iMEdD: incubator for Media Education and Development

INTRO TO DATA JOURNALISM

Kelly Kiki

k.kiki@imedd.org

Department of Statistics

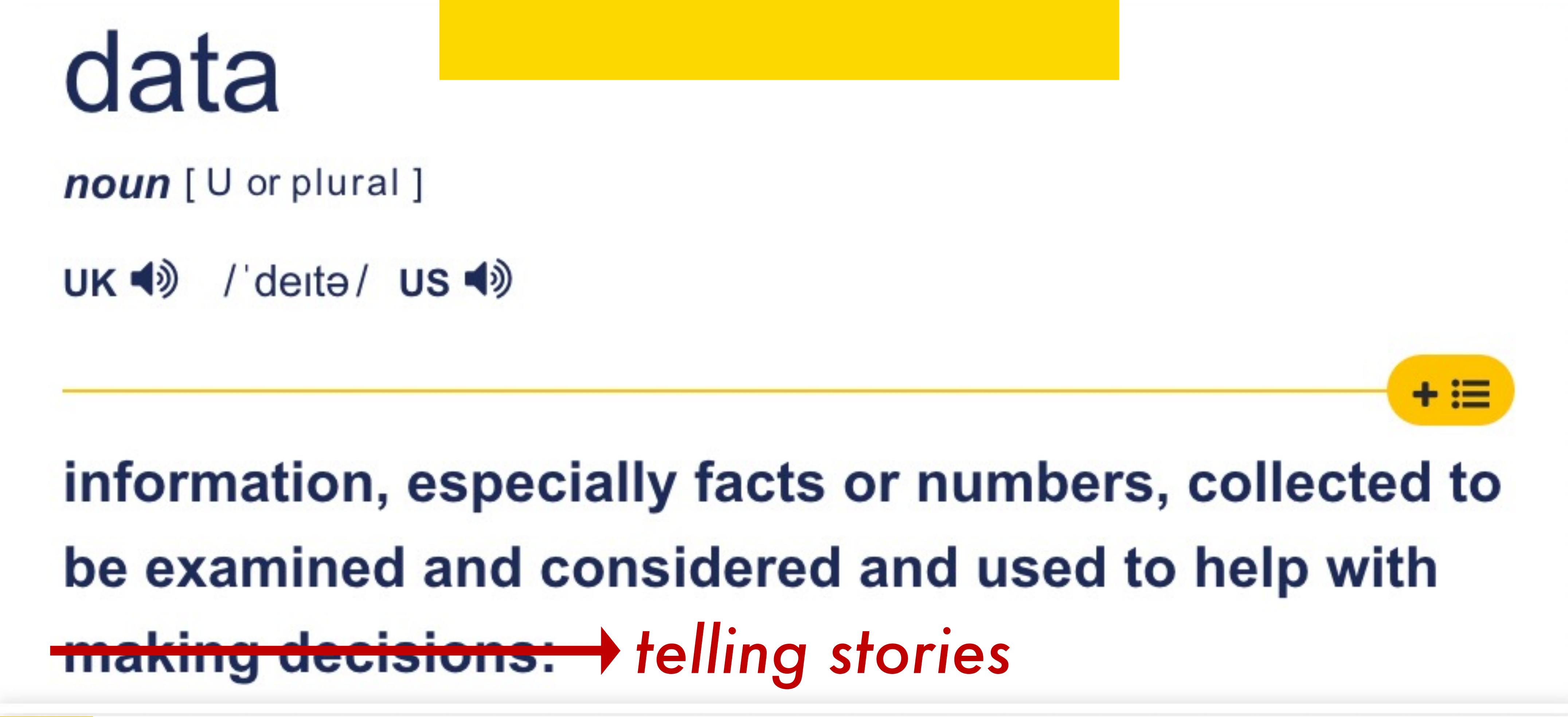
Athens University of Economics and Business

Athens, 1 December 2023

Founding Donor ΙΣΝ / SNF ΙΔΡΥΜΑ ΣΤΑΥΡΟΣ ΝΙΑΡΧΟΣ STAVROS NIARCHOS FOUNDATION

Special thanks to my colleague Thanasis Troboukis for contributing to this presentation.

WHAT IS DATA & DATA JOURNALISM



data

noun [U or plural]

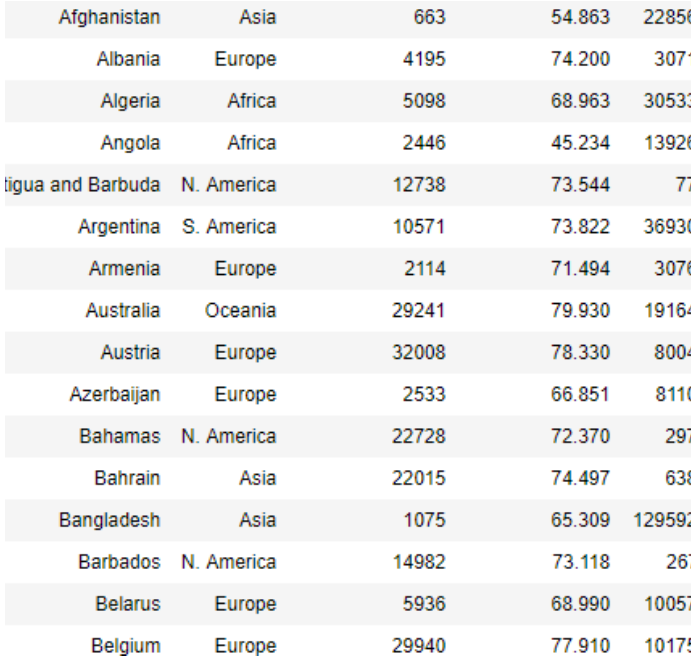
UK 🔊 / ˈdeɪtə / US 🔊

information, especially facts or numbers, collected to be examined and considered and used to help with ~~making decisions:~~ → *telling stories*

Cambridge Dictionary

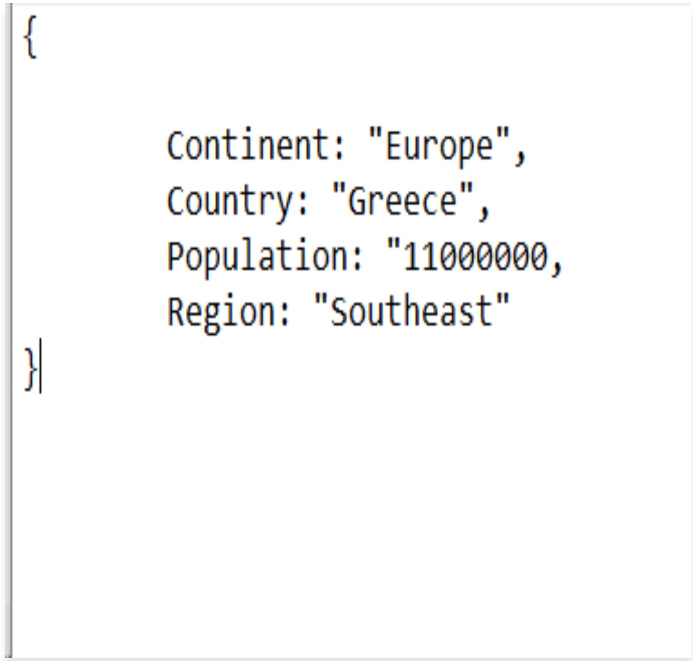
DATA TYPES & FORMATS

STRUCTURED (.csv, .xlsx files, databases)



Afghanistan	Asia	663	54.863	2285
Albania	Europe	4195	74.200	307
Algeria	Africa	5098	68.963	3053
Angola	Africa	2446	45.234	1392
tigua and Barbuda	N. America	12738	73.544	7
Argentina	S. America	10571	73.822	3693
Armenia	Europe	2114	71.494	307
Australia	Oceania	29241	79.930	1916
Austria	Europe	32008	78.330	800
Azerbaijan	Europe	2533	66.851	811
Bahamas	N. America	22728	72.370	29
Bahrain	Asia	22015	74.497	63
Bangladesh	Asia	1075	65.309	12959
Barbados	N. America	14982	73.118	26
Belarus	Europe	5936	68.990	1005
Belgium	Europe	29940	77.910	1017

SEMI-STRUCTURED (.json files, HTML)

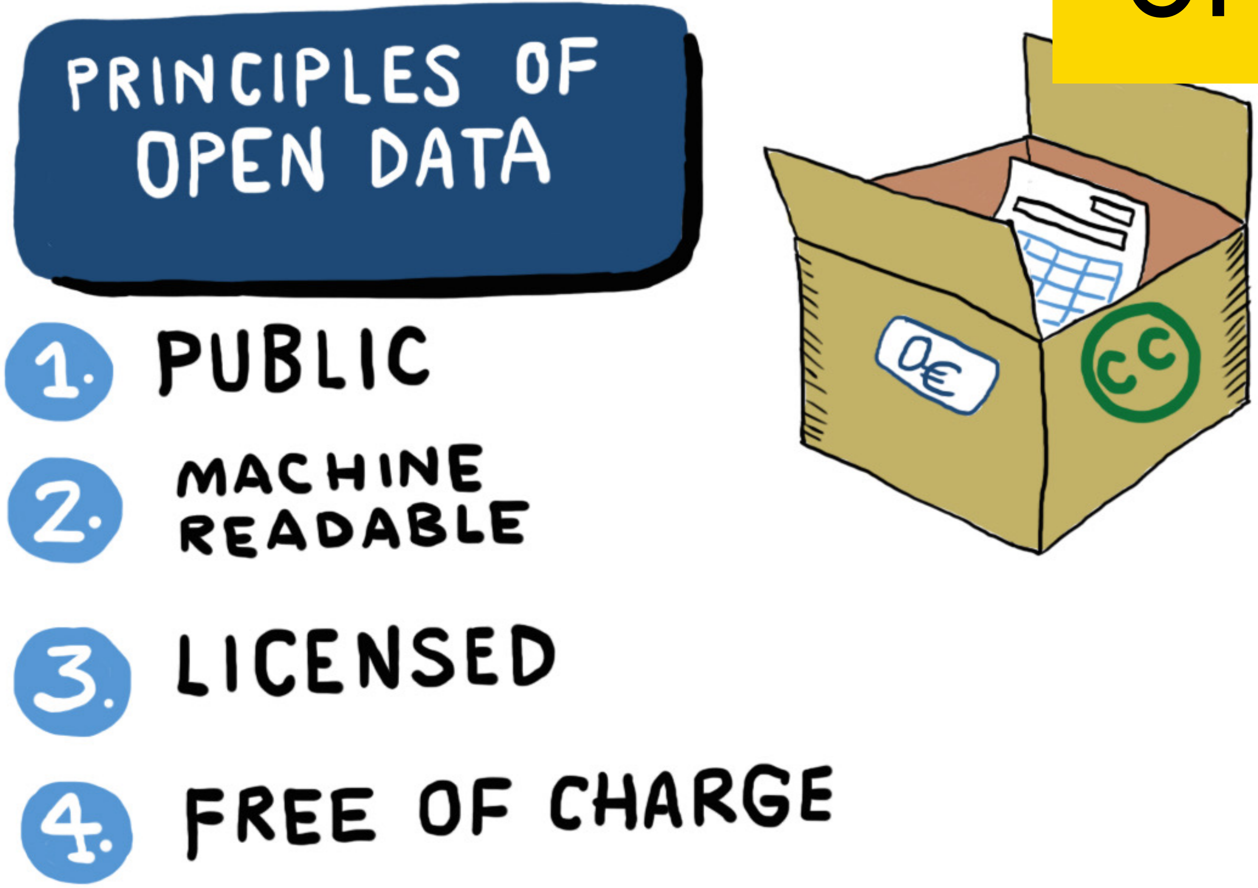


```
{

        Continent: "Europe",
        Country: "Greece",
        Population: "11000000,
        Region: "Southeast"

}
```

UNSTRUCTURED (text files, audio, video)

www.iMEdD.org

COMMON CASES OF LOOKING LIKE THEY ARE OPEN. BUT THEY AREN'T.



Picture: Marker Wizards, available on [Helsinki Region Infoshare]()

1 Preprocessed figures found in press releases, emails and/or other documents communicated to journalists

2 Data enclosed in .pdf files uploaded on the Internet

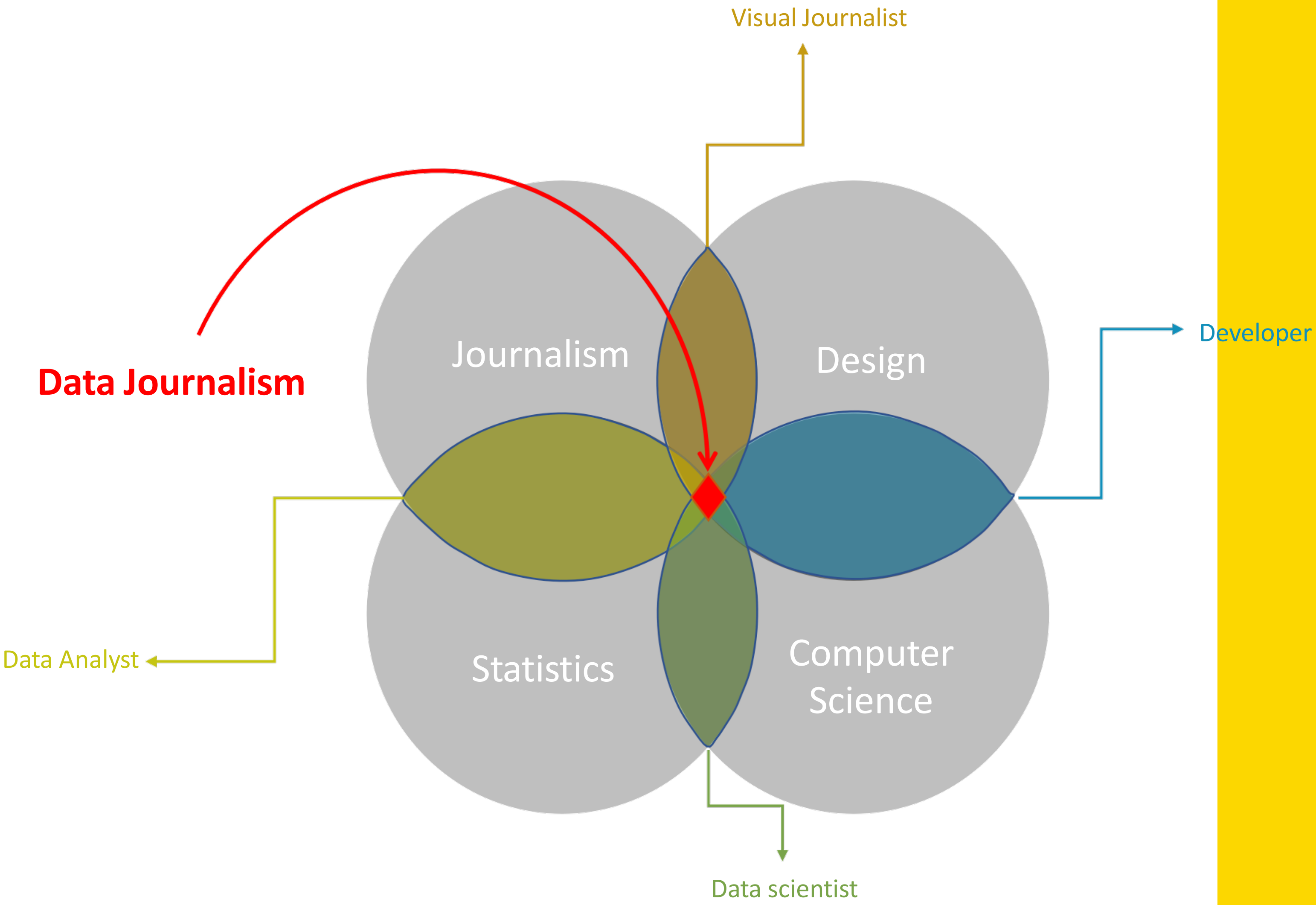
3 Unstructured data found on websites (.html pages), especially when no permission of reuse is given

4 Unconsistently maintained and/or undocumented data

WHAT IS DATA JOURNALISM

> “ What was once a garage band has now grown big enough to make up an orchestra.

