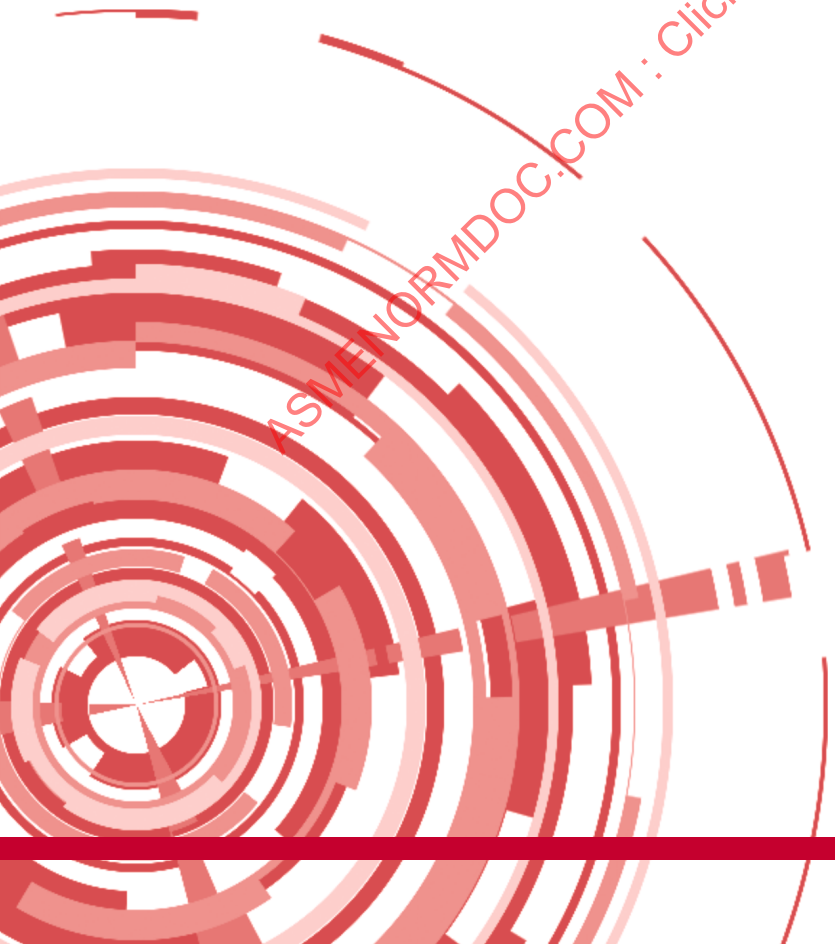
ASME STB-1-2020

Guideline on Big Data/Digital Transformation Workflows and Applications for the Oil and Gas Industry



ASME
SETTING THE STANDARD

STB-1-2020

GUIDELINE ON BIG DATA/DIGITAL TRANSFORMATION WORKFLOWS AND APPLICATIONS FOR THE OIL AND GAS INDUSTRY

Prepared by:

Barbara Thompson, P.E.
TeamBS, LLC.

ASME
SETTING THE STANDARD

Date of Issuance: December 31, 2020

This publication was prepared by ASME Standards Technology, LLC (“ASME ST-LLC”) and sponsored by The American Society of Mechanical Engineers (“ASME”), Petroleum Division.

The American Society of Mechanical Engineers
Two Park Avenue, New York, NY 10016-5990
ISBN No. 978-0-7918-7399-1

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FOREWORD

Guideline Description

This guideline is the culmination of the efforts of ASME industry professionals in oil and gas to define Big Data and its useful applications to upstream, midstream and downstream businesses.

The concept of Big Data intimidates decision-makers and business leadership. While the introduction of new digital technologies is the cornerstone of the industry's digital transformation, use of Big Data requires the sharing of asset-specific or operational data of a size and scope that is difficult to grasp. Industry data may also be of poor quality, requiring significant effort and resources to cleanse and aggregate prior to its analysis. In addition, the gathering, aggregation, analysis, and data storage and maintenance is very expensive. The industry needs a standardized and efficient end-to-end workflow to validate the quality of the data, and subsequently, the accuracy of the results from the analysis.

The goal of this document is to alleviate that intimidation by providing a framework for understanding and a workflow for utilizing data analytical techniques to solve business problems; without requiring the reader to be a full-time statistician or data scientist professional.

Who Should Use this Guideline?

The design of this Guideline is intended to be universal in its application to Big Data challenges in the oil and gas industry. It is written not only for oil and gas professionals who are beginners to Big Data techniques but also for data professionals looking to contribute to unique oil and gas applications.

Specific users can be characterized as:

Citizen Data Scientist: a subject matter expert in engineering, operations, supply chain, planning, project management or operations that requires data insights.

Early Career Engineer: a young professional that is looking to improve his or her career by adding a data dimension to their problem solving.

Data Scientist Professional: a data science professional that is looking to apply his or her deep experience in data analytics by learning the unique sets of data and operational challenges of the oil and gas industry.

The author acknowledges, with deep appreciation, the activities of the ASME volunteers and staff who have provided valuable technical input, advice and assistance with review of, commenting on, and editing of, this document, particularly the activities of the Peer Review Group (PRG) consisting of Michael Wells, Thalia Kruger, Amit Kumar, Mete Mutlu, Brian Webster, Kathryn Hyam, and the activities of the ASME Petroleum Division Leaders and Petroleum Division Big Data Task Group consisting of Jim Kaculi, Ed Marotta, John O'Brien, and Jamie Hart.

Established in 1880, ASME is a professional not-for-profit organization with more than 100,000 members and volunteers promoting the art, science and practice of mechanical and multidisciplinary engineering and allied sciences. ASME develops codes and standards that enhance public safety, and provides lifelong learning and technical exchange opportunities benefiting the engineering and technology community. Visit https://www.asme.org/ for more information.

ASME ST-LLC is a not-for-profit Limited Liability Company, with ASME as the sole member, formed in 2004 to carry out work related to new and developing technology. ASME ST-LLC's mission includes meeting the needs of industry and government by providing new standards-related products and services, which advance the application of emerging and newly commercialized science and technology, and providing the research and technology development needed to establish and maintain the technical relevance of codes and standards. Visit http://asmestllc.org/ for more information.

1 PURPOSE, DEFINITIONS AND REFERENCES

1.1 Scope

This guideline explains the current use and application of data analytics and data science in the oil and gas industry. It is designed to provide guidance on how to utilize data analytics and machine learning/artificial intelligence (ML/AI) to address a given business need, resulting in value-creation. Within the guidelines will be descriptions of various data analytics techniques and the recommended tools for the respective techniques.

1.1.1 How to Use This Guideline

(a) Basics

This document is designed as both a "how-to" and a reference document. The chapters are arranged in a building block sequence that parallels the journey of a business data project. These building blocks should help provide a roadmap to data-driven projects.

Each chapter is also a standalone reference for its respective topic. The reader can reference these individual chapters as needed to fill gaps in his or her understanding of either data techniques or oil and gas.

(b) Chapters

In this first chapter, the user is introduced to definitions and acronyms common to both oil and gas and to the data analytics strategies highlighted in this guide.

Chapter 2 provides background to the various types of data in the oil and gas industry, including how they are curated, described, used, and safeguarded.

Chapter 3 describes the Digital Facility and the related oil and gas activities and operations that produce data. It also defines the types of data produced by each operation and potential applications for data-driven insights.

Chapter 4 explores the types of data analytics tools available to a data project leader and how they can be utilized. This chapter is where the user can find information about topics such as regression models, classification, machine learning, optimization, and data visualization.

Chapter 5 is a description of a full data project. The user will learn how to structure a project, assemble the appropriate team resources, frame the questions to be answered, model and execute the project, and usefully deploy the resulting model.

Two sets of Appendices are included for references and details that supplement the discussions of the chapters. The Mandatory Appendices are presented to help understand data projects and the team members required to execute them. Nonmandatory Appendices are included to provide the user with additional definitions, sample case studies as examples, and a description of relevant certifications available if the reader wishes to increase his or her knowledge and proficiency in data analytics.

1.2 Definitions

The definitions and acronyms in this section are selected based on relevance to computer science, data analytics, computer programming, project management principles, oil and gas terms, and mathematics.

Term	Definition
Engineering, Procurement, Construction (EPC) Contractor Company	a business enterprise that engineers, procures, and constructs operational facilities, buildings or other structures.
5S Lean Process	a workplace organization method that uses a list of five Japanese words *seiri* (sort), *seiton* (set in order), *seisō* (shine), *seiketsu* (standardize), and *shitsuke* (sustain) to describe how to organize a work space for efficiency and effectiveness by identifying and storing the items used, maintaining the area and items, and sustaining the new order. The decision-making process usually comes from a dialogue about standardization, which builds understanding among employees of how they should do the work.
activity[4]	part of a task in the User Guide; describes actions to perform a task
analytics[8][17]	the discovery, interpretation, and communication of meaningful patterns in data. It also entails applying data patterns towards effective decision-making for potentially beneficial business outcomes. Is the synthesis of knowledge from information. Analytics relies on the simultaneous application of statistics, computer programming, operations research, and ML/AI to quantify performance.
Application Program Interface (API)[16]	a set of programming standards and instructions for accessing or building web-based software applications.
application[18]	computer software that enables a computer to perform a certain task
Approved For Construction (AFC)	drawings and documents that have been designed, reviewed and ready to fabricate or manufacture.
ASCII American Standard Code for Information Interchange	a set of digital codes representing letters, numerals, and other symbols, widely used as a standard format in the transfer of text between computers.
asset integrity	activities and processes executed to monitor, maintain and repair capital assets of an operating facility to ensure the assets continue to perform safely and as designed. These activities and processes are used to extend the life of the asset.
asset management	the practice of maintenance, inspection, repair and surveillance of equipment (objects) within a physical operation.
augmented reality (AR)[6]	an interactive experience of a real-world environment where the objects that reside in the real world are enhanced by computer-generated perceptual information, sometimes across multiple sensory modalities, including visual, auditory, haptic, somatosensory and olfactory.
average squared error	in statistics, a measure of the average or mean of the squares of the errors – which are the square of the difference between estimated values and actual values. See also, mean squared error

Term	Definition
Big Data[17]	extremely large datasets primarily in the characteristics of volume, variety, velocity, and/or variability that require a scalable architecture for efficient storage, manipulation, and analysis.
Big Data Application Provider[17]	the entity that executes a life cycle to meet security and privacy requirements as well as System Orchestrator-defined requirements
Big Data engineering[17]	includes advanced techniques that harness independent resources for building scalable data systems.
Big Data Scientist[3]	person responsible for developing algorithms to operationalize big data. See also Data Scientist.
central limit theorem	in probability theory, in many situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution, even if the original variables are not normally distributed.
central tendencies	another term for averages in a probability distribution.
classification analysis[3]	a systematic process for obtaining important and relevant information about data, also called meta data (data about data).
cloud[3]	a broad term that refers to any internet-based application or service that is hosted remotely.
cloud computing[16]	a distributed computing system hosted and running on remote servers and accessible from anywhere on the internet.
columnar database or column-oriented database[16]	a database that stores data by column rather than by row. In a row-based database, a row might contain a name, address and phone number. In a column-oriented database, all names are in one column, addresses in another and so on. A key advantage of a columnar database is faster hard disk access.
computer generated data[2]	data generated by computers such as log files
confidence interval[16]	a range of values which is likely to contain the population parameter of interest with a given level of confidence.
confusion matrix[5]	a tool that defines the performance of a classification model on a test data for which the true values are known. This matrix uses the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) to evaluate the performance.
correlation analysis[16]	a means to determine a statistical relationship between variables, often for the purpose of identifying predictive factors among the variables. A technique for quantifying the strength of the linear relationship between two variables.
CRISP-DM methodology[4]	the general term for all concepts developed and defined in CRISP-DM
cross-validation[5]	the process of splitting the labelled data into training set and testing set, and after the model has been trained with the training set, testing the model using the testing set to judge the validity of the model. The results in determine the utility of the model applied to previously unseen data.
dashboard[3]	a graphical representation of the analyses performed by the algorithms
data access[3]	the act or method of viewing or retrieving stored data.
data analyst[16]	a person responsible for the tasks of modelling, preparing and cleaning data for the purpose of deriving actionable information from it.

Term	Definition
data analytics[16]	the process of examining large data sets to uncover hidden patterns, unknown correlations, trends, customer preferences and other useful business insights. The end result might be a report, an indication of status or an action taken automatically based on the information received.
data architecture[3]	how enterprise data is structured, depending on the end result required. Data architecture has three stages or processes: conceptual representation of business entities. the logical representation of the relationships among those entities, and the physical construction of the system to support the functionality.
data center[3]	a physical facility that houses a large number of servers and data storage devices. Data centers might belong to a single organization or sell their services to many organizations.
data cleansing[16]	the process of reviewing and revising data to delete duplicate entries, correct misspelling and other errors, add missing data and provide consistency.
data collection[3]	any process that captures any type of data.
data consumer[17]	end users or other systems that use the results of the Big Data Application Provider.
data custodian[3]	a person responsible for the database structure and the technical environment, including the storage of data.
data governance[17]	the overall management of the availability, usability, integrity, and security of the data employed in an enterprise.
data integration	the process whereby multiple sources of data are combined into a format that can be used on a common platform for analysis and deployment.
data integrity[3]	a corporate metric to measure the accuracy, completeness, timeliness, and validity of the data.
data locality[17]	refers to the data processing occurring at the location of the data storage.
data migration[3]	the process of moving data between different storage types or formats, or between different computer systems.
data mining[2]	the process of finding certain patterns or information from data sets
data model, data modeling[3]	defines the structure of the data for the purpose of communicating between functional and technical team members to present data required for business processes. It can also refer to the communication of a plan to develop how data is stored and accessed among application development team members.
data ownership	considers data as intellectual property, it is the legal property of the owner and cannot be used without the owner's permission.
data point[3]	an individual item on a graph or a chart.
data provider[17]	source of new data or information feeds into data systems
data quality[3]	the measure of data to determine its reliability for decision making, planning, or operations.
data science[3]	a generally accepted term as a discipline that incorporates statistics, data visualization, computer programming, data mining, machine learning, and database engineering to solve complex problems.

Term	Definition
data scientist[17]	a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes in the analytics life cycle.
data security	the practice of protecting data from destruction, misuse, or unauthorized access.
data set[16]	a collection of data, very often in tabular form.
data structure[3]	a specific way of storing and organizing data.
data wrangling[5]	the process of acquiring data from multiple sources, cleaning the data (removing/replacing missing/redundant data), combining the data to acquire only required fields and entries, and preparing the data for easy access and analysis.
Database (DB) [3]	a digital collection of data and the structure around which the data is organized. The data is typically entered into and accessed via a database management system (DBMS).
database management system[3]	software that collects and provides access to data in a structured format.
decision tree[5][10]	a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. By taking a series of decisions along the branches, the user ultimately reaches the desired result at one of the leaves.
demographic data[3]	data relating to the characteristics of a human population.
dependent variable[5]	a variable that is under test and changes with respect to the change in an independent variable.
descriptive analytics[16]	understanding a data set through business intelligence and visualization techniques such as pie charts, bar charts, or line graphs, tables
digital transformation	the theory and practice of creating digital and data driven processes to replicate traditional workflows and processes in an enterprise, project or operating facility.
digital twin[6]	a virtual model of a process, product or service, which has the capability to learn and adapt to its environment; connecting the physical and digital worlds with digitized data. A digital twin continuously learns and updates itself from multiple sources to represent its near real-time status, working condition or position throughout its lifecycle.
dimensionality reduction[5]	the process of reducing the number of features or dimensions of a training set without losing critical information from the data in order to increase the model's performance.
document management[3]	the practice of tracking and storing electronic documents and scanned images of paper documents.
downstream	refining of crude oil and natural gas and marketing/delivery to consumers
event[16]	a set of outcomes of an experiment (a subset of the sample space) to which a probability is assigned.
event analytics[3]	presents the series of steps that led to an action.
factor analysis	a multivariate technique to describe, if possible, the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities, called *factors*

Term	Definition
graph database[3]	a NoSQL database which uses the graph data model comprised of vertices, which is an entity such as a person, place, object or relevant piece of data and edges, which represent the relationship between two nodes. It provides index-free adjacency, meaning that every element is directly linked to its neighbor element.
Hadoop[3]	open source software library project administered by the Apache Software Foundation. Apache defines Hadoop as "a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model."
HDFS (Hadoop Distributed File System) [3]	the storage layer of Hadoop, is a distributed, scalable, Java-based file system adept at storing large volumes of unstructured data
histogram[18]	a graphical representation of the distribution of a set of numeric data, usually a vertical bar graph
independent variable	also known as the attribute, this variable is data that is collected and used to calculate dependent variables (targets) through statistical means.
IoT (Internet of Things)[16]	the network of physical objects or "things" embedded with electronics, software, sensors and connectivity to enable it to achieve greater value and service by exchanging data with the manufacturer, operator and/or other connected devices via the Internet. Each "thing" is uniquely identifiable through its embedded computing system but is able to interoperate within the existing Internet infrastructure.
JavaScript[18]	a scripting language designed in the mid-1990s for embedding logic in web pages, but which later evolved into a more general-purpose development language.
join	an operation that connects two data frames by linking matching rows with the same entries into a large data frame
K-Means	an unsupervised method of describing clusters within a data set by determining mean values among groups or clusters of data through iterative calculations.
K-nearest neighbors (KNN)	a simplified supervised machine learning tool that assists in classification of data for regression analysis by grouping values in a data set within k number of groups and associating the data points in each group to its nearest neighbor in the data set.
kriging	the analysis and synthesis of spatial data collected at different spatial scales and over different spatial supports.
linear regression	a supervised machine learning tool that attempts to model the relationship in data set through a linear equation.
logarithm	a relationship between numbers using exponents as the tool to define the relationship. Typically used in data sets with multiple orders of magnitude.
logistic regression	a binary relationship in models that measures the probability of an event, for example Pass/Fail or Yes/No, typically used in fraud detection.
machine learning (ML)	a category of statistical tools that develop models to improve processes through predictive analytics.
mean[16]	the weighted average of data.

Term	Definition
mean squared error	in statistics, a measure of the mean or average of the squares of the errors – which are the square of the difference between estimated values and actual values. See also, average squared error
MetOcean data	refers to the combined wind, wave, current and atmospheric conditions at a specific location.
midstream	gathering and transportation of crude oil and gas via pipelines, truck tankers, or vessel tankers
mode[16]	the measurement that occurs most often in a data set.
model fitting[5]	process of checking the accuracy, performance or predictive power of a model, using statistical metrics on the data set.
model lifecycle[6]	a process that documents the initial model structure, track the quality, recalibrates and maintains the model, supports training activities, and evaluates the business benefits of the model over time
model selection[5]	the process of selecting a statistical model from a set of alternative models for a problem set that results in the right balance between approximation and estimation errors.
modeling[11]	the process of developing a model.
normal distribution	a probability distribution that is symmetric and contains a higher frequency of data about the mean.
NoSQL[16]	a broad class of database management systems identified by non-adherence to the widely used relational database management system model. NoSQL databases are not built primarily on tables and generally do not use SQL for data manipulation. Database management systems that are designed to handle large volumes of data and are often well-suited for big data systems because of their flexibility and distributed-first architecture needed for large unstructured databases.
object databases[3]	data stored in the form of objects, as used by object-oriented programming. They are different from relational or graph databases and most of them offer a query language that allows object to be found with a declarative programming approach.
optimization[15]	determines the values of a set of independent variables (e.g. temperatures and pressures) that minimize a real-valued objective function that is usually based on cost.
ordinal logistic regression[16]	assumes that the coefficients that describe the relationship between, say, the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category and all higher categories, etc.
outlier[5]	unusual observations in data that lie at an abnormal distance from where majority of the data points are located.
output[4]	the tangible result of performing a task
percentiles[5]	media measure used in statistics indicating the value below which a given percentage of a group of observations falls. Percentiles are used in robust statistical testing, for instance to compare two distributions.
phase[4]	a term for the high-level part of the CRISP-DM process model; consists of related tasks
Predictive Analysis/Analytics[16]	the use of statistical functions on one or more historical data sets to predict trends or future events.

