



What is data?

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

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

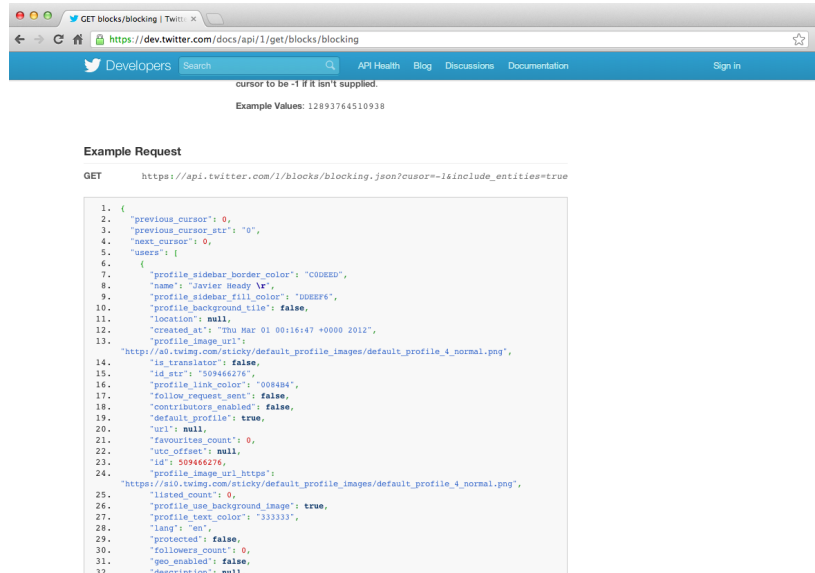
Quantitative: Height, weight, blood pressure

What do data look like?

```
@HWI-EAS121.4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGCTCGTGCCTGACAAGACAGGGG
+HWI-EAS121.4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX`aZ`aZM`Z`YRa]YSG[ [ZREQLHESHDNDHNDHMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121.4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTCGCCGACGACAGGCAGCGGTTCAGCCTGCGCTTTGGCCTTGGCCTTCGGAAA
+HWI-EAS121.4:100:1783:1611#0/1
a^^^\\`^^^a^^`a^_`a_]`^a_____`_^^`X_]XTV_\\]NX_XVX]]_TTTTG[VTHPN]VFDZ
@HWI-EAS121.4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTCTTA
+HWI-EAS121.4:100:1783:322#0/1
abaa`^aaaaabbbbaabbbbbb`bbbb`bbbbbbb`bbbaV`a``a``]``aT]a_V\\]]`^a`]a_abbaV__
@HWI-EAS121.4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCATATTCTCCGGTGTGTGGTTTAACCGATCATCGCGCATTACTTCCCGGCTGC
+HWI-EAS121.4:100:1783:1394#0/1
``^[aa/b`^`[aabbb][`a`abbb`a``bbbbabababaaab_VZa`^__bab_X`[a`HV_[_]`^_X`T_VQO
@HWI-EAS121.4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGCTTCTGCTTGAAAAAATACAT
+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_[ ]ZTDUZUSOPX]]POP\\GS\\WSHHD
@HWI-EAS121.4:100:1783:455#0/1
GGGTAAATCAGGGACAATGTAATGGCTGCACAAAAAATACATCTTTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121.4:100:1783:455#0/1
abb`babbabababbbbbb`bbbbbb`b`b`\\abbbabbbbabbbbbb`bb`ab`O`bab`Q`bbabaa`
```

http://brianknaus.com/software/srtoolbox/s_4_1_sequence80.txt

What do data look like?



