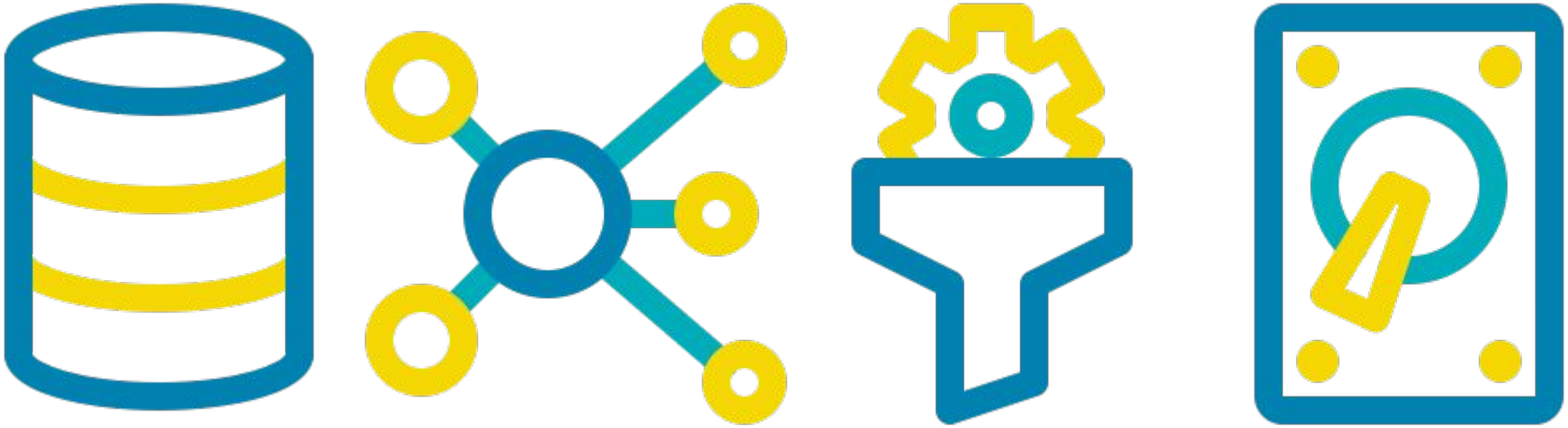
Overview of Big Data (3Vs)

(or is it 4Vs …)

Suzanne Little

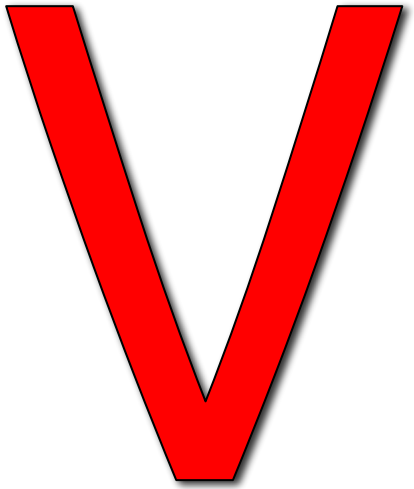
You may have heard the term but just how big is “big data”?



