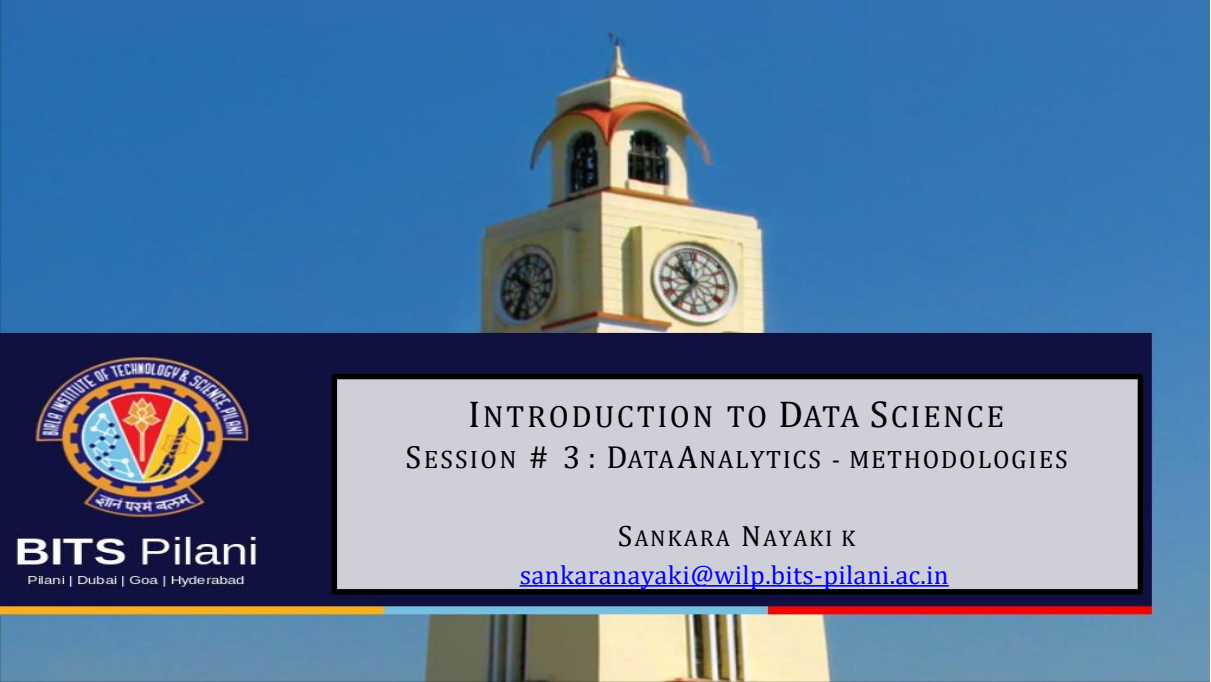


BITS Pilani
Pilani | Dubai | Goa | Hyderabad

# INTRODUCTION TO DATA SCIENCE

## SESSION # 3 : DATA ANALYTICS - METHODOLOGIES

SANKARA NAYAKI K
sankaranayaki@wilp.bits-pilani.ac.in

The instructor is gratefully acknowledging the authors who made their course materials freely available online.

References:


- Introducing Data Science by Cielen, Meysman and Ali
- Storytelling with Data by Cole Nussbaumer Knaflic; Wiley
- Introduction to Data Mining by Tan, Steinbach and Vipin Kumar
- The Art of Data Science by Roger D Peng and Elizabeth Matsui
- Python Data Science Handbook: Essential tools for working with data by Jake VanderPlas

# Data Analytics

- Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making.
- 4 different types of analytics
  1. Descriptive Analytics
  2. Diagnostic Analytics
  3. Predictive Analytics
  4. Prescriptive Analytics

# Data Analytics Methodologies

- Use standard methodology to ensure a good outcome.
  1. CRISP-DM
  2. SEMMA
  3. BIG DATA LIFE CYCLE
  4. SMAM

# CRISP-DM

- Cross Industry Standard Process for Data Mining
- conceived around 1996
- 6 high-level phases
- Used in IBM SPSS Modeler tool

# Why Should There be a Standard Process?

- Framework for recording experience
  - Allows projects to be replicated
- Aid to project planning and management
- “Comfort factor” for new adopters
  - Demonstrates maturity of Data Mining
  - Reduces dependency on “stars”
- Encourage best practices and help to obtain better results

