Visualizations in the Analysis Workflow

ASQ Calgary (section 409)

Thomas Speidel, P.Stat., Data Scientist

2020/01/27 | A copy of this presentation is available on GitHub: https://github.com/tspeidel/ASQ | Views are my own.

Agenda

In this presentation, I will provide an overview of the role that visualizations play in today's world of Big Data, AI (artificial intelligence), and ML (machine learning). Furthermore, I will illustrate how the second objective of visualizations addresses key principles of data and information quality.

Part 1. Establishing a Common Dictionary: Quality, Big Data, AI

Part 2. Visualizations: Objectives and Constraints

Part 3. Applications and Good Practices

Data Manifesto

Data, in an by themselves, do not directly create knowledge. While efforts to collect and store data have increased dramatically in recent years, nearly absent is a focus on knowledge creation. In the age of **Big Data**, the availability of vast amounts of information can coexist with the **absence of knowledge**.

Data cannot speak for themselves. It is in when we interpret data that knowledge is created. My focus is on **bridging the gap between data and knowledge creation**. That gap is filled by statistics and evidence based decision making.

Part 1
A Common Dictionary

Quality is not Binary

Statistics Canada defines quality along 6-dimensions.

1. Relevance
2. **Accuracy**: the degree to which the information correctly describes the phenomena it was designed to measure
3. Timeliness
4. Accessibility
5. **Interpretability**: the information necessary to interpret and utilize it appropriately
6. **Coherence**: the degree to which it can be successfully brought together with other information

> *Quality must be built in at each phase of the process*

[1] *Government of Canada, S. C. (2009). Statistics Canada Quality Guidelines*

How do we assess the 6-dimensions?

Principle of **Intelligent transparency**

Accessible, comprehensible, usable, assessable.

- The consumer of the information needs to be aware of key quality aspects, irrespective of role.
- Some dimensions are obvious (e.g. timeliness), others require analyses
- The use of visualization play a key role in both discovering and communicating the quality of the information.

[1] *Royal Society. (n.d.). Final report - Science as an open enterprise*

> *... and it almost always happens that data that are ideally fit for one use are marginally fit for a second and poorly fit for a third.*

Redman, T. C. (2001). Data quality: The Field guide.

Big Data

We have lots and lots of data! We don't need to worry about quality or interpretation!

- Big Data largely refers to the management of vast quantities of data (4V's Volume, Velocity, Variety, Veracity).
- Main challenges of Big Data are: management, relevance, coherence, accuracy.
- Relevance, coherence, accuracy are challenges of **any** data. Thus, **management** is the differentiator (cluster computing, cloud ecosystems, model management, metadata management, GPU, etc).

What About AI?

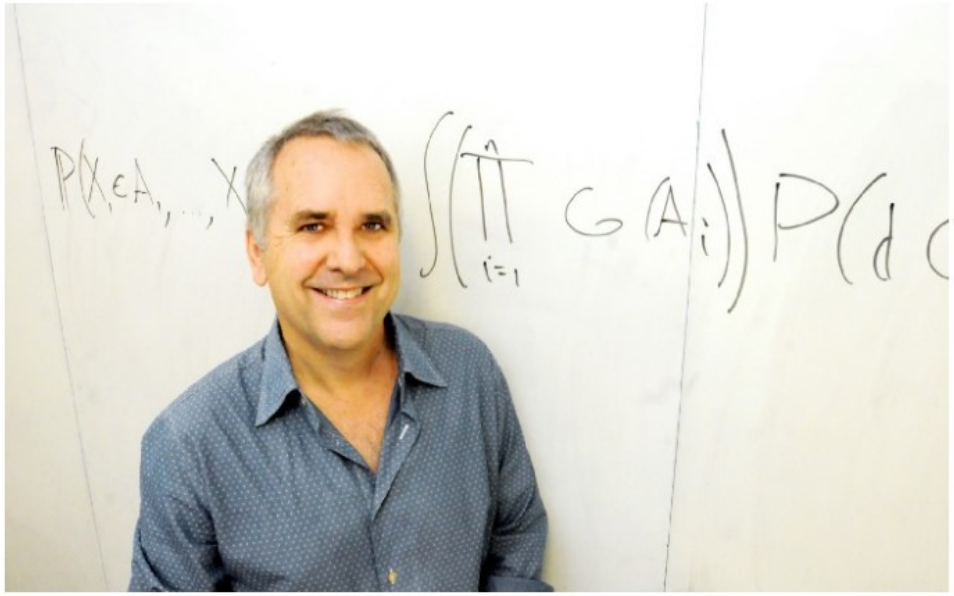


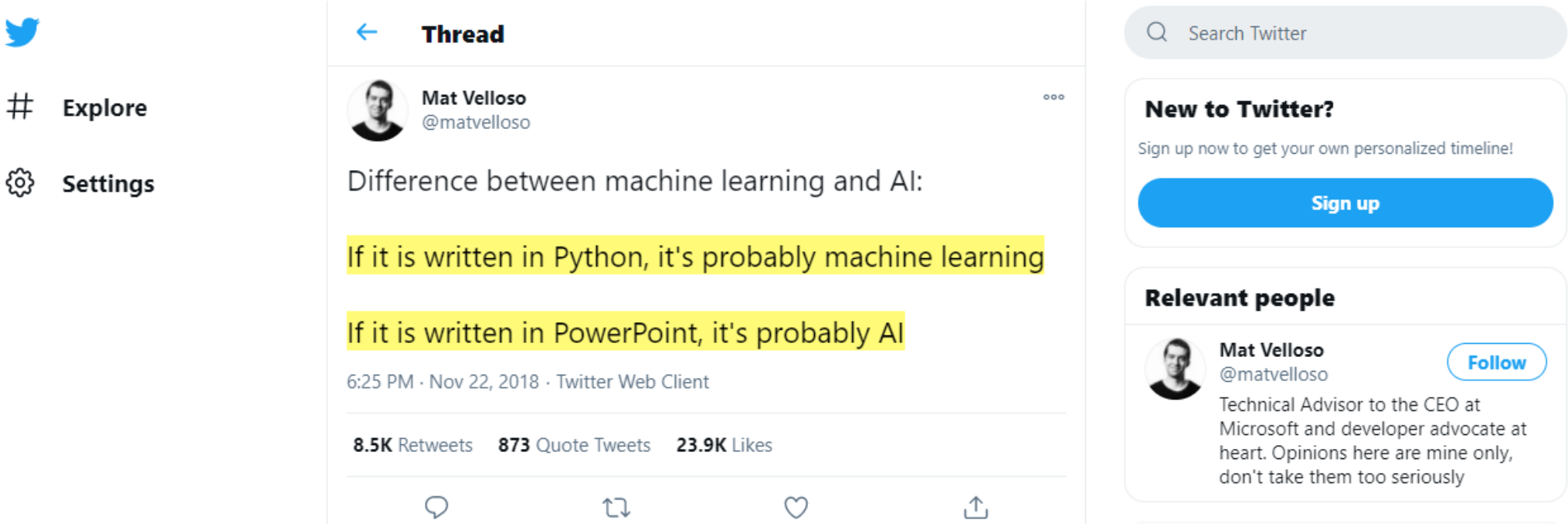
Photo credit: Peg Skorpinski

Michael Jordan is an American scientist, professor at the University of California, Berkeley and researcher in machine learning, statistics, and artificial intelligence. He is one of the leading figures in machine learning, and in 2016 Science reported him as the world's most influential computer scientist. (Wikipedia).

- The science of developing computer systems that can perform tasks normally requiring human intelligence.
- "*Most of what is being called "AI" today, is what has been called "Machine Learning" (ML) for the past several decades. ML is an algorithmic field that blends ideas from statistics, computer science and many other disciplines to design algorithms that process data, make predictions and help make decisions*".
- *Rebranding of well established ideas and technologies*.
- "AI" used as an intellectual wildcard "*one that makes it difficult to reason about the scope and consequences of emerging technology*".

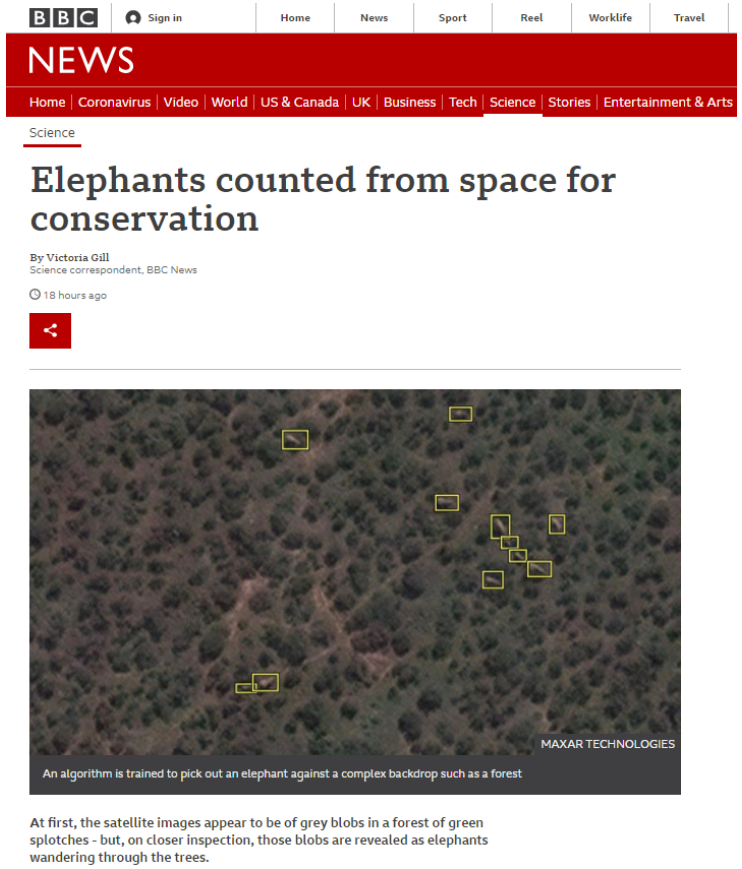
Jordan, M. (2018, April 30). Artificial intelligence — The revolution hasn't happened yet. Medium.

… and what practitioners think…



Velloso, M. (2018).

What we often mean by AI (and where it works well)



- Humans label thousands of satellite images of elephants
- A deep learning algorithm (a type of Artificial Neural Network) is trained on the labelled images
- If the training is satisfactory, the algorithm is applied on new data *of the same stream*
- Besides the algorithm, the data pipeline and technology stack are important aspects

Gill, V. (2021, January 21). Elephants counted from space for conservation. BBC News.

… and where it ~~works well~~ doesn't work well

> *In short, the biggest difference between AI then and now is that the necessary computational capacity, raw volumes of data, and processing speeds are readily available so the technology can shine.*

Hammond, C. (2015). Why artificial intelligence is succeeding: Then and now.

