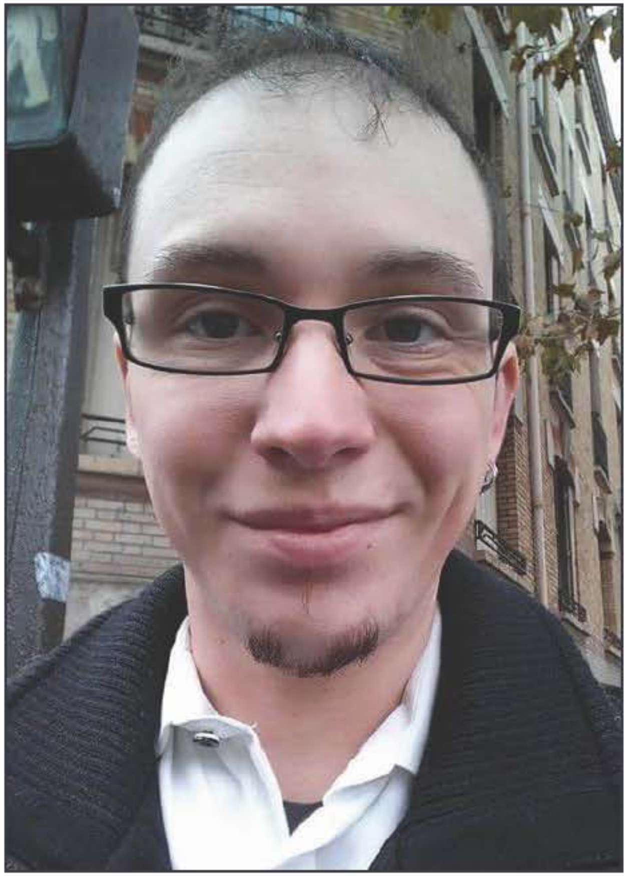
MassMutual DSDP 2017:

INTRODUCTION TO DATA VISUALIZATION

June 8, 2017

R. Jordan Crouser & Amelia McNamara
Statistical & Data Sciences
Smith College

People



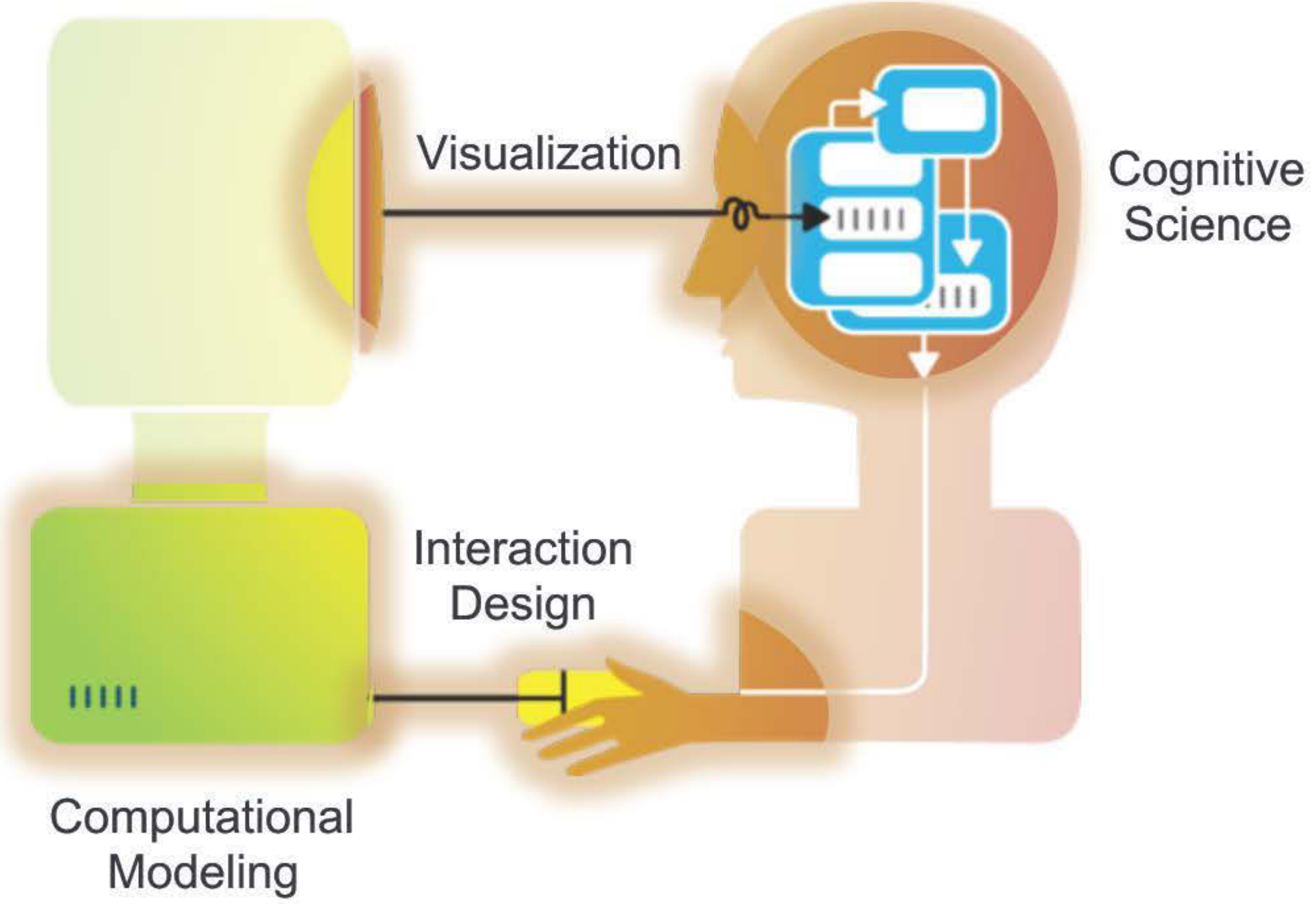
Jordan

(computer scientist)

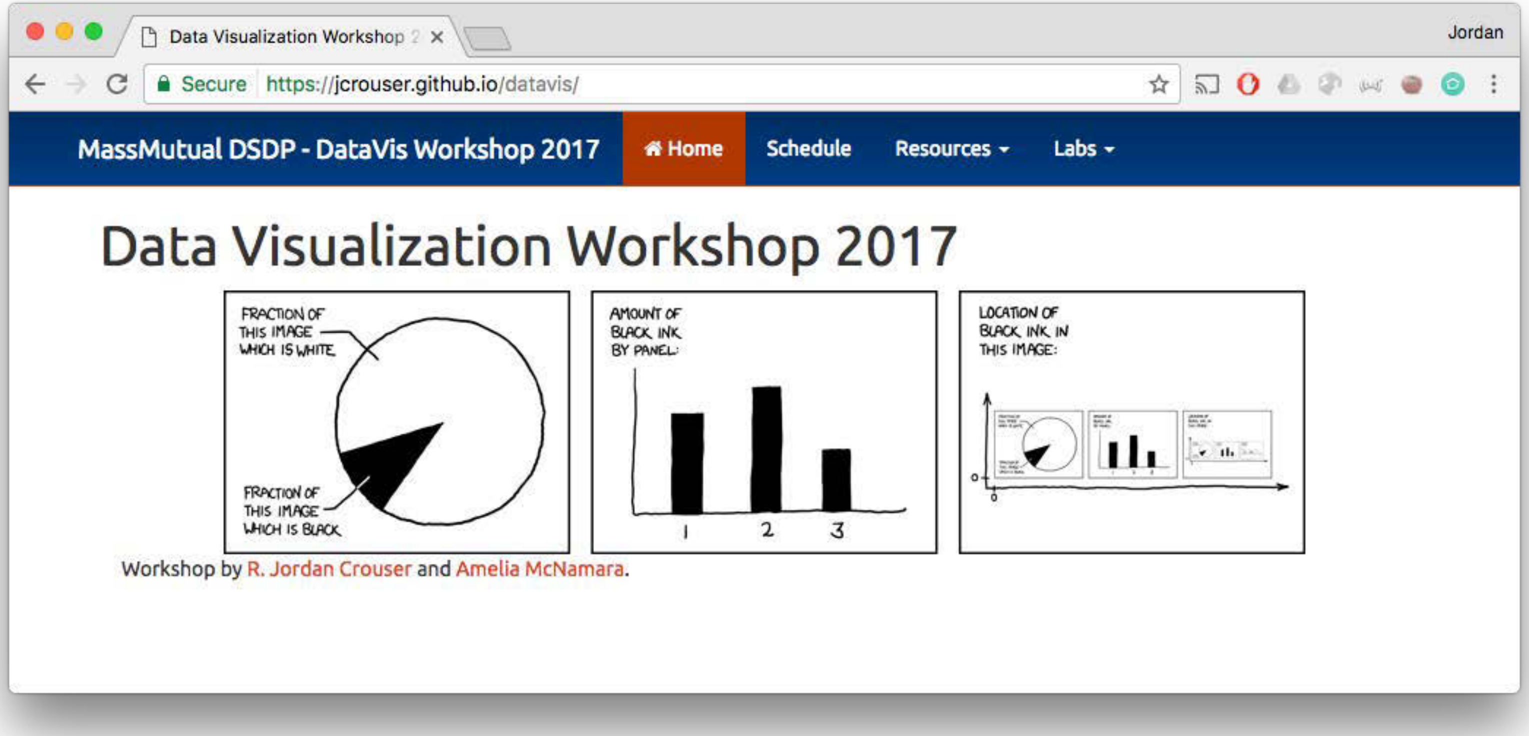
Amelia

(statistician)

Our research (broadly)



Housekeeping



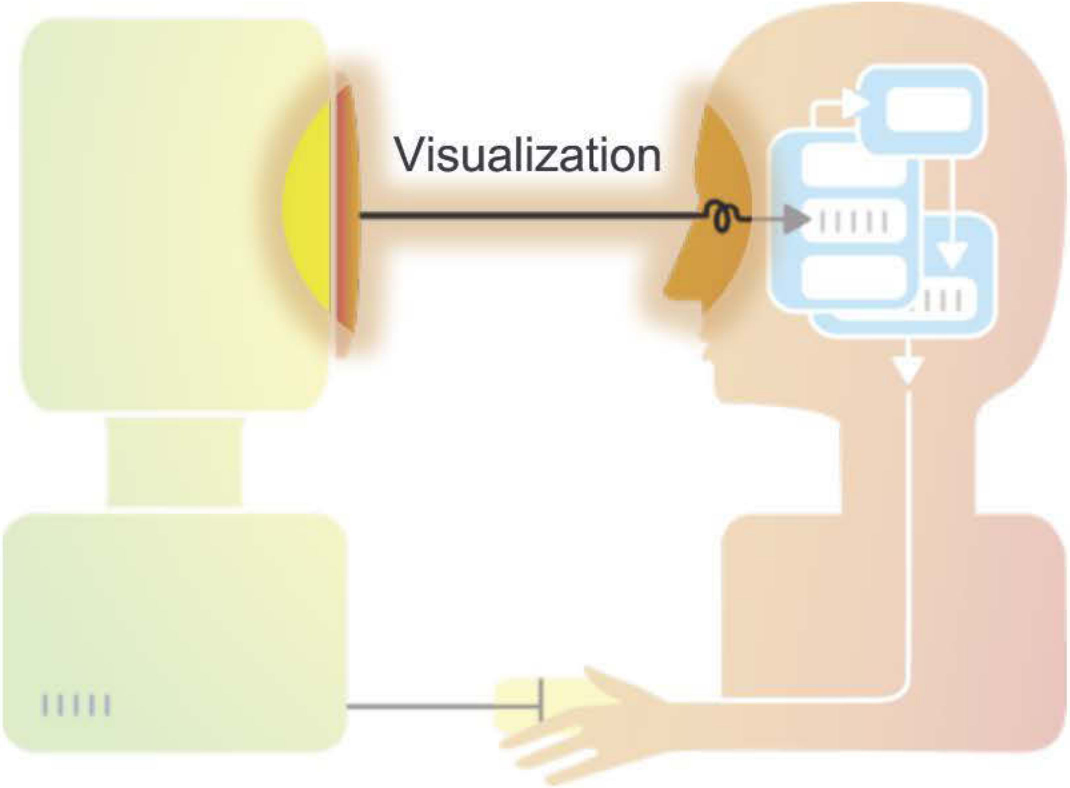
Data Visualization Workshop 2
Jordan
Secure
https://jcrouser.github.io/datavis/
MassMutual DSDP - DataVis Workshop 2017
Home
Schedule
Resources
Labs
FRACTION OF THIS IMAGE WHICH IS WHITE
FRACTION OF THIS IMAGE WHICH IS BLACK
AMOUNT OF BLACK INK BY PANEL:
1
2
3
LOCATION OF BLACK INK IN THIS IMAGE:
0
0

Data Visualization Workshop 2017

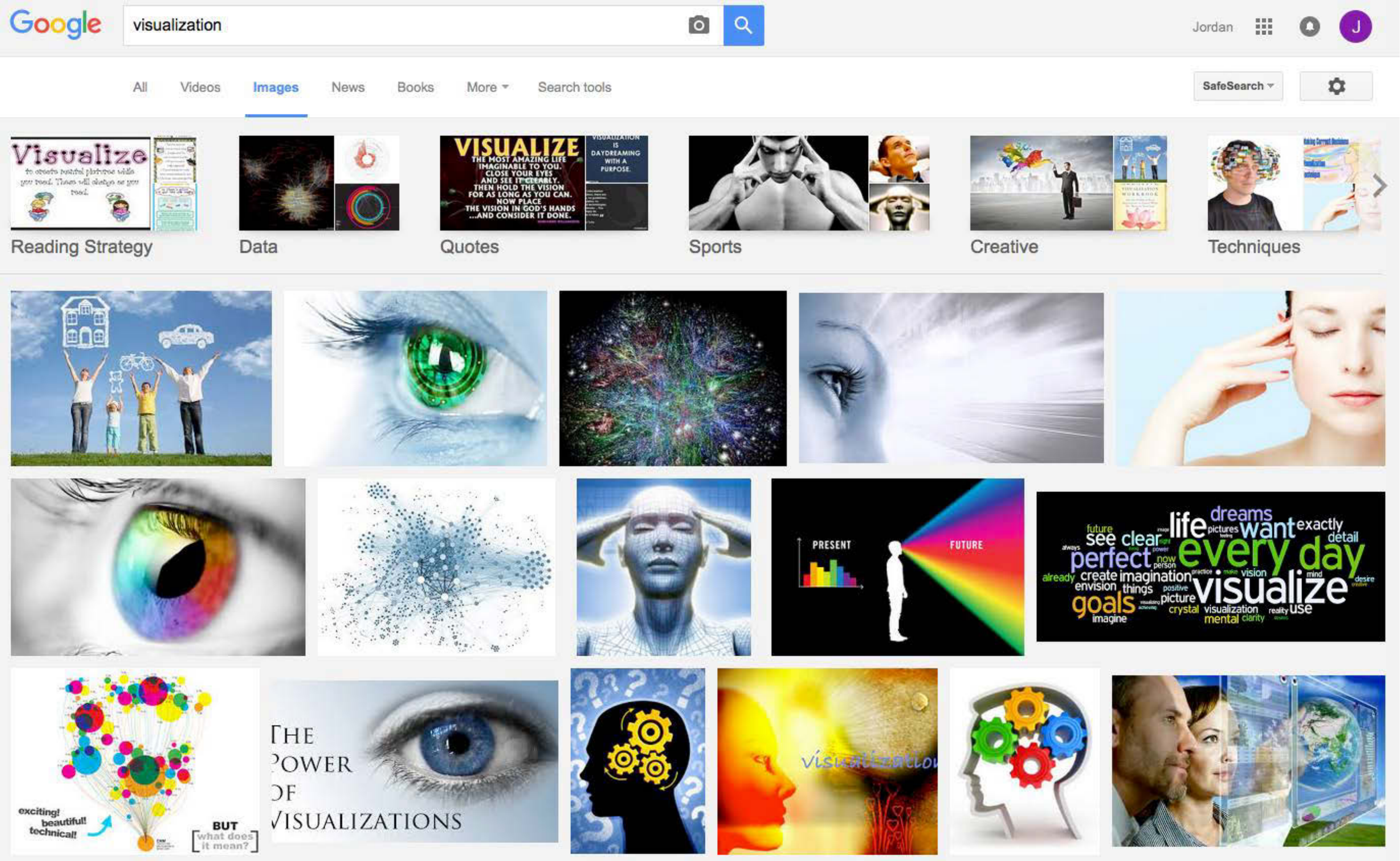
Workshop by R. Jordan Crouser and Amelia McNamara.

jcrouser.github.io/datavis

About this course



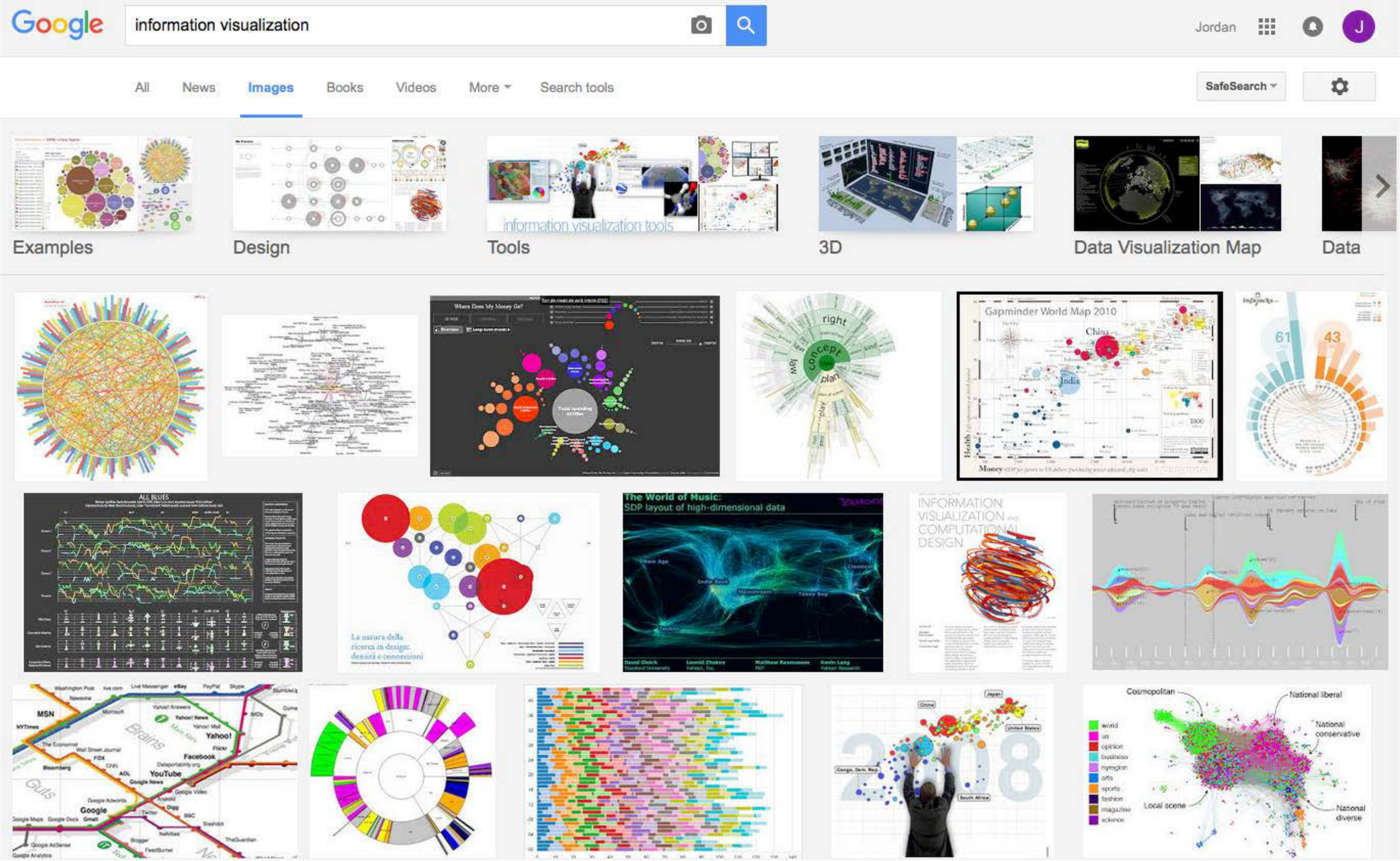
What is visualization?



visualization

Google
Jordan
J
All
Videos
Images
News
Books
More
Search tools
SafeSearch
Reading Strategy
Data
Quotes
Sports
Creative
Techniques
Visualize
VISUALIZE
THE MOST AMAZING LIFE IMAGINABLE TO YOU. CLOSE YOUR EYES AND SEE IT CLEARLY. THEN HOLD THE VISION FOR AS LONG AS YOU CAN. NOW PLACE THE VISION IN GOD'S HANDS ...AND CONSIDER IT DONE.
VISUALIZATION IS DAYDREAMING WITH A PURPOSE.
PRESENT
FUTURE
dreams life want exactly detail future see clear perfect every day already create imagination vision envision things picture visualize goals crystal visualization mental clarity reality use imagine
exciting! beautiful! technical!
BUT what does it mean?
THE POWER OF VISUALIZATIONS
visualization

What is visualization?

