Storytelling with Data

Nam Wook Kim

Mini-Courses — January @ GSAS
2018

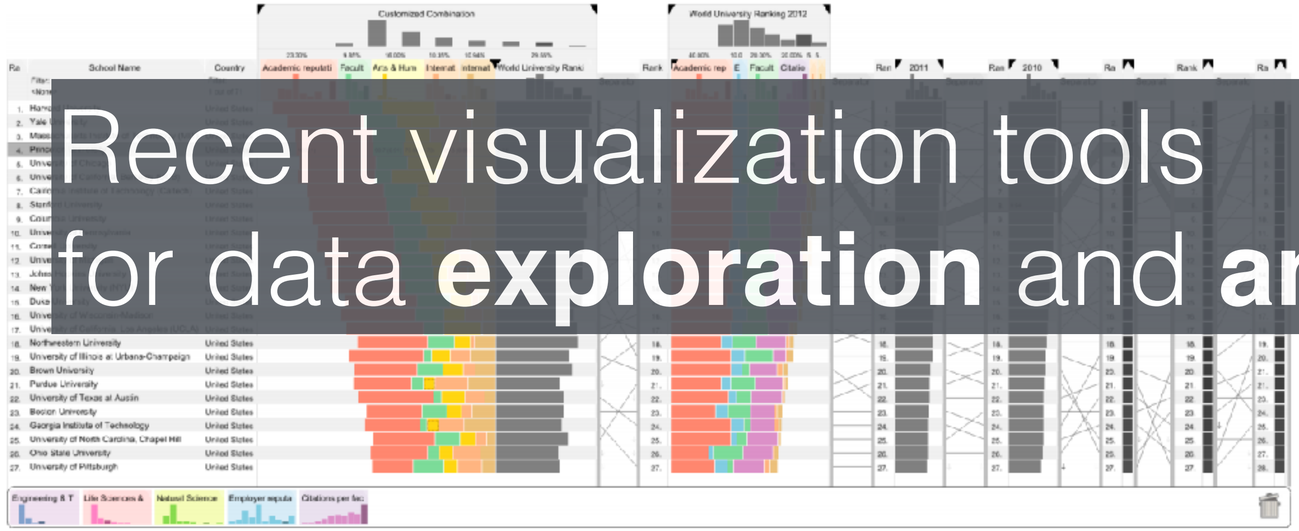
Goal

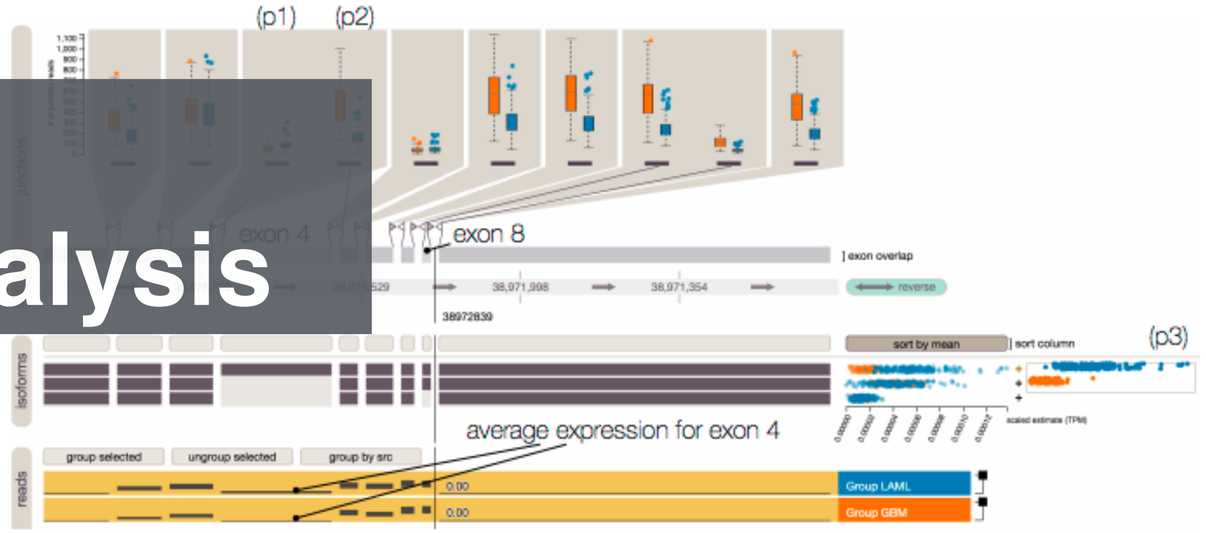
From exploratory analysis
to communication

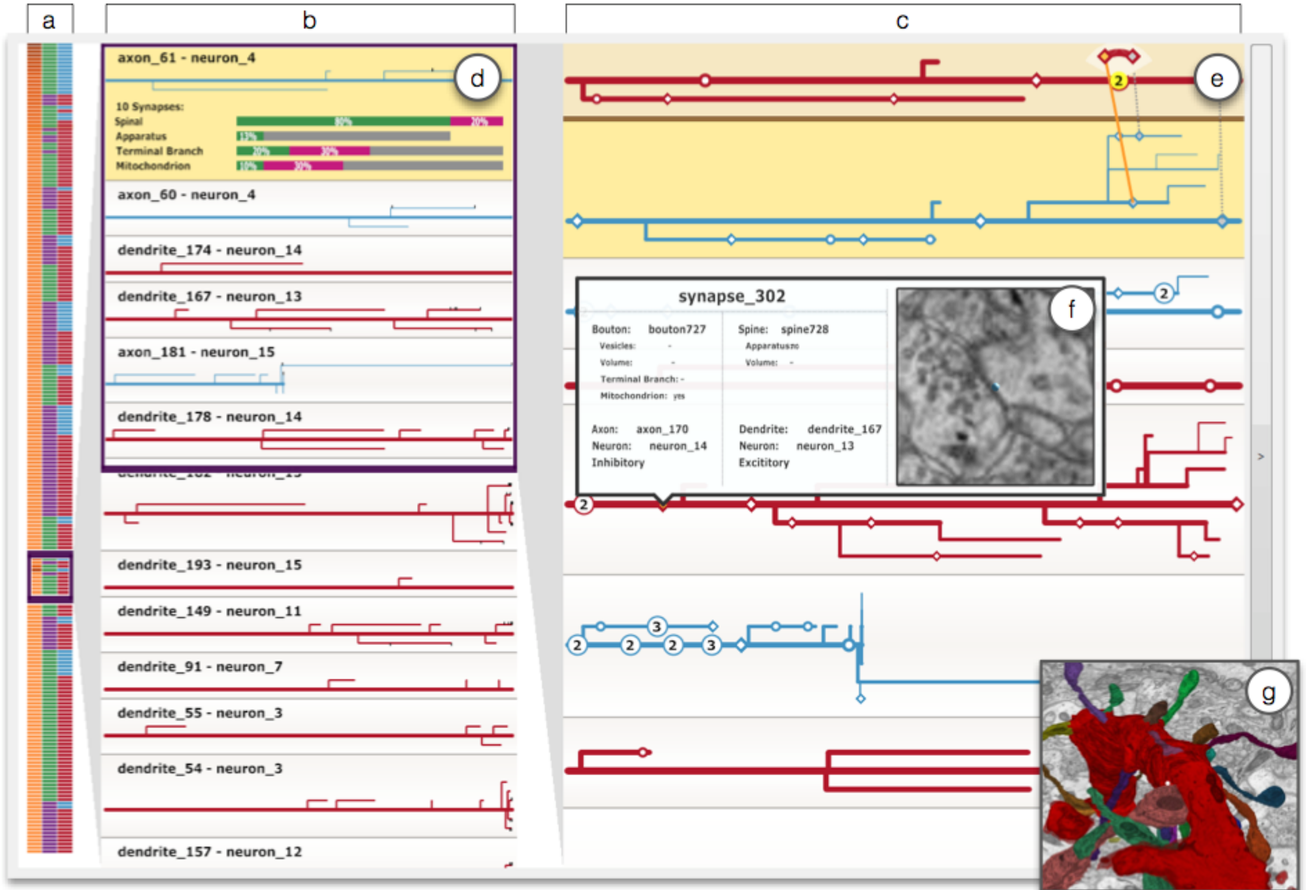
Topics

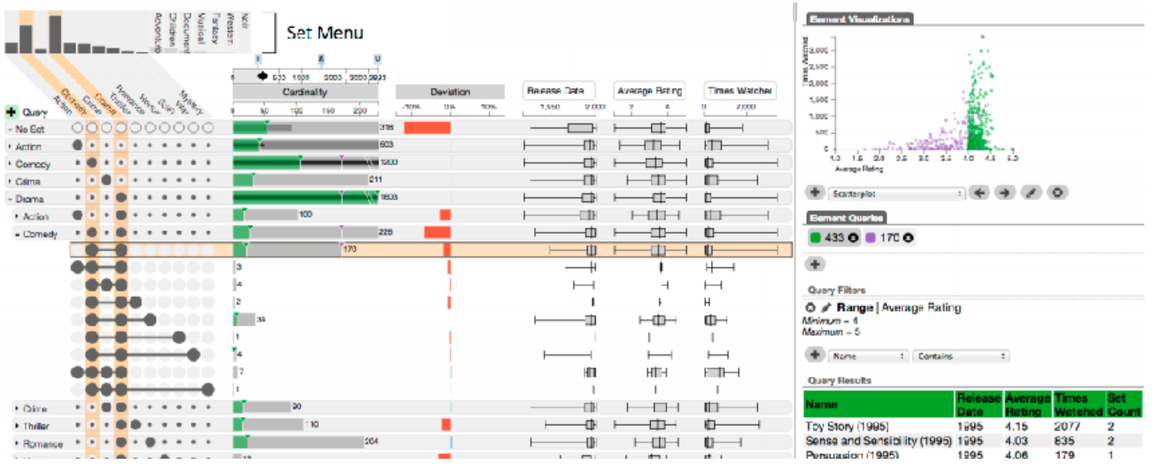
- Visualization for communication
- Data-Driven Storytelling
- Presentation & Storytelling in Tableau

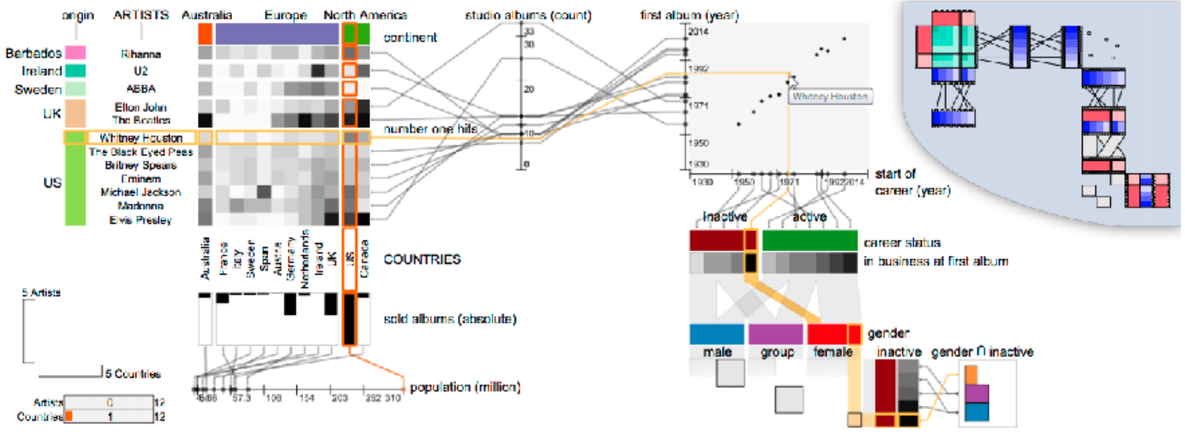
Recent visualization tools for data **exploration** and **analysis**











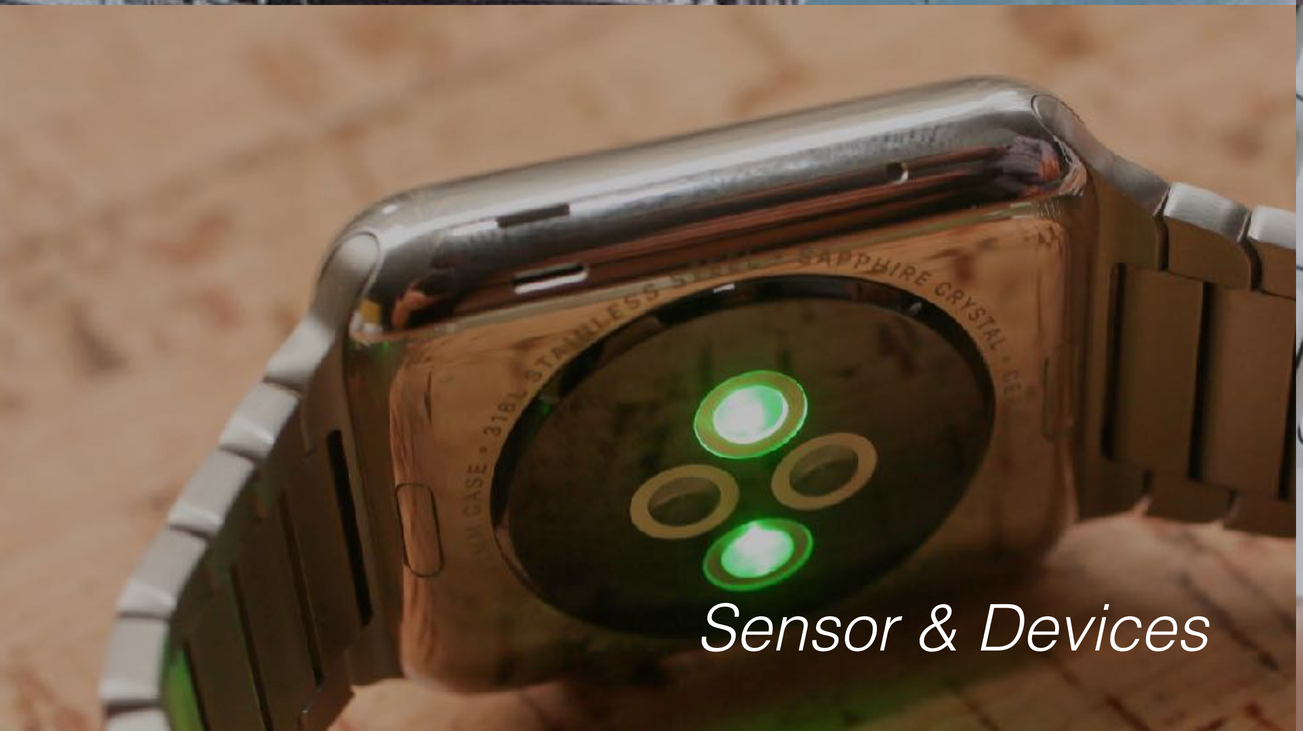
Data → Explore → ?

- Data Centered
- Domain Experts
- Discovering Insights

Need for communicating large, complex, and diverse data.

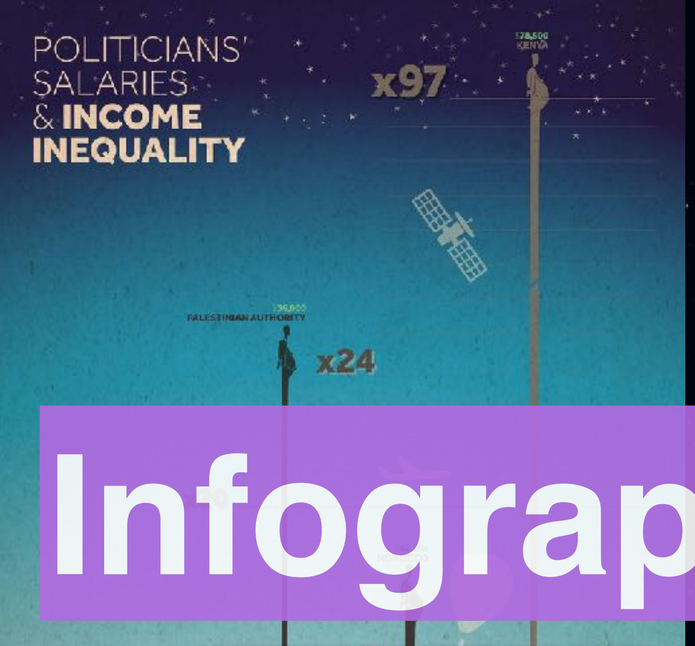
Open Government

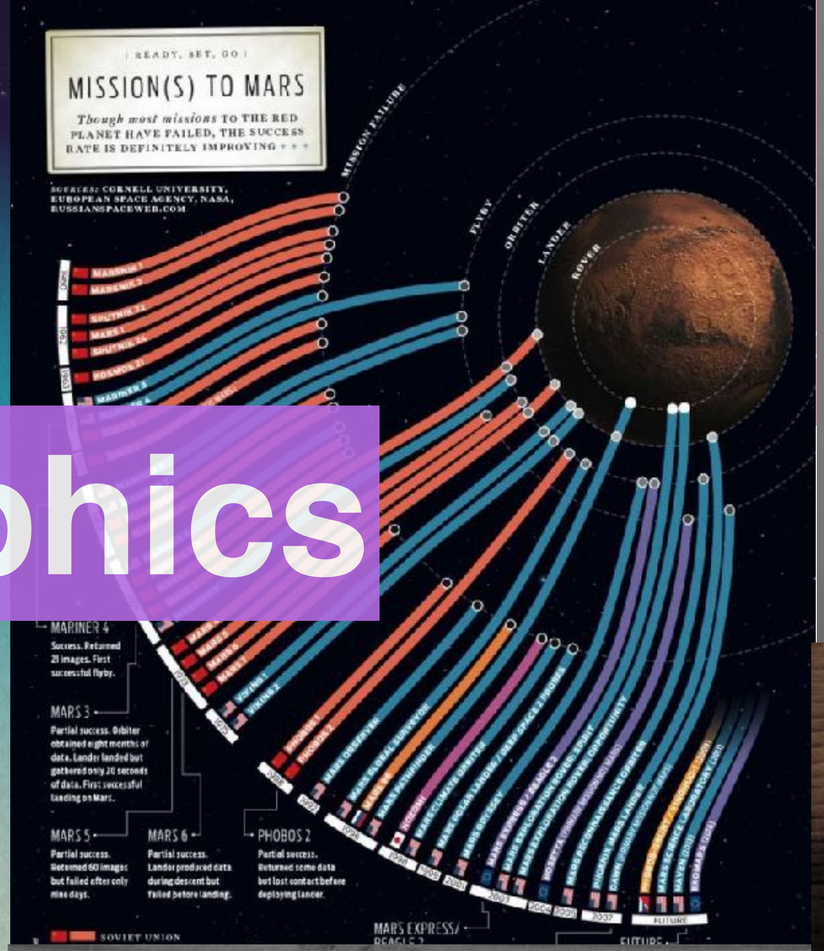
Health & Medicine



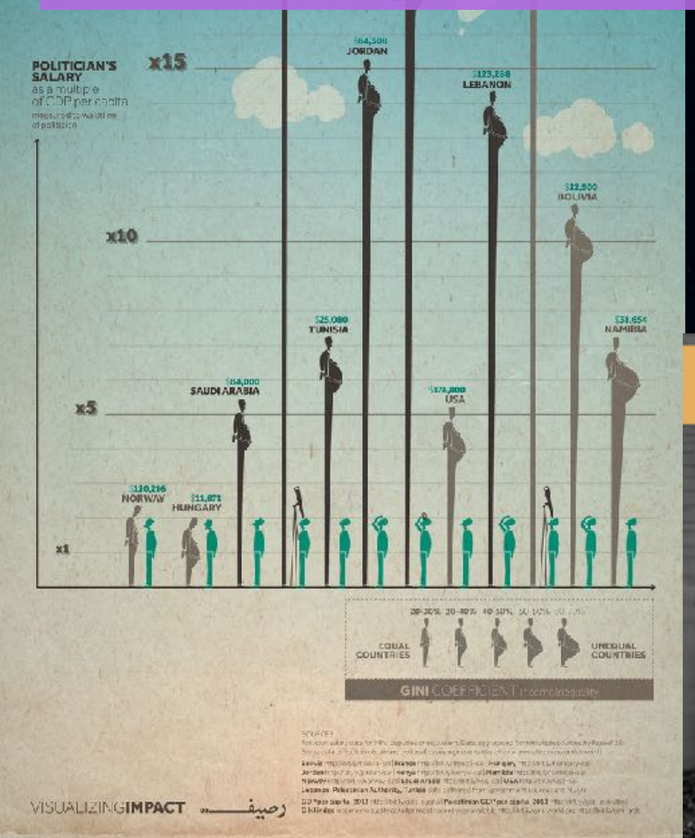
Sensor & Devices

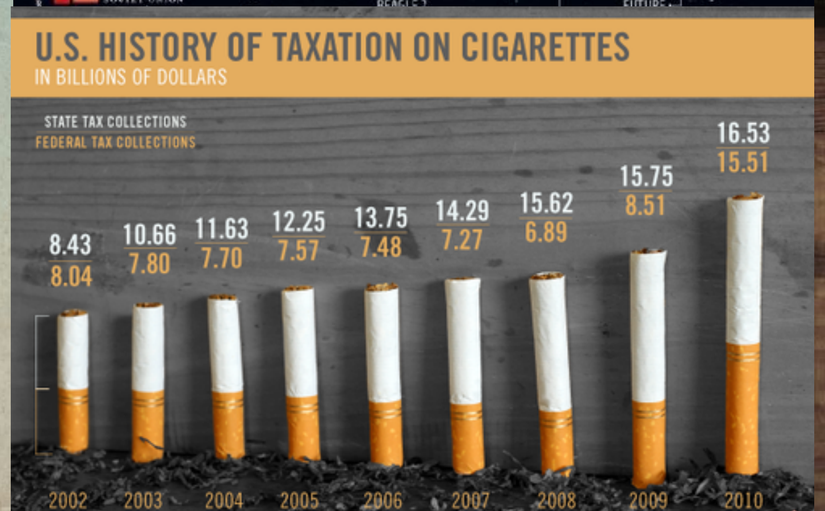
Social Media

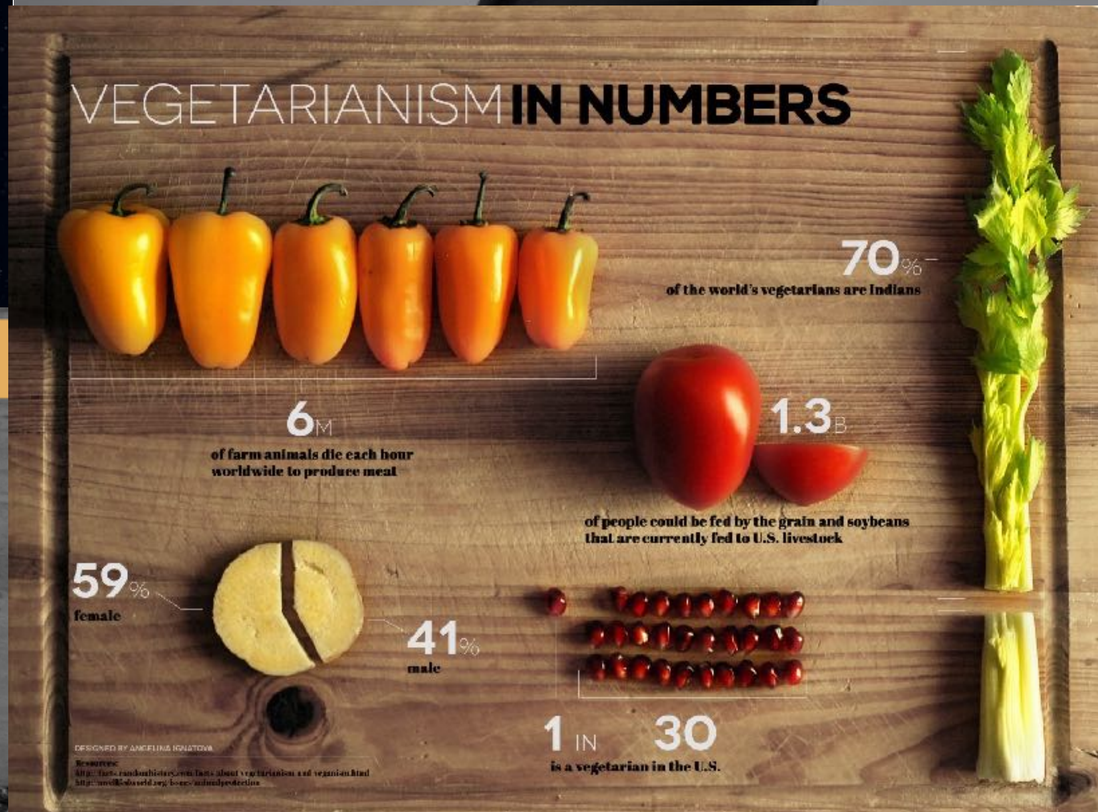




Infographics







Storytelling

Hans Rosling

tapestry

Storytelling is the most powerful way to put ideas into the world today.

Robert McKee

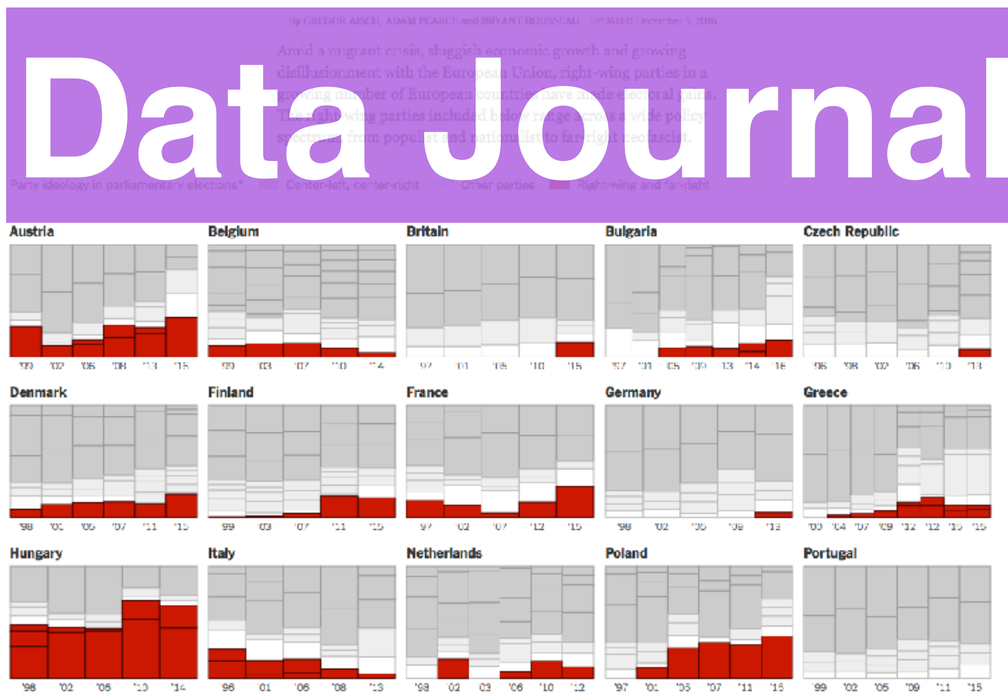
cole nussbaumer knaflic

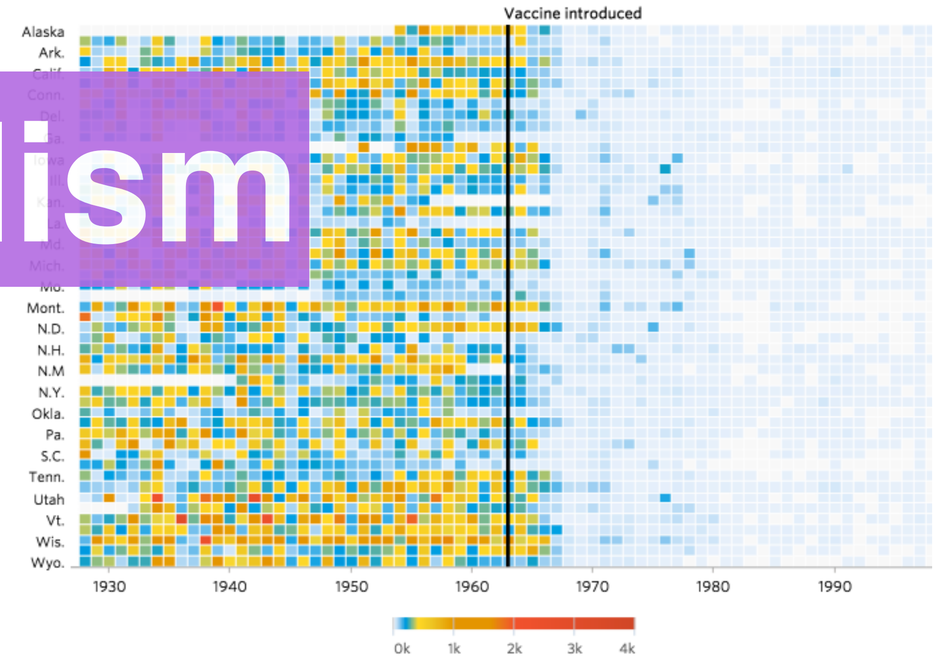
storytelling with data

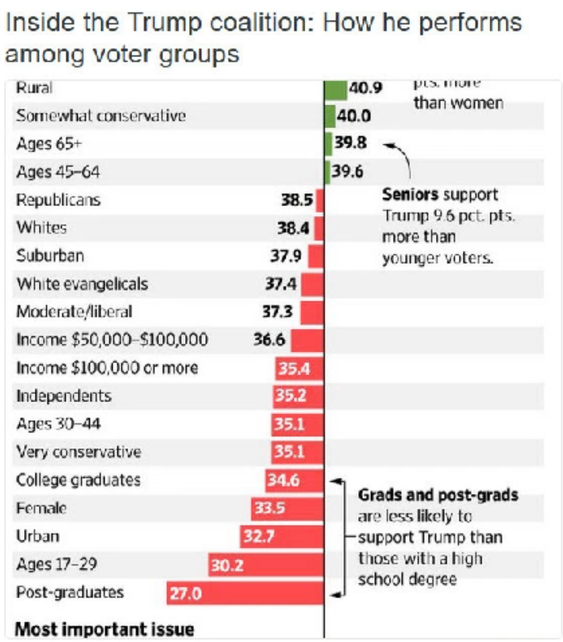
a data visualization guide for business professionals

