UNIV2V96 Transfer Research Initiative

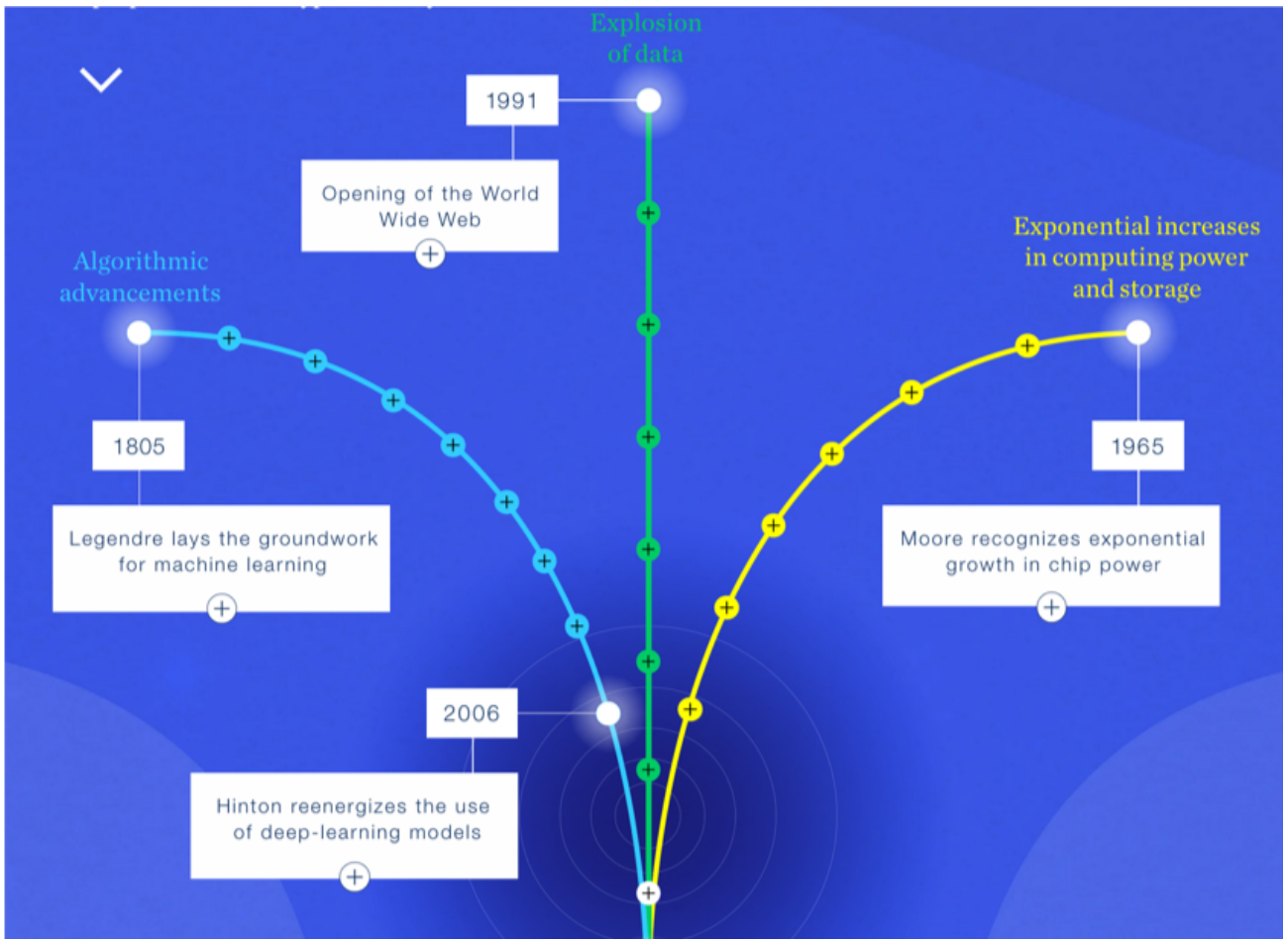
Data Science in Practice: Data Visualization

Karl Ho
School of Economic, Political and Policy Sciences
University of Texas at Dallas

Overview:

1. Why Data Science? Why now?
2. Data fluency
3. Types of Data Science
4. Data Science Roadmap
5. Data Visualization
    1. How to make a different chart?
    2. File formats to know (to become a computer graphic expert)
    3. Seeing data, visual thinking
    4. Cognitive Science: how to create a chart

Why Data Science? Why now?



McKinsey & Co., An Executive's Guide to AI

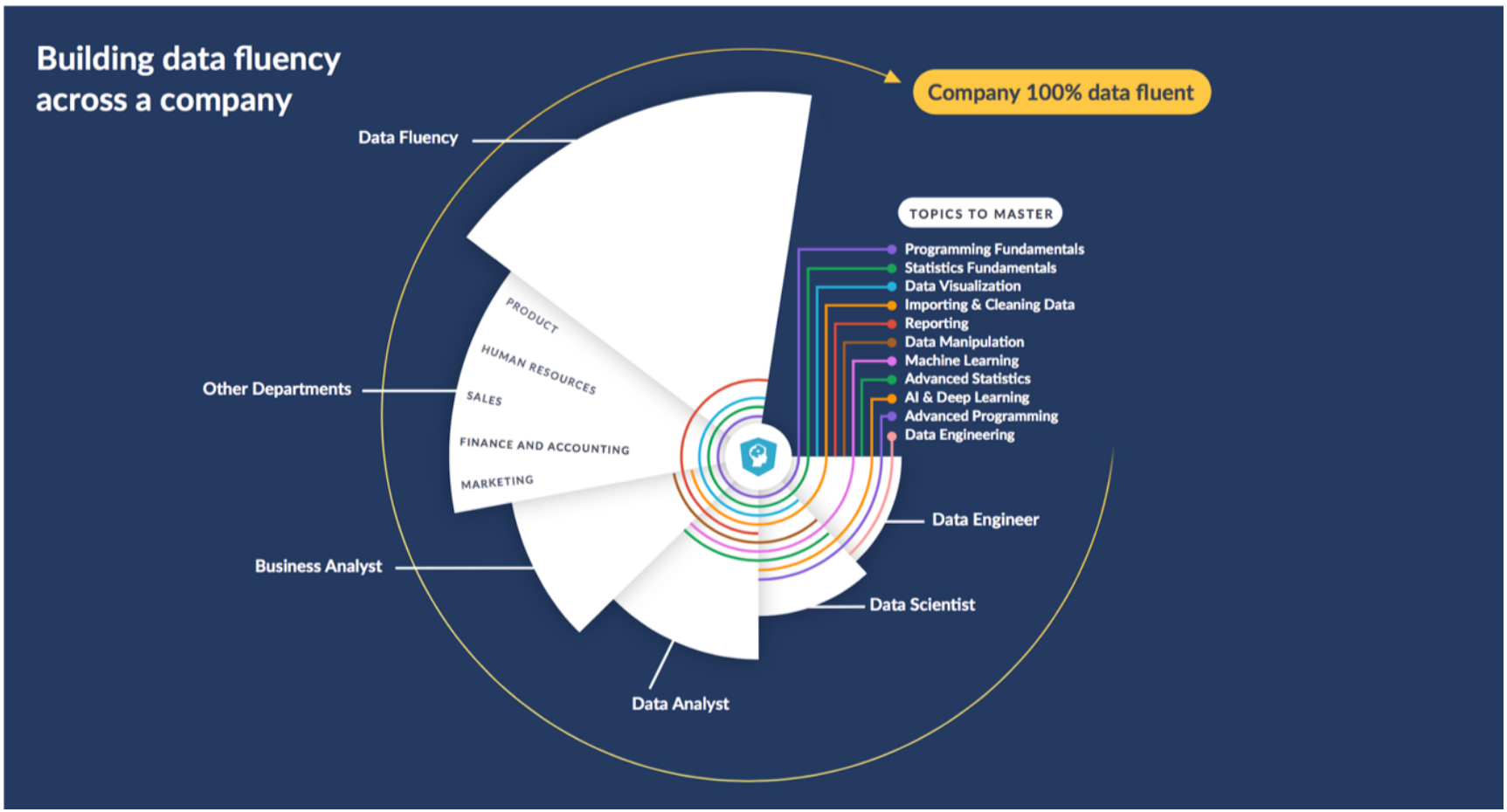
A Minute on the Internet in 2019

Estimated data created on the internet in one minute

Sources: Lori Lewis & Officially Chad via Visual Capitalist

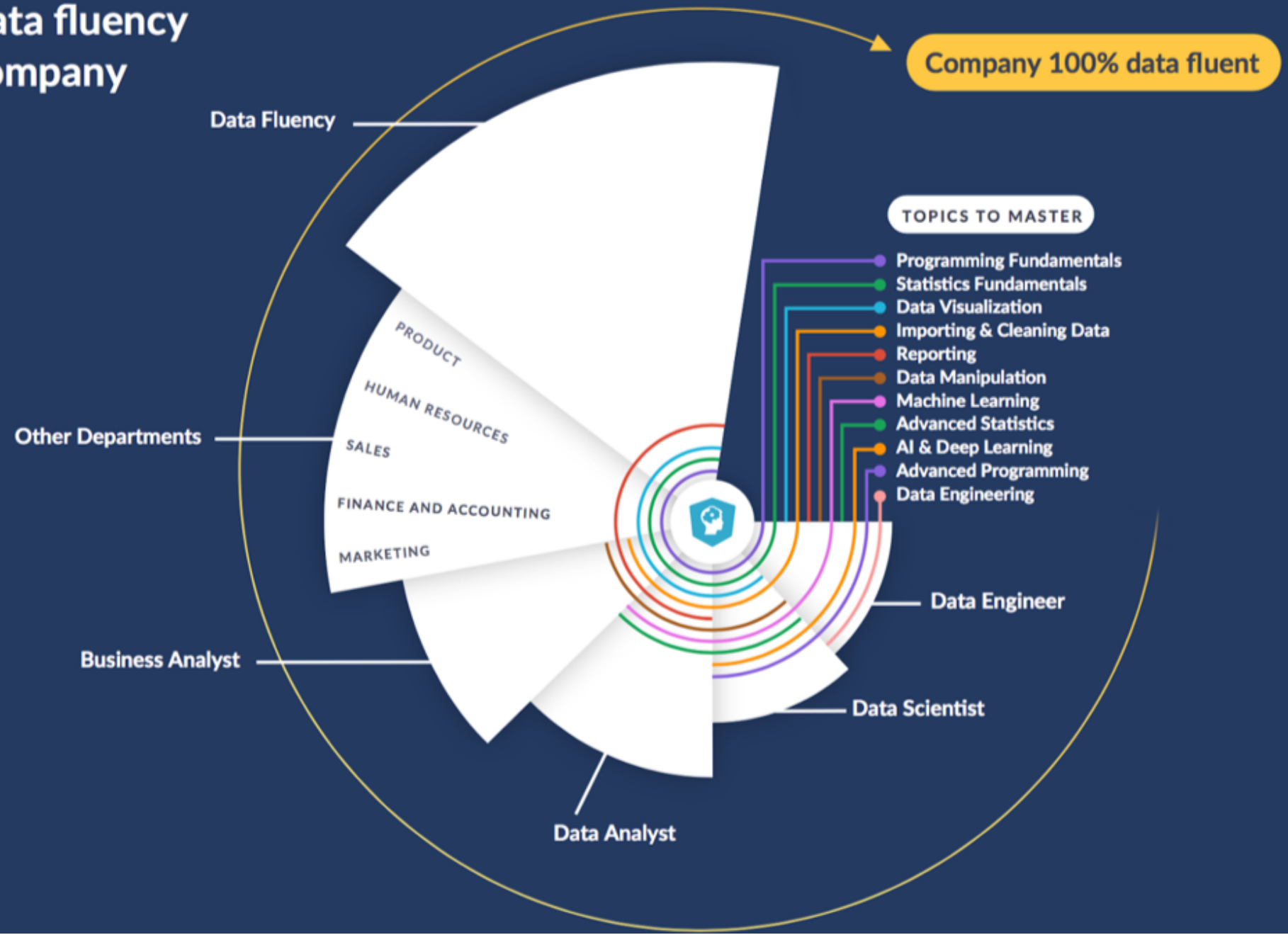
statista

Data fluency



Hugo Bowne-Anderson. 2019. "What 300 L&D leaders have learned about building data fluency"

