What is data?

Jeffrey Leek, Assistant Professor of Biostatistics
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

Raw versus processed data

Raw data

- The original source of the data
- Often hard to use for data analyses
- Data analysis *includes* processing
- Raw data may only need to be processed once

http://en.wikipedia.org/wiki/Raw_data

Processed data

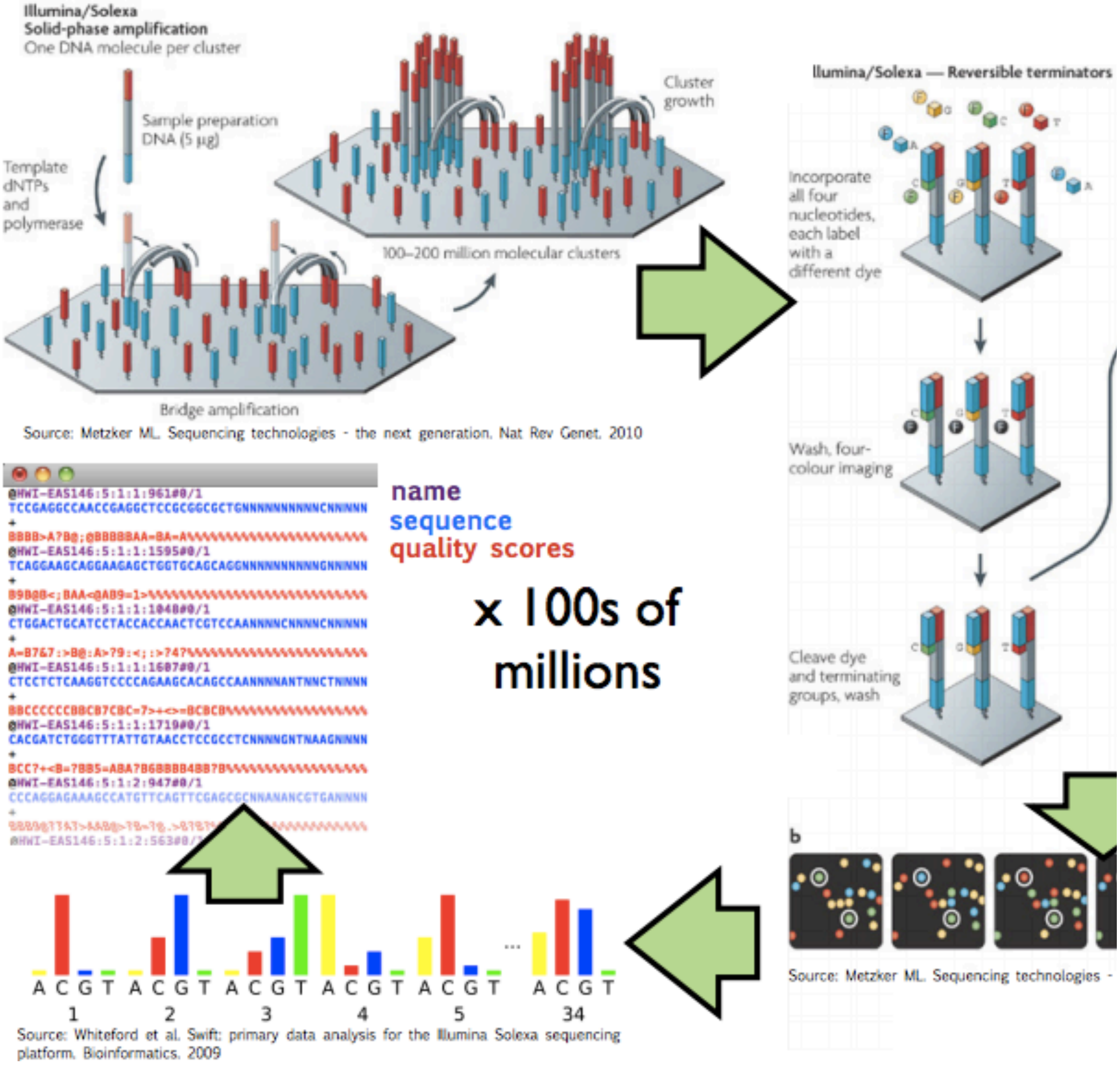
- Data that is ready for analysis
- Processing can include merging, subsetting, transforming, etc.
- There may be standards for processing
- All steps should be recorded

http://en.wikipedia.org/wiki/Computer_data_processing

An example of a processing pipeline

http://www.illumina.com.cn/support/sequencing/sequencing_instruments/hiseq_1000.asp

An example of a processing pipeline


Illumina/Solexa
Solid-phase amplification
One DNA molecule per cluster
Sample preparation
DNA (5 μg)
Template
dNTPs
and
polymerase
Cluster
growth
100–200 million molecular clusters
Bridge amplification
llumina/Solexa — Reversible terminators
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
b
A C G T A C G T A C G T A C G T A C G T ... A C G T
1 2 3 4 5 34

Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

```
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNNCNNNNN
+
BBBB>A?B@;@BBBBBAA=BA=A%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGGNNNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNCNNNNCNNNNN
+
A=B7&7:>B@:A>?9:<;:>?4?%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCCAGAAGCACAGCCAANNNNANTNNCTNNNN
+
BBCCCCCCBBCB7CBC=7>+<>=BCBCB%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCCTCNNNNGNTNAAGNNNN
+
BCC7+<B=7BB5=ABA7B6BBBB4BB7B%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAGAAAGCCATGTTCAGTTCGAGCGCNNANANCGTGANNNN
+
BBB@?7%7>AAB@>?B=?@.>B?B?%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:563#0/
```

name
sequence
quality scores

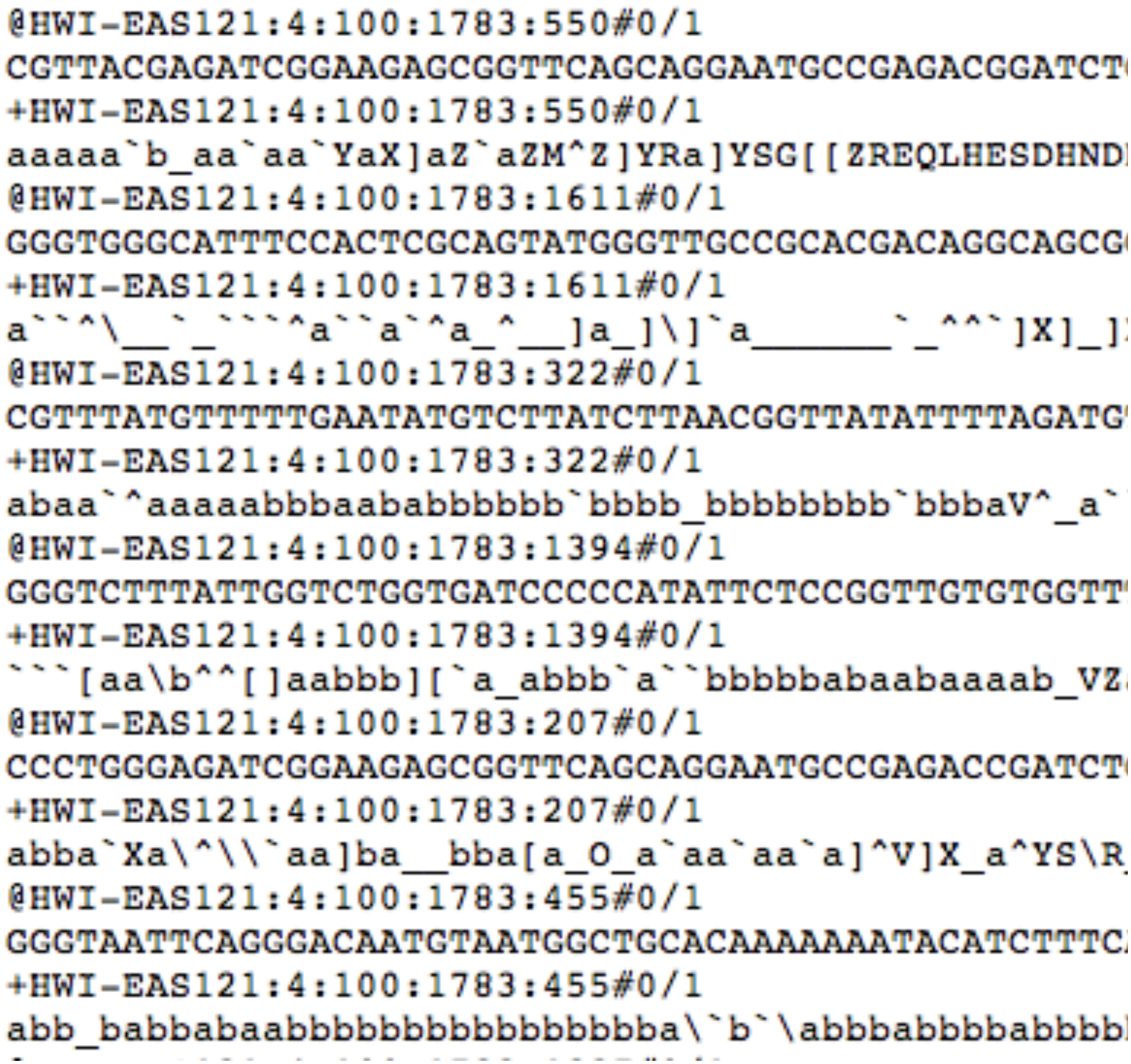
x 100s of millions

Source: Metzker ML. Sequencing technologies -

Source: Whiteford et al. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics. 2009