Part 2
Visualizations: Objectives and Constraints

Two objectives

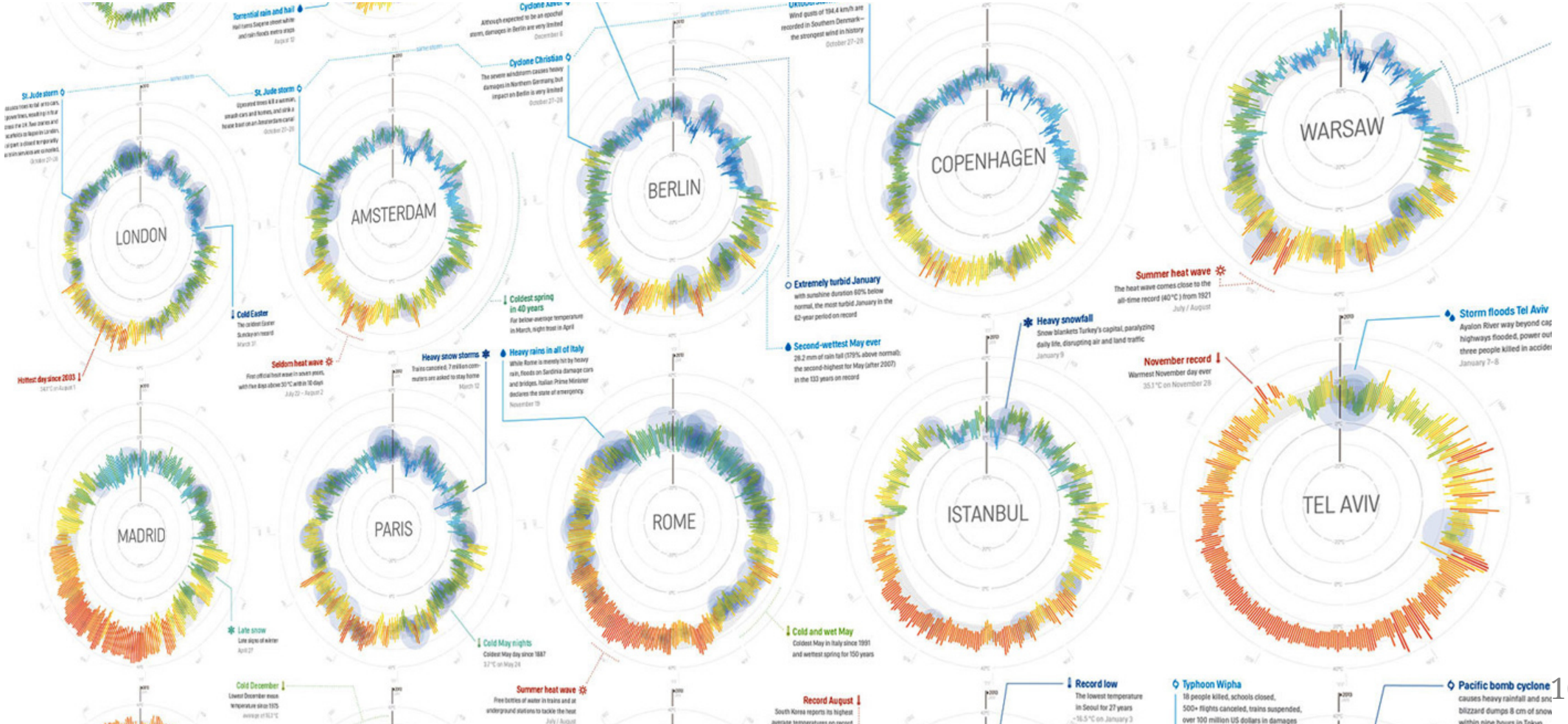
Communication (first goal): display a convincing pattern, attract reader's attention

Discovery (second goal): observe deviations from our expectations

[1] *Gelman, Unwin. 2013*

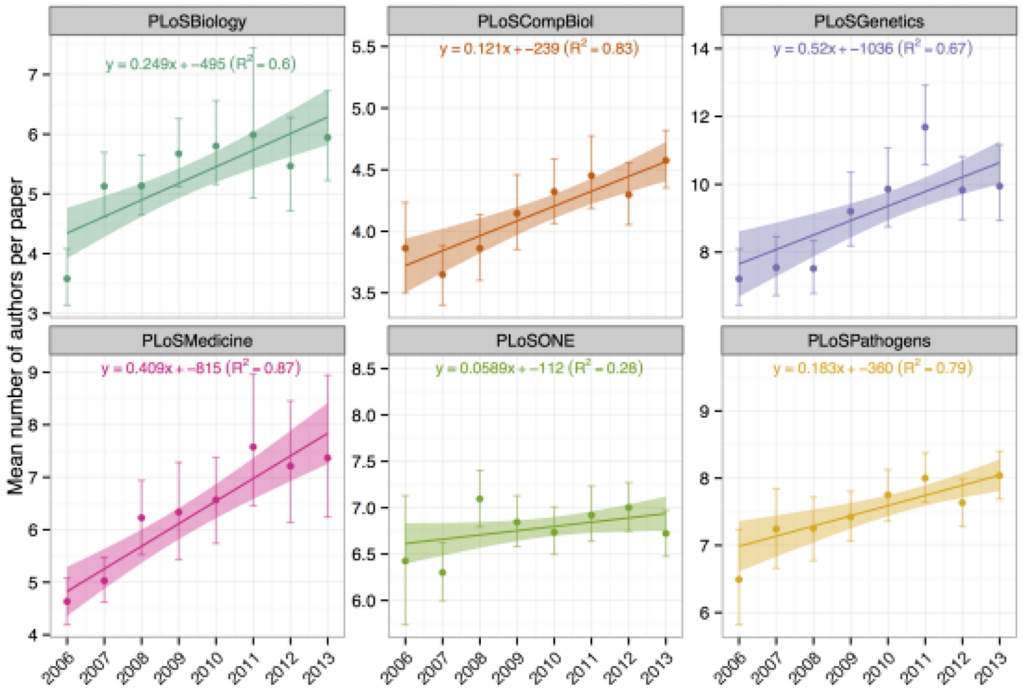
First Goal of Visualizations: Communication

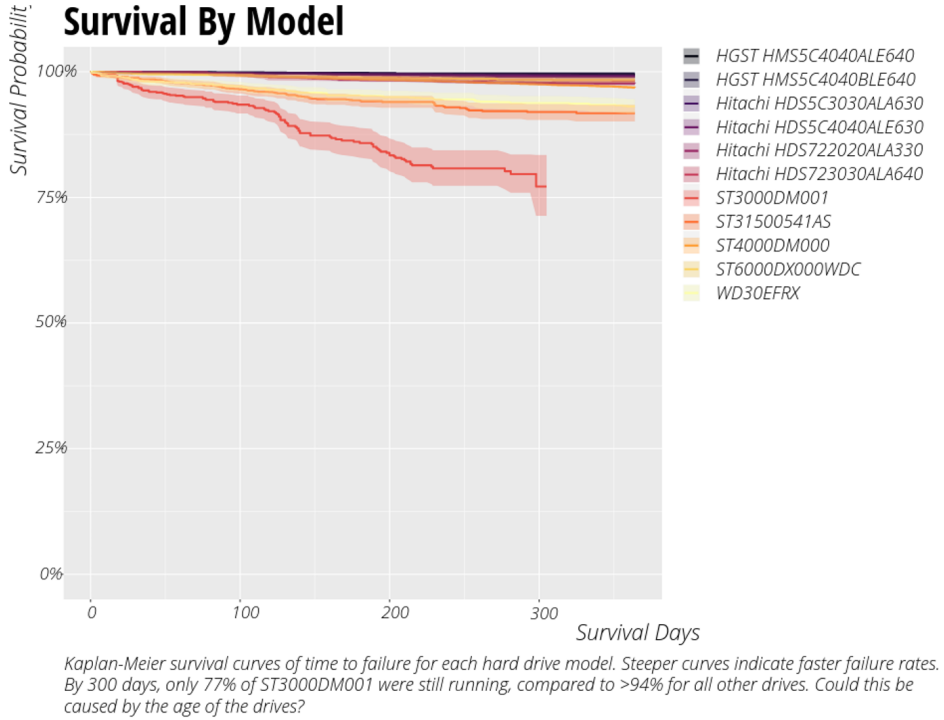
To summarize, communicate and engage (telling a story)



Second Goal of Visualizations: Discovery

To identify, understand, highlight features or relationships (helping the readers make up their mind)





Kaplan-Meier survival curves of time to failure for each hard drive model. Steeper curves indicate faster failure rates. By 300 days, only 77% of ST3000DM001 were still running, compared to >94% for all other drives. Could this be caused by the age of the drives?

[1] *Gelman, Unwin. 2013*

To Persuade or to Inform?

Why do we have to choose? Because we **cannot** achieve both effectively at the same time (*Spiegelhalter, Pearson, Short. 2011*).

To Persuade

To Inform

Nutrition Facts

Serving Size 1 cup (228g)
Servings Per Container 2

Amount Per Serving	
Calories 250	Calories from Fat 110
	% Daily Value*
Total Fat 12g	**18%**
Saturated Fat 3g	**15%**
Trans Fat 3g	
Cholesterol 30mg	**10%**
Sodium 470mg	**20%**
Total Carbohydrate 31g	**10%**
Dietary Fiber 0g	**0%**
Sugars 5g	
Protein 5g	
Vitamin A	4%
Vitamin C	2%
Calcium	20%
Iron	4%

* Percent Daily Values are based on a 2,000 calorie diet. Your Daily Values may be higher or lower depending on your calorie needs.

	Calories	2,000	2,500
Total Fat	Less than	65g	80g
Sat Fat	Less than	20g	25g
Cholesterol	Less than	300mg	300mg
Sodium	Less than	2,400mg	2,400mg
Total Carbohydrate		300g	375g
Dietary Fiber		25g	30g

~~Visualizations for Data Storytelling~~
~~(First Goal: Communication)~~

Part 3: Applications and Good Practices

Why Should We Care?

1. Because data visualization is an integral and fundamental part of the Data Science/ML/AI workflow. In most cases, visualizations help us uncover fundamental problems with the way we use data.

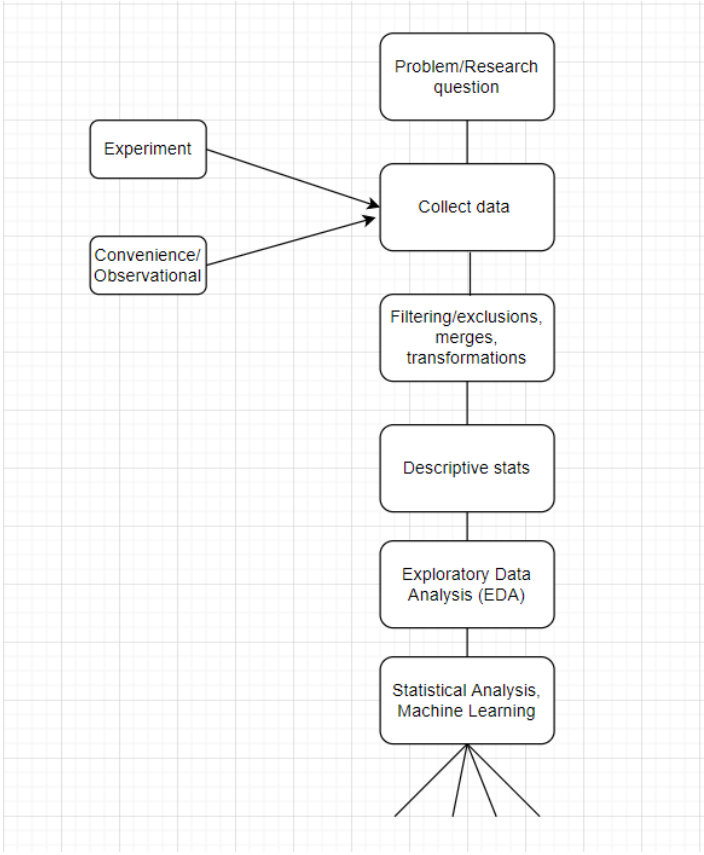
2. Because in a world of Big Data, AI, ML, IoT, **veracity** is more important than ever: visualization remain one of the main tool by which we assess data and information accuracy.