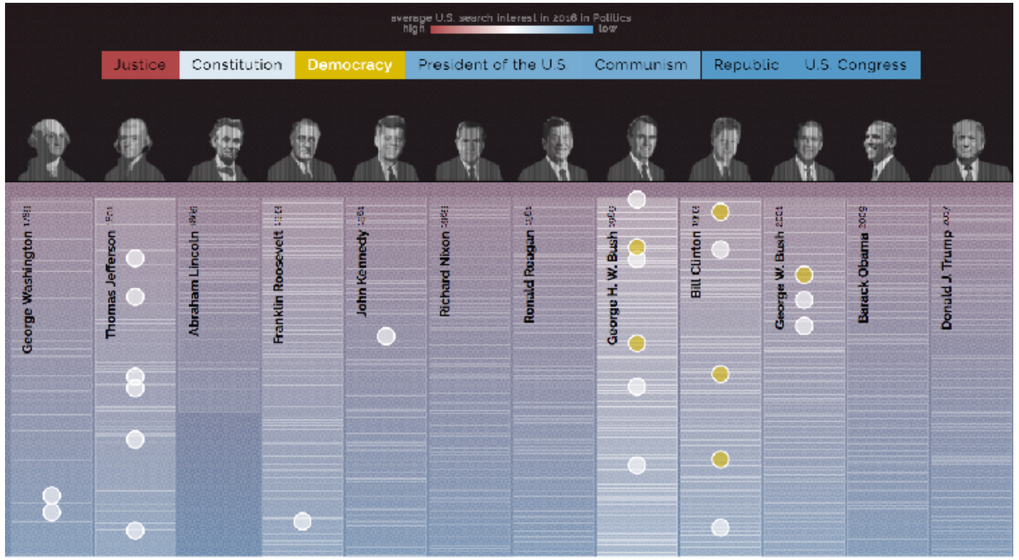
WILEY

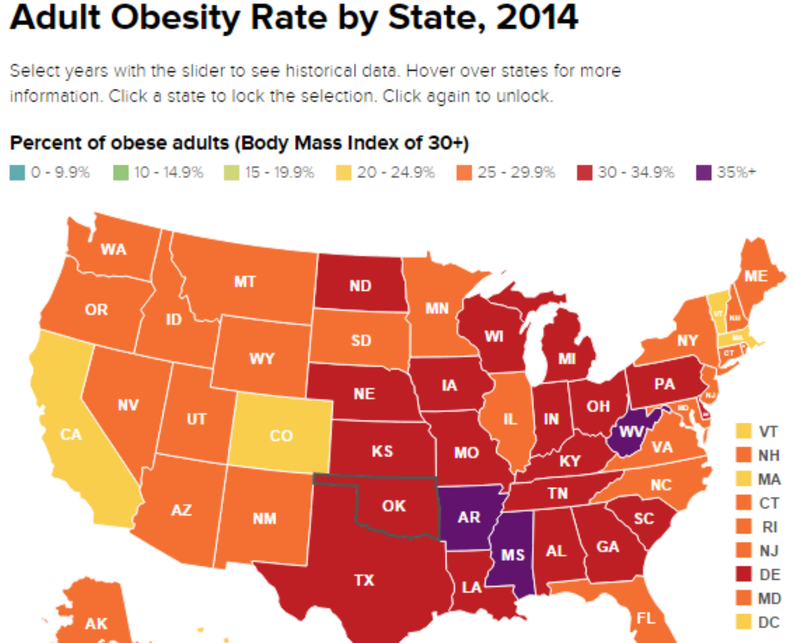
Data Journalism

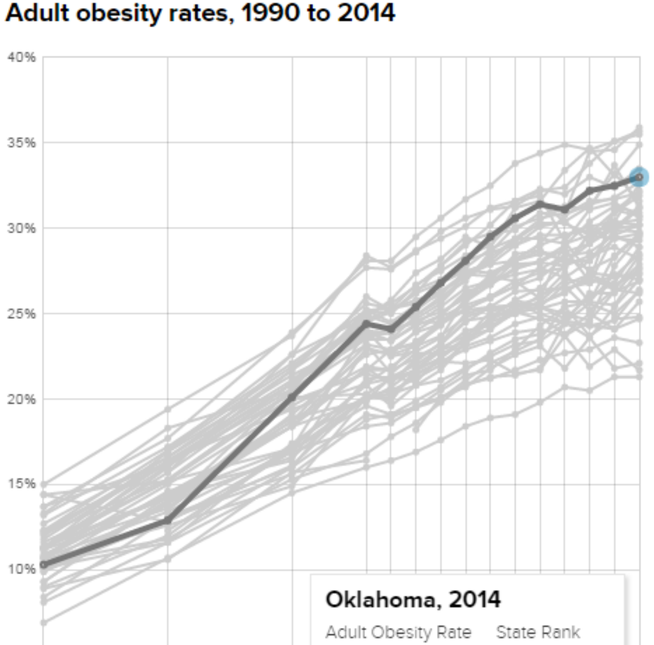




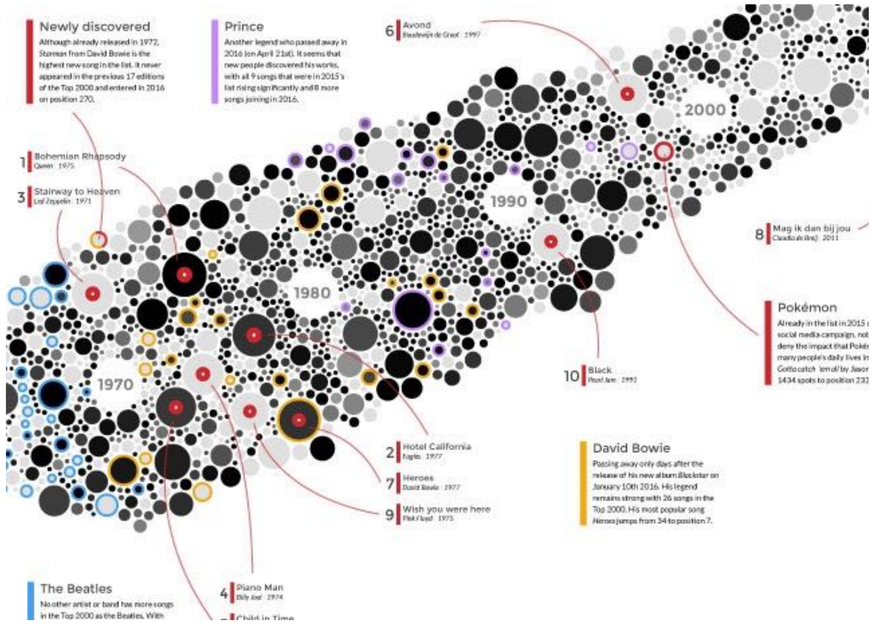


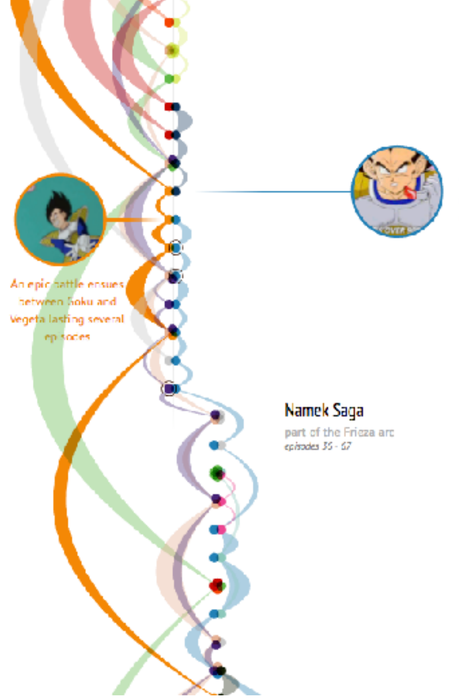


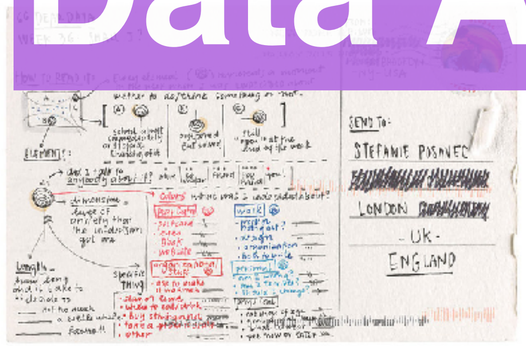


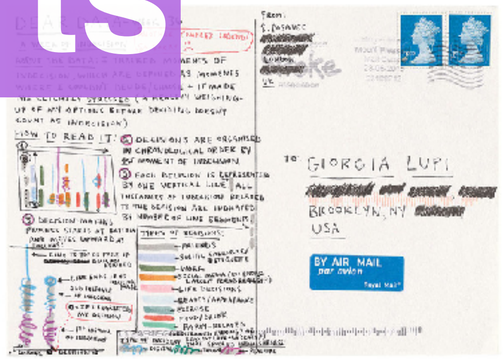


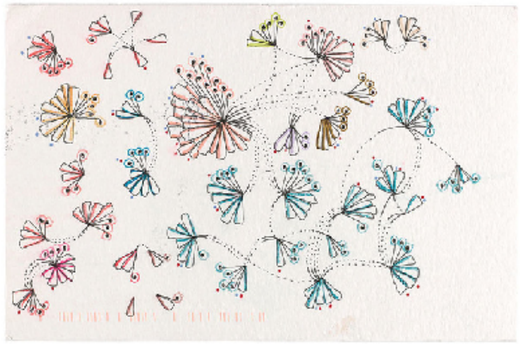
Data Arts

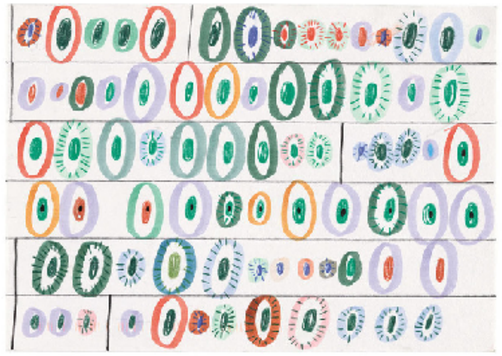


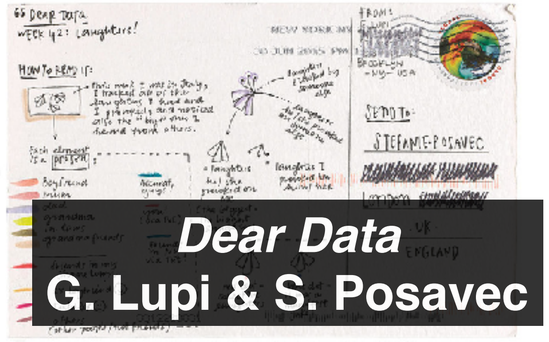


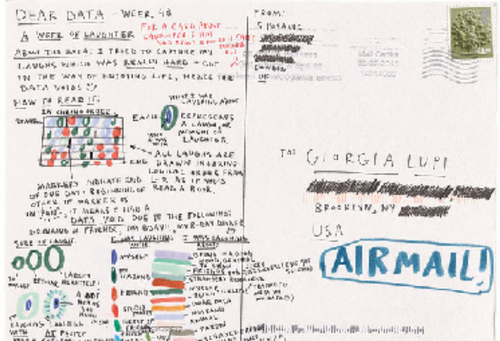




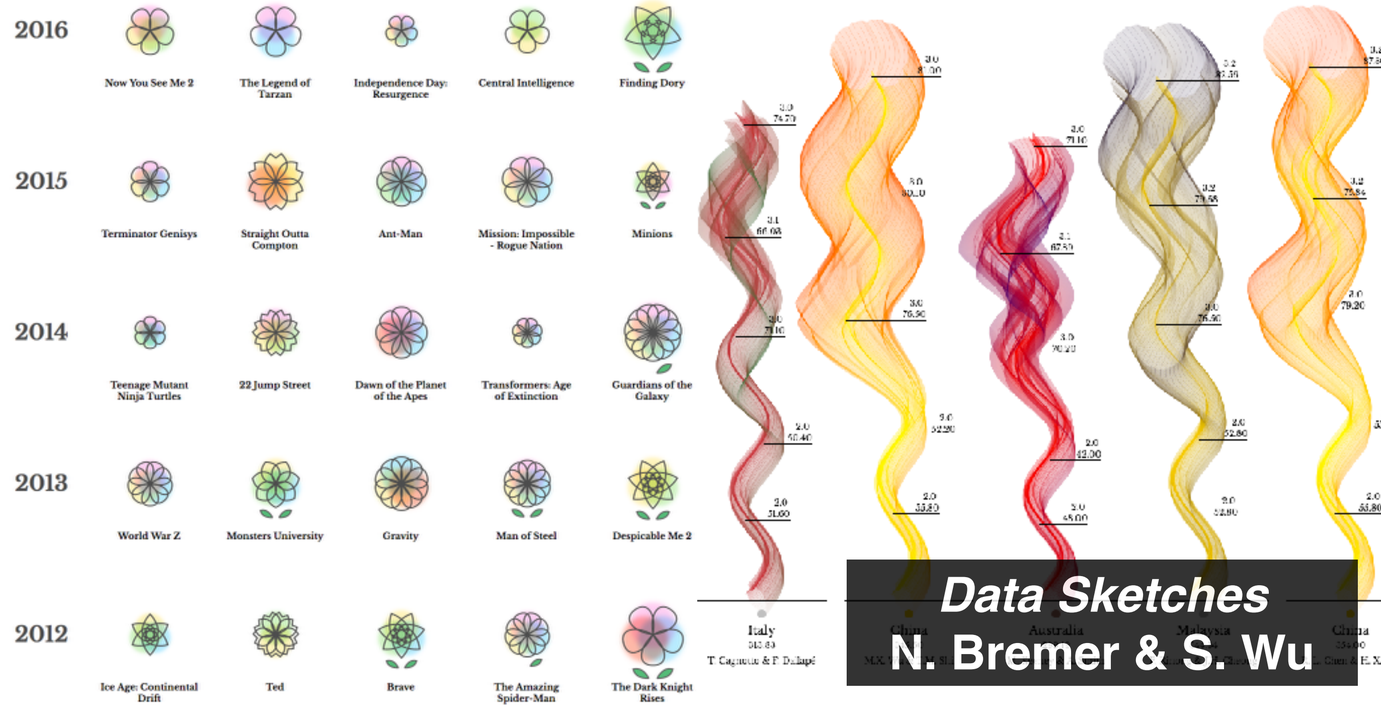








Dear Data
G. Lupi & S. Posavec



Data Sketches
N. Bremer & S. Wu

Data → Explore → Communicate

- Data Centered
- Domain Experts
- Discovering Insights

- Human Centered
- General Audience
- Conveying Messages

New Challenges

What is beyond data exploration?

What is a communication-oriented visualization?

How can we better support communication of data?

How can we tell compelling stories with data?

“

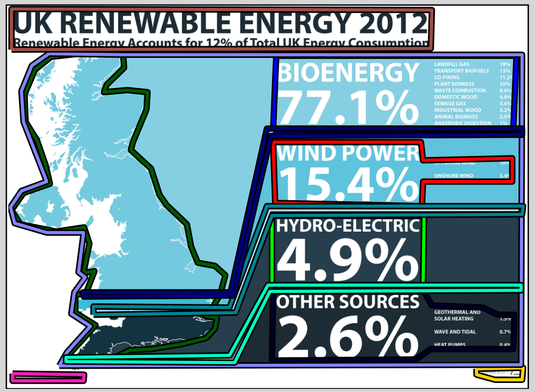
Presentation—specifically, its **use of elements from storytelling**—is the next logical step in visualization research and should be a focus of at least **equal importance** with **exploration** and **analysis**

— Kosara & Mackinlay 2013

Research:

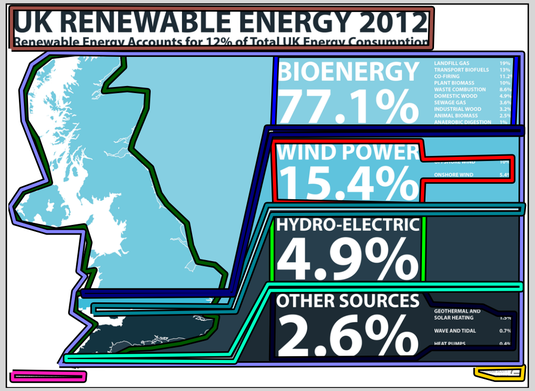
What makes a visualization communicative?

LABELED VISUALIZATION DATABASE

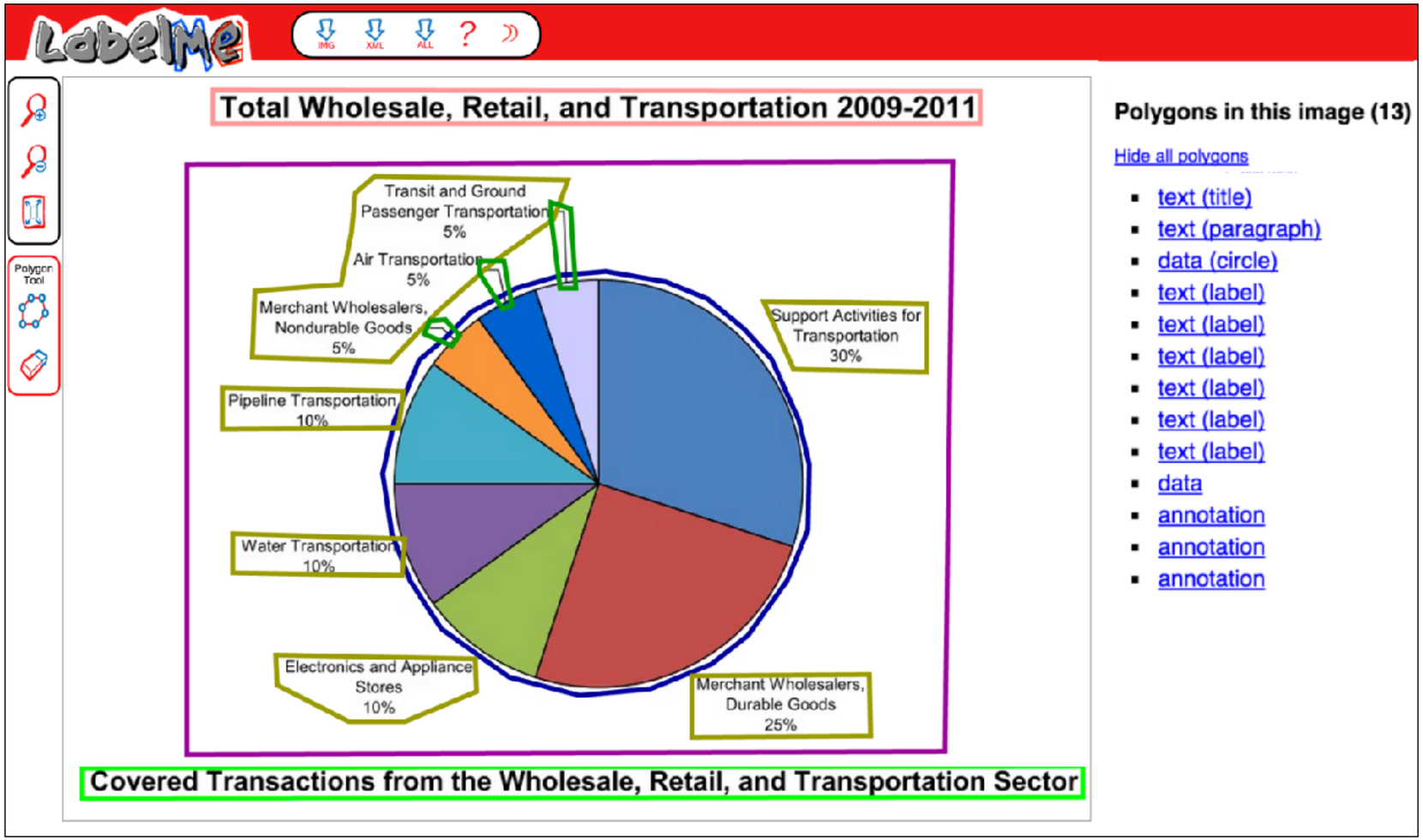


393 visualizations

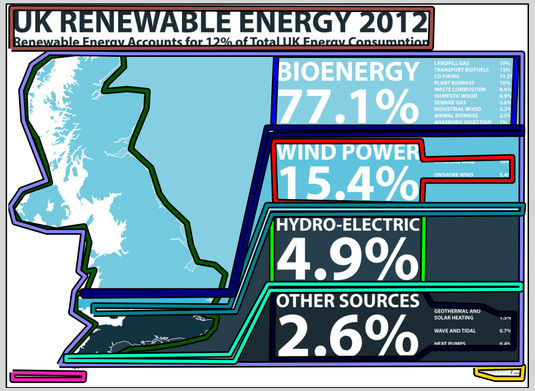
LABELED VISUALIZATION DATABASE



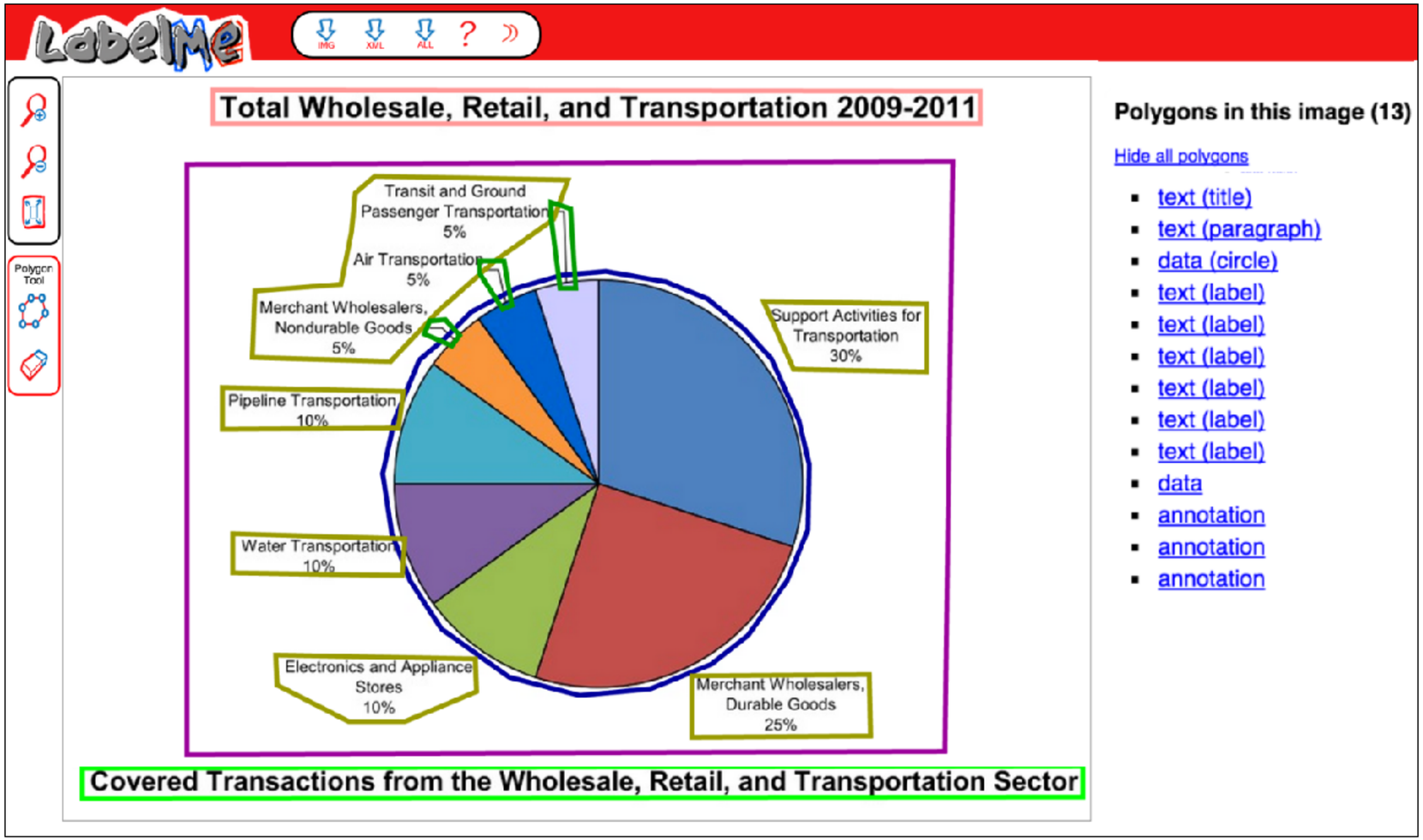
393 visualizations



Labeled Visualization Database



393 visualizations



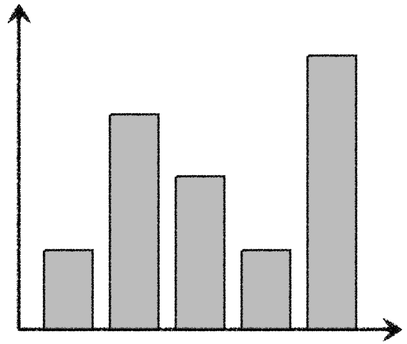
\+

Data-Ink Ratio and Visual Density

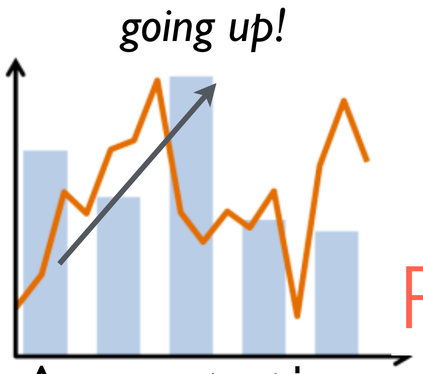
Human Recognizable Objects

Data and Message Redundancy

Examples of Redundant Encoding

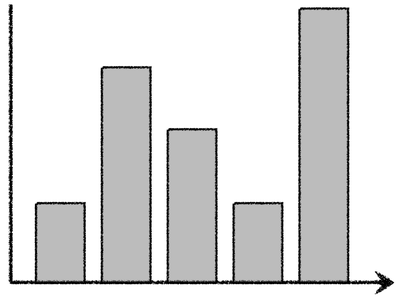


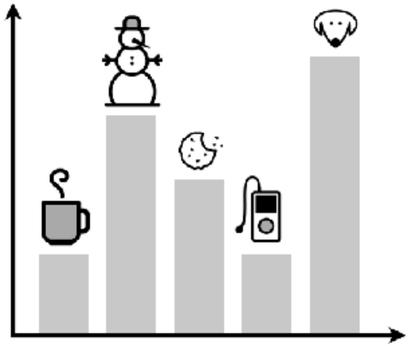
Examples of Redundant Encoding



Annotation

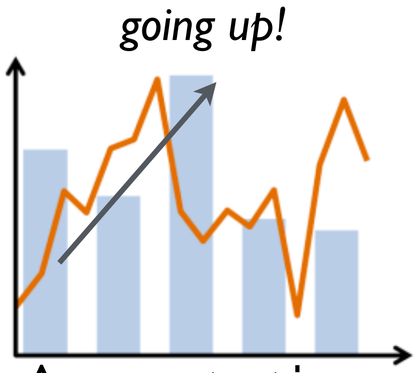
Message Redundancy



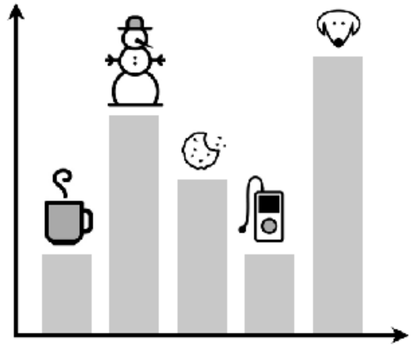


Human recognizable objects

Examples of Redundant Encoding

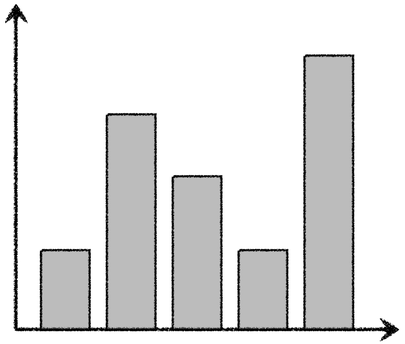


Annotation

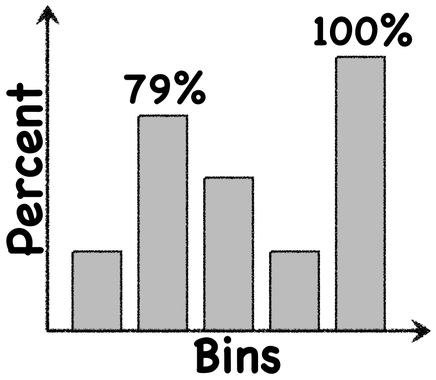


Human recognizable objects

Message Redundancy



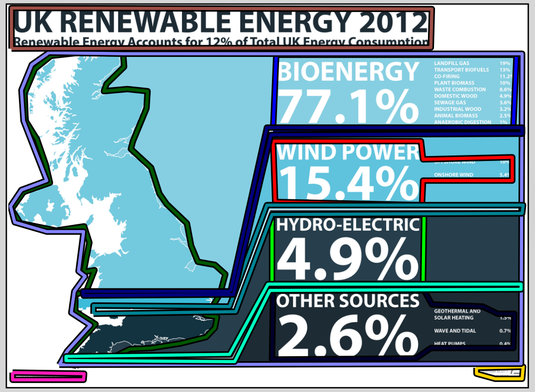
Data Redundancy



Data Labels

Experiment Procedure

Labeled Visualization Database

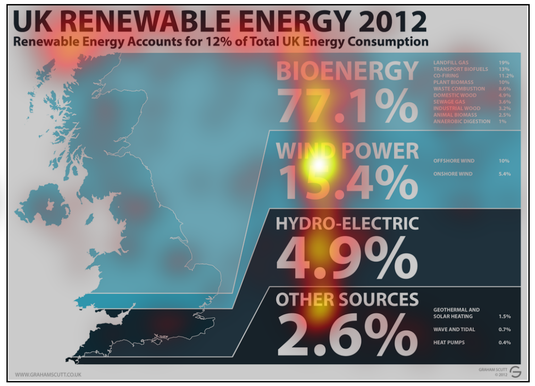


393 visualizations

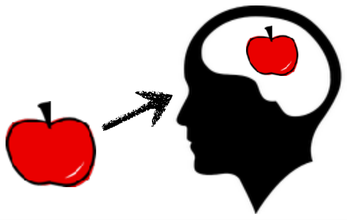
33 participants

"Encoding"

