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

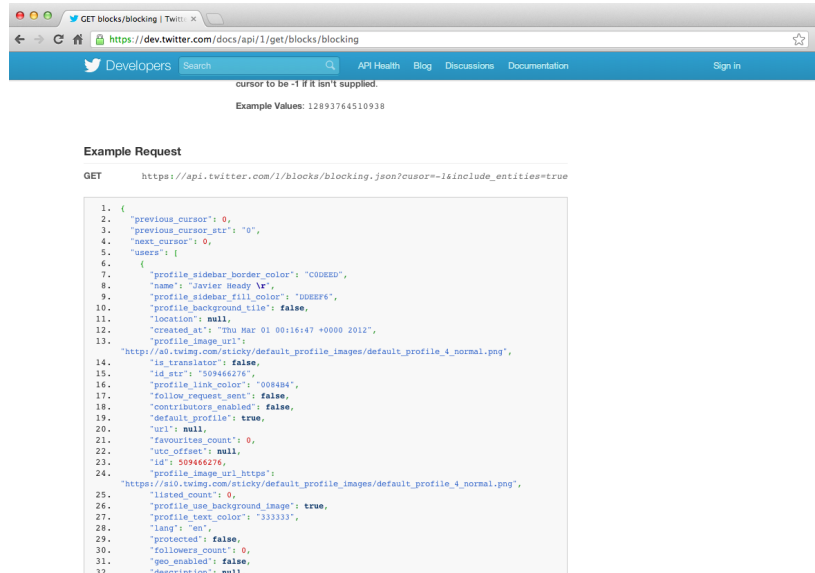**Qualitative**: Country of origin, sex, treatment   unordered

Quantitative: Height, weight, blood pressure   usually on a continuous scale, ordered

What do data look like?

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a``^\__`_```^a``a`^a_^__]a_]\]`a_____`_^^`]X]_]XTV_\]]NX_XVX]]_TTTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbbaababbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V\]]_]^a`]a_abbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTACTTCCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZa_^___bab_X`[a\HV_[_]_[^_X\T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAACA
+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
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAAATACATCTTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
```

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

# What do data look like?



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

# What do data look like?



```
------------------------------ ALLERGIES ------------------------------        ------------------------------ MEDICATION HISTORY ------------------------------
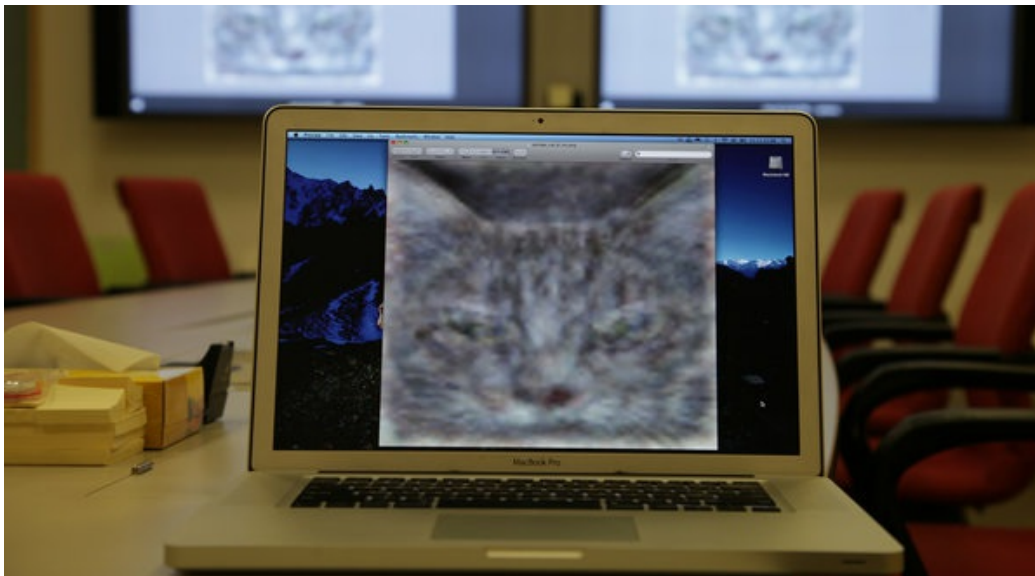
ast Updated: 01 Dec 2011 @ 0851                                                Last Updated: 11 Apr 2011 @ 1737

                                                                               Medication: AMLODIPINE BESYLATE 10MG TAB
llergy Name:        TRIMETHOPRIM                                               Instructions: TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TABLET FOR
ocation:            DAYT29                                                     GRAPEFRUIT JUICE--
ate Entered:        09 Mar 2011                                                Status: Active
eaction:                                                                       Refills Remaining: 3
llergy Type:        DRUG                                                       Last Filled On: 20 Aug 2010
A Drug Class:       ANTI-INFECTIVES,OTHER                                      Initially Ordered On: 13 Aug 2010
bserved/Historical: HISTORICAL                                                 Quantity: 45
omments:            The reaction to this allergy was MILD (NO SQUELAE)         Days Supply: 90
                                                                               Pharmacy: DAYTON
llergy Name:        TRAMADOL                                                   Prescription Number: 2718953
ocation:            DAYT29
ate Entered:        09 Mar 2011                                                Medication: IBUPROFEN 600MG TAB
eaction:            URINARY RETENTION                                          Instructions: TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WITH FOO
llergy Type:        DRUG                                                       Status: Active
A Drug Class:       NON-OPIOID ANALGESICS                                      Refills Remaining: 3
bserved/Historical: HISTORICAL                                                 Last Filled On: 20 Aug 2010
omments:            gradually worsening difficulty emptying bladder            Initially Ordered On: 01 Jul 2010
```

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?

What do data look like?

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