

Visualization for Data Science

CMPT 733

Steven Bergner
sbergner@cs.sfu.ca

Outline

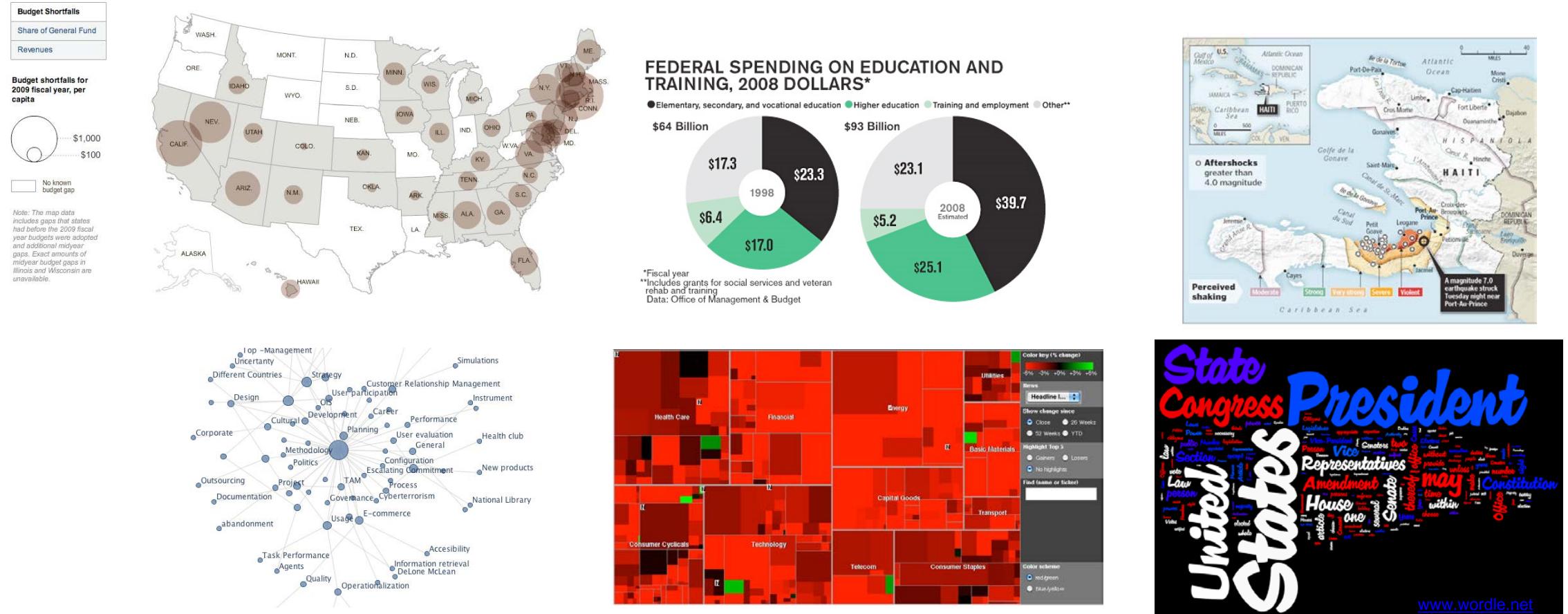
- Visualization: What, Why, and How?
- Motivational examples and goals
- Design principles



vi·su·al·ize

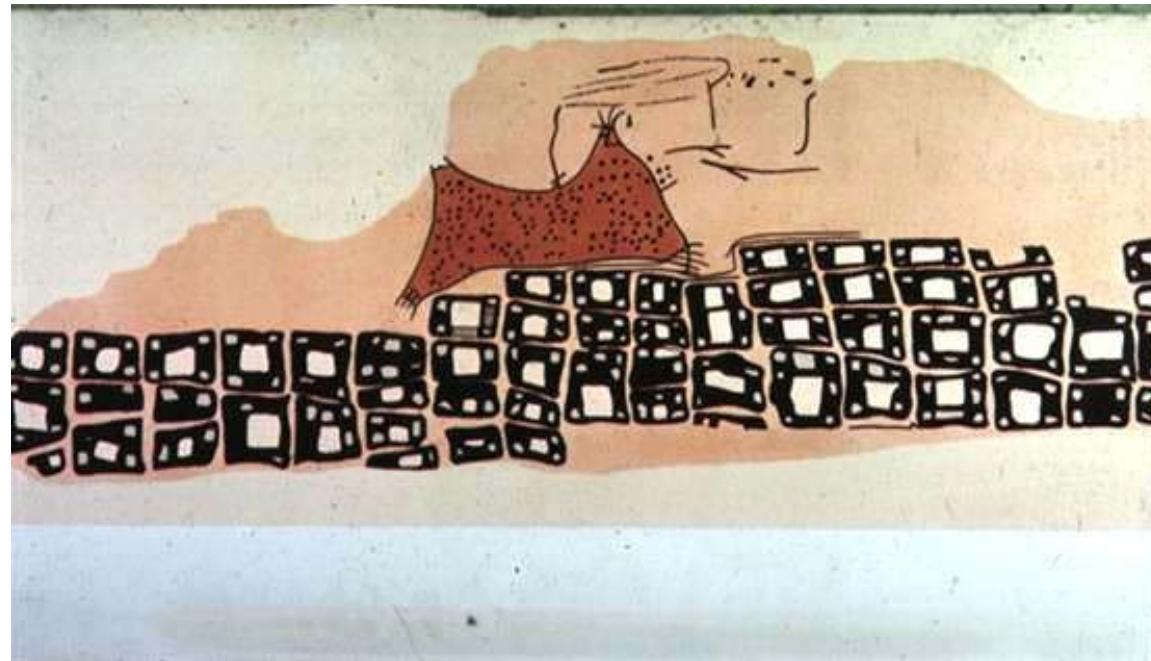
1. To form a mental image of
2. To make visible

