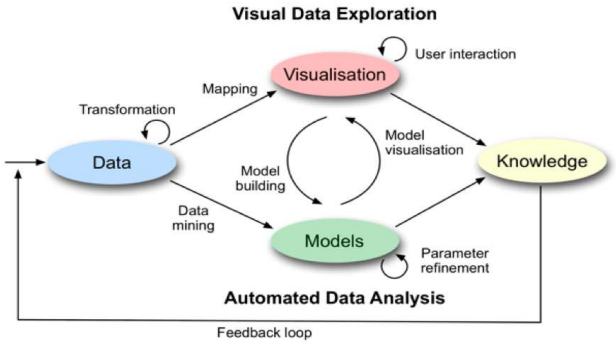
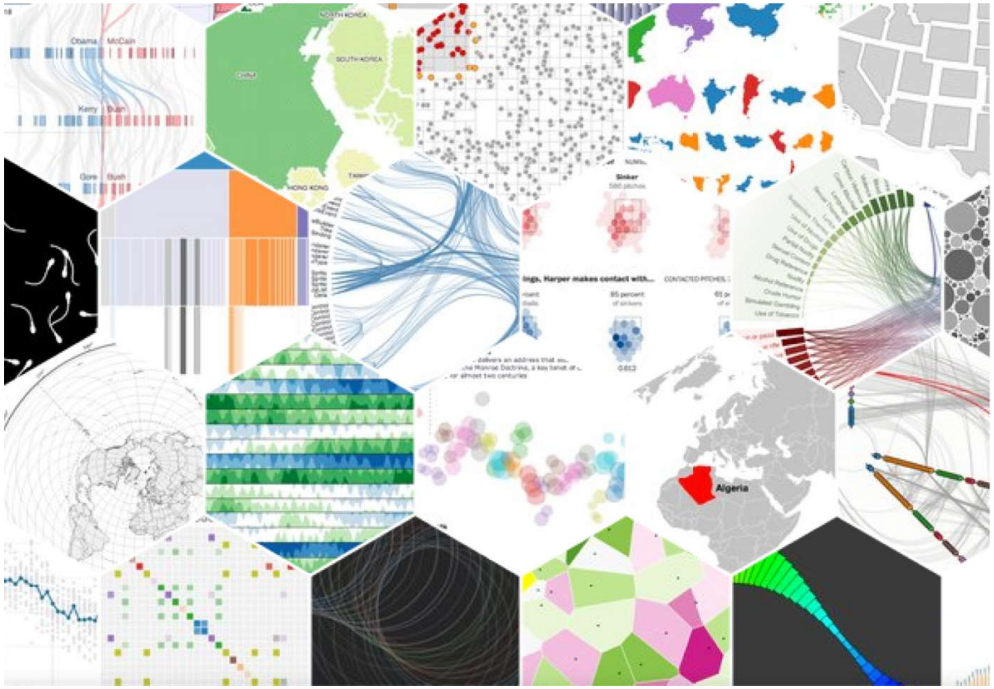


1052057: Visual Analytics Fall 2019

Giuseppe Santucci



Course Introduction

Thanks to Enrico Bertini, John Stasko, Robert Spence, Ross Ihaka, Marti Hearst

Outline

- Facts about the VA course
- Historical examples
- Definitions
  - The Power of Information Visualization
  - Visual Analytics
- The problem and the involved issues

Outline

- Facts about the course
- Historical examples
- Definitions
  - The Power of Information Visualization
  - Visual Analytics
- The problem and the involved issues

Course resources

- Giuseppe Santucci home page (news, office hours, etc.)
  - **https://piazza.com/uniroma1.it/fall2019/1052057/home**
  - Follow the link for this course
  - Enroll!

- The course has no textbooks; I have used content from:
  1. Robert Spence : Information Visualization – Design for interaction 2nd Ed. - Pearson Prentice Hall
  2. Colin Ware : Information Visualization, Second Edition: Perception for Design 2nd Ed. – Elsevier
  3. Stephen Few : Show me the numbers – Analytics Press
  4. Illuminating The Path - PNNL - Pacific Northwest National Laboratory (download from Piazza)
  5. Solving Problems with Visual Analytics (download from Piazza)

- Office Hours: Monday 14.30-16.30 Via Ariosto 25 room B218
  - Always have a look at news before coming !

Program: a) Number visualization

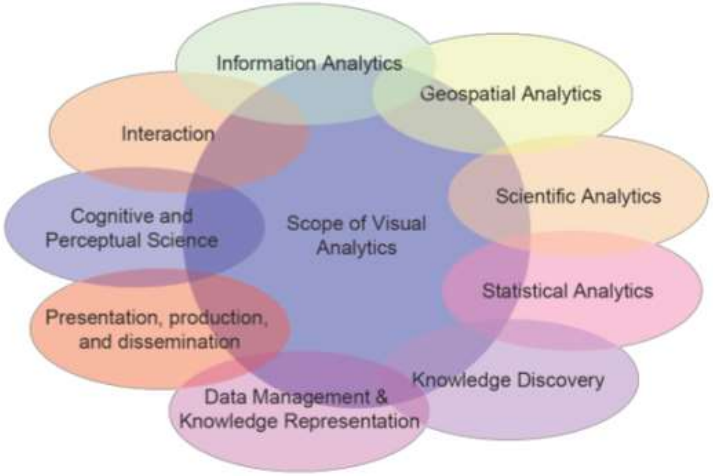
- Introduction
- Common errors & lies
- Kinds of number
- Table and graphs

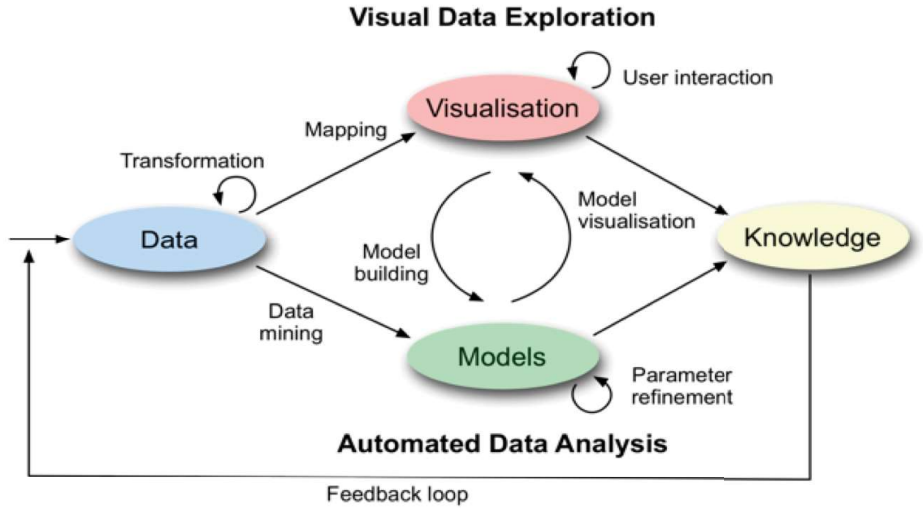
Program: b) Information visualization

- Introduction
- Representation
  - Encoding
  - Type of data – Univariate / Multivariate data
  - Data and relationships
  - Perceptual issues
- Presentation
  - Space limitation
  - Time limitation
- Interaction
  - Continuous interaction
  - Stepped interaction
- Case studies
- d3.js (By Marco Angelini and Simone Lenti)

Program: c) Visual Analytics

- Introduction
  - applications, tools, research challenges
- Representation and Transformation
  - data wrangling
  - dimensionality reduction
  - data reduction/summarization/modeling
- Presentation and Interaction
  - visualization techniques/scalability
  - space-time
  - visual data mining
  - interaction techniques





What we are ***not*** covering

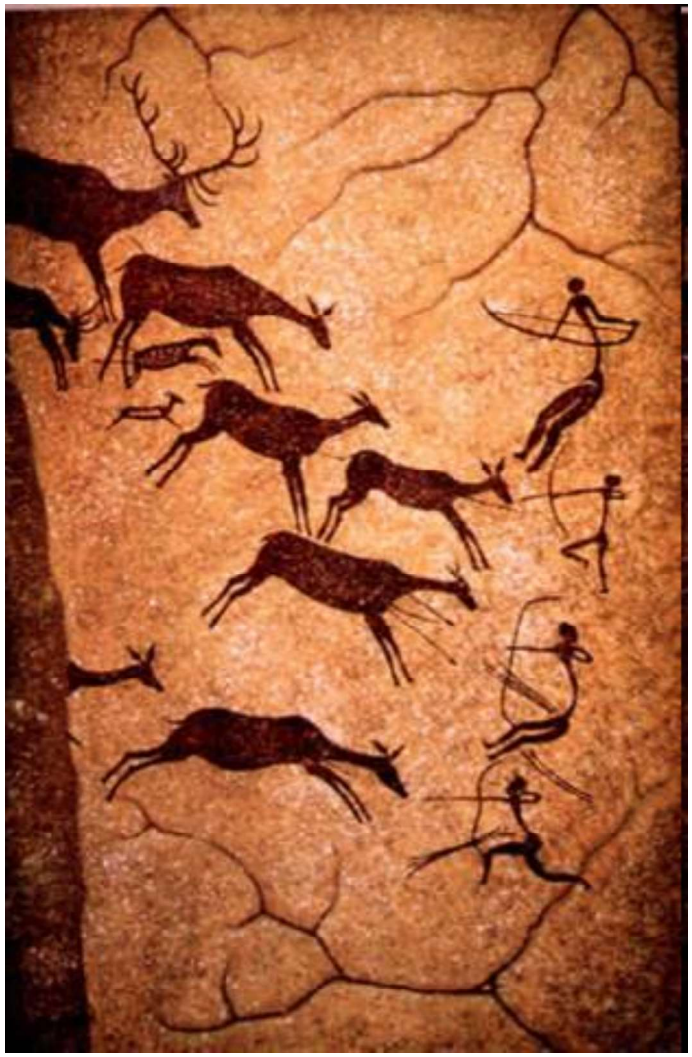
- Scientific visualization
- Cartography (maps)
- Education
- Games
- Computer graphics in general

Outline

- Facts about the course
- Historical examples
- Definitions
  - The Power of Information Visualization
  - Visual Analytics
- The problem and the involved issues

Visualization?

- Old stuff…



Visualization ?

1. Problem solving / Analyzing
2. Explaining
3. Making decisions

Problem Solving/Analyzing

Mystery: what was causing a cholera epidemic in London in 1854?

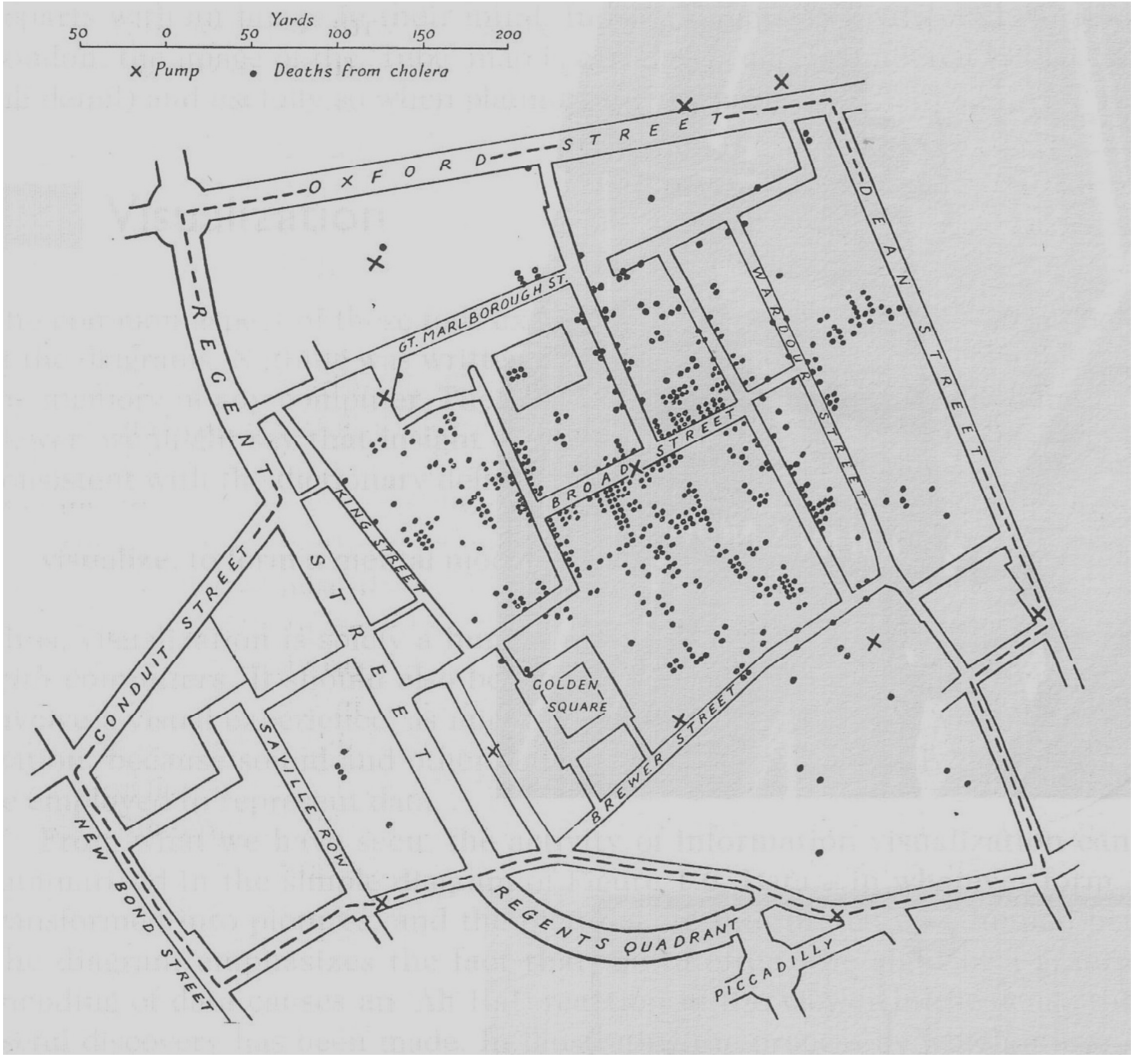
Visualization for Problem Solving

Illustration of Dr. John Snow (1854)

Dots indicate location of deaths

X indicate the location of water pumps

From Visual Explanations by Edward Tufte, Graphics Press, 1997



Visualization for Problem Solving

The actual John Snow pub in London close to the water pump !!!

John Snow deducted that the cholera epidemic was caused by a contaminated water pump !!!
Closing that pump quickly solved the problem

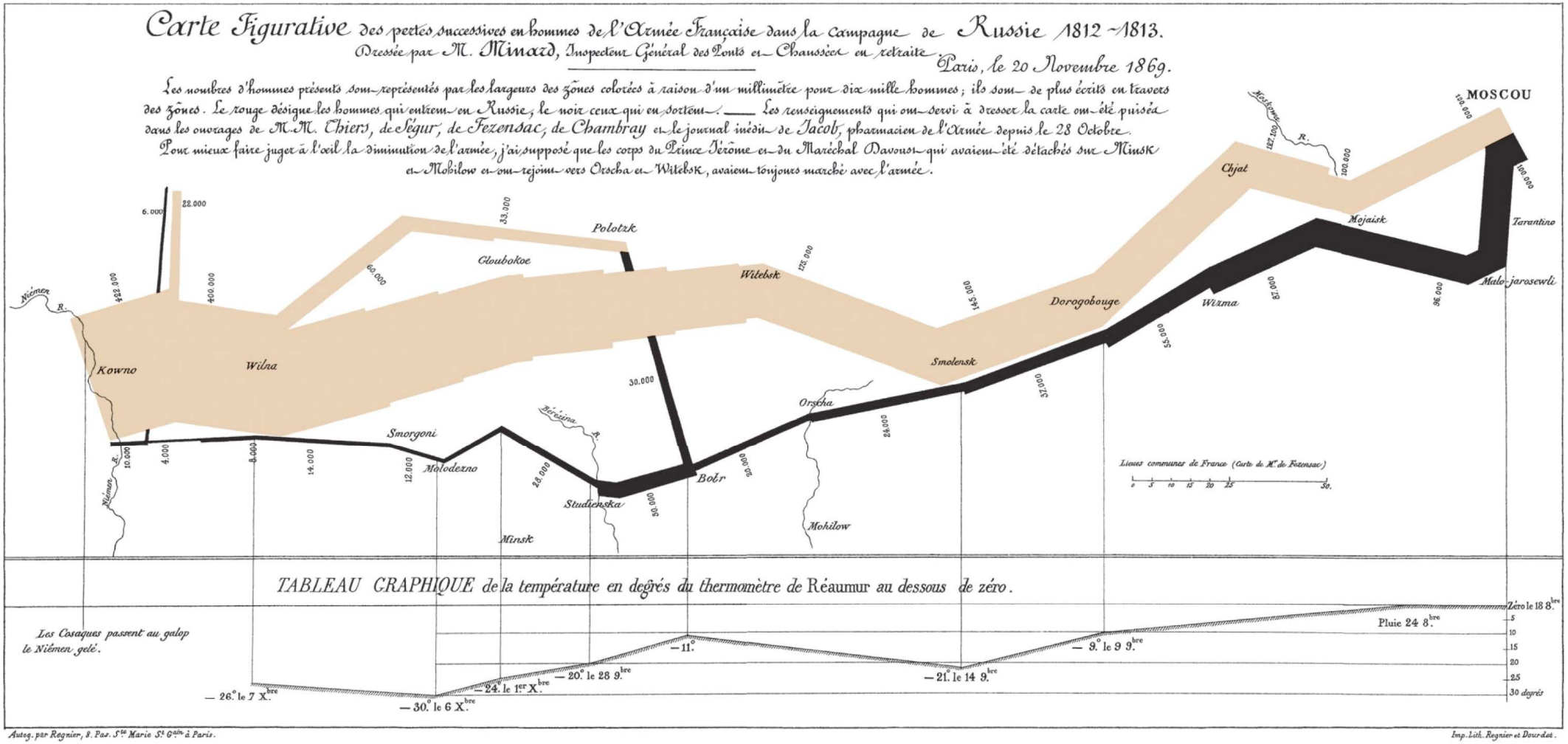
B.T.W., workers at the nearby brewery were noted to be relatively free of cholera…

Explaining

What happened during the Napoleon's Russian Campaign?

