

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](http://www.science.smith.edu/~jcrouser)

# Some housekeeping

- Workshop website:

**[jcrouser.github.io/MassMutual-DataVis](https://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

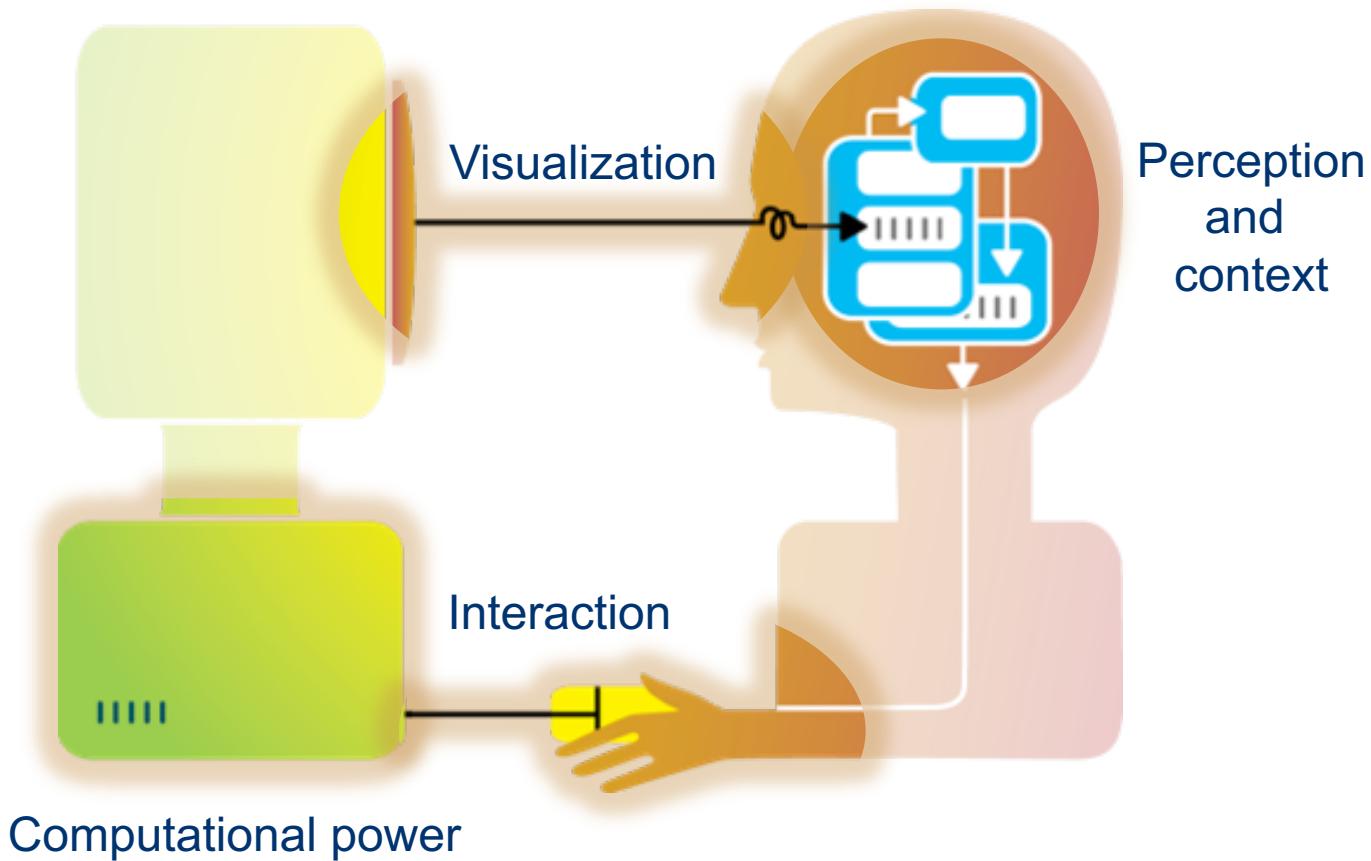
3. Opportunities to  
get to know  
your new team

# What I do: analytical tools for messy data



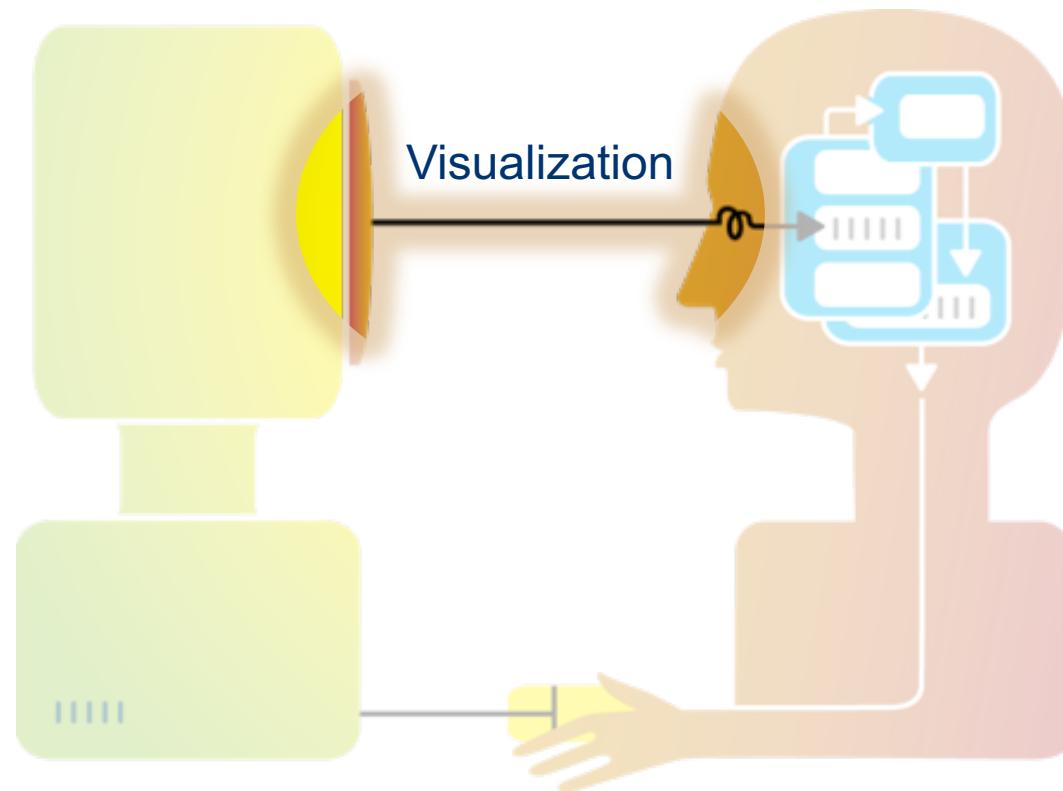
# Big idea behind my research

Humans and machines have **complimentary strengths**



# Focus of this workshop

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



# 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?

A screenshot of a Google search results page for the query "visualization". The top navigation bar shows "Jordan" as the user, with icons for camera, search, and profile. Below the bar, there are tabs for All, Videos, Images (which is selected), News, Books, More, and Search tools. To the right are buttons for SafeSearch and settings. The main content area displays a grid of images under several categories: Reading Strategy, Data, Quotes, Sports, Creative, and Techniques. Below these are two rows of more images, including a colorful eye, a brain network, a person's head with hands on their temples, a bar chart, a word cloud with "visualize", a network graph, a blue eye with text, a head silhouette with gears, a brain with gears, and a couple looking at a globe.

# What is visualization?

Google search results for "information visualization".

Search bar: information visualization

Navigation: All, News, **Images**, Books, Videos, More, Search tools, SafeSearch, Settings

