Storytelling with data

Skills Course
Class 1: February 16, 2018
Nitze 517

Nikos Tsafos (nikostsafos@jhu.edu)

Why visualization matters

(explore data)

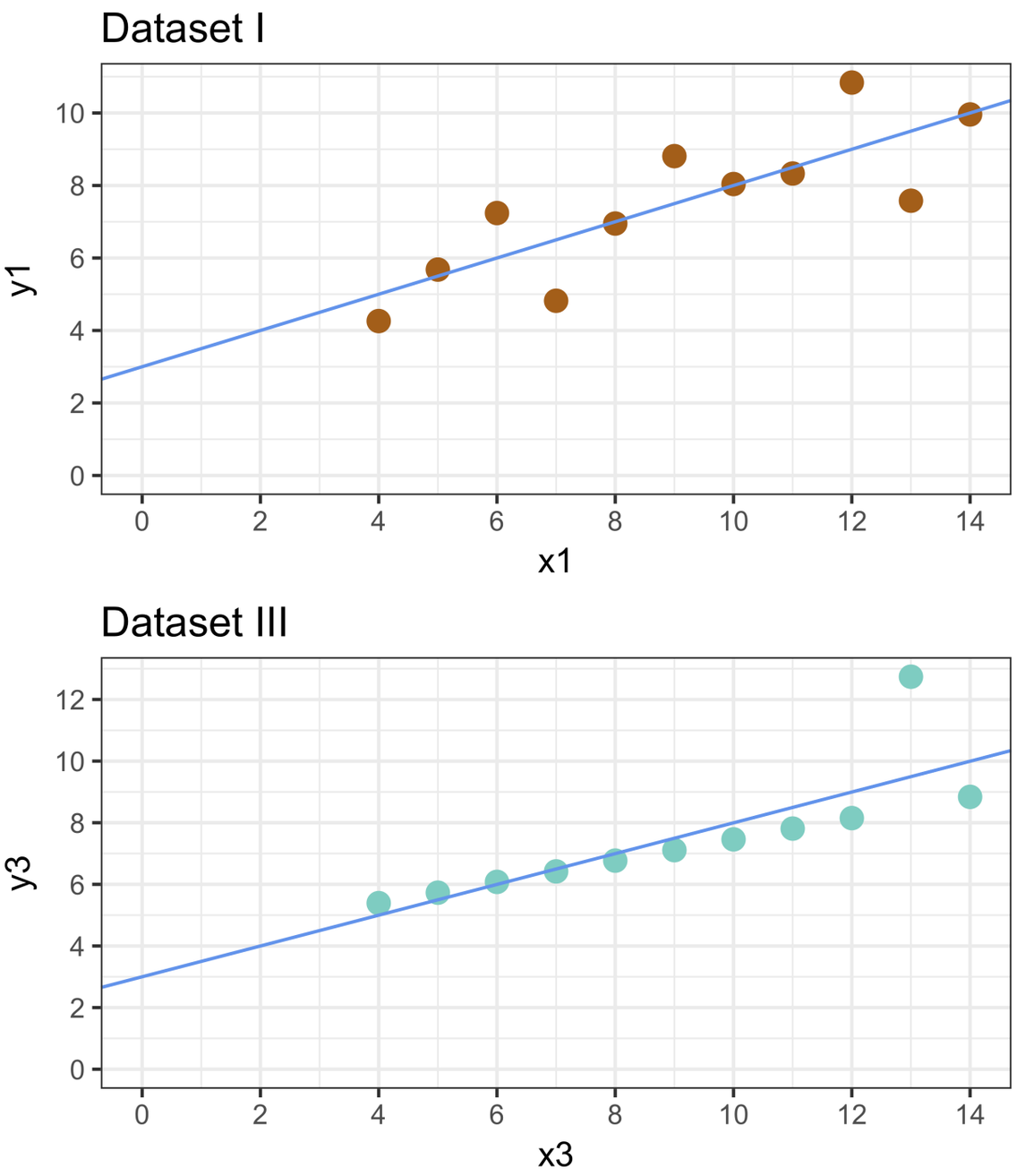
Four data sets with identical summary statistics…

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

Mean of x	9.00
Mean of y	7.50
Regression line	Y = 3 + 0.5X
Correlation between x and y	0.816
R-square	0.667

Adapted from Anscombe's quartet

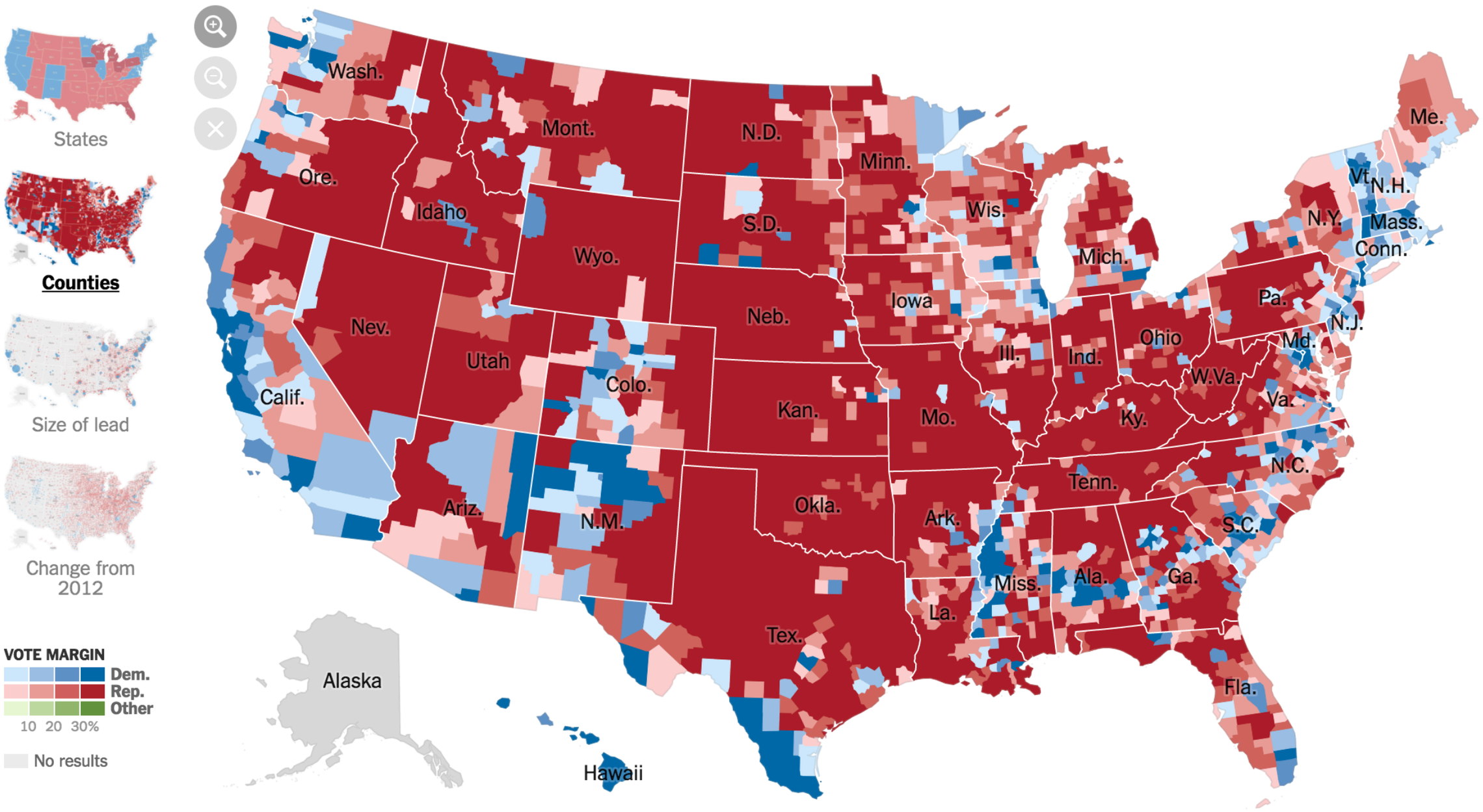
… but totally different relationships



Adapted from Anscombe's quartet

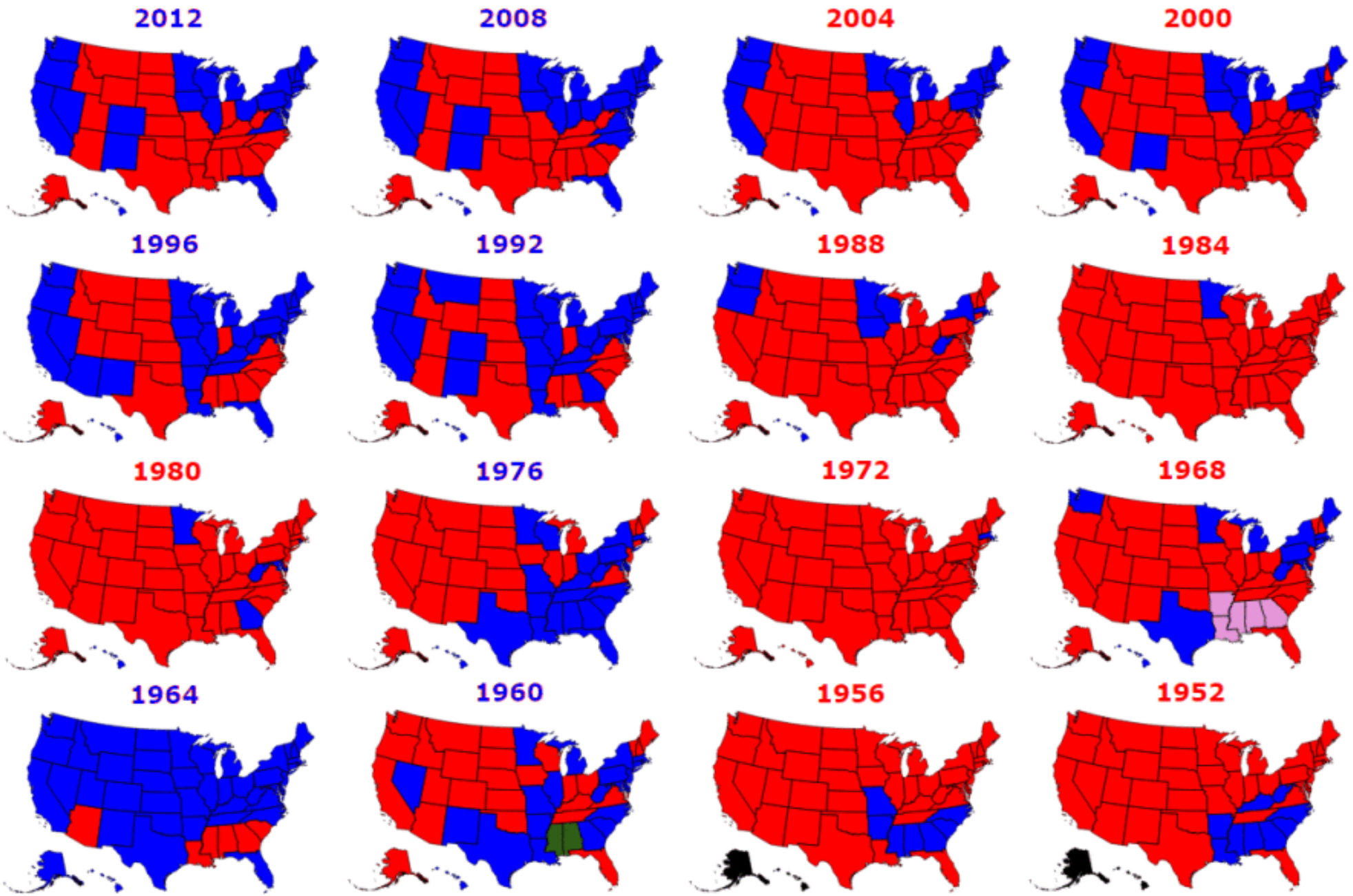
Why visualization matters (information density)

Over 3,000 data points (counties), plus attributes, in a one map



Source: The New York Times, 2016 Presidential Election Results

Sixty years of US presidential elections in one image



Source: Five great 20th century Presidential landslides

What this course covers

Course objectives

(What works) Understand the fundamentals of good design and data visualization.

(How to do it) Understand the software and principles that one can use to create good design.

(Let's do it) Gain hands on experience (and feedback) on creating a compelling visual story.

Course outline

Class 1. Fundamentals of design and data visualization

Class 2. Workshop (Learn by doing)

Class 3. Presentations (Show your work)

Outline for today's class

From concept to design How will the visualization be shared (presentation, online, print, web, etc.)? Is it a repeated exercise or a one-off? Who is the audience? What is the purpose of the visualization? What's the story?

Design principles How does our mind process visual information? Cover topics such as the Gestalt principles, spatial distortions (e.g. subway maps), color, type (font), weight (bold, transparent), annotations, headlines, notes, etc.

Design examples What makes visualization compelling? Having reviewed principles of visual design, we will explore actual visualizations from the real world; what works and what doesn't? We will rely on several websites for inspiration.

Design options What options do we have to visualize information? Cover conventional and emerging chart/visualization types: bar, area, and line charts; choropleth and other maps; sparklines; small multiples; bump charts; etc. For inspiration, we will rely on http://datavizproject.com/.

From concept to design

Five questions before you start

How will the visualization be shared (presentation, online, print, web, etc.)?

What works on a projector (with voice over) might not work on Twitter.

Is it a repeated exercise or a one-off?

A weekly or monthly report might look different than a one-off presentation.

Who is the audience?

How much context do they need; what do want to see; what will they find persuasive?

What is the purpose of the visualization?

Is it to merely present information; or make a point, draw attention to a conclusion?

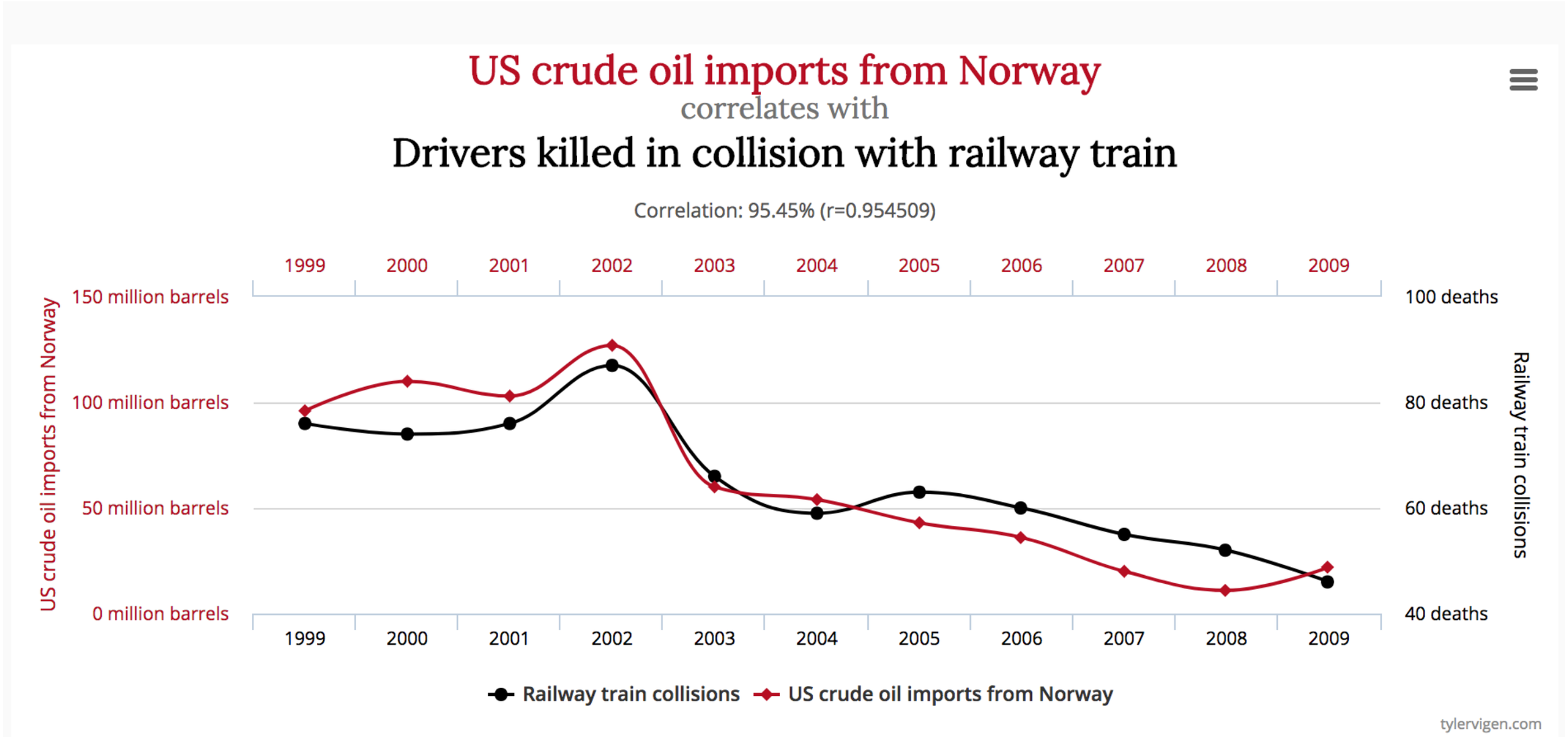
What's the story?

What do you want the reader to take away? What idea are you trying to challenge?

First, a few “rules”

1. Have something to say

Garbage in, garbage out (aka data don't always prove something)



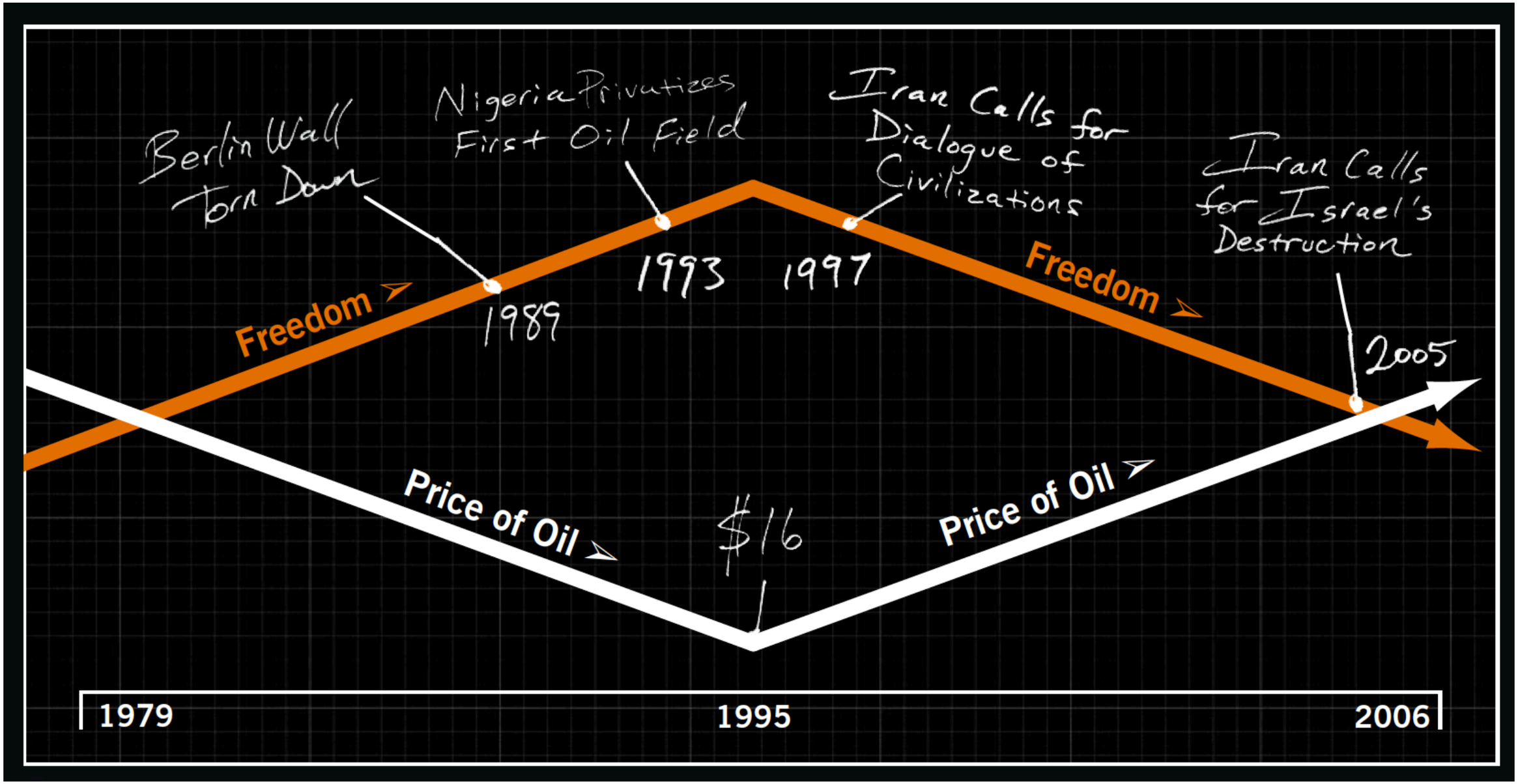
Data sources: Dept. of Energy and Centers for Disease Control & Prevention

Source: Tyler Vigen, Spurious Correlations

2. Be intellectually honest

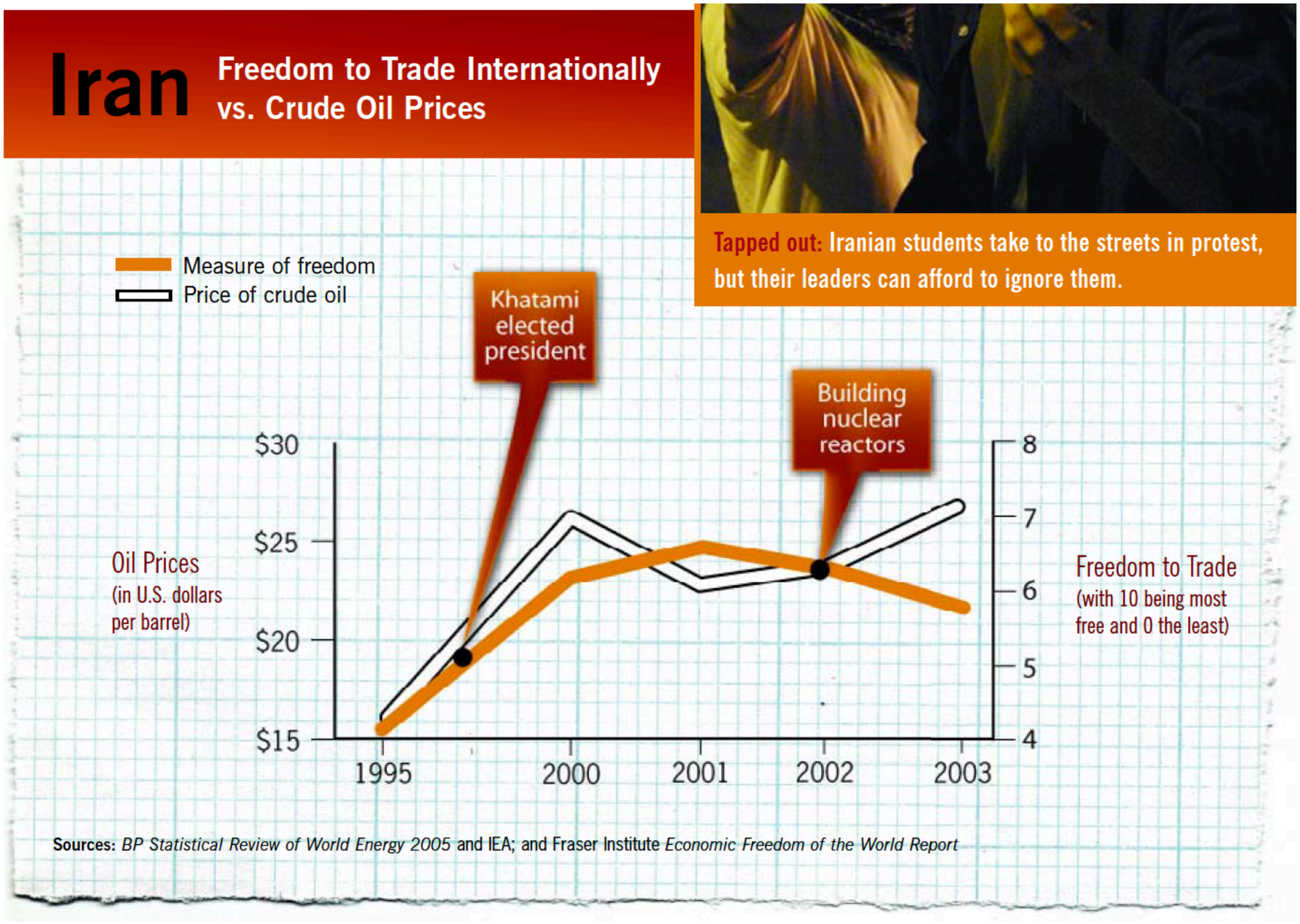
(contextualize; represent reality)

