Quantified Self

Introduction to Data Visualization

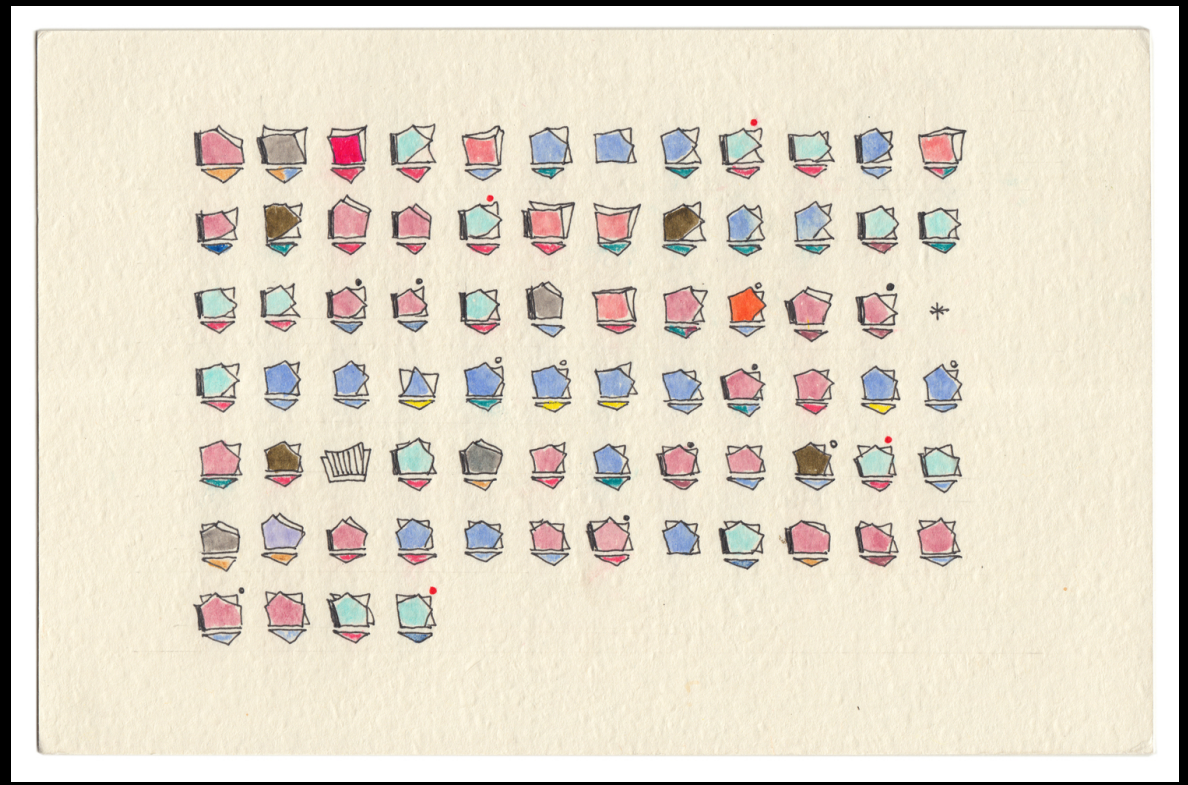
The Graduate Center at CUNY | Summer 2018

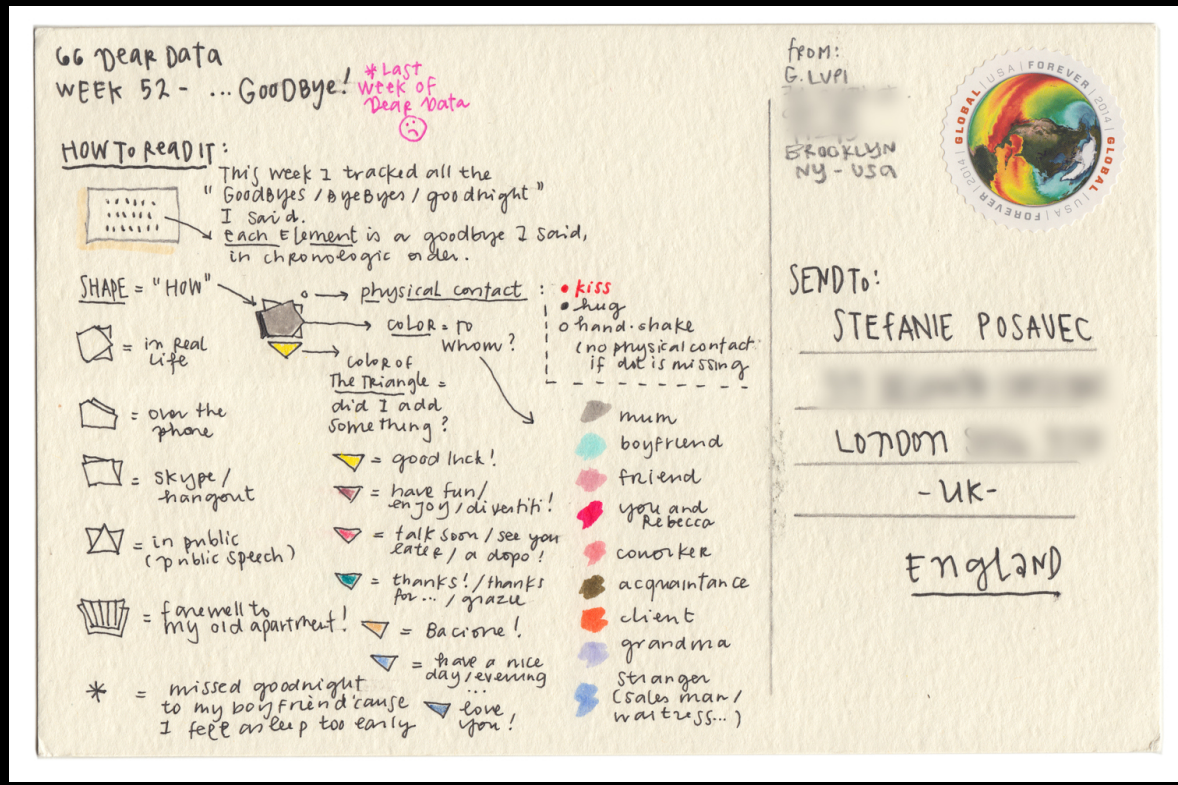
June 5, 2018

Five Minute Reflection

What data do you keep track of? Why? What do you do with the data?

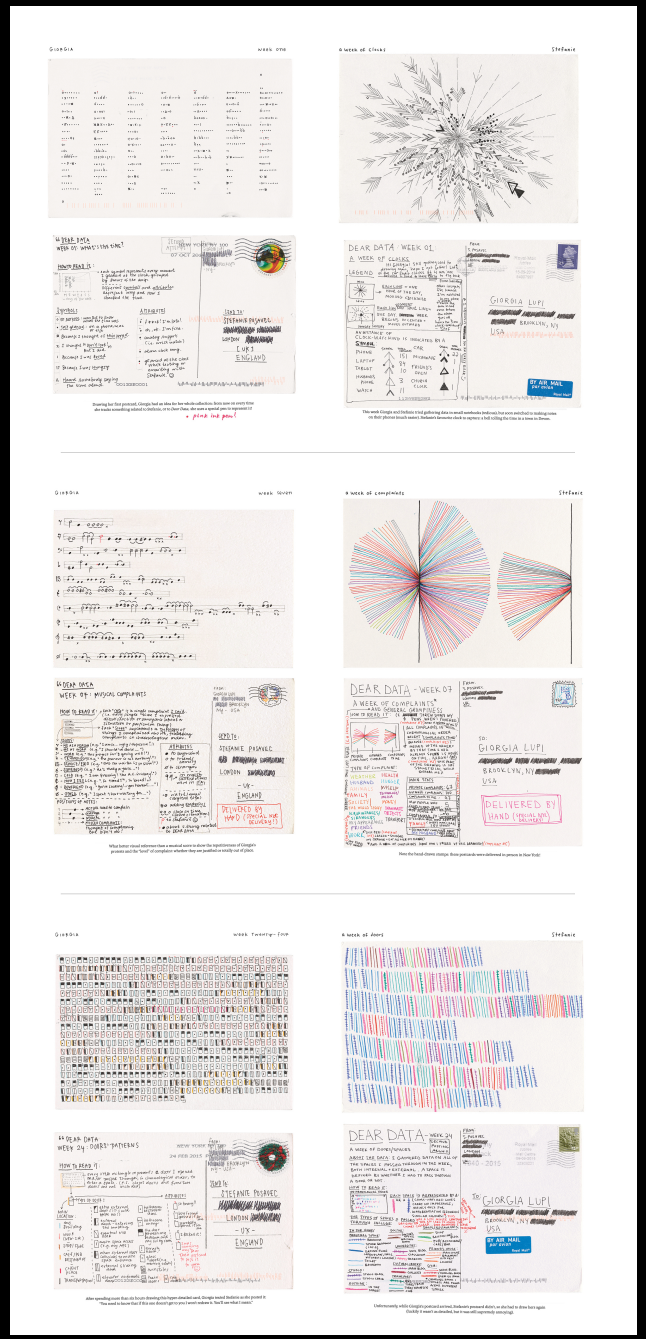
Quantified Self: tries to incorporate technology into data acquisition on aspects of a person's daily life (Wikipedia).





