

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?

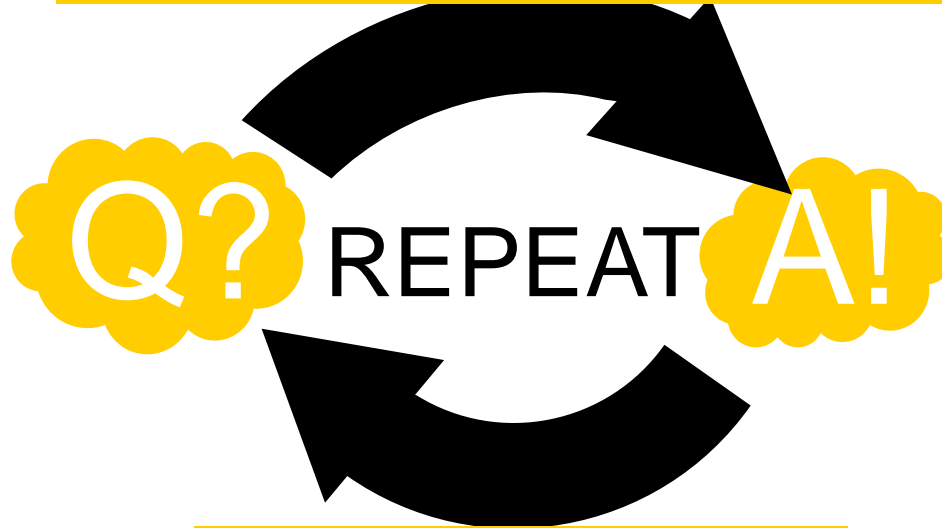
(HINT: Usually, because you have a question you need to answer!)

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)

CHOOSING A QUESTION

“How has language evolved over time?”

“What will the weather be like next month?”

“Are the right people seeing my advertisements?”

“What is the current temperature?”

A PROBLEM OF SCALE

**CHALLENGING
TO FIND DATA**

“How has language evolved over time?”

“What will the weather be like next month?”

“Are the right people seeing my advertisements?”



**NOT AS
INTERESTING**

“What is the current temperature?”

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

Qualtrics: Online Survey Sc...
www.qualtrics.com

qualtrics

Ask Questions
Get Answers

Collect → Analyze

Create your free account

University Student Satisfac...
https://www.surveymonkey.com/blog/en/univers...

SurveyMonkey®

1. How well do the professors teach at this university?

- Extremely well
- Quite well
- Moderately well
- Slightly well
- Not at all well

2. How effective is the teaching outside your major program?

- Extremely effective
- Very effective
- Moderately effective
- Slightly effective

This survey asks students to assess educational, social, and university. How effective is the teaching? Is their academic safe on campus? When you get to know what your students programs and your overall enrollment and retain students. Or questions if you want to know how students experienced sp

Use Cases | Use Cases | Google Con | How it work | Home | Go...
www.google.com/insights/consumersurveys/home

Google consumer surveys

Google consumer surveys

Report Custom Insights

What would most influence your decision to buy clothes online instead of in-store?
Results for respondents with demographics. Weighted by Age, Gender, Region. (1,138 responses) @ Winner statistically significant. @

Response	Percentage	Confidence Interval
Free shipping	40.5%	(39.7 - 41.3)
Online discounts	24.9%	(23.7 - 26.1)
Ability to return in store	16.4%	(15.4 - 17.4)
Free returns	16.1%	(15.1 - 17.1)

To find out what people really think, just ask the Internet.

When you want answers to your business questions, you need to reach everyday people — not just those who choose to participate in research panels.

The Ridiculously Power

To create a survey using the University Student Satisfaction sign in to SurveyMonkey. You'll be able to choose the template



CROWDSOURCING DATA COLLECTION

The screenshot shows a web browser window with the Amazon Mechanical Turk interface. The browser's address bar displays the URL <https://www.mturk.com/mturk/welcome>. The page title is "Amazon Mechanical Turk".

The main content area is titled "HITs containing 'short survey'" and shows "11-20 of 49 Results". Below the title, there is a "Sort by:" dropdown menu set to "HITs Available (most first)" and a "GO!" button. To the right of the dropdown are links for "Show all details" and "Hide all details". Further right are navigation links: "First << Previous < 1 2 3 4 5 > Next >> Last".

The list of HITs is displayed in a table-like format with alternating light blue and light green rows. Each row contains the following information:

- HIT Title:** "Answer a short survey about Work Team Dynamics" (for the first two rows) and "Short Survey" (for the third row).
- Requester:** "Whitney Ohmer" (for the first two rows) and "David Tannenbaum" (for the third row).
- HIT Expiration Date:** "Oct 12, 2014 (2 weeks 5 days)" (for the first two rows) and "Oct 12, 2014 (2 weeks 5 days)" (for the third row).
- Reward:** "\$0.25" (for the first two rows) and "\$0.10" (for the third row).
- Time Allotted:** "60 minutes" (for the first two rows) and "60 seconds" (for the third row).
- HITs Available:** "1" (for all three rows).

Each row also includes a "Request Qualification (Why?)" link and a "View a HIT in this group" link.

The bottom of the screenshot shows the start of a fourth row with the title "Short survey about website experience (on average it takes 13 minutes)", a requester name, an expiration date of "Oct 6, 2014 (1 week 6 days)", a reward of "\$1.50", and a time allotted of "15 minutes".

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

Wikipedia:Statistics: Revision history

en.wikipedia.org/w/index.php?title=Wikipedia:Statistics&action=history

Create account Login

Project page [Talk](#) [Read](#) [Edit](#) [View history](#)

Wikipedia:Statistics: Revision history

[View logs for this page](#)

Browse history

From year (and earlier): From month (and earlier): [Tag filter:](#)

For any version listed below, click on its date to view it.

SENSORS

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



HOW TO OBTAIN DATA?

