

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

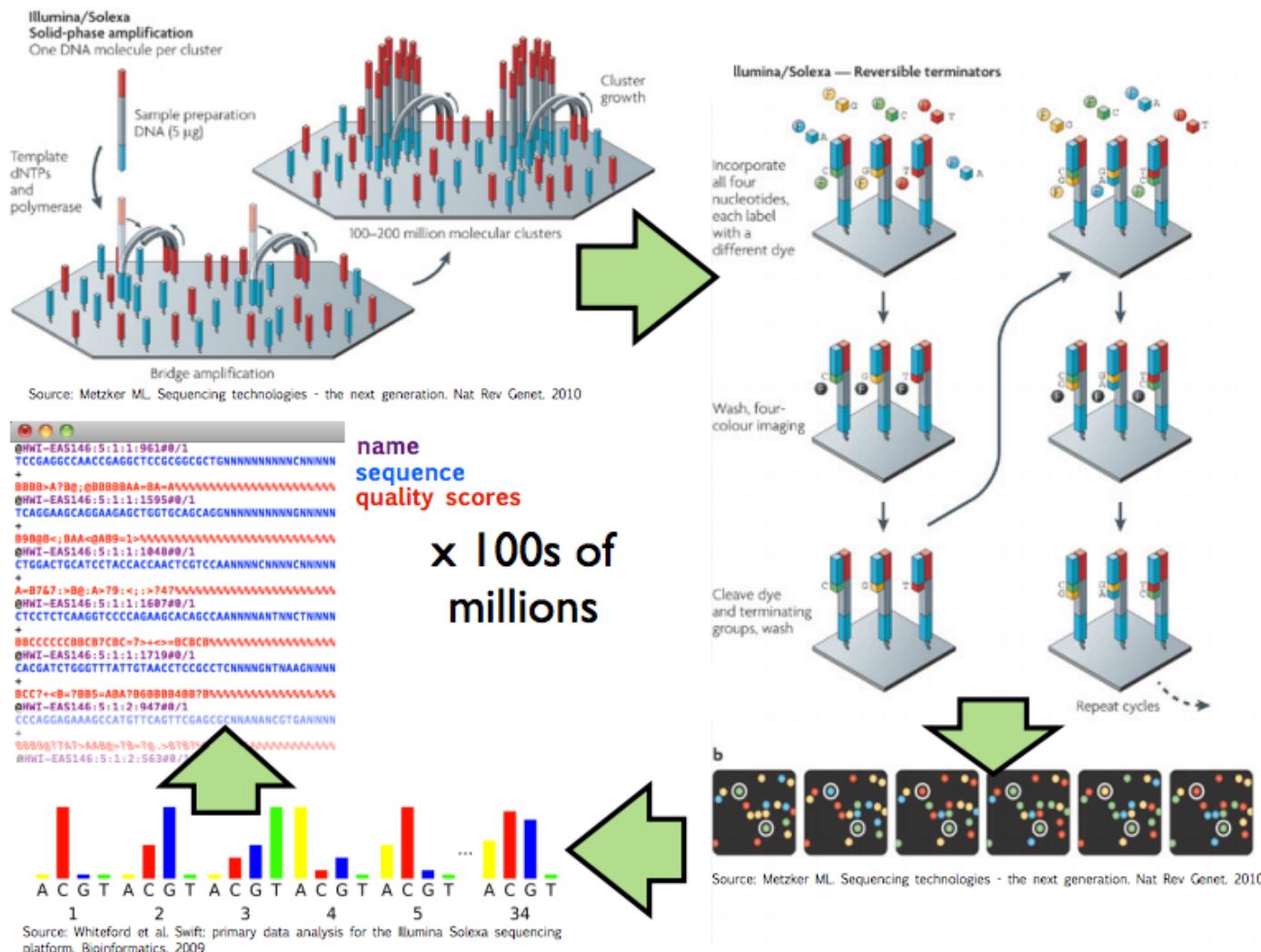
http://en.wikipedia.org/wiki/Computer_data_processing

An example of a processing pipeline



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

An example of a processing pipeline



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

What do raw data look like?

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

What do raw data look like?

