BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science

Data Analytics Methodologies

Dr. Ramakrishna Dantu
Associate Professor, BITS Pilani

innovate achieve lead

Disclaimer and Acknowledgement

- The content for these slides has been obtained from books and various other source on the Internet
- I here by acknowledge all the contributors for their material and inputs.
- I have provided source information wherever necessary
- I have added and modified the content to suit the requirements of the course

Introduction to Data Science

Data Analytics Module Topics

- Defining Analytics
- Types of data analytics
  - Descriptive, Diagnostic
  - Predictive, Prescriptive
- Data Analytics – methodologies
  - KDD
  - CRISP-DM
  - SEMMA
  - SMAM
  - BIG DATA LIFE CYCLE
- Analytics Capacity Building
- Challenges in Data-driven decision making

Data Analytics Methodologies

innovate achieve lead

What exactly is a methodology?

Definition of 'methodology'

methodology

Collins COBUILD

Word Frequency

(mɛθədɒlədʒi)

Word forms: plural **methodologies**

VARIABLE NOUN

A **methodology** is a system of methods and principles for doing something, for example for teaching or for carrying out research.

[*formal*]

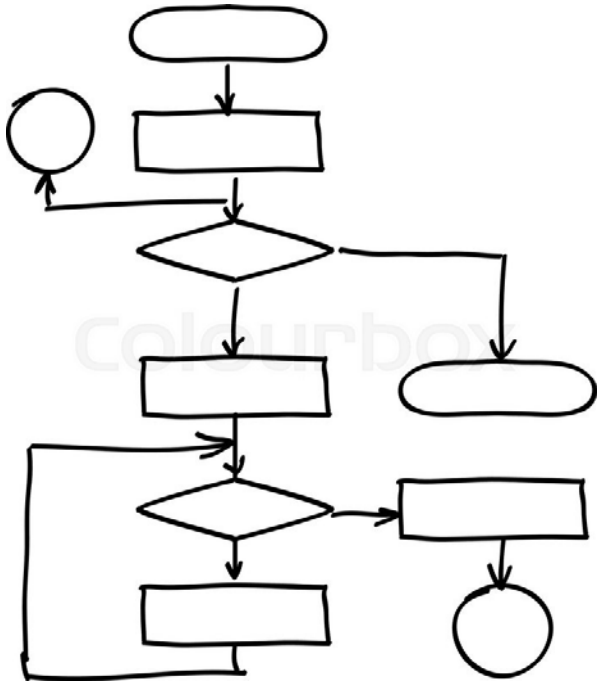
Teaching methodologies vary according to the topic.

In their own work they may have favored the use of methodology different from mine.

Synonyms: practice, style, approach, technique More Synonyms of **methodology**

methodological (mɛθədəlɒdʒɪkᵊl) **ADJECTIVE** [*usually ADJECTIVE noun*]

...theoretical and methodological issues raised by the study of literary texts.



Source: Data Mining and Predictive Analytics Daniel Larose & Chantal Larose

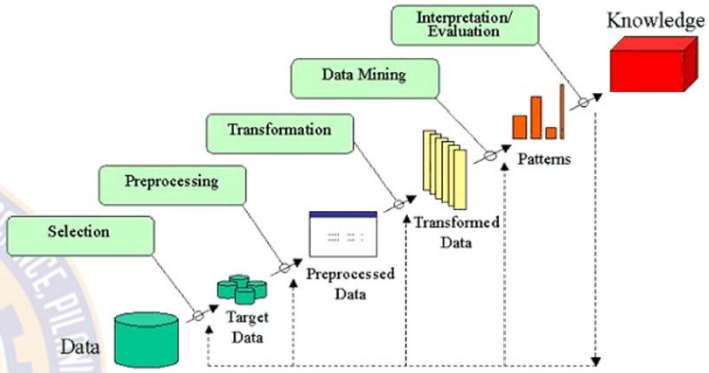
Knowledge Discovery in Databases

Knowledge Discovery in Databases

innovate achieve lead

CRISP-DM Methodology

- The Knowledge Discovery in Databases (or KDD) refers to a process of finding knowledge from data in the context of large databases



- It does this by applying data mining methods (algorithms) to extract what is deemed knowledge:
  - In accordance with the specifications of measures and thresholds, using a database along with any required preprocessing, subsampling, and transformations of that database

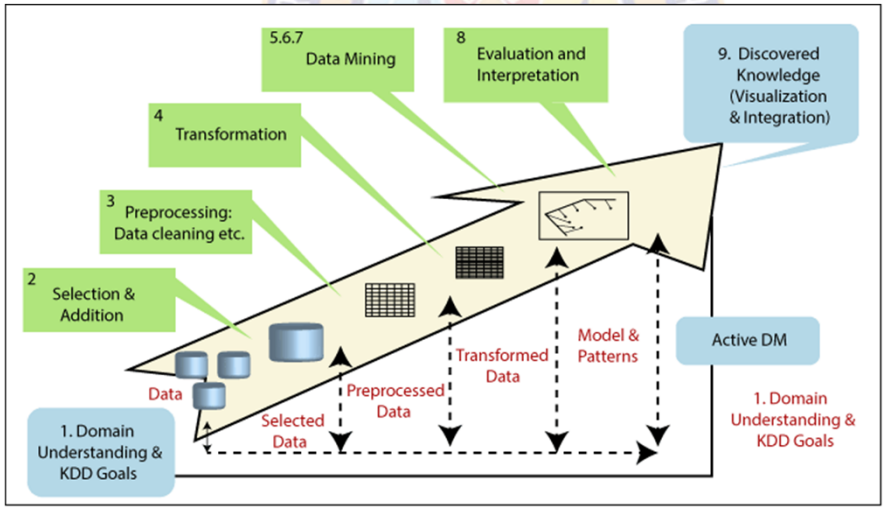
Source: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html
Image source: https://www.javatpoint.com/kdd-process-in-data-mining

Knowledge Discovery in Databases

innovate achieve lead

KDD Process

- KDD requires relevant prior knowledge and brief understanding of application domain and goals
- KDD process is iterative and interactive in nature
- There are nine different steps in seven different phases of this model



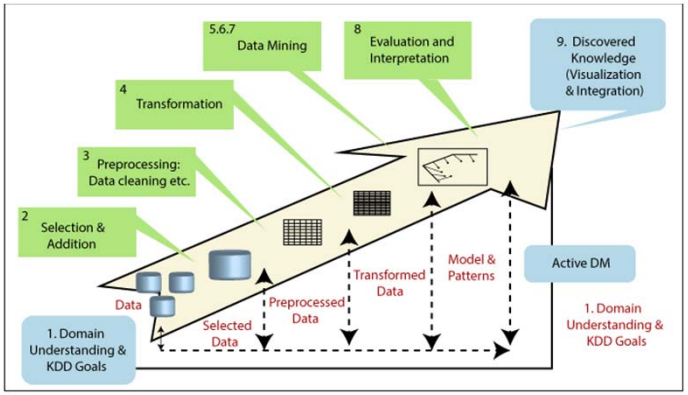
Source: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html
Image source: https://www.javatpoint.com/kdd-process-in-data-mining

Knowledge Discovery in Databases

innovate achieve lead

KDD Process

- Pre KDD process (Developing Domain Understanding)
  - This is the first stage of KDD process in which goals are defined from customer's view point and used to develop and understanding about application domain and its prior knowledge
  - Involves developing an of understanding of:
    - The application domain
    - The relevant prior knowledge and
    - The goals of the end-user
- Post KDD process (Using Discovered Knowledge)
  - This understanding must be continued by:
    - The knowledge consolidation
    - Incorporation this knowledge into the system



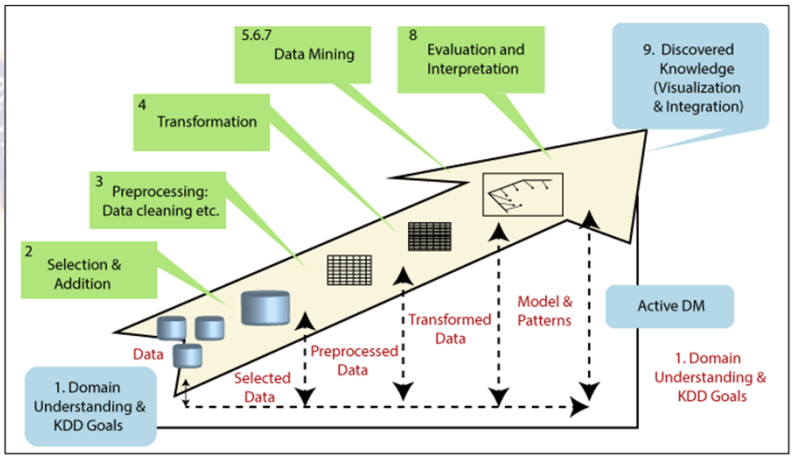
Source: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

Knowledge Discovery in Databases

innovate achieve lead

KDD Process

- Selection:
  - This is the second stage of KDD process which focuses creating on target data set and subset of data samples or variables.
  - It is an important stage because knowledge discovery is performed on all these.
- Pre-processing
  - This is the third stage of KDD process which focuses on target data cleaning and pre-processing to obtain consistent data without any noise and inconsistencies
  - In this stage strategies are developed for handling such type of noisy and inconsistent data.
- Transformation
  - This is the fourth stage of KDD process which focuses on the transformation of data using dimensionality reduction and other transformation methods