10 seconds / image

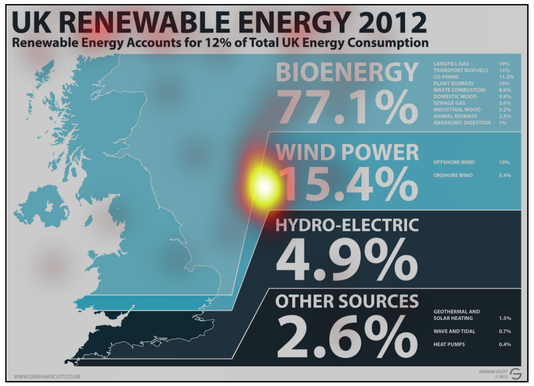


Eye-Tracking Data

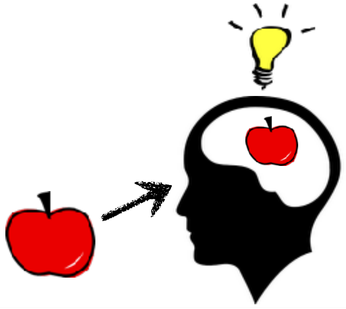


"Recognition"

2 seconds / image



Eye-Tracking Data



"Recall"

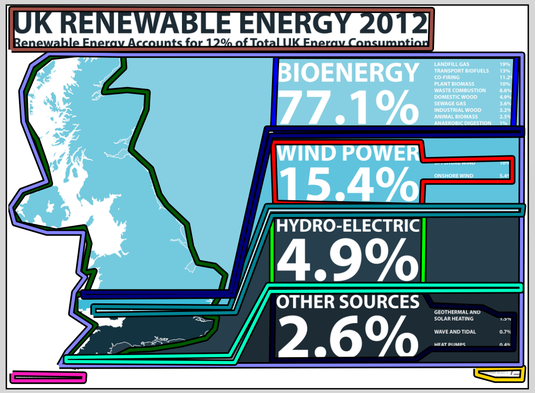
20 minutes

Text Descriptions

~60 minutes

Experiment Procedure

Labeled Visualization Database

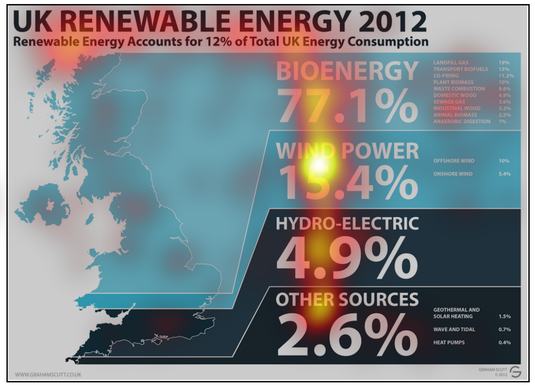


393 visualizations

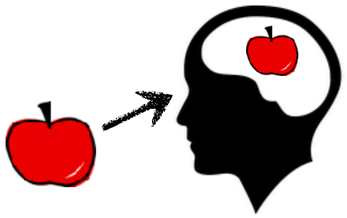
33 participants

"Encoding"

10 seconds / image



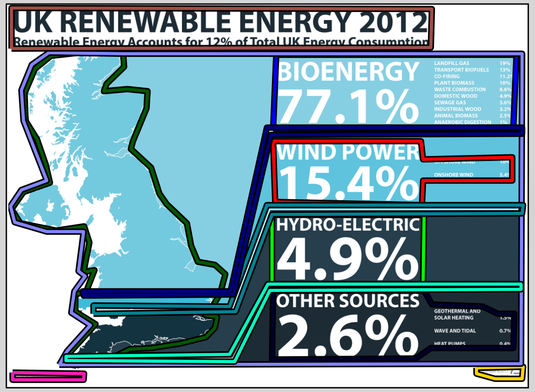
Eye-Tracking Data



20 minutes

Experiment Procedure

Labeled Visualization Database

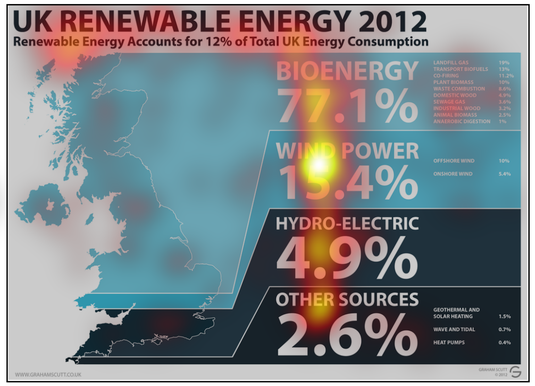


393 visualizations

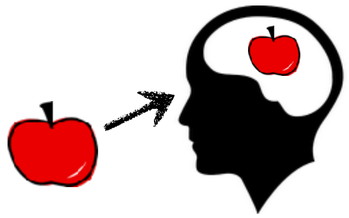
33 participants

"Encoding"

10 seconds / image

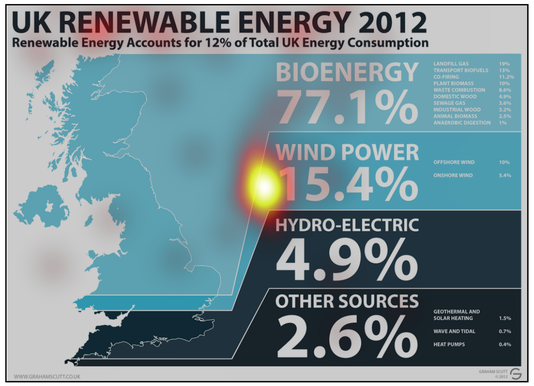


Eye-Tracking Data

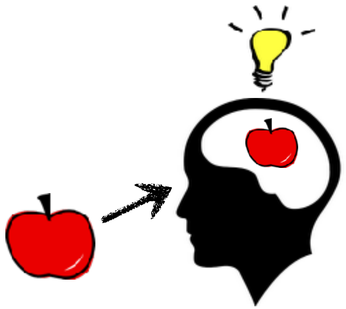


"Recognition"

2 seconds / image



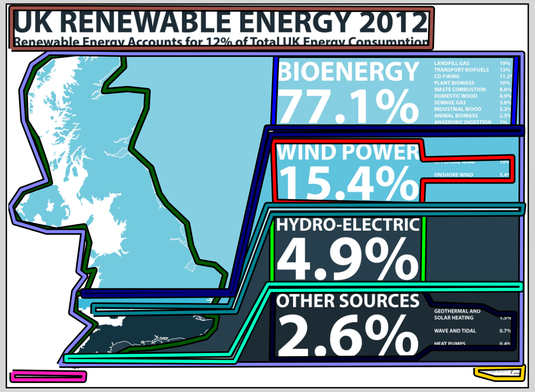
Eye-Tracking Data



10 minutes

Experiment Procedure

Labeled Visualization Database

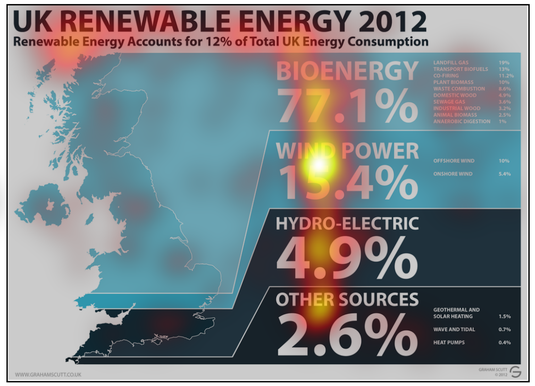


393 visualizations

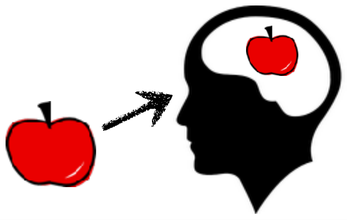
33 participants

"Encoding"

10 seconds / image

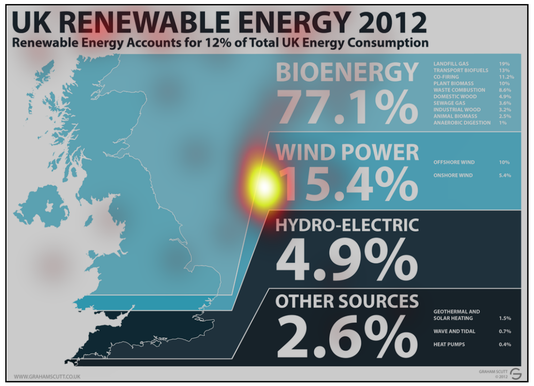


Eye-Tracking Data

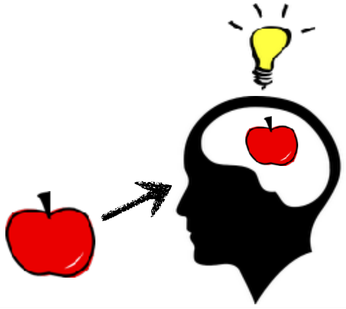


"Recognition"

2 seconds / image



Eye-Tracking Data



"Recall"

20 minutes

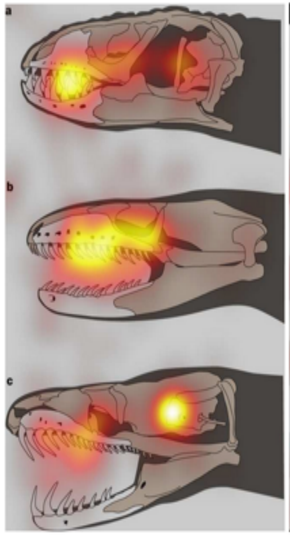
Text Descriptions

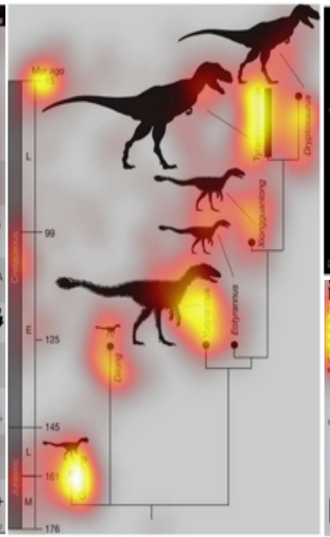
20 minutes

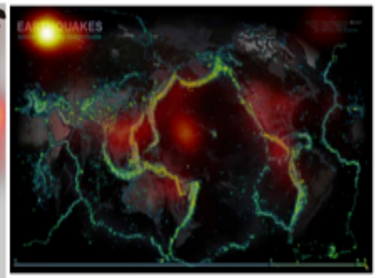
Results

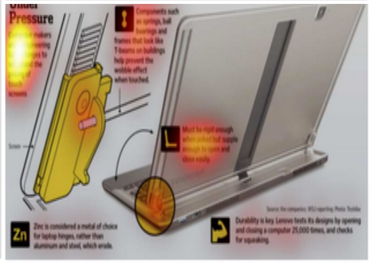
Encoding (10 sec)

Most Memorable

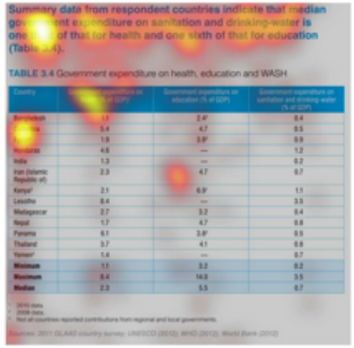


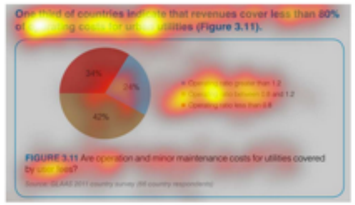


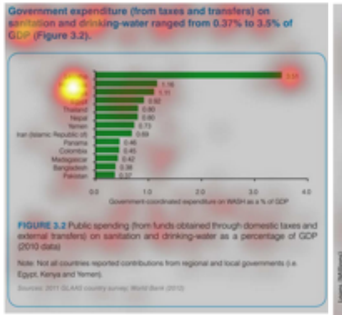


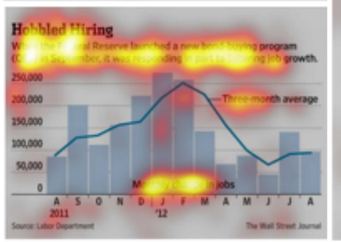


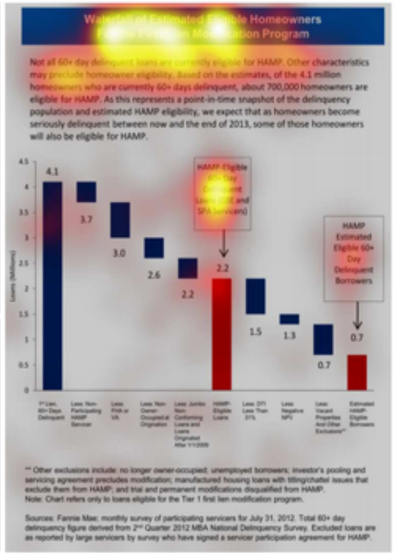
Least Memorable









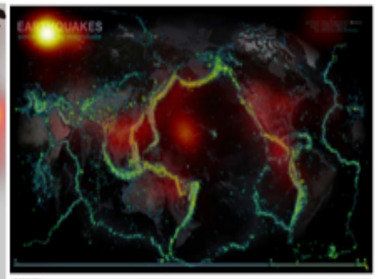


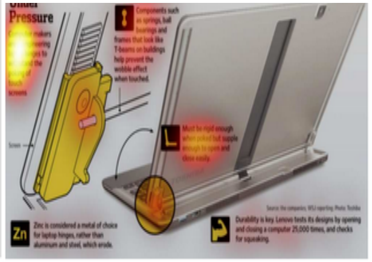
Most Memorable

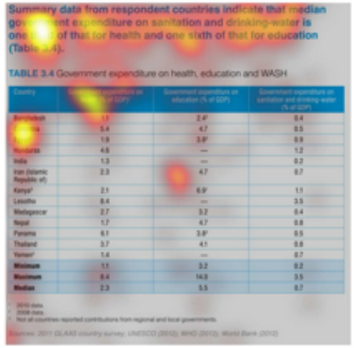
Least Memorable

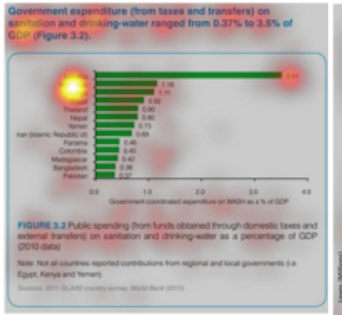
Encoding (10 sec)

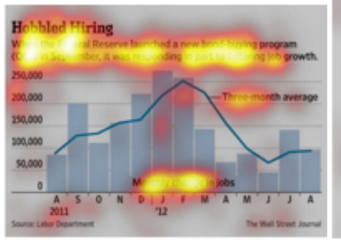
Recognition

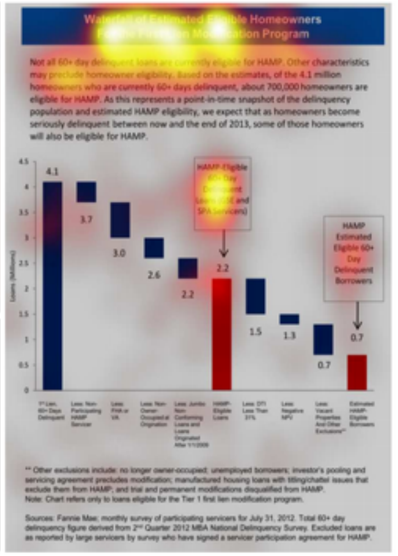


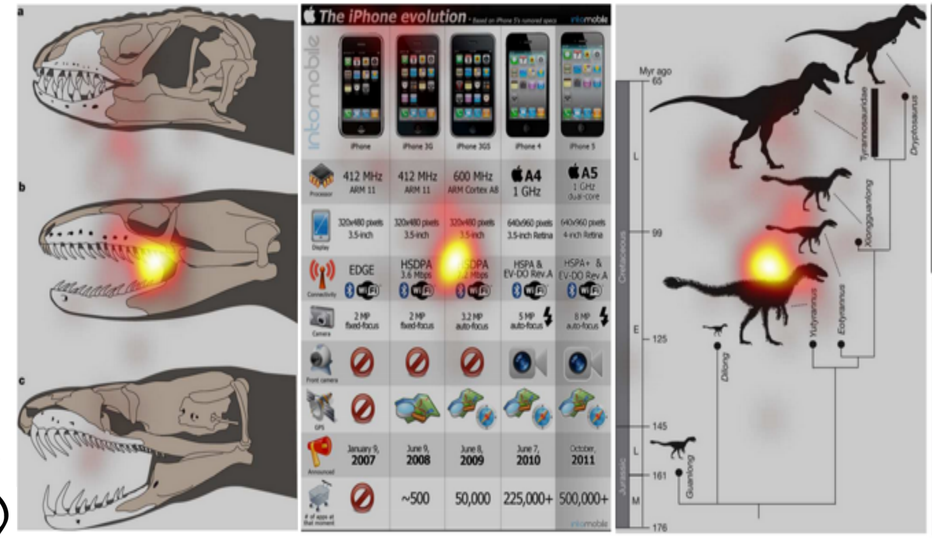


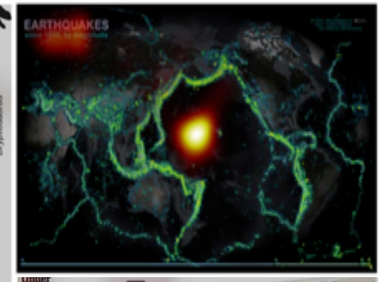


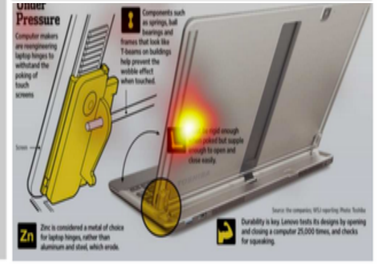


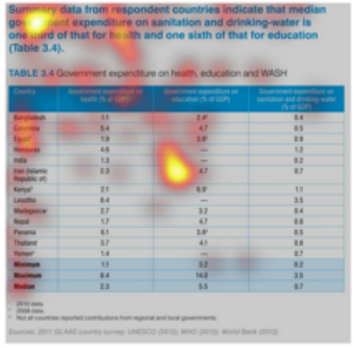


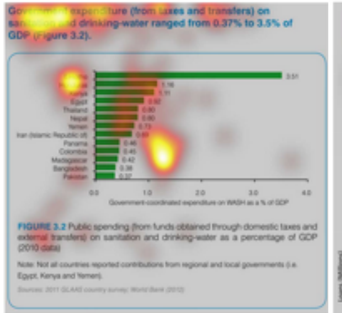


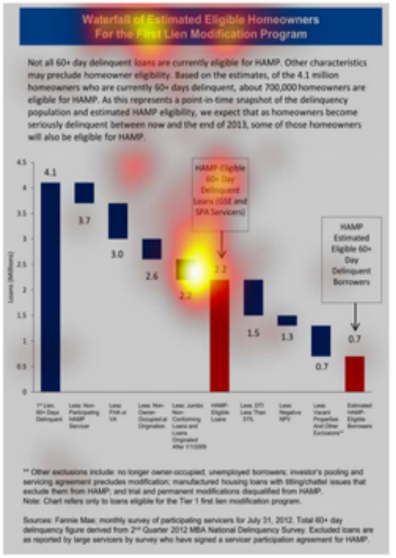












Most Memorable

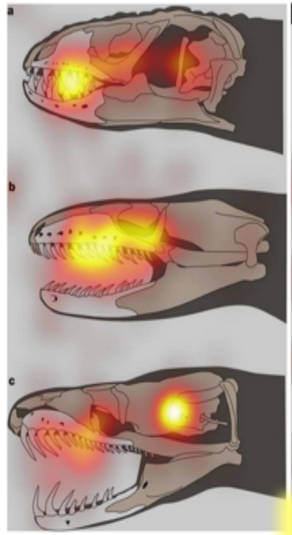
Least Memorable

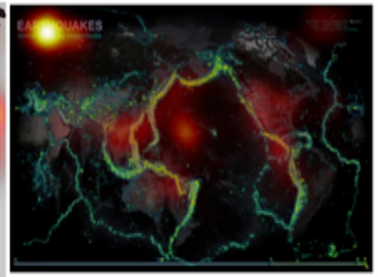
Encoding (10 sec)

