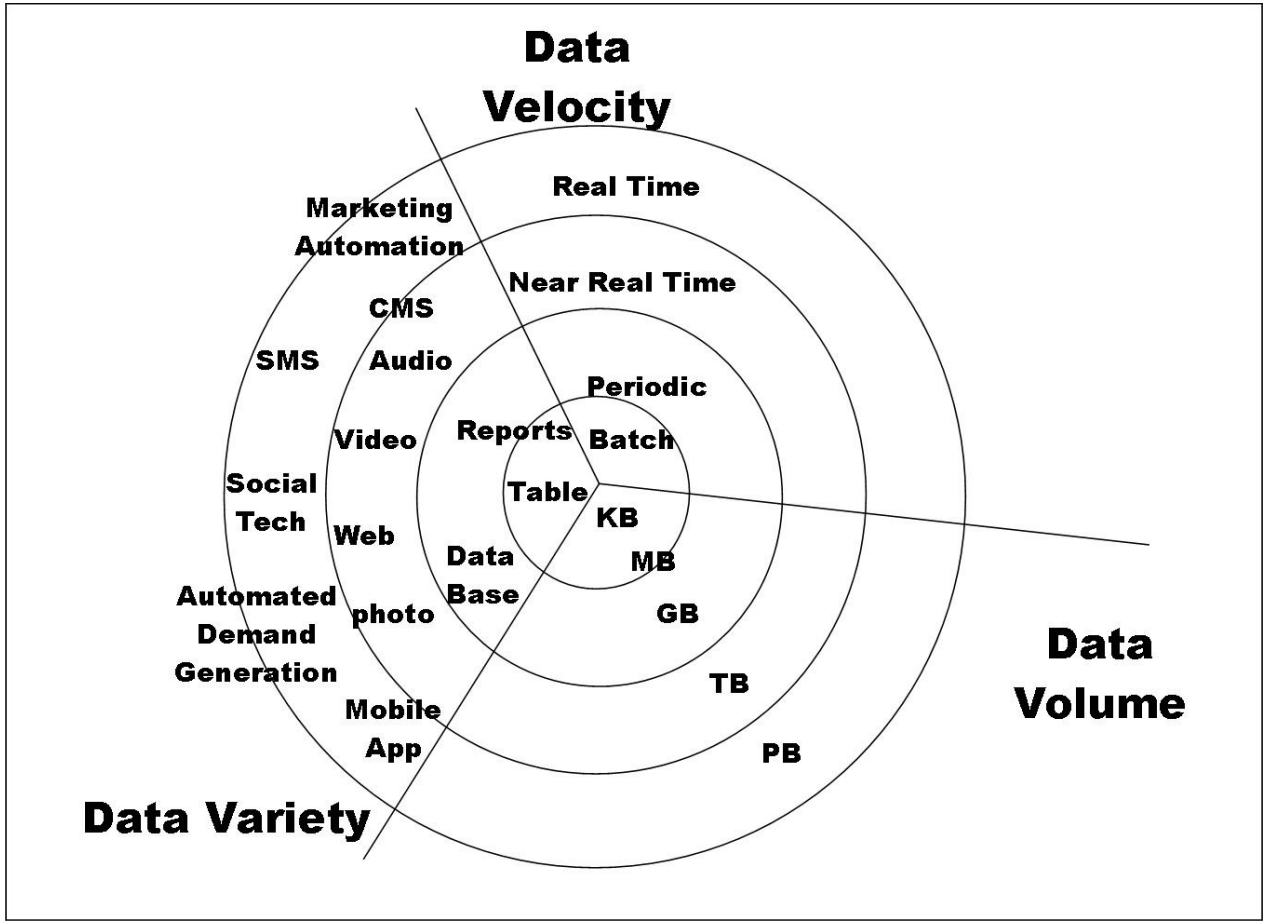
Unit I
Introduction to
Big Data Analytics