http://www.dear-data.com/all/

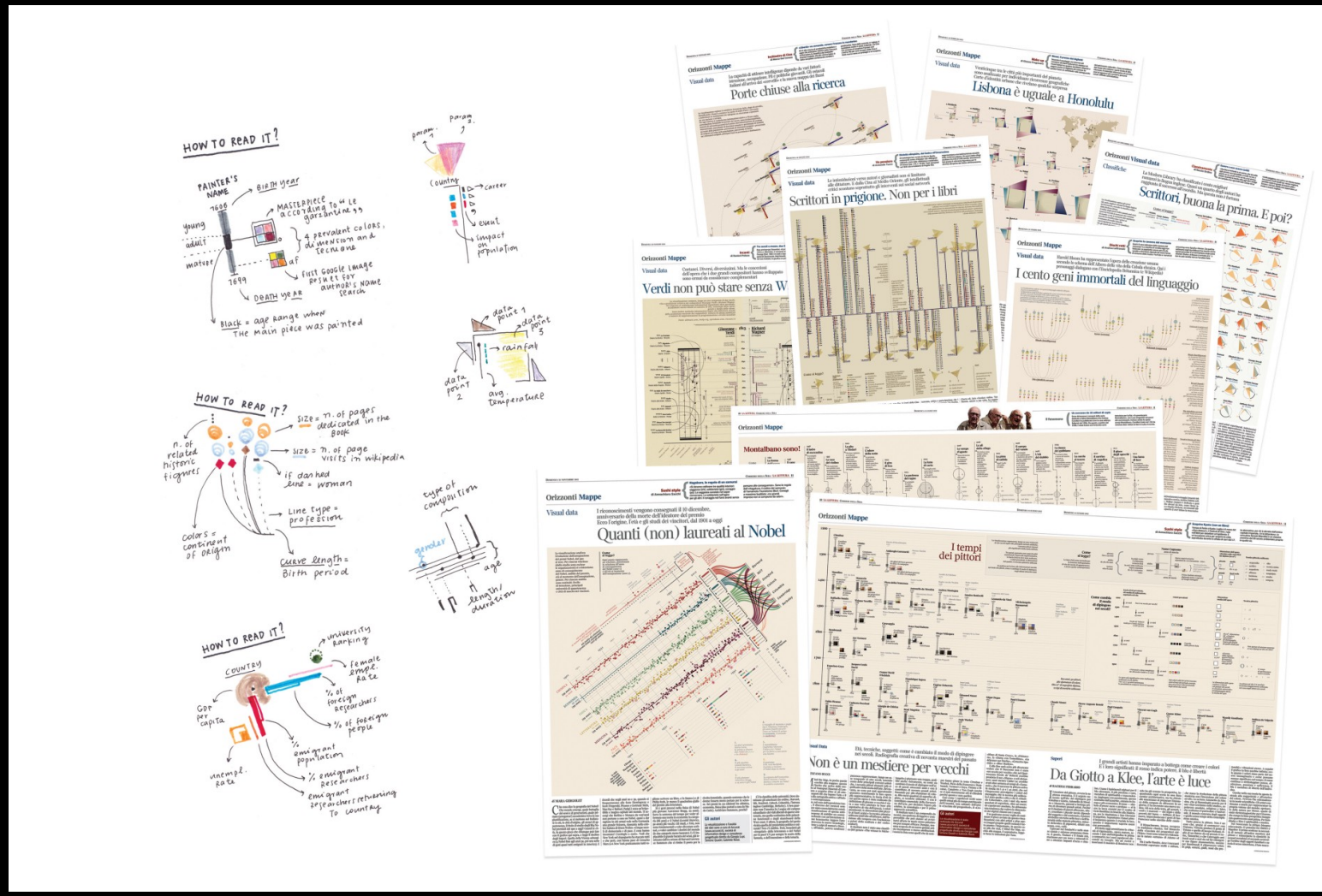
Dear Data 1 & 2



For all its flaws, data represents:

- Knowledge
- Behavior
- People
- And other phenomena

Ways Forward

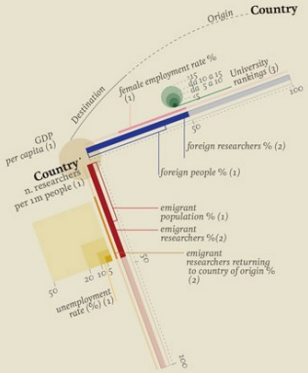


Brain drain

The phenomena of so-called «brain drain» is explored through a map showing incoming and outgoing flows of researchers in 16 countries. Using a series of parametres, the map is an attempt to discover the motivations that move researchers from one country to another. Each country is visualized through the representation of: GDP per capita, female employment rate, overall unemployment rate, university rankings, percentage of foreign researchers, percentage of overall foreign population, percentage of emigrant researchers, percentage of overall emigrant population, percentage of researchers returning to their country of origin, and the main countries researchers come from and move to.

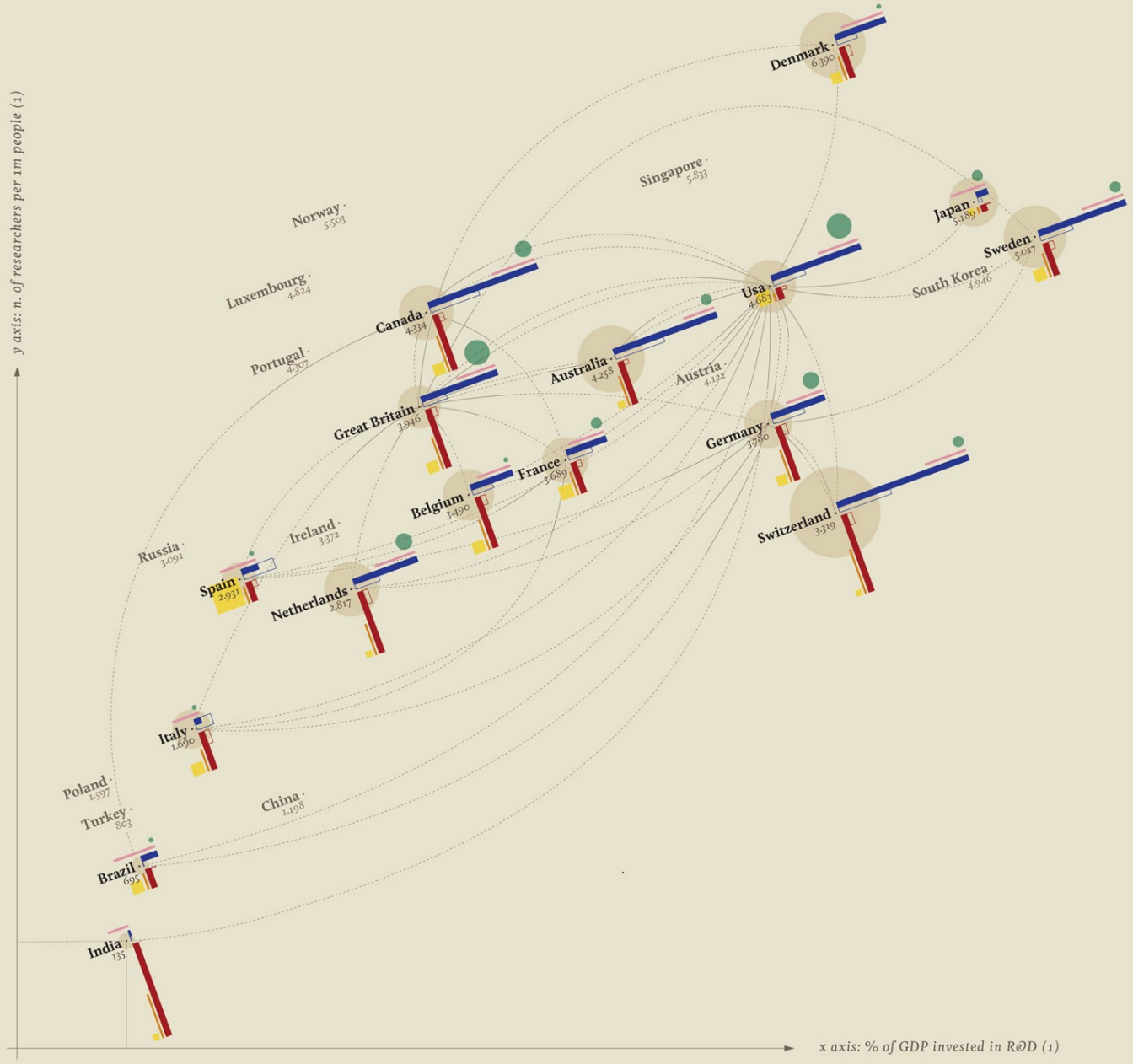
How to read it?

The countries are positioned according to:
% of GDP invested in R&D (x axis)
+ n. of researchers per 1m people (y axis)
The analysis is based on the following data

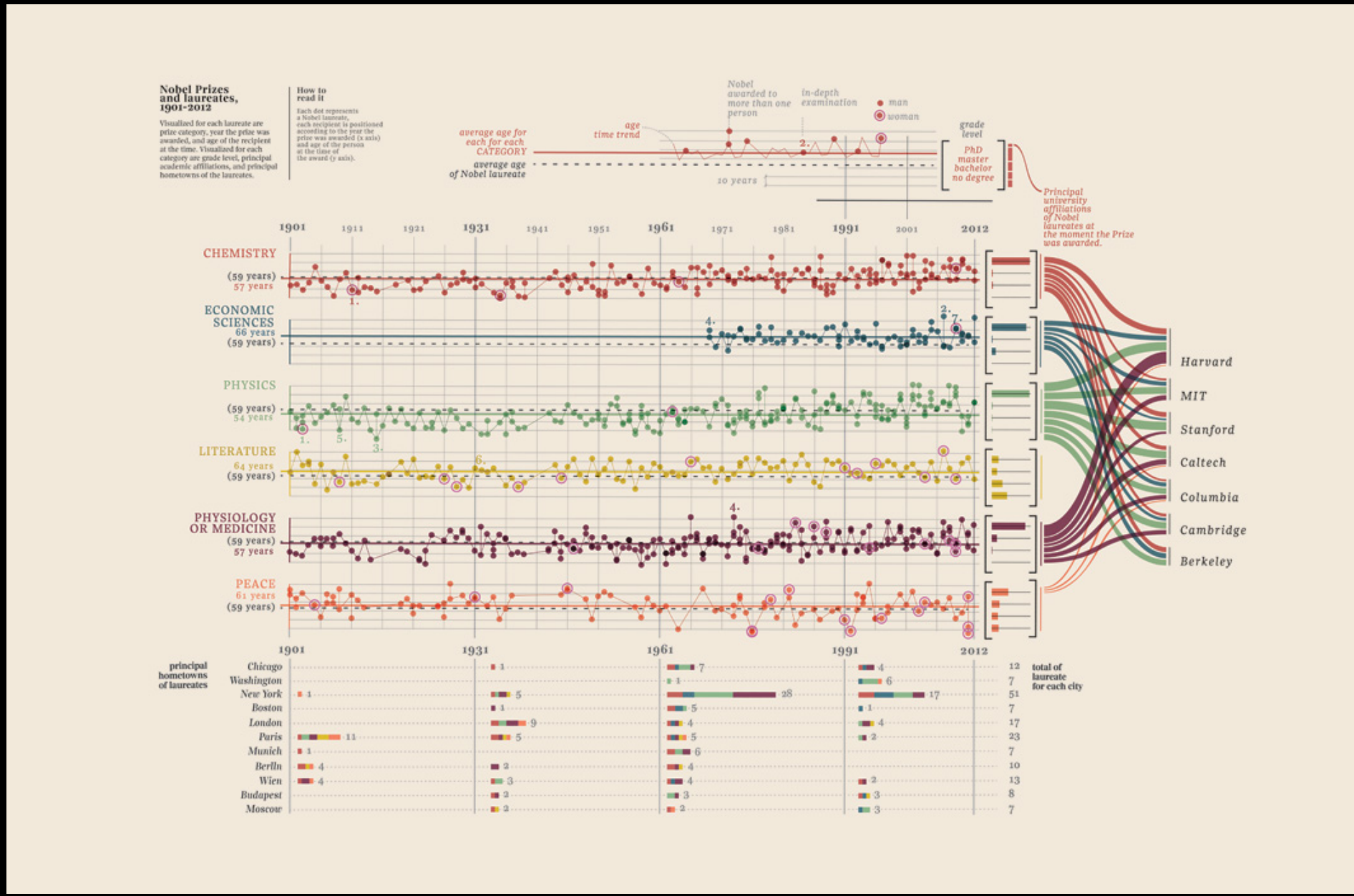


(1) World Bank (2005-2010, worldbank.org)
(2) Foreign Born Scientists: Mobility Patterns for Sixteen Countries (2012 paper by Chiara Franzoni, Giuseppe Scellato and Paula Stephan, nber.org)
(3) Times Higher Education World University Rankings (2011-2012 timeshighereducation.co.uk)