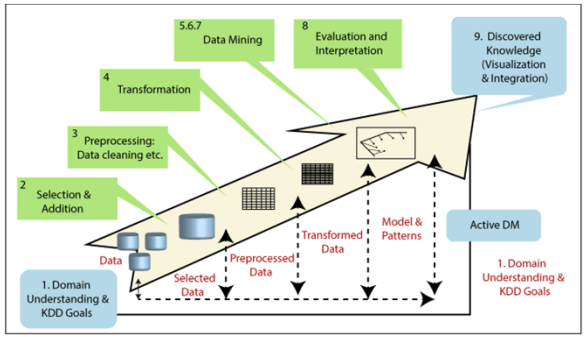
Source: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html
Image source: https://www.javatpoint.com/kdd-process-in-data-mining

Knowledge Discovery in Databases

innovate achieve lead

KDD Process

- Data Mining - this stage consists of three steps:
  - Choosing the suitable data mining task
    - In this step of KDD process an appropriate data mining task is chosen based on particular goals that are defined in first stage
    - The examples of data mining method or tasks are classification, clustering, regression and summarization etc.
  - Choosing a suitable data mining algorithm
    - In this step, one or more appropriate data mining algorithms are selected for searching different patterns from data
    - There are number of algorithms present today for data mining but appropriate algorithms are selected based on matching the overall criteria for data mining
  - Employing the data mining algorithm
    - In this step, the selected algorithms are implemented



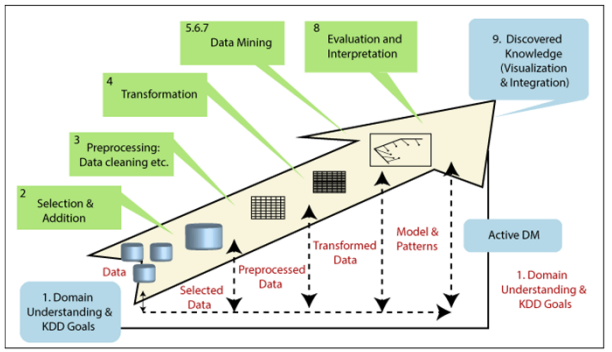
Source: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html
Image source: https://www.javatpoint.com/kdd-process-in-data-mining

Knowledge Discovery in Databases

innovate achieve lead

KDD Process

- Interpretation/Evaluation
  - This is the eighth step of KDD process that focuses on interpretation and evaluation of mined patterns
  - This step may involve in visualization of extracted patterns
- Post KDD Process (Using discovered knowledge)
  - In this final step of KDD process the discovered knowledge is used for different purposes
  - The discovered knowledge can be used by the interested parties or can be integrate with another system for further action



Source: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html
Image source: https://www.javatpoint.com/kdd-process-in-data-mining

CRISP-DM

CRoss-***I***ndustry ***S***tandard ***P***rocess for ***D***ata ***M***ining

Data Analytics Methodologies

innovate achieve lead

CRISP-DM Methodology

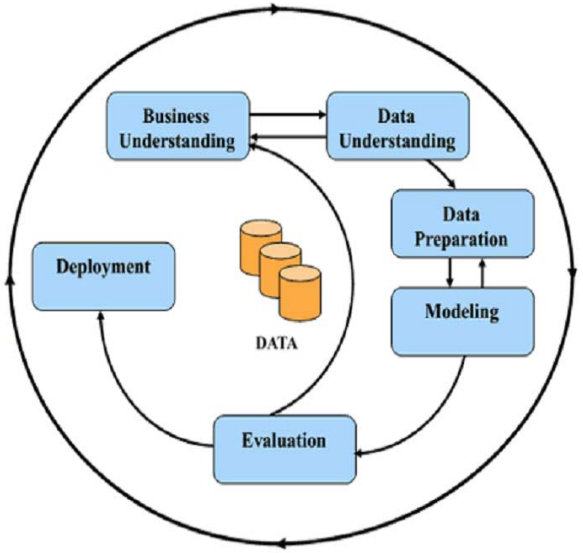
- Stands for "***CR***oss-***I***ndustry ***S***tandard ***P***rocess for ***D***ata ***M***ining"
- CRISP-DM was conceived in 1996 and became a European Union project under the ESPRIT funding initiative in 1997
  - European Strategic Program on Research in Information Technology
- The project was led by five companies: Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR Corporation and OHRA, an insurance company.
  - ISL, later acquired and merged into SPSS. The computer giant NCR Corporation produced the Teradata data warehouse and its own data mining software.
- It is an open standard process model that describes common approaches used by data mining experts
- Provides a nonproprietary, technology-agnostic, and structured approach for fitting data mining project into the general problem-solving strategy
- This model is an idealized sequence of six stages
- In practice, it will often be necessary to backtrack to previous activities and repeat certain actions

Source: Data Mining and Predictive Analytics Daniel Larose & Chantal Larose

Data Analytics Methodologies

innovate achieve lead

CRIPS-DM Methodology



- A data mining project is conceptualized as a life cycle consisting of six phases
- The phase-sequence is *adaptive*
  - That is, the next phase in the sequence depends on the outcomes associated with the previous phase
- Dependencies between phases are indicated by the arrows
  - For instance, suppose that we are in the modeling phase
  - Depending on the behavior and characteristics of the model, we may have to return to the data preparation phase for further refinement before moving forward to the model evaluation phase

Source: Data Mining and Predictive Analytics Daniel Larose & Chantal Larose
Image Source: https://towardsdatascience.com/why-using-crisp-dm-will-make-you-a-better-data-scientist-66efe5b72686

Data Analytics Methodologies

innovate achieve lead

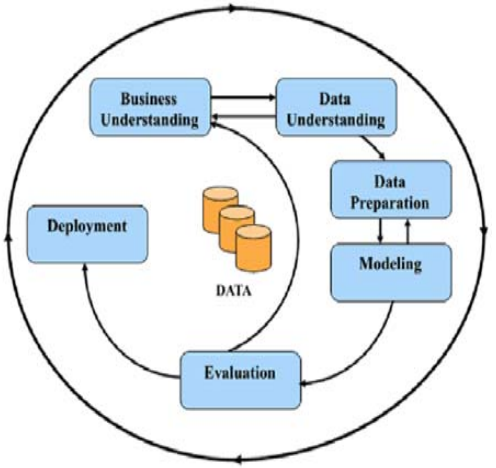
Six Phases of CRIPS-DM Methodology

- **Business/Research Understanding Phase**
  - This is the first phase of CRISP-DM process
  - It focuses on and uncovers important factors including success criteria, business and data mining objectives and requirements as well as business terminologies and technical terms.

  a) First, clearly enunciate the project objectives and requirements in terms of the business or research unit as a whole

  b) Then, formulate these goals and restrictions into a data science problem definition

  c) Finally, prepare a preliminary strategy for achieving these objectives



Source: Data Mining and Predictive Analytics Daniel Larose & Chantal Larose

Data Analytics Methodologies

innovate achieve lead

Six Phases of CRIPS-DM Methodology

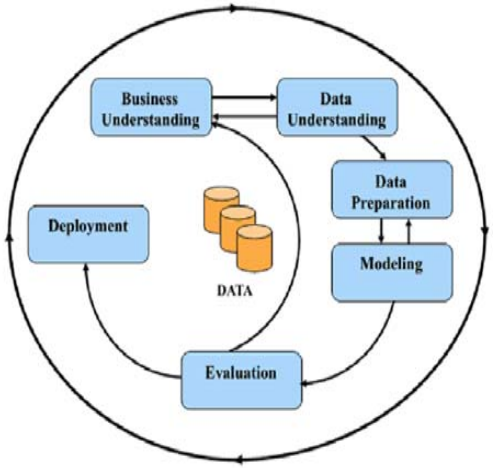
- **Data Understanding Phase**
  - This phase of CRISP-DM process focuses on data collection, exploring of data, checking quality, to get insights into the data to form hypotheses for hidden information

  a) First, collect the data

  b) Then, use exploratory data analysis to familiarize yourself with the data, and discover initial insights

  c) Evaluate the quality of the data

  d) Finally, if desired, select interesting subsets that may contain actionable patterns



Source: Data Mining and Predictive Analytics Daniel Larose & Chantal Larose

Data Analytics Methodologies

innovate achieve lead

Six Phases of CRIPS-DM Methodology

- **Data Preparation Phase**
  - This phases focuses on selection and preparation of final data set
  - This phase may include many tasks such as records, table and attributes selection as well as cleaning and transformation of data.

a) This labor-intensive phase covers all aspects of preparing the final data set, which shall be used for subsequent phases, from the initial, raw, dirty data

b) Select the cases and variables you want to analyze, and that are appropriate for your analysis.

c) Perform transformations on certain variables, if needed.

d) Clean the raw data so that it is ready for the modeling tools.