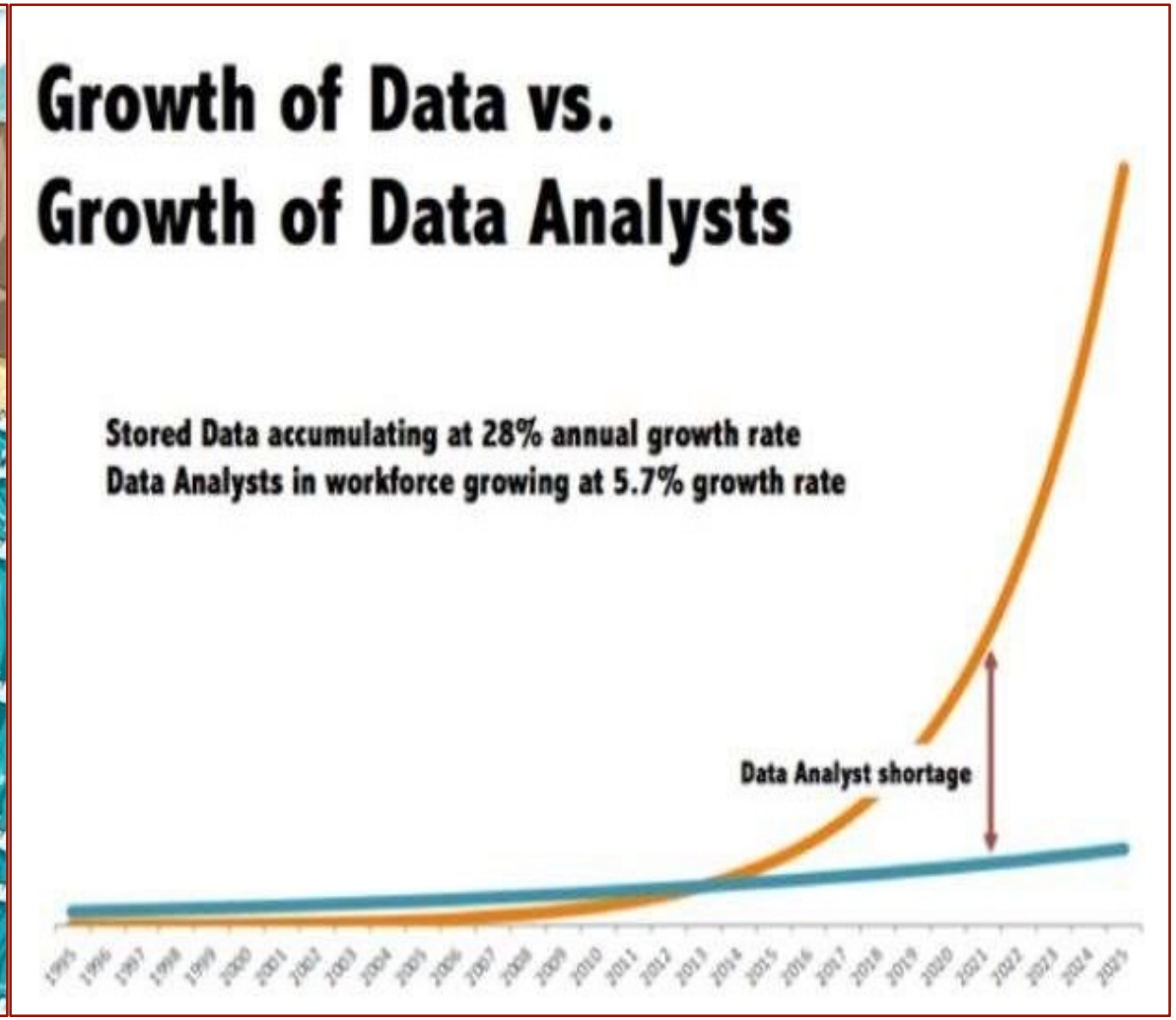
Dr Purnima Gandhi

Data vs. Big Data

Big Data characteristics



Why we are talking about Big data and data science?



