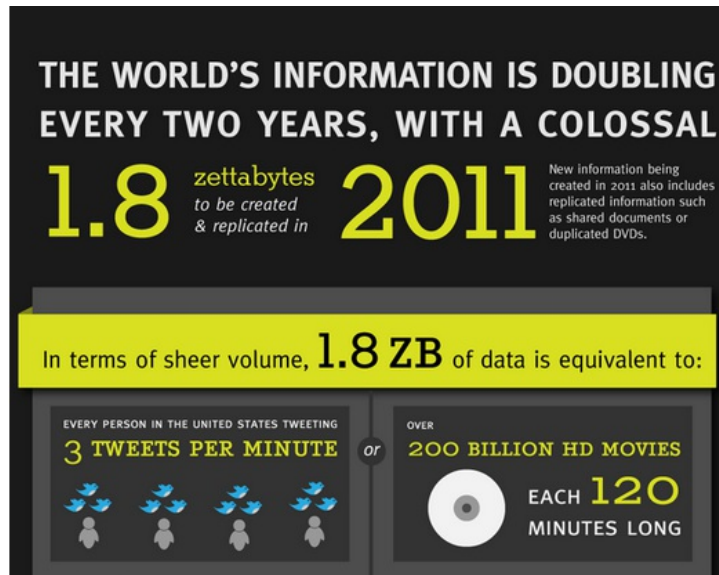




What about big data?

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

How much is there?

