

# Big Data

# Metric Prefixes

SI prefixes								V · T · E
Prefix		1000 <sup>m</sup>	10 <sup>n</sup>	Decimal	English word		Since <sup>[n 1]</sup>	
name	symbol				short scale	long scale		
yotta	Y	1000 <sup>8</sup>	10 <sup>24</sup>	1 000 000 000 000 000 000 000 000	septillion	quadrillion	1991	
zetta	Z	1000 <sup>7</sup>	10 <sup>21</sup>	1 000 000 000 000 000 000 000	sextillion	thousand trillion	1991	
exa	E	1000 <sup>6</sup>	10 <sup>18</sup>	1 000 000 000 000 000 000	quintillion	trillion	1975	
peta	P	1000 <sup>5</sup>	10 <sup>15</sup>	1 000 000 000 000 000	quadrillion	thousand billion	1975	
tera	T	1000 <sup>4</sup>	10 <sup>12</sup>	1 000 000 000 000	trillion	billion	1960	
giga	G	1000 <sup>3</sup>	10 <sup>9</sup>	1 000 000 000	billion	thousand million	1960	
mega	M	1000 <sup>2</sup>	10 <sup>6</sup>	1 000 000	million		1960	
kilo	k	1000 <sup>1</sup>	10 <sup>3</sup>	1 000	thousand		1795	
hecto	h	1000 <sup>2/3</sup>	10 <sup>2</sup>	100	hundred		1795	
deca	da	1000 <sup>1/3</sup>	10 <sup>1</sup>	10	ten		1795	
		1000 <sup>0</sup>	10 <sup>0</sup>	1	one		–	

# 2017 This Is What Happens In An Internet Minute

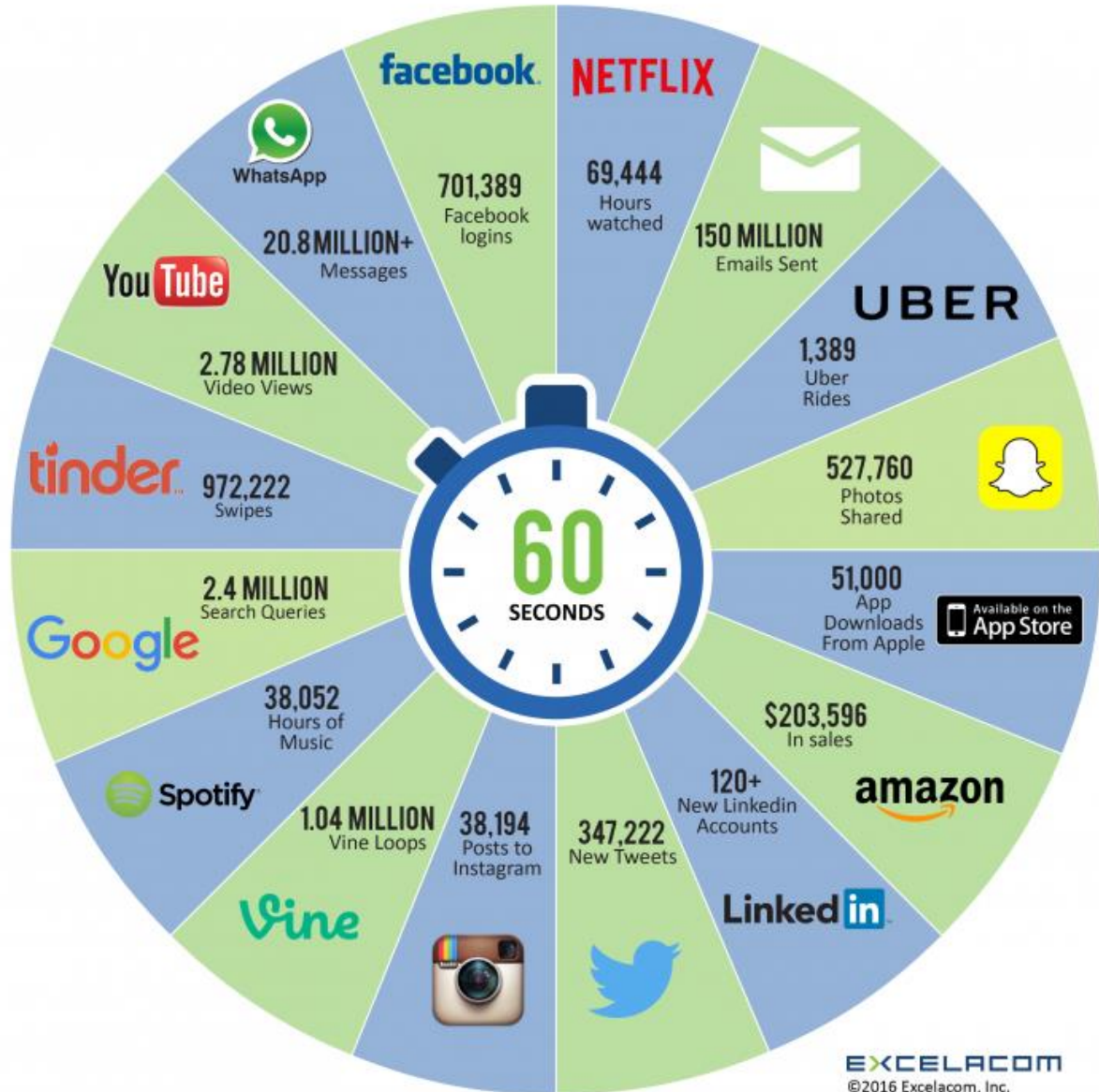
# 2018 This Is What Happens In An Internet Minute <sup>3</sup>