Source: https://bit.ly/31HBHuQ

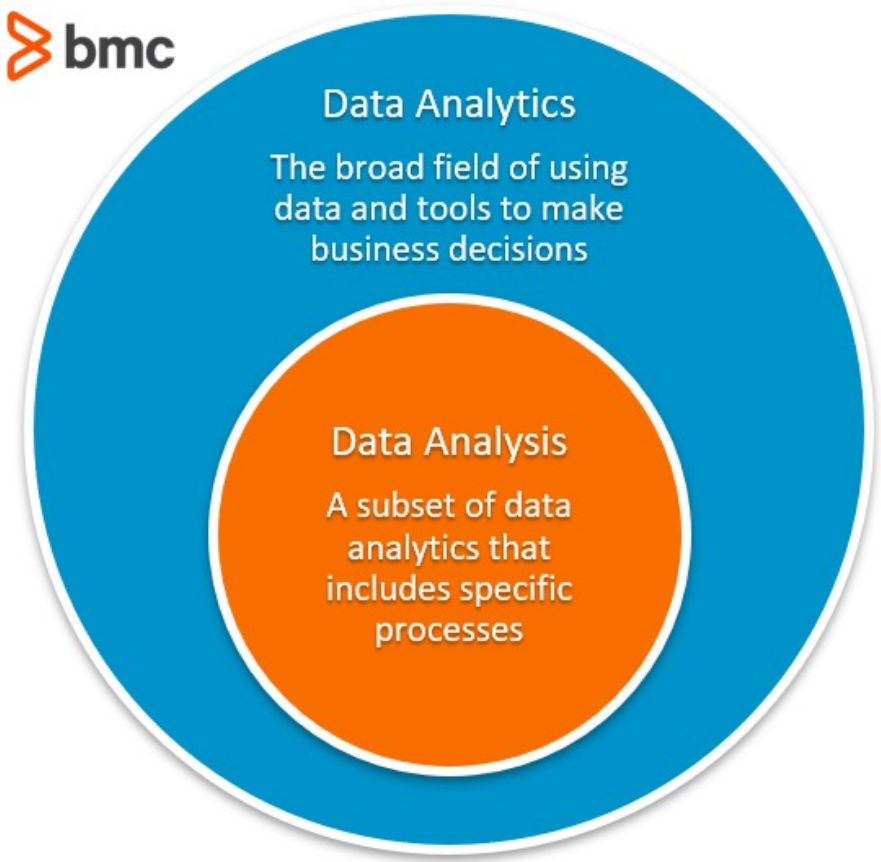
Traditional data vs. Big Data

	Traditional data	Big data
Volume	GB	Constantly updated (PB to TB currently..)
Data generated rate	Per hour, day, …	More rapid
Structure	Structured	Semi-structured , unstructured
Data source	Centralized	Fully distributed
Data integration	Easy	Difficult
Data store	RDBMS	HDFS, NoSQL
Access	Interactive	Batch or real time/ near real time

Analysis vs Analytics

- Mathematics and statistics
- Aggregation and Statistics
  - Data warehousing and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling
- Machine learning
- Artificial intelligence

Many more…….



