Taller de fundamentos de visualización de datos

Hola, este es el material que usé para el taller en la Tarugo#4 (octubre ‘19) y aunque algunas cosas se entienden bien, este contenido no está pensado para ser auto explicativo.

Cualquier pregunta a javi@tinybird.co espero que os sirva de ayuda y por lo menos os obligue a pensar un poco la próxima vez que representéis datos en pantalla.

Un saludo
Javi Santana

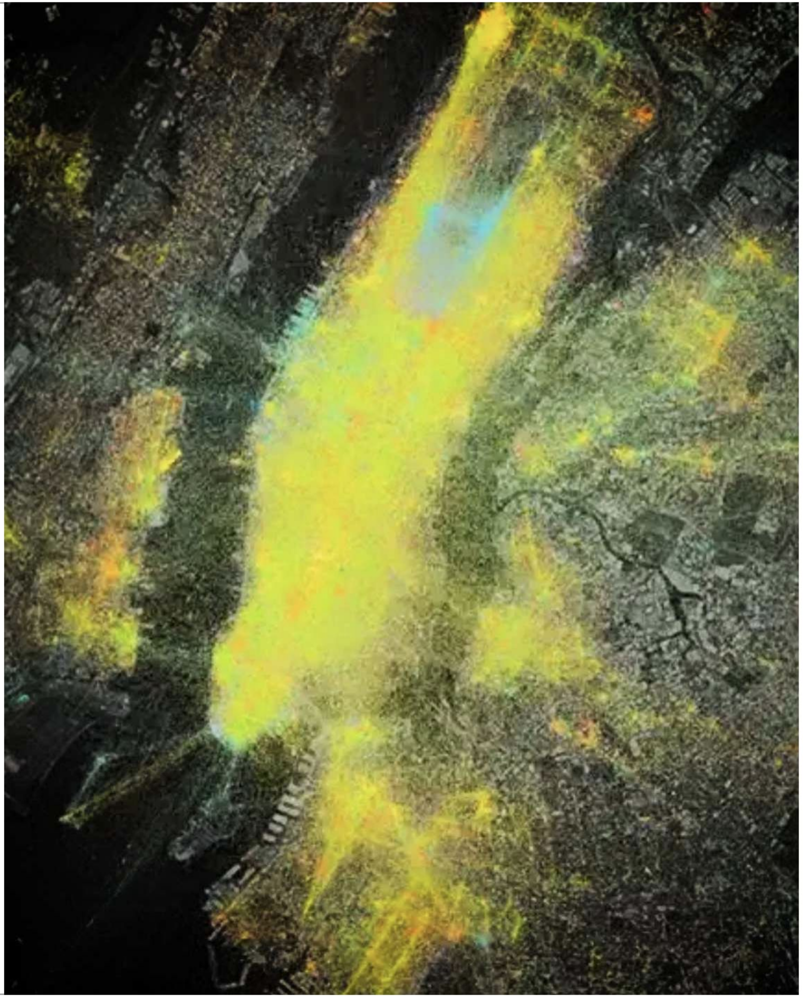
Data Visualization

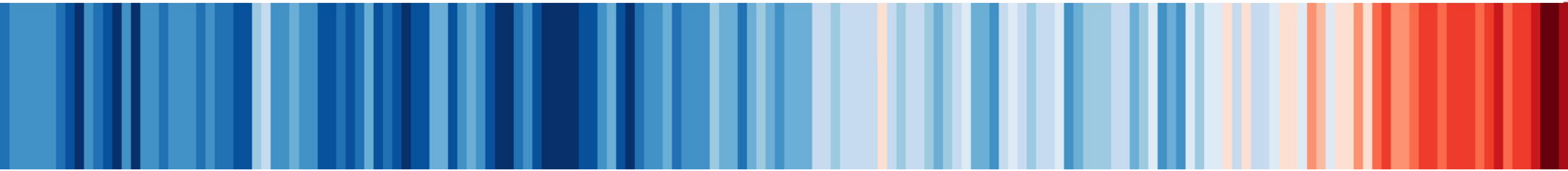
by Tinybird

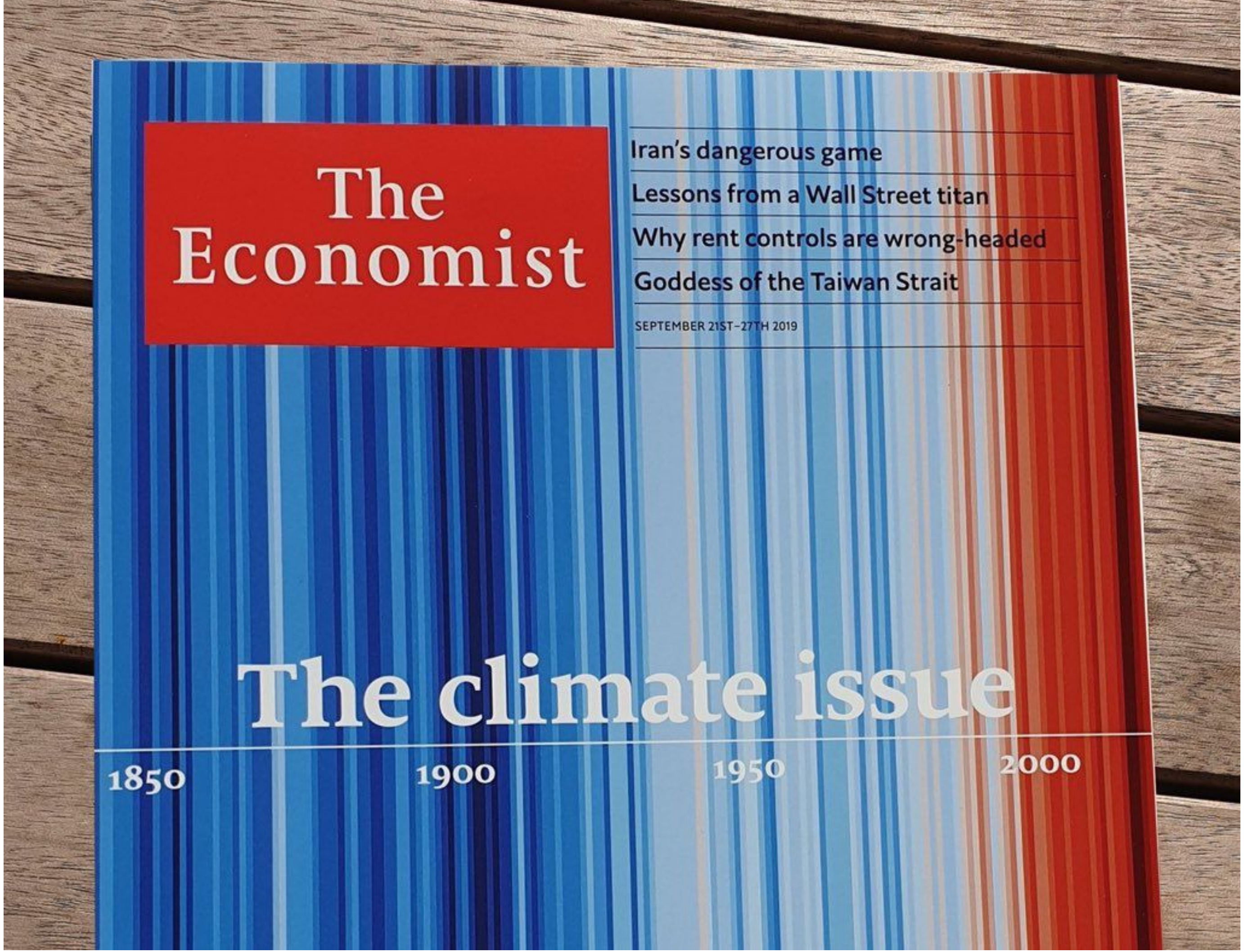
CHECKING IN | AUG. 27, 2019

Ten Years On, Foursquare Is Now Checking In to You Even the company is still trying to figure out whether that's "cool or creepy."

By James D. Walsh







The Economist

Iran's dangerous game

Lessons from a Wall Street titan

Why rent controls are wrong-headed

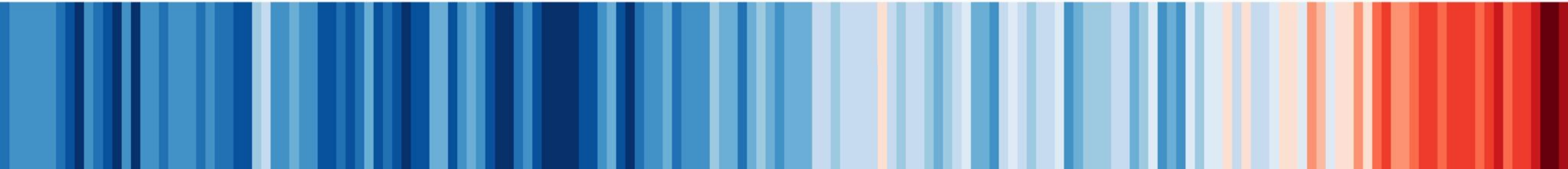
Goddess of the Taiwan Strait

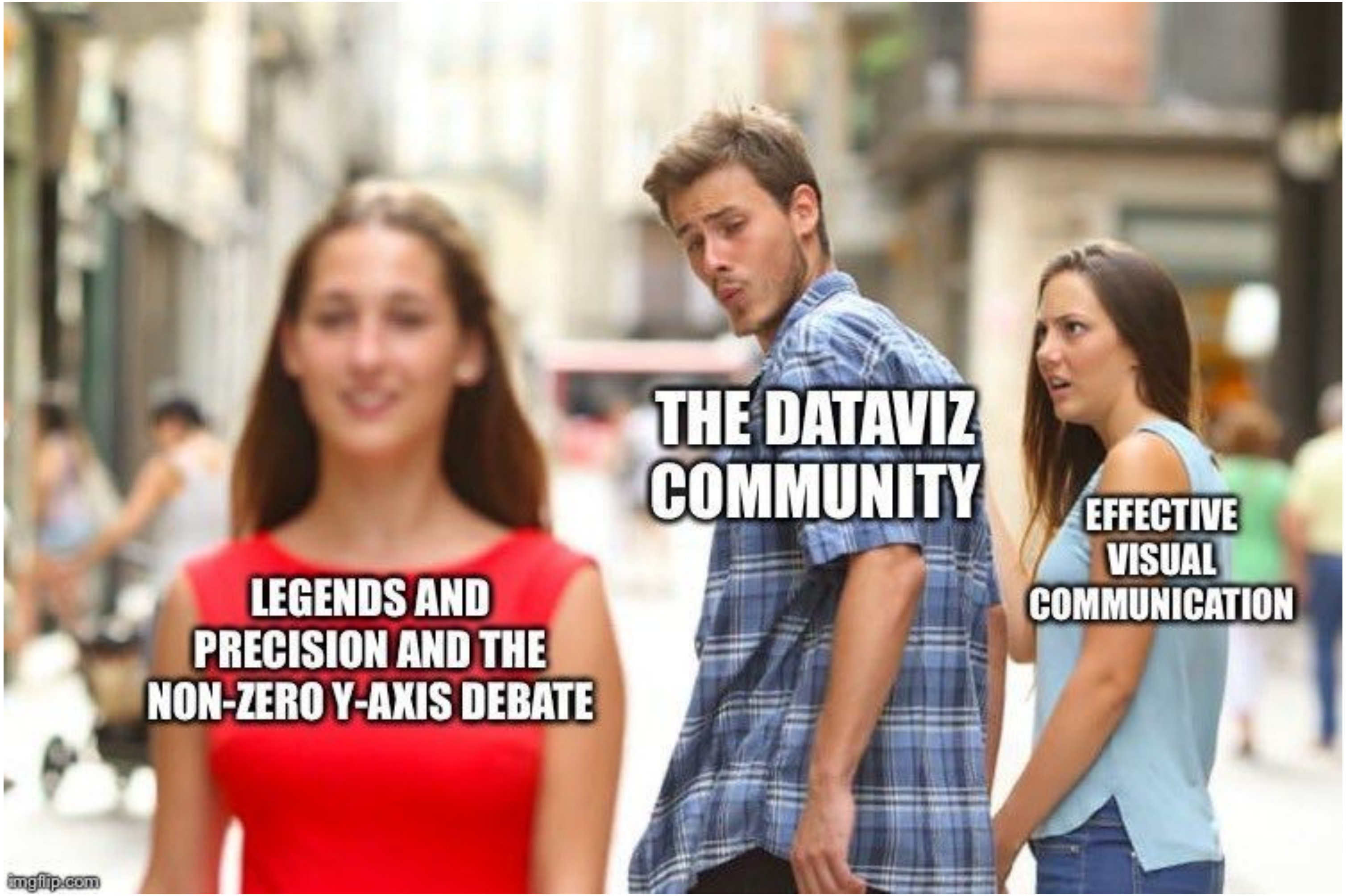
SEPTEMBER 21ST–27TH 2019

The climate issue

“I was looking for a way to communicate to audiences that aren't used to seeing graphs, or axes, or labels — things that we see day-to-day, but are complicated to them. It may look too mathematical to them, so it turns them off straight away.”

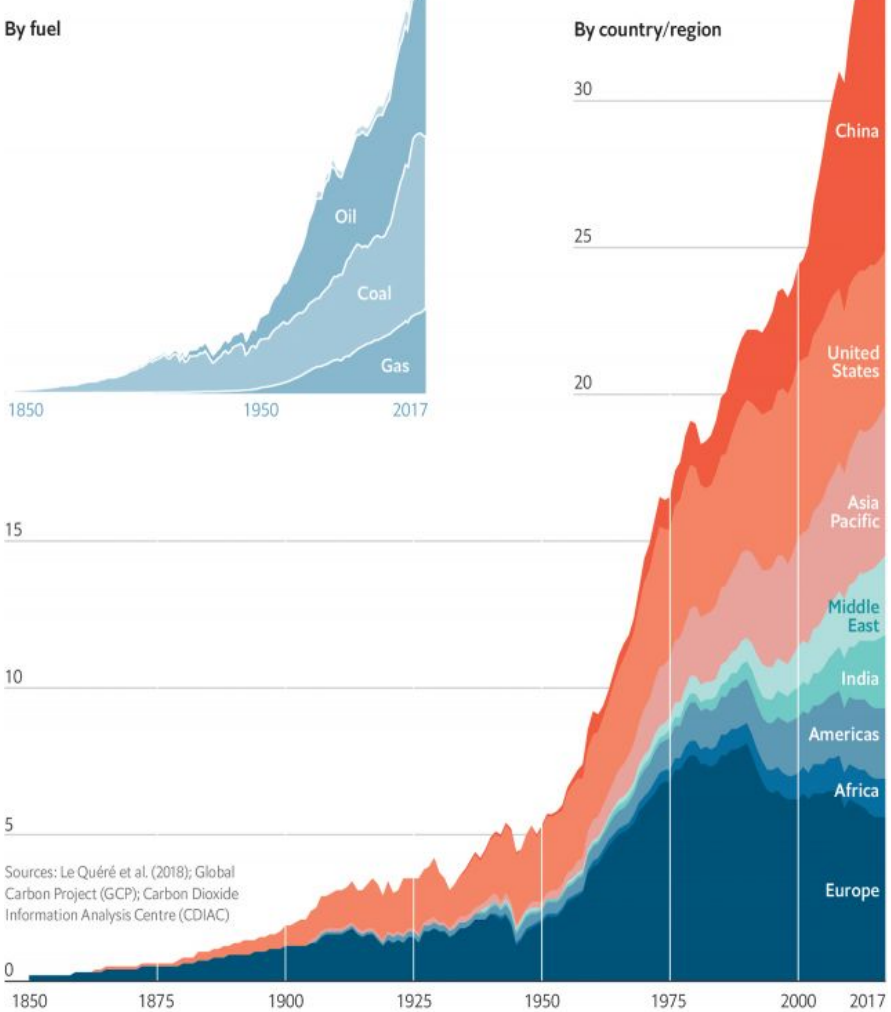
— @ed_hawkins




THE DATAVIZ COMMUNITY
LEGENDS AND PRECISION AND THE NON-ZERO Y-AXIS DEBATE
EFFECTIVE VISUAL COMMUNICATION
imgflip.com

https://twitter.com/jburnmurdoch/status/1182190780145979392

CO_2 emissions, gigatonnes



Sources: Le Quéré et al. (2018); Global Carbon Project (GCP); Carbon Dioxide Information Analysis Centre (CDIAC)

The Economist

01.
What’s Data vis

02.
The Basis

03.
Examples

04.
Conclusions

05.
Must reads

06.
Hands on

01.

What’s Data visualization?

What’s Data visualization?

Wikipedia

Data visualization is the graphical representation of data

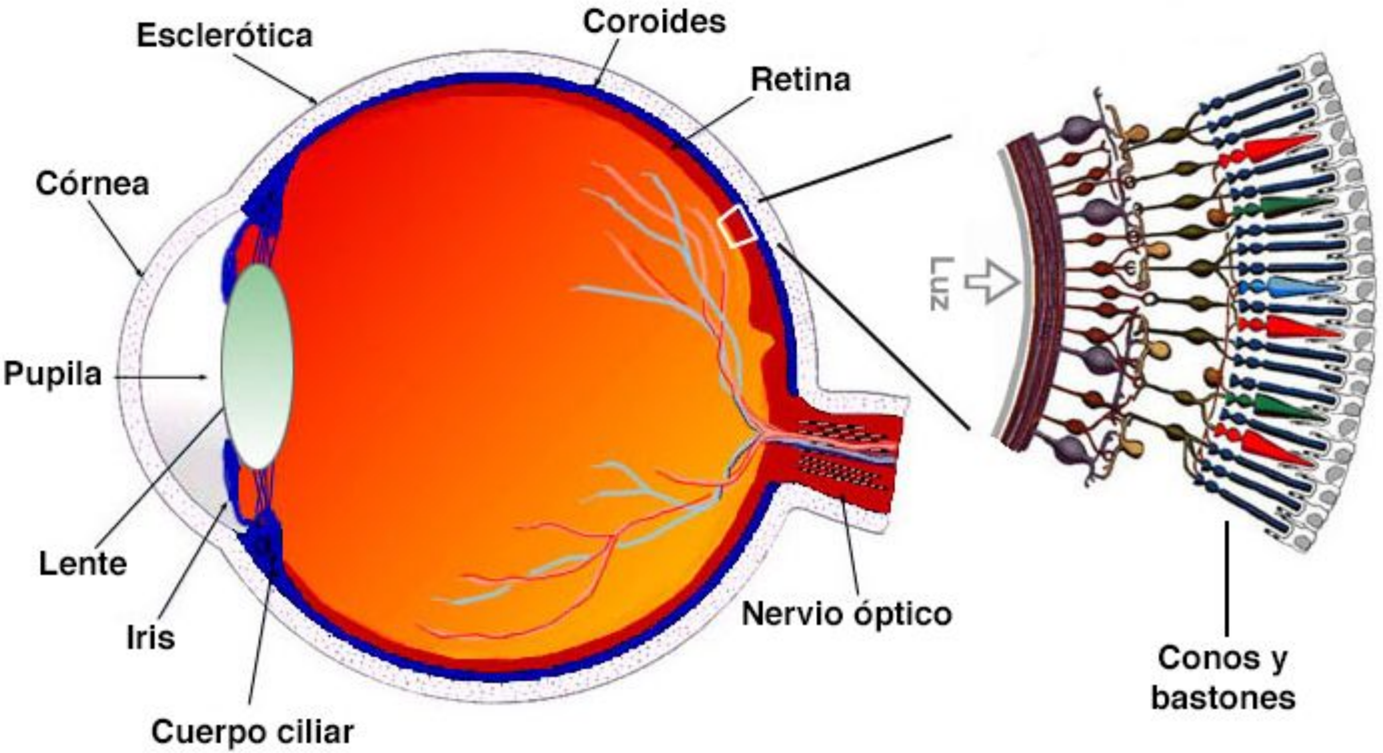
What’s Data visualization?