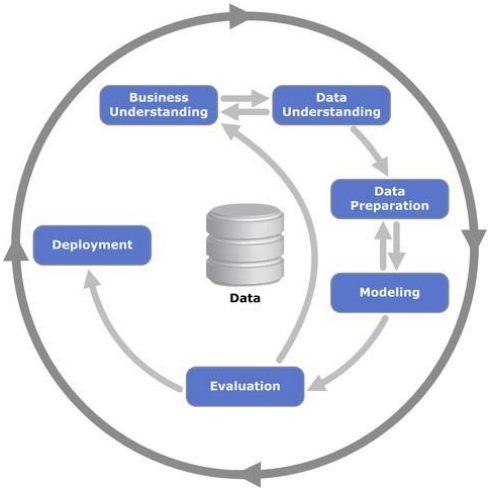
# CRISP-DM: Phases

- ## Business Understanding
  - Understanding project objectives and requirements; Data mining problem definition
- ## Data Understanding
  - Initial data collection and familiarization; Identify data quality issues; Initial, obvious results
- ## Data Preparation
  - Record and attribute selection; Data cleansing

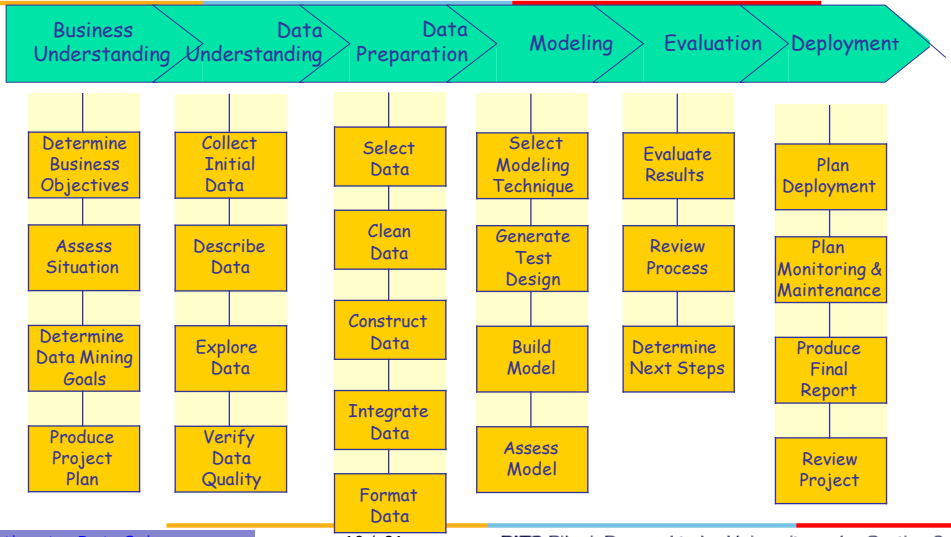
# CRISP-DM: Phases

- **Modelling**
  - Run the data mining tools
- **Evaluation**
  - Determine if results meet business objectives; Identify business issues that should have been addressed earlier
- **Deployment**
  - Put the resulting models into practice; Set up for continuous mining of the data

# CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

# Phases & Tasks



# Why CRISP – DM ?

- The data mining process must be reliable and repeatable by people with little data mining skills
- CRISP-DM provides a uniform framework for
  - guidelines
  - experience documentation
- CRISP-DM is flexible to account for differences
  - Different business/agency problems
  - Different data

# References – CRISP-DM

- https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-project html
- https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome

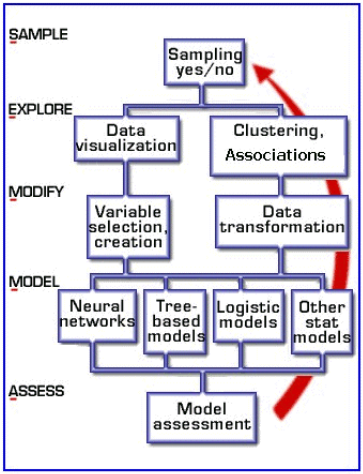
# SEMMA

- SAS Institute
- Sample, Explore, Modify, Model, Assess
- 5 stages

# SEMMA

1. Sample – This stage consists on sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly. This stage is pointed out as being optional.
2. Explore – This stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.
3. Modify – This stage consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.
4. Model – This stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.
5. Assess – This stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.