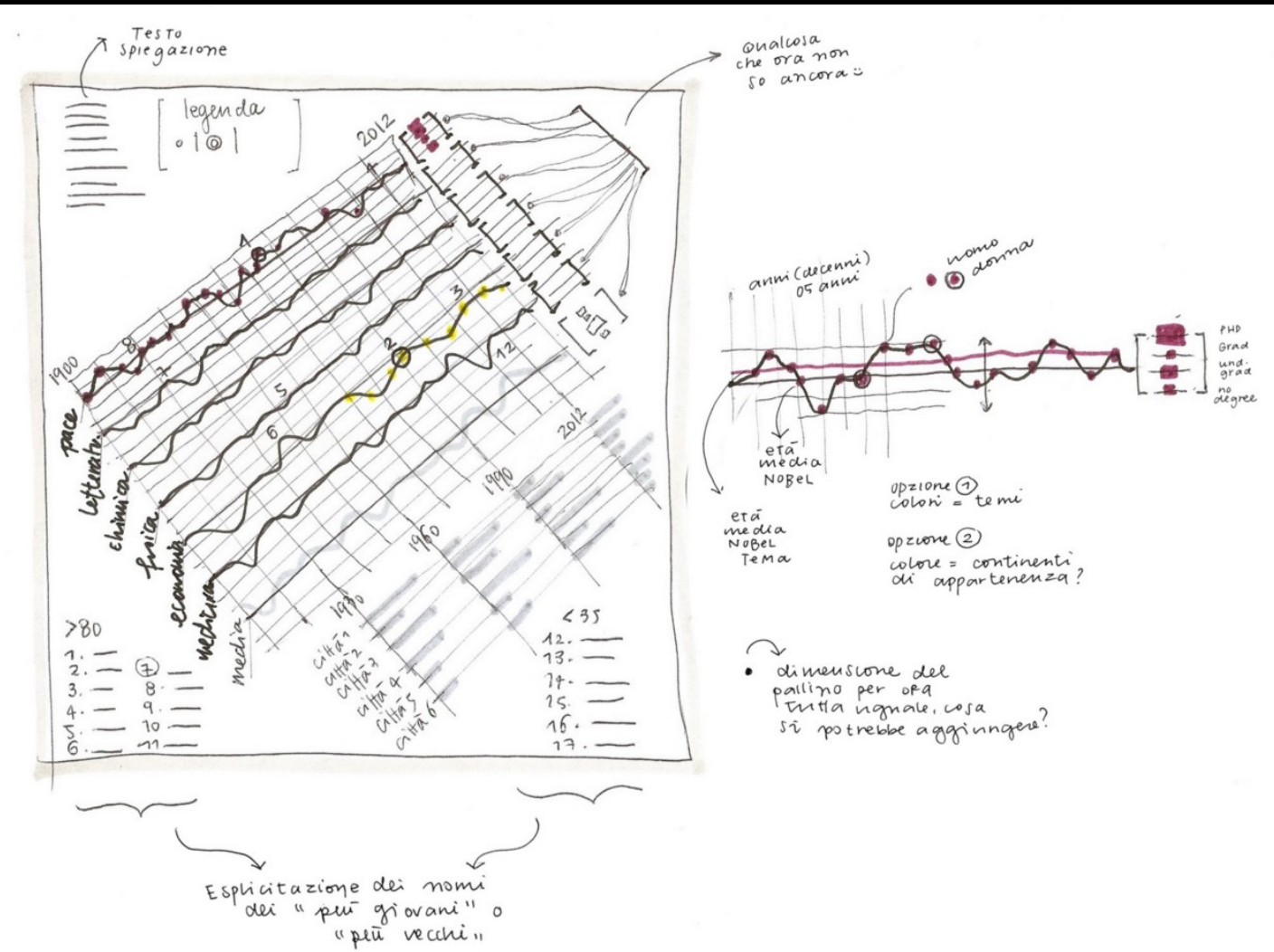
The visualization has been designed and produced by Accurat (www.accurat.it), and was originally published in italian on La Lettura the sunday cultural supplement of Corriere della Sera.



But Provide a Roadmap



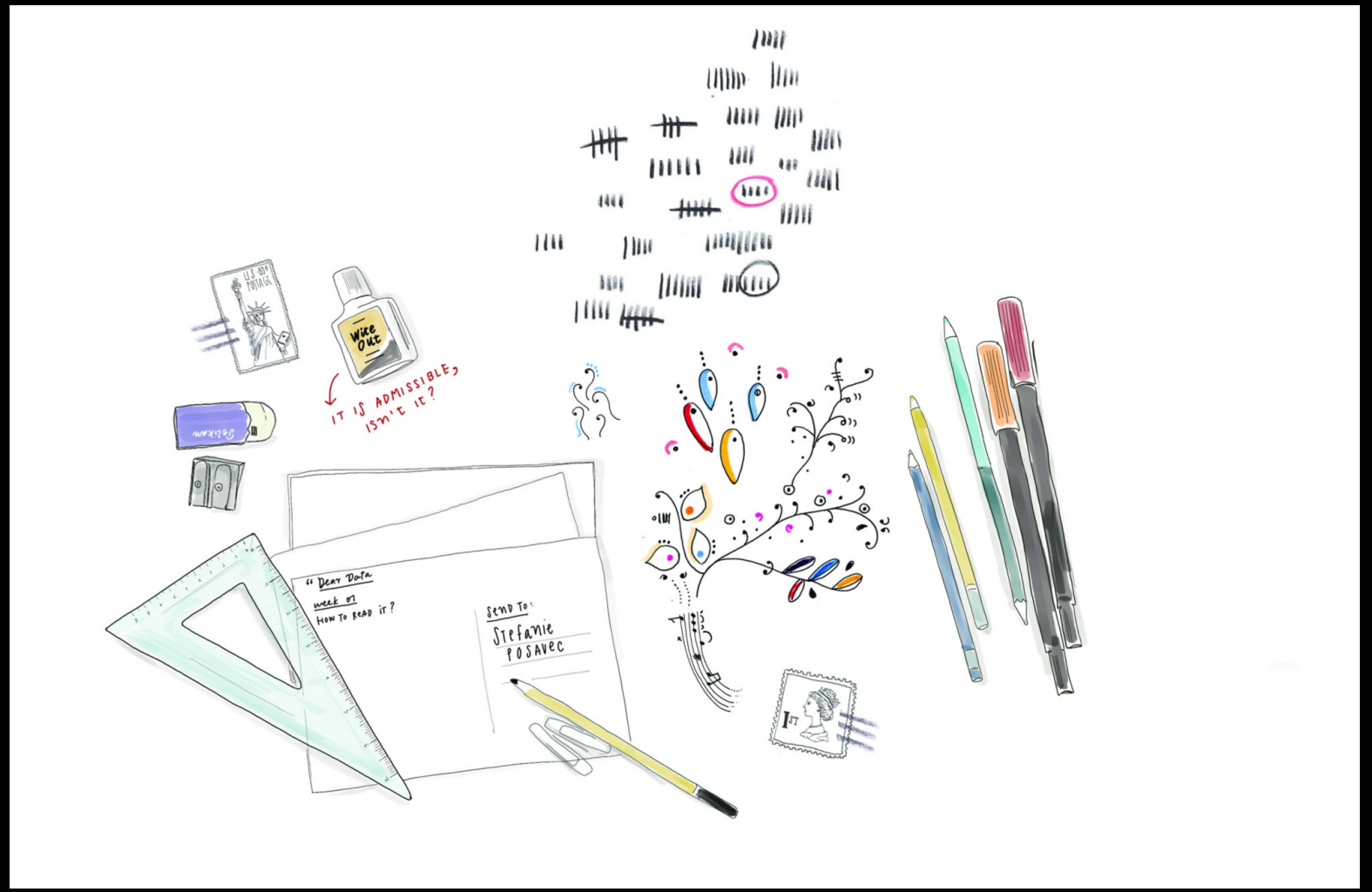
Learn the Standards then Move Beyond



CONTEXT

- Know the Data
- Collect it yourself if possible
- Understand how it was collected
- Really get to know the data
  - Transform it
  - Clean it
  - Sample from it

Data will always be imperfect



Data

Introduction to Data Visualization

The Graduate Center at CUNY | Summer 2018

June 6, 2018

Five minute reflection
Sort the following people into no more than 3 categories (feel free to google them):

- Leo Tolstoy
- Beyoncé Knowles
- Nathan Yau
- Haruki Murakami
- Oprah Winfrey
- Victor Hugo
- Miriam Makeba
- Michael Jackson
- J.K. Rowling
- Lev Manovich
- William Playfair

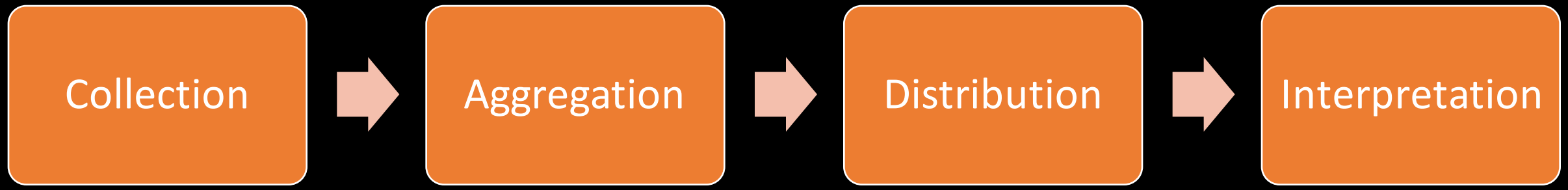
Data: plural of datum.

Datum: something given in an argument; the starting point on which to build; that which is assumed

- Rosenberg 2013, Miriam Webster

Raw Data: ??

Making Data



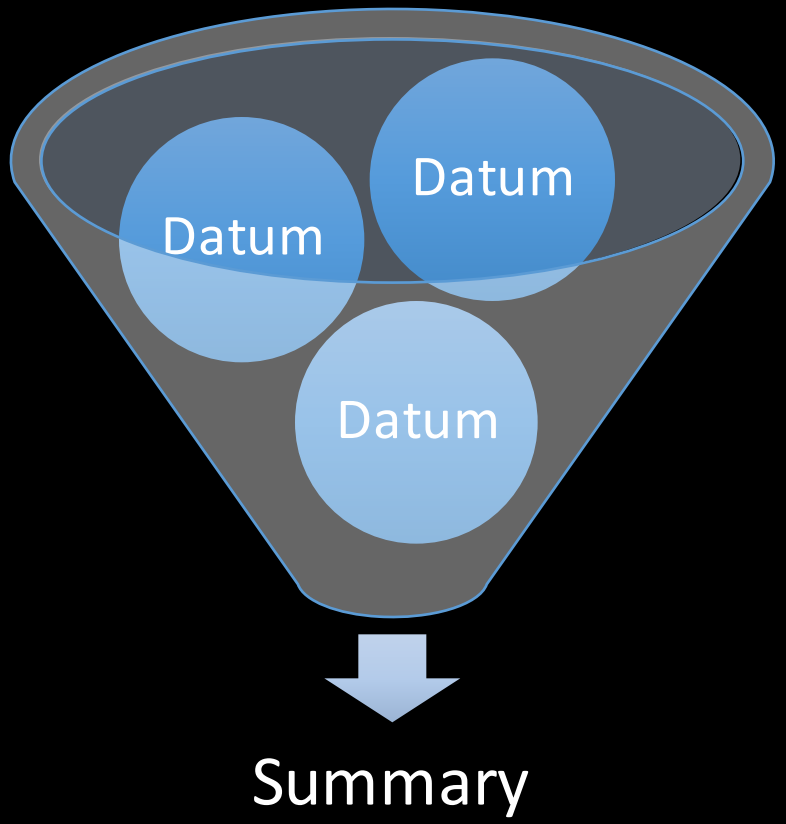
Collection

- Surveys
- Observations
- Measurements

João Batista Neto
https://commons.wikimedia.org/w/index.php?curid=43063497

Aggregation

- Disaggregated Instances
- Counts
- Mean/Median/Mode
- Min/Max/Range
- Quartiles



