432 Class 25 Slides

github.com/THOMASELOVE/2020-432

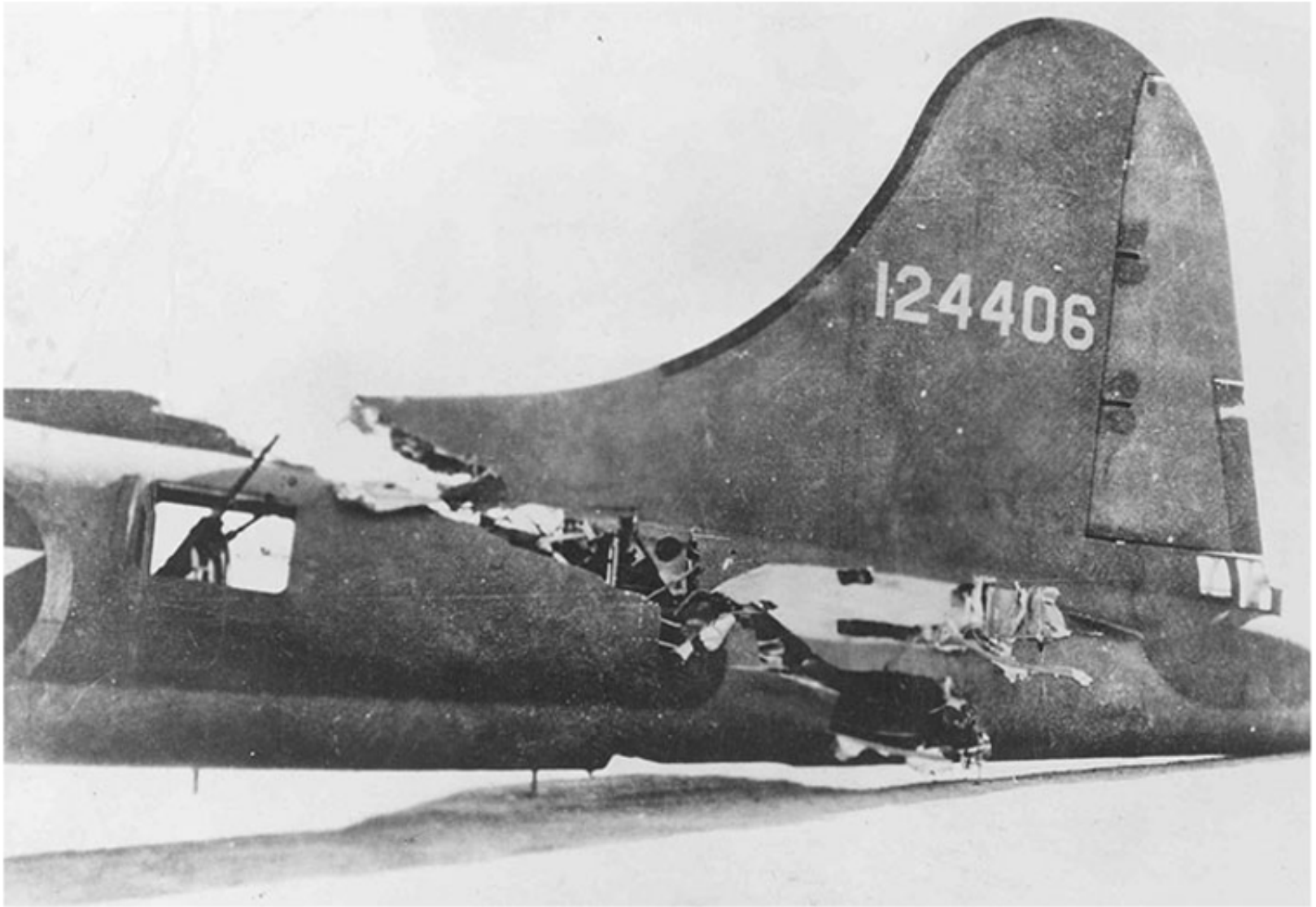
2020-04-23

Sources for Today's Material include

- Howard Wainer *Visual Revelations*
- Andrew Gelman and `andrewgelman.com`
- Christopher Gandrud and his book *Reproducible Research with R and R Studio*
- Karl Broman *Creating Effective Figures and Tables* at tinyurl.com/graphs2017
- Edward Tufte and `edwardtufte.com`
- http://www.datavis.ca/gallery/index.php
- https://www.boredpanda.com/world-war-2-aircraft-survivorship-bias-abraham-wald/
- plus this link to financialgazette.co/zw

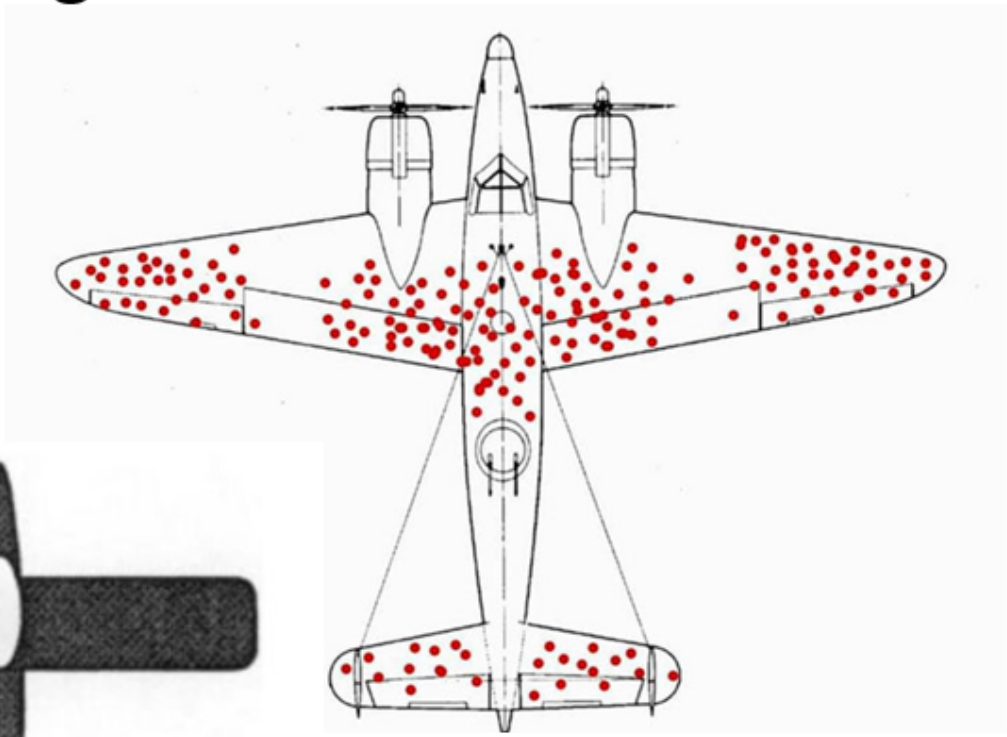
The Abraham Wald Story

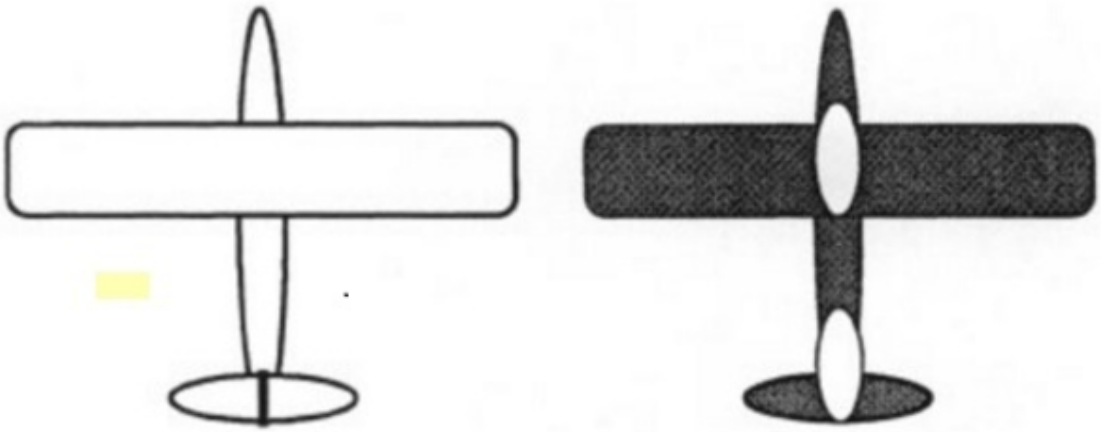
During WWII, a group of statisticians had a difficult task to solve

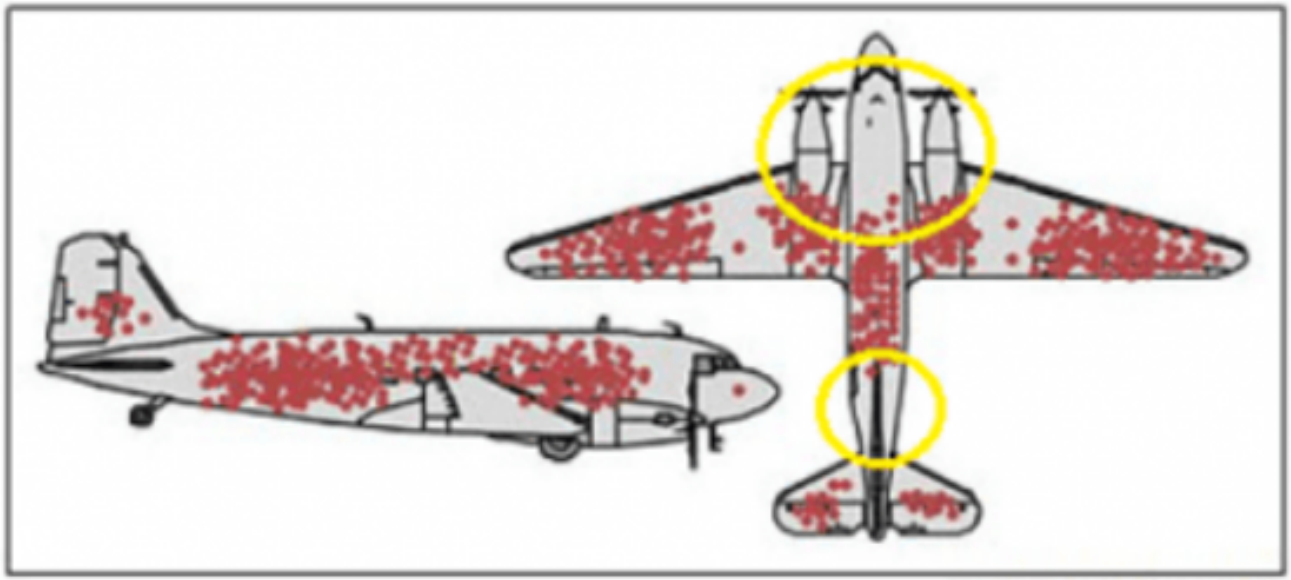


They were asked to evaluate and determine which parts of the aircraft needed to be up-armored in order to minimize the damage of the planes from enemy fire.

