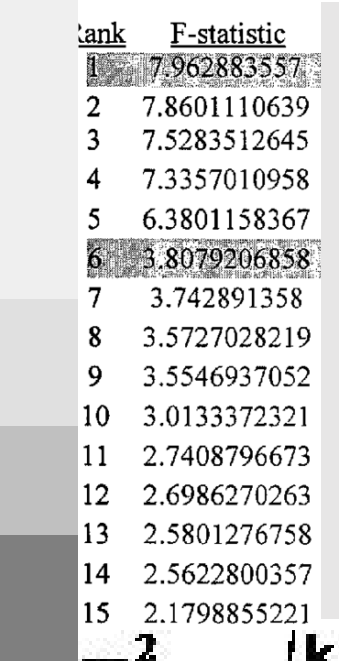
Data Visualization



Prof. Dr. Javier Valdes
Javier.valdes@th-deg.de

IAI - Institut für Angewandte Informatik
Institute for Applied Informatics

Technische Hochschule Deggendorf
Technologie Campus Freyung
Grafenauer Str. 22, D- 94078 Freyung,
Germany

Tel.: +49 8551 91764 40
Fax: +49 8551 91764 69

Part 1: Introduction to data Visualization

Our objectives

1. Learn the main reasons and applications
2. Identify data visualization as a tool
3. Learn the basic concepts of visualization
4. Introduction to data visualization in R
5. Generation of scientific reports

In other words:

Learn to choose which graphics for which situations
How to generate high quality graphs

Materials

Information

Knowledge

Practice

An extensive comparative study of cluster validity indices

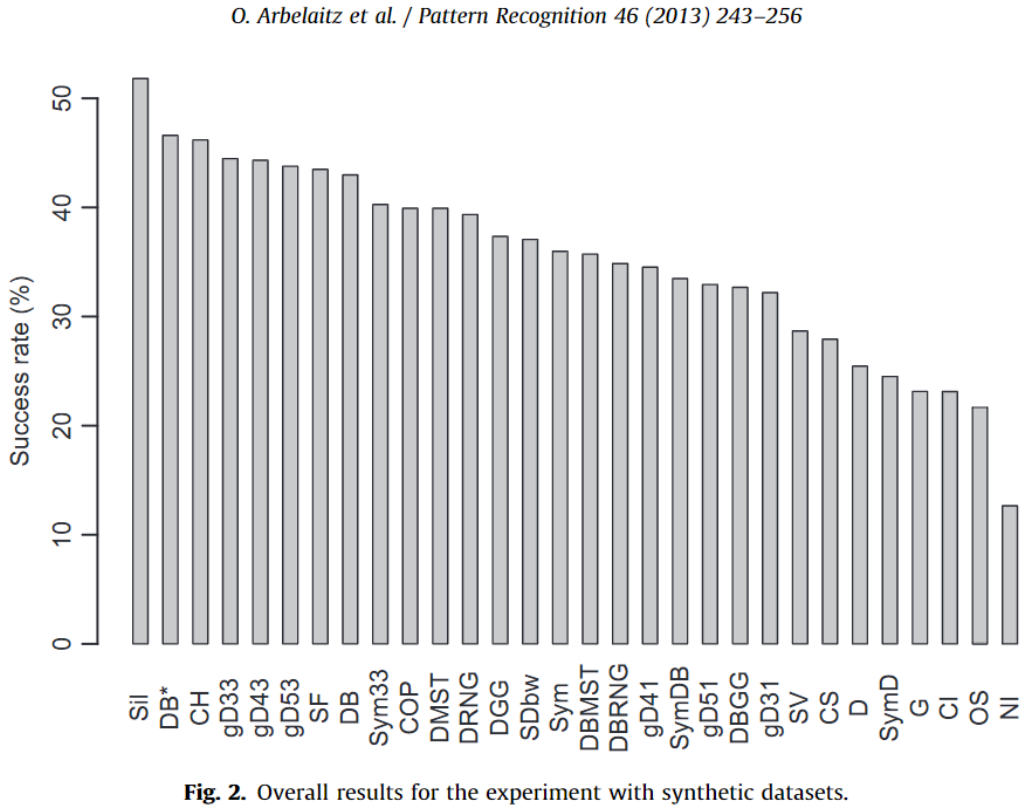


Fig. 2. Overall results for the experiment with synthetic datasets.

Arbelaitz, Olatz, et al. "An extensive comparative study of cluster validity indices." *Pattern Recognition* 46.1 (2013): 243-256.

An extensive comparative study of cluster validity indices

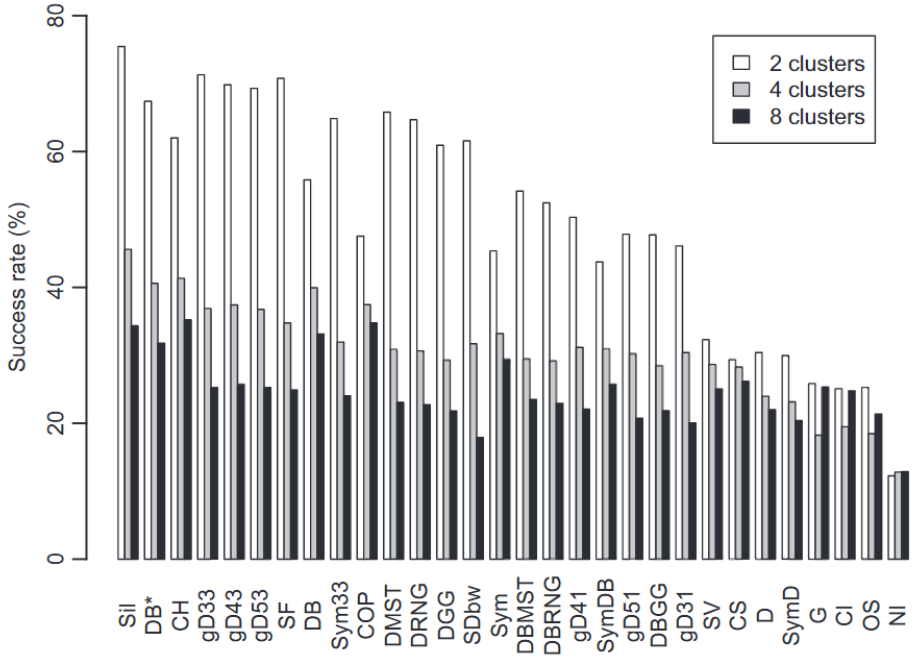


Fig. 4. Results for synthetic datasets broken down by number of clusters.

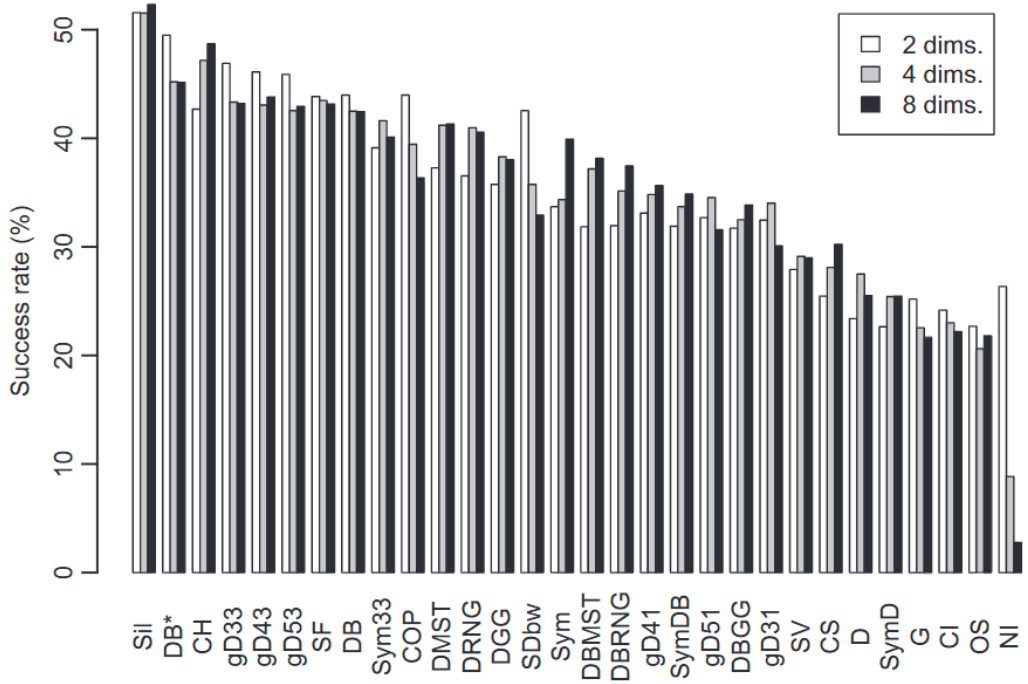
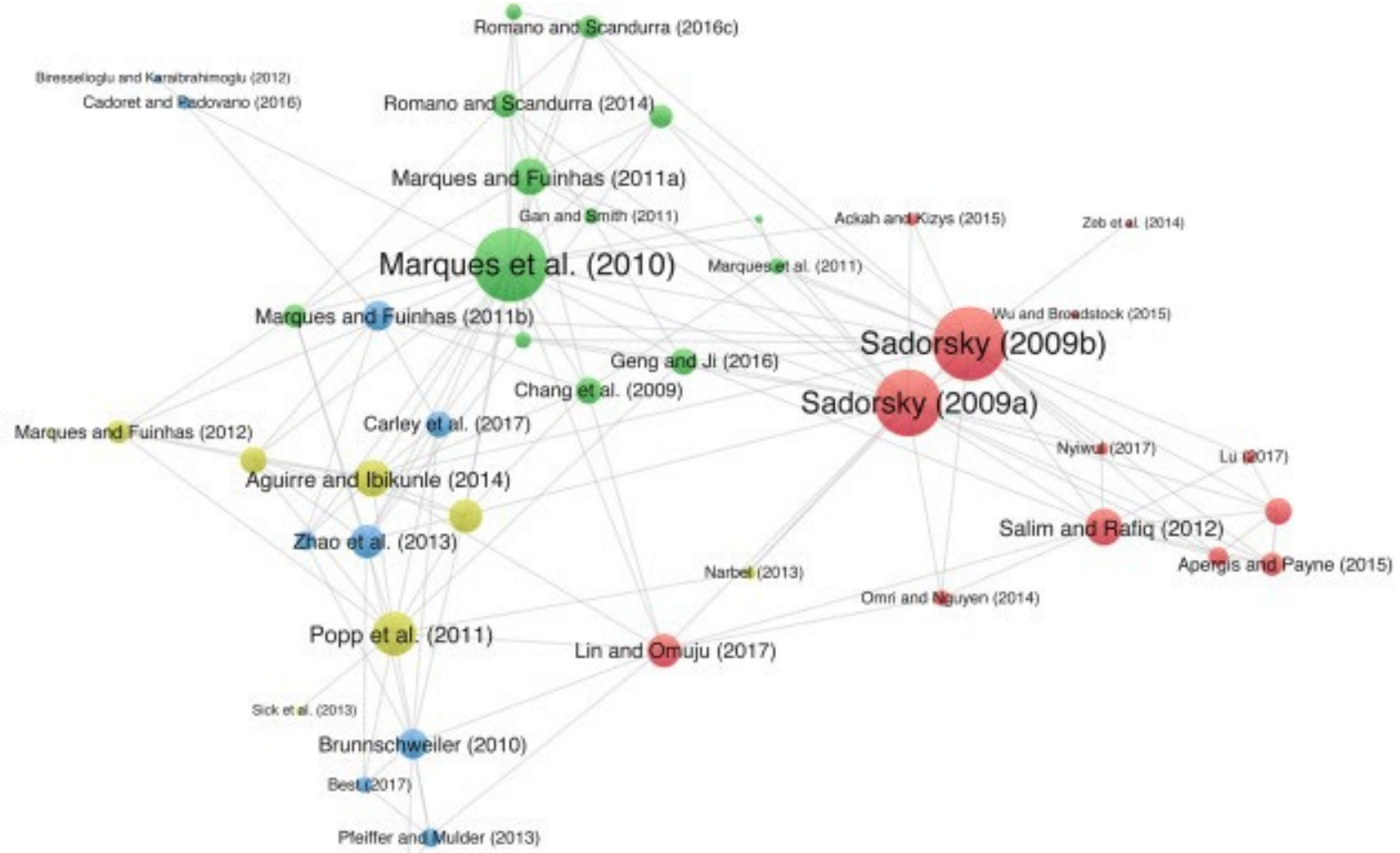


Fig. 5. Results for synthetic datasets broken down by dimensionality.

Arbelaitz, Olatz, et al. "An extensive comparative study of cluster validity indices." *Pattern Recognition* 46.1 (2013): 243-256.

Empirical determinants of renewable energy deployment: A systematic literature review



Bourcet, Clémence. "Empirical determinants of renewable energy deployment: a systematic literature review." *Energy Economics* (2019): 104563.

