Number Visualization

(and Information Visualization and Visual Analytics as well)

Giuseppe Santucci

University of Rome "La Sapienza"
santucci@dis.uniroma1.it

Thanks to:
Ross Ihaka (very inspiring lectures)

Number visualization ?

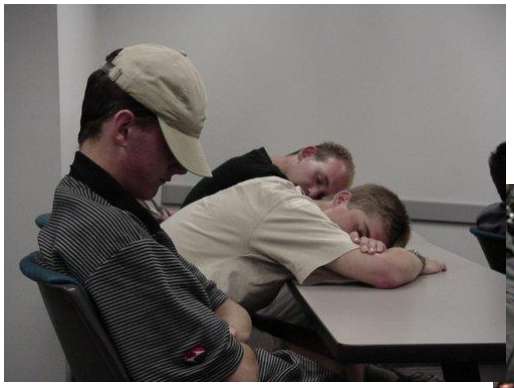
- Obviously information visualization is, in general, about numbers
- In some cases, however, the number part is relevant, and the use of tables and graphs to communicate **quantitative** information is commonplace in business today (pie chart, diagrams, boxplots, etc.)

- Actual software applications allows for easy (?) development of different typologies of charts
- I will discuss the basic relationships and the logical steps that allows for moving from quantitative data to suitable visualizations

Types of Data

- Quantitative (allows arithmetic operations)
  - *123, 29.56, …*
- Categorical (group, identify & organize; no arithmetic)
  - Nominal (name only, no ordering)
    - *Direction: North, East, South, West*
  - Ordinal (ordered, not measurable)
    - *First, second, third …*
    - *Hot, warm, cold*
  - Interval (starts out as quantitative, but it is made categorical by subdividing into ordered ranges)
    - *0-999, 1000-4999, 5000-9999, 10000-19999, …*
  - Hierarchical (successive inclusion)
    - *Region: Continent > Country > State > City*
    - *Animal > Mammal > Horse*
- Relationships
  - Correlation
  - ....

uhmmm...



- Boring ?
- I do agree !
- I changed my mind !
- It is plenty of books that teach about quantitative data and how to show it (see references)
- Read all of them! I'll go for another way...

Outline
(basically what you have **NOT** to do)

- An introductive example
- Good and bad graphs
  - Basic rules
  - Some additional considerations
- Visual issues
  - Quantitative perception (basic rules)
- Information Visualization

A lotto game

- Forms of lotto are played world-wide and many people have theories about how to make money at the game
- We will examine a particular lotto game, to see whether it might be possible to play it profitably
- The game we'll look at is the daily pick-it lottery run by the state of New Jersey in the USA

Lotto rules

- Each player selects a number between 000 and 999
- A winning number is selected by independently picking three digits between 0 and 9 at random
- All players that hold the winning number split the prize money for the game

Available data

- The results of the games (winning number and winning amount) are publicly available
- Does this data contain information which will enable us to choose a profitable strategy for this game?
- We will use the results of 254 consecutive games to look for a profitable strategy

The data (254 values)

(winning number, winning amount)

- (810, $190.0), (156, $120.5), (140, $285.5), (542, $184.0), (507, $384.5),
- (972, $324.5), (431, $114.0), (981, $506.5), (865, $290.0), (499, $869.5),
- (020, $668.5), (123, $83.0), (356, $188.0), (015, $449.0), (011, $289.5),
- (160, $212.0), (507, $466.0), (779, $548.5), (286, $260.0), (268, $300.5),
- (698, $556.5), (640, $371.5), (136, $112.5), (854, $254.5), (069, $368.0),
- (199, $510.0), (413, $102.0), (192, $206.5), (602, $261.5), (987, $361.0),
- (112, $167.5), (245, $187.0), (174, $146.5), (913, $205.0), (828, $348.5),
- (539, $283.5), (434, $447.0), (357, $102.5), (178, $219.0), (198, $292.5),
- (406, $343.0), (079, $332.5), (034, $532.5), (089, $445.5), (257, $127.0),
- (662, $557.5), (524, $203.5), (809, $373.5), (527, $142.0), (257, $230.5),
- (008, $482.5), (446, $512.5), (440, $330.0), (781, $273.0), (615, $171.0),
- (231, $178.0), (580, $463.5), (987, $476.0), (391, $290.0), (267, $176.0),
- (808, $195.0), (258, $159.5), (479, $296.0), (516, $177.5), (964, $406.0),
- (742, $182.0), (537, $164.5), (275, $137.0), (112, $191.0), (230, $298.0),
- (310, $110.0), (335, $353.0), (238, $192.5), (294, $308.5), (854, $287.0),
- (309, $203.5), (026, $377.5), (960, $211.5), (200, $342.0), (604, $259.0),
- (841, $231.0), (659, $348.0), (735, $159.0), (105, $130.5), (254, $176.0),
- (117, $128.5), (751, $159.0), (781, $290.0), (937, $335.0), (020, $514.0),
- (348, $191.0), (653, $304.5), (410, $167.0), (468, $257.0), (077, $640.0),
- (921, $142.0), (314, $146.0), (683, $356.0), (000, $96.0), (963, $295.0),

Visualizing the data

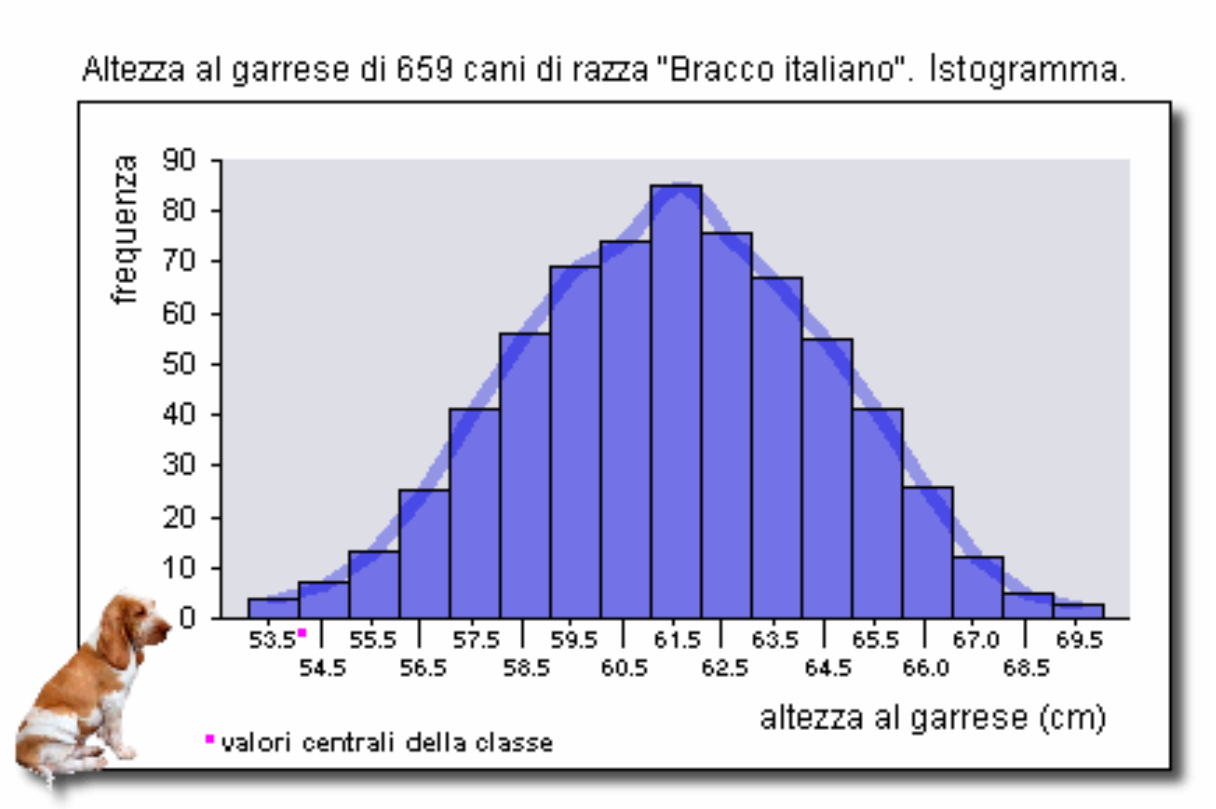
- Humans can really only make sense of three or four numbers at a time
- By representing the values in a graphical form we make it easier to handle large numbers of values
- Using visualizations should make it possible to learn more about this data

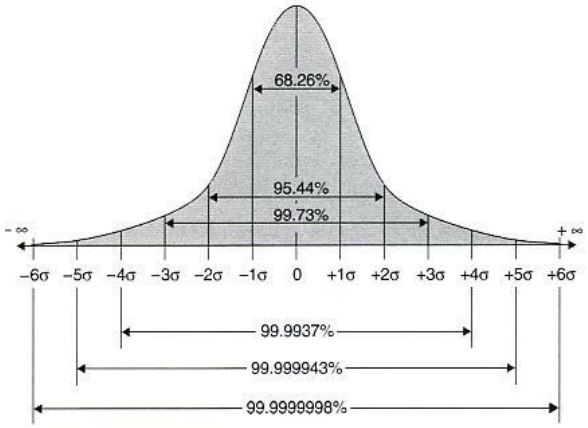
- We have NOT to **lie** or make **noise** !!!

User task and visualization

- One approach to making money at “Pick It” is to try to select numbers which are more likely to win
- We can look at the distribution of the winning numbers to see whether some ranges of values are more like to produce a winner than others
- One way to do this is to produce a histogram of the winning numbers

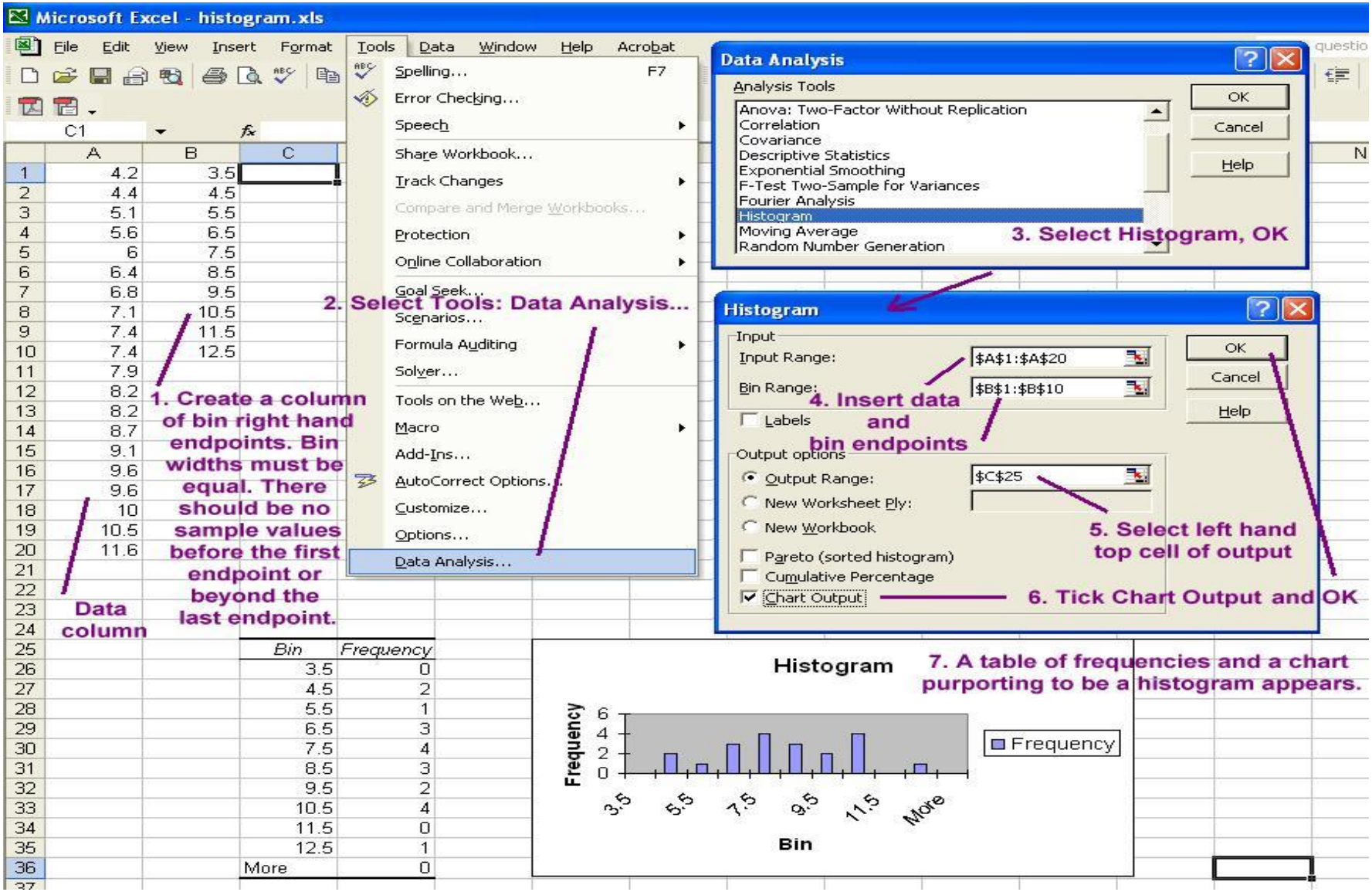
Histogram example



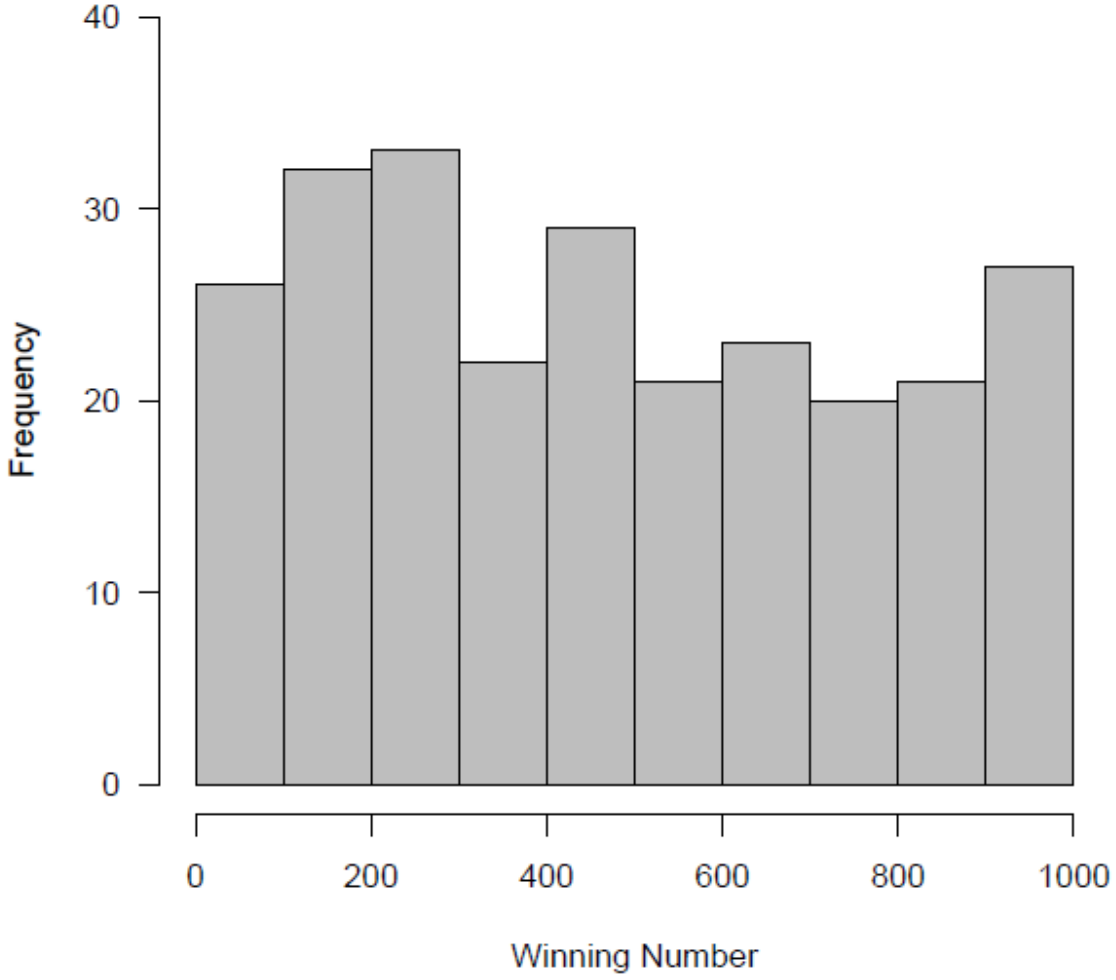


bin

Excel and histograms



Data distribution

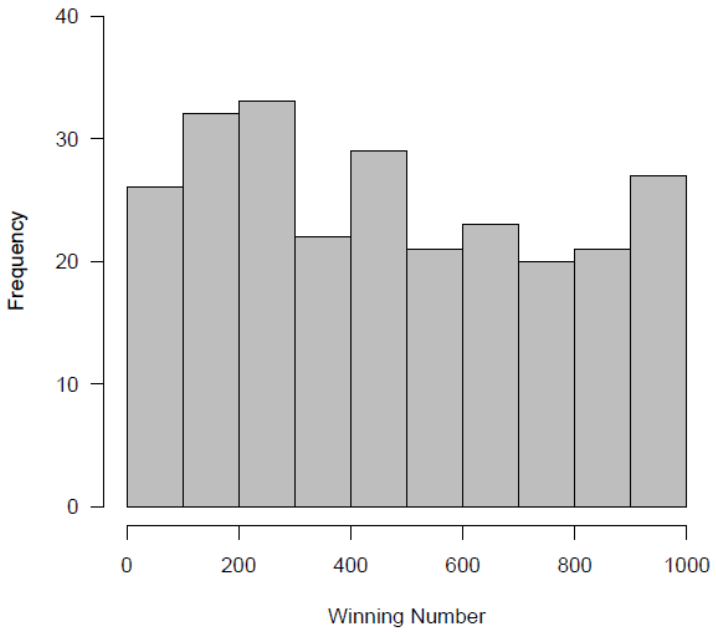


Is the bin size ok?

Analysis

- It looks there tend to be more winners in the region from 100 to 300 than in other regions

- This suggests that we might be best to choose numbers in this range

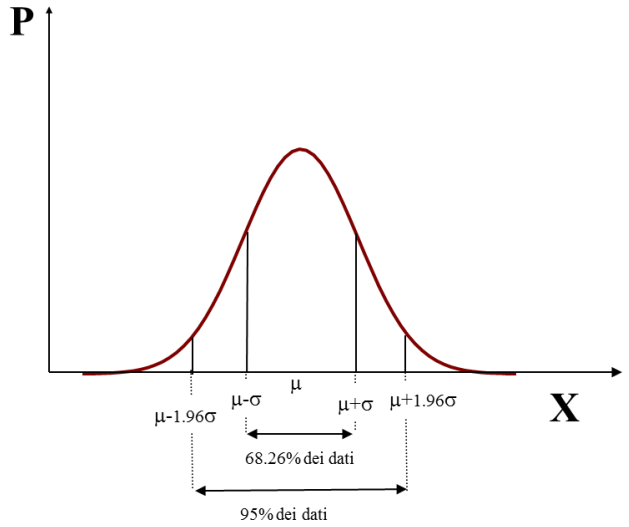


We are telling lies...
(wrong number understanding)

- Even if the winning numbers are chosen randomly we can expect some “random variability” among them
- To judge the significance of what we see in the histogram we have to recall some formal statistical theory

The mean is not enough !

- There are 254 values. We would expect the number of values in each cell to be approximately: 25.4 = 254/10
- Such a number **is a random variable** as well, with normal distribution
- 95% of the observations fall within +/- 2σ



Looking for new insights

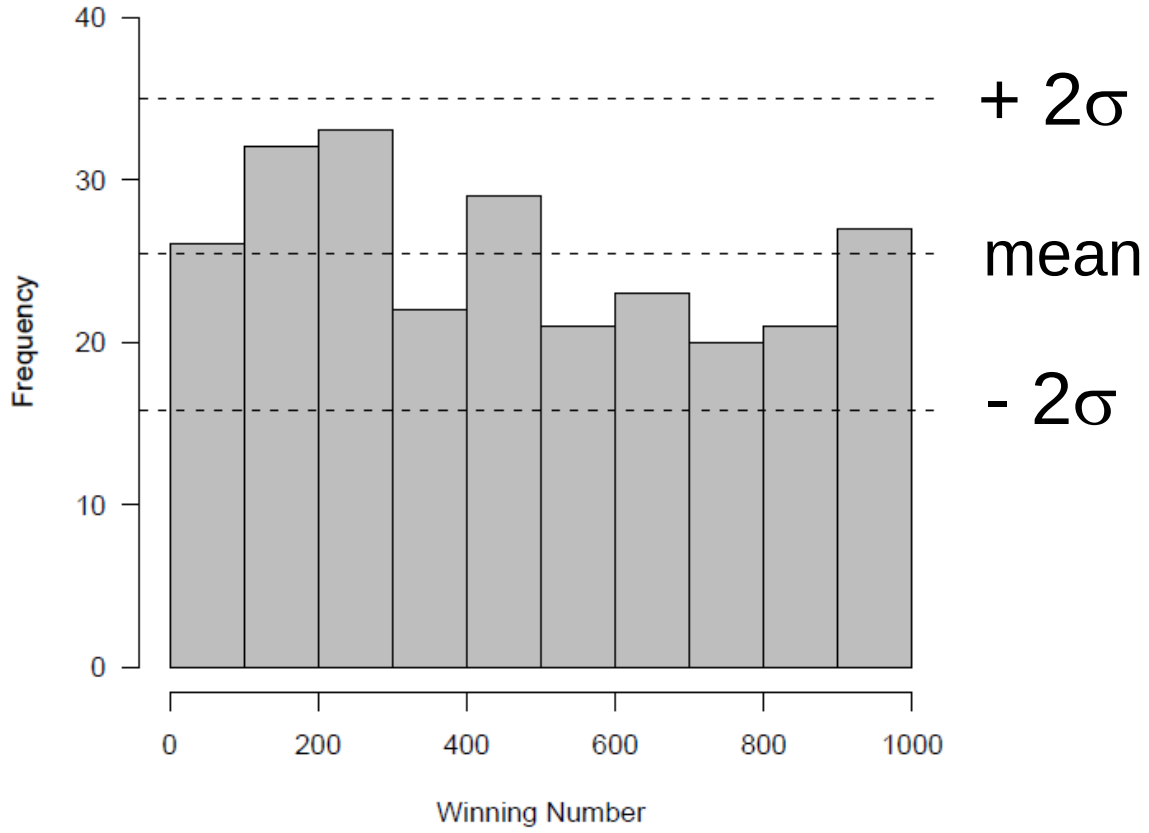
- The histogram shows that there is a wide (more than 2σ) range amounts won in the game

- It might be possible to choose the numbers which win larger amounts

- We search for relationship between ticket number and winning amount

- A scatter plot is the natural way to look for such a relationship.

Better number visualization

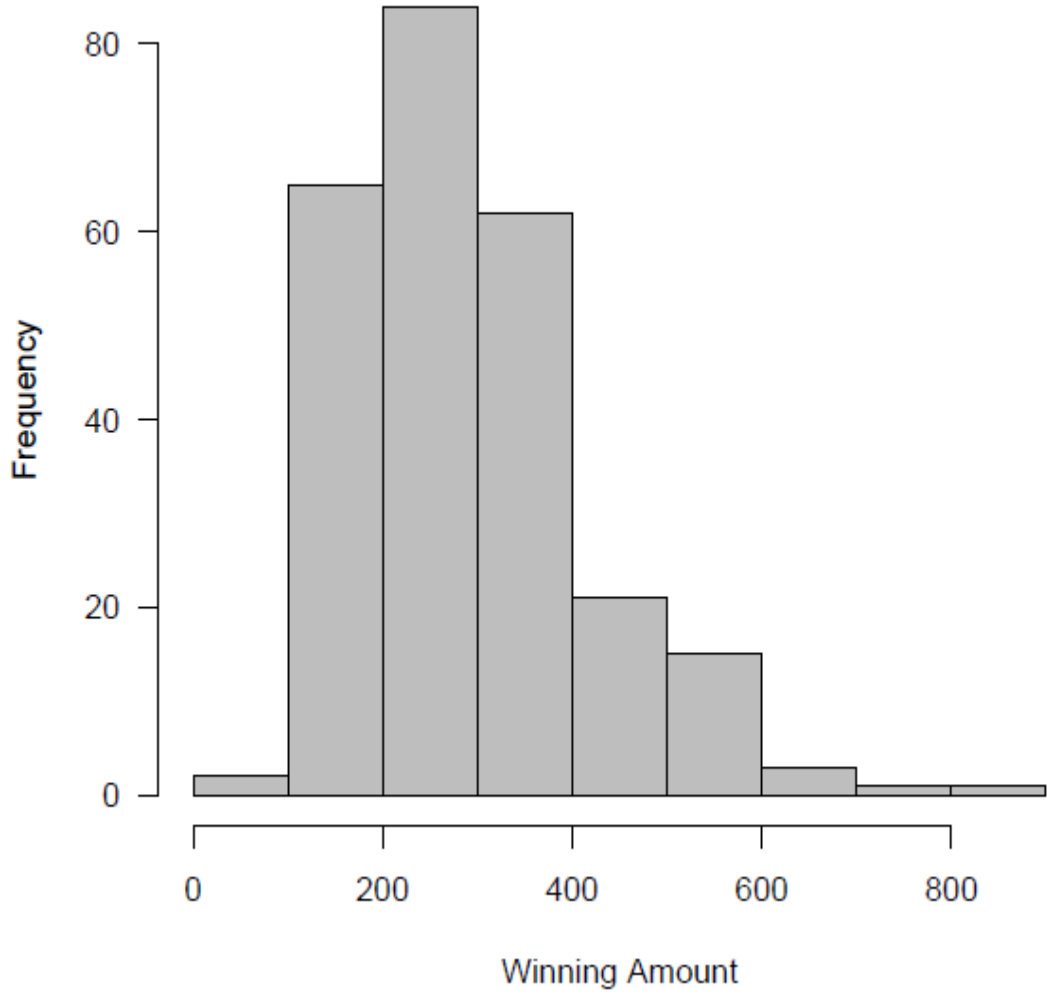


- Variance <u>analysis</u> AND <u>visualization</u>

Conclusions and new task

- Winning numbers are totally random
- It makes no sense to look for a " lucky " number
- We can change our task:
  - to increase the amount won !
- So we study the distribution of winning amount