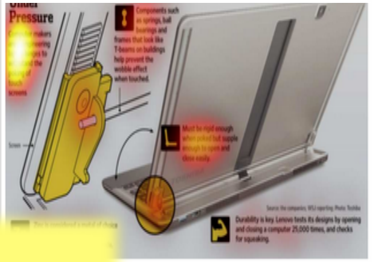
Recognition

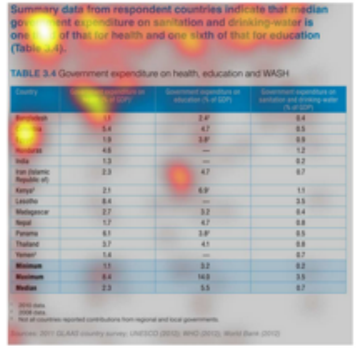
Visual Associations

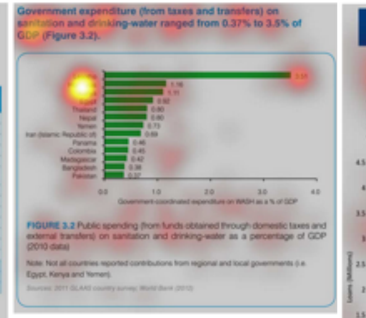
Semantic Associations

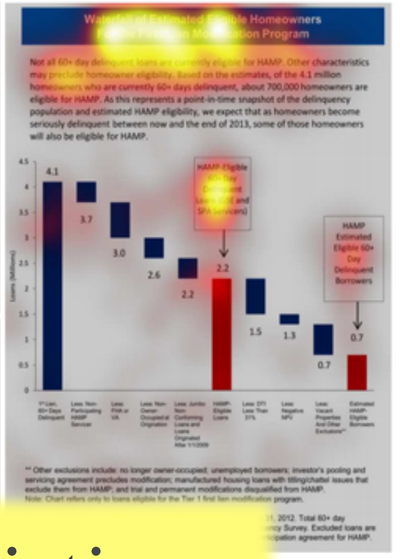


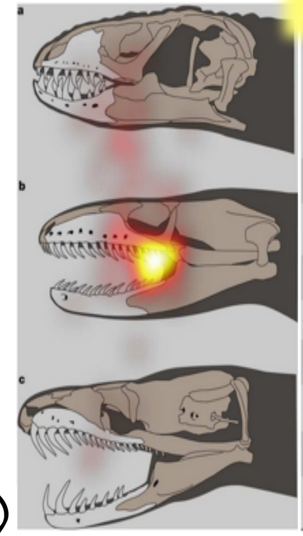


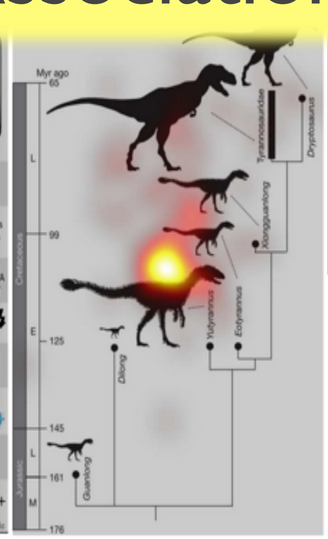


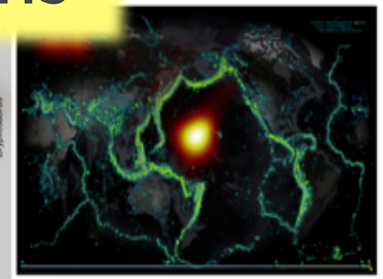


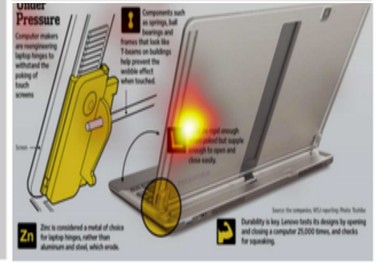


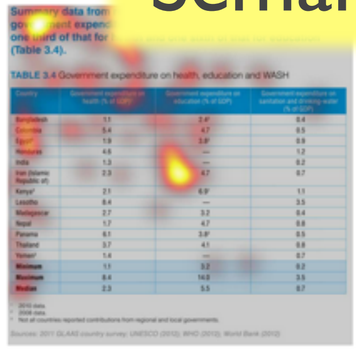


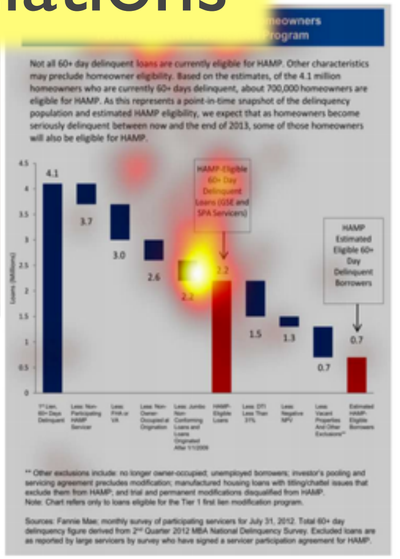


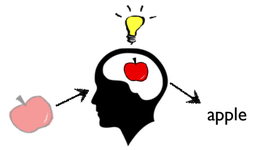




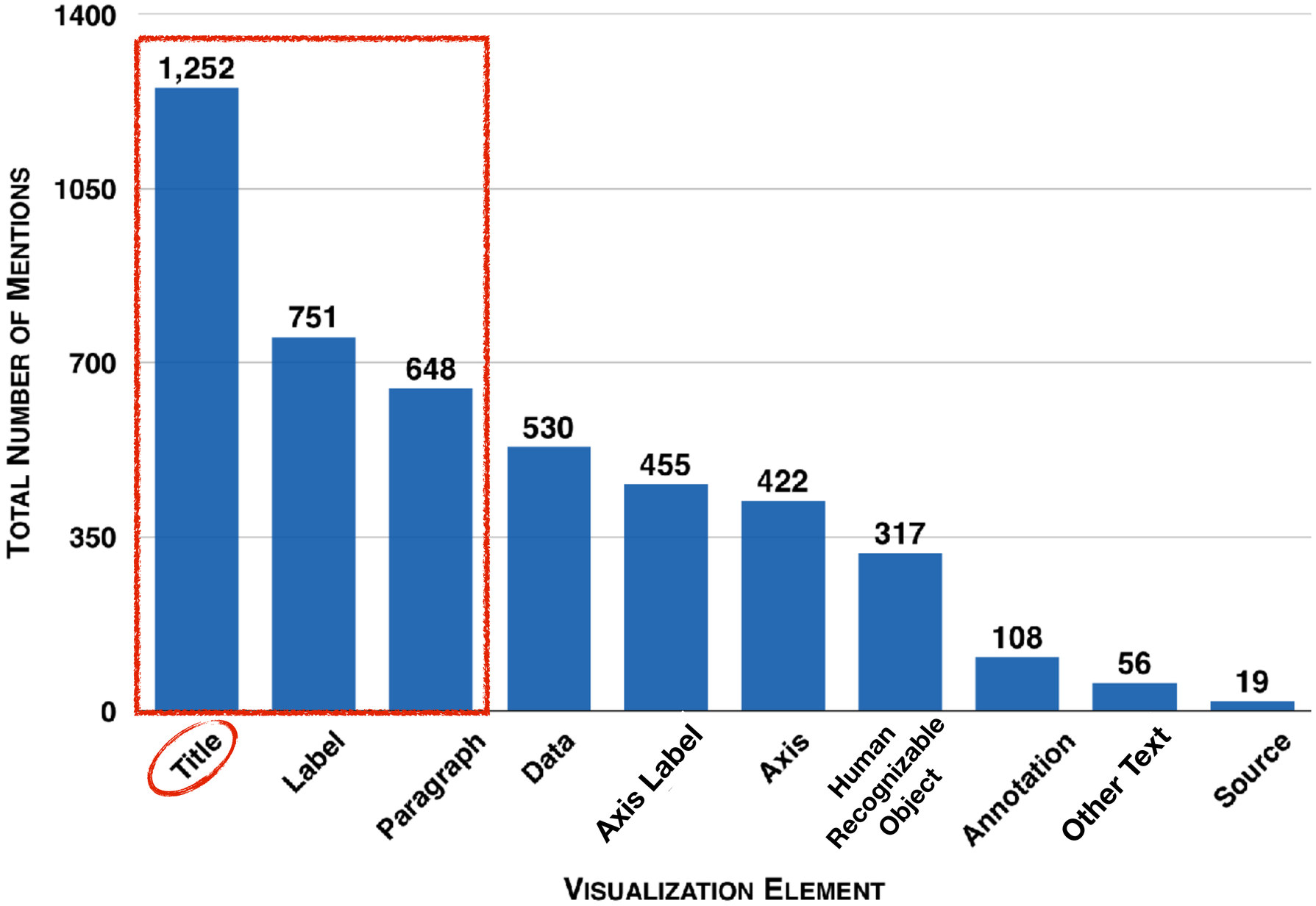


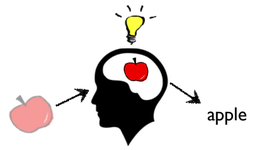




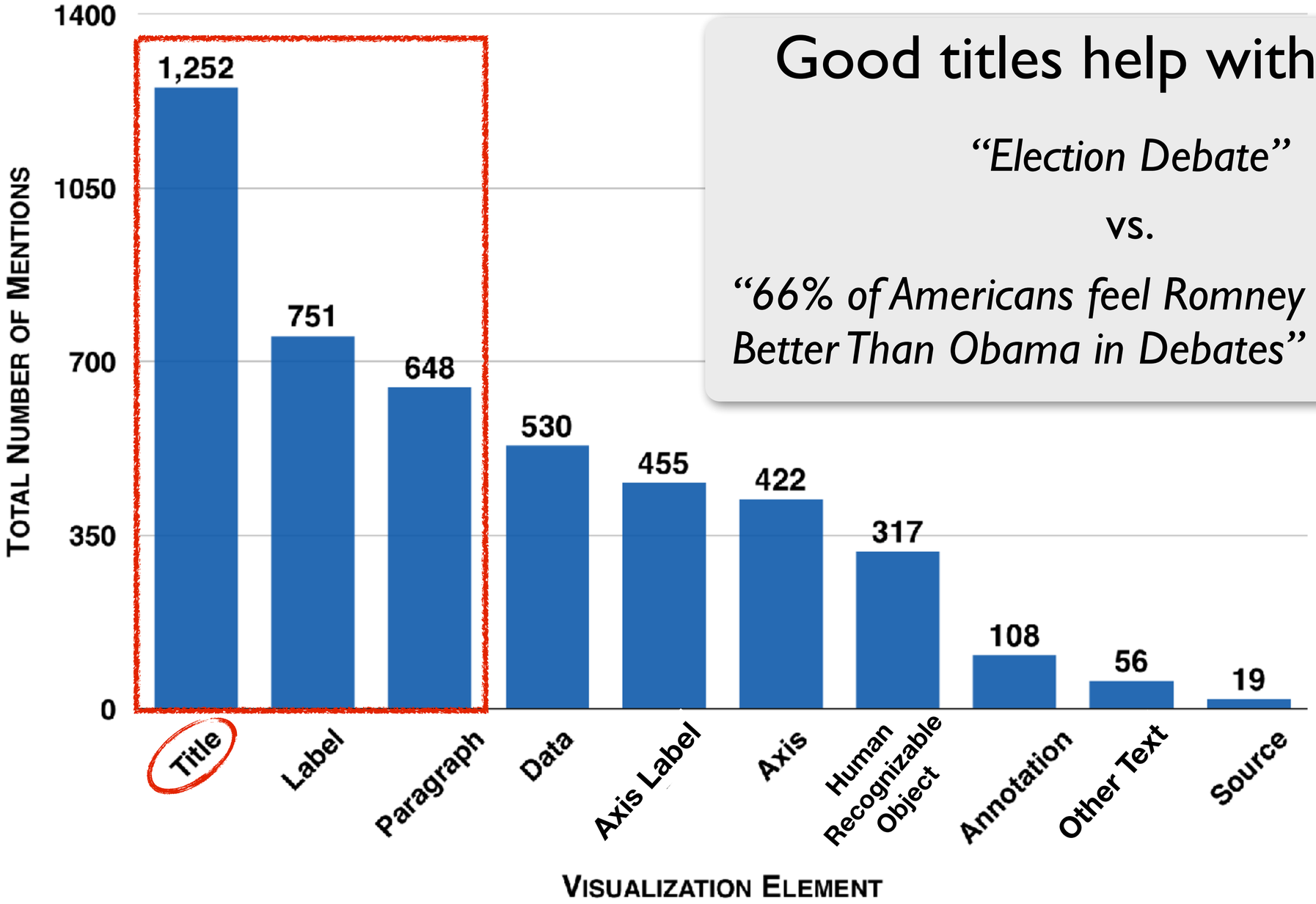

apple

RECALL


1400
1050
700
350
0
TOTAL NUMBER OF MENTIONS
1,252
751
648
530
455
422
317
108
56
19
Title
Label
Paragraph
Data
Axis Label
Axis
Human Recognizable Object
Annotation
Other Text
Source
VISUALIZATION ELEMENT



RECALL



Good titles help with recall!

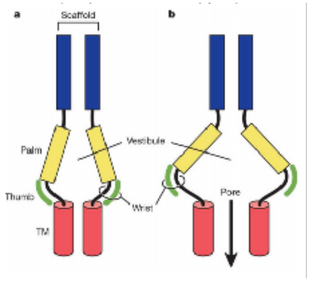
"Election Debate"

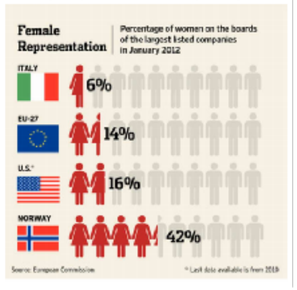
vs.

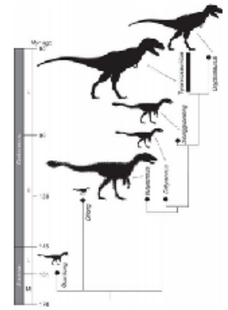
"66% of Americans feel Romney Performed Better Than Obama in Debates"

LOW QUALITY DESCRIPTION

HIGH QUALITY DESCRIPTION







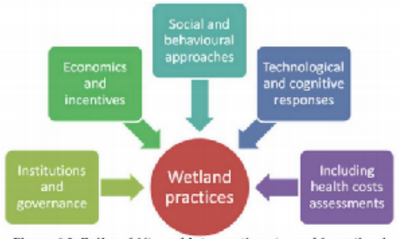
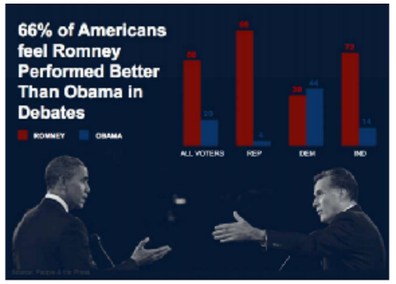
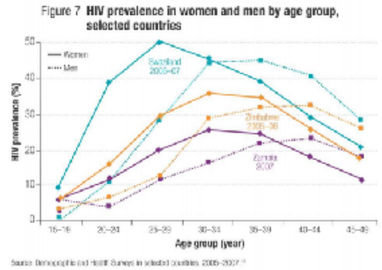
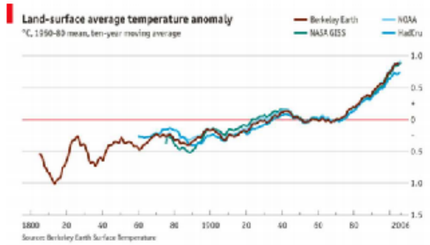
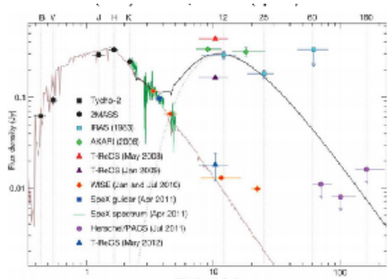


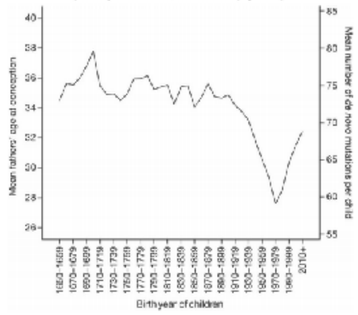
Figure 6.2: Policy shifts and interventions to enable wetland practices to accommodate notions of ecosystem services and human health

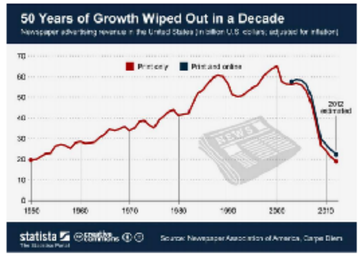


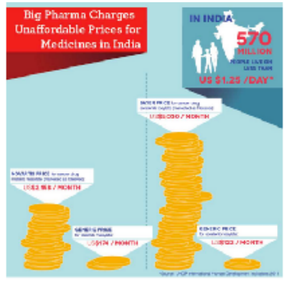


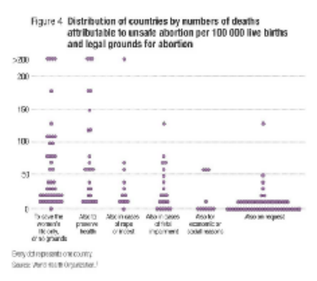










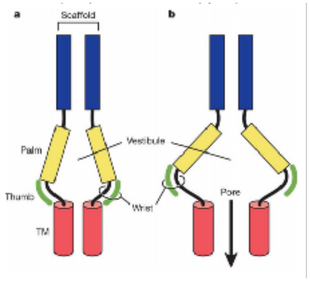


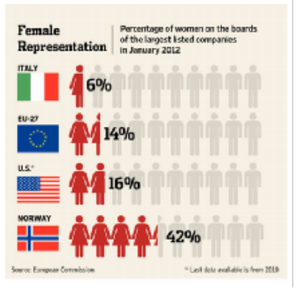
MEMORABLE

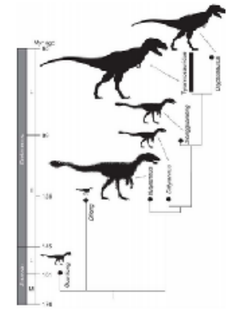
LOW QUALITY DESCRIPTION

HIGH QUALITY DESCRIPTION

FORGETTABLE







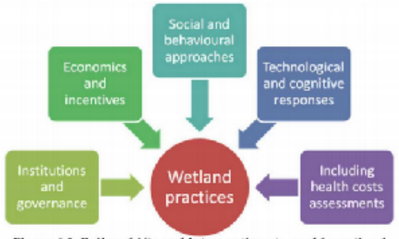
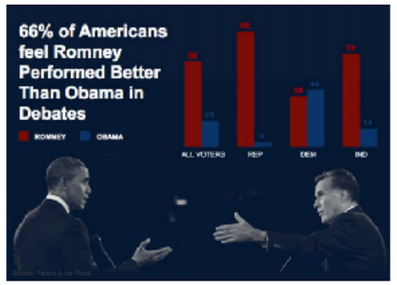
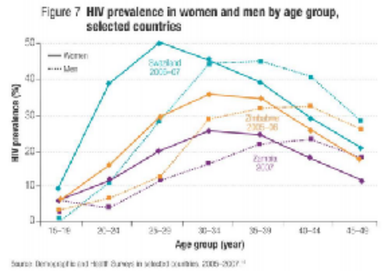
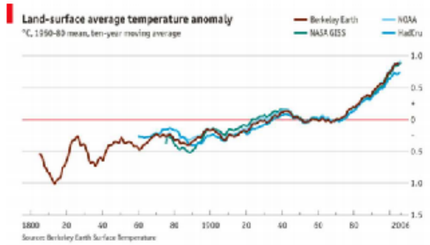
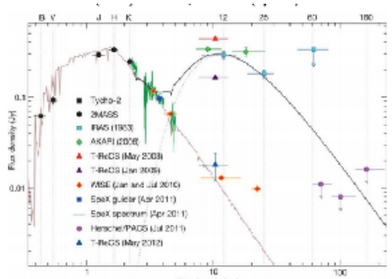


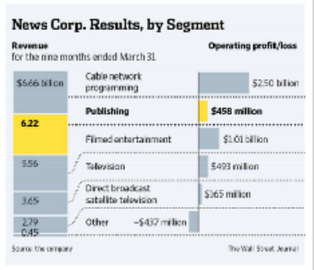
Figure 6.2: Policy shifts and interventions to enable wetland practices to accommodate notions of ecosystem services and human health

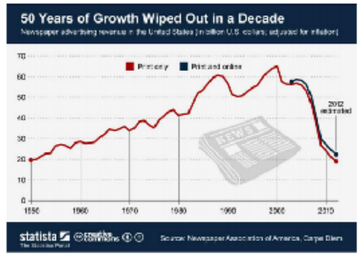


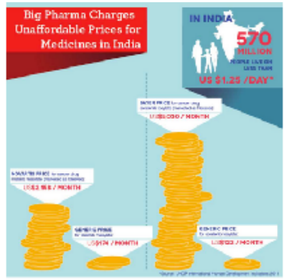


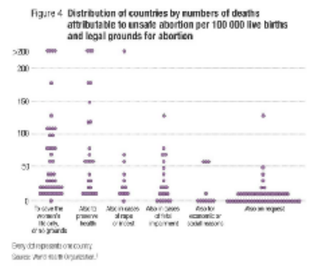












MEMORABLE

"Effective"

FORGETTABLE

LOW QUALITY DESCRIPTION

HIGH QUALITY DESCRIPTION

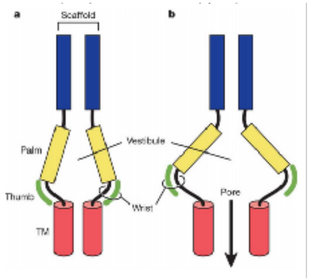
Visual Associations

Semantic Associations

Message redundancy

Data redundancy

Titles/annotations



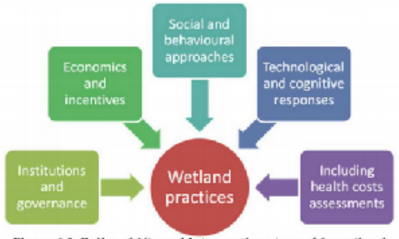
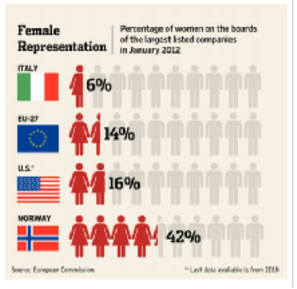
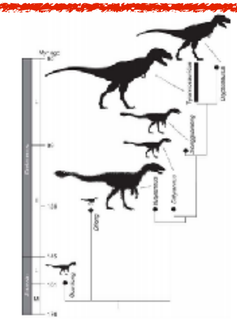
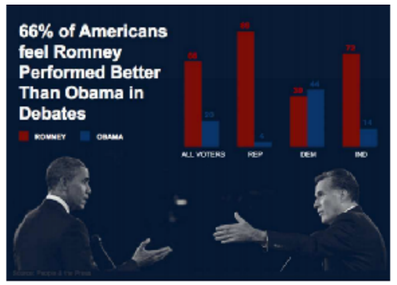
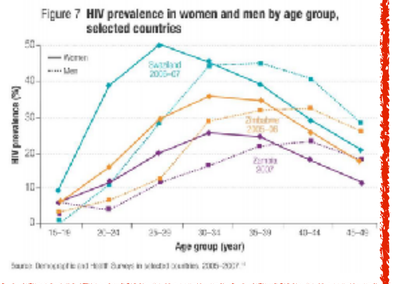


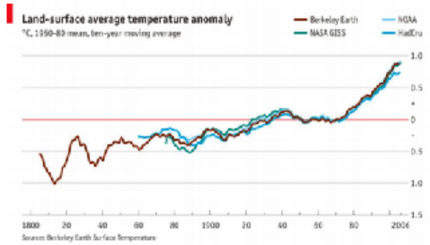
Figure 6.2: Policy shifts and interventions to enable wetland practices to accommodate notions of ecosystem services and human health

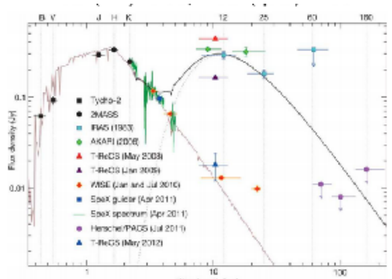


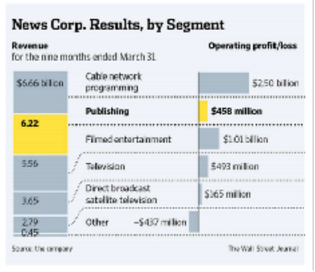


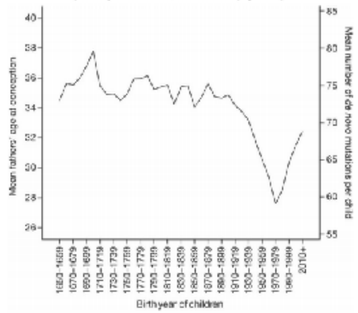


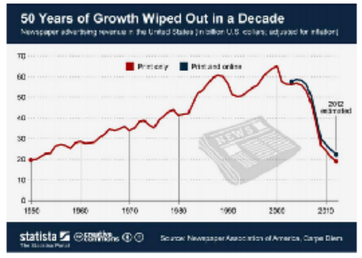


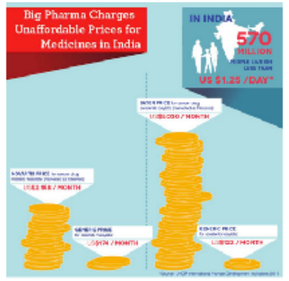


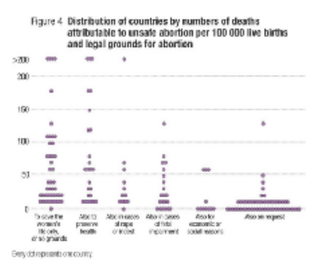




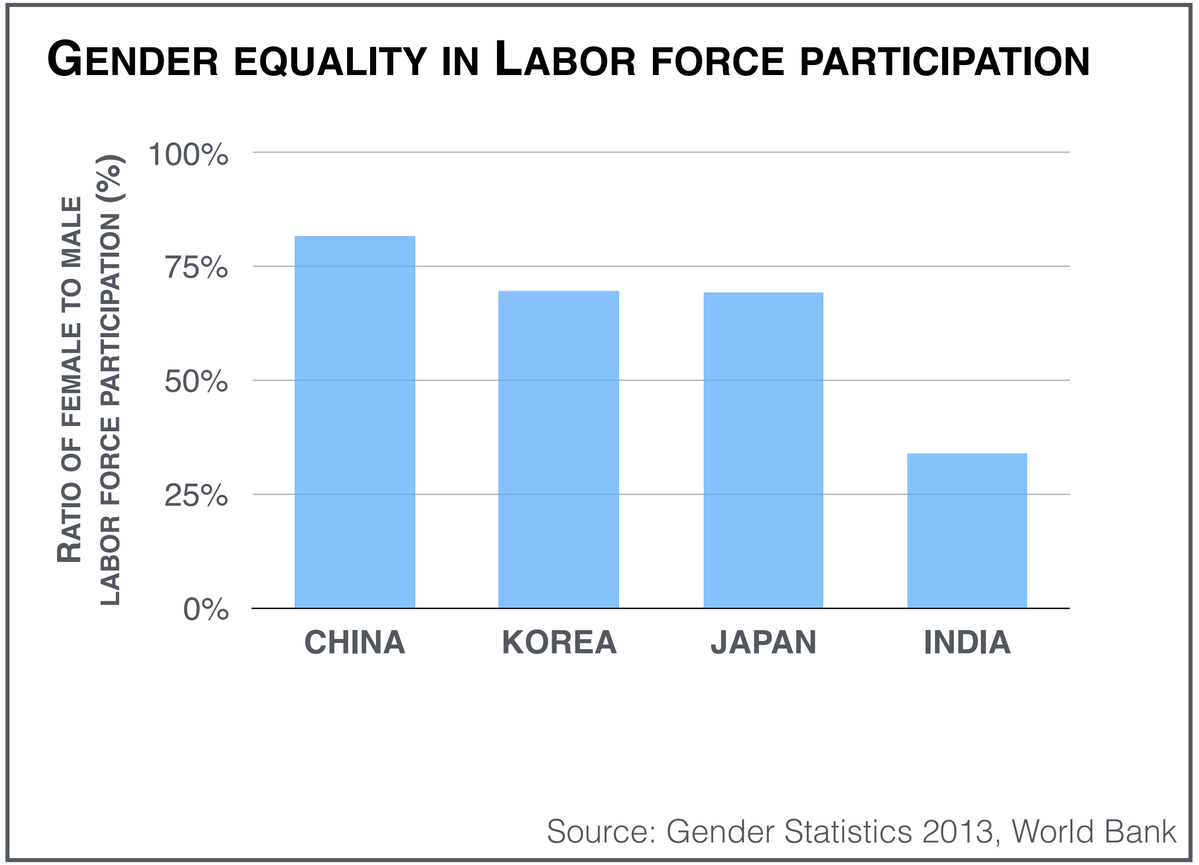




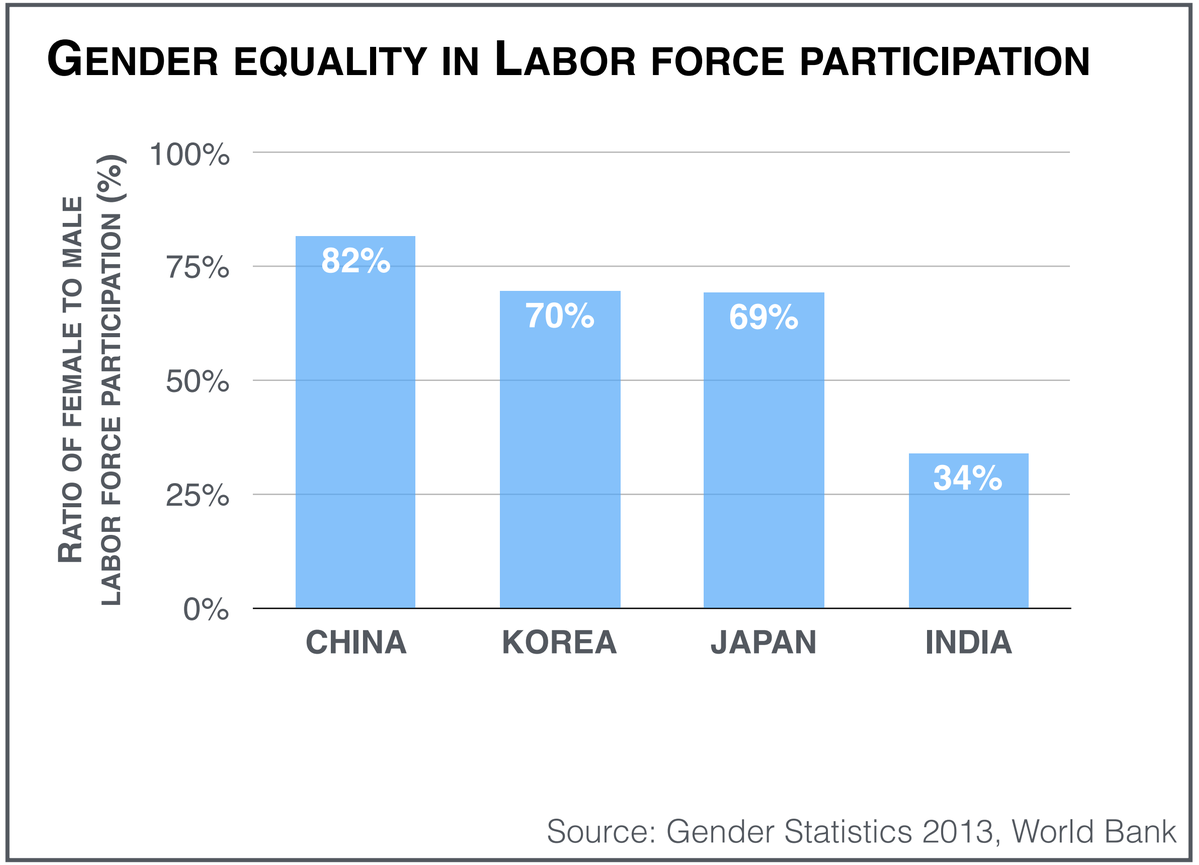




Examples


GENDER EQUALITY IN LABOR FORCE PARTICIPATION
RATIO OF FEMALE TO MALE LABOR FORCE PARTICIPATION (%)
100%
75%
50%
25%
0%
CHINA
KOREA
JAPAN
INDIA
Source: Gender Statistics 2013, World Bank

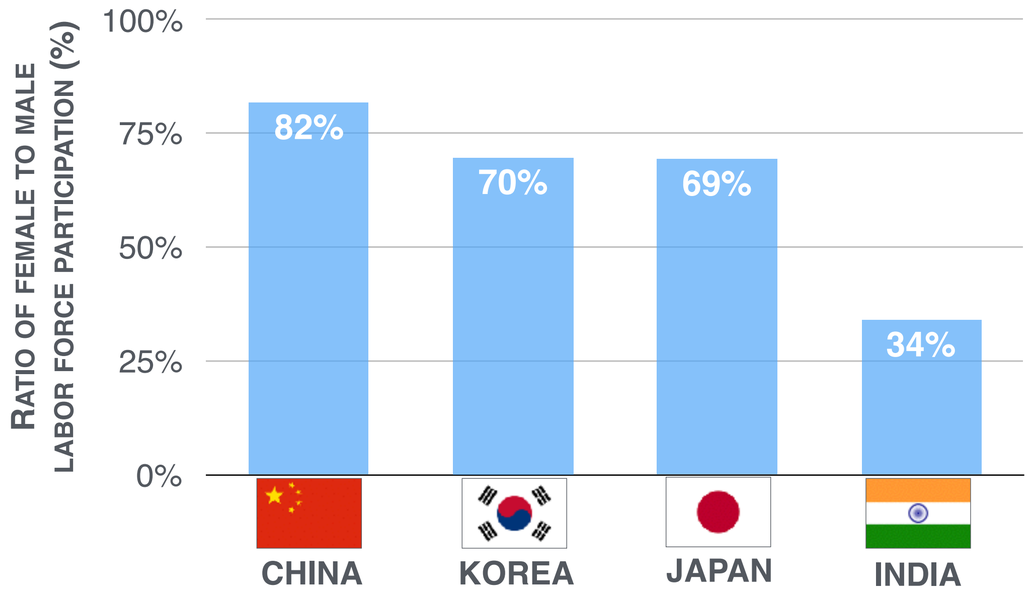
ORIGINAL



DATA REDUNDANCY

GENDER EQUALITY IN LABOR FORCE PARTICIPATION

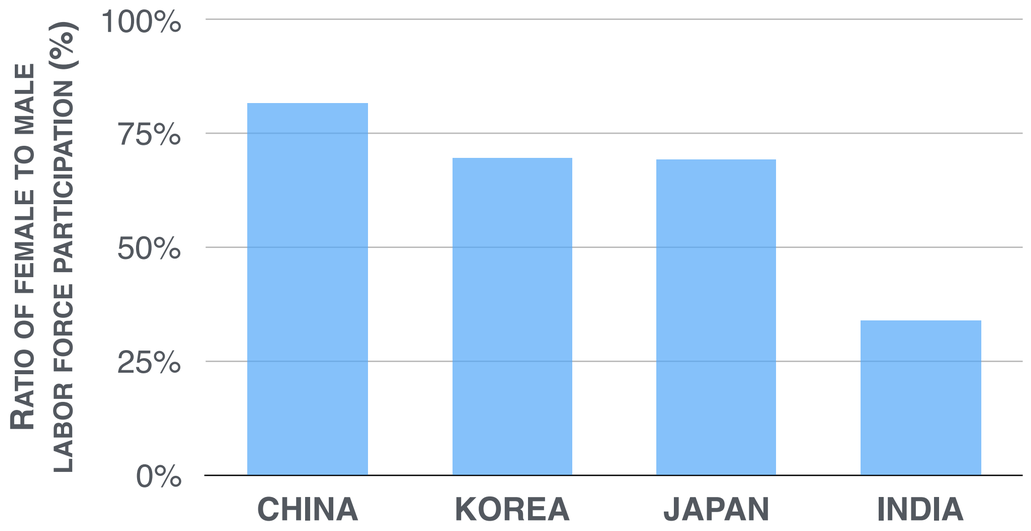
CHINA LEADS IN FEMALE LABOR FORCE PARTICIPATION WHEREAS INDIA LAGS SIGNIFICANTLY BEHIND AMONG OTHER ASIAN-PACIFIC COUNTRIES IN 2013.



