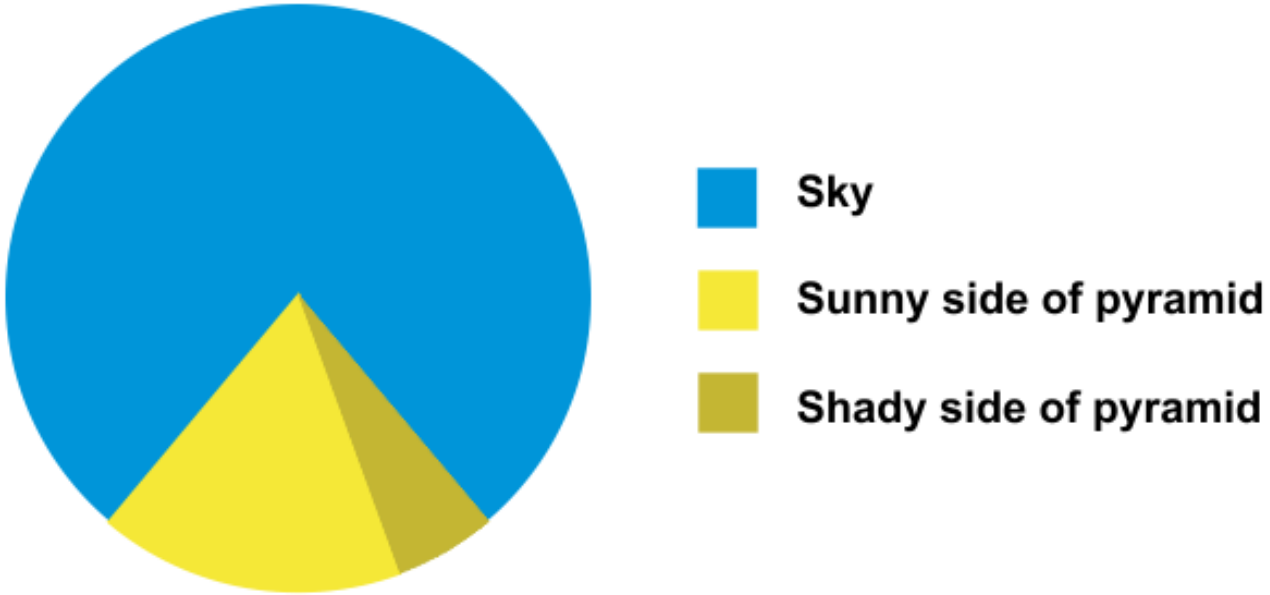
A lightning quick introduction to data visualization

Andrew Heiss, PhD

Brigham Young University

West Coast Nonprofit Data Conference

April 27, 2018



Plan for today

Why visualize data?

Types of visualizations

Aesthetics and design

How do I do all this?

Take a sad plot and make it better

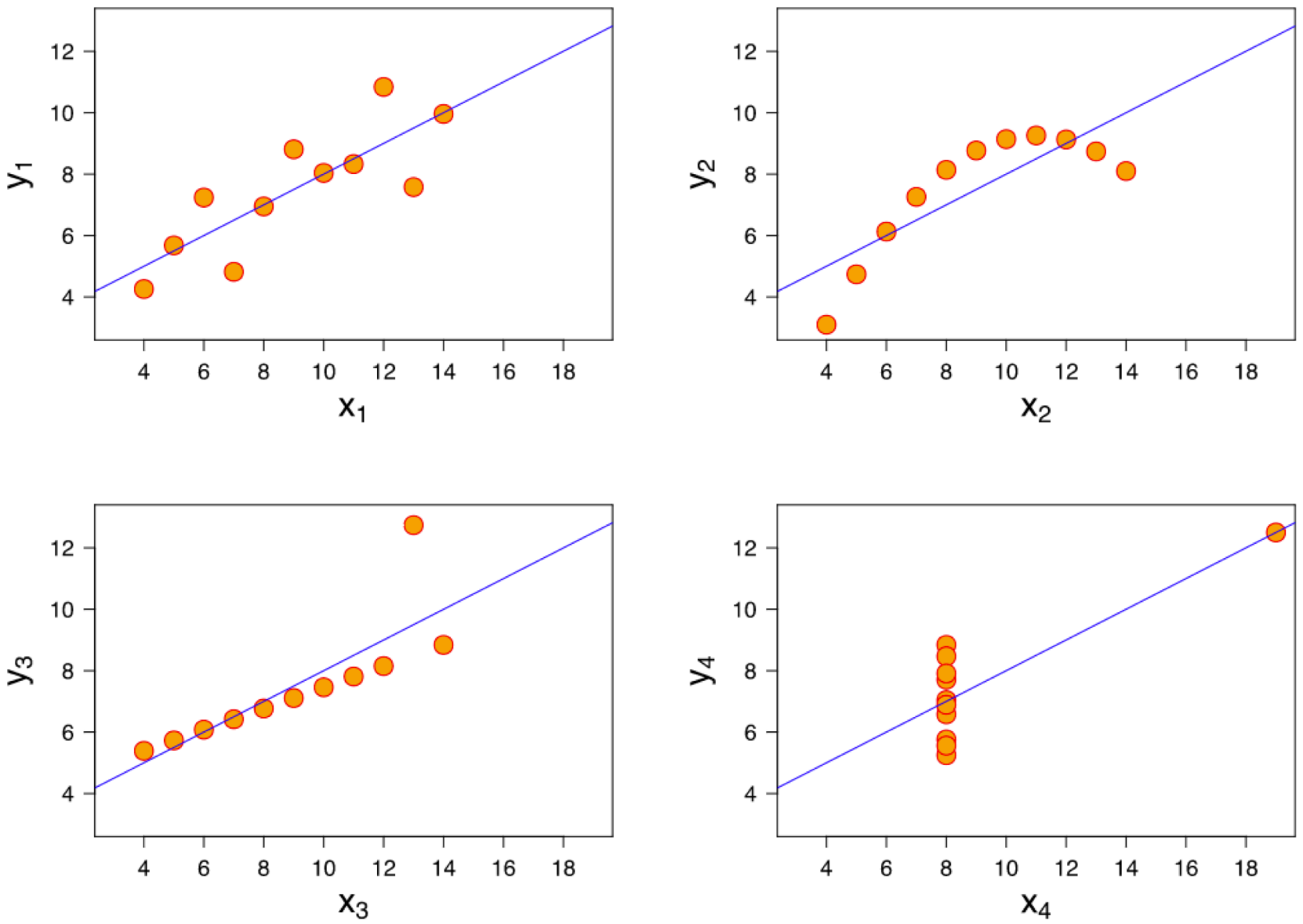
talks.andrewheiss.com/wcnpd18

Why visualize data?

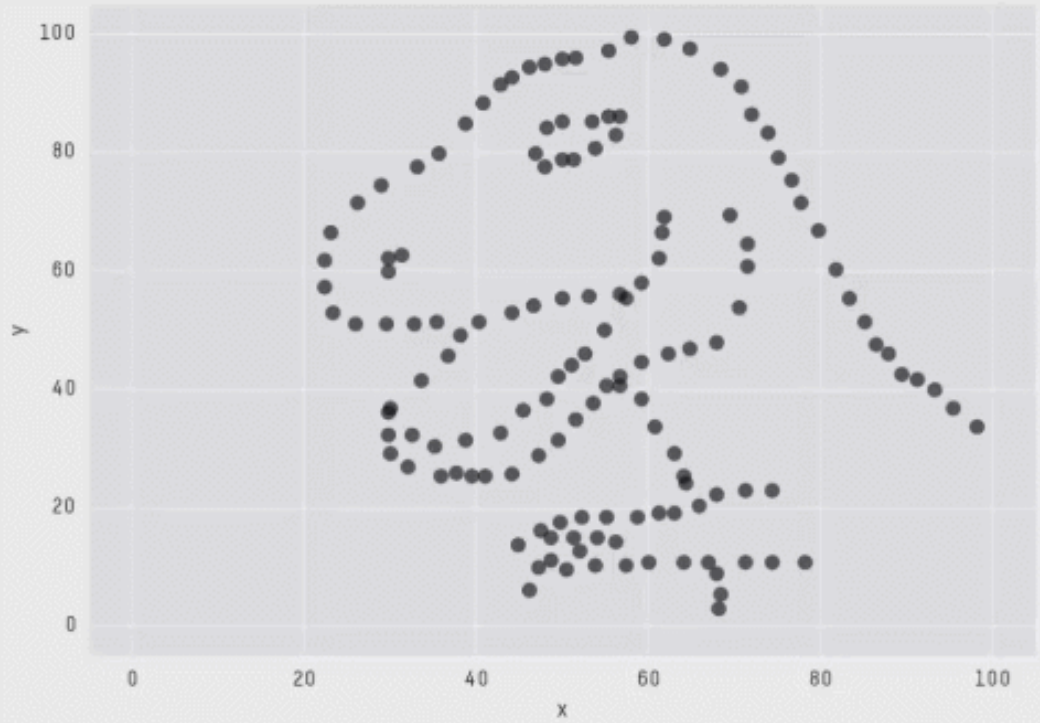
Theories are only stories until you have some data.

Data alone cannot tell stories or prove theories.

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


y1
x1
y2
x2
y3
x3
y4
x4
4
6
8
10
12
14
16
18
4
6
8
10
12

Never trust summary statistics alone



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Humans are visual creatures

frontiers in NEUROSCIENCE

REVIEW ARTICLE
published: 22 August 2014
doi: 10.3389/fnins.2014.00265

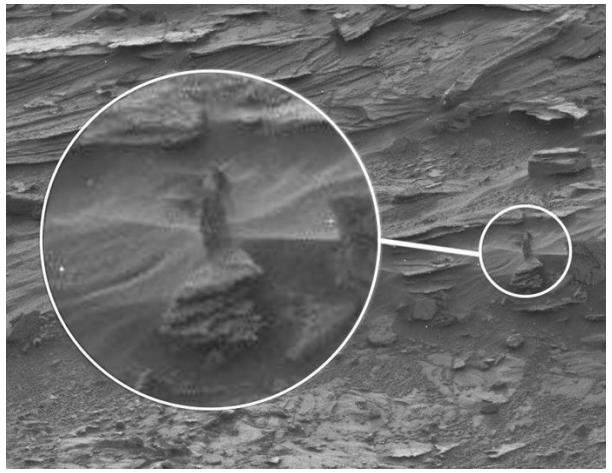
Superior pattern processing is the essence of the evolved human brain

Mark P. Mattson [1,2]*

[1] Laboratory of Neurosciences, National Institute on Aging Intramural Research Program, Baltimore, MD, USA
[2] Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, MD, USA

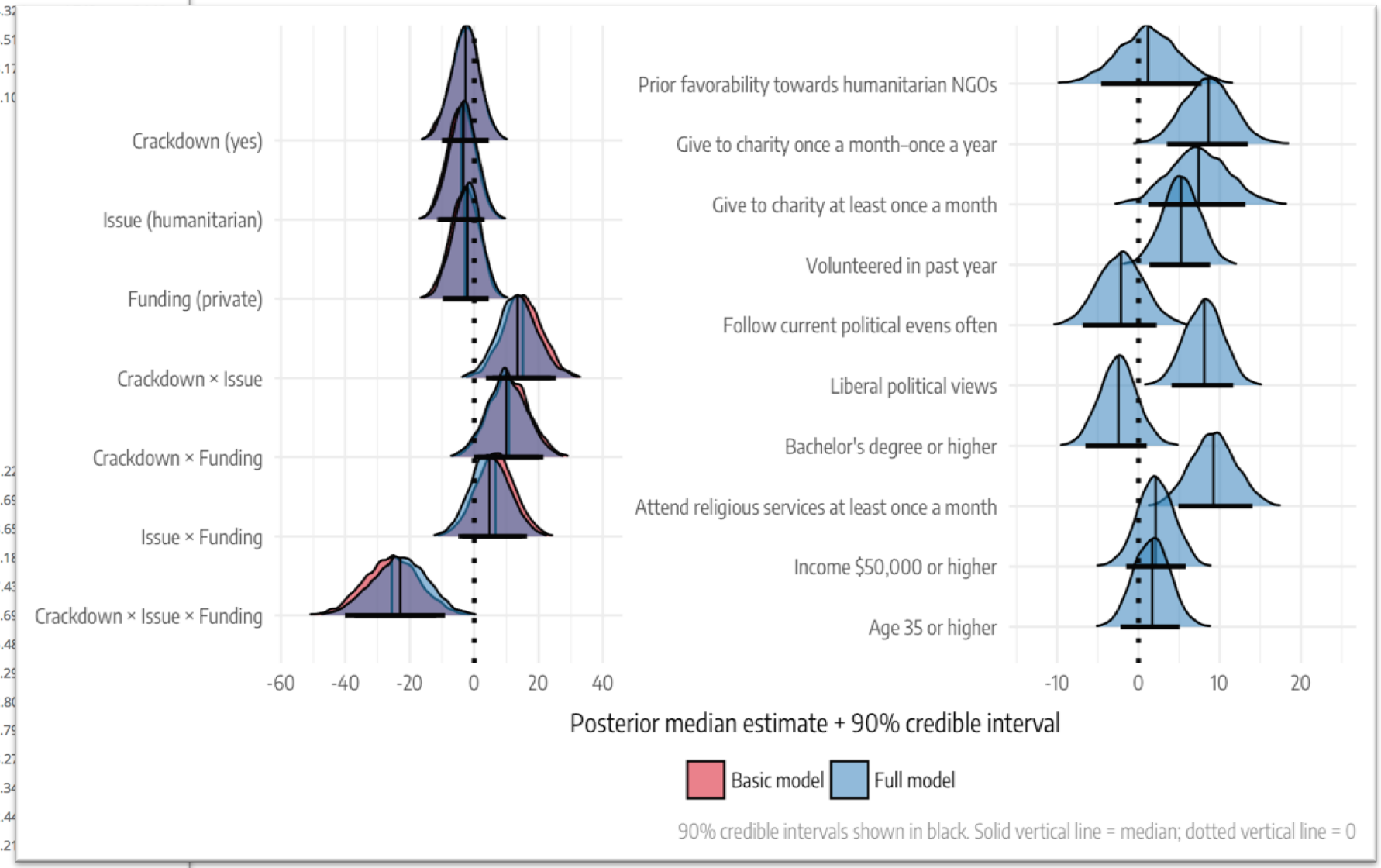
Edited by:
J. Michael Williams, Drexel University, USA

Humans have long pondered the nature of their mind/brain and, particularly why its capacities for reasoning, communication and abstract thought are far superior to other species, including closely related anthropoids. This article considers superior pattern



