Data, Environment and Society

Lecture 7: Visualization

September 19, 2019

Instructor: Duncan Callaway

GSI: Salma Elmallah

Machine learning used to help tell which wildfires will burn out of control

New technique could help authorities conduct triage in multiple-blaze scenarios

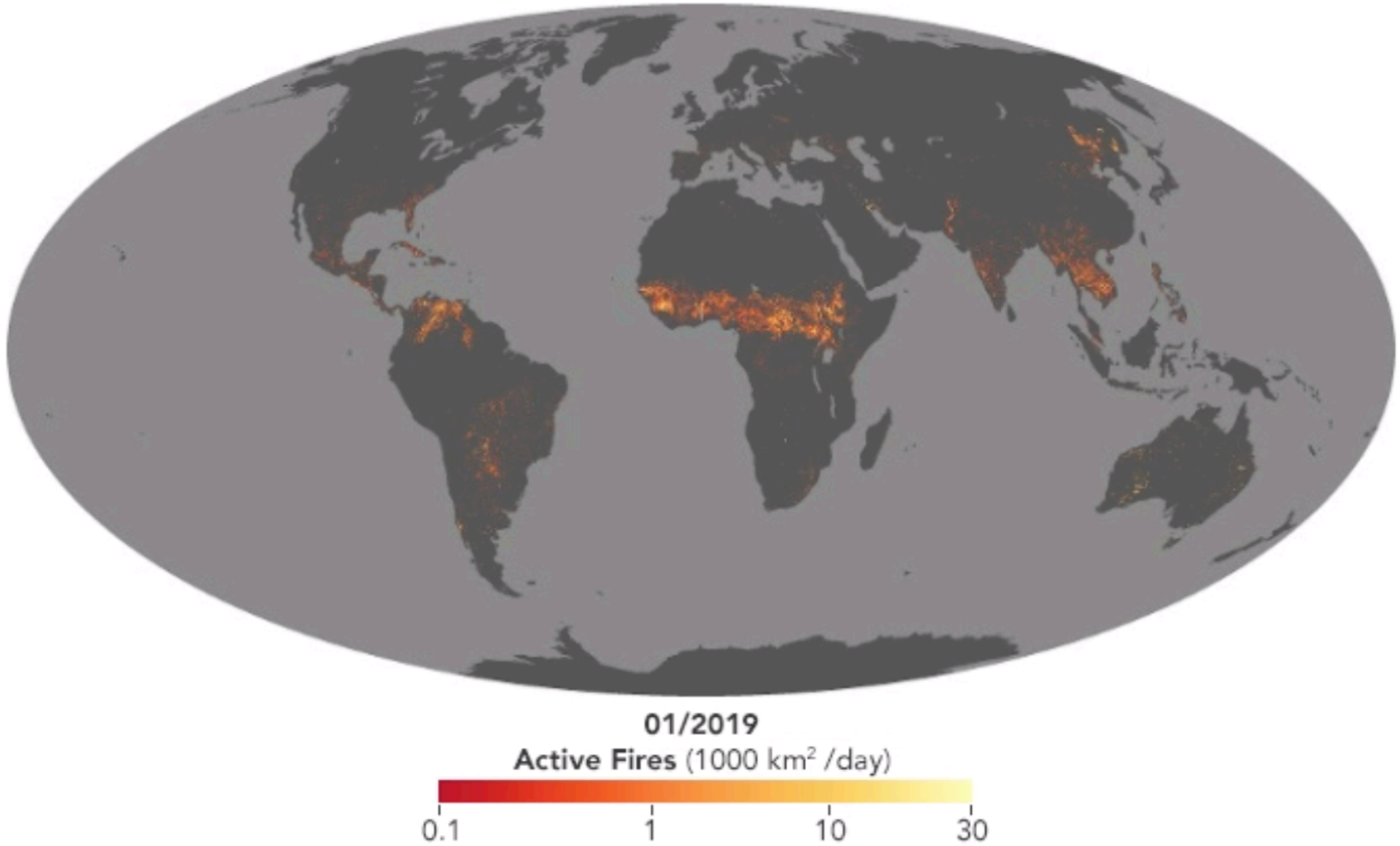
Date: September 17, 2019

Source: University of California - Irvine

Summary: Scientists have developed a new technique for predicting the final size of a wildfire from the moment of ignition.

Building a Long-Term Record of Fire

Editor's Note: Read more about studying Earth's fires with satellites in Part 1. This story was written as part of the 20th anniversary celebration of this website.



Announcements

- Kammen lecture will be moved (not next Thursday, stay tuned)
- HW2 due today
- HW3 released today
- We will be posting polls to get feedback on homework, lecture and labs, stay tuned.

Outline and objectives for today

- Tidy up discussion of EDA and data cleaning from Lecture 6
- Principles of visualization –

  → develop an understanding of what factors to consider as you make a new plot
- Types of plots

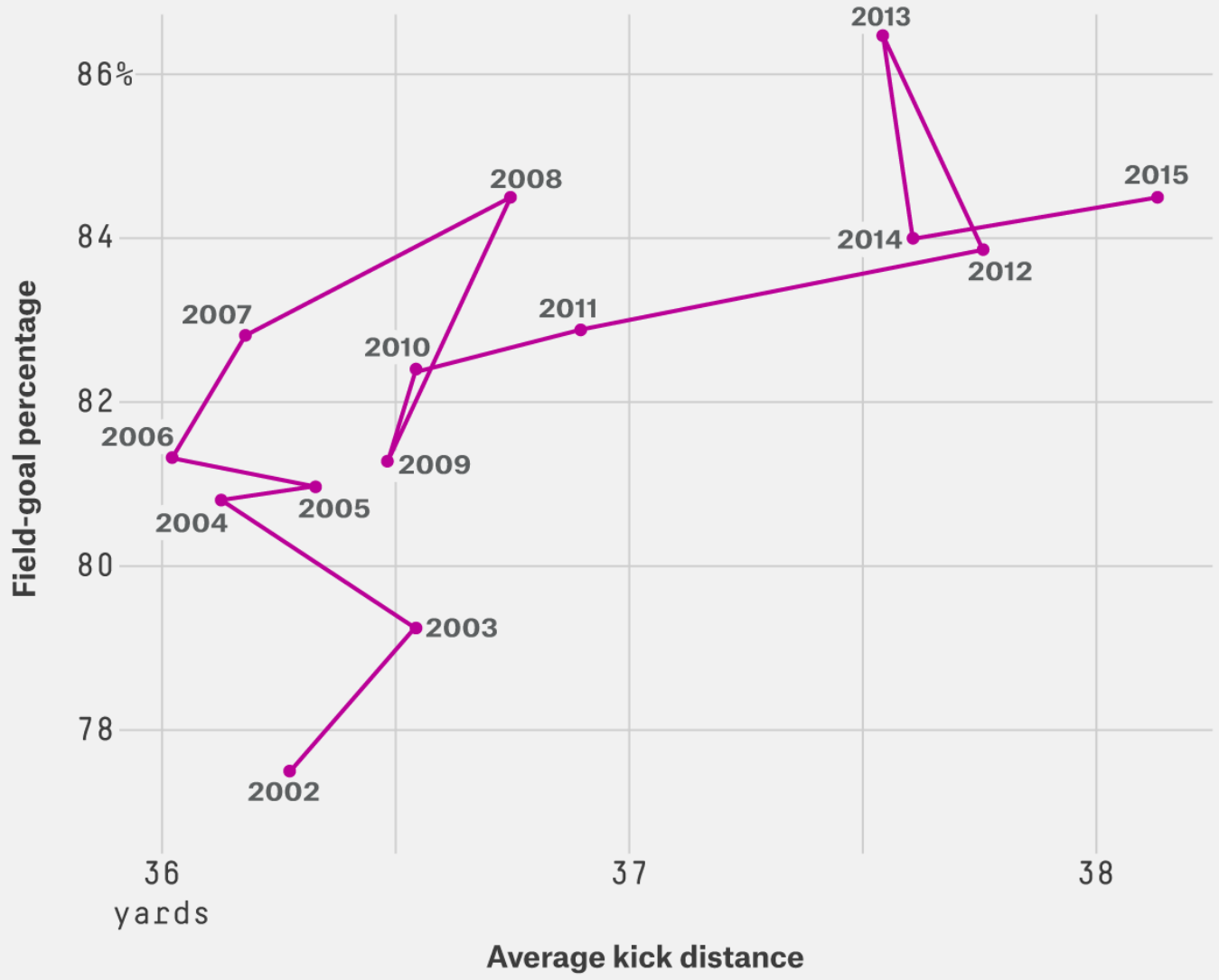
  → building a library of ways to consider plotting your data
- Throughout, (hopefully) have fun looking at cool charts.

Let’s just look at some cool pictures

- What jumps out quickly?
- What did the creator do to make it easier for you to understand their story?
- How many categories of data are displayed in a two dimensional plot?

Kickers are taking longer attempts than ever

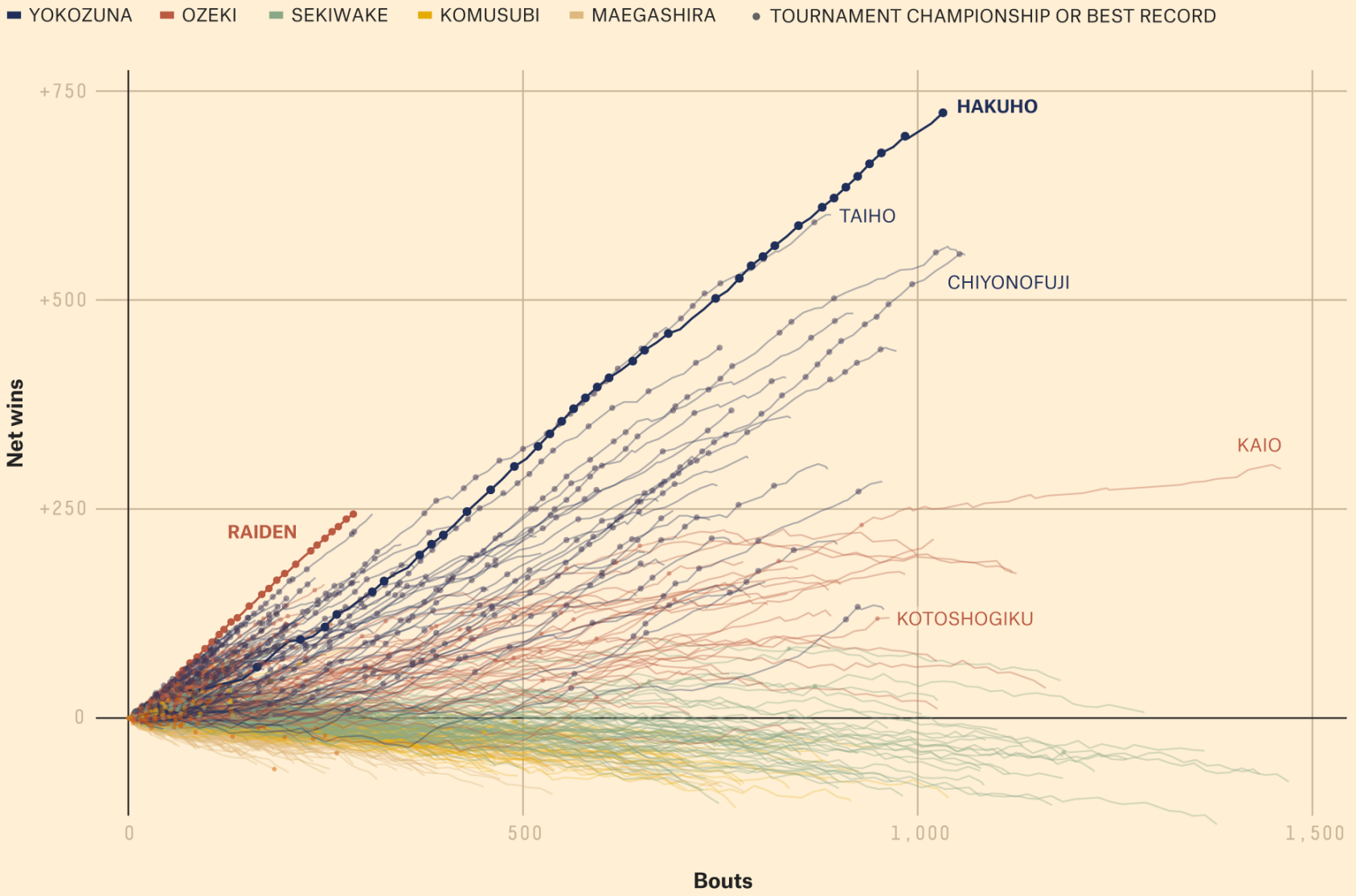
Field-goal percentage for all kicks vs. average kick distance



FIVETHIRTYEIGHT SOURCE: ESPN

The many types of sumo careers

Cumulative net wins vs. total bouts for sumo wrestlers in the top division by rank

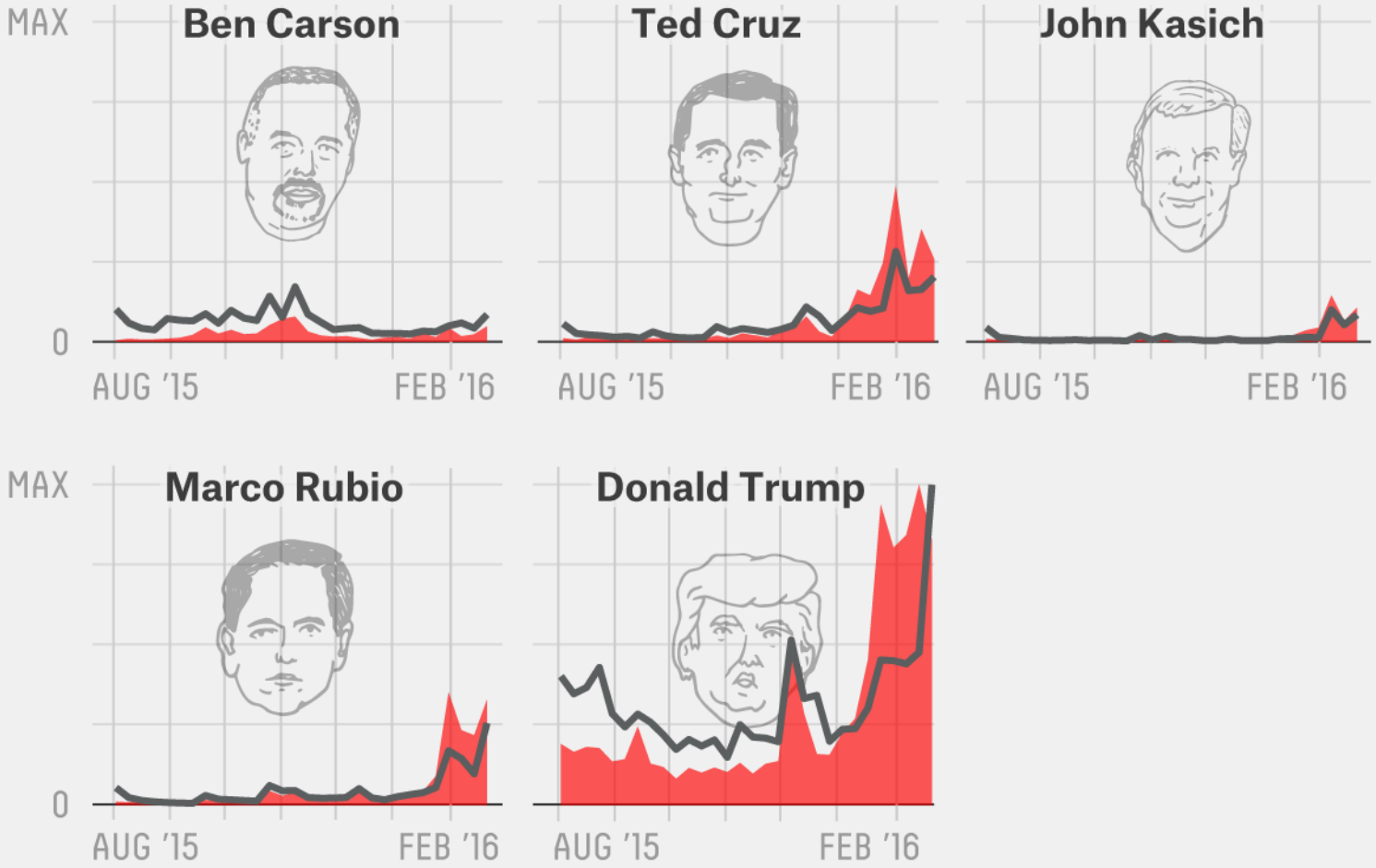


FIVETHIRTYEIGHT

SOURCE: SUMO REFERENCE

Trump continues to dominate both news coverage and Google searches

Relative Google searches and articles appearing on Google News through the week of Feb. 27, 2016

