

DATA COLLECTION

Slides 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?

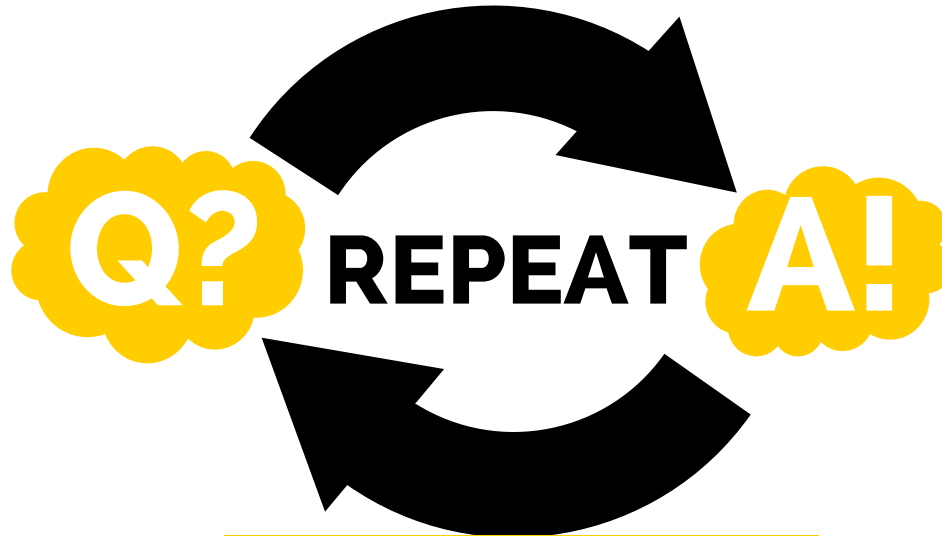
(HINT: Usually, because you have a question you need to answer!)

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'll revisit this later in the course...

**(...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




Ask Questions
Get Answers



Collect Analyze

Create your form

The Ridiculously Power



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

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




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%	(±3.0 / -3.0)
Online discounts	24.9%	(±2.8 / -2.8)
Ability to return in store	16.4%	(±2.4 / -2.5)
Free returns	16.1%	(±2.1 / -2.0)

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.




CROWDSOURCING DATA COLLECTION

Amazon Mechanical Turk - x

https://www.mturk.com/mturk/welcome

HITs containing 'short survey'

11-20 of 49 Results

Sort by:  [Show all details](#) | [Hide all details](#) [First](#) << [Previous](#) < 1 2 3 4 5 > [Next](#) >> [Last](#)

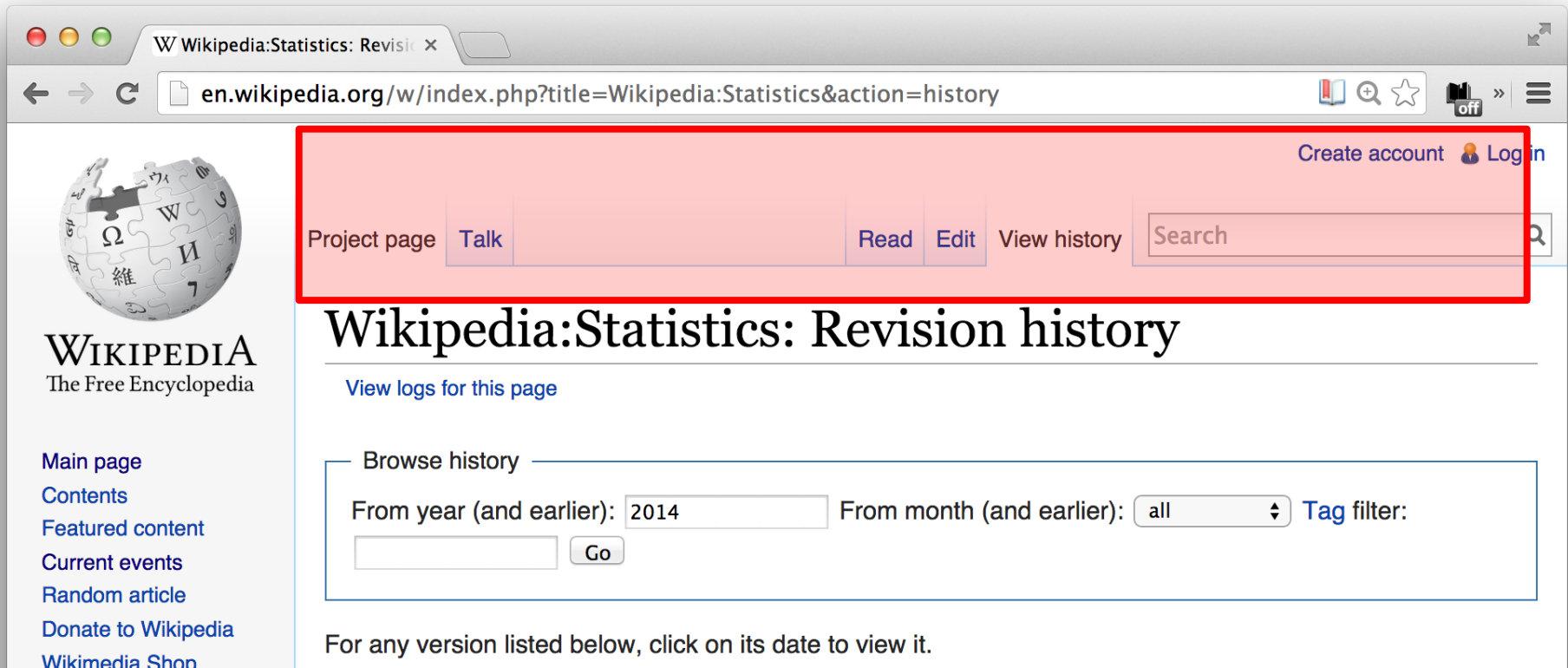
Answer a short survey about Work Team Dynamics		Request Qualification (Why?) View a HIT in this group
Requester: Whitney Ohmer	HIT Expiration Date: Oct 12, 2014 (2 weeks 5 days)	Reward: \$0.25
Time Allotted: 60 minutes		HITs Available: 1
Answer a short survey about Work Team Dynamics		Request Qualification (Why?) View a HIT in this group
Requester: Whitney Ohmer	HIT Expiration Date: Oct 12, 2014 (2 weeks 5 days)	Reward: \$0.25
Time Allotted: 60 minutes		HITs Available: 1
Short Survey		View a HIT in this group
Requester: David Tannenbaum	HIT Expiration Date: Oct 12, 2014 (2 weeks 5 days)	Reward: \$0.10
Time Allotted: 60 seconds		HITs Available: 1
Short survey about website experience (on average it takes 13 minutes)		Not Qualified to work on this HIT (Why?) View a HIT in this group

