Rome, 17th September 2020

Visual analytics for Data Quality

Marco Angelini
angelini@diag.uniroma1.it

A.WA.RE
Advanced **V**isualization & **V**isual **A**nalytics **RE**search group at Sapienza

SAPIENZA
UNIVERSITÀ DI ROMA

whoami

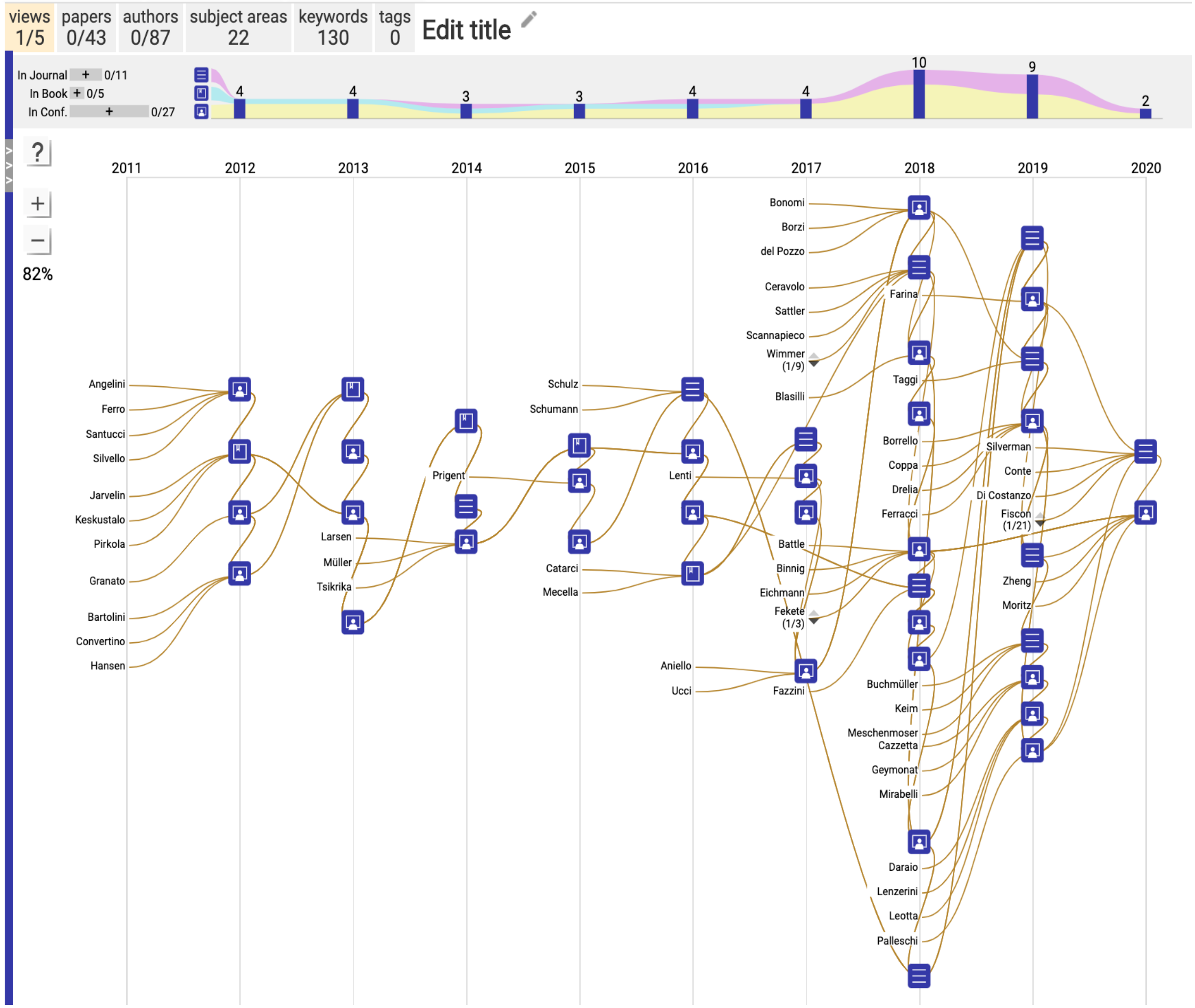
Marco Angelini

Post-doctoral researcher in Predictive Visual Analytics

Mail: angelini@diag.uniroma1.it

Website:

https://sites.google.com/dis.uniroma1.it/angelini



2012

2017

2020

Ph.D.

Post-doctoral position

A (very) simple question

- How many 3s?
- You have 4 seconds to answer……..

Game over!

So ?

- Time was not sufficient?

- You can answer this question in less than 0.2 seconds!
- Let’s try again…

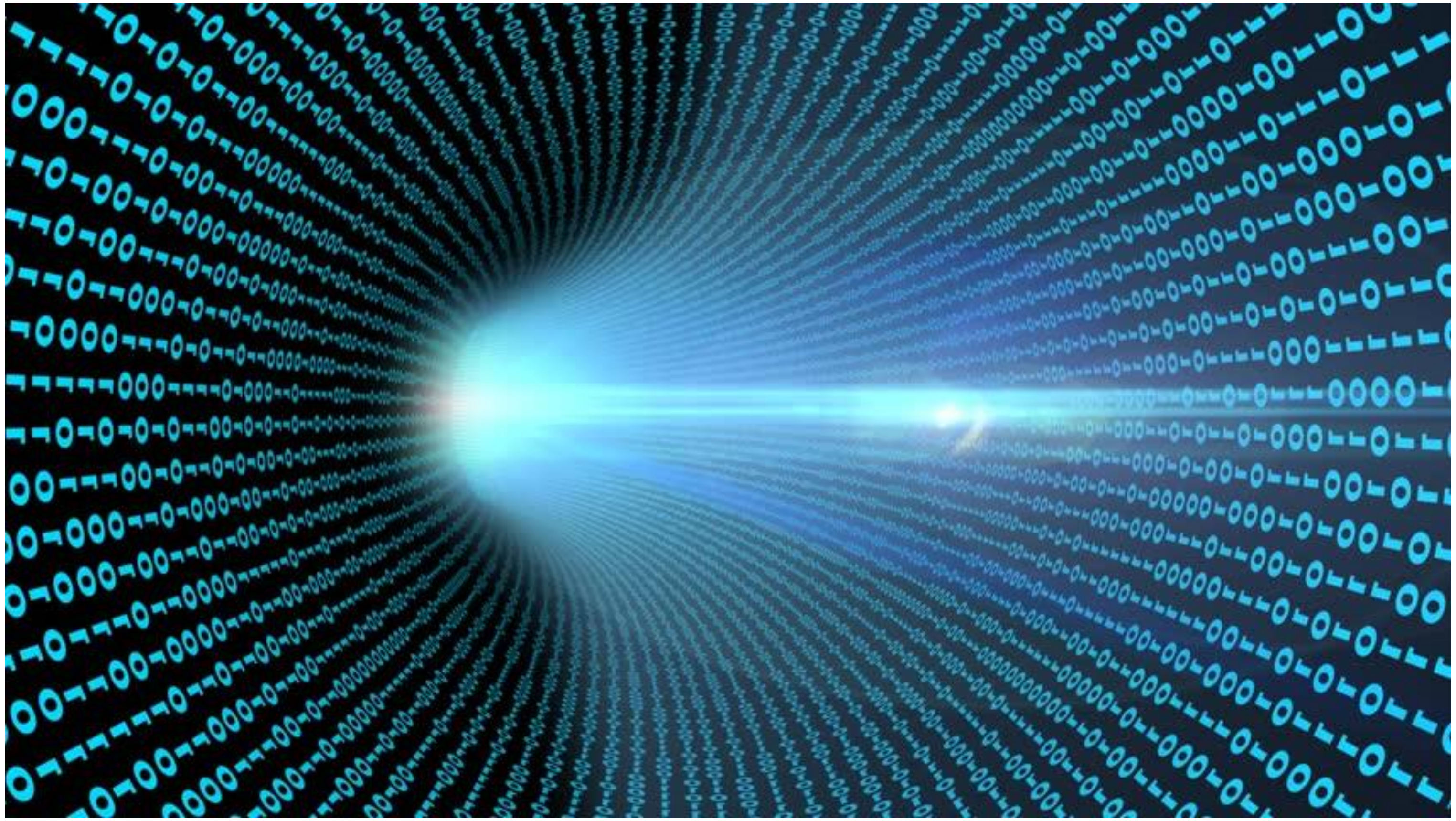
4587576268086099280839826980 28

7479762962628678971877436719 47

7465887867589673296672876820 85

- Color is pre-attentive(it pops up)
- It does not require any cognitive effort

Lots of Data!



Ben Chams - Fotolia

data
quality
management
reporting
Data
staff
place
recording
procedures
processes
collection
reviewed
updates
needs
covering
information
responsibility
may
support
top
changes
working
supported
results
level
defined
associated
clearly
commitment
needed
service
regularly
body's
arrangements
external
standards
governance
taken
senior
corporate
example
ensure
body
basis
set
training
used
agreement
Systems
necessary
delivery
form
reported
charged
complete
requirements
risk
address
guidance
programme
action
business
Policies
provision
regular
policy
timely
analysis
systems
periodically
monitoring
assessed
day
objectives
practice
relevant
manipulation
clear
work
policies

Data Quality

- Understanding the gross structure of the datasets (how many columns, how many rows, etc.)

- How big is the dataset, how many attributes, how is the data organized, etc?

- Internalizing the dataset attributes (columns)
  - what type of data is in each column?
  - Is it categorical, quantitative, and ordinal, etc?
  - What are the most frequent values?

- Discovering relationships among the attributes and structure within the table
  - how are the columns related?
  - Are there duplications among the columns, implicit relationships, and implicit structure within the table?

Data Quality

- Finding invalid and missing values,
  - Invalid values occur when items are miss-keyed, when data is carelessly entered, or when data is inconsistently collected.
  - Missing values occur when data attributes are dropped as part of the data extraction process, important fields are ignored and not populated by data entry clerks, or when data tables are expanded as part of system maintenance but never populated.

- Discovering zeros and other suspicious values such as 99 or 99999. These values are often indicative of coding problems in the data collection process and may require manual investigation.

- Identifying duplicated rows and column. Errors in data extraction routines often manifest themselves by

Data Quality

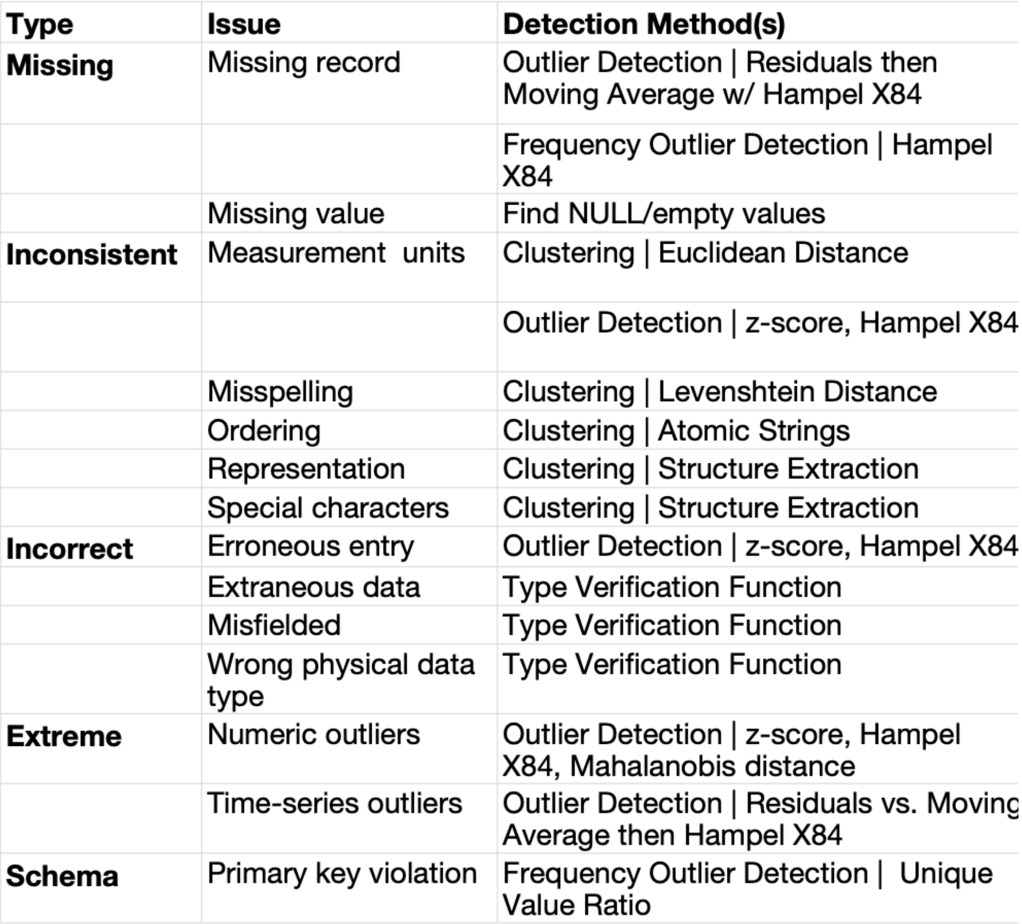
Data profiling

Data Quality Measurement

Data cleansing

Data Quality Monitoring

- Accuracy
- Completeness
- Coherence
- Relevance
- Timeliness
- Accessibility
- interpretability



Type	Issue	Detection Method(s)
Missing	Missing record	Outlier Detection | Residuals then Moving Average w/ Hampel X84
		Frequency Outlier Detection | Hampel X84
	Missing value	Find NULL/empty values
Inconsistent	Measurement units	Clustering | Euclidean Distance
		Outlier Detection | z-score, Hampel X84
	Misspelling	Clustering | Levenshtein Distance
	Ordering	Clustering | Atomic Strings
	Representation	Clustering | Structure Extraction
	Special characters	Clustering | Structure Extraction
Incorrect	Erroneous entry	Outlier Detection | z-score, Hampel X84
	Extraneous data	Type Verification Function
	Misfielded	Type Verification Function
	Wrong physical data type	Type Verification Function
Extreme	Numeric outliers	Outlier Detection | z-score, Hampel X84, Mahalanobis distance
	Time-series outliers	Outlier Detection | Residuals vs. Moving Average then Hampel X84
Schema	Primary key violation	Frequency Outlier Detection | Unique Value Ratio

.

You saw a lot of it during these days….

How can an analyst be helped ?

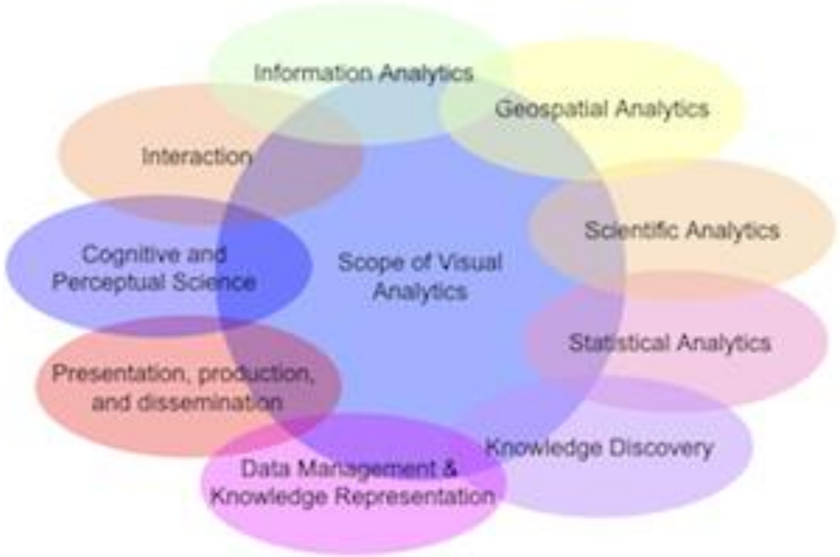
- Managing complexity of this workflow
- Some of the indicator are easy to compute but requires explanation to a user
- Some of the indicators need analyzing data in detail and recognize the behavior
- Exploring data require good skills

Visual Analytics: definition

Visual Analytics is the science of analytical reasoning supported by interactive visual interfaces.
the complex nature of many problems makes it indispensable to include human intelligence at an early stage in the data analysis process.

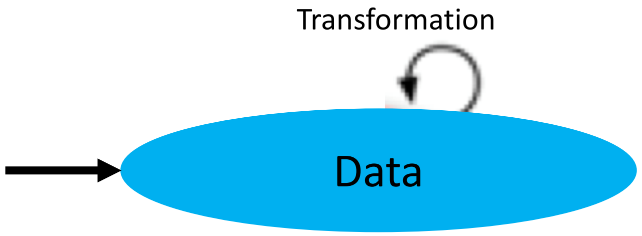
Visual Analytics methods allow decision makers to combine their human flexibility, creativity, and background knowledge with the enormous storage and processing capacities of today's computers to gain insight into complex problems.

Using advanced visual interfaces, humans may directly interact with the data analysis capabilities of today's computer, allowing them to make well-informed decisions in complex situations.

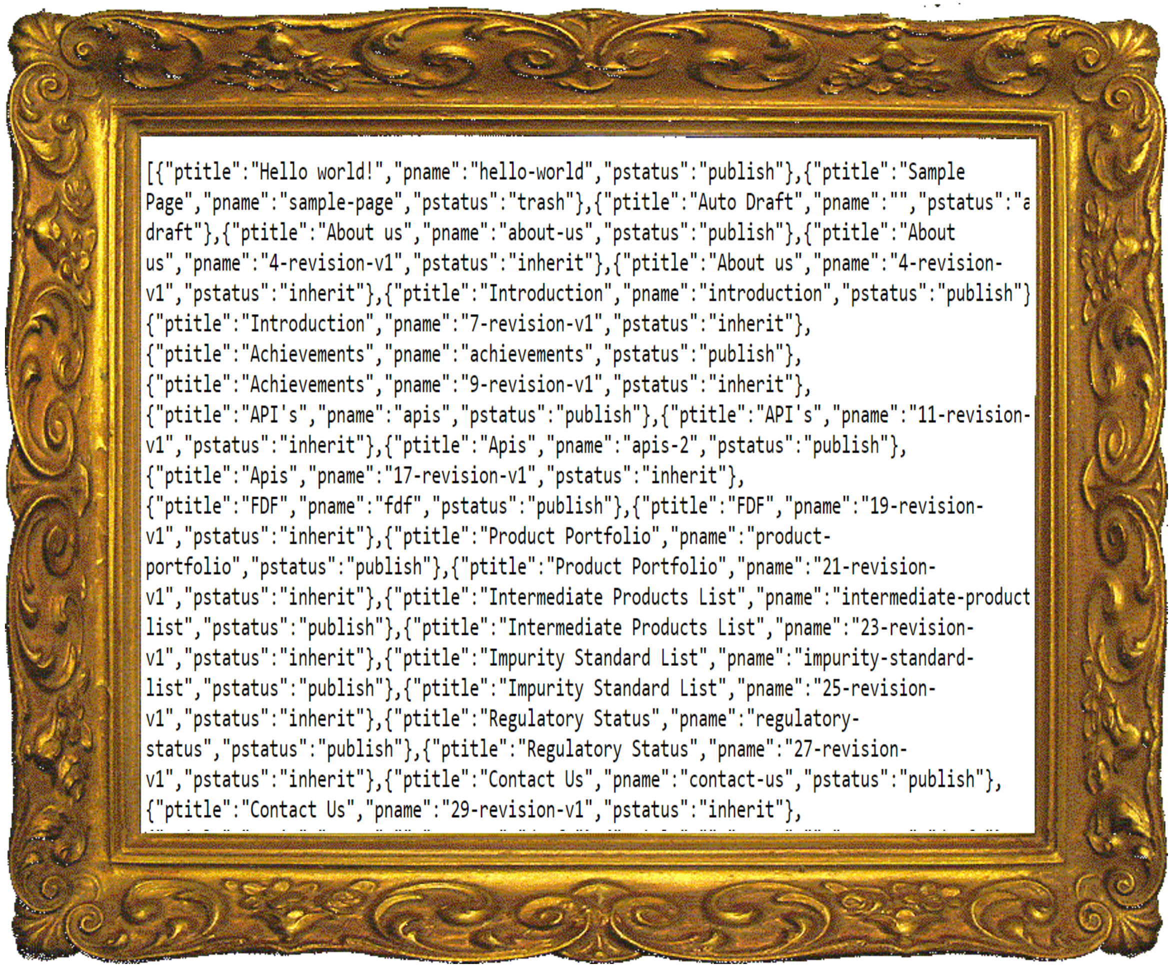

Information Analytics
Geospatial Analytics
Interaction
Scientific Analytics
Cognitive and Perceptual Science
Scope of Visual Analytics
Statistical Analytics
Presentation, production, and dissemination
Knowledge Discovery
Data Management & Knowledge Representation

Thomas, J., Cook, K.: Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE-Press (2005)

Visual Analytics

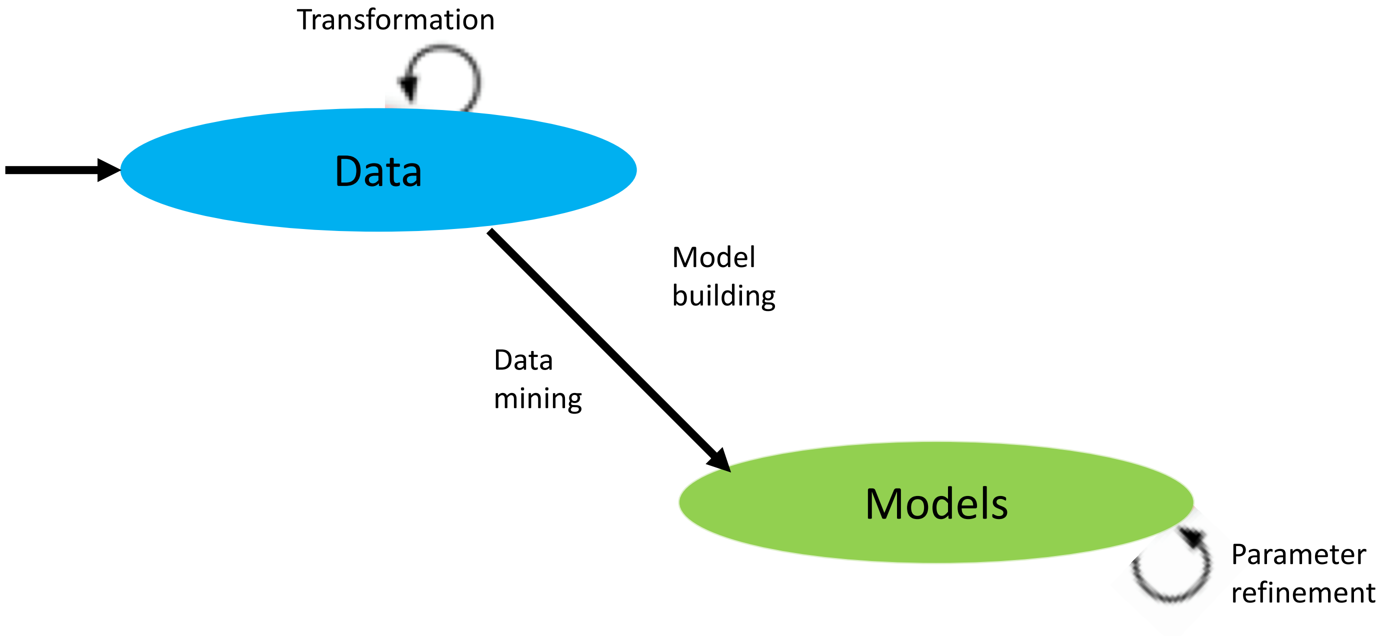


Data are ~~beautiful~~ ugly



```
[{"ptitle":"Hello world!","pname":"hello-world","pstatus":"publish"},{"ptitle":"Sample
Page","pname":"sample-page","pstatus":"trash"},{"ptitle":"Auto Draft","pname":"","pstatus":"a
draft"},{"ptitle":"About us","pname":"about-us","pstatus":"publish"},{"ptitle":"About
us","pname":"4-revision-v1","pstatus":"inherit"},{"ptitle":"About us","pname":"4-revision-
v1","pstatus":"inherit"},{"ptitle":"Introduction","pname":"introduction","pstatus":"publish"}
{"ptitle":"Introduction","pname":"7-revision-v1","pstatus":"inherit"},
{"ptitle":"Achievements","pname":"achievements","pstatus":"publish"},
{"ptitle":"Achievements","pname":"9-revision-v1","pstatus":"inherit"},
{"ptitle":"API's","pname":"apis","pstatus":"publish"},{"ptitle":"API's","pname":"11-revision-
v1","pstatus":"inherit"},{"ptitle":"Apis","pname":"apis-2","pstatus":"publish"},
{"ptitle":"Apis","pname":"17-revision-v1","pstatus":"inherit"},
{"ptitle":"FDF","pname":"fdf","pstatus":"publish"},{"ptitle":"FDF","pname":"19-revision-
v1","pstatus":"inherit"},{"ptitle":"Product Portfolio","pname":"product-
portfolio","pstatus":"publish"},{"ptitle":"Product Portfolio","pname":"21-revision-
v1","pstatus":"inherit"},{"ptitle":"Intermediate Products List","pname":"intermediate-product
list","pstatus":"publish"},{"ptitle":"Intermediate Products List","pname":"23-revision-
v1","pstatus":"inherit"},{"ptitle":"Impurity Standard List","pname":"impurity-standard-
list","pstatus":"publish"},{"ptitle":"Impurity Standard List","pname":"25-revision-
v1","pstatus":"inherit"},{"ptitle":"Regulatory Status","pname":"regulatory-
status","pstatus":"publish"},{"ptitle":"Regulatory Status","pname":"27-revision-
v1","pstatus":"inherit"},{"ptitle":"Contact Us","pname":"contact-us","pstatus":"publish"},
{"ptitle":"Contact Us","pname":"29-revision-v1","pstatus":"inherit"},
```

Visual Analytics



Variables and indicators

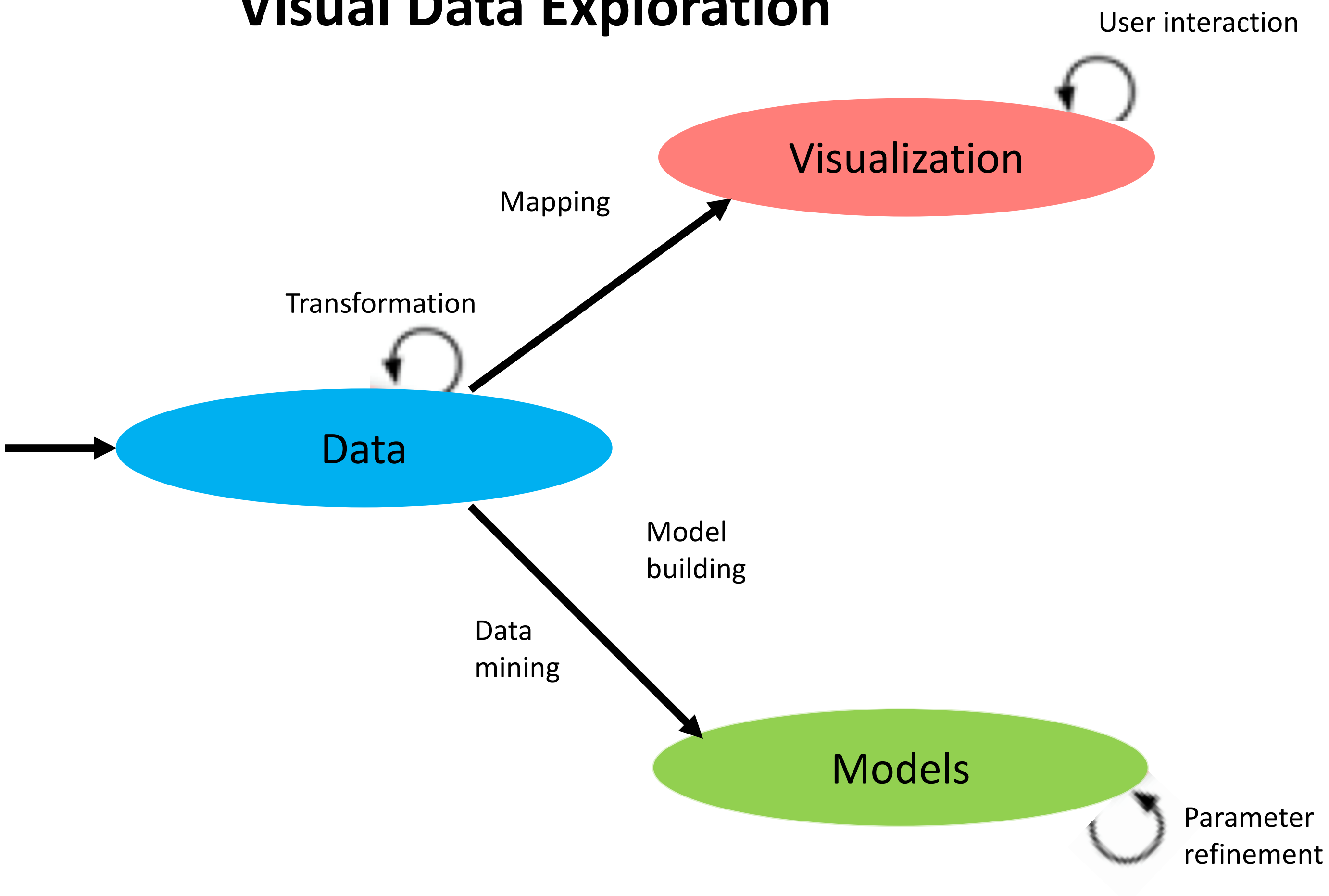
The Analytics

Models

Innovation			Education			Research		
INPUT	**OUTPUT**	**ENV. VAR.**	**INPUT**	**OUTPUT**	**ENV. VAR.**	**INPUT**	**OUTPUT**	**ENV. VAR.**
% pop with higher education	GDP per capita	Foundation year	Total number of enrolled students ISCED5-8	Graduates at ISCED 5-7 (national, foreign and total graduates)	University hospital	Total academic staff (Full Time Equivalent)	Total number of documents published in scholarly journals indexed in Scopus	Ph.D. intensity (students ISCED8/students ISCED5-8)
business R&D exp	patent number	Region of establishment (NUTS3; NUTS 2 and country)	Total academic staff (Full Time Equivalent)	Graduates at ISCED 5-7 area F09 (medicine)	Ph.D. intensity (students ISCED8/students ISCED5-8) in FoE 09 Medicine	Total academic staff (HC)	Normalized Impact	Ph.D. intensity (students ISCED8/students ISCED5-8) in FoE 09 Medicine
% pop lifelong learning activities	revenues		Students enrolled at ISCED 5-7 (national, foreign and total students)	Graduates at ISCED 8—area F 09 (medicine)	Ratio foreign/national students ISCED5-7	Academic staff—ISCED-F 09 (HC)	High Quality Publications Ratio of publications that an institution publishes in the most influential scholarly journals of the world	Total students enrolled/Total academic staff (HC)
high-tech empl in manuf	added value of high-tech industries		Students enrolled at ISCED 8—(distinguished in national, foreign and total students enrolled)		Ph.D. intensity (students ISCED8/students ISCED5-8)	Number of administrative staff (FTE)	Excellence rate indicates the amount (in %) of an institution's scientific output that is included into the set of the 10% of the most cited papers	

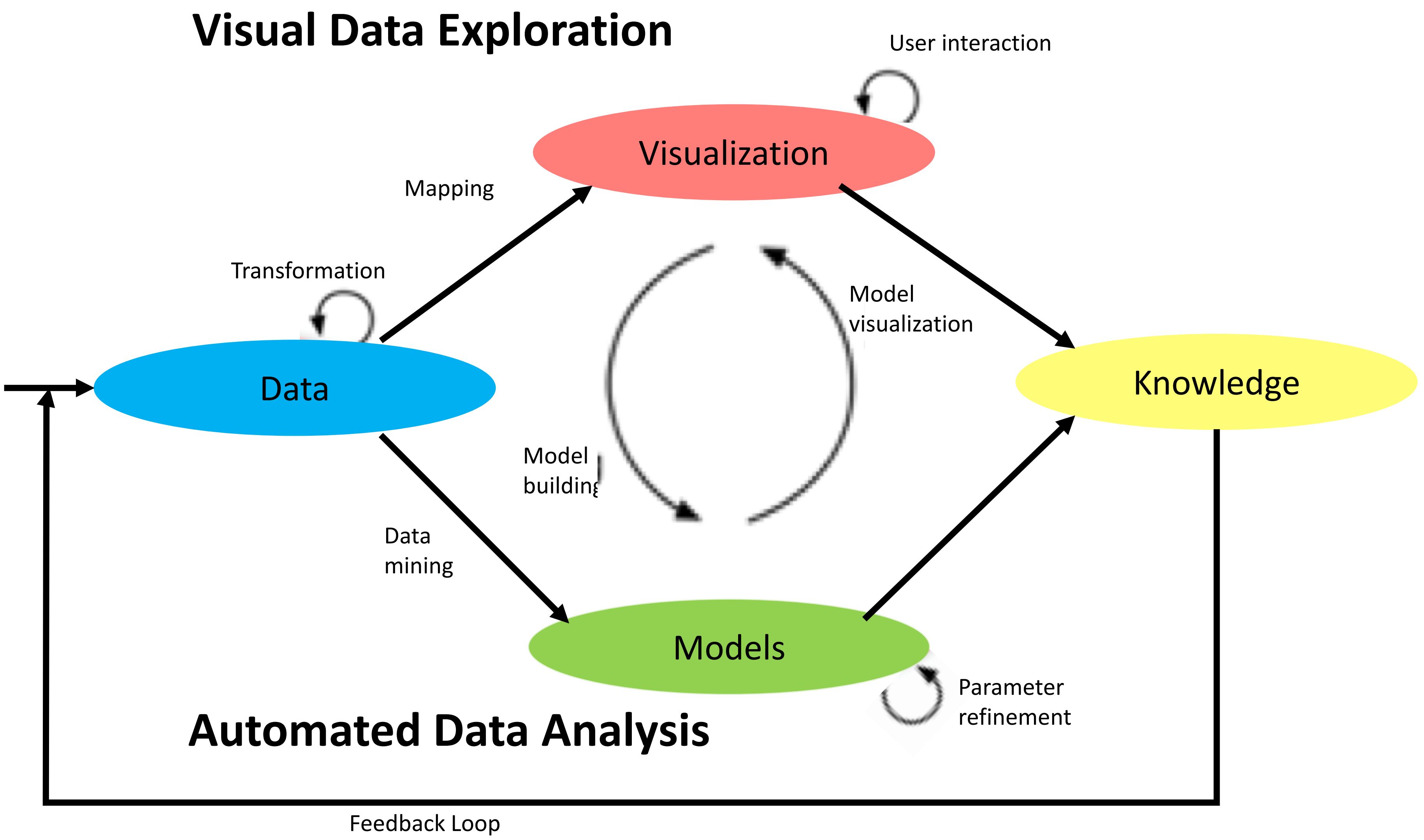
Visual Analytics

Visual Data Exploration



Visualization

Visual Analytics



Data Quality: a survey

667 software tools dedicated to “data quality” (still emerging market)

- half (50.82 %) of the DQ tools were domain specific, which means they were either dedicated to specific types of data or built to measure the DQ of a proprietary tool.
- 16.67 % of the DQ tools focused on data cleansing without a proper DQ measurement strategy
- Most surveyed tools supported data profiling to some extent
- did not find a tool that implements a wider range of DQ metrics for the most important DQ dimensions as proposed in research.
- Identified metric implementations have several drawbacks: some are only applicable on attribute-level (e.g., no aggregation), some require a gold standard that might not exist, and some have implementation errors.