COLLECT IT

- OBSERVATION
- SURVEYS
- LOGGING
- SENSORS
- CROWDSOURCING

FIND OR EXTRACT IT

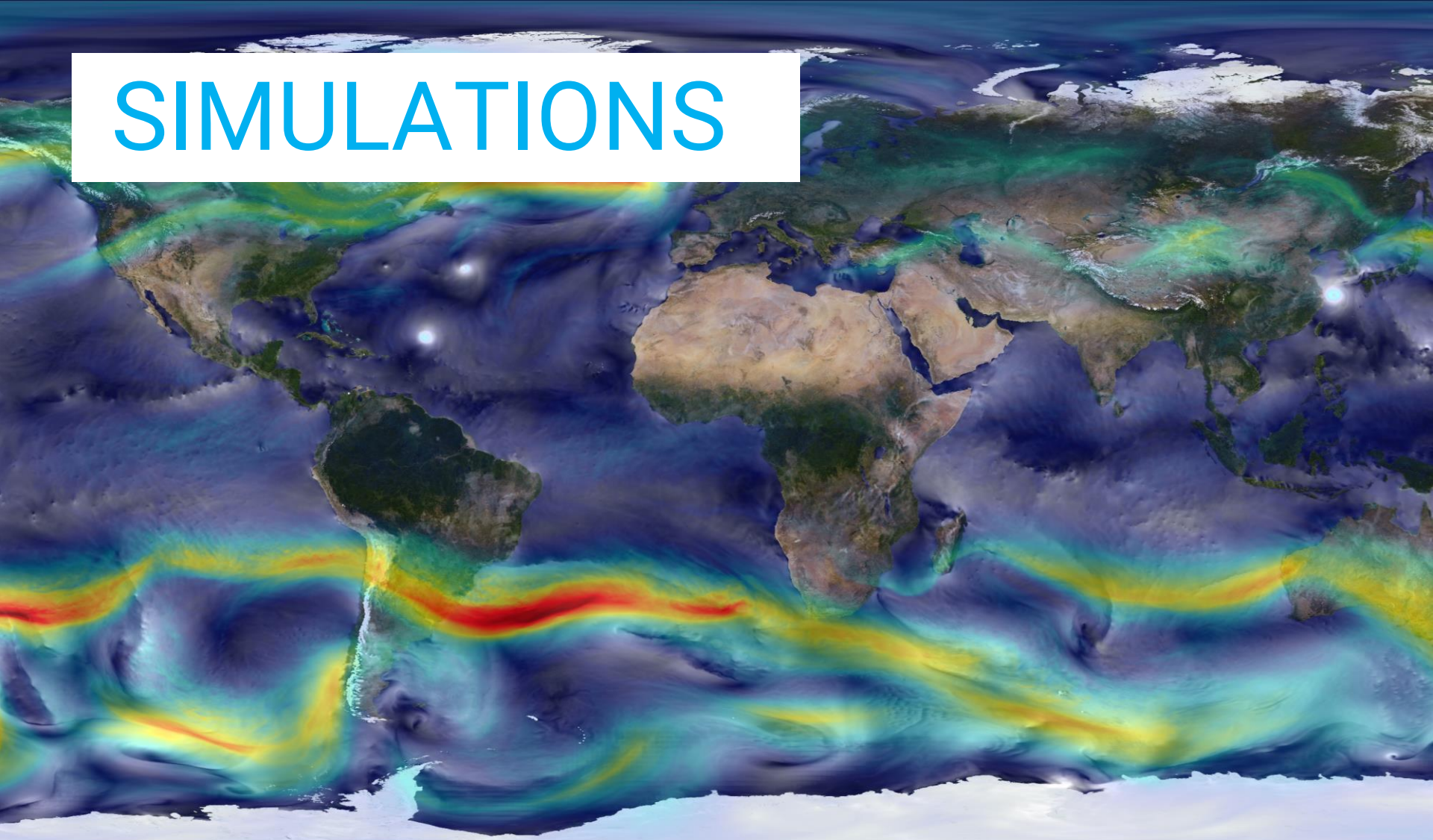
- OPEN CORPUSES
- DATA RETAILERS
- APIS
- SCRAPING THE WEB

GENERATE IT

- SIMULATIONS

GENERATING DATA

SIMULATIONS



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

The Upshot

EDITED BY DAVID LEONHARDT
FOLLOW US: [f](#) [t](#) [RSS](#)
GET THE UPSHOT IN YOUR INBOX

SHARE

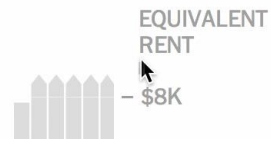
Is It Better to Rent or Buy?

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



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

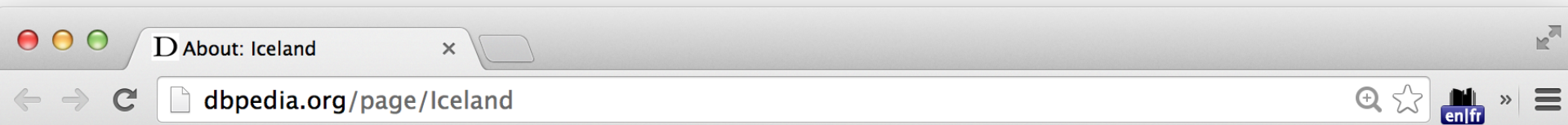
GENERATE IT

- SIMULATIONS

FINDING AND EXTRACTING EXISTING DATA

LARGE OPEN CORPUSES

DBPEDIA



About: Iceland

An Entity of Type : [place](#), from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](#)