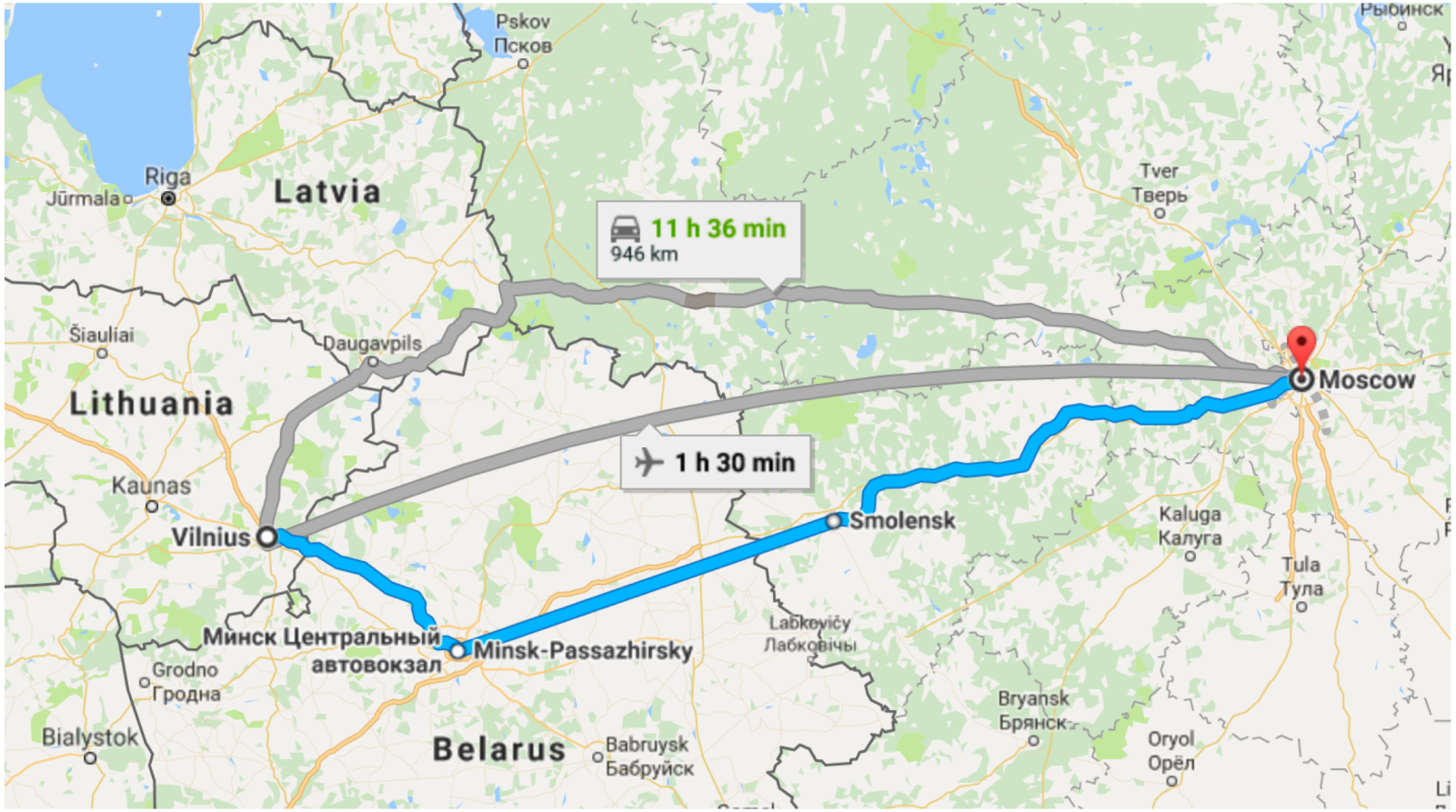
@FacesPics

	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	20.207	20.641	22.033	4.32		
	(1.597)	(2.239)	(3.049)	(4.51		
Crackdown (yes)	3.797	2.636	-2.692	3.17		
	(2.254)	(3.055)	(4.337)	(2.10		
Issue (humanitarian)		-0.993	-4.030			
		(3.094)	(4.312)			
Funding (private)			-2.867			
			(4.444)			
Crackdown × Issue		2.418	15.090			
		(4.366)	(5.969)			
Crackdown × Funding			10.831			
			(6.274)			
Issue × Funding			6.595			
			(6.013)			
Crackdown × Issue × Funding			-25.544			
			(8.628)			
Prior favorability towards humanitarian NGOs				1.22		
				(3.69		
Give to charity once a month–once a year				8.65		
				(3.18		
Give to charity at least once a month				7.43		
				(3.69		
Volunteered in past year				5.48		
				(2.29		
Follow current political evens often				-2.80		
				(2.79		
Liberal political views				8.27		
				(2.34		
Bachelor's degree or higher				-2.44		
				(2.21		
Attend religious services at least once a month				9.817	9.833	9.230
				(2.696)	(2.724)	(2.796)



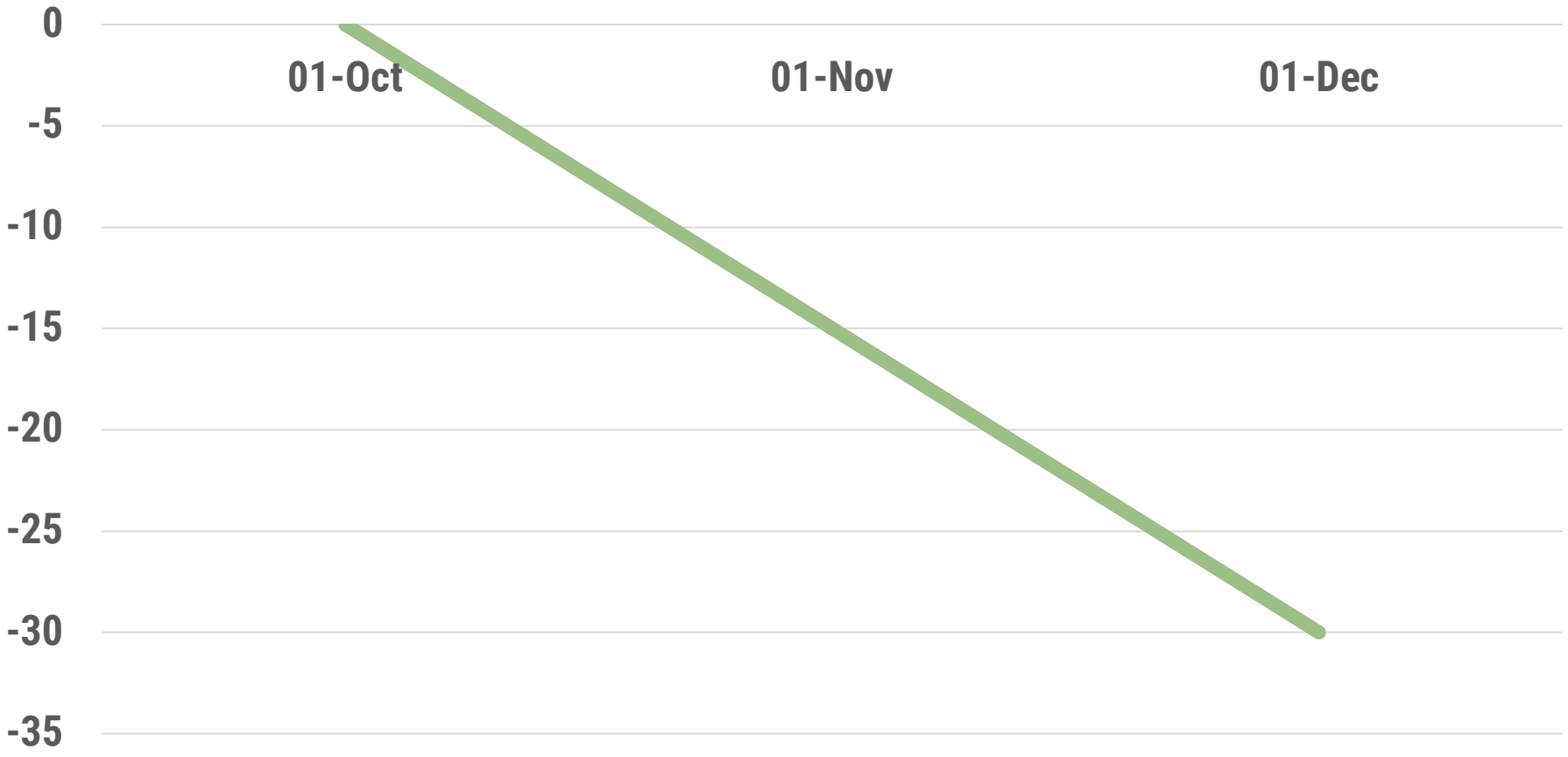
The War of 1812 (the European one)

Long distance!



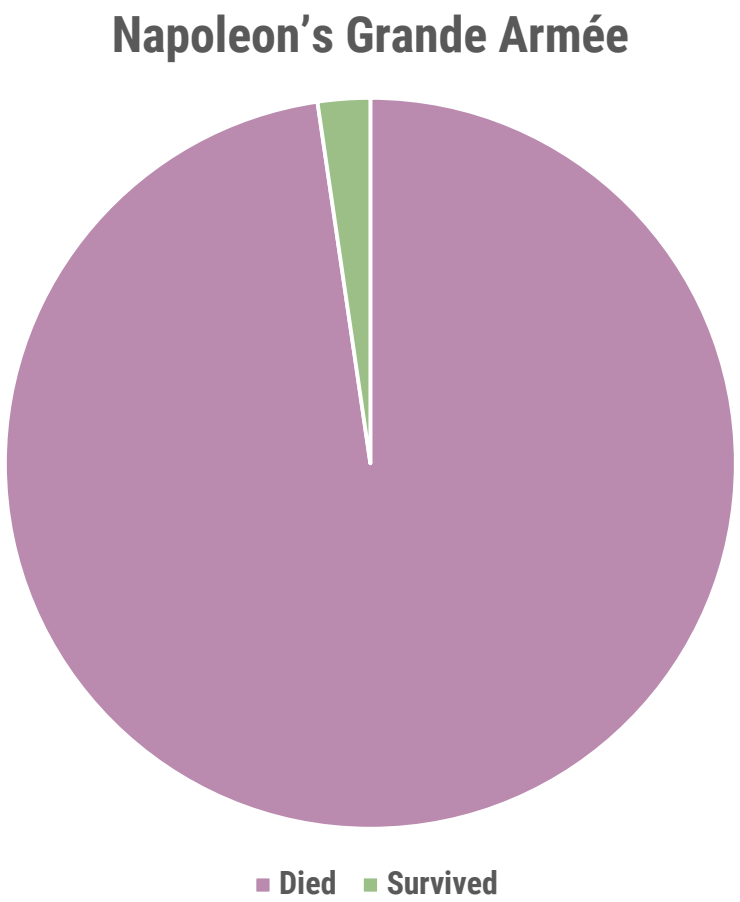
The War of 1812 (the European one)

Very cold!



The War of 1812 (the European one)

Very sad!



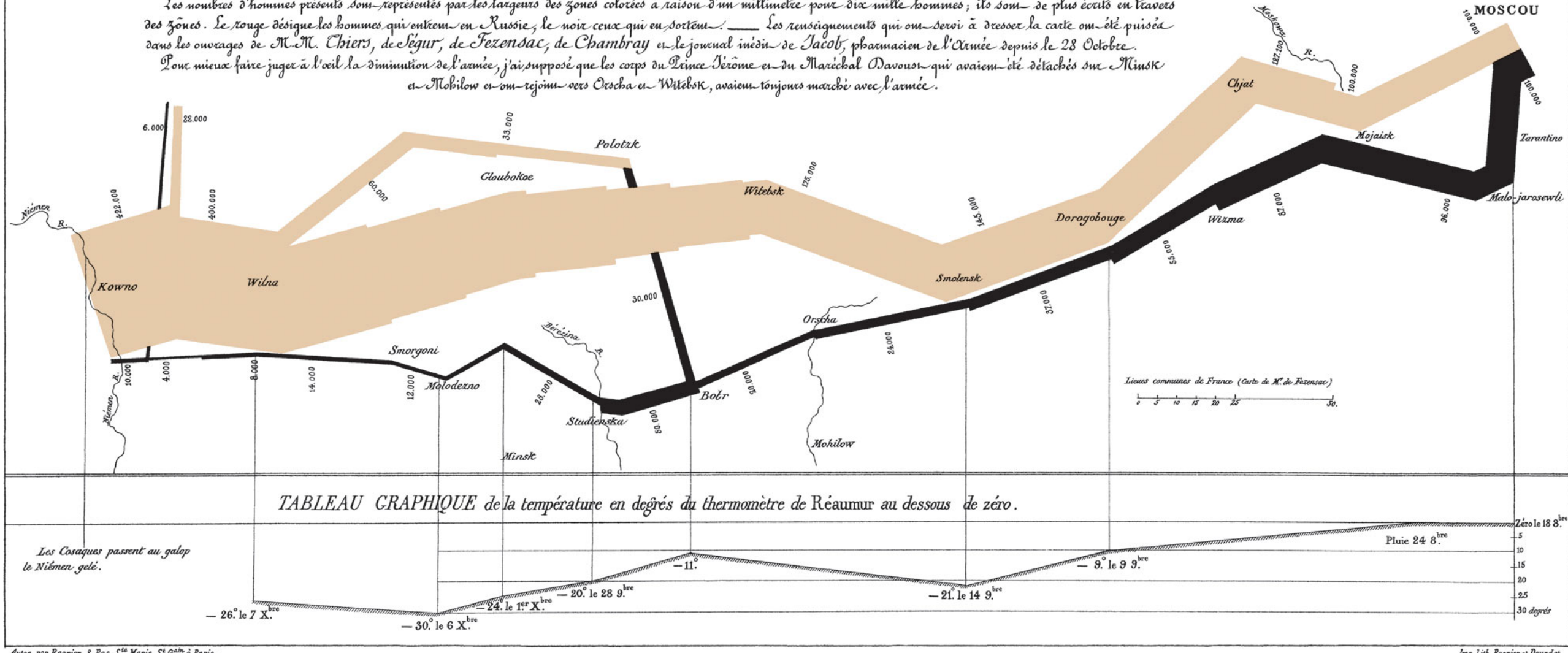
Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. ___ Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.



Autog. par Regnier, 8. Pas. St.te Marie St. Gain. à Paris.

Imp. Lith. Regnier et Dourdet.

Characteristics of graphical excellence

1. “... the well-designed presentation of interesting data—a matter of substance, statistics, and design.”