LEARNING R

OBTAINING R

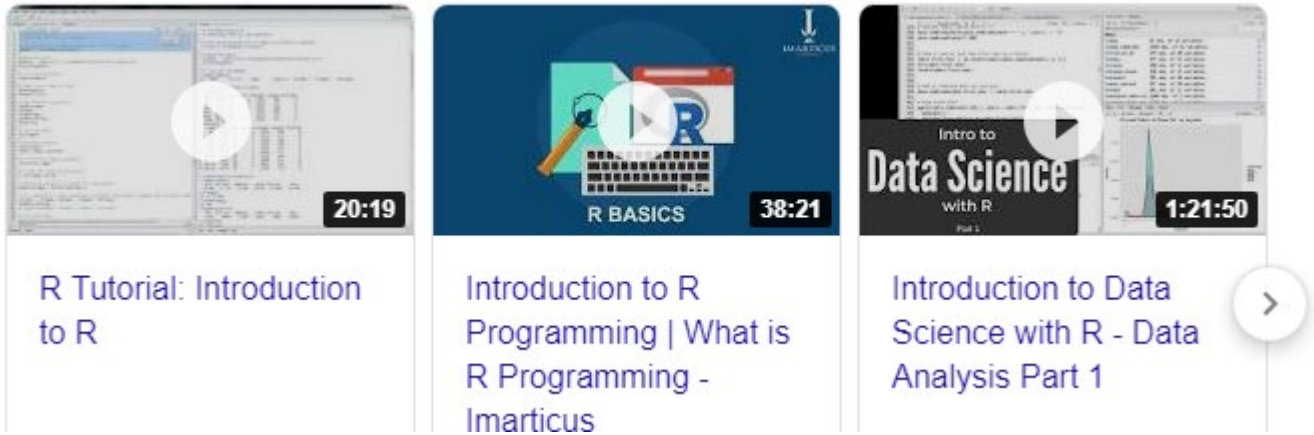
- Comprehensive R Archive Network:

  http://cran.r-project.org

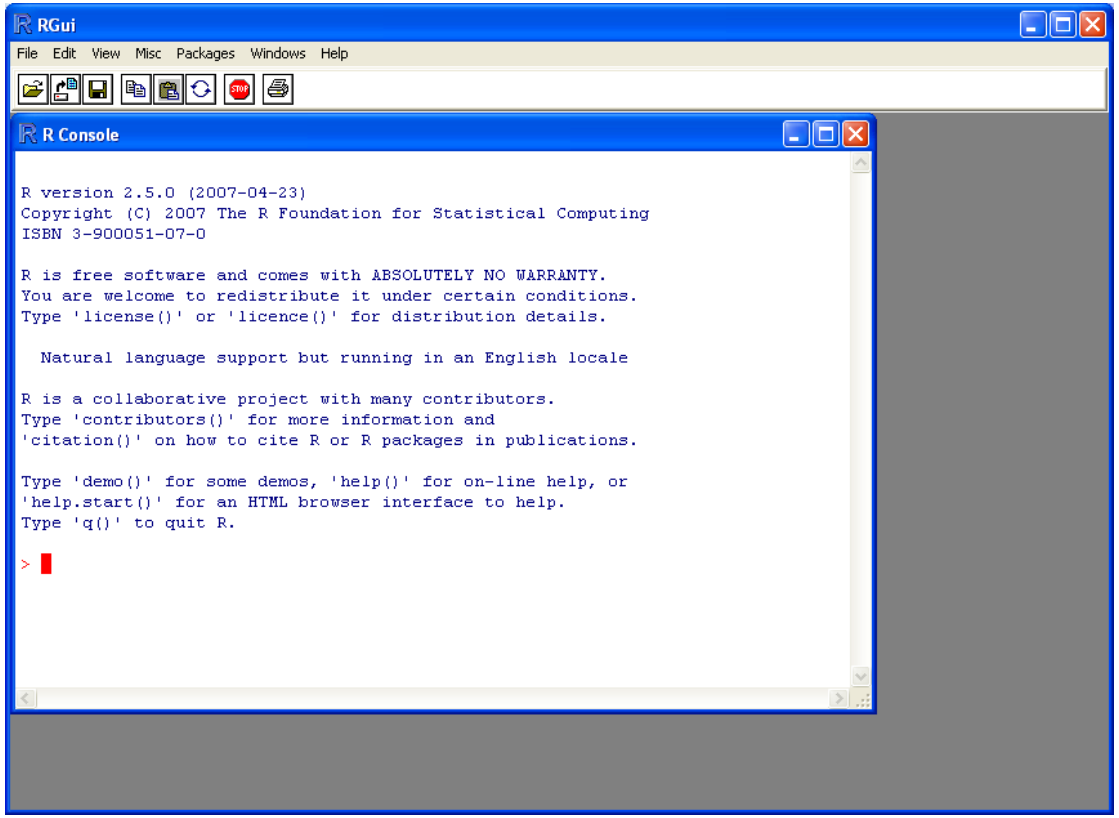
- Courses:

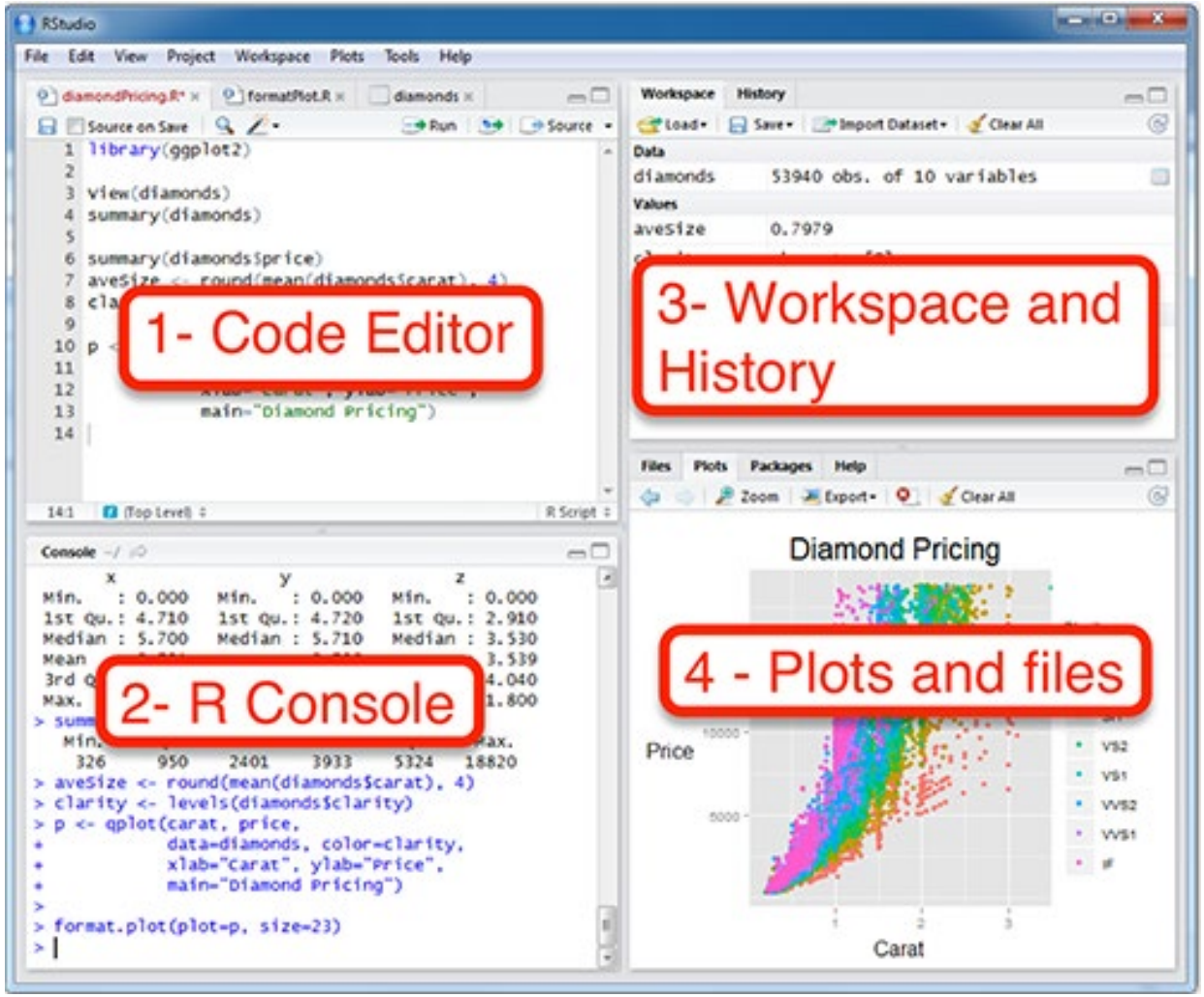
  https://www.datacamp.com/

- Videos:



R vs. Rstudio


RGui
File Edit View Misc Packages Windows Help
R Console
R version 2.5.0 (2007-04-23)
Copyright (C) 2007 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.


RStudio
File Edit View Project Workspace Plots Tools Help
1- Code Editor
3- Workspace and History
2- R Console
4 - Plots and files
Diamond Pricing
Price
Carat