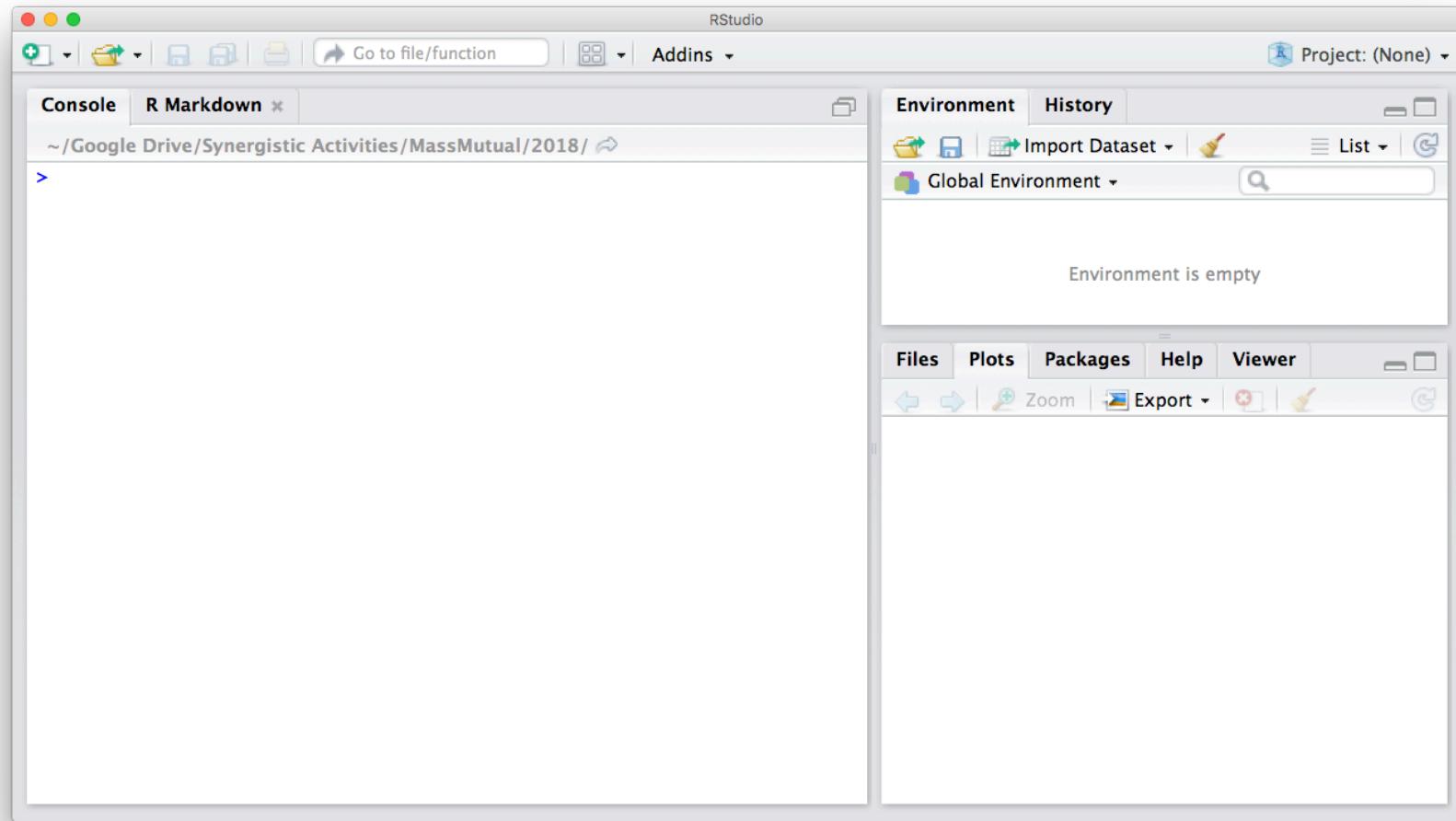
Profile: Jordan

Grid view: 12 items

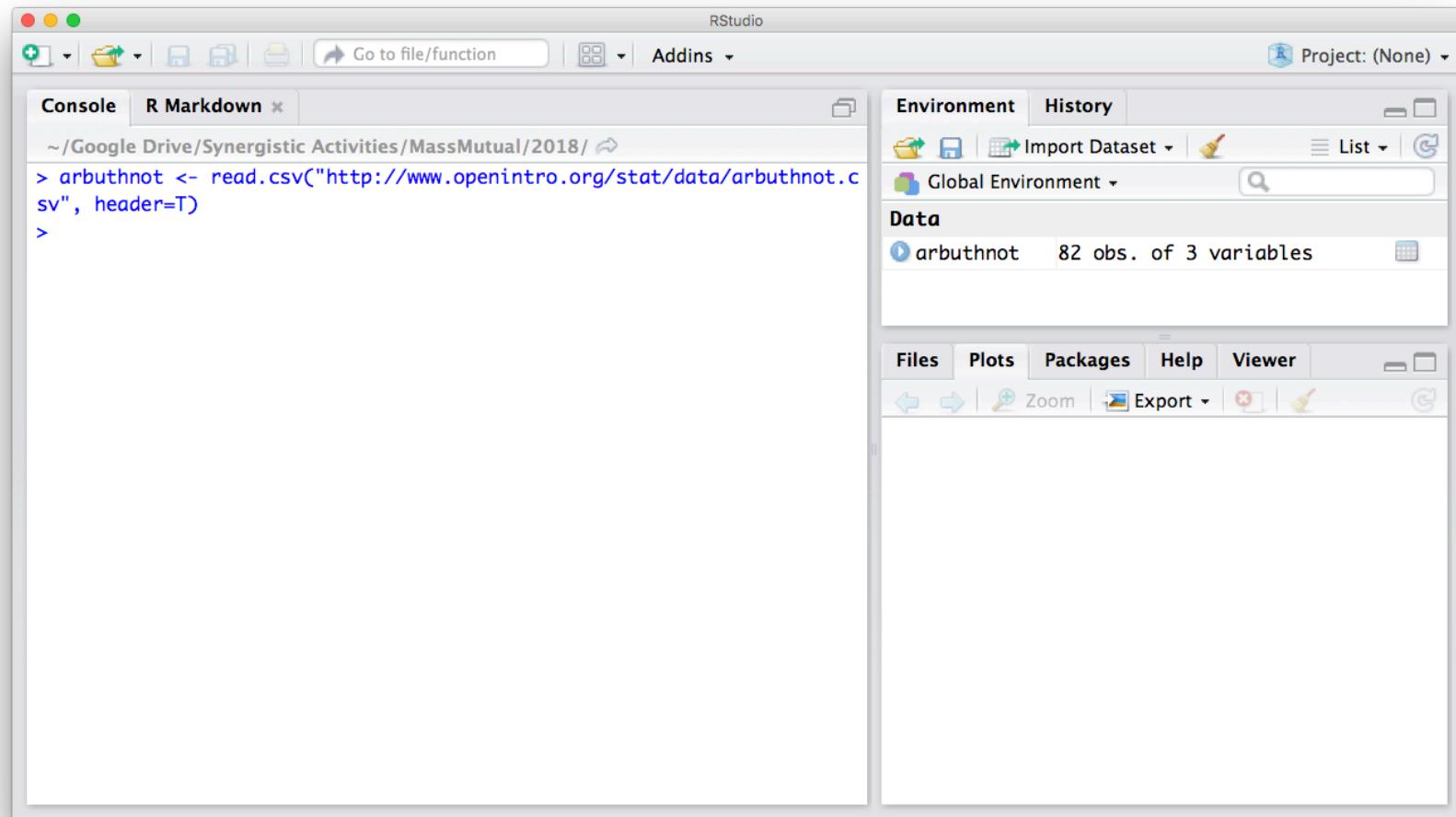
Items:

- Examples:** A circular sunburst chart showing various categories.
- Design:** A complex network diagram with many nodes and connections.
- Tools:** A collection of images showing different visualization interfaces.
- 3D:** A 3D visualization interface showing data in a spatial context.
- Data Visualization Map:** A world map where each country's color represents its GDP per capita.
- Data:** A circular sunburst chart showing data distribution.
- Network visualization:** A large circular network graph with many colored nodes and lines.
- Money flow visualization:** A circular sunburst chart showing money flowing between countries.
- Conceptual map:** A circular tree map where the size of each node represents its importance.
- World Map 2010:** A world map showing GDP per capita in US dollars.
- Infographic:** A circular sunburst chart with large numbers (61, 43) indicating specific data points.
- Music visualization:** A complex network diagram showing relationships between musical genres.
- Information visualization design:** A circular sunburst chart with a central hub and radiating segments.
- Line chart:** A line chart showing data over time with multiple colored series.
- Media network visualization:** A network diagram showing connections between various media outlets like MSN, NYTimes, and BBC.
- Category distribution visualization:** A circular sunburst chart showing the breakdown of categories.
- Scatter plot:** A scatter plot with large numerical labels (2, 8) and data points.
- Political spectrum visualization:** A circular sunburst chart showing political affiliation across different regions.

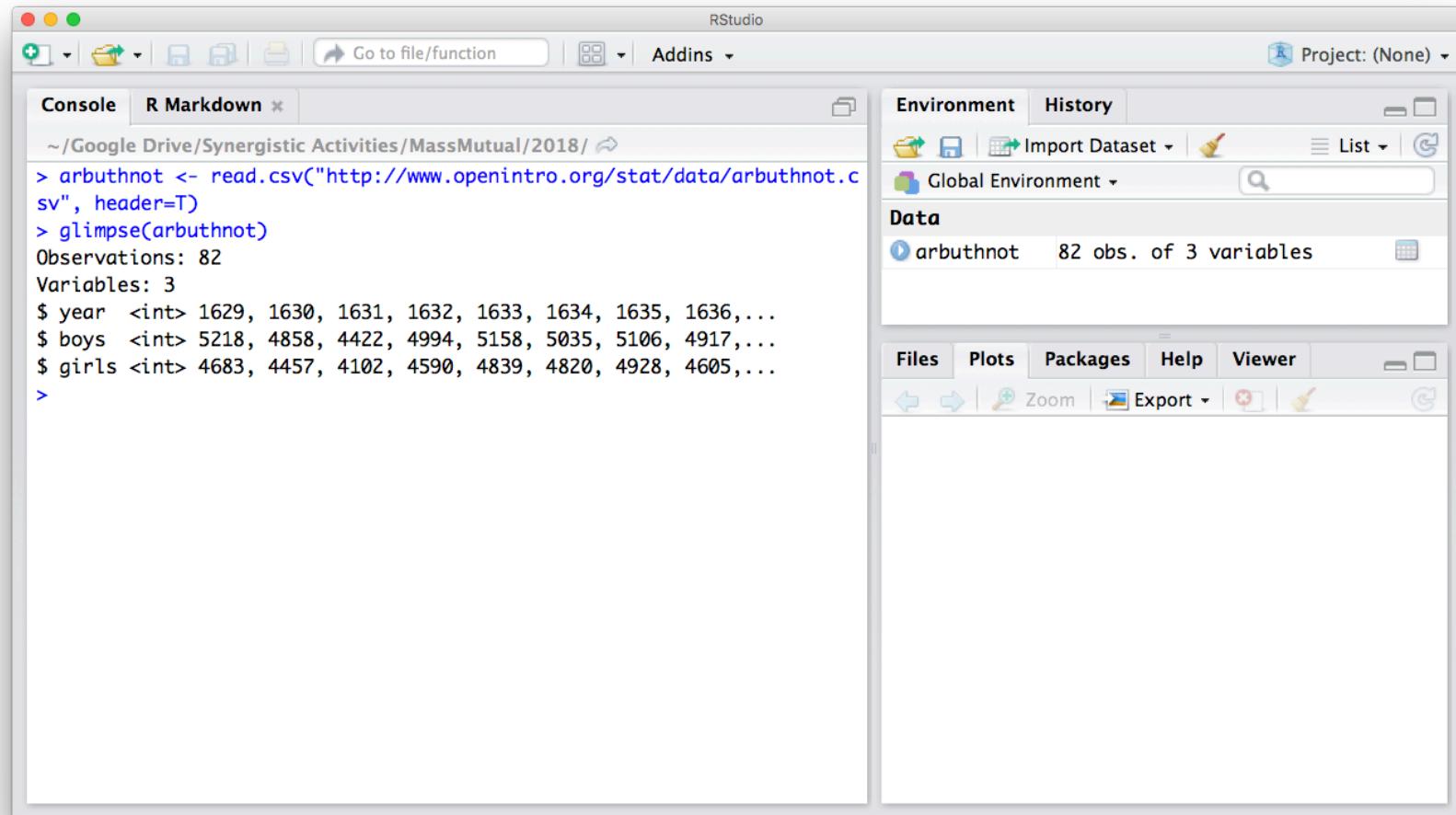
# Flashback...