Diagram of all of the places where the planes were damaged the most







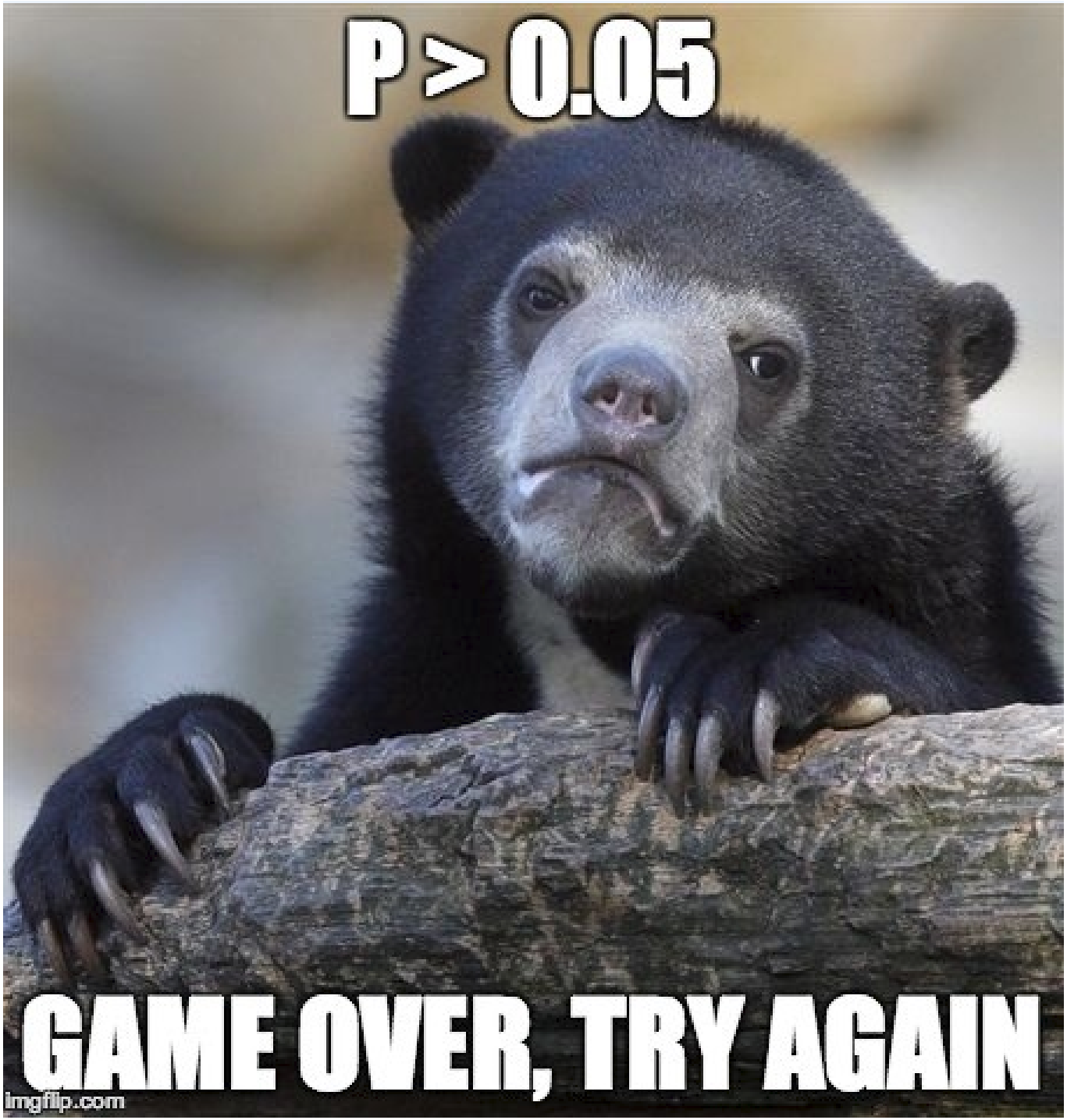
Credit: Cameron Moll

Gentlemen, you need to put more armour-plate where the holes aren't because that's where the holes were on the airplanes that didn't return - Abraham Wald 1942.

Presenting Research

- Usually, this is highly abridged
  - Slide shows
  - Abstracts
  - Journal articles
  - Books
  - Websites

Your job is usually to announce the findings and try to convince us that the results are correct.

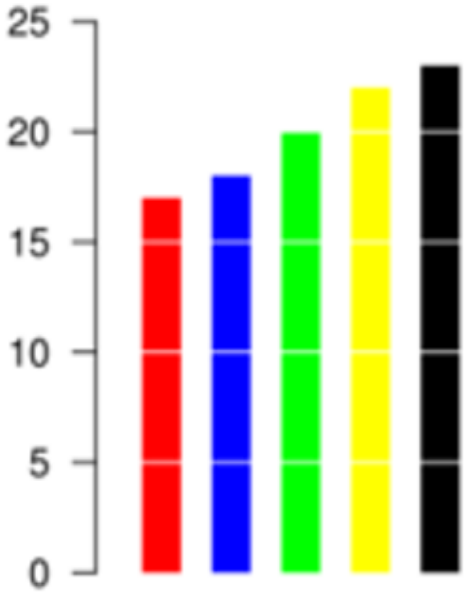

P > 0.05
GAME OVER, TRY AGAIN
imgflip.com

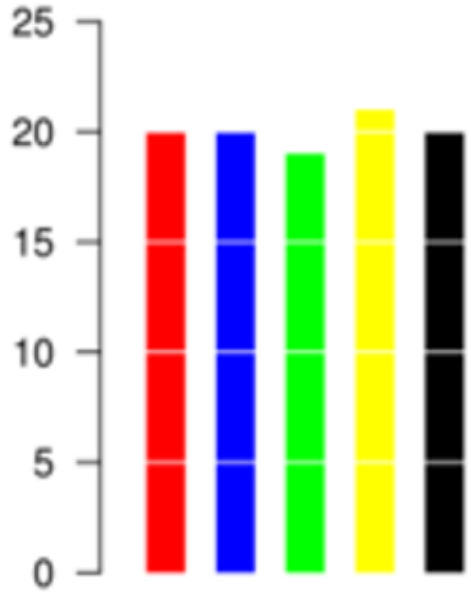
You Have Ten Minutes?

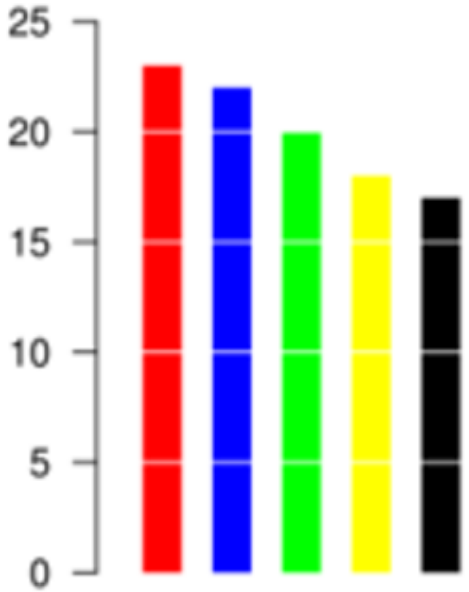
- No time for subtlety.
- Round, a lot.
- Edit, ruthlessly.
  - One pass through software ("default options") is never enough.
  - Better for people to leave the table hungry than stuffed.
- Have something to say, and say it clearly.
- Some possibilities are never a good choice.

Stay Away from the Pie!

Which of these three bar graphs describes the same data as pie graph A?



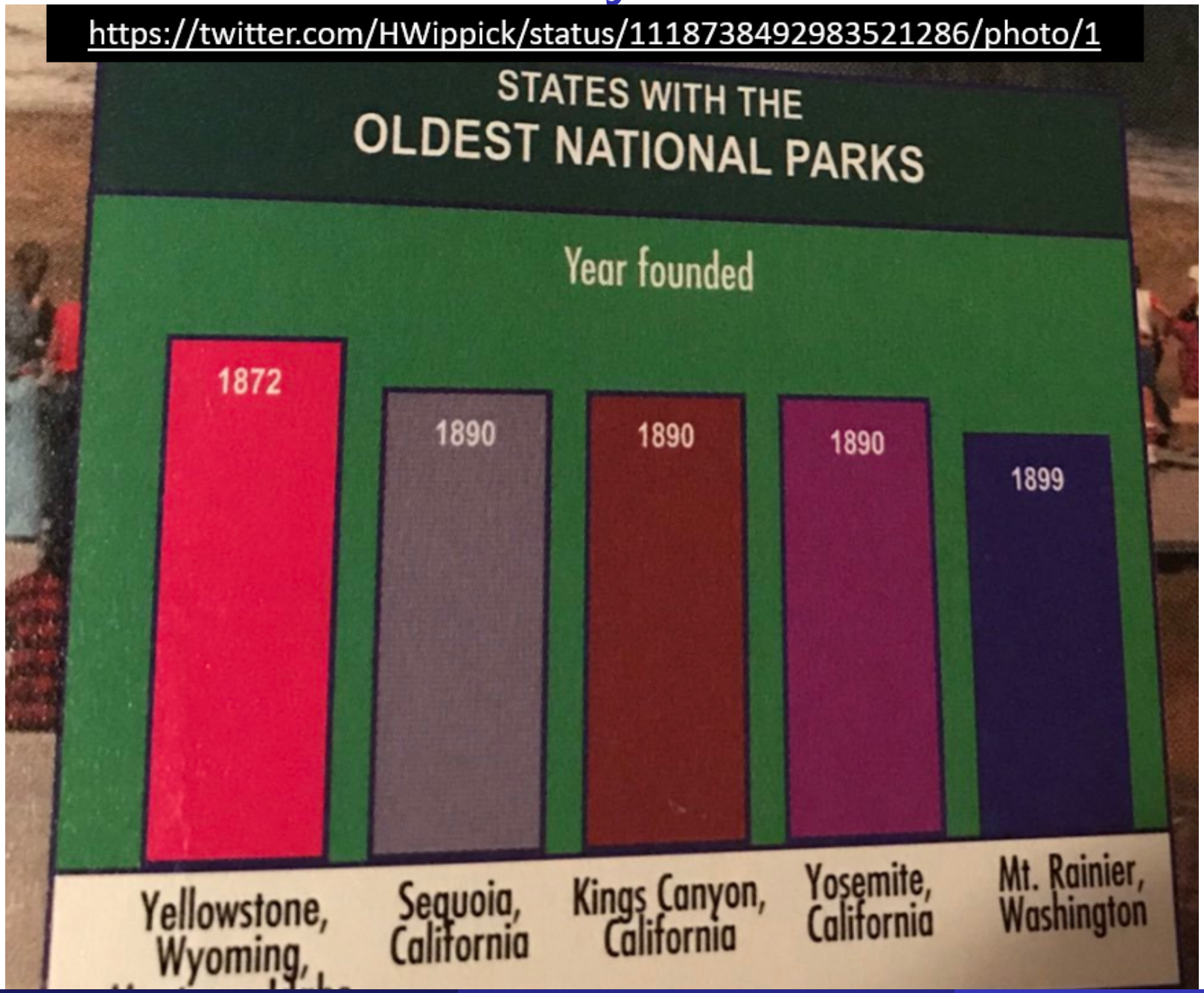




Stay Away from the Pie!

