DATA COLLECTION

Petra Isenberg

Slides originally by WESLEY WILLETT

VISUAL ANALYTICS

WHERE DOES DATA COME FROM?

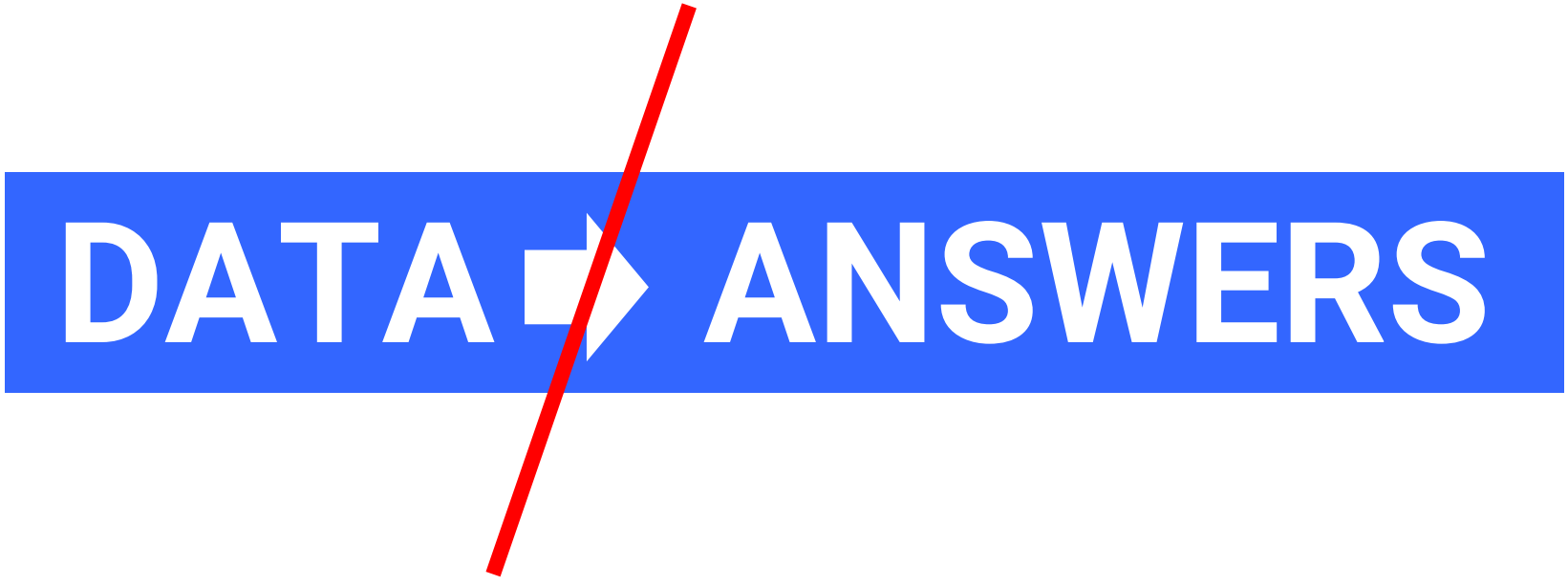
We tend to think of data as a thing…

in a database…

somewhere…

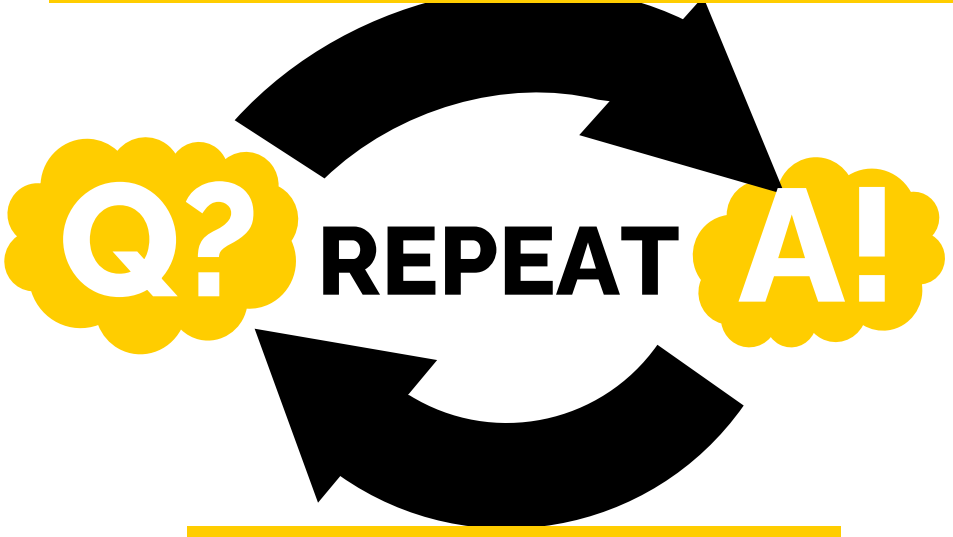
WHY DO YOU NEED DATA?

(HINT: Usually, because you have a question you need to answer!)


DATA
ANSWERS

ANALYSIS IS A CYCLE

GATHERING DATA, APPLYING STATISTICAL TOOLS, AND CONSTRUCTING GRAPHICS TO ADDRESS QUESTIONS



INSPECT "ANSWERS" AND ASSESS NEW QUESTIONS

(SOMETIMES YOU'LL ALREADY START WITH DATA...)

"EXPLORATORY DATA ANALYSIS"

JOHN TUKEY

We already saw this...

(...BUT OFTEN YOU START
WITH A QUESTION AND NEED
TO COLLECT DATA TO FIT IT)

HOW TO OBTAIN DATA?

COLLECT IT

- OBSERVATION
- SURVEYS
- LOGGING
- SENSORS
- CROWDSOURCING

FIND OR EXTRACT IT

- OPEN CORPUSES
- DATA RETAILERS
- APIS
- SCRAPING THE WEB

GENERATE IT

- SIMULATIONS

ALL OF THESE HAVE

PROS/CONS

THIS LIST IS NOT EXHAUSTIVE

This lecture is intended to expose you to just a few useful data sources and collection methods.

COLLECTING DATA

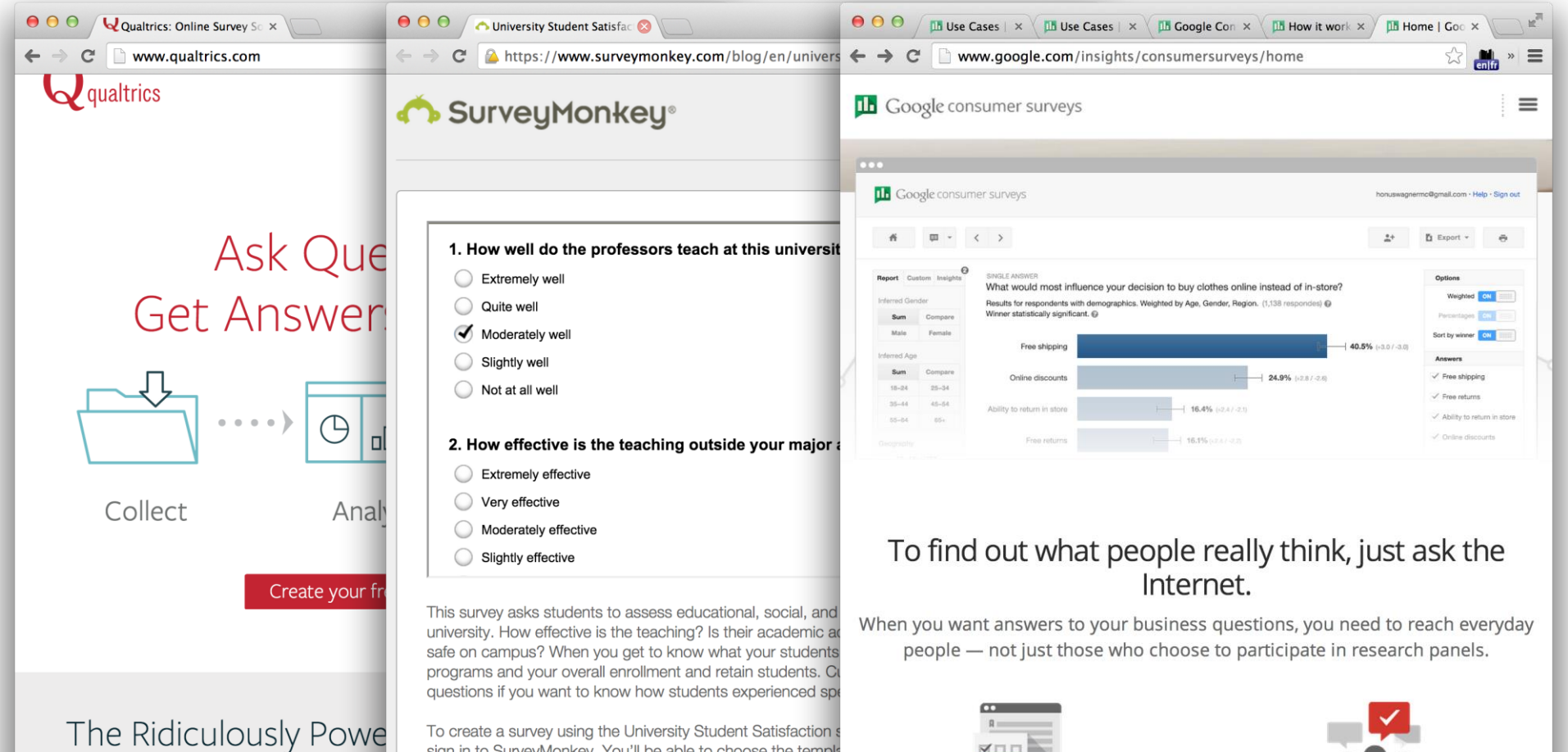
Choosing the best way to capture information you need.

SURVEYS

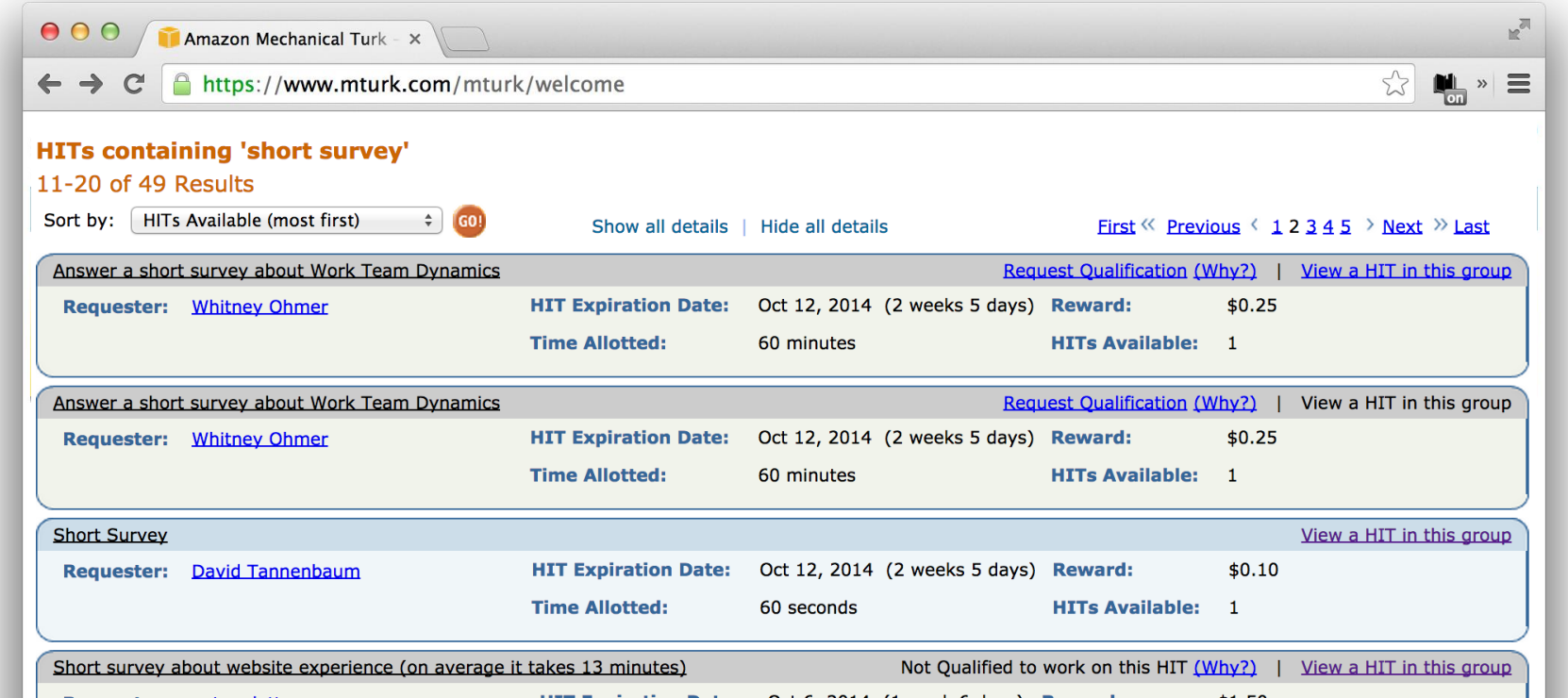
Paper surveys / In person interviews

STILL ONE OF THE BEST WAYS TO GET DETAILED DATA OR DATA ABOUT SENSITIVE SUBJECTS

SURVEYS ONLINE



CROWDSOURCING DATA COLLECTION



WEB LOGGING

Tracking Visits, Click-Throughs, and Traffic Patterns and other measures of User Activity.

- Google Analytics
- Open Web Analytics
- and many others…

EDITS & ACCESSS LOGS ON WIKIPEDIA

SENSORS

- Weather stations
- Personal activity trackers
- Cameras
- Mobile phones

HOW TO OBTAIN DATA?