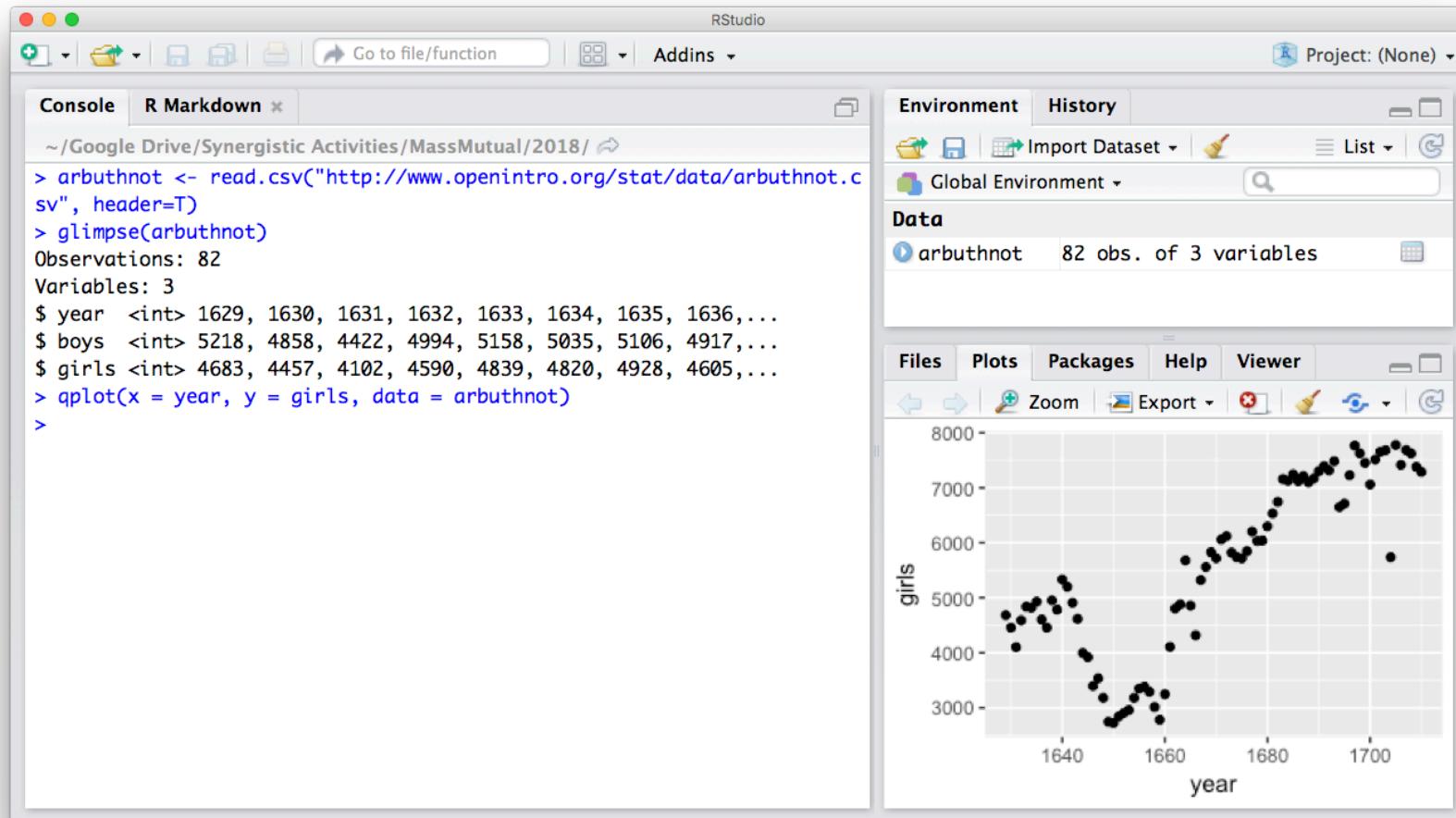
# Flashback...



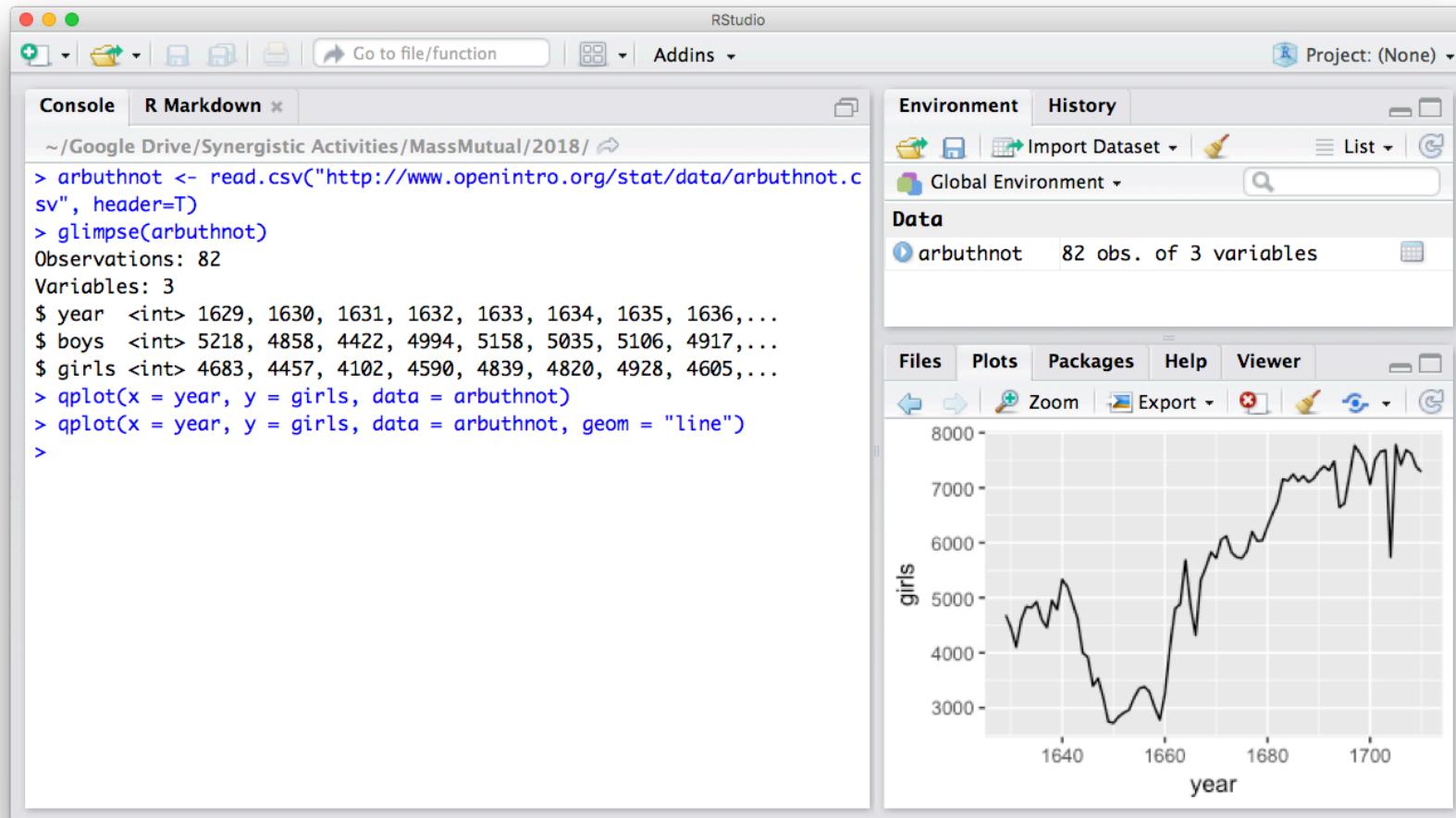
# Flashback...



# Flashback...



# Flashback...



# Question

What makes these  
“visualizations” **useful?**

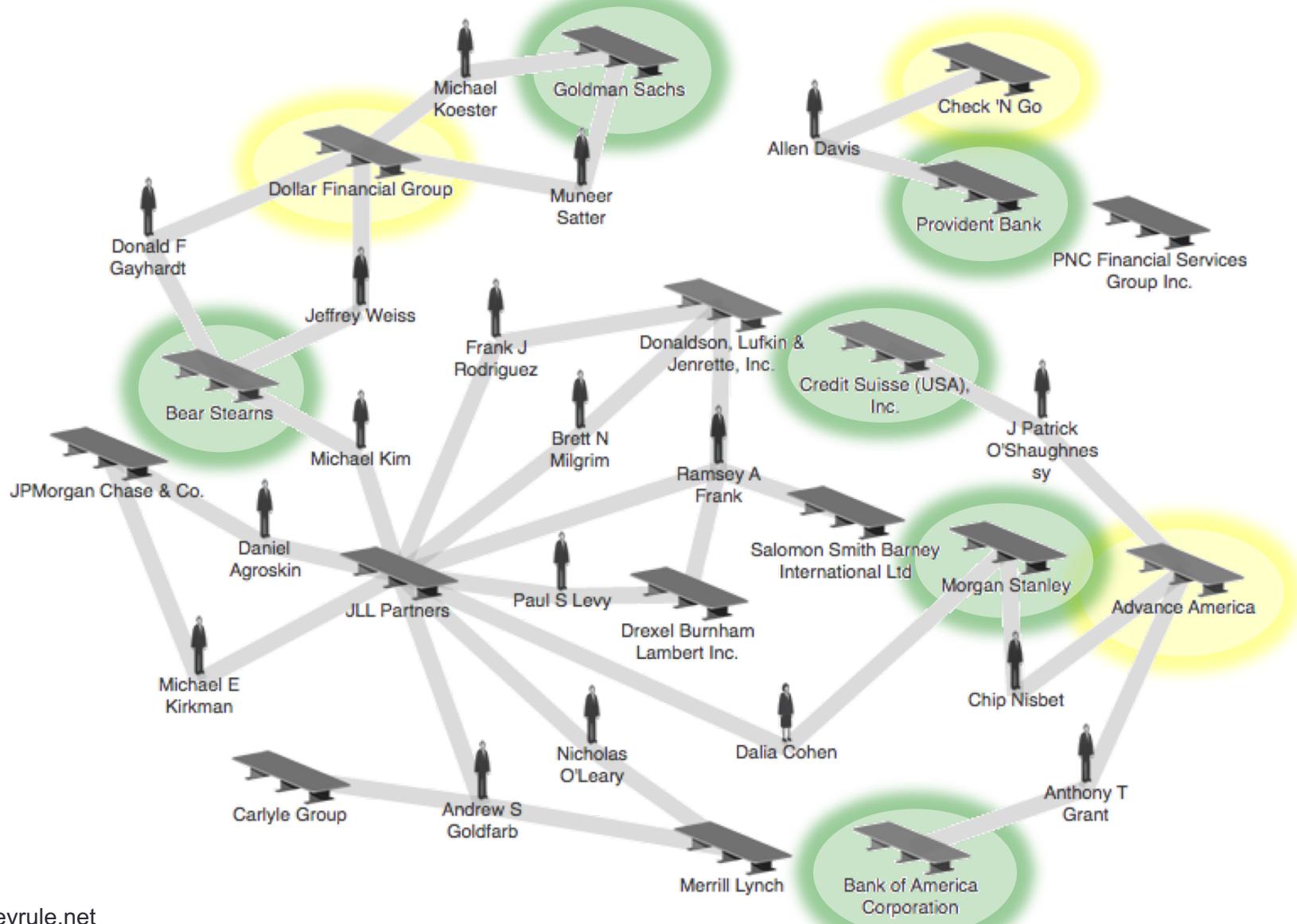


# Do they help you spot trends?



More info here: [http://en.wikipedia.org/wiki/1854\\_Broad\\_Street\\_cholera\\_outbreak](http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak)

