جامعة الإسكندرية
ALEXANDRIA UNIVERSITY
كلية الحاسبات وعلوم البيانات

LEC.2.

Introduction to Data Science

Data, Information, and Knowledge:

Data: are any facts, numbers, or text that can be processed by a computer. today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- **operational or transactional data** such as, sales, cost, inventory, payroll, and accounting
- **nonoperational data**, such as industry sales, forecast data, and macro economic data
- **meta data** - data about the data itself, such as logical database design or data dictionary definitions

- **Information**

  The patterns, associations, or relationships among all this data can provide information.

  For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

- **Knowledge**

  Information can be converted into knowledge about historical patterns and future trends.

  For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

What is

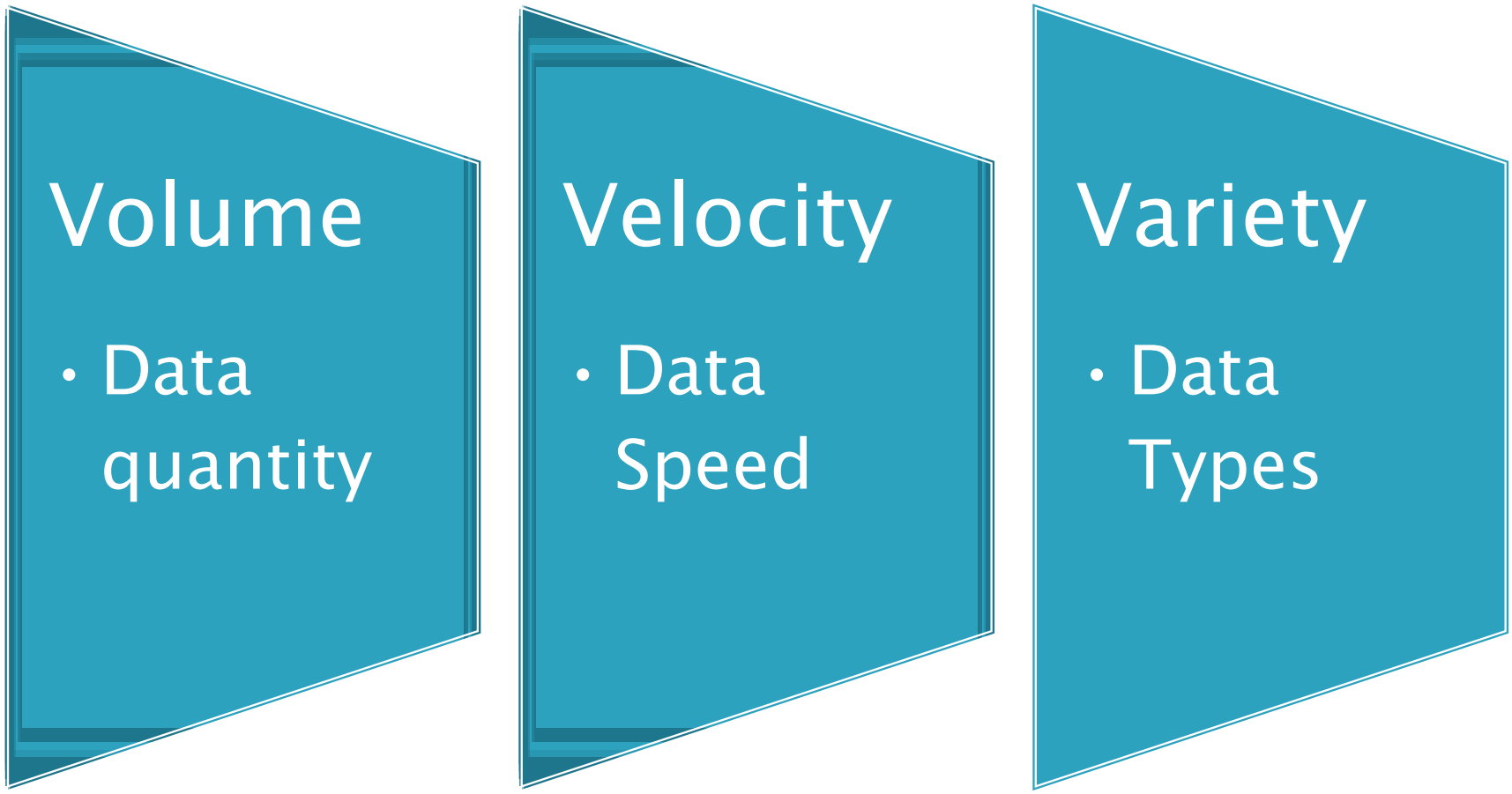
What makes data, "*Big" Data*?