2. Complex ideas communicated with clarity, precision, and efficiency.

3. That which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.

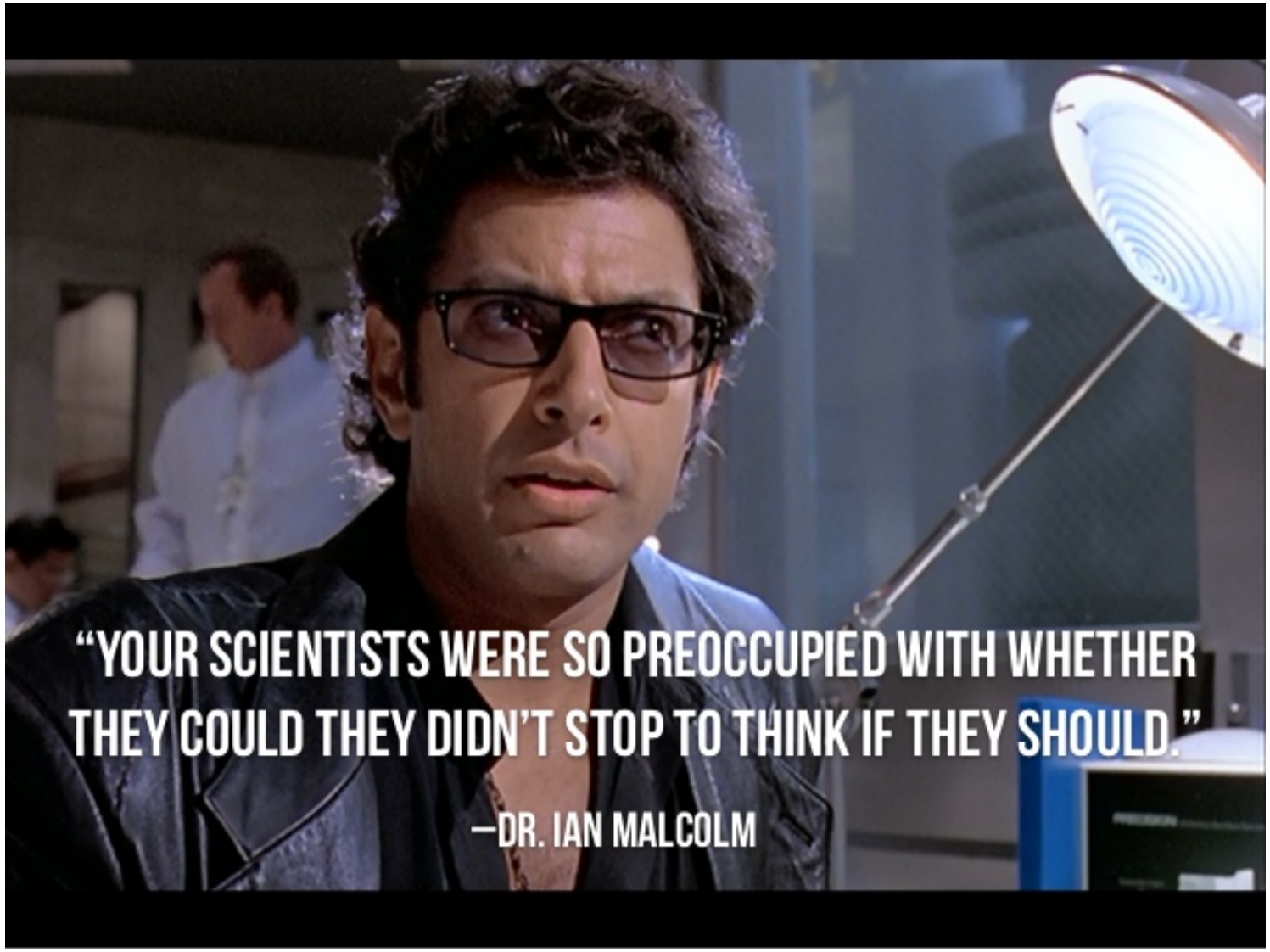
4. Nearly always multivariate.

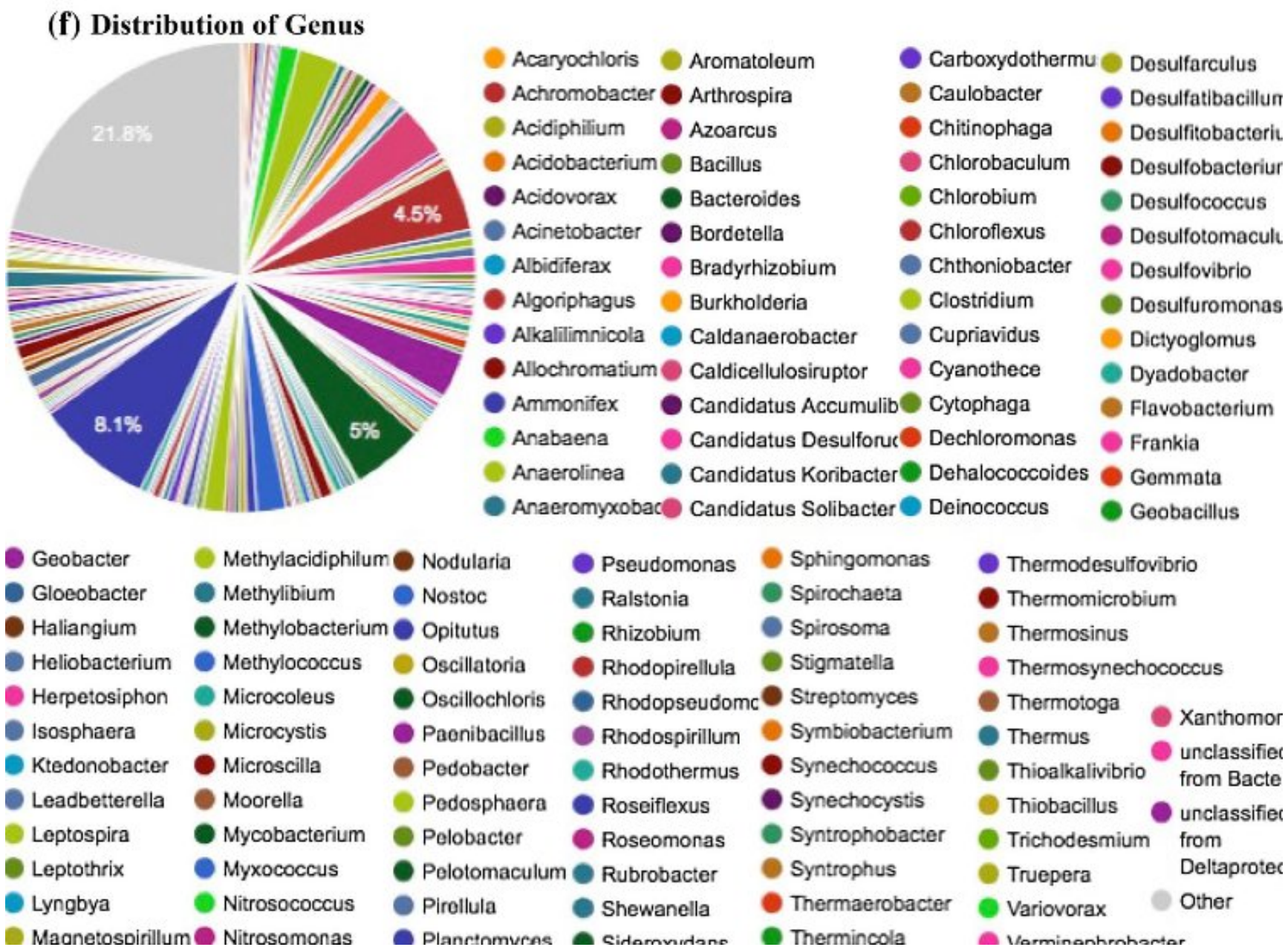
5. Requires telling the truth about the data.

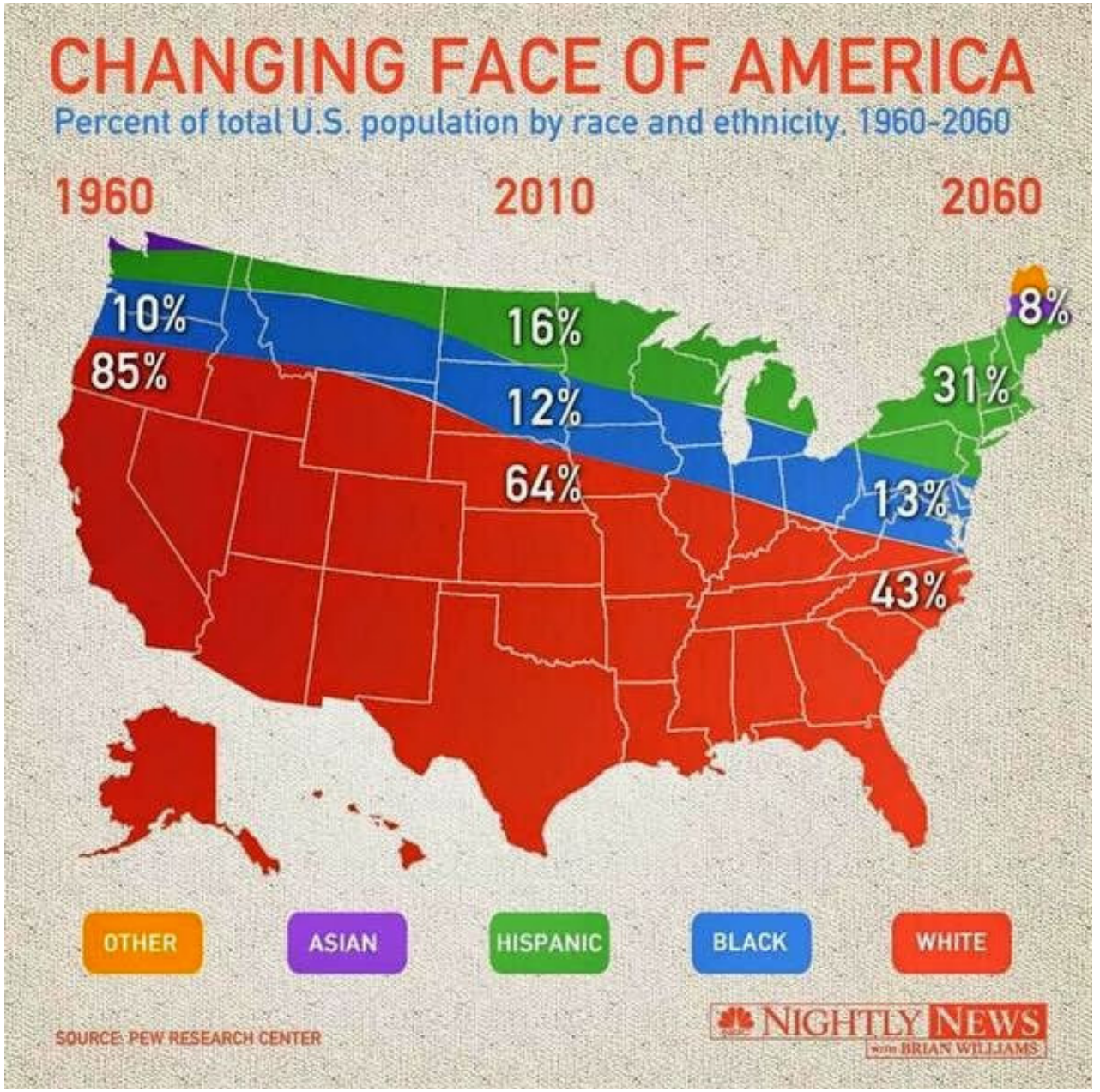
What makes Minard's graph so great?

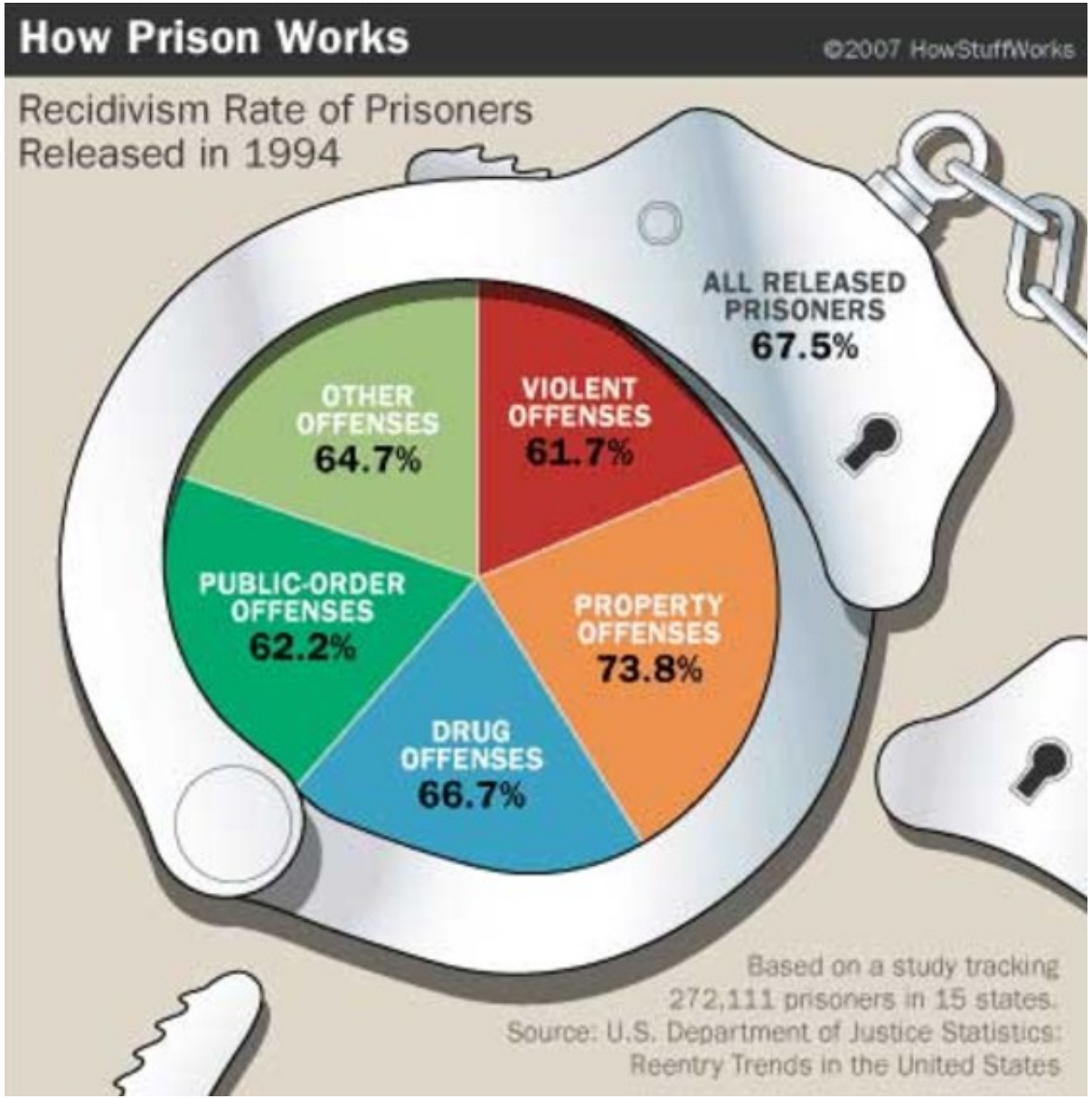
$$\text{Data:Ink ratio} = \frac{\text{Ink}_{\text{Data}}}{\text{Ink}_{\text{Total}}}$$

We forget this!