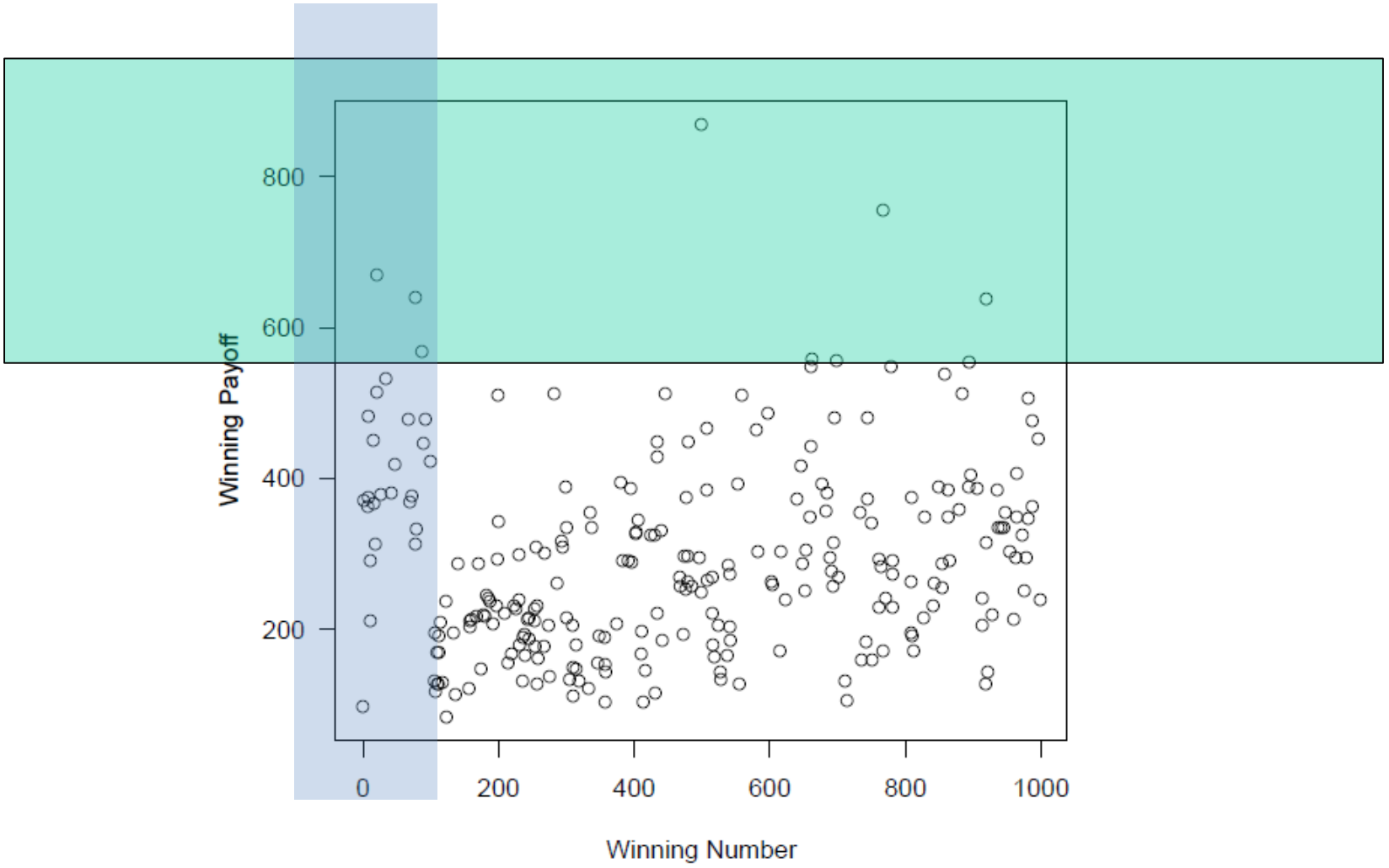
New visualization



Looking for new insights

- The histogram shows that there is a wide (more than 2σ) range amounts won in the game
- It *might* be possible to choose the numbers which win larger amounts
- We search for relationship between ticket number and winning amount
- A scatter plot is the natural way to look for such a relationship

New visualization

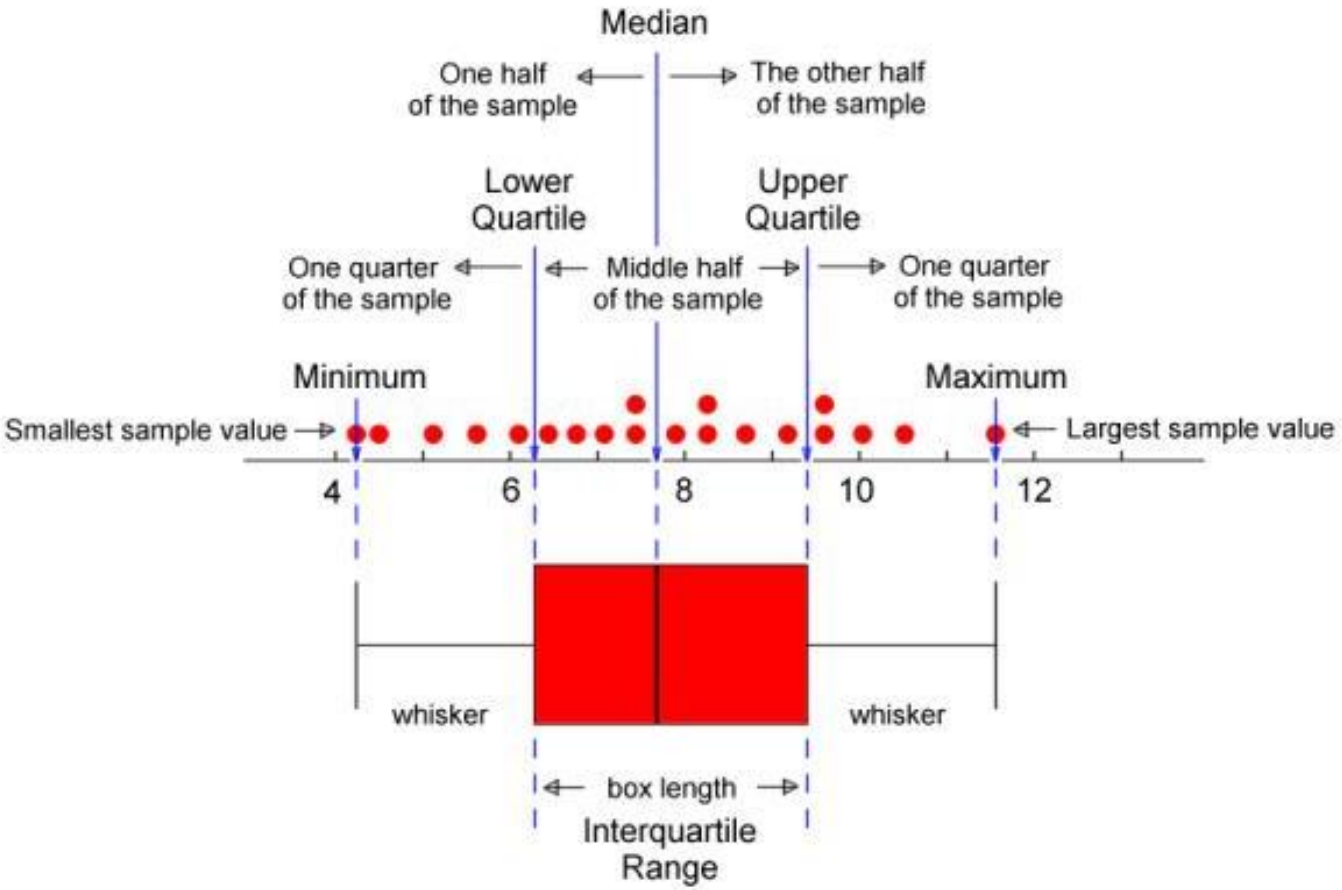


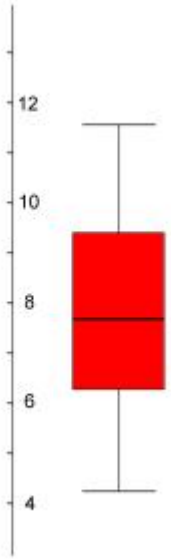
Insights from the scatterplot

- The winning amounts in a band to the left of the plot appear to generally be higher than those in the rest of the plot

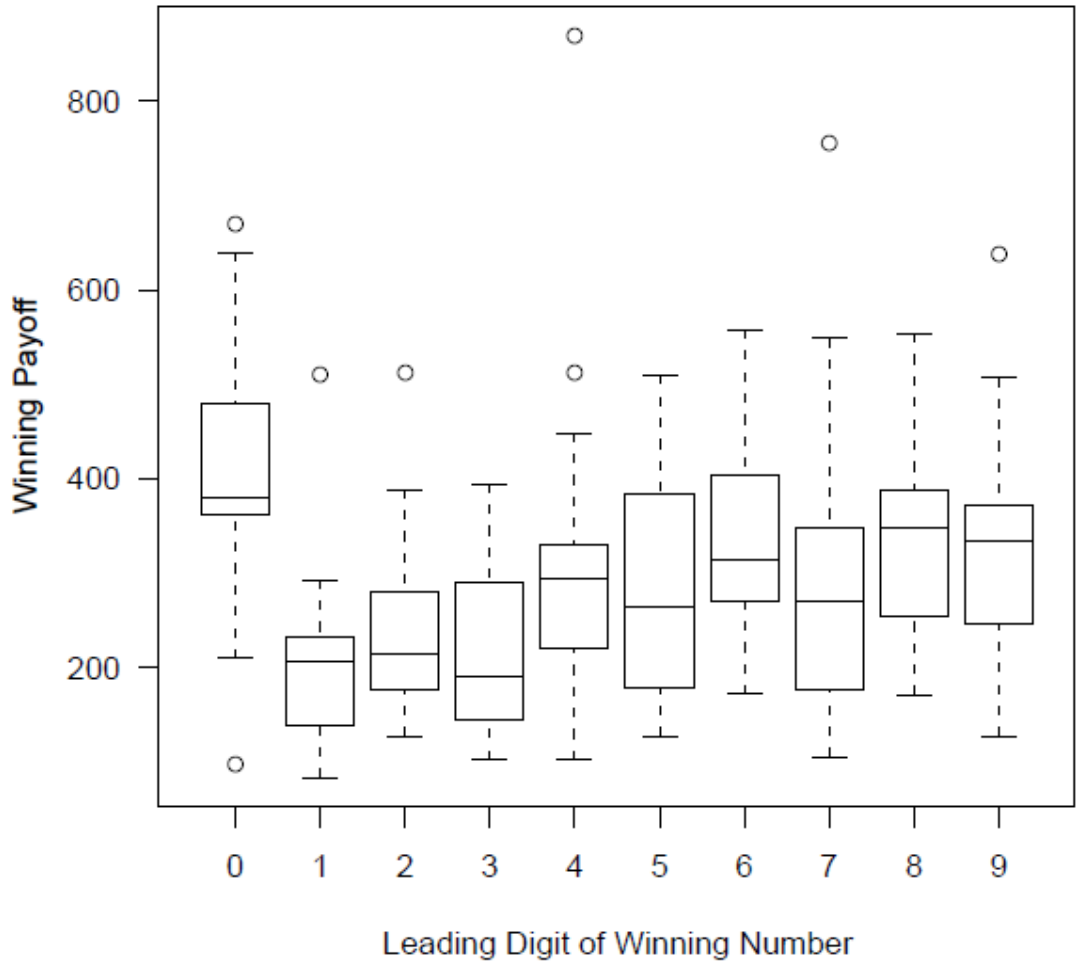
- We can investigate this further by separating the values into groups according to the first digit of the ticket number and drawing box plots for each group

Boxplot


Median
One half of the sample
The other half of the sample
Lower Quartile
Upper Quartile
One quarter of the sample
Middle half of the sample
One quarter of the sample
Minimum
Maximum
Smallest sample value
Largest sample value
4
6
8
10
12
whisker
whisker
box length
Interquartile Range


12
10
8
6
4

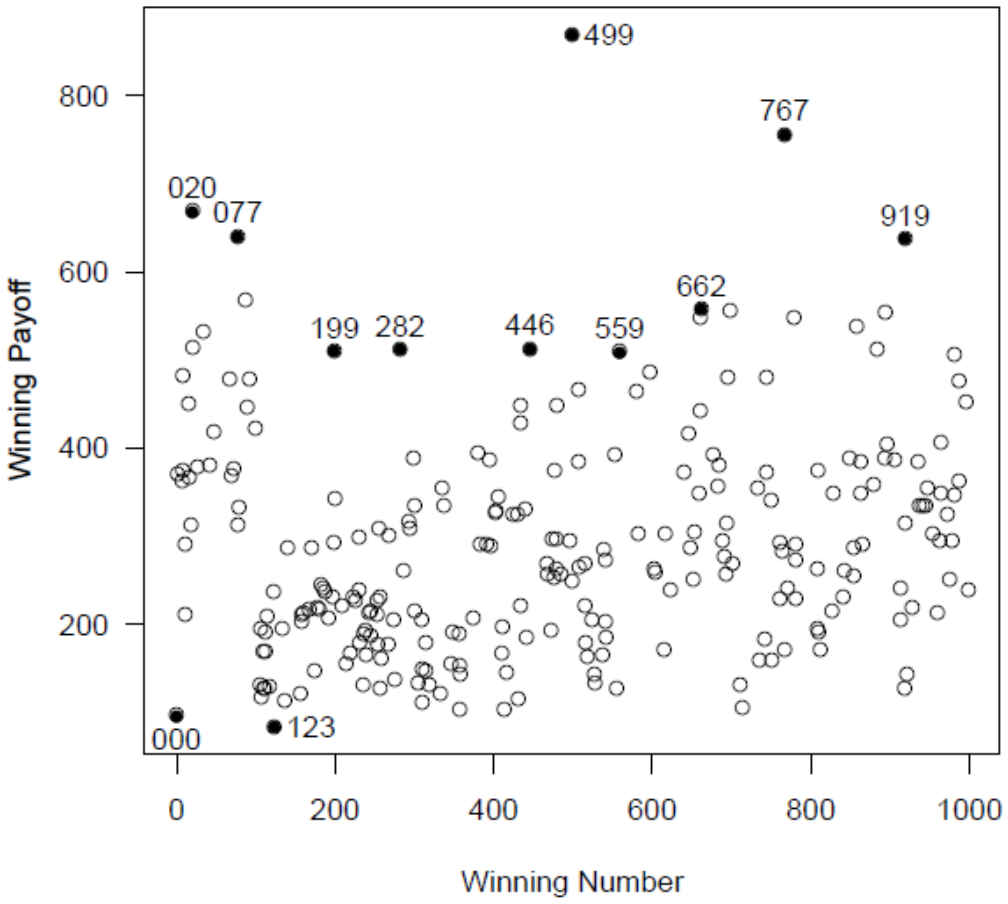
Lottery's boxplots



New insights

- Tickets with a leading zero digit clearly tend to produce larger winnings
- It is also apparent that there are some very large and some very small winning amounts
- It is probably of interest to identify the ticket numbers corresponding to these extremes

High and low winning numbers



Lotto strategy

- While winning numbers are non predictable, players' choices are!

- Choose numbers which are less likely to be chosen by other players
- Then, when you win (if), you will tend to win more
- Possible ways to choose:
  - Choose a number with a leading zero
  - Choose a number with repeated digits
  - Avoid “obvious” numbers like, e.g. 000, 123, 246, . . .

Lessons learned

- Define clearly the task
- Use basic visualizations
  - bar charts
  - scatterplots
  - boxplots
- Be ready to switch among them
- Look for precise values when needed
- Do not lie !

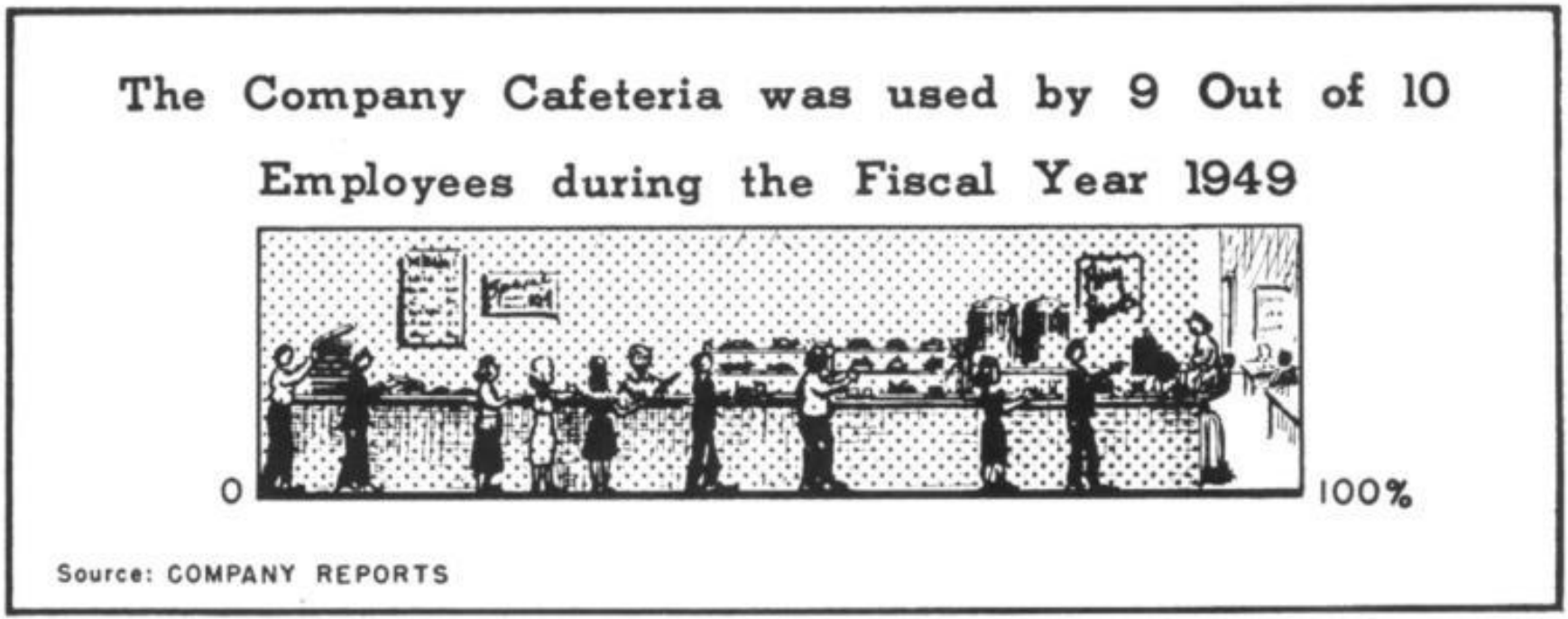
Outline
(basically what you have NOT to do)

- An introductive example
- Good and bad graphs
  - Basic rules
  - Some additional considerations
- Visual issues
  - Quantitative perception (basic rules)
- Information Visualization

Rule 0:
Do not use diagrams when handling few numbers

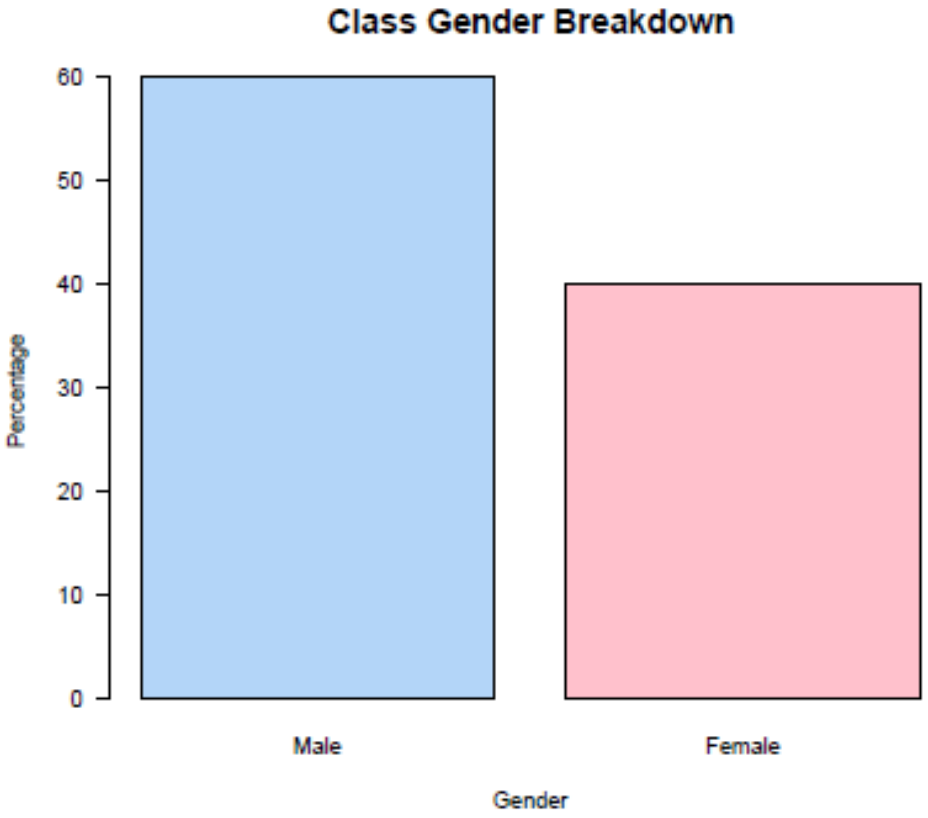
- It does not make sense to use graphs to display very small amounts of data
- The human brain is quite capable of grasping one two, or even three values

Rule 0 violation (and also rule 2)



90%

Rule 0 violation

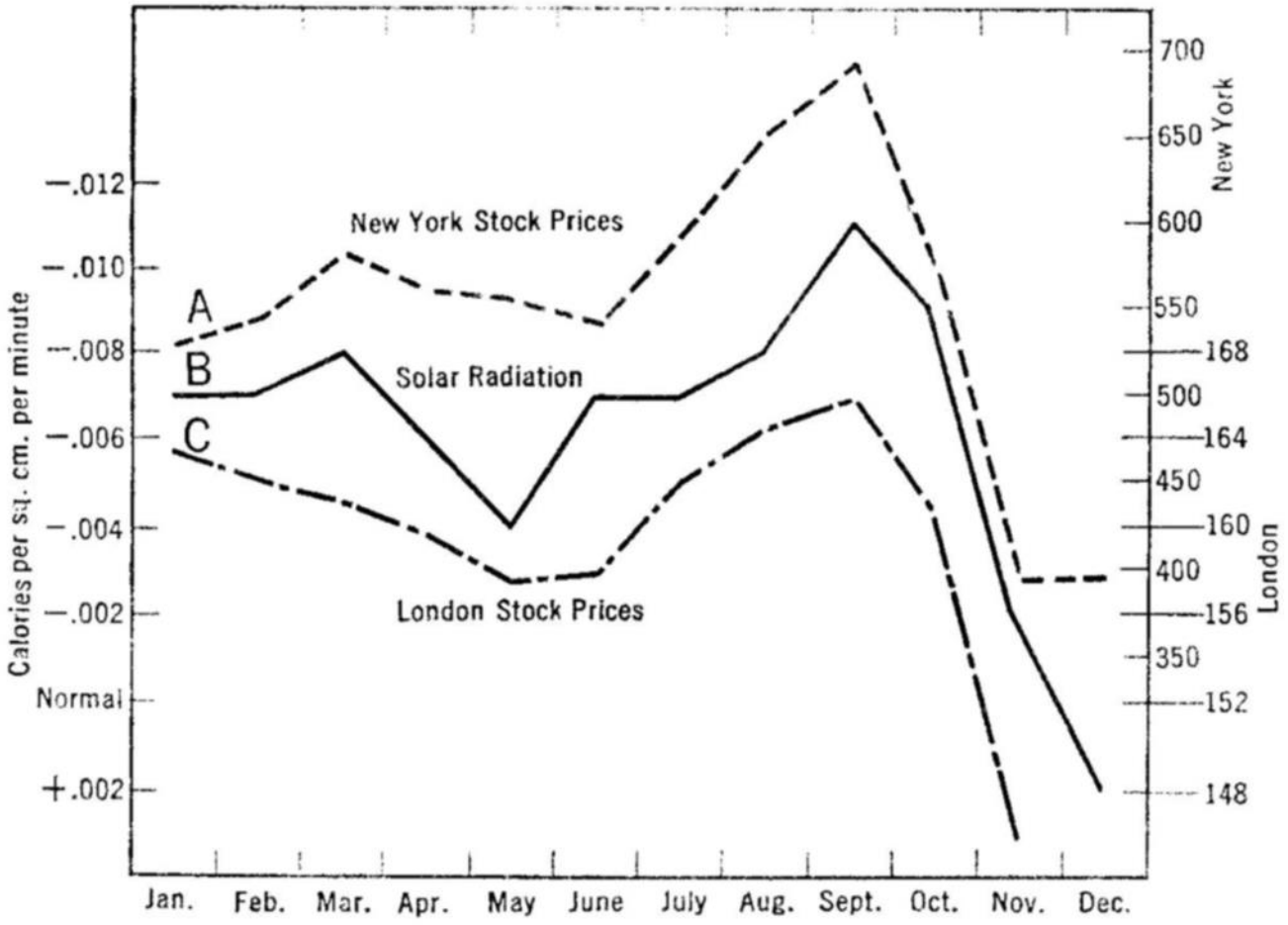


Male 60%
Female 40%

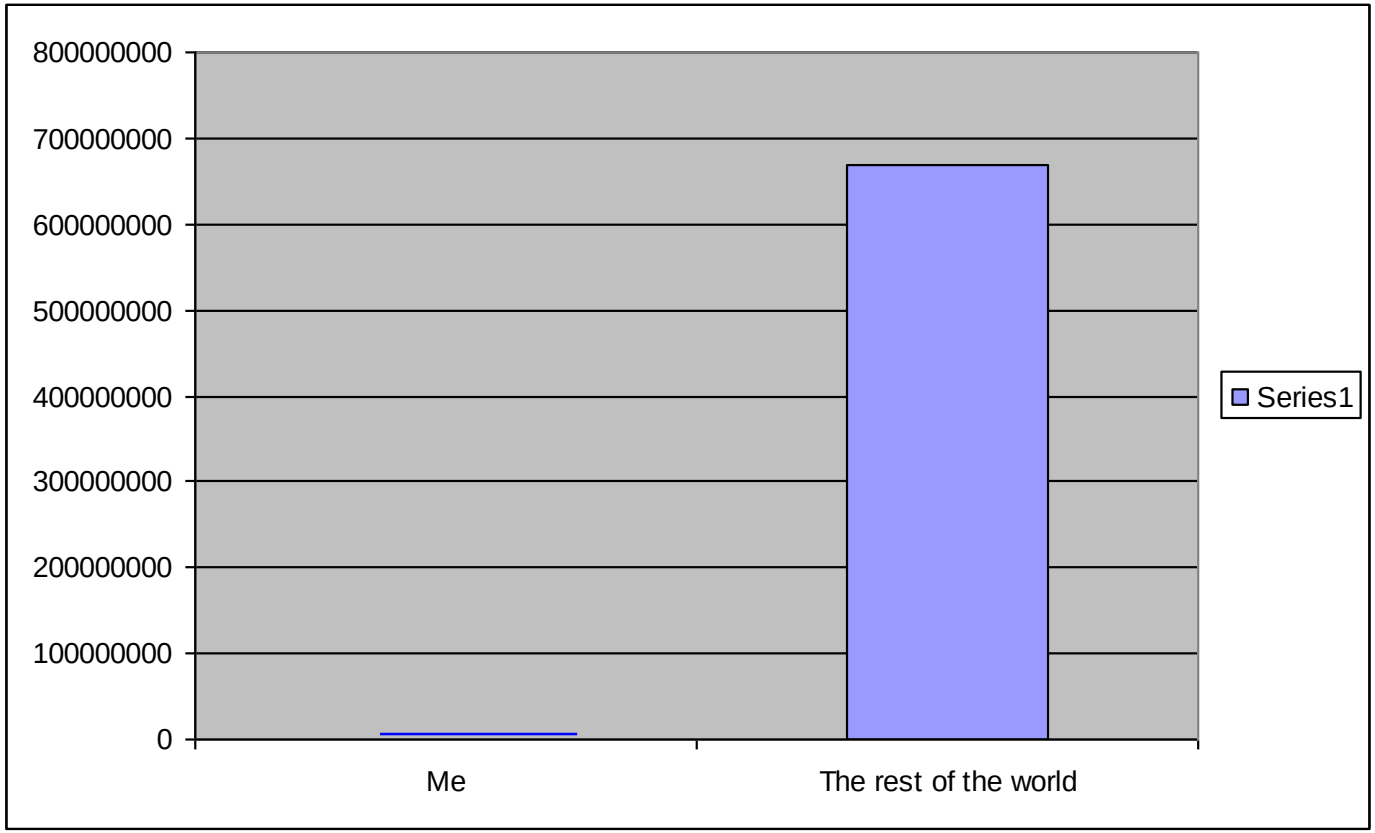
Rule 1:
Insure data quality / significance

- Graphs are only as good as the data they display
- No amount of creativity can produce a good graph from dubious or non relevant data

Rule 1 violation



Role 1 violation (and also rule 0)



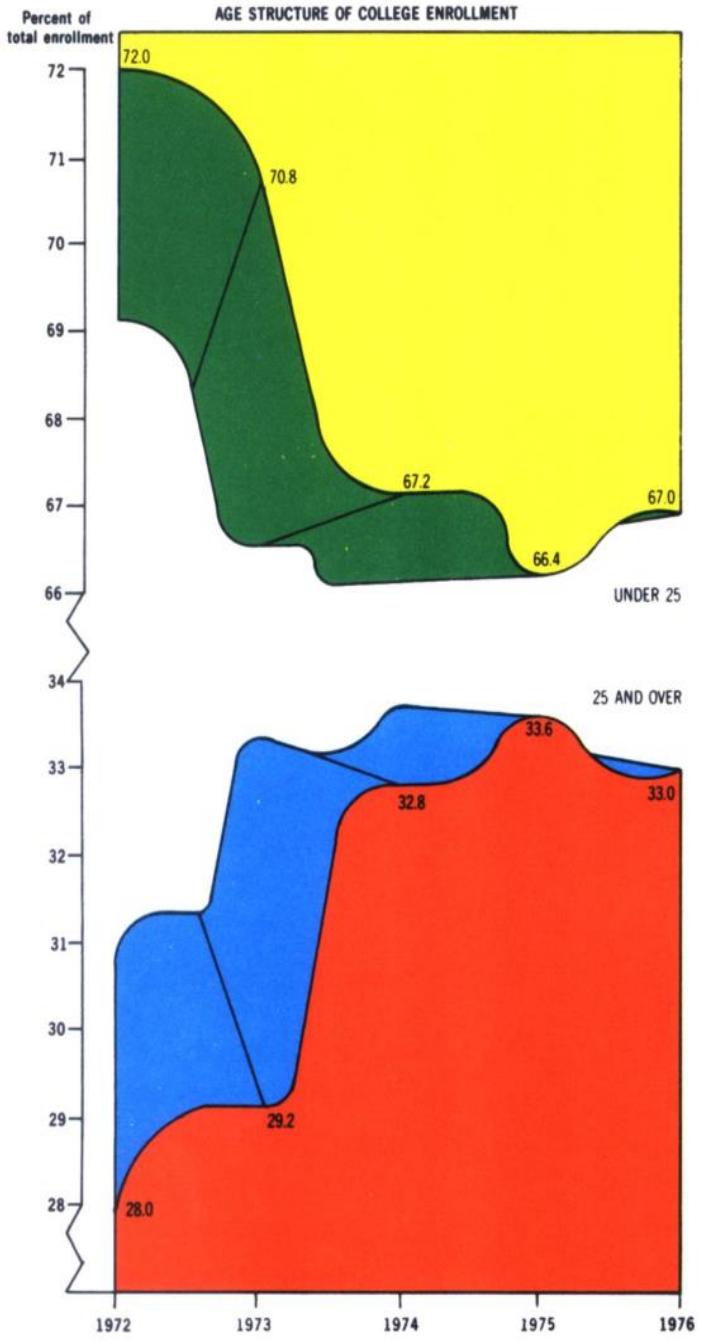
Not very significant data but good example of distortion

Rule 2:
Insure chart simplicity

- Graphs should be no more complex than the data which they portray

- Unnecessary complexity can be introduced by
  - irrelevant decorations
  - colors
  - 3d effects
  - ...
- These are collectively known as “chartjunk”

- For a very comprehensive set of chartjunk effects look at Microsoft Excel
  - the later the version the larger the set !