ata fluency
ompany



Data fluency

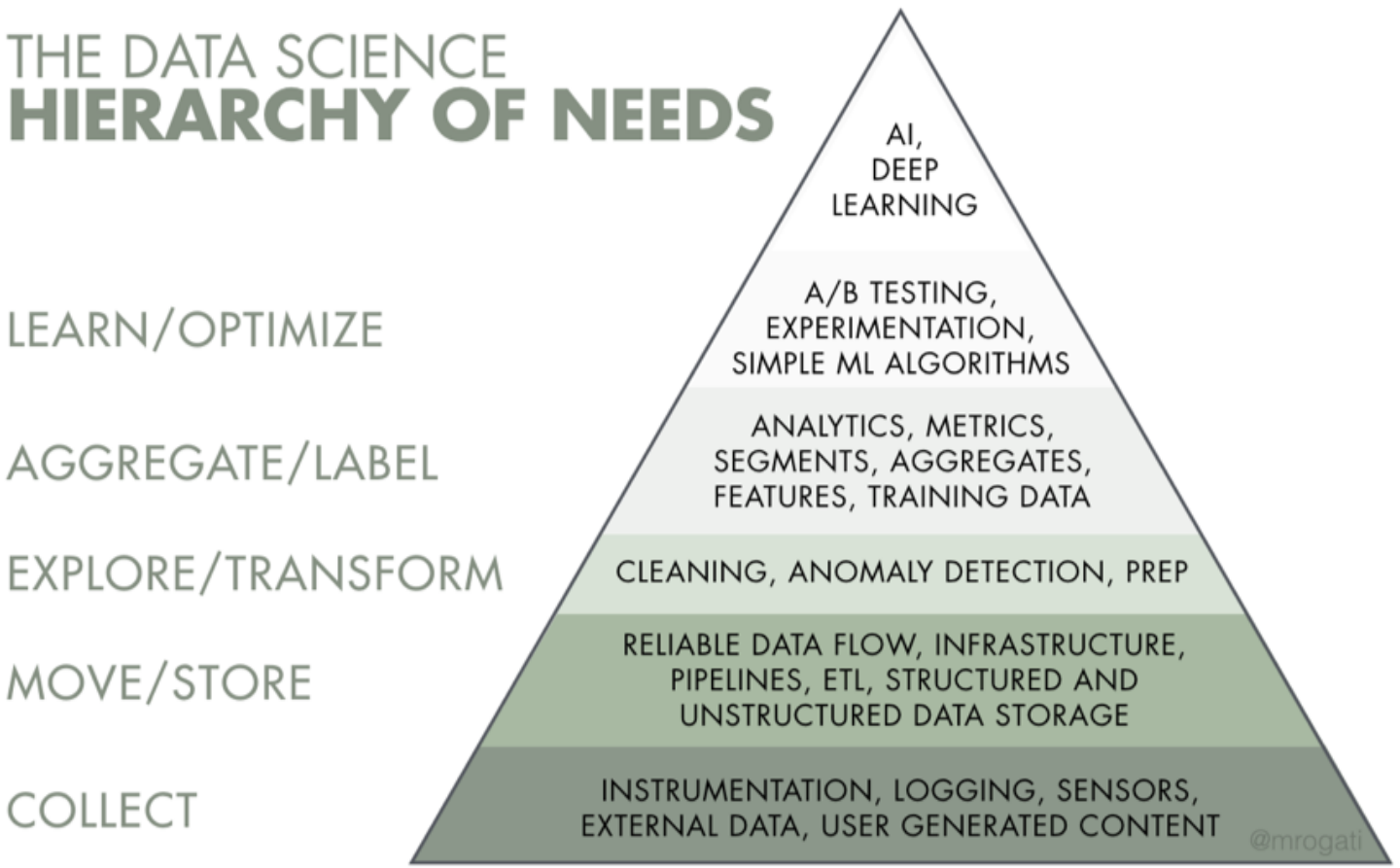
Everybody has the data skills and literacy to understand and perform data driven documents and tasks

Danger of immature data fluency

Types of Data Science

1. Business intelligence (Descriptive analytics)
2. Machine learning (Predictive analytics)
3. Decision making (prescriptive analytics)

Rogati AI hierarchy of needs



Data Science Roadmap

1. Introduction - Data theory
2. Data methods
3. Statistics
4. Programming
5. Data Visualization
6. Information Management
7. Data Curation
8. Spatial Models and Methods
9. Machine Learning
10. NLP/Text mining

Graphic file formats

Raster	Vector
JPEG	SVG
GIF	EPS
PNG	PDF
BMP	

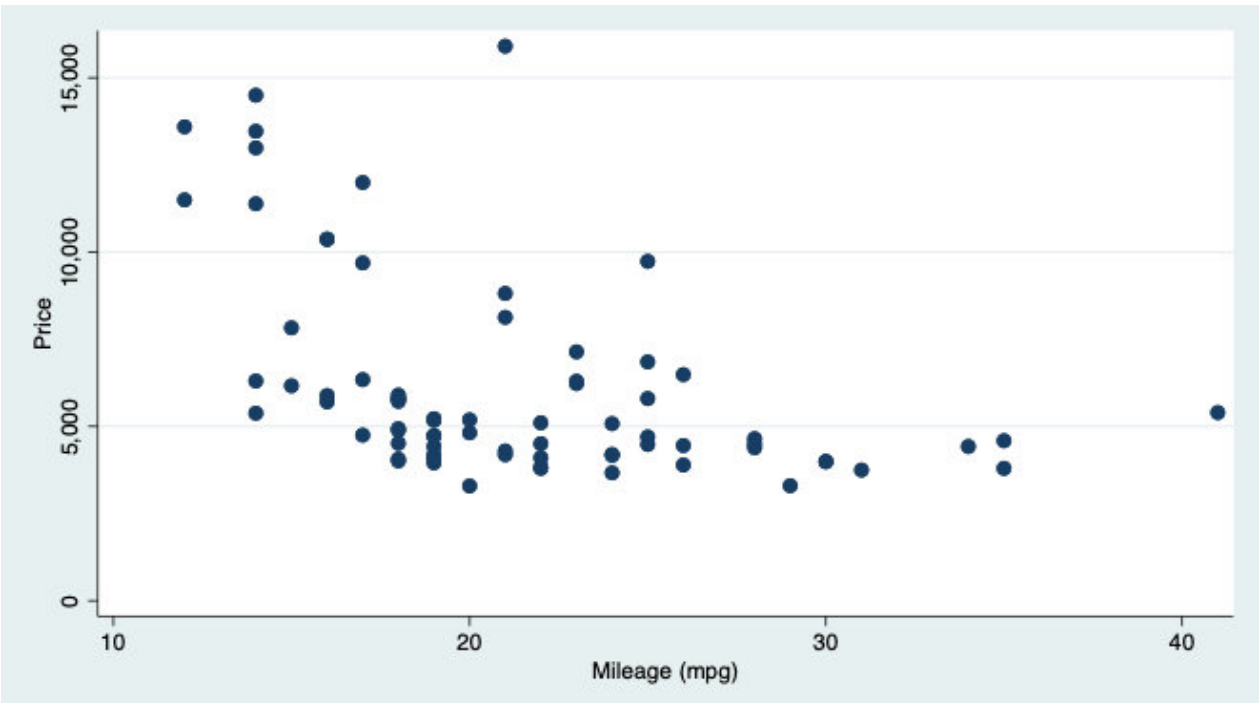
File format choice in Data Visualization

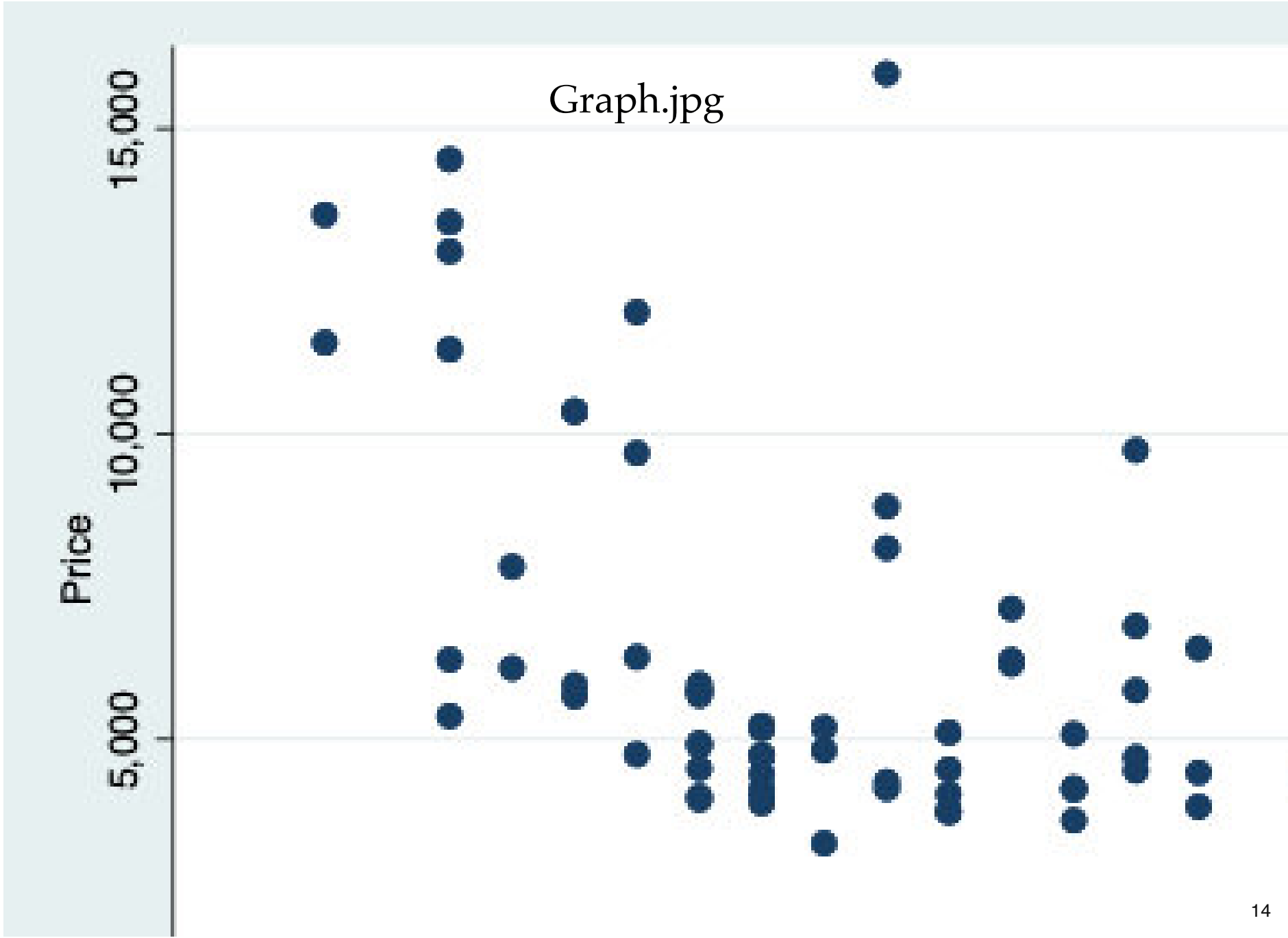
Data scientists specialized in data visualization usually prefer vector files in particular SVG or PDF. Quality is the huge factor and specialists will not tolerate pixelized edges or staggered graphics when the chart or graphic is zoomed in.