Lisa Ehrlinger1, 2, Elisa Rusz1 , and Wolfram Wöß1 ,A SURVEY OF DATA QUALITY MEASUREMENT AND MONITORING TOOLS, Preprint, 2019

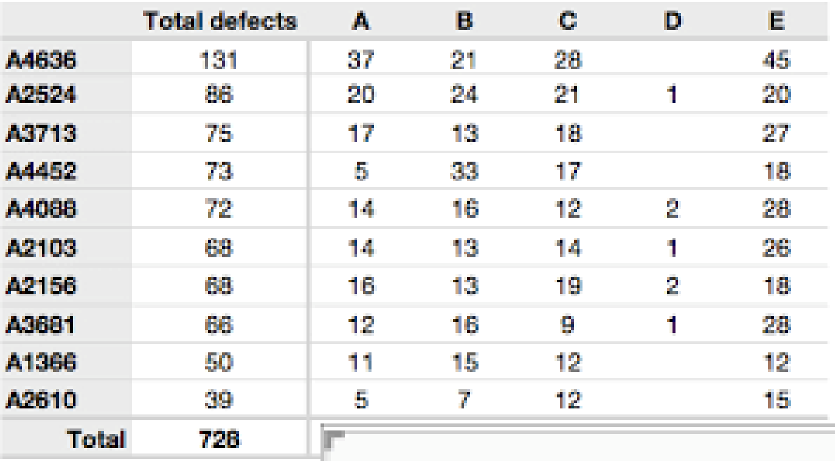
Data Quality & visualization: survey

- Very low presence of visual environments (in contrast with other “fields”)

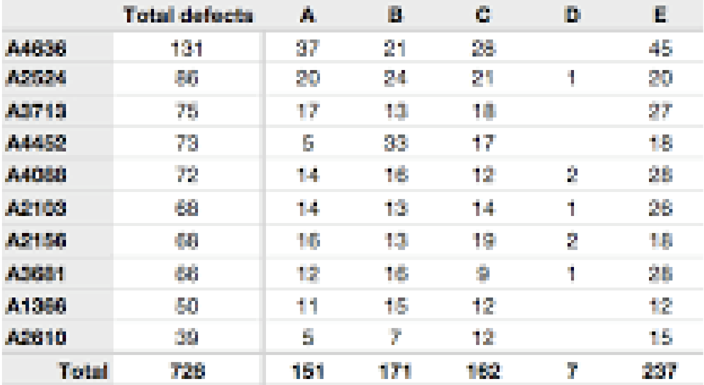
- the authors list 9 usability criteria for the GUI, but in the evaluation they only distinguish between (g) representing “not user friendly GUI” and a (G) for “user-friendly GUI” with drag and drop functionality.

Lisa Ehrlinger1, 2, Elisa Rusz1 , and Wolfram Wöß1 ,A SURVEY OF DATA QUALITY MEASUREMENT AND MONITORING TOOLS, Preprint, 2019

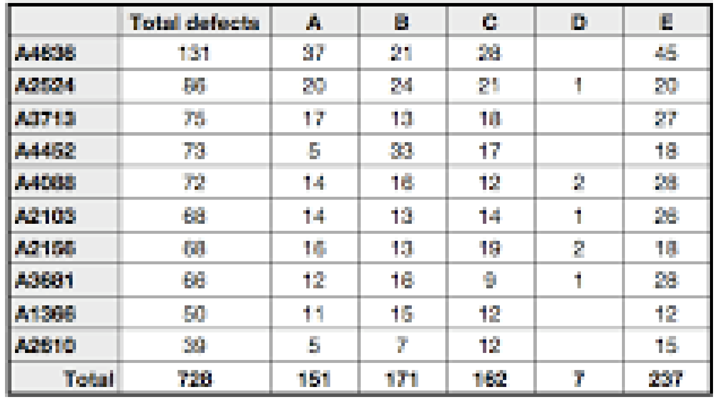
Visualization 4 Data Quality: Tables are kings…



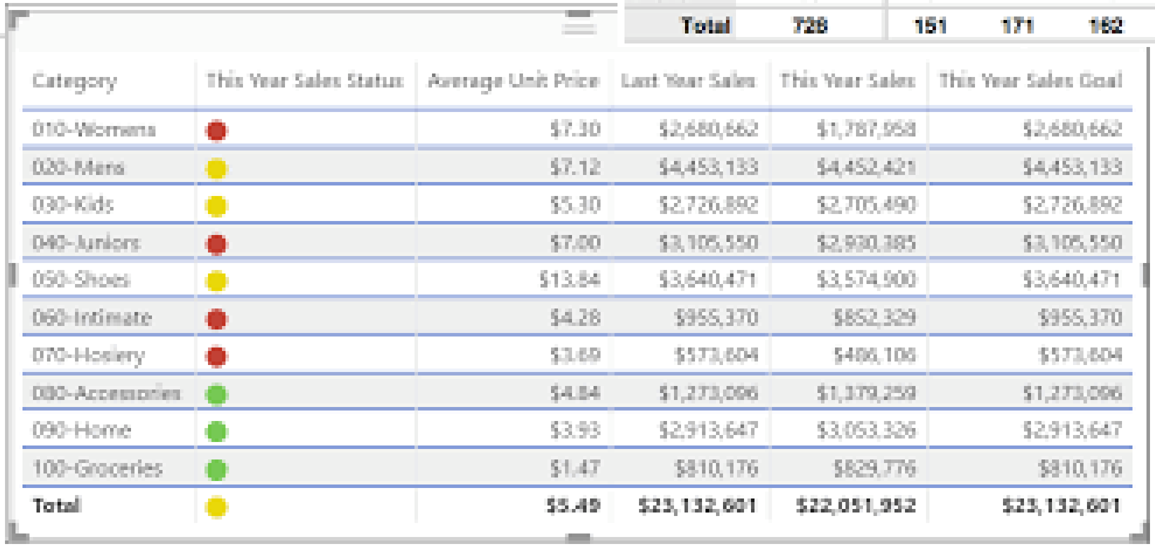
	Total defects	A	B	C	D	E
A4636	131	37	21	28		45
A2524	86	20	24	21	1	20
A3713	75	17	13	18		27
A4452	73	5	33	17		18
A4088	72	14	16	12	2	28
A2103	68	14	13	14	1	26
A2156	68	16	13	19	2	18
A3681	66	12	16	9	1	28
A1366	50	11	15	12		12
A2610	39	5	7	12		15
Total	728					



	Total defects	A	B	C	D	E
A4636	131	37	21	28		45
A2524	86	20	24	21	1	20
A3713	75	17	13	18		27
A4452	73	5	33	17		18
A4088	72	14	16	12	2	28
A2103	68	14	13	14	1	26
A2156	68	16	13	19	2	18
A3681	66	12	16	9	1	28
A1366	50	11	15	12		12
A2610	39	5	7	12		15
Total	728	151	171	162	7	237



	Total defects	A	B	C	D	E
A4636	131	37	21	28		45
A2524	86	20	24	21	1	20
A3713	75	17	13	18		27
A4452	73	5	33	17		18
A4088	72	14	16	12	2	28
A2103	68	14	13	14	1	26
A2156	68	16	13	19	2	18
A3681	66	12	16	9	1	28
A1366	50	11	15	12		12
A2610	39	5	7	12		15
Total	728	151	171	162	7	237



Category	This Year Sales Status	Average Unit Price	Last Year Sales	This Year Sales	This Year Sales Goal
010-Womens	🔴	$7.30	$2,680,662	$1,787,958	$2,680,662
020-Mens	🟡	$7.12	$4,453,133	$4,452,421	$4,453,133
030-Kids	🟡	$5.30	$2,726,892	$2,705,490	$2,726,892
040-Juniors	🔴	$7.00	$3,105,550	$2,930,385	$3,105,550
050-Shoes	🟡	$13.84	$3,640,471	$3,574,900	$3,640,471
060-Intimate	🔴	$4.28	$955,370	$852,329	$955,370
070-Hosiery	🔴	$3.69	$573,604	$486,106	$573,604
080-Accessories	🟢	$4.84	$1,273,096	$1,379,259	$1,273,096
090-Home	🟢	$3.93	$2,913,647	$3,053,326	$2,913,647
100-Groceries	🟢	$1.47	$810,176	$829,776	$810,176
Total	🟡	$5.49	$23,132,601	$22,051,952	$23,132,601

…but with problems

- No overview provided (only planar indicators)

- Structured, but difficult to intercept the changes

- Scale very bad with data cardinality/dimensionality

A step back: Visualization literacy

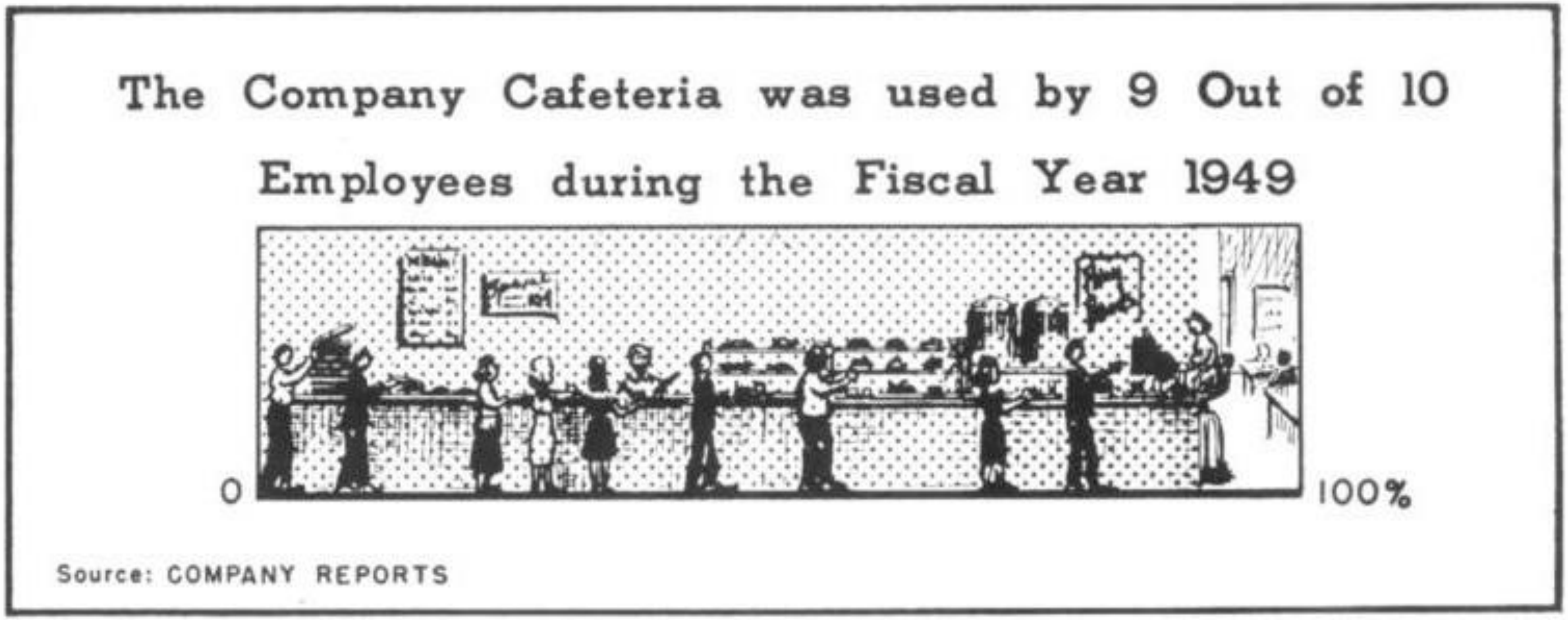
Informal approach

- Rules for different kind of information
- Data quality has mostly prdominant numerical information (e.g. indicators, retios, etc..)
- …with some exceptions

Numerical information: 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

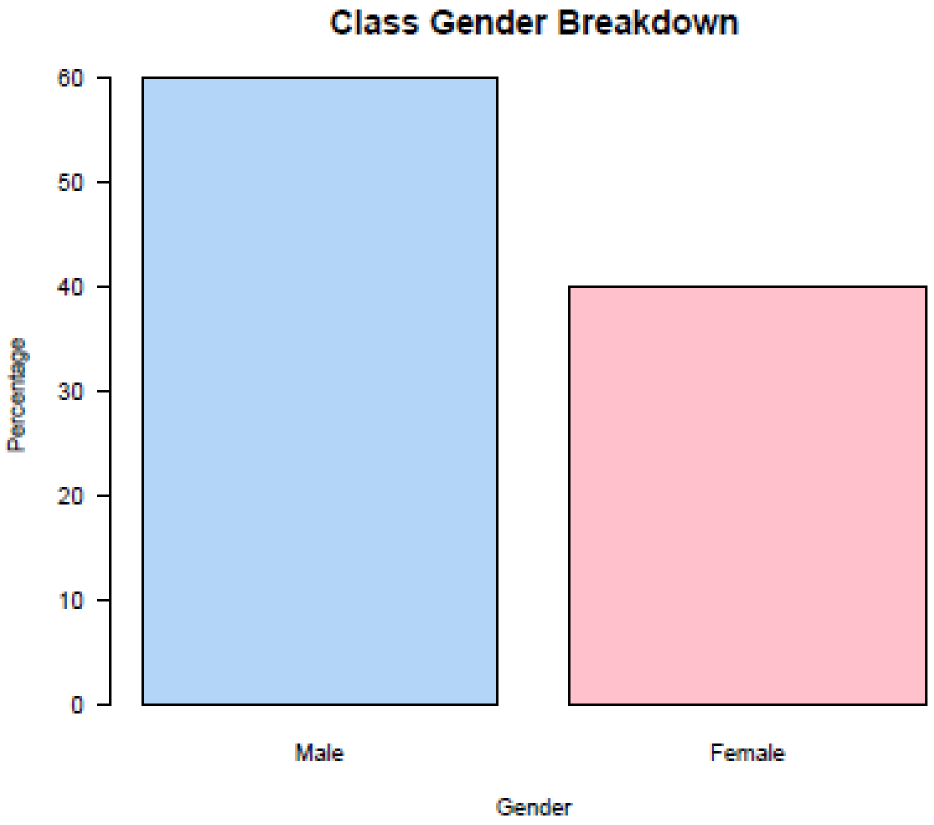
Rule 0 violation (and also rule 2)



The Company Cafeteria was used by 9 Out of 10 Employees during the Fiscal Year 1949

Source: COMPANY REPORTS

Rule 0 violation

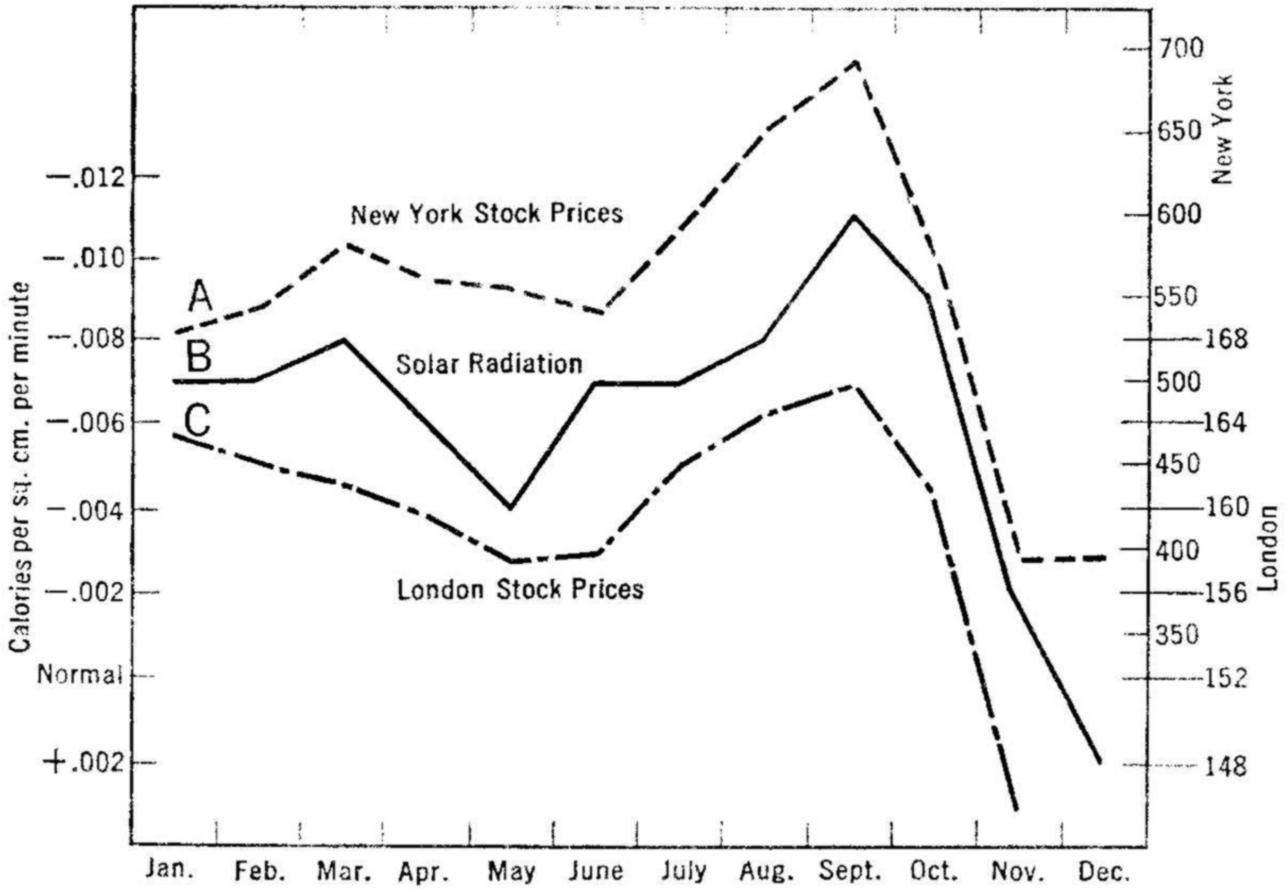


Male	60%
Female	40%

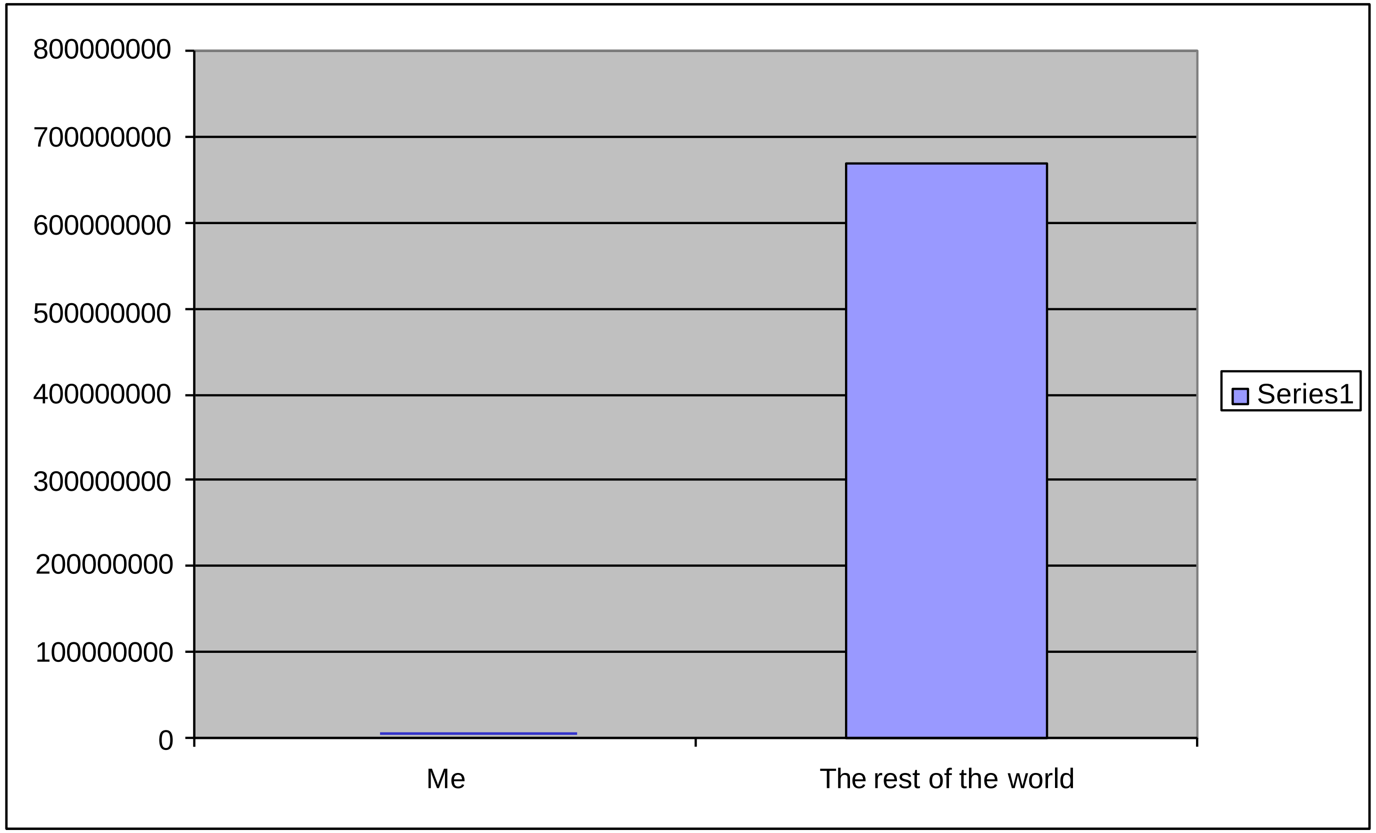
Numerical Information: Rule 1

- **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



Rule 1 violation (and also rule 0)



Not very significant data but a good example of distortion

Numerical Information: Rule 2: Insure chart simplicity

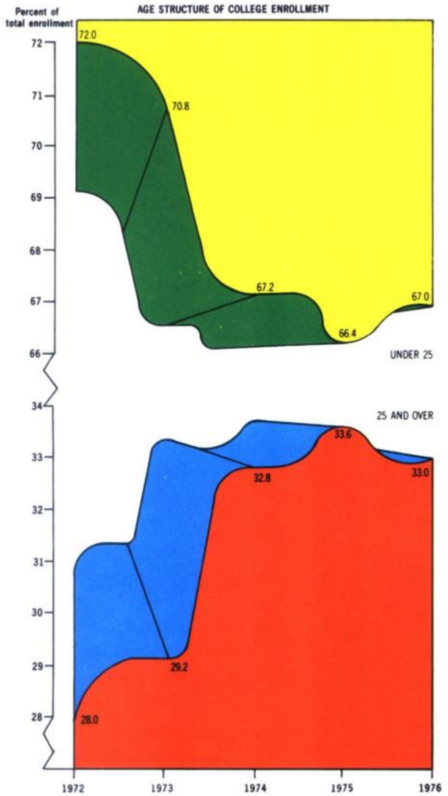
- Graphs should be no more complex than the data which they portray

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

- For a very comprehensive set of chart junk effects look at Microsoft Excel
  - the more recent the version the larger the set !

Rule 2 violation (and also rule 3)

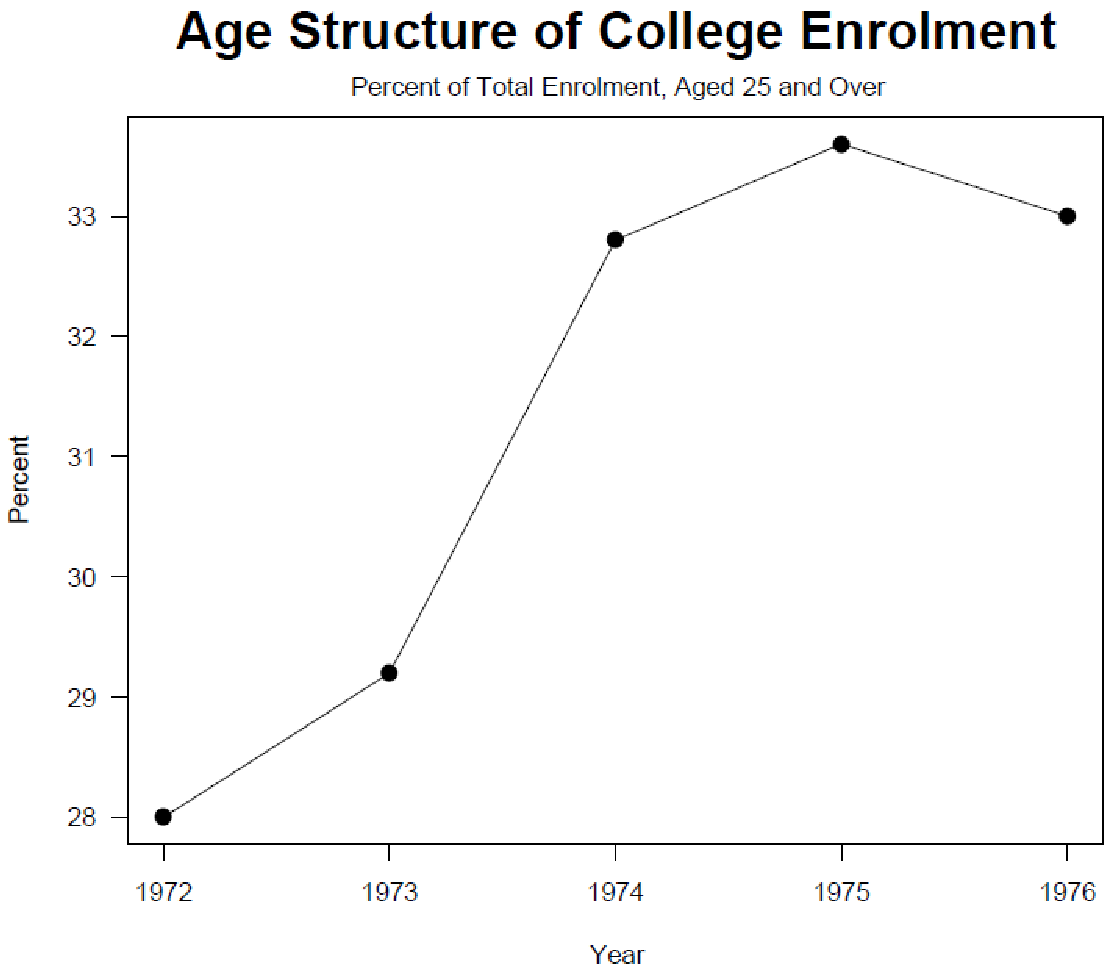
Rule 3 violation



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

- 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)
  - split y axis
  - odd interpolation
- Designers of this graph are now working in the Microsoft Excel's team, inspiring the new Excel's versions ...

Same data…

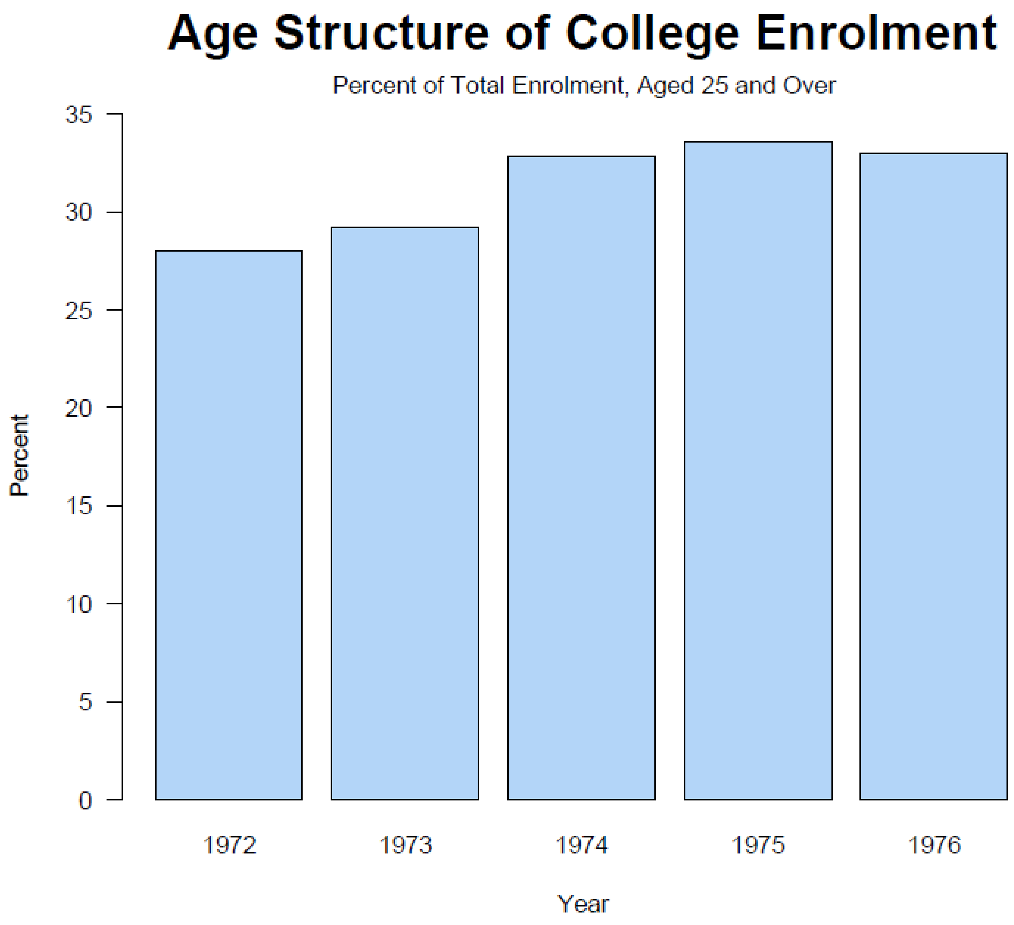


Age Structure of College Enrolment
Percent of Total Enrolment, Aged 25 and Over
Percent
33
32
31
30
29
28
1972
1973
1974
1975
1976
Year

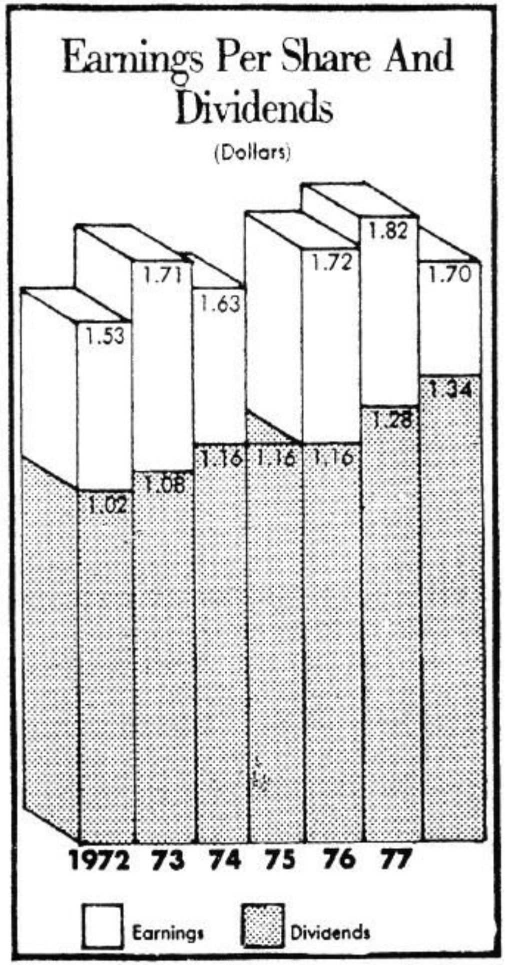
The same data...