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 Jr",
9.       "profile_sidebar_fill_color": "DDEEFF",
10.      "profile_background_title": 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.      "favorites_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-OPIOID ANALGESICS	Initially Ordered On:	01 Jul 2010
Observed/Historical:	HISTORICAL	Quantity:	300
Comments:	gradually worsening difficulty emptying bladder		

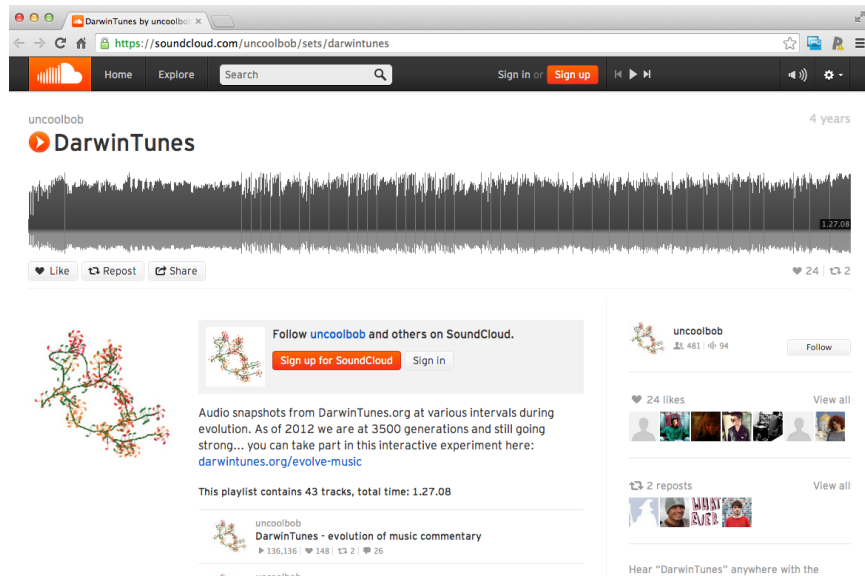
<http://blue-button.github.com/challenge/>

What do data look like?



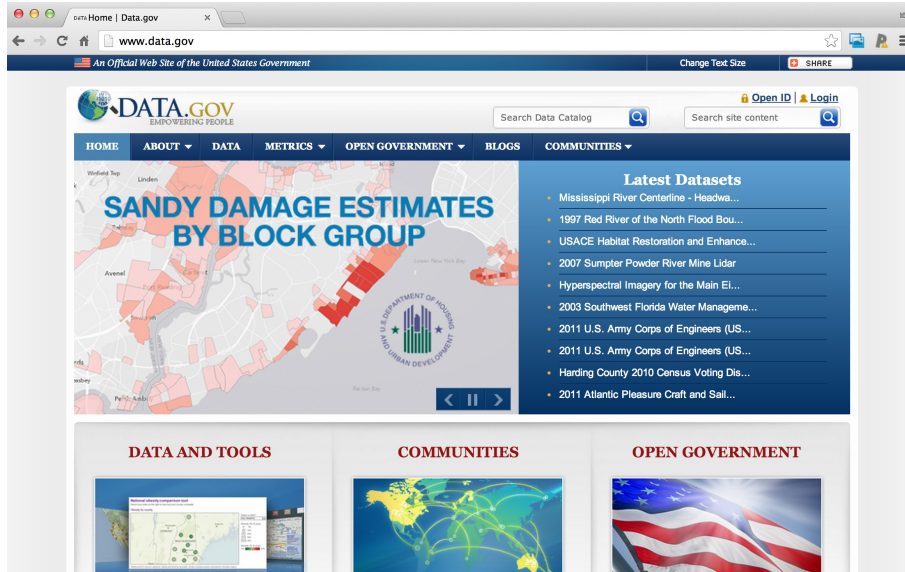
http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html?pagewanted=all&_r=0

What do data look like?