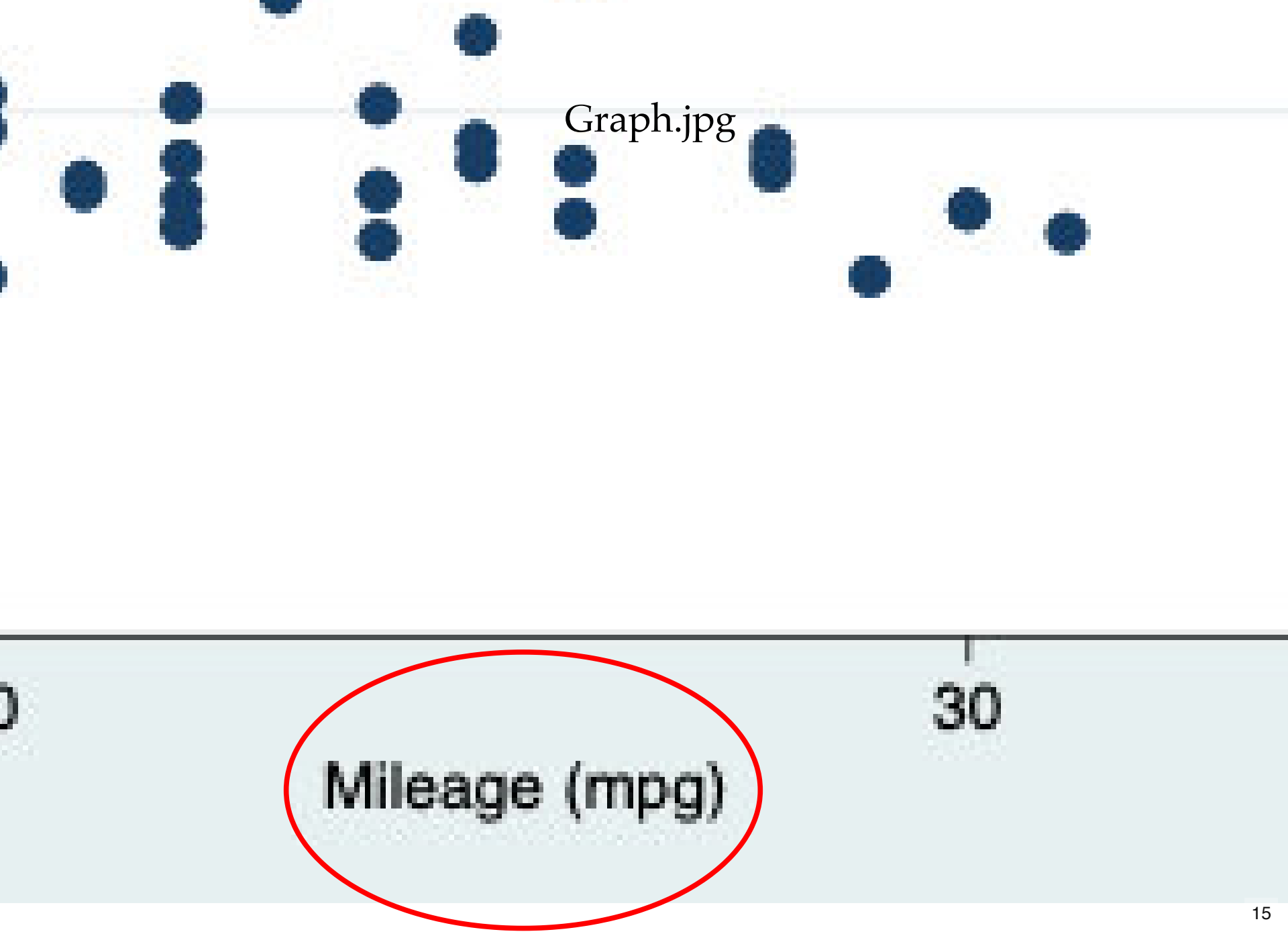
R, Stata and most dataviz programs allow options for file exporting or specification for resolutions. SVG and PDF should be the first choices unless resolutions can be set at above 300 or even 600.

File format makes a huge difference!