Distribution

- Tables (csv/tsv/xls)
- Text (txt/json)
- Databases (SQL/Odata)
- Shape File (shp/kmz)

Data Features

- Items: Each instance with all of its attributes
- Attributes: The features of each item
- Semantics: Real world meaning
- Type: structural or mathematical interpretation

AT001	AT002	AT003	AT004	AT005	AT006	AT007	AT008
Leo Tolstoy	0	M	Russian	Author	Yes	1828	20,000
Beyoncé Knowles	1	F	American	Musician	No	1981	2,500,000
Nathan Yau	1	M	American	Statistician	Yes	NULL	150,000
Haruki Murakami	1	M	Japanese	Author	No	1949	300,000
Oprah Winfrey	1	F	American	Entertainer	No	1954	2,000,000
Victor Hugo	0	M	French	Author	No	1802	50,000
Miriam Makeba	0	F	South African	Musician	No	1932	100,000
Michael Jackson	0	M	American	Musician	No	1958	3,000,000
J.K. Rowling	1	F	British	Author	No	1965	10,000,000
Lev Manovich	1	M	Russian	Researcher	Yes	1960	250,000
William Playfair	0	M	Scottish	Economist	Yes	1759	30,000

Items

- Leo Tolstoy
- Beyoncé Knowles
- Nathan Yau
- Haruki Murakami
- Oprah Winfrey
- Victor Hugo
- Miriam Makeba
- Michael Jackson
- J.K. Rowling
- Lev Manovich
- William Playfair

Attributes

- Living
- Gender
- Nationality
- Occupation
- Status as Academic
- Birth Year
- Salary

Semantics

ID	Referent
AT001	Name
AT002	Living
AT003	Gender
AT004	Nationality
AT005	Occupation
AT006	Academic
AT007	Birth Year
AT008	Salary

Types

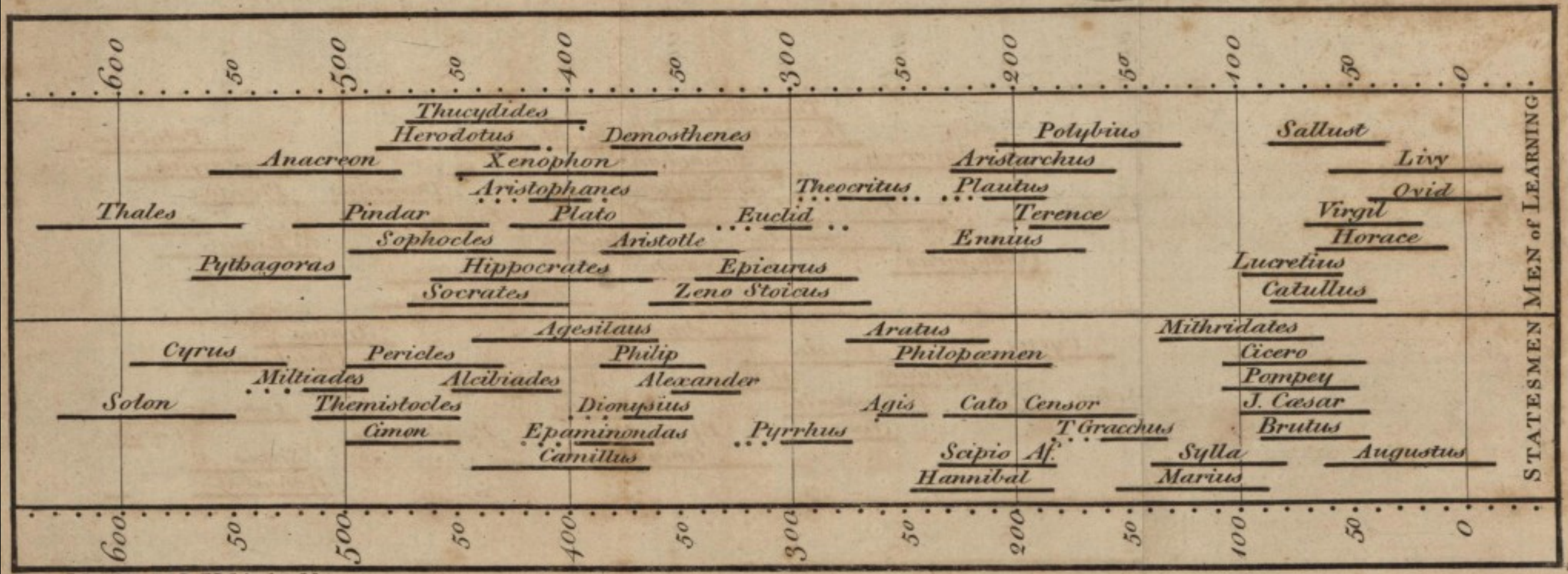
Referent	Data Type
Name	String
Living	Boolean
Gender	Binary Code
Nationality	String
Occupation	String
Academic	Binary Code
Birth Year	Ordinal Number
Salary	Decimal Number

Data is Never “Raw”

- Attributes are never perfect
- Always have bins
- Always have breaks

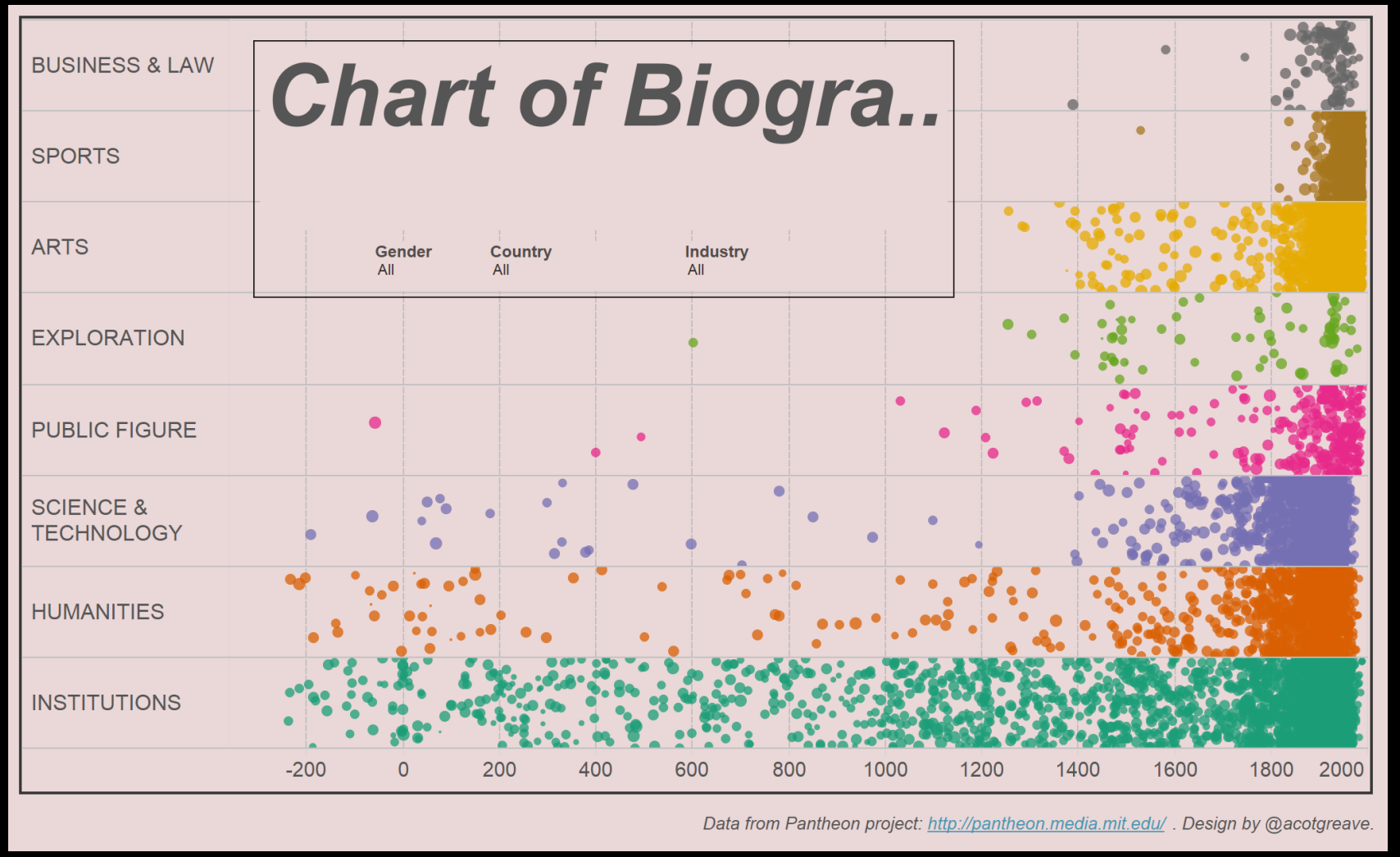
- Collection
- Storage
- Transmission

A Specimen of a Chart of Biography.



J. Priestley LL.D. F.R.S. inv.t et del.

Priestly's Chart Remade



https://public.tableau.com/views/Priestley/ChartofBiography?:embed=y&:display_count=yes&:showTabs=y&:showVizHome=no

Data Types – Chart of Biography

- Chart of Biography:
  - Items: people
  - Attributes: Name (nominal), area of influence (nominal), era (continuous)
  - Timeline/ Gantt Chart
  - Position, length, linear scale
- Remake
  - Items: people
  - Attributes: Name (nominal), area of influence (nominal), era (continuous)
  - Timeline/ Scatter Plot /Bar Chart?
  - Position, Color, linear scale

Three ingredients Three ways

Marks

- Points
- Lines
- Polygons (Areas)

