

Introduction To Big Data

<https://github.com/bobbylovemovie/trainbigdata>

bobbylovemovie@gmail.com

Big Data
Big Data Analytics
Data Science
Machine Learning
Artificial Intelligence
Deep Learning



What is Big Data ?

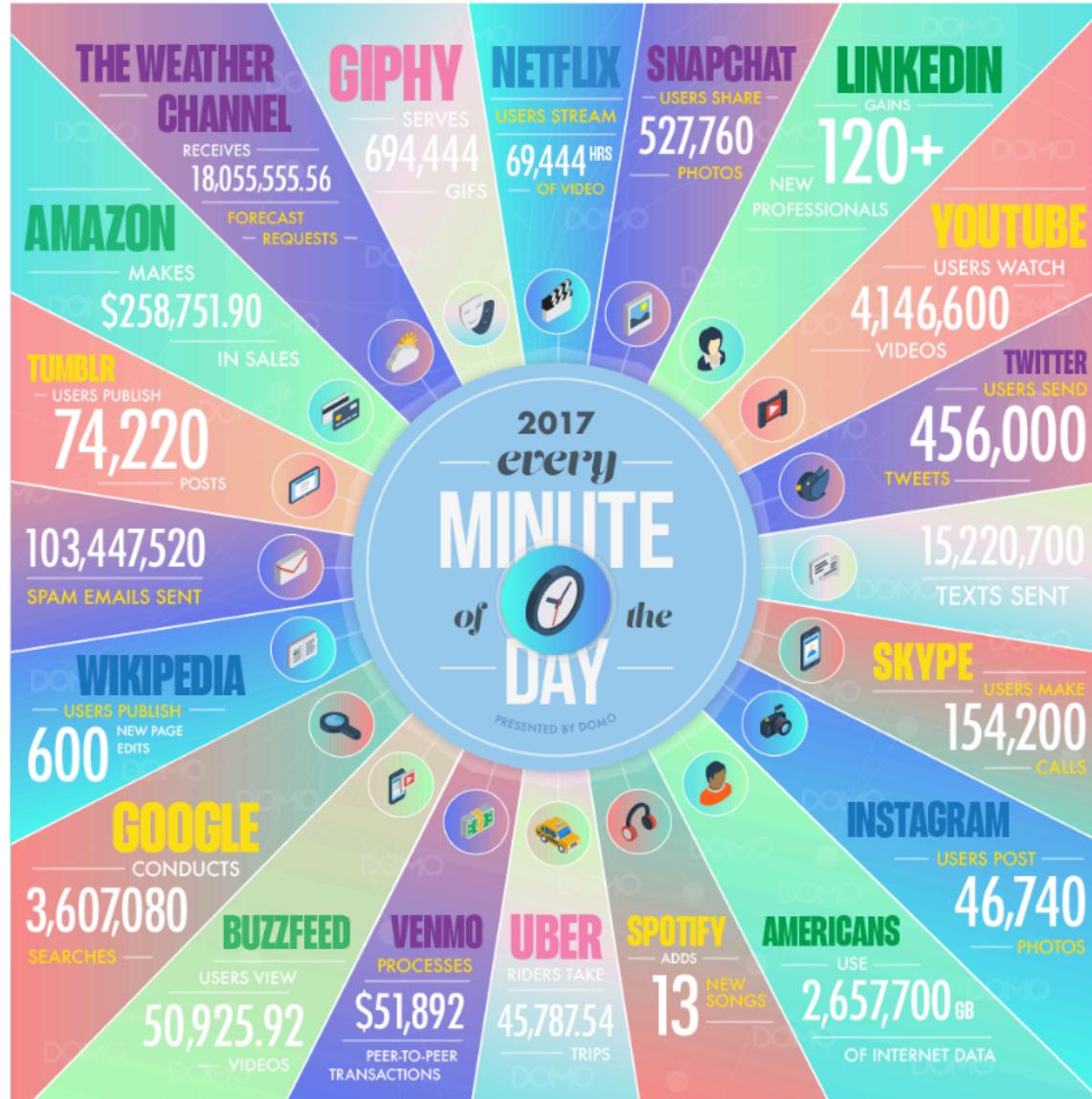
“ Big data is data that exceeds the processing capacity of conventional database systems.