Graph.jpg

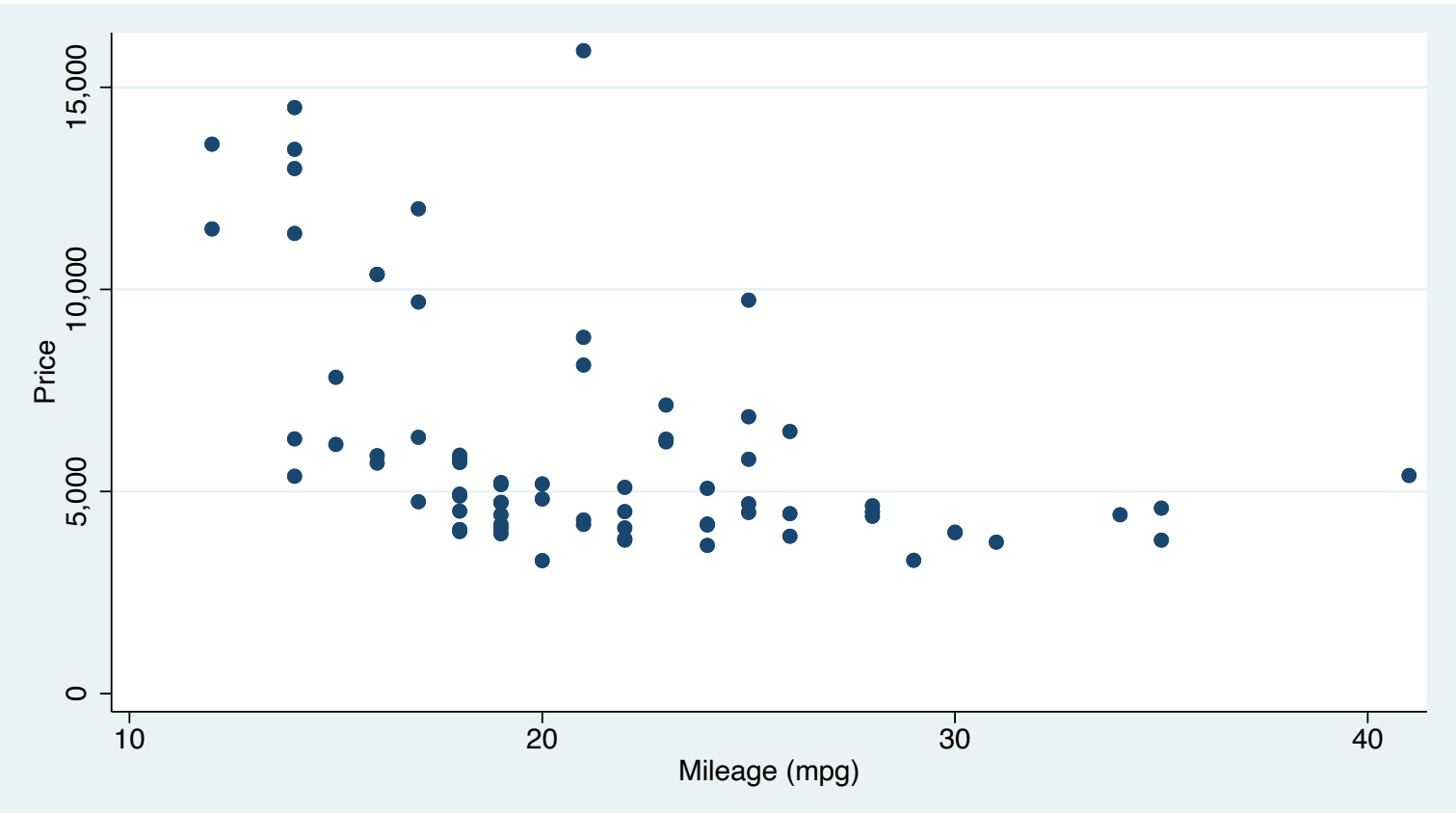


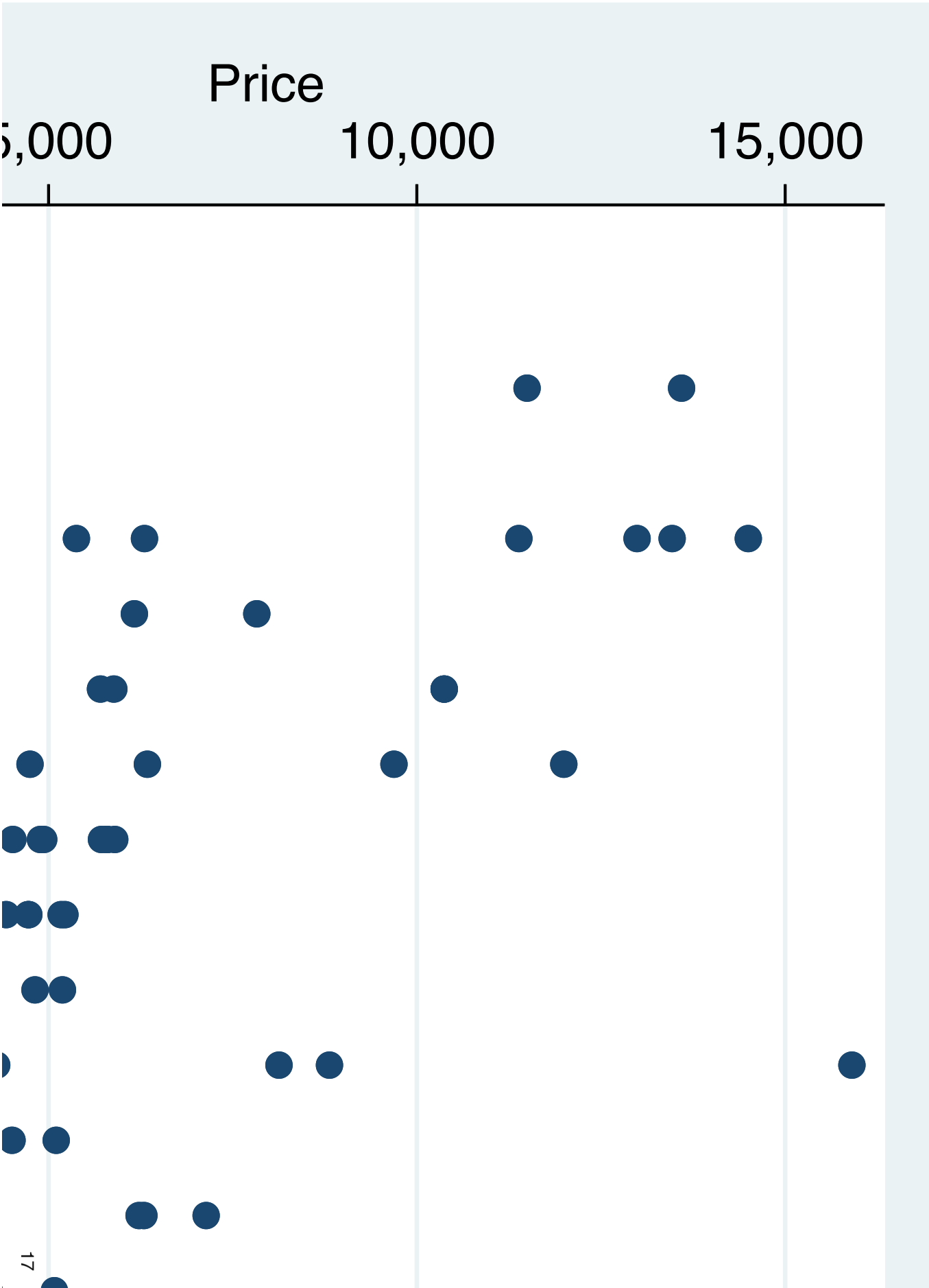

Graph.jpg
Price
5,000
10,000
15,000


Graph.jpg
30
Mileage (mpg)

File format makes a huge difference!

Graph.svg




Price
,000
10,000
15,000

15,000
10,000

What is Complexity?

Source: Ammer, Ralph. 2017. "Make me think!" *Medium* https://blog.prototypr.io/make-me-think-90b46aa50513

What is Simplicity?

Source: Ammer, Ralph. 2017. "Make me think!" *Medium* https://blog.prototypr.io/make-me-think-90b46aa50513

Should the technology grow?

Or should the user?

- Ammer 2018

Cognitive Scientist: The Visualization Process

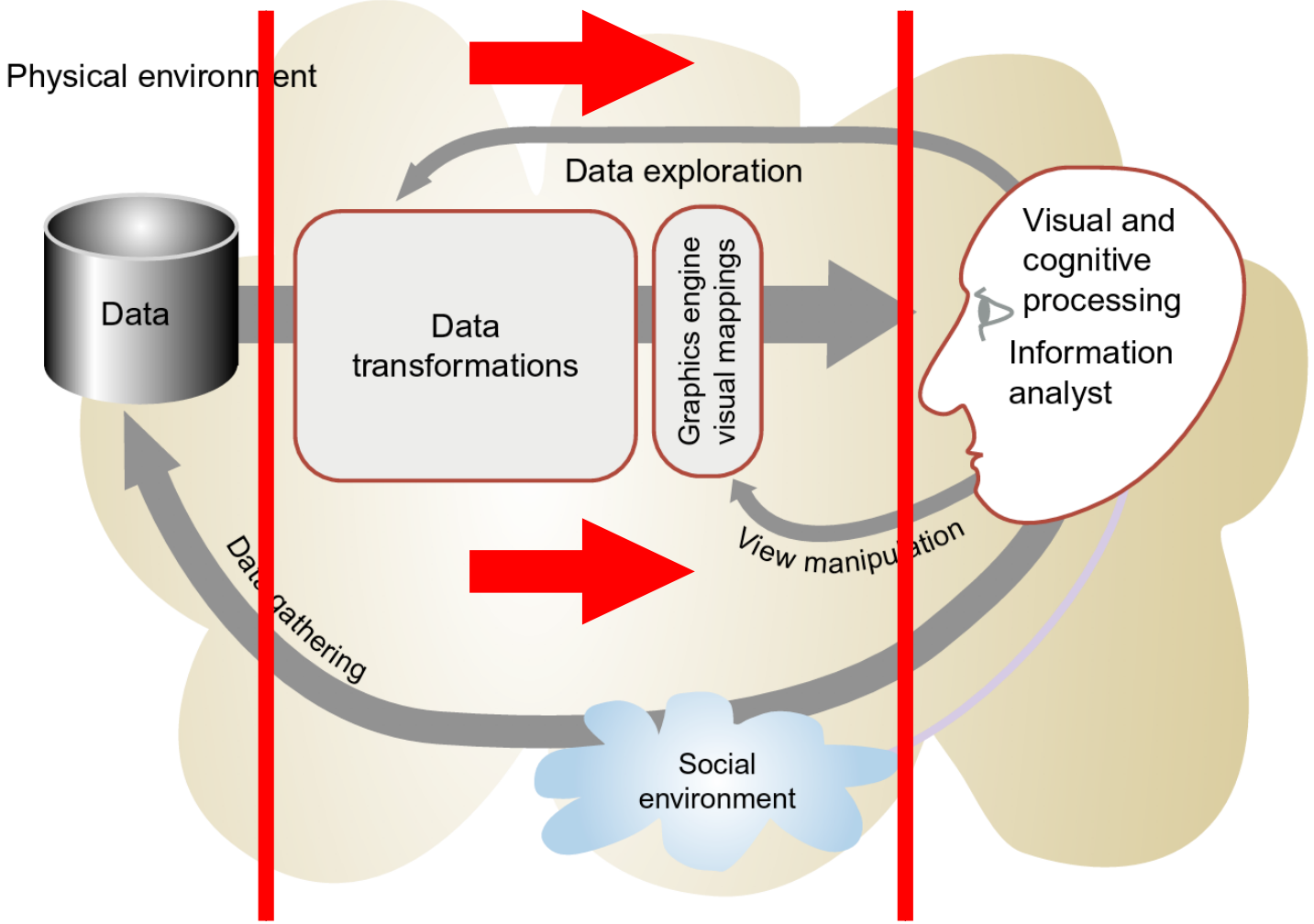


Figure 1.2 The visualization process.

Source: Ware, Colin 2012.

Visualization Stages

1. The collection and storage of data.
2. A preprocessing stage designed to transform the data into something that is easier to manipulate.
3. Mapping from the selected data to a visual representation, which is accomplished through computer algorithms that produce an image on the screen.
4. The human perceptual and cognitive system (the perceiver).

Cognitive Science Guidelines

1. Design graphic representations of data by taking into account human sensory capabilities in such a way that important data elements and data patterns can be quickly perceived.
2. Important data should be represented by graphical elements that are more visually distinct than those representing less important information.

Cognitive Science Guidelines

3. Greater numerical quantities should be represented by more distinct graphical elements.
4. Graphical symbol systems should be standardized within and across applications.
5. Where two or more tools can perform the same task, choose the one that allows for the most valuable work to be done per unit time.

Cognitive Science Guidelines

6. Consider adopting novel design solutions only when the estimated payoff is substantially greater than the cost of learning to use them.
7. Unless the benefit of novelty outweighs the cost of inconsistency, adopt tools that are consistent with other commonly used tools.

Cognitive Science Guidelines

8. Effort spent on developing tools should be in proportion to the profits they are expected to generate. This means that small-market custom solutions should be developed only for high-value cognitive work.

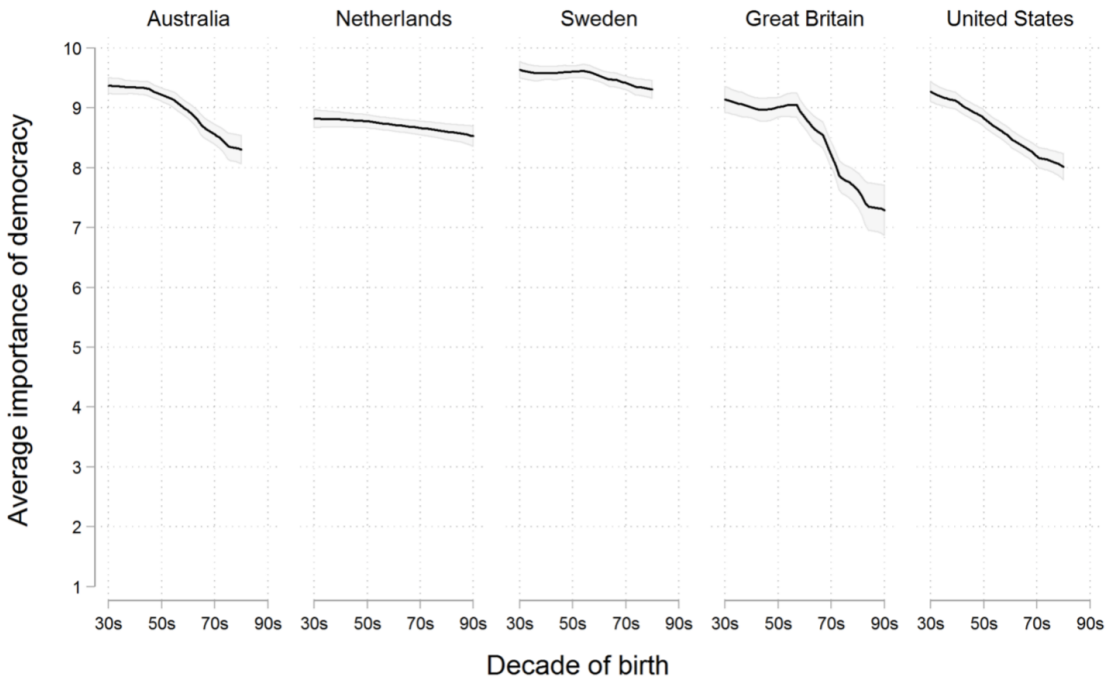
Cognitive Science

- Very important foundation for research and design of visualization of data
- Humans have very similar visual systems.
- Visual system is tuned to receive data in certain ways.

- Perceptions can be learnt.