A number is data visualization

Emails sent
34,796,235,769
Templates created by users
3,361,313
Email addresses added to the system
270,000,000
15,445,001 campaigns sent
YEAR 2012
MailChimp
Support tickets
404,437
Support emails
225,611
Chats
178,826
Volume by day of the week, email and chat support
Total orders
26
Total pizzas
918
$$$
Total slices
7,344
Slices wasted
93
Slices per person
2.5
1.4 MILLION
184,000
250 PROGRAMS
300 YMC COMMUNIT
llion people who backed a project in 2012
72
acked two or
jects
50,047
people backe
more project
le backed 100 or more projects
29,943
18,382
923,482
584,394
1.54
57,2
643
62,9
485
67%
of the world's working hours are worked by women
43%
of the developing world's workforce is female
18%
38%
10%
48%
INDEX
65
37
ONE HUNDRED AND NINETY-SIX
THIRTEEN
VOLVO
TEN
3

What’s Data visualization?

A number is data visualization

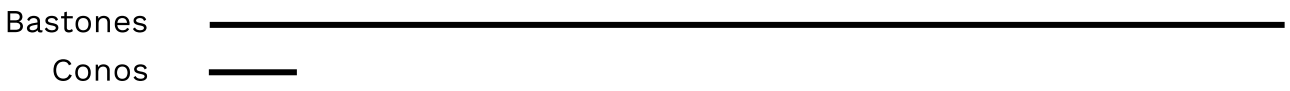

Esclerótica
Coroides
Retina
Córnea
Pupila
Lente
Iris
Cuerpo ciliar
Nervio óptico
Luz
Conos y bastones

What’s Data visualization?

Perception

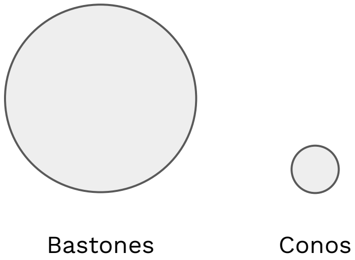
120M bastones vs **7M** conos

Perception II



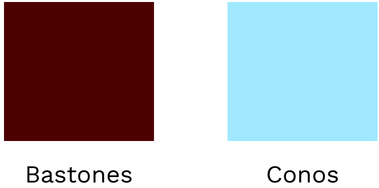
What’s Data visualization?

Perception II



What’s Data visualization?

Perception III



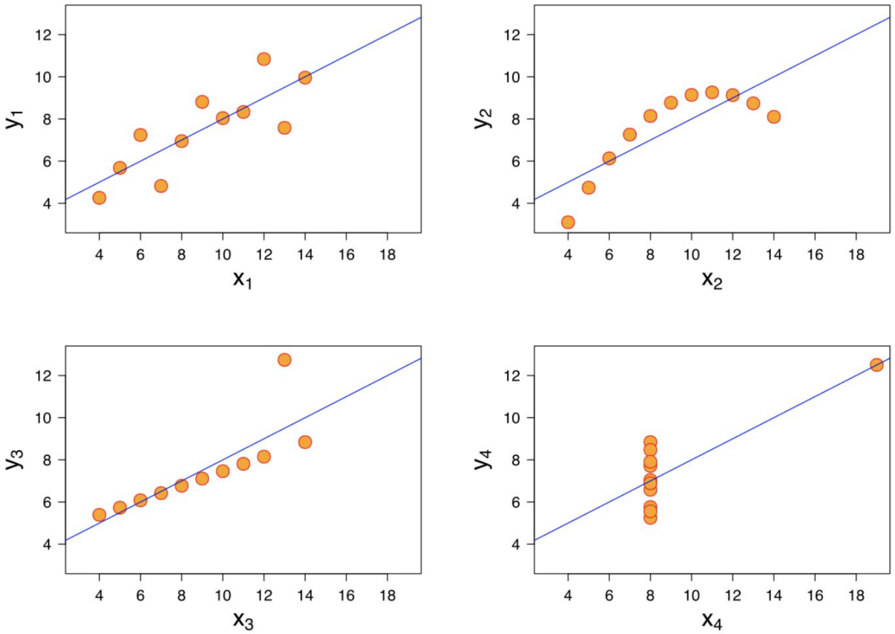
What’s Data visualization?

Perception (and IV)

Bastones	Madrid -> Valencia
Conos	Madrid -> Leganés

What’s Data visualization?

Why do we need a graphical representation if we have numbers and feelings



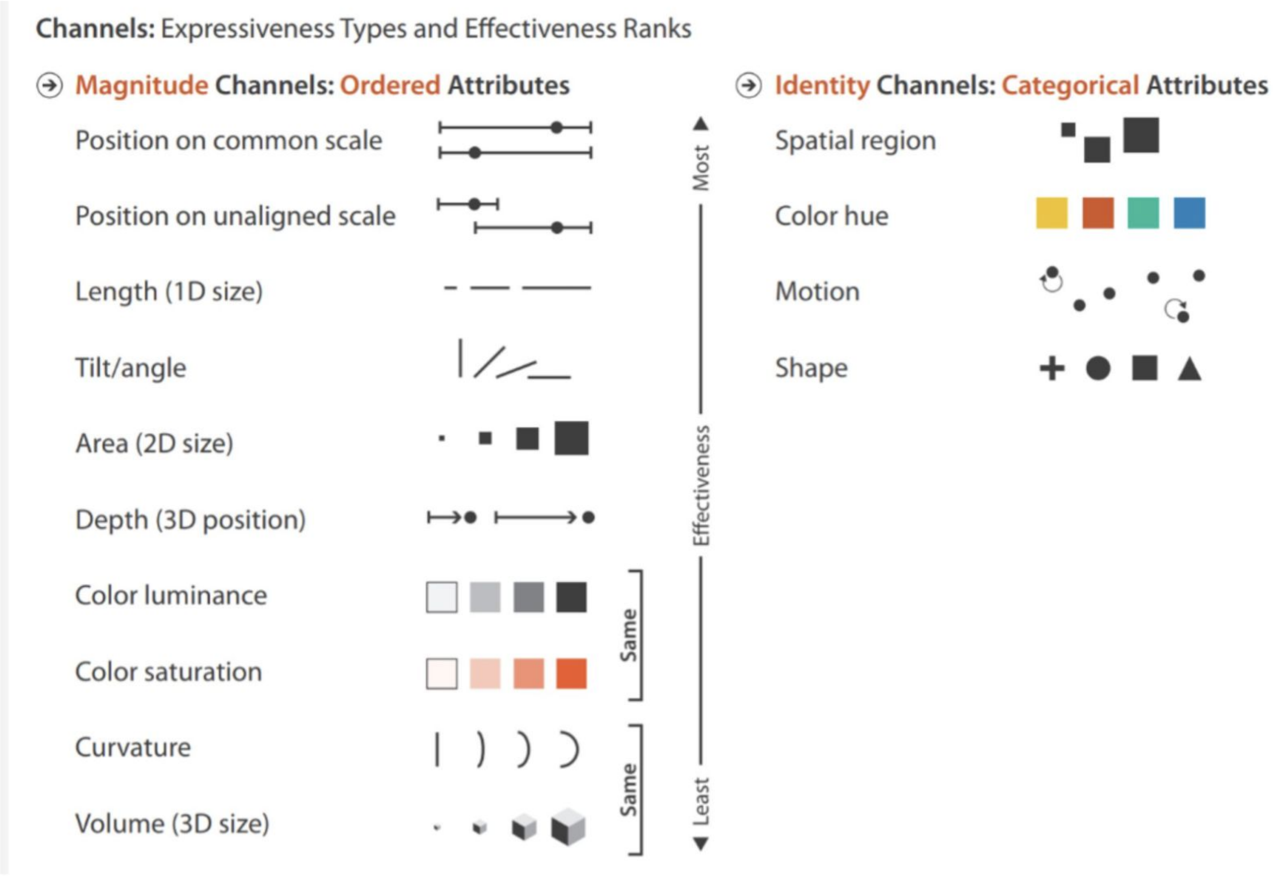
02.

The basis

The basis

How to represent different kind of data

Magnitudes: age, speed...
Categories: political party, football team...



https://twitter.com/DataRemixed/status/1136090210113019904

http://hint.fm/wind/

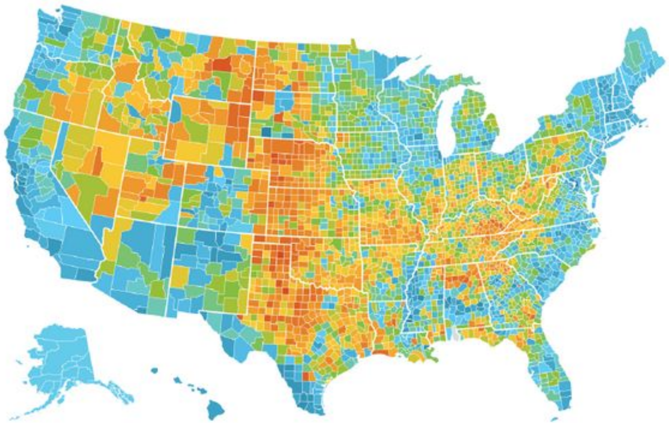
20
40
60
80
120
140
160
180
M.P.H.
VDO
12
3
6
9
VDO
8
9
10
11

The basis

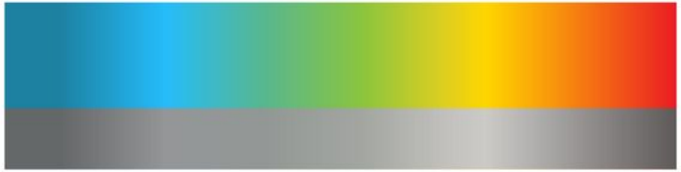
How to represent different kind of data

- Position and size: easy to understand
- Color: not really

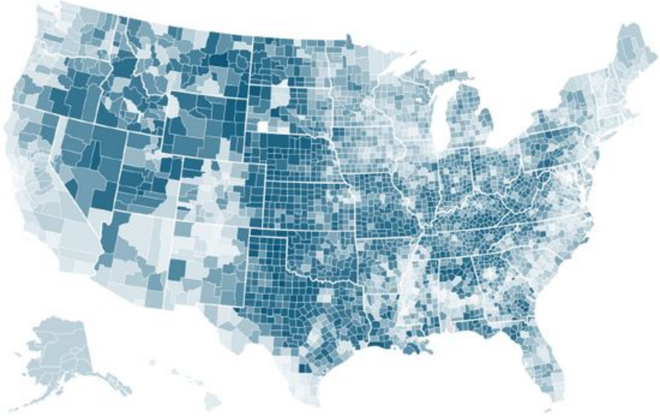
NOT IDEAL



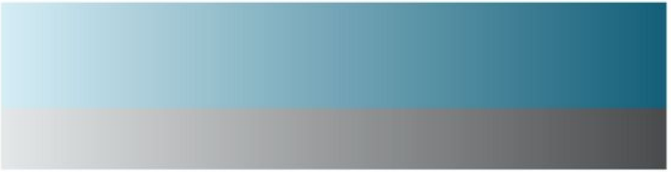
HUE-BASED GRADIENT



BETTER

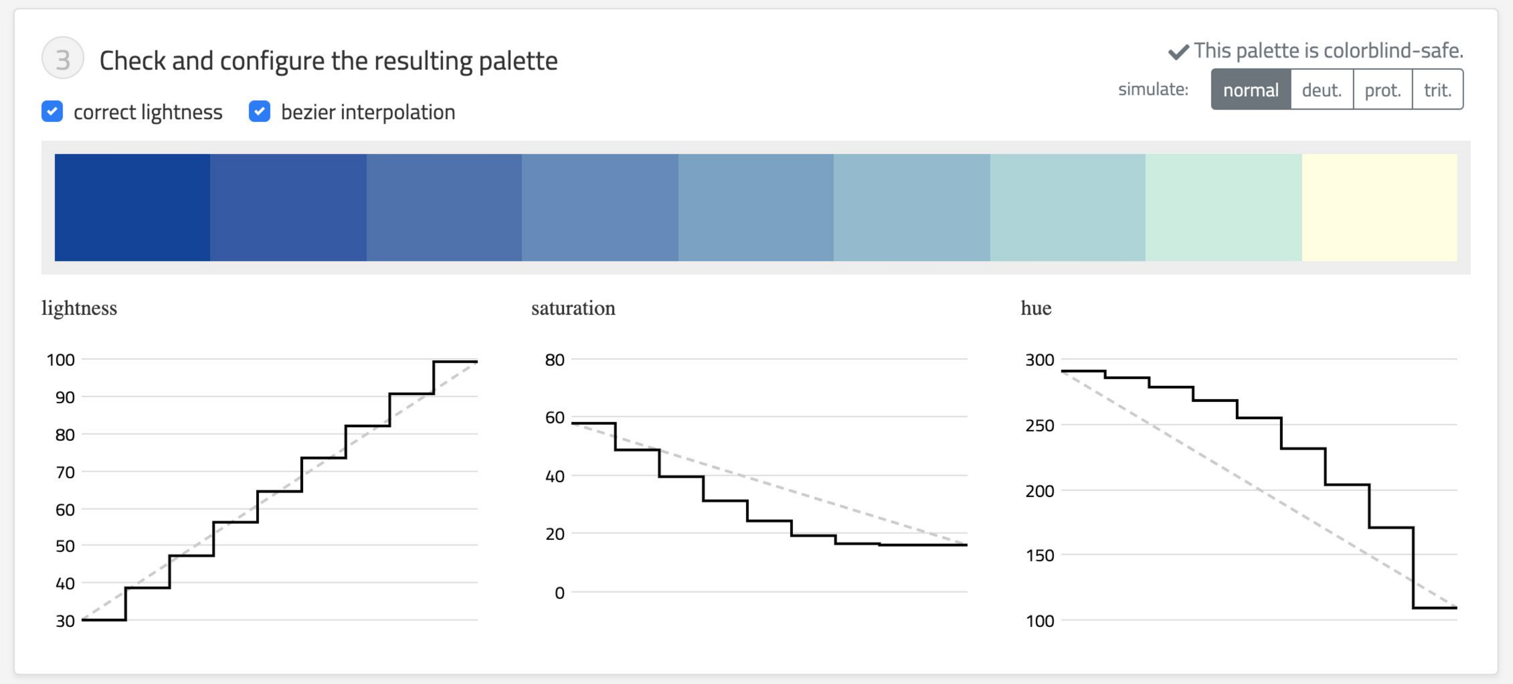


LIGHTNESS-BASED GRADIENT



Basics

Using color to represent **magnitudes**

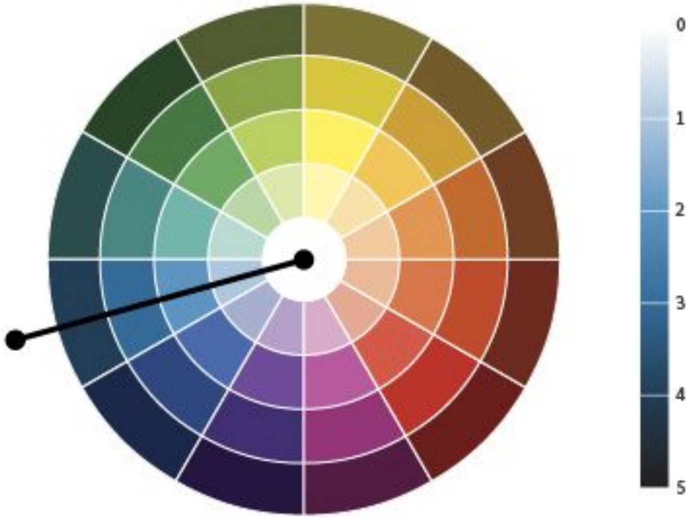


3 Check and configure the resulting palette

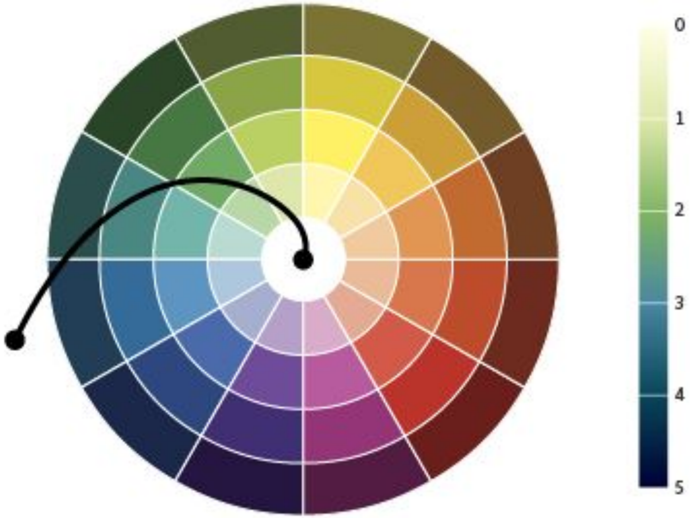
☑ correct lightness ☑ bezier interpolation

✓ This palette is colorblind-safe.

simulate: [✓ normal] deut. prot. trit.