http://www.cbcb.umd.edu/~hcorrada/CMSC858B/lectures/lect22_seqIntro/seqIntro.pdf

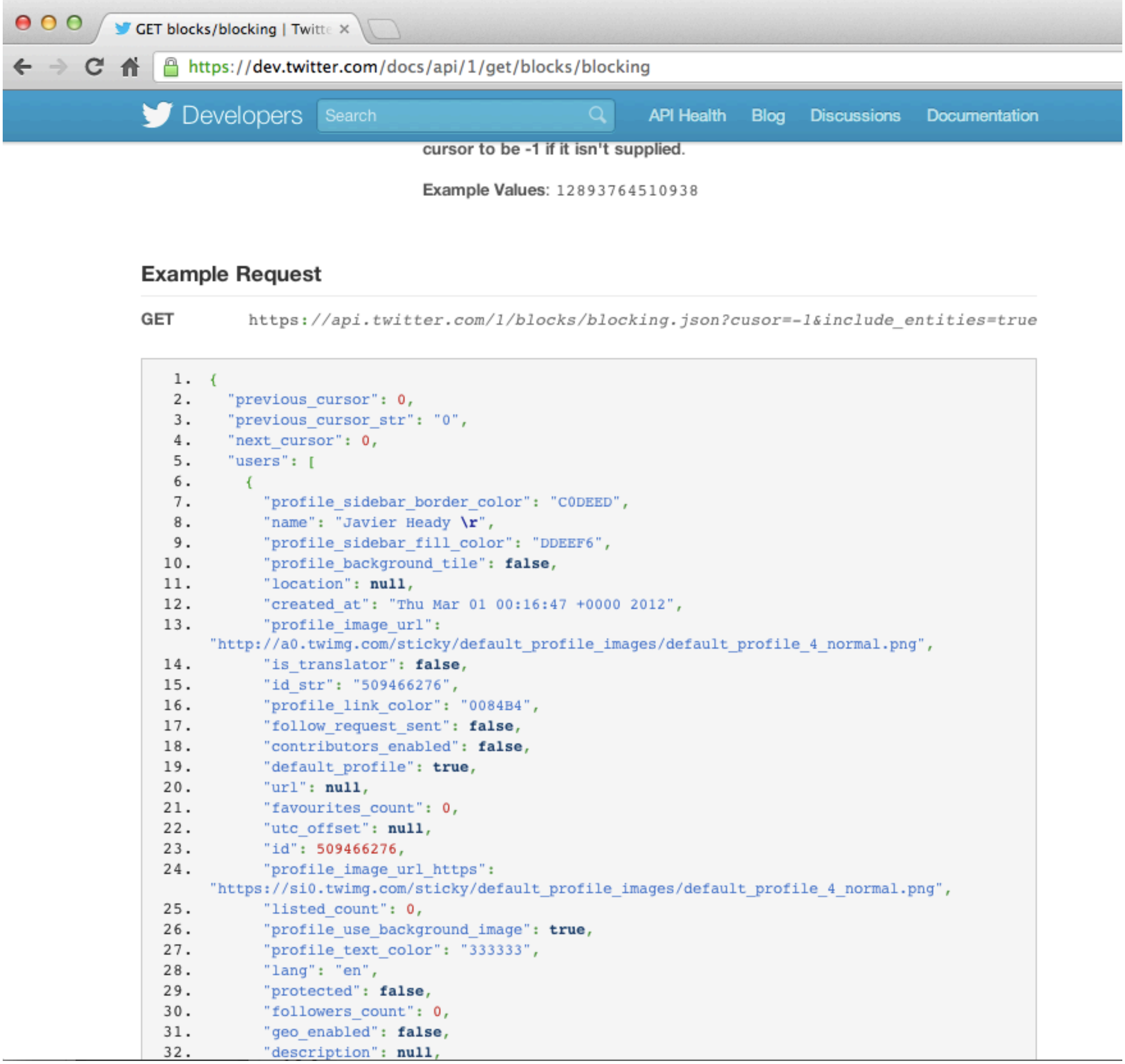
What do raw data look like?



~~~
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCT
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHND
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCG
+HWI-EAS121:4:100:1783:1611#0/1
a``^\__`_```^a``a`^a_^__]a_]\]`a________`_^^`]X]_]
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATG
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbbaababbbbbb`bbbb_bbbbbbbb`bbbaV^_a`
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTT
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZ
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT
+HWI-EAS121:4:100:1783:207#0/1
abba`Xa\^\\`aa]ba__bba[a_O_a`aa`aa`a]^V]X_a^YS\R
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAAATACATCTTTC
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbbbbbbbba\`b`\abbbabbbbabbbb
~~~

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

What do raw data look like?