How NOT to present a visual story 1/5



Source: Friedman (2009), The First Law of Petropolitics

How NOT to present a visual story 2/5

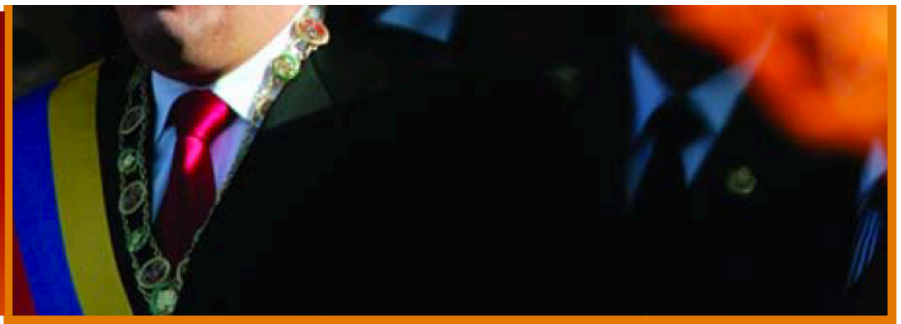


Source: Friedman (2009), The First Law of Petropolitics

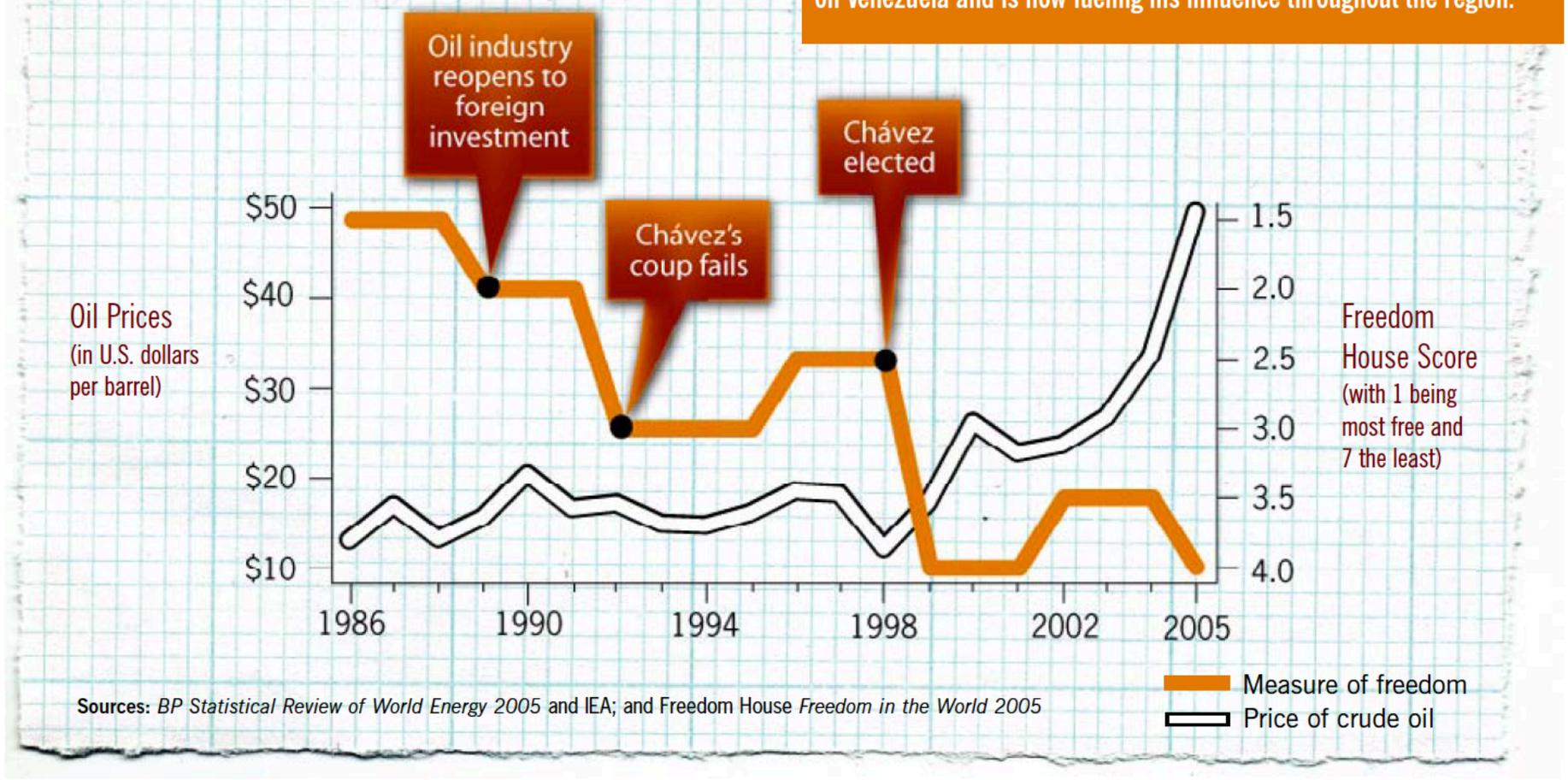
How NOT to present a visual story 3/5

Venezuela

Freedom House *Freedom in the World* Rankings vs. Crude Oil Prices



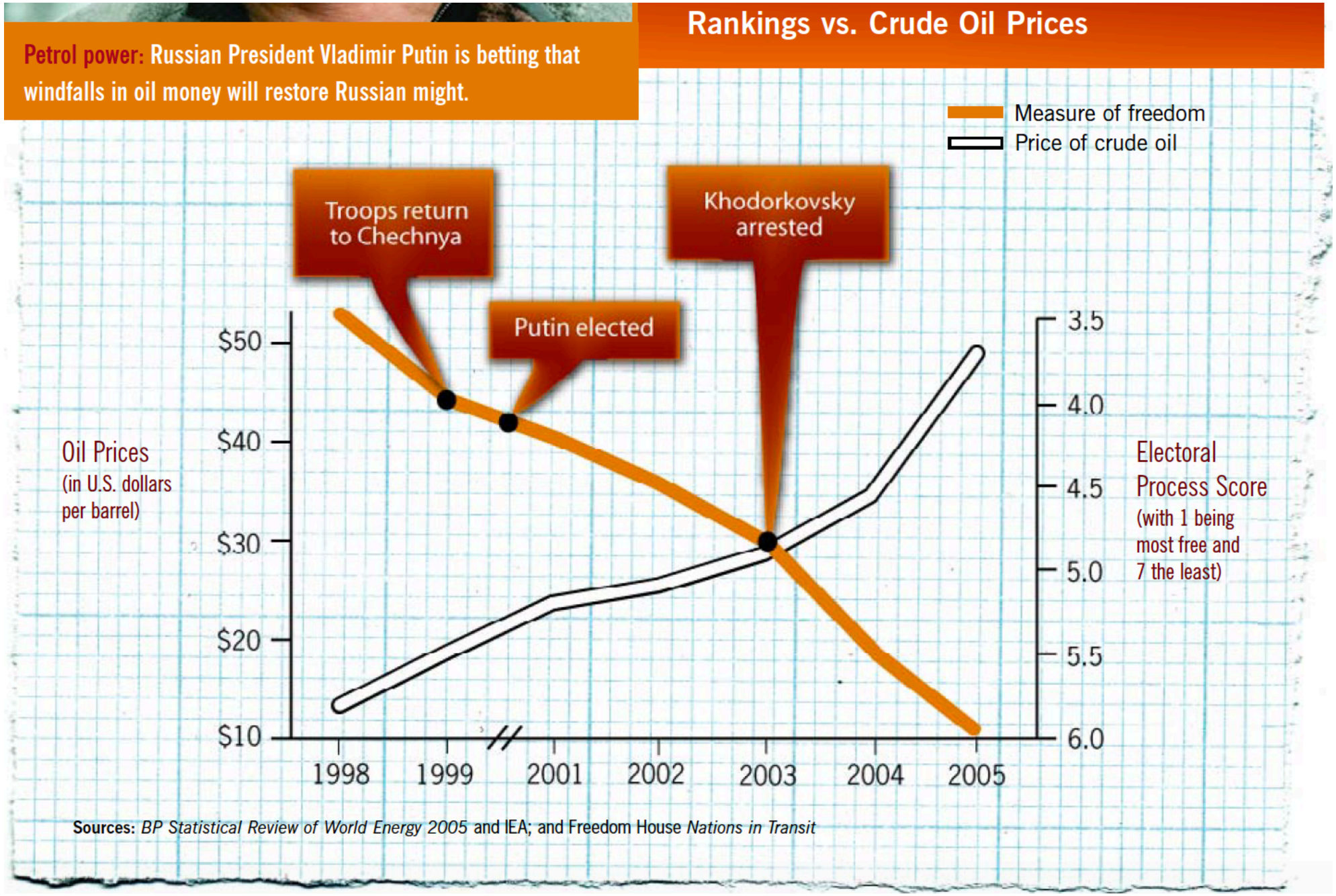
Firestarter: The rising price of oil has strengthened Hugo Chávez's grip on Venezuela and is now fueling his influence throughout the region.



Sources: *BP Statistical Review of World Energy 2005* and IEA; and Freedom House *Freedom in the World 2005*

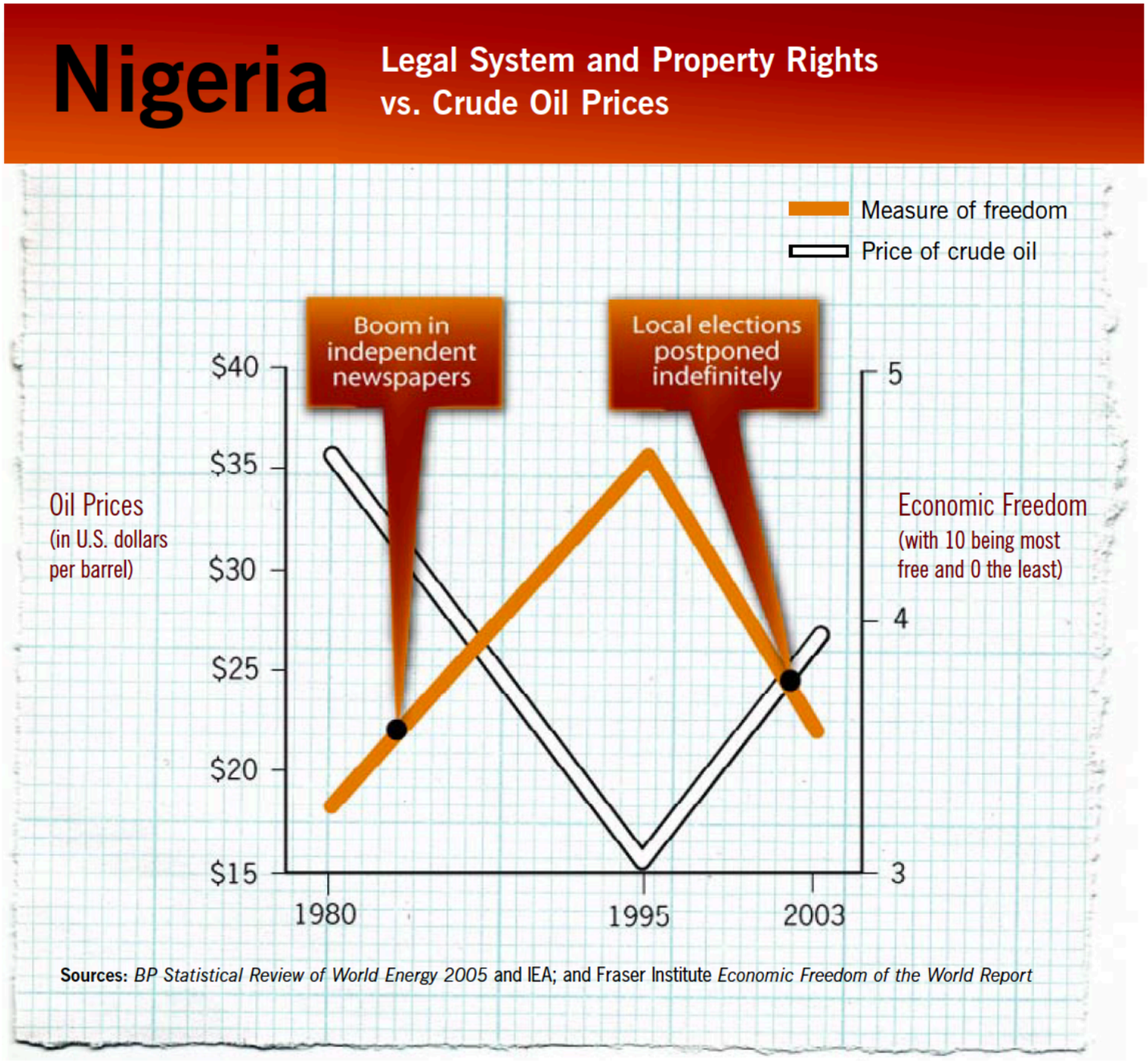
Source: Friedman (2009), The First Law of Petropolitics

How NOT to present a visual story 4/5



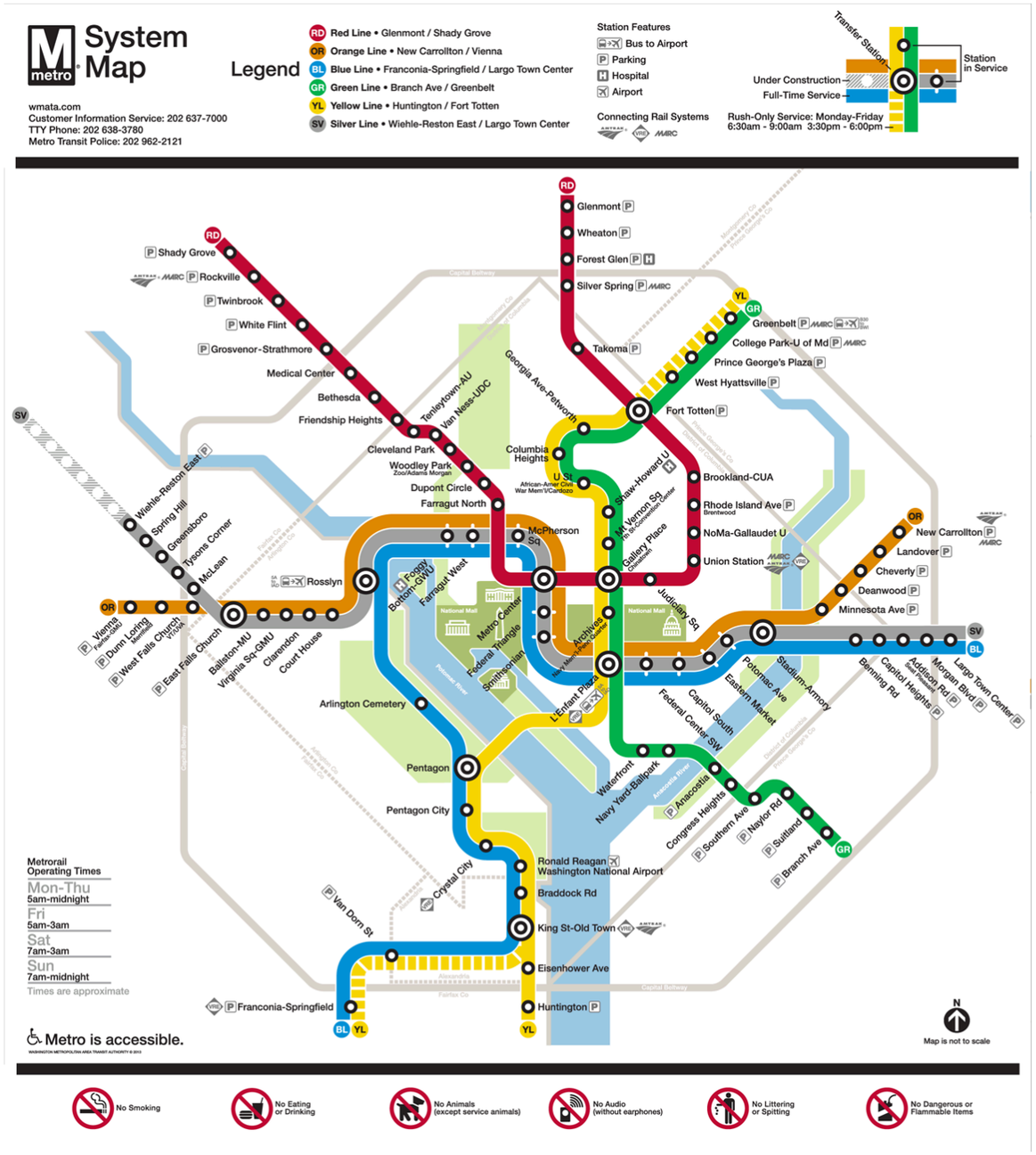
Source: Friedman (2009), The First Law of Petropolitics

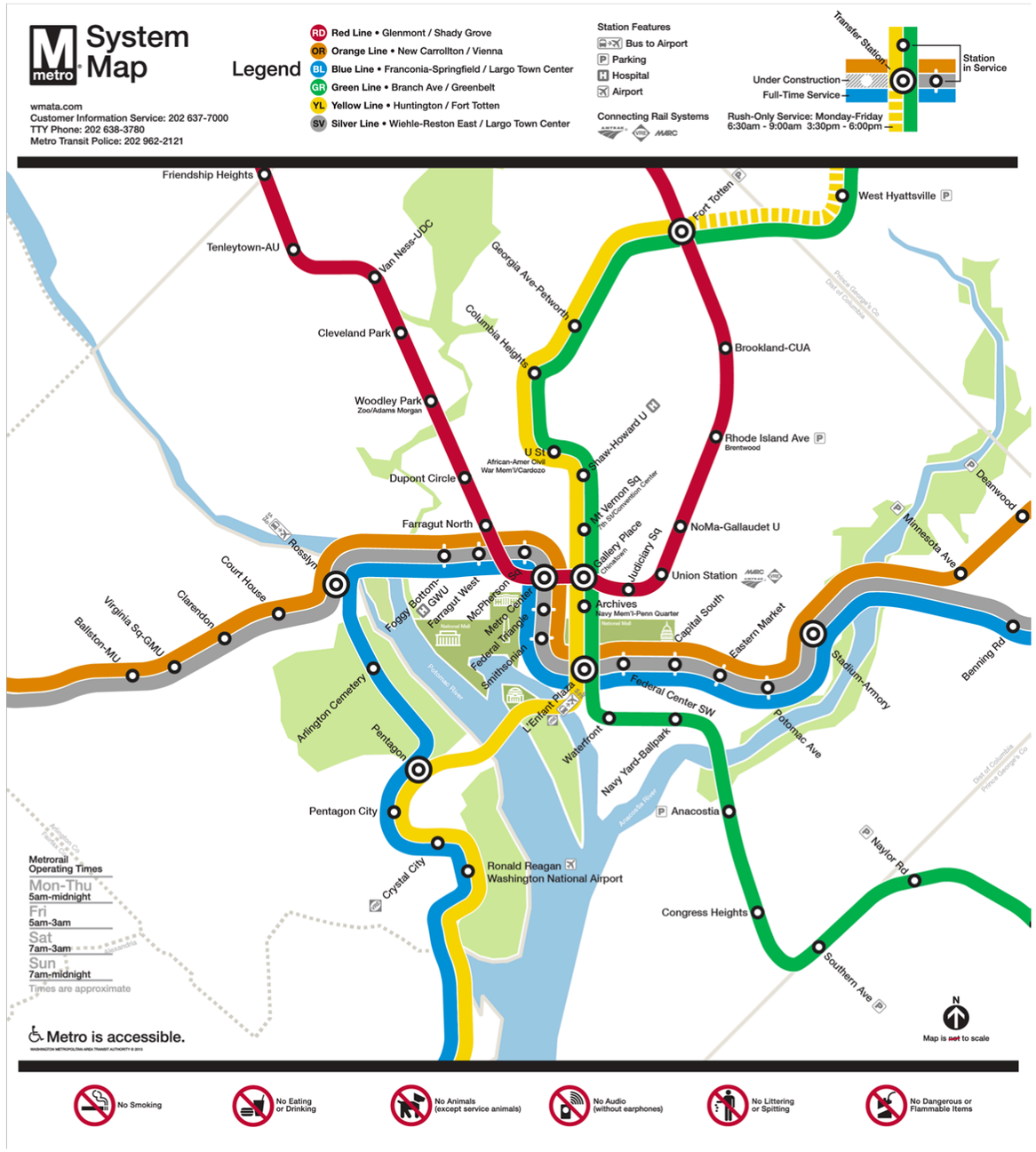
How NOT to present a visual story 5/5



Source: Friedman (2009), The First Law of Petropolitics

But some distortion sometimes helps





Source: Peter Dovak, Washington Metro Map to Scale

3. Make ink count

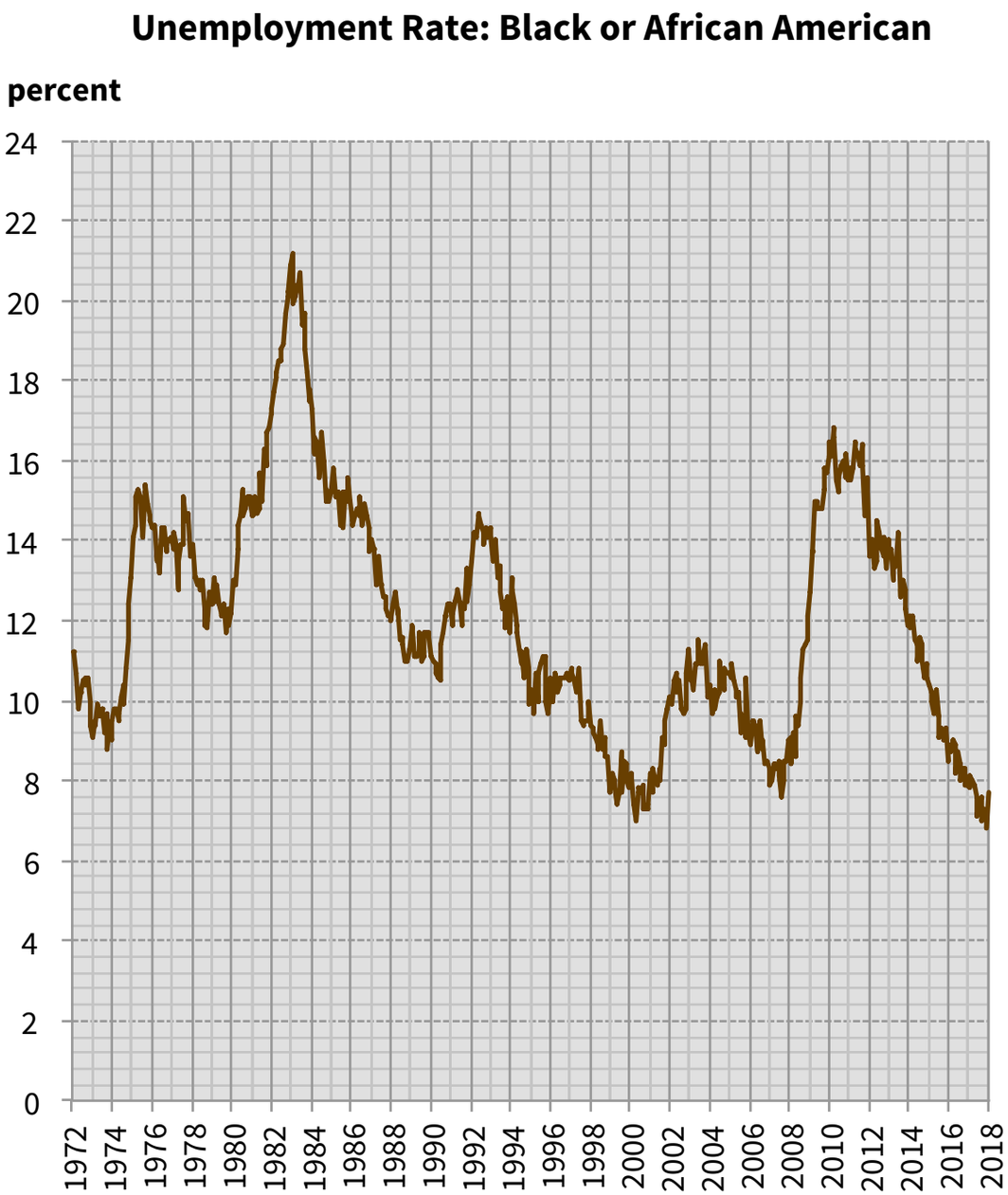
$$\text{Data-ink ratio} = \frac{\text{data-ink}}{\text{total ink used to print the graphic}}$$