COLLECT IT

- OBSERVATION
- SURVEYS
- LOGGING
- SENSORS
- CROWDSOURCING

FIND OR EXTRACT IT

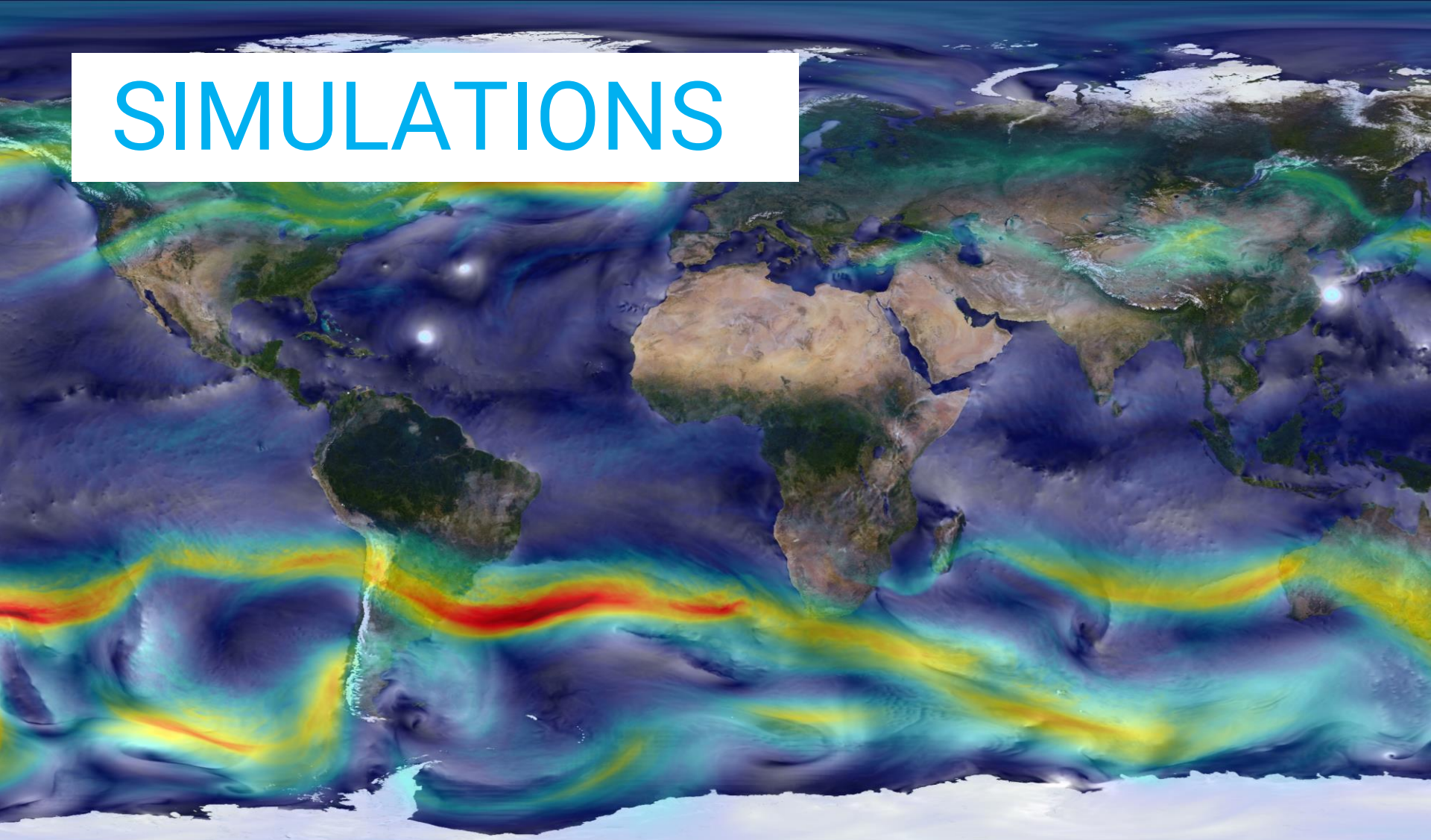
- OPEN CORPUSES
- DATA RETAILERS
- APIS
- SCRAPING THE WEB

GENERATE IT

- SIMULATIONS

GENERATING DATA

SIMULATIONS



http://www.nasa.gov/content/a-portrait-of-global-winds/

The Upshot

EDITED BY DAVID LEONHARDT

Is It Better to Rent or Buy?

By MIKE BOSTOCK, SHAN CARTER and ARCHIE TSE

The choice between buying a home and renting one is among the biggest financial decisions that many adults make. But the costs of buying are more varied and complicated than for renting, making it hard to tell which is a better deal. To help you answer this question, our calculator takes the most important costs associated with buying a house and computes the equivalent monthly rent. RELATED ARTICLE

Home Price

A very important factor, but not

EQUIVALENT RENT
– $8K

If you can rent a similar home for less than ...

HOW TO OBTAIN DATA?

COLLECT IT

- OBSERVATION
- SURVEYS
- LOGGING
- SENSORS
- CROWDSOURCING

FIND OR EXTRACT IT

- OPEN CORPUSES
- DATA RETAILERS
- APIS
- SCRAPING THE WEB

GENERATE IT

- SIMULATIONS

FINDING AND EXTRACTING EXISTING DATA

LARGE OPEN CORPUSES

DBPEDIA

DBpedia | Browse using | Formats | Faceted Browser | Sparql Endpoint

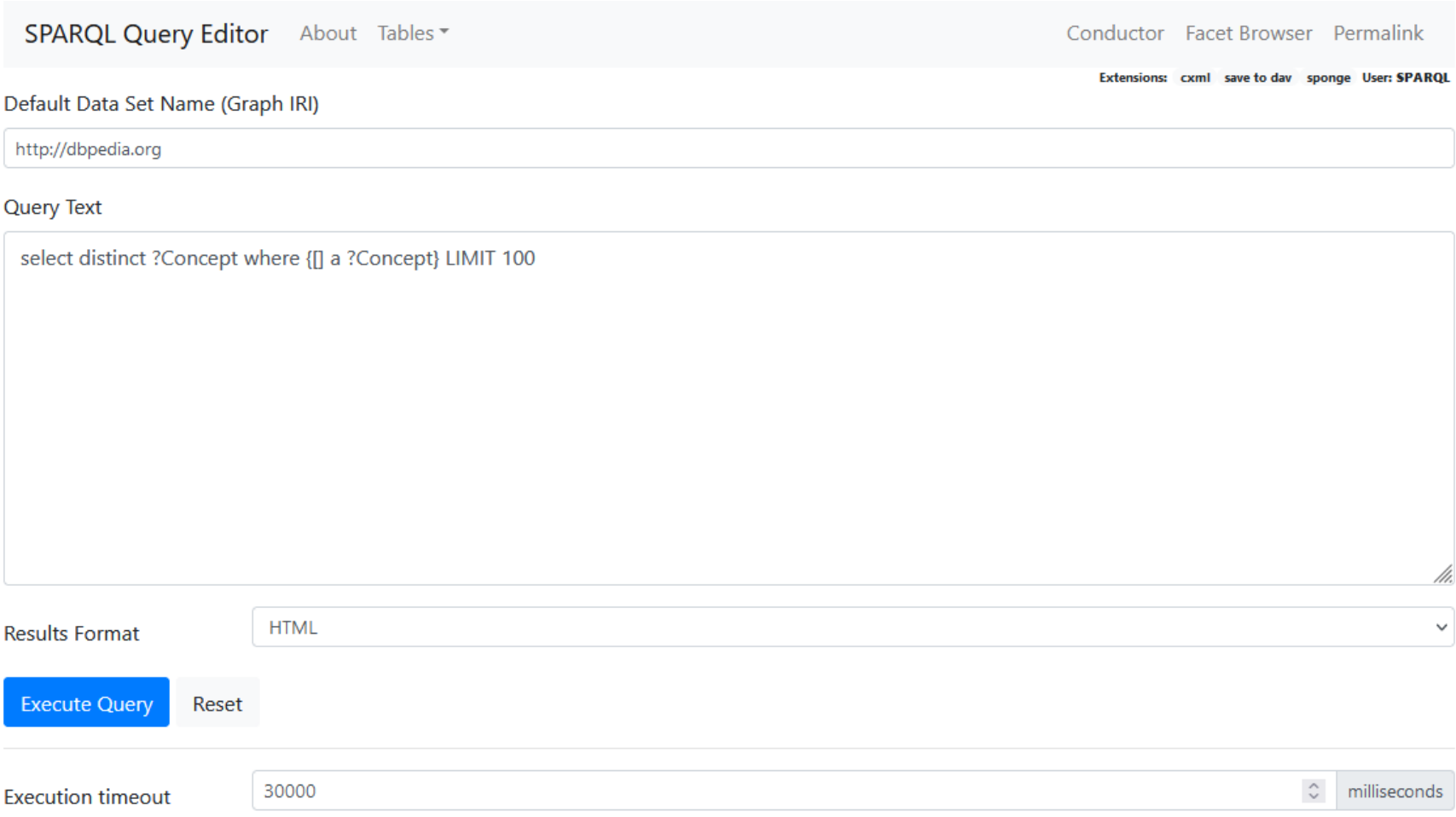
About: Iceland

An Entity of Type: populated place, from Named Graph: http://dbpedia.org, within Data Space: dbpedia.org

Iceland (Icelandic: Ísland; ['istlant]) is a Nordic island country in the North Atlantic Ocean and the most sparsely populated country in Europe. Iceland's capital and largest city is Reykjavík, which (along with its surrounding areas) is home to over 65% of the population. Iceland is the only part of the Mid-Atlantic Ridge that rises above sea level, and its central volcanic plateau is erupting almost constantly. The interior consists of a plateau characterised by sand and lava fields, mountains, and glaciers, and many glacial rivers flow to the sea through the lowlands. Iceland is warmed by the Gulf Stream and has a temperate climate, despite a high latitude just outside the Arctic Circle. Its high latitude and marine influence keep summers chilly, and most of its islands have a polar

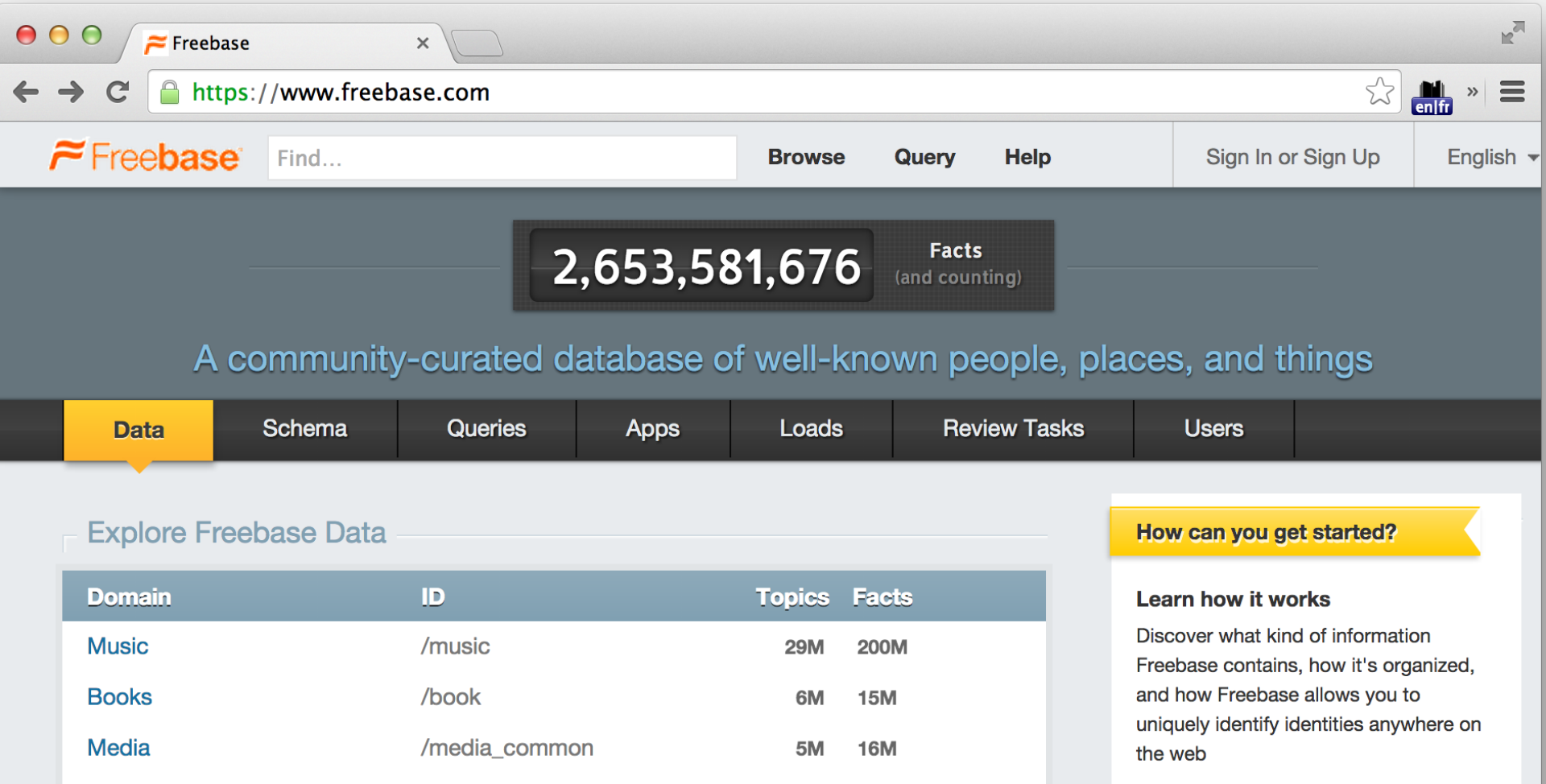
Property	Value
dbo:PopulatedPlace/area	• 102775.0 • 102819.93799222886
dbo:PopulatedPlace/populationDensity	• 3.5 • 3.552139858590502
dbo:abstract	• Iceland (Icelandic: Ísland; ['istlant]) is a Nordic island country in the North Atlantic Ocean and the most sparsely populated country in Europe. Iceland's capital and largest city is Reykjavík, which (along with its surrounding areas) is home to over 65% of the population. Iceland is the only part of the Mid-Atlantic Ridge that rises above sea level, and its central volcanic plateau is erupting almost constantly. The interior consists of a plateau characterised by sand and lava fields, mountains, and glaciers, and many glacial rivers flow to the sea through the lowlands. Iceland is warmed by the Gulf Stream and has a temperate climate, despite a high latitude just outside the Arctic