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



The screenshot shows a web browser window displaying the Wikipedia page for "Wikipedia:Statistics: Revision history". The browser's address bar shows the URL: `en.wikipedia.org/w/index.php?title=Wikipedia:Statistics&action=history`. The page features a navigation bar with links for "Project page", "Talk", "Read", "Edit", and "View history", along with a search box. The main heading is "Wikipedia:Statistics: Revision history", and there is a link to "View logs for this page". Below this, a "Browse history" section includes a form for filtering by year (set to 2014) and month (set to all), with a "Go" button. The page also includes a sidebar with navigation links and a footer with the text "For any version listed below, click on its date to view it."

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 Search

Wikipedia:Statistics: Revision history

[View logs for this page](#)

Browse history

From year (and earlier): 2014 From month (and earlier): all Tag filter:

Go

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

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

SENSORS

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



HOW TO OBTAIN DATA?

COLLECT IT

- OBSERVATION
- SURVEYS
- LOGGING
- SENSORS
- CROWDSOURCING

FIND OR EXTRACT IT

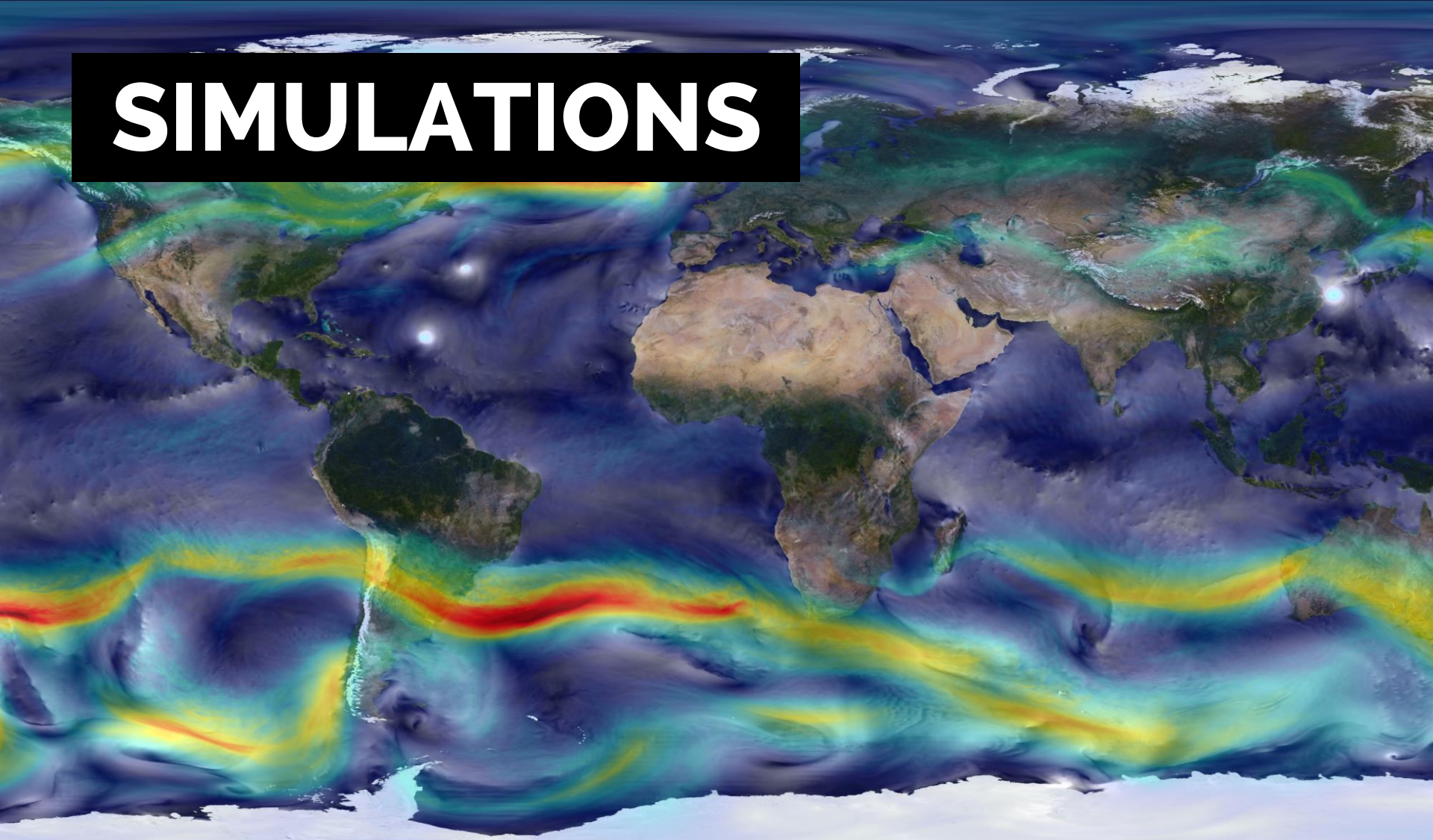
- OPEN CORPUSES
- DATA RETAILERS
- APIS
- SCRAPING THE WEB

GENERATE IT

- SIMULATIONS

GENERATING DATA

SIMULATIONS



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



The Upshot

EDITED BY DAVID LEONHARDT

FOLLOW US:

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

dbpedia.org/page/Iceland

About: Iceland

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

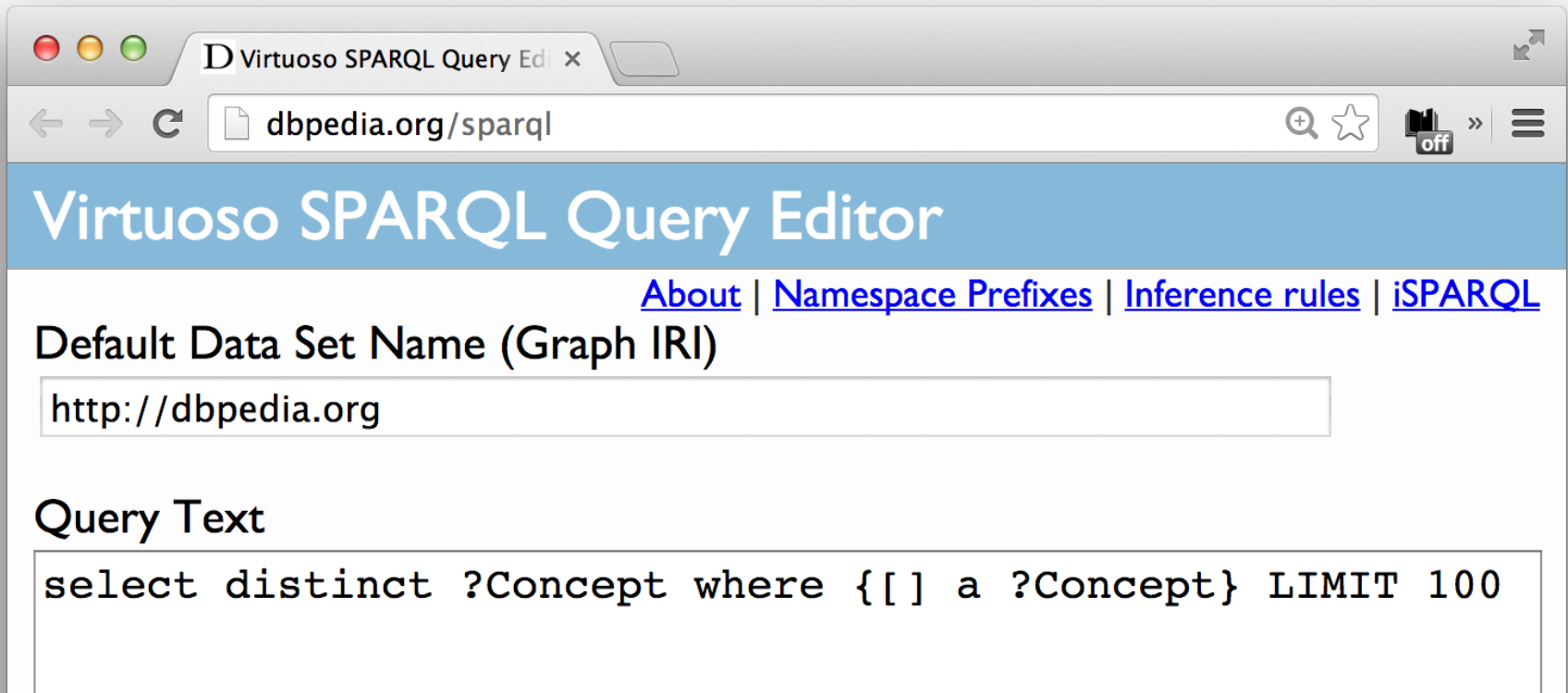


Iceland /ˈaɪslənd/ (Icelandic: Ísland [ˈistlant]), sometimes referred to in full as the Republic of Iceland (Lýðveldið Ísland), is a Nordic island country marking the juncture between the North Atlantic and the Arctic Ocean, on the Mid-Atlantic Ridge. The country has a population of 325,671 and a total 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 screenshot of a web browser window displaying the Virtuoso SPARQL Query Editor. The browser's address bar shows the URL `dbpedia.org/sparql`. The page title is "Virtuoso SPARQL Query Editor". Below the title, there are navigation links: [About](#), [Namespace Prefixes](#), [Inference rules](#), and [iSPARQL](#). The "Default Data Set Name (Graph IRI)" field contains the text `http://dbpedia.org`. The "Query Text" field contains the SPARQL query: `select distinct ?Concept where {[] a ?Concept} LIMIT 100`.

Virtuoso SPARQL Query Editor

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

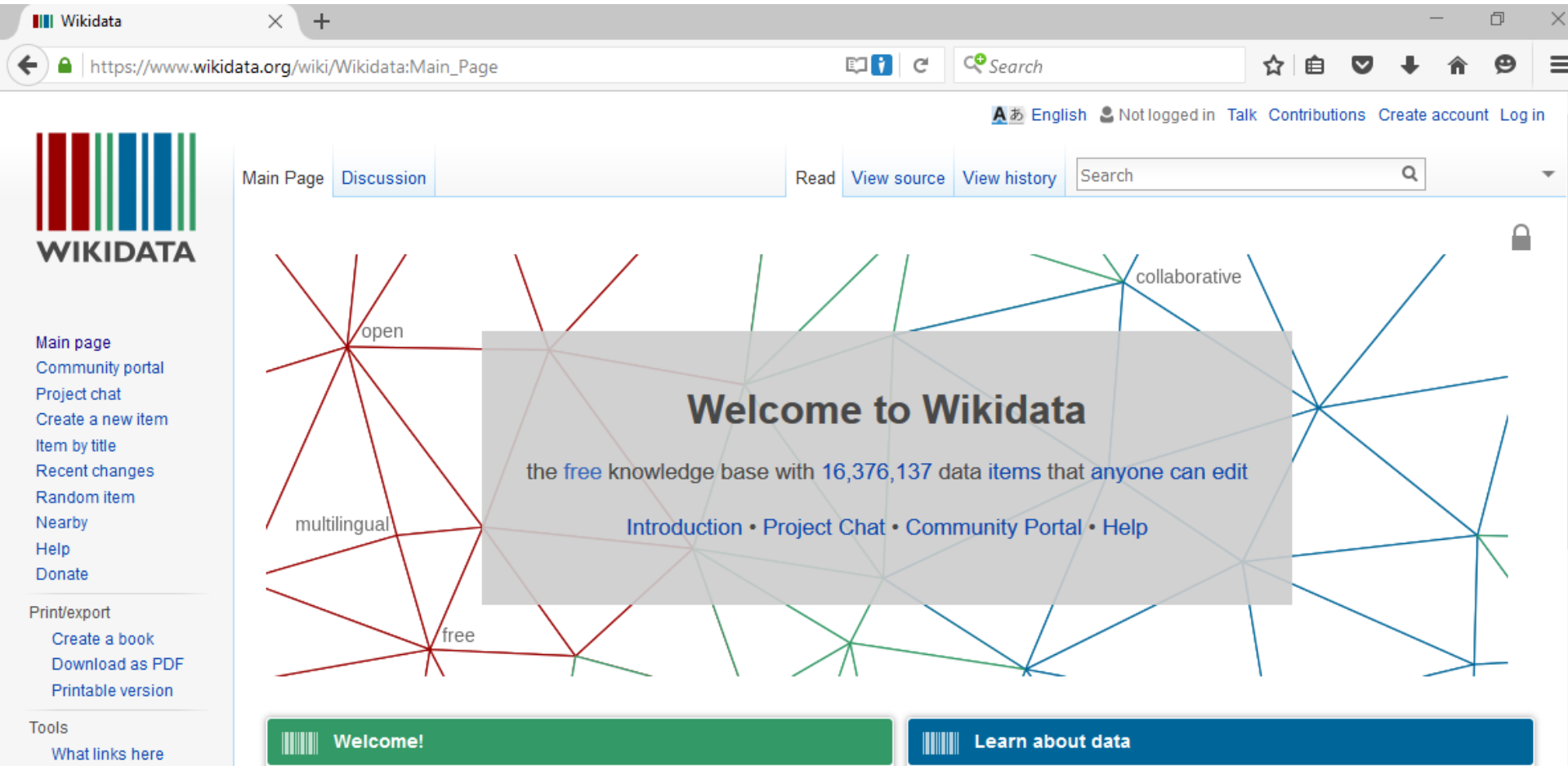
Default Data Set Name (Graph IRI)