The screenshot shows a web browser window with the following details:

- Title Bar:** GET blocks/blocking | Twitter
- URL:** https://dev.twitter.com/docs/api/1/get/blocks/blocking
- Header Bar:** Developers, Search, API Health, Blog, Discussions, Documentation, Sign in
- Content Area:**
 - A note: "cursor to be -1 if it isn't supplied."
 - An "Example Values" field containing the value: 12893764510938
 - A section titled "Example Request" with a code example:

```
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": "CODEED",
8.       "name": "Javier Heady \r",
9.       "profile_sidebar_fill_color": "DDEEF6",
10.      "profile_background_tile": false,
11.      "location": null,
12.      "created_at": "Thu Mar 01 00:16:47 +0000 2012",
13.      "profile_image_url":
14.        "http://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://si0.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 raw data look like?

ALLERGIES		MEDICATION HISTORY
Last Updated: 01 Dec 2011 @ 0851		Last Updated: 11 Apr 2011 @ 1737
Allergy Name:	TRIMETHOPRIM	Medication: AMLODIPIINE BESYLATE 10MG TAB
Location:	DAYT29	Instructions: TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TABLE GRAPEFRUIT JUICE--
Date Entered:	09 Mar 2011	Status: Active
Action:		Refills Remaining: 3
Allergy Type:	DRUG	Last Filled On: 28 Aug 2010
A 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
Allergy Name:	TRAMADOL	Pharmacy: DAYTON
Location:	DAYT29	Prescription Number: 2718953
Date Entered:	09 Mar 2011	
Action:	URINARY RETENTION	Medication: IBUPROFEN 600MG TAB
Allergy Type:	DRUG	Instructions: TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WT
A Drug Class:	NON-OPIOID ANALGESICS	Status: Active