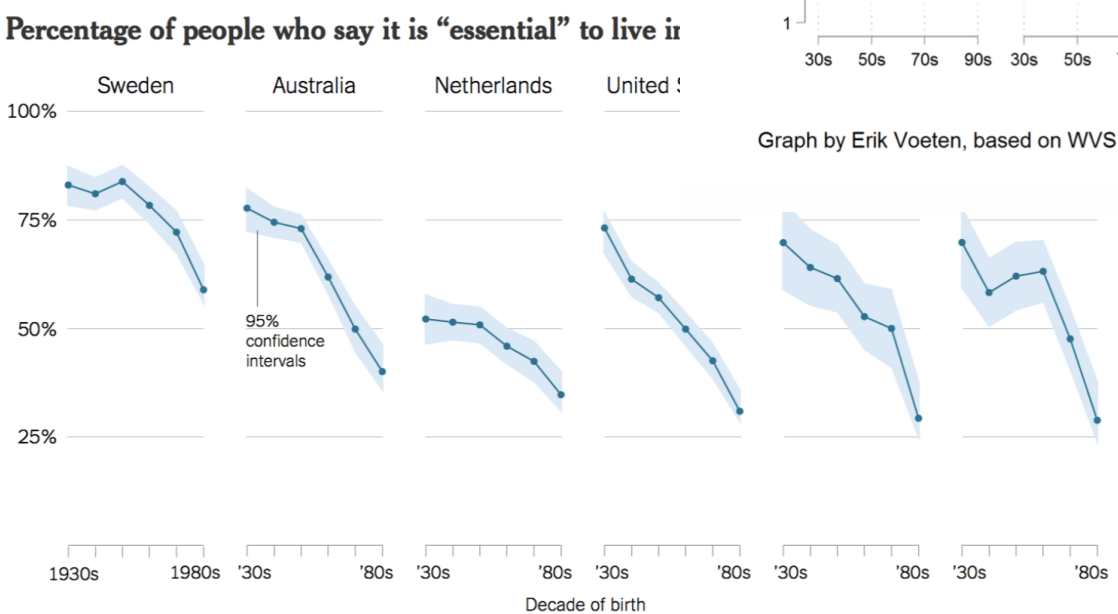
Scale matters!

Figure 1.9: Perhaps the crisis has been overblown. (Erik Voeten.)



Graph by Erik Voeten, based on WVS 5

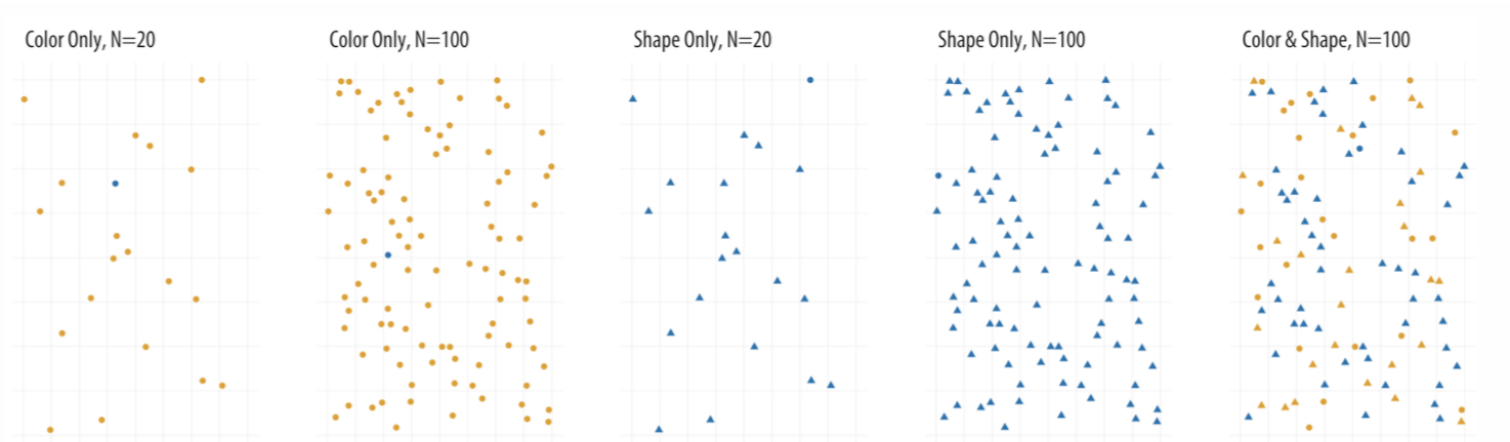
Figure 1.8: A crisis of faith in democracy? (New York



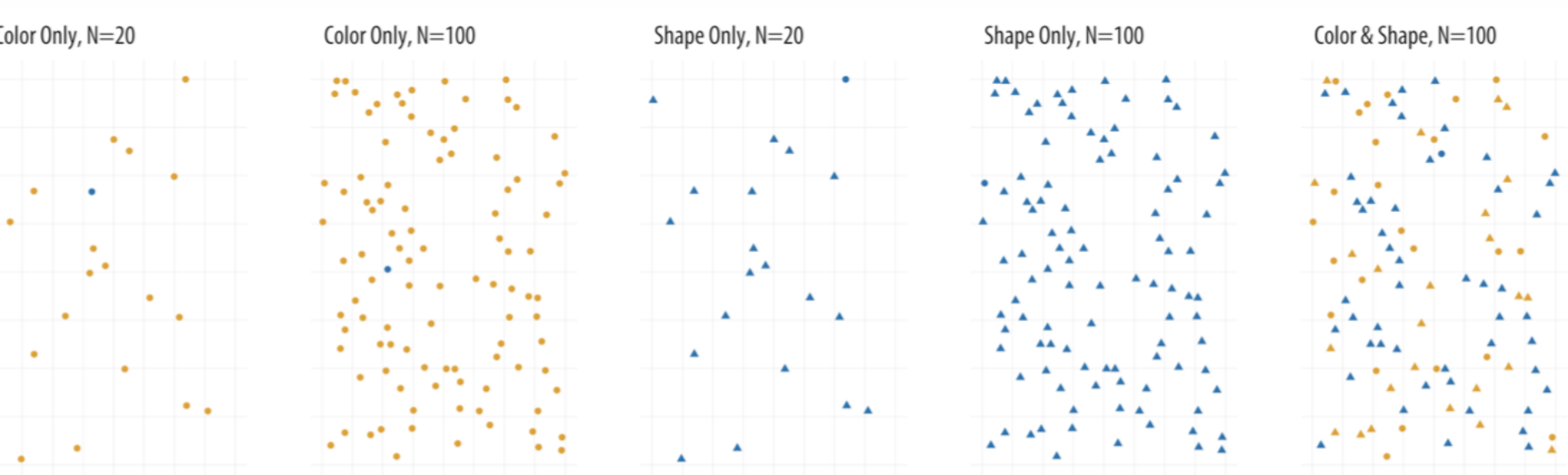
Source: Yascha Mounk and Roberto Stefan Foa, "The Signs of Democratic Deconsolidation," Journal of Democracy | By The New York Times

Pre-attentive pop-out

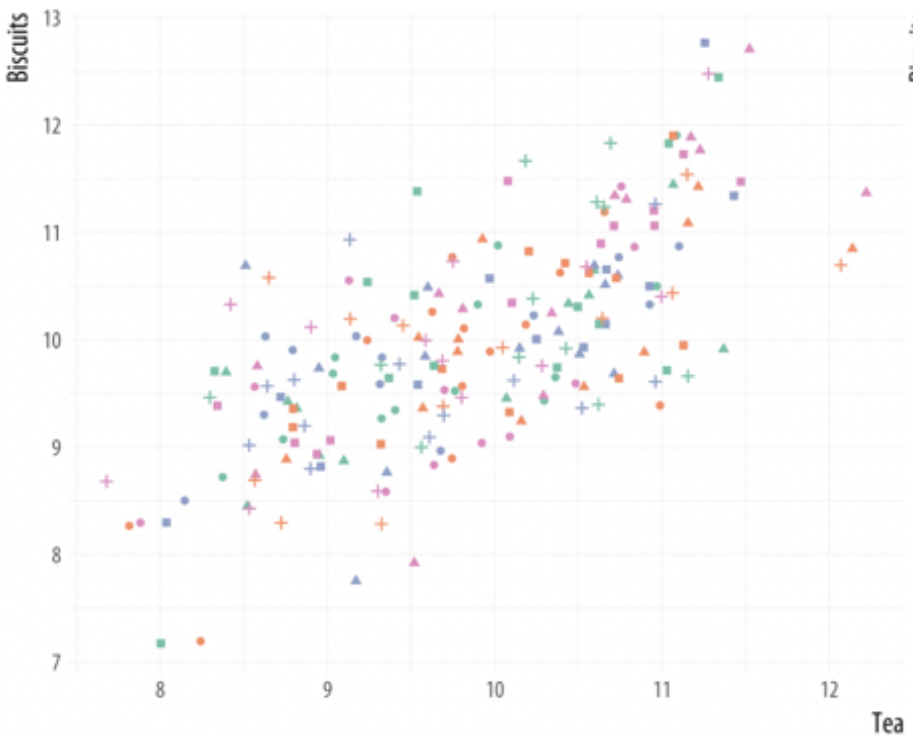
Some objects in our visual field are easier to see than others (e.g. color, shapes).

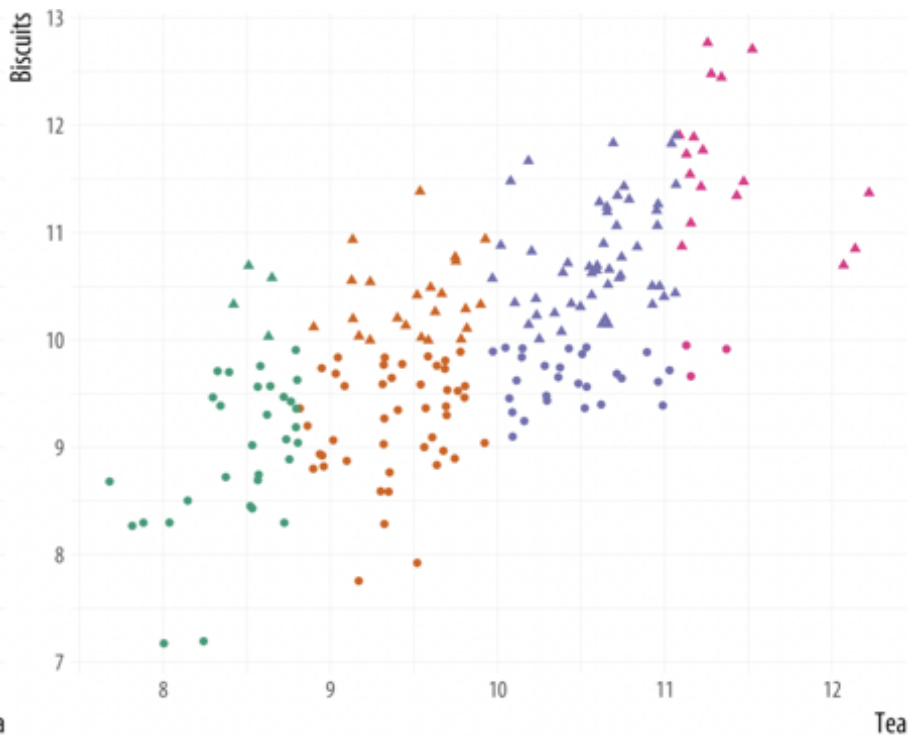


Pre-attentive pop-out



Pre-attentive pop-out



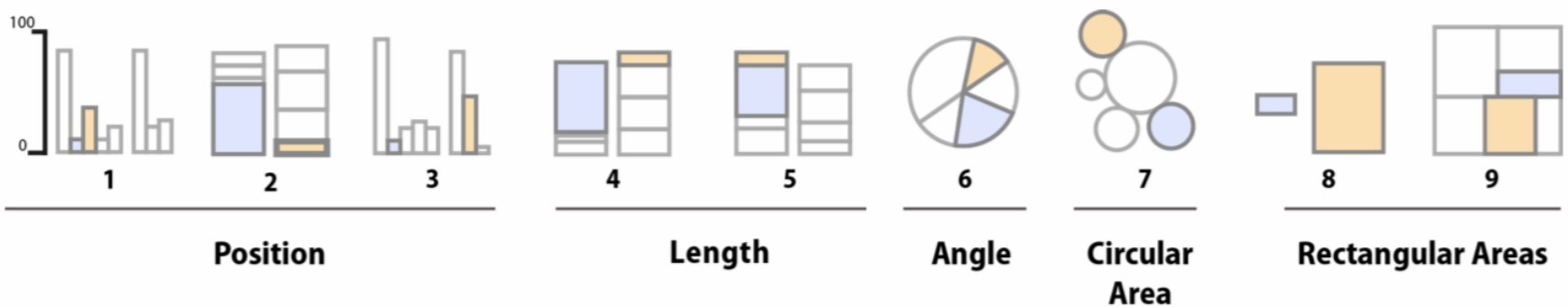


Even harder. Why?

