Data Visualization for Data Science Practitioners

Olaf Menzer

Senior Data Scientist

[1] Pacific Life, Newport Beach
[2] University of California, Santa Barbara

Advanced Analytics Meetup, Irvine, April 17th 2019

Bio

2006: Research Assistant at Max Planck Institute in Jena, Germany… started working with climate data and maps

2011: M.Sc. in Bioinformatics in Jena, Germany

2011: Moved to California, UC Santa Barbara; Research in Ecosystem Science, Statistics

2015: Ph.D. in Geographic Information Science, UCSB

2015-2018: Predictive Analyst, Ingram Micro, Irvine

Since 2018: Data Scientist, Pacific Life, Newport Beach

Jena, Germany

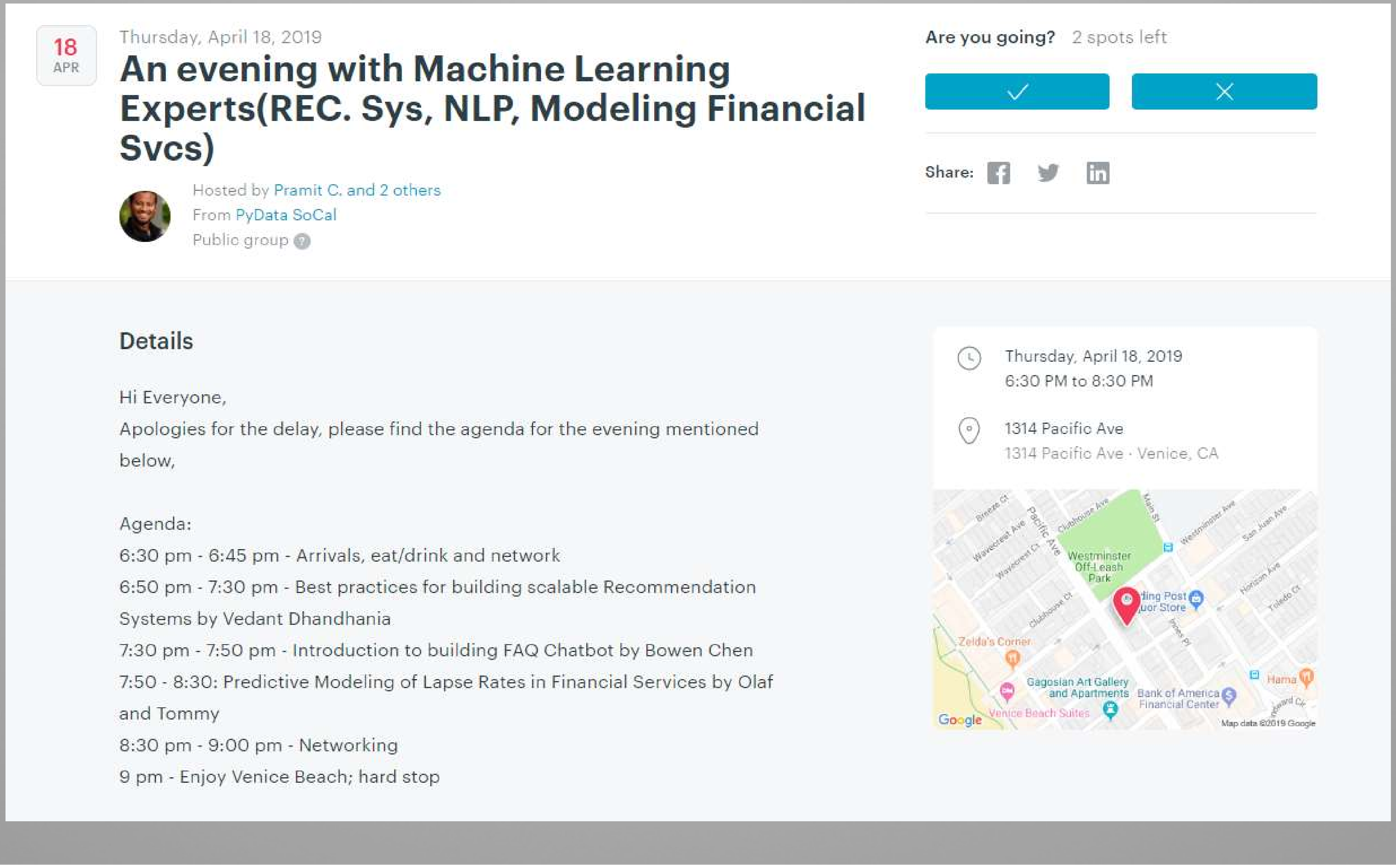
UC Santa Barbara

Pacific Life HQ

Quick Commercial

PyData Socal Meetup in Venice, tomorrow Thursday, 6.30pm

https://www.meetup.com/PyData-SoCal/events/259853926/



18
APR

Thursday, April 18, 2019

An evening with Machine Learning Experts(REC. Sys, NLP, Modeling Financial Svcs)

Hosted by Pramit C. and 2 others
From PyData SoCal
Public group ?

Are you going? 2 spots left

✓ ✕

Share:

Details

Hi Everyone,

Apologies for the delay, please find the agenda for the evening mentioned below,

Agenda:
6:30 pm - 6:45 pm - Arrivals, eat/drink and network
6:50 pm - 7:30 pm - Best practices for building scalable Recommendation Systems by Vedant Dhandhania
7:30 pm - 7:50 pm - Introduction to building FAQ Chatbot by Bowen Chen
7:50 - 8:30: Predictive Modeling of Lapse Rates in Financial Services by Olaf and Tommy
8:30 pm - 9:00 pm - Networking
9 pm - Enjoy Venice Beach; hard stop

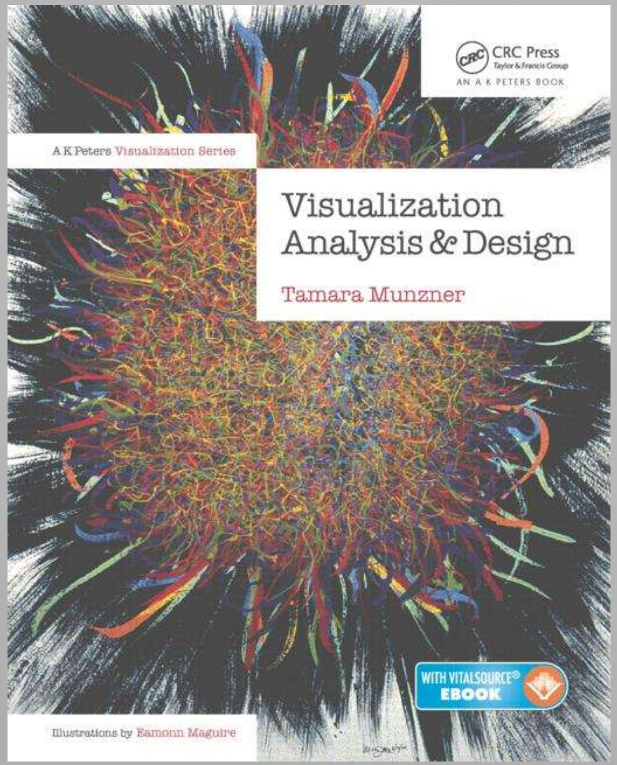
Thursday, April 18, 2019
6:30 PM to 8:30 PM

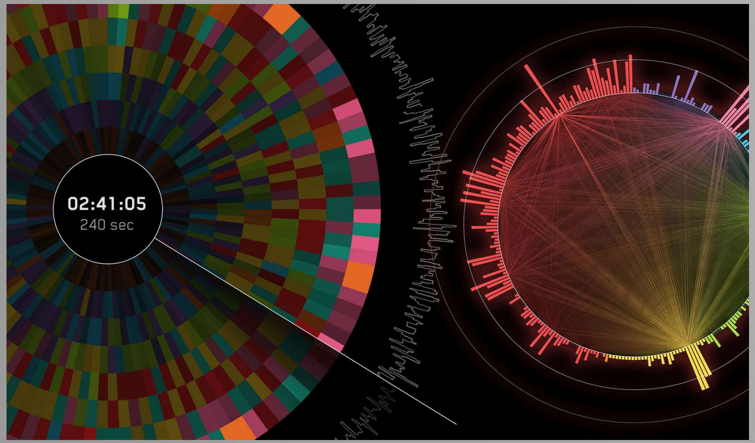
1314 Pacific Ave
1314 Pacific Ave · Venice, CA

Westminster Off-Leash Park
Zelda's Corner
Gagosian Art Gallery and Apartments
Bank of America Financial Center
Venice Beach Suites
Google
Map data ©2019 Google

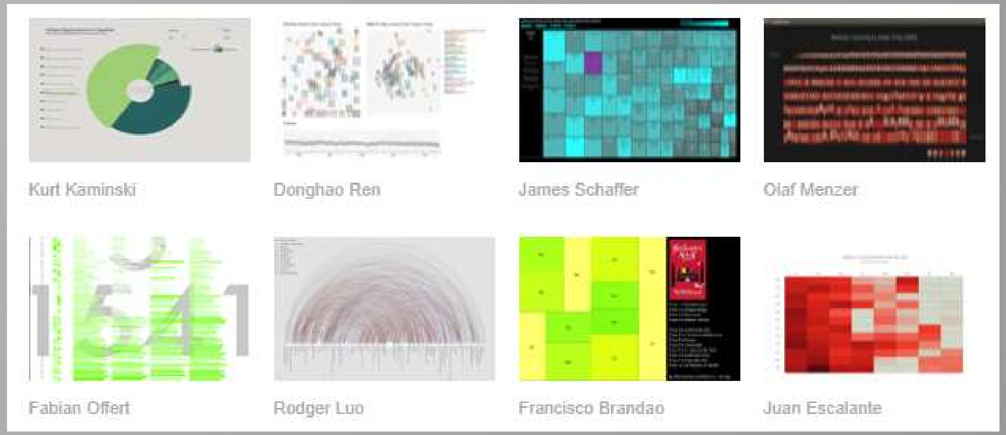
Data Visualization Credits

Tamara Munzner,
Visualization Analysis & Design, 2014


CRC Press
Taylor & Francis Group
AN A K PETERS BOOK
A K Peters Visualization Series
Visualization Analysis & Design
Tamara Munzner
WITH VITALSOURCE® EBOOK
Illustrations by Eamonn Maguire