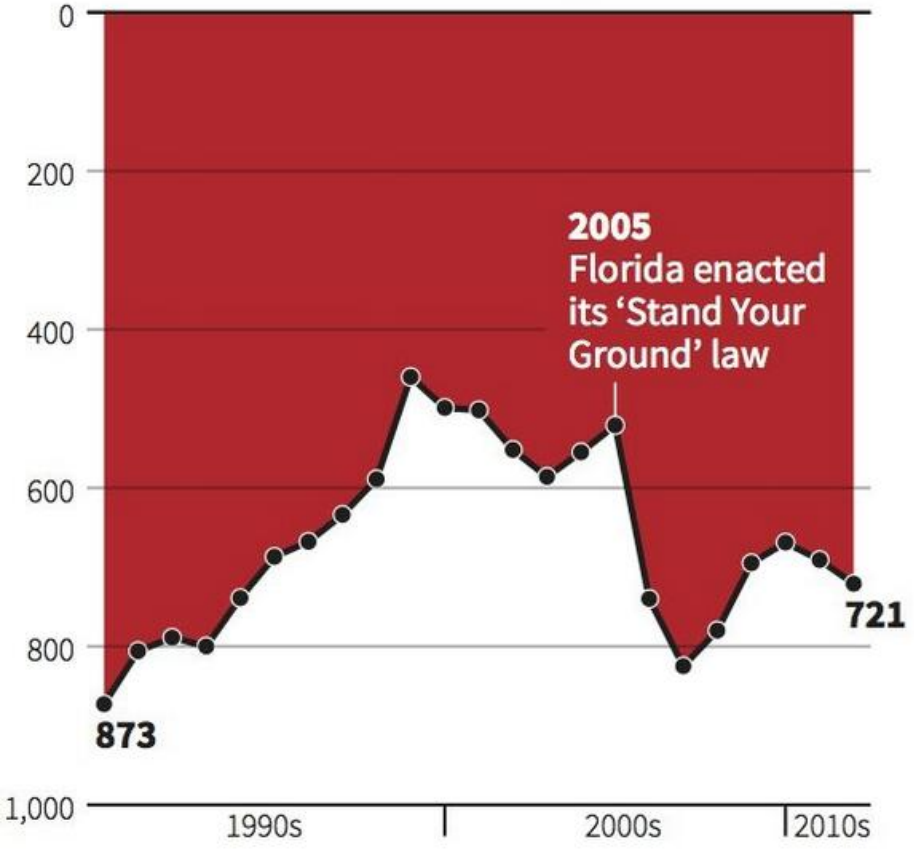

(f) Distribution of Genus
21.8%
4.5%
8.1%
5%
Acaryochloris
Achromobacter
Acidiphilium
Acidobacterium
Acidovorax
Acinetobacter
Albidiferax
Algoriphagus
Alkalilimnicola
Allochromatium
Ammonifex
Anabaena
Anaerolinea
Anaeromyxobac
Aromatoleum
Arthrospira
Azoarcus
Bacillus
Bacteroides
Bordetella
Bradyrhizobium
Burkholderia
Caldanaerobacter
Caldicellulosiruptor
Candidatus Accumulib
Candidatus Desulforuc
Candidatus Koribacter
Candidatus Solibacter
Carboxydothermu
Caulobacter
Chitinophaga
Chlorobaculum
Chlorobium
Chloroflexus
Chthoniobacter
Clostridium
Cupriavidus
Cyanothece
Cytophaga
Dechloromonas
Dehalococcoides
Deinococcus
Desulfarculus
Desulfatibacillun
Desulfitobacteriu
Desulfobacteriur
Desulfococcus
Desulfotomaculu
Desulfovibrio
Desulfuromonas
Dictyoglomus
Dyadobacter
Flavobacterium
Frankia
Gemmata
Geobacillus
Geobacter
Gloeobacter
Haliangium
Heliobacterium
Herpetosiphon
Isosphaera
Ktedonobacter
Leadbetterella
Leptospira
Leptothrix
Lyngbya
Magnetospirillum
Methylacidiphilum
Methylibium
Methylobacterium
Methylococcus
Microcoleus
Microcystis
Microscilla
Moorella
Mycobacterium
Myxococcus
Nitrosococcus
Nitrosomonas
Nodularia
Nostoc
Opitutus
Oscillatoria
Oscillochloris
Paenibacillus
Pedobacter
Pedosphaera
Pelobacter
Pelotomaculum
Pirellula
Planctomyces
Pseudomonas
Ralstonia
Rhizobium
Rhodopirellula
Rhodopseudomc
Rhodospirillum
Rhodothermus
Roseiflexus
Roseomonas
Rubrobacter
Shewanella
Sideroxydans
Sphingomonas
Spirochaeta
Spirosoma
Stigmatella
Streptomyces
Symbiobacterium
Synechococcus
Synechocystis
Syntrophobacter
Syntrophus
Thermaerobacter
Thermincola
Thermodesulfovibrio
Thermomicrobium
Thermosinus
Thermosynechococcus
Thermotoga
Thermus
Thioalkalivibrio
Thiobacillus
Trichodesmium
Truepera
Variovorax
Verminephrobacter
Xanthomor
unclassifiec from Bacte
unclassifiec from Deltaprotec
Other


CHANGING FACE OF AMERICA
Percent of total U.S. population by race and ethnicity, 1960-2060
1960
2010
2060
10%
85%
16%
12%
64%
8%
31%
13%
43%
OTHER
ASIAN
HISPANIC
BLACK
WHITE
SOURCE: PEW RESEARCH CENTER
NIGHTLY NEWS
with BRIAN WILLIAMS


How Prison Works
©2007 HowStuffWorks
Recidivism Rate of Prisoners Released in 1994
ALL RELEASED PRISONERS
67.5%
OTHER OFFENSES
64.7%
VIOLENT OFFENSES
61.7%
PUBLIC-ORDER OFFENSES
62.2%
PROPERTY OFFENSES
73.8%
DRUG OFFENSES
66.7%
Based on a study tracking 272,111 prisoners in 15 states.
Source: U.S. Department of Justice Statistics: Reentry Trends in the United States

Gun deaths in Florida

Number of murders committed using firearms



Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

REUTERS

Types of visualizations

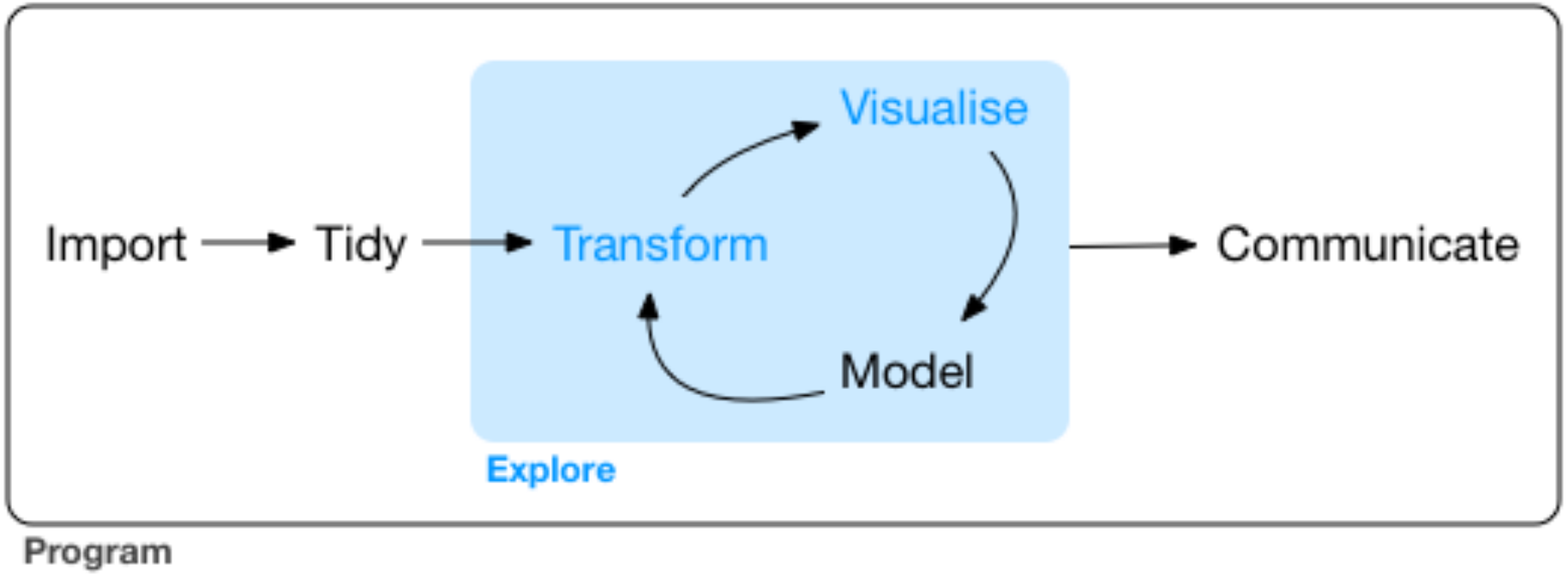
Exploratory visualizations

Academic-ish
Quick scatterplots, histograms, other charts to help understand your data

Explanatory visualizations

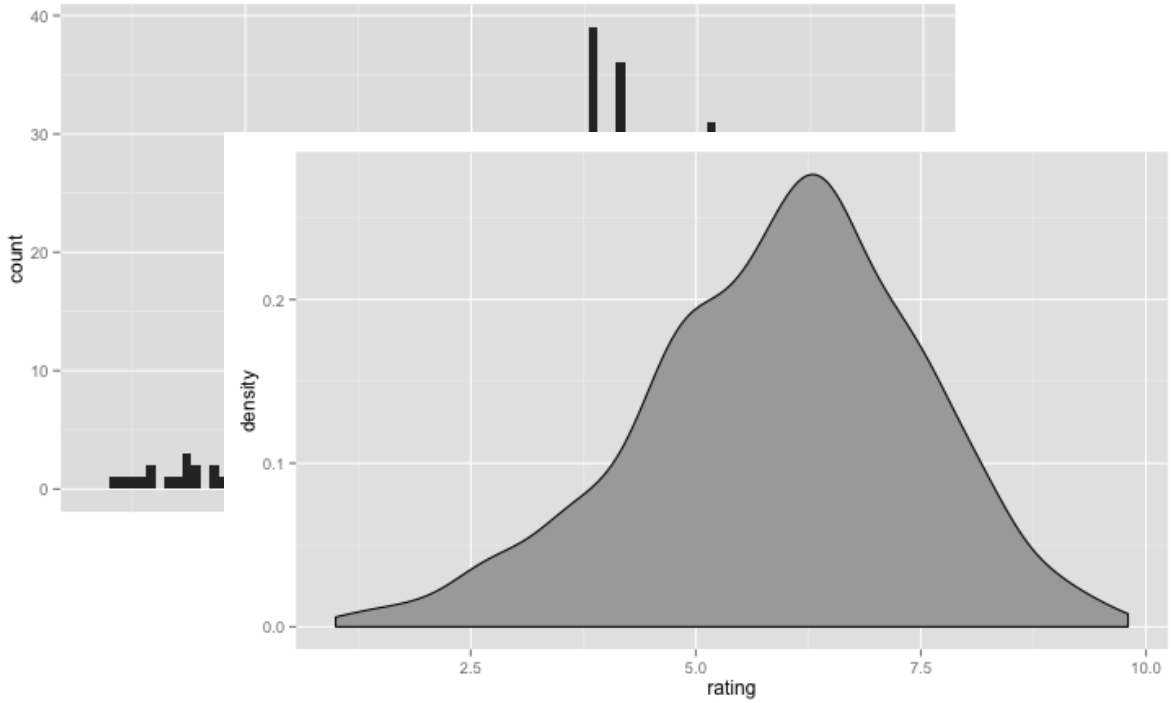
Publishable
Consumable by the general public; Vox, NYT, Washington Post, FiveThirtyEight, etc.