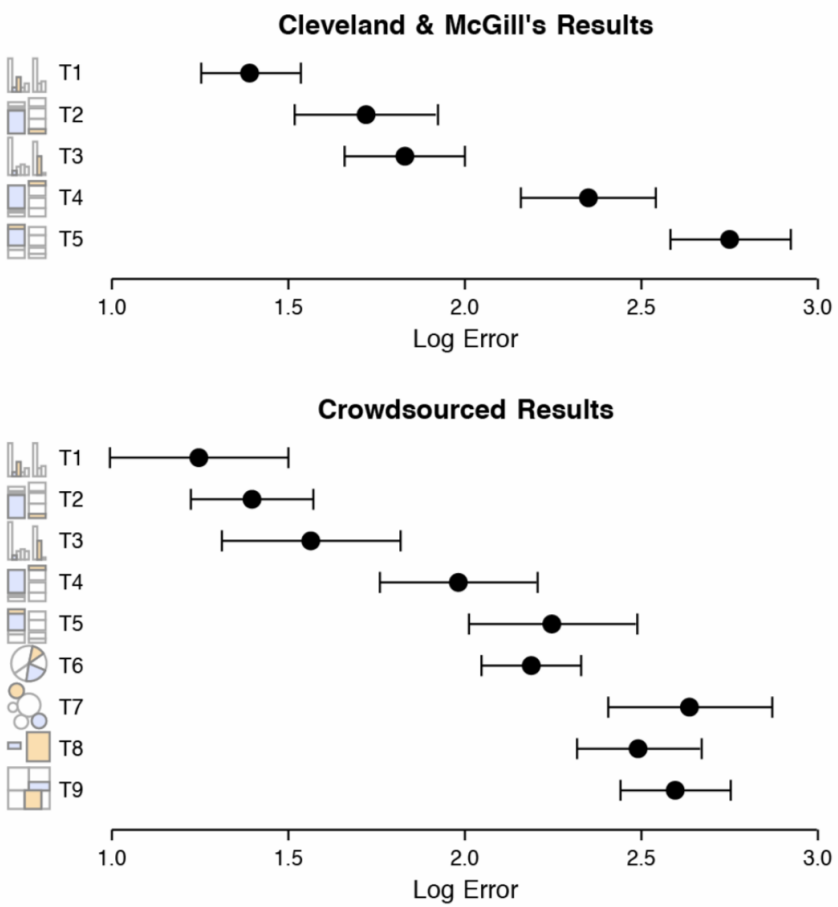
Cleveland and McGill Experiments

In the 1980s, William S. Cleveland and Robert McGill conducted some experiments identifying and ranking theses tasks for different types of graphics.

Cleveland and McGill Experiments



Cleveland and McGill Experiments



Comparison:

- relative position
- on a common scale
- by length
- by angles
- by areas
- by volume

Miller's Law

"The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information"

- George Miller 1956

What do Cognitive Scientist teach us about visuals?

Not all graphical elements are born equal, or equally clear to our eyes.

Humans are much better at judging line length than angle or grayscale.

- Cleveland & McGill 1987

Graphical elements used to encode data:

1. More accurate
    1. Position on a plane
    2. Line length
    3. Angle & slope
    4. Area
    5. Volume
2. Less Accurate
    1. Color

- Adoph, Cleveland & McGill 1987

Graphical elements used to encode data:

1. More accurate
    1. Position on a plane
    1. Position on a plane
    2. Line length
    3. Angle & slope
    4. Area
    5. Volume
2. Less Accurate
    1. Color
    1. Color

but not multiple.

Single color can stand out, ...

- Adoph, Cleveland & McGill 1987

Graphical elements used to encode data:

Adolph and Cleveland:

- Use more visible elements (line, points on plane) for important variables
- Use color for qualitative data (e.g. countries)

Angular data encoding

Which band

is

what?

Source: Cleveland 1985

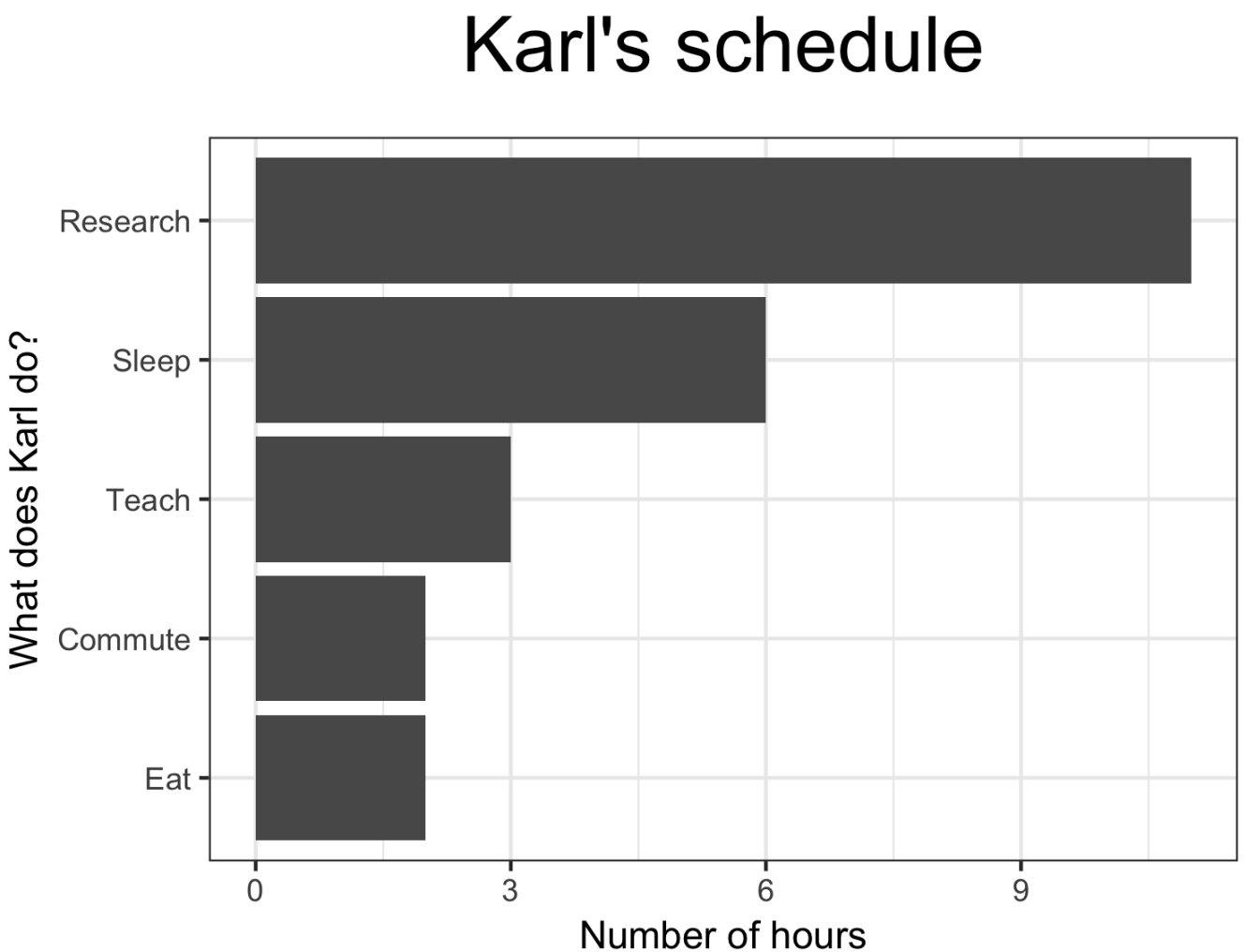
Angular data encoding

Which band
is
what?

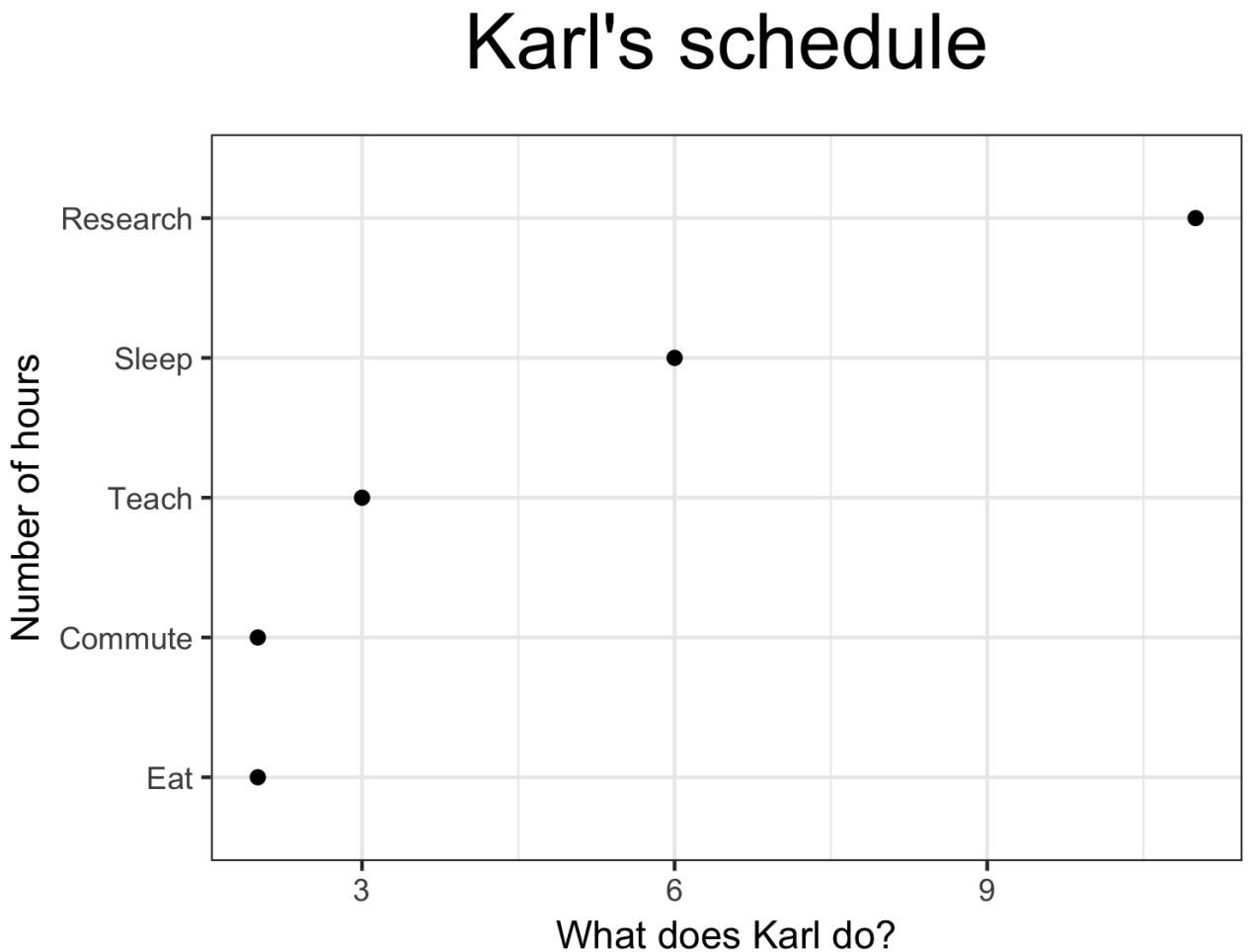
https://www.utdallas.edu/~kyho/present/googlecharts/piechart/