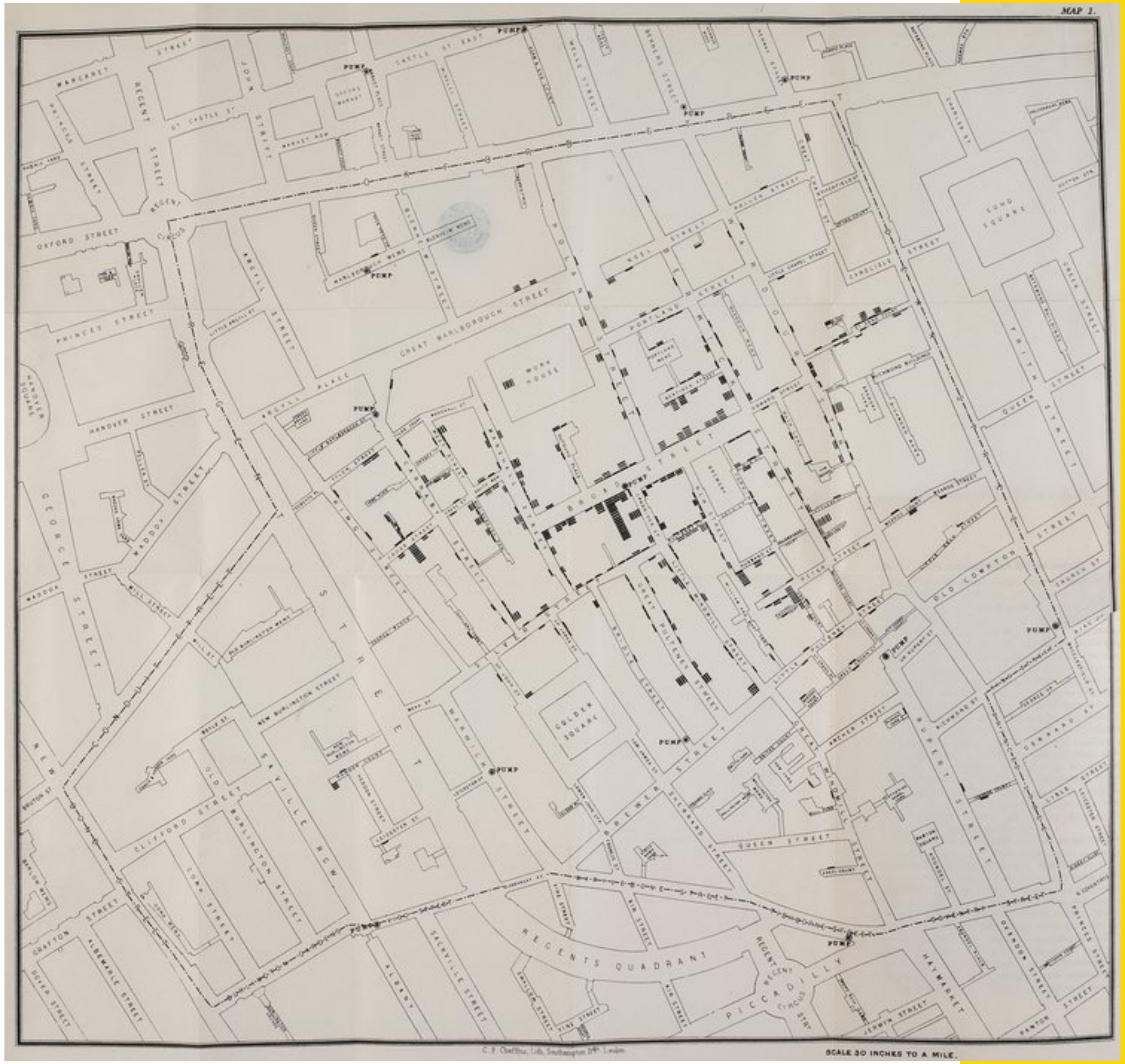
Sarah Cohen (2021),
Ways of Doing Data Journalism



DATA JOURNALISM IS NOT THAT NEW

1855

One of the most influential urban medical maps of the 19th century, by John Snow. His essay showed the spread of cholera in Soho, London.

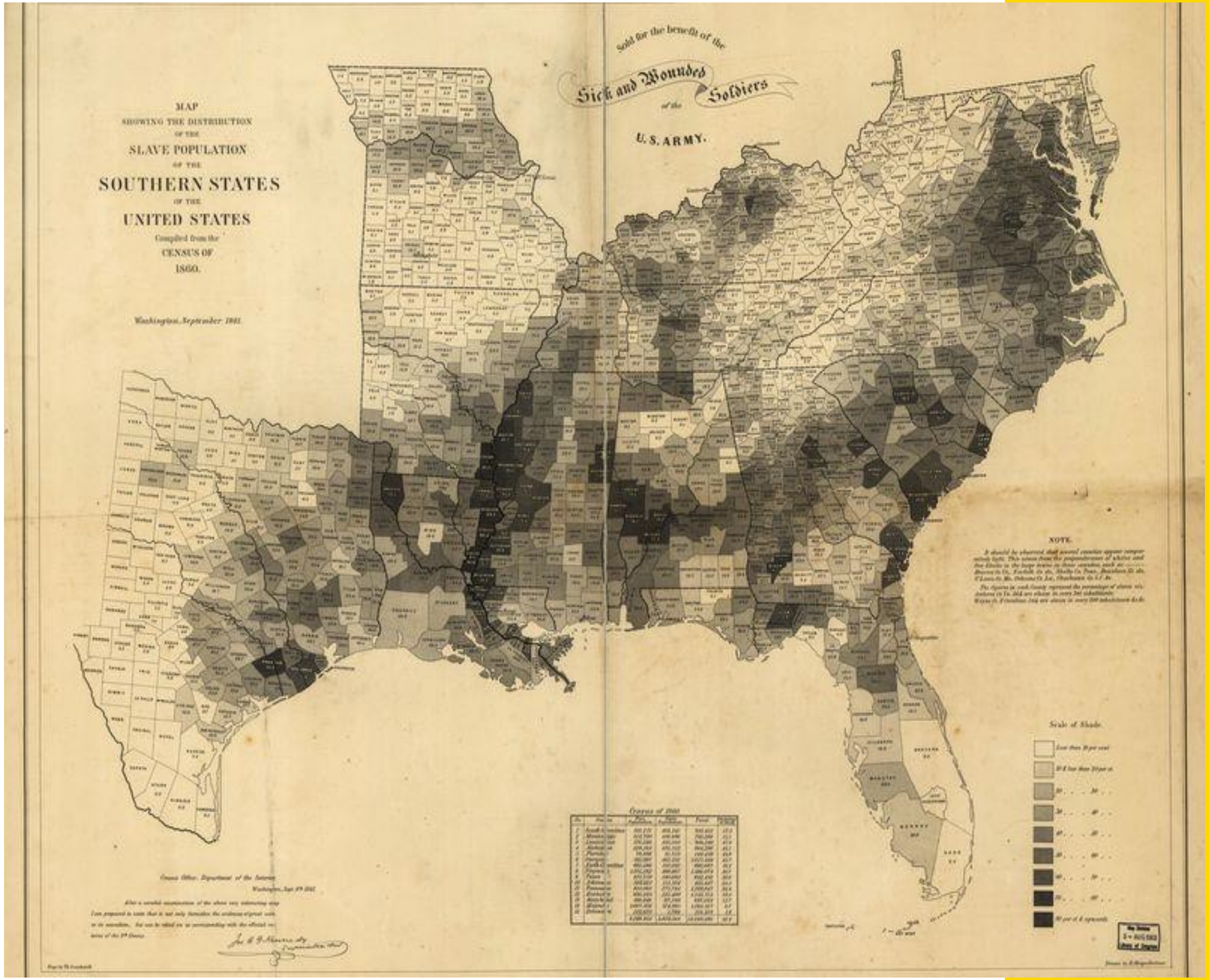


British Library

DATA JOURNALISM IS NOT THAT NEW

1860

A choropleth map showing the distribution of the slave population in the southern states of the Unided States of America.

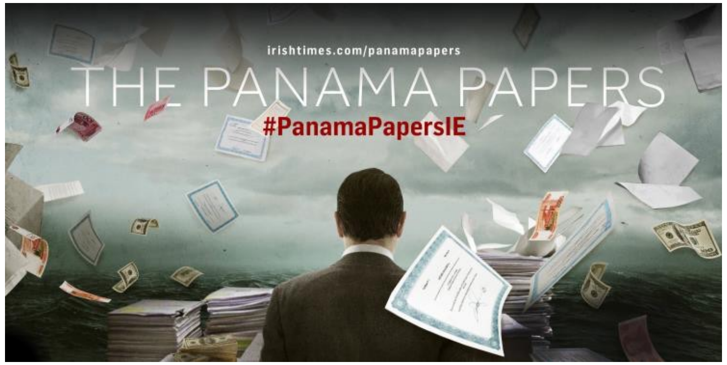


Library of Congress, Geography and Map Division

WHY WE NEED DATA JOURNALISM

Data analysis, processing and visualization can expose patterns hidden beyond what eyes can see. This can lead to important stories for the public interest.

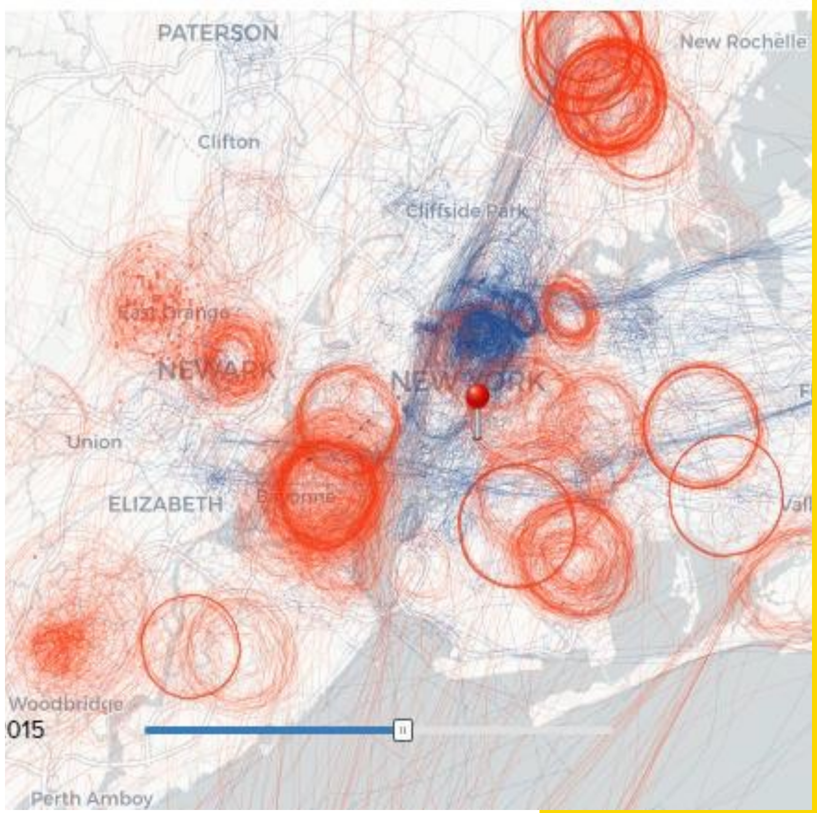
> “When information was scarce, most of our efforts were devoted to hunting. Now that information is abundant, processing is more important.
>
> *Philiip Meyer, Journalist*
> *datajournalism.com*



An ICIJ Investigation

PANDORA PAPERS

The largest investigation in journalism history exposes a shadow financial system that benefits the world’s most rich and powerful. Read all...



