SAKARYA ÜNİVERSİTESİ

BIG DATA

TOO BIG TO IGNORE

SÜMEYYE KAYNAK

OUTLINE

The evolution of Big Data

What is Big Data?

Why Big Data Matters?

5V’s of Big Data

Lifecycle of Big Data

Big data analytics categories

Big data applications

Challenges of Big Data

GRADING SCALE

- Midterm 20%
- Assignment 20%
- Quiz 10%
- Final 50%

MATERIALS

Online material

- Stanford University, Computer Science Course
- Available online at: http://infolab.stanford.edu/~ullman/mmds.html

(Mining of Massive Datasets)

MATERIALS

Book

1. **Hands-on big data modelling**, James Lee, Tao Wei, Suresh Kumar Mukhiya
2. Big Data Fundamentals, Thomas Erl, Wajid Khattak, and Paul Buhler
3. An introduction to Data Science, Jeffrey Stanton, Syracuse University, 2013.
4. The white book of Big Data, Fujitsu, 2021
5. Mining of Massive Datasets Anand Rajaraman Kosmix, Inc. Jeffrey D. Ullman Stanford Univ.
6. R Programming for Data Science (2016), Roger D. Peng, https://bookdown.org/rdpeng/rprogdatascience/

PROJECT

1. **Hands-on big data modelling**, James Lee, Tao Wei, Suresh Kumar Mukhiya
2. Kaggle
3. Freelancing

PROJECT

1. **Hands-on big data modelling**, James Lee, Tao Wei, Suresh Kumar Mukhiya
2. Kaggle
3. Freelancing

MATERIALS

Book

1. **Hands-on big data modelling**, James Lee, Tao Wei, Suresh Kumar Mukhiya
2. Big Data Fundamentals, Thomas Erl, Wajid Khattak, and Paul Buhler
3. An introduction to Data Science, Jeffrey Stanton, Syracuse University, 2013.
4. The white book of Big Data, Fujitsu, 2021
5. Mining of Massive Datasets Anand Rajaraman Kosmix, Inc. Jeffrey D. Ullman Stanford Univ.
6. R Programming for Data Science (2016), Roger D. Peng, https://bookdown.org/rdpeng/rprogdatascience/

PROJECT

1. Hands-on big data modelling, James Lee, Tao Wei, Suresh Kumar Mukhiya
2. **Kaggle**
3. Freelancing

PROJECT

1. Hands-on big data modelling, James Lee, Tao Wei, Suresh Kumar Mukhiya
2. **Kaggle**
3. Freelancing

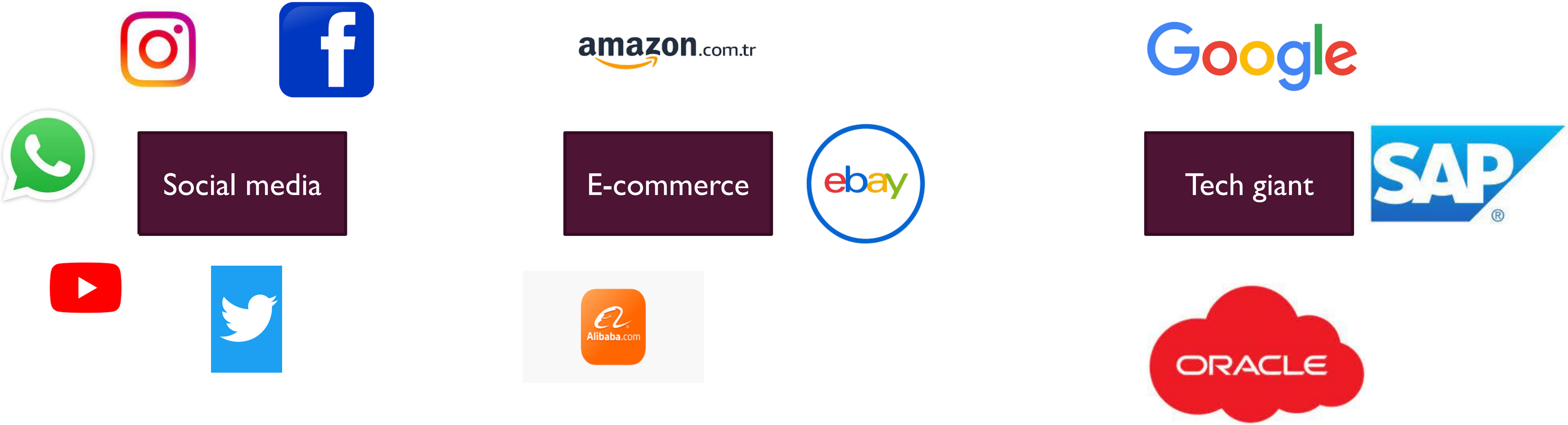
PROJECT

1. Hands-on big data modelling, James Lee, Tao Wei, Suresh Kumar Mukhiya
2. Kaggle
3. **Freelancing**

THE EVOLUTION OF BIG DATA

- In 2010 the term ‘Big Data’ was virtually unknown, but by mid-2011 it was widely touted as the latest trend.

THE EVOLUTION OF BIG DATA



Social media

E-commerce

Tech giant

BIG DATA

- Big Data involves having more data than you can handle with the computing power you already have, and you cannot easily scale your current computing environment.
- The definition of Big data therefore continues to evolve with time and advances in technology.

BIG DATA

Big data;

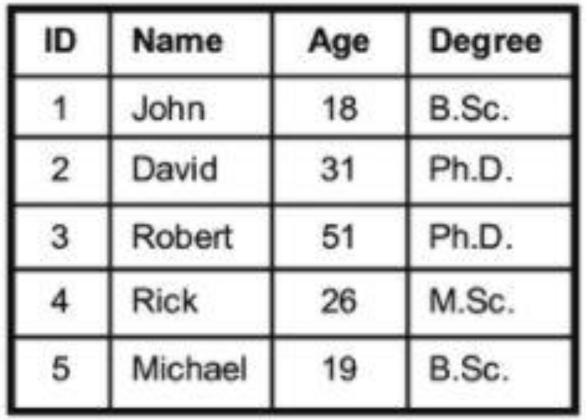
- It is a set of structured, semi-structured and unstructured data produced in high volume, speed and diversity.
- Are the raw material of knowledge.

DATA TYPES

- Structured data
- Semi-structured data
- Unstructured data

DATA TYPES

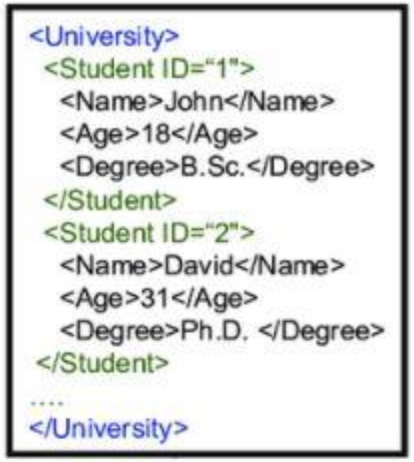
- Structured data: It refers to all types of data that are easy to model, input, store, query, manipulate and visualize.



ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

DATA TYPES

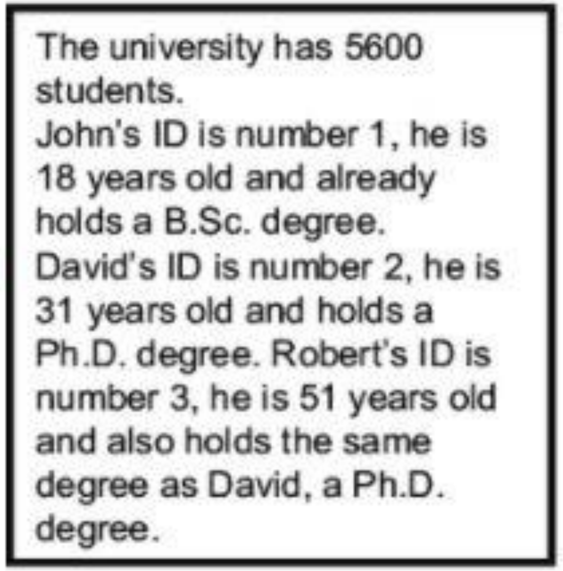
- Semi-structured data is a form of structured data that does not obey the tabular structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data.
- XLM, JSON



```
<University>
 <Student ID="1">
   <Name>John</Name>
   <Age>18</Age>
   <Degree>B.Sc.</Degree>
 </Student>
 <Student ID="2">
   <Name>David</Name>
   <Age>31</Age>
   <Degree>Ph.D. </Degree>
</Student>
....
</University>
```

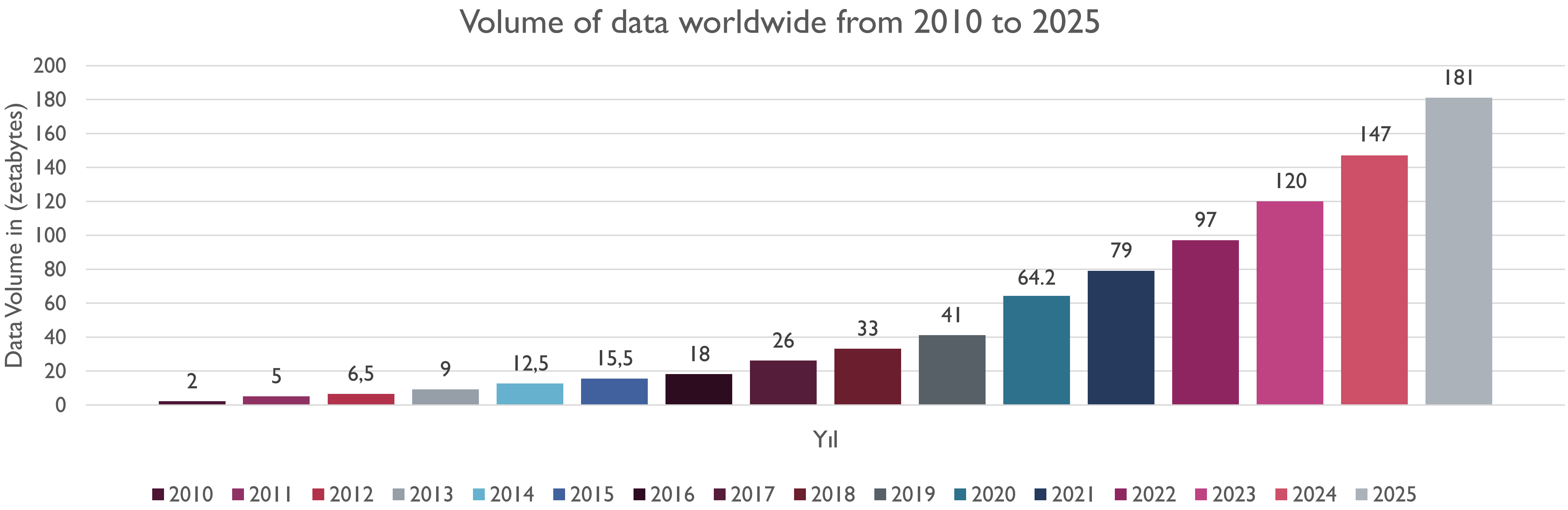
DATA TYPES

- Unstructured data means that it is datasets that are presented and stored in an undefined format.



The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

WHY BIG DATA



Volume of data worldwide from 2010 to 2025

WHY BIG DATA

- Amazon has taken advantage of its Big Data to create an extremely accurate representation of what products a customer should buy.

WHY BIG DATA

- Amazon has taken advantage of its Big Data to create an extremely accurate representation of what products a customer should buy.

WHY BIG DATA

- Amazon has taken advantage of its Big Data to create an extremely accurate representation of what products a customer should buy.
- The results are real and measurable, and they offer a practical advantage for a customer.

WHY BIG DATA

Amazon;

A customer buying a jacket in a snowy region. Why not suggest purchasing gloves to match, or boots, as well as a snow shovel, an ice melt, and tire chains?

Big Data analytics is able to interpret trends and bring understanding to the purchasing process by simply looking at what customers are buying, where they are buying it, and what they have purchased in the past.

WHY BIG DATA

Google;

aims to use Big Data to its fullest extent, to judge search results, predict Internet traffic usage, and service customers with Google’s own applications.

WHY BIG DATA

- Organizations that research earthquakes, weather, and global climates benefit from concept of Big Data.

WHY BIG DATA

- Small and medium businesses have access to scores of publicly available data, including most of the Web and social networking sites.
- Several hosted services have also come into being that can offer the computing power, storage, and platforms for analytics, changing the Big Data analytics market into a “pay as you go” entity.

WHY BIG DATA

- The momentum behind Big Data continues to be driven by the realization that large unstructured data sources can deliver almost immeasurable value.

WHY BIG DATA

Capturing Big Data's value;

- A potential annual value of $300 billion was realized in U.S. healthcare- more than double the total annual healthcare spending in Spain.
- A potential annual value of £250 billion has been realized for Europe's public sector administration.

WHY BIG DATA

1. Reduction in cost
2. Reduced production time
3. Development of new products
4. Smart decision-making

BIG DATA STATISTICS

It is estimated that the big data market value will reach $103 billion by 2023.

It is estimated that 97.2% of companies have started investing in big data technology.

Internet users create 2.5 quintillion bytes of data every day.

IDC's Digital Universe Survey in 2012 revealed that only 0.5% of the data was actually analyzed.

BIG DATA STATISTICS

In 2019, the number of active Facebook users was 2.3 billion.

300 hours of new video are uploaded and displayed every minute. Youtube's data has exceeded 1 billion gigabytes.

Google records 1.2 trillion search volumes per year, with 40,000 search queries sent every second.