> *The transformation can only be accomplished by man, not by hardware (computers, gadgets, automation, new machinery). A company cannot buy its way into quality* - Edward Deming (1982).

[1] *Kuonen, D. (2018). (Big) Data as the Fuel and Analytics as the Engine of the Digital Transformation*

Why Should We Care?

What Workflow?



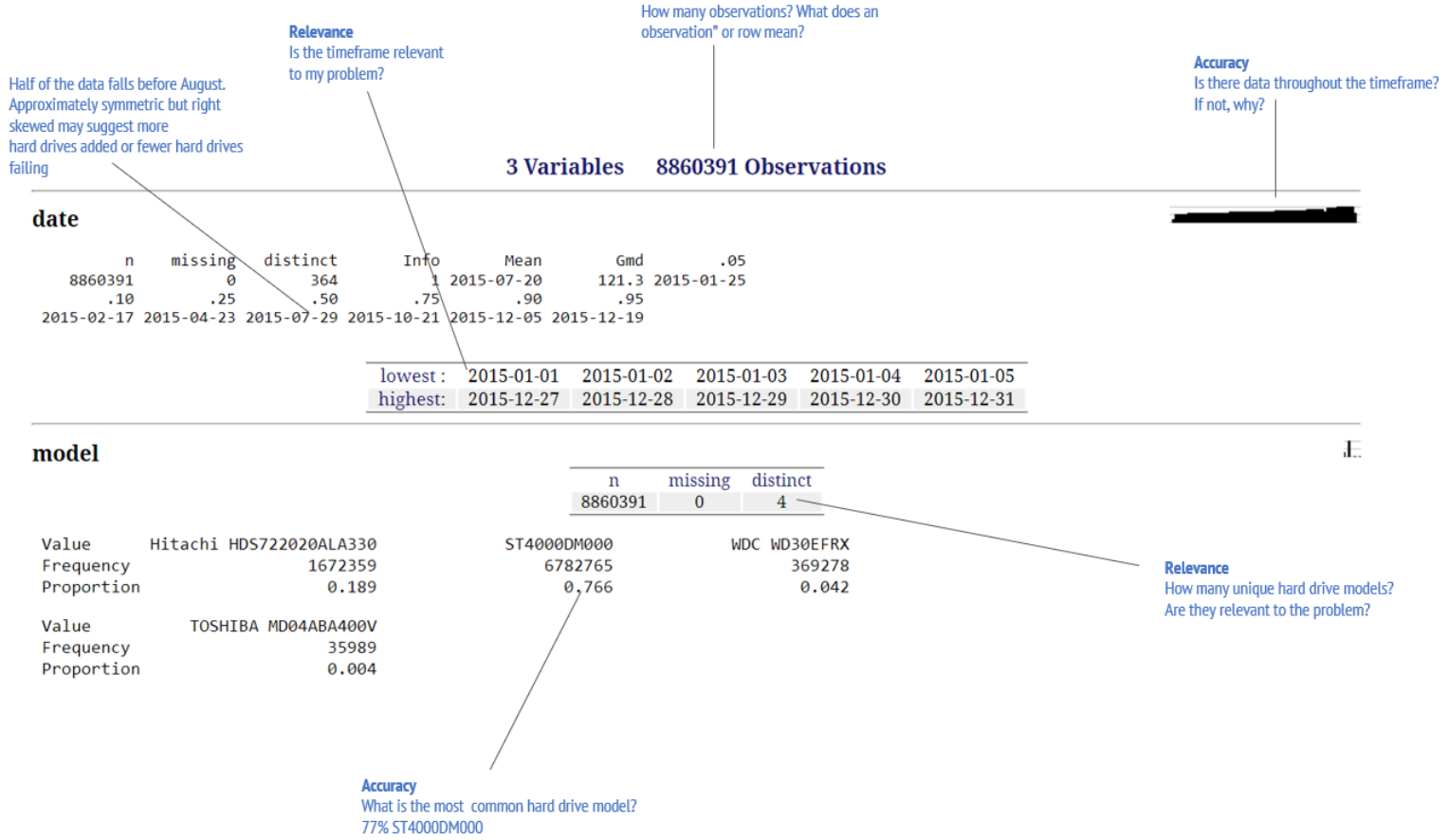
Summary Statistics

Critically examine the data to see if it meets your expectations. Some key questions we asked ourselves:

- What is an "observation" (row)?
- Does the data capture the process or problem I want to understand?
- Does the data capture well-established domain knowledge?
- Are there gaps in the data? If so, why?
- Are there extremes and outliers? Can we tell them apart?

No need to be a Data Scientist to ask ourselves these questions!

Summary Statistics

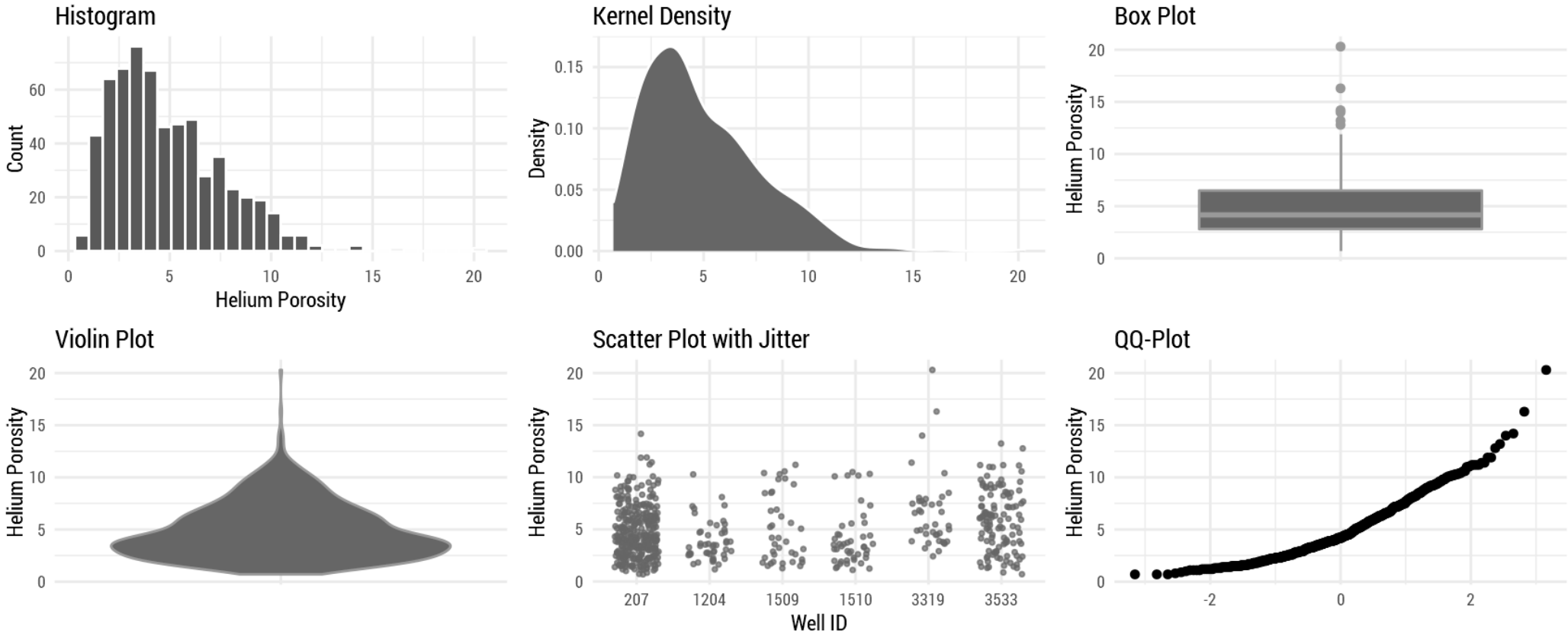


Population:
4 hard drive models observed daily over the year 2015 (one observation per drive model per day).

Jargon:
Panel or longitudinal data.

Relevance:
Is this relevant to the problem at hand?

Alternative Univariate Visualizations

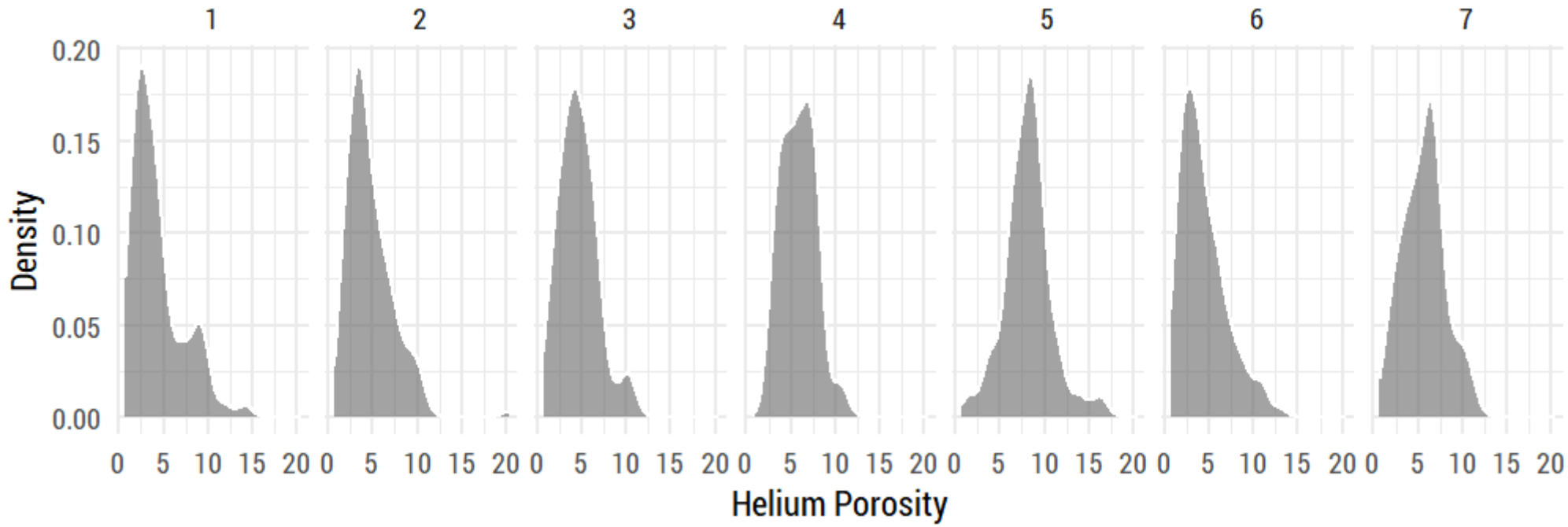


Bivariate Visualizations

> *Numbers become evidence by being in relation to.*
> -Edward R. Tufte, Visual Explanations: Images & Quantities, Evidence & Narrative (1997)