Spies in the Skies, Buzzfeed News

DATA JOURNALISTS' SKILLSET & WORKFLOW

THE SKILLSET OF A DATA JOURNALIST

Background knowledge

- Fact-checking
- Reporting
- Research/Investigation experience

Tech skills

- Command of spreadsheet-related softwares (Excel, Google Sheets, Open Office etc)
- Coding skills/ Use of a programming language (Python, R etc)
- Basic HTML
- Ability to learn quickly how to use new tools

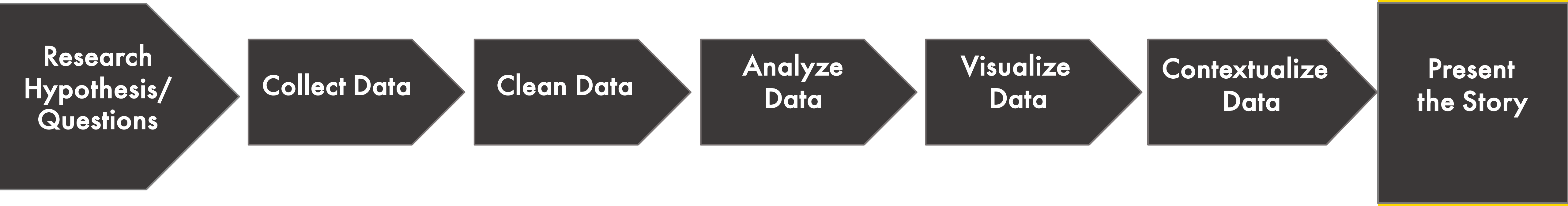
Math-related skills

- Descriptive & Comparative Statistics
- Understand concepts of Predictive Statistics and Algorithms to collaborate with data-scientists

Interpersonal competencies

- Love for detail
- High frusstration tolerance
- Problem solver attitude
- Team player!

THE DATA JOURNALISM WORKFLOW



Research Hypothesis/ Questions

Collect Data

Clean Data

Analyze Data

Visualize Data

Contextualize Data

Present the Story

0. RESEARCH QUESTIONS

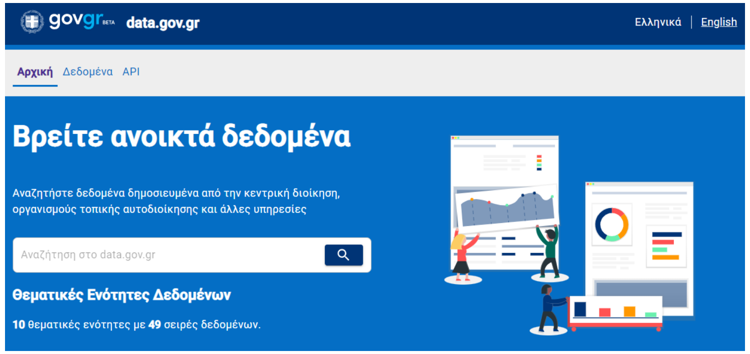
- Define the research hypothesis you need to confirm or reject.
- Wirte down specific questions you need to answer on your way to prove or disapprove your general idea.
- Know exactly what you are looking for, so that you can ask your data the proper questions once you get them.

1. DATA COLLECTION

- Find structured and machine-readable data available to download on online repositories/databases.
- Scrape the web / Extract data from .pdf and other text files published.
- Use an API service available to request data.
- Make a data request to relevant insittutions, public bodies and/or third-party partners.

ONLINE OFFICIAL DATA REPOSITORIES

Ελληνική Στατιστική Αρχή


Βρείτε ανοικτά δεδομένα

Αναζητήστε δεδομένα δημοσιευμένα από την κεντρική διοίκηση, οργανισμούς τοπικής αυτοδιοίκησης και άλλες υπηρεσίες

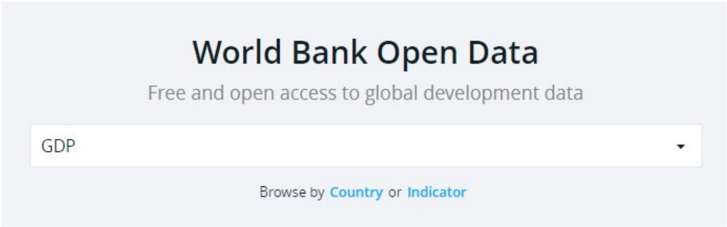
Θεματικές Ενότητες Δεδομένων

10 θεματικές ενότητες με 49 σειρές δεδομένων.

eurostat

Your key to European statistics

European Commission > Eurostat > Data > Database


World Bank Open Data

Free and open access to global development data

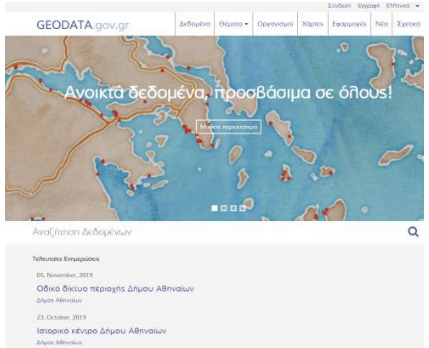
GDP

Browse by Country or Indicator

UNdata A world of information

energy

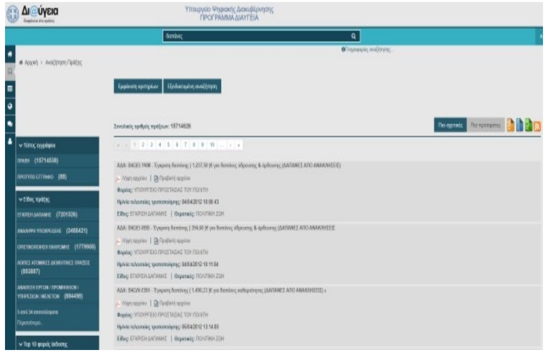
32 databases - 60 million records

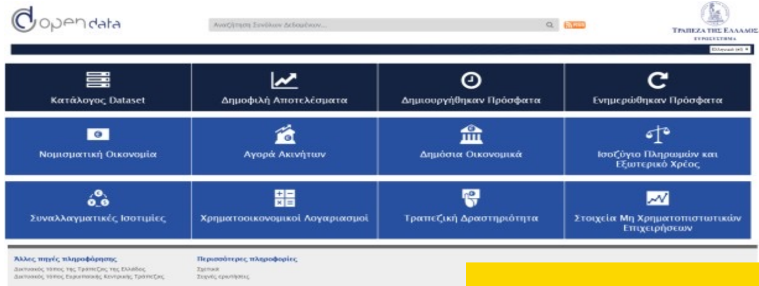

Ανοικτά δεδομένα, προσβάσιμα σε όλους!

Τελευταίες Ενημερώσεις

05, November, 2019
Οδικό δίκτυο περιοχής Δήμου Αθηναίων
Δήμος Αθηναίων

23, October, 2019
Ιστορικό κέντρο Δήμου Αθηναίων
Δήμος Αθηναίων





2. DATA CLEANING

- Explore your dataset, understand your data, check them out, make methodological decisions and corrections.
- Prepare your data so that you can analyze them: Manipulate the data to end up with a structured quantitative dataset which includes all the variables you wish to study.

3. DATA ANALYSIS

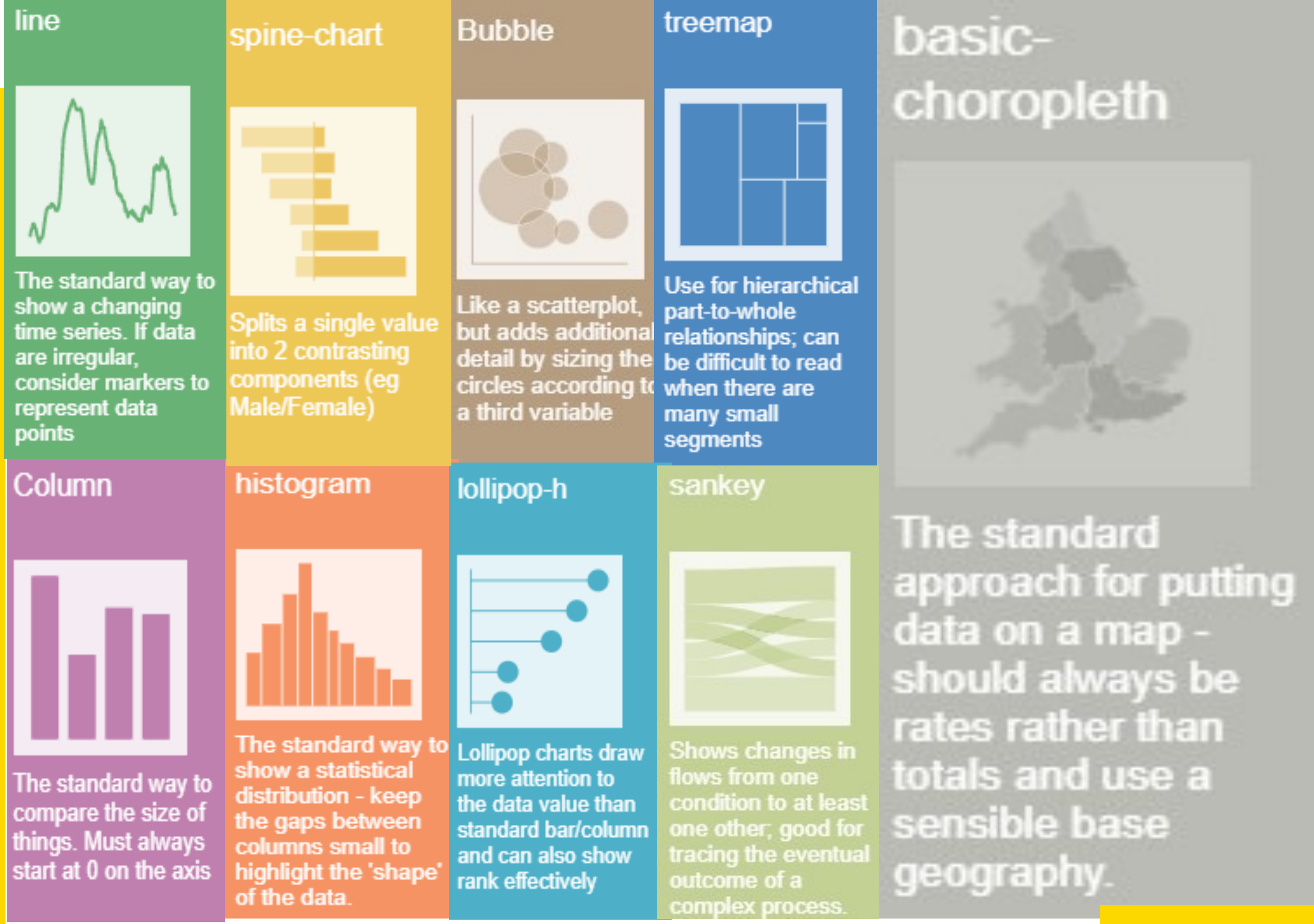
- Do not rely on absolute numbers.
- Normalize your data in order to compare different populations.
- Mind that outliers might influence your averages.
- Calculate percentage change when looking for growth in a time period compared to another.
- Always remember that correlation does not mean causation.

4. DATA VISUALIZATION

Choose the chart type

it fits to what you wish to

visualize



line

The standard way to show a changing time series. If data are irregular, consider markers to represent data points

spine-chart

Splits a single value into 2 contrasting components (eg Male/Female)

Bubble

Like a scatterplot, but adds additional detail by sizing the circles according to a third variable

treemap

Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small segments

basic-choropleth

The standard approach for putting data on a map - should always be rates rather than totals and use a sensible base geography.

Column

The standard way to compare the size of things. Must always start at 0 on the axis

histogram

The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

lollipop-h

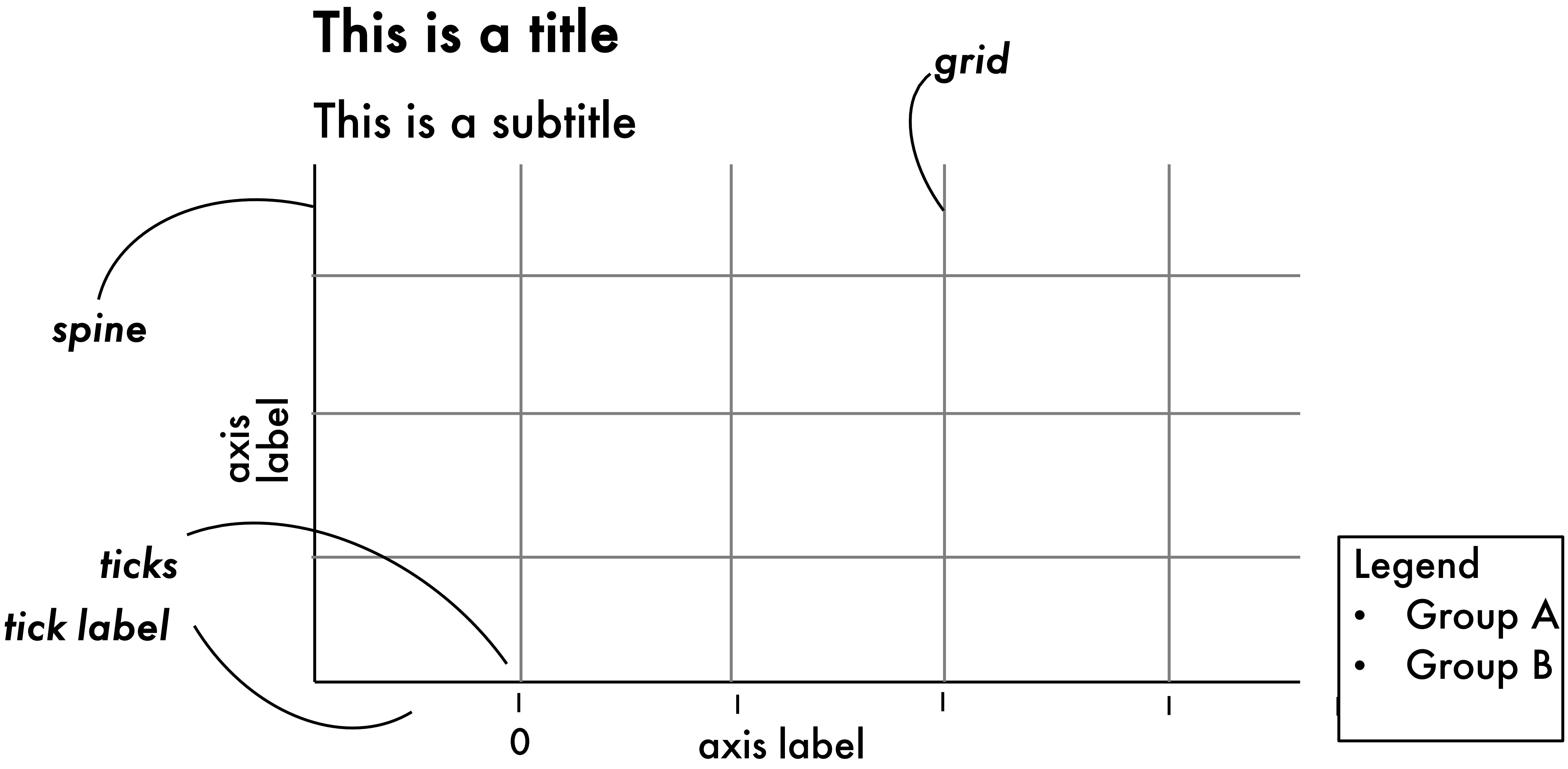
Lollipop charts draw more attention to the data value than standard bar/column and can also show rank effectively

sankey

Shows changes in flows from one condition to at least one other; good for tracing the eventual outcome of a complex process.