Big data

Characterised by 3 ‘V’s

- Volume
- Variety
- Velocity



a 4th V is sometimes added …

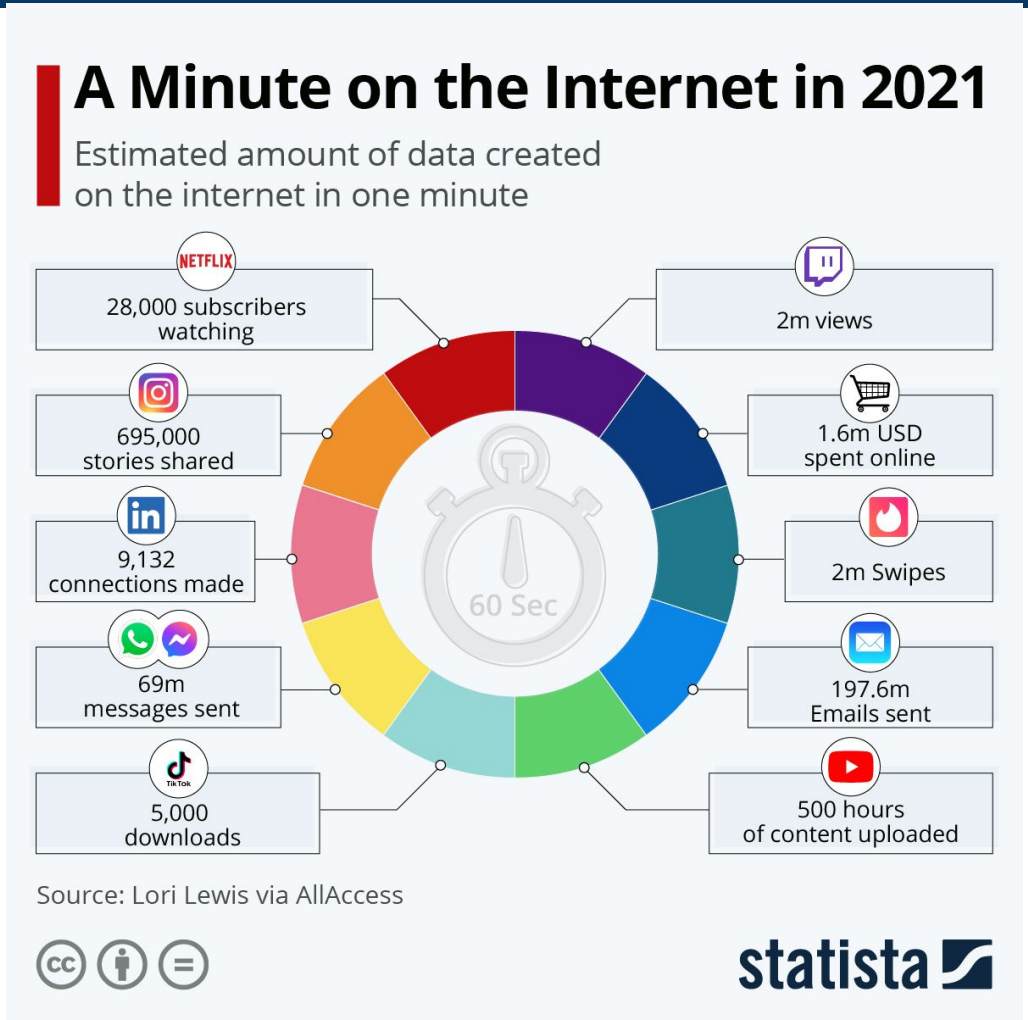
April 1st post 42 Vs of Big data!

Big data: Volume

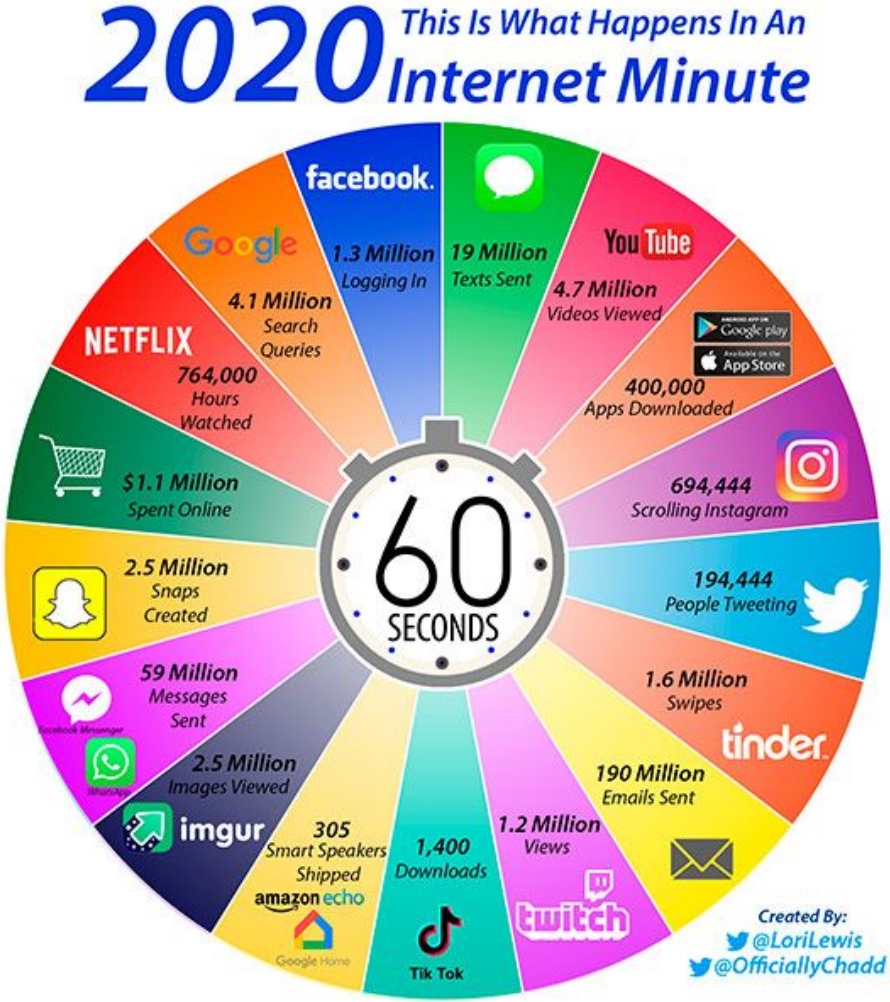
- Refers to the **amount** of data
- For big data, varies from terabytes to petabytes to zetabytes
- Can you open the whole dataset in your PC? Probably not big.
- In 2008, Google was already processing 20,000 terabytes of data (20 petabytes) a day
- In 2018, Google processes 40,000 searches per second!
- Social media produces vast quantities of data per minute!