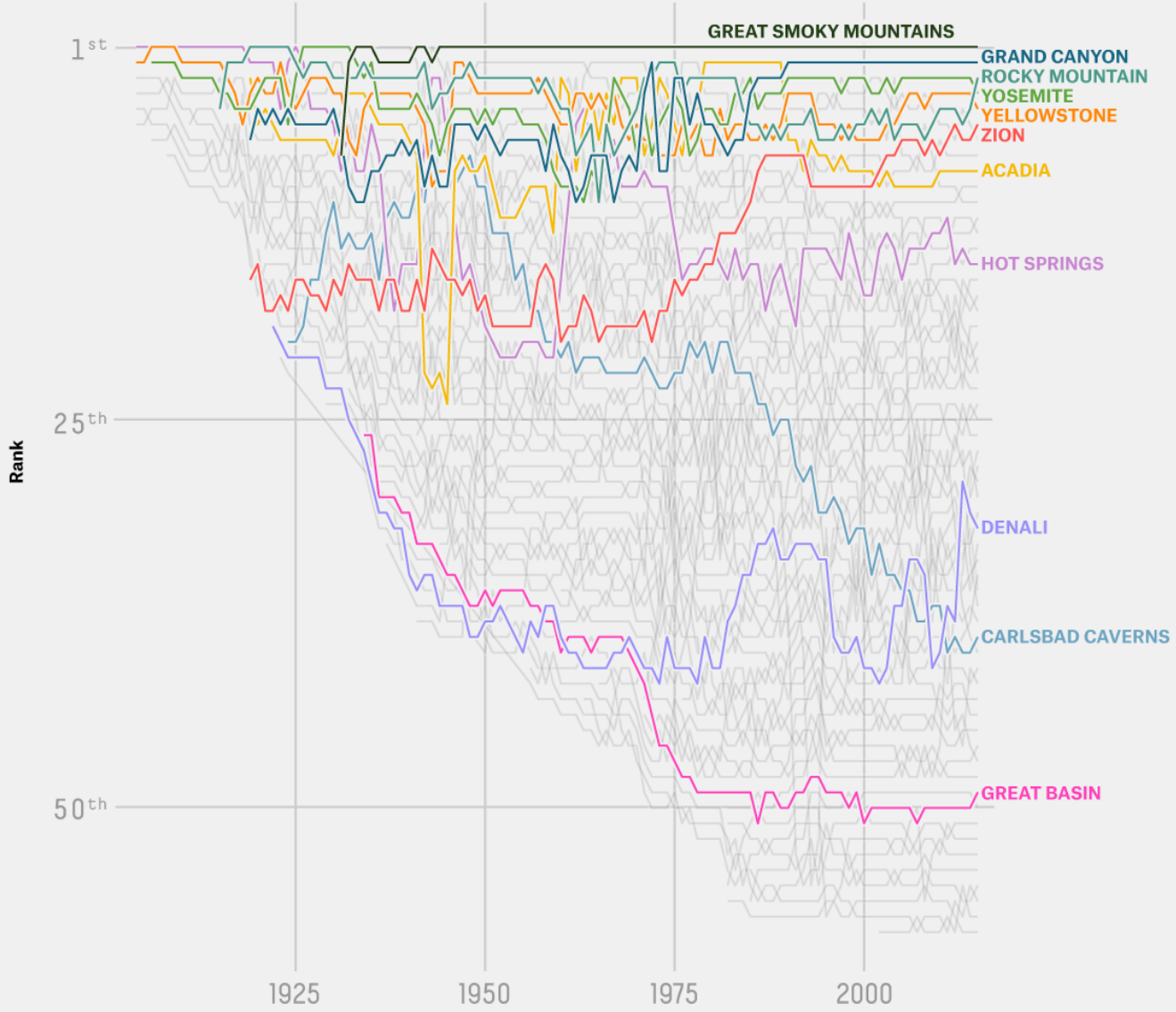


FIVETHIRTYEIGHT

SOURCE: GOOGLE TRENDS

The most popular national parks

National parks ranked by number of visitors in a given year

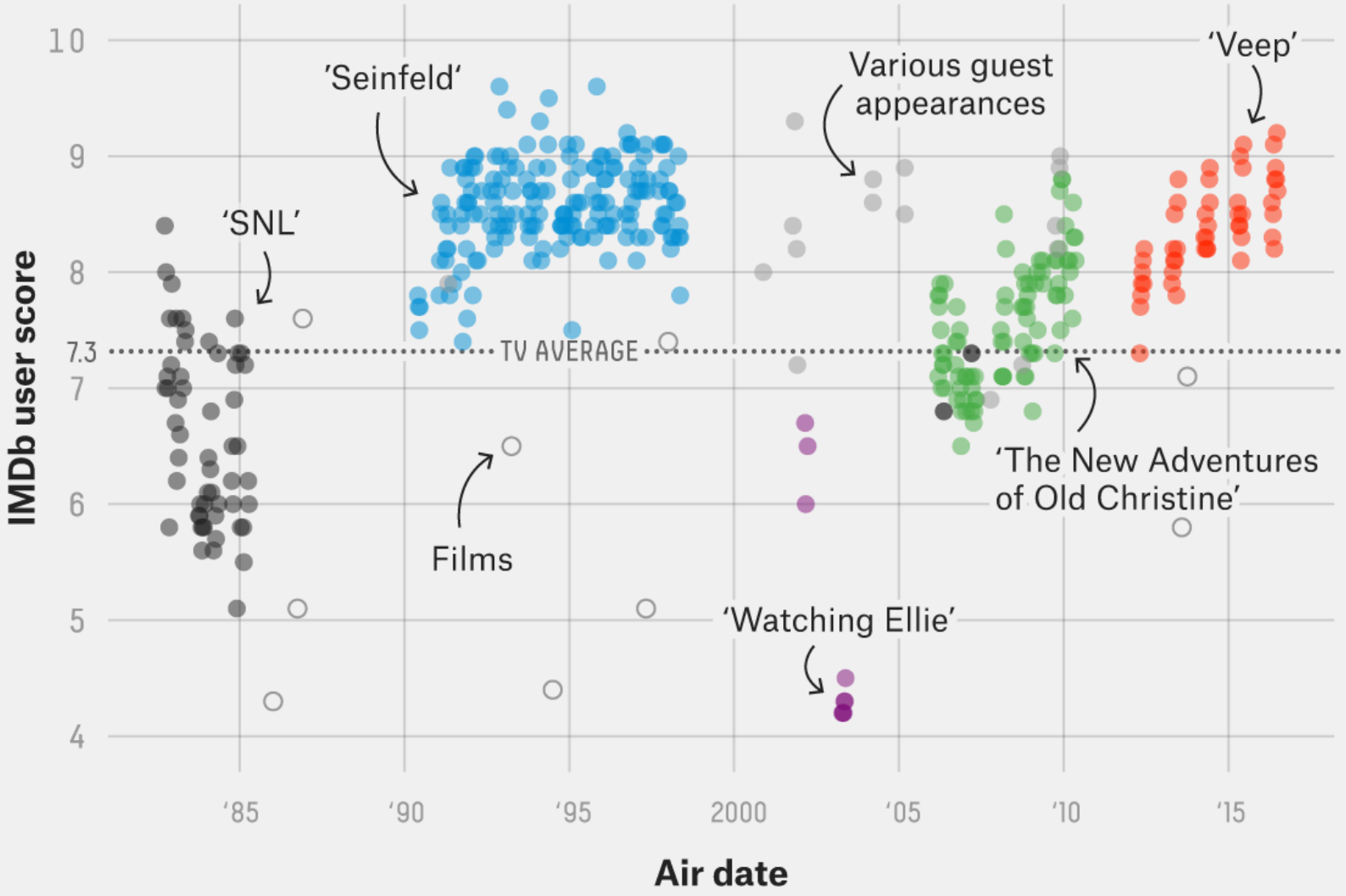


FIVETHIRTYEIGHT

SOURCE: NATIONAL PARK SERVICE

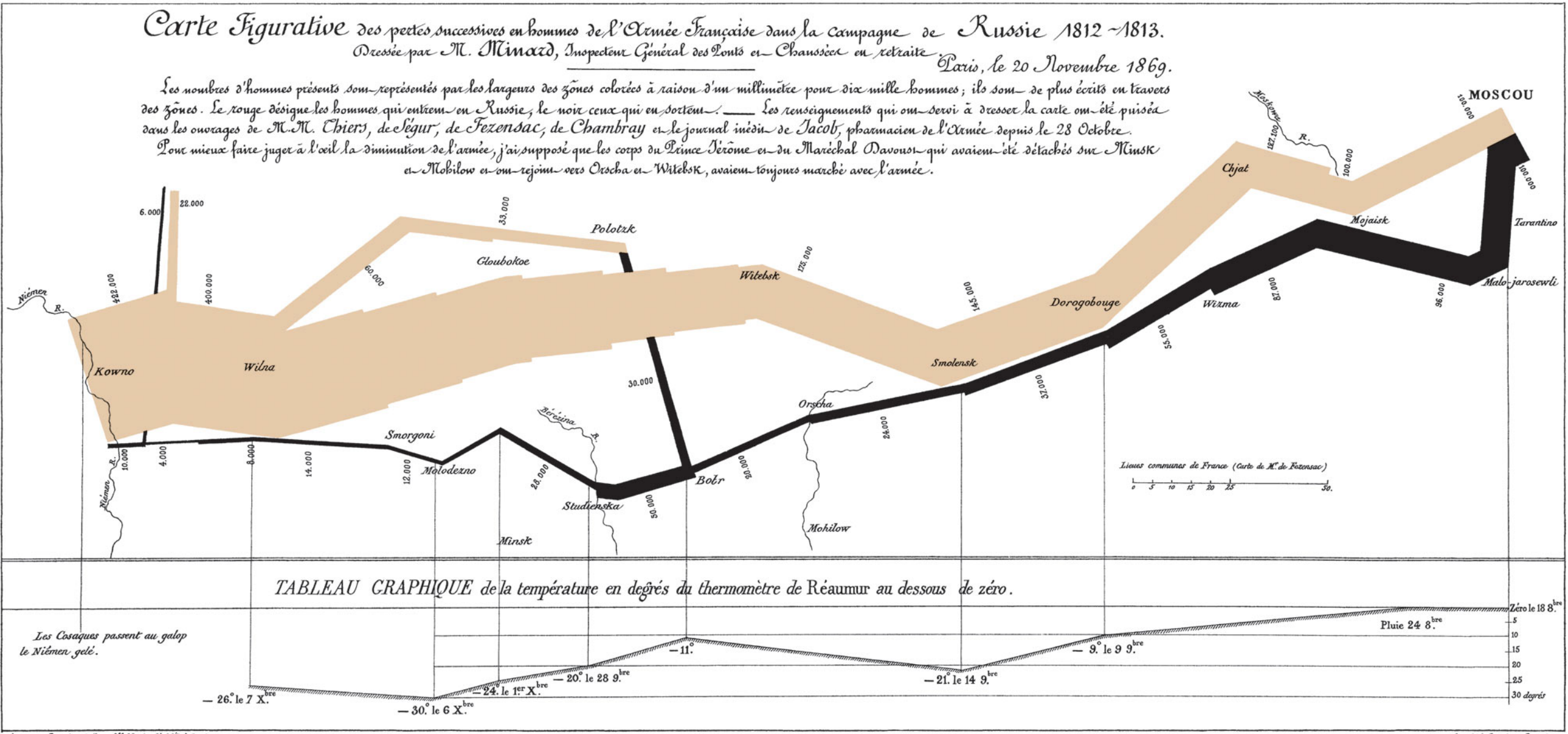
Julia Louis-Dreyfus is good at almost everything

IMDb ratings for appearances by Louis-Dreyfus



FIVETHIRTYEIGHT

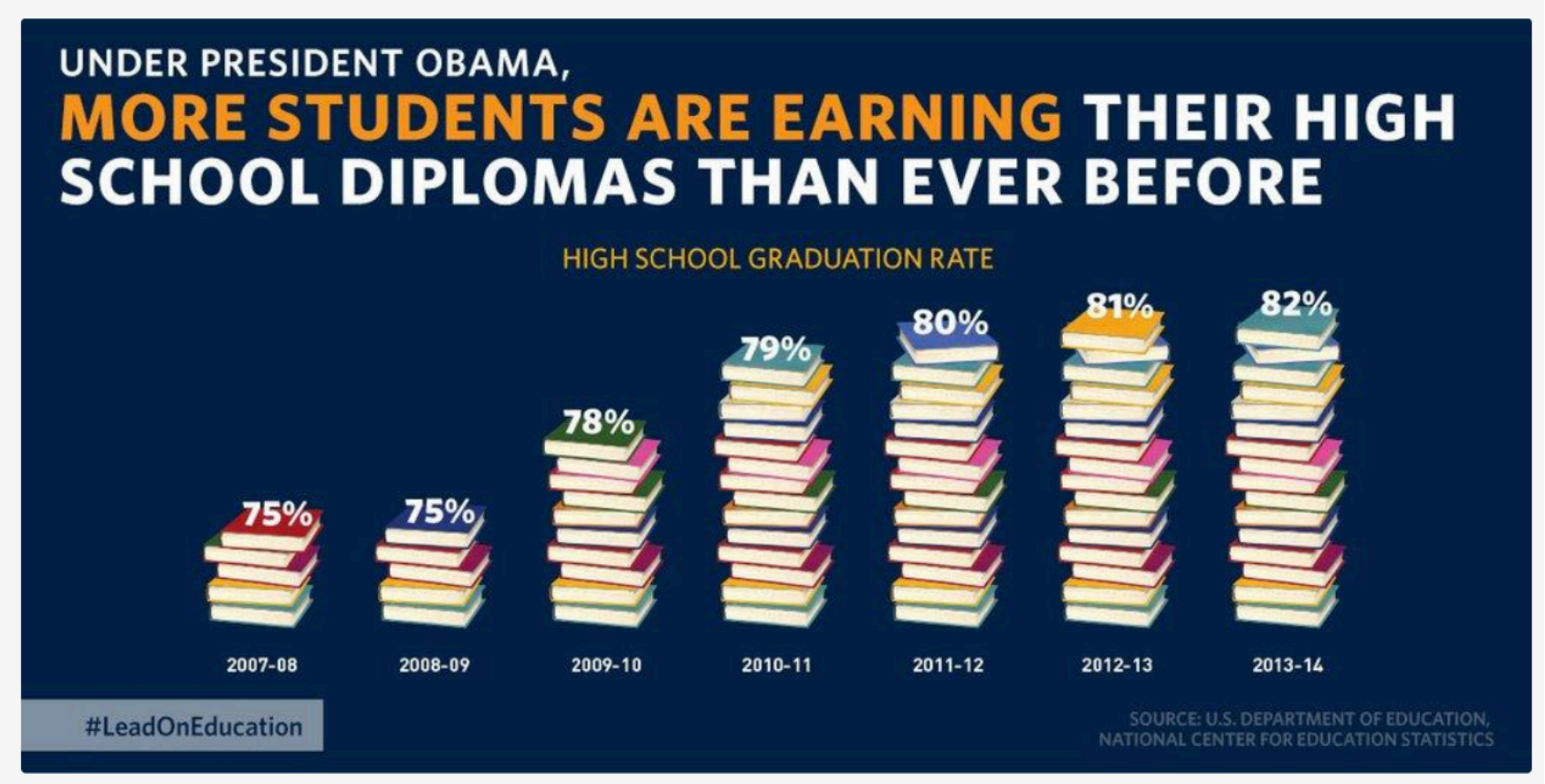
SOURCE: IMDB


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'oeil 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 en ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
Niemen R.
Kowno
Wilna
Smorgoni
Molodezno
Minsk
Studienska
Bérézina R.
Bobr
Orscha
Mohilow
Polotzk
Gloubokoe
Witebsk
Chjat
Dorogobouge
Smolensk
Wixma
Mojaisk
Moskowa R.
MOSCOU
Tarantino
Malo-jarosewli
Lieues communes de France (Carte de M.r de Fezensac)
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
— 26.° le 7 X.bre
— 30.° le 6 X.bre
— 24.° le 1.er X.bre
— 20.° le 28 9.bre
— 11.°
— 21.° le 14 9.bre
— 9.° le 9 9.bre
Pluie 24 8.bre
Zéro le 18 8.bre
5
10
15
20
25
30 degrés
Autog. par Regnier, 8. Pas. Ste Marie St Gain à Paris.
Imp. Lith. Regnier et Dourdet.

Principles of visualization

- Scale
- Conditioning
- Perception
- Transformation
- Context
- Smoothing

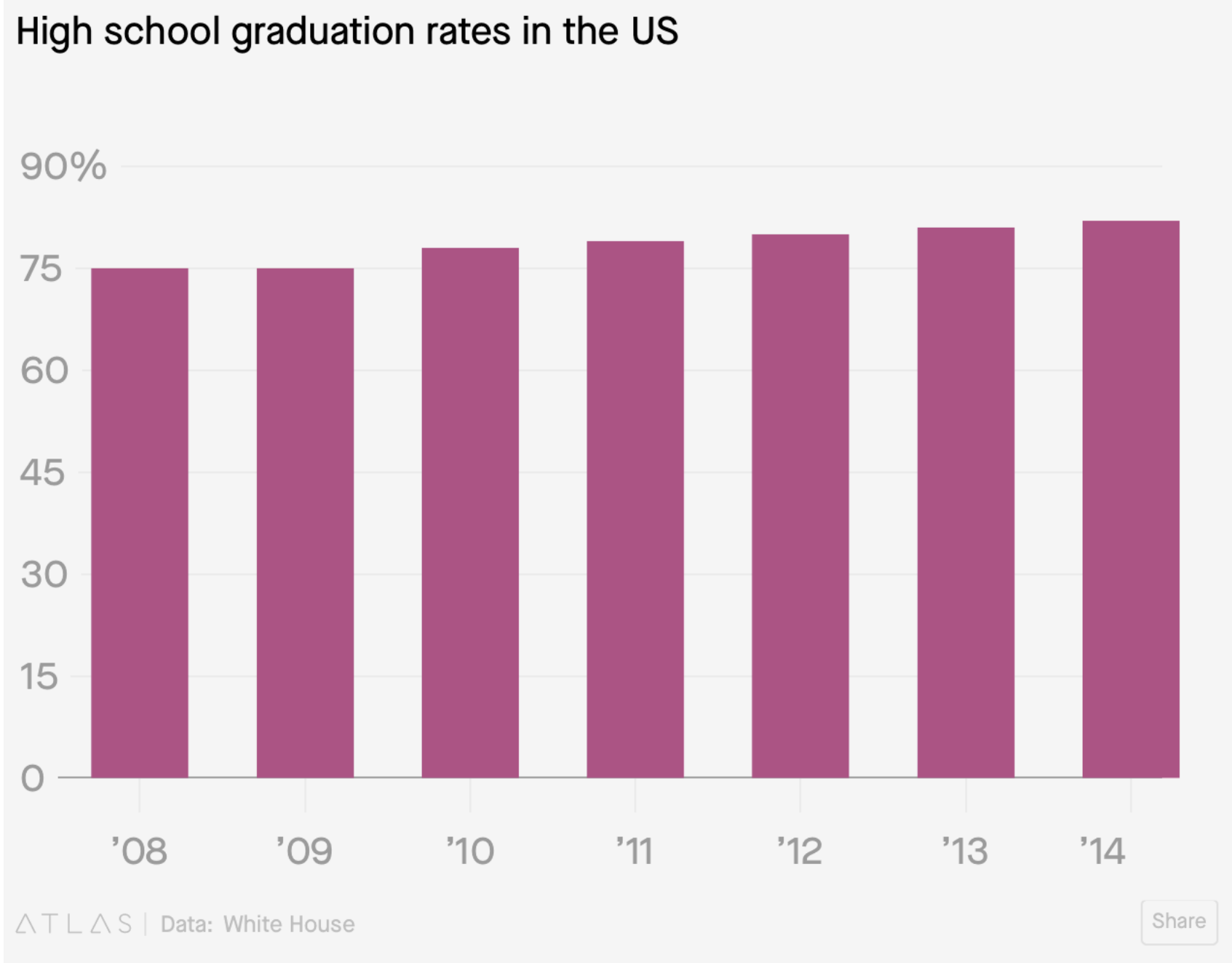
Scale



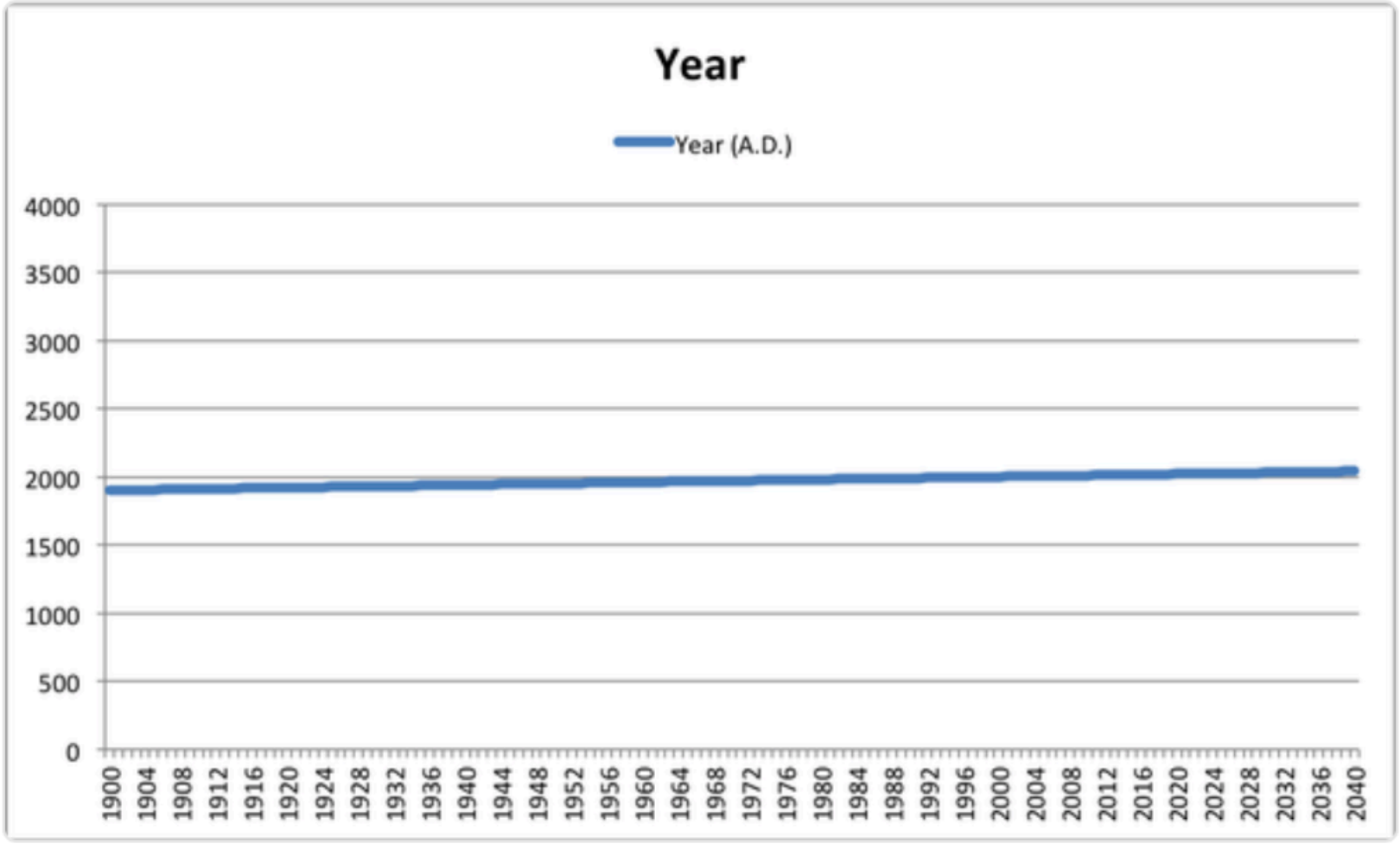
- How would you fix this chart?

Scale, ctd

- Fixed?
  - Not necessarily – though y-axis starts at zero (often a good thing) what we really want to know is how things are changing over time.