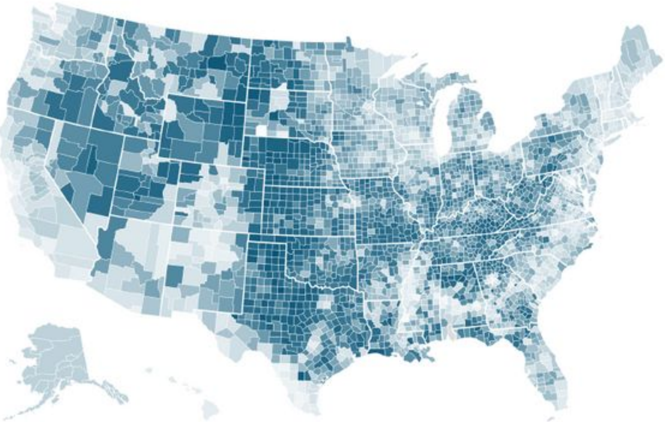
Single-Hue Sequential Color Scale



Multi-Hue Sequential Color Scale

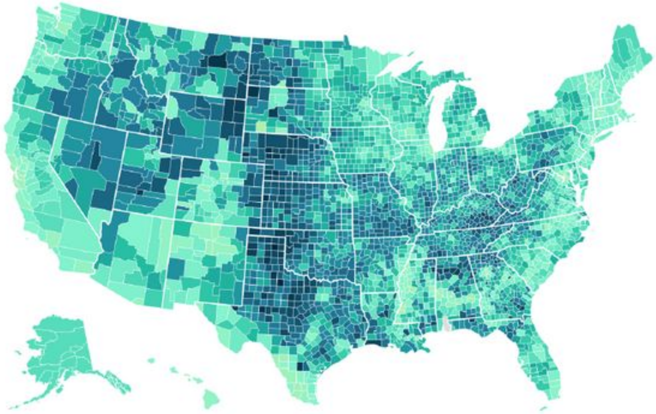
https://theblog.adobe.com/the-power-of-the-palette-why-color-is-key-in-data-visualization-and-how-to-use-it/

NOT SO BAD



ONE HUE

EVEN BETTER

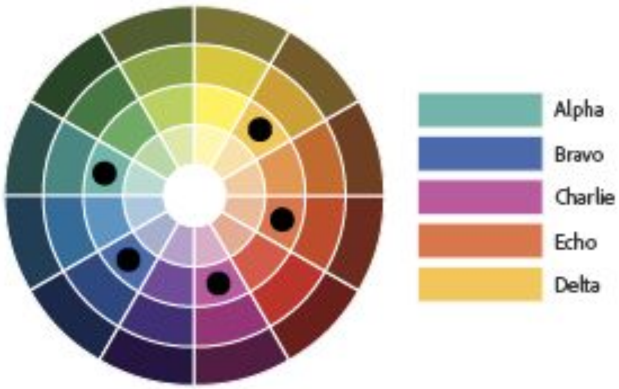


TWO HUES

https://blog.datawrapper.de/colors/

Basics

Using color to represent **categories**



The best of both worlds.

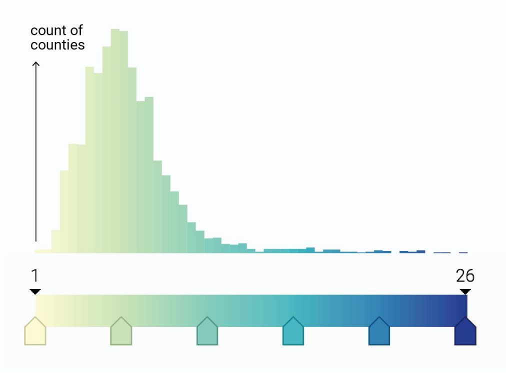
The basis

Partitions

- Differentiate colors is hard so you need to make differences more explicit

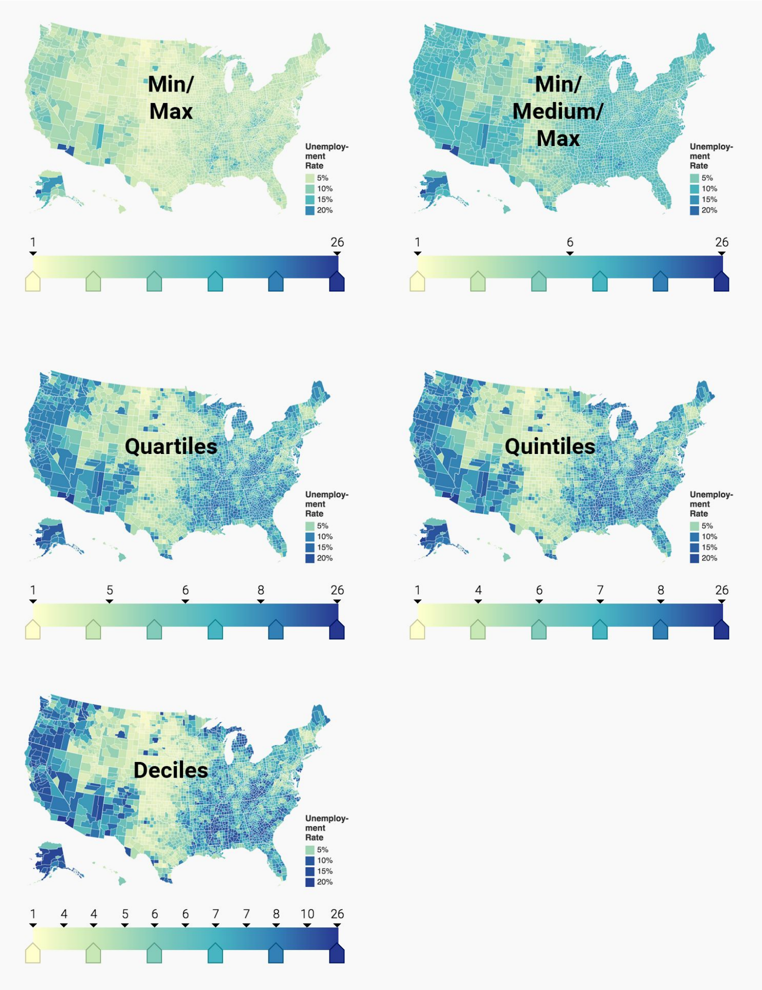
Basics

Partitions



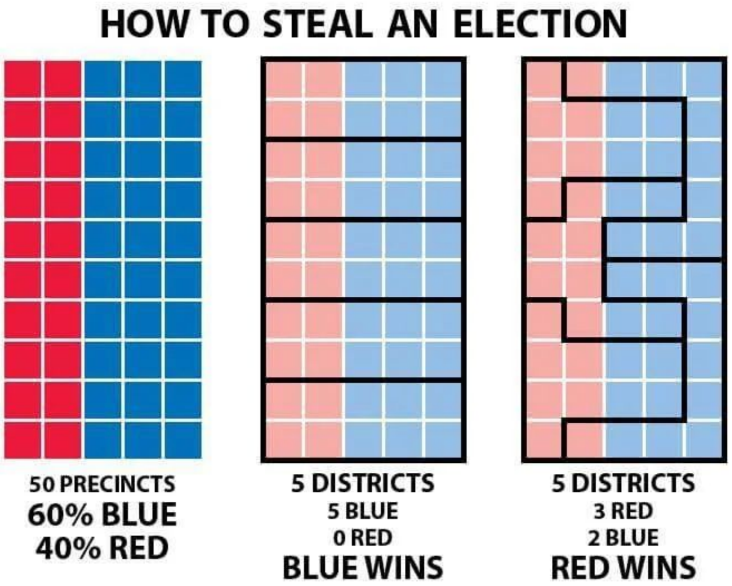
Basics

Partitions



What’s Data visualization?

Statistics + Visual representation



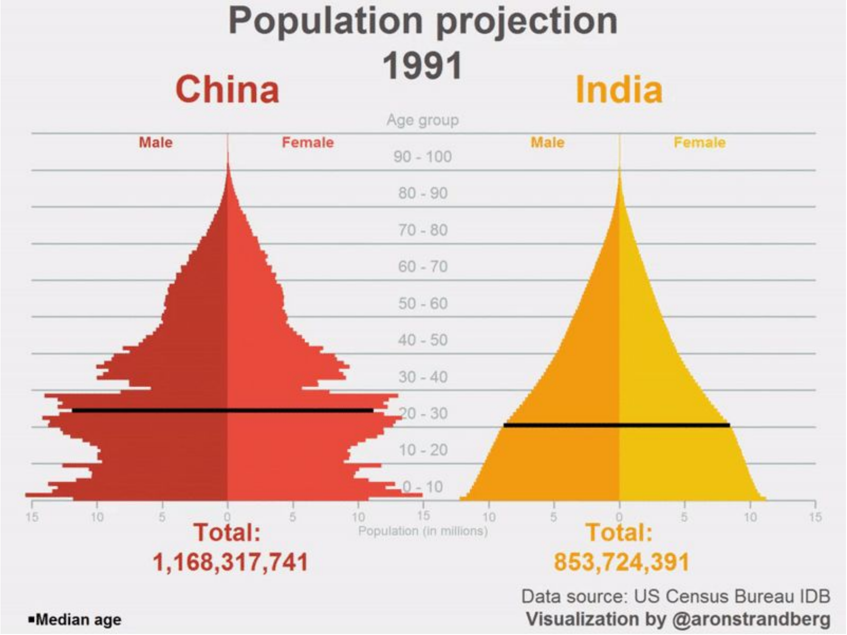
Color is art

No matter what stupid combination you choose there is always someone who made art with those colors

Live proof

What’s Data visualization?

Statistics + Visual representation



Ask yourself this question

What do I want to achieve with the visualization ?

- See global patterns (big data)
- Explore the data (big data, BI)
- Measure / make decisions (dashboards)
- Show results (scientist paper)
- Understand the data (data scientist)
- Tell a story (newspapers)
- Wow factor (newspapers, marketing, usually pretty shitty)

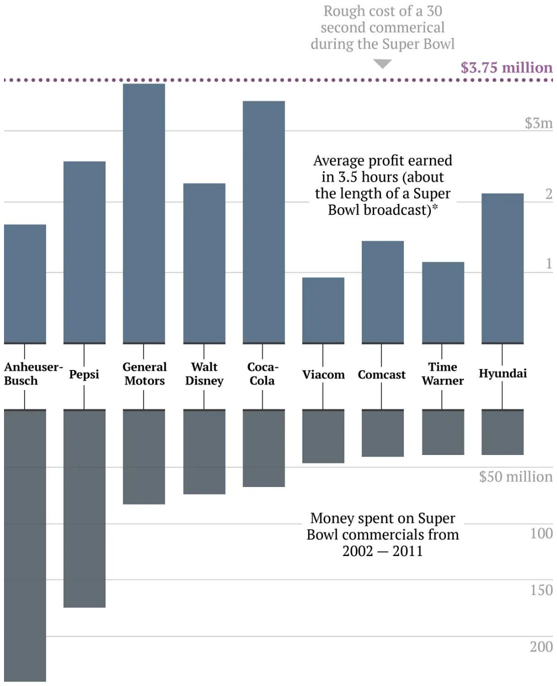
- “Business insights” -> BULLSHIT 99%, excel graphs 99%

03

Examples

Examples

Bar chart



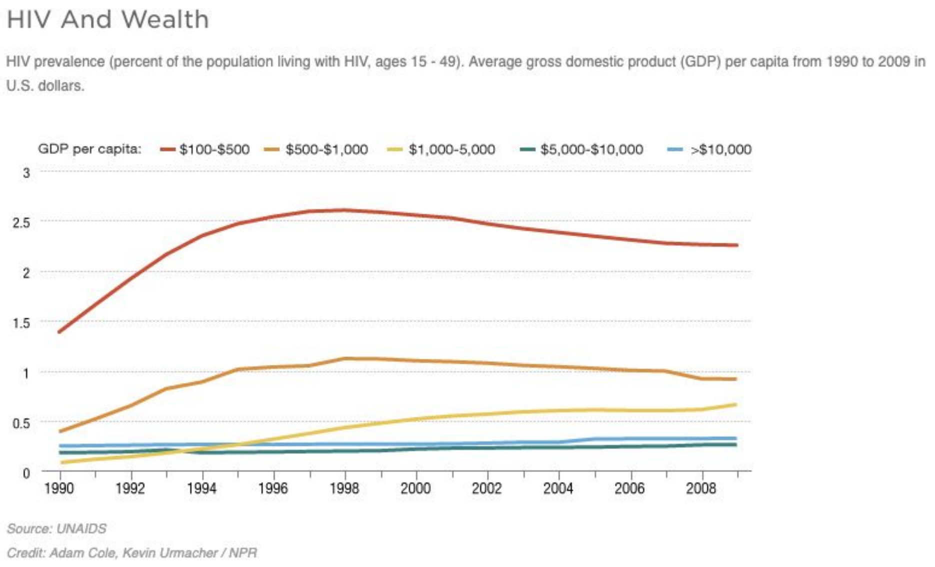
Note: this number is determined by taking the latest fiscal year earnings for each company, dividing by the number of hours in a year, then multiplying by 3.5.

Ritchie King | Quartz

Data: Compiled by Factset, Kantar Media, New York Times

Examples

Line chart



HIV And Wealth

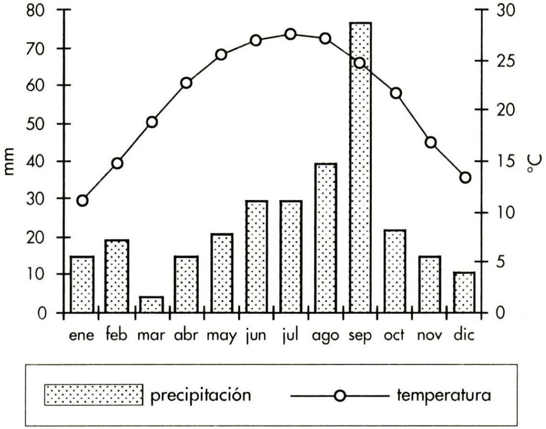
HIV prevalence (percent of the population living with HIV, ages 15 - 49). Average gross domestic product (GDP) per capita from 1990 to 2009 in U.S. dollars.

Source: UNAIDS

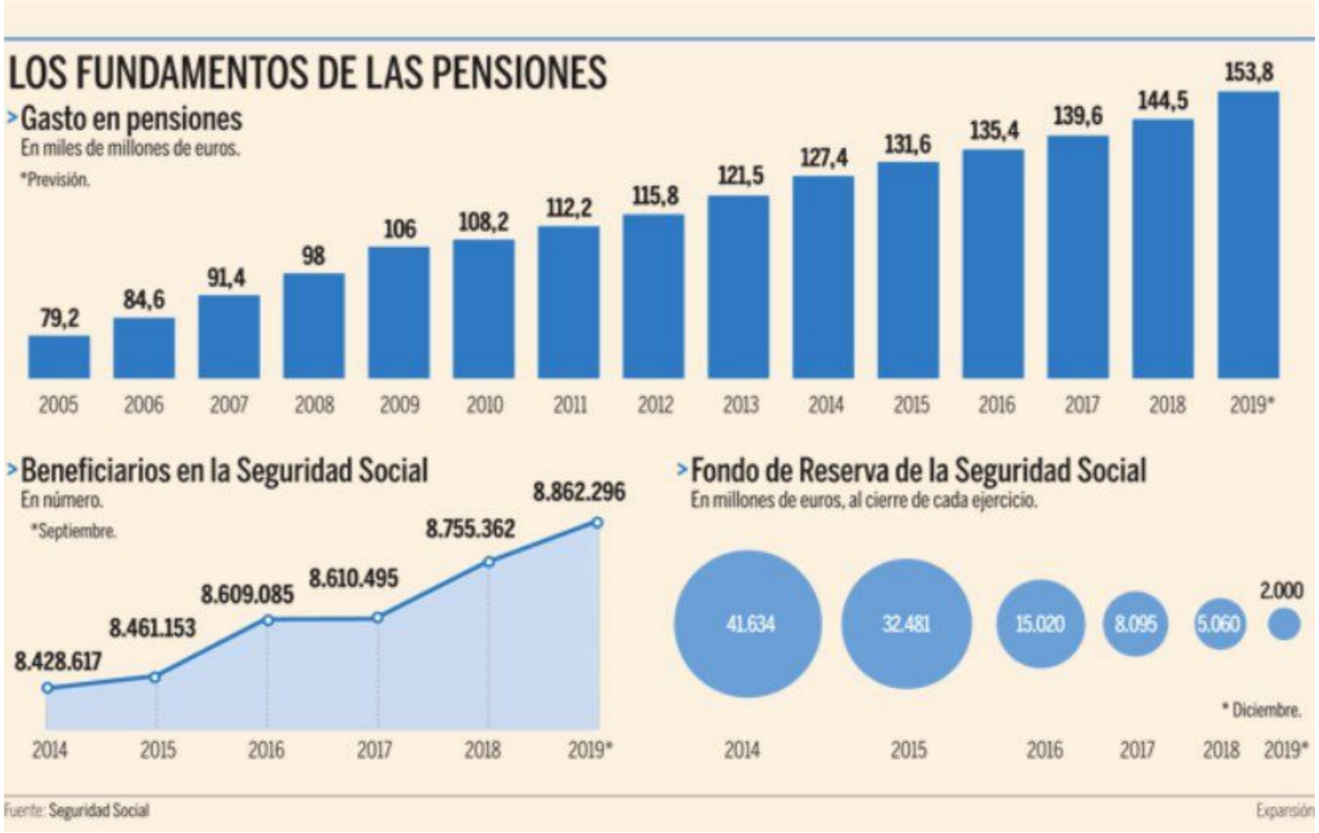
Credit: Adam Cole, Kevin Urmacher / NPR

Examples

Line + bar chart



No dejes que la realidad te estropee una buena visualización



LOS FUNDAMENTOS DE LAS PENSIONES

> Gasto en pensiones

En miles de millones de euros.

*Previsión.

> Beneficiarios en la Seguridad Social

En número.

*Septiembre.

> Fondo de Reserva de la Seguridad Social

En millones de euros, al cierre de cada ejercicio.

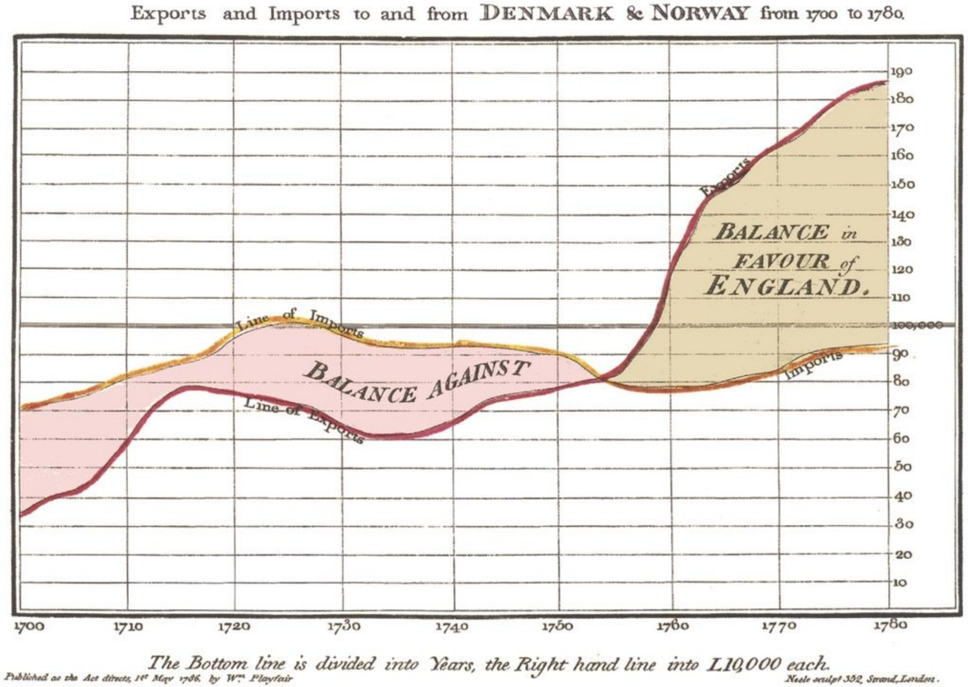
* Diciembre.

