

# What is data?

Jeffrey Leek

May 17, 2016

# Definition of data

Data are values of qualitative or quantitative variables, belonging to a set of items.

<http://en.wikipedia.org/wiki/Data>

# Definition of data

Data are values of qualitative or quantitative variables, belonging to a set of items.

<http://en.wikipedia.org/wiki/Data>

**Set of items:** Sometimes called the population; the set of objects you are interested in

# Definition of data

Data are values of qualitative or quantitative variables, belonging to a set of items.

<http://en.wikipedia.org/wiki/Data>

**Variables:** A measurement or characteristic of an item.

# Definition of data

Data are values of qualitative or quantitative variables, belonging to a set of items.

<http://en.wikipedia.org/wiki/Data>

**Qualitative:** Country of origin, sex, treatment

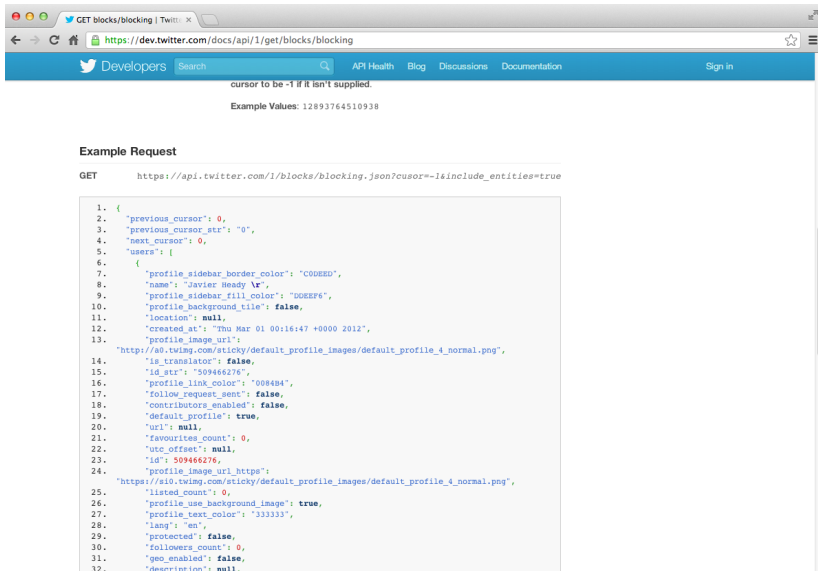
**Quantitative:** Height, weight, blood pressure

## What do data look like?

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGACGGATCTCGTATGCCGCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]az`aZM`Z]YRa]YSG[[ZREQLHESDHNDHNDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTGCCGCACGACAGGCAGCGGTTCAGCCTGCGCTTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a```\_`_````a``a`a`a`_`_]a_]`]\`a`_____`_`^^]X]_]XTV\_]]NX_XVX]]_TTTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbaababbbbbbb`bbbb`bbbbbbb`bbbaV`a``a``]``aT]a__V\]]_]`a`]a`abbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCATATTCTCCGGTGTGTGGTTTAACCGATCATCGCGCATTAATTCCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
````[aa\b`^`^][aabbb][`a`abbb`a``bbbbbababaaaab_VZa`_`__bab_X`[a\HV[_]_[^_X\T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGCTCTTCTGCTTGAAAAAAAAAACA
+HWI-EAS121:4:100:1783:207#0/1
abba`Xa\^`\`aa]ba__bba[a_O`a`aa`aa`a]^V]X_a^YS\R\_H_[ ]\ZTDUZZUSOPX]]POP\GS\WSHHD
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAATACATCTTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121:4:100:1783:455#0/1
abb`babbababbbbbbbbbbbbbbbba\b`b`abbbabbbbabbbbbbaabbbbbb`bb`ab`O`bab`Q`bbabaa`a
```

[http://brianknaus.com/software/srtoolbox/s\\_4\\_1\\_sequence80.txt](http://brianknaus.com/software/srtoolbox/s_4_1_sequence80.txt)

# What do data look like?



GET blocks/blocking | Twitter X

https://dev.twitter.com/docs/api/1/get/blocks/blocking

Developers Search API Health Blog Discussions Documentation Sign in

cursor to be -1 if it isn't supplied.