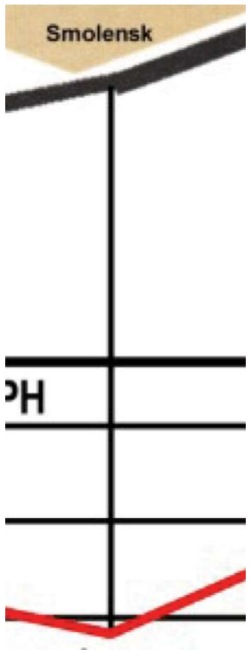
Russian campaign of 1812

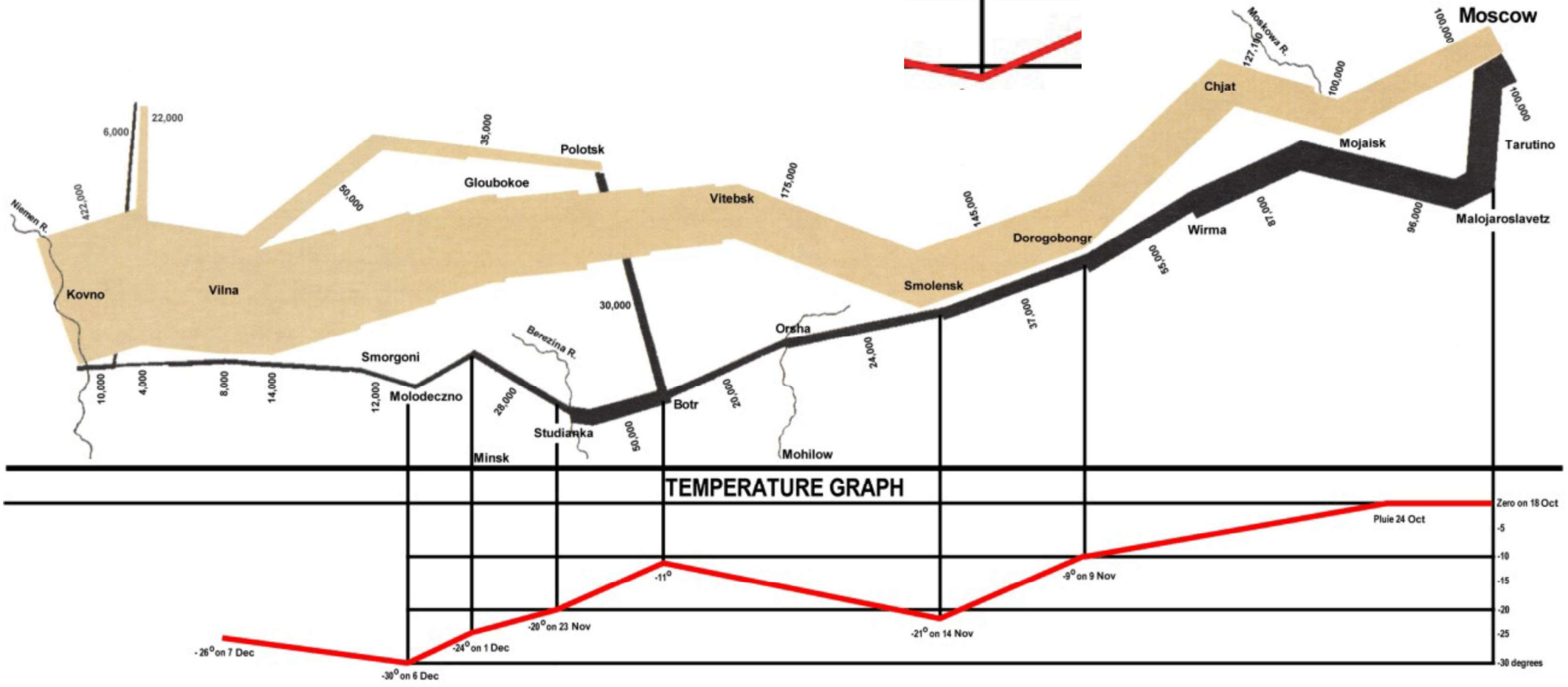
Charles Joseph Minard (1781 – 1870)



- size of the army
- location on a 2D surface
- direction of the army's movement
- temperature
- location of major river crossings

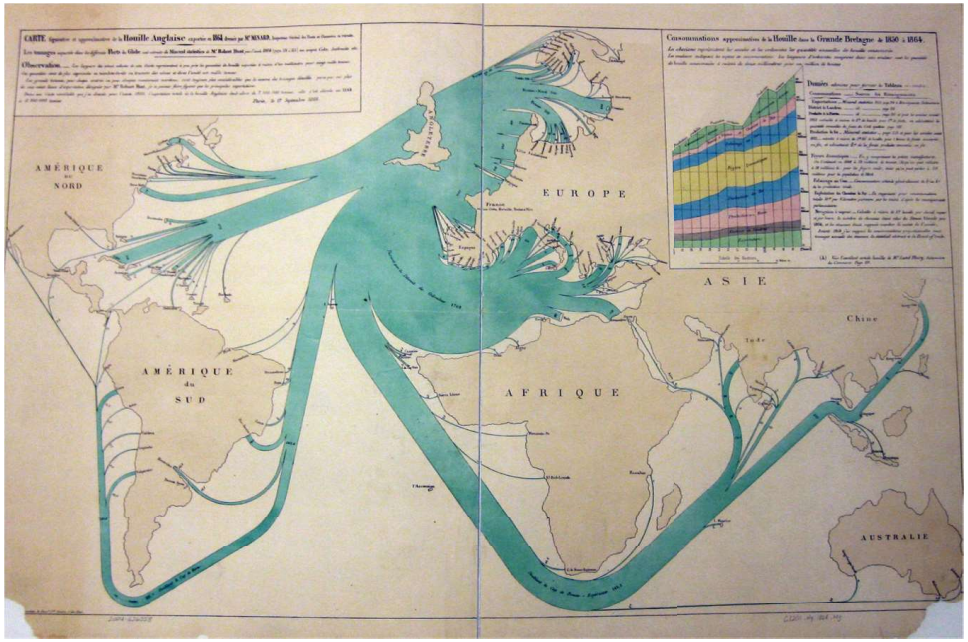
Thickness= number of soldiers
Brown=going to Moscow
Black=coming back
Red=temperature

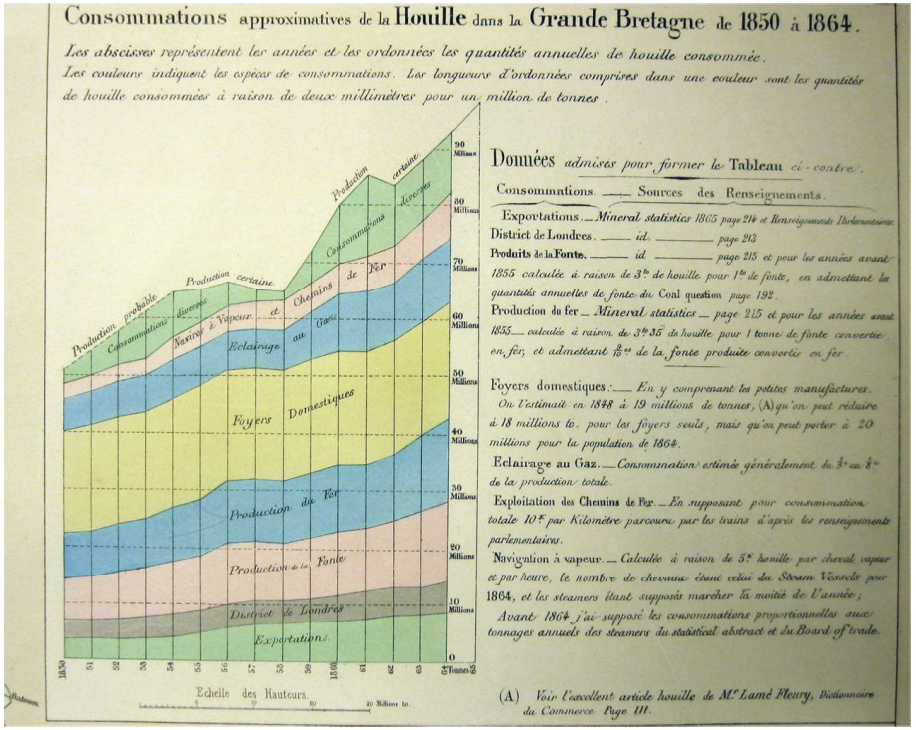




More Minard's maps can be seen at:

http://cartographia.wordpress.com/category/charles-joseph-minard/

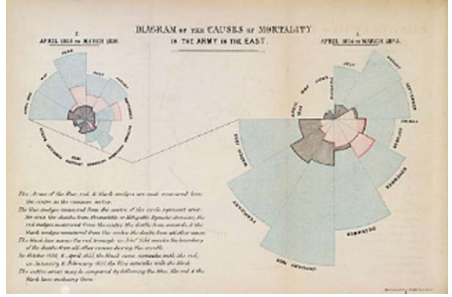


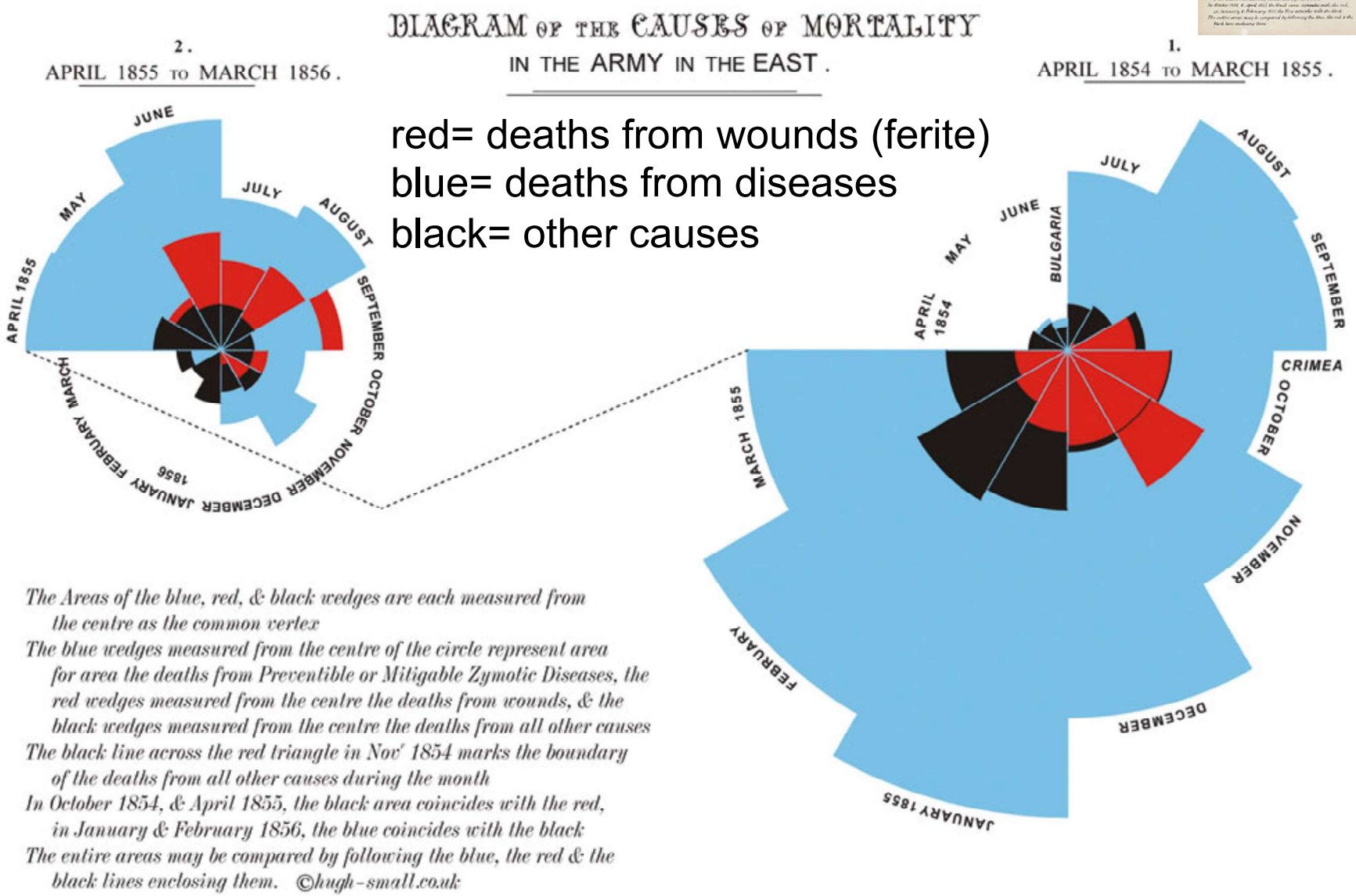


Explaining!!

(several modern visualizations have been inspired by Minard's work)

Time oriented visualization
Florence Nightingale in 1858 during the Crimean War





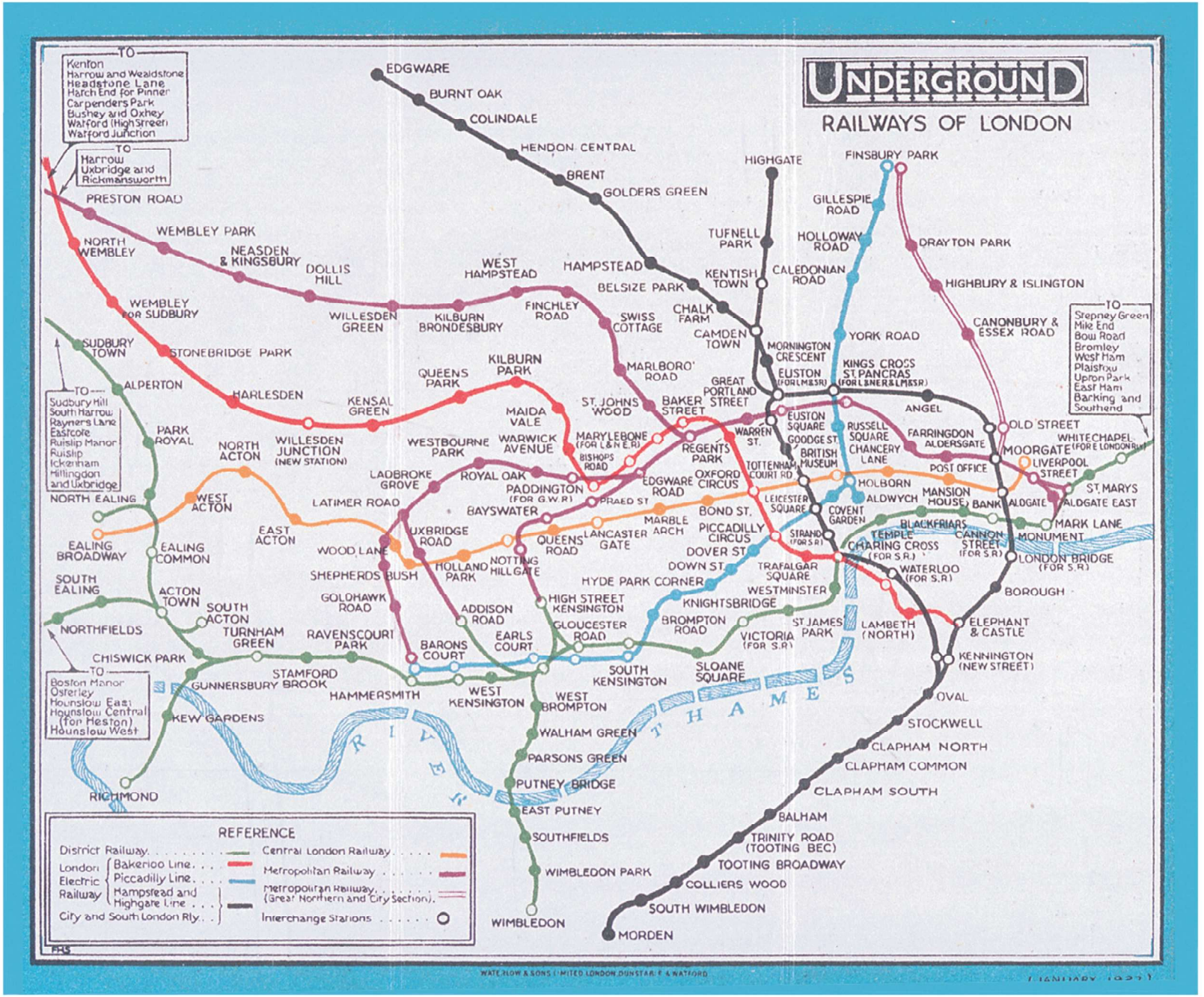
In March 1855 the Sanitary Commission arrived in Turkey, improving the water supply, sewage removal, and ventilation

Visualization for Making decision

Traveling in London by underground

How can I get Queens Park from Victoria?

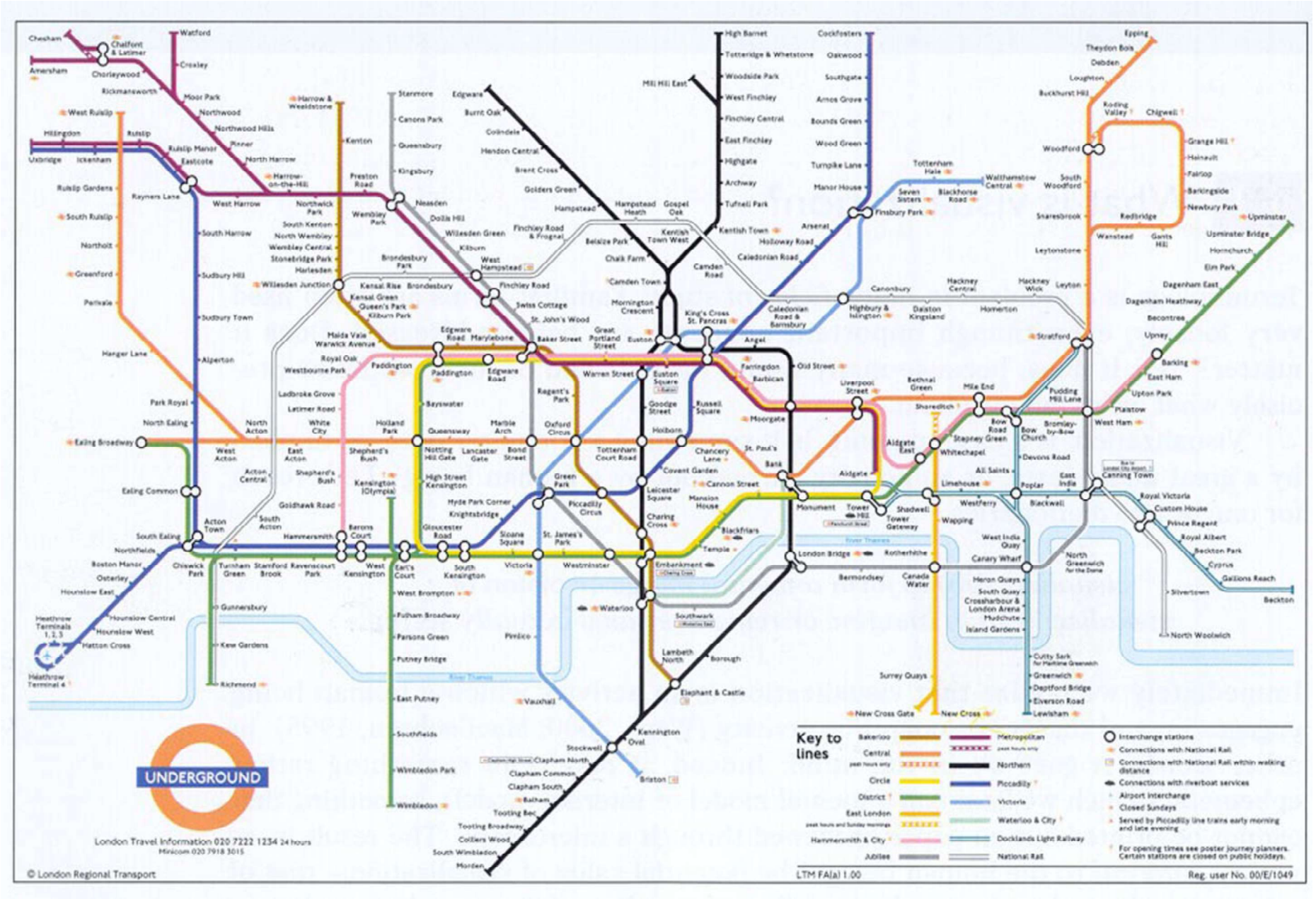
London Underground Map 1927



The Harry Beck's idea

- Real position (when traveling in underground) does not matter
- Only station sequences matter together with their connections
- Beck proposed a "distorted" map
- Actually all the underground maps in the world follow the Beck's approach
- He got a little payment (London underground was not sure about the idea)
- Still true right now: infovis people do not become rich…
- Visual analytics too...

London Underground Map 1990s



Outline

- Facts about the course
- Historical examples
- Definitions
  - The Power of Information Visualization
  - Visual Analytics
- The problem and the involved issues

Moving to the present time

- What is Information Visualization ?
- First of that, what is Visualization ?
- Visualize: to form a mental model or mental image of something
- It is a **cognitive activity** and it has nothing to do with computers

What is Information Visualization?

“Transformation of the symbolic into the geometric”

(McCormick et al., 1987)

“... finding the artificial memory that best supports our natural means of perception.“

(Bertin, 1983)

“Information visualization is the use of computer-supported, *interactive*, visual representations of *abstract data* to *amplify cognition*.”

(Card et al., 1999)

What is Information Visualization?

Information visualization is the use of *computer-supported*, *interactive*, visual representations of *abstract data* to *amplify cognition*.