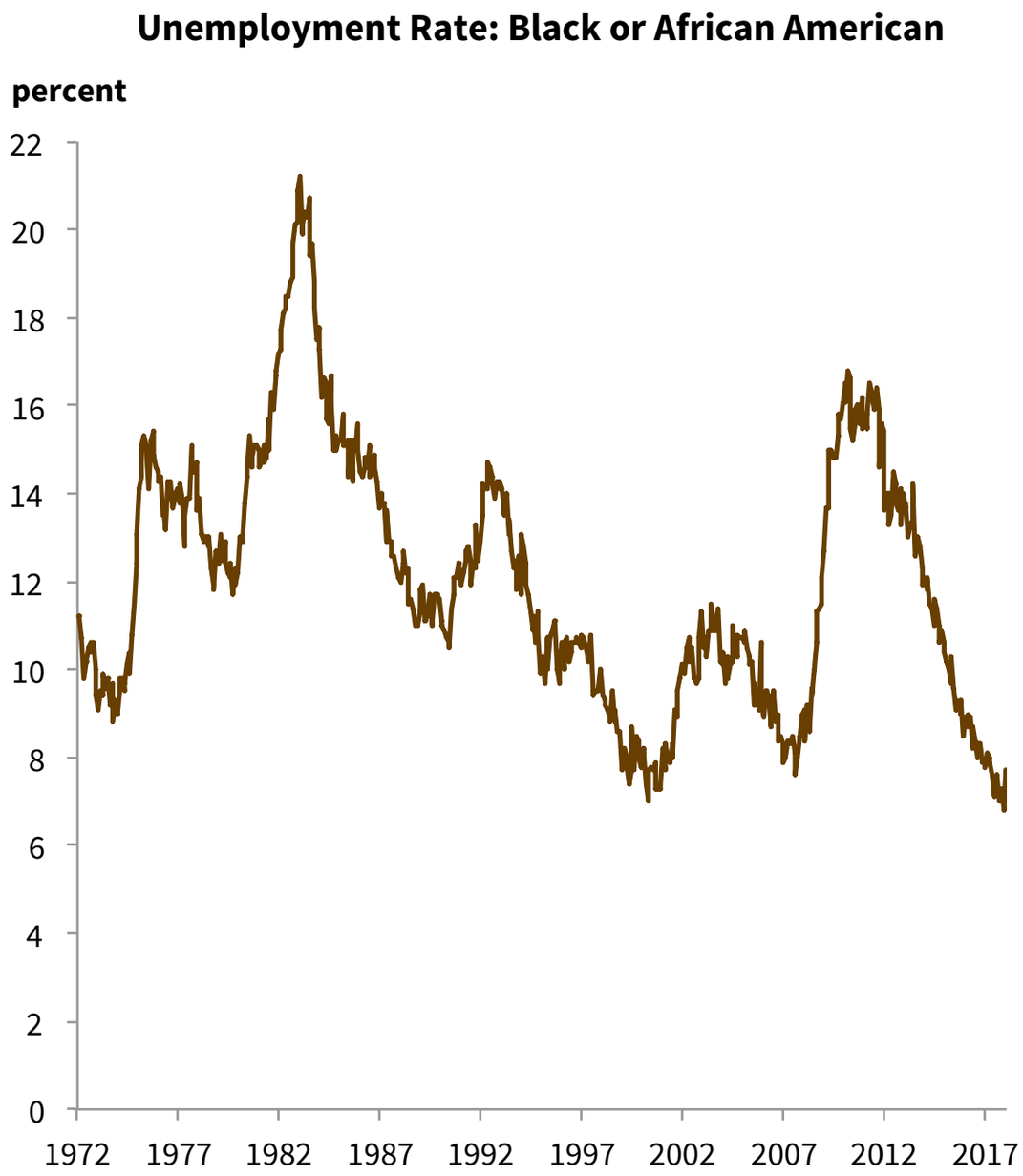
= proportion of a graphic's ink devoted to the non-redundant display of data-information

= 1.0 – proportion of a graphic that can be erased without loss of data-information.

Source: Tufte, Edward (1983). The Visual Display of Quantitative Information, p. 93

Low data-ink ratio vs. high data-ink ratio





High data-ink ratio (aka: every pixel conveys vital information)

PLAYER	P	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	PF	+/-
Kevin Durant	F	28	10	3	9	33.3	0	1	0	4	4	100	1	5	6	6	1	0	2	1	6
Draymond Green	PF	35	17	8	16	50	1	5	20	0	0	0	1	7	8	11	3	1	1	2	14
Zaza Pachulia	C	15	12	4	5	80	0	0	0	4	4	100	2	1	3	1	1	0	0	3	-4
Klay Thompson	G	34	25	10	14	71.4	5	6	83.3	0	0	0	0	3	3	0	3	0	0	1	16
Stephen Curry	PG	28	17	7	13	53.8	3	7	42.9	0	0	0	1	3	4	8	2	2	0	0	6
Kevon Looney		7	0	0	1	0	0	0	0	0	0	0	1	0	1	1	0	1	0	0	-2
Andre Iguodala		26	6	3	7	42.9	0	3	0	0	0	0	1	5	6	3	1	0	0	3	16
Patrick McCaw		13	0	0	1	0	0	1	0	0	0	0	0	1	1	1	1	0	0	2	1
JaVale McGee		7	7	2	2	100	0	0	0	3	4	75	1	2	3	0	1	0	1	0	11
David West		16	13	6	6	100	0	0	0	1	1	100	1	3	4	1	2	0	1	5	11
Shaun Livingston		14	4	1	3	33.3	0	0	0	2	2	100	0	2	2	3	1	1	1	1	13
Omri Casspi		6	2	1	3	33.3	0	0	0	0	0	0	2	1	3	0	0	0	0	0	-1
Nick Young		4	9	3	5	60	3	4	75	0	0	0	0	0	0	0	0	0	0	0	-2
		240	122	48	85	56.5	12	27	44.4	14	15	93.3	11	33	44	35	16	5	6	18	17

Source: NBA, Box score for GSW-SAS, February 10, 2018 (link)

4. Anticipate

Present the facts…

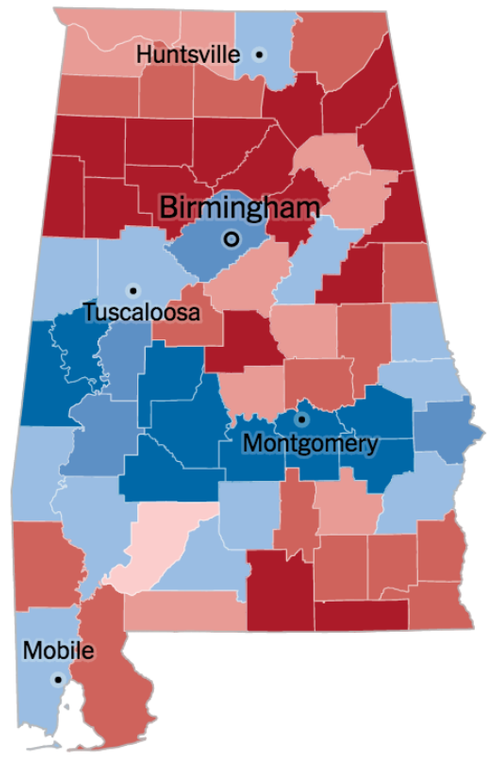
By MATTHEW BLOCH, NATE COHN, JOSH KATZ and JASMINE LEE DEC. 12, 2017, 11:59 PM ET

CANDIDATE	PARTY	VOTES	PCT.
✔ Doug Jones	Democrat	671,151	49.9%
Roy Moore	Republican	650,436	48.4
Total Write-Ins	—	22,819	1.7

100% reporting (2,220 of 2,220 precincts)

Doug Jones, a Democrat, won the special election on Tuesday to fill the United States Senate seat vacated by Jeff Sessions, now the attorney general. Mr. Jones aimed to create a lead in the urban counties that include Birmingham and Montgomery, and across a band of largely black counties. Strong support for Roy S. Moore, the Republican, was expected in rural, mostly white parts of the state.

One critical battleground was a trio of smaller, whiter cities: Mobile, Tuscaloosa and Huntsville. Late Tuesday night, Mr. Jones led by a large margin in Mobile County, and he had won Tuscaloosa County and Madison County, home of Huntsville.



VOTE SHARE

Dem. 50 60 70% Rep. 50 60 70% No results

COUNTY	JONES	MOORE	WRITE-INS	RPT.
Jefferson	149,522	66,309	3,710	100%
Madison	65,664	46,313	3,446	100
Mobile	62,253	46,725	1,539	100

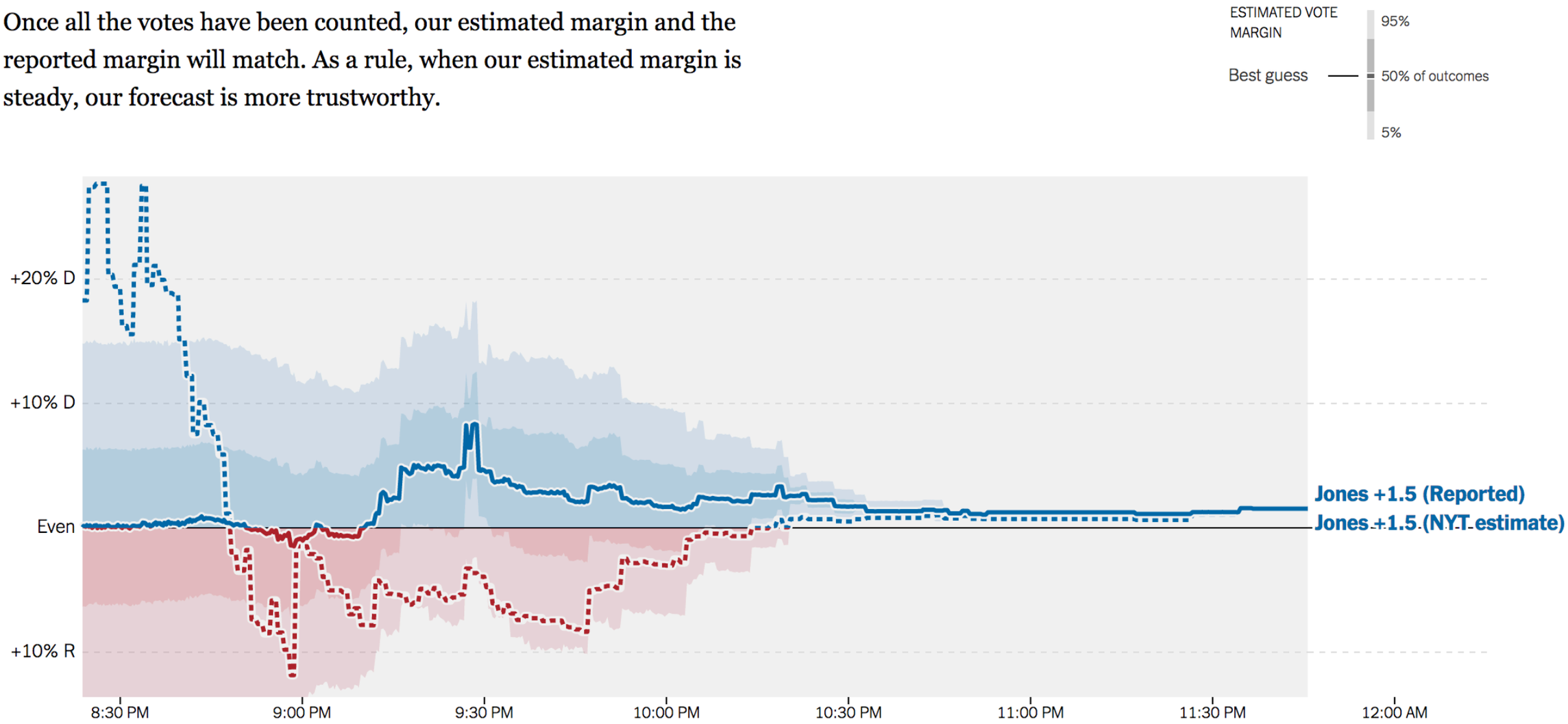
+ Show all

Source: The New York Times, Coverage of Alabama Senate Special Election (December 12, 2017)

But also understand what the reader really cares about

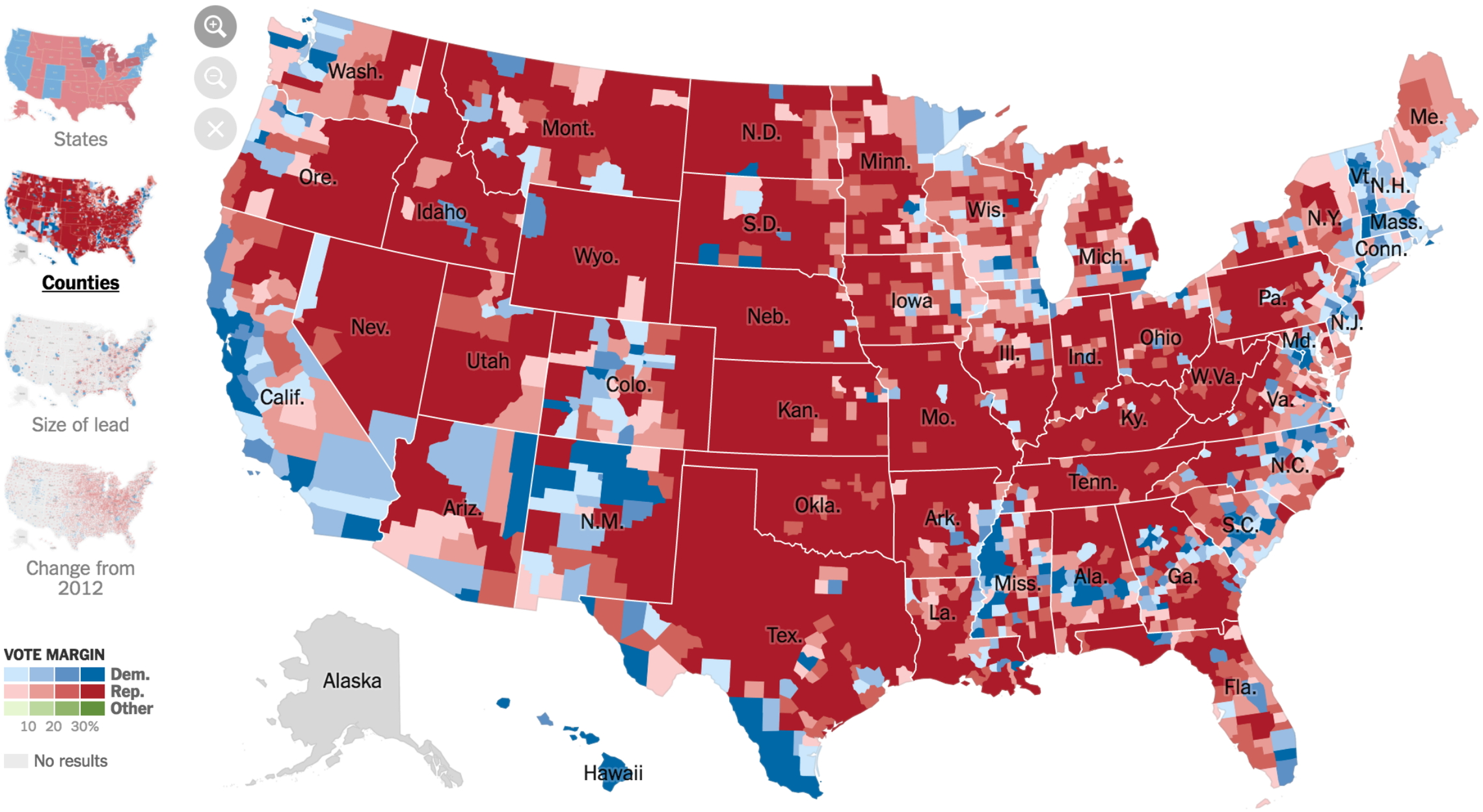
Projected Vote Margin

Once all the votes have been counted, our estimated margin and the reported margin will match. As a rule, when our estimated margin is steady, our forecast is more trustworthy.



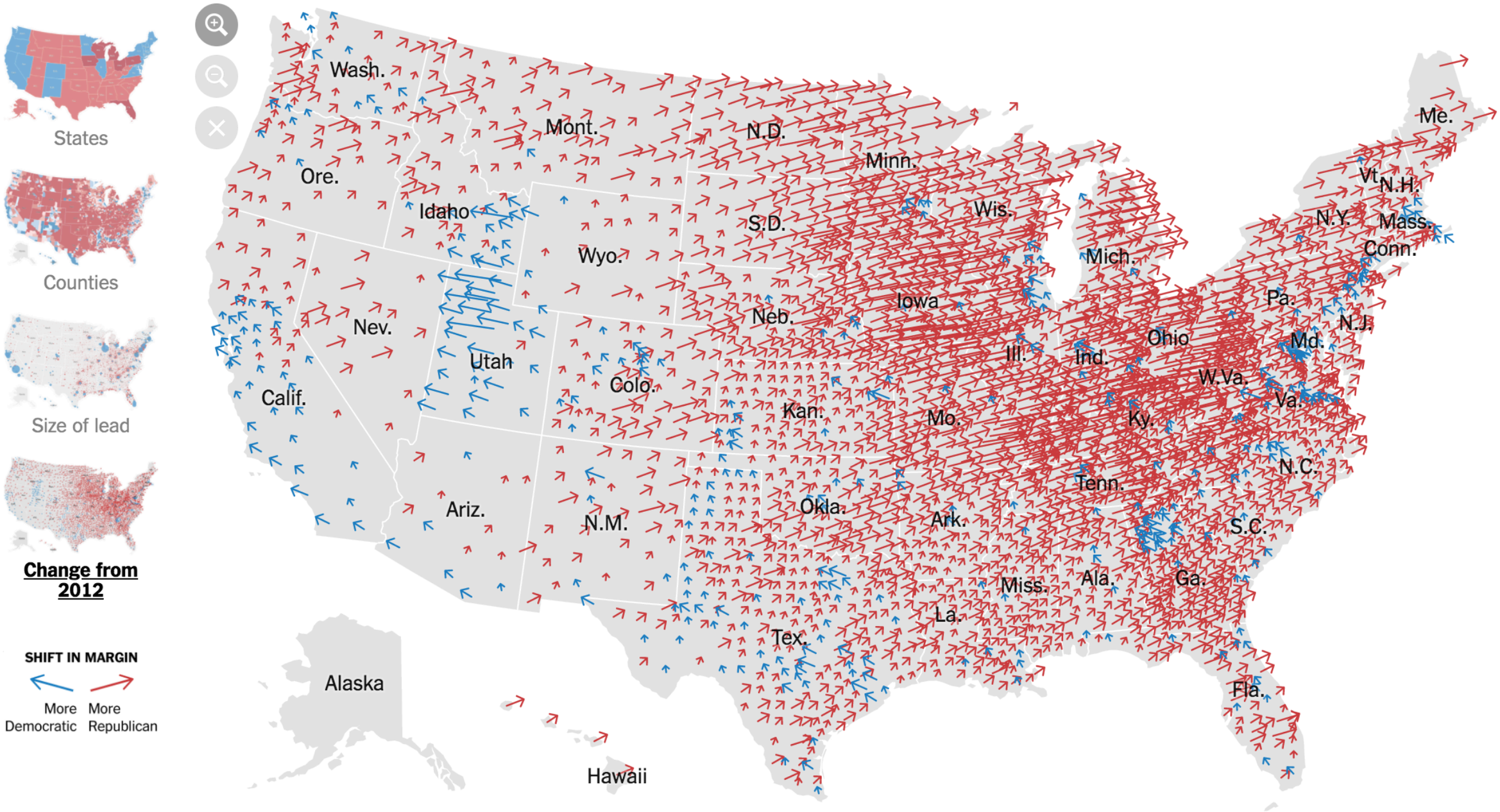
Source: The New York Times, Coverage of Alabama Senate Special Election (December 12, 2017)