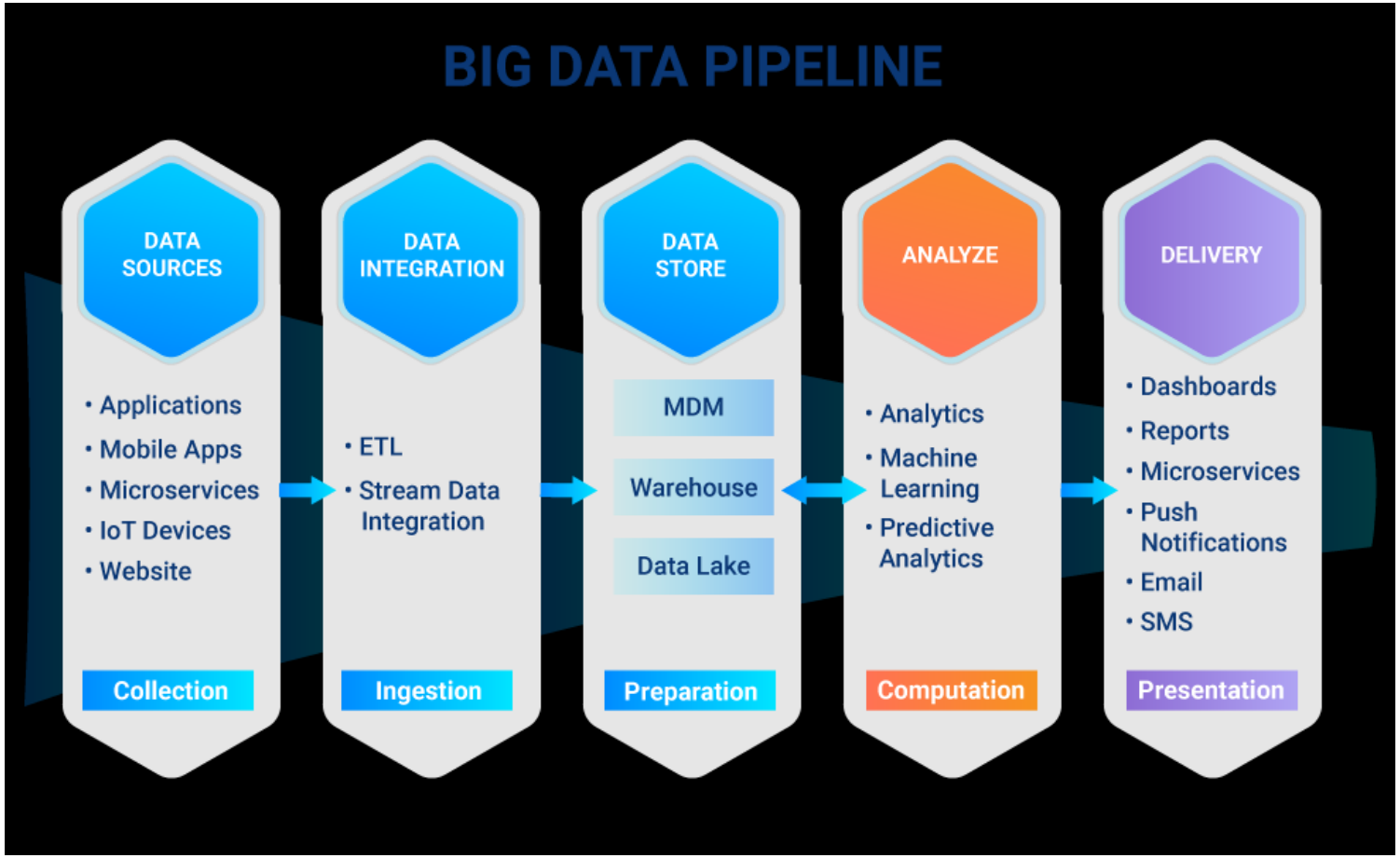
Big Data

- No single standard definition…
- "***Big Data***" is data
- Whose scale, diversity, and complexity require new architecture, techniques, algorithms, & analytics to manage it and extract value & hidden knowledge from it…
- "Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze." -The McKinsey Global Institute, 2012

Big Data pipeline

- Big Data Analytics is interdisciplinary and emerging technology
- BDA is not strait forward
- The term "data pipeline" describes a **set of processes** that move data from one place to another place. ... Big data pipelines can also use the same transformations and load data into a variety of depositories, including relational databases, data lakes, and data warehouses.

Big Data pipeline



A big data analytics cycle can be described by the following stage −

1. **Business Problem Definition**
2. **Data Identification**
3. **Data Acquisition & Filtering**
4. **Data Extraction**
5. **Exploratory Data Analysis**
6. **Data Preparation for Modeling and Assessment**
7. **Data Visualization**
8. **Analysis of Results**

What is data science?