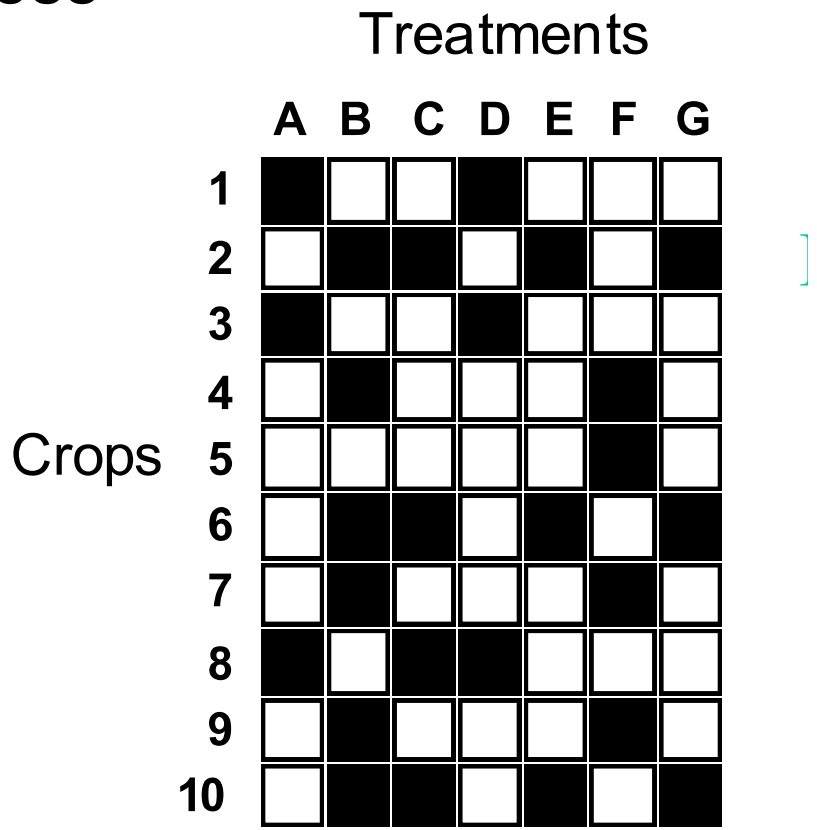
[Card et al. '99]

…computer supported and interactive

- **Computer-supported**
  - Even if beautiful examples of paper based visualizations exist the actual understanding of information visualization (infovis) is about computer based visualization, **but we have to always remember that a cognitive activity is involved in the process**
- **Interactive**
  - To exploit the full power of infovis techniques interaction is mandatory. The user must be allowed for manipulating the visualization to better reach his goals

Interaction example

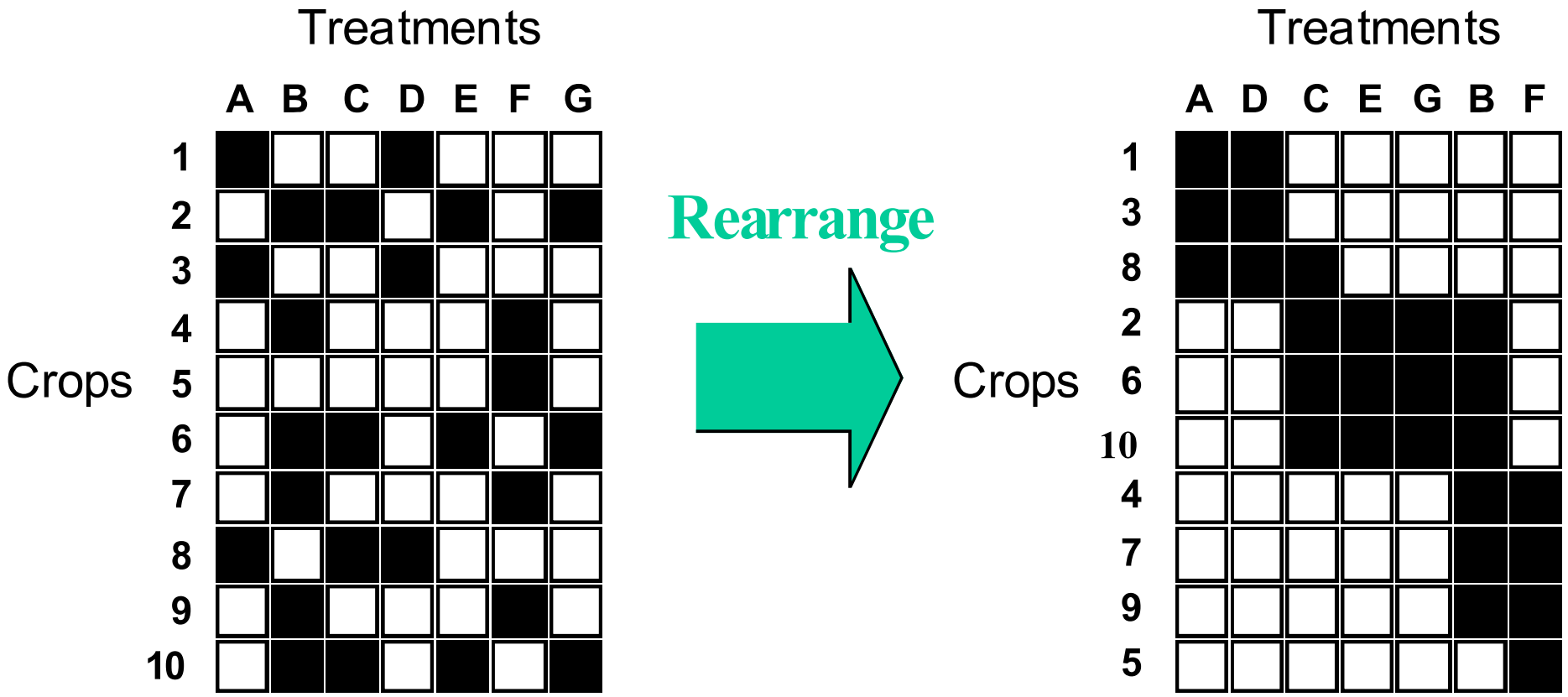
- Agronomists are experimenting 7 treatments (anti-parasite, fertilizer, etc.) on 10 different crops (piantagioni)
- A black square indicates success
- Does this visualization help?



Interaction example

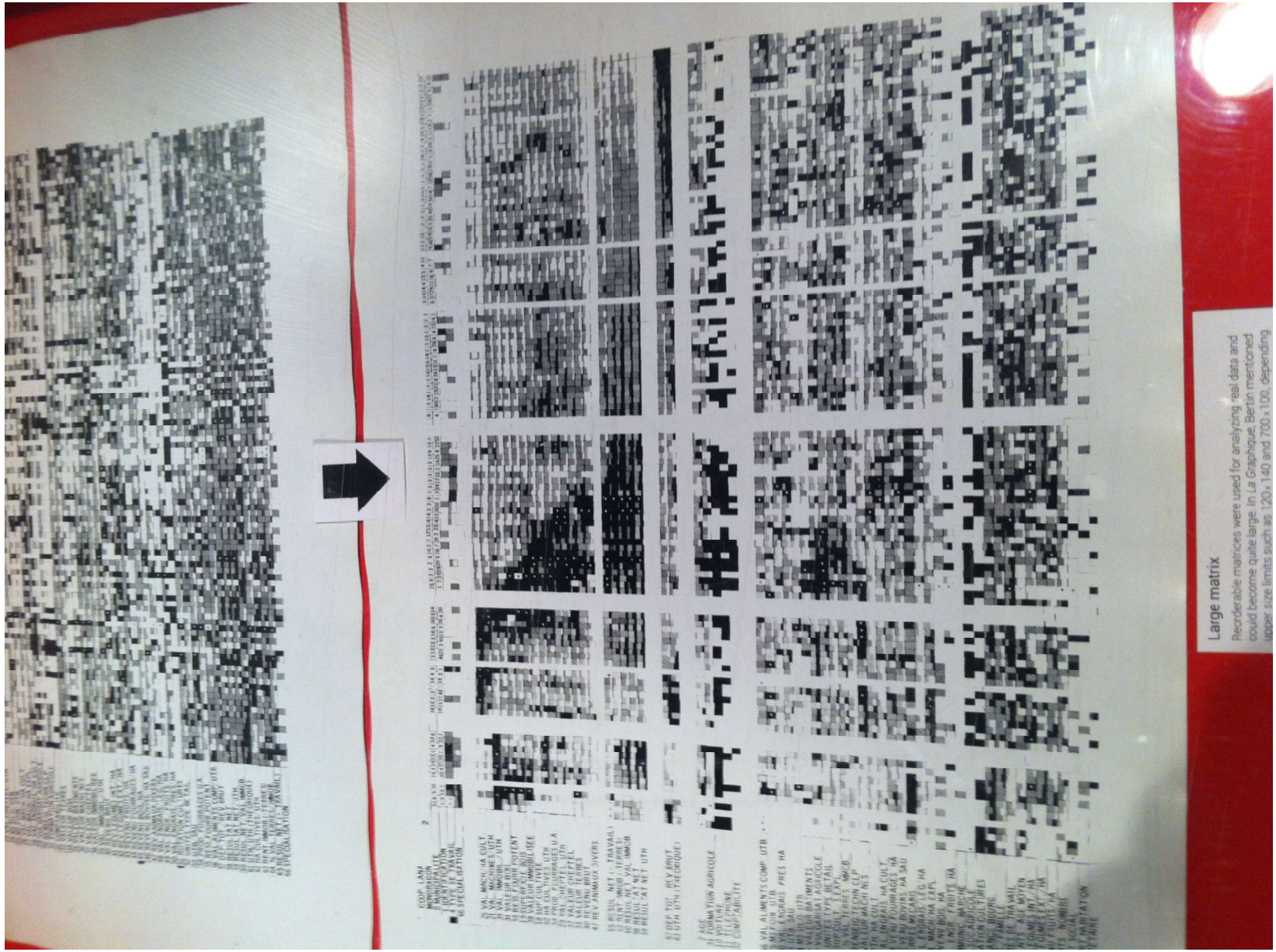
Let's rearrange the columns

- thanks to Jacques *Bertin* (27 July 1918 – 3 May 2010)

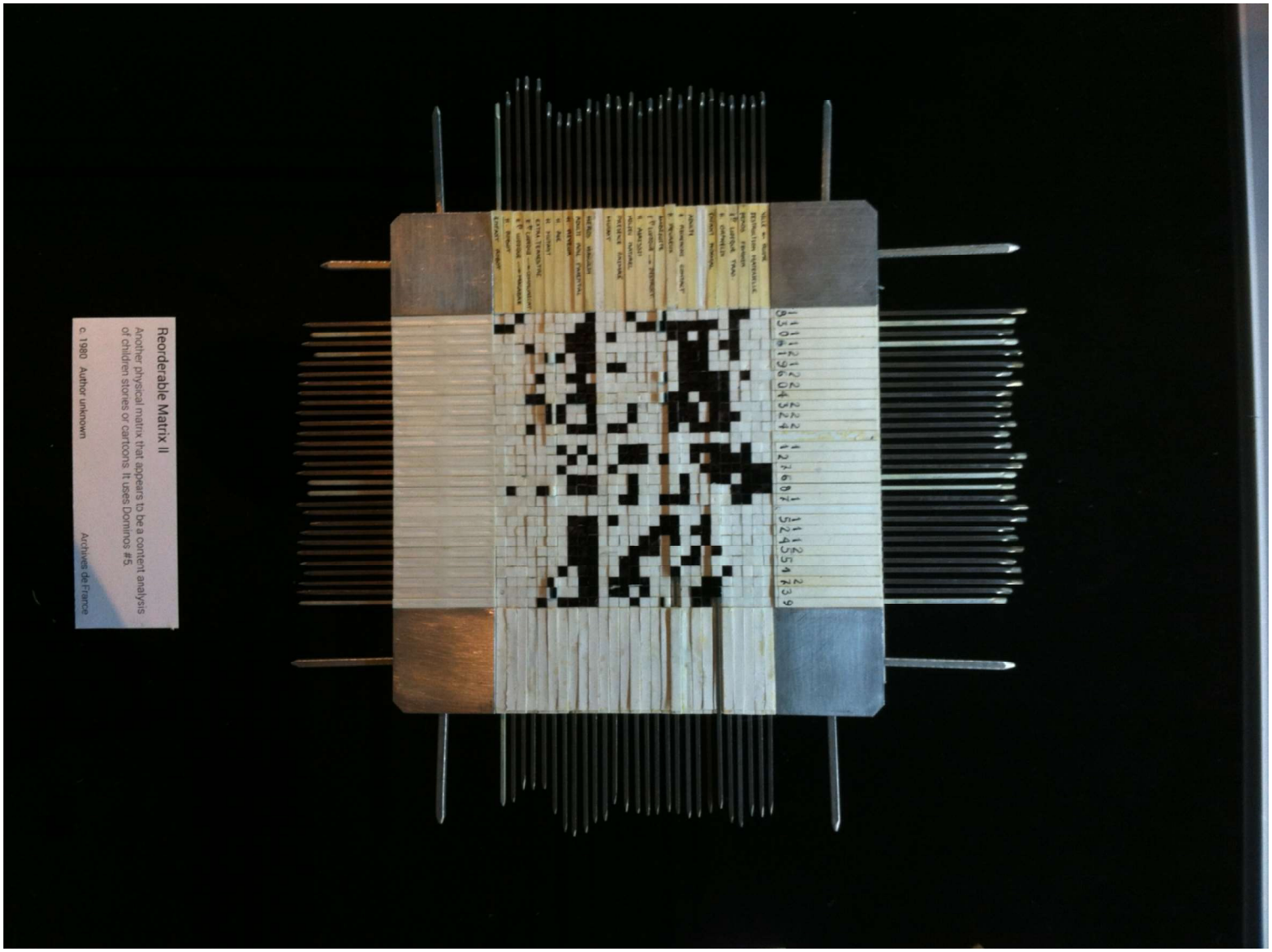


That raises the possibility of optimizations! E.g., arranging crops 1, 3, 8 close each other...

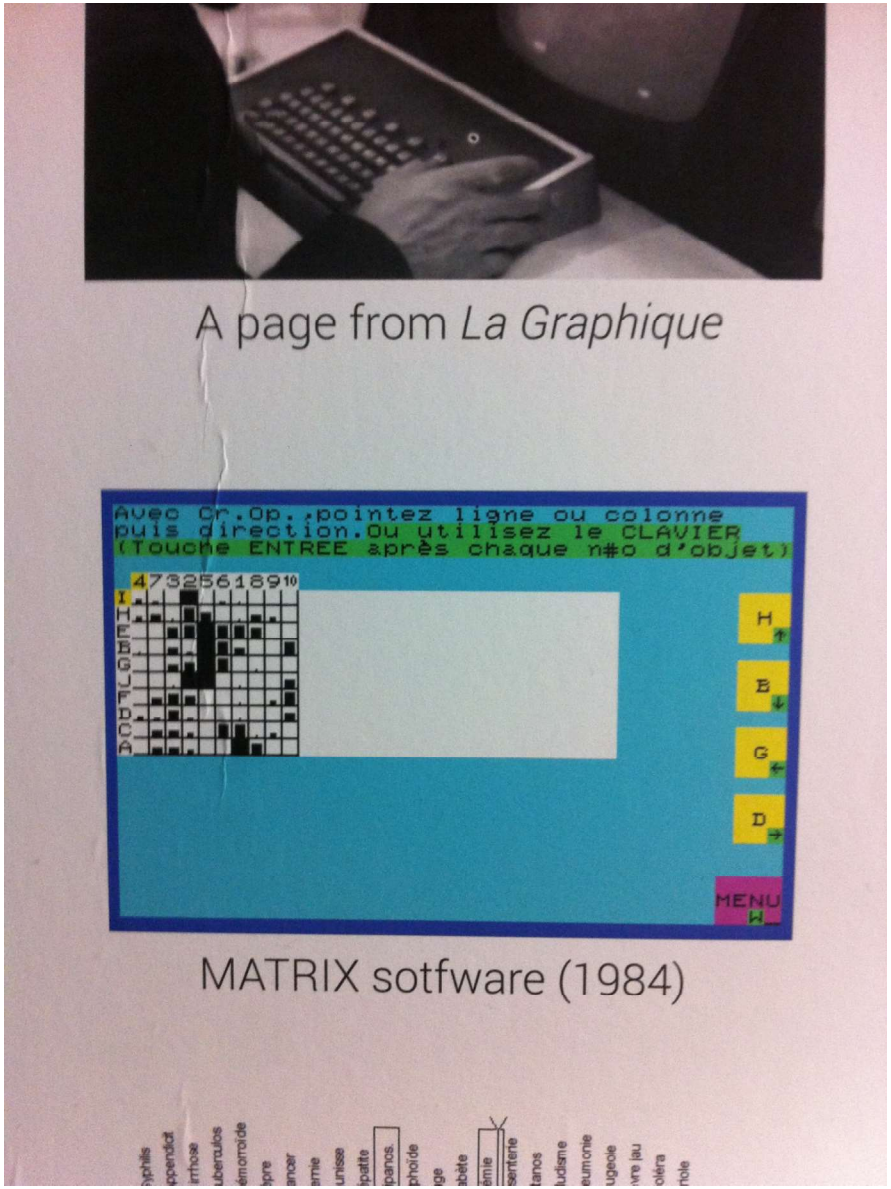
Bertin’s work (drugs and mental illness)



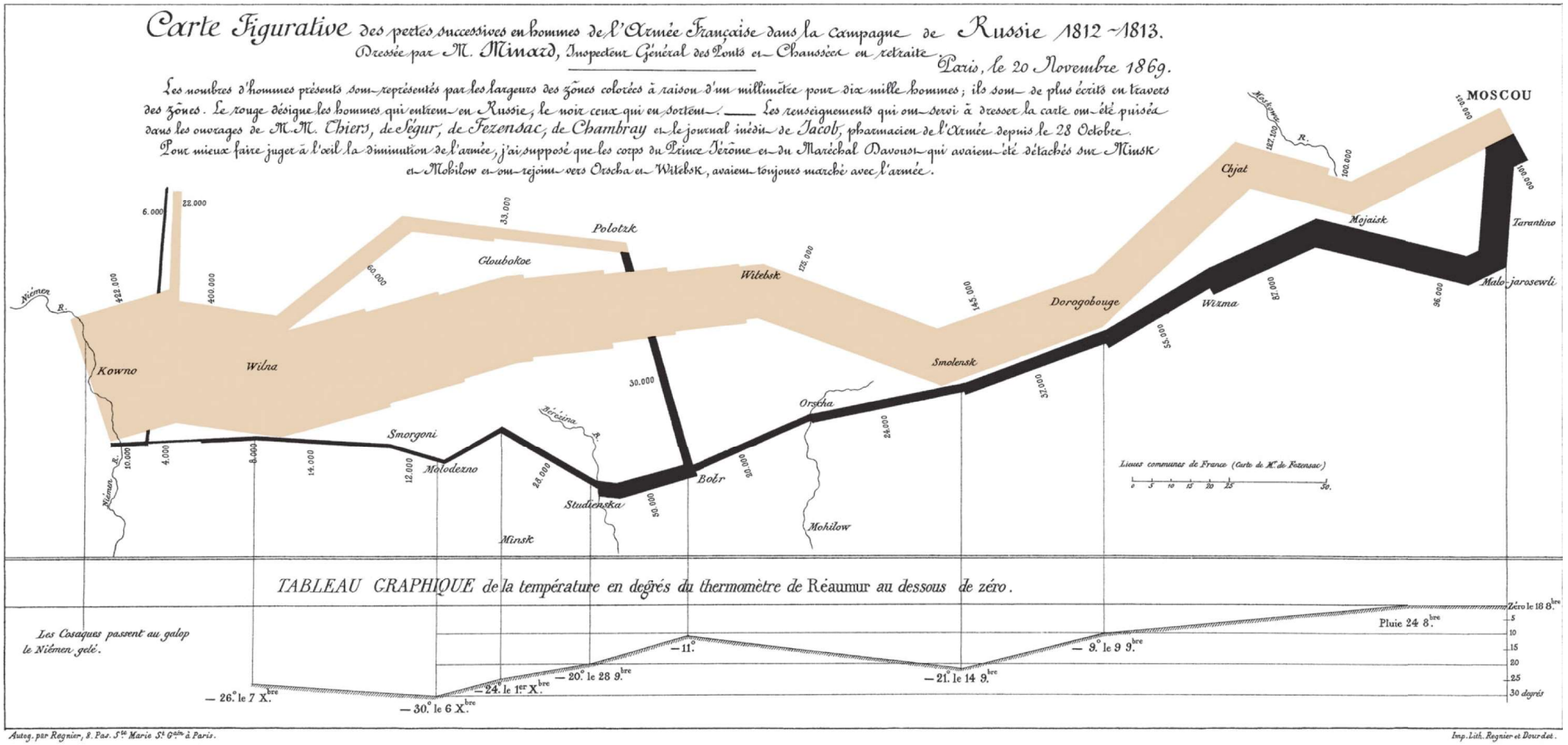
Manually!


Reorderable Matrix II
Another physical matrix that appears to be a content analysis of children stories or cartoons. It uses Dominos #5
c. 1980 Author unknown
Archives de France

Computer !

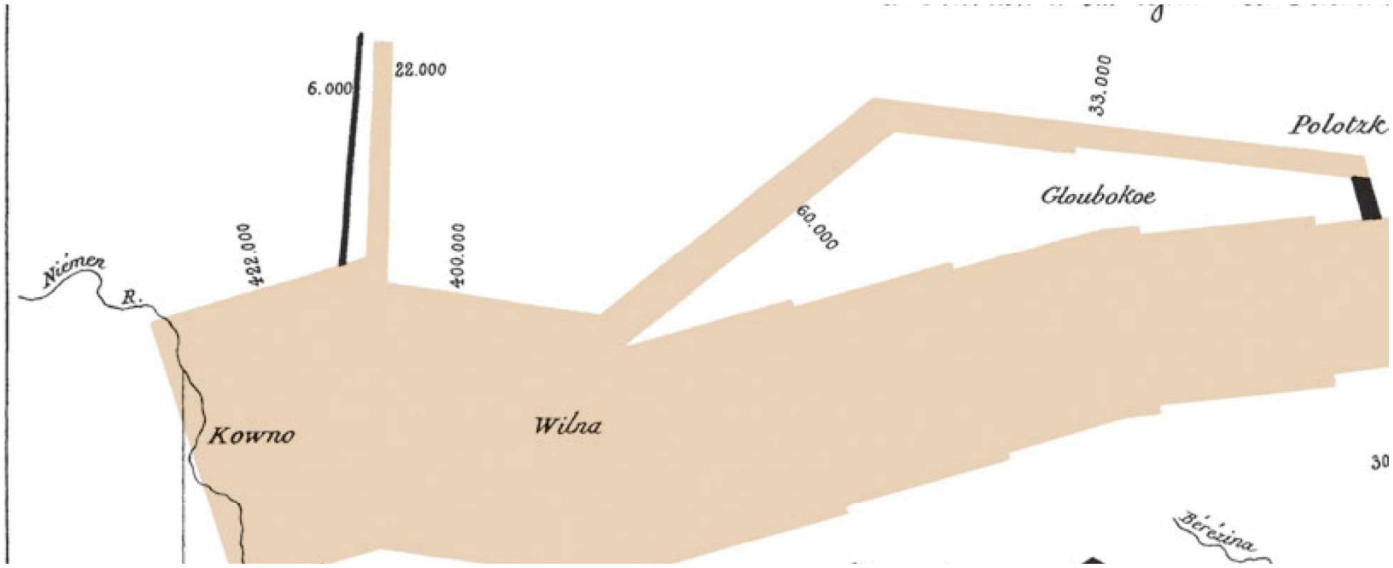

A page from *La Graphique*
Avec Cr.Op. pointez ligne ou colonne
puis direction.Ou utilisez le CLAVIER
(Touche ENTREE après chaque n#o d'objet)
H
B
G
D
MENU
MATRIX sotfware (1984)

How can computer and interaction improve this viz?