Source: Gender Statistics 2013, World Bank

DATA & MESSAGE REDUNDANCY

GENDER EQUALITY IN LABOR FORCE PARTICIPATION

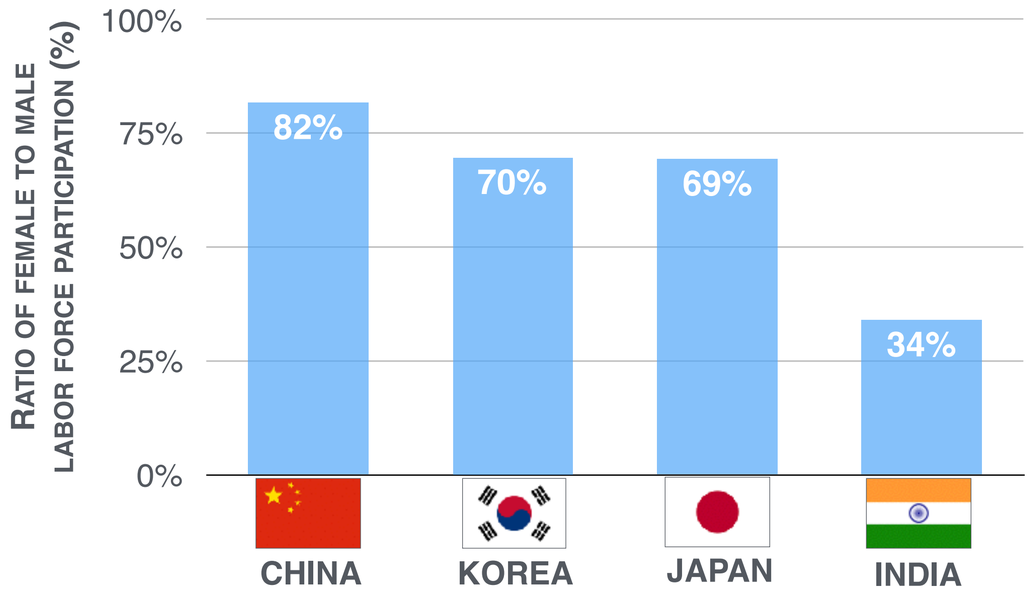


Source: Gender Statistics 2013, World Bank

ORIGINAL

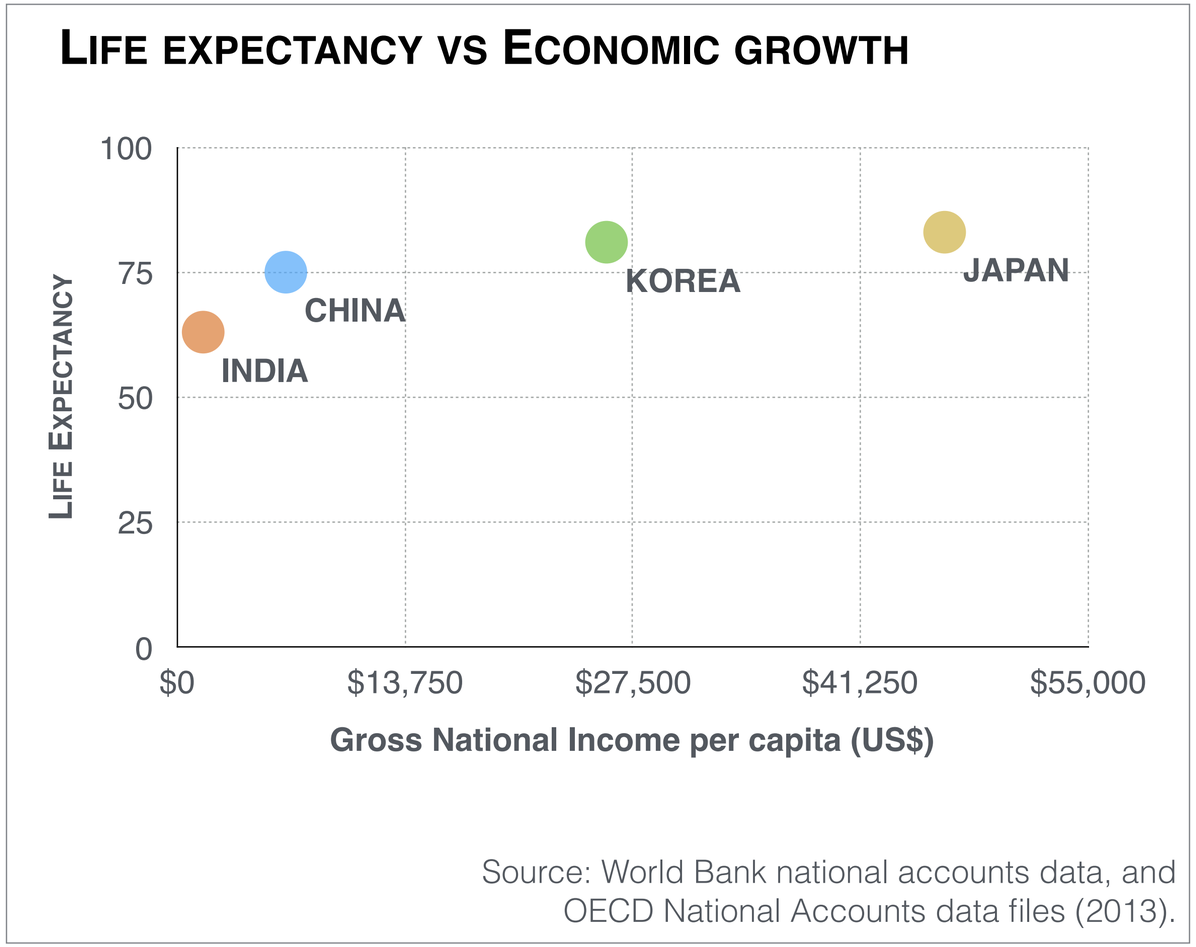
GENDER EQUALITY IN LABOR FORCE PARTICIPATION

CHINA LEADS IN FEMALE LABOR FORCE PARTICIPATION WHEREAS INDIA LAGS SIGNIFICANTLY BEHIND AMONG OTHER ASIAN-PACIFIC COUNTRIES IN 2013.

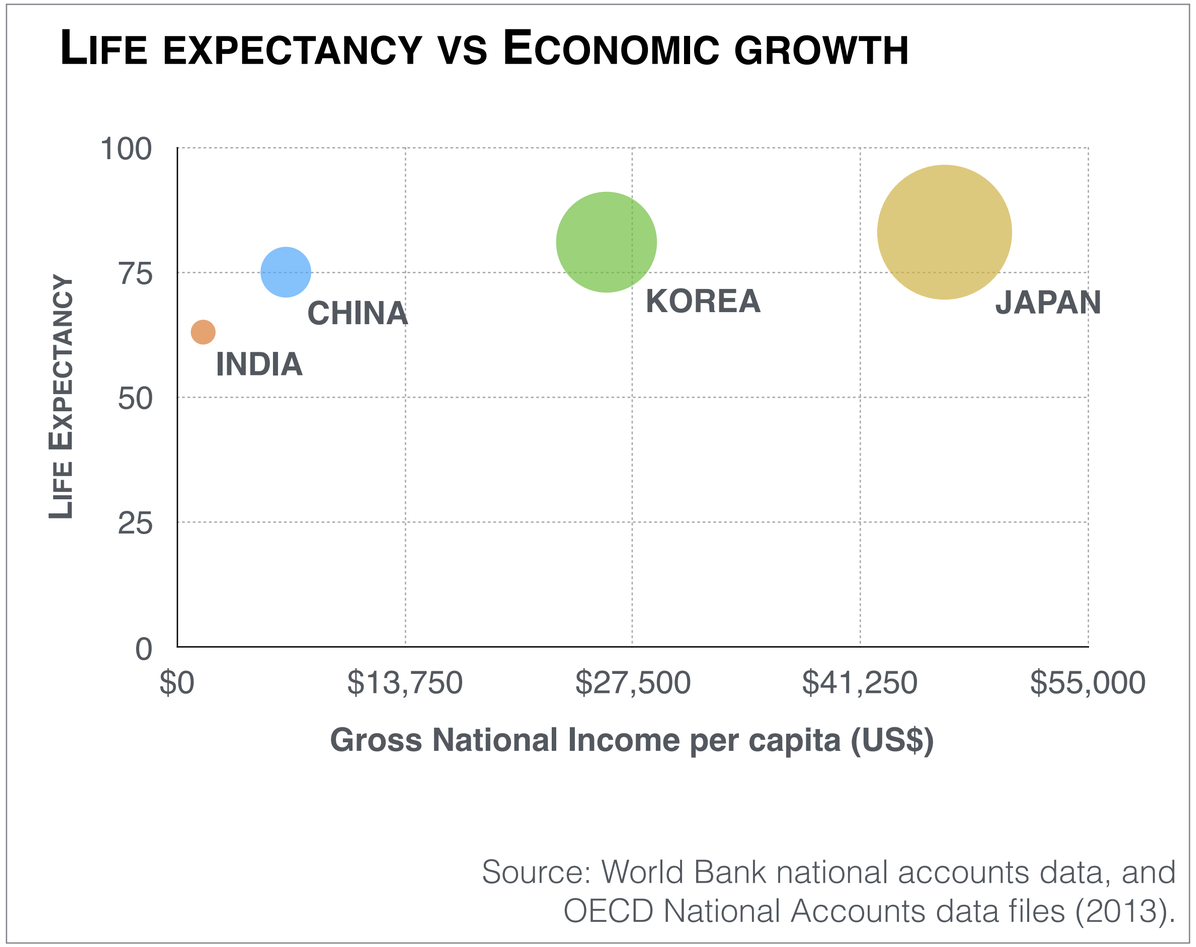


Source: Gender Statistics 2013, World Bank

DATA & MESSAGE REDUNDANCY


LIFE EXPECTANCY VS ECONOMIC GROWTH
100
75
50
25
0
LIFE EXPECTANCY
CHINA
INDIA
KOREA
JAPAN
$0
$13,750
$27,500
$41,250
$55,000
Gross National Income per capita (US$)
Source: World Bank national accounts data, and
OECD National Accounts data files (2013).

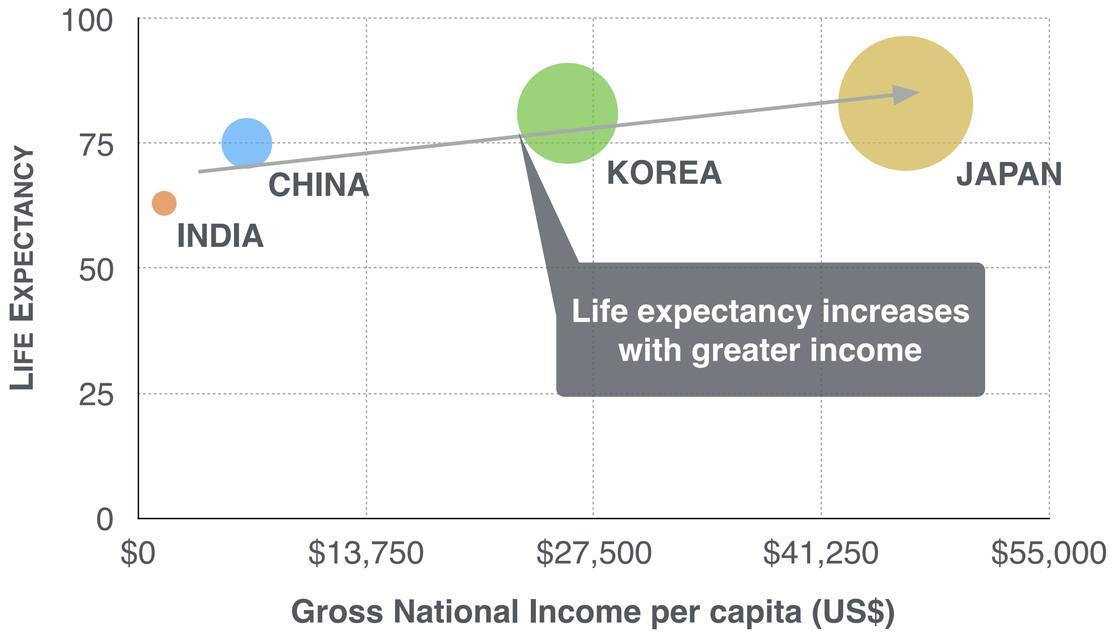
ORIGINAL



DATA REDUNDANCY

LIFE EXPECTANCY VS ECONOMIC GROWTH

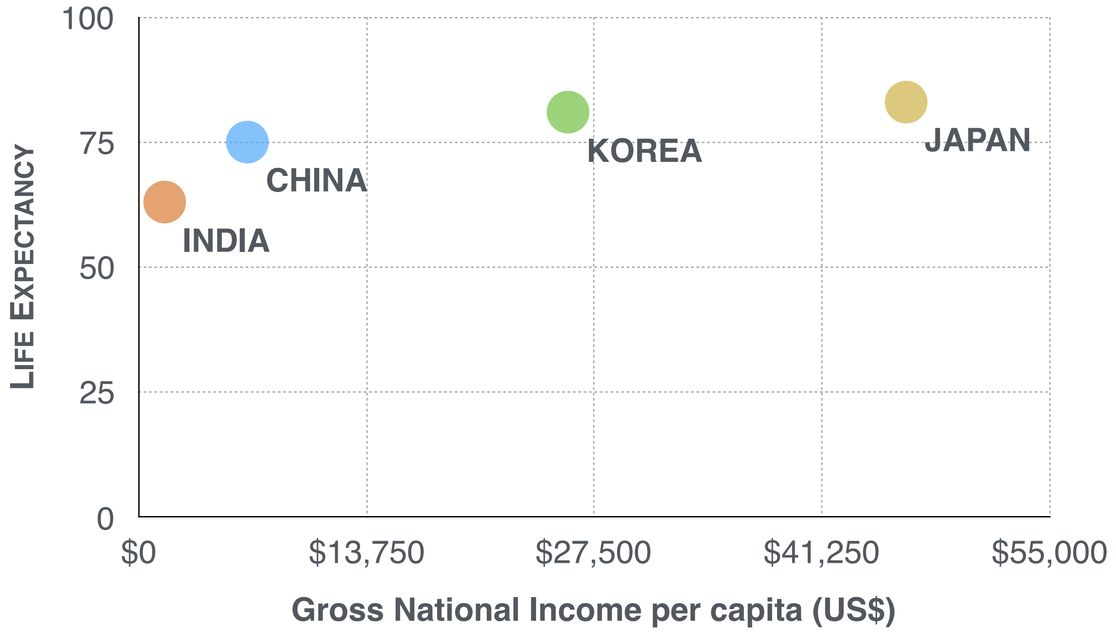
RELATION BETWEEN LIFE EXPECTANCY AND GROSS NATIONAL INCOME OF ASIAN-PACIFIC COUNTRIES IN 2013



Source: World Bank national accounts data, and OECD National Accounts data files (2013).

DATA & MESSAGE REDUNDANCY

LIFE EXPECTANCY VS ECONOMIC GROWTH

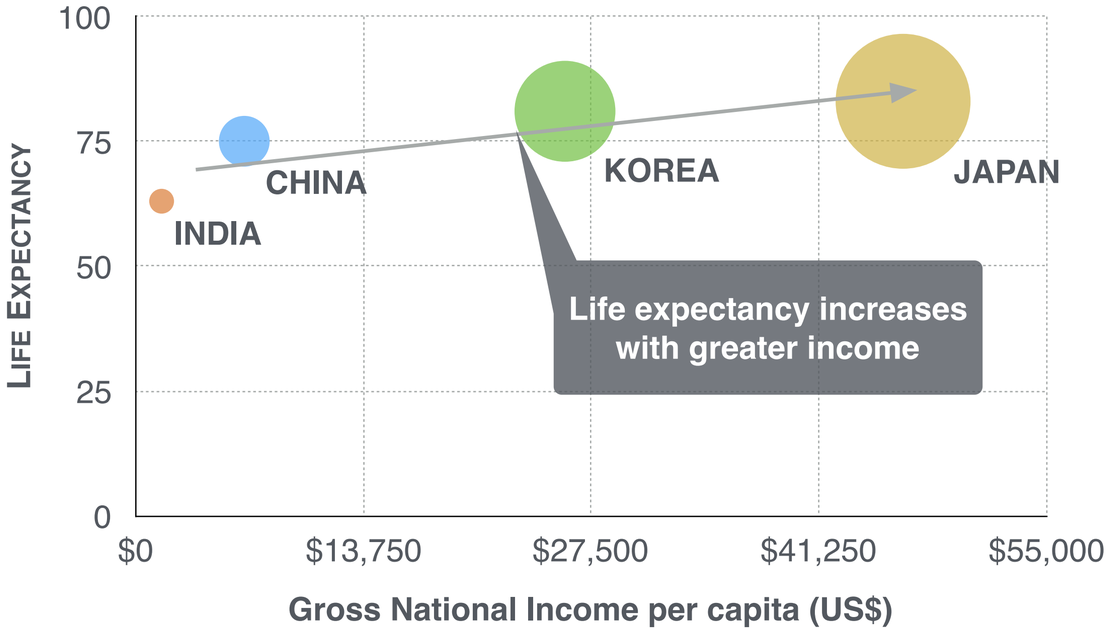


Source: World Bank national accounts data, and OECD National Accounts data files (2013).

ORIGINAL

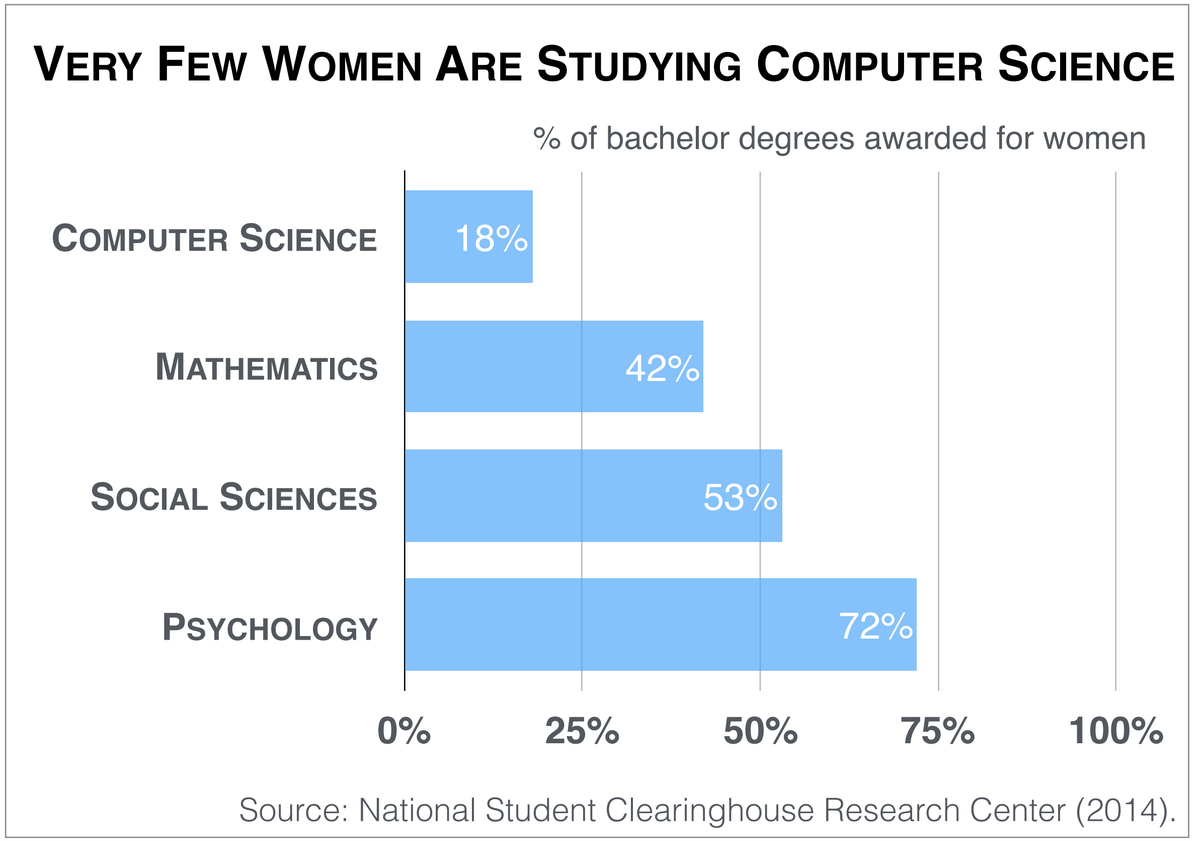
LIFE EXPECTANCY VS ECONOMIC GROWTH

RELATION BETWEEN LIFE EXPECTANCY AND GROSS NATIONAL INCOME OF ASIAN-PACIFIC COUNTRIES IN 2013

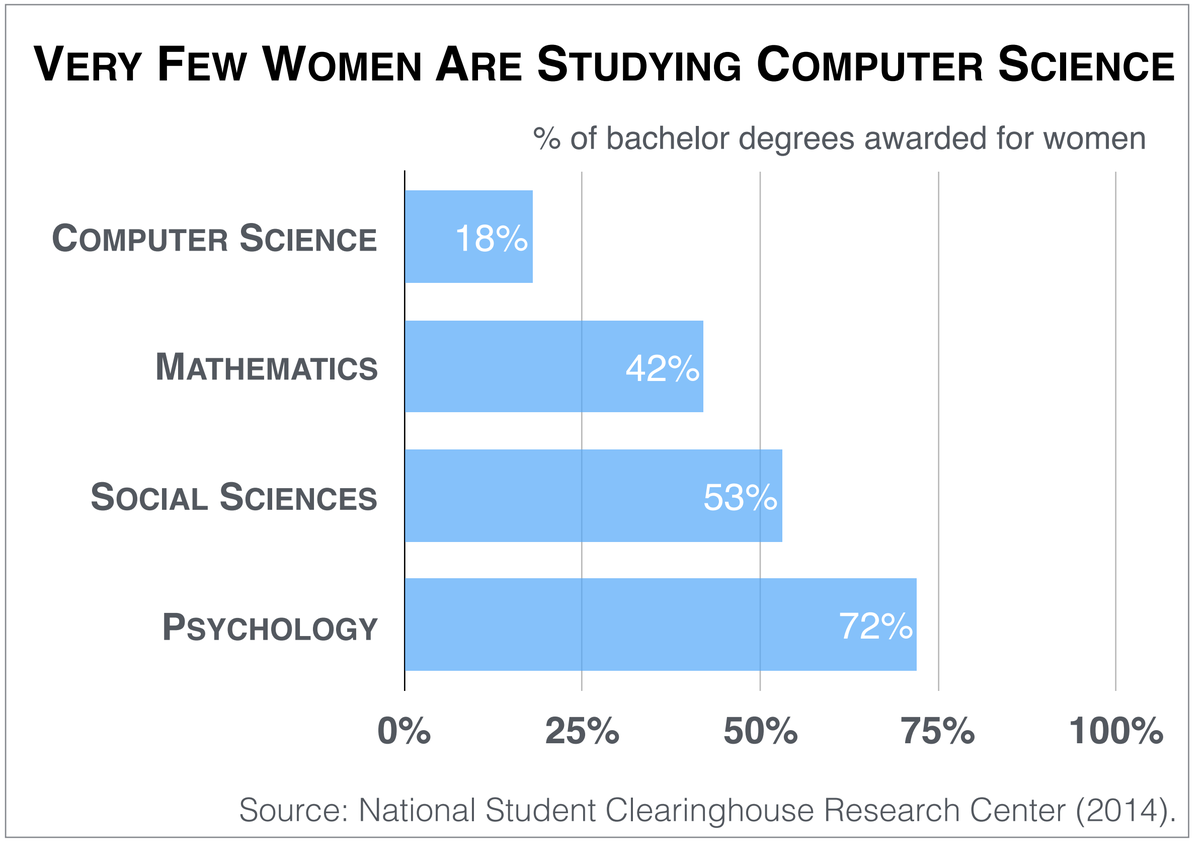


Source: World Bank national accounts data, and OECD National Accounts data files (2013).

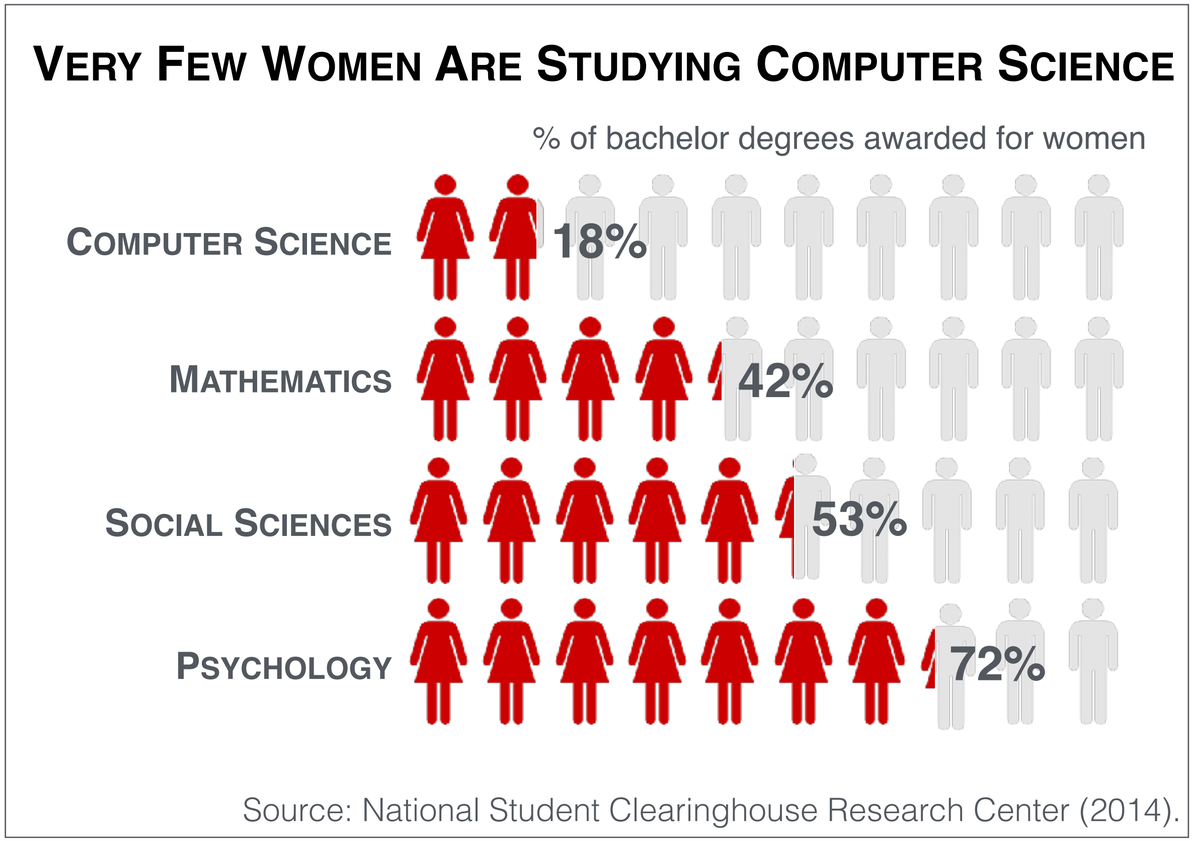
DATA & MESSAGE REDUNDANCY



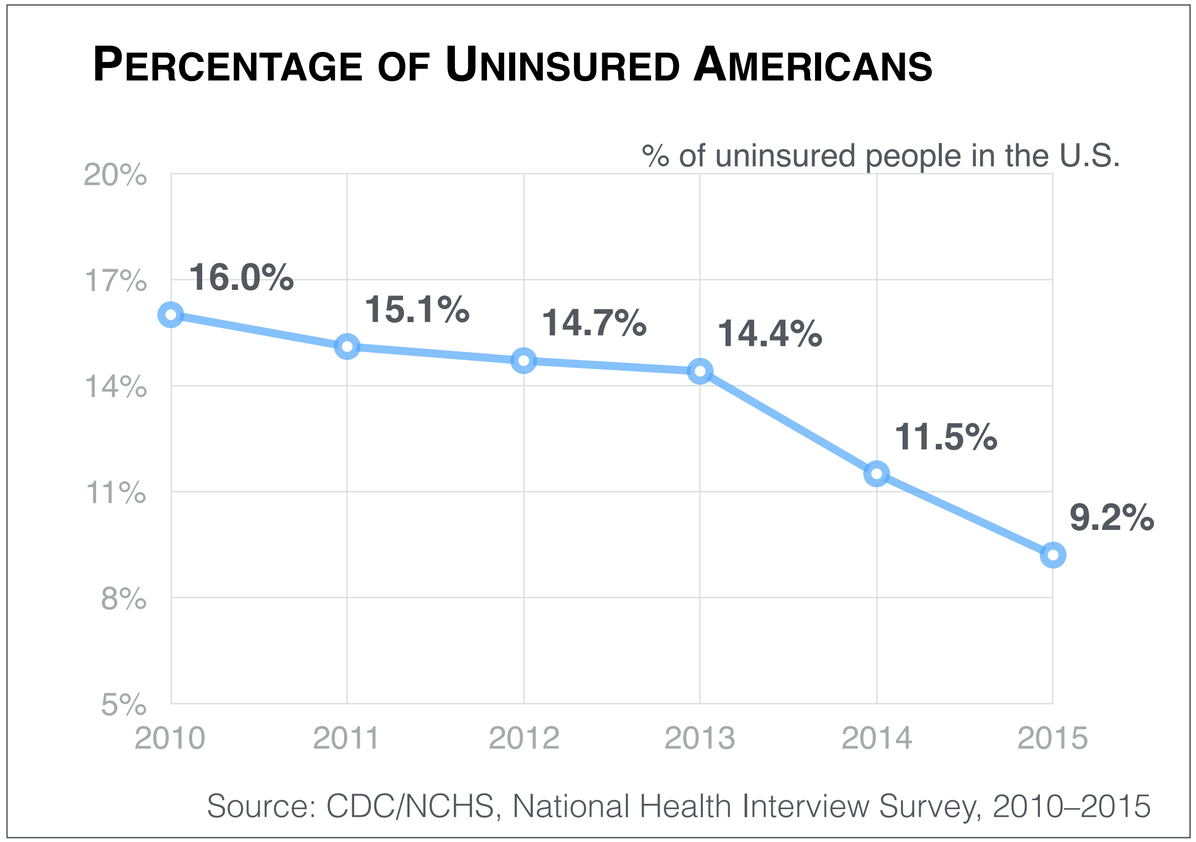
WITHOUT PICTOGRAM



WITHOUT PICTOGRAM

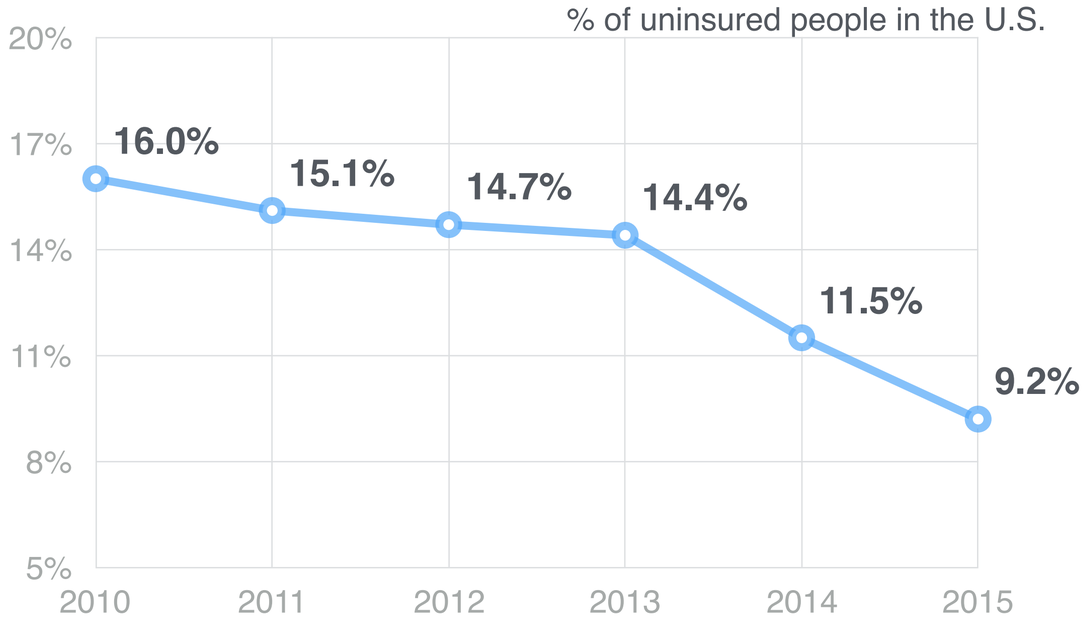


WITH PICTOGRAM


PERCENTAGE OF UNINSURED AMERICANS
% of uninsured people in the U.S.
20%
17%
14%
11%
8%
5%
16.0%
15.1%
14.7%
14.4%
11.5%
9.2%
2010
2011
2012
2013
2014
2015
Source: CDC/NCHS, National Health Interview Survey, 2010–2015

