**BigData One:**

● 1 terabyte (TB) = 1012 bytes = 1000 GB.

● 1 petabyte (PB) = 1015 bytes = 1 million GB.

● 1 exabyte (EB) = 1018 bytes = 1 billion GB.

● 1 zettabyte (ZB) = 1021 bytes = 1000 billion GB.

● 1 yottabyte (YB) = 1024 bytes = million billion GB.

Characterizing Big Data by various V’s:

**Volume: scale of data**

**Variety: different forms of data**

**Velocity: analysis of streaming data**

**Veracity: uncertainty of data**

Other V’s:

**Variability:**

**Visualization:**

**Value:**

**BigData Two:**

**Data science process:**

1. **Collecting and (pre)processing:** 
   1. **Structured data**
   2. **Unstructured data: natural language text. Video, audio**
   3. **Semi-structured data: JSON, XML, CSV**
2. **Cleaning and analyzing:**
   1. **Cleaning the data**
   2. **Exploratory data analysis**
   3. **Models & algorithms**
3. **Data product, communicating**

**Machine learning vs data mining:**

Data mining: Computational data analysis to find interesting properties from data. Emphasis is on analyzing current data.

Machine learning: A class of computational analysis methods that build (and update) a

general data model based on known data. Emphasis is on analyzing future data. Two aspects, training or fitting and prediction.

**Hadoop Distributed File System (HDFS)：**

Amdahl's law:

maximum parallel speedup using n computers: 1 / (f + (1 - f) / n), f is the fraction of code that can not be parallelized.