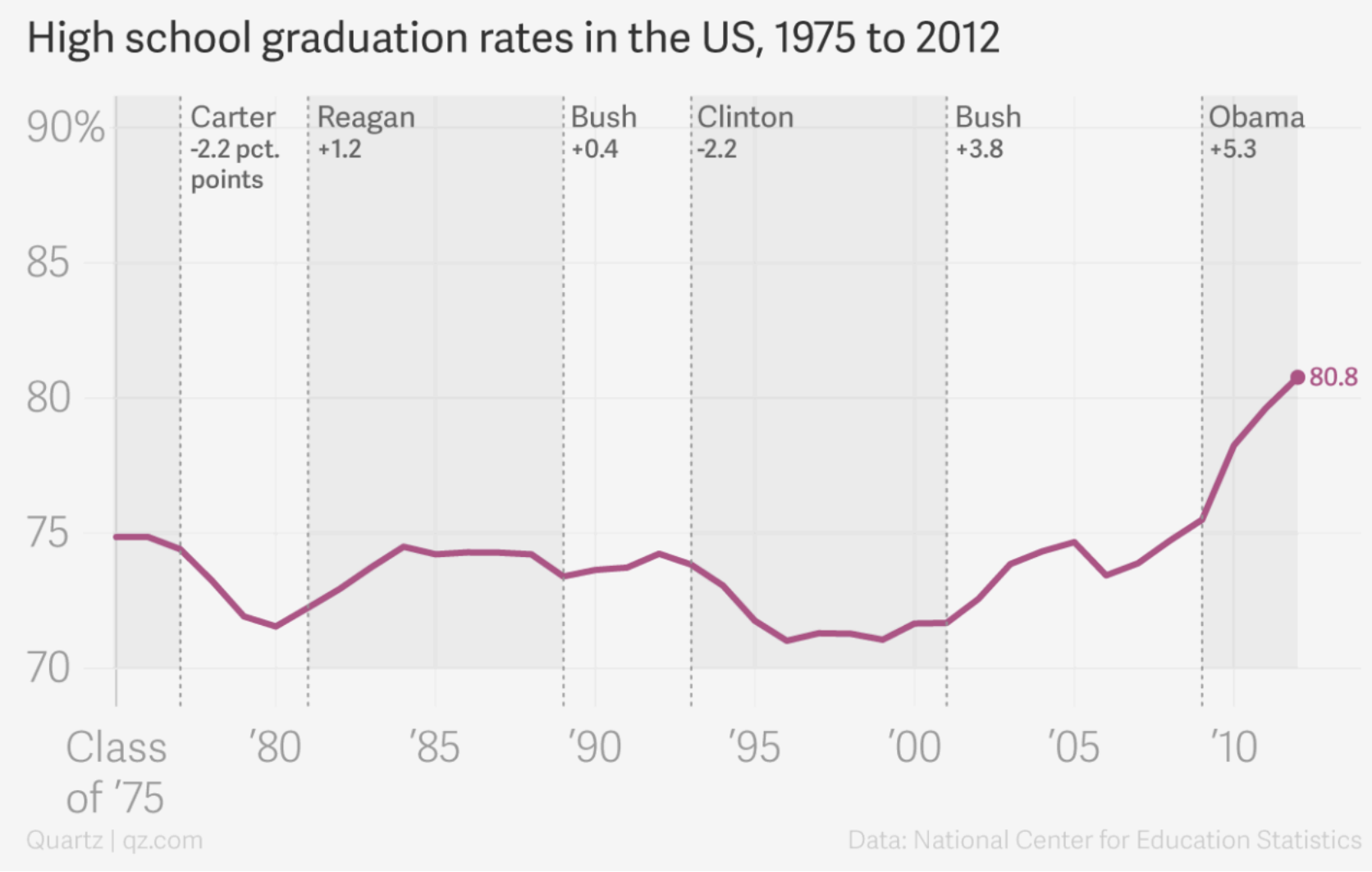
Scale: we shouldn't *always* start at zero…



Scale, ctd

- Education plot fixed?
  - We see comparisons to other administrations
  - We see changes over a longer period of time

(note, these data are from a slightly different data set than the data in the first two graduation charts)



Principles of visualization

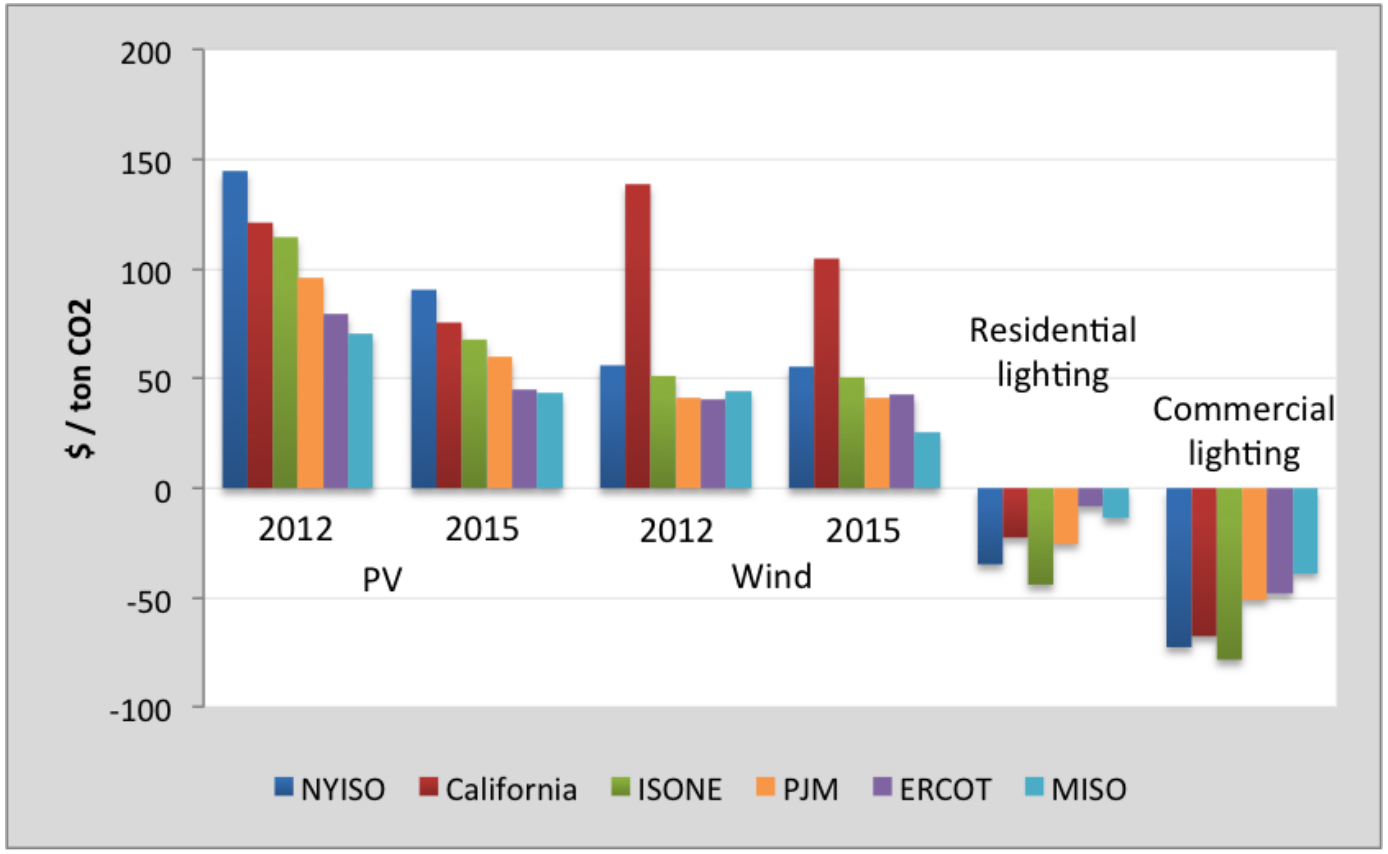
- Scale
- Conditioning
- Perception
- Transformation
- Context
- Smoothing

Conditioning

- What does conditioning mean, in the context of visualization?
  - To subgroup the data to see relationships between different groups of interest

Conditioning example

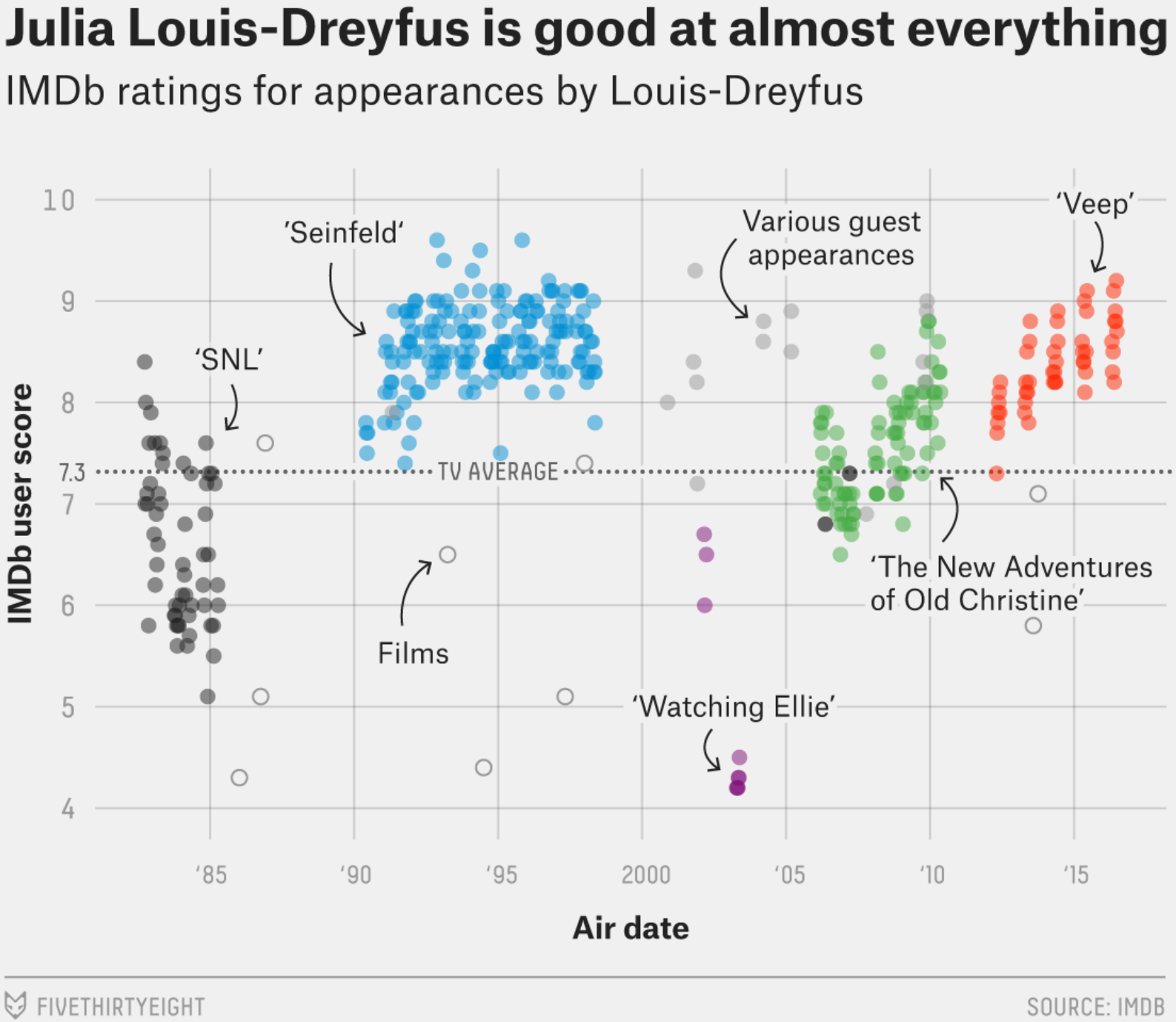
Figure 7: Marginal abatement costs across technologies and regions



Notes: This figure plots the total marginal abatement cost, in dollars per ton of carbon dioxide, from different region-technology combinations. Regions are ordered from highest to lowest marginal abatement cost for PV.

Taken from Callaway, Fowlie and McCormick JAERE (2018)

Revisiting Julia



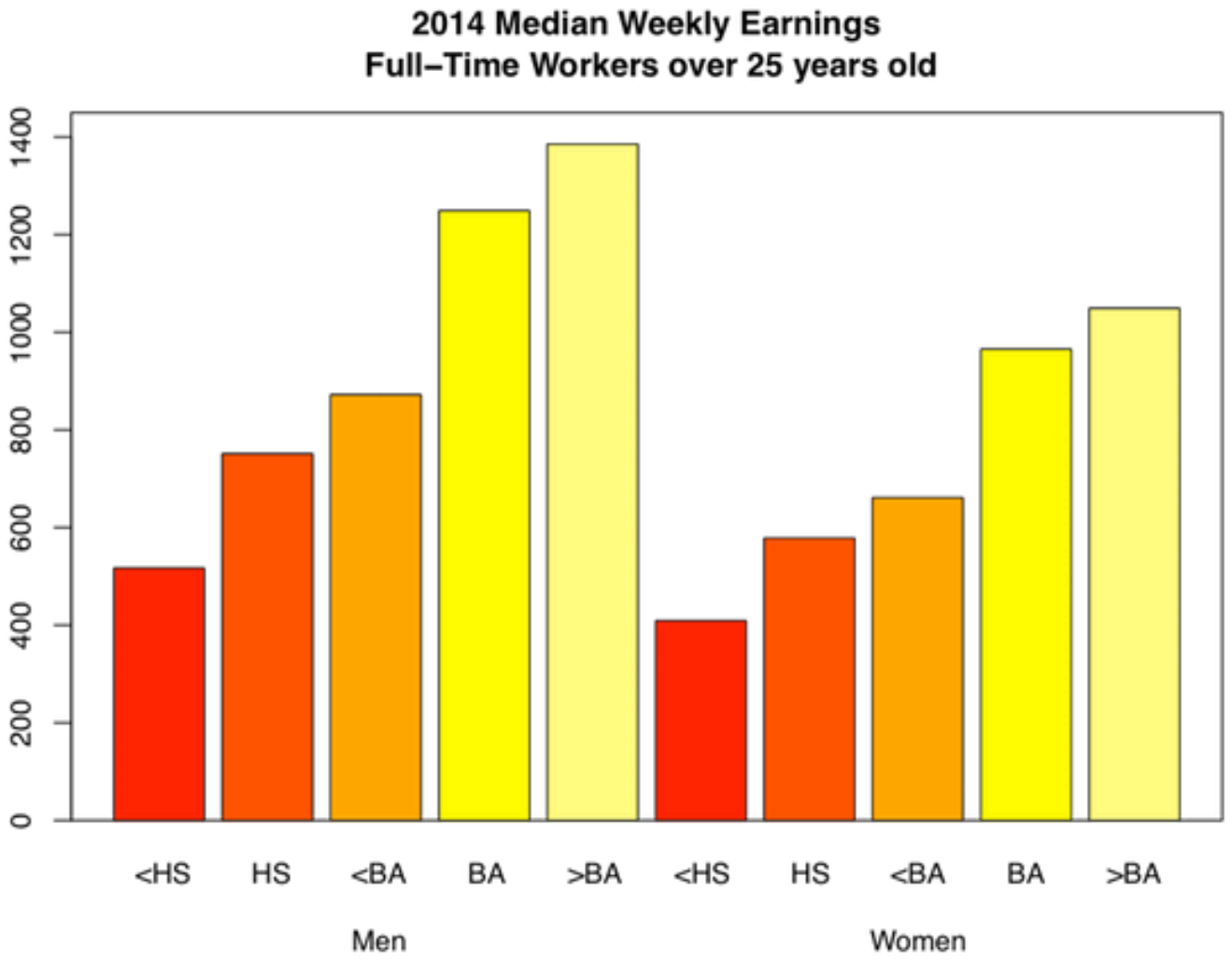
Principles of visualization

- Scale
- Conditioning
- Perception
- Transformation
- Context
- Smoothing

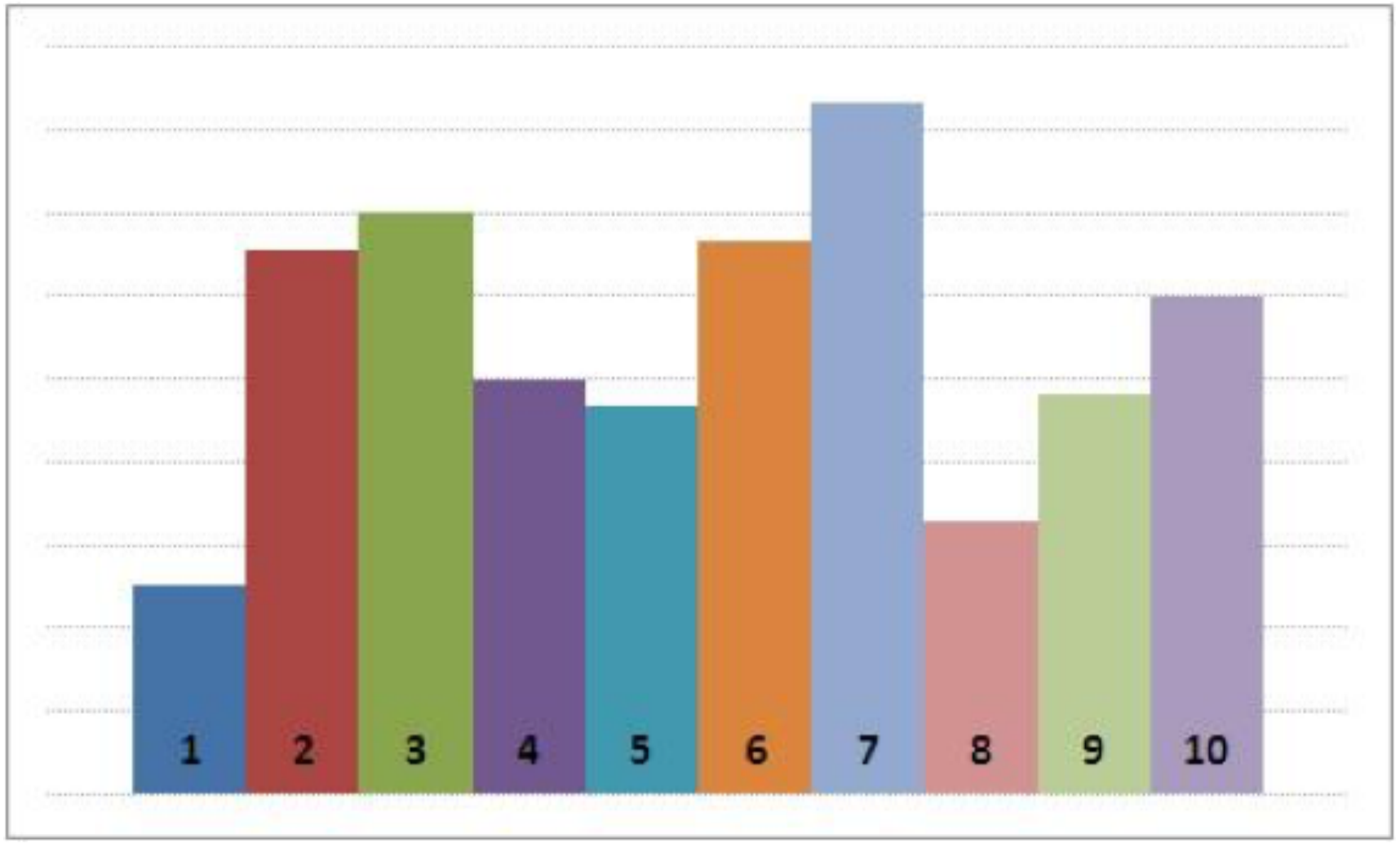
Perception

- Some principles:
  - We perceive some colors more strongly than others (especially greens)
  - We perceive lighter shaded areas as larger than darker shaded ones.
- An under-appreciated issue: color-blindness.

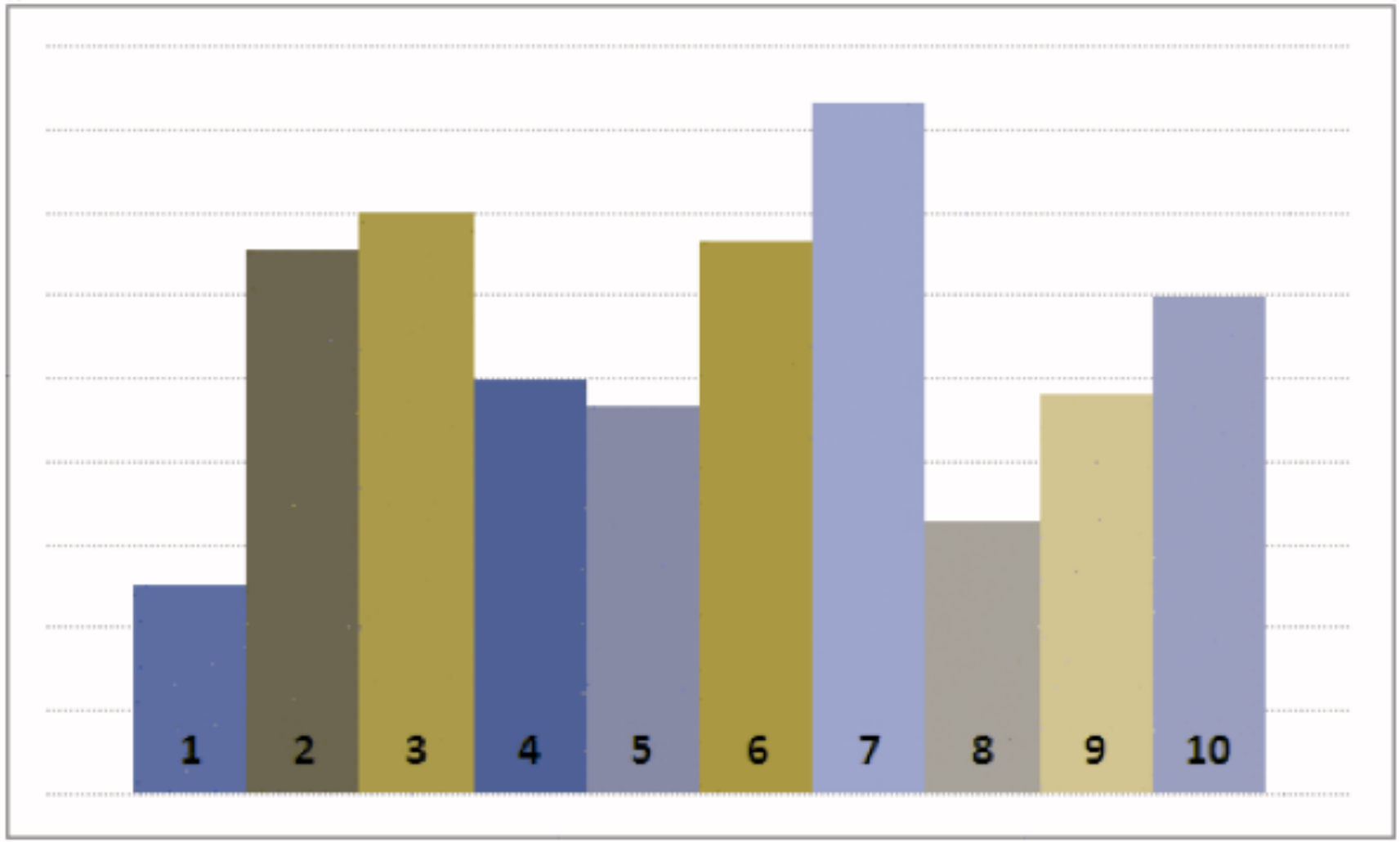
What might you change to adjust perception?