Fuente: Seguridad Social

Expansión

Examples

Area chart



Examples

Area chart + line

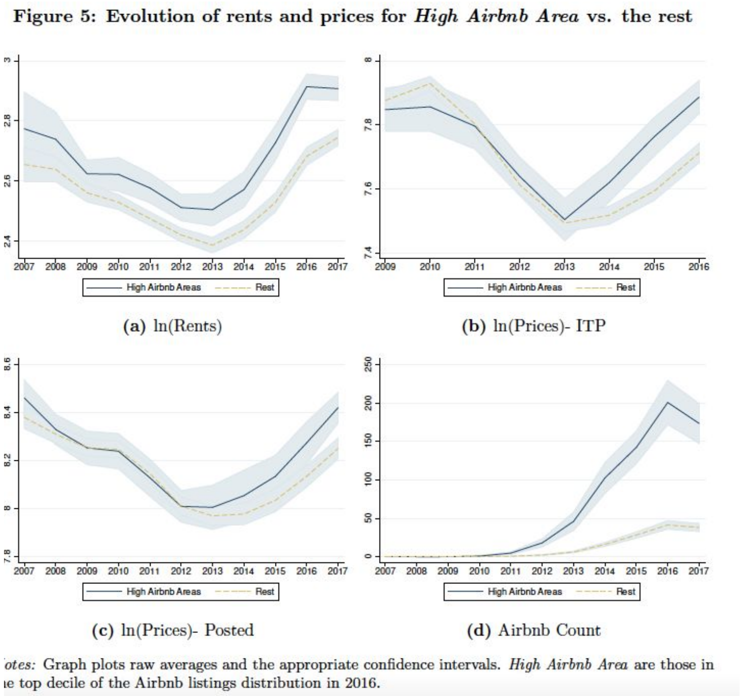


Figure 5: Evolution of rents and prices for *High Airbnb Area* vs. the rest

(a) ln(Rents)

(b) ln(Prices)- ITP

(c) ln(Prices)- Posted

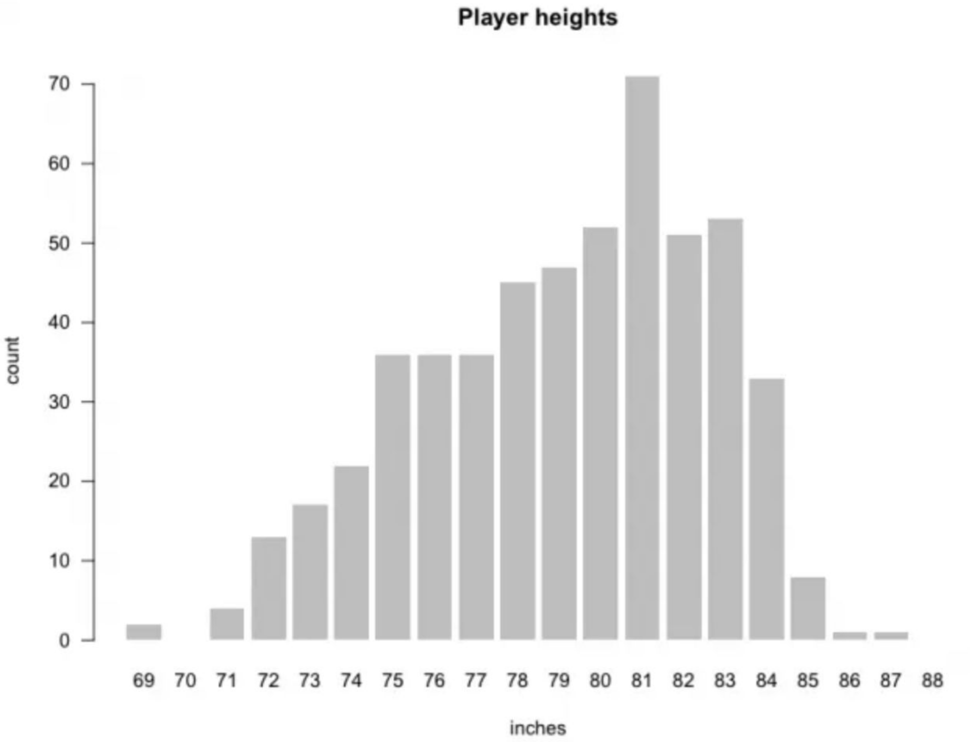
(d) Airbnb Count

otes: Graph plots raw averages and the appropriate confidence intervals. *High Airbnb Area* are those in e top decile of the Airbnb listings distribution in 2016.

https://nadaesgratis.es/admin/el-impacto-de-airbnb-en-el-mercado-de-vivienda-de-barcelona/attachment/191023-airbnb

Examples

Histogram

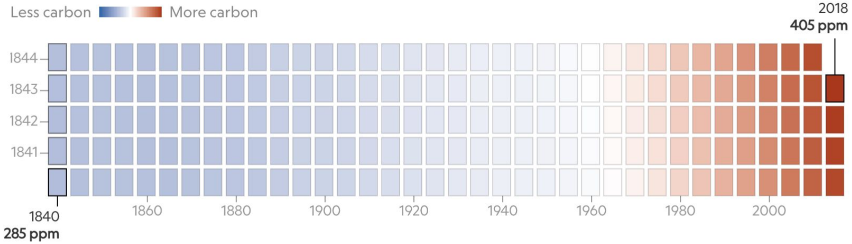


Examples

Heatmap (be careful with heatmaps)

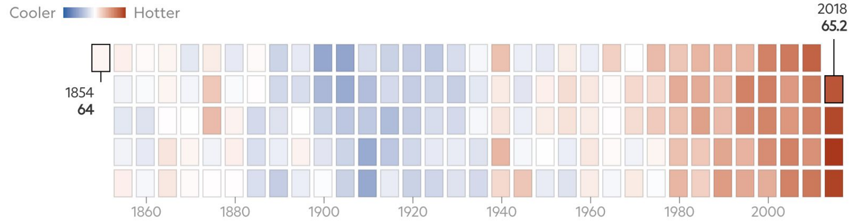
Since the Industrial Revolution, hundreds of billions of tonnes of **carbon have entered our atmosphere**.

Atmospheric carbon compared with a 20th-century average, parts per million



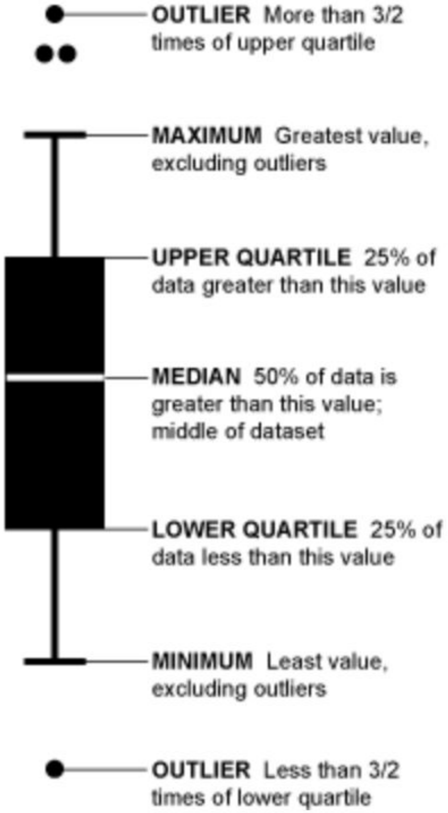
The **surface of the sea** is getting hotter.

Sea-surface temperature compared with a 20th-century average, degrees Fahrenheit



Examples

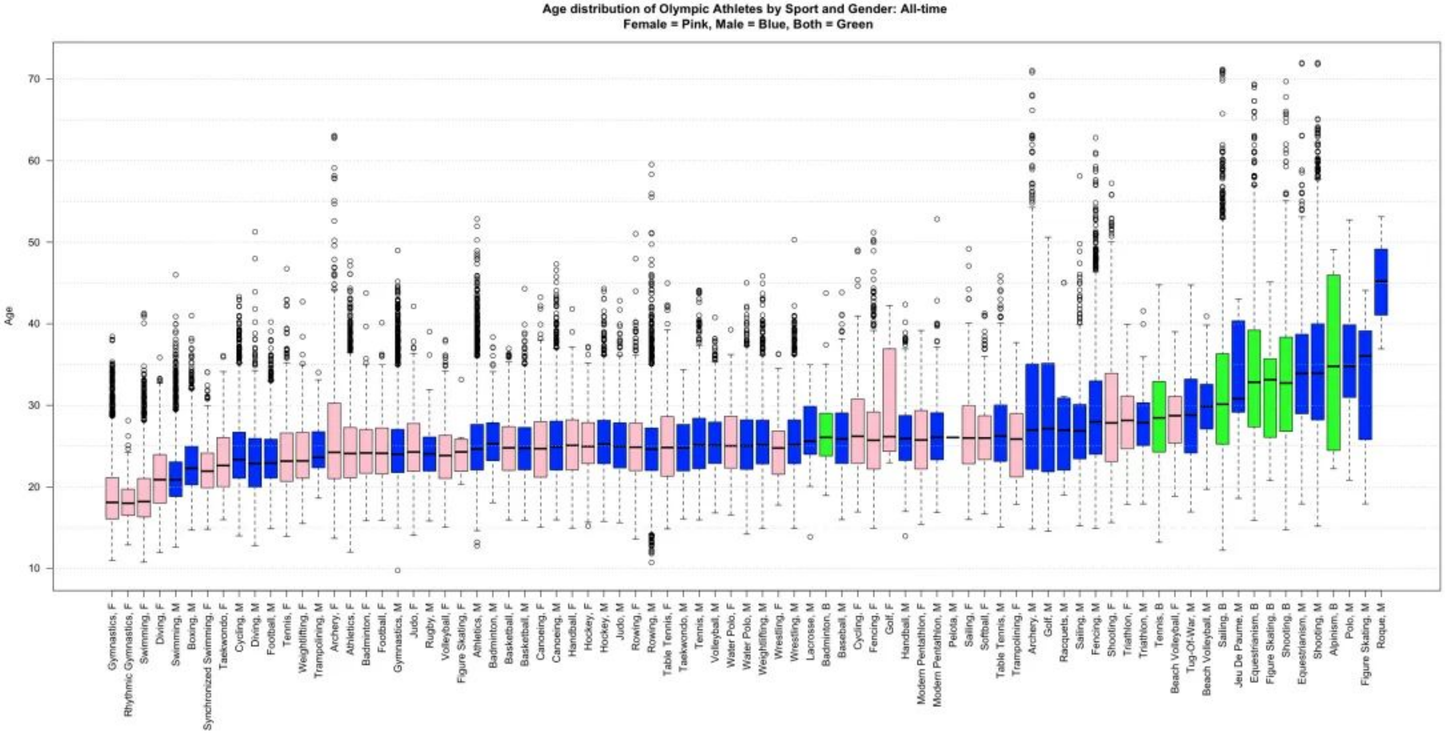
Boxplot



OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile

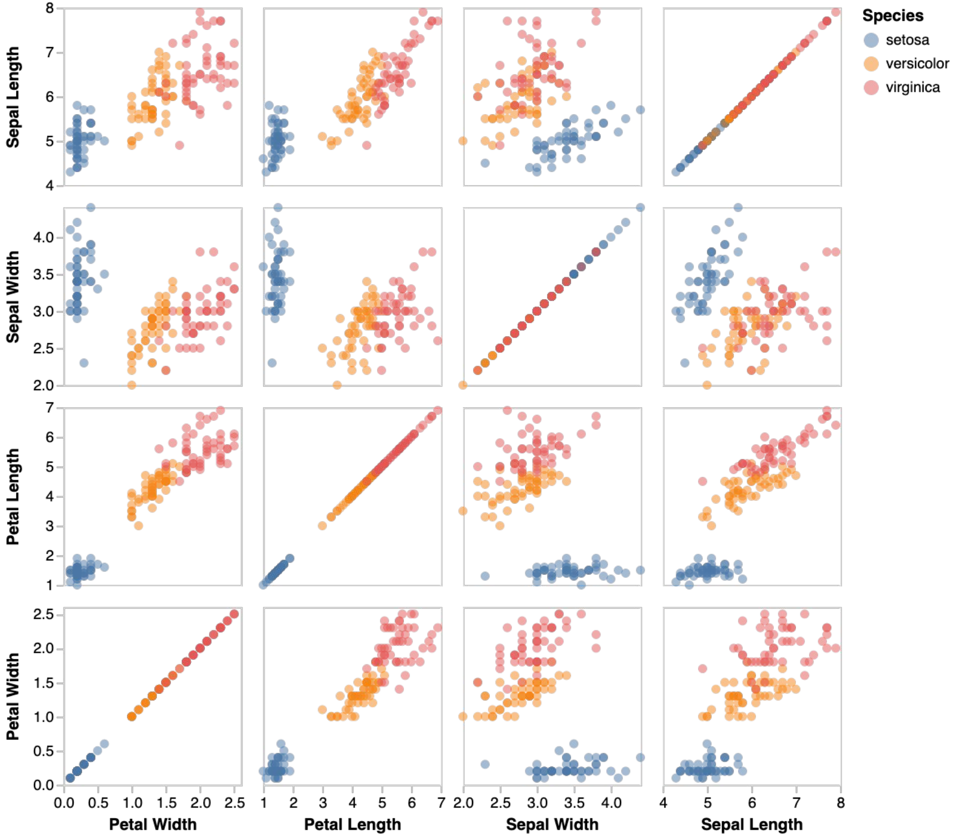
Examples

Boxplot

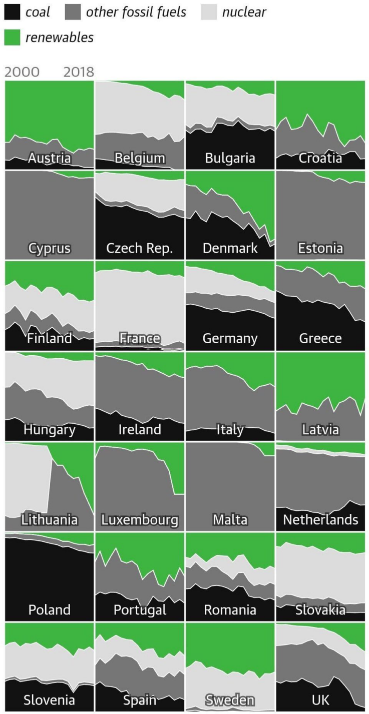


Examples

Scatterplot



Power generation by source (2000-2018)