Term	Definition
predictive modeling[3]	the process of developing a model that will most likely predict a trend or outcome.
prescriptive analytics[16]	builds on predictive analytics by including actions and make data-driven decisions by looking at the impacts of various actions.
probability[16]	the likelihood of a given event's occurrence, which is expressed as a number between 1 and 0.
probability distribution[16]	a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range. Probability distributions may be discrete or continuous.
public data[3]	data sets that are created with public funding or that are available to the public without requiring a license
Python	an object oriented programming language that is commonly used in data management and analytics
quartiles	values that divide data into quarters according to how they fall on the number line in ranking of 25% segments with respect to the segment's relationship to the median number of the data. The first quartile is 25% below the median. The third quartile is 25% above the median.
query	the process of interrogating a database to get information or relationships of information within the database.
query language	a language for specifying database queries.
R[16]	an open source programming language used for statistical computing and graphics. It is a GNU project which is similar to the S language. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques and is highly extensible. It is one of the most popular languages in data science.
r squared (R^2) [5]	a measure of how well a regression model fits the data by measuring the relationship between the linear model and the independent variables.
random forest[5]	a combination of many decision trees in a single model.
range[16]	difference between the largest and smallest measurement in a data set.
real-time data[16]	data that is stored as it is created, processed, stored, analyzed and visualized within milliseconds
reference data[3]	describes an object and its properties. The object may be physical or virtual.
regularization	a regression technique that keeps all independent variables in the model but forces the coefficients for less important independent variables to zero by adding a penalty to the residual sum of squares calculation.
risk analysis[3]	the application of statistical methods on one or more datasets to determine the likely risk of a project, action, or decision.
sample[16]	a data set which consists of only a portion of the members from some population. Sample statistics are used to draw inferences about the entire population from the measurements of a sample.
scripting[18]	the use of a computer language that, can be run directly with no need to first compile it to binary code
sentiment analysis[18]	a natural language processing algorithm that attaches mathematical scores to written words to calculate how people feel about certain topics by analyzing those words

Term	Definition
server[3]	a physical or virtual computer that serves requests for a software application and delivers those requests over a network.
standard deviation[16]	the positive square root of the variance
stepwise analysis	used in Regression models, steps through the model, either by moving forward and incrementally adding coefficients, or backward by including all coefficients and incrementally removing coefficients. The decision by the analysis to add or remove a coefficient is determined by its goodness of fit.
storage[3]	any means of storing data persistently.
structured data[17]	data that has a predefined data model or is organized in a predefined way.
sum of squares[16]	a process that helps express the total variation that can be attributed to various factors.
systems engineering[14]	a transdisciplinary and integrative approach to enable the successful realization, use, and retirement of engineered systems, using systems principles and concepts, and scientific, technological, and management methods.
t-test	a test statistic that is used to determine whether there is a significant difference between the mean of two groups.
table	data organized in rows and columns
Tableau[18]	a commercial data visualization package often used in data science projects.
target variable	another term for independent variable
text analytics[3]	the application of statistical, linguistic, and machine learning techniques (natural language processing) on text-based sources to derive meaning or insight.
time series[5]	one-dimensional data indexed by time
time series analysis[18]	analyzing well-defined data obtained through repeated measurements over time. The data must be well defined and measured at successive points in a time series.
training (a model)	the process to evaluate the effectiveness of a model by dividing the data set into a Training Set and a Validation Set (typically a 70/30 ratio). If after running a successful model with the Training set, the Validation set returns the same results with the successful model, the model is said to be "trained".
training data set[5]	data that is used for training a machine learning model. Typically, data is separated (randomly) into a 70/30 split of training/validation data.
transactional Data[16]	data that relates to the conducting of business, such as accounts payable and receivable data or product shipments data. Other examples of transactional data are Purchase orders, maintenance notifications, and work orders.
unstructured data[17]	data that does not have a predefined data model or is not organized in a predefined way.
value[17]	refers to the inherent wealth, economic and social, embedded in any dataset.
verification[13]	the process of confirming that a model is correctly implemented with respect to the conceptual model (it matches specifications and assumptions deemed acceptable for the given purpose of

Term	Definition
	application). During verification, the model is tested to find and fix errors in the implementation of the model.
visualization[16]	a visual abstraction of data designed for the purpose of deriving meaning or communicating information more effectively.
weather data[3]	any facts or numbers about the state of the atmosphere including temperature, humidity, wind speed, rain, snow or pressure.

1.3 References

[a] Tellenbach B., Rennhard M., Schweizer R. (2019) Security of Data Science and Data Science for Security. In: Braschler M., Stadelmann T., Stockinger K. (eds) Applied Data Science. Publisher: Springer, Cham. https://doi.org/10.1007/978-3-030-11821-1_15

[b] Wheatley, Malcom (May 29, 2013) "Underground Analytics: The Value of When an Oil Pump Fails", www.duoline.blogspot.com/2013/06/underground-analytics-value-in.html?m=1

[c] https://www.cisecurity.org/controls/

[d] https://library.seg.org/seg-technical-standards

[e] https://www.osha.gov/leadingindicators/docs/OSHA_Leading_Indicators.pdf and https://www.osha.gov/shpguidelines/program-evaluation.html#:~:text=Lagging%20indicators%20generally%20track%20worker,or%20illnesses%20before%20they%20occur.

[f] https://www.mckinsey.com/business-functions/operations/our-insights/big-data-and-the-supply-chain-the-big-supply-chain-analytics-landscape-part-1#

[g] https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

[h] Jones, Edward R., (2018), "The Art and Science of Data Analytics", (Unpublished)

[i] https://www.linkedin.com/pulse/10-best-data-analytics-bi-platforms-tools-2020-bernard-marr/.

[j] Certified Analytics Professional (CAP®) Examination Study Guide, Publisher: Institute for Operations and Management Sciences (INFORMS), 5521 Research Park Drive, Suite 200, Cantonsville, MD 21228

[k] Meridje, Yacine, (2020), "Data Driven Permian Basin Production Forecasting using Machine Learning". (Unpublished Capstone Project, Texas A&M University Masters Data Analytics Program)

2 DATA STRUCTURE AND MANAGEMENT

2.1 Introduction

This chapter provides an overview of structured and unstructured data and the respective databases used in the oil and gas industry. Understanding the types and their usage is a critical and key component to properly executed data analytics projects. Data should ideally be accessed, securely stored, verified and validated.

2.2 Structured Data

2.2.1 Types and Usage

Structured data is the most recognizable form of data for technical and business professionals. It is made up of alpha-numeric characters and usually accessed and stored in tables and matrices of rows and columns.

Structured data may be used for descriptive, predictive and prescriptive analytical techniques as described in Chapter 4.

2.2.2 Databases

(a) Relational
Relational Databases, managed by software programs called Relational Database Management Systems (RDBMS), contain digital data in rows and columns, or tables. They are characterized by Index and Key variables used to join one or more similar databases. Each row is called a record and each column is a variable or attribute. All data is stored among tables based on relations. Examples of relational databases include address books, accounts payable, and oil well production reports.

Examples of RDBMS programs are Oracle Database, MySQL, Microsoft SQL, Excel Spreadsheets, Microsoft Access, and IBM DB2.

(b) Object-Oriented
Object-Oriented Databases are managed by software programs called Object Oriented Database Management Systems (ODBMS). They are organized by objects that have Structure, Classes and Identities.

The Structure of an ODBMS is defined by the properties of a Message Interface, Method Execution, and Variables. Objects are further defined by Classes. A Class is a real-world entity. The variables can be shared from one Object to another within the structure. Each action in an ODBMS is related to its original object rather than organized as tables. Because of this object relationship rather than mapping to columns and rows, ODBMS can handle more complex data with many varied relationships.

Examples of ODBMS include many open source programs and can be accessed by object-oriented programming languages like JAVA, C++, and Python.

2.2.3 Examples

(a) Table 2-1 lists examples of structured data types and potential analytical techniques.

Table 2-1: Data Description Table

Data Type	Examples	Measurement Scale	Summary Statistics	Predictive/Modeling Tools
Float	0.01, 1.2, 3.95, …n.n	Continuous, Intervals	Mean, standard deviation, median, mode, correlation	Regression, Classification, Linear, Logarithmic
Binary	Yes/No, 1,0, Male/Female	Nominal	Mode, Chi-squared	Logistic
Ordinal	0,1, 2,..n	Ordinal		Regression, Logistic
Count	1, 2,..n	Ratio	Mean, standard deviation, median, mode, correlation, geometric mean, harmonic mean	Poisson
Categorical	Title, Labels	Nominal	Mode, Chi-squared	Regression, Classification, Logistic

(b) Figure 2-1 is an example of a Relational Database. The two tables are related by the Account ID column.

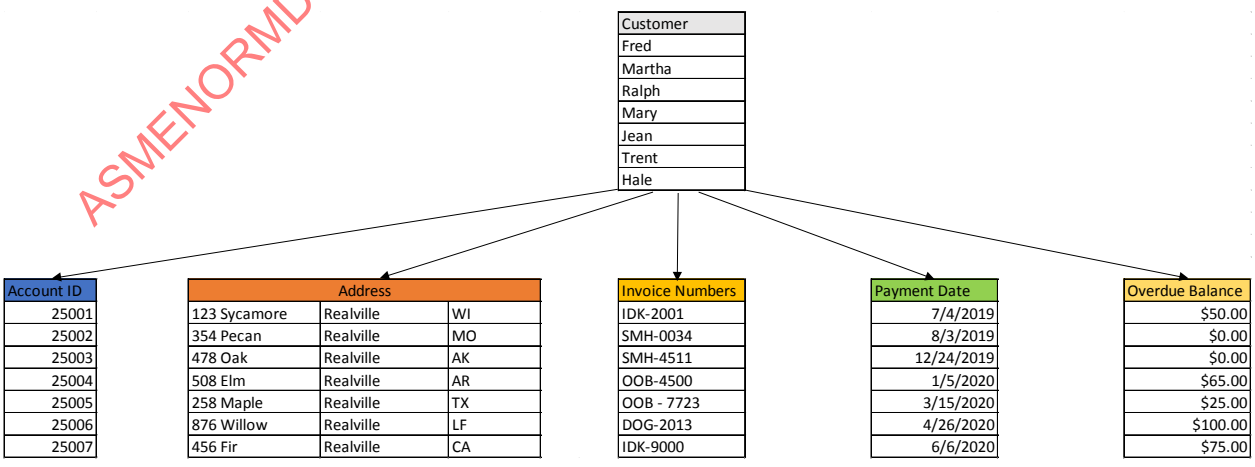
Figure 2-1: Relational Database Example

Account ID	Name	Address	City	State
25001	Fred	123 Sycamore	Realville	WI
25002	**Martha**	**354 Pecan**	**Realville**	**MO**
25003	**Ralph**	**478 Oak**	**Realville**	**AK**
25004	Mary	508 Elm	Realville	AR
25005	Jean	258 Maple	Realville	TX
25006	Trent	876 Willow	Realville	LF
25007	**Hale**	**456 Fir**	**Realville**	**CA**

Account ID	Invoice Number	Payment Date	Overdue Balance
25003	**IDK-2001**	**7/4/2019**	**$50.00**
25007	**SMH-0034**	**8/3/2019**	**$0.00**
25002	**SMH-4511**	**12/24/2019**	**$0.00**
35016	OOB-4500	1/5/2020	$65.00
55001	OOB - 7723	3/15/2020	$25.00
89120	DOG-2013	4/26/2020	$100.00
42015	IDK-9000	6/6/2020	$75.00

(c) Figure 2-2 is an example of an Object-Oriented Database, using the same data as example (b). Each of the headers describe the Class of the Object, which are then related through object-oriented programming techniques.

Figure 2-2: Object-Oriented Database Example



2.3 Unstructured Data

2.3.1 Types and Usage

Unstructured data is data that cannot be defined into objects, tables or easily recognized structures. Text and photographs comprise most unstructured data, but it might also contain dates, numbers, facts, and references that do not organically organize themselves. When individuals discuss Big Data, the unstructured data is the typical reference point for that discussion. In addition to the lack of structure, the amount of data to be stored is enormous and possibly makes up 80% of most organizations' data.

Data mining techniques for unstructured data require tools that can process natural language and/or recognize patterns, and large quantities of both. In this sense, the data does contain a form of structure, but it is not easily recognizable.

2.3.2 Respective Databases

(a) Data Lakes
Also known as data warehouses or data swamps, data lakes are large repositories of data that are unstructured, or sometimes referred to as "raw." Due to the storage requirements, most data lakes are stored in the cloud unless the organization owning the data has considerable onsite storage.

Examples of off-site data lakes are Google Cloud, Amazon S3 or Apache Hadoop. A Hadoop is an open-source software for reliable, scalable and distributed computing. It provides massive storage capabilities and impressive computing power. It is not a programming language but rather an ecosystem that facilitates the moving and organization of Big Data. Hadoop-powered storage provides the capability for storing information derived by Internet of Things (IoT).

(b) NoSQL
NoSQL databases, originally referred as "non-SQL" or "non-relational" database, store data in non-tabular form. Included in this set of databases are the previously discussed ODBMS, Key-value stores, Document stores, and Graph databases. Each of these types of databases uses a unique method to map data, documents, dictionaries or relationships through tags.

Examples of NoSQL include Apache Ignite, Couchbase, Oracle NoSQL, Amazon DynamoDB, and many others. These databases arrange data based on correlations of values rather than tables.

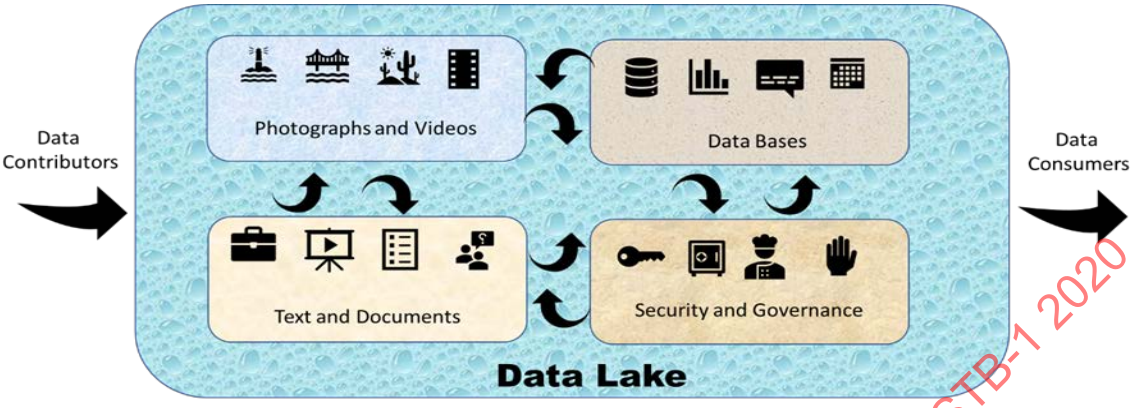
(c) Graph Databases
Graph databases are a unique subset of NoSQL databases that are gaining popularity for complex data mapping. They map data elements on a chart or graph and have finite numbers of relations. Graph databases have nodes with data and edges that describe relationships. Each node can have many edges and therefore described many relationships. These databases are suited for data sets with a wide variety of both structured and unstructured data.

Examples of Graph databases include AllegroGraph, Neoj4, and Infinite Graph. These databases use languages to manage the data such as SPARQL, Java, and CYPHER.

2.3.3 Examples

(a) Data Lakes
Figure 2-3 is an example of Data Lake. Note the variety and number of data sources to be managed and distributed.

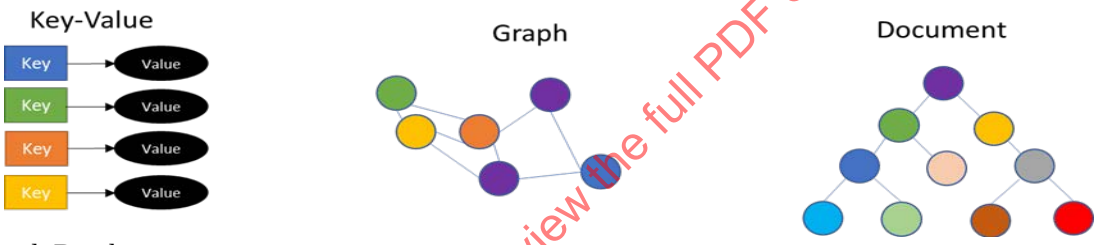
Figure 2-3: Data Lake Example



(b) NoSQL Types

The common types of NoSQL Data Types are shown in Figure 2-4.

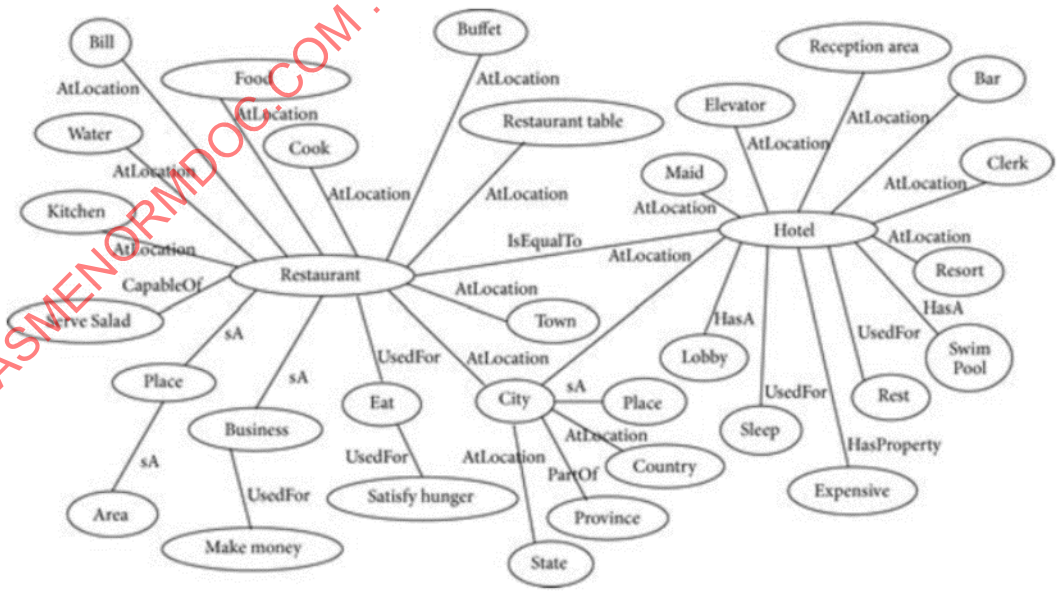
Figure 2-4: NoSQL Data Types



(c) Graph Databases

Figure 2-5 shows details of the Graph Databases with emphasis on the Nodes and Edges.

Figure 2-5: Graph Database



Source: Agarwal, Basant & Mittal, Namita & Bansal, Pooja & Garg, Sonal. (2015). Sentiment Analysis Using Common-Sense and Context Information. Computational intelligence and neuroscience. 2015. 715730. 10.1155/2015/715730

2.4 Security and Governance of Data

2.4.1 Responsibility of the Enterprise

The enterprise team of engineers, planners, project managers and data professionals are responsible for solving business challenges using both internal and external sources of data. Many of these internal sources are proprietary to the business and are key to the competitive advantage of the business. Some external sources are subscription-based and should ideally be treated as confidential. The Data Scientist has a core responsibility to safeguard the confidentially-provided information. Guarding this data requires both active protection (e.g. not sharing with other individuals not expressly governed by the same confidentiality) and passive protection (e.g. following corporate protocols, using protection software, backing up data sets).

2.4.2 Key Concepts of Information Security

According to the Federal Information Security Management Act of 2002 (2012), the term "information security" means protecting information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability (CIA):

(a) **Confidentiality** requires the implementation of authorized restrictions on access and disclosure, which includes measures for protecting personal privacy and proprietary information.
(b) **Integrity** means guarding against improper modification or destruction and includes information, non-repudiation and authenticity.
(c) **Availability** means ensuring timely and reliable access to and use of information.

2.4.3 Data Protection

Big Data requires the processing and storage of large amounts of data. To process that data efficiently, it should ideally be available in unencrypted form. The obvious drawback is that hackers can steal or corrupt the data, or internal users can accidentally modify, destroy or corrupt data.

An example:

> A central Contracts and Procurement organization identified that the company could procure less expensive pump seals than was being currently procured. The CP organization decided to implement the change for seals that had different materials of construction than what had been identified in the bills of materials (BOM) and material masters (MM) in a refinery's Enterprise Resource Planning (ERP) system. The CP organization made changes to the refinery's data systems so the lower cost seals would be purchased in the future. Unfortunately, the CP personnel neither used an approved MOC process to make changes nor engaged the discipline engineers at the refinery responsible for asset integrity and safety of the refinery. Several months later, site personnel discovered the change in the BOMs and Material Masters. The new seal materials identified by the CP organization were incompatible with the process fluids, which would have eventually resulted in a catastrophic failure of the seals and possibly fire and personnel injuries.
>
> All of the changes that had been implemented had to be returned to the original configuration, which doubled the cost of the original implementation. Furthermore, the incorrect seals that had been installed since the data change had been made had to be removed from service and replaced with seals that had the appropriate material compatibility.

To safeguard the data, it should ideally be stored in encrypted form, then unencrypted for processing tasks. The hacker will have to steal the keys to encryption or "find" the data during processing tasks, which is much more difficult.

Another approach is to be able to process the data within the encrypted space. This is a relatively new approach that is gaining acceptance and usage.

For more information on this topic, please see https://www.cisecurity.org/controls/ [c].