5V'S OF BIG DATA

1. Variety
2. Velocity
3. Volume
4. Veracity
5. Value

5V’S OF BIG DATA

1. **Variety**
2. Velocity
3. Volume
4. Veracity
5. Value

5V’S OF BIG DATA

1. Variety
2. **Velocity: reflects speed at which this data is generated and changes.**
3. Volume
4. Veracity
5. Value

5V'S OF BIG DATA

1. Variety
2. Velocity: reflects speed at which this data is generated and changes.
3. **Volume**
4. Veracity
5. Value

5V'S OF BIG DATA

1. Variety
2. Velocity: reflects speed at which this data is generated and changes.
3. Volume
4. **Veracity**
5. Value

5V'S OF BIG DATA

1. Variety
2. Velocity: reflects speed at which this data is generated and changes.
3. Volume
4. Veracity
5. **Value**

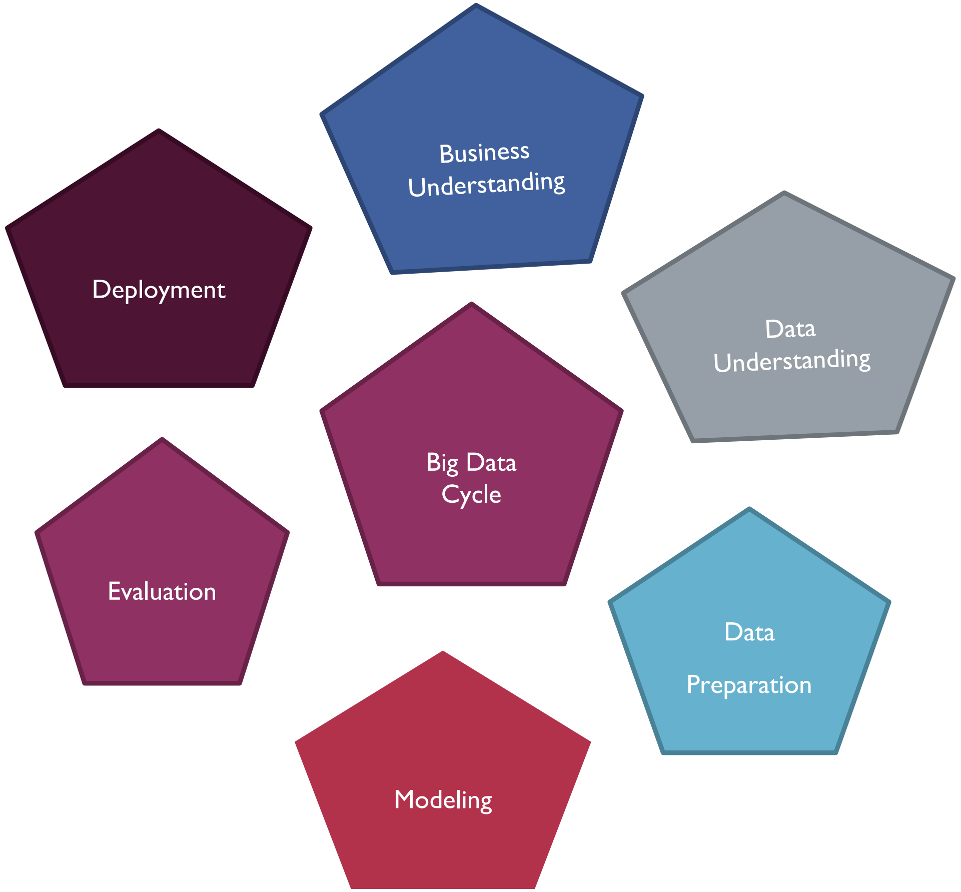
WHERE DOES BIG DATA COME FROM

- Computer generated
  - Application server logs (web sites, games, internet)
  - Sensor data (weather, atmospheric science, astronomy, smart grids)
  - Images/videos (traffic, security cameras, military surveillance)

- Human generated
  - Blogs/reviews/emails/pictures/scientific research/medical records
  - Social graphs: Facebook, contacts, twitter

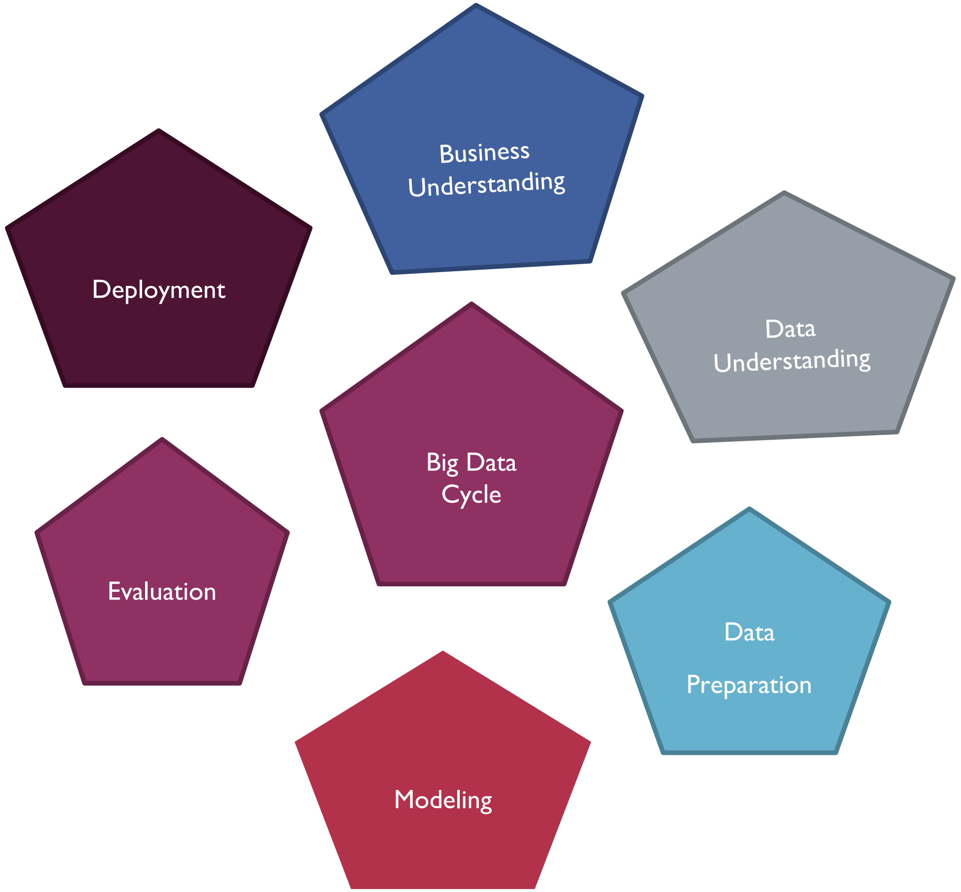
LIFE CYCLE OF BIG DATA

1. Business Case Evaluation
2. Data Identification
3. Data preparation
4. Model planning
5. Model building
6. Evaluation of the model
7. Implementation of the model