- size of the army
- location on a 2D surface
- direction of the army's movement
- temperature
- location of major river crossings

What about forecasting? (but this is Visual Analytics...)



?

Moscow

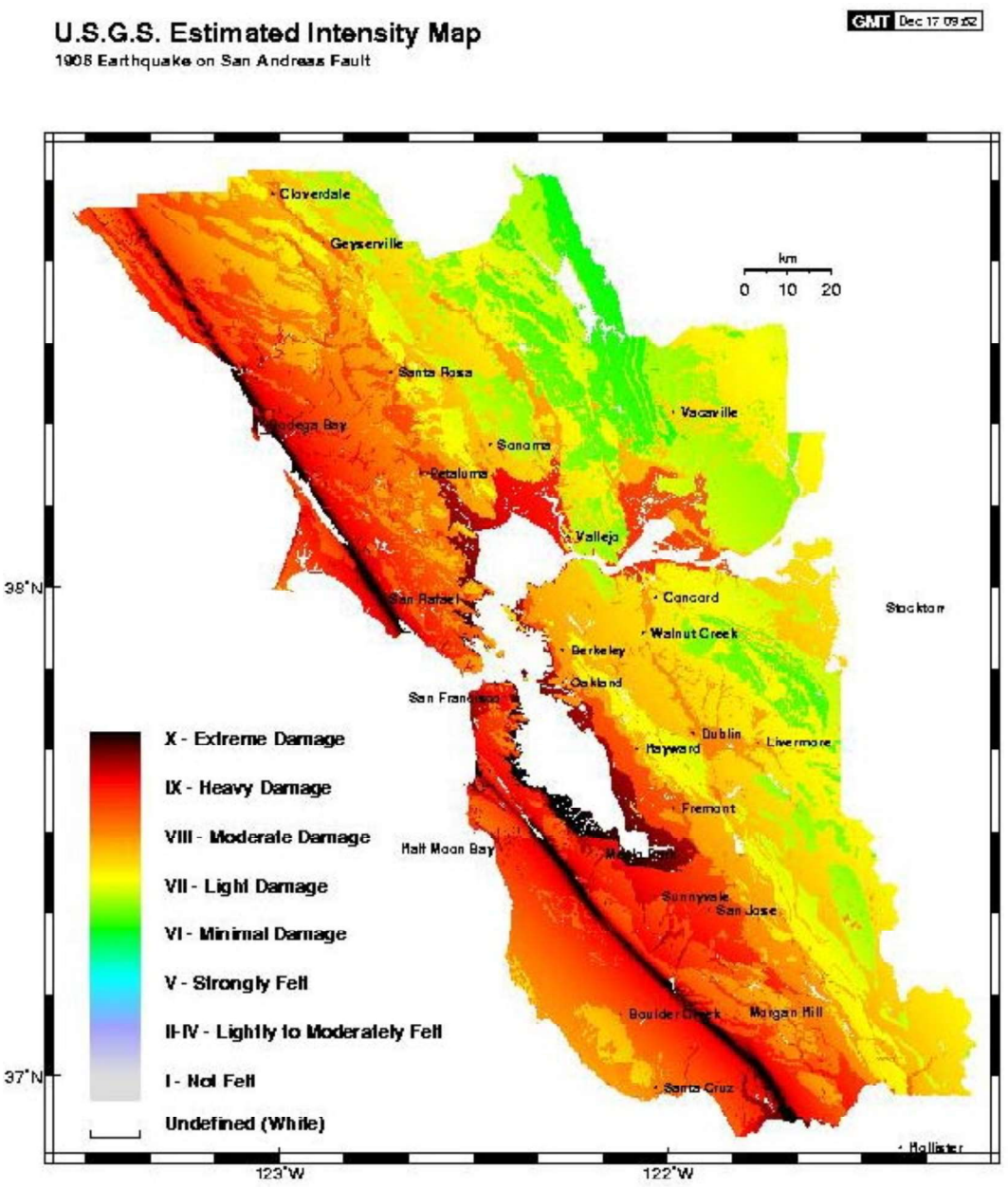
Distance to Moscow

- size of the army
- location on a 2D surface
- direction of the army's movement
- temperature
- location of major river crossings

…it is about abstract data

- Abstract data
  - Information visualization is about visualizing abstract data, i.e., data that does not refer to physical situation. In other words it is NOT scientific visualization/geographic visualization
- Scientific visualization primarily relates to and represents something physical or geometric
- Examples
  - Air flow over a wing
  - Weather over Italy
  - Torrents inside a tornado
  - Organs in the human body
  - Molecular bonding…

Scientific/geographic visualization

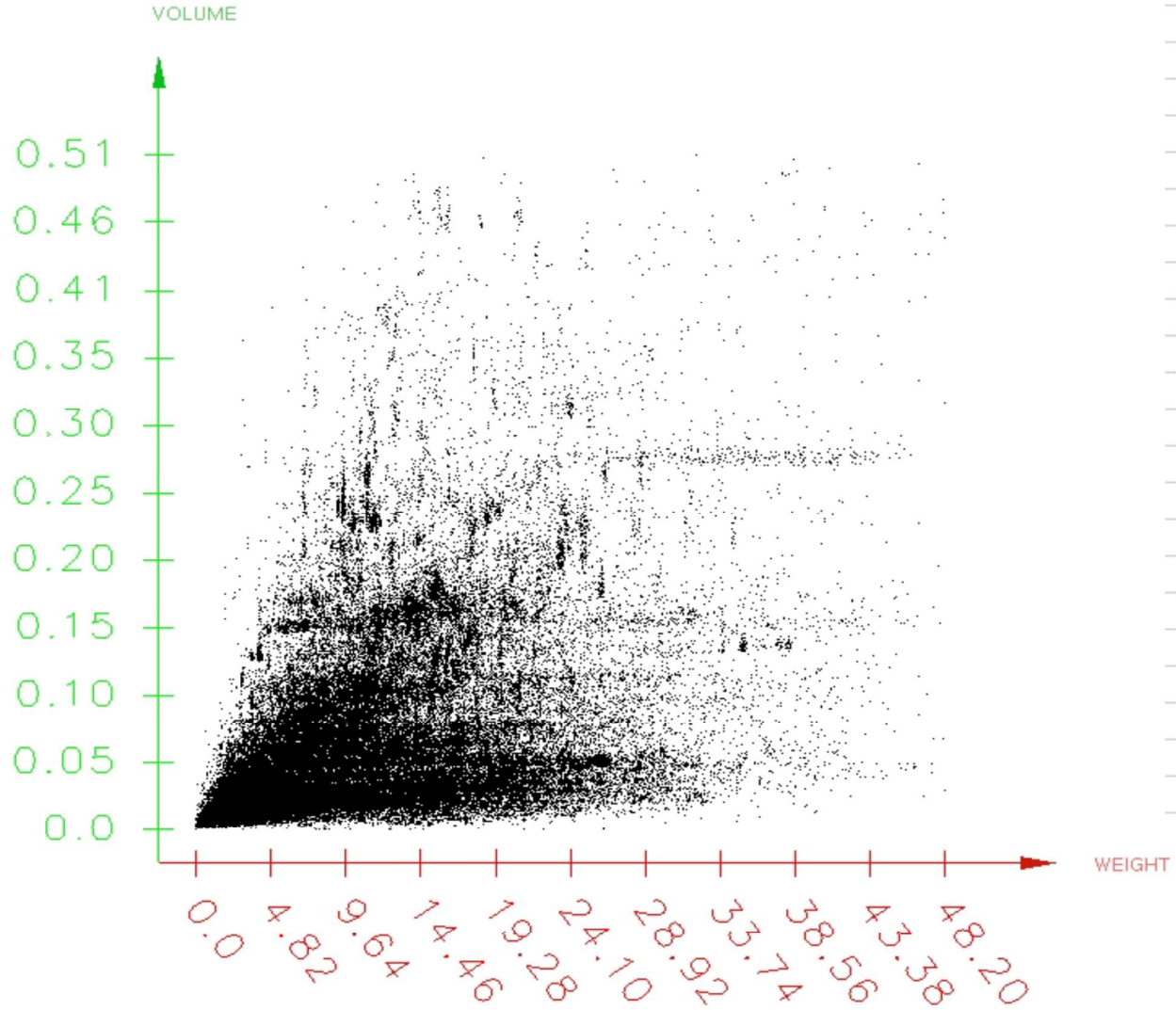


Earthquake intensity

…abstract data

- What the visualized “information” is?
  - Items that do not have a direct **physical/visual** correspondence (or such a correspondence is not relevant for the application)
  - Examples: sport statistics, stock trends, query results, software data, etc…
- Items are represented on a 2D / 3D physical space using their numerical characteristics (attributes)
- The visualization is useful for analysis and decision-making (not just fun or colors)
- E.g., Postal parcels
  - Shipping date
  - Volume
  - Weight
  - Sender country
  - Receiver country
  - …

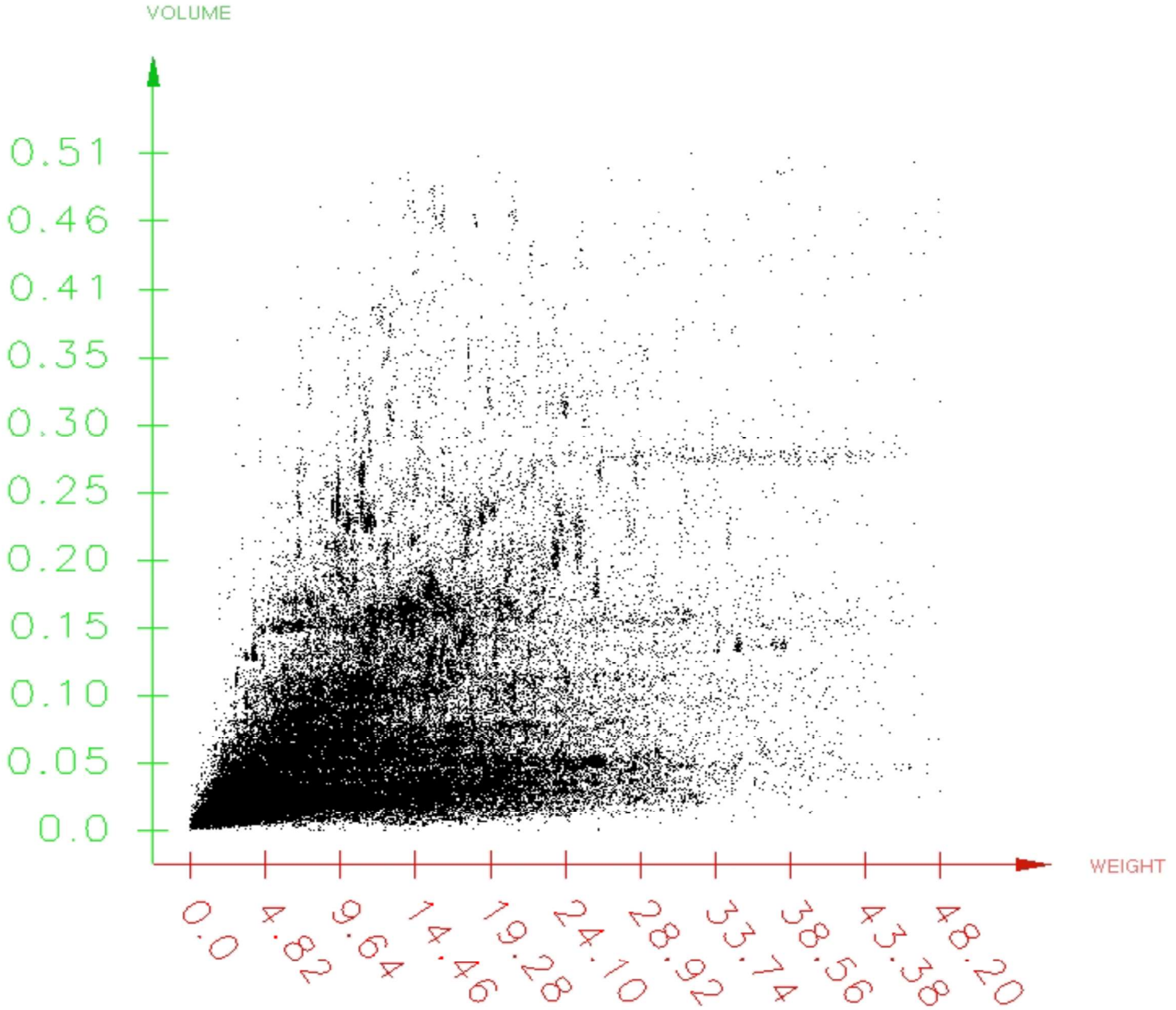
Abstract data



peso	volum
4.1700	0.0307
10.5100	0.0181
2.6500	0.0171
2.9900	0.0521
1.9000	0.0070
4.9300	0.0644
1.7300	0.0123
2.6500	0.0254
13.1400	0.1037
3.3600	0.0112
2.6000	0.0223
1.3100	0.0276
6.6800	0.1342
2.5400	0.0159
11.3400	0.0203
0.0000	0.0000
6.7800	0.0719
2.3300	0.0239
8.2400	0.1028
14.9800	0.0313
5.3300	0.0332
15.5500	0.0153
5.4100	0.0318
1.1300	0.0192
15.7800	0.1161
13.3200	0.0888

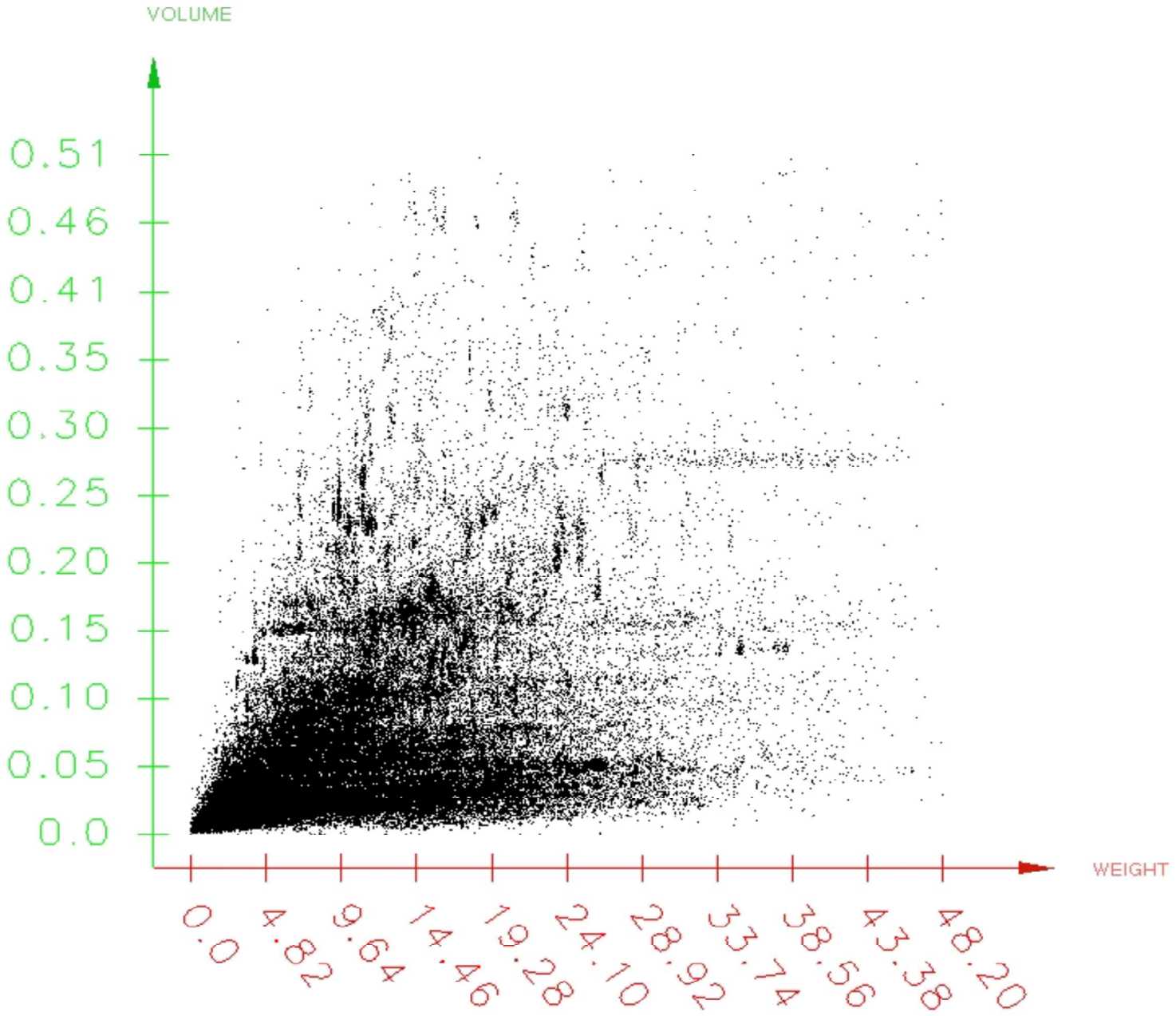
A 2D Scatterplot showing about 200.000 postal parcels

What can you discover from it?



peso	volum
4.1700	0.0307
10.5100	0.0181
2.6500	0.0171
2.9900	0.0521
1.9000	0.0070
4.9300	0.0644
1.7300	0.0123
2.6500	0.0254
13.1400	0.1037
3.3600	0.0112
2.6000	0.0223
1.3100	0.0276
6.6800	0.1342
2.5400	0.0159
11.3400	0.0203
0.0000	0.0000
6.7800	0.0719
2.3300	0.0239
8.2400	0.1028
14.9800	0.0313
5.3300	0.0332
15.5500	0.0153
5.4100	0.0318
1.1300	0.0192
15.7800	0.1161
13.3200	0.0888

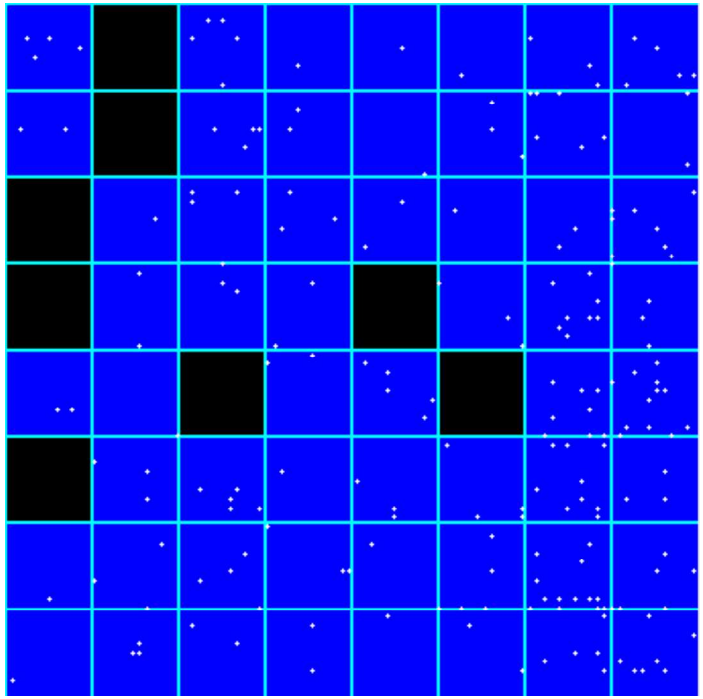
What can you NOT discover from it?