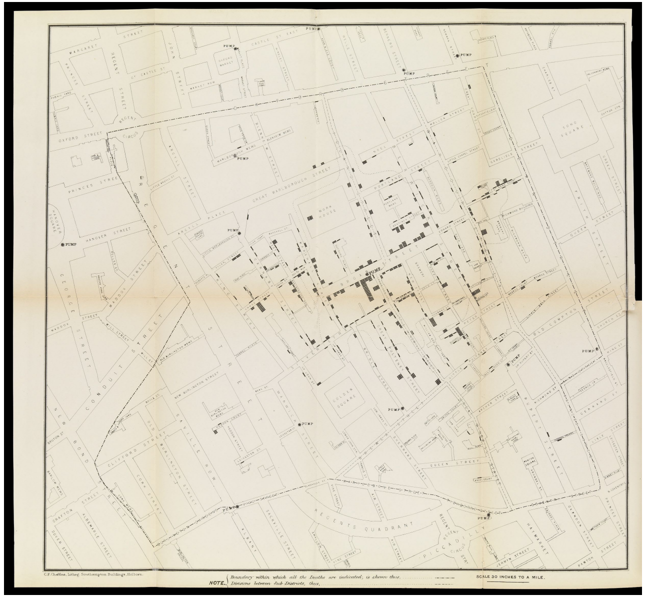


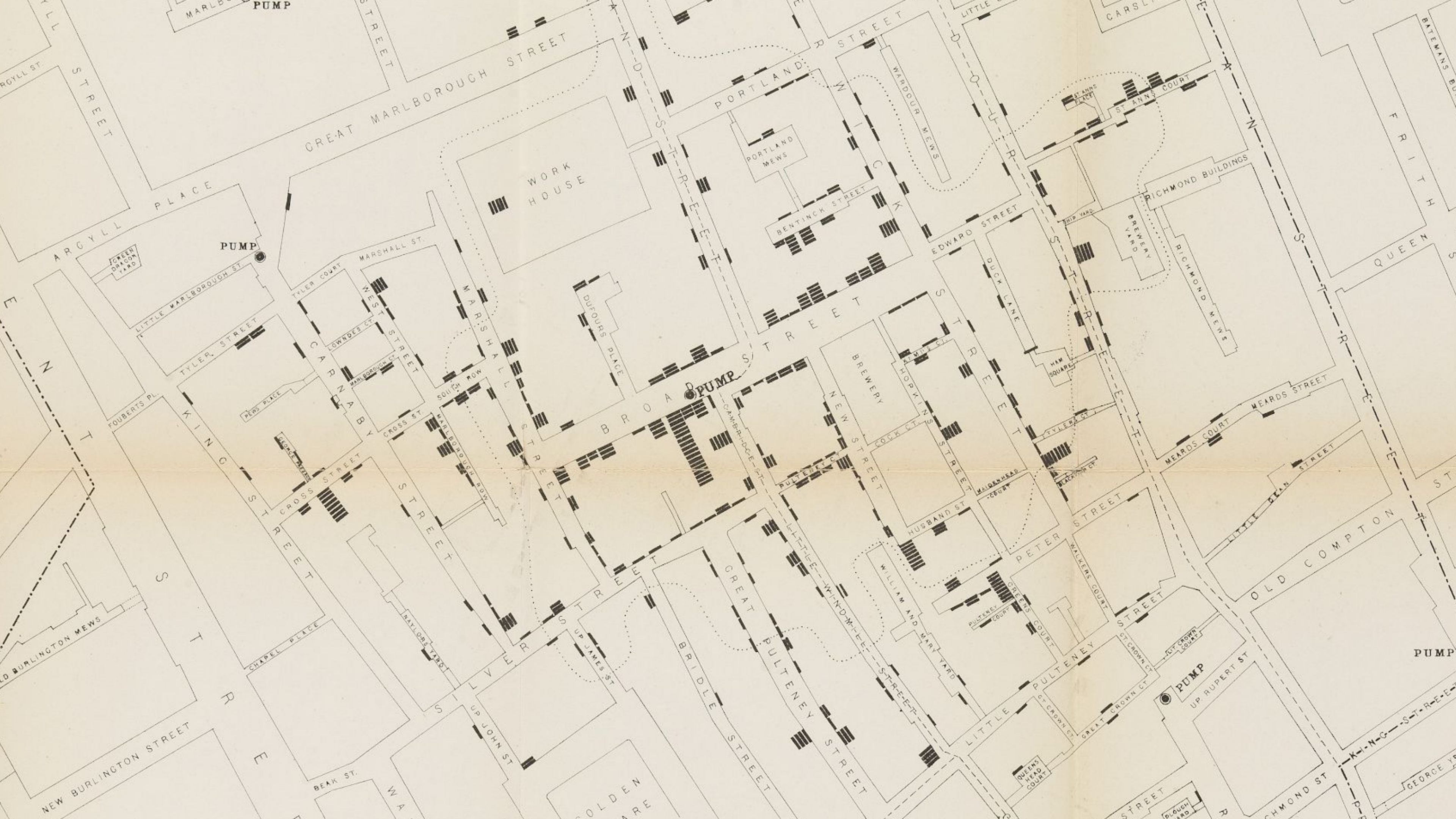
Source: Sandbag Climate Campaign

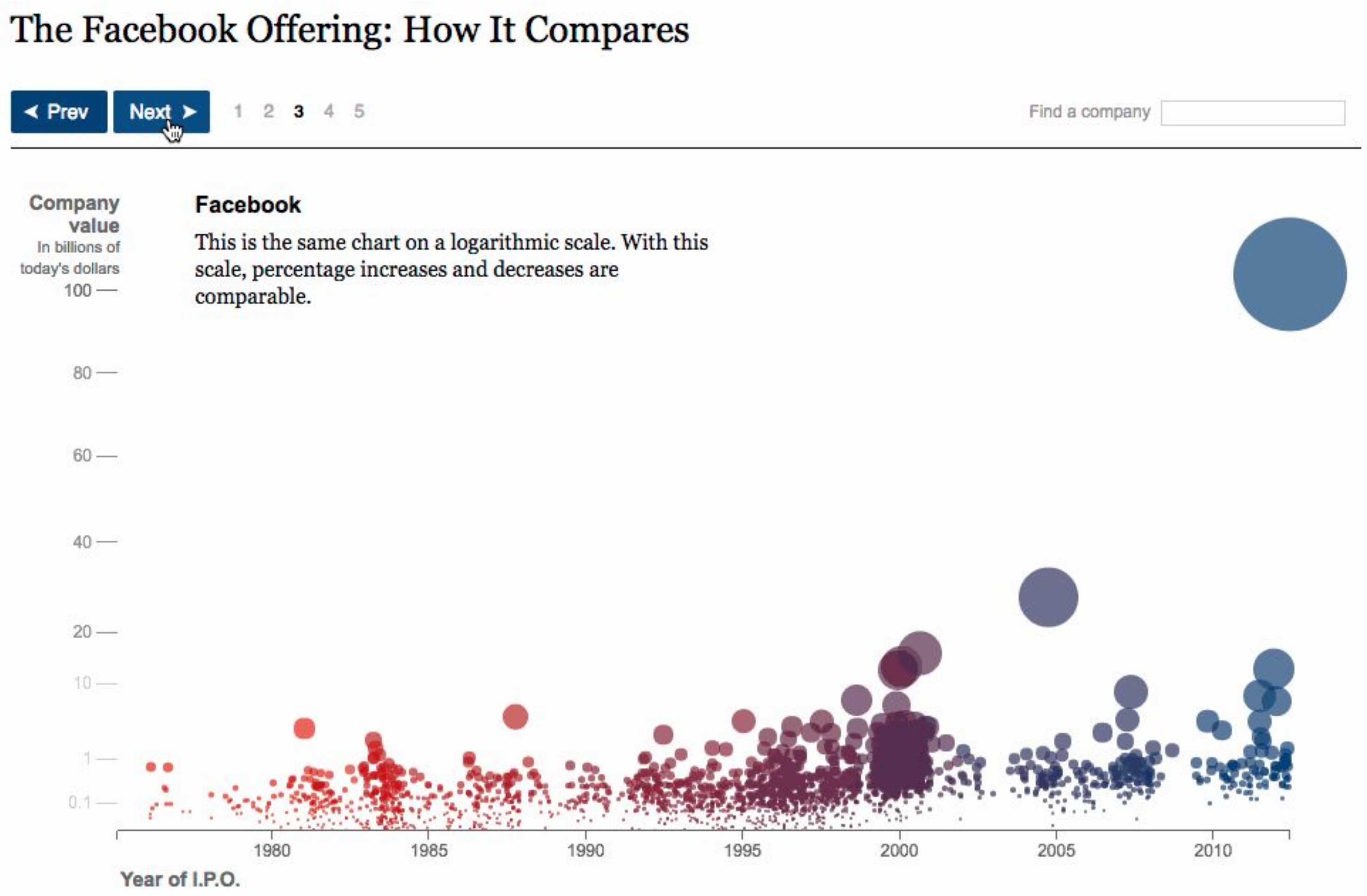
https://twitter.com/ghensel/status/1133350342819815426

Examples

Maps



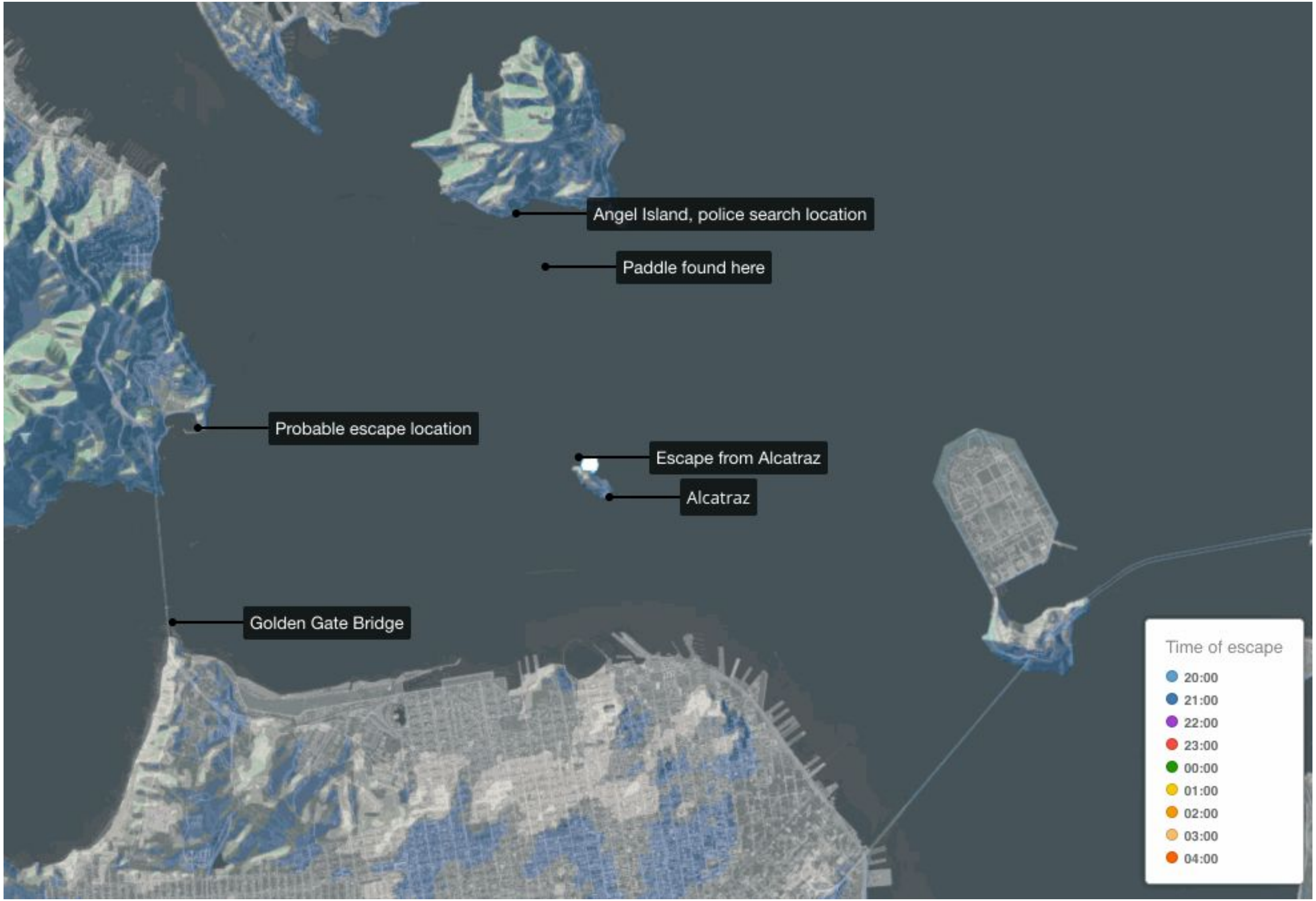

PUMP
GREAT MARLBOROUGH STREET
ARGYLL PLACE
ARGYLL ST.
PUMP
GREEN DRAGON YARD
LITTLE MARLBOROUGH ST.
TYLER STREET
TYLER COURT
MARSHALL ST.
WORK HOUSE
MARSHALL STREET
WEST STREET
LOWNDES CT.
MARLBOROUGH CT.
CARNABY STREET
SOUTH ROW
MEWS PLACE
FOUBERTS PL.
KING STREET
CROSS ST.
CROSS STREET
MARLBOROUGH ROW
DUFOURS PLACE
BROAD STREET
PUMP
CAMBRIDGE ST.
BERWICK STREET
PORTLAND STREET
PORTLAND MEWS
BENTINCK STREET
WARDOUR MEWS
EDWARD STREET
NEW STREET
BREWERY
HOPKINS STREET
KEMPS CT.
COCK CT.
PULTENEY CT.
MAIDENHEAD COURT
HUSBAND ST.
DUCK LANE
DEAN STREET
HAM SQUARE
TYLERS CT.
BLACKMAN CT.
ST ANNS PLACE
ST ANNS COURT
RICHMOND BUILDINGS
SHIP YARD
BREWERY YARD
RICHMOND MEWS
FRITH STREET
QUEEN STREET
MEARDS STREET
MEARDS COURT
LITTLE DEAN STREET
OLD COMPTON STREET
PETER STREET
WALKERS COURT
WILLIAM AND MARY YARD
GREENS COURT
PULTENEY COURT
LITTLE PULTENEY STREET
GT. CROWN CT.
GREAT CROWN CT.
LIT. CROWN COURT
PUMP
UP RUPERT ST
QUEENS HEAD COURT
KING STREET
PLOUGH YARD
LITTLE WINDMILL STREET
GREAT PULTENEY STREET
BRIDLE STREET
SILVER STREET
UP JAMES ST
UP JOHN ST
NAYLORS YARD
CHAPEL PLACE
BEAK ST.
GOLDEN SQUARE
NEW BURLINGTON STREET
OLD BURLINGTON MEWS
REGENT STREET
BATEMANS BUILDINGS
CARLISLE STREET
LITTLE CHAPEL STREET
GEORGE YARD
RICHMOND ST
PUMP



The Facebook Offering: How It Compares

Facebook

This is the same chart on a logarithmic scale. With this scale, percentage increases and decreases are comparable.


Angel Island, police search location
Paddle found here
Probable escape location
Escape from Alcatraz
Alcatraz
Golden Gate Bridge
Time of escape
● 20:00
● 21:00
● 22:00
● 23:00
● 00:00
● 01:00
● 02:00
● 03:00
● 04:00

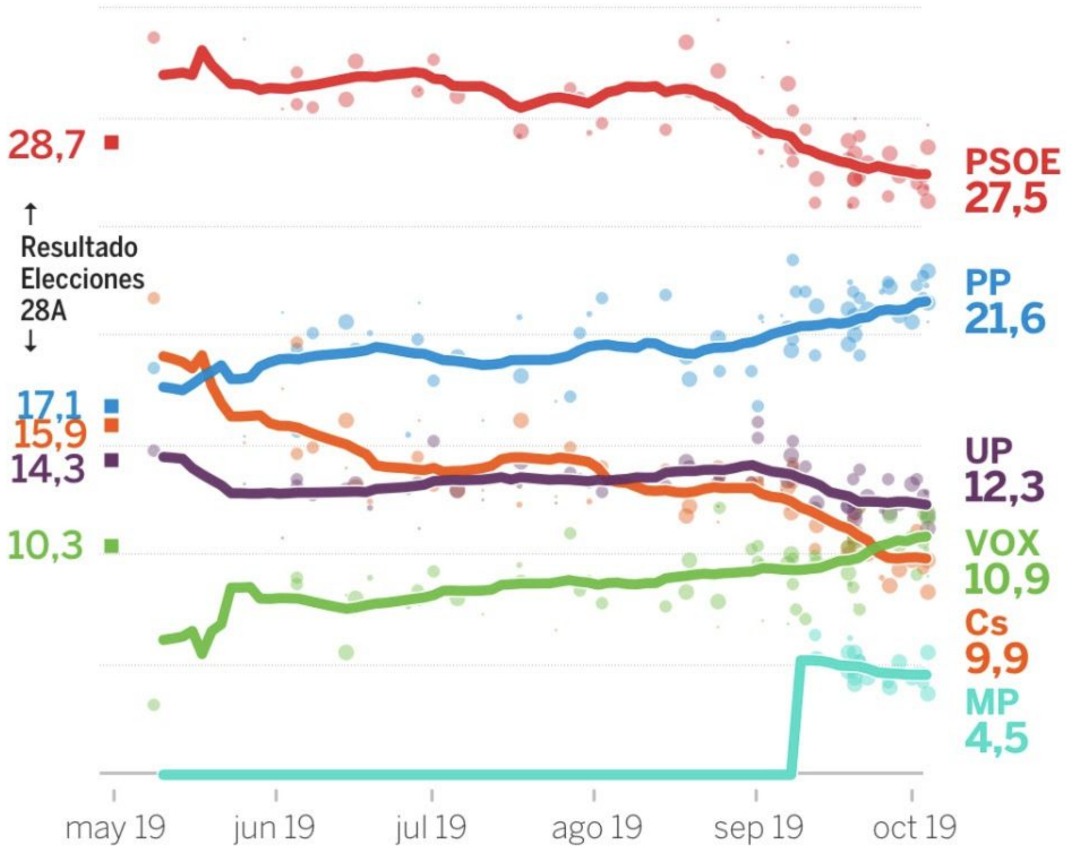
Examples

More animated maps

https://flowingdata.com/2015/12/15/a-day-in-the-life-of-americans/
http://javisantana.com/kotar/
https://team.carto.com/u/javi/me
https://javi.carto.com/builder/870c41d2-fc35-11e3-83a9-0edbca4b5057/embed

En realidad puedes hacer lo que quieras

Estimación de voto según cada encuesta,
y **promedio** a partir de todas ellas



Fuente: @kikollaneras

Elaboración propia / EL PAÍS

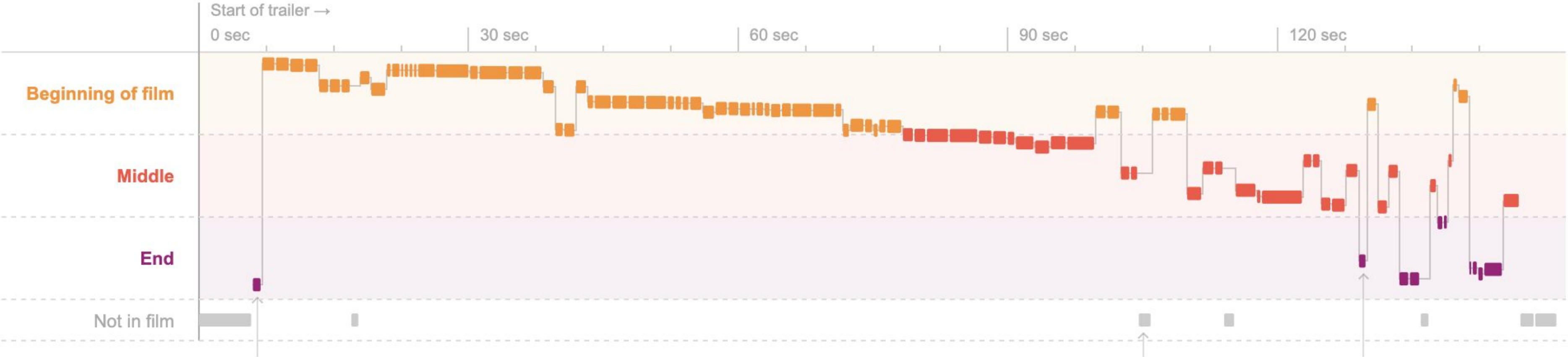
http://www.evolutionoftheweb.com/

Dissecting a Trailer: The Parts of the Film That Make the Cut

How scenes from five of the nine best picture nominees were reassembled to promote the films.

Silver Linings Playbook

"Silver Linings Playbook" follows the standard model for trailers, according to Bill Woolery, a trailer specialist in Los Angeles who once worked on trailers for movies like "The Usual Suspects" and "E.T. the Extra-Terrestrial." While introducing the movie's story and its characters, the trailer largely follows the order of the film itself.



The trailer's opening shot — an image of the family's home — appears near the end of the film, but there are similar shots near the beginning of the movie.

A handful of very short shots are never seen in the film, although most are shown from alternate camera angles.

Shots that accompany the main actors' names are also shown out of order.

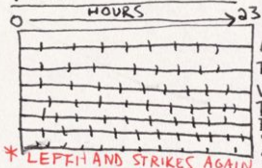
https://archive.nytimes.com/www.nytimes.com/interactive/2013/02/19/movies/awardsseason/oscar-trailers.html

DEAR DATA - WEEK 32

A WEEK OF SOUNDS

ABOUT THE DATA: EVERY HOUR I WAS AWAKE I TRACKED THE SOUNDS I HEARD AROUND ME (GENERALLY ON THE TOP OF THE HOUR, SOMETIMES LATER IF I FORGOT)

HOW TO READ IT:



* LEFTHAND STRIKES AGAIN

FOR CONTEXT:

M: ON TRAIN FROM GLASGOW
T: WORKING IN STUDIO
W: ↓ "
Th: ↓ "
F: WORKING AT HOME / OUT FOR DRINKS
S: AT HOME / ON HIGH STREET / OUT FOR DRINKS
S: AT HOME / CYCLING 30 MILES IN COUNTRYSIDE!

EACH 🟦 REPRESENTS ONE SOUND.

SOUND TYPES ARE ORGANISED AS FOLLOWS:

① 'ORGANIC' SOUNDS (SOUNDS CREATED BY PEOPLE, ANIMALS, OR NATURE)

- 🟪 A SOUND MY HUSBAND MADE: SPEAKING, DRINKING COFFEE, HUMMING, FIXING BIKE, ETC.*
- 🟦 PEOPLE'S SOUNDS: MOVEMENT: FOOTSTEPS, EATING, PUSHING CHAIRS BACK (SOUNDS OF ACTIVITY)
- 🟩 PEOPLE'S VOICES
- 🟧 RUSTLING, SHUFFLING NEWSPAPERS
- 🟩 CLANKING BOTTLES + CUTLERY
- 🟩 CRASHES, BANGS, + RATTLES
- 🟧 ROLLING WHEELS: SUITCASES, CARTS, ETC.
- 🟥 BIRDSONG
- 🟦 WIND IN TREES

② 'MACHINE' SOUNDS:

- 🟧 RUNNING WATER
- 🟩 LAPTOP
- 🟦 NEIGHBOURS' LOUD TV + MUSIC
- 🟥 APPLIANCES RUNNING: RADIATOR / BOILER, WASHING MACHINE, POWER TOOL, COFFEE MACHINE, ETC.
- ⬜ OUR FILM / TV
- 🟩 TRAINS
- 🟩 MOTOR VEHICLE
- 🟧 AIRPLANE
- 🟧 THE HUM OF ELECTRICITY
- 🟦 RECORDED MUSIC

③ UNUSUAL SOUNDS: ⬛ STEEL DRUM BAND, ARCADE GAMES, CHURCH BELLS, A HORSE! :)

* ALSO INCLUDES SOME RIDICULOUSLY STUPID SOUNDS HE WAS MAKING TO MESS W/ THE DATA!