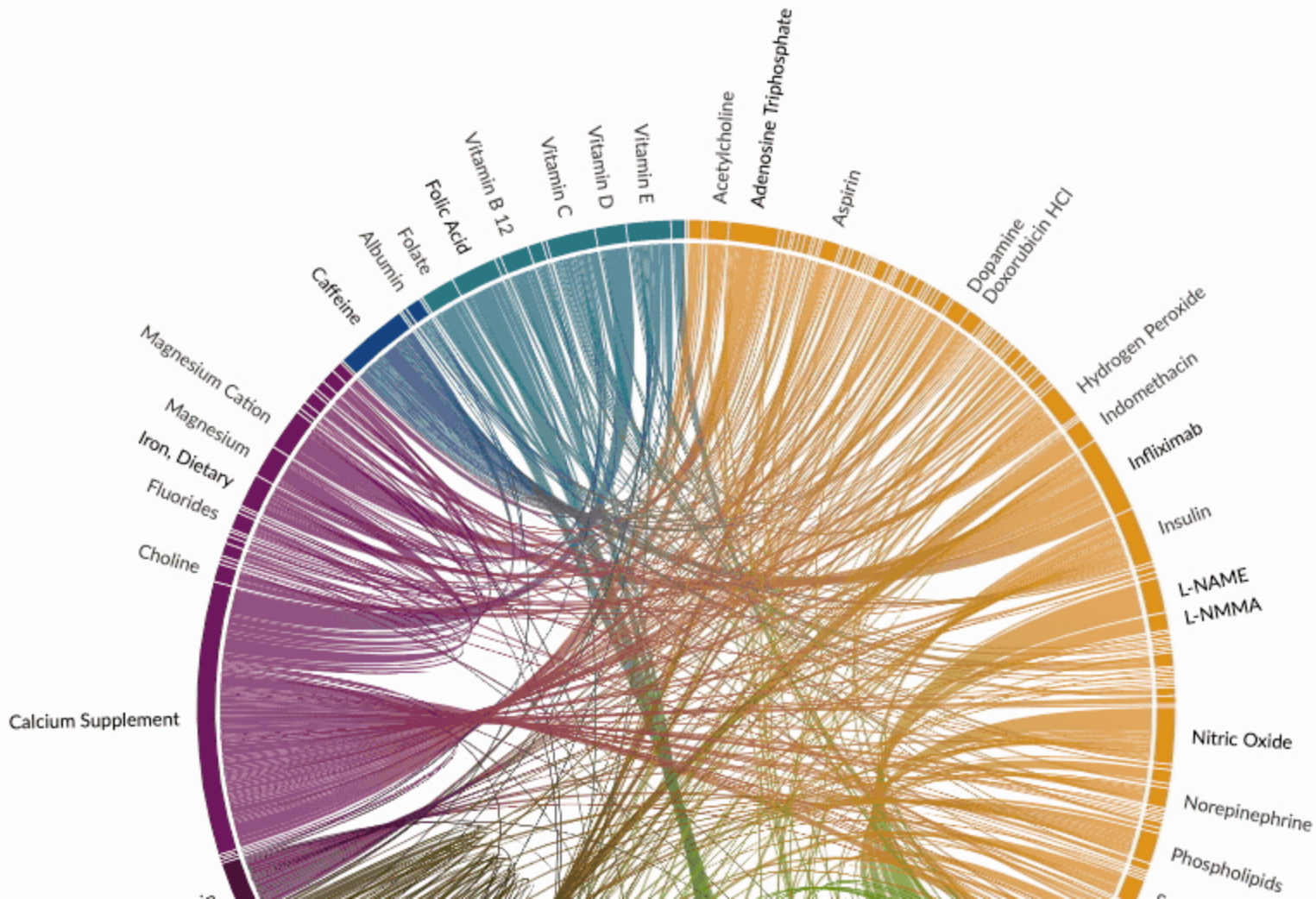
KEEP CALM AND LOVE STATISTICS

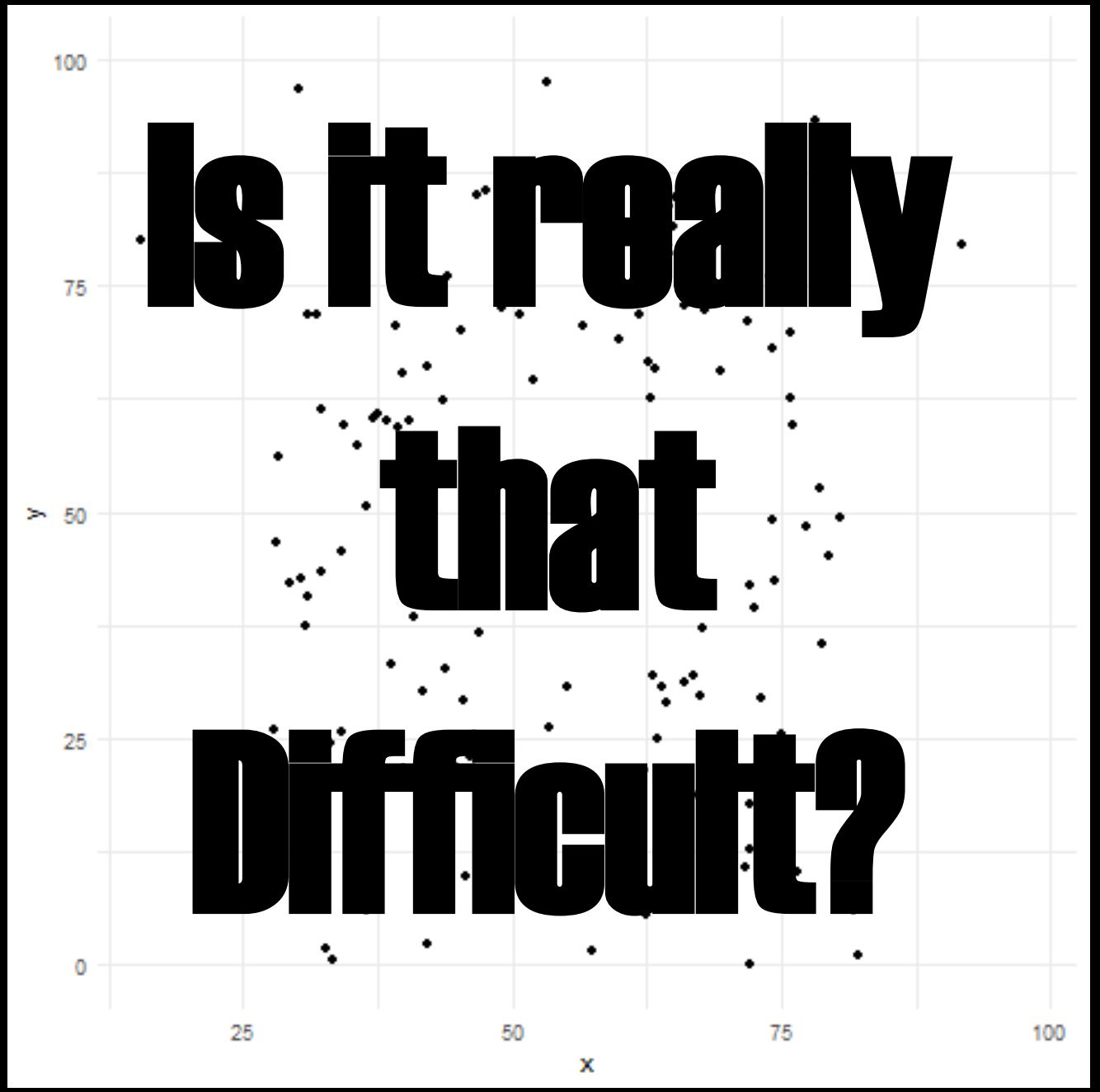

EAT
SLEEP
NETFLIX
REPEAT

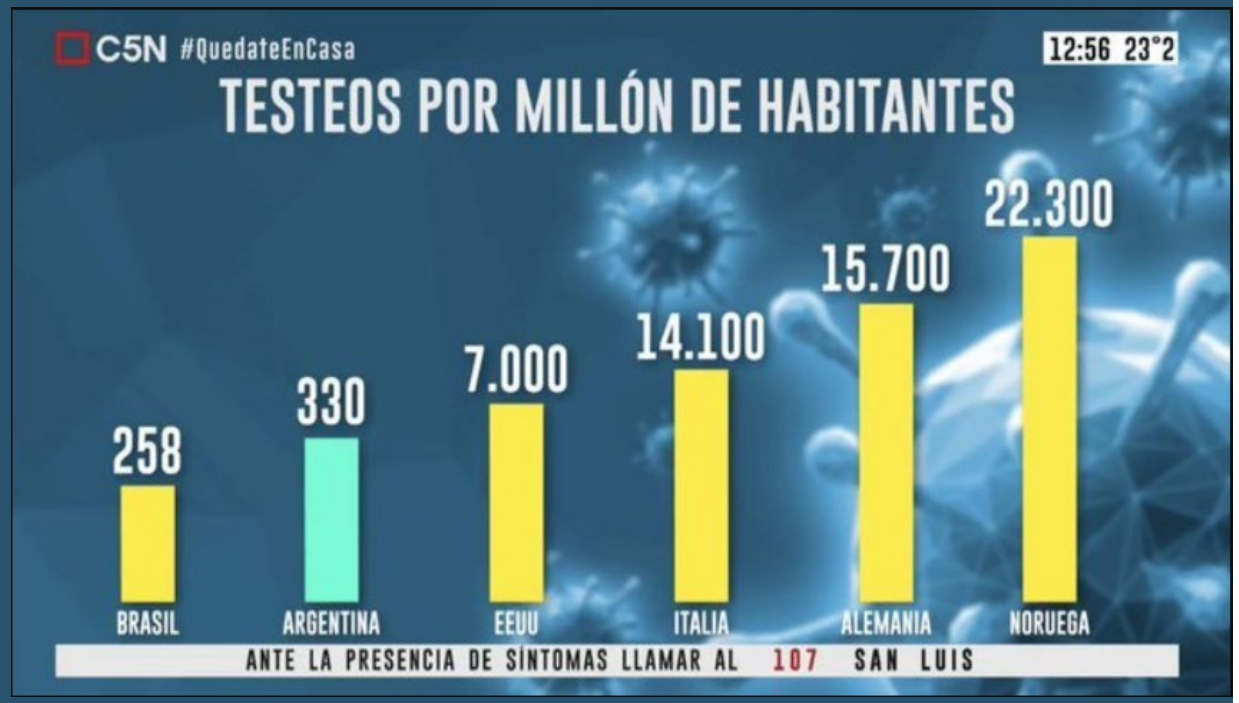
VISUALIZING SUPPLEMENT-DRUG INTERACTIONS

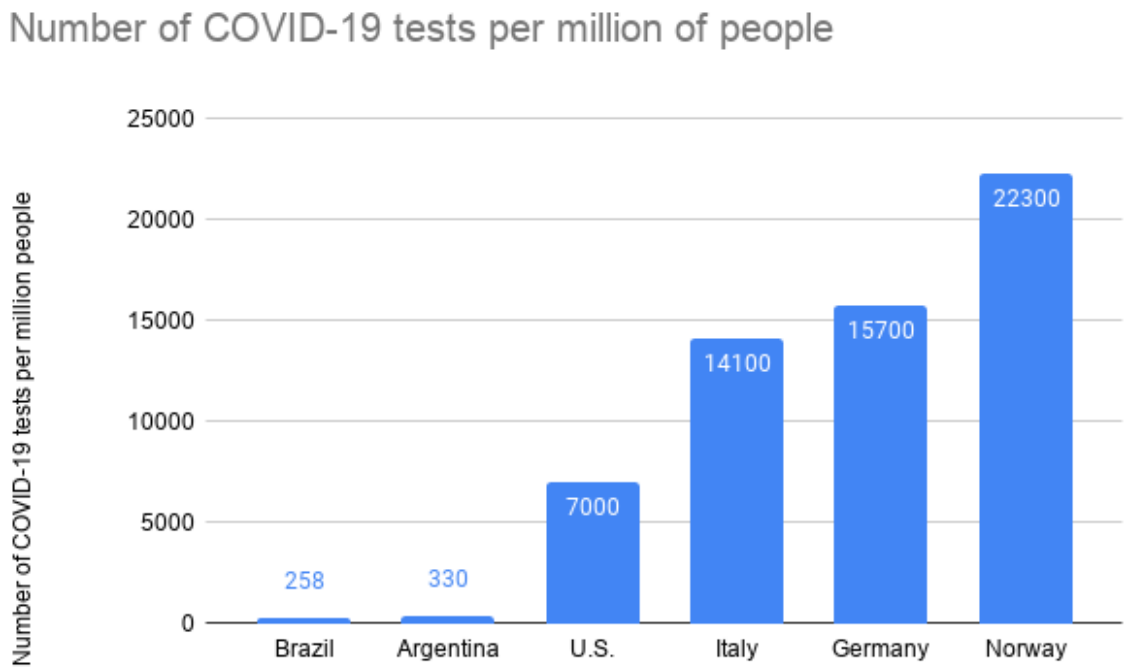
reported in the scientific literature

This chord diagram shows interactions reported in the scientific literature between a list of supplements and drugs curated by the Allen Institute for AI. The algorithm automatically extracts evidence from publications and allows users to explore the resulting database via text-based searchers here. I created this chord diagram to facilitate visual exploration of the various interactions. Supplements are grouped quasi-functionally (e.g., minerals, vitamins, fatty acids) and shown as distinct colors. Clicking on the name or outer band of any supplement or drug will link to the relevant Supp.AI webpage with additional information about the agent and it's interactions.




