MassMutual DSDP 2019:

INTRODUCTION TO DATA VISUALIZATION

June 24, 2019

R. Jordan Crouser
Assistant Professor of Computer Science
Smith College

Introductions & background

Jordan

(he / him)

- **2017 on: Asst. Prof. in CS (Smith)**
- 2015 to 2017: Visiting Asst. Prof. in SDS (Smith)
- 2013 – 2015: Research Scientist (MITLL)
- 2010 – 2013: PhD in Visual Analytics (Tufts)
- 2008 – 2010: MSc in Educational Tech. (Tufts)
- 2004 – 2008: BA in CS and Math (Smith)

For more info, visit:

www.science.smith.edu/~jcrouser

Some housekeeping

- Workshop website:

  jcrouser.github.io/MassMutual-DataVis

- Rough structure:
  - 6 modules over 3 days (AM and PM)
  - Intro → Walkthrough/Lab → Explore → Share
- Assumptions:
  - R/Rstudio installed
  - Basic proficiency in R

Learning objectives

1. Understand why data vis works (and doesn't)

2. Explore some foundational methods / tools

3. Opportunities to get to know your new team

What I do: analytical tools for **messy data**

Big idea behind my research

Humans and machines have **complimentary strengths**

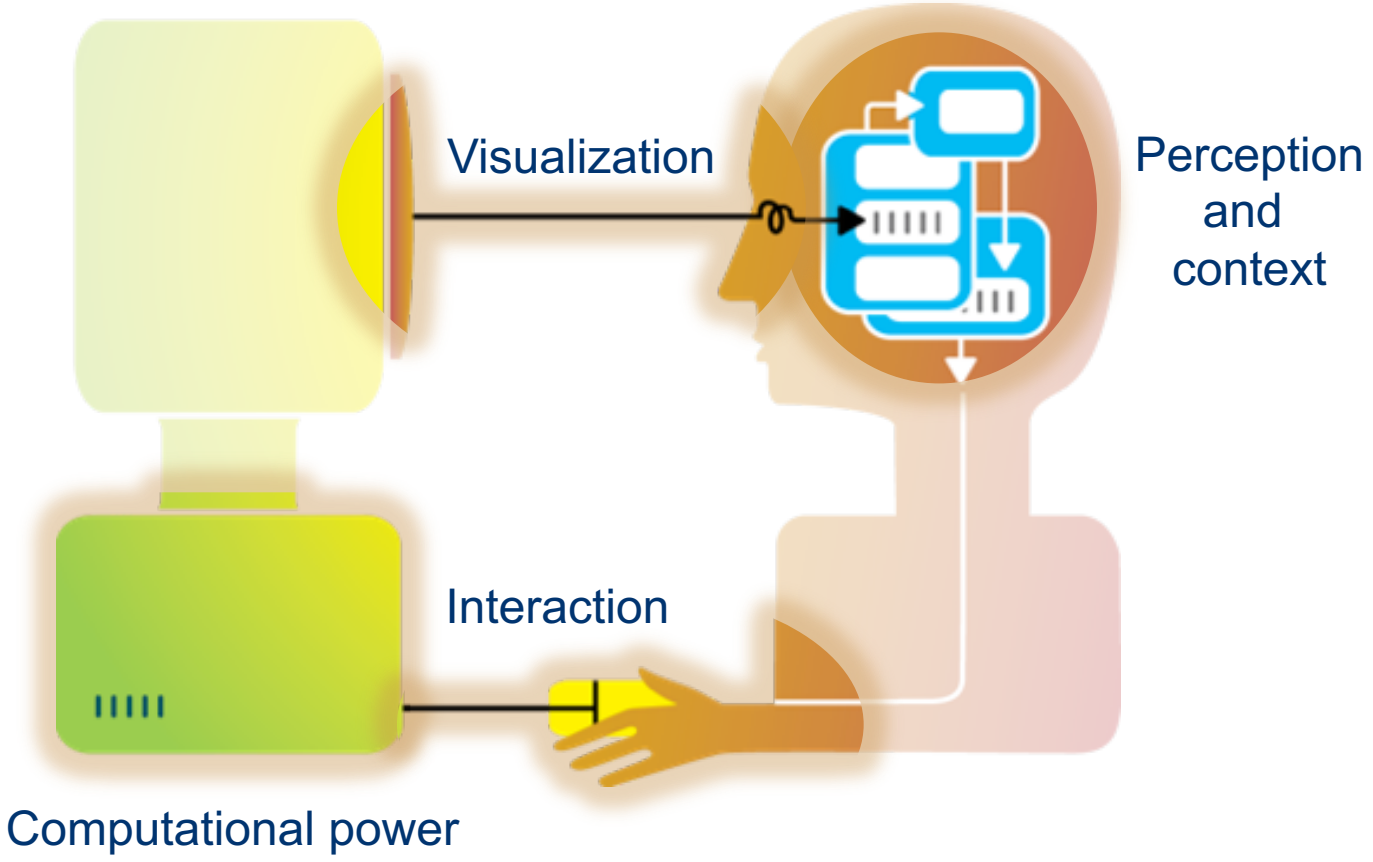


Image credit: Ali Ansari

Focus of this workshop

How do we build **visualizations** that help humans understand patterns in data?

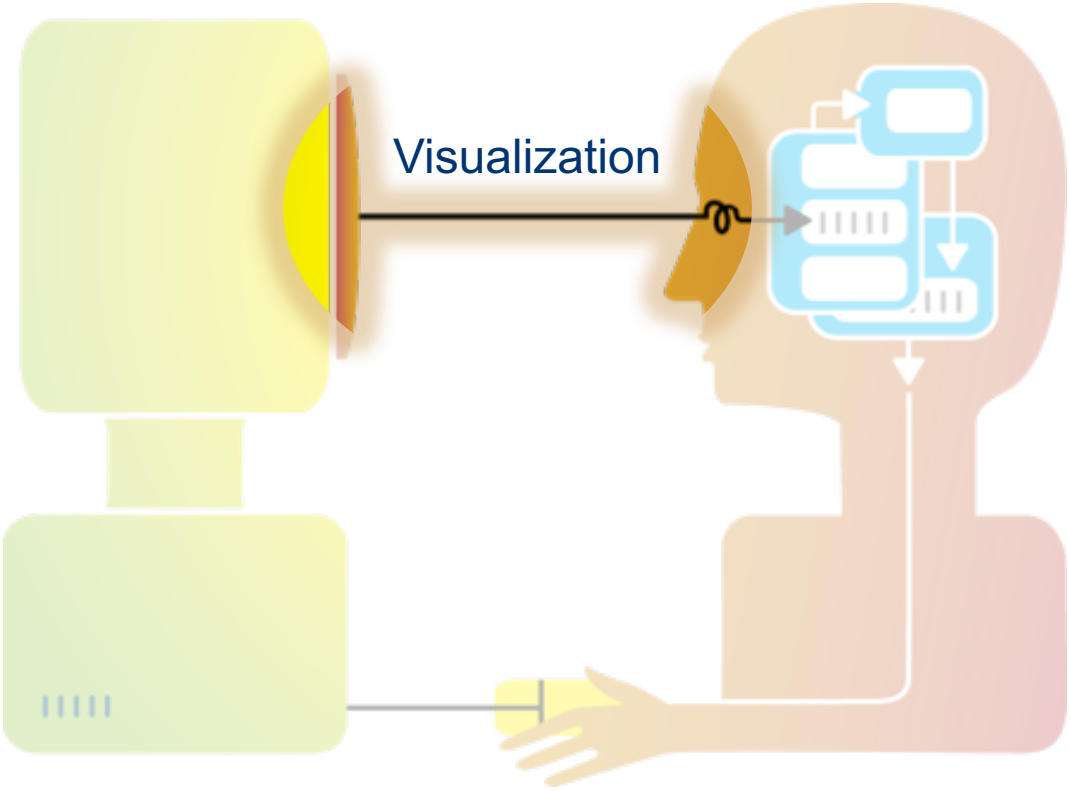


Image credit: Ali Ansari

3-minute biographies

About **you**:

- Your name and pronouns
- Your alma mater
- Your major / area of focus

3 questions:

1. What brought you to **DSDP**?
2. What's one **big thing** you hope to get out of it?
3. What's one thing about life after graduation that you find particularly **challenging / anxiety-provoking**?

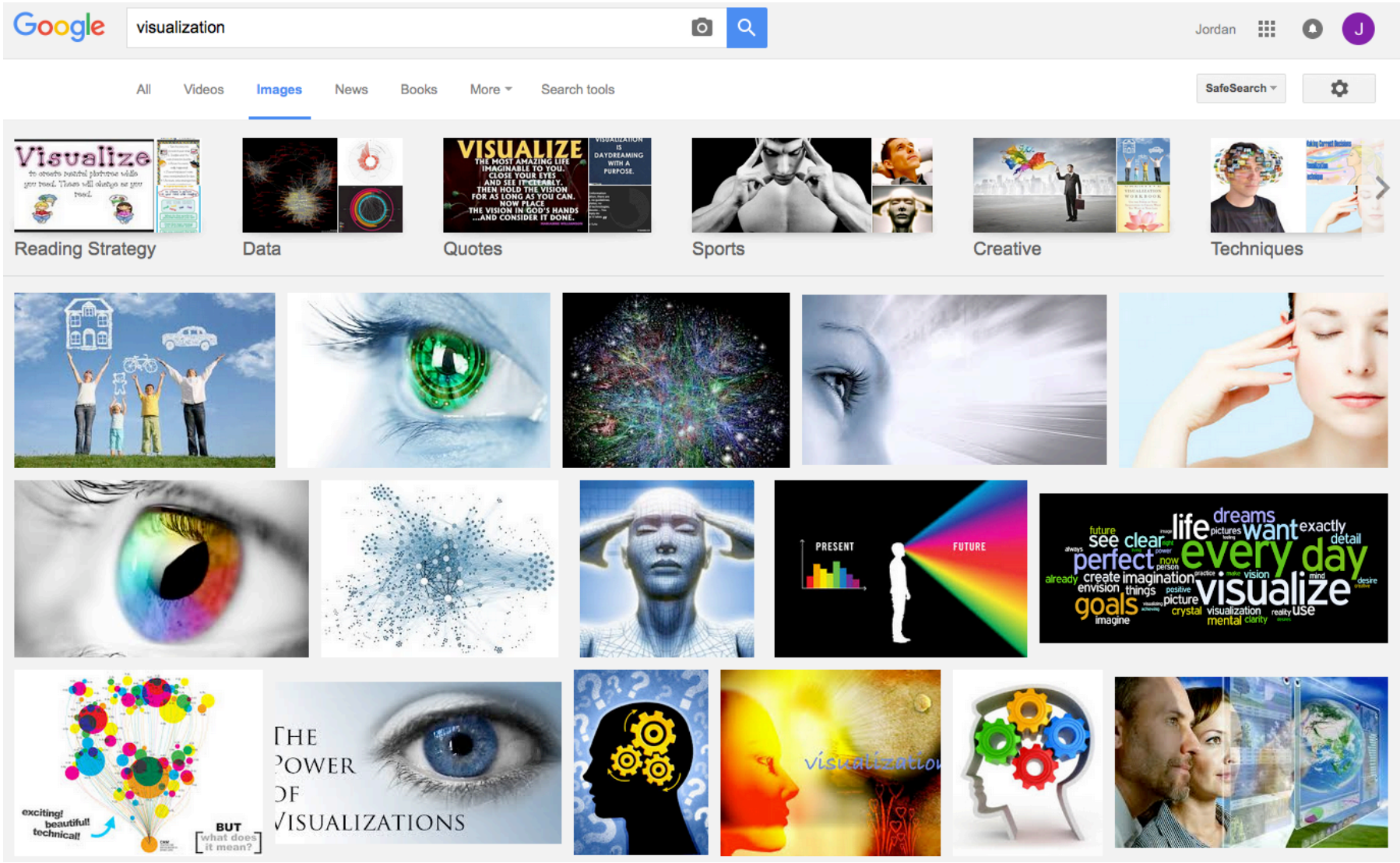
-or-

3. What's one thing about you that would probably **surprise** us?

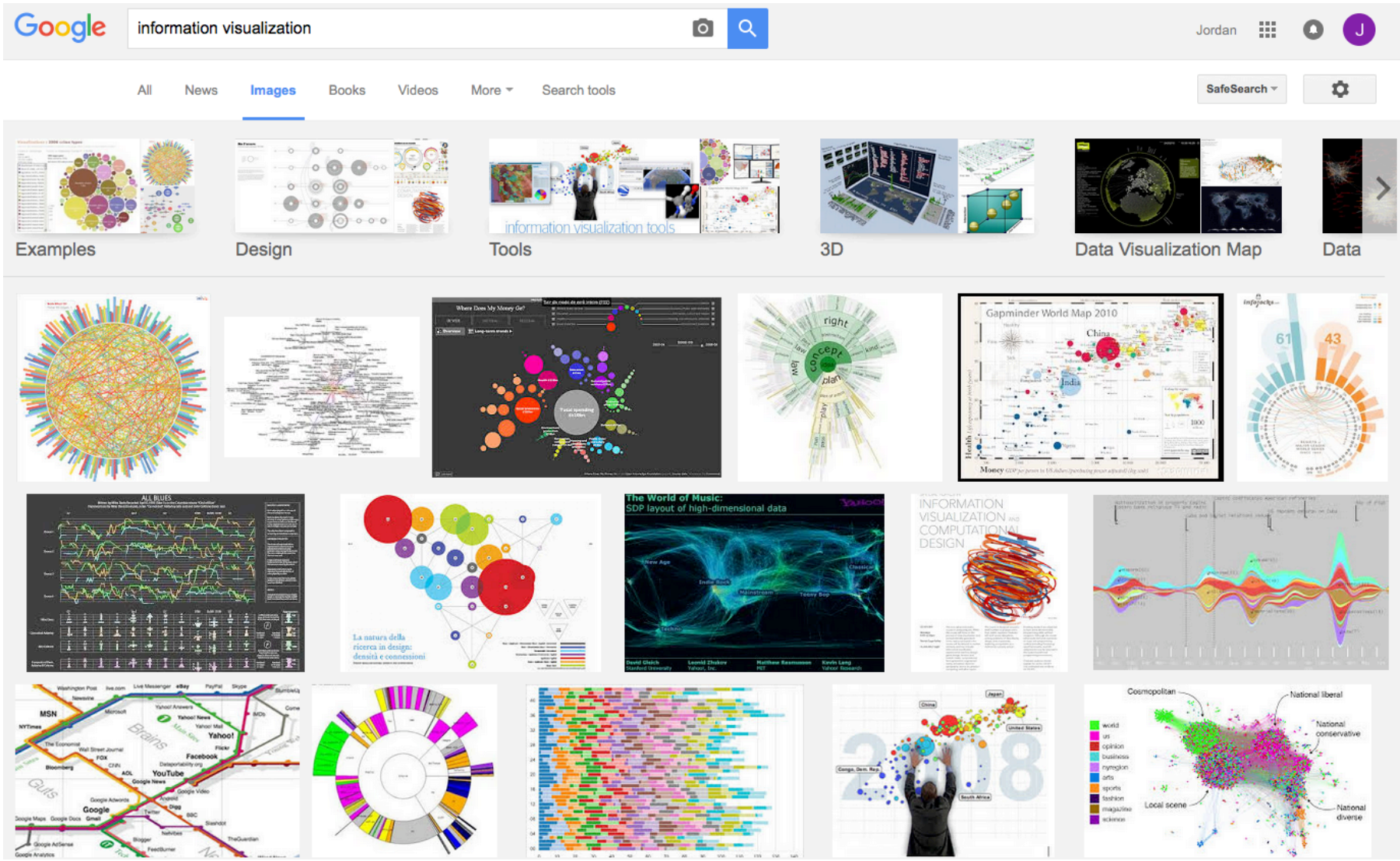
Outline

- ✓Introductions
- Visualization overview
  - Flashback to early experiences in data wrangling
  - Visualization (def.)
  - Data (def.)
  - Quick history lesson
- Graphical primitives
- Visual dimensions
- Pre-lunch activity: deconstructing data graphics
- After lunch: ggplot2 crash course

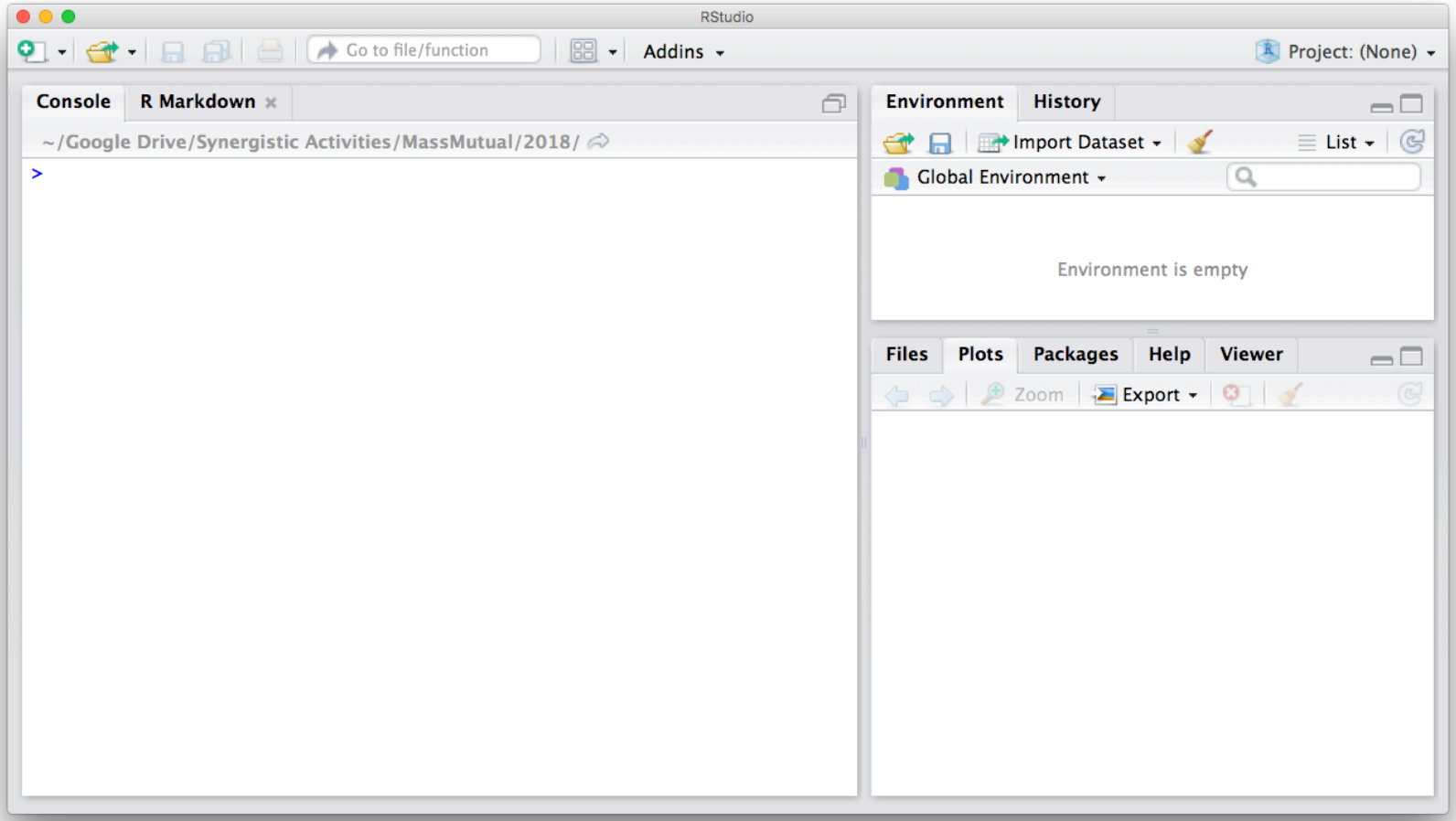
What is visualization?