QUERYING DBPEDIA

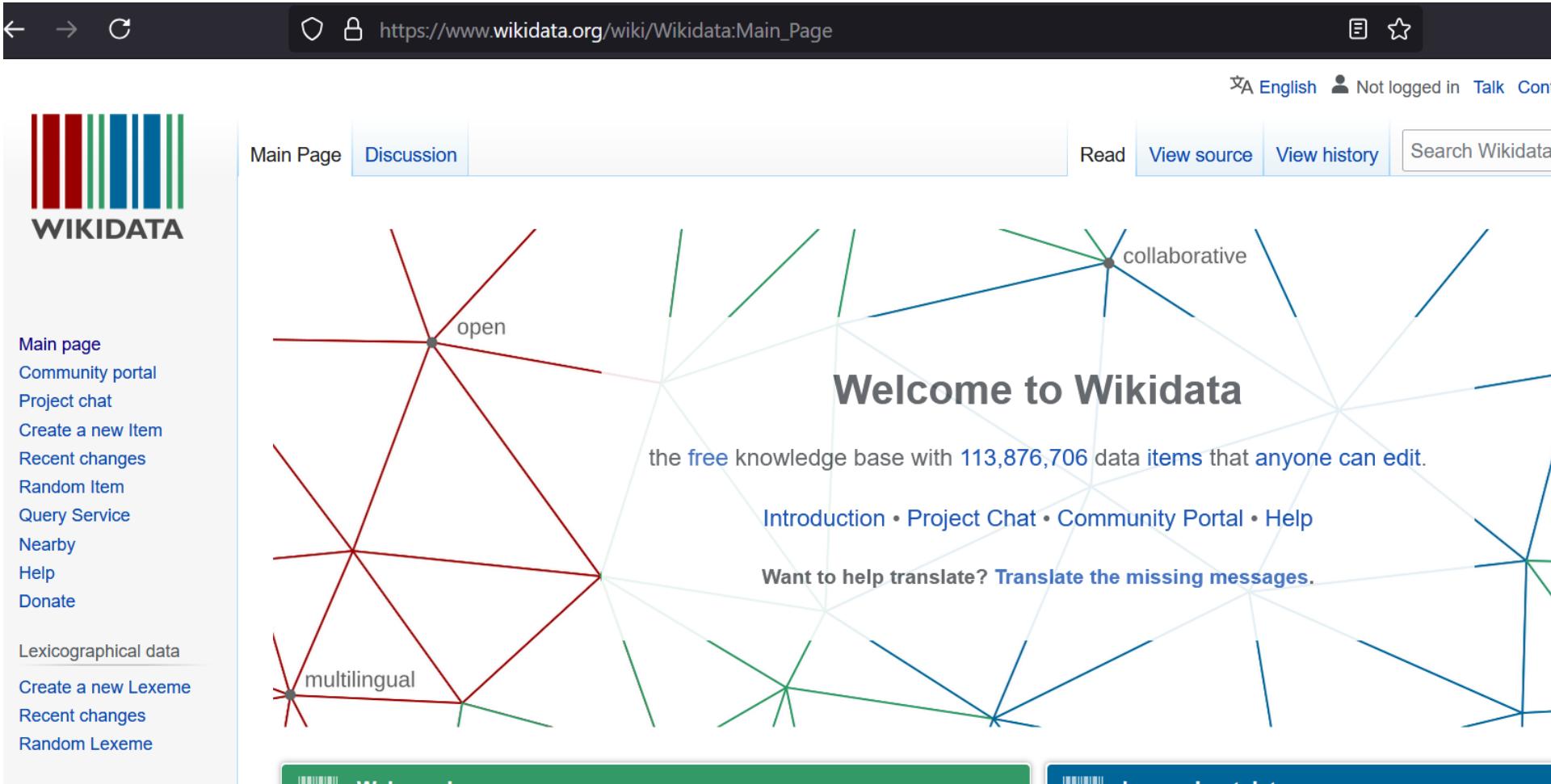


Ended in 2014

FREEBASE



WIKIDATA


https://www.wikidata.org/wiki/Wikidata:Main_Page
English
Not logged in
Talk
Main Page
Discussion
Read
View source
View history
Search Wikidata
WIKIDATA
Main page
Community portal
Project chat
Create a new Item
Recent changes
Random Item
Query Service
Nearby
Help
Donate
Lexicographical data
Create a new Lexeme
Recent changes
Random Lexeme
open
collaborative
multilingual
Welcome to Wikidata
the free knowledge base with 113,876,706 data items that anyone can edit.
Introduction • Project Chat • Community Portal • Help
Want to help translate? Translate the missing messages.

PROJECT GUTENBERG

https://www.gutenberg.org

About ▾ Search and Browse ▾ Help ▾

Quick search Go! Ways to donate PayPal

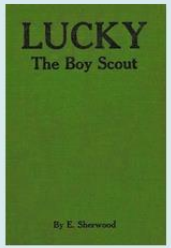
Welcome to Project Gutenberg

Project Gutenberg is a library of over 70,000 free eBooks

Choose among free epub and Kindle eBooks, download them or read them online. You will find the world's great literature here, with focus on older works for which U.S. copyright has expired. Thousands of volunteers digitized and diligently proofread the eBooks, for you to enjoy.

Pas perdus by Fagus

Troijan Helena yksityiselämäss



Lucky, the Boy Scout by Elmer

Aunuksen helmi by Simo Eronen

Storia delle scienze ad uso dei licei

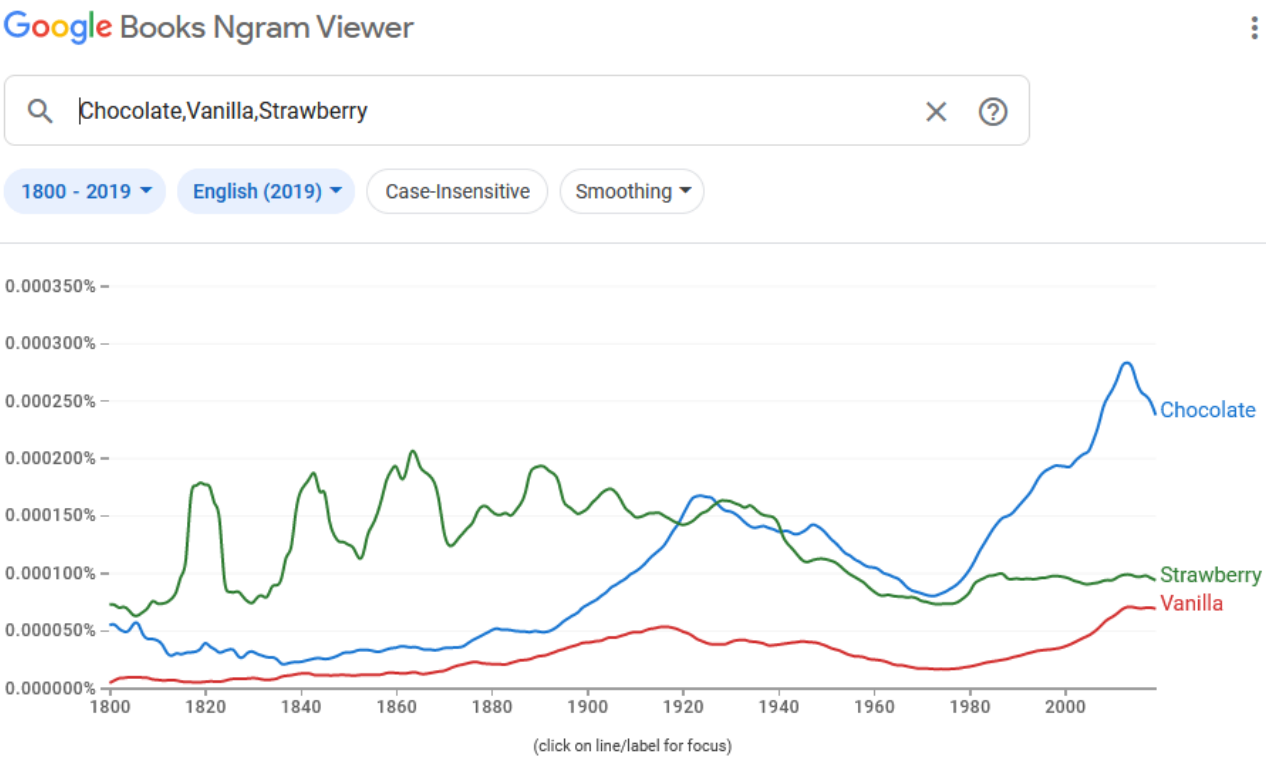
Chambers's journal of popular

Chambers's journal of popular

Le dernier rapport d'un Européen sur

Festival plays by Marguerite Merington

GOOGLE N-GRAMS



FINDING AND EXTRACTING EXISTING DATA

GOVERNMENT AND INTERNATIONAL DATA INITIATIVES

DATA.WORLDBANK.ORG

https://da

THE WORLD BANK IBRD • IDA | Data

New to this site? Start Here

MOST RECENT

Taking the right turns: Harnessing available evider and guidelines to bridge ga in gender data production

Anna Tabitha Bonfert, Heather Moy Miriam Muller, Sep 29, 2022

How citizen science can hel realize the full potential of data

Haishan Fu, Craig Hammer, Edward Anderson, Sep 28, 2022

Decades of learning and experience from the Trust

https://data.oecd.org

OECD.org | Data | Publications | More sites | News | Job vacancies

OECD Data

› Français

Find, compare and share the latest **OECD data**: charts, maps, tables and related publications …

Search

Browse by topic or country

Search tips

› Catalogue of OECD databases

OECD Data

Featured charts

- **G20 GDP** falls 0.4% in the second quarter of 2022. See news.
- **Leading indicators** continue to point to weakening growth. See news.

Quarterly GDP

Total, Percentage change, previous period, Q2 2022 or latest available

Latest news

Statistical news releases

See recent statistical news releases.

Data Insights

Discover *Data Insights* featuring data visualisations related to the Covid-19 crisis.