Present the facts…



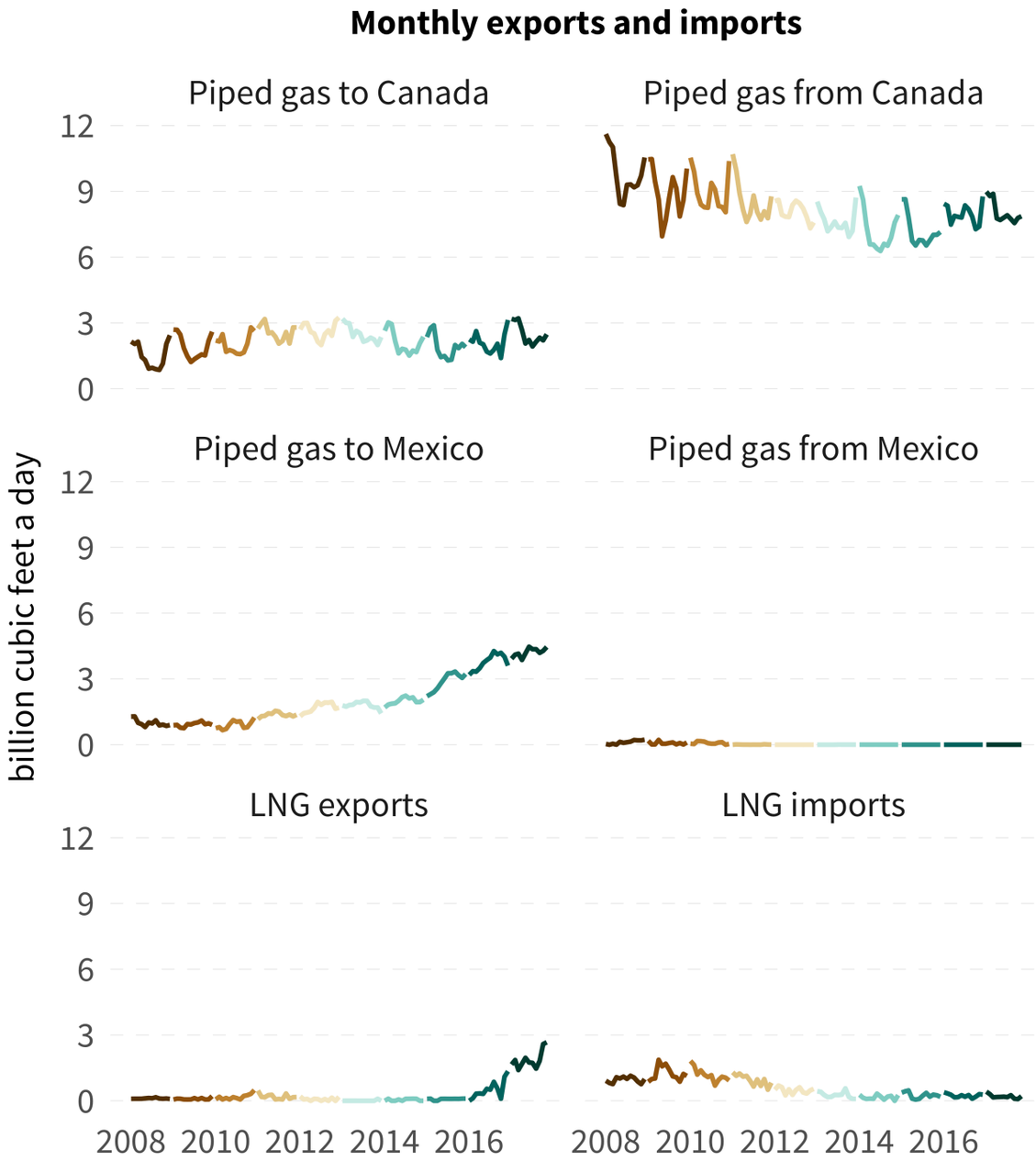
Source: The New York Times, 2016 Presidential Election Results

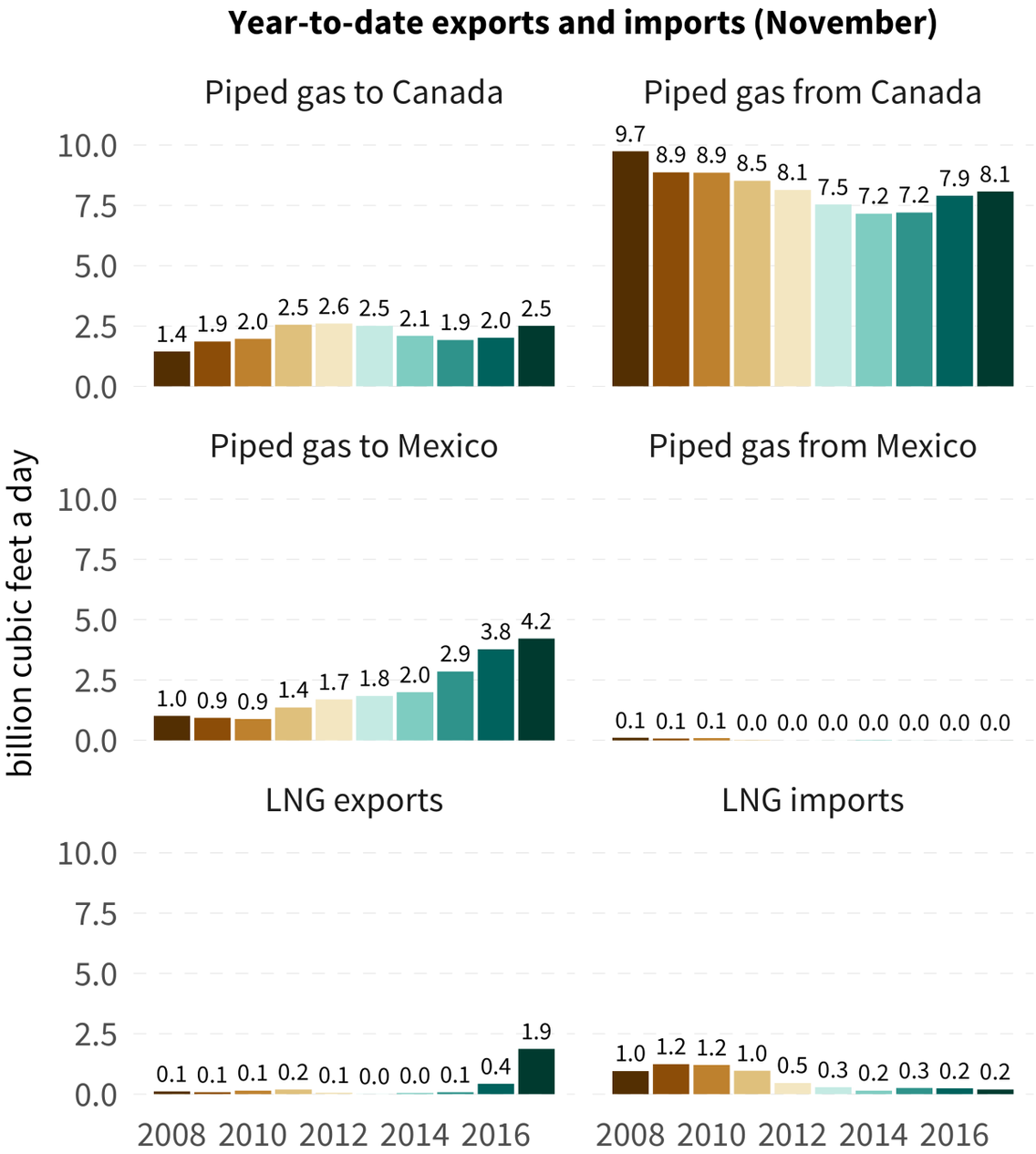
But also understand what the reader really cares about



Source: The New York Times, 2016 Presidential Election Results

Present the facts (left), anticipate the reader's needs (right)



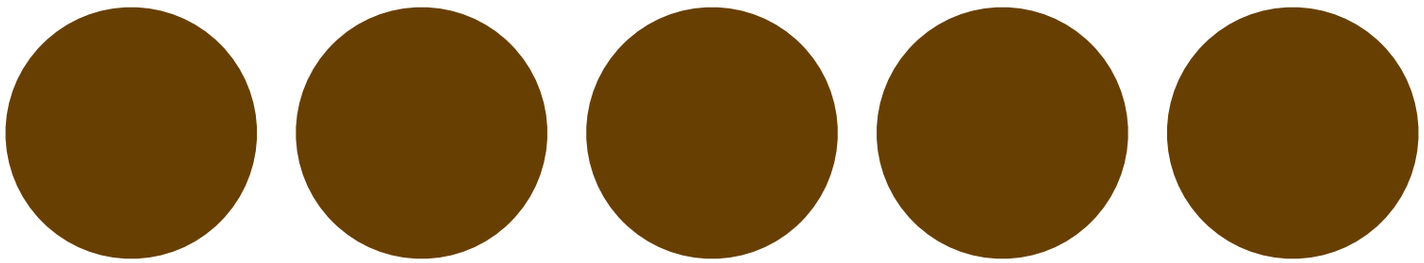


Source: Visualization by the author; data from Energy Information Administration, Gas Imports/Exports

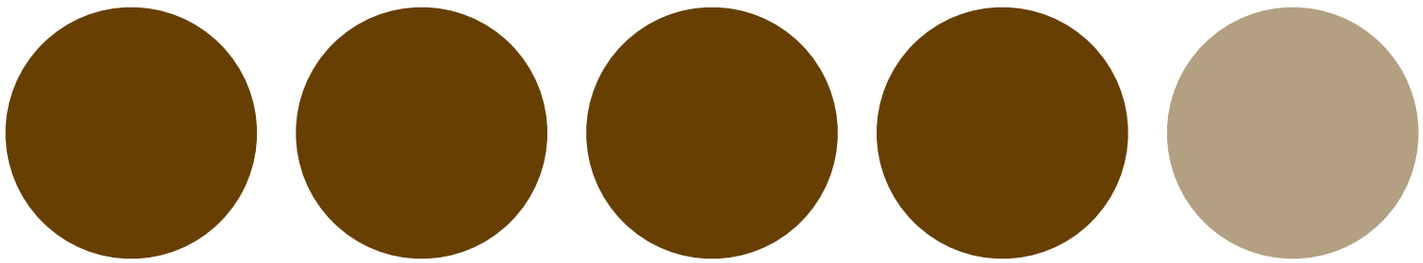
Design principles

(aka control what the eye sees)