DataVis @CalTech

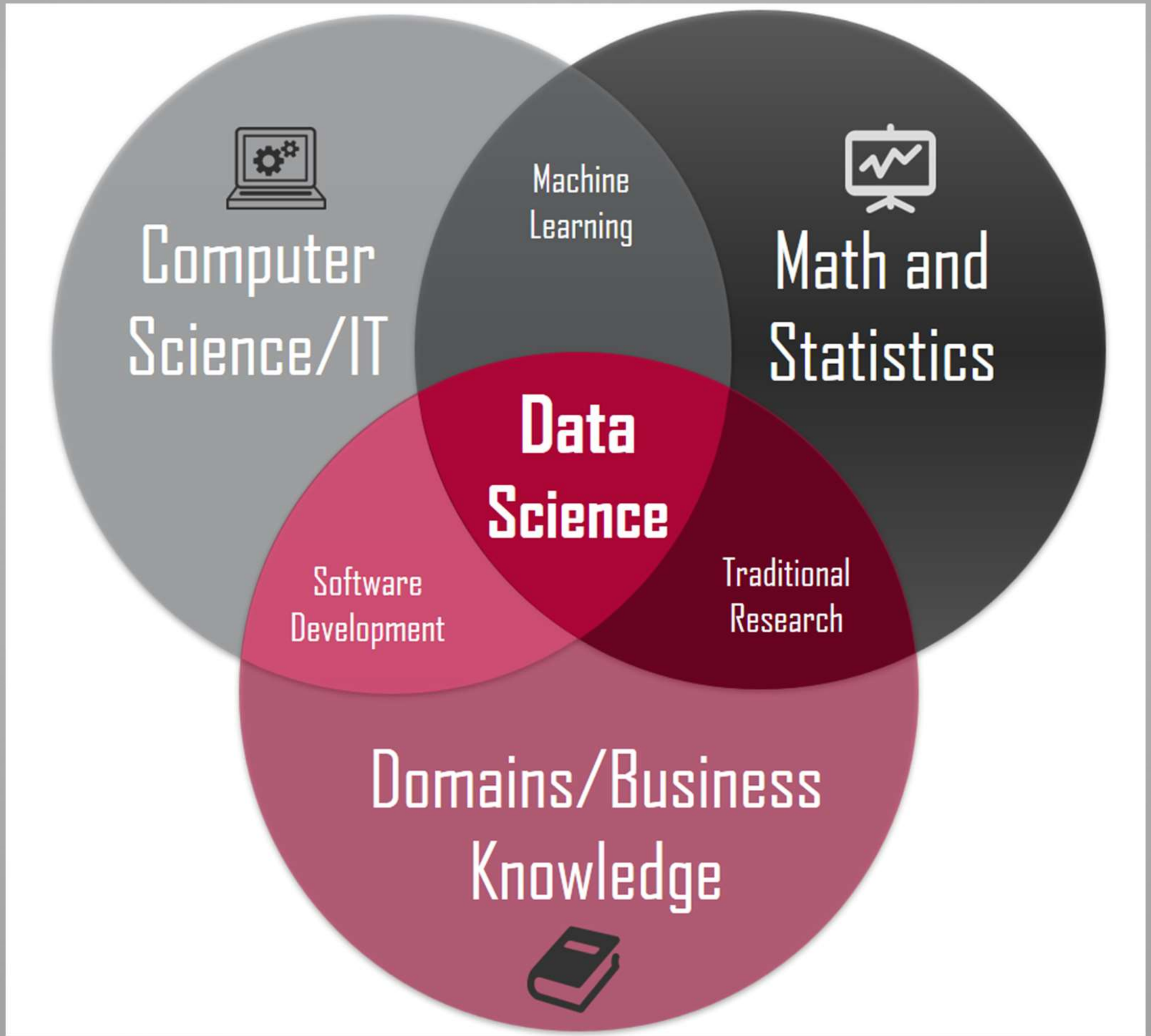


VisLab @UCSB

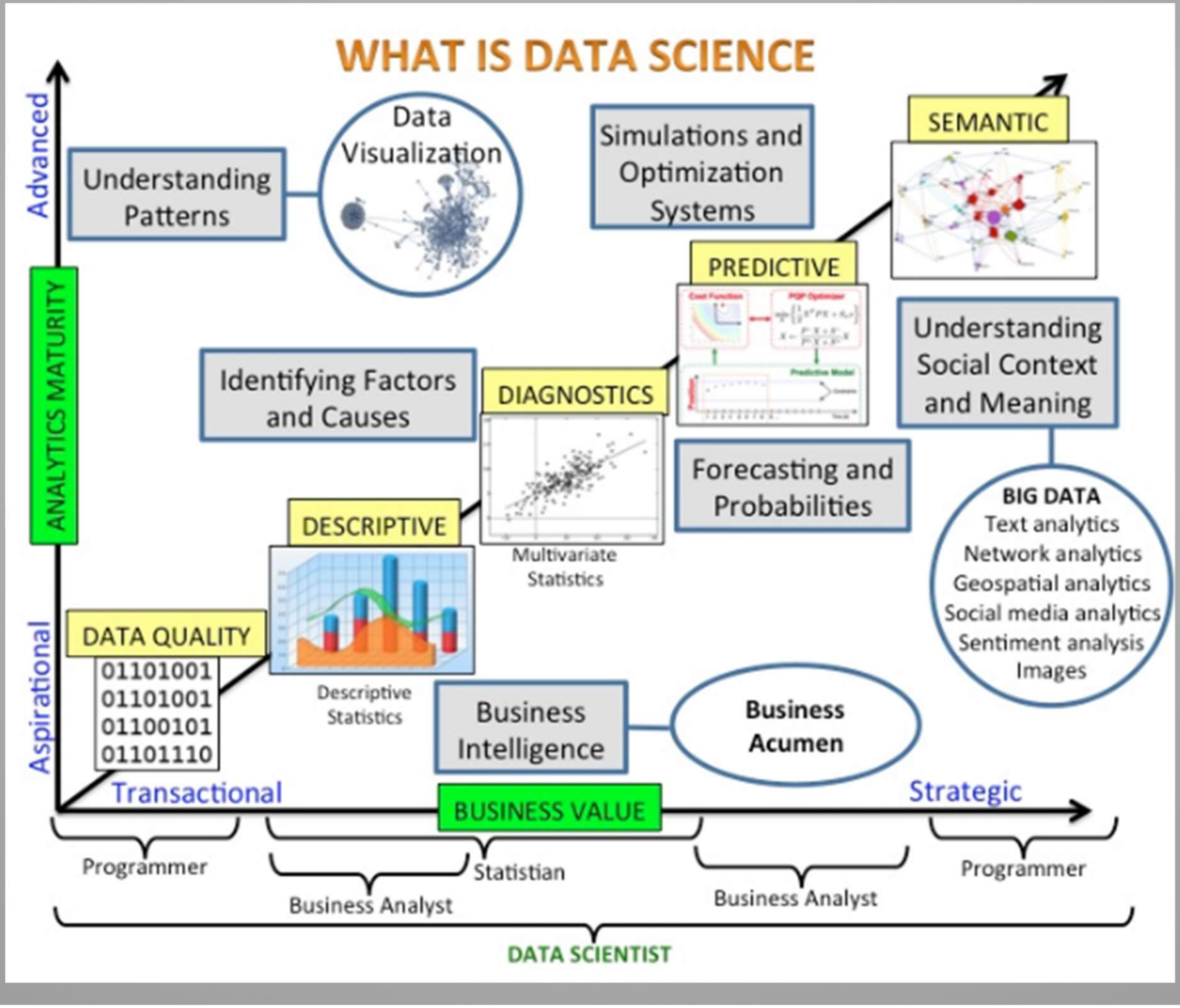
1) Data Visualization – Why do we care?
2) Data Visualization Elements & Tools
3) Advanced Graph Examples (with R code)
4) Emerging Data Visualizations
5) R Exercises

1) Data Visualization – Why do we care?
2) Data Visualization Elements & Tools
3) Advanced Graph Examples (with R code)
4) Emerging Data Visualizations
5) R Exercises

The Data Science “Skillgram”



Data Visualization in Data Science



Data Visualization – as a means to…

(1) The need to augment human capabilities to view & interpret data

(2) Understanding relationships between factors

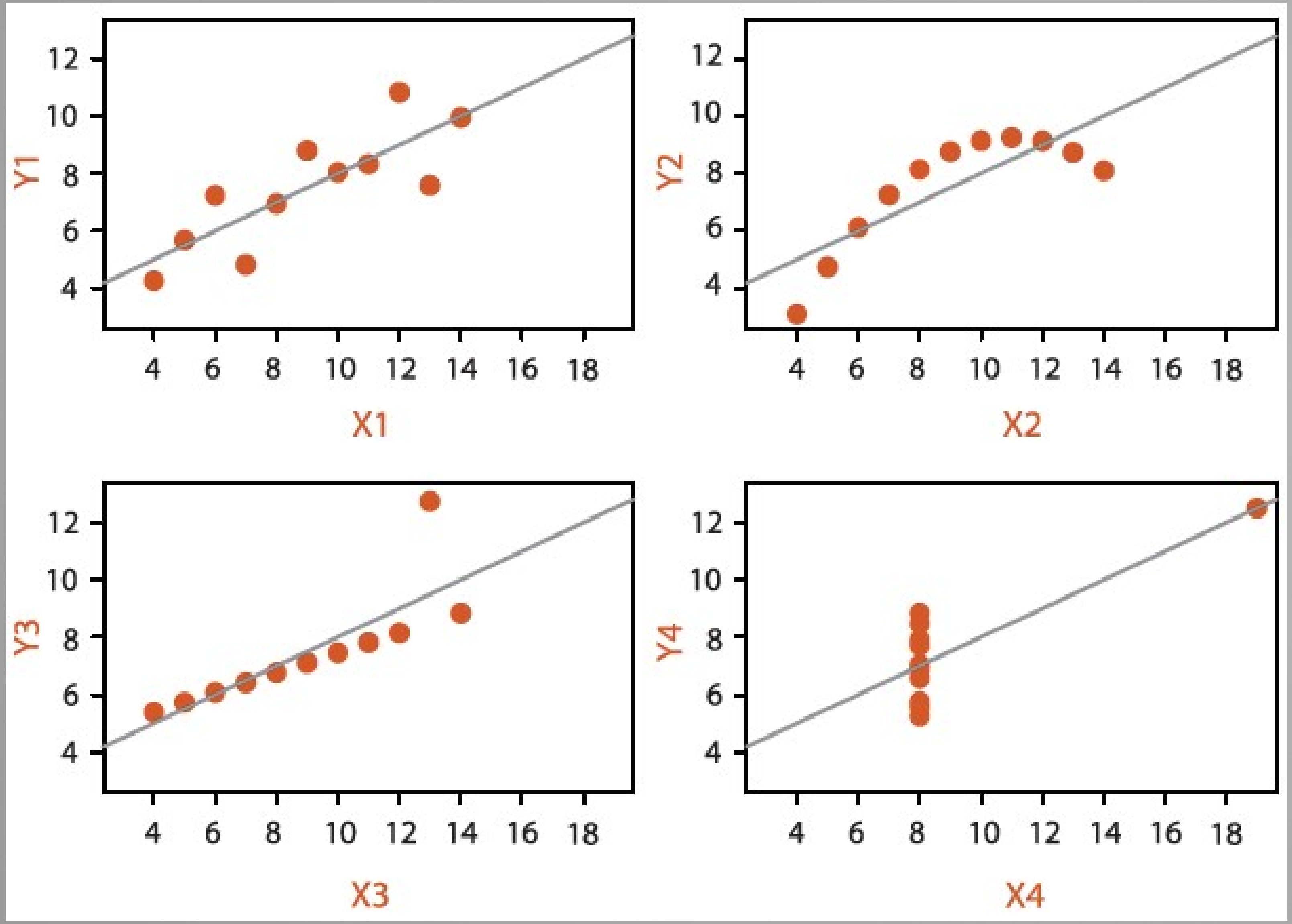
(3) Recognizing patterns

(4) Breaking down complexity / dimensionality reduction

(5) Exploratory Analysis

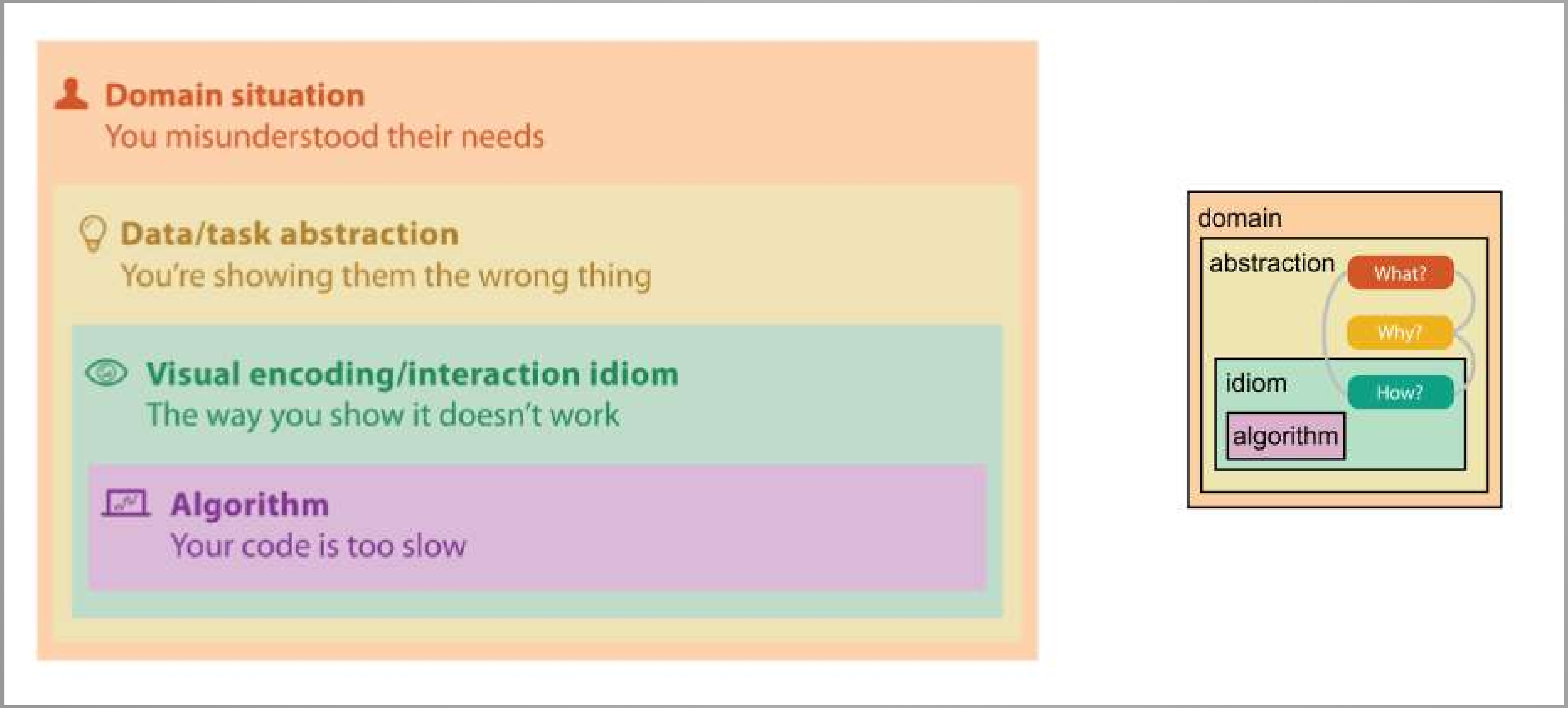
(6) Communicating results to collaborators, clients, friends, …

Anscombe’s quartet



Property	Value
Mean of x	9
Variance of x	11
Mean of y	7.5
Variance of y	4.125
Correlation	0.816
Regression Line	y = 3 + 0.5x
Coefficient of determination	0.67

Why can Data Viz be difficult?