`http://dbpedia.org`

Query Text

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


WIKIDATA



The image shows a browser window displaying the Wikidata Main Page. The browser's address bar shows the URL https://www.wikidata.org/wiki/Wikidata:Main_Page. The page features a navigation menu with options like "Main Page", "Discussion", "Read", "View source", and "View history". A search bar is present in the top right. The main content area displays a large, semi-transparent gray box with the text "Welcome to Wikidata" and "the free knowledge base with 16,376,137 data items that anyone can edit". Below this, there are links for "Introduction", "Project Chat", "Community Portal", and "Help". The background of the page is a network diagram with nodes and edges, some labeled with terms like "open", "multilingual", "free", and "collaborative".

Wikidata

https://www.wikidata.org/wiki/Wikidata:Main_Page

English Not logged in Talk Contributions Create account Log in

Main Page Discussion Read View source View history Search

Welcome to Wikidata

the free knowledge base with 16,376,137 data items that anyone can edit

[Introduction](#) • [Project Chat](#) • [Community Portal](#) • [Help](#)

open collaborative

multilingual free

Print/export

- Create a book
- Download as PDF
- Printable version

Tools

- What links here

Welcome!

Learn about data

PROJECT GUTENBERG

All Books (sorted by popul: 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](#)

[Sort by Release Date](#)

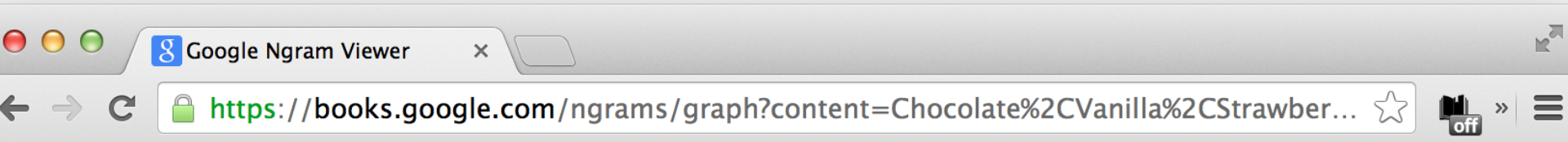


[The Kama Sutra of Vatsyayana](#)

Vatsyayana

13285 downloads

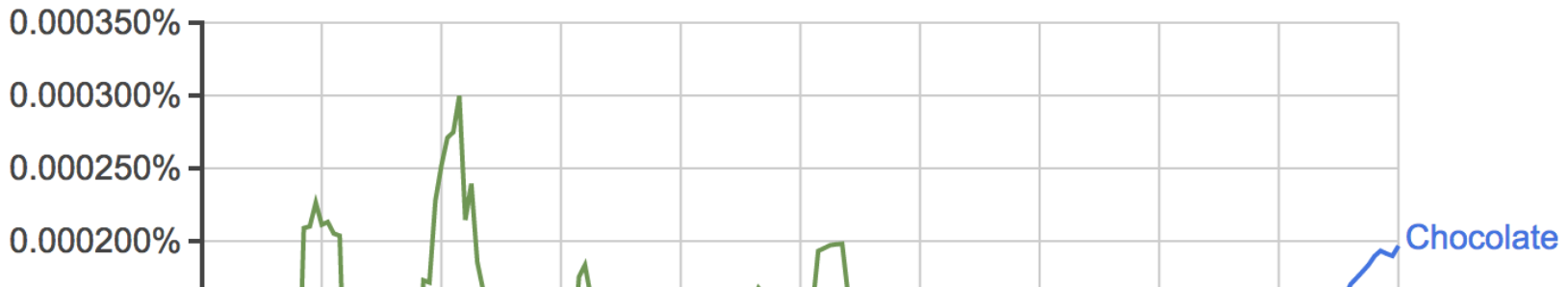
GOOGLE N-GRAMS



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 Data World Bank website. At the top, there's a navigation bar with 'Home', 'About', 'Data', 'Research', 'Learning', 'News', and 'Projects & Operations'. Below this is a 'Data' section with sub-links for 'By Country', 'By Topic', 'Indicators', 'Data Catalog', and 'Microdata'. A sidebar on the right contains a 'Find an indicator' search box, 'BROWSE DATA' by country, and 'FEATURED' articles. One featured article is titled 'Launch of a live, interactive application for the India Country Partnership Strategy'.