Age structure of College enrollment (percentage of enrolled people above 25 years)

Rule 2 violation (and also rule 3)

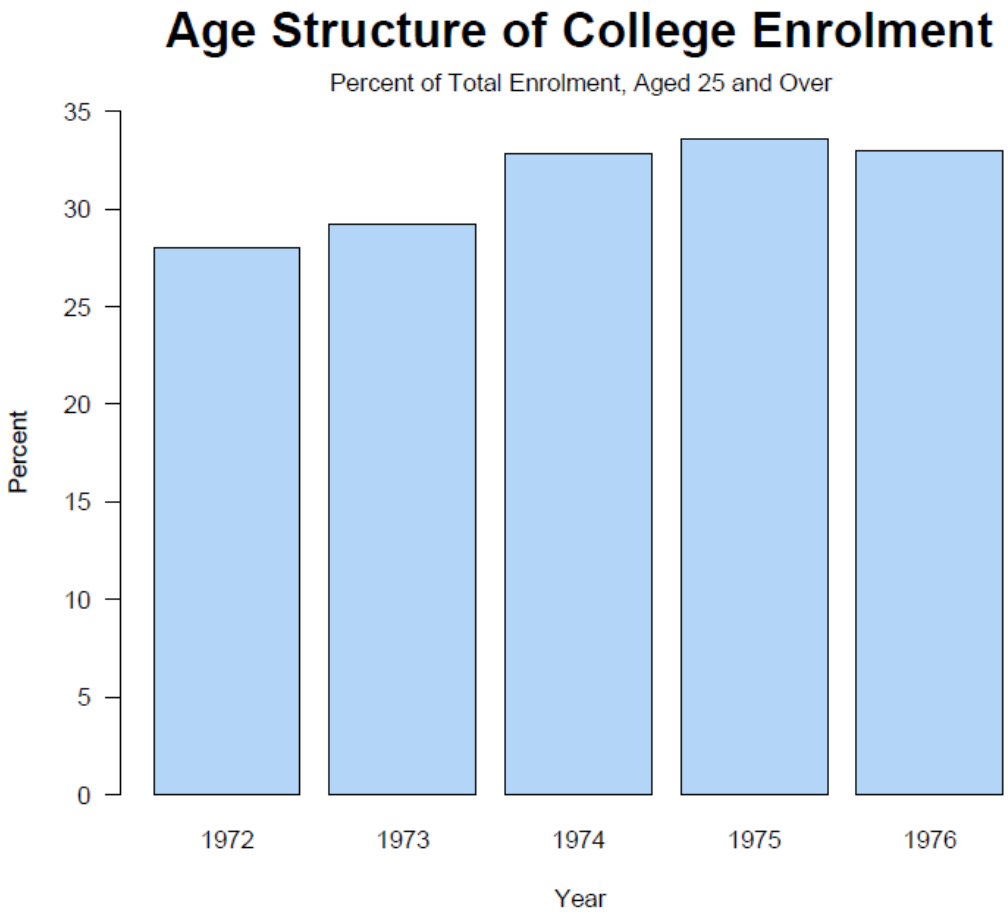
- A very good bad example!
- **Only 5 numbers on it** but
  - 4 meaningless colors
  - useless 3D
  - useless axes split
  - confusing and wrong visual attributes (size)
  - random interpolation
- Designers of this graph are now working in the Microsoft Excel's team, inspiring the new Excel's versions ...

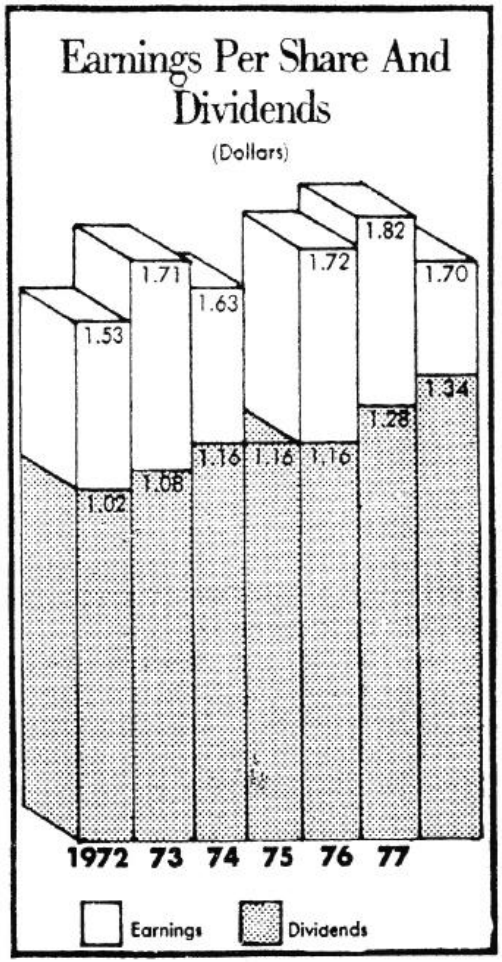
American Education Magazine

Same data...

Year	Percentage above 25
1972	28.0
1973	29.2
1974	32.8
1975	33.6
1976	33.0

Same data...


Age Structure of College Enrolment
Percent of Total Enrolment, Aged 25 and Over
Percent
35
30
25
20
15
10
5
0
1972
1973
1974
1975
1976
Year

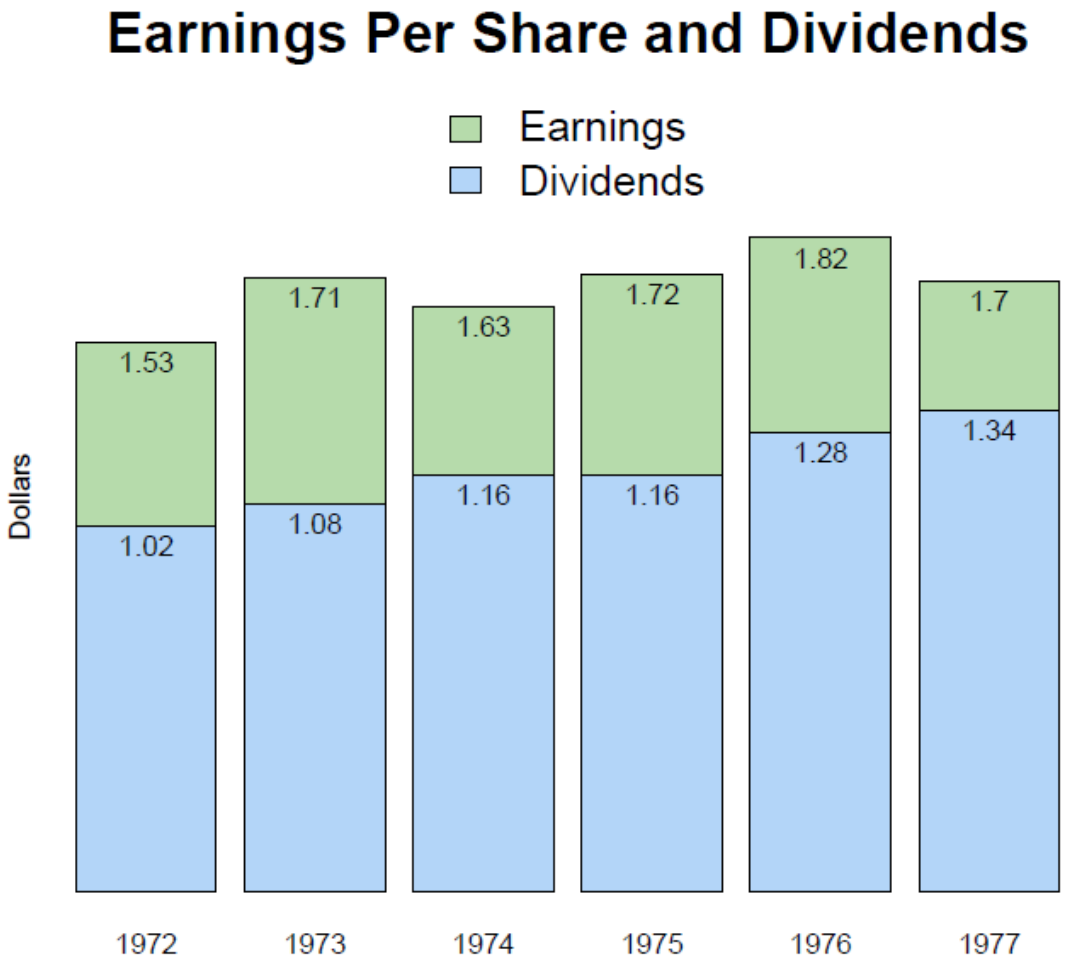


Rule 2 violation

- Why 3D?
- The extra dimension used in this graph has confused even the person who created it..

The Washington Post, 1979

The same data...



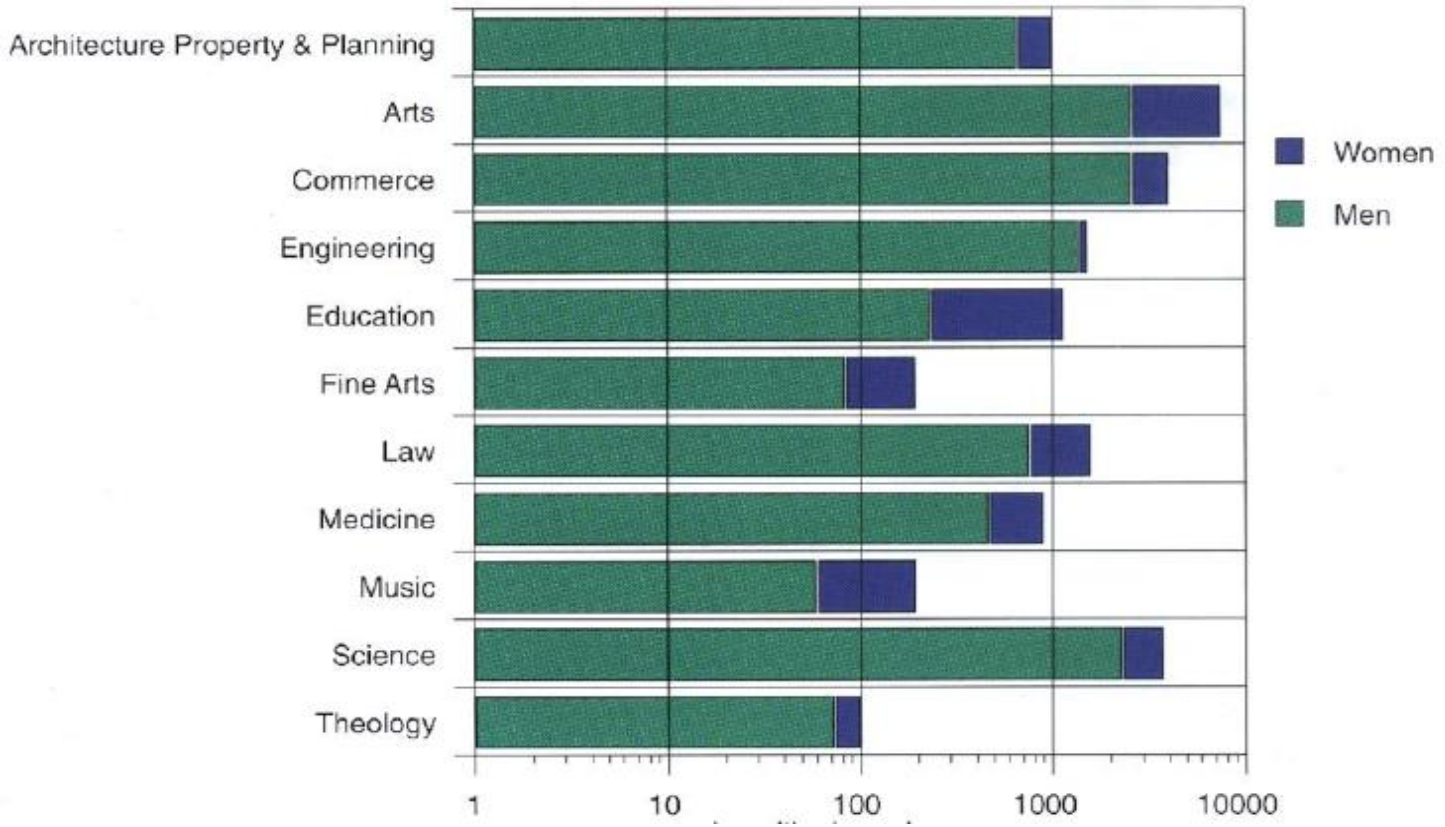
Rule 3:
Do not distort data

- Graphs should not provide a distorted picture of the values they portray

- Distortion can be:
  - deliberate
  - accidental

- Of course, it could be useful to know how to produce a graph which bends the truth...

Rule 3 violation

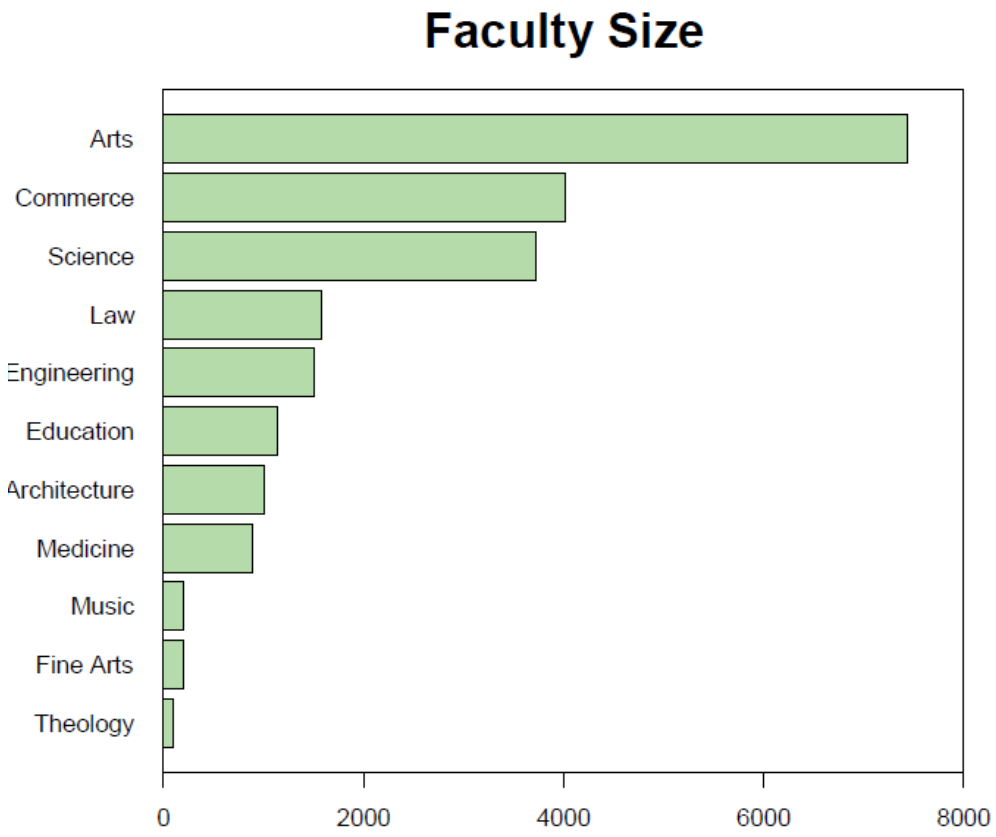


At a very quick glance:

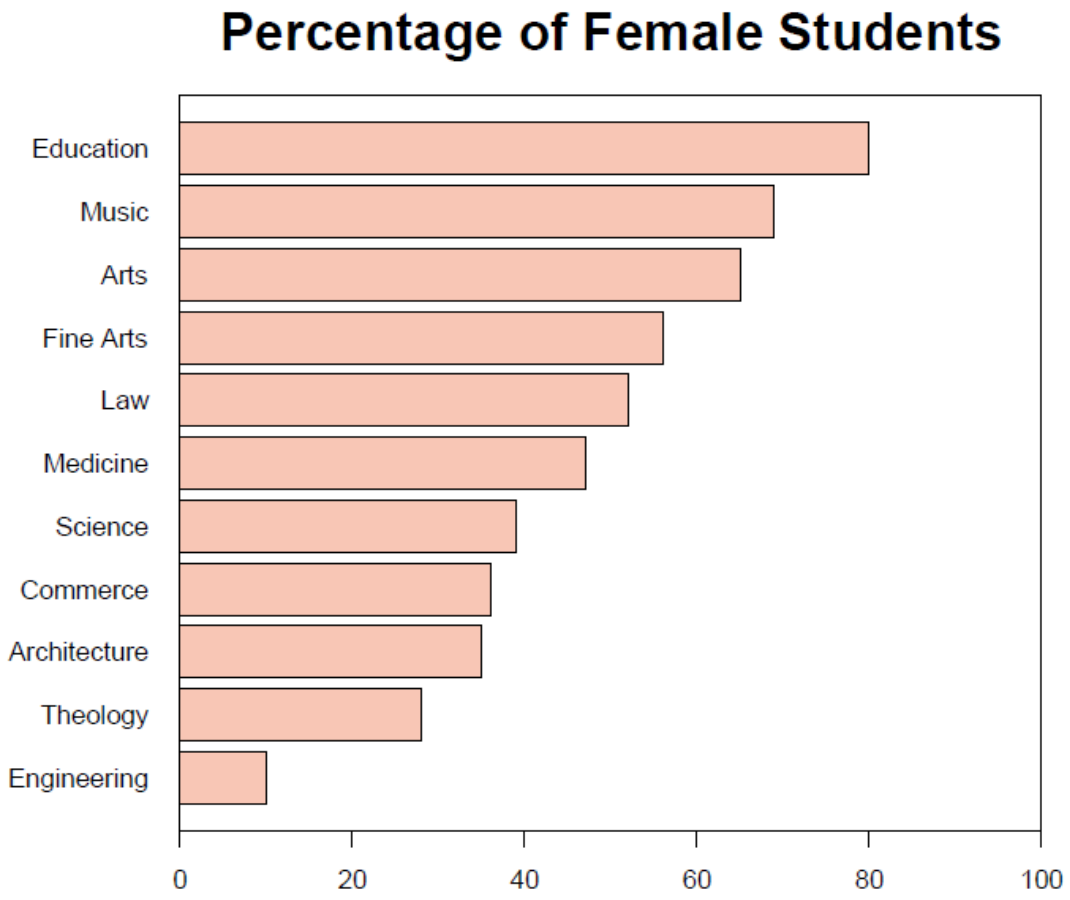
- balanced faculty population
- most male students

What's wrong with this graph?

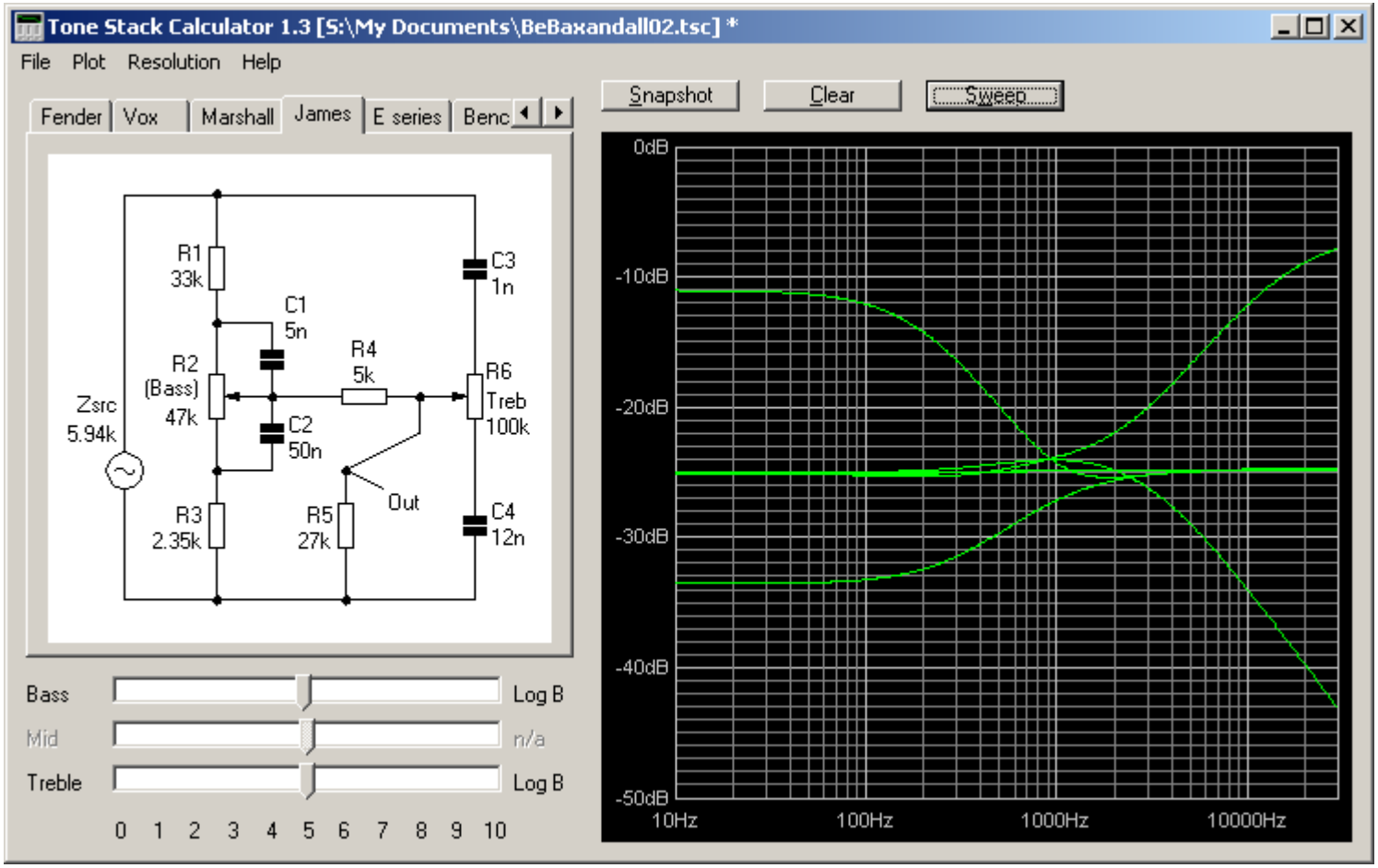
The truth : population size



The truth : male /female ratio



In other cases distortion is ok...