2.4.4 Developing Software that is Secure

Security can be designed into software during its development. The traditional software development lifecycle (SDLC) can be augmented by including security in that development. This is known as secure SDLC or SSDLC. This ensures that the code is written with security and defense architecture. SSDLC processes require domain experts in Information Technology to work alongside the Data Scientist to understand the threats and build them into software that is eventually deployed for use in the enterprise.

2.4.5 Facility Management Systems

The responsibility of enterprises to safeguard data requires overall facility management systems and periodic audits to confirm that the safeguards are in place and are working as intended. Although this is not the direct responsibility of the Data Scientist, he/she is integral to the process. The following best practices are recommended:

(a) Analysis of the system with respect to Privacy and Information Security
(b) Evaluation of legal and regulatory requirements and compliance
(c) Implementation of data breach notification laws
(d) Solutions to maximize data security are in place
(e) Examination of data process and information flows
(f) Confirmation that adequate enforcement procedures are in place and in use

3 BIG DATA IN THE OIL AND GAS INDUSTRY

3.1 Introduction

3.1.1 Overview

Big Data is the cornerstone of what is now called the Digital Facility in the oil and gas industry. This Big Data represents the physical facility and its historical data, current operations and predicted outcomes. Each component of this digital facility reflects certain unique sets of data related to its manufacture, installation, operation, inspection, surveillance, maintenance and repair.

Various types of data are unique to upstream, midstream and downstream businesses and operations – both onshore and offshore. The sources of these data sets can be proprietary, public domain, or a combination of both. The next sections provide details of the most likely, but not only, sources of data for the oil and gas professional.

Combining data sets from varied sources will require data sorting, cleansing and alignment prior to any meaningful analysis. This Chapter will focus on the data that is collected or contemplated. The tools to clean and normalize data prior to analysis are discussed in Chapter 4.

3.1.2 Oil and Gas Facility Lifecycle Digital Requirements

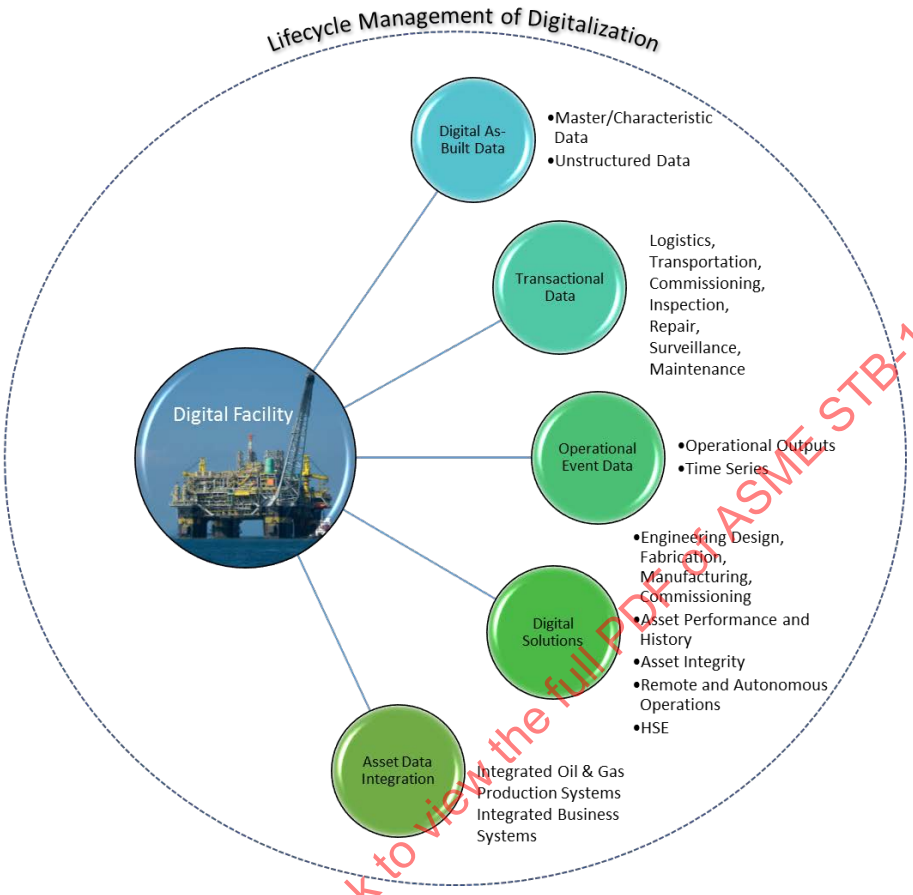
The Digital Facility is an operated facility that has a digital reflection for some of the components and activities associated with its operation. The intended outcome is to safely and cost-effectively improve and extend the Lifecycle of the facility through the harnessing of Big Data.

The Digital Facility reflects the Physical Facility. For designated components in the Physical Facility, there is a digital "component" in the Digital Facility. This philosophy is also known as a Digital Twin. The Digital Twin has two essential functions: providing operational information for analysis, and simulating operations to improve decision-making.

Delivery of the Digital Facility requires a digital understanding of the component Master information and operations. The delivery depicted in Figure 3-1 depends on:

(a) Digital As-Built and Master Data
This is physical information about each component: Master Data (characteristics such as bill of materials, dimensions, weight, location in the facility, safety characteristics) and Unstructured Data. This data includes engineering design and as-built data, 3D models, and documentation.

Figure 3-1: Digital Facility Components



(b) Transactional Data
During the life of a piece of equipment, it is transported, installed, commissioned, inspected, surveilled, maintained and repaired. This data becomes part of the description of the equipment but is different from the Master Data in the as-built state.

(c) Operational Event Data
Each component in the Digital Facility creates operational data in the form of historical event data and real-time data as Time Series Data. Within this system are the operating parameters for each piece of equipment and forecast trends based on actual operating outputs. How that data is collected and stored is defined in enterprise level data standards. These data standards will dictate the formats of all outputs for use in predictive analytic platforms.

(d) Digital Solutions
Digital solutions are developed based on the Master data for the components and the overall system, and the operating data output from activities during operations. They include both hardware and software systems that collect, store, secure, and analyze the data. Predictive analytic algorithms are produced by the digital solutions for the Digital Facility to address the following business objectives:

(1) Engineering Design, Fabrication, Manufacturing, Commissioning
(2) Asset Performance and Efficiency
(3) Asset Integrity – Maintenance, Surveillance and Repair
(4) Remote and Autonomous Operations - Inspections
(5) Health, Safety and Environment

(e) Asset Data Integration
Physical Facilities are integrated across engineering and operational disciplines. Digital Facilities are integrated across all data systems and warehousing of data (e.g. Data Lakes and ERP Systems). The Integration should ideally consider:

(1) Integrated Oil and Gas Production Systems, Pipelines, Refineries and Petrochemical Plans
(2) Integrated Business Systems
- Asset Integrity
- Procurement/ERP
- Warehousing
- Event Management

(f) Lifecycle Management of Digitalization
Digital tools and processes should ideally be maintained and upgraded throughout the lifecycle of the Digital Facility. This mirrors the lifecycle management of the Physical Facility. This management should ideally plan for the obsolescence of hardware and software tools and create a plan to manage that obsolescence.

3.1.3 Designing the Digital Facility

Digital strategies are intentional. Philosophies for a digital strategy are specified in the Conceptual phases (Assess and Select) of a project. This philosophy defines the specifications as well as to the implementation of digital tools during the Front End Engineering and Design (FEED or Define) that will carry through to the Operate Phase of the project. These two phases are required so that the Digital Facility is appropriately constructed in the Engineering, Procurement, Construction per EPCI, and Installation (EPCI) phase of the project. Although some facilities can be retrofitted with digitalization, the efficacy of the system is diminished if it is not designed from the conceptual stages of the facility.

3.1.4 Understanding the Data in Oil and Gas Activities

The remainder of this Chapter describes discrete Oil and Gas segments. Each segment contains activities and the data produced from those activities. For the purposes of this document, Digital As-built and Engineering Information Data are described as *Master Data*. Dynamic Operational Data is described as *Time Series Data*. Reports and documents are described as *Unstructured Data*. Each segment will have a brief statement on *Models*.

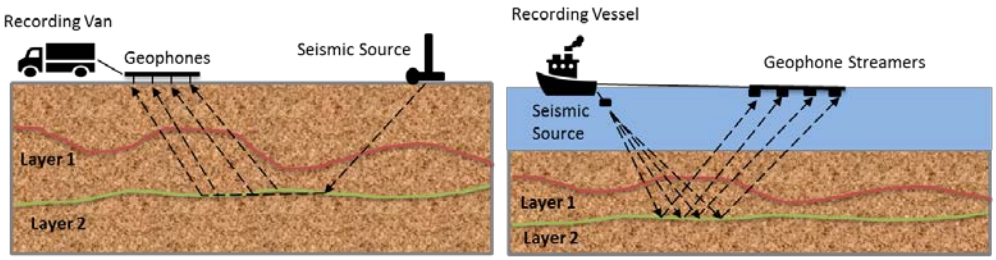
3.2 Hydrocarbon Reservoirs, Drilling, Production, Transportation and Refining

3.2.1 Activities that Produce Data

The exploration and production phases of the hydrocarbon business are sequenced from Seismic Activity – Drilling – Logging – Well Completion – Production – Transportation – Refining. The Production phase has sub phases of Development – Construction – Installation – Operation. The Production phase data relates to the Digital Facility described in Chapter 3.1.2. The phases preceding Production supply significant amounts of data that are useful for data analytics for forecasting and improving these activities.

(a) Seismic Imaging
Seismic imaging is a data-intensive activity designed to produce 3D and 4D models of the earth's subsurface to detect the existence of potential hydrocarbons, geothermal sources, and carbon sequestration destinations. A seismic source that vibrates the earth sends "shock" waves that are refracted through different rock densities and "heard" by geophones that record the waves and translate those waves into a "picture" of the subsurface. This activity is executed both on land and offshore (See Figure 3-2).

Figure 3-2: Onshore and Offshore Seismic Imaging Activities



Master Data include the measuring equipment details such as physical dimensions, wave energy propagation, power, transportation and logistics to get the equipment into the field.

Time Series Data include Magnetotellurics, Velocity, and Anisotropy that describe the recorded data in the recording truck or ship.

Unstructured Data include written reports and spatial data on location(s).

Models created include 3D, 4D, and Salt Interpretation. These are customized and unique to seismic data processing. Data Analytics and Visualization are fundamental to geological success in discovery of hydrocarbons.

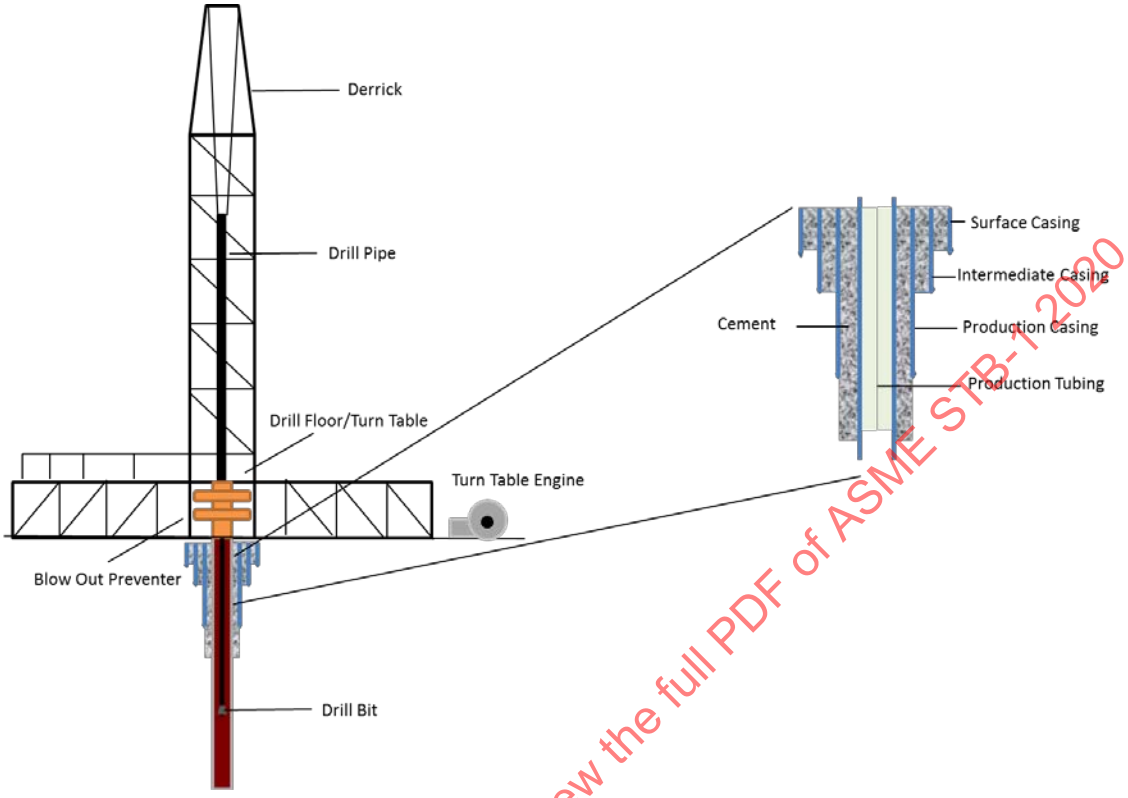
For more complete information on activities and data related to Seismic Imaging, please see https://library.seg.org/seg-technical-standards [d].

(b) Drilling and Well Construction

Drilling operations require the penetration of the subsurface to extract mineral resources using a specially designed drill bit and threaded sections of steel drill pipe. The drill pipe and bit are rotated during drilling operations to cut the subsurface rock. Cuttings from the drilling operation are returned to the surface via a drilling fluid called mud and then filtered and recirculated into the well. The drill pipe sections (40 ft. in length each) can extend up to 40,000 ft. During the activities associated with drilling, the components of the drilling system are in contact with various rock densities, porosity, pressures and temperatures. The goal of the driller is to "hit" a target depth and location as specified by the geologist and the drilling engineer. Drilling operations may use Measurement While Drilling (MWD) technology to precisely place the drill bit. A key component of the drilling operation is to safely manage pressure control through a Blowout Preventer (BOP) and the engineered weight and pressure of the drilling mud. The drilling rig or derrick is designed to hold the weight of the drill bit, drill pipe and drilling mud. A central rotary turntable on the drill floor of the drilling rig, a top drive or the drilling mud rotates the drill bit to execute the drilling and cutting of the rock. Drilling offshore requires additional equipment in the form of risers between the seabed and drilling rig to contain the drilling mud and cuttings. In shallow water, the drilling rig is on a fixed platform. In deep water, the drilling rig is floating and either moored to the seafloor or has a set of thrusters that position the drilling rig via GPS location settings.

Well Construction is the activity associated with constructing a structural support for the hydrocarbon well once drilling operations are complete. Many of the well construction activities occur during drilling to ensure the well is supported and structurally stable during drilling and other operations. The structural components are a series of large drill pipes, called casing, and cement. The final structure resembles an upside-down wedding cake. The drilling derrick and well construction casing are shown in Figure 3-3.

Figure 3-3: Drilling Derrick and Casing Components



Master Data include the weight and dimensions of each of the components, material specifications for each component, physical properties of the drilling mud, engineering design of the drilling rig and its systems, engineering design of the well casing, the geophysical properties of the rock structures, and transportation of the equipment to the location.

Time Series Data include the location information gathered as the drill bit cuts through the rock formations, changing properties of the drilling mud, reservoir navigation, pressures, temperatures, and activation of the valves BOP (if required to manage wellbore over pressure).

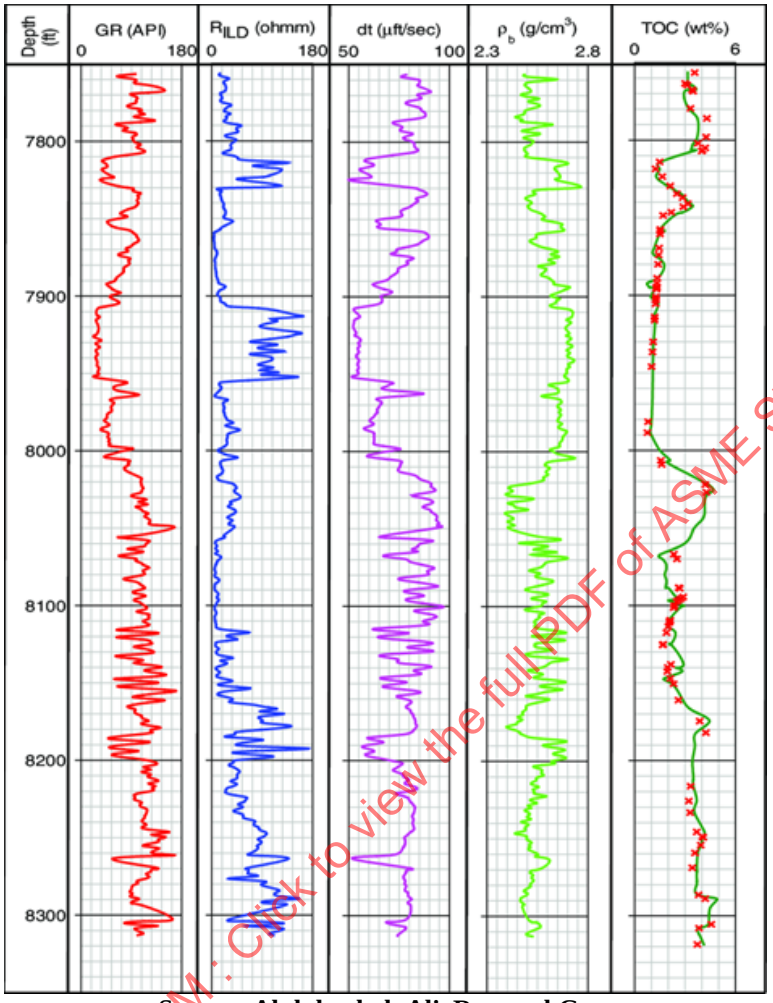
Unstructured Data include the various written drilling and general reports during daily operations and imaging/charts created during the MWD activities.

Models of the substructure provided by seismic activity are enhanced by information received during MWD activities. Data Analytics are performed on any information provided by the Time Series Data.

(c) Well Logging

Well logging is a procedure to define the reservoir penetrated by the wellbore using a wireline that measures the electrical resistivity of the rock formations along the entire length of the wellbore. Logging While Drilling (LWD) is a new technology that reduces the time required to create a log by eliminating the separate logging activity. Logging is performed from the drilling rig. The log (Figure 3-4) measures hydrocarbon saturation and formation pressure. This information leads to additional drilling and production decisions.

Figure 3-4: Example Well Log



Source: Abdulwahab Ali, ResearchGate

Master Data include the physical properties of the logging equipment and the wellbore diameter, length, and wall thickness(es).

Time Series Data include changing formation information as the logging operation proceeds along with pressures and temperatures.

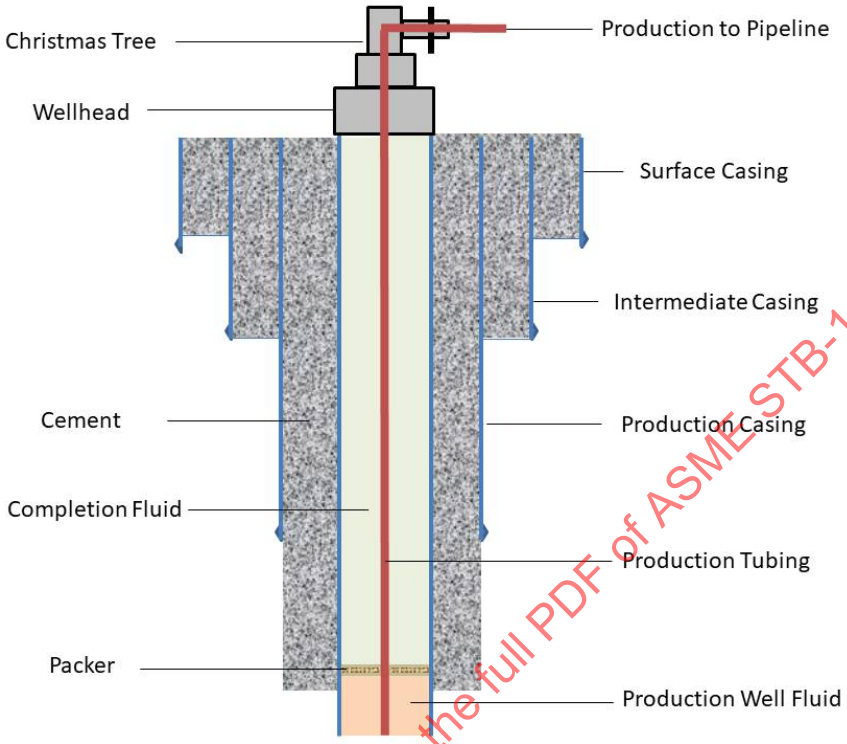
Unstructured Data include the log charts and written reports.