Created By:  
@LoriLewis  
@OfficiallyChadd

Created By:  
@LoriLewis  
@OfficiallyChadd

# 2016 What happens in an INTERNET MINUTE?



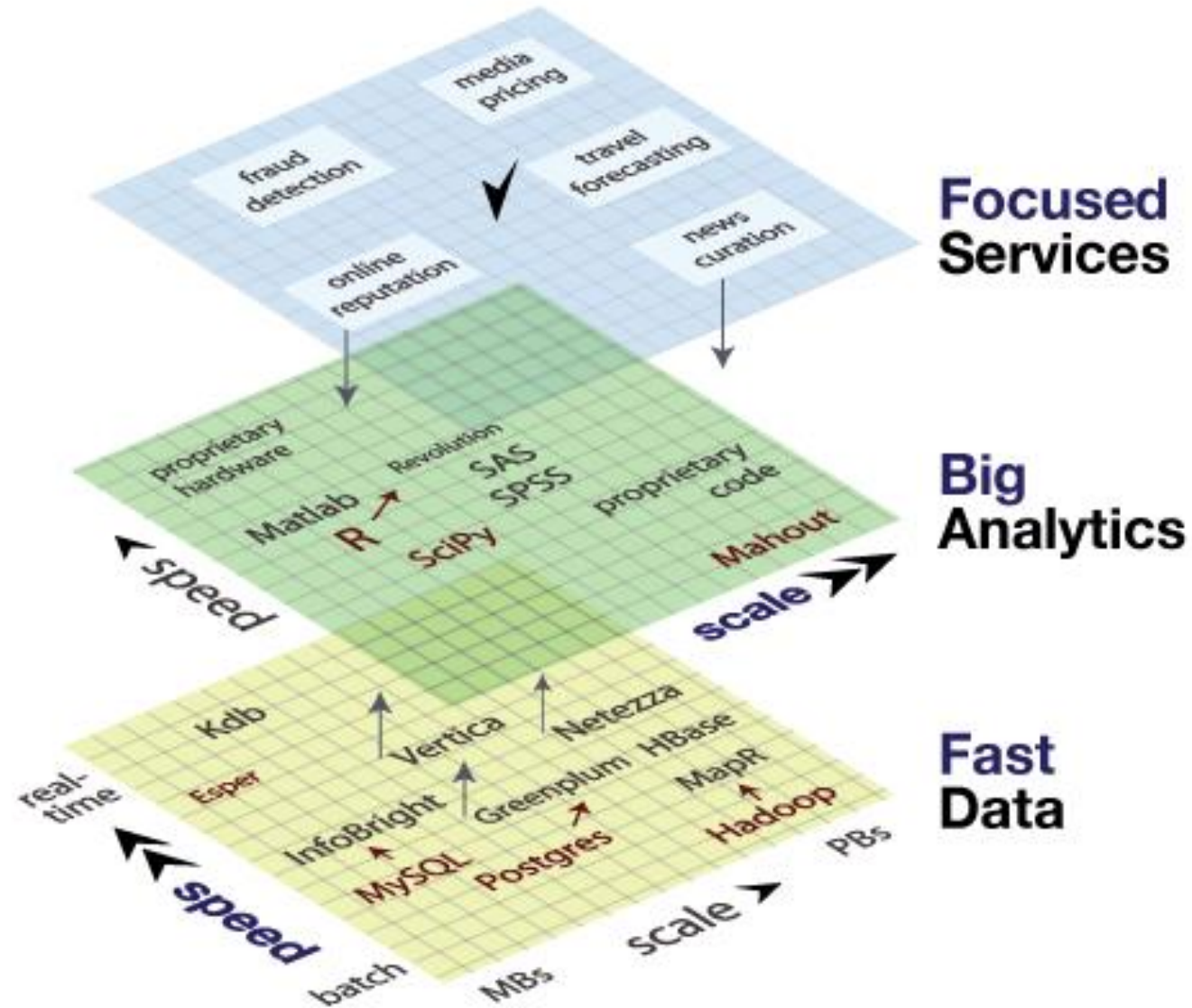
EXCELACOM  
©2016 Excelacom, Inc.