Source: Data Mining and Predictive Analytics Daniel Larose & Chantal Larose

Data Analytics Methodologies

innovate achieve lead

Six Phases of CRIPS-DM Methodology

- **Modeling Phase**
  - This phase involves selection and application of various modeling techniques
  - Different parameters are set and different models are built for same data mining problem

a) Select and apply appropriate modeling techniques.

b) Calibrate model settings to optimize results.

c) Often, several different techniques may be applied for the same data science problem

d) May require looping back to previous phase, in order to bring data into a specific format requirements of a particular modeling technique

Source: Data Mining and Predictive Analytics Daniel Larose & Chantal Larose

Data Analytics Methodologies

innovate achieve lead

Six Phases of CRIPS-DM Methodology

- **Evaluation Phase**
  - This phase involves evaluation of obtained models and deciding how to use the results
  - Interpretation of the model depends upon the algorithm and models are evaluated to review whether they achieve the objectives properly or not.

a) Evaluate the models delivered by the modeling phase for quality and effectiveness, before deploying them for use in the field

b) Determine whether the model in fact achieves the objectives set for it in phase 1.

c) Establish whether some important facet of the business or research problem has not been sufficiently accounted for.

d) Finally, come to a decision regarding the use of the modeling results.

Source: Data Mining and Predictive Analytics Daniel Larose & Chantal Larose

Data Analytics Methodologies

Six Phases of CRIPS-DM Methodology

- **Deployment Phase**
  - This phase focuses on deploying the final model and determining the usage of obtain knowledge and results
  - This phase also focuses on organizing, reporting and presenting the gained knowledge when needed

a) Model creation does not signify the completion of the project until they are used

b) Examples:
   a) A simple deployment: Generate a report.
   b) A more complex deployment: Initiate projects based on the predictive outcomes of the model

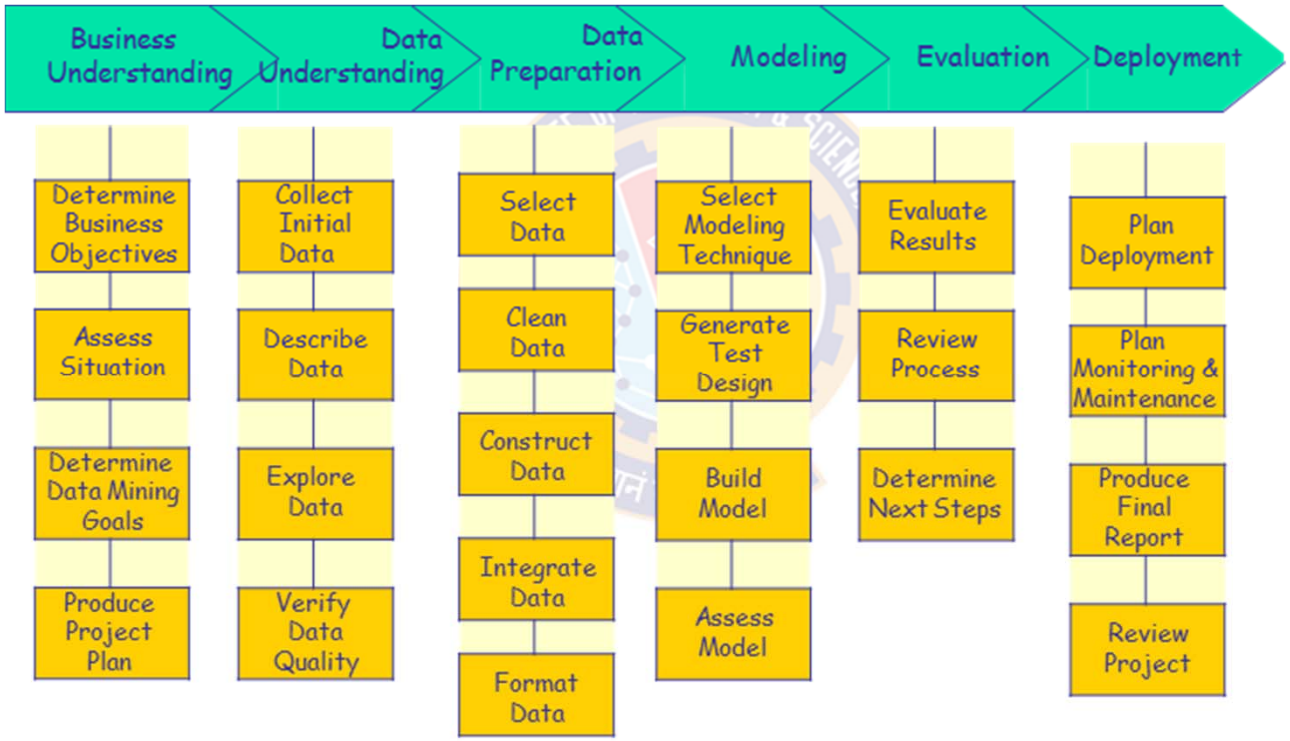
c) For businesses, the customer often carries out the deployment based on your model

Source: Data Mining and Predictive Analytics Daniel Larose & Chantal Larose

Data Analytics Methodologies

innovate achieve lead

Phases and Tasks of CRIPS-DM Methodology



SEMMA Process Model

Data Analytics Methodologies

SEMMA Process Model

- SEMMA stands for **S**ample, **E**xplore, **M**odify, **M**odel and **A**ccess
- It is a data mining method presented by the SAS Institute
- It enables data organization, discovery, development and maintenance of data mining projects
- SAS Institute defines data mining as the process of Sampling, Exploring, Modifying, Modeling, and Assessing (SEMMA) large amounts of data to uncover previously unknown patterns which can be utilized as a business advantage
- The data mining process is applicable across a variety of industries and provides methodologies for such diverse business problems, such as:
  - Fraud detection, customer retention and attrition, market segmentation, risk analysis, customer satisfaction, bankruptcy prediction, and portfolio analysis

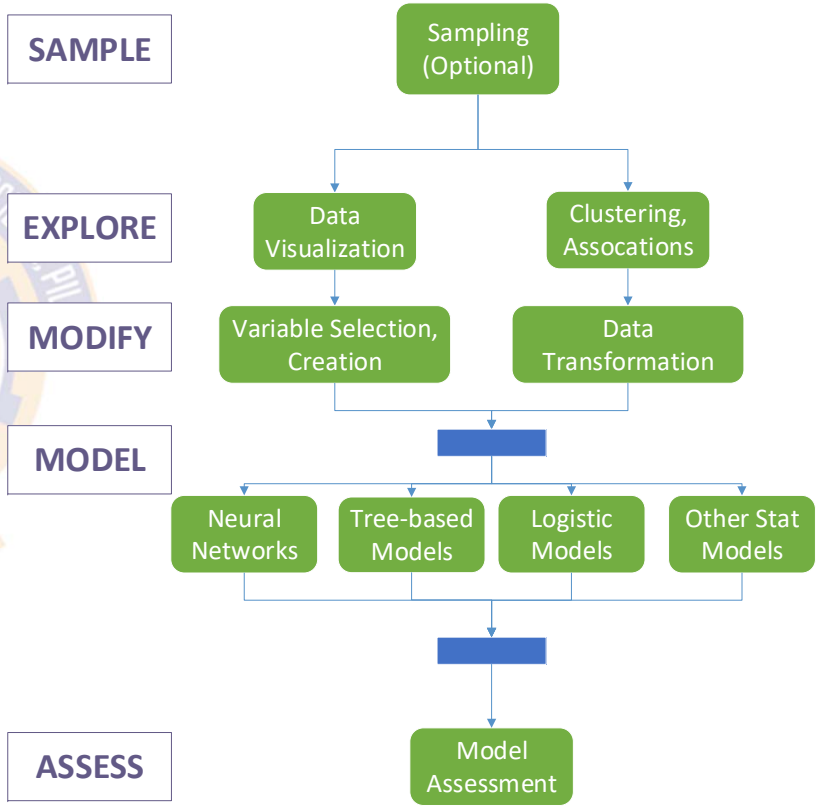
Source: Big Data Analytics Methods, 2nd Edition by Ghavami P

Data Analytics Methodologies

innovate achieve lead

SEMMA Process Model

- Five stages of SEMMA model:
  - **Sample:**
    - Involves sampling of data to extract critical information about the data while small enough to manipulate quickly.
  - **Explore:**
    - This stage focuses on data exploration and discovery.
    - It improves understanding of the data, finding trends and anomalies in the data.
  - **Modify:**
    - This stage involves modification of data through data transformations, and the model selection process.
    - This stage may evaluate anomalies, outliers and variable reduction in data.
  - **Model:**
    - The goal of this stage is to build the model and apply the model to the data.
    - Different modeling techniques may be applied as datasets have different attributes.
  - **Access:**
    - The final stage of SEMMA evaluates the reliability and usefulness of the results, performance and accuracy of the model(s)