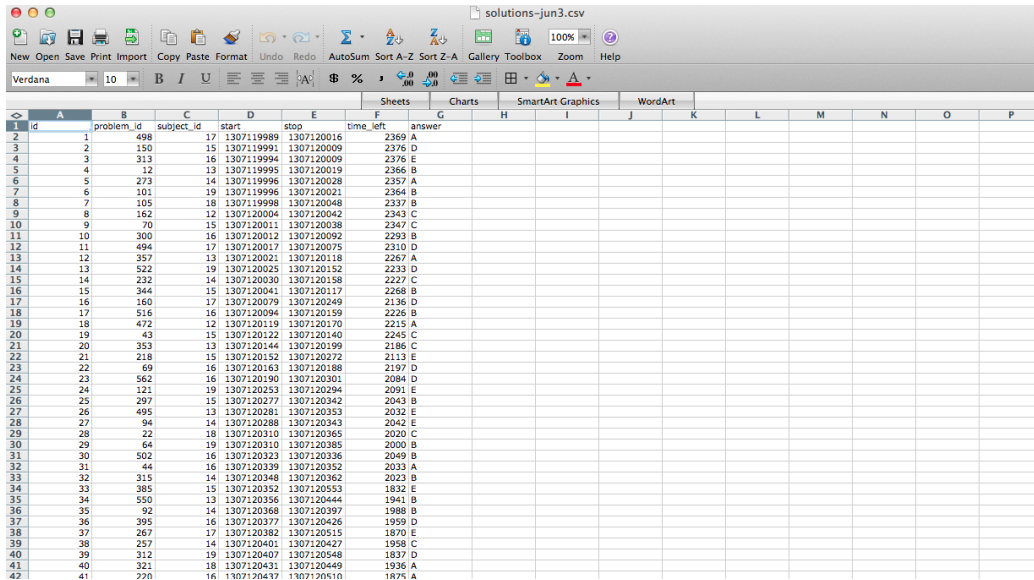
<http://www.pnas.org/content/109/30/12081.full> <https://soundcloud.com/uncoolbob/sets/darwintunes>

What do data look like?



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

What do data look like? Rarely



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	id	problem_id	subject_id	start	stop	time_left	answer									
2	1	498	17	1307119989	1307120016	2369	A									
3	2	150	15	1307119991	1307120009	2376	D									
4	3	313	16	1307119994	1307120009	2376	E									
5	4	12	13	1307119995	1307120019	2366	B									
6	5	273	14	1307119996	1307120028	2357	A									
7	6	101	19	1307119996	1307120021	2364	B									
8	7	105	18	1307119998	1307120048	2337	B									
9	8	162	12	1307120004	1307120042	2343	C									
10	9	70	15	1307120011	1307120038	2347	C									
11	10	300	16	1307120012	1307120092	2293	B									
12	11	494	17	1307120017	1307120075	2310	D									
13	12	357	13	1307120021	1307120118	2267	A									
14	13	522	19	1307120025	1307120152	2233	D									
15	14	232	14	1307120030	1307120158	2227	C									
16	15	344	15	1307120041	1307120117	2268	B									
17	16	160	17	1307120079	1307120249	2136	D									
18	17	516	16	1307120094	1307120159	2256	B									
19	18	472	12	1307120119	1307120170	2215	A									
20	19	43	15	1307120122	1307120140	2245	C									
21	20	353	13	1307120144	1307120199	2186	C									
22	21	218	15	1307120152	1307120272	2113	E									
23	22	69	16	1307120163	1307120188	2197	D									
24	23	562	16	1307120190	1307120301	2084	D									
25	24	121	19	1307120253	1307120294	2091	E									
26	25	297	15	1307120277	1307120342	2043	B									
27	26	495	13	1307120281	1307120353	2032	E									
28	27	94	14	1307120288	1307120343	2042	E									
29	28	22	18	1307120310	1307120365	2020	C									
30	29	64	19	1307120310	1307120385	2000	B									
31	30	502	16	1307120323	1307120336	2049	B									
32	31	44	16	1307120339	1307120352	2033	A									
33	32	315	14	1307120348	1307120362	2023	B									
34	33	385	15	1307120352	1307120553	1832	E									
35	34	550	13	1307120356	1307120444	1941	B									
36	35	92	14	1307120368	1307120397	1988	B									
37	36	395	16	1307120377	1307120426	1959	D									
38	37	267	17	1307120382	1307120515	1870	E									
39	38	257	14	1307120401	1307120427	1958	C									
40	39	312	19	1307120407	1307120548	1837	D									
41	40	321	18	1307120431	1307120449	1936	A									
42	41	220	16	1307120437	1307120510	1876	A									

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