BIG DATA IS BIG

Big data from its name is very big
Starting size of it at least 1 TB

Big data is the data characterized by 3 attributes: volume, variety and velocity."

Three Characteristics of Big Data V3s



1st Character of Big Data
Scale (**Volume)**

- A typical PC might have had 10 gigabytes of storage in 2000.

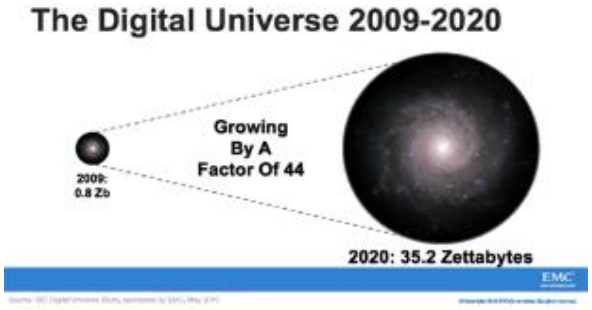
- Today, Facebook ingests 500 terabytes of new data every day.

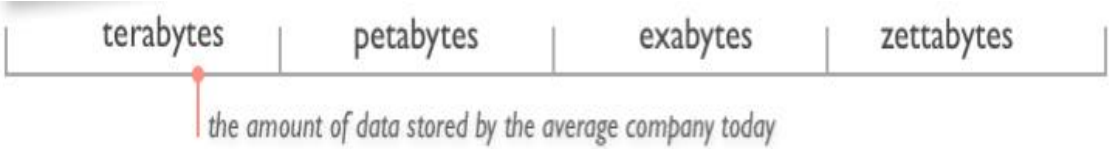
- Boeing 737 will generate 240 terabytes of flight data during a single flight across the US.

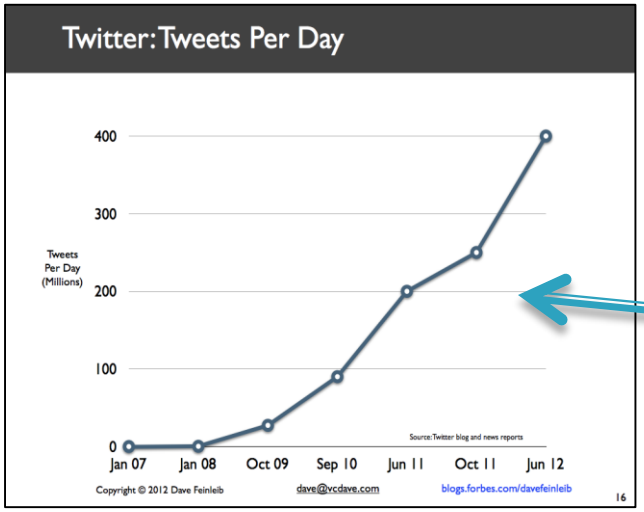
- The smart phones, the data they create and consume; sensors embedded into everyday objects will soon result in billions of new, constantly-updated data feeds containing environmental, location, and other information, including video.

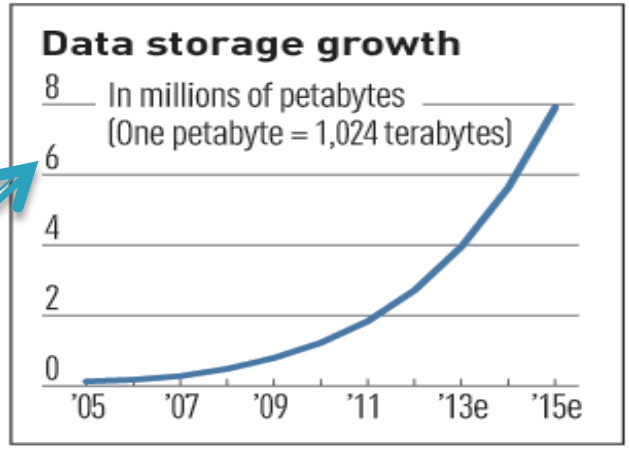
Scale (Volume)

- **Data Volume**
  - 44x increase from 2009 -2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially









Exponential increase in collected/generated data

2nd Character of Big Data
Speed (Velocity)

- Clickstreams and ad impressions capture user behavior at millions of events per second
- high-frequency stock trading algorithms reflect market changes within microseconds
- machine to machine processes exchange data between billions of devices
- infrastructure and sensors generate massive log data in real-time
- on-line gaming systems support millions of concurrent users, each producing multiple inputs per second.