Models of the substructure provided by seismic activity are enhanced by information received during logging activities. Data Analytics are performed on any information provided by the Time Series Data.

(d) Well Completion

Well Completion is a broad category of activities required to place the well into production. To introduce hydrocarbons into the wellbore to flow to the surface, the steel pipe is perforated at locations indicated by the well log information. A series of valves, sleeves and packing are inserted and locked into place along the wellbore to control and direct the flow of hydrocarbons to the surface (Figure 3-5). Some wellbores will include electrical submersible pumps (ESPs) to assist the flow with artificial lift. A pressure control system using a well head and a production Christmas tree are located on the surface. The drilling rig performs the well completion.

Figure 3-5: Completed Well Example



Master Data include the physical properties of the valves, actuators, sleeves, packers, cement, ESPs, installation equipment, and downhole tools. It will also include electrical and hydraulic control system components.

Time Series Data include the performance data of the valves and sleeves as they cycle open and closed as well as indicators of the valve and sleeve positions. Certain equipment may need monitoring of temperature and pressure. The surface pressure control equipment will monitor fluid flow rates, leaks, temperature, pressure, valve position indicators, chemical injection rates, and hydraulic and electrical control system performance.

Unstructured Data include written reports and engineering design documentation.

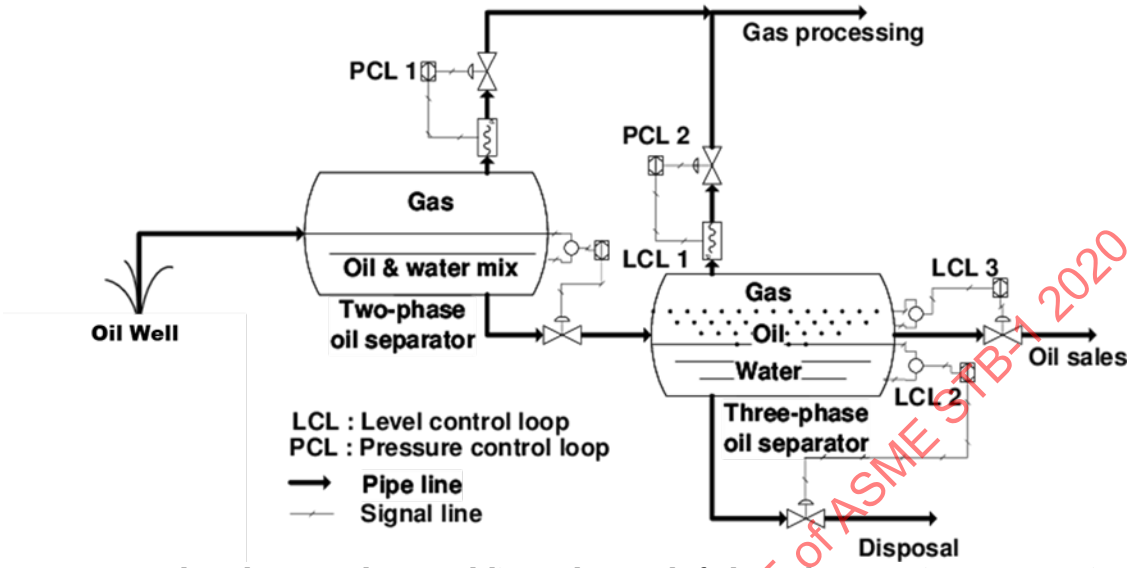
Models include predictive analytics for asset condition monitoring and maintenance/repair/replacement of components.

(e) Production, Transportation and Refining

The production phase of oil and gas operations encompasses the flow of hydrocarbons from the completed well to a production facility for initial separation and treatment prior to entering a pipeline gathering system. Pipeline systems gather oil or gas from many production facilities for eventual flow into refineries. In each stage of this journey, oil and gas are maintained at pressures between 2,000 to 15,000 psi.

Oil and gas separation is a mechanical process that separates oil, water and gas into individual components as shown in Figure 3-6. Oil is then treated with chemicals and then pumped into an oil pipeline. In remote locations, it is stored in a floating vessel and then transferred to an oil cargo vessel. Water is treated to remove any residual oil or chemicals and then either safely discharged or reinjected into a reservoir. Gas is treated to remove residual oil or water and then either flared or compressed into a gas pipeline.

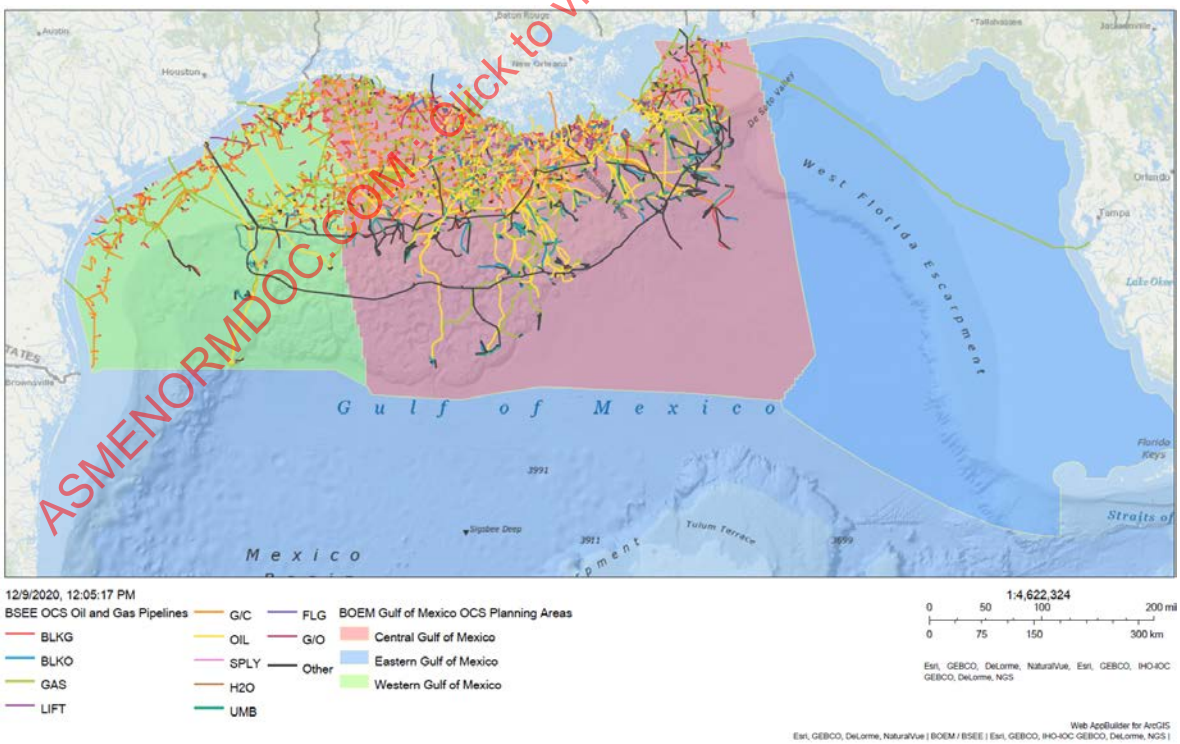
Figure 3-6: Oil and Gas Production Facility



Source: A. F. Sayda and J. H. Taylor, "Modeling and Control of Three-Phase Gravity Separators in Oil Production Facilities", Proc. American Control Conference, New York, 11-13 July 2007

Oil and gas are transported through pipelines at elevated pressures. An example of a pipeline network is shown in Figure 3-7.

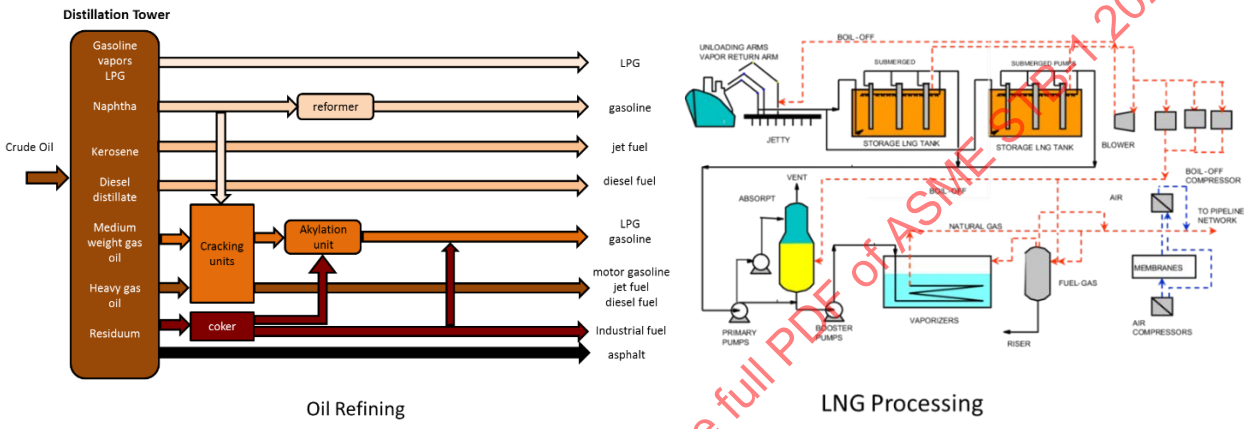
Figure 3-7: Gulf of Mexico Pipeline System



Source: Bureau of Ocean Energy Management, www.data.boem.gov/Main/Mapping.aspx

Refineries and processing plants convert oil and gas to end products through combinations of heat and chemicals. These processes are typically proprietary to the owner of the Refinery. The output of the Refinery are products for sale to consumers (gasoline, butane, propane, natural gas, diesel, jet, kerosene, coke and marine bunker fuel) or to Petrochemical plants that further process refined products into plastics and other industrial chemicals. Gas can also be liquefied through very high pressure and very low temperatures into Liquefied Natural Gas (LNG) or regasified through lower pressures and higher temperatures for transport on specialized LNG carriers where pipelines are not practical (overseas transport). Figure 3-8 shows typical processes each for oil and gas.

Figure 3-8: Oil Refining and LNG Processing



Reprinted from Journal of Loss Prevention in the Process Industries, Vol 28, O.N. Aneziris, I.A. Papazoglou, M. Konstantinidou, Z. Nivolianitou, Integrated risk assessment for LNG terminals, Page No. 13, 2014, with permission from Elsevier

Master Data, **Time Series Data**, **Unstructured Data** and **Models** are discussed in sections 3.3, 3.4 and 3.5.

3.2.2 Digital Facility Descriptions

Each of the activities described in Section 3.2.1 have opportunities for data output and collection. The discrete Digital Facilities that reflect the Physical Facility are:

(a) Drilling Rig and Associated Operations
  (1) Land Operations – Drilling Rig System
  (2) Offshore Operations – Floating/Jack-up Drilling Vessel

(b) Completed Well and Associated Pressure Control and Flowlines to a Production Facility
  (1) Land – Well Pads with flowlines to nearby Production Facilities
  (2) Offshore – Subsea Production Systems (Wet Trees) tied back to Subsea Facilities/Floating Facilities/Fixed Platforms/Land Facilities or Direct Vertical Access (Dry Trees) wells to Floating Facilities/Fixed Platforms

(c) Production Facility
  (1) Land – small skid-mounted facilities to large regional plants
  (2) Offshore – Floating Systems or Fixed Platforms

(d) Pipelines/Gathering System – Onshore and Offshore

(e) Refinery/LNG Processing Plant

3.3 Mechanical Equipment and Instrumentation

Although the facilities described in Section 3.2 have uniquely distinct functions, the mechanical equipment and instrumentation have properties and functions that are roughly similar from facility to facility. The following sections provide more detail on the main and common component systems in hydrocarbon processing. These are the physical facilities that define the digital facility. For both the physical facility and the digital facility, the mechanical equipment and associated instrumentation are assessed as a system. These systems are very large and complex.

Managing the system via physical means is time consuming. Creating the digital facility is time consuming; however, the operations are much easier once the digital system is online and operating. This is the intersection of Information Technology (IT) and Operational Technology (OT).

All of the equipment and operations discussed in subsequent sections will have accompanying unstructured data in the form of manuals, inspection reports, operational reports, and photographic information. Models produced will include predictive analytics, time series analytics, natural language processing and prescriptive analytics.

3.3.1 Pressure Control Equipment

Pressure Control Equipment is characterized by thick chambers and valves that control the flow of a fluid by maintaining the internal pressures of the fluid flow and/or resisting the external ambient pressure.

(a) Valves
Valves are opened and closed through use of actuators. These actuators are manually or remotely operated. Remote control actuators are controlled with either remotely operated vehicles, hydraulics, air or electric. Hydraulic actuators are controlled via direct hydraulics or electro-hydraulics. Data of interest in collection include valve position, cycles of opening and closing, leak detection, pressures and temperature. Christmas trees are an example of sets of valves operating as a unit.

(b) Pressure Vessels
Pressure Vessels include a large number of equipment types in production facilities and refineries. Their functions range from acting as holding tanks to sophisticated chemical processes and many subdivisions. Common types are separators, reactor columns, boilers, and heat exchangers. Data of interest are power usage, pressures, temperatures, open and closing of inlet and outlet valves, level, flow rates, and leak detection.

3.3.2 Rotating Equipment

Rotating equipment increases the pressure of a fluid to maintain flow through a pipe or flowline system. Pumps act on liquids (occasionally liquids and solids) and compressors act on gases. Data of interest are pressures, power usage, temperatures, vibration, leak detection, flow rates and fluid composition. Due to the complexity of rotating equipment, it requires frequent maintenance and has attracted a great deal of interest for predictive analytical tools for predictive maintenance, known as condition monitoring.

3.3.3 Electrical and Instrumentation

Electrical components are required to power facilities and operate the mechanical equipment functions. Instrumentation components measure and control the mechanical equipment and system functions. Because instrumentation requires power to operate, it is usually grouped and linked to electrical components.

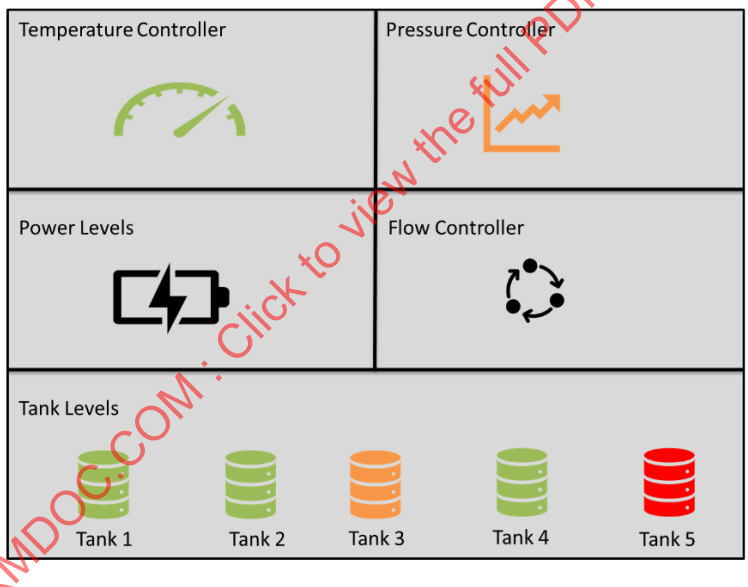
(a) Electrical
Electrical components include power generators, switchgear, transformers, cables, and many subcomponents required to operate the complete system and interface with the facility equipment and systems. Data of interest are voltage, amps, current, temperatures, signals, and radio signals.

(b) Instrumentation
Instrumentation includes the sensors, gauges, transmitters, receivers, and switches that measure the Data of interest for OT and for data collection. Data of interest are positions of indicators, levels of fluids, fire detection, chemical composition, fluid composition, gas detection, and frequencies.

3.3.4 Process Control

Process control is the operational activity of managing and controlling mechanical and chemical processes in a production facility or a refinery. This is a system function that gathers information from all equipment in the facility and relays that back to a central control facility in a network. The main management tool is a Programmable Logic Controller (PLC) which operates on a local area network at the facility and through cloud-based systems at a remote headquarters. The PLC is connected to the system via a distributed control system (DCS) that has local control of subsystems.

Figure 3-9: Process Control Graphical User Interface



Through graphical user interfaces (Figure 3-9), humans can observe and manage all aspects of the process system. Data of interest are many of the Data discussed in earlier sections: fluid composition, fluid flow, pressures, temperatures, high liquid level alarms, valve positions, power usage, fuel usage, etc.

3.3.5 Process Equipment

Process equipment is designed to alter hydrocarbons through mechanical means (pumping, separating) or chemical means (heat, chemical injections, catalysts, pressure). Many of the pressure vessels discussed in 3.3.1(b) are part of the process equipment and its systems.

Data of interest are measured in Process Control and include temperatures, pressures, fluid and gas composition, power usage, feedstock usage, feedstock properties, produced fluid and gas properties, system status, and liquid level alarms.

3.3.6 Civil/Structural

Civil and structural components include the foundations and structures that support all oil and gas equipment. Foundations are fabricated with concrete, earth or steel. Structures are fabricated from concrete and steel. These materials are used both onshore and offshore. Offshore structures include fixed steel platforms, floating steel vessels and platforms, concrete gravity-based structures and floating concrete structures. Offshore also includes subsea foundations fabricated from steel that support subsea production equipment such as manifolds, distribution equipment, and flowline connection systems.

Data of interest for structures are strains, stresses, vibration, corrosion, and deflection. The performance of foundations is influenced by the geotechnical considerations and long-term soil condition. Soil can be disturbed by water, drought, earthquakes, subsea currents, and wave interaction in shallow water.

3.4 Pipelines/Storage

Pipelines and local flowlines are the main transport mechanism for hydrocarbons. At either end of the pipeline, the hydrocarbons are stored in large tanks. Along the length of pipelines, inline compressor or pump stations maintain pressures and gas/fluid flow rates. Local field flowlines are gathered into manifolds that connect into larger pipeline systems. Many pipelines are buried to safeguard them and prevent disruption of roads and rail. Pipelines eventually transport to “tank farms” or to refineries/LNG Plants.

Data of interest for pipelines include wall thickness, fluid composition, flow rates, temperatures and pressures. A measurement system using devices called pigs are pumped within pipelines (active and inactive) to measure wall thickness or to clean out debris or build up. These measurement systems should ideally be done remotely due to the length of pipelines and the fact that they are buried and not easily accessible. Many pipelines are very old (more than 50 years) and their integrity is in question. Pipelines need remote data systems for integrity and asset condition monitoring because they are buried and out of sight and reach.

3.5 Operations

Operational activities have been covered in part by previous sections relating to drilling activities and process control. Other specific activities not included previously that are less structured but still important to the digital facility include staffing levels and reports via formal means and informal means (emails, texts). In addition to managing the activities, the operations team conducts safety drills, stand downs for safety, testing, training, and routine maintenance. These activities are part of the overall data lake for the oil and gas facility.

3.6 MetOcean

MetOcean data requirements are specific to offshore oil and gas activities, although onshore production facilities and refineries are subject to inclement weather conditions. Floating drilling rigs and permanent floating production vessels are designed to operate in persistent waves, wind and currents. Daily drilling and production operations on floating vessels are designed by engineers to operate within “sea state” windows. Operators of floating vessels, therefore, measure wind speeds and directions, current speeds and directions, wave heights, periods and directions, air temperature and precipitation. Operators of floating vessels are also concerned that long term exposure to sea states creates fatigue in the floating structures, their mooring systems, and the riser systems that bring hydrocarbons on board the floating production vessel.

In addition to the MetOcean data collection, Data of interest to measure are stresses, strains, vessel motions, and corrosion on floating vessel structures and mooring lines (chain, wire rope, synthetics).

3.7 Health and Safety

The Health and Safety of workers in oil and gas applies to office workers, manufacturing facility workers, and workers in drilling, production and refining. Companies that employ the workers should ideally at a minimum provide safe environments and work processes according to government and industry regulations. For example, in the United States, this is governed by agencies such as the Occupational Safety and Health Administration (OSHA), Pipeline and Hazardous Materials Safety Administration (PHMSA), US Coast Guard, Bureau of Environmental Enforcement (BSEE), and the Bureau of Ocean Energy Management (BOEM). Most states also have local safety agencies. In the United Kingdom, this is governed by the Health and Safety Executive (HSE), and many European countries are governed by European Union OSHA (EU-OSHA). The American Petroleum Institute (API) also provides guidance on safe work practices.

The Health and Safety of workers is considered the "license" to operate in the oil and gas industry. Leading corporations have high standards for health and safety and expect to be best-in-class performers. Health and Safety are communicated with leading and lagging metrics that help indicate the dedication to safe working conditions and track incidents to provide transparency to customers and regulators.