FROM:
S POSAVEC
[redacted]
LONDON
[redacted]
UK

£1.00

LONDON EC1A 1BB
APR 2

TO:
GIORGIA LUPI
[redacted]
BROOKLYN, NY [redacted]
USA

BY AIR MAIL
par avion
Royal Mail®

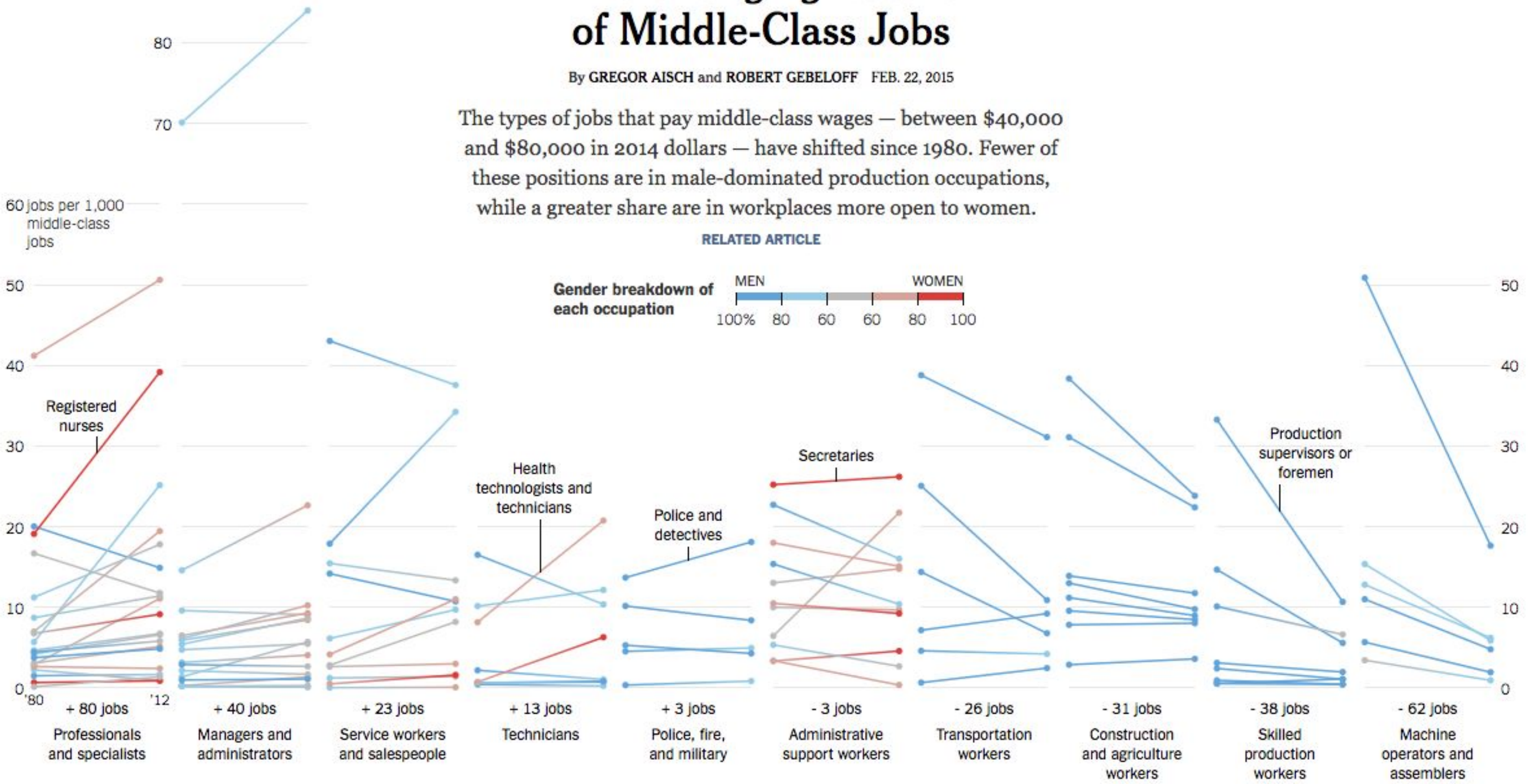
http://www.stefanieposavec.com/dear-data-about

The Changing Nature of Middle-Class Jobs

By **GREGOR AISCH** and **ROBERT GEBELOFF** FEB. 22, 2015

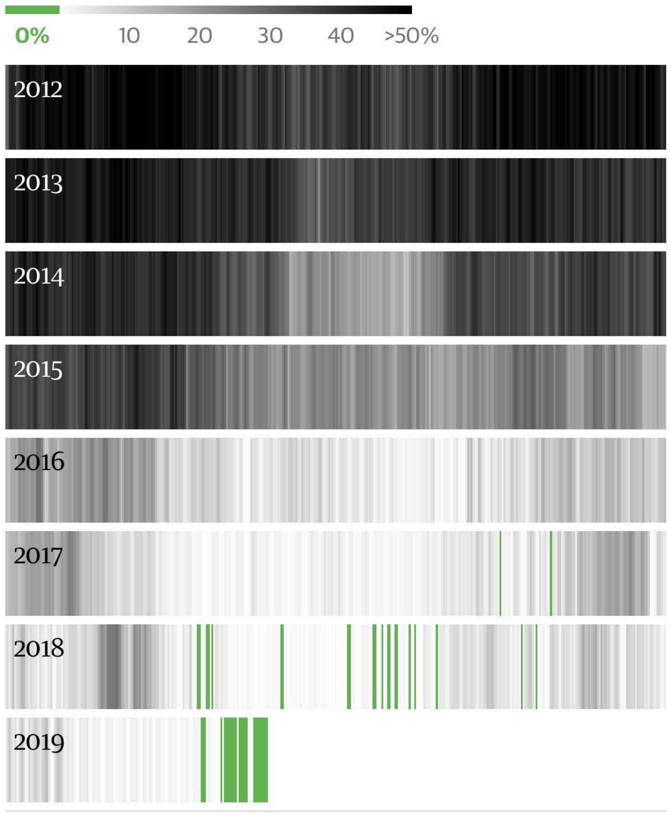
The types of jobs that pay middle-class wages — between $40,000 and $80,000 in 2014 dollars — have shifted since 1980. Fewer of these positions are in male-dominated production occupations, while a greater share are in workplaces more open to women.

RELATED ARTICLE

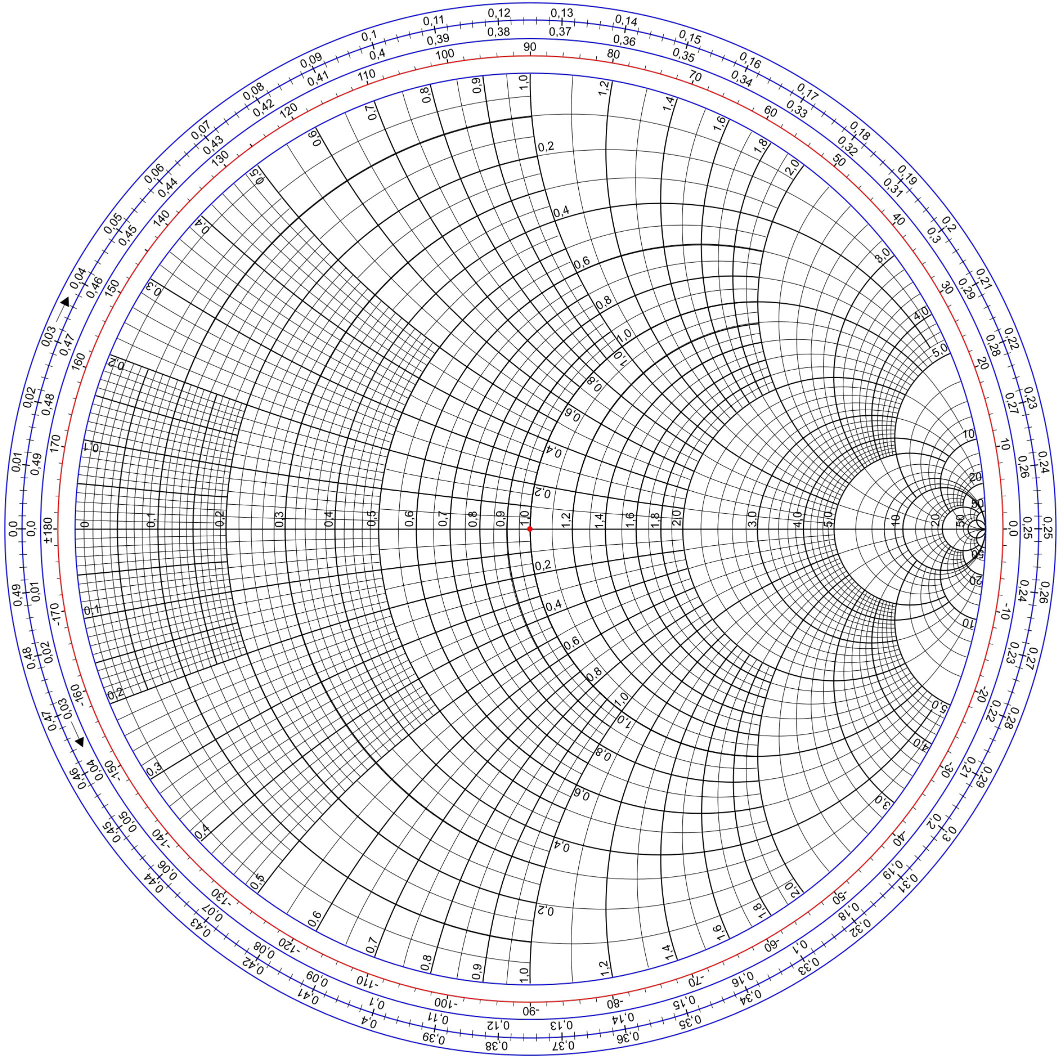


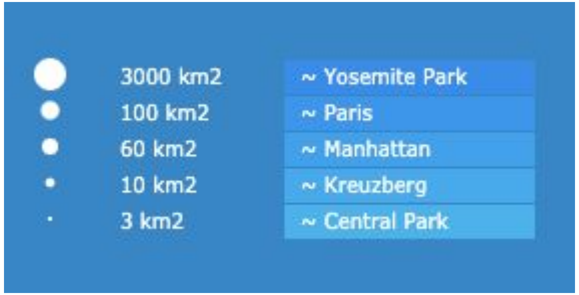
Britain is rapidly phasing out coal

Daily share of Britain's power generated by burning coal

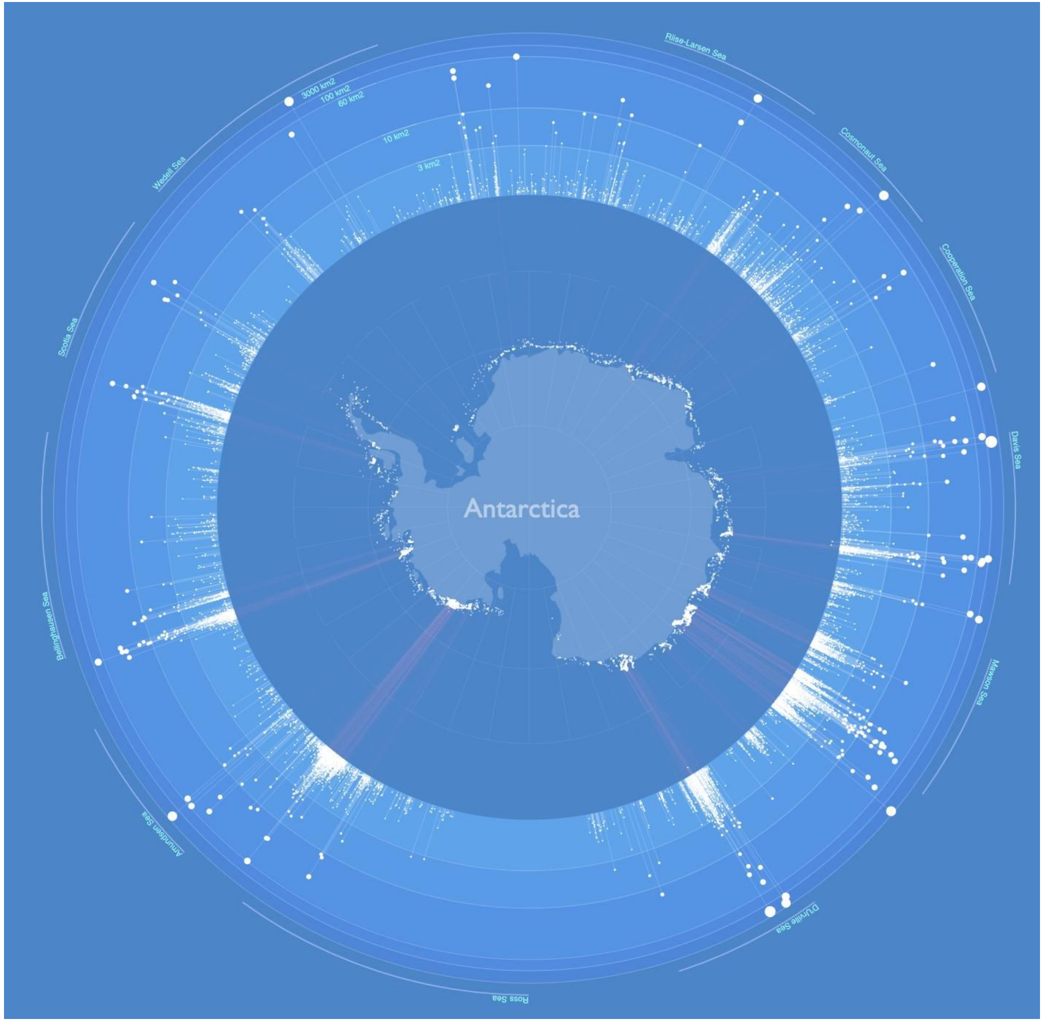


https://twitter.com/EmmaFidler/status/1132347203031326722

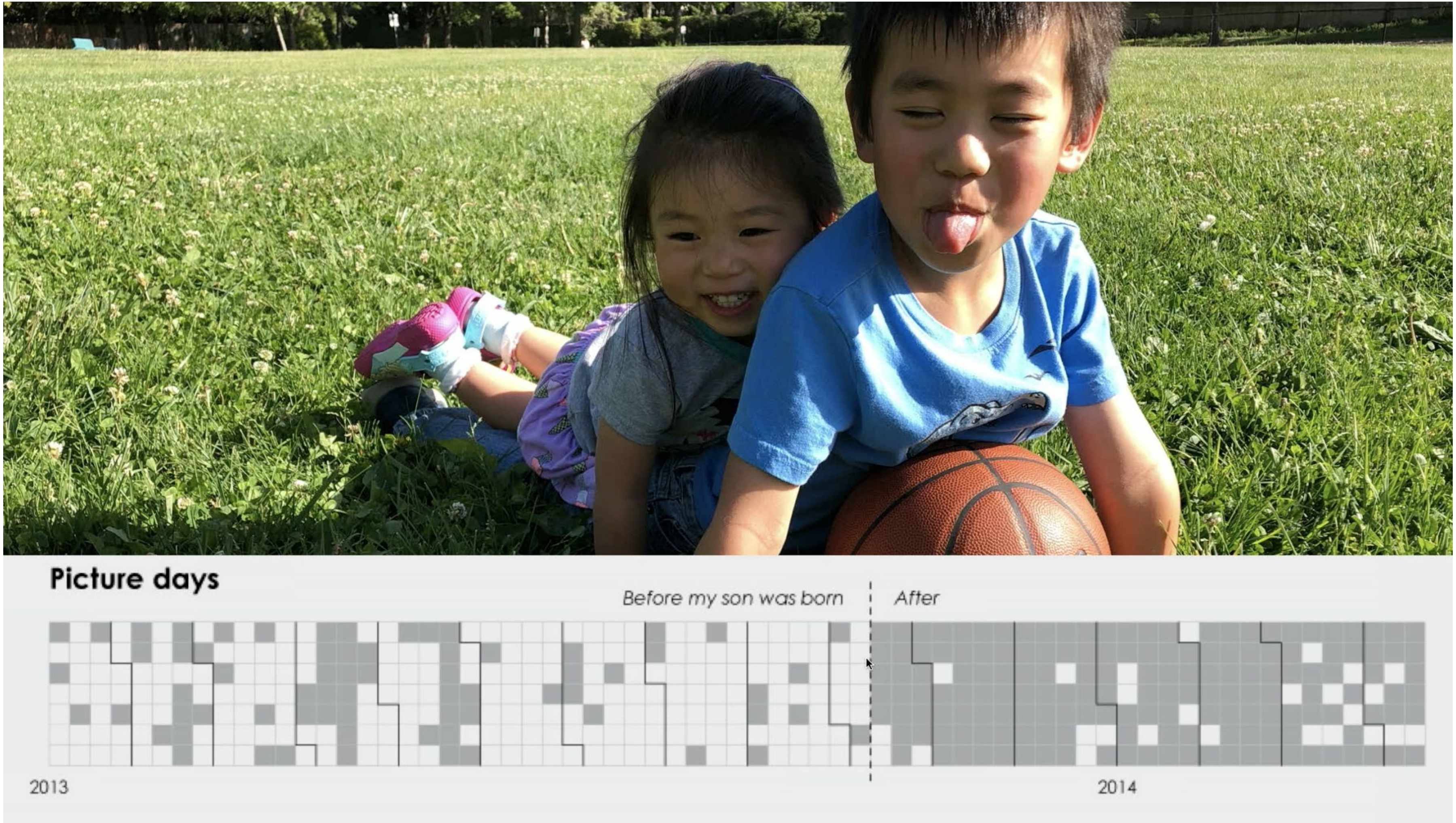

0,0
0,01
0,02
0,03
0,04
0,05
0,06
0,07
0,08
0,09
0,1
0,11
0,12
0,13
0,14
0,15
0,16
0,17
0,18
0,19
0,2
0,21
0,22
0,23
0,24
0,25
0,26
0,27
0,28
0,29
0,3
0,31
0,32
0,33
0,34
0,35
0,36
0,37
0,38
0,39
0,4
0,41
0,42
0,43
0,44
0,45
0,46
0,47
0,48
0,49
±180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1,0
1,2
1,4
1,6
1,8
2,0
3,0
4,0
5,0
10
20
50



●	3000 km2	~ Yosemite Park
●	100 km2	~ Paris
●	60 km2	~ Manhattan
•	10 km2	~ Kreuzberg
·	3 km2	~ Central Park



https://twitter.com/Lagomorpho/
0384



Picture days

04

Conclusions

Me
Best chart for the data
Fancy dataviz

https://twitter.com/rebellenoire/status/1181909242204295169

Conclusions

Tufte

- Less is usually more (ink-data ratio)
- Graphical excellence consists of complex ideas communicated with clarity, precision, and efficiency
- Keep it proportional! “Lie Factor = *size of effect shown in graphic* divided by *size of effect in data*”
- You don’t have to use a graphic when there isn’t much data — a table is often better.
- Pie charts are useless. Period. DO NOT USE PIE CHARTS.