Speed (Velocity)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- **Examples**
  - **E-Promotions:** 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

3rd Character of Big Data
Complexity (Varity)

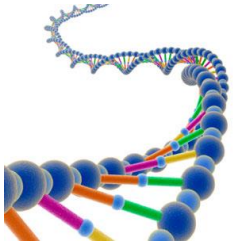
- Big Data isn't just numbers, dates, and strings. Big Data is also geospatial data, 3D data, audio and video, and unstructured text, including log files and social media.

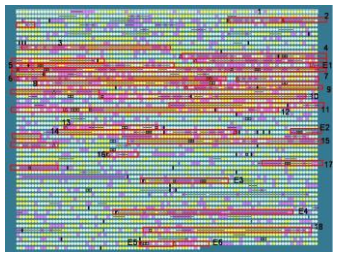
- Traditional database systems were designed to address smaller volumes of structured data, fewer updates or a predictable, consistent data structure.

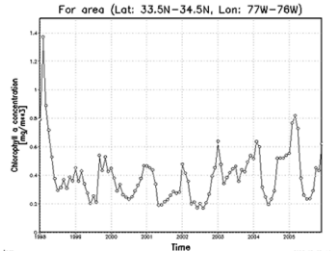
- Big Data analysis includes different types of data

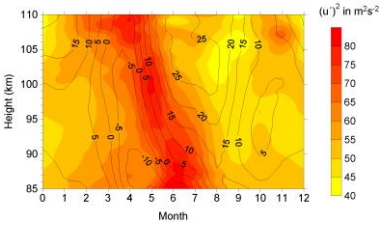
Complexity (Varity)

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data



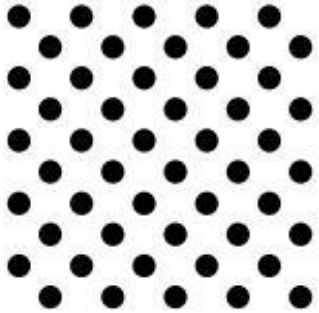
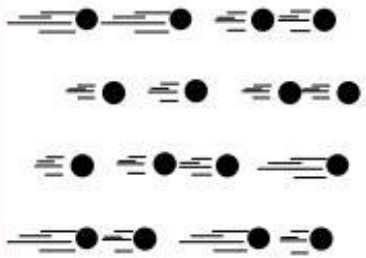
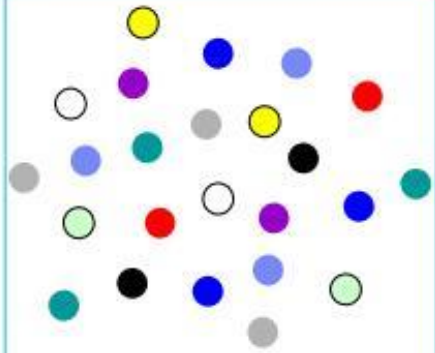
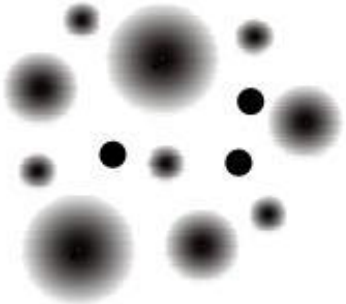






To extract knowledge➔ all these types of data need to linked together

Some Make it 4V's

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

What is “big data”?

- "Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization” (Gartner 2012)

- Complicated (intelligent) analysis of data may make a small data “appear” to be “big”

- Bottom line: Any data that exceeds our current capability of processing can be regarded as “big”