BAD TITLE

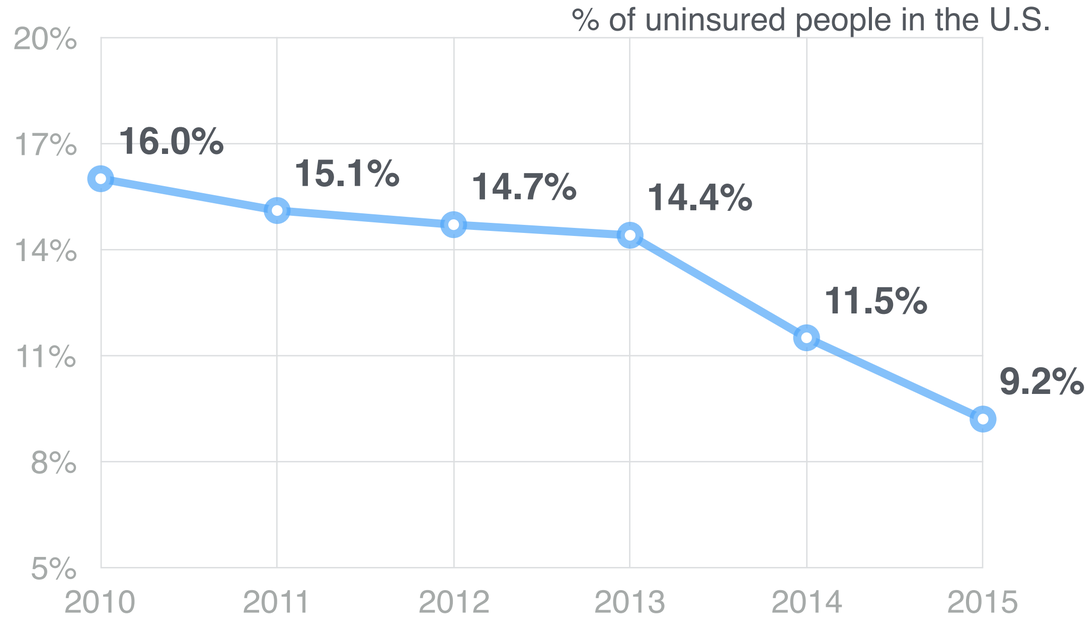
Percentage of Uninsured Americans



Source: CDC/NCHS, National Health Interview Survey, 2010–2015

Bad Title

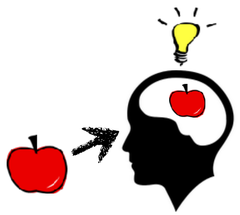
America's Uninsured Rate Dips below 10%



Source: CDC/NCHS, National Health Interview Survey, 2010–2015

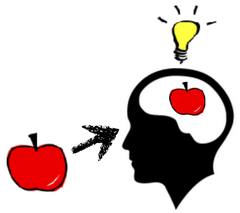
Good Title

Summary

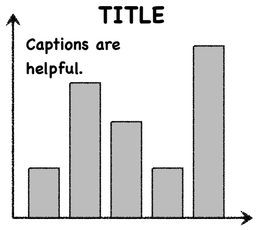


Visualizations that are memorable “at-a-glance” have memorable content.

Summary

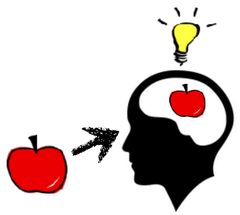


Visualizations that are memorable “at-a-glance” have memorable content.

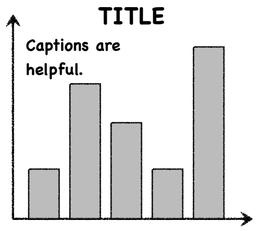


Titles and text are key elements in a visualization and help recall the message.

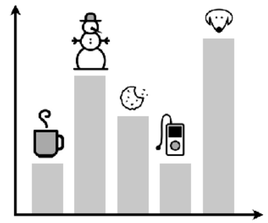
Summary



Visualizations that are memorable “at-a-glance” have memorable content.

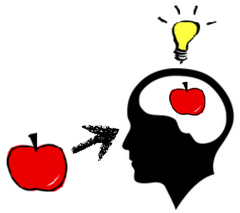


Titles and text are key elements in a visualization and help recall the message.

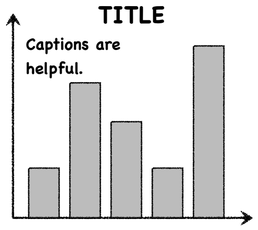


Human recognizable objects (e.g., pictograms) can help with the recognition or recall of a visualization.

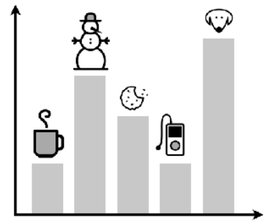
Summary



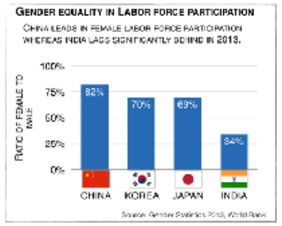
Visualizations that are memorable "at-a-glance" have memorable content.



Titles and text are key elements in a visualization and help recall the message.



Human recognizable objects (e.g., pictograms) can help with the recognition or recall of a visualization.

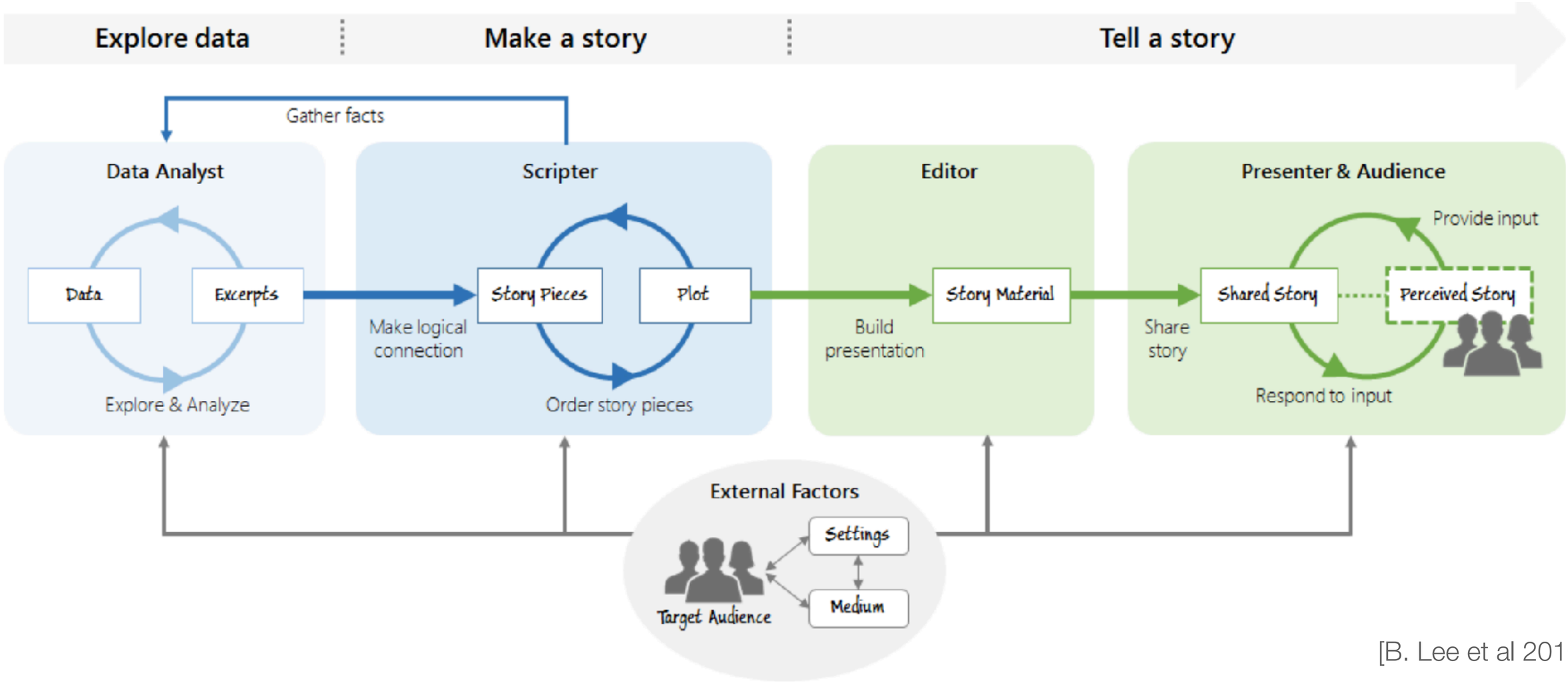


Redundancy helps with visualization recall and understanding.

A single picture cannot tell compelling stories

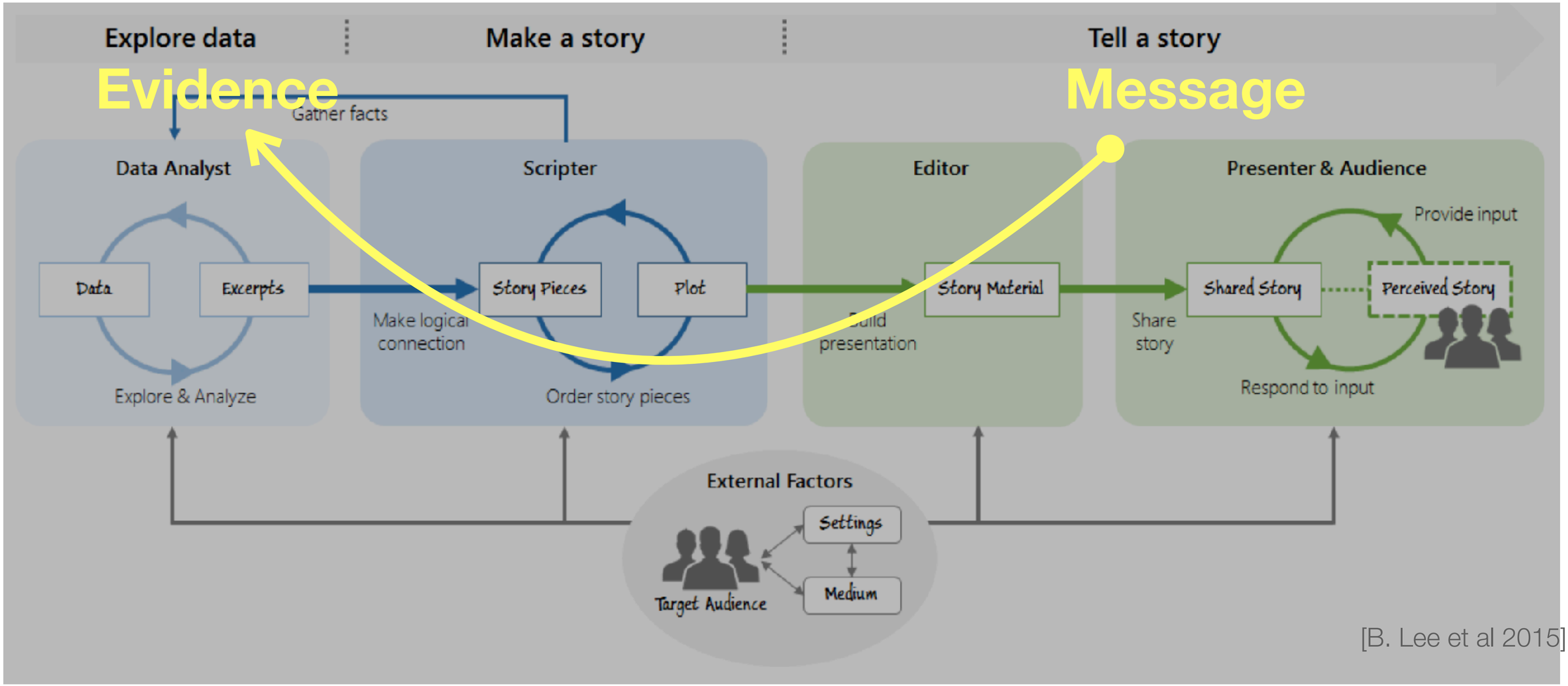
Data-Driven Storytelling

Storytelling process: transforming data into visually shared stories.

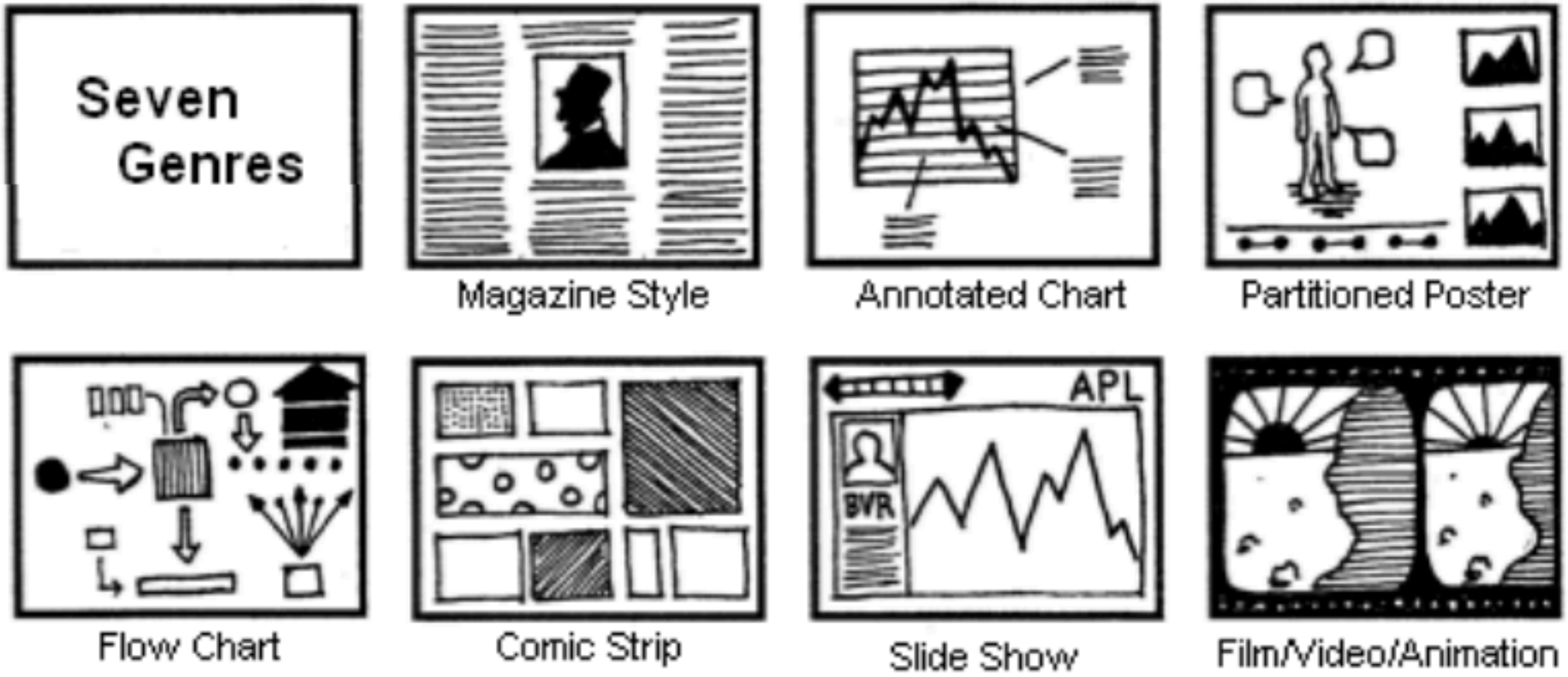


[B. Lee et al 2015]

In reality…



[B. Lee et al 2015]

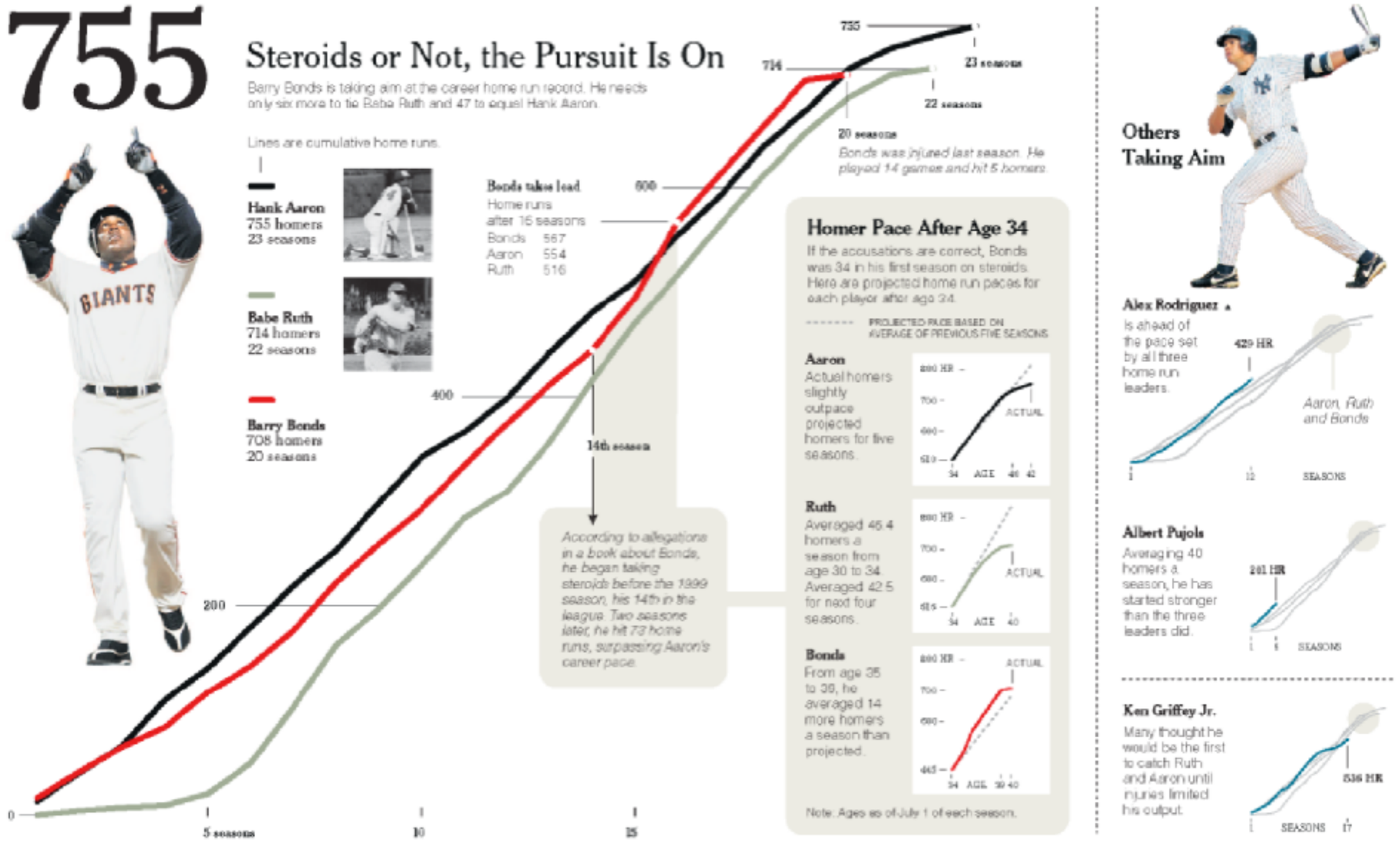


Mostly combined…

[Segel & Heer 2010]

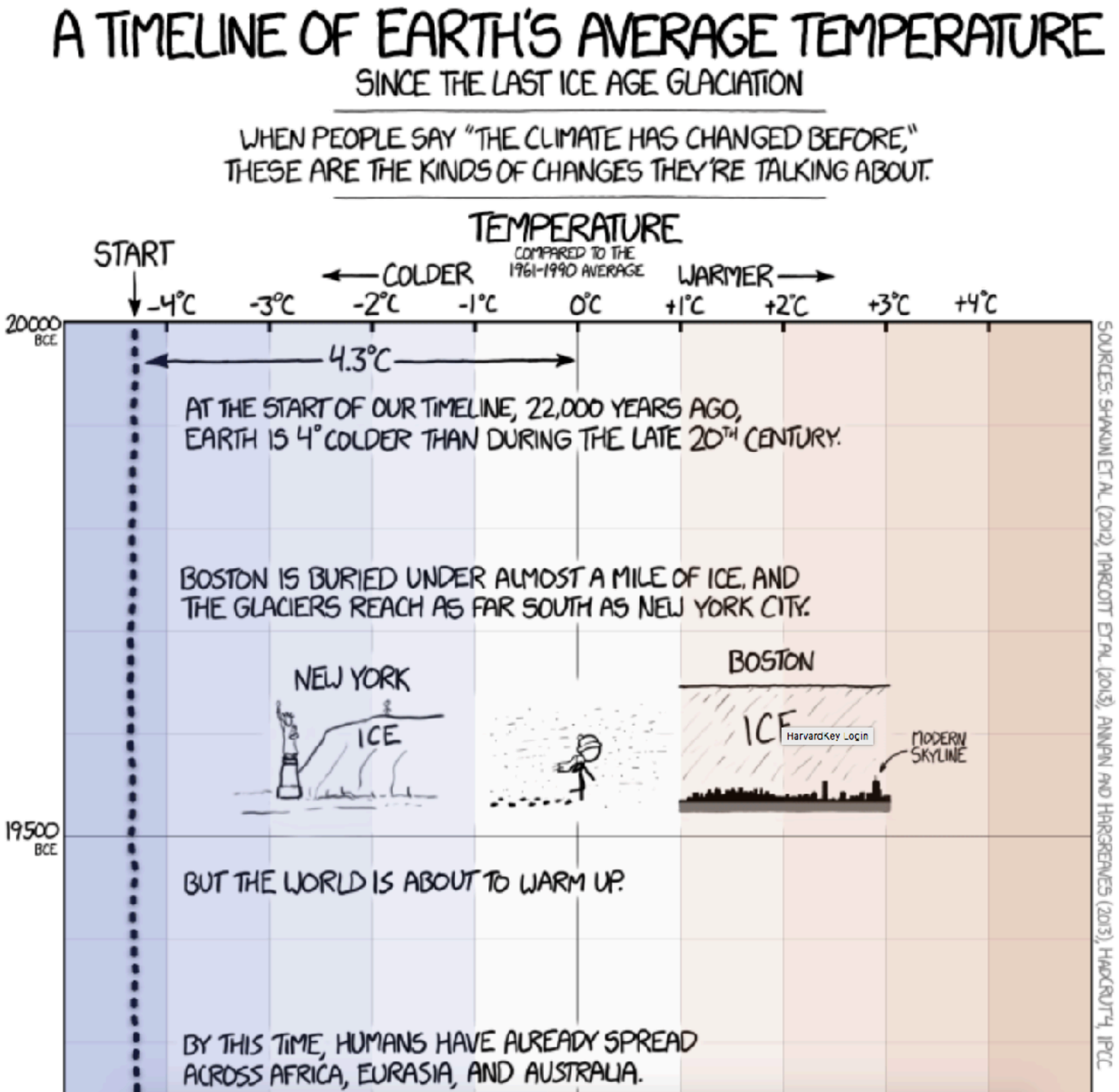
Examples

Magazine



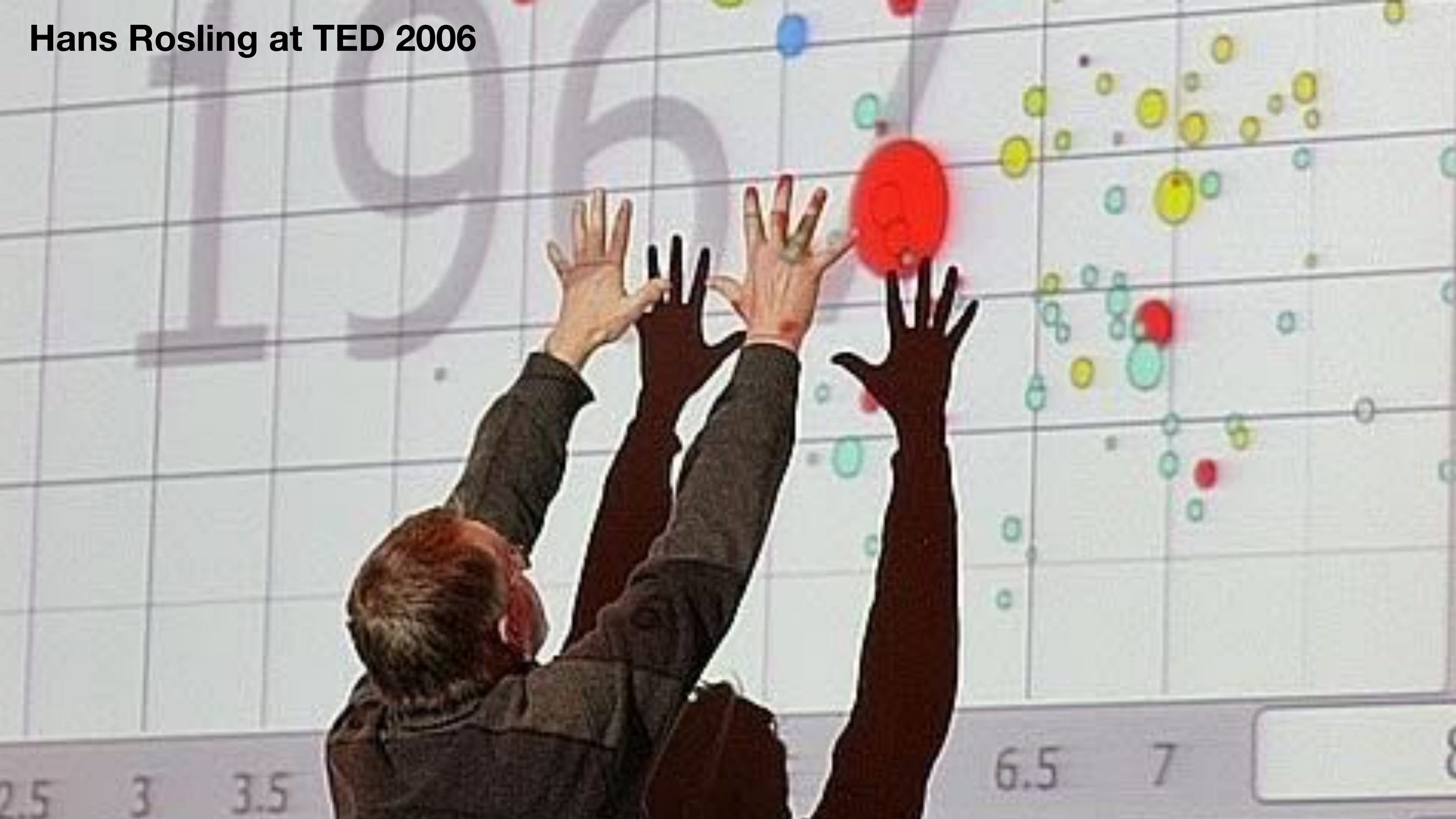
New York Times 2006

Comics



https://xkcd.com/1732/

Hans Rosling at TED 2006



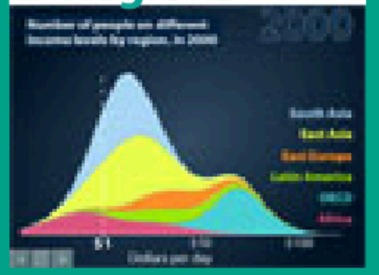
Human Development Trends 2005

Interactive presentation of some of the messages in the Human Development Report 2005