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

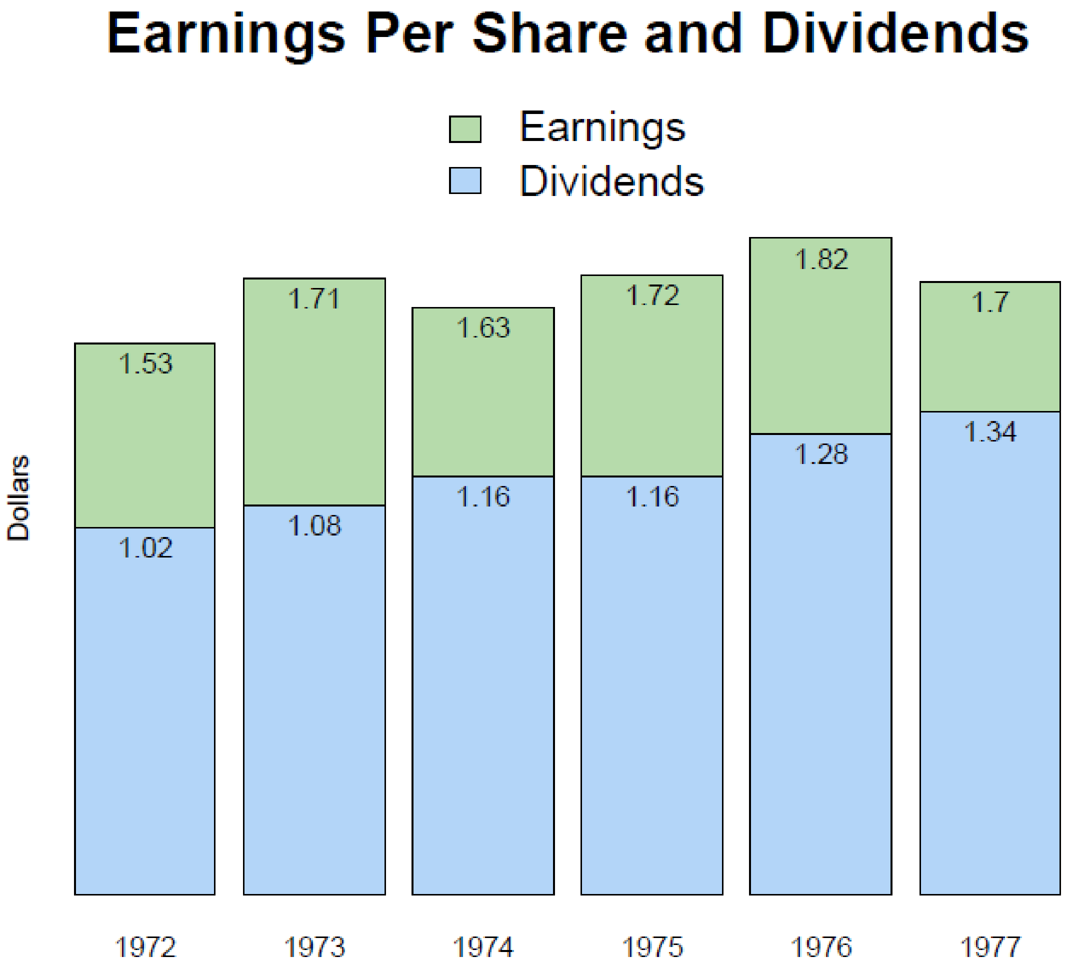
Rule 2 violation



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

The Washington Post, 1979

The same data...

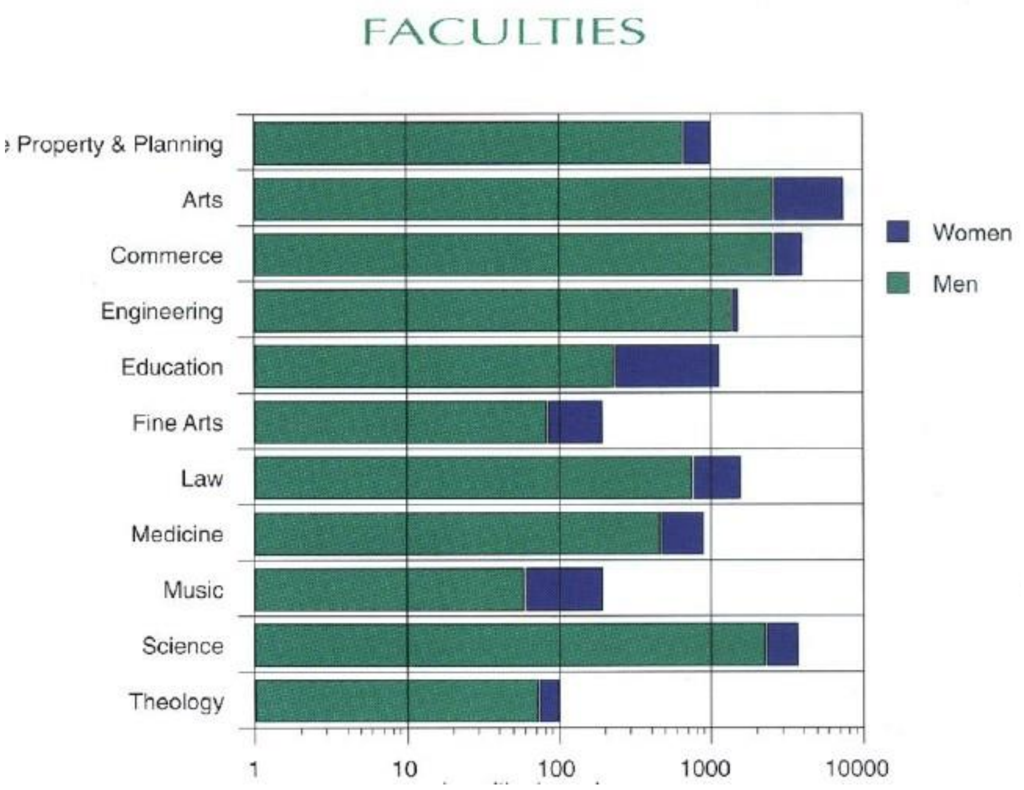


Numerical Information: Rule 3

- **Do not distort data in a confusing way**

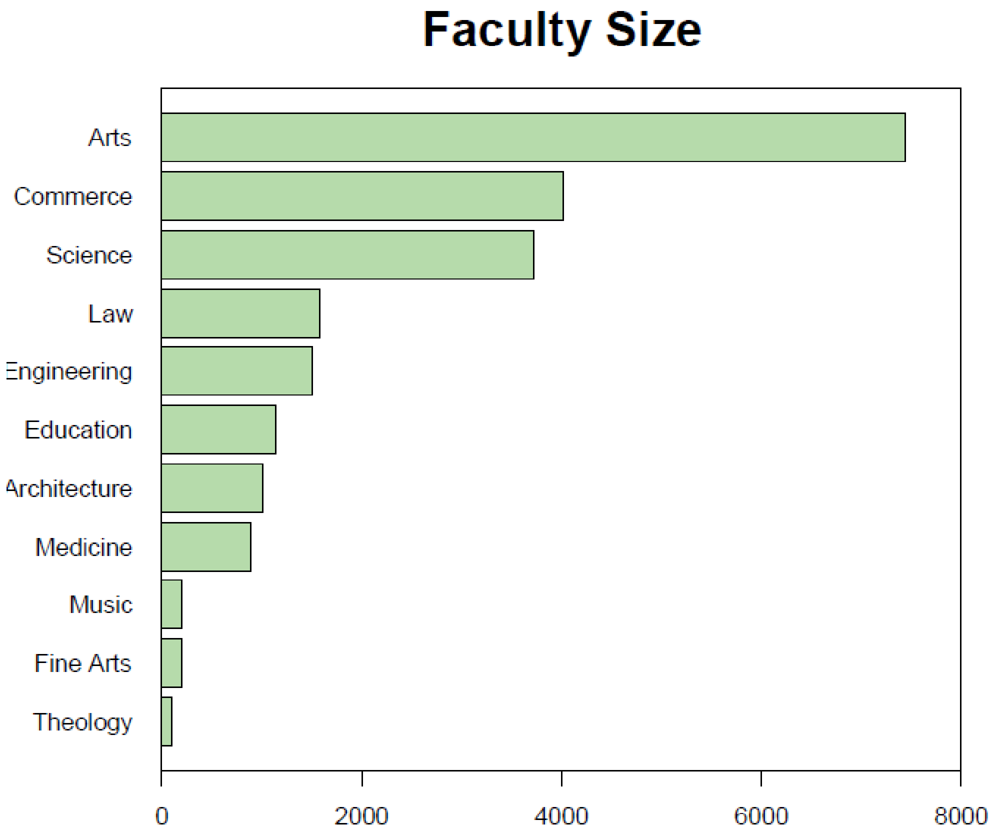
- Graphs should not provide a distorted picture of the values they portray
- Distortion can be either deliberate or accidental
- Of course, it could be useful to know how to produce a graph which bends the truth...

Rule 3 violation

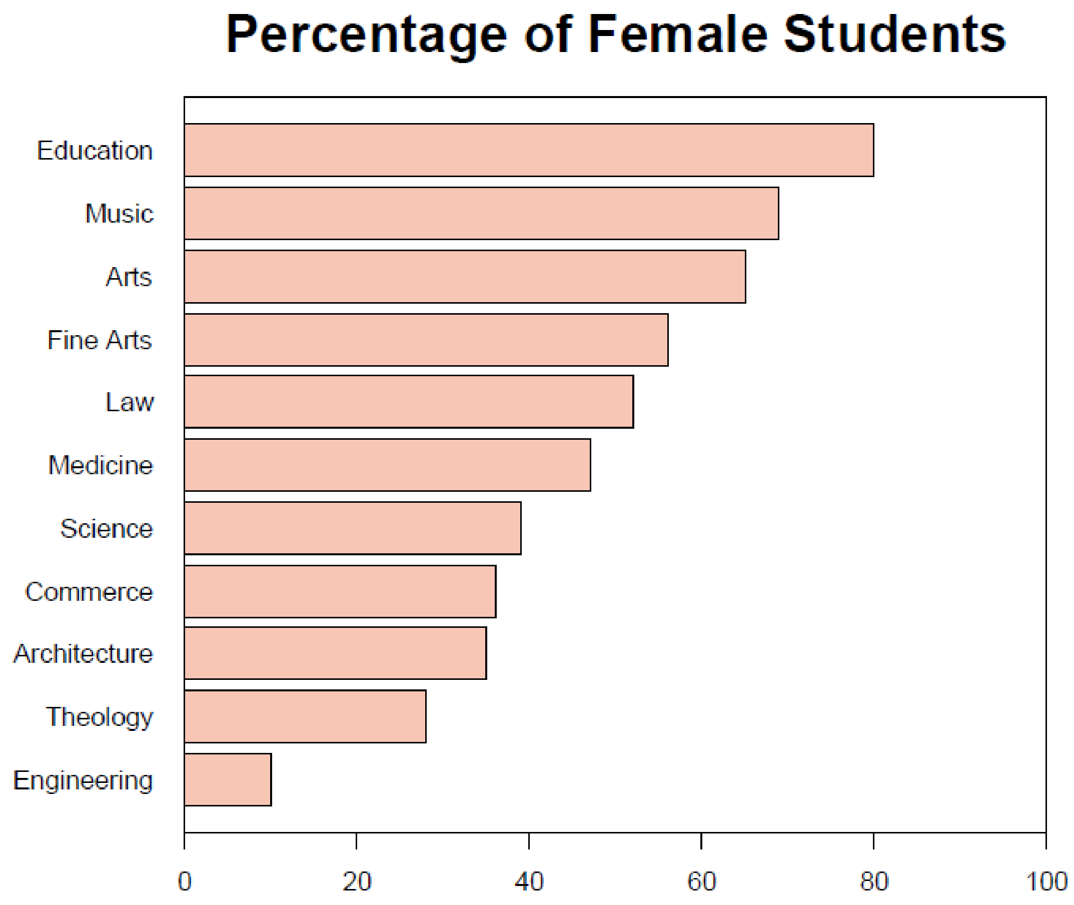


- At a very quick glance:
  - balanced faculty population
  - most male students
- What is wrong with this graph?
- The X scale is logarithmic!

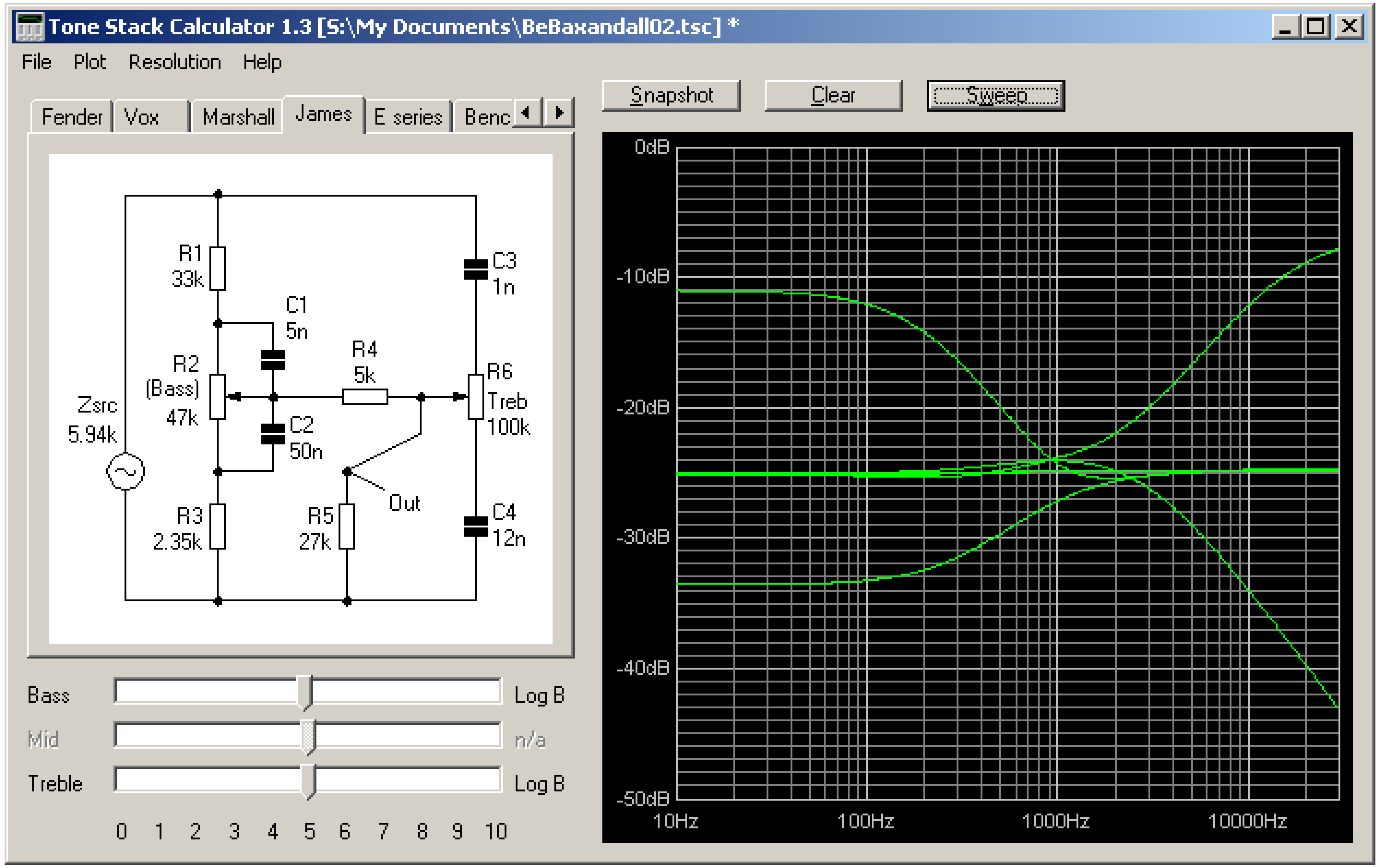
The truth : population size



The truth : female /male ratio



In other cases distortion **is** ok…



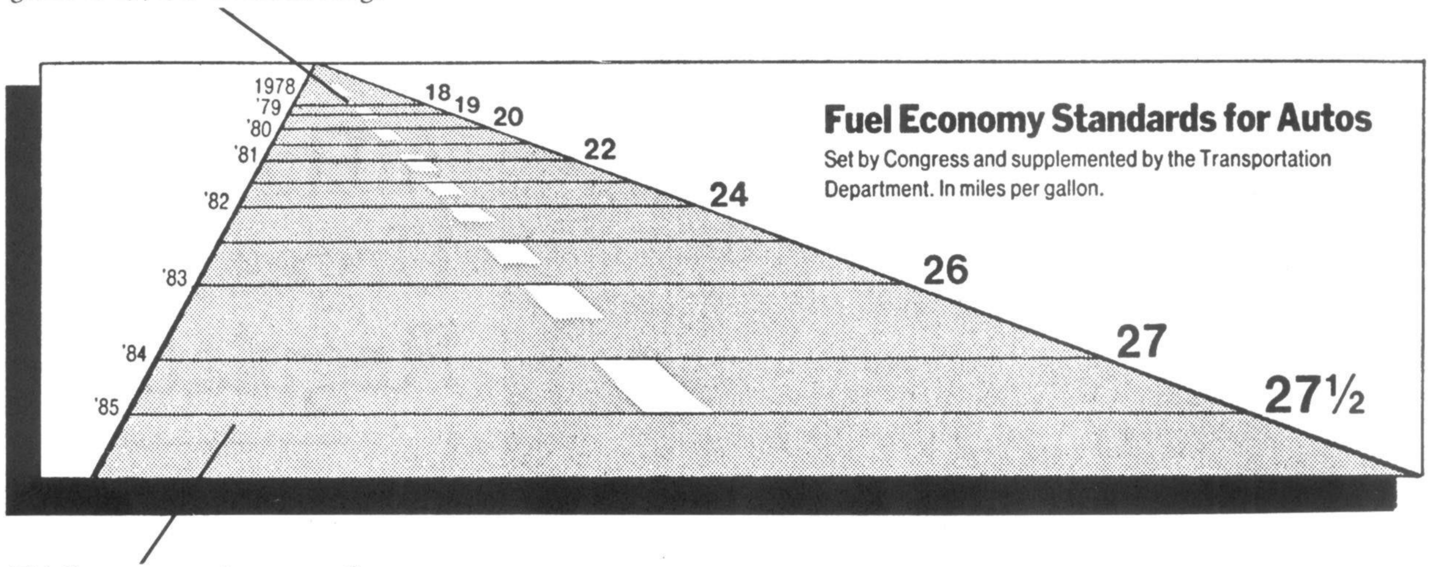
The lie factor

- The visual pioneer Ed Tufte of Yale University has defined a “lie factor” as a measure of the amount of distortion in a graph
- The lie factor is defined to be: Lie Factor =

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

Measuring distortion through the lie factor

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

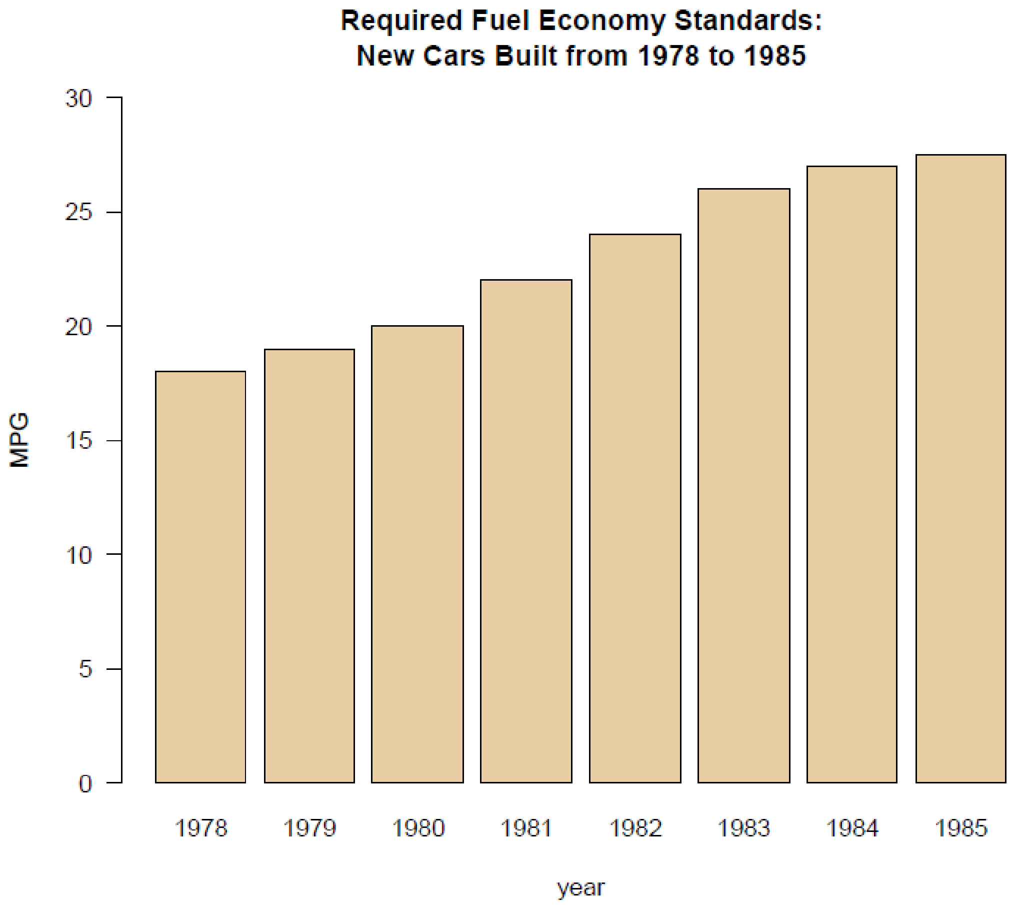


This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

Graph effect = 5.3/0.6=8.8 Data effect = 27.5/18=1.52

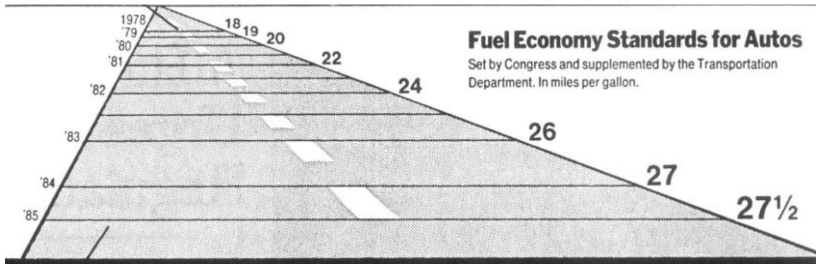
Lie Factor = 8.8/1.52 = **5.8**

The same data with **lie** factor=1



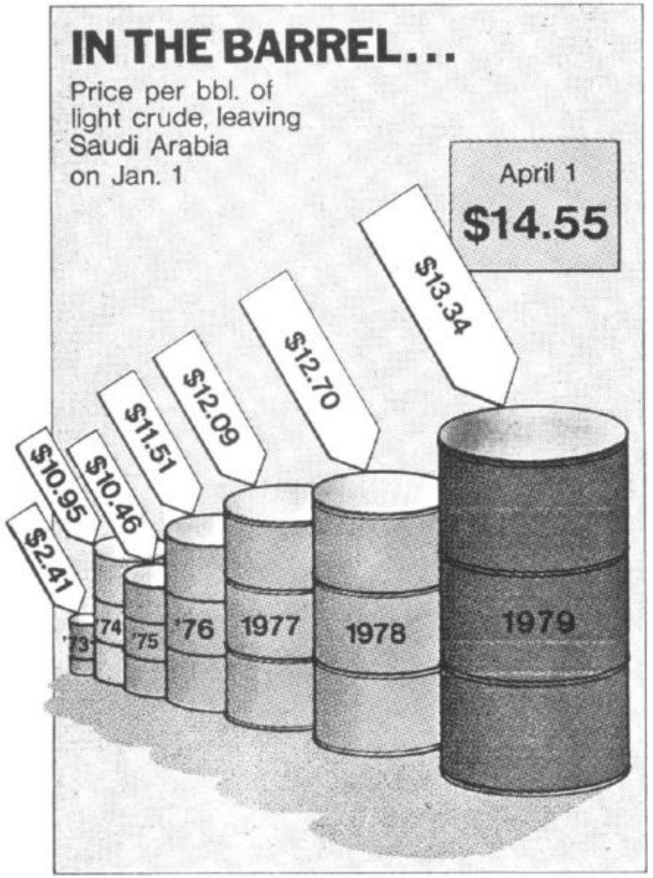
Common Sources of Distortion

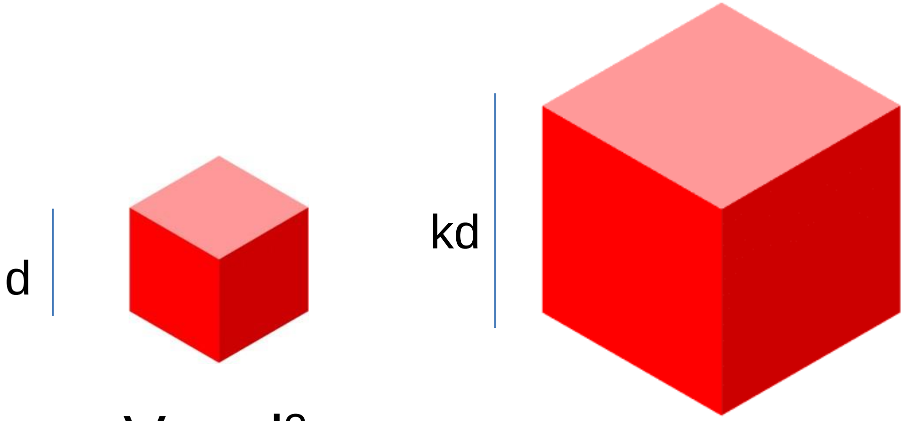
- The use of image perspective is a common source of distortions in graphs



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

Distortion through non linear volumes





$V_1 = d^3$

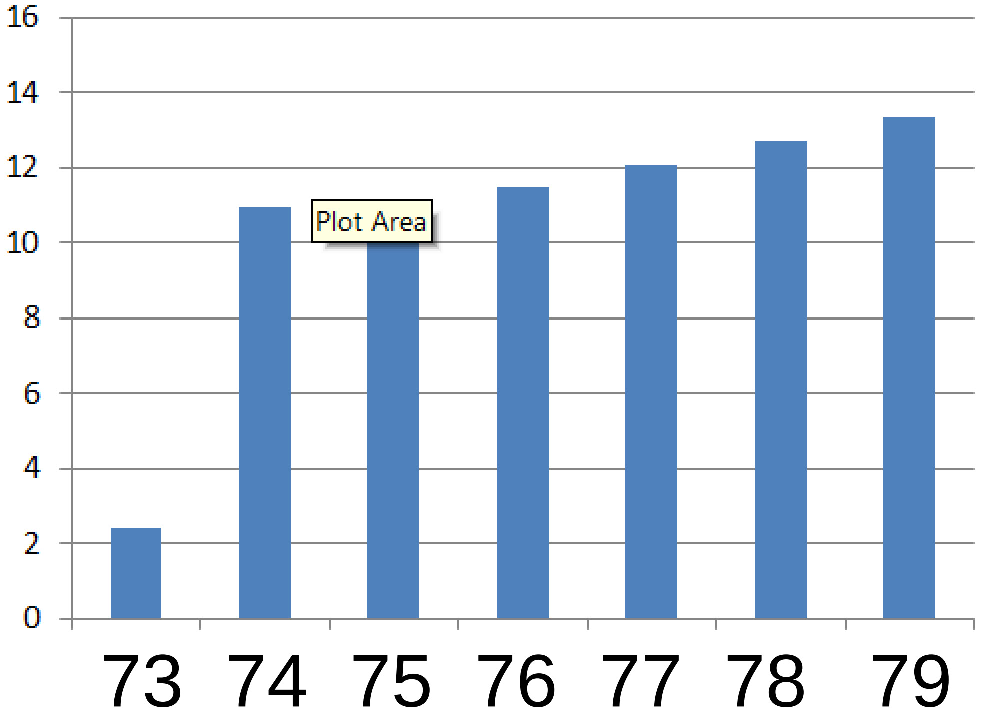
$V_2 = k^3 d^3$

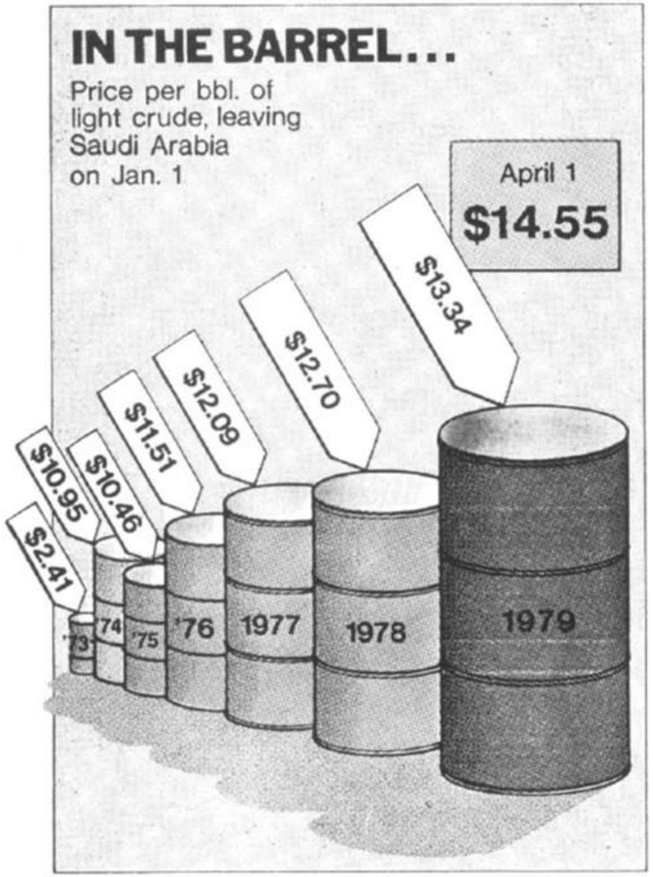
Graph effect = $V_2/V_1 = k^3d^3/d^3 =$ **k^3**
Data effect = kd/d = **k**
Lie Factor = k^3/k = **k^2**