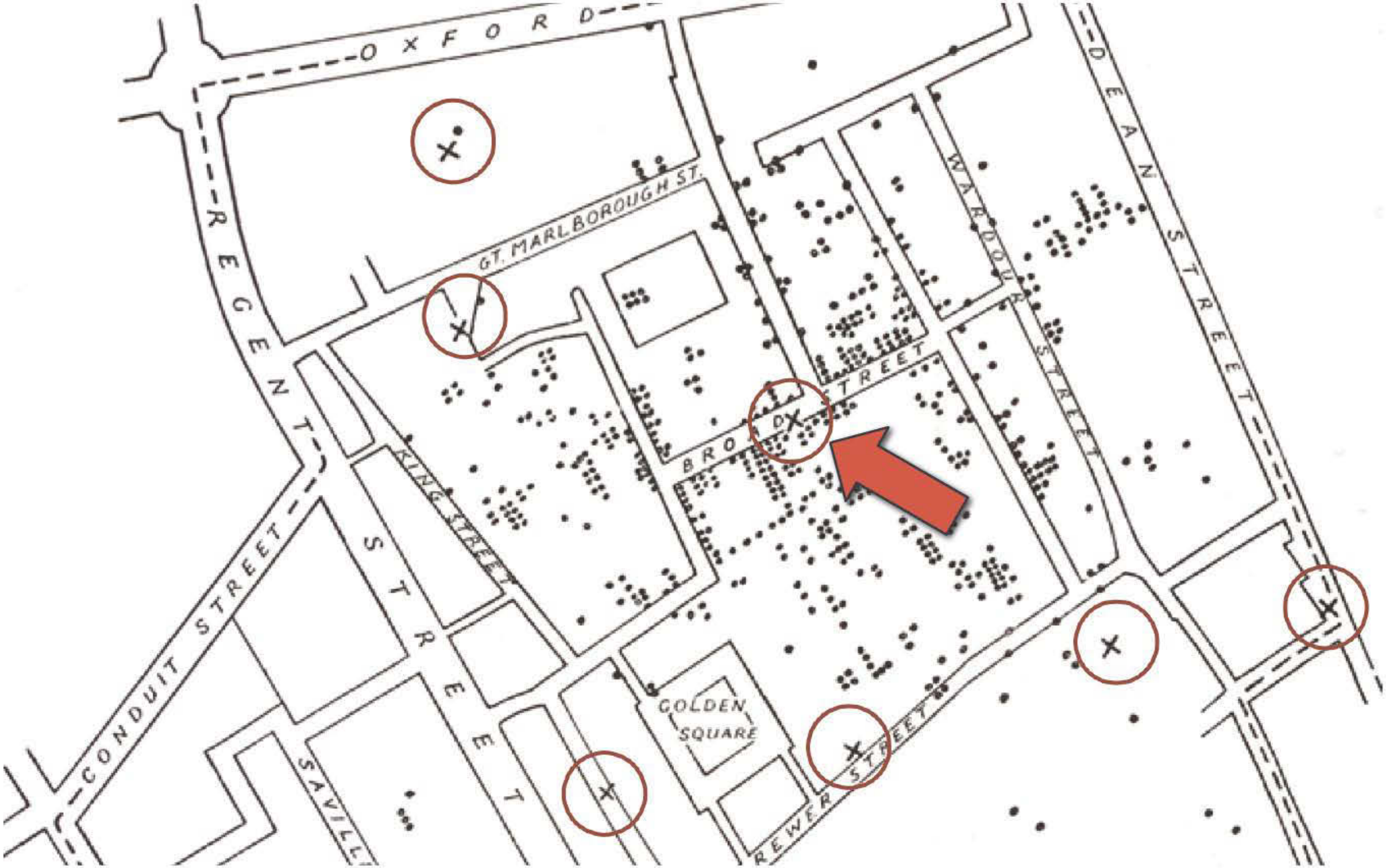


information visualization

Perhaps a more helpful question:

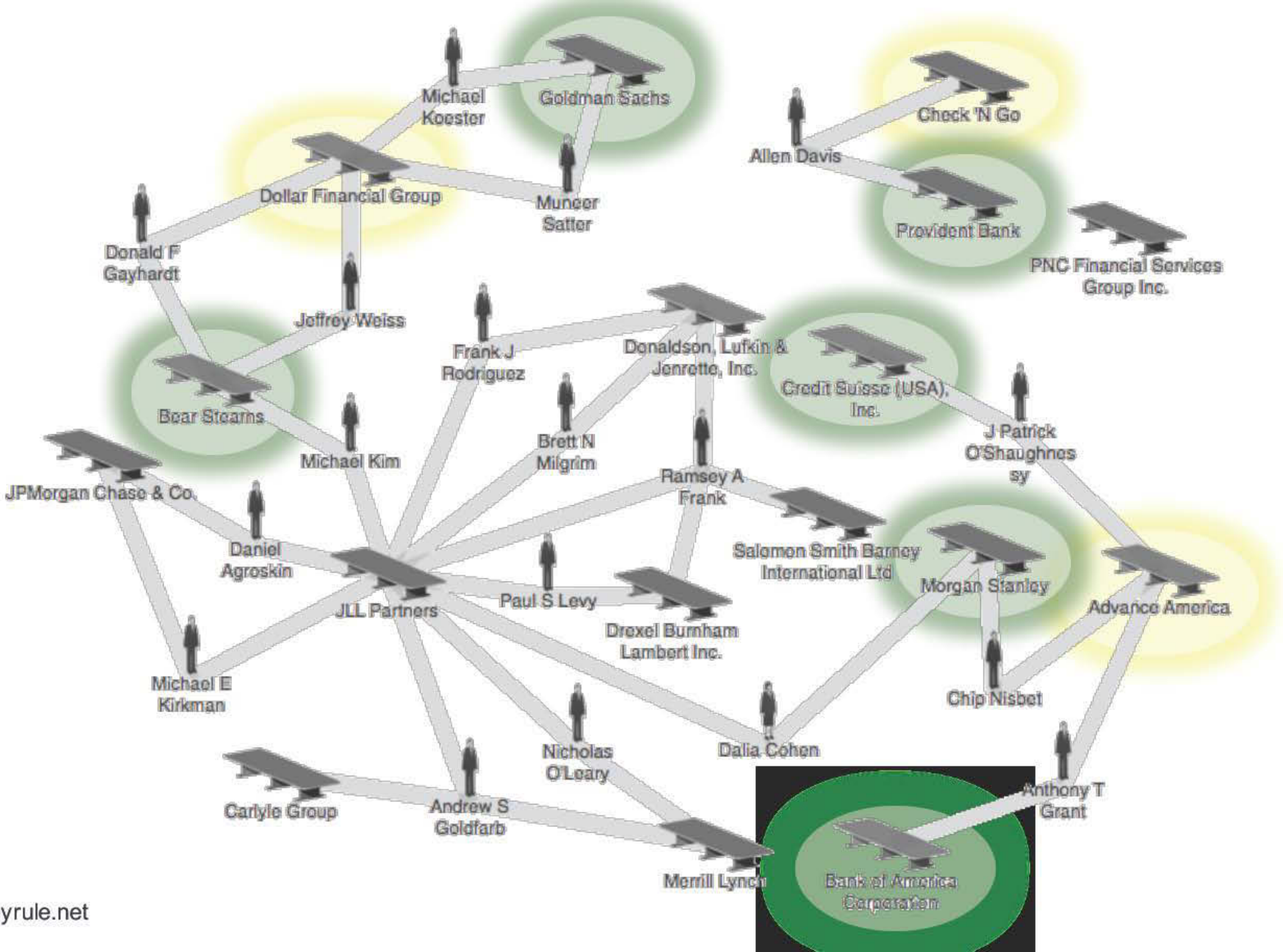
What are some ways
a “visualization” can be **useful**?

Does it help you spot trends?



More info here: http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

Does it help you explore?



Credit: theyrule.net

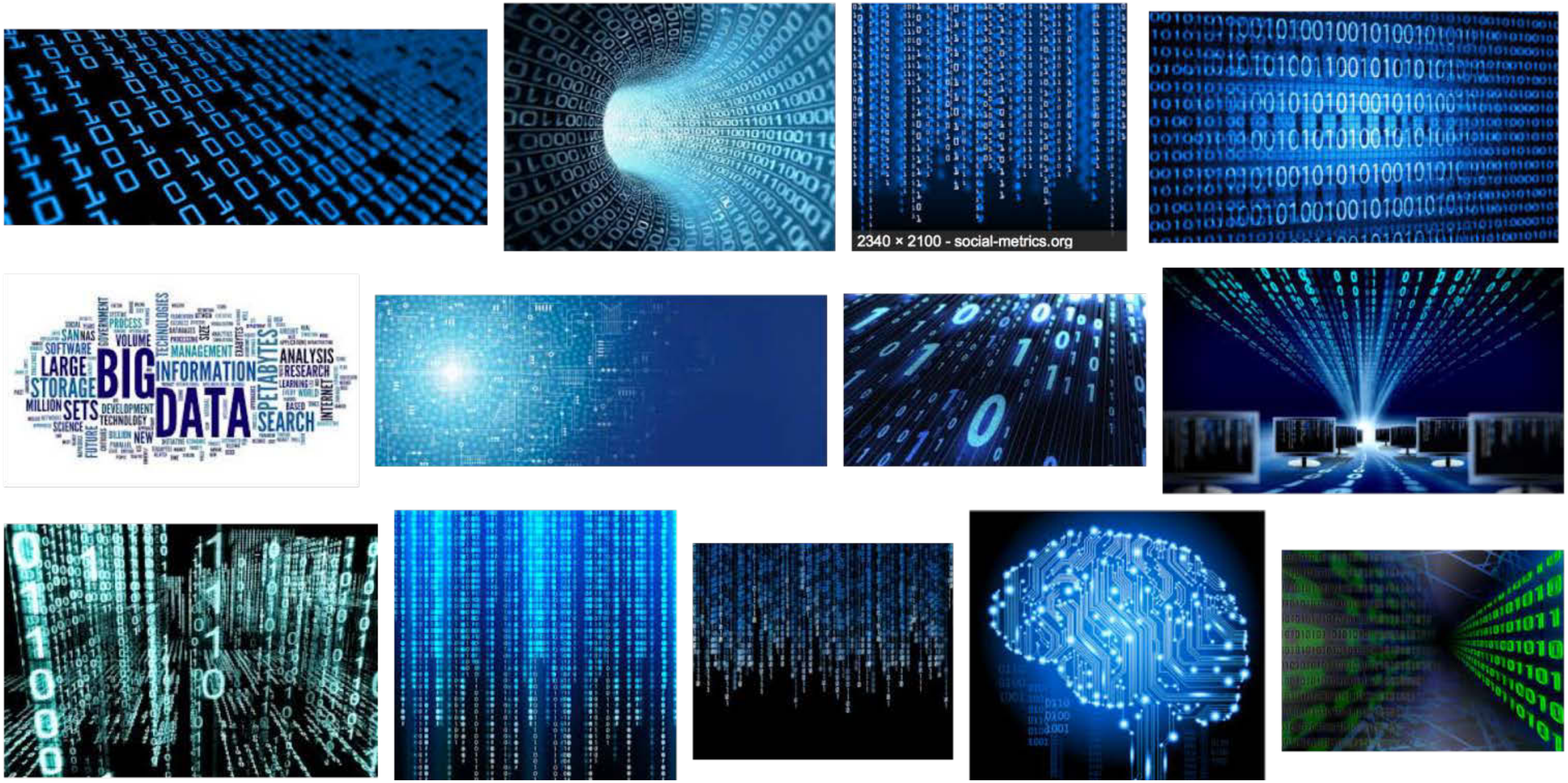
Does it tell a story?

Hans Rosling’s 200 Countries, 200 Years, 4 Minutes – The Joy of Stats – BBC Four
https://www.youtube.com/watch?v=jbkSRLYSojo

Visualization (def.)

Visual **representations** of data that reinforce human **cognition**

Wait… what is “data”?



Data: a definition

Data is a set of *variables* that capture various aspects of the world:

Tuition rates, enrollment numbers, public vs. private, etc.

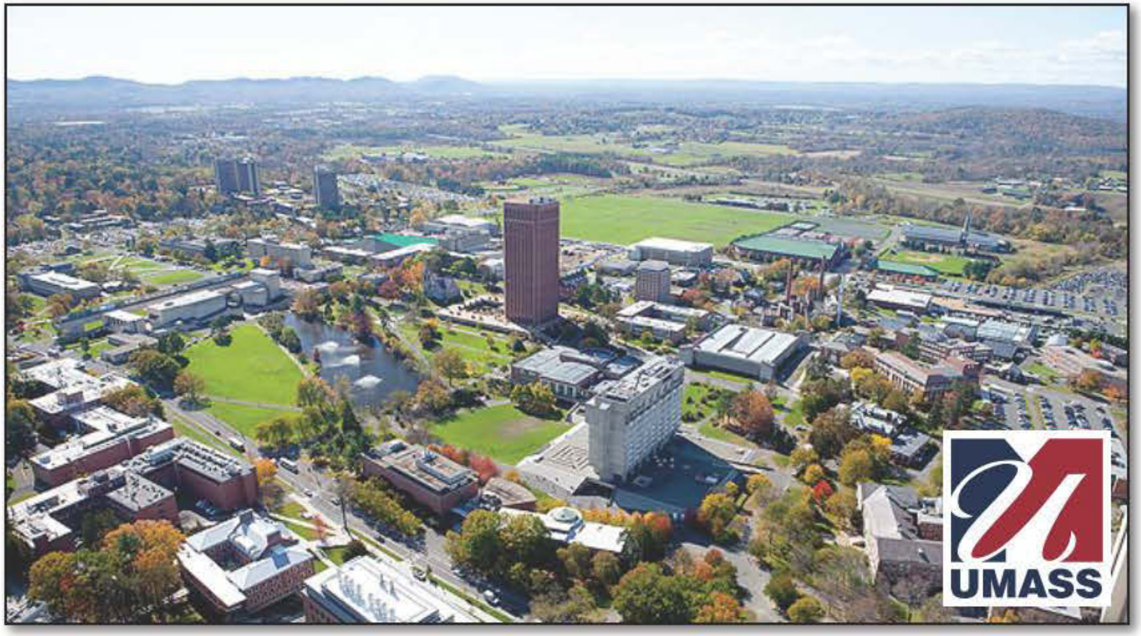
Data: a definition

A dataset also contains a set of *observations* (also called *records)* over these variables. For example:

tuition = $46,288, *enrollment* = 2,563, *private, etc.*

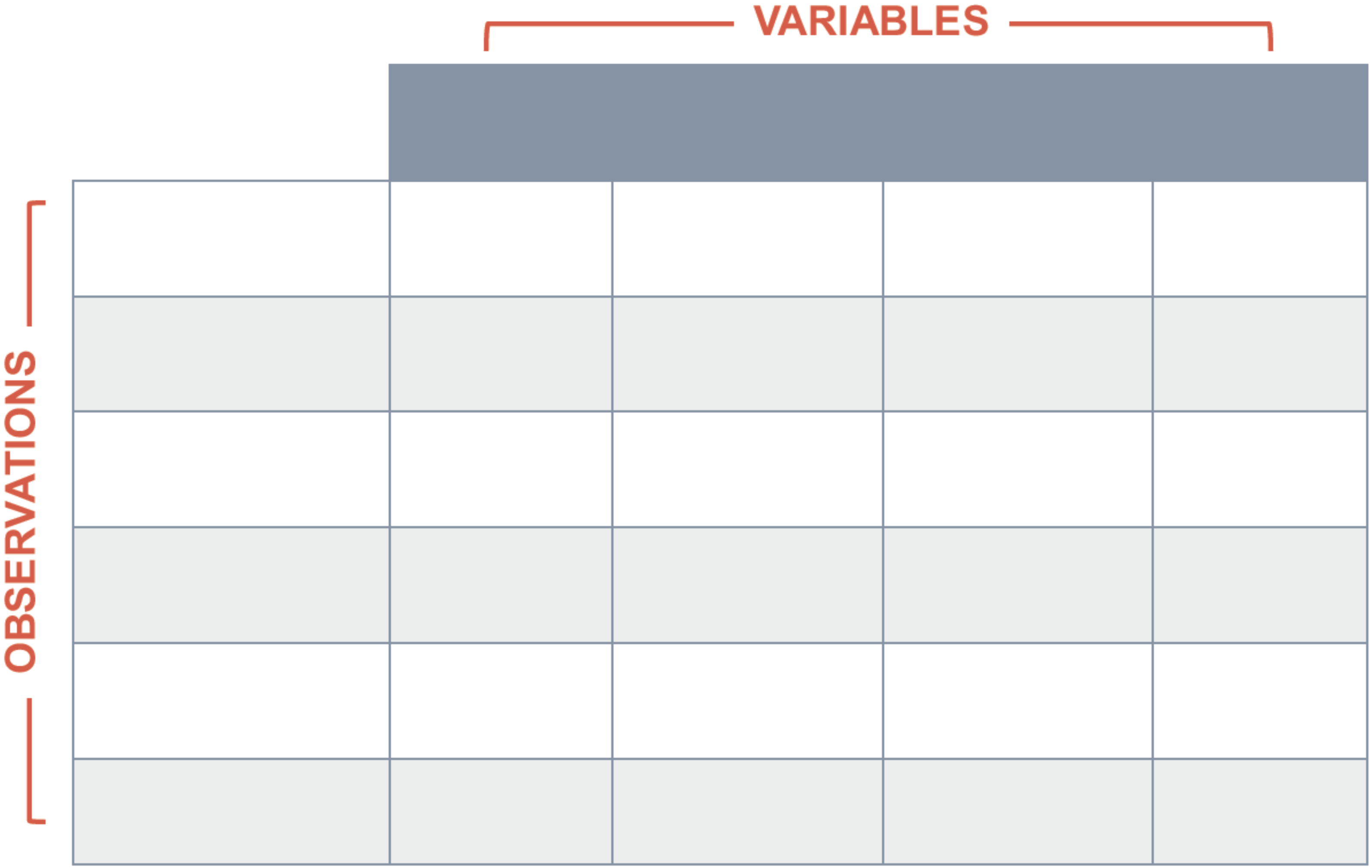
Data: a definition

A dataset also contains a set of *observations* (also called *records)* over these variables. For example:



tuition = $16,115, *enrollment* = 28,635, *public, etc.*

One way to think about this:



VARIABLES

OBSERVATIONS

Another way to think about this

```
class school_obs:
     def __init__(tuition, enrollment,
                  pub_or_priv):
          self.tuition = tuition
          self.enrollment = enrollment
          self.pub_or_priv = pub_or_priv
```

VARIABLES

OBSERVATIONS

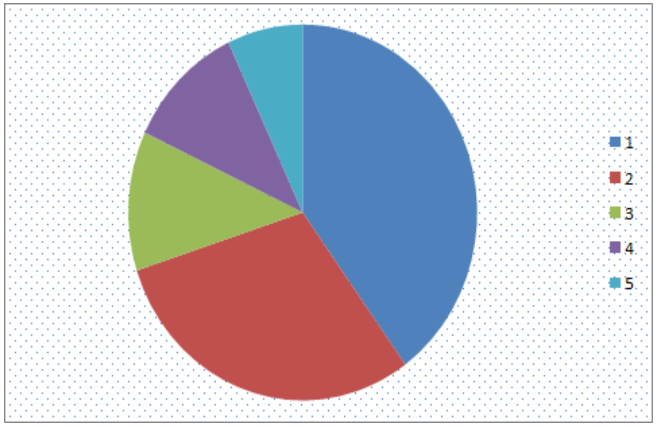
Why is this important?

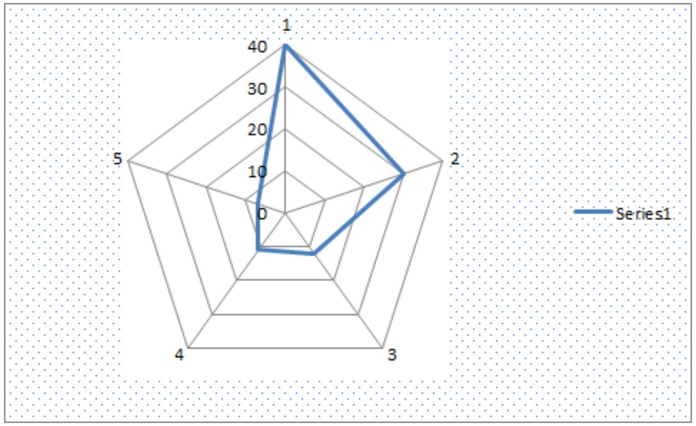
- Data have dimensions
- Visualizations have dimensions, too
- To build visualizations, we need to **map** data dimensions to visual dimensions

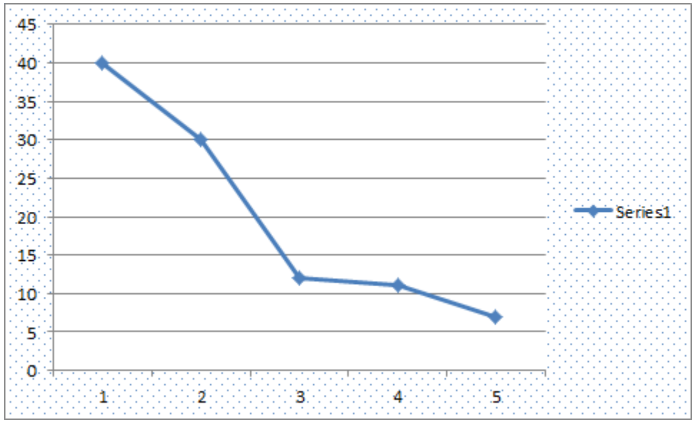
Key question for this course

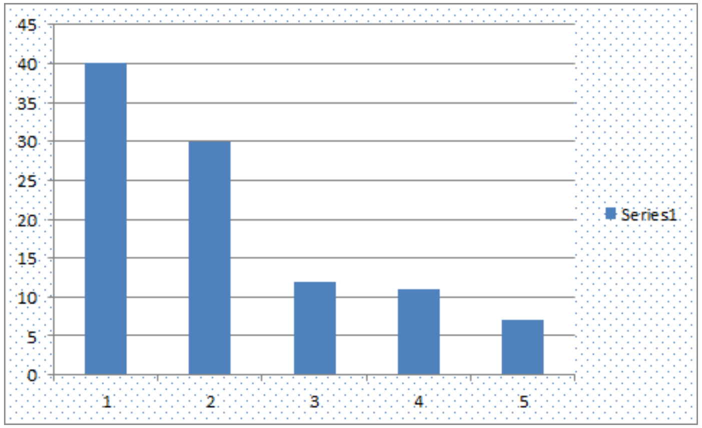
Which **data dimension** should be mapped