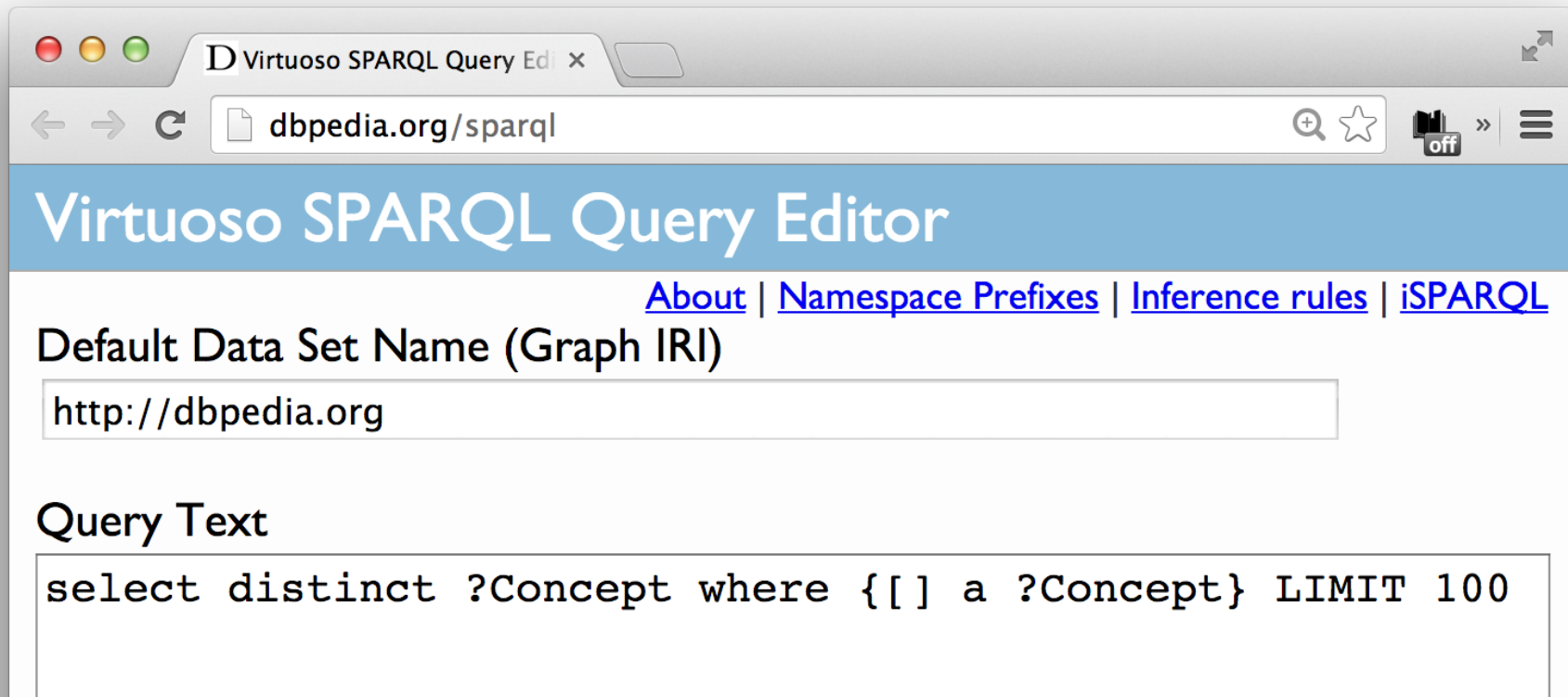


Iceland /ˈaɪslənd/ (Icelandic: Ísland [ˈistlant]), sometimes referred to in full as the Republic of Iceland (Lýðveldið Ísland), is a Nordic island country marking the juncture between the North Atlantic and the Arctic Ocean, on the Mid-Atlantic Ridge. The country has a population of 325,671 and a total area of 103,000 km² (40,000 sq mi), which makes it the most sparsely populated country in Europe.

Property

Value

QUERYING DBPEDIA



The image shows a browser window with the title "Virtuoso SPARQL Query Editor". The address bar shows "dbpedia.org/sparql". The page has a blue header with the title "Virtuoso SPARQL Query Editor". Below the header, there are navigation links: "About", "Namespace Prefixes", "Inference rules", and "iSPARQL". The main content area has a section titled "Default Data Set Name (Graph IRI)" with a text input field containing "http://dbpedia.org". Below that is a section titled "Query Text" with a text area containing the SPARQL query: "select distinct ?Concept where {[] a ?Concept} LIMIT 100".

Virtuoso SPARQL Query Editor

[About](#) | [Namespace Prefixes](#) | [Inference rules](#) | [iSPARQL](#)

Default Data Set Name (Graph IRI)

Query Text

```
select distinct ?Concept where {[] a ?Concept} LIMIT 100
```


FREEBASE

The screenshot shows the Freebase website interface. At the top, there is a navigation bar with the Freebase logo, a search input field labeled "Find...", and links for "Browse", "Query", and "Help". On the right side of the navigation bar, there are links for "Sign In or Sign Up" and "English". Below the navigation bar, a large banner displays the number "2,653,581,676" in a dark box, with the text "Facts (and counting)" to its right. Below the banner, a tagline reads "A community-curated database of well-known people, places, and things". At the bottom of the banner area, there is a horizontal menu with several tabs: "Data" (highlighted in yellow), "Schema", "Queries", "Apps", "Loads", "Review Tasks", and "Users".

Explore Freebase Data

Domain	ID	Topics	Facts
Music	/music	29M	200M
Books	/book	6M	15M
Media	/media_common	5M	16M

How can you get started?

Learn how it works

Discover what kind of information Freebase contains, how it's organized, and how Freebase allows you to uniquely identify identities anywhere on the web