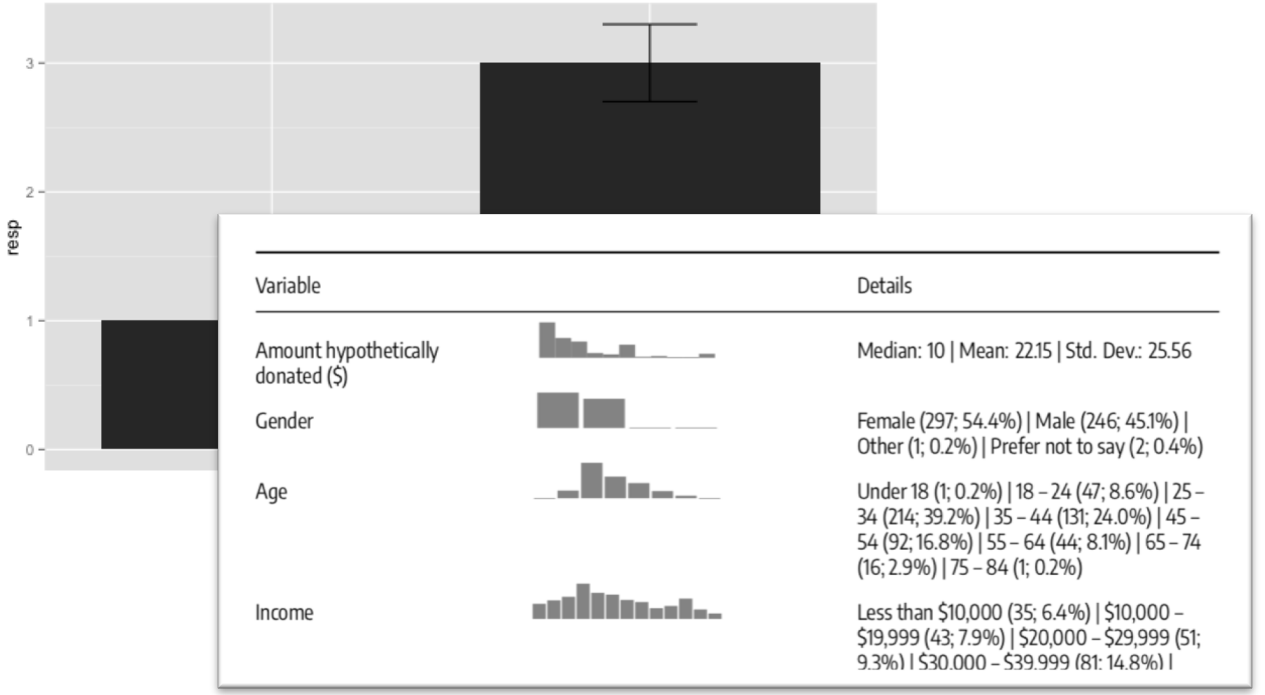
Exploratory **Explanatory**



Exploratory data analysis

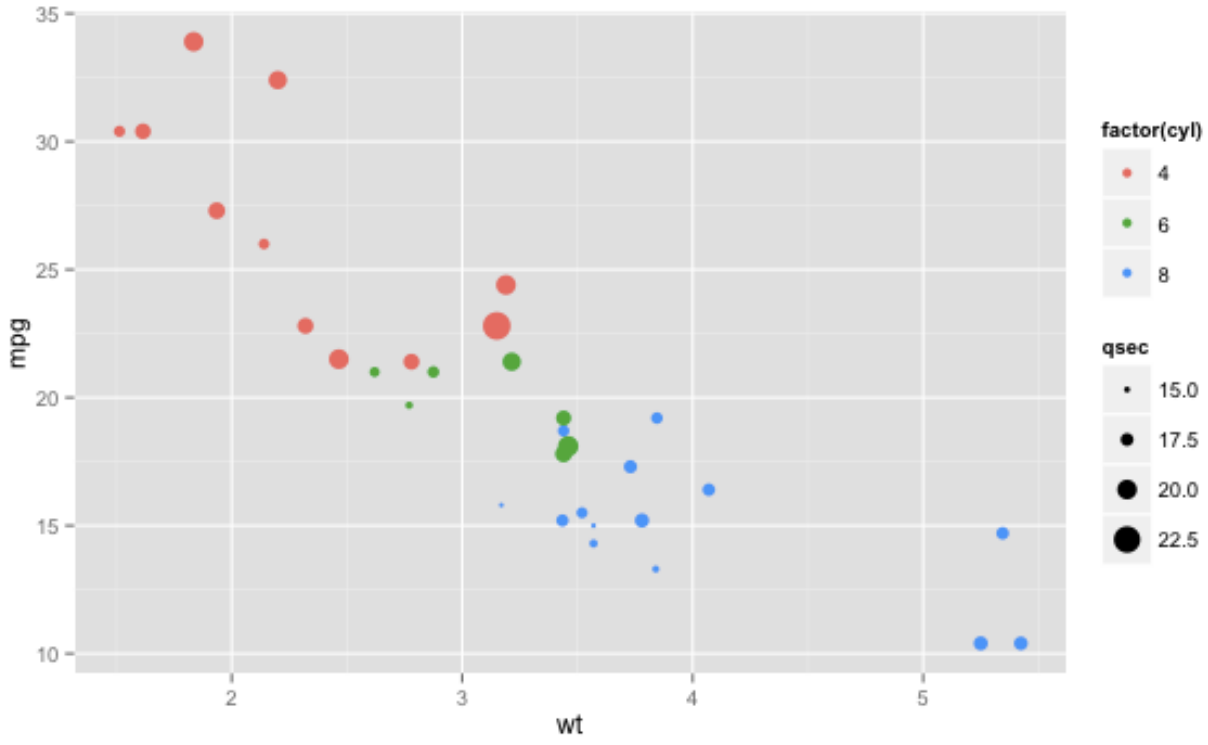
Visualize every variable individually

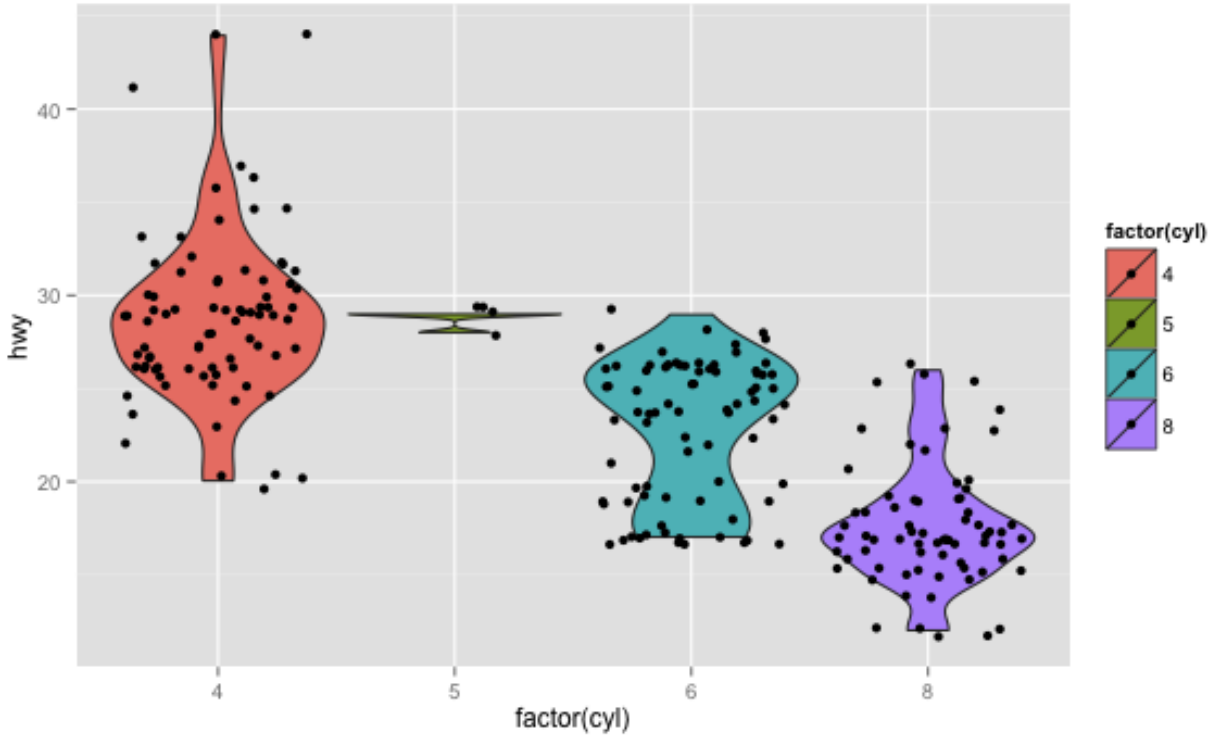




Exploratory data analysis

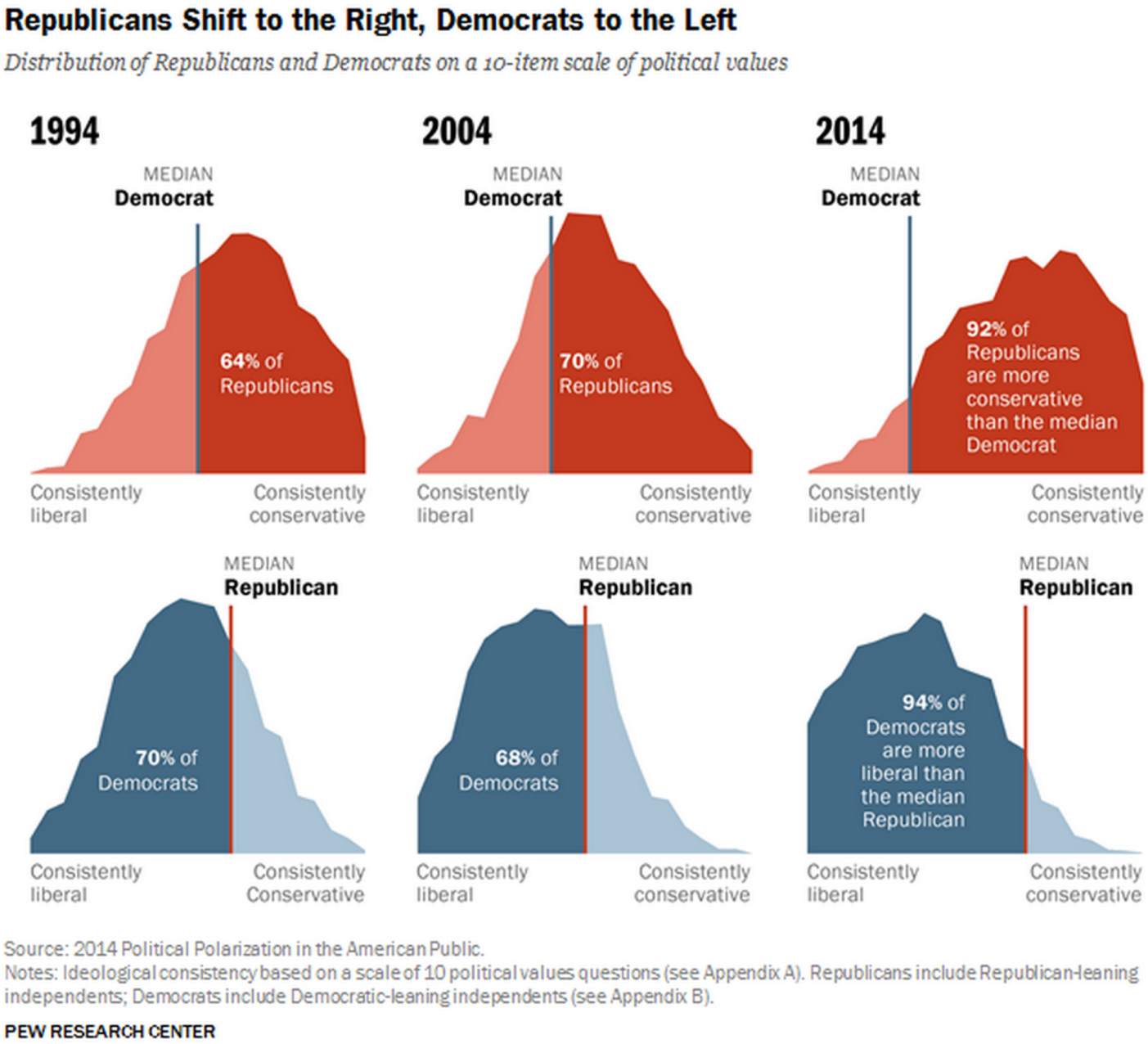
Visualize relationships between each variable





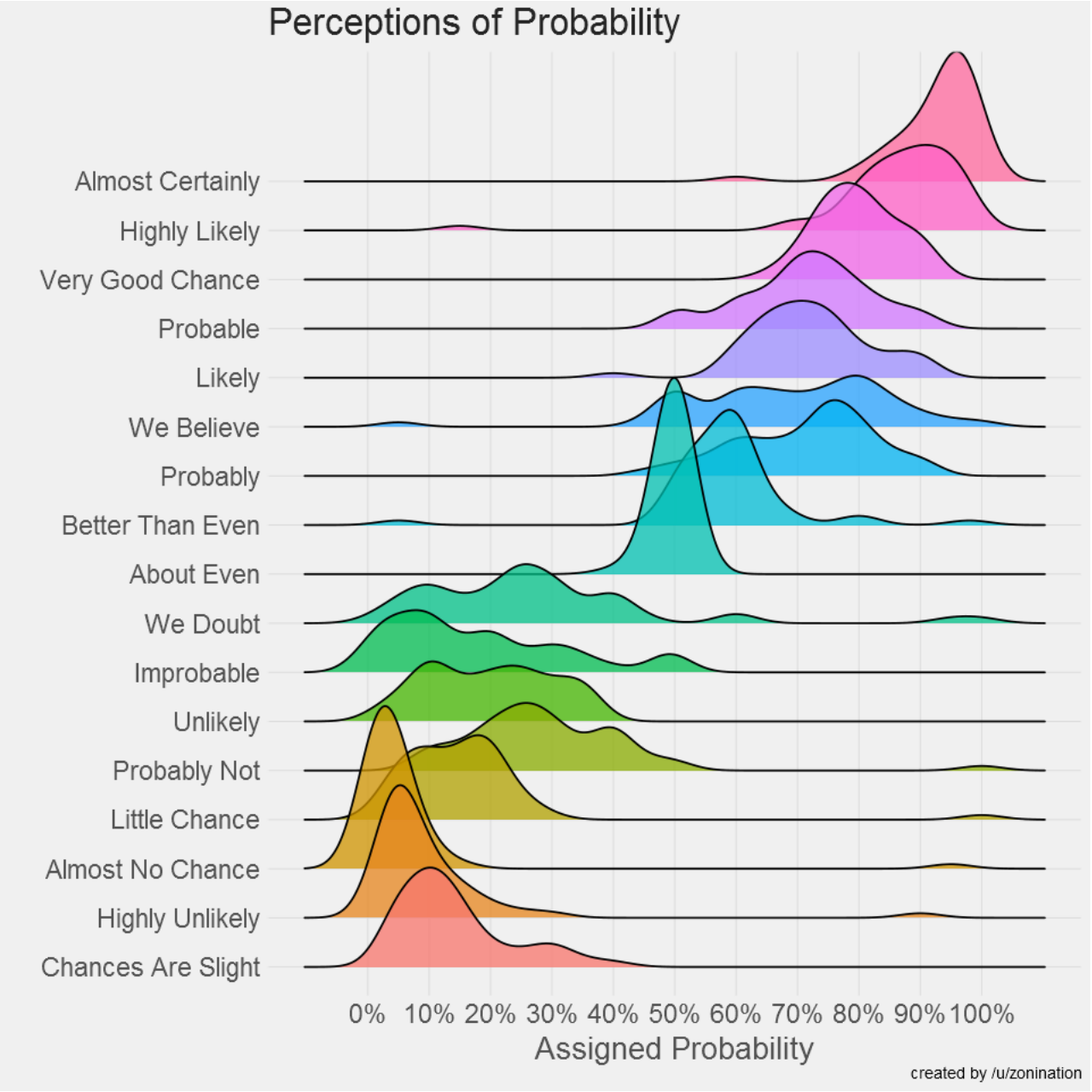
Explanatory data analysis

Annotate and tell a story



Explanatory data analysis

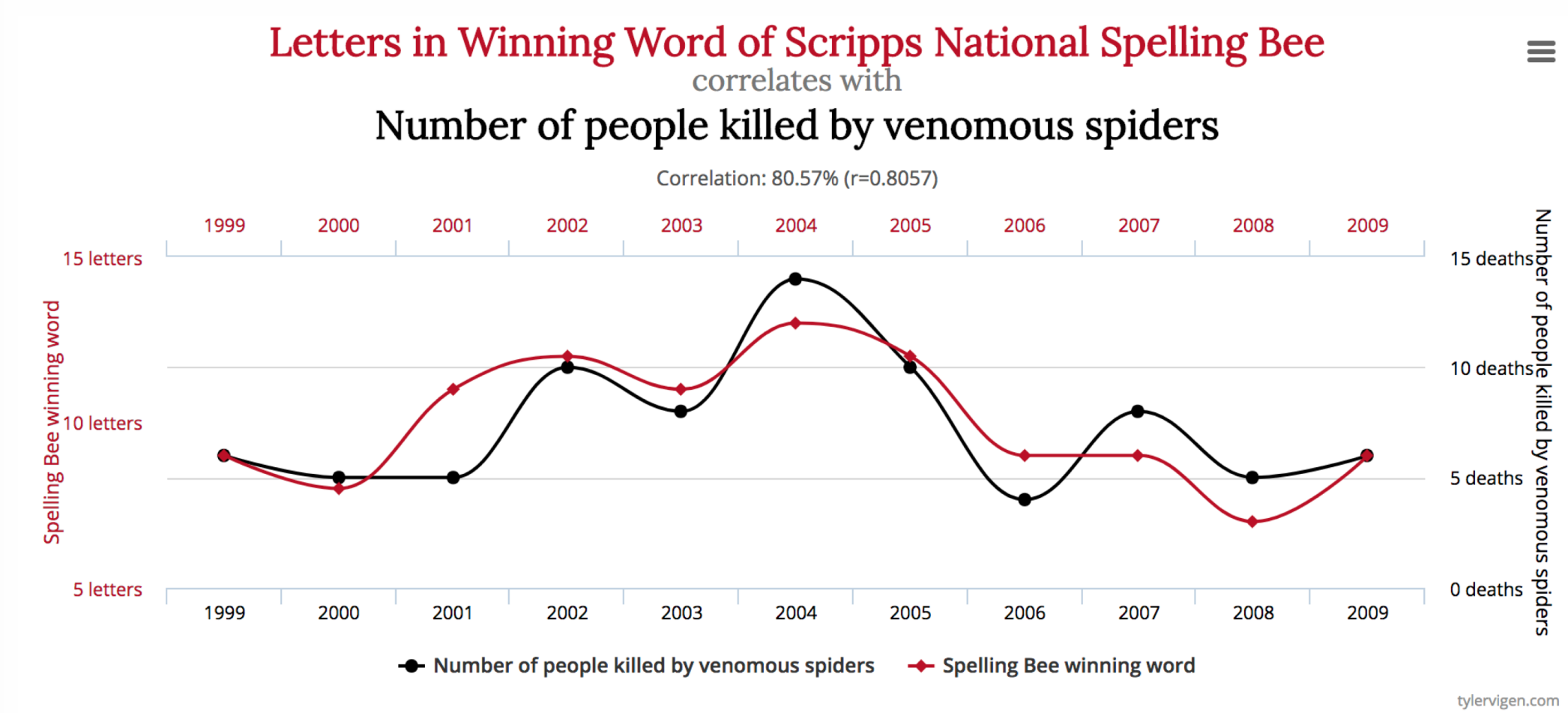
Annotate and tell a story



Which chart type do I use?