Is it really that Difficult?
100
75
50
25
0
y
25
50
75
100
x

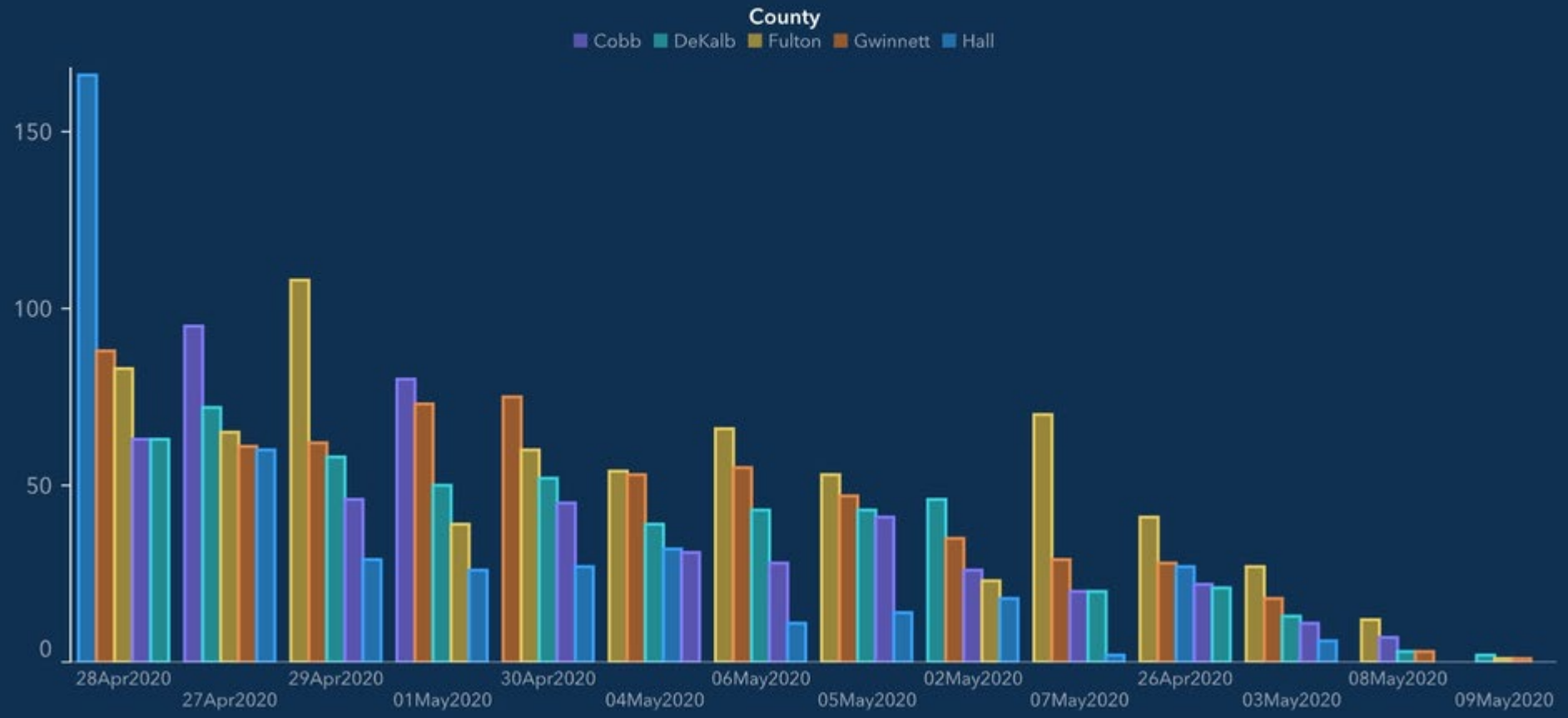




https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9

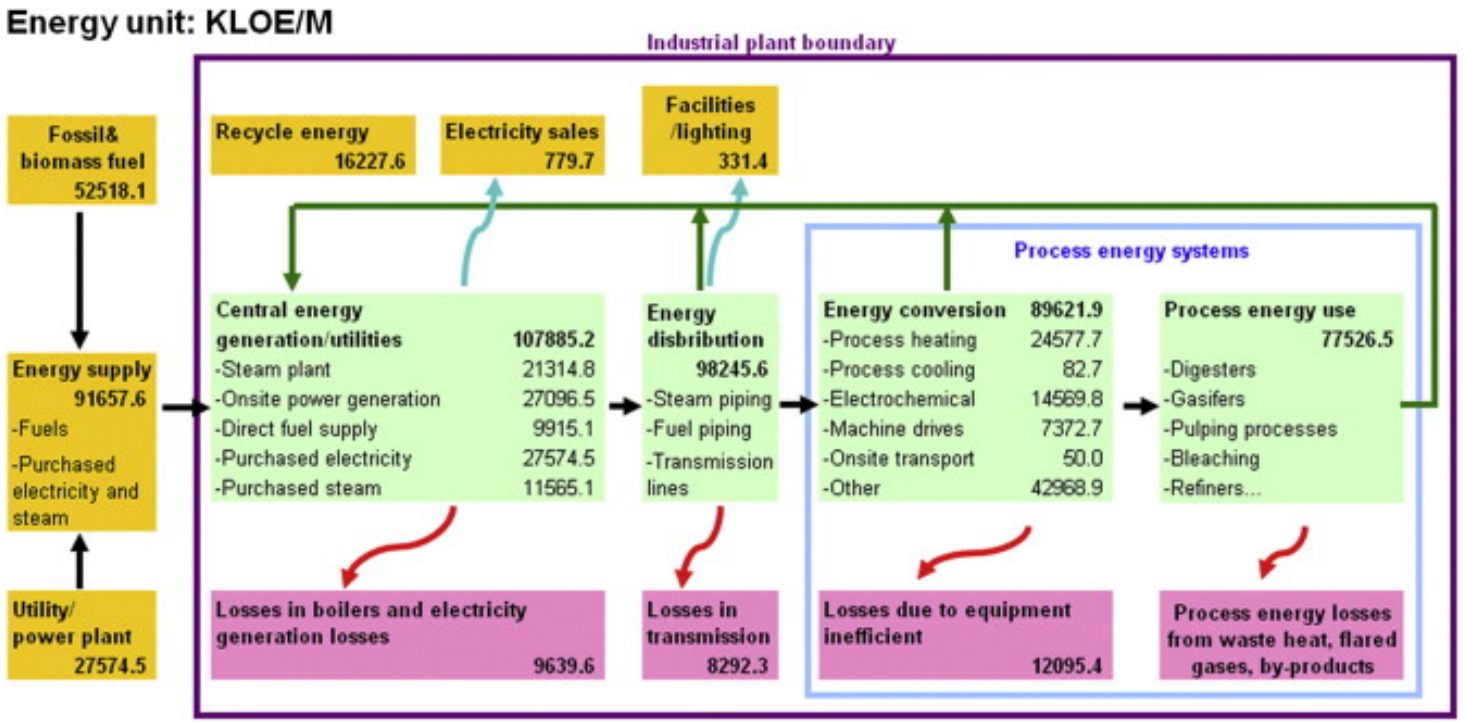
Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.



https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9

Energy flow analysis in pulp and paper industry



Hong, Gui-Bing, et al. "Energy flow analysis in pulp and paper industry." *Energy* 36.5 (2011): 3063-3068.

Energy flow analysis in pulp and paper industry

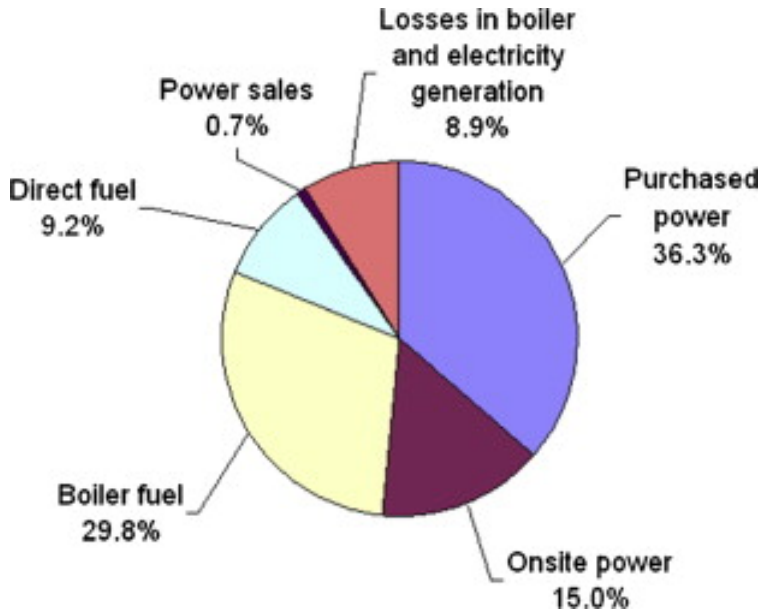


Fig. 3. Primary energy use distribution of Taiwanese pulp and paper industry

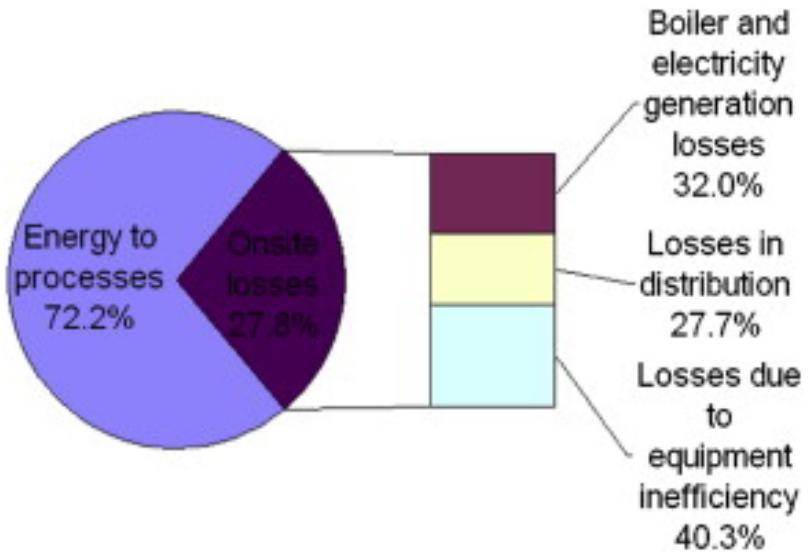
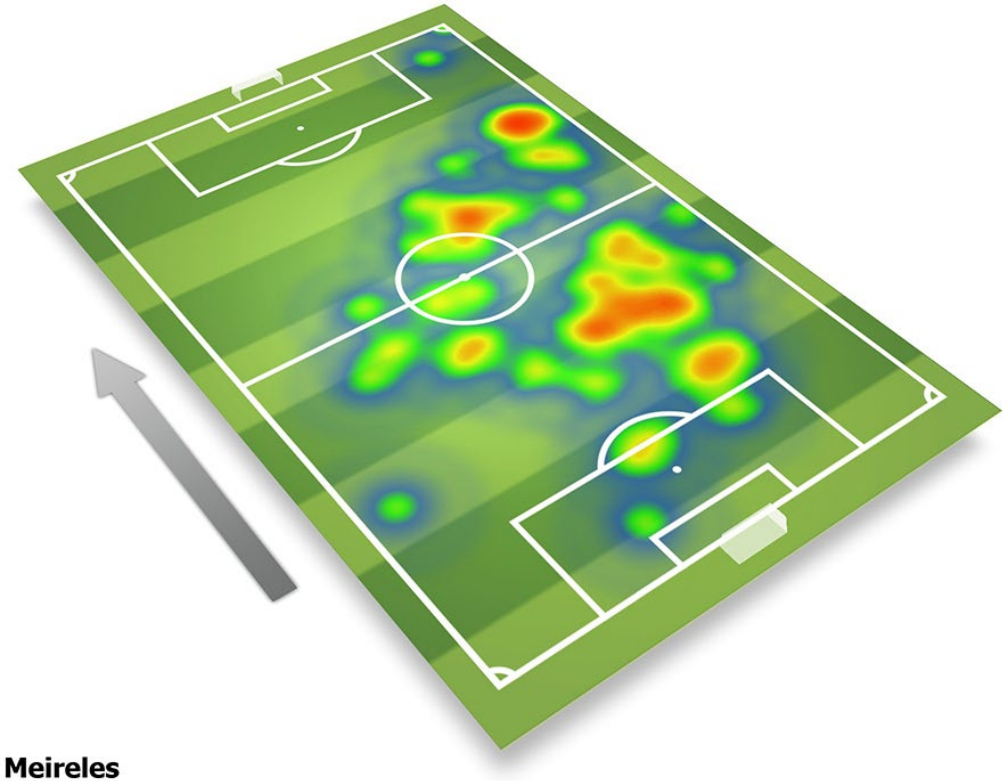
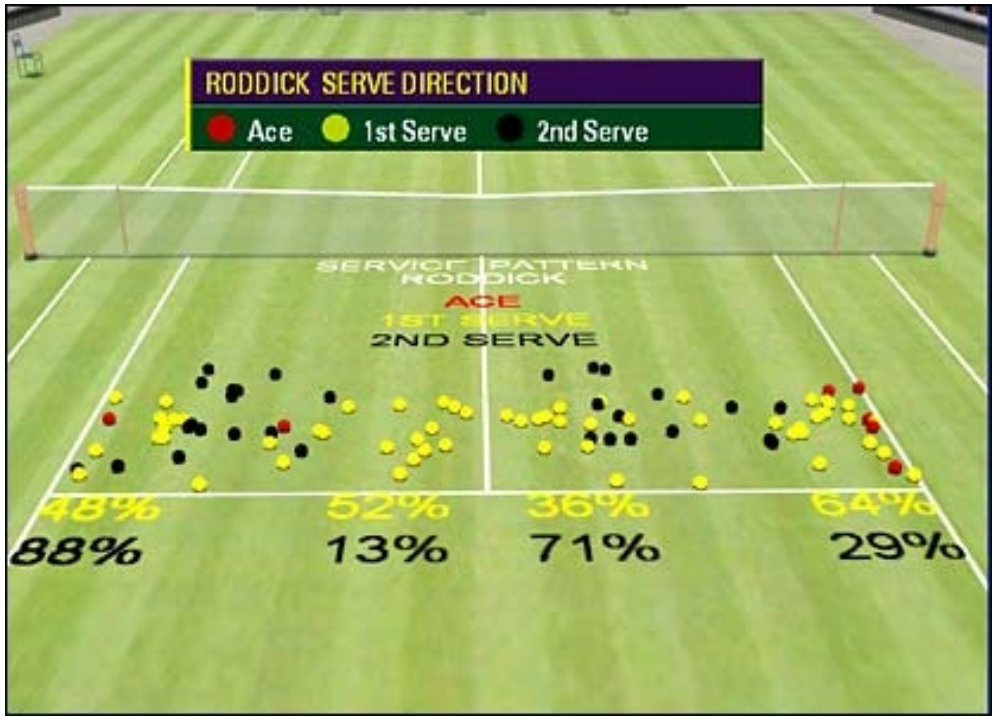


Fig. 4. Onsite energy loss profile of Taiwanese pulp and paper industry.

Hong, Gui-Bing, et al. "Energy flow analysis in pulp and paper industry." *Energy* 36.5 (2011): 3063-3068.



http://www.optasports.de/media/299686/speigel-meireles-heatmap.jpg



http://newsimg.bbc.co.uk/media/images/41263000/jpg/_41263791_roddickserve_stats416.jpg

Ways Writers Use Misleading Graphs

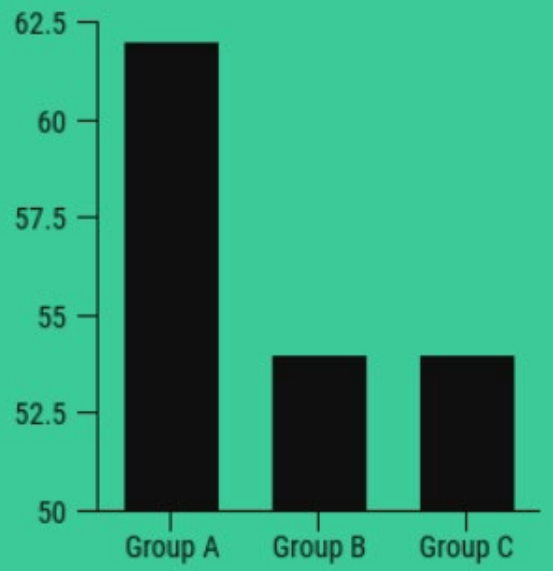
What are the main errors that we can make when generating a graph?

Discuss as a group which errors are most typical and list the results on the Etherpad by degree of severity.

Time: 5 min

1 OMITTING THE BASELINE

In most cases, the baseline for a graph is 0. But writers can skew how data is perceived by making the baseline a different number. This is known as a "truncated graph".



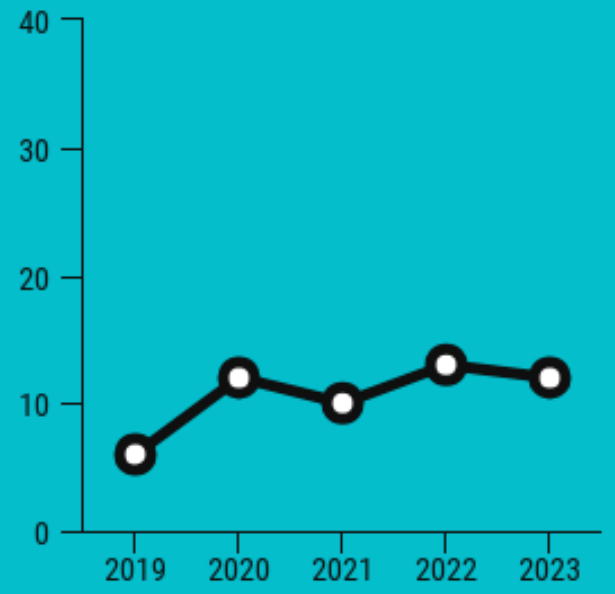
MISLEADING

- Starting the vertical axis at 50 makes a small difference between groups seem massive
- Group A looks much larger than Groups B and C

https://venngage.com/blog/misleading-graphs/

2 MANIPULATING THE Y-AXIS

Expanding or compressing the scale on a graph can make changes in data seem more or less significant than they actually are.