Density maps

4 data items are plotted on the same pixel:**d=4**

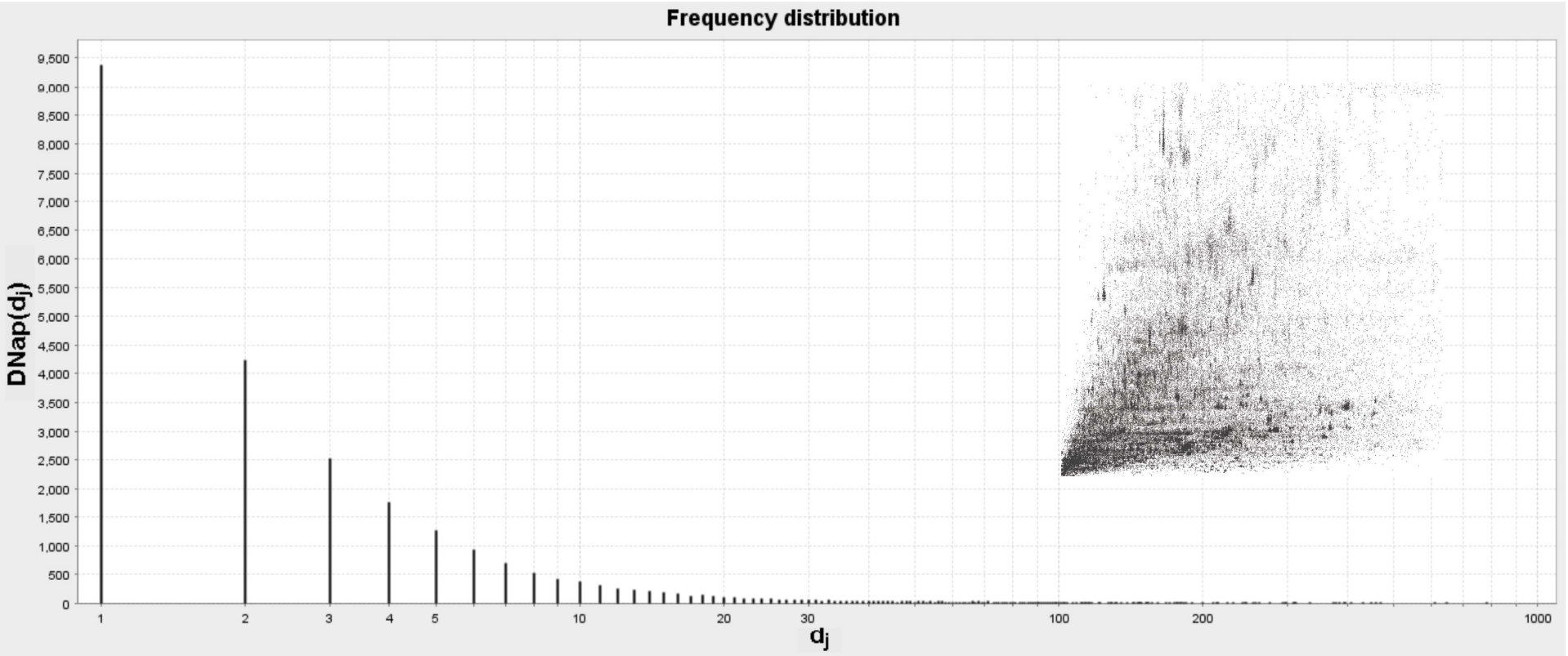
empty pixel



8x8 pixels

we can **map** the density values to a 256 levels grey or color scale

In the example we borrow the Keim&Kriegel [KK95] color scale, presenting a monotonically increasing brightness

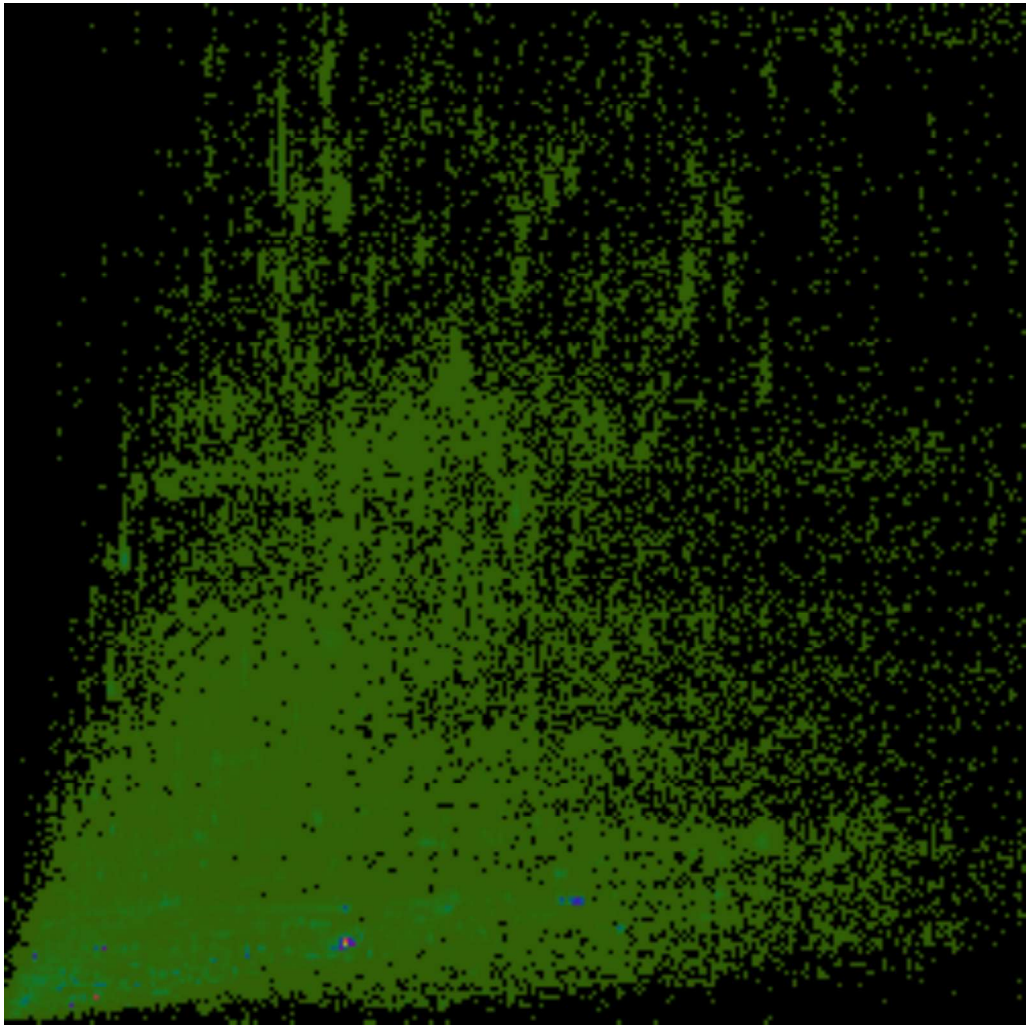


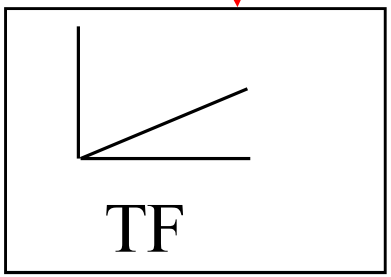
126 different data densities = { 1, 2, … , 1,633 }

?

256 Color Codes = { 0,1, 2, … , 255}

$$ColorCode(d) = Round\left[255\,\frac{d - d_{\min}}{d_{\max} - d_{\min}}\right]$$

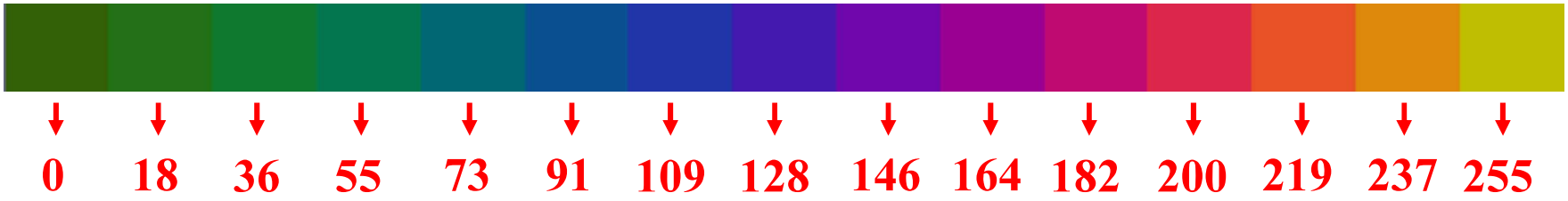




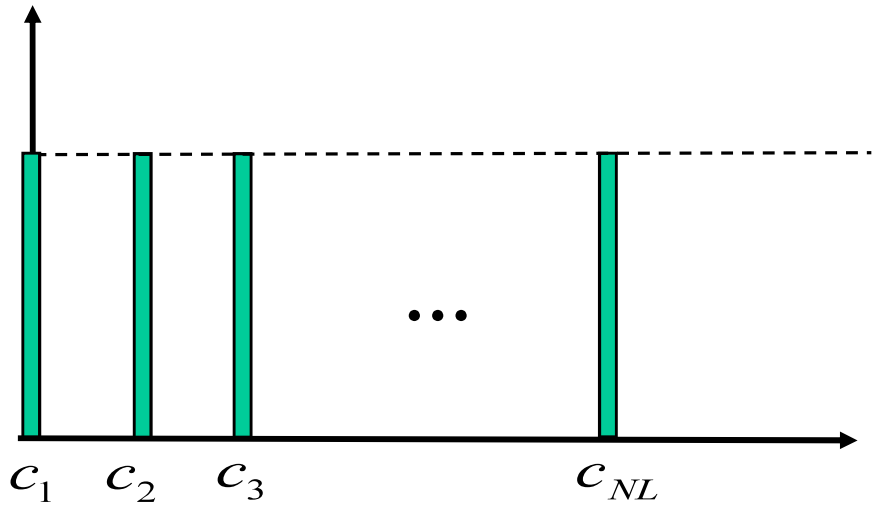
Linear?

Uniform scale mapping

We use a reduced color scale, e.g. with 15 codes (N_L=15)

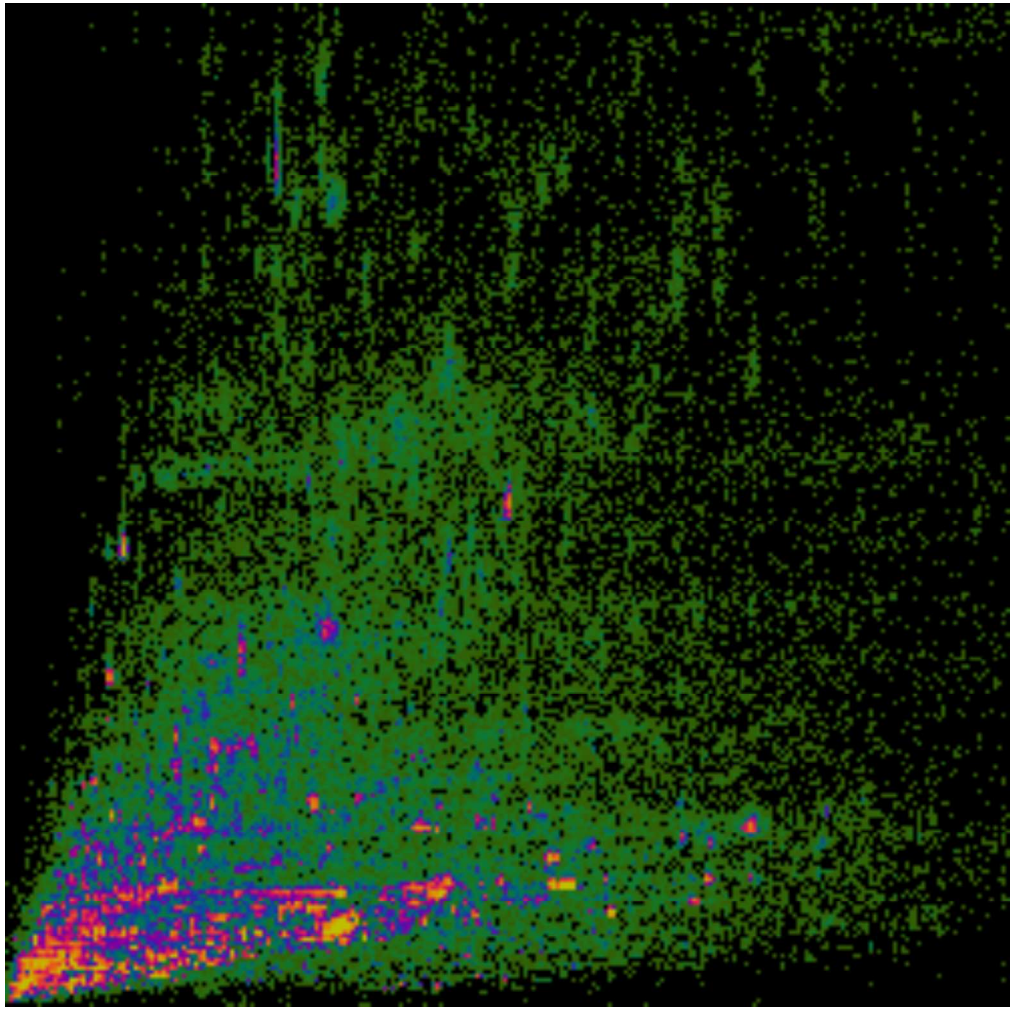


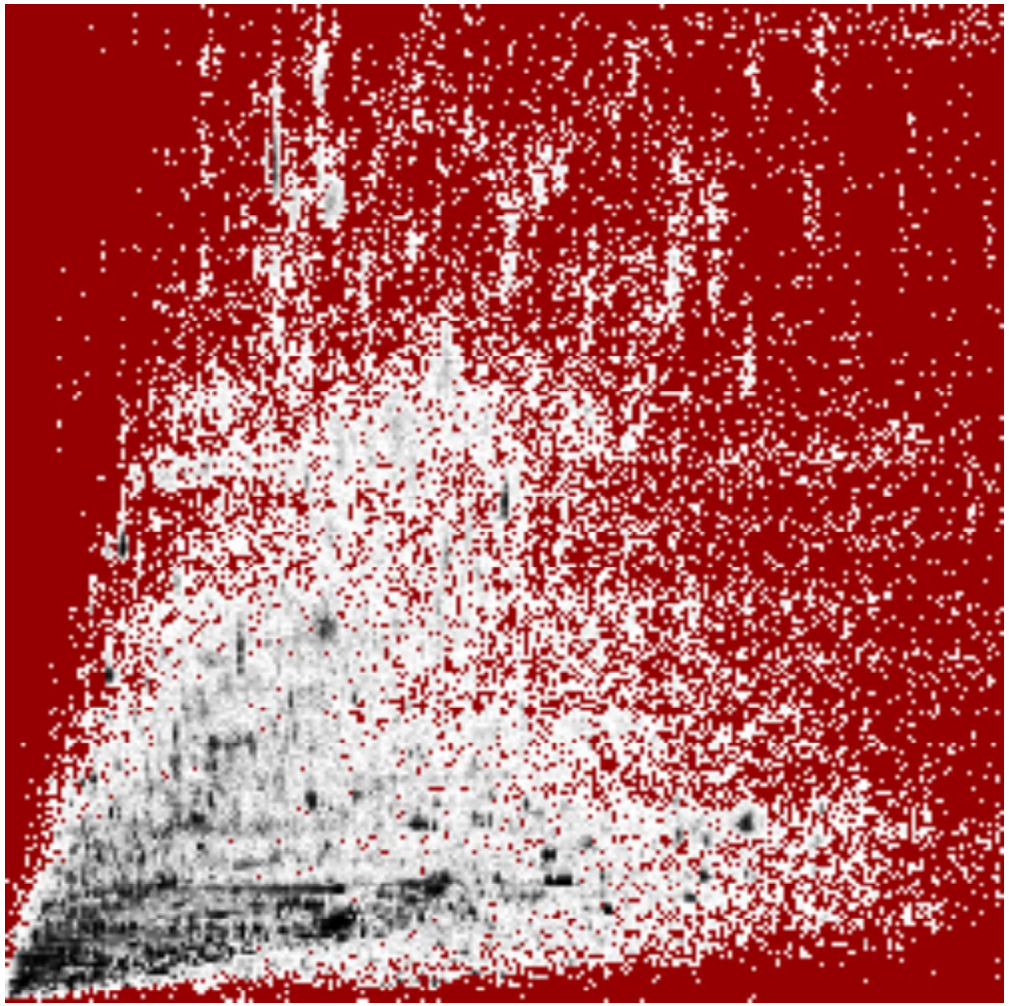
This implies that **different** density values will be **necessarily** represented by the **same** color code: to reduce the degradation the mapping is performed through an algorithm that tries to assign to each code the same number of pixels



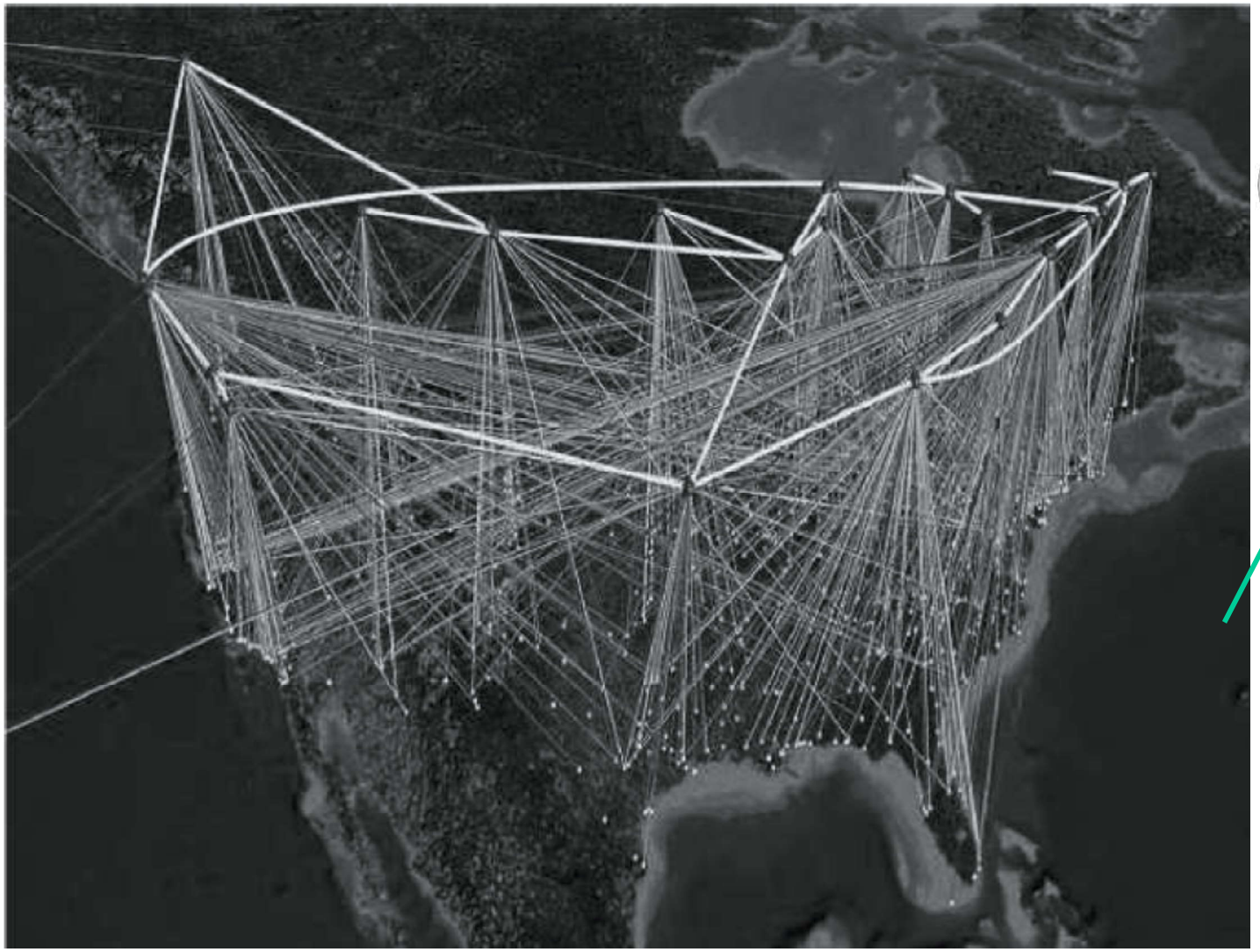
Target color code frequency distribution

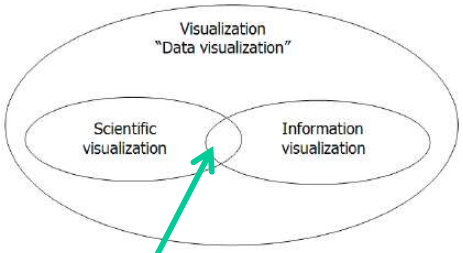
45





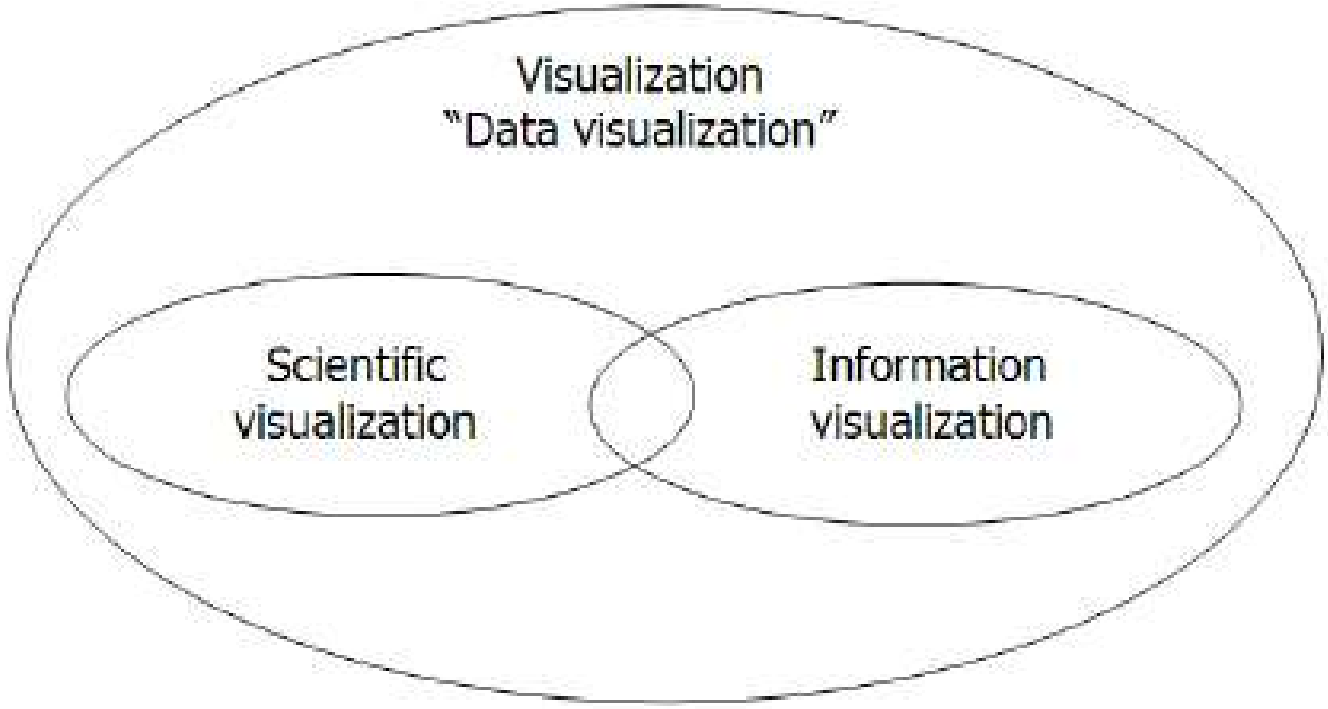
Mixed visualization





Byte traffic into the ANS/NSFNET T3 backbone in 1993

Overview


Visualization
"Data visualization"
Scientific
visualization
Information
visualization

… it amplifies cognition

Amplify cognition using the human vision

- Highest bandwidth sense
- Fast, parallel
- Pattern recognition
- Pre-attentive
- Extends memory and cognitive capacity
  - Multiplication test
- People think visually (I see… means also I understand in most languages)
- Amplify cognition
  - Presenting data in the right way, taking into account the way in which human vision system works can greatly improve the comprehension of complex phenomena
- Three very simple examples (put away pencil and paper…)

Amplify cognition

- Example: multiplication (Card, Moran, & Shneiderman.)