# 2019 This Is What Happens In An Internet Minute



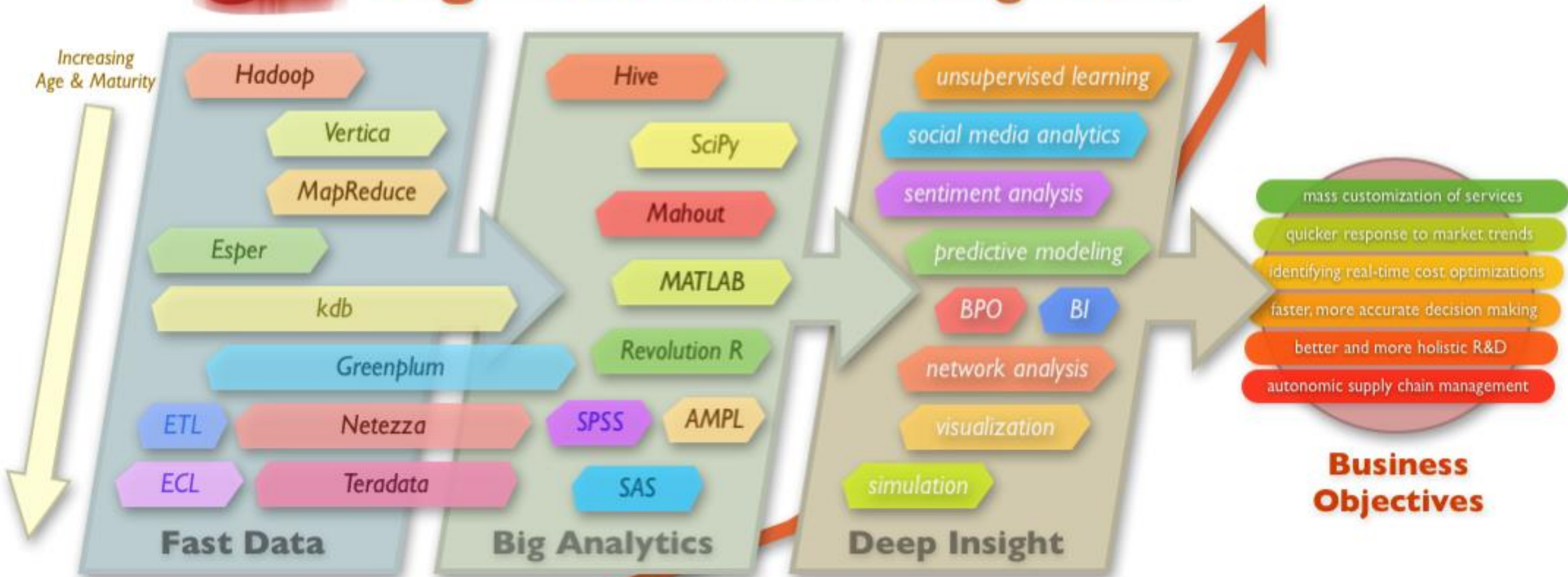
Created By:  
@LoriLewis  
@OfficiallyChadd