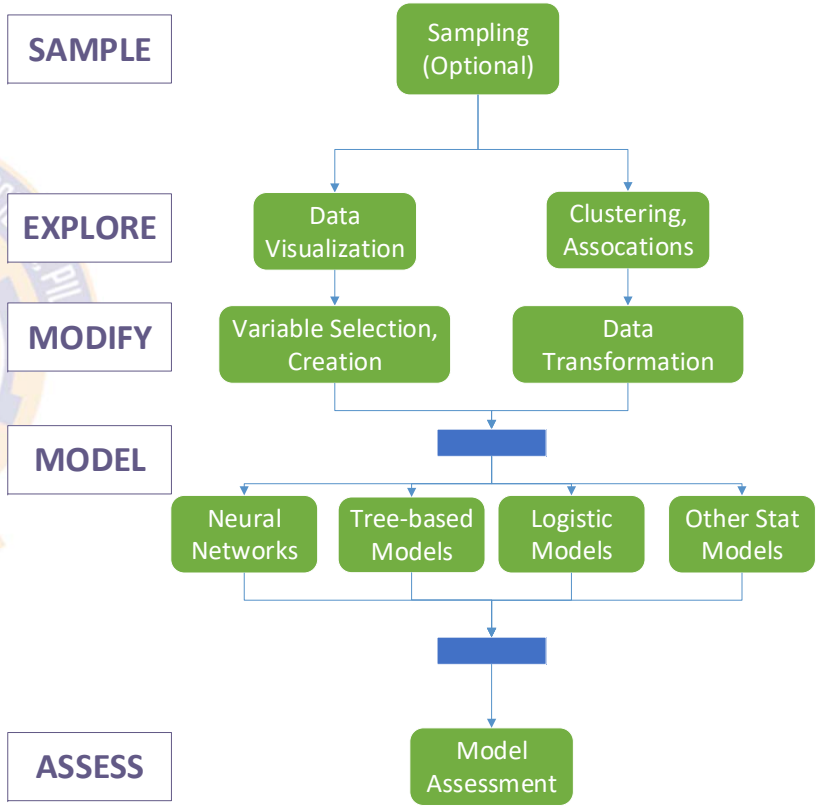
Source: SAS® Enterprise Miner™ 14.3: Reference Help
Big Data Analytics Methods, 2nd Edition by Ghavami P

Data Analytics Methodologies

innovate achieve lead

SEMMA Process Model

- SEMMA is actually not a data mining methodology
- It is a logical organization 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



Source: SAS® Enterprise Miner™ 14.3: Reference Help
Big Data Analytics Methods, 2nd Edition by Ghavami P

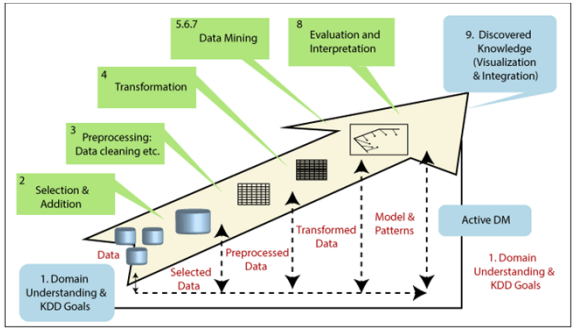
Comparison of Methodologies

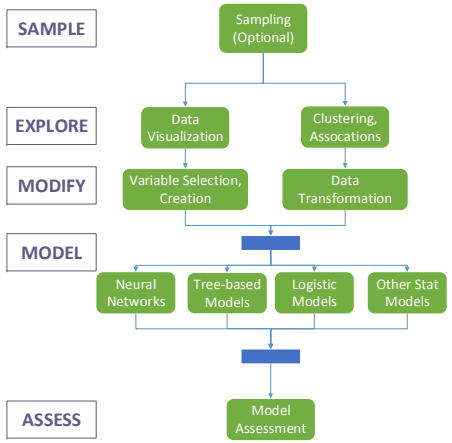
Data Analytics Methodologies

Comparison of KDD Vs. SEMMA

- By comparing KDD with SEMMA stages we can see that they are very similar

SEMMA	KDD
Sample phase	Can be identified with *Selection*
Explore phase	Can be identified with *Preprocessing*
Modify phase	Can be identified with *Transformation*
Model phase	Can be identified with *Data Mining*
Assess phase	Can be identified with *Interpretation/Evaluation*





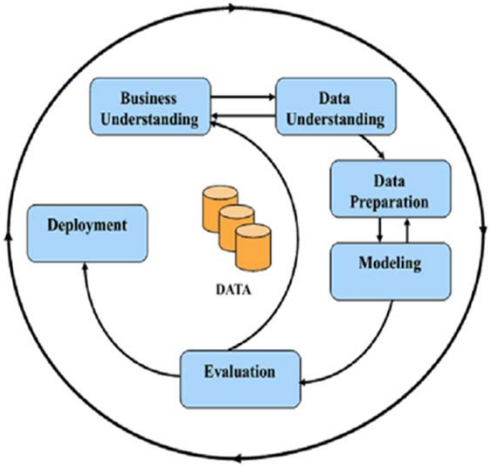
- Close examination tells us that the five stages of the SEMMA process can be seen as a practical implementation of the five stages of the KDD process, since it is directly linked to the SAS Enterprise Miner software.

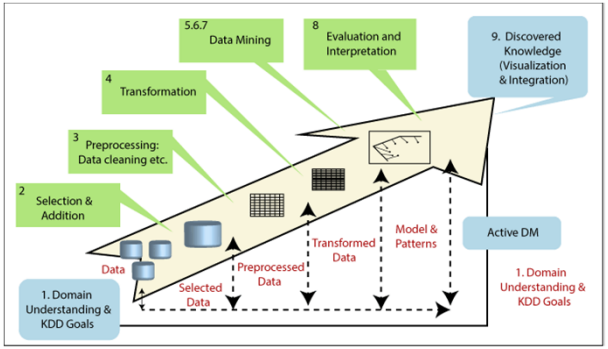
Data Analytics Methodologies

innovate achieve lead

Comparison of KDD Vs. CRIPS-DM

- Comparing the KDD stages with the CRISP-DM stages is not as straightforward as in SEMMA
- CRISP-DM methodology incorporates the steps that, as referred above, must precede and fo process that is to say:
- The Business Understanding phase
  - Can be identified with the development of an understanding of the application domain, the relevant prior knowle the end-user;
- The Data Understanding phase
  - Can be identified as the combination of Selection and Pre processing;
- The Deployment phase
  - Can be identified with the consolidation by incorporating this knowledge into the system.
- The Data Preparation phase
  - Can be identified with Transformation;
- The Modeling phase
  - Can be identified with Data Mining;
- The Evaluation phase
  - Can be identified with Interpretation/Evaluation.





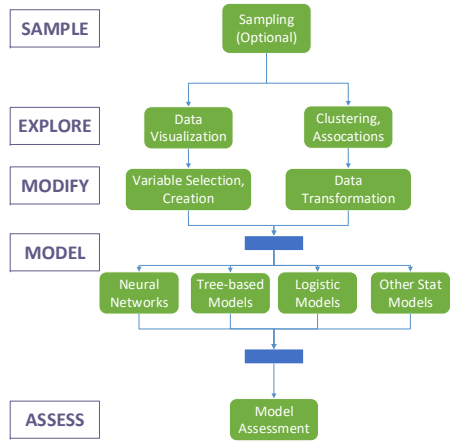
Data Analytics Methodologies

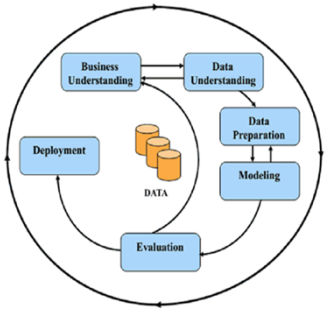
innovate achieve lead

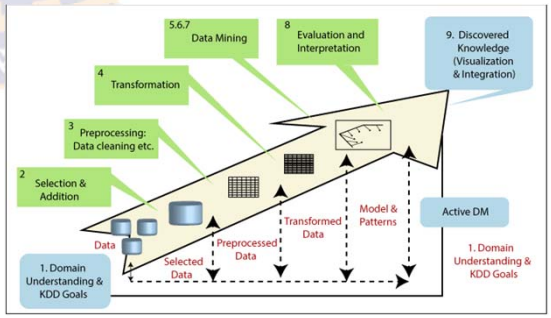
KDD Vs. SEMMA Vs. CRIPS-DM

Table 1. Summary of the correspondences between KDD, SEMMA and CRISP-DM

KDD	SEMMA	CRISP-DM
Pre KDD	----------	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	----------	Deployment



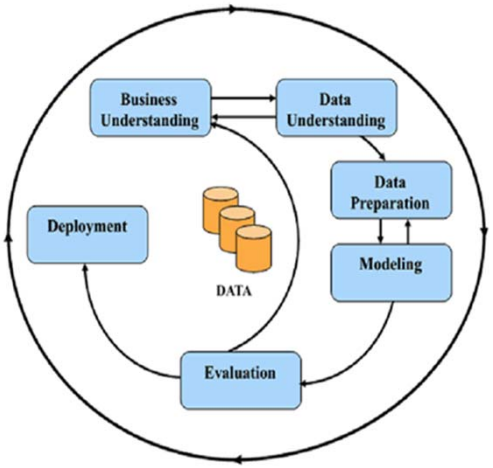




Data Analytics Methodologies

Drawbacks of CRISP-DM-1

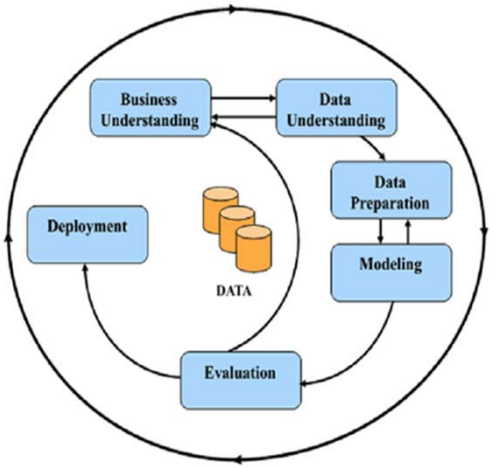
- One major drawback is that the model no longer seems to be actively maintained
- The official site, CRISP-DM.org, is no longer being maintained
- Furthermore, the framework itself has not been updated for working with new technologies, such as big data
- Big data technologies means that there can be additional effort spend in the data understanding phase
  - For example, as the business grapples with the additional complexities that are involved in the shape of big data sources.