to which **visual dimension?**



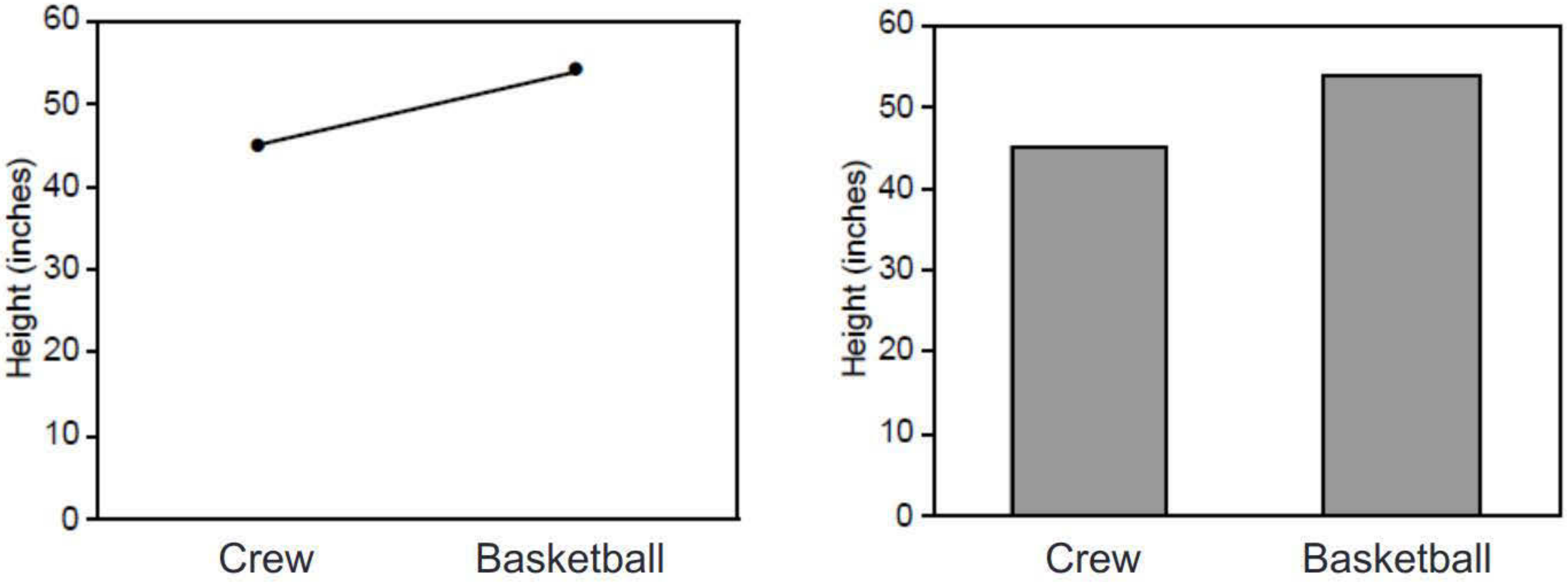




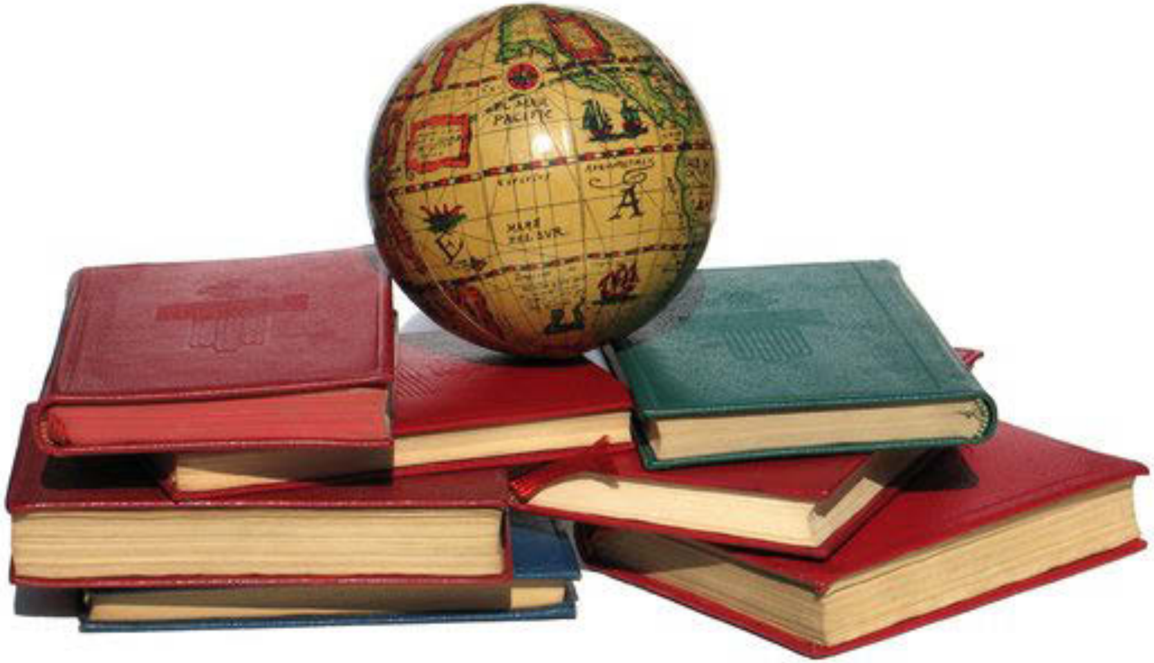


Answer: it depends

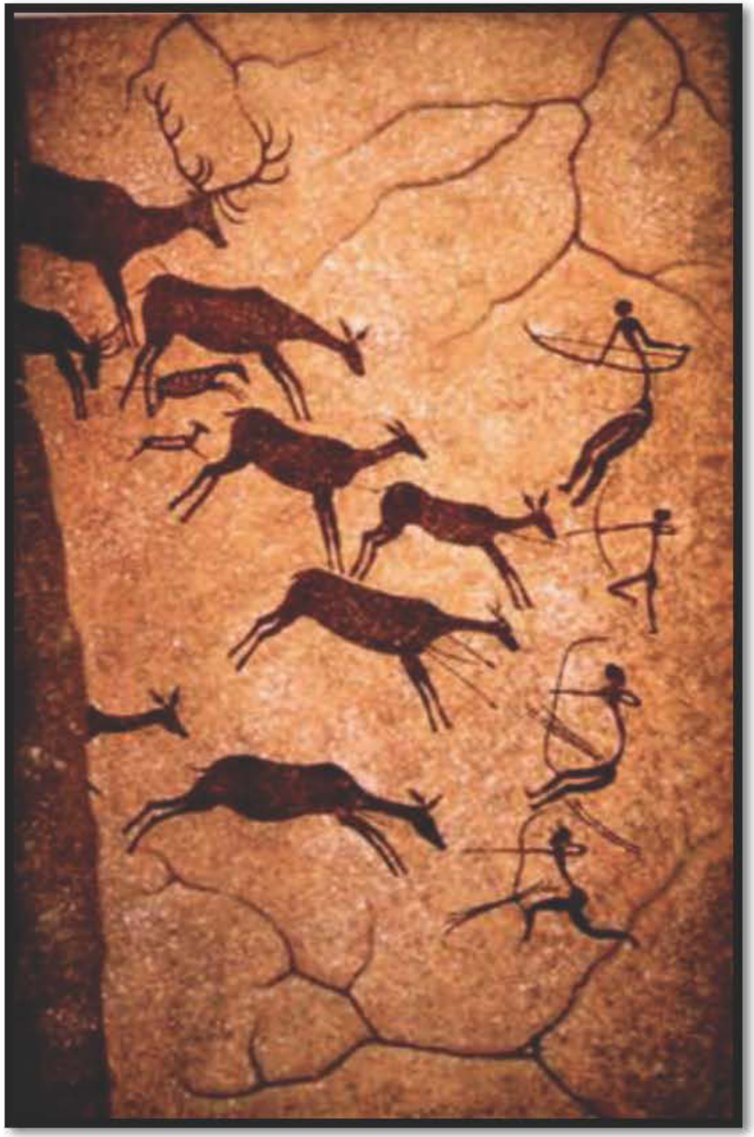
Average Height for Youth Sports Participants



A quick history lesson…

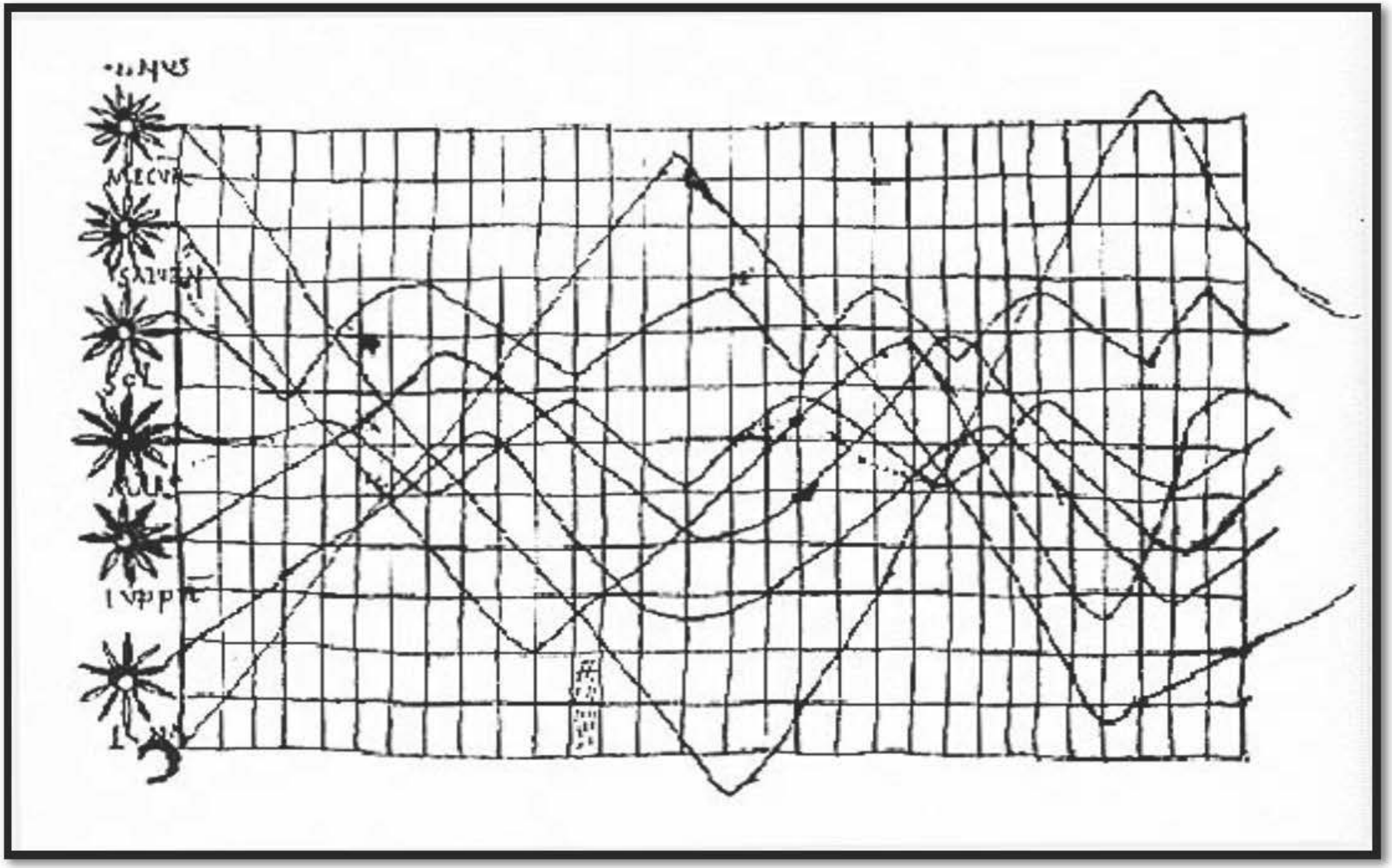


(Incomplete) History of Visualization: 15,000BC



15,000 BC. Laxcaux, France

(Incomplete) History of Visualization: 900s

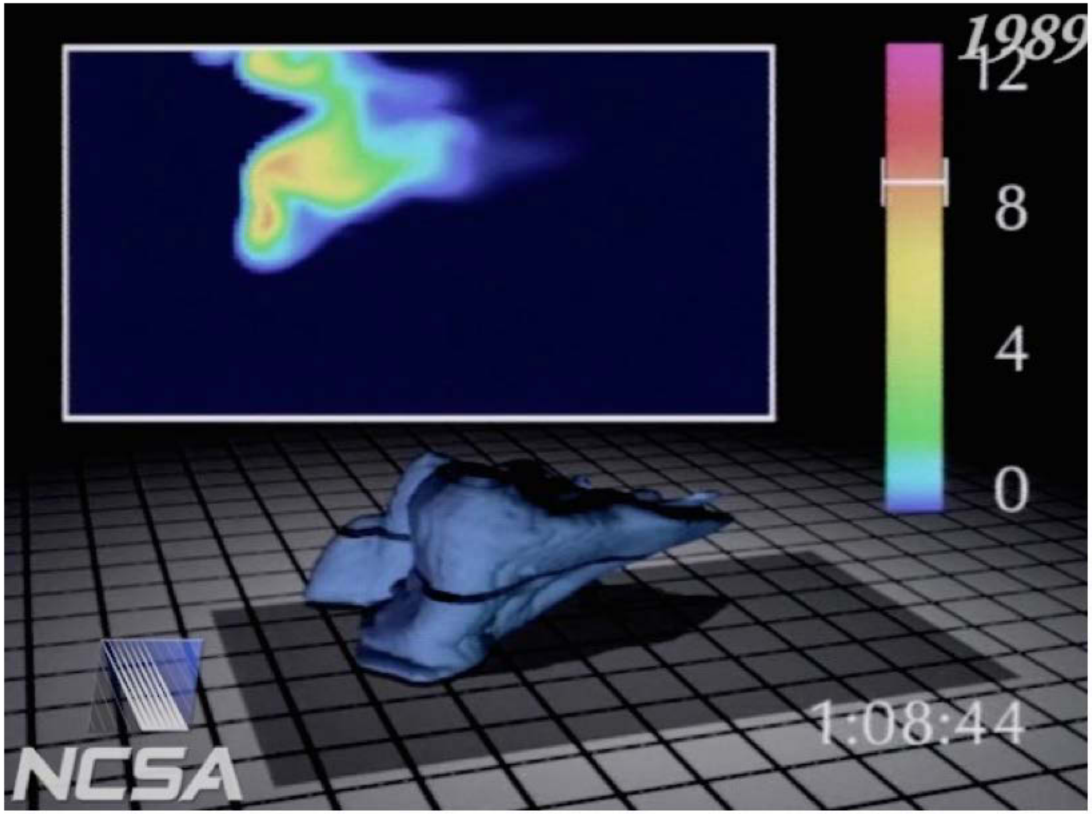


"De cursu per zodiacum", illustrator unknown

(Incomplete) History of Visualization: 1970s

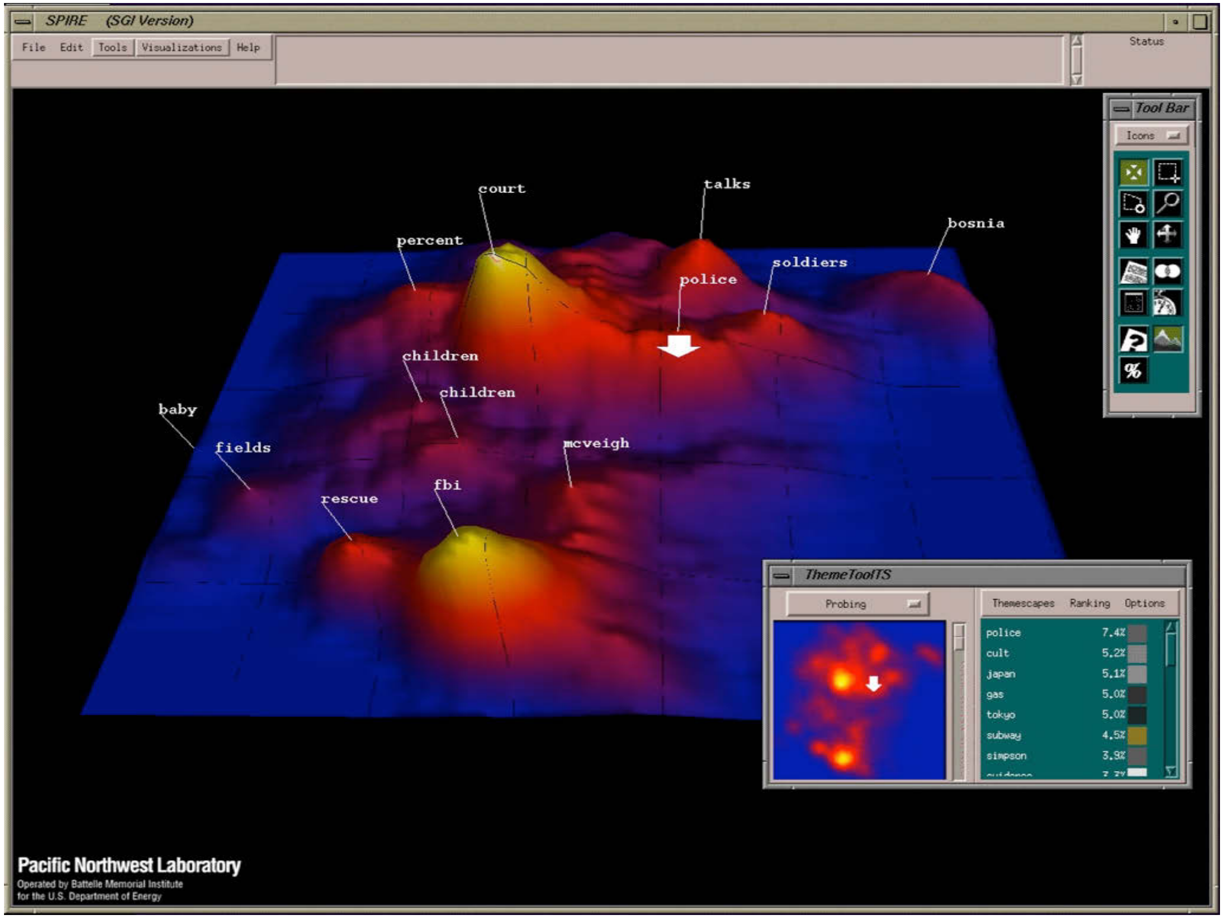
- CAD/CAM, building cars, planes, chips
- Starting to think about: 3D, animation, edu, medicine

(Incomplete) History of Visualization: 1980s



- Scientific visualization, physical phenomena
- Starting to think about: photorealism, entertainment

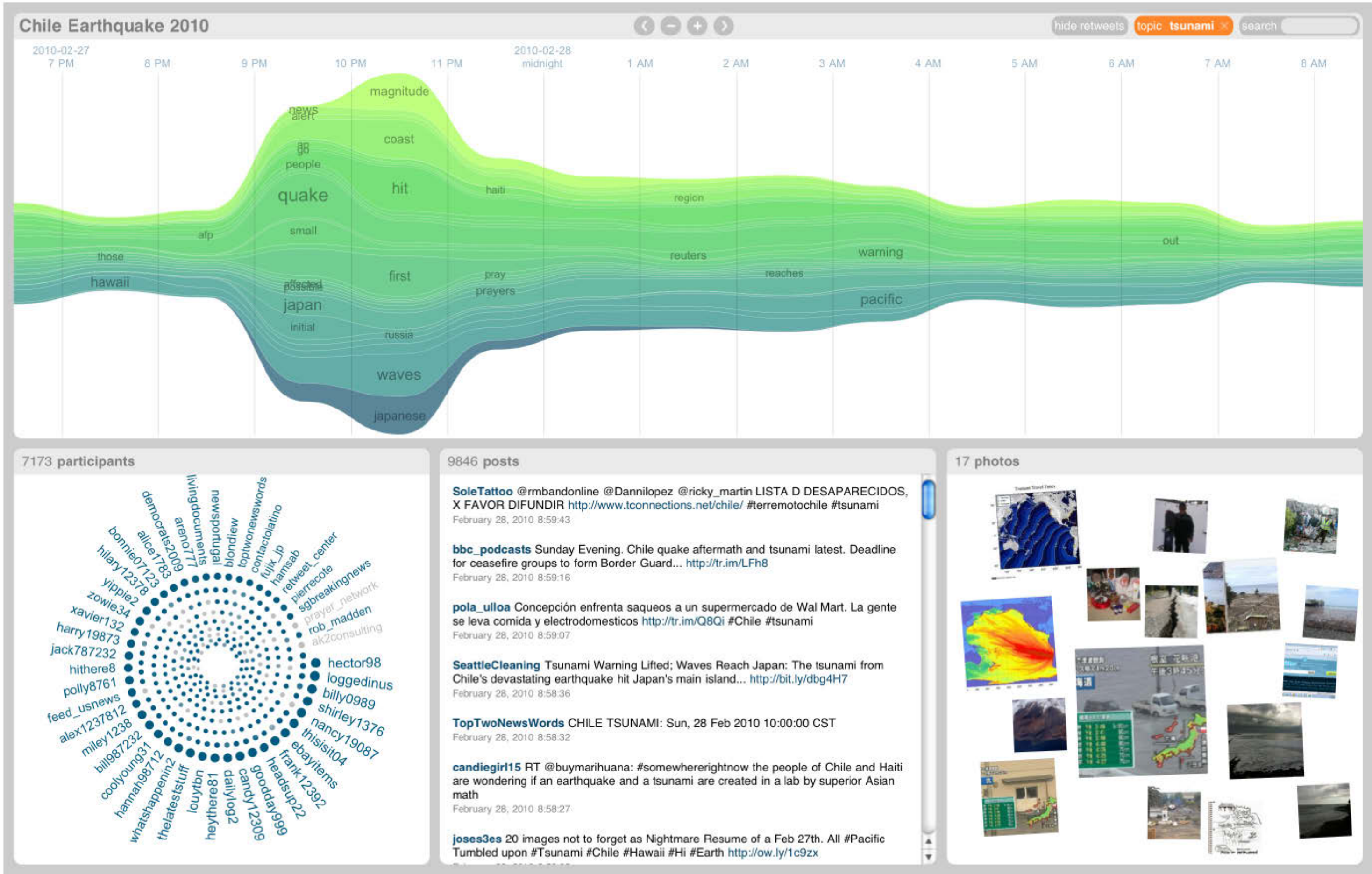
(Incomplete) History of Visualization: 1990s



police	7.4%
cult	5.2%
japan	5.1%
gas	5.0%
tokyo	5.0%
subway	4.5%
simpson	3.9%

- Information visualization, storytelling
- Starting to think about: online spaces, interaction

(Incomplete) History of Visualization: 2000s



9846 posts

SoleTattoo @rmbandonline @Dannilopez @ricky_martin LISTA D DESAPARECIDOS, X FAVOR DIFUNDIR http://www.tconnections.net/chile/ #terremotochile #tsunami
February 28, 2010 8:59:43

bbc_podcasts Sunday Evening. Chile quake aftermath and tsunami latest. Deadline for ceasefire groups to form Border Guard... http://tr.im/LFh8
February 28, 2010 8:59:16

pola_ulloa Concepción enfrenta saqueos a un supermercado de Wal Mart. La gente se leva comida y electrodomesticos http://tr.im/Q8Qi #Chile #tsunami
February 28, 2010 8:59:07

SeattleCleaning Tsunami Warning Lifted; Waves Reach Japan: The tsunami from Chile's devastating earthquake hit Japan's main island... http://bit.ly/dbg4H7
February 28, 2010 8:58:36

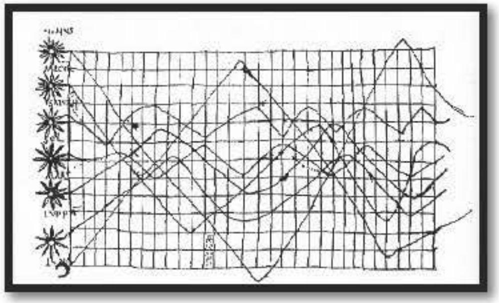
TopTwoNewsWords CHILE TSUNAMI: Sun, 28 Feb 2010 10:00:00 CST
February 28, 2010 8:58:32

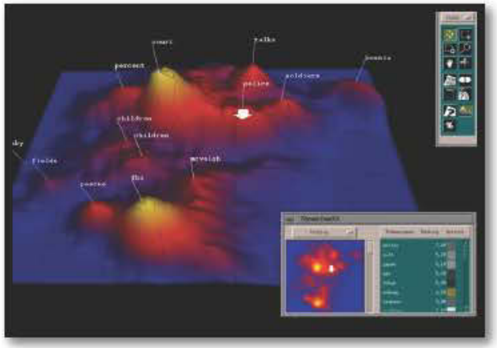
candiegirl15 RT @buymarihuana: #somewhererightnow the people of Chile and Haiti are wondering if an earthquake and a tsunami are created in a lab by superior Asian math
February 28, 2010 8:58:27

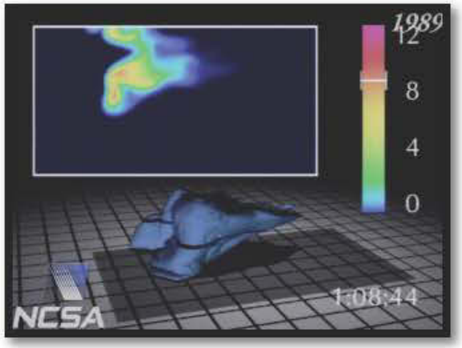
joses3es 20 images not to forget as Nightmare Resume of a Feb 27th. All #Pacific Tumbled upon #Tsunami #Chile #Hawaii #Hi #Earth http://ow.ly/1c9zx

- Coordination across multiple views, interaction
- Starting to think about: sensemaking, provenance

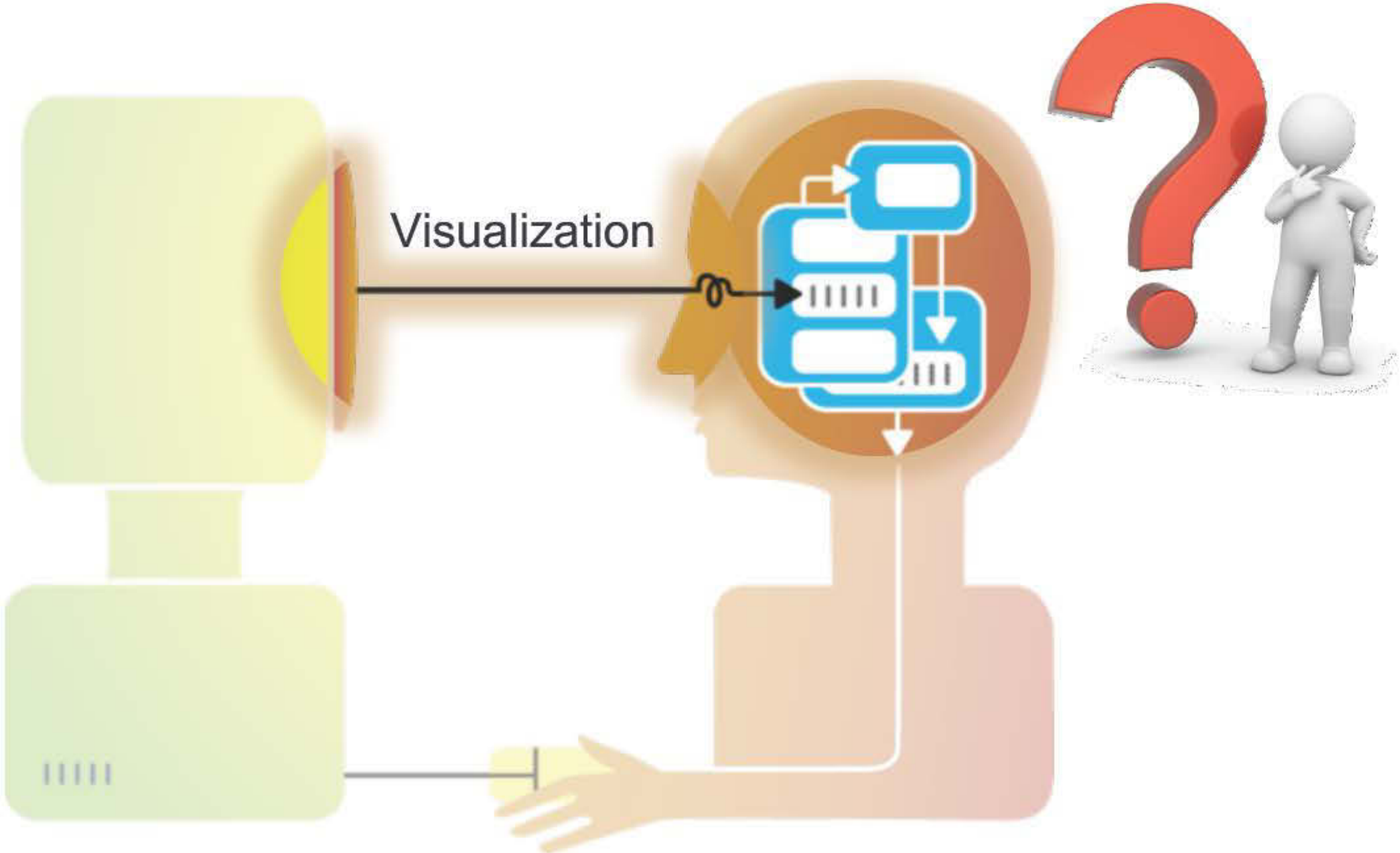
Discussion: what are they all trying to do?