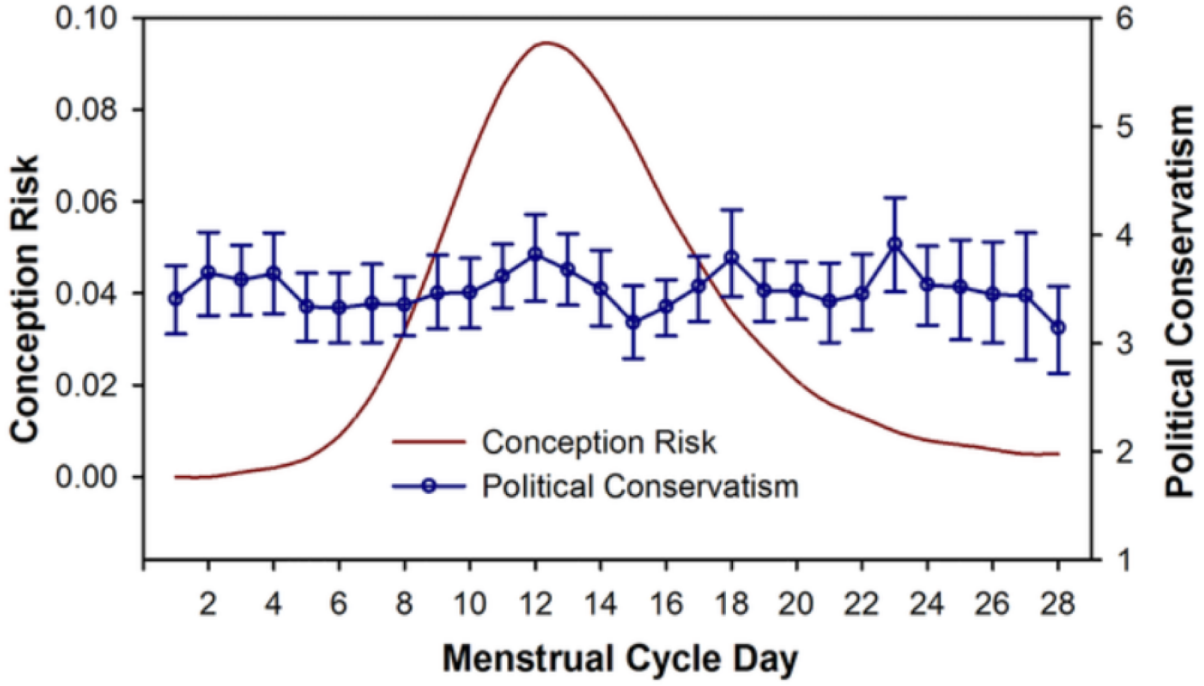
datavizcatalogue.com

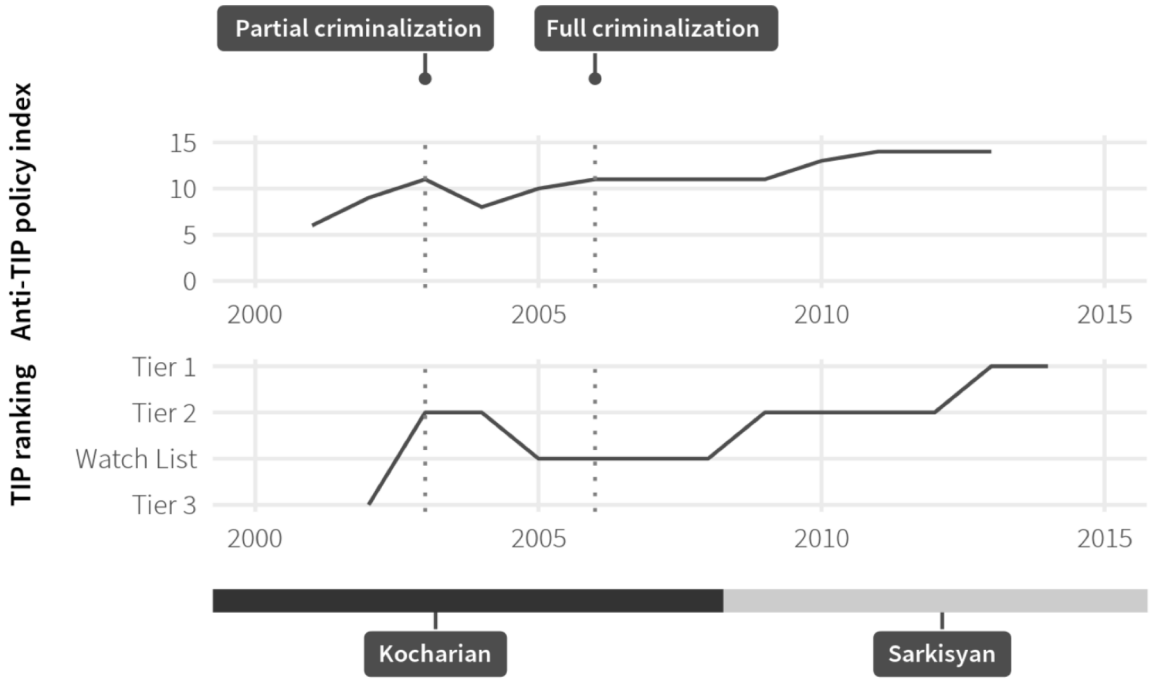
Dataviz gotchas: multiple y-axes



Data sources: National Spelling Bee and Centers for Disease Control & Prevention

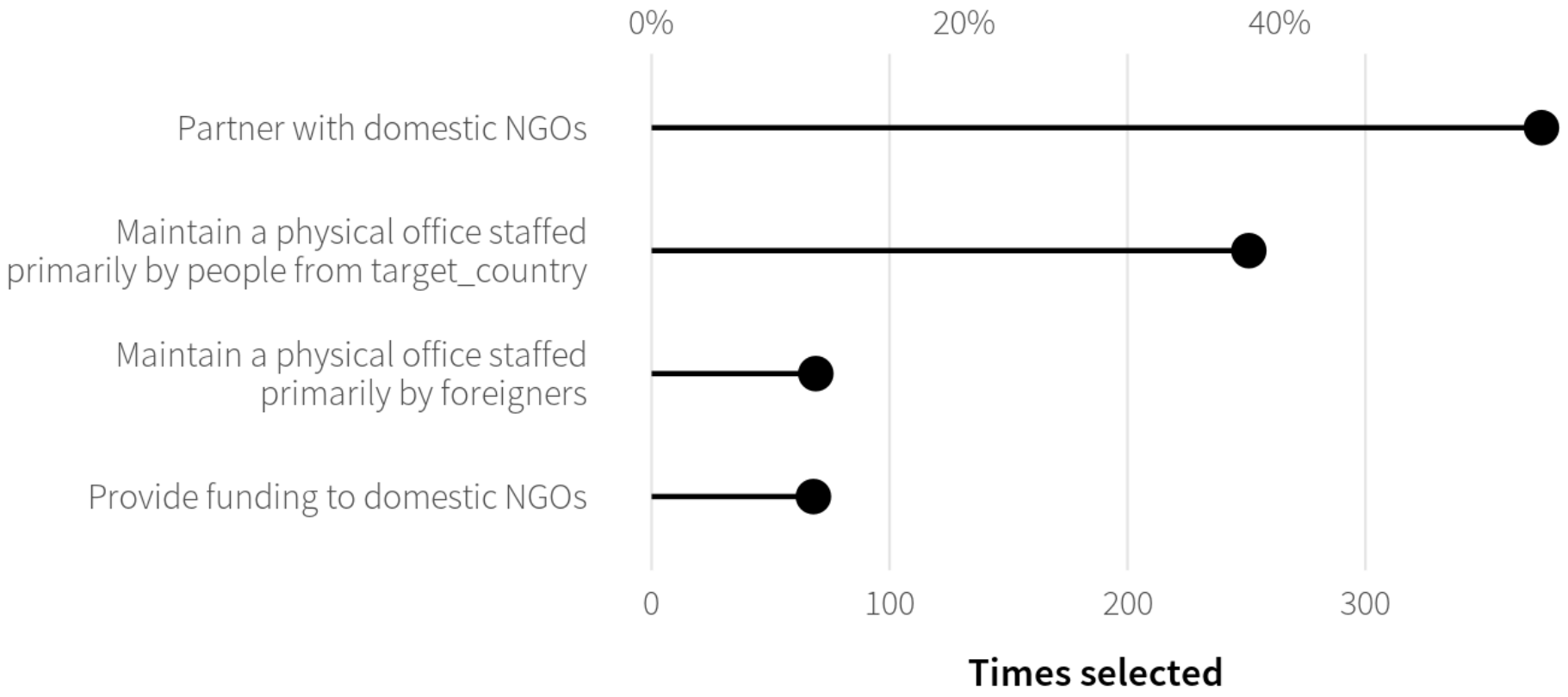
Dataviz gotchas: multiple y-axes



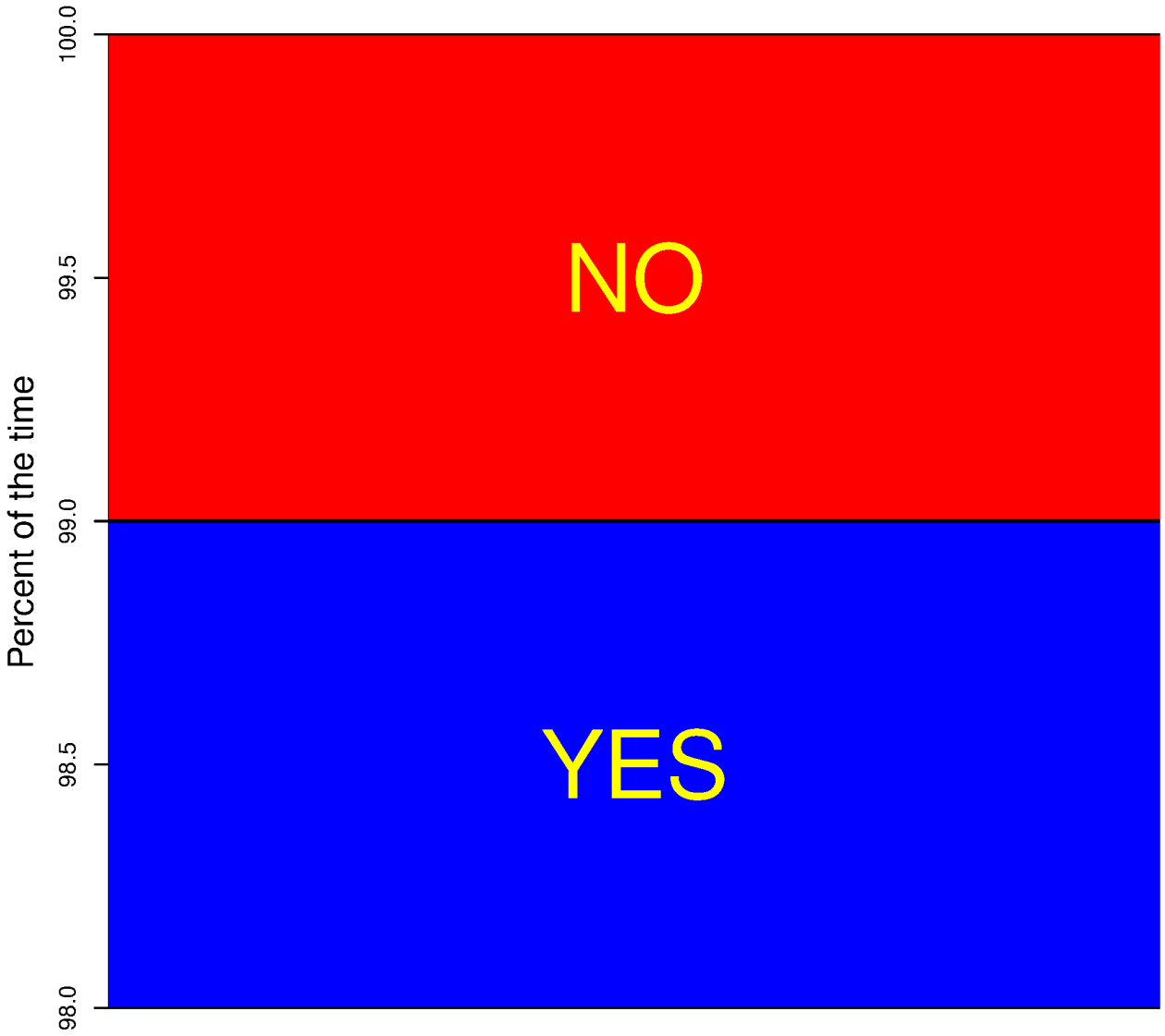


Dataviz gotchas: multiple y-axes

It’s okay if both axes measure the same thing



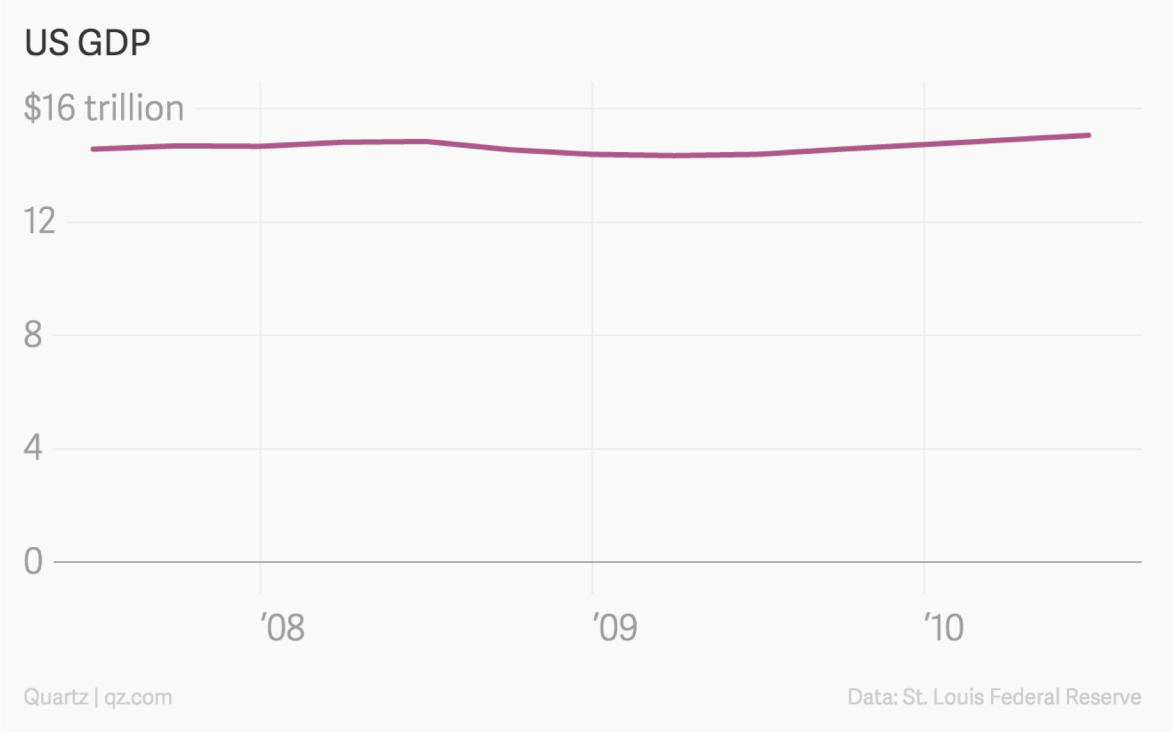
Is truncating the Y–axis misleading?