Lie Factor = **Data effect^2**

Lie factor = $(14.55/2.41)^2 = 6^2 = 36$

The same data





Distortion through areas

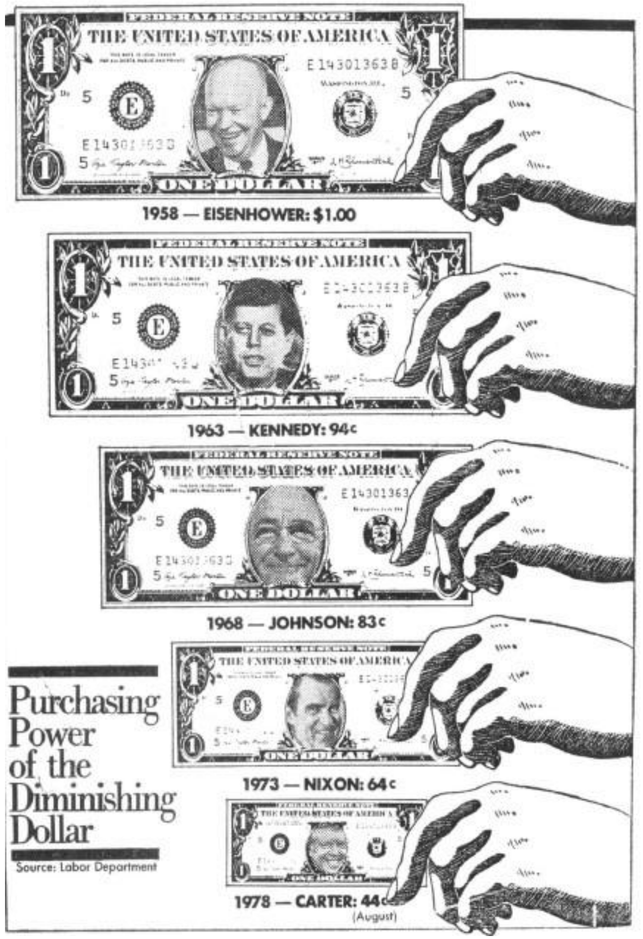
1

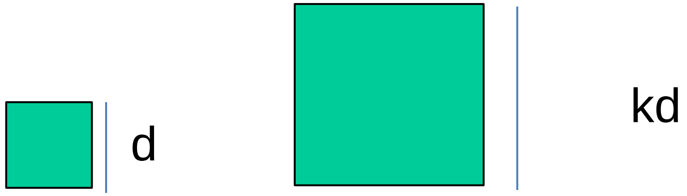
0.94

0.83

0.64

0.46



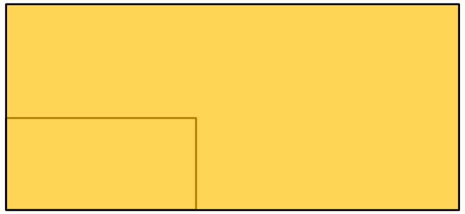


Graph effect = $A_2/A_1 = k^2d^2/d^2 = \mathbf{k^2}$

Data effect = $kd/d = \mathbf{k}$

Lie Factor = $k^2/k = \mathbf{k}$

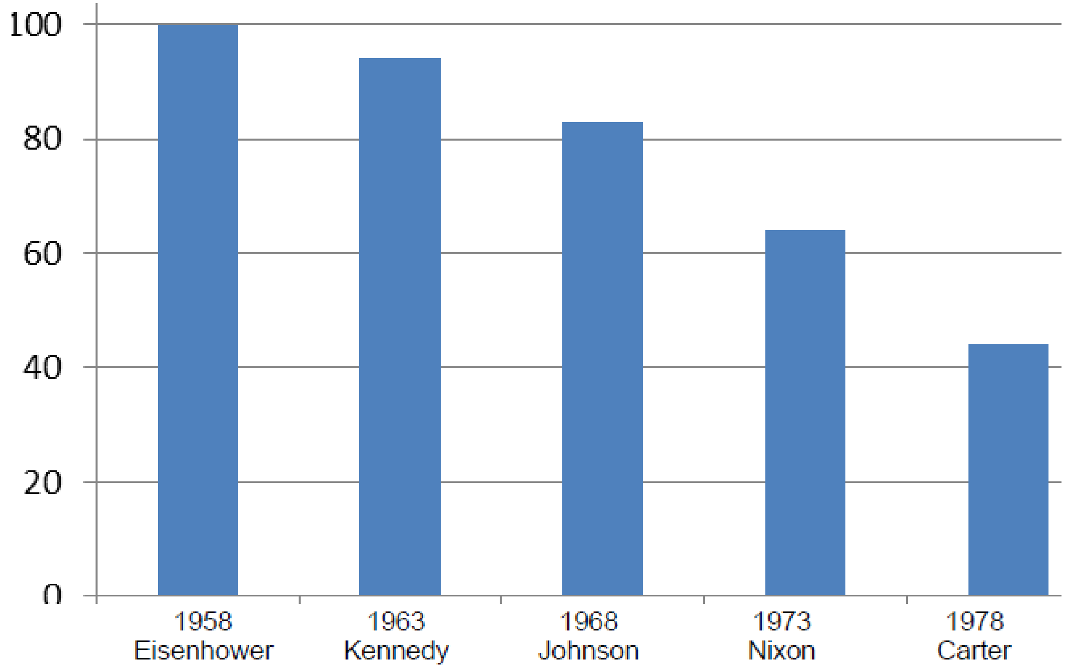
Lie factor = **Data effect**



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

The same data with lie factor=1

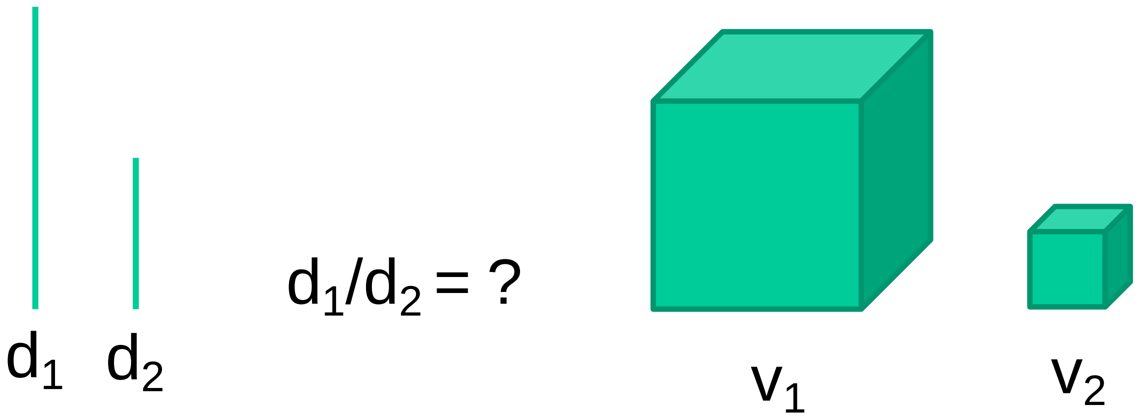
Note that in a histogram you are comparing **lengths**, not **areas**



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

Encoding numerical values

- Human beings are better in comparing lengths than areas or volumes

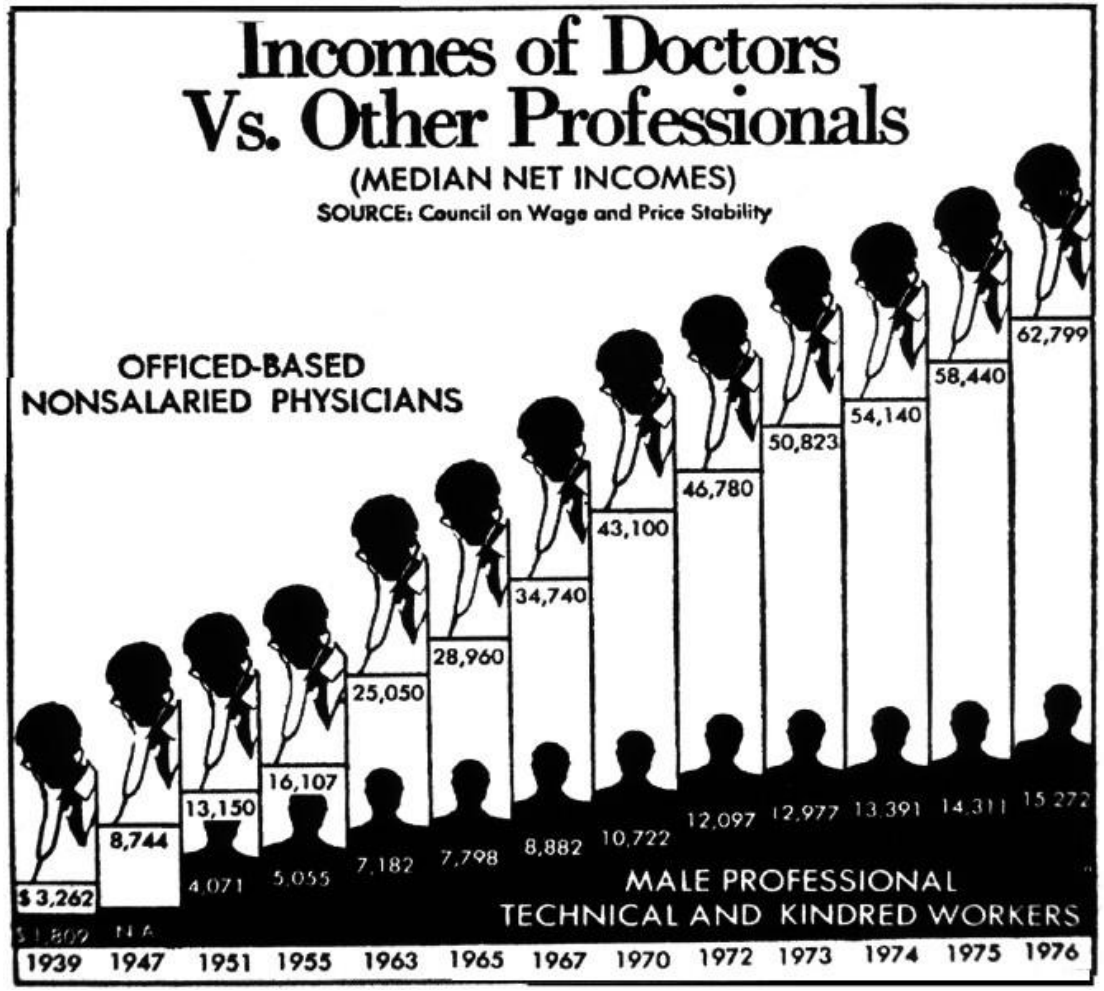


$d_1/d_2 = ?$

$V_1/V_2 = ?$

- So, using volume or area **instead** of length is **wrong**!
- Or it is an intentional lie!

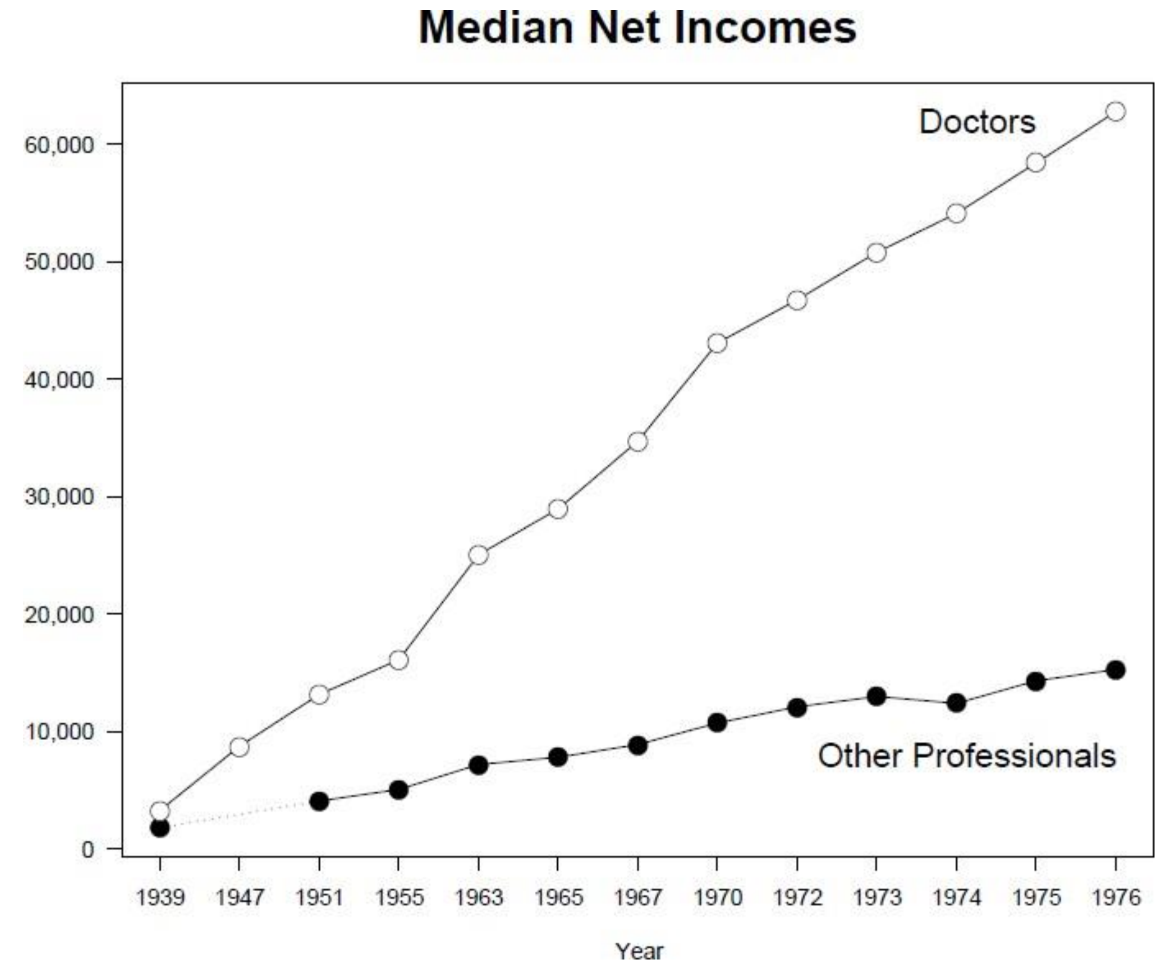
Distortion (deliberate?)



What's wrong with this graph?

A part of the chart junk

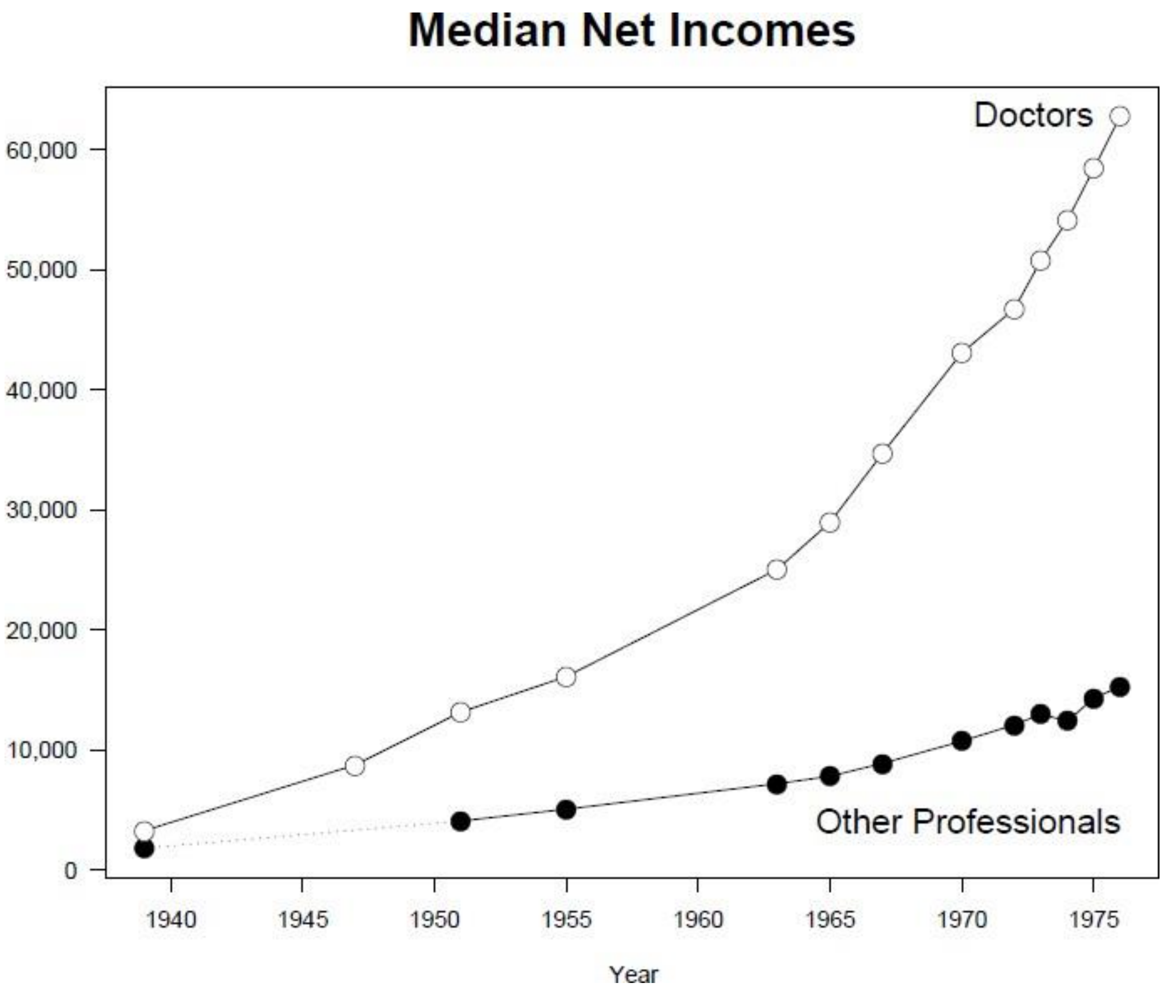
Presented data



It suggests
a linear trend

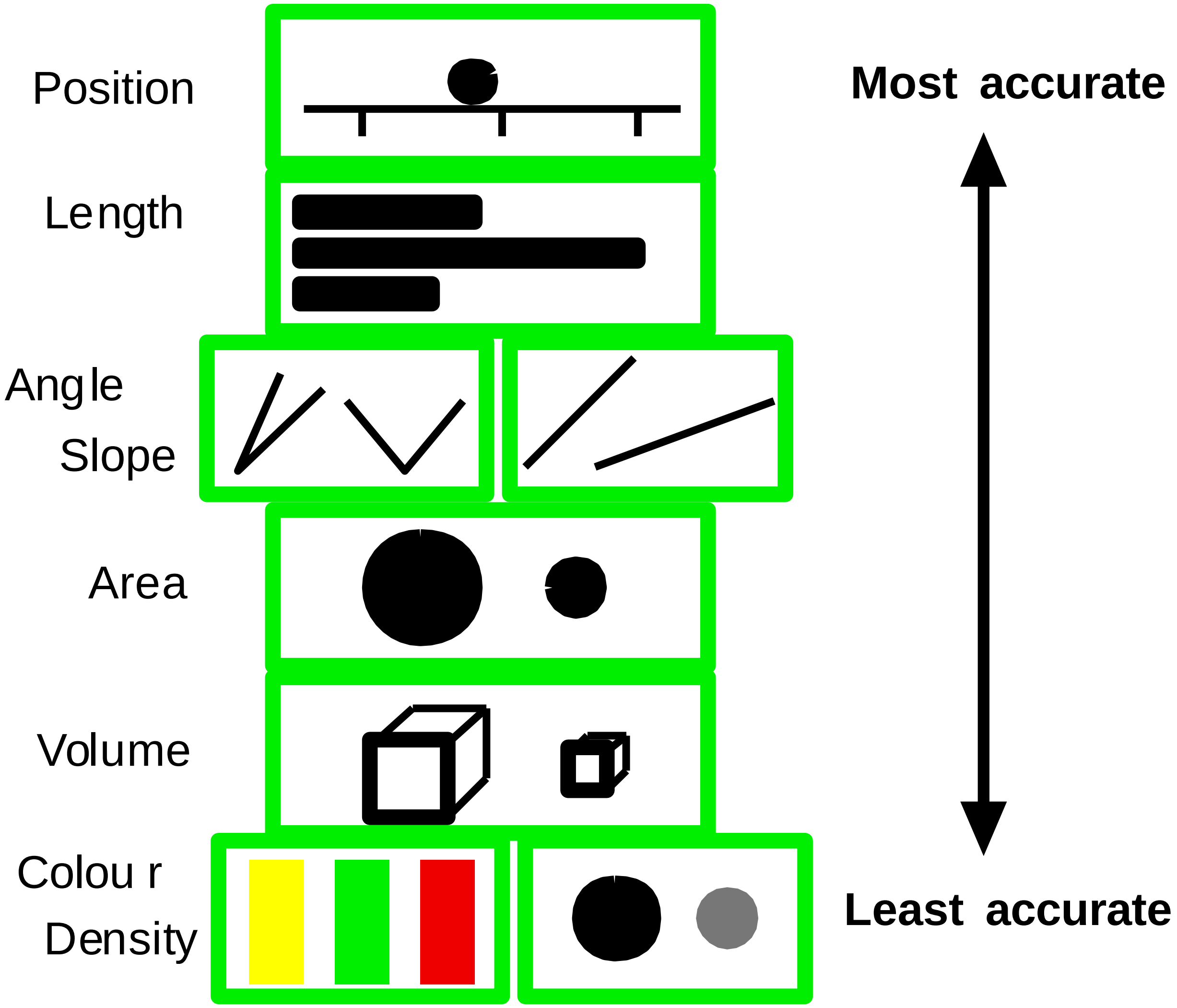
What is wrong
with it?

Real data...



The time scale
was not uniform!

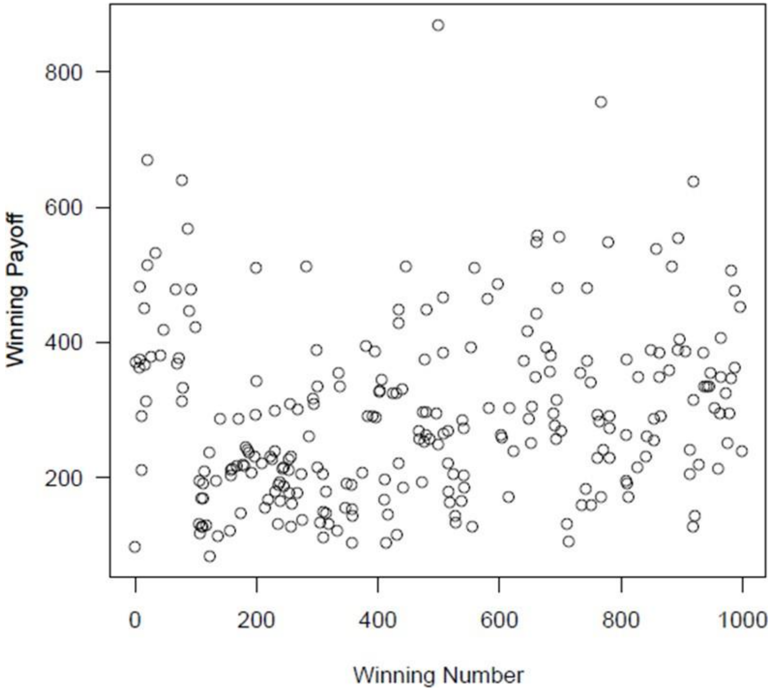
Now the exponential
trend is clear



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

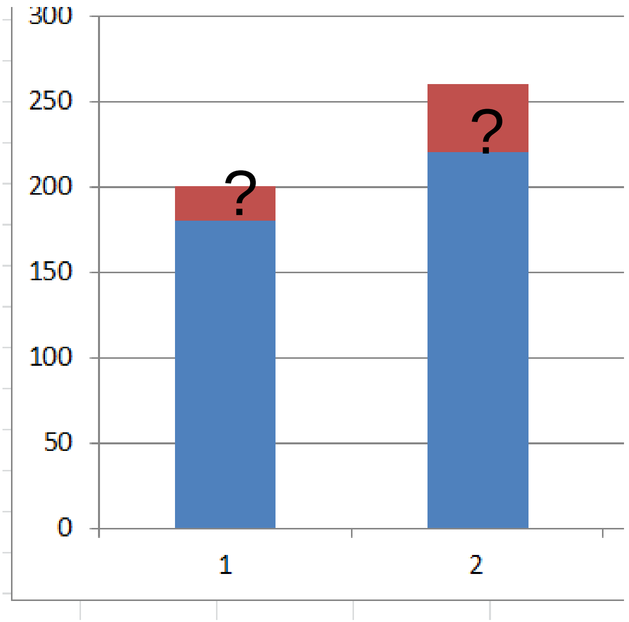
Position

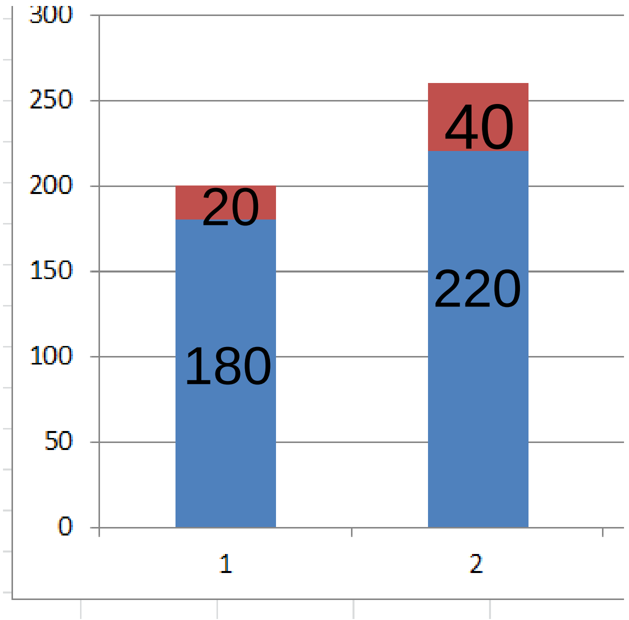
- It works fine



Length?

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

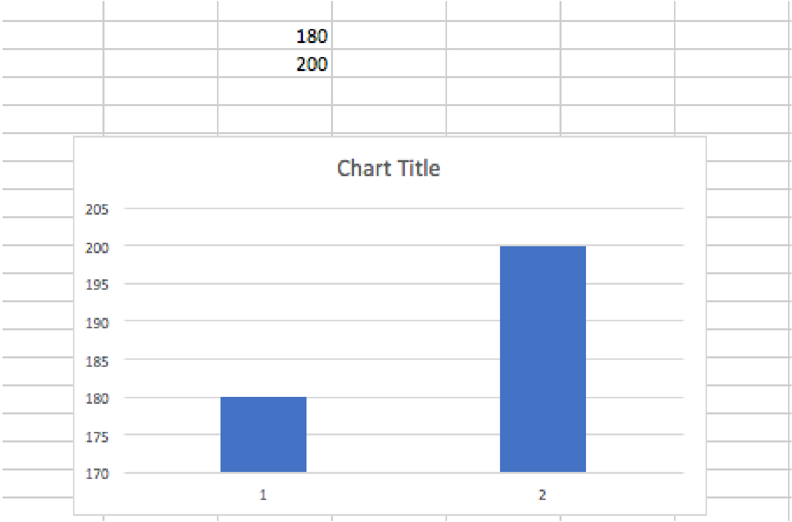




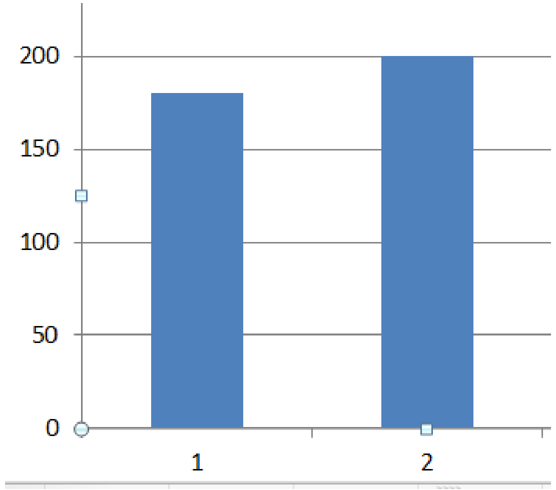
It makes sense to explicitly add figures

Length?

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



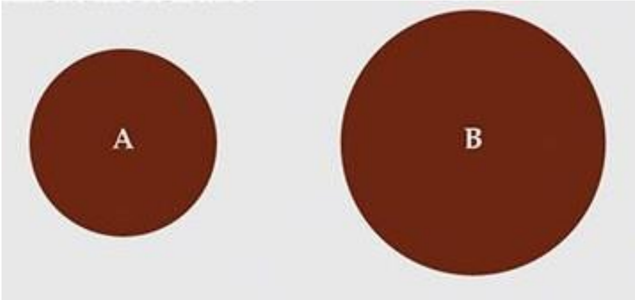
Automatically produced by Excel



The reality

Areas: some new surprising issues

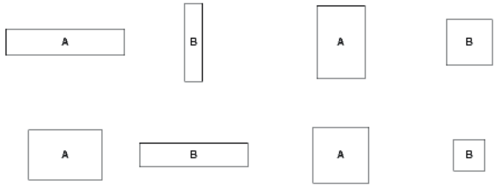
- Human being are very bad in estimating area ratios



- What is the ratio between this two circles A/B ?

  25% 35% 40% 45% 50% 55% 60% 70% ?

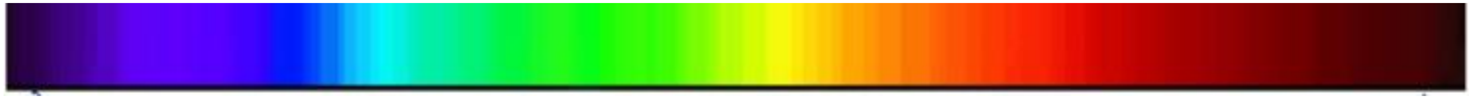
- What is the shape that produces the biggest error?



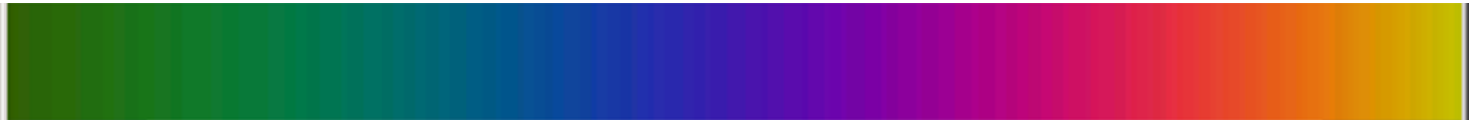
- The square!
- Perceptual Guidelines for Creating Rectangular Treemaps (Nicholas Kong et al., Infovis 2010)

Colors / Numerical data

- 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!!!)
  - It is a very common error (I did it as well, 20 years ago...)



rainbow scale



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

Other choices (Colin Ware)

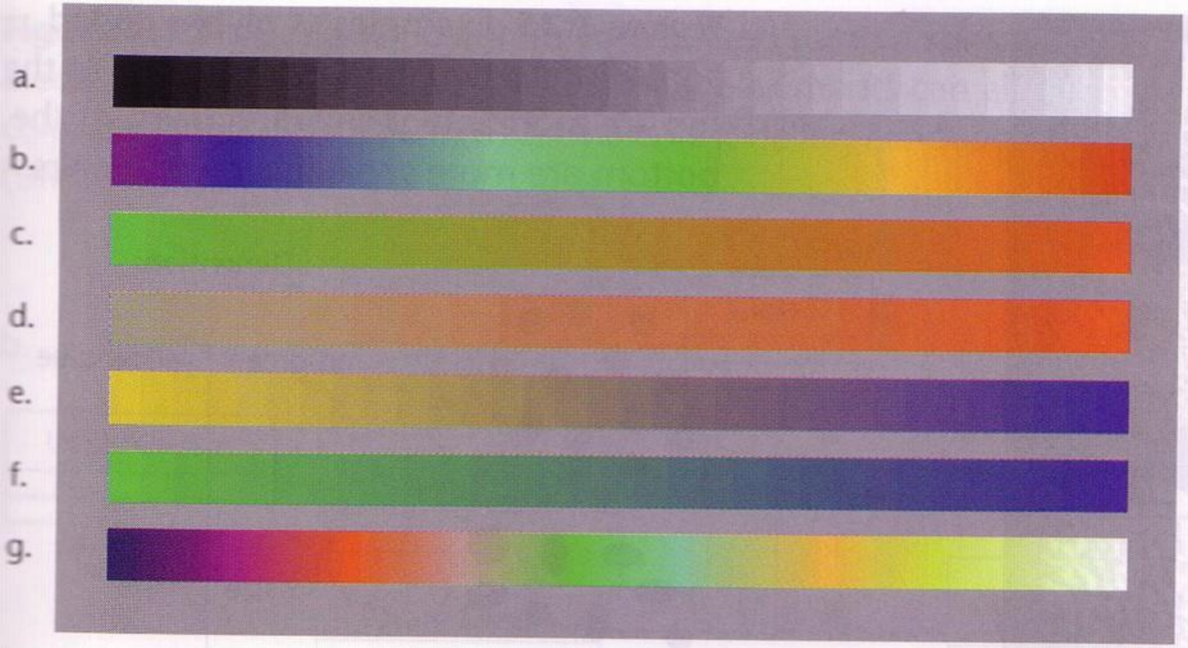
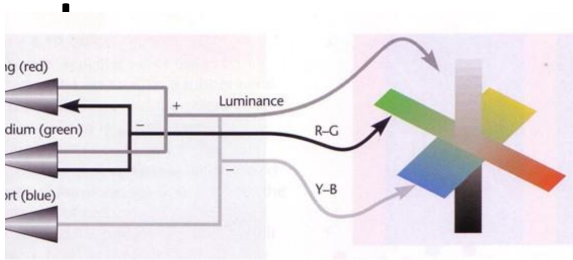
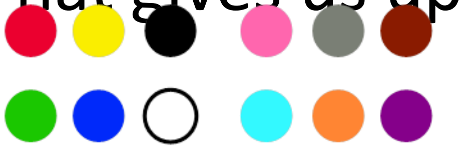


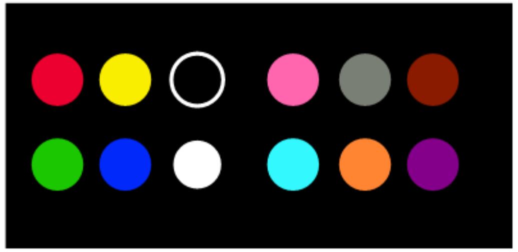
Figure 4.24
Seven different color sequences: (a) Gray scale. (b) Spectrum approximation. (c) Red-green. (d) Saturation. (e) and (f) Two sequences that will be perceived by people suffering from the most common forms of color blindness. (g) A sequence of colors in which each color is lighter than the previous one.