MISLEADING

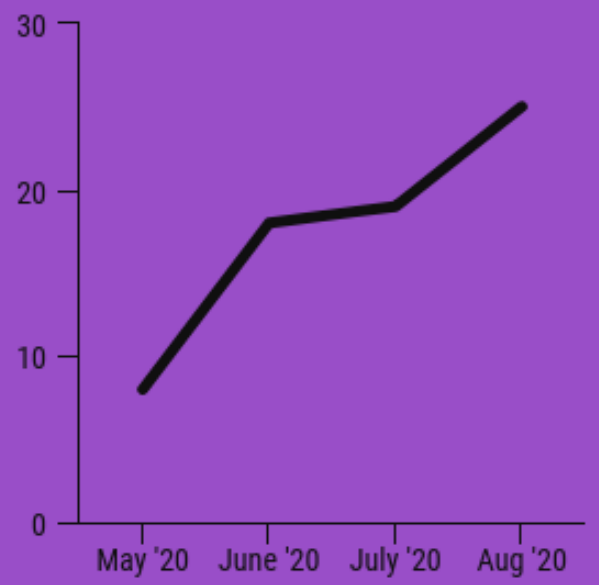
- The scale is disproportionate to the data, making the change over time seem small

https://venngage.com/blog/misleading-graphs/

3

CHERRY PICKING DATA

Writers may only include certain data points on their graphs to reinforce their narratives. This can create a false impression of the data.



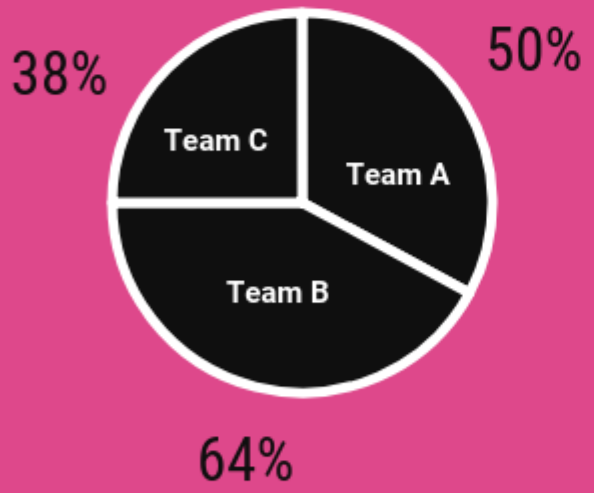
MISLEADING

- Only a few months out of the year are graphed, depicting an upward trends

https://venngage.com/blog/misleading-graphs/

4 USING THE WRONG GRAPH

The type of graph you use should depend on the type of data you want to visualize. Using the wrong type of graph can skew the data. Writers will sometimes use the wrong type of graph on purpose.



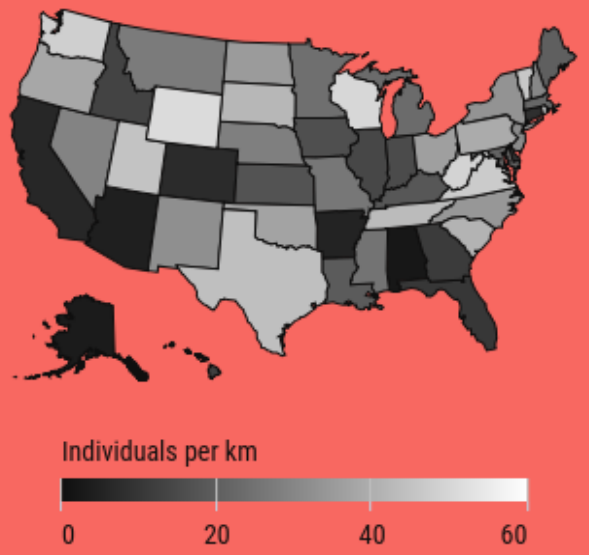
MISLEADING

- Pie charts are used to compare parts of a whole, not the difference between groups
- A different type of graph should be used to compare the three teams

https://venngage.com/blog/misleading-graphs/

5 GOING AGAINST CONVENTIONS

Over time, we have developed standards for how data is visualized. Flipping those conventions can make a graph confusing or misleading to readers.

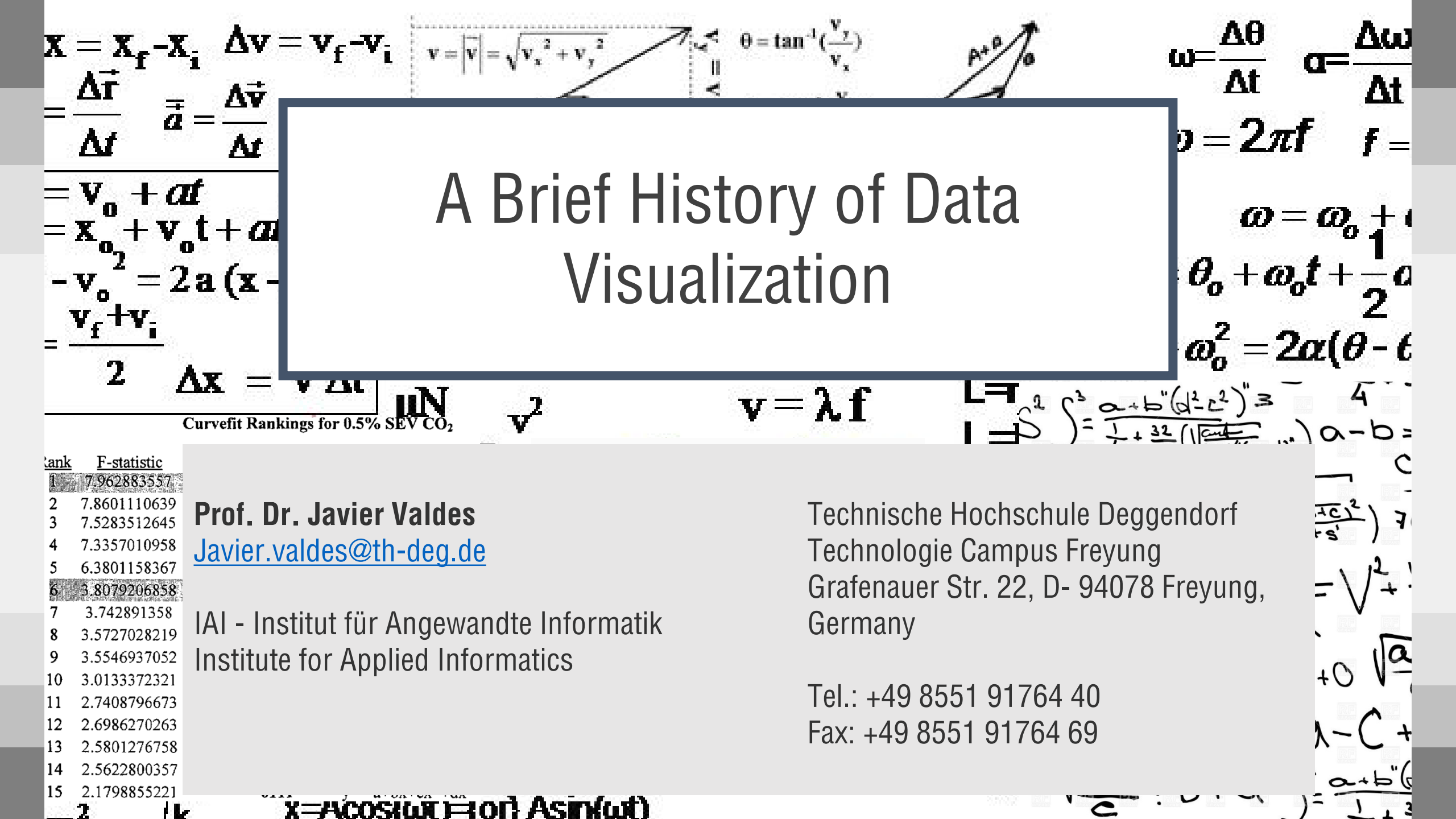


MISLEADING

- Normally, darker shades are associated with density on a map but here, dark has been used to depict lower population density
- This graph can confuse and mislead readers, who expect dark to represent a higher population density