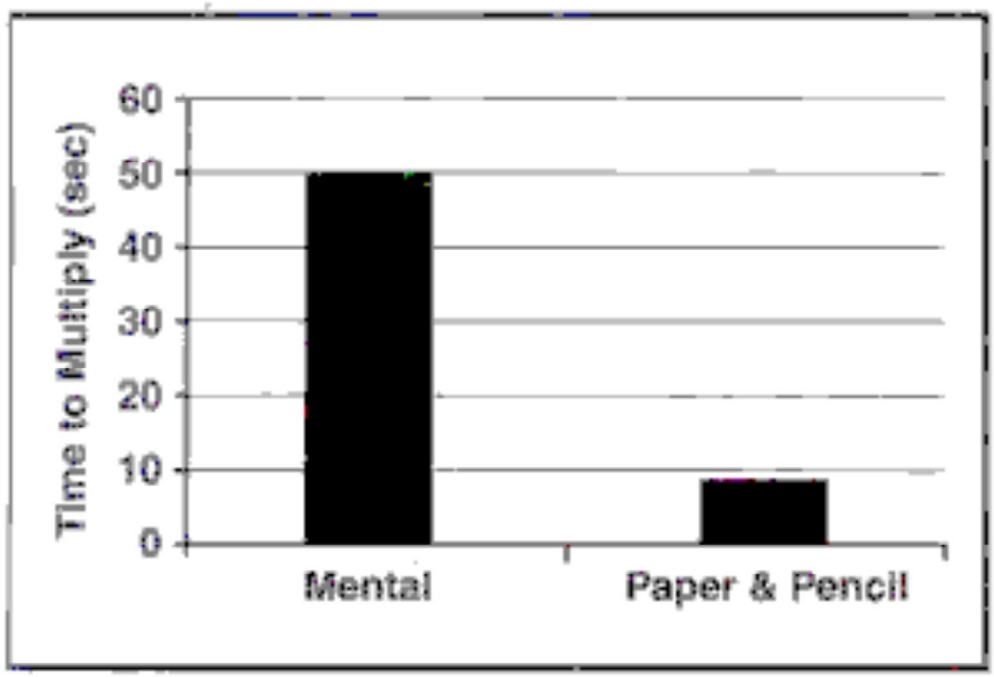
- In your head, multiply 35 x 95

Amplify cognition

- Now do it on paper

Visual Aids for Thinking

- People are 5 times faster with the visual aid



Three simple questions

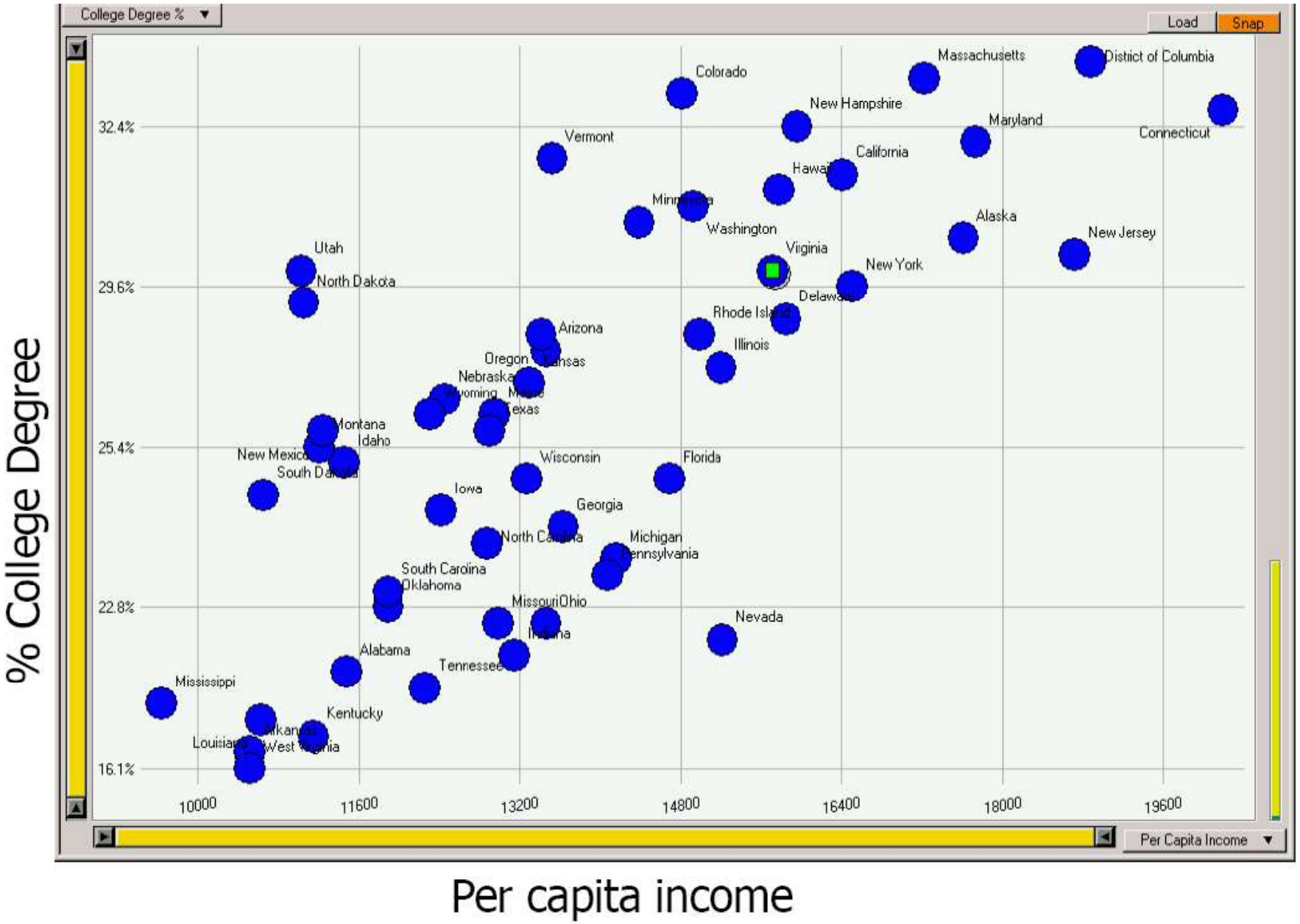
Table - StateData ()

Load Snap

State	College Degree %	Per Capita Income
Alabama	20.6%	11486
Alaska	30.3%	17610
Arizona	27.1%	13461
Arkansas	17.0%	10520
California	31.3%	16409
Colorado	33.9%	14821
Connecticut	33.8%	20189
Delaware	27.9%	15854
District of Columbia	36.4%	18881
Florida	24.9%	14698
Georgia	24.3%	13631
Hawaii	31.2%	15770
Idaho	25.2%	11457
Illinois	26.8%	15201
Indiana	20.9%	13149
Iowa	24.5%	12422
Kansas	26.5%	13300
Kentucky	17.7%	11153
Louisiana	19.4%	10635
Maine	25.7%	12957
Maryland	31.7%	17730
Massachusetts	34.5%	17224
Michigan	24.1%	14154
Minnesota	30.4%	14389
Minnesota	30.4%	14389
Mississippi	19.9%	9648
Missouri	22.3%	12989
Montana	25.4%	11213
Nebraska	26.0%	12452
Nevada	21.5%	15214
New Hampshire	32.4%	15959
New Jersey	30.1%	18714
New Mexico	25.5%	11246
New York	29.6%	16501
North Carolina	24.2%	12885
North Dakota	28.1%	11051
Ohio	22.3%	13461
Oklahoma	22.8%	11893
Oregon	27.5%	13418
Pennsylvania	23.2%	14068
Rhode Island	27.5%	14981
South Carolina	23.0%	11897
South Dakota	24.6%	10661
Tennessee	20.1%	12255
Texas	25.5%	12904
Utah	30.0%	11029
Vermont	31.5%	13527
Virginia	30.0%	15713
Washington	30.9%	14923
West Virginia	16.1%	10520
Wisconsin	24.9%	13276
Wyoming	25.7%	12311

Which state has the highest % college degree?
Highest Income? Relationship between college and income?

The quick answers



Visualization reveals data!

- Anscombes' quartet: we have 4 samples of 11 (x,y) points
- same linear model
- same mean: 7.5
- same variance: 4.127
- same correlation: 0.816

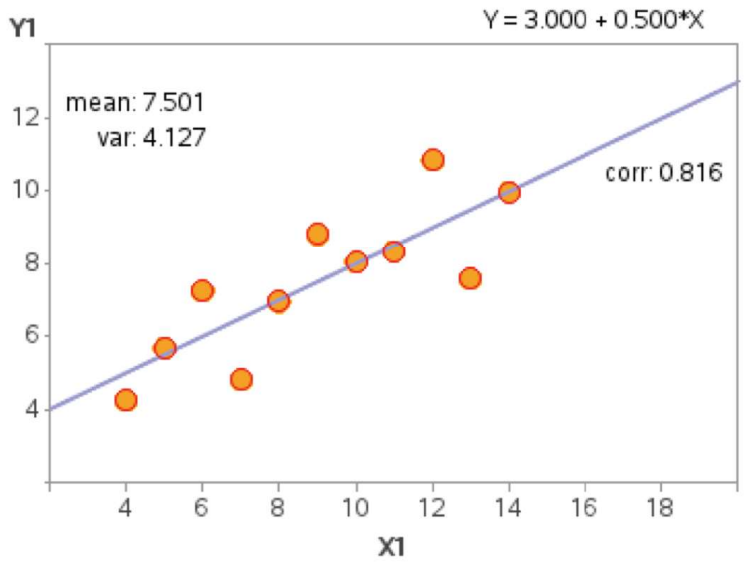
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

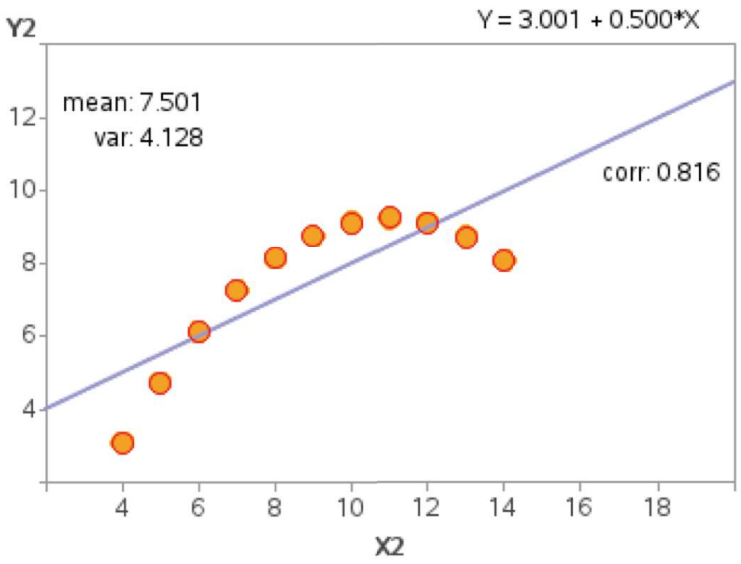
- Can you figure out how samples look like?

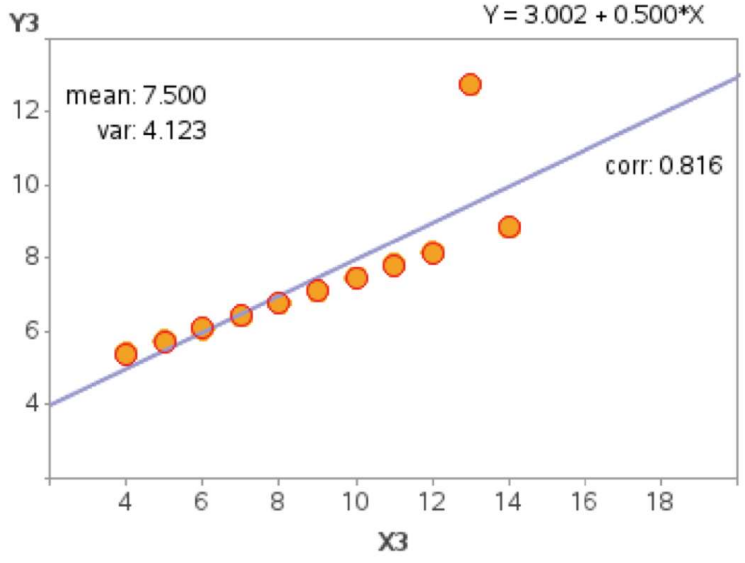
Here is the data

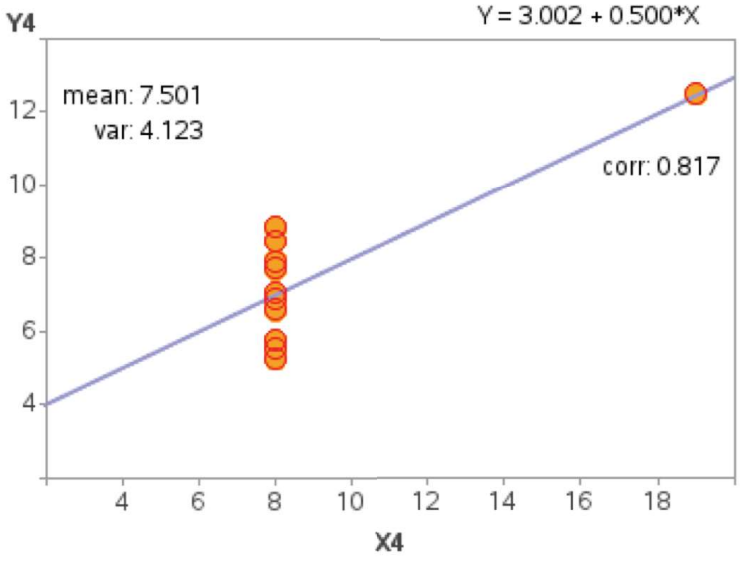
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Visualization reveal data!









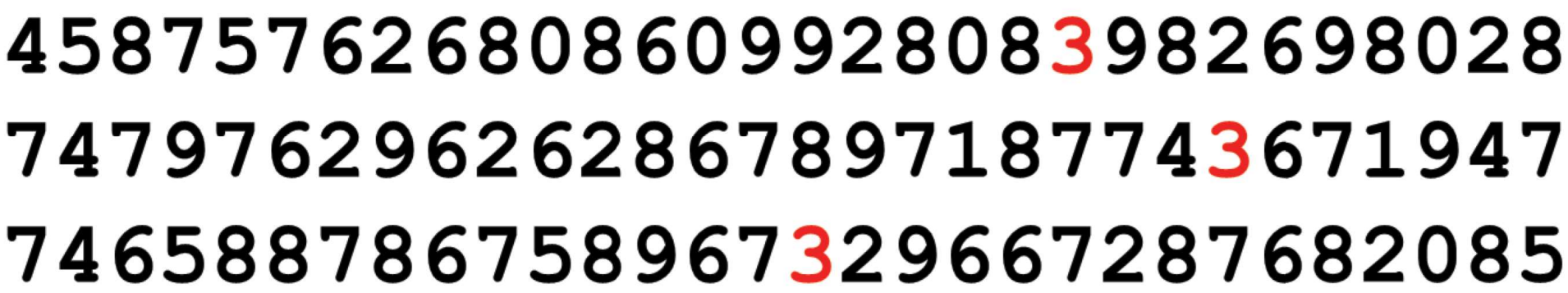
One (very) simple question

- How many 3s here ?
- You have 4 seconds…

Game over!

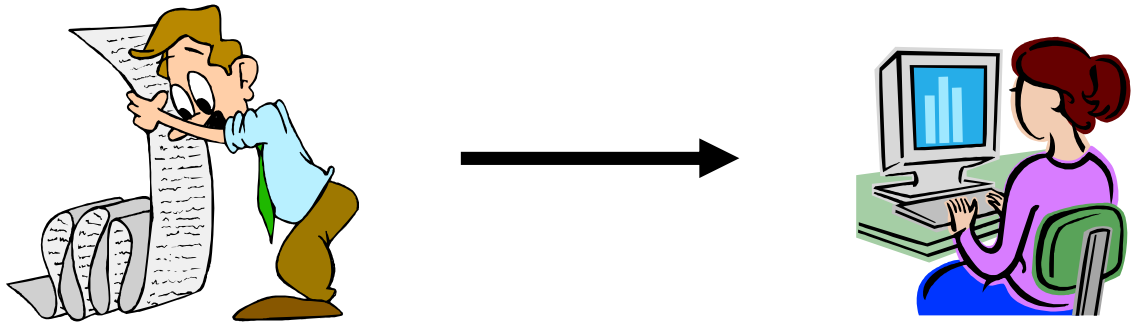
So ?

- Time is not enough?
- You can do that in less than half of a second !



- Color is pre-attentive (pops up)
- No cognitive effort is required
- A lot of issues are already clear
- Most of the people ignore them...