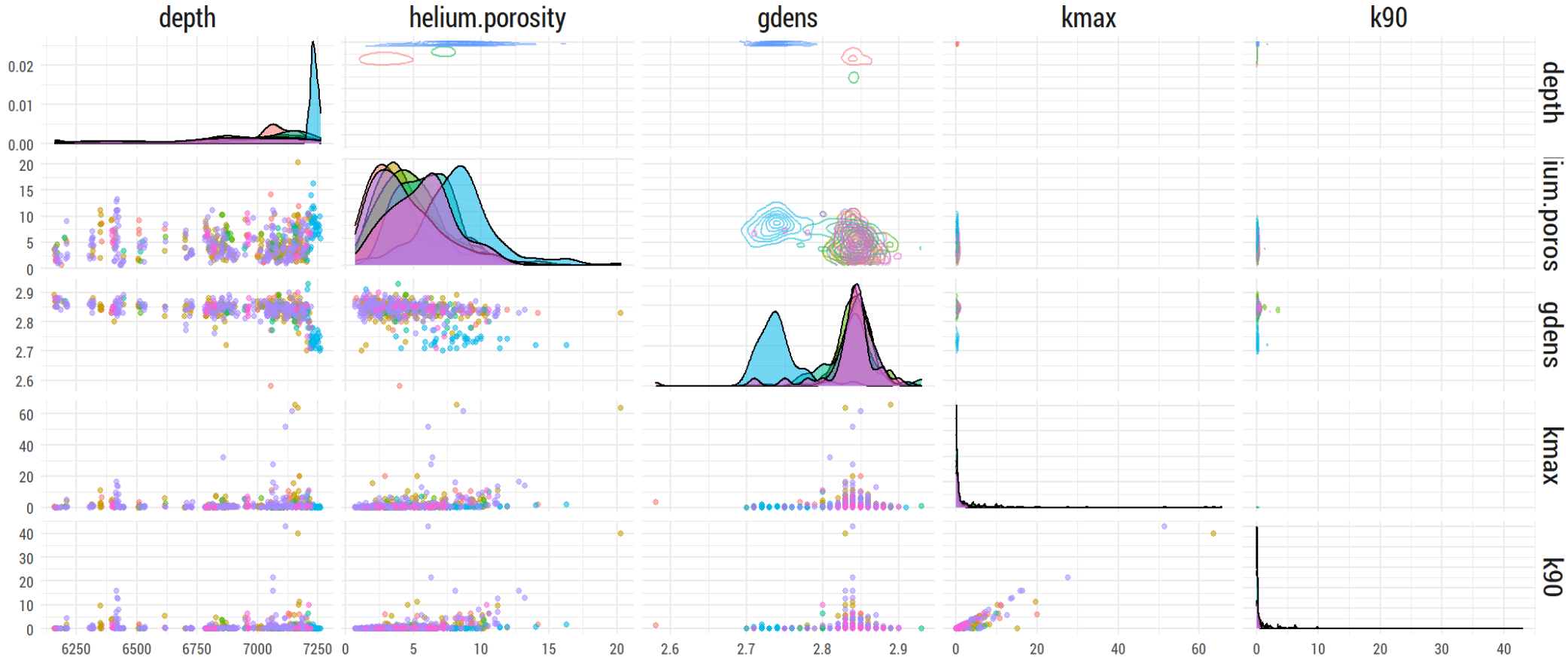
Kernel Density

Kernel density plots have the unique advantage that they are self-standardizing. They are also more effective than histograms. This makes them useful to **compare distributions** when the units aren't the same (not here).



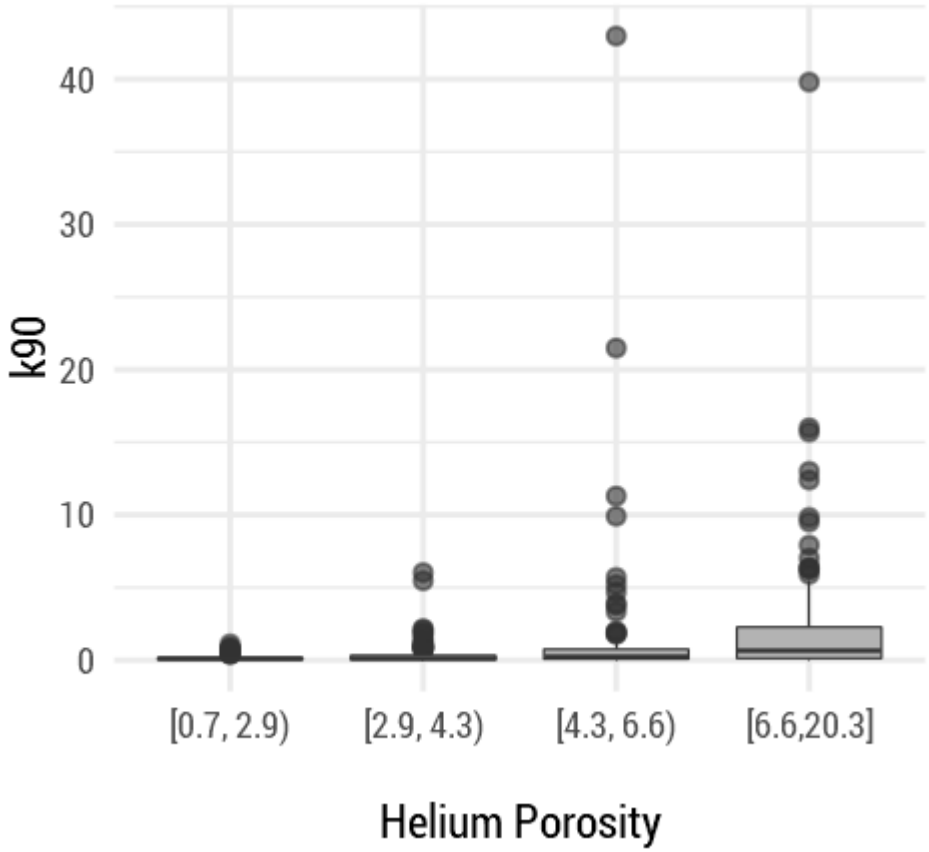
Scatterplot Matrix or Cross Plot

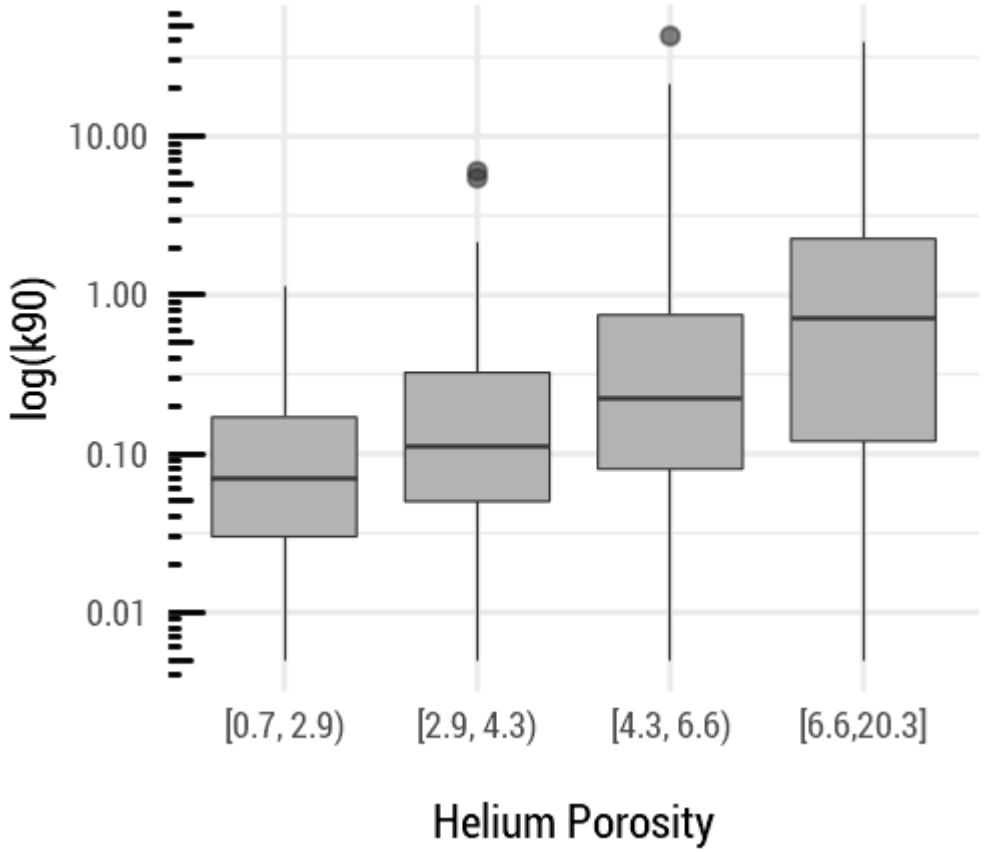
One of the very first plots you ant to make. Shows all pairwise combinations. Here, I've cheated by removing a few outliers.



Boxplots

Here we *cut* (bin) `helium.porosity` into quartiles (P25, P50, P75).





Scatterplots With 3 Variables

Scatterplots With 4 Variables

Correlation Matrix

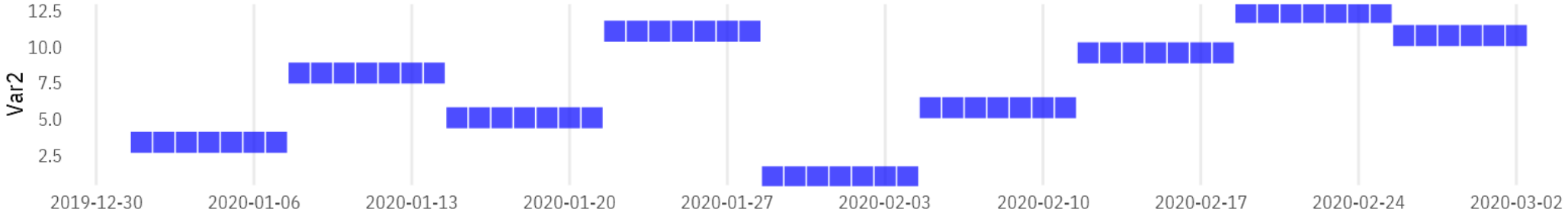
A unitless measure of the strength of the association between pairs of variables.