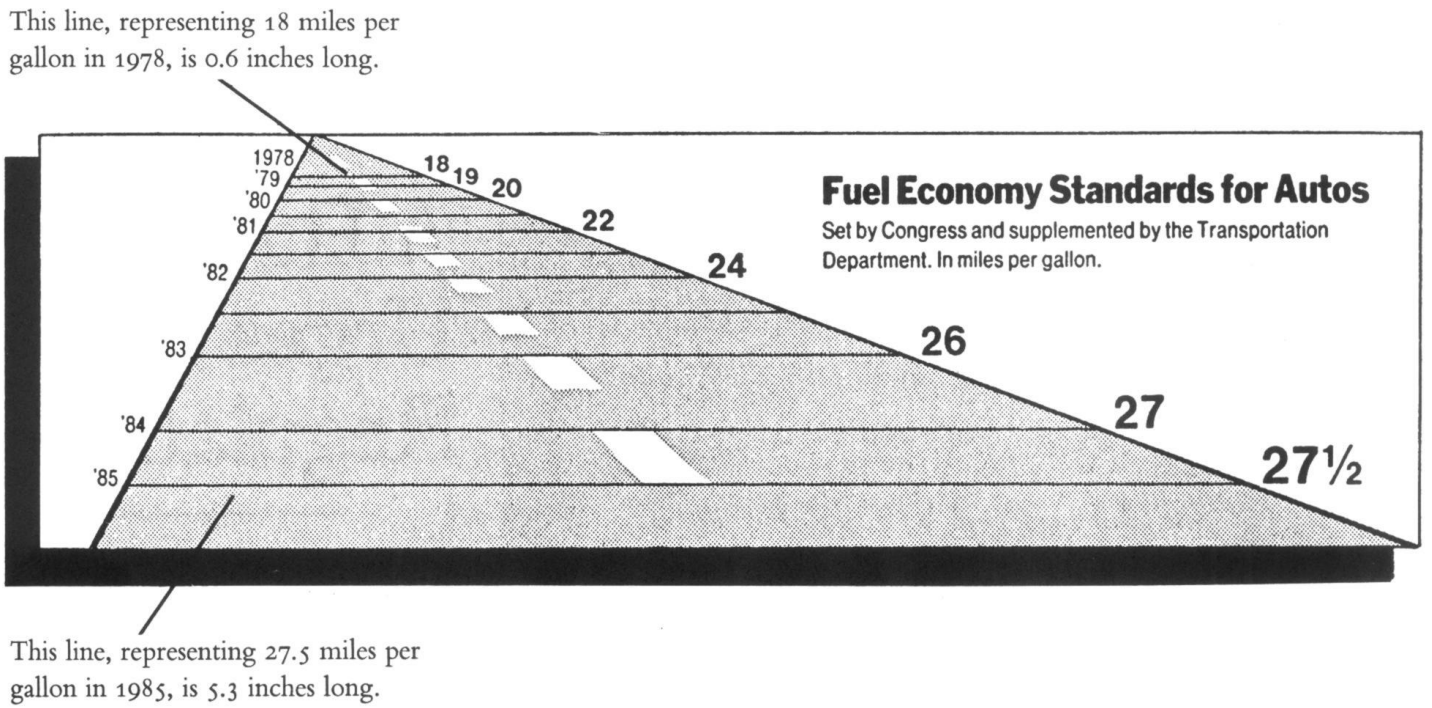


The lie factor

- Ed Tufte of Yale University has defined the “lie factor” as a measure of the amount of distortion
- Lie Factor =

  size of effect in graphic / size of effect in data
- If the lie factor is greater than 1, the graph is exaggerating the size of the effect

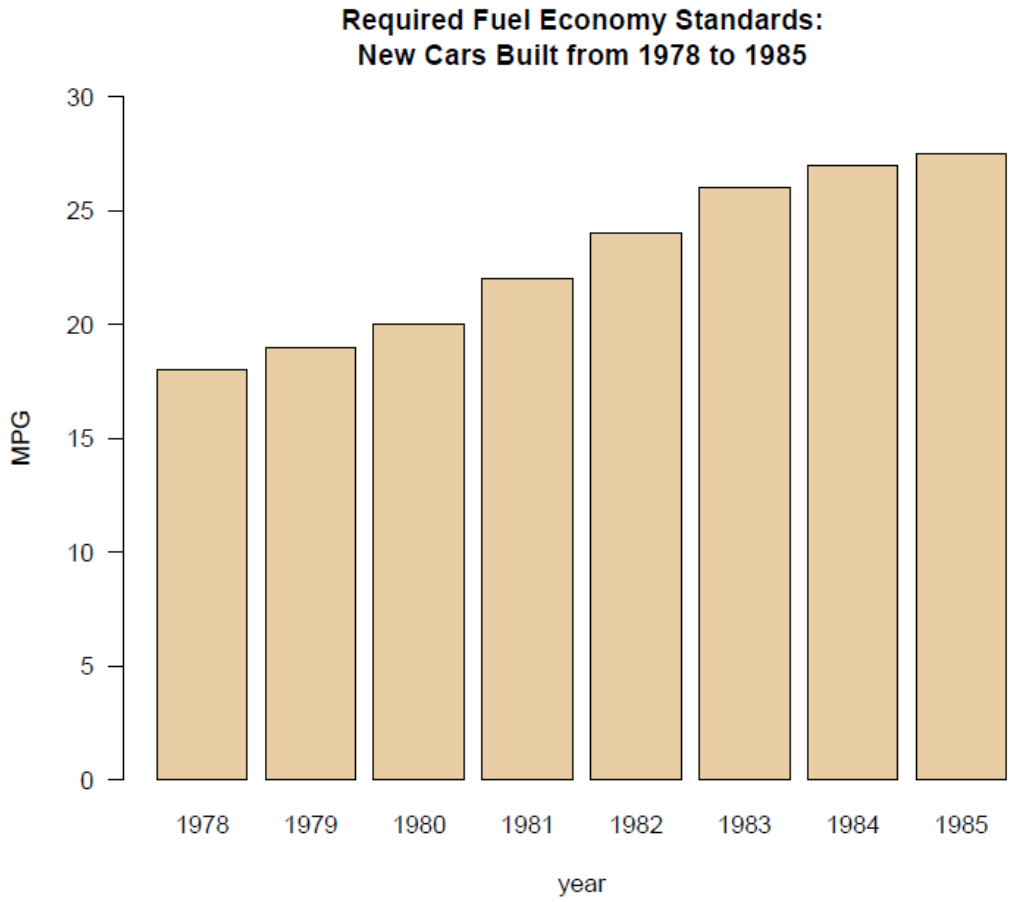
Measuring distortion through the lie factor (miles per gallon across years)



$$\text{Data Effect} = \frac{27.5 - 18}{18} = 0.53, \qquad \text{Graph Effect} = \frac{5.3 - .6}{.6} = 7.83,$$

$$\text{Lie Factor} = 14.8$$

The same data with lie factor=1 (and following the previous roles)

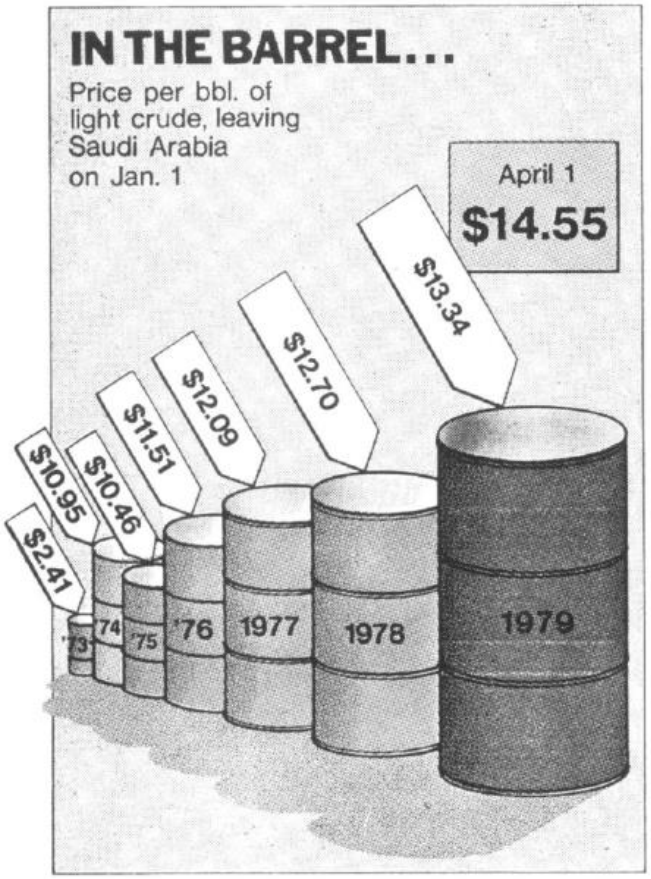


Common Sources of Distortion

- The use of 3 dimensional “effects” is a common source of distortions in graphs

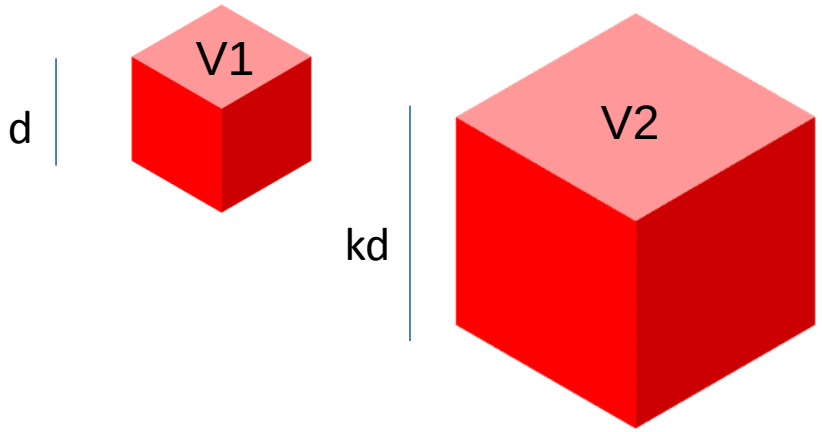
- Another common source is the inappropriate (or deliberate?) use of linear scaling when using area or volume to represent values

Distortion through volumes



$V1=d^3$
$V2=k^3d^3$

$V1/V2 = k^3$
$kd/d = k$



Lie factor= $k^3/k = k^2 =$
size_of_effect_in_data²

Distortion through areas

kd

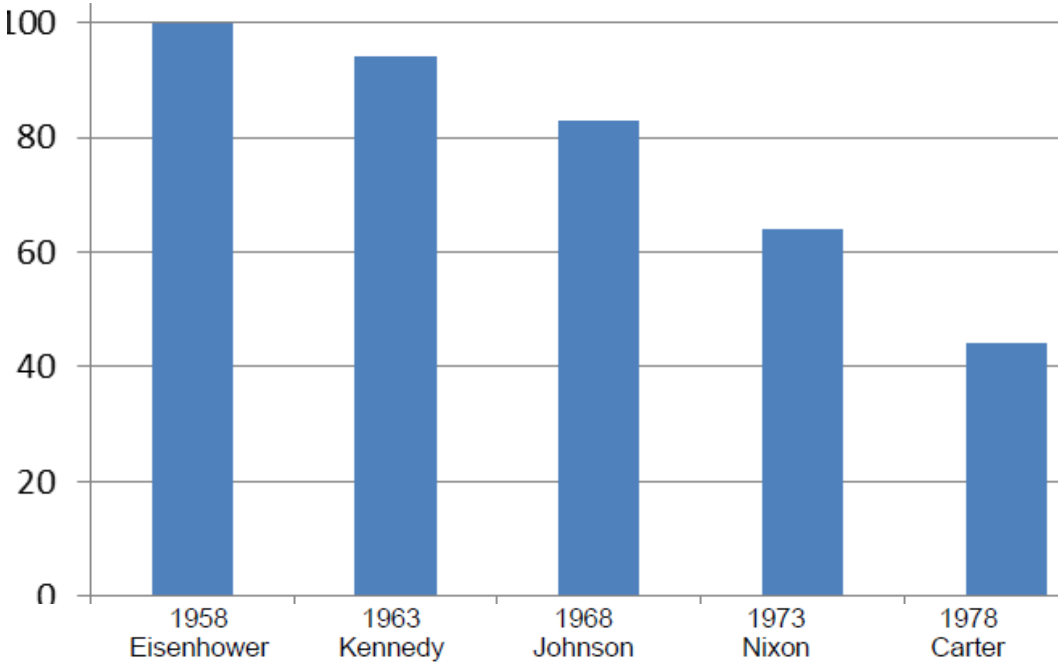
Lie factor = $k^2/k = k =$ **size_of_effect_in_data**

d

Is the bottom dollar roughly half the size of the top one?

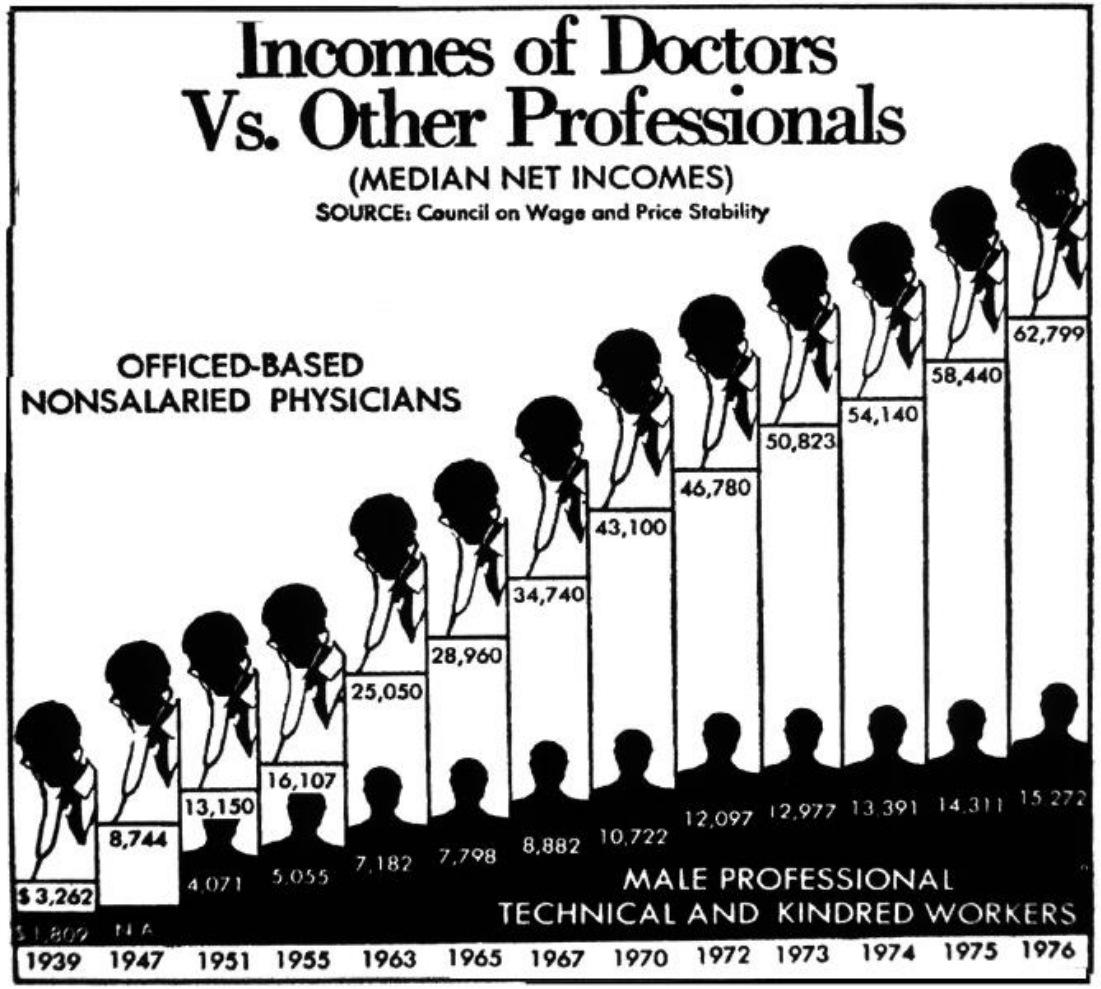
The same data with lie factor = 1

Note that in a histogram you are comparing **lengths**, not **areas**



This is why it is better to use thin bars...

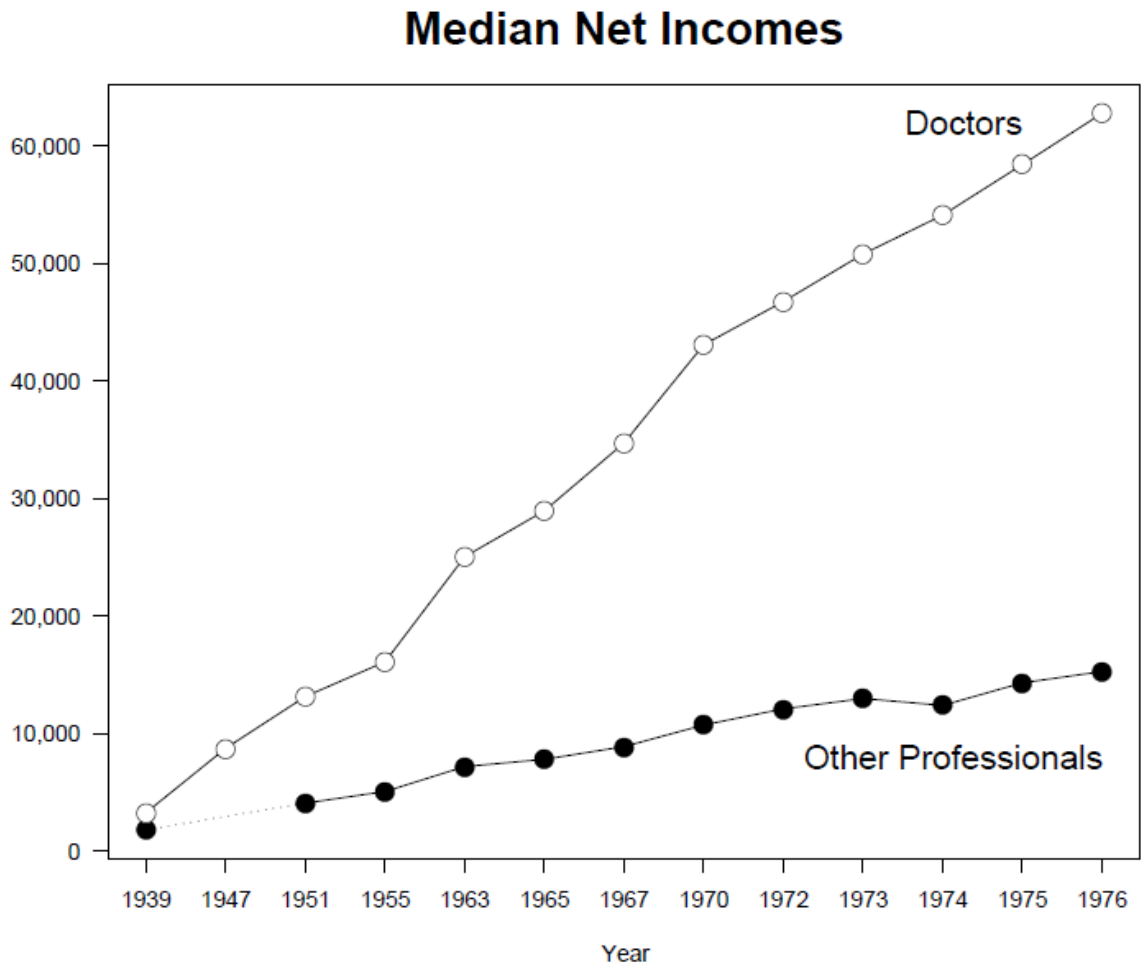
Distortion (deliberate?)



What's wrong with this graph?

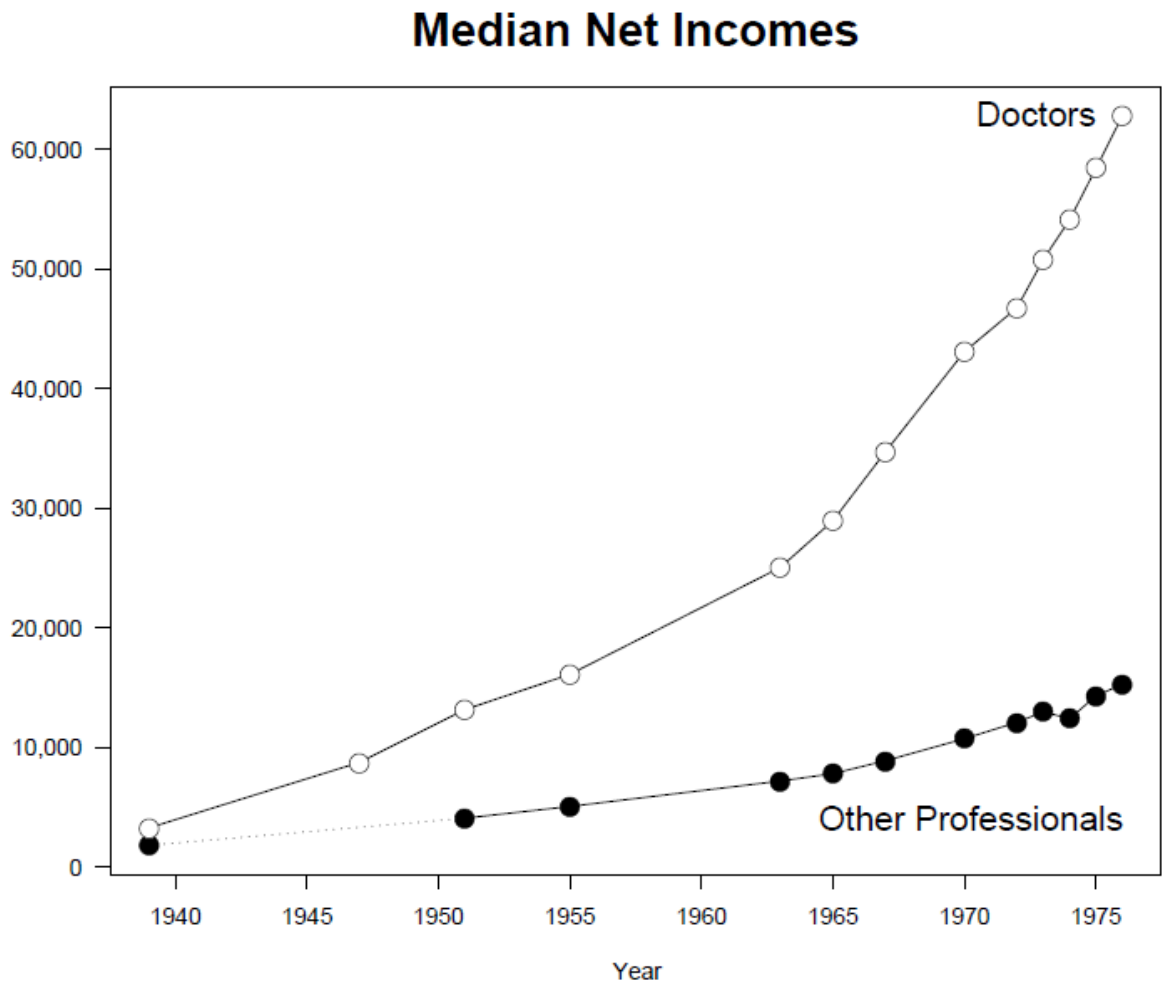
Neglecting chartjunk...

Removing chartjunk



It suggests
a linear trend

Real data...



The time scales were different!

Exponential trend !

One of the best graph lie...

- The cover story, "Why does college have to cost so much?" shows a large graph superimposed on a scene from the Cornell campus. There are two jagged lines running across the graph
  - "Cornell's Tuition" = MONEY
  - "Cornell's Ranking"= QUALITY
- The clear impression is that students are paying more for far less
- What is wrong with it?

The lie

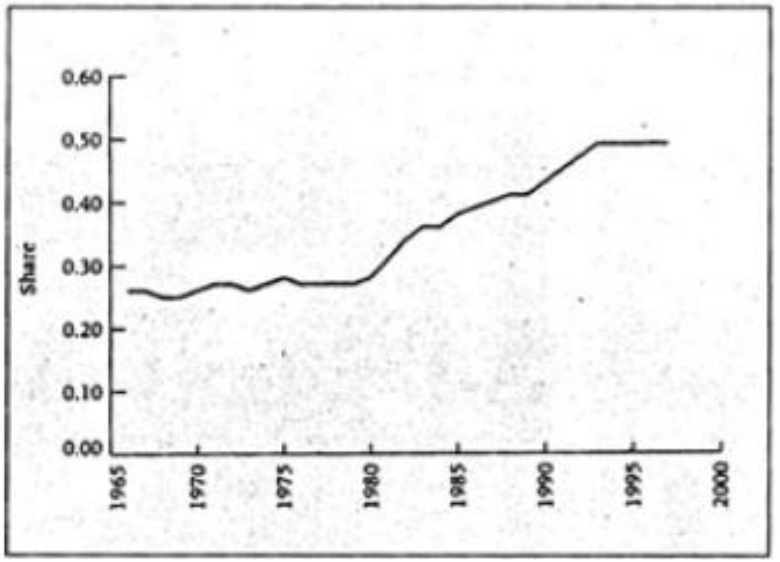
- The ranking graph covers an 11 year period, the tuition graph 35 years, yet they are shown simultaneously (the same apparent width) on the same horizontal "scale".

- The vertical scale for tuition and ranking could not possibly have common units, but the ranking graph is placed under the tuition graph creating the impression that cost exceeds quality.

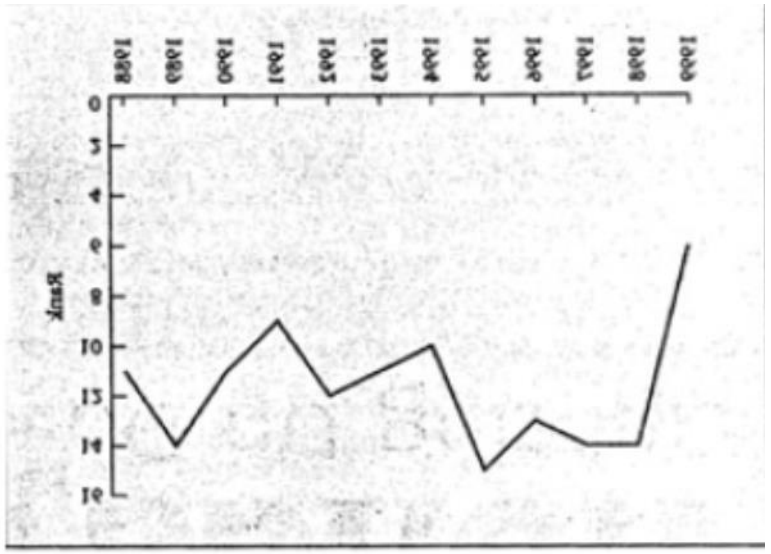
- The differing time units are cleverly disguised by printing them rotated 90°.

- And here is the masterstroke: the sharp "drop" in the ranking graph over the past few years actually represents the fact that Cornell's rank has IMPROVED from 15th TO 6th ...