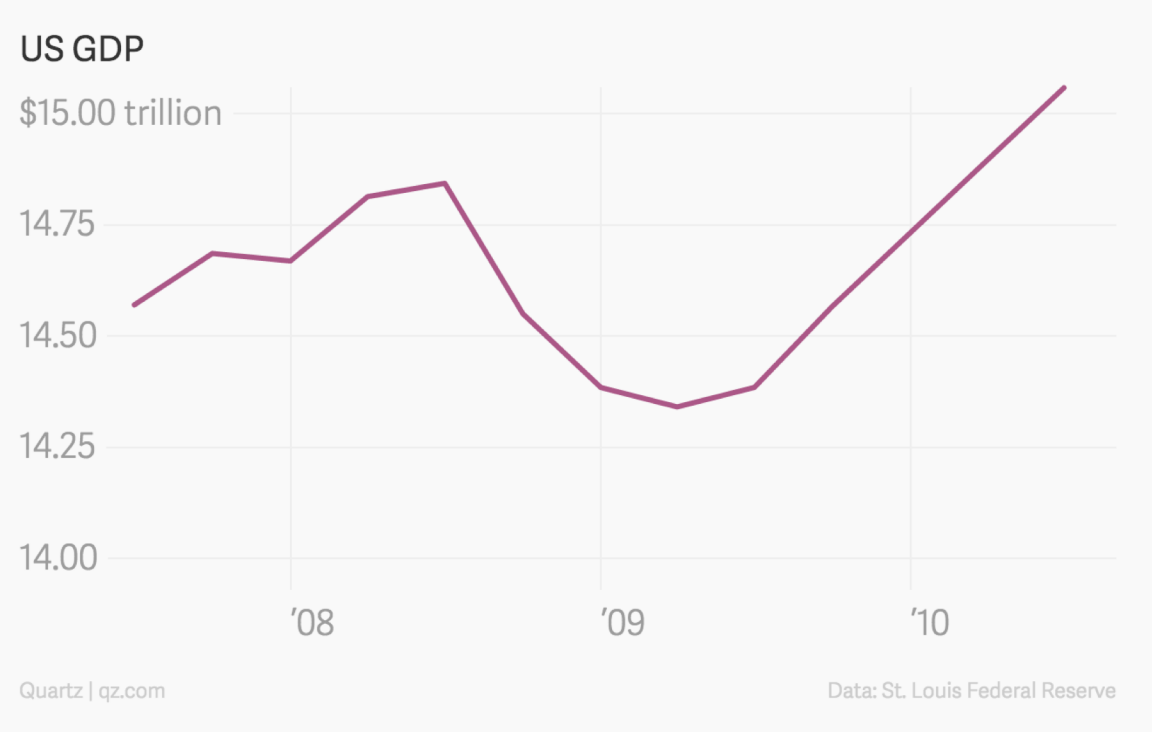


Dataviz gotchas: truncated y-axes

Truncation is okay sometimes!

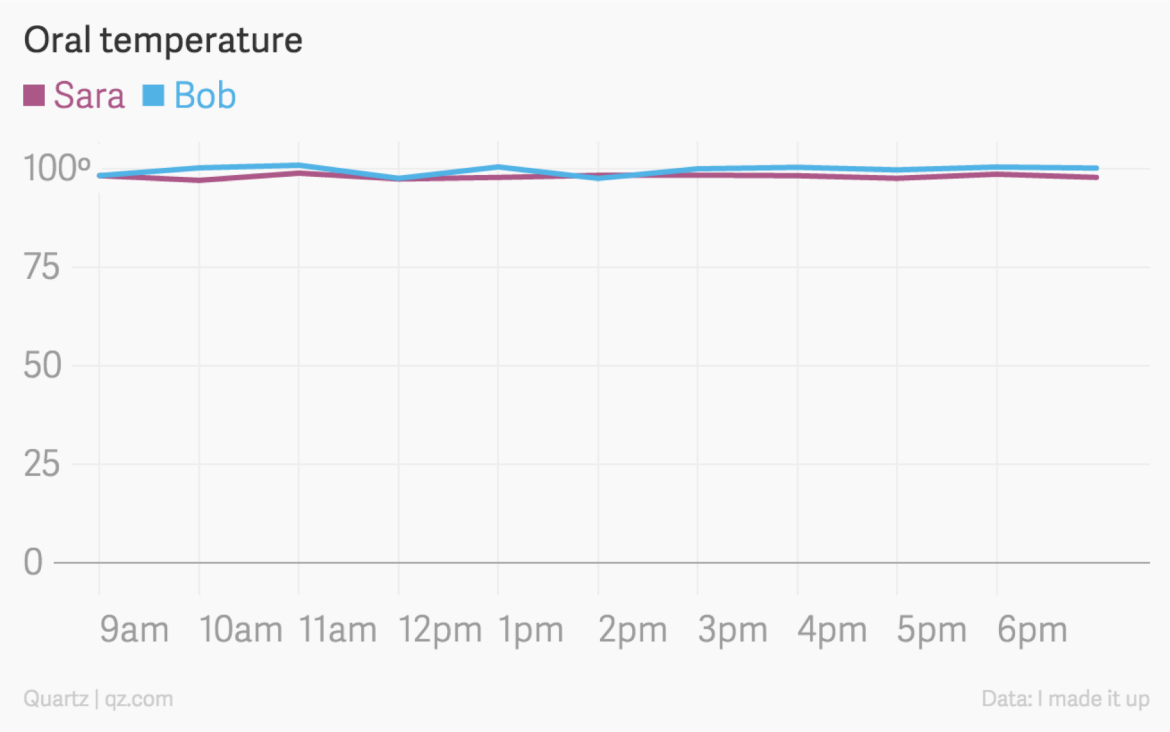
…when small movements matter

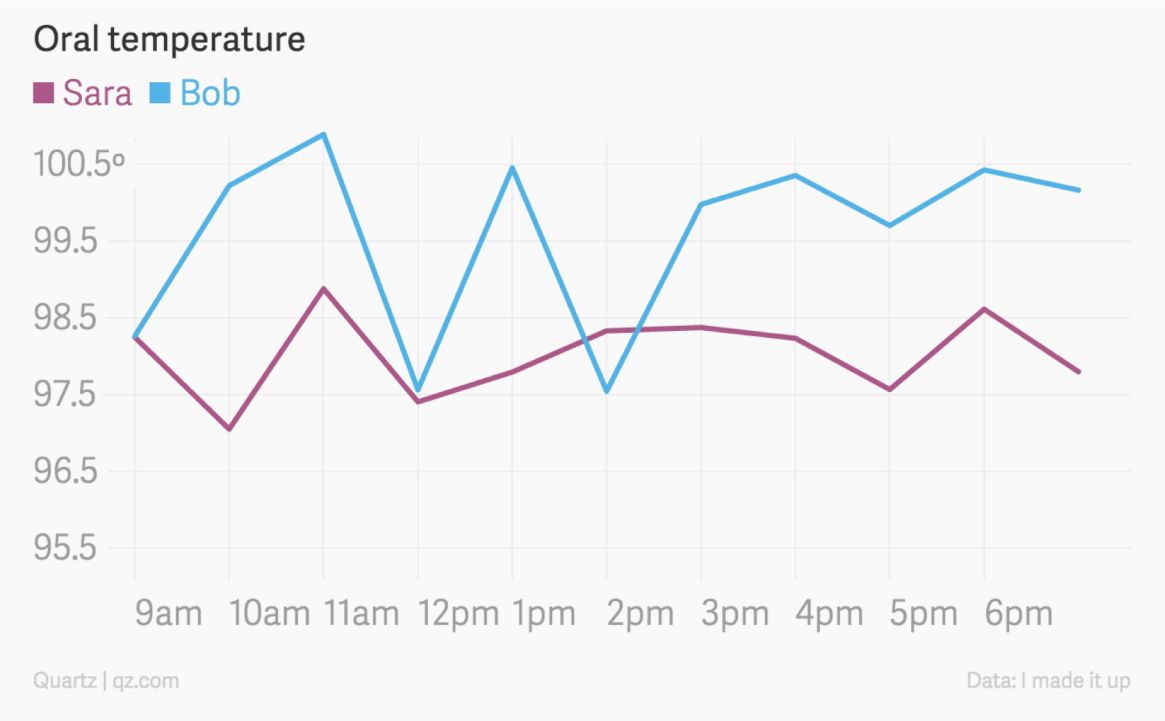




Truncation is okay sometimes!

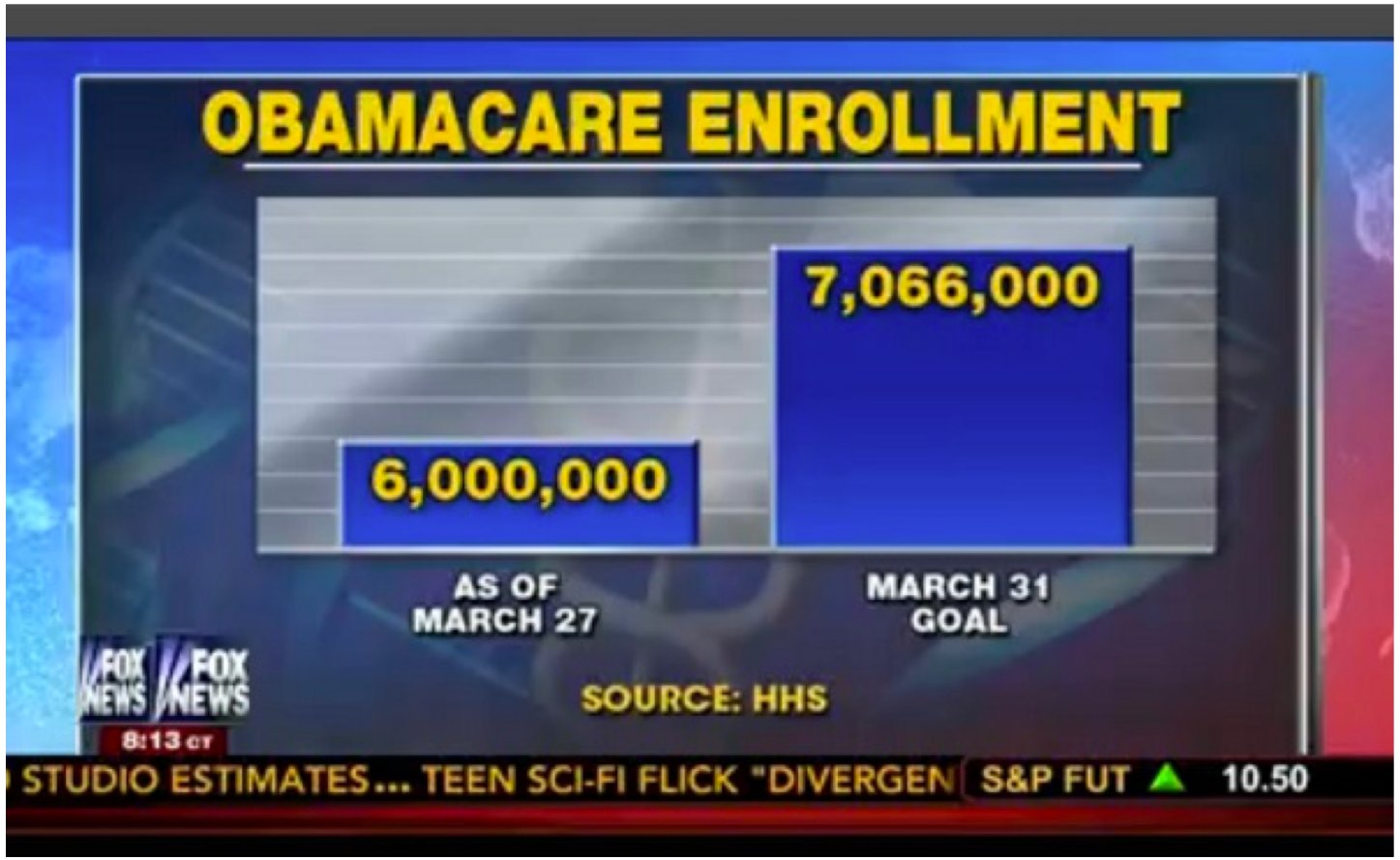
…when zero values are impossible

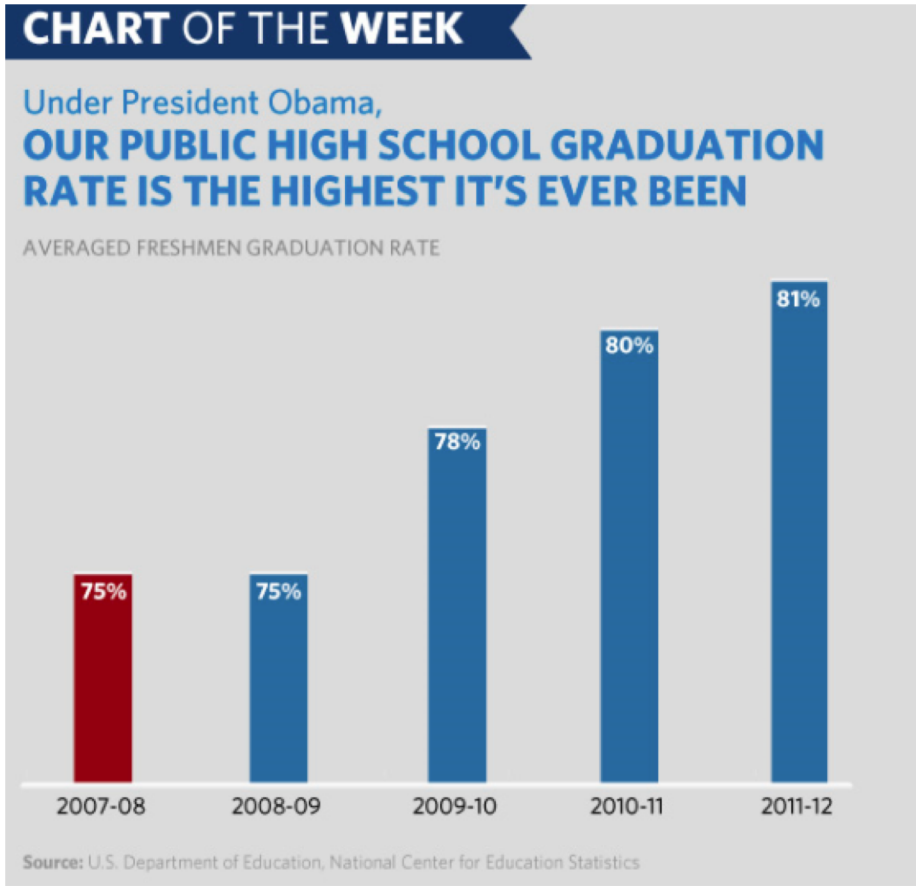




Truncation is okay sometimes!