Sources of data

Data from internet

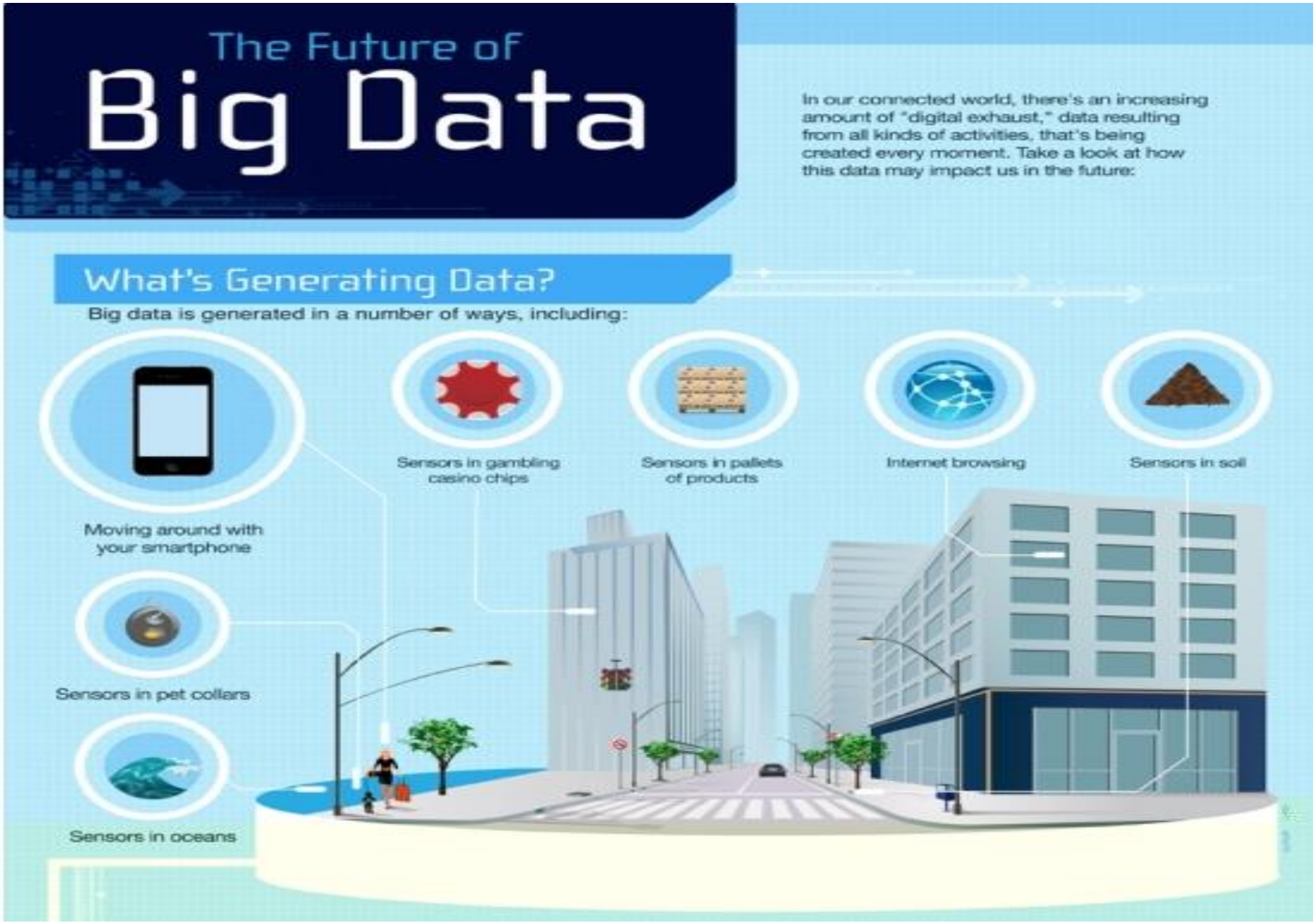
Data from military corporations

Hospitals data

NASA corporation data

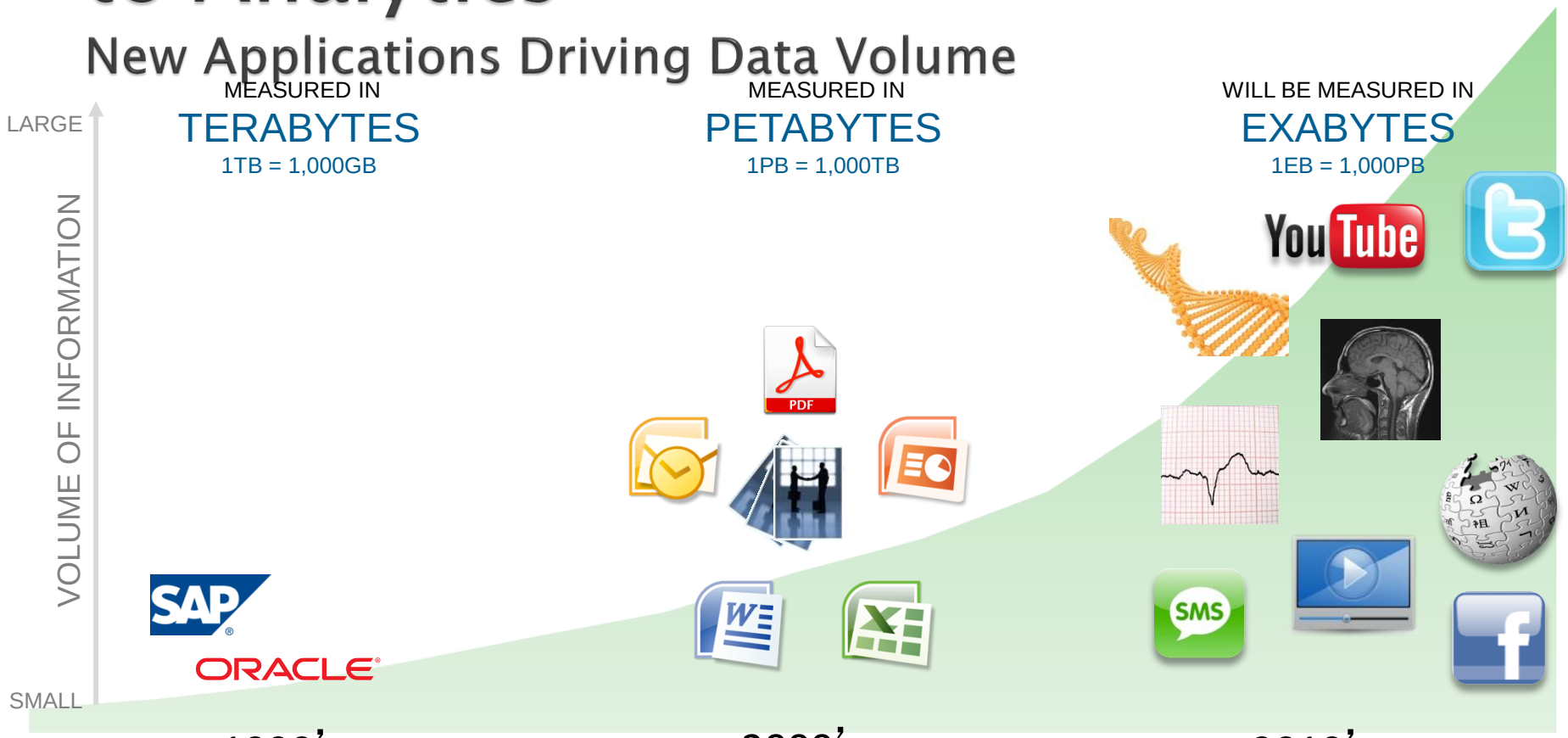
And so on...