Style

- Position
- Color
- Size

Bins & Breaks

- How many can be perceived
- How many can be understood

- Here:
  - 3 sizes
  - 3 widths

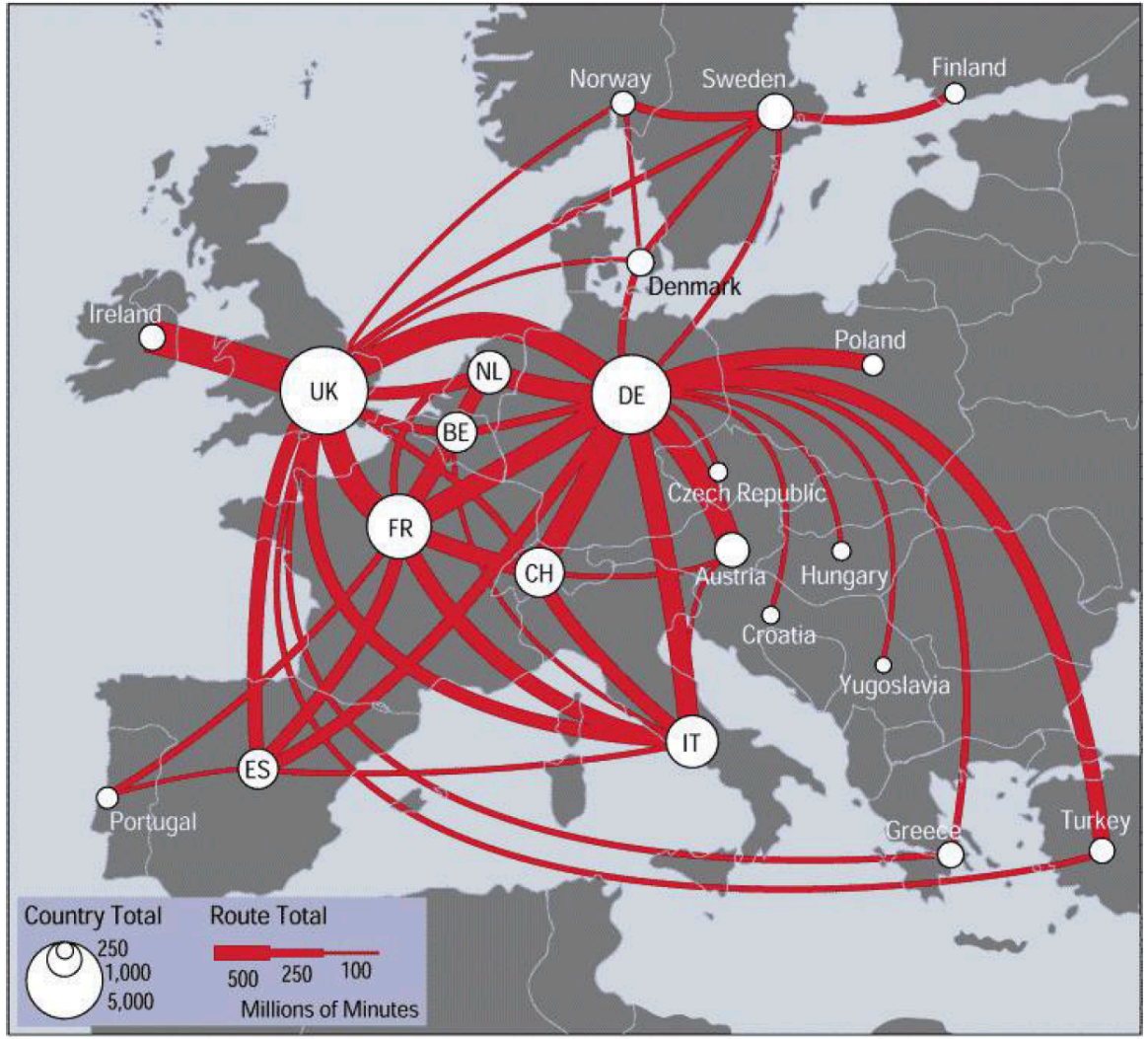
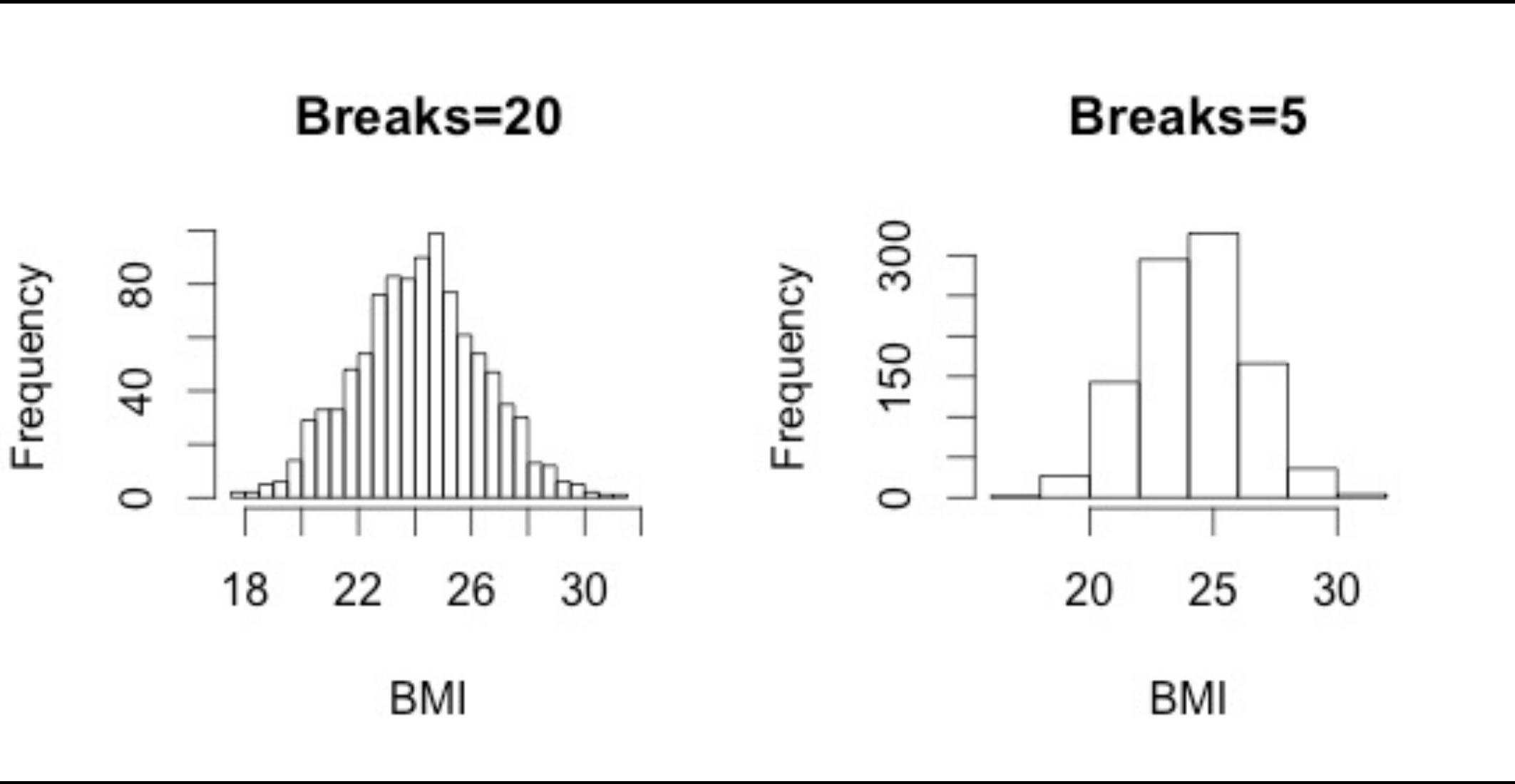


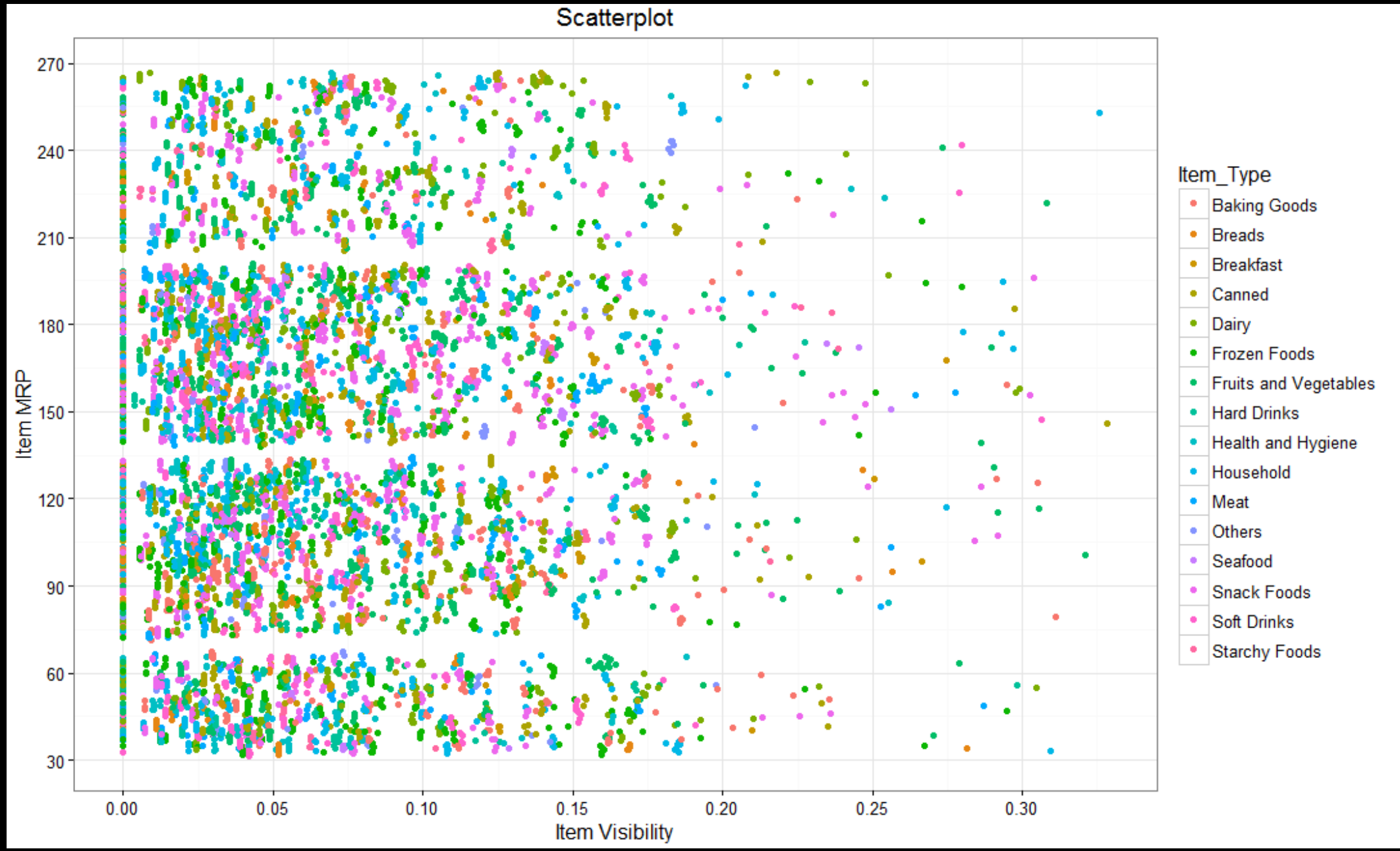
Figure 5.9. Linewidth has a limited number of discriminable bins.