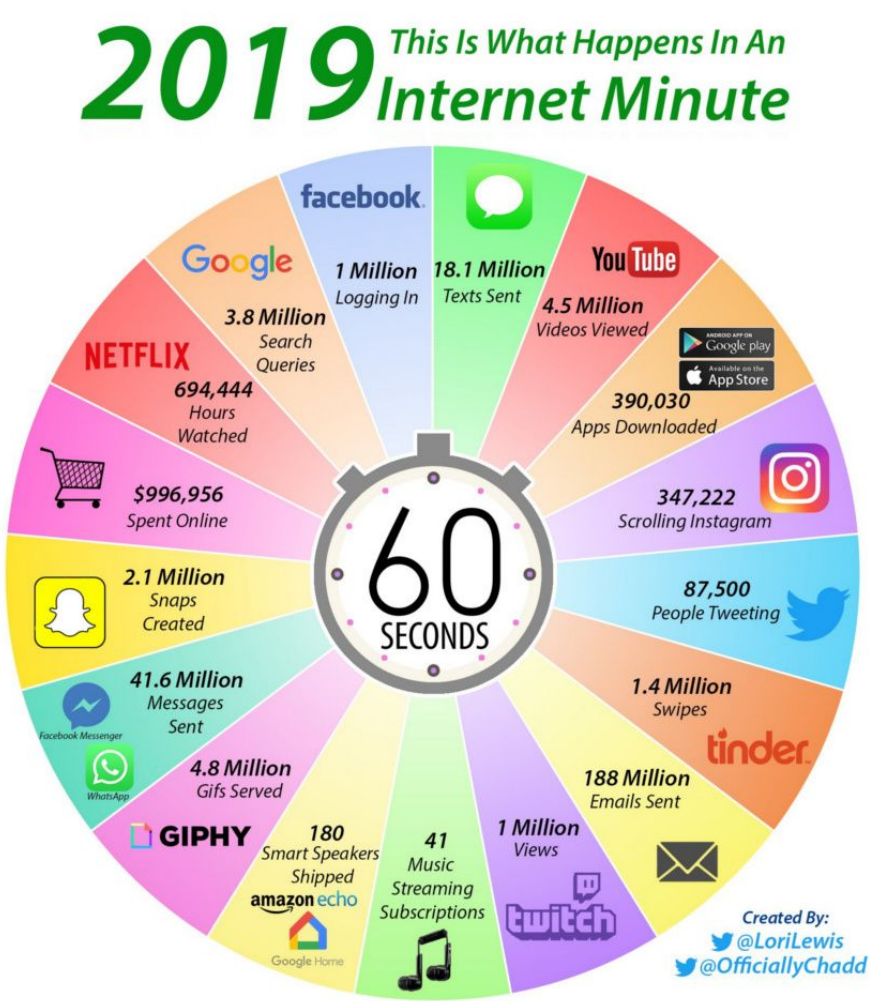
https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#672fd38760ba