WIKIDATA



Main Page Discussion

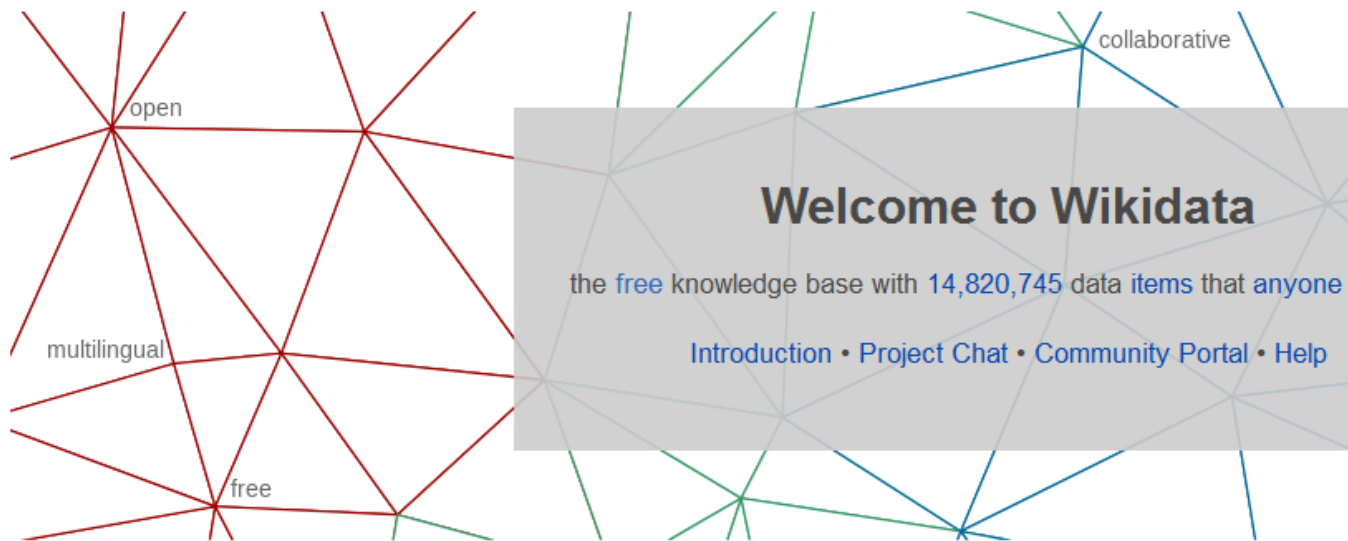
- Main page
- Community portal
- Project chat
- Create a new item
- Item by title
- Recent changes
- Random item
- Help
- Donate

Print/export

- Create a book
- Download as PDF
- Printable version

Tools

- What links here
- Related changes
- Special pages
- Permanent link
- Page information



Welcome!

Wikidata is a free linked database that can be read and edited by both humans and machines.

Wikidata acts as central storage for the **structured data** of its Wikimedia sister projects including Wikipedia

Learn about data

New to the wonderful world of Wikidata? Get up to speed and feeling comfortable with Wikidata.

PROJECT GUTENBERG

All Books (sorted by popularity) x

www.gutenberg.org/ebooks/search/?sort_order=downloads

Project Gutenberg offers 46,845 free ebooks to download.

Donate Flattr this!

[Search](#) [Latest](#) [Terms of Use](#) [Bookmarks](#) [Donate?](#) [Mobile](#)

Search Project Gutenberg. <: Help

All Books (sorted by popularity)

A

[Sort Alphabetically](#)

3

[Sort by Release Date](#)

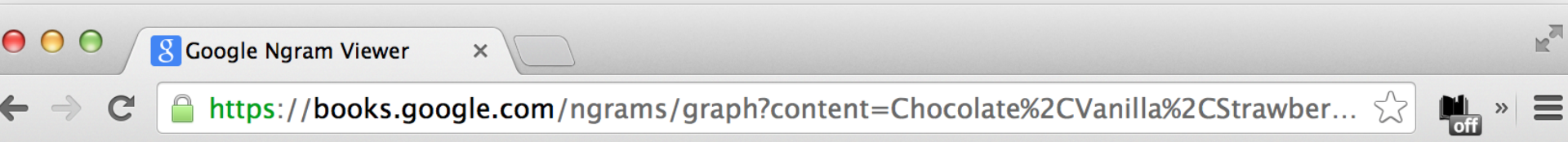


[The Kama Sutra of Vatsyayana](#)

Vatsyayana

13285 downloads

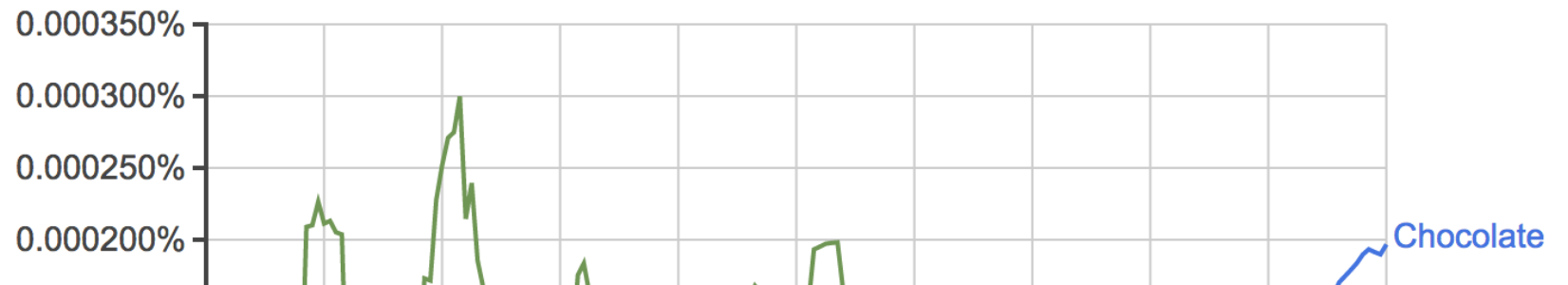
GOOGLE N-GRAMS



Google books Ngram Viewer

Graph these comma-separated phrases: case-insensitive

between and from the corpus with smoothing of . [Search lots of books](#)



FINDING AND EXTRACTING EXISTING DATA

GOVERNMENT AND INTERNATIONAL
DATA INITIATIVES

DATA.WORLDBANK.ORG

The screenshot shows the World Bank Data website. At the top, it says "Data | The World Bank" and "data.worldbank.org". The main header includes "THE WORLD BANK" logo and "Working for a World Free of Poverty". Navigation tabs include Home, About, Data, Research, Learning, News, and Projects & Operations. A "Data" section is highlighted in red. Below it, there are tabs for "By Country", "By Topic", "Indicators", "Data Catalog", and "Microdata". A language selector shows "English" selected. The main content area features a "World Bank Open Data" section and a "Find an indicator" search box. A "BROWSE DATA" section includes "By Country" and "Indicators". A "FEATURED" section lists "World Development Indicators", "Open Finances", "Projects & Operations", and "Open Government Toolkit".