# Do they help you explore?



# 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.*

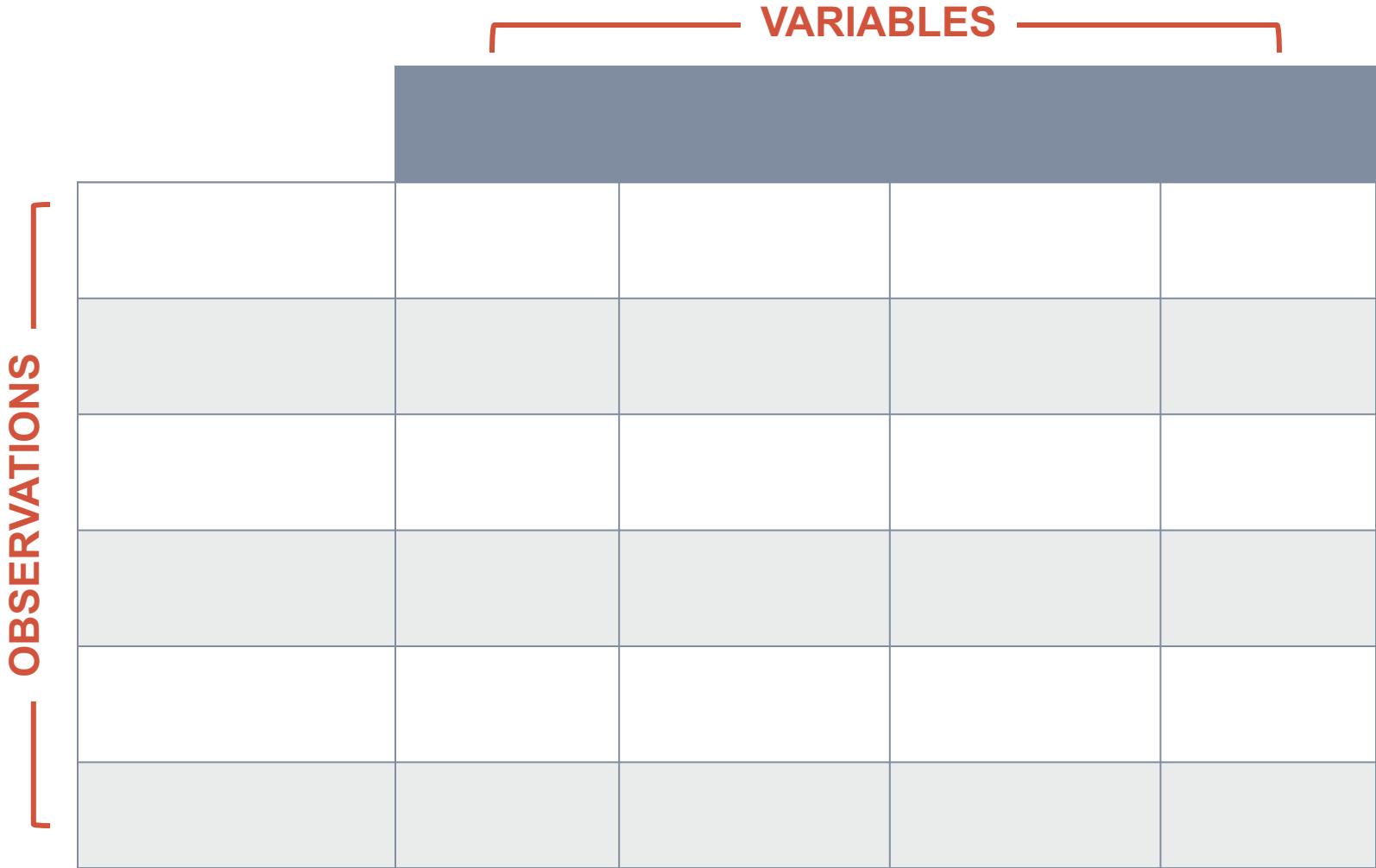
# Data (def.)

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:

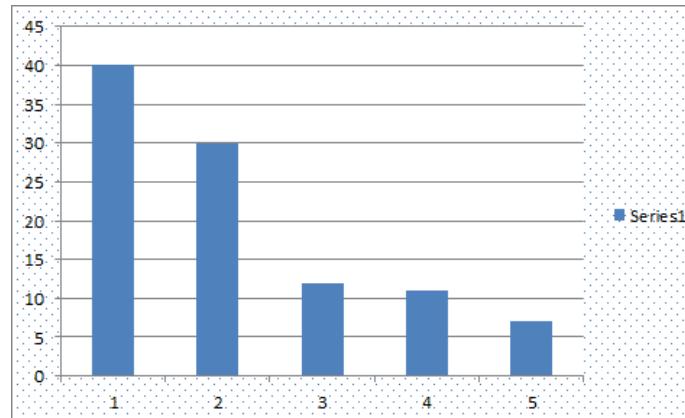
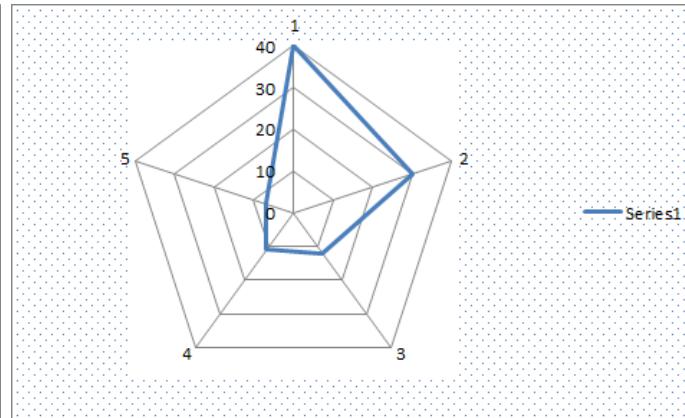
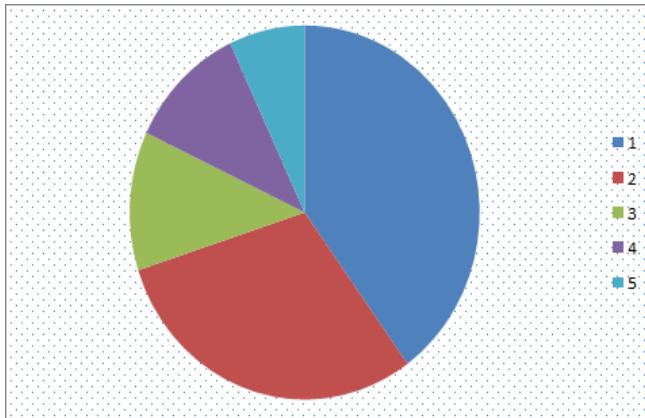


# Why is this important?

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

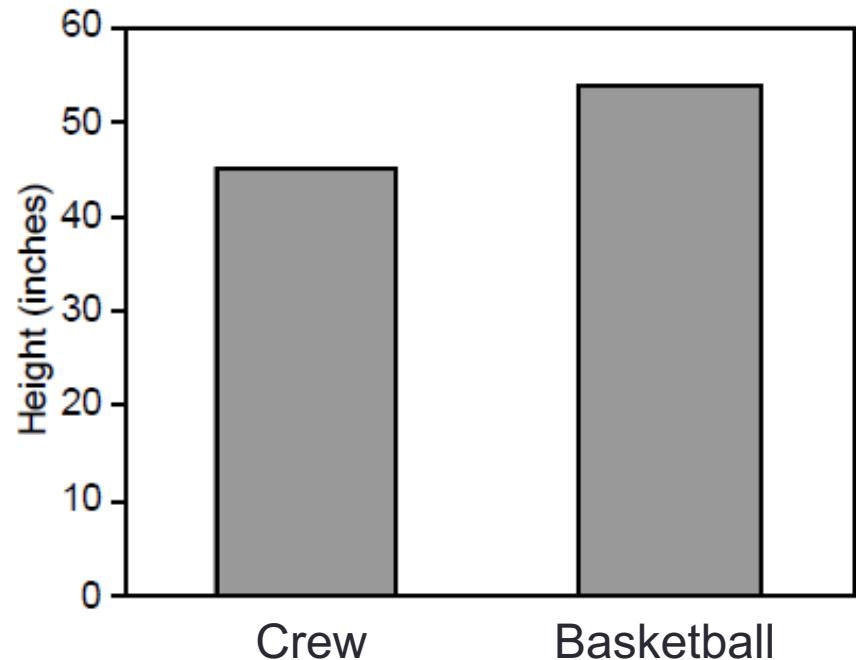
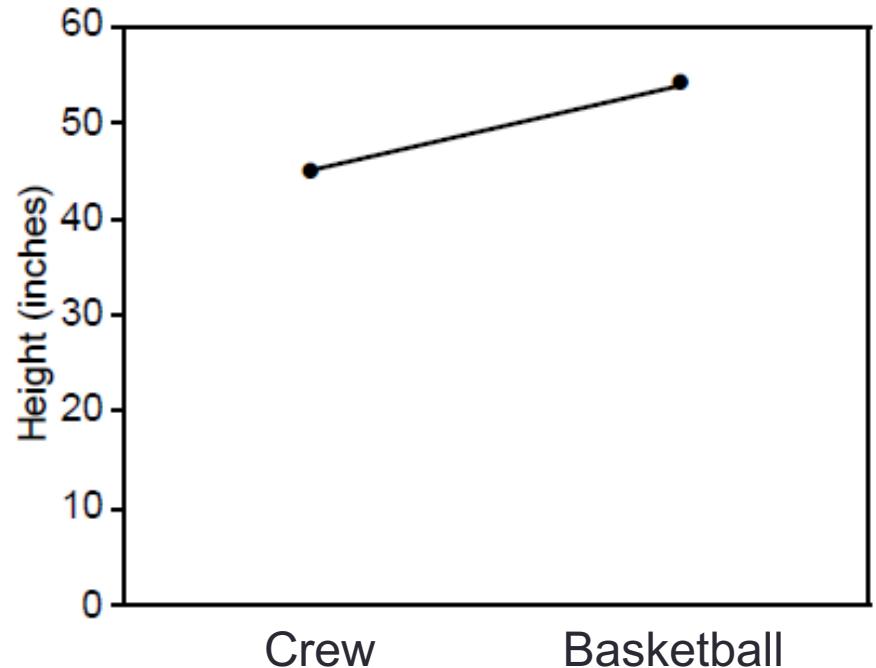
# Key question for this workshop

Which **data dimension** should be mapped  
to which **visual dimension**?