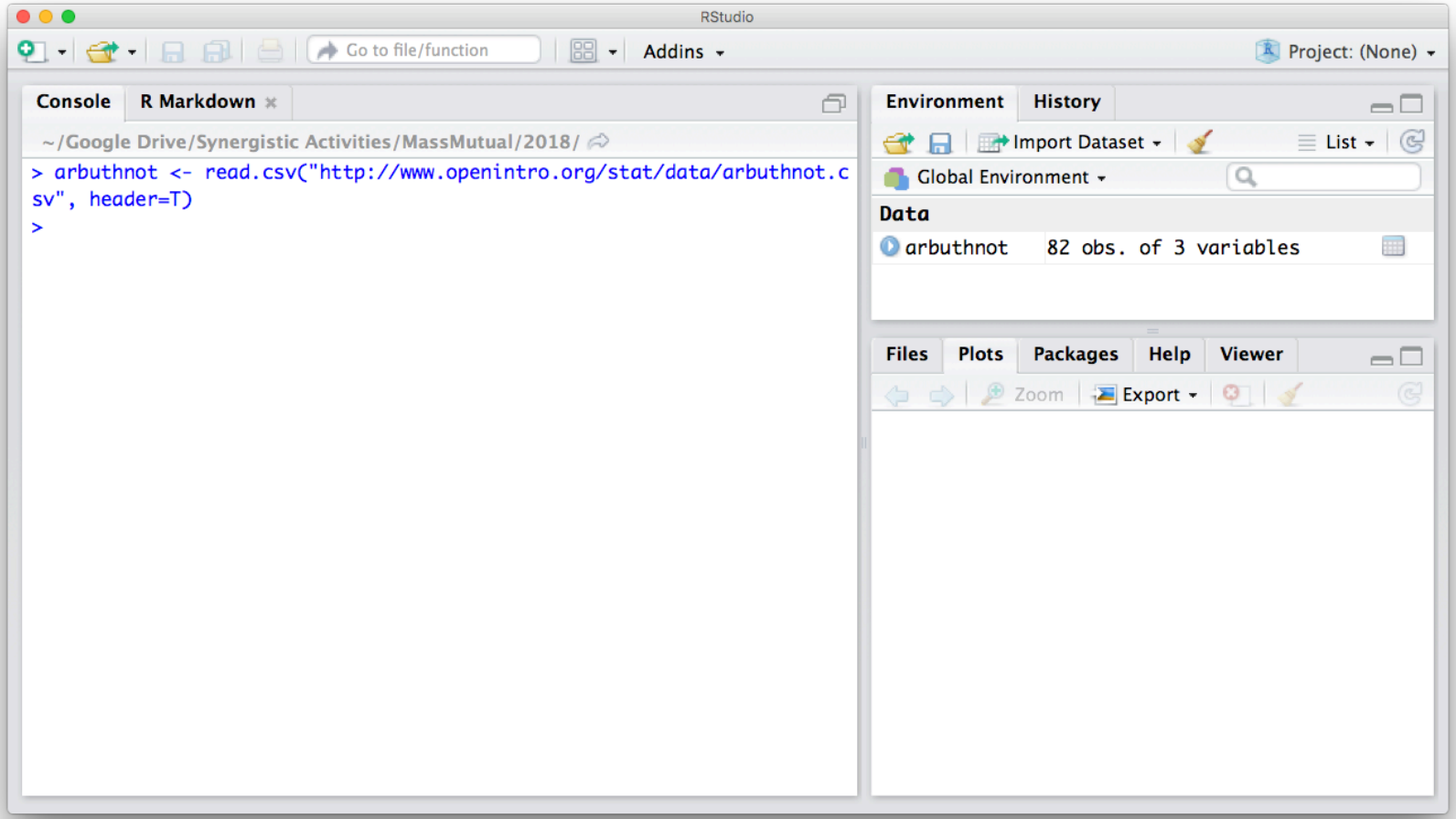
What is visualization?



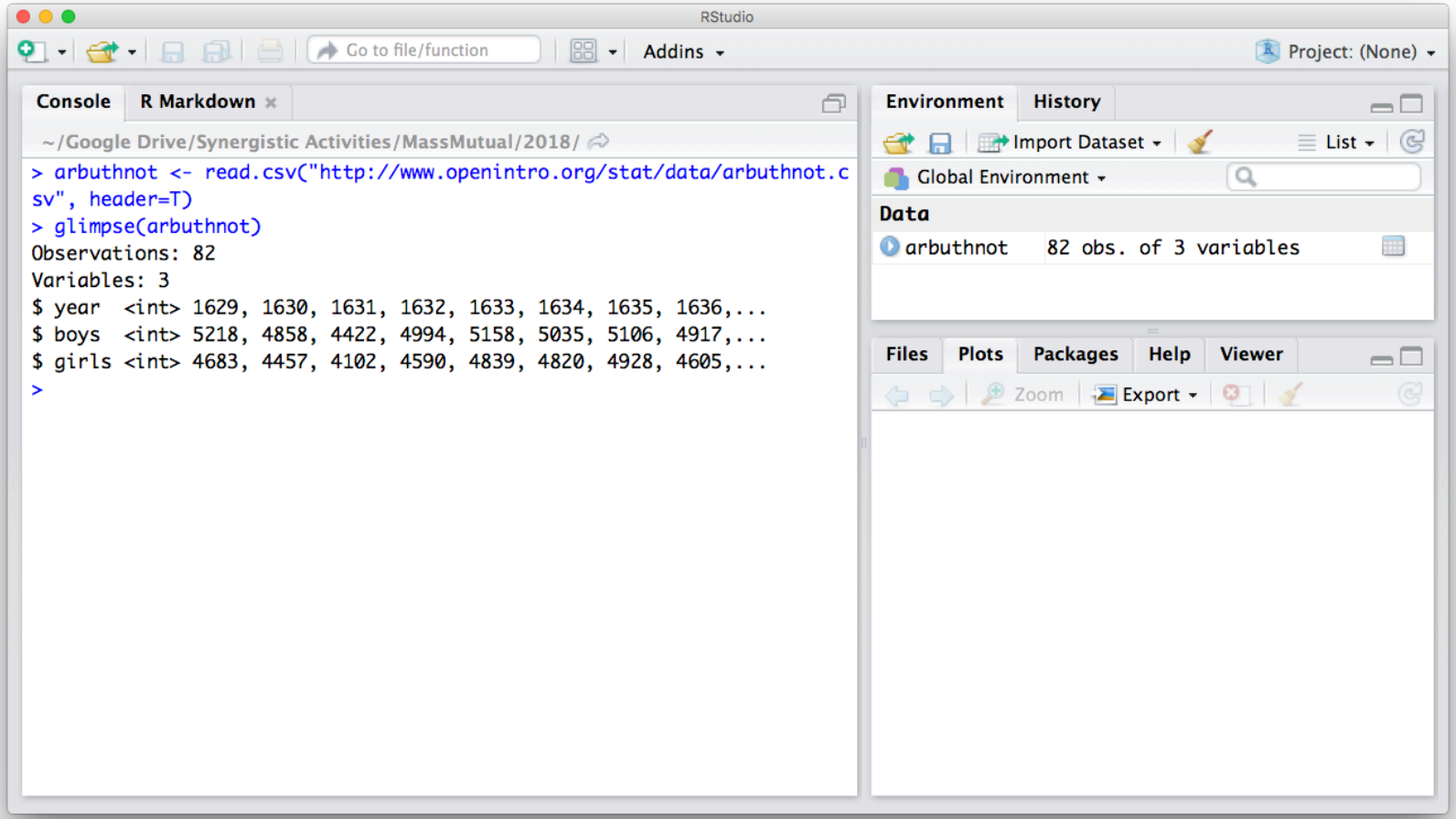
Flashback…


RStudio
Go to file/function
Addins
Project: (None)
Console
R Markdown
~/Google Drive/Synergistic Activities/MassMutual/2018/
>
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Zoom
Export

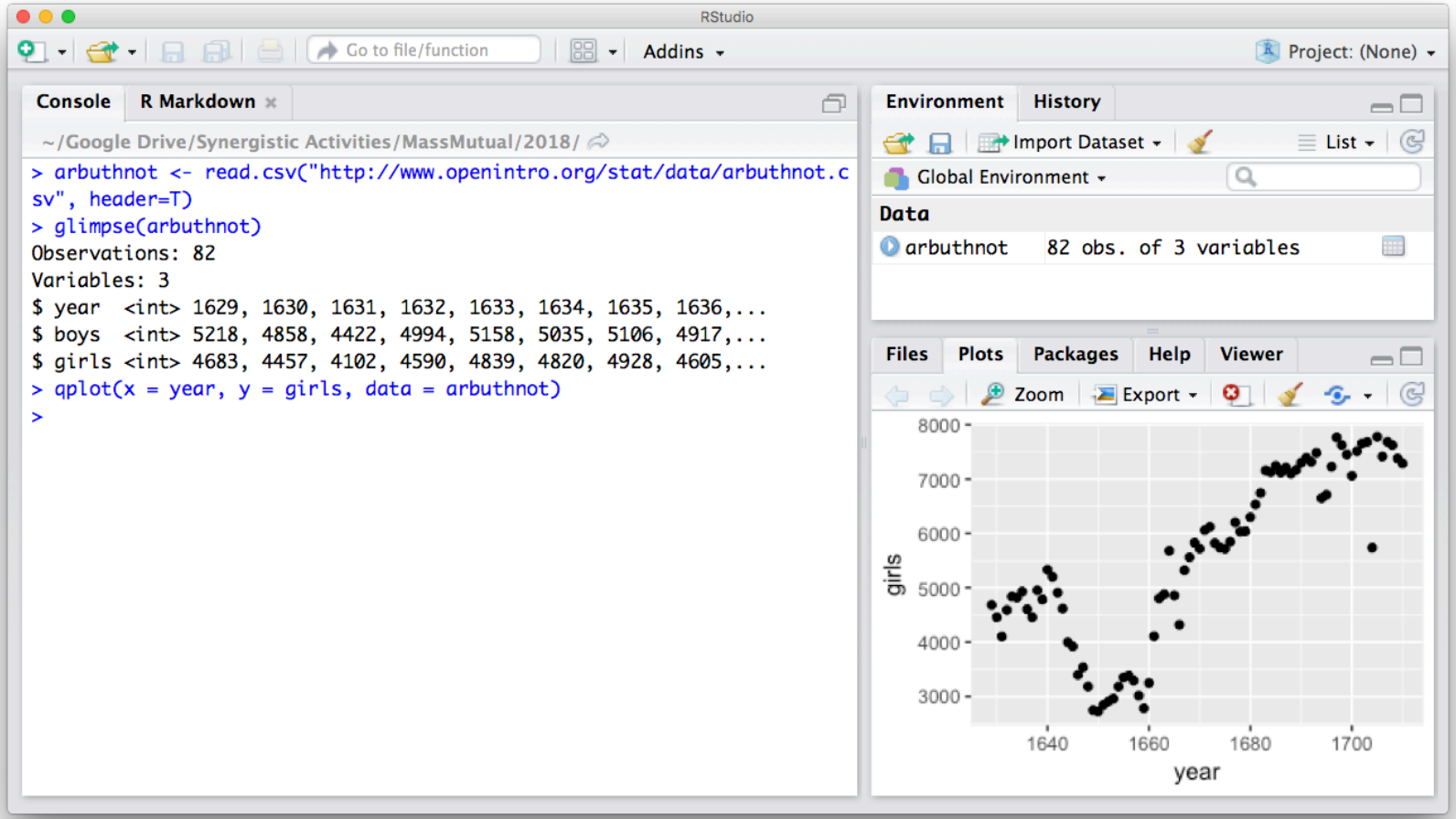
Flashback…


RStudio
Go to file/function
Addins
Project: (None)
Console
R Markdown
~/Google Drive/Synergistic Activities/MassMutual/2018/
> arbuthnot <- read.csv("http://www.openintro.org/stat/data/arbuthnot.csv", header=T)
>
Environment
History
Import Dataset
List
Global Environment
Data
arbuthnot
82 obs. of 3 variables
Files
Plots
Packages
Help
Viewer
Zoom
Export

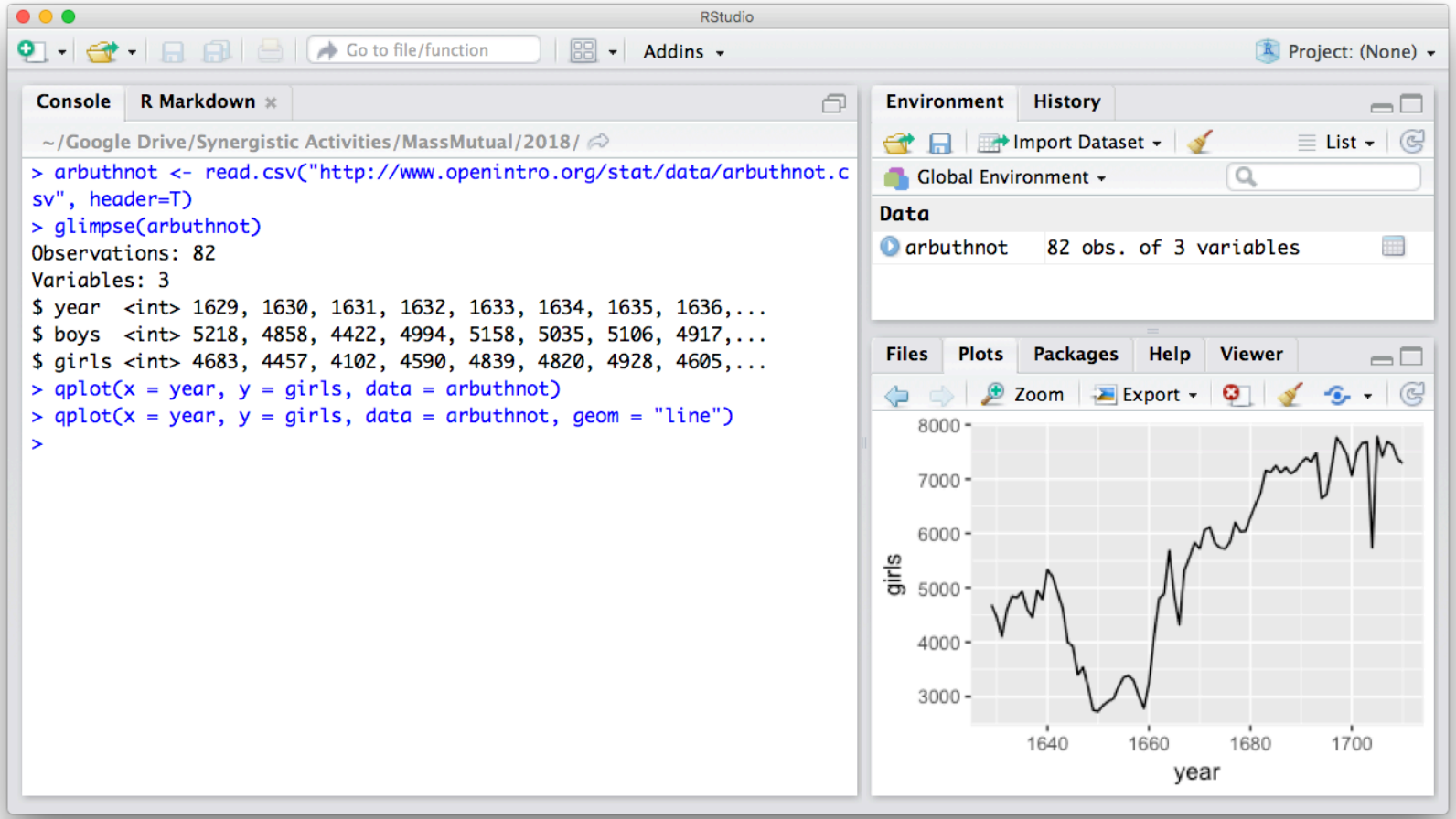
Flashback…


RStudio
Go to file/function
Addins
Project: (None)
Console
R Markdown
~/Google Drive/Synergistic Activities/MassMutual/2018/
> arbuthnot <- read.csv("http://www.openintro.org/stat/data/arbuthnot.csv", header=T)
> glimpse(arbuthnot)
Observations: 82
Variables: 3
$ year <int> 1629, 1630, 1631, 1632, 1633, 1634, 1635, 1636,...
$ boys <int> 5218, 4858, 4422, 4994, 5158, 5035, 5106, 4917,...
$ girls <int> 4683, 4457, 4102, 4590, 4839, 4820, 4928, 4605,...
>
Environment
History
Import Dataset
List
Global Environment
Data
arbuthnot
82 obs. of 3 variables
Files
Plots
Packages
Help
Viewer
Zoom
Export

Flashback…


RStudio
Go to file/function
Addins
Project: (None)
Console
R Markdown
~/Google Drive/Synergistic Activities/MassMutual/2018/
> arbuthnot <- read.csv("http://www.openintro.org/stat/data/arbuthnot.csv", header=T)
> glimpse(arbuthnot)
Observations: 82
Variables: 3
$ year <int> 1629, 1630, 1631, 1632, 1633, 1634, 1635, 1636,...
$ boys <int> 5218, 4858, 4422, 4994, 5158, 5035, 5106, 4917,...
$ girls <int> 4683, 4457, 4102, 4590, 4839, 4820, 4928, 4605,...
> qplot(x = year, y = girls, data = arbuthnot)
>
Environment
History
Import Dataset
List
Global Environment
Data
arbuthnot
82 obs. of 3 variables
Files
Plots
Packages
Help
Viewer
Zoom
Export
8000
7000
6000
5000
4000
3000
girls
1640
1660
1680
1700
year

Flashback…


RStudio
Go to file/function
Addins
Project: (None)
Console
R Markdown
~/Google Drive/Synergistic Activities/MassMutual/2018/
> arbuthnot <- read.csv("http://www.openintro.org/stat/data/arbuthnot.csv", header=T)
> glimpse(arbuthnot)
Observations: 82
Variables: 3
$ year <int> 1629, 1630, 1631, 1632, 1633, 1634, 1635, 1636,...
$ boys <int> 5218, 4858, 4422, 4994, 5158, 5035, 5106, 4917,...
$ girls <int> 4683, 4457, 4102, 4590, 4839, 4820, 4928, 4605,...
> qplot(x = year, y = girls, data = arbuthnot)
> qplot(x = year, y = girls, data = arbuthnot, geom = "line")
>
Environment
History
Import Dataset
List
Global Environment
Data
arbuthnot
82 obs. of 3 variables
Files
Plots
Packages
Help
Viewer
Zoom
Export
8000
7000
6000
5000
4000
3000
girls
1640
1660
1680
1700
year