GET blocks/blocking | Twitte ×
https://dev.twitter.com/docs/api/1/get/blocks/blocking
Developers
Search
API Health
Blog
Discussions
Documentation

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

Example Values: `12893764510938`

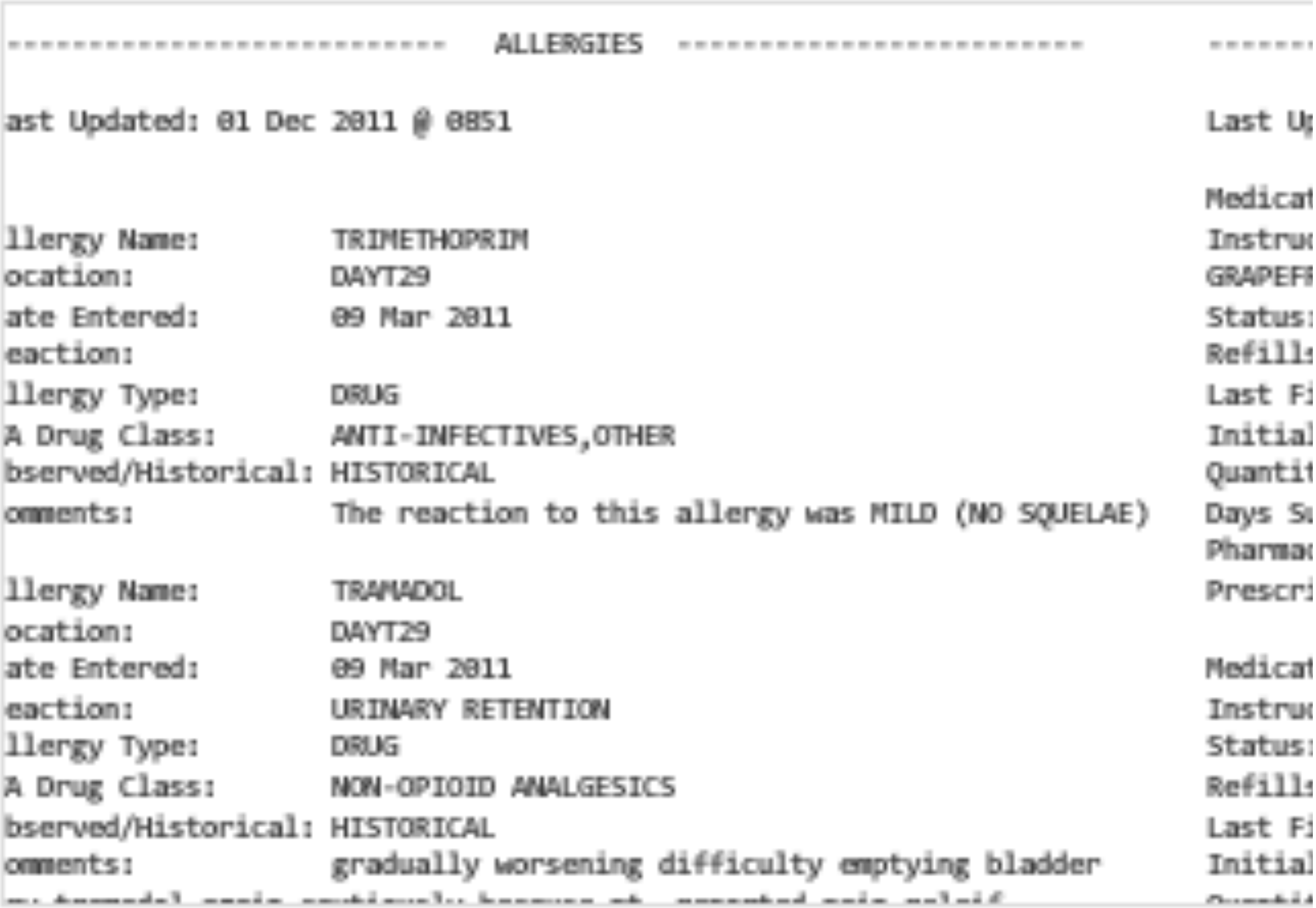
Example Request

GET `https://api.twitter.com/1/blocks/blocking.json?cusor=-1&include_entities=true`

```
{
  "previous_cursor": 0,
  "previous_cursor_str": "0",
  "next_cursor": 0,
  "users": [
    {
      "profile_sidebar_border_color": "C0DEED",
      "name": "Javier Heady \r",
      "profile_sidebar_fill_color": "DDEEF6",
      "profile_background_tile": false,
      "location": null,
      "created_at": "Thu Mar 01 00:16:47 +0000 2012",
      "profile_image_url":
"http://a0.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
      "is_translator": false,
      "id_str": "509466276",
      "profile_link_color": "0084B4",
      "follow_request_sent": false,
      "contributors_enabled": false,
      "default_profile": true,
      "url": null,
      "favourites_count": 0,
      "utc_offset": null,
      "id": 509466276,
      "profile_image_url_https":
"https://si0.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
      "listed_count": 0,
      "profile_use_background_image": true,
      "profile_text_color": "333333",
      "lang": "en",
      "protected": false,
      "followers_count": 0,
      "geo_enabled": false,
      "description": null,
```

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

What do raw data look like?