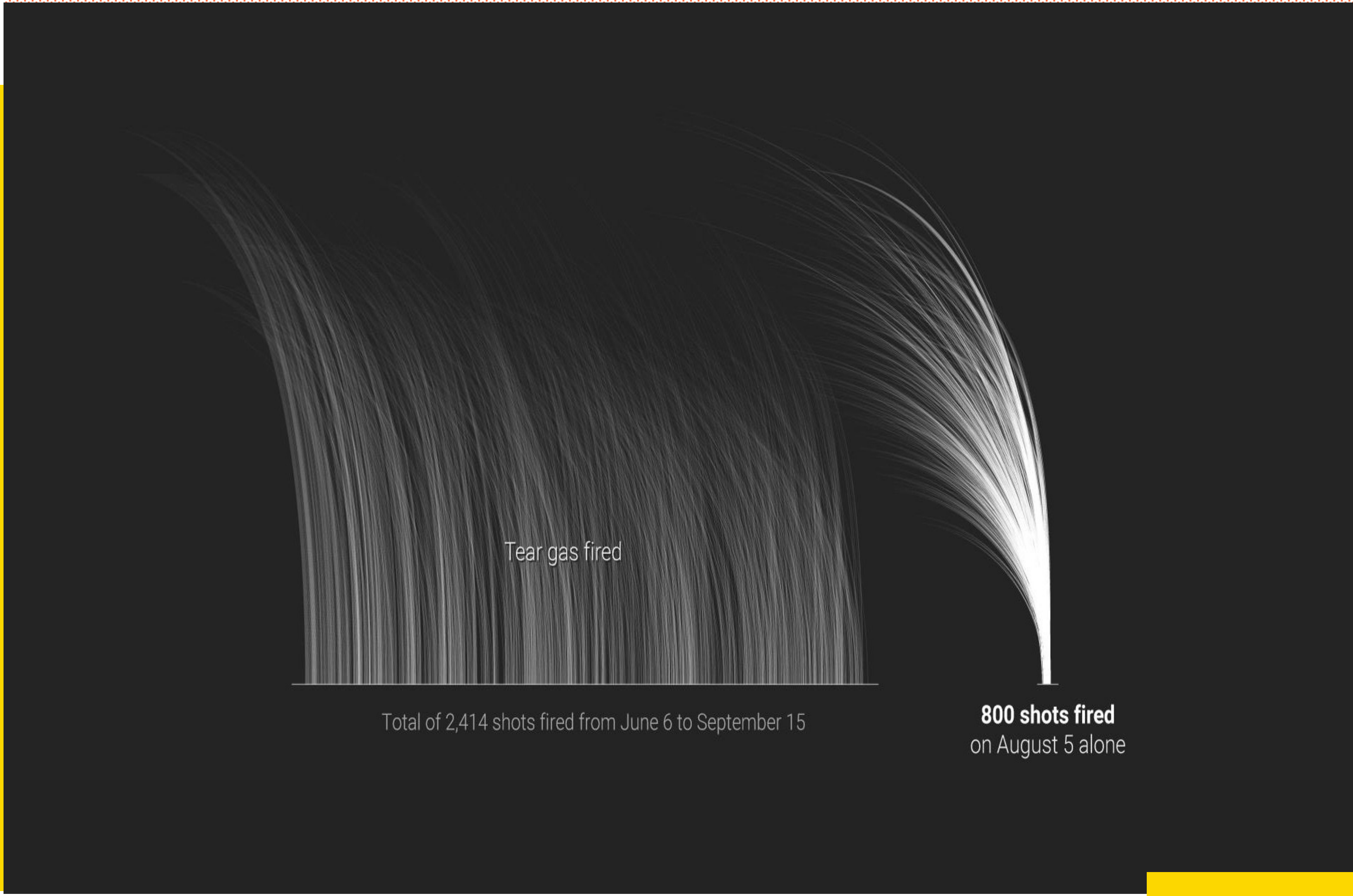
[Financial Times Visual Vocabulary](https://)

THE ANATOMY OF A CHART



SOURCE: Here is where your hyperlinked data source goes.
Notes: Here is where you underline whatever else your reader needs to know.

THEN, START BEING MINIMAL



South China Morning Post

5. CONTEXTUALIZE DATA

- Interview a datapoint. Datapoints are usually humans. Talk to humans.
- Talk to experts and ask them to elaborate your findings.
- Combine your data findings with context knowledge coming out of your datasets.

6. PRESENT YOUR STORY

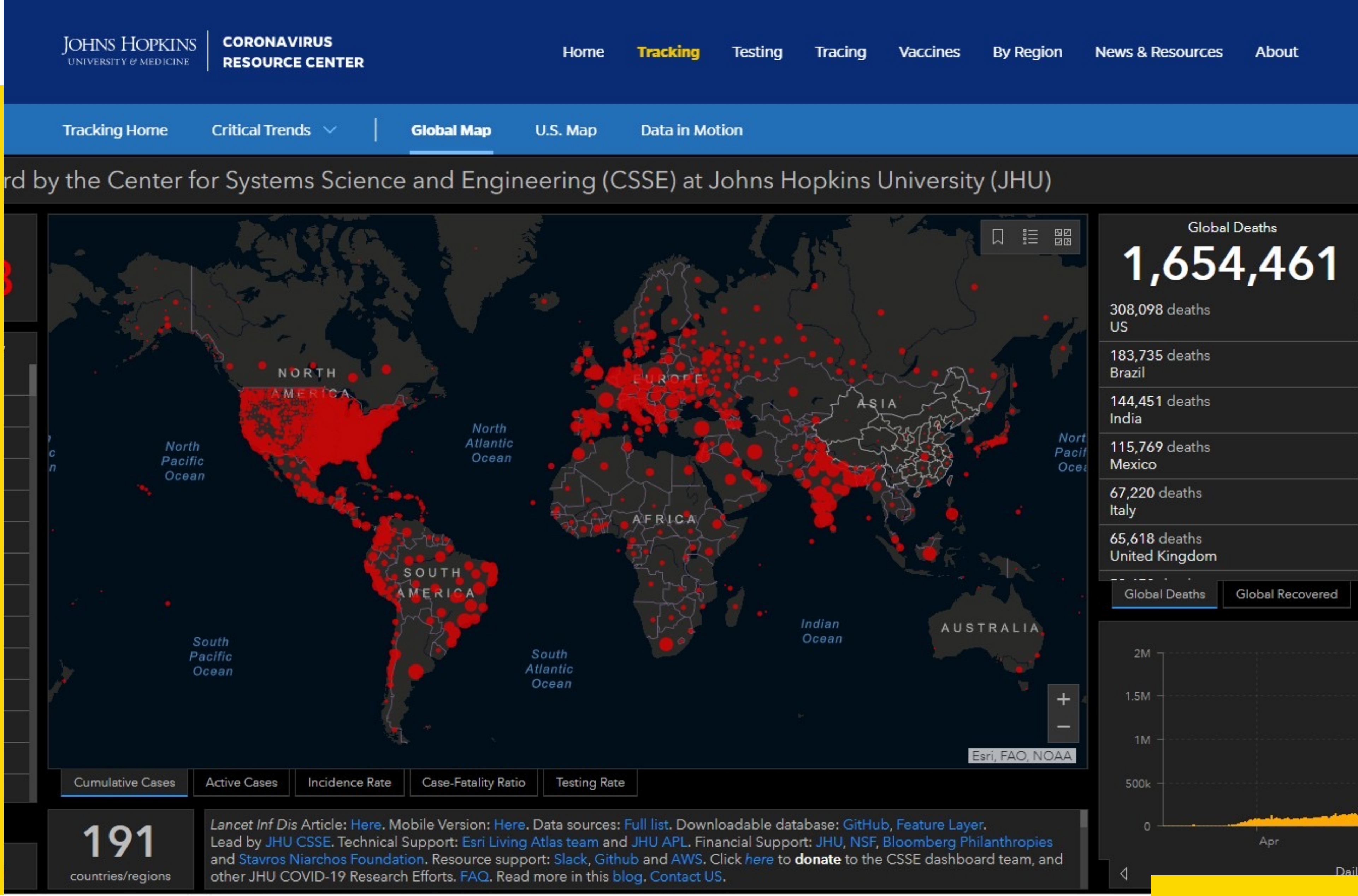
- Tell the story that derives from your data and/or report on a story that your data verify. But do not forget to:
- Be accountable. Along with your story, provide your methodology, open your datasets and your data processing to the public.

HOW A STATISTICIAN CAN BE PART OF A DATA JOURNALISM PROJECT

- **Become a journalist's source:** communicate your research findings, share your data and work with journalists who are interested to report on those for the best understanding of the general pubic
- **Offer your help:** collaborate and share your expertise to aid data journalists in comprehending the significance of their findings, guiding them on how to enhance their analysis, and alerting them to potential pitfalls they should navigate carefully.
- **Participate actively:** become a member of a team and investigate collaboratively; please note that if you plan to work or contribute as a data journalist, you will be expected to do more than anallyze already collected and cleaned data neatly delivered to you.

IN THE RECENT YEARS, DATA HAS BEEN THE NEWS

Coronavirus Research Center, JHU



rd by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)

Global Deaths

1,654,461

308,098 deaths
US

183,735 deaths
Brazil

144,451 deaths
India

115,769 deaths
Mexico

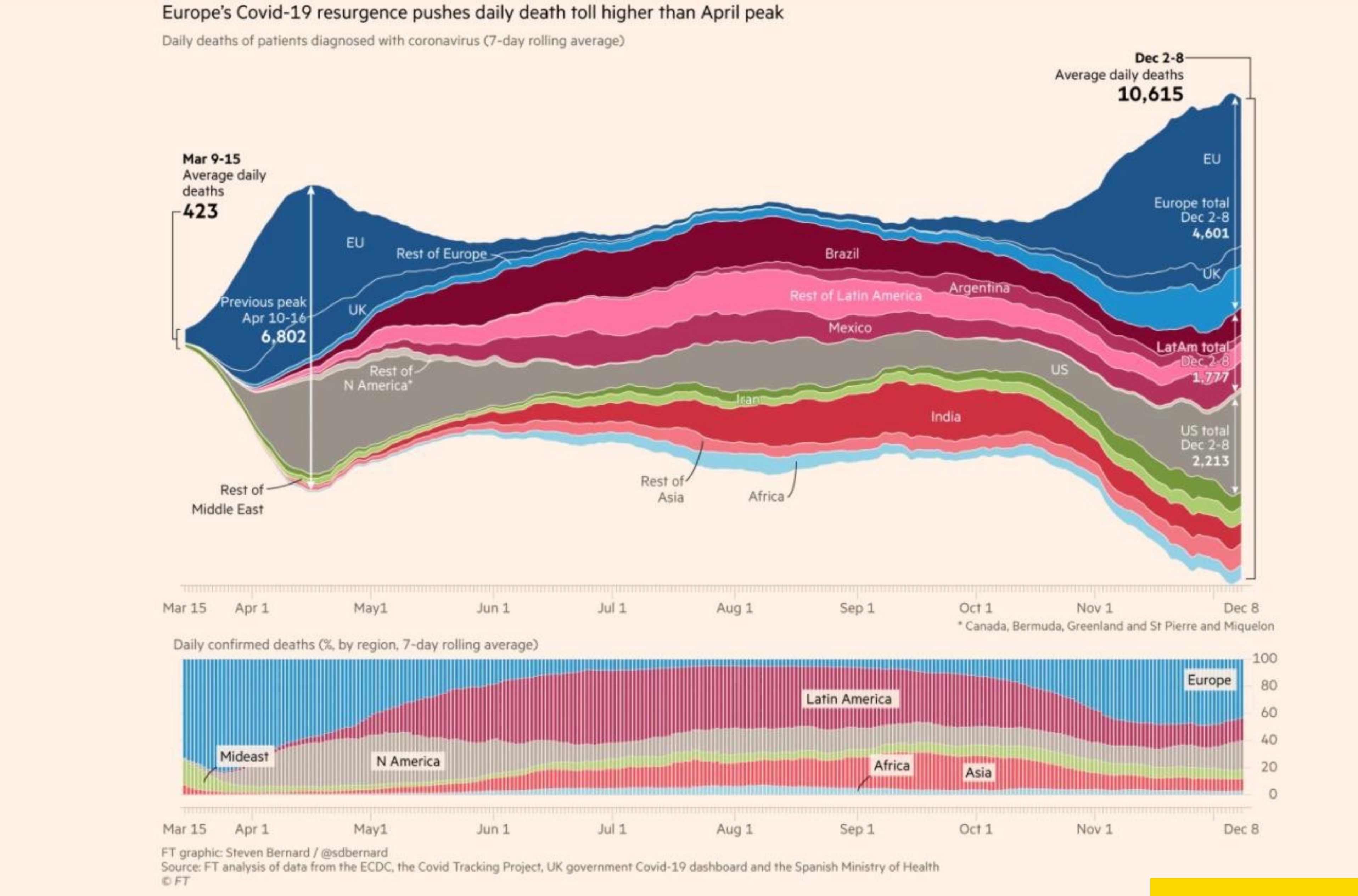
67,220 deaths
Italy

65,618 deaths
United Kingdom

191
countries/regions

Lancet Inf Dis Article: Here. Mobile Version: Here. Data sources: Full list. Downloadable database: GitHub, Feature Layer. Lead by JHU CSSE. Technical Support: Esri Living Atlas team and JHU APL. Financial Support: JHU, NSF, Bloomberg Philanthropies and Stavros Niarchos Foundation. Resource support: Slack, Github and AWS. Click *here* to **donate** to the CSSE dashboard team, and other JHU COVID-19 Research Efforts. FAQ. Read more in this blog. Contact US.