Data Analytics Methodologies

Drawbacks of CRISP-DM-1

- The methodology itself was conceived in mid 1990s
- Gregory Piatetsky of KDNuggets says:
  - *"CRISP-DM remains the most popular methodology for analytics, data mining, and data science projects, with 43% share in latest KDnuggets Poll, but a replacement for unmaintained CRISP-DM is long overdue."*

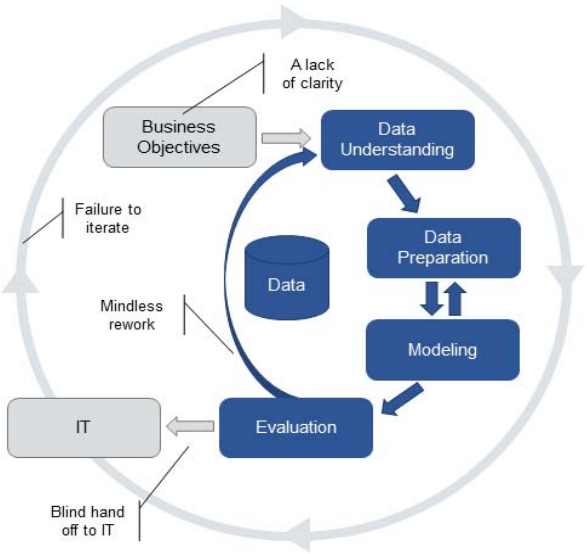


Data Analytics Methodologies

innovate achieve lead

Drawbacks of CRISP-DM-2

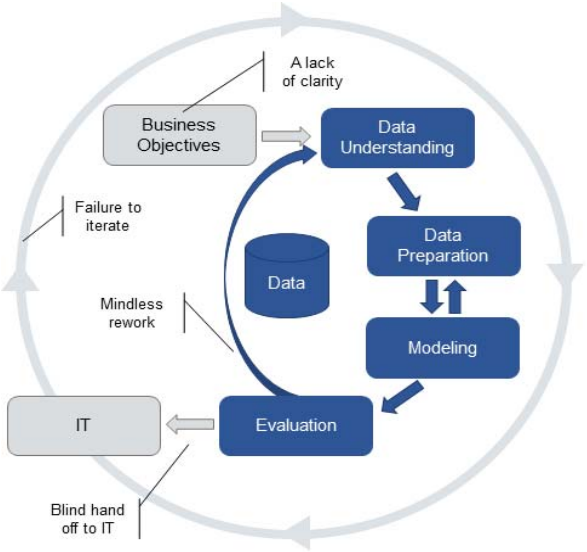
- CRISP-DM is a great framework and its use on projects helps focus them on delivering real business value
- CRISP-DM has been around a long time so many projects are implemented using CRISP-DM
- However, projects do take shortcuts
- Methodology translate into frequent friction points between the business and IT professionals
- Some of these shortcuts make sense but too often they result in projects using a corrupted version of the approach like the one shown in Figure



Data Analytics Methodologies

Advantages of CRISP-DM

- A lack of clarity
  - Rather than drilling down into the details to get clarity on both the business problem and the analytic approach, the project team work superficially with business goals and use some metrics to measure success.
- Mindless rework:
  - Most try to find new data or new modeling techniques rather than working with their business partners to re-evaluate the business problem.
- Blind hand-off to IT:
  - Some analytic teams don't think about deployment and operationalization of their models at all
  - They fail to recognize that the models they build will have to be applied to live data in operational data stores or embedded in operational systems.
- Failure to iterate
  - Analytic professionals know that models age and that models need to be kept up to date if they are to continue to be valuable.



innovate achieve lead

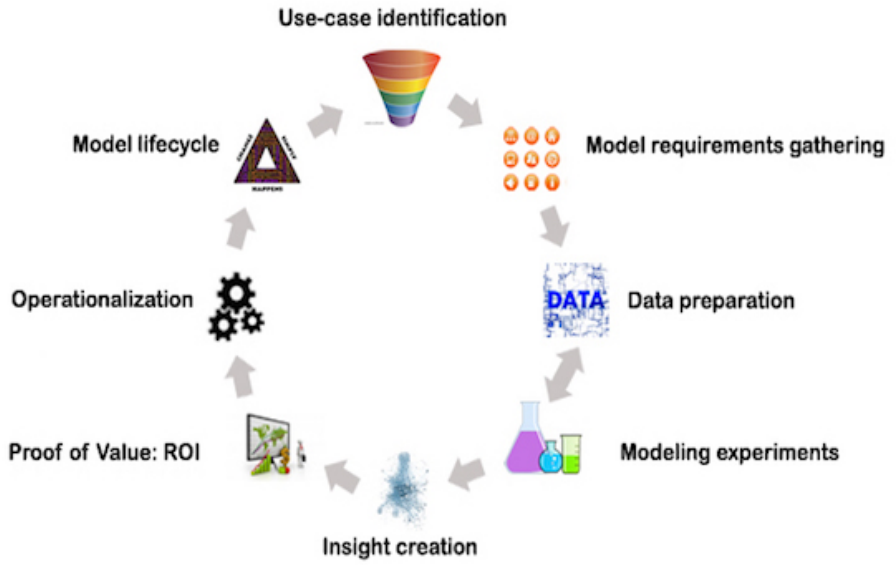
Standard Methodology for Analytical Models (SMAM)

Data Analytics Methodologies

innovate achieve lead

Standard Methodology for Analytics Models (SMAM)

- Here we will look into another methodology called SMAM, which is supposed to overcome those friction points



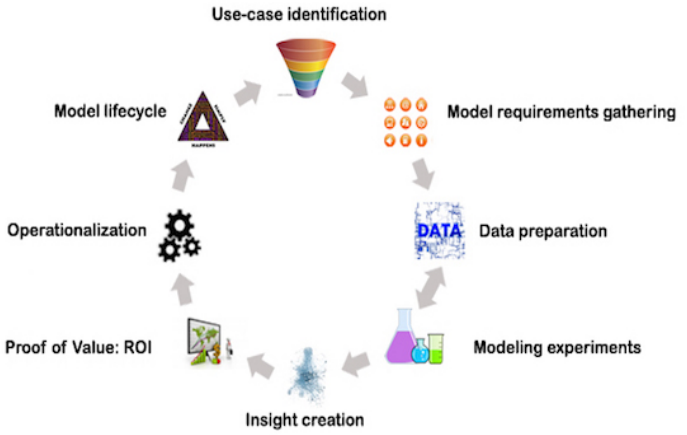
Source: https://www.kdnuggets.com/2015/08/new-standard-methodology-analytical-models.html

Data Analytics Methodologies

innovate achieve lead

Standard Methodology for Analytics Models (SMAM)

Phase	Description
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
Modeling 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



Source: https://www.kdnuggets.com/2015/08/new-standard-methodology-analytical-models.html

Data Analytics Methodologies

innovate achieve lead

Standard Methodology for Analytics Models (SMAM)

- Use case identification phase
  - Describes the brainstorming/discovery process of looking at the different areas where models may be applicable
  - It involves educating business parties involved about what analytical modeling is, what realistic expectations are from the various approaches and how models can be leveraged in the business
  - Discussions on the use-case identification involve topics around data availability, model integration complexity, analytical model complexity and model impact on the business
  - From a list of identified use-cases in an area, the one with the best ranking on above mentioned criteria should be considered for implementation
  - Parties involved in this phase are (higher) management to ensure the right goal setting, IT for data availability, the involved department for model relevance checking and the data scientists to infuse the analytical knowledge and creative analytical idea provisioning
  - The result of this phase is a chosen use-case, and potentially a roadmap with the other considered initiatives on a timeline.

Source: https://www.kdnuggets.com/2015/08/new-standard-methodology-analytical-models.html

Data Analytics Methodologies

Standard Methodology for Analytics Models (SMAM)

- Model requirements gathering
  - In this phase, requirements are gathered for the use-case selected in the previous phase
  - These requirements include:
    - Conditions for the cases/customers/entities considered for scoring
    - Requirements about the availability of data, initial ideas around model reporting can be explored and finally, ways that the end-users would like to consume the results of the analytical models
    - Parties involved in this phase are people from the involved department(s), the end-users and the data scientists
    - The result of this phase is a requirements document.