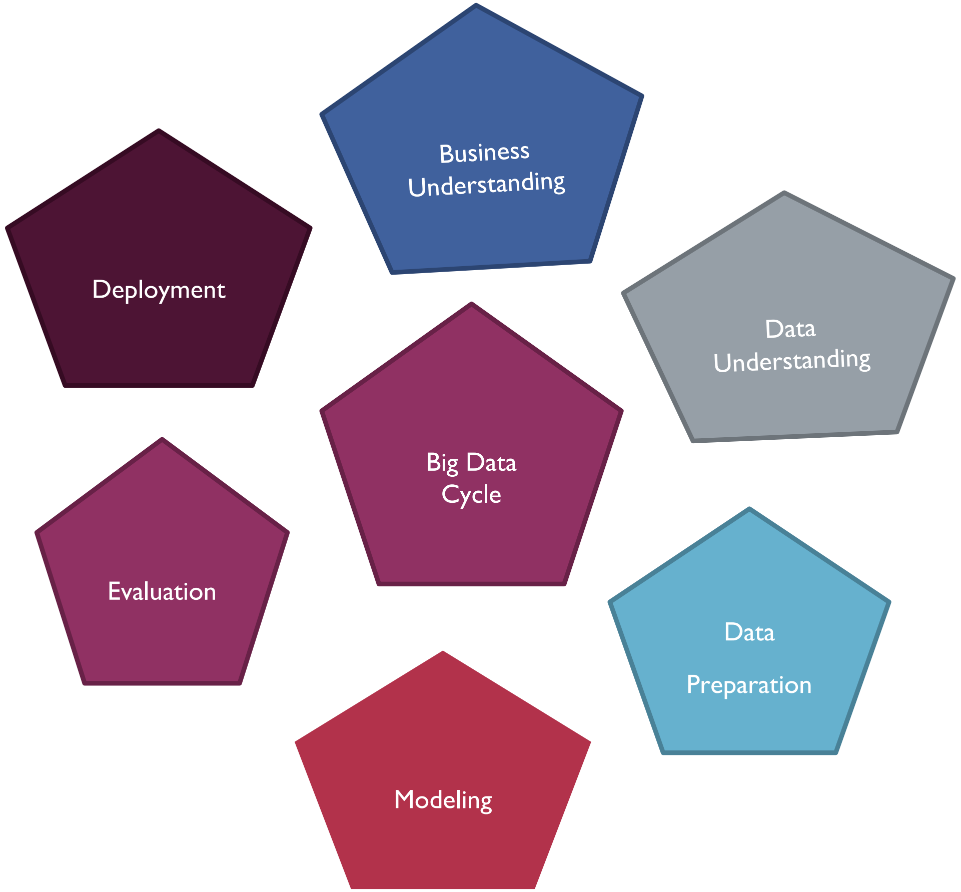
LIFE CYCLE OF BIG DATA

1. **Business Case Evaluation**
2. Data Identification
3. Data preparation
4. Model planning
5. Model building
6. Evaluation of the model
7. Implementation of the model



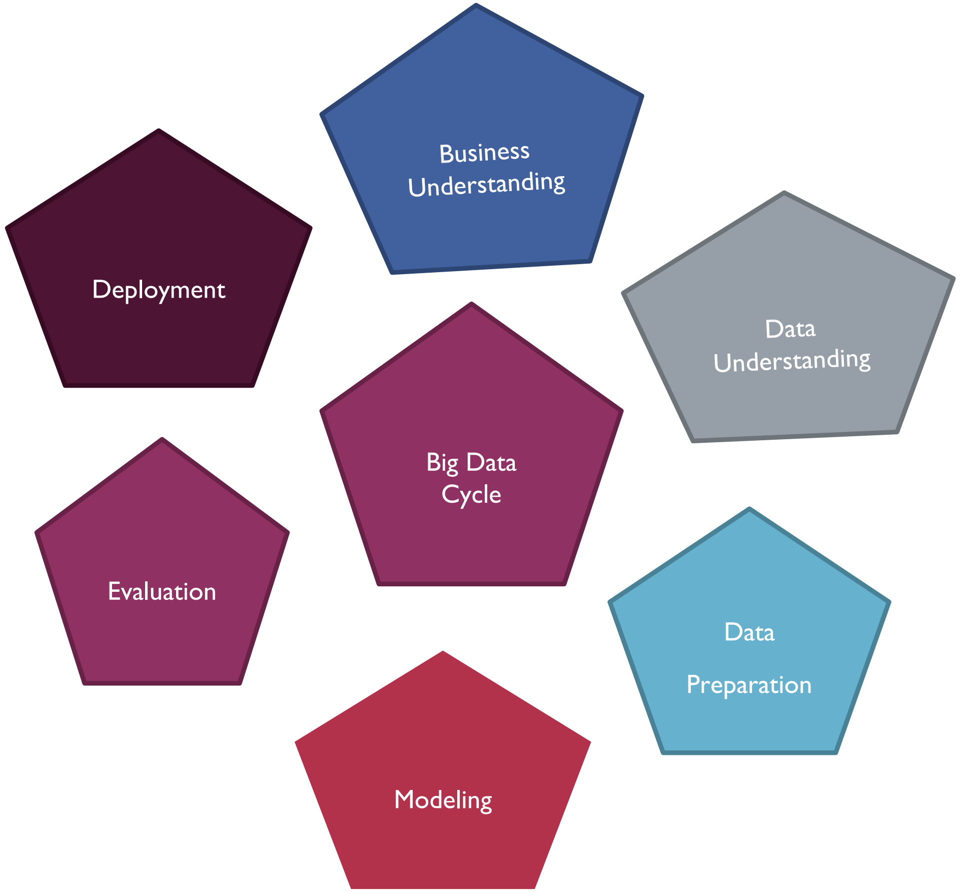
LIFE CYCLE OF BIG DATA

1. Business Case Evaluation
2. **Data Identification**
3. Data preparation
4. Model planning
5. Model building
6. Evaluation of the model
7. Implementation of the model