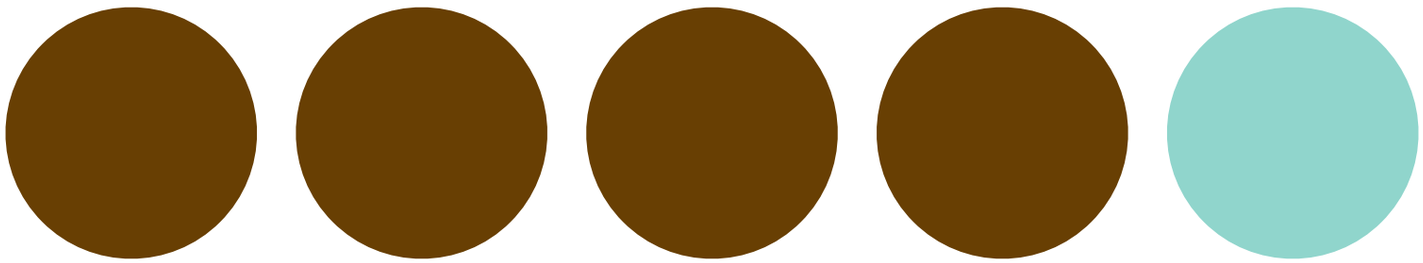
Control what the eye sees: No direction



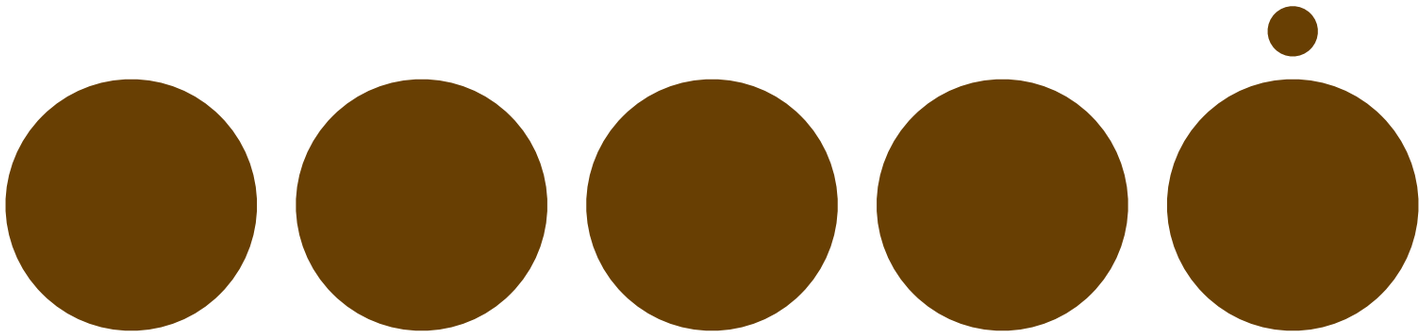
Control what the eye sees: Weight



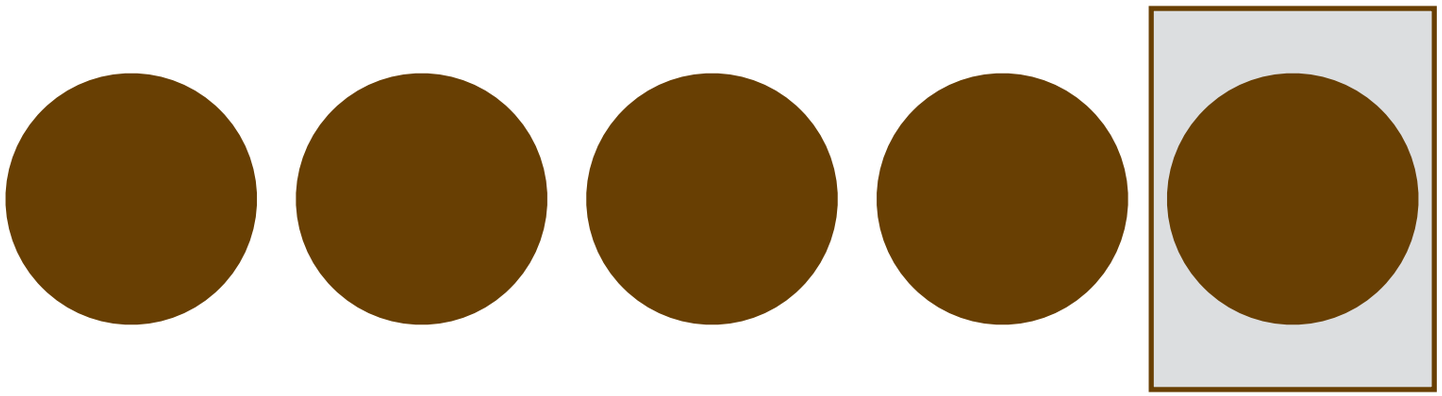
Control what the eye sees: Color



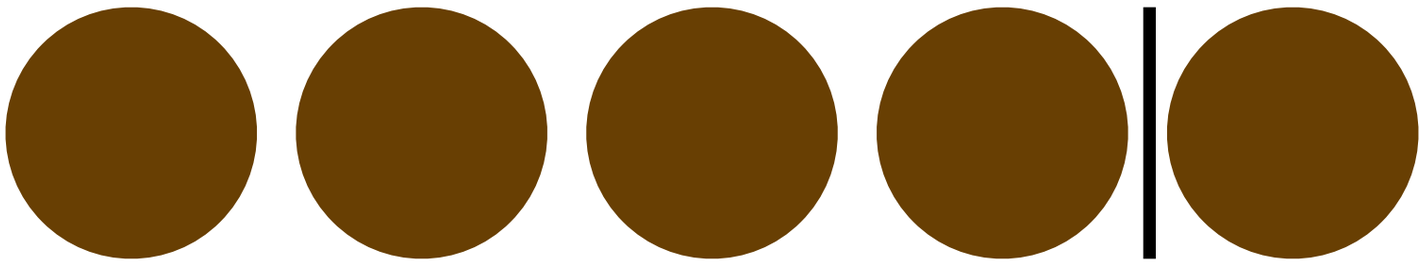
Control what the eye sees: Annotation



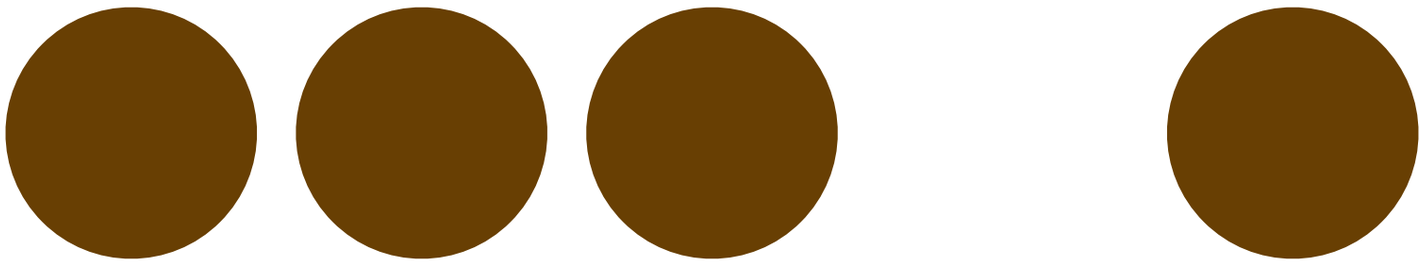
Control what the eye sees: Enclosure



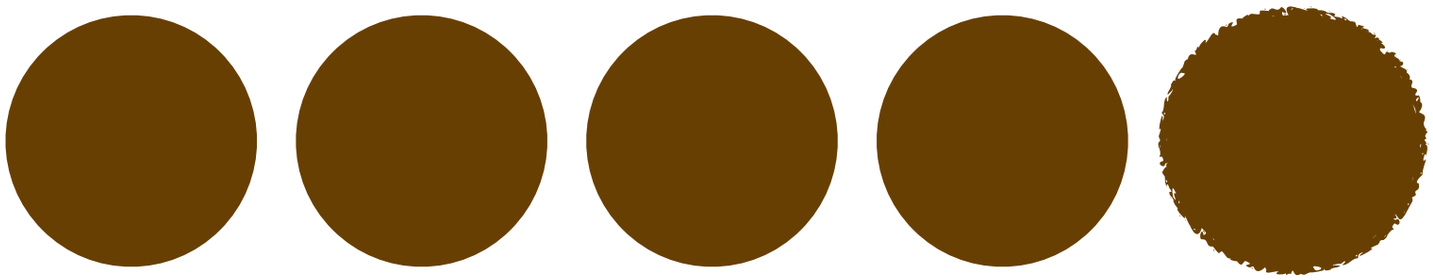
Control what the eye sees: Separation



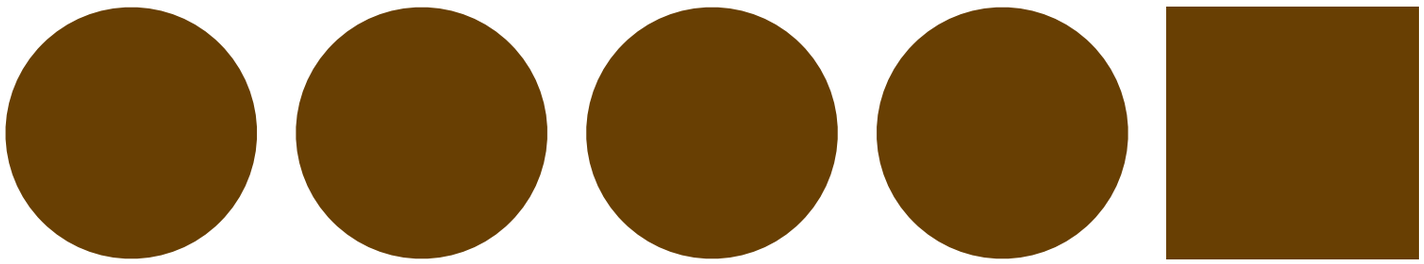
Control what the eye sees: Space



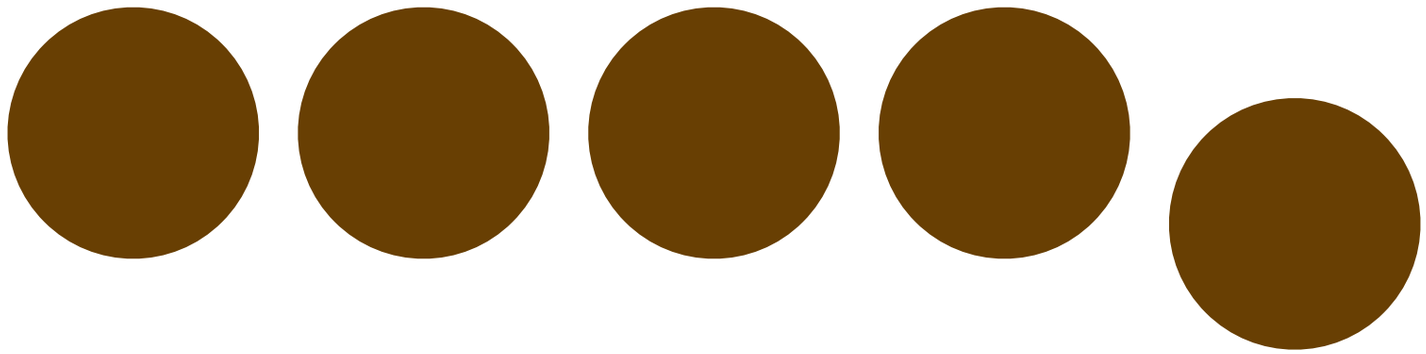
Control what the eye sees: Texture



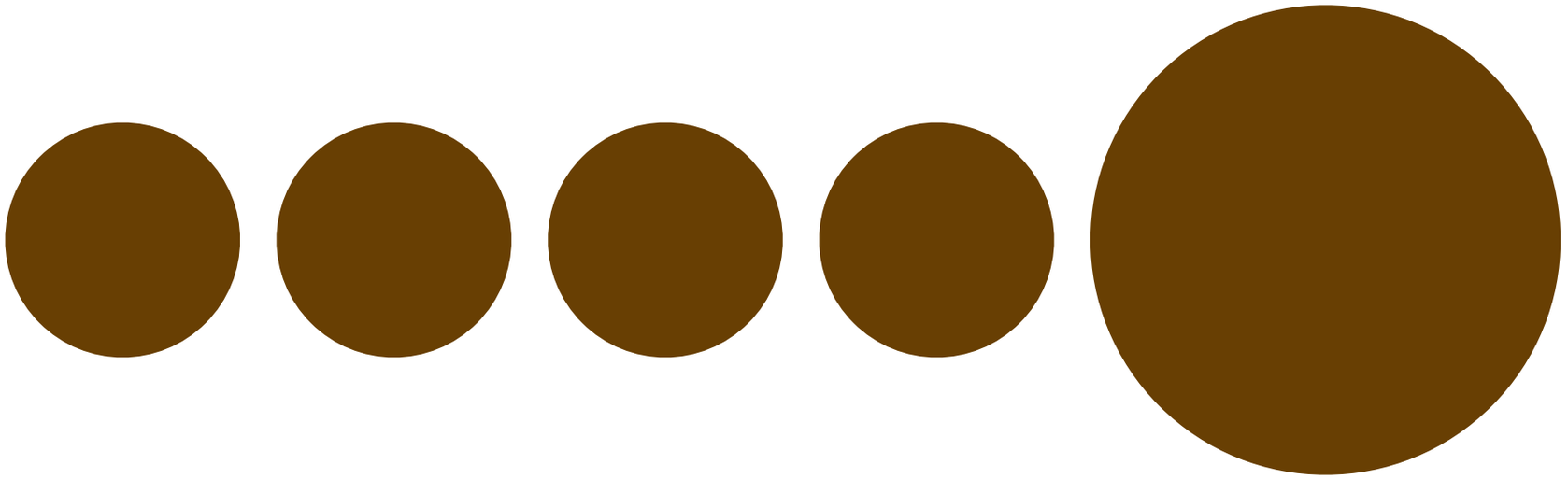
Control what the eye sees: Form



Control what the eye sees: Direction



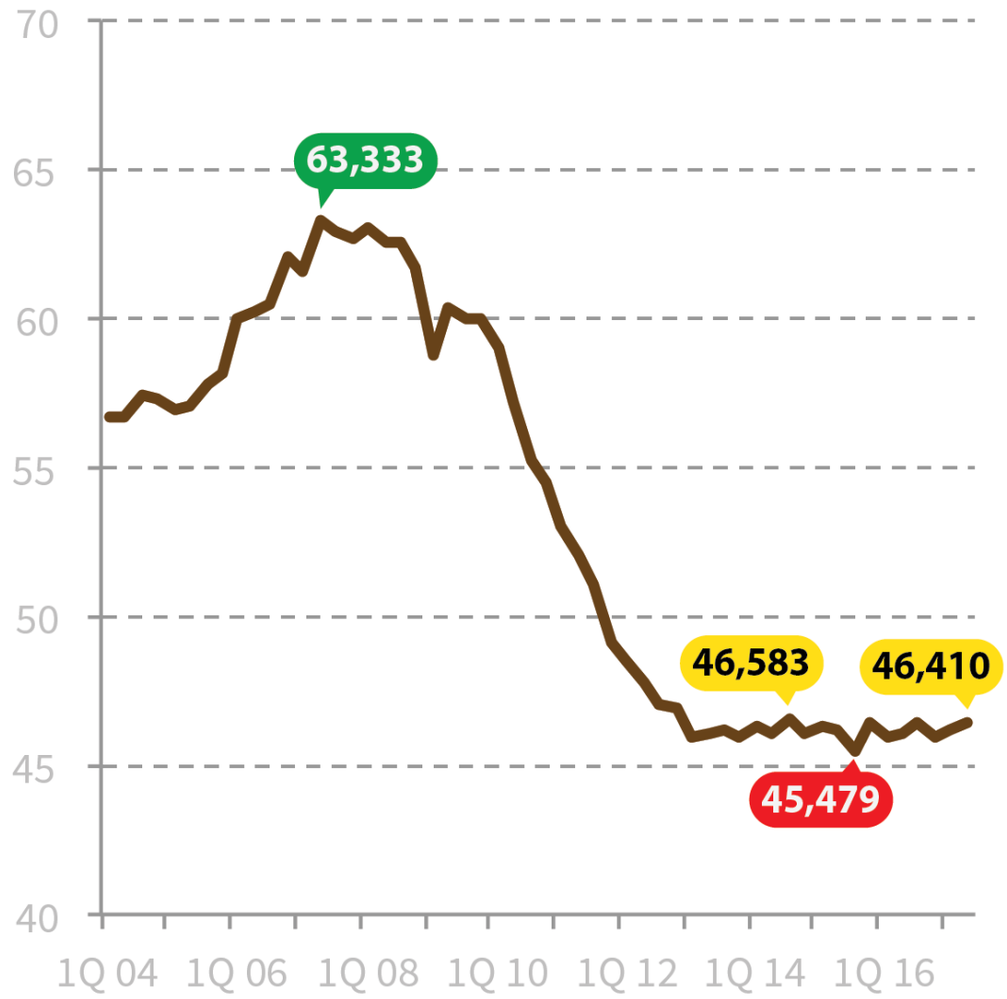
Control what the eye sees: Size



Design principles (in practice)

Greek GDP stagant; still below 2014 Q3

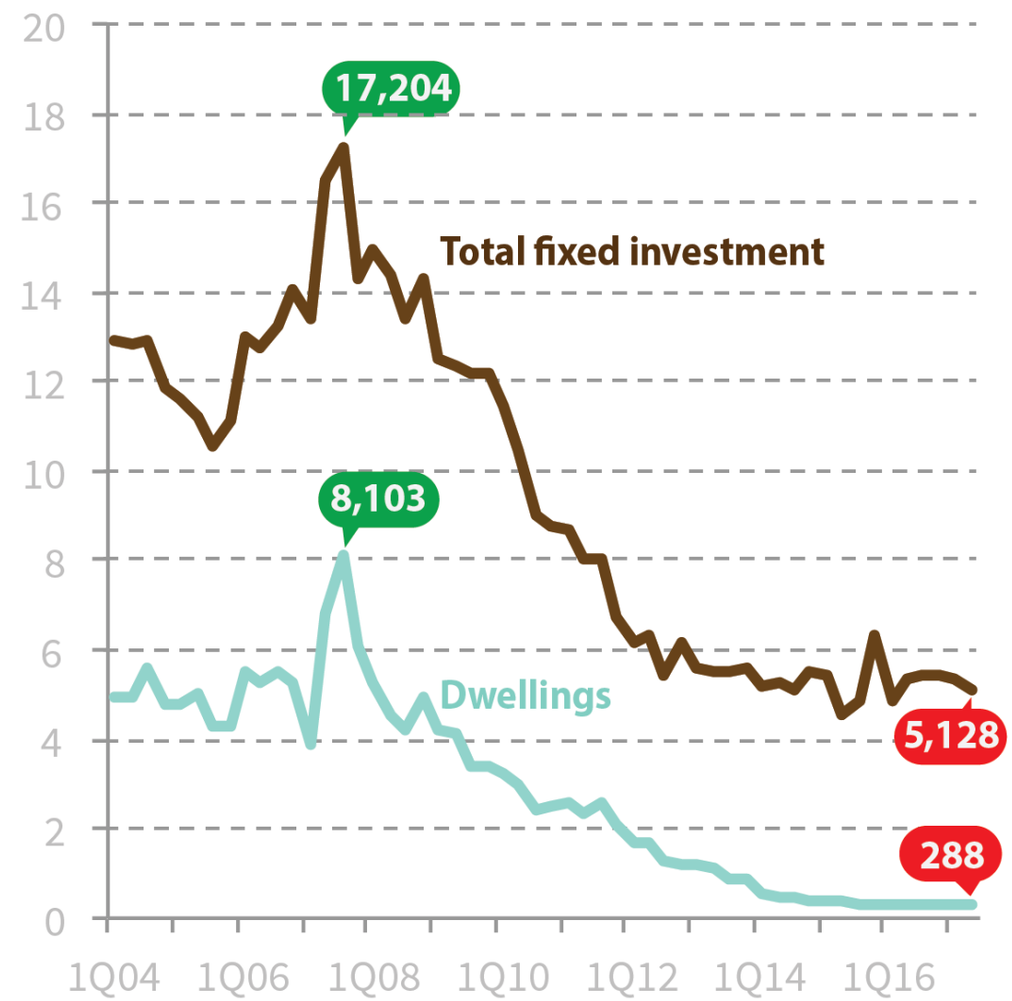
in billion €, chain-linked volumes of reference year 2010



Seasonally adjusted data. Source: Hellenic Statistical Authority

Greek investment -70% vs. 2007 Q3

in billion €, chain-linked volumes of reference year 2010

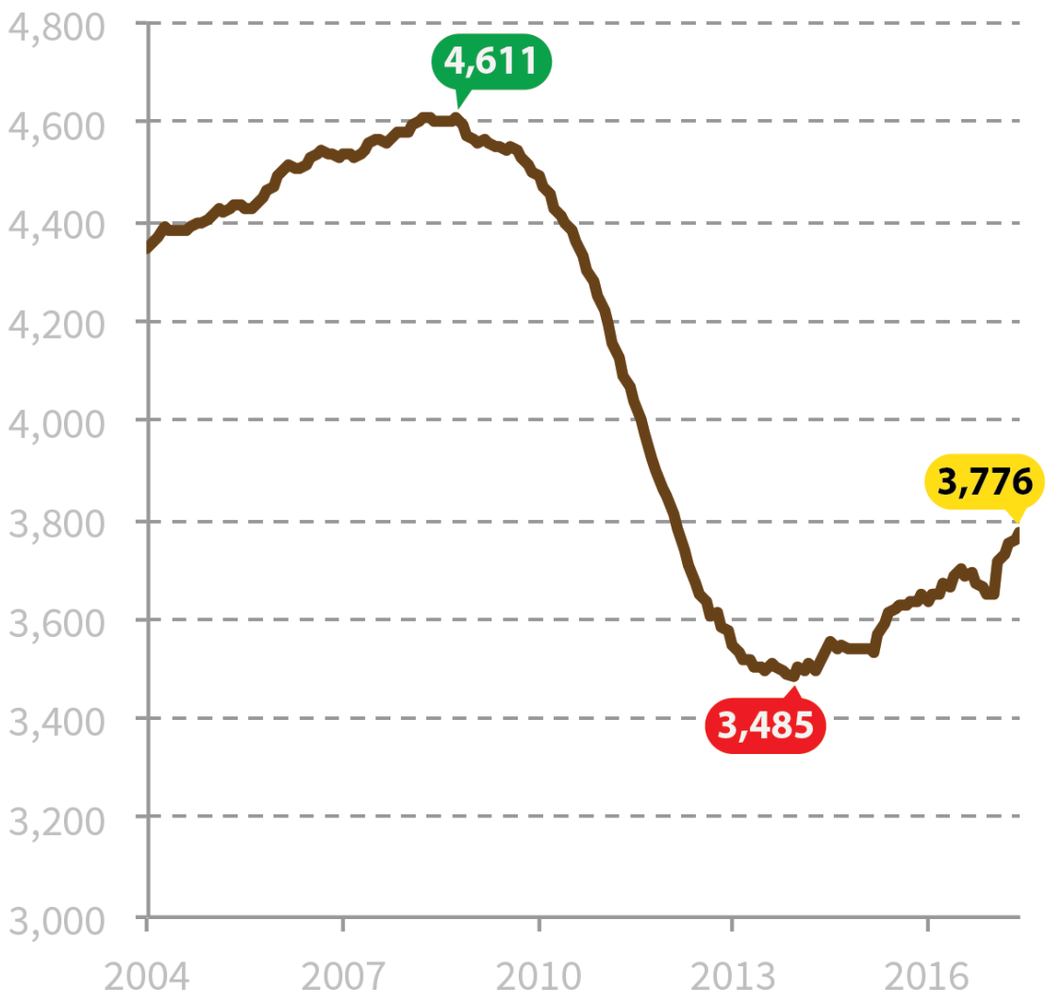


Seasonally adjusted data. Source: Hellenic Statistical Authority

Source: Visualization by the author (link)

Greek employment +8.4% vs. 2013

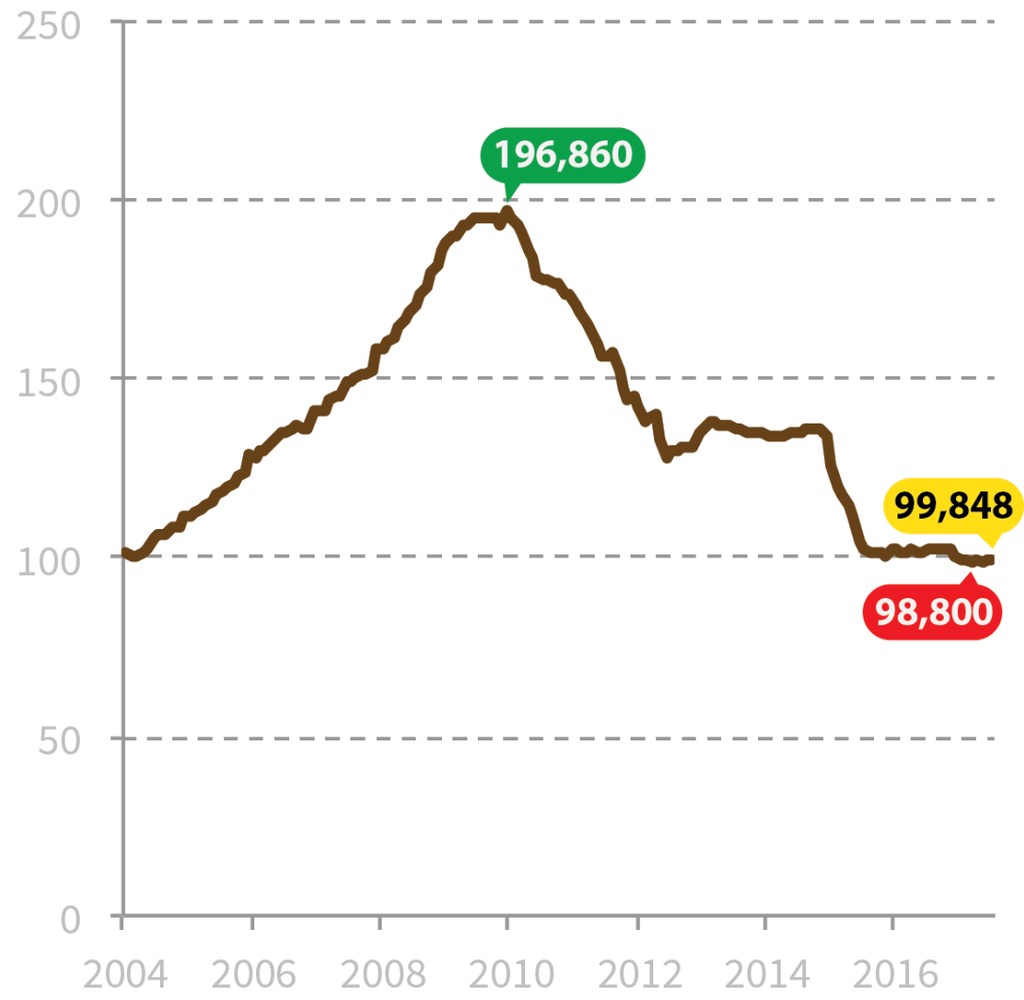
in thousands



Seasonally adjusted data. Source: Hellenic Statistical Authority

Greek deposits -49% vs. Dec 2009

in billion €



Households and non-profit institutions. Source: Bank of Greece

Source: Visualization by the author (link)

Design examples

ian bremmer @ianbremmer · Feb 12

% worried about their healthcare system, January 2018

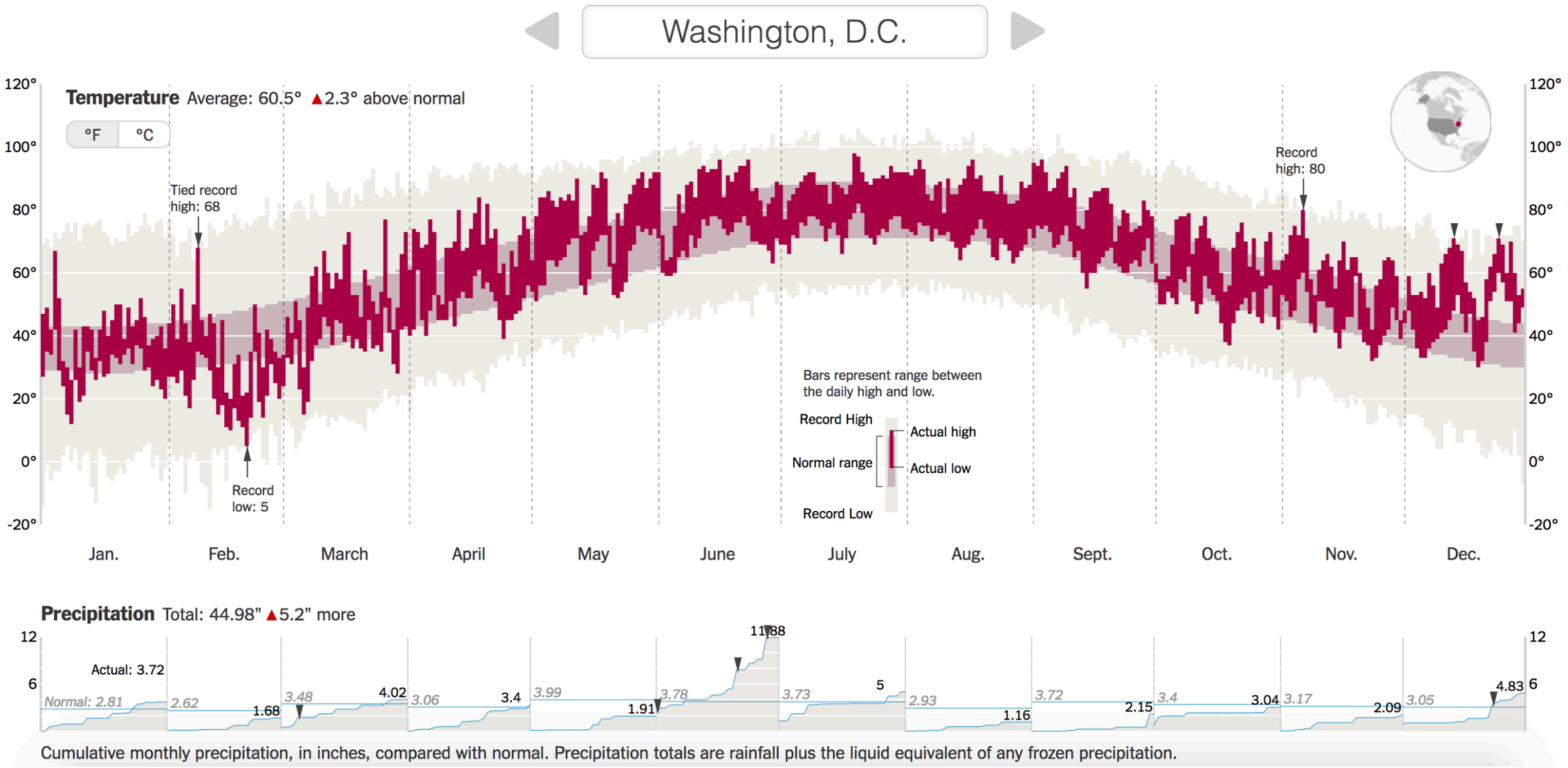
Hungary 72%
Poland 62%
Brazil 46%
UK 42%
China 38%
US 38% (lower than I expected)
Russia 27%
Japan 17%
France 14%
India 14%
Germany 13%
S Korea 6%
Turkey 3%

Ipsos

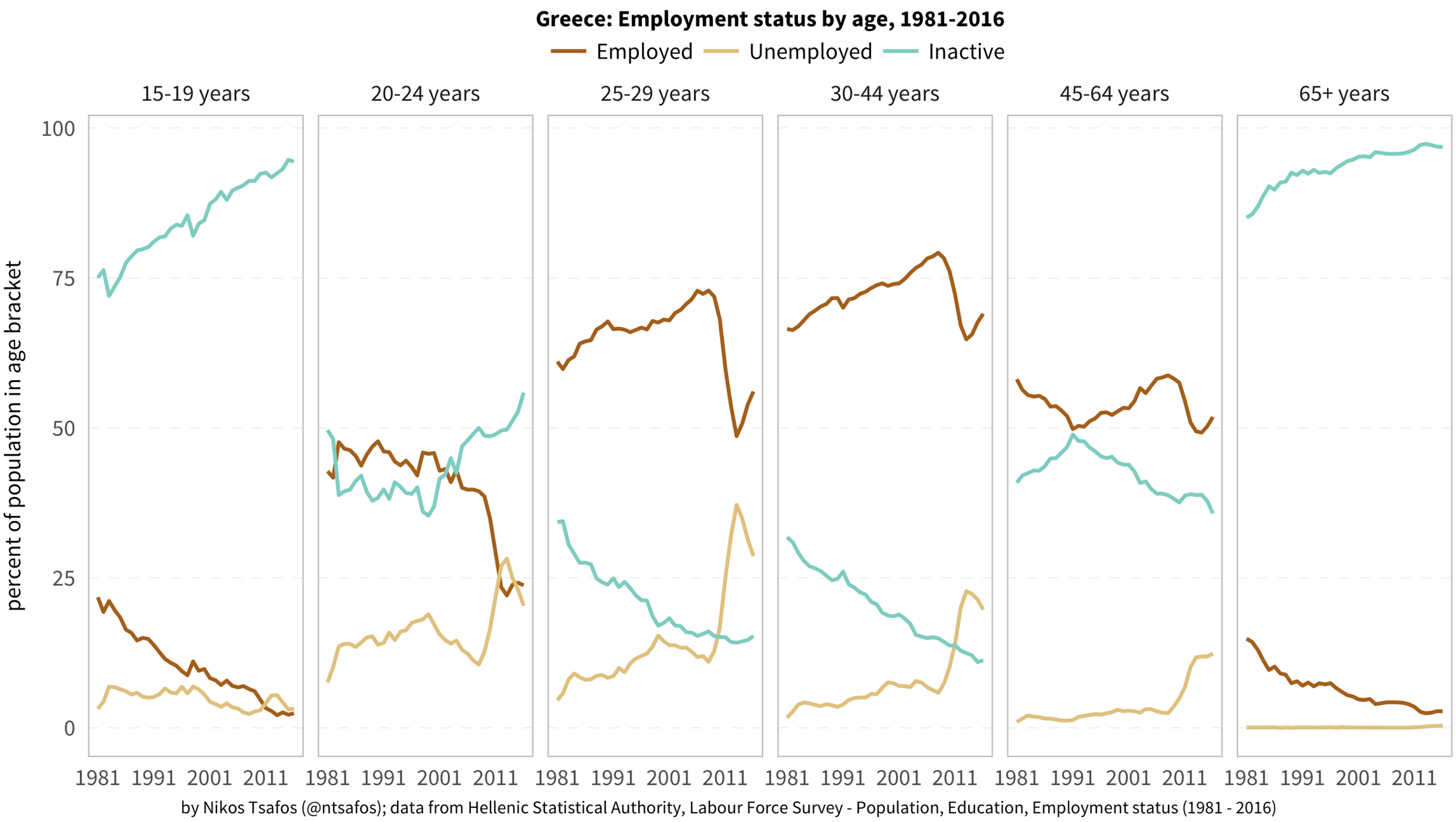
35 205 348

U.S. | How Much Warmer Was Your City in 2015?