Colors /Categorical

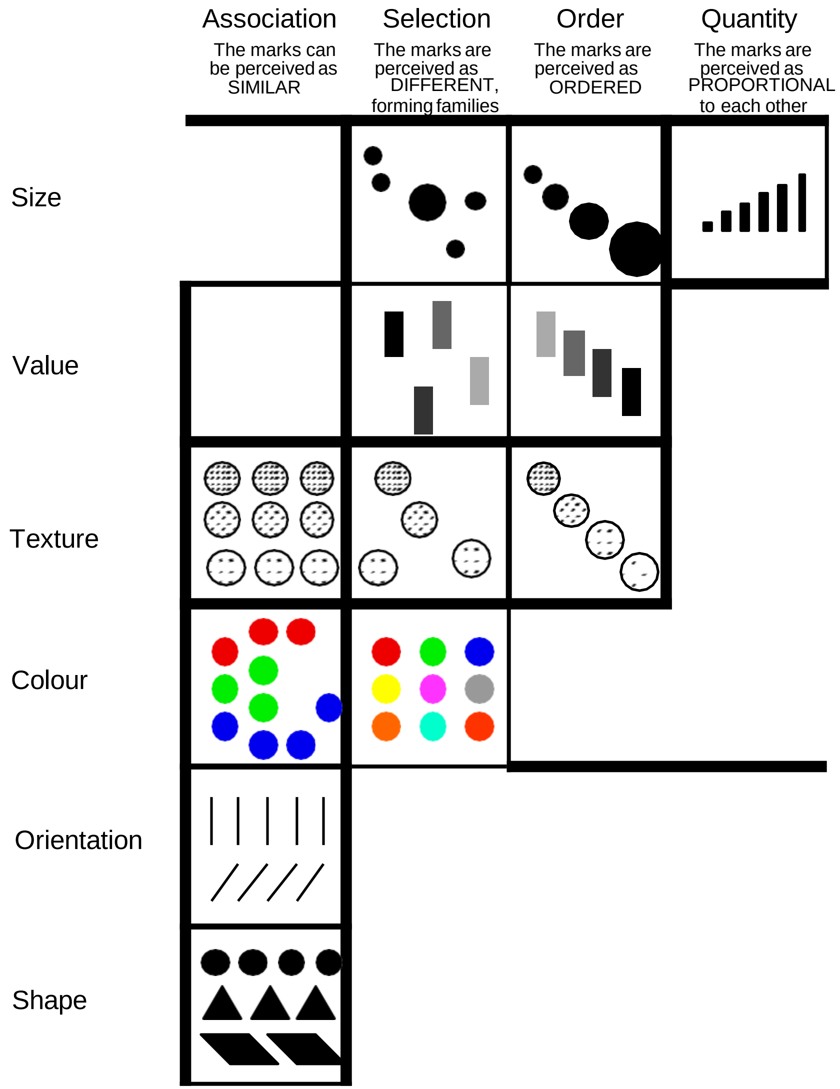


- Colors are fine with categorical data
- Do not reinvent the wheel (again)
- The Ewald Hering idea is that there are only 6 elementary colors arranged in three pairs
  - black-white
  - red-green
  - yellow-blue
- That gives us up to asily distinguishable (11!)





12 Colors
for labeling



Association
The marks can be perceived as SIMILAR

Selection
The marks are perceived as DIFFERENT, forming families

Order
The marks are perceived as ORDERED

Quantity
The marks are perceived as PROPORTIONAL to each other

Size

Value

Texture

Colour

Orientation

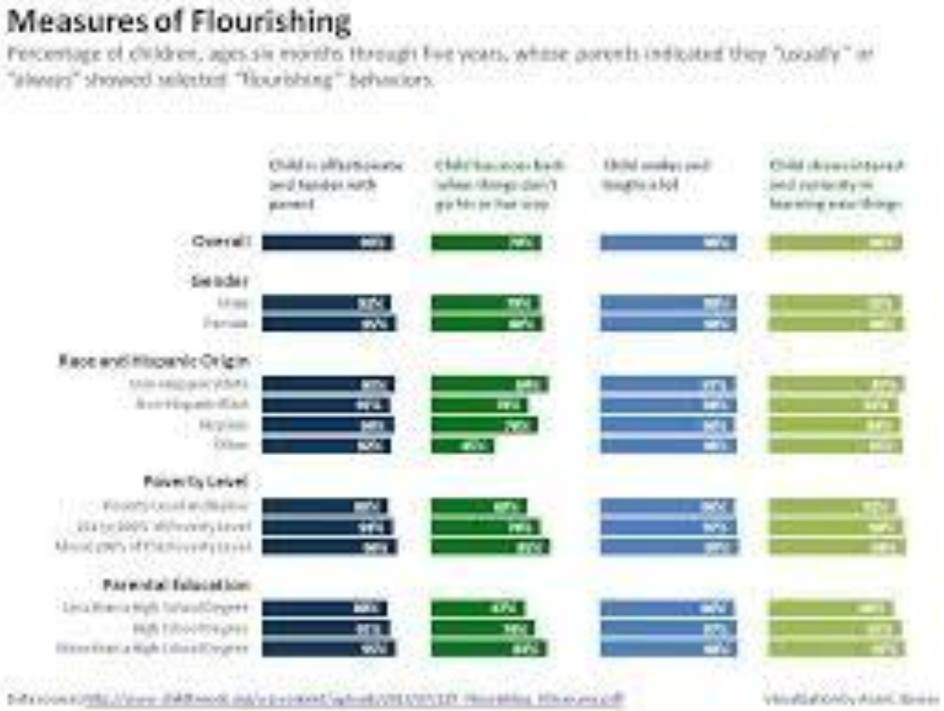
Shape

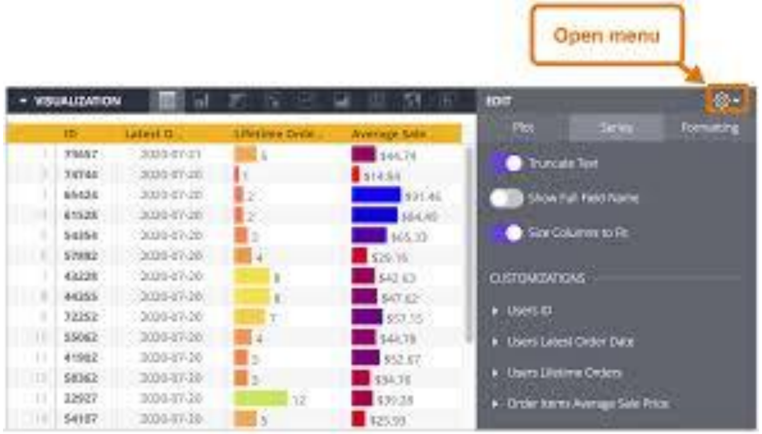
Interpretation of Bertin’s guidance regarding the suitability of various encoding methods to support common tasks

Some new considerations

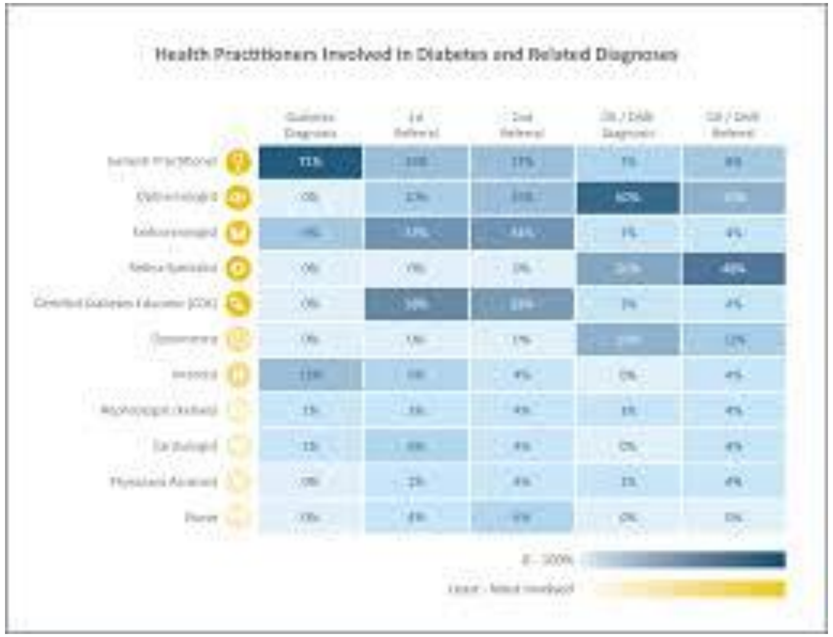
- Chartjunk is not the unique enemy...
- Before PCs building graphs was a matter of paper and pencil
  - requiring time and effort
  - pushing you to better understand :
    - the meaning of numbers
    - the graph purpose
    - the graph organization
    - …
- now, with Excel (or MatplotLib, or general chart software 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...)

Visualization literacy: better tables





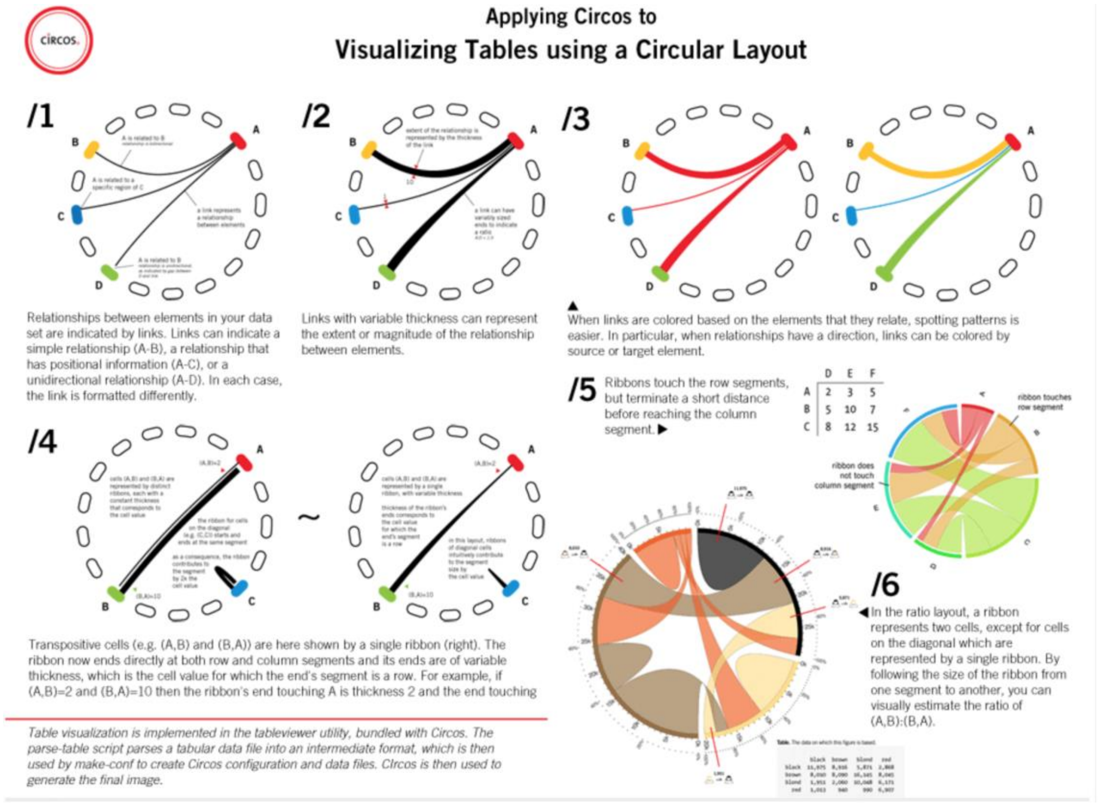
	ID	Latest O...	Lifetime Orde...	Average Sale...
1	79657	2020-07-21	5	$44.74
2	74744	2020-07-20	1	$14.84
3	65426	2020-07-20	2	$91.46
4	61528	2020-07-20	2	$84.40
5	54354	2020-07-20	3	$65.33
6	57982	2020-07-20	4	$29.16
7	43228	2020-07-20	8	$42.63
8	44255	2020-07-20	8	$47.62
9	72252	2020-07-20	7	$57.15
10	55062	2020-07-20	4	$44.78
11	41982	2020-07-20	5	$52.67
12	58362	2020-07-20	3	$34.78
13	22927	2020-07-20	12	$39.28
14	54187	2020-07-20	5	$25.99



Health Practitioners Involved in Diabetes and Related Diagnoses

	Diabetes Diagnosis	1st Referral	2nd Referral	DR / DME Diagnosis	DR / DME Referral
General Practitioner	71%	[illegible]	17%	7%	8%
Ophthalmologist	0%	17%	[illegible]	60%	[illegible]
Endocrinologist	[illegible]	37%	41%	7%	4%
Retina Specialist	0%	0%	0%	[illegible]	49%
Certified Diabetes Educator (CDE)	0%	39%	[illegible]	7%	4%
Optometrist	0%	0%	1%	[illegible]	12%
Internist	11%	0%	4%	0%	4%
Nephrologist (Kidney)	1%	1%	4%	1%	4%
Cardiologist	1%	[illegible]	4%	0%	4%
Physician Assistant	0%	2%	4%	1%	4%
Nurse	0%	4%	[illegible]	0%	0%

Visualization literacy: better **THAN** tables



CIRCOS.
/1
A
B
C
D
A is related to B
A is related to a specific region of C
a link represents a relationship between elements
A is related to B
/2
extent of the relationship is represented by the thickness of the link
a link can have variably sized ends to indicate a ratio
/3
/4
(A,B)=2
(B,A)=10
cells (A,B) and (B,A) are represented by distinct ribbons, each with a constant thickness that corresponds to the cell value
the ribbon for cells on the diagonal (e.g. (C,C)) starts and ends at the same segment
as a consequence, the ribbon contributes to the segment by 2x the cell value
~
cells (A,B) and (B,A) are represented by a single ribbon, with variable thickness
thickness of the ribbon's ends corresponds to the cell value for which the end's segment is a row
in this layout, ribbons of diagonal cells intuitively contribute to the segment size by the cell value
/5
ribbon touches row segment
ribbon does not touch column segment
/6

Applying Circos to

Visualizing Tables using a Circular Layout

/1 Relationships between elements in your data set are indicated by links. Links can indicate a simple relationship (A-B), a relationship that has positional information (A-C), or a unidirectional relationship (A-D). In each case, the link is formatted differently.

/2 Links with variable thickness can represent the extent or magnitude of the relationship between elements.

/3 ▲ When links are colored based on the elements that they relate, spotting patterns is easier. In particular, when relationships have a direction, links can be colored by source or target element.

/4 Transpositive cells (e.g. (A,B) and (B,A)) are here shown by a single ribbon (right). The ribbon now ends directly at both row and column segments and its ends are of variable thickness, which is the cell value for which the end's segment is a row. For example, if (A,B)=2 and (B,A)=10 then the ribbon's end touching A is thickness 2 and the end touching

/5 Ribbons touch the row segments, but terminate a short distance before reaching the column segment. ▶

	D	E	F
A	2	3	5
B	5	10	7
C	8	12	15

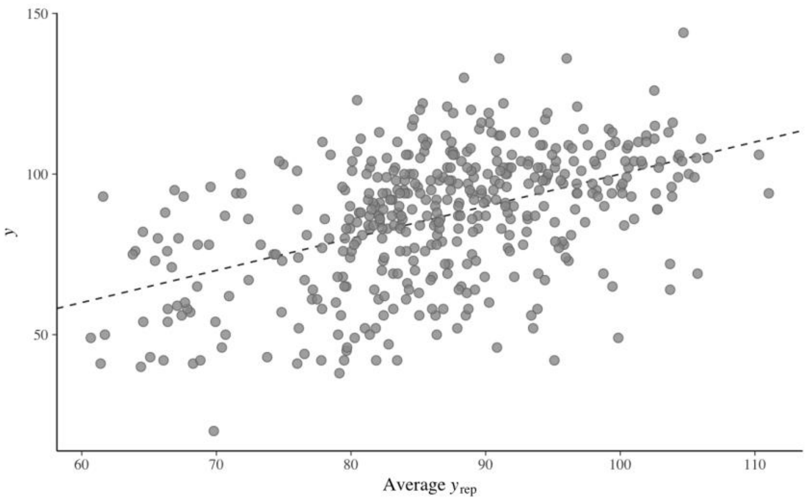
/6 ◀ In the ratio layout, a ribbon represents two cells, except for cells on the diagonal which are represented by a single ribbon. By following the size of the ribbon from one segment to another, you can visually estimate the ratio of (A,B):(B,A).

Table. The data on which this figure is based.

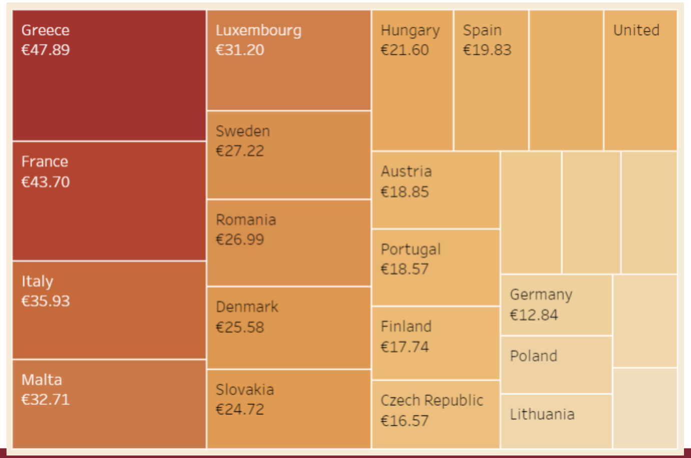
	black	brown	blond	red
black	11,975	8,916	5,871	2,868
brown	8,010	8,090	16,145	8,045
blond	1,951	2,060	10,048	6,171
red	1,013	940	990	6,907

Table visualization is implemented in the tableviewer utility, bundled with Circos. The parse-table script parses a tabular data file into an intermediate format, which is then used by make-conf to create Circos configuration and data files. CIrcos is then used to generate the final image.

Chords: relations

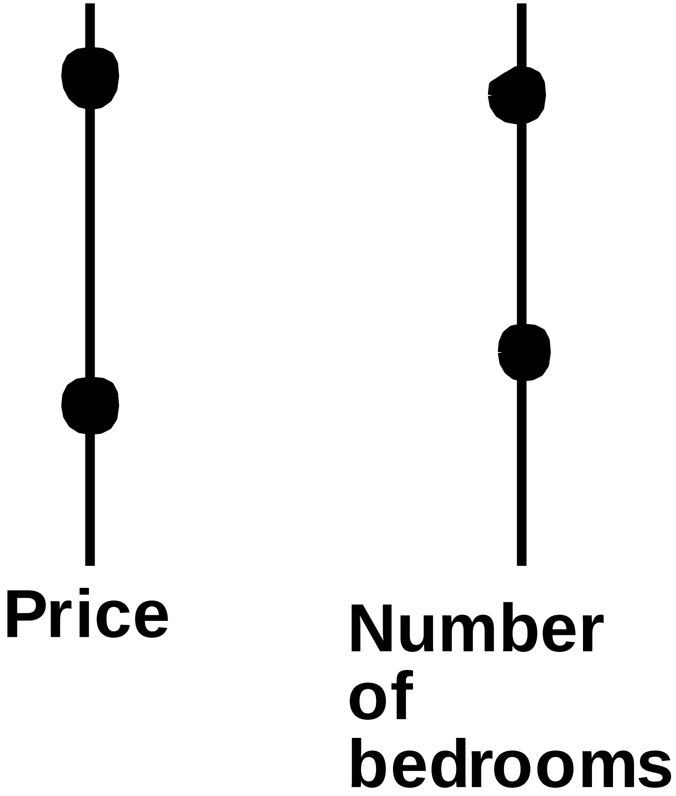


Scatterplots: correlations

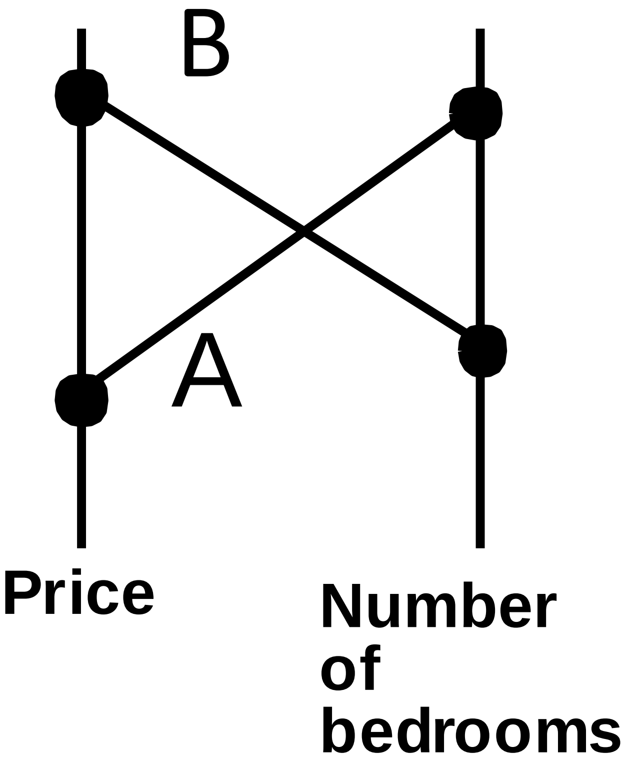


Treemaps: hierarchy / part to whole

Parallel coordinates

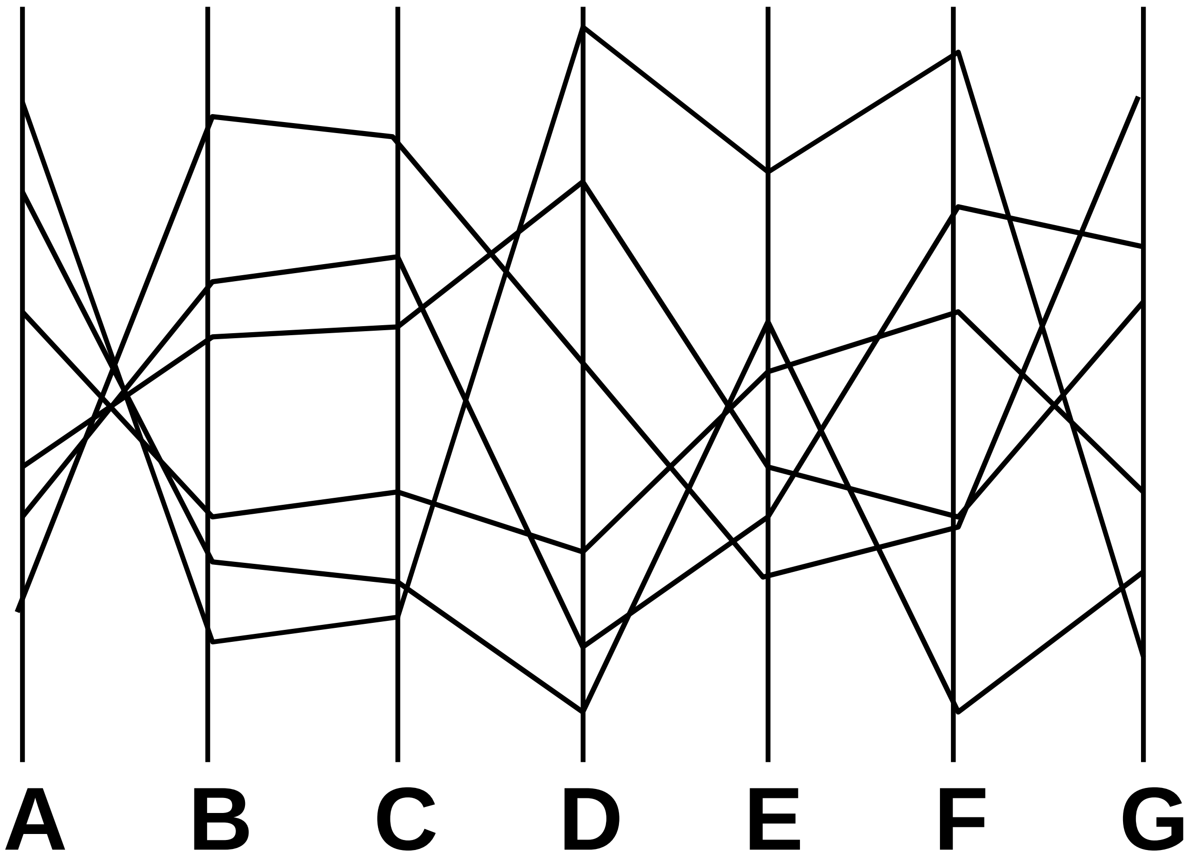


An alternative representation to the scatterplot in which the two attribute scales are presented in parallel, thereby requiring two points to represent each house



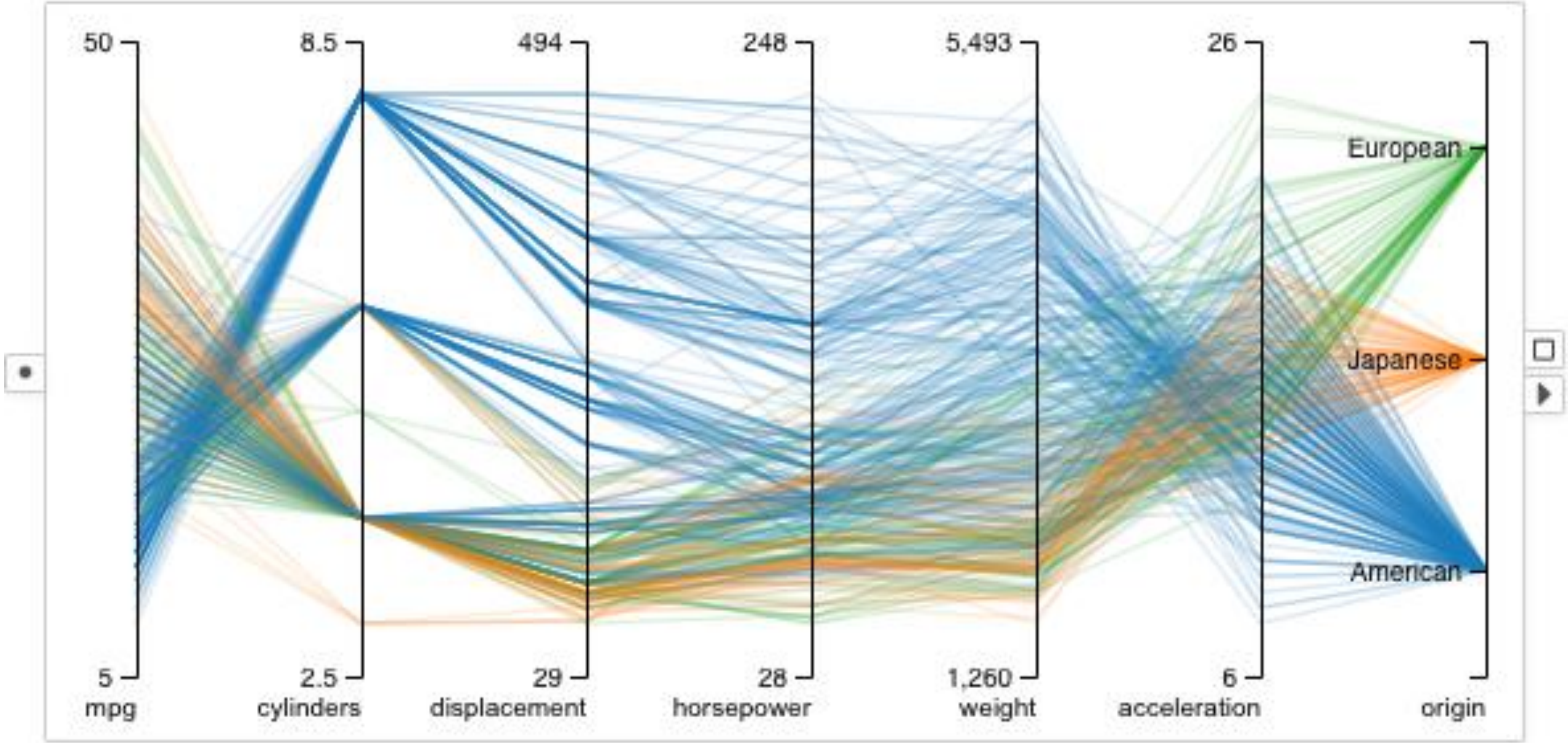
To avoid ambiguity the pair of points representing a house are joined and labelled

Parallel coordinates

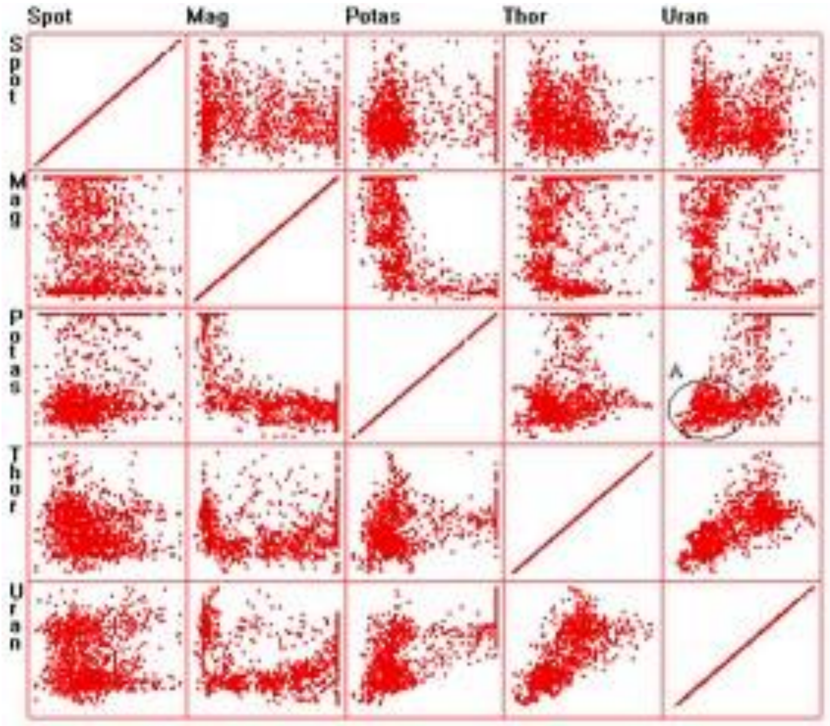


A parallel coordinate plot for six objects, each characterised by seven attributes. The trade-off between A and B, and the correlation between B and C, are immediately apparent. The trade-off between B and E, and the correlation between C and G, are not

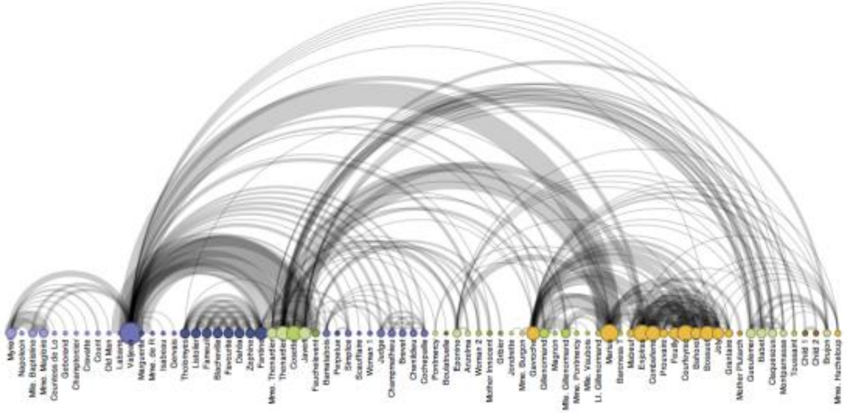
Parallel coordinates

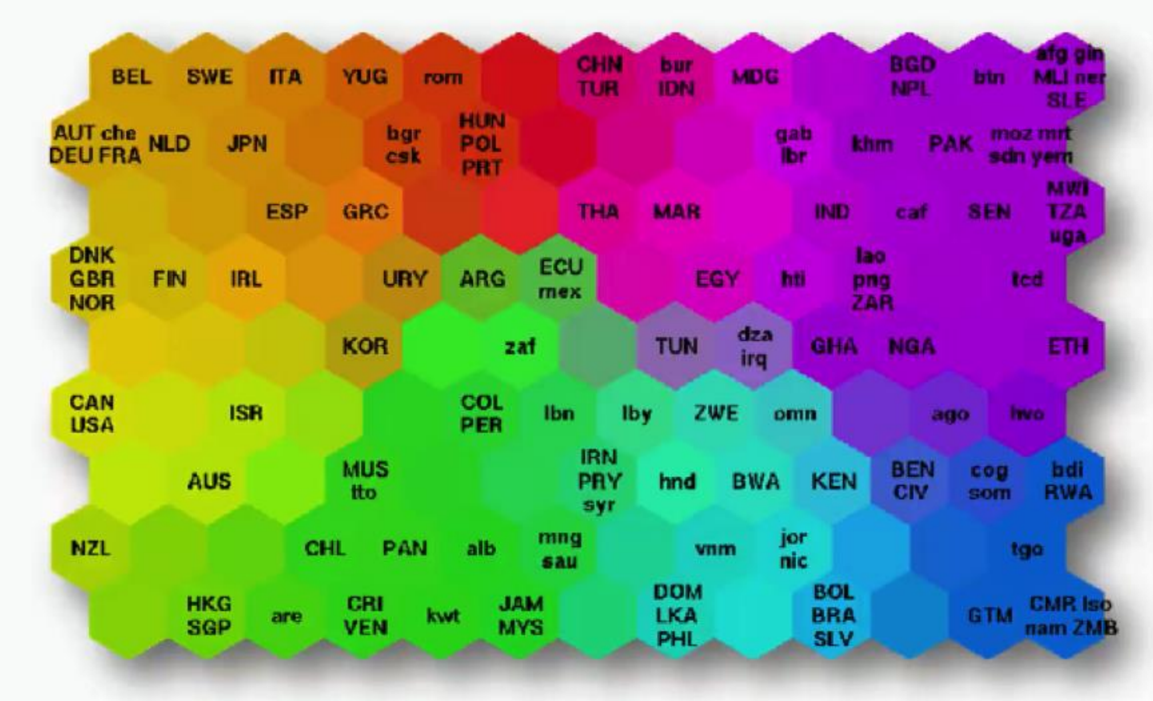


More advanced visualizations



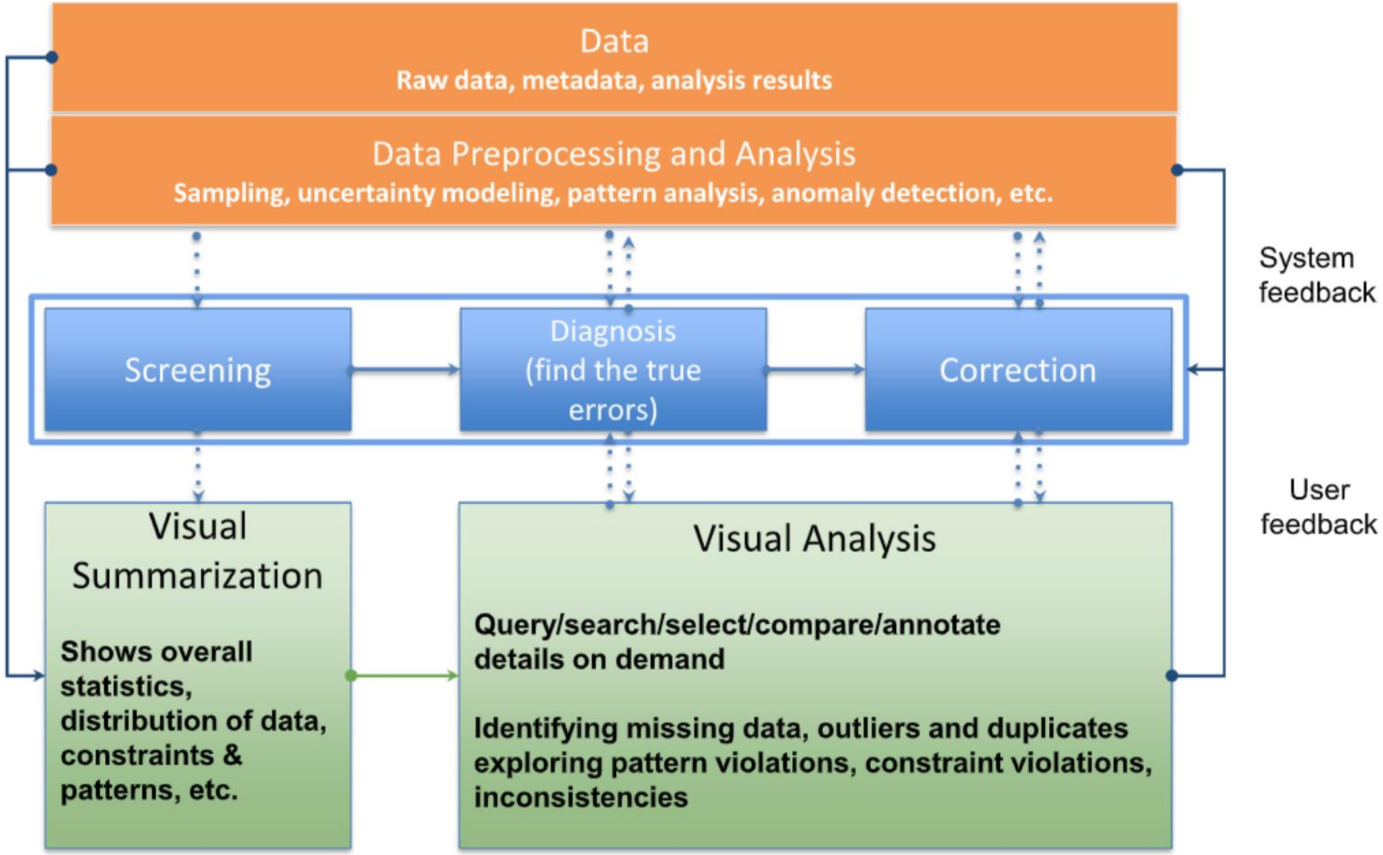
Scatterplot-matrix





Self Organizing Maps

Visual Analytics 4 Data Quality (VA4DQ): an overview



Data

Raw data, metadata, analysis results

Data Preprocessing and Analysis

Sampling, uncertainty modeling, pattern analysis, anomaly detection, etc.