The real data



Money

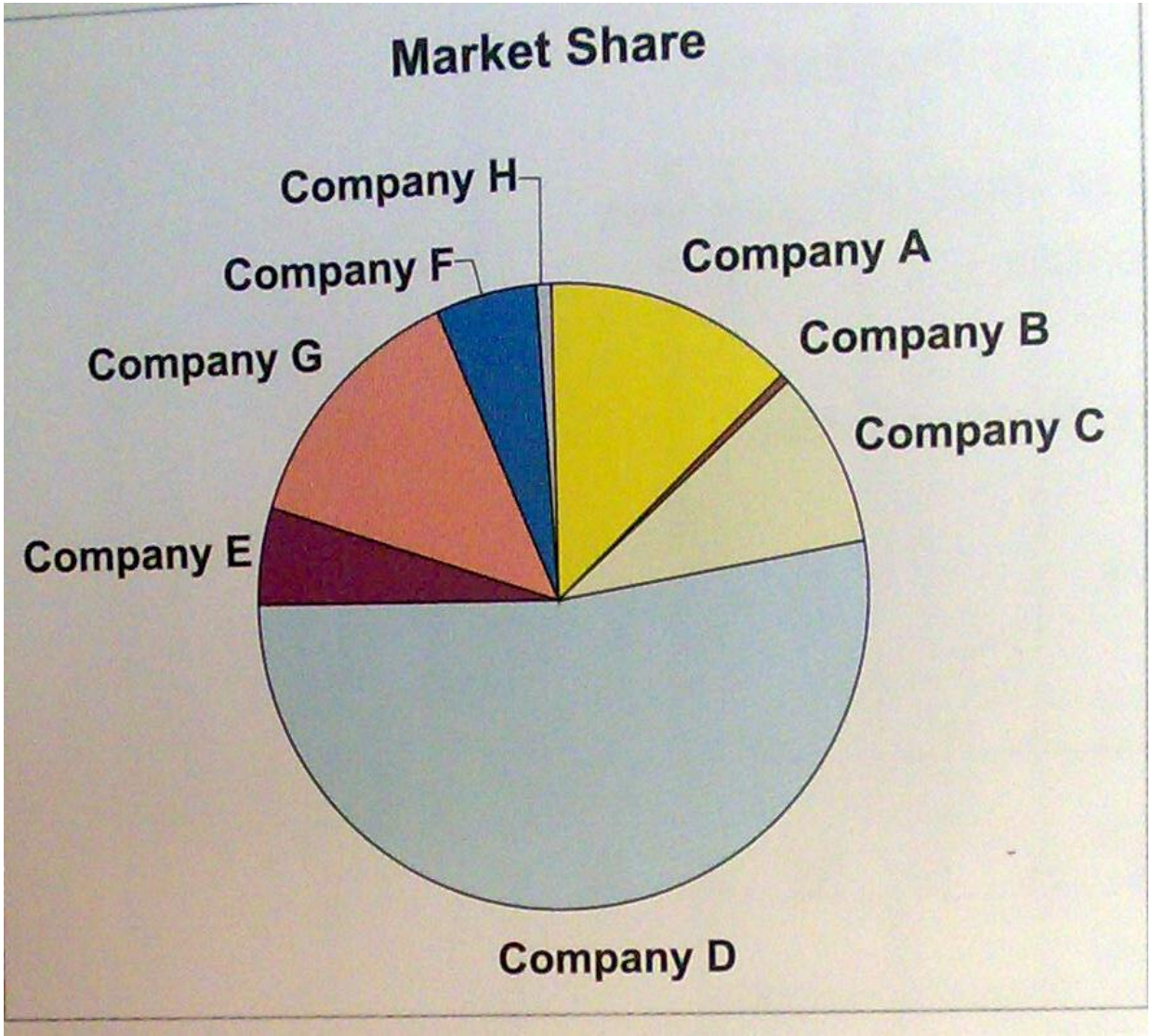


Rank

Outline
(basically what you have NOT to do)

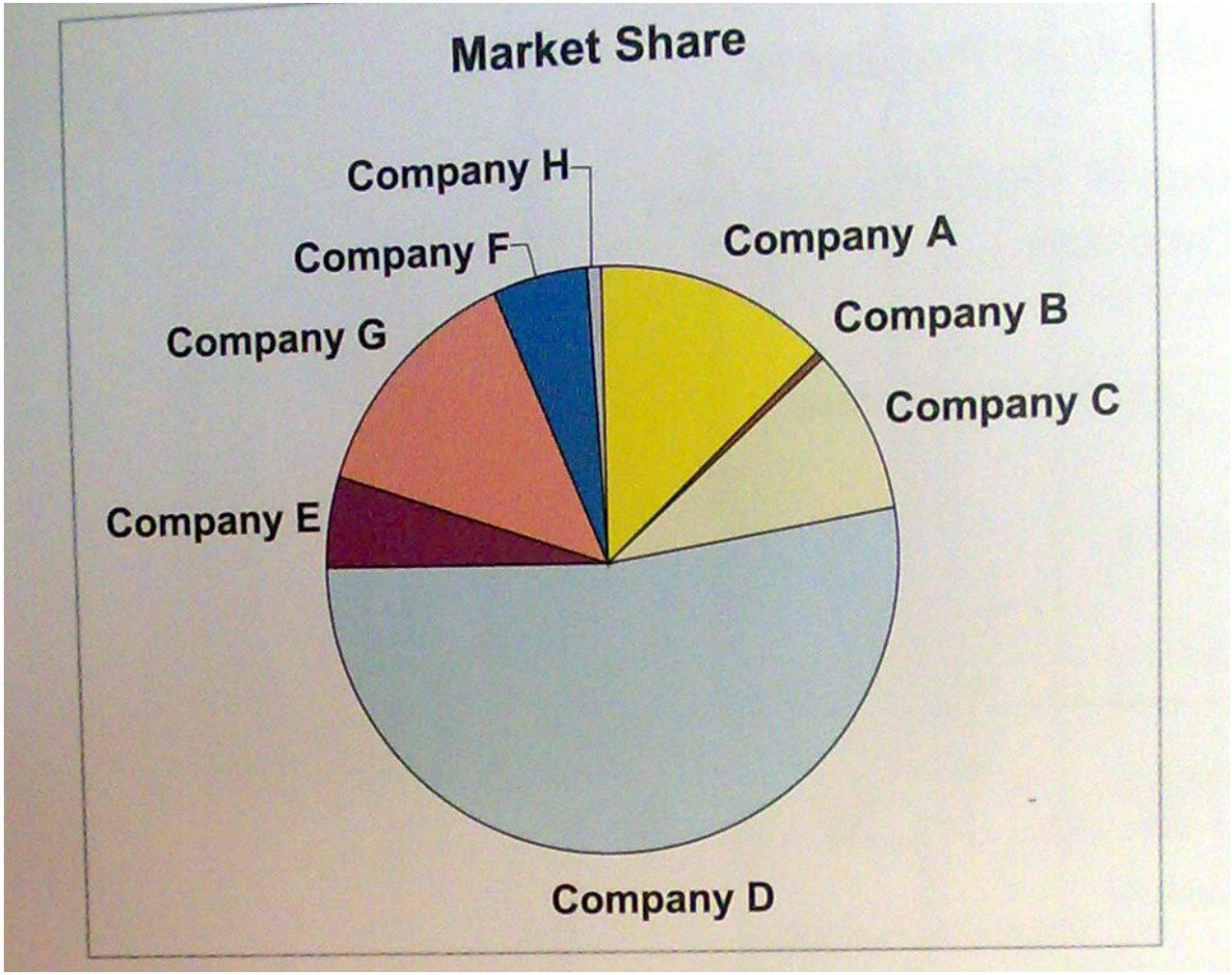
- An introductive example
- Good and bad graphs
  - Basic rules
  - Some additional considerations
- Visual issues

The last example: our company against the world!



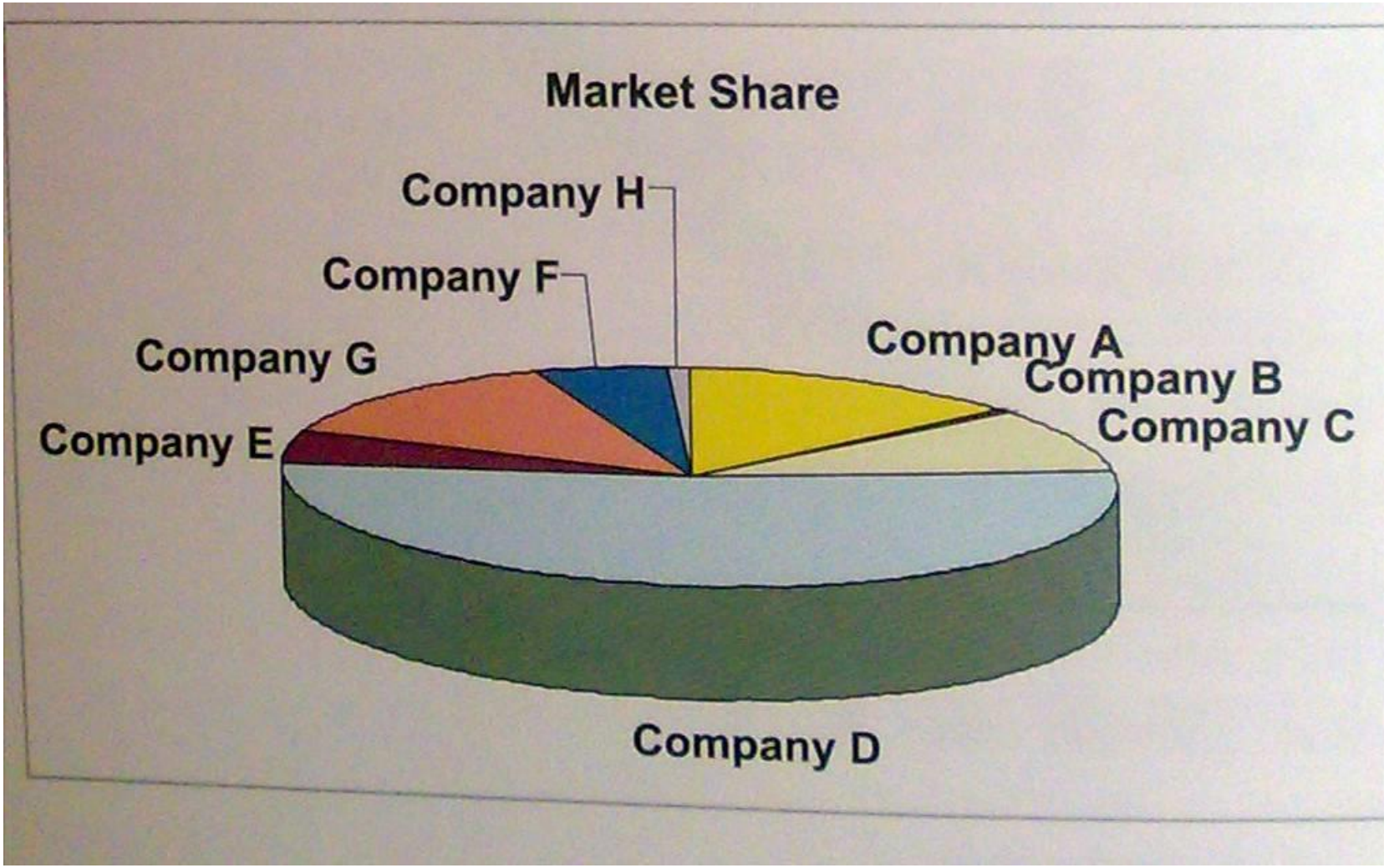
- What is the purpose of this chart?
- Comparison !
- What is wrong whit it?

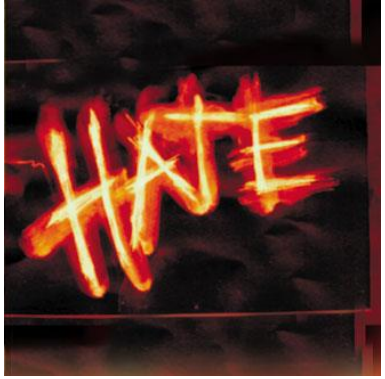
The last example



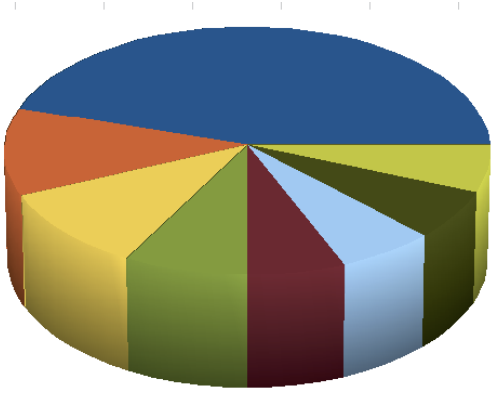
- Is the order clear?
- Which is my company?
- Who is bigger G or A?

Even worst : 3D!!!

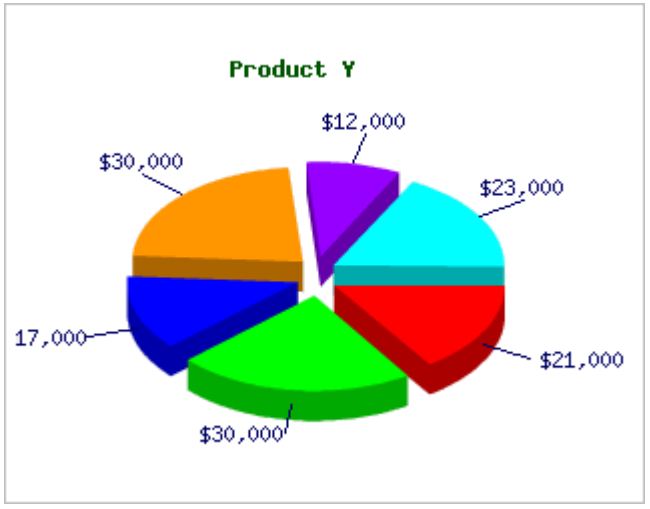


I  pie charts!

HATE

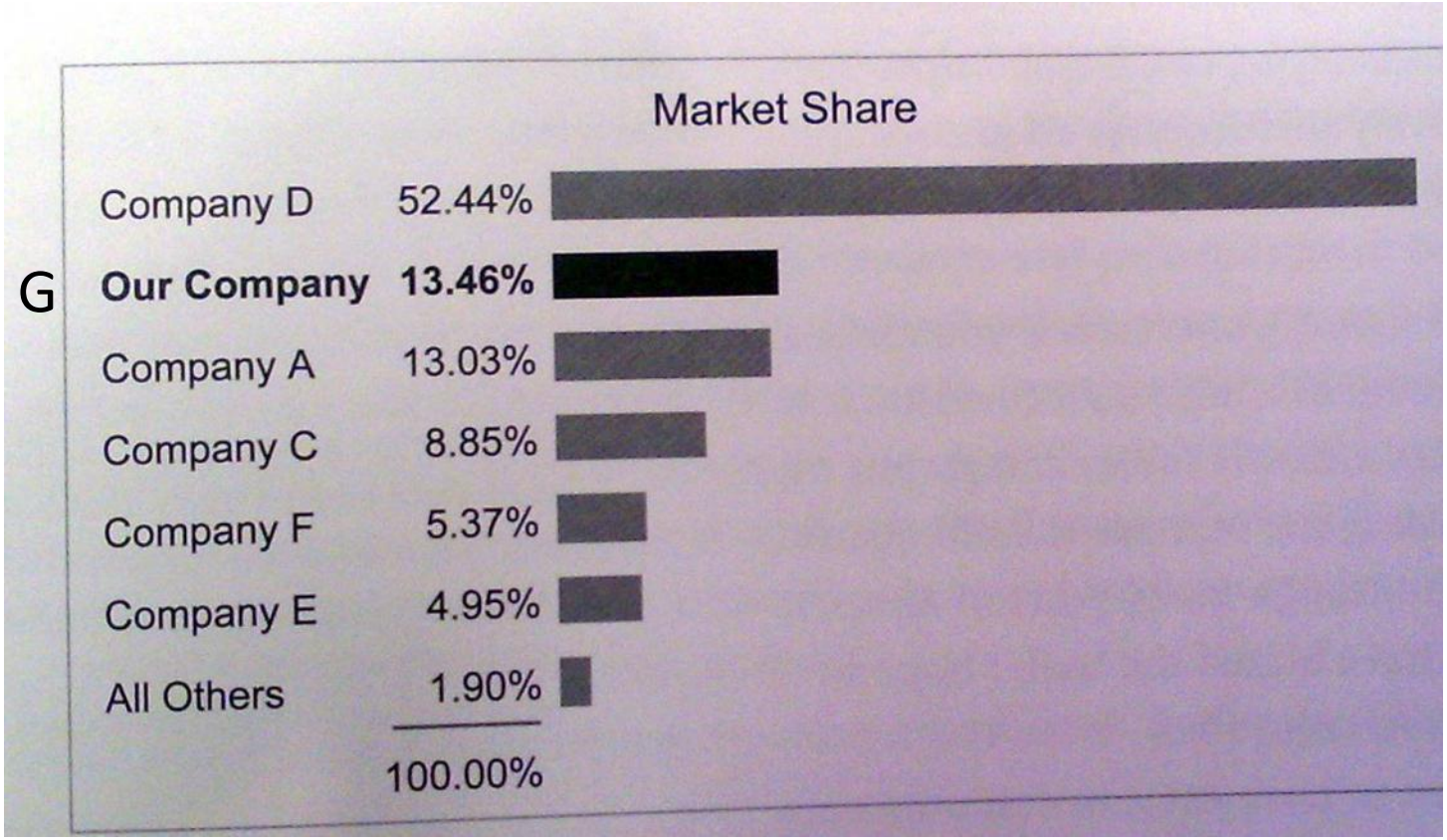



MySpa
Faceb
Orkut
hi5
Xanga
Live Sp
Classm
Bebo


Product Y
$12,000
$30,000
$23,000
17,000
$21,000
$30,000

At least most of them...

A better solution



G

If you have ordering (ranking) alternatives think about that!

Chartjunk is not the unique enemy...

- Before PCs, building graphs was a matter of paper and pencil
  - requiring time and effort
  - pushing you to better understand :
    - the meaning of numbers
    - the graph purpose
    - the graph organization
    - …
- now, with Excel you can produce graphs so fast that you might loose control...
  - you select predefined solutions
  - you might not understand how the graph is built (row, columns, headings, ...)
  - you can make mistakes (e.g., missing a row...)

So...

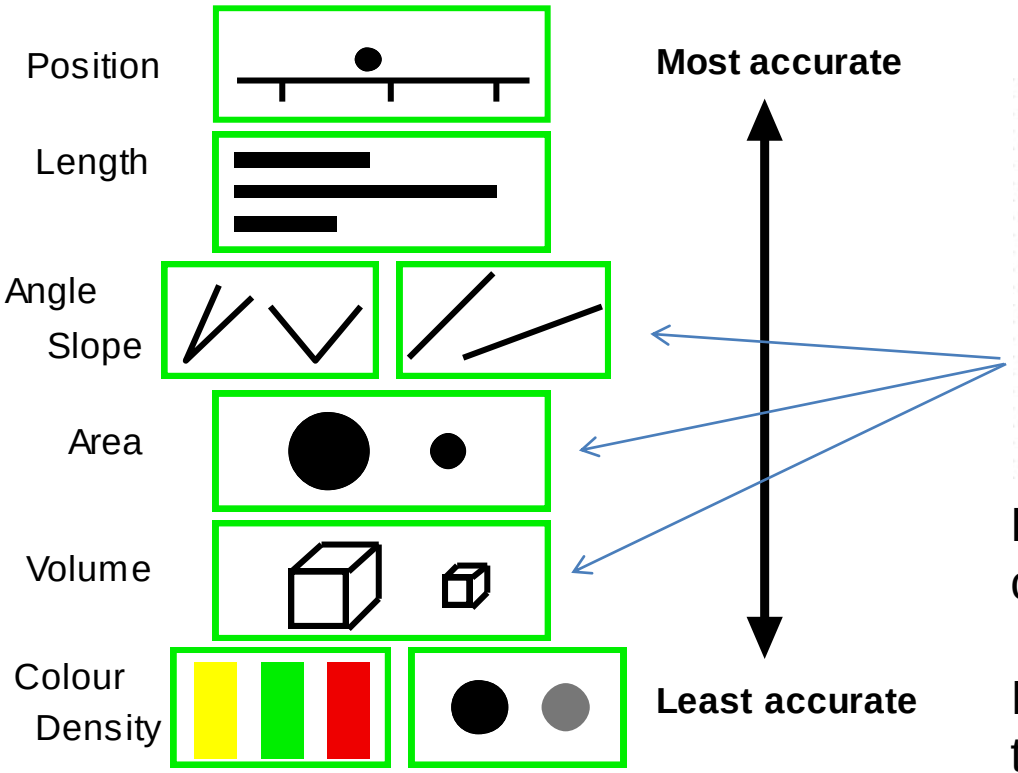
1. Look at the numbers and at the task
2. Plan a graph (even on the paper!)
3. Look for an Excel implementation of your design
4. If step 3 fails, proceed without Excel ! You can also consider more serious visualization tools, e.g., R (http://cran.r-project.org/bin/windows/base/) or MatLab (www.mathworks.it)

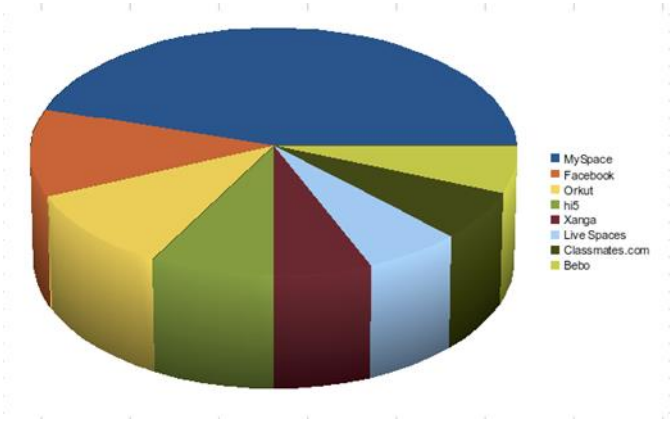
Outline
(basically what you have NOT to do)

- An introductive example
- Good and bad graphs
  - Basic rules
  - Some additional considerations
- Visual issues
  - Quantitative perception (basic rules)
- Information Visualization

Why do I  pie-charts?

The relative difficulty of assessing **quantitative** value as a function of visual encoding mechanism, as established by Cleveland and McGill

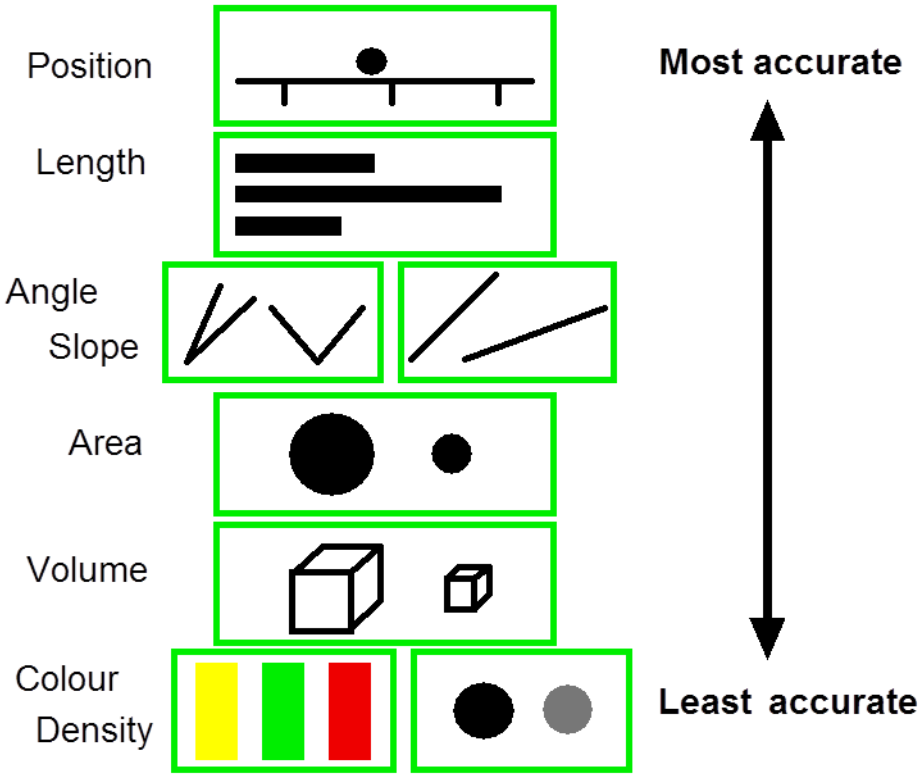




Pie-charts discards the two first choices

I do NOT see ANY reason to use them

What about quantitative comparison?

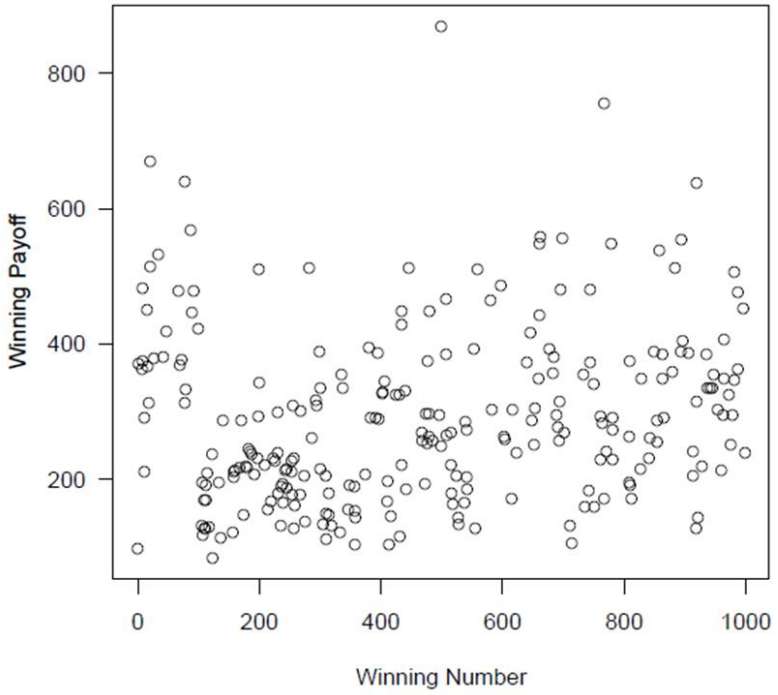


Use position and length
Avoid angles
Avoid areas
Avoid volumes

Use colors carefully

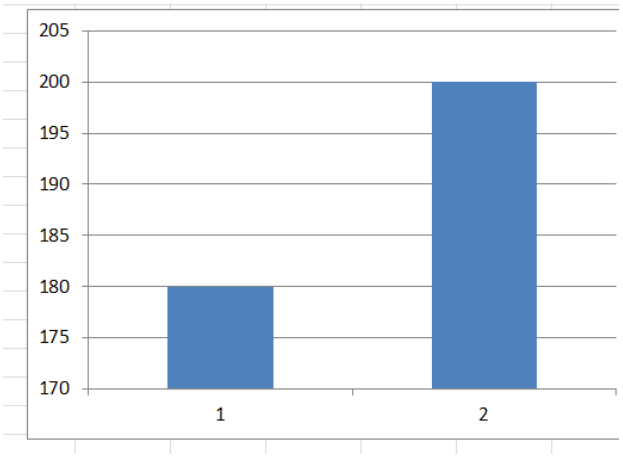
Position

- It works fine

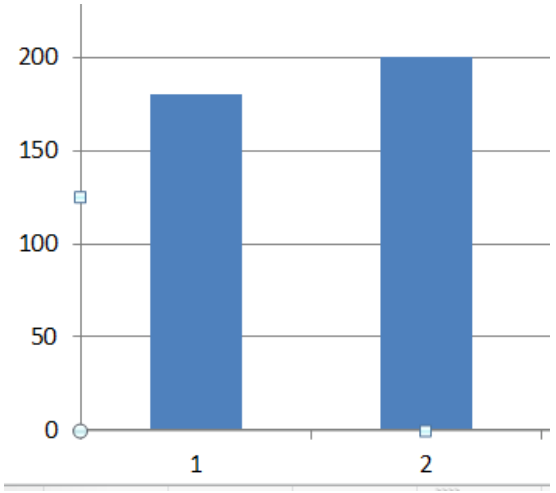


Length?

- Length is fine as well , but use the right scale!



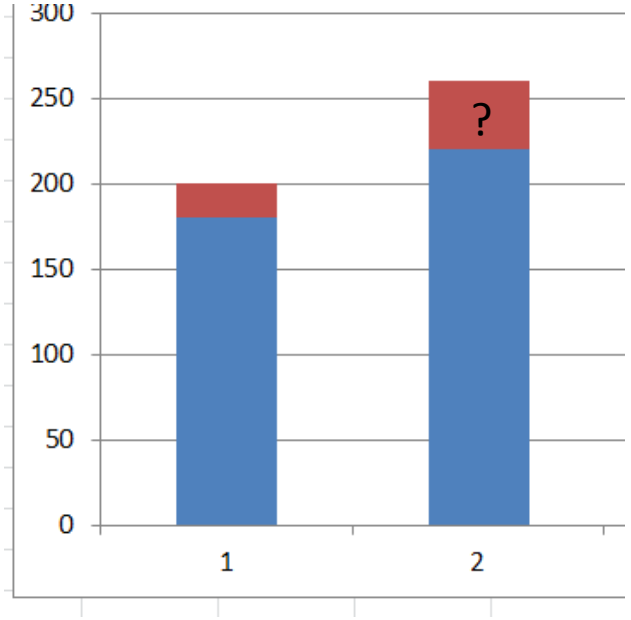
Automatically produced by Excel

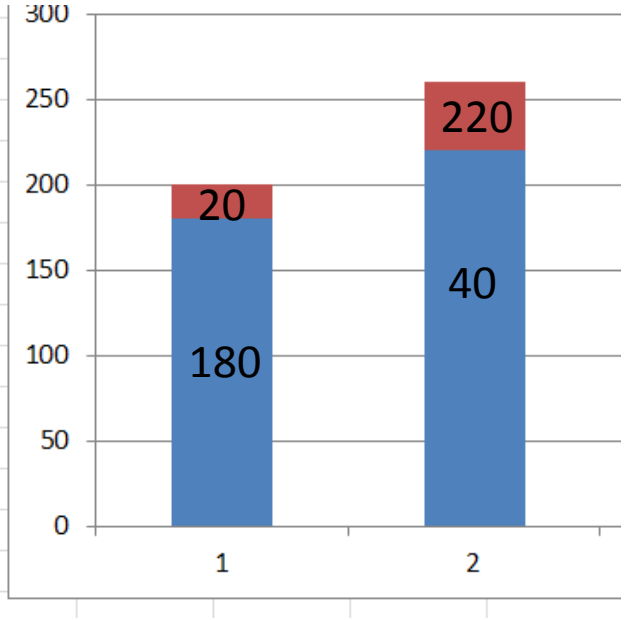


The reality

Length?