SHARE



Source: New York Times, How Much Warmer Was Your City in 2015? (February 19, 2016)



Source: Visualization by the author; data from Hellenic Statistical Authority

How Trump compares with past presidents

392 DAYS | 4 YEARS | 8 YEARS

● Approval rating ○ Disapproval rating ○ Net approval

Barack Obama 2009-17

George W. Bush 2001-09

Bill Clinton 1993-2001

George H.W. Bush 1989-93

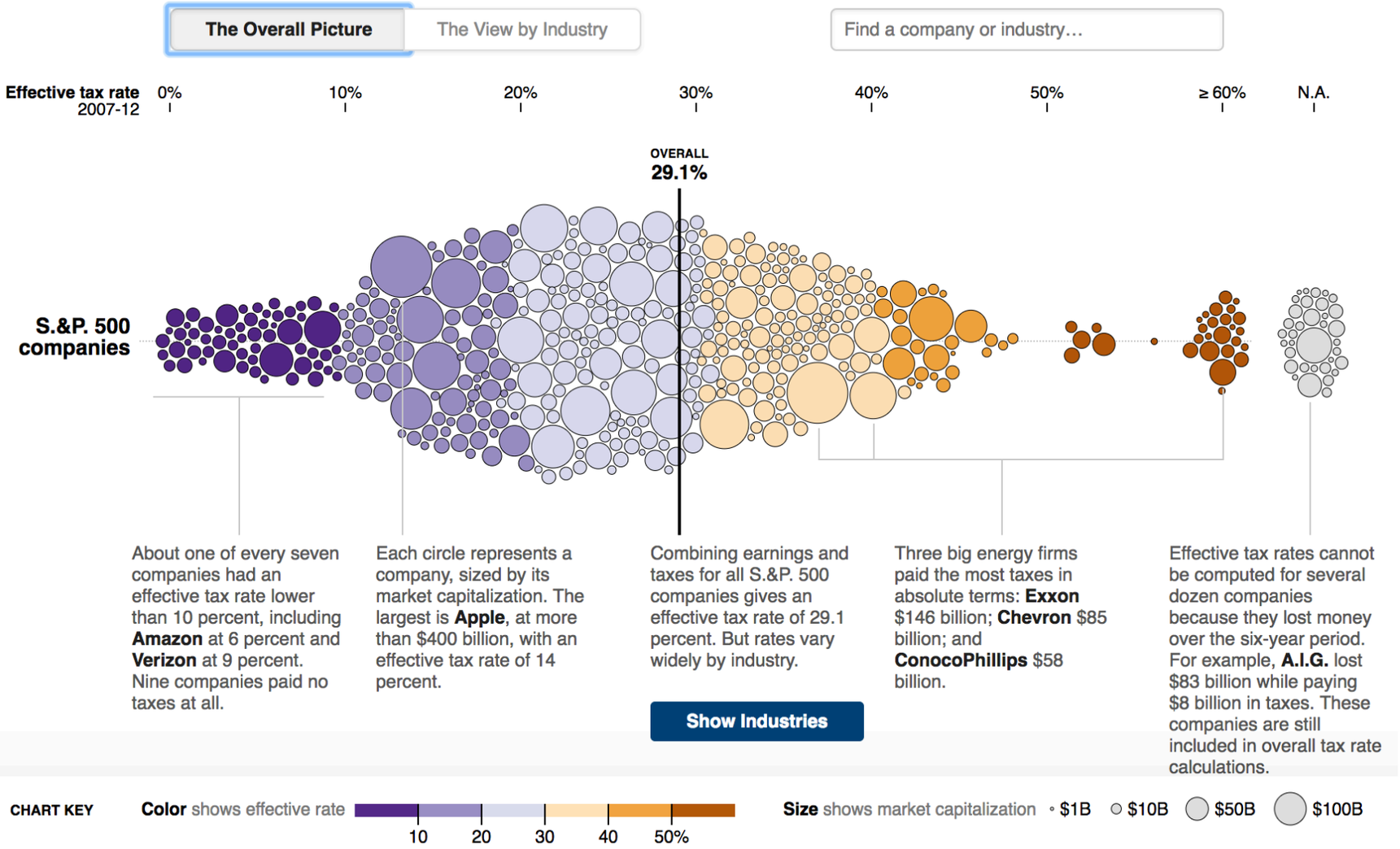
Ronald Reagan 1981-89

Jimmy Carter 1977-81

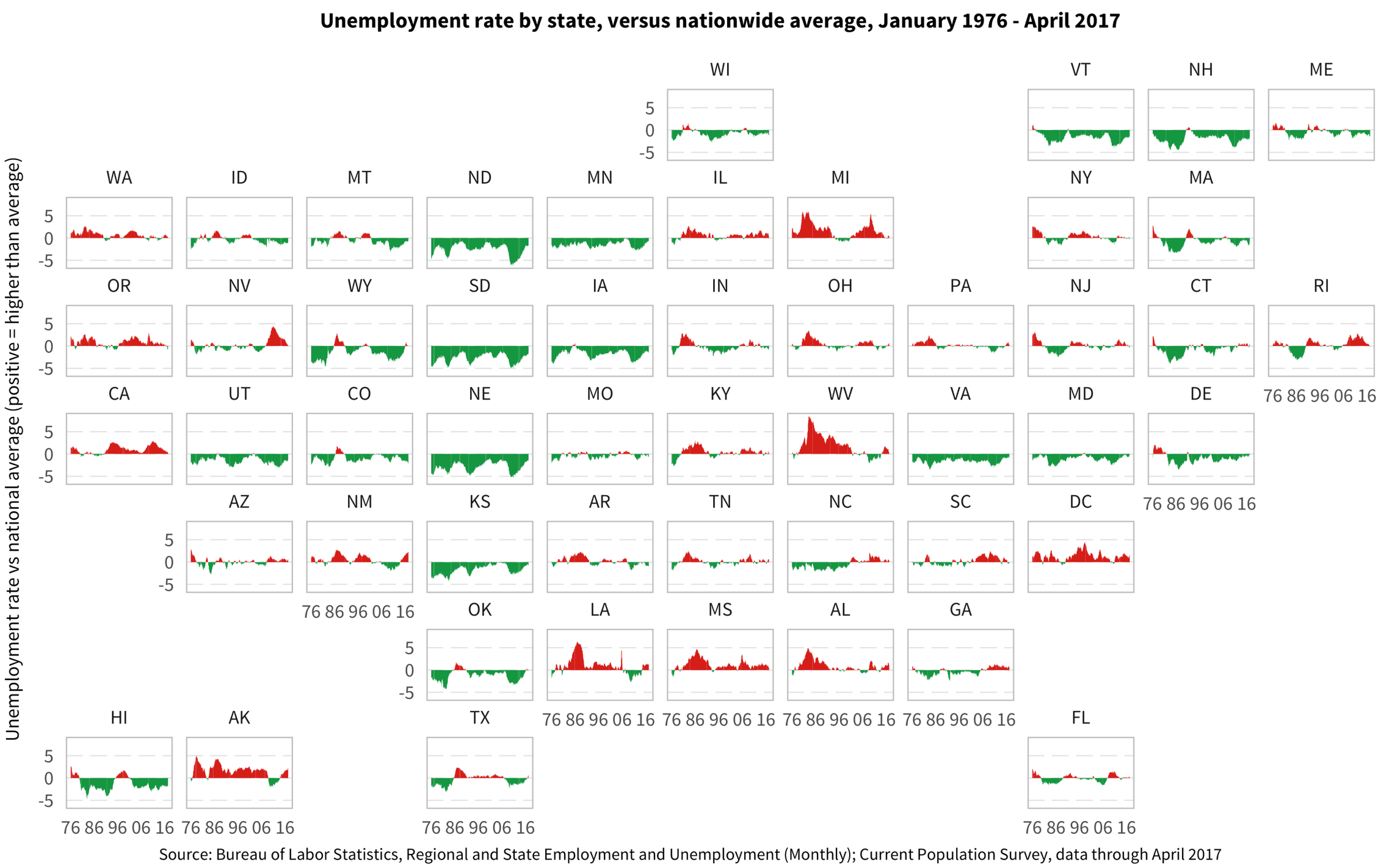
Source: FiveThirtyEight, How popular is Donald Trump?

Across U.S. Companies, Tax Rates Vary Greatly

Last week, in a Congressional hearing, Apple got grilled for its low-tax strategy. But not every business can copy that approach. Here is a look at what S.&P. 500 companies paid in corporate income taxes — federal, state, local and foreign — from 2007 to 2012, according to S&P Capital IQ. Related Article »



Source: The New York Times, Across U.S. Companies, Tax Rates Vary Greatly (May 25, 2013)



Source: Visualization by the author; data from Bureau of Labor Statistics

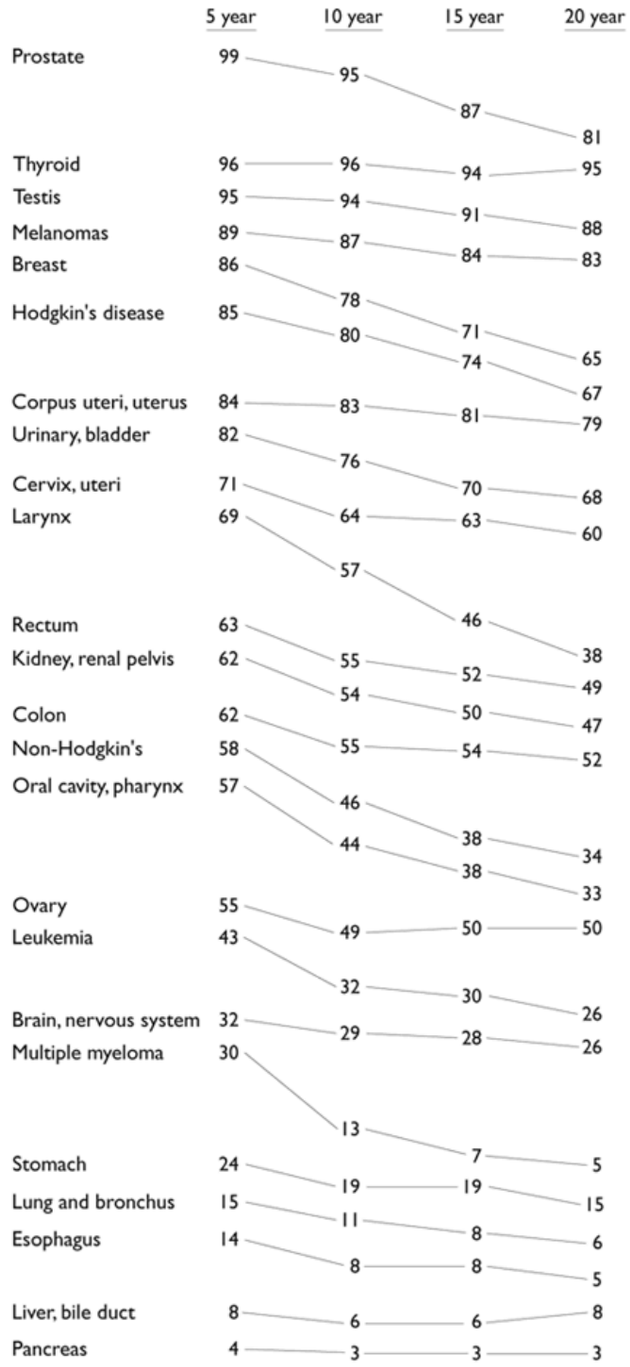
Current Receipts of Government as a Percentage of Gross Domestic Product, 1970 and 1979

Source: Tufte, Edward (1983). The Visual Display of Quantitative Information, p. 158

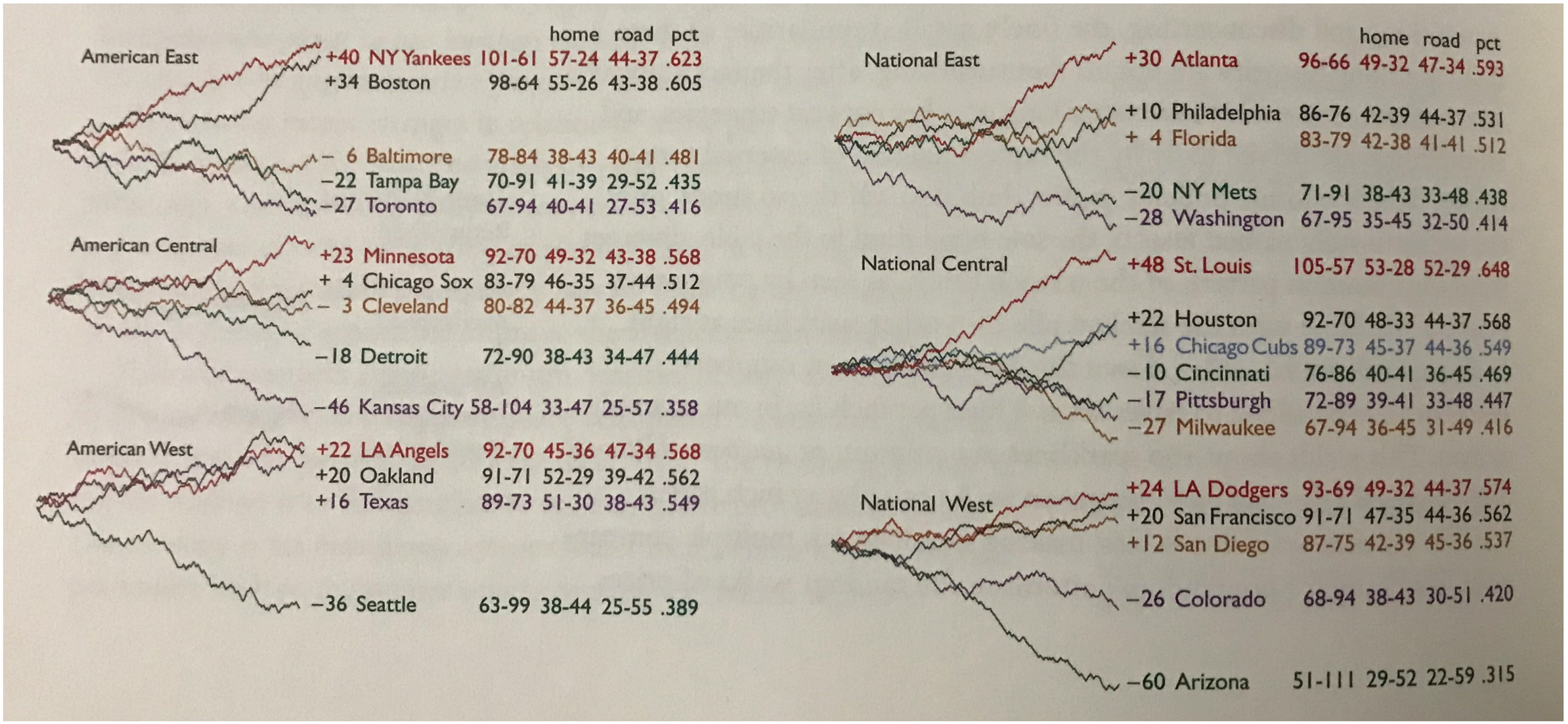
	Relative survival rate, % (SE)			
	5 years	10 years	15 years	20 years
Cancer site				
Oral cavity and pharynx	56·7 (1·3)	44·2 (1·4)	37·5 (1·6)	33·0 (1·8)
Oesophagus	14·2 (1·4)	7·9 (1·3)	7·7 (1·6)	5·4 (2·0)
Stomach	23·8 (1·3)	19·4 (1·4)	19·0 (1·7)	14·9 (1·9)
Colon	61·7 (0·8)	55·4 (1·0)	53·9 (1·2)	52·3 (1·6)
Rectum	62·6 (1·2)	55·2 (1·4)	51·8 (1·8)	49·2 (2·3)
Liver and intrahepatic bile duct	7·5 (1·1)	5·8 (1·2)	6·3 (1·5)	7·6 (2·0)
Pancreas	4·0 (0·5)	3·0 (0·5)	2·7 (0·6)	2·7 (0·8)
Larynx	68·8 (2·1)	56·7 (2·5)	45·8 (2·8)	37·8 (3·1)
Lung and bronchus	15·0 (0·4)	10·6 (0·4)	8·1 (0·4)	6·5 (0·4)
Melanomas	89·0 (0·8)	86·7 (1·1)	83·5 (1·5)	82·8 (1·9)
Breast	86·4 (0·4)	78·3 (0·6)	71·3 (0·7)	65·0 (1·0)
Cervix uteri	70·5 (1·6)	64·1 (1·8)	62·8 (2·1)	60·0 (2·4)
Corpus uteri and uterus, NOS	84·3 (1·0)	83·2 (1·3)	80·8 (1·7)	79·2 (2·0)
Ovary	55·0 (1·3)	49·3 (1·6)	49·9 (1·9)	49·6 (2·4)
Prostate	98·8 (0·4)	95·2 (0·9)	87·1 (1·7)	81·1 (3·0)
Testis	94·7 (1·1)	94·0 (1·3)	91·1 (1·8)	88·2 (2·3)
Urinary bladder	82·1 (1·0)	76·2 (1·4)	70·3 (1·9)	67·9 (2·4)
Kidney and renal pelvis	61·8 (1·3)	54·4 (1·6)	49·8 (2·0)	47·3 (2·6)
Brain and other nervous system	32·0 (1·4)	29·2 (1·5)	27·6 (1·6)	26·1 (1·9)
Thyroid	96·0 (0·8)	95·8 (1·2)	94·0 (1·6)	95·4 (2·1)
Hodgkin's disease	85·1 (1·7)	79·8 (2·0)	73·8 (2·4)	67·1 (2·8)
Non-Hodgkin lymphomas	57·8 (1·0)	46·3 (1·2)	38·3 (1·4)	34·3 (1·7)
Multiple myeloma	29·5 (1·6)	12·7 (1·5)	7·0 (1·3)	4·8 (1·5)
Leukaemias	42·5 (1·2)	32·4 (1·3)	29·7 (1·5)	26·2 (1·7)

Rates derived from SEER 1973–98 database (both sexes, all ethnic groups).[12]
NOS=not otherwise specified.

Table 4: **Most recent period estimates of relative survival rates, by cancer site**



Source: Tufte, Edward. Cancer survival rates: tables, slopegraphs, barcharts

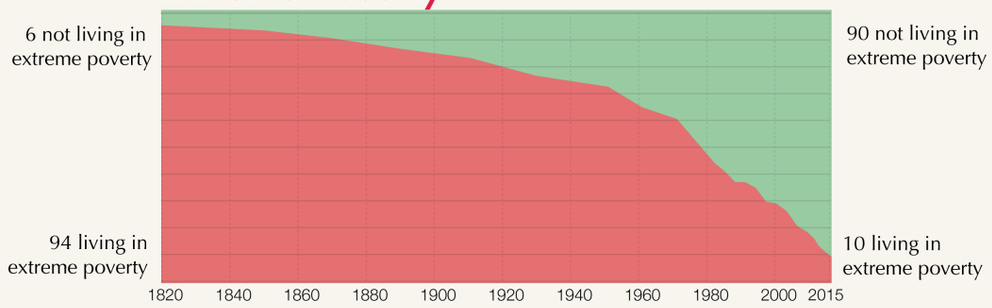


Source: Tufte, Edward (1983). The Visual Display of Quantitative Information, p. 174

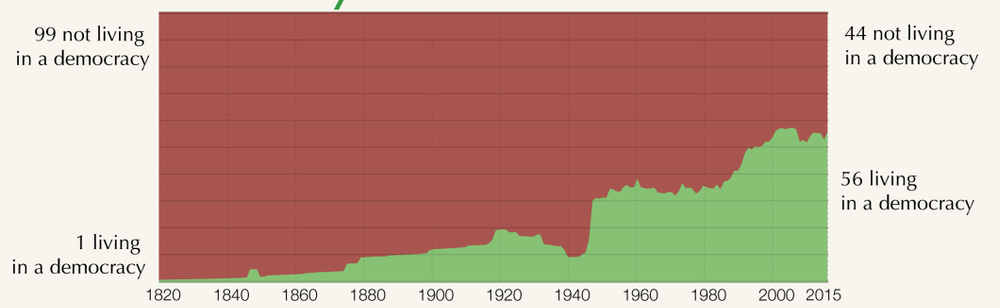
The World as 100 People
over the last two centuries

Our World in Data

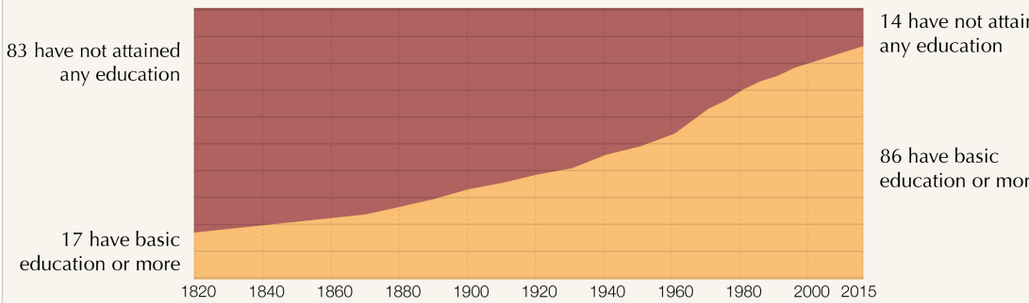
Extreme Poverty



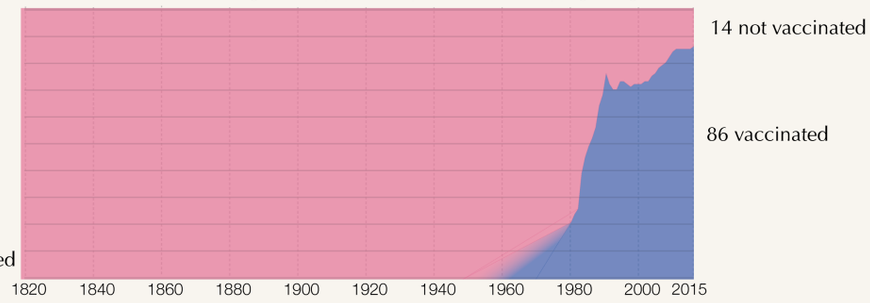
Democracy



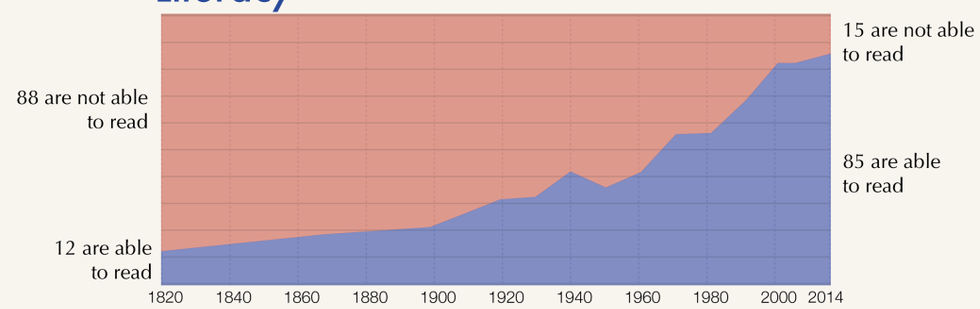
Basic Education



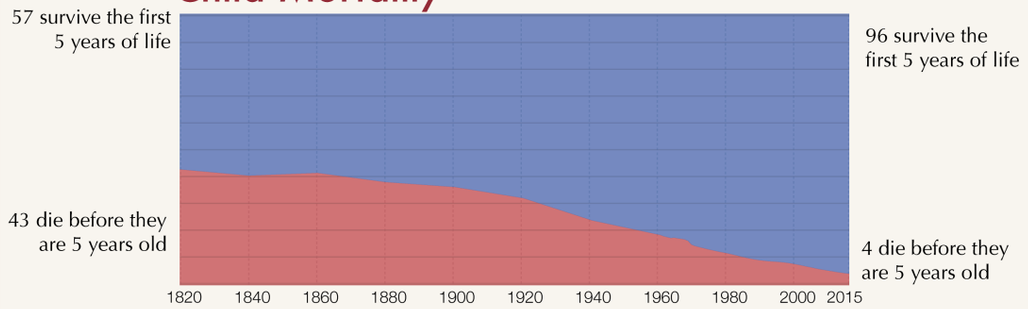
Vaccination against diphtheria, pertussis (whooping cough), and tetanus