Information visualization, statistics and data mining, Visual Analytics

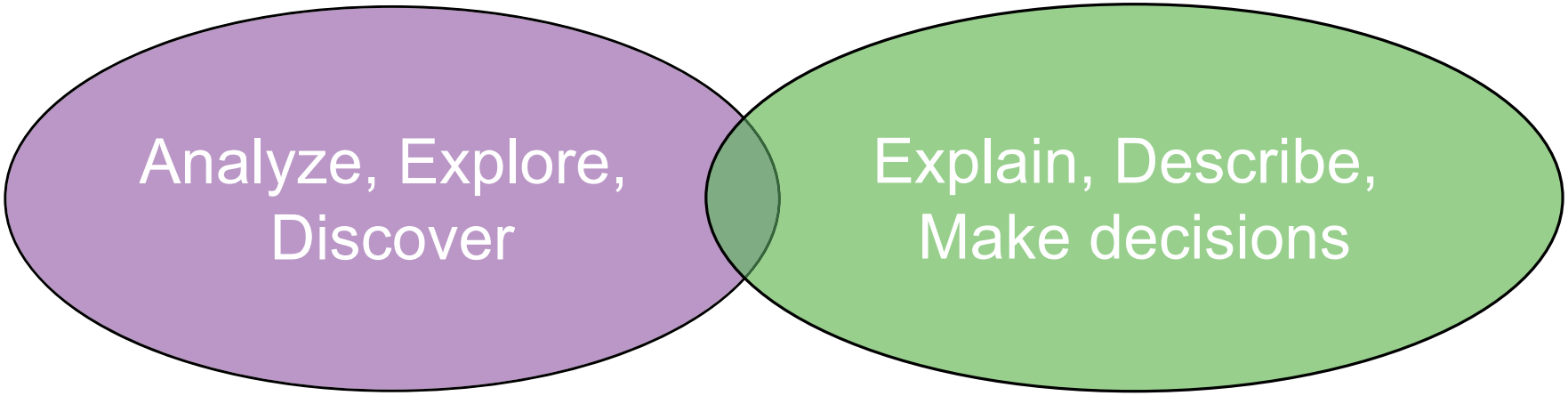


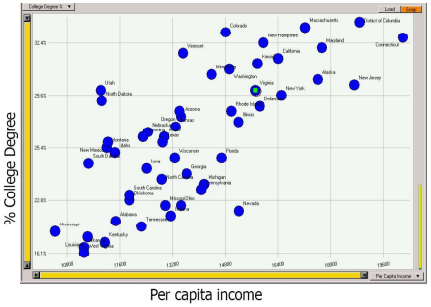
1. Infovis is perfect for exploration, when we don't know exactly what to look at. It supports vague goals

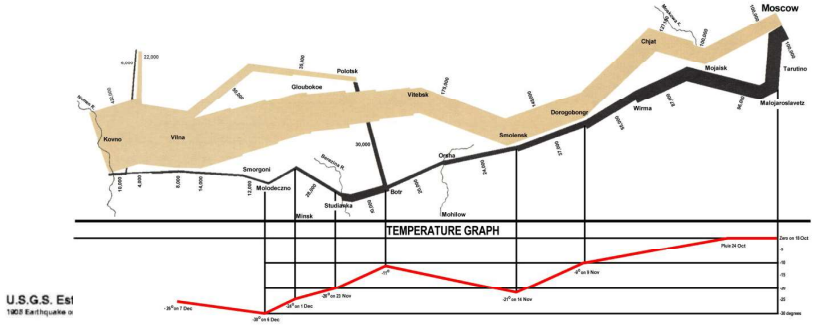
2. Infovis is very good in explaining complex data and to support decisions

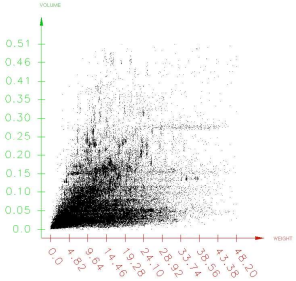
- Other approaches to data analysis
  - Statistics: strong verification but does not support exploration and vague goals (**and remember the Anscombes' quartet !**)
  - Data mining: actionable and reliable but black box, not interactive, question-response style
  - **Visual analytics** (formerly Visual Data Mining) is trying to join the two worlds

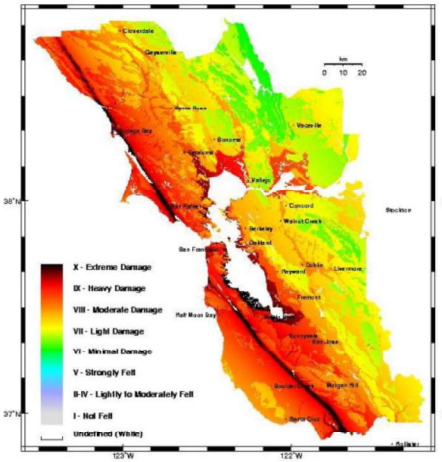
Visualization: Two Primary Goals





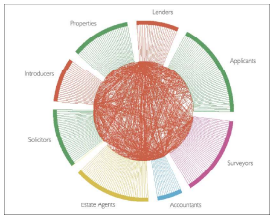


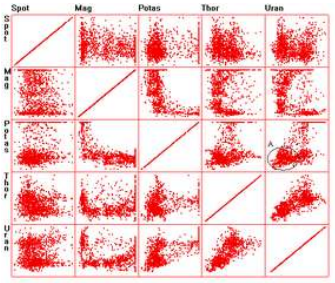


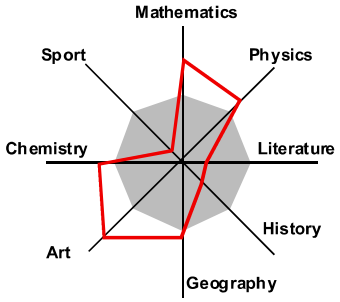


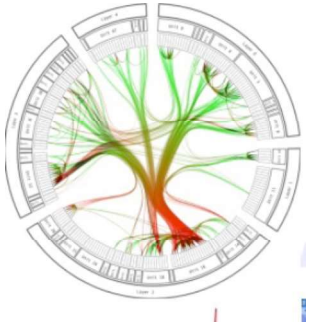
Canonical steps in infovis – STEP 1

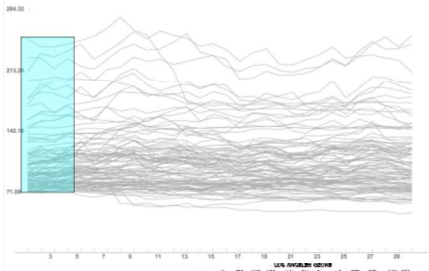
DATA → Internal **Representation**

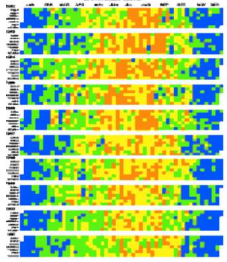


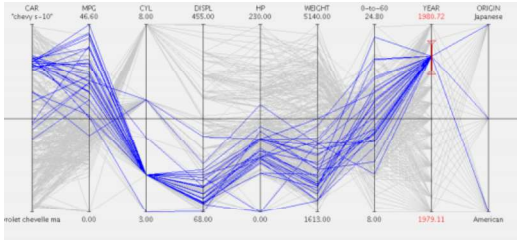


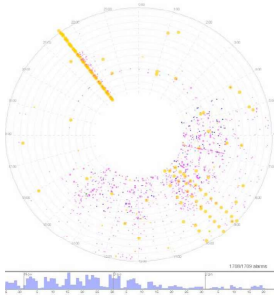


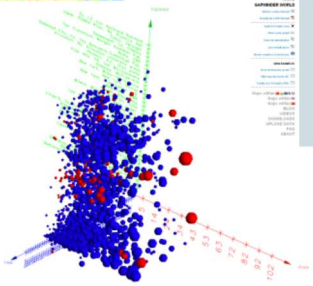


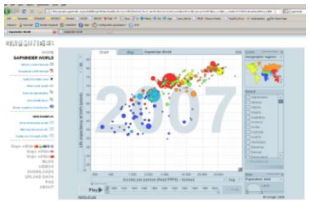


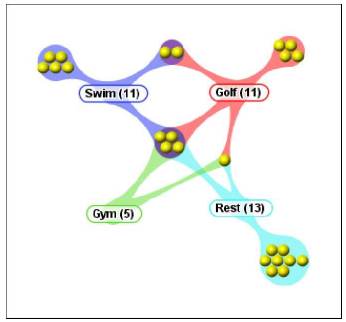












Encoding of values
- Univariate data
- Bivariate data
- Trivariate data
- Multidimensional data

Encoding of relations

Temporal data

Map & Diagrams

Graphs/Trees

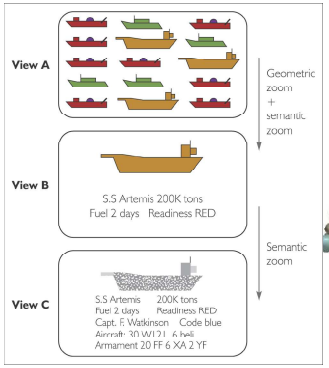
Data streams

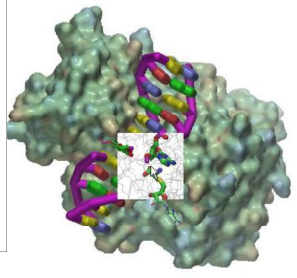
Canonical steps in infovis – STEP 2

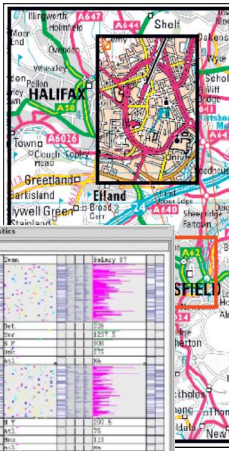
Internal **Representation**

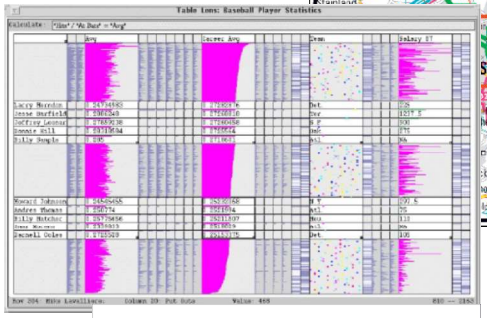
Presentation

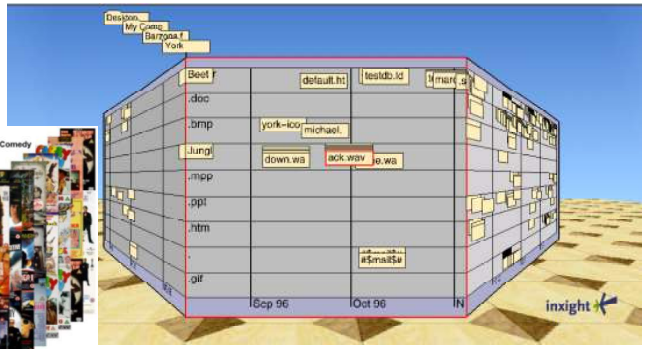
- Space limitations
  - Scrolling
  - Overview + details
  - Distortion
  - Suppression
  - Zoom & pan
  - Semantic zoom
- Time limitation
- Perceptual issues
- Cognitive issues

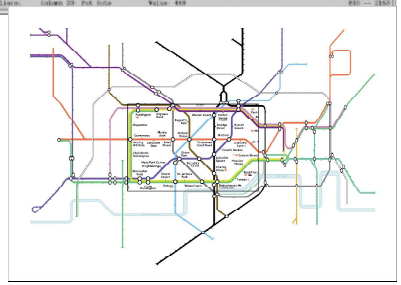












Problem solved!

We have (~) agreed and (~) mature solutions for

Presentation

Representation

of a large variety of data

So the problem seems solved

But…

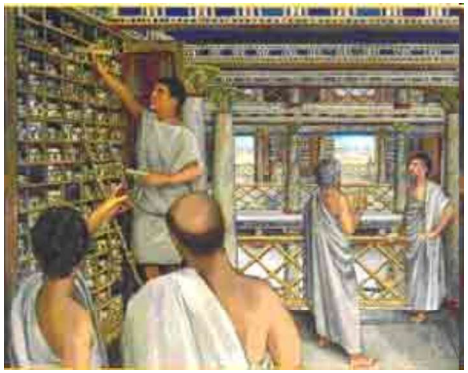
Data size and complexity !

- 100 million FedEx transactions per day
- 150 mi r day
- 300 mi
- 50 billi
- 600 bil
- 1 trillio oogle), corresp
- Google day

kilobyte, megabyte, gigabyte, terabyte, petabyte …

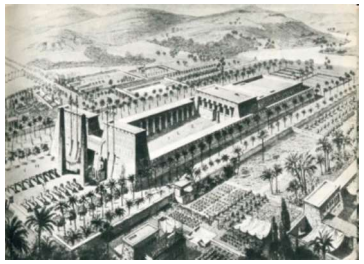
A petabyte ?

- 1 petabyte = 2^{50} = ~ 10^{15} (1000 trillions)
- How big is it?
- A quick comparison with one of the worst data loss in the story: the Alexandria library fire (270 a.d. ?)
- How many bytes were lost?
  - We are neglecting the quality...



Lost data

- Averaging data reported by historical writers we can assume about 50.000 lost books (~2^{16})



- Assuming each book size as Dante's Divina Commedia, 500.000 chars (~2^{19}), it results in

$$2^{16} * 2^{19} = 2^{35} = \sim 32 \text{ gigabytes}$$

- 6% of my laptop hard disk

What is the nowadays situation?

- The new Bibliotheca Alexandrina stores millions of books (2^{20})

and, incidentally, a modern fire prevention system...

- There are about 90.000 libraries in Europe and about 250.000 (~2^{18}) in the world:

libraries * books * chars =

$2^{18} * 2^{20} * 2^{19} = 2^{57} = 2^{7} * 2^{50}$ = ~128 petabytes

- For the sake of simplicity and removing duplicates (how many copies of Divina Commedia are around?) we can conclude the calculation to

~1 petabyte of chars

1 petabyte !

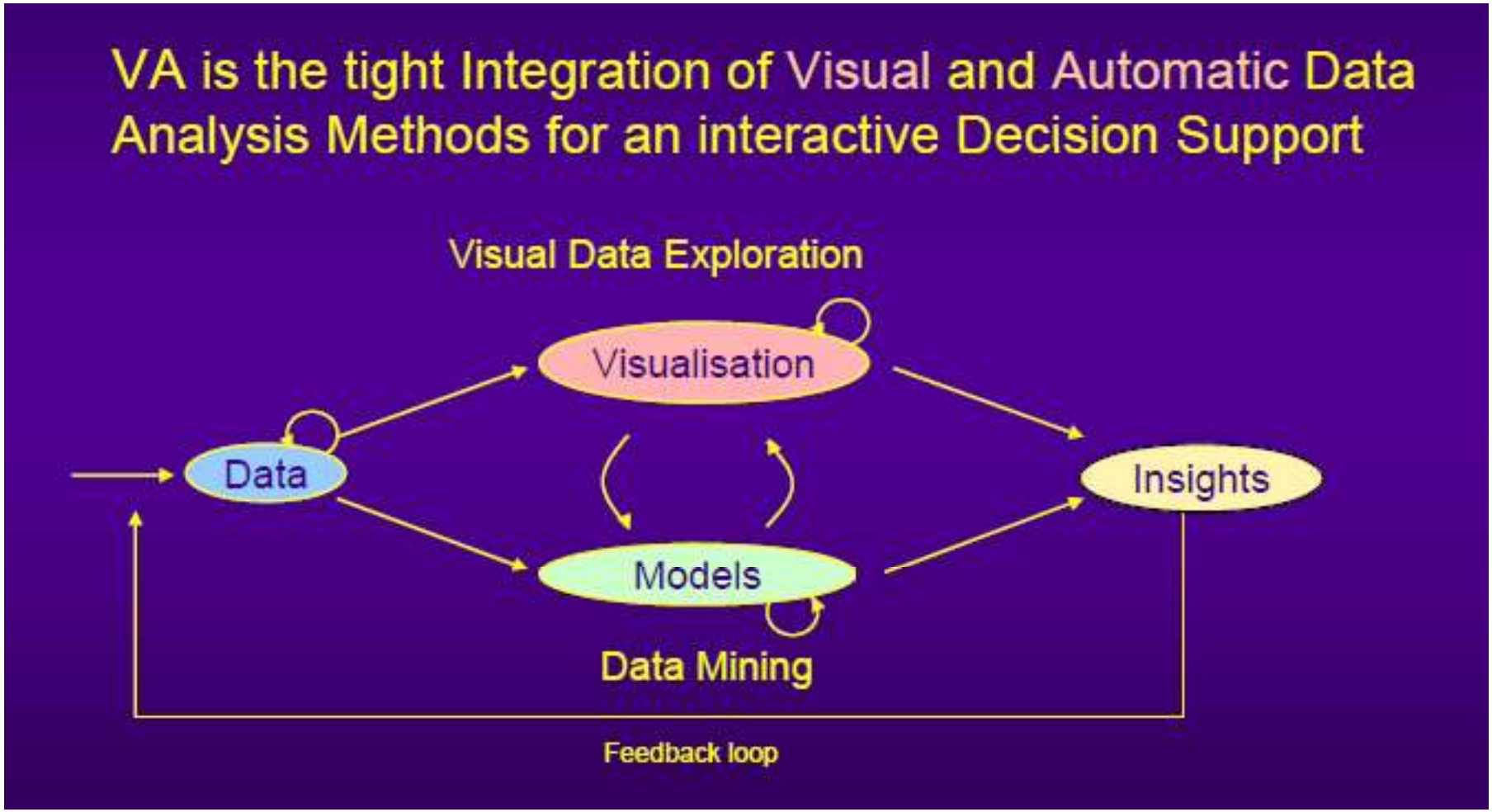
- **The entire written works of humankind, from the beginning of recorded history, in all languages...**

- Now we have a better intuition of what a petabyte is…

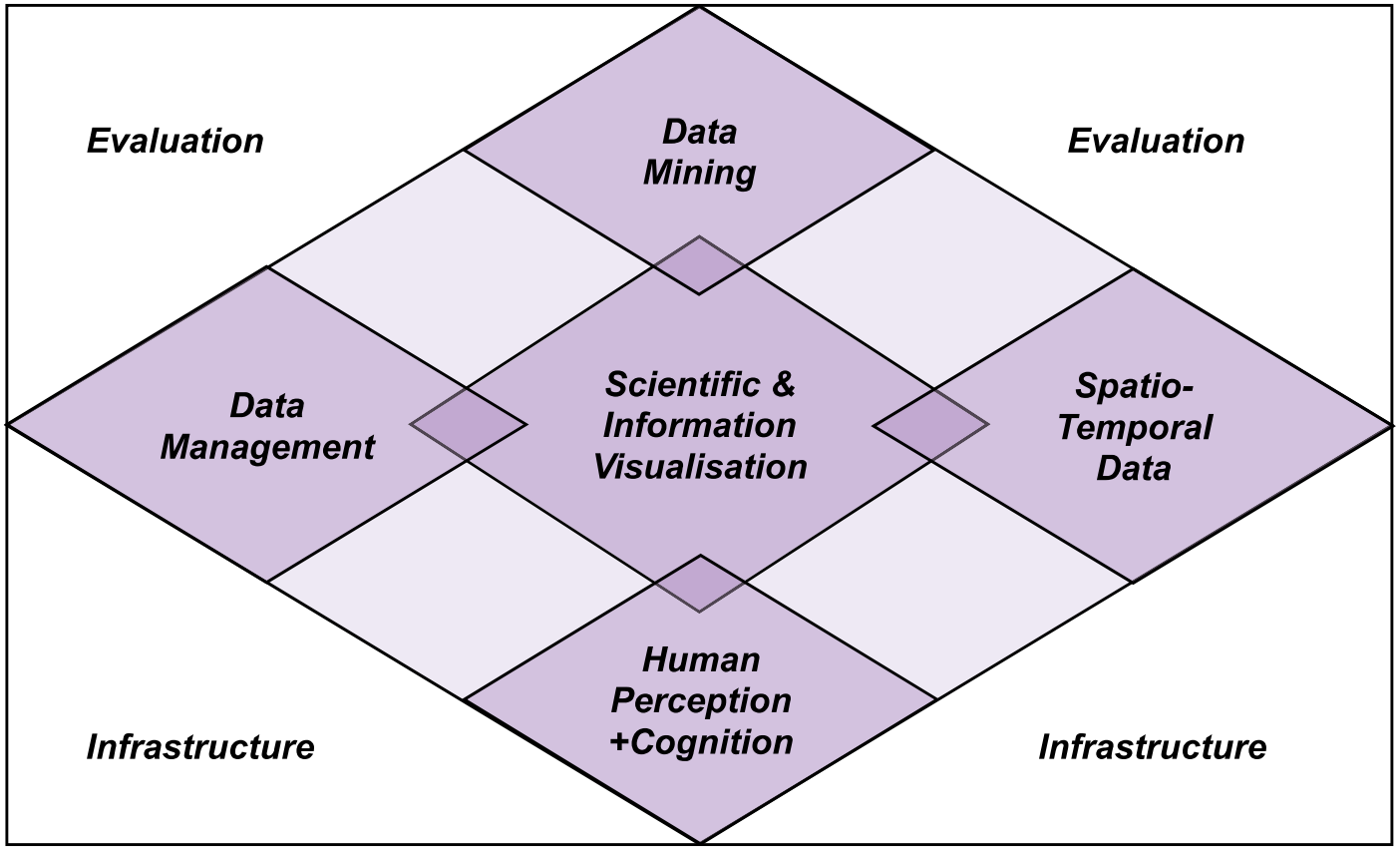
- What are the challenges of managing petabytes of data?
  - **Not** the **storage**
  - **Not** the **retrieval** (if you just need to retrieve a book)

- Challenges come from **effectively using** such immense wealth of data (without being overwhelmed). It means:
  - understanding it
  - discovering patterns, insights, and trends
  - making decisions

Visual Analytics



VA is highly interdisciplinary

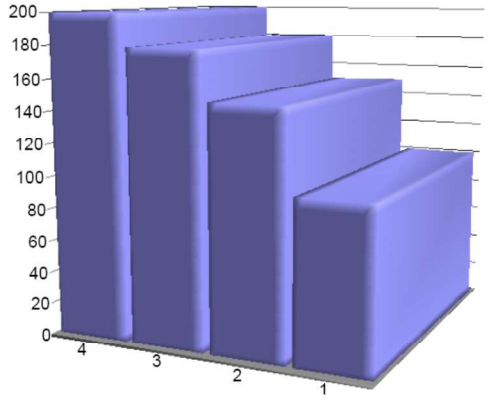


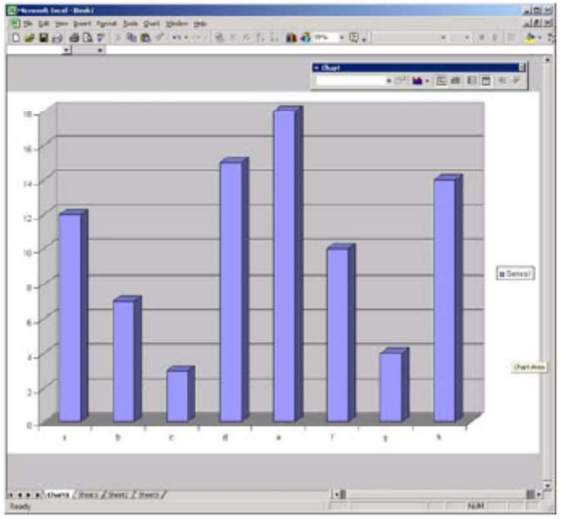
Each component presents challenging issues

Outline

- Facts about the course
- Historical examples
- Definitions
  - The Power of Information Visualization
  - Visual Analytics
- The problem and the involved issues

So, let's visualize





- Please, GET RID of those darn 3D bars !!!!!