# The Emerging Big Data Stack





# Big Data: The Moving Parts



From <http://blogs.zdnet.com/Hinchcliffe>

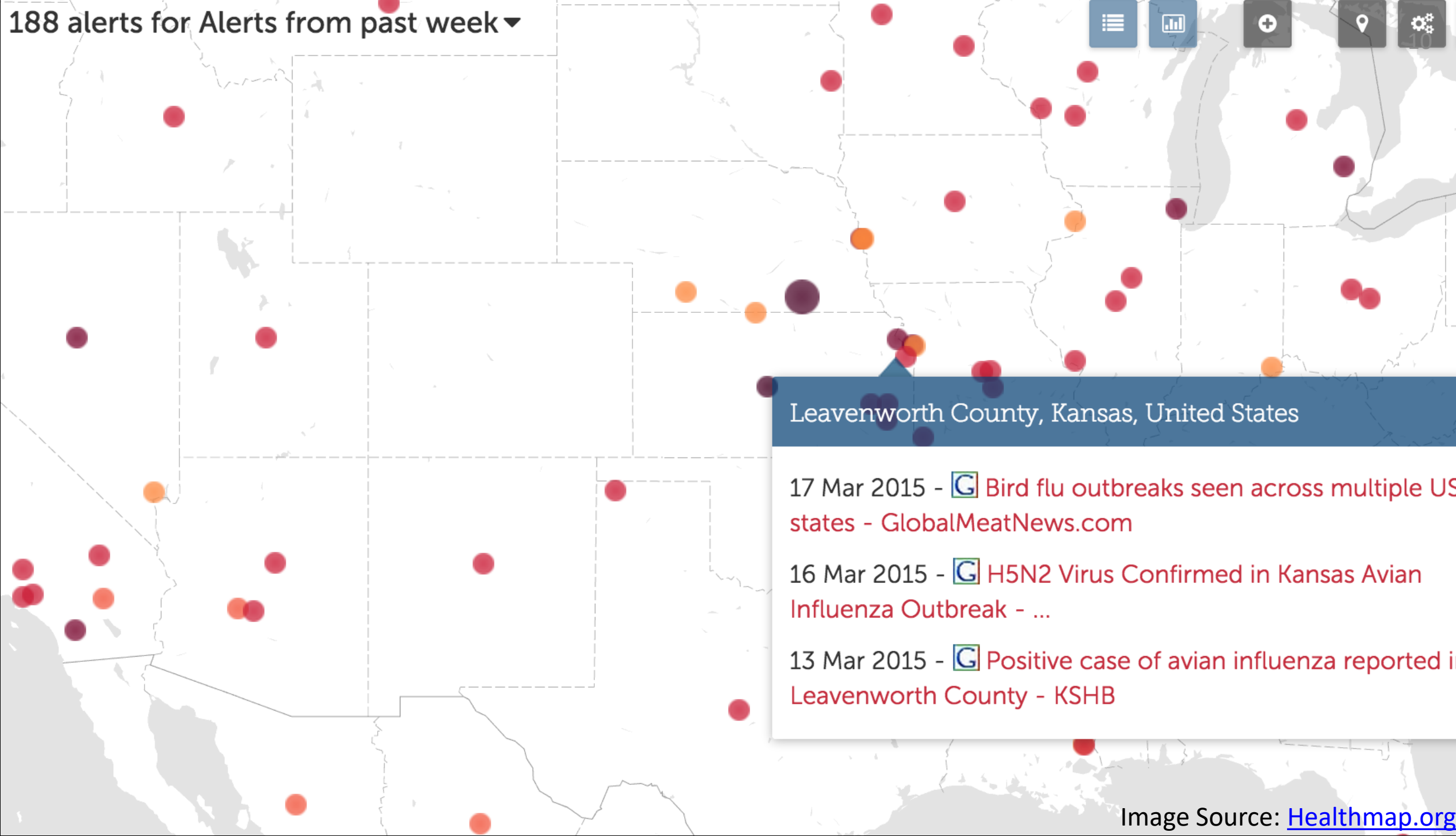
the growth of data will be exponential for the foreseeable future