The screenshot shows the Data Lab - OECD website. It features the OECD logo and the tagline 'BETTER POLICIES FOR BETTER LIVES'. The main navigation includes 'OECD Home', 'About', 'Countries', 'Topics', 'Statistics', and 'Newsroom'. A grid of topic links is visible, including 'Agriculture and fisheries', 'Education', 'Innovation', 'Regulatory reform', 'Science and technology', 'Social and welfare issues', 'Tax', and 'Trade'. Below the navigation, there are several data visualization cards for 'Product Market Regulation', 'OECD-FAO Agricultural Outlook', 'Environmental Outlook', and 'Climate Change'. A large black banner at the bottom of the page reads 'DATA.OECD.ORG'.

GOVERNMENT INITIATIVES

WWW.DATA.GOV (US)

DATA.GOV.UK

DATA.GOV.BE

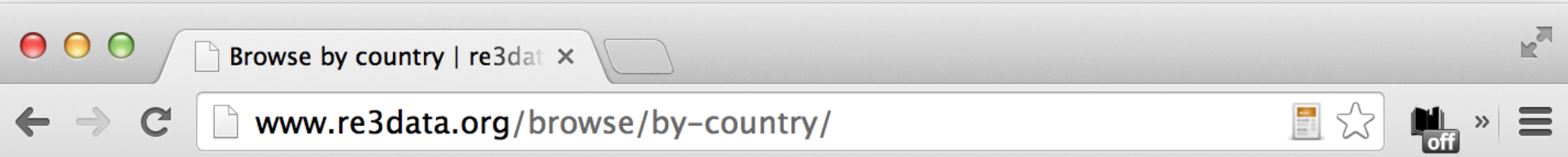
The screenshot shows the homepage of Data.gov. At the top, there is a navigation bar with "DATA" and "TOPICS". Below this, the main heading reads "The home of the data" in a large, bold font. Underneath, a sub-heading says "Here you will find data, tools, web and mobile applications". A search bar is visible on the right side. 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 logo "DATA.GOV.UK Beta" 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 "Show Search Facets >". The main heading is "19422 Results". The first result is titled "Live traffic information from the Highway" by 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" by 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 search results page on Data.gov.be. The header includes the logo "Data.gov.be" and a navigation bar with "HOME", "CONDITIONS D'UTILISATION", "DATA", "APPS", "IDÉES", and "FORUM". 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 "Show Search Facets >". The main heading is "19422 Results". The first result is titled "Live traffic information from the Highway" by 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" by 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...".

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



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, the French national open data portal. It features a search bar with the text 'Recherche', a navigation menu with categories like 'Agriculture et alimentation', 'Culture', 'Économie et Emploi', 'Éducation et Recherche', 'International et Europe', 'Logement, Développement Durable et Énergie', 'Santé et Social', 'Société', and 'Territoires et Transports', and a 'Partagez, les données' button. The foreground window is opendata.paris.fr/explore/, the Paris Open Data portal. It has a large 'PARISDATA' logo with the 'MAIRIE DE PARIS' logo to its right. Below the logo is a navigation menu with 'Les données', 'Les Data Challenges', 'L'API', 'La licence', 'La démarche', and 'Le forum'. There is also a 'Cartographe' link. A search bar contains the text 'Trouver un jeu de données...' and a 'Filtres' button. Below the search bar, the text 'Zones de rencontre' is visible. A purple 'Déplacement' button is partially visible on the right side of the foreground window.

[HTTP://OPENDATA.PARIS.FR/EXPLORE/](http://opendata.paris.fr/explore/)

INITIATIVES IN GERMANY

HTTPS://WWW.GOVDATA.DE



Daten Dokumente Apps Infr

Das Datenportal für Deutschland

Open Government: Verwaltungsdaten transparent, offen und frei nutzbar

Nach Datensit

OffenesDresden.de

Nächste Treffen

Jeden 1. und 3. Mittwoch im Monat bist auch Du herzlich eingeladen. Gern kannst Du auch schon vorher bescheid sagen, das Du kommst, via [meetup](#).

- 17.2.2016 19:00
open data meetup Dresden
Wir treffen uns, um gemeinsam an Civic Apps für Dresden zu arbeiten. Neueinsteigerinnen und Menschen, die nicht programmieren, sind immer herzlich willkommen! [Finde uns auf meetup.com](#)
Ort: [GCHQ](#)
- 3.3.2016 19:00
open data meetup Dresden
Wir treffen uns, um gemeinsam an Civic Apps für Dresden zu arbeiten. Neueinsteigerinnen und Menschen, die nicht programmieren, sind immer herzlich willkommen. [Finde uns auf meetup.com](#)
Ort: [GCHQ](#)

Was ist OpenData und offenesdresden.de?

Wir wollen mit öffentlichen Geldern erzeugte Daten für alle veröffentlichen und sie benutzen.

- [Open Data Portal des Freistaats Sachsen](#)
- Wikipedia: [Open Data](#)
- [Govdata](#): Das Datenportal für Deutschland
- [Open Data Network](#): Netzwerk zur Förderung von Open Government, Open Data, Transparenz und Partizipation
- [Open Knowledge Foundation](#): Werkzeuge für Offene Daten & Digitale Gemeingüter
- [Aktuelle Informationen und offene Treffen \(Twitter\)](#)
- [Open Knowledge Lab Dresden](#)
- [Mailingliste für Dresden](#)

Herzlich willkommen auf de

Nach drei Jahren haben wir GovData grundlegend erneuert. Probieren Sie es aus und geben Sie uns gerne ein [Feedback](#), v Die dazugehörige Pressemeldung können Sie im [Blog](#) lesen.

Stöbern Sie in diesen Katego



Bevölkerung (1510)

Bildung und Wissenschaft

und Gesundheit

(255)

(702)

und Wohnen

und Tourismus

HTTP://OFFENESDRESDEN.DE/

Tweets

Follow



Open Data Dresden @OpenDataDresden 4 Feb
Cooler Datensatz über alle Leute, die die #TSA pro Jahr verklagt (claimed) haben. [dhs.gov/tsa-claims-data](#) #OpenData



Open Data Dresden @OpenDataDresden 3 Feb
[@mechko](#) Wenn bekannt ist, dass es sich um den die OpenData Strategie der Stadt handelt, findet sich auch der Hinweis [ratsinfo.dresden.de/getfile.php?id...](#) Expand



Open Data Dresden @OpenDataDresden 3 Feb
[@mechko](#) Tja, herrlich. Da zeigt die Recherchefunktion vom Ratsinfo-System #Dresden wieder was sie kann. Nämlich nichts! #fail Info folgt.