Visualization: To convey information through visual representations



Visualization Goals

Map

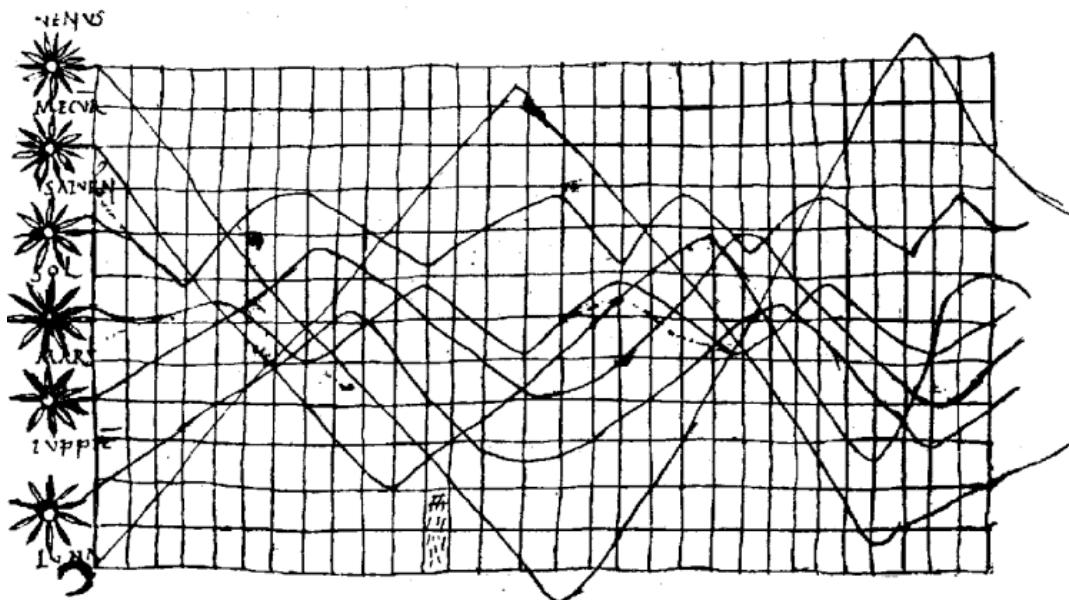


Konya town map, Turkey, c. 6200 BC

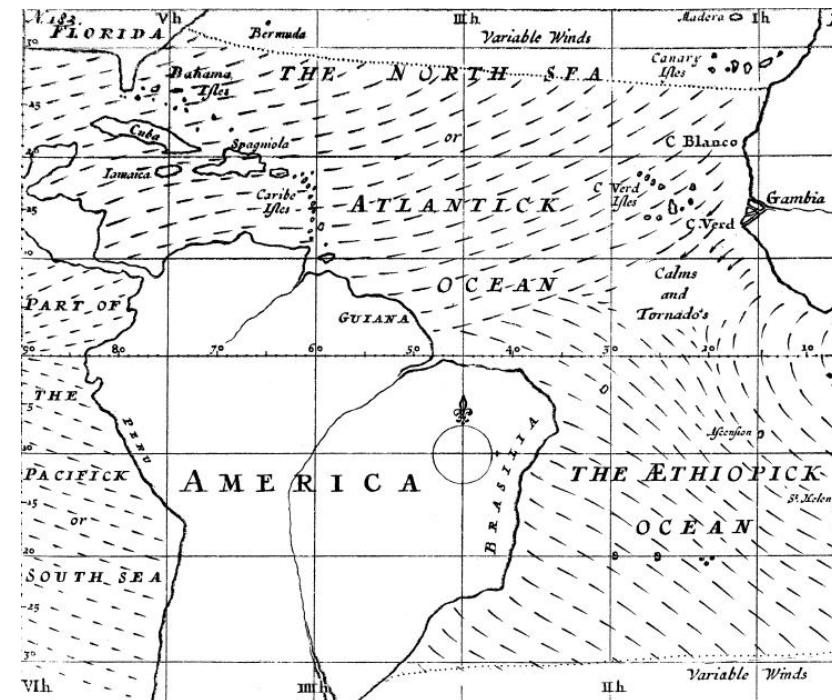


Anaximander's Map of the World
Anaximander of Miletus, c. 550 BC

Map

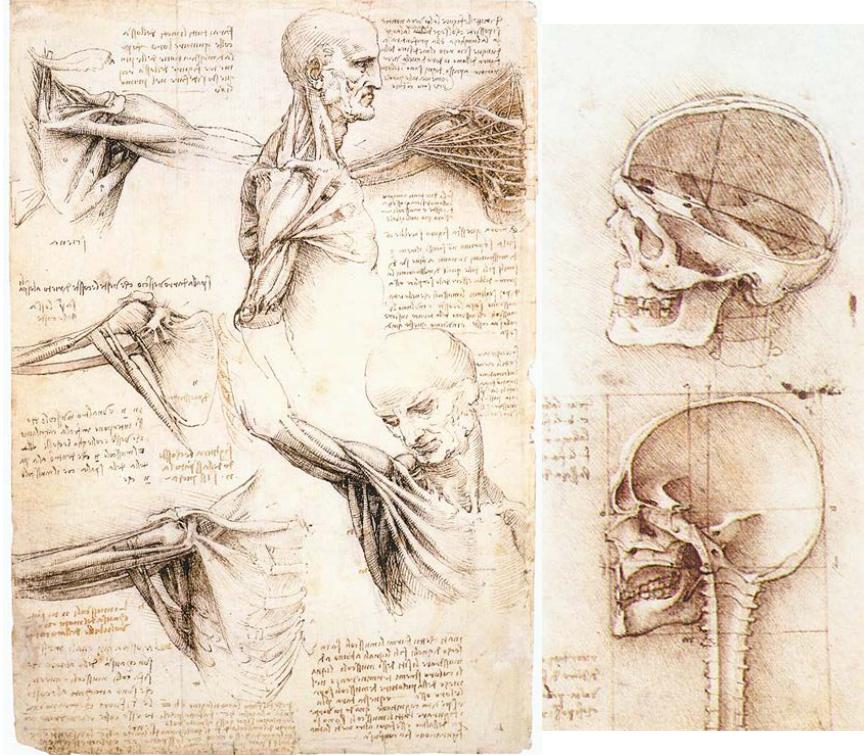


Planetary Movement Diagram, c. 950



Halley's Wind Map, 1686

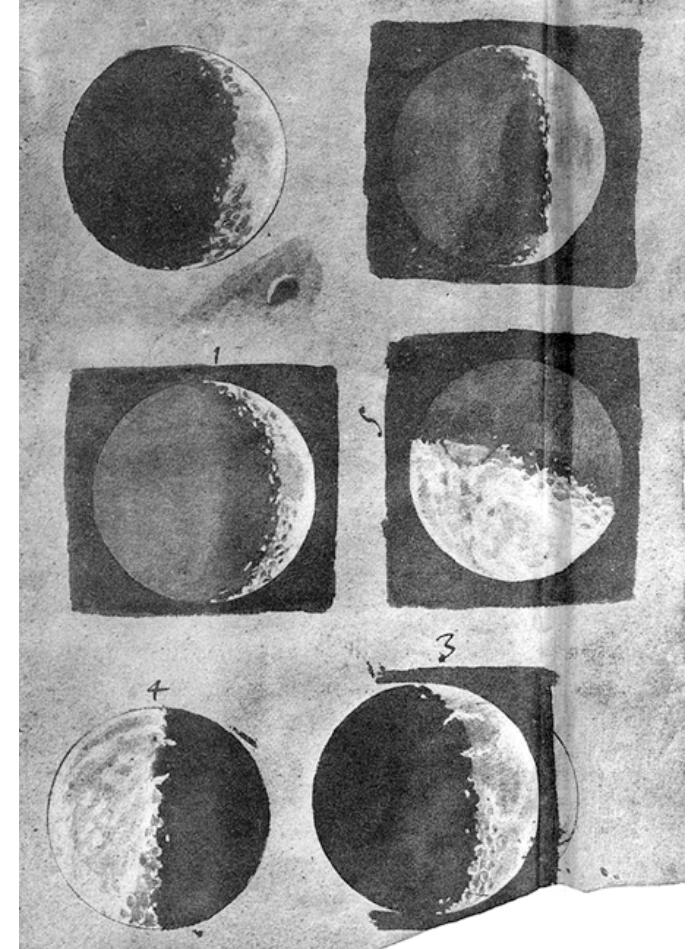
Record



Leonardo Da Vinci, ca. 1500

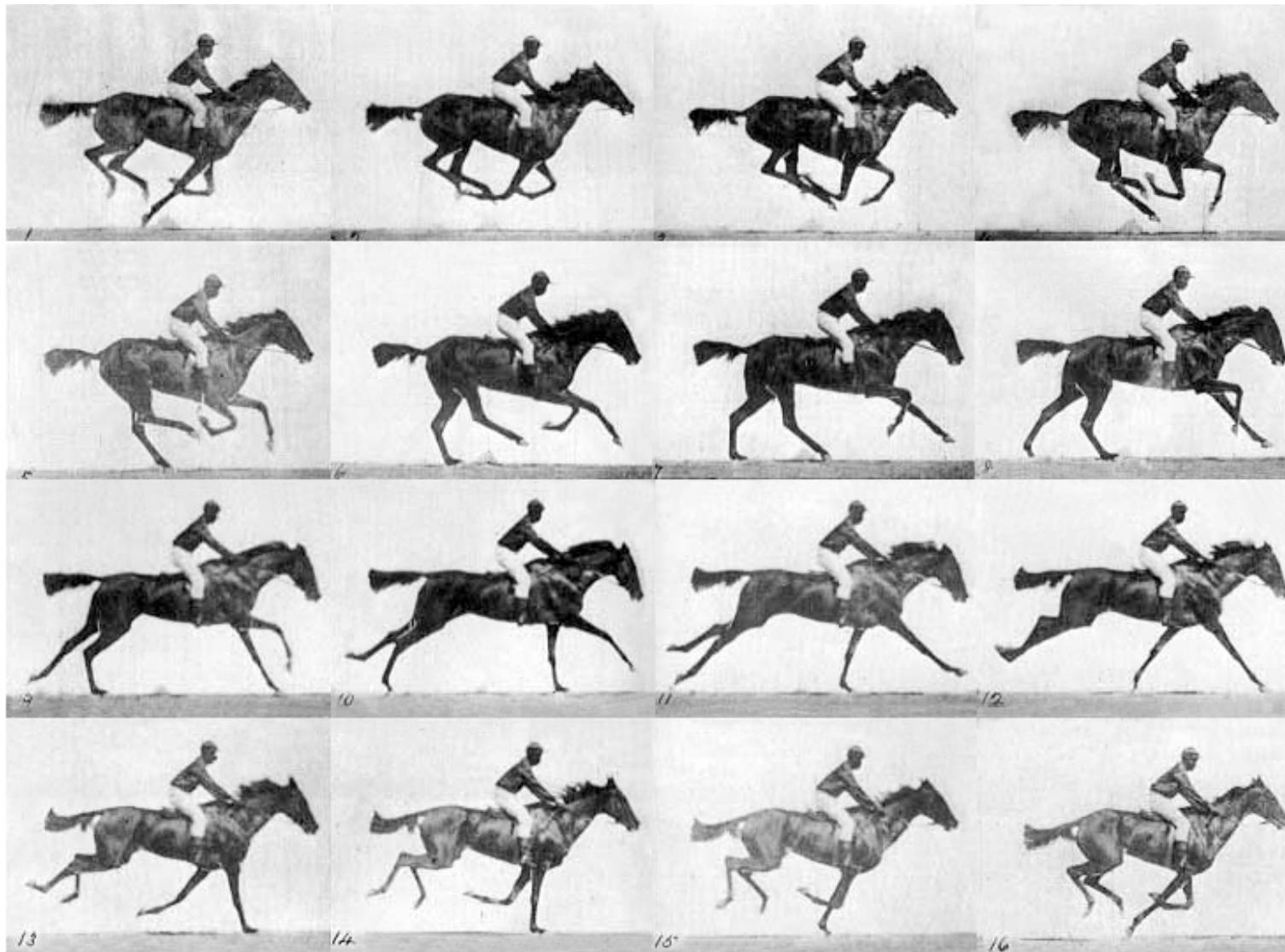


William Curtis (1746-1799)



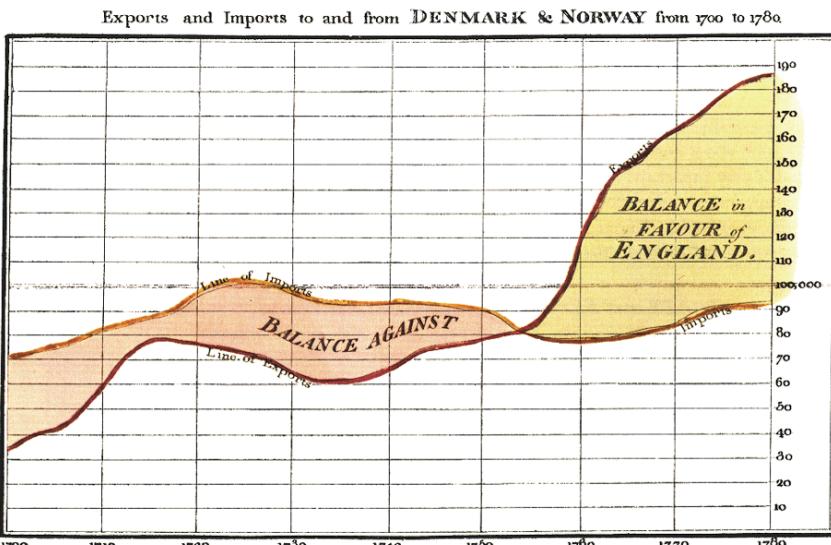
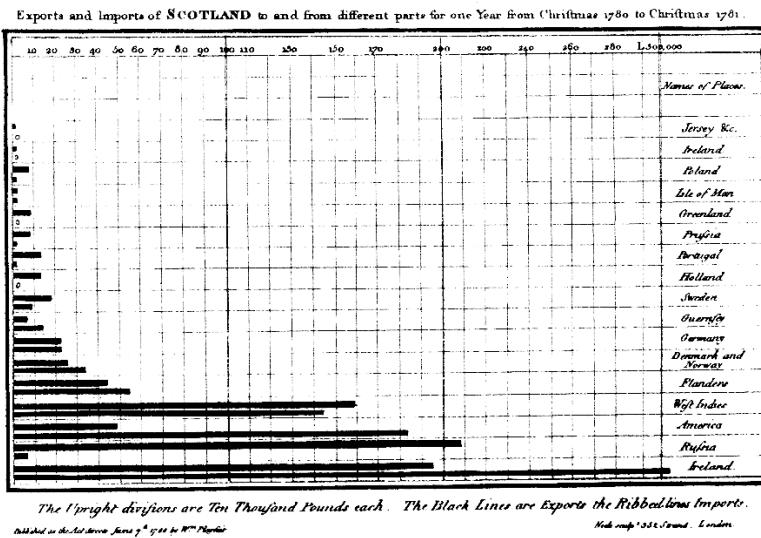
Galileo Galilei, 1616