Conclusions

Santana

- La visualización de datos normalmente sirve para saber qué preguntas hacer, no para resolverlas
- “All non-trivial abstractions, to some degree, are leaky” (Joel)
- No hay una forma buena de visualizar algo
- No te dejes engañar por las pintas de una visualización
- A hacer visualización se aprende haciendo visualizaciones (y copiando y probando)
- No quieres ser este tipo de persona:

> “No dejes que la realidad te estropee una buena visualización”

05.

Must reads

Must reads

- Visual display of quantitative information (Tufte)
- @flowingdata
- @MKrzywinski
- @lisacrost
- Dataviz project https://datavizproject.com
- https://nadaesgratis.es data analysis + vis

06.

HANDS ON! (Santana’s way)

Before start

About the visualization

This visualization could be done with a library in a few minutes, we are trying to show how easy is to visualize data with no tools.

It might be too easy for experienced developers. The code is not the important part.

Use case

NO2 in Madrid

- Lots of public data in a reusable format
- This is a big problem, we are all breathing this thing right now (72 ug/m^3 today)

Use case

The process

- Get Data -> Visualize

Use case

The process - the data

- Understand the source data: format, meaning and so on
- Understand the actual data: histograms, max, min...
- Clean the data
- Prepare the data to be visualized
  - Depending on the data size this step is mandatory

Use case

Some [random] data tools

- Cleaning: Trifacta // Open refine // Python // R // ...
- Prepare data: Any database SQL database (Postgres) / CSVKit / Python...
- Understand the data: excel // pandas...
- Serve: S3 // github // datasette // carto // tinybird

Santana's choice:

- Python // clickhouse // vega (tinybird)

Use case

The process - the visualization

- Decide what we want to show
- Go for a visualization type, remember
    - See global patterns
    - Explore the data
    - Measure / make decisions
    - Tell a story
    - Wow factor
- Iterate 100 times

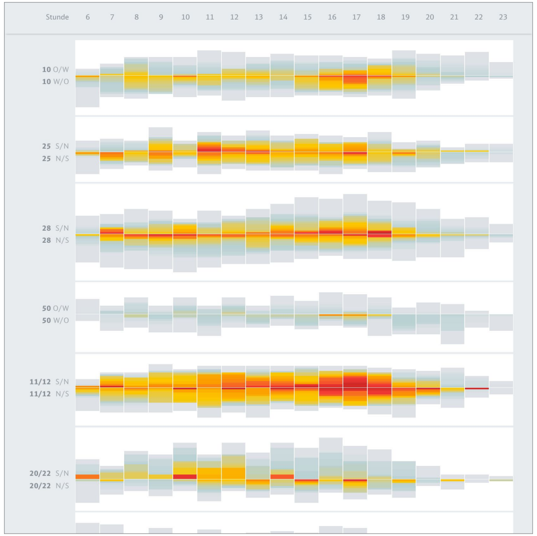
Use case

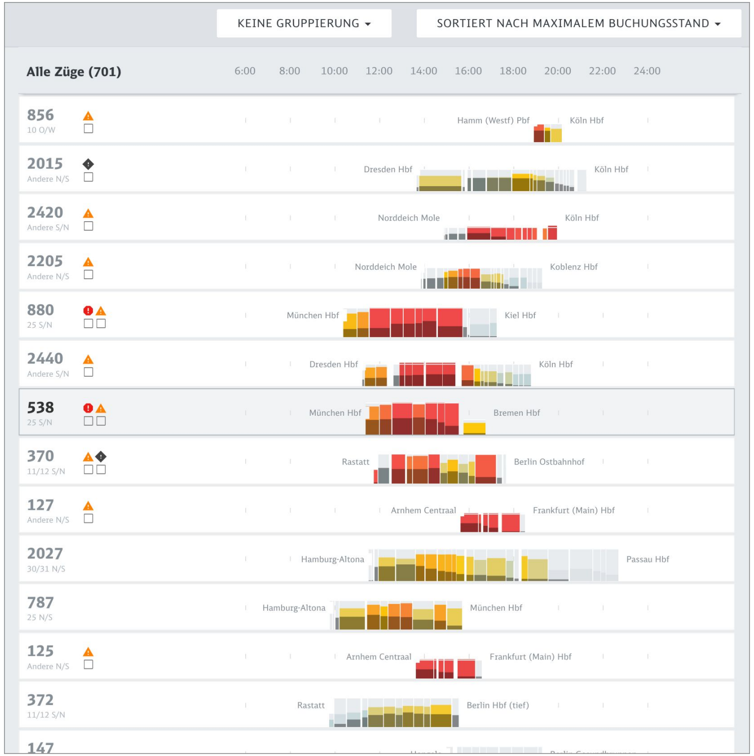
Some [random] vis tools

- Vega, D3, observable
- Pandas (mathplotlib)
- Datastudio/tableu/qlik/....

Santana’s choice:

- Vanilla js, d3, pandas





Alle Züge (701)

856 10 O/W ⚠ □ Hamm (Westf) Pbf Köln Hbf

2015 Andere N/S ◆ □ Dresden Hbf Köln Hbf

2420 Andere S/N ⚠ □ Norddeich Mole Köln Hbf

2205 Andere N/S ⚠ □ Norddeich Mole Koblenz Hbf

880 25 S/N ❗⚠ □□ München Hbf Kiel Hbf

2440 Andere S/N ⚠ □ Dresden Hbf Köln Hbf

538 25 S/N ❗⚠ □□ München Hbf Bremen Hbf

370 11/12 S/N ⚠◆ □□ Rastatt Berlin Ostbahnhof

127 Andere N/S ⚠ □ Arnhem Centraal Frankfurt (Main) Hbf

2027 30/31 N/S Hamburg-Altona Passau Hbf

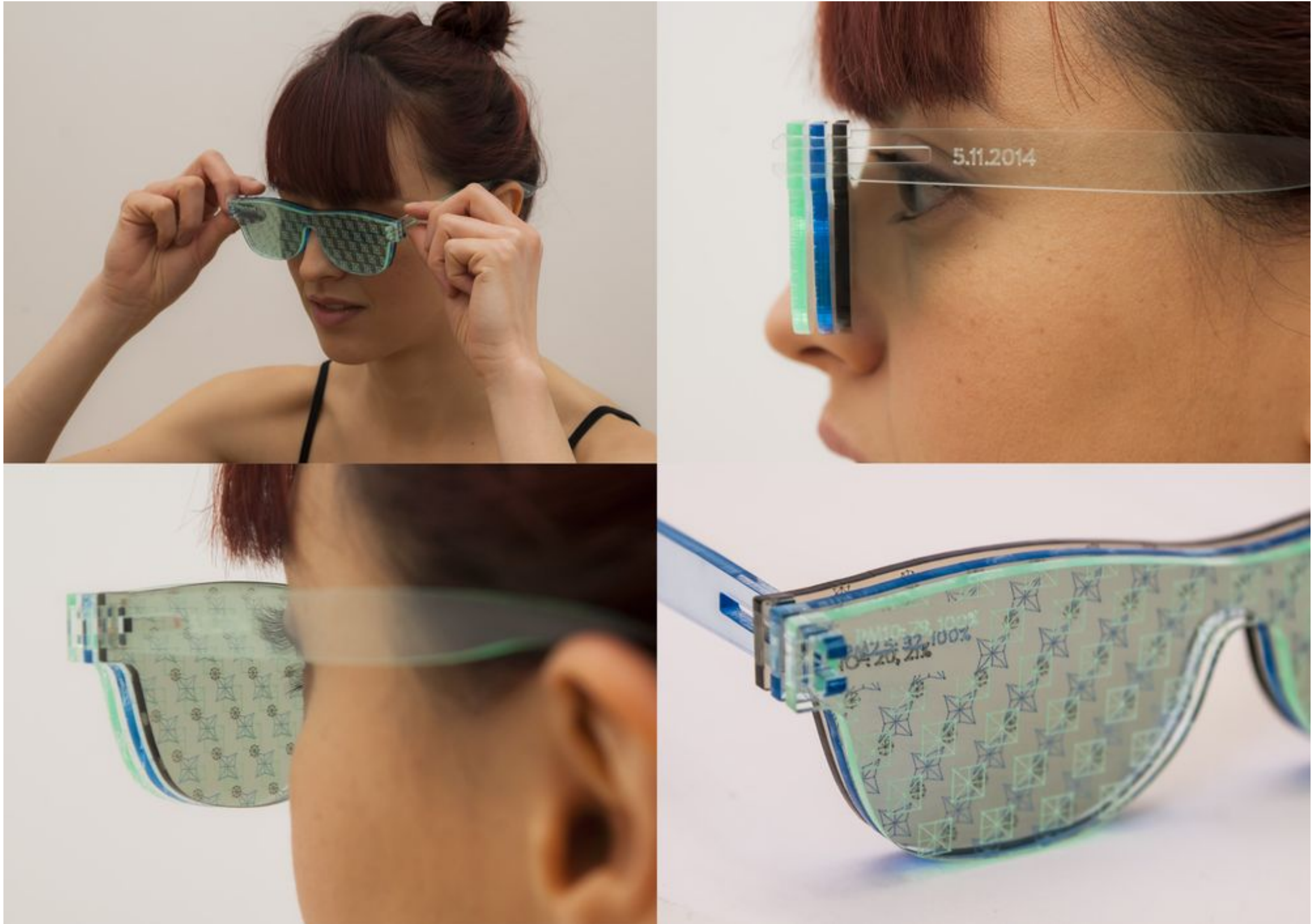
787 25 N/S Hamburg-Altona München Hbf

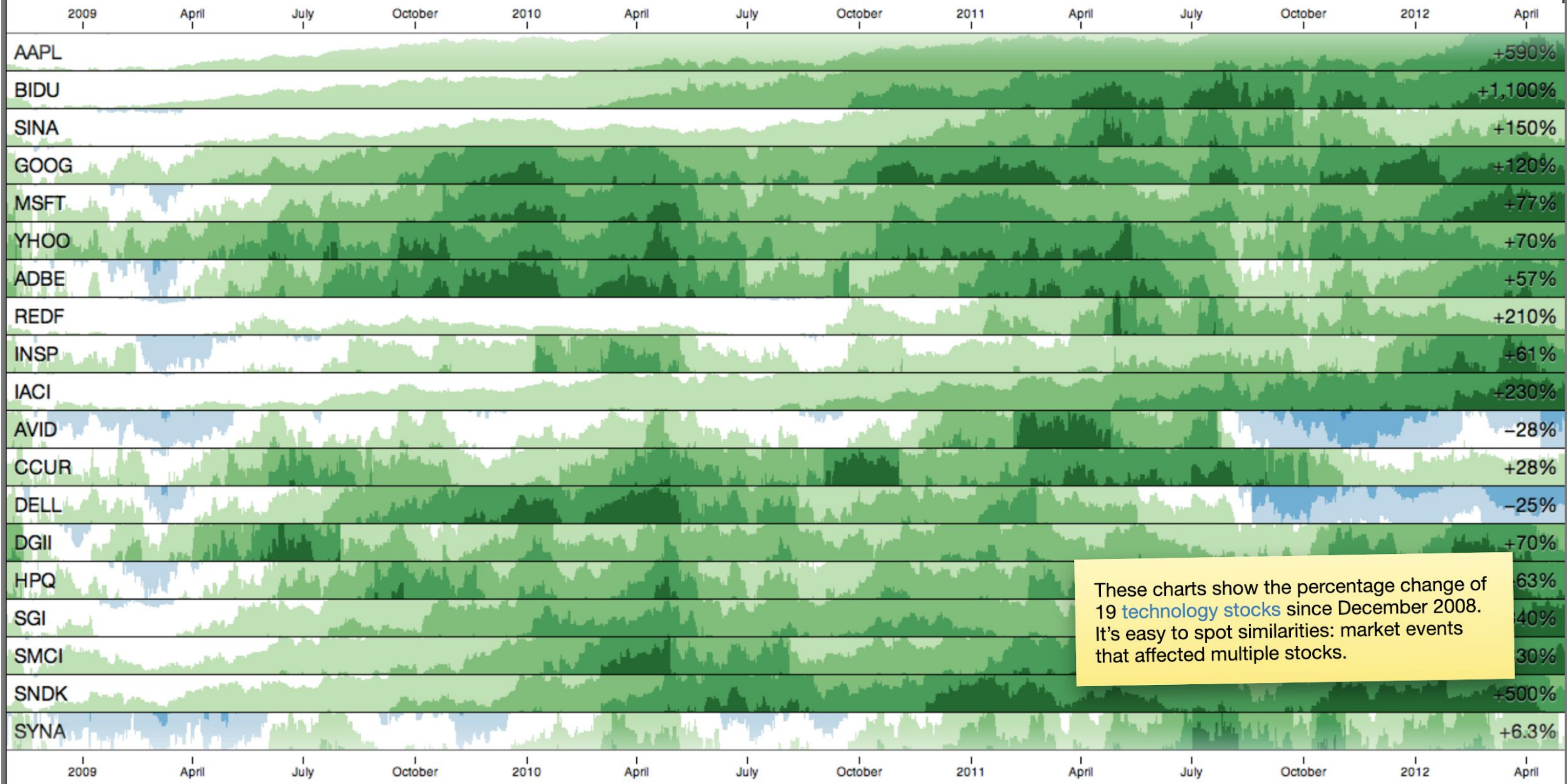
125 Andere N/S ⚠ □ Arnhem Centraal Frankfurt (Main) Hbf

372 11/12 S/N Rastatt Berlin Hbf (tief)

147

https://twitter.com/moritz_stefaner/status/1055112691671539713


5.11.2014
32, 100%
20, 21%



These charts show the percentage change of 19 technology stocks since December 2008. It's easy to spot similarities: market events that affected multiple stocks.