Example Values: 12893764510938

### Example Request

GET https://api.twitter.com/1/blocks/blocking.json?cursor=-1&include\_entities=true

```
1. {
2.   "previous_cursor": 0,
3.   "previous_cursor_str": "0",
4.   "next_cursor": 0,
5.   "users": [
6.     {
7.       "profile_sidebar_border_color": "C0DEED",
8.       "name": "Javier Heady \r",
9.       "profile_sidebar_fill_color": "D0E0F6",
10.      "profile_background_tile": false,
11.      "location": null,
12.      "created_at": "Thu Mar 01 00:16:47 +0000 2012",
13.      "profile_image_url":
14.        "https://a0.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
15.      "is_translator": false,
16.      "id_str": "509466276",
17.      "profile_link_color": "0084B4",
18.      "follow_request_sent": false,
19.      "contributors_enabled": false,
20.      "default_profile": true,
21.      "url": null,
22.      "favourites_count": 0,
23.      "utc_offset": null,
24.      "id": "509466276",
25.      "profile_image_url_https":
26.        "https://s10.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
27.      "listed_count": 0,
28.      "profile_use_background_image": true,
29.      "profile_text_color": "333333",
30.      "lang": "en",
31.      "protected": false,
32.      "followers_count": 0,
33.      "geo_enabled": false,
34.      "description": null,
```

https:

//dev.twitter.com/docs/api/1/get/blocks/blocking

# What do data look like?

ALLERGIES		MEDICATION HISTORY	
Last Updated: 01 Dec 2011 @ 0851		Last Updated: 11 Apr 2011 @ 1737	
Allergy Name:	TRIMETHOPRIM	Medication:	AMLODIPINE BESYLATE 10MG TAB
Location:	DAYT29	Instructions:	TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TABLET FOR GRAPEFRUIT JUICE--
Date Entered:	09 Mar 2011	Status:	Active
Reaction:		Refills Remaining:	3
Allergy Type:	DRUG	Last Filled On:	20 Aug 2010
Drug Class:	ANTI-INFECTIVES, OTHER	Initially Ordered On:	13 Aug 2010
Observed/Historical:	HISTORICAL	Quantity:	45
Comments:	The reaction to this allergy was MILD (NO SQUELAE)	Days Supply:	90
		Pharmacy:	DAYTON
		Prescription Number:	2718953
Allergy Name:	TRAMADOL	Medication:	IBUPROFEN 600MG TAB
Location:	DAYT29	Instructions:	TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WITH FOOD
Date Entered:	09 Mar 2011	Status:	Active
Reaction:	URINARY RETENTION	Refills Remaining:	3
Allergy Type:	DRUG	Last Filled On:	20 Aug 2010
Drug Class:	NON-OPIOD ANALGESICS	Initially Ordered On:	01 Jul 2010
Observed/Historical:	HISTORICAL	Quantity:	300
Comments:	gradually worsening difficulty emptying bladder		

<http://blue-button.github.com/challenge/>

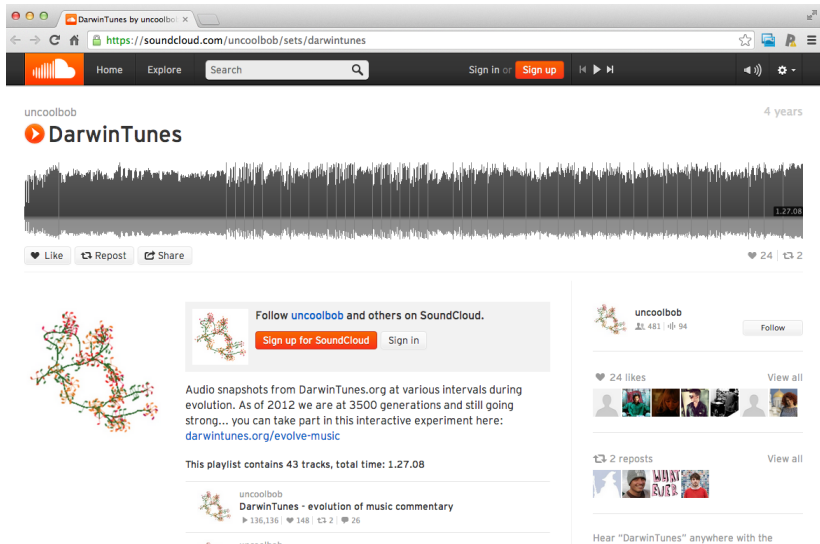


# What do data look like?



[http://www.nytimes.com/2012/06/26/technology/  
in-a-big-network-of-computers-evidence-of-machine-learning  
html?pagewanted=all&\\_r=0](http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html?pagewanted=all&_r=0)