[FT Coronavirus Tracker](#)



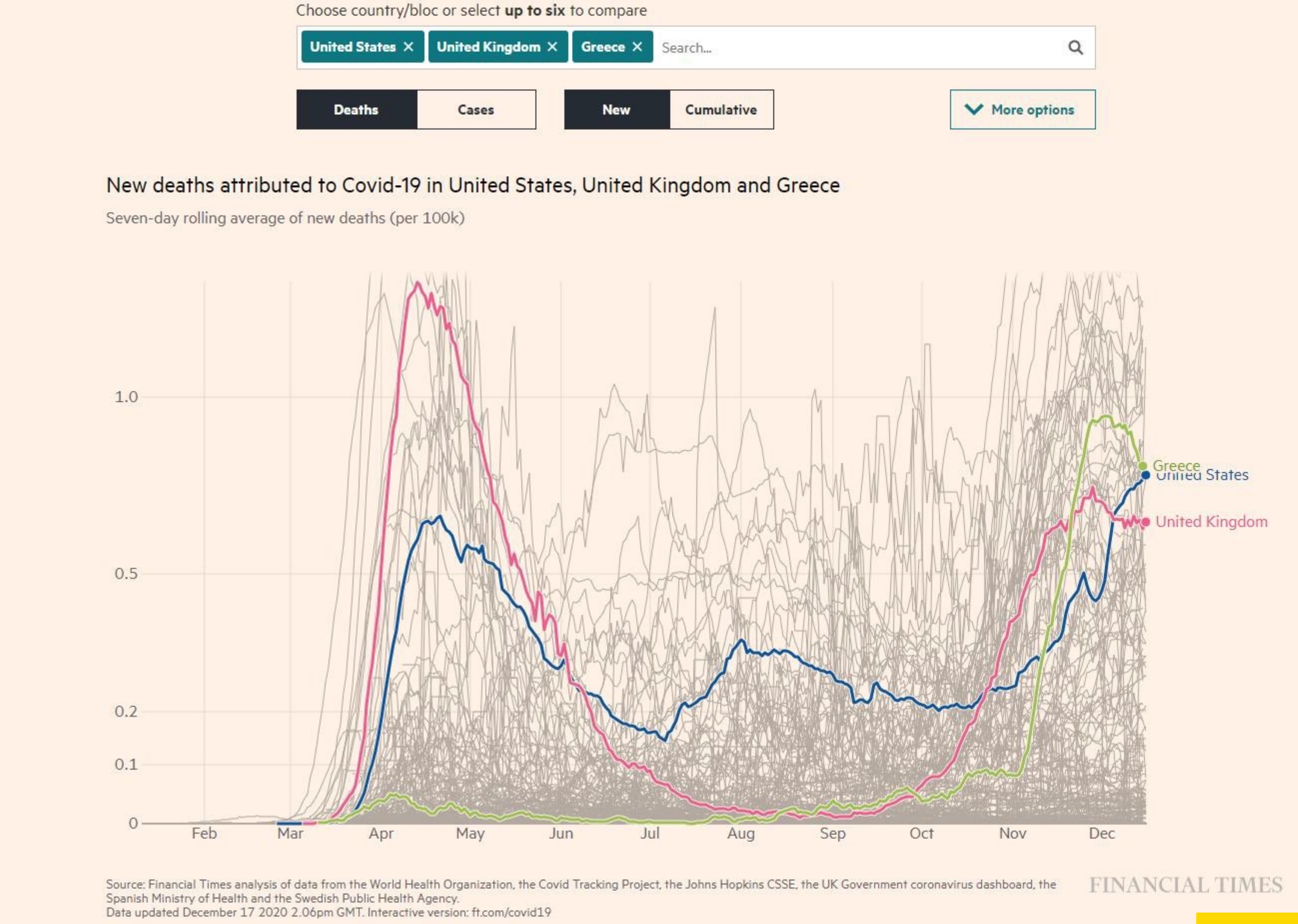
Europe's Covid-19 resurgence pushes daily death toll higher than April peak

Daily deaths of patients diagnosed with coronavirus (7-day rolling average)

* Canada, Bermuda, Greenland and St Pierre and Miquelon

FT graphic: Steven Bernard / @sdbernard
Source: FT analysis of data from the ECDC, the Covid Tracking Project, UK government Covid-19 dashboard and the Spanish Ministry of Health
© FT

FT Coronavirus Tracker



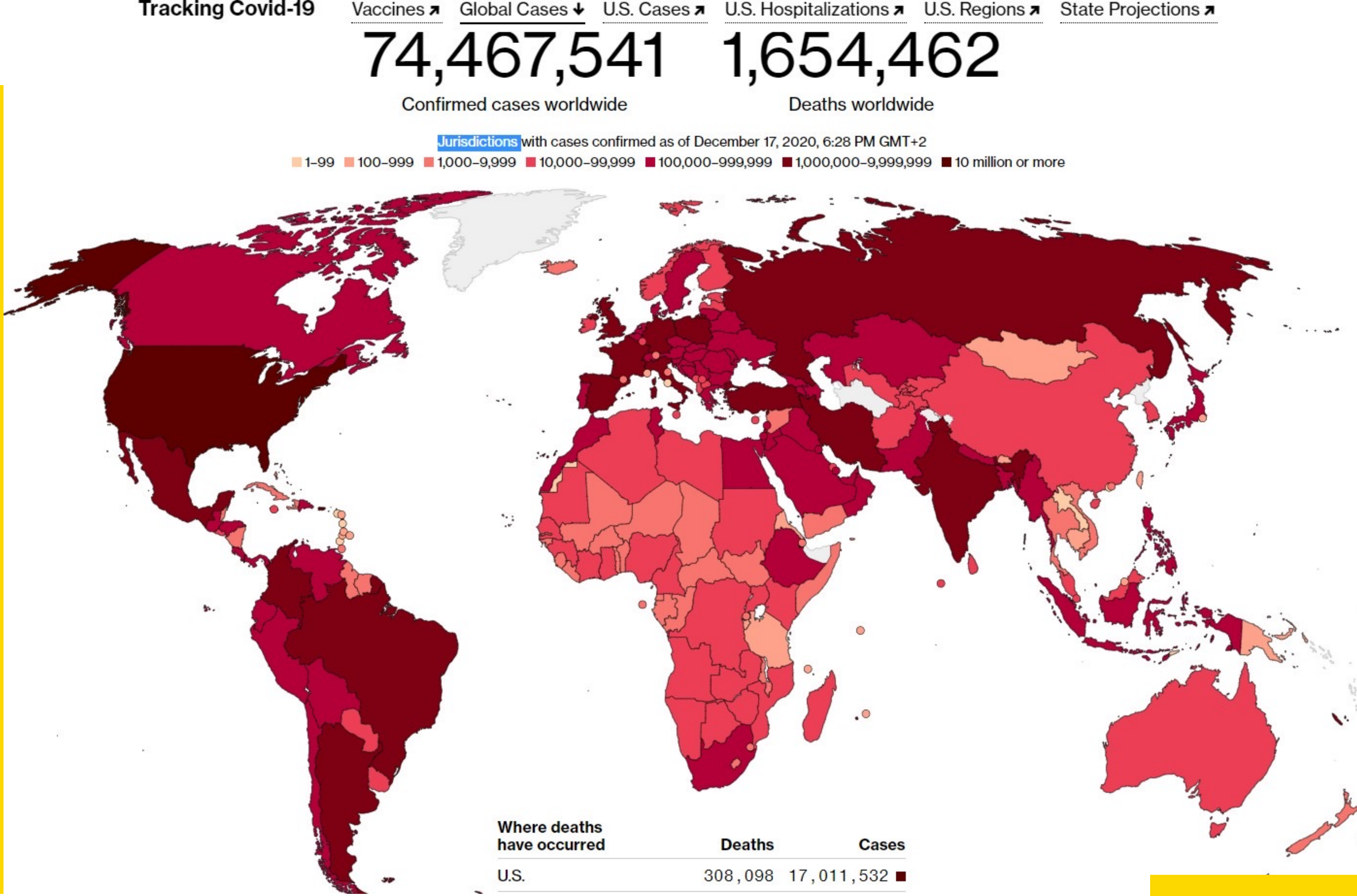
Choose country/bloc or select **up to six** to compare

New deaths attributed to Covid-19 in United States, United Kingdom and Greece

Seven-day rolling average of new deaths (per 100k)

Source: Financial Times analysis of data from the World Health Organization, the Covid Tracking Project, the Johns Hopkins CSSE, the UK Government coronavirus dashboard, the Spanish Ministry of Health and the Swedish Public Health Agency.
Data updated December 17 2020 2.06pm GMT. Interactive version: ft.com/covid19

Bloomberg, Tracking COVID-19



Tracking Covid-19

Vaccines ↗ Global Cases ↓ U.S. Cases ↗ U.S. Hospitalizations ↗ U.S. Regions ↗ State Projections ↗