Last Value Carried Forward

Date	Var1	Var2
2020-01-15	2.7	3.4
2020-01-14	6.7	3.4
2020-01-05	2.6	5.1
2020-01-10	1.7	8.2
2020-01-16	9.0	3.4
2020-01-04	2.3	5.1
2020-01-11	9.4	8.2
2020-01-13	6.0	8.2
2020-01-06	4.4	5.1
2020-01-08	3.4	8.2
2020-01-07	2.7	8.2
2020-01-01	9.9	5.1
2020-01-12	4.6	8.2
2020-01-09	2.7	8.2
2020-01-02	6.2	5.1
2020-01-03	2.0	5.1
...	...	...

- Common when data is collected for reporting/regulatory purposes
- May not be so obvious to detect (sort order)
- Big data mirage: effective sample size for `Var2` is n=3 (not 16!)
- Problem: if undetected, can damage analysis and conclusions
    - Analysts need to watch out
    - Data providers should not assume values need to be filled/in

Last Value Carried Forward



> *Information displays should be documentary, comparative, causal and explanatory, quantified, multivariate, exploratory, skeptical.*
>
> -Edward R. Tufte, Visual Explanations: Images & Quantities, Evidence & Narrative (1997)

General Graphing

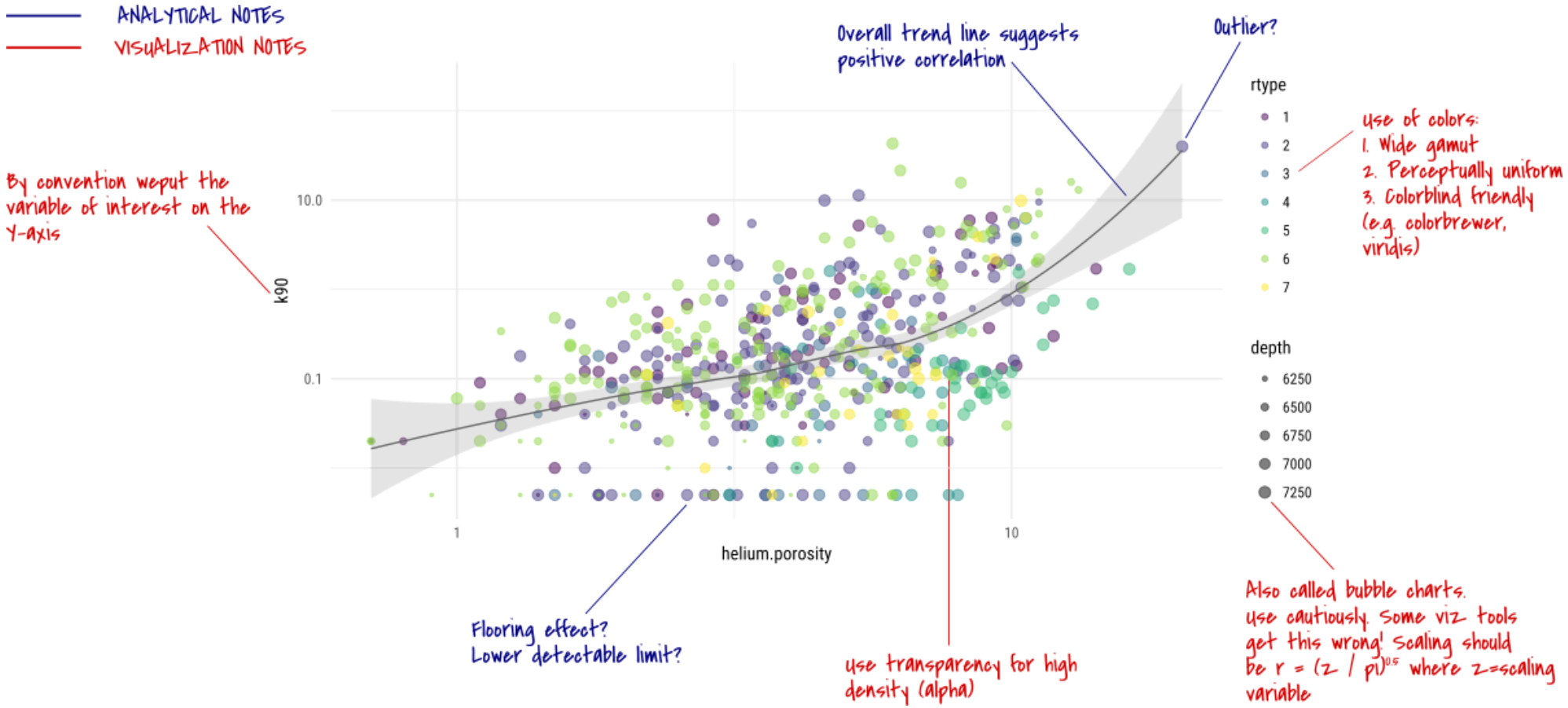
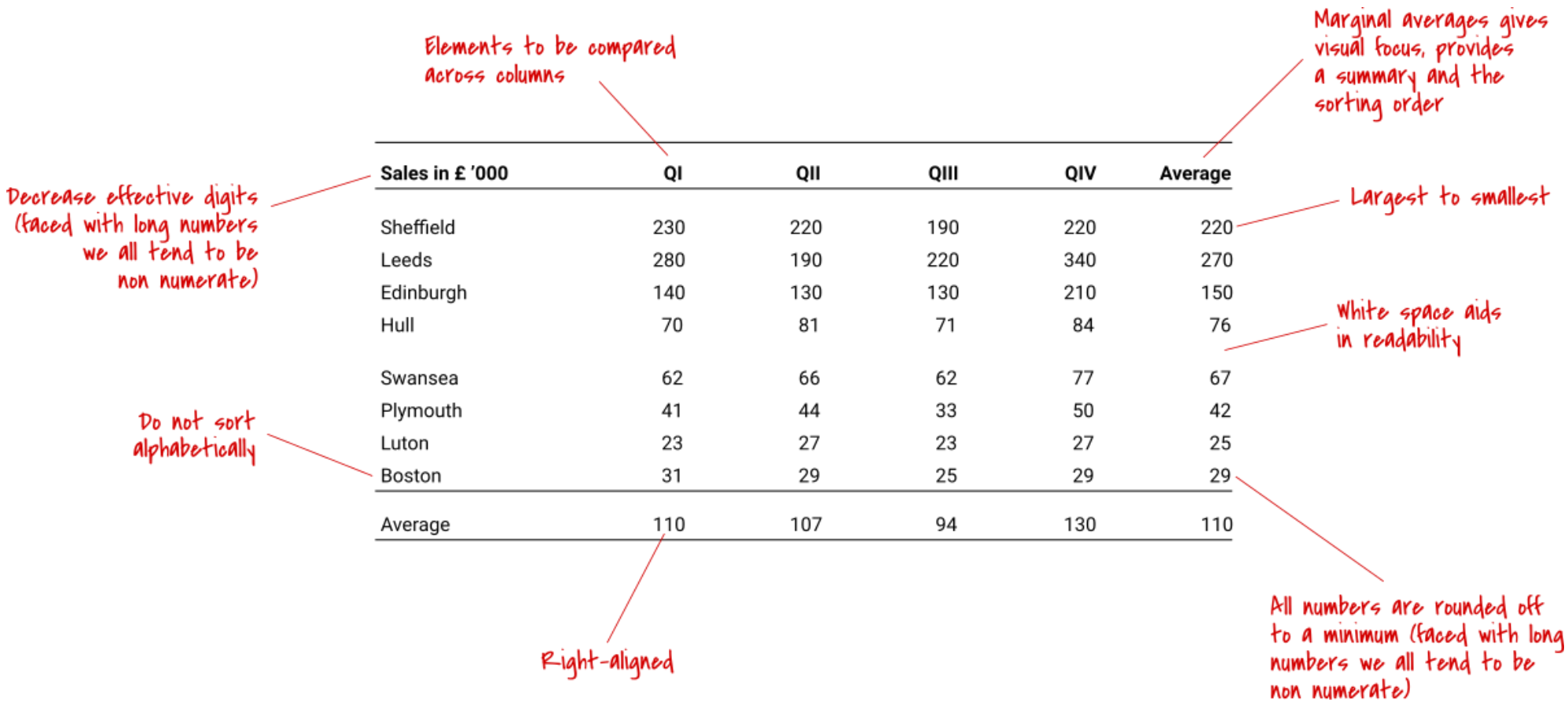


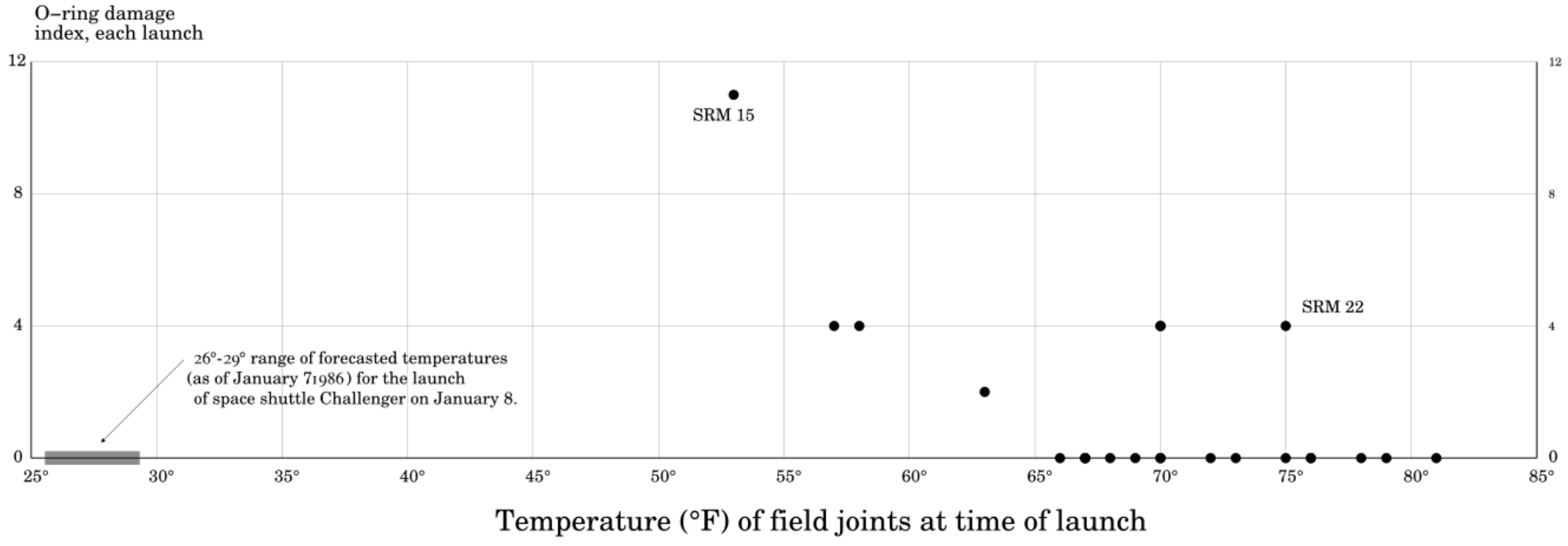
Table Tips



Sales in £ '000	QI	QII	QIII	QIV	Average
Sheffield	230	220	190	220	220
Leeds	280	190	220	340	270
Edinburgh	140	130	130	210	150
Hull	70	81	71	84	76
Swansea	62	66	62	77	67
Plymouth	41	44	33	50	42
Luton	23	27	23	27	25
Boston	31	29	25	29	29
Average	110	107	94	130	110