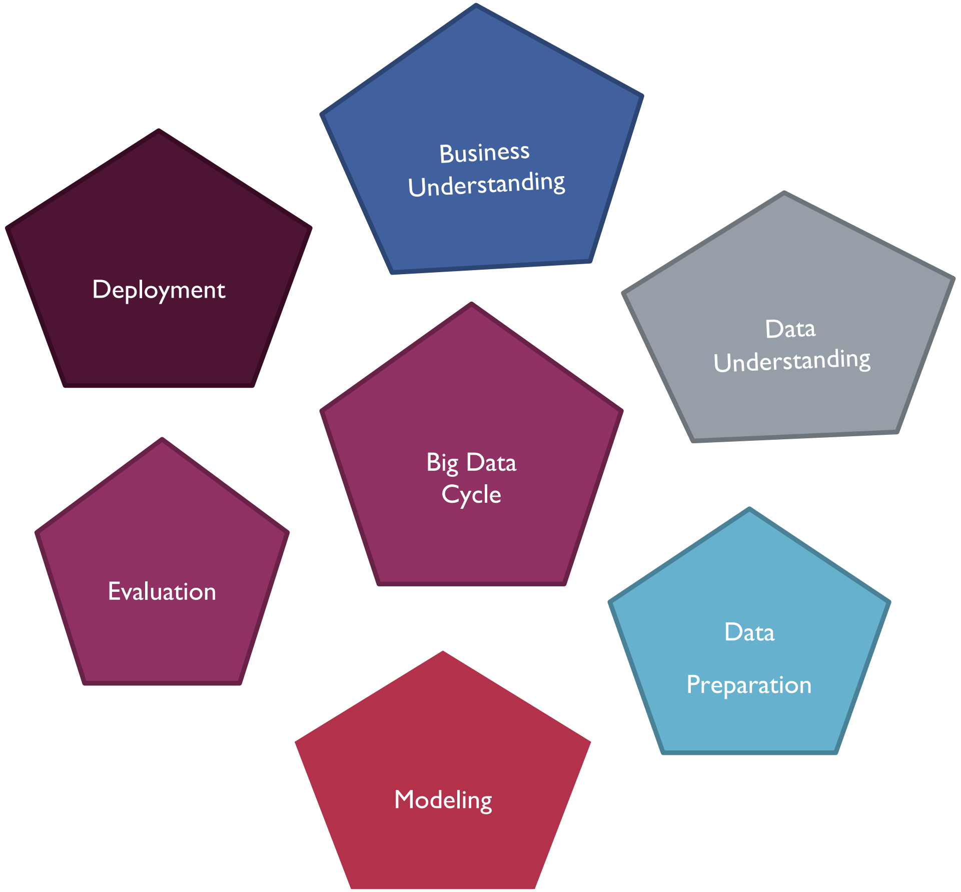
LIFE CYCLE OF BIG DATA

1. Business Case Evaluation
2. Data Identification
3. **Data preparation**
4. Model planning
5. Model building
6. Evaluation of the model
7. Implementation of the model



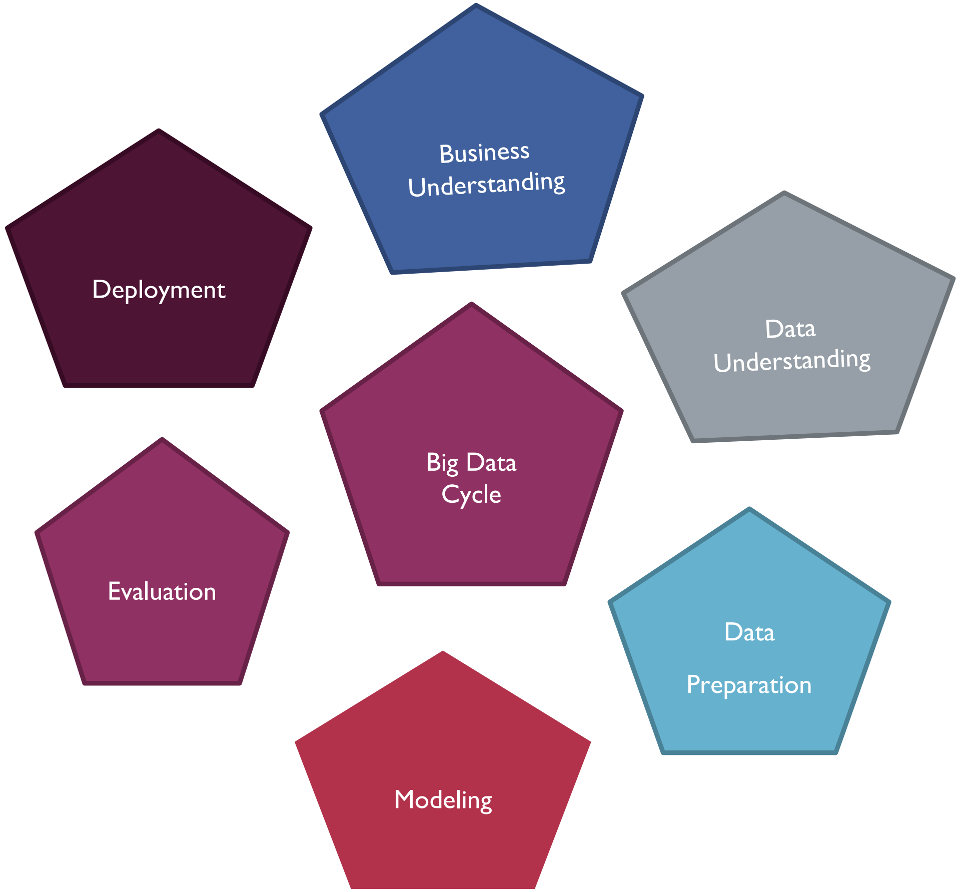
LIFE CYCLE OF BIG DATA

1. Business Case Evaluation
2. Data Identification
3. Data preparation
4. **Model planning**
5. Model building
6. Evaluation of the model
7. Implementation of the model



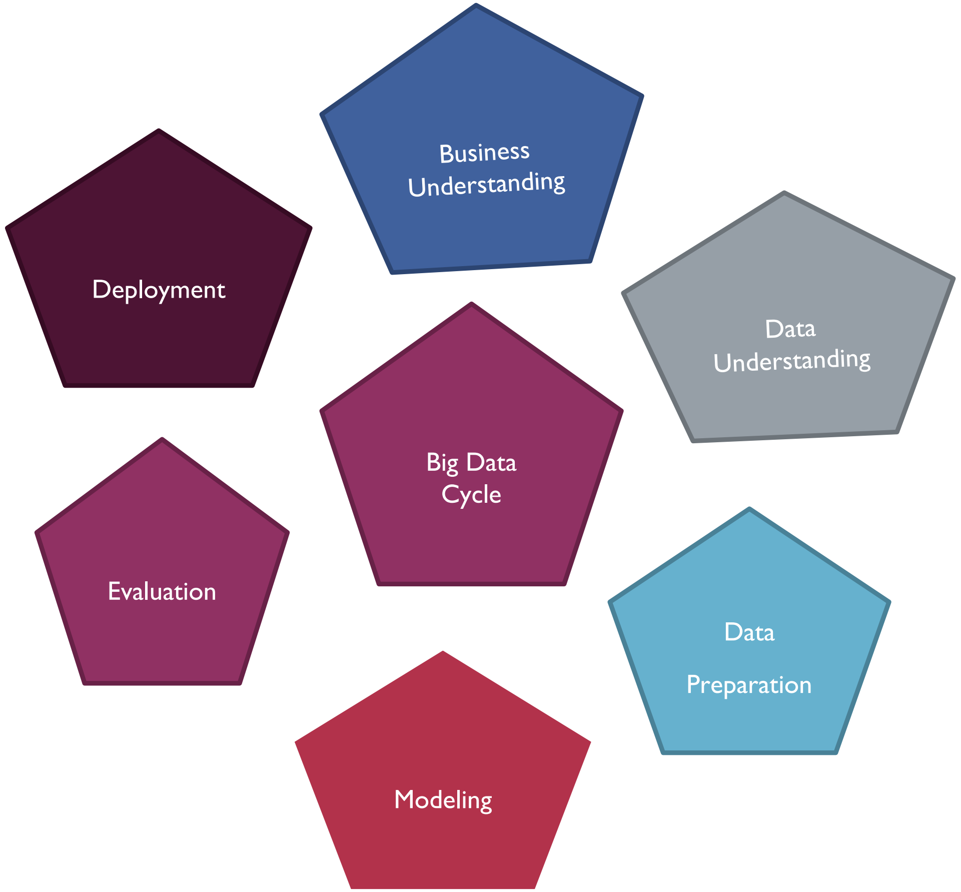
LIFE CYCLE OF BIG DATA

1. Business Case Evaluation
2. Data Identification
3. Data preparation
4. Model planning
5. **Model building**
6. Evaluation of the model
7. Implementation of the model