Not color blind

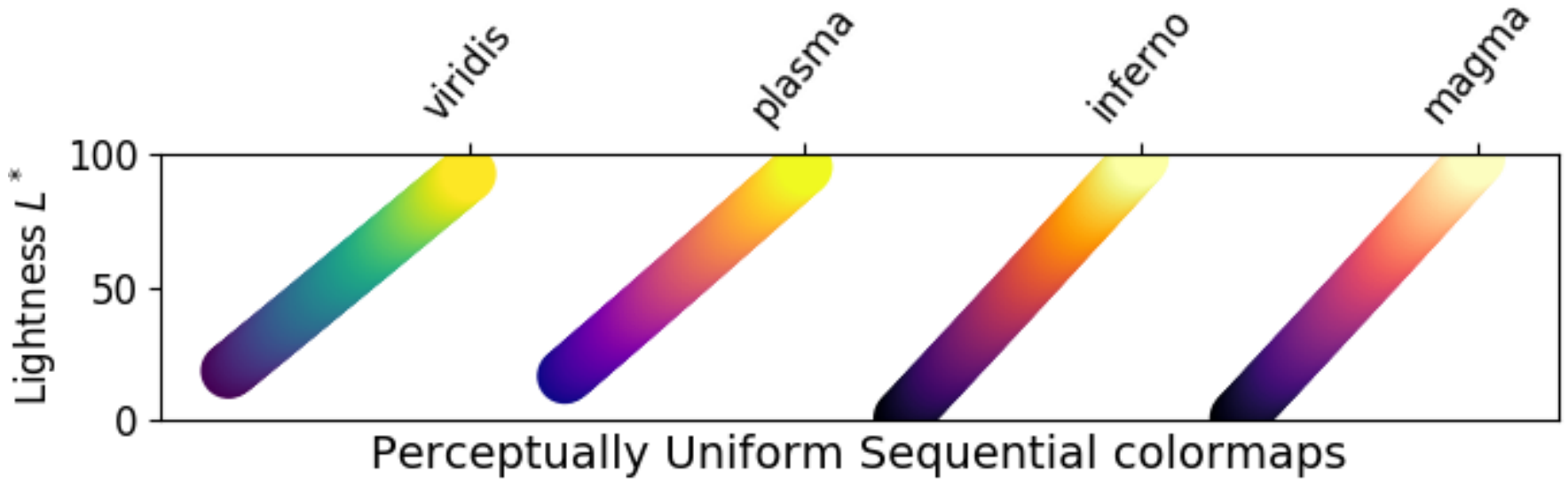


Color blind

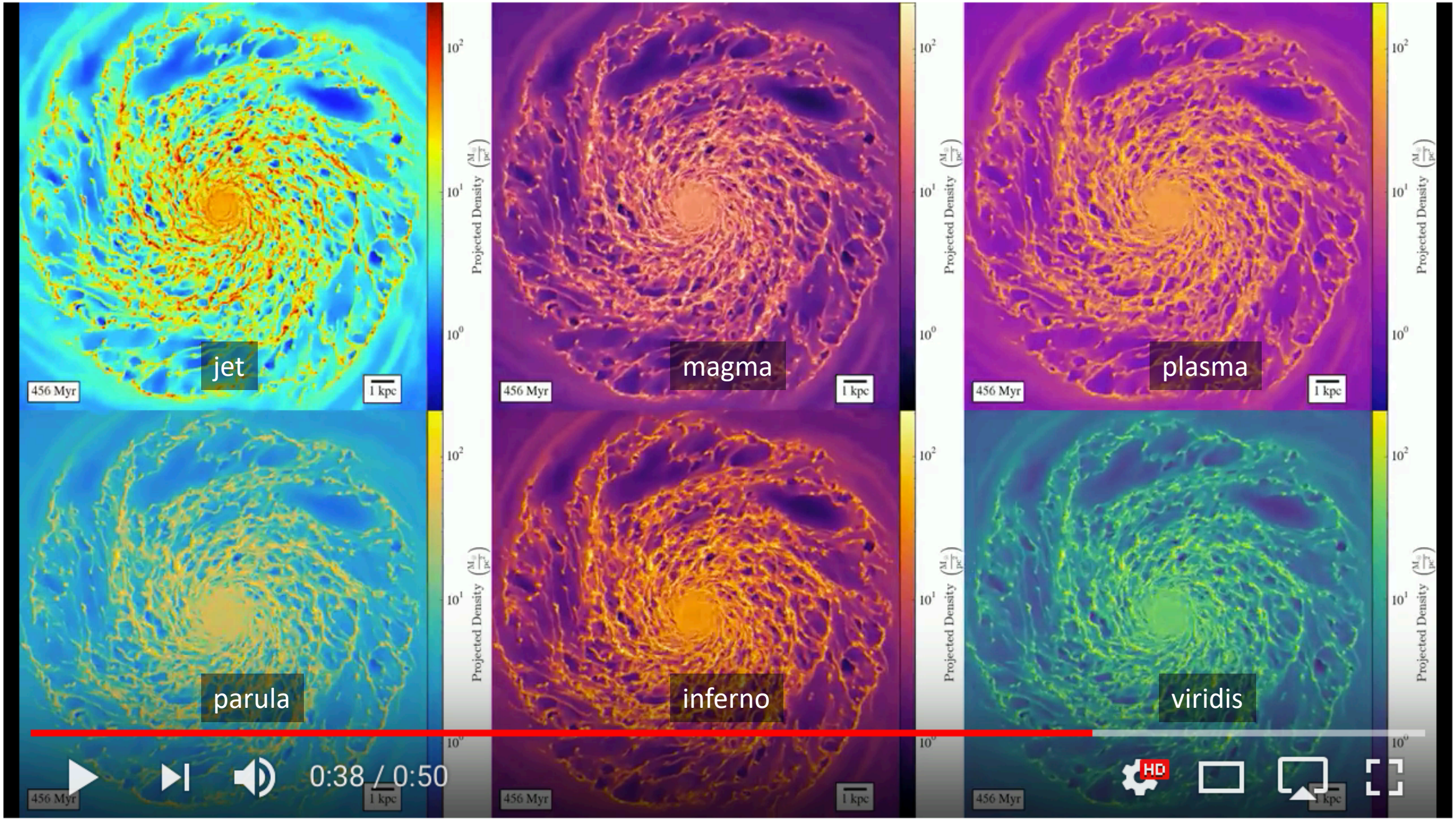


Perceptually uniform Color palettes – quantitative data

- Perceptually uniform: equal steps in data are perceived as equal steps in the color space
- A few perceptually uniform matplotlib color maps



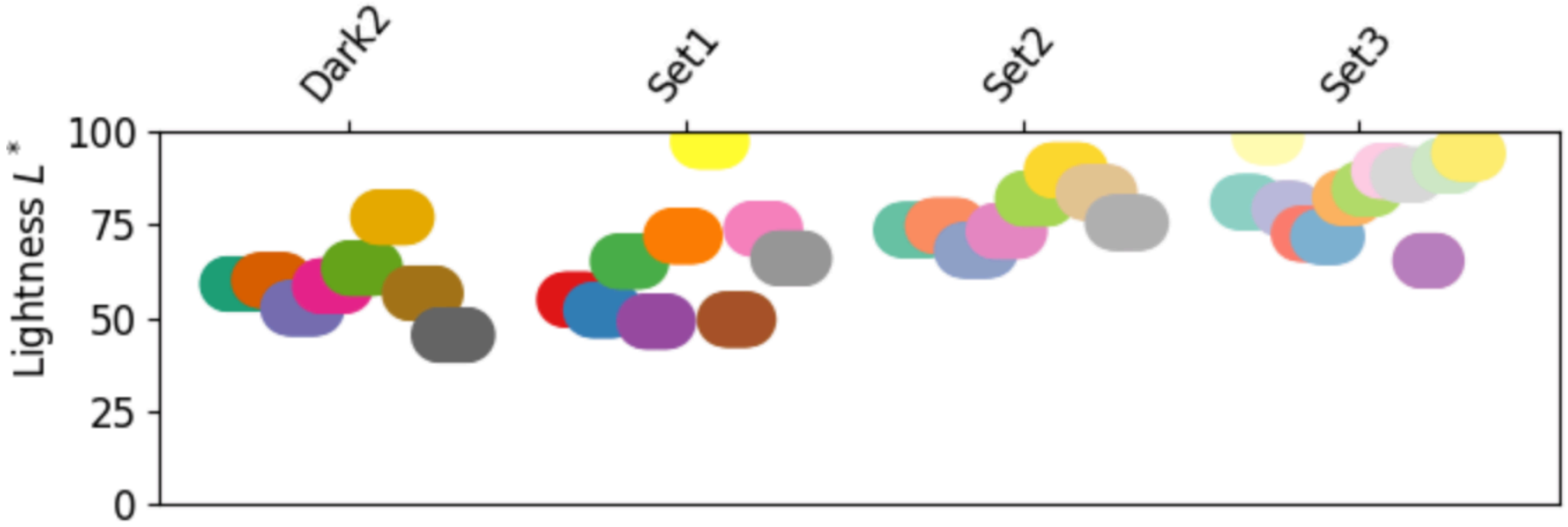
- Good when you want to represent a quantity in space



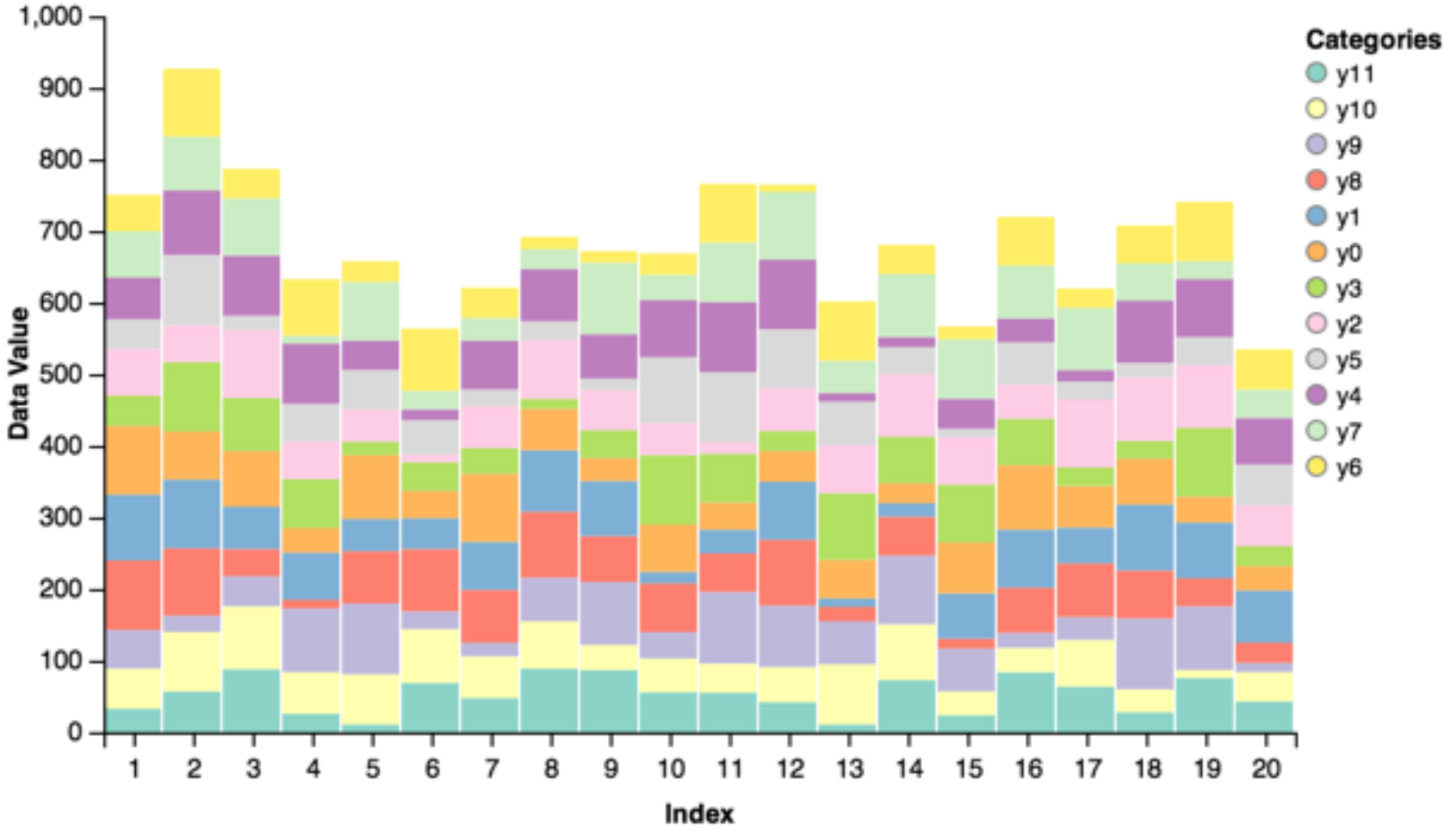
Nathan Goldbaum's galaxy formation simulation video

Qualitative / categorical colormaps

- Often used for representing quantity in an area or line – colors distinguish between different *categories* rather than quantities.
- A few matplotlib quantitative colormaps:



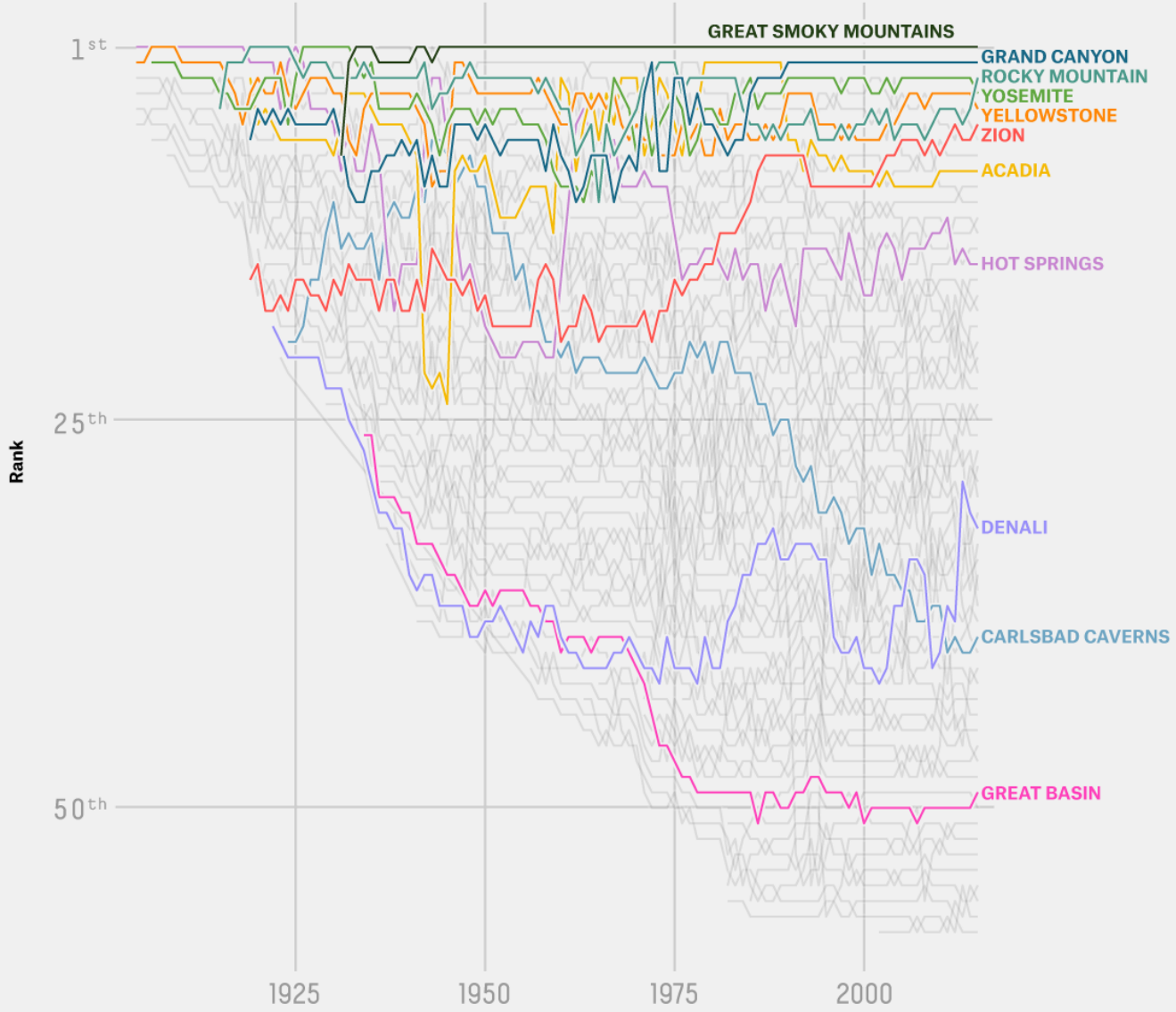
Example: Set3 palette for categorical data



https://vincent.readthedocs.io/en/latest/charts_library.html

The most popular national parks

National parks ranked by number of visitors in a given year



FIVETHIRTYEIGHT

SOURCE: NATIONAL PARK SERVICE

Colormap summary

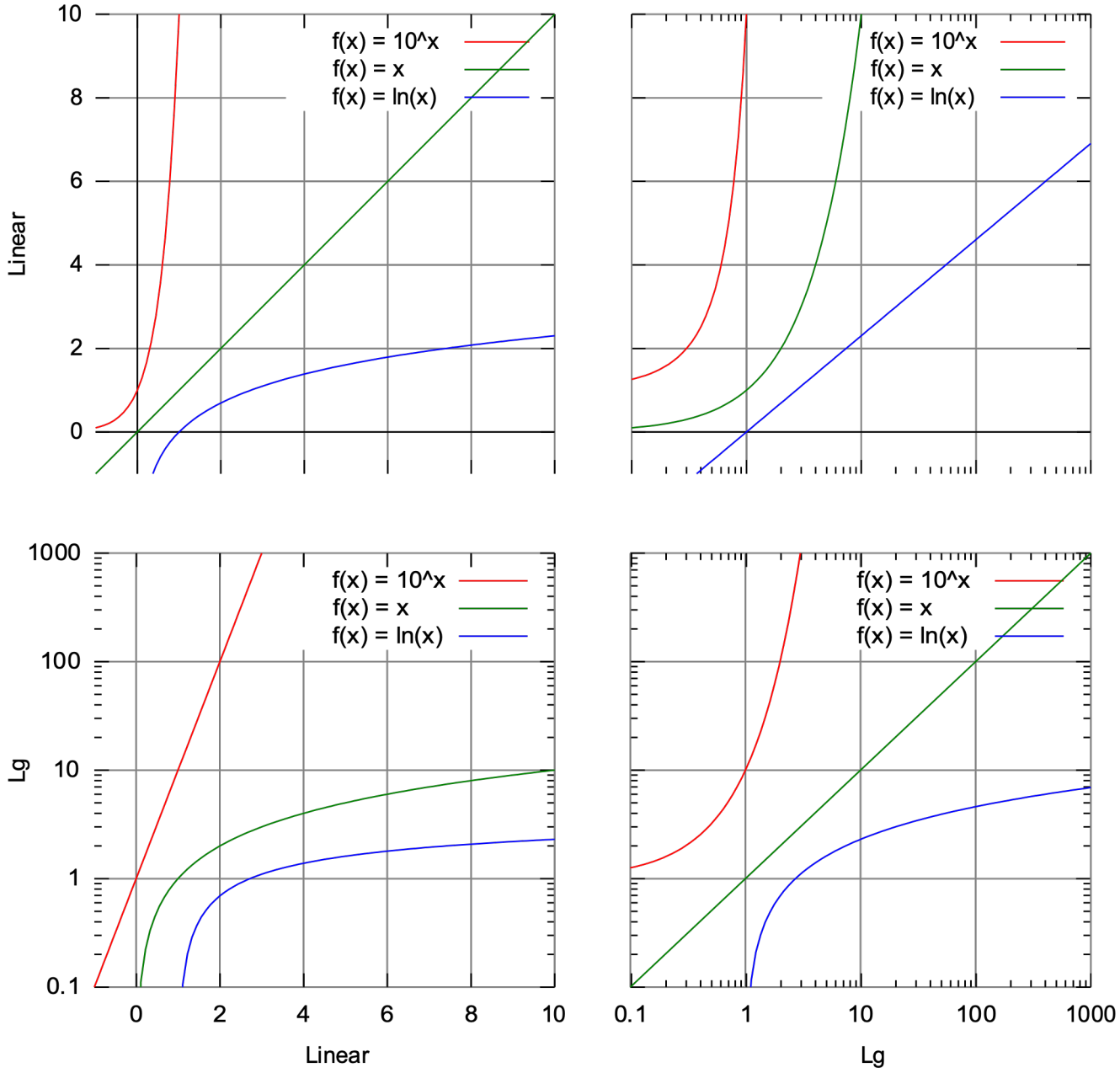
- Consider for displaying quantitative variables like temperature, concentration, etc
  - plasma, 
  - viridis, 
  - inferno, 
  - magma 
- Consider for displaying categorical data
  - set1, 
  - set2, 
  - set3, 
  - dark2 