So, let's visualize

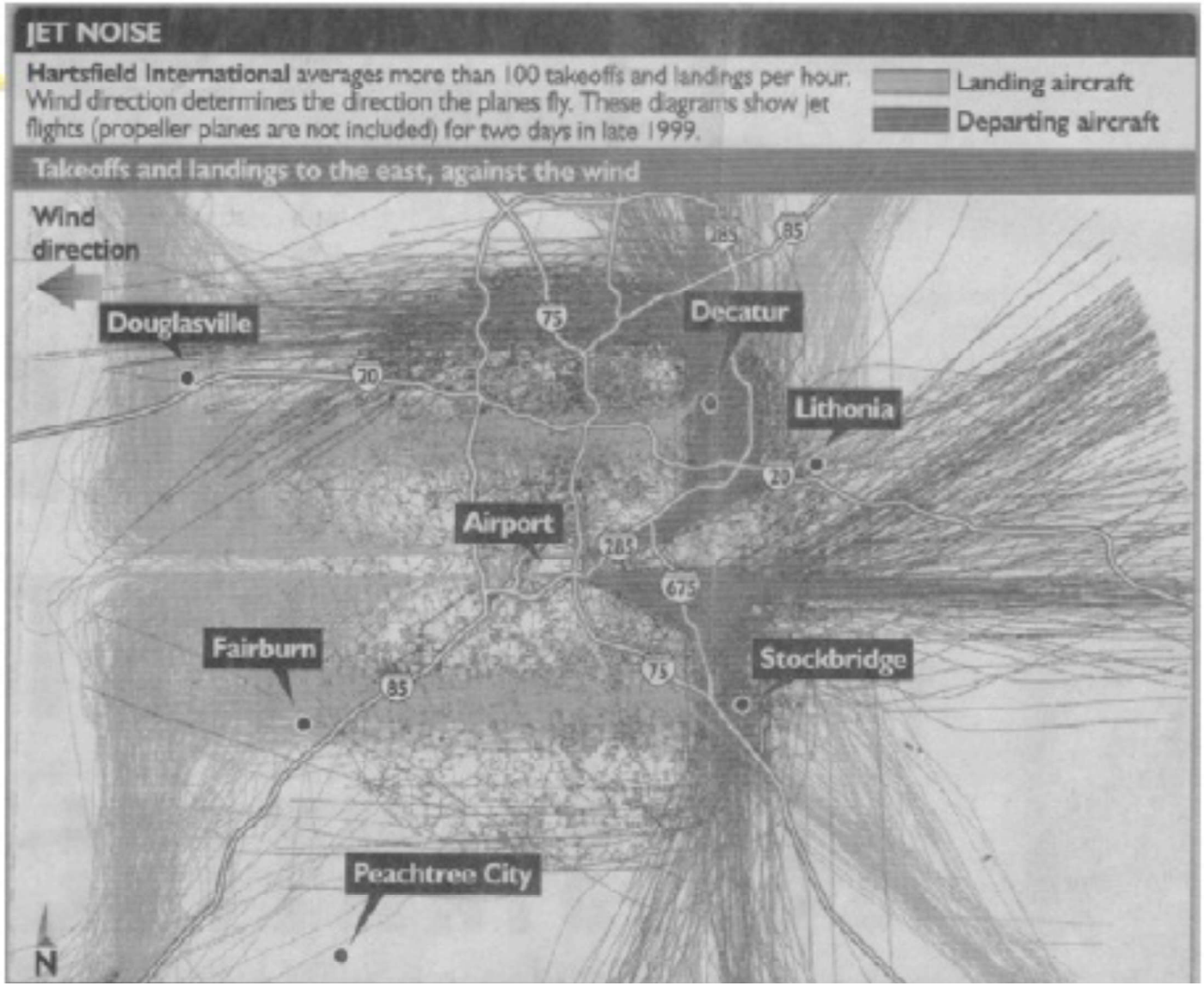
- Or worse yet…

So, let's visualize

- Make a simple decision here …

So, let's visualize

Atlanta Flight Traffic



Atlanta Journal
April 30, 2000

...

So, let's visualize

Power Costs

Average cost per month to use

Wall Street Journal
August 16, 2001

- …

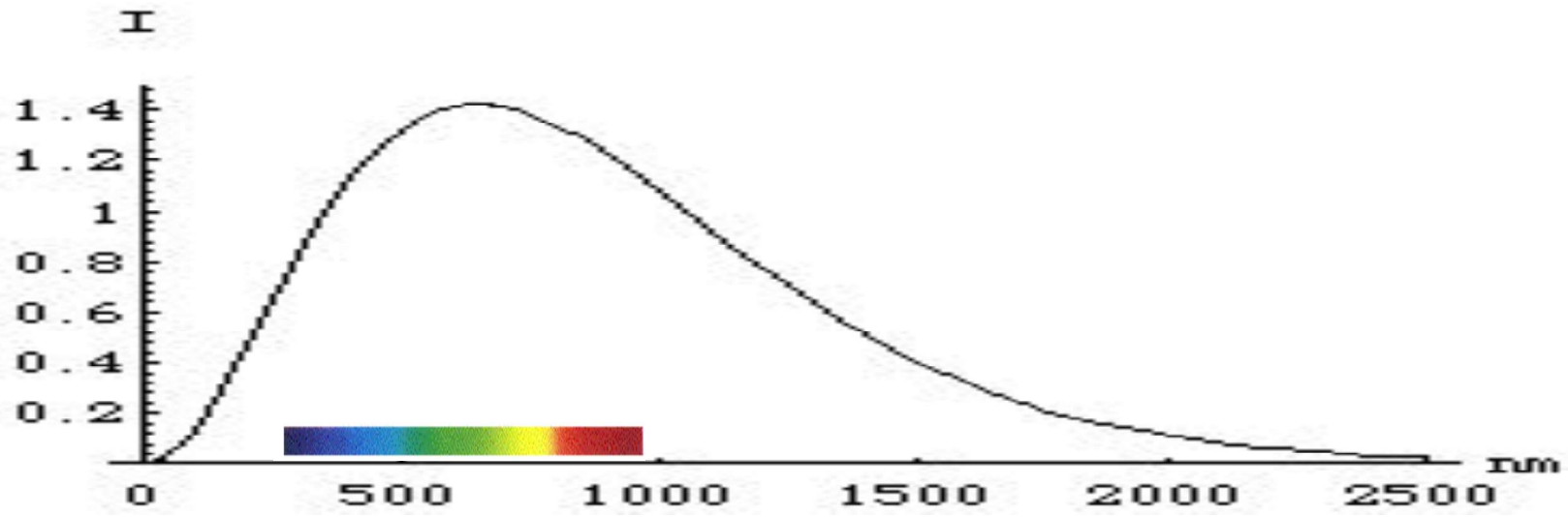
Visualization is not enough

- Perceptual, cognitive issues

- Why a green laser is so strong ?
  - Big battery?
  - More power ?

Visualization is not enough

- ?
- Human eyes are more sensible to green (555 nm) than low-red and hi-blue !
- More means 100 times!!!!



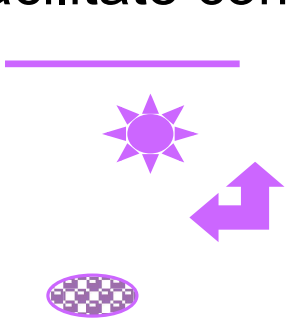
Human Perceptual Facilities

Use the eye for pattern recognition; people are good at

- scanning
- recognizing
- remembering images

Graphical elements facilitate comparisons via

- length
- shape
- orientation
- texture



Animation shows changes across time

Color and other pre-attentive features helps make distinctions

But … How many colors can human eyes distinguish in pre-attentive way? (about 6 / 12 ☹ …)

Focusing problem / effect

Most people see the red

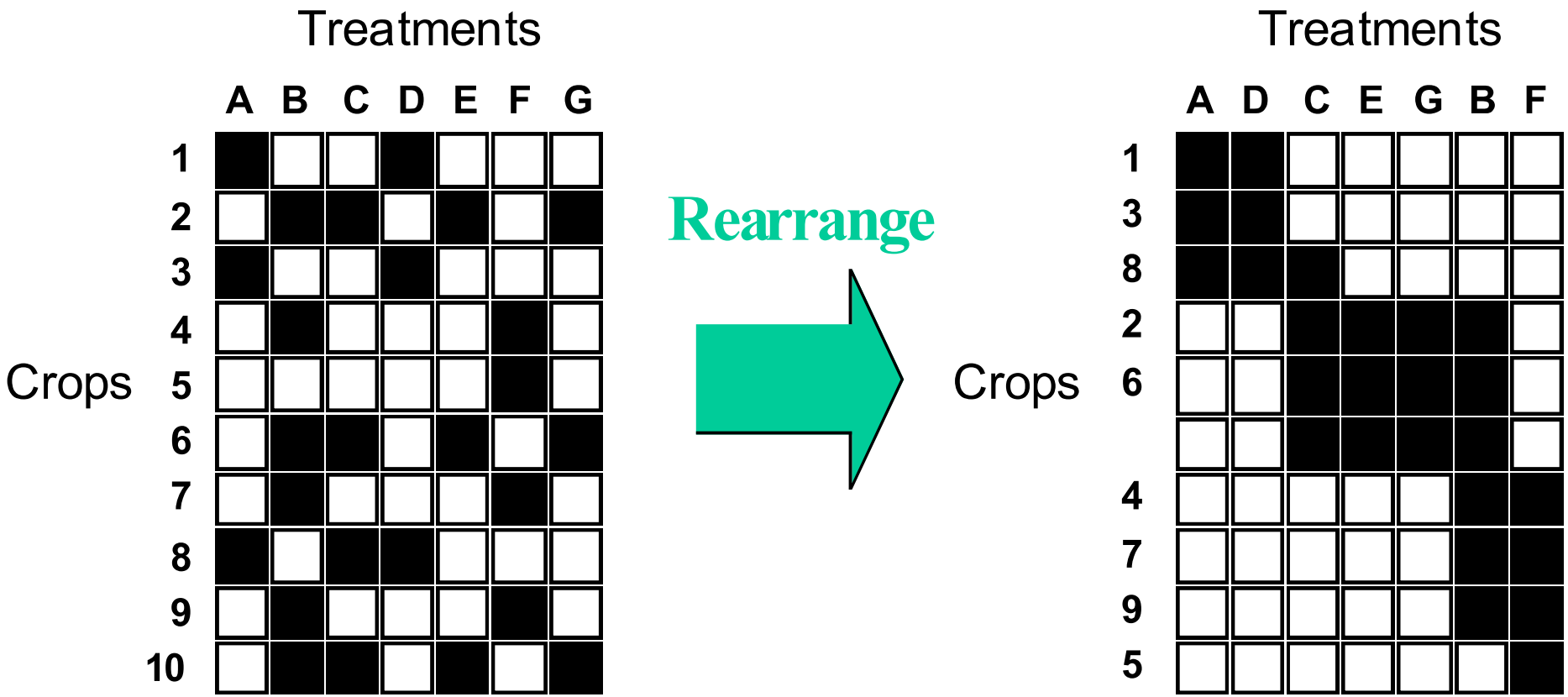
Closer than the blue

But some see the

Opposite effect

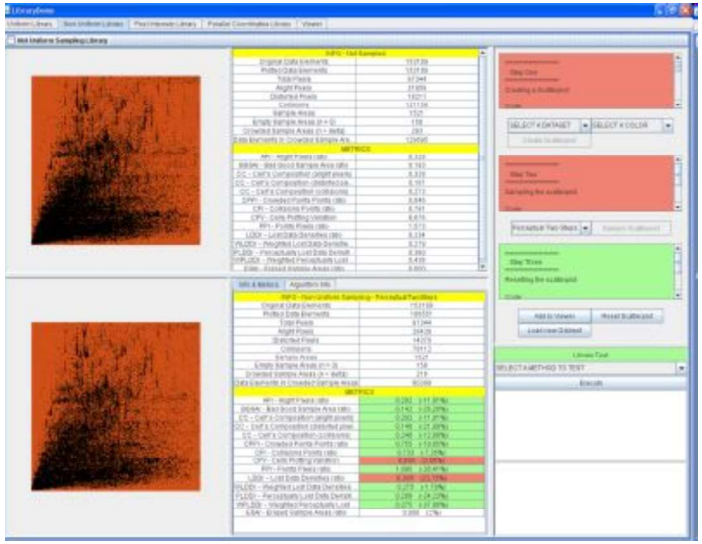
Visualization is not enough

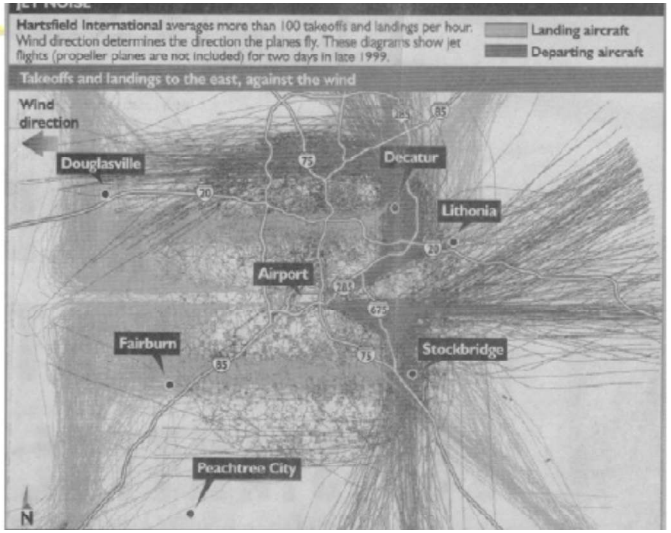
- Interaction is not a plus !
- And algorithms as well



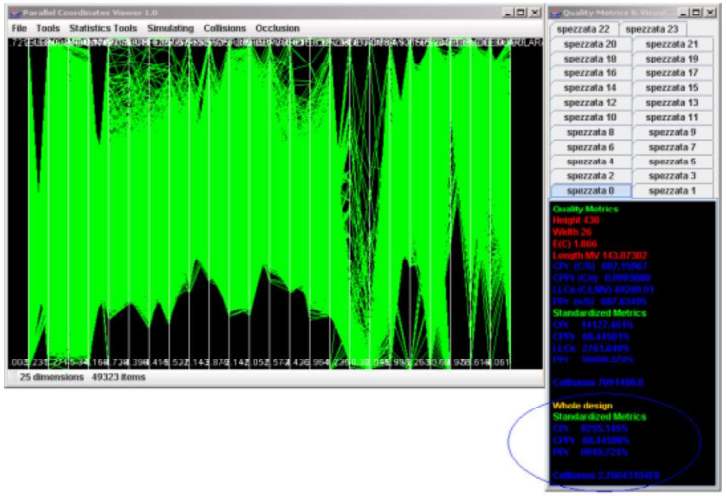
Visualization is not enough

- Data are complex (a lot of attributes) and dataset can be very large !
- How to manage complexity?

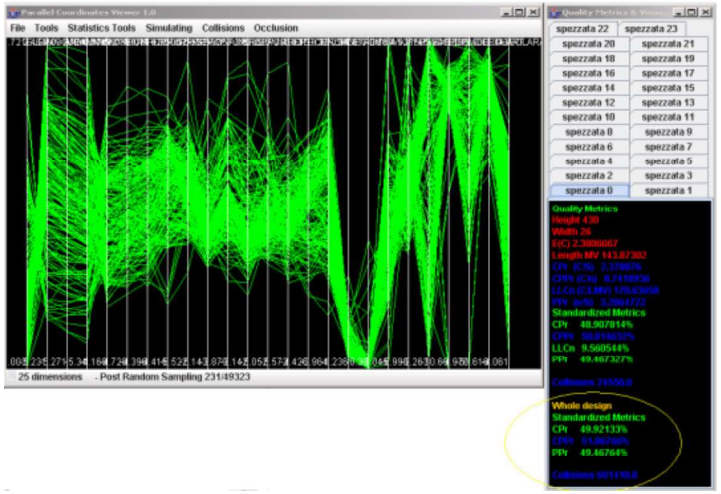




Rappresentazione iniziale



Rappresentazione dopo il campionamento guidato mediante la metrica



User tasks

- User task (and skill) must be considered !
- Learning time: visualizations are not simple!!!!!
- Evaluation of visual system is not trivial
- Answer this question: Do you know the answer?
  - If yes,
    - Presentation, communication, education
  - If no,
    - Exploration, analysis
    - Problem solving, planning,
    - Aid to thinking, reasoning

Visualization is not enough

- How develop the visualization ?
- From scratch or using an infovis toolkit ? E.g., d3.js !

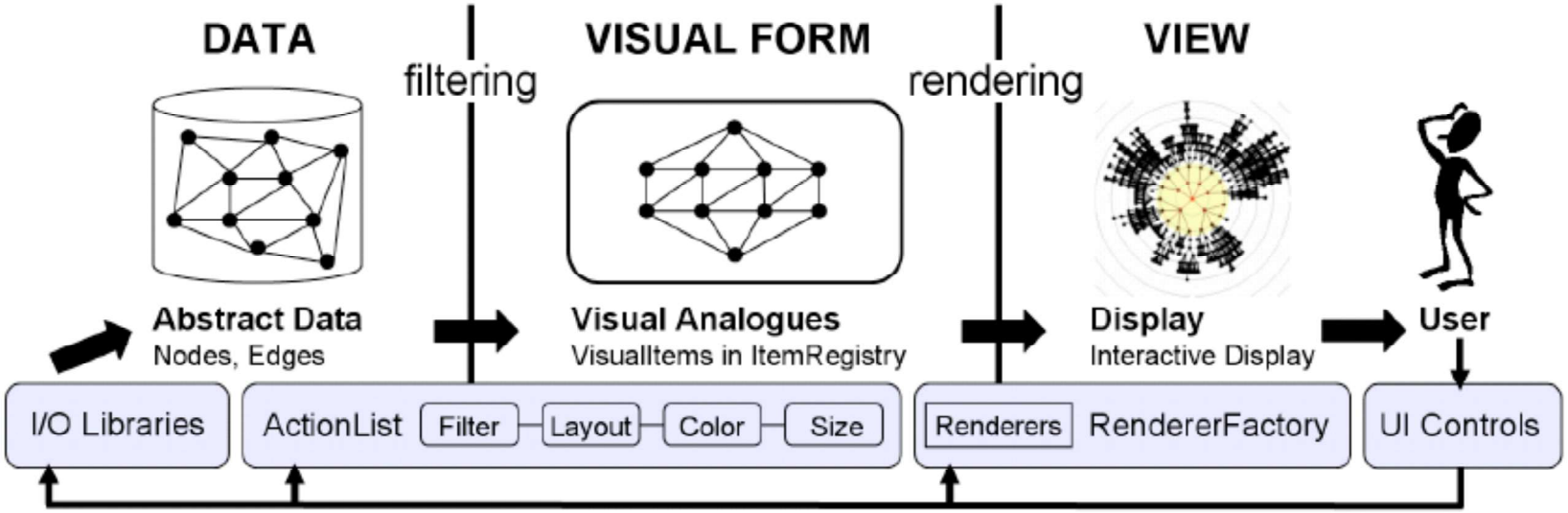


Figure 2. The prefuse visualization framework. Lists of composable actions filter abstract data into visualizable content and assign visual properties (position, color, size, font, *etc*). Renderer modules, provided on a per-item basis by a RendererFactory, draw the VisualItems to construct interactive Displays. User interaction can then trigger changes at any point in the framework.