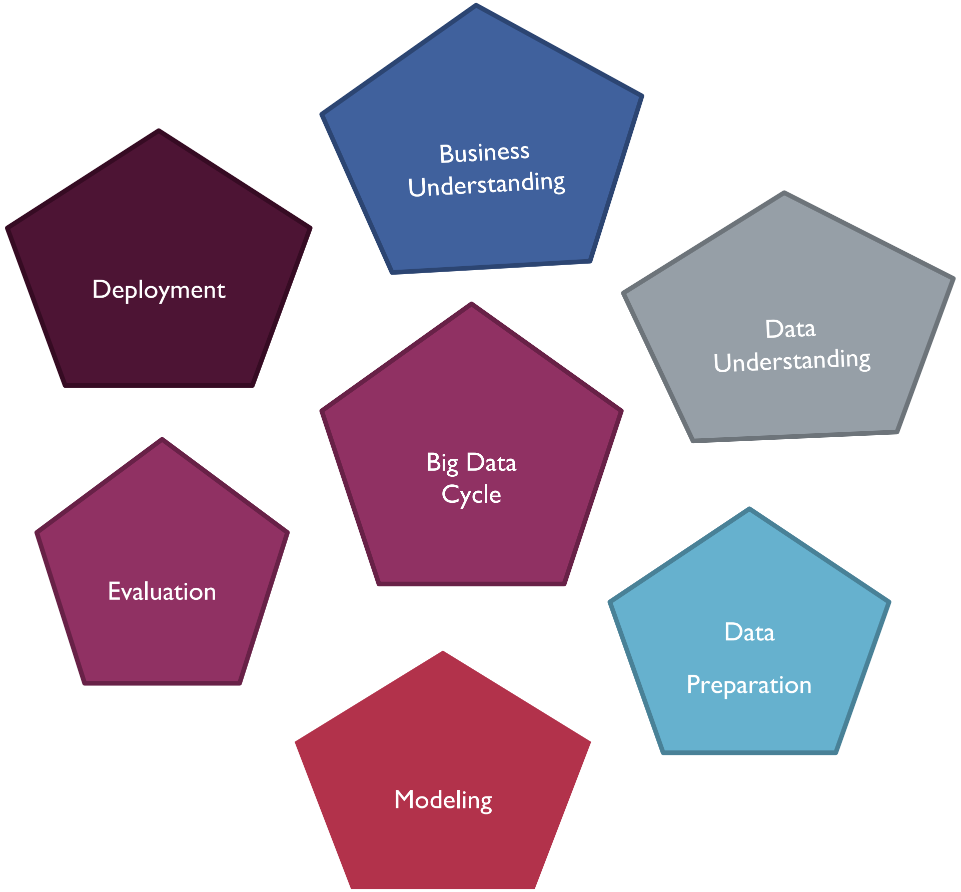
LIFE CYCLE OF BIG DATA

1. Business Case Evaluation
2. Data Identification
3. Data preparation
4. Model planning
5. Model building
6. **Evaluation of the model**
7. Implementation of the model



LIFE CYCLE OF BIG DATA

1. Business Case Evaluation
2. Data Identification
3. Data preparation
4. Model planning
5. Model building
6. Evaluation of the model
7. **Implementation of the model**



BIG DATA ANALYTICS CATEGORIES

- Data analytics enable data-driven decision-making with scientific backing so that decision can be based on factual data and not simply on past experience or intuition alone. There are four general categories of analytics that are distinguished by the results they produce:
  - Descriptive analytics
  - Diagnostic analytics
  - Predictive analytics
  - Prescriptive analytics

BIG DATA ANALYTICS CATEGORIES

- **Descriptive analytics:** Answer questions about events that have already occurred.
  - What was the sales volume over the past 12 months?
  - What is the number of support calls received as categorized by severity and geographic location?
  - What is the monthly commission earned by each sales agent?

BIG DATA ANALYTICS CATEGORIES

- **Diagnostic analytics:** Determine the cause of a phenomenon that occurred in the past using questions that focus on the reason behind the event.
  - Why were Q2 sales less than Q1 sales?
  - Why have there been more support calls originating from the Eastern region than from the Western region?
  - Why was there an increase in patient re-admission rates over the past three months?

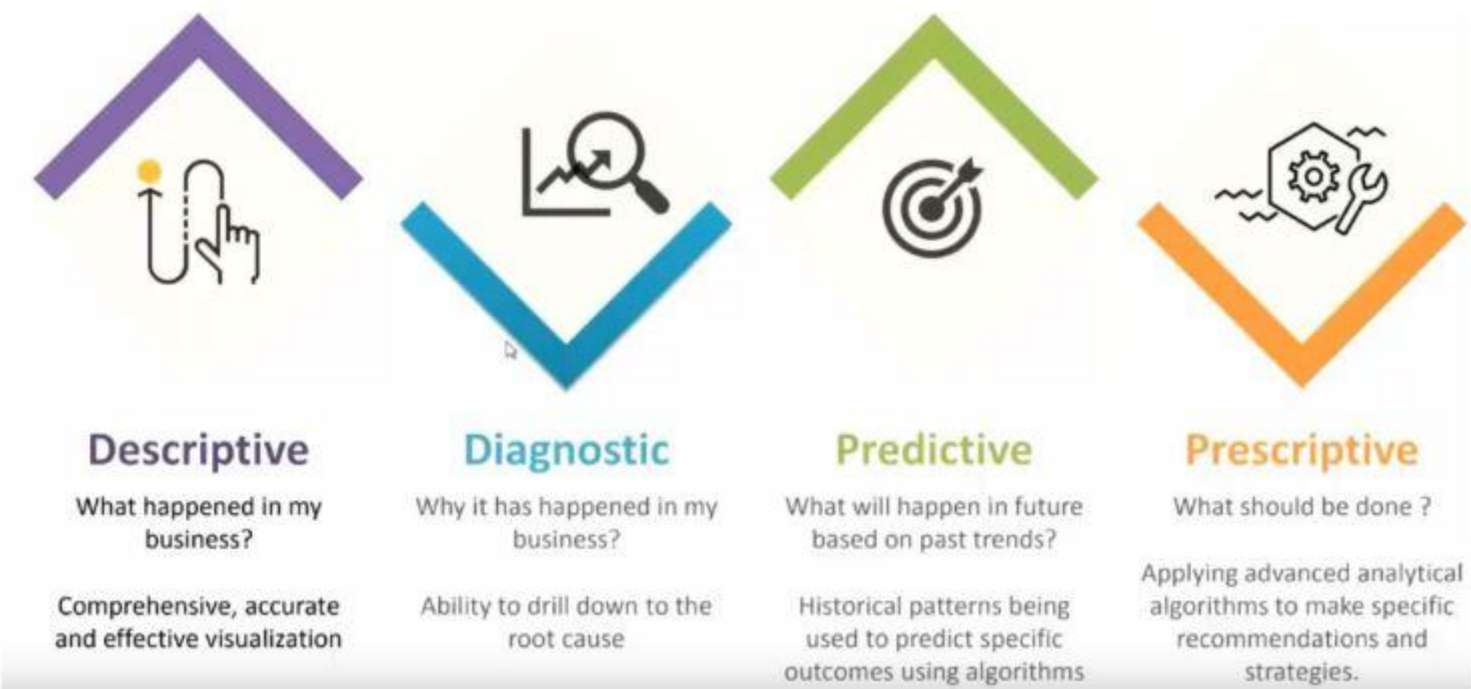
BIG DATA ANALYTICS CATEGORIES

- **Predictive analytics:** Determine the outcome of an event that might occur in the future.
  - What are the chances that a customer will default on a loan if they have missed a monthly payment?
  - What will be the patient survival rate if Drug B is administered instead of Drug A?
  - If a customer has purchased Products A and B, what are the chances that they will also purchase Product C?