the amount of data stored by the average company today

Google Trends  
healthmap.org

188 alerts for Alerts from past week ▾



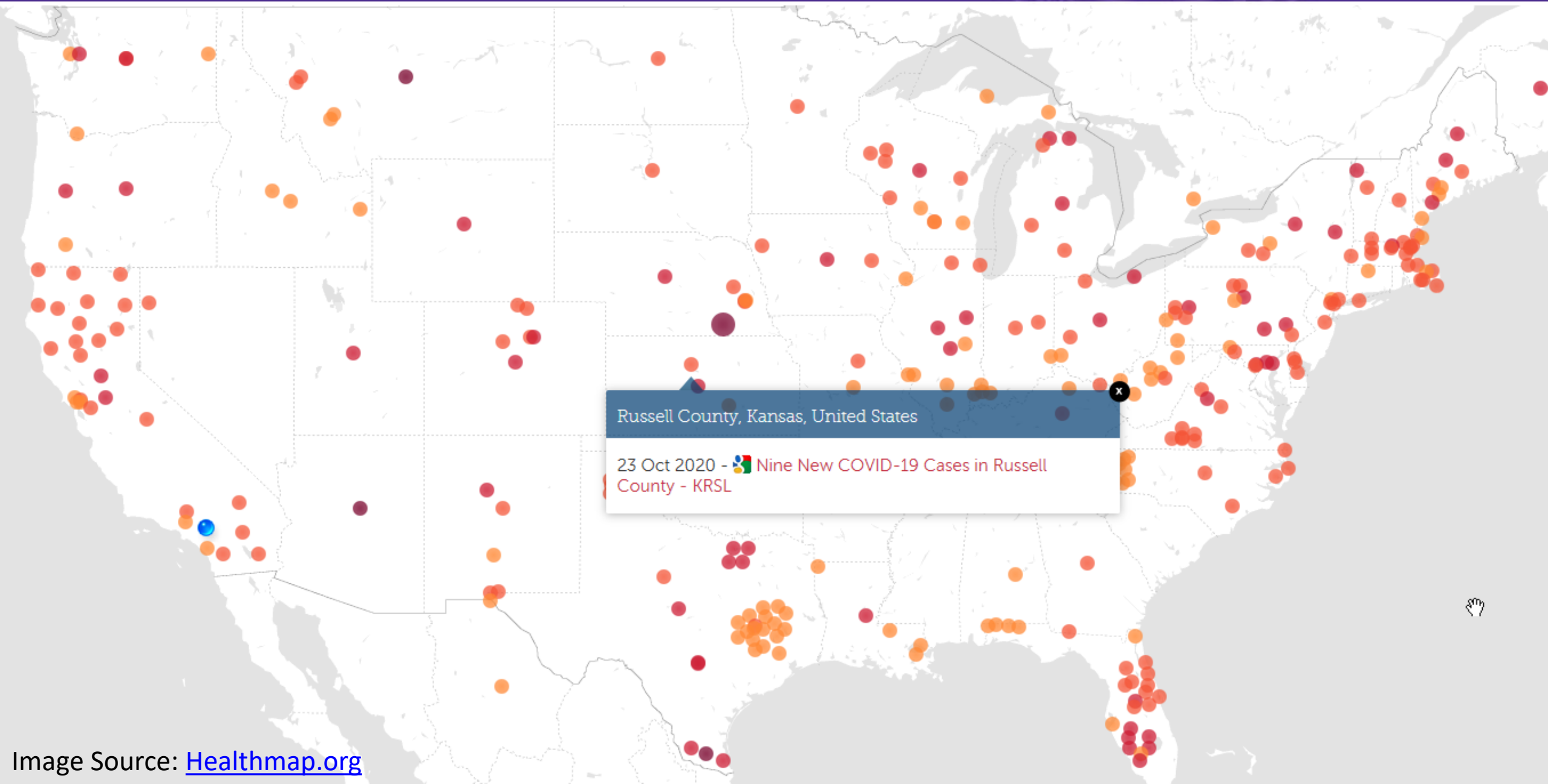
Leavenworth County, Kansas, United States