- The lookup of precise number might be difficult if the position is not evident (e.g., stacked bar chart)



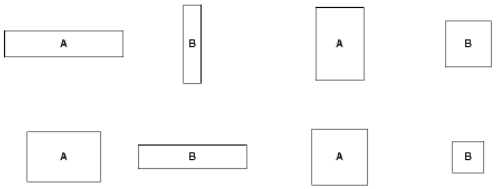


It makes sense to explicitly add figures

Areas: some new surprising issues

- Human being are bad in estimating area ratios

- What is the ratio between this two circles?
  35% 40% 45% 50% 55% 60% ?
- What is the shape that produces the biggest error?



- The square!
- Perceptual Guidelines for Creating Rectangular Treemaps (Nicholas Kong et al., Infovis 2010)

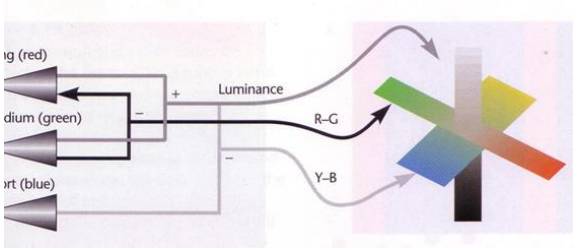
Colors

- Someone already thought how to associate quantitative values to colors and different choices are available
- Do not reinvent the wheel
- (The rainbow scale does not work)

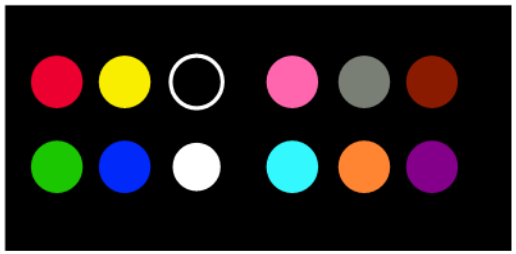
rainbow scale

HSI color model
(Keim and Kriegel) - Issues in visualizing
large databases. Proc. of the IFIP working conference
on Visual database Systems, 1995

Colors



- Colors are fine with categorical data
- Do not reinvent the wheel (again)
- The Hearing idea (1920) is that there are only 6 elementary colors arranged in three pairs
- That gives us up to 12 (6+6) colors easily distinguishable (11!)



12 Colors
for labeling

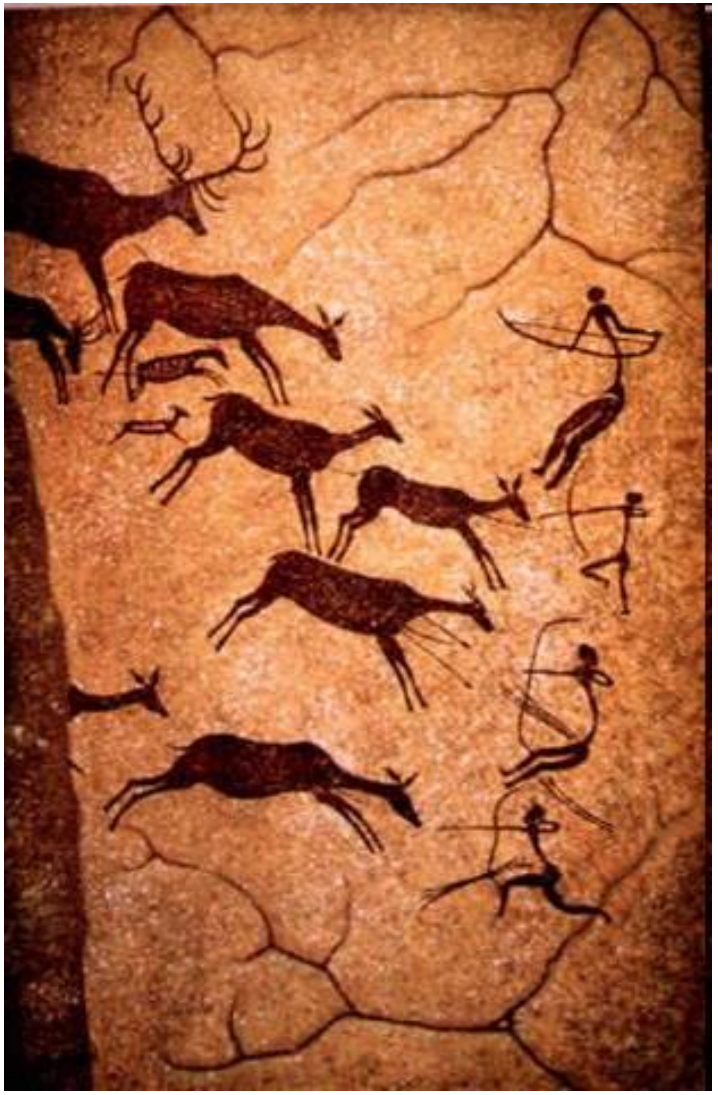
Outline
(basically what you have NOT to do)

- An introductive example
- Good and bad graphs
  - Basic rules
  - Some additional considerations
- Visual issues
  - Quantitative perception (basic rules)
- Information Visualization

Information Visualization?

Old stuff…

Lascaux walls
(1500 BC)



Visualization ?

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

Problem Solving/Analyzing

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

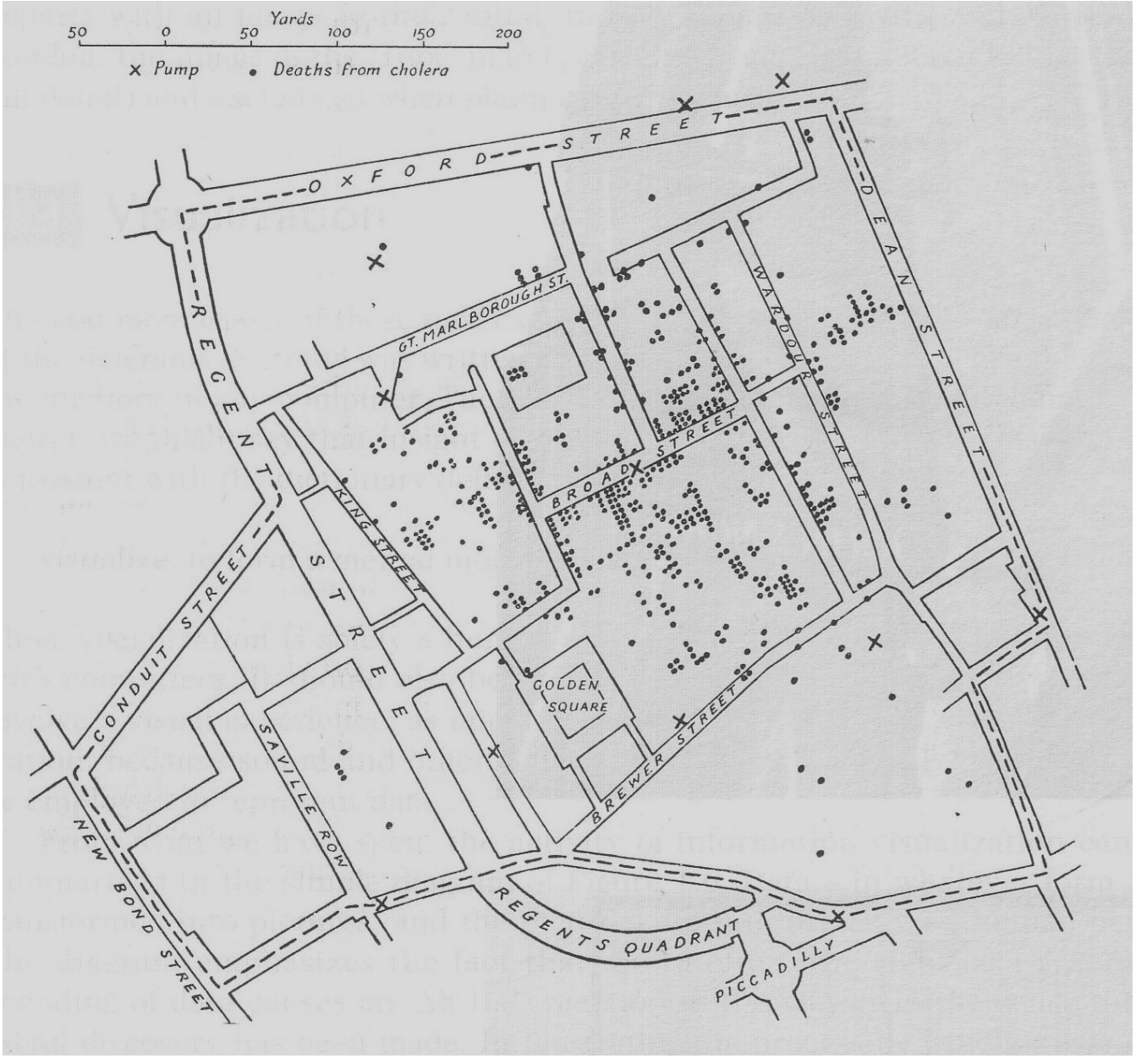
Visualization for Problem Solving

Illustration of Dr. John Snow (1854).

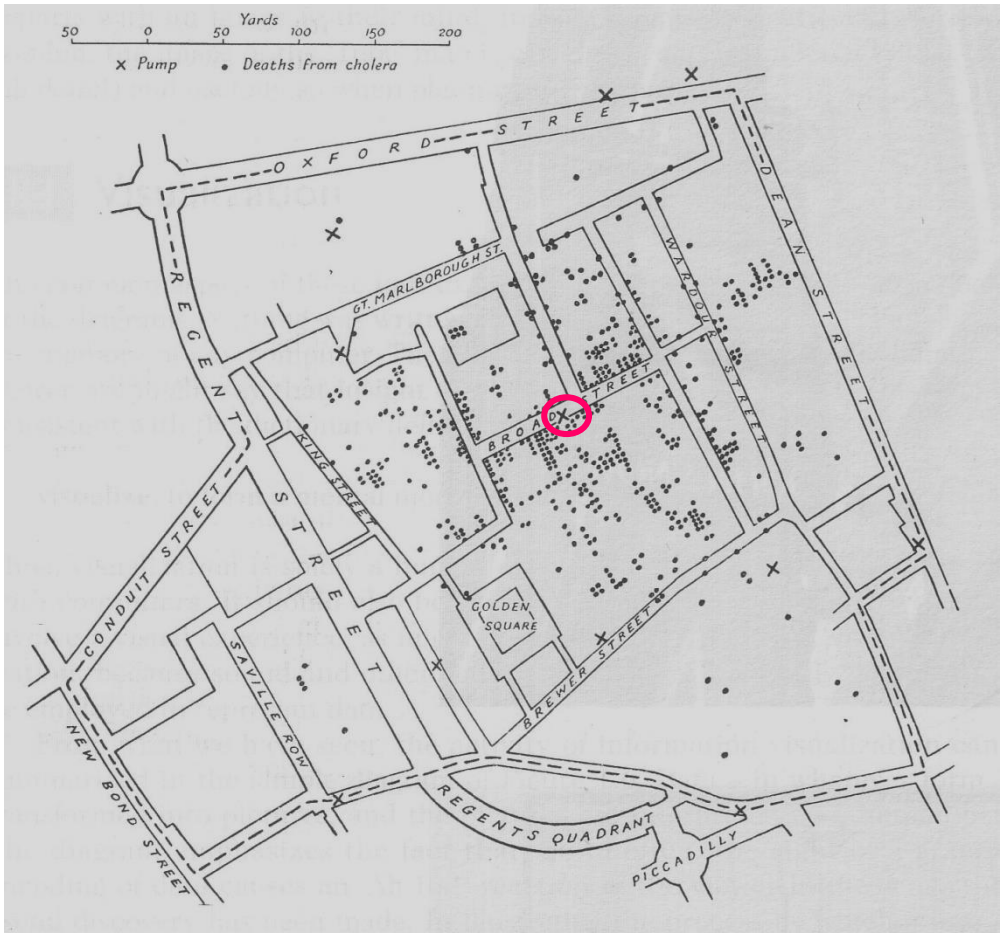
Dots indicate location of deaths.

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 !!!

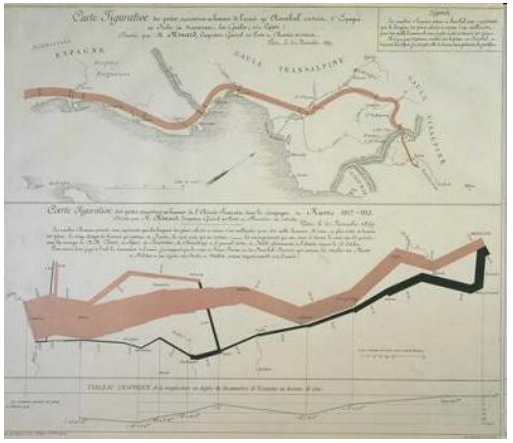
John Snow deducted that the cholera epidemic was caused by a bad water pump !!!
Closing that pump quickly solved the problem

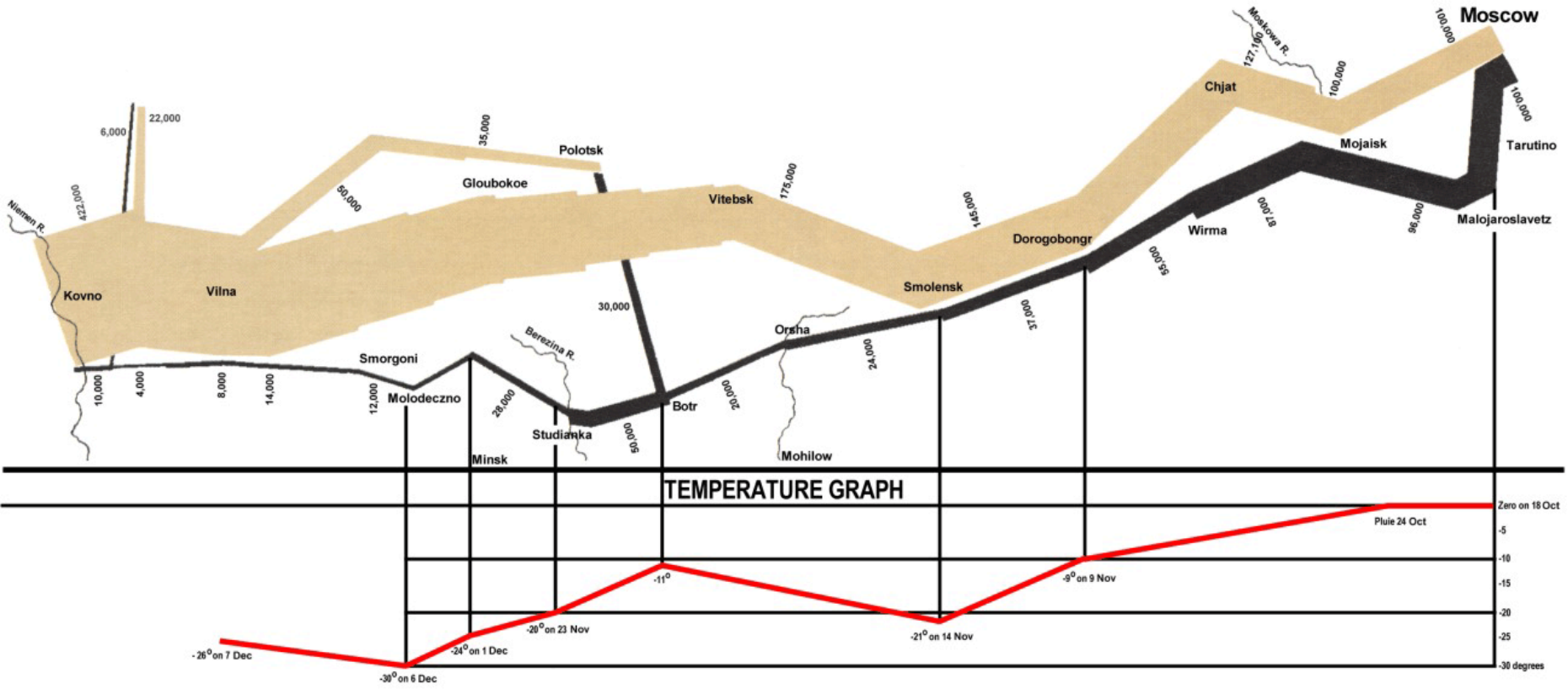
B.T.W., workers at the nearby brewery were noted to be relatively free of cholera…

Explaining

What happened during the Napoleon's Russian Campaign?

The Charles Joseph Minard's map (1861)





Visualization for Making decision

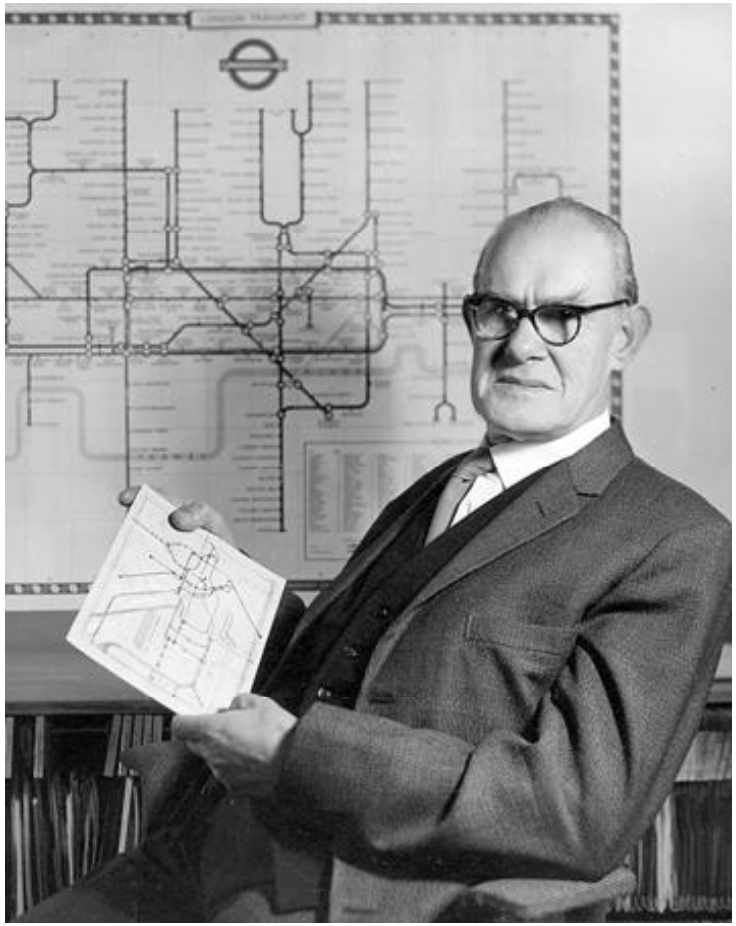
Traveling in London by underground

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…



London Underground Map 1990s

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?

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

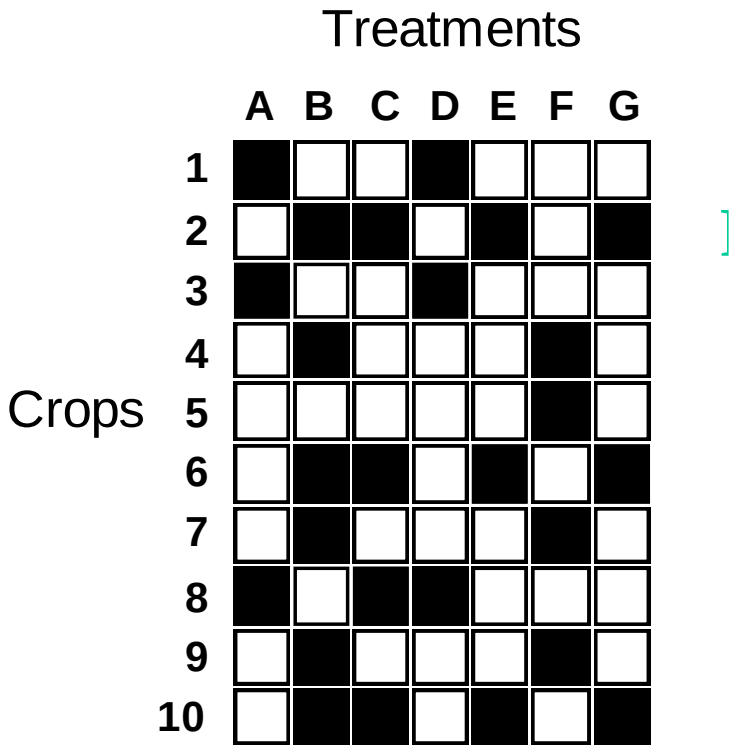
[Card et al. ‘99]

…computer supported and interactive

- **Computer-supported**
  - Even 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

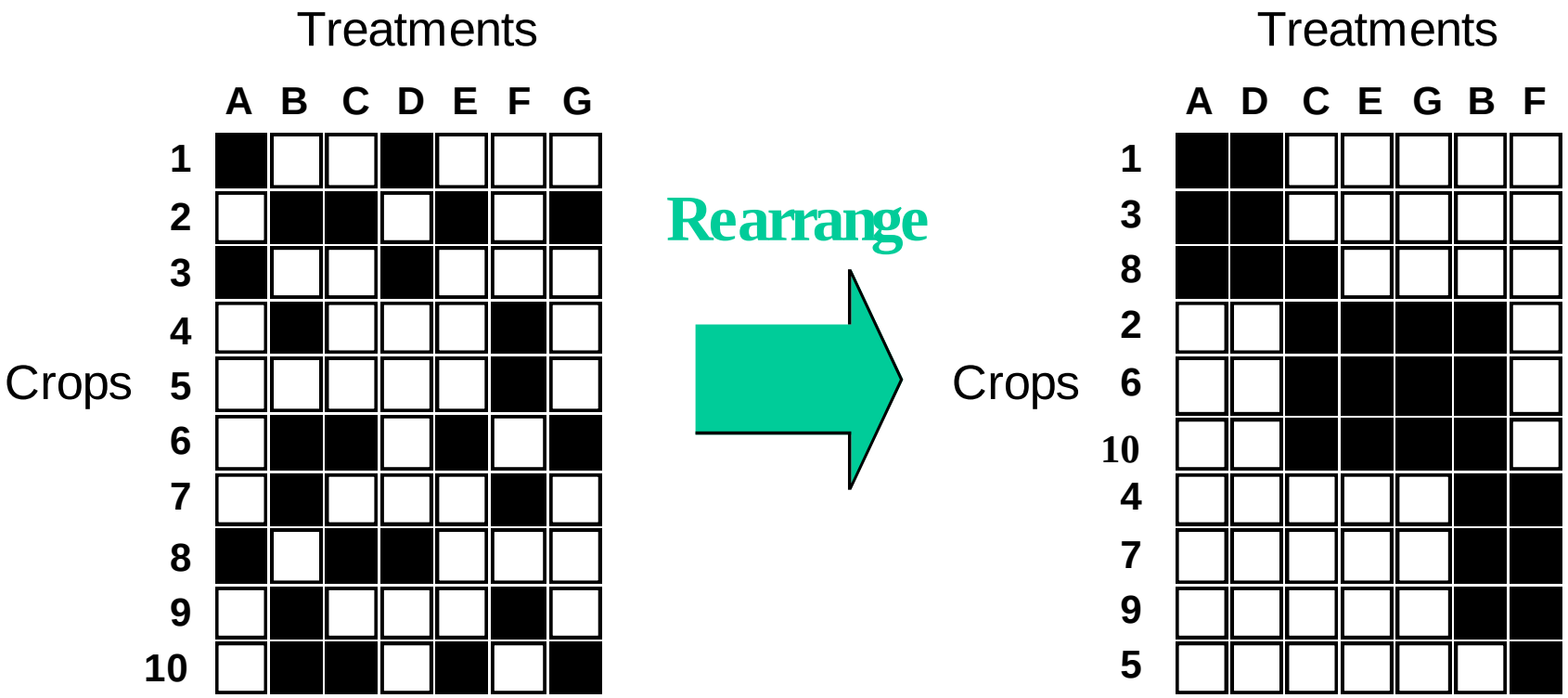
Interaction example

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



Interaction example

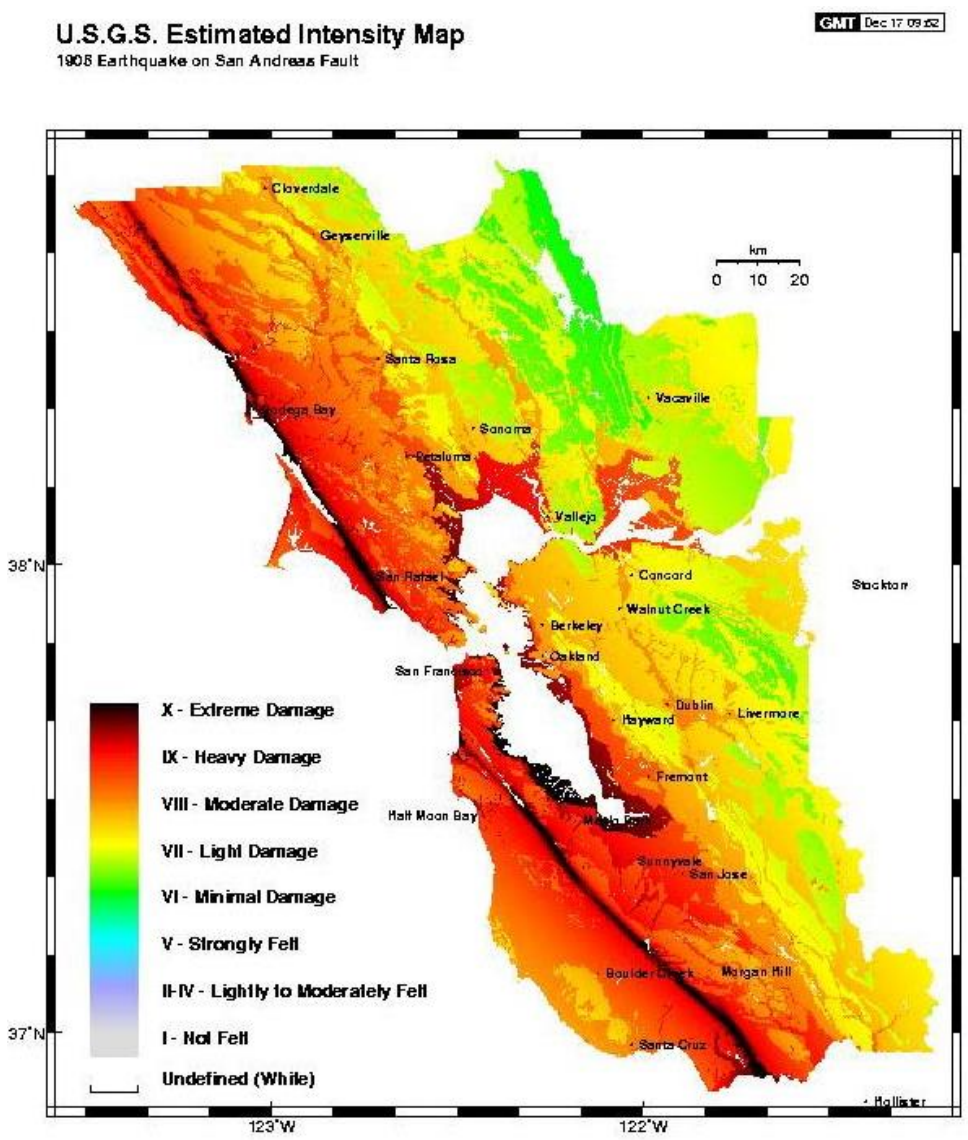
- Let's rearrange the columns



…it is about abstract data

- Abstract data
  - Information visualization is about visualizing abstract data, i.e., presenting images 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