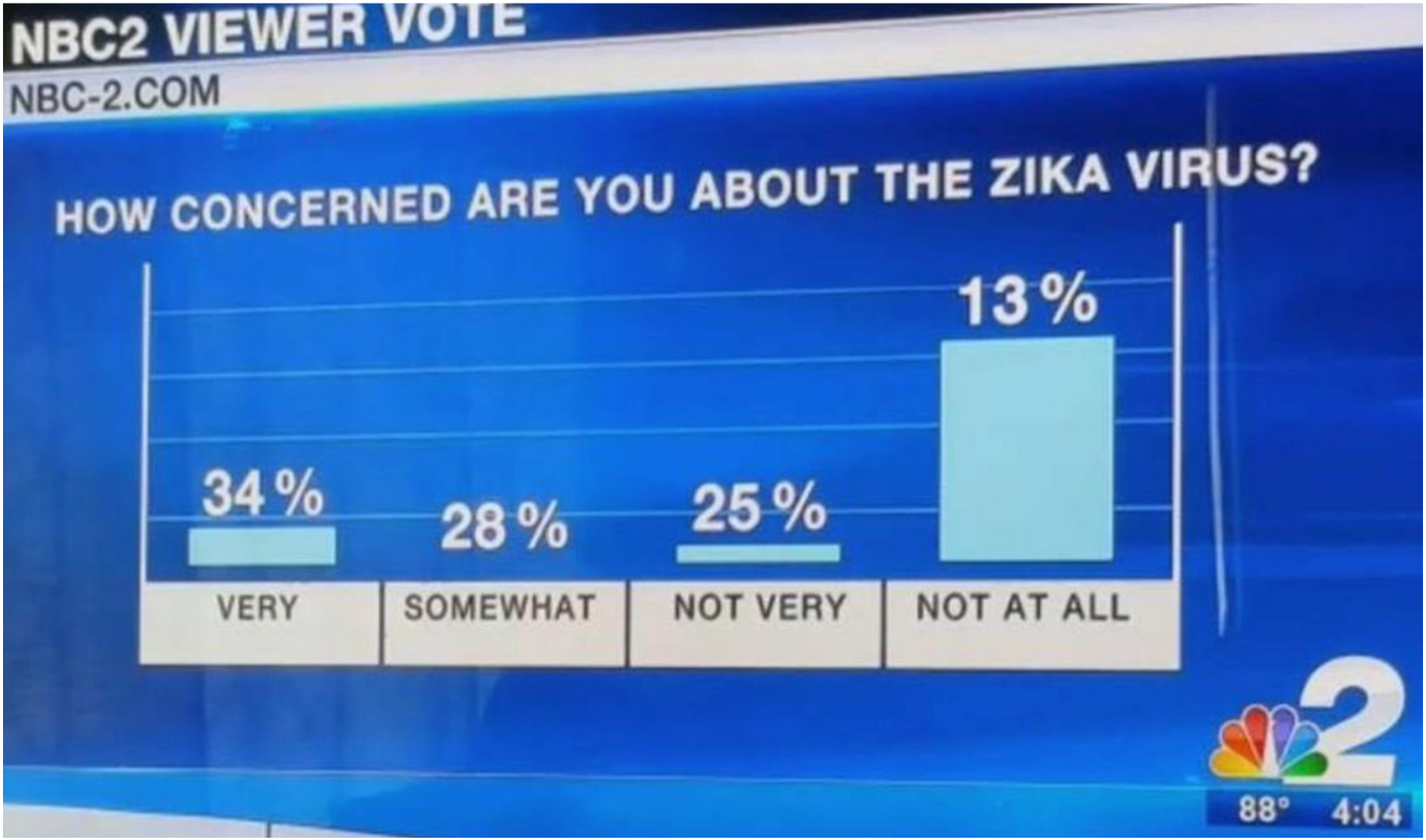
Not that bars are always better

https://twitter.com/HWippick/status/1118738492983521286/photo/1

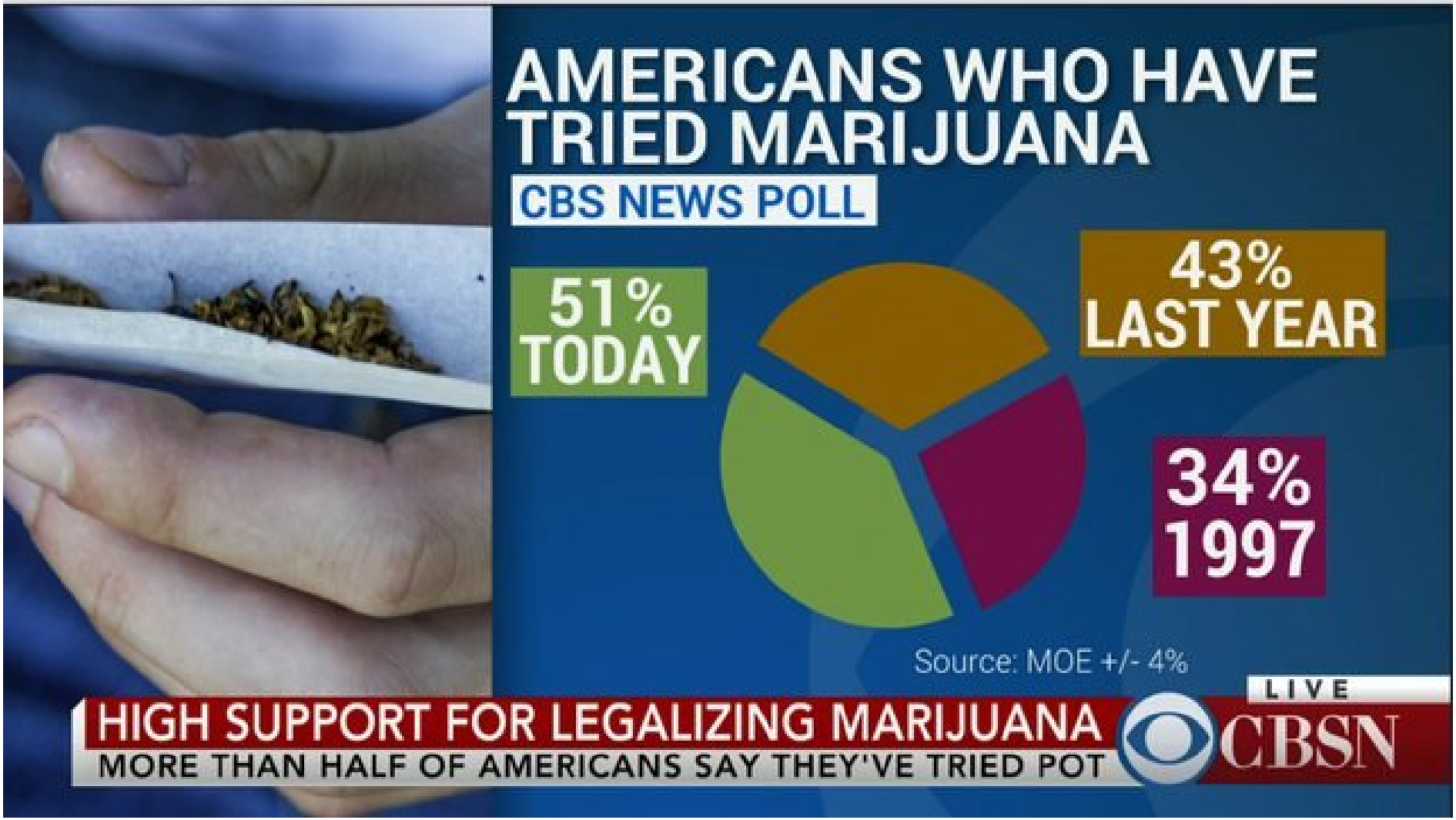


Not that bars are always better

https://twitter.com/asher_rosinger/status/1119278062804328448/photo/1

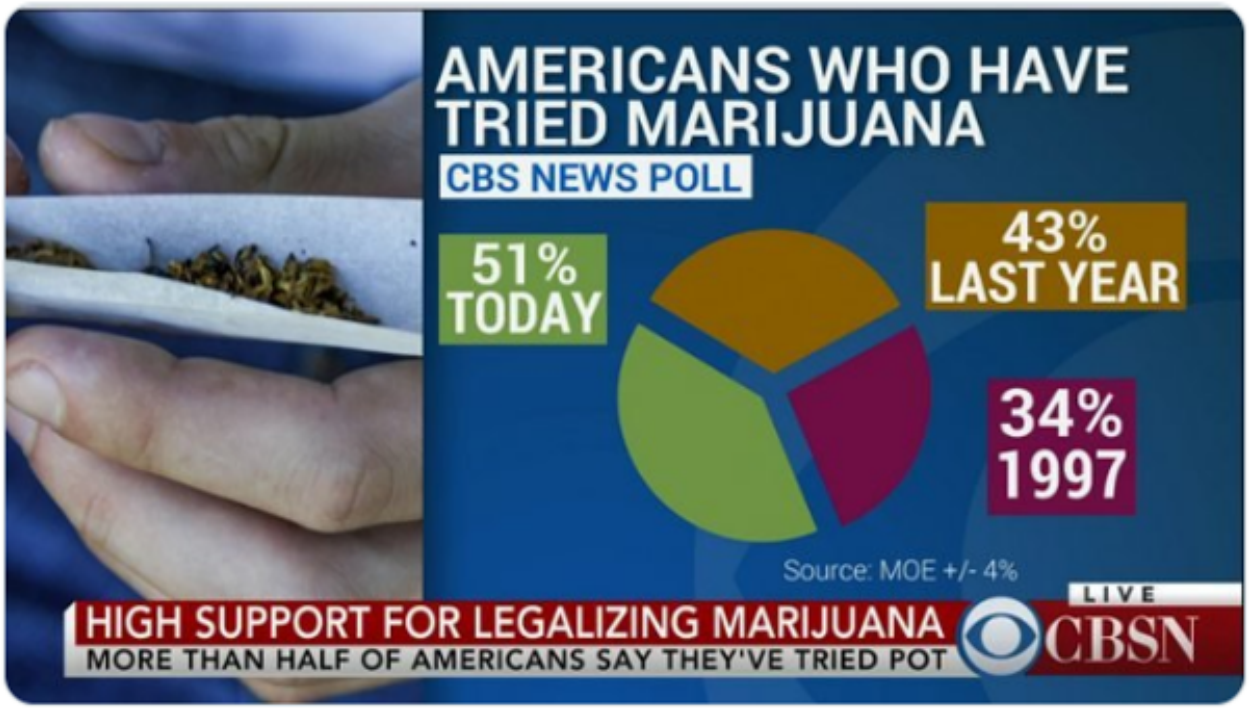


What's Wrong With This Picture?