BIG DATA ANALYTICS CATEGORIES

- **Prescriptive analytics:** It is built upon the results of predictive analytics by prescribing actions that should be taken. The focus is not only on which prescribed option is best to follow, but why.
  - Among three drugs, which one provides the best results?
  - When is the best time to trade a particular stock?

BIG DATA ANALYTICS CATEGORIES



Descriptive

What happened in my business?

Comprehensive, accurate and effective visualization

Diagnostic

Why it has happened in my business?

Ability to drill down to the root cause

Predictive

What will happen in future based on past trends?

Historical patterns being used to predict specific outcomes using algorithms

Prescriptive

What should be done ?

Applying advanced analytical algorithms to make specific recommendations and strategies.

BIG DATA APPLICATIONS

- Education
- Healthcare
- Government
- Entertainment and Media
- Weather
- Transportation
- Banking

BIG DATA IN EDUCATION

- **Customized and dynamic learning programs:** Customized programs and schemes to benefit individual students can be created using the data collected on the bases of each student's learning history. This improves the overall student results.
- **Reframing course material:** Reframing the course material according to the data that is collected on the basis of what a student learns and to what extent by real-time monitoring of the components of a course is beneficial for the students.
- **Grading:** New advancements in grading systems have been introduced as a result of a proper analysis of student data.
- **Career Prediction:** Appropriate analysis and study of every student's records will help understand each student's progress, strengths, weaknesses, interests, and more. It would also help in determining which career would be the most suitable for the student in future.

BIG DATA IN HEALTHCARE

- No unnecessary diagnosis and treatment cost reduce
- Epidemic prediction and help to decide preventative measures
- Detection of diseases in early stage
- Medical results of past medicines and evidence-base more accurate prescription.

BIG DATA IN GOVERNMENT

Welfare Schemes

- In making faster and informed decisions regarding various political programs
- To identify areas that needs attention
- To stay up to date in the field of agriculture by keeping track of all existing land and livestock
- To overcome national challenges such as unemployment, terrorism, energy resources exploration

Cyber Security

- Deceit recognition
- To catch tax evaders

BIG DATA IN ENTERTAINMENT AND MEDIA

- Predicting the interests of audiences
- Effective targeting of the advertisements
- Optimized or on-demand scheduling of media streams in digital media distribution platforms
- Getting insights from customer reviews

BIG DATA IN WEATHER

- Forecasting
- To study global warming
- In understanding the patterns of natural disasters
- To make necessary preparations in the case of crises
- To predict the availability of usable water

BIG DATA IN TRANSPORTATION

- **Route planning:** To estimate users’ needs on different routes and on multiple modes of transportation and then utilize route planning to reduce their wait time.
- **Congestion management and traffic control:** Real-time estimation of congestion and traffic patterns. E.g. using Google Maps to locate the least traffic-prone routes.
- **Safety level of traffic:** To use the real-time processing of big data and predictive analysis to identify accident-prone areas and help reduce accidents.

BIG DATA IN BANKING

- Misuse of credit/debit cards
- Venture credit hazard treatment
- Business clarity
- Customer statistics alteration
- Money laundering
- Risk mitigation

CHALLENGES OF BIG DATA

1. Data collection and storage
2. Scalability
3. Data integrity
4. Visualization of data
5. Analysis of data

CHALLENGES OF BIG DATA

1. **Data collection and storage**
    1. Too much data size and volume
    2. Insufficient storage space
    3. Management and maintenance costs

    Solution: Cloud Computing, compression algorithms,

CHALLENGES OF BIG DATA

1. **Data integrity**
    1. Different types of data need to be managed and processed
    2. Data from different sources needs to be mapped.
    3. Data quality needs to be improved.

CHALLENGES OF BIG DATA

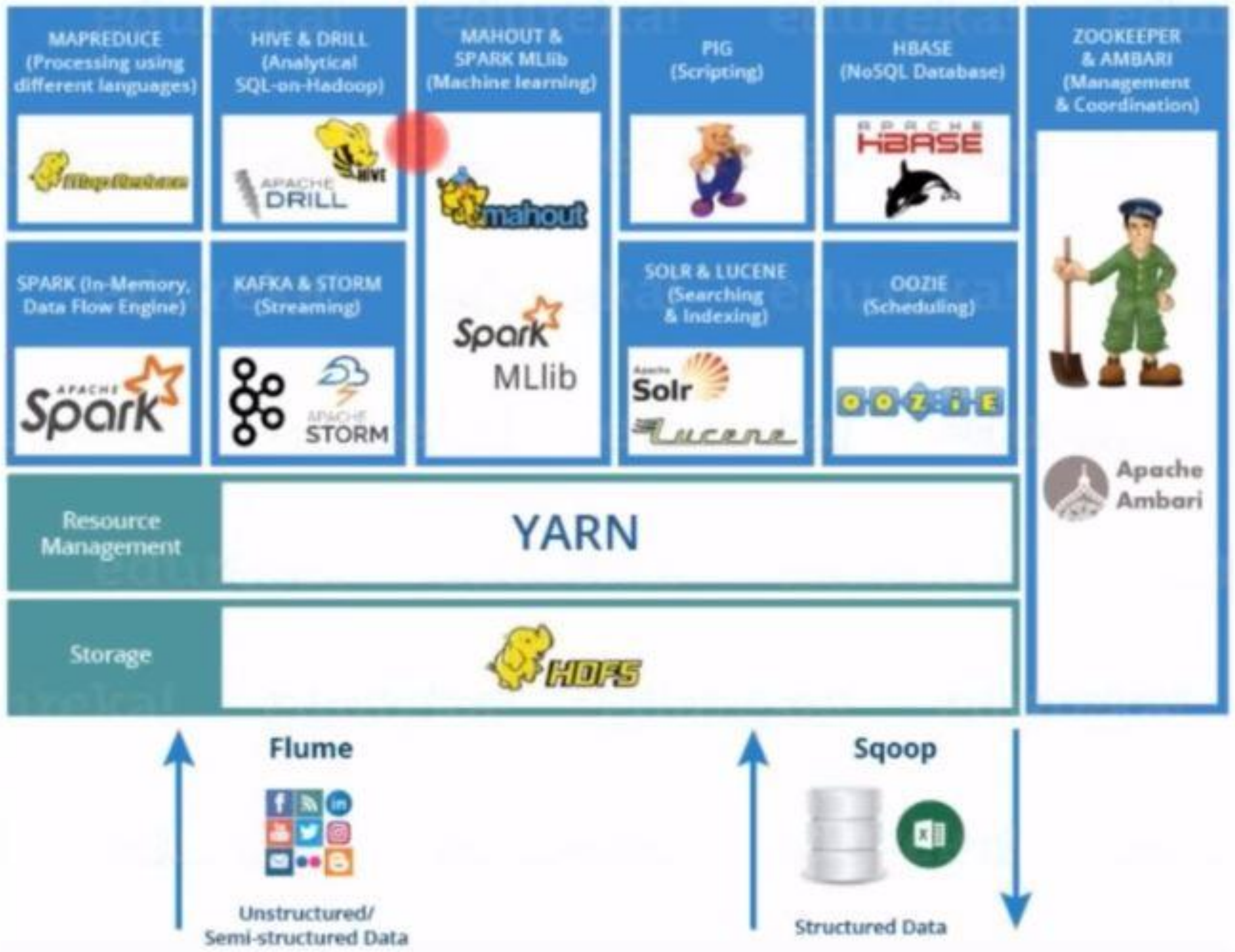
1. **Visualization of data**
    1. The results of the processed big data should be presented visually in a meaningful way.