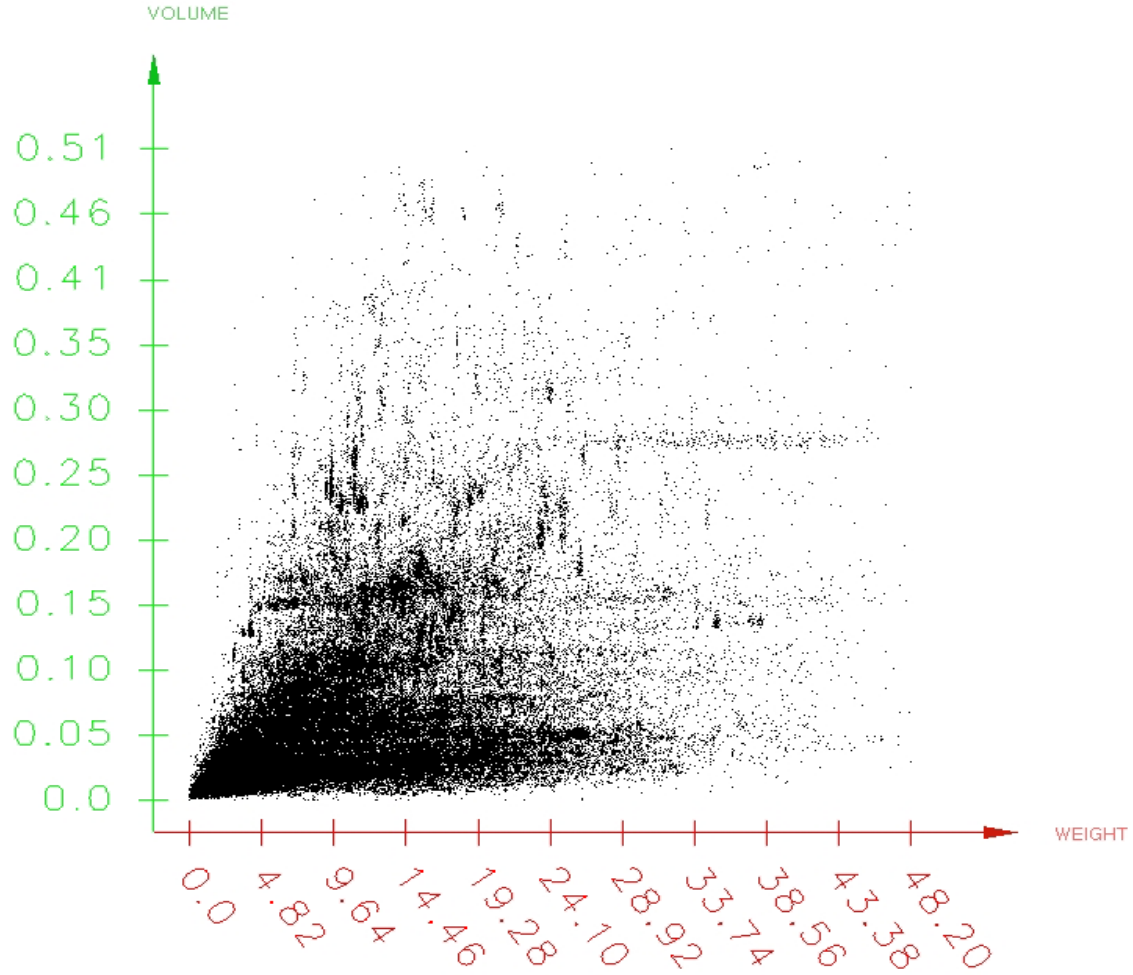


Earthquake intensity

…abstract data

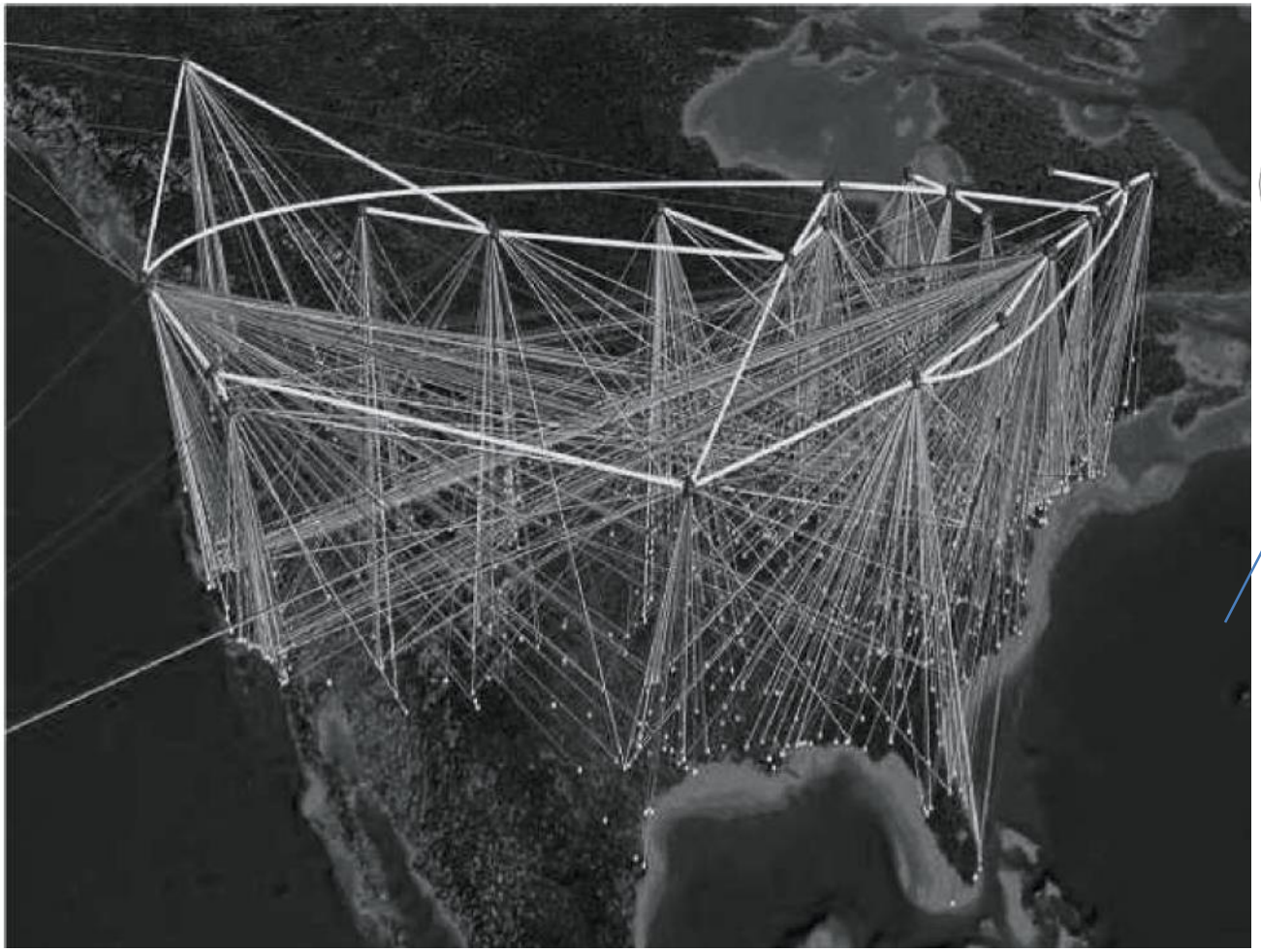
- What is “information”
  - Items that do not have a direct **physical/visual** correspondence (or such a correspondence is not relevant for the application)
  - Examples: sport statistics, stock trends, query results, software data, etc…
- Items are represented on a 2D / 3D physical space using their 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
  - …

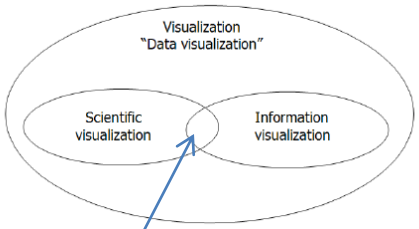
Abstract data



A 2D Scatterplot showing about 200.000 postal parcels

Mixed visualization





Byte traffic into the ANS/NSFNET T3 backbone in 1993

The power of visualization

Three simple questions

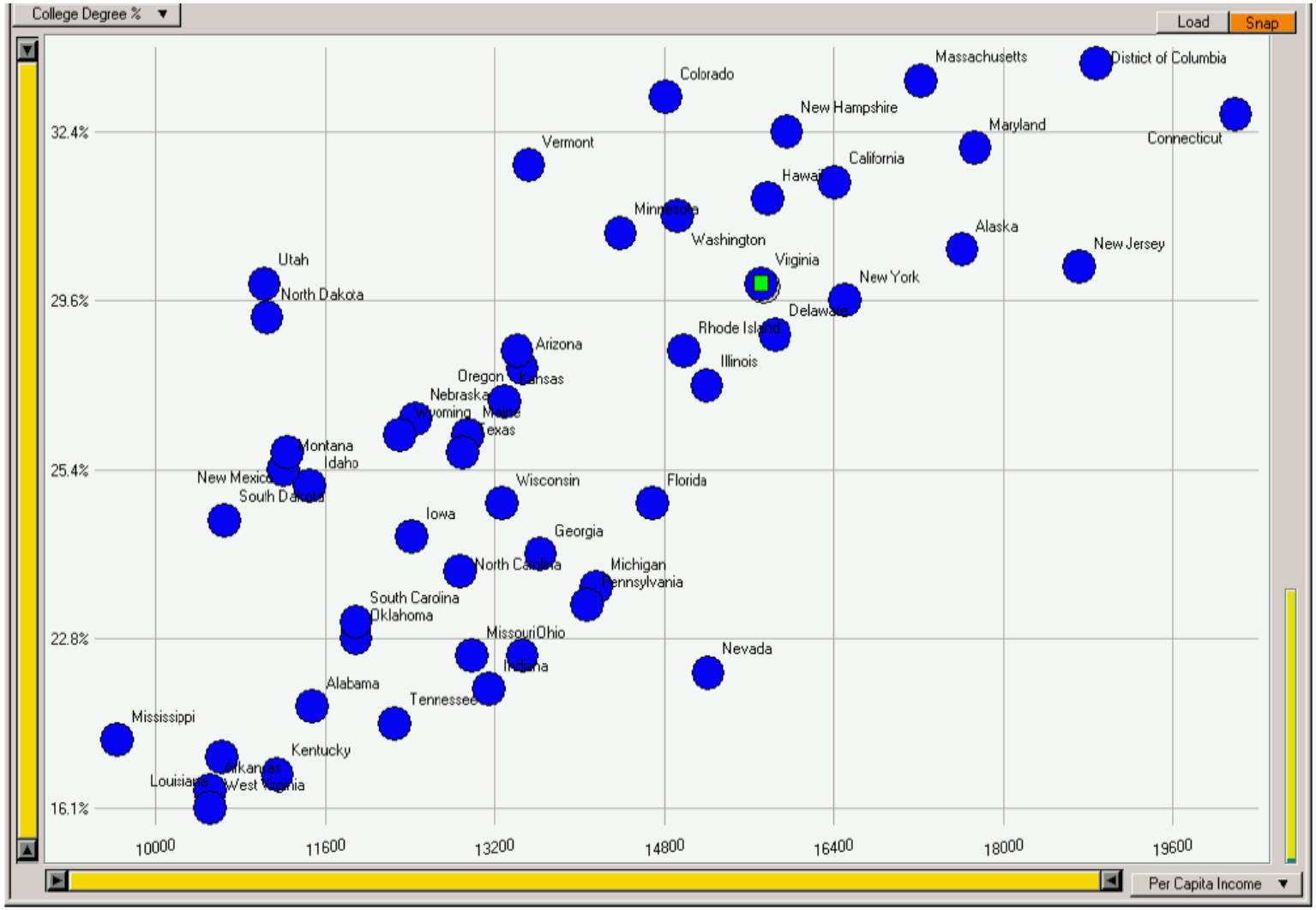
Table - StateData ()

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
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



One (very) simple question

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

Game over!

So ?

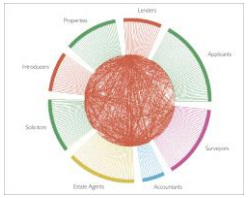
- Time is not enough?
- You can do that in less than one second !

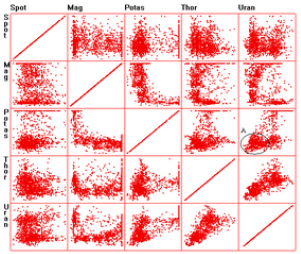
4587576268086099280839826980287479762962862867897187743671947746588786758967329667287682085

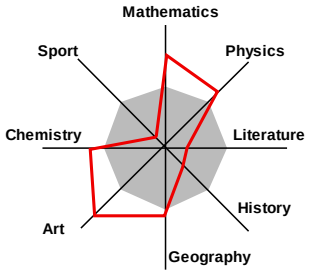
- Color is pre-attentive (pops up)
- No cognitive effort is required
- A lot of issues are already clear
- Most of people ignore them...

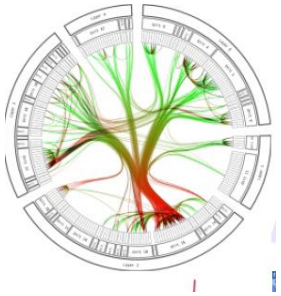
Canonical steps in infovis – STEP 1

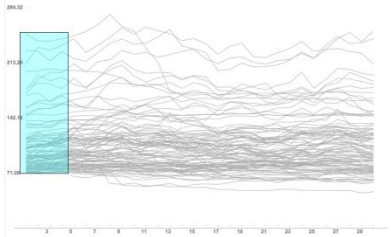
DATA → Internal **Representation**

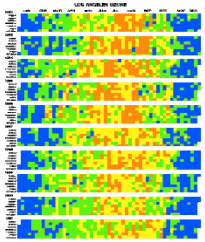


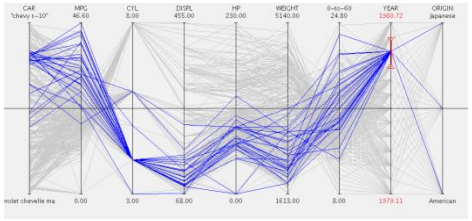












Encoding of values
- Univariate data
- Bivariate data
- Trivariate data
- Multidimensional data

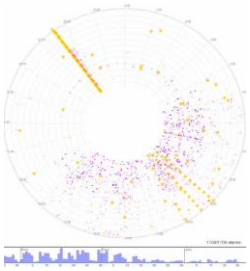
Encoding of relations

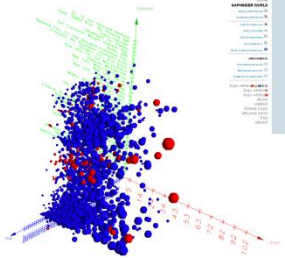
Temporal data

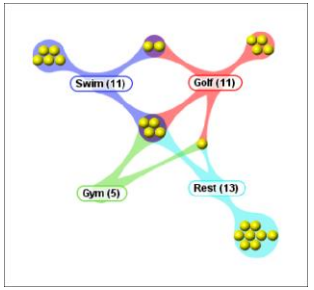
Map & Diagrams

Graphs/Trees

Data streams

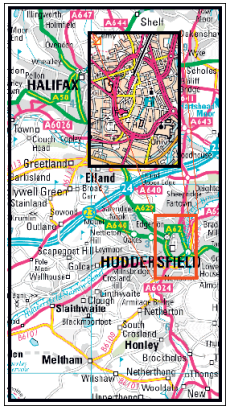


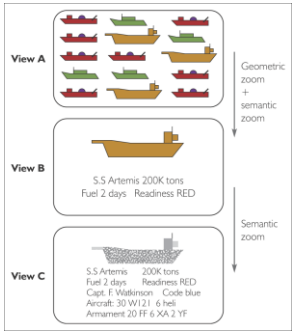




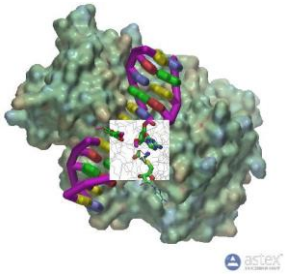
Canonical steps in infovis – STEP 2

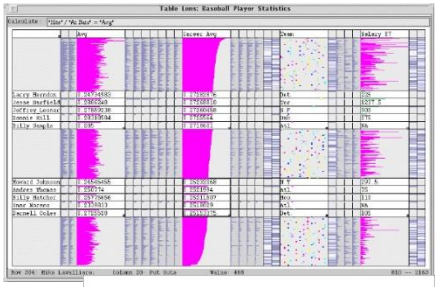
Internal
Representation

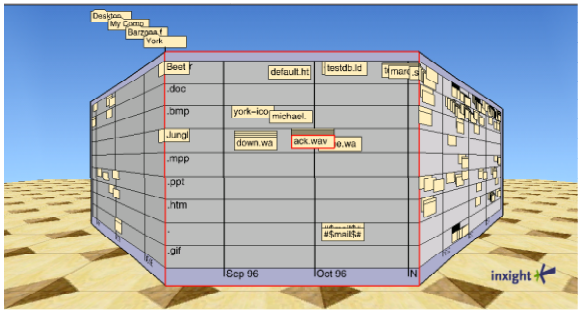


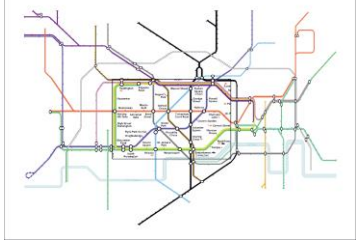


Presentation









- 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 million VISA credit card transactions per day
- 300 million long distance ATT calls per day
- 50 billion e-mails per day
- 600 billion IP packets per day
- 1 trillion (10^{12}) of web pages (according to Google), corresponding to about 3 petabytes of data
- Google processes 20 petabytes of data per day

kilobyte, megabyte, gigabyte, terabyte, petabyte …

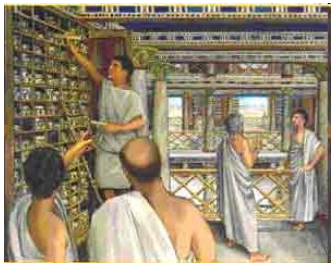
A petabyte ?

- 1 petabyte = 2^{50} = ~ 10^{15} (1000 trillions)

- How big is it?

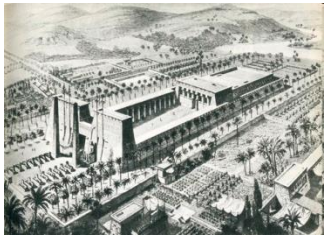
- A quick comparison with one of the worst data loss in the story: the Alexandria library fire (270 a.d. ?)

- 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

$2^{16} * 2^{19} = 2^{35}$ = ~32 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

Rescuing information

•In different situations people need to exploit and to use hidden information resting in unexplored large data sets

- decision-makers
- analysts
- engineers
- emergency response teams
- ...

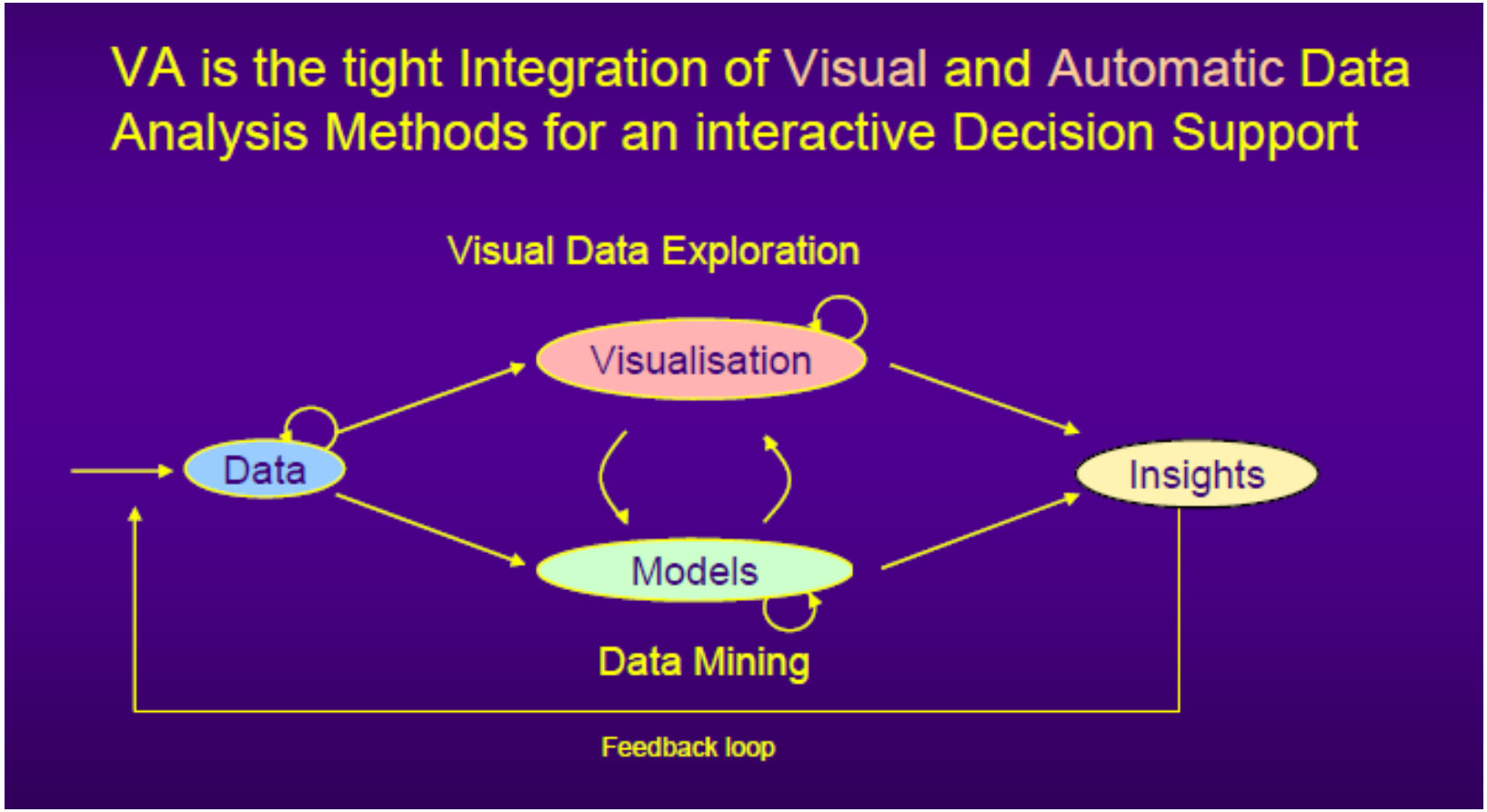
•Several techniques exist devoted to this aim

- Automatic analysis techniques (e.g., data mining)
- Manual analysis techniques (e.g., Information visualization)

•Petabyte datasets require a joint effort:

Visual Analytics

Visual Analytics



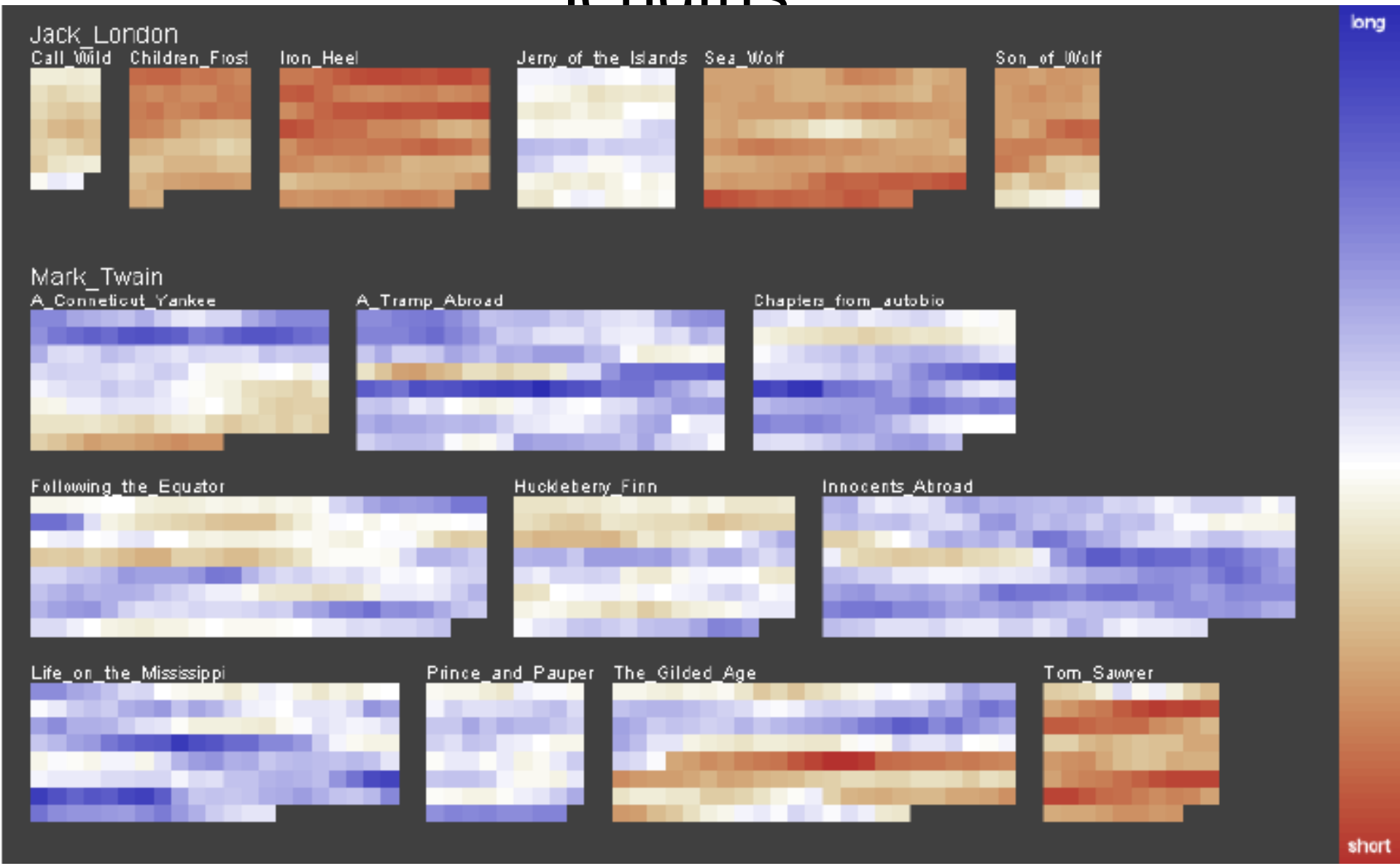
A Visual Analytics example (Group 1)