Dorsa Amir
@DorsaAmir

Easily the funniest data viz I've ever seen.
(1) First of all, it's a pie chart.
(2) The total is greater than 100%.
(3) The relevant categories are today, last year, and... the year 1997?
(4) The margin of error (MOE) is listed as the source.



7:06 PM · Apr 17, 2019 · Twitter Web Client

Clearly Communicating Quantitative Information

- Are the most important elements or relationships visually most prominent?
- Are the elements, symbol shapes and colors consistent with their use in previous graphs?
- Are all of the graphical elements necessary to convey the relationships?
- Are the graphical elements accurately positioned and scaled?

Source: http://www.datavis.ca/gallery/index.php

What are you trying to do?

- Is this **information visualization** (grabby, visually striking - dramatize the problem to draw the casual viewer in deeper)
- Or **statistical graphics** (reveal patterns and discrepancies for viewers who are already interested in the problem)

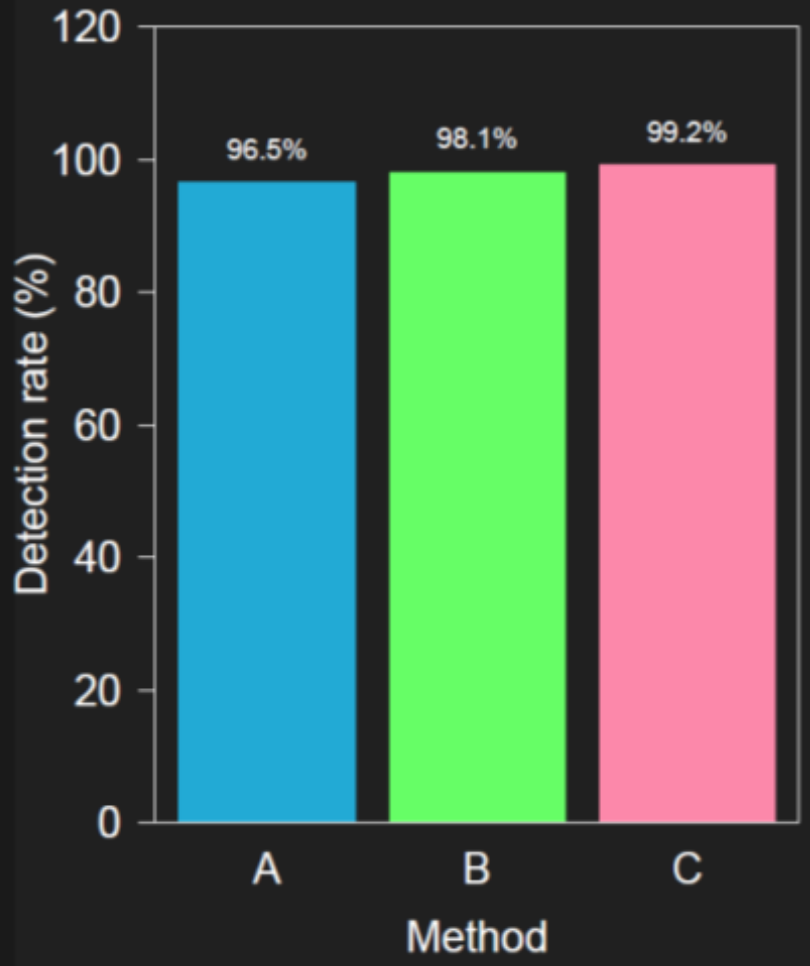
Make tradeoffs carefully - meaningful choices.

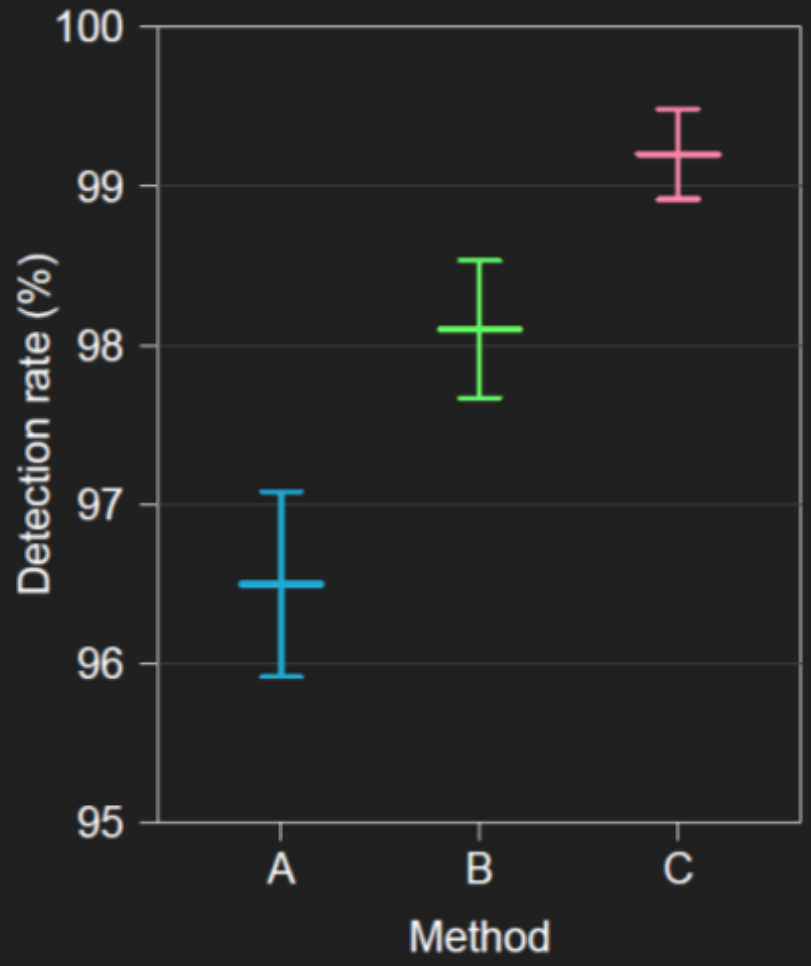
from Karl Broman (tinyurl.com/graphs2017)

Displaying data well

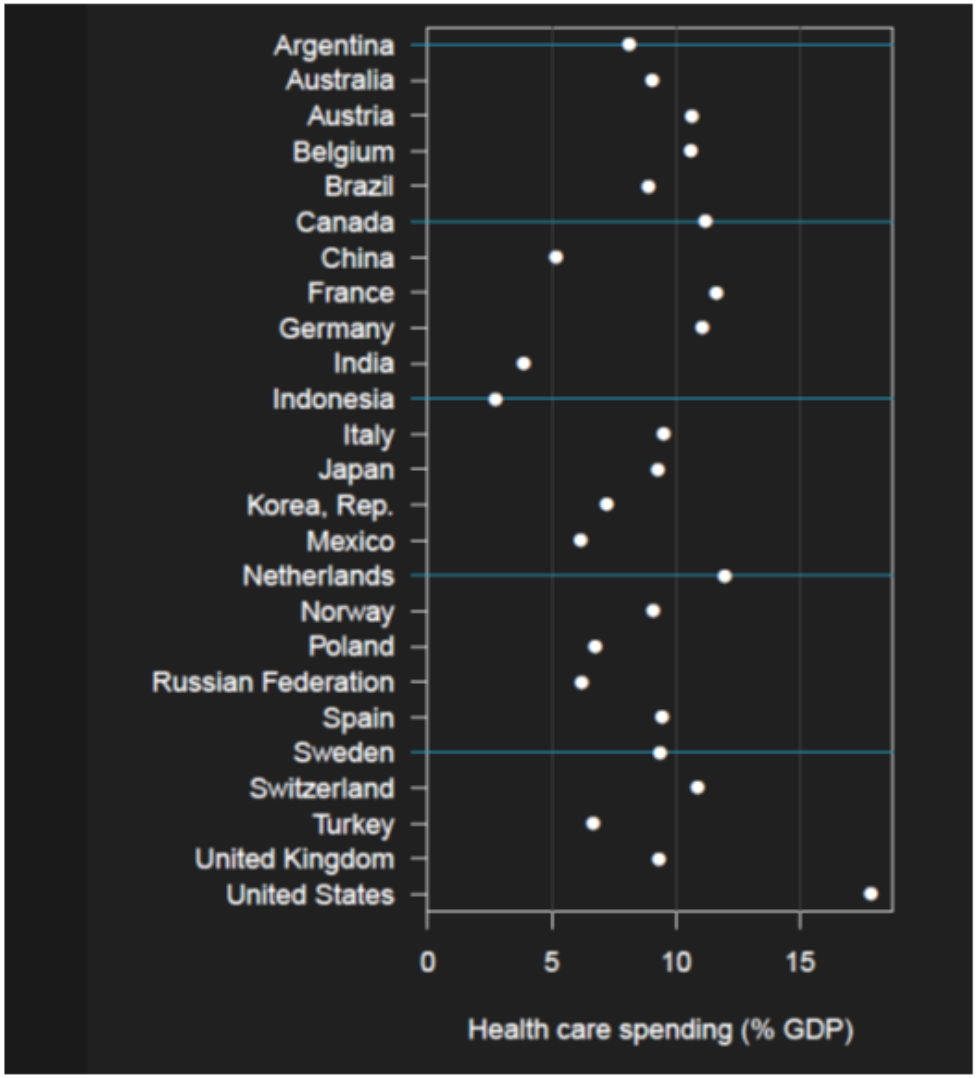
- Be accurate and clear.
- Let the data speak.
  - Show as much information as possible, taking care not to obscure the message.
- Science not sales.
  - Avoid unnecessary frills (esp. gratuitous 3d).
- In tables, every digit should be meaningful. Don't drop ending 0's.

Must you include 0?