Engineering design is critical to the safety of operations of facilities. Hazardous Identification and Assessment (HAZID) and Hazardous Operations (HAZOPS) are conducted during the design and engineering phases to identify and prevent or mitigate unsafe conditions and responses by workers.

Data of interest in lagging indicators are near misses, first aids, lost time incidents, property damage, stand downs, workers compensation insurance costs, and fatalities. Leading indicators include results of HAZIDs, HAZOPS, safety training, employee safety observations, and job safety analyses. The data collected contains unstructured reports, training documents and engineering design documents.

For more information, https://www.osha.gov/leadingindicators/docs/OSHA_Leading_Indicators.pdf and https://www.osha.gov/shpguidelines/program-evaluation.html#:~:text=Lagging%20indicators%20generally%20track%20worker,or%20illnesses%20before%20they%20occur. [e]

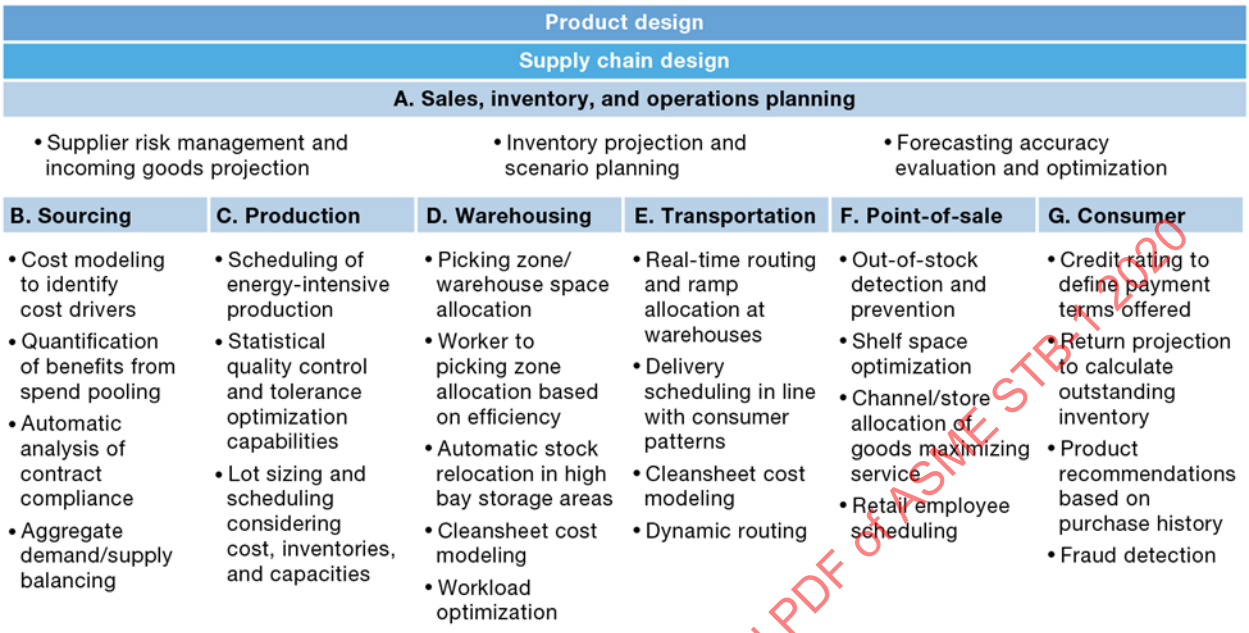
3.8 Supply Chain

Supply Chain is a business function that is closely related to engineering design and the successful construction and operation of oil and gas facilities. Each system, subsystem and component must be designed by engineers according to specifications, requisitioned by the engineer to Supply Chain for purchase, manufactured or fabricated according to the specifications' quality plans, tested, transported to site, commissioned and then started. Data is collected on each item or raw material for each step of the process. This quantitative data is created and stored on the ERP system of companies and then communicated to suppliers.

Scheduling and planning are critical to the success of a project and the eventual startup of the facility. The logistics of manufacturing/fabrication and then delivery to site requires data-intensive planning that includes costs, time to fabricate, time to transport, time to test and commission. Prescriptive analytics that optimize logistics are critical to meeting multiple and complex deadlines. Quality inspections at each stage of the activity are report-intensive and create volumes of unstructured data. Predictive analytics on suppliers can reduce risk by analyzing performance and preparing forecasts. Figure 3-10 provides an overview of Supply Chain Analytics considerations.

For more information, see https://www.mckinsey.com/business-functions/operations/our-insights/big-data-and-the-supply-chain-the-big-supply-chain-analytics-landscape-part-1# [f]

Figure 3-10: Supply Chain Analytics Landscape



Source: McKinsey & Company, "Big data and the supply chain: The big-supply-chain analytics landscape (Part 1)," February 2016, www.mckinsey.com. Reproduced with permission.

3.9 Special Note to this Chapter

An example of the data required for a digital twin of a pump is included as Appendix I-1. This figure is displayed similarly to Figure 3-1 and represents the details of potential information desired for a pump in a facility. The pump is representative of many components (objects) in a system. Future revisions of this Guideline contemplate expanding this chart to create a complete list of equipment and component systems from petroleum industry activities and facilities.

4 METHODS OF ANALYSIS

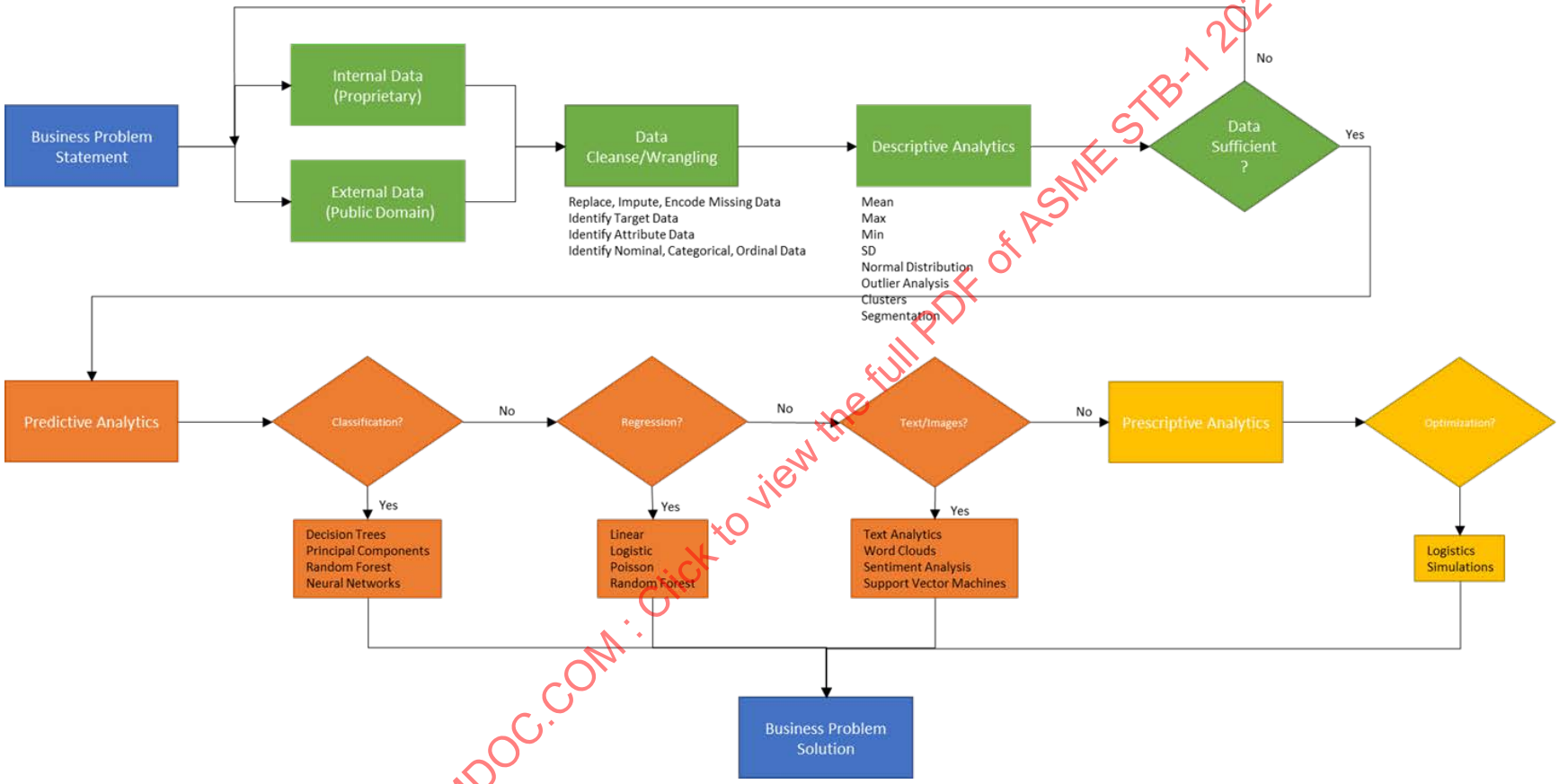
4.1 General Information on How and When to Use These Methods

A myriad of tools and methods have been created and used in statistics and analytics. New tools are created almost daily either with commercial software platforms or in programming languages such as Python and R. The number of tools, theories and methods can be overwhelming and intimidating without context and structure to their usage. This Chapter is designed to provide a structure and context for understanding what the tools are and how they can best be applied by an engineer.

Figure 4-1 is a simple flowchart that walks through a data analytics categorization journey. It represents a high-level, best-practice overview for analytics projects and is the basis for the chapter discussions ahead.

For a more detailed view of the data journey, please see Appendix II-1 for a very detailed journey map, which covers all of these topics. It resembles a subway map with junctions that have many waypoints and different destinations. This is representative of the variety of options in analytical tools and processes. It is common to journey down one or more paths of this map with the same data set to arrive at the best solution. For some data sets, several options are used and provide valid results. Using a combination of options to arrive at a solution is called ensemble analysis.

Figure 4-1: Data Analytics Journey Overview



4.2 Descriptive Analytics and Data Mining

4.2.1 Importance and Objectives

Proper use or modeling of data sets is dependent on reliable data. Even modestly-sized data sets have missing values, incorrectly keyed-in values, and outliers. Computational tools must be instructed to identify categorical versus nominal values. Most analytical methods are not valid unless the data set is normally distributed. Finally, with Big Data, much of it is unstructured text, recordings, and photography. Up to 80% of data in an organization is text based. The practice of cleaning, identifying and filtering data is called Data Mining. Reliable analytical solutions require thorough application of Data Mining techniques. Data Mining is represented by the green boxes in the Figure 4-1 flowchart.

Table 4-1 is provided to illustrate Data Mining through the lens of 5S Lean techniques. This is a step by step process that should be repeated until the data set is ready for analysis. The first two steps – *Sort* and *Set in Order* – are crucial in the cleansing or "wrangling" of data sets.

Table 4-1: 5S Lean Approach to Data Mining

5S Category	Data Analytics 5S Process
Sort	Examine available data for suitability
	Discard data that is not useful or meaningful, partition data
	Identify and remove data that is misleading or may lead to incorrect modeling
Set in Order	Confirm and partition data
	Identify variables as categorical or nominal
	Identify variables as target or attributes
	Replace, impute and encode missing data
Shine	Create Descriptive Analytics to assess data distribution, means, standard deviations, etc.
	Identify and remove outliers
	Run training models and assess results
	Utilize visualization techniques to further assess
	Run validation model to check for fitness and parsimony
	Rerun model or model(s) until results are stable and validated
Standardize	Check the analytics tools on related but different data sets for validation
	Present results in a concise and business setting to management and stakeholders
	Revise and update according to stakeholder requests
	Complete written descriptions and instructions for use
	Issue final model for use by designated stakeholders
Sustain	Review and update the model as required on a regularly scheduled basis
	Inform stakeholders of updates to model and consequences of updates
	Consider re-modeling as new techniques and or new data become available

4.2.2 General Statistical Descriptors

The *Shine* step creates a set of descriptors for the data set that describes summary statistics. The common descriptors are mean, maximum value, minimum value, and standard deviation. These describe the Central Tendency or Central Limit Theorem of the data. Additional descriptors are mean squared error, average squared error, quartiles, and t-tests. These describe the Measures of Dispersion of the data. Histograms and Box Plots are excellent tools to visualize the distribution of data and determine normal attributes verses skewness. These characteristics are investigated for Single Variable or Univariate data sets.

Multivariate data sets, those with more than one variable, can also be investigated for relationships between the variables. Matrices of correlations and covariance are developed to investigate these relationships. Correlation coefficients describe the mathematical dependencies between the variables.

Some multivariate sets of data can be described using a technique called Clustering. Clustering uses mathematical techniques to segment sets of variables by commonalities that generally describe the segment. Clustering is commonly used to describe large data sets with multiple levels of relationships.

As suggested by Figure 4-1, the Shine process is repeated until the data set is fully described and understood. Further Shine steps apply to Predictive Analytics through running models and validating results.

4.2.3 Descriptive Analytical Tools

Whether using basic statistical functions or a sophisticated statistical program like SAS, the data scientist can produce tables with Central Tendency Data and Measures of Dispersion. Visual tools depicted in Figure 4-2 graphically represent examples of tools that quickly identify the character of the data that augment the tables. Visual tools also provide greater insights that tables cannot supply.

Depicted in Figure 4-2 are:

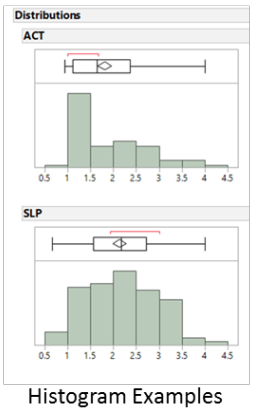
(a) Two histograms, one is normal, the other is skewed. The data in the skewed histogram must be further examined prior to analysis and potentially have the data transformed with linear or nonlinear functions to introduce normality. *For example, a data analyst will want sufficient data for analysis and a normal distribution will confirm that sufficient data has been collected.*

(b) A scatter plot with two obvious outliers. The outliers must be examined by looking at the specific data points and explained by a subject matter expert or discounted as errors. *For example, when measuring fluid properties from a well stream, the instrumentation was malfunctioning that particular day. The operator would know that information and have logged it in a report.*

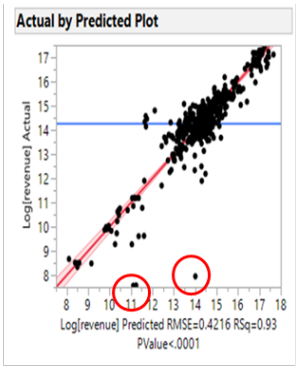
(c) A scatter plot with three clusters defined using the K-means algorithm. These clusters can then be described based on characteristics of each cluster. *For example, using a report of location, drilling techniques, well depth, fluid temperature, fluid pressure and clustering to find segments of locations with similar characteristics. These locations could inform a driller of what is required and expected at the next location that is similar to one of the desired segments.*

(d) A scatter plot and associated histogram for bivariate relationships. Multivariate analysis techniques require multivariate normality, therefore this Bivariate Distribution is useful in determining suitability for analytical modeling. *For example, reviewing temperature and pressure readings from a pressure vessel to determine correlations and covariance.*

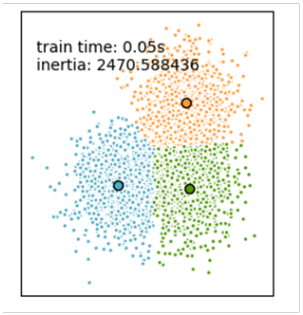
Figure 4-2: Descriptive Analytic Tool Examples



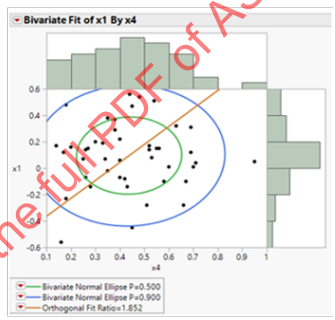
Histogram Examples



Scatterplot with Outliers



Clustering
Source: https://scikit-learn.org/stable/modules/clustering.html#k-means



Example of Bivariate Distribution

4.3 Predictive Analytics

4.3.1 Importance and Objectives

Predictive Analytics are depicted in the orange section of the flow chart in Figure 4-1. Tools used in Predictive Analytics discover correlations between independent variables or *attributes*, and the desired variable, or *target*. Correlations can be identified and calculated to forecast future events; hence they are *predictive*.

The ability to understand the nature of the data and how to model it is a key skill described in the 5S Shine activity in Table 4-1. Predictive analytical tools assist in developing, testing, and validating these models. This activity is known as training the model.

Predictive Analytics [b] are divided into two main model types:

(a) Supervised
- (1) Contain both targets and attributes
- (2) Train the model by minimizing the error between observed targets and the model predictions
- (3) Examples (For reference on targets, see Table 2-1)

- Linear and Nonlinear Regression – forecasting an Interval Target
- Logistic Regression – forecasting a Nominal Target
- Neural Networks – forecasting an Interval or Nominal Target
- Decision Trees – Interval or Nominal Targets
- Ensemble Models – Interval or Nominal Targets

(b) Unsupervised Models

(1) Contain only attributes
(2) Trained by minimizing a well-defined optimization function
(3) Text data mining is an unsupervised application
(4) Examples

- Principal Components Analysis
- Factor Analysis
- K Nearest Neighbor Cluster Analysis
- Sentiment and Opinion Analysis
- Ensemble Models from Unsupervised Models

4.3.2 Regression Problems and Solutions

Regression Analytics model the relationship between a target variable and its attribute(s). Linear regression involves developing parameters of an equation with an intercept and the slope of a line. Multiple linear regression has one target but multiple attributes.

Logistic regressions calculate the maximum likelihood of a target response based on binary variables: yes/no, male/female, etc.

In addition to parameters, regressions have hyperparameters – Model Attributes, Transformations and Interactions that are accounted for in the solution equations. Most regression models are rated on their ability to demonstrate Goodness of Fit and/or Validation. Goodness of Fit measures how well the model forecasts results. When assessing empirical operations event data, validation is a technique whereby the model data is randomly divided (usually 70/30) into a training and holdout set. The training model develops the equation; it is then tested on the holdout set to validate it. Validation is a key component of 5S Shine.

Examples of Regression Analyses:

(a) *Given the historical oil price and other economic indices concurrent with the prices, predict the price of oil based on the other economic indices. (Linear)*
(b) *Given the number of cycles of a valve opening and closing, use the rate of open and close to predict when it will begin to leak. (Logistic)*

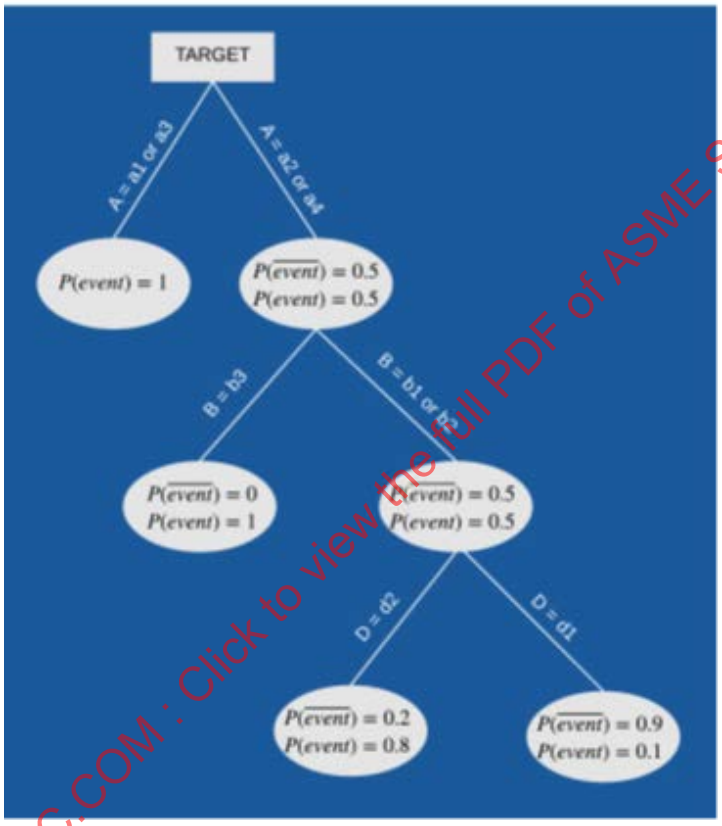
4.3.3 Classification Problems and Solutions

Classification analytics are versatile algorithms that predict results using tools like Decision Trees, Random Forests and Neural Networks. Also gaining acceptance are Support Vector Machines and K Nearest Neighbors. Each of these tools has unique equations and properties to categorize data. Because these models may vary in results with the same input data, it is often common and prudent to run and Ensemble Case with 2 or 3 of these tools. These tools work by generating rules for categories within the data and selecting the best category for the next data point based on the rules generated. The ultimate tool to measure Classification is the Confusion Matrix. A few models are discussed here. For more information, please see https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623 [g].