Source: Google charts based on JavaScript

Better


Karl's schedule
What does Karl do?
Research
Sleep
Teach
Commute
Eat
0
3
6
9
Number of hours

Better better



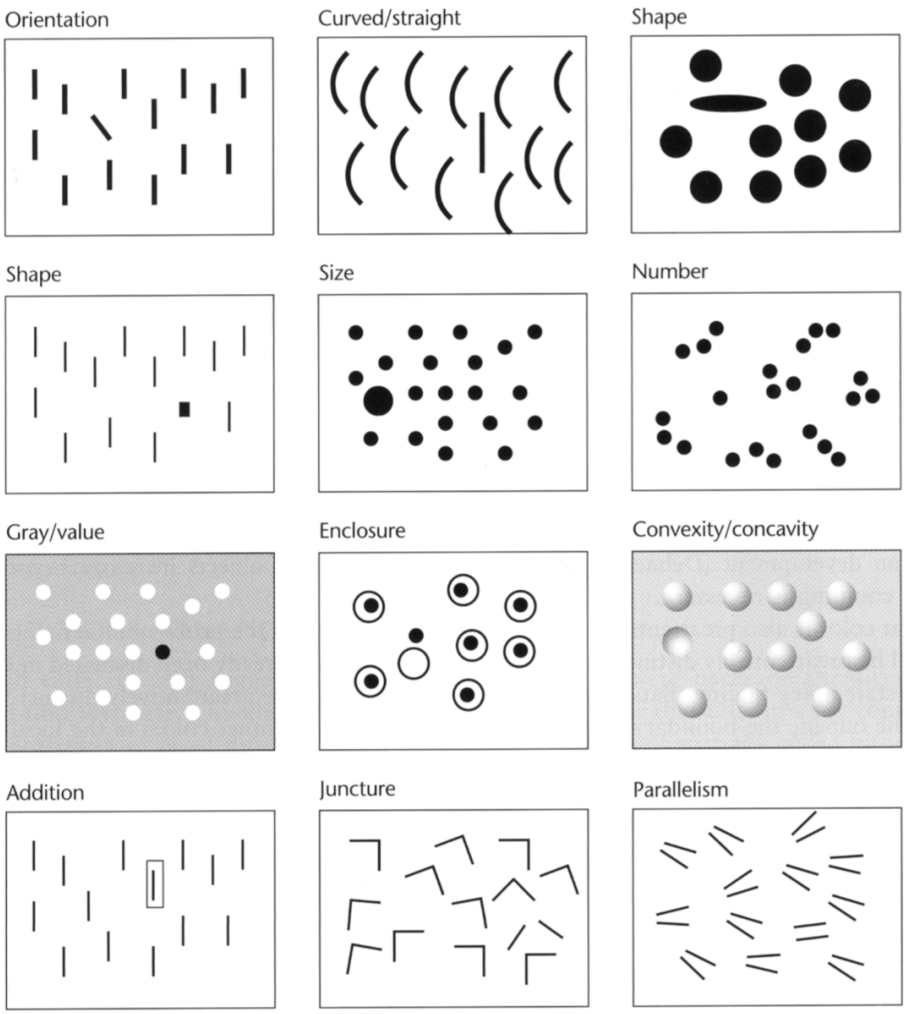
Find the 3's

85689726984689762689764358922659865986554897689269898
02462996874026557627986789045679232769285460986772098
90834579802790759047098279085790847729087590827908754
98709856749068975786259845690243790472190790709811450
85689726984689762689764458922659865986554897689269898

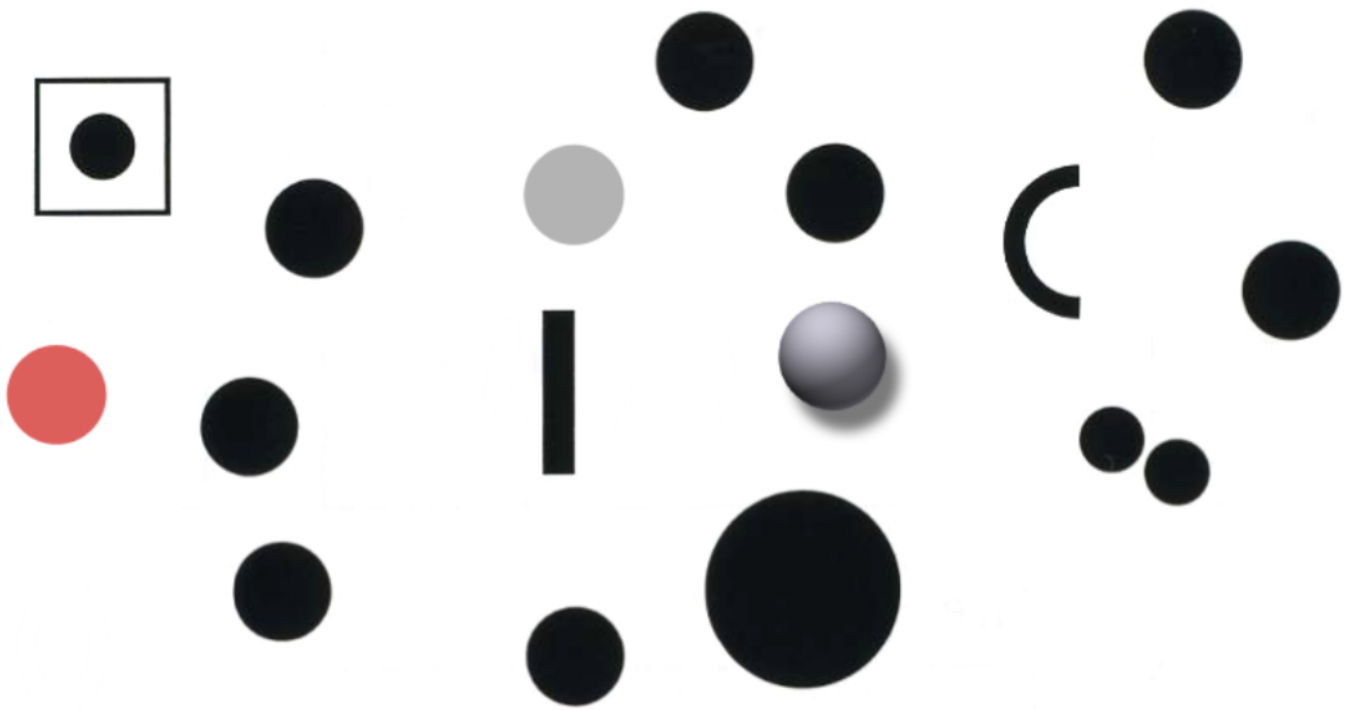
Our brains process color differences "pre-attentively" – fast & effortlessly.

- Colin Ware

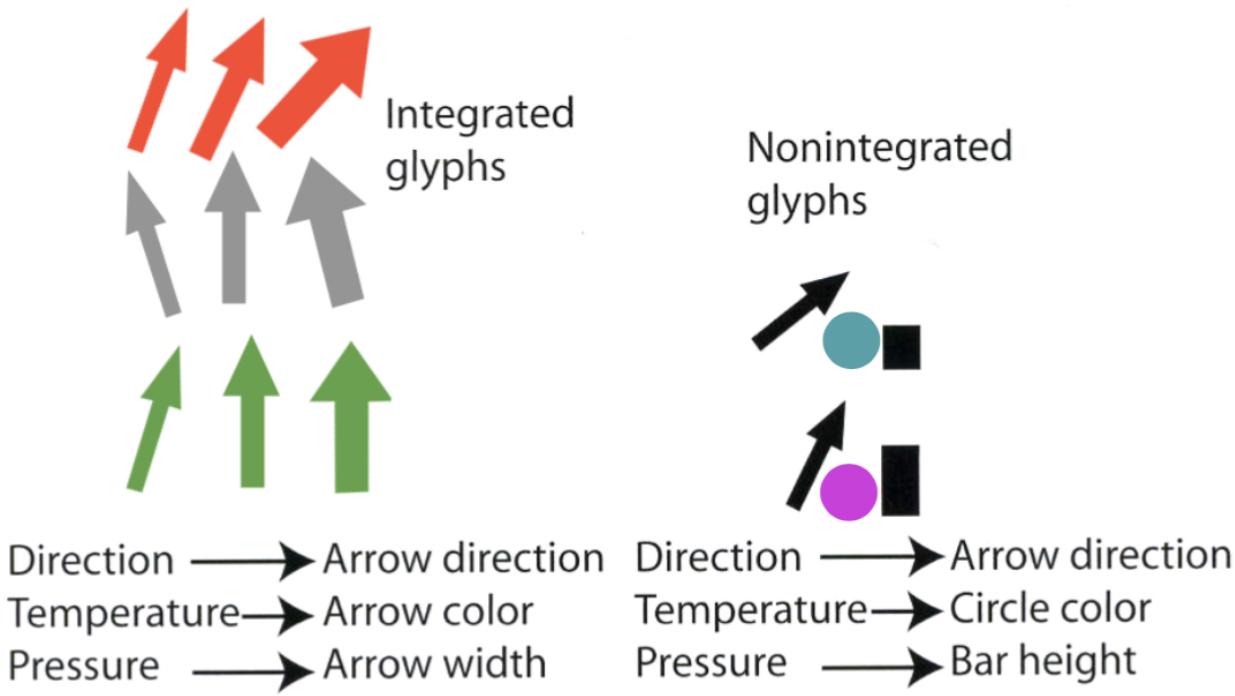
Shapes too


Orientation
Curved/straight
Shape
Shape
Size
Number
Gray/value
Enclosure
Convexity/concavity
Addition
Juncture
Parallelism

Again, not all are equal...



