UNIT I
The V's of Big Data

The V's of Big Data

- Big Data is also **data** but with a **huge size**.

- Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time.

- In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

- No single standard definition…

"**Big Data**" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

- Velocity
- Volume
- Value
- Variety
- Veracity

Velocity

- First let's talk about velocity.
  - Obviously, velocity refers to the speed at which vast amounts of data are being generated, collected and analyzed.
- Every day the number of emails, twitter messages, photos, video clips, etc. increases at lighting speeds around the world.
- Every second of every day data is increasing.

- Not only must it be analyzed,
  - but the speed of transmission, and access to the data must also remain instantaneous to allow for real-time access to website, credit card verification and instant messaging.
- Big data technology allows us now to analyze the data while it is being generated, without ever putting it into databases.
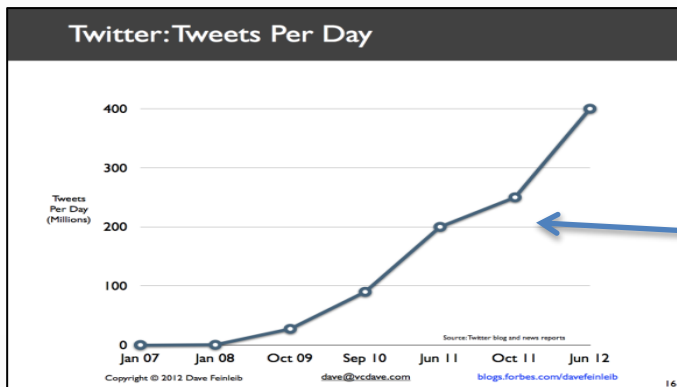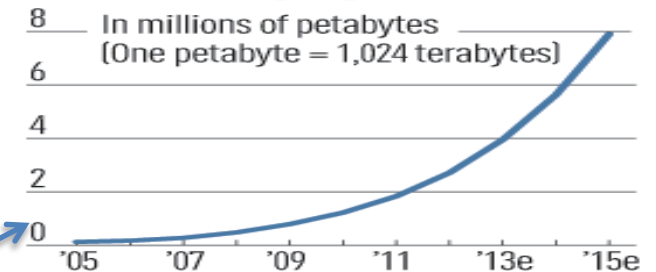
- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially

The Digital Universe 2009-2020

Growing By A Factor Of 44

2009: 0.8 Zb

2020: 35.2 Zettabytes

Data storage growth

In millions of petabytes
(One petabyte = 1,024 terabytes)

Twitter: Tweets Per Day

Exponential increase in collected/generated data

Value

- When we talk about value, we're referring to the worth of the data being extracted.

- Having endless amounts of data is one thing, but unless it can be turned into value it is useless.

- The most important part of embarking on a big data initiative
  - is to understand the costs and benefits of collecting analyzing the data
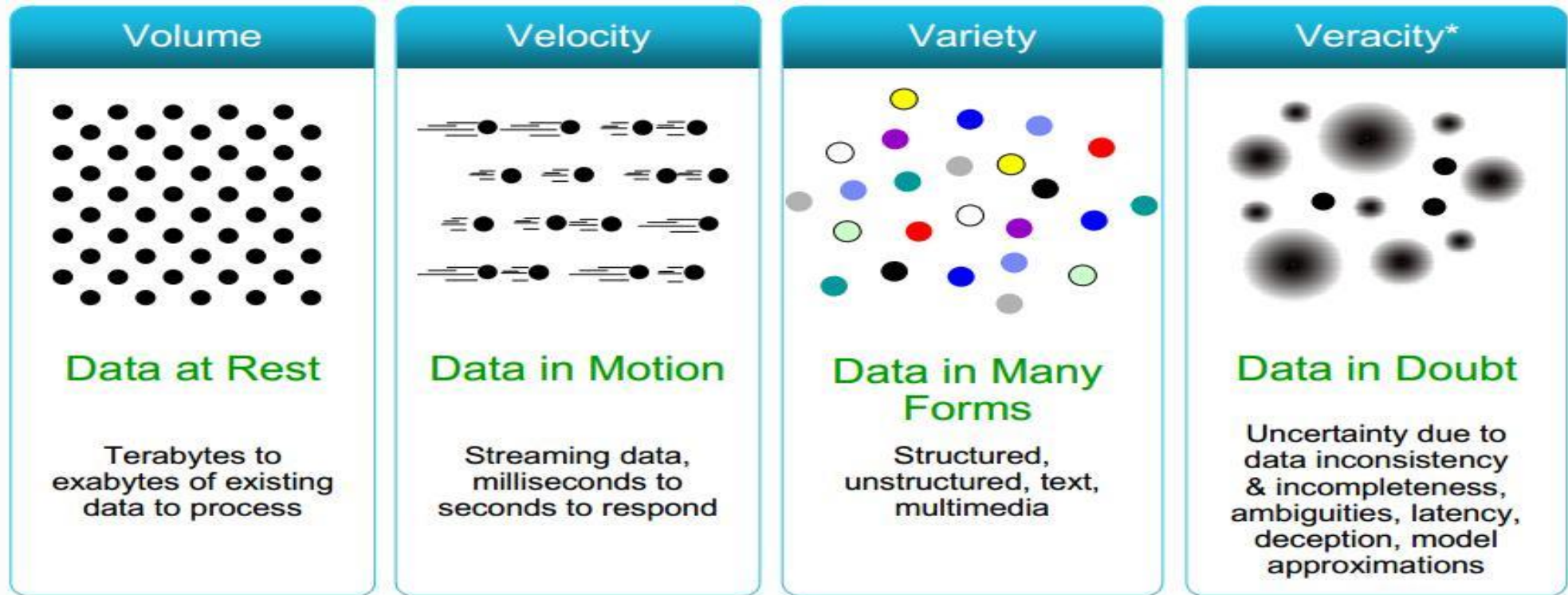 - to ensure that ultimately the data that is reaped can be monetized.

- Variety is defined as the different types of data we can now use.
- Data today looks very different than data from the past.
- We no longer just have structured data (name, phone number, address, financials, etc) that fits nice and neatly into a data table.
- Today's data is unstructured. In fact, 80% of all the world's data fits into this category, including photos, video sequences, social media updates, etc.
- New and innovative big data technology is now allowing structured and unstructured data to be harvested, stored, and used simultaneously.

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …

- Streaming Data
  - You can only scan the data once

- A single application can be generating/collecting many types of data

- Big Public Data (online, weather, finance, etc)

- Veracity is the quality or trustworthiness of the data. Just how accurate is all this data?
  - For example, think about all the Twitter posts with hash tags, abbreviations, typos, etc., and the reliability and accuracy of all that content. Gleaning loads and loads of data is of no use if the quality or trustworthiness is not accurate.
  - Another good example of this relates to the use of GPS data. Often the GPS will "drift" off course as you peruse through an urban area. Satellite signals are lost as they bounce off tall buildings or other structures. When this happens, location data has to be fused with another data source like road data, or data from an accelerometer to provide accurate data.

Models

- https://www.coursehero.com/file/p75930n8/Volume-Scale-Data-Volume-44x-increase-from-2009-2020-From-08-zettabytes-to-35zb/

Thank You