Domain situation
You misunderstood their needs

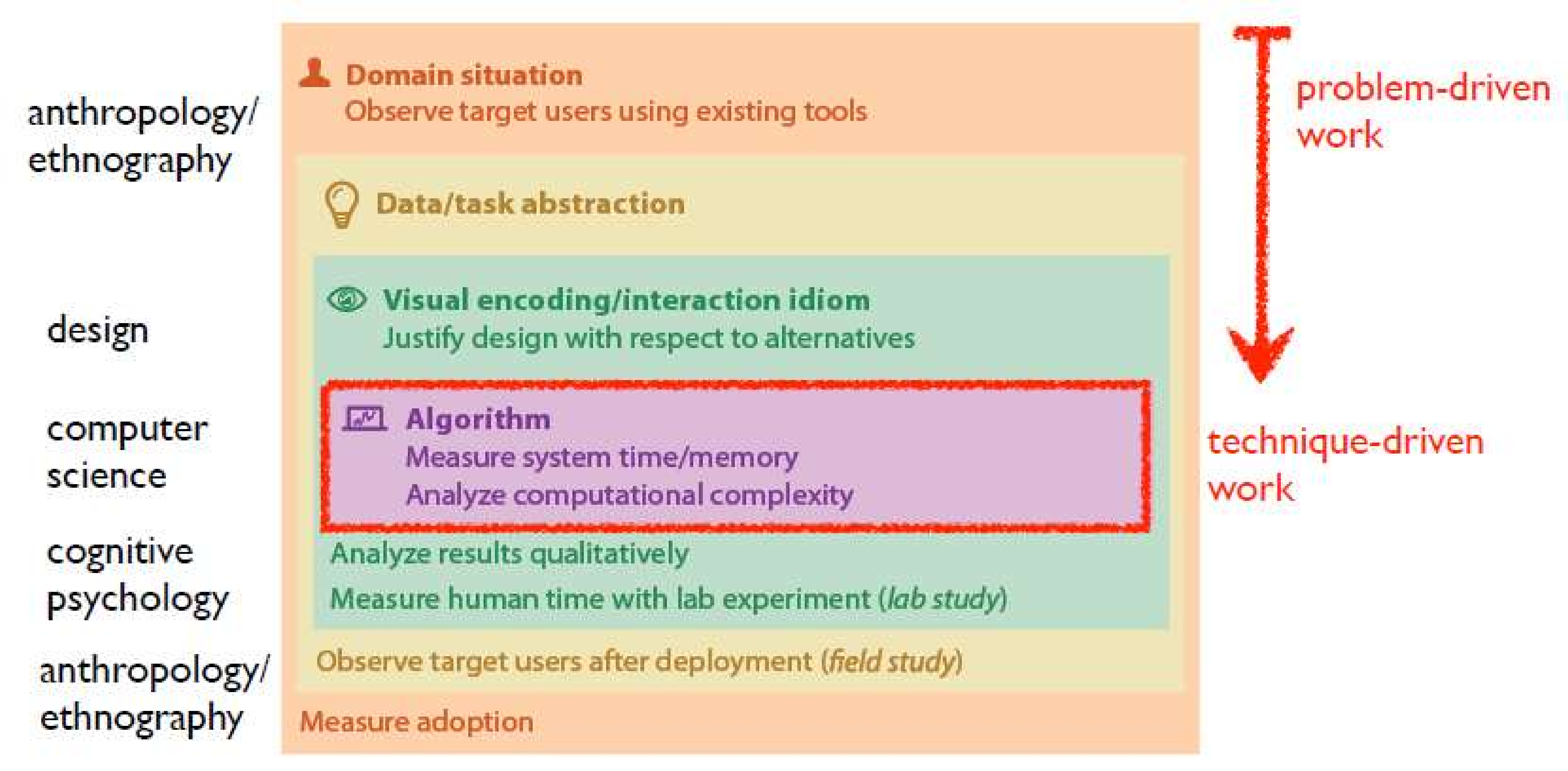
Data/task abstraction
You're showing them the wrong thing

Visual encoding/interaction idiom
The way you show it doesn't work

Algorithm
Your code is too slow

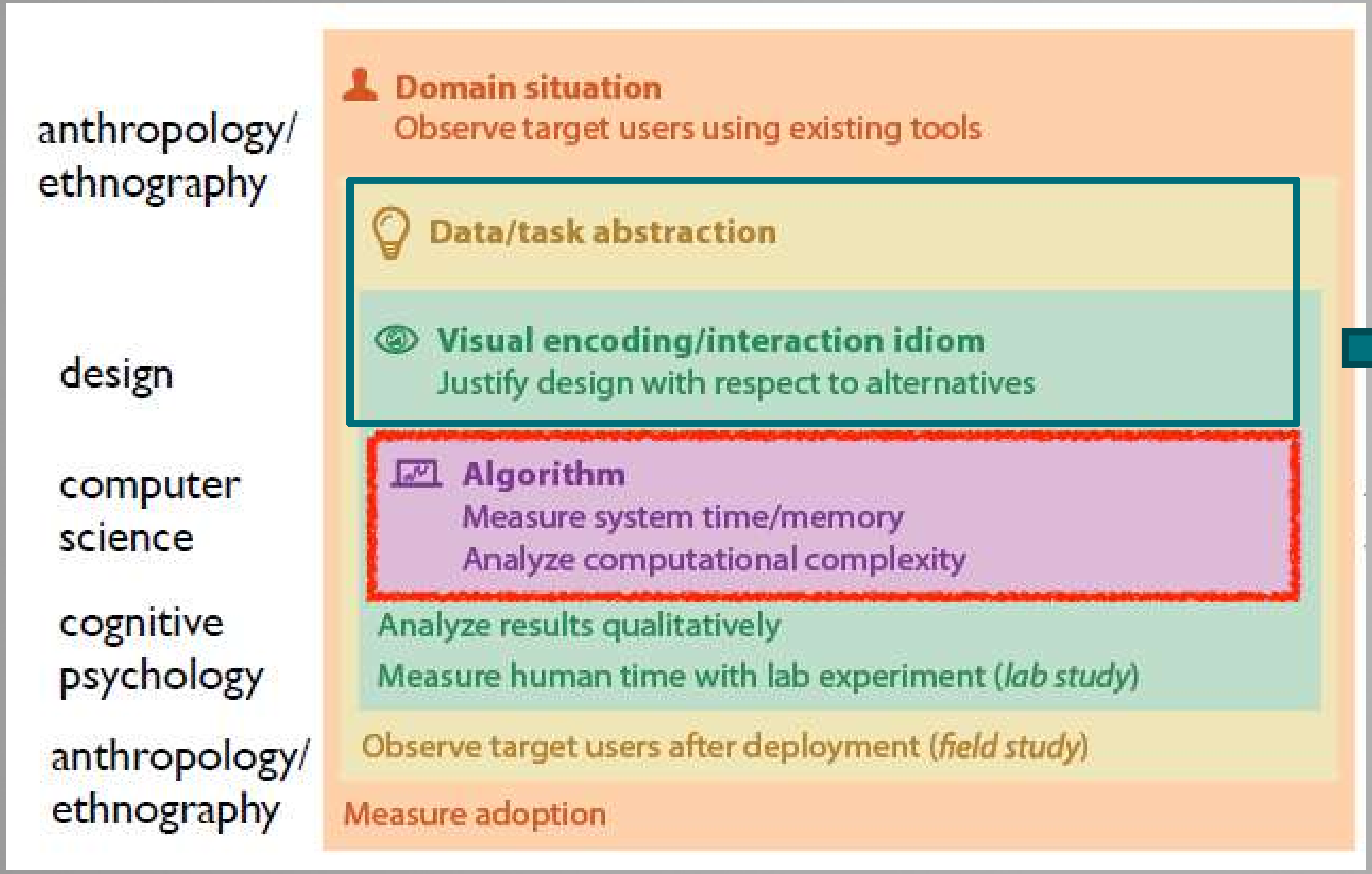
A Nested Model for Visualization Design and Validation. Tamara Munzner. IEEE TVCG (Proc. InfoVis 2009), 15(6):921-928, 2009

What makes a good graph successful?


anthropology/ ethnography
design
computer science
cognitive psychology
anthropology/ ethnography
Domain situation
Observe target users using existing tools
Data/task abstraction
Visual encoding/interaction idiom
Justify design with respect to alternatives
Algorithm
Measure system time/memory
Analyze computational complexity
Analyze results qualitatively
Measure human time with lab experiment (*lab study*)
Observe target users after deployment (*field study*)
Measure adoption
problem-driven work
technique-driven work

A Nested Model for Visualization Design and Validation. Tamara Munzner. IEEE TVCG (Proc. InfoVis 2009), 15(6):921-928, 2009

What makes a good graph successful?



anthropology/ ethnography

Domain situation
Observe target users using existing tools

Data/task abstraction

design

Visual encoding/interaction idiom
Justify design with respect to alternatives

computer science

Algorithm
Measure system time/memory
Analyze computational complexity

cognitive psychology

Analyze results qualitatively
Measure human time with lab experiment (*lab study*)

anthropology/ ethnography

Observe target users after deployment (*field study*)

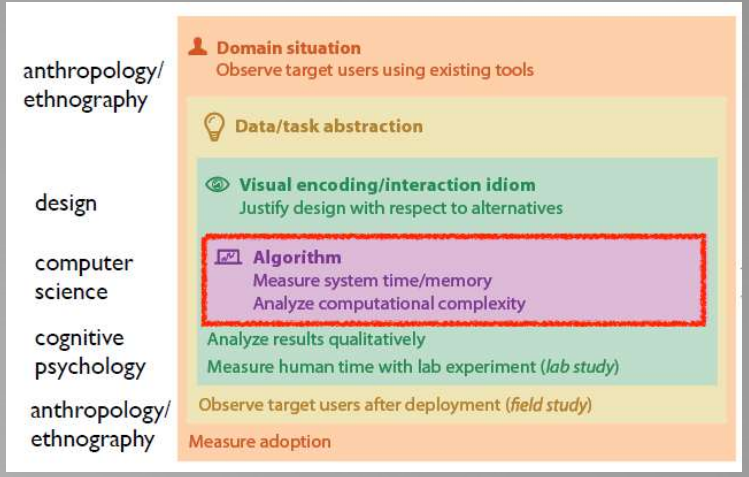
Measure adoption

Good in theory,

But… what next?

A Nested Model for Visualization Design and Validation. Tamara Munzner. IEEE TVCG (Proc. InfoVis 2009), 15(6):921-928, 2009

Practical considerations



anthropology/ ethnography

Domain situation
Observe target users using existing tools

Data/task abstraction

design

Visual encoding/interaction idiom
Justify design with respect to alternatives

computer science

Algorithm
Measure system time/memory
Analyze computational complexity

cognitive psychology

Analyze results qualitatively
Measure human time with lab experiment (*lab study*)

anthropology/ ethnography

Observe target users after deployment (*field study*)
Measure adoption

To make good graphs, practitioners are faced with:

1) Tasks that require skills in Computer Science, Design, Psychology, etc.

2) Too many choices in graph types and platforms

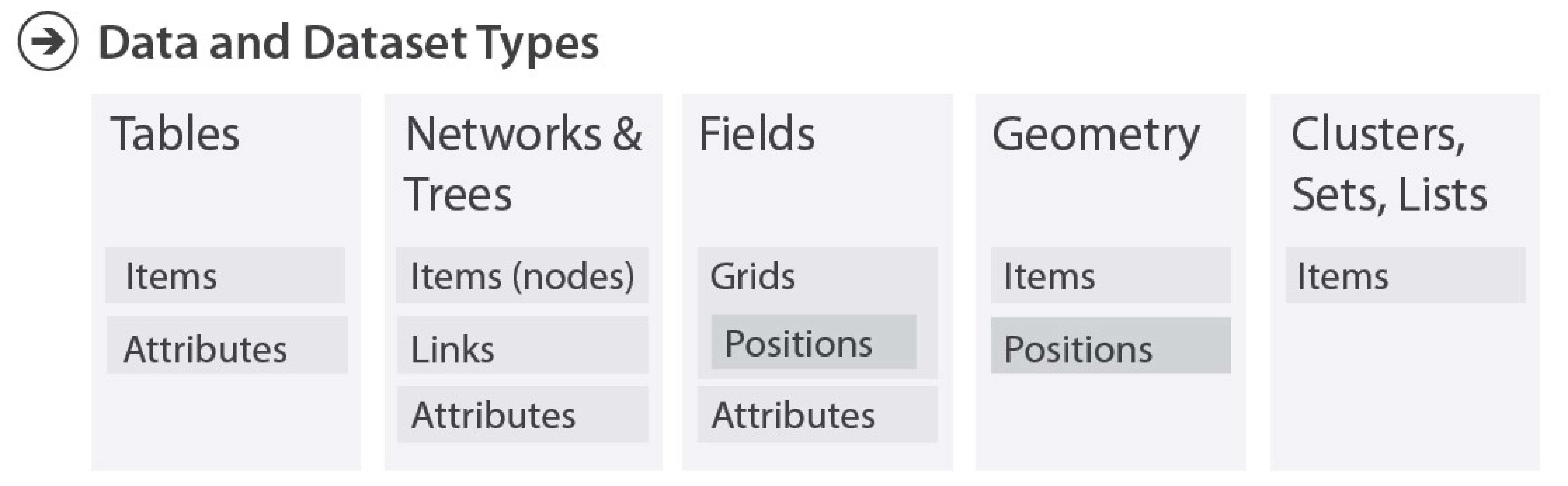
3) Searching for the **template that is most suitable** for the data

4) Having to test if the graph represents information correctly

5) Finally, **developing a narrative** / story (at the least, in form of a headline)

1) Data Visualization – Why do we care?
2) Data Visualization Elements & Tools
3) Advanced Graph Examples (with R code)
4) Emerging Data Visualizations
5) R Exercises

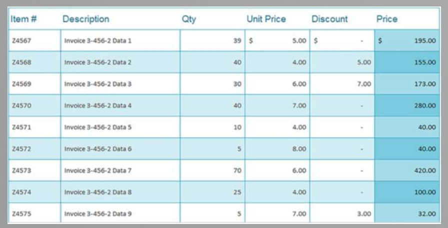
I. Data Representation and Abstraction (“*What?*”)



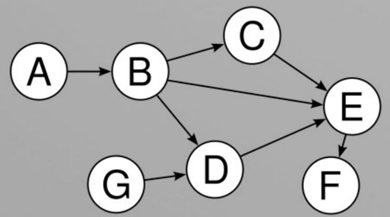
Data and Dataset Types

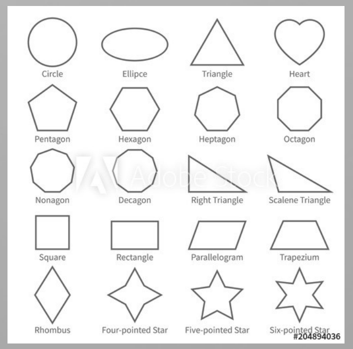
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

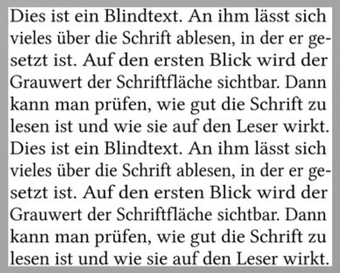
(Munzner (2018) Visualization Analysis & Design)