[1] *Recreated from Ehrenberg, 1981*

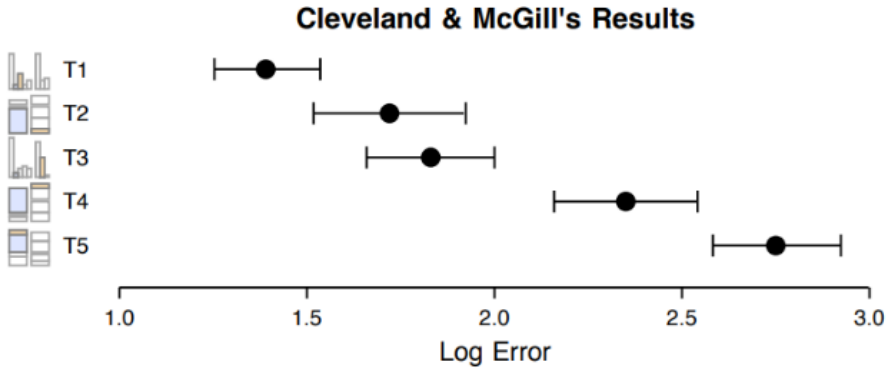
Data-to-ink ratio

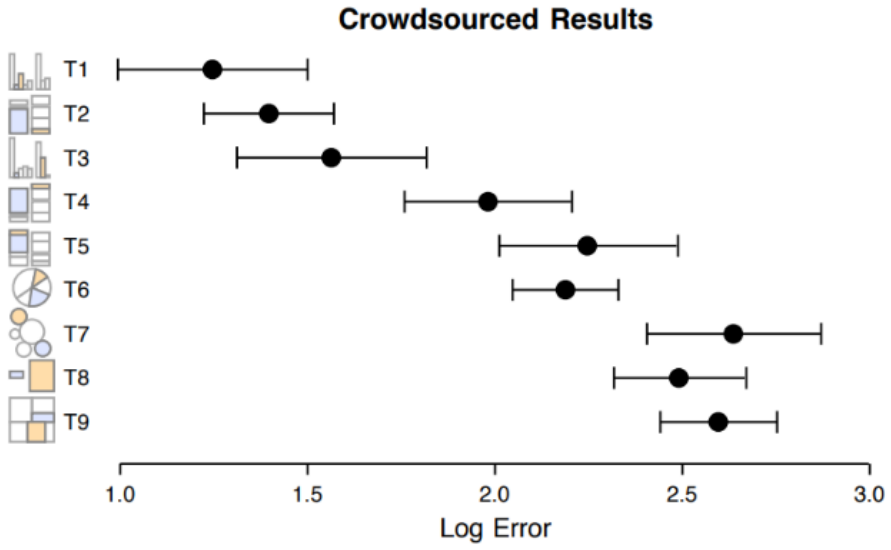


[1] *Recreated from Tufte, 1983. Visual Explanations: Images and Quantities, Evidence and Narrative*

The Type of Chart Affects Visual Judgement

Experiments performed to quantify visual judgment of graphs: 25 years later, results have not changed.





[1] *Cleveland and McGill, 1984* and *Heer and Bostok, 2010*

Avoid 3D

Why

Removes clarity, adds obfuscation, inhibits comprehension, does not help retain information, third dimension is usually inexistent (and if it did exist, you would try to avoid it).

Exceptions

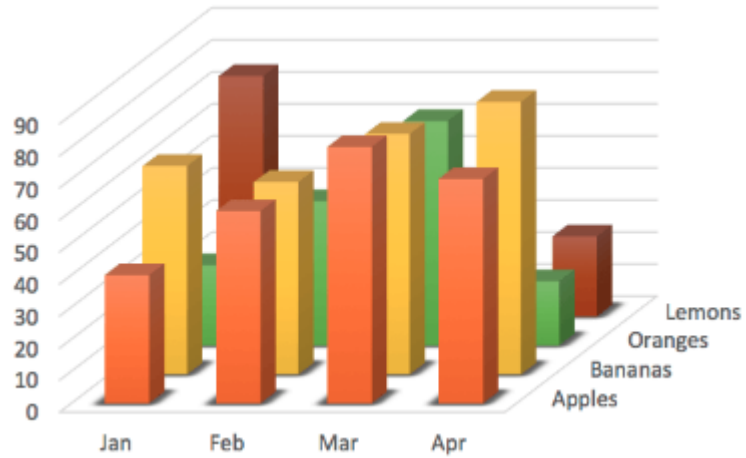
Spatial data, interaction surfaces.

Advantages

A reader is more likely to remember a 3D graph than a plain graph.

Disadvantage

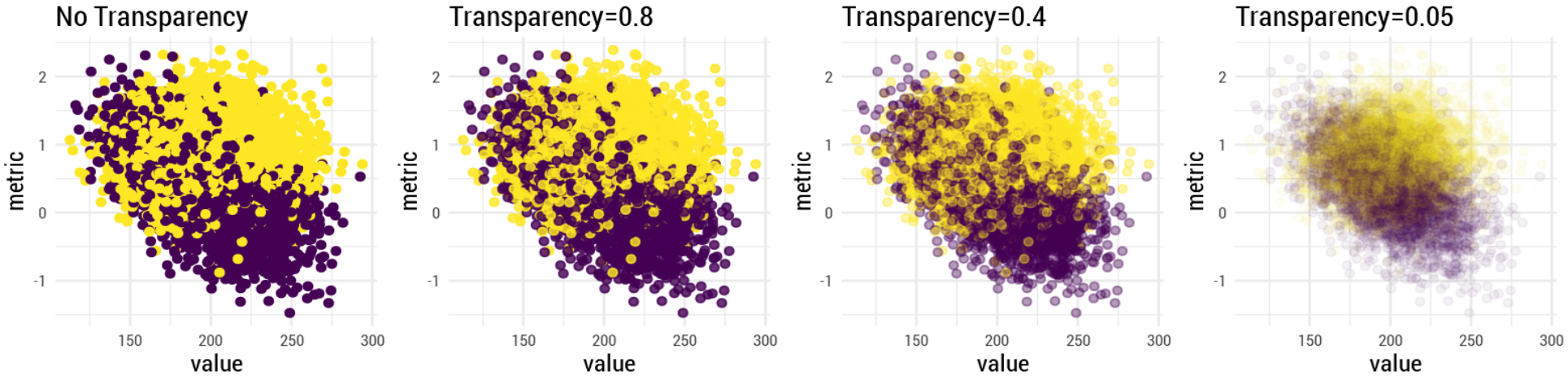
A reader is more likely to be confused by a 3D graph than a plain graph.



Visualizing High Density Data

Good...

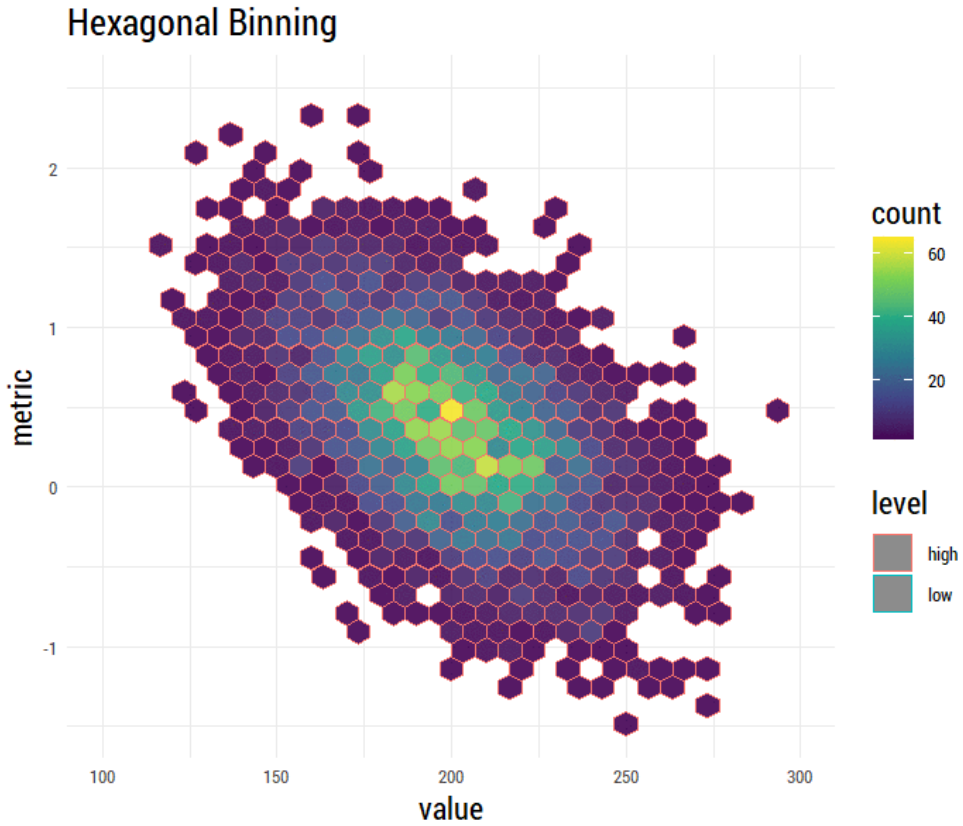
Add **transparency** (alpha) or reduce symbol size. Easy and quick to implement: reveals underlying patterns but may hide outliers.