Question

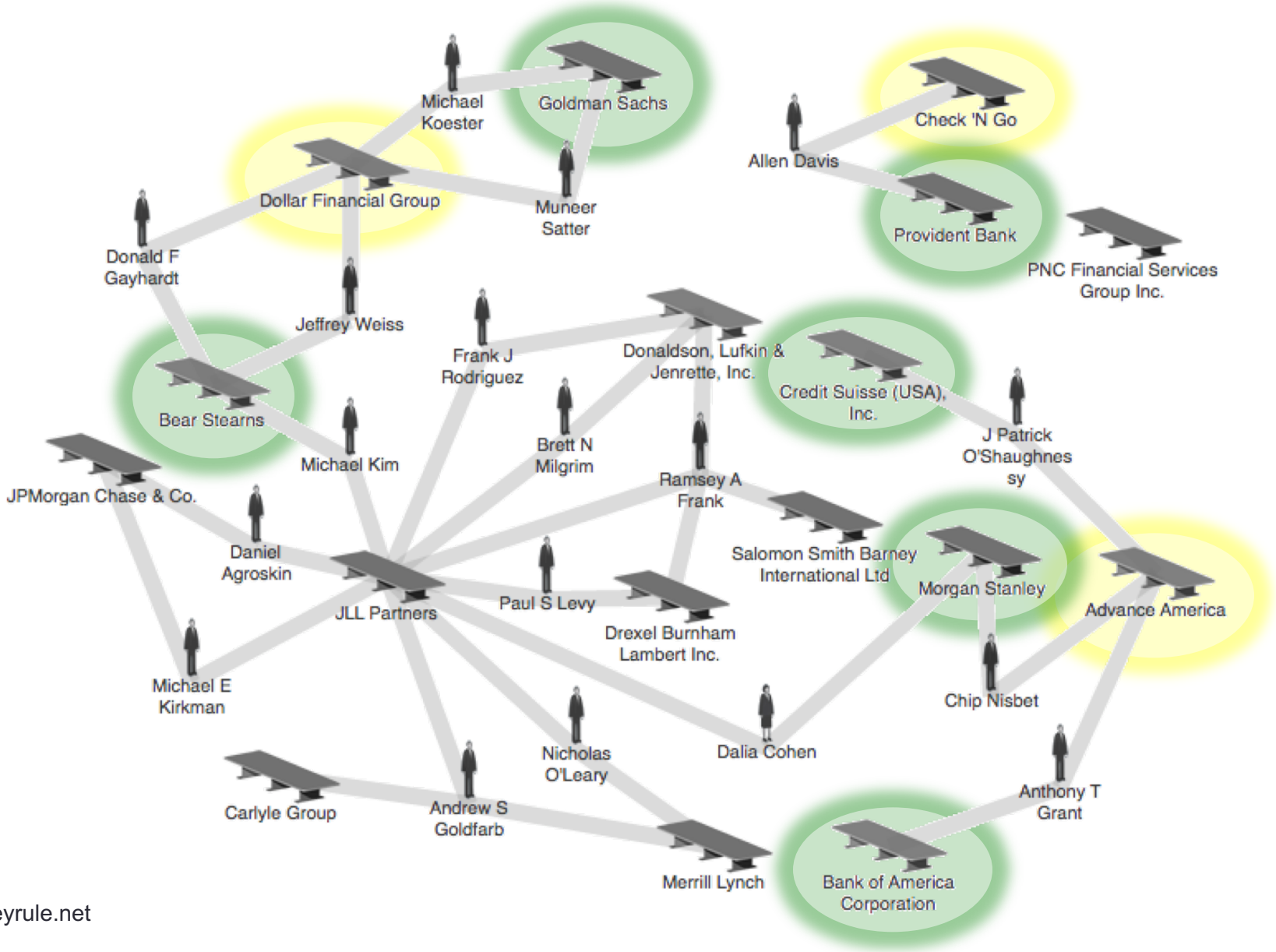
What makes these

"visualizations" **useful**?

Do they help you spot trends?

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

Do they help you explore?



Credit: theyrule.net

Do they 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**

Data (def.)

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

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

Data (def.)

and a corresponding set of *observations* (a.k.a. *records)* over these variables. For example:

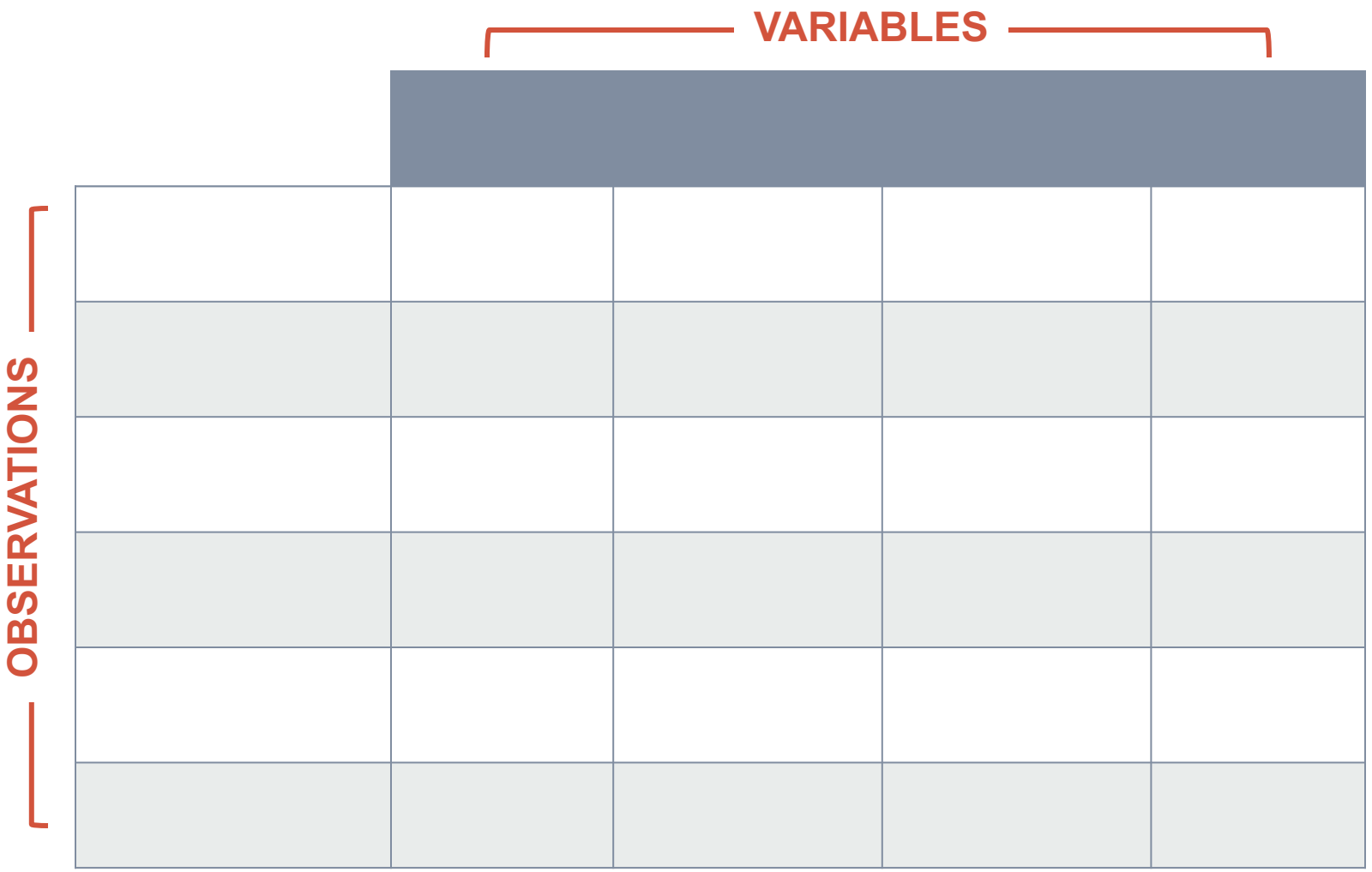
tuition = $46,288, *enrollment* = 2,563, *private, etc.*

Data (def.)

and a corresponding set of *observations* (a.k.a. *records)* over these variables. For example:

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

One way to think about this:



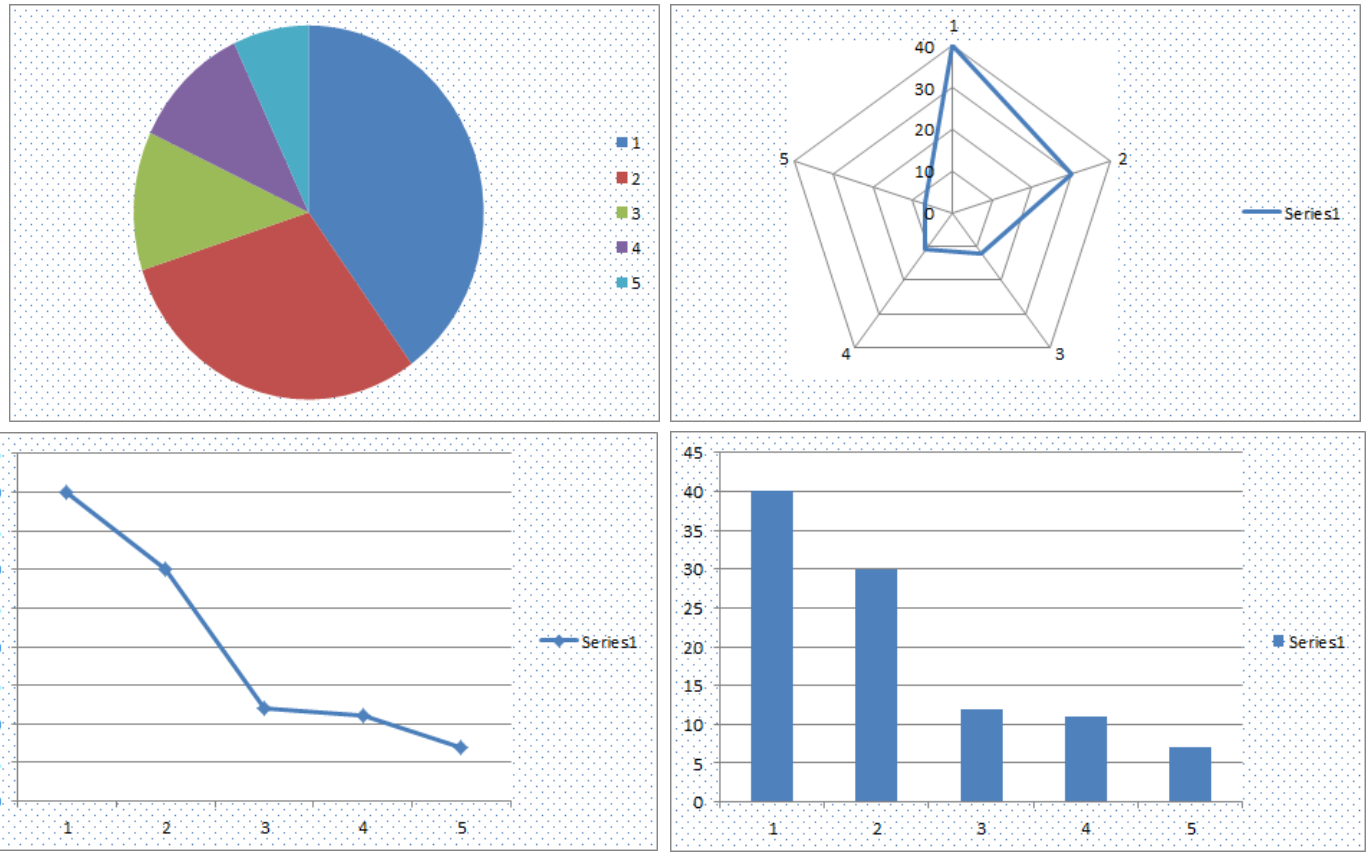
Why is this important?