(a) Decision Trees

Decision trees are easily interpreted models that perform similar analyses as regression models, however they can work with a mixture of data types. They are also able to operate with missing data. The target appears at the top of the tree, and the attributes are split along the tree according to binary choices based on current data. The disadvantage is that they can become too complex and overfit the data (that is, the model can only be used for that specific data set). An example is shown in Figure 4-3.

Figure 4-3: Decision Tree



Source: Edward R. Jones, The Art and Science of Data Analytics, 2018 [h]

(b) Random Forests

Random Forests are an ensemble of Decision Trees. Decision Trees are built sequentially with a set of data that does not change. If the data changes slightly, the Decision Tree results are changed. To account for these changes, a Random Forest is constructed by building many trees, each from a randomly selected dataset and then averaging their forecasts. This is referred to as ensemble modeling.

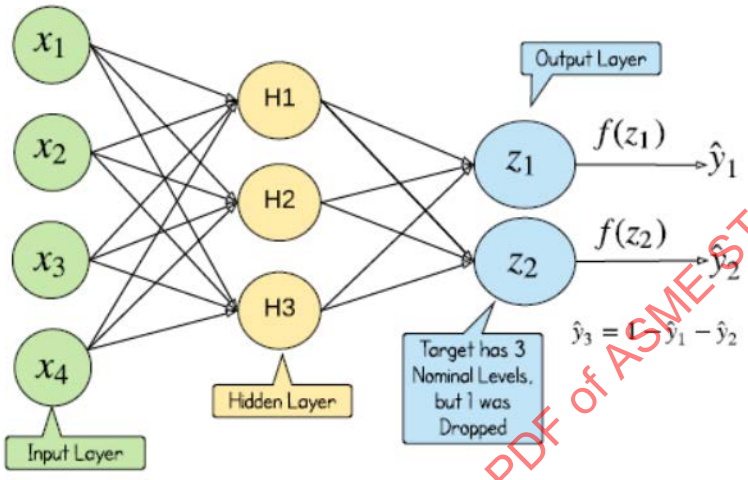
Random Forests are very accurate but take an immense amount of computing power.

(c) Neural Networks

Neural networks are built on step functions that are patterned after research performed on the human brain in the 1930's and 40's. Based on the work of McCulloch and Pitts, the equation mirror is shown in Figure 4-4. The neurons of the brain, which receive +1 or -1 signals from neuron to neuron are duplicated and processed to define and answer from multiple inputs. In Figure 4-4, the input and output layers are seen, but the hidden layers conduct the calculations. Because the hidden layers cannot be revealed, it makes it difficult to explain results.

The networks are the backbone of Artificial Intelligence. For example, Neural networks are good tools for asset condition monitoring and predictive maintenance due to their ability to train and adapt. *They are also used for Process Control.*

Figure 4-4: Neural Network



Source: Edward R. Jones, The Art and Science of Data Analytics, 2018 [h]

4.3.4 Unstructured Data Problems and Solutions

Unstructured data is the largest component of Big Data and is potentially the hardest to data mine and comprehend. The tools and training for text analytics and graphic data are complex and understood by a small subset of business professionals. From Bernard Marr, internationally best-selling author, keynote speaker, futurist, and strategic business and technology advisor, July 19, 2020, *"There are many challenges associated with big data. When I work with business leaders to help them develop their data strategy, often the first pitfalls or challenges they think of are around technology or skills. In other words, without the technical infrastructure, in-house knowledge, or vast budgets of companies like Amazon or Facebook, many business leaders think the advantages of big data are beyond their grasp."*

Storage alternatives and programming tools are advancing rapidly and democratically to address Big Data and its unstructured nature. Machine learning is at the core of the following solutions:

(a) Text Mining – Natural Language Processing (NLP) techniques are prevalent and accessible. These tools provide insights using word clouds, sentiment analysis, topic analysis and web scraping to search for more data. *For example, data gleaned from internal emails, texts and reports can be mined for certain key words and other information on activities and potentially point to upcoming high potential incidents.*
(b) Image Recognition – the ability to "search" photographs to find patterns and match them against known patterns. Facial recognition software is one example. *In the oil field, image recognition can be used to search external coatings for signs of corrosion, damage or tampering.*

4.3.5 Time Series

Time Series analysis is a forecasting tool for data that follows a sequence over time. Auto-Regressive Moving Average (ARMA) and Auto-Regressive Integrated Moving Average (ARIMA) are examples of special models that detect the underlying structure of the data and other factors that determine a forecast. These forecasts support decision making for the next time period(s) by determining the contributing factors to seasonality, trend variations, and fluctuations.

Common examples of Time Series forecasts include process pressure, temperatures, fluid composition, commodity prices, oil and gas production forecasts and stock market behavior(s).

4.4 Prescriptive Analytics

4.4.1 Importance and Objectives

Prescriptive Analytics have traditionally been called Optimization calculations. Prescriptive Analytics are crucial to Supply Chain Management logistics.

Prescriptive Analytics also refer to applying the decisions of Predictive Analytics, and then reviewing outcomes, revising and then re-applying them. The main objective is the best outcome. *For example, Prescriptive Analytics has been deployed to predict when and why an Electrical Submersible Pump (ESP) will fail and recommend the necessary actions to prevent the failure.*

4.4.2 Optimization Problems and Solutions

Oil and gas assets are large, complex and use vast resources of people, cash and energy to operate. Minimizing cost and maximizing revenues/outputs are the goal of every organization. Decisions must be based on mathematical algorithms. The algorithms for Optimization use:

(a) Objective Function – describes the desired outcome to maximize targets such as revenue or manufacturing output
(b) Constraints – describes limitations on the system to be calculated such as fuel available, weather windows, etc.

The Optimization software program calculates up to dozens of constraints in any operation to provide a more complete picture of what is required and to use resources correctly without waste and schedule delay.

4.4.3 Simulation Problems and Solutions

Process Simulation provides the ability to train operators, optimize output, predict incidents and prepare for unknowns. These software tools are based on the concept of the Digital Twin. The Digital Twin can be as simple as one major component of a system (jet engines), or the entire system as a Digital Facility (see section 3.1.2). Facilities that can simulate complete operations can enhance safety by lowering or eliminating staffing required to operate. The goal of the offshore upstream industry is to mimic this approach to minimally staff offshore facilities.

4.5 Application Program Interfaces

4.5.1 Importance and Objectives

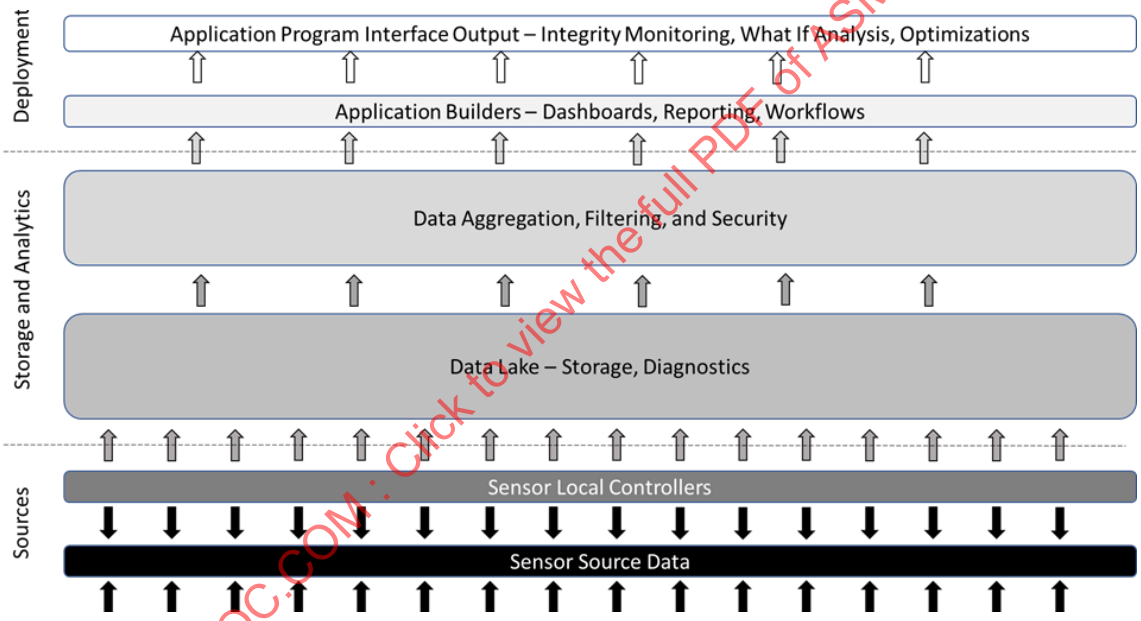
Application Program Interfaces (APIs) are the last step in the analytics software tool portfolio. APIs allow the analytical model to be translated into an actionable tool for others that did not write the

software or create the model. This is part of the 5S Standardize activities to deploy a useful tool. At this stage, data becomes a Business Problem Solution as referenced in Figure 4-1.

4.5.2 Implementation

Figure 4-5 represents the journey of data from the source to actionable information. At the source (called "edge" in computing) data is generated from the instruments and equipment discussed previously in Chapter 3. Local controllers manage the operation at the edge and transfer the data to the storage repository (data lake). The data is filtered, cleansed, secured, aggregated and modeled (machine learning). The next step is to transform the data into "meaning" through APIs that can be interpreted both by humans monitoring dashboards and by computers developing actions based on the data models. These dashboards provide real time visualization tools in a format that is easily understood by subject matter experts on the sources that do not need to understand the analytics engine providing the visualization.

Figure 4-5: Data Journey from Source to API



APIs are constructed using programming languages such as Python, Java or MATLAB. Real-Time Specification for Java is a popular set of interfaces that allow real world computing

4.6 Visualization Tools

Not all data analytics need to be captured in an API or dashboard for use in an actionable tool or machine interface. Many business applications for the engineering team are available to describe the results of analytical projects to inform business leadership. These tools have been designed for a range of practitioners in analytics – novice to expert. The objective is to have a business-friendly representation of the data analysis and for it to express meaning, not just charts and graphs. Collectively, these tools are known as Business Intelligence (BI).

For more information and reviews of these tools, please see this article by Bernard Marr:

https://www.linkedin.com/pulse/10-best-data-analytics-bi-platforms-tools-2020-bernard-marr/ [i].

5 DATA ANALYTICS PROJECT WORKFLOWS

5.1 Introduction

5.1.1 CRISP-DM

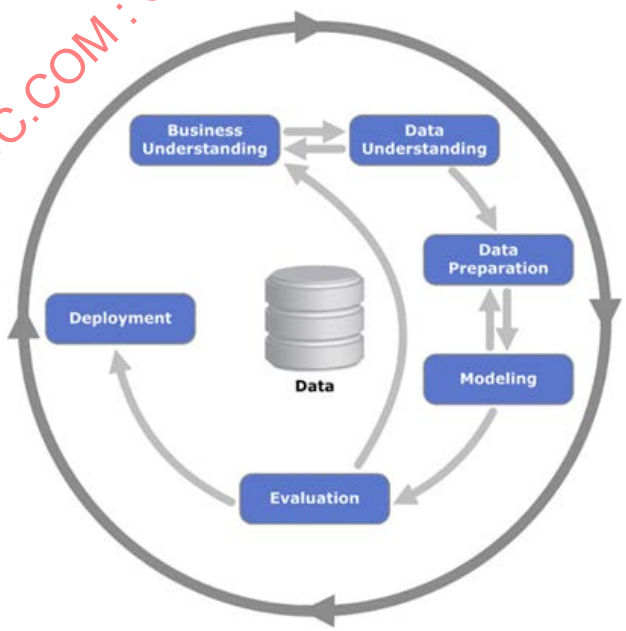
Even with the most fantastic instrumentation and data analytics tools, the effort to mine and represent data is meaningless without providing value to the business enterprise. This value is questioned by business leaders daily as they confront the avalanche and enthusiasm of the Big Data world. Data is no different than any other business process: it must follow a standard to establish credibility and reliability for the business enterprise. This Chapter describes data business processes developed to provide credibility and utility to the business leader who wants data insights to add value, not just cost and headcount.

In 1996, data and business professionals assembled and produced the first standard called Cross Industry Standard Process for Data Mining (CRISP-DM). This was adopted by multiple industries because it established:

(a) Business Understanding – turning objectives into a data analytics project
(b) Data Understanding – research and understanding of the data sets
(c) Data Preparation – see Table 4-1 – 5S Approach to Data Mining
(d) Modeling – see Chapter 4 – Methods of Analysis
(e) Evaluation – reviewing the model for the best results
(f) Deployment – using the data and model in operations

Graphically described in Figure 5-1, this is a cyclical process that will require repetition and iteration until the data sets are suitable for modeling; the modeling produces reliable results, and the ultimate business challenge has been addressed.

Figure 5-1: CRISP-DM Business Process



Source: CRISP-DM Business Process Certified Analytics Professional (CAP®) Examination Study Guide, Publisher: Institute for Operations and Management Sciences (INFORMS)

5.1.2 INFORMS and the Job Task Analysis

Building on the CRISP-DM, the Institute for Operations Research and the Management Sciences (INFORMS) developed an in-depth Job Task Analysis (JTA) that expanded on the principles of CRISP-DM to develop the Certified Analytics Professional curriculum. [j]

INFORMS promotes best practices and advances in operations research, management science, and analytics to improve operational processes, decision-making, and outcomes through an array of highly-cited publications, conferences, competitions, networking communities, and professional development services." Nonmandatory Appendix B provides more detail on the Certified Analytics Professional (CAP^{R}) and how to learn about INFORMS.

The JTA framework created by INFORMS is the basis for this Chapter and how to execute data analytics projects for oil and gas projects.

5.1.3 Structure, Roles and Responsibilities

By combining Six Sigma techniques and the JTA, a Supplier-Input-Project-Output-Customer (SIPOC) chart has been created for this Guideline in Mandatory Appendix II-2. This structure describes the JTA with a systems view that identifies the entities in a data project and high-level roles.

Projects with this level of detail require specific assignments of job roles and responsibilities. These roles for team members (described as example Job Functions), Mandatory Appendix II-3, have been described in a Responsible-Accountable-Consultative-Informed (RACI) chart in Mandatory Appendix II-4. References to these Appendix materials are included throughout Chapter 5.

5.1.4 Value to the Enterprise

The following examples are real world business challenges that were solved using data analytics techniques. Each of these challenges were determined to have added value to the enterprise.

(a) Enabling Normally Unattended Facilities (NUF)
The offshore facility of the future is likely to look substantially different than today. The objective of reducing risk and adding value provides a large opportunity to make the industry substantially safer as well as cost competitive.

Every human hour spent offshore on a live hydrocarbon facility is a risk manhour. We manage and mitigate that risk but the opportunity to eliminate it is a substantial prize.

As we strive towards net zero, reducing or eliminating Tier 1 emissions is critical. We can massively impact this by reducing transit trips for human operation.

Eliminating facilities to safely sustain human life allows for significant reductions in Capital Expenditures, (CAPEX) and Operating Expenditures (OPEX). For example:

(1) A change from a 64 individual human operation to NUF can eliminate 14 Million Risk Manhours; 64 humans x 365 days x 25 years.
(2) Just eliminating 60% is a reduction of 8.4 Million Risk Manhours for one facility.
(3) Multiple operators have now published studies showing opportunity reductions of 30% on CAPEX and 22% on OPEX over traditional facilities.
(4) Potential savings: $300MM per facility (this will vary on size, complexity and regional locations, it could be more or less).

The Digital Facility described in Chapter 3 provides the vehicle to produce the NUF.

(b) Supply Chain - Tracking
The complexity of global supply chain is an immense challenge. A facility can procure millions of items from hundreds of countries and thousands of locations over its lifetime. Tracking, locating and maximizing this offers significant opportunities. For example:

(1) Tracking alone, knowing what, where and how to deliver in an expedited manner remains one of the outstanding opportunities.
(2) The value comes from improving and optimizing use of components across an enterprise not just a facility.
(3) Losses from misplacement, theft or just lack of awareness is estimated to cost the oil and gas industry $200-300MM annually.
(4) Downtime through slow or lack of availability can exceed $2-3 Billion a year.

This is solved through appropriate tagging, tracking, standardization, and a robust digital system. Predictive Analytic techniques and machine learning are reliable tools for this type of analysis.

(c) Supply Chain - Optimization
The ability to maximize throughput and delivery offers substantial improvements in economic performance. One extra Liquefied Natural Gas (LNG) shipment over plan could mean improved sales of $20MM. For example:

1% improvement on production throughput:
(1) 7,500 BBL a day X 0.001 = 75BBLS x 365 x 10 years = 273,750 BBLs
(2) Multiply by 1,000 wells @ $40 per barrel = approximately $11.0 B

The challenge is to get from the 1-2 well application to the 1,000 and beyond. Scaling and adoption of solutions is a business challenge that should be championed from the top down by business leaders and bottom up from the subject matter experts and data professionals. Prescriptive Analytics that optimize operations are the best tools for supply chain challenges.

(d) Data Driven Permian Basin Production Forecasting Using Machine Learning
The next several sections will discuss an example data-driven solution for forecasting success in drilling in the Permian Basin in the United States.

5.2 Business Problem Framing

5.2.1 Description

Business problems confront leadership daily. Not all of them are solved by data analytics, but for those that are, the data professionals should be available as resources to judge the applicability of a data-based solution. Leadership needs to articulate the issues confronting the business, a basic knowledge of how the business operates, and a list of internal and external stakeholders. Ideally, the business leaders will produce a problem statement document, identified business needs and constraints, and general agreement from the management team on the need to solve this problem (See Mandatory Appendix II-2).

(a) Per INFORMS, Business Problem Framing includes:

(1) Obtain or receive the problem statement and usability requirements
(2) Identify stakeholders
(3) Determine whether the problem is amenable to an analytics solution
(4) Refine the problem statement and delineate constraints
(5) Define an initial set of business benefits
(6) Obtain stakeholder agreement on the business problem statement

(b) From the INFORMS CAP[R] Handbook, the Business Framing should consider the following questions:

(1) Who: are the stakeholders who satisfy one or more of the following with respect to the project: funding, using, creating, or affected by the project's outcome?
(2) What: problem/function is the project meant to solve/perform?
(3) Where: does the problem occur? Or where does the function need to be performed? Are the physical and spatial characteristics articulated?
(4) When: does the problem occur, or function need to be performed? When does the project need to be completed?
(5) Why: does the problem occur, or function need to occur?

5.2.2 Team Member Roles

The Responsible Party in Business Framing is the Data Analytics Lead supported by the other team members as shown in Table 5-1. All members have a role, but the most active are the Data Analytics Lead and the Executive with the Business Challenge. Key to success at this stage is articulating the business challenge.

Table 5-1: RACI Chart for Business Framing

	Data Analytics Team Lead	Data Engineer	Data Scientist	SME Engineer	SME Product Line	IT	Department Head	Technology / R&D	Project Manager	Finance	Marketing	Executive	Sponsor
1.0 Business Problem Framing													
1.1 Receive and refine the business problem	R	I	I	C	C	C	A	C	C	C	C	C	C
1.2 Identify stakeholders	R	I	I	I	C	I	A	I	C	I	I	I	I
1.3 Determine whether the problem is amenable to analytics solution	R	C	C	C	C	I	A	I	I	I	I	I	I
1.4 Refine business problem and delineate constraints	R	C	C	C	C	C	A	I	I	I	I	I	I
1.5 Define initial set of business benefits	R	I	I	C	C	C	A	C	I	C	C	C	C
1.6 Obtain stakeholder agreement on problem statement	R	I	I	C	C	C	A	C	C	C	C	C	C

Legend:
R - Process Responsibility
R - Improvement Responsibility
A - Process Accountability
C - Process/Improvement Consult
I - Process/Improvement Inform

5.2.3 Example Business Challenge – Permian Basin Production Forecasting

(The following is used with permission of Yacine Meridji, MS Data Analytics Class of 2020, Texas A&M University [k]. His full capstone project is included in Nonmandatory Appendix A and represents a fully developed data project. The details of his project will be revealed according to the INFORMS phases. He is the Data Scientist referenced in the discussion)

After a century of oil and gas production, the Permian Basin still has 180 Billion barrels of technically recoverable production. Knowing where to drill and what techniques to use to enhance current recoveries requires an examination of vast amounts of data to build these predictions.

(a) The Business Framing question for this project:

(1) What is driving production in the Permian Basin?
(2) Is it Localized? Why are certain counties better than others?
(3) What Operators and Basins can we analyze?