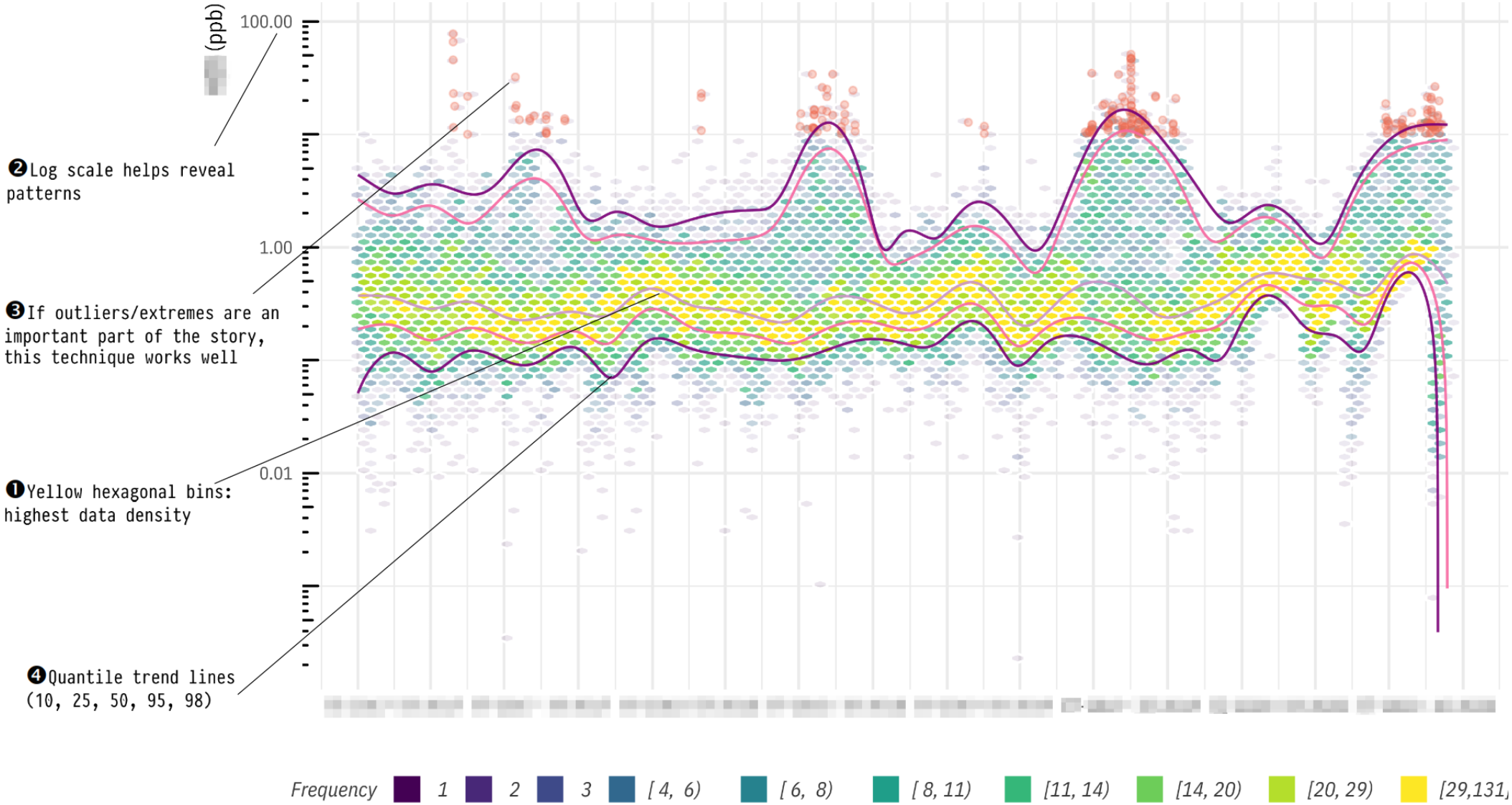


Visualizing High Density Data

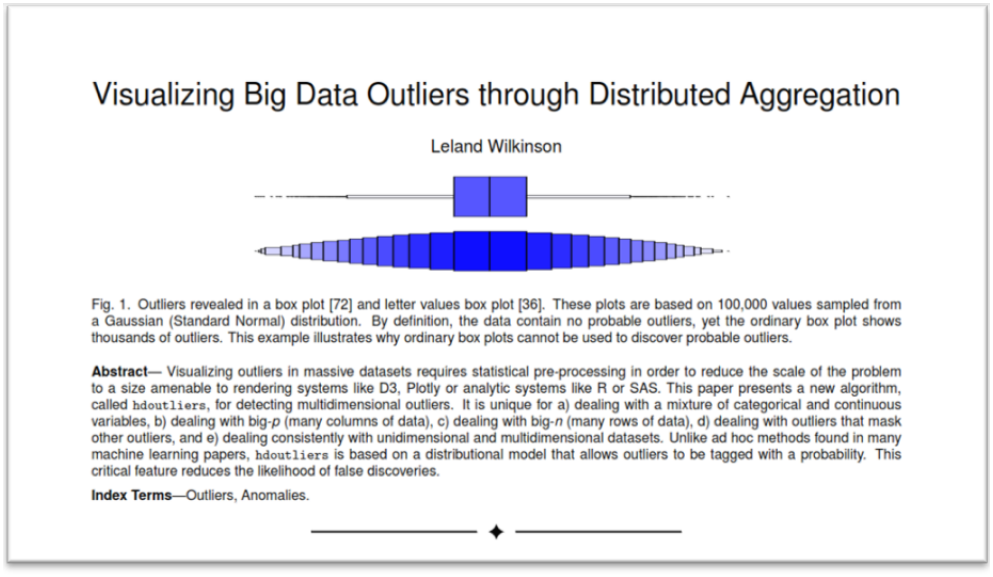
Better...

Use **binning techniques** for high density data (they've been around since the mid 80's!). A little harder to read, but won't miss outliers and sparse data.




(ppb)
100.00
1.00
0.01
❷Log scale helps reveal patterns
❸If outliers/extremes are an important part of the story, this technique works well
❶Yellow hexagonal bins: highest data density
❹Quantile trend lines (10, 25, 50, 95, 98)
Frequency
1
2
3
[4, 6)
[6, 8)
[8, 11)
[11, 14)
[14, 20)
[20, 29)
[29,131]

Visualizing High Density Data



Visualizing Big Data Outliers through Distributed Aggregation

Leland Wilkinson

Fig. 1. Outliers revealed in a box plot [72] and letter values box plot [36]. These plots are based on 100,000 values sampled from a Gaussian (Standard Normal) distribution. By definition, the data contain no probable outliers, yet the ordinary box plot shows thousands of outliers. This example illustrates why ordinary box plots cannot be used to discover probable outliers.

Abstract— Visualizing outliers in massive datasets requires statistical pre-processing in order to reduce the scale of the problem to a size amenable to rendering systems like D3, Plotly or analytic systems like R or SAS. This paper presents a new algorithm, called `hdoutliers`, for detecting multidimensional outliers. It is unique for a) dealing with a mixture of categorical and continuous variables, b) dealing with big-*p* (many columns of data), c) dealing with big-*n* (many rows of data), d) dealing with outliers that mask other outliers, and e) dealing consistently with unidimensional and multidimensional datasets. Unlike ad hoc methods found in many machine learning papers, `hdoutliers` is based on a distributional model that allows outliers to be tagged with a probability. This critical feature reduces the likelihood of false discoveries.

Index Terms—Outliers, Anomalies.

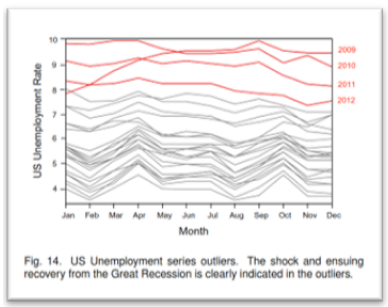


Fig. 14. US Unemployment series outliers. The shock and ensuing recovery from the Great Recession is clearly indicated in the outliers.

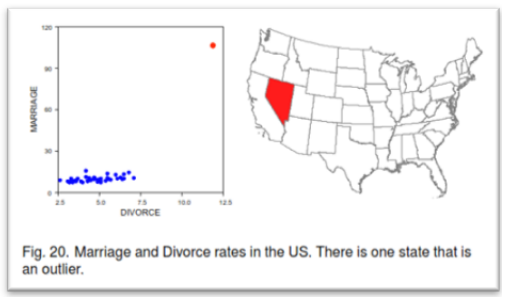


Fig. 20. Marriage and Divorce rates in the US. There is one state that is an outlier.

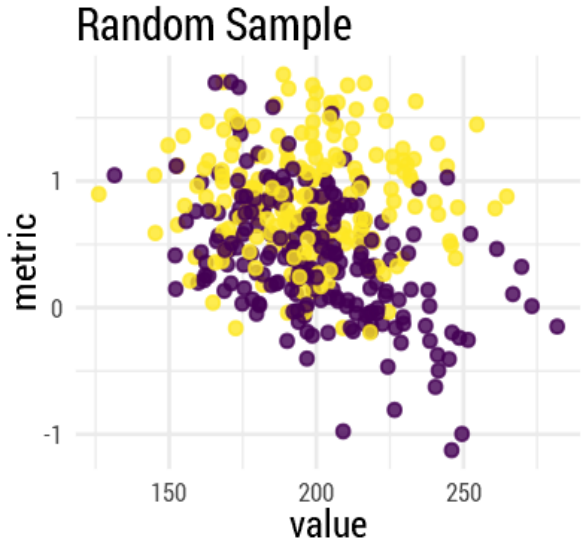
- It is often the case that researchers have developed methods to deal with a problem.
- It may not be cost-effective for mainstream or special purposes tools (e.g. Excel, historians, Tableau) to implement novel methods.
- Analysts may want to consider getting the help of a Data Scientist or learning to use tools such as R or Python where these methods are usually implemented as libraries/packages

[1] *Wilkinson, L. (2018). Visualizing big data outliers through distributed aggregation. IEEE Transactions on Visualization and Computer Graphics, 24(1), 256-266*

Visualizing High Density Data

Best...

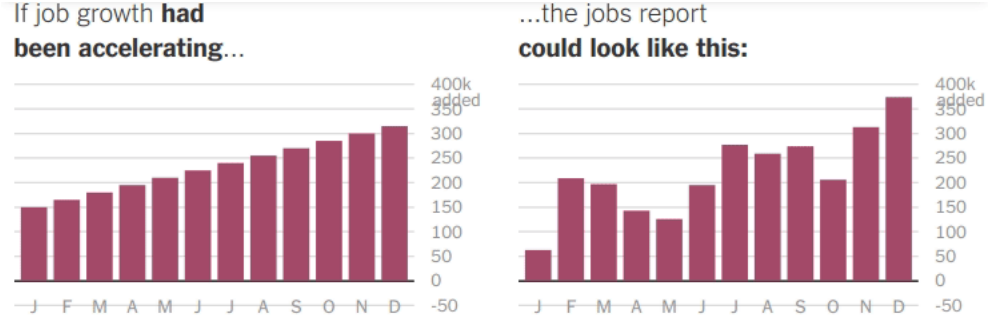
Critically assess whether you need BIG data in the first place: unless outliers are the main focus of analysis, it's often satisfactory to **sample** the data.



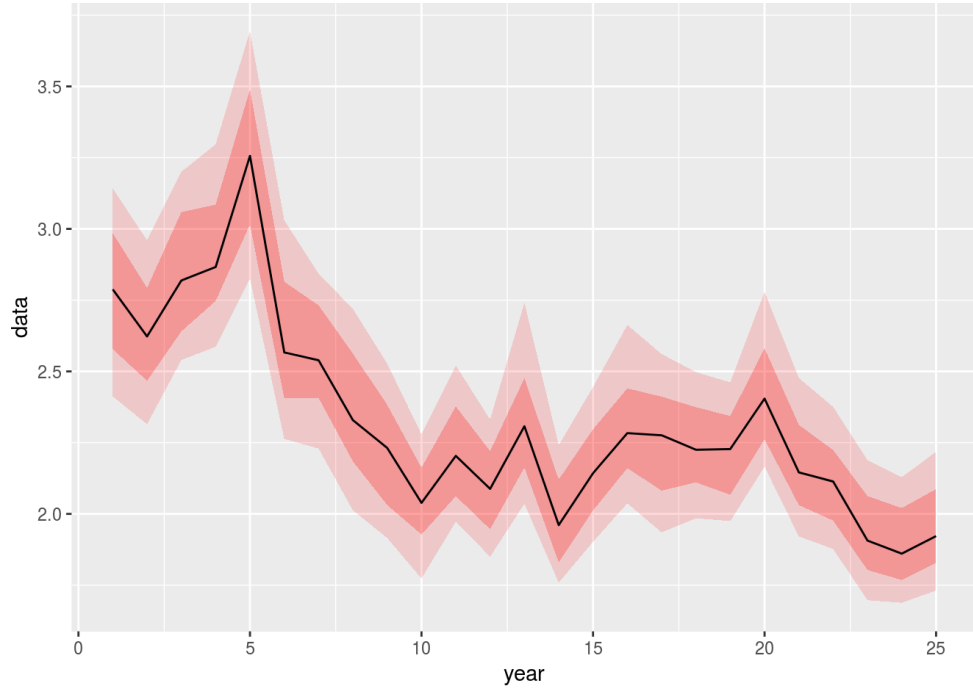
[1] *If you are unconvinced, ask yourself: do I really have the whole population despite a huge dataset? Big data provides no protection against bias*

Show the Uncertainty

We need to visualize the uncertainty in order to make optimal decisions.