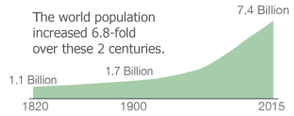
Literacy



Child Mortality

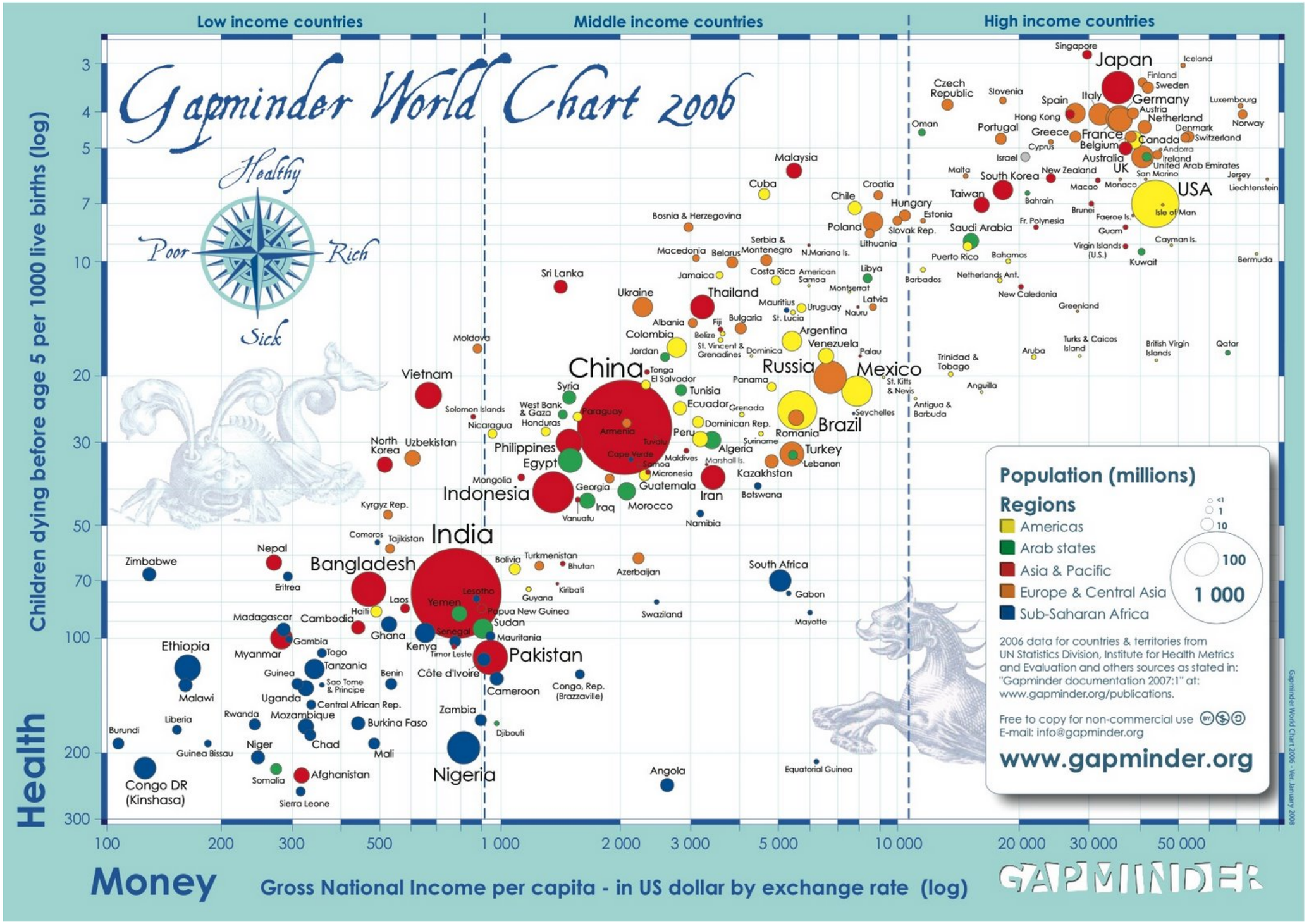


Data sources:
Extreme Poverty: Bourguignon & Morrison (2002) up to 1970 – World Bank 1981 and later (2015 is a projection).
Vaccination: WHO (Global data are available for 1980 to 2015 – the DPT3 vaccination was licenced in 1949)
Education: OECD for the period 1820 to 1960. IIASA for the time thereafter.
Literacy: OECD for the period 1820 to 1990. UNESCO for 2004 and later.
Democracy: Politiy IV index (own calcluation of global population share)
Colonialism: Wimmer and Min (own calcluation of global population share)
Continent: HYDE database
Child mortality: up to 1960 own caluclations based on Gapminder; World Bank thereafter



All these visualizations are from OurWorldInData.org an online publication that presents the empirical evidence on how the world is changing.

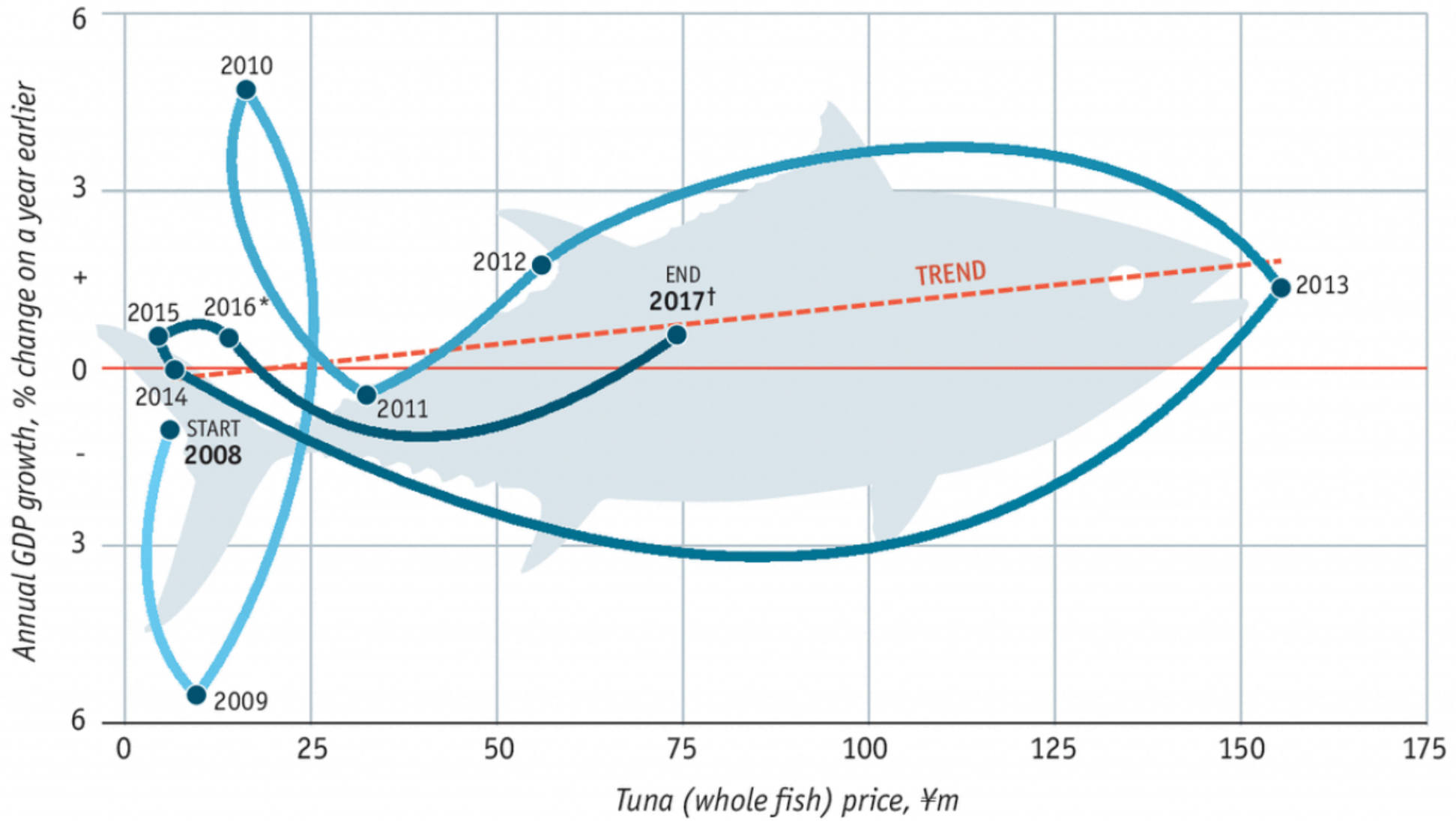
Source: Max Roser (2017), The short history of global living conditions and why it matters that we know it



Source: Hans Rosling, Wealth Equals Health

Economy of scales

Tsukiji market, first tuna prices v Japan GDP growth, 2008-17



Sources: IMF; press reports

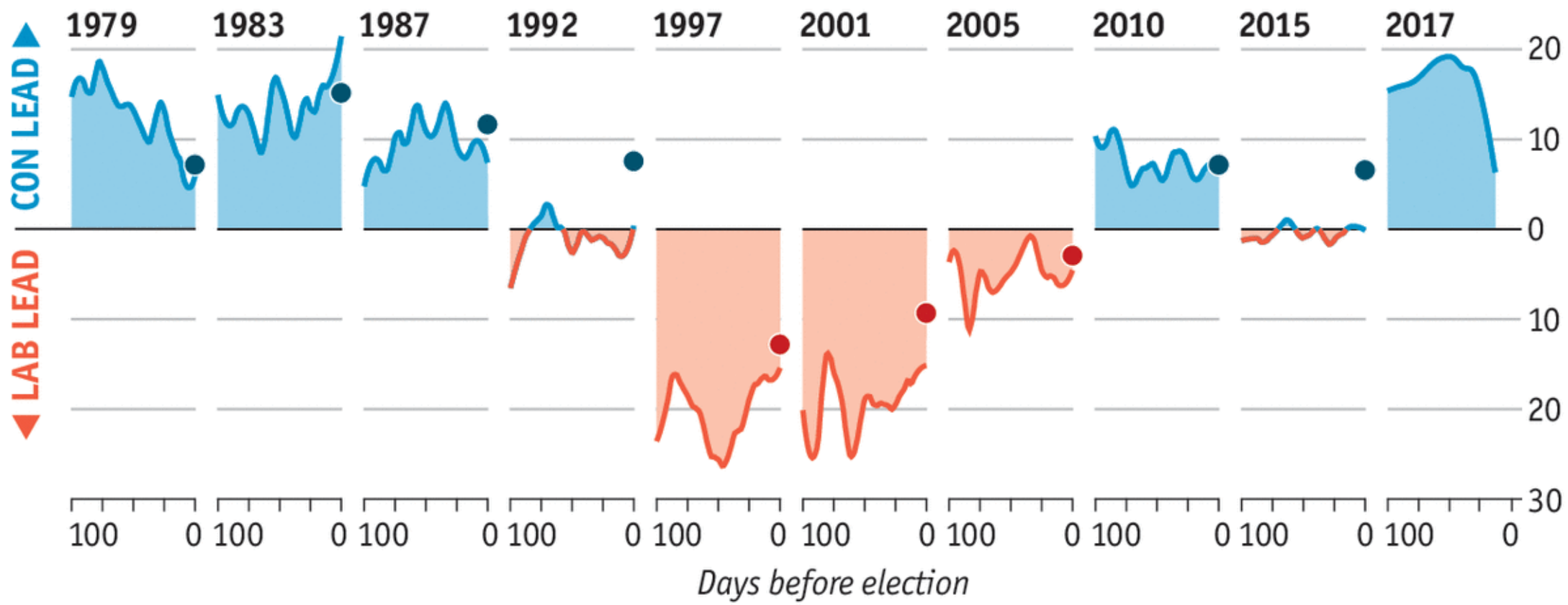
*GDP estimate †GDP forecast

Economist.com

Source: The Economist, Can tuna prices predict Japan's GDP growth?

Landslides and slip-ups

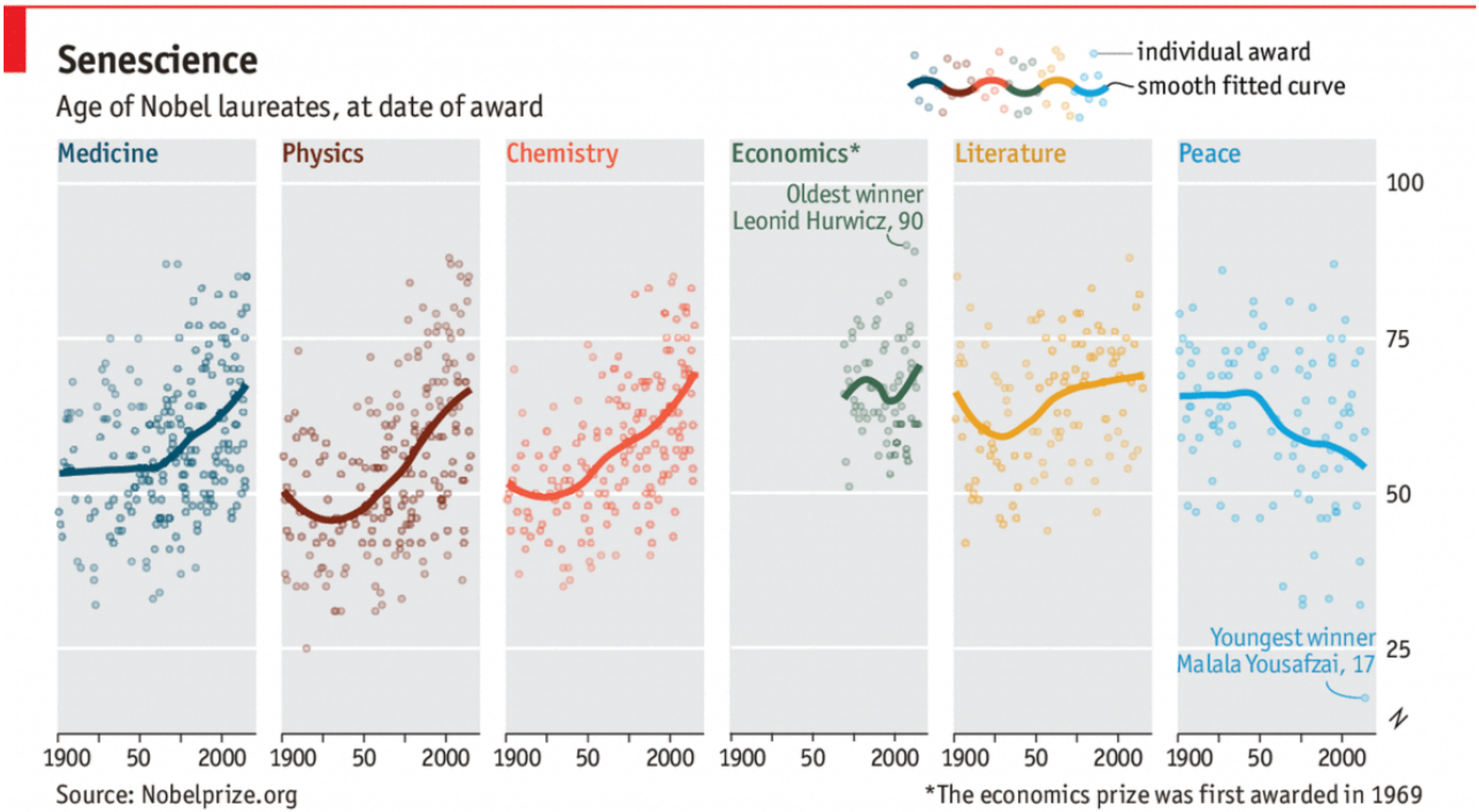
Britain, polling gap between Conservative and Labour
100 days before general elections, percentage points



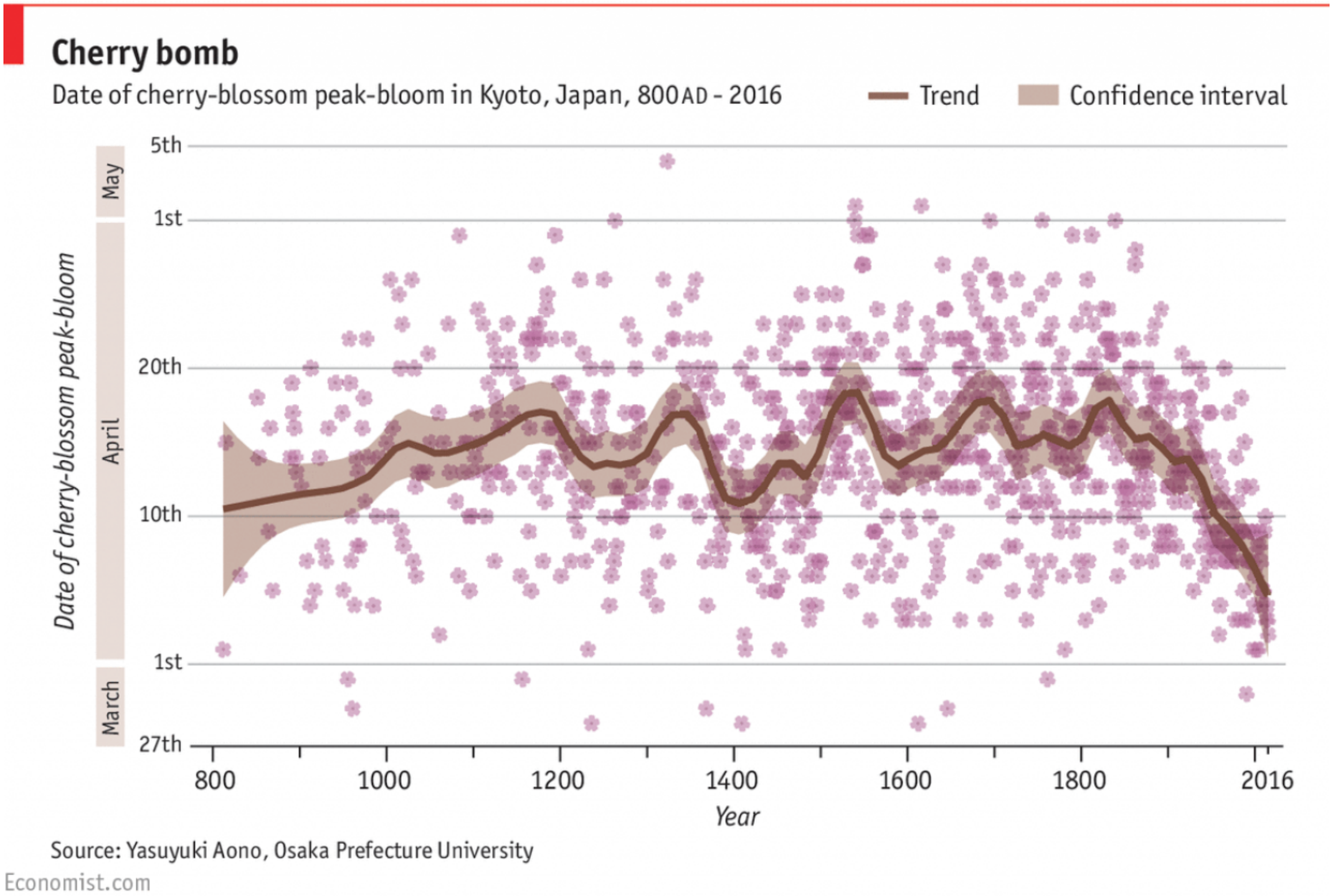
Sources: Britain Elects; ComRes; ICM; Opinium; YouGov; Will Jennings; *The Economist*

Economist.com

Source: The Economist, Are British pollsters headed towards another miss?