Home About **Data** Research Learning News Projects & Operations

Data

By Country By Topic Indicators Data Catalog Microdata

This page in English Español Français العربية 中文

World Bank Open Data: free and open access to data about development in countries

Engagement Area
Outcomes
Indicators
Projects

Launch of a live, interactive application for the India Country Partnership Strategy

Posted on 22 Sep 2014

Global child mortality rate dropped 49% since 1990

Emi Suzuki | Posted on 16 Sep 2014

Taking a closer look at youth-related data: regional trends, differences

Hiroko Maeda | Posted on 2 Sep 2014

View all news | View all blogs »

The screenshot shows the OECD Data Lab website. At the top, it says "Data Lab - OECD" and "www.oecd.org/statistics/datalab/#d.en.227006". The main header includes the "OECD" logo and "BETTER POLICIES FOR BETTER LIVES". Navigation tabs include OECD Home, About, Countries, Topics, Statistics, and Newsroom. A "Follow us" section includes links for E-mail Alerts, Blogs, RSS, Twitter, Facebook, and YouTube. Below the navigation, there are lists of topics such as Agriculture and fisheries, Education, Innovation, Regulatory reform, etc. A "topics:" section lists various categories like Agriculture and fisheries, Competition, Development, etc. The main content area features several data visualizations, including "Product Market Regulation", "OECD-FAO Agricultural Outlook", "Economic Outlook projections", "Employment Outlook", "Environmental Outlook", and "Climate Change".

Data Lab - OECD
www.oecd.org/statistics/datalab/#d.en.227006

OECD

BETTER POLICIES FOR BETTER LIVES

Follow us
E-mail Alerts Blogs RSS Twitter Facebook YouTube

OECD Home **About** Countries Topics Statistics Newsroom

- > Agriculture and fisheries
- > Bribery and corruption
- > Chemical safety and biosafety
- > Competition
- > Corporate governance
- > Development
- > Economy
- > Education
- > Employment
- > Environment
- > Finance
- > Green growth and sustainable development
- > Health
- > Industry and entrepreneurship
- > Innovation
- > Insurance and pensions
- > International migration
- > Internet
- > Investment
- > Public governance
- > Regional, rural and urban development
- > Regulatory reform
- > Science and technology
- > Social and welfare issues
- > Tax
- > Trade

topics : all

- Agriculture and fisheries
- Competition
- Development
- Economy
- Education
- Employment
- Energy
- Environment
- Finance
- Green
- growth and sustainable
- development
- Health
- Industry

Product Market Regulation

OECD-FAO Agricultural Outlook

Economic Outlook projections

Employment Outlook

Environmental Outlook

Climate Change

Adult Skills

Urban Explorer

Facebook on iPhone

DATA.OECD.ORG

GOVERNMENT INITIATIVES

WWW.DATA.GOV (US)

DATA.GOV.UK

DATA.GOV.BE

The screenshot shows the homepage of Data.gov. At the top, there is a navigation bar with the Data.gov logo and the text "DATA TOPICS". Below this, a large blue banner contains the text "The home of the data" and "Here you will find data, tools, web and mobile applications". A search bar is visible on the right side of the page. At the bottom, there is a section titled "BROWSE TOPICS" with a search input field containing the text "Health Care Provider Charge".

The screenshot shows the search results page on Data.gov.uk. The header includes the Data.gov.uk logo and the tagline "Opening up Government". Below the header, there is a search bar with the text "Search for data..." and a link to "conduct map based search". A green bar indicates "19422 Results". The first result is titled "Live traffic information from the Highway" and is from the "Highways Agency". The description states: "Live traffic information data showing traffic information on the road network in England, maintained by the Highways Agency. August 2013 Following a change of...". The second result is titled "Learning Aim Reference Service" and is from the "Skills Funding Agency". The description states: "Learning Aim Reference Service (LARS) service will offer a 'Quick facility, allowing users to search by most commonly used fields full set of search fields will still...".

The screenshot shows the datasets page on Data.gov.be. The header includes the Data.gov.be logo and the text "Beta". Below the header, there is a navigation bar with the text "HOME CONDITIONS D'UTILISATION DATA APPS IDÉES FORUM". A search bar is visible. Below the search bar, there is a section titled "Liste de sets de données disponibles comme 'open data'". There are three dropdown menus for "Catégorie", "Type", and "Granularité", all set to "- Tout -". An "Appliquer" button is next to the dropdowns. Below the dropdowns, there is a table with the following data:

Titre	Catégorie	Type
Zones de stationnement voirie 2013	Mobilité	Téléchargement
Usages TIC des ménages wallons	TIC	Téléchargement Service web
Usages TIC des citoyens wallons	Population, Economie, TIC	Téléchargement Service web
UDP Mars 2013 par commune	Energie, Pouvoirs publics	Téléchargement
UDP Mai 2013 par commune	Energie, Pouvoirs publics	Téléchargement

NEW DATA INITIATIVES JUST TO TRACK ALL THE DATA INITIATIVES

Browse by country | re3dat x

← → ↻ www.re3data.org/browse/by-country/

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Home Search **Browse** Suggest FAQ

About Schema Contact Imprint

INITIATIVES IN FRANCE

[HTTP://DATA.GOUV.FR](http://data.gouv.fr)

The image shows two overlapping browser windows. The background window is the homepage of data.gouv.fr, featuring a search bar, a navigation menu with categories like 'Agriculture et alimentation' and 'Culture', and a 'Partagez, les données' section. The foreground window is the 'Open Data Paris' website, displaying the 'PARISDATA' logo, a navigation menu with 'Les données', 'Les Data Challenges', 'L'API', 'La licence', 'La démarche', and 'Le forum', and a search bar with the text 'Trouver un jeu de données...'. A 'Zones de rencontre' section is partially visible at the bottom.