Where is all this data coming from ?



Opportunities for a New Approach to Analytics

New Applications Driving Data Volume



1990's (RDBMS & DATA WAREHOUSE)

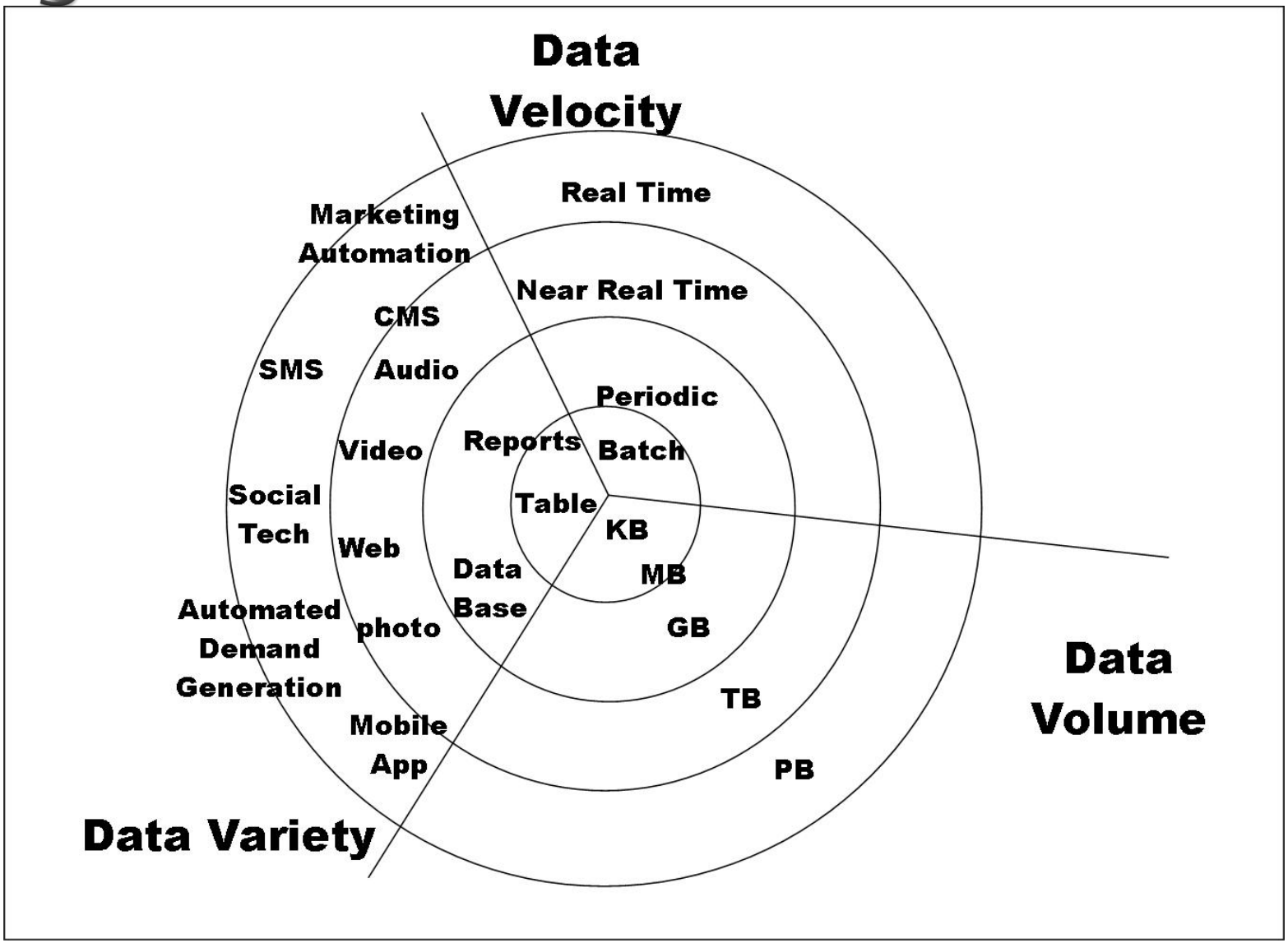
2000's (CONTENT & DIGITAL ASSET MANAGEMENT)