74,467,541
Confirmed cases worldwide

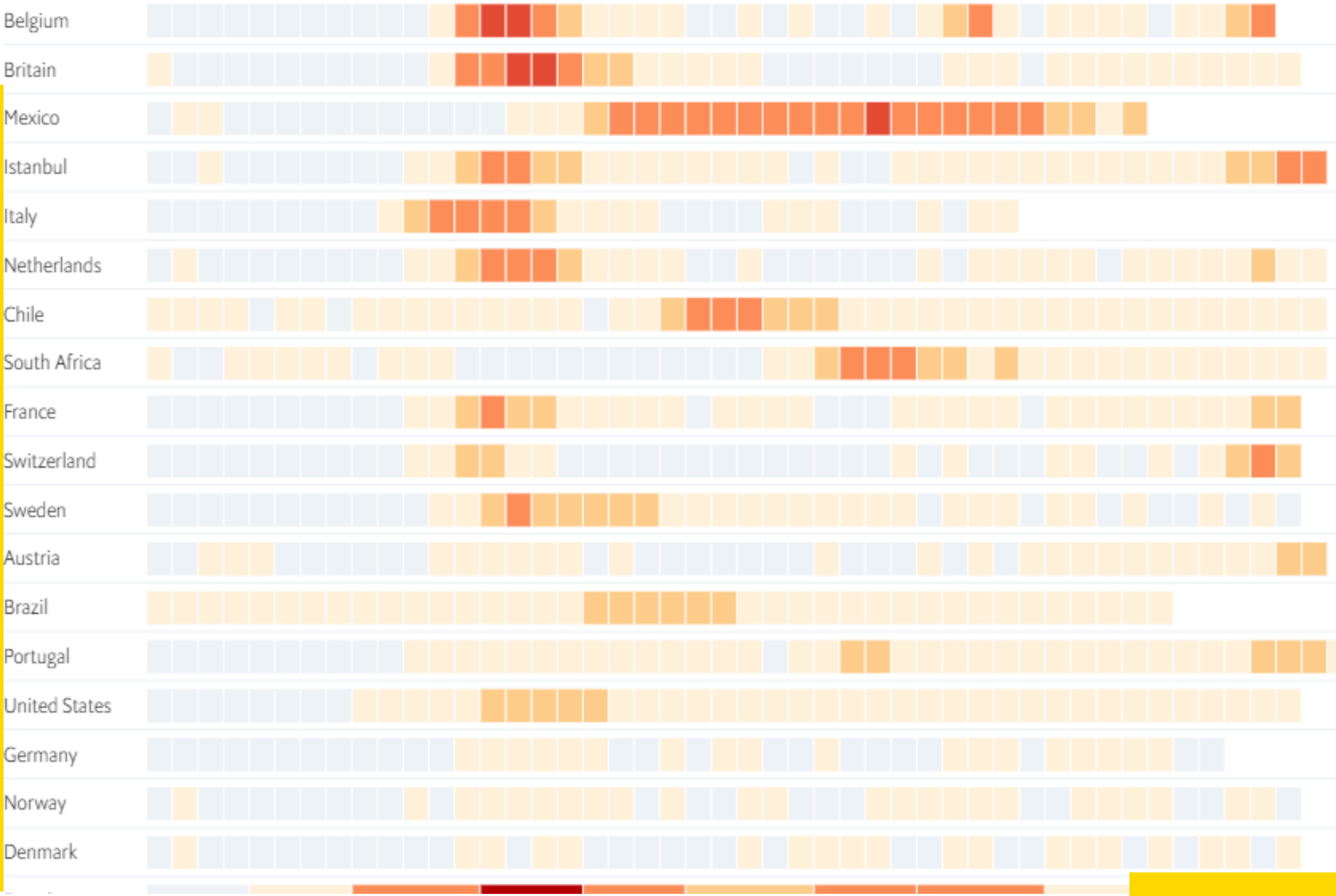
1,654,462
Deaths worldwide

Jurisdictions with cases confirmed as of December 17, 2020, 6:28 PM GMT+2

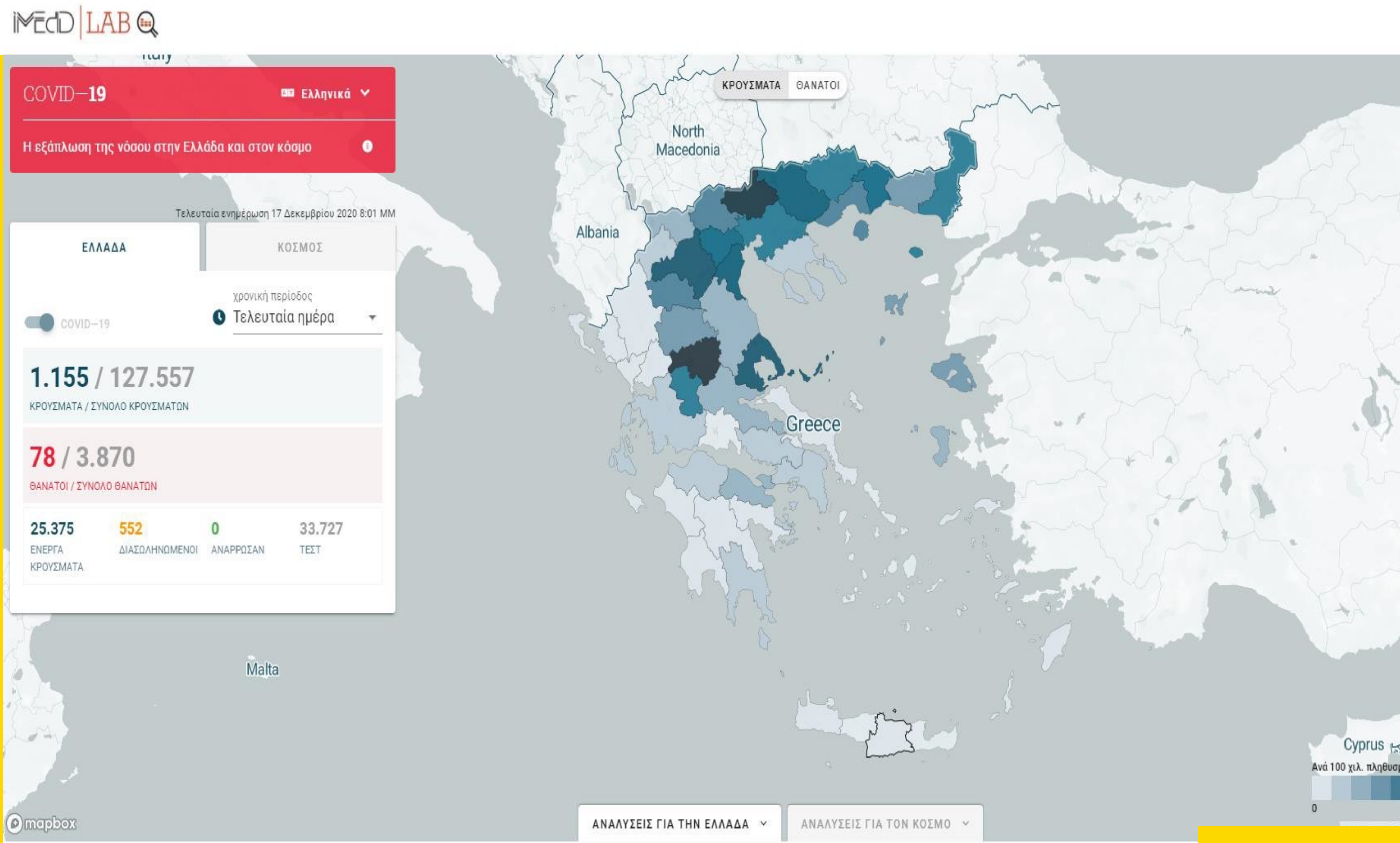
■ 1–99 ■ 100–999 ■ 1,000–9,999 ■ 10,000–99,999 ■ 100,000–999,999 ■ 1,000,000–9,999,999 ■ 10 million or more

Where deaths have occurred	Deaths	Cases
U.S.	308,098	17,011,532 ■

Economist, Tracking excess deaths across countries



iMEdD Lab, COVID-19: The spread of the disease in Greece and worldwide



COVID–19

Η εξάπλωση της νόσου στην Ελλάδα και στον κόσμο

Τελευταία ενημέρωση 17 Δεκεμβρίου 2020 8:01 ΜΜ

χρονική περίοδος: Τελευταία ημέρα

1.155 / 127.557
ΚΡΟΥΣΜΑΤΑ / ΣΥΝΟΛΟ ΚΡΟΥΣΜΑΤΩΝ

78 / 3.870
ΘΑΝΑΤΟΙ / ΣΥΝΟΛΟ ΘΑΝΑΤΩΝ

25.375 ΕΝΕΡΓΑ ΚΡΟΥΣΜΑΤΑ

552 ΔΙΑΣΩΛΗΝΩΜΕΝΟΙ

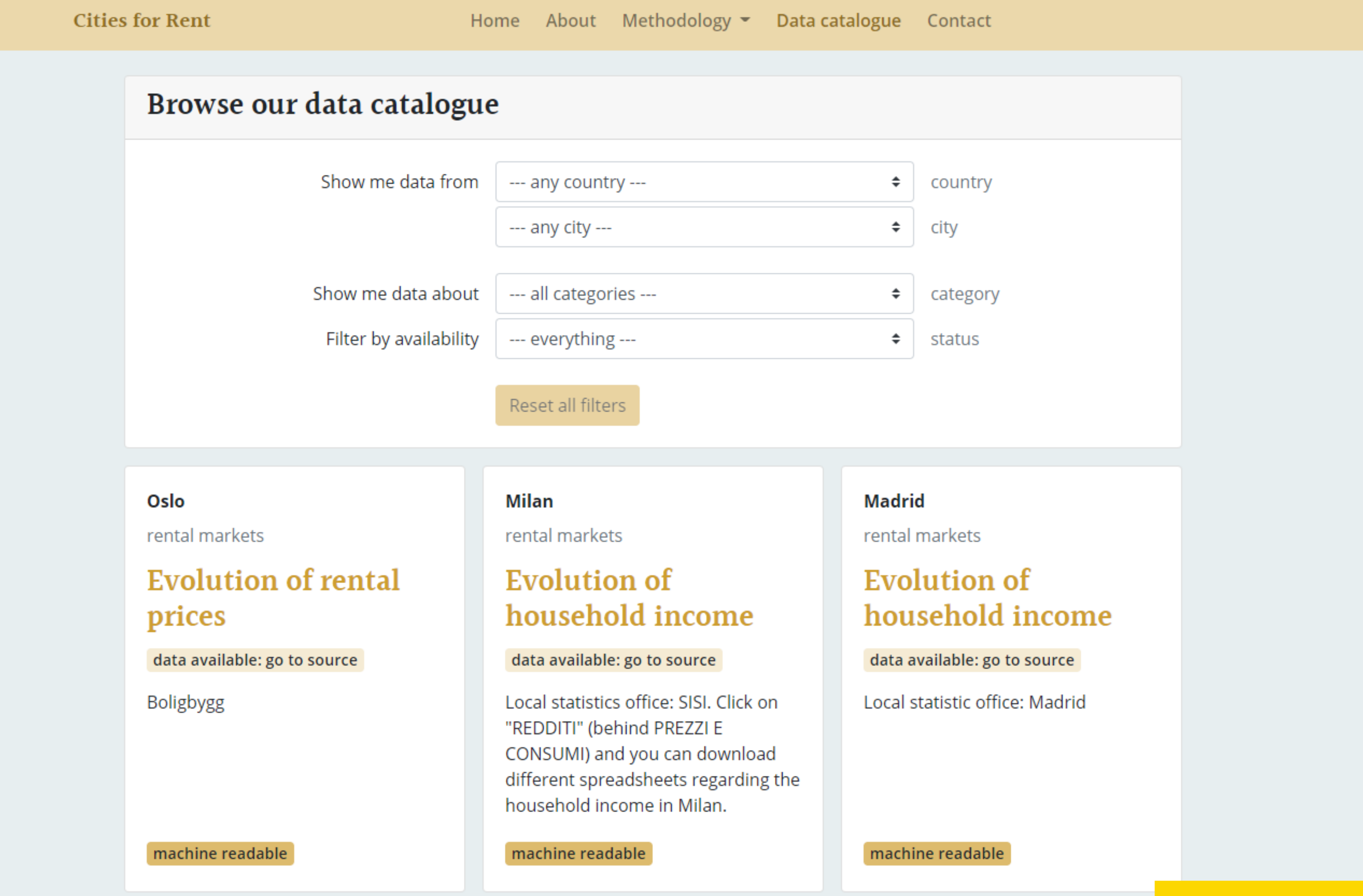
0 ΑΝΑΡΡΩΣΑΝ

33.727 ΤΕΣΤ

BEFORE AND BEYOND COVID-19

Why the superrich are inevitable, The Pudding

Cities for Rent, Arena for Journalism in Europe



Browse our data catalogue

Show me data from --- any country --- country

--- any city --- city

Show me data about --- all categories --- category

Filter by availability --- everything --- status

Reset all filters

Oslo

rental markets

Evolution of rental prices

data available: go to source

Boligbygg

machine readable

Milan

rental markets

Evolution of household income

data available: go to source

Local statistics office: SISI. Click on "REDDITI" (behind PREZZI E CONSUMI) and you can download different spreadsheets regarding the household income in Milan.

machine readable

Madrid

rental markets

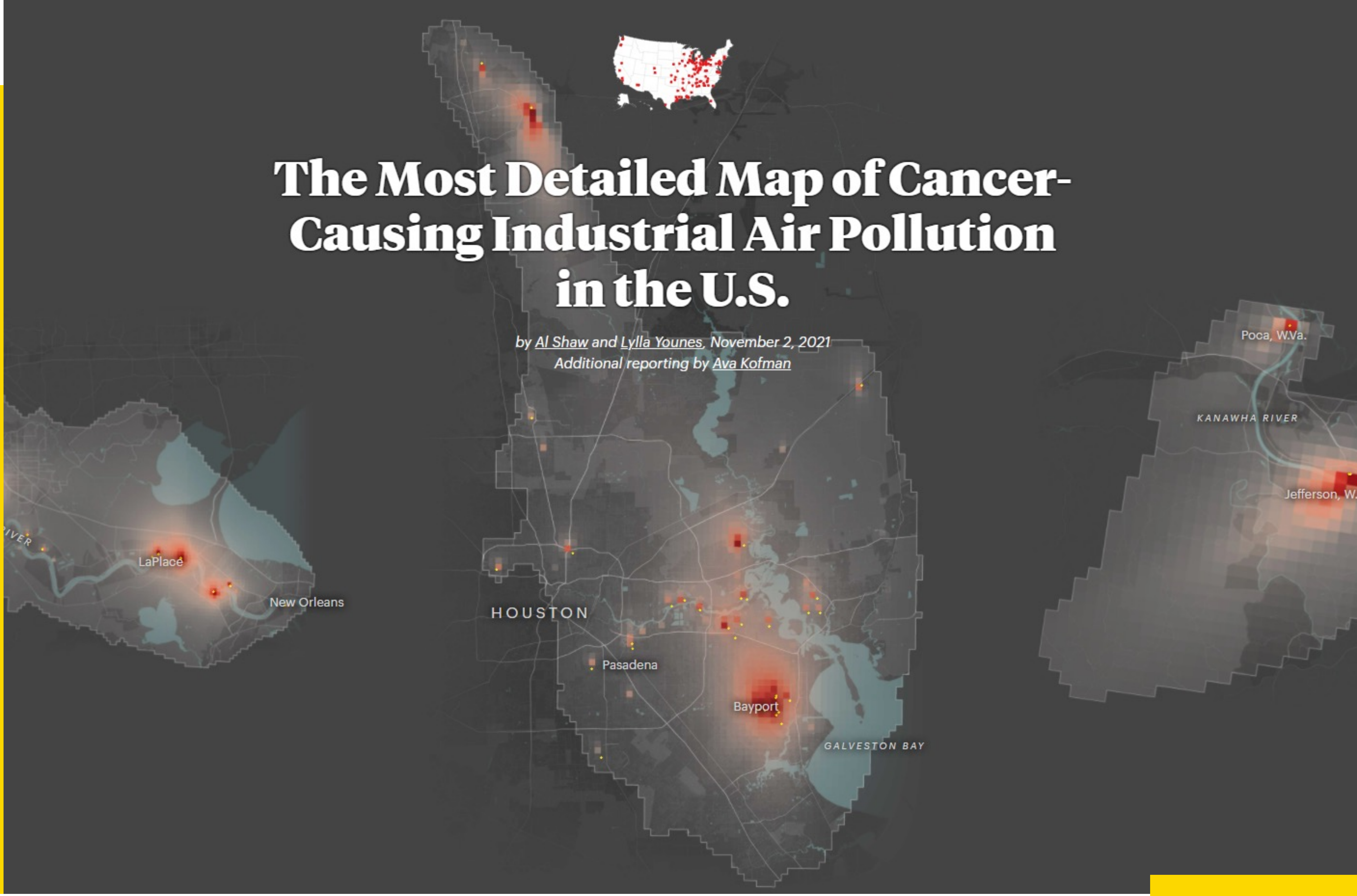
Evolution of household income

data available: go to source

Local statistic office: Madrid

machine readable

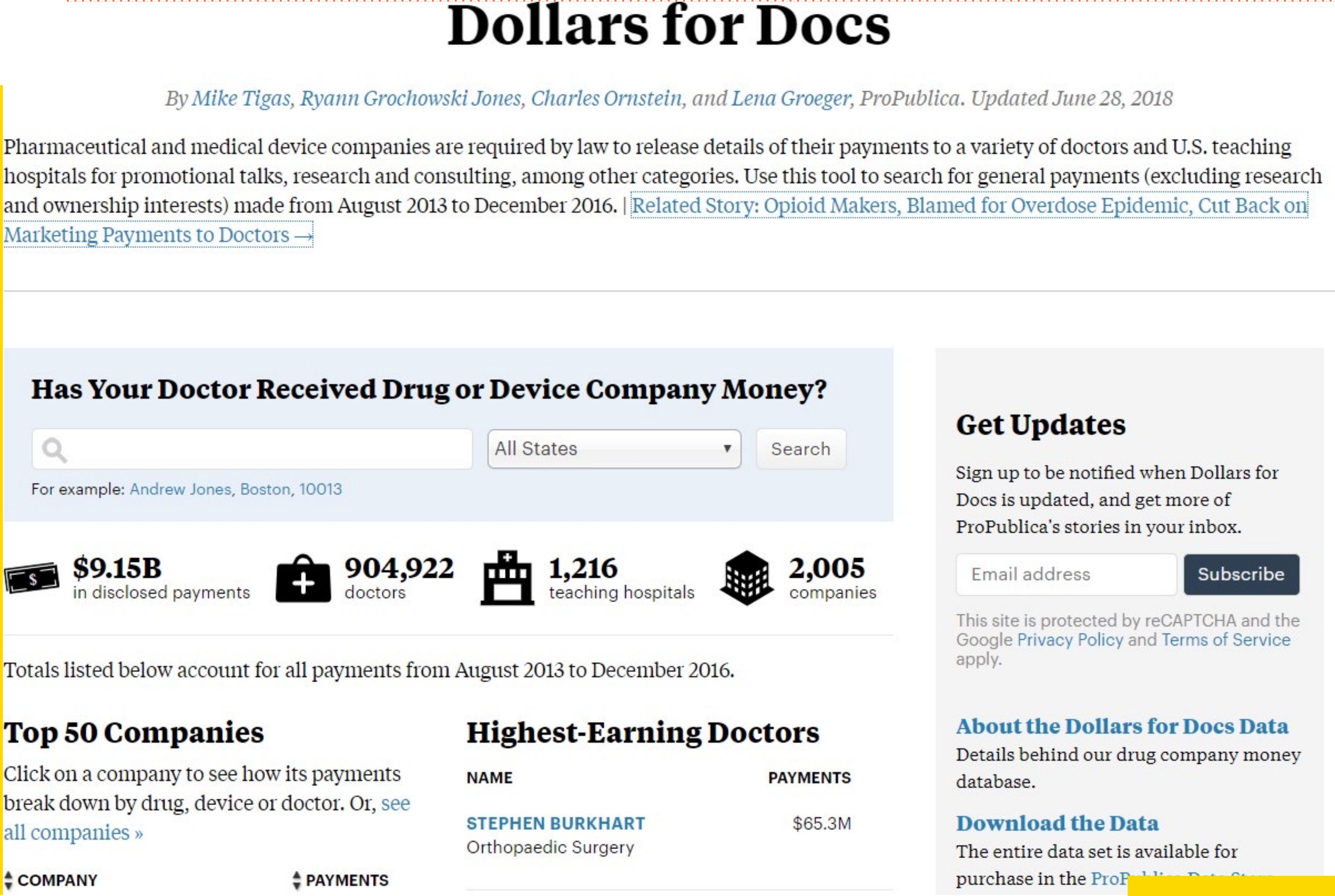
The Most Detailed Map of Cancer, ProPublica