**FINDING AND EXTRACTING
EXISTING DATA**

OTHER PUBLIC DATA REPOSITORIES

MORE REPOSITORIES

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>

FINDING AND EXTRACTING EXISTING DATA

DATA RETAILERS

DATA RETAILERS

AZURE DATA MARKET

<https://datamarket.azure.com/browse/data>

FACTUAL

<http://www.factual.com/>

AND AGAIN, THERE ARE MANY, MANY MORE...

FINDING AND EXTRACTING EXISTING DATA

APIS

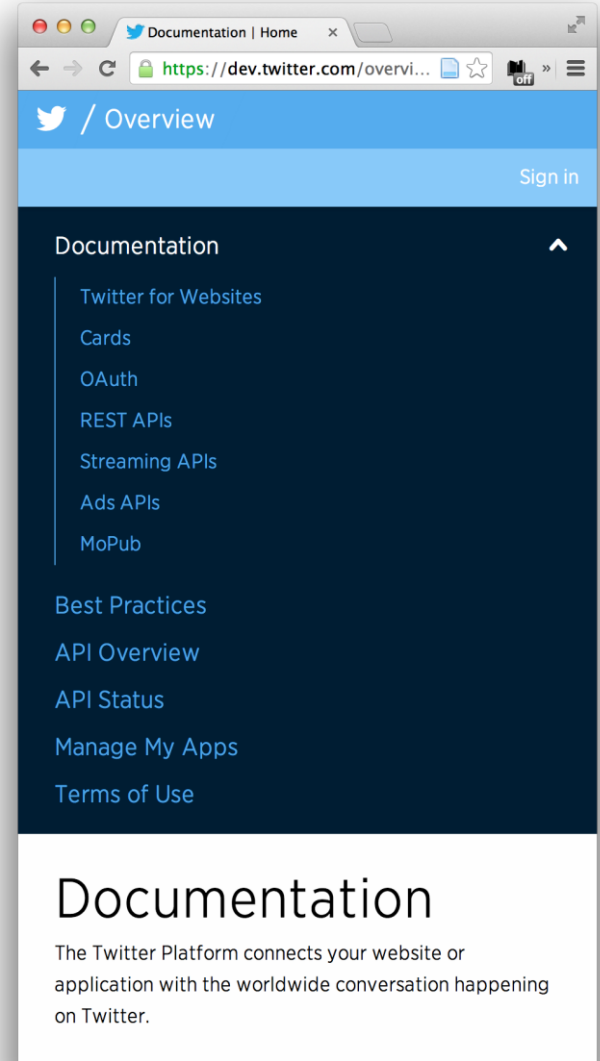
TWITTER

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

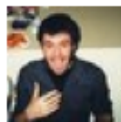
Show Sources (showing 8 of 10 sources loaded)

Show Tweets

All Ordinary People Journalists / Bloggers Organizations Uncategorized Eyewitnesses

All Exclude RTs Images & Videos

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

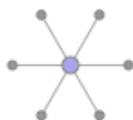


Myself in 160 characters: Schizophrenic. Also a criminologist

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

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)



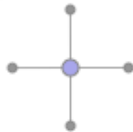
Post-punk, proto-utopian, anarchist, activist, musician, blogger, student, failed comedian.

<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

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)

ECHONEST APIS

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

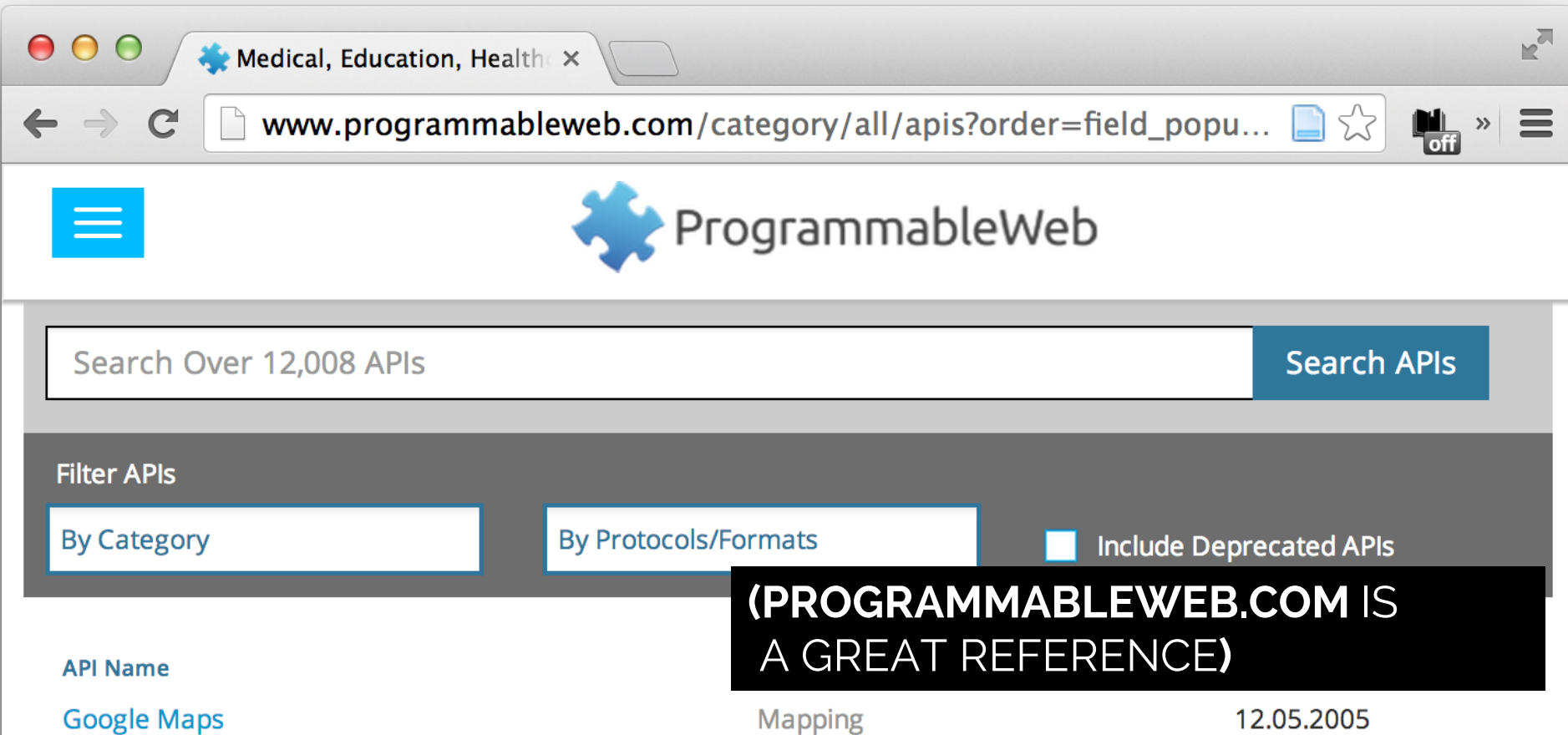
(Incredibly detailed Song, Album, Artist data for millions of musicians)