https://venngage.com/blog/misleading-graphs/

A Brief History of Data Visualization

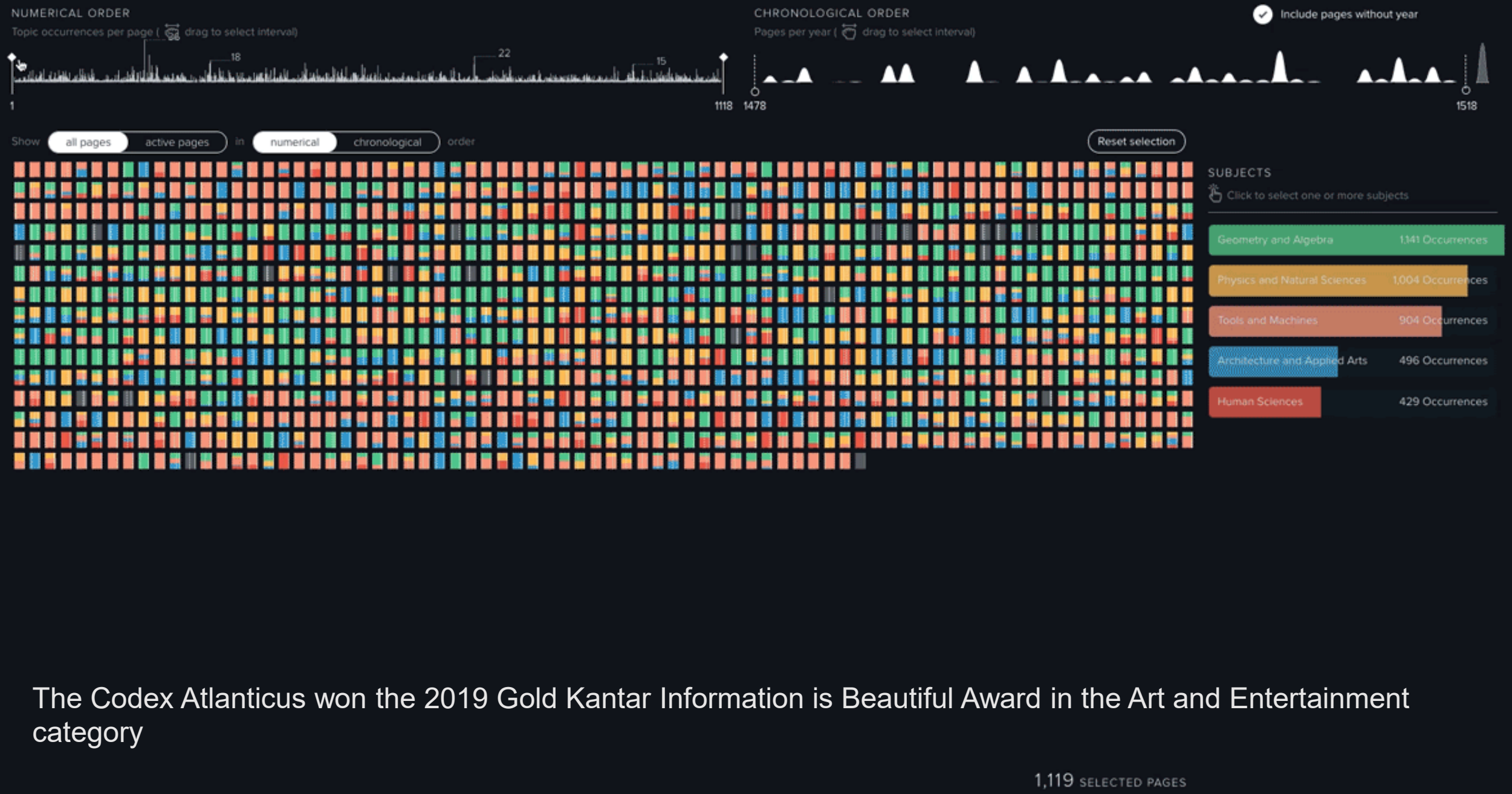


Prof. Dr. Javier Valdes
Javier.valdes@th-deg.de

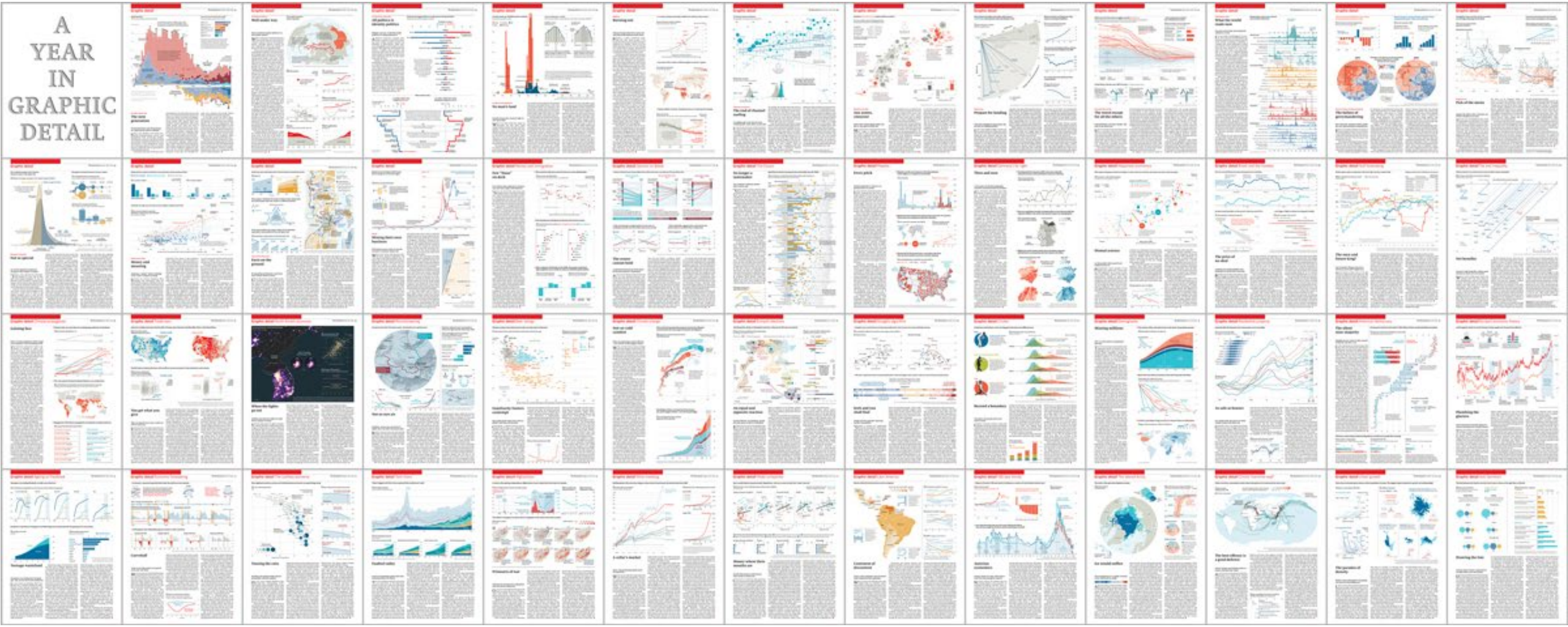
IAI - Institut für Angewandte Informatik
Institute for Applied Informatics

Technische Hochschule Deggendorf
Technologie Campus Freyung
Grafenauer Str. 22, D- 94078 Freyung,
Germany

Tel.: +49 8551 91764 40
Fax: +49 8551 91764 69

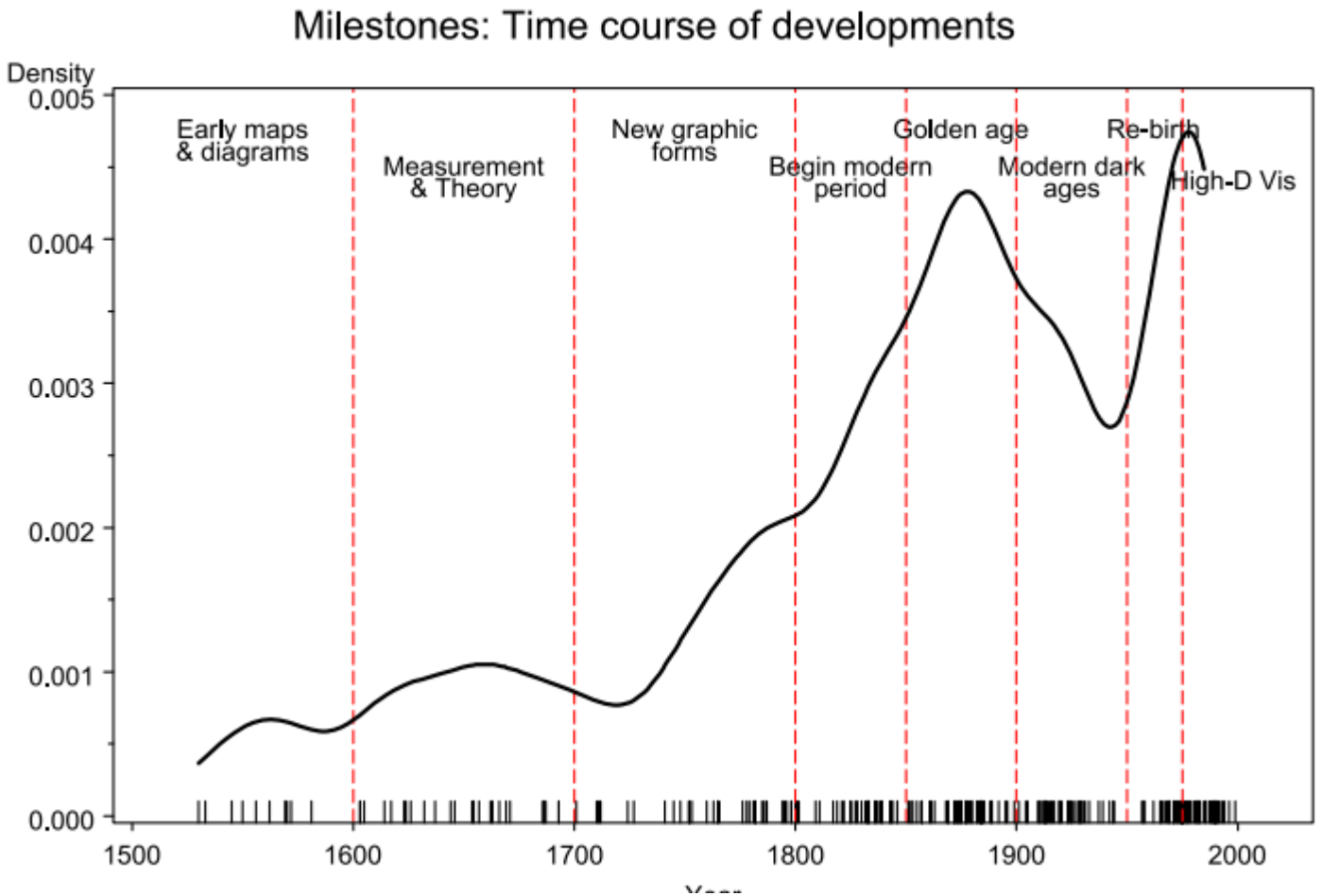


The Codex Atlanticus won the 2019 Gold Kantar Information is Beautiful Award in the Art and Entertainment category



The Economist print version started a new section for their Graphic Detail series in 2018

A Brief History of Data Visualization



Michael Friendly, 2008. A Brief History of Data Visualization, in: Handbook of Data Visualization, Springer

Pre-17th Century: Early Maps and Diagrams

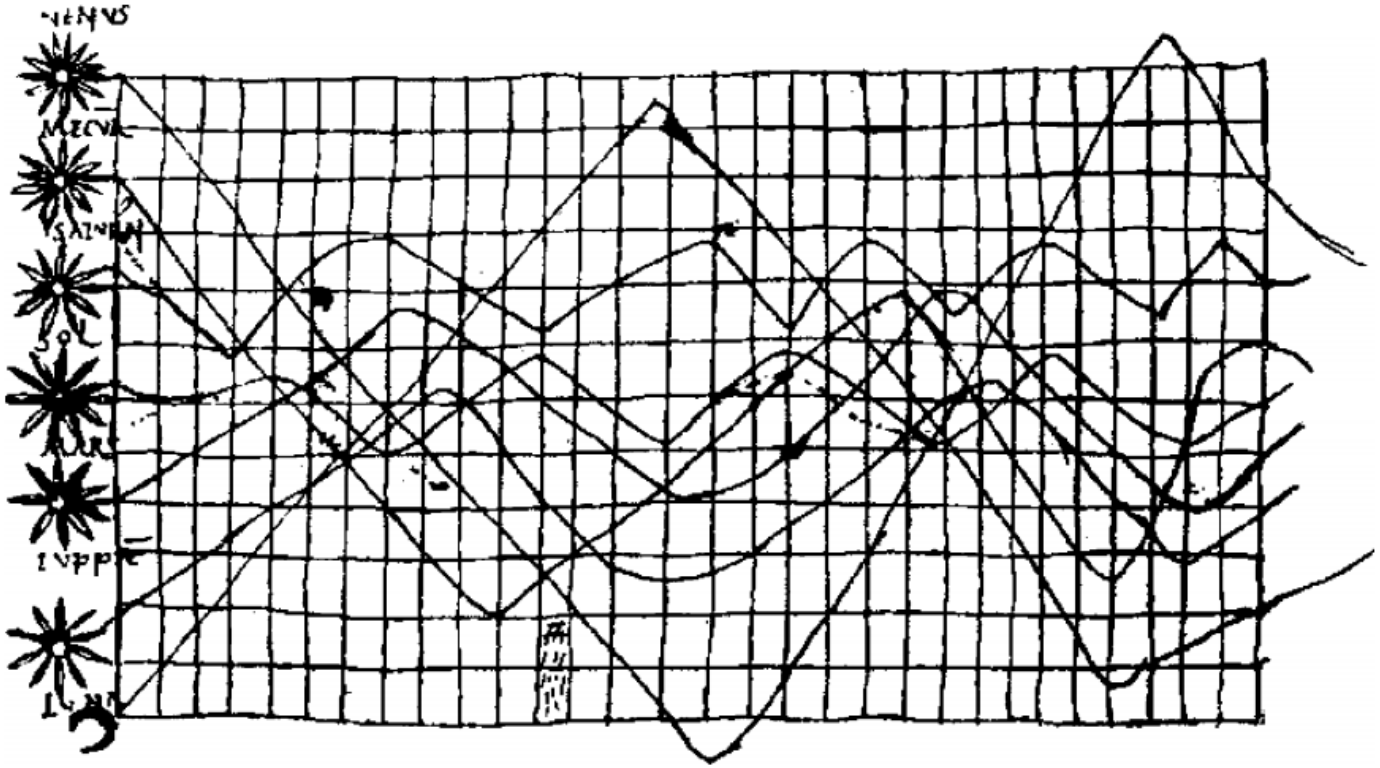


Figure 1.2. Planetary movements shown as cyclic inclinations over time, by an unknown astronomer, appearing in a 10th-century appendix to commentaries by A.T. Macrobius on Cicero's *In Somnium Sciponis*. *Source*: Funkhouser (1936, p. 261)

- Positions of stars: Diagrams, geometric diagrams
- Notion of a coordinate system
- Graph paper
- Proto-bar graph
- Ttrigonometric tables
- Modern cartographic atlas

Michael Friendly, 2008. A Brief History of Data Visualization, in: Handbook of Data Visualization, Springer

Exercice

Divide into seven groups and read together the section on the evolution of data visualization.

List the most important points as a table on the Etherpad dedicated to your section.

Time: 25 min

Michael Friendly, 2008. A Brief History of Data Visualization, in: Handbook of Data Visualization, Springer

Exercice

Generate a glossary of all terms referring to data visualization techniques. To do this

1. Start by selecting a list of terms that appear in your section.
2. look for those terms that are repeated in several sections.
3. Find definitions that satisfy the views of all the groups that identified that term.

Time: 15 min

The objective is not to copy and paste definitions from the Internet, but to write the text in their own words. Be creative!!!

Data Visualization & Statistics

Curvefit Rankings for 0.5% SEV CO_2

Rank	F-statistic
1	7.962883557
2	7.8601110639
3	7.5283512645
4	7.3357010958
5	6.3801158367
6	3.8079206858
7	3.742891358
8	3.5727028219
9	3.5546937052
10	3.0133372321
11	2.7408796673
12	2.6986270263
13	2.5801276758
14	2.5622800357
15	2.1798855221

Prof. Dr. Javier Valdes
Javier.valdes@th-deg.de

IAI - Institut für Angewandte Informatik
Institute for Applied Informatics

Technische Hochschule Deggendorf
Technologie Campus Freyung
Grafenauer Str. 22, D- 94078 Freyung,
Germany

Tel.: +49 8551 91764 40
Fax: +49 8551 91764 69

WHY STATISTICS?