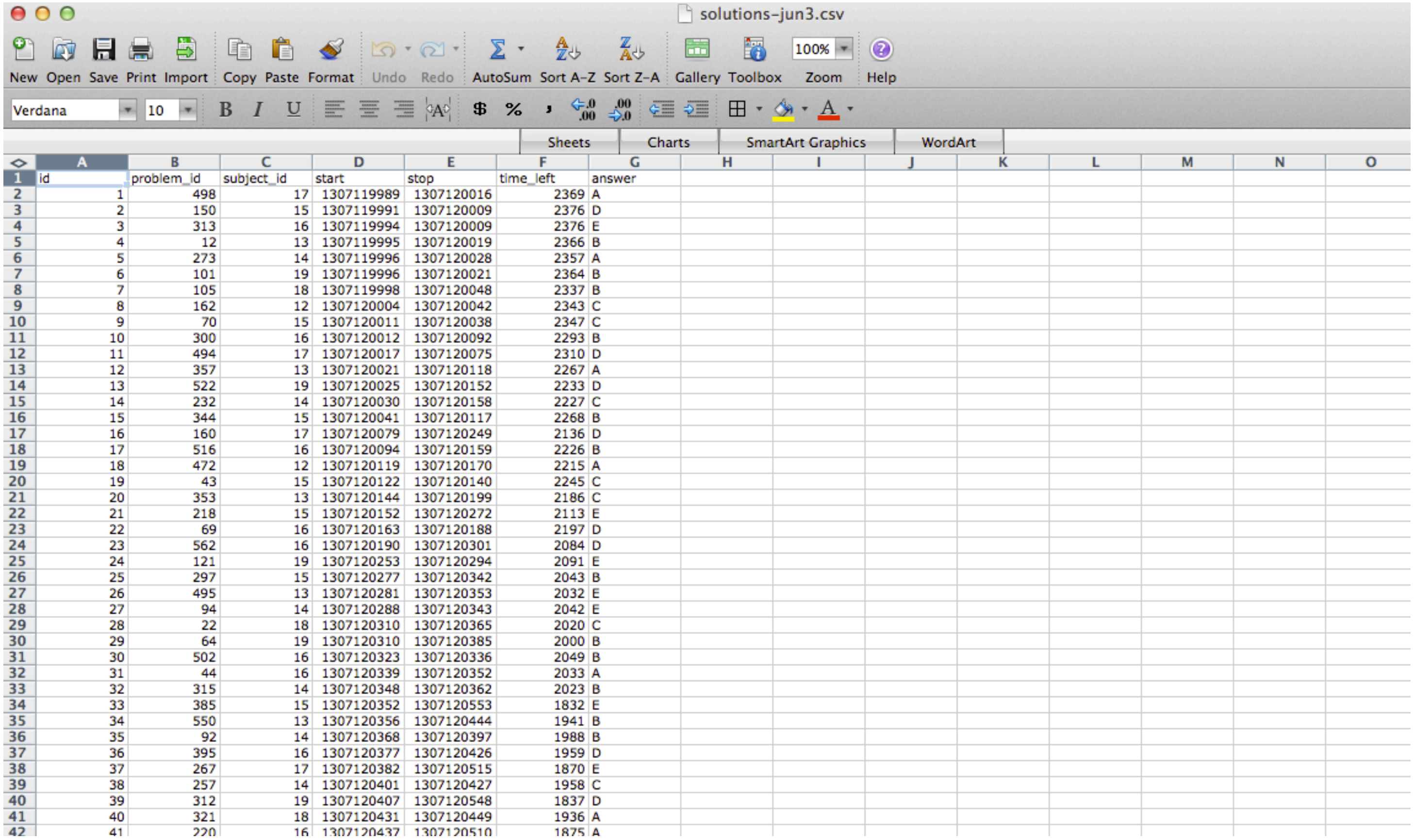
```
---------------------------  ALLERGIES  -------------------------        -------

ast Updated: 01 Dec 2011 @ 0851                                          Last Up

                                                                         Medicat
llergy Name:        TRIMETHOPRIM                                         Instruc
ocation:            DAYT29                                               GRAPEFR
ate Entered:        09 Mar 2011                                          Status:
eaction:                                                                 Refills
llergy Type:        DRUG                                                 Last Fi
A Drug Class:       ANTI-INFECTIVES,OTHER                                Initial
bserved/Historical: HISTORICAL                                           Quantit
omments:            The reaction to this allergy was MILD (NO SQUELAE)   Days Su
                                                                         Pharmac
llergy Name:        TRAMADOL                                             Prescri
ocation:            DAYT29
ate Entered:        09 Mar 2011                                          Medicat
eaction:            URINARY RETENTION                                    Instruc
llergy Type:        DRUG                                                 Status:
A Drug Class:       NON-OPIOID ANALGESICS                                Refills
bserved/Historical: HISTORICAL                                           Last Fi
omments:            gradually worsening difficulty emptying bladder      Initial
```

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

What do processed data look like?