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

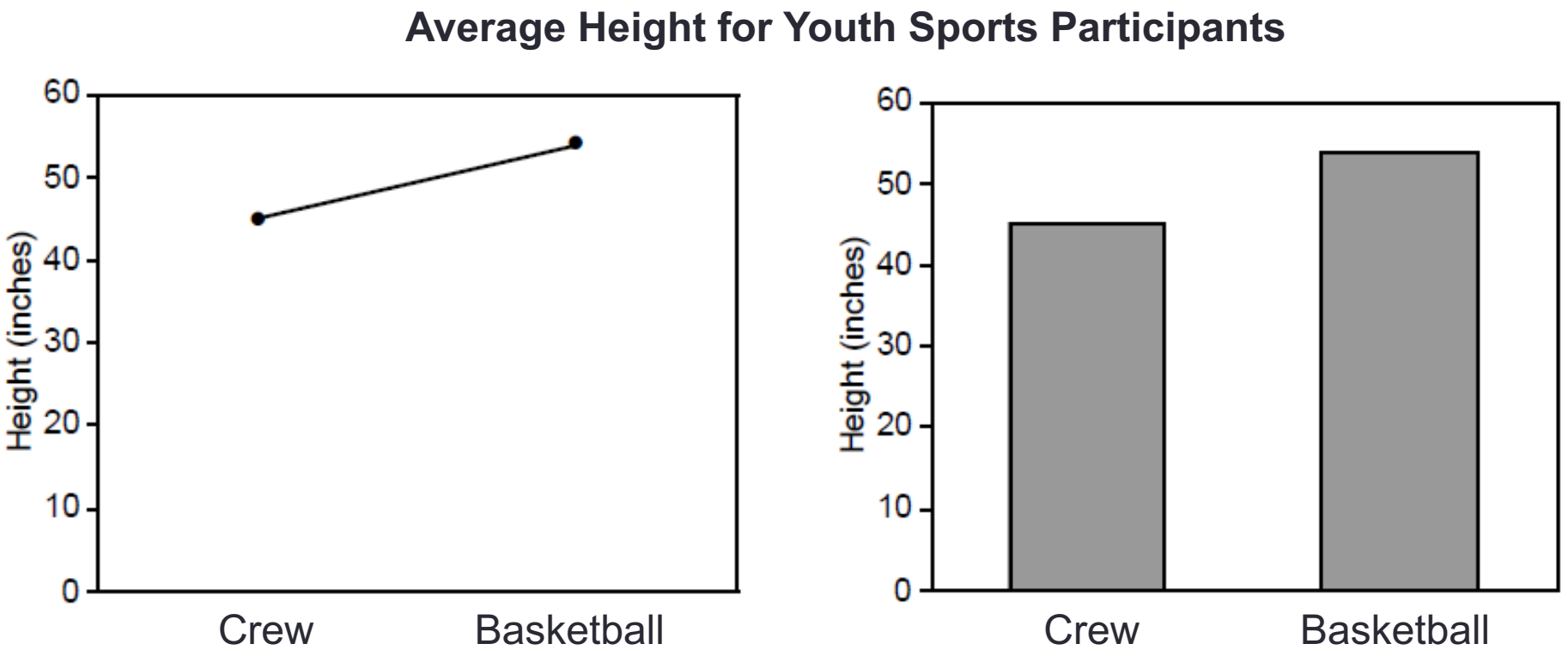
Key question for this workshop

Which **data dimension** should be mapped

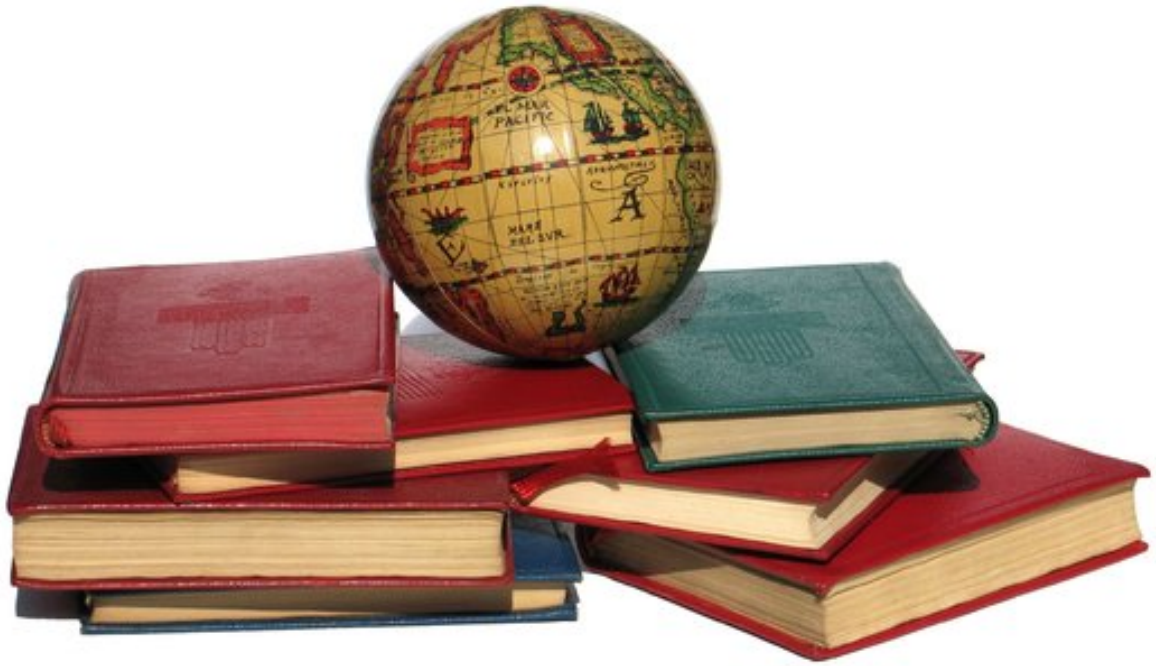
to which **visual dimension?**



Answer: it depends



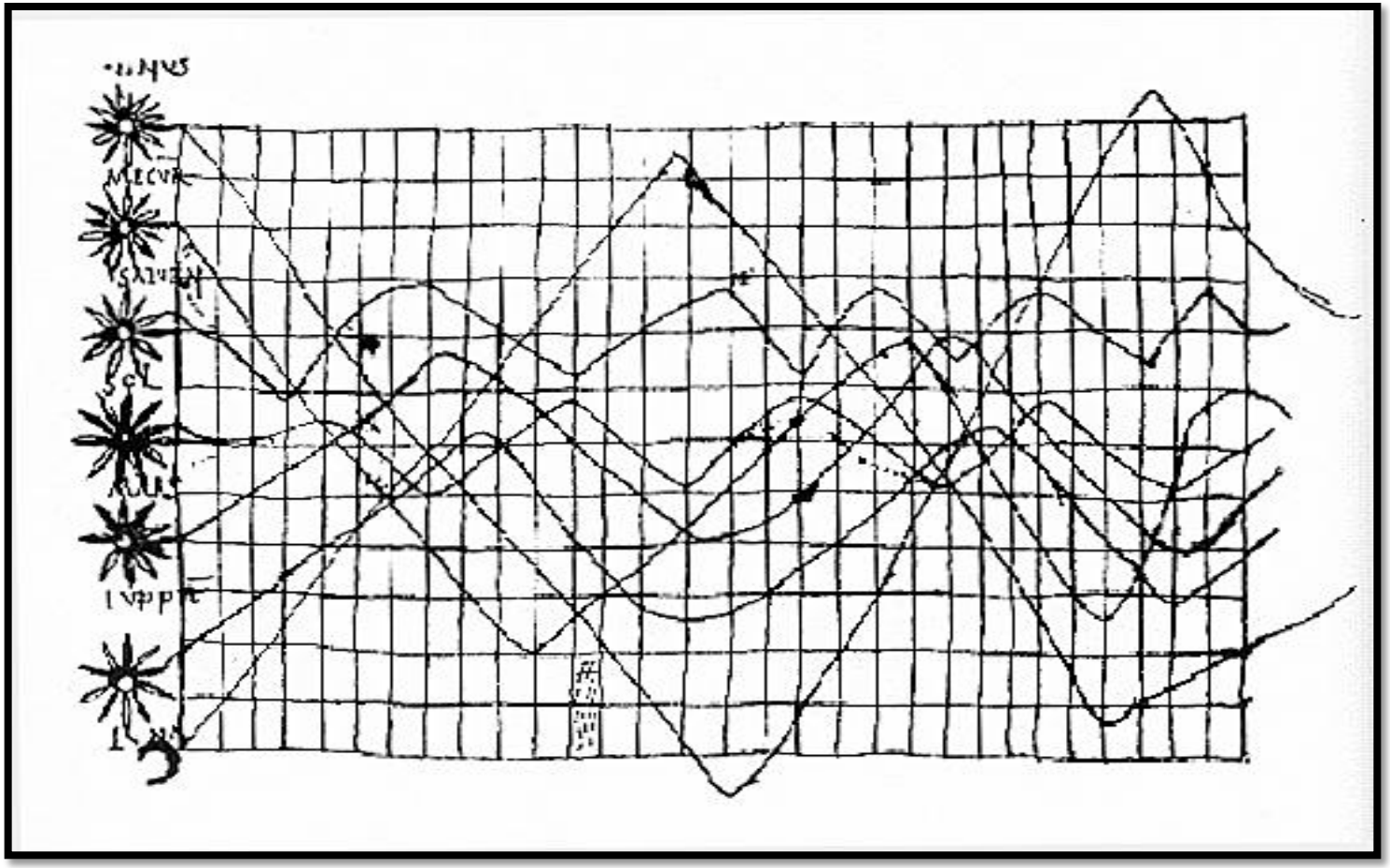
A quick history lesson…



(Incomplete) History of Visualization: 15,000BC

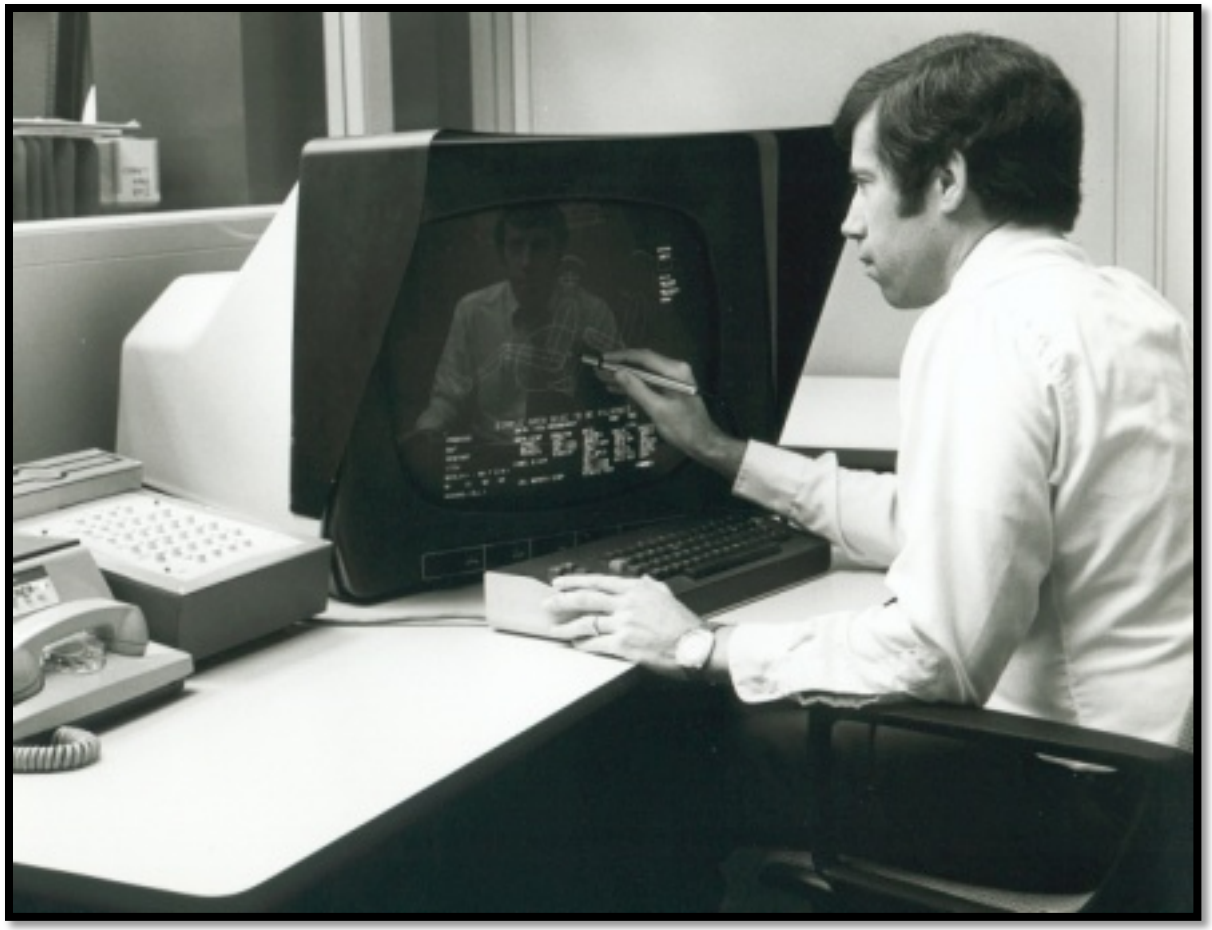
15,000 BC. Laxcaux, France

(Incomplete) History of Visualization: 900s



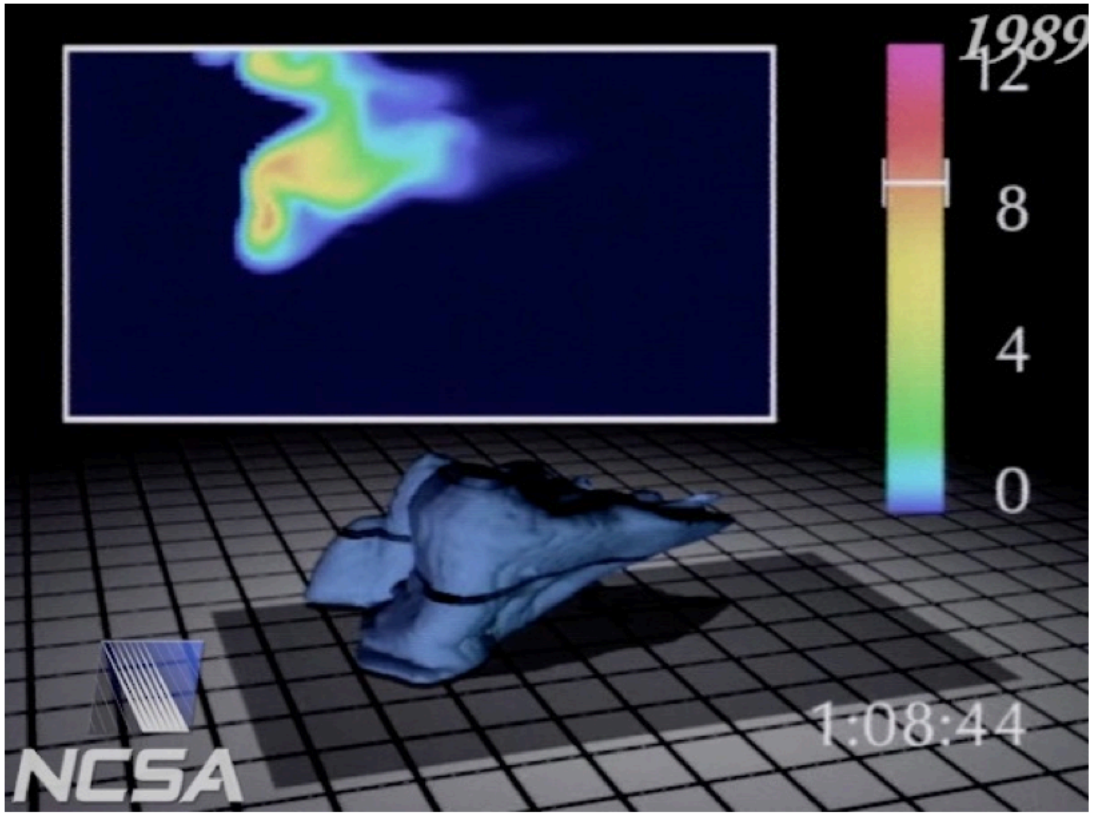
“De cursu per zodiacum”, illustrator unknown

(Incomplete) History of Visualization: 1970s



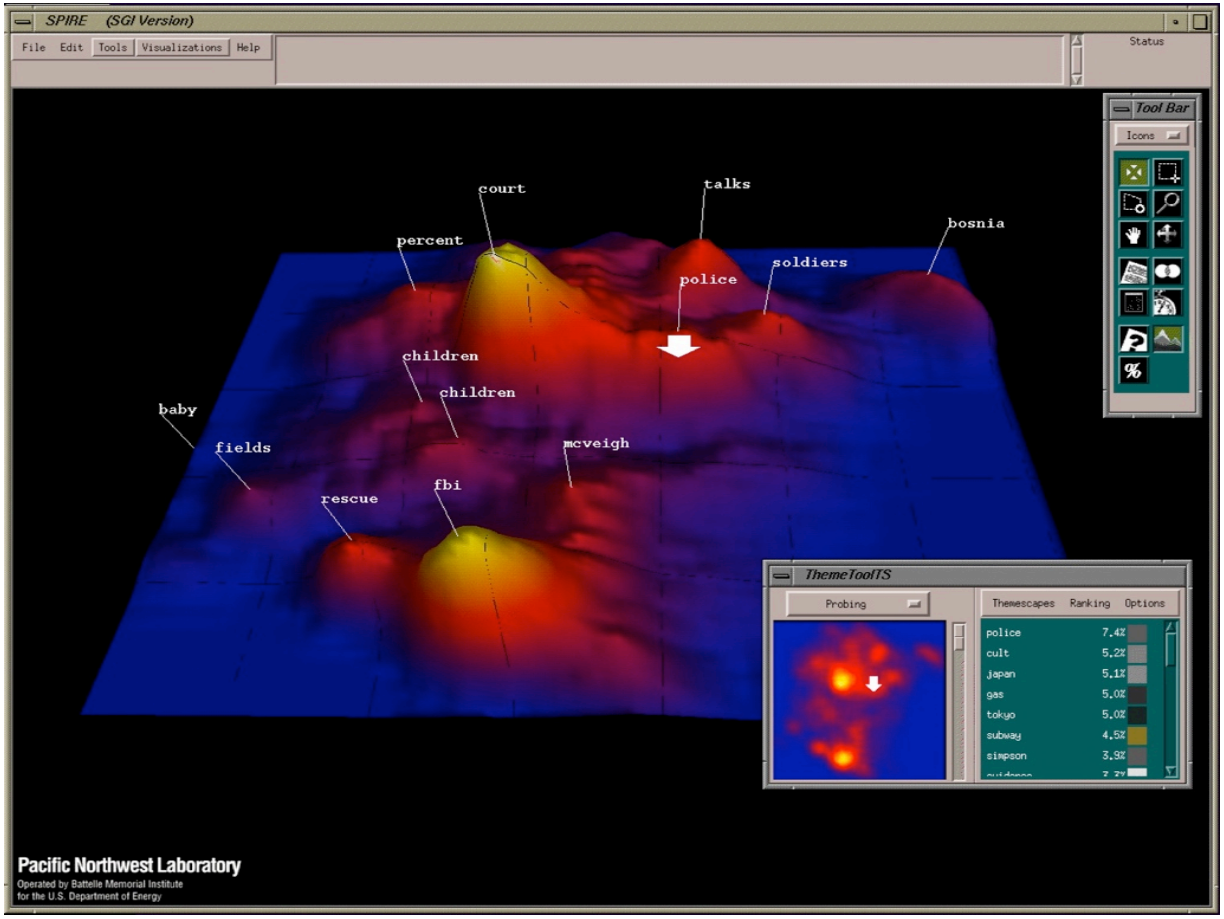
- CAD/CAM, building cars, planes, chips
- Starting to think about: 3D, animation, edu, medicine

(Incomplete) History of Visualization: 1980s



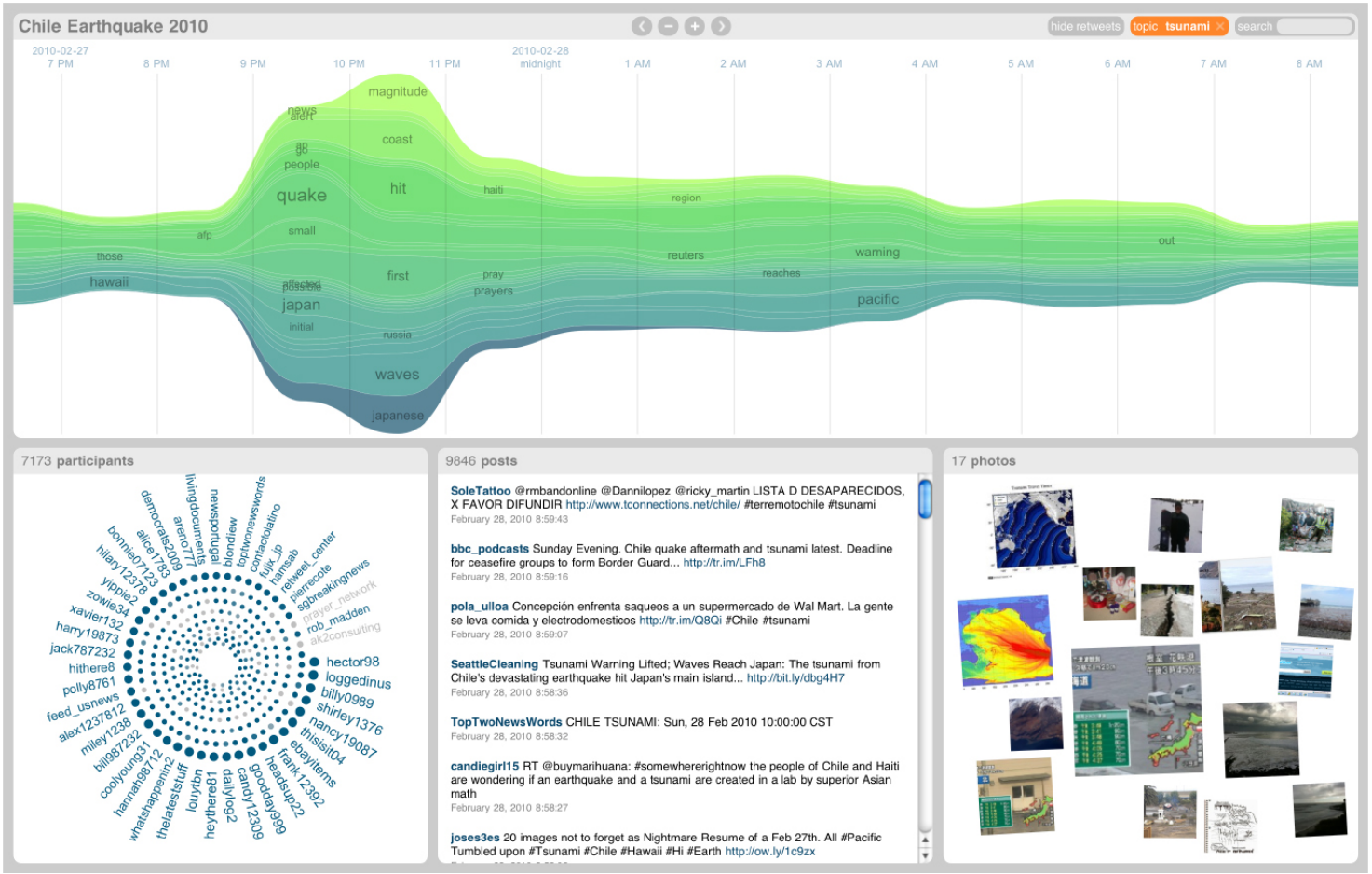
- Scientific visualization, physical phenomena
- Starting to think about: photorealism, entertainment

(Incomplete) History of Visualization: 1990s



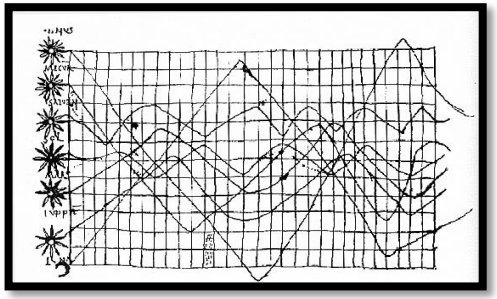
- Information visualization, storytelling
- Starting to think about: online spaces, interaction

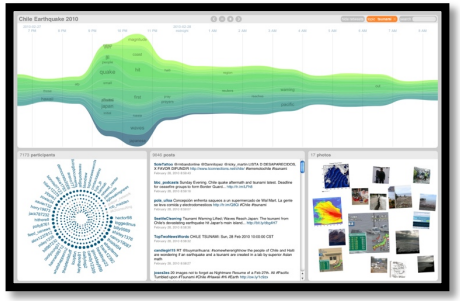
(Incomplete) History of Visualization: 2000s

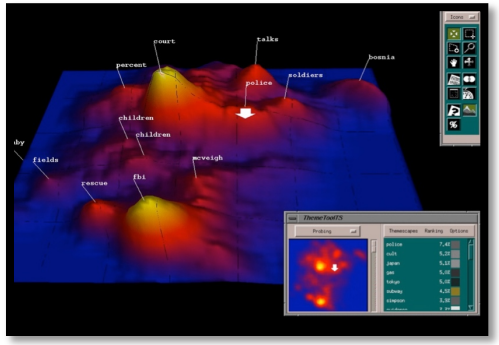


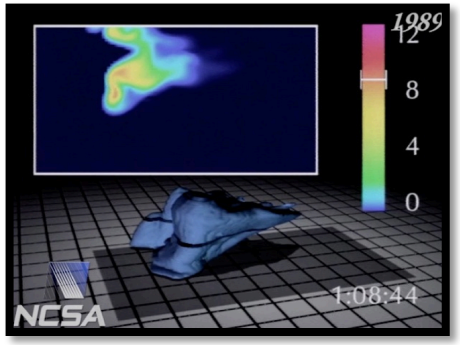
- Coordination across multiple views, interaction
- Starting to think about: sensemaking, provenance

Discussion: what are they all trying to do?