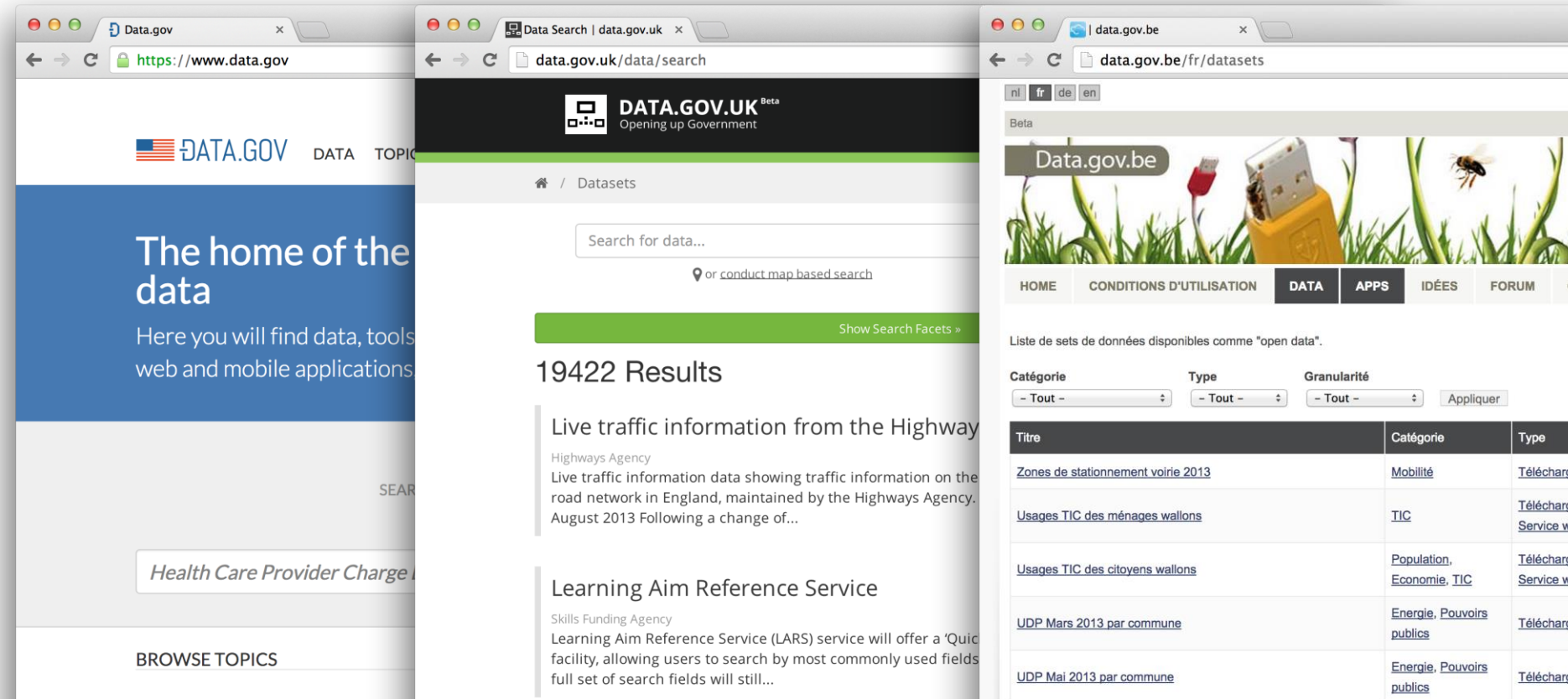
Statistical resources

Database access

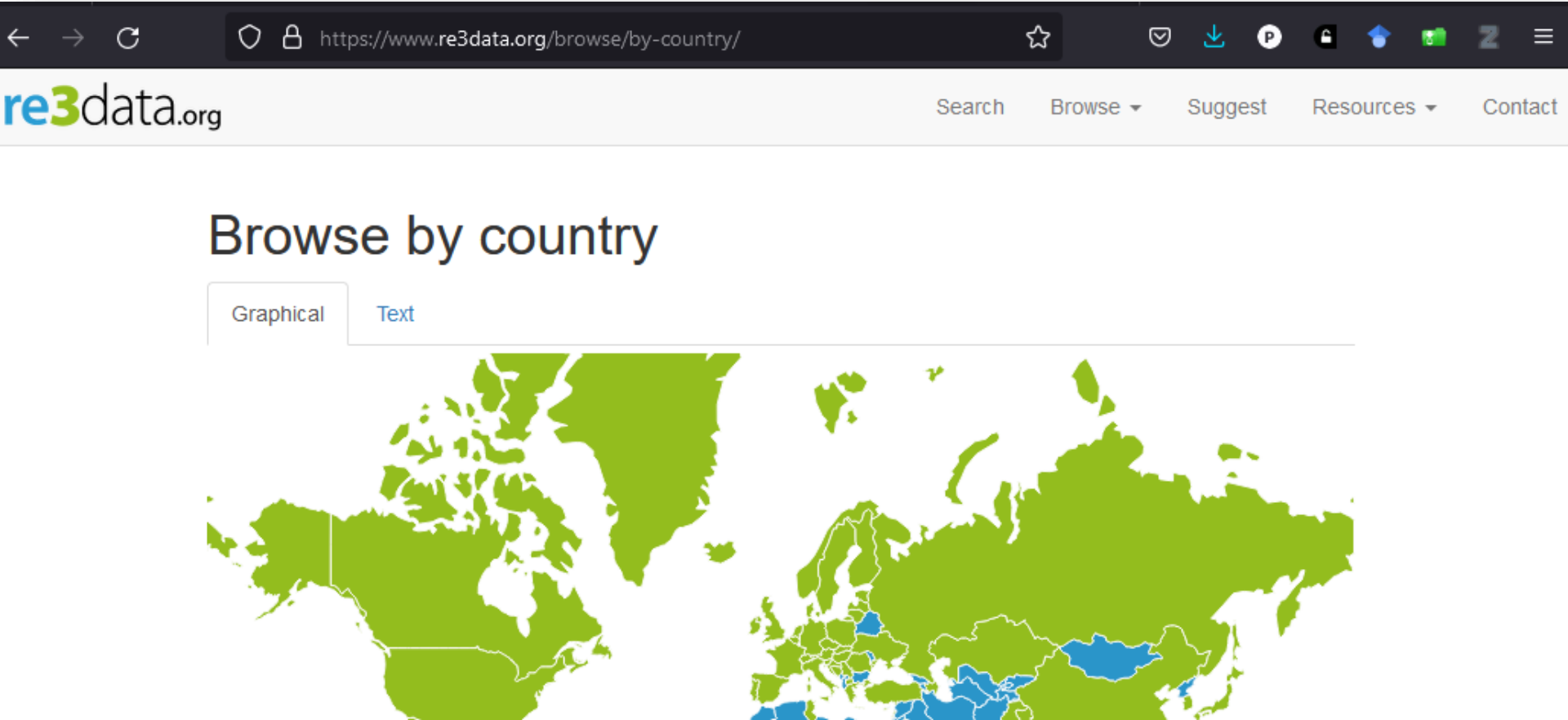
DATA.OECD.ORG

GOVERNMENT INITIATIVES

WWW.DATA.GOV (US) **DATA.GOV.UK** **DATA.GOV.BE**

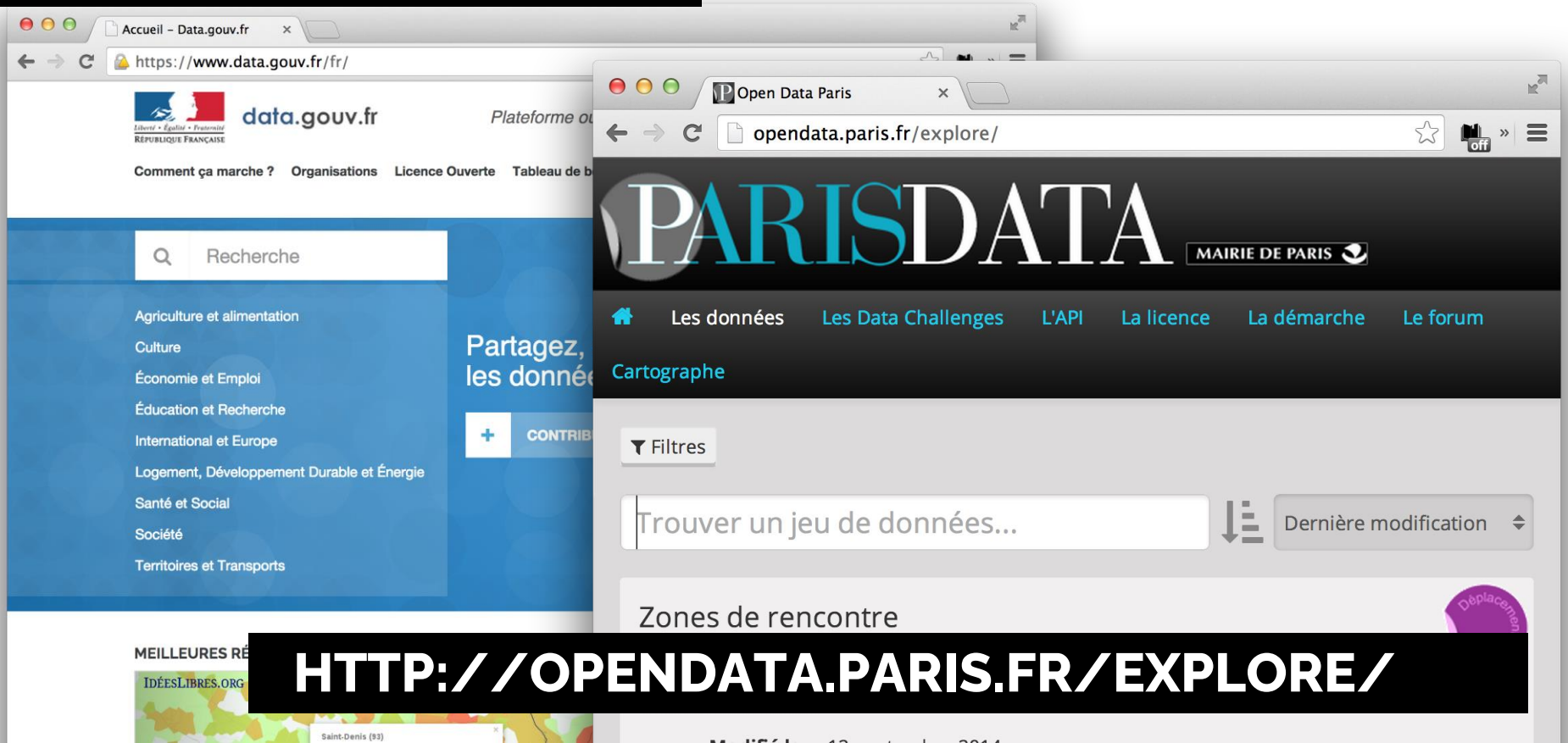


NEW DATA INITIATIVES JUST TO TRACK ALL THE DATA INITIATIVES



INITIATIVES IN FRANCE

HTTP://DATA.GOUV.FR



HTTP://OPENDATA.PARIS.FR/EXPLORE/

FINDING AND EXTRACTING EXISTING DATA

OTHER PUBLIC DATA REPOSITORIES

MORE REPOSITORIES OF PUBLIC DATA SETS

VISUALIZING.ORG
http://visualizing.org/data/browse

AMAZON PUBLIC DATA HOSTING
http://aws.amazon.com/publicdatasets/

GOOGLE PUBLIC DATA
http://www.google.com/publicdata/directory

KAGGLE
https://www.kaggle.com/

FINDING AND EXTRACTING EXISTING DATA

DATA RETAILERS

DATA RETAILERS

FACTUAL
http://www.factual.com/

Dawex
https://www.dawex.com/en/

Datamean
https://datmean.com/

Weather stations, …

AND AGAIN, THERE ARE MANY, MANY MORE…

FINDING AND EXTRACTING EXISTING DATA

APIS

TWITTER - X / TECH / ELON MUSK

Twitter announces new API pricing, posing a challenge for small developers

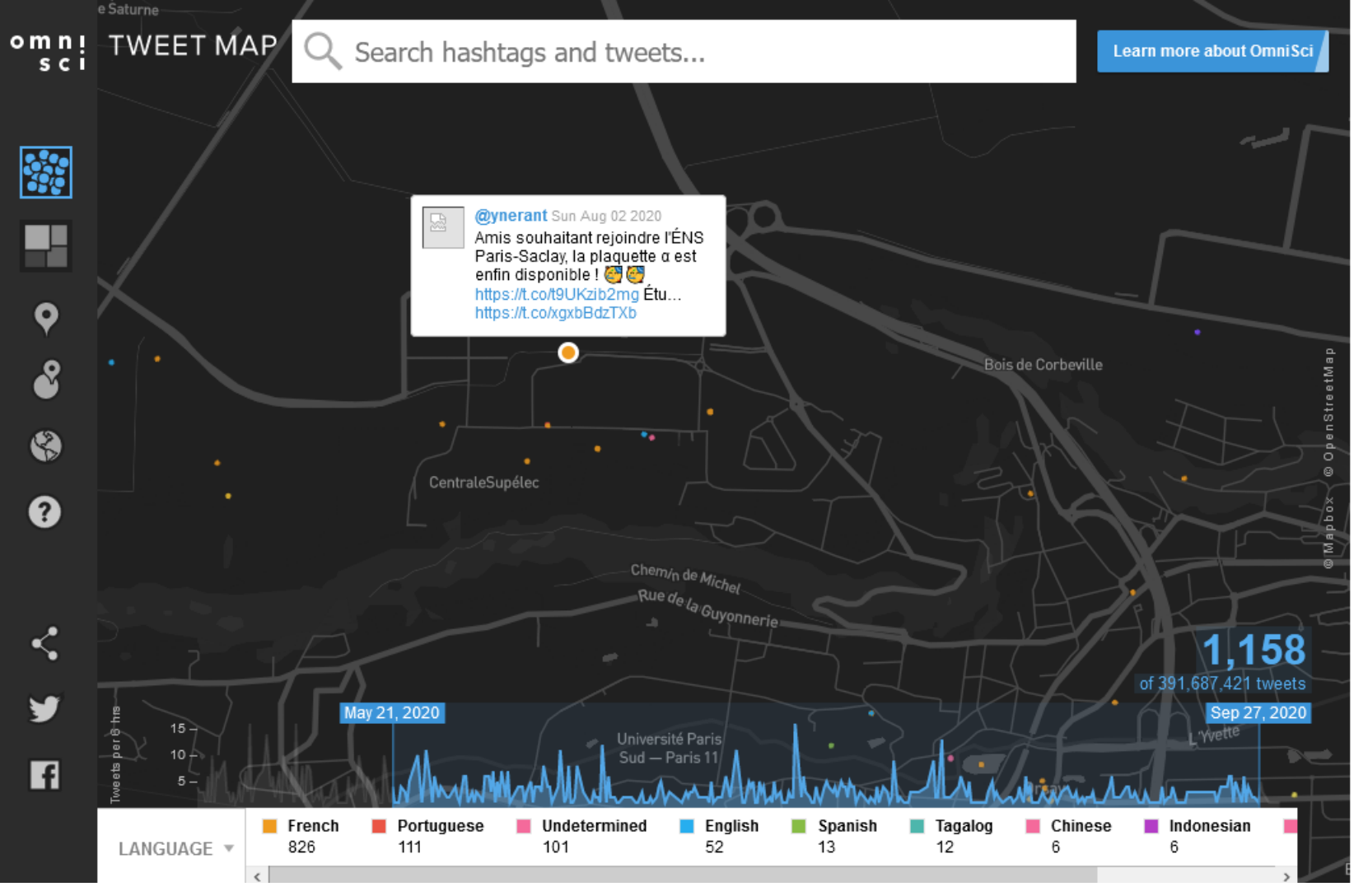
Illustration: Alex Castro / The Verge

/ After announcing it would be changing its API rules in February, Twitter has now detailed how free access to its API will work in the future.

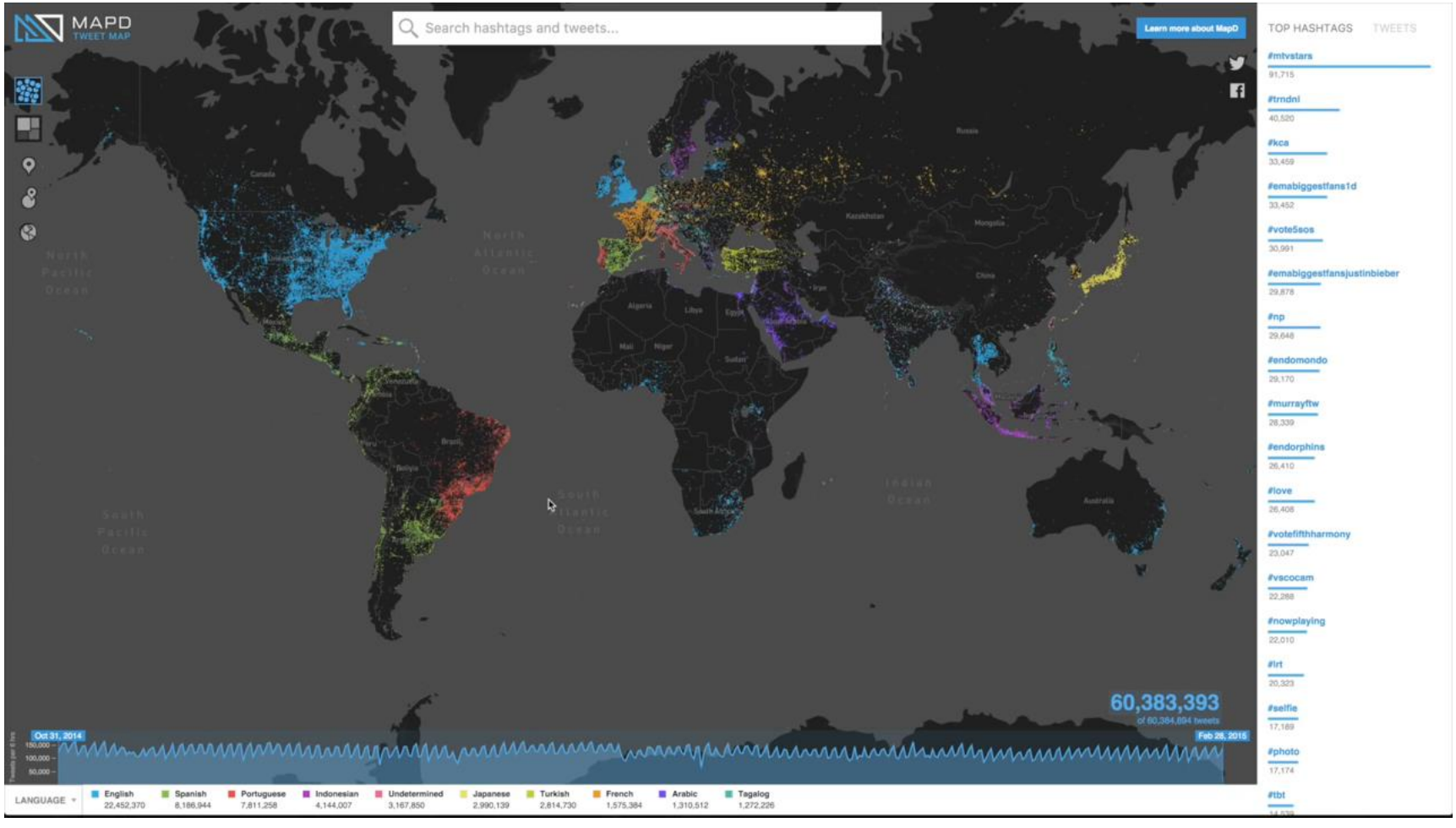
By Jon Porter, a reporter with five years of experience covering consumer tech releases, EU tech policy, online platforms, and mechanical keyboards.