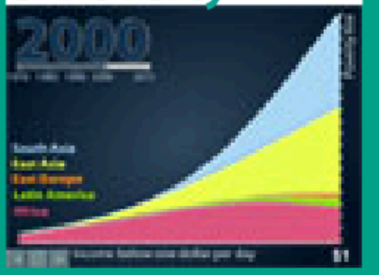
English
Dansk
Portuguese
Suomi
Français
Deutsch

1 Income

2 Regions



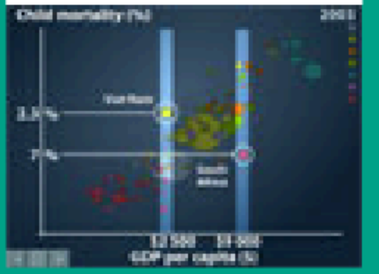
3 Poverty



4 Health

5 Countries

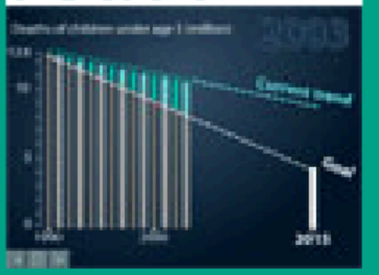
6 Differences



7 Trends

8 Gaps

9 Deaths

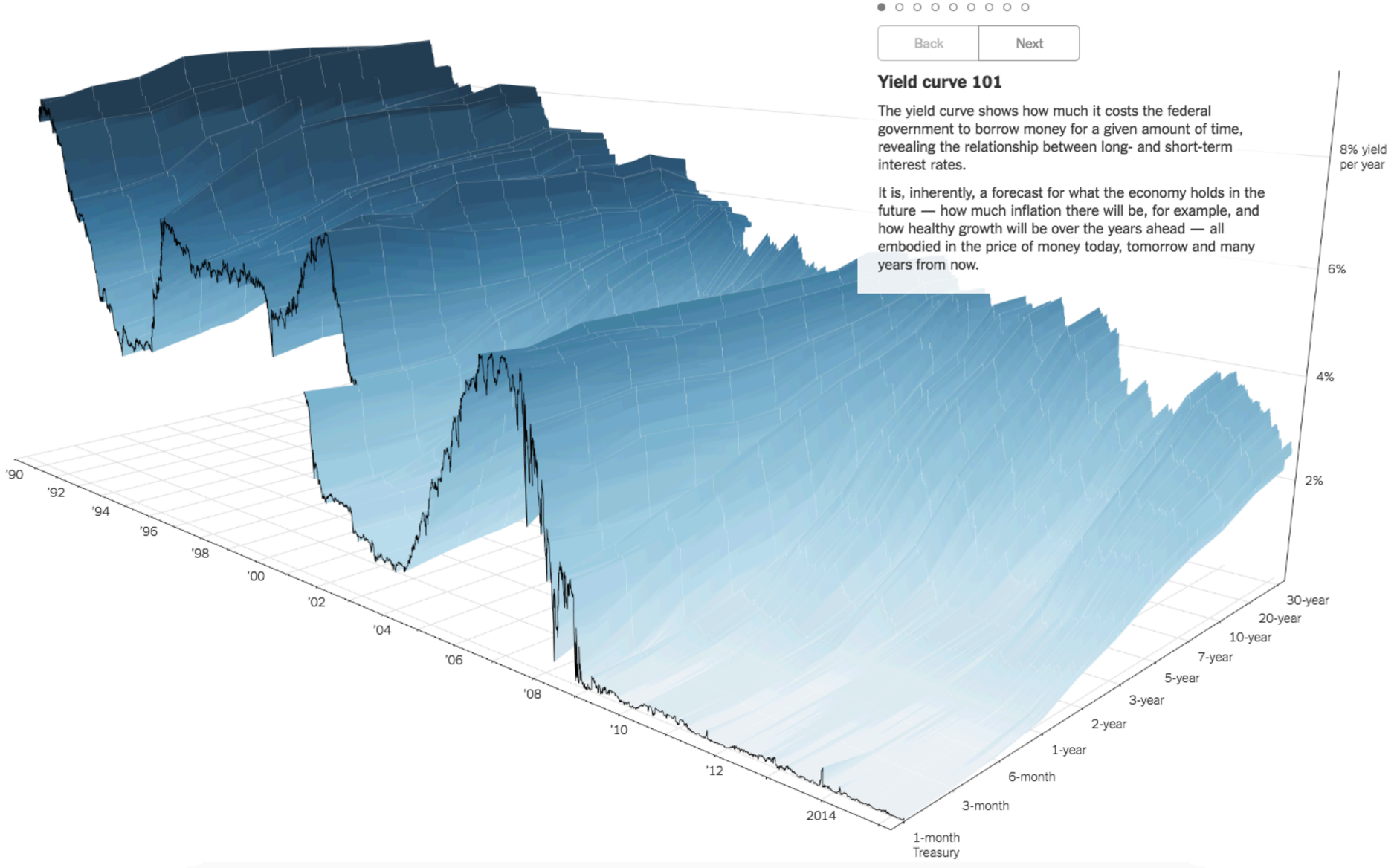


Produced in collaboration with:
GAPMINDER
www.gapminder.org

English translation: Claes Johansson, UNDP

A 3-D View of a Chart That Predicts The Economic Future: The Yield Curve

By **GREGOR AISCH** and **AMANDA COX** MARCH 18, 2015



Published: February 2, 2010

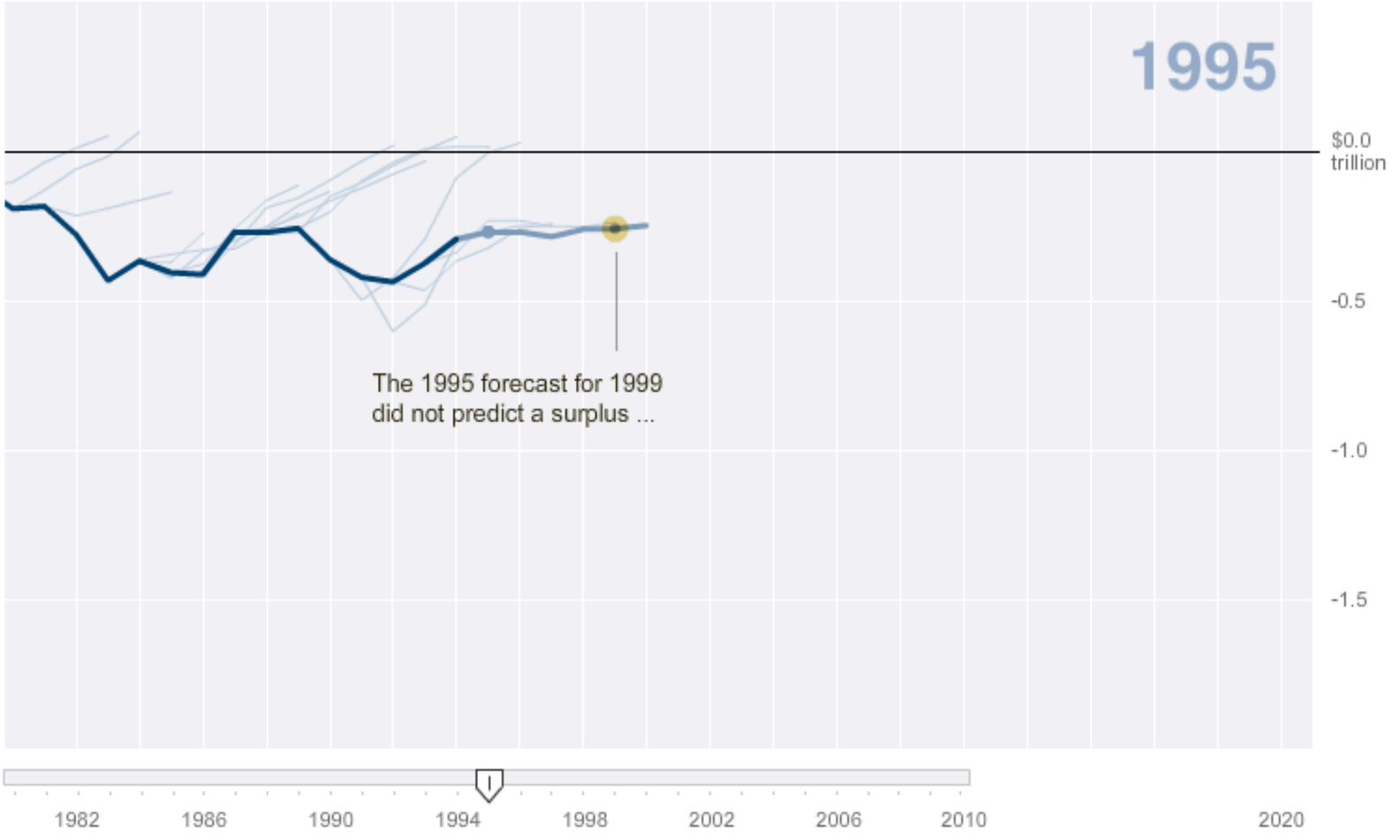
Budget Forecasts, Compared With Reality

Just two years ago, surpluses were predicted by 2012. How accurate have past White House budget forecasts been?

Past forecasts

Even that may be an understatement. In the last 30 years, about 80 percent of four-year deficit forecasts have been too optimistic.

The early Clinton budgets — which failed to predict the surpluses that were generated, in part, by a stock market bubble — are the only major exception.



By AMANDA COX | Send Feedback

Source: Office of Management and Budget

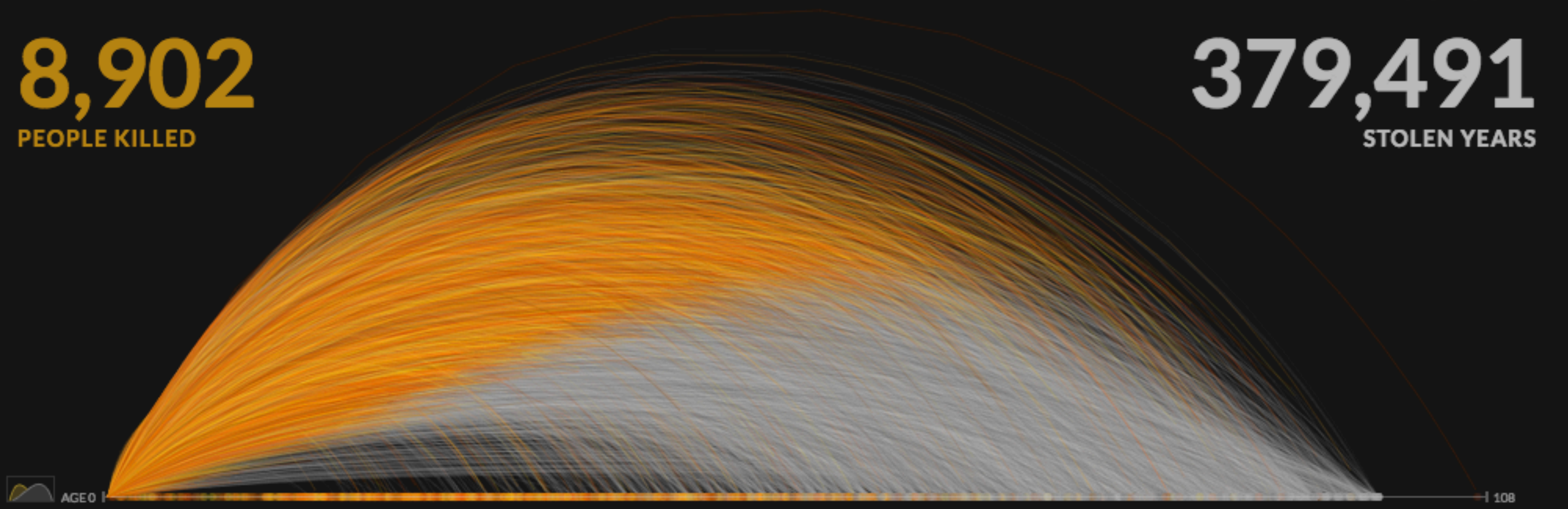
TWITTER LINKEDIN SHARE

U.S. GUN DEATHS IN 2013 2010

OCTOBER

8,902
PEOPLE KILLED

379,491
STOLEN YEARS



THE Fallen OF World War II

Are there more?

[Segel & Heer 2010]

R2
D3

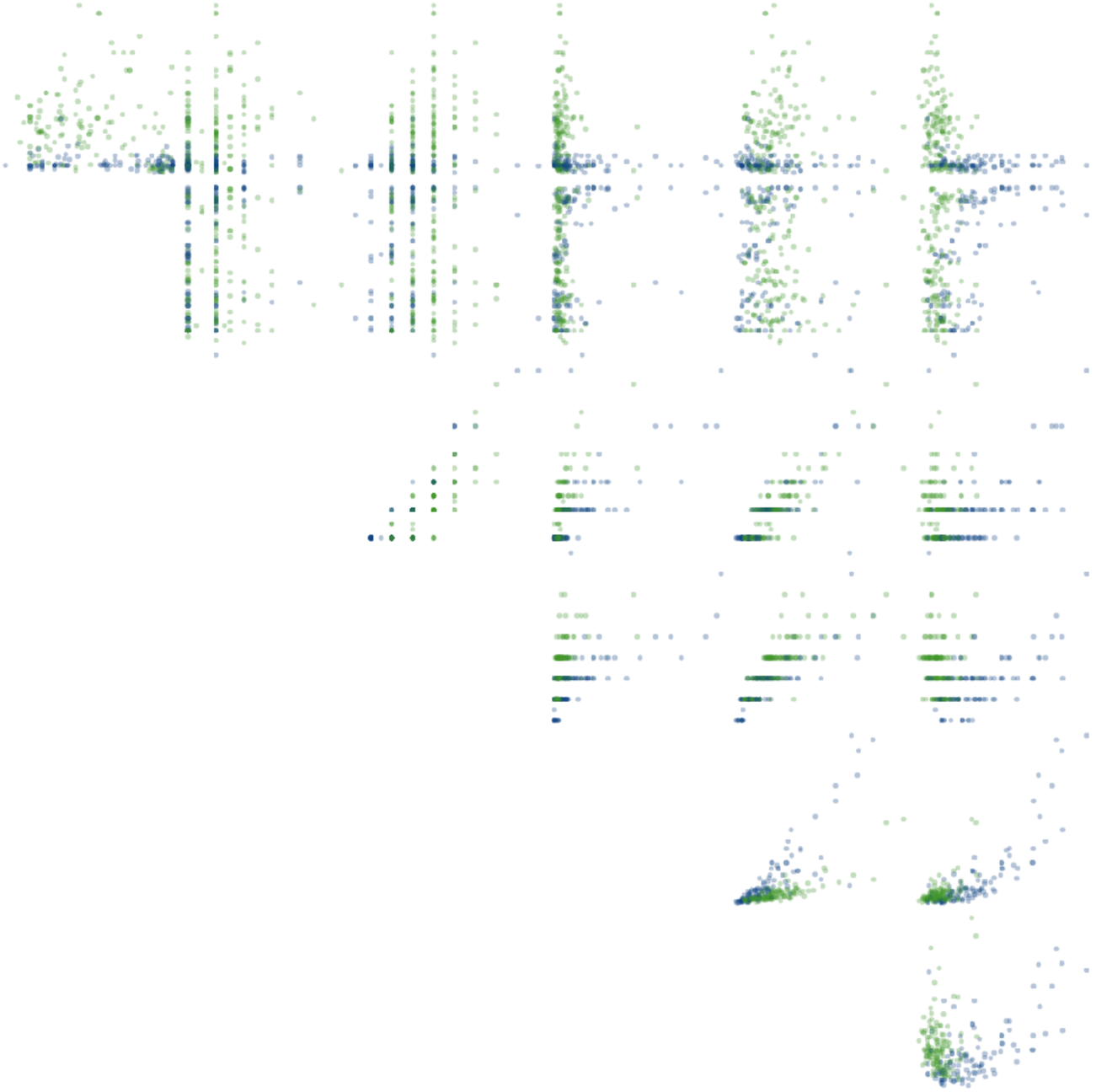
A visual introduction to machine learning

English

In machine learning, computers apply **statistical learning** techniques to automatically identify patterns in data. These techniques can be used to make highly accurate predictions.

Keep scrolling. Using a data set about homes, we will create a machine learning model to distinguish homes in New York from homes in San Francisco.

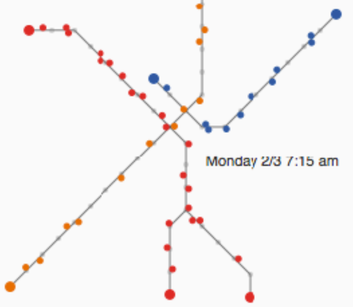
SCROLL



Visualizing MBTA Data

An interactive exploration of Boston's subway system

Mike Barry and Brian Card - June 10, 2014



Star 676 Share 6.9K Tweet G+ Share

Boston's Massachusetts Bay Transit Authority (MBTA) operates the 4th busiest subway system in the U.S. after New York, Washington, and Chicago. If you live in or around the city you have probably ridden on it. The MBTA recently began publishing substantial amount of subway data through its public APIs. They provide the full schedule in General Transit Feed Specification (GTFS) format which powers Google's transit directions. They also publish realtime train locations for the Red, Orange, Blue, and Green lines. The following visualizations use data captured from these feeds for the entire month of February, 2014. Green Line data became available in October, 2014 so is not shown here. Also, working with the MBTA, we were able to acquire per-minute entry and exit counts at each station measured at the turnstiles used for payment.

We attempt to present this information to help people in Boston better understand the trains, how people use the trains, and how the people and trains interact with each other.

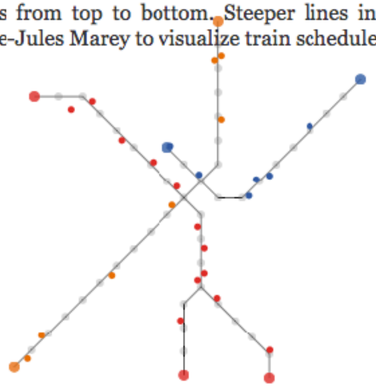
The Trains

In a typical weekday, trains make approximately 1150 trips on the red, orange, and blue lines starting at 5AM and continuing through 1AM the next morning. On Saturdays trains make 870 trips and on Sundays they make 760.

Average Number of Trips per Day

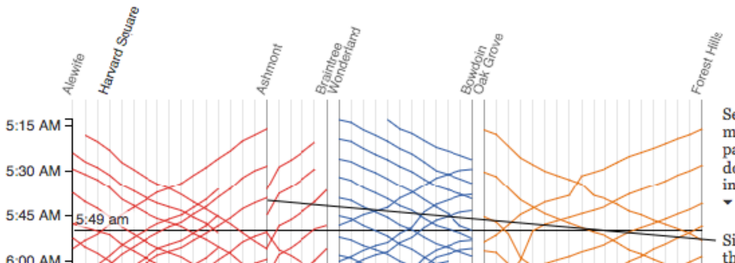
	Weekdays	Saturdays	Sundays
Red	450	350	300
Orange	320	260	220
Blue	380	260	240
Total	**1150**	**870**	**760**

To better understand how the trains operate on a typical day, below are all trips that trains took on the red, orange, and blue lines on Monday February 3 2014. Each vertical line represents a station, and time extends from top to bottom. Steeper lines indicate slower trains. This visualization was first used by Étienne-Jules Marey to visualize train schedules and is typically called a "Marey Diagram."



Locations of each train on the red, blue, and orange lines at 5:49

Subway Trips on Monday February 3, 2014



Service starts at 5AM on Monday morning. Each line represents the path of one train. Time continues downward, so steeper lines indicate slower trains.

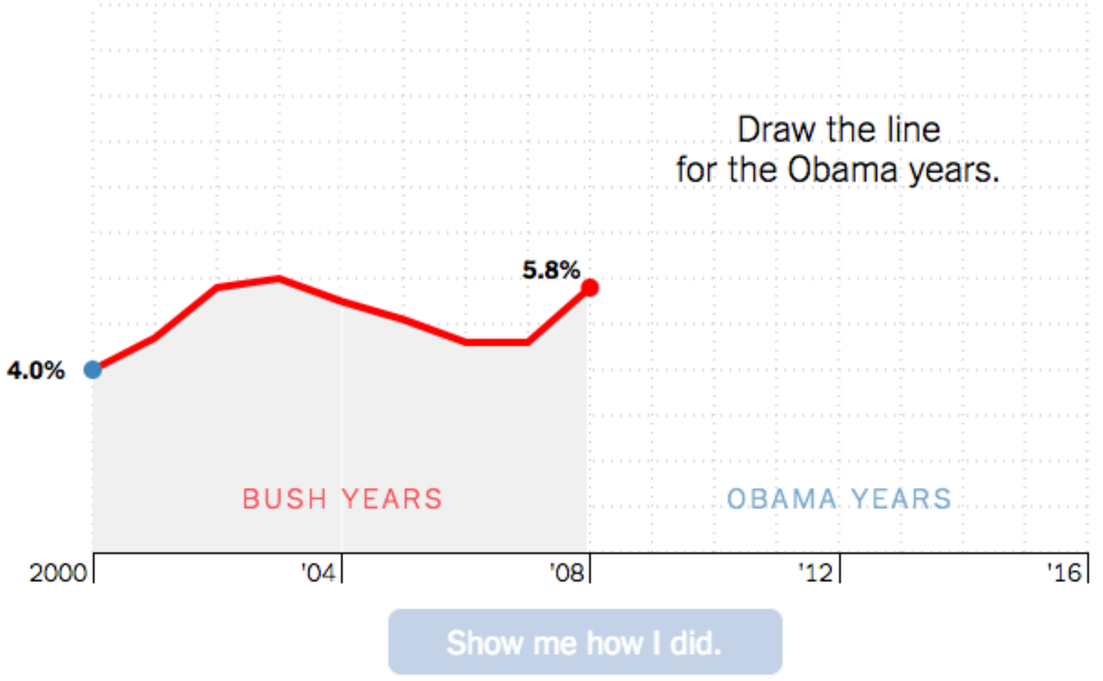
Since the red line splits, we show the Ashmont branch first then the

You Draw It: What Got Better or Worse During Obama's Presidency

By LARRY BUCHANAN, HAEYOUN PARK and ADAM PEARCE JAN. 15, 2017

Draw your guesses on the charts below to see if you're as smart as you think you are.

Under President Obama, the **unemployment rate** ...



Salubrious Nation: a game-y look at U.S. health

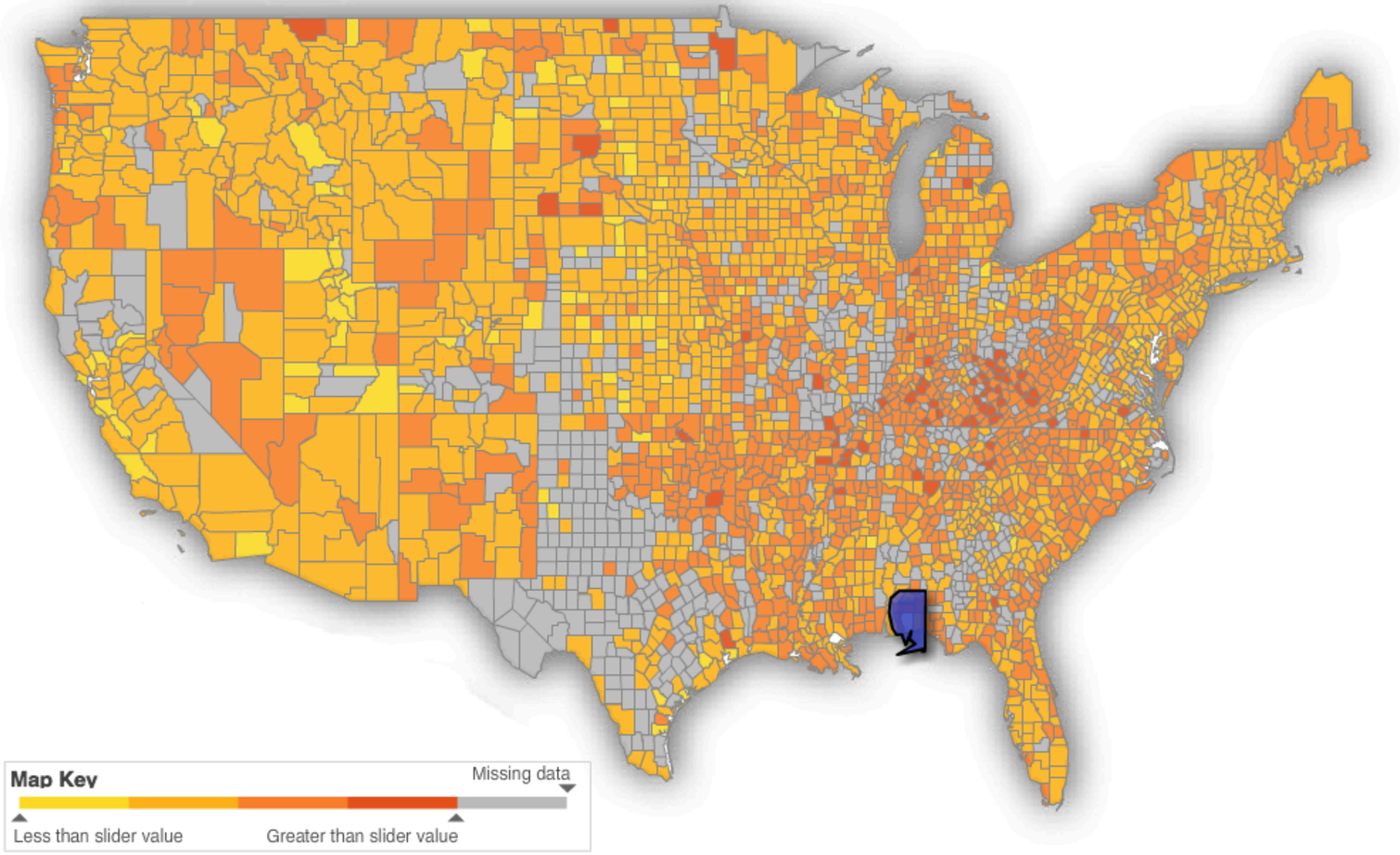
Level (1 / 8)

Adult Smoking
(Percentage rate)

Guess the health parameters for **Santa Rosa, FL**.
Hover your mouse over a county for demographic hints.