**The data is too big, moves too fast,
or doesn't fit the structures of your database architectures.**

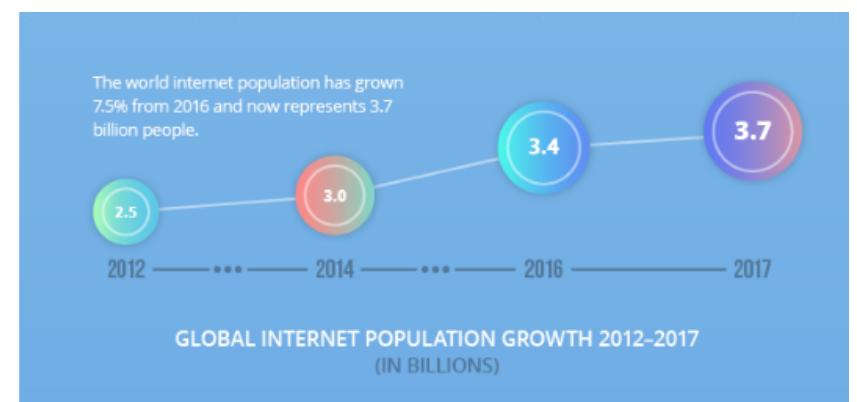
**To gain value from this data,
you must choose an alternative way to process it. ”**

Big Data Now: O'Reilly Media

Why Big Data ?



DATA NEVER SLEEP 4.0



Three Characteristics of Big Data



Volume

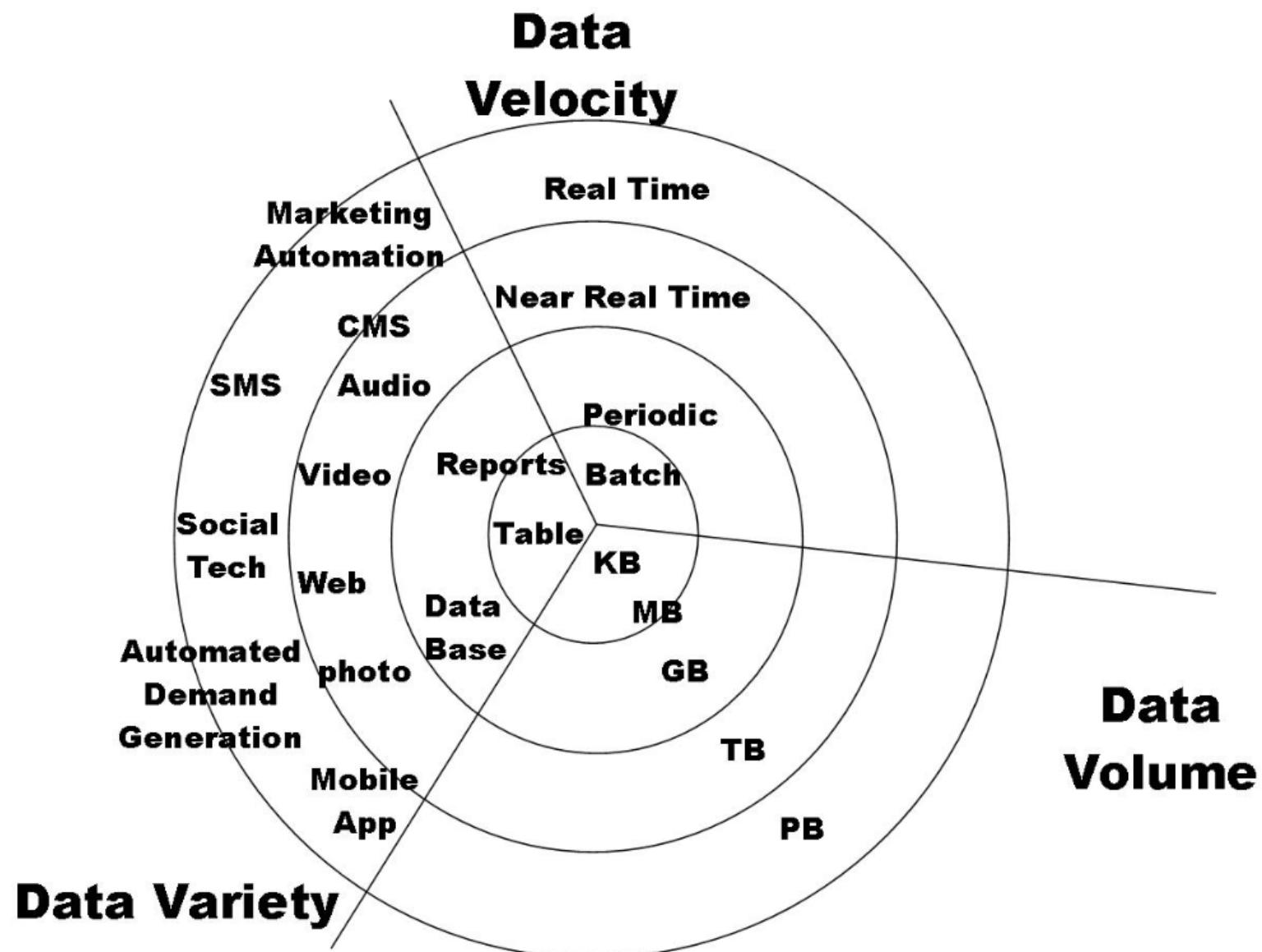
- Volumes of data are larger than those conventional relational database infrastructures can cope with