Record

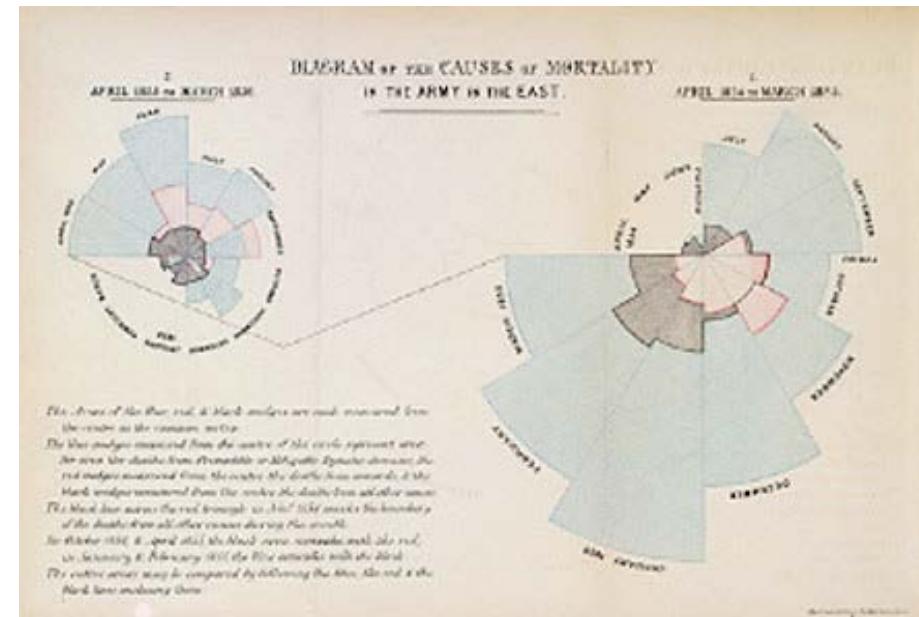


E. J. Muybridge, 1878

Abstract



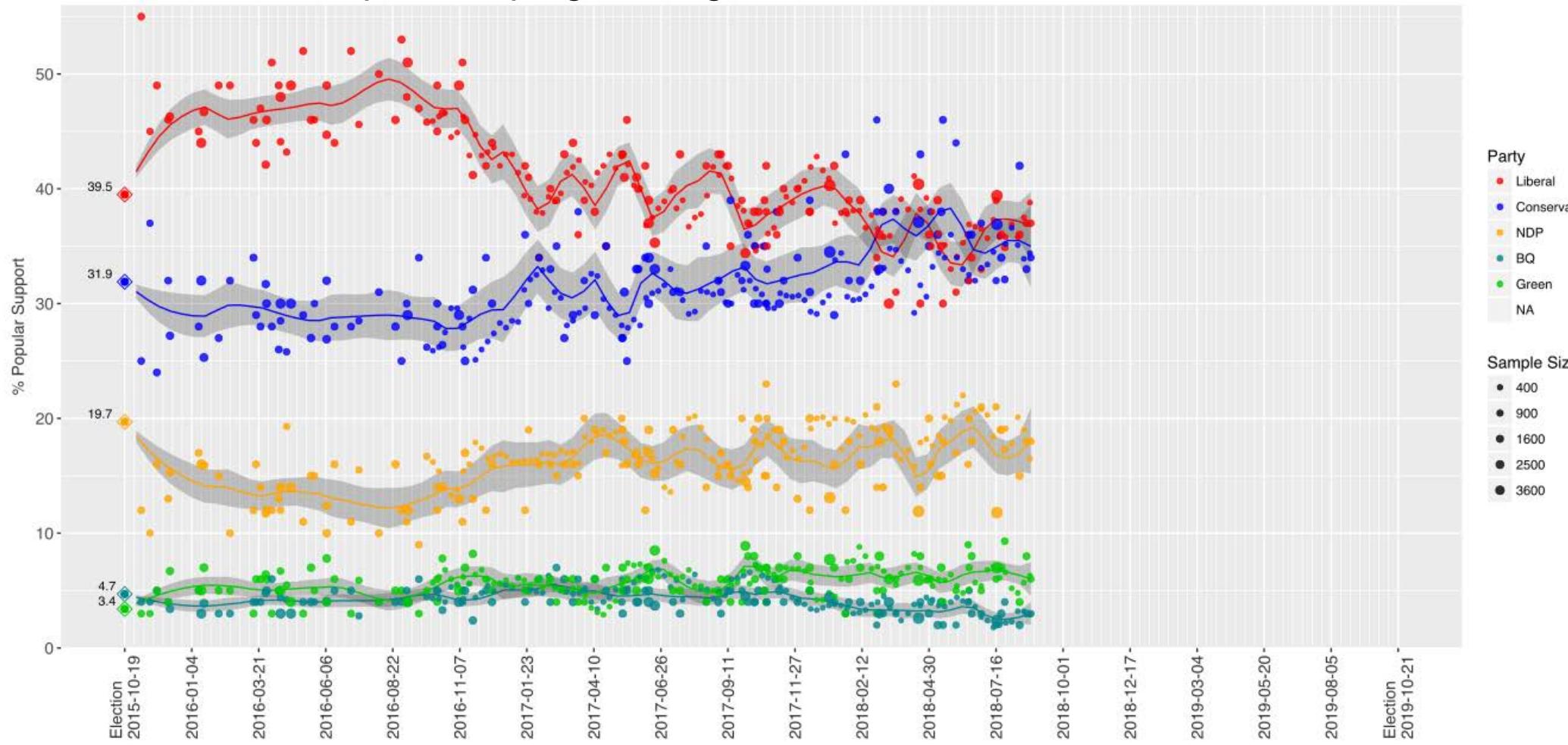
W. Playfair, 1786



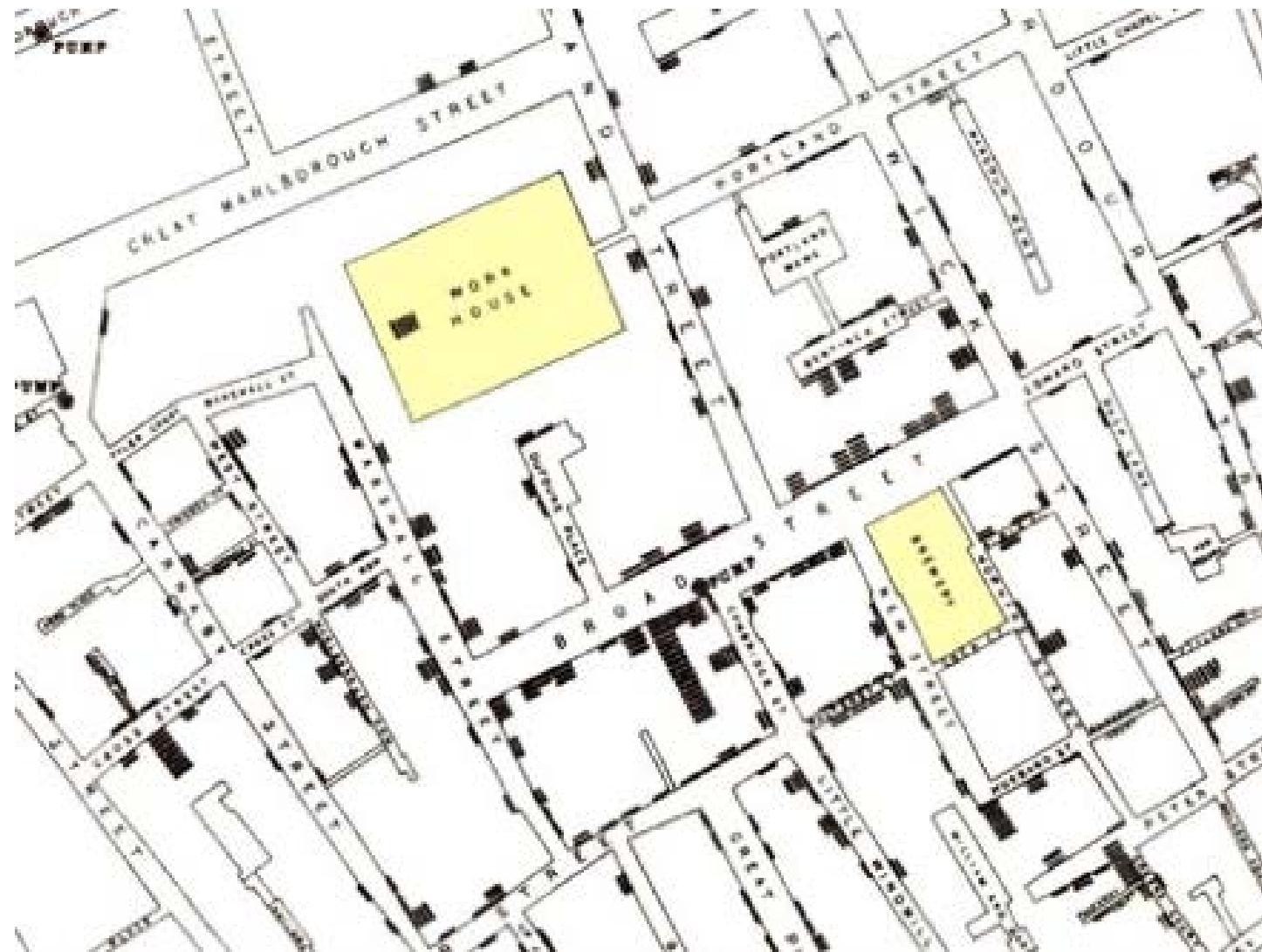
F. Nightingale, 1856

Abstract

Canadian pre-campaign voting intentions for the federal election 2019



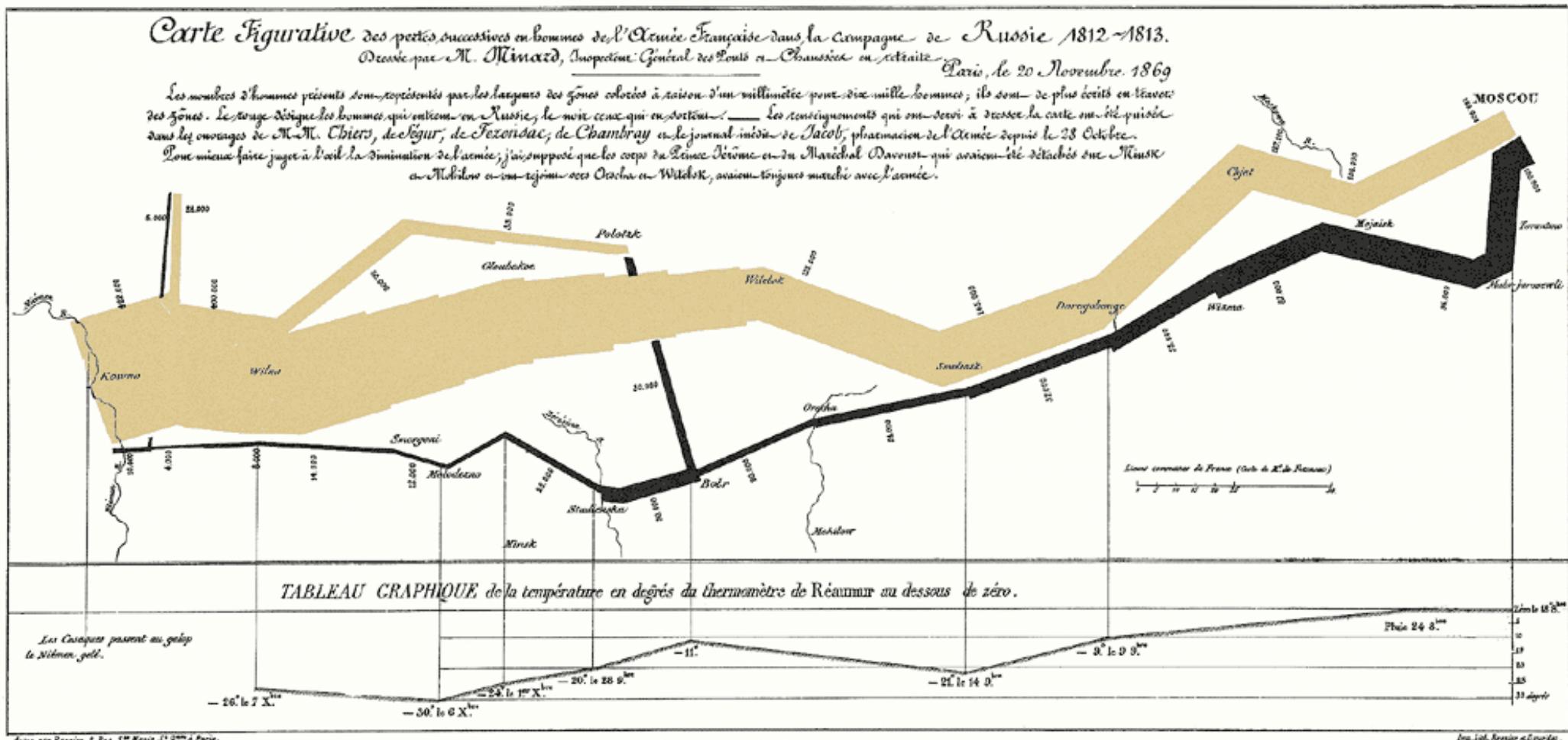
Discover



John Snow, 1854

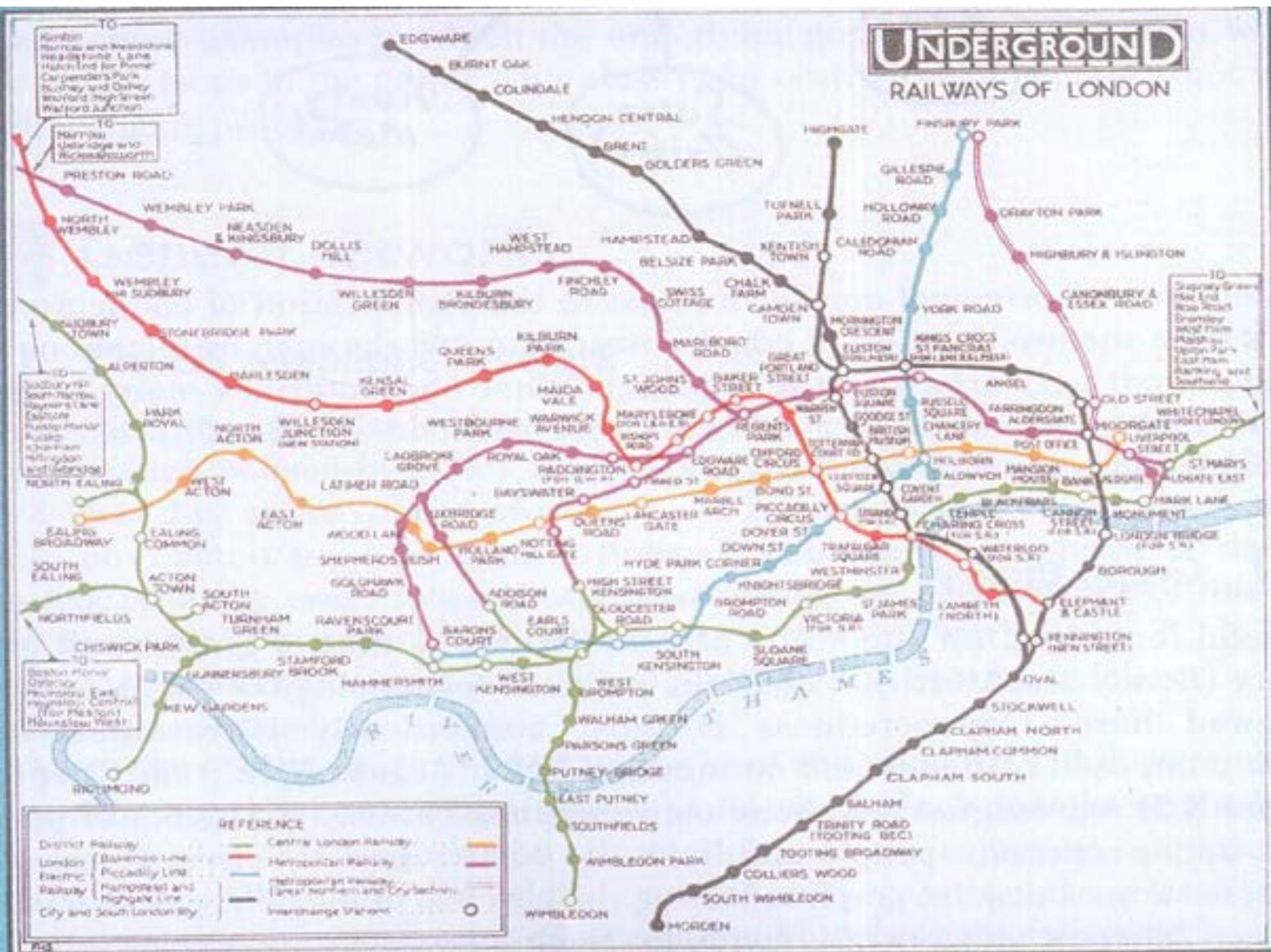
E. Tufte, Visual Explanations, 1997

Discover



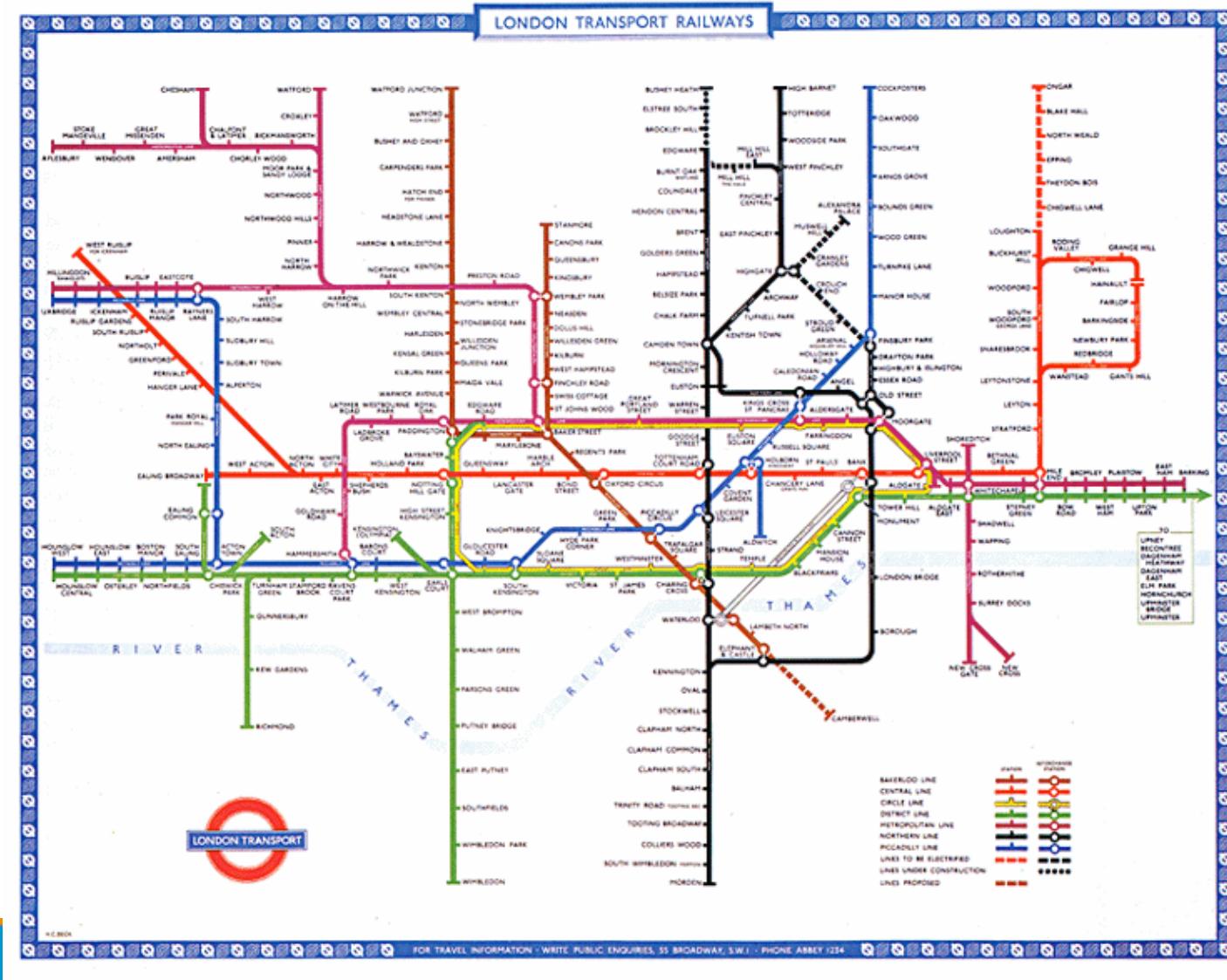
C.J. Minard, 1869

Clarify



London Subway Map, 1927

Clarify

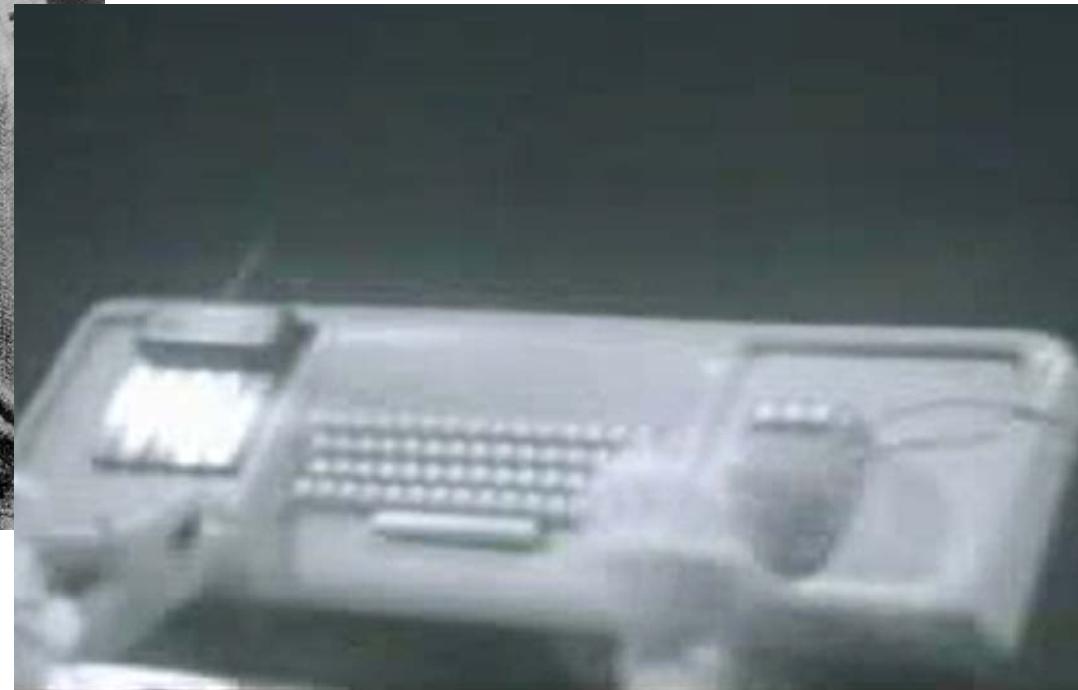


Harry Beck, 1933

Interact



Ivan Sutherland, Sketchpad, 1963



Doug Engelbart, 1968

