



# Motivation and pre-requisites

**Jeffrey Leek**  
Johns Hopkins Bloomberg School of Public Health

## About this course

- This course covers the basic ideas behind getting data ready for analysis
  - Finding and extracting raw data
  - Tidy data principles and how to make data tidy
  - Practical implementation through a range of R packages
- What this course depends on
  - The Data Scientist's Toolbox
  - R Programming
- What would be useful
  - Exploratory analysis
  - Reporting Data and Reproducible Research

~ might want to plot during cleaning

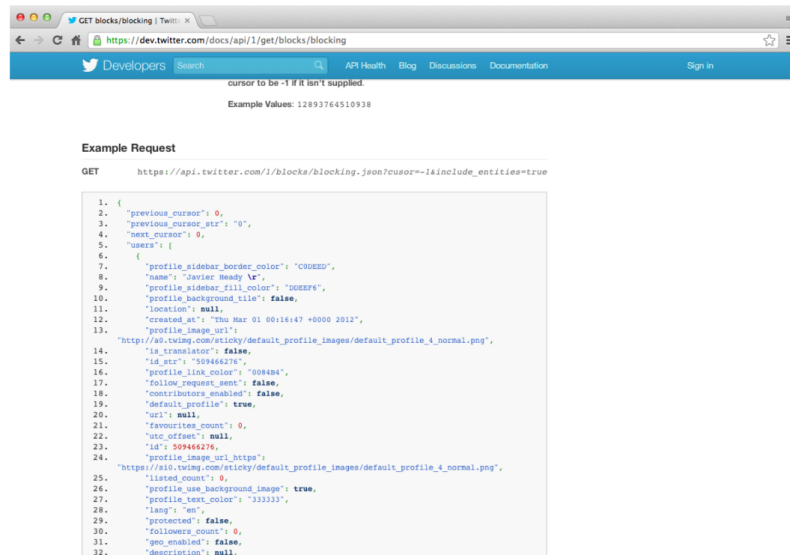
↳ Often skipped in classes  
↳ Tidy = easy to use

3/10

## Fasta file - gene seq

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

# What does data really look like?



```
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": "Dariusz Huczyk",
9.       "profile_sidebar_fill_color": "D0E2F6",
10.      "profile_background_tile": false,
11.      "location": null,
12.      "created_at": "Thu Mar 01 00:16:47 +0000 2012",
13.      "profile_image_url": "http://s3.amazonaws.com/twimg.com/avatar/default_profile_images/default_profile_4_normal.png",
14.      "is_translator": false,
15.      "id_str": "509466276",
16.      "profile_link_color": "0084B4",
17.      "follow_request_sent": false,
18.      "contributors_enabled": false,
19.      "default_profile": true,
20.      "url": null,
21.      "favourites_count": 0,
22.      "utc_offset": null,
23.      "id": 509466276,
24.      "profile_image_url_https": "http://s3.amazonaws.com/twimg.com/avatar/default_profile_images/default_profile_4_normal.png",
25.      "listed_count": 0,
26.      "profile_use_background_image": true,
27.      "profile_text_color": "333333",
28.      "lang": "en",
29.      "protected": false,
30.      "followers_count": 0,
31.      "geo_enabled": false,
32.      "description": null,
```

JSON

Neatly structured, but  
need to reorganize for  
analysis

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

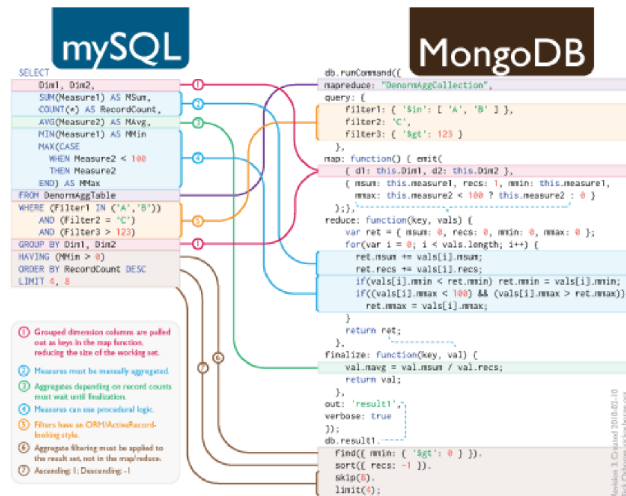
# What does data really look like?

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
Allergy Name:	TRAMADOL	Prescription Number:	2718953
Location:	DAYT29	Medication:	IBUPROFEN 600MG TAB
Date Entered:	09 Mar 2011	Instructions:	TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WITH FOOD
Reaction:	URINARY RETENTION	Status:	Active
Allergy Type:	DRUG	Refills Remaining:	3
Drug Class:	NON-OPIOID ANALGESICS	Last Filled On:	20 Aug 2010
Observed/Historical:	HISTORICAL	Initially Ordered On:	01 Jul 2010
Comments:	gradually worsening difficulty emptying bladder		

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

Free-text instructions

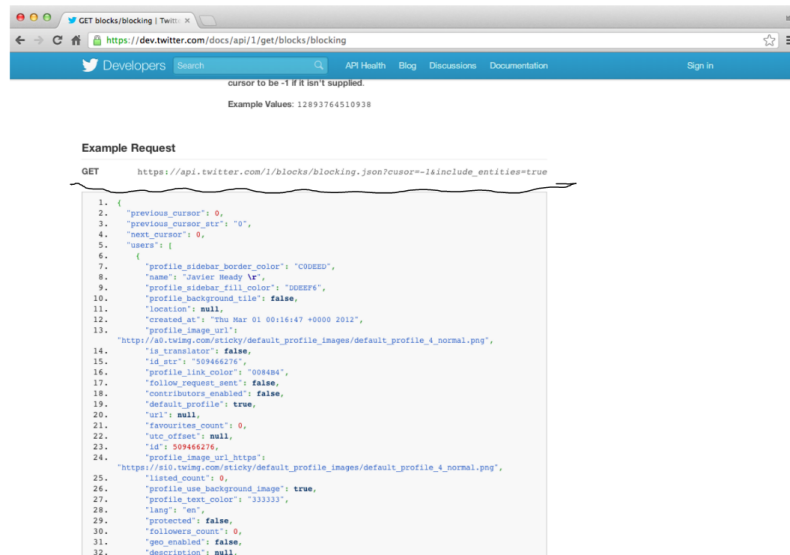
# Where is data?



↳ Data might need to be extracted (ex. from a database)

<http://rickosborne.org/blog/2010/02/infographic-migrating-from-sql-to-mapreduce-with-mongodb/>

# Where is data?



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

Trying to get info from a website (Twitter API)



# Where is data?



<https://data.baltimorecity.gov/>

↳ OPEN Data  
↳ Has lots of data files

## The goal of this course

Pipeline

Raw data -> Processing script -> tidy data -> data analysis -> data communication

most stats/machine learning classes

this class