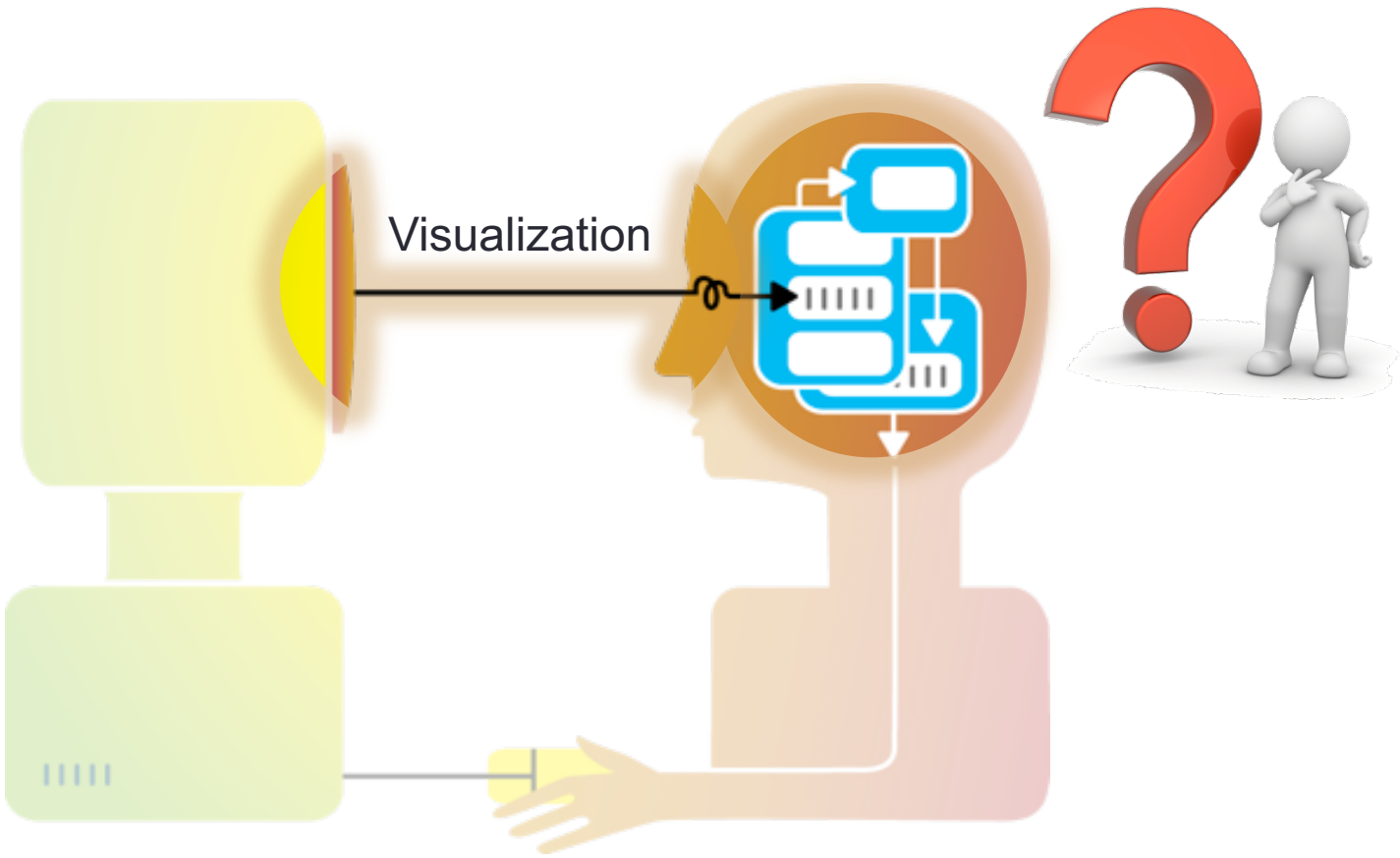




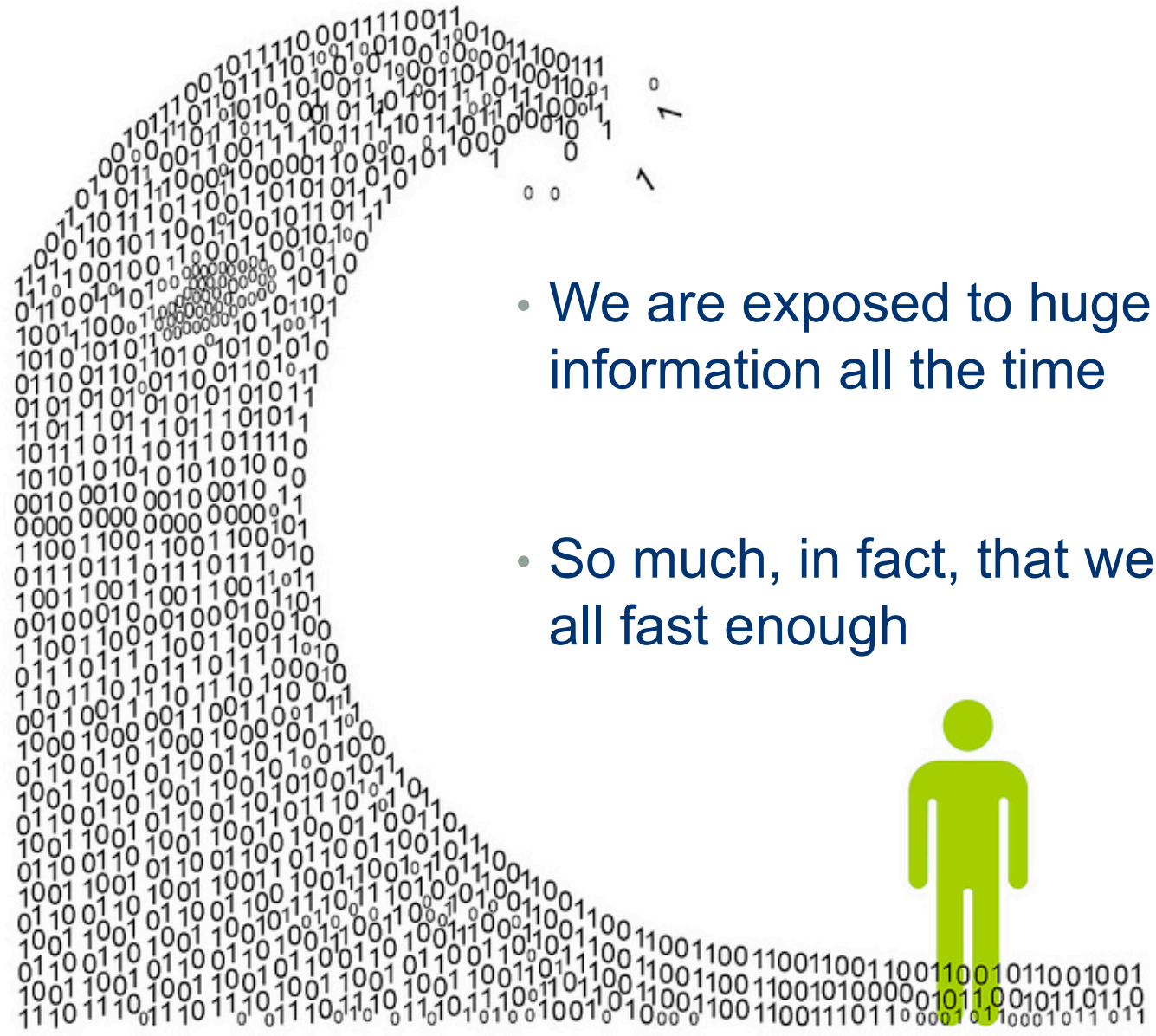




Visualization helps shape *mental models*



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

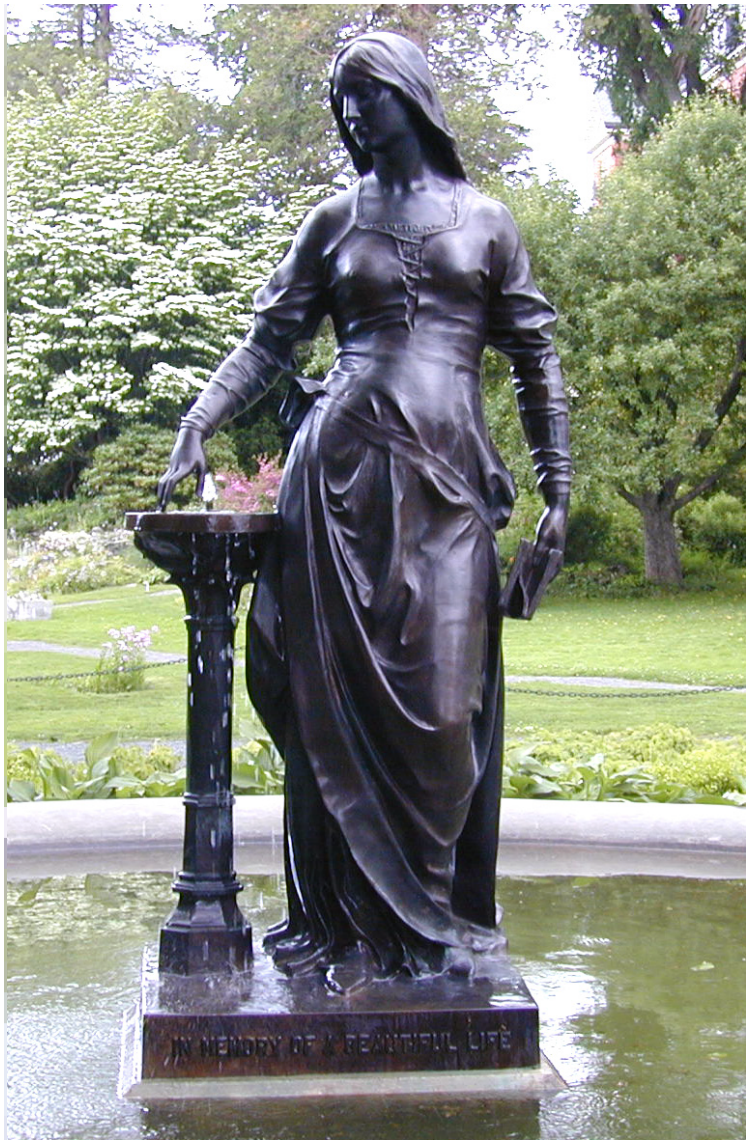
Mental Models: a Sketch

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?

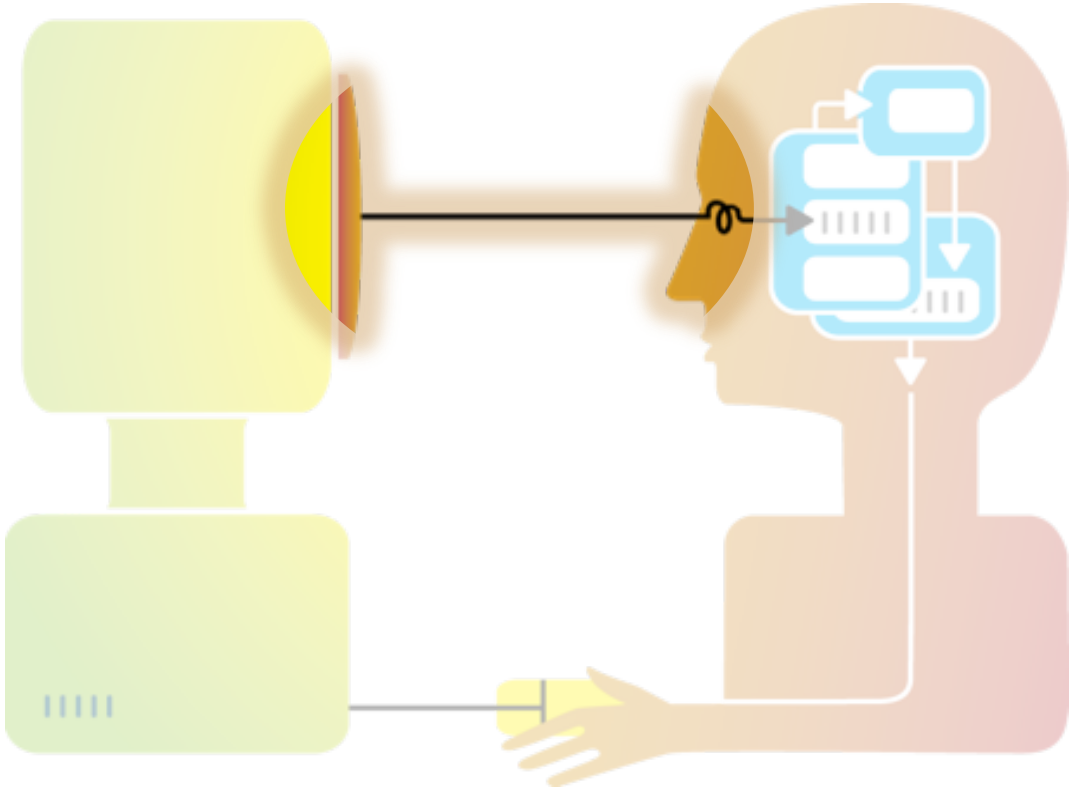
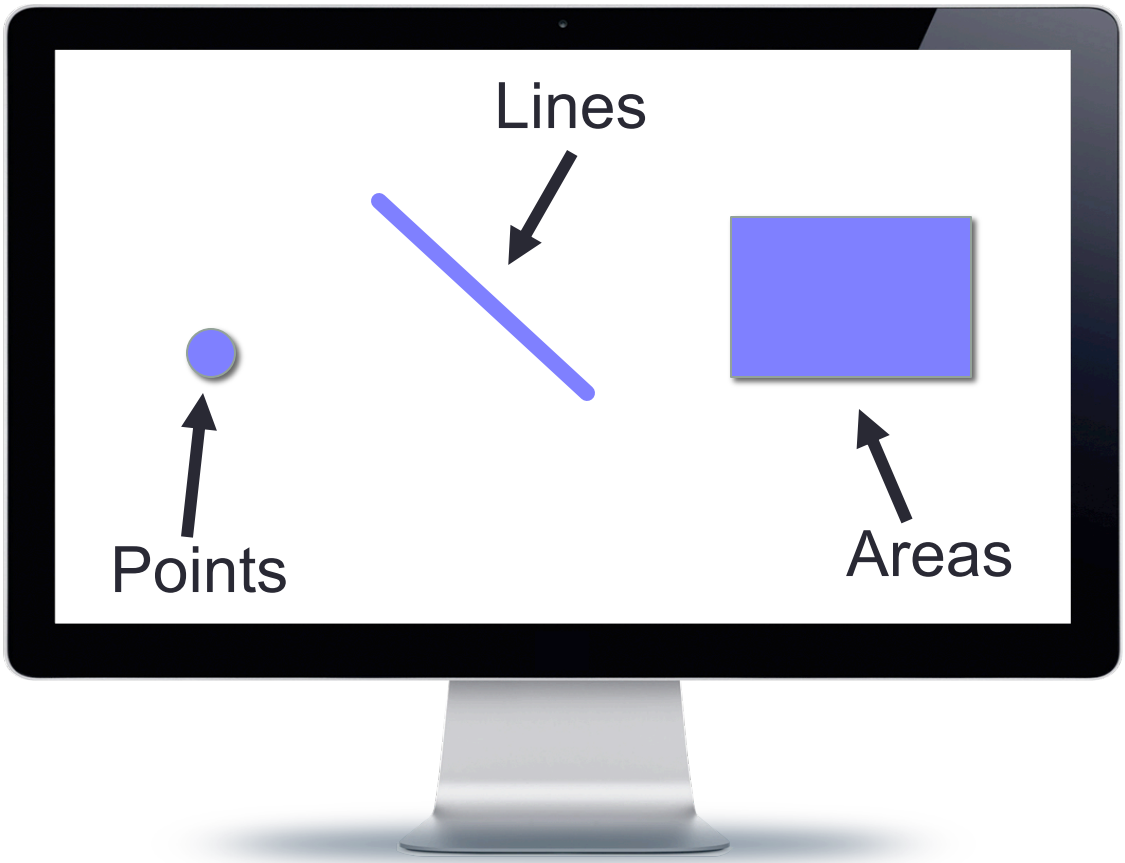


Image credit: Ali Ansari

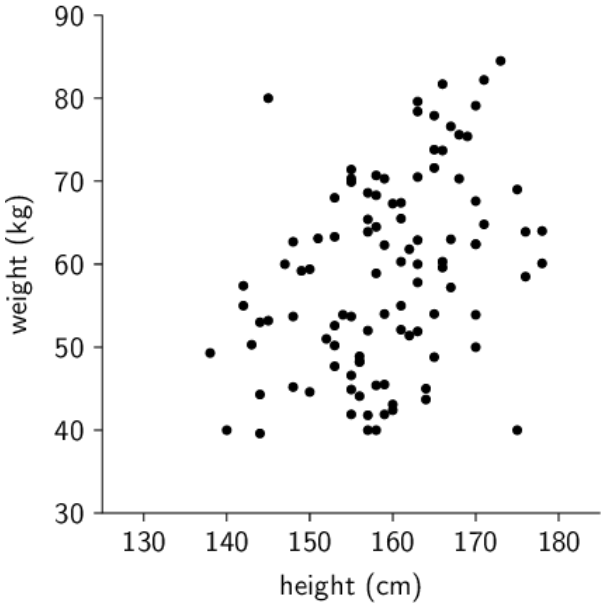
Graphical primitives

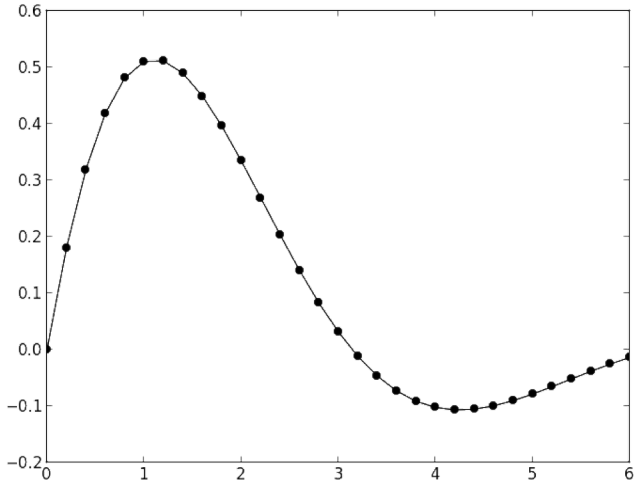
The images we draw are composed of marks: like ink



Visual dimension: position

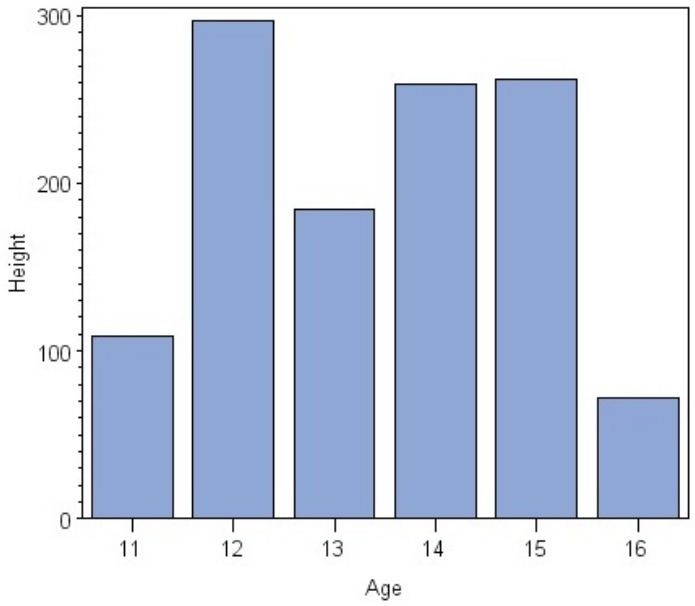
- Encode information using **where** the mark is drawn
- Some examples:

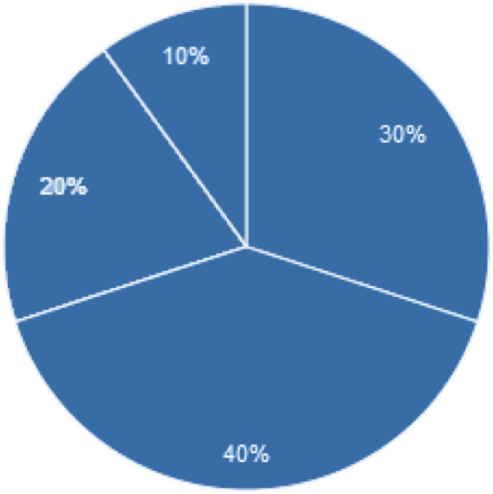




Visual dimension: size

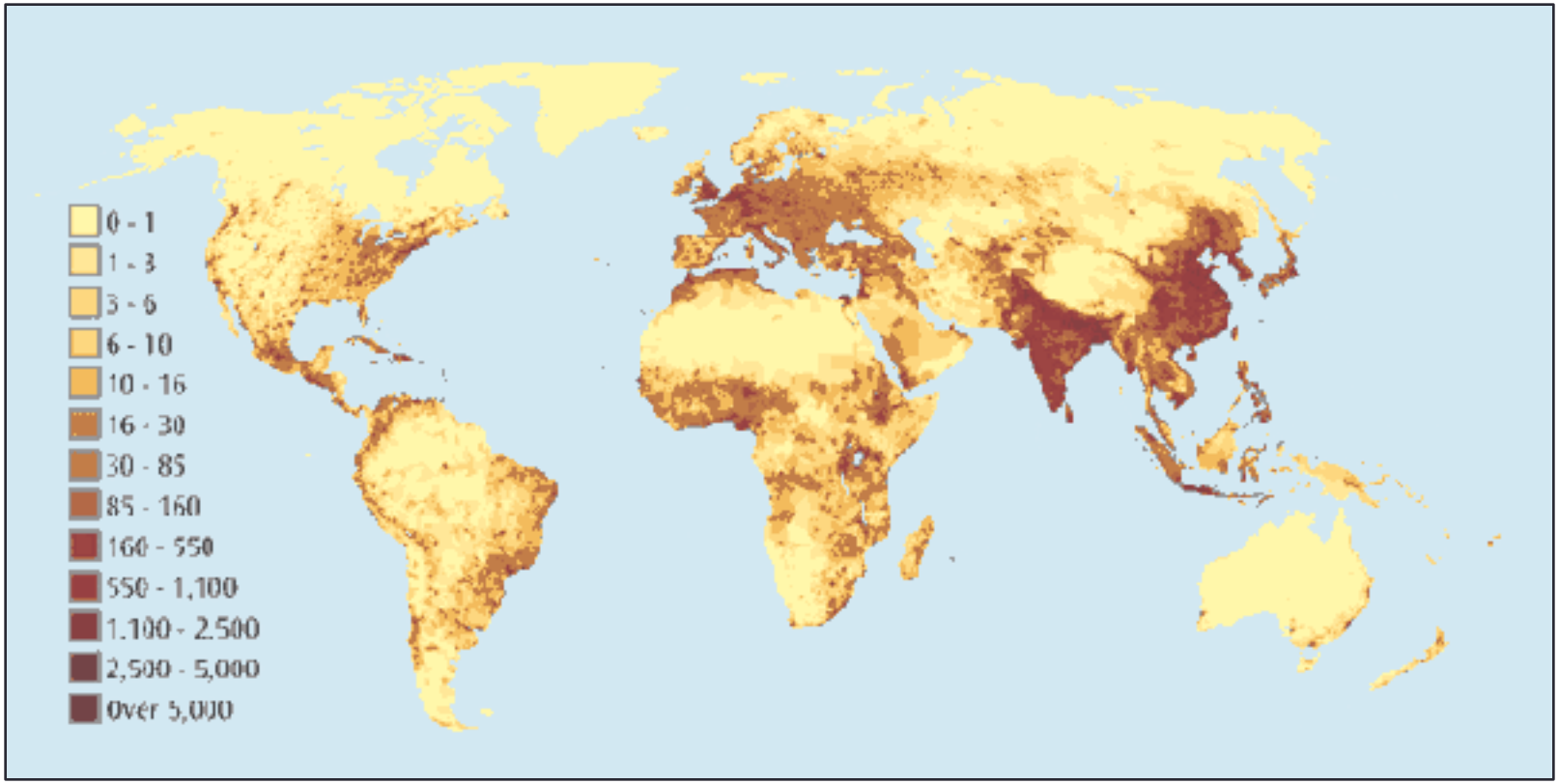
- Encode information using **how big** the mark is drawn
- Examples:





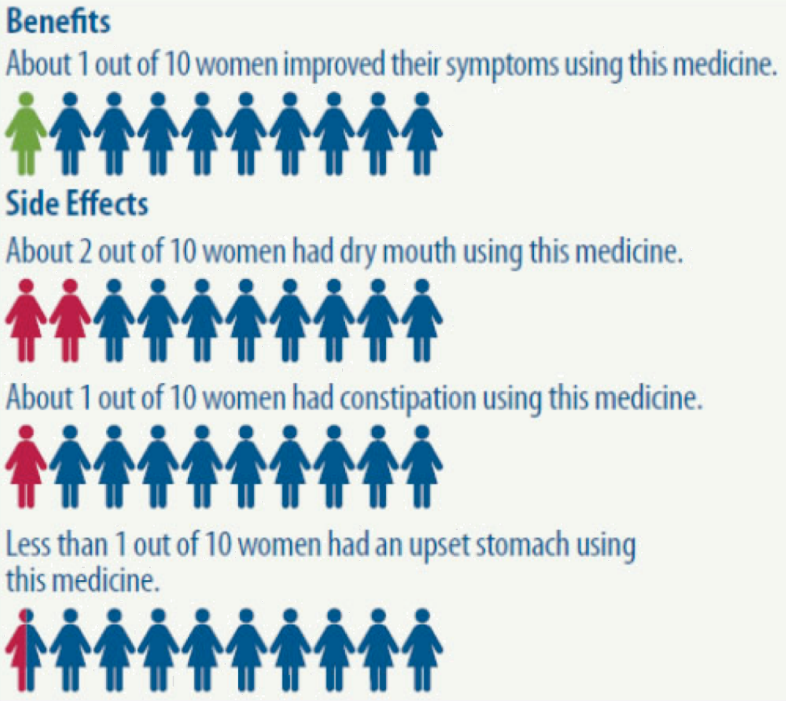
Visual dimension: value

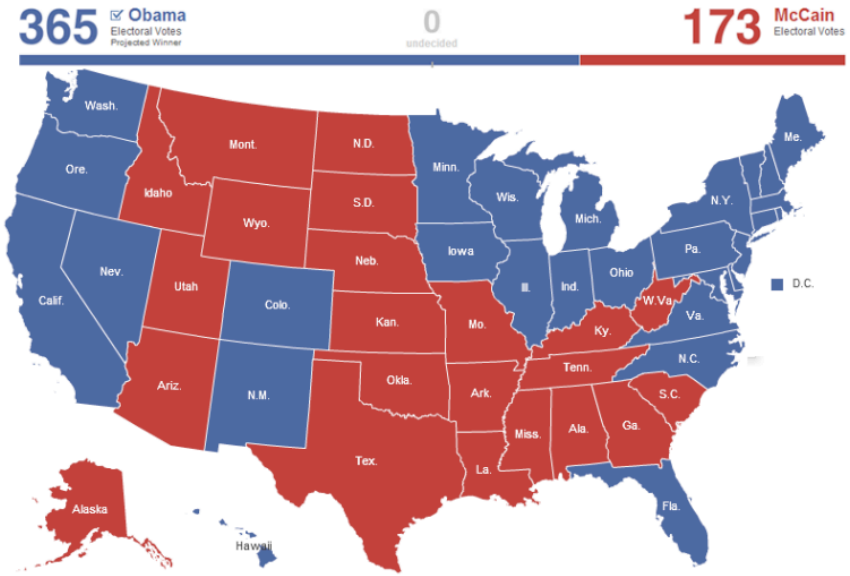
- Encode information using **how dark** the mark is drawn
- Example:



Visual dimension: color

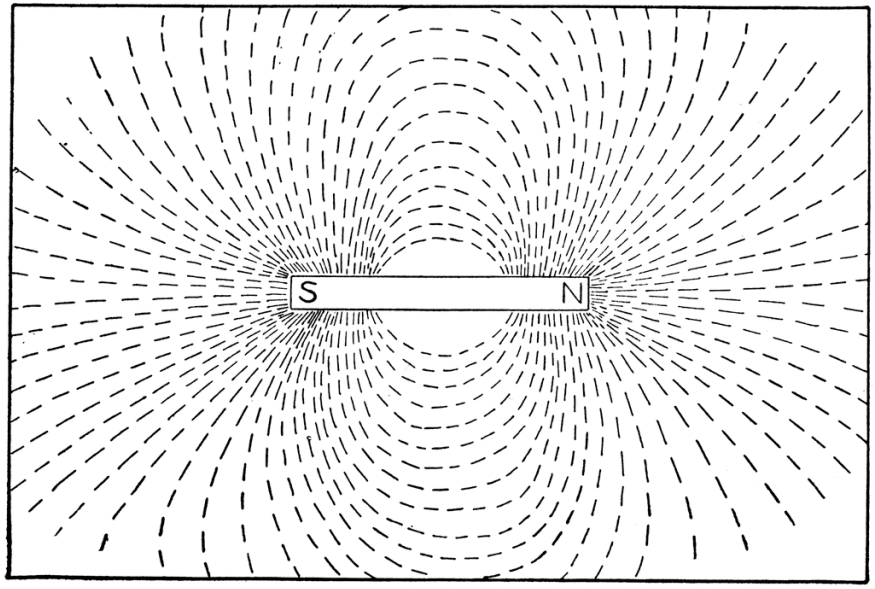
- Encode information using the **hue** of the mark
- Examples:





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: visual dimensions & data type

		Categorical	Numerical
Position			
Size			
Value			
Color			
Orientation			
Shape			

Jacques Bertin, *Semiologie Graphique* (Semiology of Graphics), 1967.

Pre-lunch activity: deconstructing graphics

1. Find a data visualization you think is interesting
   - Some ideas: NYTimes, VisualisingData.com, Visual.ly
   - Remember to cite your source!

2. Identify the following:
   - What is the **data** that's being visualized? Where did it come from?
   - Which **data dimensions** are mapped to which **visual dimensions**?
   - How does this **shape your understanding** of the data?
   - If you **liked** the visualization: what is it doing **well**?
   - If you **disliked** the visualization: what would you **change**?

Discussion

What makes a **good** encoding?

Principle 1: expressiveness

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

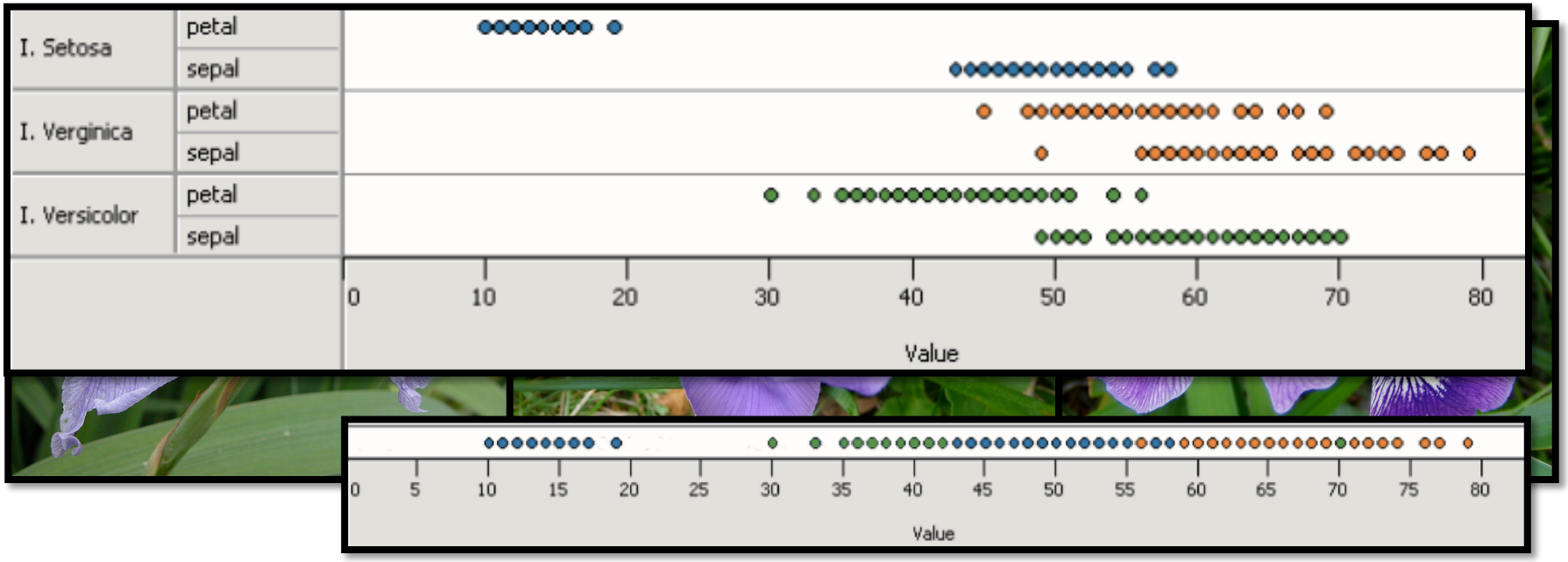
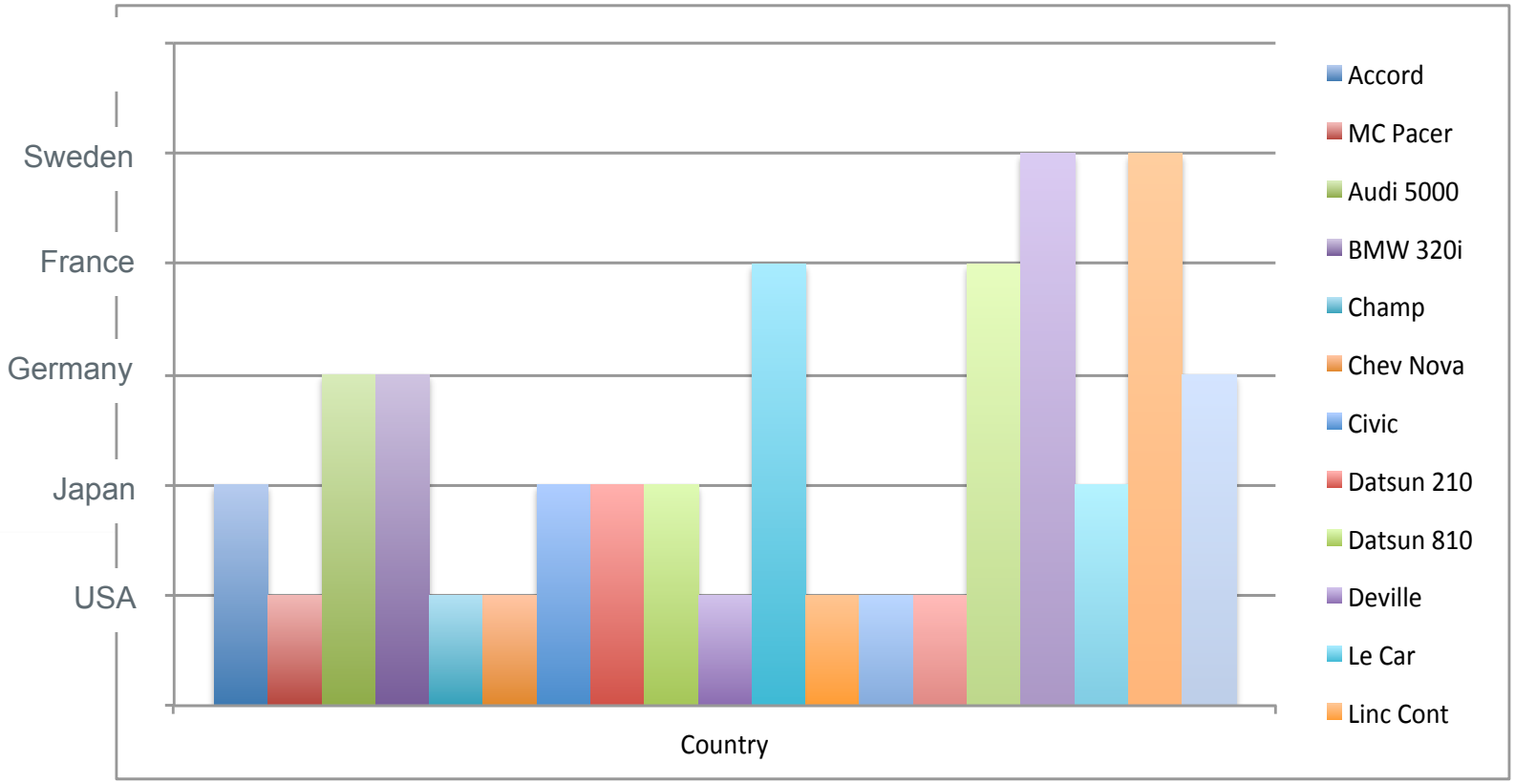


Fig. Courtesy of M Krzywinski

Principle 1: expressiveness

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

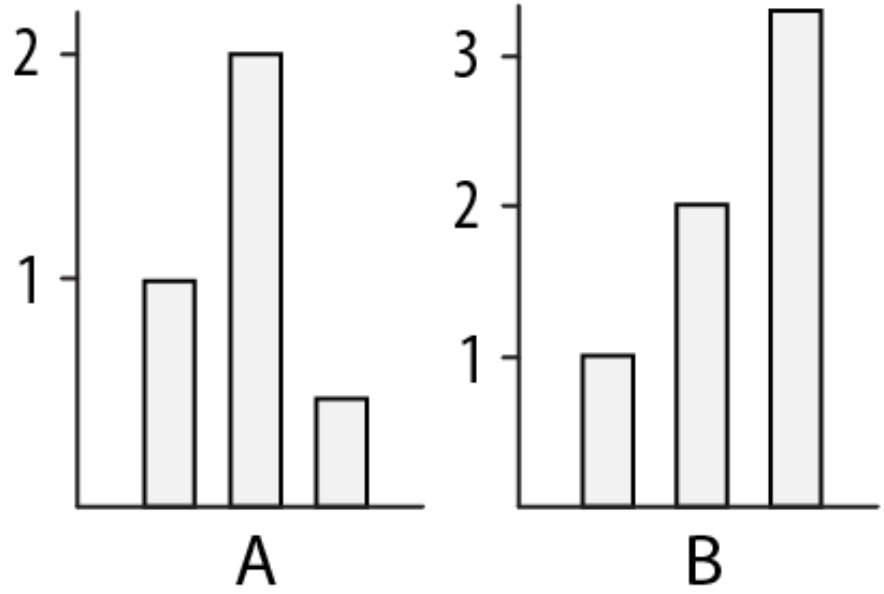


Adapted from Mackinlay J (1986) Automating the design of graphical presentations of relational information.