2010's (NO-SQL & KEY/VALUE)

Big Data

- Big Data is any data that is expensive to manage and hard to extract value from
  - Volume
    - The size of the data
  - Velocity
    - The latency of data processing relative to the growing demand for interactivity
  - Variety and Complexity
    - the diversity of sources, formats, quality, structures.

Big Data


Data
Velocity
Real Time
Near Real Time
Periodic
Batch
Marketing
Automation
CMS
SMS
Audio
Video
Reports
Table
Social
Tech
Web
KB
MB
GB
TB
PB
Data
Base
Automated
Demand
Generation
photo
Mobile
App
Data
Volume
Data Variety

Types of Data We Have:

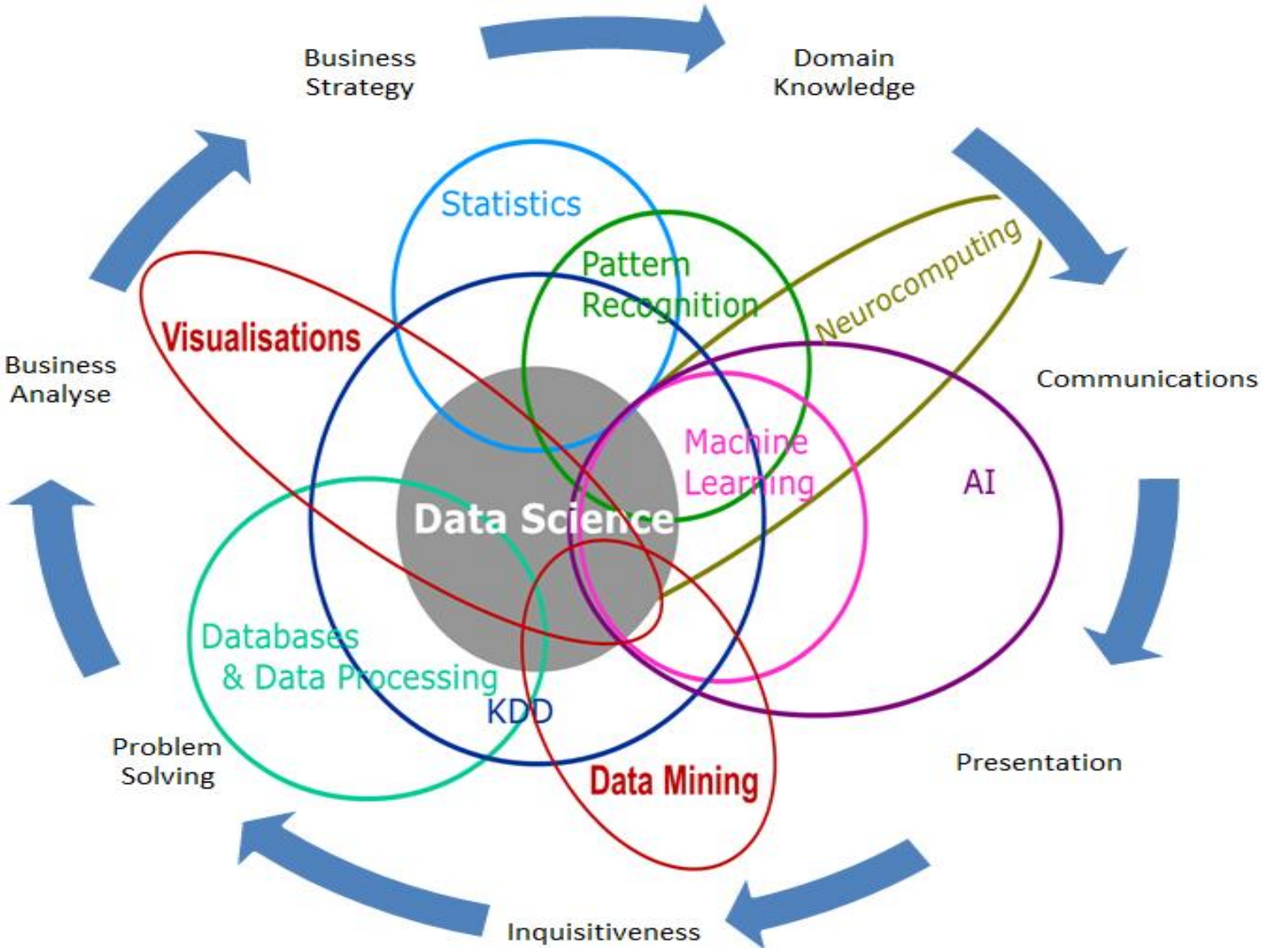
- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data