[HTTP://OPENDATA.PARIS.FR/EXPLORE/](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

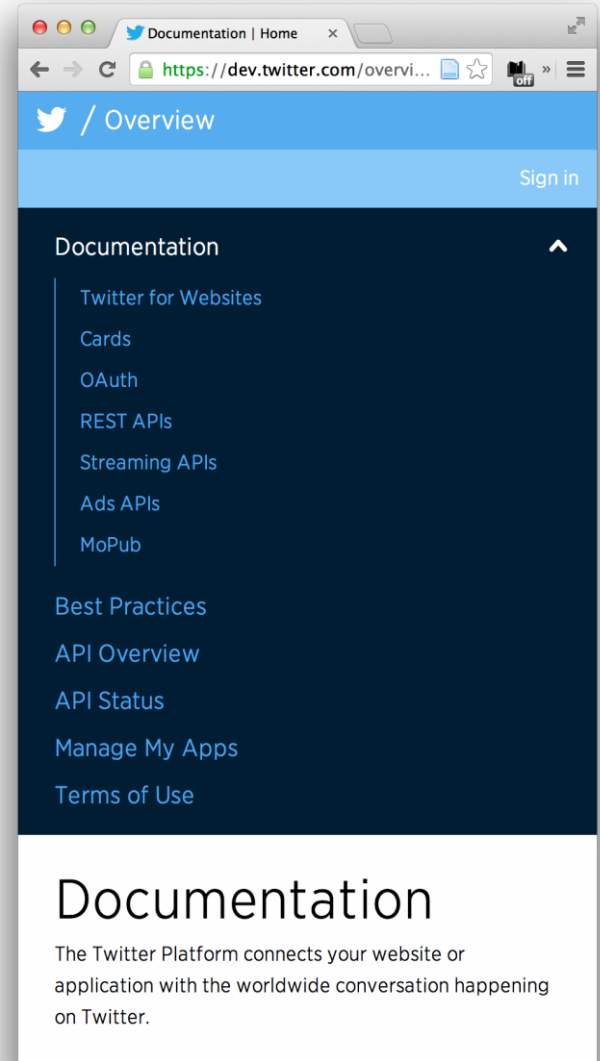
TWITTER

Streaming APIs (live data by users and by topics)

The “Firehose” (all of live twitter)

Complete Archives via “Gnip” and eventually (maybe) the US Library of Congress

[HTTPS://DEV.TWITTER.COM](https://dev.twitter.com)



CNN #COP17 ECOSPHERE PROJECT

The CNN #COP17 ECOSPHERE
Project launched on 14
November 2011.

This is a timeline of how the
ECOSPHERE develops in the
build-up to the COP17
Conference in Durban.

◀ Back to ECOSPHERE



1276 tweets

[HTTP://CNN-ECOSPHERE.COM/](http://CNN-ECOSPHERE.COM/)

◀ 14 Nov 2011 ▶

ECOSPHERE TIMELINE

Learn More
about the project

CNN.com

Tottenham Riots

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

Search

(enter search terms here)

Search

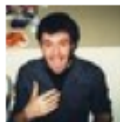
Sort

times retweeted

Show Sources (showing 8 of 10 sources loaded)

All Ordinary People Journalists / Bloggers Organizations Uncategorized Eyewitnesses

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

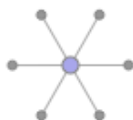


Myself in 160 characters: Schizophrenic. Also a criminologist

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,

London, United Kingdom



Journo/Blogger

41

RTed

56

Klout

Show Tweets

All Exclude RTs Images & Videos

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>

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,



Ordinary Person

23

RTed

49

Klout

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 "hubs" is a racist and an idiot. #tottenham #tottenhamriots

SRSR

[DIAKOPOULOS ET AL. 2012]

GOOGLE EARTH ENGINE

[HTTPS://EARTHENGINE.GOOGLE.ORG/](https://earthengine.google.org/)

1984

2012

MORE APIS

(APPPLICATION PROGRAMMING INTERFACES)

NEW YORK TIMES APIS

<http://developer.nytimes.com/>

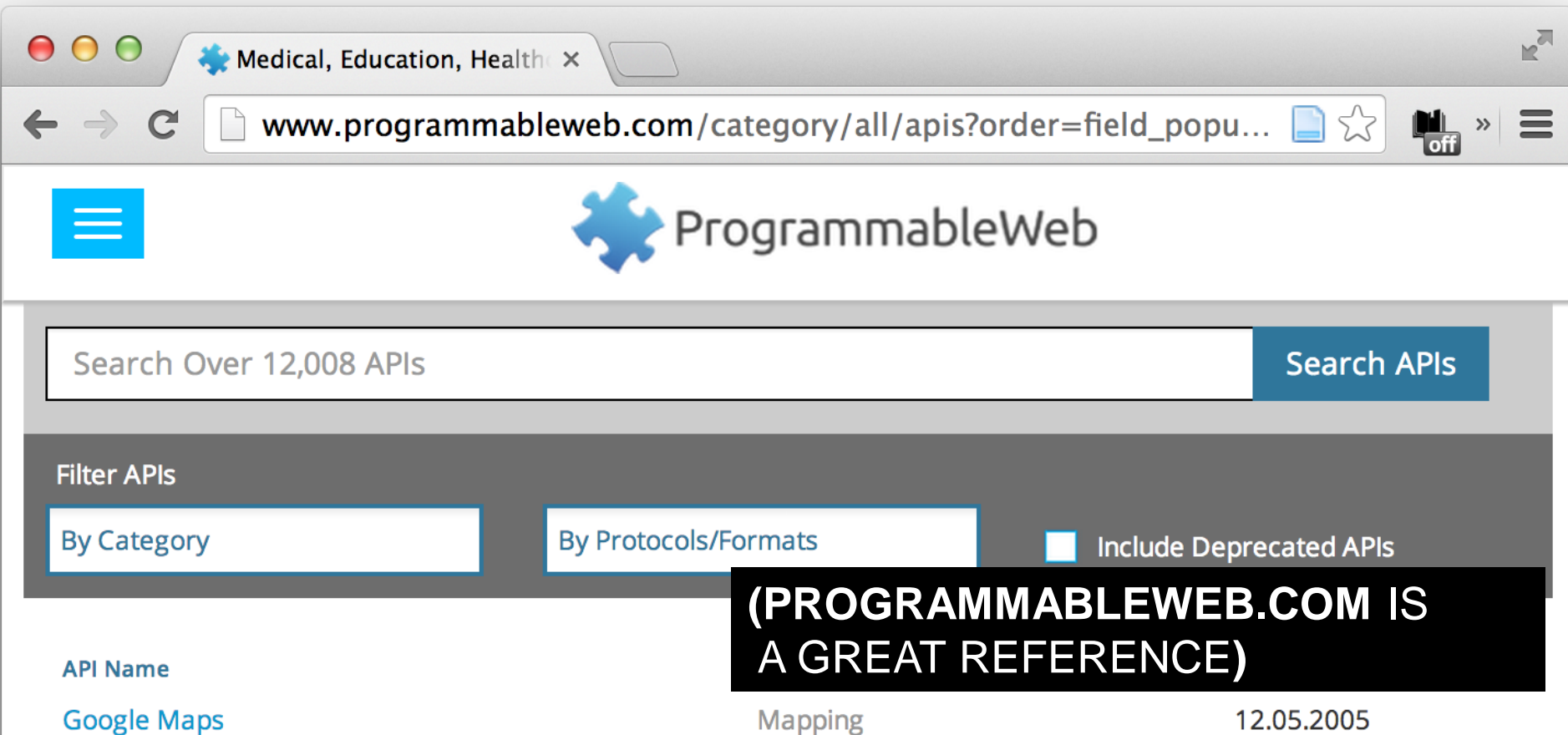
(Archival news articles from 1851, books, movies, geographical, and political data)