17 Mar 2015 - [G](#) Bird flu outbreaks seen across multiple US states - GlobalMeatNews.com

16 Mar 2015 - [G](#) H5N2 Virus Confirmed in Kansas Avian Influenza Outbreak - ...

13 Mar 2015 - [G](#) Positive case of avian influenza reported in Leavenworth County - KSHB





# Big Data Uses

- Topic Modeling
- Natural Language Processing
- Analytics & Data Forecasting
- Sentiment Analysis & Crowdsourcing
- Information Visualization
- Thematic Mapping

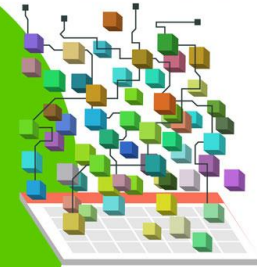
### 40 ZETTABYTES

[ 43 TRILLION GIGABYTES ]  
of data will be created by 2020, an increase of 300 times from 2005



### It's estimated that 2.5 QUINTILLION BYTES

[ 2.3 TRILLION GIGABYTES ]  
of data are created each day



## Volume SCALE OF DATA

6 BILLION PEOPLE  
have cell phones



WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least  
**100 TERABYTES**  
[ 100,000 GIGABYTES ]  
of data stored



# The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015  
**4.4 MILLION IT JOBS**  
will be created globally to support big data, with 1.9 million in the United States



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

**150 EXABYTES**  
[ 161 BILLION GIGABYTES ]

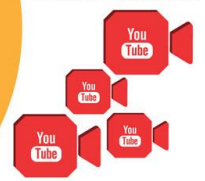


**30 BILLION PIECES OF CONTENT**  
are shared on Facebook every month



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



## Variety DIFFERENT FORMS OF DATA

**400 MILLION TWEETS**  
are sent per day by about 200 million monthly active users



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



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

## Velocity ANALYSIS OF STREAMING DATA

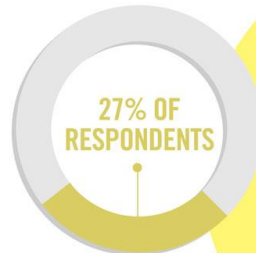
By 2016, it is projected there will be  
**18.9 BILLION NETWORK CONNECTIONS**  
— almost 2.5 connections per person on earth



**1 IN 3 BUSINESS LEADERS**  
don't trust the information they use to make decisions



Poor data quality costs the US economy around  
**\$3.1 TRILLION A YEAR**



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

## Veracity UNCERTAINTY OF DATA

# 40 ZETTABYTES

[ 43 TRILLION GIGABYTES ]  
of data will be created by 2020, an increase of 300 times from 2005