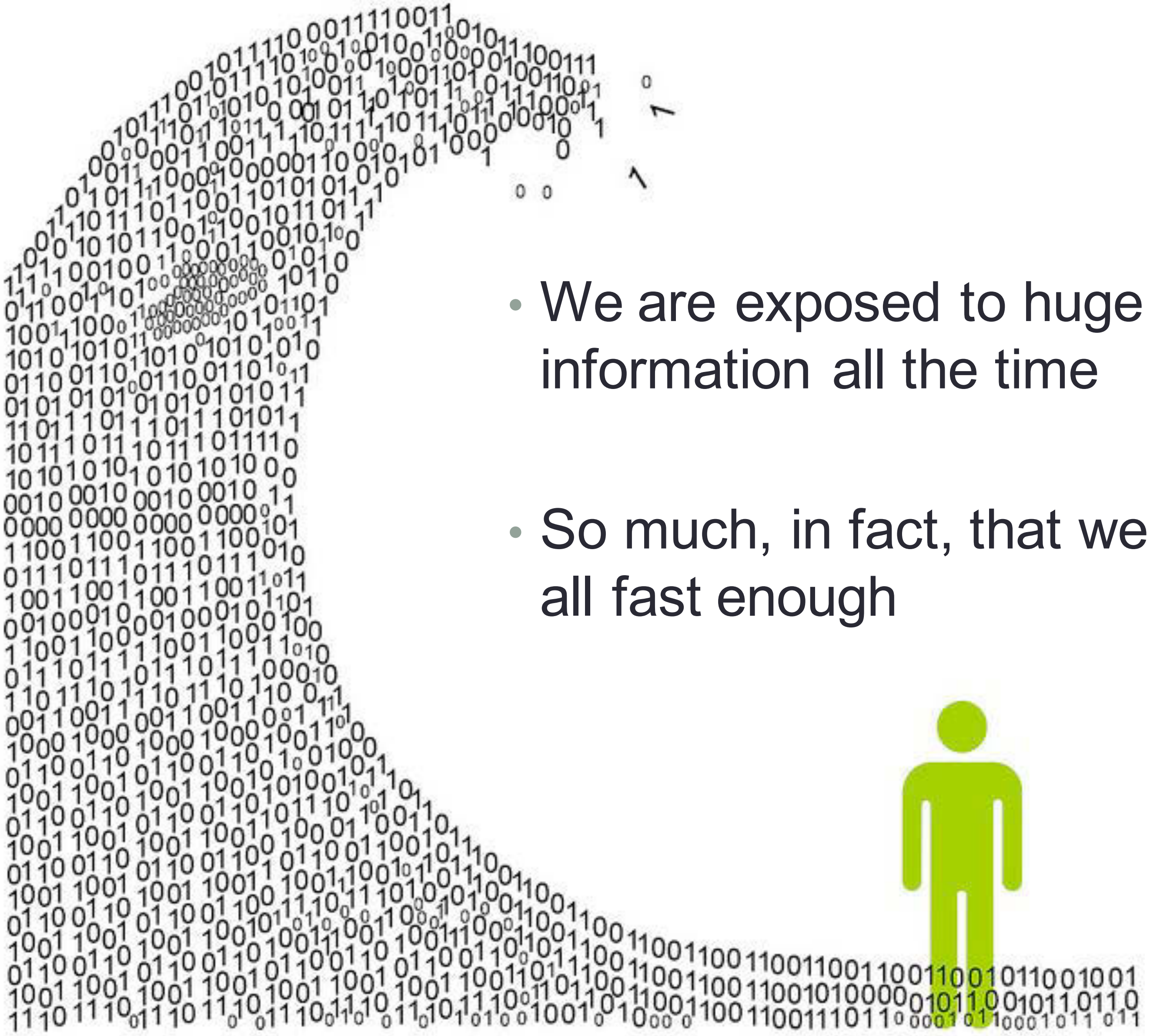




Visualization helps shape *mental models*



Information overload



- We are exposed to huge amounts of information all the time

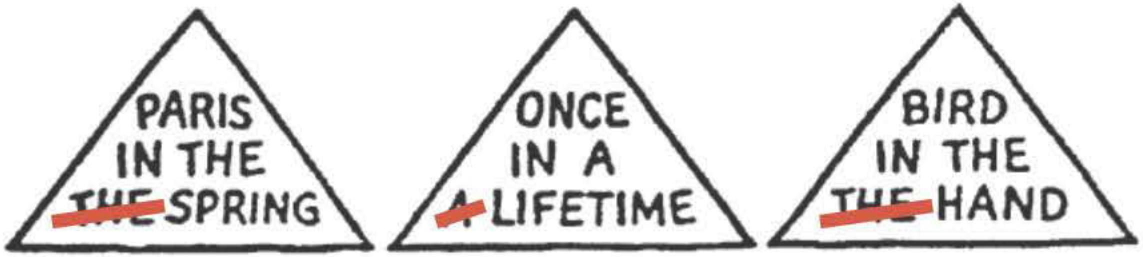
- So much, in fact, that we can’t process it all fast enough

Mental models

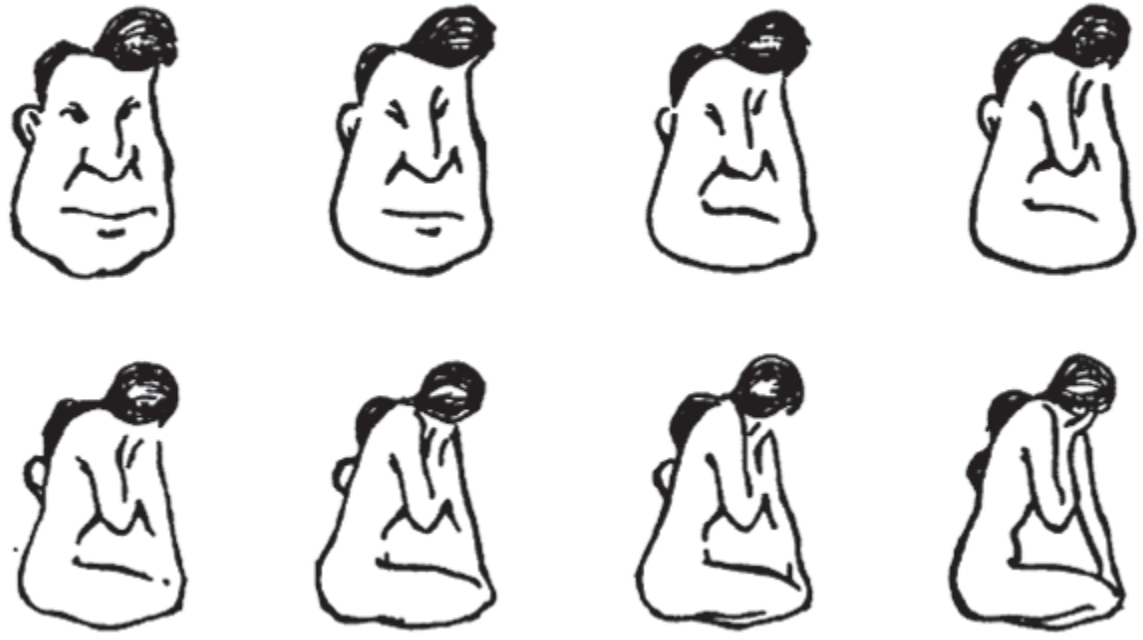
To cope, we construct **mental models:** abstracted, simplified versions of the world that are more manageable

Mental Models: a Sketch

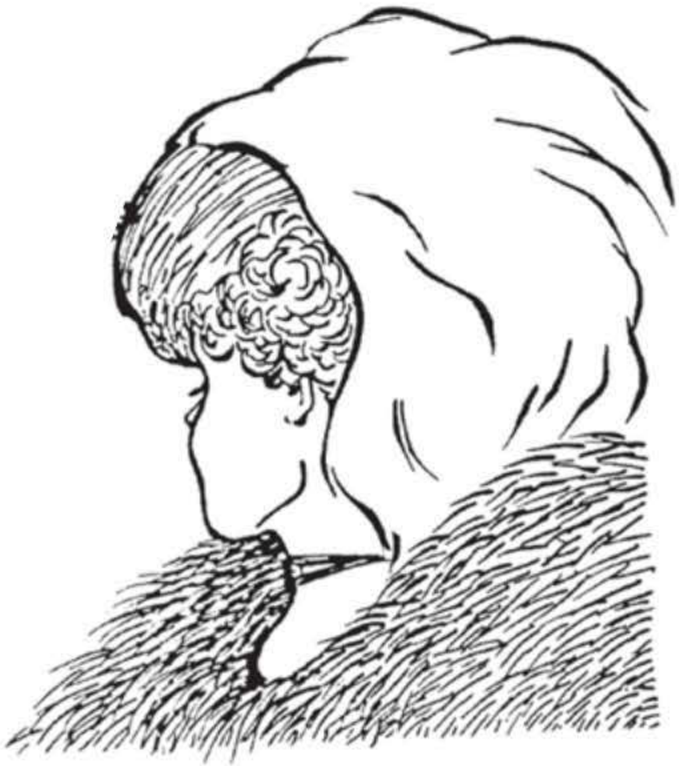
1. We tend to see what we expect to see



2. Mental models form quickly, & update slowly



3. New information gets incorporated into the existing model



4. Initial exposure interferes with accurate perception

Blur size

128px
64px
32px
16px
8px
None

The good, the bad, and the ugly…

The good:

- Well-tuned mental models let us process information quickly
- Frees up more processing power to synthesize information

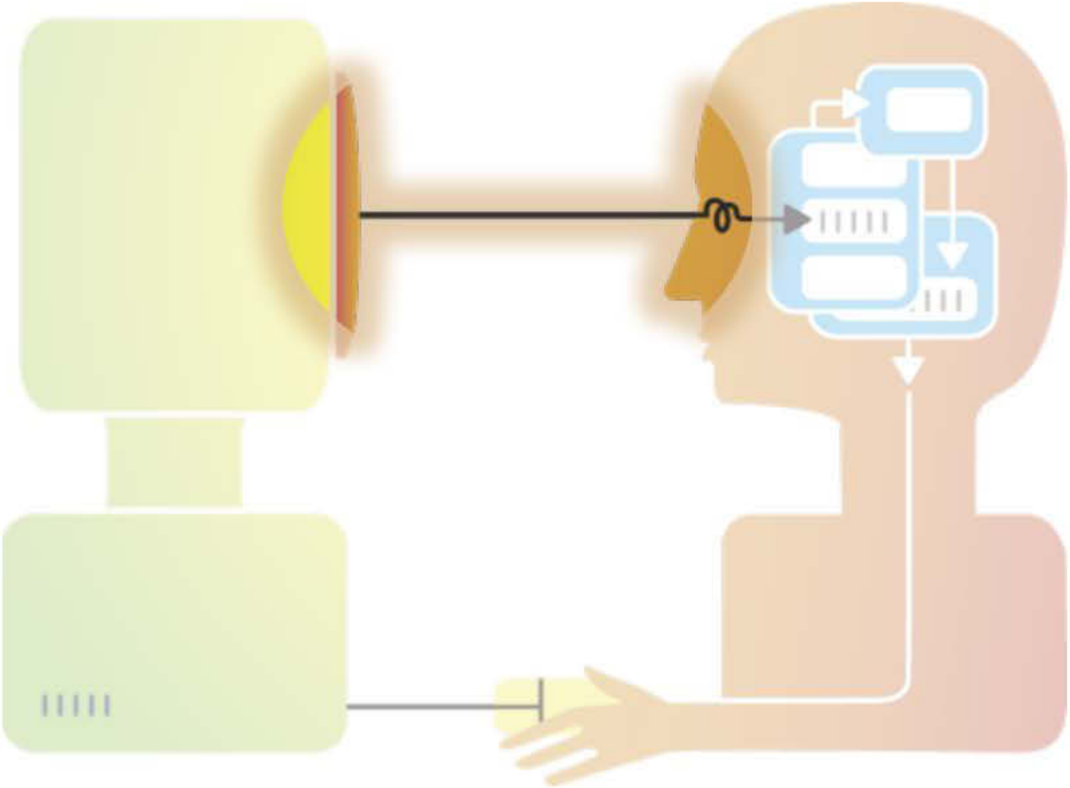
The bad:

- People (esp. experts) tend not to notice information that contradicts their mental model
- A “fresh pair of eyes” can be beneficial

The ugly:

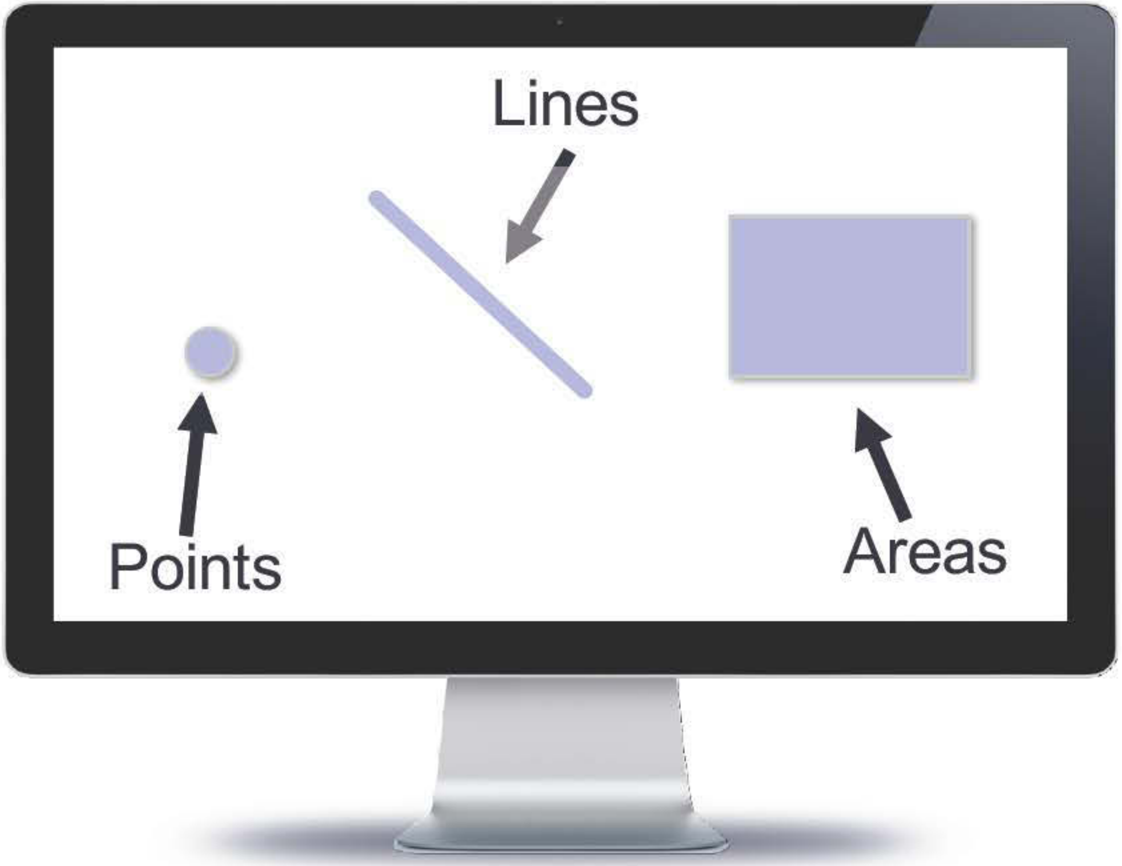
- Mental models are unavoidable: everyone has them, and they’re all different
- **Key:** be aware of how mental models form, how they shape perception, and how to support (or challenge) them

So what do we have to work with?



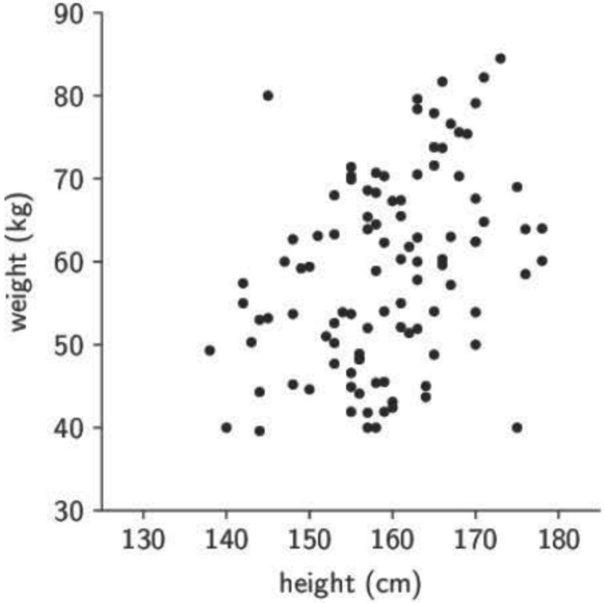
Graphical primitives

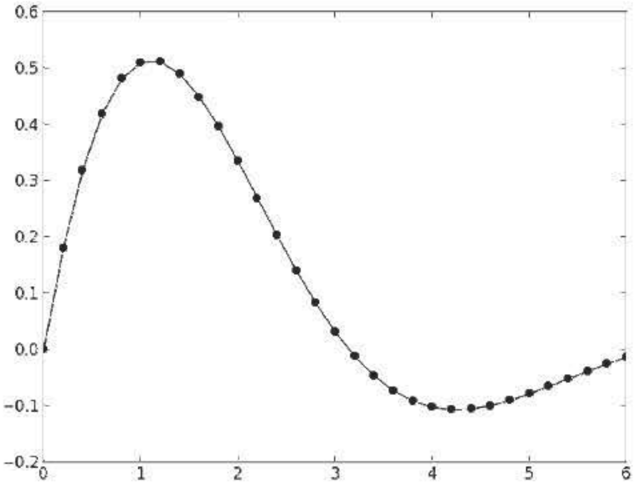
The images we draw are composed of marks: like ink

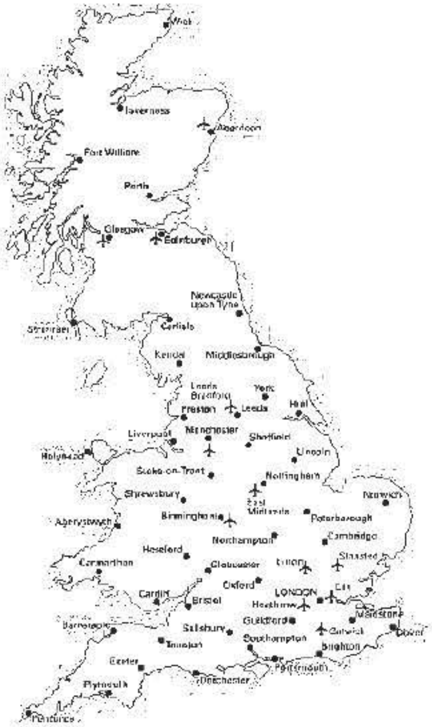


Visual dimension: position

- Encode information using **where** the mark is drawn
- Some examples:

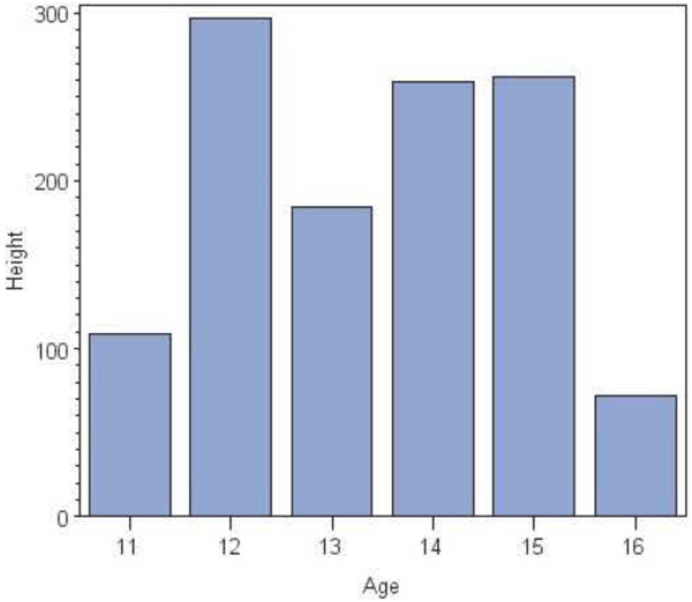


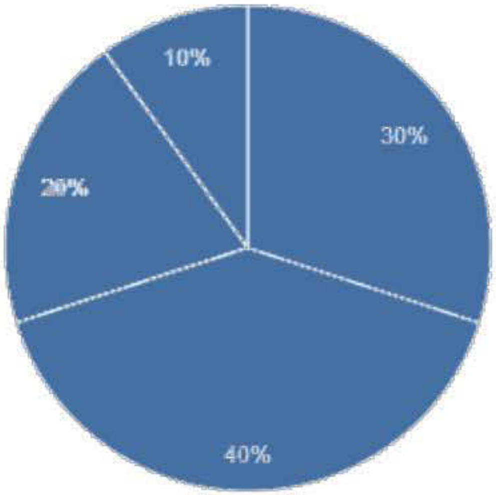


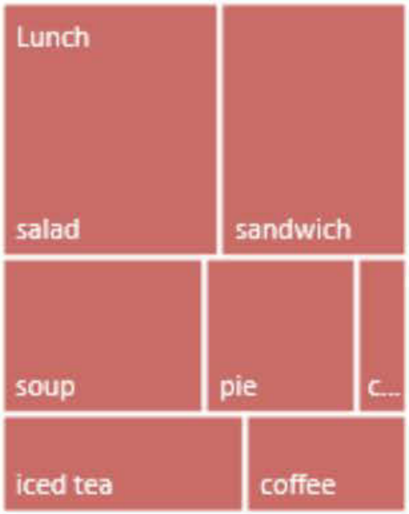


Visual dimension: size

- Encode information using **how big** the mark is drawn
- Examples:

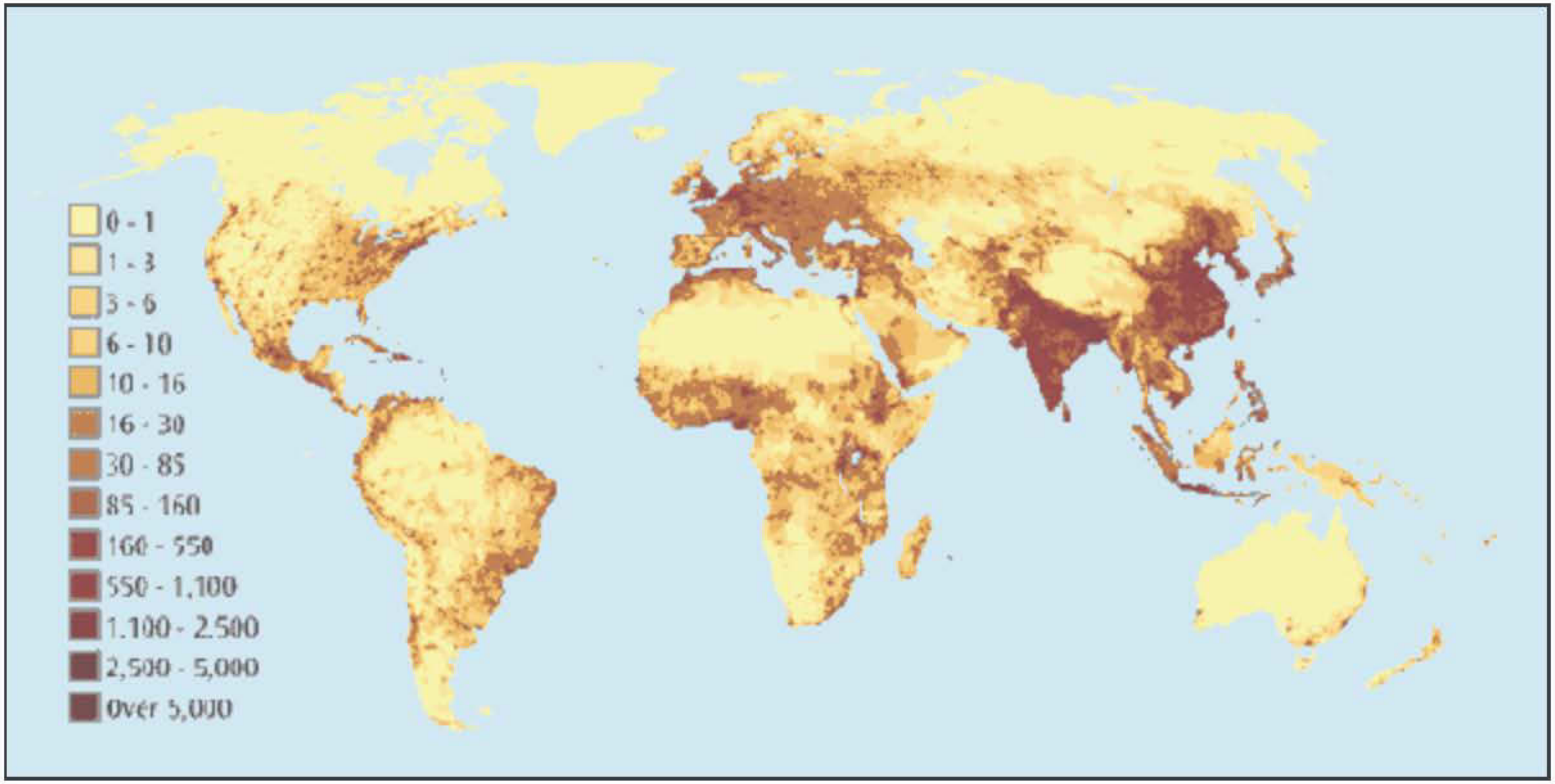






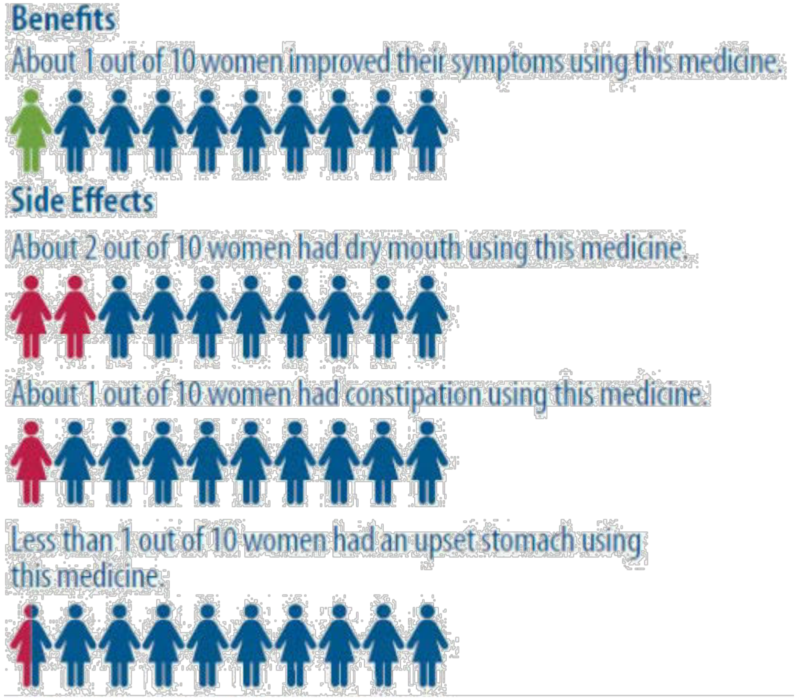
Visual dimension: value

- Encode information using **how dark** the mark is drawn
- Example:



Visual dimension: color

- Encode information using the **hue** of the mark
- Examples:



Benefits

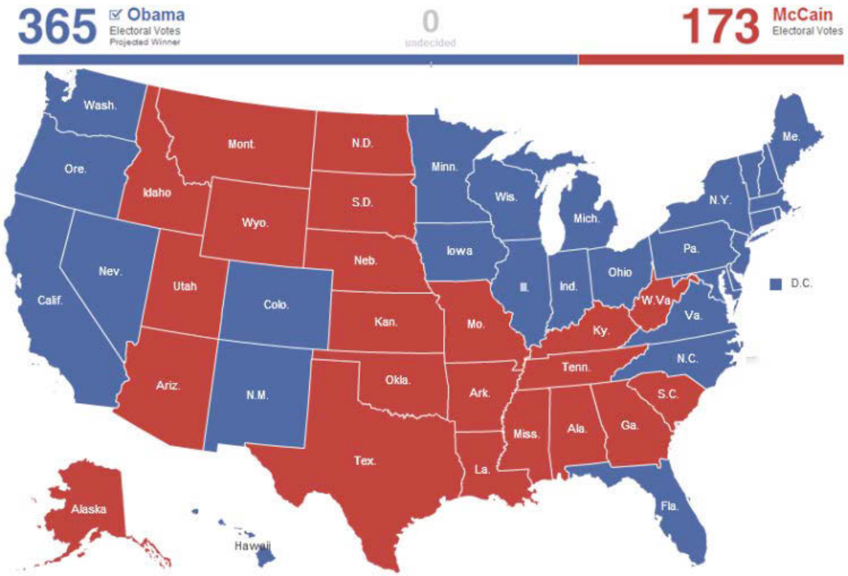
About 1 out of 10 women improved their symptoms using this medicine.

Side Effects

About 2 out of 10 women had dry mouth using this medicine.

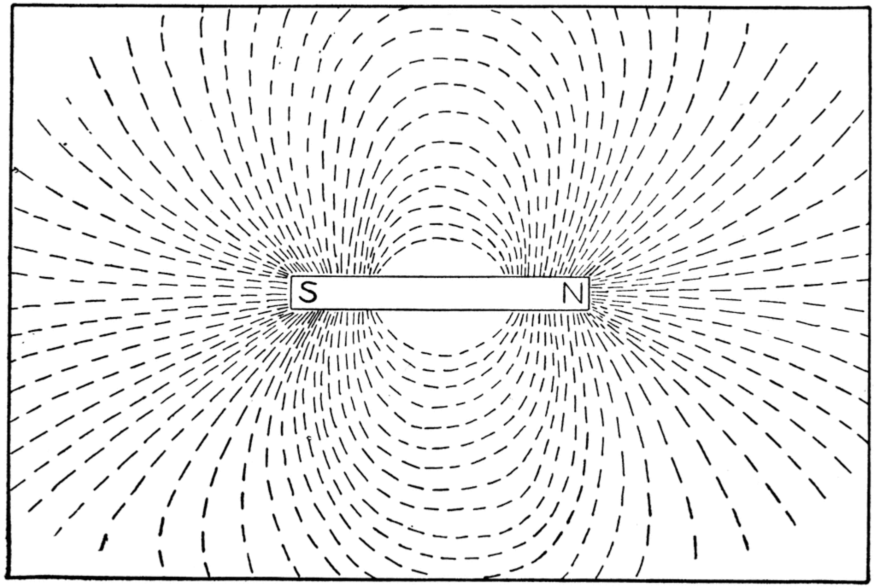
About 1 out of 10 women had constipation using this medicine.

Less than 1 out of 10 women had an upset stomach using this medicine.



Visual dimension: orientation

- Encode information using how the mark is **rotated**
- Examples:



Visual dimension: shape

- Encode information using how the mark is **shaped**
- Examples:

Discussion

What makes a **good** encoding?

Principle 1: expressiveness

- Encodes **all** the facts
- Example:

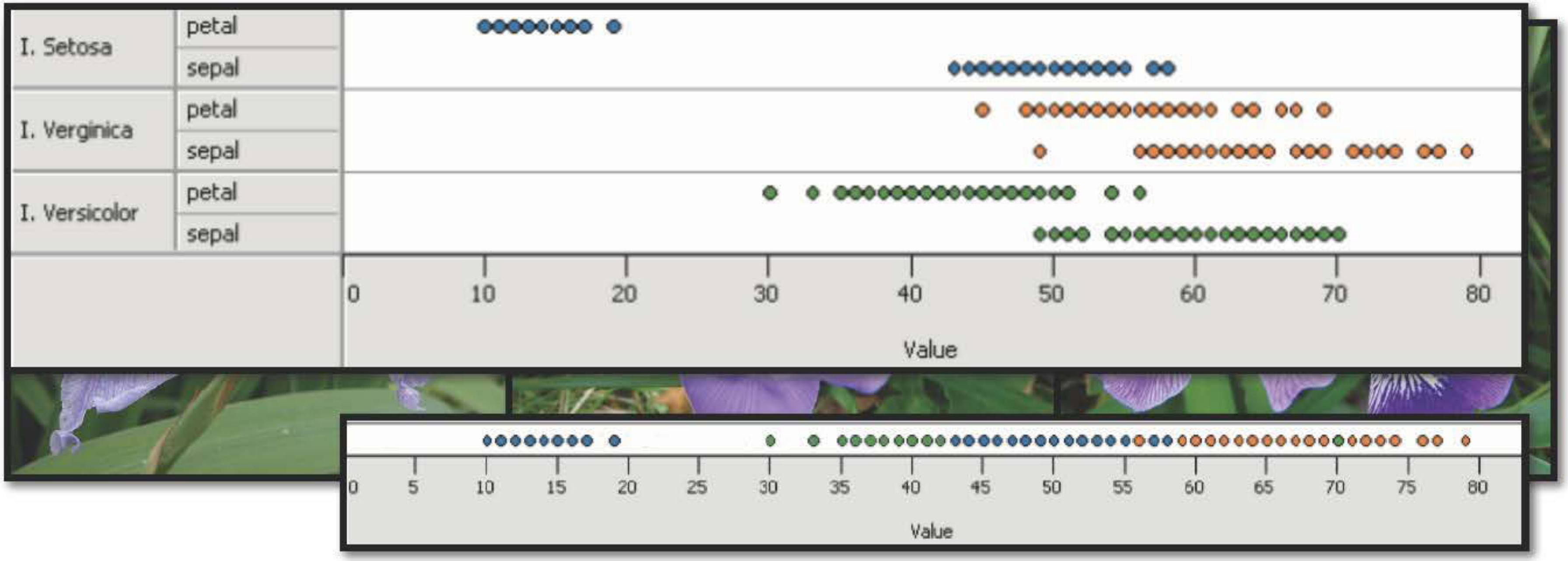
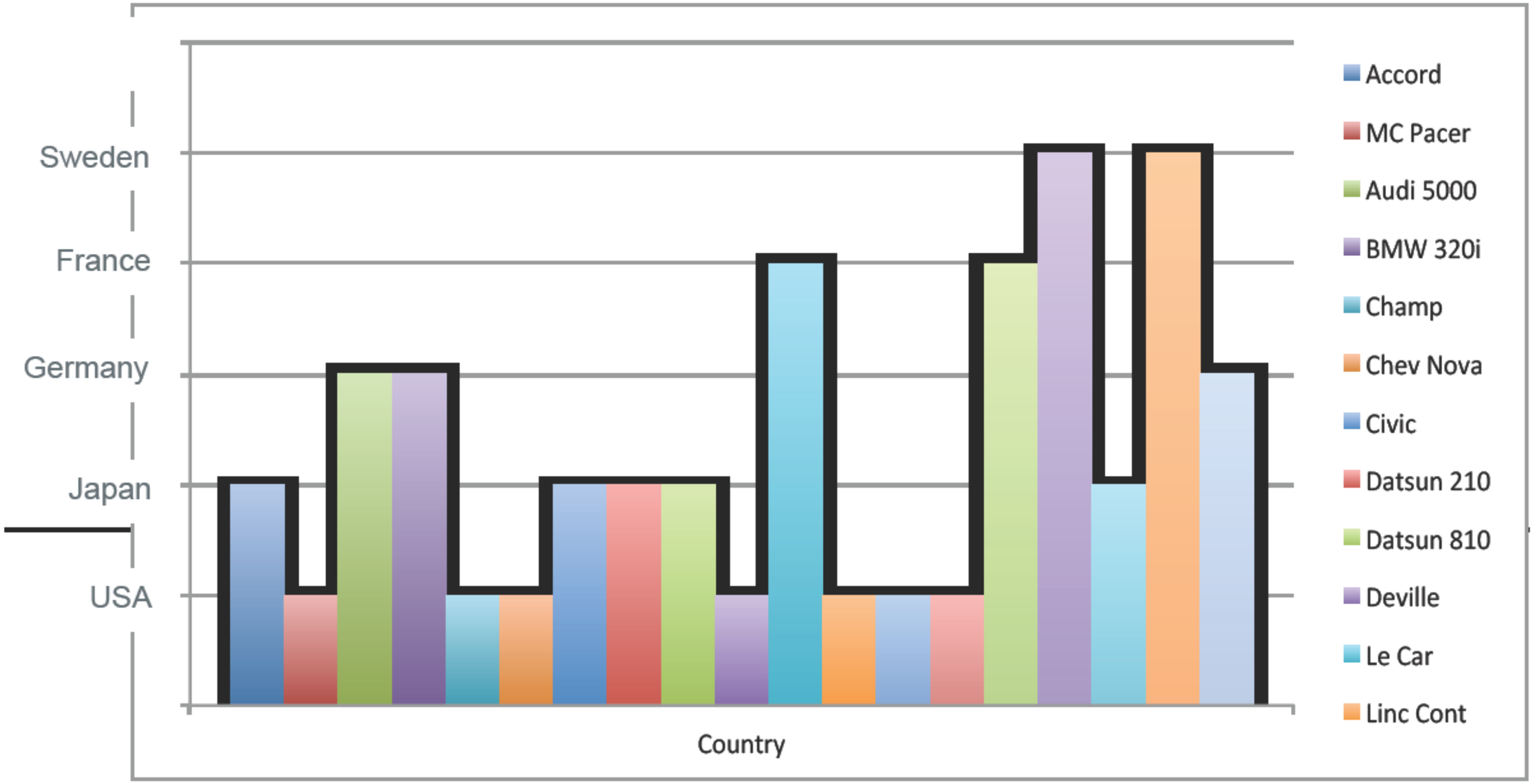


Fig. Courtesy of M Krzywinski

Principle 1: expressiveness

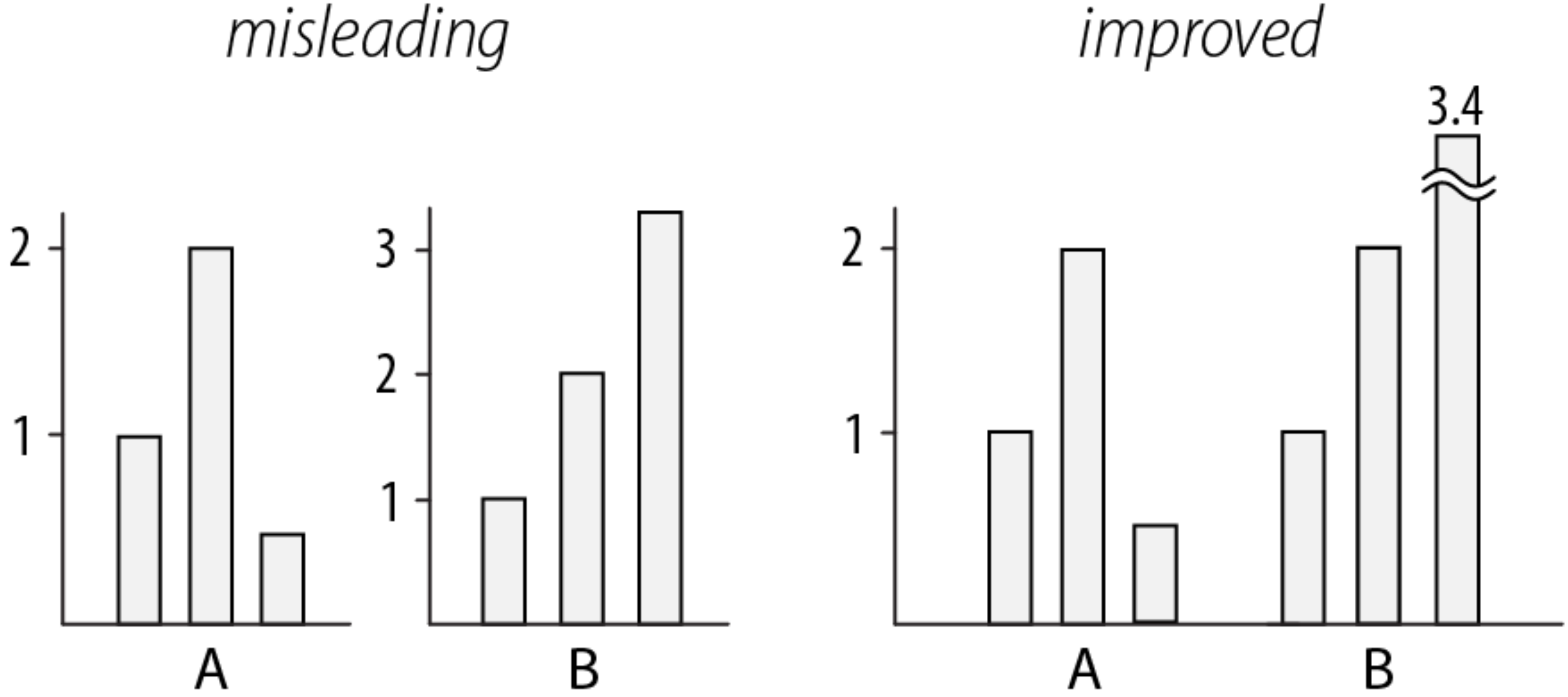
- Encodes **only** the facts
- Example:



Adapted from Mackinlay J (1986) Automating the design of graphical presentations of relational information.

Principle 2: consistency

- Use **consistent axes** when comparing charts

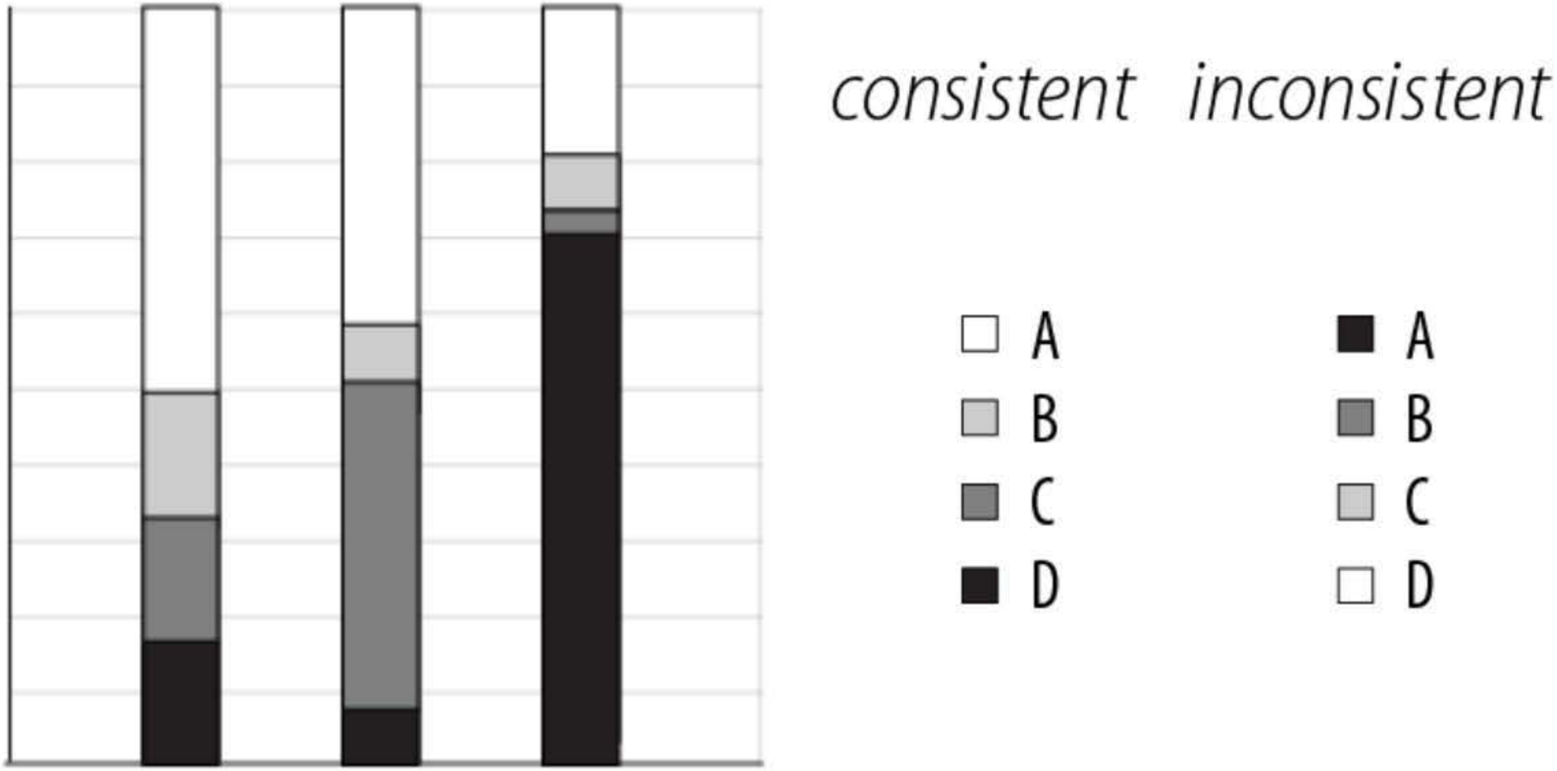


Raina SZ, et al. (2005) Evolution of base-substitution gradients in primate mitochondrial genomes. Genome Res 15: 665-673.