Principles of visualization

- Scale
- Conditioning
- Perception
- Transformation
- Context
- Smoothing

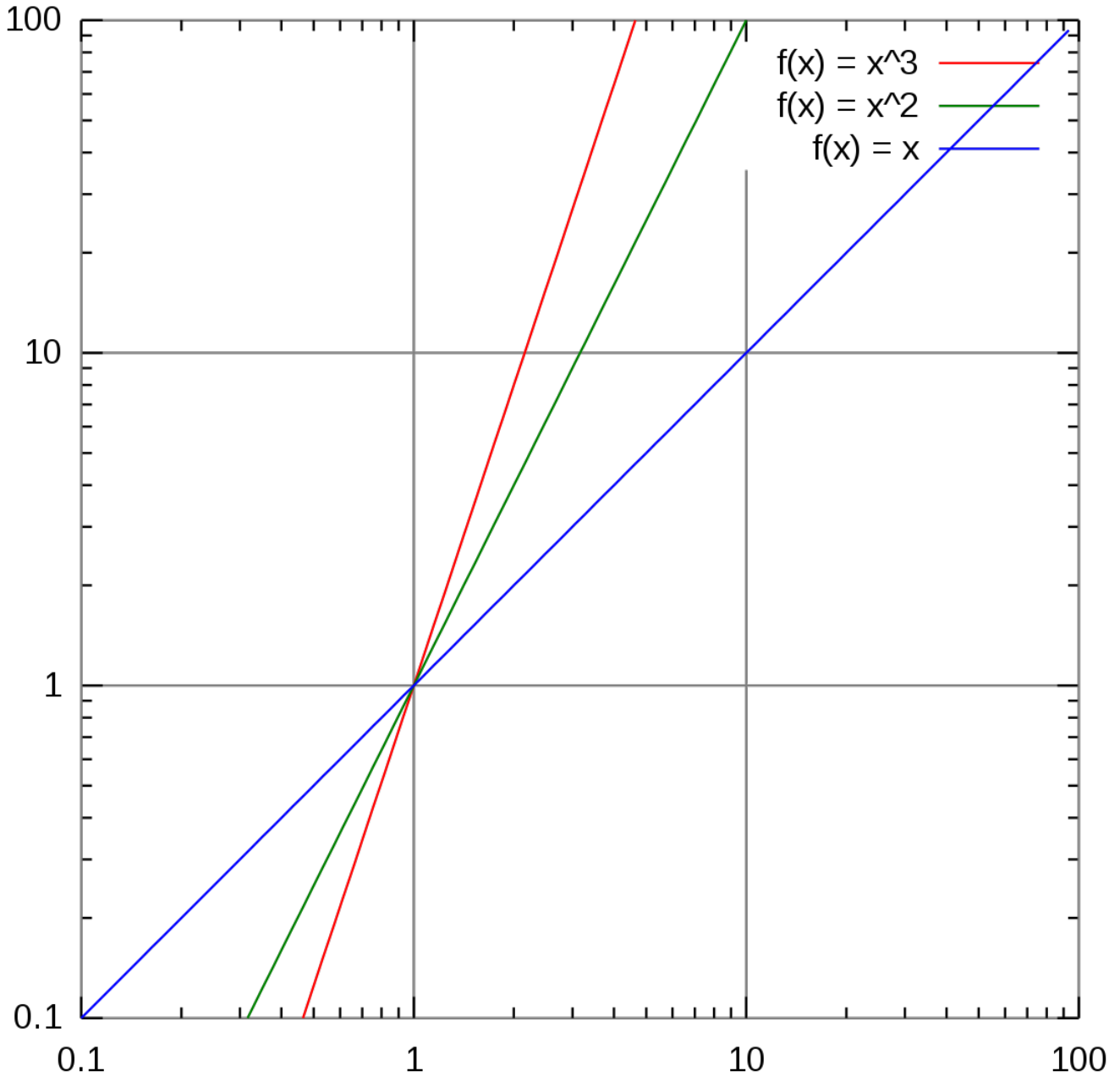
Transformation

- Log transform is the “swiss army knife” transformation
- Squashes down big numbers; makes it easier to see small numbers on the same axis.



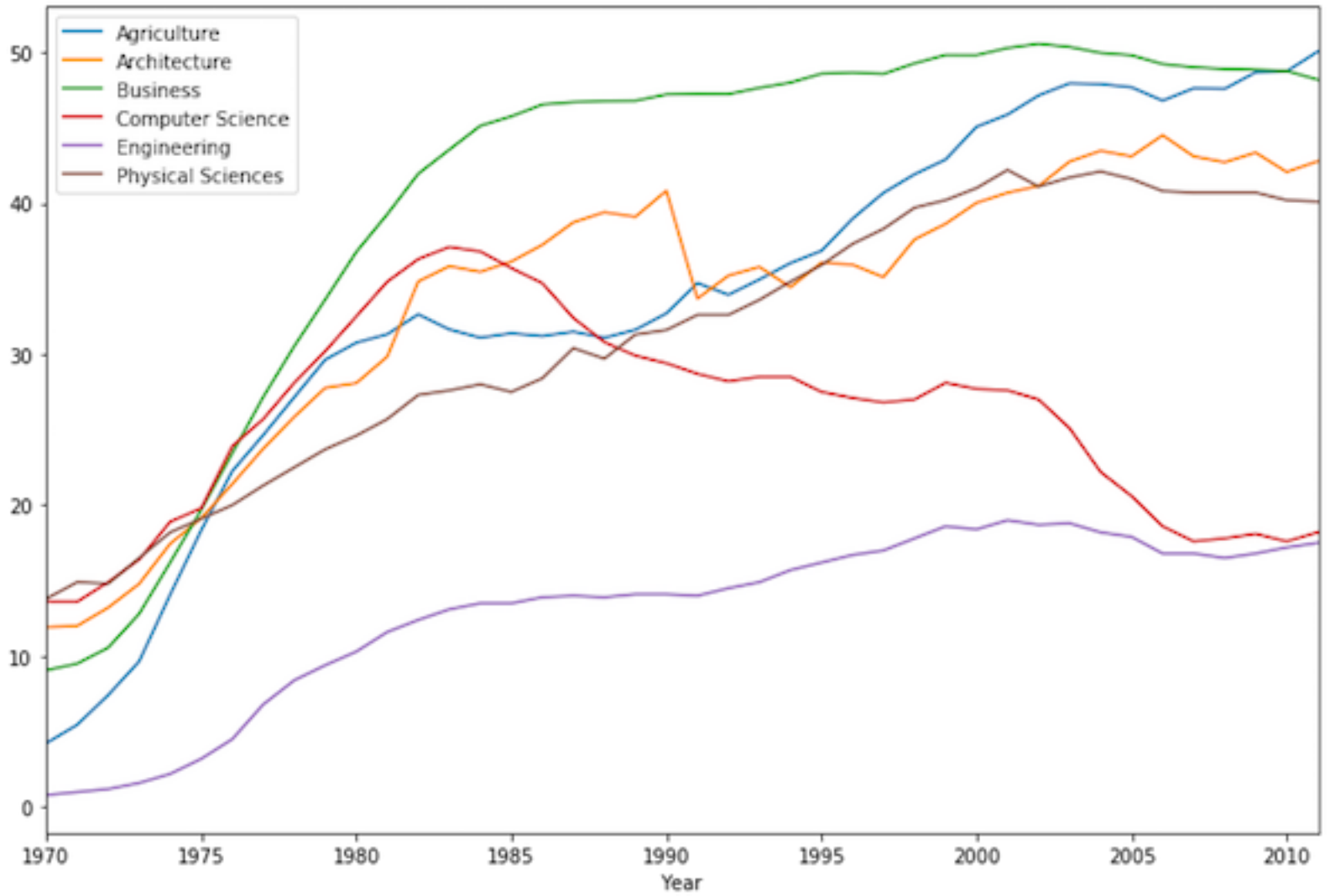
By Autopilot - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=10733854

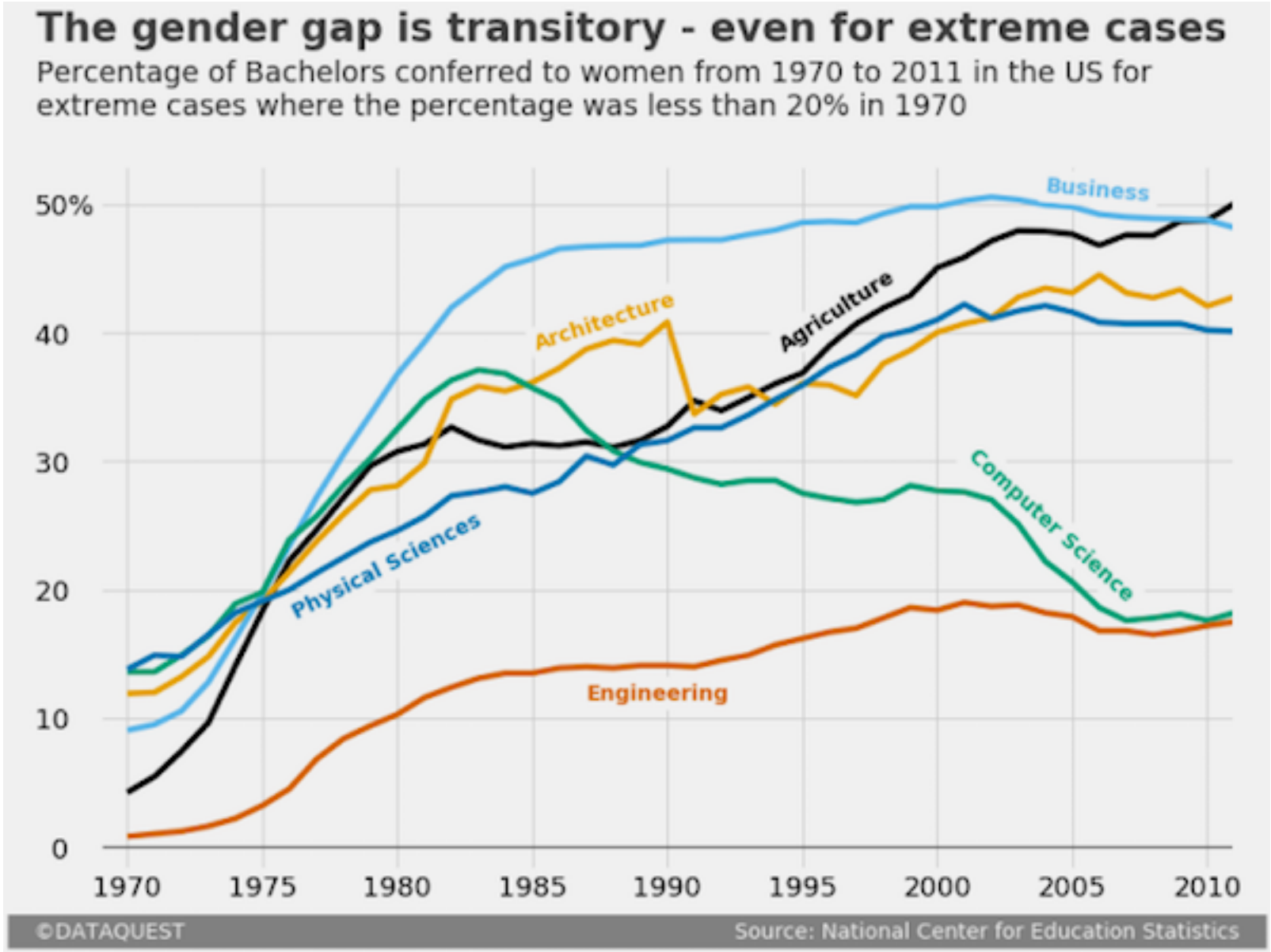
Power laws



Principles of visualization

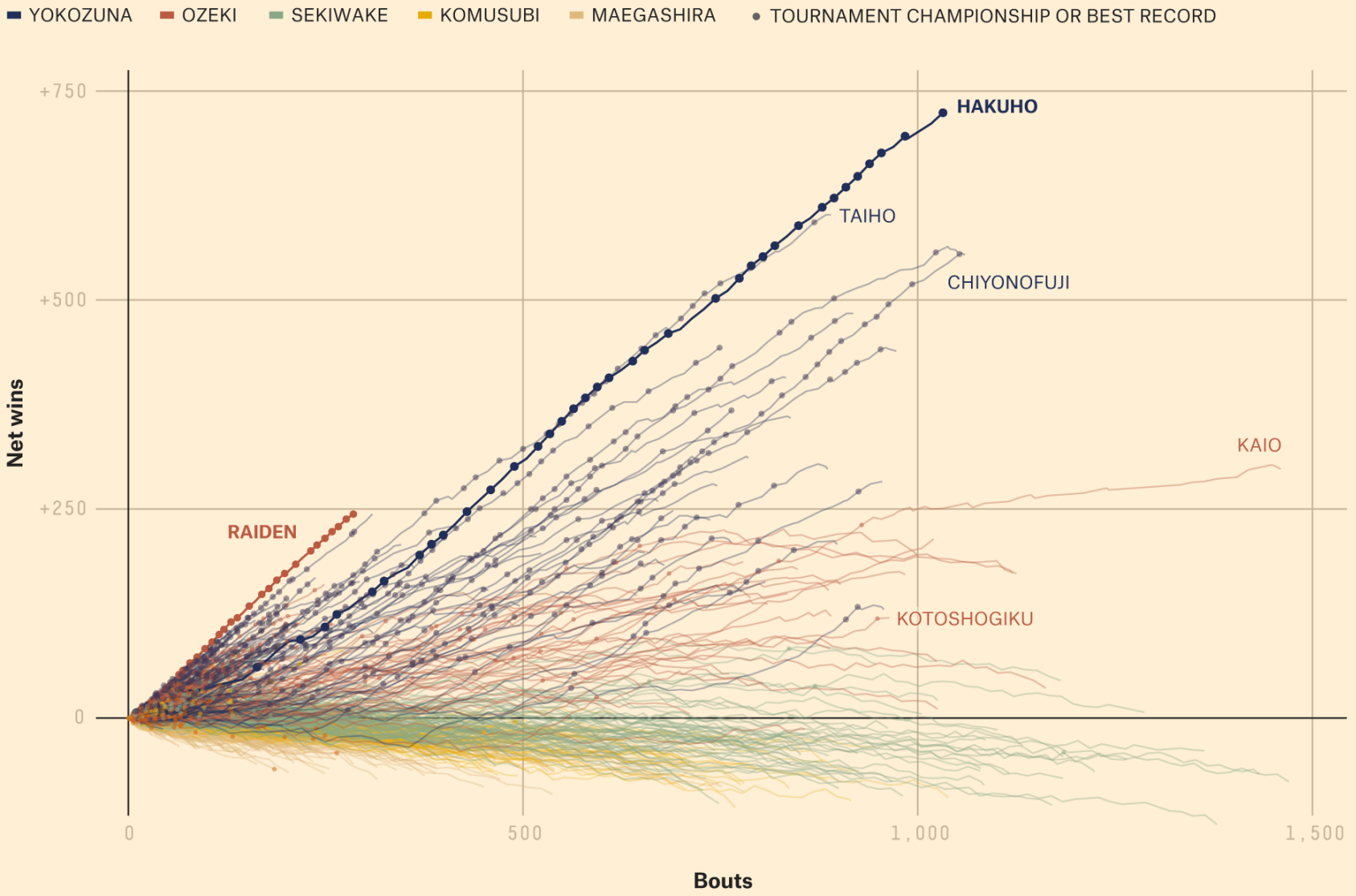
- Scale
- Conditioning
- Perception
- Transformation
- Context
- Smoothing


Agriculture
Architecture
Business
Computer Science
Engineering
Physical Sciences
50
40
30
20
10
0
1970
1975
1980
1985
1990
1995
2000
2005
2010
Year


The gender gap is transitory - even for extreme cases
Percentage of Bachelors conferred to women from 1970 to 2011 in the US for extreme cases where the percentage was less than 20% in 1970
50%
40
30
20
10
0
1970
1975
1980
1985
1990
1995
2000
2005
2010
Business
Architecture
Agriculture
Physical Sciences
Computer Science
Engineering
©DATAQUEST
Source: National Center for Education Statistics

The many types of sumo careers

Cumulative net wins vs. total bouts for sumo wrestlers in the top division by rank