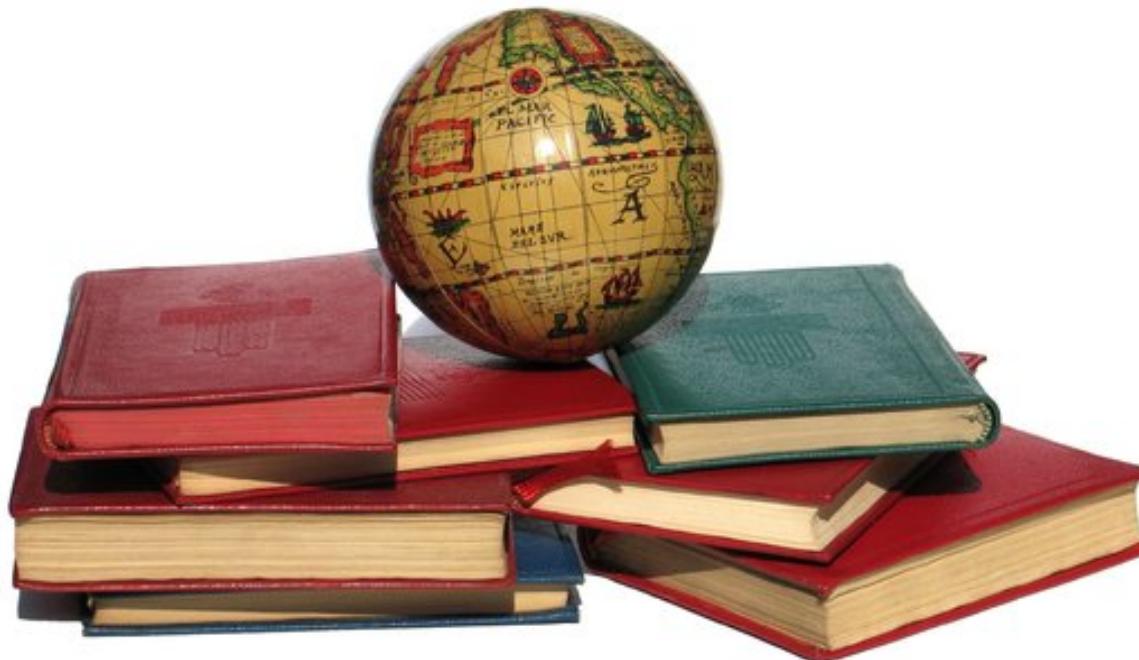


# Answer: it depends

Average Height for Youth Sports Participants



# A quick history lesson...

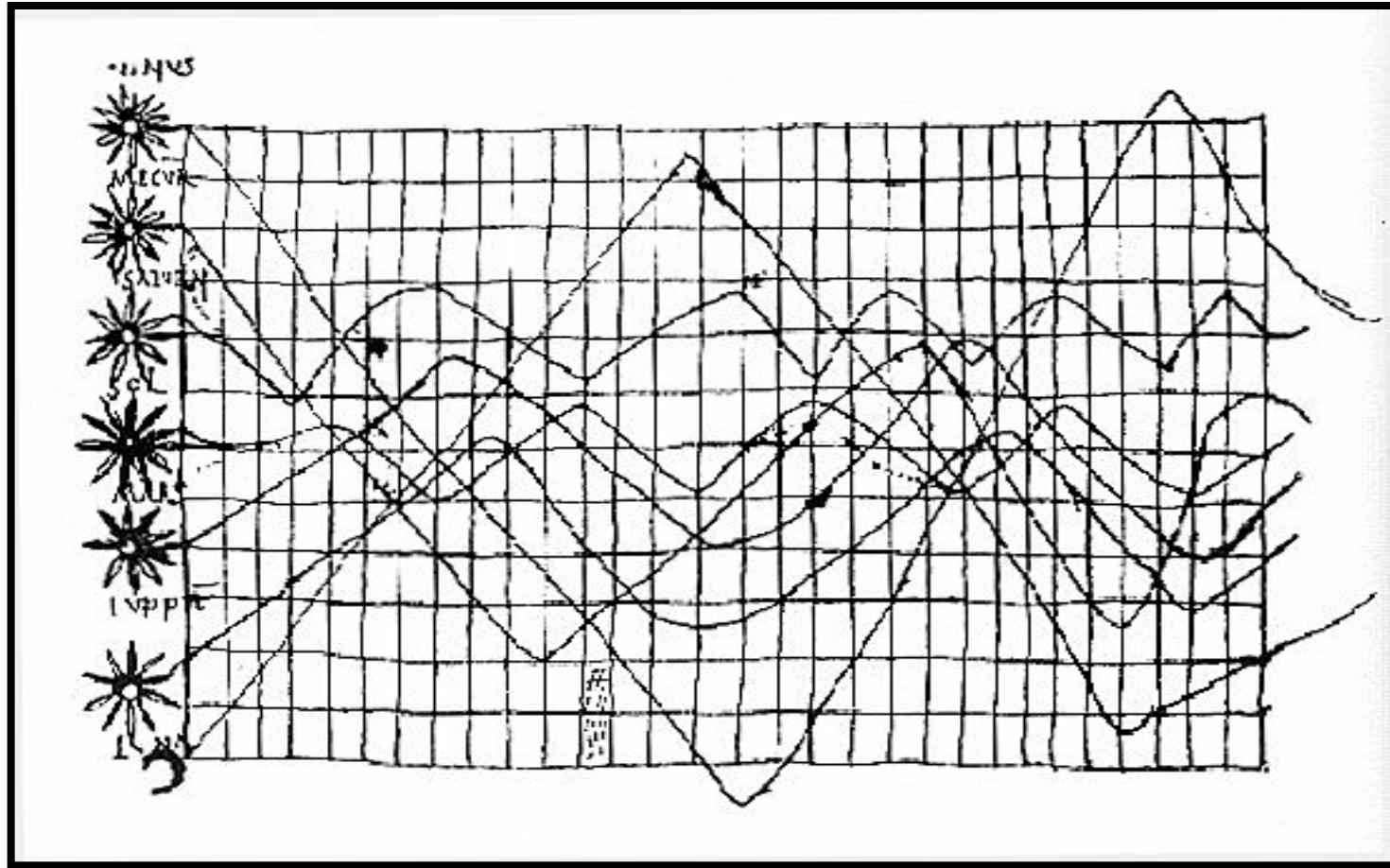


# (Incomplete) History of Visualization: 15,000BC



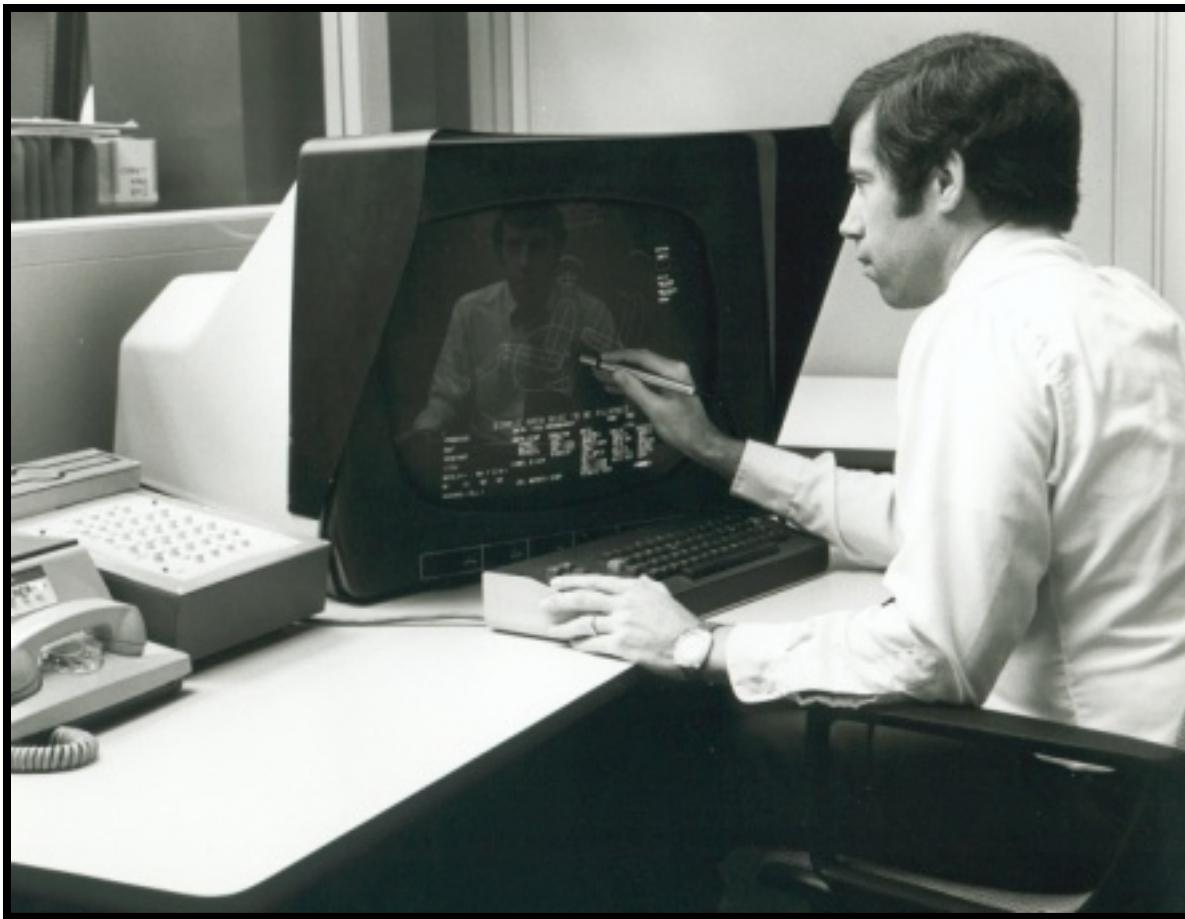
15,000 BC. Laxcaux, France

# (Incomplete) History of Visualization: 900s



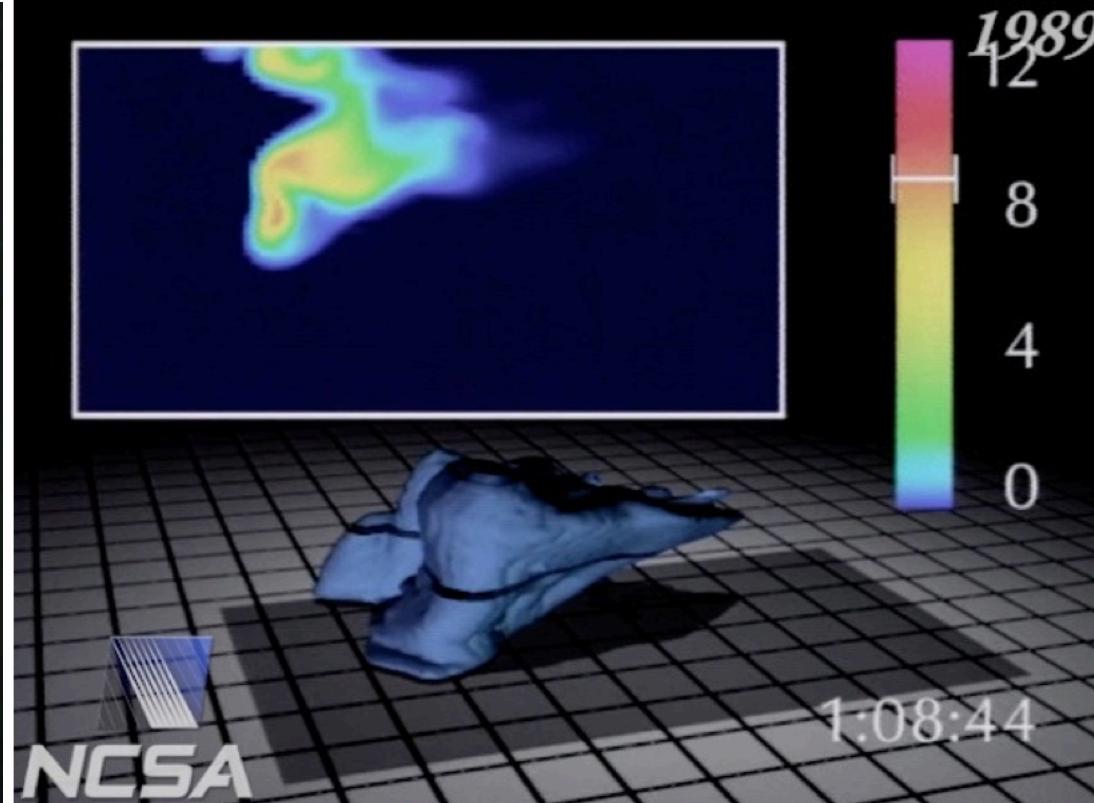
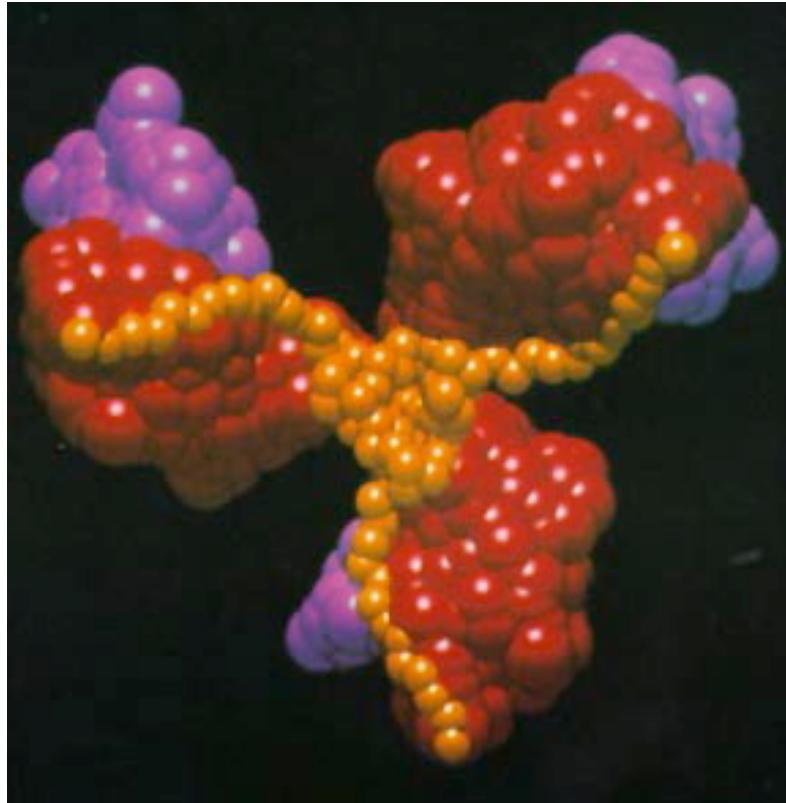
"De cursu per zodiacum", illustrator unknown

# (Incomplete) History of Visualization: 1970s



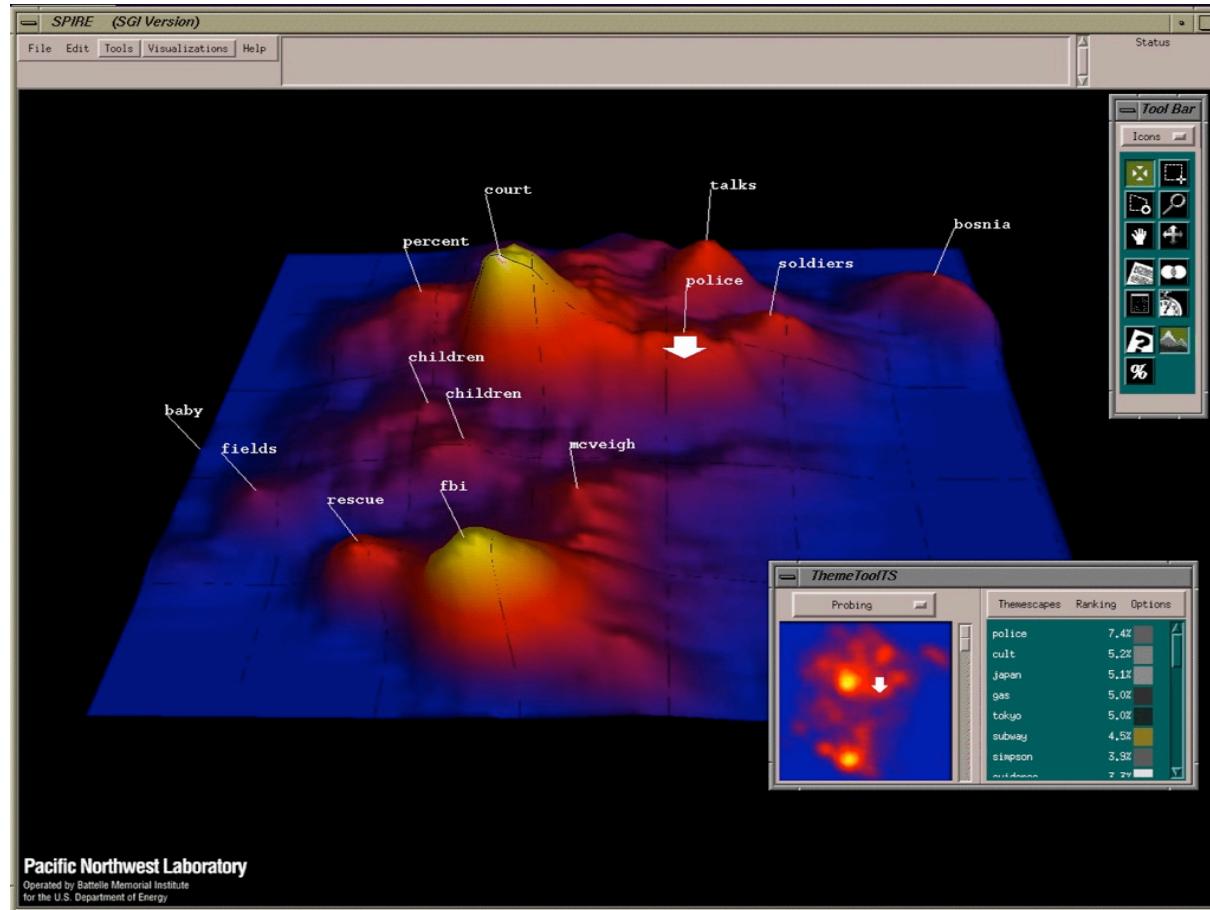
- CAD/CAM, building cars, planes, chips