# It's estimated that 2.5 QUINTILLION BYTES

[ 2.3 TRILLION GIGABYTES ]  
of data are created each day

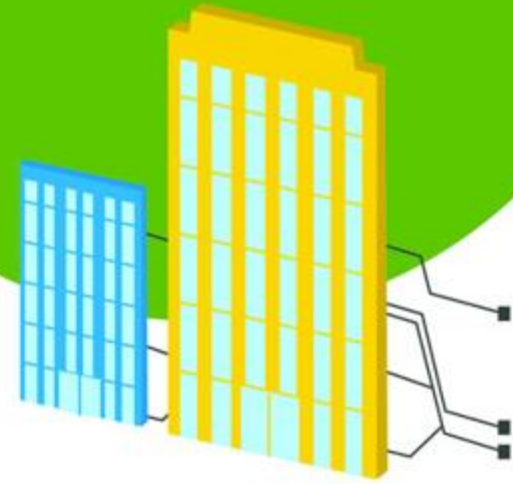


# Volume SCALE OF DATA

 **6 BILLION PEOPLE**  
have cell phones



WORLD POPULATION: 7 BILLION



Most companies in the U.S. have at least

# 100 TERABYTES

[ 100,000 GIGABYTES ]  
of data stored

T  
H  
C  
I

From  
hist  
stor  
and  
But  
mas

As a  
brea

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

**150 EXABYTES**

[ 161 BILLION GIGABYTES ]



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



# Variety

## DIFFERENT FORMS OF DATA

**30 BILLION PIECES OF CONTENT**

are shared on Facebook every month



**400 MILLION TWEETS**

are sent per day by about 200 million monthly active users



The New York Stock Exchange captures

**1 TB OF TRADE INFORMATION**

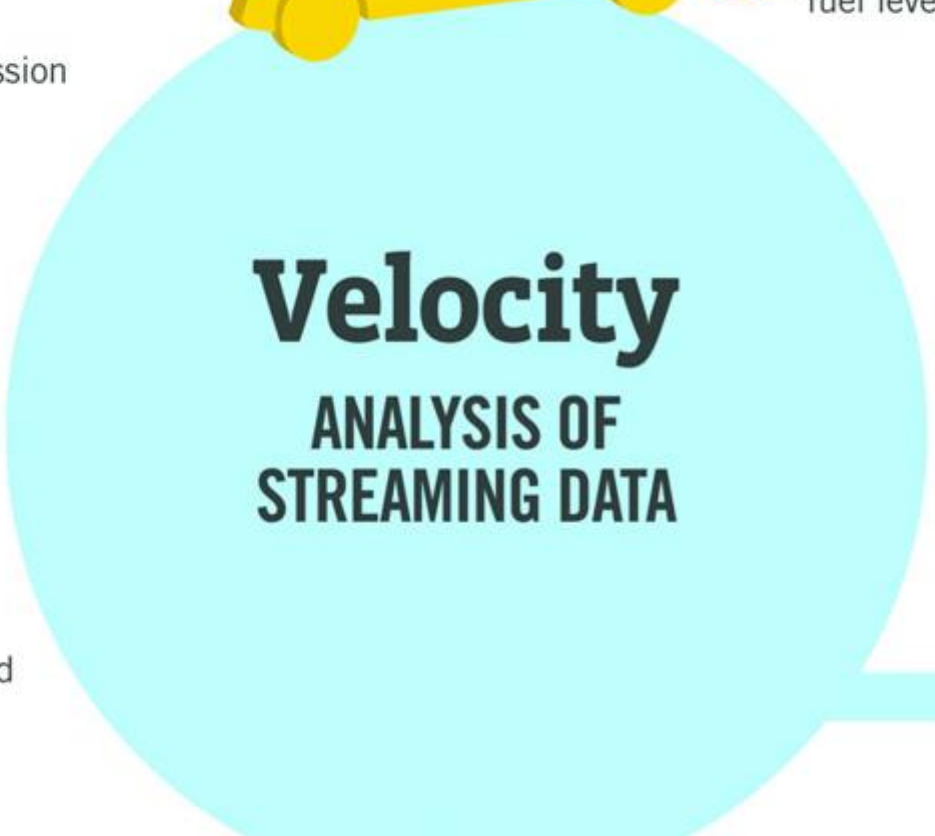
during each trading session



Modern cars have close to

**100 SENSORS**

that monitor items such as fuel level and tire pressure



# Velocity

## ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be

**18.9 BILLION NETWORK CONNECTIONS**

– almost 2.5 connections per person on earth



velocity

Depend  
data e  
interna  
social  
mobile  
adapt t  
custom  
infrastr

By 2016  
**4.4 M**  
will be  
with 1.