# What do data look like?



The screenshot shows a web browser window displaying a SoundCloud page for a user named 'uncoolbob'. The browser's address bar shows the URL <https://soundcloud.com/uncoolbob/sets/darwintunes>. The page header includes navigation links for 'Home', 'Explore', and a search bar, along with a 'Sign in or Sign up' button. The main content area features the 'uncoolbob' profile name, a play button icon, and the title 'DarwinTunes'. Below this is a large audio waveform representing the entire set, with a duration of 1:27:08. Under the waveform are buttons for 'Like', 'Repost', and 'Share', and a summary of 24 likes and 2 reposts. A promotional banner encourages following 'uncoolbob' and others on SoundCloud, with a 'Sign up for SoundCloud' button. To the left of the banner is a colorful, abstract image of a branching structure with red and orange nodes. The text below the banner describes 'DarwinTunes.org' as an interactive experiment in the evolution of music, mentioning 3500 generations as of 2012 and providing a link to [darwinTunes.org/evolve-music](http://darwinTunes.org/evolve-music). It also states that the playlist contains 43 tracks with a total time of 1:27:08. Below this is a section for 'uncoolbob's 'DarwinTunes - evolution of music commentary', showing 136,136 plays, 148 likes, and 26 reposts. On the right side of the page, there is a user profile for 'uncoolbob' with 481 followers and 94 following, a 'Follow' button, and a list of 24 likes and 2 reposts, each with a small profile picture. At the bottom of the page, a partial sentence reads 'Hear "DarwinTunes" anywhere with the'.

<http://www.pnas.org/content/109/30/12081.full>

<https://soundcloud.com/uncoolbob/sets/darwintunes>

# What do data look like?

The screenshot shows the Data.gov website in a web browser. The browser's address bar displays "www.data.gov". The website's header includes the Data.gov logo with the tagline "EMPOWERING PEOPLE", a search bar for the "Data Catalog", and links for "Open ID" and "Login". A navigation menu contains links for HOME, ABOUT, DATA, METRICS, OPEN GOVERNMENT, BLOGS, and COMMUNITIES. The main content area features a large map titled "SANDY DAMAGE ESTIMATES BY BLOCK GROUP" showing a coastal area with red and orange shaded regions. To the right of the map is a "Latest Datasets" section with a list of data items. Below the main content are three smaller sections: "DATA AND TOOLS", "COMMUNITIES", and "OPEN GOVERNMENT", each with a representative image.

DATA.GOV  
EMPOWERING PEOPLE

Search Data Catalog

Search site content

Open ID Login

HOME ABOUT DATA METRICS OPEN GOVERNMENT BLOGS COMMUNITIES

**SANDY DAMAGE ESTIMATES BY BLOCK GROUP**

U.S. DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT

**Latest Datasets**

- Mississippi River Centerline - Headwa...
- 1997 Red River of the North Flood Bou...
- USACE Habitat Restoration and Enhance...
- 2007 Sumpter Powder River Mine Lidar
- Hyperspectral Imagery for the Main EI...
- 2003 Southwest Florida Water Manageme...
- 2011 U.S. Army Corps of Engineers (US...
- 2011 U.S. Army Corps of Engineers (US...
- Harding County 2010 Census Voting Dis...
- 2011 Atlantic Pleasure Craft and Sail...

**DATA AND TOOLS**

**COMMUNITIES**

**OPEN GOVERNMENT**

<http://www.data.gov/>

## What do data look like? Rarely

</

# The data is the second most important thing

- ▶ The most important thing in data science is the question
- ▶ The second most important is the data
- ▶ Often the data will limit or enable the questions
- ▶ But having data can't save you if you don't have a question