Source: https://www.kdnuggets.com/2015/08/new-standard-methodology-analytical-models.html

Data Analytics Methodologies

innovate achieve lead

Standard Methodology for Analytics Models (SMAM)

- Data preparation
  - In this phase, discussions revolve around data access, data location, data understanding, data validation, and creation of the modeling data
  - This is a phase where IT/data administrators/DBA's closely work together with the data scientist to help prepare the data in a format consumable by the data scientist
  - The process is agile; the data scientist tries out various approaches on smaller sets and then may ask IT to perform the required transformation in large
  - As with the CRISP-DM, the previous phase, this phase and the next happen in that order, but often iterate back and forth
  - Parties involved
    - The involved parties are IT/data administrators/DBA/data modelers and data scientists
  - The results of this phase should be the data scientist being convinced that with the data available, a model is viable, as well as the scoring of the model in the operational environment

Source: https://www.kdnuggets.com/2015/08/new-standard-methodology-analytical-models.html

Data Analytics Methodologies

Standard Methodology for Analytics Models (SMAM)

- Modeling experiments
  - In this phase, data scientists play with the data; crack the nut; trying to come up with the solution that is both cool, elegant and working
  - Results are not immediate; progress is obtained by evolution and by patiently collecting insights to put them together in an ever evolving model
  - At times, the solution at hand may not look viable anymore, and an entire different angle needs to be explored, seemingly starting from scratch
  - The data scientist may need to connect to end-user to validate initial results, or to have discussion to get ideas which can be translated into testable hypotheses/ model features
  - The result of this phase is an analytical model that is evaluated in the best possible way with the (historic) data available as well as a reporting of these facts.

Source: https://www.kdnuggets.com/2015/08/new-standard-methodology-analytical-models.html

Data Analytics Methodologies

Standard Methodology for Analytics Models (SMAM)

- Insight creation
  - Dashboards and visualization are critically important for the acceptance of the model by the business
  - This outcomes of this phase are analytic reporting and operational reporting
  - Analytical reporting refers to any reporting on data where the outcome (of the analytical model) has already been observed – **descriptive modeling**
  - This data can then be used to understand the performance of the model and the evolution of performance over time
  - Operational reporting refers to any reporting on the data where the outcome has not yet been observed – **predictive modeling**
  - This data can be used to understand what the model predicts for the future in an aggregated sense and is used for monitoring purposes
  - The involved parties are the end-users, business department, and the data scientists
  - The result of this phase is a set of visualizations and dashboards that provide a clear view on the model effectiveness and provide business usable insights.

Source: https://www.kdnuggets.com/2015/08/new-standard-methodology-analytical-models.html

Data Analytics Methodologies

innovate achieve lead

Standard Methodology for Analytics Models (SMAM)

- Proof of value (ROI)
  - This phase involves deploying the model on a small scale to ensure return on investment
  - If the ROI is positive, the business will be convinced that they can trust the model
  - A decision can be made if the model should be deployed or not
- Operationalization
  - This phase involves putting the model into production and operationalization of the model by closely working with the IT department
  - The model is handed over the maintenance team for support

Source: https://www.kdnuggets.com/2015/08/new-standard-methodology-analytical-models.html

Data Analytics Methodologies

innovate achieve lead

Standard Methodology for Analytics Models (SMAM)

- Model lifecycle
  - An analytical model in production will not be fit forever
  - Depending on how fast the business changes, the model performance degrades over time.
  - Generally, two types of model changes can happen: refresh and upgrade
  - In a model refresh, the model is trained with more recent data, leaving the model structurally untouched
  - The model upgrade is typically initiated by the availability of new data sources and the request from the business to improve model performance by the inclusion of the new sources
  - The involved parties are the end-users, the operational team that handles the model execution, the IT/data administrators/DBA for the new data and the data scientist

Source: https://www.kdnuggets.com/2015/08/new-standard-methodology-analytical-models.html

Big Data Life Cycle

Data Analytics Methodologies

innovate achieve lead

Big Data Life Cycle

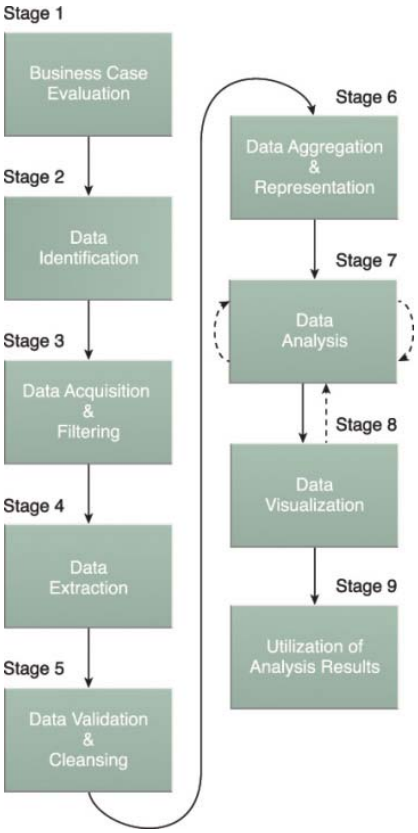
- Big Data analysis differs from traditional data analysis primarily due to the volume, velocity, variety, and veracity characteristics of the data
- Massive amounts of data - It is big – typically in terabytes or even petabytes
- Varied data formats such as video files to SQL databases to text data
  - It could be a traditional database, it could be video data, log data, text data or even voice data
- Data that comes in at varying frequency – from days to minutes - It keeps increasing as new data keeps flowing in
- To address these distinct aspects of Big Data, a step-by-step methodology is needed to organize and manage the tasks and activities associated with Big Data analysis
- These activities and tasks involve acquiring, processing, analyzing and repurposing data
- In addition to the lifecycle, it is also important to consider the issues of training, education, tooling and staffing of a data analytics team

Source: Big Data Fundamentals: Concepts, Drivers & Techniques by Wajid Khattak; Paul Buhler; Thomas Erl

Data Analytics Methodologies

Big Data Life Cycle

- The Big Data analytics lifecycle can be divided into nine stages:
  1) Business Case Evaluation
  2) Data Identification
  3) Data Acquisition & Filtering
  4) Data Extraction
  5) Data Validation & Cleansing
  6) Data Aggregation & Representation
  7) Data Analysis
  8) Data Visualization
  9) Utilization of Analysis Results



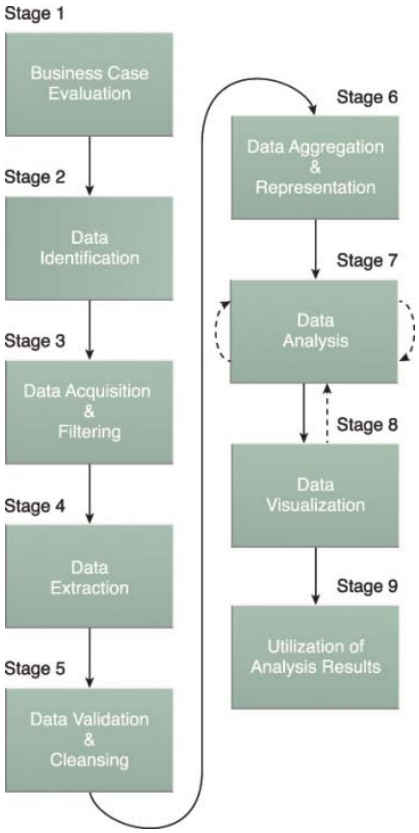
Source: Big Data Fundamentals: Concepts, Drivers & Techniques by Wajid Khattak; Paul Buhler; Thomas Erl

Data Analytics Methodologies

innovate achieve lead

Big Data Life Cycle

- Business Case Evaluation
  - This stage involves creating, assessing, and approving a business case prior to proceeding to next stage
  - The business case presents:
    - Clear requirements
    - Justification for the analytics project
    - Motivation
    - Goals for carrying out the analysis
    - Budget
    - Timeline
    - Resources



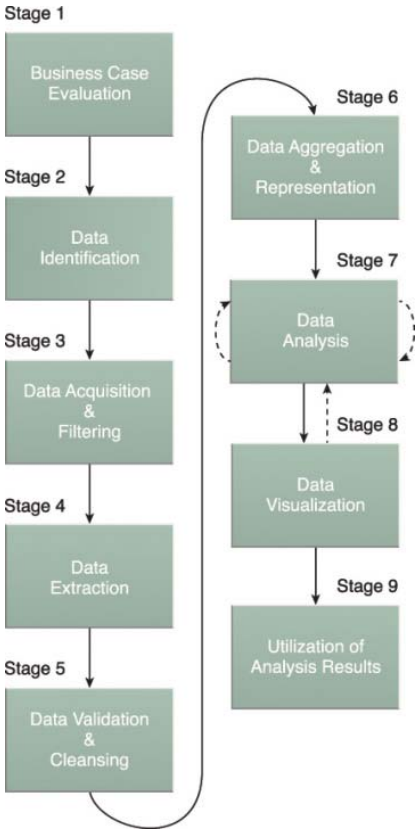
Source: Big Data Fundamentals: Concepts, Drivers & Techniques by Wajid Khattak; Paul Buhler; Thomas Erl

Data Analytics Methodologies

innovate achieve lead

Big Data Life Cycle

- Data Identification
  - This stage involves identifying the data and their sources required for the analysis
  - A wider variety of data sources may increase the probability of finding hidden patterns and correlations
    - For example, it benefits to look into as many types of resources as possible, particularly when it's not clear exactly what to look for
    - However, there is a tradeoff between the value gained and the investment in terms of resources, time, and budget
  - These dataset can be internal or external to the enterprise depending on the project objectives
  - Internal data sources include
    - Data Warehouse, Data marts, Operational Systems, etc.,.
  - External data sources include
    - Third-party datasets such as market trends, stock data analysis; Government supplied datasets, and other publicly available datasets.
    - Sometimes, data may be needed to be harvested using automated tools from weblogs (blogs) and other content-based web sites