Screening

Diagnosis (find the true errors)

Correction

System feedback

User feedback

Visual Summarization

Shows overall statistics, distribution of data, constraints & patterns, etc.

Visual Analysis

Query/search/select/compare/annotate details on demand

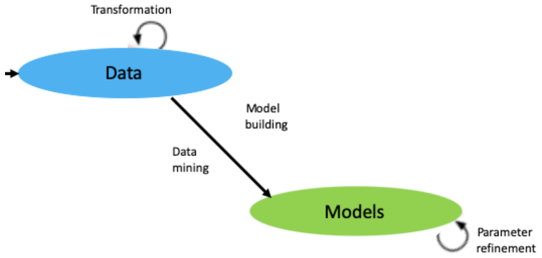
Identifying missing data, outliers and duplicates exploring pattern violations, constraint violations, inconsistencies

Liu, S., Andrienko, G., Wu, Y., Cao, N., Jiang, L., Shi, C., ... & Hong, S. (2018). **Steering data quality with visual analytics: The complexity challenge**. *Visual Informatics*, *2*(4), 191-197.

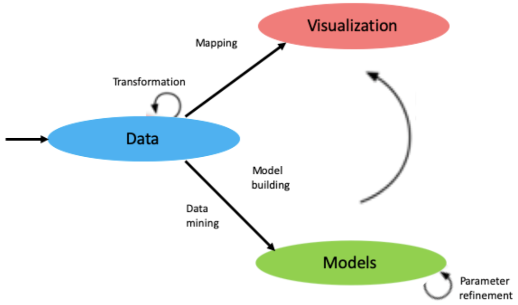
VA4DQ: an overview

The flavor of integration can be quite different

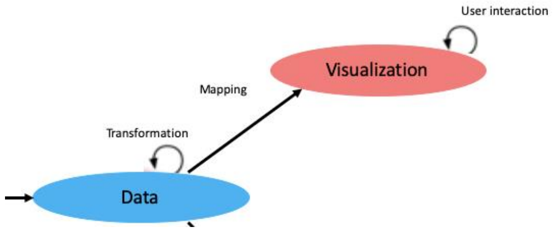
Data



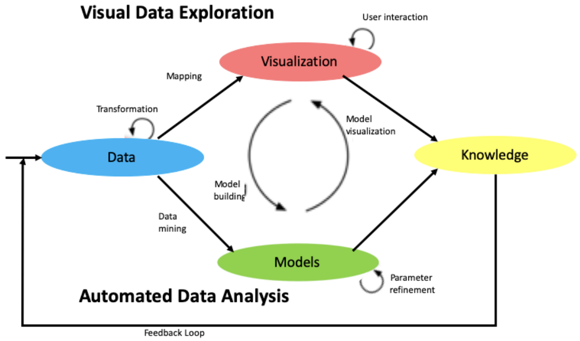
Only numerical



Visualization of results



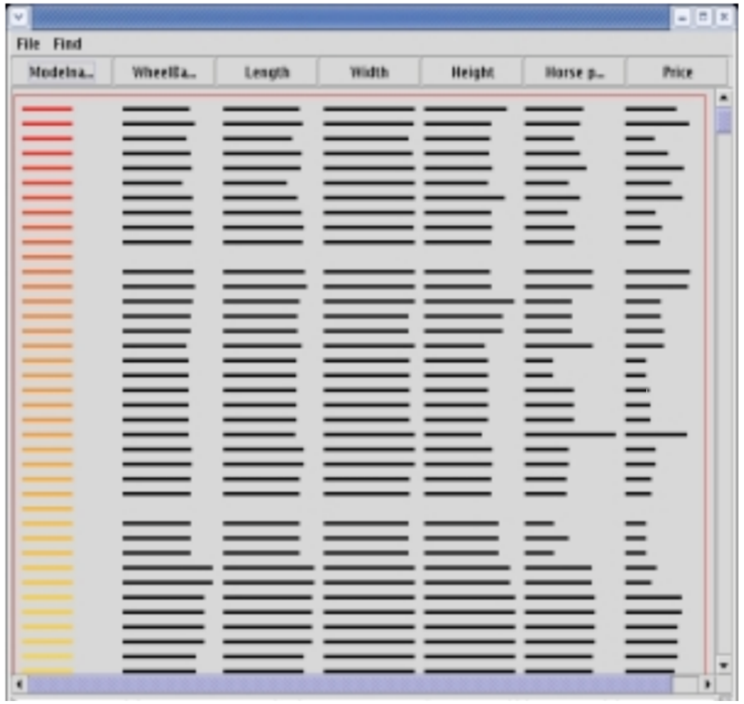
Visualization with Basic interaction



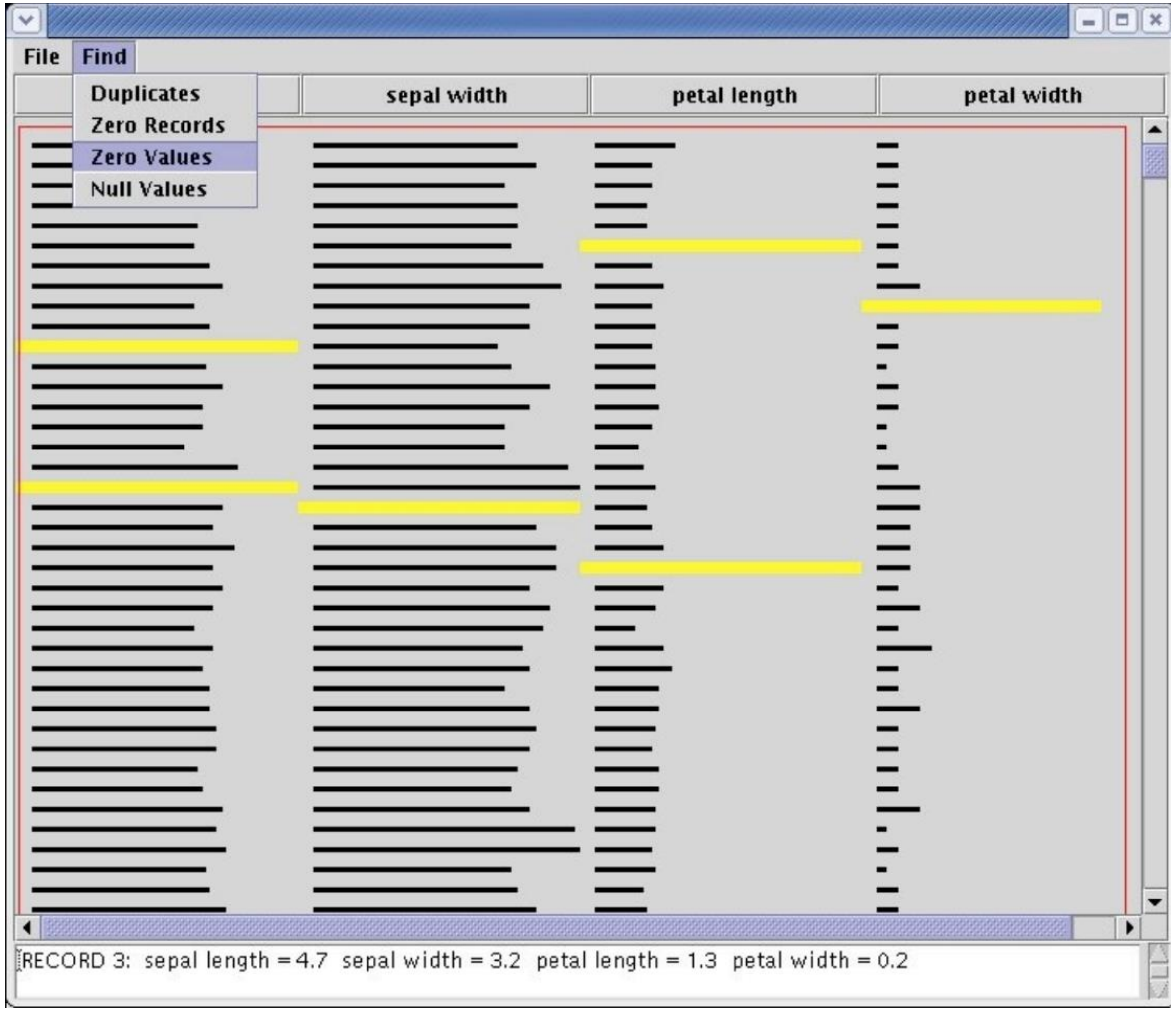
Visual Analytics

What state-of-the-art research proposes for ”Data Quality” ?

DaVis: a tool for visualizing Data Quality



Sulo, Rajmonda, Stephen Eick, and Robert Grossman. "DaVis: a tool for visualizing data quality." *Posters Compendium of InfoVis* 2005 (2005): 45-46.

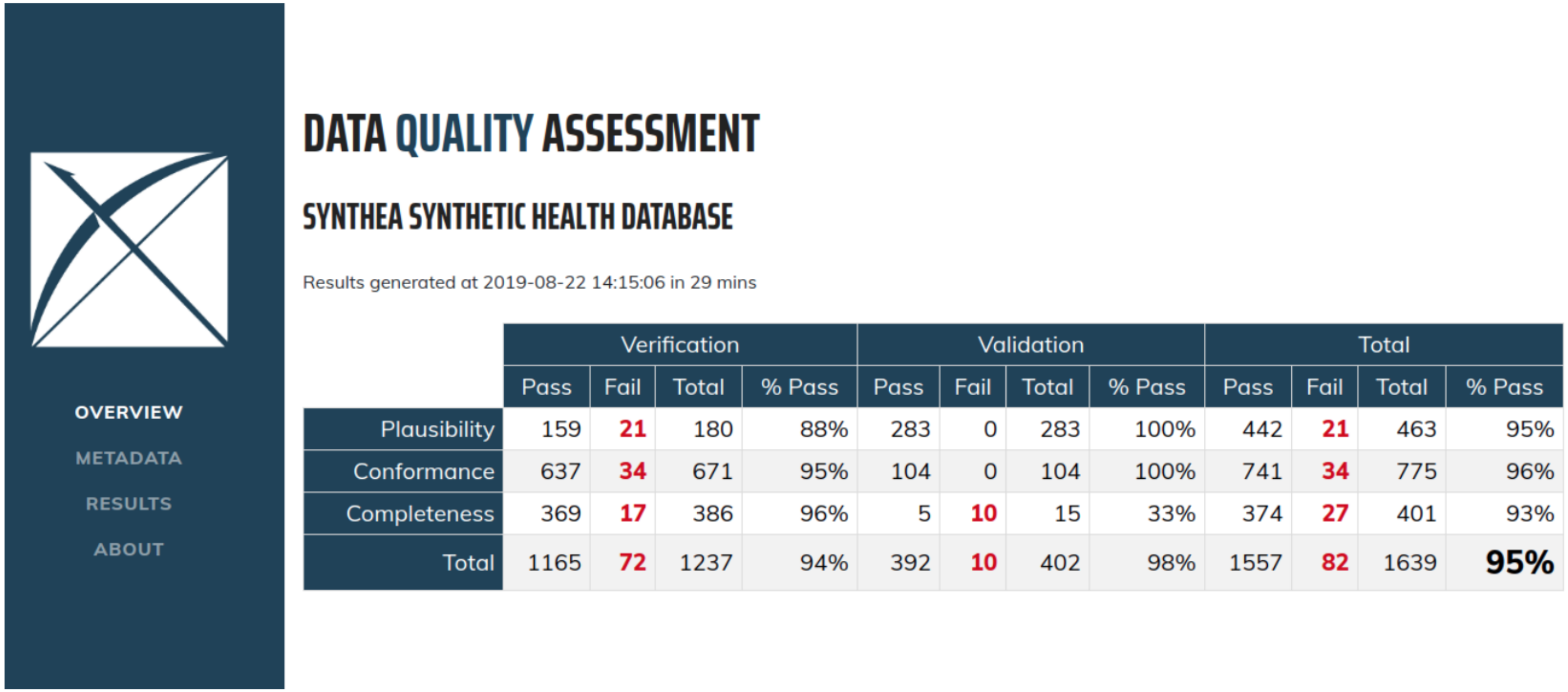


RECORD 3: sepal length = 4.7 sepal width = 3.2 petal length = 1.3 petal width = 0.2

Yellow represents zero values

Visual Data Quality Dashboard

- It uses R for computaing indicators
- Strongly oriented at reporting
- Stiil oriented at tables (less scalability)



DATA QUALITY ASSESSMENT

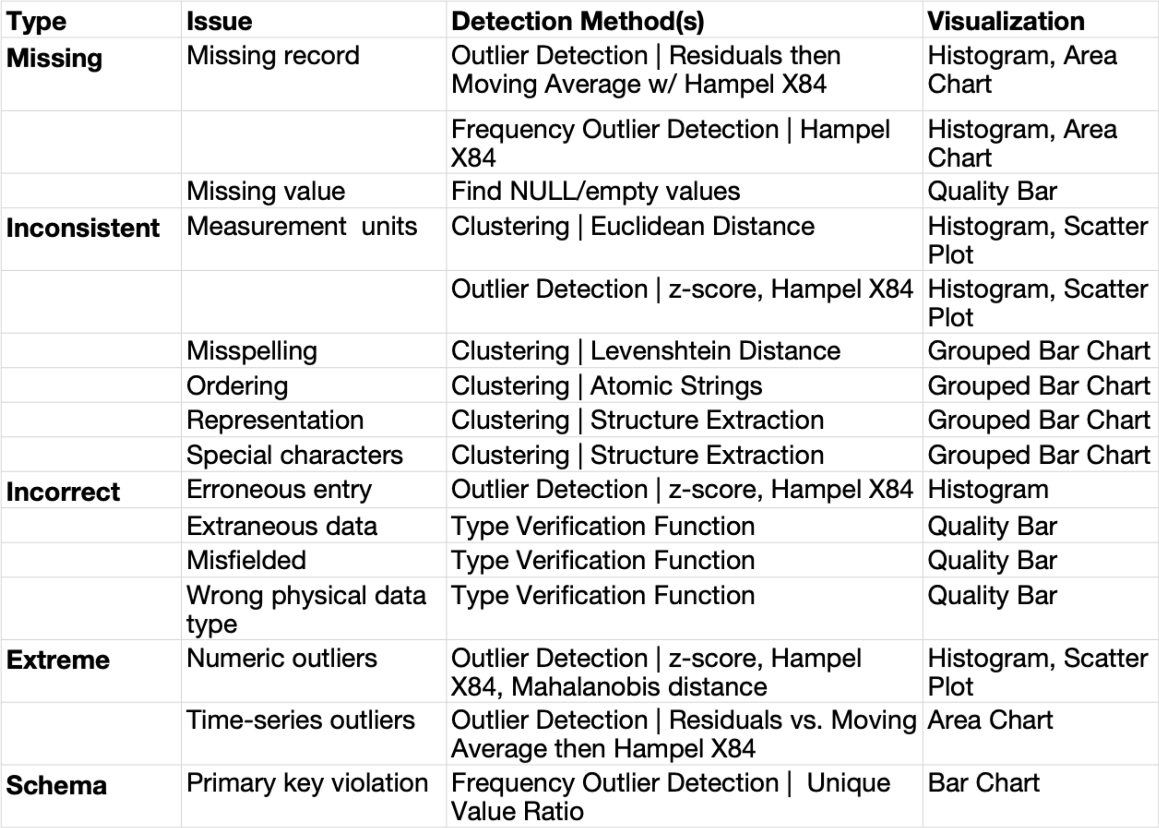
SYNTHEA SYNTHETIC HEALTH DATABASE

Results generated at 2019-08-22 14:15:06 in 29 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	159	21	180	88%	283	0	283	100%	442	21	463	95%
Conformance	637	34	671	95%	104	0	104	100%	741	34	775	96%
Completeness	369	17	386	96%	5	10	15	33%	374	27	401	93%
Total	1165	72	1237	94%	392	10	402	98%	1557	82	1639	**95%**

https://github.com/OHDSI/DataQualityDashboard

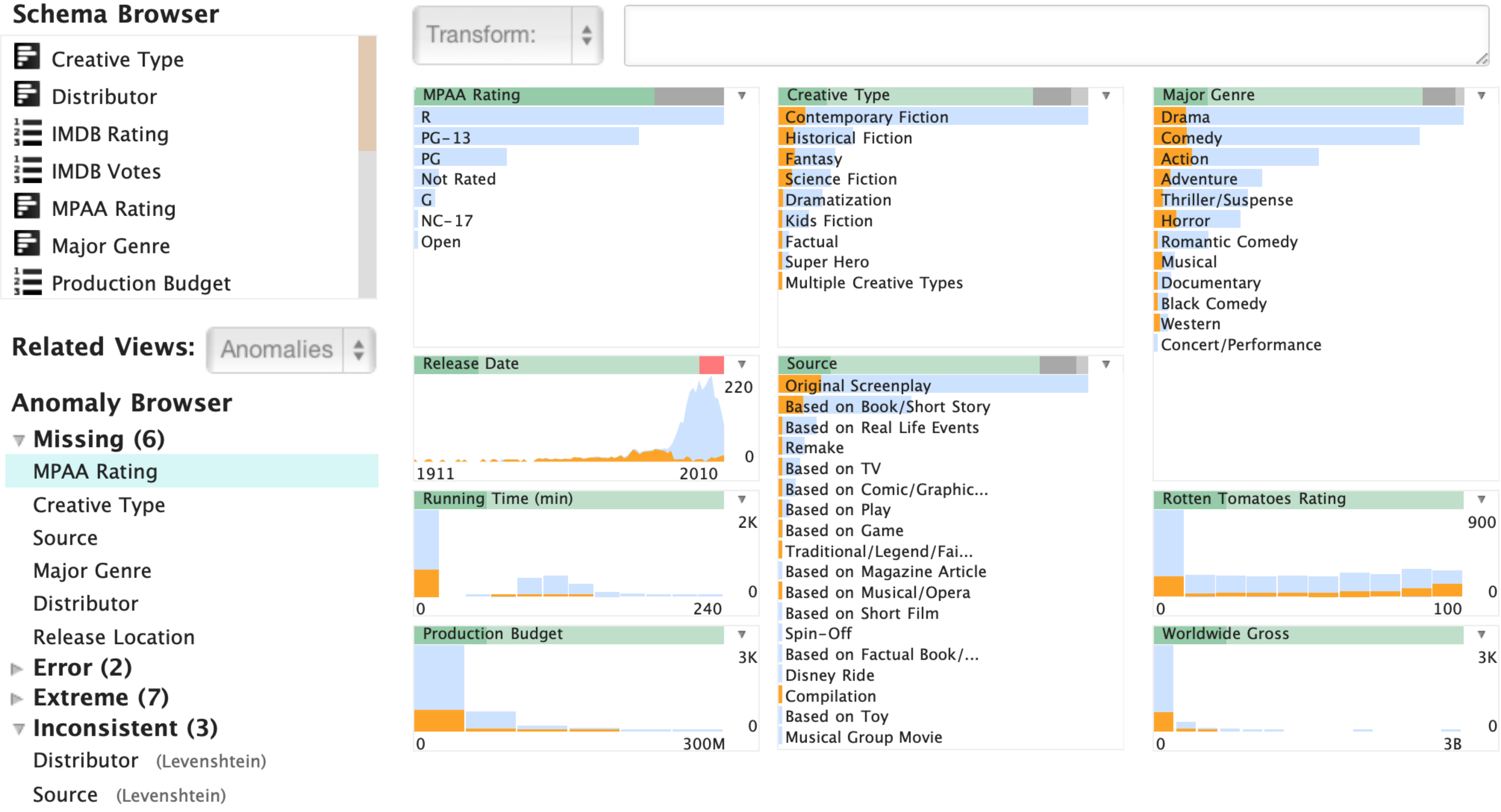
Profiler



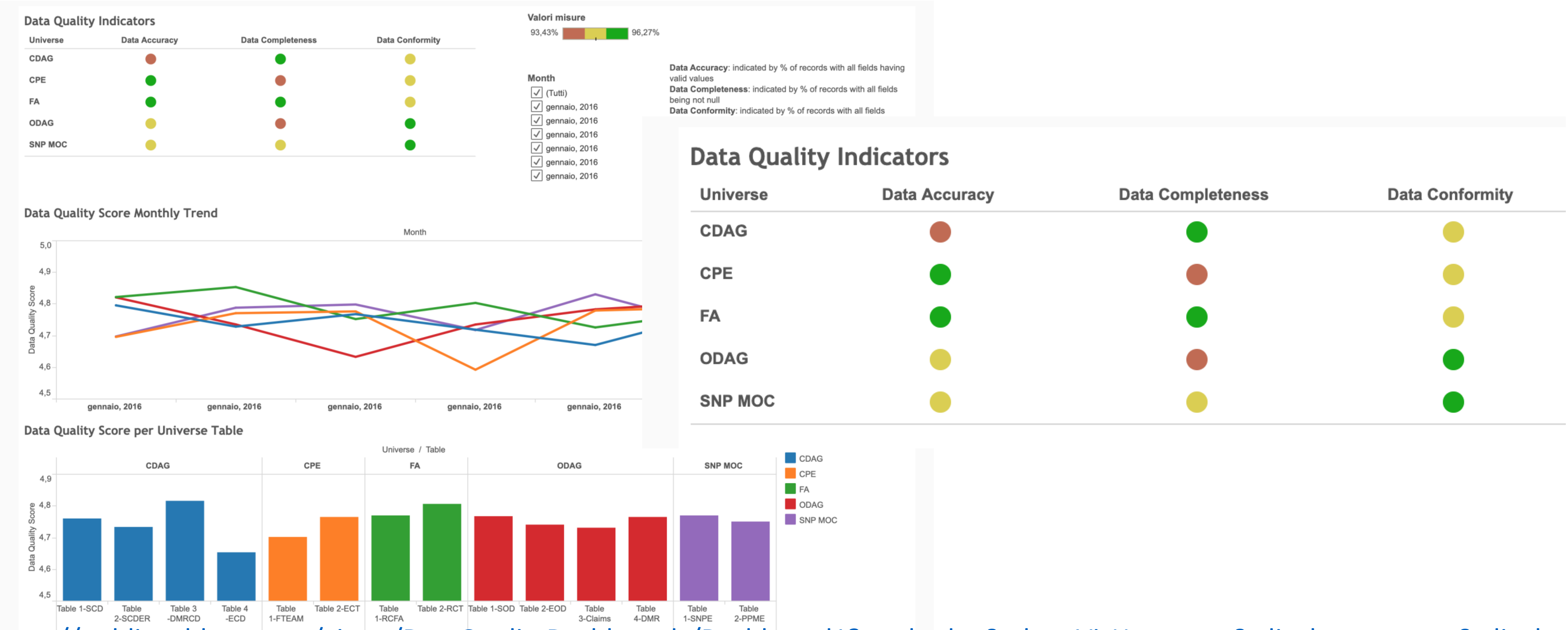
Type	Issue	Detection Method(s)	Visualization
Missing	Missing record	Outlier Detection | Residuals then Moving Average w/ Hampel X84	Histogram, Area Chart
		Frequency Outlier Detection | Hampel X84	Histogram, Area Chart
	Missing value	Find NULL/empty values	Quality Bar
Inconsistent	Measurement units	Clustering | Euclidean Distance	Histogram, Scatter Plot
		Outlier Detection | z-score, Hampel X84	Histogram, Scatter Plot
	Misspelling	Clustering | Levenshtein Distance	Grouped Bar Chart
	Ordering	Clustering | Atomic Strings	Grouped Bar Chart
	Representation	Clustering | Structure Extraction	Grouped Bar Chart
	Special characters	Clustering | Structure Extraction	Grouped Bar Chart
Incorrect	Erroneous entry	Outlier Detection | z-score, Hampel X84	Histogram
	Extraneous data	Type Verification Function	Quality Bar
	Misfielded	Type Verification Function	Quality Bar
	Wrong physical data type	Type Verification Function	Quality Bar
Extreme	Numeric outliers	Outlier Detection | z-score, Hampel X84, Mahalanobis distance	Histogram, Scatter Plot
	Time-series outliers	Outlier Detection | Residuals vs. Moving Average then Hampel X84	Area Chart
Schema	Primary key violation	Frequency Outlier Detection | Unique Value Ratio	Bar Chart

Sean Kandel, Ravi Parikh, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. Profiler: integrated statistical analysis and visualization for data quality assessment. In Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '12). Association for Computing Machinery, New York, NY, USA, 547–554. DOI:https://doi.org/10.1145/2254556.2254659

Profiler


Schema Browser
Creative Type
Distributor
IMDB Rating
IMDB Votes
MPAA Rating
Major Genre
Production Budget
Related Views:
Anomalies
Anomaly Browser
Missing (6)
MPAA Rating
Creative Type
Source
Major Genre
Distributor
Release Location
Error (2)
Extreme (7)
Inconsistent (3)
Distributor (Levenshtein)
Source (Levenshtein)
Transform:
MPAA Rating
R
PG-13
PG
Not Rated
G
NC-17
Open
Creative Type
Contemporary Fiction
Historical Fiction
Fantasy
Science Fiction
Dramatization
Kids Fiction
Factual
Super Hero
Multiple Creative Types
Major Genre
Drama
Comedy
Action
Adventure
Thriller/Suspense
Horror
Romantic Comedy
Musical
Documentary
Black Comedy
Western
Concert/Performance
Release Date
220
0
1911
2010
Source
Original Screenplay
Based on Book/Short Story
Based on Real Life Events
Remake
Based on TV
Based on Comic/Graphic...
Based on Play
Based on Game
Traditional/Legend/Fai...
Based on Magazine Article
Based on Musical/Opera
Based on Short Film
Spin-Off
Based on Factual Book/...
Disney Ride
Compilation
Based on Toy
Musical Group Movie
Running Time (min)
2K
0
0
240
Production Budget
3K
0
0
300M
Rotten Tomatoes Rating
900
0
0
100
Worldwide Gross
3K
0
0
3B

General Environments: Tableau



Data Quality Indicators

Universe	Data Accuracy	Data Completeness	Data Conformity
CDAG	🔴	🟢	🟡
CPE	🟢	🔴	🟡
FA	🟢	🟢	🟡
ODAG	🟡	🔴	🟢
SNP MOC	🟡	🟡	🟢

Valori misure

93,43% 96,27%

Month

☑ (Tutti)
☑ gennaio, 2016
☑ gennaio, 2016
☑ gennaio, 2016
☑ gennaio, 2016
☑ gennaio, 2016
☑ gennaio, 2016

Data Accuracy: indicated by % of records with all fields having valid values
Data Completeness: indicated by % of records with all fields being not null
Data Conformity: indicated by % of records with all fields

Data Quality Score Monthly Trend

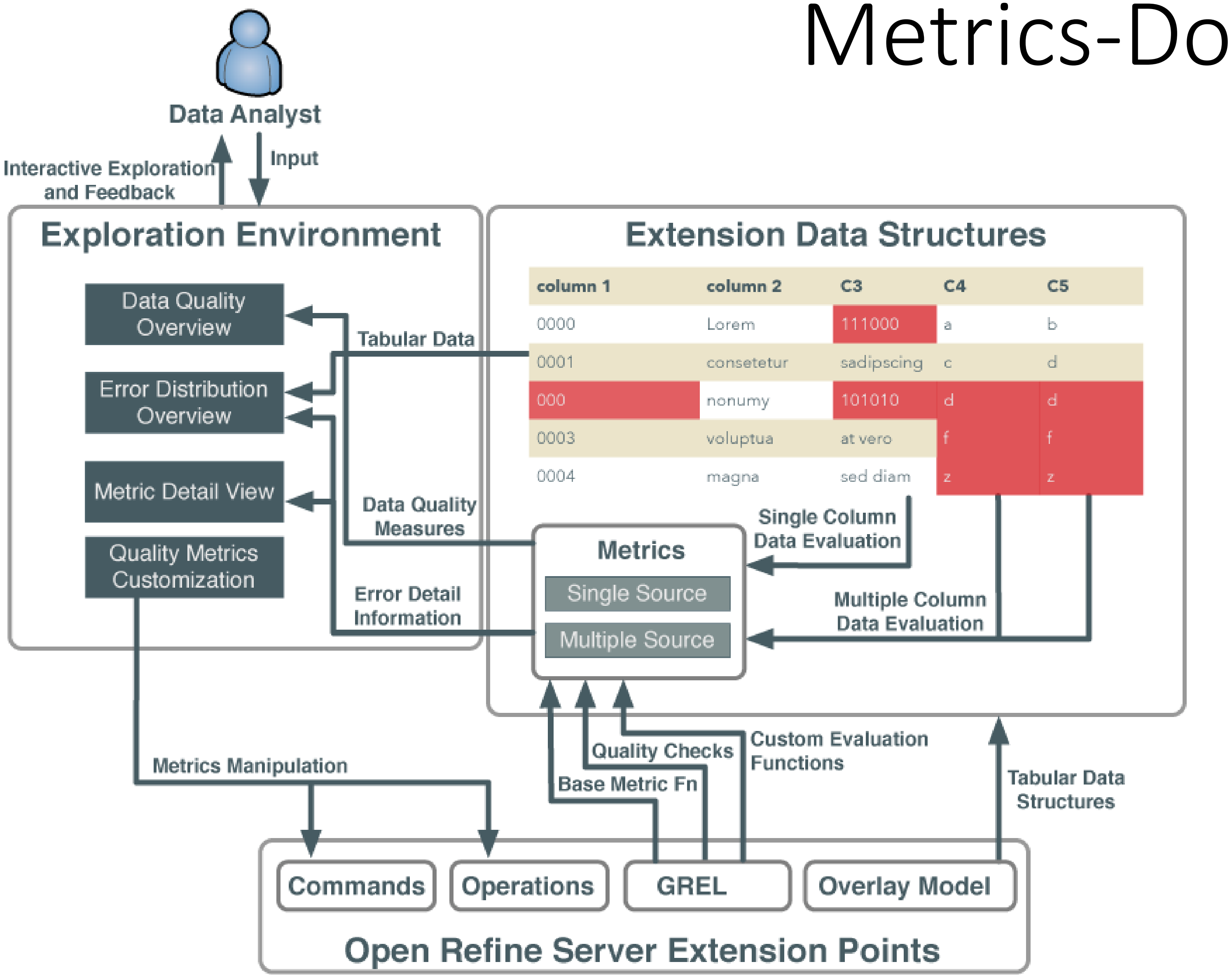
Data Quality Score per Universe Table

Data Quality Indicators

Universe	Data Accuracy	Data Completeness	Data Conformity
CDAG	🔴	🟢	🟡
CPE	🟢	🔴	🟡
FA	🟢	🟢	🟡
ODAG	🟡	🔴	🟢
SNP MOC	🟡	🟡	🟢

https://public.tableau.com/views/DataQualityDashboards/Dashboard1?:embed=y&:showVizHome=no&:display_count=y&:display_static_image=y&:bootstrapWhenNotified=true