Velocity

- Rate at which data flows in is much faster.
 - Mobile event and interaction by users.
 - Video, image , audio from users

Variety

- the source data is diverse, and doesn't fall into neat relational structures eg. text from social networks, image data, a raw feed directly from a sensor source.



Big Data = Volume, Variety and Velocity (3Vs)

Amount of new data stored varies across geography

New data stored¹ by geography, 2010
Petabytes



Volume

¹ New data stored defined as the amount of available storage used in a given year; see appendix for more on the definition and assumptions.

SOURCE: IDC storage reports; McKinsey Global Institute analysis

Velocity

30 billion pieces of content are shared on Facebook every month.



4 billion hours of video are watched on YouTube each month



400 Million Tweets are sent per day
200M monthly active users



Source: IBM

Variety



People to Machine



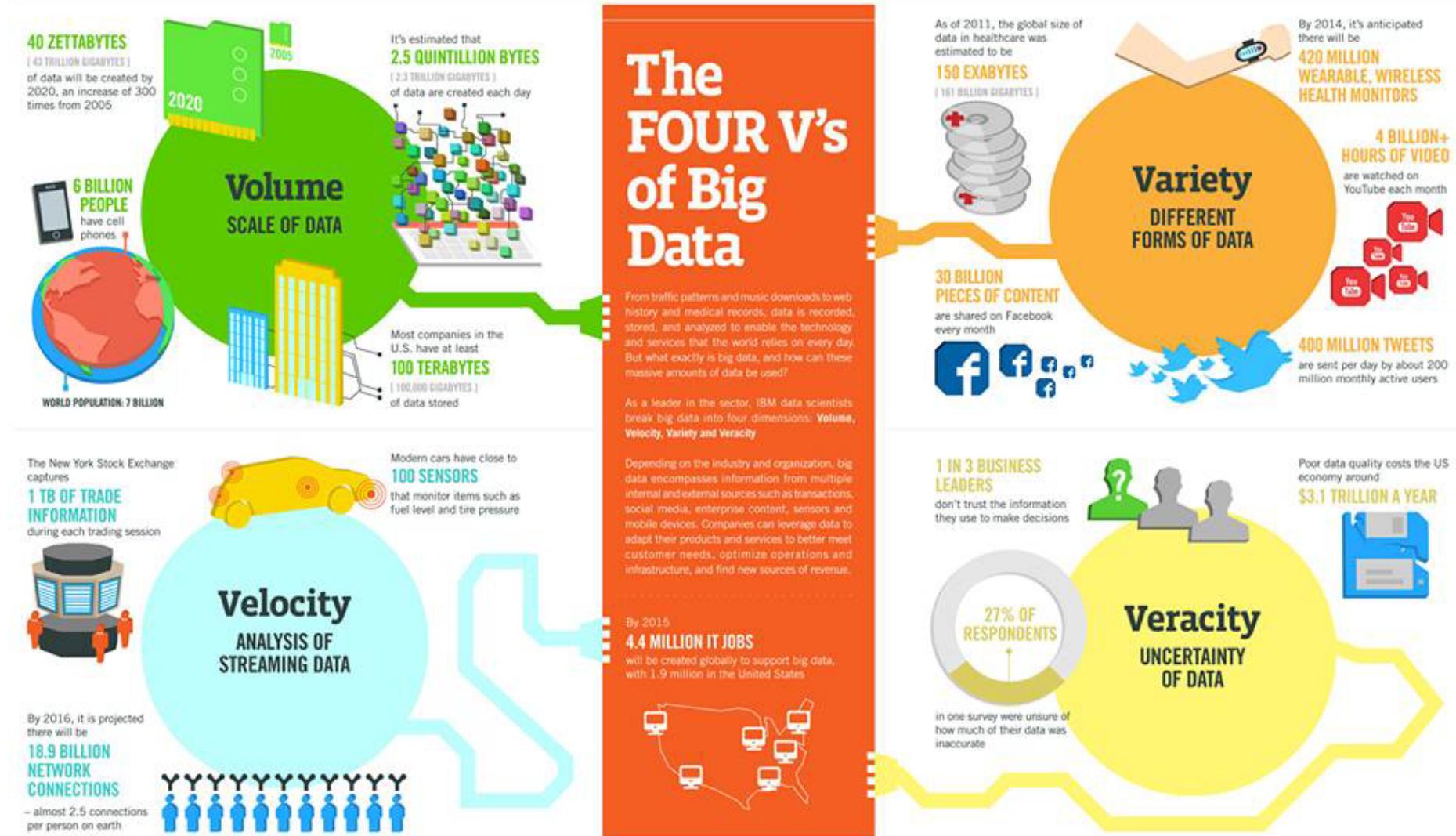
Machine to Machine



Scale

1000 kilobytes = 1 Megabyte
1000 Megabytes = 1 Gigabyte
1000 Gigabytes = 1 Terabyte
1000 Terabytes = 1 Petabyte
1000 Petabytes = 1 Exabyte
1000 Exabytes = 1 Zettabyte
1000 Zettabytes = 1 Yottabyte
1000 Yottabytes = 1 Brontobyte
1000 Brontobytes = 1 Geopbyte

4Vs of Big Data



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MPTEC, QAS