CHALLENGES OF BIG DATA

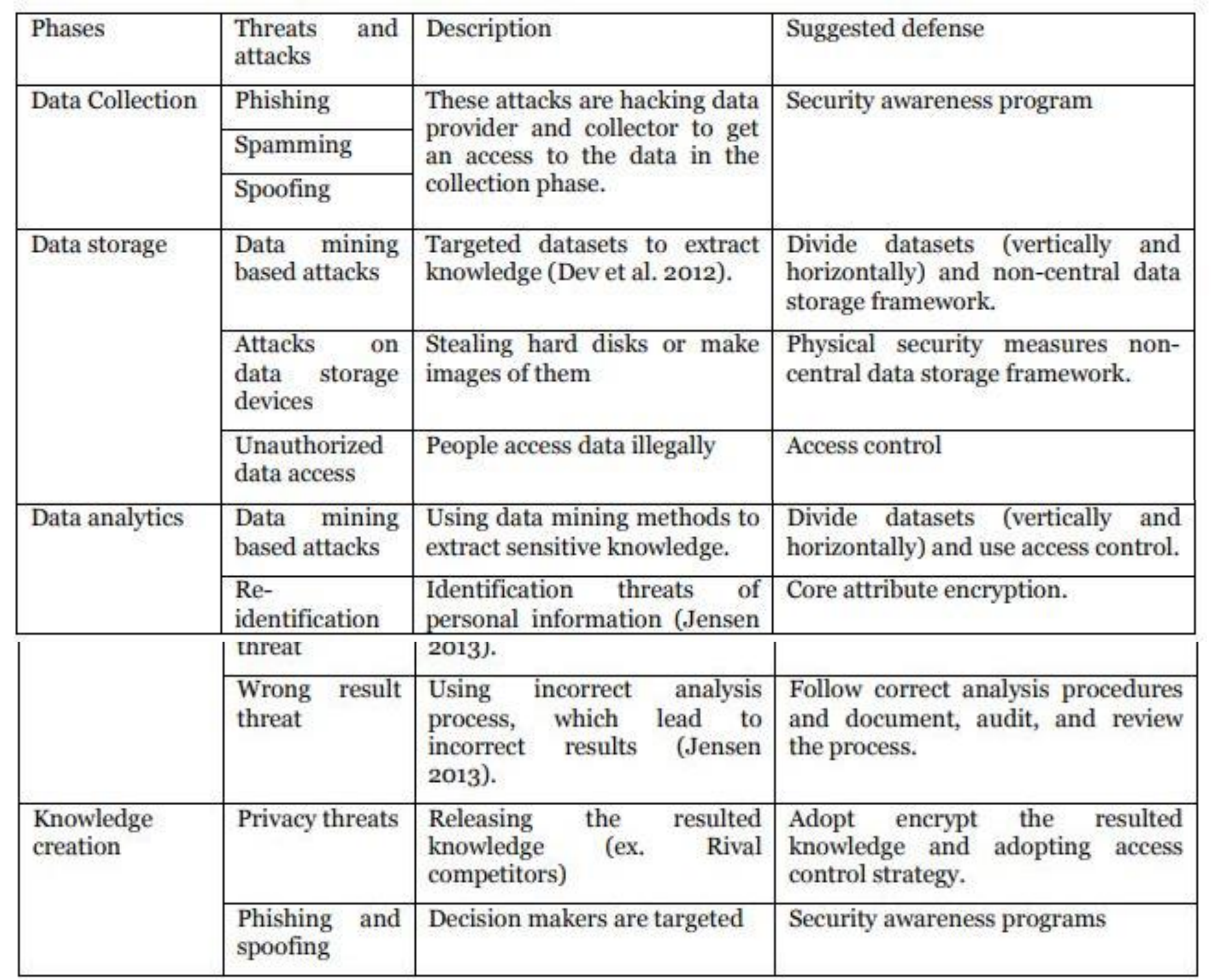
1. **Scalability**
    1. It is the system's ability to adapt to increasing demands in terms of processing.

CHALLENGES OF BIG DATA

1. **Scalability**
    1. It is the system's ability to adapt to increasing demands in terms of processing.


MAPREDUCE (Processing using different languages)
MapReduce
HIVE & DRILL (Analytical SQL-on-Hadoop)
APACHE DRILL
HIVE
MAHOUT & SPARK MLlib (Machine learning)
mahout
Spark MLlib
PIG (Scripting)
HBASE (NoSQL Database)
APACHE HBASE
ZOOKEEPER & AMBARI (Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
APACHE Spark
KAFKA & STORM (Streaming)
APACHE STORM
SOLR & LUCENE (Searching & Indexing)
Apache Solr
Lucene
OOZIE (Scheduling)
OOZIE
Apache Ambari
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data

BIG DATA THREATS AND ATTACKS



Phases	Threats and attacks	Description	Suggested defense
Data Collection	Phishing	These attacks are hacking data provider and collector to get an access to the data in the collection phase.	Security awareness program
	Spamming		
	Spoofing		
Data storage	Data mining based attacks	Targeted datasets to extract knowledge (Dev et al. 2012).	Divide datasets (vertically and horizontally) and non-central data storage framework.
	Attacks on data storage devices	Stealing hard disks or make images of them	Physical security measures non-central data storage framework.
	Unauthorized data access	People access data illegally	Access control
Data analytics	Data mining based attacks	Using data mining methods to extract sensitive knowledge.	Divide datasets (vertically and horizontally) and use access control.
	Re-identification threat	Identification threats of personal information (Jensen 2013).	Core attribute encryption.
	Wrong result threat	Using incorrect analysis process, which lead to incorrect results (Jensen 2013).	Follow correct analysis procedures and document, audit, and review the process.
Knowledge creation	Privacy threats	Releasing the resulted knowledge (ex. Rival competitors)	Adopt encrypt the resulted knowledge and adopting access control strategy.
	Phishing and spoofing	Decision makers are targeted	Security awareness programs