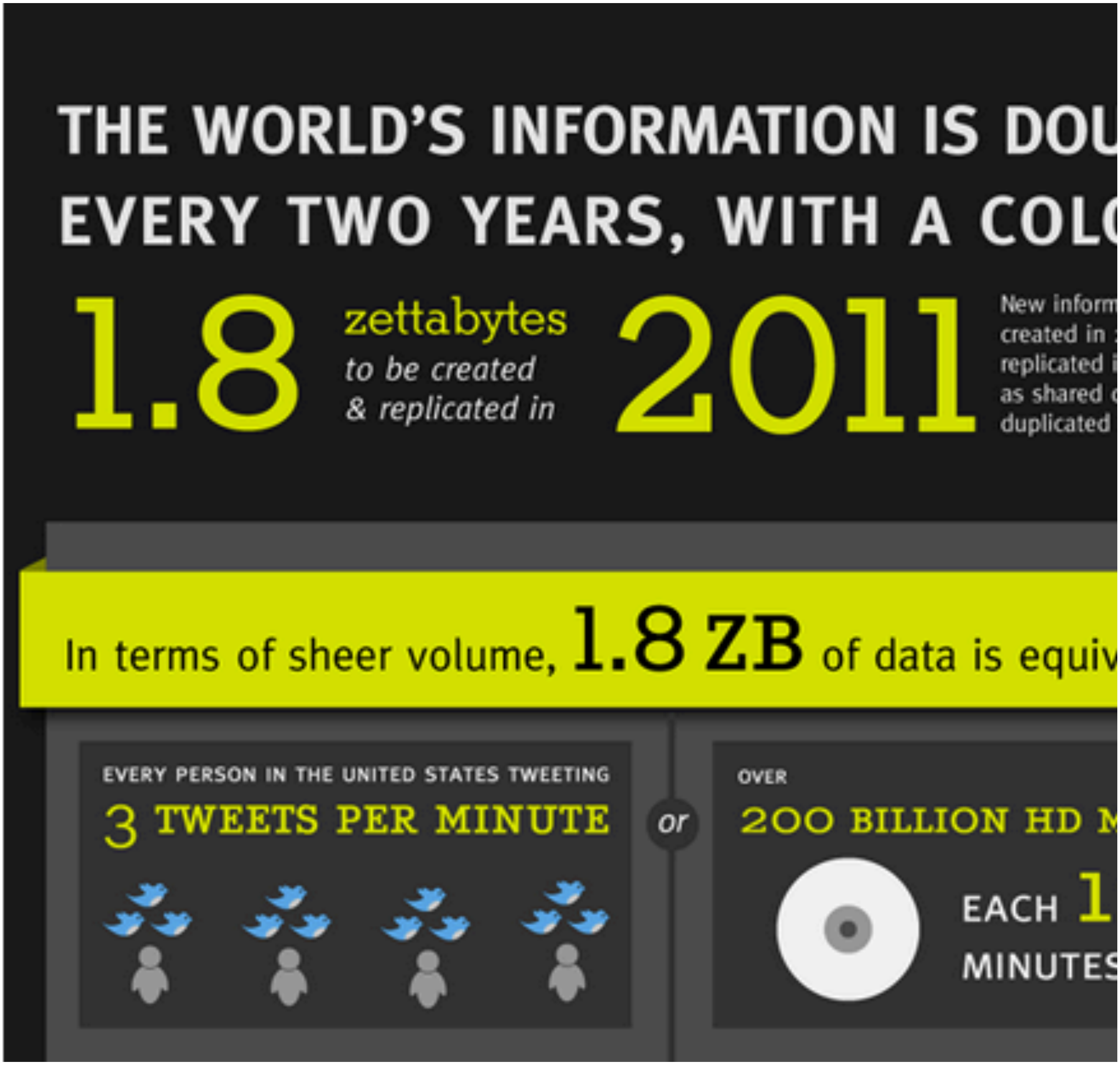
	A	B	C	D	E	F	G
1	id	problem_id	subject_id	start	stop	time_left	answer
2	1	498	17	1307119989	1307120016	2369	A
3	2	150	15	1307119991	1307120009	2376	D
4	3	313	16	1307119994	1307120009	2376	E
5	4	12	13	1307119995	1307120019	2366	B
6	5	273	14	1307119996	1307120028	2357	A
7	6	101	19	1307119996	1307120021	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	2226	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	1875	A

1. Each variable forms a column
2. Each observation forms a row
3. Each table/file stores data about one kind of observation (e.g. people/hospitals).

http://vita.had.co.nz/papers/tidy-data.pdf

Leek, Taub, and Pineda 2011 PLoS One

How much is there?