[play Engelbart.mov]

Interact



M. Wattenberg, 2005

Interact

A Peek Into Netflix Queues

Examine Netflix rental patterns, neighborhood by neighborhood, in a dozen cities. Some titles with distinct patterns are *Mad Men*, *Obsessed* and *Last Chance Harvey*. [Comments \(131\)](#)

100 titles that were frequently rented from Netflix in 2009

[◀ Previous](#)

[Next ▶](#)

Most rented

Least rented

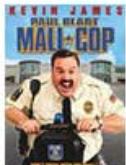
Change how movies are sorted

Most rented

Alphabetical

By metascore

Paul Blart: Mall Cop

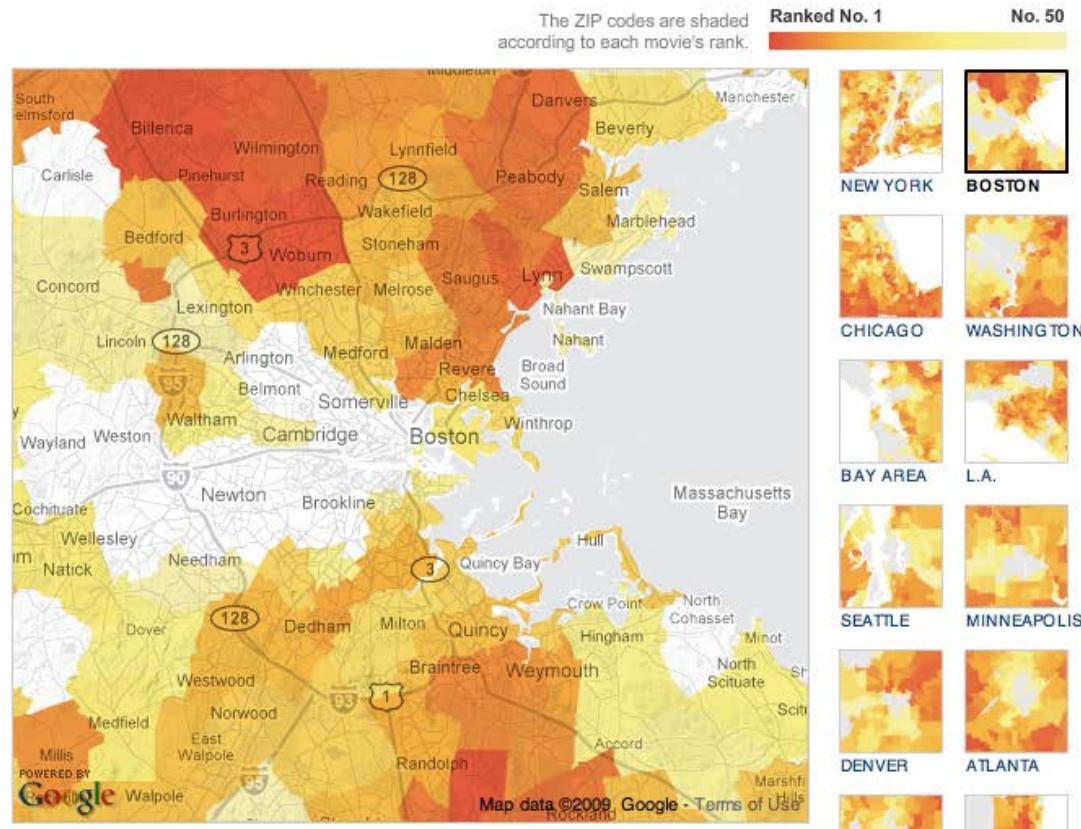


39

Metacritic score

100=loved by critics, 0=hated

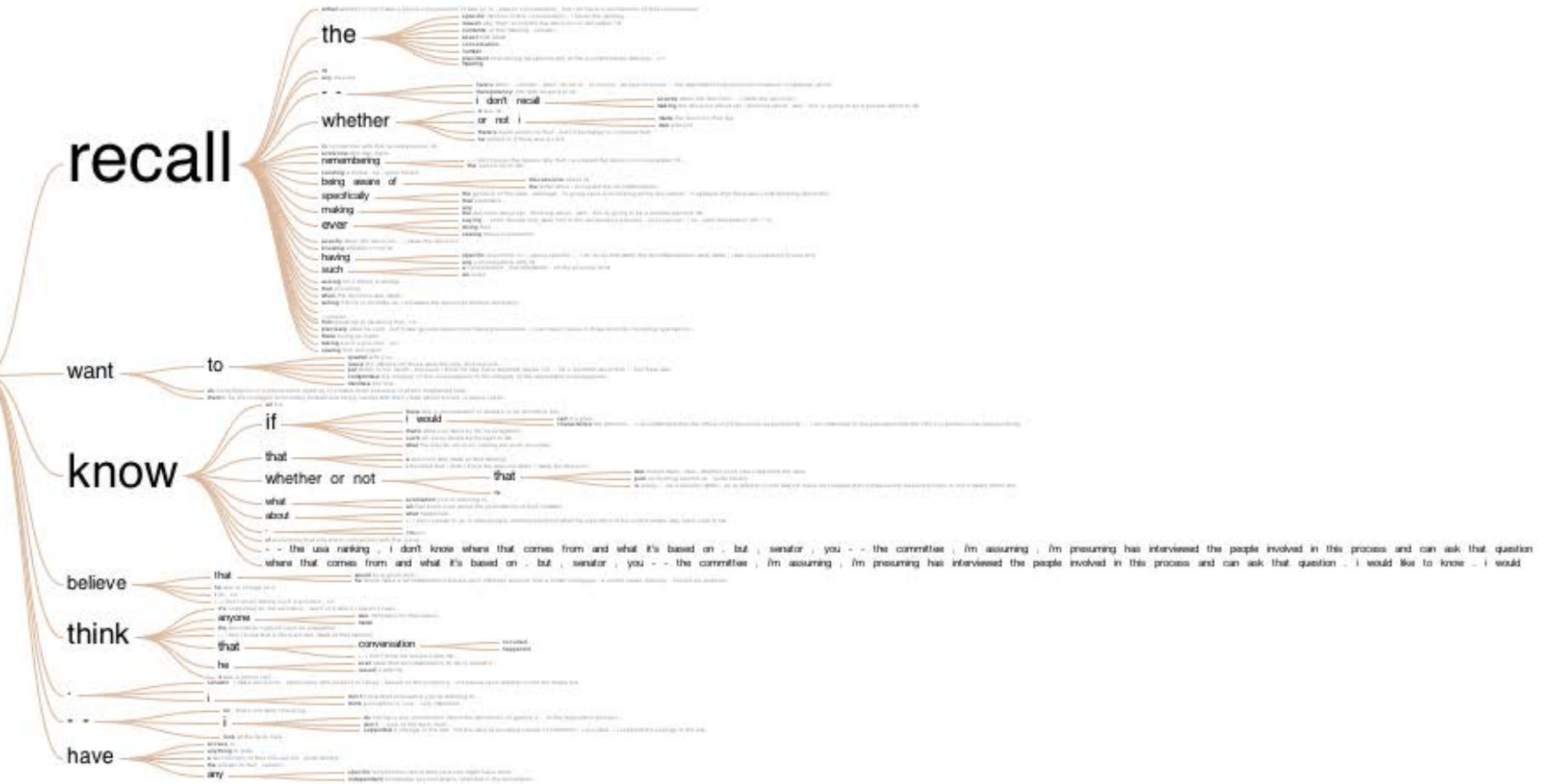
[Read Rest of NYT Review »](#)