# SEMMA


SAMPLE
Sampling yes/no
EXPLORE
Data visualization
Clustering, Associations
MODIFY
Variable selection, creation
Data transformation
MODEL
Neural networks
Tree-based models
Logistic models
Other stat models
ASSESS
Model assessment

# SEMMA

- “SEMMA is not a data mining methodology but rather a logical organisation of the functional tool set of SAS Enterprise Miner for carrying out the core tasks of data mining.

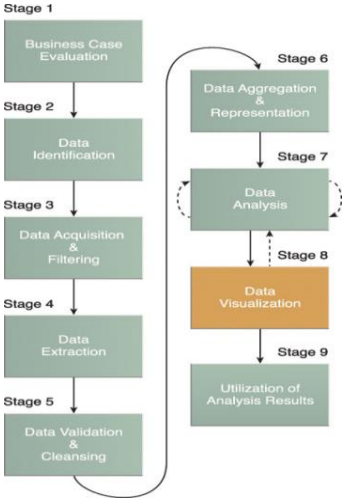
- Enterprise Miner can be used as part of any iterative data mining methodology adopted by the client. Naturally steps such as formulating a well defined business or research problem and assembling quality representative data sources are critical to the overall success of any data mining project.

- SEMMA is focused on the model development aspects of data mining.”

# References – SEMMA

- https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en
- http://jesshampton.com/2011/02/16/semma-and-crisp-dm-data-mining-methodologies/

# BIG DATA LIFE CYCLE & STAGES

# BIG DATA LIFE CYCLE

- **Data Acquisition**, acquiring information from a rich and varied data environment,
- **Data Awareness**, connecting data from different sources into a coherent whole, including modeling content, establishing context, and insuring searchability,
- **Data Analytics**, using this contextual data to answer questions about the state of your organization,
- **Data Governance**, establishing a framework for providing for the provenance, infrastructure and disposition of that data.

# BIG DATA LIFE CYCLE (PHASES)

*Phase 1: Foundations*

*Phase 2: Acquisition*

*Phase 3: Preparation*

*Phase 4: Input and Access*

*Phase 5: Processing*

*Phase 6: Output and Interpretation*

*Phase 7: Storage*

*Phase 8: Integration*

*Phase 9: Analytics and Visualisation*

*Phase 10: Consumption*

*Phase 11: Retention, Backup, and Archival*

*Phase 12: Destruction*

# BIG DATA LIFE CYCLE

1. Foundation : understanding and validating data requirements, solution scope, roles and responsibilities, data infrastructure preparation, technical and non-technical considerations, and understanding data rules in an organisation.
2. Data Acquisition refers to collecting data. Data sets can be obtained from various sources. These sources can be internal and external to the business organisations
3. Data preparation phase, the collected data — in raw format- is cleaned or cleansed
4. Data input refers to sending data to planned target data repositories, systems, or applications.
5. Data Processing phase starts with processing the raw form of data. Convert data into a readable format giving it the form and the context. After completion of this activity, we can interpret the data using the data analytics tools

# BIG DATA LIFE CYCLE

6. In the data output phase, the data is in a format which is ready for consumption by the business users. We can transform data into usable formats such as plain text, graphs, processed images, or video files.
7. Storing data in designed and designated storage units. These units are part of the data platform - capacity, scalability, security, compliance, performance and availability.
8. In Big Data, there may be a need for the integration of stored data to different systems for various purposes.
9. Integrated data can be useful and productive for data analytics and visualisation.
10. Once data analytics takes place, then the data is turned into information ready for consumption by the internal or external users, including customers of the business organisation.