1. Statistics is defined as a branch of mathematics or science that deals with the collection, analysis and interpretation of numerical information.
2. Statistics changes numbers into information…
3. Statistics is the art and science of deciding:
    1. What is the appropriate data to collect,
    2. Deciding how to collect them efficiently
    3. and then using them to give information (answer questions and make decisions)
4. Statistics is making decisions when there is uncertainty

Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics

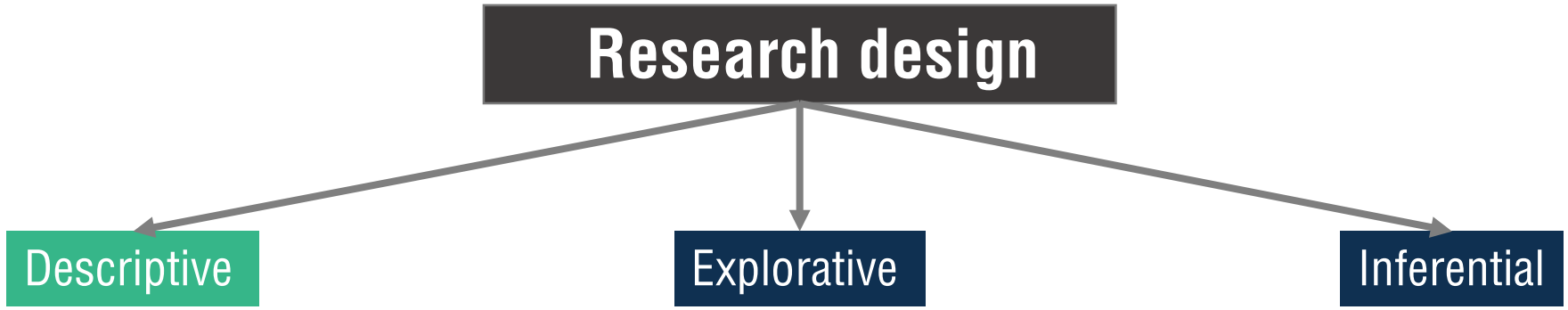
An Introduction to Statistic Analysis and Data visualization

Uses of dataviz in Stats

- I do not understand the data!
- How do I choose my dependent and explanatory variables?
- What statistical tools should I use?
- How can I compare different methods/models?
- How do I show my results?

WHY STATISTICS?

Procedure for the determination of the study design



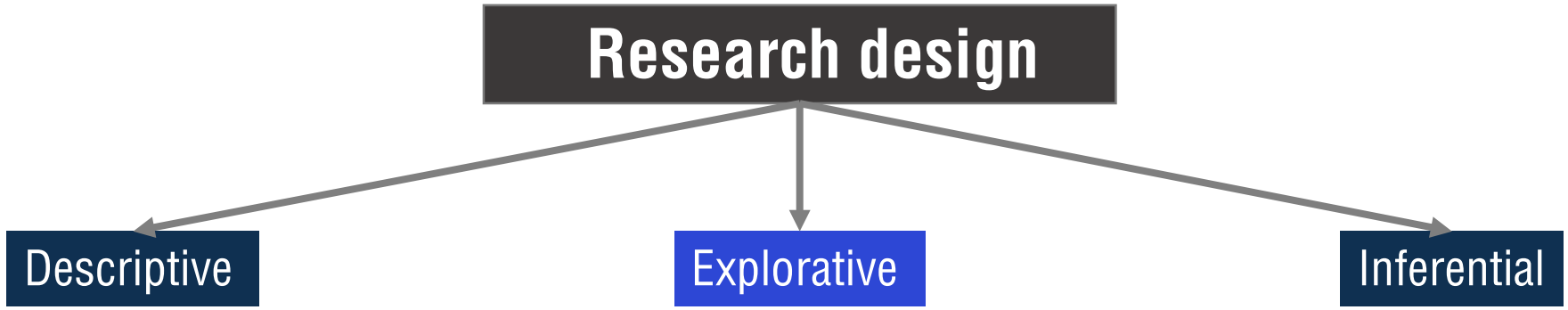
Descriptive

"In a descriptive study, the facts relevant to the topic of study are recorded and described as accurately as possible, but no relationships between variables are examined."

Homburg & Krohmer 2009, S. 250

WHY STATISTICS?

Procedure for the determination of the study design



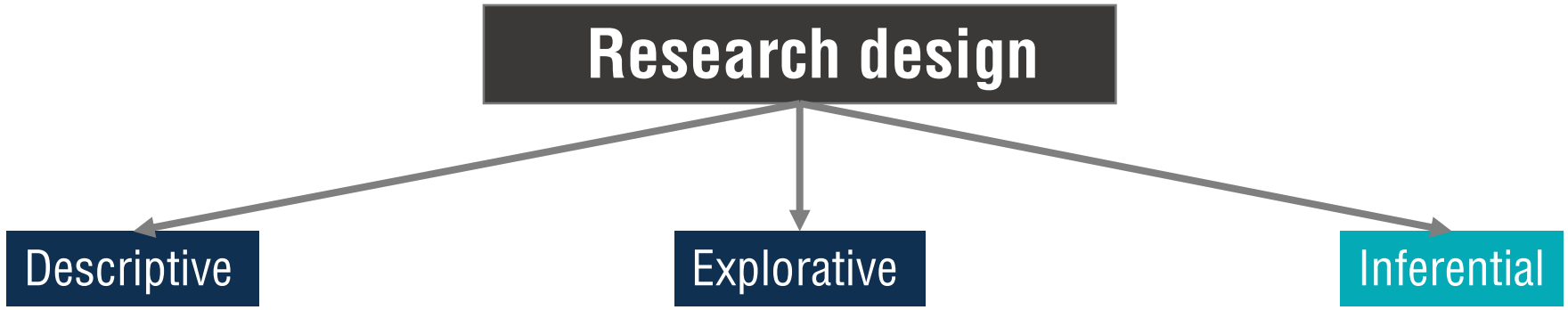
Explorative

"An exploratory investigation serves to first understand and structure the (usually still relatively unexplored) topic of investigation in detail. Relationships among the variables under consideration may be explored, and the exploratory nature of the investigation is manifested in the fact that no hypotheses about such relationships are formulated before the data analysis is conducted."

Homburg & Krohmer 2009, S. 250

WHY STATISTICS?

Procedure for the determination of the study design



Inferential

"In the context of an explicative investigation, the focus is on causes of observed phenomena. Accordingly, it is concerned with relationships between variables, but here they are considered on the basis of hypotheses formulated in advance."

Homburg & Krohmer 2009, S. 250

Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics

Basic Statistics Concepts

- Mean (average)
- Median
- Mode
- Covariance
- Variance
- Standard deviation
- z-value
- Percentiles
- Probability distribution
- Sampling

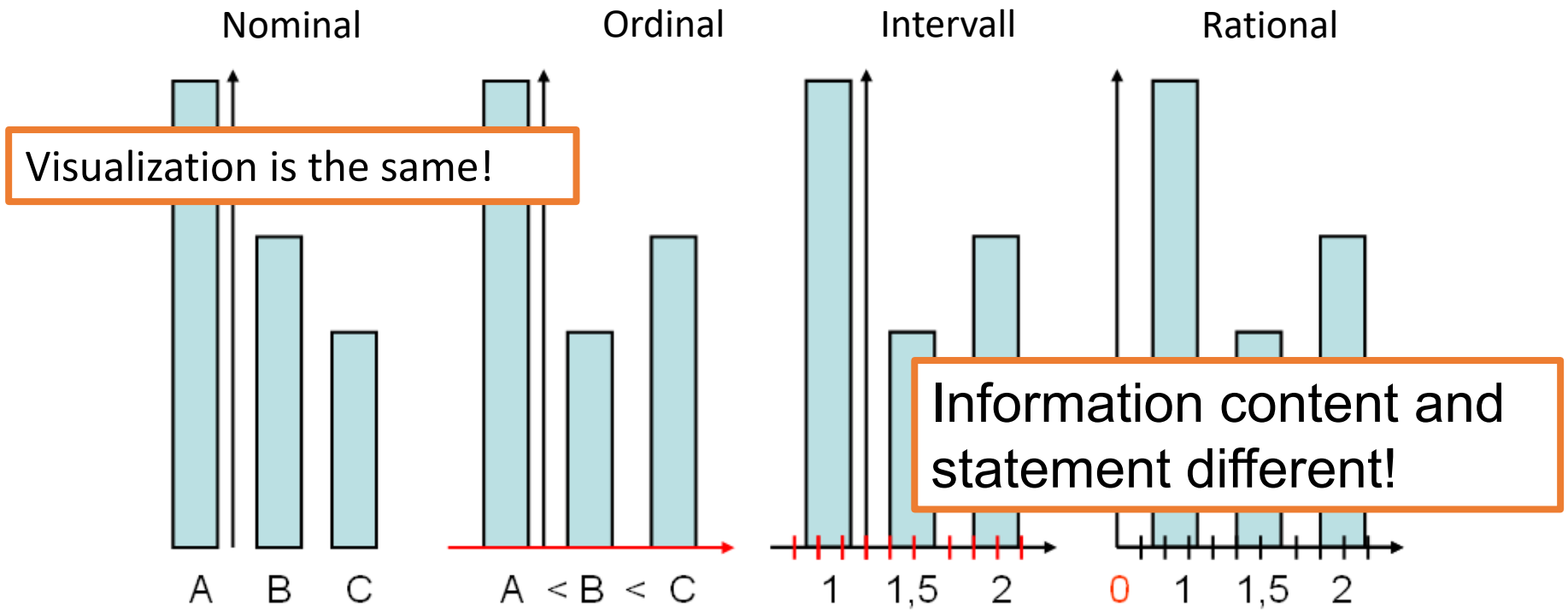
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics

Scales - Overview

Scales	Information	Description	Delimitation	Central tendency	dispersion	Example
Nominal	A = B und A ≠ B	Classification / codification of qualitative characteristics -> Real numbers are possible (code)	Designation of qualitative characteristics, the expression of which is not subject to ranking	Frequency Density		Sex job Name Postal code …
Ordinal	A = B und A ≠ B A > B > C	Ordered values / ranking -> Real numbers are possible	Ranking of qualitative characteristics, the expression of which is subject to natural ranking	Frequency Density Median Quartiles Percentile	Median-Deviation Range Interquartile range	School marks Efficiency levels Sport …
Intervall	A = B und A ≠ B A > B > C d = A + B d = B – A Gleichheit von Intervallen	Scale with constant distances and arbitrary zero point	Measurement of quantitative characteristics with constant distances and arbitrary zero point	Frequency Density Median Quartiles Percentile Arithm. Mean	Additionally: Standard deviation Variance	Day of the month IQ …
Rational	A = B und A ≠ B A > B > C d = A + B d = B – A c = A * B c = A / B Equality of ratios	Scale with true zero point -> Exactly comparable measuring ratios	Measurement of quantitative characteristics with natural zero point	Frequency Density Median Quartiles Percentile Arithm. Mean Geom. Mean Harmon. Mean	Additionally: Coefficient of variation	Length Area Weight Volume …

Scales - Summary

In addition to the evaluation and interpretation, the visualisation of the data must also be considered



https://de.wikipedia.org/wiki/Skalenniveau#/media/File:Skalenniveau.png

Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics
Basic Statistics

Exercises

Exercice 1

Indicate the scale level for the following examples!

Indicate the scale level for the following examples!

	Nominal	Ordinal	Interval	Ratio
Sex				
Social class				
Income tax				
Temperature in C				
Wind speed m/s				
Body weight				
School grades				
Exam points				
Time				
Eye colour				
Duration of treatment at the doctor				
Preferences				

Indicate the scale level for the following examples!

	Nominal	Ordinal	Interval	Ratio
Sex	X			
Social class		X		
Income tax				X
Temperature in C			X	
Wind speed m/s				X
Body weight				X
School grades		X		
Exam points				X
Time			X	
Eye colour	X			
Duration of treatment at the doctor				X
Preferences		X		

Exercice 2

Consider methods of empirical measurement for the following variables or empirical observations!

Consider methods of empirical measurement for the following variables or empirical observations!

	Methode(n)	Skala(en)
Internet Speed		
Cognitive stress		
Precipitation		
Volume		
Sportiness		

Consider methods of empirical measurement for the following variables or empirical observations!