Communicate

118
hits

recall
i don't
know



“Many Eyes”, M. Wattenberg 2007

Communicate

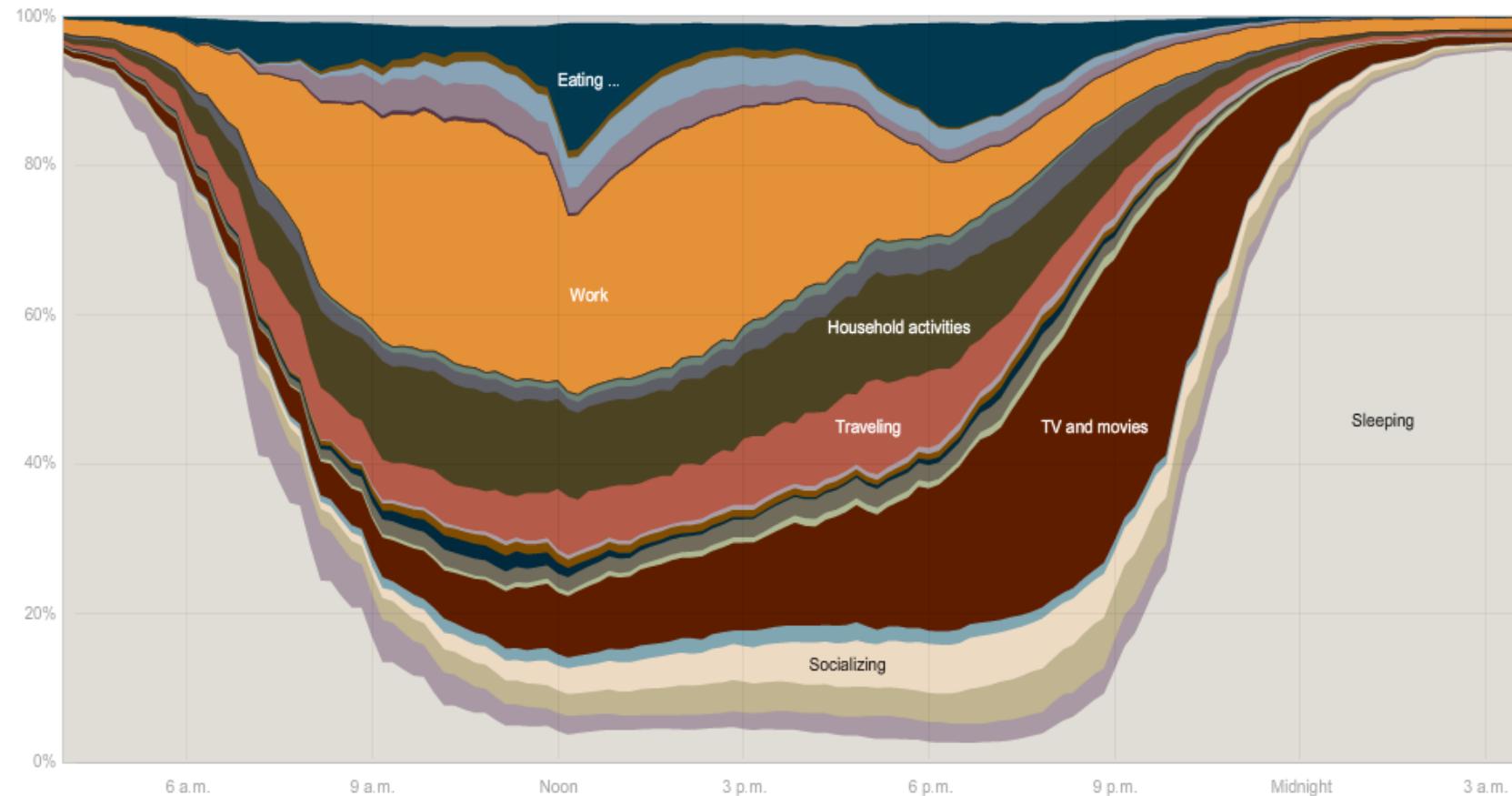
How Different Groups Spend Their Day

The American Time Use Survey asks thousands of American residents to recall every minute of a day. Here is how people over age 15 spent their time in 2008. [Related article](#)

Everyone

Sleeping, eating, working and watching television take up about two-thirds of the average day.

Everyone	Employed	White	Age 15-24	H.S. grads	No children
Men	Unemployed	Black	Age 25-64	Bachelor's	One child
Women	Not in lab...	Hispanic	Age 65+	Advanced	Two+ children



Inspire / Tell a Story



Hans Rosling, TED 2006

Visualization

- To convey information through visual representations

Map

Record

Abstract

Discover

Clarify

Interact

Communicate

Inspire

Goals

- Insight and analysis
 - Extract the information content
 - Make things and relationships visible
 - Analyze the data by means of the visual representation
- Communication
 - Allow the non-expert to understand
 - Guide the expert into the right direction
- Exploration
 - Interactive control
 - Use visual representation to understand the phenomena
- “The purpose of computing is insight not numbers”
(Hamming 1962)

What is Visualization?

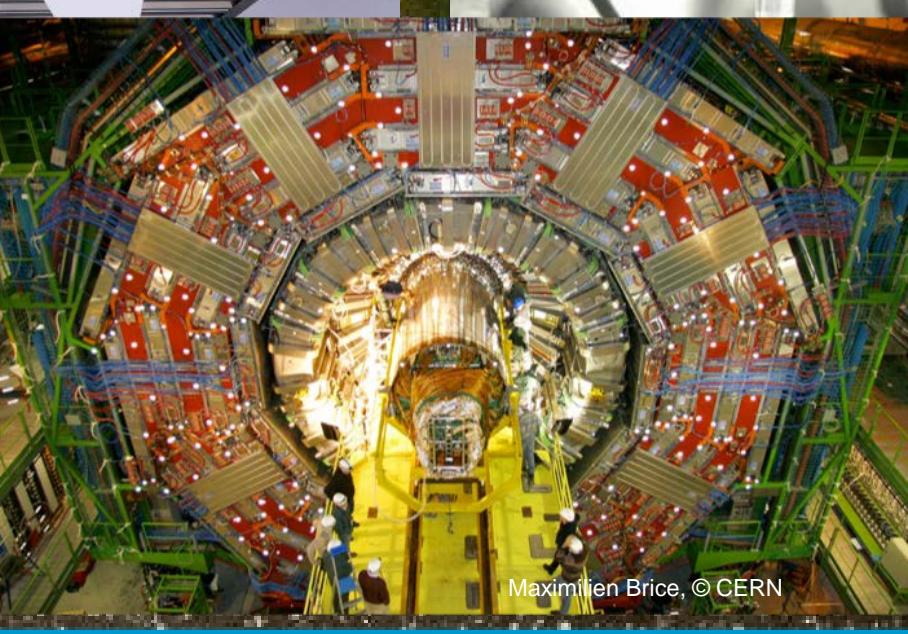
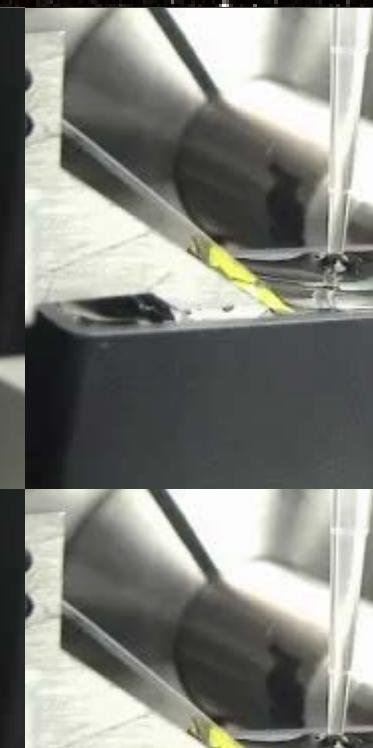
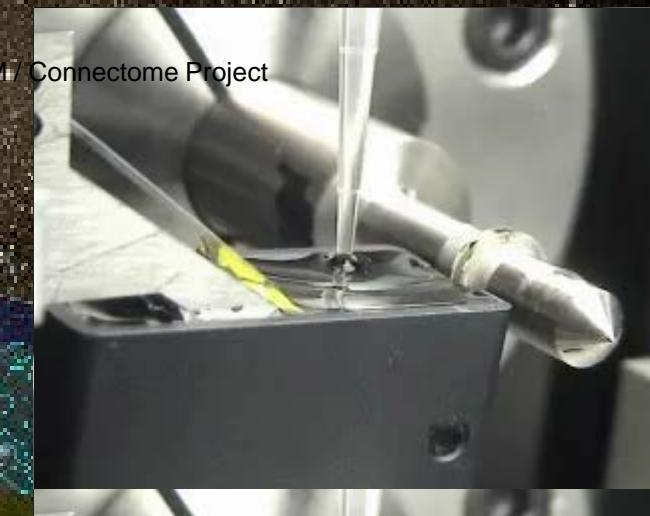
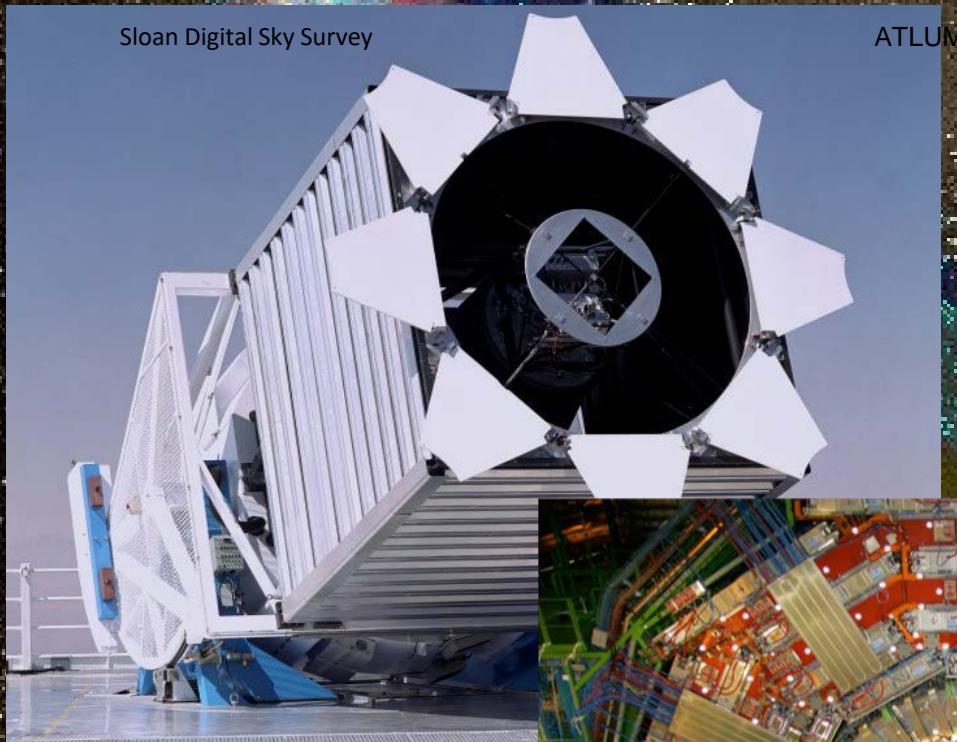
- What?
- Why?
- Who?
- How?

Information Explosion / Big Data

The collage consists of five screenshots arranged in a grid-like fashion:

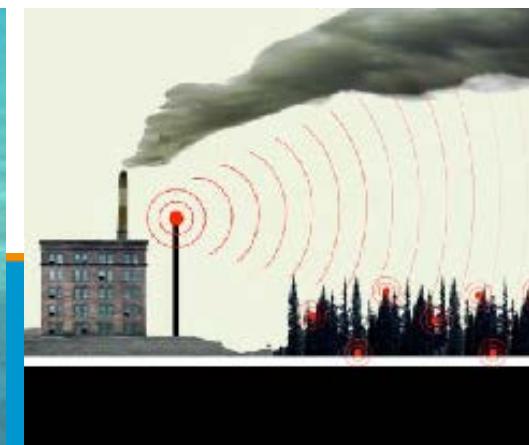
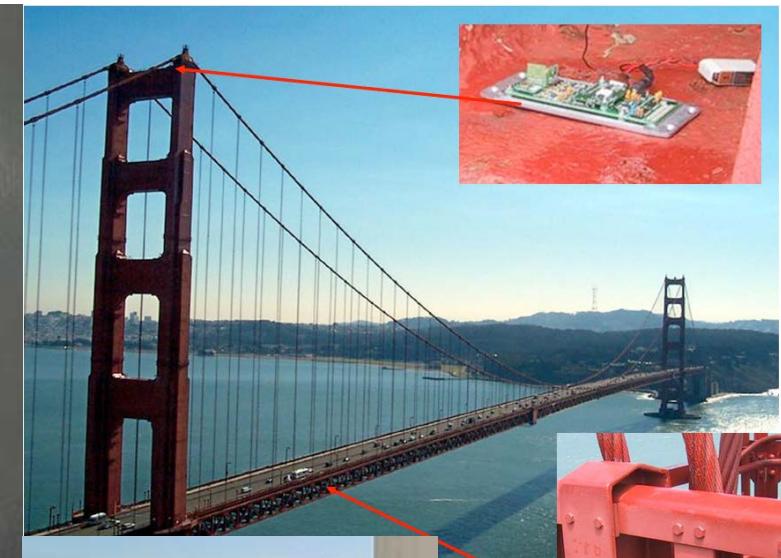
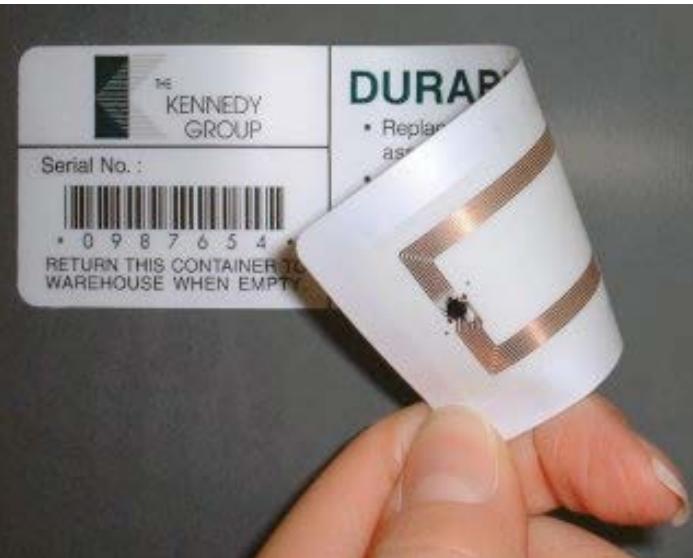
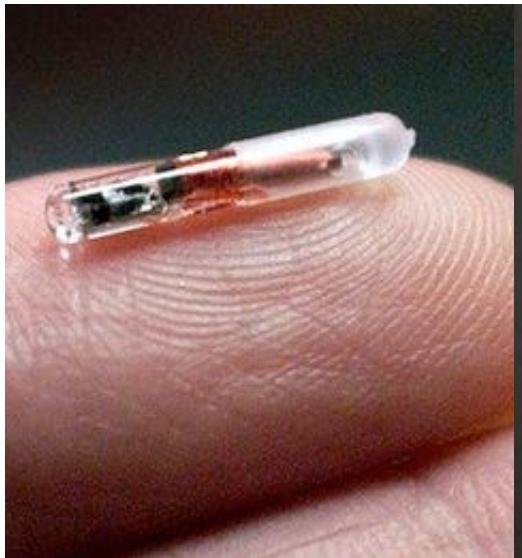
- Google Reader (1000+)**: Shows a screenshot of Google Reader with over 1000 items in the inbox, displaying various news articles and blog posts.
- Digg / All News, Videos, & Images**: Shows the Digg homepage with a large list of news items, videos, and images, including a prominent story about Barack Obama's win in South Carolina.
- Wikipedia**: Shows the Wikipedia homepage in English, Deutsch, Polski, Nederlands, and Português, highlighting the collaborative nature of the encyclopedia.
- facebook**: Shows a Facebook group page for "Barack Obama for President in 2008" with a large image of Barack Obama and many user posts.
- twitter**: Shows the Twitter profile of hpfister (@hpfister) with 140 tweets, displaying a stream of tweets from users like guykawasaki and timoreilly.

Instrument Data Explosion



“The Industrial Revolution of Data”

Joe Hellerstein, UC Berkeley



Limits of Cognition



Daniel J. Simons and Daniel T. Levin, Failure to detect changes to people during a real world interaction, 1998

“It is things that make us smart.”

Donald Norman



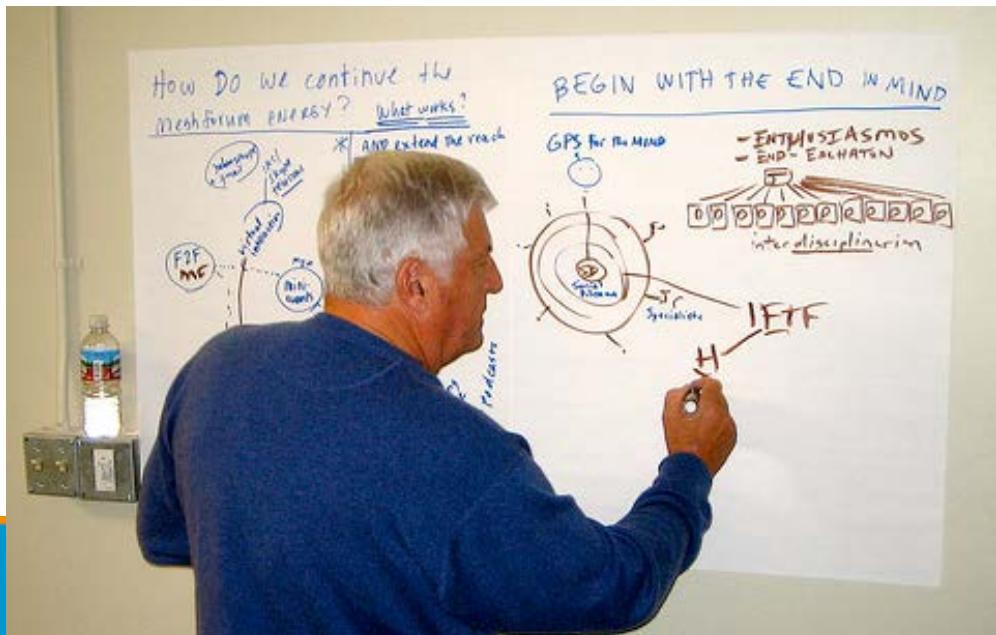
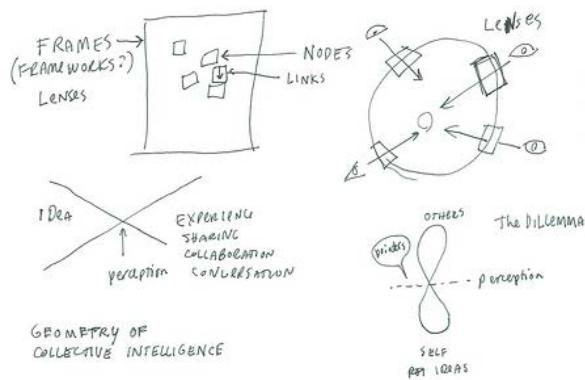
“It is things that make us smart.”

Donald Norman



“It is things that make us smart.”

Donald Norman



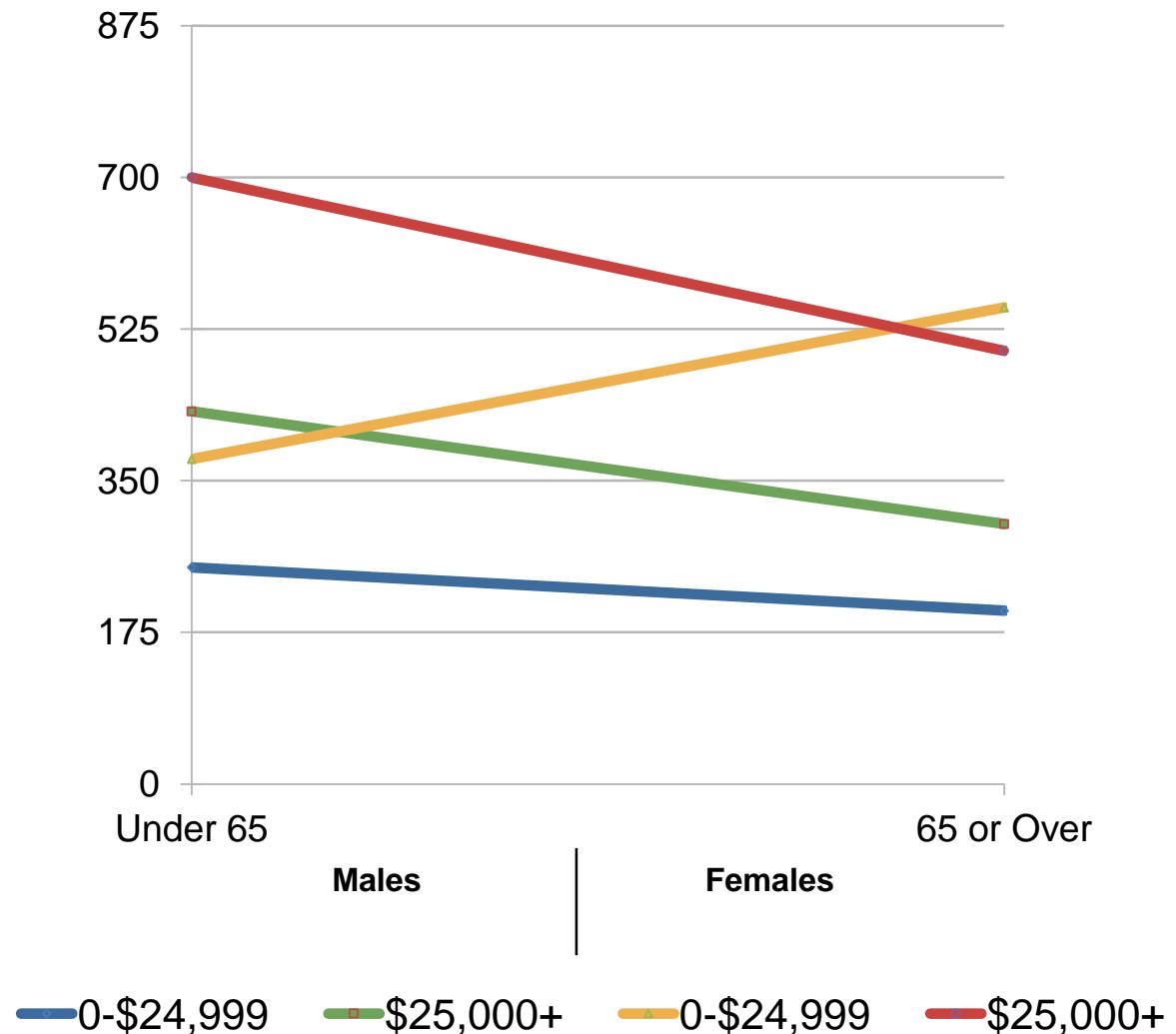
Visual Thinking Collection, Dave Grey

Mental Queries

Which gender or income level group shows different effects of age on triglyceride levels?

	Males		Females	
Income Group	Under 65	65 or Over	Under 65	65 or Over
0-\$24,999	250	200	375	550
\$25,000+	430	300	700	500

Visual Queries

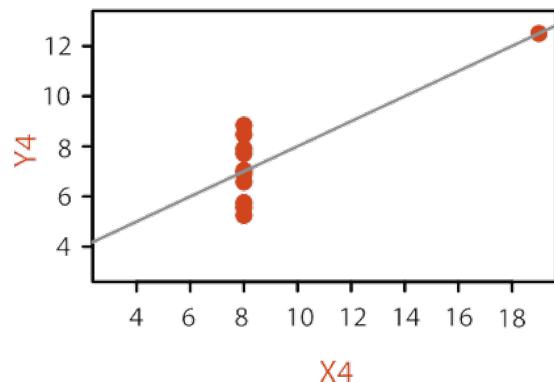
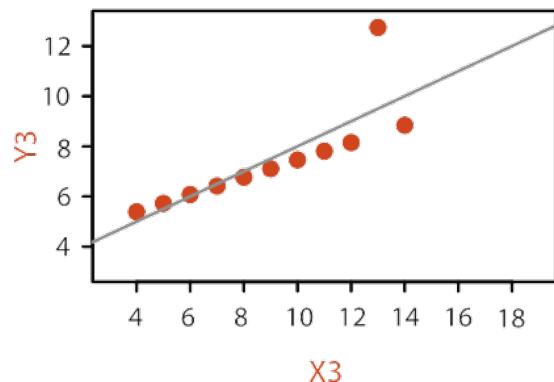
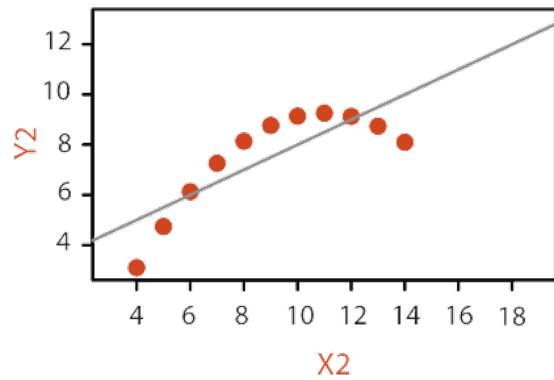
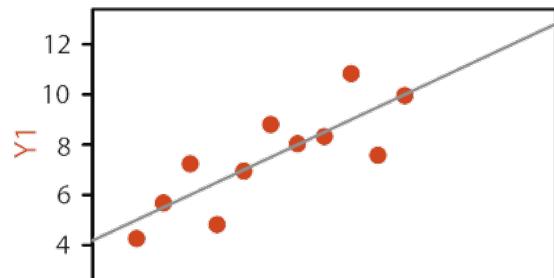


Visualization

- Helps us think
- Reduces load on working memory
- Offloads cognition
- Uses the power of human perception

Why use an external representation?