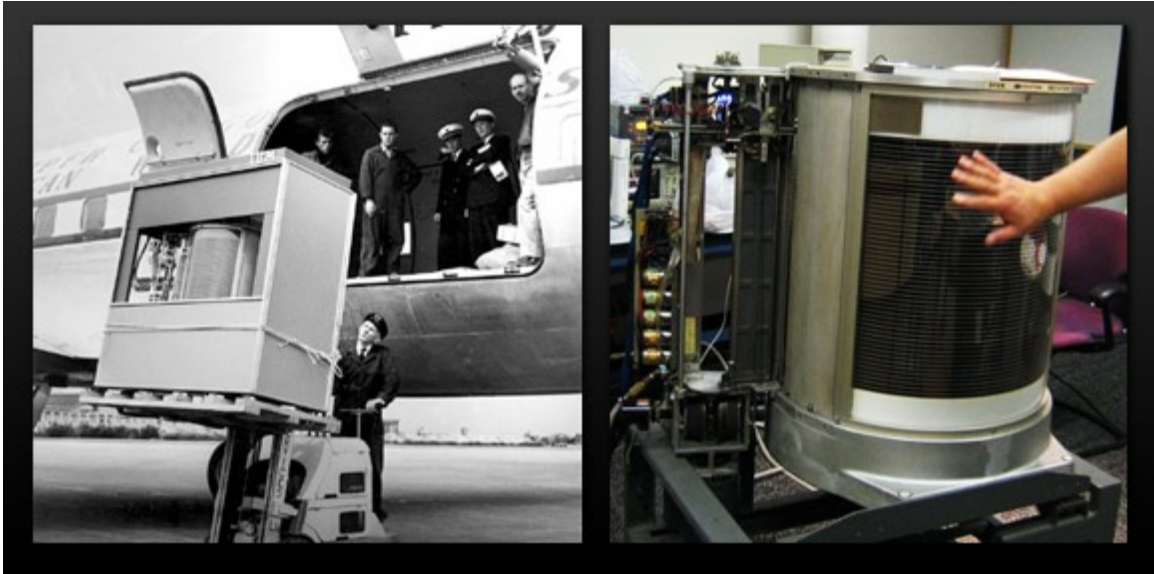


<http://mashable.com/2011/06/28/data-infographic/>

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

Search or A

Physics > Physics and Society

Planetary-Scale Views on an Instant-Messaging Network

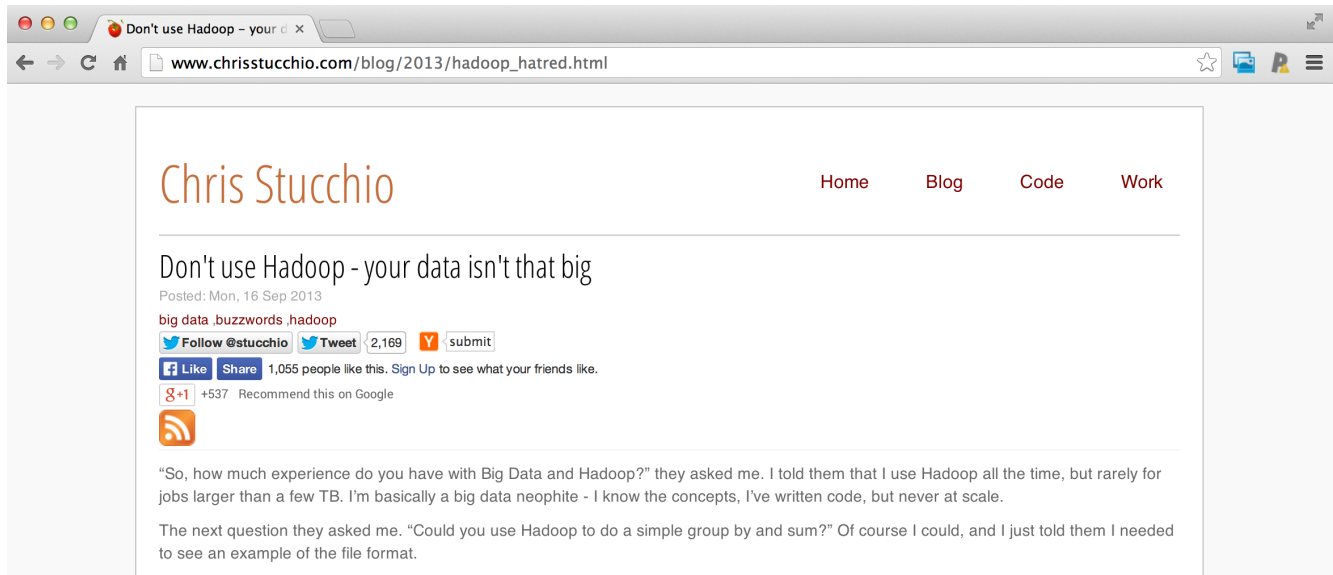
Jure Leskovec, Eric Horvitz

(Submitted on 6 Mar 2008)

We present a study of anonymized data capturing a month of high-level communication activities within the whole of the Microsoft Messenger instant-messaging system. We examine characteristics and patterns that emerge from the collective dynamics of large numbers of people, rather than the actions and characteristics of individuals. The dataset contains summary properties of 30 billion conversations among 240 million people. From the data, we construct a communication graph with 180 million nodes and 1.3 billion undirected edges, creating the largest social network constructed and analyzed to date. We report on multiple aspects of the dataset and synthesized graph. We find that the graph is well-connected and robust to node removal. We investigate on a planetary-scale the oft-cited report that people are separated by "six degrees of separation" and find that the average path length among Messenger users is 6.6. We also find that people tend to communicate more with each other when they have similar age, language, and location, and that cross-gender conversations are both more frequent and of longer duration than conversations with the same gender.

[Leskovec and Horvitz WWW '08](#)

Big or small - you need the right data



The screenshot shows a web browser window with the address bar displaying www.chrisstucchio.com/blog/2013/hadoop_hatred.html. The page content includes the author's name "Chris Stucchio" in orange, navigation links for "Home", "Blog", "Code", and "Work", and the article title "Don't use Hadoop - your data isn't that big". Below the title, it says "Posted: Mon, 16 Sep 2013" and lists tags "big data", "buzzwords", and "hadoop". Social sharing buttons for Twitter (Follow @stucchio, Tweet 2,169), Facebook (Like, Share, 1,055 likes), and Google+ (+537) are present, along with a "submit" button. An RSS feed icon is also visible. The article text begins with a quote: "So, how much experience do you have with Big Data and Hadoop?" they asked me. I told them that I use Hadoop all the time, but rarely for jobs larger than a few TB. I'm basically a big data neophyte - I know the concepts, I've written code, but never at scale. The next question they asked me. "Could you use Hadoop to do a simple group by and sum?" Of course I could, and I just told them I needed to see an example of the file format.

http://www.chrisstucchio.com/blog/2013/hadoop_hatred.html

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](#)