The Most Detailed Map of Cancer-Causing Industrial Air Pollution in the U.S.

by Al Shaw and Lylla Younes, November 2, 2021
Additional reporting by Ava Kofman

Dollars for Docs, ProPublica News Apps



Dollars for Docs

By Mike Tigas, Ryann Grochowski Jones, Charles Ornstein, and Lena Groeger, ProPublica. Updated June 28, 2018

Pharmaceutical and medical device companies are required by law to release details of their payments to a variety of doctors and U.S. teaching hospitals for promotional talks, research and consulting, among other categories. Use this tool to search for general payments (excluding research and ownership interests) made from August 2013 to December 2016. | Related Story: Opioid Makers, Blamed for Overdose Epidemic, Cut Back on Marketing Payments to Doctors →

Has Your Doctor Received Drug or Device Company Money?

All States ▾ Search

For example: Andrew Jones, Boston, 10013

$9.15B in disclosed payments

904,922 doctors

1,216 teaching hospitals

2,005 companies

Totals listed below account for all payments from August 2013 to December 2016.

Top 50 Companies

Click on a company to see how its payments break down by drug, device or doctor. Or, see all companies »

⬍ COMPANY	⬍ PAYMENTS

Highest-Earning Doctors

NAME	PAYMENTS
STEPHEN BURKHART Orthopaedic Surgery	$65.3M

Get Updates

Sign up to be notified when Dollars for Docs is updated, and get more of ProPublica's stories in your inbox.

Email address Subscribe

This site is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

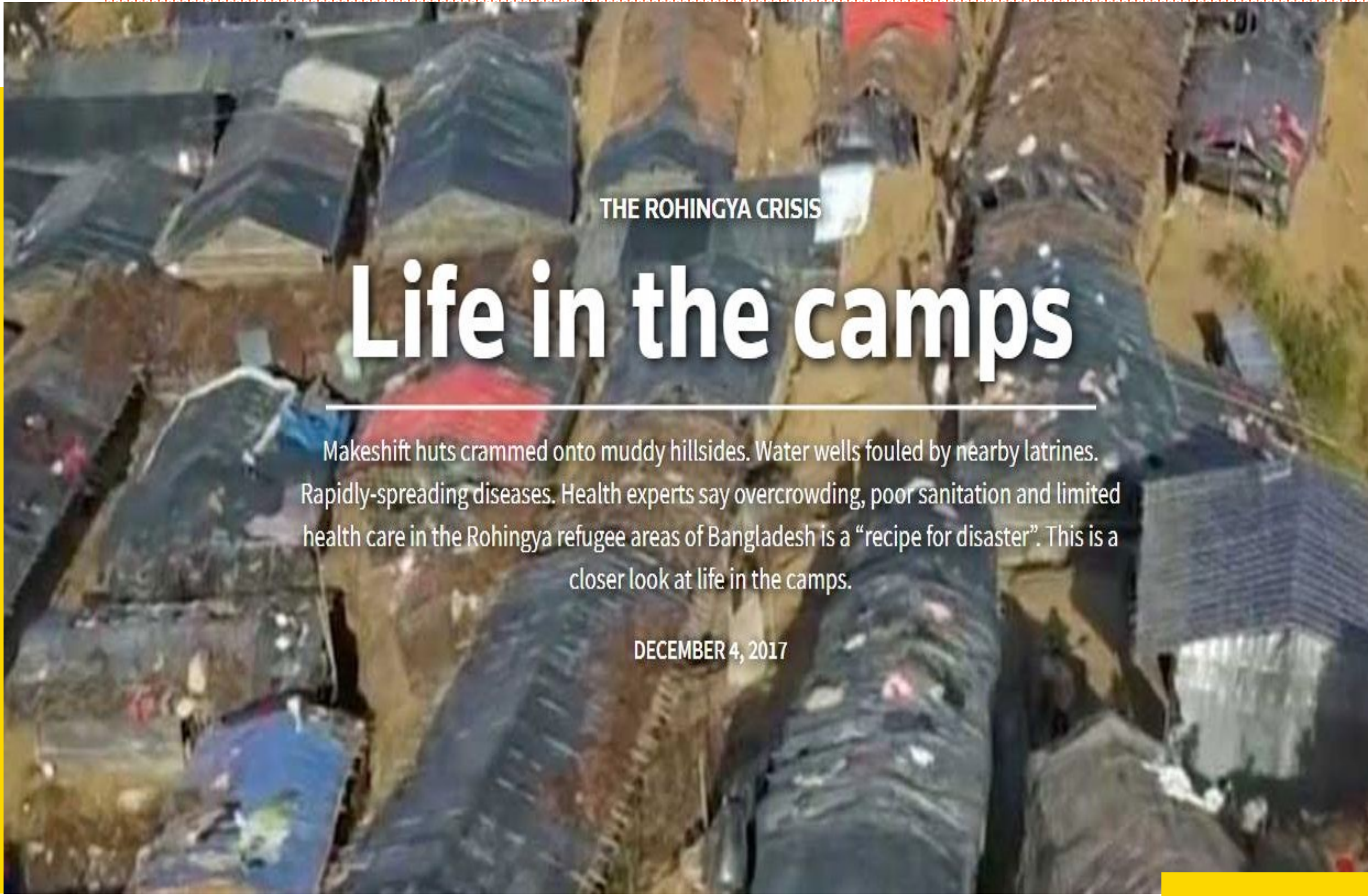
About the Dollars for Docs Data

Details behind our drug company money database.

Download the Data

The entire data set is available for purchase in the ProP

Life in the camps, Reuters Graphics



THE ROHINGYA CRISIS

Life in the camps

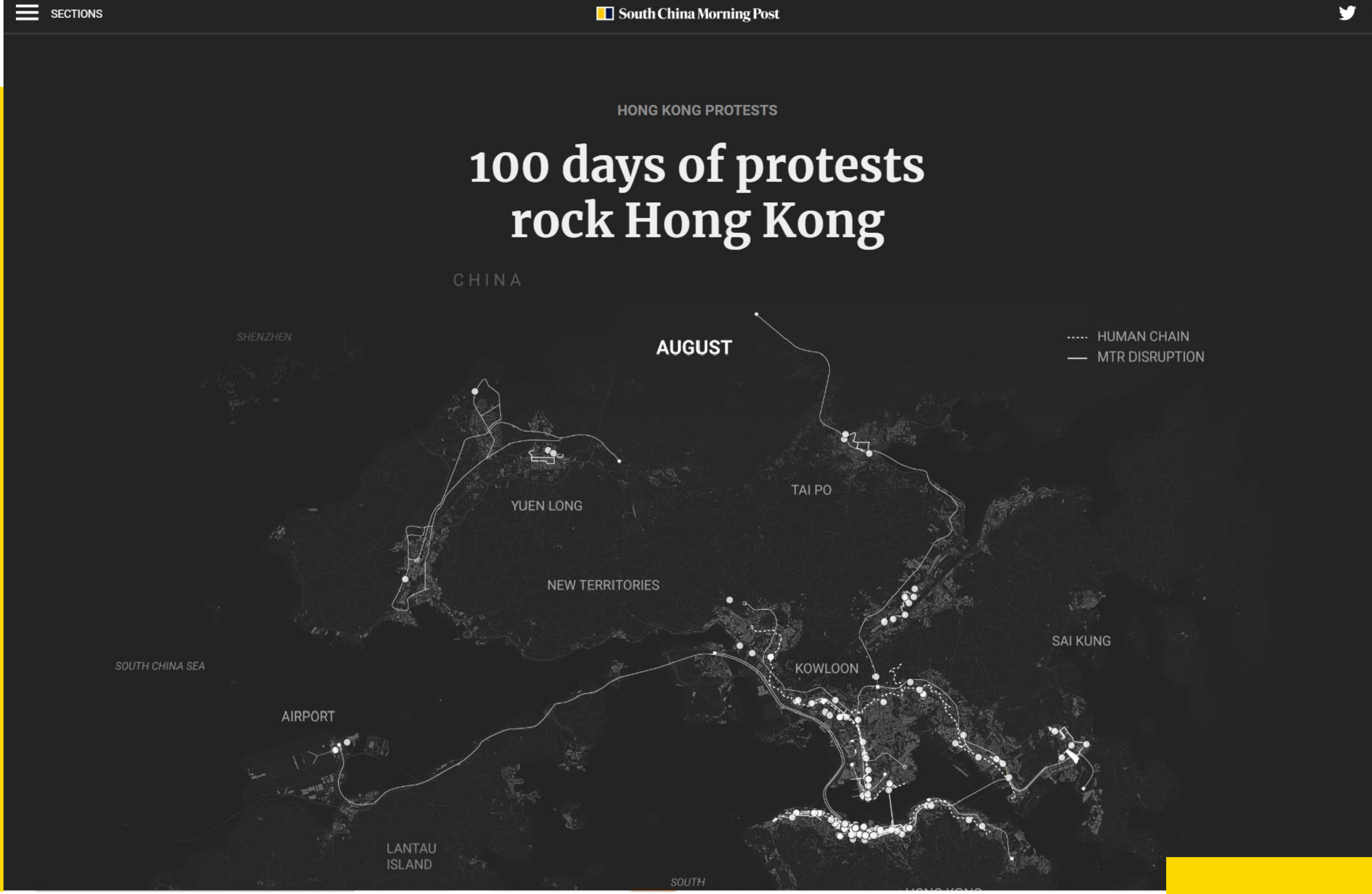
Makeshift huts crammed onto muddy hillsides. Water wells fouled by nearby latrines. Rapidly-spreading diseases. Health experts say overcrowding, poor sanitation and limited health care in the Rohingya refugee areas of Bangladesh is a “recipe for disaster”. This is a closer look at life in the camps.