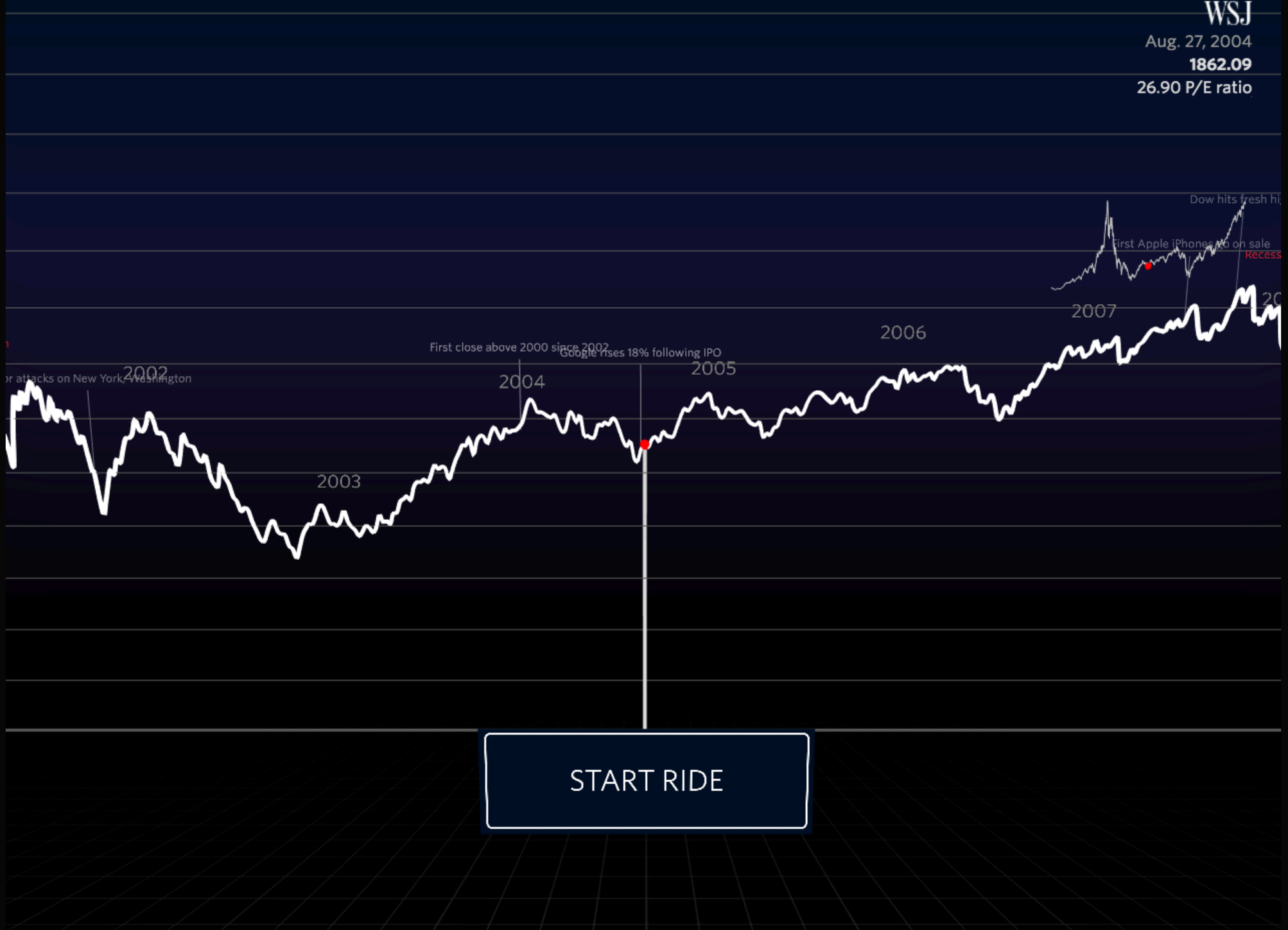
Guess: 23.5 (1 — 46) Submit

Your Score: 0
High Score: 86



Please contact Nick Diakopoulos (nicholas.diakopoulos@gmail.com) with comments or suggestions.

digital dreams
0 minutes
49 seconds
0
X 547
0


WSJ
Aug. 27, 2004
1862.09
26.90 P/E ratio
Dow hits fresh hi
First Apple iPhones to on sale
Recess
2007
2006
First close above 2000 since 2002
Google rises 18% following IPO
2005
2004
2003
2002
r attacks on New York, Washington
START RIDE

Epoch
000,000

Learning rate: 0.03

Activation: Tanh

Regularization: None

Regularization rate: 0

Problem type: Classification

DATA

Which dataset do you want to use?

Ratio of training to test data: 50%

Noise: 0

Batch size: 10

REGENERATE

FEATURES

Which properties do you want to feed in?

X_1

X_2

X_1^2

X_2^2

X_1X_2

$\sin(X_1)$

$\sin(X_2)$

2 HIDDEN LAYERS

4 neurons

2 neurons

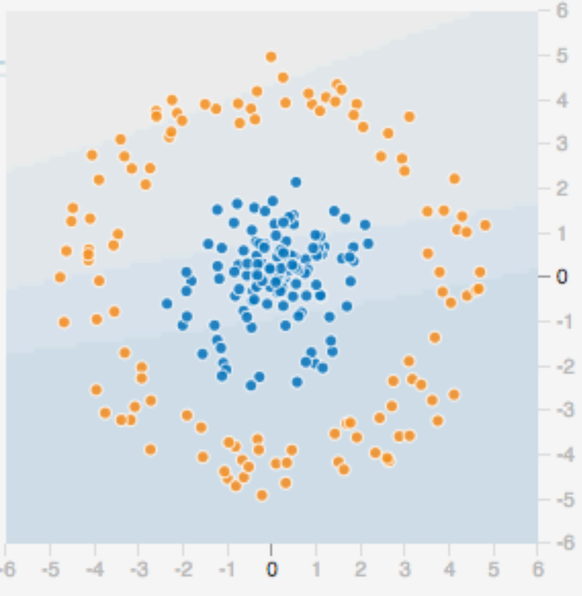
*This is the output from one **neuron**. Hover to see it larger.*

*The outputs are mixed with varying **weights**, shown by the thickness of the lines.*

OUTPUT

Test loss 0.513

Training loss 0.500

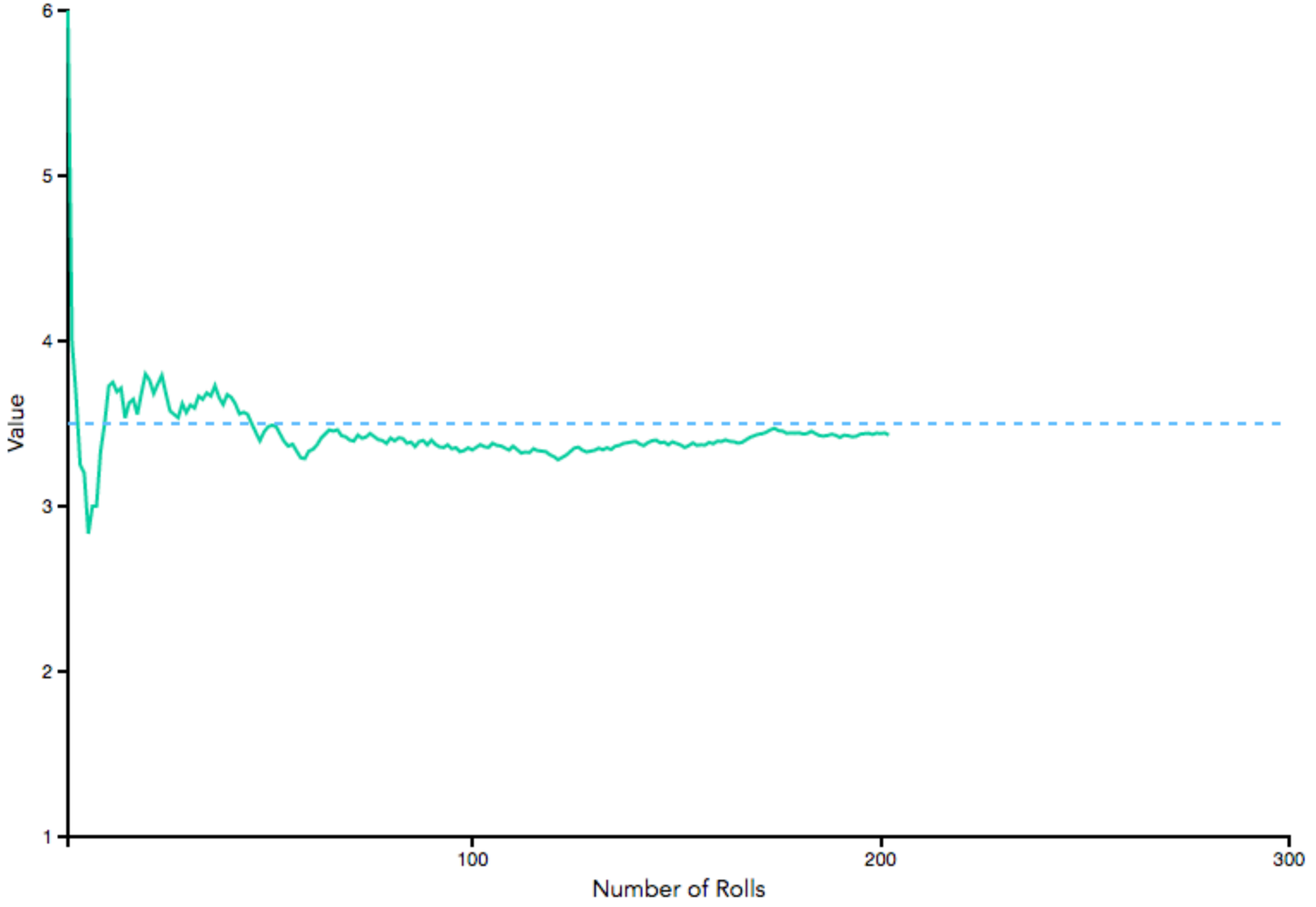


Colors shows data, neuron and weight values.

-1 0 1

Show test data

Discretize output



Expectation

The expectation is a statistic that attempts to summarize the average of a random experiment. It can be interpreted as the long term average outcome of repeating the experiment many times. Mathematically, it is defined as the probability weighted sum of all outcomes in the sample space, denoted,

$$E[X] = \sum_{x \in \Omega} xP(x)$$

Consider the random experiment of rolling a fair die and watch as the running average converges to the expectation of 3.5.

Roll the Die

Roll 100 times

Make the die unfair by adjusting the blue bars below and observe how this changes the expectation.

EXPLORABLE EXPLANATIONS

Lion cubs play-fight to learn hunting skills. Rats play to learn social & emotional skills. Monkeys play to learn cognitive skills, to practice problem-solving and creativity.

And yet, in the last century, we humans have convinced ourselves that play is useless, and learning is *supposed* to be boring. Gosh, no wonder we're all so miserable.

Welcome to **Explorable Explanations**, a hub for learning through play! We're a disorganized "movement" of artists, coders & educators who want to *reunite play and learning*.

Let's get started! Check out these **3 random Explorables**:

★ Fireflies biology

a small, serene simulation of self-synchronization

Published: February 26, 2010

Fractions of a Second: An Olympic Musical

At the Olympics, the blink of an eye can be all that separates the gold medalist from the 10th-place finisher. In some events, this is obvious. But in others, with athletes racing one by one, the closeness of the race is harder to perceive. Listen to the differences below.

Alpine skiing

The women's downhill course was extremely tiring, and, because it was more challenging than the men's course, it ended up separating the skiers by much larger margins. This pattern appears in the two speed events: the downhill and the super-G.

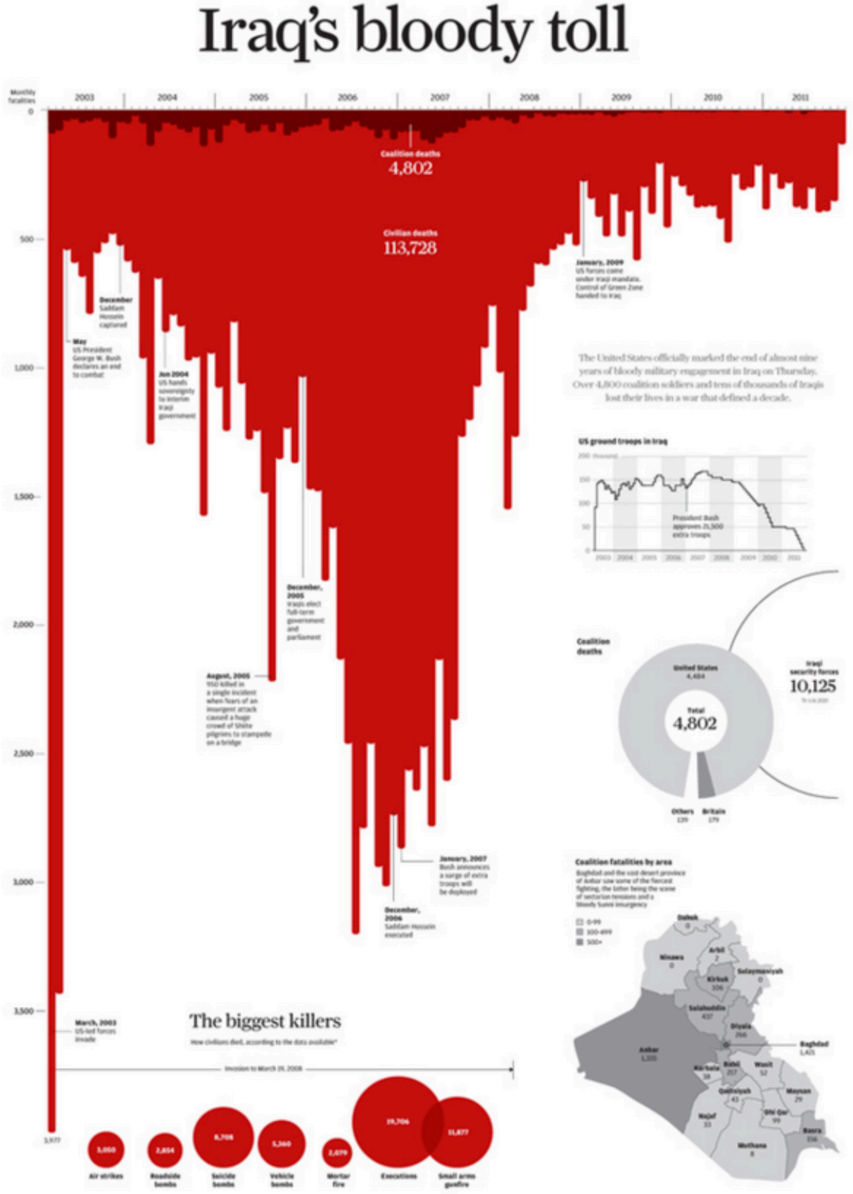
		WINNING TIME	SECONDS BEHIND GOLD MEDALIST 0 0.25 0.5 0.75 1 1.25
PLAY ▶	Women's Downhill	1:44.19	
PLAY ▶	Men's Downhill	1:54.31	
PLAY ▶	Women's Super-G	1:20.14	
PLAY ▶	Men's Super-G	1:30.34	
PLAY ▶	Women's Super Combined	2:09.14	
PLAY ▶	Men's Super Combined	2:44.92	
PLAY ▶	Women's Slalom	1:42.89	
PLAY ▶	Men's Slalom	1:39.32	
PLAY ▶	Women's Giant Slalom	2:27.11	
PLAY ▶	Men's Giant Slalom	2:37.83	

Skeleton, Bobsled and Luge

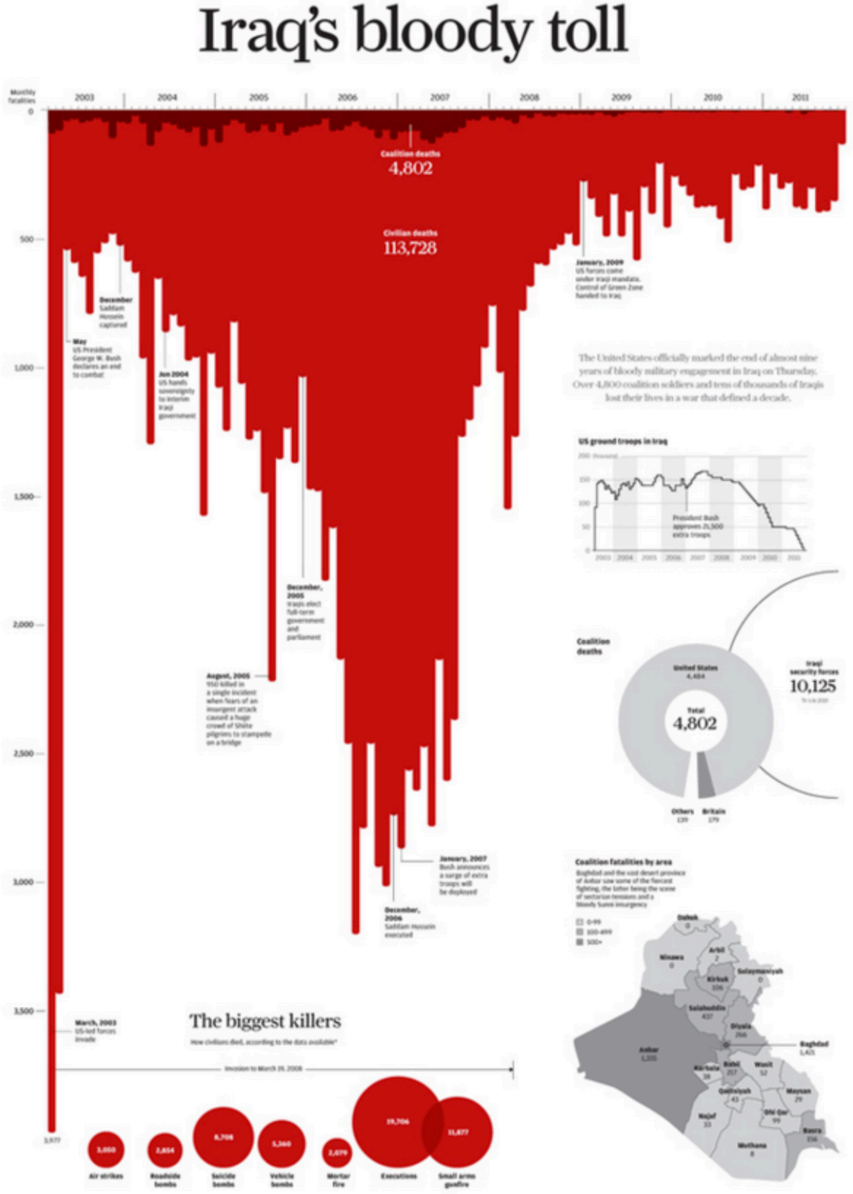
In percentage terms, the men's skeleton had one of the tightest finishes in Vancouver, with only .07 of a second separating the top two finishers across a three-and-a-half-minute run. But the difficult track produced speeds higher than expected, and many sliding events had relatively large gaps between gold and silver.

		WINNING TIME	SECONDS BEHIND GOLD MEDALIST 0 0.25 0.5 0.75 1 1.25
PLAY ▶	Men's Skeleton	3:29.73	
PLAY ▶	Women's Skeleton	3:35.64	
PLAY ▶	Men's Two-Man Bobsled	3:26.65	
PLAY ▶	Men's Four-Man Bobsled	3:24.46	
PLAY ▶	Women's Bobsled	3:32.28	
PLAY ▶	Men's Singles Luge	3:13.085	
PLAY ▶	Women's Singles Luge	2:46.524	
PLAY ▶	Doubles Luge	1:22.705	

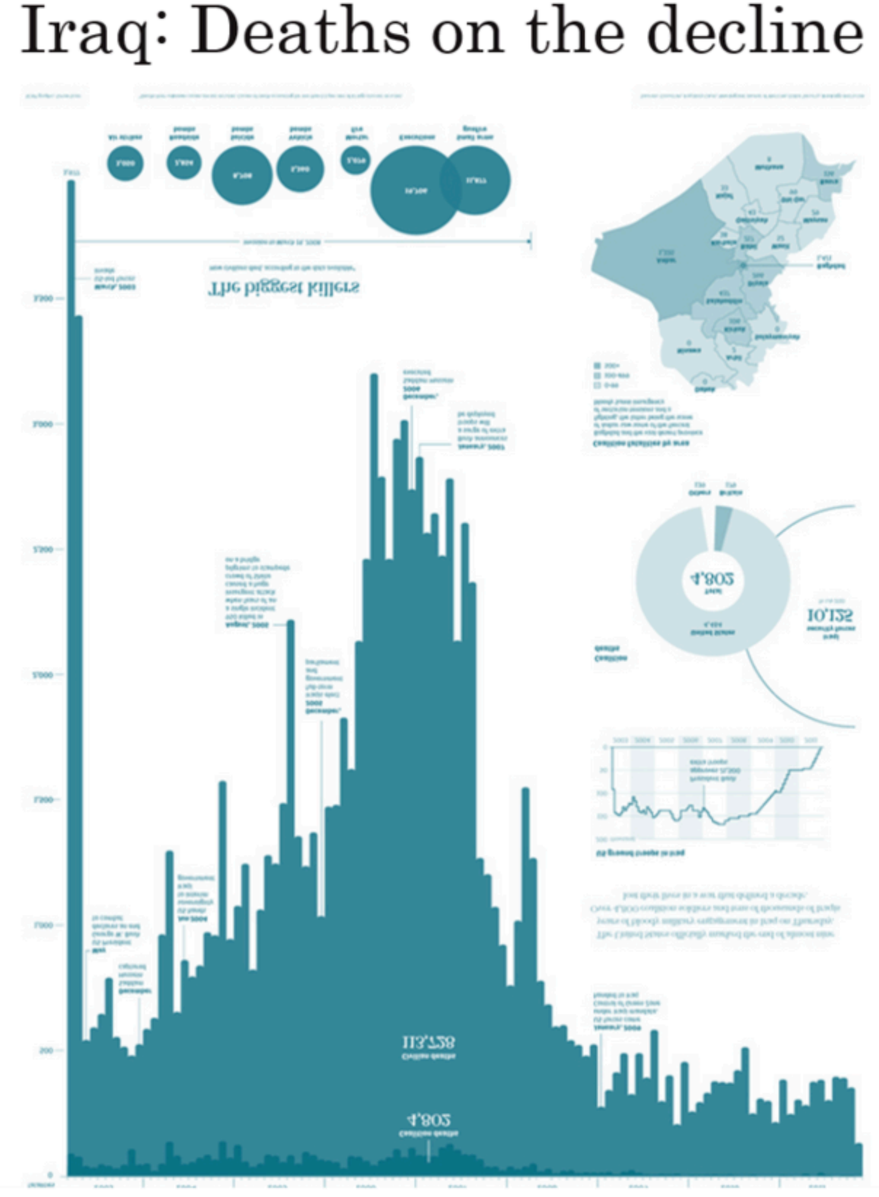
Storytelling, Double-edged Sword?



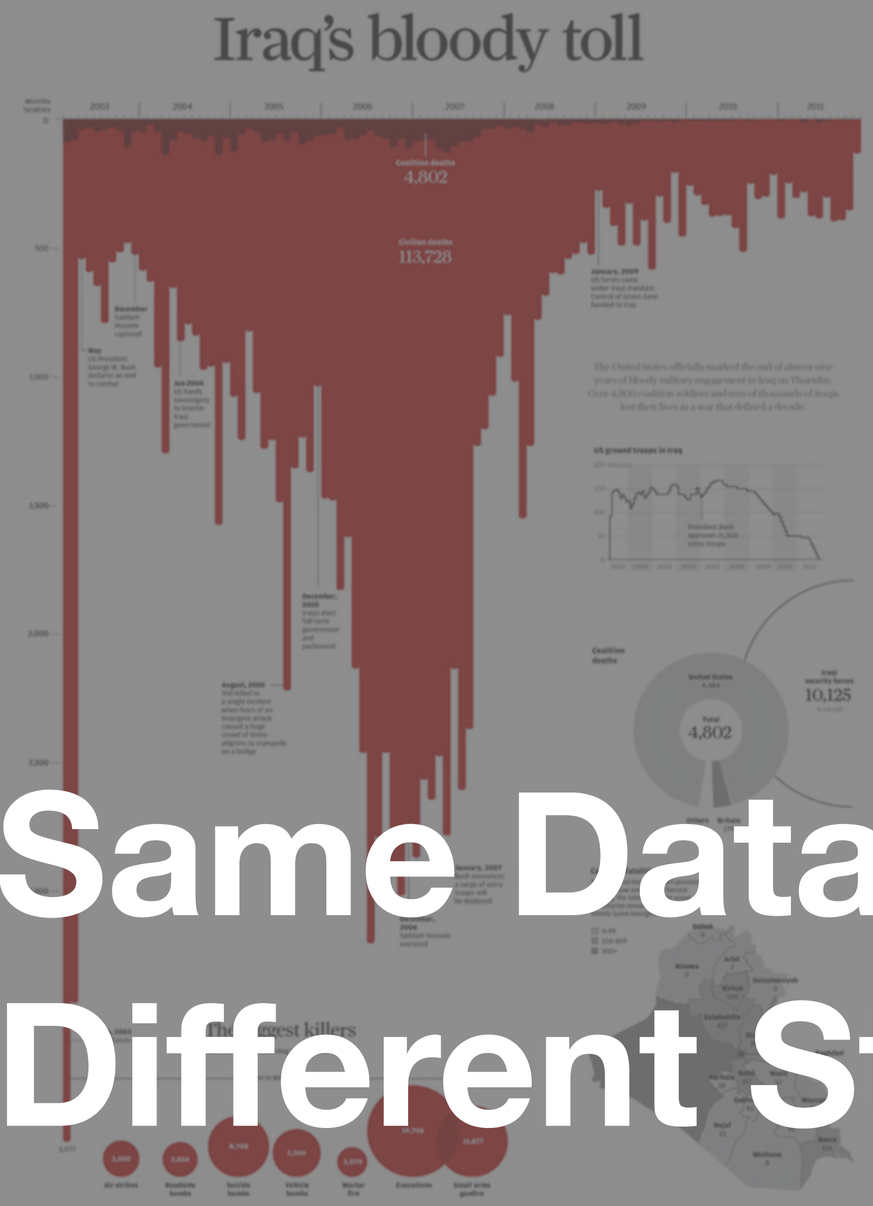
[South China Morning Post 2011]

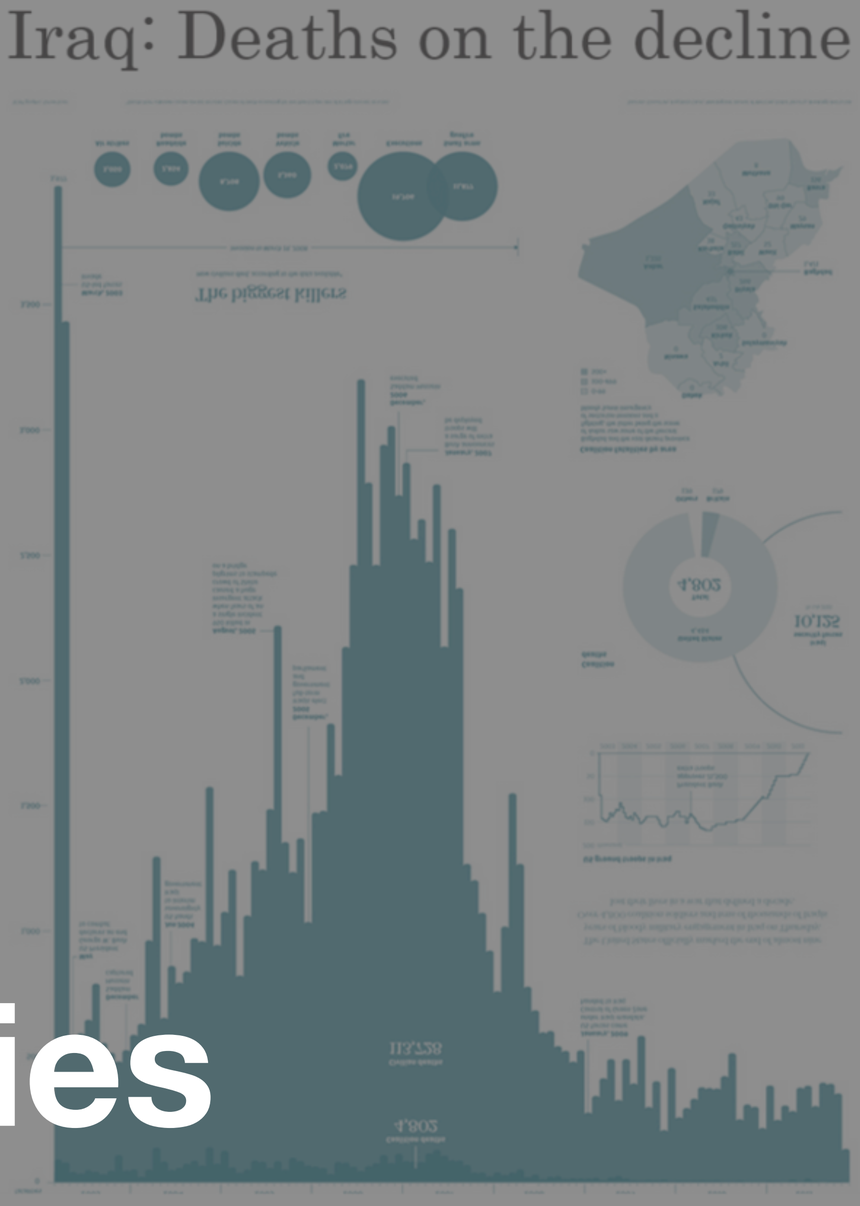


[South China Morning Post 2011]



Flipped upside down….





Same Data, Different Stories

[South China Morning Post 2011]

Flipped upside down….

Presentation & Storytelling in Tableau

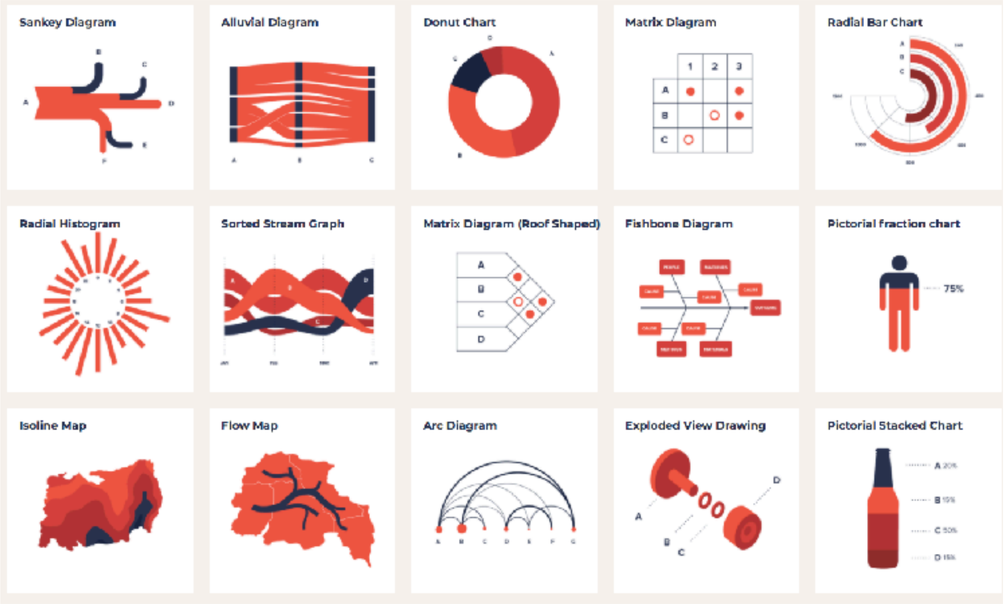
Dashboard

Annotation

Story Points

Next

Advanced visualizations



10 min break