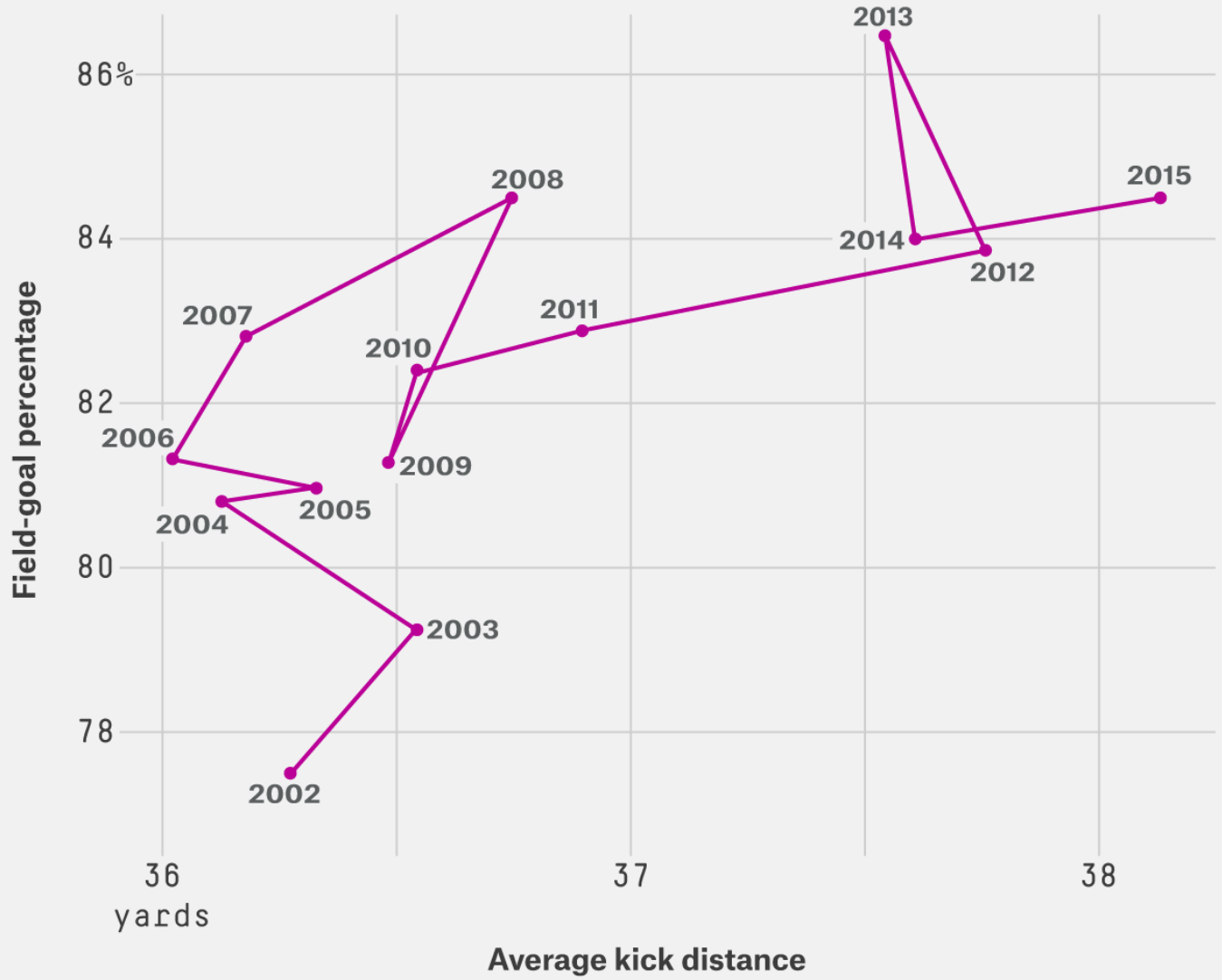


FIVETHIRTYEIGHT

SOURCE: SUMO REFERENCE

Kickers are taking longer attempts than ever

Field-goal percentage for all kicks vs. average kick distance



FIVETHIRTYEIGHT SOURCE: ESPN

Types of plots

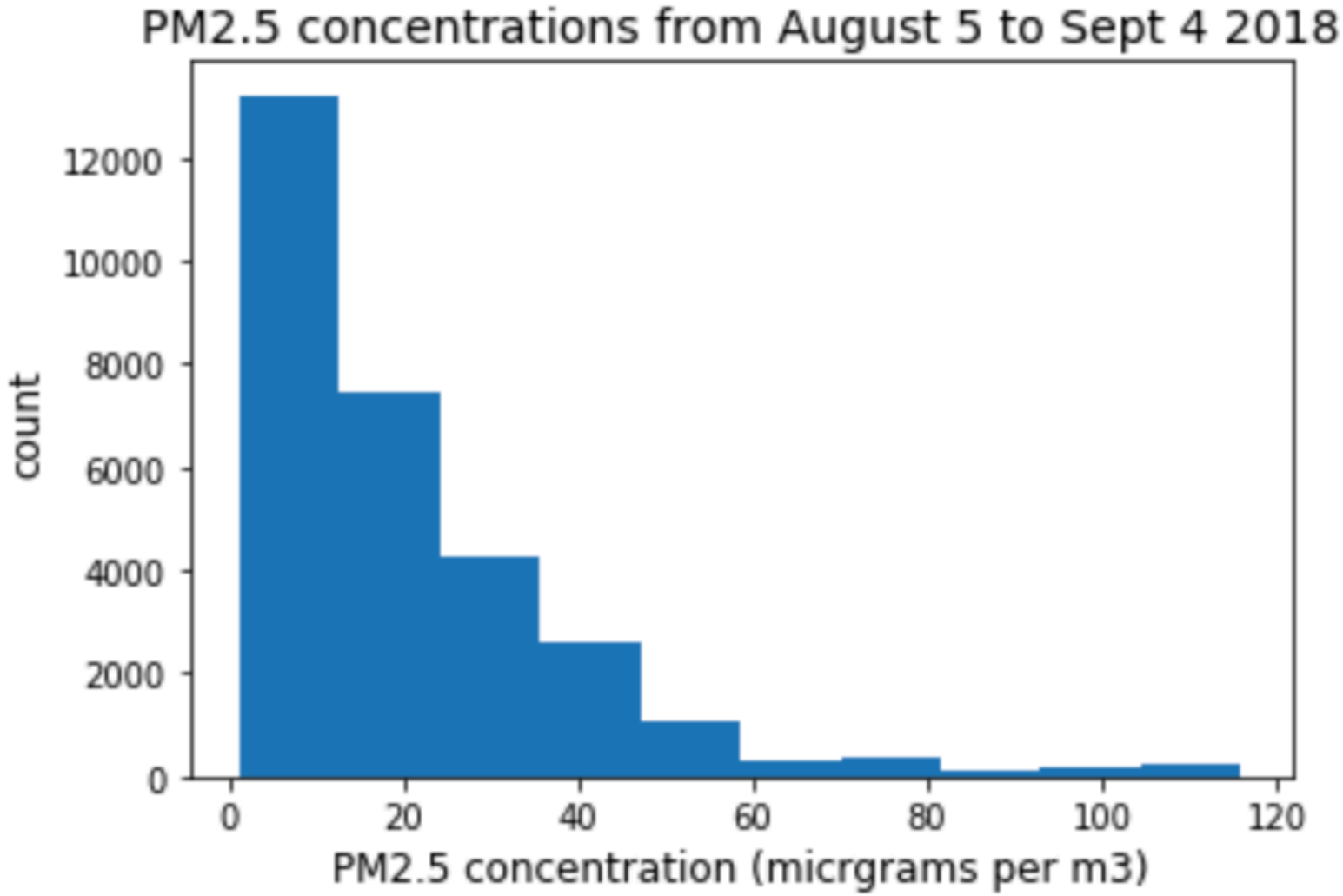
(lecture 7 stopped here; lecture 8 started here)

Plot types we'll talk about today

- Distributions:
  - Histogram
  - Kernel density
  - Box and whisker
- Scatter plots
- Line plots
- Bar charts
- Three dimensional plots
  - Geographic maps
  - Heat maps, contour maps
  - Scatter plots

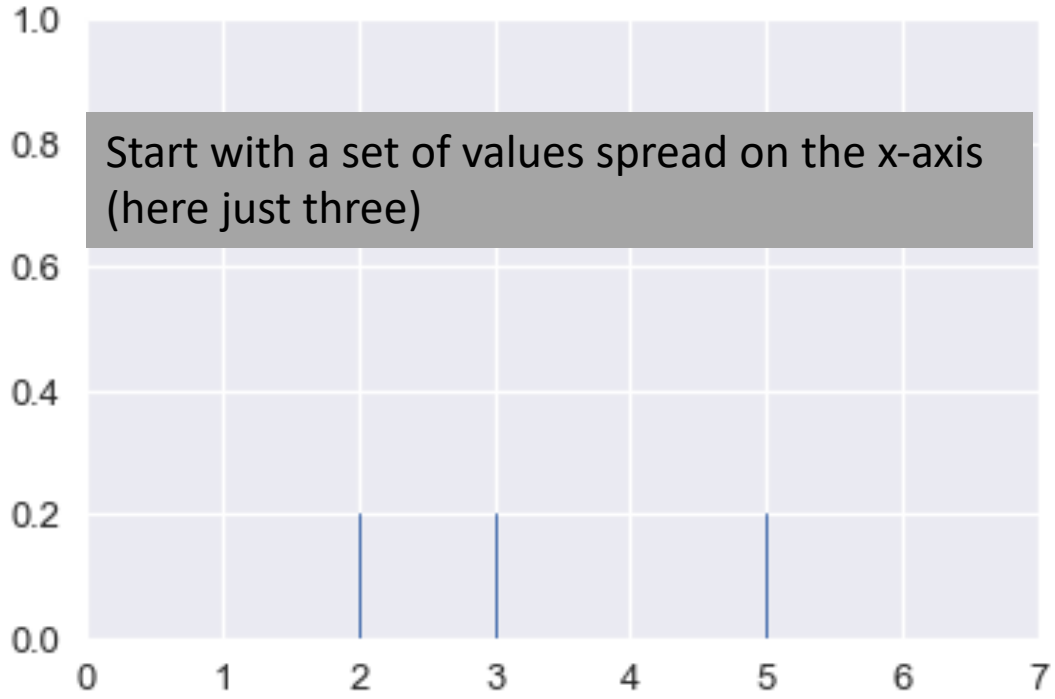
Histograms

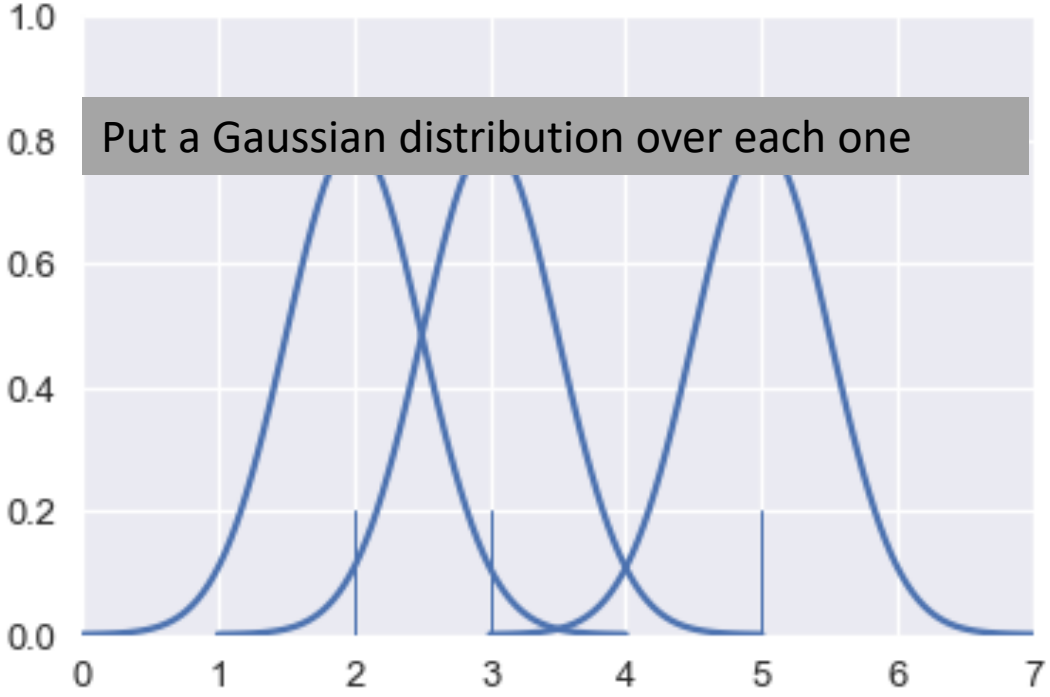
- Plots count (number of observations) versus a single variable

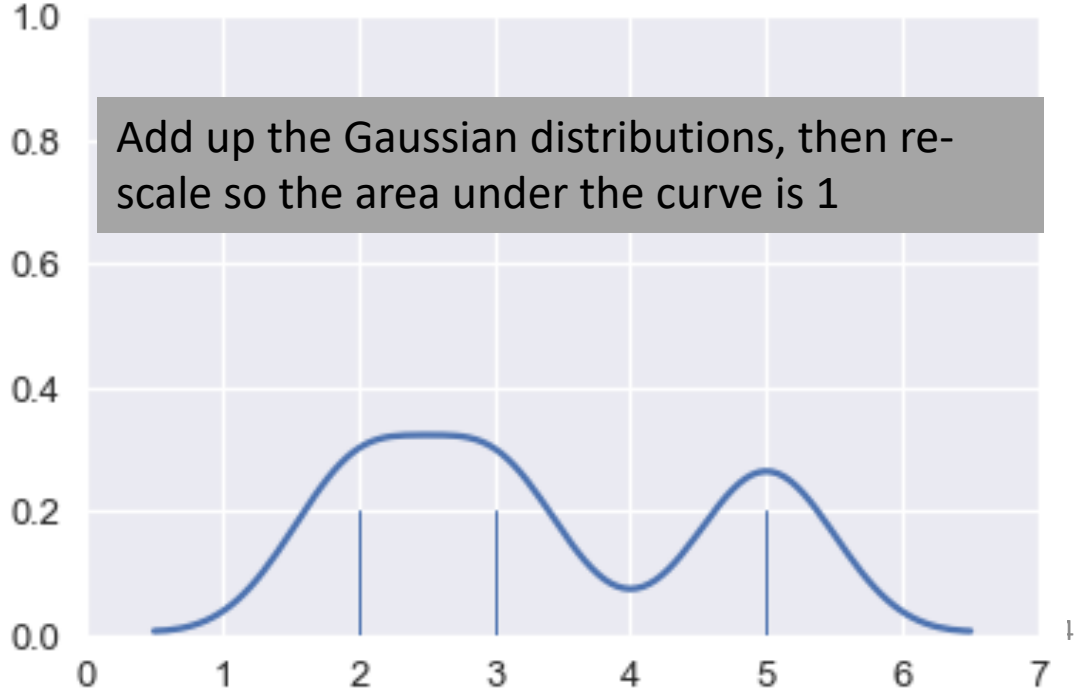


Ecole Bilingue data from PurpleAir

Smoothing with kernel density estimates





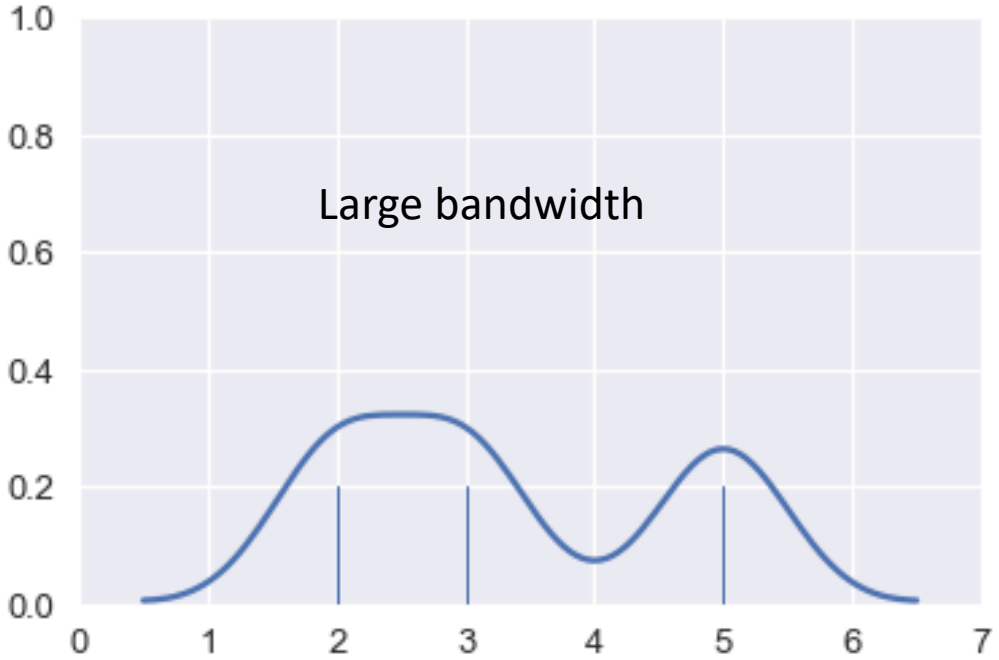


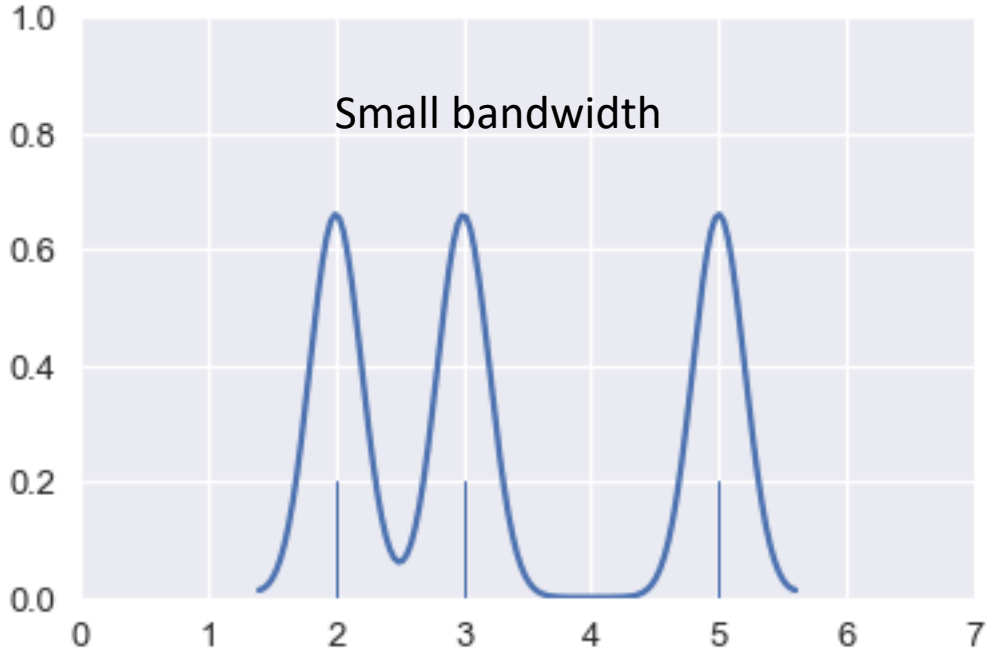
Why normalize the kernel density estimate curve so the area = 1?

- The KDE is an estimate of the probability distribution of values in the population
- Definition of the probability distribution is that the probability of an event happening in a particular *range* is the integral of the curve over that range
- Total probability can’t be greater than one, so we normalize.

The KDE “bandwidth”

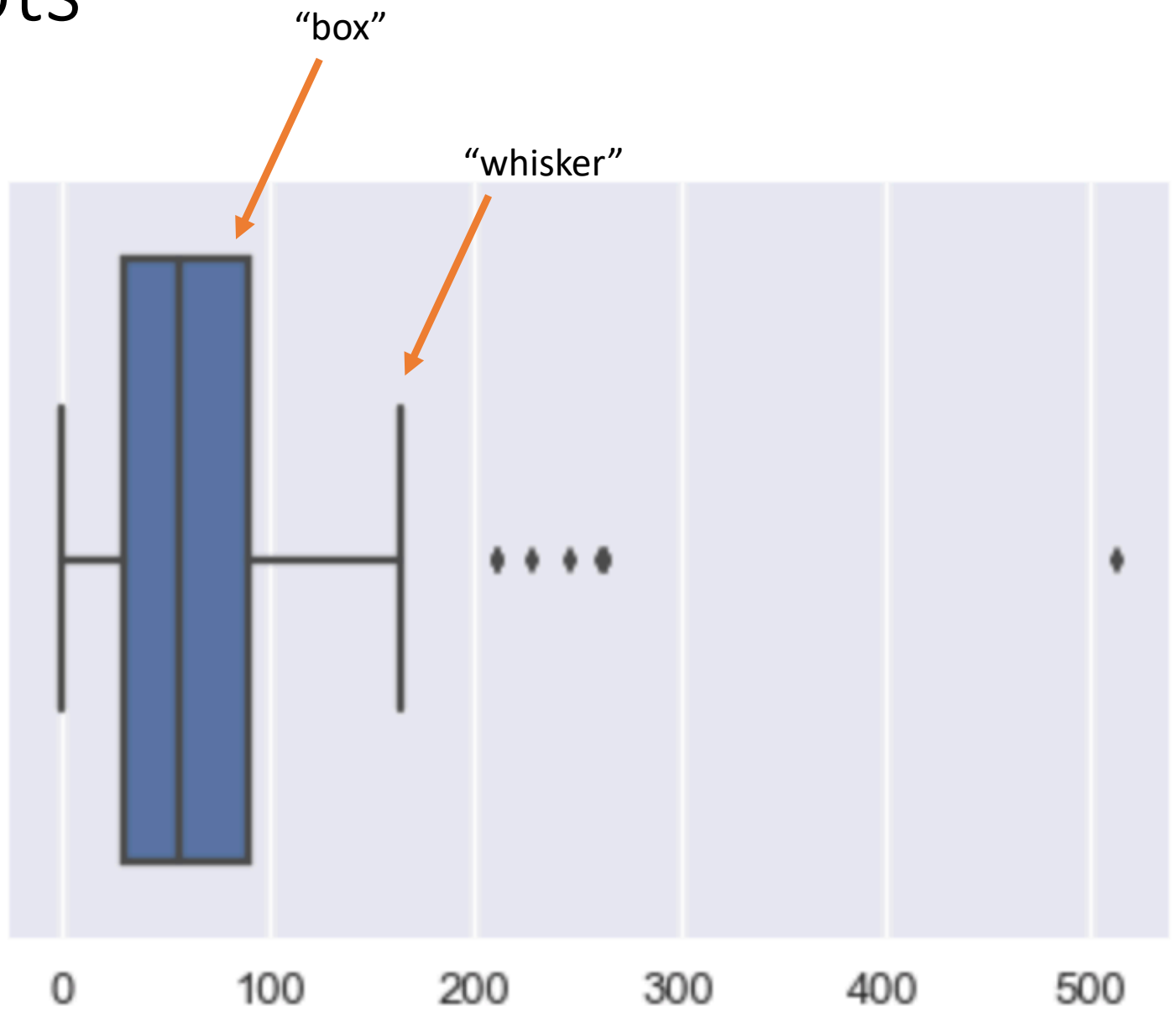
- We can vary the width of the Gaussian distribution we overlay on each point. We call this width the bandwidth
- What would be the effect on the final KDE plot or reducing or growing the bandwidth?