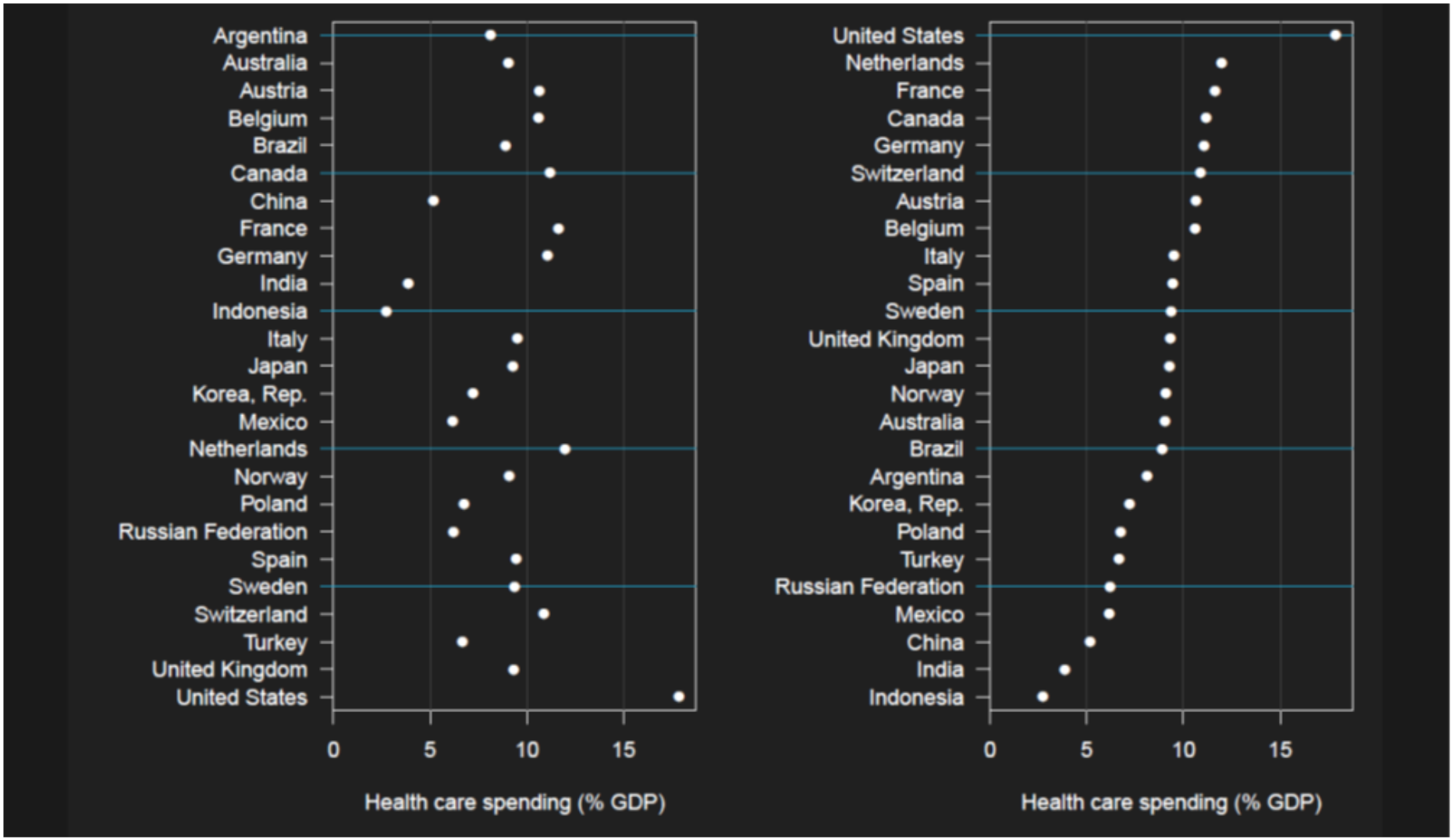


From Karl Broman...



Karl Broman, "Creating Effective Figures and Tables" at tinyurl.com/graphs2017

Don’t sort alphabetically



Karl Broman, “Creating Effective Figures and Tables” at tinyurl.com/graphs2017

You Have Ten Minutes?

- No time for subtlety.
- Round, a lot.
- Edit, ruthlessly.
  - One pass through software ("default options") is never enough.
  - Better for people to leave the table hungry than stuffed.
- Have something to say, and say it clearly.
- Stay away from the pie.

Data Visualization: Napoleon's Russian Campaign

Wainer: Chapter 4 of *Visual Revelations*

CHAPTER 4 Three Graphic Memorials

"*Hear, forget; see, remember.*" The wisdom of this ancient Confucian saying is apparent. Memorable memorials are visual. Who can ever forget the tragedy chronicled by the austere black granite wall that is the Vietnam Memorial? It is massive in form and content, built from the space taken by the more than 58,000 names inscribed upon it. As the loss of life increases, so too does the height of the wall, and the emotions it evokes. It is a very personal thing. William A. Atwell, Terry Lee Dillard, Ward K. Patton, Jerry Lee Graves, Edward J. Downs, John E. Rice, Jack M. Strong—these names join with thousands of others to form the wall. The interaction of the monument with those who come to it, whether to seek out a particular name or to picnic, often becomes part of the diverse images we take away with us. The tragedy of Vietnam written in the small becomes large and indelible.

The History

It's 1812.

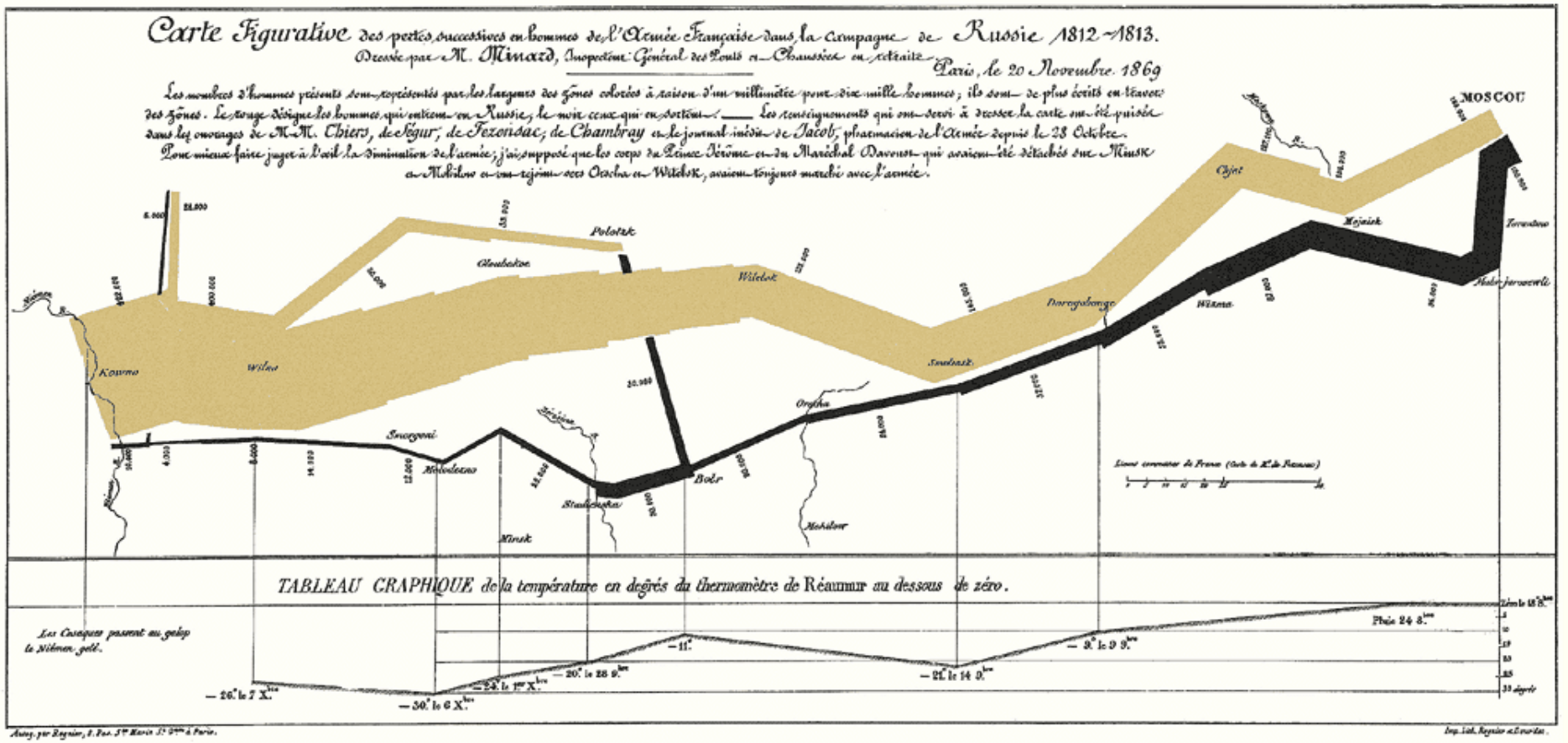
- Napoleon has most of Europe (outside of the United Kingdom) under his control.
- But he cannot break through the defenses of the U.K., so he decides to place an embargo on them.
- The Russian Czar, Alexander, refuses to participate in the embargo.

So Napoleon gathers a massive army of over 400,000 to attack Russia in June 1812.

- Meanwhile, Russia has a plan. As Napoleon's troops advance, the Russian troops burn everything they pass.