Item #	Description	Qty	Unit Price	Discount	Price
Z4567	Invoice 3-456-2 Data 1	39	$ 5.00	$ -	$ 195.00
Z4568	Invoice 3-456-2 Data 2	40	4.00	5.00	155.00
Z4569	Invoice 3-456-2 Data 3	30	6.00	7.00	173.00
Z4570	Invoice 3-456-2 Data 4	40	7.00	-	280.00
Z4571	Invoice 3-456-2 Data 5	10	4.00	-	40.00
Z4572	Invoice 3-456-2 Data 6	5	8.00	-	40.00
Z4573	Invoice 3-456-2 Data 7	70	6.00	-	420.00
Z4574	Invoice 3-456-2 Data 8	25	4.00	-	100.00
Z4575	Invoice 3-456-2 Data 9	5	7.00	3.00	32.00

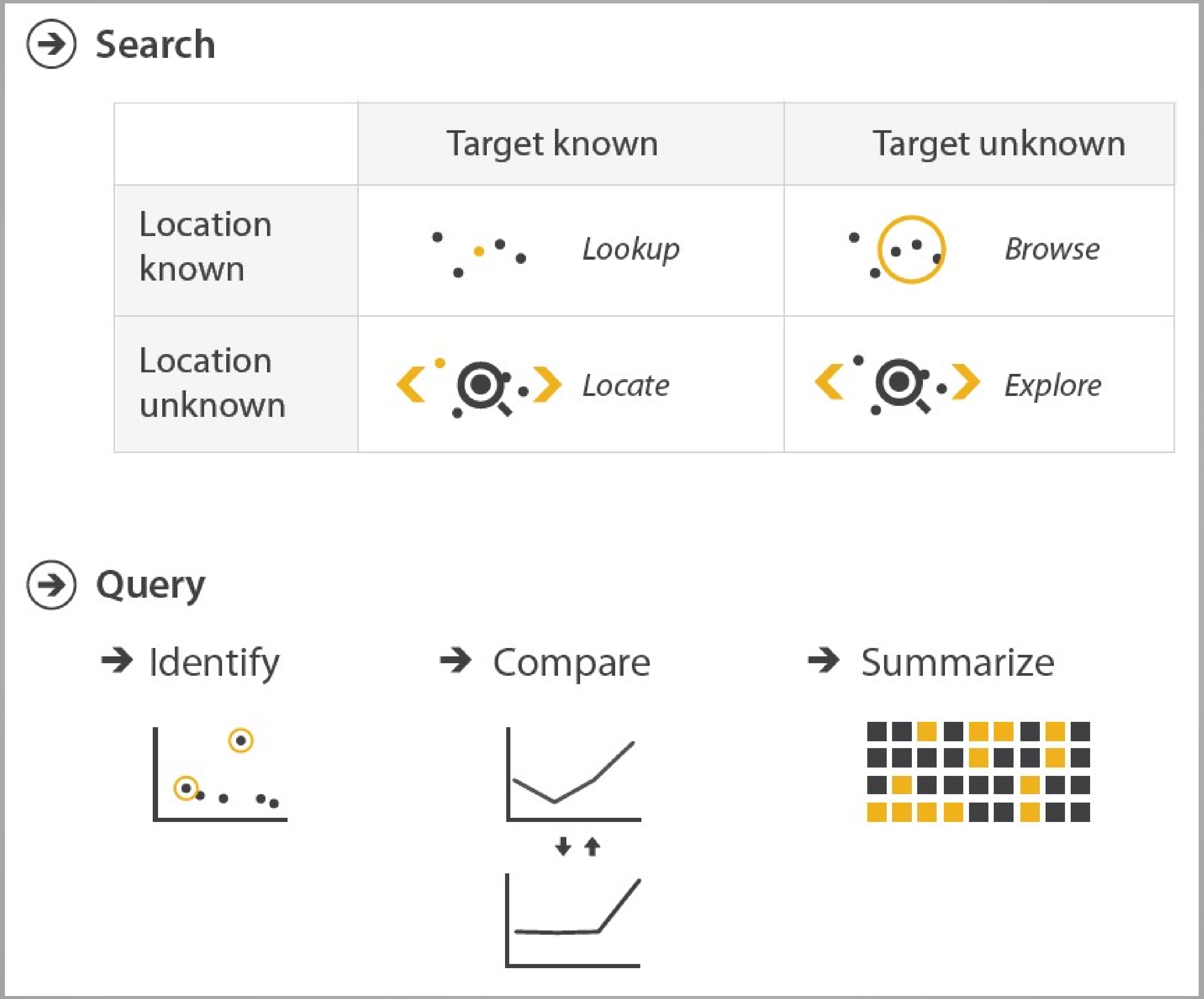






Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schriftfläche sichtbar. Dann kann man prüfen, wie gut die Schrift zu lesen ist und wie sie auf den Leser wirkt. Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schriftfläche sichtbar. Dann kann man prüfen, wie gut die Schrift zu lesen ist und wie sie auf den Leser wirkt.

II. Visualization Actions (“*Why?*”)



Search

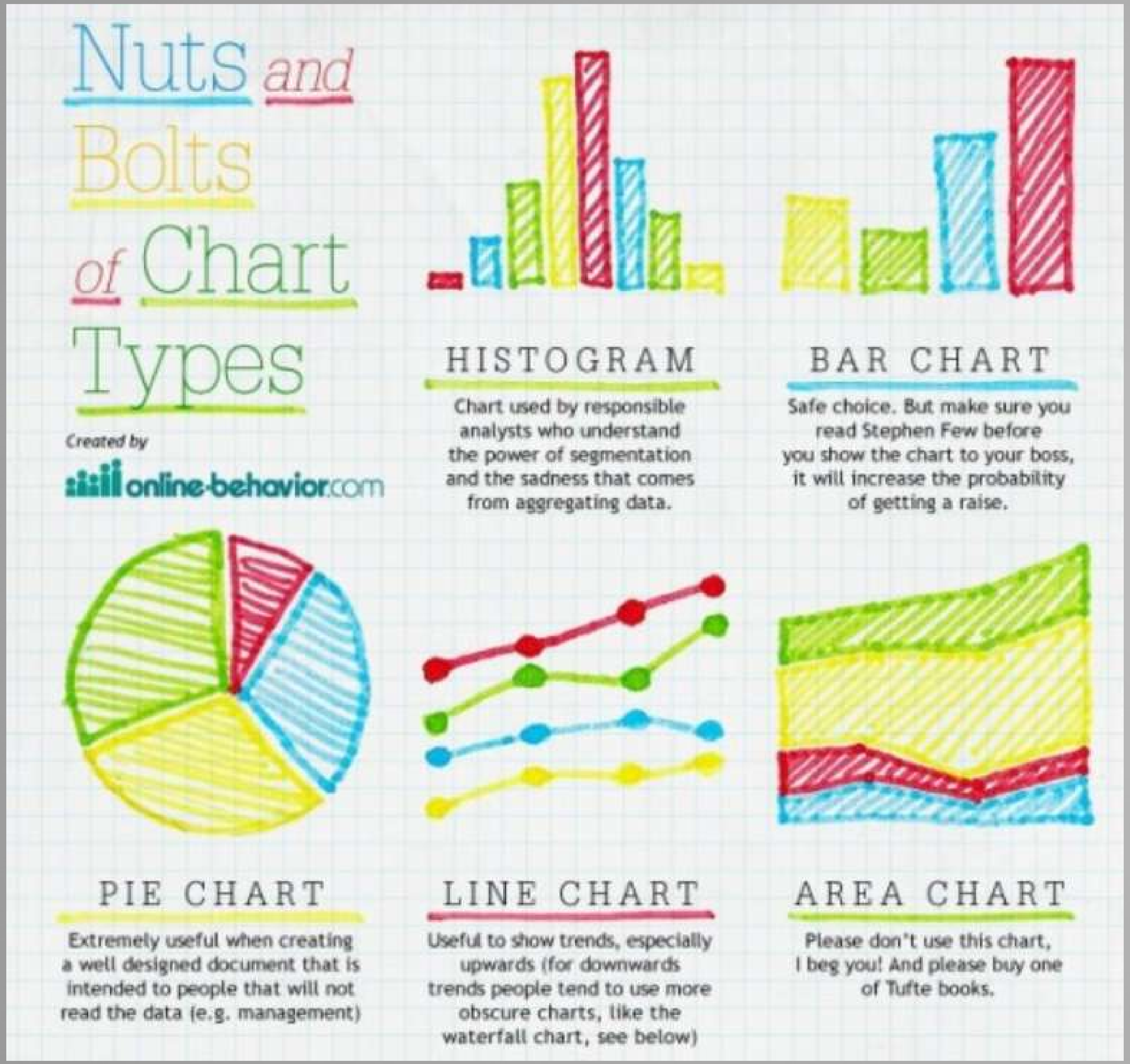
	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*

Query

- Identify
- Compare
- Summarize

(Munzner (2018),
Visualization Analysis & Design)

III. *Most* Widely Used Visualization Tools (*“How?”*)



Nuts and Bolts of Chart Types

Created by online-behavior.com

HISTOGRAM

Chart used by responsible analysts who understand the power of segmentation and the sadness that comes from aggregating data.

BAR CHART

Safe choice. But make sure you read Stephen Few before you show the chart to your boss, it will increase the probability of getting a raise.

PIE CHART

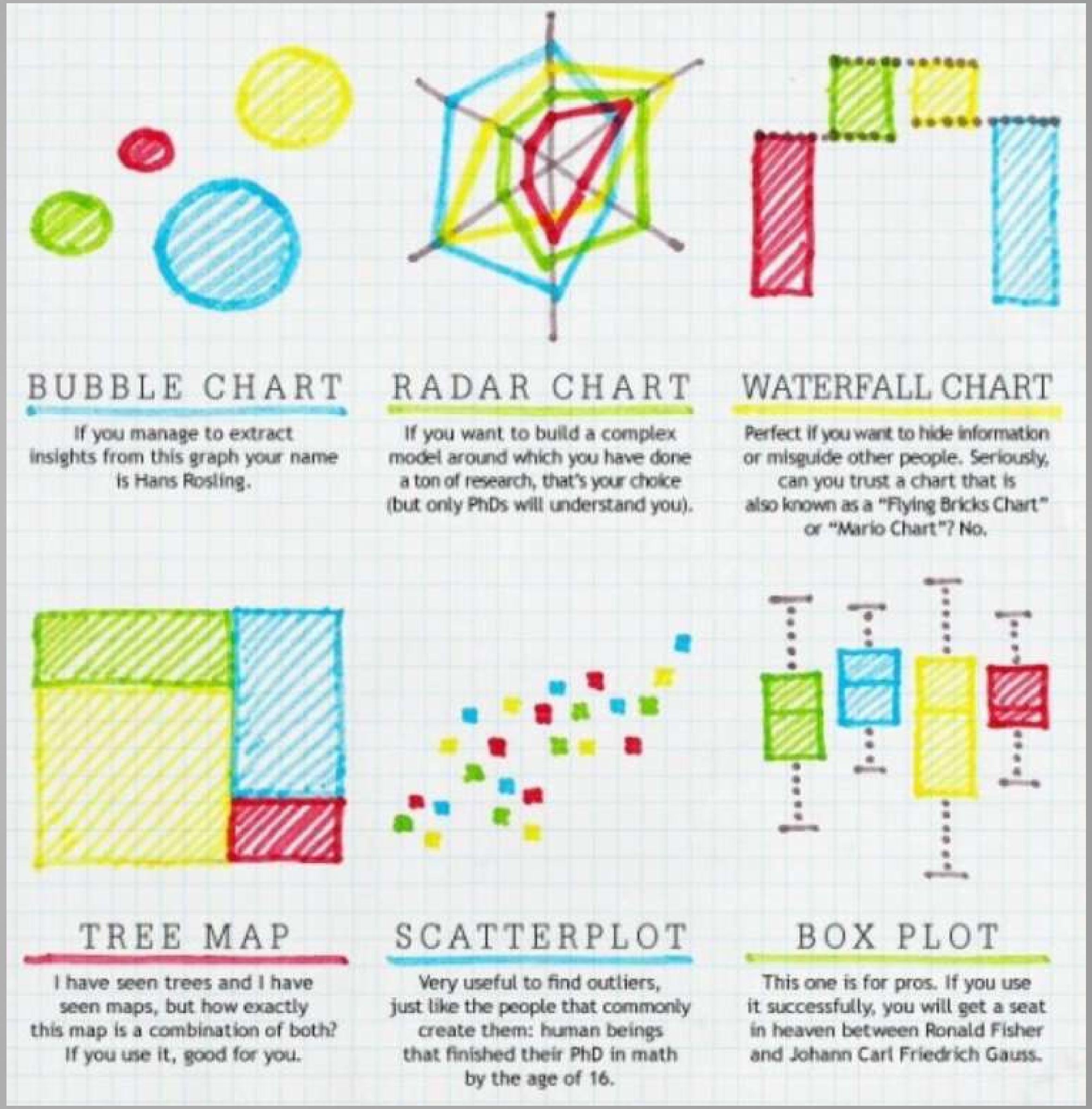
Extremely useful when creating a well designed document that is intended to people that will not read the data (e.g. management)

LINE CHART

Useful to show trends, especially upwards (for downwards trends people tend to use more obscure charts, like the waterfall chart, see below)

AREA CHART

Please don't use this chart, I beg you! And please buy one of Tufte books.



BUBBLE CHART

If you manage to extract insights from this graph your name is Hans Rosling.

RADAR CHART

If you want to build a complex model around which you have done a ton of research, that's your choice (but only PhDs will understand you).

WATERFALL CHART

Perfect if you want to hide information or misguide other people. Seriously, can you trust a chart that is also known as a “Flying Bricks Chart” or “Mario Chart”? No.

TREE MAP

I have seen trees and I have seen maps, but how exactly this map is a combination of both? If you use it, good for you.