Munzer 2015

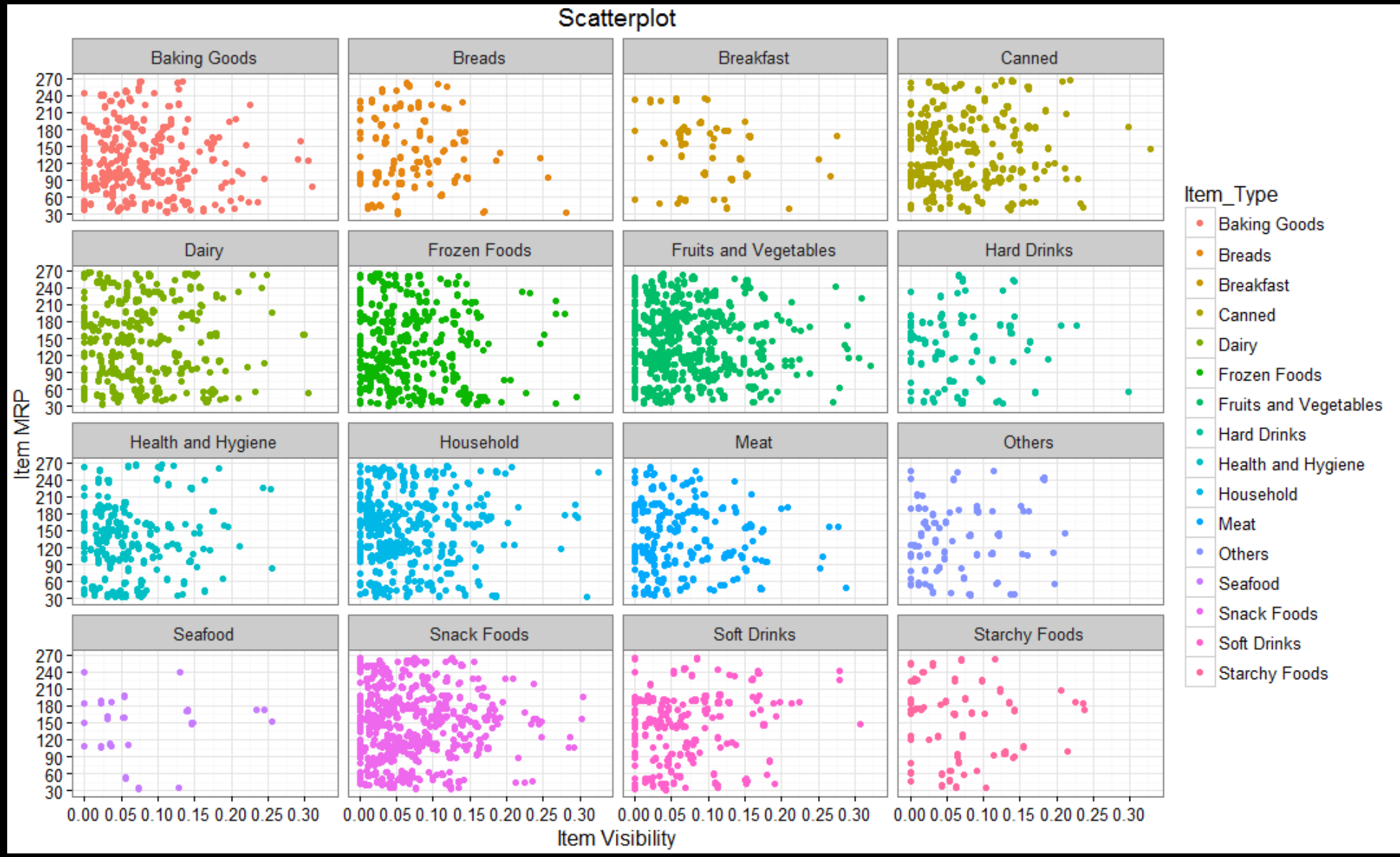
Bins & Breaks


Breaks=20
Frequency
0
40
80
18
22
26
30
BMI
Breaks=5
Frequency
0
150
300
20
25
30
BMI

Bins & Breaks


Scatterplot
Item MRP
270
240
210
180
150
120
90
60
30
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Item Visibility
Item_Type
Baking Goods
Breads
Breakfast
Canned
Dairy
Frozen Foods
Fruits and Vegetables
Hard Drinks
Health and Hygiene
Household
Meat
Others
Seafood
Snack Foods
Soft Drinks
Starchy Foods

Bins & Breaks: Small Multiples



Data Manipulation

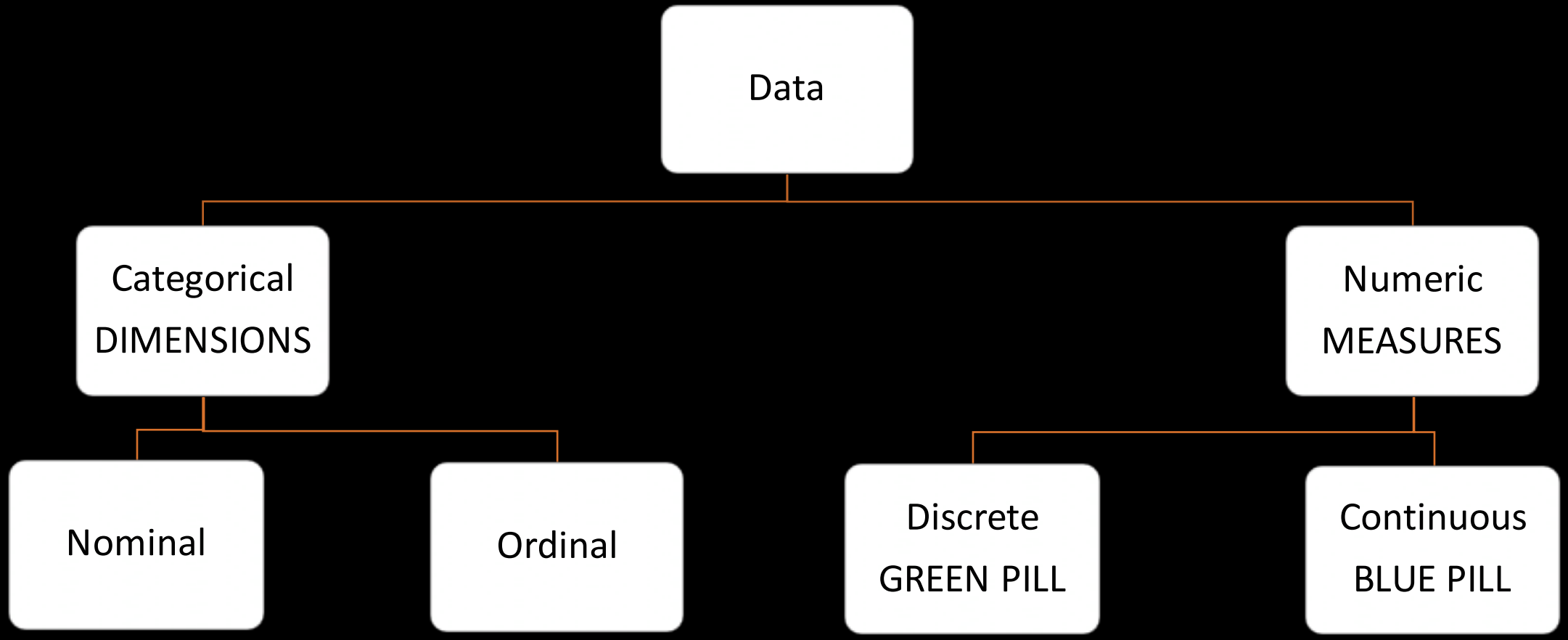
Introduction to Data Visualization

The Graduate Center at CUNY | Summer 2018

June 7, 2018

Five Minute Reflection

Data visualization, from the need for precision to the very notion of “data” is flawed. From a Humanities perspective, what can be gained by continuing to do visualizations? What is lost? Do you have any suggestions for a way forward to mediate some of the flaws?



Examples:
Occupation,
Nationality,
Labels

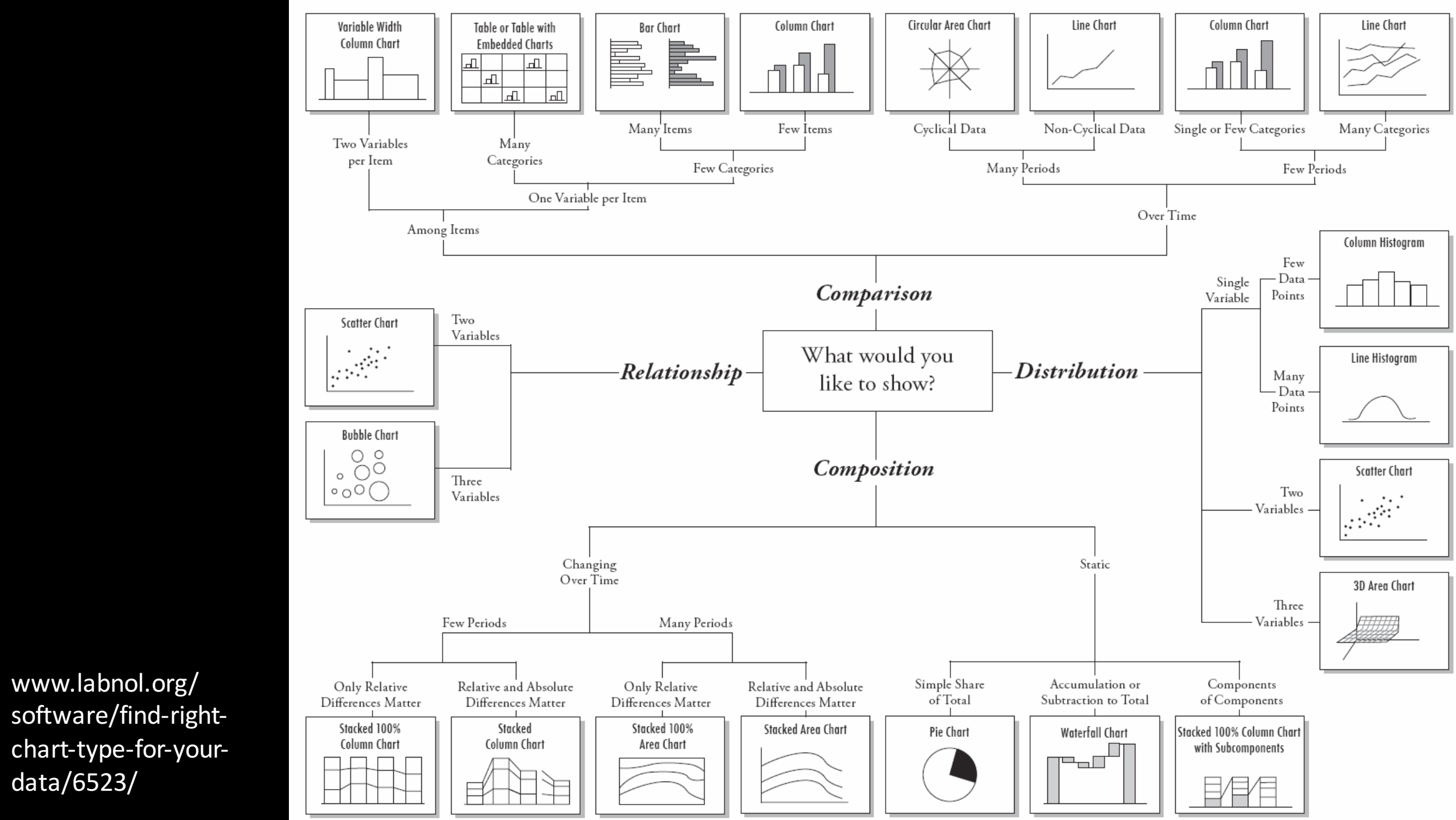
Examples:
Days of week,
Place in race,
Grades

Examples:
Children,
Books read,
Complaints

Examples:
Weight,
Distance,
(Time)