M. Krzwinski, behind every great visualization is a design principle, 2012

Principle 2: consistency

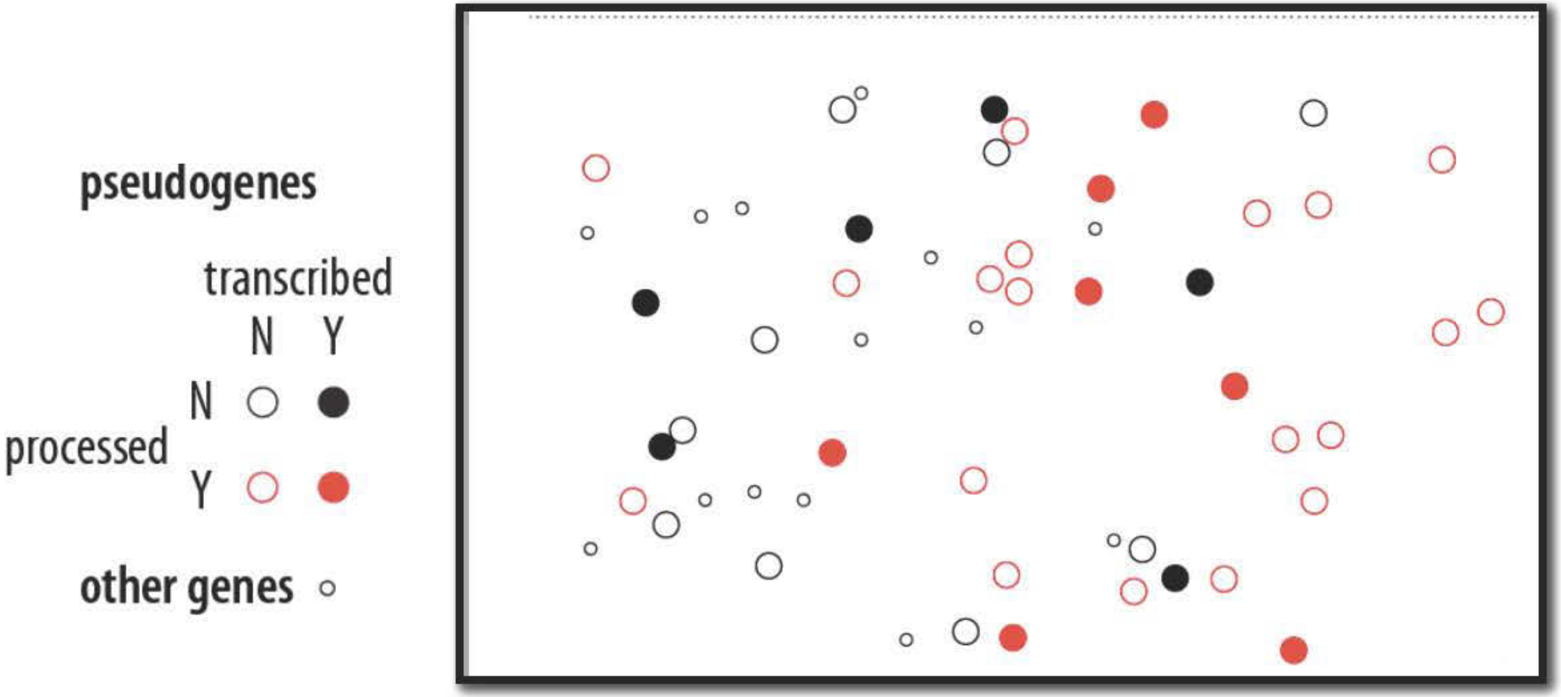
- A note on **legends**: order items according to appearance



M. Krzwinski, behind every great visualization is a design principle, 2012

Principle 2: consistency

- Visual variation should **reflect and enhance** the underlying variation in the data
- Avoid **visually similar** encodings for independent variables
- Example:



Principle 2: consistency

- Uniform size and alignment reduces visual complexity and aids interpretation
- Example:

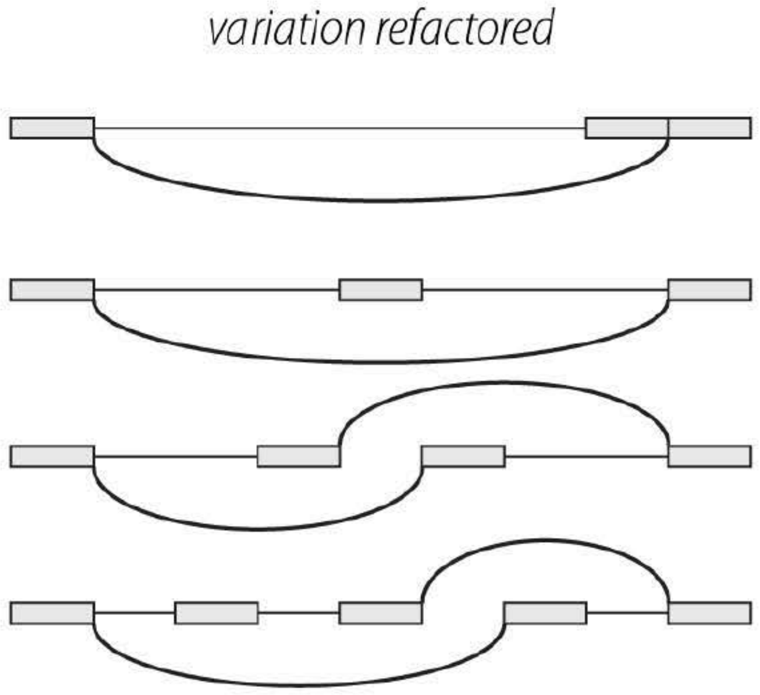
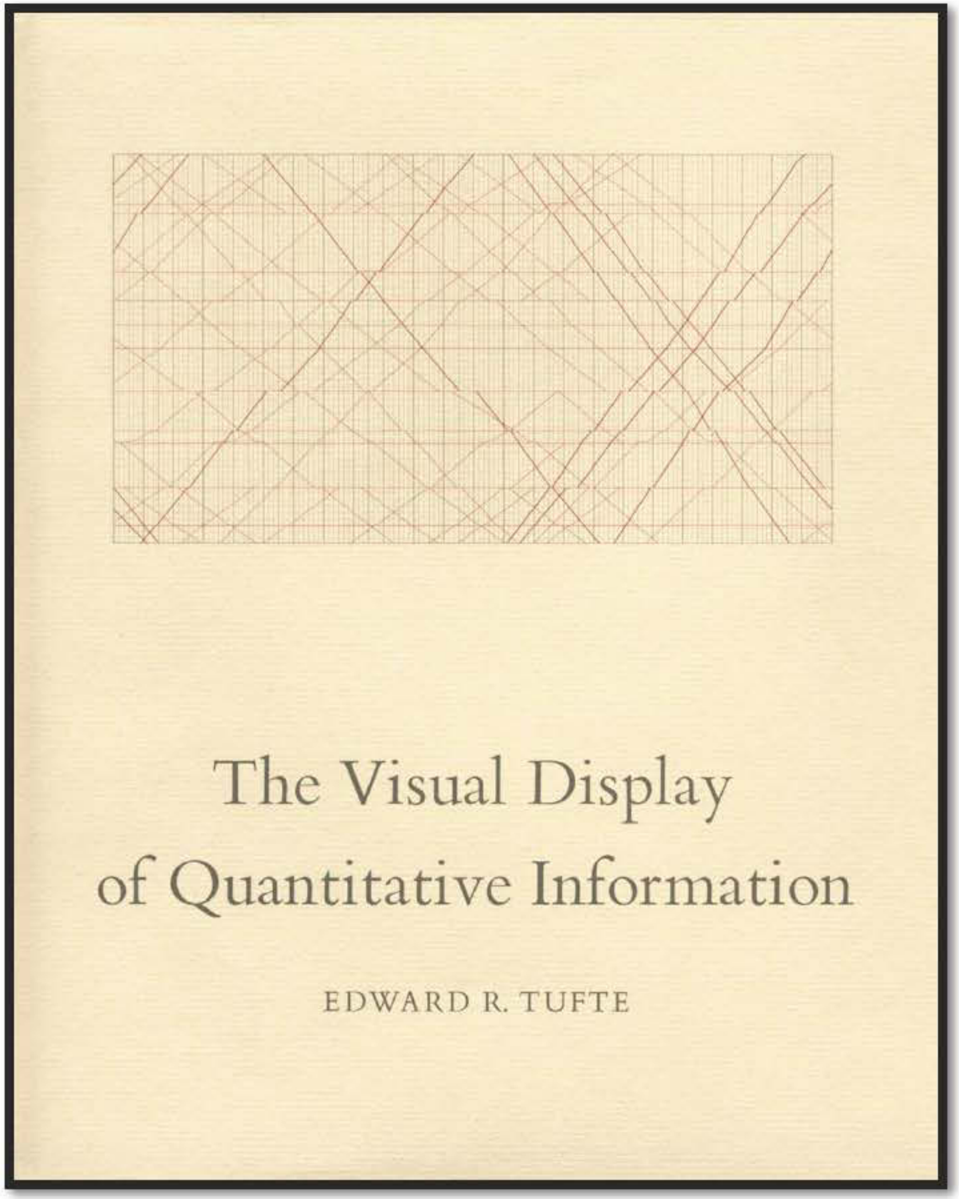


Fig. 1: Sharov AA et al. (2005) Genome-wide assembly and analysis of alternative transcripts in mouse. Genome Res 15: 748-754.
Fig. 2: M. Krzwinski, behind every great visualization is a design principle, 2012

Tufte, 1983

“Above all else, show the data.”



The Visual Display of Quantitative Information

EDWARD R. TUFTE

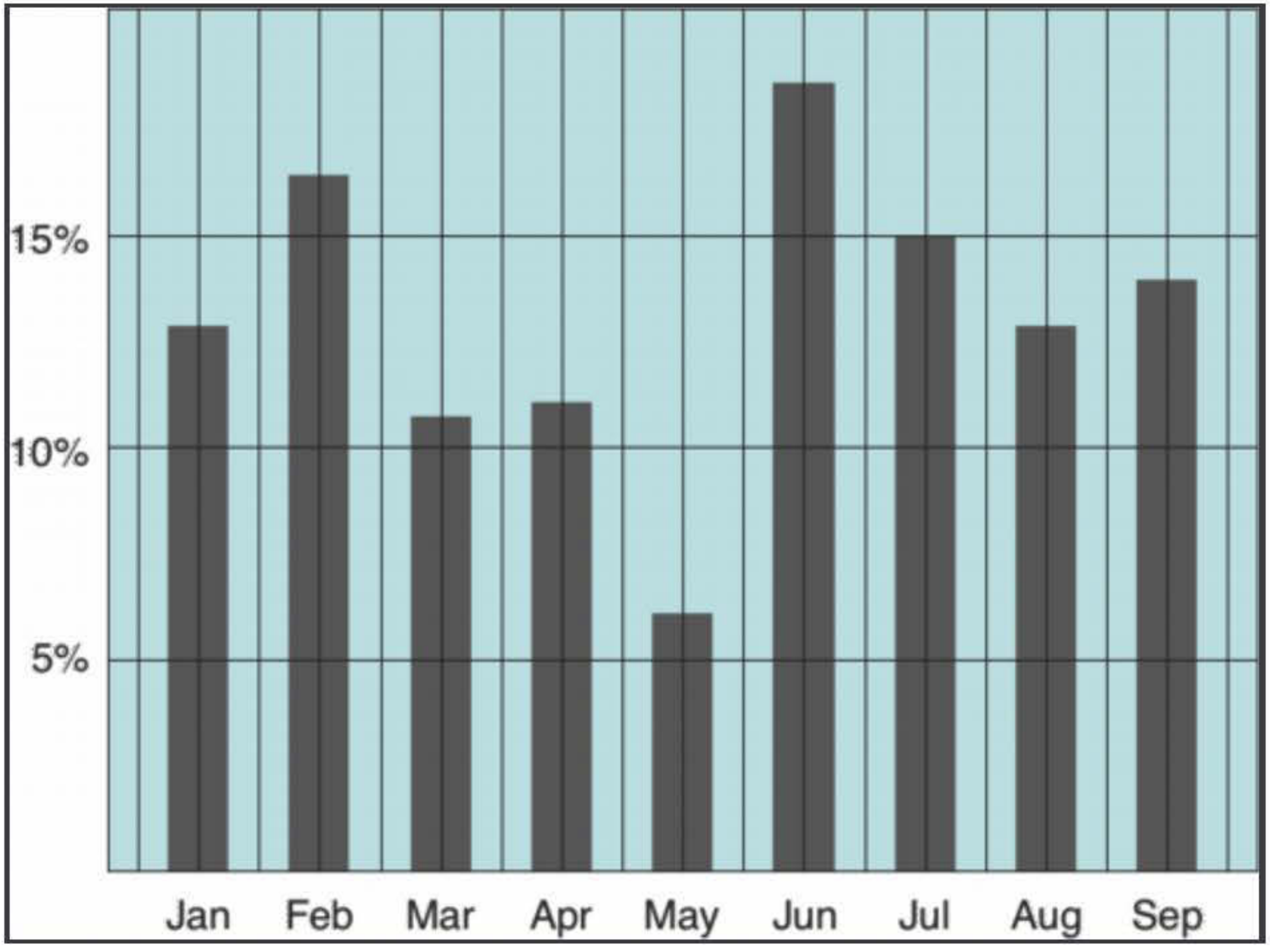
Tufte, 1983

$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used to print the graphic}}$$

= proportion of a graphic's ink devoted to the non-redundant display of data-information

= 1 - proportion of a graphic that can be erased

Tufte: maximize the data-ink ratio



Familiar example

11th St
Greenw
3rd Ave
2nd Ave
4th Ave
1st Ave
E 14th St
E 9th St
FDR Dr S
and Tunnel
Hudson St
Varick St
Thompson St
Mercer St
Lafayette St
Prince St
E 4th St
E 6th St
Avenue C
Avenue D
FDR Dr N S
12th Ave
6th Ave
Broadway
Lafayette St
E Houston St
Stanton St
Delancey St
Pitt St
New York
Mott St
Bowery
Allen St
Grand St
Madison St
12th Ave
Broadway
FDR Dr N
FDR Dr S
Pearl St
FDR Dr N
South St
erry
John St
Water St
278
Brooklyn Qu
Flushi

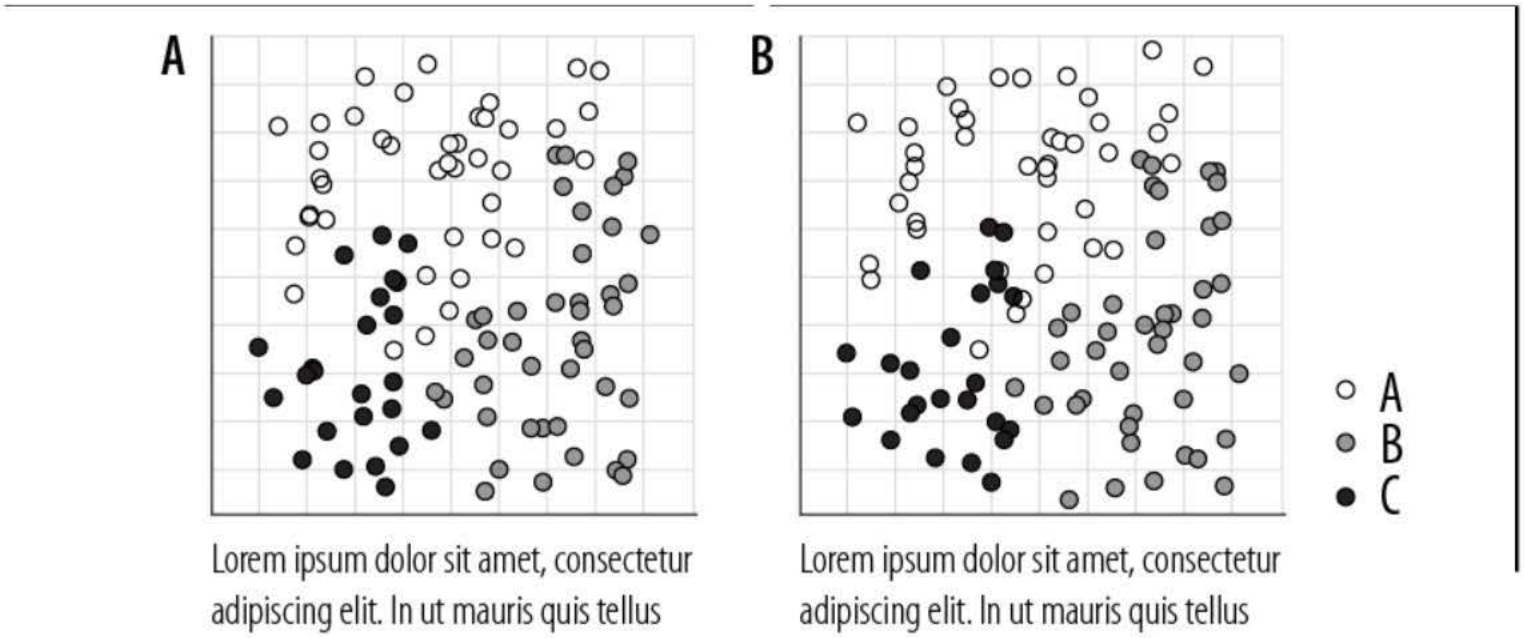
Walker St
Church St
Frederick Dou
Joss Landry
Cham-
ber St
Reade St
Lafayette St
West St
Joe Dorso Hwy
South Dorso St
South St
Brooklyn Bridge
478
278

Discussion

- What do you think of the data-ink ratio?
- Consider ways to **maximize** it…

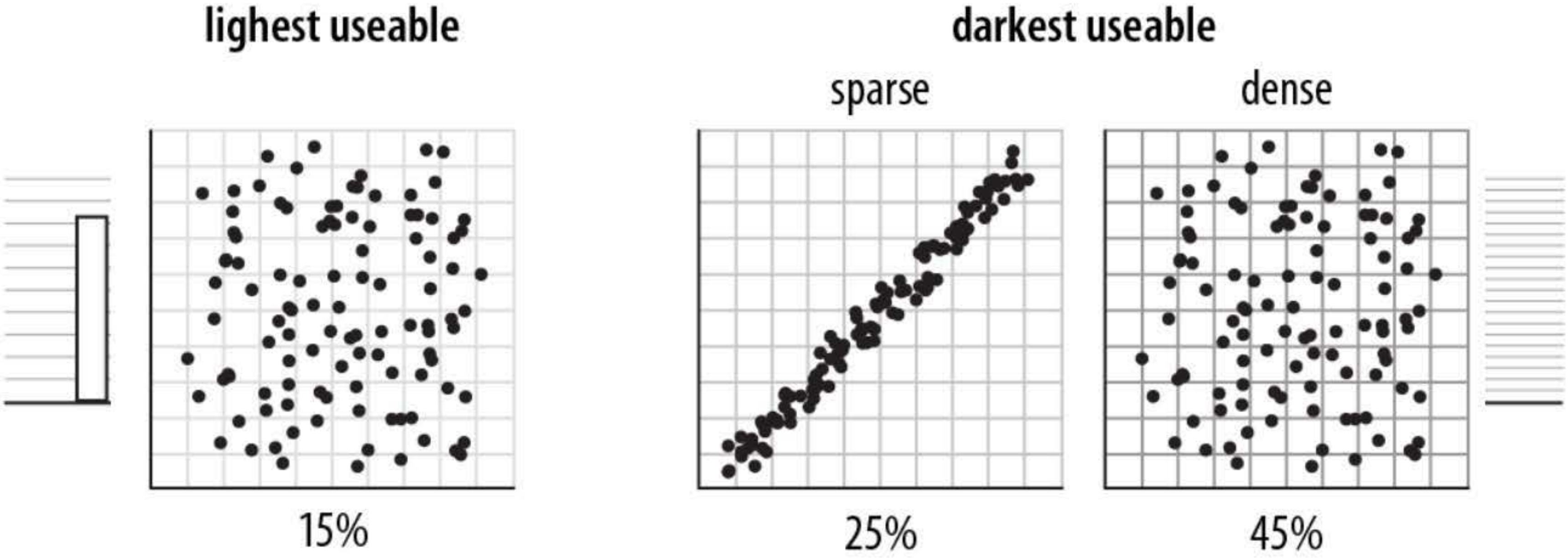
Principle 3: importance ordering

- Avoid unnecessary containment and repetition
- Example



Principle 3: importance ordering

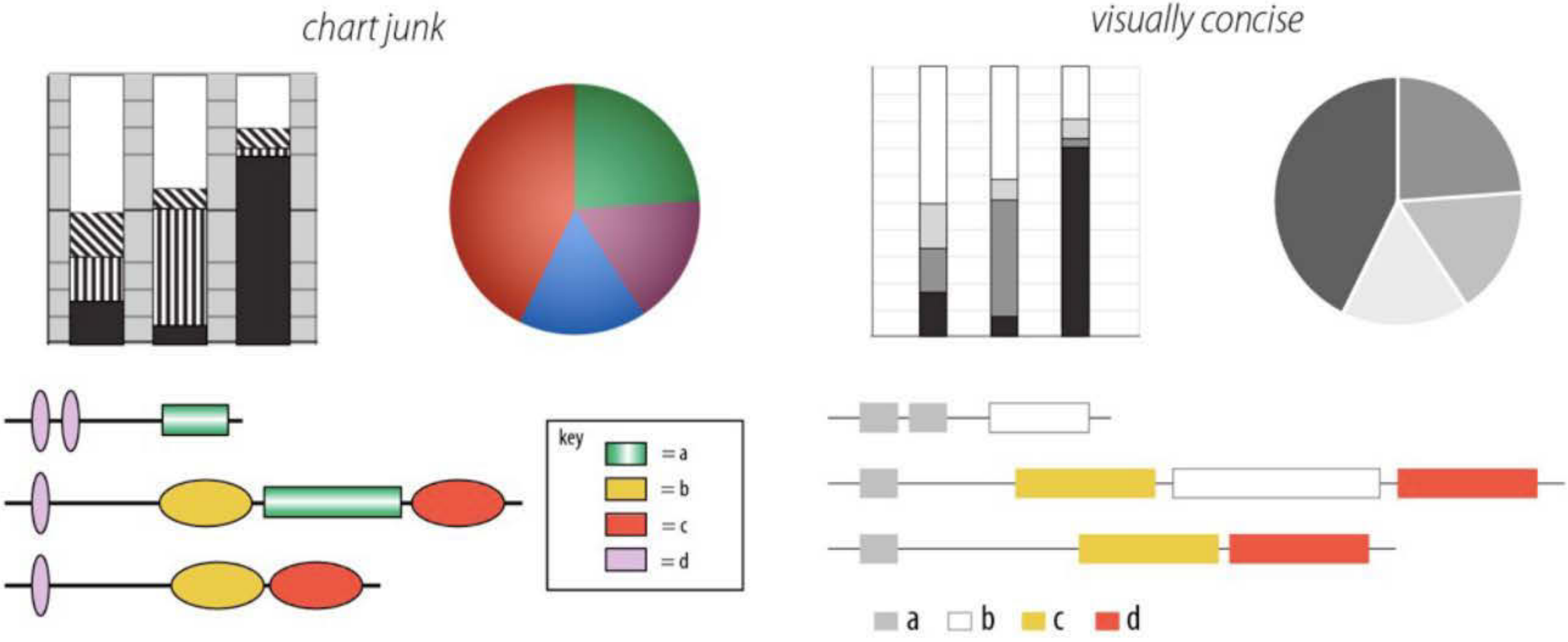
- Navigational aids shouldn’t compete with data
- Avoid: heavy **axes**, **error bars** and **glyphs**



Heer J, Bostock M (2010) Crowdsourcing graphical perception: using mechanical turk to assess visualization design. Proceedings of the 28th international conference on Human factors in computing systems. Atlanta, Georgia, USA: ACM. pp. 203-212.