Observed/Historical:	HISTORICAL	Refills Remaining: 3
Comments:	gradually worsening difficulty emptying bladder	Last Filled On: 28 Aug 2010
		Initially Ordered On: 01 Jul 2010

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

What do processed data look like?

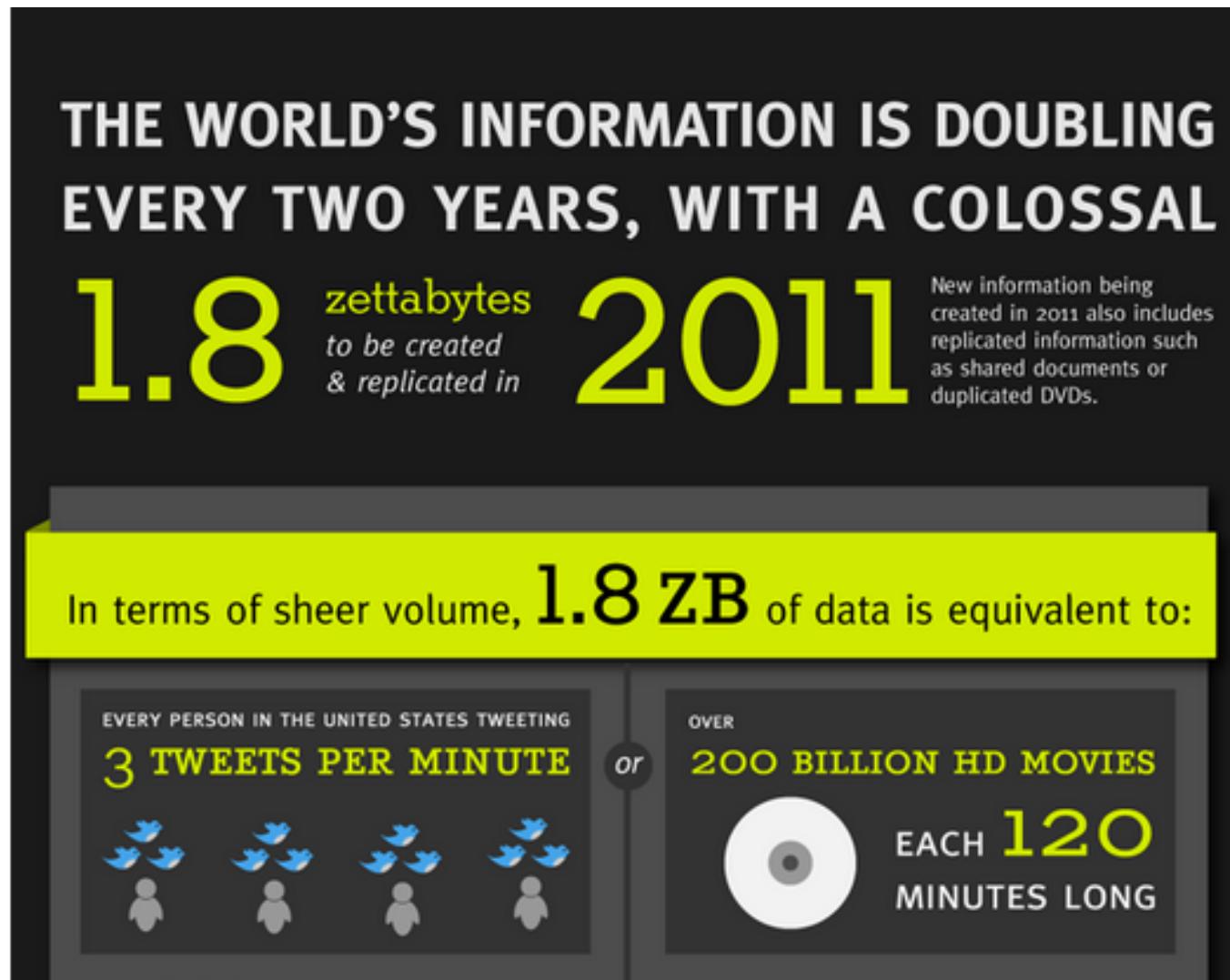
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
		problem_id	subject_id	start	stop	time_left	answer									
2	1	198	17	1307119889	1307120016	2369	A									
3	2	150	15	1307119991	1307120009	2376	D									
4	3	313	16	1307119994	1307120009	2376	E									
5	4	12	13	1307119995	1307120019	2366	B									
6	5	273	14	1307119995	1307120028	2357	A									
7	6	101	19	1307119996	1307120021	2364	B									
8	7	105	18	1307119998	1307120048	2337	B									
9	8	162	12	1307120000	1307120042	2343	C									
10	9	70	15	1307120011	1307120038	2347	C									
11	10	300	16	1307120012	1307120092	2293	B									
12	11	494	17	1307120013	1307120075	2310	D									
13	12	357	13	1307120014	1307120116	2247	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	1307120099	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	1307120191	2084	D									
25	24	121	19	1307120224	1307120256	2091	E									
26	25	297	15	1307120277	1307120342	2043	D									
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	1307120358	1307120444	1941	B									
36	35	92	14	1307120368	1307120397	1988	B									
37	36	395	16	1307120374	1307120406	1959	D									
38	37	257	17	1307120373	1307120515	1014	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?

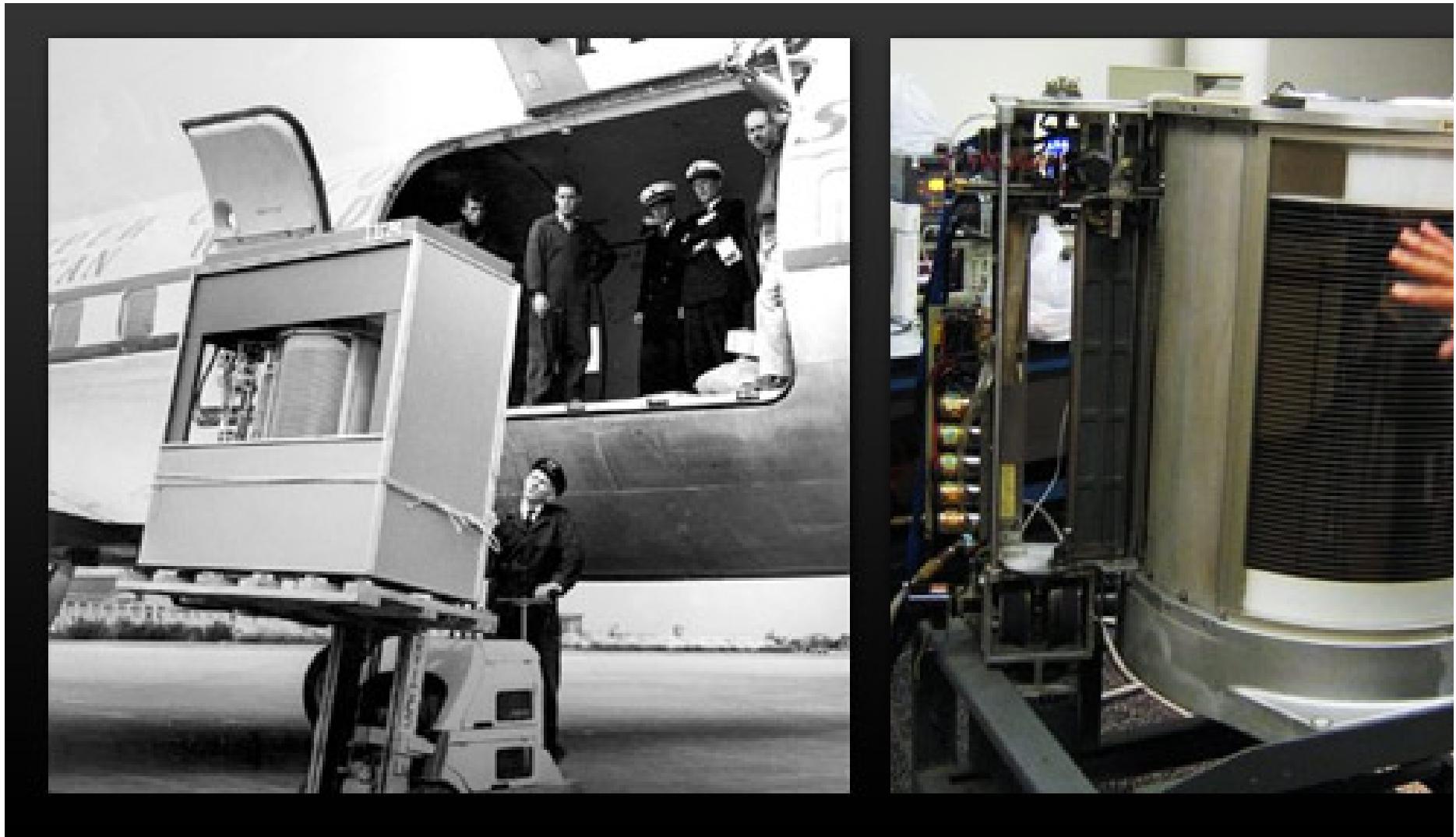


<http://mashable.com/2011/06/28/data-infographic/>

So what about big data?



Depends on your perspective