…but **never** on bar charts!



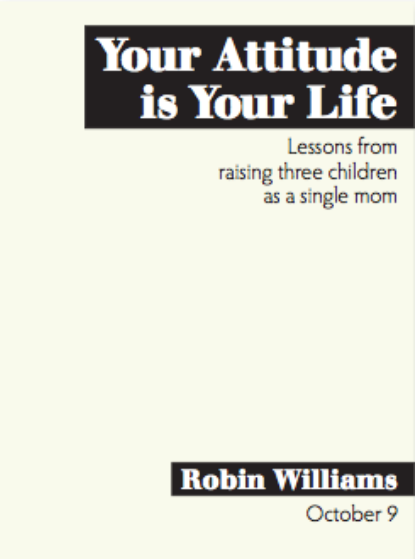


Aesthetics and design

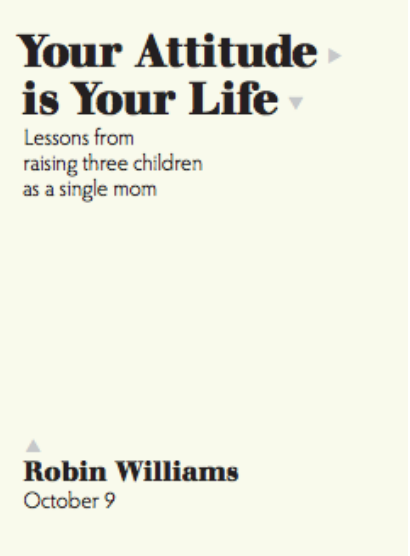
Fonts

Nice sans serif fonts; move away from Arial

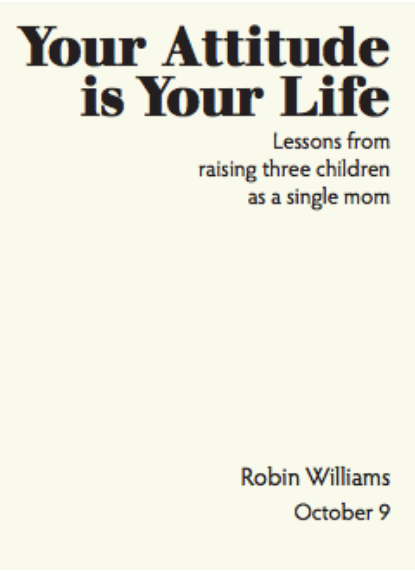
Design principles



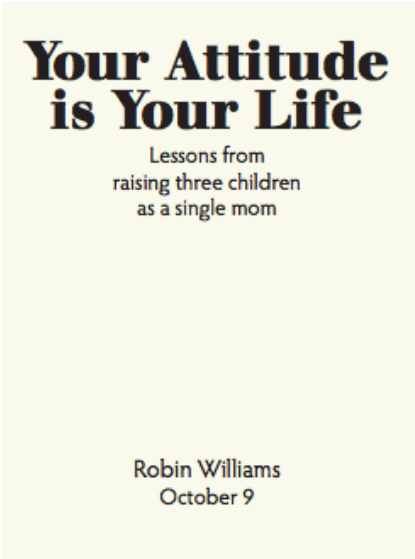
Contrast



Repetition



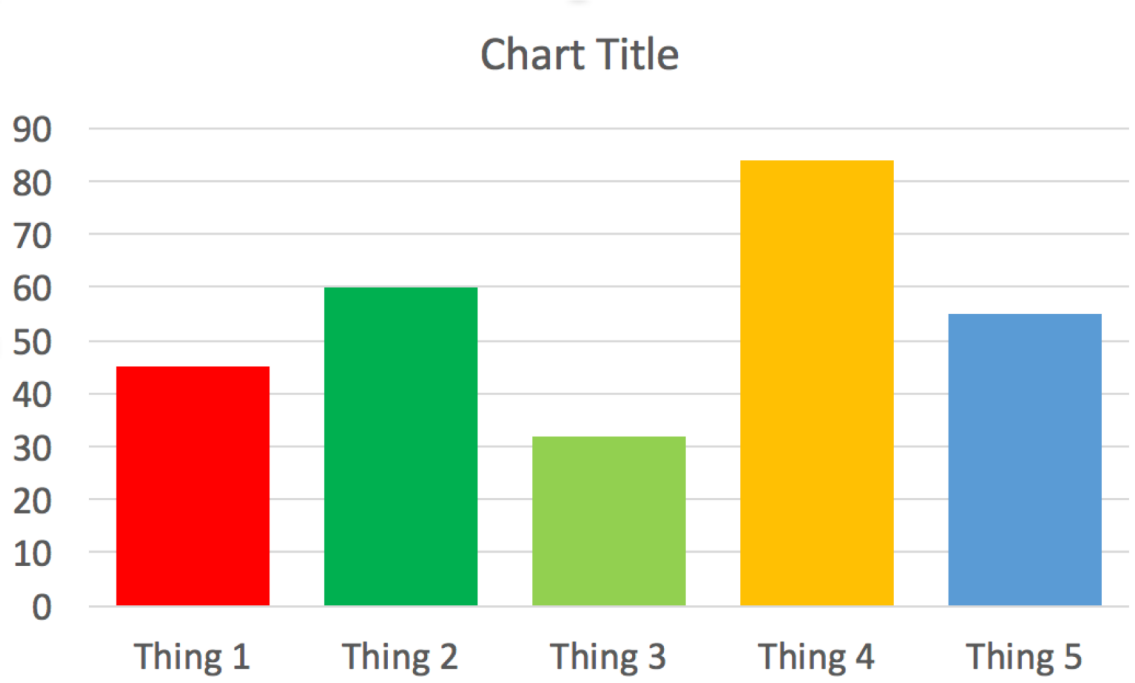
Alignment

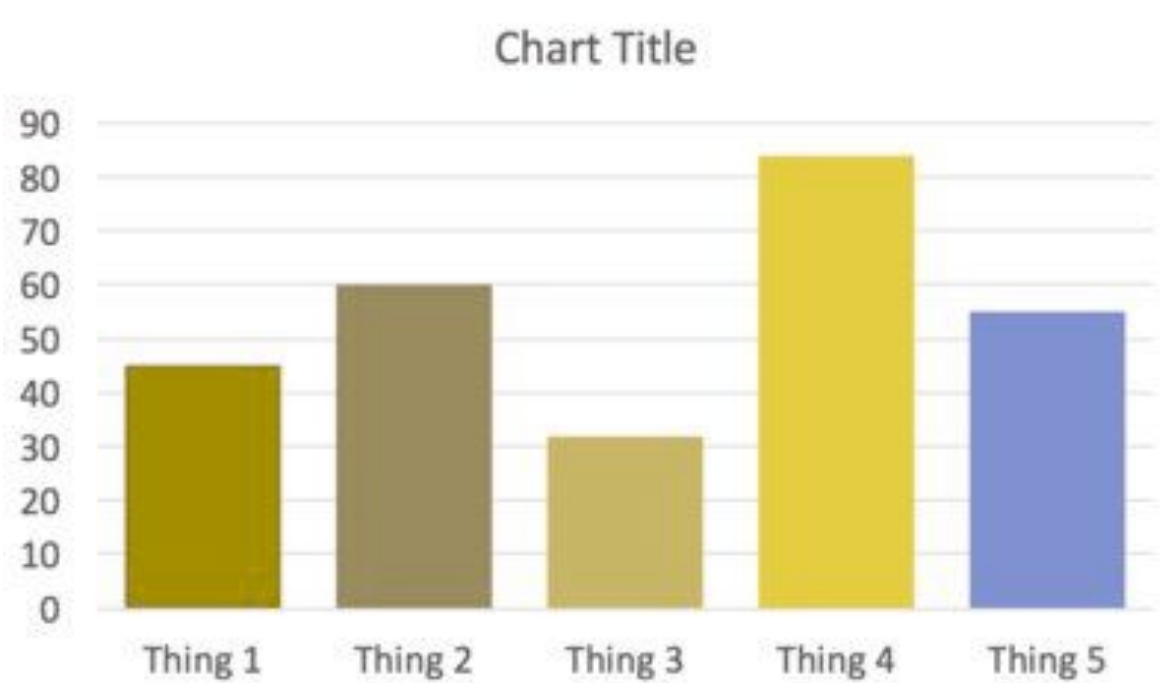


Proximity

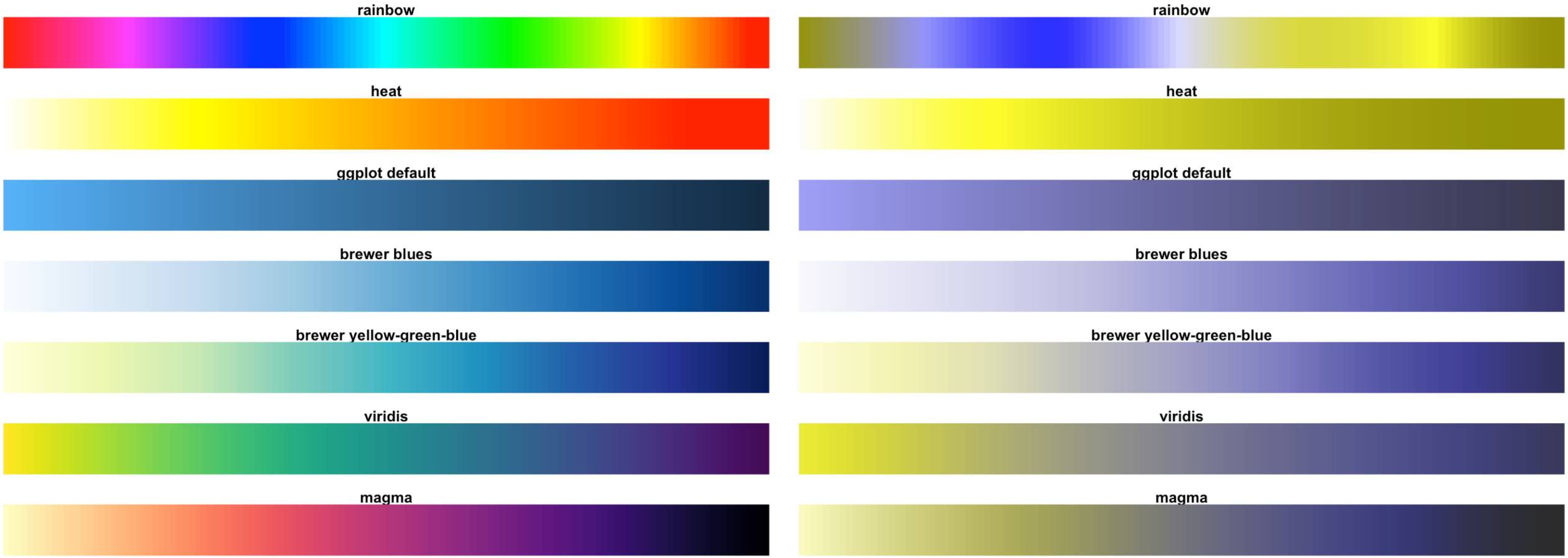
Colors

Be aware of colorblind users!



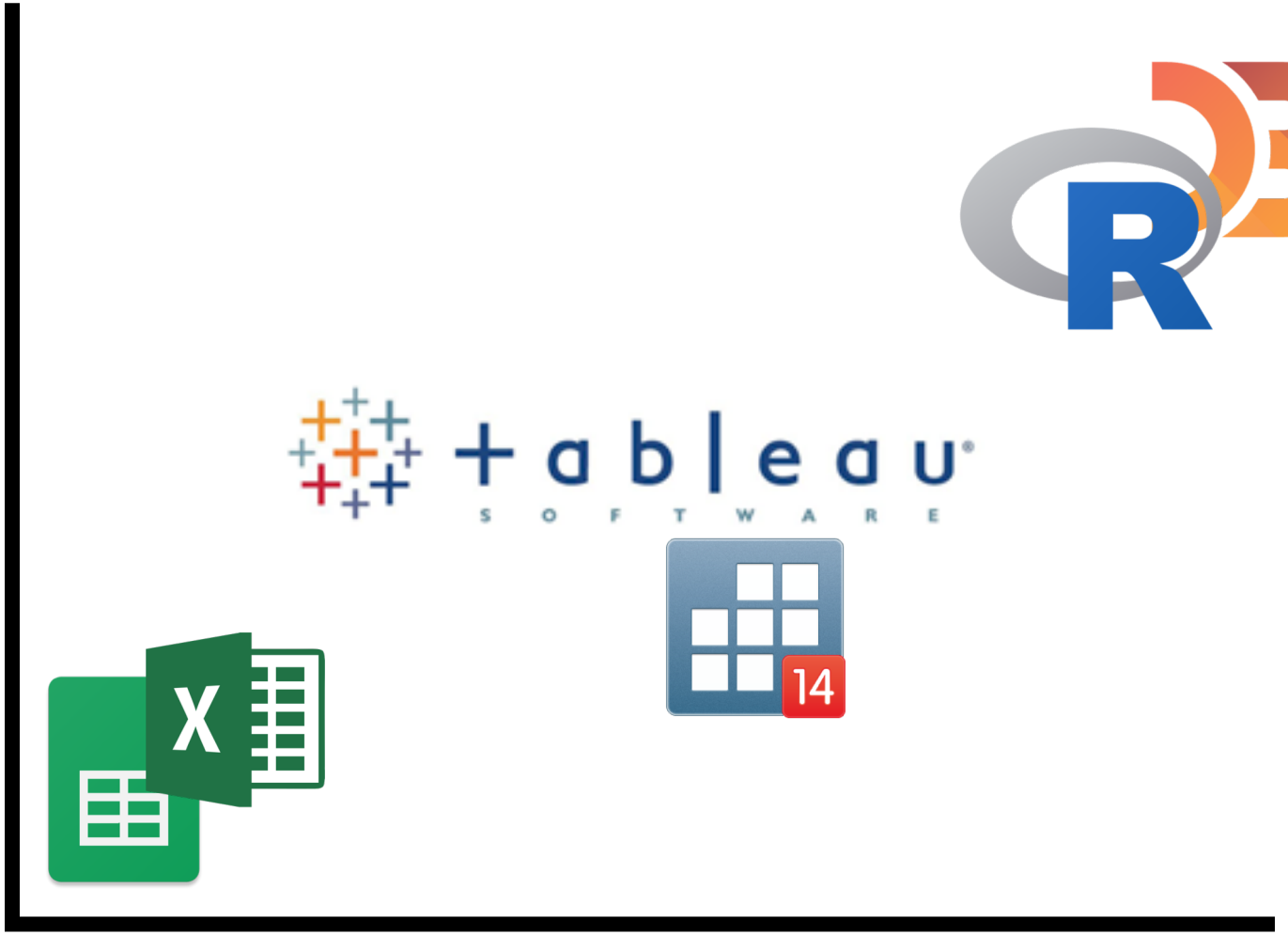


Choose universally accessible palettes



How do I do all this?

Flexibility and power



Barrier to entry (amount of coding required)

Take a sad plot and make it better

talks.andrewheiss.com/wcnpd18