THE WORLD'S INFORMATION IS DOU
EVERY TWO YEARS, WITH A COLO

1.8 zettabytes to be created & replicated in 2011

New inform
created in
replicated i
as shared o
duplicated

In terms of sheer volume, 1.8 ZB of data is equiv

EVERY PERSON IN THE UNITED STATES TWEETING 3 TWEETS PER MINUTE

OVER 200 BILLION HD M

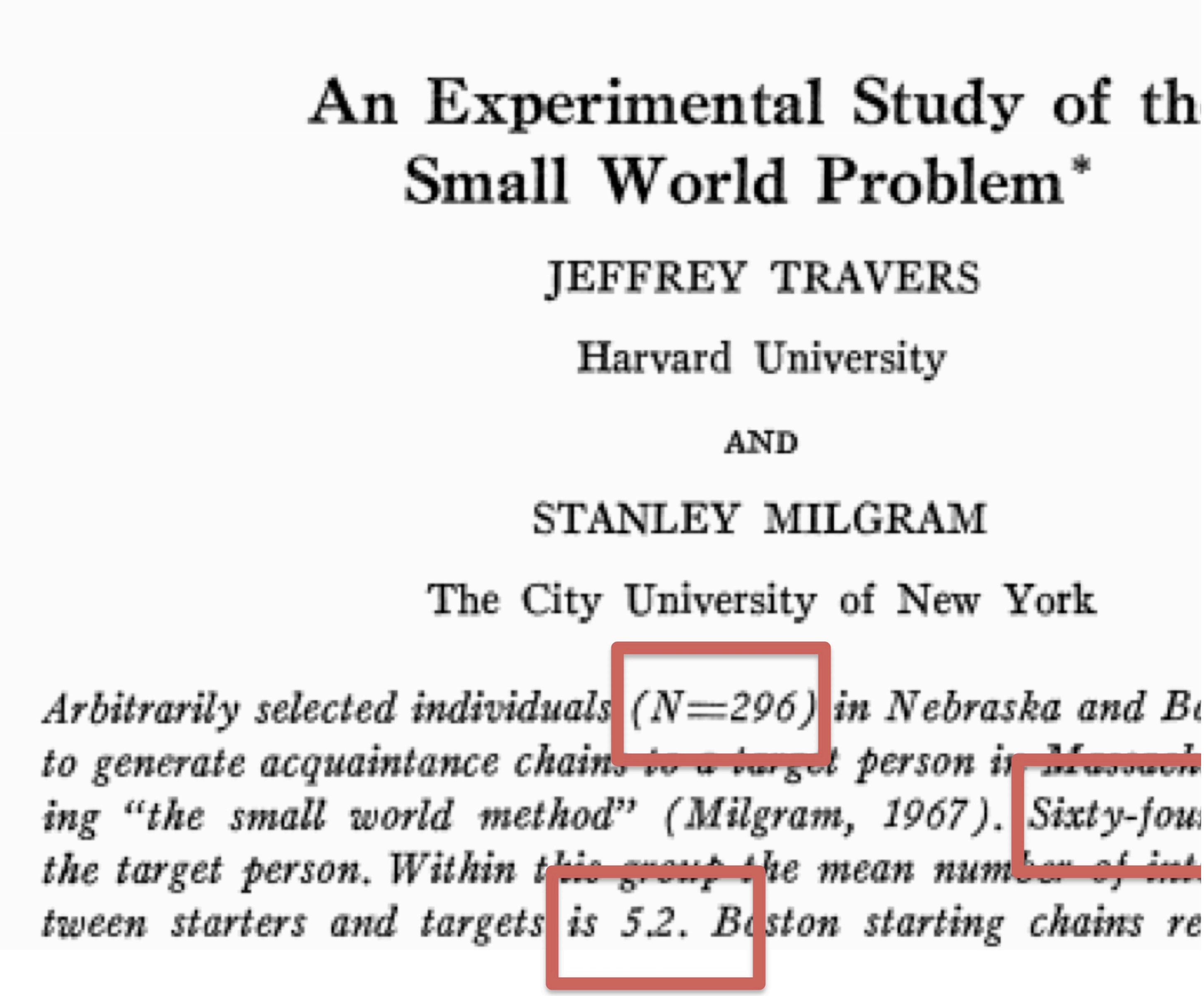
EACH 1
MINUTES

http://mashable.com/2011/06/28/data-infographic/

So what about big data?

Depends on your perspective

Why big data now?