What is Data Science?

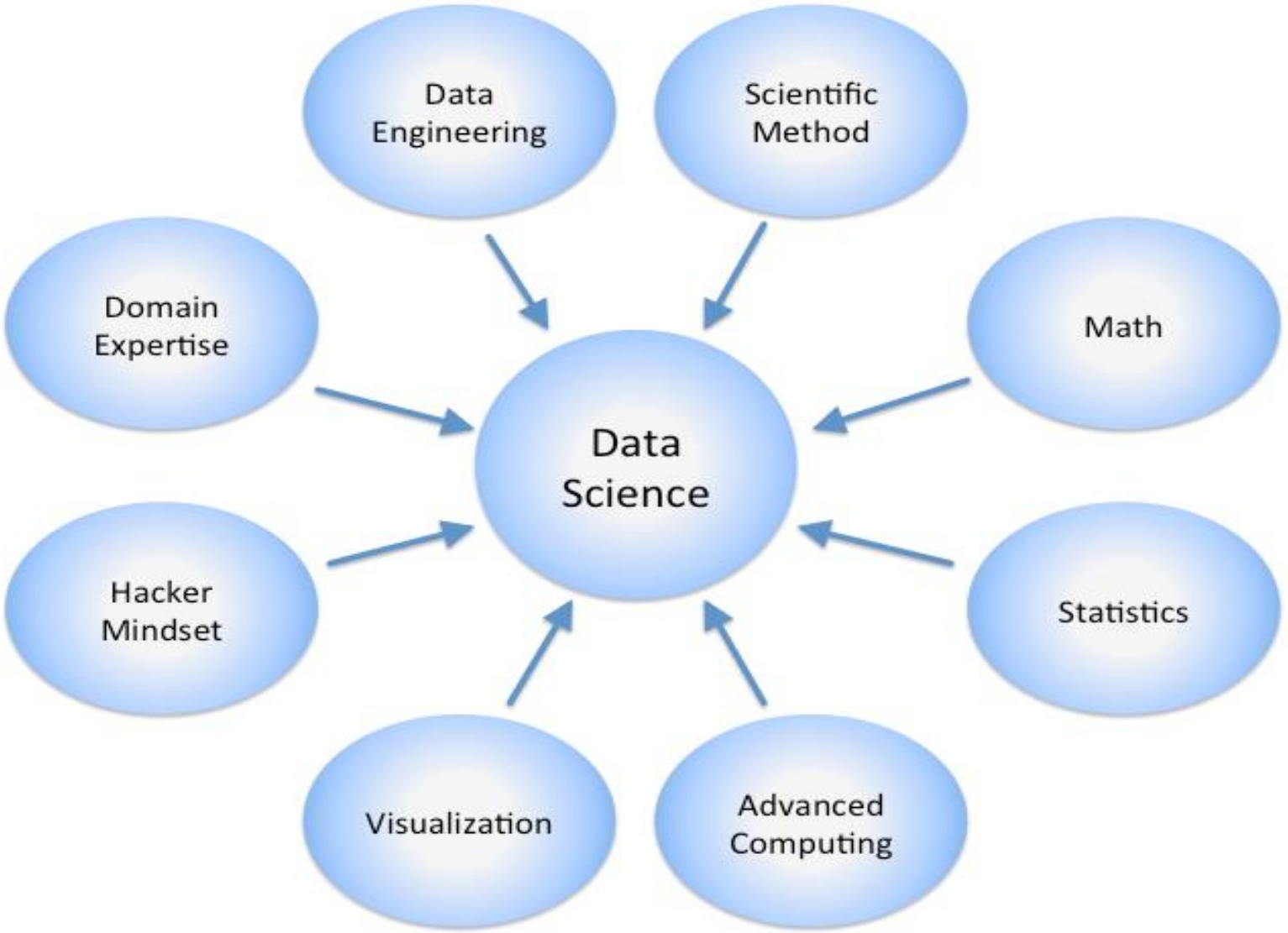
- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- Data science principles apply to all data – big and small