Source: The Economist, Greying of the Nobel laureates

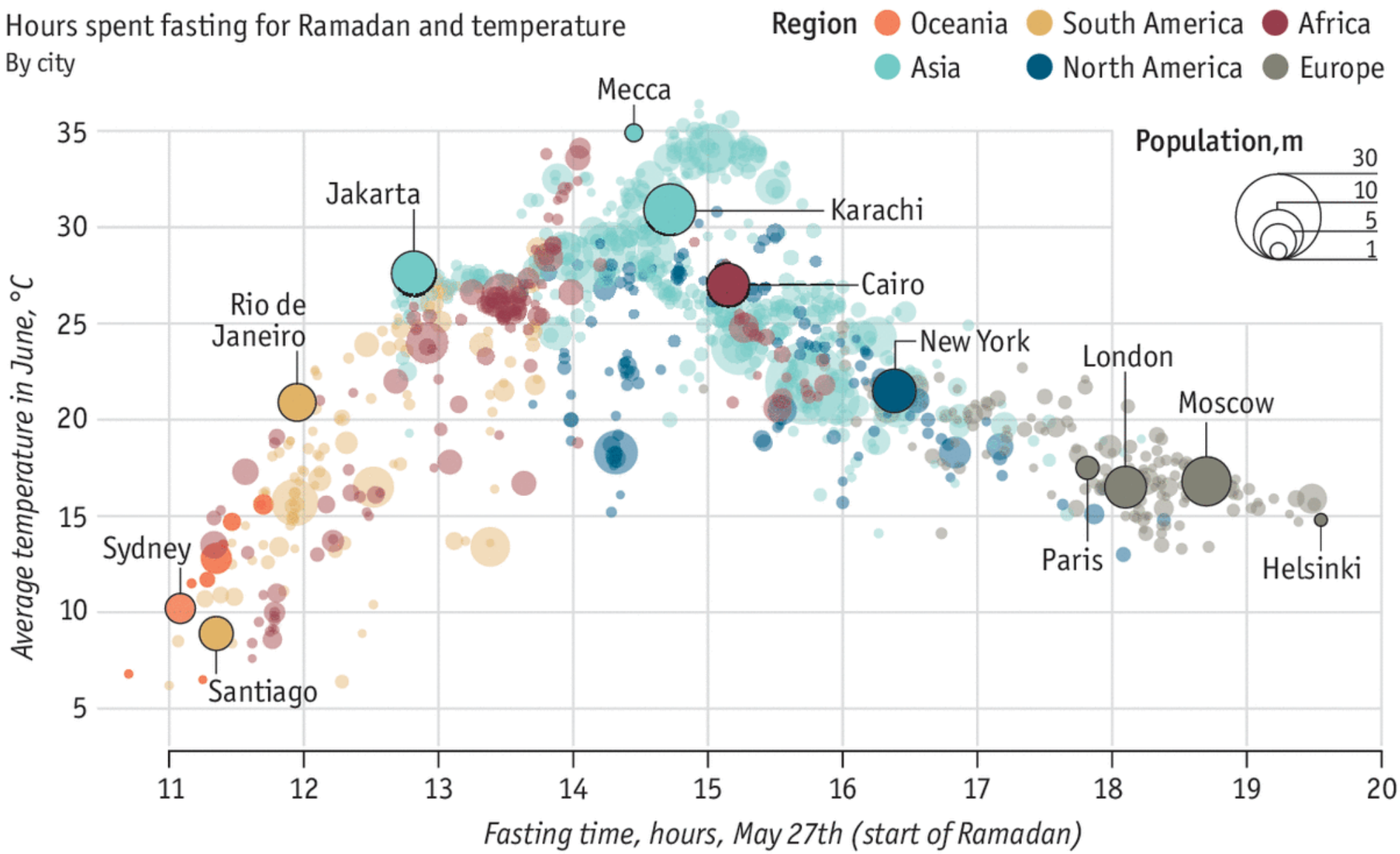


Source: The Economist, Japan's cherry blossoms are emerging increasingly early

Cruel summer

Hours spent fasting for Ramadan and temperature

By city



Sources: Al Adhan, WorldClim

Economist.com

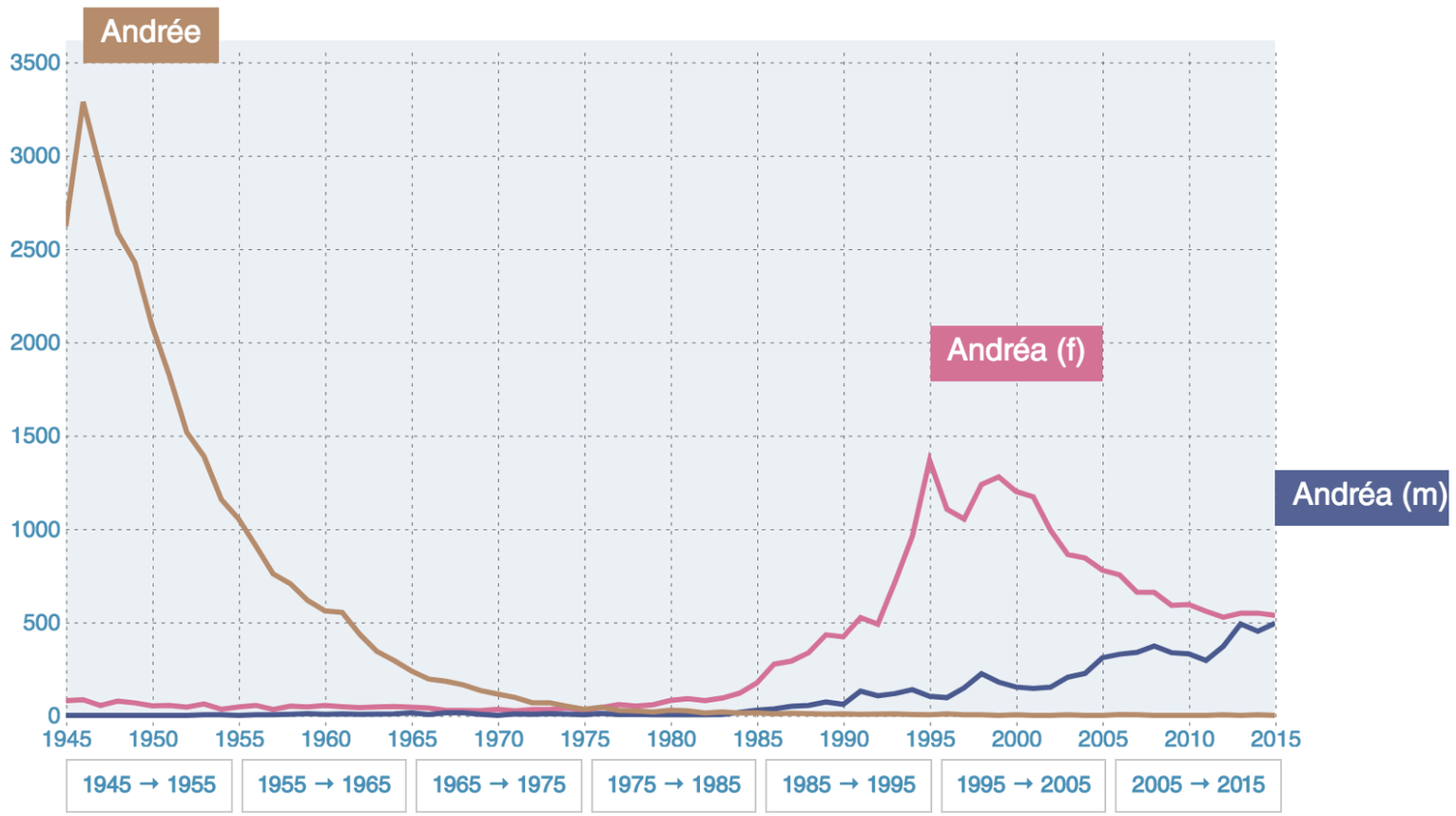
Source: The Economist, Which cities are the toughest to observe Ramadan in?

De 1945 à 2015 :

70 ans de prénoms en France

Recherche...

Aaron Abdallah Abdel Abdelaziz Abdelkader Abdelkrim Abdellah Abdoulaye Abel Achille Adam Adel Adelaide Adèle Adelie Adeline Adem Adil Adrian Adriana Adriano Adrien Adrienne Agathe Agnès Ahmed Ahmet Aicha Aida Aime Aimée Aissa Aissata Alain Alan Alban Albane Albert Albin Aldo Alessandro Alessio Alex Alexander Alexandra Alexandre Alexandrine Alexane Alexia Alexiane Alexis Alfred Ali Alice Alicia Alienor Aline Alison Alix Alix Alizée Alois Alphonse Alya Alyssa Amadou Amal Amalia Amanda Amandine Amar Amaury Ambre Ambrine Ambroise Amel Amélia Amélie Ameline Amina Aminata Amine Amir Amira Amy Anae Anael Anaelle Anaïs Anas Anastasia Anatole André Andréa Andréa Andreas Andrée Andrew Andy Ange Angel Angela Angèle Angelina Angeline Angelique Angelo Angie Anick Anis Anissa Anita Anna Annabelle Anne-Cécile Anne-Charlotte Anne-Claire Anne-Gaelle Anne-Laure Anne-Lise Anne-Marie Anne-Sophie Anne Annette



Partagez ces résultats avec ce lien :

https://dataaddict.fr/prenoms/#andrea-f-f,andrea

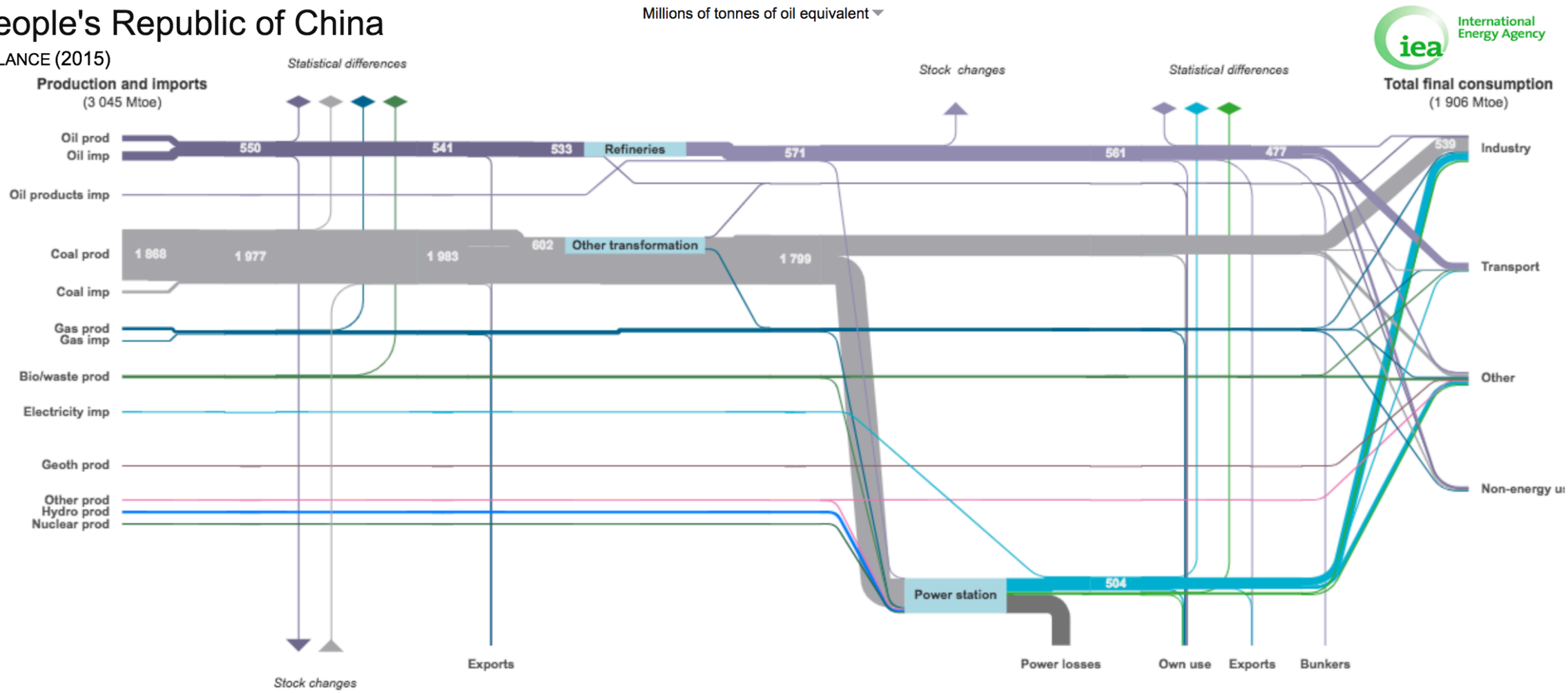
Share 501 Tweeter Star

Les prénoms sélectionnés sont les plus courants en France, ils ont été donnés au moins 2000 fois entre 1945 et 2015. Source :Insee - Fichier des prénoms (Édition 2016)

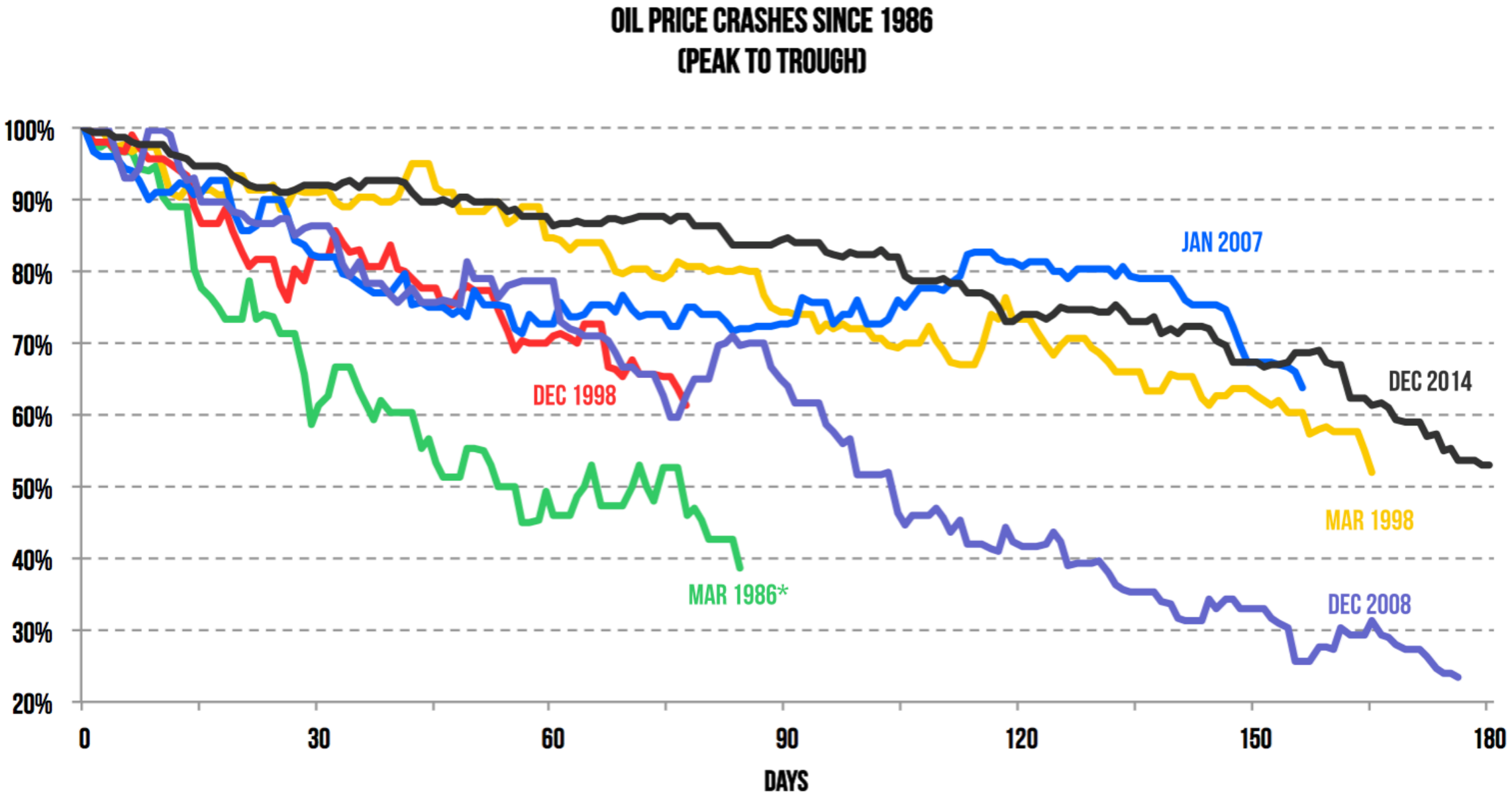
Source: https://dataaddict.fr/prenoms/#andrea-f-f,andrea-m-h,andree-f

People's Republic of China

BALANCE (2015)



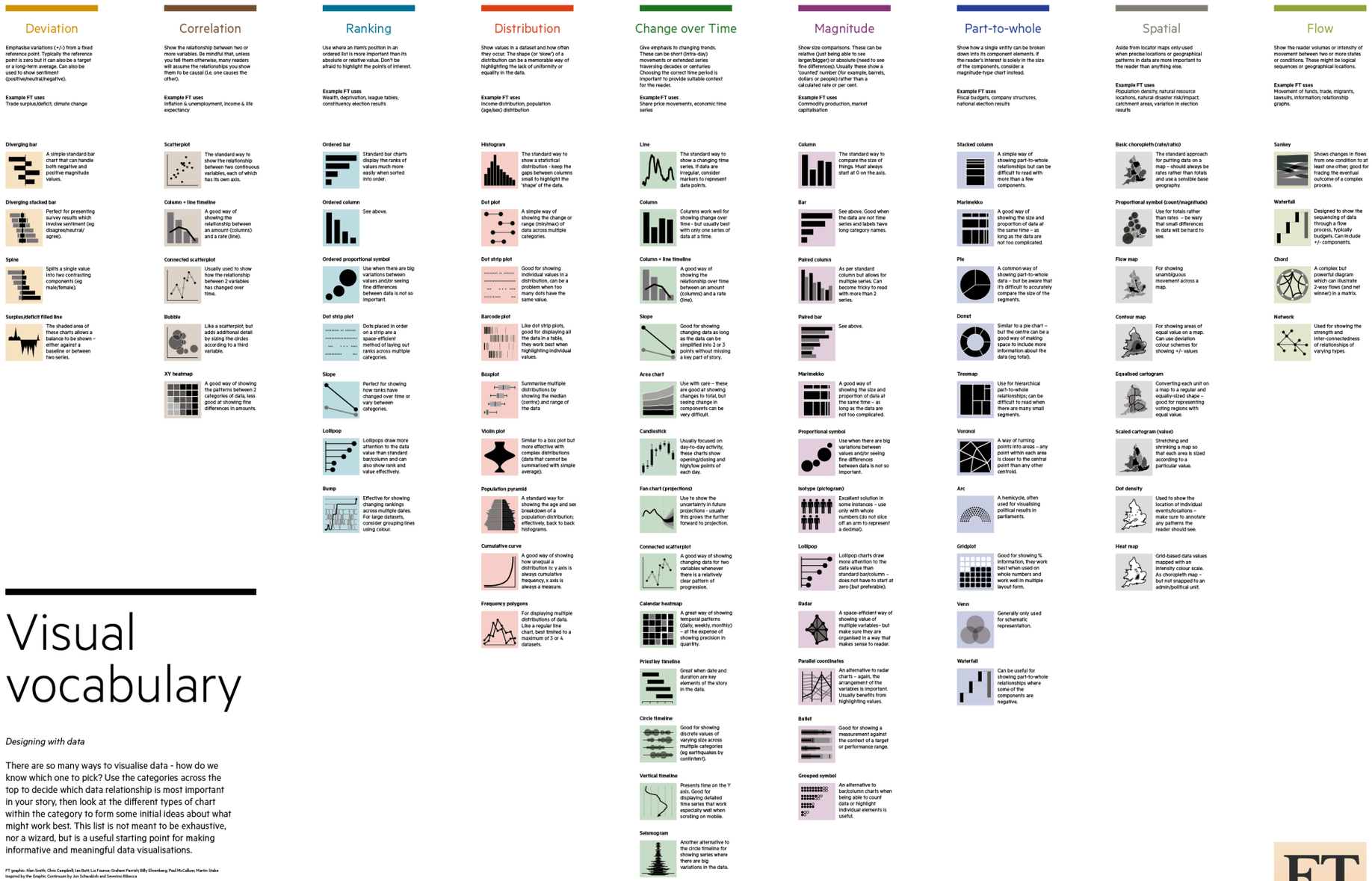
Source: IEA, People's Republic of China energy balance



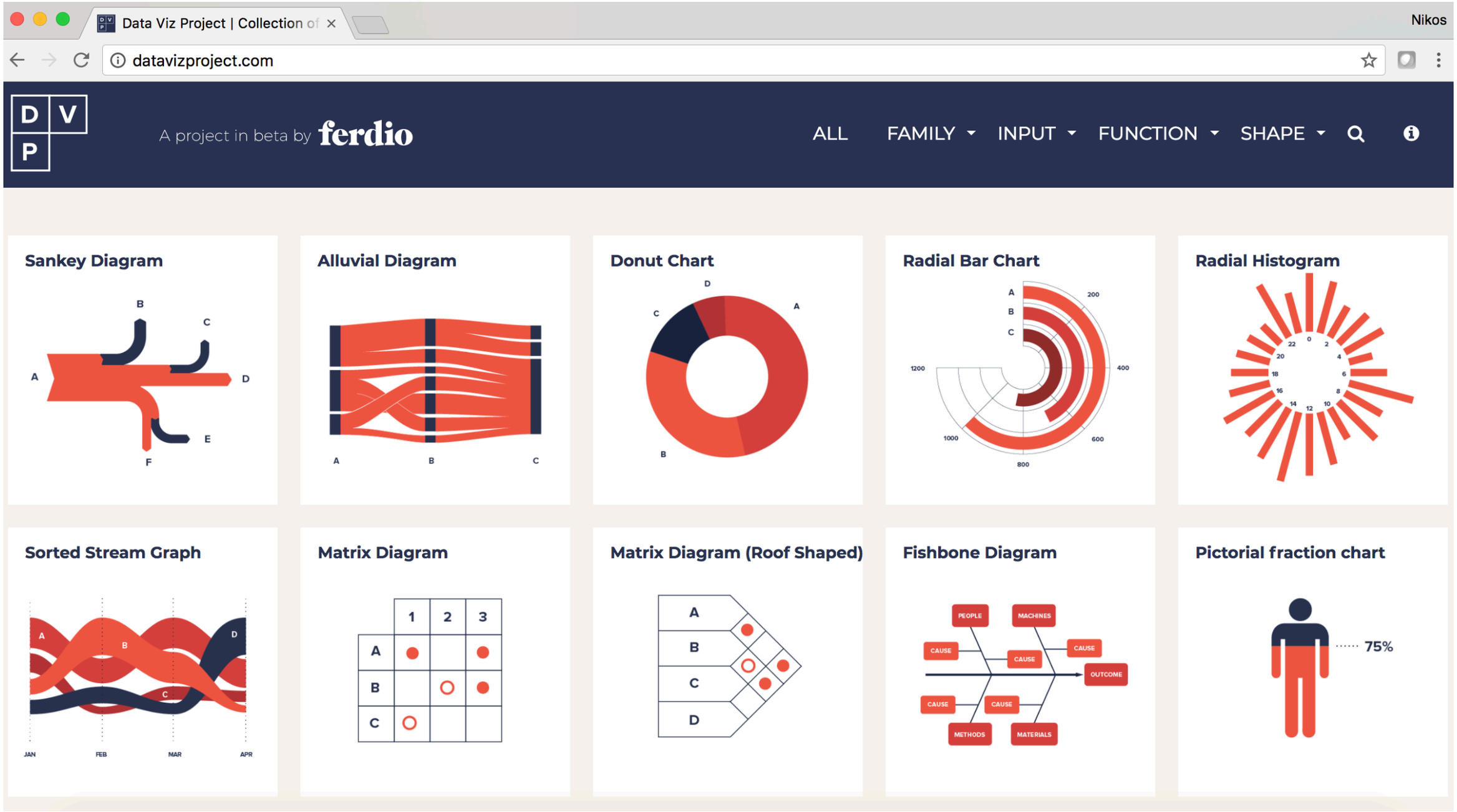
SOURCE: ENALYTICA BASED ON ENERGY INFORMATION ADMINISTRATION. DATA STARTS IN JANUARY 1986, THUS UNDERSTATING THE MAGNITUDE OF THE 1986 DROP

Design options

Choosing a graph: The FT approach



Source: Financial Times, FT-Interactive



Source: http://datavizproject.com/

In summary

Plan	How will the visualization be shared? Is it a repeated exercise or a one-off? Who is the audience? What is the purpose of the visualization? What's the story?
Experiment	How should I present this information? What information should I show and what leave out? Can I transform the data to make it more legible?
Finalize	What is the data-ink ratio? What do I want to draw attention to? What colors/forms/annotations will clarify the story? Is this visualization intellectually honest?

References / Further Reading

Graphs and theory

https://www.edwardtufte.com/tufte/

https://github.com/d3/d3/wiki/gallery

http://vizualize.tumblr.com/

https://www.informationisbeautifulawards.com/

http://mbtaviz.github.io/

http://datavizproject.com/

Journalism

https://www.economist.com/blogs/graphicdetail

https://projects.fivethirtyeight.com/

https://twitter.com/pewresearch/media

https://flowingdata.com/tag/new-york-times/

https://www.nytimes.com/section/upshot

Videos

Hans Rosling, TED talk

Neil Halloran, The Fallen of World War II (YouTube).

For next class, bring:

1. Computer
2. Data, preliminary visualizations