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, 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".

(PROGRAMMABLEWEB.COM IS A GREAT REFERENCE)

API Name
Google Maps
Mapping
12.05.2005

FINDING AND EXTRACTING EXISTING DATA

SCRAPING THE WEB

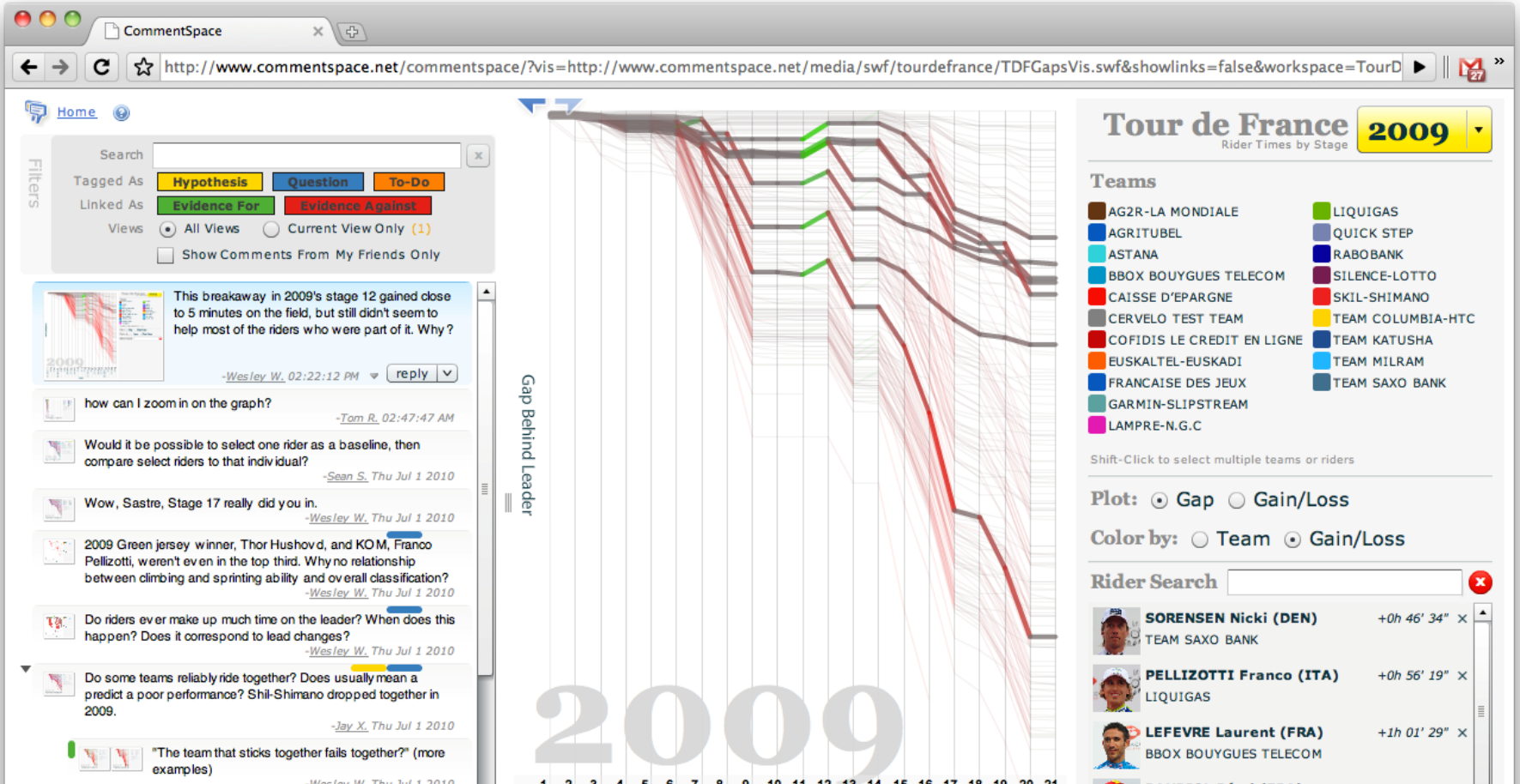
WHY SCRAPE?

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...





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

f t g+ y u

EN

Sunday July 27th, 2014

Stage 21
Évry / Paris Champs-Élysées

STAGE FINISHED

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	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