Deriving new values from the dataset for ad-hoc visualization

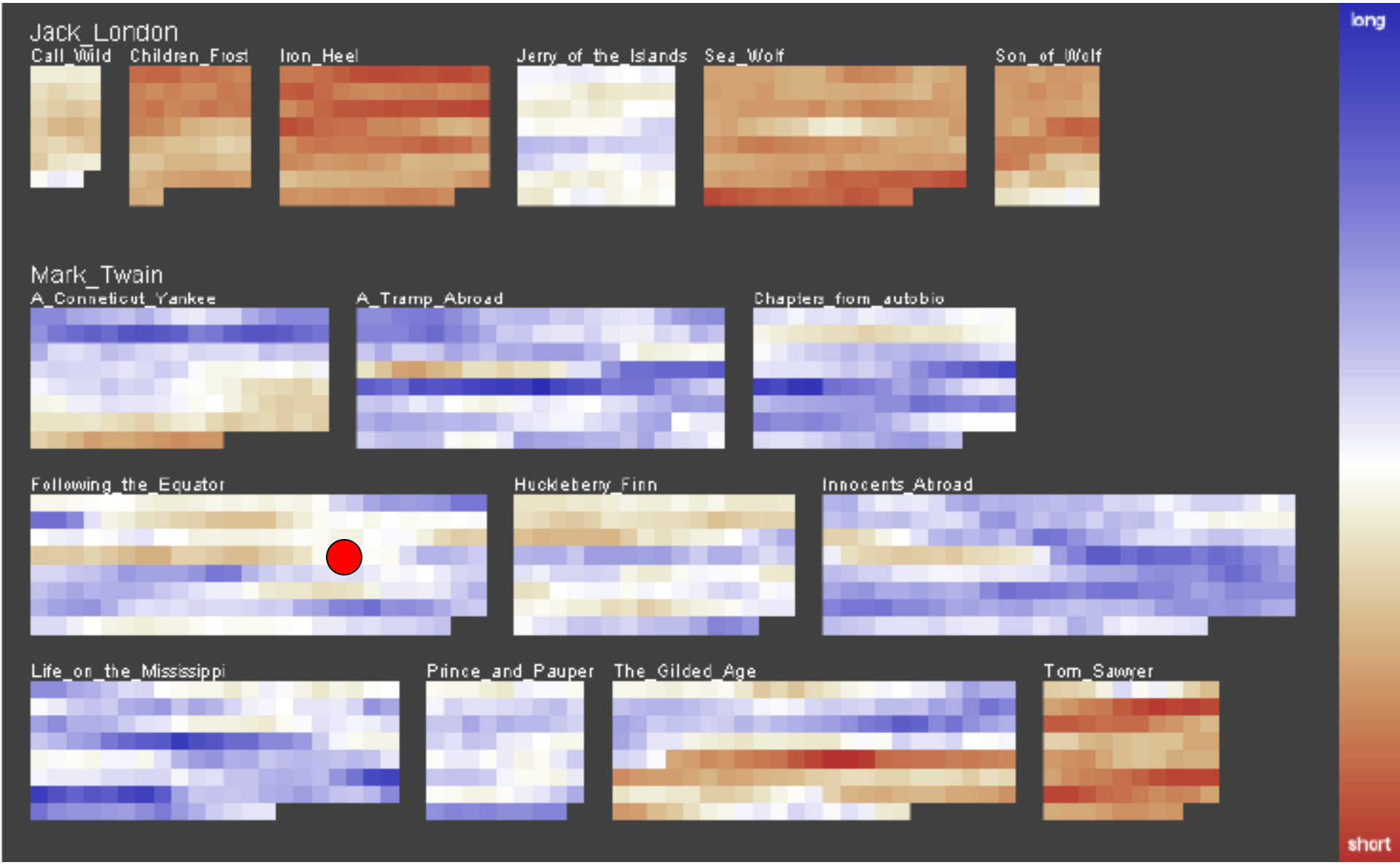
- How to visually compare J. London and M. Twain books ?
- [D. A. Keim and D. Oelke. Literature Fingerprinting: A New Method for Visual Literary Analysis. 2007 IEEE Symp. on Visual Analytics Science and Technology (VAST '07)]

1. Split the book in several text block (e.g., pages, paragraph, sentences)
2. Measure, for each text block, a relevant feature (e.g., average sentence length, word usage, etc.)
3. Associate the relevant feature to a visual attribute (e.g., color)
4. Visualize it

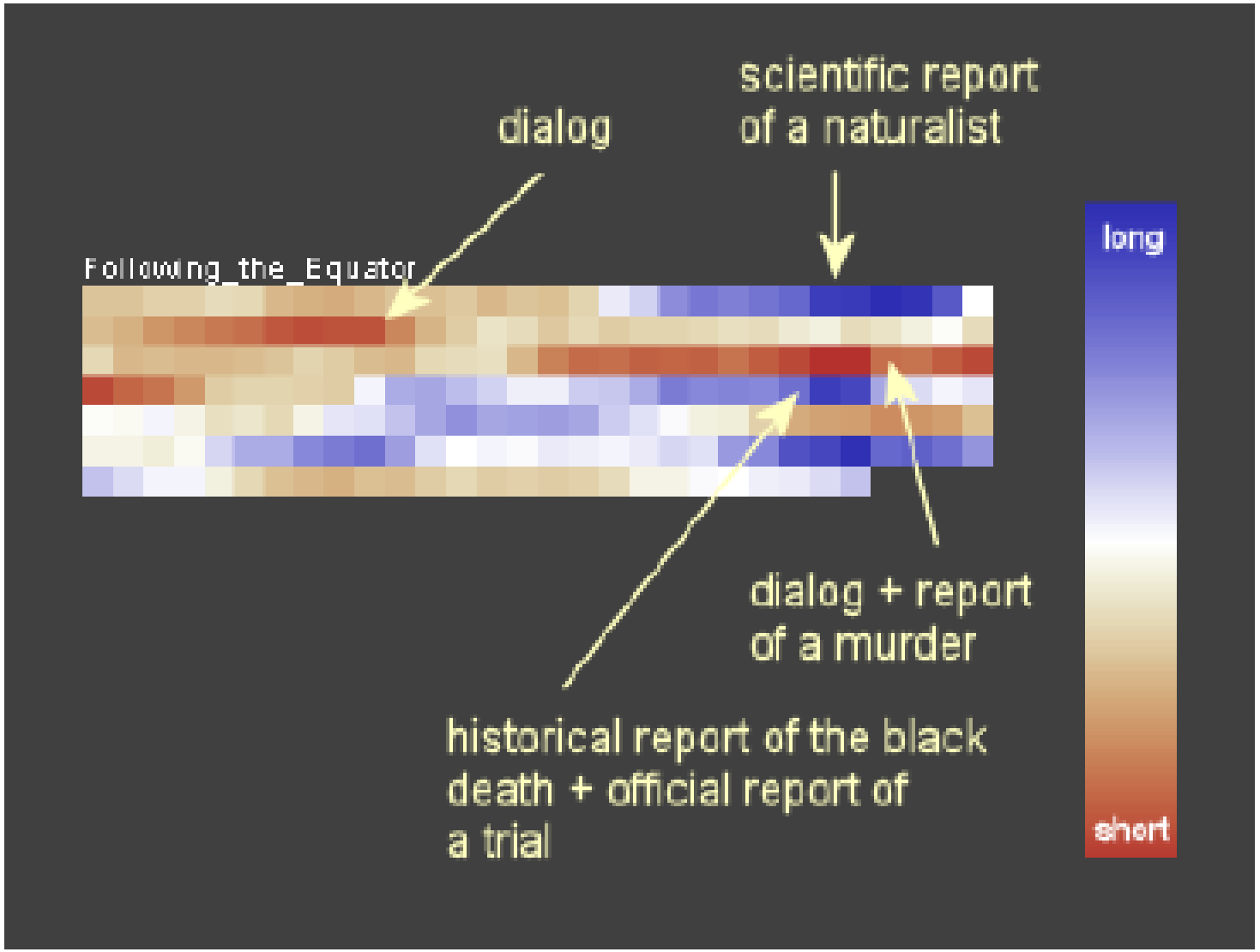
J.London vs M.Twain average sentence lengths



User interaction (a non uniform book?)



Details of a book



What about the Bible?

The Vismaster CA project

Partners

Austria	Danube University Krems Vienna University of Technology VRVis Research Centre
Finland	University of Helsinki
France	INRIA ILOG Business Objects
Germany	Fraunhofer IGD Fraunhofer IAIS University of Konstanz University of Rostock Universität Stuttgart DFKI
Greece	Centre for Research and Technology Hellas
Italy	University of Roma University of Pisa University of Bari
The Netherlands	University of Technology Eindhoven University of Groningen Intl. Institute of Geo-Information Science and Earth Observation
Norway	University of Bergen
Sweden	University of Linköping
Switzerland	University of Zürich University of Fribourg
United Kingdom	Lancaster University City University University of Leeds

Contact

Scientific Coordinator: Prof. Dr. Daniel A. Keim

Address	University of Konstanz Department of Computer and Information Science Box 78 78457 Konstanz, Germany
Phone	+49 7531 88-3161
Fax	+49 7531 88-3062
E-mail	Daniel.Keim@uni-konstanz.de
Web	http://infovis.uni-konstanz.de

Coordinator: **Dr. Jörn Kohlhammer**

Address	Fraunhofer IGD Fraunhoferstr. 5 64283 Darmstadt, Germany
Phone	+49 6151 155-646
Fax	+49 6151 155-139
E-mail	Joern.Kohlhammer@igd.fraunhofer.de
Web	http://www.igd.fraunhofer.de

JOIN THE COMMUNITY!

This project is open to new community participants. You can join one or more of six working groups of interest: data mining, data management, perception and cognition, spatio-temporal analysis, visual analytics infrastructure and evaluation.

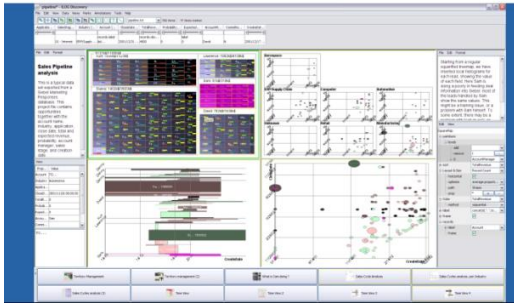
If you are interested to join the European Visual Analytics community, please send an E-mail to Jörn Kohlhammer!

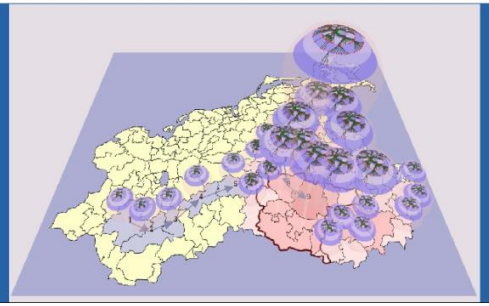
VisMaster

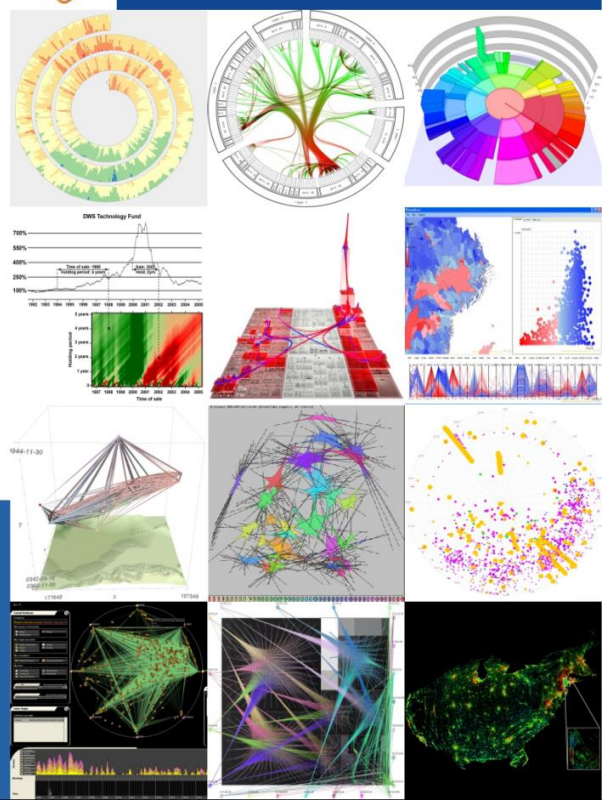
Visual Analytics - Mastering the Information Age

The project VisMaster CA acknowledges the financial support of the Future and Emerging Technologies (FET) Programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number 225429.

http://www.vismaster.eu







The Pamoptesec Project Consortium

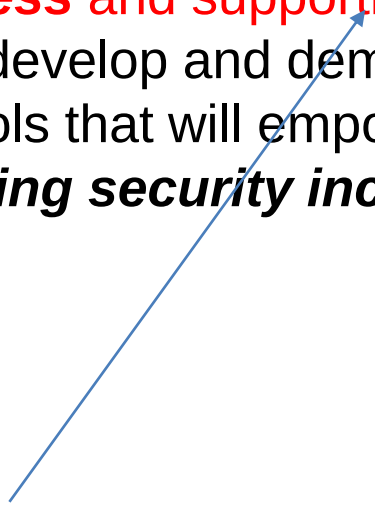
1. Institut Mines-Telecom (IMT) **France** - A group of prestigious **higher education** establishments under the Ministry of Industry will coordinate the overall project
2.RHEA **Belgium** - A software and engineering **consulting company** specialising in space and other cutting-edge technologies, for the technical organisation and monitoring of the technical progress.
3.Alcatel-Lucent Bell Labs **France** - A **research company** of optical components, networks architectures and data analytics,
4.Epistematica **Italy** – A company operating in the **market of IT services** and software applications
5.The Research Center of Cyber Intelligence and Information Security (CIS) **Italy** - Cis belongs to Sapienza **University of Rome**, which is one of the largest and oldest universities in Italy
6.The Hamburg University of Technology (TUHH) **Germany** - A **young university** with a 10-year involvement in technology for ontologies,
7.SUPELEC **France** – One of the most famous and prestigious **higher education institutes** of engineering
8.ACEA **Italy** – **Energy and water provider** in Rome. It will provide the test bed for the project

Objectives

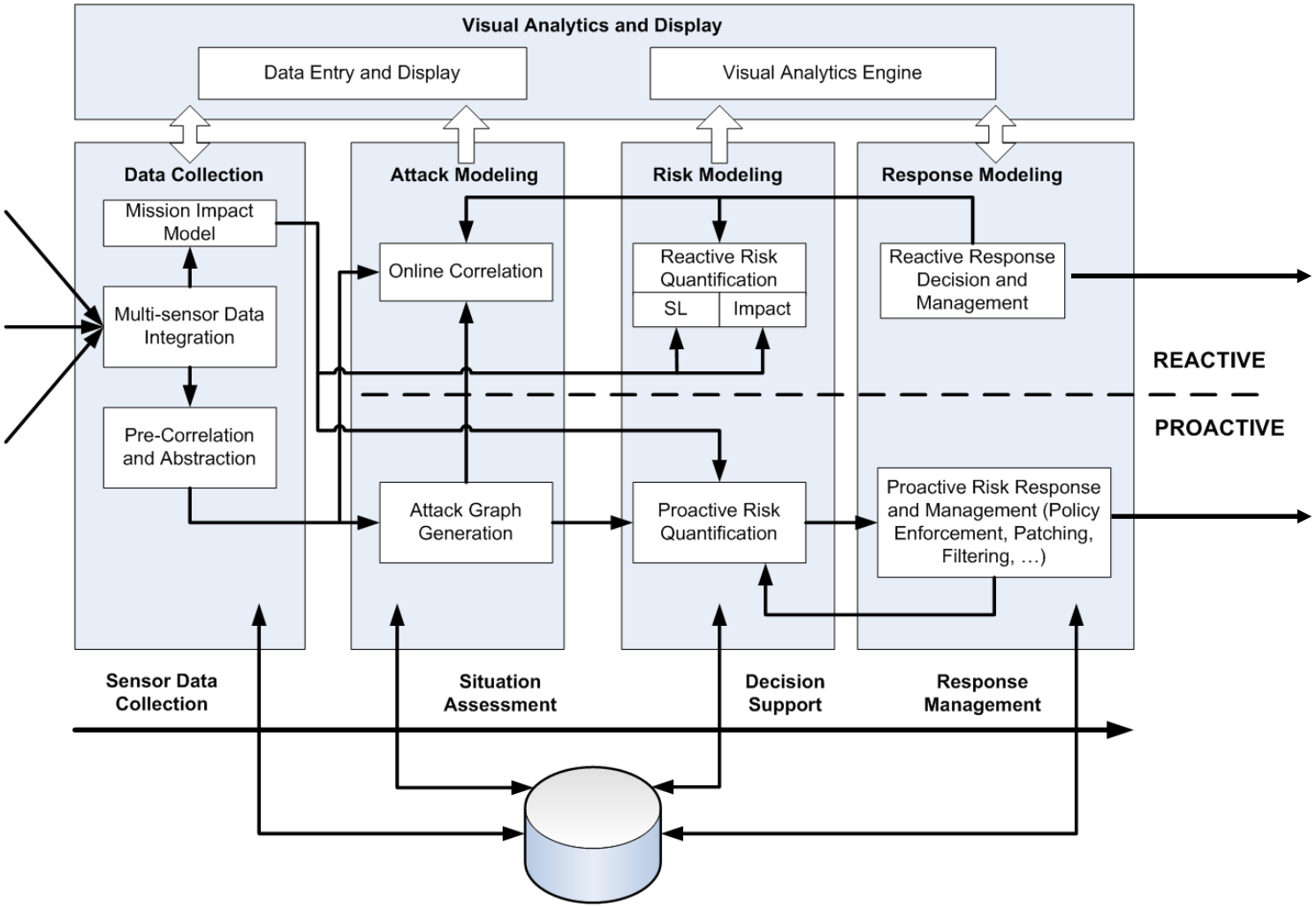
•Objective ICT-2013.1.5 Trustworthy ICT

c) Development, demonstration and innovation in cyber security

"This activity addresses the application of technologies to ***increase the level of cyber security*** in Internet. This includes the development and demonstration of technologies, methodologies and processes to ***prevent, detect, manage and react*** to cyber incidents in real-time, and to support the breach notifications, improving the ***situational awareness*** and supporting the ***decision making*** process. It will also develop and demonstrate advanced technologies and tools that will empower users, notably individuals and SMEs, in ***handling security incidents*** and ***protecting their privacy***."



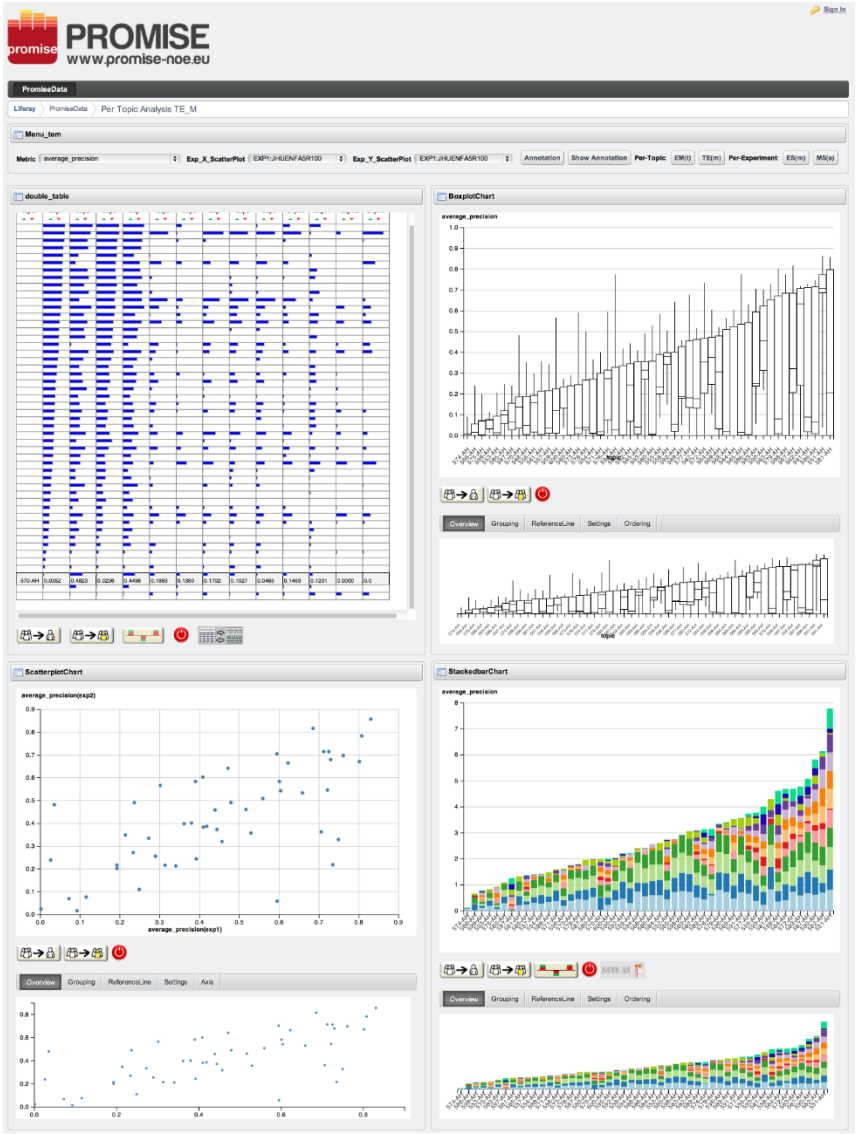
Architecture



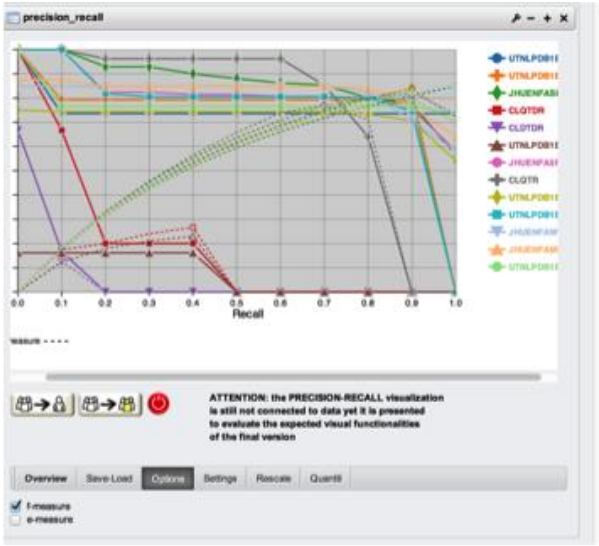
Visual analytics

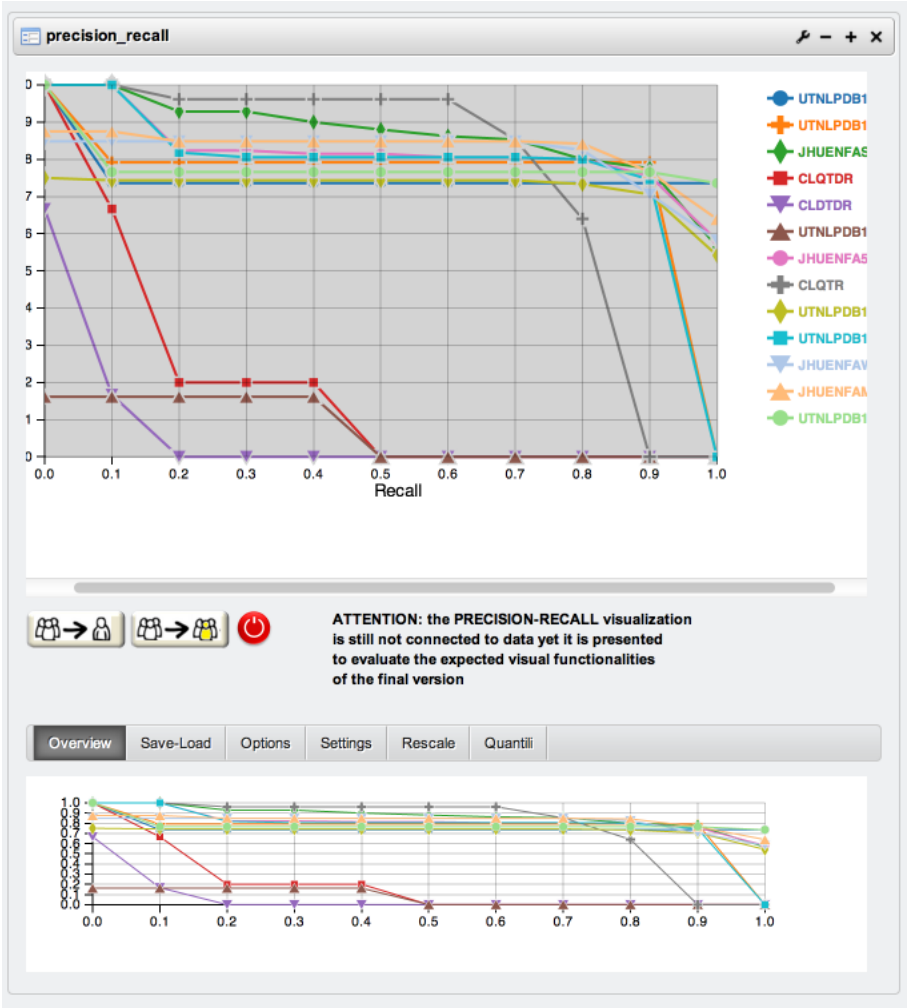
- **Data Entry and Display:** provides the general display environment enabling the operator to interact with the various modules supporting the operational flow.
- **Visual Analytics Engine:** provides the operator with access to all data within the database for deep analysis.

Reusing Promise Results

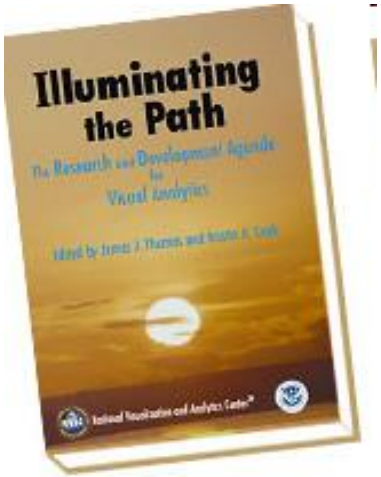


Reusing Promise Results



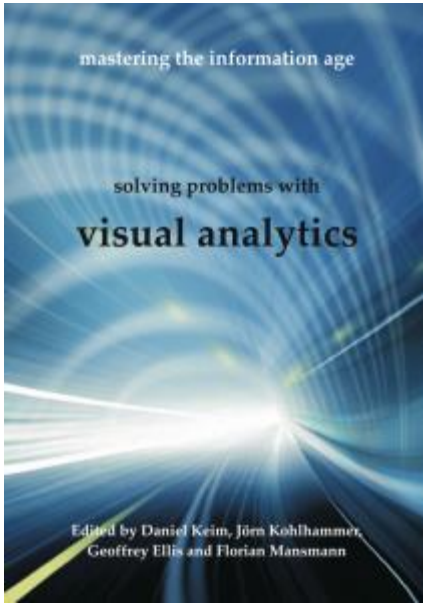


The new (European) book on VA



- Illuminating the path : The Research and Development Agenda for Visual Analytics
  - 2005, focusing on USA homeland security

- Managing the Information Age Solving Problems with Visual Analytics
  - One of the major outcome of Vismaster
  - Availble for free at:
  - http://www.vismaster.eu/



Books worth to read

- Stephen Few - Show me the number - Analytic press
- Stephen Few - Now You See It: Simple Visualization Techniques for Quantitative Analysis - Analytic press
- Robert Spence - Information Visualization: Design for Interaction (2nd Edition) - Addison-Wesley (ACM Press)
- Colin Ware - Information Visualization, Third Edition: Perception for Design (Interactive Technologies) - Morgan Kaufmann
- Managing the Information Age Solving Problems with Visual Analytics (http://www.vismaster.eu/)