Charles Minard's original map

Napoleon's disastrous Russian Campaign of 1812

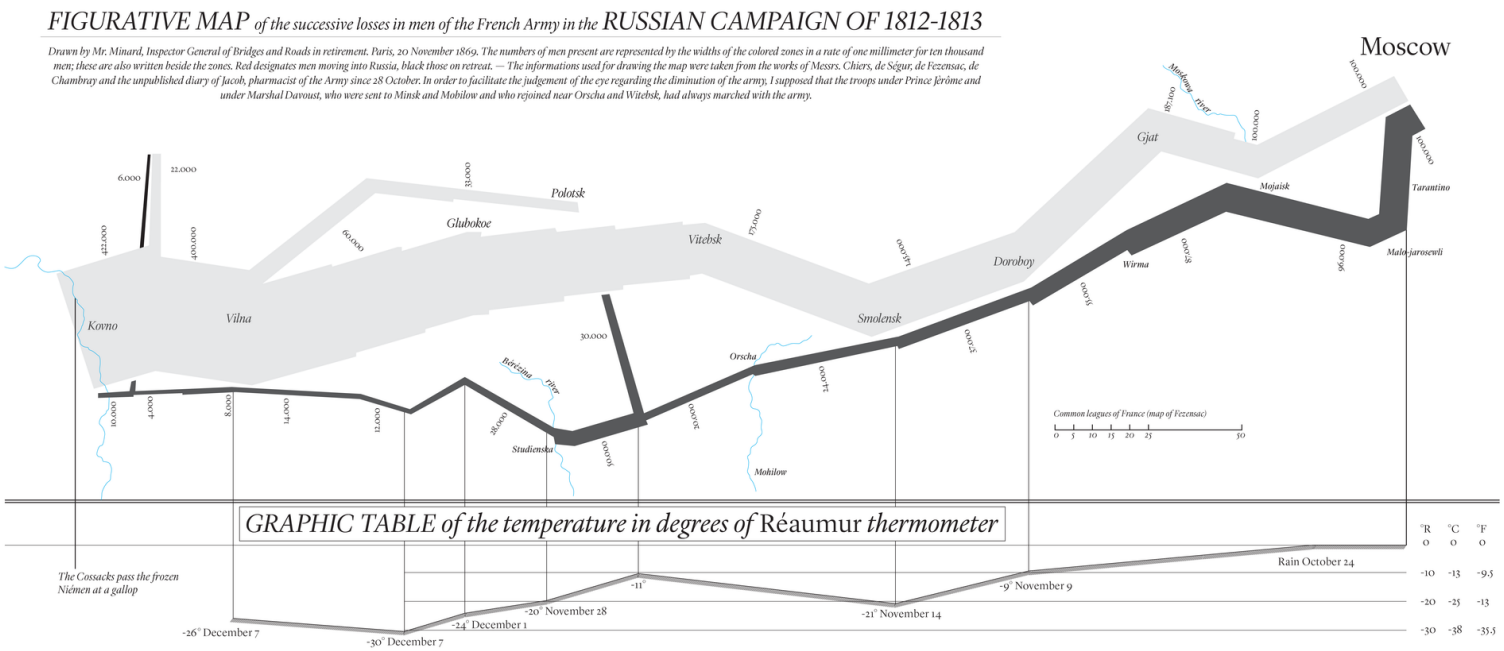


Wainer: Chapter 4 [b]

Napoleon's Russian Campaign

Memorializing that portion of the generation of young French men lost in Napoleon's ill-fated Russian campaign was surely part of Charles Joseph Minard's motivation in the construction of his famous 1869 graphic. Minard's plot, shown in figure 1, depicts the movement of the French army from the time it crossed the Polish-Russian border with 422,000 men in June of 1812. The shrinking size of the army is characterized by the progressive narrowing of the broad band stretching across the map. In the original scale, each millimeter of its width represents 10,000 men. When the army reached Moscow in September, only 100,000 remained. The city was deserted, and the army began its retreat, depicted by the darker line below. It is linked to the temperature scale showing quantitatively the depths of the Russian winter. The banks of the Berezina River were littered with the bodies of the 22,000 men who perished as the November temperature dropped to −20°. When the remainder of the army crossed into Poland as the year ended, only 10,000 men remained.

A Modern Redrawing of Minard's Original Map



Source: By Iñigo Lopez - Own work, CC BY-SA 4.0, at this link

What are we looking at?

- The numbers of Napoleon's troops by location (longitude)
  - Organized by group (at one point they divided into three groups) and direction (advance, then retreat)
- The path that his troops took to Moscow and back again
- The temperature experienced by his troops when winter settled in on the return trip
- Historical context, as shown in the passage of time
- Geography (for example, river crossings)

Wainer: Chapter 4 [c]

The story of the tragedy is clear. We can see the bodies frozen into the snow. Marey told how this graph "brought tears to the eyes of all France."[1] No wonder; there were few families unaffected.

Minard's depiction of Napoleon's Russian campaign has been characterized as perhaps "the best statistical graphic ever drawn."[2] Why? It is not the quality of the pen stroke, although it certainly passes muster in that regard. It is the importance and richness of the data. A single page carries six variables that tell the evocative story of where and how thousands of men died. Its poignancy is heightened through the immediate and graphic answer to the question, Compared to what? Ten thousand men returned. A lot or a few? Opposing the returning trickle against the departing torrent answers the question. The difference between them measures the tragedy. But nowhere does the shrinking distance between two lines depict a more touching tragedy than in my next example.

I'll spare you that next example in favor of showing you some of the work of Edward Tufte.

	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: Hermann Brenner, "Long-term survival rates of cancer patients achieved by the end of the 20th century: a period analysis," *The Lancet*, 360 (October 12, 2002), 1131-1135.

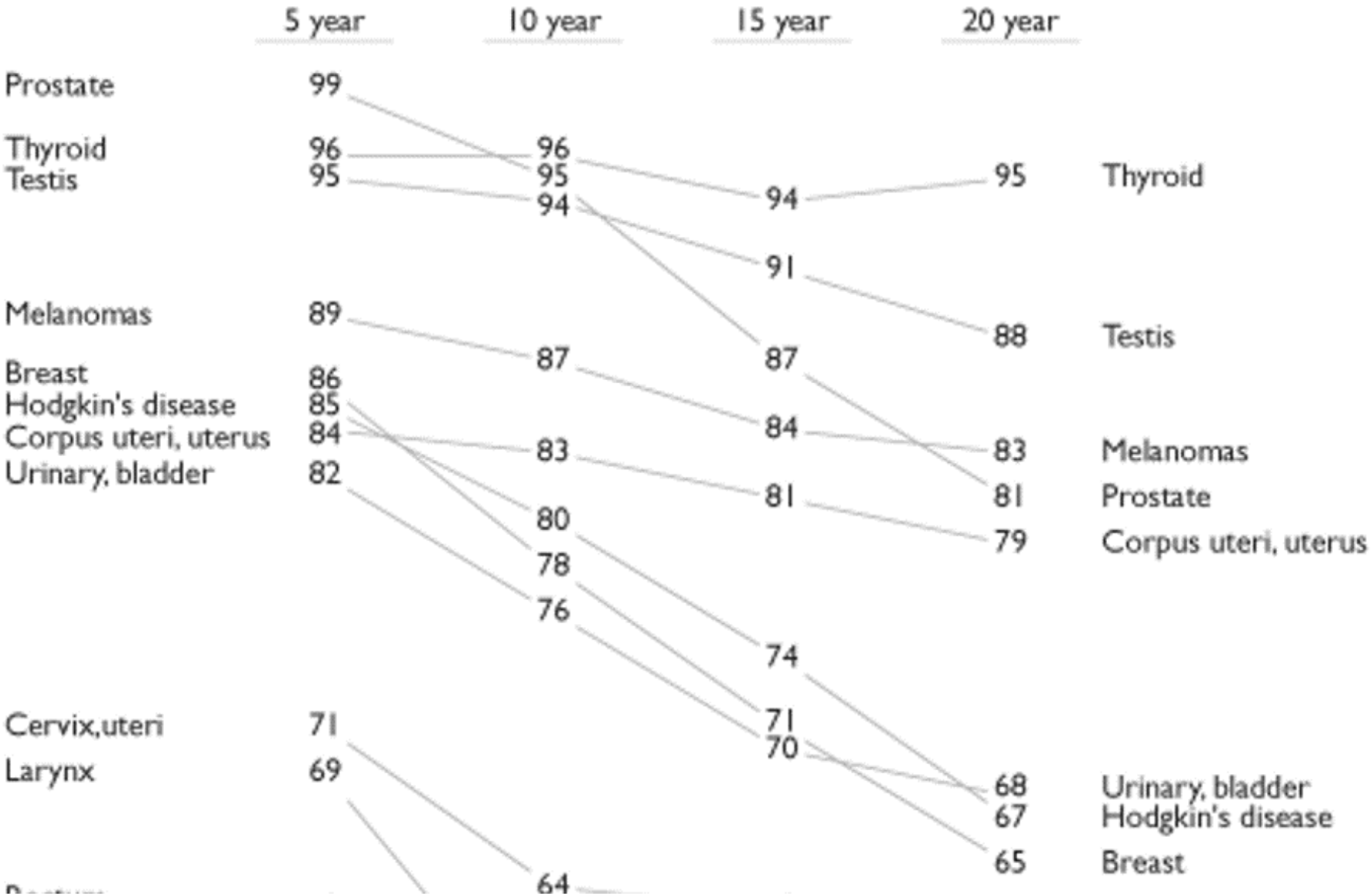
edwardtufte.com

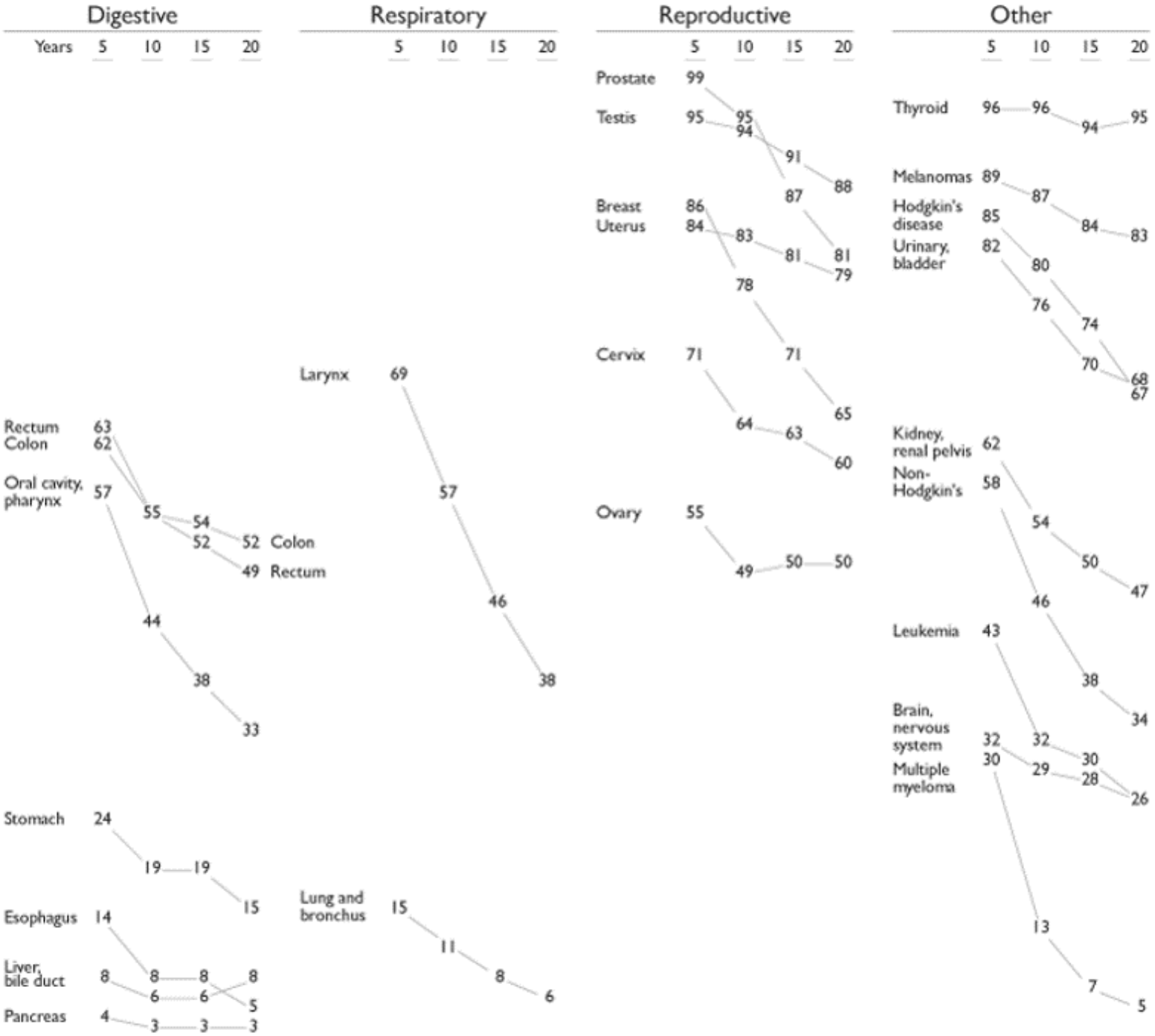
Estimates of relative survival rates, by cancer site

	% survival rates and their standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
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
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary, bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix, uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney, renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity, pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain, nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver, 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	1.5	2.7	0.6	2.7	0.8

edwardtufte.com

Slopegraphs!




