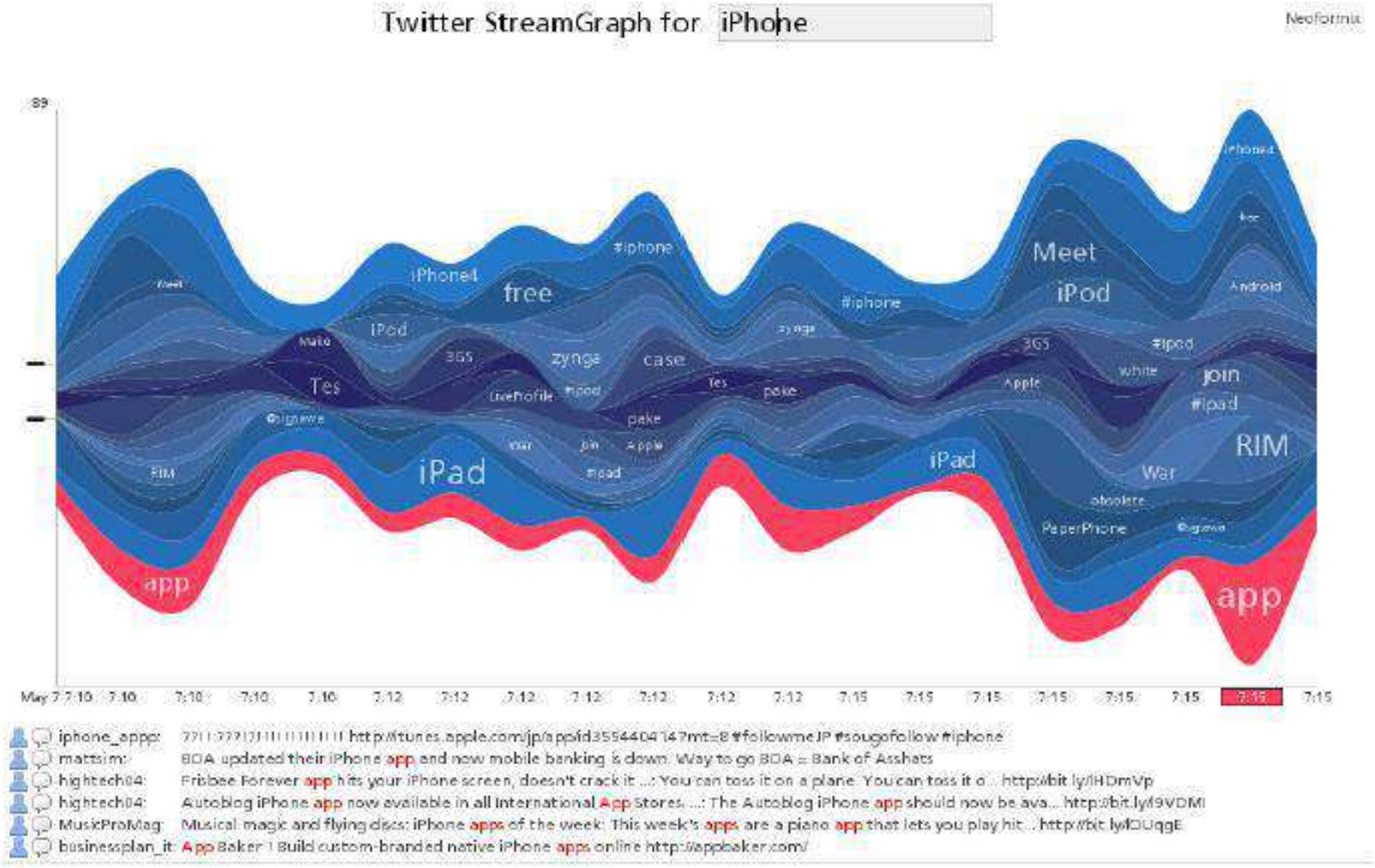
ThemeRivers



Visual DM



StreamGraph:



Visual DM

Mapscapes