	Method (s)	Scale(s)
Internet Speed	Measuring the Down- und Upload rate	Ratio scale
Cognitive stress	Survey Brain activity measurement (EEG)	Ordinal- or Interval scale Ratio scale
Precipitation	Measure (precipitation sensor)	Ratio scale
Volume	Measure (microphone)	Ratio scale
Sportiness	Competition Measure (time) Observation	Ordinal scale Ratio scale Ordinal scale

Assignment

Prof. Dr. Javier Valdes
Javier.valdes@th-deg.de

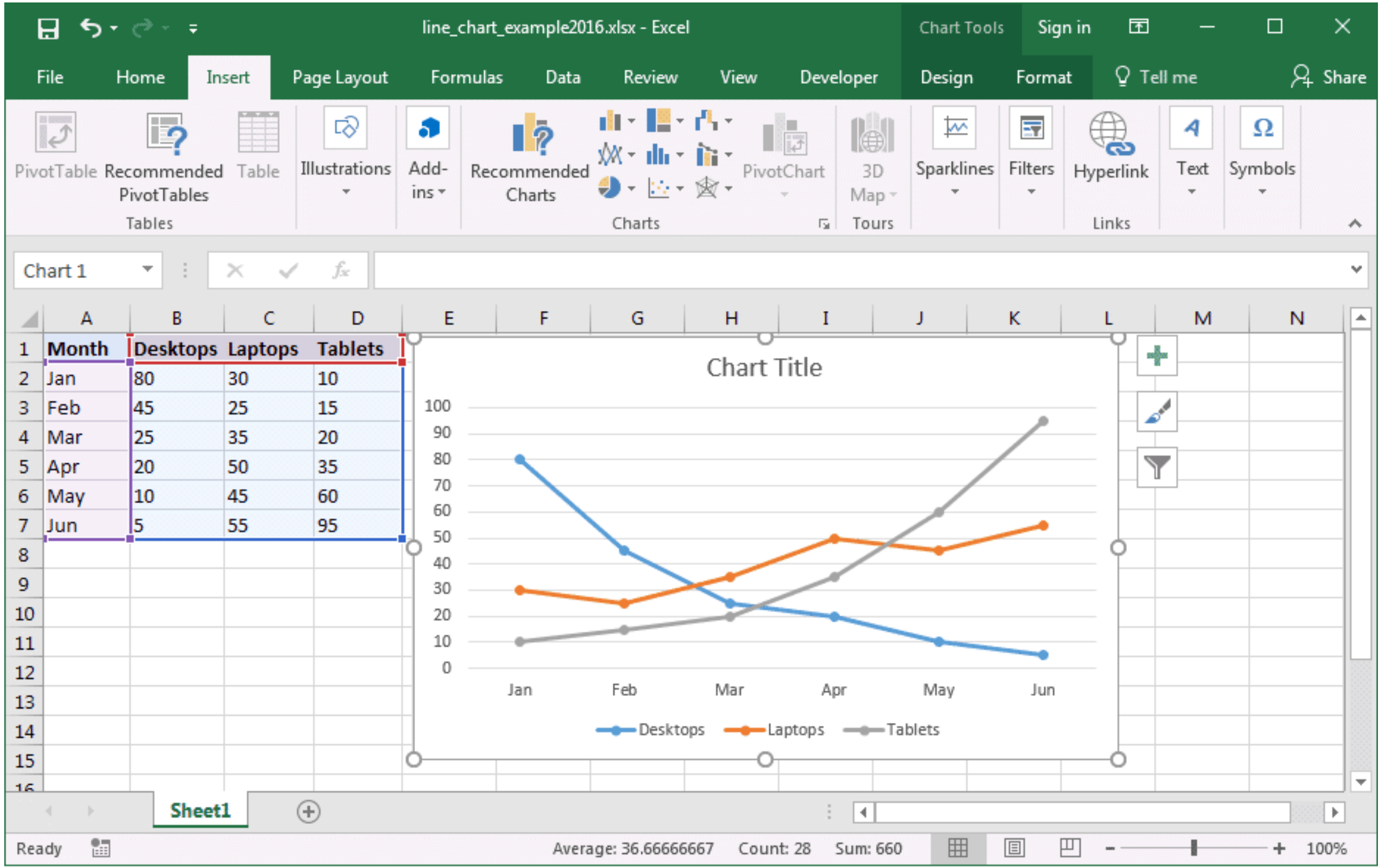
IAI - Institut für Angewandte Informatik
Institute for Applied Informatics

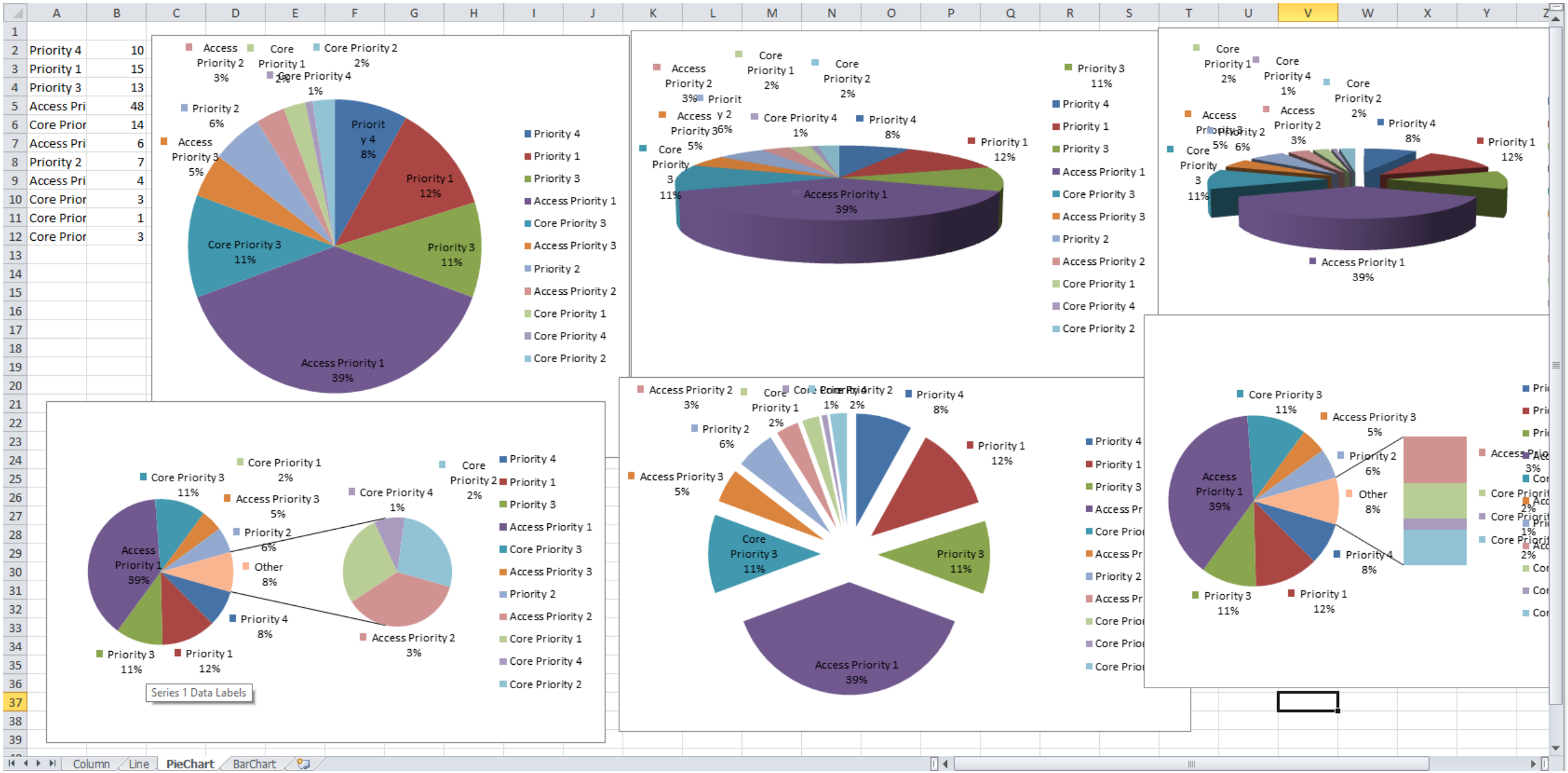
Technische Hochschule Deggendorf
Technologie Campus Freyung
Grafenauer Str. 22, D- 94078 Freyung,
Germany

Tel.: +49 8551 91764 40
Fax: +49 8551 91764 69

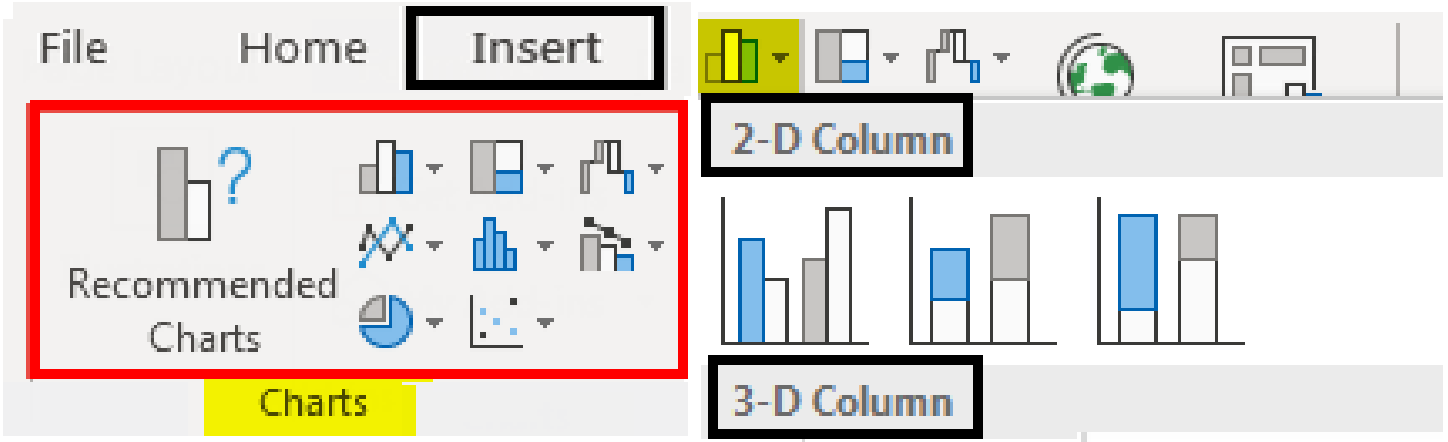
Curvefit Rankings for 0.5% SEV CO_2

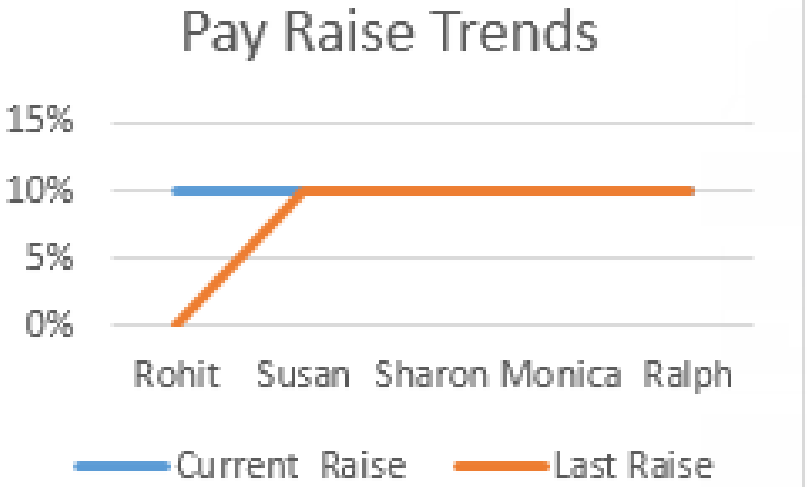
Rank	F-statistic
1	7.962883557
2	7.8601110639
3	7.5283512645
4	7.3357010958
5	6.3801158367
6	3.8079206858
7	3.742891358
8	3.5727028219
9	3.5546937052
10	3.0133372321
11	2.7408796673
12	2.6986270263
13	2.5801276758
14	2.5622800357
15	2.1798855221

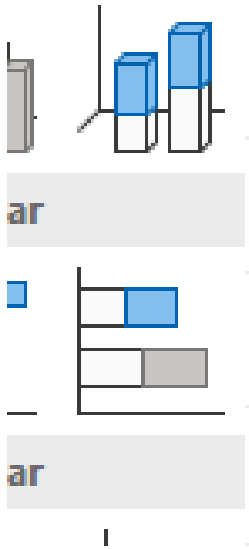

line_chart_example2016.xlsx - Excel
Chart Tools
Sign in
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Developer
Design
Format
Tell me
Share
PivotTable
Recommended PivotTables
Table
Tables
Illustrations
Add-ins
Recommended Charts
PivotChart
Charts
3D Map
Tours
Sparklines
Filters
Hyperlink
Text
Symbols
Links
Chart 1
Month
Desktops
Laptops
Tablets
Jan 80 30 10
Feb 45 25 15
Mar 25 35 20
Apr 20 50 35
May 10 45 60
Jun 5 55 95
Chart Title
100
90
80
70
60
50
40
30
20
10
0
Jan
Feb
Mar
Apr
May
Jun
Desktops
Laptops
Tablets
Sheet1
Ready
Average: 36.66666667
Count: 28
Sum: 660
100%

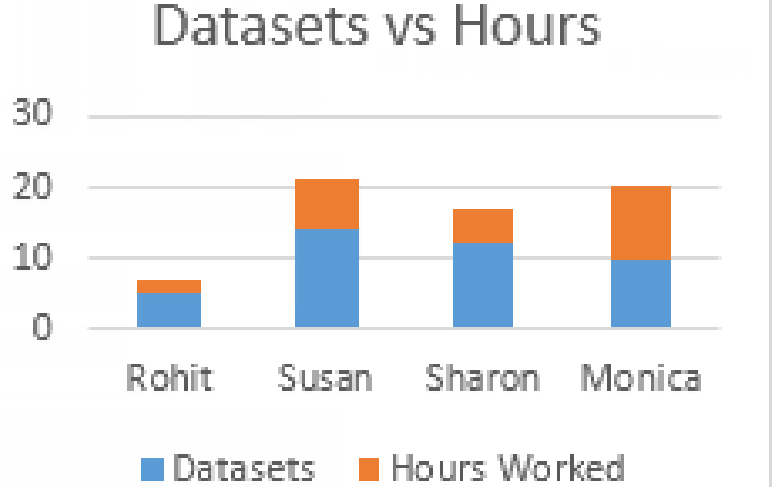

Priority 4 10
Priority 1 15
Priority 3 13
Access Pri 48
Core Prior 14
Access Pri 6
Priority 2 7
Access Pri 4
Core Prior 3
Core Prior 1
Core Prior 3
Access Pri 1
Access Priority 2 3%
Core Priority 1 2%
Core Priority 2 2%
Core Priority 4 1%
Priority 2 6%
Access Priority 3 5%
Priority 4 8%
Priority 1 12%
Priority 3 11%
Core Priority 3 11%
Access Priority 1 39%
Series 1 Data Labels
Other 8%

Types of Graphs in Excel










1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40

“The longer you live in the past,
the less future you have to enjoy.”

— Robert Tew