# BIG DATA LIFE CYCLE

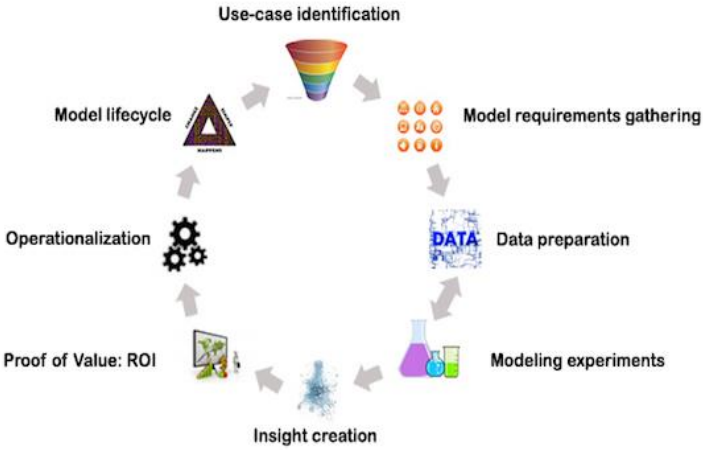
11 We know that critical data must be backed up for protection and meeting industry compliance requirements.

12 There may be regulatory requirements to destruct a particular type of data after a certain amount of times.

# References – BIGDATA LIFE CYCLE

- https://www.sciencedirect.com/science/article/pii/S1877050920315465
- https://medium.com/illumination-curated/big-data-lifecycle-management-629dfe16b78d

# SMAM


Use-case identification
Model requirements gathering
Data preparation
Modeling experiments
Insight creation
Proof of Value: ROI
Operationalization
Model lifecycle
DATA

# SMAM- PHASES

| | |
|---|---|
| Use-case identification | Selection of the ideal approach from a list of candidates |
| Model Requirements gathering | Understanding the conditions required for the model to function |
| Data preparation | Getting the data ready for the modeling |
| Modelling experiments | Scientific experimentation to solve the business question |
| Insight creation | Visualization and dashboarding to provide insight |
| Proof of Value: ROI | Running the model in a small scale setting to prove the value |
| Operationalization | Embedding the analytical model in operational systems |
| Model lifecycle | Governance around model lifetime and refresh |

# SMAM- REFERENCES

- https://www.kdnuggets.com/2015/08/new-standard-methodology-analytical-models.html
- https://www.linkedin.com/pulse/standard-methodology-analytical-models-olav-laudy/?trk=prof-post

# Analytics Capacity Building

- Implement project plans
- Elevate from project → initiative
- Achieve expeditionary execution of strategy, responding to emerging conditions .
- Use analytics to support strategy, innovation, and organizational development

# Challenges in Data-driven decision making

1. **Discrimination :** Algorithmic discrimination can come from various sources. First, the data used to train algorithms may have **biases that lead to discriminatory decisions**. Second, discrimination may arise from the use of a particular algorithm.
2. **Lack of transparency:** Transparency refers to the capacity to understand a computational model and therefore contribute to the **attribution of responsibility for consequences** derived from its use. A model is transparent if a person can easily observe it and understand it.
3. **Violation of privacy:** Reports and studies have focused on the misuse of users' personal data and on data aggregation by entities such as data brokers, which have direct implications for people's privacy.
4. **Digital literacy:** It is extremely important that we devote resources to digital and computer literacy programs for all citizens, from children to the elderly. If we do not, it will be very difficult (if not impossible) as a society to **make decisions about technologies** that we do not understand.

# Challenges in Data-driven decision making

5. **Fuzzy responsibility:** As more and more decisions that affect millions of people are made automatically by algorithms, we must be clear about who is responsible for the consequences of these decisions. Transparency is often considered a fundamental factor in the **clarity of attribution of responsibility**.
6. **Lack of ethical frameworks:** Algorithmic data-based decision-making processes generate important **ethical dilemmas regarding what actions are appropriate** in light of the inferences made by algorithms. It is therefore essential that decisions be made in accordance with a clearly defined and accepted ethical framework.
7. **Lack of diversity:** Given the variety of cases in which algorithms can be applied for decision-making, it is important to reflect on the **frequent lack of diversity** in the teams that generate such algorithms. So far, data-based algorithms and artificial intelligence techniques for decision-making have been developed by homogeneous groups of IT professionals.

# Using CRISP-DM to Predict Car Prices

**Business Problem#1**

Imagine for example a used car dealer who needs estimates what the price of a used car could be. The car dealer could be interest in predicting the price of a car based on its attributes. More precise, we try to answer to the following 3 business questions:

•Is the price of a car related to the horsepower?

•Is the price of a care related to the length of the car?

•Can the price of a car be predicted based in its attribute with reasonable accuracy?

***Note: The CRISP-DM process starts with the understanding of the business problem.***

# THANK YOU