Digestive
Years 5 10 15 20
Rectum 63
Colon 62
Oral cavity, pharynx 57
55
54
52
52 Colon
49 Rectum
44
38
33
Stomach 24
19
19
15
Esophagus 14
Liver, bile duct 8 8 8 8
6 6 5
Pancreas 4 3 3 3
Respiratory
5 10 15 20
Larynx 69
57
46
38
Lung and bronchus 15
11
8
6
Reproductive
5 10 15 20
Prostate 99
Testis 95 95
94
91
88
87
Breast 86
Uterus 84 83
81
81
79
78
Cervix 71
71
65
64
63
60
Ovary 55
49
50
50
Other
5 10 15 20
Thyroid 96 96 94 95
Melanomas 89
87
84
83
Hodgkin's disease 85
Urinary, bladder 82
80
76
74
70
68
67
Kidney, renal pelvis 62
Non-Hodgkin's 58
54
50
47
46
Leukemia 43
38
34
Brain, nervous system 32 32 30
Multiple myeloma 30 29 28 26
13
7
5

edwardtufte.com

In addition to slopegraphs, consider **sparklines**: intense, simple, word-sized graphics

The most common data display is a noun accompanied by a number.

For example, a medical patient's current level of glucose is reported in a clinical record as a word and number:

---

glucose 6.6

---

sparklines: intense, simple, word-sized graphics

Placed in the relevant context, a single number gains meaning. Thus, the most recent measurement of glucose should be compared with earlier measurements for the patient. This data-line shows the path of the last 80 readings of glucose:

 glucose 6.6

sparklines: intense, simple, word-sized graphics

Lacking a scale of measurement, this free-floating line is de-quantified. At least we do know the value of the line's right-most data point, which corresponds to the most recent value of glucose, the number recorded at far right. Both representations of the most recent reading are tied together with a color accent:

sparklines: intense, simple, word-sized graphics

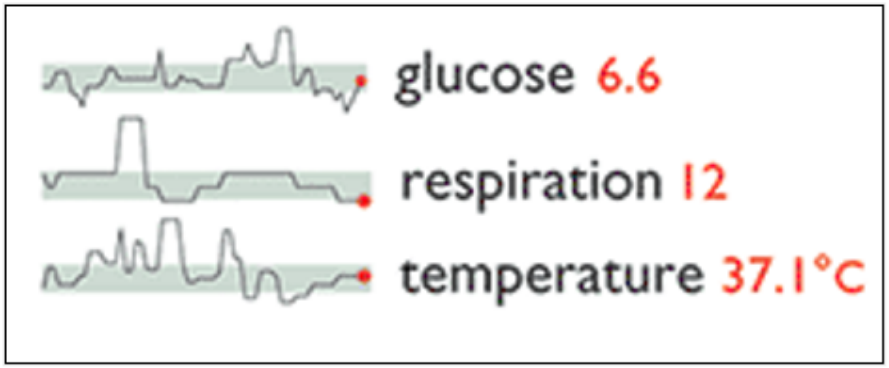
Some useful context is provided by showing the normal range of glucose, here as a gray band. Compared to normal limits, readings

above the band horizon are elevated, those below reduced:

 glucose 6.6 or glucose  6.6

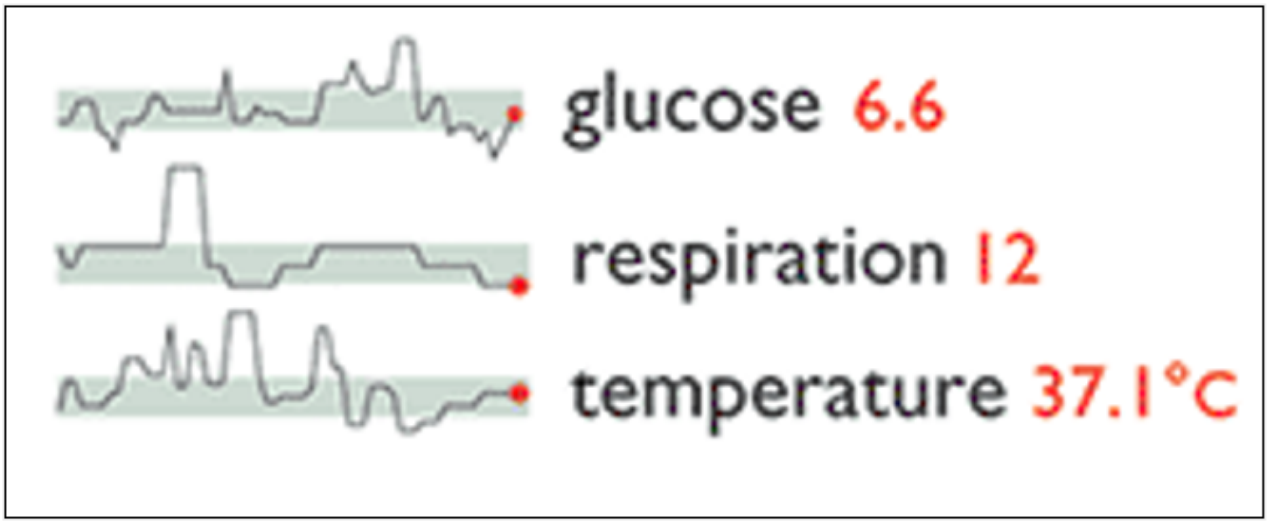
sparklines: intense, simple, word-sized graphics

For clinical analysis, the task is to detect quickly and assess wayward deviations from normal limits, shown here by visual deviations outside the gray band. Multiplying this format brings in additional data from the medical record; a stack, which can show hundreds of variables and thousands of measurements, allows fast effective parallel comparisons:



sparklines: intense, simple, word-sized graphics

These little data lines, because of their active quality over time, are named sparklines—small, high-resolution graphics usually embedded in a full context of words, numbers, images. Sparklines are datawords: data-intense, design-simple, word-sized graphics.



edwardtufte.com

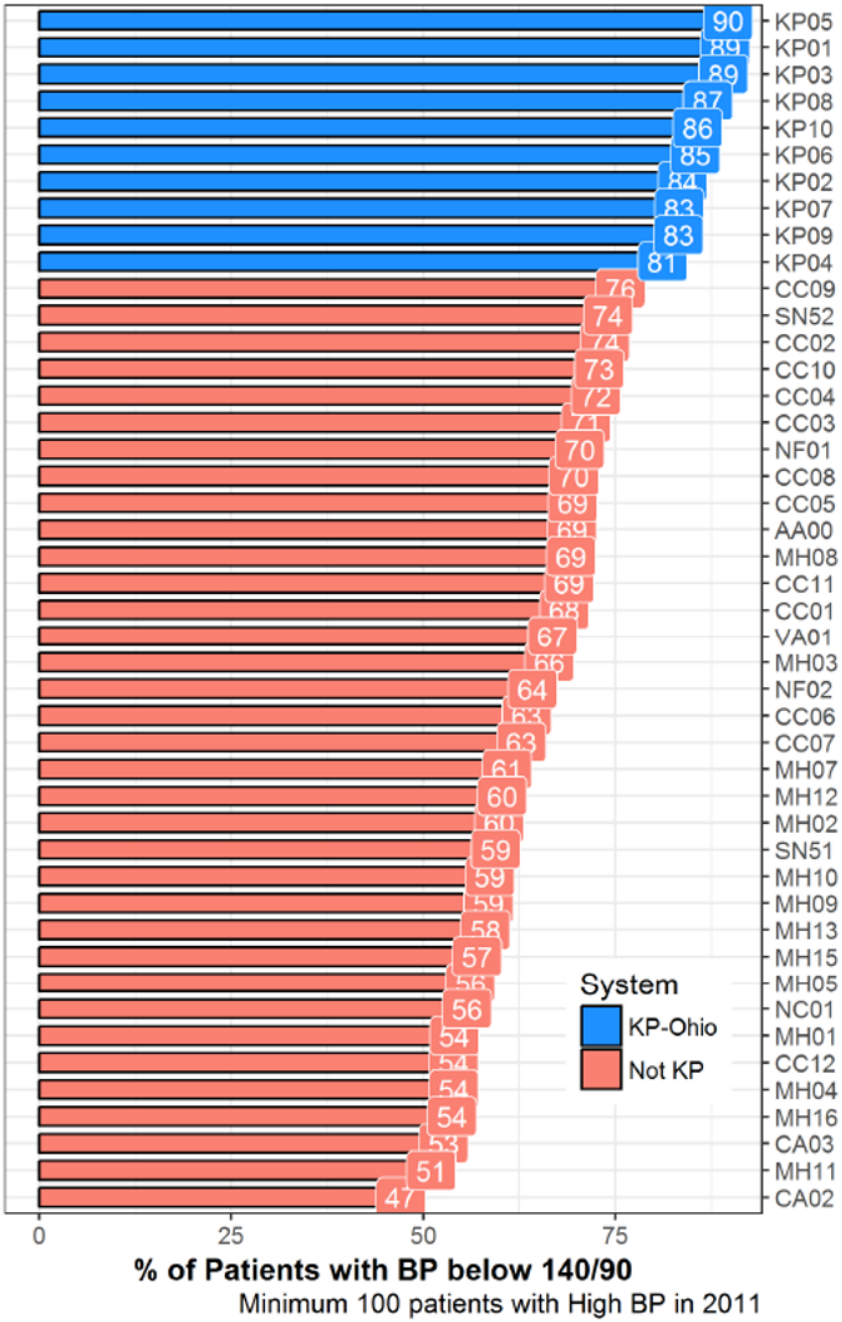
Do as I do? (Better Health Partnership work)