Data Type Dictates Visualization Type

- Comparison
- Distribution
- Relationship
- Composition

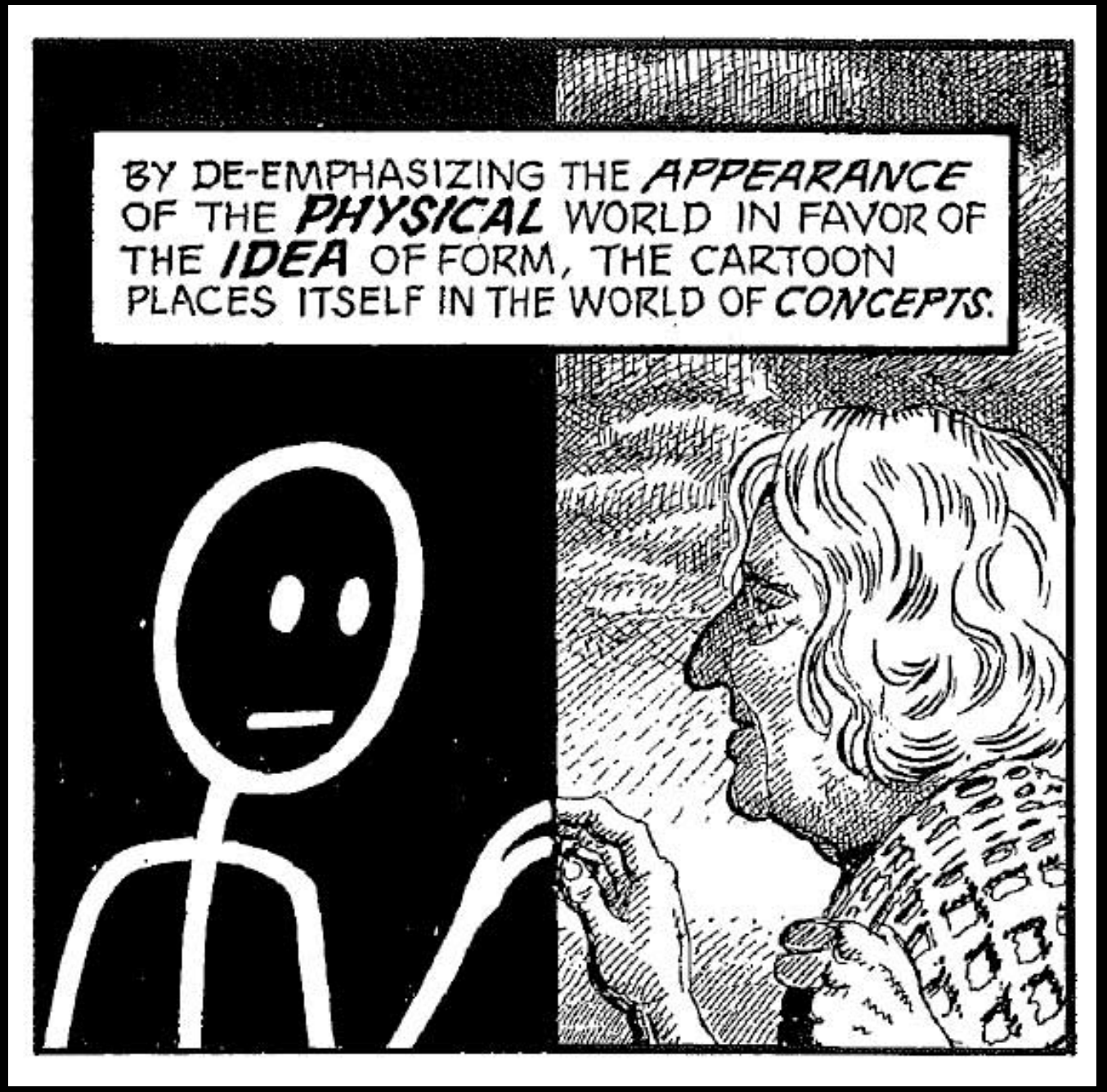


www.labnol.org/software/find-right-chart-type-for-your-data/6523/

Tools to help you pick

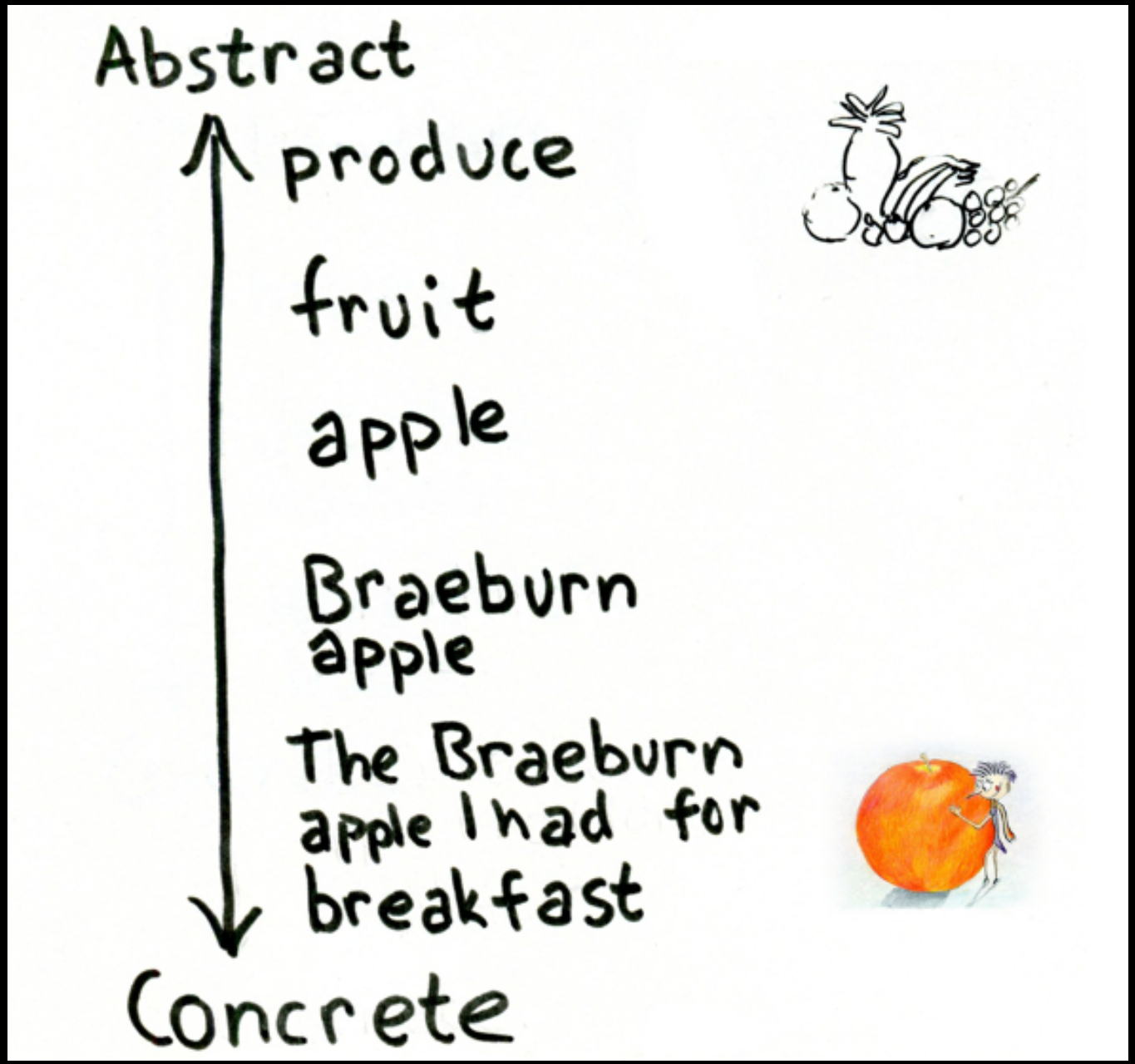
- annkemery.com/essentials
- datavizcatalogue.com

Data Visualization: Concepts or Realities



McCloud, 1994

Hiyakawa's Ladder of Abstraction



Mathematical Representations

Data itself is an abstraction from reality, but in order to make comparisons between different abstractions, they must be transformed into complimentary formats

Normalization

- Adjusting data to a “common” scale
  - Divide by total for each unit
  - i.e., Population reported as percentages is comparable, counts are not

- Shifted and scaled for comparison (log, exponent)
- Normal Distribution
- Quantiles

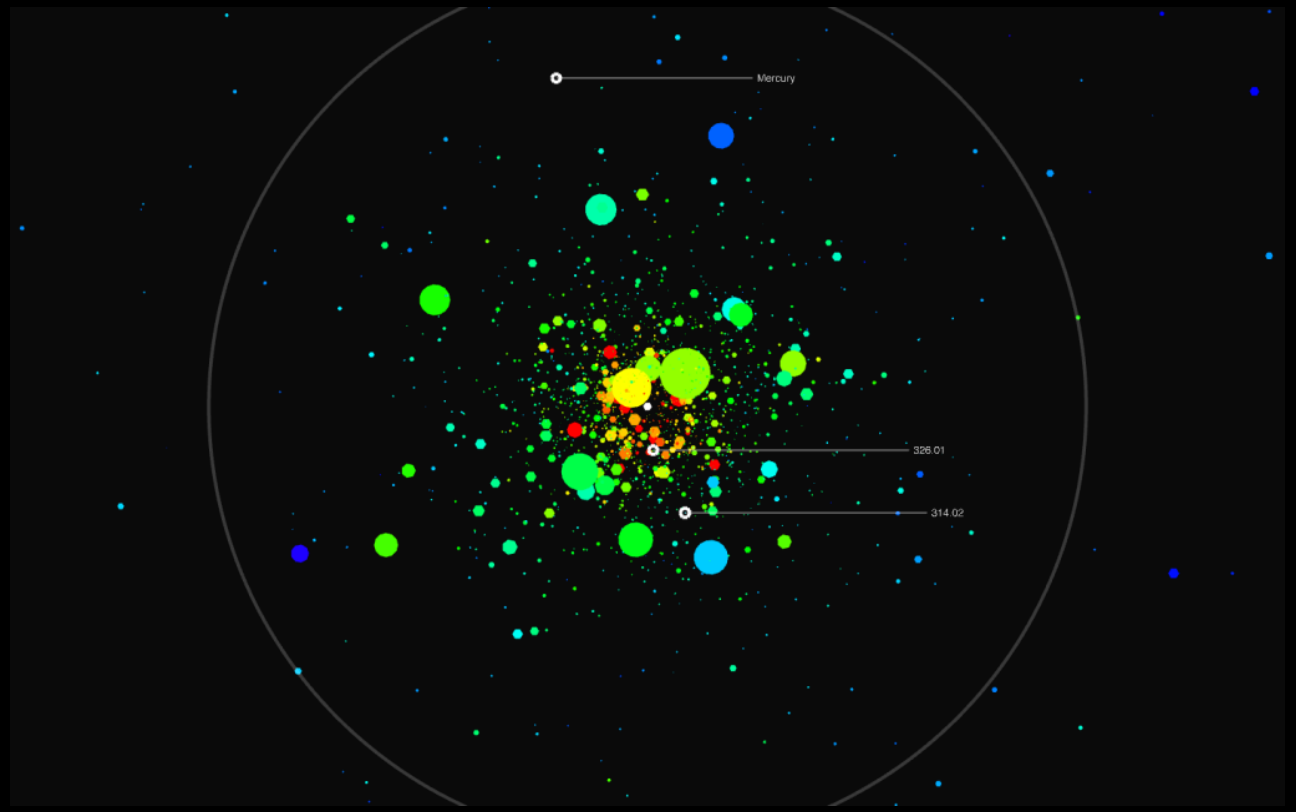
Averages

1, 3, 4, 8, 11, 11, 15

- **Mean**: simple average (7.57)
  - sum of all items ÷ total number of items
- **Median**: number in the middle (8)
- **Mode**: most common number (11)
- **Range**: 1 to 15 (14)