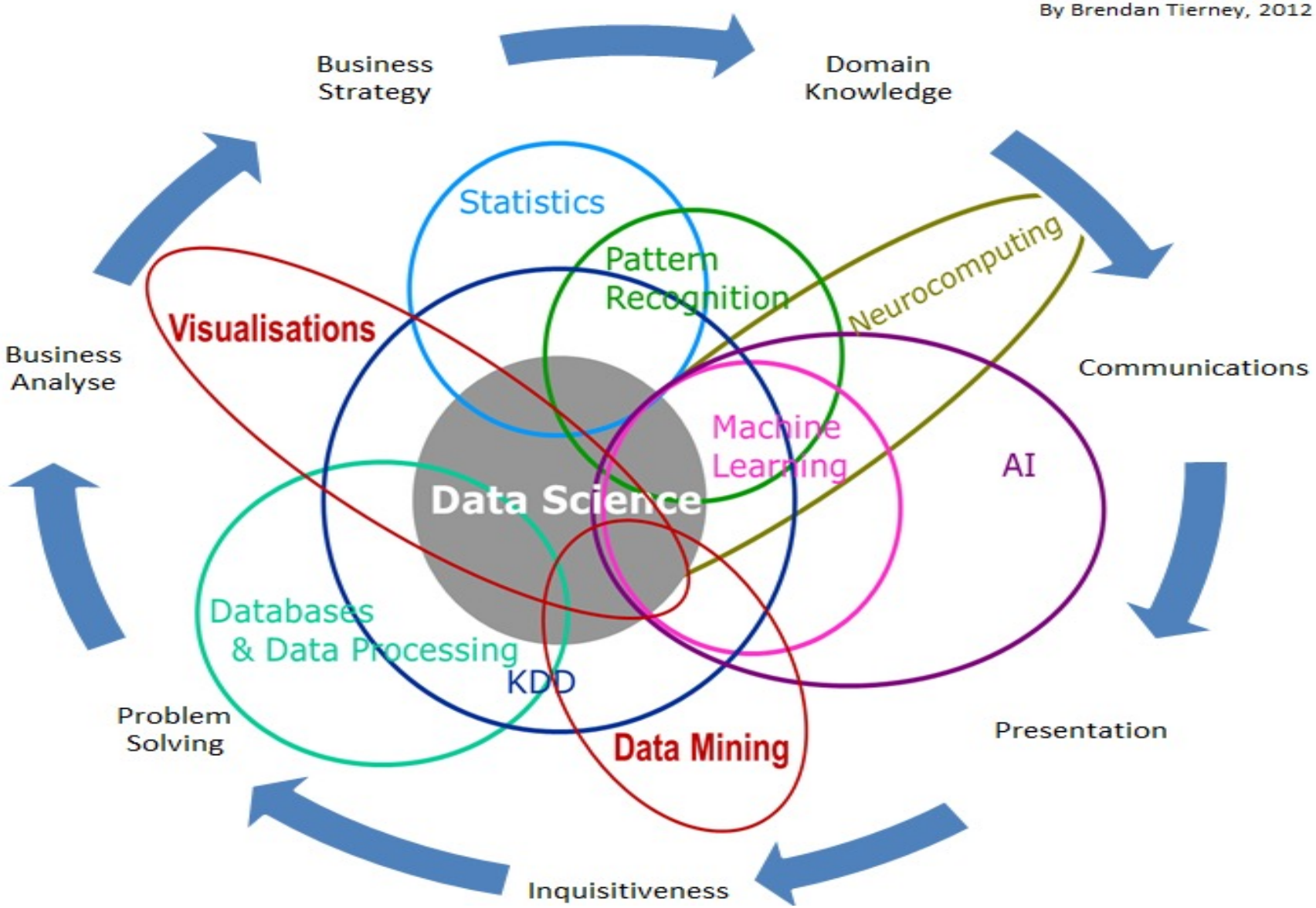
A multi-disciplinary and emerging field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

Goal - Turn data (Data science principles apply to all data – big and small) into data products and create business value.

Source: https://bit.ly/30dekJB

Data Science Is Multidisciplinary

By Brendan Tierney, 2012



Big Data and Data Science

- "… the sexy job in the next 10 years will be statisticians," Hal Varian, Google Chief Economist
- The U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018. McKinsey Global Institute's June 2011
- New Data Science institutes being created or repurposed – NYU, Columbia, Washington, UCB,...
- New degree programs, courses, boot-camps:
  - e.g., at Berkeley: Stats, I-School, CS, Astronomy…
  - One proposal (elsewhere) for an MS in "Big Data Science"