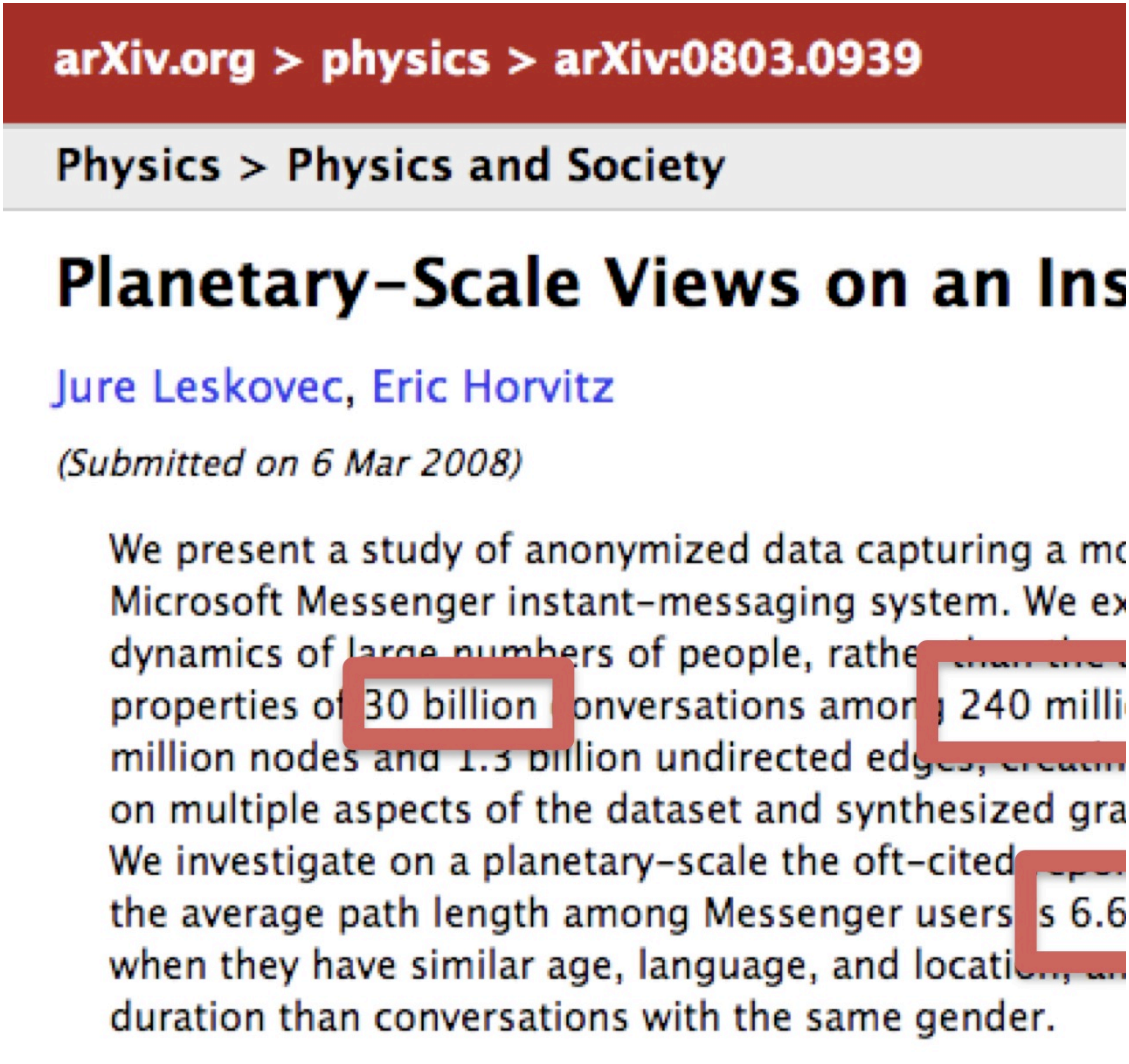
An Experimental Study of th
Small World Problem*

JEFFREY TRAVERS
Harvard University
AND
STANLEY MILGRAM
The City University of New York

*Arbitrarily selected individuals (N=296) in Nebraska and Bo
to generate acquaintance chains to a target person in Massach
ing "the small world method" (Milgram, 1967). Sixty-fou
the target person. Within this group the mean number of inte
tween starters and targets is 5.2. Boston starting chains re*

Travers and Milgram (1969) Sociometry

Why big data now?



arXiv.org > physics > arXiv:0803.0939

Physics > Physics and Society

Planetary-Scale Views on an Ins

Jure Leskovec, Eric Horvitz

(Submitted on 6 Mar 2008)

We present a study of anonymized data capturing a mo
Microsoft Messenger instant-messaging system. We ex
dynamics of large numbers of people, rather than the
properties of 30 billion conversations among 240 milli
million nodes and 1.3 billion undirected edges, creatin
on multiple aspects of the dataset and synthesized gra
We investigate on a planetary-scale the oft-cited
the average path length among Messenger users is 6.6
when they have similar age, language, and location, an
duration than conversations with the same gender.

Leskovec and Horvitz WWW '08

Big or small - you need the right data

> “The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data...”

Tukey

Big or small - you need the right data

> “ ...no matter how big the data are. ”

Leek