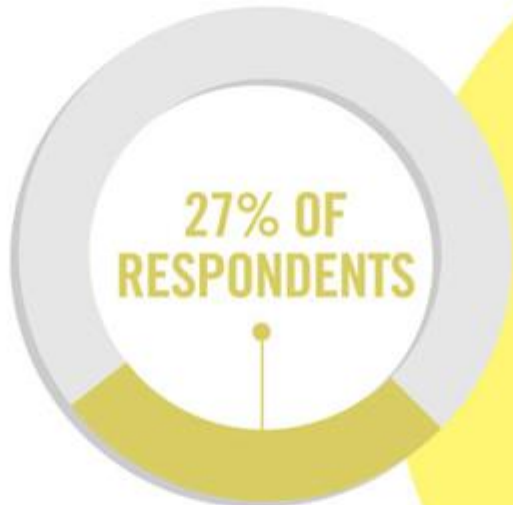
### 1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



Poor data quality costs the US economy around

**\$3.1 TRILLION A YEAR**



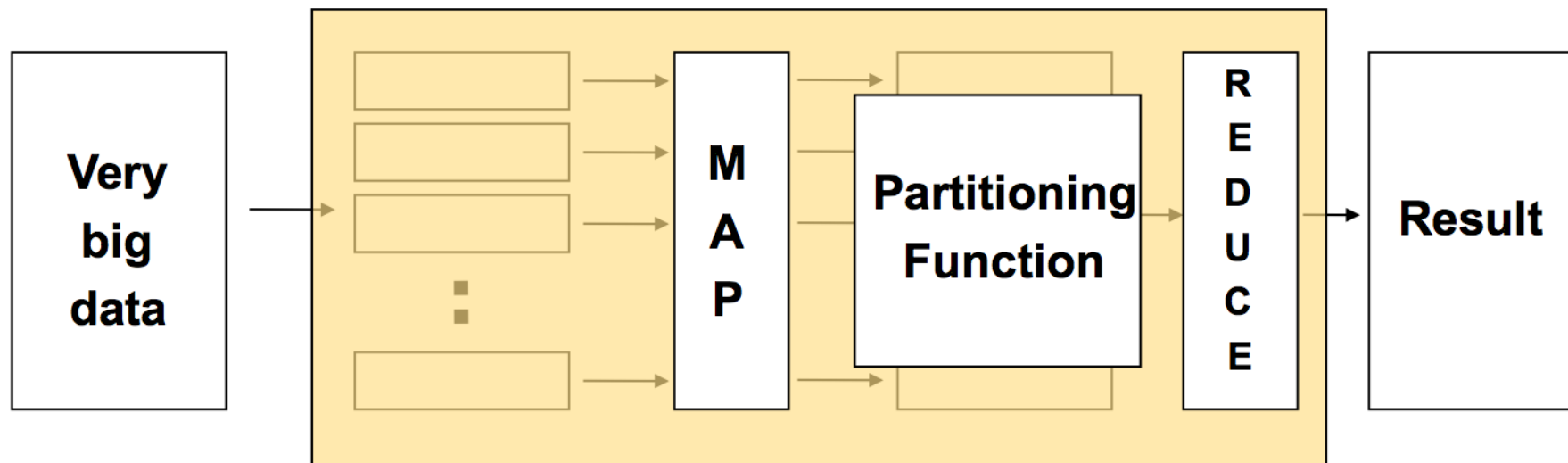
in one survey were unsure of how much of their data was inaccurate

# Veracity

UNCERTAINTY OF DATA



# MAP REDUCE



- Map:

- ★ Accepts *input* key/value pair
- ★ Emits *intermediate* key/value pair

- Reduce :

- ★ Accepts *intermediate* key/value\* pair
- ★ Emits *output* key/value pair



# Word Count

The Overall MapReduce Word Count Process

edureka!

