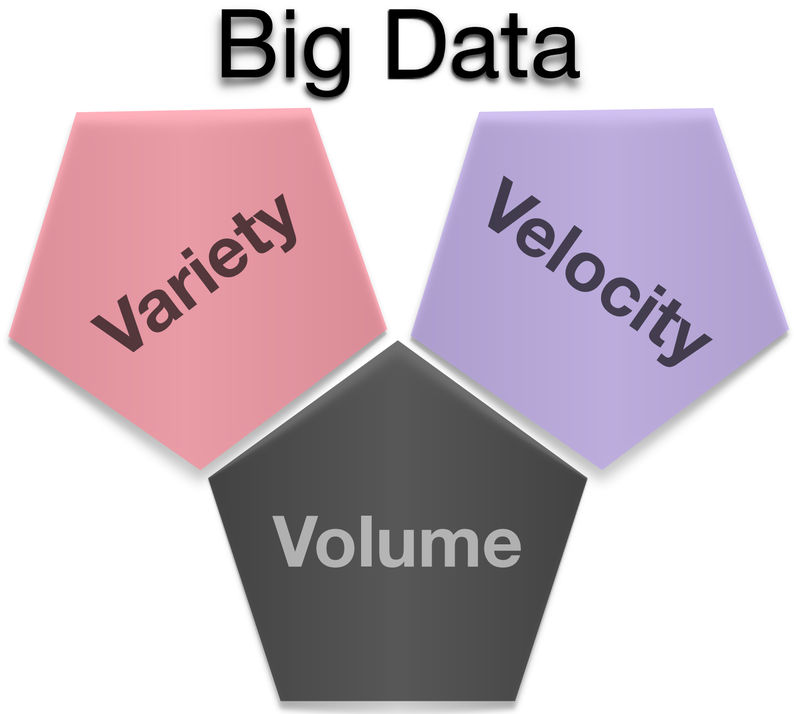
What Is Big Data?

- Big Data is all about large and complex data sets, which can be both structured and unstructured.
- It is impossible to “shovel” Big Data with structuring and analytics.
- It is a special field for a set of technologies to:
  - Obtain and clean
  - Process and analyze
  - Visualize and communicate
  - A large amount of structured and unstructured information

Big Data and Vs

- The initial definition of Big Data, was 3 Vs.
  - Volume
  - Variety
  - Velocity



Volume

- The amount of data:
  - The most straightforward V for big data.
  - How much data we have now?
    - <1 ZB in 2010 (1ZB = 10^{21} bytes, or 10^{9} TB)
    - 33 ZB in 2018
    - 59 ZB in 2020 (was predicted to be 40 ZB)
    - 175 ZB in 2025.

~~Into~~ In the Digital Era

- People's daily lives
  - 4.66 billion active internet users worldwide - 59.5 % of the population as of January 2021[1].
  - 4.32 billion (92.6% of users) accessed the internet via mobile devices.
  - 55 million of tweets/day[2]
- Scientific discovery
  - LHC (the Large Hadron Collider): 90 PB/year, 25 PB/year extra for non LHC[3] (1PB = 10^6 GB)
  - LSST (Large Synoptic Survey Telescope)： 20 Tb/night, 15 PB/year

1: source: https://www.statista.com/statistics/617136/digital-population-worldwide/
2: source: https://www.oberlo.com/blog/twitter-statistics
3: https://home.cern/science/computing/storage

Variety

- The structure of data
  - Structured data, is only part of data we have.
    - Structured text, tweets, pictures, videos
  - Semi-Structured data
    - Semantic Web: RDF, XML
  - Unstructured data
    - Simple text files, emails, voicemails, hand-written text, recordings, etc.

Velocity

- The speed of generating data
  - The New York Stock Exchange generates one TB of new trade data, per day.
  - Meta (Facebook) generates 500 TB of photo and video, messages, comments, per day.
  - A Jet Engine generate 10 TB, per 30 minutes of flight. The flights, generate many 1000 TB, per day.
- How fast the data is generated and processed to meet the demands, determines real potential.