AIRBNB
GUESTS BOOK
747 STAYS
AMAZON
SHOPPERS SPEND
$455K
X
USERS SEND
360K
TWEETS
6.3M
SEARCHES
HAPPEN ON
GOOGLE
WHATSAPP
USERS SEND
41.6M
MESSAGES
LINKEDIN
USERS SUBMIT
6,060
RESUMES
VIEWERS WATCH
43 YEARS
OF STREAMING
CONTENT
3,720
USERS DOWNLOAD
INSTAGRAM
THREADS
CYBER-
CRIMINALS
LAUNCH 30 DDOS ATTACKS
EVERY MINUTE
01:00
OF THE DAY
PRESENTED BY
DOMO
CHATGPT
USERS SEND
6,944 PROMPTS
INSTAGRAM
USERS SEND
694K REELS VIA DM
FACEBOOK
USERS LIKE
4M POSTS
DOORDASH
DINERS PLACE
$122K
IN ORDERS
FANS STREAM A
TAYLOR SWIFT
SONG 69.4K TIMES
VENMO
USERS SEND
$463K
IN PAYMENTS
GLOBAL
INTERNET
USERS SPEND
25.1M
HOURS ONLINE
THE AVERAGE
PERSON
PRODUCES
102 MB
OF DATA
PEOPLE TRADE
$398M
IN TREASURY
BONDS
PEOPLE SEND
241M
EMAILS
TWITCH
USERS WATCH
48K
HOURS OF CONTENT



https://www.statista.com/chart/25443/estimated-amount-of-data-created-on-the-internet-in-one-minute/


2020 This Is What Happens In An Internet Minute
facebook.
1.3 Million Logging In
19 Million Texts Sent
YouTube
4.7 Million Videos Viewed
400,000 Apps Downloaded
694,444 Scrolling Instagram
194,444 People Tweeting
1.6 Million Swipes
tinder
190 Million Emails Sent
1.2 Million Views
twitch
1,400 Downloads
Tik Tok
305 Smart Speakers Shipped
amazon echo
Google Home
2.5 Images Viewed
imgur
59 Million Messages Sent
2.5 Million Snaps Created
$1.1 Million Spent Online
NETFLIX
764,000 Hours Watched
4.1 Million Search Queries
Google
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd


2019 This Is What Happens In An Internet Minute
facebook
1 Million Logging In
18.1 Million Texts Sent
YouTube
4.5 Million Videos Viewed
390,030 Apps Downloaded
347,222 Scrolling Instagram
87,500 People Tweeting
1.4 Million Swipes
tinder
188 Million Emails Sent
1 Million Views
twitch
41 Music Streaming Subscriptions
180 Smart Speakers Shipped
amazon echo
Google Home
4.8 Million Gifs Served
GIPHY
41.6 Million Messages Sent
2.1 Million Snaps Created
$996,956 Spent Online
NETFLIX
694,444 Hours Watched
3.8 Million Search Queries
Google
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd

Big data: Variety

- Refers to **differing** types and data sources
- Structured, semi-structured and unstructured data
- Organisations may need to combine data from many different sources
- With the proliferation of analytics and sensor data, the variety of data is expanding rapidly
- Q: How many digital cameras do you own?

Big data: Velocity

- Data in motion **-- dynamic**, temporal
- We may need to process data as it arrives
  - ...because we cannot store such volumes
  - ...because we need timely processing
- Related notion of latency (lag-time)
  - The time between data being generated and processed
  - Some applications, such as fraud detection are highly time-sensitive.

Big data: The 4th V - Veracity

- Introducing the notion of data **uncertainty**
  - Veracity refers to how reliable the data is
- Is the data correct?
- Is it out-of-date?
- Is it complete?
- Data cleansing helps greatly
  - Using techniques such as data fusion, stochastic models
  - But with the huge volumes of data being generated – errors will slip in

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**

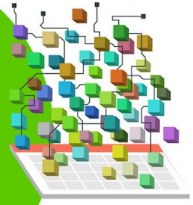
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs,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

Volume
SCALE 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



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

Variety
DIFFERENT FORMS OF DATA

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

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

Velocity
ANALYSIS OF STREAMING DATA

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

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

Veracity
UNCERTAINTY OF DATA

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

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

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

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

http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Resources

Video: Introduction to Big Data, O'Reilly

Caesar Wu and Rajkumar Buyya and Kotagiri Ramamohanarao (2016) "Big Data Analytics = Machine Learning + Cloud Computing", https://arxiv.org/abs/1601.03115 or version at https://learning.oreilly.com/library/view/big-data/9780128093467/B9780128053942000015.xhtml