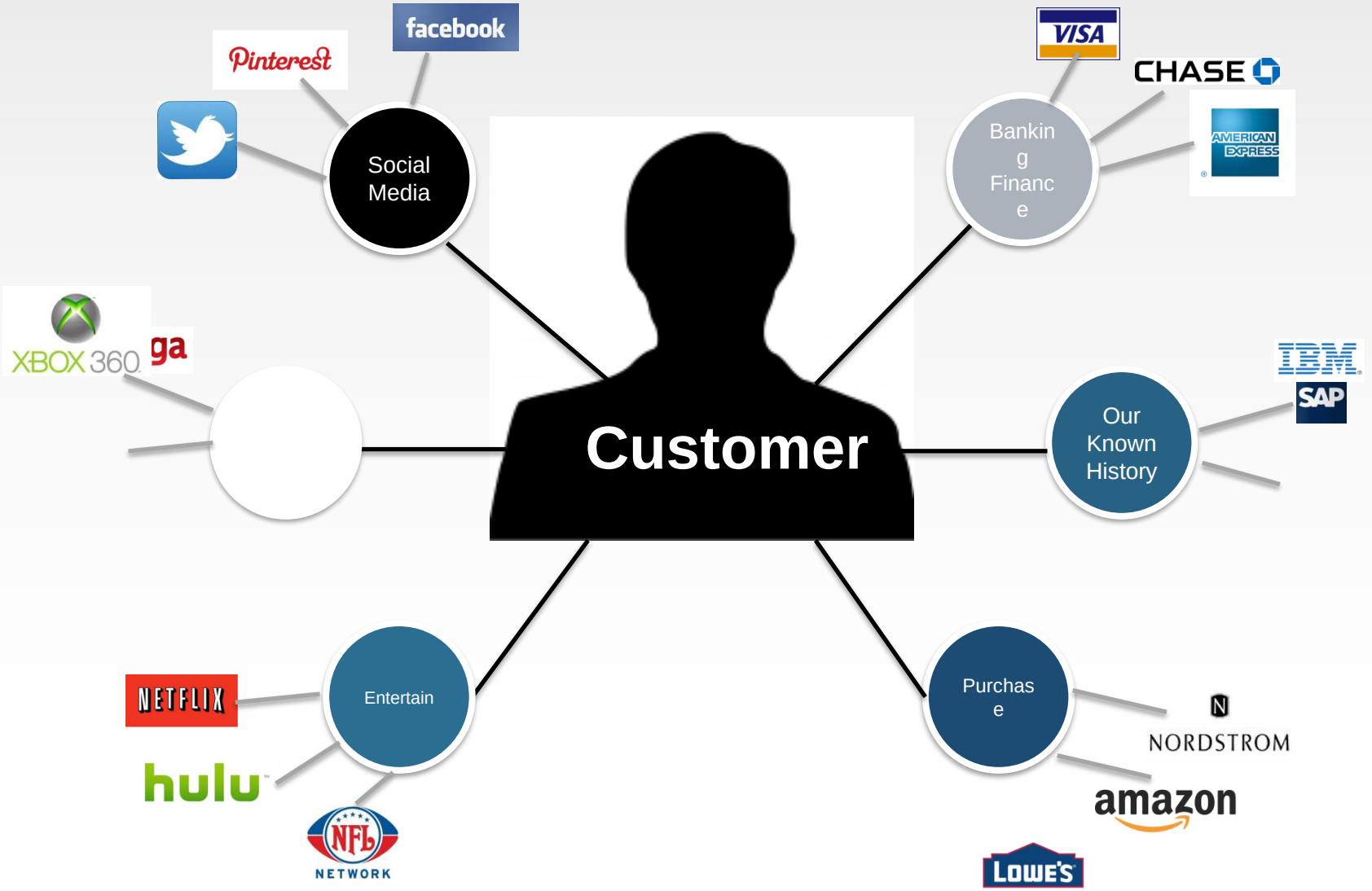
Big Data

What's Big Data?

- **No single definition**
- Collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Purchases at department/grocery stores
  - Bank/Credit Card transactions
  - Social Networks
- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

A Single View to the Customer



Types of Big Data

- Big Data' could be found in three forms:
  - **Structured**
  - **Unstructured**
  - **Semi-structured**
- It can also be classified as:
  - Relational Data (Tables/Transaction/Legacy Data)
  - Text Data (Web)
  - Semi-structured Data (XML)
  - Graph Data
    - Social Network, Semantic Web (RDF), …
  - Streaming Data

Structured Data

- Any data that can be stored, accessed and processed in the form of fixed format