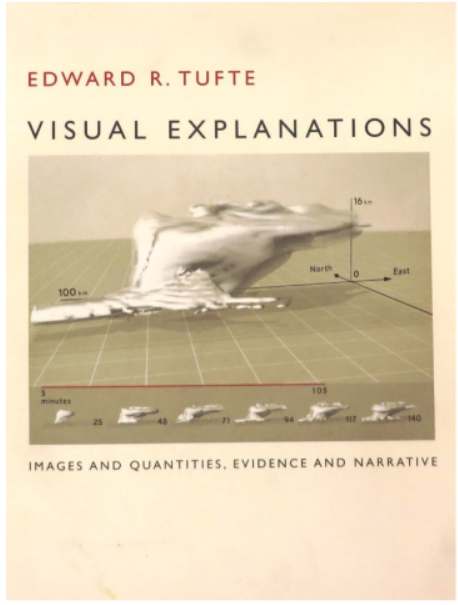
If you squint, you can see evidence of the upward trend, with stronger growth apparent in the second half of the year. But month by month, you wouldn't have any way of knowing if it was a true acceleration, or just a false signal generated by sampling error.



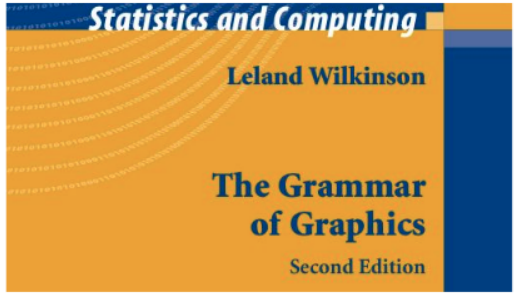
[1] *New York Times. How Not to Be Misled by the Jobs Report, 2014*
[2] *Chato Osio and Hamon. 2017*

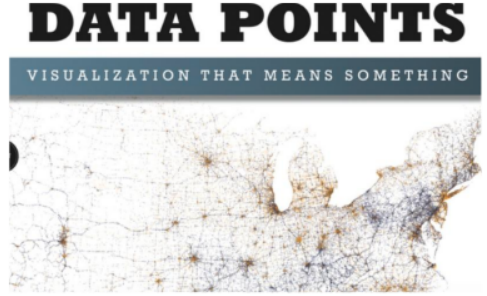
> *An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem*
>
> -John Tukey

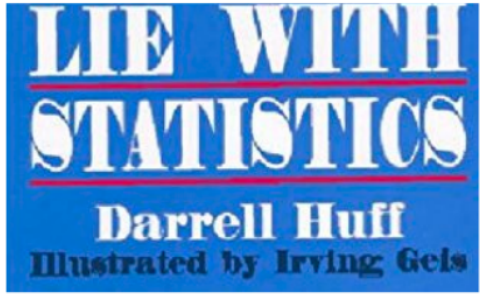
8. References

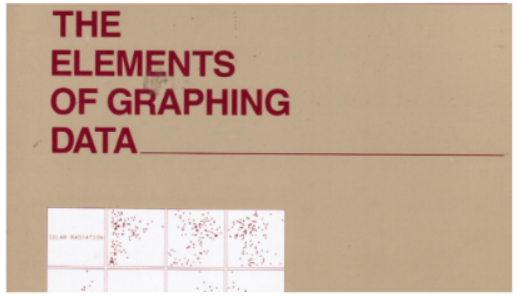

EDWARD R. TUFTE
VISUAL EXPLANATIONS
IMAGES AND QUANTITIES, EVIDENCE AND NARRATIVE

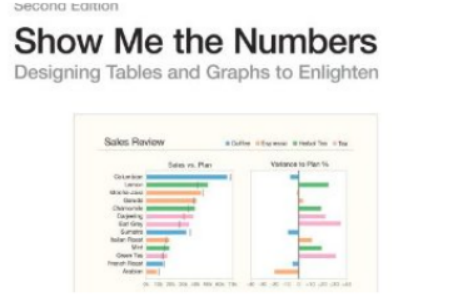

EDWARD TUFTE
BEAUTIFUL EVIDENCE

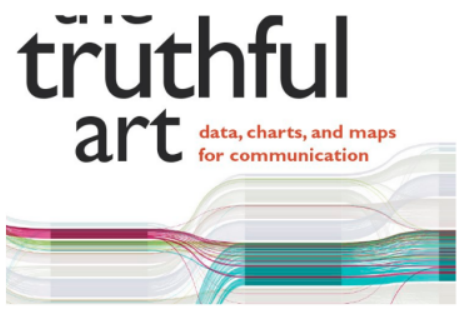

Statistics and Computing
Leland Wilkinson
The Grammar of Graphics
Second Edition

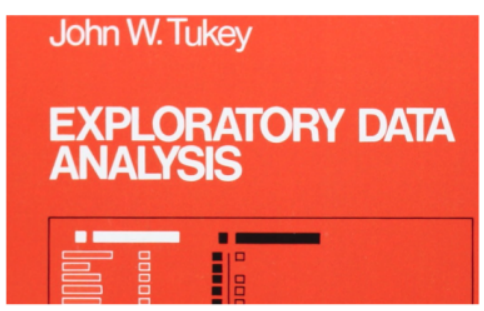

DATA POINTS
VISUALIZATION THAT MEANS SOMETHING

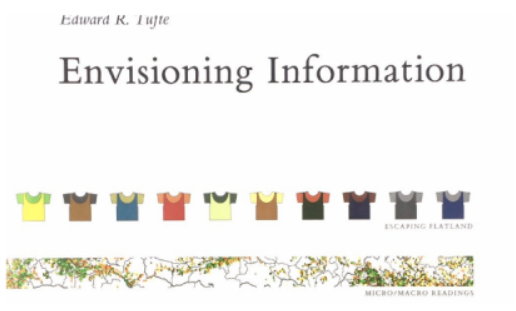

LIE WITH STATISTICS
Darrell Huff
Illustrated by Irving Geis

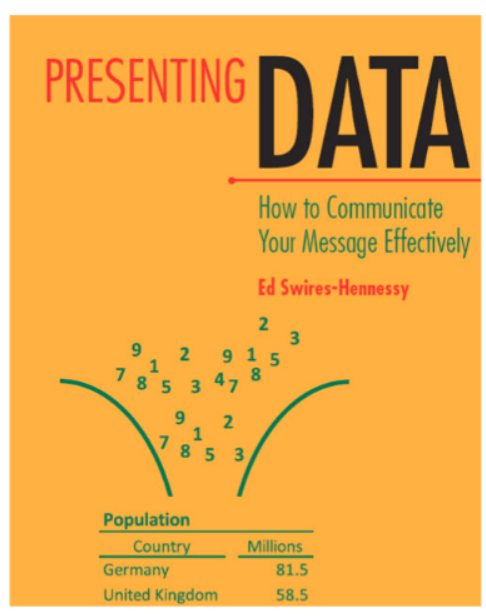

THE ELEMENTS OF GRAPHING DATA

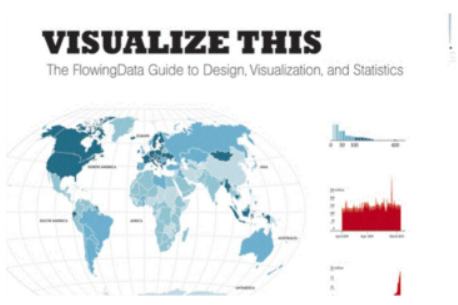

Second Edition
Show Me the Numbers
Designing Tables and Graphs to Enlighten

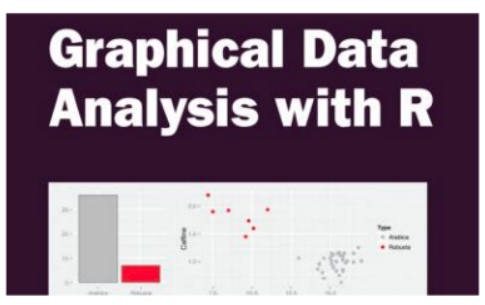

truthful art
data, charts, and maps for communication

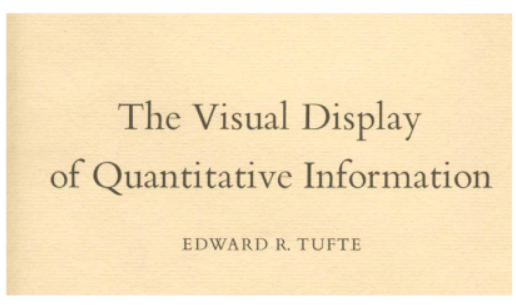

John W. Tukey
EXPLORATORY DATA ANALYSIS


Edward R. Tufte
Envisioning Information


PRESENTING DATA
How to Communicate Your Message Effectively
Ed Swires-Hennessy
Population
Country
Millions
Germany
81.5
United Kingdom
58.5


VISUALIZE THIS
The FlowingData Guide to Design, Visualization, and Statistics


Graphical Data Analysis with R


The Visual Display of Quantitative Information
EDWARD R. TUFTE

Author	Title
Tufte, E. R. (1983)	The Visual Display Of Quantitative Information
Tufte, E. R. (1990)	Envisioning Information
Tukey, J. W. (1977)	Exploratory Data Analysis
Few, S. (2012)	Show Me The Numbers. Designing Tables And Graph To Enlighten.
Chiasson T., Gregory D. et al. (2014)	Data + Design
Ehrenberg, A.S.C. (1981)	The Problem Of Numeracy. The American Statistician, Vol. 35, N. 2
Wilkinson, L. (2005)	The Grammar Of Graphics
Simon, H.A. (1996)	The Sciences Of The Artificial.
Cleveland, W.S. (1985)	The Elements Of Graphing Data.
Yau, N. (2011)	Visualize This. The Flowingdata Guide To Design, Visualization And Statistics
Good P.I., Hardin J.W. (2012)	Common Errors In Statistics (And How To Avoid Them)
Gelman A., Unwin A. (2012)	Infovis And Statistical Graphics: Different Goals, Different Looks

Thank you

Thomas Speidel
thomas@speidel.ca
ca.linkedin.com/in/speidel/
alternative-stats.netlify.com

This presentation and most of the graphs were produced in R, a programming language and software environment for statistical computing and graphics. The Xaringan package was used for as the presentation template.