- Starting to think about: 3D, animation, edu, medicine

# (Incomplete) History of Visualization: 1980s



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

# (Incomplete) History of Visualization: 1990s



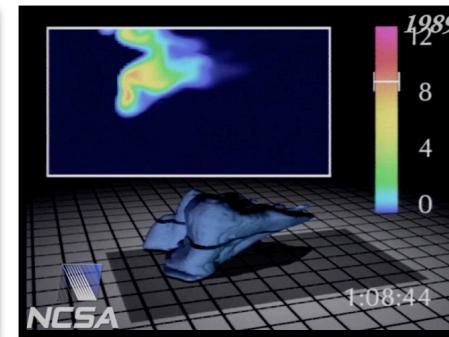
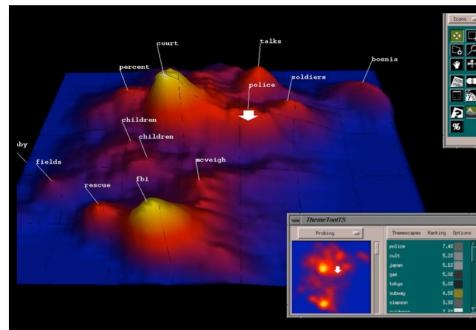
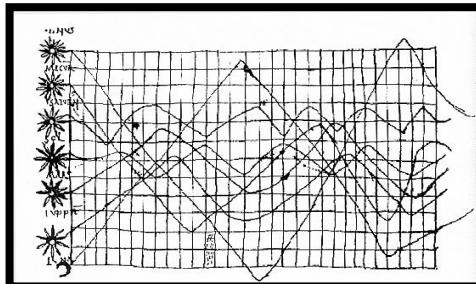
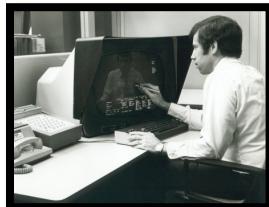
- Information visualization, storytelling
- Starting to think about: online spaces, interaction

## (Incomplete) History of Visualization: 2000s

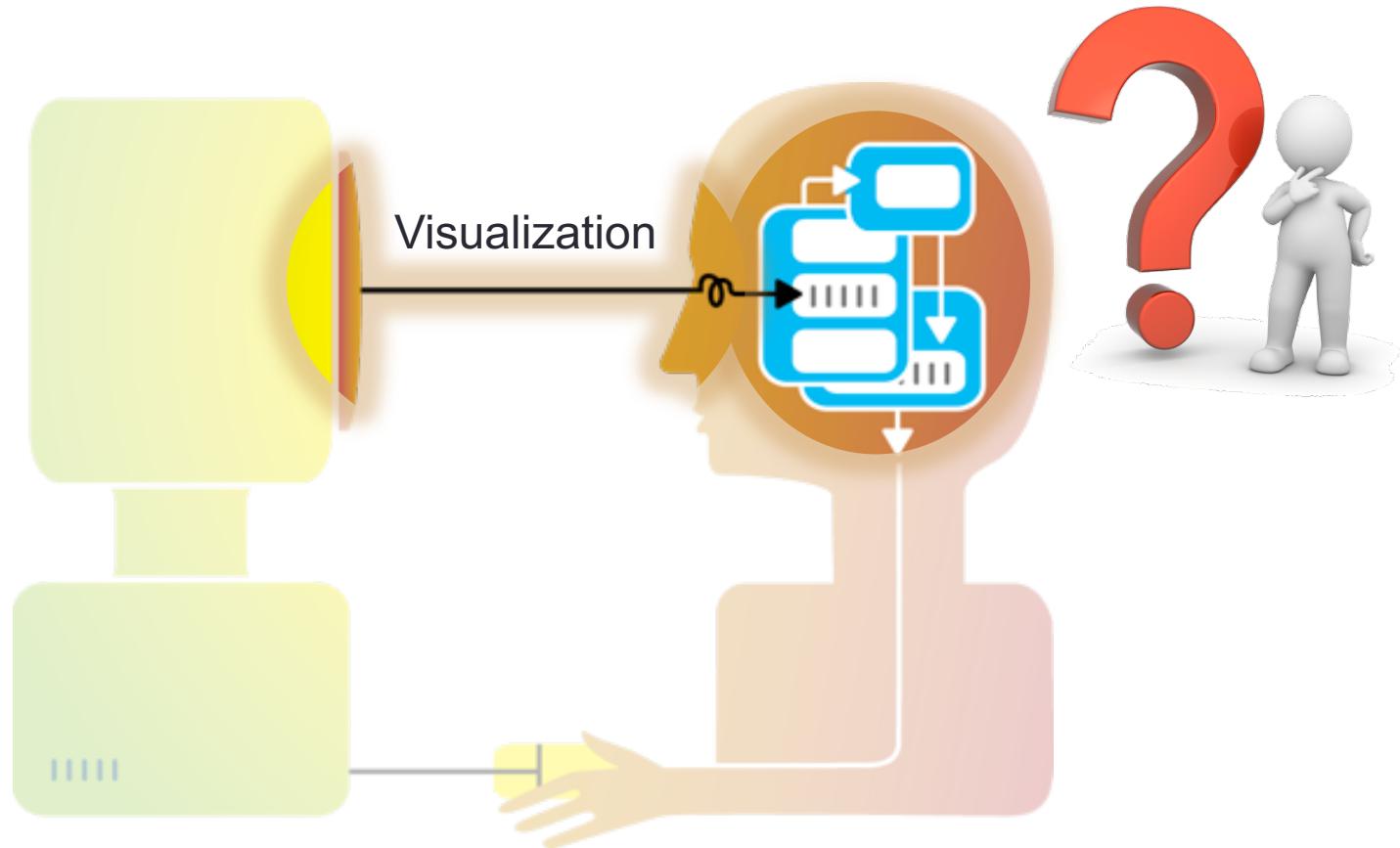


- Coordination across multiple views, interaction
  - Starting to think about: sensemaking, provenance