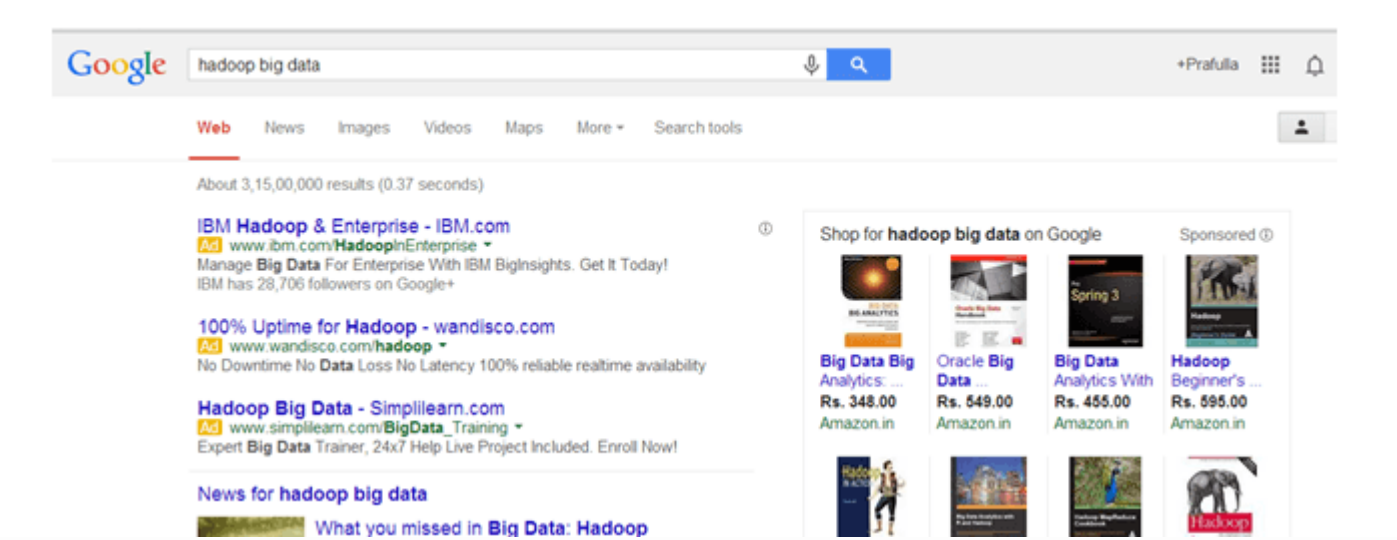
  e.g. Data of employees in a database

Employee_ID	Employee_Name	Gender	Department	Salary
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000

Unstructured Data

- Any data with unknown form or unknown structure

  e.g. Google search output



Semi-structured Data

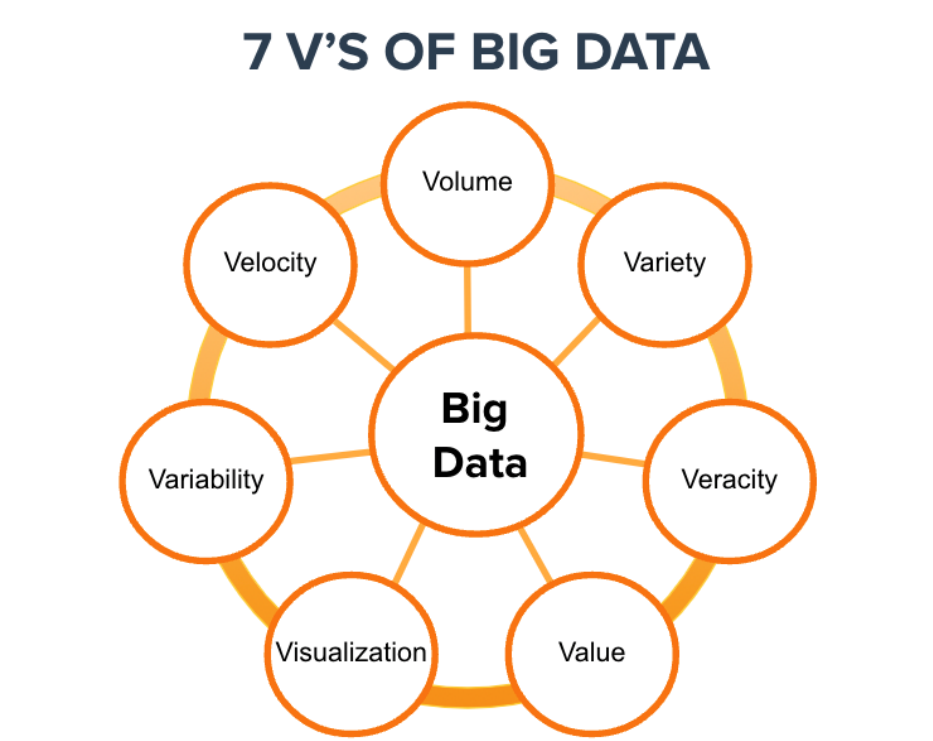
- Can be obtained from structured and unstructured data

  e.g. Data represented in an XML file.

```
<rec><name>PrashantRao</name><gender>Male</gender><age>35</age></rec>
<rec><name>Seema R.</name><gender>Female</gender><age>41</age></rec>
<rec><name>Satish Mane</name><gender>Male</gender><age>29</age></rec>
 <rec><name>Subrato Roy</name><gender>Male</gender><age>26</age></rec>
 <rec><name>Jeremiah J.</name><gender>Male</gender><age>35</age></rec>
```