Mar 30, 2023, 11:35 AM GMT+2 | 18 Comments / 18 New

- GET /2/tweets/search/stream
- GET /2/tweets/search/stream/rules
- POST /2/tweets/search/stream/rules



https://www.omnisci.com/demos/tweetmap



https://www.youtube.com/watch?v=QhE3befmFSI

Tottenham Riots

402 sources sharing 551 tweets matching "tottenhamriots" or "tottenham"

Search (enter search terms here) Search

Sort # times retweeted

Show Sources (showing 8 of 10 sources loaded)

All | Ordinary People | Journalists / Bloggers | Organizations | Uncategorized | Eyewitnesses

Show Tweets

All | **Exclude RTs** | Images & Videos

Daniel Carr, @daniel_carr (2 years, 3 months old)

Myself in 160 characters: Schizophrenic. Also a criminologist

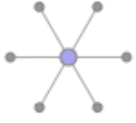
London, United Kingdom

Journo/Blogger | 41 RTed | 56 Klout

Network Sketch

213 Followers

163 Following



Friends' Locations

London, GB 34.48%

Glasgow, GB 5.17%

Manchester, GB 3.45%

Top Entities Mentioned Historically

Bruce Grove, Tottenham Hale, London, BBC, Haha,

31 Tweets

#tottenham #tottenhamriots Fire near Bruce Grove Station, larger one towards Lordship Lane
Aug. 6, 2011, 11:27 p.m.

#tottenham #tottenhamriots @MrsCheddies by Bruce Grove I mean north of previous fires, on High Rd towards Lordship Lane
Aug. 6, 2011, 11:24 p.m.

@hackneyhive yeah around that area there are 2 fires, one small now, one very large #tottenham #tottenhamriots

Aidan Rowe, @Aidan_Rowe (1 year, 2 months old)

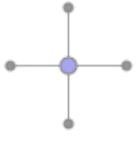
Post-punk, proto-utopian, anarchist, activist, musician, blogger, student, failed comedian.
http://redwriters1.blogspot.com

Ordinary Person | 23 RTed | 49 Klout

Network Sketch

215 Followers

395 Following



Friends' Locations

Dublin, IE 43.48%

London, GB 4.35%

Cork, IE 1.74%

Top Entities Mentioned Historically

Oslo, BBC, Dublin, Dermot Mulqueen, Johann Hari,

5 Tweets

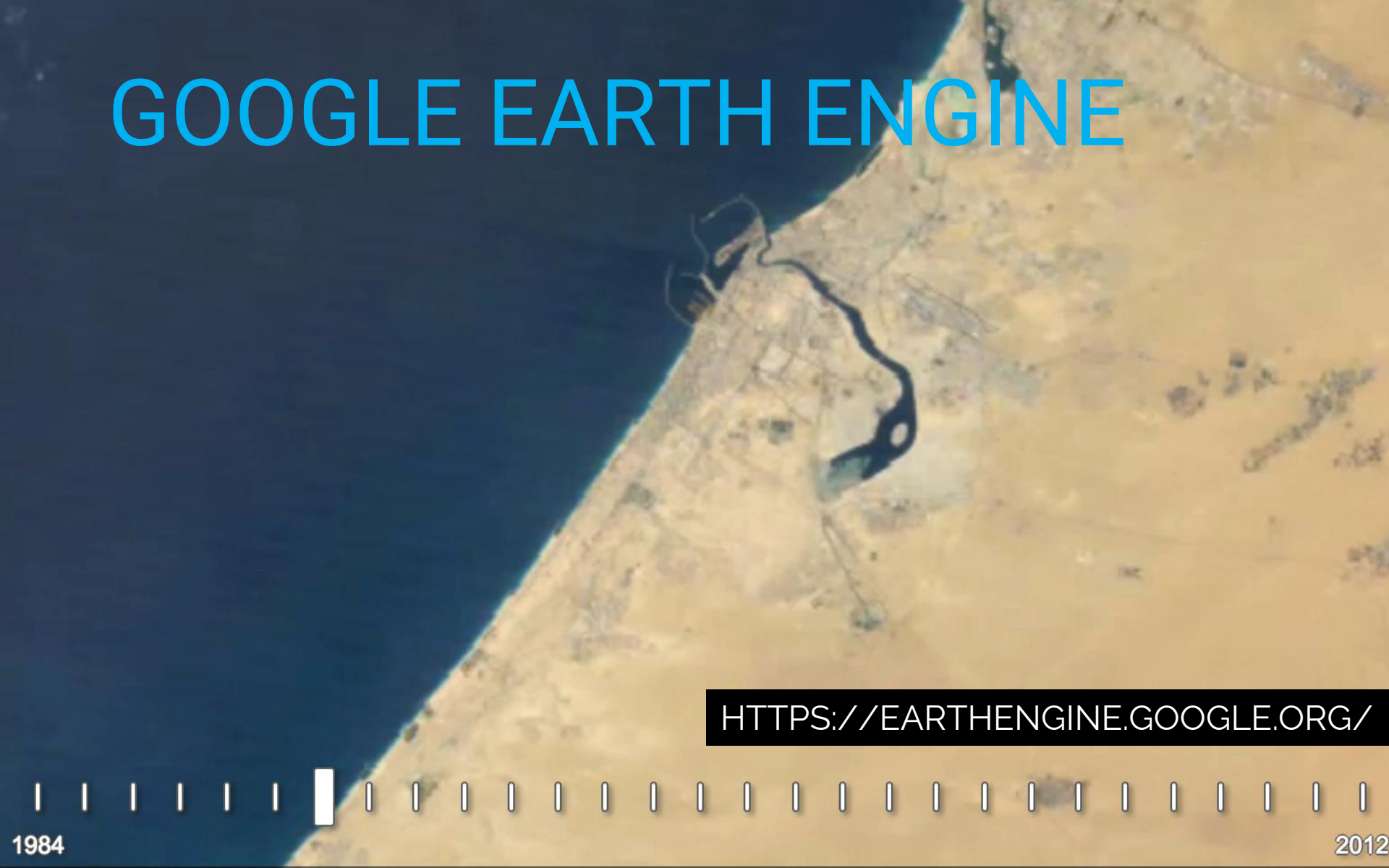
"Why couldn't the people in #Tottenham just have held a nice dignified protest for us to ignore?" - Liberals #tottenhamriots
Aug. 7, 2011, 12:49 a.m.

Any reports of arrests? #tottenham #tottenhamriots Hope everyone is safe. #acab
Aug. 7, 2011, 12:06 a.m.

Anyone using the words "mindless", "hooligans" or "thugs" is a racist and an idiot. #tottenham #tottenhamriots

SRSR [DIAKOPOULOS ET AL. 2012]

GOOGLE EARTH ENGINE



MORE APIS

(APPLICATION PROGRAMMING INTERFACES)

NEW YORK TIMES APIS

http://developer.nytimes.com/

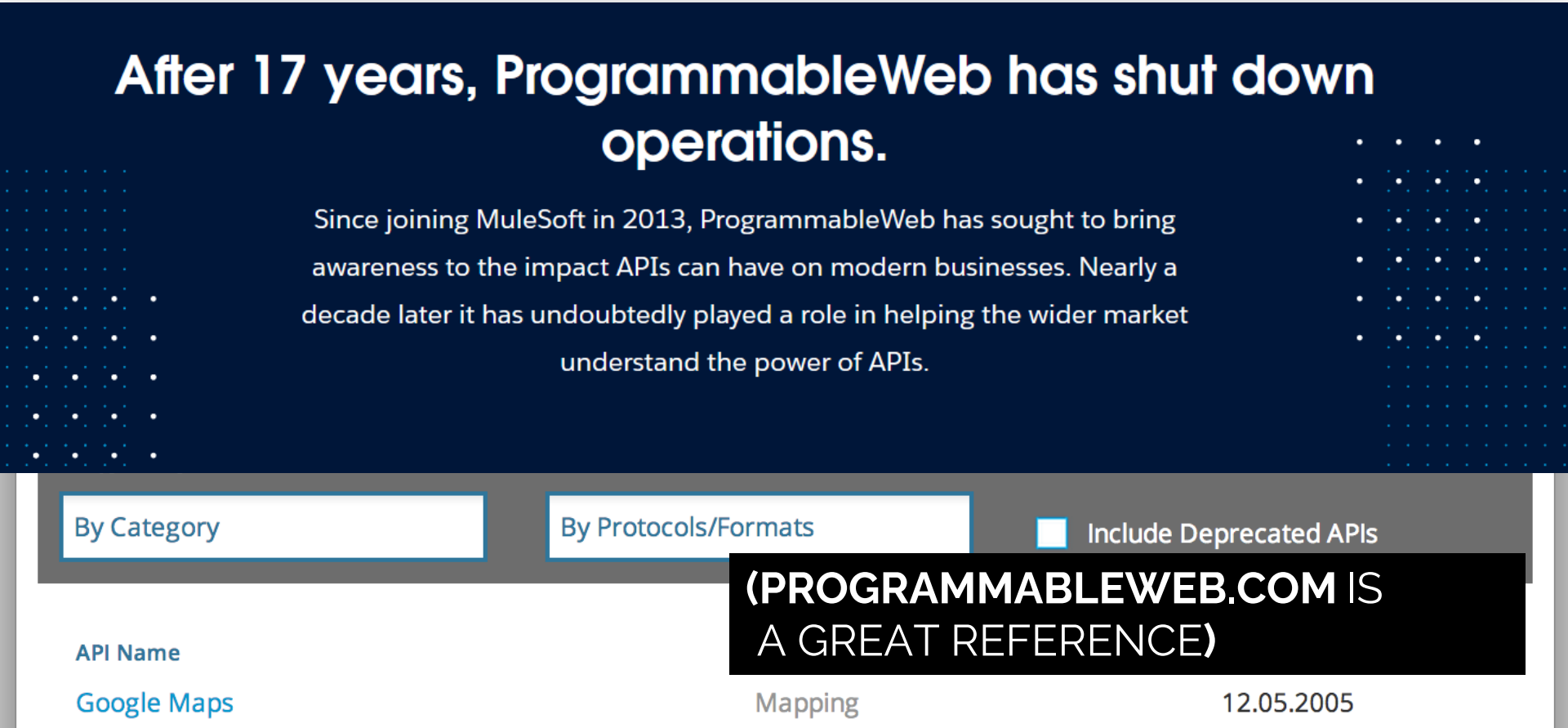
(Archival news articles from 1851, books, movies, geographical, and political data)

OPEN STREET MAP

http://wiki.openstreetmap.org/wiki/API

(Detailed location and map data for the whole world)

AND THE LIST GOES ON!



(PROGRAMMABLEWEB.COM IS A GREAT REFERENCE**)**

https://apis.guru/

Check out our other open-source projects

Star 3.4k Tweet

Help support the work that we do by contributing to our Open Collective campaign!

Support APIs.guru

Filter 2,529 APIs

Search...

1Forge Finance APIs

Stock and Forex Data and Realtime Quotes

Events API

Events API API logo

1Password Events API Specification.

1Password Connect

1Password Connect API logo

REST API interface for 1Password Connect.

Authentiq API

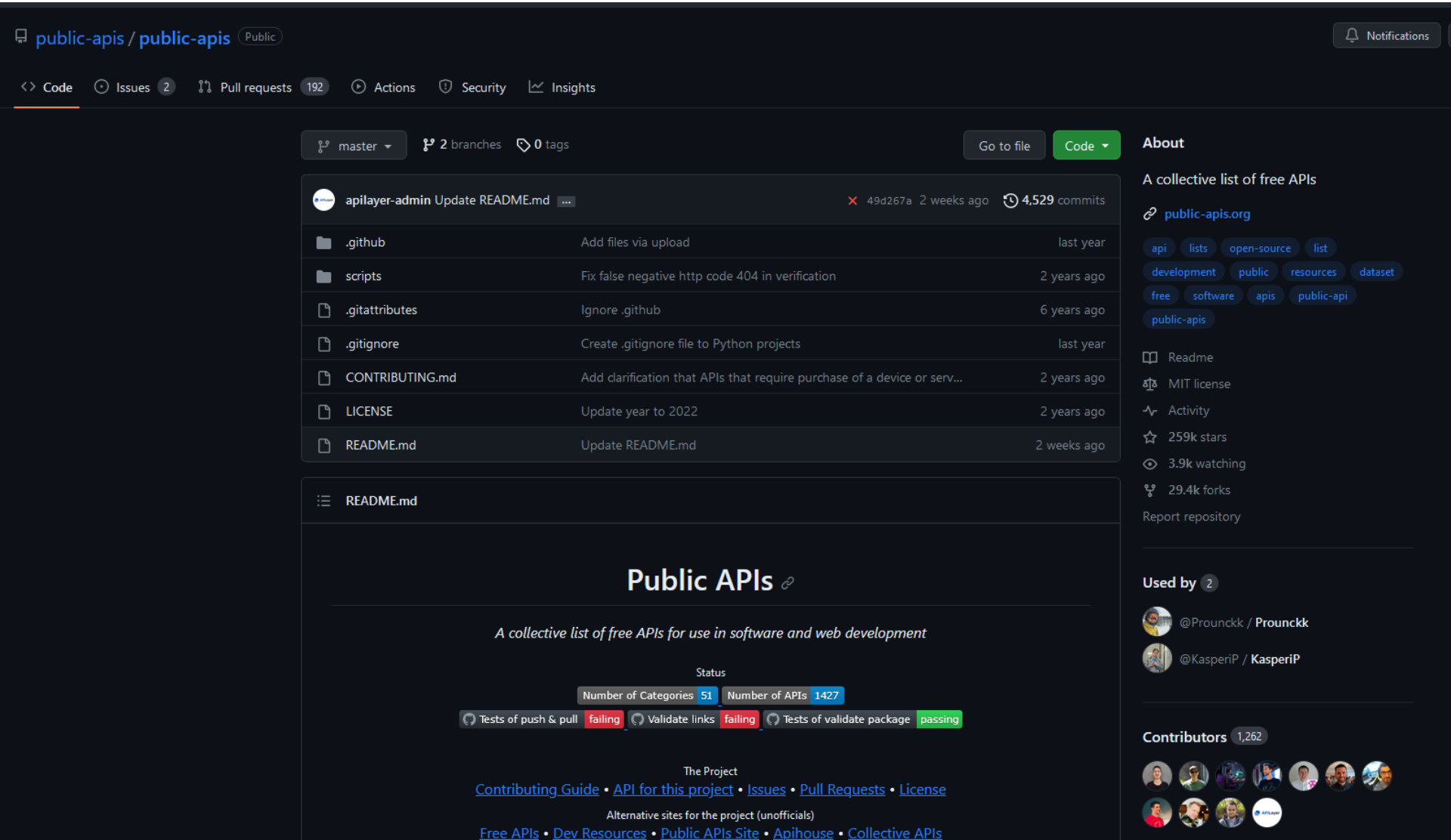
Strong authentication, without the passwords.