# 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

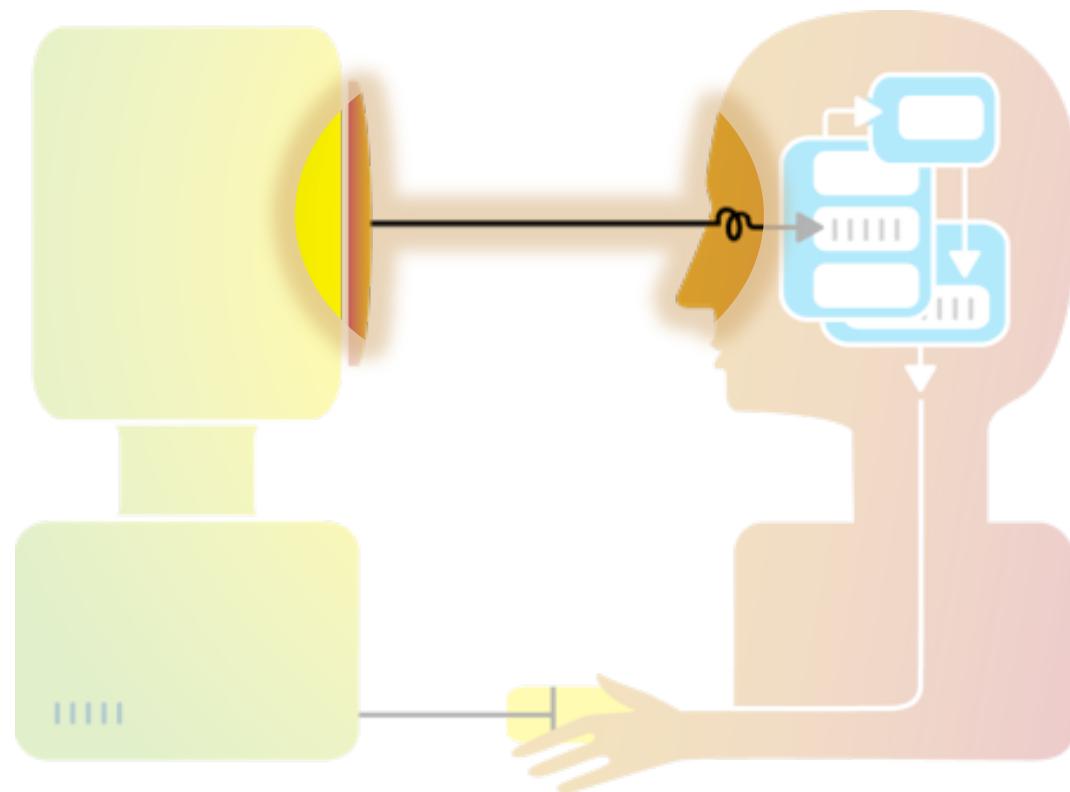
## The bad:

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

## The ugly:

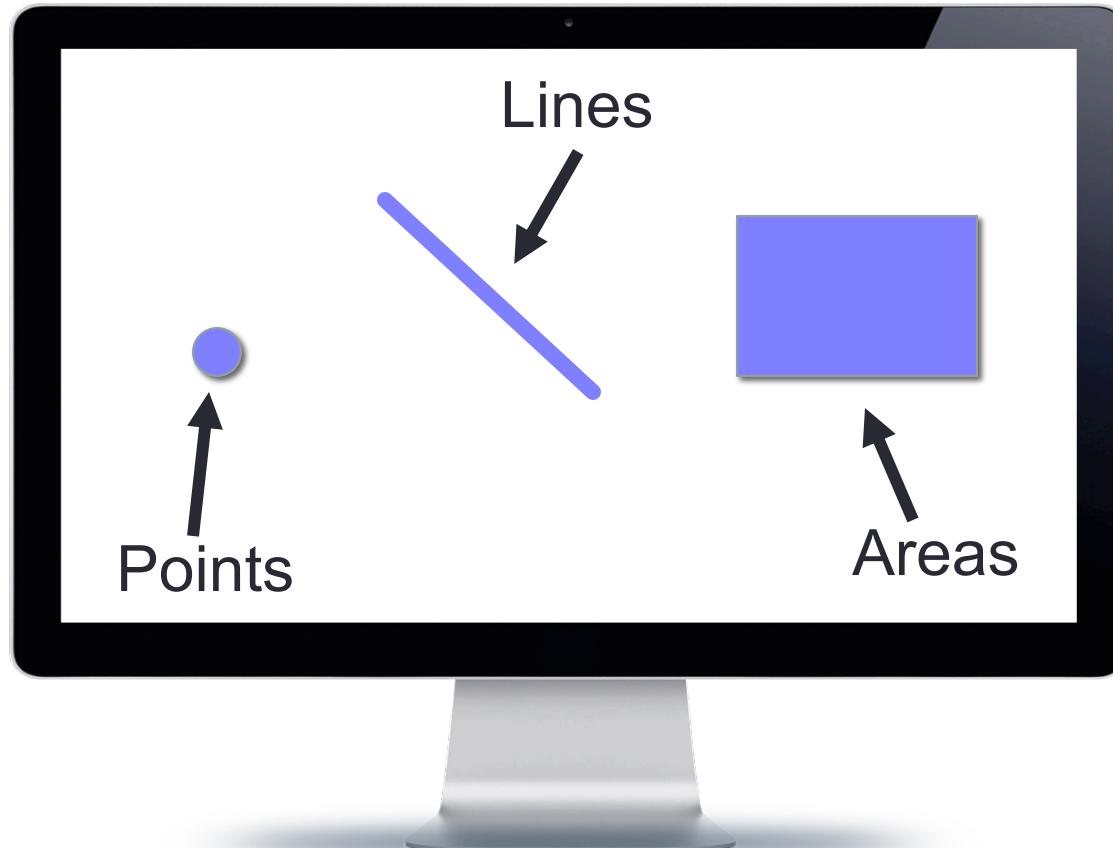
- Mental models are unavoidable: everyone has them, and they’re all different
- **Key:** be aware of how mental models form, how they shape perception, and how to support (or challenge) them

# So what do we have to work with?



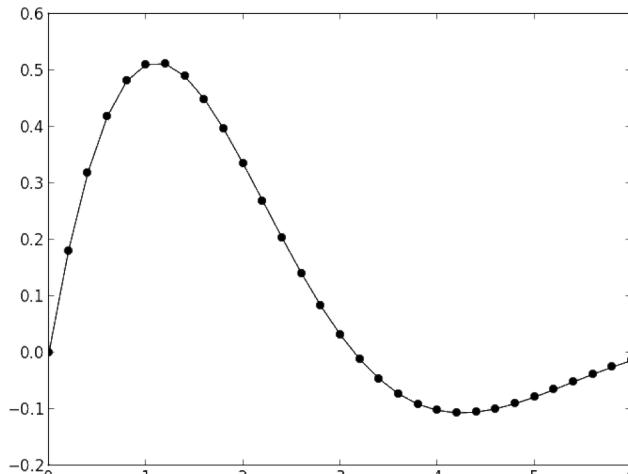
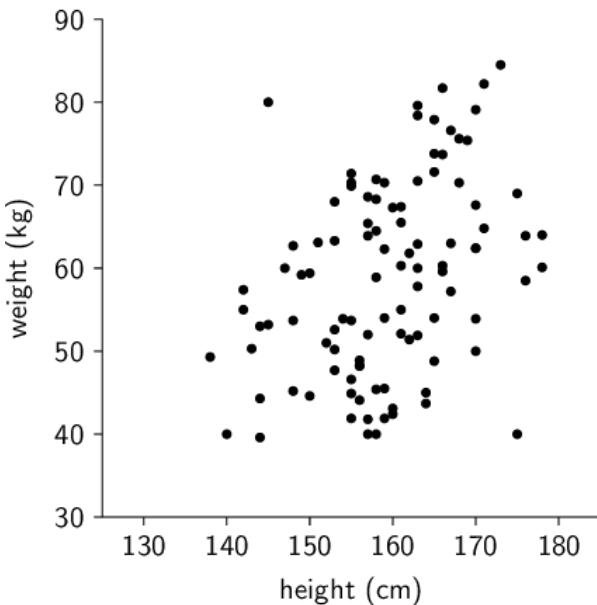
# Graphical primitives