Why big data now?

An Experimental Study of the Small World Problem*

JEFFREY TRAVERS

Harvard University

AND

STANLEY MILGRAM

The City University of New York

Arbitrarily selected individuals ($N=296$) in Nebraska and Boston are asked to generate acquaintance chains to a target person in Massachusetts, employing “the small world method” (Milgram, 1967). Sixty-four chains reach the target person. Within this group the mean number of intermediaries between starters and targets is 5.2. Boston starting chains reach the target

[Travers and Milgram \(1969\) Sociometry](#)

Why big data now?

[arXiv.org > physics > arXiv:0803.0939](#)

Physics > Physics and Society

Planetary-Scale Views on an Instant-Messaging Netw

Jure Leskovec, Eric Horvitz

(Submitted on 6 Mar 2008)

We present a study of anonymized data capturing a month of high-level communication activ
Microsoft Messenger instant-messaging system. We examine characteristics and patterns tha
dynamics of large numbers of people, rather than the actions and characteristics of individual
properties of 30 billion conversations among 240 million people. From the data, we construct
million nodes and 1.3 billion undirected edges, creating the largest social network constructe
on multiple aspects of the dataset and synthesized graph. We find that the graph is well-conne
We investigate on a planetary-scale the oft-cited $\langle k \rangle = 6$, that people are separated by ``six de
the average path length among Messenger users is 6.6. We also find that people tend to comm
when they have similar age, language, and location, and that cross-gender conversations are
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](#)