Context

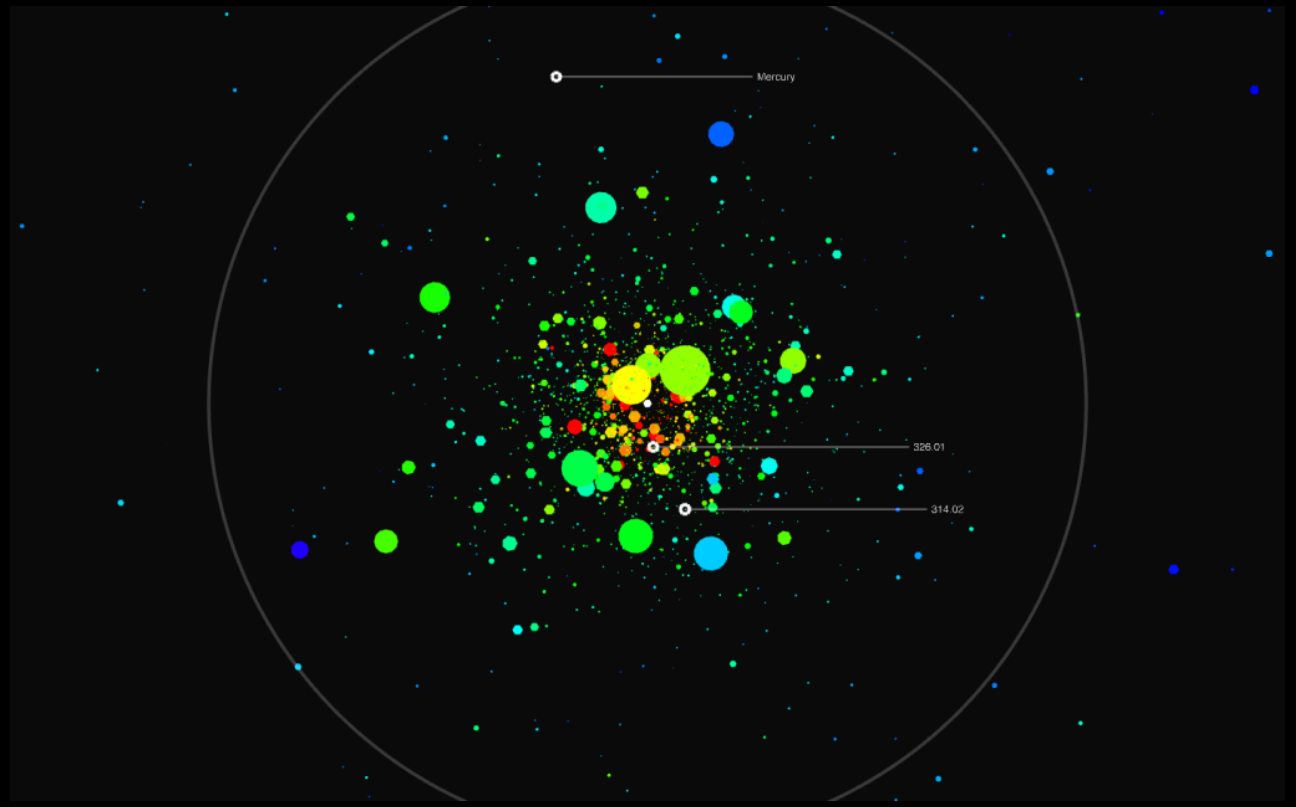
- Narrative around the data is as important as the data itself
- Allows reader to participate in your visualization
- Situates the reader so they know what to look for



Context

- Narrative around the data is as important as the data itself
- Allows reader to participate in your visualization
- Situates the reader so they know what to look for

1236 exoplanets identified by the NASA's Kepler mission.



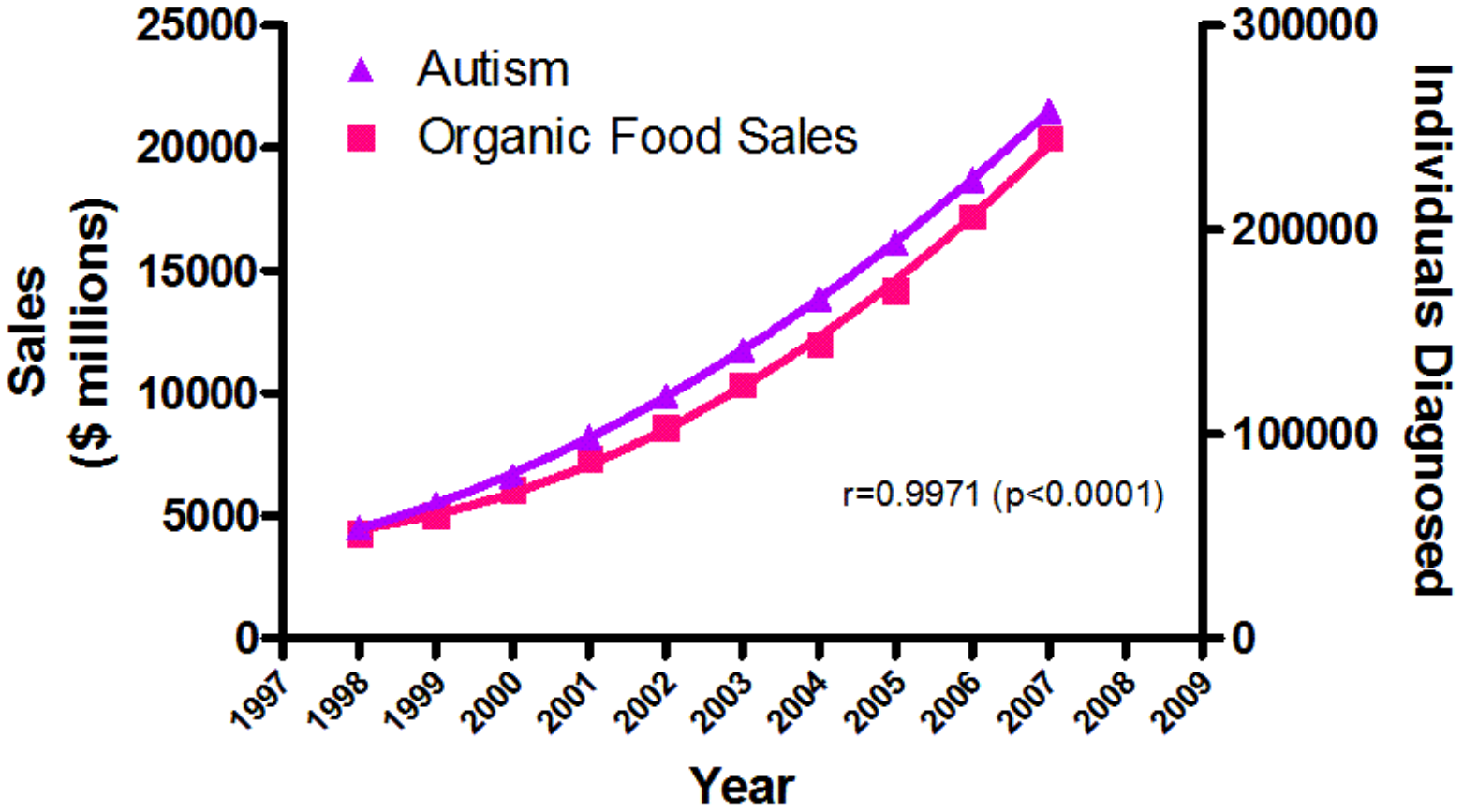
http://blog.blprnt.com/blog/blprnt/data-in-an-alien-context-kepler-visualization-source-code

Correlation is not Causation

Correlation/Causation

- Correlation:
  - two variables change at the same rate
- Causation:
  - One causes the other
  - Both are caused by an outside variable

The real cause of increasing autism prevalence?



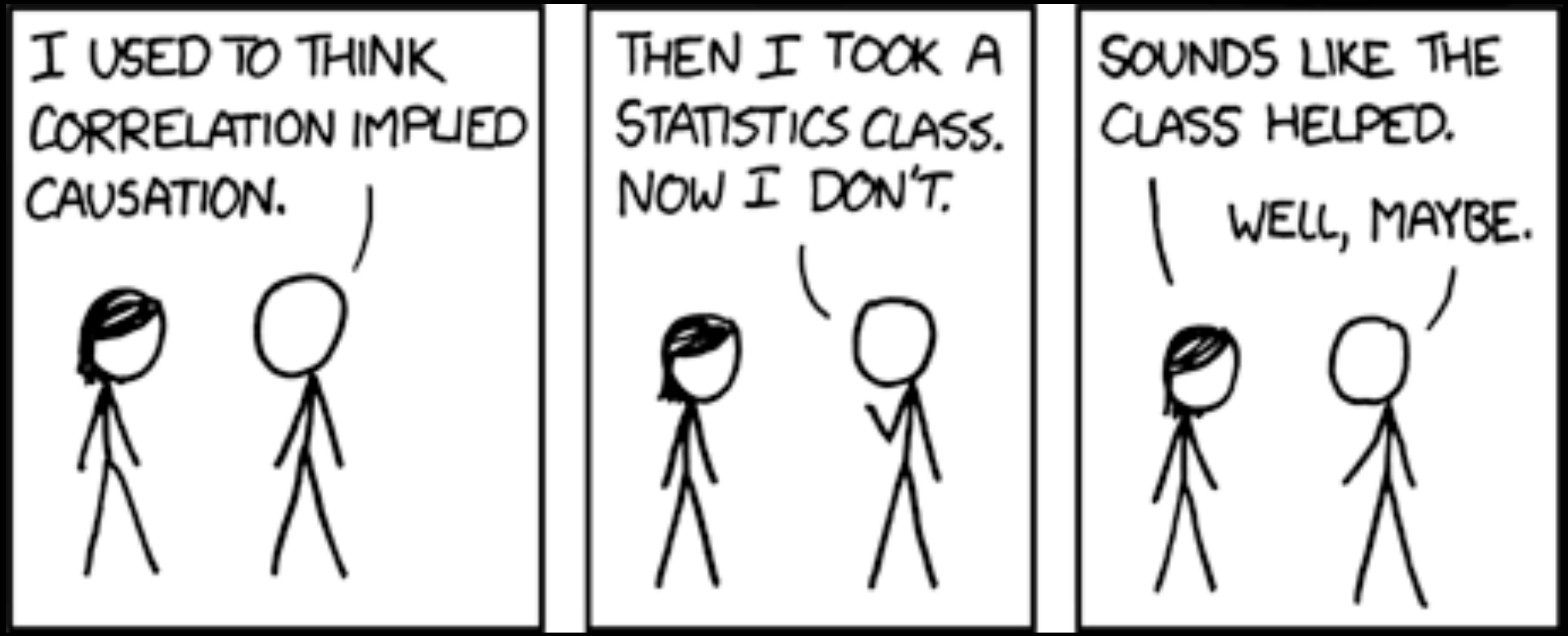
Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

Correlation/Causation Traps

- Using Proxies (particularly with social phenomena)
- Looking at data without an experiment
- Allowing pre-existing biases to guide outcome

Ways to Avoid Causation from Correlation

- Situate visualizations in context
- Think about claims from both directions
- Support claims with articles and research


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.