(b) The Data Scientist collected the following information from oil and gas operators:

(1) In 2018 US became the #1 oil producer with the most contribution coming from the Permian Basin
(2) Technology was key to improved well performance
(3) Permian has nearly 3,000 ft. depth of good quality shales and a vast aerial extent of 75,000 square miles
(4) Improved economics are needed for investment
(5) Existing infrastructure should be leveraged
(6) Better recovery factors are desired than 10-15%

5.3 Analytics Problem Framing

5.3.1 Description

The Analytics Team Lead is the interface between the data experts and the business leaders/experts. This person must decide the actual ability to reframe the business problem to an analytics problem. During this phase, the data science team works to produce a Problem Statement Framing Document that discusses success metrics. This document is shared with and agreed to by all stakeholders. (See Mandatory Appendix II-2).

(a) Per INFORMS, the steps are:

(1) Reformulate a problem statement as an analytics problem
(2) Develop a proposed set of drivers and relationships to inputs
(3) State the set of assumptions related to the problem
(4) Define key metrics of success
(5) Obtain stakeholder agreement on the approach

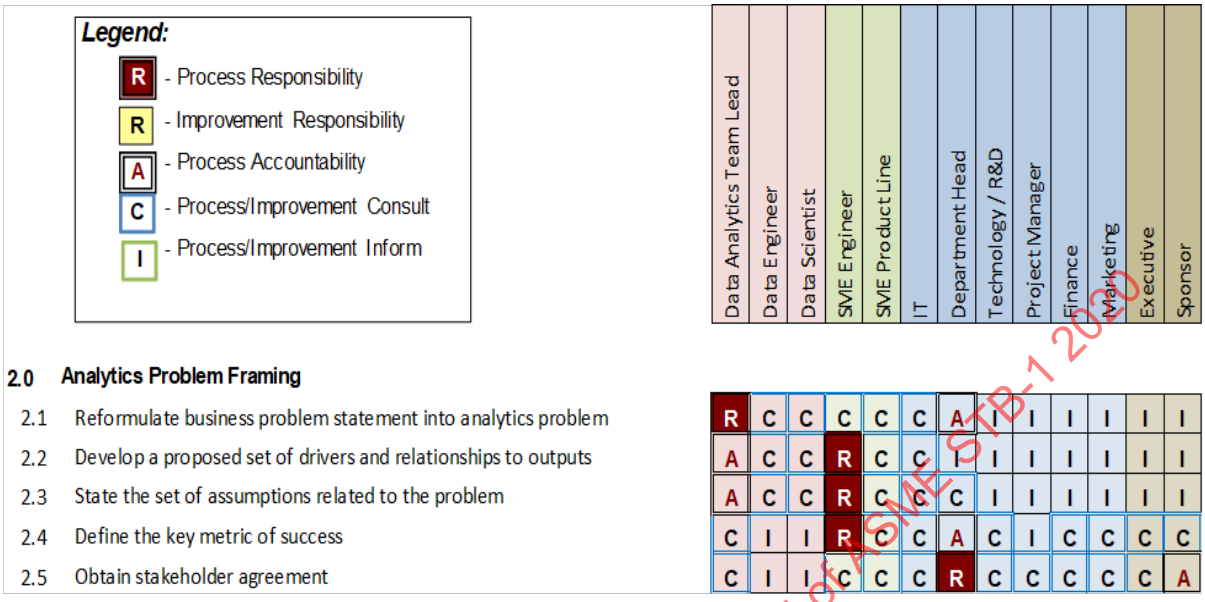
(b) Restated as questions:

(1) What result do we want?
(2) Who will act?
(3) What will they do?
(4) What will change in the organization as a result of the new information generated?

5.3.2 Team Member Roles

The Data Analytics Team Lead now shares responsibility with an SME Engineer and the Department Head for the product or services, and should work together to find alignment on what is needed and helpful to all stakeholders as shown in Table 5-2.

Table 5-2: Analytics Problem Framing RACI Chart



5.3.3 Example Business Challenge - Permian Basin Forecasting Model Continued

As this document returns to the example challenge of the Permian Basin Forecasting Model, the question of how to frame the Business Challenge is addressed.

The Data Scientist hypothesized that lateral length and proppant amount influence production improvements. His challenge, then, was to examine production data from several fields and validate this hypothesis. He spent several months researching and collecting from data sources that were available in the public domain. His data targets needed information on the lateral extent of horizontal wells and the quantity of proppant used to fracture the wells.

The Data Scientist performed the project independently, but identified the team members required in a normal business team environment:

(a) Data Scientist	Himself
(b) SME Engineer	Reservoir Engineer, Geologist
(c) SME Production Line	Completions Engineer

5.4 Data

5.4.1 Description

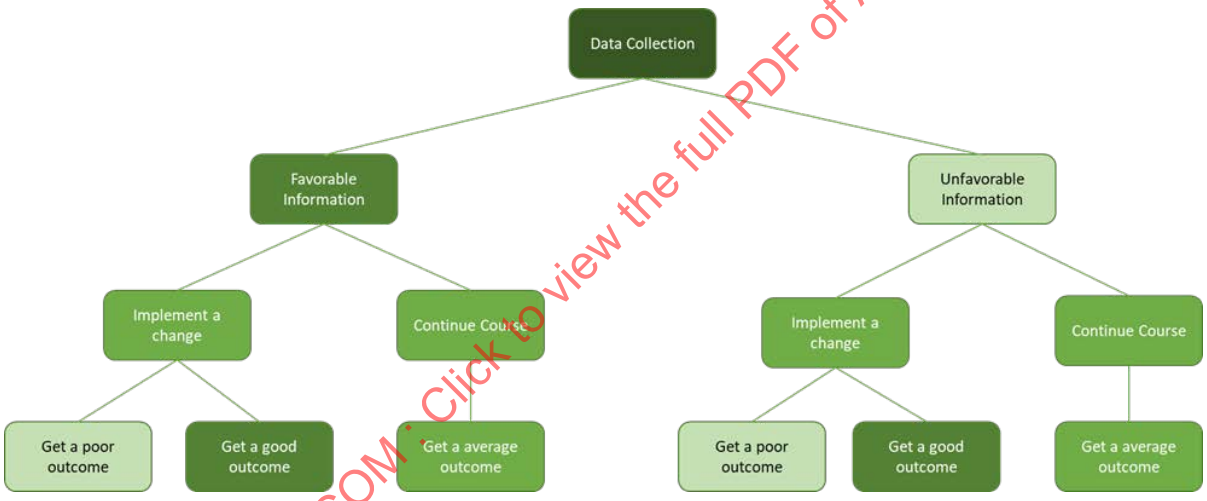
Analytics is the process of making data meaningful for business decision-making. Several steps are required to collect the data, understand it, cleanse it and validate it for modeling purposes. The team determines if it has access to data and an adequate IT system for cleaning and exploration of the data. As data is explored, it is matched against business and analytical problem statements. The data set(s) are documented and prepared as a workable set for analytical methods. The analytical team may need to refine the problem statements based on the available data and its suitability for analysis. Once this step is completed, the team gets approval to proceed with the analyses (see Mandatory Appendix II-2).

(a) From INFORMS, the steps are:

(1) Identify and prioritize data needs and resources
(2) Identify means of data collection and acquisition
(3) Determine how and why to harmonize, rescale, clean and share data
(4) Identify ways of discovering relationships in the data
(5) Determine the documentation and reporting of findings
(6) Use data analysis results to refine business and analytics problem statements

(b) This can be a lengthy and iterative process. In a timeline of activities for data projects, more than 70% of the time spent is on data "wrangling" activities to get a data set that is prepared for analysis. INFORMS suggests a decision tree approach. This is time consuming, but well worth the time devoted once modeling commences. An example decision tree on data would look like Figure 5-2 with the objective of gathering favorable data. This process should continue until the data set is complete and cleansed of missing or outlier values that cannot be explained.

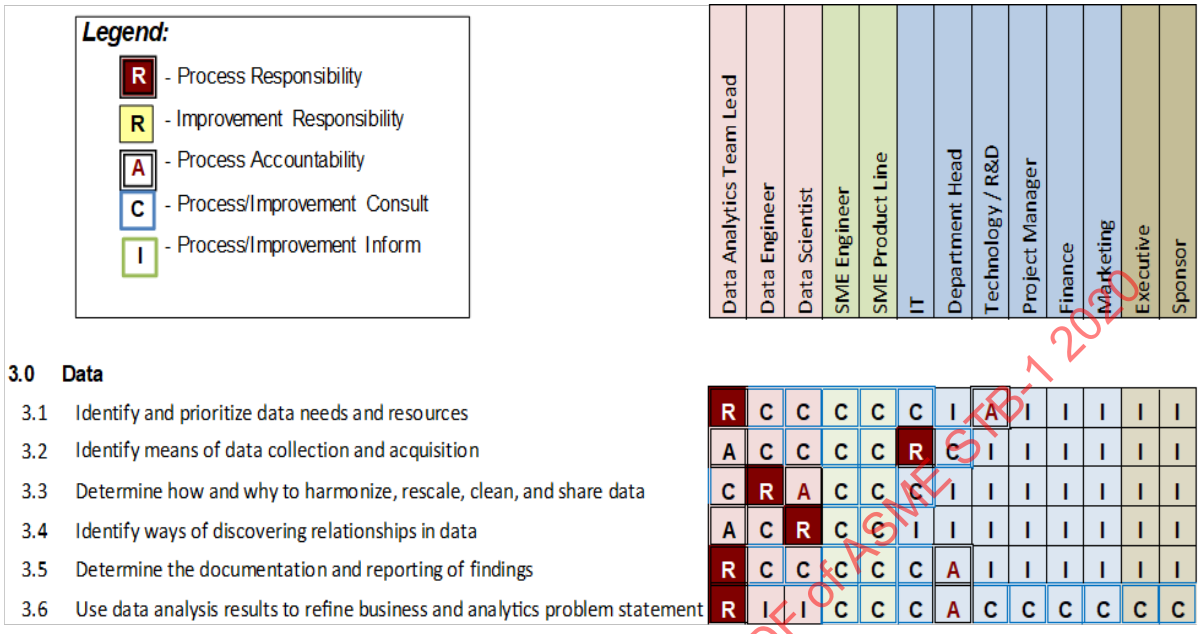
Figure 5-2: Data Selection Decision Tree



5.4.2 Team Member Roles

This team has a larger set of Responsible Role Players. The Data Analyst Lead is joined by both the Data Engineer and Data Scientist along with a leading role by IT as shown in Table 5-3. Data collection and storage can place burdens on the business operating systems of a company, therefore cooperation and leadership from IT is crucial.

Table 5-3: Data RACI Chart



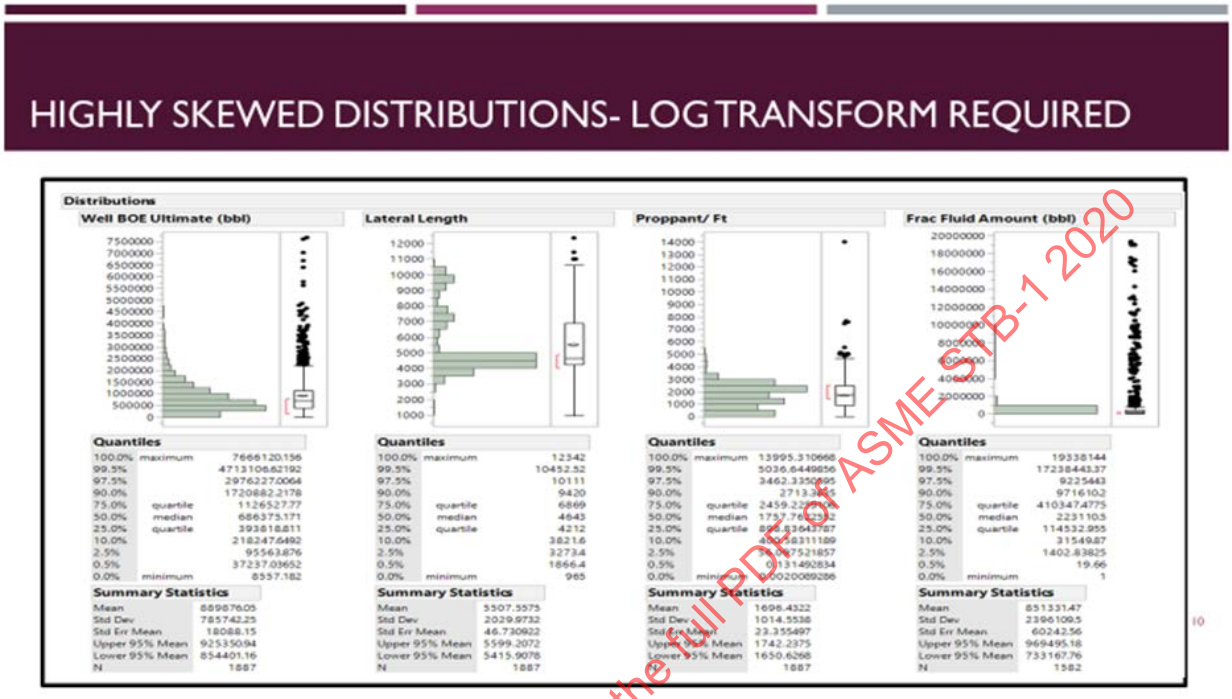
5.4.3 Example Business Challenge - Permian Basin Forecasting Model Continued

Returning to the Example Business Challenge, The Data Scientist identified several sources of data that he could obtain from public domain and from a company that had data that was not proprietary. The sources:

(a) Obtained the data from Tomlinson Geophysical Services and will be using n=1887 wells in the Delaware Basin
(b) Same vintage data all wells drilled and completed 2007-2018
(c) 14 dependent variables were used to predict EUR (Estimated Ultimate Recovery)
(d) Several missing data points were removed, some replaced with the mean values
(e) Used n=1512 wells post data clean up
(f) Focused on top two reservoirs: Wolfcamp and Bonespring (most prolific)
(g) All wells are in the Delaware Basin (West Permian Basin)

The Data Scientist understood that data cleanup was required and did so using replacement techniques discussed in Chapter 2. During the process of descriptive statistical review, he noticed skewed histograms and elected to log transform the data to obtain the distribution normality required for model development (see Figure 5-3).

Figure 5-3: Descriptive Analytics for Permian Basin Data Sets



5.5 Methodology Approach and Selection

5.5.1 Description

The Data Scientist has the depth of knowledge and experience to decide on the methodology approach based on the business framing and the available data. More than one approach is likely available to the team and they can use all available approaches to combine into an ensemble. The team needs to evaluate software tools, skill sets of the team, and the desired results of the analysis. (See Mandatory Appendix II-2).

(a) The INFORMS approach is concise and indicates that some activities are repeated until a successful conclusion:

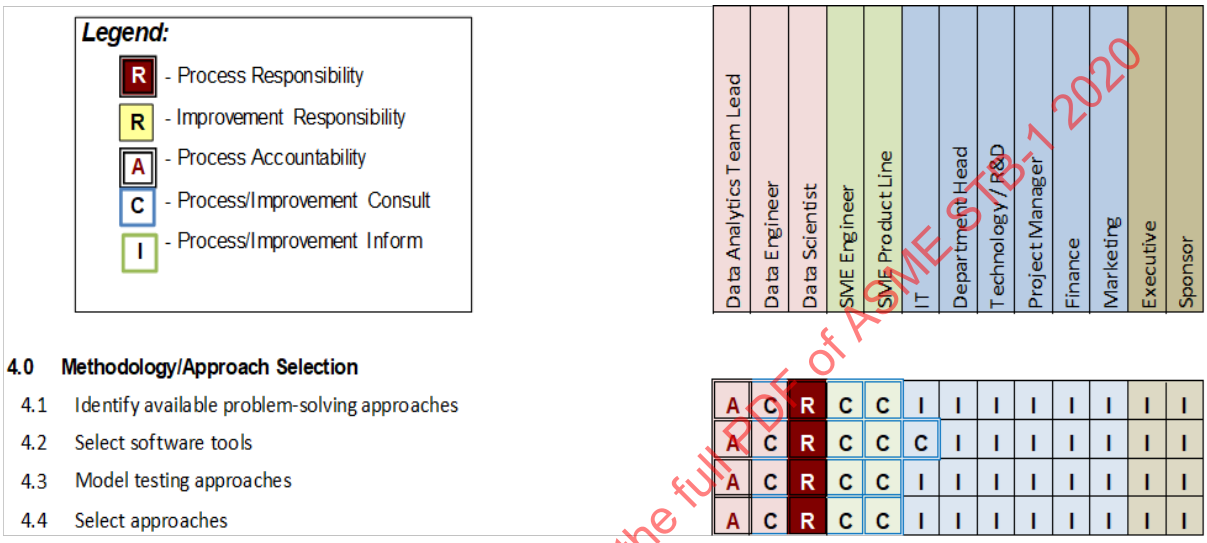
(1) Identify available problem-solving approaches (methods)
(2) Select software tools
(3) Test approaches (methods)
(4) Select approaches (methods)

(b) Referring to Figure 4-1, the methodology and output are dependent on the problem to be solved. Is the problem requiring Predictions on what could happen, or Prescriptions on what should happen? Does the problem require both? Success with the selection is highly dependent on the experience, wisdom and perseverance of the team.

5.5.2 Team Member Roles

The Responsibility at this phase shifts to the Data Scientist, as shown in Table 5-4. This is the individual with the experience and judgement to select the model strategy. The Data Scientist will rely heavily on support from the Data Engineer and the SMEs in Engineering and Product line.

Table 5-4: Methodology Selection RACI Chart



5.5.3 Example Business Challenge - Permian Basin Forecasting Model Continued

As the Example Business Challenge continues, the Data Scientist developed a plan for the analysis after examining the data and reviewing the goal to find the Estimated Ultimate Recovery (EUR). He also wanted to explore more business outcomes for a potential operator. His final approach:

(a) Use machine learning to identify key controllable well parameters that drive production
(b) Improve operational efficiency, reduce cost and enhance well deliverability
(c) Forecast productivity from existing and future wells
(d) Identify AOI (Area of Interest) true economic potential
(e) Use the model findings to optimize field development plans
(f) Identify outliers and possible re-completion candidates
(g) Sweet spot mapping
(h) Identify key players in the basin and why are they successful.

5.6 Model Building and Testing

5.6.1 Description

This is the crucial activity of the entire process: building a model that produces meaningful results. The success of the model is directly dependent on completion of the previous steps. With the input from the Analytics Problem Statement, the data set, and the IT system, the software can be deployed to run the models and evaluate the results. The data should be divided into training and validation sets and findings should be calibrated. All of the steps, decisions and results should ideally be documented, including limitations and constraints as well as the actual results (see Mandatory Appendix II-2).
The INFORMS approach anticipates iterative activities to achieve the result:

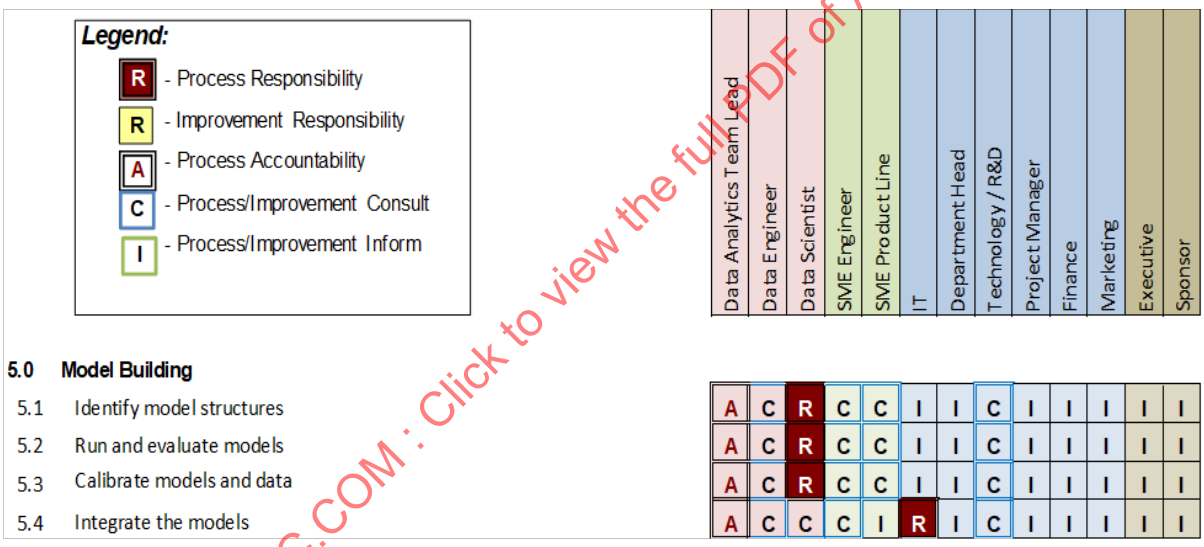
(a) Identify and build effective model structures to help solve the business problem
(b) Run and evaluate the models
(c) Calibrate models and data
(d) Integrate the models

Part of the validation of the model testing will require an interim review of the Analytics and Business Framing leadership. The objective of this phase is to find the model or models that best answers the question and is validated, reliable and deployable. It is reasonable to assume that rework will be required to develop the best model(s).