\$ % .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

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

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

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

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

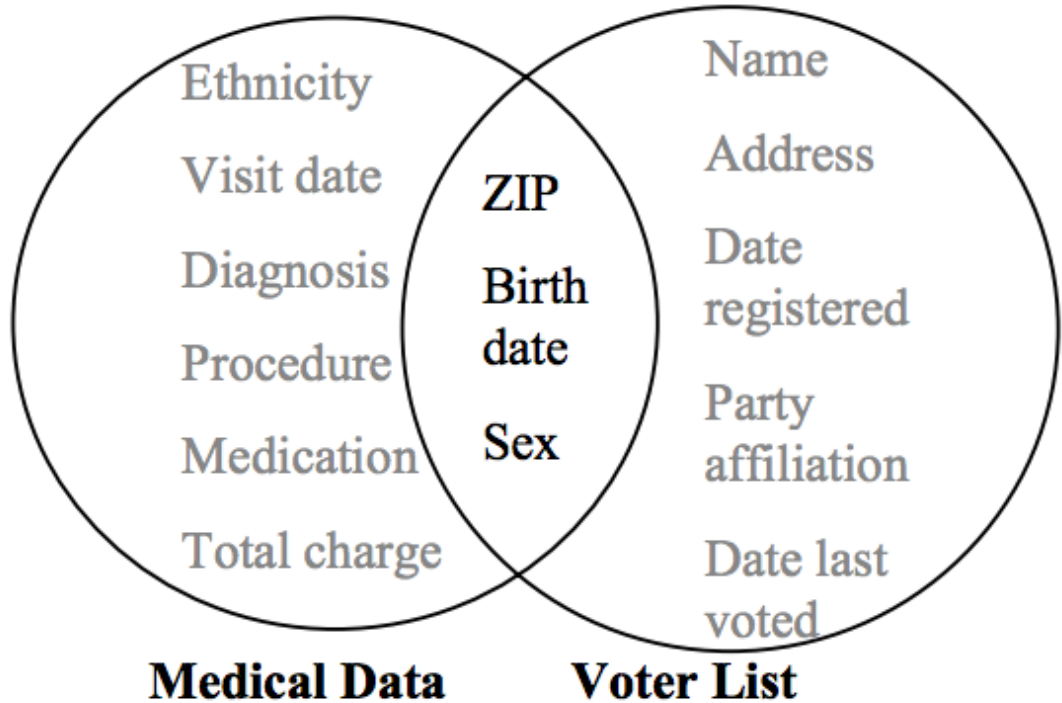
- Be sure you know who owns the data.
- Think early on about whether or not you'll need to publish or (re)share data.
- Be careful you aren't violating copyright, especially when scraping.

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

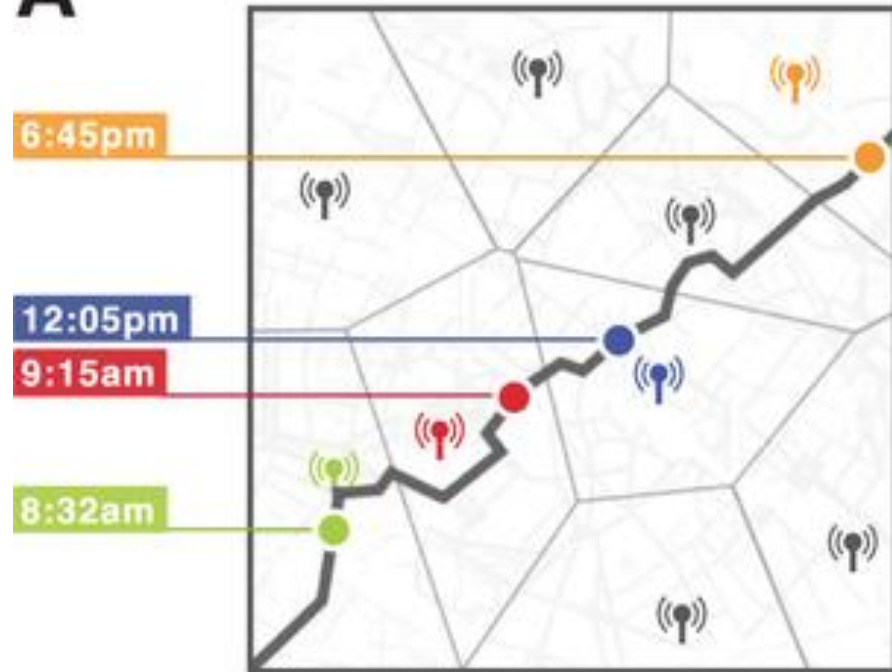


[L. Sweeney. 2002]

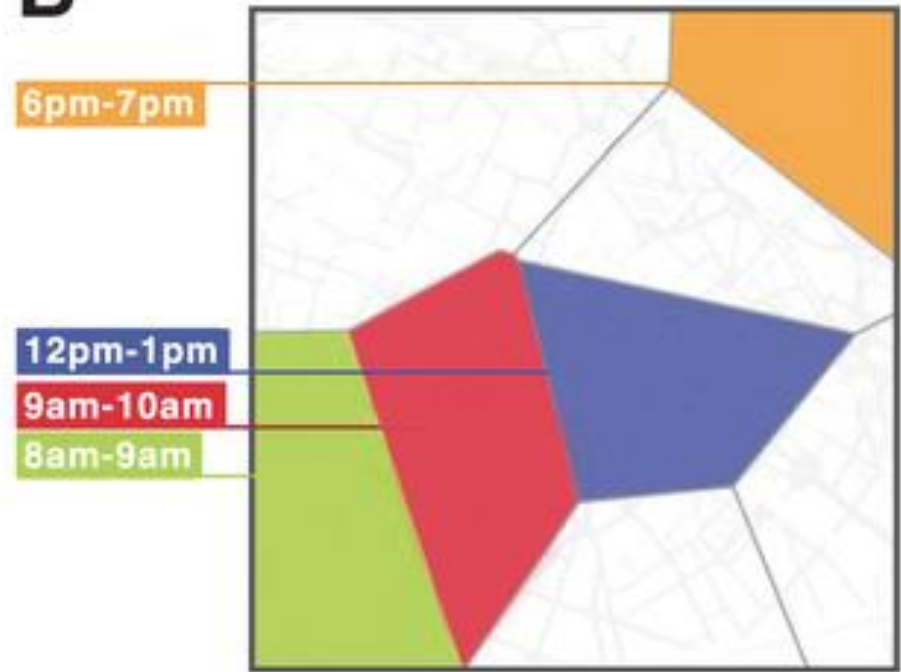
[k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY](#)

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 tracked via social networks [Fowler JH, Christakis NA (2008) *BMJ* 337:859-862]

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-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 see ...
content than one person can view ...
stories, and activities undertaken by ...
primary manner by which people see content that ...
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 ...

“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

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

(...AND WE'LL
BE HAPPY TO DISCUSS
OR SUGGEST MORE)

NEXT UP

TUTORIAL 2 – BUILDING A WEB SCRAPER

TOMORROW

DATA CLEANING

STATISTICS

BEFORE NEXT CLASS

INSTALL :



OpenRefine (formerly Google Refine)

<http://openrefine.org/>

DATA COLLECTION

Slides by WESLEY WILLETT

VISUAL ANALYTICS