Box and whisker plots

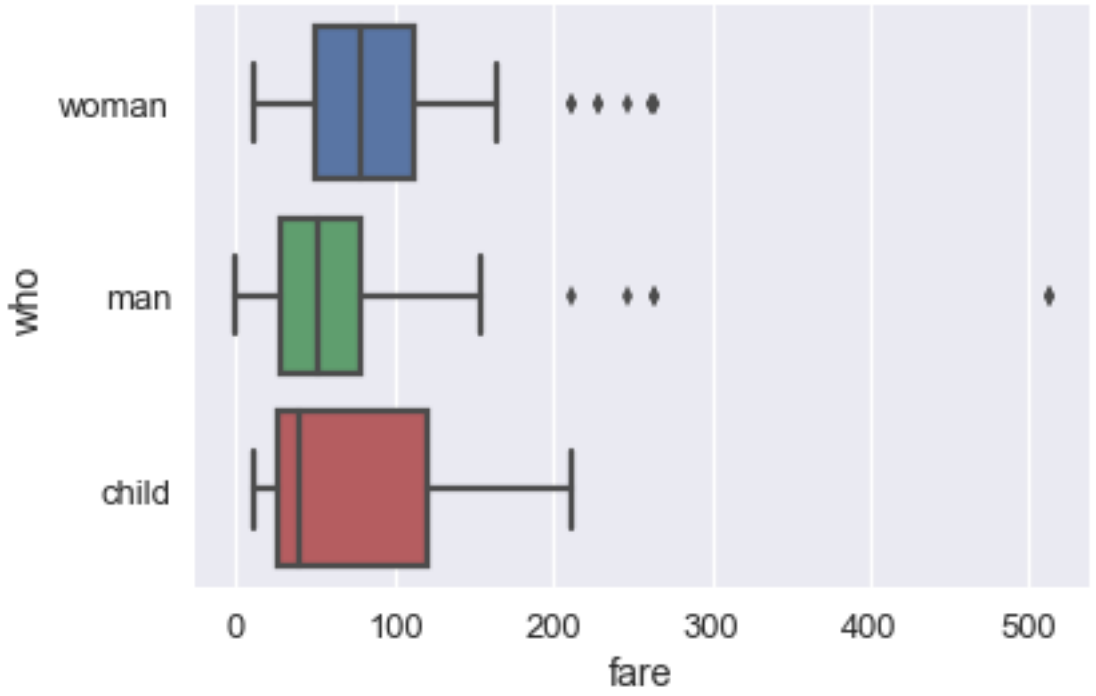
- Display similar information to a histogram, but abstracted:
  - Median
  - “Interquartile range” (IQR; 25th to 75th percentile)
  - Whiskers located (default)
    - 1.5 times the IQR away from the 25th or 75th percentile, or
    - At the last data point on that side of the distribution
  - Data points more than 1.5 times the IQR away from the 25th or 75th percentile are shown individually

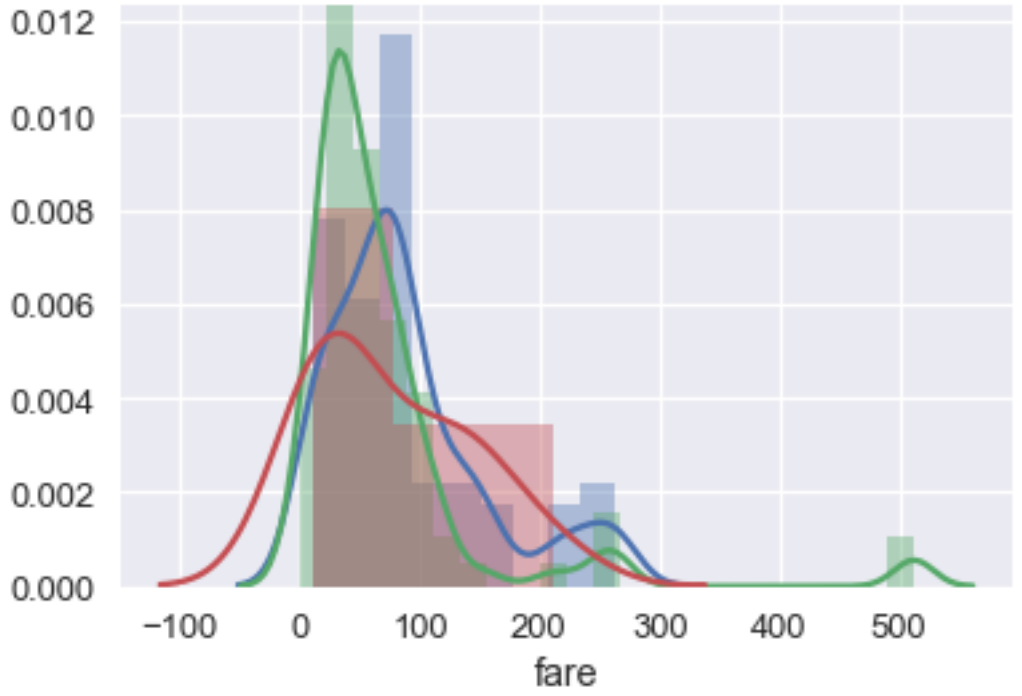


Quick notes on box and whisker

- Nice for comparing distributions
- They gloss over a lot of detail
- Parameter defaults are adjustable:
  - 25^{th} and 75^{th} percentile → Xth and Yth percentile (you can choose)
  - Whiskers at 1.5 times IQR → Z times IQR, or Xth and Yth percentile (you can choose)
- But most boxplots done with the standard parameter values we list here.

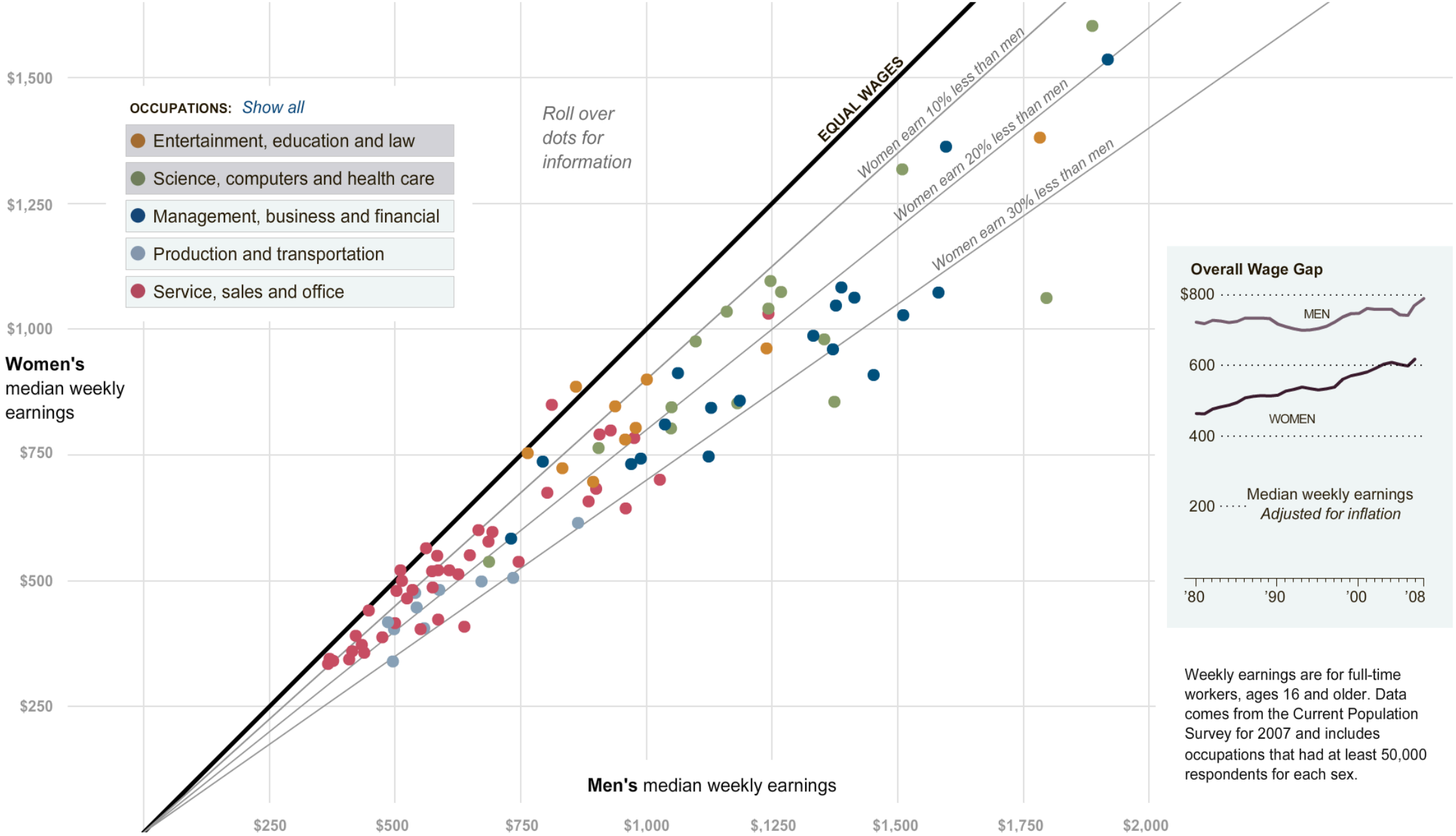
Same distributions, different ways to plot





DS100 textbook

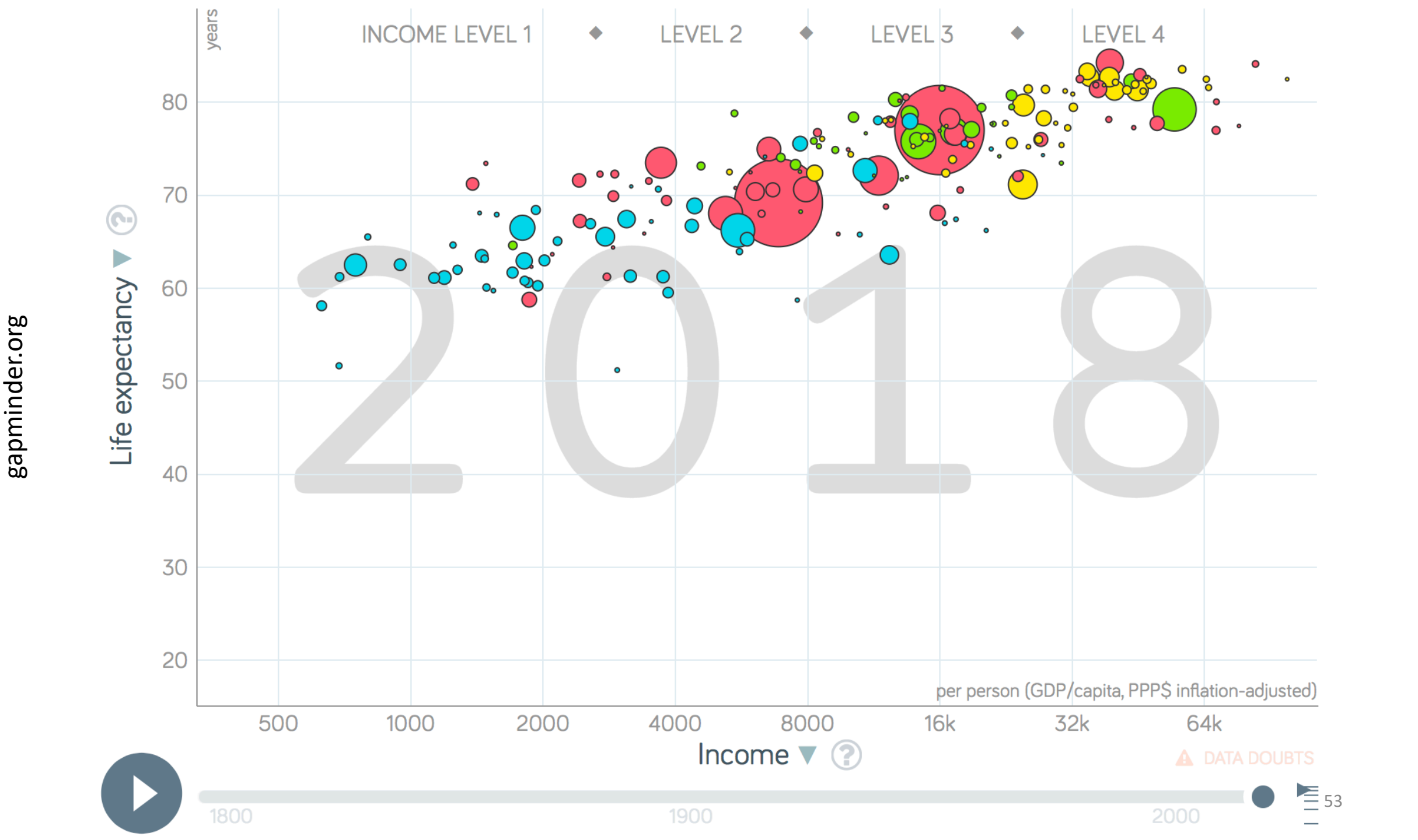
Scatter plots



https://archive.nytimes.com/

Scatter plots – things to adjust

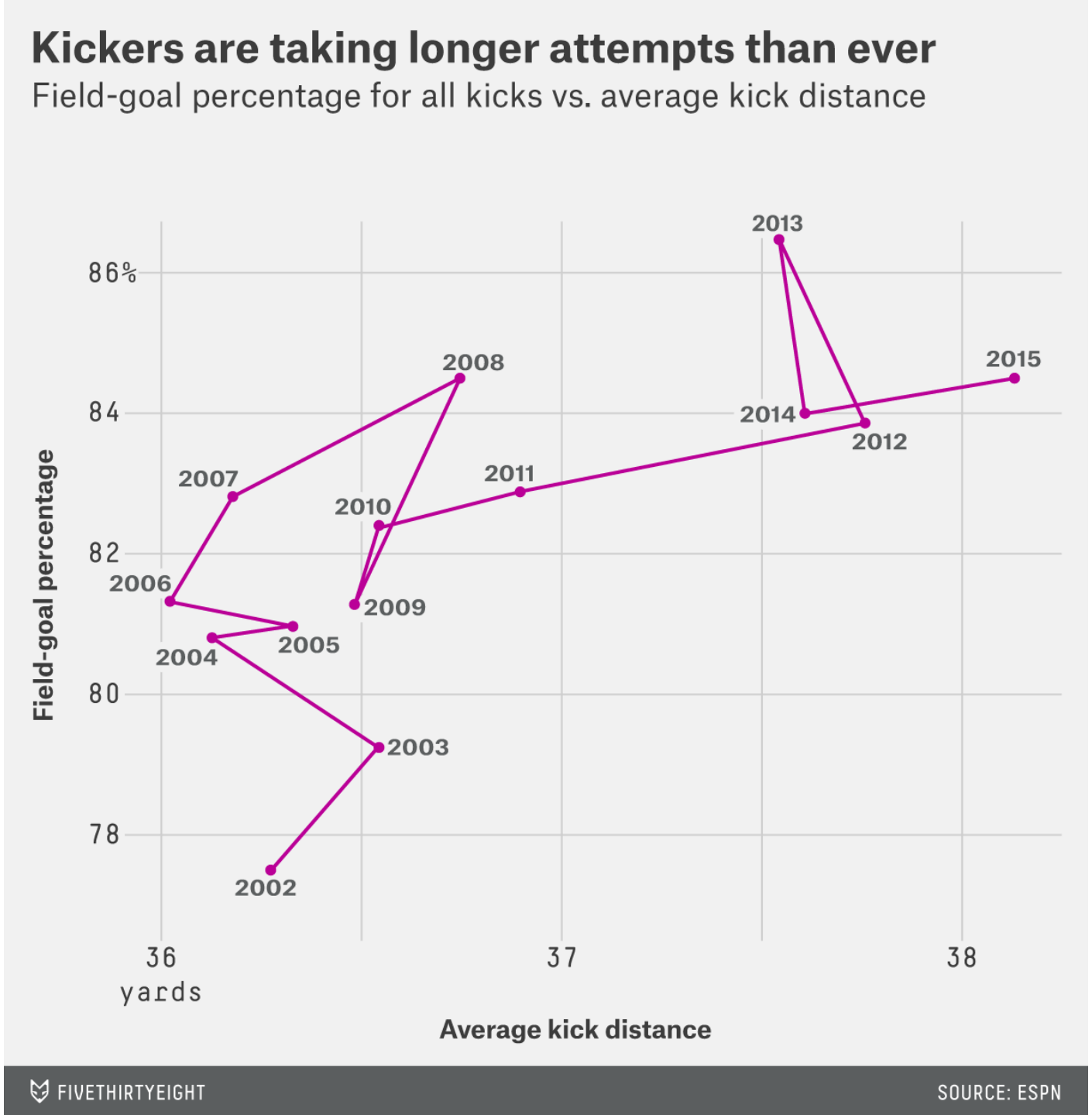
- Symbol shape → category
- Symbol size → additional quantitative data
- Symbol color → additional quantitative data or category
- Labels → context


gapminder.org
years
INCOME LEVEL 1
LEVEL 2
LEVEL 3
LEVEL 4
Life expectancy
80
70
60
50
40
30
20
2018
per person (GDP/capita, PPP$ inflation-adjusted)
500
1000
2000
4000
8000
16k
32k
64k
Income
DATA DOUBTS
1800
1900
2000

Line plots

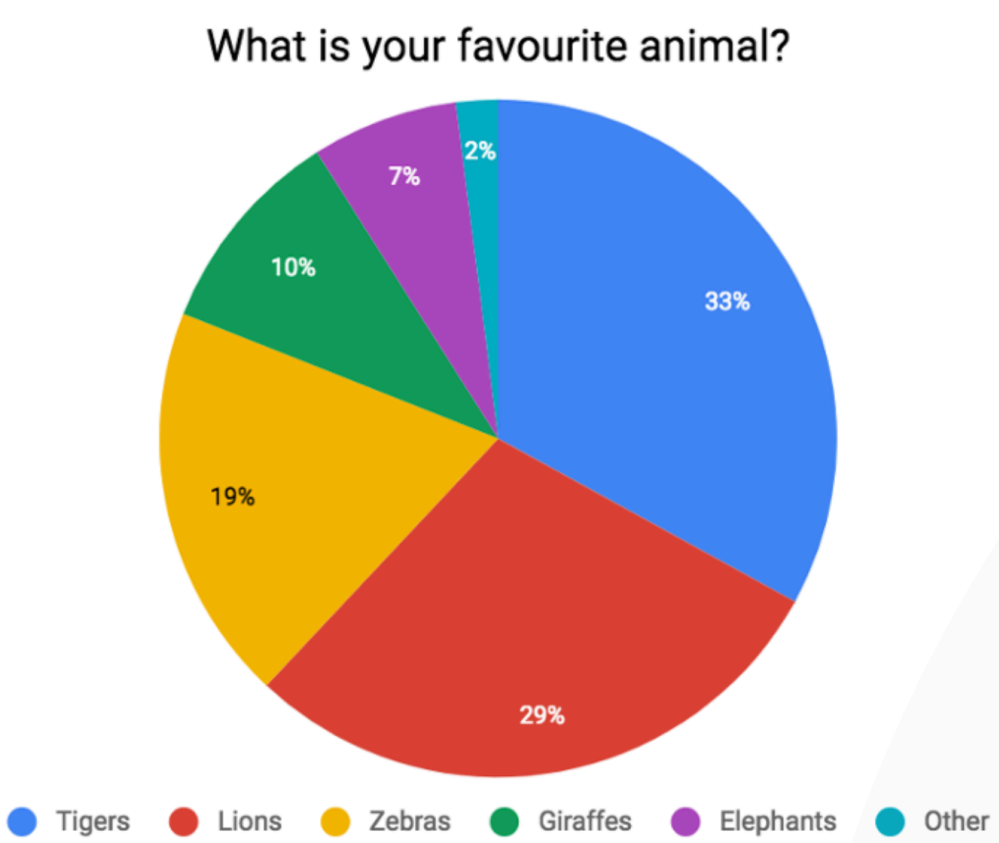
- Versus scatter, line plots connect measurements ”in order”
  - In cases where measurements are taken infrequently in time, this might help to visually interpolate what happened between measurements
  - It also gives a clear sense of sequence or correlation