Principle 3: importance ordering

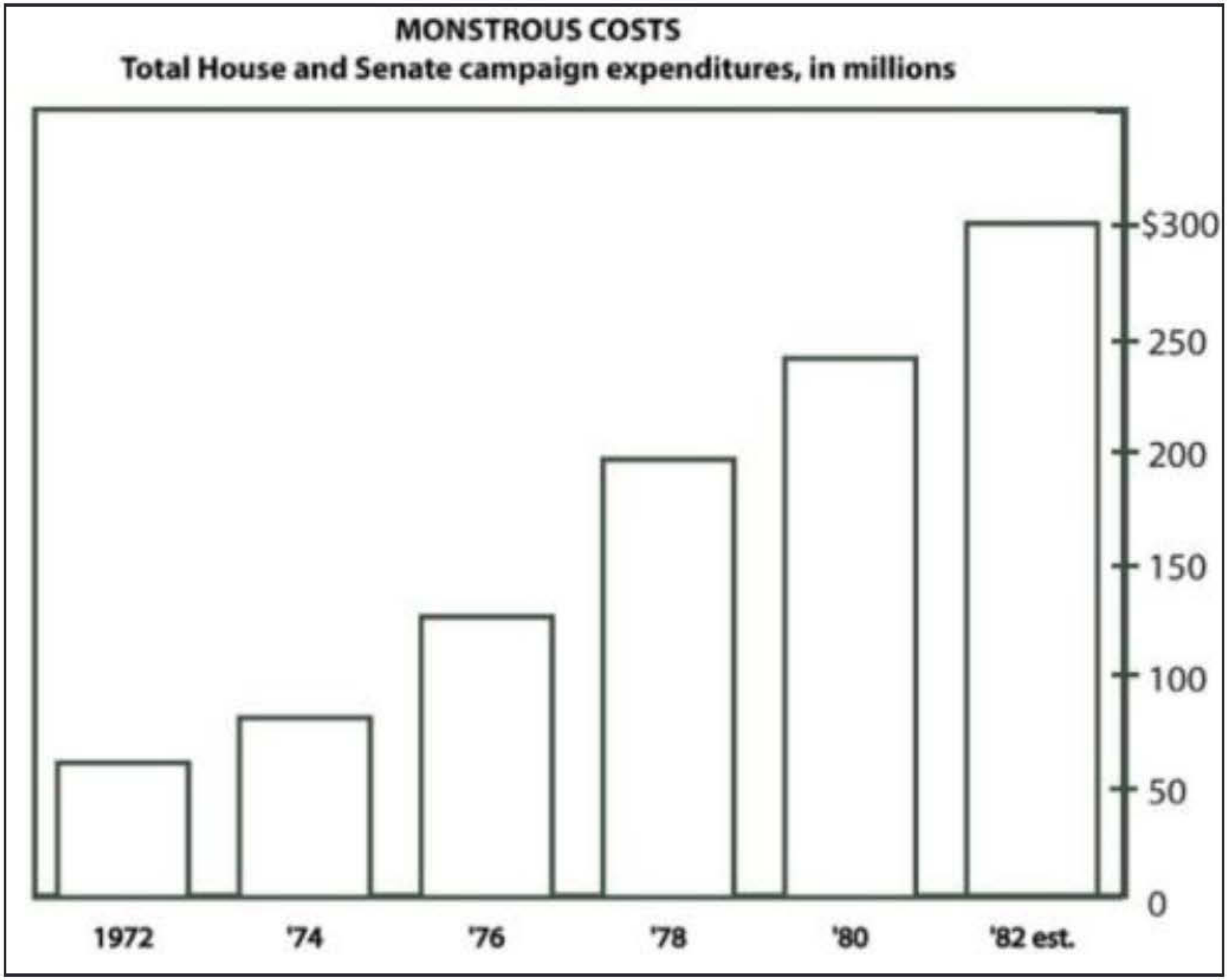
- Simplify, simplify, simplify…



Sharov AA, et al (2006) Genome Res 16: 505-509.
Peterson J, et al. (2009) Genome Res 19: 2317-2323.
Thomson NR, et al. (2005) Genome Res 15: 629-640.
DB, Ko MS (2005) Genome Res 15: 748-754.

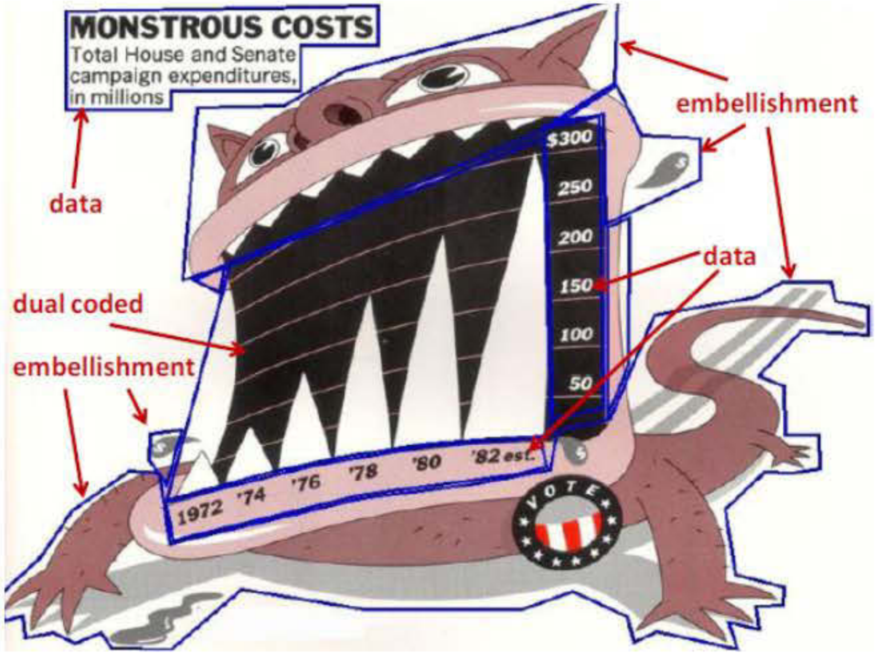
M. Krzwinski, behind every great visualization is a design principle, 2012

A caveat: “chart junk” and recall



Bateman et al. “Useful Junk? The Effects of Visual Embellishmenton Comprehension and Memorability of Charts”, CHI 2010

Chart junk and eye gaze



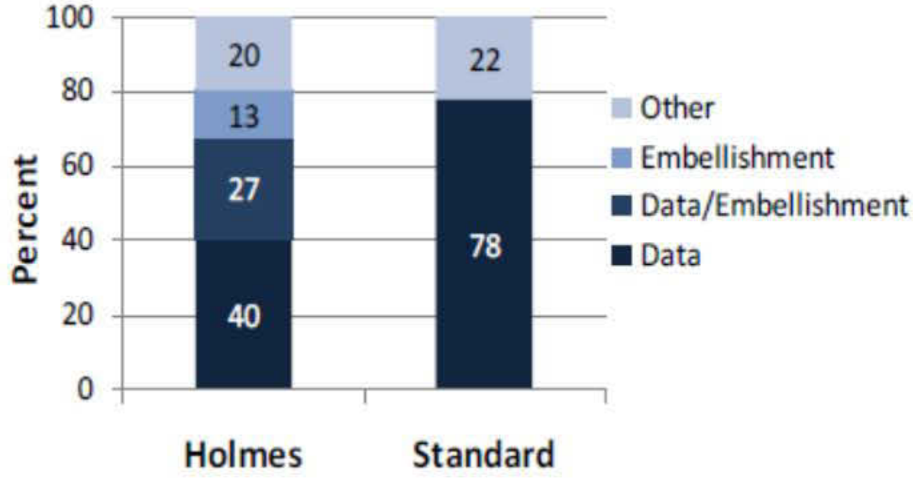


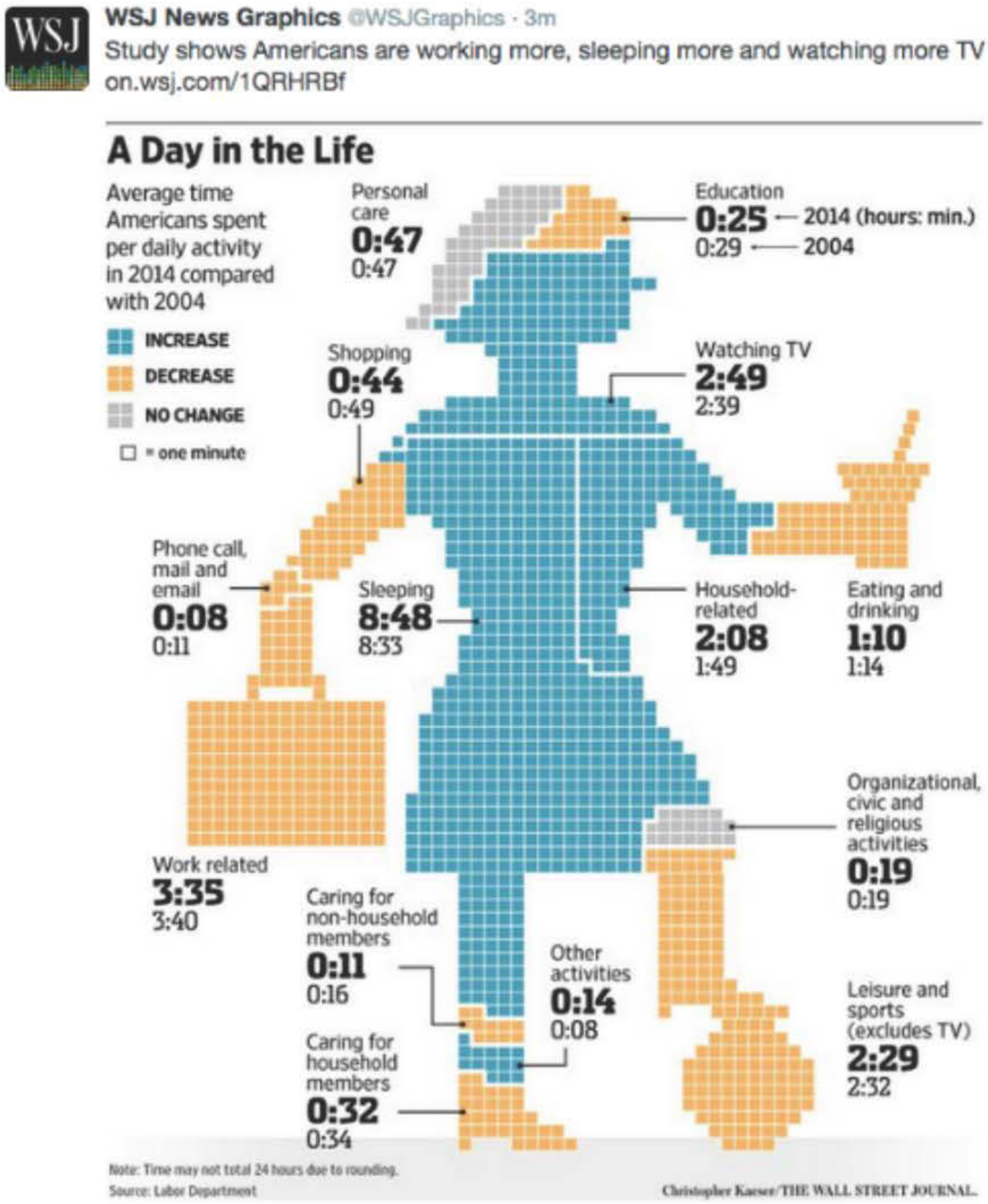
Figure 9. Percentage of on-screen time spent looking at different chart elements for Holmes and Plain charts.

Lab 1: Deconstructing Data Graphics

- Break into groups of 2-3 people, and go to:

  jcrouser.github.io/datavis/lab1.html

- During this lab, we'll **critique** some professionally-made visualizations using these principles
- Try to think about the following questions:
  - What is the **first thing you notice** about this visualization?
  - What **point** is this visualization trying to make?
  - Who is the **intended audience**?
  - What is the visualization **doing well**?
  - What **problems** do you see with the visualization design?
  - **Why** do you think the designer made those choices?



WSJ News Graphics @WSJGraphics · 3m

Study shows Americans are working more, sleeping more and watching more TV on.wsj.com/1QRHRBf

Note: Time may not total 24 hours due to rounding.

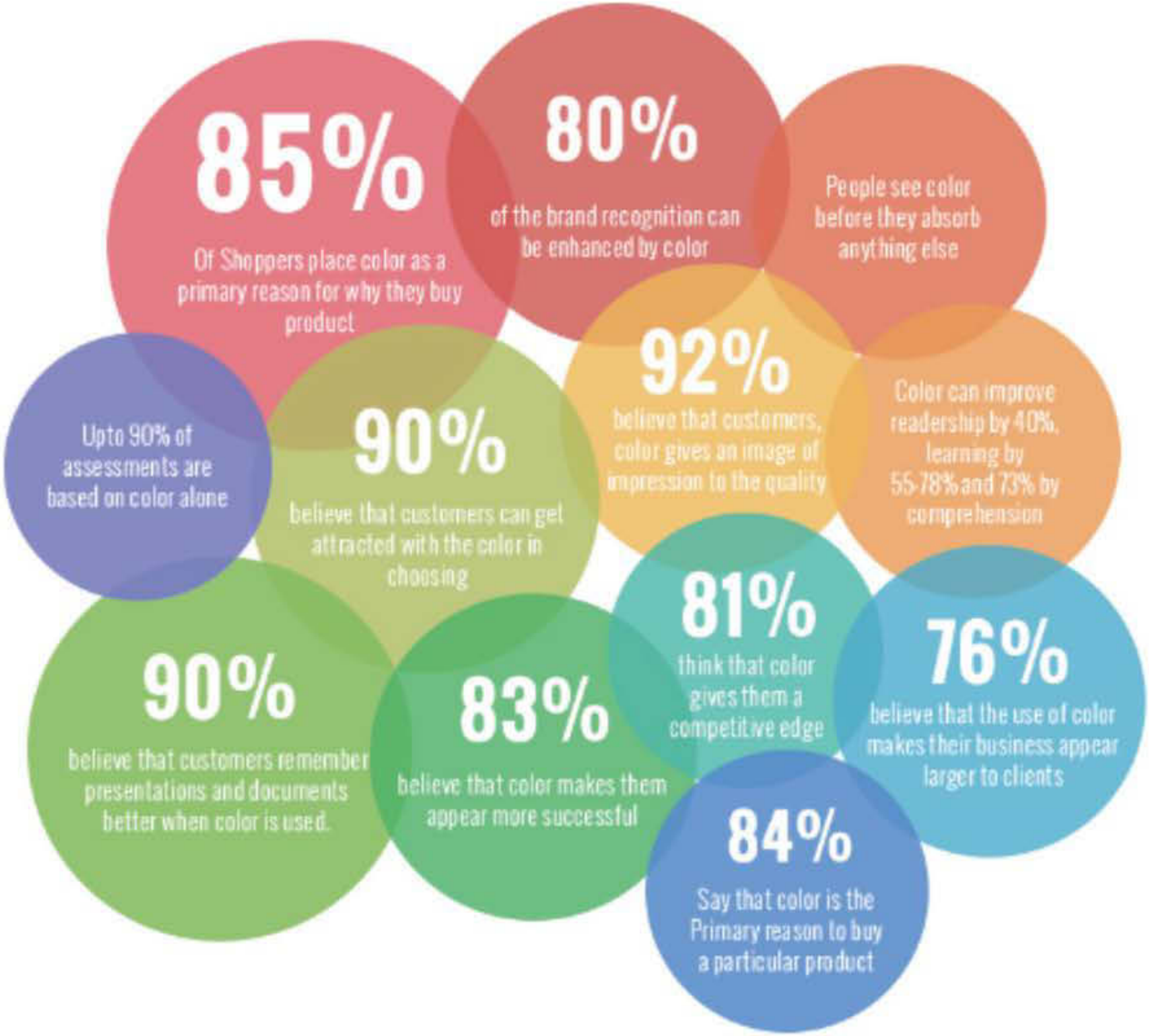
Source: Labor Department

Christopher Kaeser/THE WALL STREET JOURNAL.

What your

BRAND COLOR

SAYS ABOUT YOUR **BUSINESS**



85%

Of Shoppers place color as a primary reason for why they buy product

80%

of the brand recognition can be enhanced by color

People see color before they absorb anything else

92%

believe that customers, color gives an image of impression to the quality

Color can improve readership by 40%, learning by 55-78% and 73% by comprehension

Upto 90% of assessments are based on color alone

90%

believe that customers can get attracted with the color in choosing

90%

believe that customers remember presentations and documents better when color is used.

83%

believe that color makes them appear more successful

81%

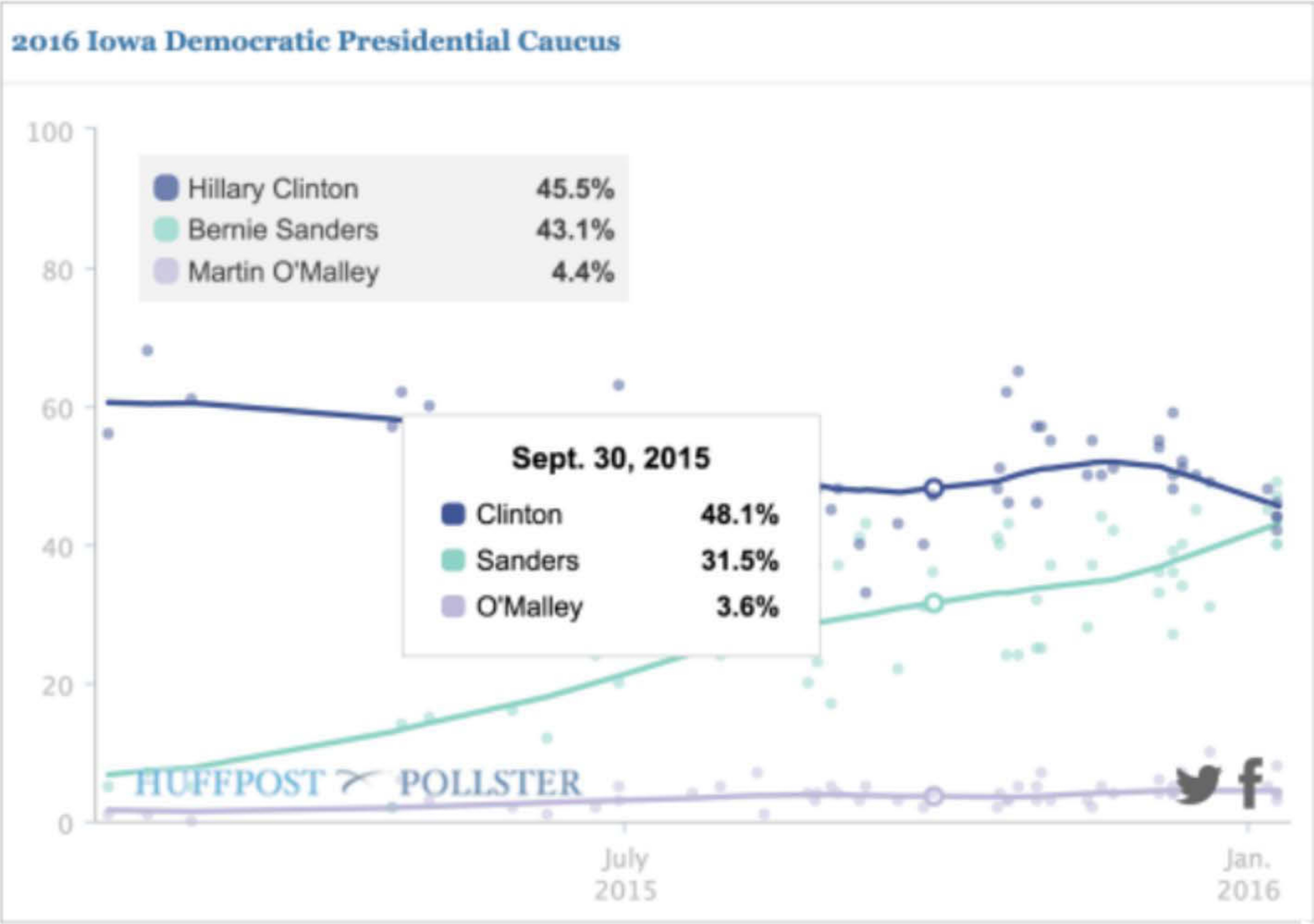
think that color gives them a competitive edge

76%

believe that the use of color makes their business appear larger to clients

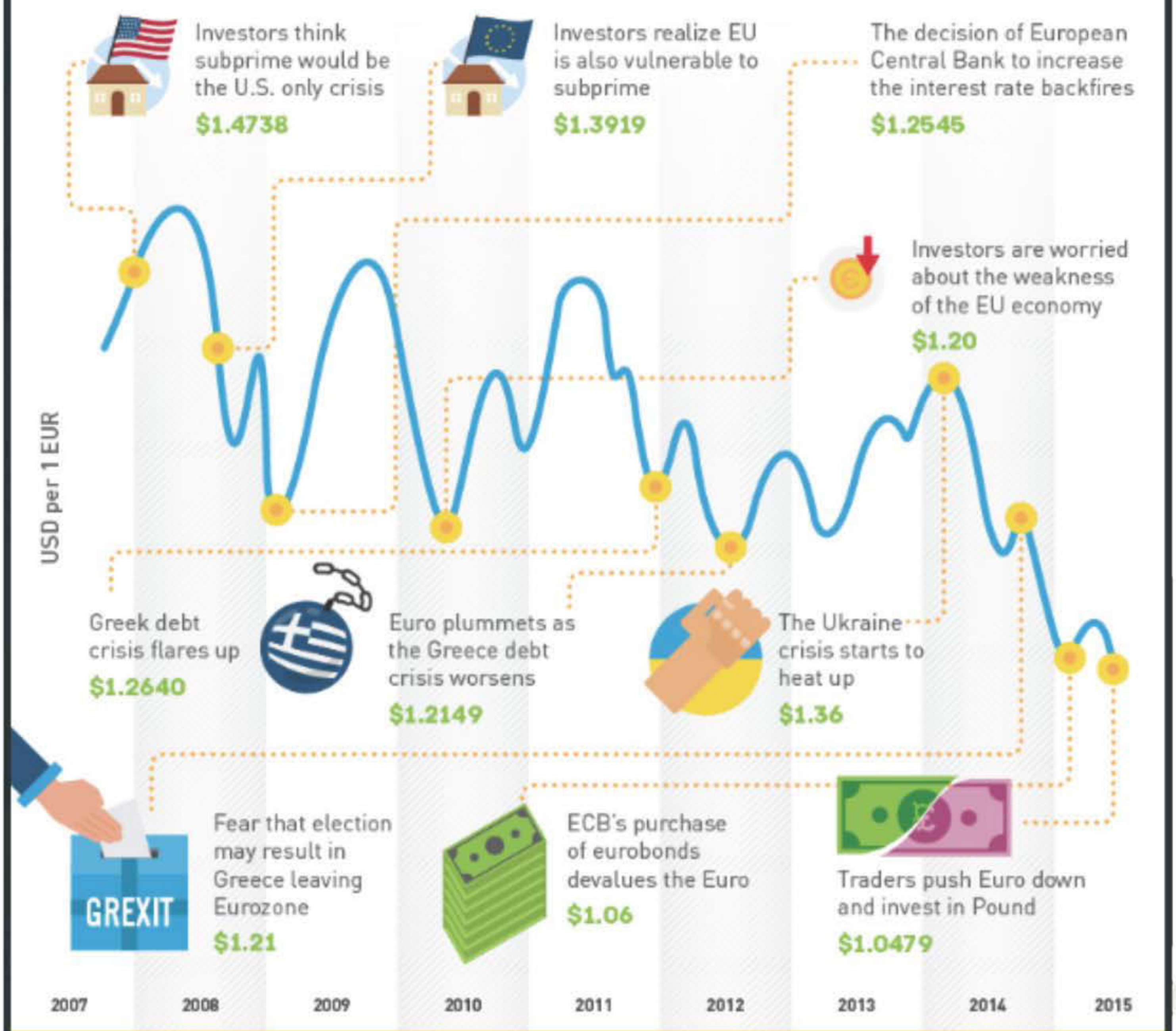
84%

Say that color is the Primary reason to buy a particular product



2016 Iowa Democratic Presidential Caucus

EVENTS CONTRIBUTING TO DROP OF EURO



Investors think subprime would be the U.S. only crisis
$1.4738

Investors realize EU is also vulnerable to subprime
$1.3919

The decision of European Central Bank to increase the interest rate backfires
$1.2545