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

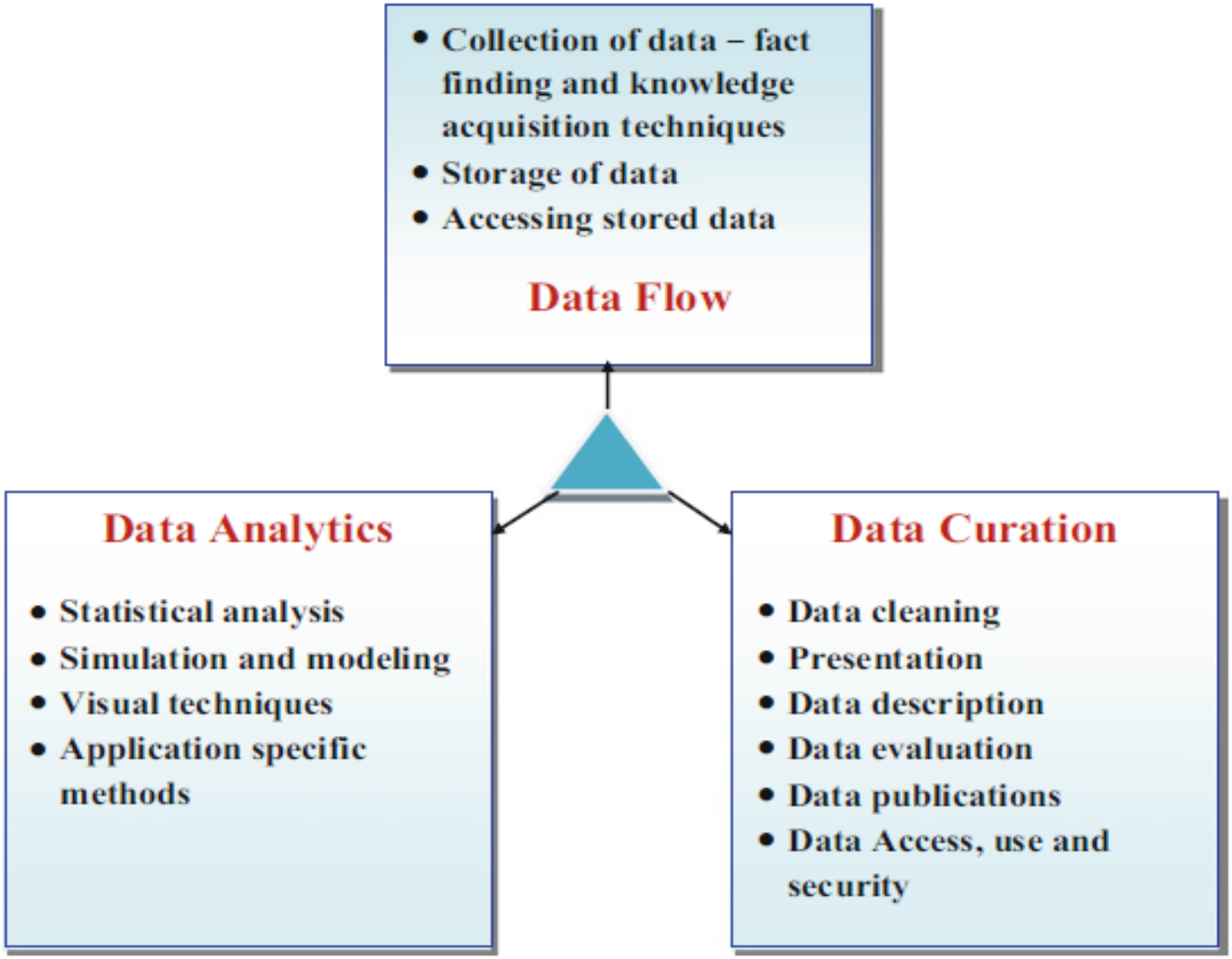
Data Science Is Multidisciplinary



Data Science



Data Science Activities:



Real Life Examples

- Companies learn your secrets, shopping patterns, and preferences

  For example, can we know if a person is diabetic, even if he/she doesn't want us to know?

- Data Science and election (2008, 2012) 1 million people installed the Obama Facebook app that gave access to info on "friends"