DECEMBER 4, 2017

Spies in the Skies, BuzzFeed News

Mark Pernice for BuzzFeed News

100 days of protests rock Hong Kong, South China Morning Post

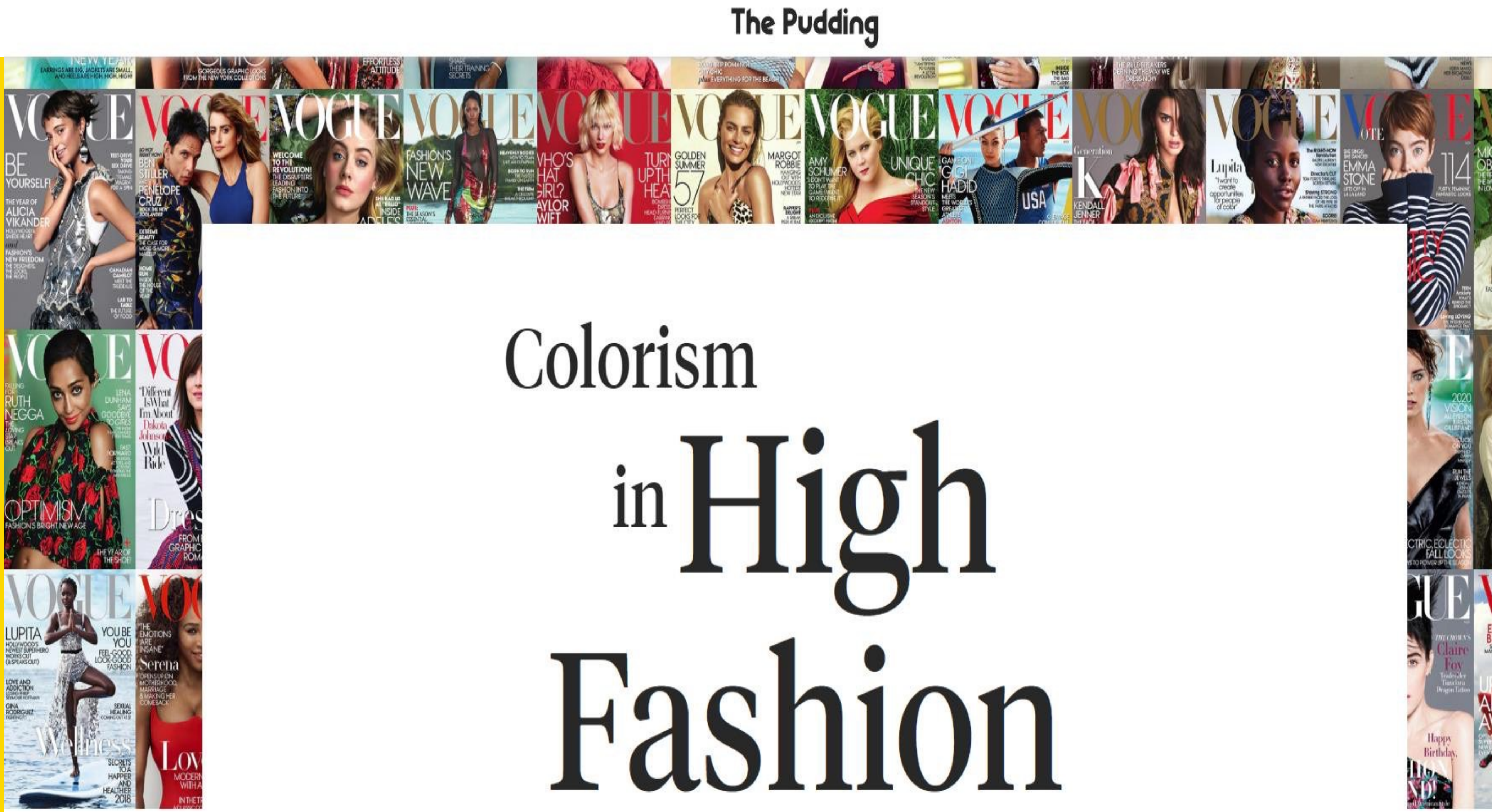

SECTIONS

South China Morning Post

HONG KONG PROTESTS

100 days of protests rock Hong Kong

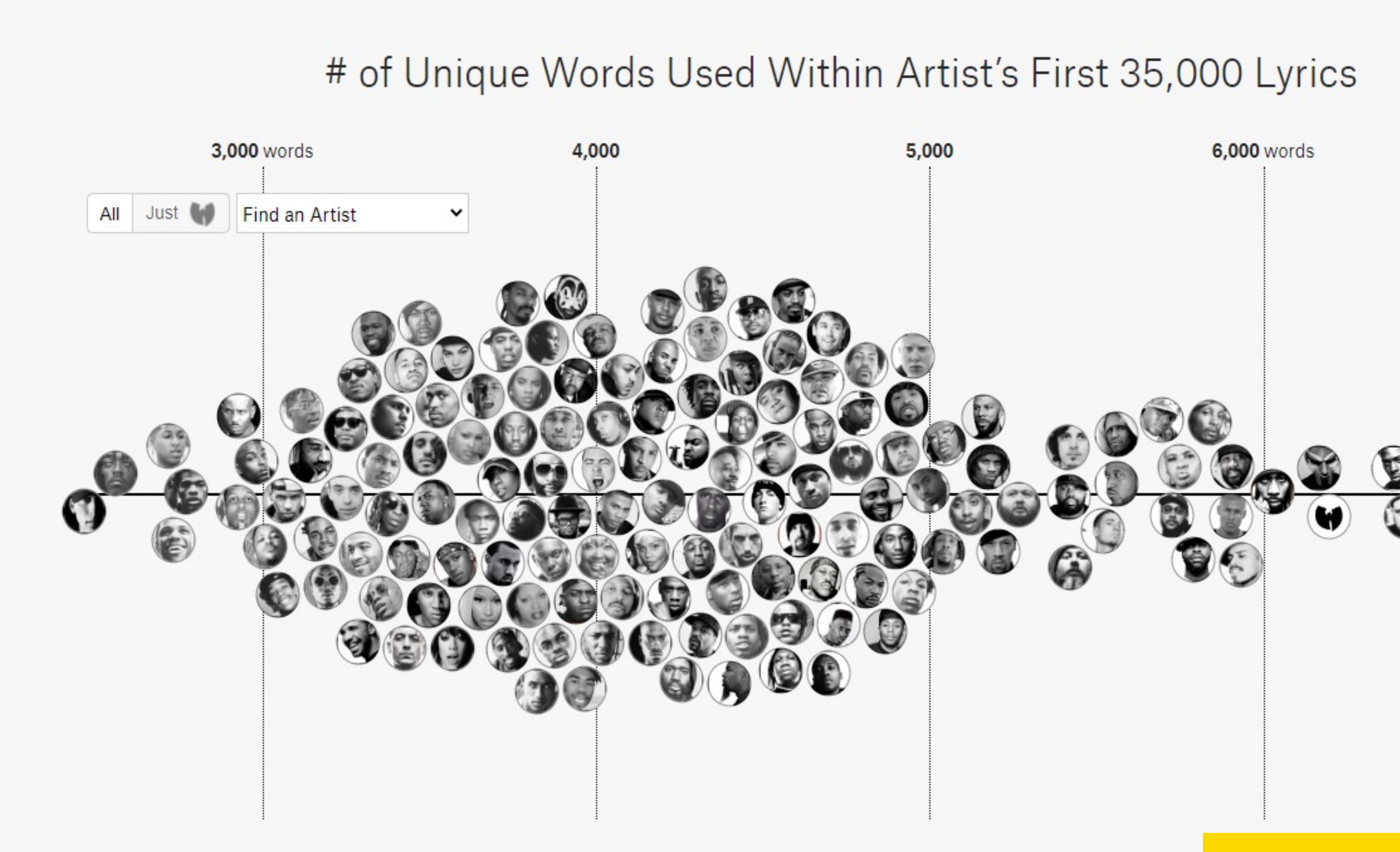
Colorism in High Fashion, The Pudding



The Pudding

Colorism in High Fashion

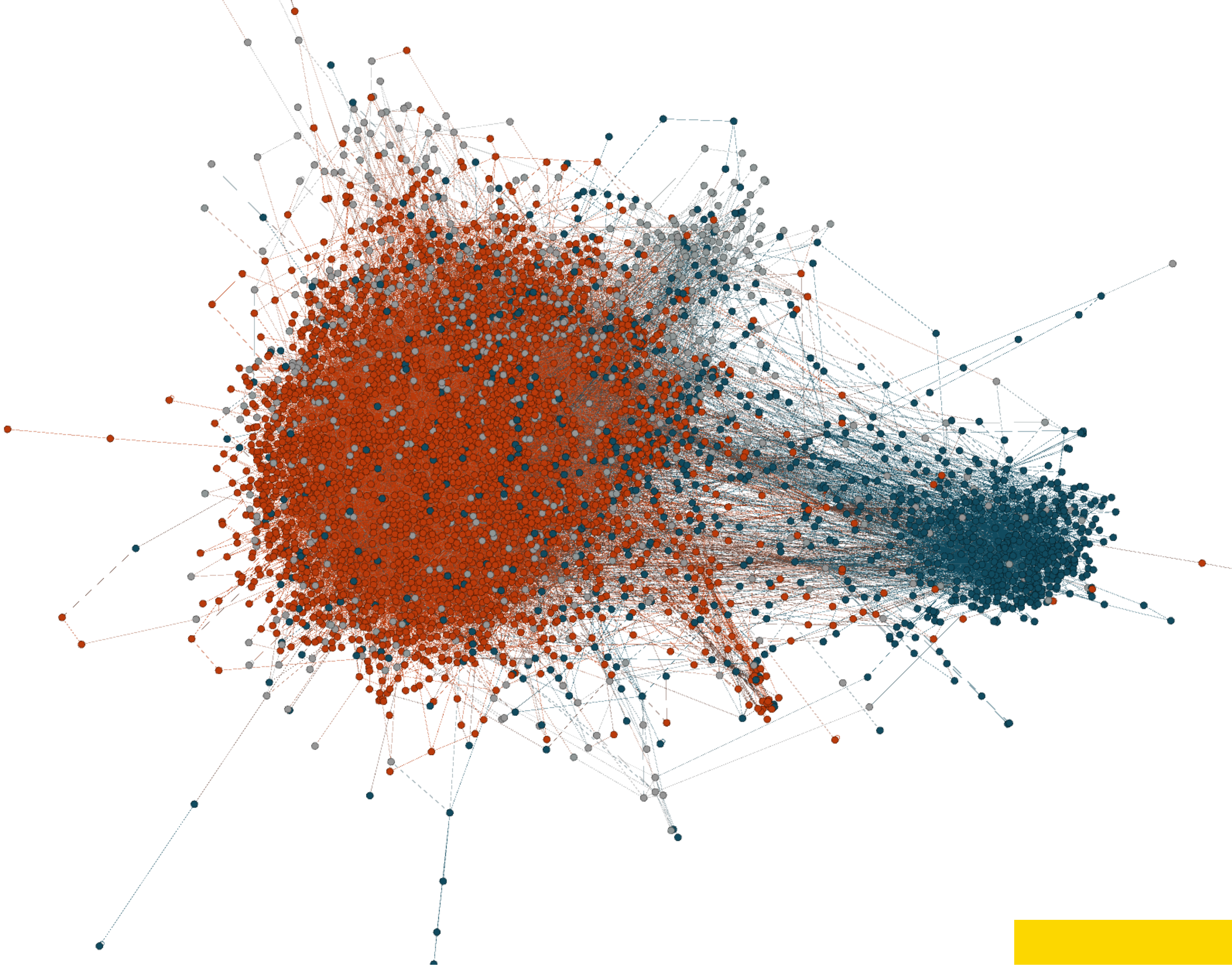
The Largest Vocabulary In Hip Hop, The Pudding



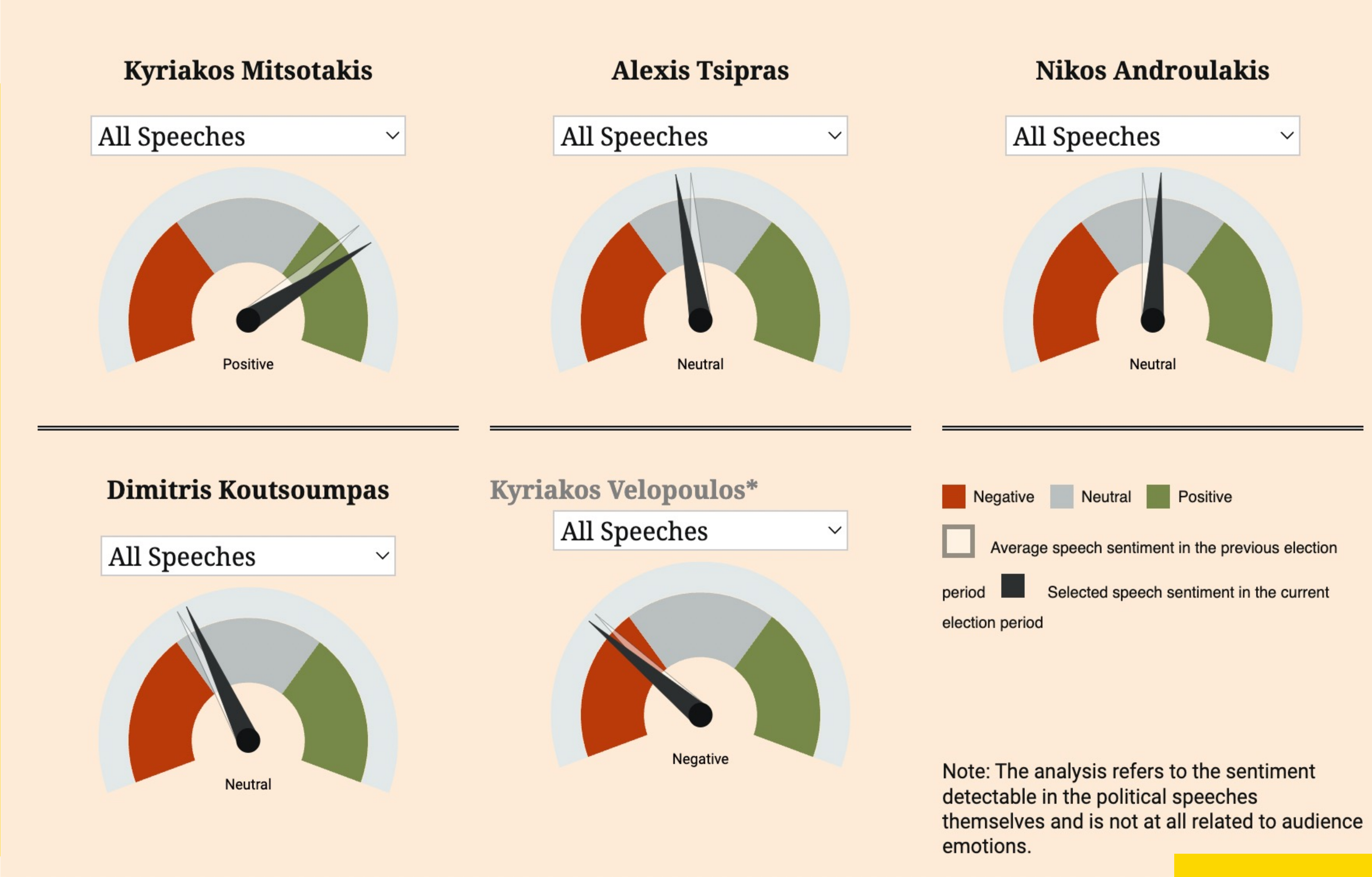
# of Unique Words Used Within Artist's First 35,000 Lyrics

[The Greek wiretapping scandal on Twitter](), iMEdD Lab

in partnership with Datalab



Analyzing the Greek political discourse: a collaboration between humans and AI, iMEdD Lab



Kyriakos Mitsotakis

All Speeches

Alexis Tsipras

All Speeches

Nikos Androulakis

All Speeches

Dimitris Koutsoumpas

All Speeches

Kyriakos Velopoulos*

All Speeches

🟥 Negative ⬜ Neutral 🟩 Positive

□ Average speech sentiment in the previous election period ■ Selected speech sentiment in the current election period

Note: The analysis refers to the sentiment detectable in the political speeches themselves and is not at all related to audience emotions.

BOOKMARKS' LIST

READINGS & TOOLS

- Bounegru, L. & Gray, J. (2021). *The Data Journalism Handbook: Towards a Critical Data Practice*. Amsterdam: Amsterdam University Press. Openly available by EJC, datajournalism.com & Google News Initiative: https://datajournalism.com/read/handbook/two (Accessed: February 12, 2023). Openly available in Greek by iMEdD Lab: https://datahandbook.lab.imedd.org/el/

- Soma, J. (n.d) "Python's Not (Just) For Unicorns", *Little Columns*. Available: http://littlecolumns.com/learn/python/

- DataCamp.com, https://www.datacamp.com/

- DataJournalism.com, https://datajournalism.com/

- Datawrapper, https://www.datawrapper.de/

- Flourish, https://flourish.studio/

- FT Vocabulary, ft.com/vocabulary

- QJIS, https://www.qgis.org/en/site/ & QJIS Tutorial, http://www.qgistutorials.com/el/

THANK YOU!