SCATTERPLOT

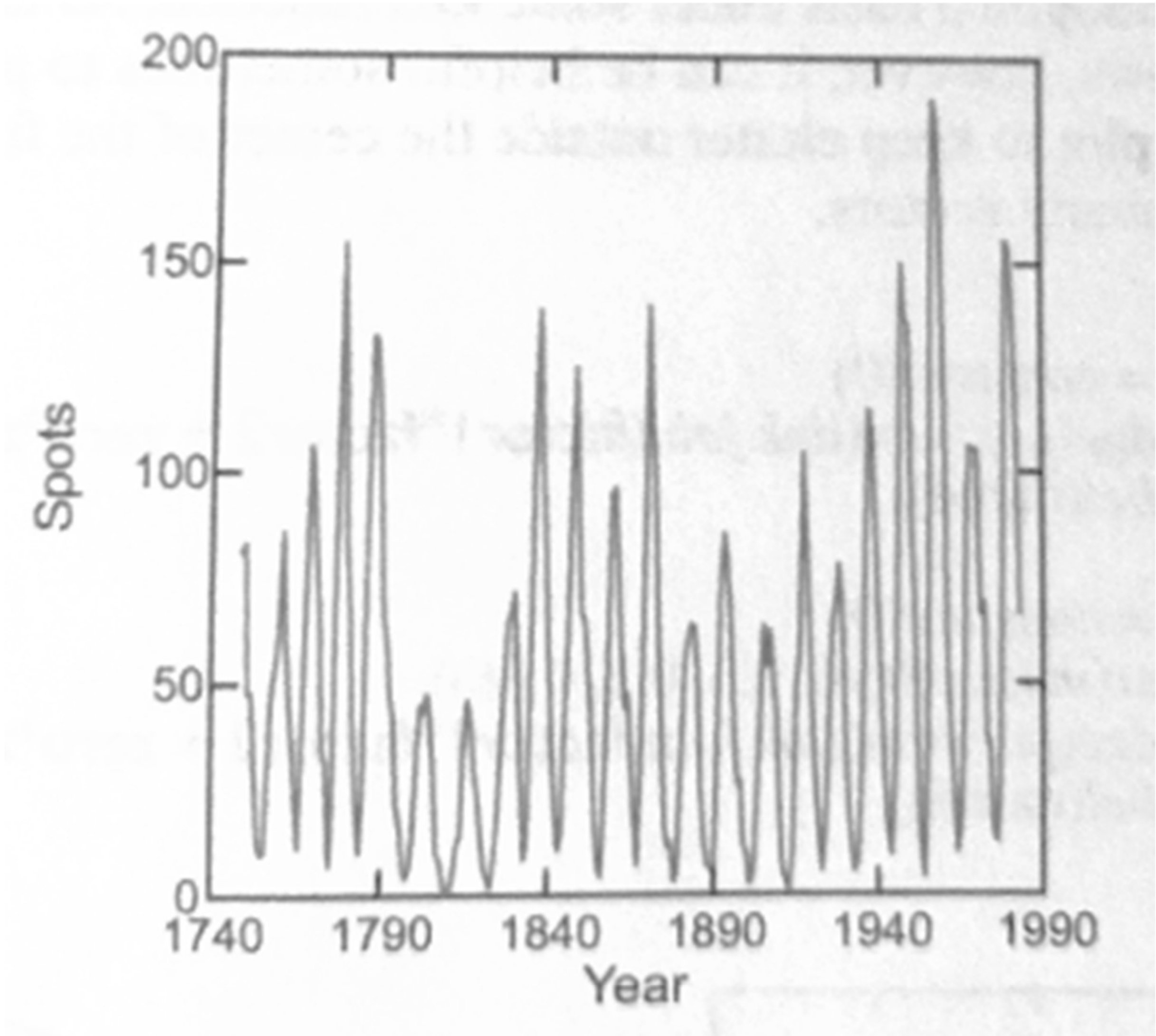
Very useful to find outliers, just like the people that commonly create them: human beings that finished their PhD in math by the age of 16.

BOX PLOT

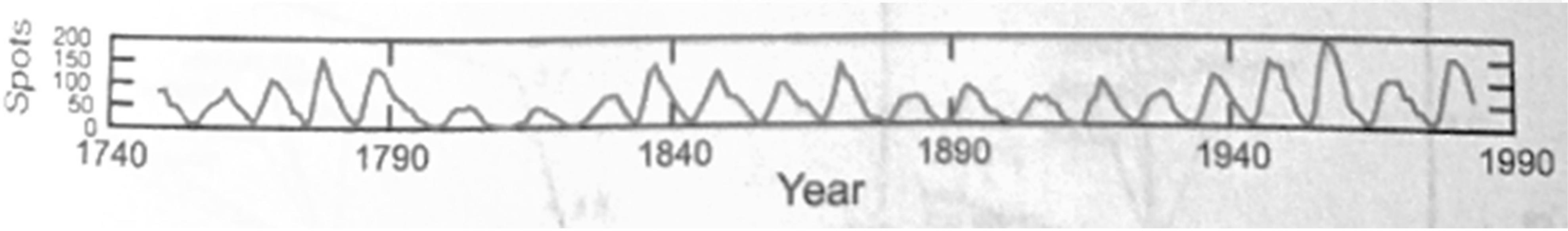
This one is for pros. If you use it successfully, you will get a seat in heaven between Ronald Fisher and Johann Carl Friedrich Gauss.

(visual.ly; online-behavior.com)

Time Series example #1: 250 years of Sunspots data

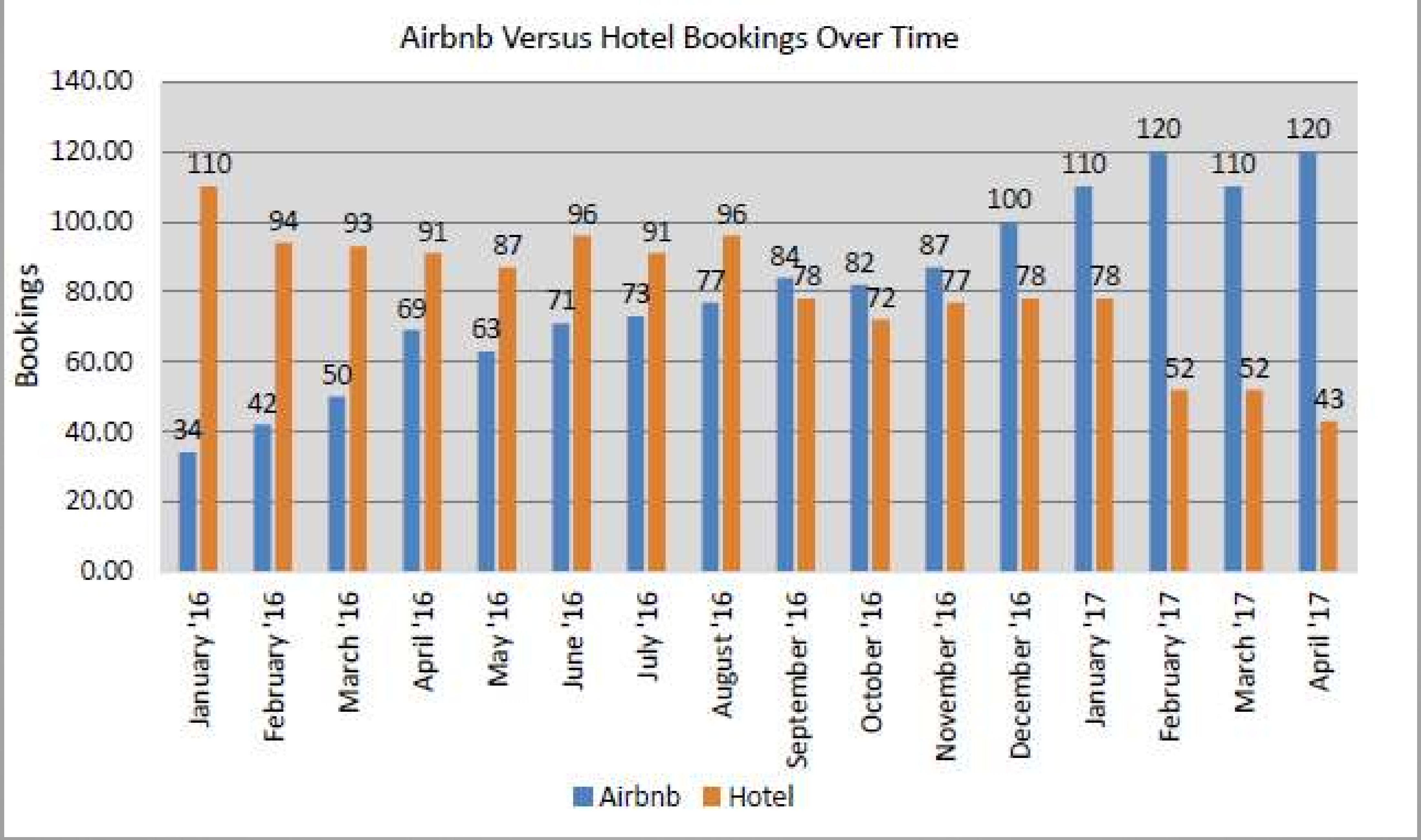


(Wilkinson, 2005,
“The Grammar of Graphics)



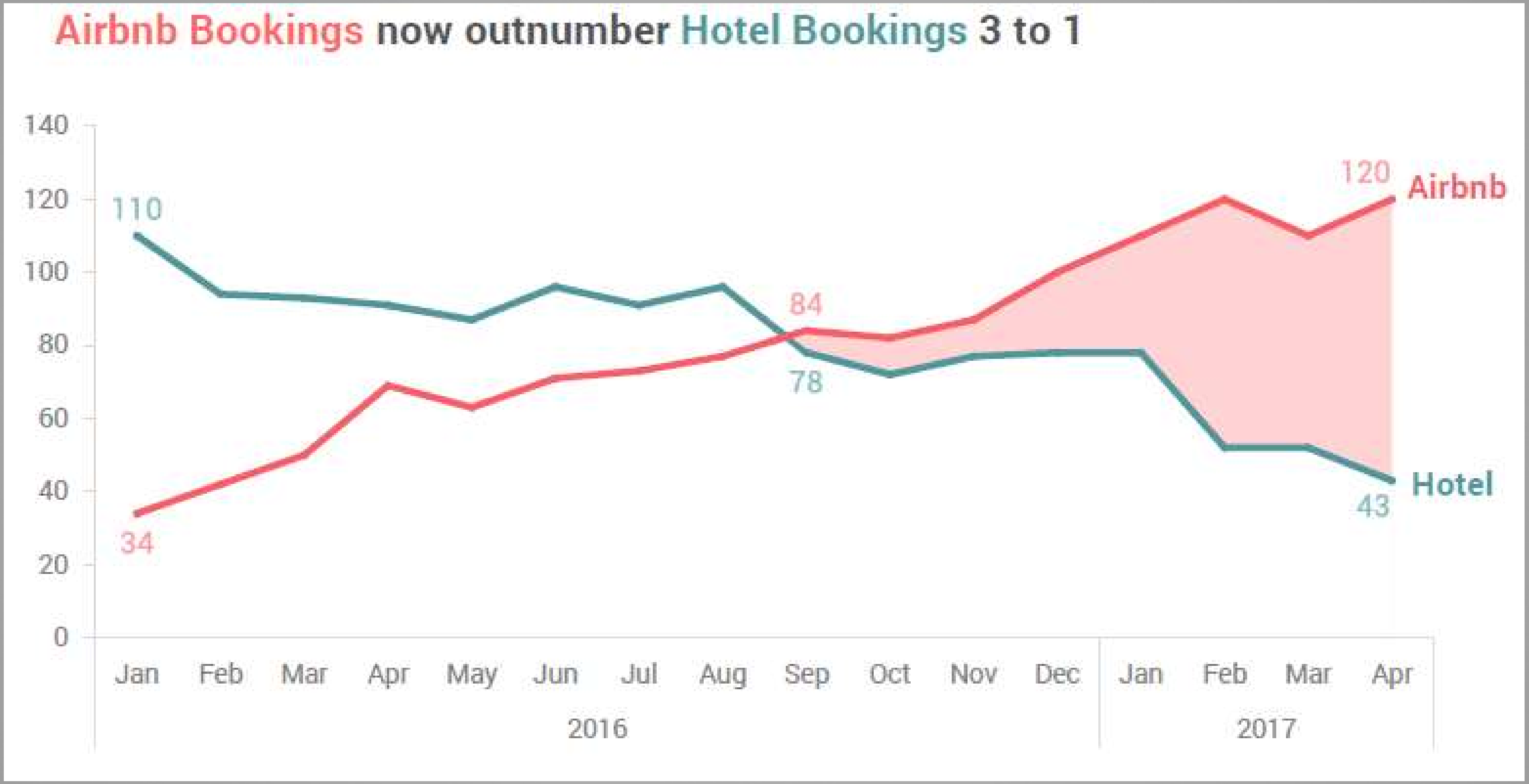
“Stretch transformation” shows periodicity of 11 years much more clearly

III. How to improve your Graph?



Isaac Reyes, DataSeer, as seen at ODSC West in San Francisco 2018

III. How to improve your Graph?



Isaac Reyes, DataSeer, as seen at ODSC West in San Francisco 2018

Visual Design Aspects (~Tufte)

Graphical displays must:

1) Show the data
2) Present many numbers in a small space (cf. data – ink ratio)
3) Make large data sets coherent
4) Encourage the eye to compare different pieces of data
5) Reveal the data at several levels of detail
6) Serve a clear purpose: description, exploration, tabulation or decoration
7) Be closely integrated with statistical and verbal descriptions of a data set.
8) …

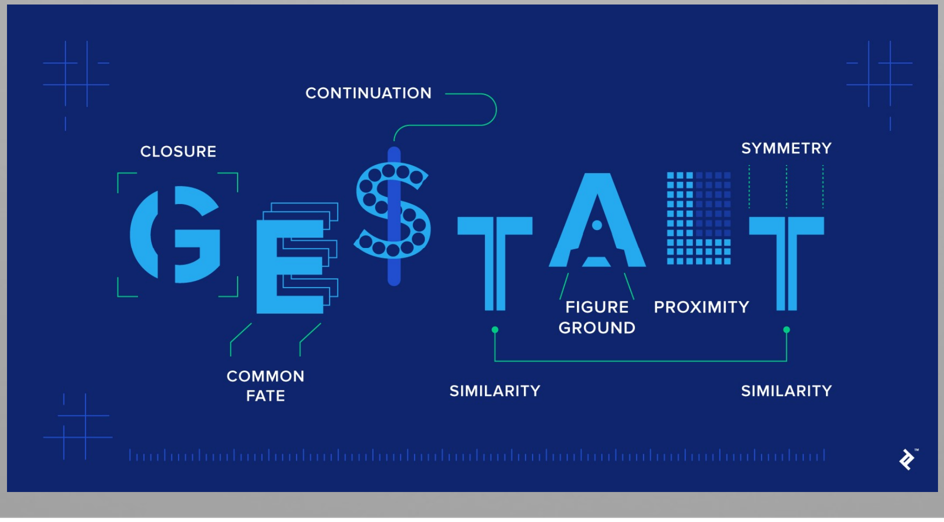
(*Tufte E. - The Visual Display of Quantitative Information*)

Other considerations

1) Who is the audience of your graph?
   → Always know your audience.