IBM

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005



6 BILLION PEOPLE

have cell phones



WORLD POPULATION: 7 BILLION



2005

Volume SCALE OF DATA



It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day



Most companies in the
U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

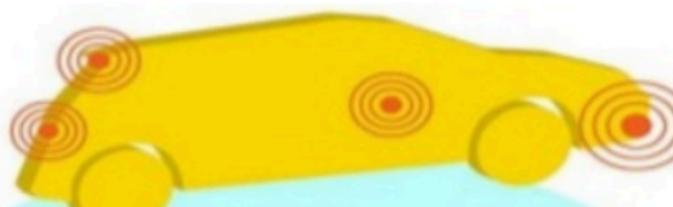
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION during each trading session



By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth



Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

Velocity

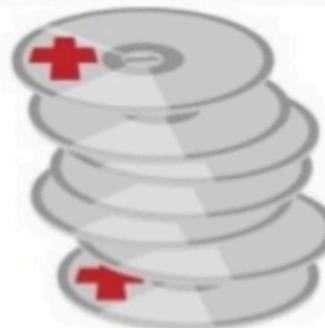
ANALYSIS OF STREAMING DATA



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]



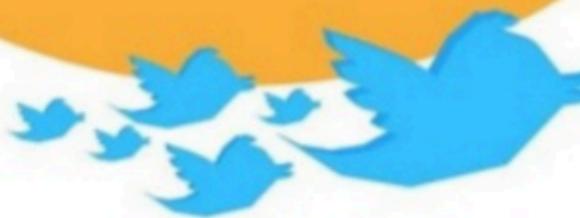
**30 BILLION
PIECES OF CONTENT**

are shared on Facebook every month



Variety

DIFFERENT
FORMS OF DATA

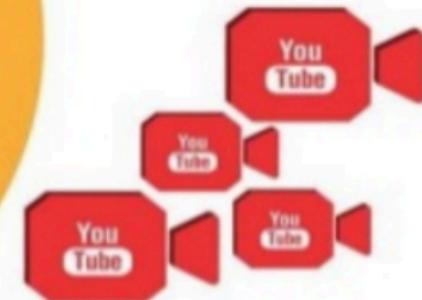


By 2014, it's anticipated there will be

**420 MILLION
WEARABLE, WIRELESS
HEALTH MONITORS**

**4 BILLION+
HOURS OF VIDEO**

are watched on YouTube each month



400 MILLION TWEETS

are sent per day by about 200 million monthly active users

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



27% OF
RESPONDENTS

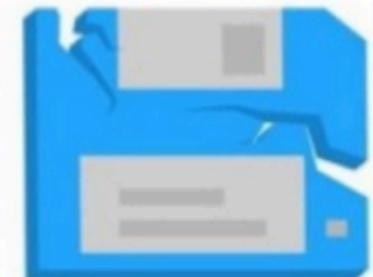
in one survey were unsure of how much of their data was inaccurate