The images we draw are composed of marks: like ink



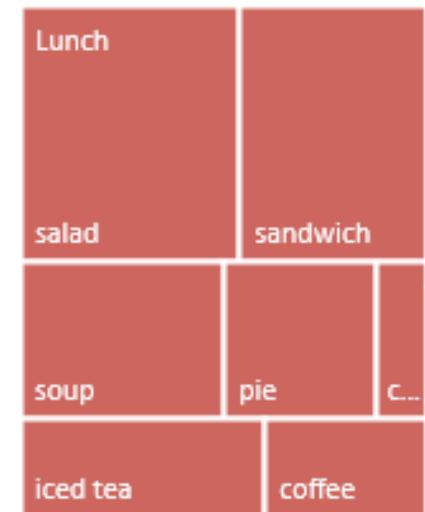
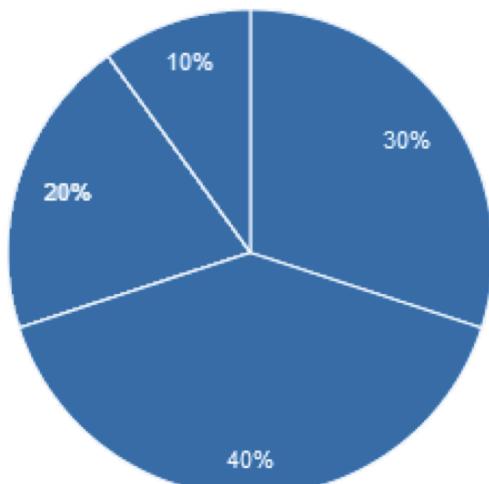
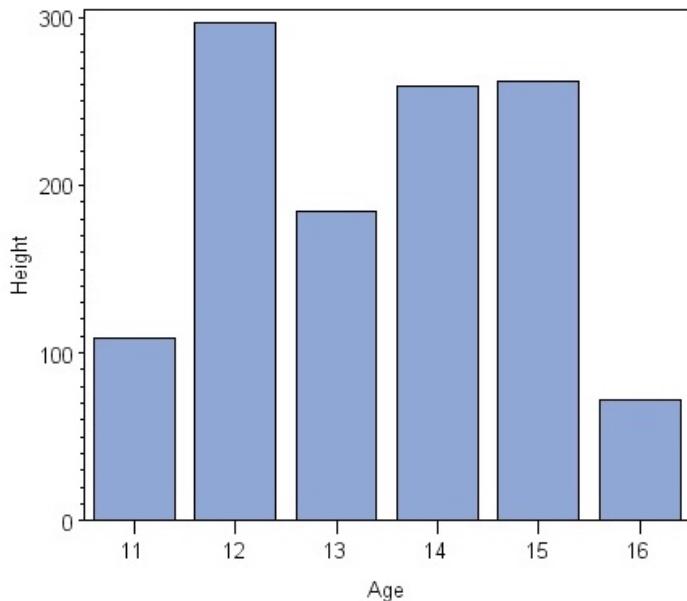
# Visual dimension: position

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



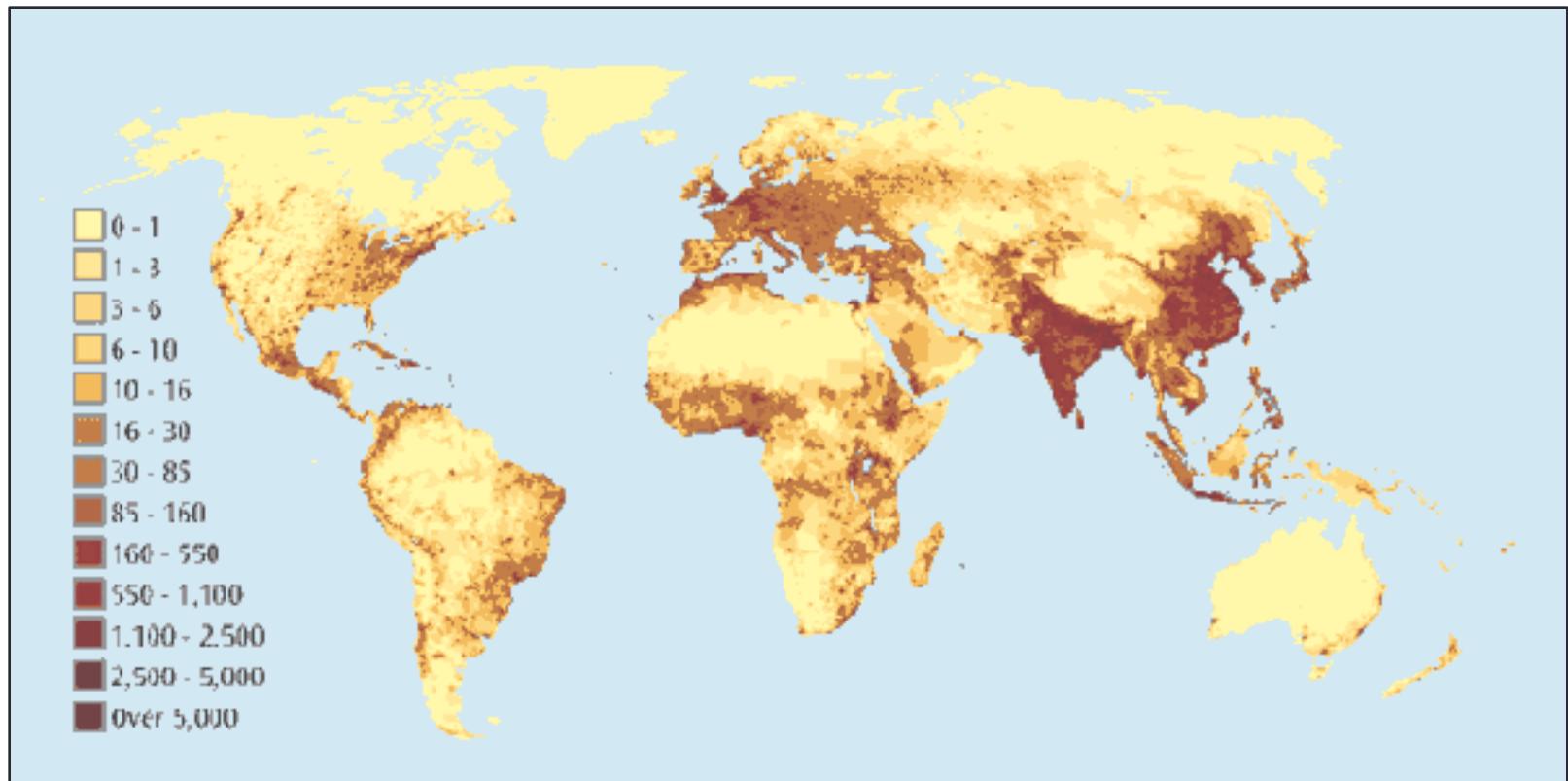
# Visual dimension: size

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



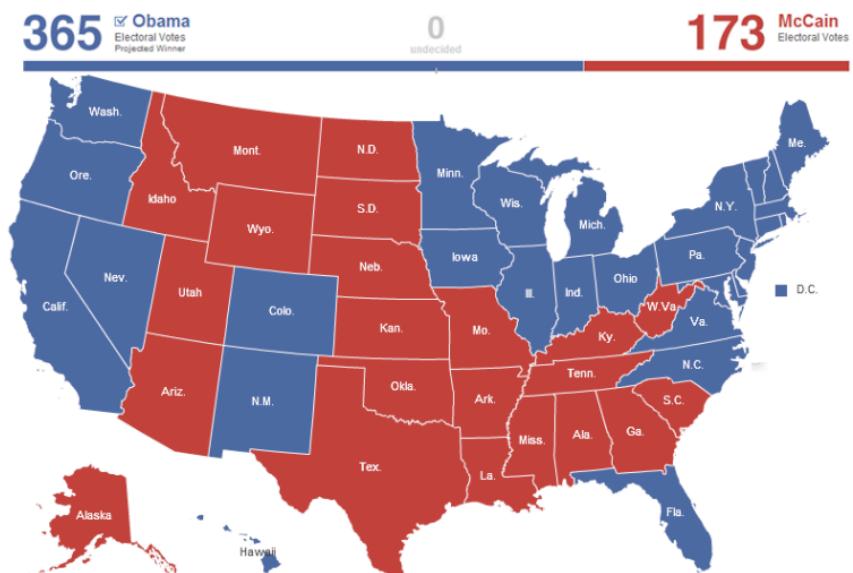
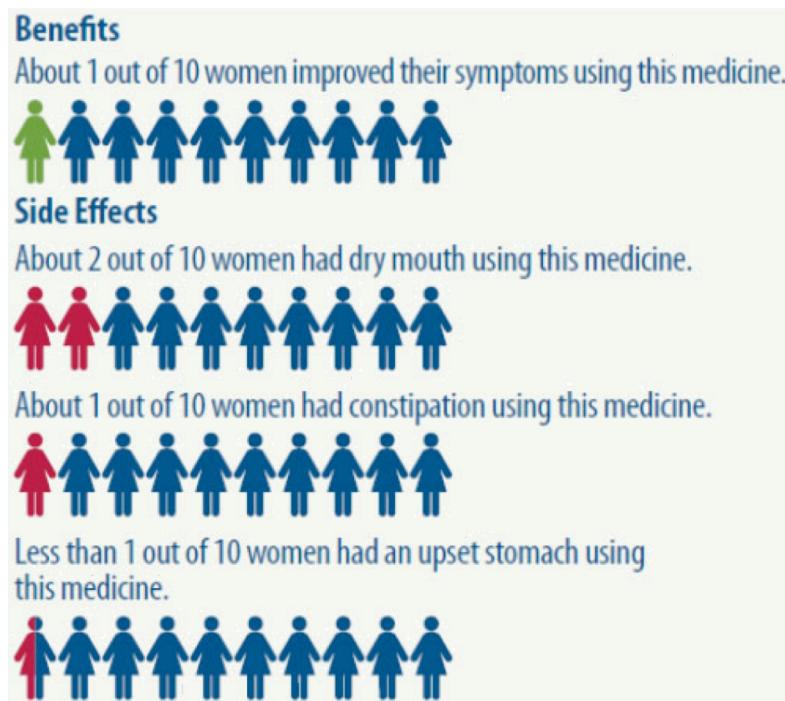
# Visual dimension: value

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



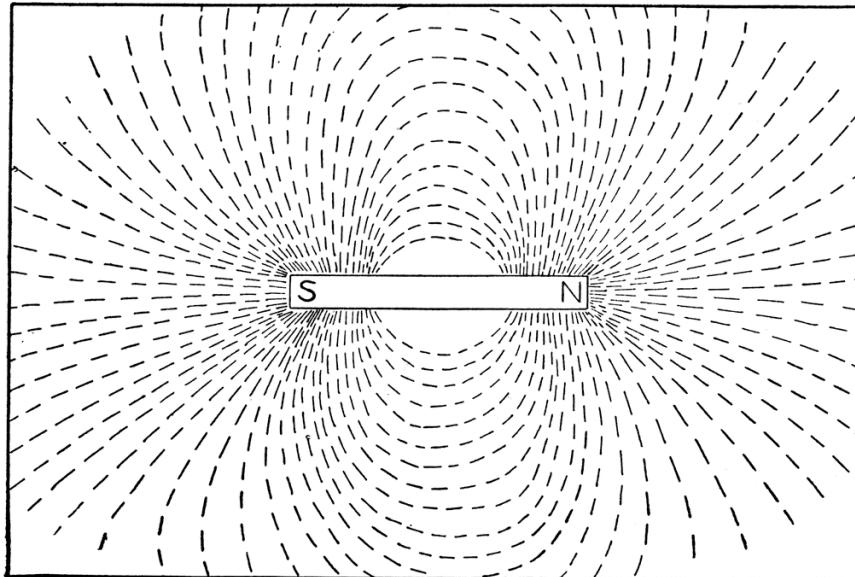
# Visual dimension: color

- 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, *Semioleogie Graphique*  
(Semiology of Graphics), 1967.

# Take-home: 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**?

# After lunch



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