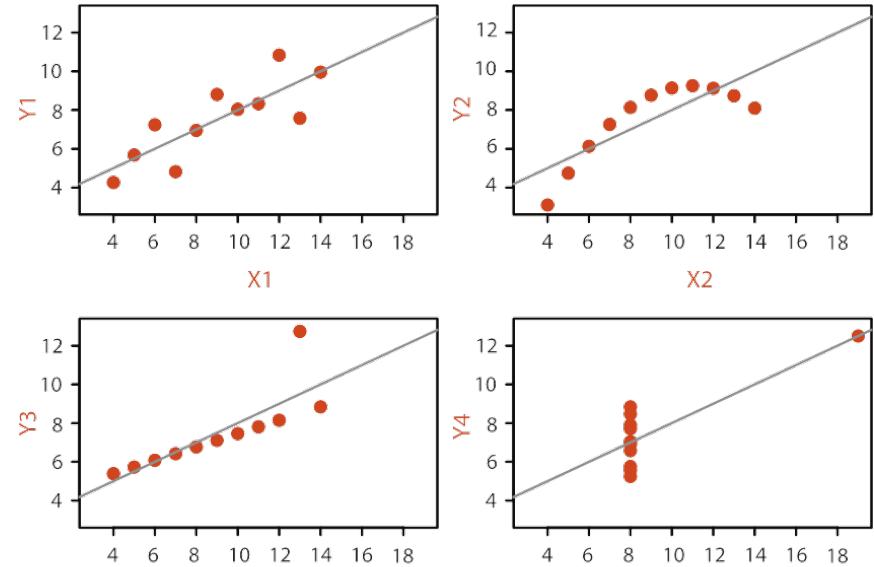
- Replace cognition with perception



	I	II	III	IV	
x	x	x	x	x	
y	y	y	y	y	
10	8,04	10	9,14	10	7,46
8	6,95	8	8,14	8	6,77
13	7,58	13	8,74	13	12,74
9	8,81	9	8,77	9	7,11
11	8,33	11	9,26	11	7,81
14	9,96	14	8,1	14	8,84
6	7,24	6	6,13	6	6,08
4	4,26	4	3,1	4	5,39
12	10,84	12	9,13	12	8,15
7	4,82	7	7,26	7	6,42
5	5,68	5	4,74	5	5,73
SUM	99,00	82,51	99,00	82,51	99,00
AVG	9,00	7,50	9,00	7,50	9,00
STDEV	3,32	2,03	3,32	2,03	3,32

[F. J. Anscombe, 1973]

Why represent all the data?



- Summaries lose information, details matter
 - Confirm expected and find unexpected patterns
 - Assess validity of statistical model

	I		II		III		IV	
	x	y	x	y	x	y	x	y
10	8,04	10	9,14	10	7,46	8	6,58	
8	6,95	8	8,14	8	6,77	8	5,76	
13	7,58	13	8,74	13	12,74	8	7,71	
9	8,81	9	8,77	9	7,11	8	8,84	
11	8,33	11	9,26	11	7,81	8	8,47	
14	9,96	14	8,1	14	8,84	8	7,04	
6	7,24	6	6,13	6	6,08	8	5,25	
4	4,26	4	3,1	4	5,39	19	12,5	
12	10,84	12	9,13	12	8,15	8	5,56	
7	4,82	7	7,26	7	6,42	8	7,91	
5	5,68	5	4,74	5	5,73	8	6,89	
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Defining Visualization (Vis)

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

[“Visualization Analysis and Design” by T. Munzner, 2014]

Why have a human in the loop?

- Not needed when automatic solution is trusted
- Good for ill-specified analysis problems
 - Common setting: “What questions can we ask?”

Why have a human in the loop?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

Munzner, T. (2014)

Long-term use • Exploratory analysis of scientific data

- Presentation of known results

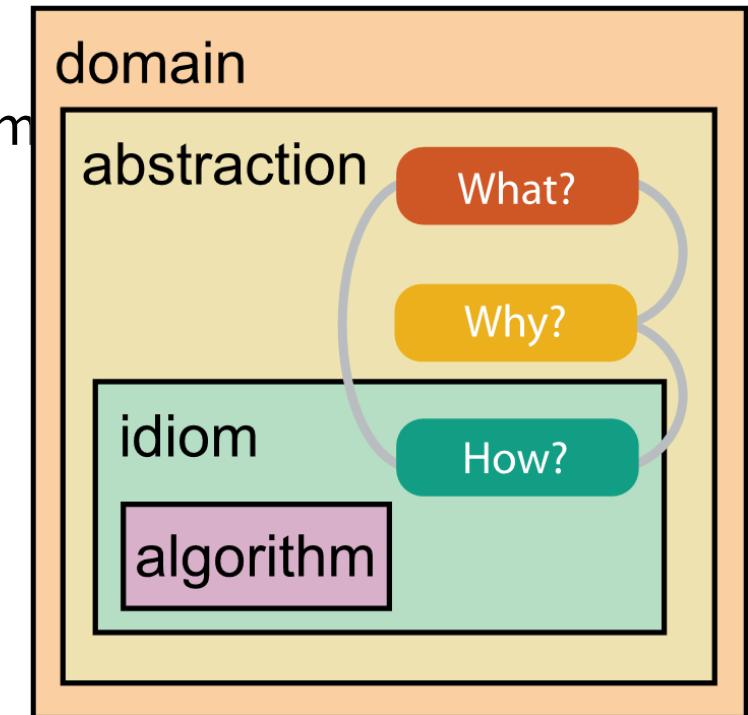
Short-term use • For **developers** of automatic solutions:

- Understand requirements for model development
- Refine/debug and determine parameters

- For **end users** of automatic solutions: verify, build trust

Analysis framework: four levels

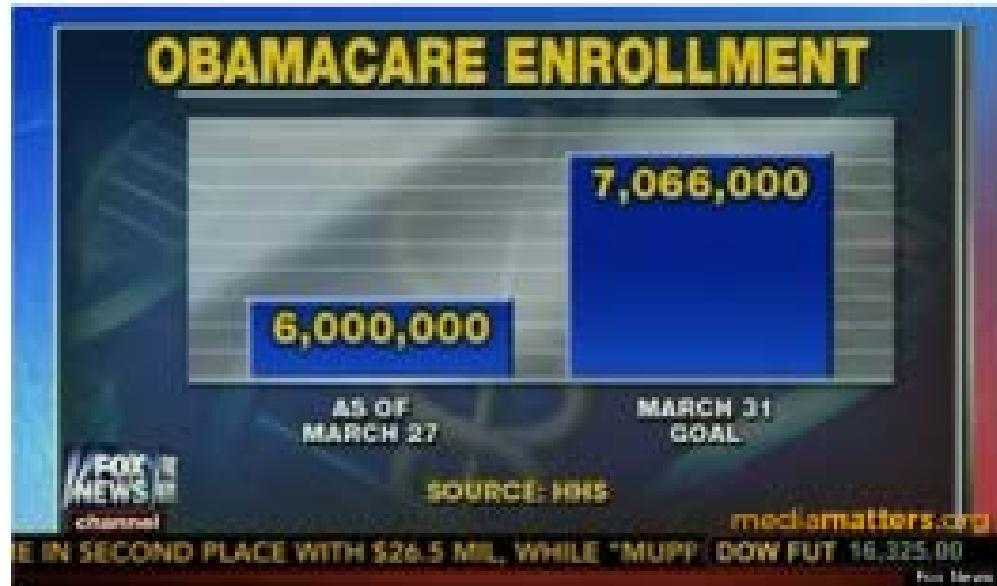
- **Domain** situation: Who are the target users?
- **Abstraction**: Translate from specifics of domain to vocabulary of vis
- **What** is shown? *Data abstraction*
 - Don't just draw what you're given: transform to new form
- **Why** is the user looking at it? *Task abstraction*
- **How** is it shown? *Idiom*
 - Visual encoding idiom: How to draw
 - Interaction idiom: How to manipulate
- **Algorithm**: efficient computation



[A Nested Model of Visualization Design and Validation.
Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]

Pitfalls

- WTF Visualizations (<http://viz.wtf>)
- Without **knowing the principles**, you might make a lot of mistakes like this!



Understand Data, Task, and Encoding

What?

Datasets

→ Data Types

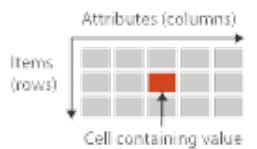
→ Items → Attributes → Links → Positions → Grids

→ Data and Dataset Types

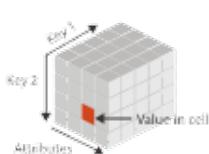


→ Dataset Types

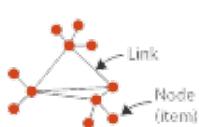
→ Tables



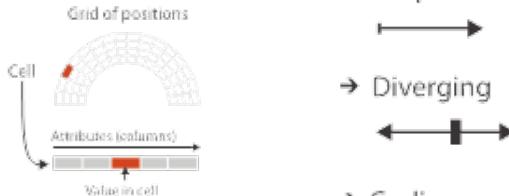
→ Multidimensional Table



→ Networks



→ Fields (Continuous)



→ Trees



→ Geometry (Spatial)



Attributes

→ Attribute Types

→ Categorical

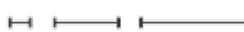


→ Ordered

→ Ordinal



→ Quantitative



→ Ordering Direction

→ Sequential



→ Diverging



→ Cyclic



Data Types

- Items and attributes as rows and columns of tables
- Position and time are special attributes
- Spatial data on grids makes computation easier

→ Dataset Availability

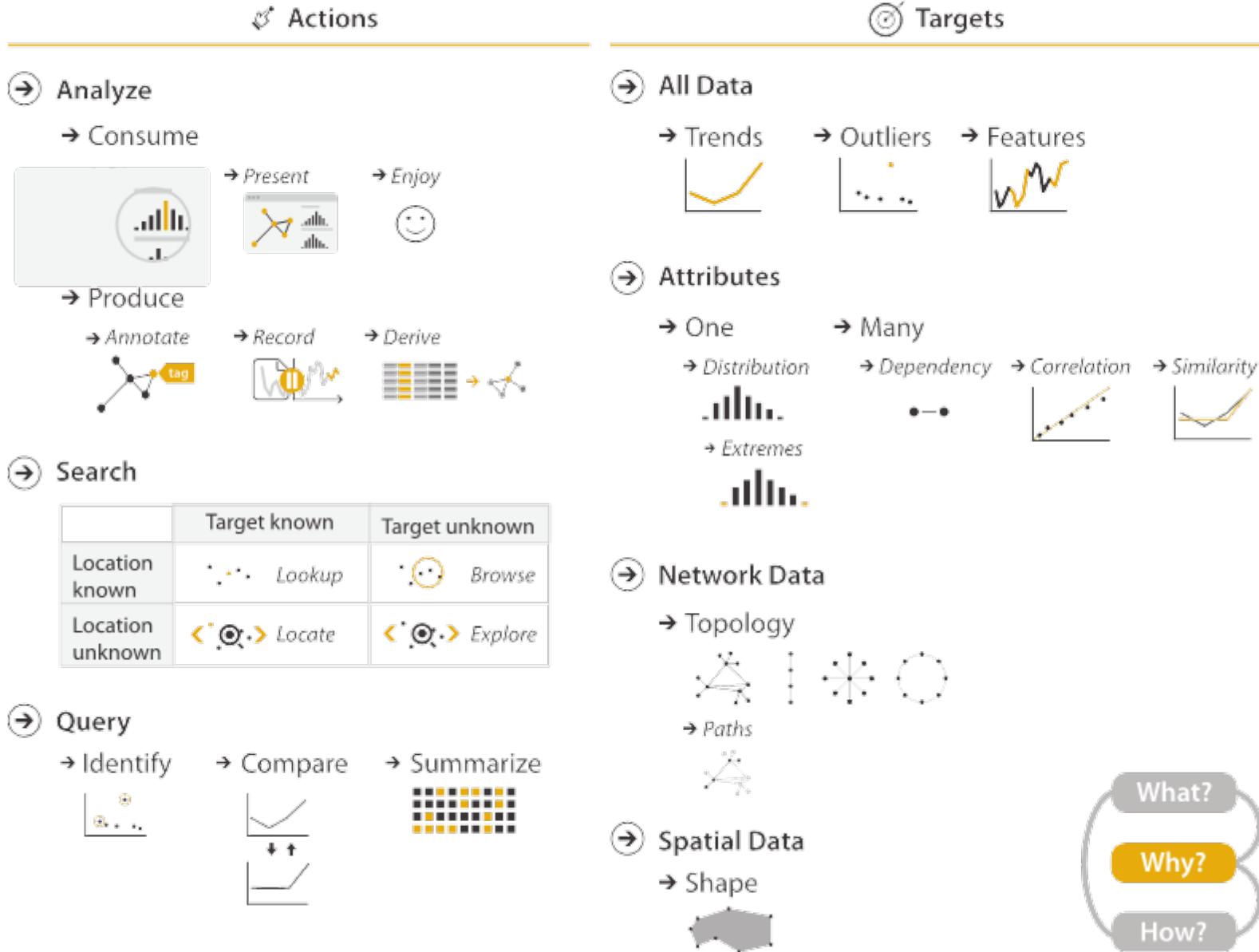
→ Static



→ Dynamic



Why?