Veracity

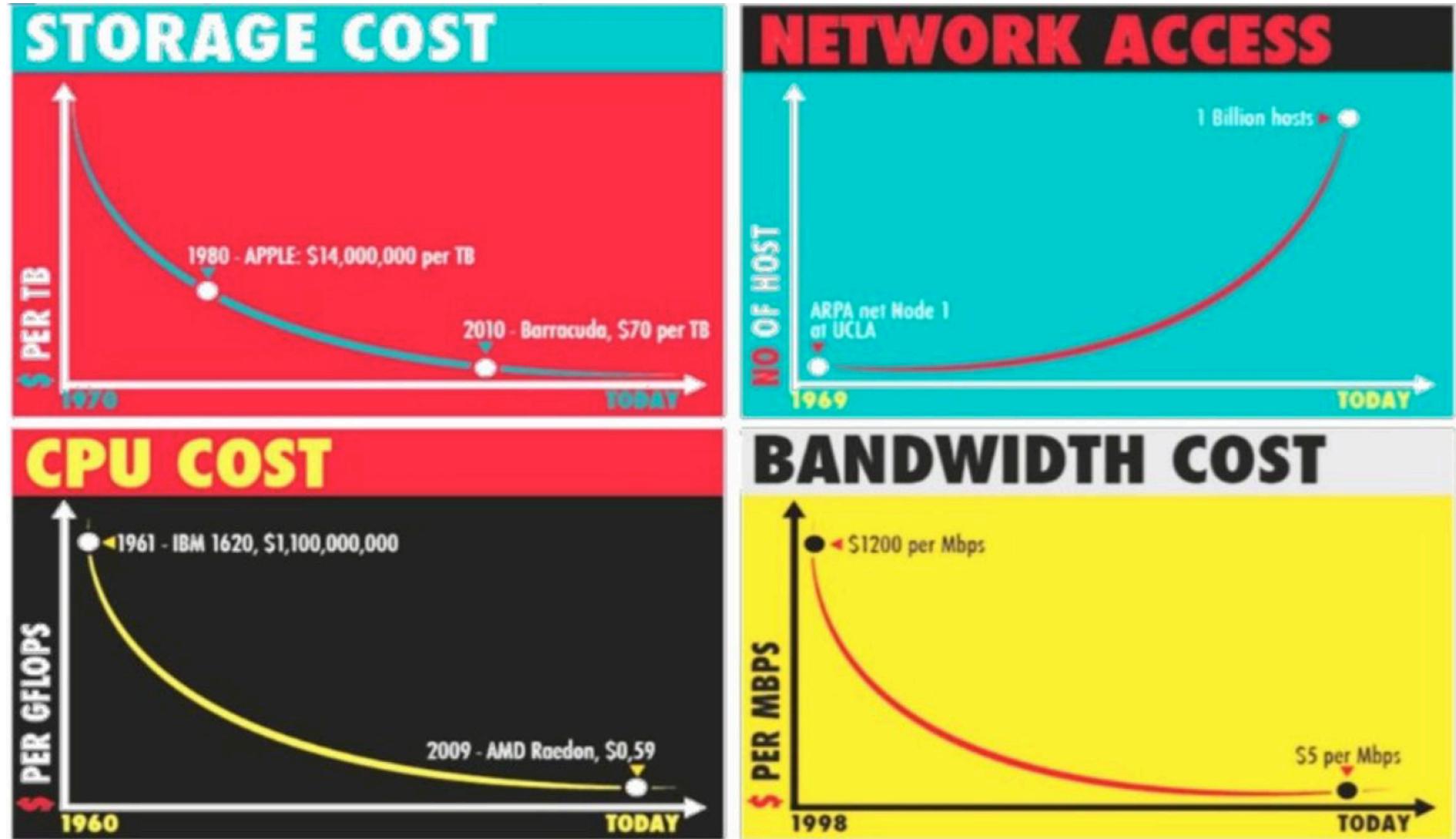
UNCERTAINTY OF DATA

Poor data quality costs the US economy around

\$3.1 TRILLION A YEAR



Big Data : Why Now?



Source: William EL KAIM, Enterprise Architecture and Technology Innovation 21