Exhibit A. Better Health Partnership: Patient Characteristics, 2016

	Diabetes		Heart Failure		High Blood Pressure	
# Health Systems	8		3		9	
# Primary Care Practices	72		35		79	
# Primary Care Providers	748		477		832	
# of Patients	51,305		8,488		183,745	
	Better Health Partnership	Range by Practice	Better Health Partnership	Range by Practice	Better Health Partnership	Range by Practice
Insurance (%)						
Medicare	**47.6**	15 - 77	**67.2**	49 - 93	**52.6**	13 - 83
Commercial (and Veterans)	**37.2**	0 - 62	**20.8**	3 - 31	**36.0**	2 - 60
Medicaid	**12.7**	0 - 70	**10.5**	0 - 39	**9.4**	0 - 68
Uninsured	**2.5**	0 - 36	**1.6**	0 - 6	**2.0**	0 - 38
Race / Ethnicity (%)						
White	**64.3**	2 - 98	**63.6**	2 - 100	**69.4**	2 - 99
Black or African-American	**28.1**	1 - 97	**31.8**	0 - 98	**25.8**	0 - 98
Hispanic or Latino	**4.7**	0 - 68	**2.9**	0 - 49	**2.7**	0 - 59
Other Race / Ethnicity	**2.8**	0 - 9	**1.7**	0 - 4	**2.1**	0 - 8
Demographics						
Average Age	**60.4**	52 - 66	**69.1**	58 - 79	**64.4**	51 - 73
% Female	**38.7**	1 - 75	**50.4**	32 - 75	**40.9**	2 - 76
% Low Income*	**26.6**	0 - 82	**31.5**	0 - 81	**33.2**	0 - 83
% Low Education*	**25.9**	0 - 83	**28.3**	0 - 75	**30.0**	0 - 86
% living in Cleveland	**33.9**	0 - 93	**42.5**	0 - 87	**31.3**	0 - 94
% in Cuyahoga County	**56.1**	0 - 100	**74.9**	7 - 100	**55.7**	0 - 99
Population Health						
% with BP below 140/90	**75.6**	56 - 90	**78.2**	70 - 92	**72.0**	36 - 85
% with BMI below 30	**33.3**	16 - 47	**47.8**	29 - 57	**46.4**	9 - 57
% Not Using Tobacco	**76.6**	42 - 92	**87.4**	63 - 96	**77.1**	12 - 93

* Living in neighborhoods with median income < $33,000; with high school graduation rate < 83%.

Patients with Good BP Control (%), 2011



Social Determinants and % Achieving Good BP Control, 2013

134,421 Northeast Ohio Adults with Hypertension in 2013

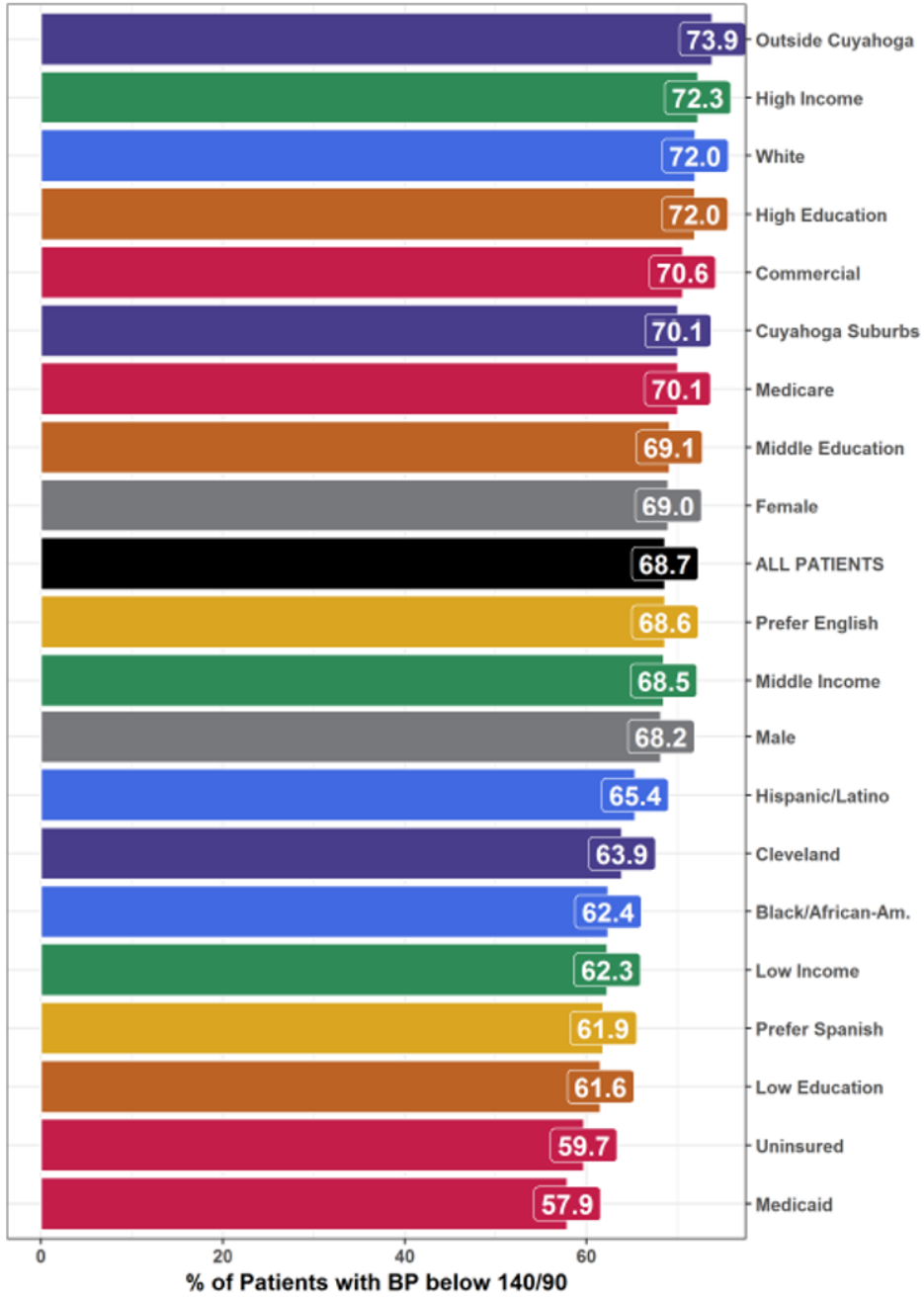
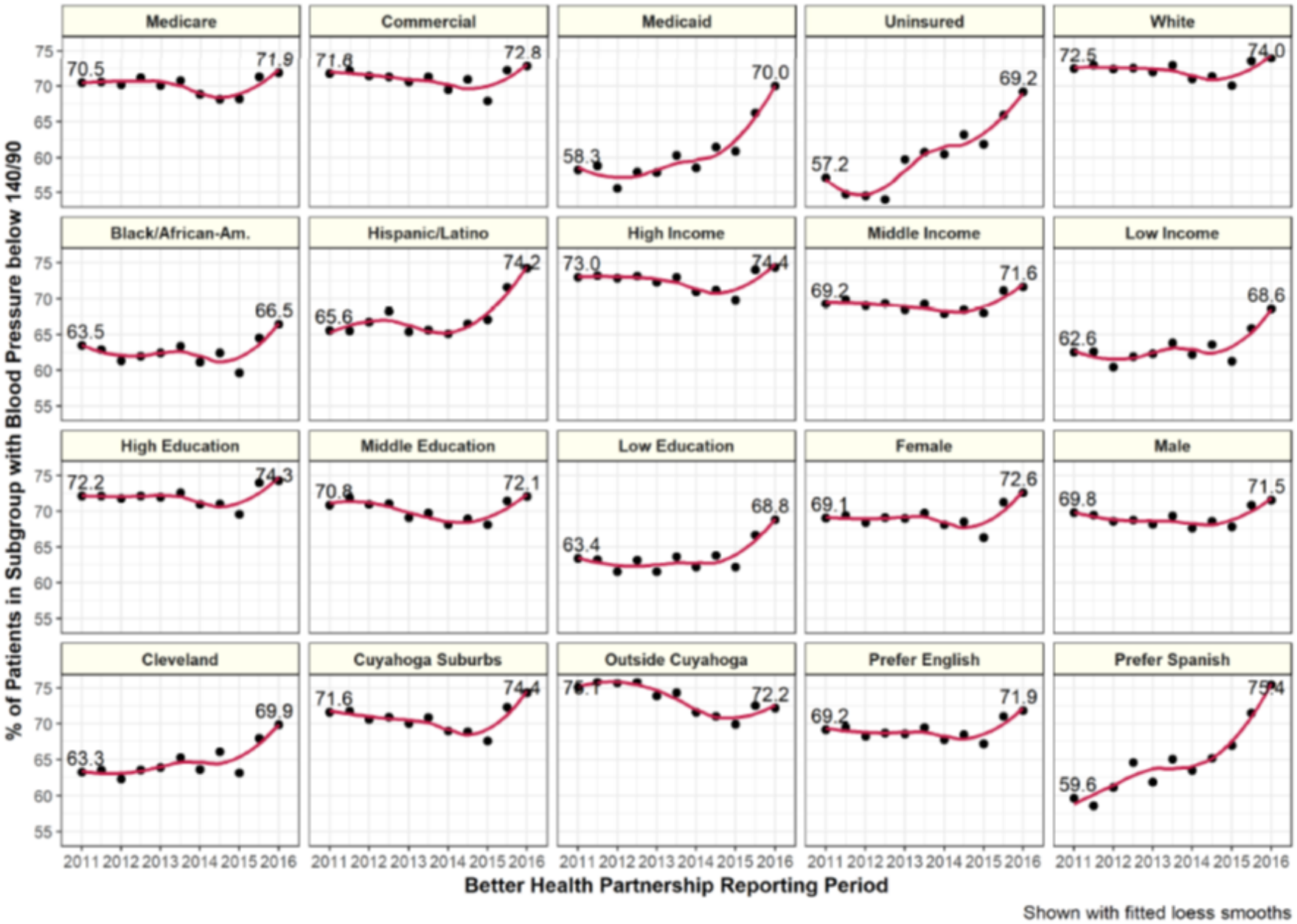


Exhibit E. Trends in Good Blood Pressure Control (< 140/90) by Subgroup, 2011-2016



Shown with fitted loess smooths