2) What is the target medium? A laptop, a large projector, a magazine, a print poster or a handout?
   → Consider resolution, color depth, format, etc.

3) Gestalt principles from Psychology Research of human perception



1) Data Visualization – Why do we care?
2) Data Visualization Elements & Tools
3) Advanced Graph Examples (with R code)
4) Emerging Data Visualizations
5) R Exercises

Beyond line charts, scatterplots and histograms …

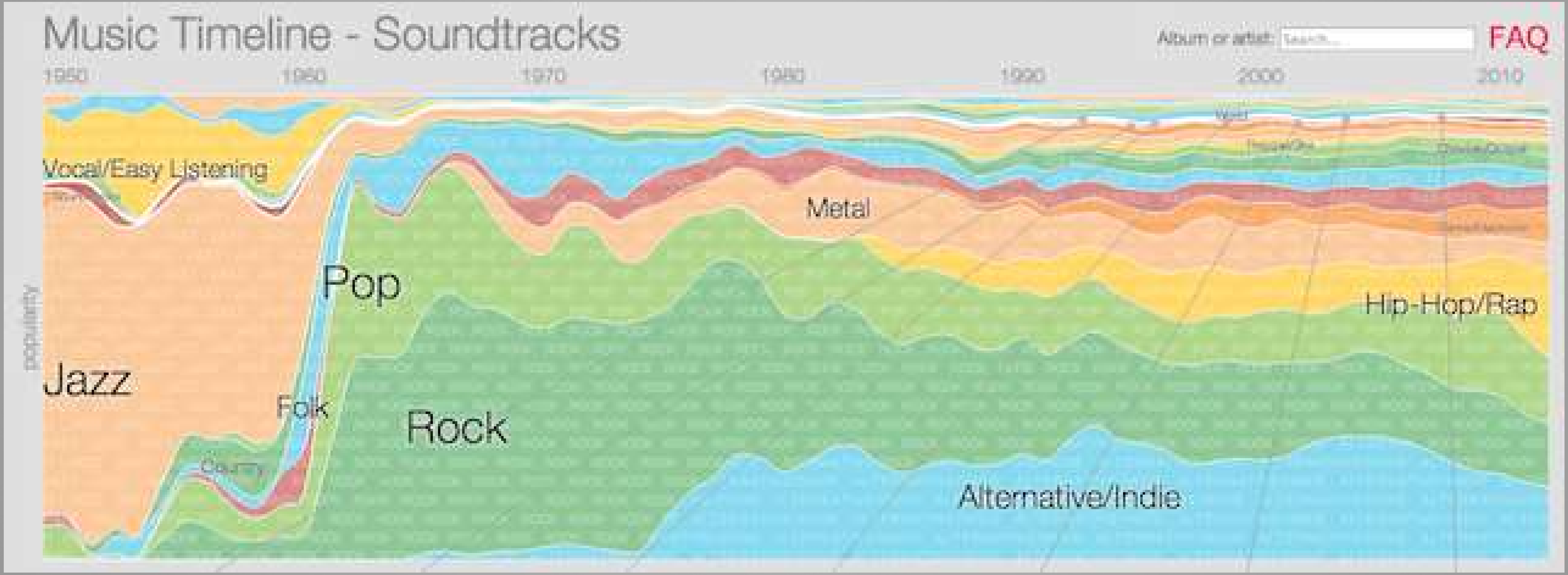
References

https://www.r-graph-gallery.com/

https://flowingdata.com/

https://plot.ly/r/shiny-gallery/

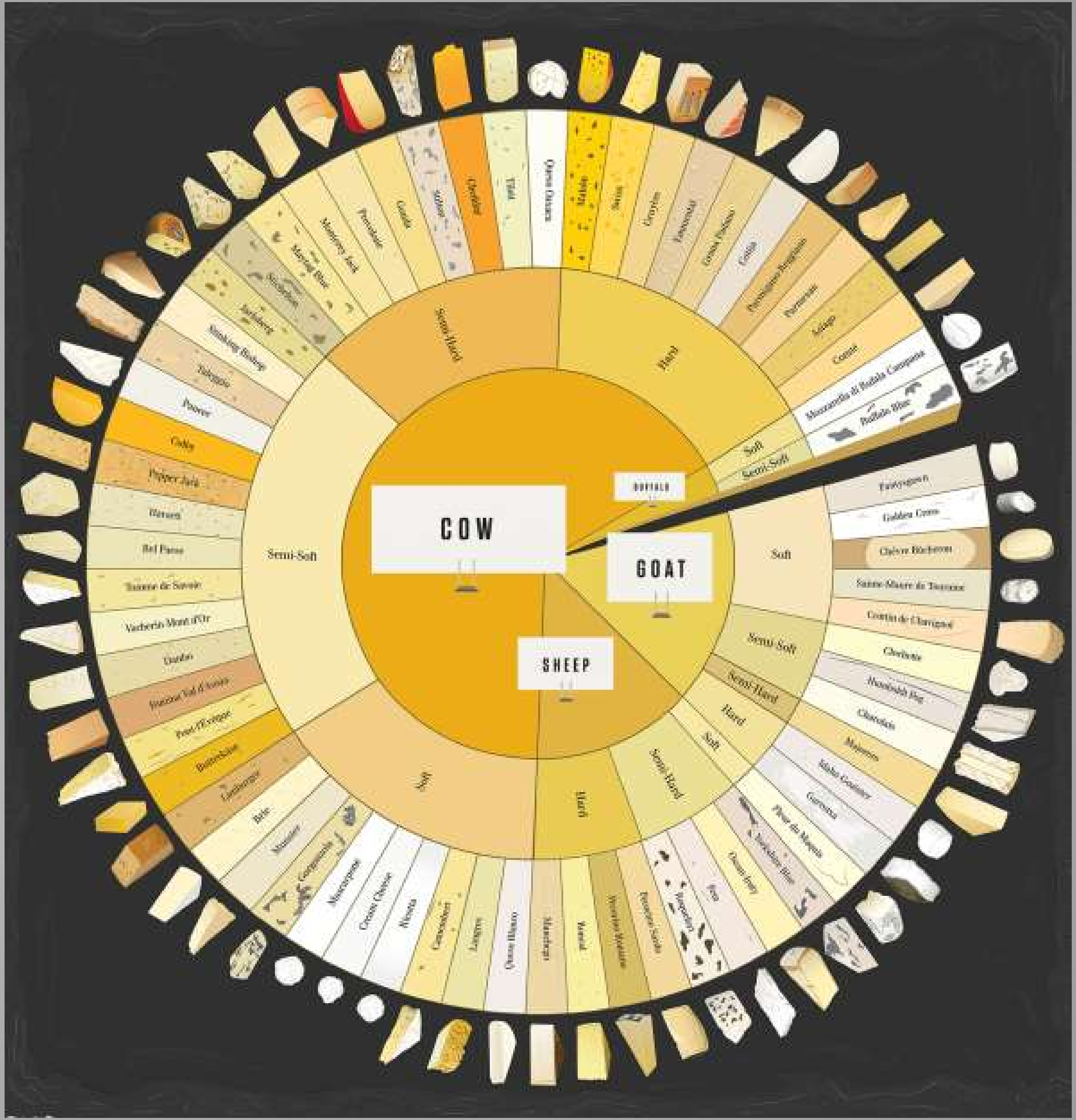
[1] Streamgraph



Source

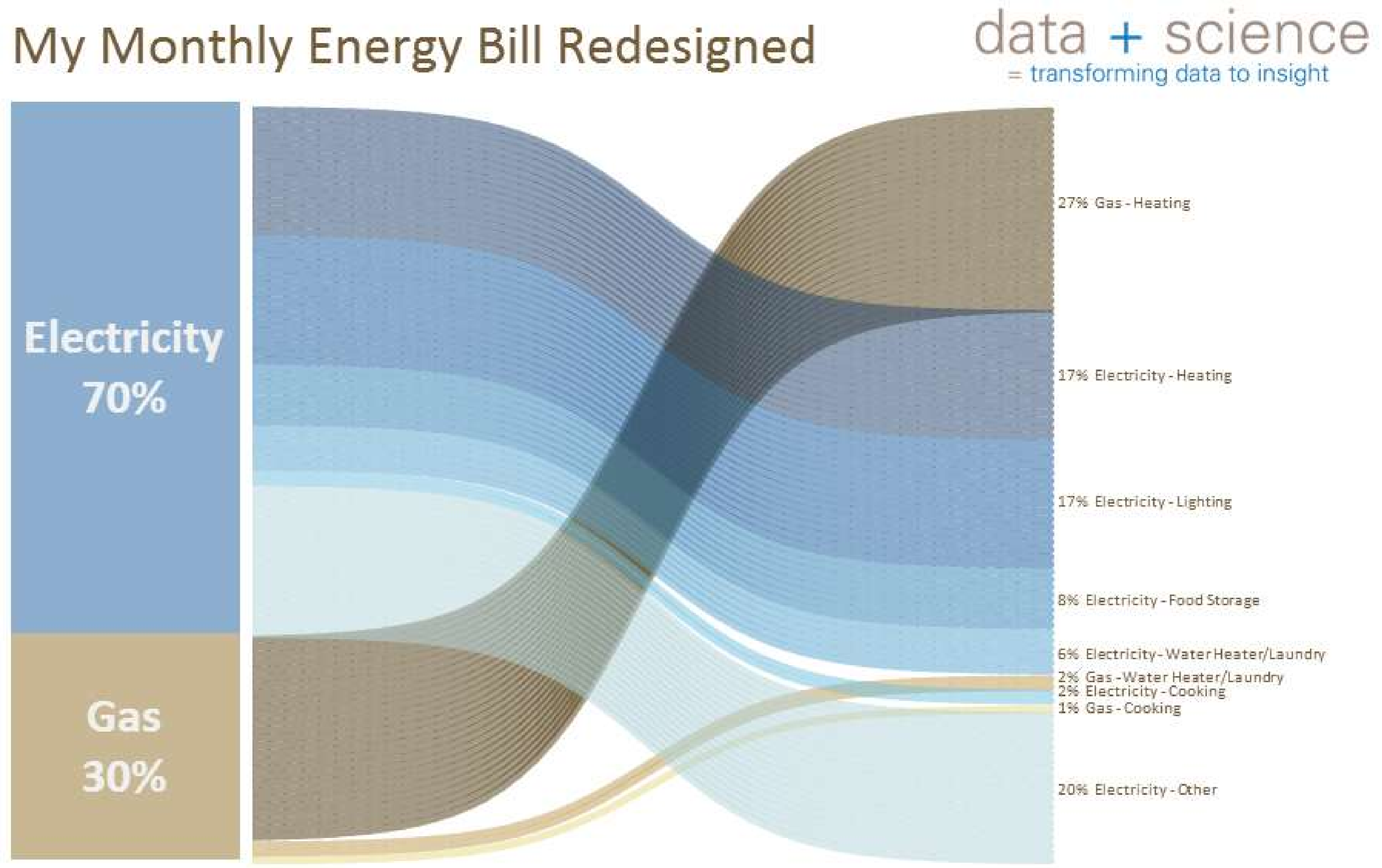
[2] Sunburst

… or how to tell the difference between 66 sorts of cheese.