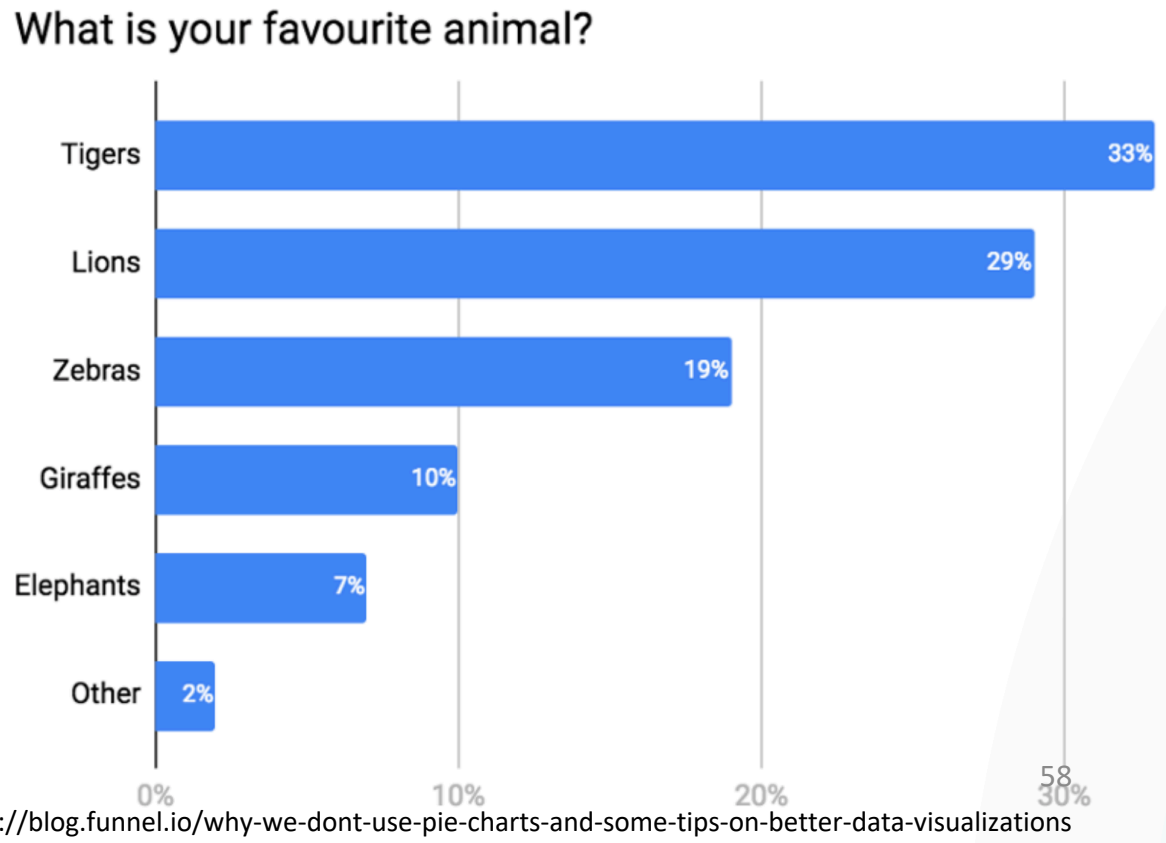
538


Kickers are taking longer attempts than ever
Field-goal percentage for all kicks vs. average kick distance
86%
84
82
80
78
Field-goal percentage
2013
2008
2015
2014
2012
2007
2011
2010
2006
2009
2005
2004
2003
2002
36
yards
37
38
Average kick distance
FIVETHIRTYEIGHT
SOURCE: ESPN

Bar charts

- Display quantity of interest across many different categories.
- You can do similar things with Pie charts, but they require comparing area (two dim.) which is harder than comparing length (one dim)

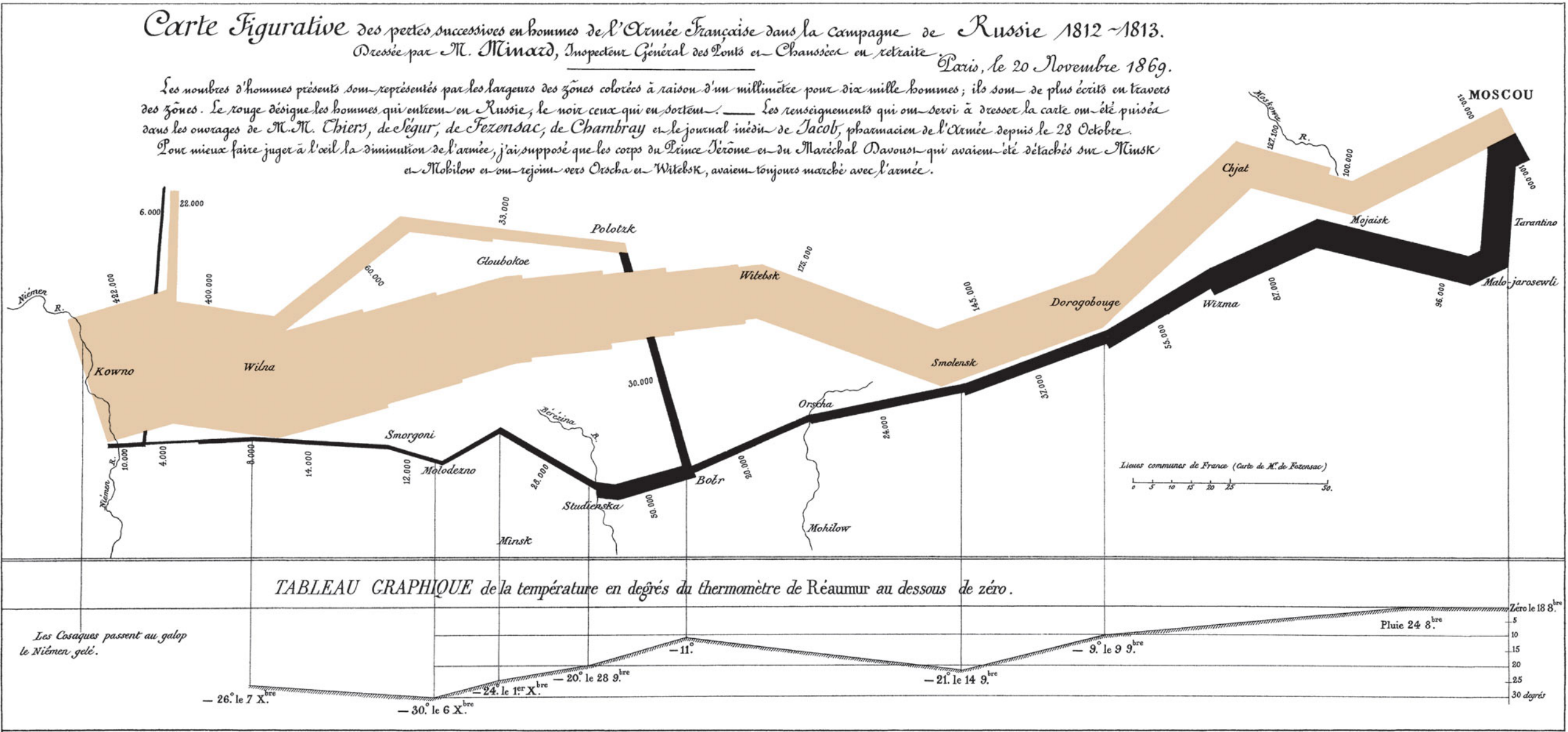




https://blog.funnel.io/why-we-dont-use-pie-charts-and-some-tips-on-better-data-visualizations

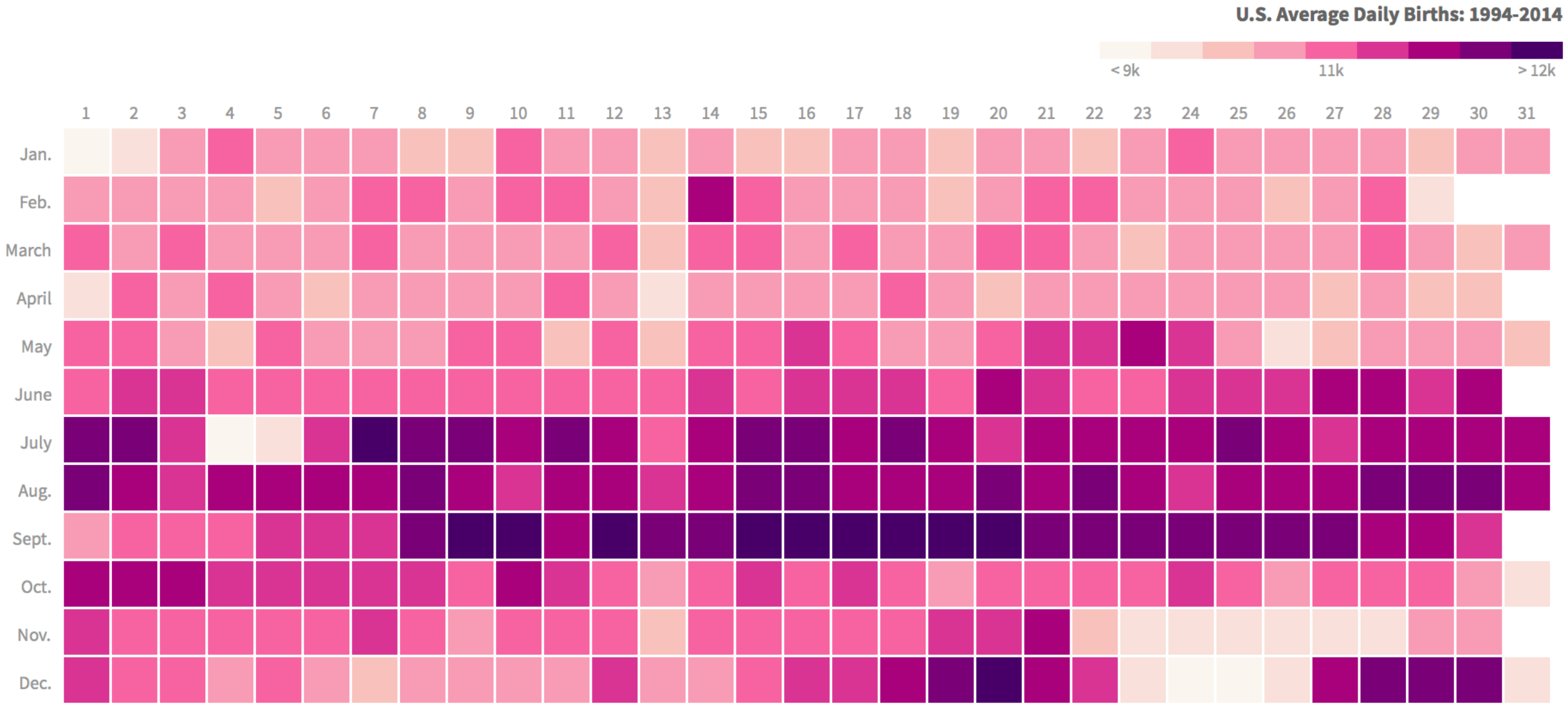
Three (and more) dimensional data

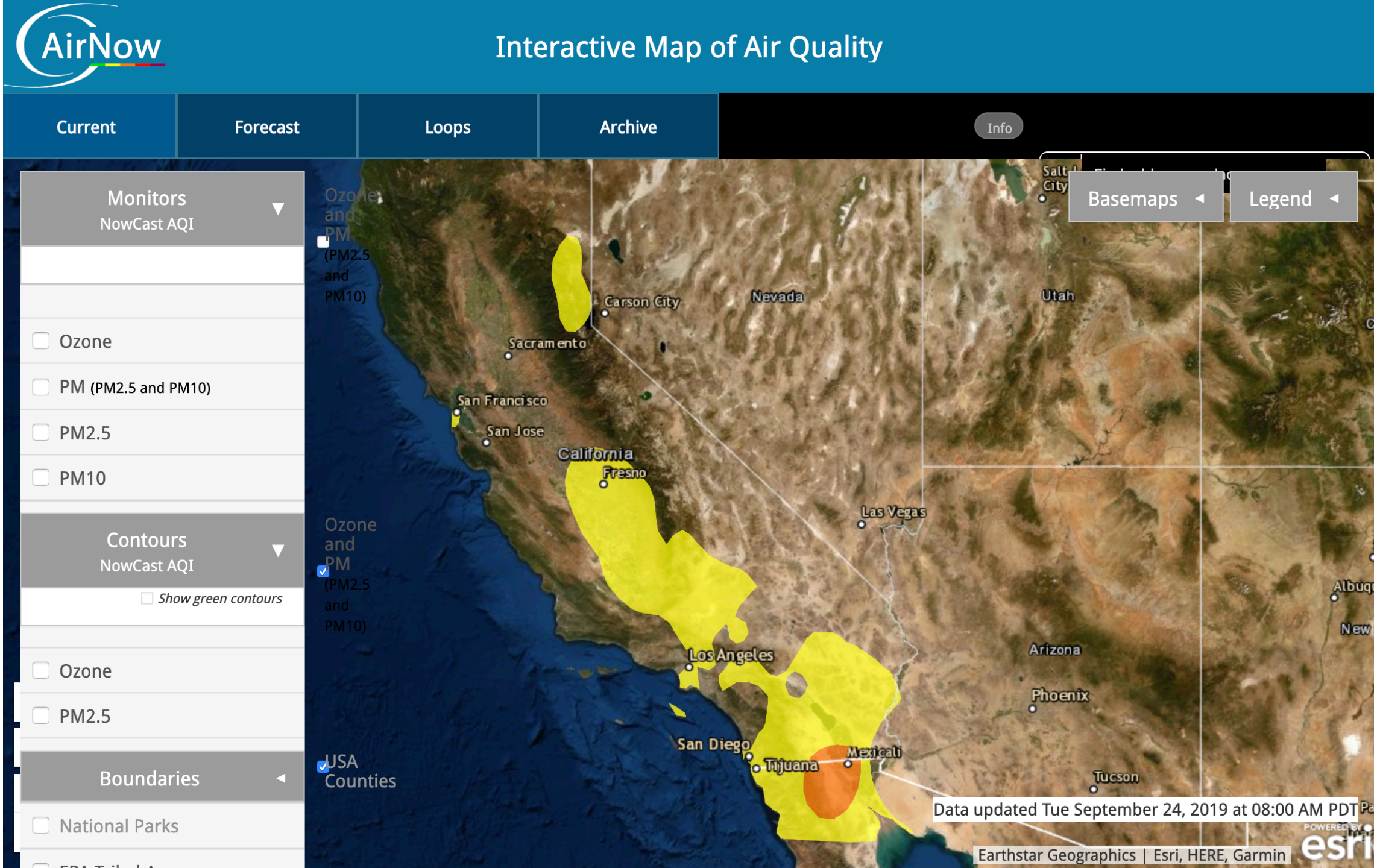
- Heat maps
- Contour plots
- Scatter plots with symbol size


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'oeil 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 en ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
Niemen R.
Kowno
Wilna
Smorgoni
Molodezno
Minsk
Studienska
Bérézina R.
Bobr
Orscha
Mohilow
Polotzk
Gloubokoe
Witebsk
Dorogobouge
Smolensk
Chjat
Moskowa R.
Mojaisk
Wizma
MOSCOU
Tarantino
Malo-jarosewli
6.000
22.000
422.000
400.000
33.000
60.000
30.000
125.000
145.000
127.100
100.000
100.000
87.000
96.000
55.000
32.000
24.000
10.000
4.000
6.000
14.000
12.000
28.000
50.000
30.000
Lieues communes de France (Carte de M.r de Fezensac)
0 5 10 15 20 25 50.
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
— 26.° le 7 X.bre
— 30.° le 6 X.bre
— 24. le 1er X.bre
— 20.° le 28 9.bre
— 11.°
— 21. le 14 9.bre
— 9.° le 9 9.bre
Pluie 24 8.bre
Zéro le 18 8.bre
5
10
15
20
25
30 degrés
Autog. par Regnier, 8. Pas. Ste Marie St Gain à Paris.
Imp. Lith. Regnier et Dourdet.

HOW POPULAR IS YOUR BIRTHDAY?

Two decades of American birthdays, averaged by month and day.




AirNow
Interactive Map of Air Quality
Current
Forecast
Loops
Archive
Info
Monitors
NowCast AQI
Ozone
PM (PM2.5 and PM10)
PM2.5
PM10
Contours
NowCast AQI
Show green contours
Ozone
PM2.5
Boundaries
National Parks
Ozone and PM (PM2.5 and PM10)
Ozone and PM (PM2.5 and PM10)
USA Counties
Basemaps
Legend
Salt Lake City
Carson City
Nevada
Utah
Sacramento
San Francisco
San Jose
California
Fresno
Las Vegas
Albuq
New
Arizona
Phoenix
Los Angeles
San Diego
Tijuana
Mexicali
Tucson
Data updated Tue September 24, 2019 at 08:00 AM PDT
Earthstar Geographics | Esri, HERE, Garmin
esri

Customizing plots

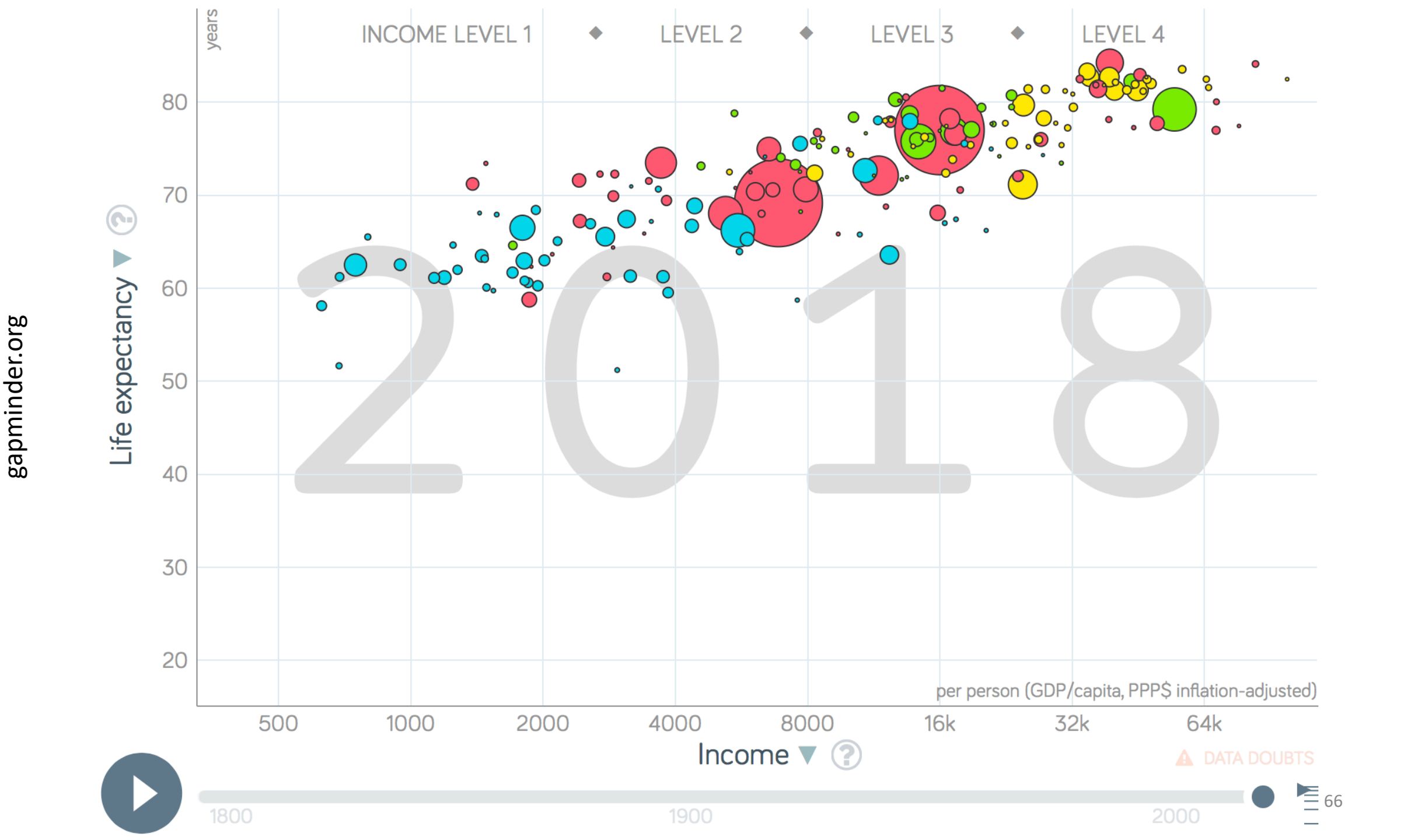
- Where to use text
  - At a minimum: x-label, y-label, title
  - Might also be relevant: legend, labeling features on the figure
- Symbols on line and scatter plots
- Colors – we've discussed already…
- Line width – use your judgment
- Font size
  - Title largest
  - Axis labels medium
  - Tick labels smallest
- Font type
  - Use sans-serif!

Fonts

- Should be readable in intended format
  - Paper: fonts can be 1-2 points less than text of paper
  - Presentations: fonts should be easily readable from back of room
- Font type:

  Serif font, `Serif font`

  Sans serif font, Sans serif font
- Best to use sans serif for figures


gapminder.org
years
INCOME LEVEL 1
LEVEL 2
LEVEL 3
LEVEL 4
Life expectancy
80
70
60
50
40
30
20
2018
per person (GDP/capita, PPP$ inflation-adjusted)
500
1000
2000
4000
8000
16k
32k
64k
Income
DATA DOUBTS
1800
1900
2000
66