Characteristics of Big Data (8V's)

- Volume, Velocity, Variety, Veracity (+ Value, Variability and Visualization)



Volume:

- The name Big Data itself is related to a size which is enormous.
- Size of data plays a crucial role in determining value out of data.
- Also, whether a particular data can be considered as a Big Data or not, is dependent upon the volume of data.
- Hence, '**Volume**' is one characteristic which needs to be considered while dealing with Big Data.

Velocity:

- Big Data Velocity deals with the speed at which data flows in from sources like
  - business processes, application logs, networks, social media sites, sensors, Mobile devices, etc.

- How fast the data is generated and processed
  - to meet the demands
  - to determine the real potential of the data.
  - to get the insights and predictions

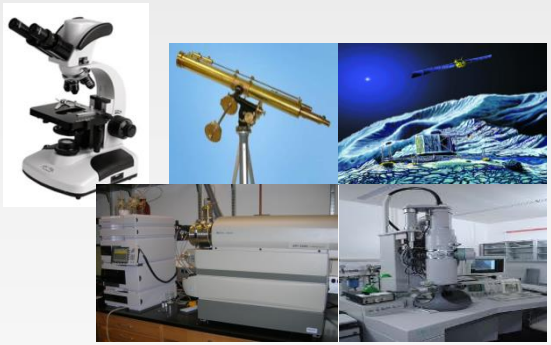
Velocity

- Online Data Analytics
- Late decisions ➔ missing opportunities

- **Examples**
  - **ePromotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

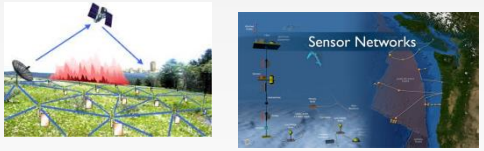
Real-time/Fast Data

Social media and networks
(all of us are generating data)



Scientific instruments
(collecting all sorts of data)

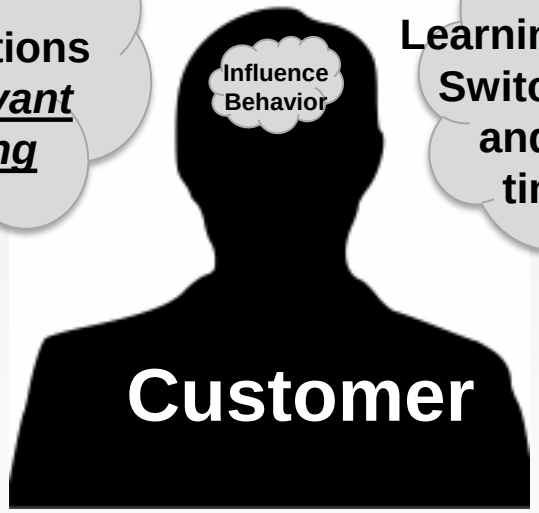
Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

- Collecting the data is not big challenge now
- But the ability to manage, analyze, summarize, visualize, and discover knowledge from it in a timely manner and in a scalable fashion

Real-Time Analytics/Decision Requirement



Product Recommendations that are ***Relevant*** & ***Compelling***

Learning why Customers Switch to competitors and their offers; in time to Counter

Improving the Marketing Effectiveness of a Promotion while it is still in Play

Friend Invitations to join a Game or Activity that expands business

Preventing Fraud as it is ***Occurring*** & preventing more proactively

Variety of Data

- Is because of heterogeneous sources and the nature of data, both structured and unstructured.
- During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications.
- Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications.
- This variety of unstructured data poses certain issues for storage, mining and analyzing data.

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

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

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

Veracity of Data

- Refers to the inconsistency
- Abnormalities, Noise, Trustworthiness, and accuracy of data

- **Value:** Ability to provide valuable insights and create value for businesses and organizations (Usefulness of the data to achieve a specific goal or objective)

- **Variability:** Variation in the data

- **Visualization:** How can we visualize the insights and predictions through graphs, charts, etc

Lifecycle of Big Data

- The general categories of activities involved with big data processing are:
  - Ingesting data into the system
  - Persisting the data in storage
  - Computing and Analyzing data
  - Visualizing the result

Benefits of Big Data Processing

- Businesses can utilize outside intelligence while taking decisions
- Improved customer service
- Early identification of risk to the product/services, if any
- Better operational efficiency