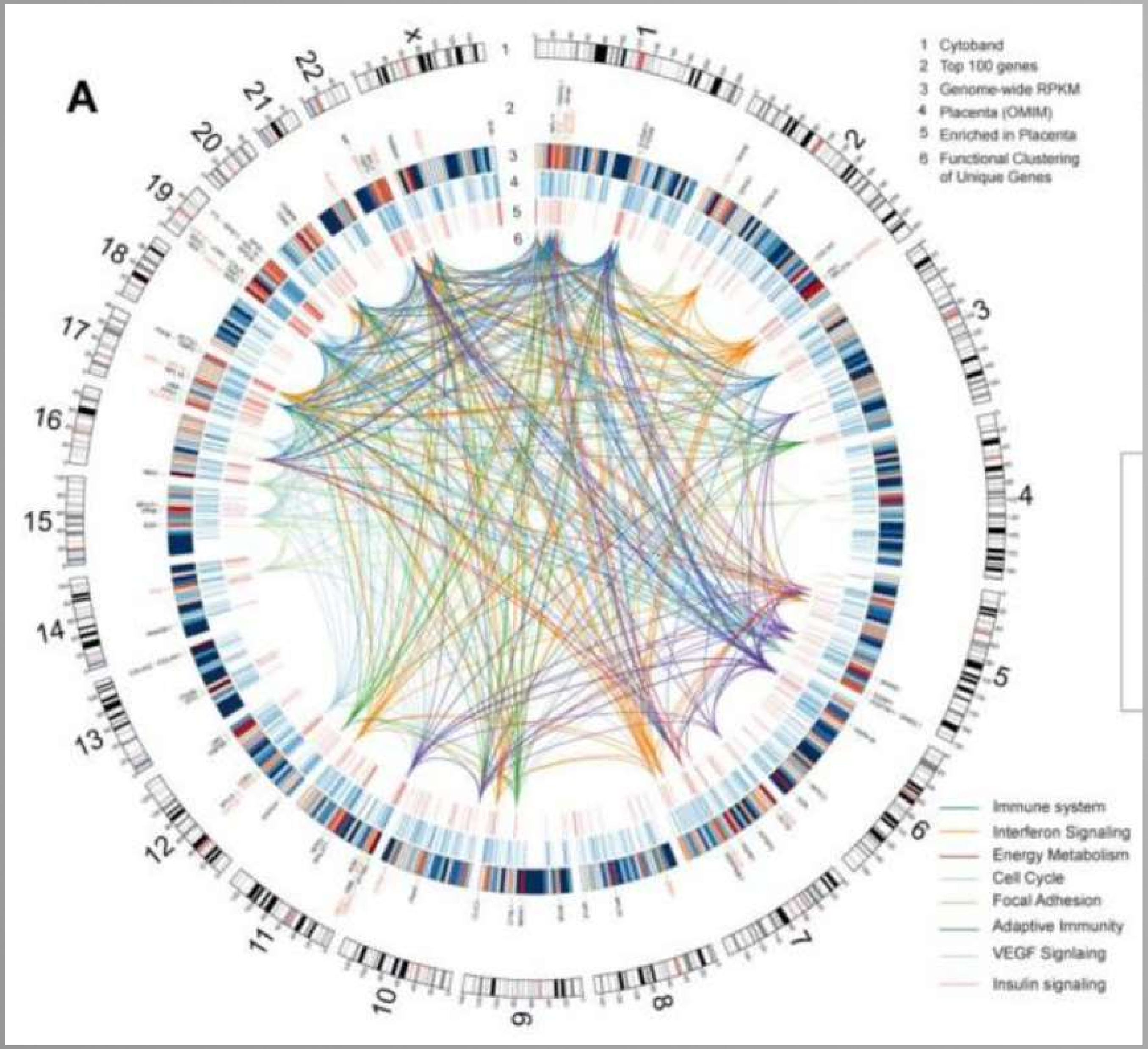
[Source](#)

[3] Sankey Chart



Source

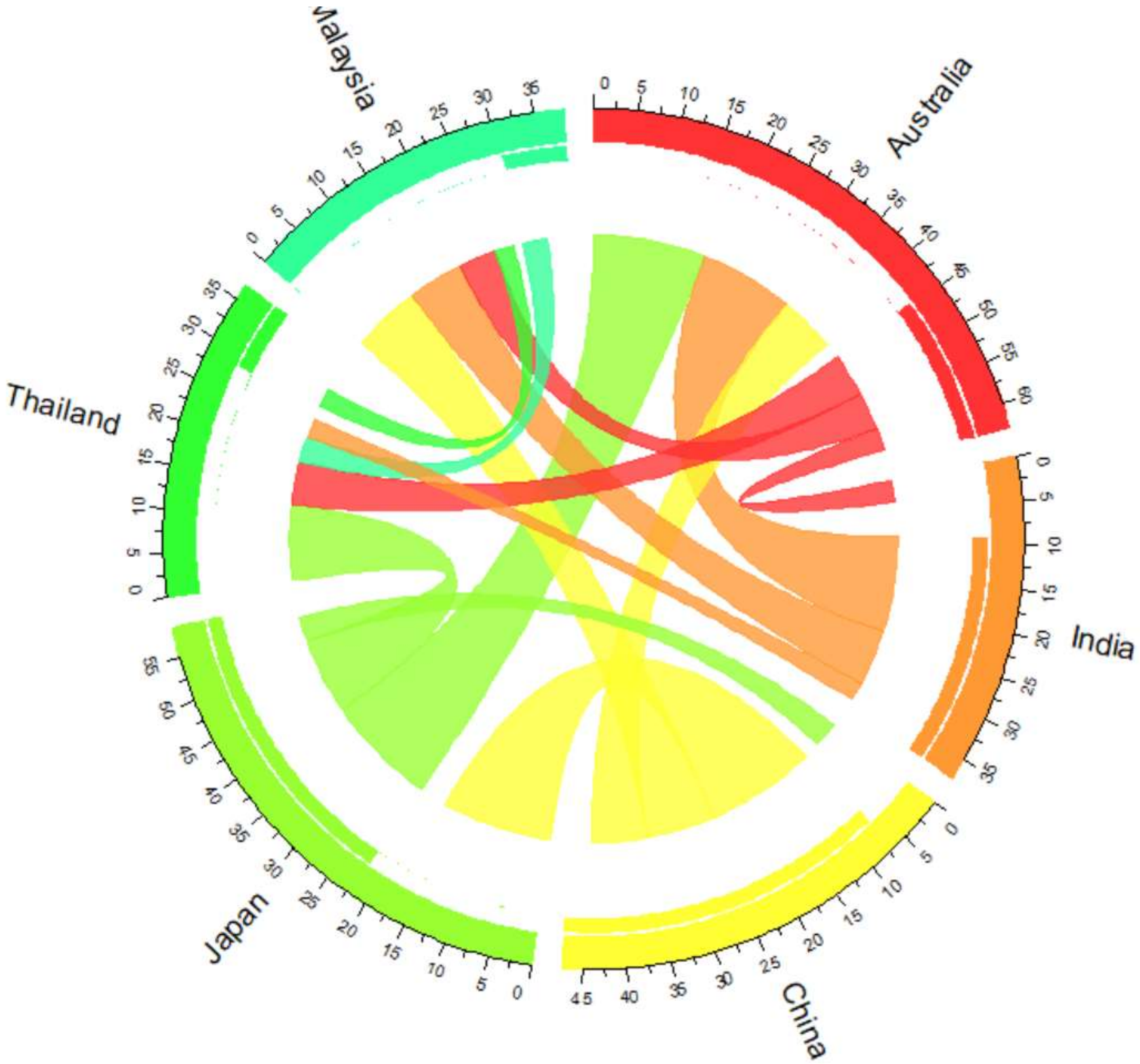
[4] Circular Plot / Chord Diagram



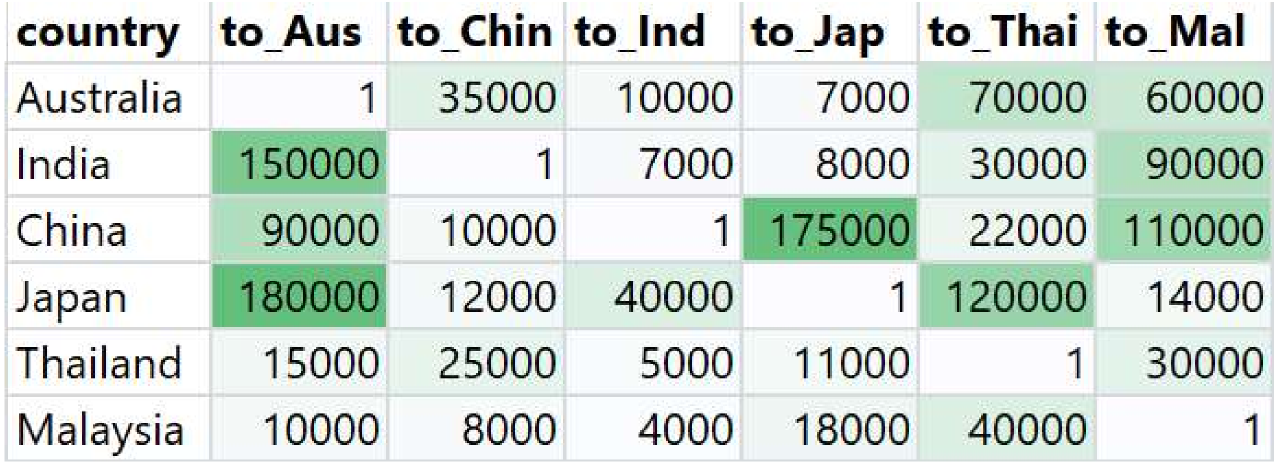
- Ability to show large amount of detail
- Heavily used in genomics research
- Interactivity can be key

- Source: circos.ca

[4] Circular Plot / Chord Diagram [R]



- shows and differentiates inbound & outbound migration patterns per country,
- magnitudes are easier to compare than on a map (due to fixed reference)



country	to_Aus	to_Chin	to_Ind	to_Jap	to_Thai	to_Mal
Australia	1	35000	10000	7000	70000	60000
India	150000	1	7000	8000	30000	90000
China	90000	10000	1	175000	22000	110000
Japan	180000	12000	40000	1	120000	14000
Thailand	15000	25000	5000	11000	1	30000
Malaysia	10000	8000	4000	18000	40000	1

Source

Advanced Graphs in R

In summary:

1) **tradeoff** between showing enough detail and cognitive load

2) Tendency to adding more complexity and dimensions

3) Complex graphs in this section are a good example for visualizing more data rich relationships between variables…

4) … but need a lot of preparation, crafting, and often a detailed description to convey a story

1) Data Visualization – Why do we care?
2) Data Visualization Elements & Tools
3) Advanced Graphs in R
4) **Emerging Data Visualizations**
5) R Exercises

“Tomorrow’s Earth” (*Science* June 28th 2018)

To fight critics, animal researchers urge more transparency *p. 1392*

Seeing how an eye sees *p. 1447*

Human mortality at extreme age *p. 1459*

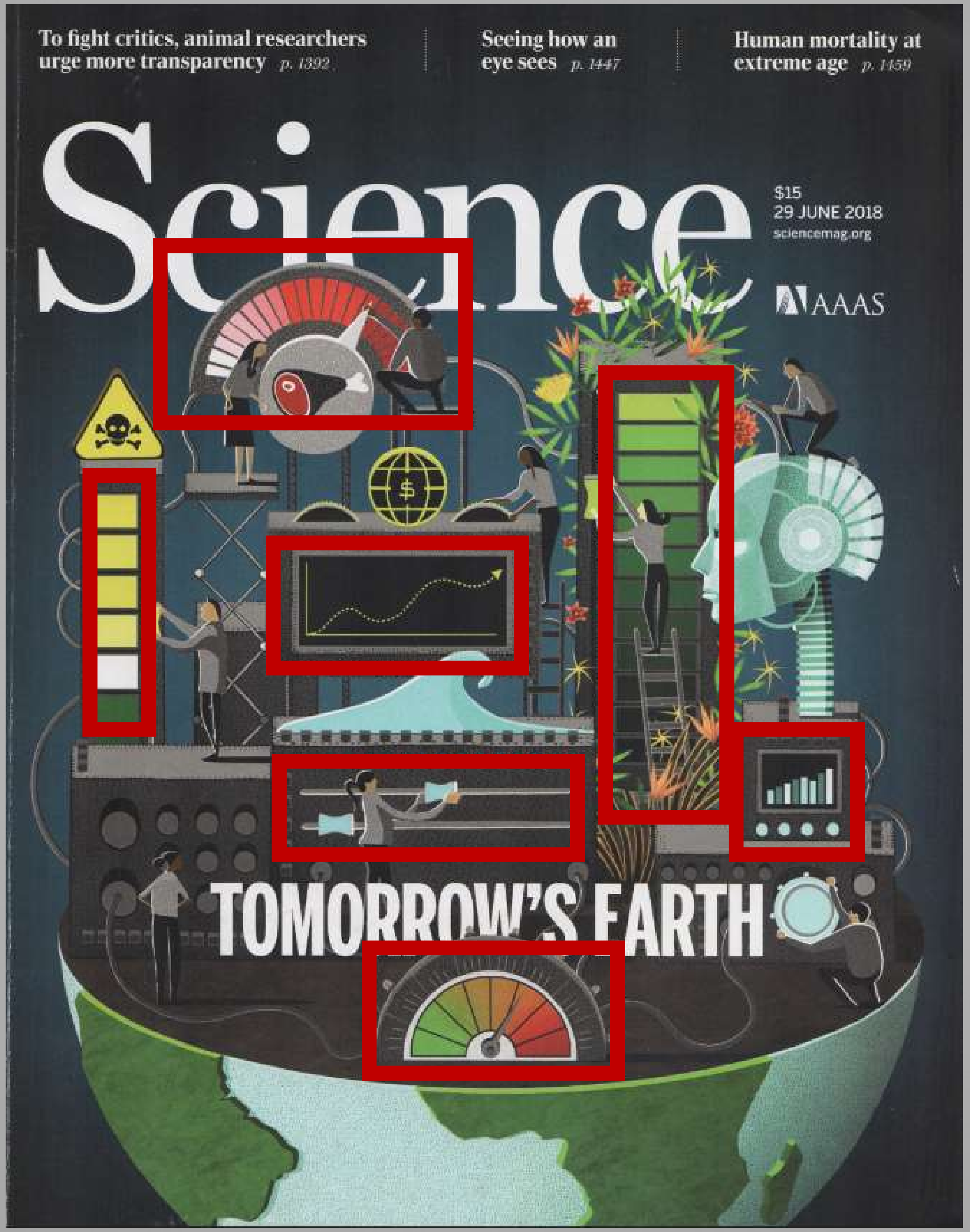
Science

$15
29 JUNE 2018
sciencemag.org

AAAS

TOMORROW’S EARTH

“Tomorrow’s Earth” (*Science* June 28th 2018)



To fight critics, animal researchers urge more transparency p. 1392

Seeing how an eye sees p. 1447

Human mortality at extreme age p. 1459

__Tomorrow’s Earth will be__

More quantitative.

More digital.

Therefore, more visual.