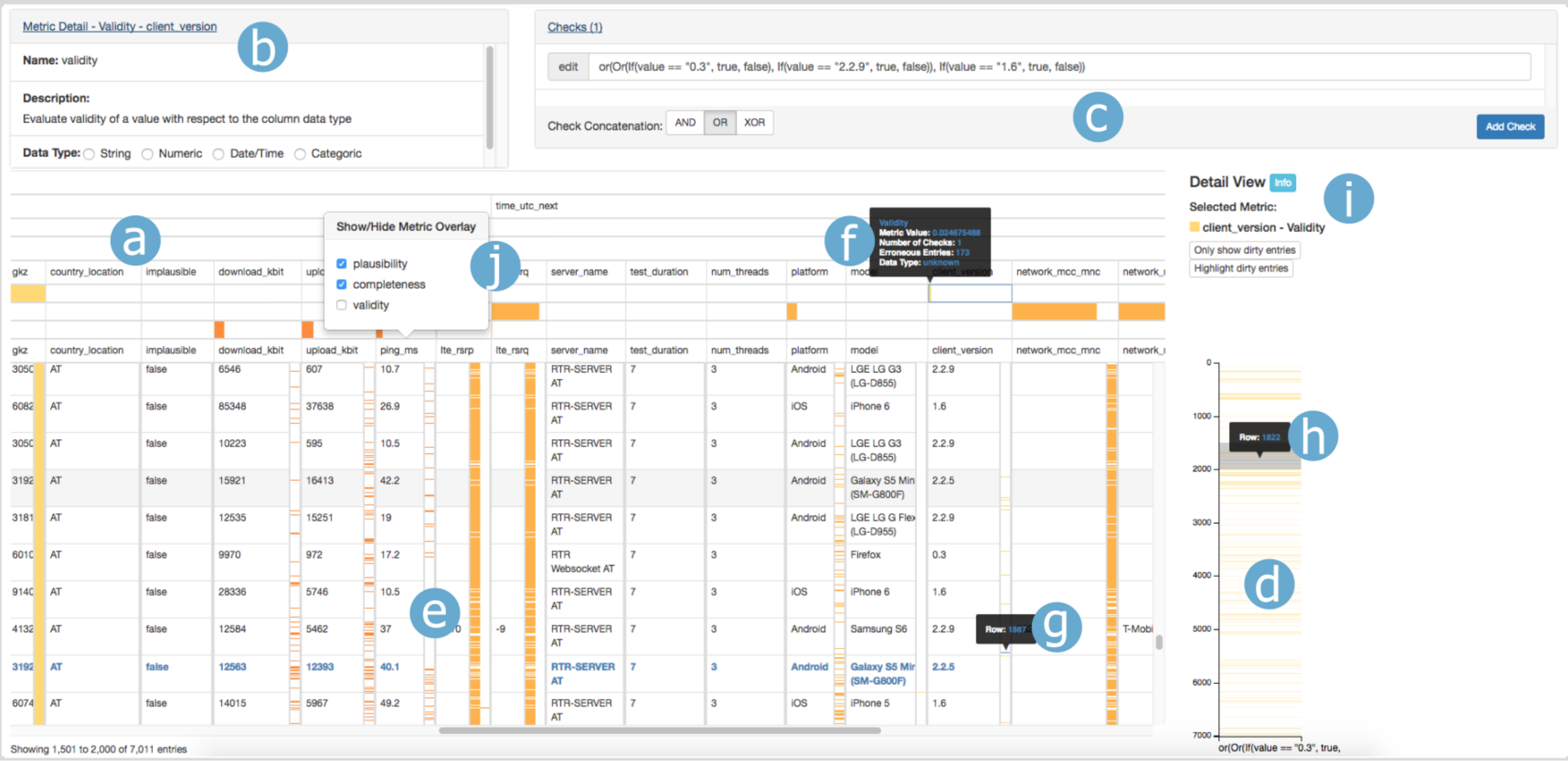
Metrics-Doc



column 1	column 2	C3	C4	C5
0000	Lorem	111000	a	b
0001	consetetur	sadipscing	c	d
000	nonumy	101010	d	d
0003	voluptua	at vero	f	f
0004	magna	sed diam	z	z

Christian Bors, Theresia Gschwandtner, Simone Kriglstein, Silvia Miksch, and Margit Pohl. 2018. Visual Interactive Creation, Customization, and Analysis of Data Quality Metrics. <i>J. Data and Information Quality</i> 10, 1, Article 3 (May 2018), 26 pages. DOI:https://doi.org/10.1145/3190578

Metrics-Doc



Metric Detail - Validity - client_version

Name: validity

Description:
Evaluate validity of a value with respect to the column data type

Data Type: ○ String ○ Numeric ○ Date/Time ○ Categoric

Checks (1)

```
or(Or(If(value == "0.3", true, false), If(value == "2.2.9", true, false)), If(value == "1.6", true, false))
```

Check Concatenation: AND OR XOR

☑ plausibility
☑ completeness
☐ validity

time_utc_next

Validity
Metric Value: 0.024675488
Number of Checks: 1
Erroneous Entries: 173
Data Type: unknown

gkz	country_location	implausible	download_kbit	upload_kbit	ping_ms	lte_rsrp	lte_rsrq	server_name	test_duration	num_threads	platform	model	client_version	network_mcc_mnc	network_i
3050	AT	false	6546	607	10.7			RTR-SERVER AT	7	3	Android	LGE LG G3 (LG-D855)	2.2.9		
6082	AT	false	85348	37638	26.9			RTR-SERVER AT	7	3	iOS	iPhone 6	1.6		
3050	AT	false	10223	595	10.5			RTR-SERVER AT	7	3	Android	LGE LG G3 (LG-D855)	2.2.9		
3192	AT	false	15921	16413	42.2			RTR-SERVER AT	7	3	Android	Galaxy S5 Min (SM-G800F)	2.2.5		
3181	AT	false	12535	15251	19			RTR-SERVER AT	7	3	Android	LGE LG G Flex (LG-D955)	2.2.9		
6010	AT	false	9970	972	17.2			RTR Websocket AT	7	3		Firefox	0.3		
9140	AT	false	28336	5746	10.5			RTR-SERVER AT	7	3	iOS	iPhone 6	1.6		
4132	AT	false	12584	5462	37	0	-9	RTR-SERVER AT	7	3	Android	Samsung S6	2.2.9		T-Mobi
3192	AT	false	12563	12393	40.1			RTR-SERVER AT	7	3	Android	Galaxy S5 Mir (SM-G800F)	2.2.5		
6074	AT	false	14015	5967	49.2			RTR-SERVER AT	7	3	iOS	iPhone 5	1.6		

Showing 1,501 to 2,000 of 7,011 entries

Detail View

Selected Metric:

🟨 client_version - Validity

or(Or(If(value == "0.3", true,

(a) *quality metrics overview distribution* heatmaps
(b) the metric information view and
(c) customization tabs
(d) the *metric detail view*

(e) the tabular *raw data view* enhanced with *error*
(f) mouseover tooltips provide detail information on
(g) metrics
(h) data errors

(j) metric distribution heatmaps can be enabled and disabled individually

Integrated VA environment

High level analysis of data:

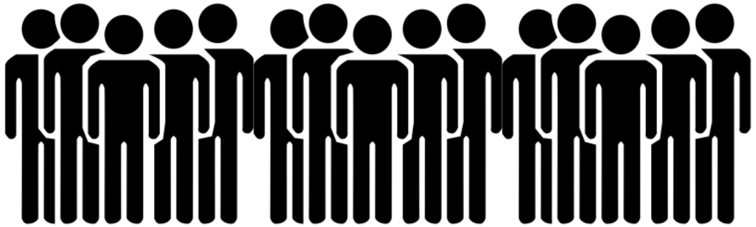
- Looking for correlation among data column
- Spotting outliers
- Exploiting coordination for understanding trends
- On research production data

Introduction

Research evaluation

- transition from a traditional evaluation model, based on bibliometric indicators of publications and citations
- modern evaluation, characterized by a multiplicity of distinct, complementary dimensions

Demand side (those that ask for research assessment) including an increase of institutional and internal assessments

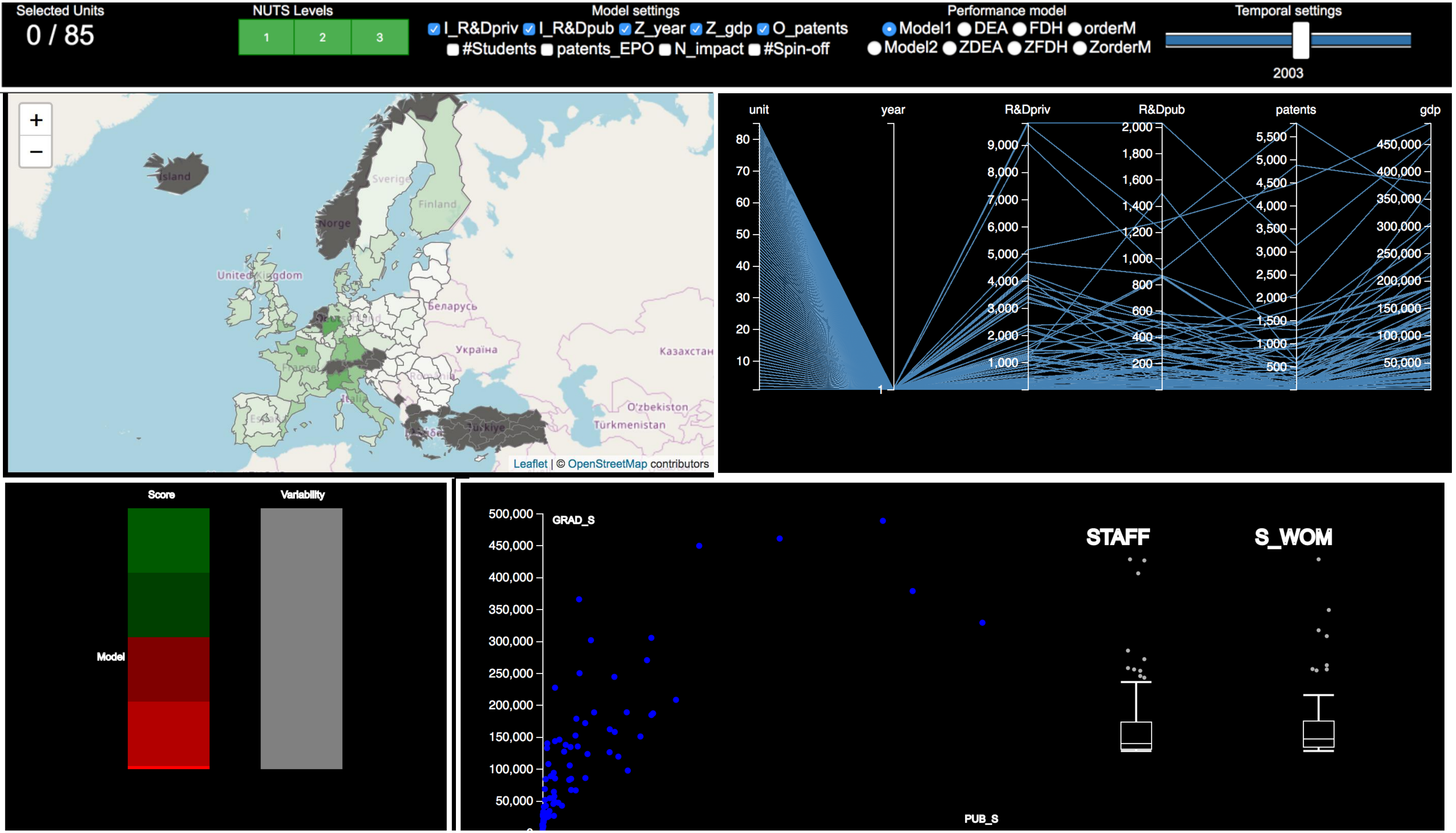


Supply side (those that offer research assessment) including proliferation of rankings, development of Altmetrics, open access repositories, new assessment tools and desktop bibliometrics

Scholars the increase of “publish or perish” pressure, impact on:

- incentives, behaviour and misconduct, and increasing critics against traditional bibliometric indicators
- the assessment process (increasing the complexity of the research assessment)
- the indicators’ development.

Visualization



Selected Units 0 / 85

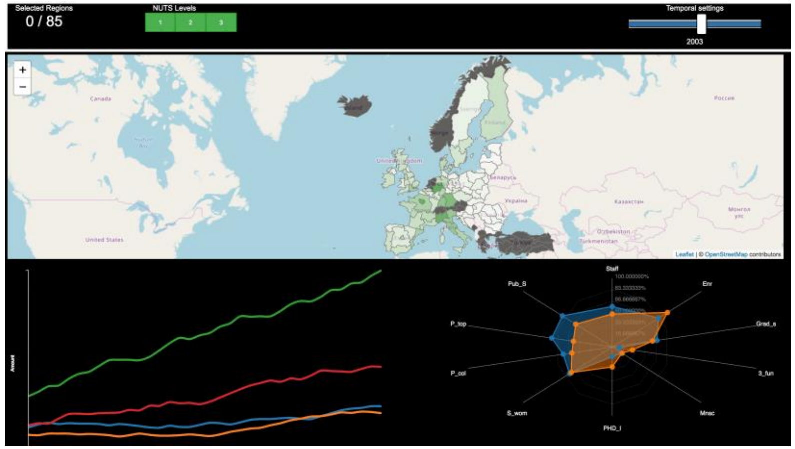
NUTS Levels 1 2 3

Model settings ☑ I_R&Dpriv ☑ I_R&Dpub ☑ Z_year ☑ Z_gdp ☑ O_patents □ #Students □ patents_EPO □ N_impact □ #Spin-off

Performance model ● Model1 ● DEA ● FDH ● orderM ● Model2 ● ZDEA ● ZFDH ● ZorderM

Temporal settings 2003

Visualization



Selected Regions 0 / 85

NUTS Levels 1 2 3

Temporal settings 2003

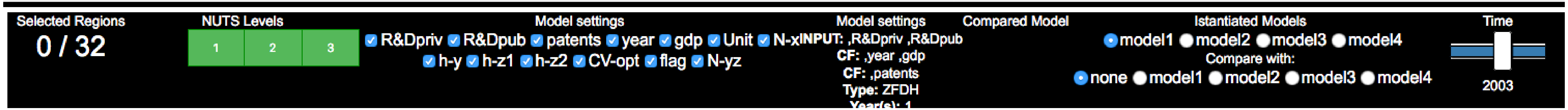
Angelini, M., Daraio, C., Lenzerini, M. *et al.* Performance model's development: a novel approach encompassing ontology-based data access and visual analytics. *Scientometrics* (2020). https://doi.org/10.1007/s11192-020-03689-x

Command bar

Selected model characteristics

Compared model characteristics

Temporal slider



Selected Regions
0 / 32

NUTS Levels
1 | 2 | 3

Model settings
☑ R&Dpriv ☑ R&Dpub ☑ patents ☑ year ☑ gdp ☑ Unit ☑ N-x
☑ h-y ☑ h-z1 ☑ h-z2 ☑ CV-opt ☑ flag ☑ N-yz

Model settings
INPUT: ,R&Dpriv ,R&Dpub
CF: ,year ,gdp
CF: ,patents
Type: ZFDH
Year(s): 1

Compared Model

Istantiated Models
◉ model1 ● model2 ● model3 ● model4
Compare with:
◉ none ● model1 ● model2 ● model3 ● model4

Time
2003

units

NUTS level
Selector

Selectable dimensions:

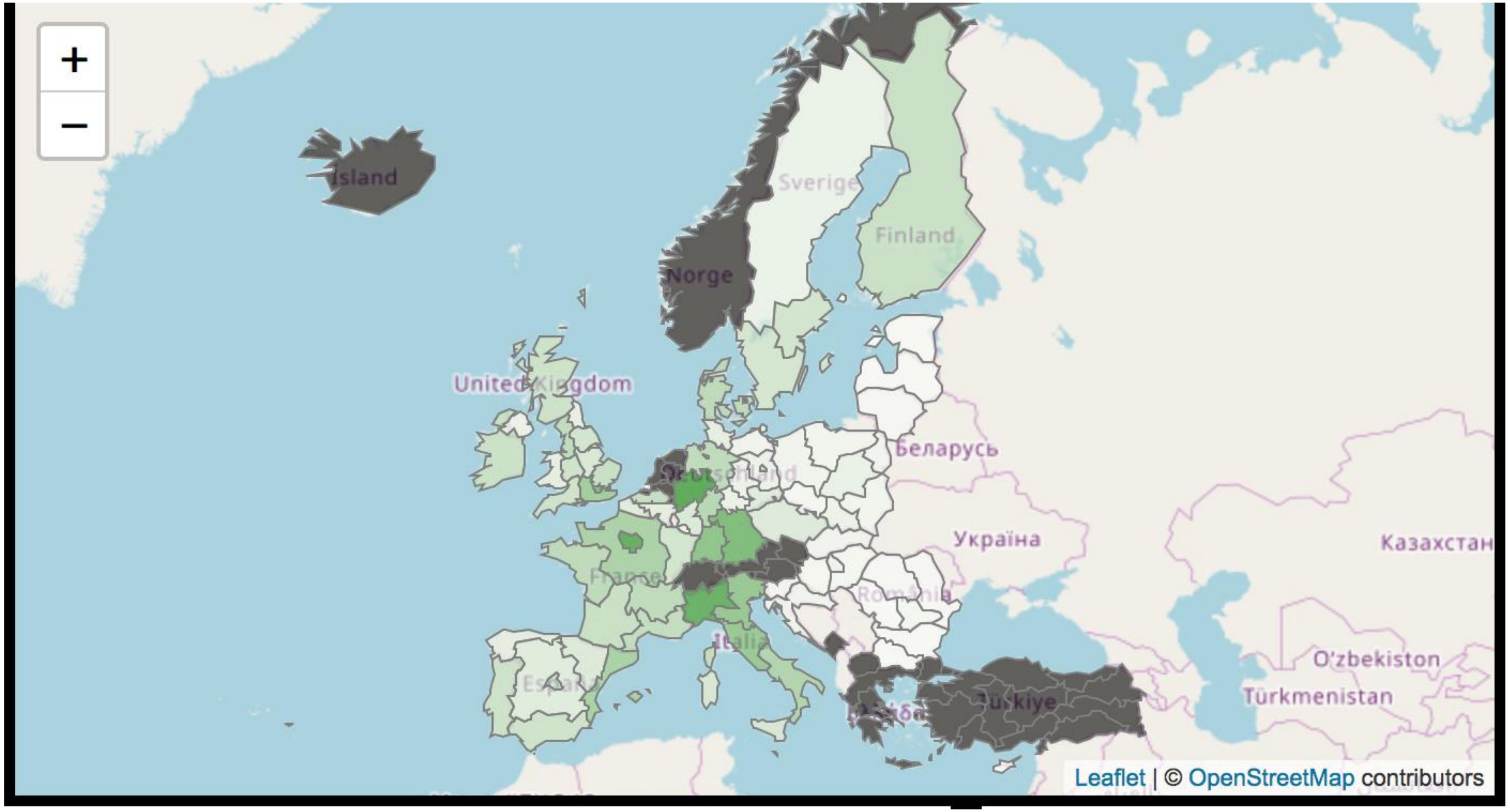
(add/remove dimensions
In the Parallel coordinates
For better readability)

Selectable models; TOP: **actual** model used

BOTTOM (Compare with). A model that stay
Fixed and will be compared with the **actual** one.

If you choose «none», no comparison will
Be made

Visualization: geographic view



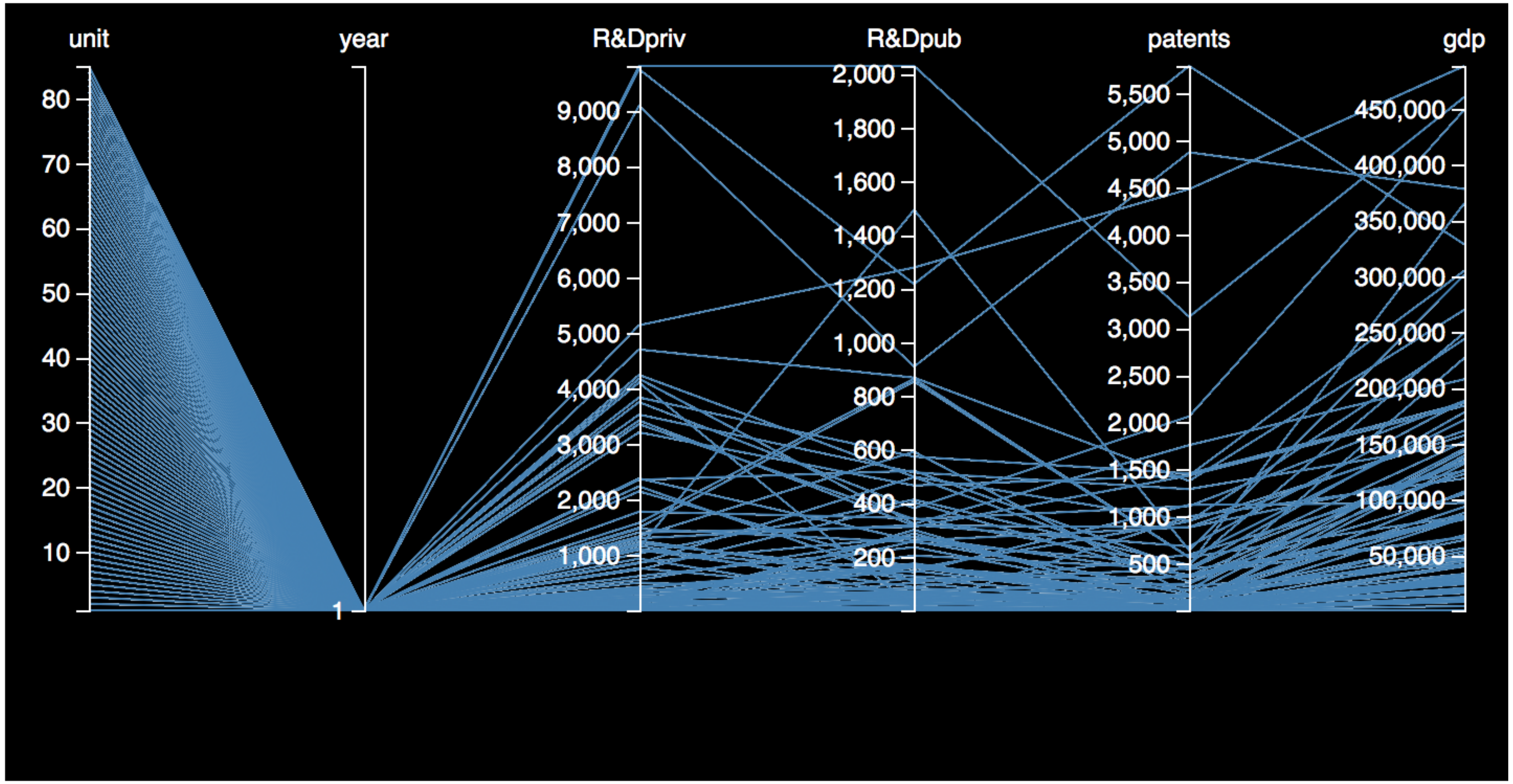
- The user can see one dimension, mapped with a color scale, for the right geographic aggregation level (NUTS levels)

🟩🟩🟩🟩🟩

low high

⬛ = no value

Visualization: parallel coordinates



- Each dimension is represented as a vertical axis

- Each tuple is a line (in blue) that pass for the value of each dimension

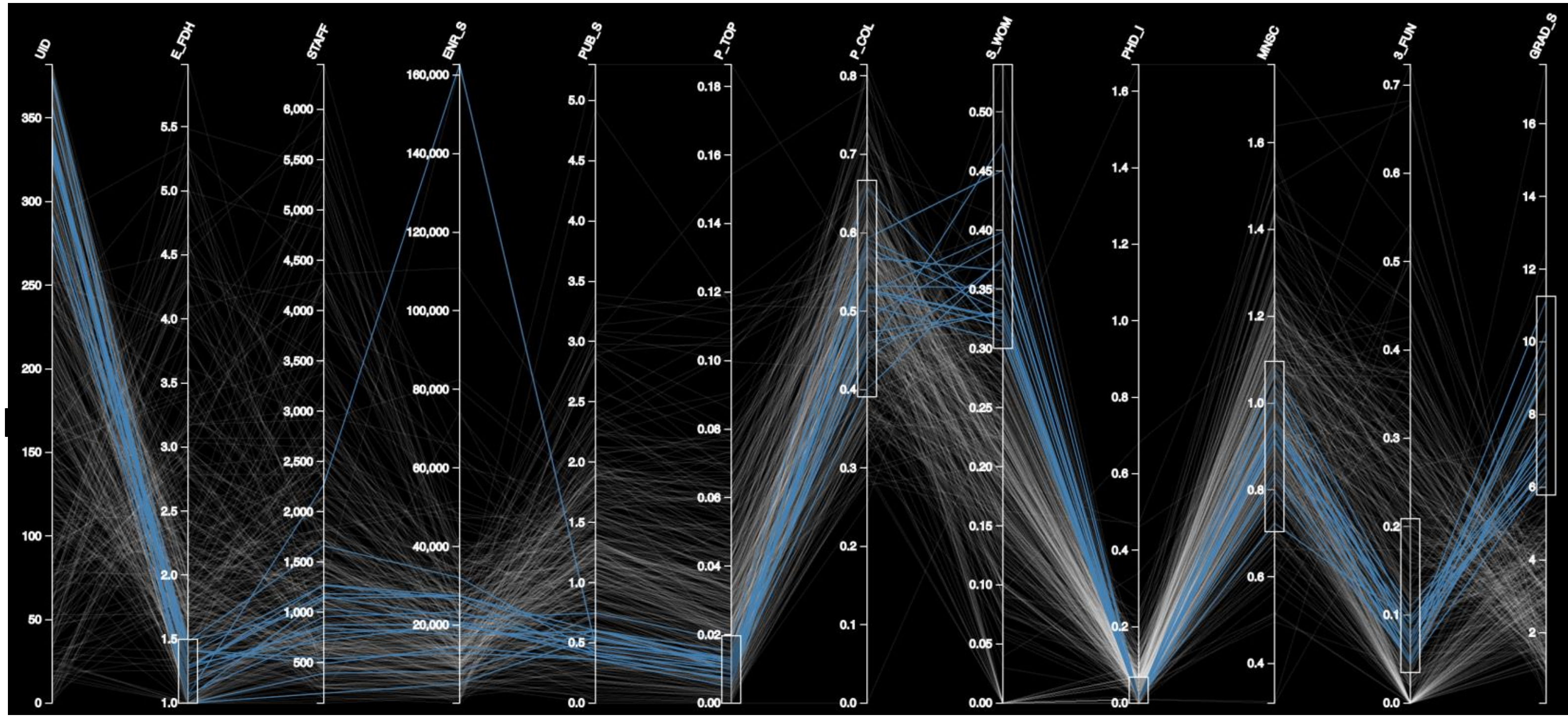
- We can see ALL the data (a lot) in a Compact view

- We can inspect for correlation between dimensions

Example: relations among dimensions

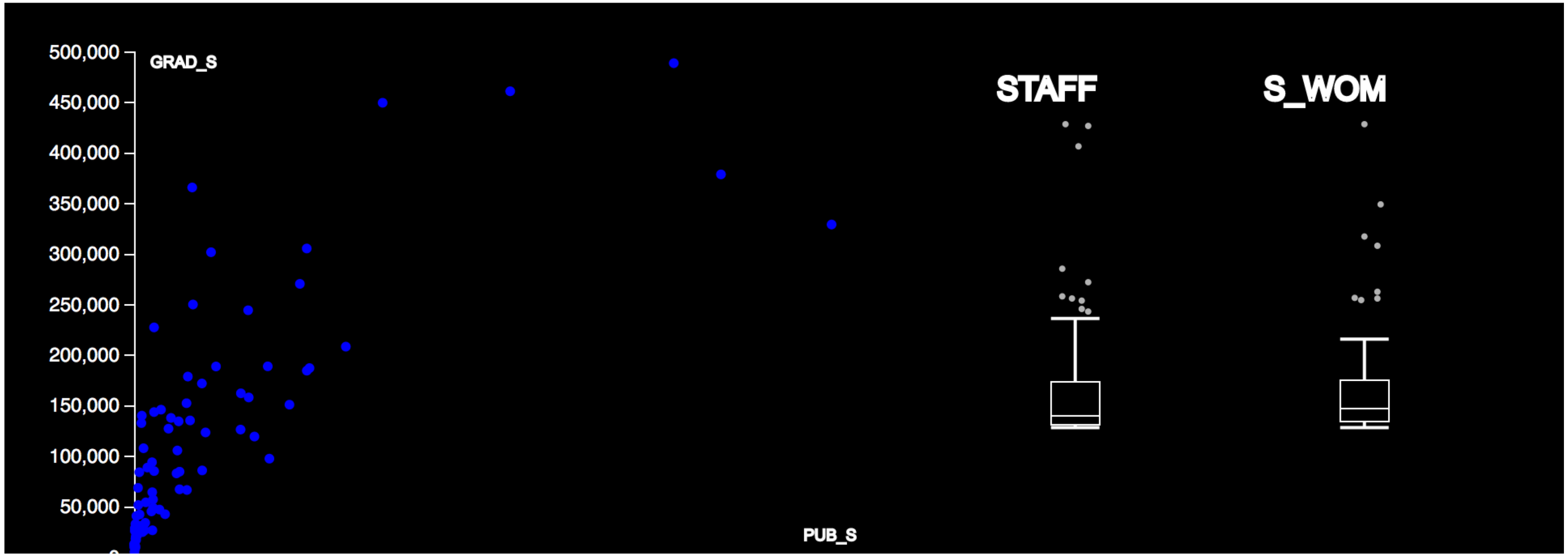
Axes (12 dimensions)

- UID institution id
- E_FDH is the FDH (in)efficiency score
- STAFF number of academic staff
- ENR_S enrolled students per academic staff
- PUB_S number of publications in WoS (fractional count) per academic staff
- P_TOP number of publications in top 10% of highly cited journals per academic staff
- P_COL percentage of papers done with international collaborations
- S_WOM share of women professors on total academic staff
- PHD_I PhD intensity
- MNCS Mean Normalized Citation
- 3_FUN share of third party funds
- GRAD_S is total number of graduates per academic staff



Among the most efficient units in teaching and research (i.e. E_FDH = [1 1.5]) there are those teaching oriented institutions (with the highest values of GRAD_S) in which the S_WOM is the highest ([0.30-0.50]): these are universities with almost zero PhD intensity that are able nevertheless to produce a small fraction of P_TOP publications with MNCS around the world average