Platform API

Platform API API logo

The REST API specification for Ably.

https://github.com/public-apis/public-apis



public-apis / public-apis Public

Notifications

Code Issues 2 Pull requests 192 Actions Security Insights

master 2 branches 0 tags Go to file Code

apilayer-admin Update README.md ... 49d267a 2 weeks ago 4,529 commits

.github	Add files via upload	last year
scripts	Fix false negative http code 404 in verification	2 years ago
.gitattributes	Ignore .github	6 years ago
.gitignore	Create .gitignore file to Python projects	last year
CONTRIBUTING.md	Add clarification that APIs that require purchase of a device or serv...	2 years ago
LICENSE	Update year to 2022	2 years ago
README.md	Update README.md	2 weeks ago

README.md

Public APIs

A collective list of free APIs for use in software and web development

Status

Number of Categories 51 Number of APIs 1427

Tests of push & pull failing Validate links failing Tests of validate package passing

The Project

Contributing Guide • API for this project • Issues • Pull Requests • License

Alternative sites for the project (unofficials)

Free APIs • Dev Resources • Public APIs Site • Apihouse • Collective APIs

About

A collective list of free APIs

public-apis.org

api lists open-source list development public resources dataset free software apis public-api public-apis

Readme

MIT license

Activity

259k stars

3.9k watching

29.4k forks

Report repository

Used by 2

@Prounckk / Prounckk

@KasperiP / KasperiP

Contributors 1,262

FINDING AND EXTRACTING EXISTING DATA

SCRAPING THE WEB

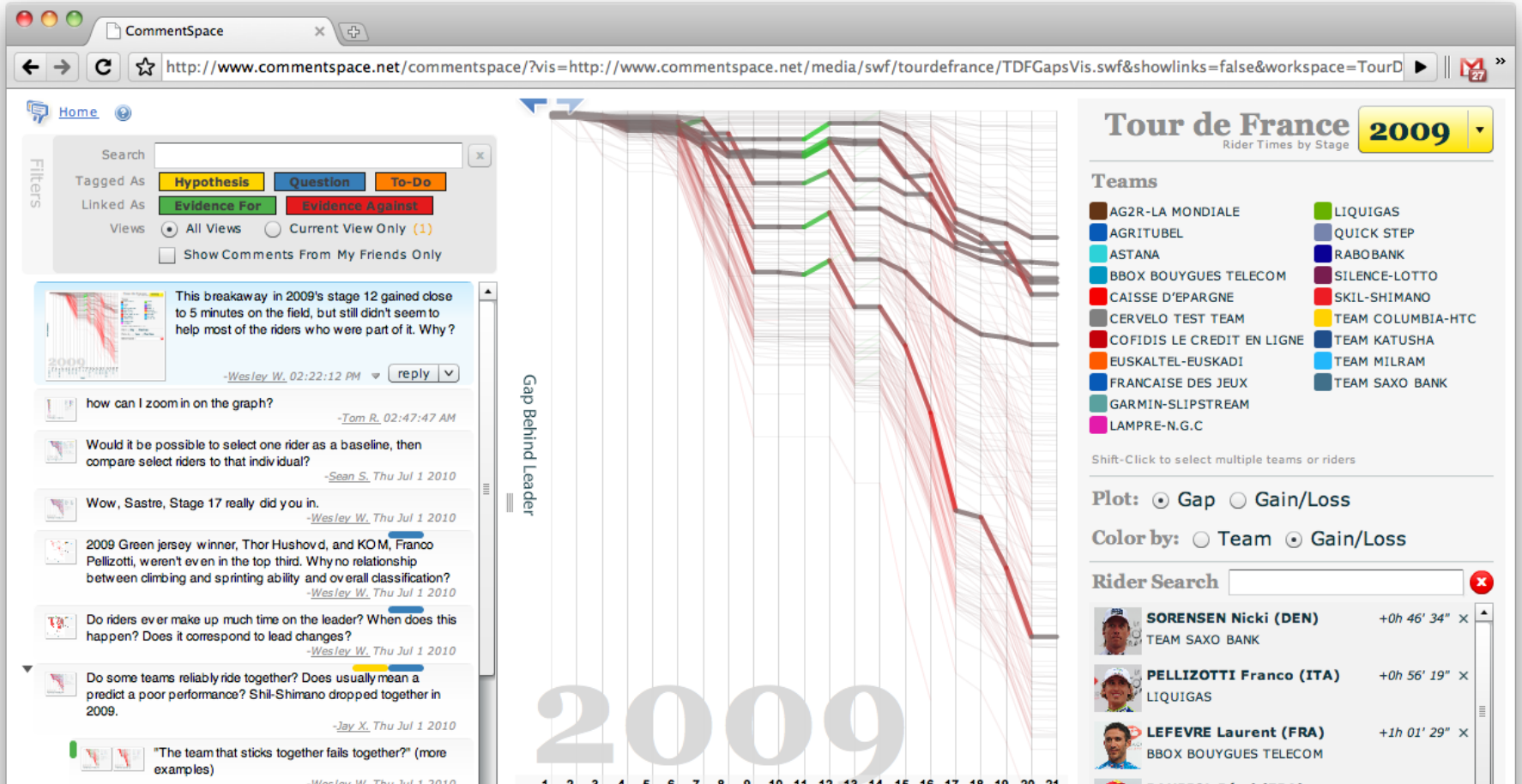
WHY SCRAPE?

No API exists for the data you want
(can't access the right data, wrong format, etc.)

Simplicity – Usually don't need to authenticate, no rate-limiting, etc.

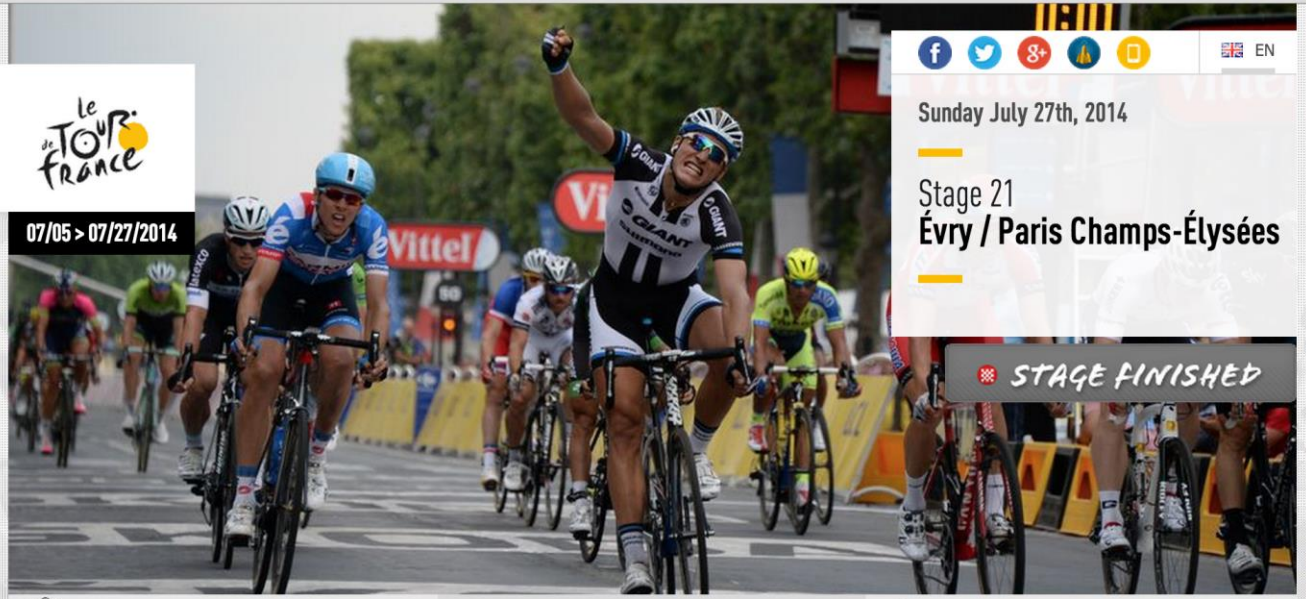
Want to capture context of pages or relationship between them.

FOR EXAMPLE...


CommentSpace
http://www.commentspace.net/commentspace/?vis=http://www.commentspace.net/media/swf/tourdefrance/TDFGapsVis.swf&showlinks=false&workspace=TourD
Home
Filters
Search
Tagged As
Hypothesis
Question
To-Do
Linked As
Evidence For
Evidence Against
Views
All Views
Current View Only (1)
Show Comments From My Friends Only
This breakaway in 2009's stage 12 gained close to 5 minutes on the field, but still didn't seem to help most of the riders who were part of it. Why?
-Wesley W. 02:22:12 PM
reply
how can I zoom in on the graph?
-Tom R. 02:47:47 AM
Would it be possible to select one rider as a baseline, then compare select riders to that individual?
-Sean S. Thu Jul 1 2010
Wow, Sastre, Stage 17 really did you in.
-Wesley W. Thu Jul 1 2010
2009 Green jersey winner, Thor Hushovd, and KOM, Franco Pellizotti, weren't even in the top third. Why no relationship between climbing and sprinting ability and overall classification?
-Wesley W. Thu Jul 1 2010
Do riders ever make up much time on the leader? When does this happen? Does it correspond to lead changes?
-Wesley W. Thu Jul 1 2010
Do some teams reliably ride together? Does usually mean a predict a poor performance? Shil-Shimano dropped together in 2009.
-Jay X. Thu Jul 1 2010
"The team that sticks together fails together?" (more examples)
Gap Behind Leader
2009
Tour de France
Rider Times by Stage
2009
Teams
AG2R-LA MONDIALE
AGRITUBEL
ASTANA
BBOX BOUYGUES TELECOM
CAISSE D'EPARGNE
CERVELO TEST TEAM
COFIDIS LE CREDIT EN LIGNE
EUSKALTEL-EUSKADI
FRANCAISE DES JEUX
GARMIN-SLIPSTREAM
LAMPRE-N.G.C
LIQUIGAS
QUICK STEP
RABOBANK
SILENCE-LOTTO
SKIL-SHIMANO
TEAM COLUMBIA-HTC
TEAM KATUSHA
TEAM MILRAM
TEAM SAXO BANK
Shift-Click to select multiple teams or riders
Plot: Gap Gain/Loss
Color by: Team Gain/Loss
Rider Search
SORENSEN Nicki (DEN)
TEAM SAXO BANK
+0h 46' 34"
PELLIZOTTI Franco (ITA)
LIQUIGAS
+0h 56' 19"
LEFEVRE Laurent (FRA)
BBOX BOUYGUES TELECOM
+1h 01' 29"

Classifications stage 21 -

www.letour.fr/le-tour/2014/us/stage-21/classifications.html



le TOUR de France

07/05 > 07/27/2014

EN

Sunday July 27th, 2014

Stage 21

Évry / Paris Champs-Élysées

STAGE FINISHED

19:16 Top 5

19:14 The winner is... Marcel Kittel

19:10 All together with 3km to go

THE RACE | ROUTE | CLASSIFICATIONS | TEAMS | VIDEOS & PHOTOS | HISTORY | STORE

Search

PARIS TOURS 12/10/2014

PREVIOUS

SUNDAY, JULY 27TH - STAGE 21 137.5km

Évry / Paris Champs-Élysées

NEXT

OVERALL STAGE

Individual | Points | Team | Climber | Youth | Combative

Overall individual time classification

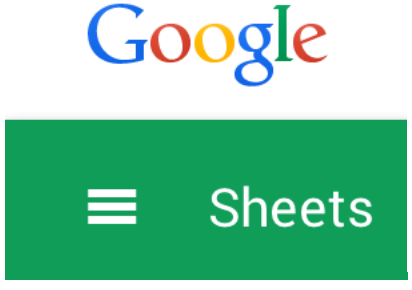
Total distance covered: **3660.5 KM**

RANK	RIDER	RIDER NO.	TEAM	TIMES	GAP
1.	NIBALI Vincenzo	41	ASTANA PRO TEAM	**89h 59' 06"**	
2.	PÉRAUD Jean-Christophe	81	AG2R LA MONDIALE	**90h 06' 43"**	+ 07' 37"
3.	PINOT Thibaut	127	FDJ.FR	**90h 07' 21"**	+ 08' 15"
4.	VALVERDE BELMONTE Alejandro	11	MOVISTAR TEAM	**90h 08' 46"**	+ 09' 40"
5.	VAN GARDEREN Tejay	141	BMC RACING TEAM	**90h 10' 30"**	+ 11' 24"
6.	BARDET Romain	82	AG2R LA MONDIALE	**90h 10' 32"**	+ 11' 26"
7.	KONIG Leopold	201	TEAM NETAPP-ENDURA	**90h 13' 38"**	+ 14' 32"
8.	ZUBELDIA AGIRRE Haimar	169	TREK FACTORY RACING	**90h 17' 03"**	+ 17' 57"
9.	TEN DAM Laurens	67	BELKIN PRO CYCLING	**90h 17' 17"**	+ 18' 11"
10.	MOLLEMA Bauke	61	BELKIN PRO CYCLING	**90h 20' 21"**	+ 21' 15"
11.	ROLLAND Pierre	151	TEAM EUROPCAR	**90h 22' 13"**	+ 23' 07"
12.	SCHLECK Frank	161	TREK FACTORY RACING	**90h 24' 54"**	+ 25' 48"
13.	VAN DEN BROECK Jurgen	131	LOTTO-BELISOL	**90h 33' 07"**	+ 34' 01"
14.	TROFIMOV Yury	29	TEAM KATUSHA	**90h 35' 47"**	+ 36' 41"
15.	KRUIJSWIJK Steven	64	BELKIN PRO CYCLING	**90h 37' 21"**	+ 38' 15"
16.	FEILLU Brice	211	BRETAGNE - SECHE ENVIRONNEMENT	**90h 43' 05"**	+ 43' 59"
17.	HORNER Christopher	114	LAMPRE - MERIDA	**90h 43' 37"**	+ 44' 31"
18.	NIEVE ITURRALDE Mikel	5	TEAM SKY	**90h 45' 37"**	+ 46' 31"
19.	GADRET John	13	MOVISTAR TEAM	**90h 46' 36"**	+ 47' 30"

SOMETIMES YOU DON’T NEED A SCRAPER!

A few tips and tricks…

PULLING DATA TABLES FROM THE WEB



IMPORTHTML

Imports data from a table or list within an HTML page.

Demographics of India

From Wikipedia, the free encyclopedia

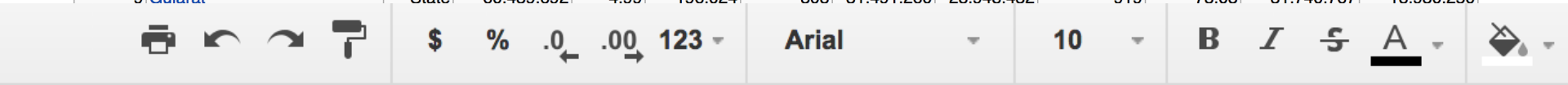
This article is about the people from India. For other uses, see Indian (disambiguation).