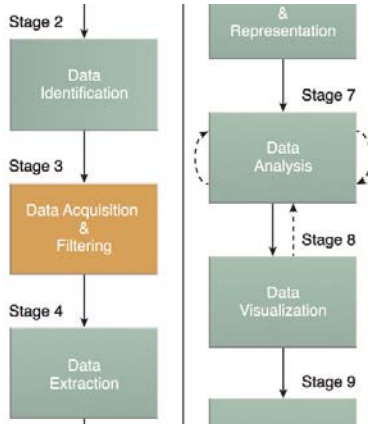


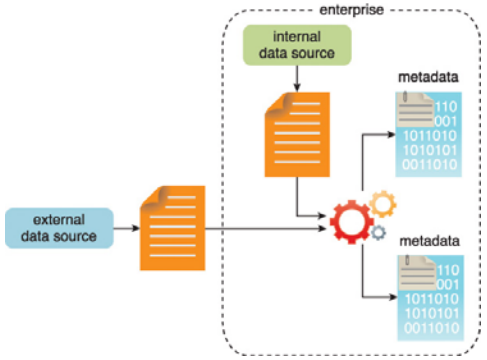
Source: Big Data Fundamentals: Concepts, Drivers & Techniques by Wajid Khattak; Paul Buhler; Thomas Erl

Data Analytics Methodologies

Big Data Life Cycle

- Data Acquisition & Filtering
  - Here, data is gathered from all of the data sources that were identified during the previous stage
  - Depending on the type of data source, data may come as a collection of files, such as data purchased from a third-party data provider, or may require API integration, such as with Twitter
  - In many cases, some or most of the acquired data may be irrelevant (noise) and can be discarded as part of the filtering process
  - Metadata can be added via automation to data from both internal and external data sources to improve the classification and querying
  - For example:
    - Dataset size and structure, source information, date and time of creation or collection and language-specific information
  - Metadata should be machine-readable so that it can be passed forward along subsequent analysis stages
  - This helps maintain the origins of data throughout the Big Data analytics lifecycle, which helps to establish and preserve data accuracy and quality



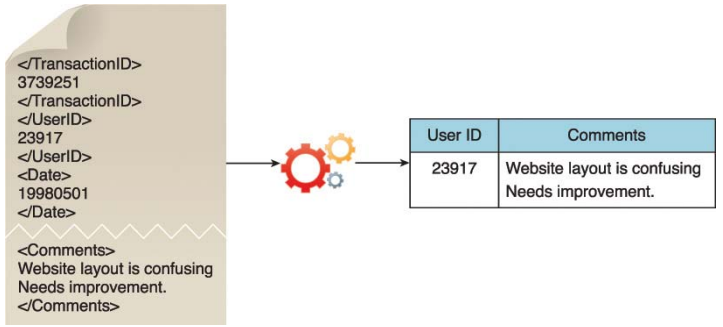


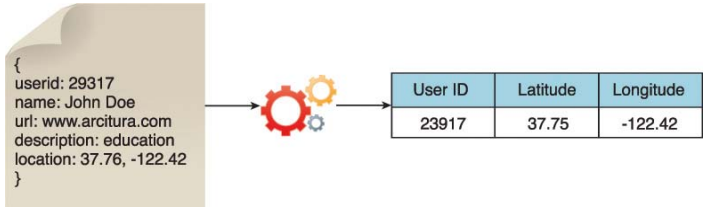
Source: Big Data Fundamentals: Concepts, Drivers & Techniques by Wajid Khattak; Paul Buhler; Thomas Erl

Data Analytics Methodologies

Big Data Life Cycle

- Data Extraction and Transformation
  - Due to disparate external data sources, data may arrive in a format incompatible with the Big Data solution
    - XML, JSON Files
    - Tab or comma or other delimited files
    - Raw text files
  - The extent of extraction and transformation required depends on the types of analytics and capabilities of the Big Data solution
    - For example: Merging fields, Separating data fields into individual components, etc





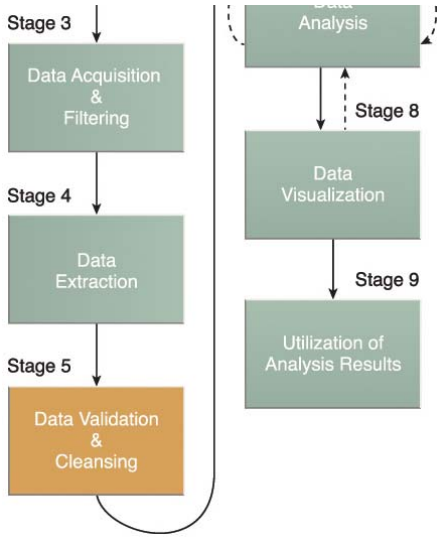
Source: Big Data Fundamentals: Concepts, Drivers & Techniques by Wajid Khattak; Paul Buhler; Thomas Erl

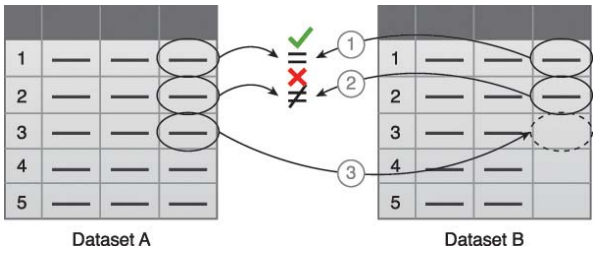
Data Analytics Methodologies

innovate achieve lead

Big Data Life Cycle

- Data Validation & Cleansing
  - Internal data usually has a pre-defined structured and pre-validated
  - External data can be unstructured and without any indication of validity
  - This stage is dedicated to establishing complex validation rules and removing any known invalid data
  - For example:
    - Redundant data across different datasets
      - This redundancy can be exploited to explore interconnected datasets in order to assemble validation parameters and fill in missing valid data
  - The data validation and cleansing can be a batch process during ETL or it can be a real-time in-memory operation





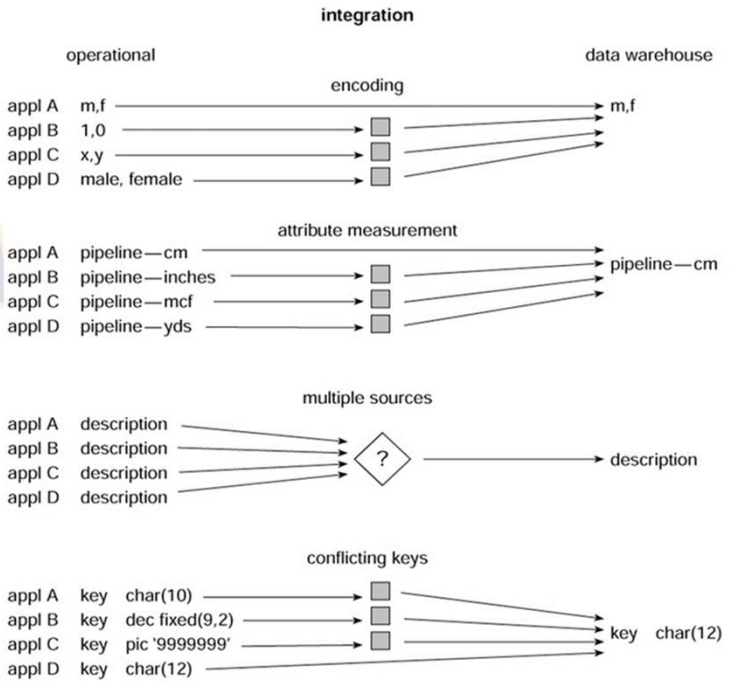
Source: Big Data Fundamentals: Concepts, Drivers & Techniques by Wajid Khattak; Paul Buhler; Thomas Erl

Data Analytics Methodologies

innovate achieve lead

Big Data Life Cycle

- Data Aggregation & Representation
  - This stage is dedicated to integrating multiple datasets together to arrive at a unified view.
  - Data spread across multiple datasets may need to be joined together via common fields such as Date or ID
  - Data may need to be reconciled because same data fields may appear in multiple datasets
    - E.g., date of birth
  - Data aggregation can become complicated because of differences in:
    - Data Structure – Although the data format may be the same, the data model may be different
    - Semantics – A value that is labeled differently in two different datasets may mean the same thing, for example "surname" and "last name"



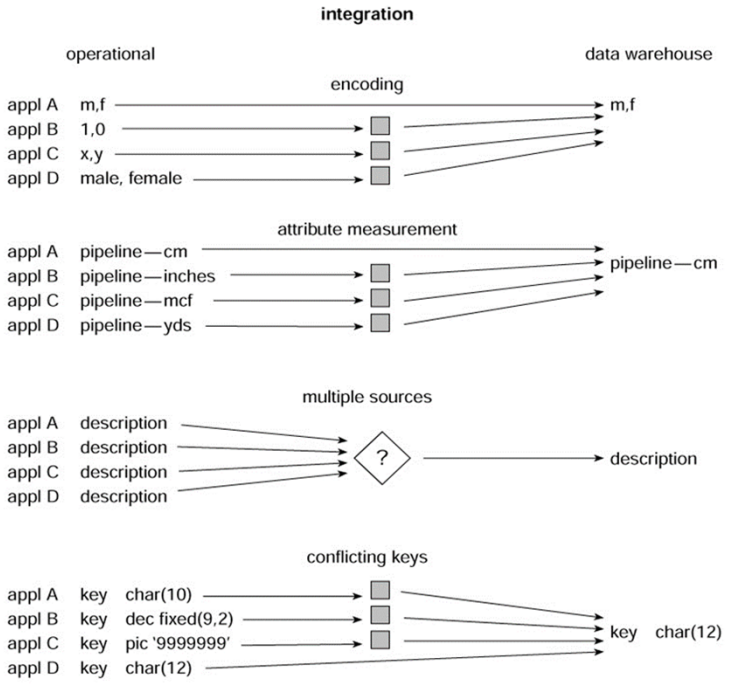
Source: Big Data Fundamentals: Concepts, Drivers & Techniques by Wajid Khattak; Paul Buhler; Thomas Erl