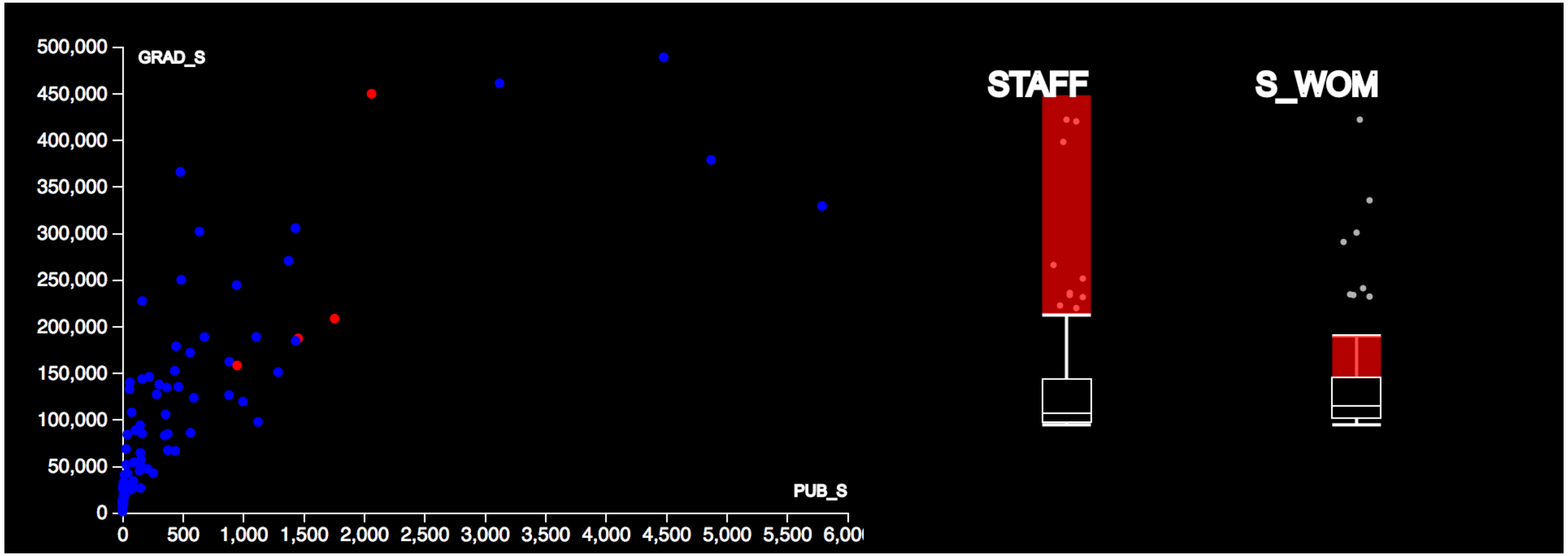
Visualization: scatterplot + boxplot



- The user see a 2D scatterplot looking again for correlation among couples of dimensions

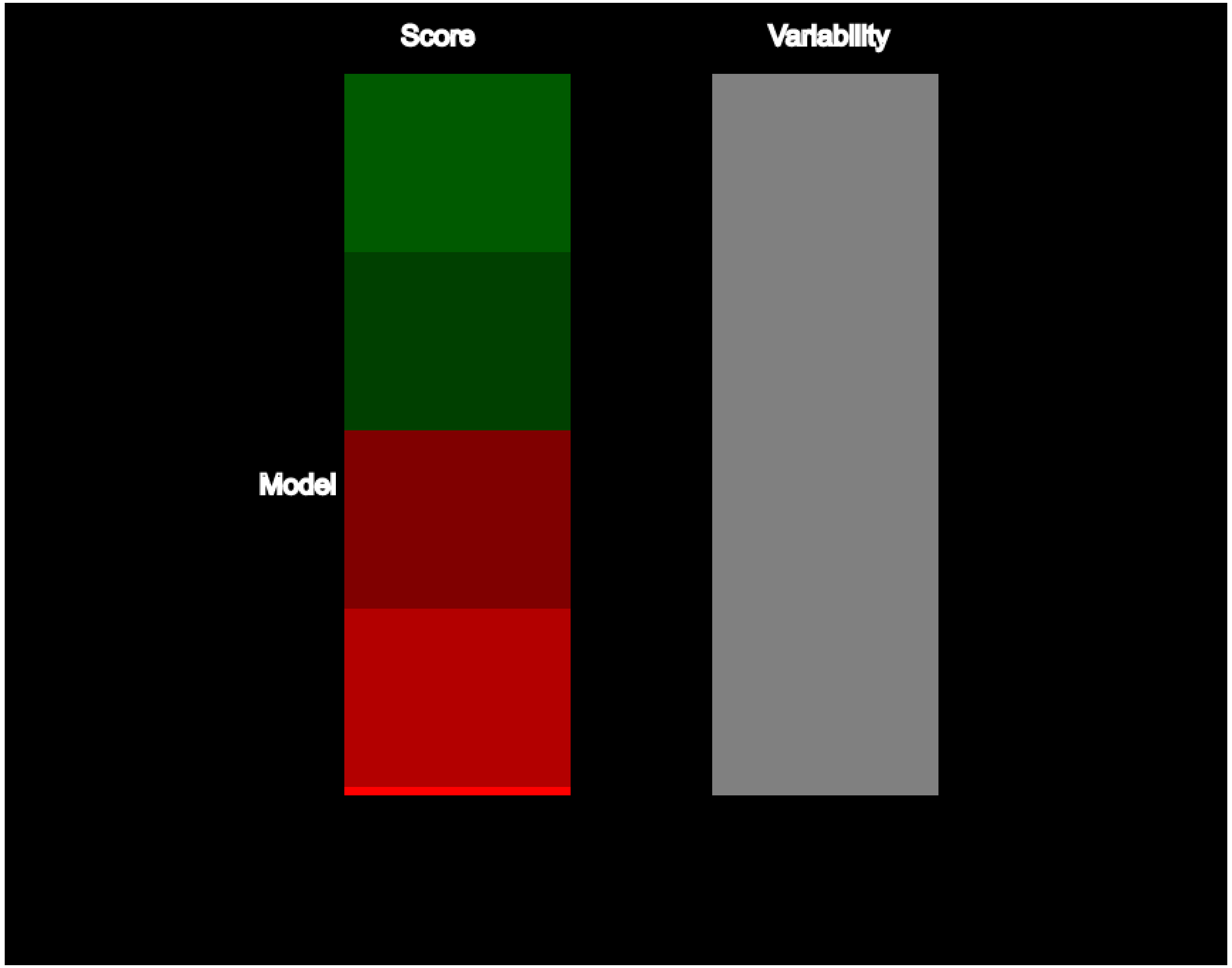
- She can filter from the boxplot (on the right) istantiated on different dimensions from the ones in the scatterplot

Example: relations among entities



Example of data filtering: with respect to all the units, the selection is composed by high outliers for academic staff (STAFF) and the 4th quartile for percentage of women staff (S_WOM); the resulting points are highlighted in red in the scatter plot, and the unit can be identified by mouse-hover.

Visualization: model performance



- Units are ordered (from top to bottom with respect to their performance score according to the selected model

- The color (green: GOOD scores, red=BAD scores) tell us How the units behave according to the model

- The second bargram (GREY) is used to calculate the variability of the model (how much the scores change if we perturb the model removing or changing one of its Parameters)

GREEN= good changes
RED= bad changes

For the same units!

All together: Demo

The Knowledge from both worlds

Data exploration: allowing comprehension of analyzed data

Knowledge

Additional analysis possible, e.g. Comparative analysis:
Identify inputs causing most uncertainty – direct research or information gathering

- Check the effect of model assumptions on model output
- Model simplification - identify inputs that do not affect the output, therefore redundant
- Better understanding of the model - what causes what.
- Corroboration or falsification
- Identifying errors - are there unexpected relationships between inputs and outputs?

assist with the decision making process

What-if analysis (testing different models/performance indicators/combinations of them)

Integrated VA environment: limitations

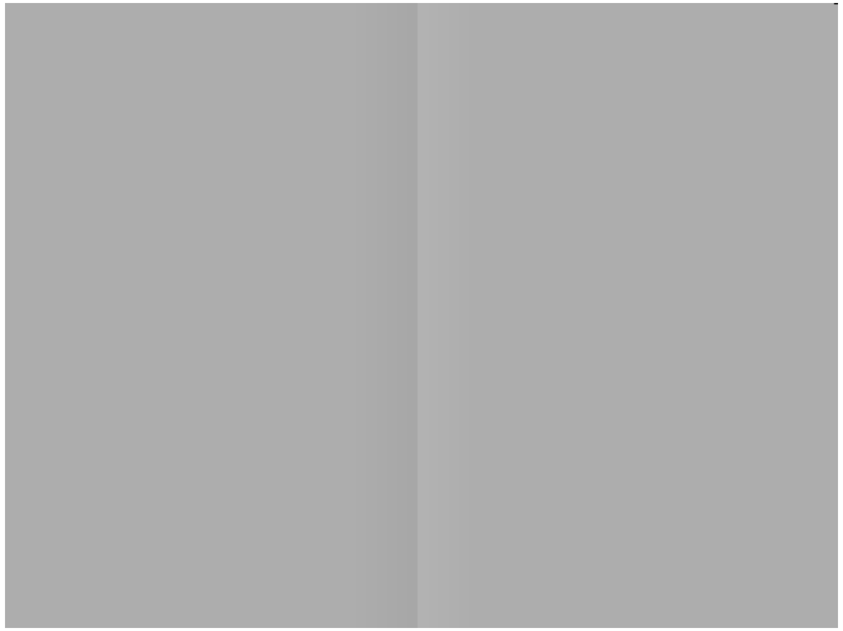
- Still working on completing data quality checks
- Modifications to dataset must be done outside of the tool
- Management of the data schema to improve
- Still in development

Visualization is a vast discipline

sharing
the
your
posting
Obama
McCain
Kerry
Bush
Gore
Bush

Cornsweet effect

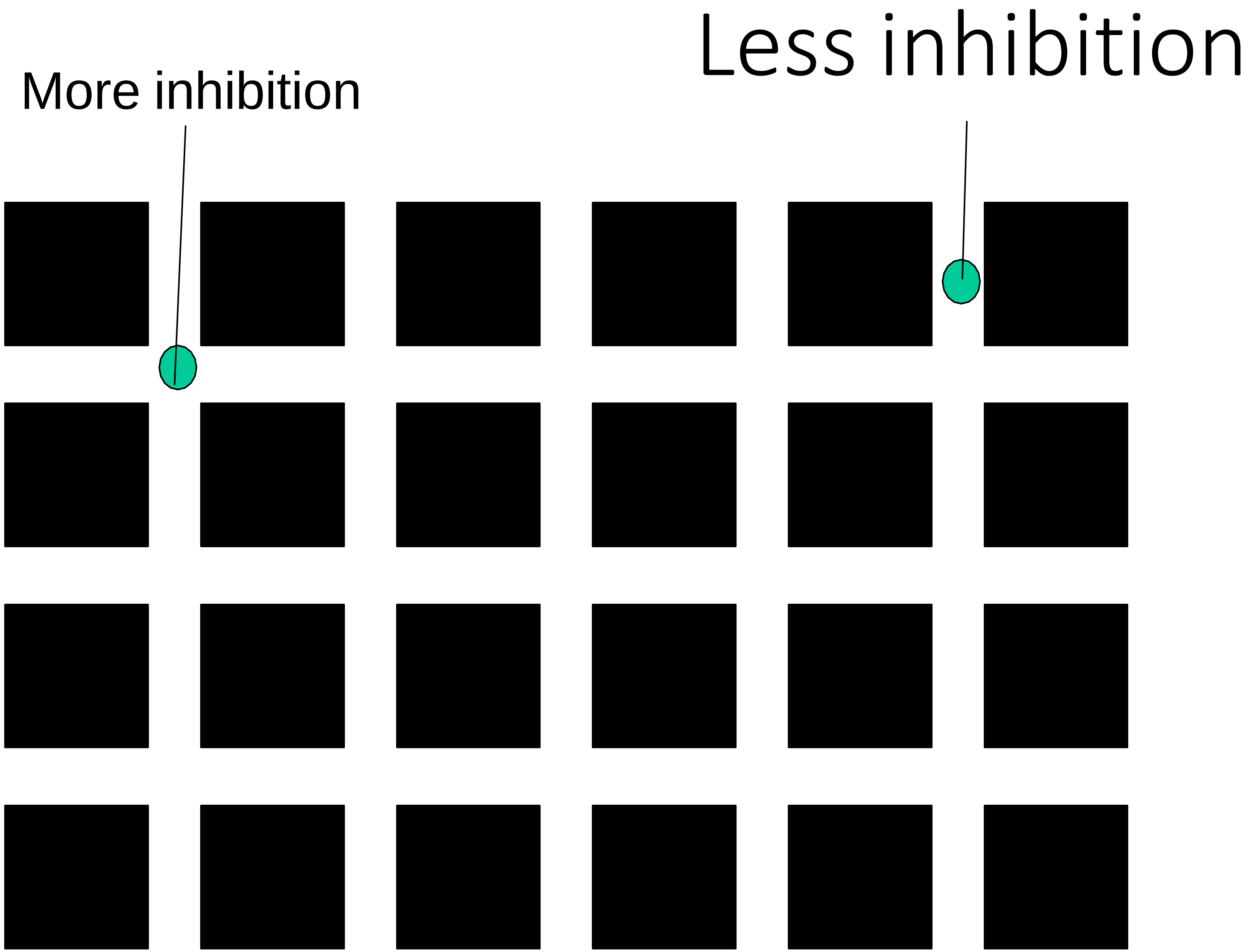
- Suitable shading creates edges and difference in lightness
- What is the darker side?



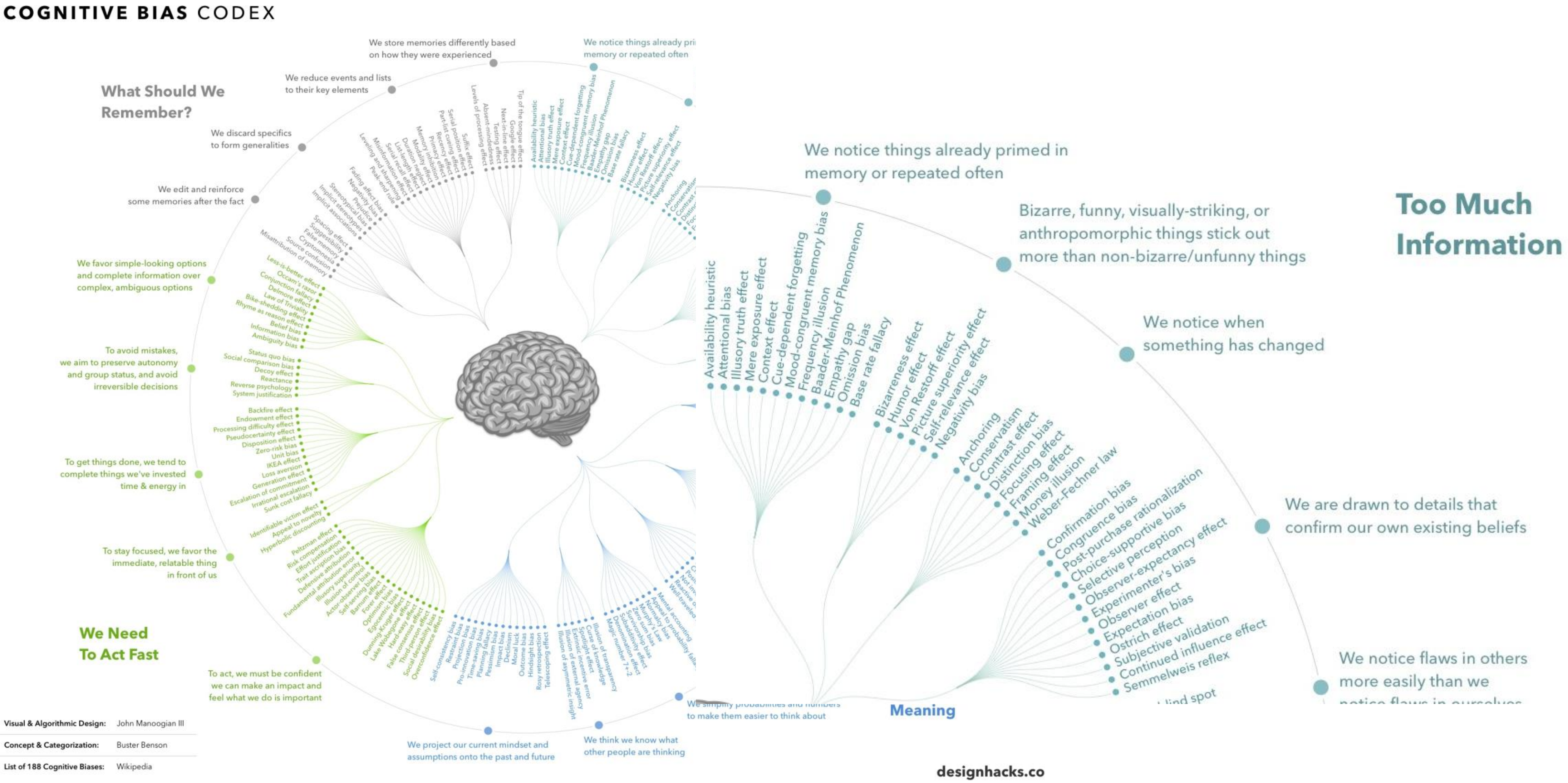
Cornsweet effect

- No one…

Human perceptual errors


Less inhibition
More inhibition

Human cognitive biases



Visual & Algorithmic Design: John Manoogian III

Concept & Categorization: Buster Benson

List of 188 Cognitive Biases: Wikipedia

designhacks.co

Research Challenges

Monitoring the Data quality evolution during time

Move from numeric representation (tables) to more exploratory analysis

- requirement of different visual paradigms to support different tasks

Better support Human intervention into the analysis (not just reporting)

Better support explanation of results

Research Challenges

- **Scalability**: existing visual data cleansing methods cannot be scaled to large scale datasets.
    - sample only a small subset of the whole training set. The challenge here is how to develop effective sampling methods that can both keep the data density and preserve important data such as influential points, outliers, and exceptions.

- there is a lack of effective quality metrics to measure the quality of different types of data such as textual data, images, videos, graph data, and trajectory data

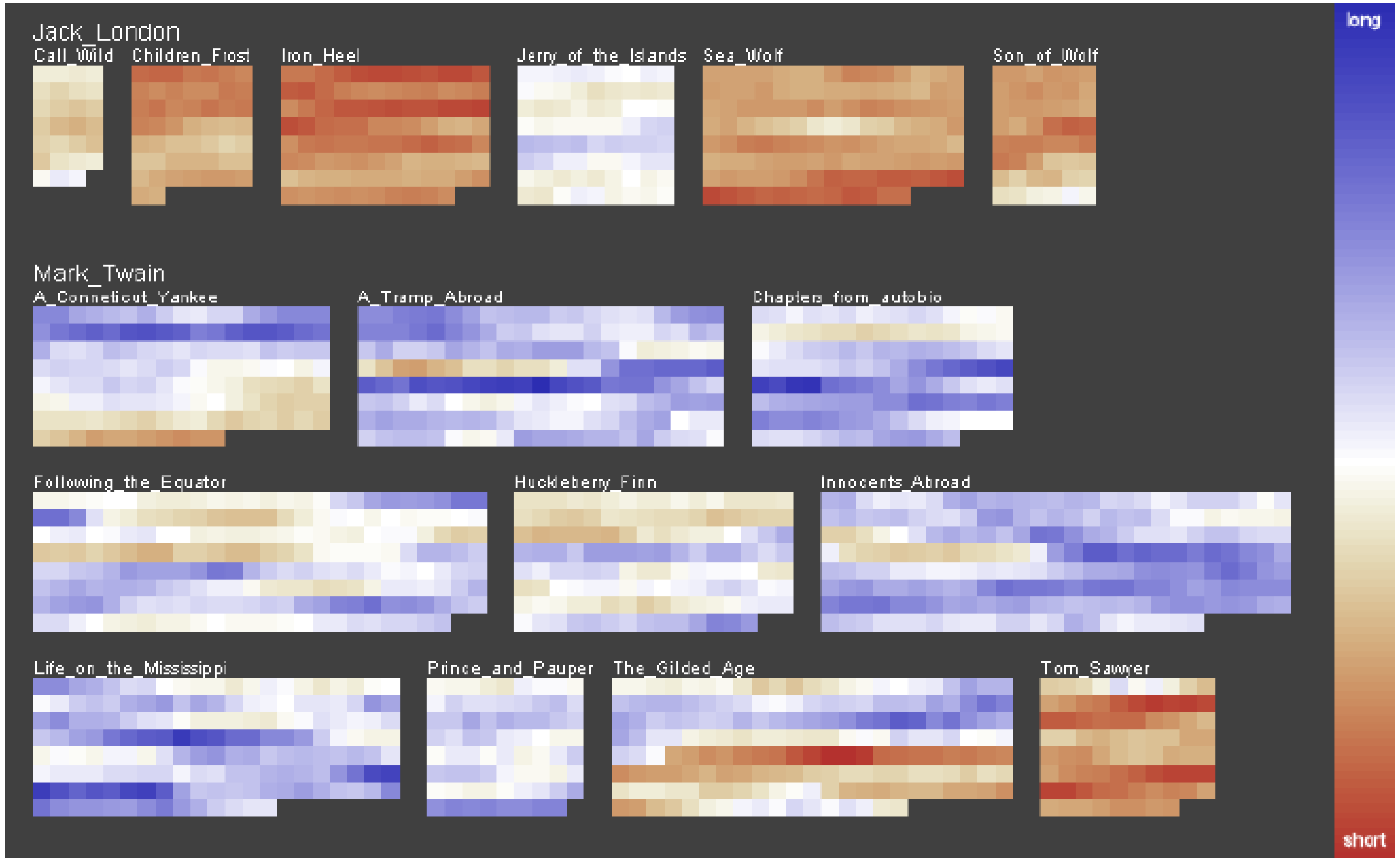
- the analyst often needs to examine multiple types of data and correct the errors among them.
    - designing an integrated interface to visually illustrate the distributions of different types of data

Research Challenges

Textual data

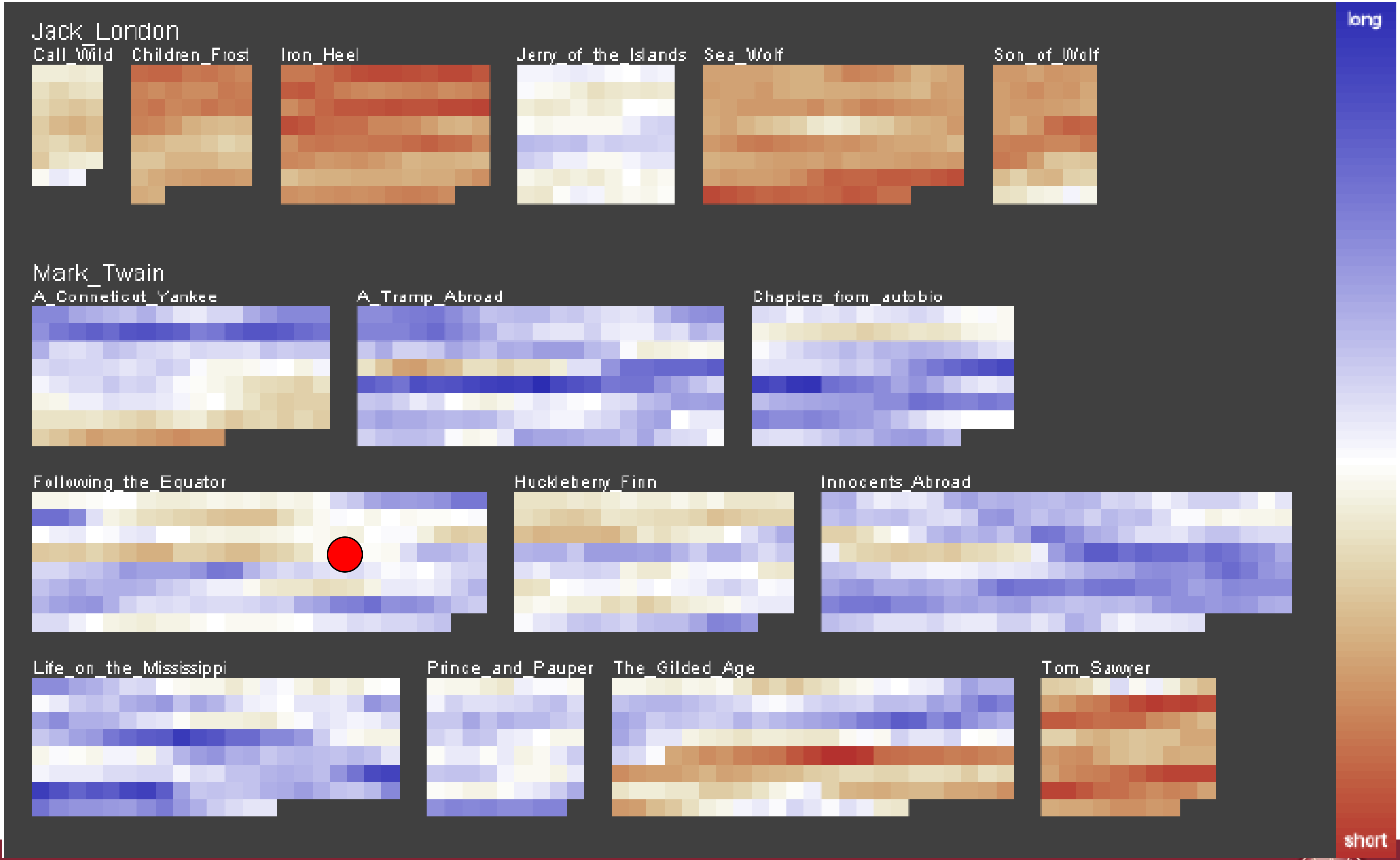
- Although textual data is widely used in many lines of work, data quality problems for such type of unstructured data remain largely unexplored.
  - This is because, due to the unstructured nature of textual documents, quality management for textual data is challenging.
- textual data often contains several data fields and mixes the useful information with irrelevant information.
- Therefore, it is important to remove the irrelevant information, which is still a hot research topic in the area of information retrieval.
- text corpora may contain text strings of different distributions, such as different lengths and language usages.
- Another challenge is how to effectively improve the quality of a text corpora with inconsistent data distributions

J.London vs M.Twain average sentence lengths

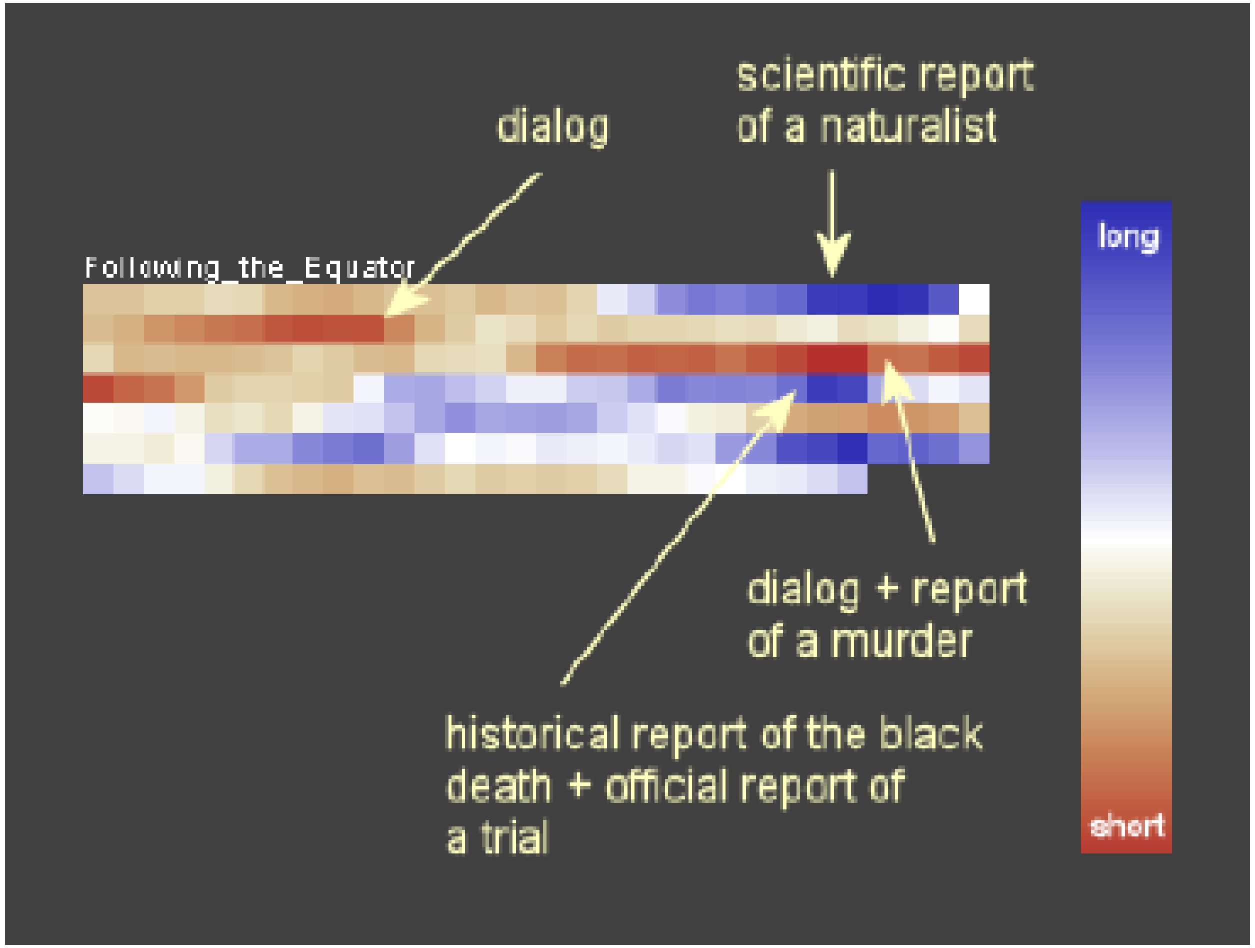


Keim, Daniel A., and Daniela Oelke. "Literature fingerprinting: A new method for visual literary analysis." *2007 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2007.

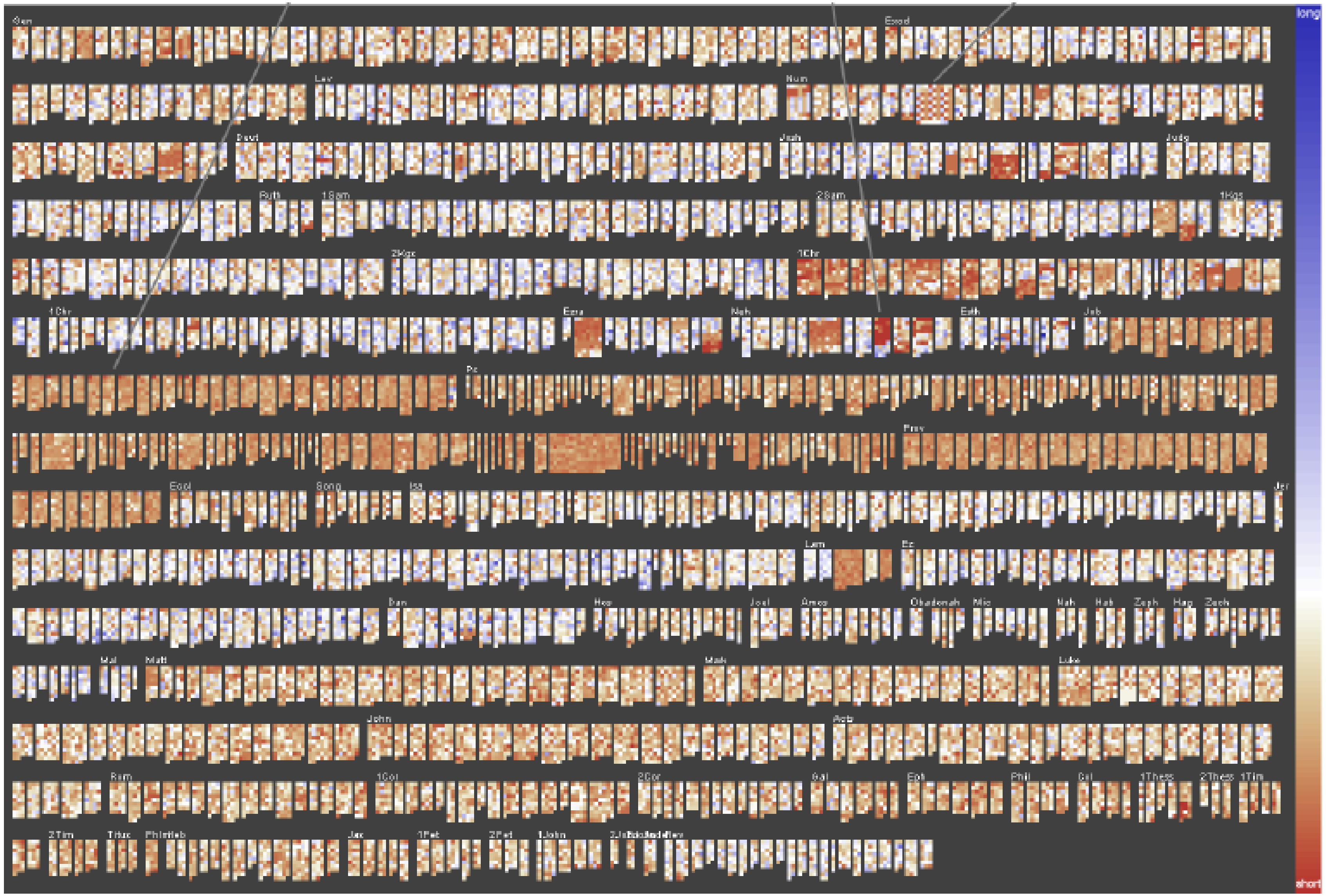
User interaction (a non uniform book?)



Details of a book



What about the Bible?



Research Challenges

Limitations of perception/cognition.

- limitations on visual perception and cognition (restricted field of view) limited working memory in cognition

- explore a mixed initiative mechanism which seamlessly integrates system initiative guidance and user initiative guidance for better human machine intelligence,

Difficulty in understanding uncertainty and its implications.

- Uncertainty might arise in any stage of a data cleaning process, and propagate in subsequent stages

- understanding the uncertainty and its implications would be generally difficult without a proper visual guidance.

Thank you for your attention !

Marco Angelini
angelini@diag.uniroma1.it

A.WA.RE

Advanced **V**isualization & **V**isual **A**nalytics **RE**search group at Sapienza

ANY QUESTIONS?