The **demographics of India** are inclusive of the second most populous country in the world, with over 1.21 billion people (2011 census), more than a sixth of the world's population. Already containing 17.5% of the world's population, India is projected to be the world's most populous country by 2025, surpassing China, its population reaching 1.6 billion by 2050.[4][5] Its population growth rate is 1.41%, ranking 102nd in the world in 2010.[6] Indian population reached the billion mark in 2000.

Demographics of India	
Population	1,236,344,631 (July 2014 est.)[1] (2nd)
Growth rate	1.51% (2009 est.) (93rd)
Birth rate	20.22 births/1,000 population (2013 est.)
Death rate	7.4 deaths/1,000 population (2013 est.)
Life expectancy	68.89 years (2009 est.)
• **male**	67.46 years (2009 est.)
• **female**	72.61 years (2009 est.)
Fertility rate	2.44 children born/woman (SRS 2011)
Infant mortality rate	44 deaths/1,000 live births (2011 est.)
Age structure	

Population distribution in India by states

Rank	State / Union Territory	Type	Population	% [18]	Area [19] (km²)	Density (/km²)	Males	Females	Sex Ratio [20]	Literacy	Rural [21] Population	Urban [21] Population
1	Uttar Pradesh	State	199,812,341	16.50	240,928	828	104,480,510	95,331,831	912	67.68	131,658,339	34,539,582
2	Maharashtra	State	121,455,333	9.28	307,713	365	58,243,056	54,131,277	929	82.34	55,777,647	41,100,980
3	Bihar	State	103,804,637	8.60	94,163	1,102	54,278,157	49,821,295	918	61.80	74,316,709	8,681,800
4	West Bengal	State	91,276,115	7.54	88,752	1,030	46,809,027	44,467,088	950	76.26	57,748,946	22,427,251
5	Madhya Pradesh	State	72,626,809	6.00	308,245	236	37,612,306	35,014,503	931	69.32	44,380,878	15,967,145
6	Tamil Nadu	State	72,147,030	5.96	130,058	555	36,137,975	36,009,055	996	80.09	34,921,681	27,483,998
7	Rajasthan	State	68,548,437	5.66	342,239	201	35,550,997	32,997,440	928	66.11	43,292,813	13,214,375
8	Karnataka	State	61,095,297	5.05	191,791	319	30,966,657	30,128,640	973	75.36	34,889,033	17,961,529
9	Gujarat	State	60,439,692	4.99	196,024	308	31,491,260	28,948,432	919	78.03	31,740,767	18,930,250



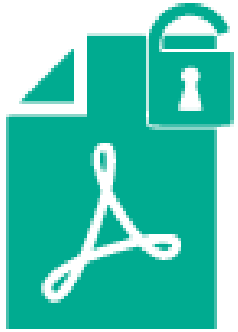
=ImportHtml("http://en.wikipedia.org/wiki/Demographics_of_India", "table",4)

	A	B	C	D	E	F
1	Rank	State / Union Territory	Type	Population	% [18]	Area [19] (km²)
2	1	Uttar Pradesh	State	199,812,341	16.5	240,928
3	2	Maharashtra	State	121,455,333	9.28	307,713
4	3	Bihar	State	103,804,637	8.6	94,163
5	4	West Bengal	State	91,276,115	7.54	88,752
6	5	Madhya Pradesh	State	72,626,809	6	308,245
7	6	Tamil Nadu	State	72,147,030	5.96	130,058

But latest version from 2018

PARSING PDFS

Tabula



Tabula is a tool
locked inside P

Extracted tabular data

2		
All Students	79,858	99%
Gender		
Male	40,492	98%
Female	39,134	99%
Ethnicity		
White	10,665	99%
Black	49,379	99%
Latino/Hispanic	13,717	98%
Asian	4,746	100%
Native American	132	99%
Multiracial	941	98%
Other Groups		
IEP	11,471	98%

Use row/columns separators

Close | Copy to clipboard as CSV | Download data

Page 3

BUILDING A WEB SCRAPER

FETCHING DATA + PARSING DATA

YOU SHOULD **SEPARATE** THESE PROCESSES **WHENEVER POSSIBLE**!

FETCHING DATA

DON'T DO EVERYTHING AT ONCE

Download complete pages and save them locally **before** you process them.

DEALING WITH PAGINATION

If results or records are spread across multiple pages, you may need to parse the page to find the link to the next page.

PARSING DATA

SERIOUSLY, DON'T DO EVERYTHING AT ONCE!

Processing data from local files means
you **don't have to get it right the first time.**

USE YOUR BROWSER'S DEVELOPER TOOLS

All modern web browsers have built-in tools that let you inspect web pages.

BE CAREFUL - YOU CAN GET YOURSELF BLOCKED

Many sites will try to slow or block heavy access (both to prevent scraping and DoS attacks)

To get around this...You can introduce delays in your scraper or scrape from multiple locations.

A FEW MORE NOTES ABOUT DATA MANAGEMENT

FORMATS AND BEST-PRACTICES

DATA FORMATS

STRUCTURED vs. UNSTRUCTURED

STRUCTURED DATA is more like what you'd find in a traditional **spreadsheet** or **database**.

UNSTRUCTURED DATA can include **raw text**, **streaming data**, even **images** or **video**.

SEMI-STRUCTURED DATA is more organized, but doesn't follow a fixed schema (e.g. DBPEDIA data)

CSV

(Comma-Separated Value)

```
1  firstName,lastName,age,streetAddress,city,state
2  John,Smith,25,21 2nd Street,New York,NY,10021,2
```

firstName	lastName	age	streetAddress	city	state	postalCode	homePhoneNumber	faxPhoneNumber	gender
John	Smith	25	21 2nd Street	New York	NY	10021	212 555-1239	646 555-4567	male

We will mostly use CSV in this course

CSV BEST PRACTICES

Remove unnecessary rows or cells

- empty cells, comments, write NA for missing values

Row	author keyword	author frequency	expert keyword	expert frequency
1	data partitioning	1	data and data management	64
2	visual knowledge discove	13	knowledge discovery	29
3	feature selection	1	features and attributes	38
4	guided visualization	1	interaction	152
5	regression	2	machine learning & statistics	55
6	model building	1	machine learning & statistics	55
7	decision support systems	1	analysis process	113
8	model validation and ana	1	machine learning & statistics	55
9	program analysis	1	analysis process	113
10	multi-variate statistics	1	multidimensional / multivariat	83
11	visual analytics	86	visual analytics	86
12	cultural heritage	2	applications	103
13	wall paintings	1	art and aesthetics	10
14	degradation	1	applications	103
15	nonnegative matrix factc	1	matrices	10
16	interactive clustering	3	clustering	50

CSV BEST PRACTICES

Splits cells if you can

If needed create a second file

First page	Last page	ous (capstone, keynote, VAST challenge, panel, poster, ...)	Abstract	Author Names
457	457	M		Donna J. Cox
6	13, 460	C	The use of critical po	James Helman;Lambertus Hesselink
14	27, 461	C	The authors discuss	Gordon V. Bancroft;Fergus Merritt;Todd Plessel;Paul G. Kelaita;R. Kevin McCabe;Al Globus
28	35, 462	C	The VIS-5D system	William L. Hibbard;David A. Santek
36	44, 462	C	The author presents	James L. Montine
45	50, 462	C	Some ideas and tech	Gregory M. Nielson;Bernd Hamann
51	58, 463	C	The use of qualitative	Yaser Yacoob
59	66	C	Visualizing the third	Del Lamb;Amit Bandopadhay
67	73	C	The animation of two	Anthony J. Maeder
74	82, 464	C	The authors propose	James V. Miller;David E. Breen;Michael J. Wozny
83	92, 465	C	The authors present	Ping-Kang Hsiung;Robert H. Thibadeau;Christopher B. Cox;Robert H. P. Dunn;Michael Wu;Pau
93	96, 467	C	The authors describe	Richard A. Becker;Stephen G. Eick;Eileen O. Miller;Allan R. Wilks
97	106, 46	C	The authors describe	Andrew J. Hanson;Pheng-Ann Heng;B. C. Kaplan
107	113	C	The authors describe	Bowen Alpern;Larry Carter;Ted Selker

CSV BEST PRACTICES

Give meaningful unique column names

	A	B	C	D
1	**ExistingFieldName**	**UserFriendlyFieldName**		
2	AccMngDpt	Department		
3	AccMngName	Account Manager		
4	CusAccMngID	Account Manager ID		
5	CusAddress	Customer Address		
6	CusCoulD	Customer Country		
7	CusID	Customer ID		
8	CusName	Customer		
9	DelAddress	Delivery Address		
10	DelDate	Delivery Date		
11	DelDesc	Delivery Description		
12	DelID	Delivery ID		
13	DelTime	Delivery Time		
14				
15				
16				

http://www.qlikviewaddict.com/2013/05/user-friendly-field-names-in-qlikview.html

Make column name casing consistent

For pandas, snake_case is recommended
→ with it you can do:

df.column_name or
df['column_name']

XML

(eXtensible Markup Language)

```
<person>
  <firstName>John</firstName>
  <lastName>Smith</lastName>
  <age>25</age>
  <address>
    <streetAddress>21 2nd Street</streetAddress>
    <city>New York</city>
    <state>NY</state>
    <postalCode>10021</postalCode>
  </address>
  <phoneNumbers>
    <phoneNumber type="home">212 555-1234</phoneNumber>
    <phoneNumber type="fax">646 555-4567</phoneNumber>
  </phoneNumbers>
  <gender>
```

firstName	lastName	age	streetAddress	city	state	postalCode	homePhoneNumber	faxPhoneNumber	gender
John	Smith	25	21 2nd Street	New York	NY	10021	212 555-1239	646 555-4567	male

JSON

(JavaScript Object Notation)

```
{
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021"
  },
  "phoneNumber": [
    {
      "type": "home",
      "number": "212 555-1239"
    },
    {
      "type": "fax",
      "number": "646 555-4567"
    }
```

firstName	lastName	age	streetAddress	city	state	postalCode	homePhoneNumber	faxPhoneNumber	gender
John	Smith	25	21 2nd Street	New York	NY	10021	212 555-1239	646 555-4567	male

YAML

(YAML Ain't Markup Language)

```yaml
---
  firstName: John
  lastName: Smith
  age: 25
  address:
        streetAddress: 21 2nd Street
        city: New York
        state: NY
        postalCode: 10021

  phoneNumber:
        -
            type: home
            number: 212 555-1234
        -
            type: fax
```

firstName	lastName	age	streetAddress	city	state	postalCode	homePhoneNumber	faxPhoneNumber	gender
John	Smith	25	21 2nd Street	New York	NY	10021	212 555-1239	646 555-4567	male

HANDLING DATA

STORING DATA

- Always keep **backups**
- **Password protect** or **encrypt** any data with personal or sensitive information

PROVENANCE

- Keep track of **where/when** data was collected
- Record any data processing steps so you (or others) can repeat them if necessary

IP, COPYRIGHT, **AND** (RE)SHARING DATA

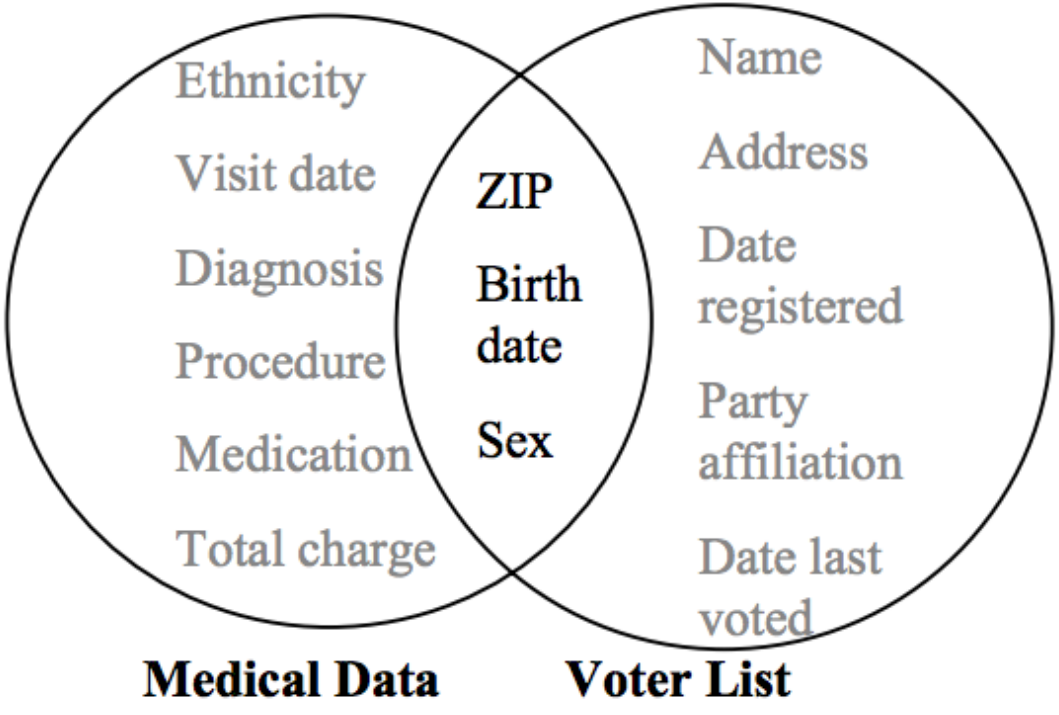
- Be sure you know who **owns** the data.
- Think early on about whether or not you'll need to **publish** or **(re)share** data.
- Be careful you **aren't violating copyright**, especially when scraping.

PRIVACY AND ANONYMIZING DATA

- Any information that could be used to identify individuals is sensitive!
- There may be legal repercussions for releasing it.
- In some cases you might need to anonymize data before sharing.

JUST REMOVING NAMES IS OFTEN NOT ENOUGH!