OPEN STREET MAP

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

(Detailed location and map data for the whole world)

AND THE LIST GOES ON!



The image shows a browser window with the URL `www.programmableweb.com/category/all/apis?order=field_popu...`. The page features the ProgrammableWeb logo and a search bar with the text "Search Over 12,008 APIs" and a "Search APIs" button. Below the search bar, there are filter options: "By Category", "By Protocols/Formats", and a checkbox for "Include Deprecated APIs". A black callout box with white text is overlaid on the bottom right of the page, stating "(PROGRAMMABLEWEB.COM IS A GREAT REFERENCE)".

Medical, Education, Health x

www.programmableweb.com/category/all/apis?order=field_popu...

Search Over 12,008 APIs Search APIs

Filter APIs

By Category By Protocols/Formats Include Deprecated APIs

API Name

Google Maps Mapping 12.05.2005

(PROGRAMMABLEWEB.COM IS A GREAT REFERENCE)

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

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 breakthrough 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

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 Pellizzotti, 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)

2009

Gap Behind Leader

Tour de France 2009
Rider Times by Stage

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)	+0h 46' 34"
	TEAM SAXO BANK	
	PELLIZZOTTI Franco (ITA)	+0h 56' 19"
	LIQUIGAS	
	LEFEVRE Laurent (FRA)	+1h 01' 29"
	BBOX BOUYGUES TELECOM	









Le Tour de France
07/05 > 07/27/2014

Facebook Twitter Google+ YouTube
EN
Sunday July 27th, 2014
Stage 21
Évry / Paris Champs-Élysées

5 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

 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

Google



Sheets

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] (km ²)	Density (/km ²)	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	Guirarat	State	60.439.692	4.99	196.024	308	31.491.260	28.948.432	919	78.03	31.740.767	18.930.250



\$ % .0 .00 123 ▾

Arial ▾

10 ▾

B *I* ~~S~~ A ▾

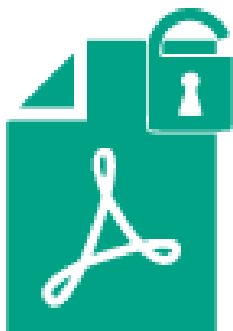


f_x | =ImportHtml("http://en.wikipedia.org/wiki/Demographics_of_India", "table", 4)

	A	B	C	D	E	F
1	Rank	State / Union Territory	Type	Population	% [18]	Area [19] (km ²)
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

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

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 discovery		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 analysis		1 machine learning & statistics	55
9	program analysis		1 analysis process	113
10	multi-variate statistics		1 multidimensional / multivariate	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 factorization		1 matrices	10
16	interactive clustering		3 clustering	50

CSV BEST PRACTICES

Splits cells if you can, if needed create a second

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 Yacoub
59	66	C	Visualizing the third	Del Lamb; Amit Bandopadhyay
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; Pat
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				
17				

!

R doesn't like spaces, special characters,
Columns names that start with a number

Better:

Customer_ID vs. Customer ID

Customer.ID vs. Customer ID

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)

```
---
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

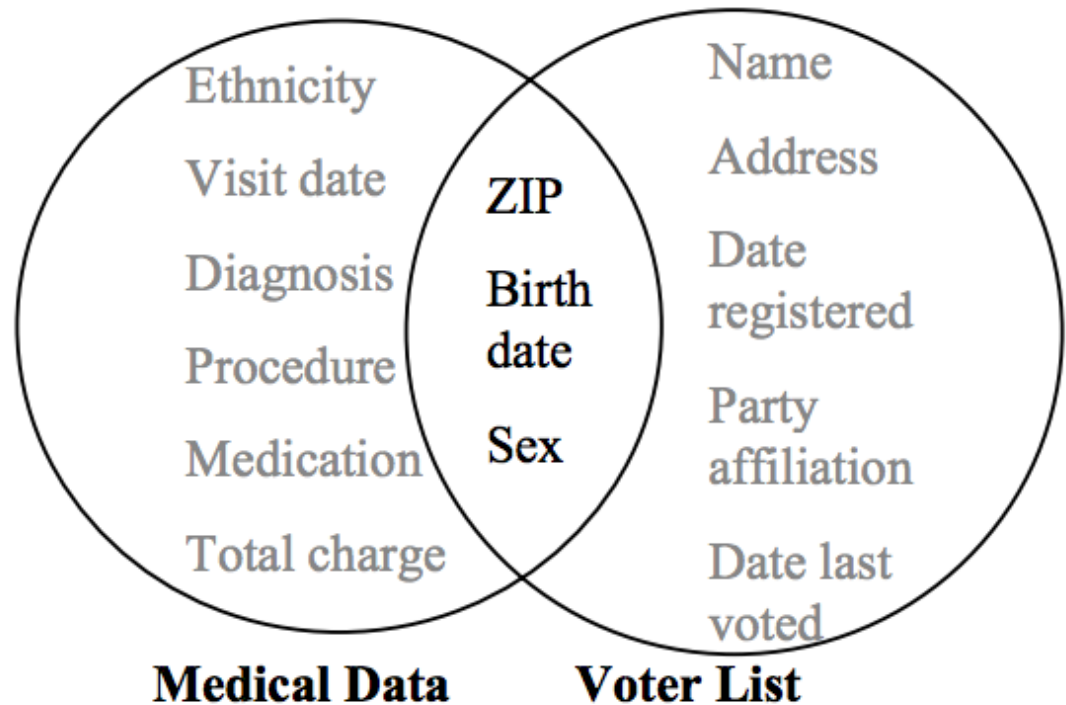
- Be sure you know who owns the data.
- Think early on about whether or not you'll need to publish or (re)share data.
- Be careful you aren't violating copyright, especially when scraping.

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

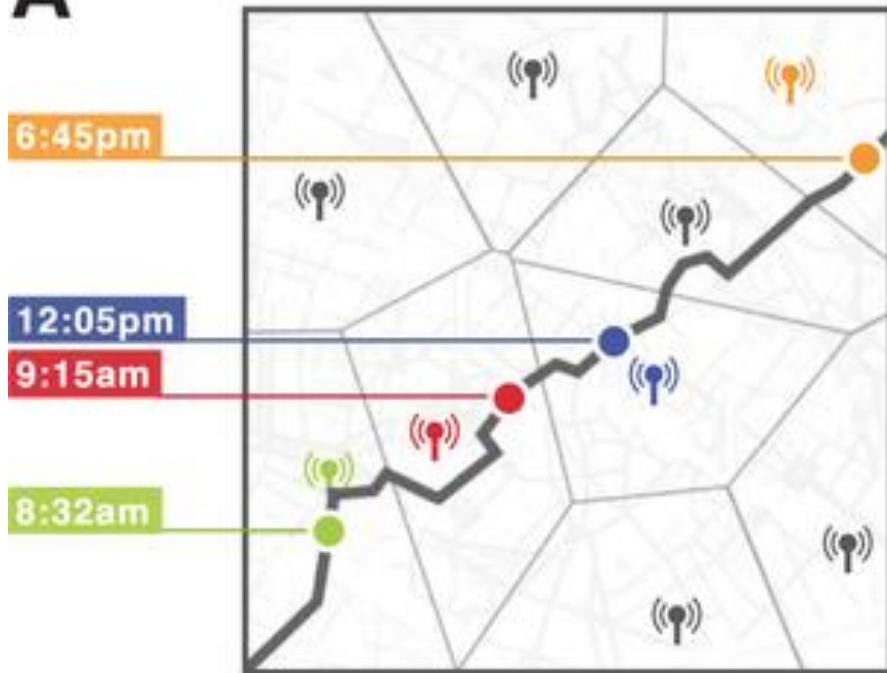


[L. Sweeney. 2002]

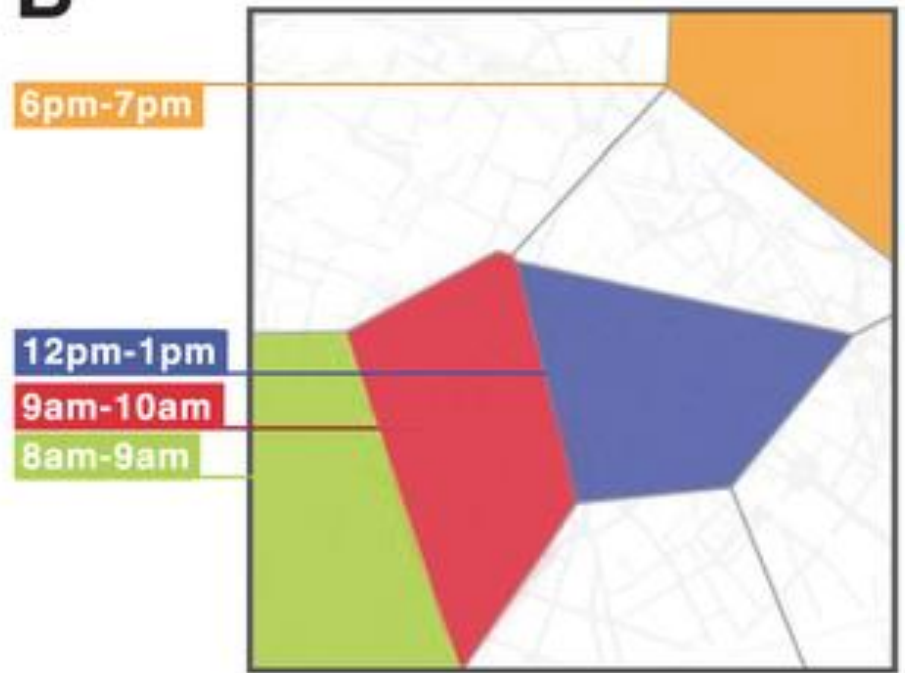
[k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY](#)

LOCATION DATA IS ESPECIALLY SENSITIVE

A



B



[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.

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 via social networks

Adam D. I. ...

Core Data
CA 94142

Edited

Emo
cont.

without the
in laboratory exper

negative emotions to others. ...
network, collected over a 20-y period ...
moods (e.g., depression, happiness) can be ...
networks [Fowler JH, Christakis NA (2008) *BMJ* 337: ...]
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 re-
duced, 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-
verbal cues are not strictly necessary for emotional contagion.

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 v ...
stories, and activities undertaken by ...
primary manner by which people see content th ...
Which content is shown or omitted in the News Feed ...
determined 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

PSYCHOLOGY
COGNITIVE SCIENCE

“EXPERIMENTING ON HUMAN BEINGS”



Dating Research from OkCupid

We Experiment On Human Beings!

July 28th, 2014 by [Christian Rudder](#)

Tweet 2,760

Share 10k

I'm the first to admit it: we might be popular, we might create a lot of great relationships, we might blab blab blab. But OkCupid doesn't

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)

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 WEEK'S CLASS

INSTALL :



OpenRefine (formerly Google Refine)

<http://openrefine.org/>