>10 Million data points in one large network graph

facebook
December 2010

When you have, say, 10M data pairs or more…

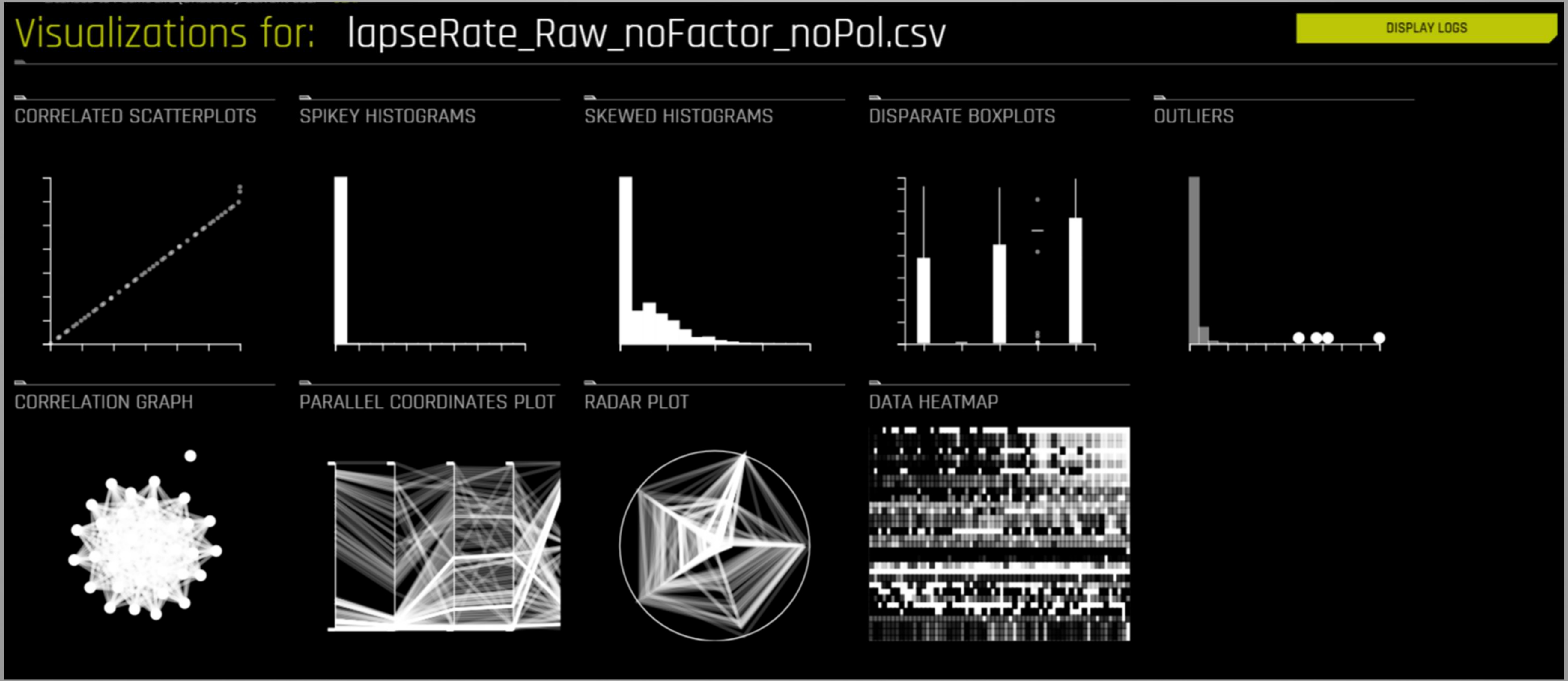
How to test if your graph
represents the data correctly?

Practical recommendations

To reduce bias in charting, the following steps can help:

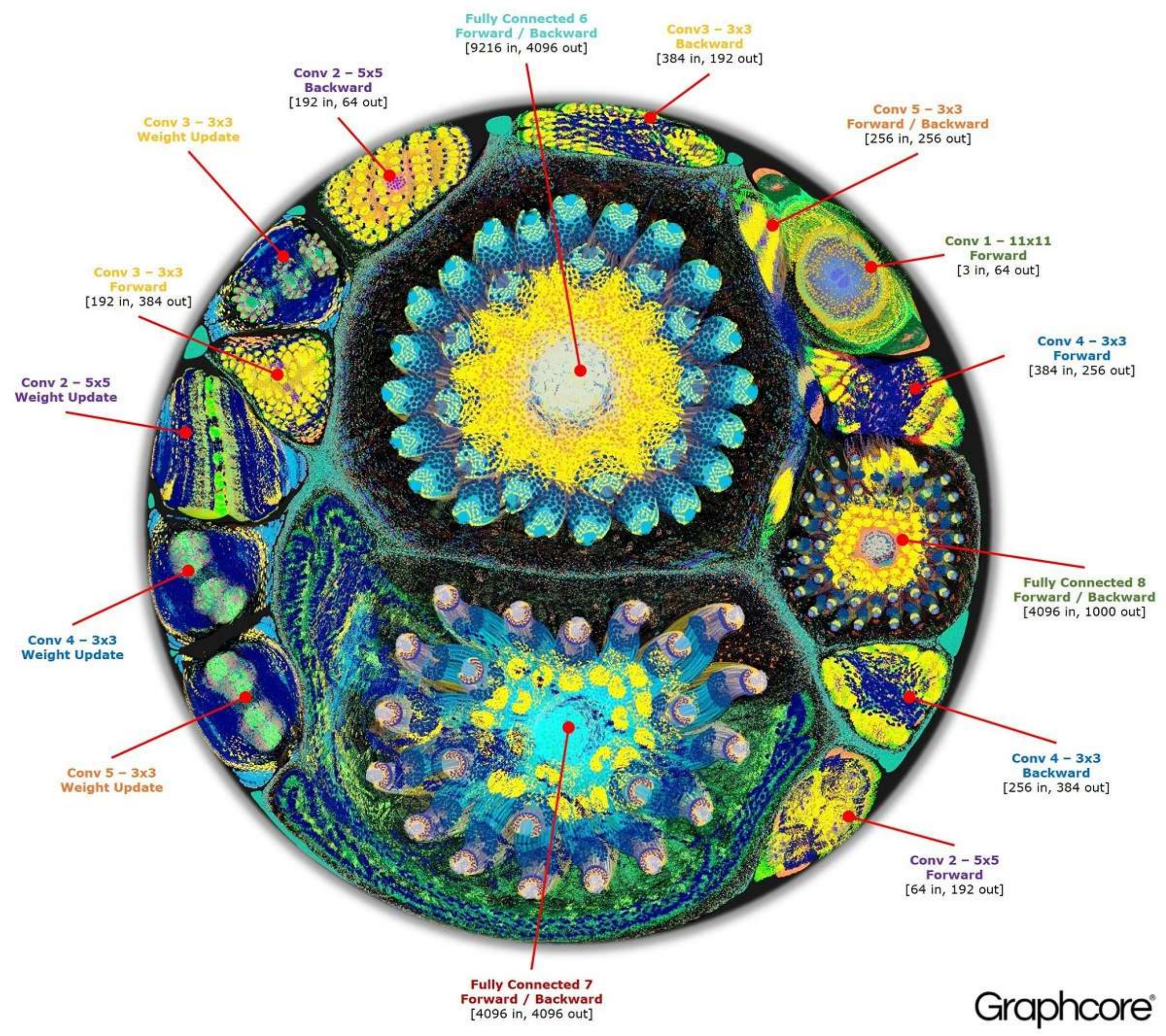
1. Plot multiple samples of the same data set.
2. Plot even more samples. Repeat.
3. Plot various aggregation levels (sum, group by, mean vs median).
4. Validate with external data sets (ideally, obtained from independent sources).
5. Use interactivity
6. Test for visual significance, e.g., with error bars —> “graphing uncertainty”.
7. Transform axis (log, log-log, sigmoid, etc.), align, stretch/compress axis ranges.

Dynamic Data Visualization has not been solved, but we're getting a lot closer → H2O.ai's "Auto Visualization" method



What will graphs look like in 5-10 years?

AlexNet



A visualization of … a computational graph of **18.7 million vertices** and **115.8 million edges**
Using a good amount of abstraction to show different IPUs (Intelligent Processing Units)

Take away

(1) Data Visualization is a highly interdisciplinary task

(2) Data Visualization needs are ubiquitous along the Data Science Stack

(3) Data Visualization is not fully automated at all

(4) New techniques offer a way to pack more data into a graph, but come at a cost (heavy manual editing, detailed descriptions, danger of cognitive overload)

Learning how to make better visualizations, and to fit visualizations to given data sets for a given audience will help us to improve our means of Visual Communication.

Further Reading

1) Munzner, T. (2017) Visualization Analysis & Design, *VIS 17 Tutorial, Phoenix AZ 10/17* (pdf slides available here http://www.cs.ubc.ca/~tmm/talks.html#vad17fullday

2) Tufte, E. (1990) Envisioning Information. Graphics Press

3) Wilkinson, L. (2005) The Grammar of Graphics. Springer

1) Data Visualization – Why do we care?
2) Data Visualization Elements & Tools
3) Advanced Graphs in R
4) Emerging Data Visualizations
5) R Exercises

Exercise 1: Biopharmaceutical App

Github link: https://github.com/olaf-menzer/biopharma-app

Example of an interactive, visual clustering app built in R shiny

It is centered around interactivity in data exploration, through the following:

- Plotting a heatmap
- A simple hierarchical clustering algorithm
- Plotting a phylogenetic tree to visualize the clusters

Exercise 2: Complex Graphs in R

Github link: https://github.com/olaf-menzer/Rgraphs

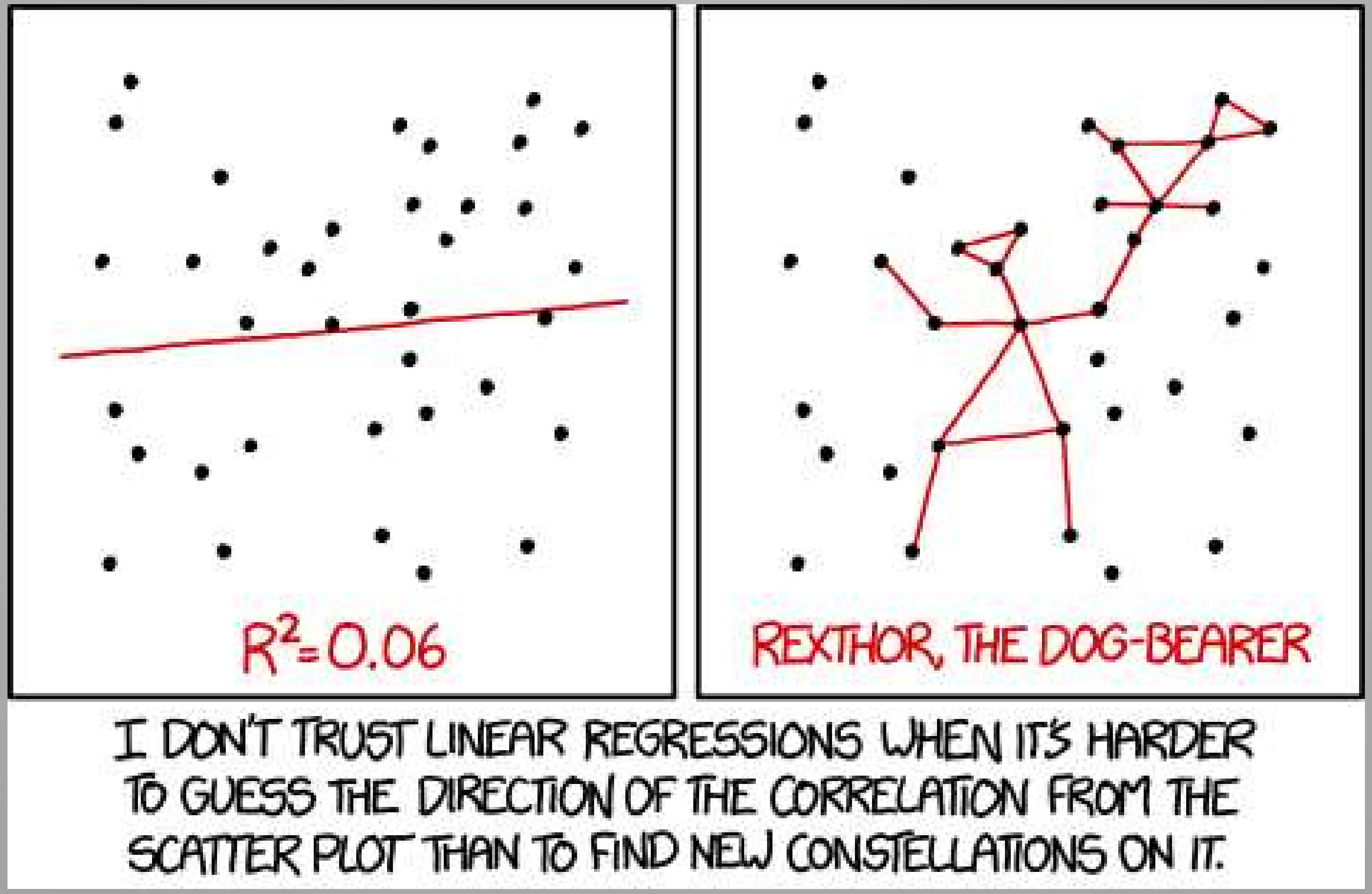
We will learn how to plot the charts in the section “Beyond line charts, scatterplots and histograms” in R, some of them with interactivity.

4 example R scripts (mini examples) for plotting:

(1) Streamgraph	(01_Streamgraph.R)
(2) Sunburst Chart	(02_SunburstR.R)
(3) Sankey Chart	(03_Sankey.R)
(4) Circular Chart	(04_Circular_Chord.R)

Example data sets given in file Explore_DataSets.R
(2) – (4) have small, coded mini example data sets built in so we can run them independent of WiFi bandwidth

Thank you.


R²=0.06
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.