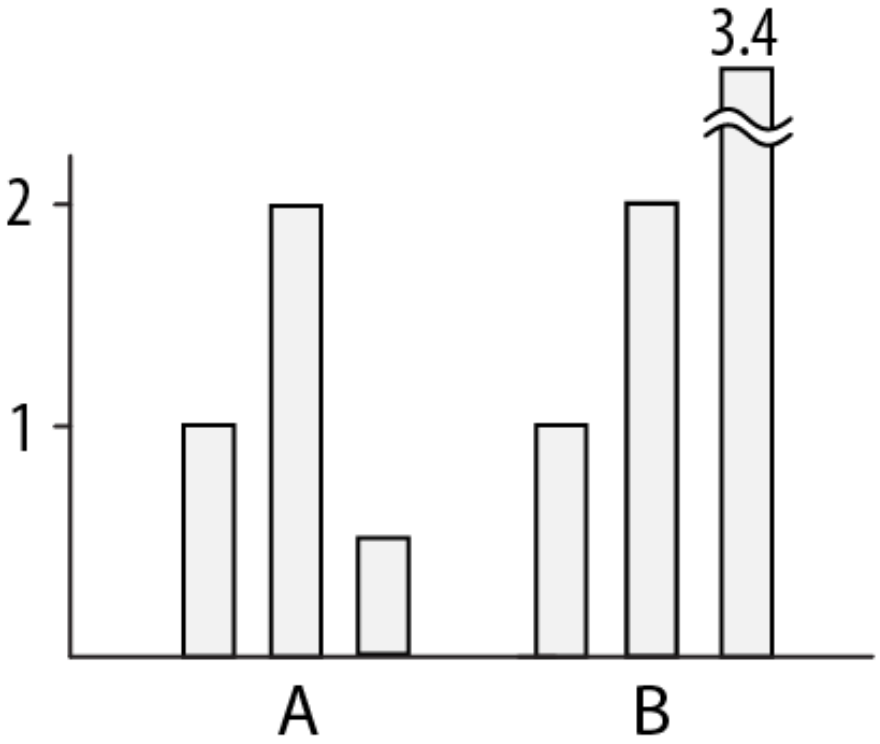
Principle 2: consistency

- Use **consistent axes** when comparing charts

misleading



improved

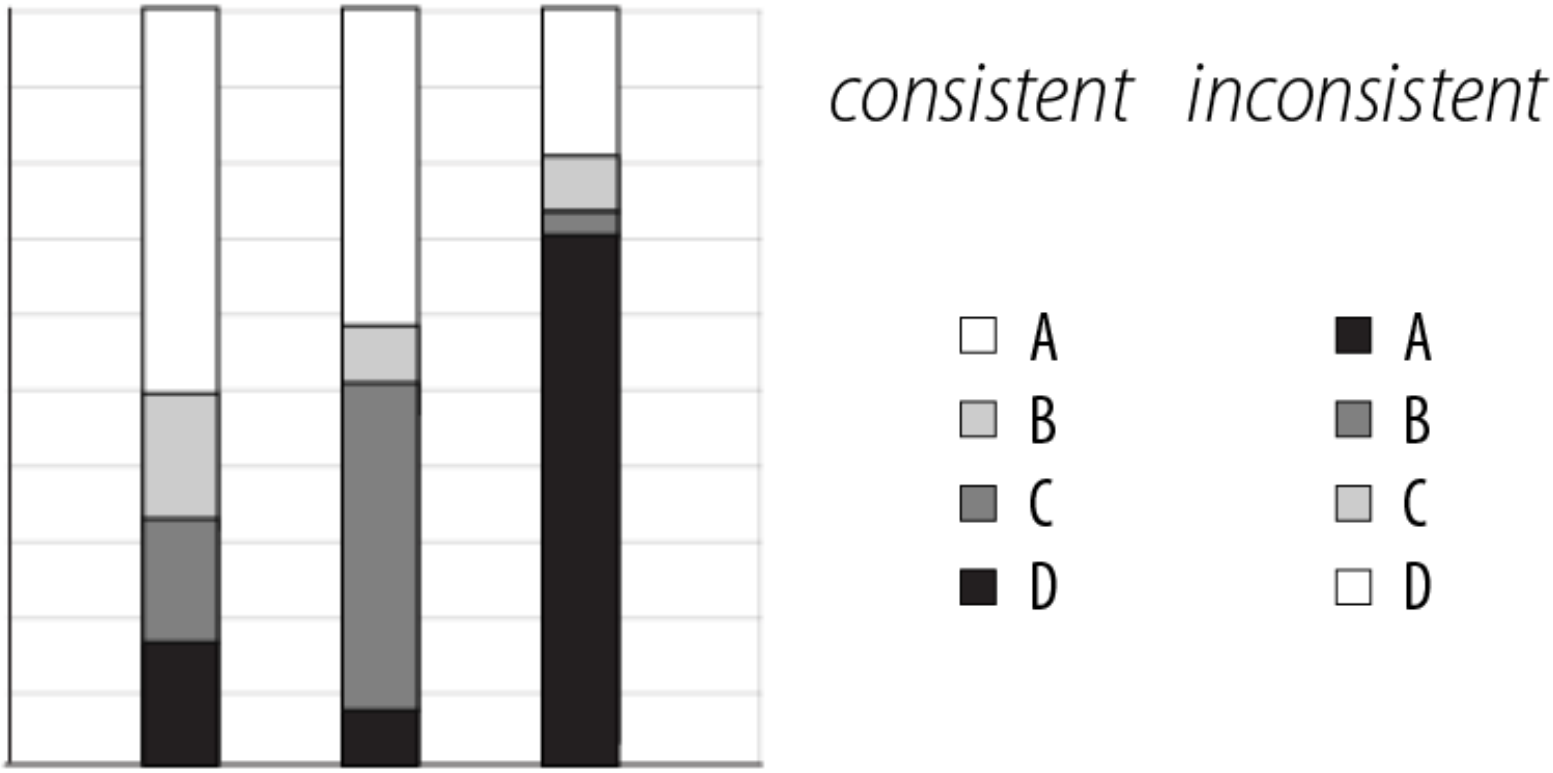


Raina SZ, et al. (2005) Evolution of base-substitution gradients in primate mitochondrial genomes. Genome Res 15: 665-673.

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

Principle 2: consistency

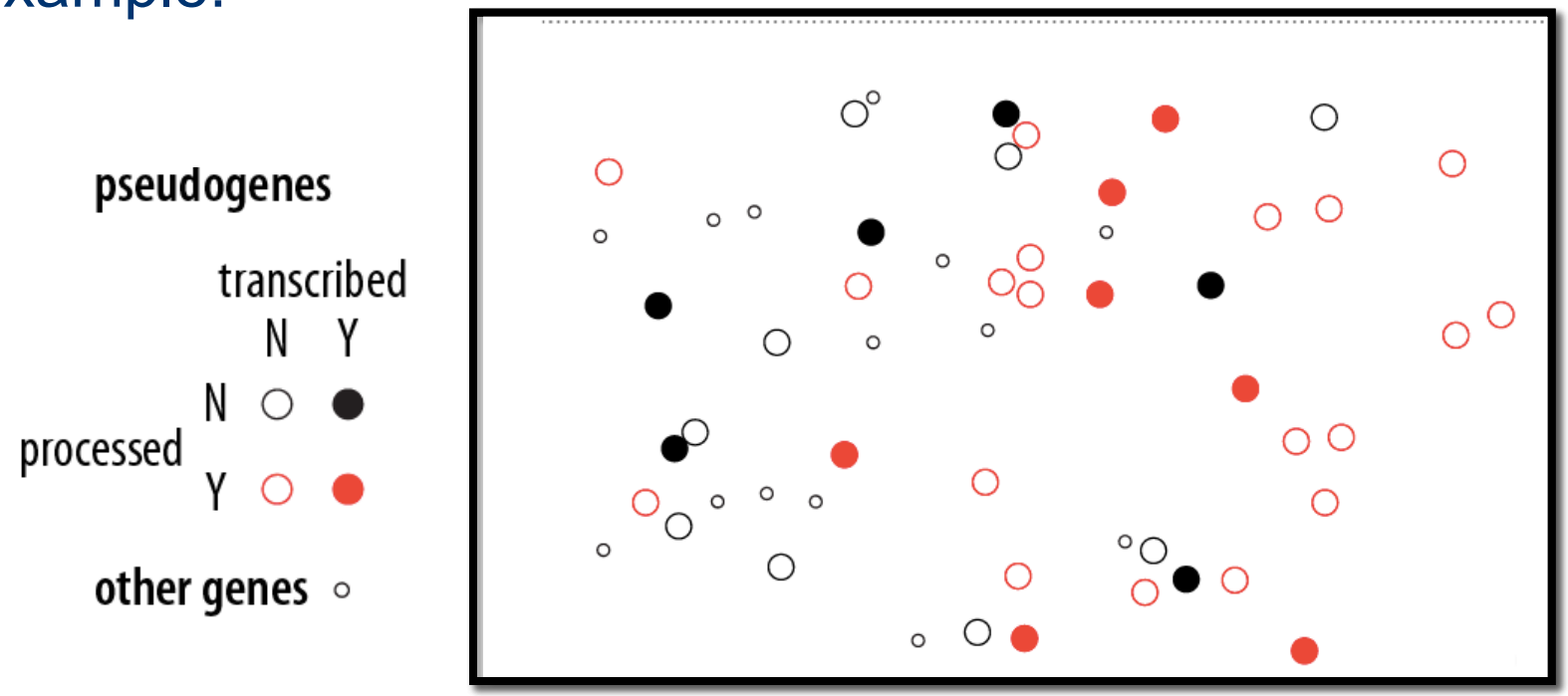
- A note on **legends**: order items according to appearance



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:

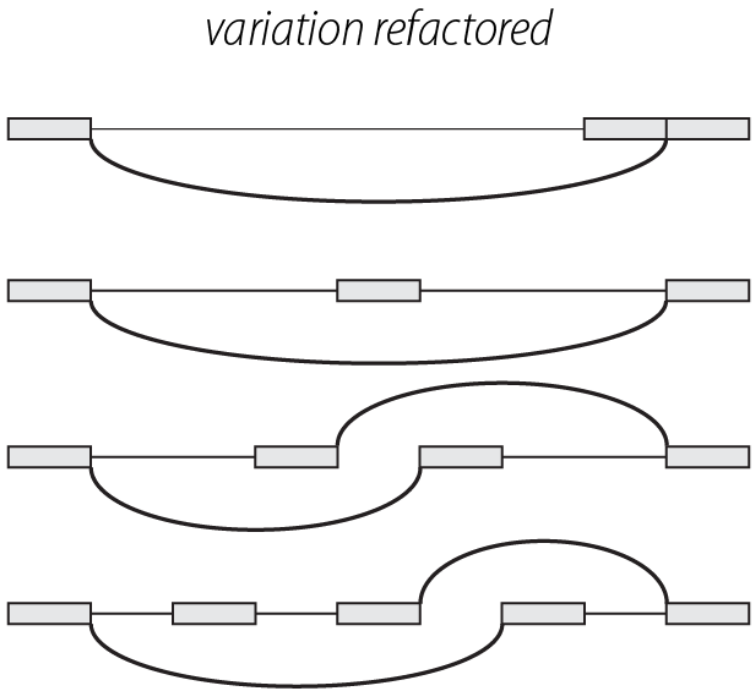
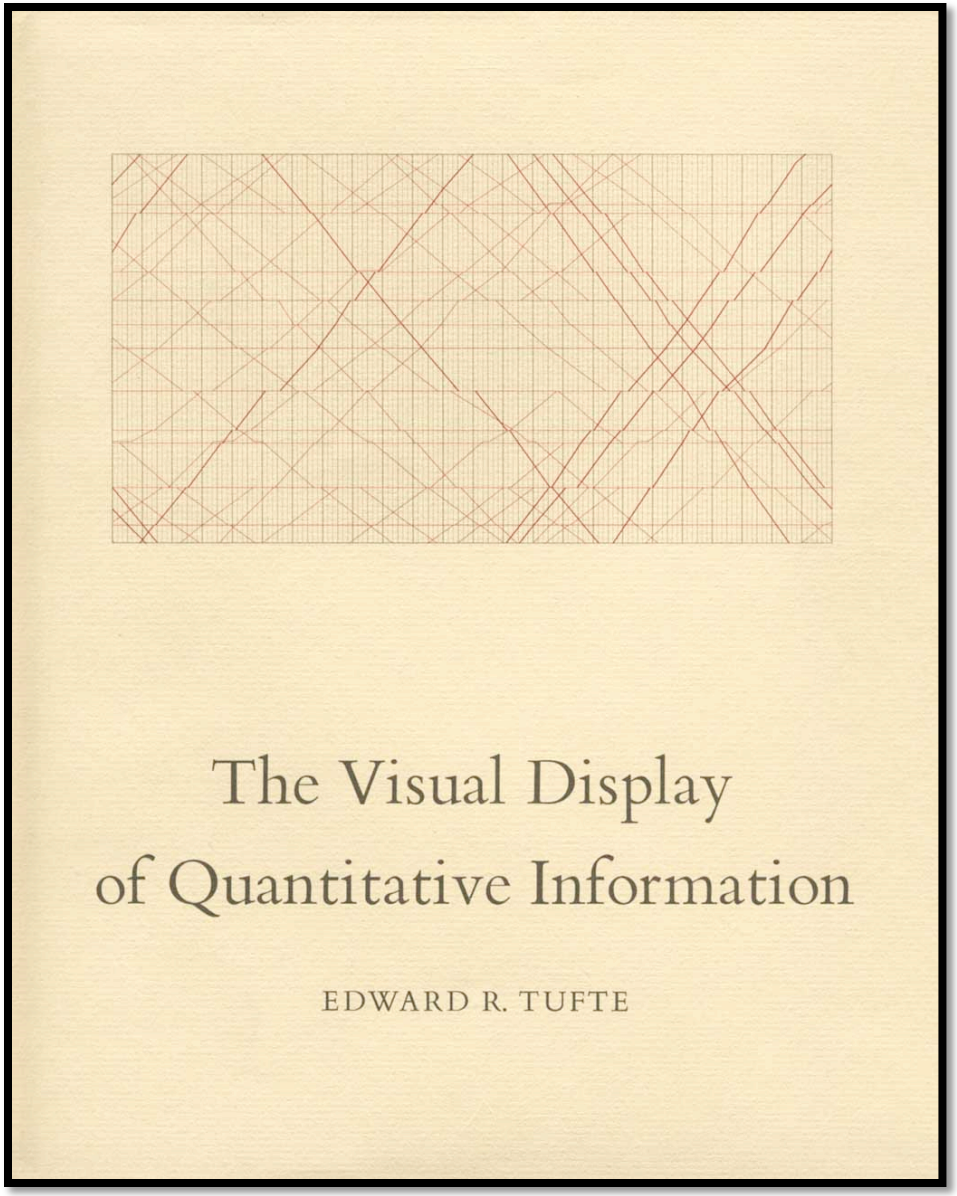


Fig. 1: Sharov AA et al. (2005) Genome-wide assembly and analysis of alternative transcripts in mouse. Genome Res 15: 748-754.
Fig. 2: M. Krzwinski, behind every great visualization is a design principle, 2012

Tufte, 1983

"Above all else, show the data."



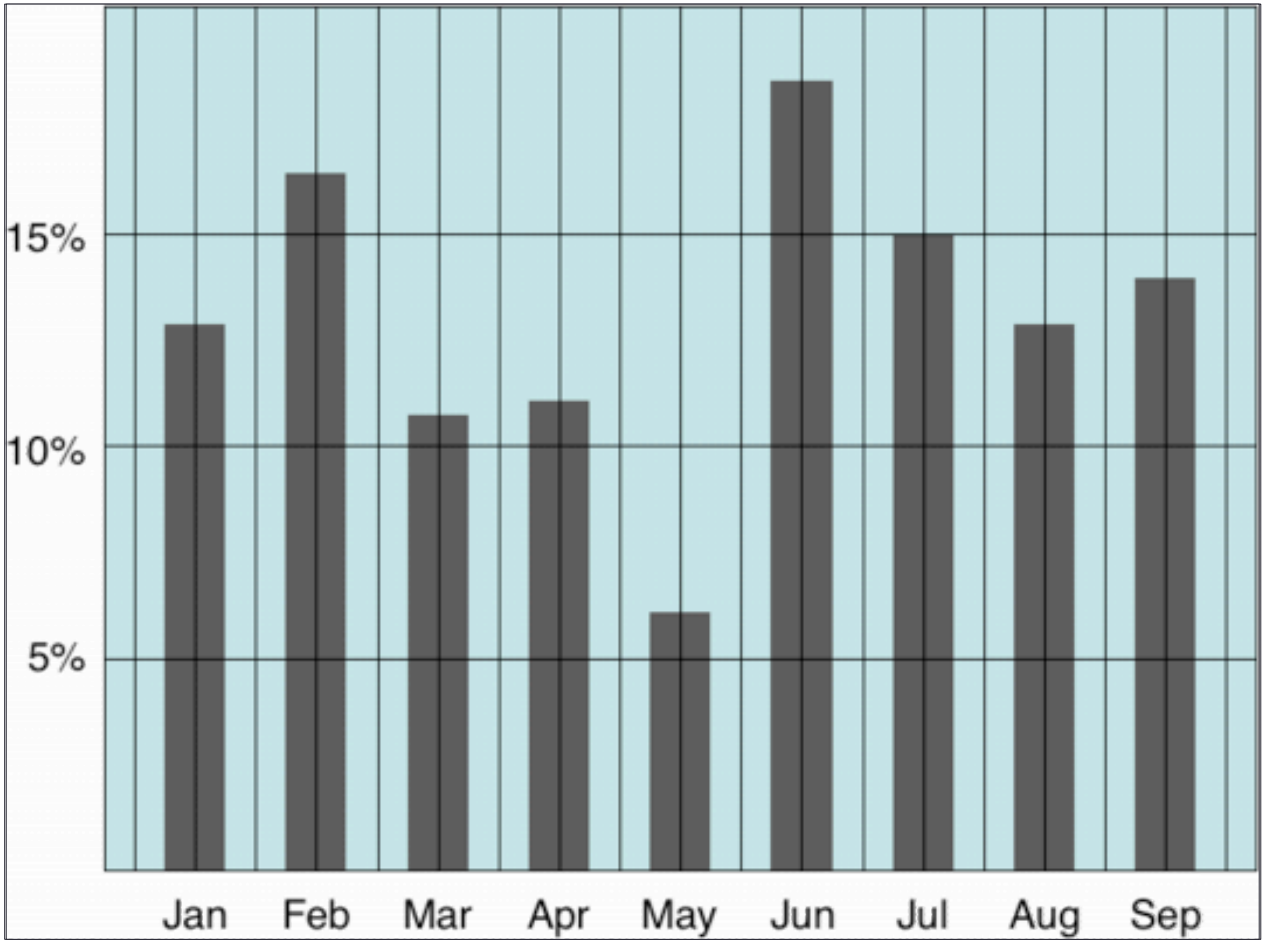
Tufte, 1983

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

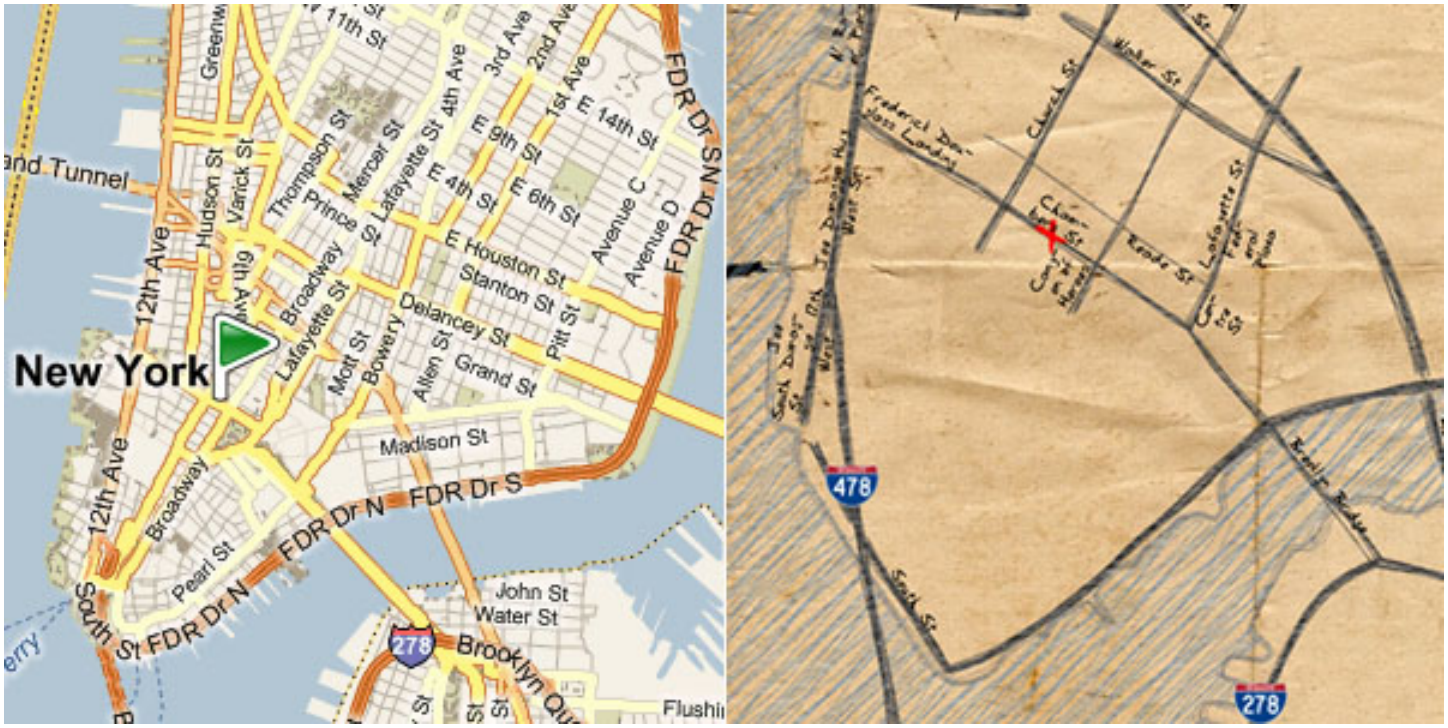
= proportion of a graphic's ink devoted to the non-redundant display of data-information