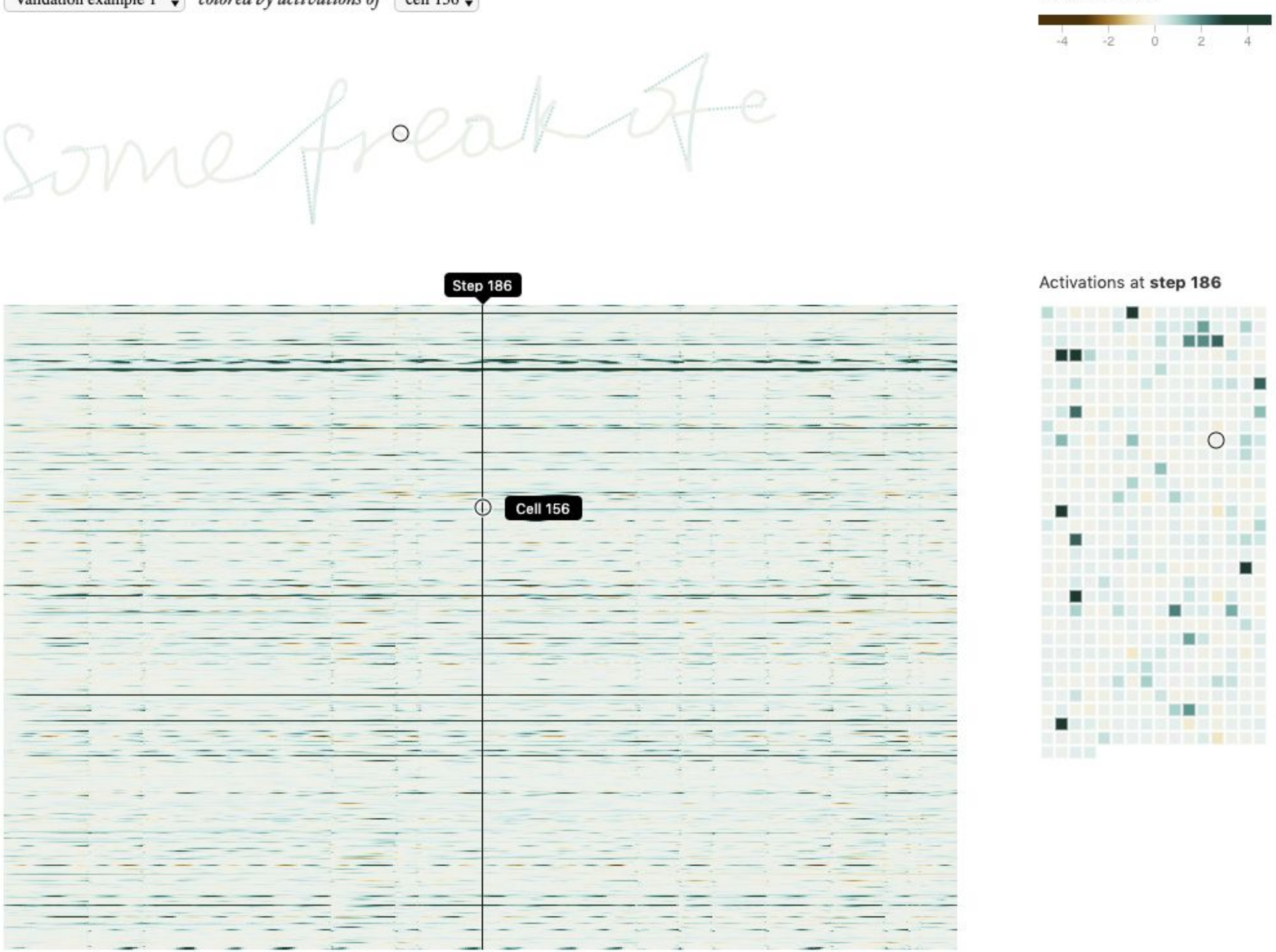
Si has vivido

5 años cerca de

Gran Vía

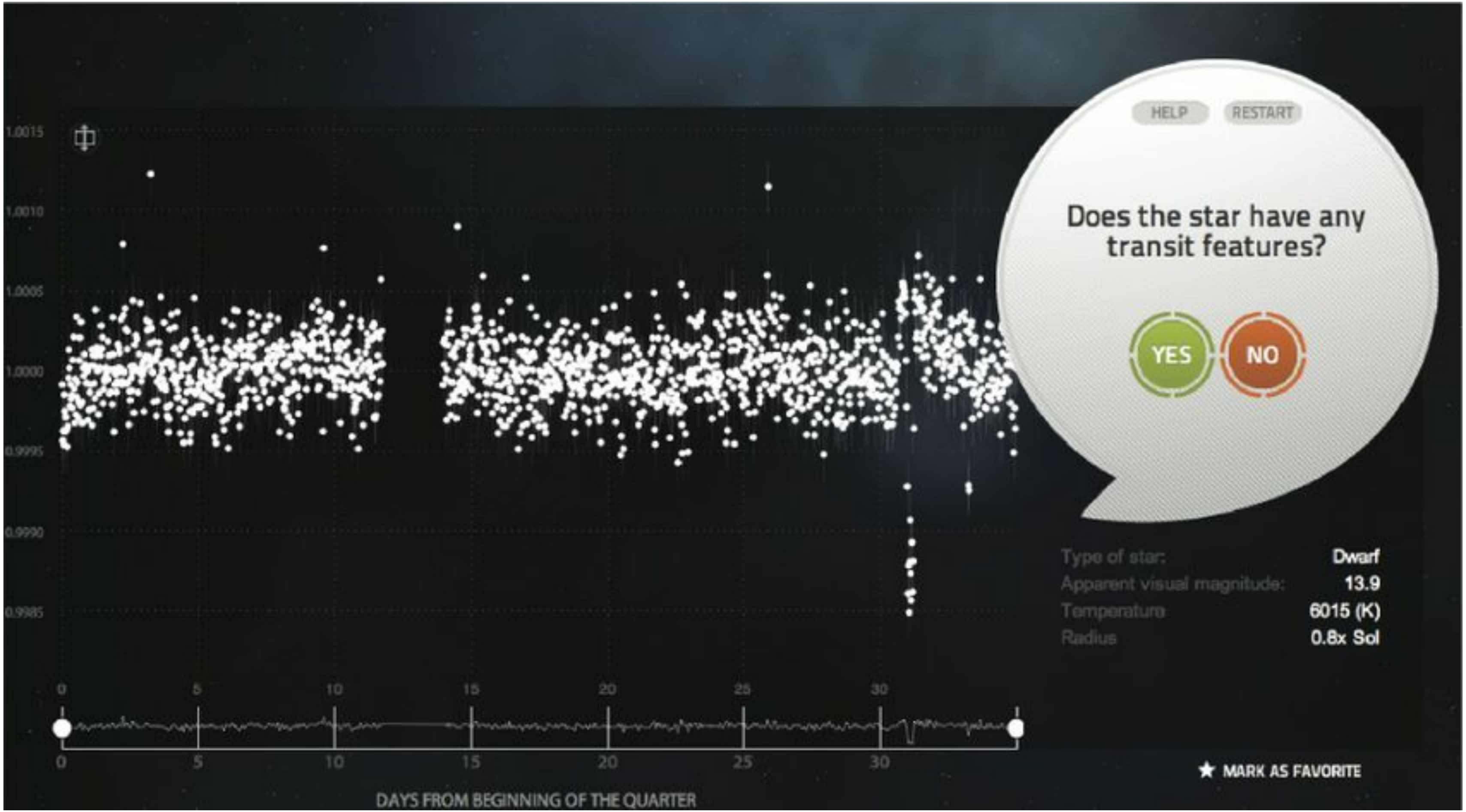
has respirado el equivalente a

Comer **3.4kg de carbonilla**



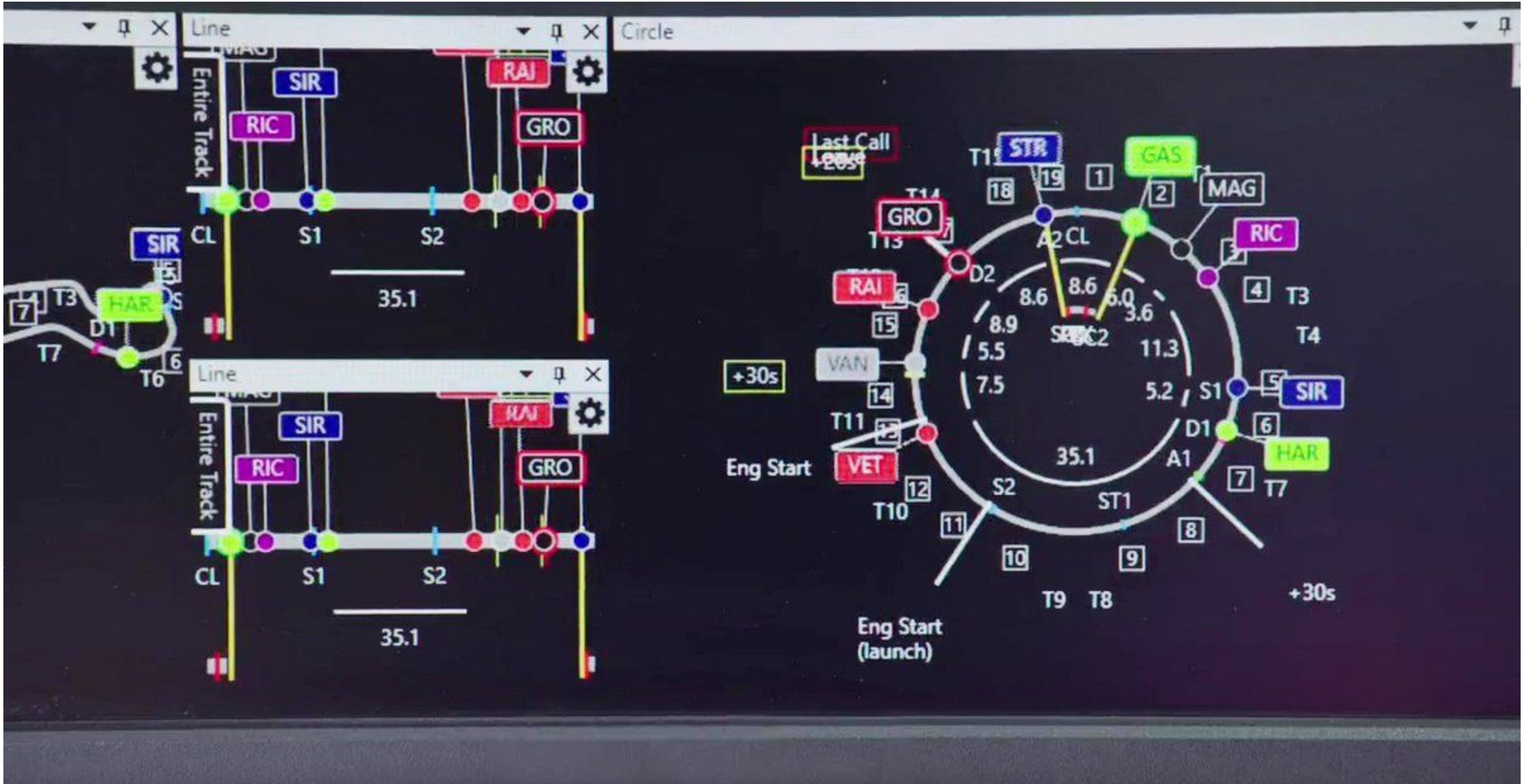
Validation example 1 colored by activations of cell 156

https://distill.pub/2016/handwriting/



Does the star have any transit features?

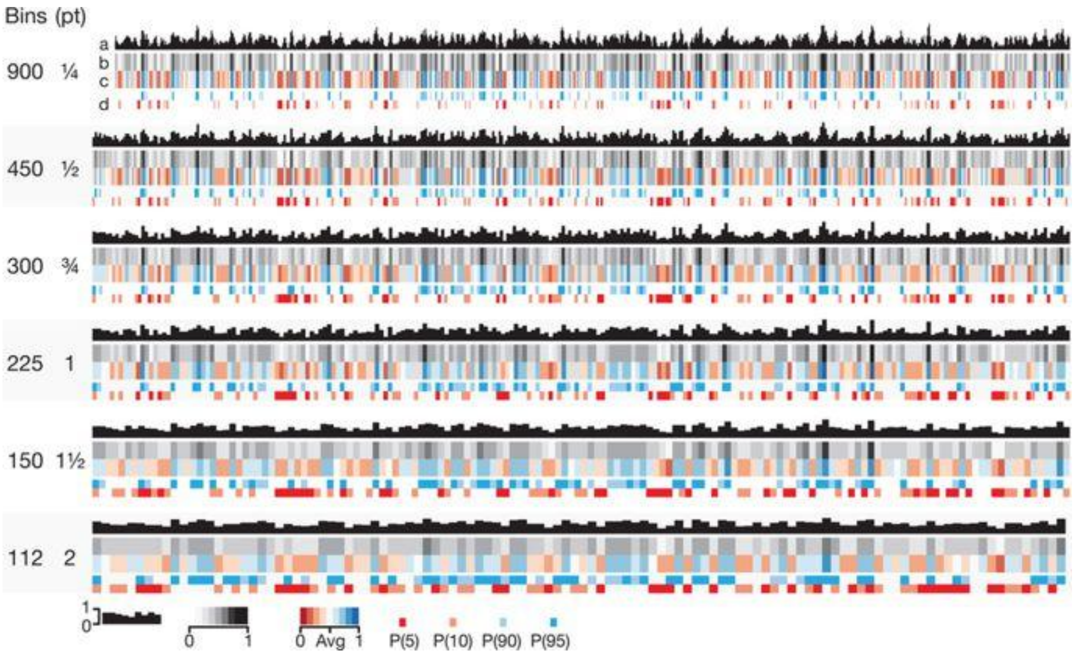
Type of star:	Dwarf
Apparent visual magnitude:	13.9
Temperature	6015 (K)
Radius	0.8x Sol



https://www.reddit.com/r/formula1/comments/baps2h/some_interesting_qualifying_data_visualization/

Use case

What kind of vis?

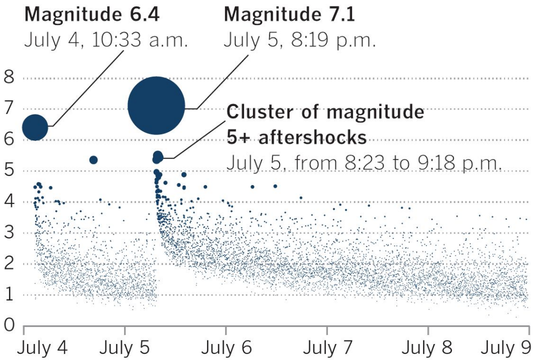


Use case

What kind of vis?

Ridgecrest aftershocks tapering off

Thousands of aftershocks have followed two larger earthquakes near Ridgecrest last week. They've followed a predictable, decreasing pattern.



Magnitude 6.4
July 4, 10:33 a.m.

Magnitude 7.1
July 5, 8:19 p.m.

Cluster of magnitude 5+ aftershocks
July 5, from 8:23 to 9:18 p.m.

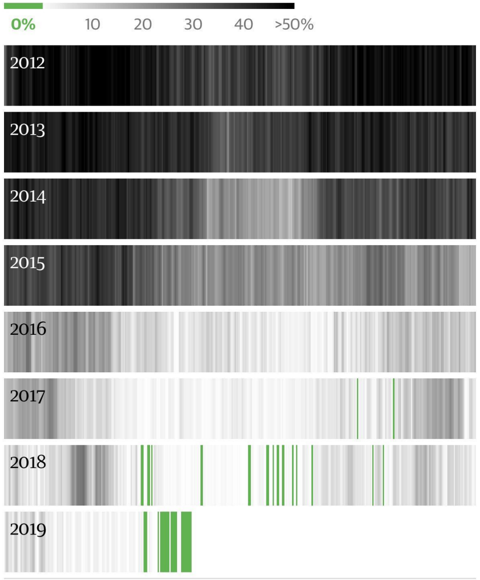
Points show earthquakes between 10 a.m. July 4 and 10 a.m. July 9 near Ridgecrest, Calif. The gap in aftershocks following the magnitude 7.1 is because instruments are unable to detect smaller earthquakes following a main shock.

Sources: U.S. Geological Survey, Caltech, Times Reporting

Chris Keller / Los Angeles Times

Britain is rapidly phasing out coal

Daily share of Britain's power generated by burning coal

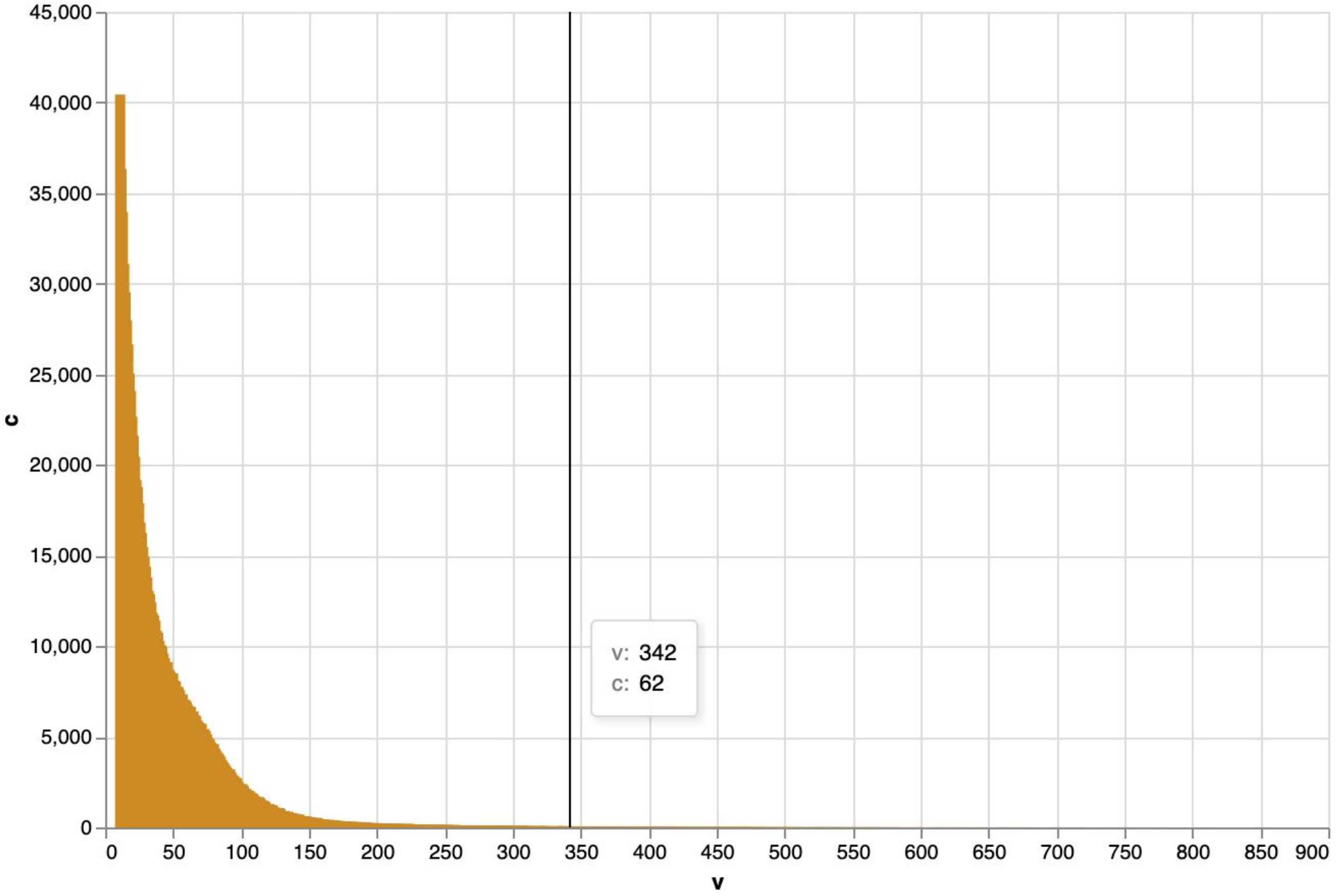


http://javisantana.com/**data_vis_workshop**/

Use case

Steps

- JSON
- Render a square
- Render the data
- Color
- More on colors
- Text


45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
c
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
v
v: 342
c: 62

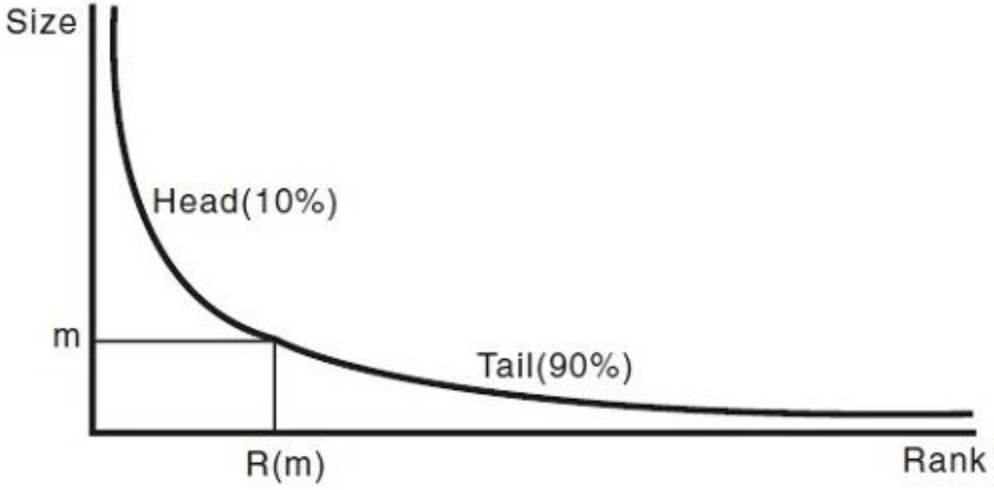
Use case

Protocolo anticontaminación

- Alerta: 180
- Aviso: 200

Use case

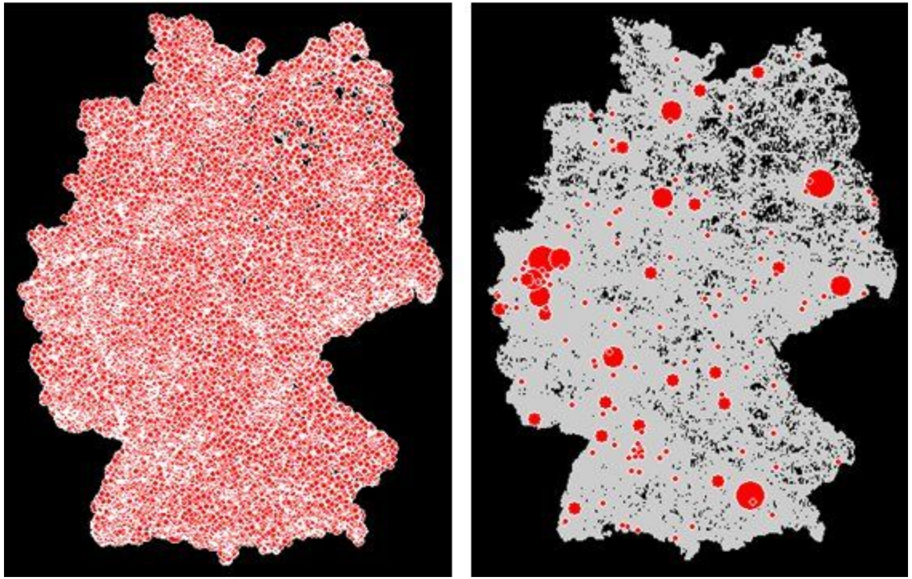
Head/tails



Use case

Head/tails

- [34,65,91,123,169,237,321]



Thanks - QA

@javisantana

@tinybirdco

Full hands on

Hands on video from data to vis

In this video I don’t work on the same visualization but the process is the same

https://www.youtube.com/watch?v=V1nbigUstGA