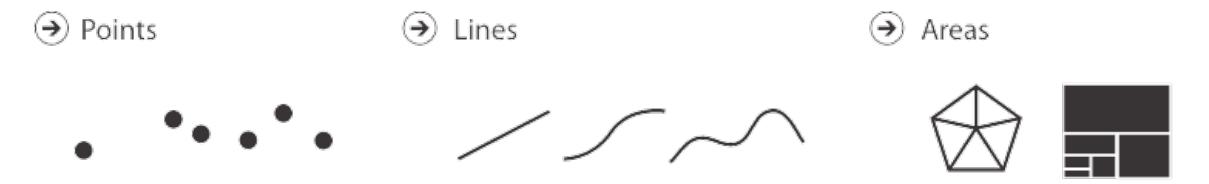
Tasks

- **Actions**
 - Analyze
 - Search
 - Query
- **Targets**
 - Item & Attributes
 - Topology & Shape

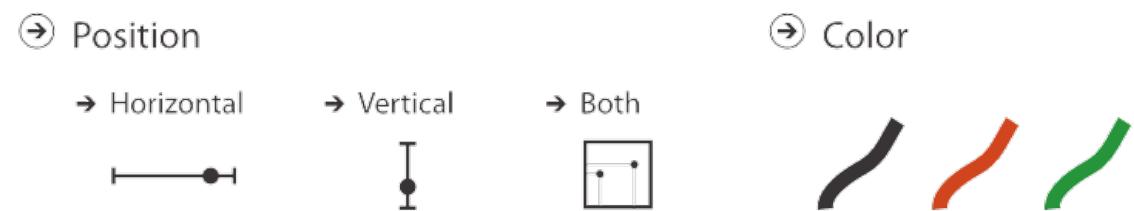


Visual Encoding – How?

- Marks
 - Geometric primitives



- Channels
 - Appearance of marks
 - Redundant coding with multiple channels possible

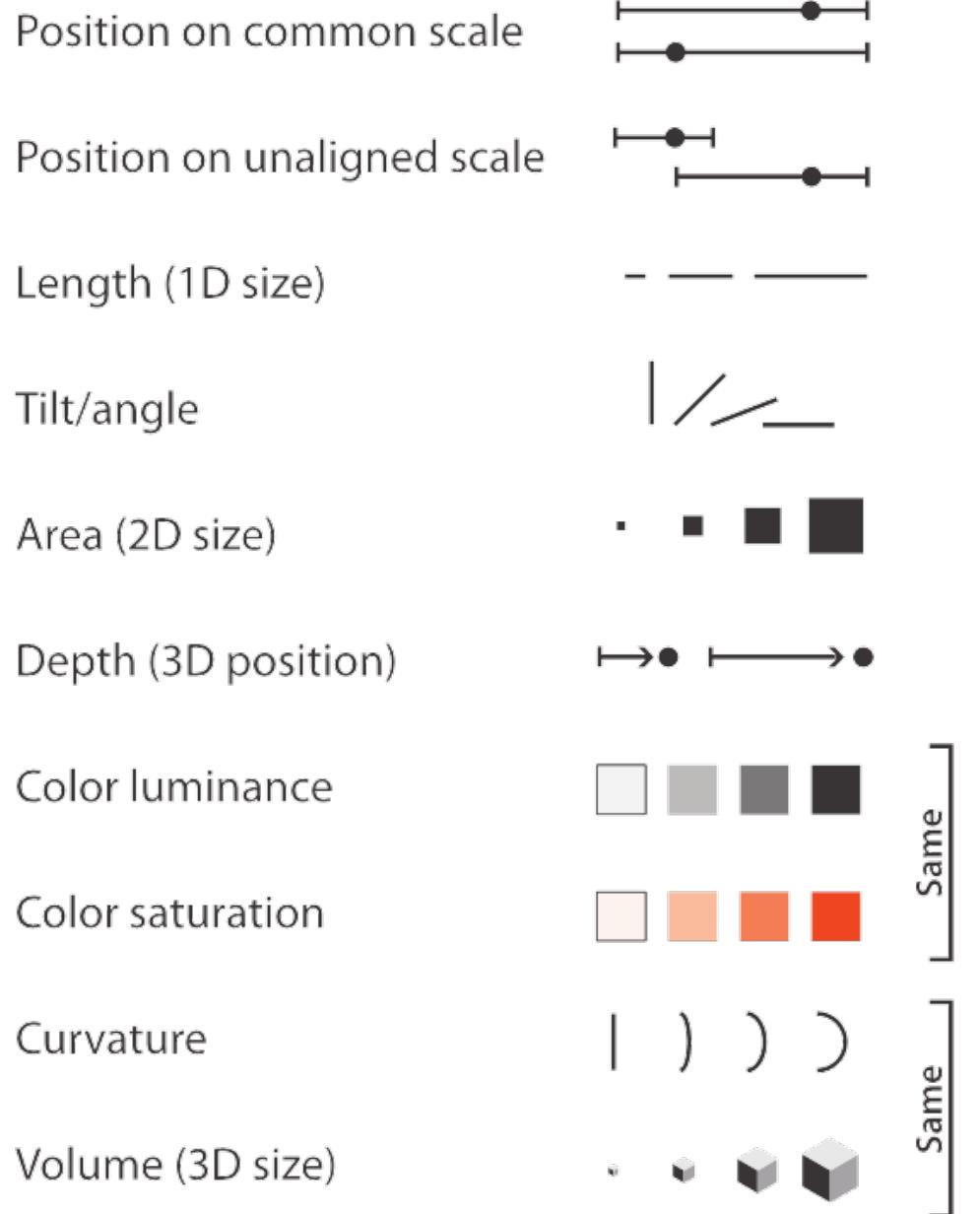


Design Principles for Task Effective Visualization

Resource limitations

- **Computational** limits
 - Processing time and system memory
- **Human** limits
 - Human attention and memory
 - Understanding abstractions
- **Display** limits
 - Pixels are precious
 - Information density tradeoff: Info encoding vs unused whitespace

→ **Magnitude Channels: Ordered Attributes**



→ **Identity Channels: Categorical Attributes**



Expressiveness principle

- **Match channel and data characteristics**

Effectiveness principle

- **Encode important attributes with higher ranked channels**

Chart Design: Simplifying

Example from Tim Bray

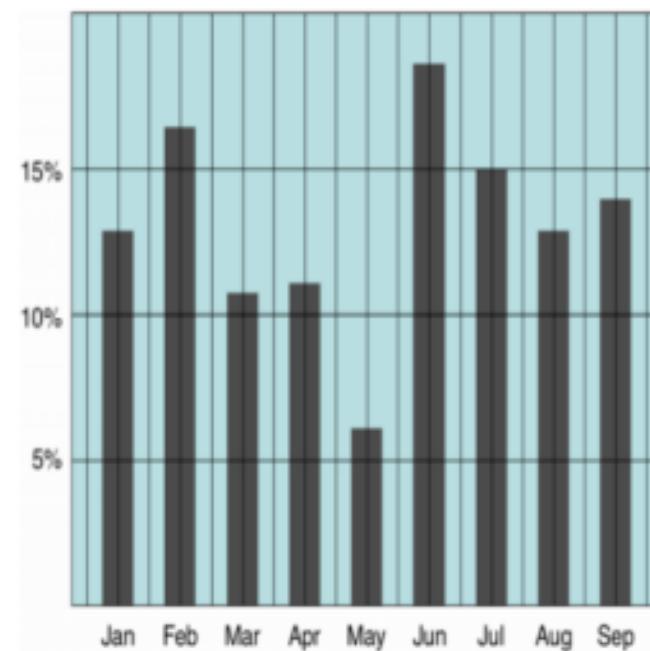
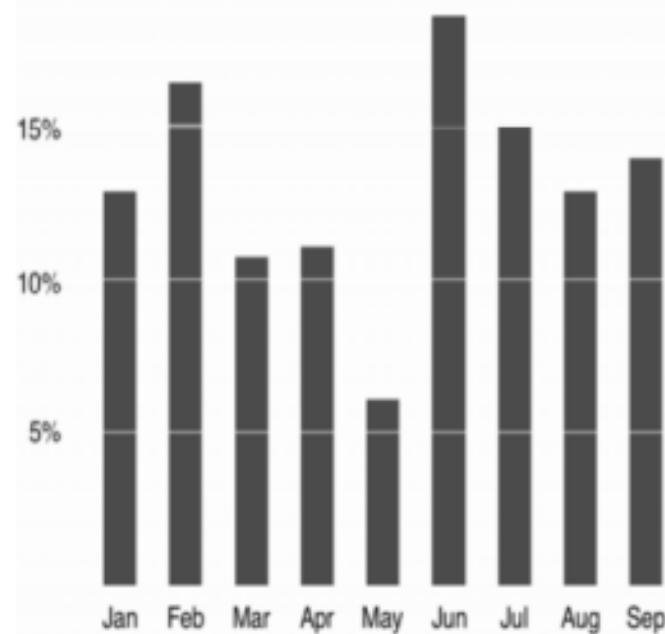


Chart Design: Simplifying

Example from Tim Bray



Principle 2: Understand Magnitudes

Which one is brighter?



Principle 2: Understand Magnitudes

Which one is longer?



Principle 3: Use Color

- **Make your visualization look beautiful**
 - Colour Lovers: <http://www.colourlovers.com>
- **Work for different kinds of data**

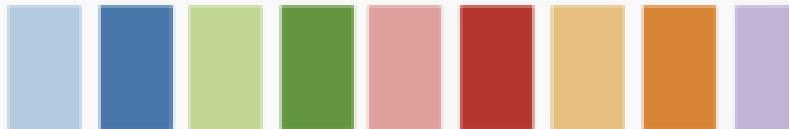
Diverging

Two sequential schemes extended out from a critical midpoint value



Categorical

Lots of contrast between each adjacent color



Principle 4: Use Structure

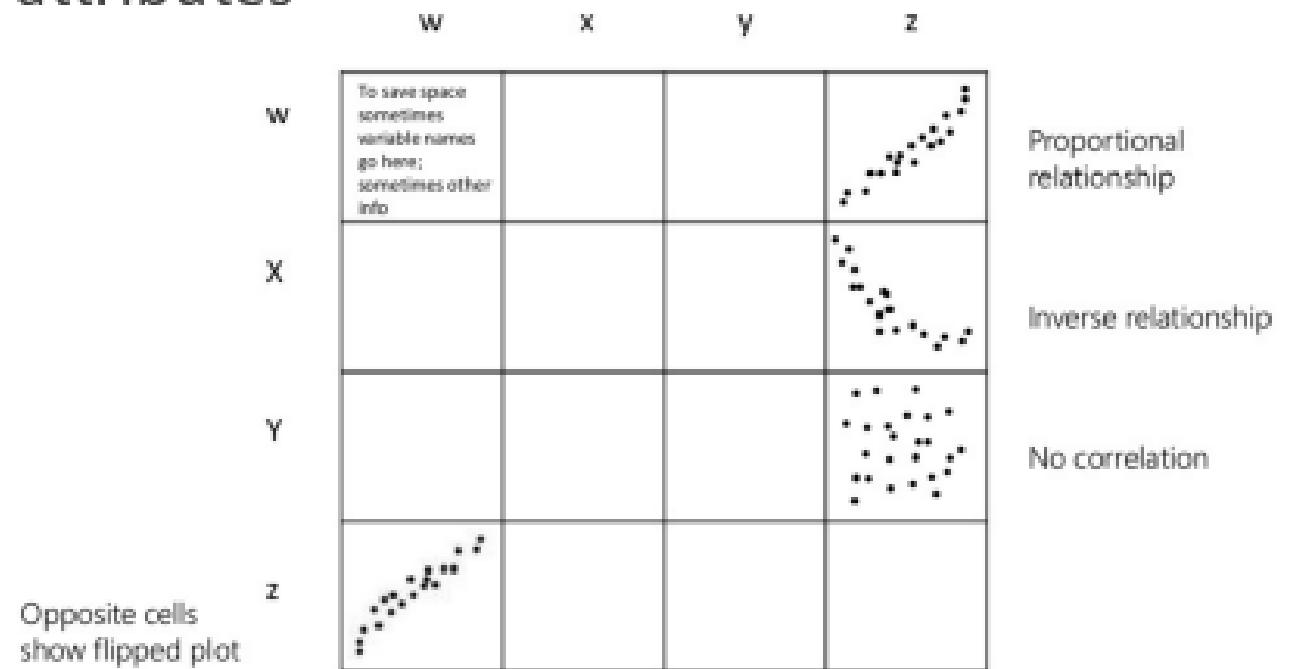
- Chart chooser: <http://labs.juiceanalytics.com>



Principle 4: Use Structure

Correlation Visualization

- Consider a table with n=4 attributes



Sources

- Tamara Munzner's ["Visualization Analysis and Design"](#), 2014
- Jiannan Wang's CMPT 733 slides, Spring 2017
- Torsten Möller's Visualization course, Spring 2018