Investors are worried about the weakness of the EU economy
$1.20

Greek debt crisis flares up
$1.2640

Euro plummets as the Greece debt crisis worsens
$1.2149

The Ukraine crisis starts to heat up
$1.36

Fear that election may result in Greece leaving Eurozone
$1.21

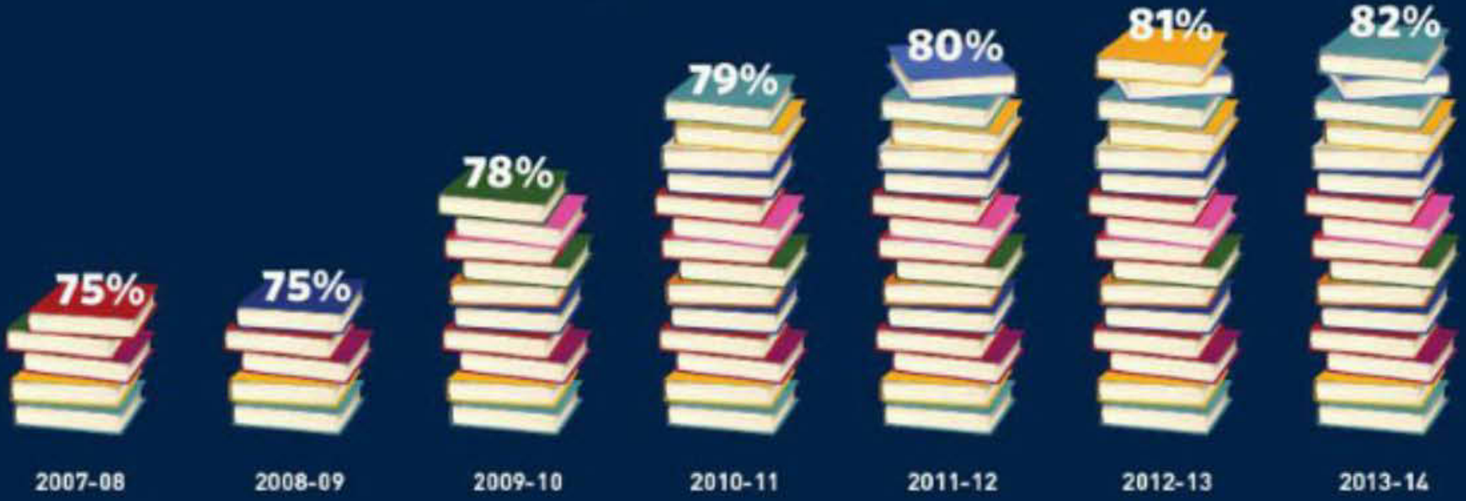
ECB's purchase of eurobonds devalues the Euro
$1.06

Traders push Euro down and invest in Pound
$1.0479

UNDER PRESIDENT OBAMA,

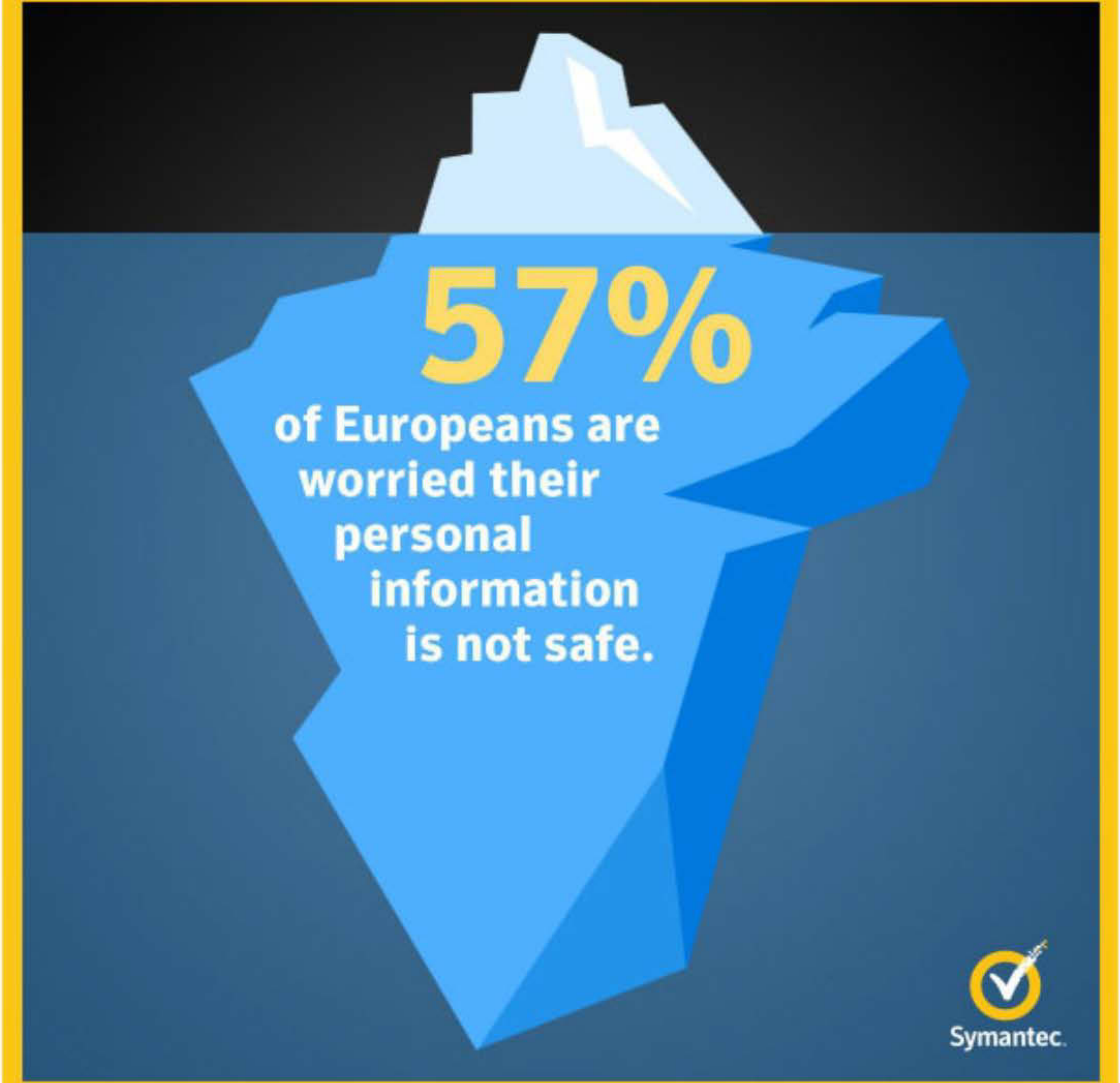
MORE STUDENTS ARE EARNING THEIR HIGH SCHOOL DIPLOMAS THAN EVER BEFORE

HIGH SCHOOL GRADUATION RATE



#LeadOnEducation

SOURCE: U.S. DEPARTMENT OF EDUCATION, NATIONAL CENTER FOR EDUCATION STATISTICS



57%

of Europeans are worried their personal information is not safe.

Symantec.

Illinois

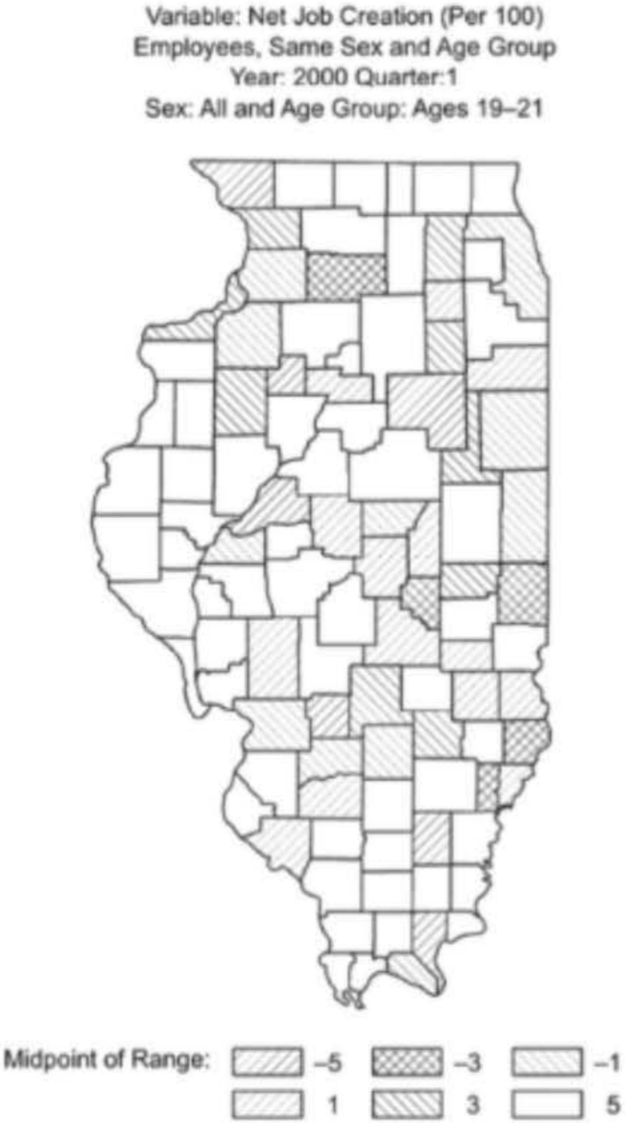

Variable: Net Job Creation (Per 100)
Employees, Same Sex and Age Group
Year: 2000 Quarter:1
Sex: All and Age Group: Ages 19–21
Midpoint of Range:
–5
–3
–1
1
3
5

Fig. 5.7 Job creation for young workers, by county, Illinois

Who do Nike sponsor?

International sports and events sponsor

The American based company is the largest sports supplier in the world, suppling equipment, shoes and apparel.

1,016

athletes sponsored worldwide

$230m

spent on athlete sponsors

485 Male Footballers

120 Male Track and Field Athletes

93 Basketball Players

72 Female Track and Field Athletes

64 Ice Hockey Players

34 Baseball Players

29 American Footballers

25 Male Golfers

31 Male Tennis Players

1 Figure Skater

21 Cricketers

10 Swimmers

1 Martial Artist

4 Female Golfers

21 Female Tennis Players

2 Cyclists

15 Female Footballers

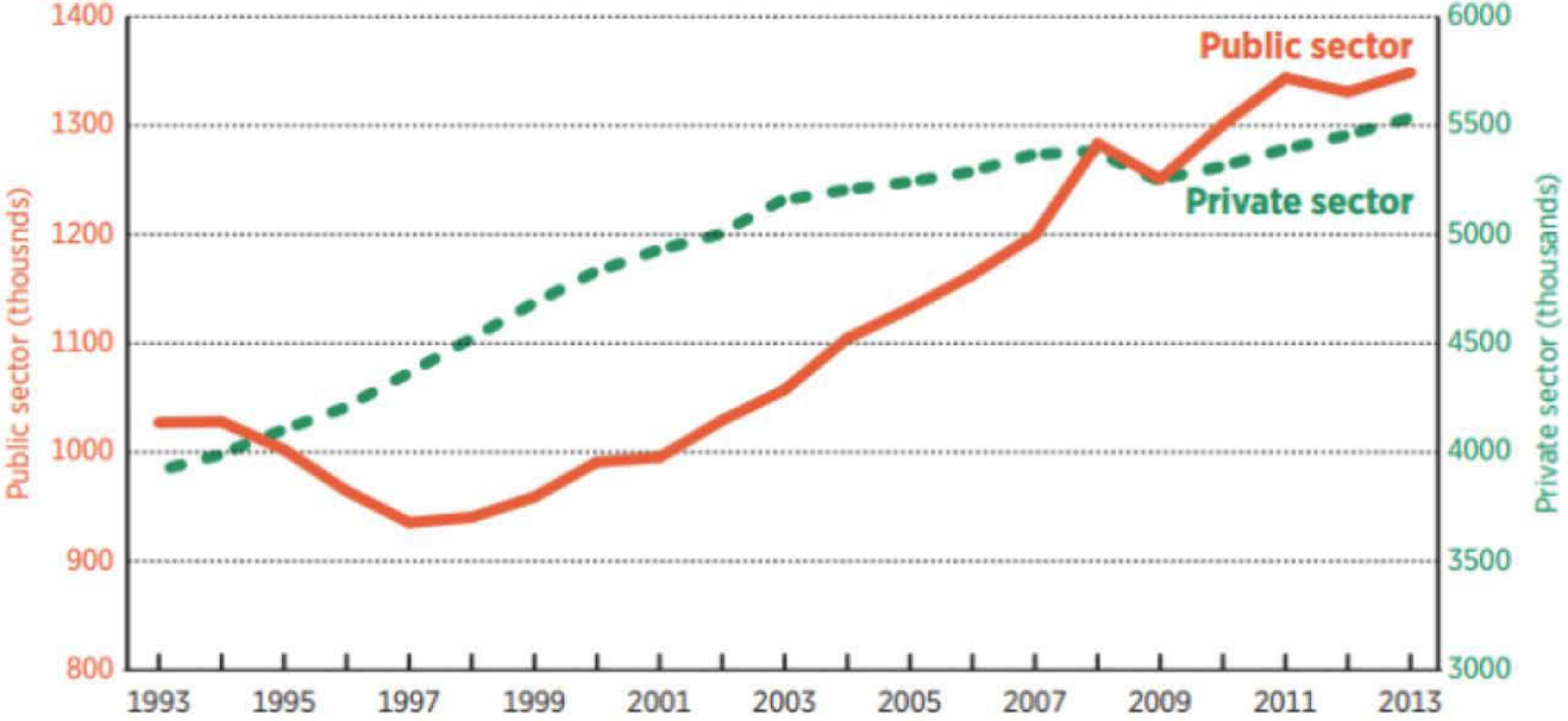
2 Skiers

8 Rugby Players

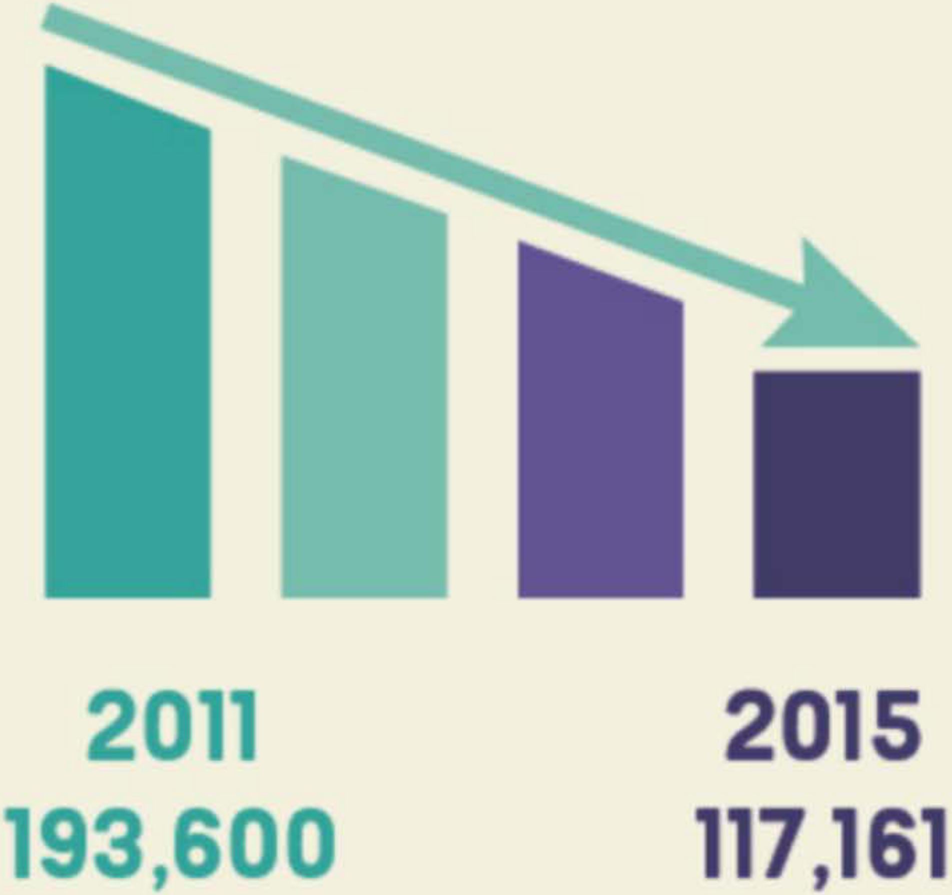
2 Squash Players

8 Boxers

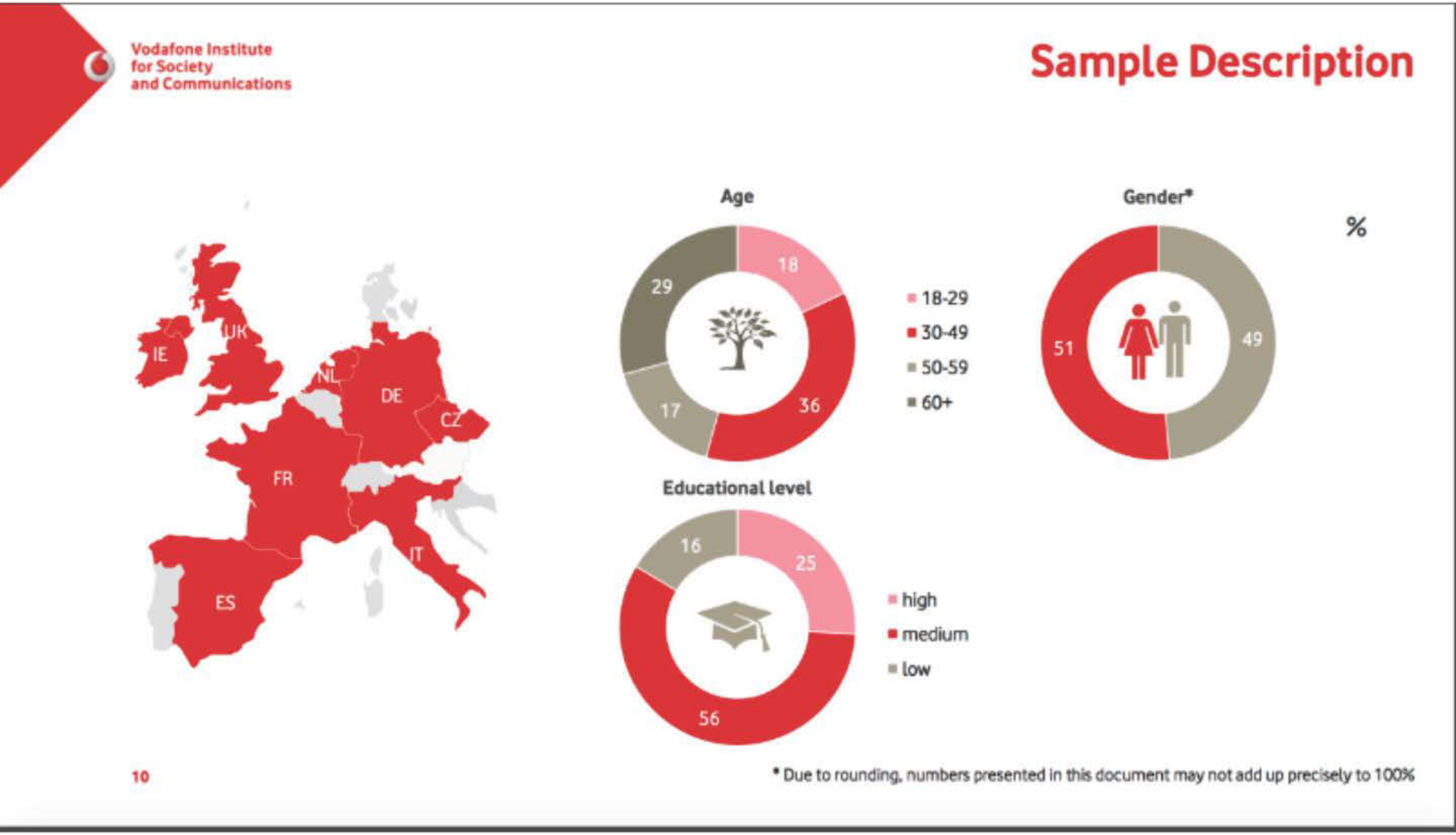
Figure 10: Public- and private-sector jobs (000s) in Ontario, 1993–2013



Source: Statistics Canada, CANSIM Table 282-0089: *employment by class of worker and sex, seasonally adjusted and unadjusted; Ontario; Public sector and private sector employees; Both sexes; Seasonally adjusted (x 1,000).*



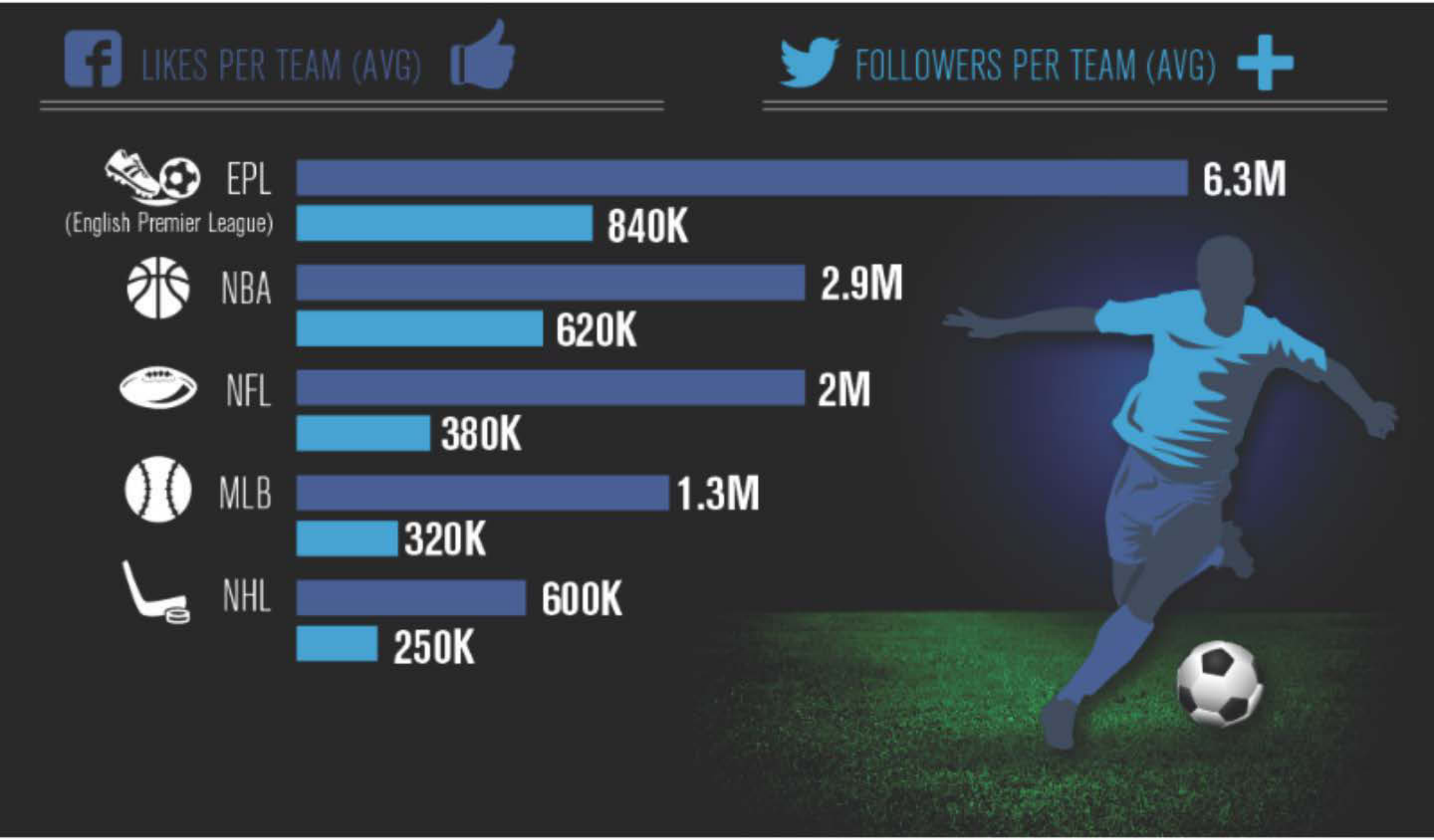
Despite the hysteria, the number of refugees in the UK has actually fallen by 76,439 since 2011.



Vodafone Institute
for Society
and Communications

Sample Description

* Due to rounding, numbers presented in this document may not add up precisely to 100%


LIKES PER TEAM (AVG)
FOLLOWERS PER TEAM (AVG)
EPL
(English Premier League)
6.3M
840K
NBA
2.9M
620K
NFL
2M
380K
MLB
1.3M
320K
NHL
600K
250K

Coming up next

- Grammar of graphics
- Introduction to `ggplot2`
- Lab: Make a Scatterplot