Multidimensional glyth

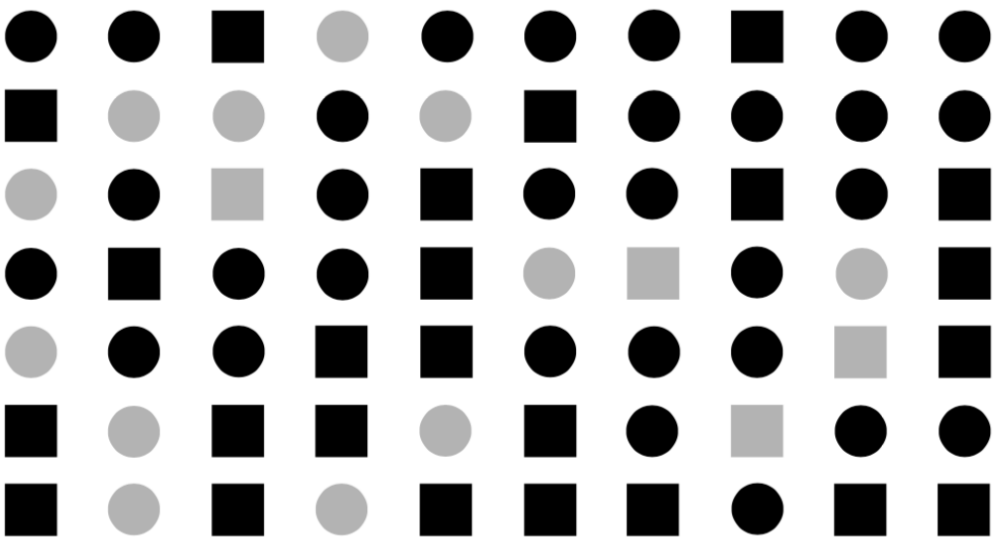


The more variables you encode to dimensions of glyphs, the harder it is to pre-attentively separate the dimensions.

Multidimensional glyth

Caveat: The more variables you encode to dimensions of glyphs, the harder it is to pre-attentively separate the dimensions

How many gray circles?



Color Science

Humans are:

- best at seeing red
- worst at seeing blue

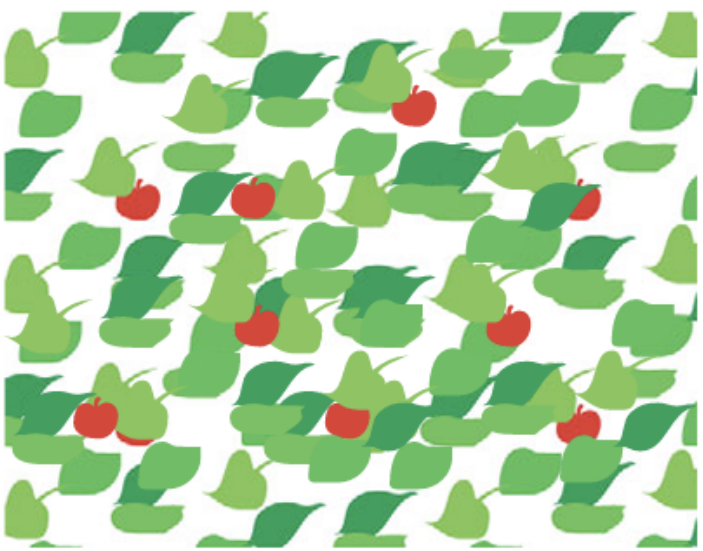
Species vary in color vision ability:

- dogs have only two cones, are red-green colorblind, and see less detail in daylight
- birds have more cones than humans – chickens have 12!

Number of cones = number of primary colors a species perceives.
Mixing the three (human) primaries in different amounts makes any color humans can see.

Color Blindness

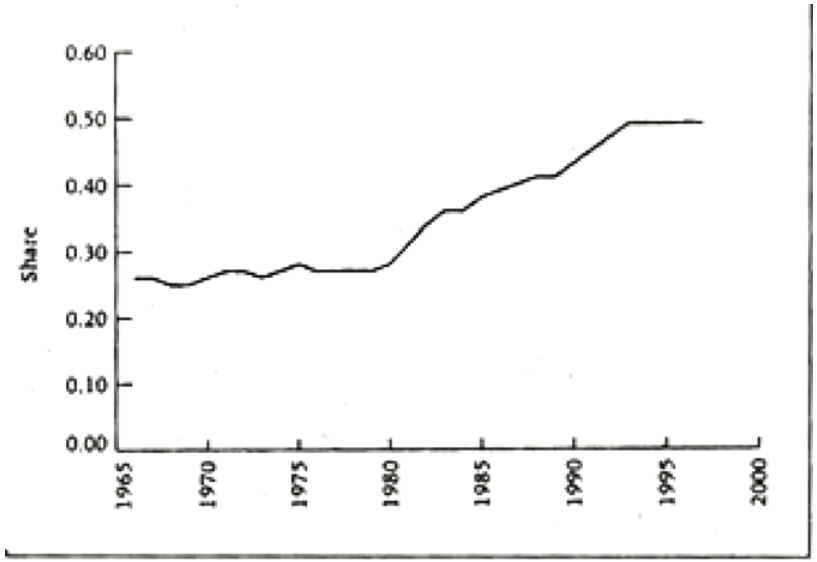
Color Blindness: the most common deficiencies are explained by lack of either the long-wavelength-sensitive cones (protanopia) or the medium-wavelength-sensitive cones (deuteranopia). Both protanopia and deuteranopia result in an inability to distinguish red and green, meaning that the cherries below are difficult for people with these deficiencies to see.



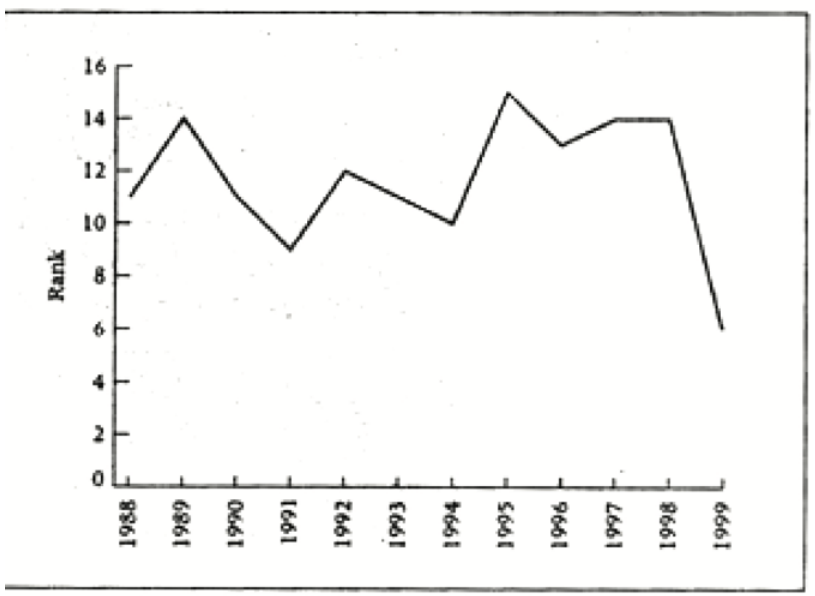
What do we learn from the image?

- Visualization facilitates hypothesis formation, inviting further inquiries into building a theory

(Colin Ware 2012, Ch. 1)



BY THE NUMBERS: OVER 35 YEARS, CORNELL'S TUITION HAS TAKEN AN INCREASINGLY LARGER SHARE OF ITS MEDIAN STUDENT FAMILY INCOME.



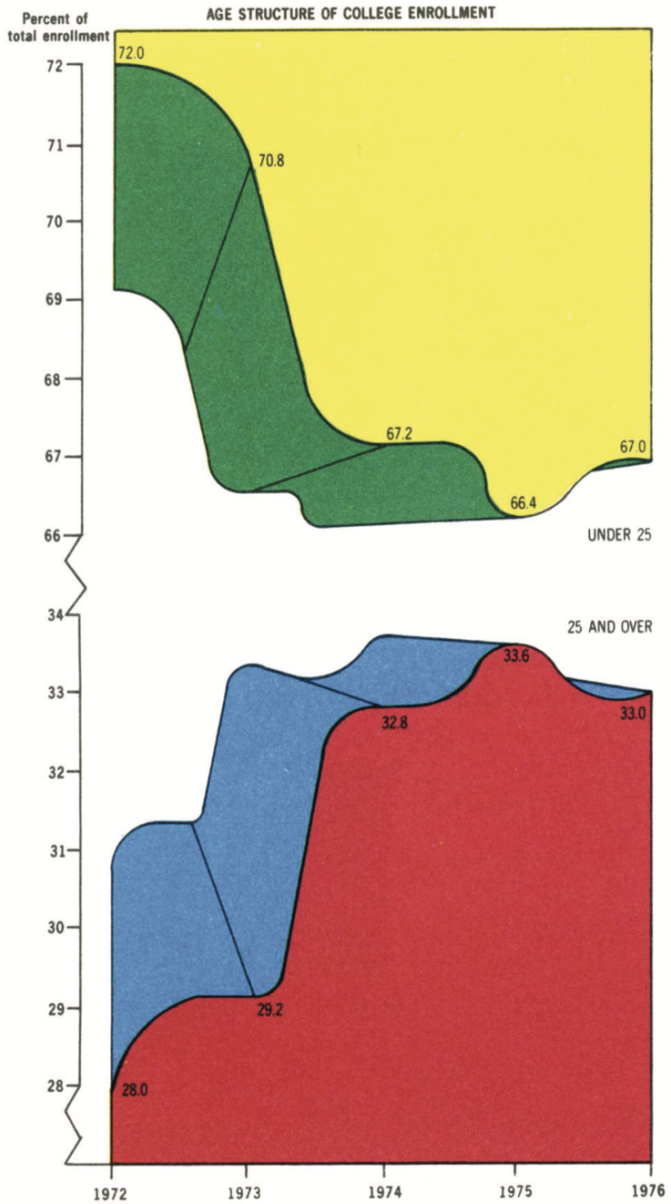
PECKING ORDER: OVER 12 YEARS, CORNELL'S RANKING IN *US NEWS & WORLD REPORT* HAS RISEN AND FALLEN ERRATICALLY.

Messages:

- Gradual rise?

- Abrupt Drop of Ranking?

Source: Chris Adolph, also Johnson, R.R. and Kuby, P.J., 2011. *Elementary statistics*. Cengage Learning.



Message:

- Age structure of college enrollment
- How much data are presented in multiple colors?

Source: Edward R. Tufte. 2001. The Visual Display of Quantitative Information. Graphics Press. 2nd ed.