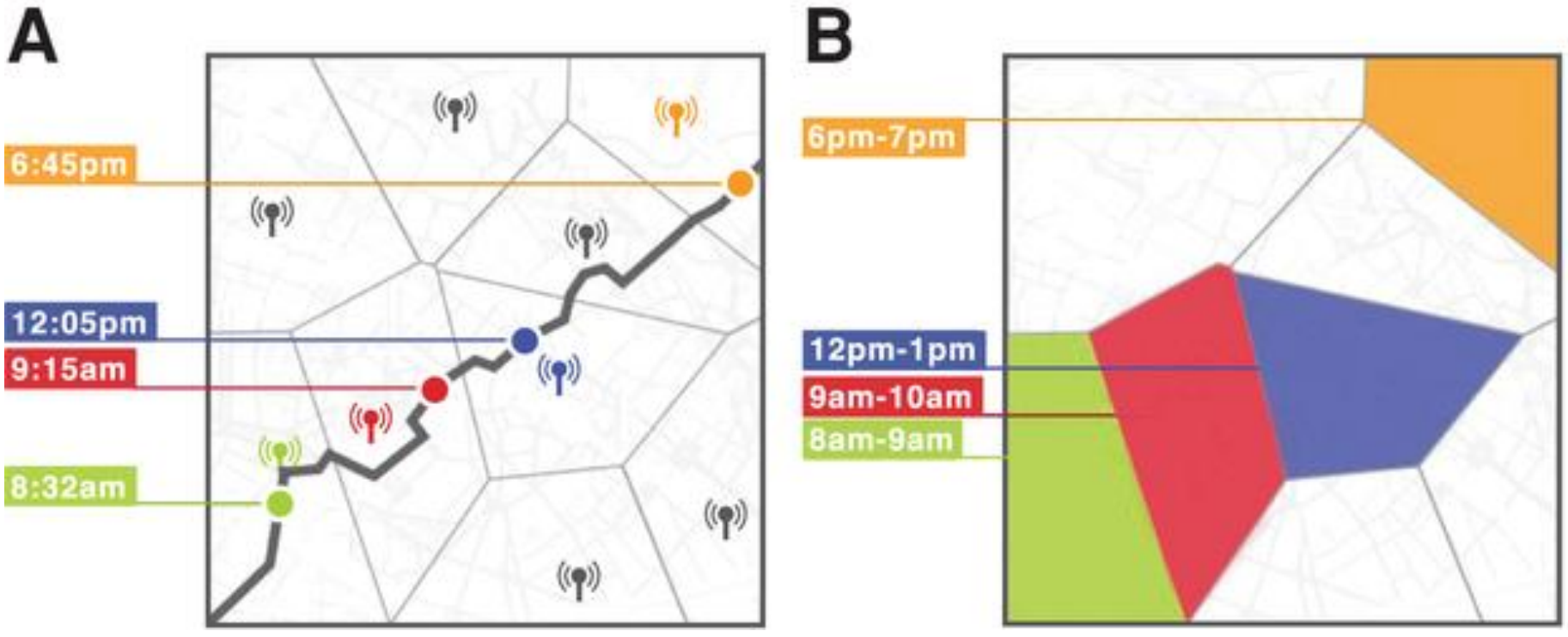
OTHER INFORMATION CAN STILL BE UNIQUE



[L. Sweeney. 2002]

k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY

LOCATION DATA IS ESPECIALLY SENSITIVE



[de Montjoye et al. 2013]
Unique in the Crowd: The privacy bounds of human mobility

REGULATIONS (ACADEMIA AND RESEARCH)

Institutional Review and Ethics Boards may need to approve experiments or data collection **before it happens.**

Studies involving people may need **informed consent.**

Consent form

PART I/ INFORMATION

Acknowledgements:

Your input is essential to our project and the research team would like to thank you for that.
Before giving your informed consent, it is important that you understand why the research is being done, what it involves and what your rights and obligations are.

"Validated Scale for Perceived Readability of Visualizations"

IDENTIFICATION

Project manager(s): Tobias Isenberg (Senior Research Scientist, Aviz Team, Inria Saclay)
Other scientists involved:

- Anne-Flore Cabouat (Master's student, Aviz Team, Inria Saclay, Université Paris-Saclay),
- Tingying He (PhD student, Aviz Team, Inria Saclay, Université Paris-Saclay),
- Petra Isenberg (Senior Research Scientist, Aviz Team, Inria Saclay).

Project team and Inria Research Centre: Aviz Team - Inria Saclay
Project name: Scale for Perceived Readability in Visualization

You are invited to take part in this project, whose the purpose is to develop a validated scale for measuring perception of readability in visualization. The Aviz team conducts research in the field of visualization and the results of these experiments will help us to improve visualization design and research in this field more generally.

This study will take place online.

The expected benefits of the project are the following: readability of a visualization has a considerable impact on a visualization's use. Researchers investigating usability, effectiveness, and acceptability of visualizations can seek to evaluate participants' perception of readability. To do so, they presently pick their own survey terms, for example "I am utterly confused", "symbols were easy to read", "I am confident in my answers", "the visualization facilitates data understanding", and so on. We currently lack a validated tool to evaluate the perceived readability of visual representations among study participants. The aim of this project is to build and validate a consistent scale to measure perceived readability.

This project has received a favorable opinion from Inria's Operational Committee for the Assessment of legal and ethical risks. For more information:
https://www.inria.fr/en/operational-committee-assesment-legal-and-ethical-risks

YOUR PARTICIPATION IN THE PROJECT

Conditions of your participation:

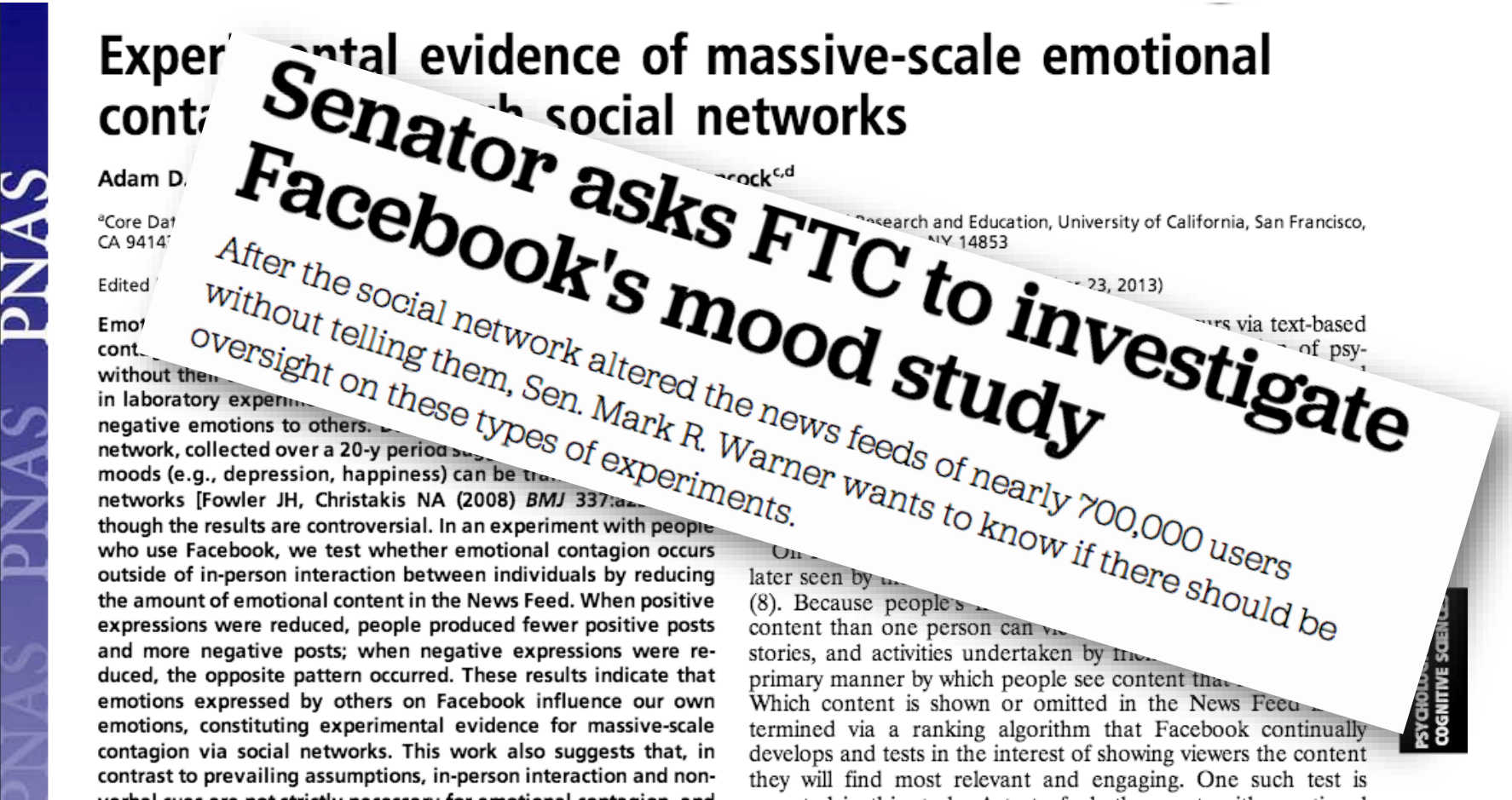
Requirements
Anyone who is a fluent speaker of English, is of legal age (18 years in most countries), and doesn't have low or impaired vision, including any form of color deficiency, can participate in this study.

REGULATIONS (INDUSTRY)

Some governments have placed limits on **how long** user data can be kept.

Some kinds of tracking (e.g., cookies) may now require **opt-in** or **notifications**.
(However this varies by country).

SOCIAL EXPERIMENTS



Experimental evidence of massive-scale emotional contagion through social networks

Adam D.

[a]Core Dat… CA 9414…

Edited …

Emot… cont… without ther… in laboratory experi… negative emotions to others. … network, collected over a 20-y period … moods (e.g., depression, happiness) can be tra… networks [Fowler JH, Christakis NA (2008) *BMJ* 337:a… though the results are controversial. In an experiment with people who use Facebook, we test whether emotional contagion occurs outside of in-person interaction between individuals by reducing the amount of emotional content in the News Feed. When positive expressions were reduced, people produced fewer positive posts and more negative posts; when negative expressions were reduced, the opposite pattern occurred. These results indicate that emotions expressed by others on Facebook influence our own emotions, constituting experimental evidence for massive-scale contagion via social networks. This work also suggests that, in contrast to prevailing assumptions, in-person interaction and non-

…ock[c,d]

…Research and Education, University of California, San Francisco, …NY 14853

…23, 2013)

…urs via text-based …of psy-

Senator asks FTC to investigate Facebook's mood study

After the social network altered the news feeds of nearly 700,000 users without telling them, Sen. Mark R. Warner wants to know if there should be oversight on these types of experiments.

later seen by … (8). Because people's … content than one person can … stories, and activities undertaken by … primary manner by which people see content that … Which content is shown or omitted in the News Feed … termined via a ranking algorithm that Facebook continually develops and tests in the interest of showing viewers the content they will find most relevant and engaging. One such test is

https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/

TECHNOLOGY

Everything We Know About Facebook's Secret Mood-Manipulation Experiment

It was probably legal. But was it ethical?

By Robinson Meyer

“EXPERIMENTING ON HUMAN BEINGS”

Dating Research from OkCupid

We Experiment On Human Beings!

July 28th, 2014 by Christian Rudder

I’m the first to admit it: we might be popular, we might create a lot of

GDPR

General Data Protection Regulation

→The world's strongest data protection rules

→ Define how organization can handle information about people (customers etc.)

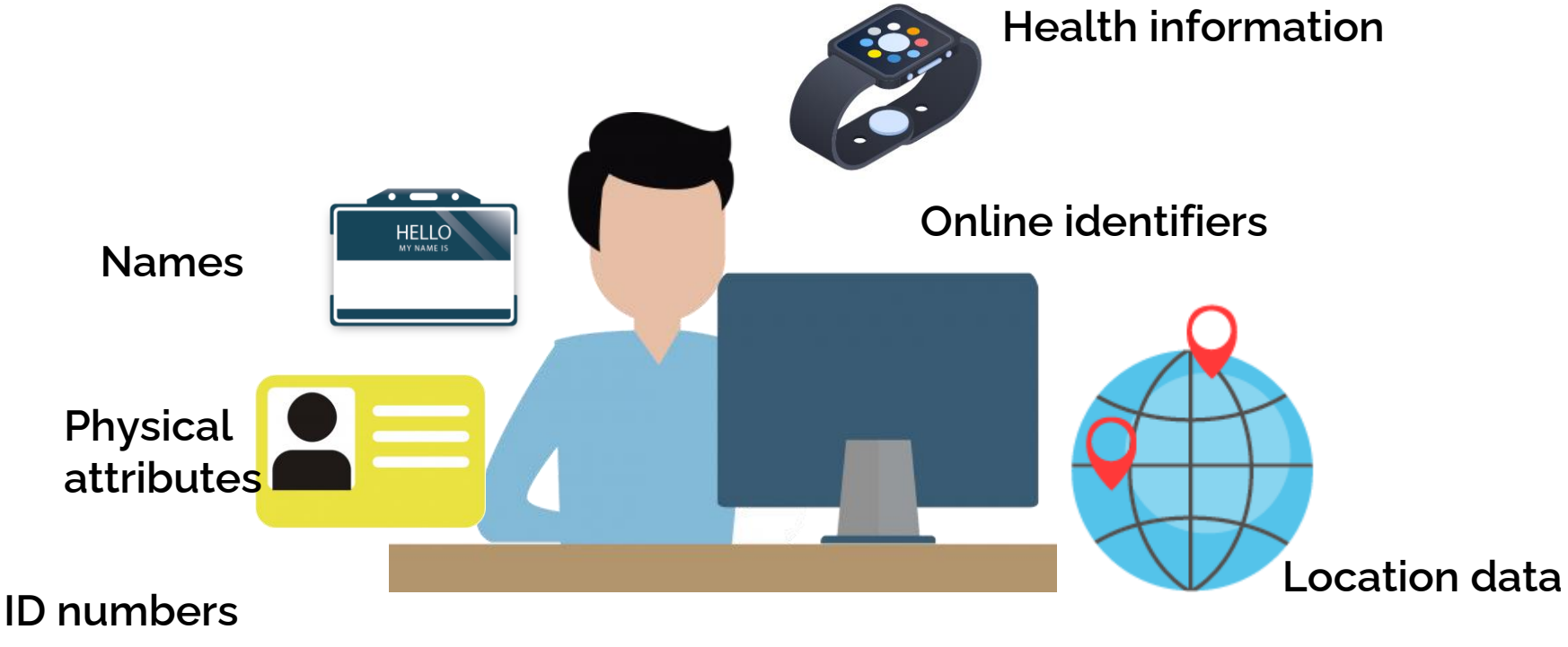
GDPR & RESEARCH

Collection of personal data

= data from which people can be identified

(data that is pseudonymous is still personal data)

PERSONAL DATA



=Any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person; (Art 4, 1)

Image credit: macrovector, freepix,

DATA PROCESSING

lawful, **fair** and **transparent**

get ethics approval

only process the minimal amount of necessary personal data; anonymize where possible

IN SUMMARY: **THERE ARE LOTS OF TOOLS AT YOUR DISPOSAL!**

COLLECT IT

- OBSERVATION
- SURVEYS
- LOGGING
- SENSORS
- CROWDSOURCING

FIND OR EXTRACT IT

- OPEN CORPUSES
- DATA RETAILERS
- APIS
- SCRAPING THE WEB

GENERATE IT

- SIMULATIONS

BEFORE NEXT CLASS

INSTALL :

OpenRefine (formerly Google Refine)
http://openrefine.org/