Data Analytics Methodologies

innovate achieve lead

Big Data Life Cycle

- Data Aggregation and Representation
  - Very often they follow different naming conventions and varied standards for data representation
    - E.g., Cusomer_Identifier, Customer_Id, Cust_Id, CID
  - Names have to be standardized and discrepancies should be resolved
  - Integrating the data involves combining all the relevant operational data into coherent data structures to be made ready for loading into the data warehouse.
  - Data integration and consolidation is a type of preprocess before other major transformation routines are applied

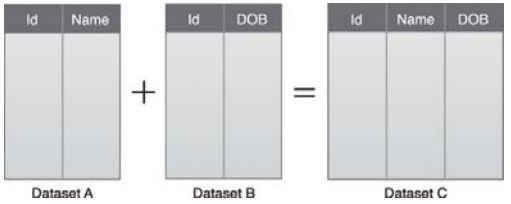


Data Analytics Methodologies

innovate achieve lead

Big Data Life Cycle

- Data Aggregation & Representation
  - Data aggregation can be a time and effort intensive operation because of large volumes of data
  - Future data analysis requirements need to be considered during this stage to help foster data reusability
  - It is important to understand that the same data can be stored in many different forms
  - One form may be better suited for a particular type of analysis than another
    - For example, data stored as a BLOB would be of little use if the analysis requires access to individual data fields
  - A standardized data structure for the Big Data solution can be used for a range of analysis techniques and projects



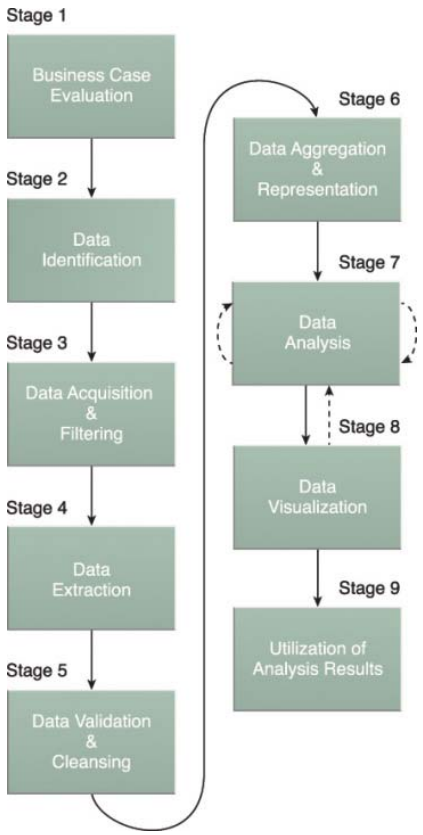
Source: Big Data Fundamentals: Concepts, Drivers & Techniques by Wajid Khattak; Paul Buhler; Thomas Erl

Data Analytics Methodologies

innovate achieve lead

Big Data Life Cycle

- Data Analysis
  - This stage is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics
  - This stage can be iterative in which case analysis is repeated until the appropriate pattern or correlation is uncovered
  - Depending on the type of analytic result required, this stage can be
    - as simple as querying a dataset to compute an aggregation for comparison
    - as challenging as combining data mining and complex statistical analysis techniques to discover patterns, anomalies, or to generate a statistical or mathematical model to depict relationships between variables.

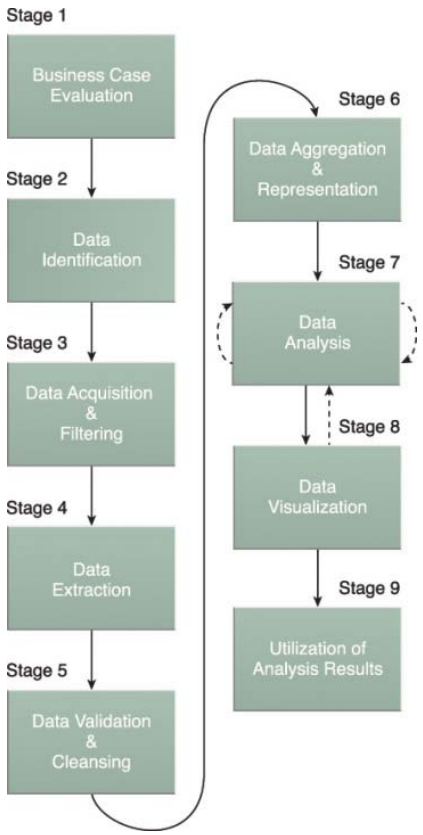


Source: Big Data Fundamentals: Concepts, Drivers & Techniques by Wajid Khattak; Paul Buhler; Thomas Erl

Data Analytics Methodologies

Big Data Life Cycle

- Data Visualization
  - This stage is dedicated to using visualization techniques and tools to graphically communicate the results of analysis for effective interpretation
  - Business users need to be able to understand the results in order to obtain value from the data analysis and subsequently have the ability to provide feedback from stage 8 back to stage 7
  - Data Visualization provides users with the ability to perform visual analysis, allowing for the discovery of answers to unknown questions
  - Data Visualization helps in presenting same results in a number of different ways, which can influence the interpretation of the results.
  - Providing a method of drilling down to comparatively simple statistics is crucial, in order for users to understand how the rolled up or aggregated results were generated.



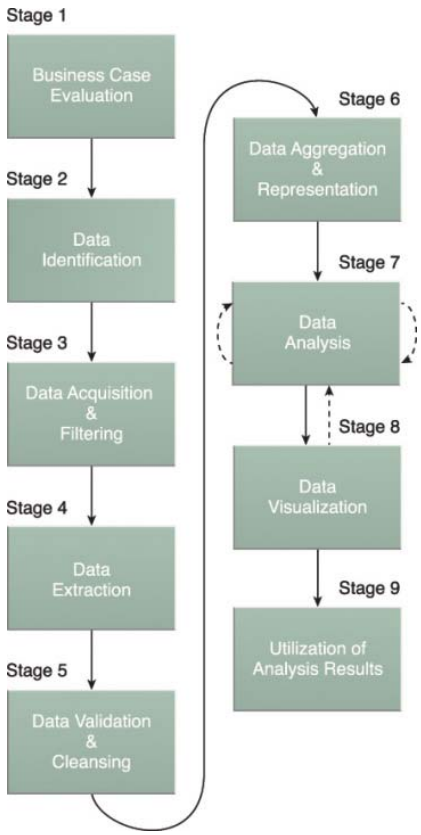
Source: Big Data Fundamentals: Concepts, Drivers & Techniques by Wajid Khattak; Paul Buhler; Thomas Erl

Data Analytics Methodologies

innovate achieve lead

Big Data Life Cycle

- Utilization of Analysis Results
  - This stage involves determining how and where processed analysis data can be further leveraged
  - The results of the analysis may produce "models" that can be used on a new set of dataset to understand patterns and relationships that exist within the data
  - A model may look like a mathematical equation or a set of rules
  - Models can be used to improve business process logic and application system logic, and they can form the basis of a new system or software program



Source: Big Data Fundamentals: Concepts, Drivers & Techniques by Wajid Khattak; Paul Buhler; Thomas Erl

Data Analytics Methodologies

Big Data Life Cycle Vs. CRISP-DM

- A comparison of stages involved in Big Data Life Cycle and CRISP-DM
  - Business Problem Definition – CRISP DM : Business understanding
  - Research -- new step
  - Human Resources Assessment – new step
  - Data Acquisition – new step
  - Data Munging – new step
  - Data Storage – new step
  - Exploratory Data Analysis – CRISP DM : Data Understanding
  - Data Preparation for Modeling and Assessment – CRISP DM : Data Preparation
  - Modeling – CRISP DM : Data Modelling
  - Implementation – CRISP DM : Evaluation & Deployment

Data Analytics Methodologies

innovate achieve lead

Big Data Life Cycle – Possible case studies

- What is the precision agriculture? Why it is a likely answer to climate change and food security?
  - https://www.youtube.com/watch?v=581Kx8wzTMc&ab_channel=Inter-AmericanDevelopmentBank
- Innovating for Agribusiness
  - https://www.youtube.com/watch?v=C4W0qSQ6A8U
- How Big Data Can Solve Food Insecurity
  - https://www.youtube.com/watch?v=4r_IxShUQuA&ab_channel=mSTARProject
- AI for AgriTech ecosystem in India- IBM Research
  - https://www.youtube.com/watch?v=hhoLSI4bW_4&ab_channel=IBMIndia
- Bringing Artificial Intelligence to agriculture to boost crop yield
  - https://www.youtube.com/watch?v=GSvT940uS_8&ab_channel=MicrosoftIndia


innovate
achieve
lead

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
ज्ञानं परमं बलम्

Thank You!