= 1 - proportion of a graphic that can be erased

Tufte: maximize the data-ink ratio



Familiar example

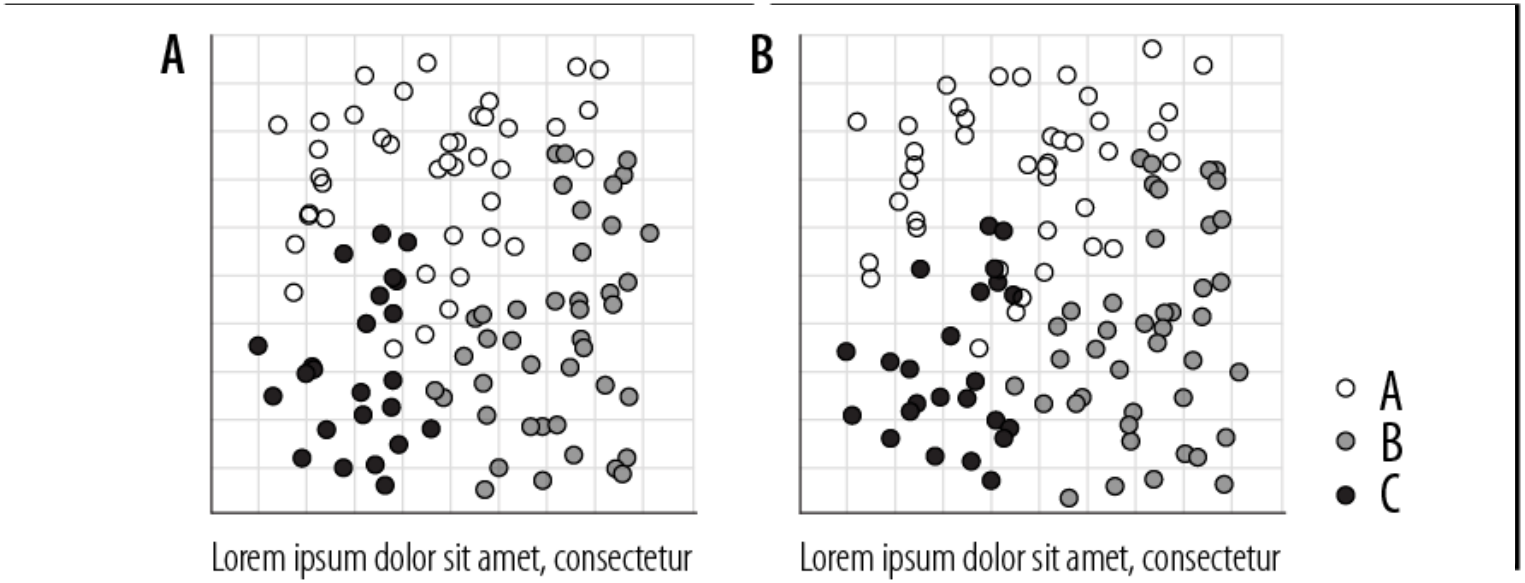


Discussion

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

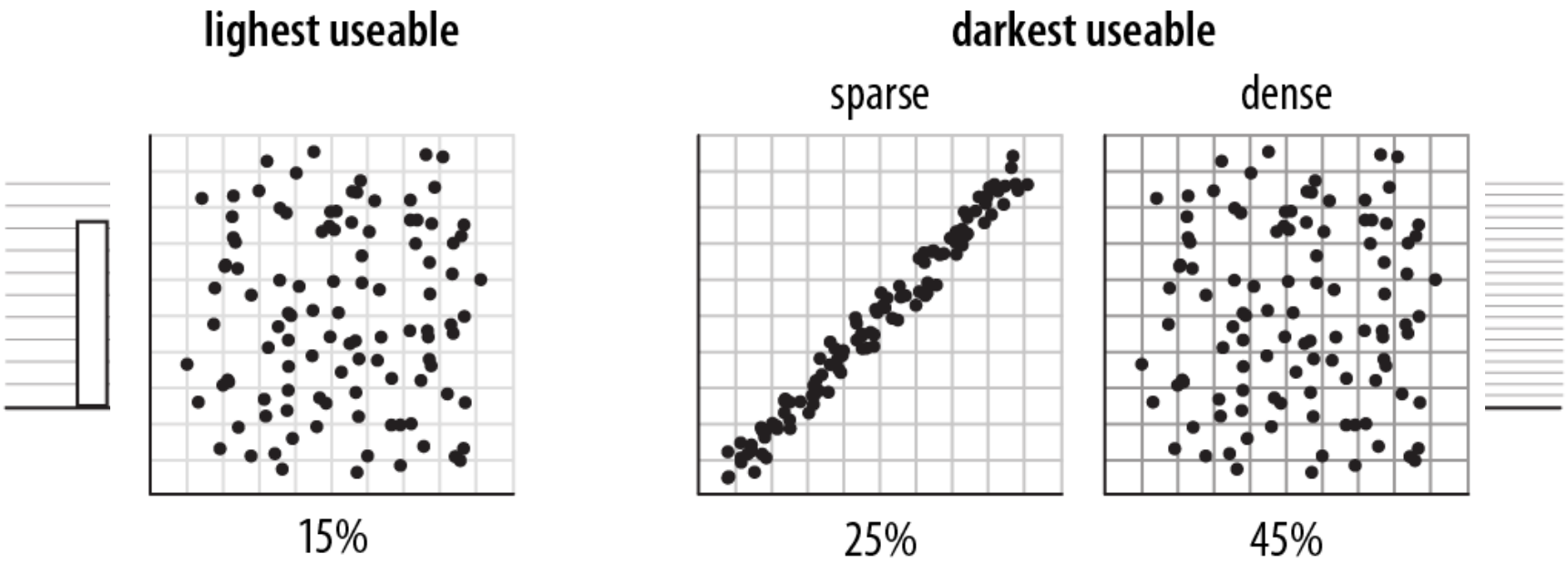
Principle 3: importance ordering

- Avoid unnecessary containment and repetition
- Example



Principle 3: importance ordering

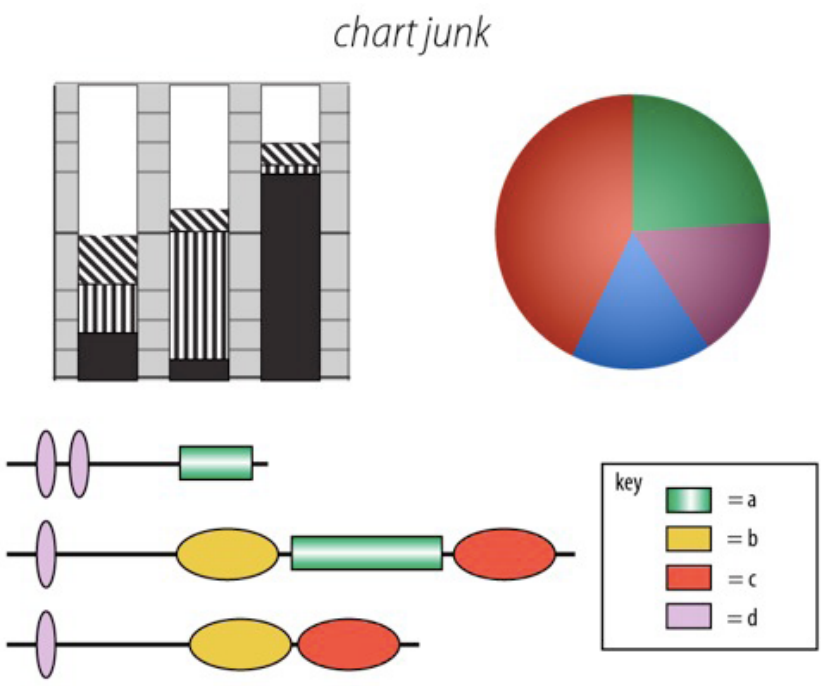
- Navigational aids shouldn't compete with data
- Avoid: heavy **axes**, **error bars** and **glyphs**

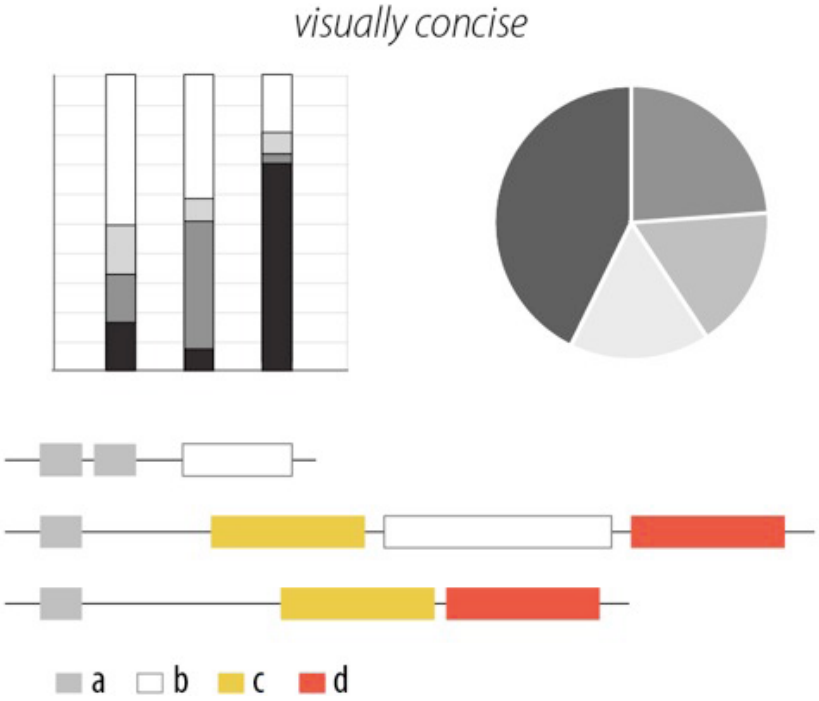


Heer J, Bostock M (2010) Crowdsourcing graphical perception: using mechanical turk to assess visualization design. Proceedings of the 28th international conference on Human factors in computing systems. Atlanta, Georgia, USA: ACM. pp. 203-212.

Principle 3: importance ordering

- Simplify, simplify, simplify…

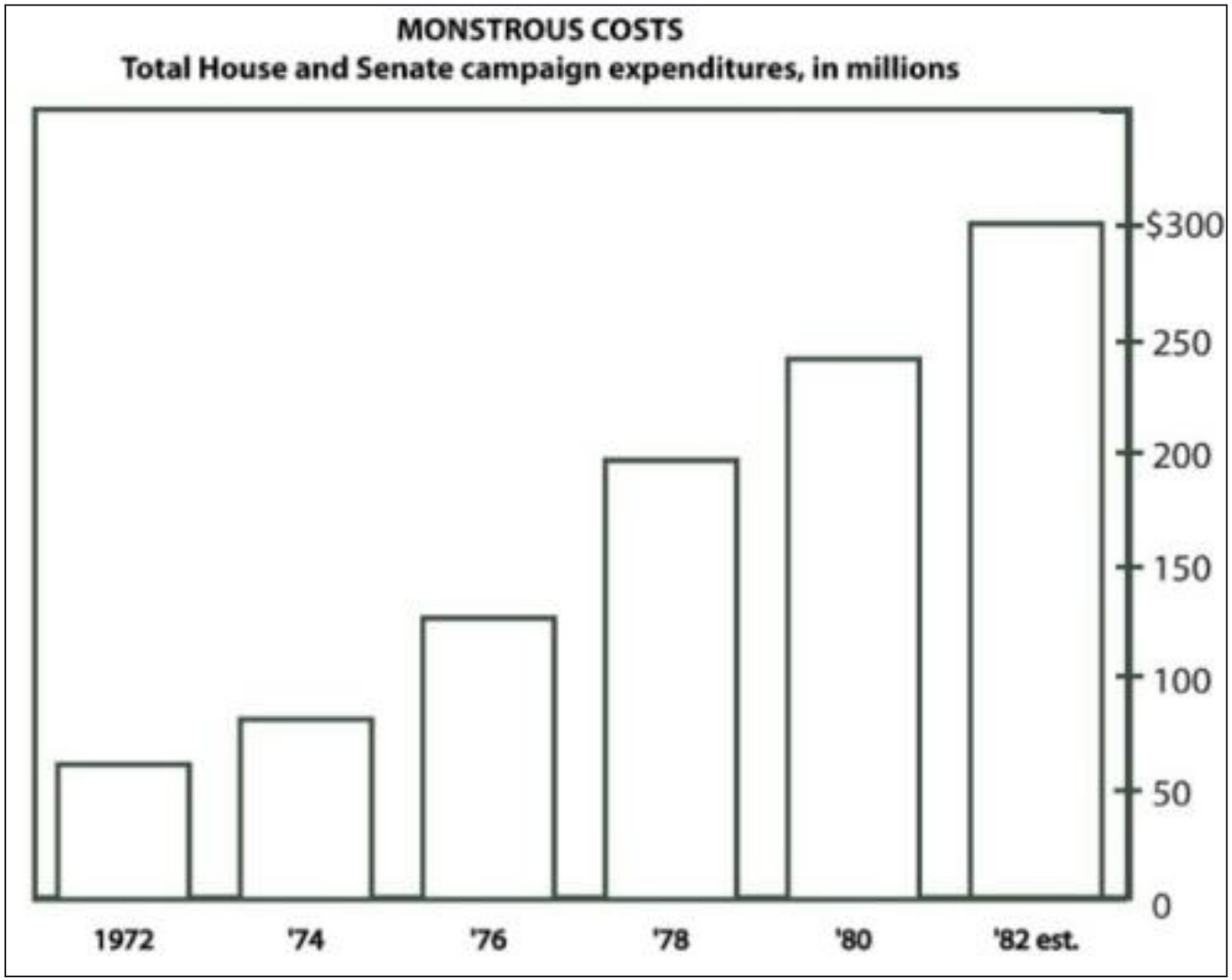




Sharov AA, et al (2006) Genome Res 16: 505-509.
Peterson J, et al. (2009) Genome Res 19: 2317-2323.
Thomson NR, et al. (2005) Genome Res 15: 629-640.
DB, Ko MS (2005) Genome Res 15: 748-754.

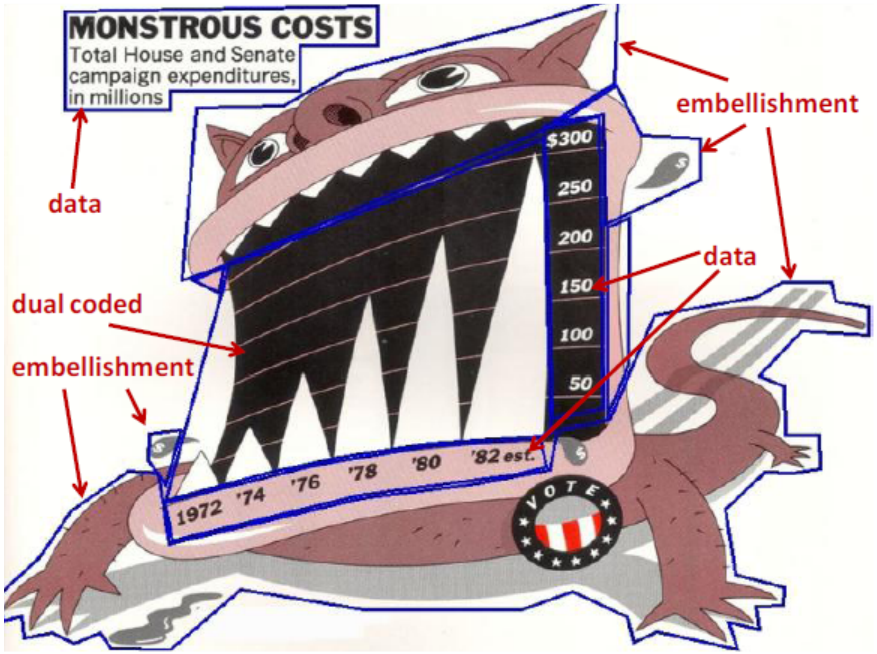
M. Krzwinski, behind every great visualization is a design principle, 2012

A caveat: “chart junk” and recall



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

Chart junk and eye gaze



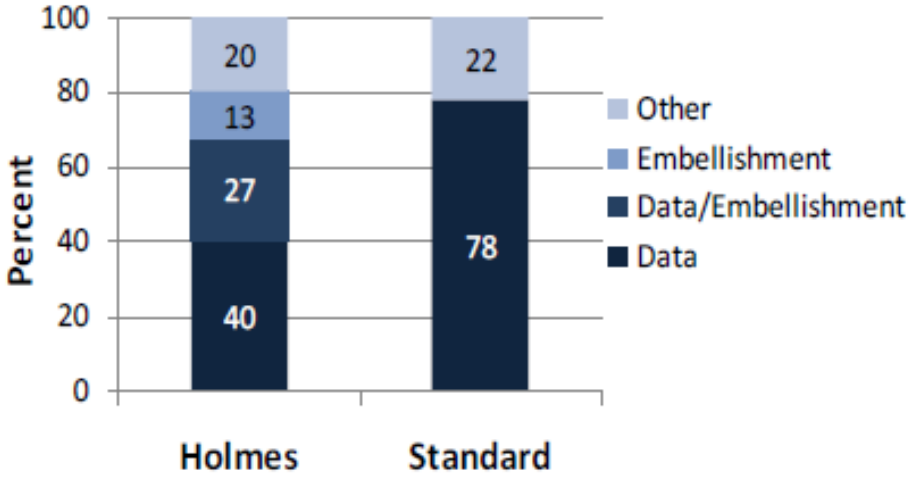


Figure 9. Percentage of on-screen time spent looking at different chart elements for Holmes and Plain charts.

After lunch

- Mini-lecture and lab: building data graphics with `ggplot2`
- **TODO** (if you haven't already):

```
> install.packages('ggplot2')
```