5.6.2 Team Member Roles

The Data Scientist continues in the lead role in collaboration with IT. The Data and SME teams continue to support the effort. The Technology/R&D leader will have a consultative role if new tools are required in hardware or software to accomplish the modeling, as shown in Table 5-5.

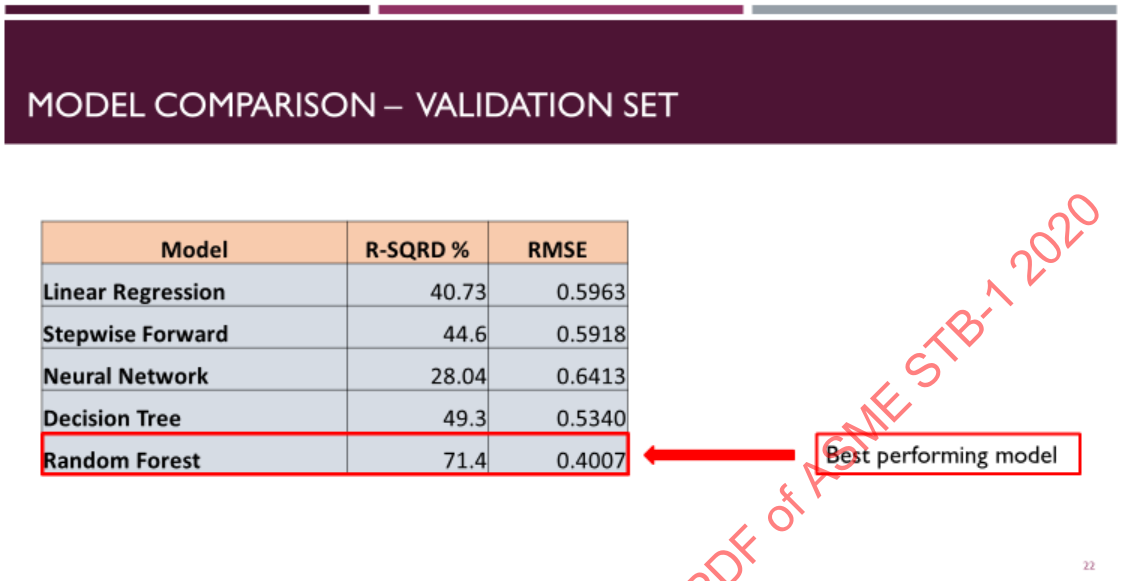
Table 5-5: Model Building and Testing RACI Chart



5.6.3 Example Business Challenge - Permian Basin Forecasting Model Continued

For this segment discussing the Example Business Challenge, the Data Scientist performed five different machine learning models to analyze the data and discover a potential answer to his hypothesis. The models were Multiple Linear Regression, Stepwise Forward Linear Regression, Neural Network, Decision Tree and a Random Forest. For each of the models, he analyzed a training set and a validation set of data. Comparing the models on maximizing R2 and minimizing Root Mean Squared Error (RMSE) for the validation set, he produced the results in Figure 5-4:

Figure 5-4: Model Comparison Results



Using these machine learning techniques, The Data Scientist confirmed his hypothesis. He developed several conclusions:

(a) Random Forest model with max depth = 20 was best model with R-Square = 71.4 % RMSE = 0.4007
(b) All models identified the statistically significant variables: Proppant amount / ft. and lateral length – which matches his hypothesis
(c) Best producing wells were longer laterals, higher fluid and proppant → hypothesis is correct
(d) The predictive model provides a more informed decision making and completion optimization
(e) Model enables E&P companies to quickly predict future well performance, ROI from the asset
(f) Model identifies underperforming assets for possible acquisition or re-completion
(g) Model allows reservoir engineers to predict future production based on completion strategy
(h) Using this model could save companies up to 4% of well cost

A notable result - his hypothesis estimated a reduction of 4% of well cost. At an average of $5M per well, this is a savings of $200,000. For a program of 100 wells, this represents $20M in savings.

5.7 Solution Deployment

5.7.1 Description

Once the models have been executed, tested and validated, they should ideally be deployed across the business enterprise to become useful. The data collected should now be converted into a stream that can be input to the model as needed to execute. The software or algorithms developed to produce the model are converted into a reusable Production Model to be used by the stakeholders in the enterprise. This stage will require more software development and systems for the user interfaces and report formats. Section 4.5 discusses this development of APIs.

5.7.2 Team Member Roles

This team is more distributed in levels of Responsibility to deploy the model. The Data Analytics Team Lead steps back in responsibility in collaboration with IT and the Department Head to build the interfaces and user interfaces. Support is required from the Data team, the SME team, the Finance and Marketing team, and the Project Manager to supply budgets and communicate the system solutions benefits, as shown in Table 5-6.

Table 5-6: Solution Deployment RACI Chart

Legend:

- R - Process Responsibility
- R - Improvement Responsibility
- A - Process Accountability
- C - Process/Improvement Consult
- I - Process/Improvement Inform

		Data Analytics Team Lead	Data Engineer	Data Scientist	SME Engineer	SME Product Line	IT	Department Head	Technology / R&D	Project Manager	Finance	Marketing	Executive	Sponsor
6.0	**Solution deployment**													
6.1	Perform business validation of the model	R	I	I	I	I	C	A	I	C	C	C	I	I
6.2	Deliver report with the findings	R	C	C	I	C	C	A	C	C	I	I	I	I
6.3	Create model, usability, and system requirements for production	C	C	C	C	C	R	A	C	C	I	I	I	C
6.4	Support deployment	C	I	I	C	C	A	R	I	I	I	I	I	C

5.7.3 Permian Basin Forecasting Model Continued

In finalizing the discussion of the Example Business Challenge, the Data Scientist addresses deployment. Since this was a capstone project and not a project embedded in a business enterprise, The Data Scientist made suggestions for deployment of this model for others to follow and left a roadmap for that deployment. He stated that:

(a) The Model will allow reservoir engineers to “plug and play” future wells and predict performance based on drilling and completion design
(b) Engineers can quantify how much each completion variable will impact production
(c) Real time knowledge of the impact of ineffective fracturing caused by operational failure
(d) Identify issues with underperforming wells

For an operator to deploy this model, they would need to write an API to take the results of the Random Forest model for each set of data entered. This API could display results in the form of a dashboard with forecasted trends and warnings of potential failures.

5.8 Model Maintenance and Recycle

5.8.1 Description

Once the model is successfully deployed, the work continues to maintain the model and confirm its utility. The business enterprise will have required outcomes and KPIs to measure the effectiveness of the model. The model will have to be periodically tested and recalibrated with the data sets as they change over time. The cost and benefit to the enterprise will be evaluated to confirm that the model continues to meet the business objectives (see Mandatory Appendix II-2).

INFORMS suggests the following activities:

(a) Document initial structure
(b) Track model quality
(c) Recalibrate and maintain the model
(d) Support training activities
(e) Evaluate the business benefit of the model over time

The model may live over several years and potentially decades. The model needs documentation to inform subsequent data and SME leadership to understand why and how the model was developed. Documentation should ideally include training on how to use the model and how to update/refresh if necessary. Business objectives change over time and the model may become obsolete and need replacement.

5.8.2 Team Member Roles

The Data Analytics Lead, SME Engineer, SME Product Line and Department Head are responsible for the support and maintenance of the model. This includes documentation and training. These activities require the support of the Data Team, IT, Technology/R&D, and the Project Manager, as shown in Table 5-7.

Table 5-7: Model Lifecycle RACI Chart

Legend:	Data Analytics Team Lead	Data Engineer	Data Scientist	SME Engineer	SME Product Line	IT	Department Head	Technology / R&D	Project Manager	Finance	Marketing	Executive	Sponsor
R - Process Responsibility													
R - Improvement Responsibility													
A - Process Accountability													
C - Process/Improvement Consult													
I - Process/Improvement Inform													
7.0 Model Lifecycle													
7.1 Document initial structure	R	C	C	C	C	C	A	C	C	I	I	I	I
7.2 Track model quality	C	C	C	R	C	C	A	C	C	I	I	I	I
7.3 Recalibrate and maintain the model	R	C	C	C	C	C	A	C	C	I	I	I	I
7.4 Support training activities	I	I	I	C	R	I	A	C	C	I	I	I	I
7.5 Evaluate the business benefits of the model over time	C	I	I	C	C	I	R	C	C	C	I	A	I

5.8.3 Example Business Challenge - Permian Basin Forecasting Model Concluded

Sustaining the Data Scientist' model will require actual use over time, re-validating the results and utility for the operator that chooses to deploy it.

The useful work of discovering the data, the significant variables in the data, and the use of a Random Forest should be documented along with any clean up and normalizing of the data that was required to make the model useful. The documentation can form the basis of training future users and updates as required to fine tune the model.

A practical follow-up to this model would be to search out different operators and offer the model as an aid for their drilling and completion activities. These operators can then add their data to the stream to validate that the model is universal in application.

5.9 The Business Solution

5.9.1 The Continuing Challenge

After the deployment of the model, the solution should ideally be monitored to respond to these basic questions:

(a) "Did our project answer the challenge and solve the problem as framed?"
- If the answer is no, then the process and model will be interrogated and improved to solve the framed problem.

(b) "Did the business challenge change during the data analytics project?"
- If the answer is yes, the project will likely be revisited and either cancelled or recycled.

(c) "Did the business challenge change after the model was deployed and operated?"
- If the answer is yes, then the team will likely revisit the business challenge and either stop the current deployment and/or launch a new project.

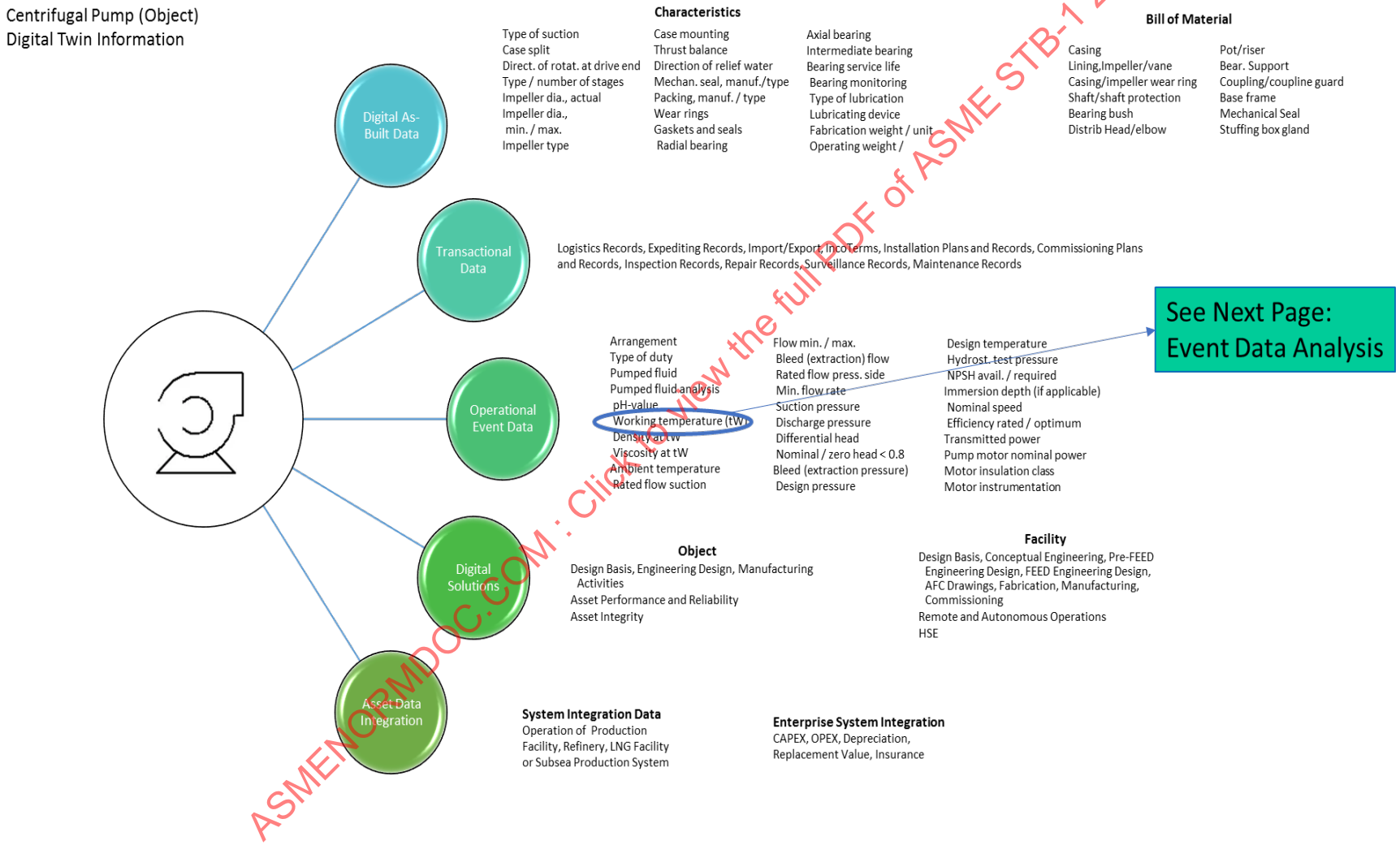
(d) "Is the model that solved the business challenge now obsolete?"
- This may provide the enterprise and the team the opportunity to update the model and/or launch a new project.

5.9.2 The Important Role of the Engineer

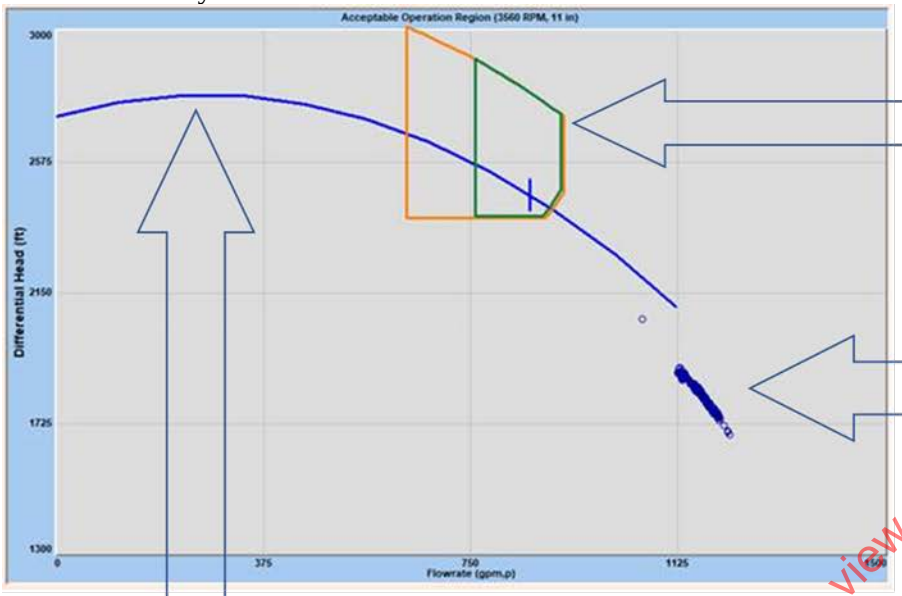
These business questions fall on a continuum of many process improvements that a healthy and thriving business executes. The role of the Engineer does not end with the data project or the deployment of the digital facility model. Change in digital technologies is to be embraced and explored if the business enterprise elects to use Big Data and digital facilities/twins to improve performance. Engineering disciplines will continue to broaden beyond designing physical systems and components. The engineers will join forces with data scientists to design and model digital systems that reflect the physical world. As businesses adopt digital facilities as the norm, engineers will need to understand and design the digital facility from the inception of a project and play a crucial role in the business challenges presented to the enterprise.

MANDATORY APPENDIX I: DATA CHARACTERIZATION CHART FOR OIL AND GAS

I-1 Digital Twin Representation Example



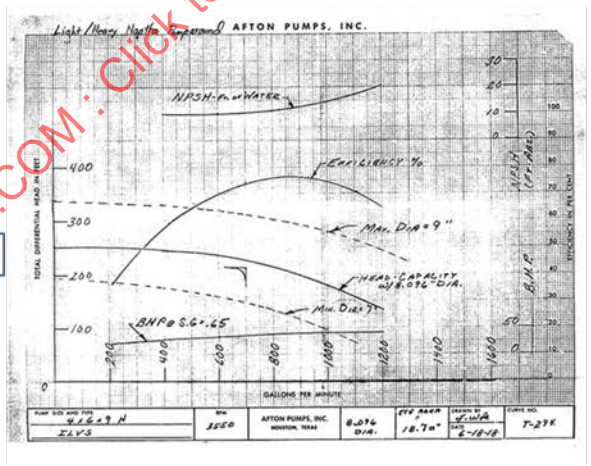
Event Data Analysis



Acceptable Operating Region
Determined by Engineering
Master Data

Pump Performance
Time Series Data

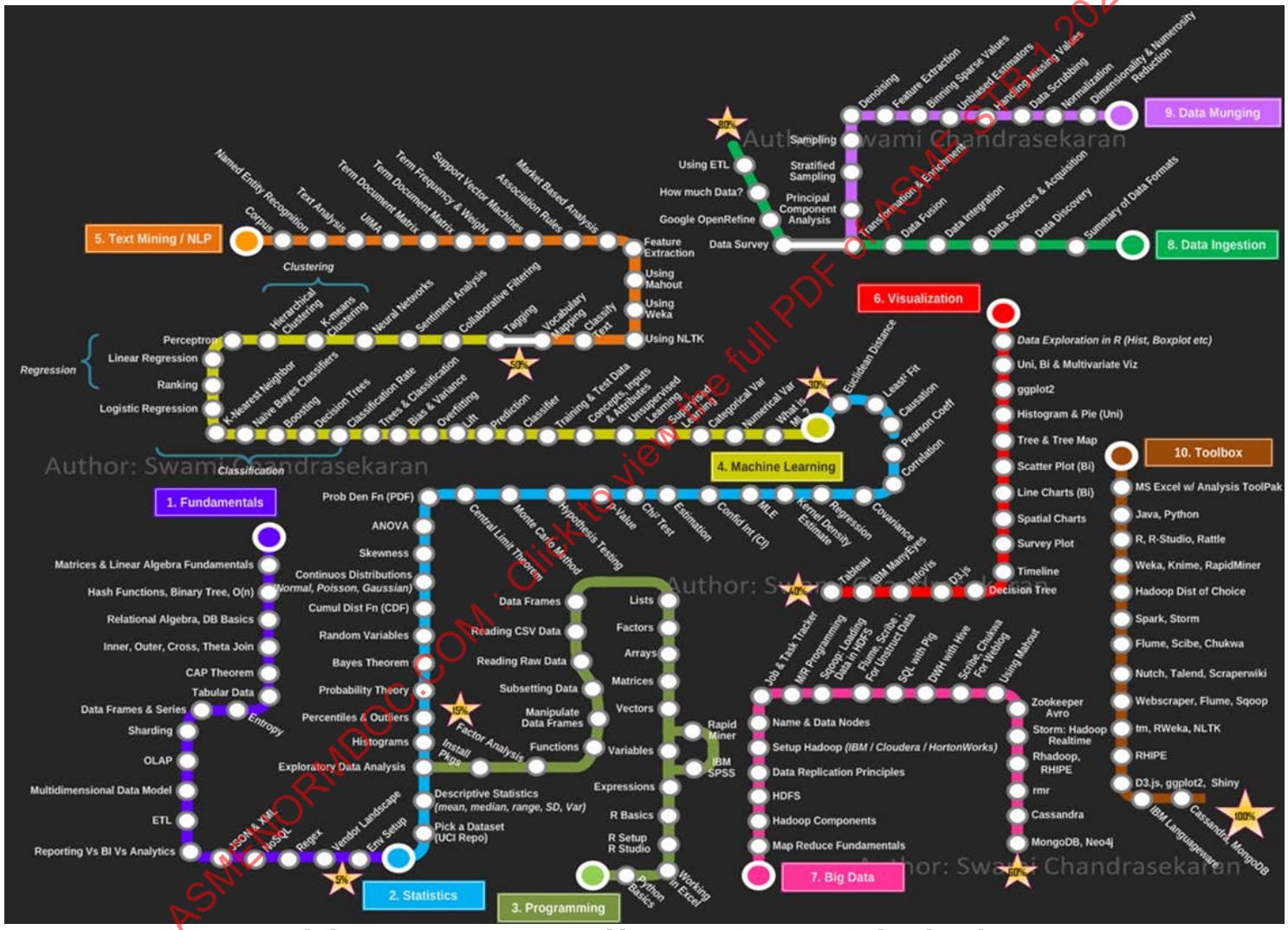
Pump Curve (Digital)
Master Data



Example of Historical
Pump Curve (Paper Document)
Master Data / Document

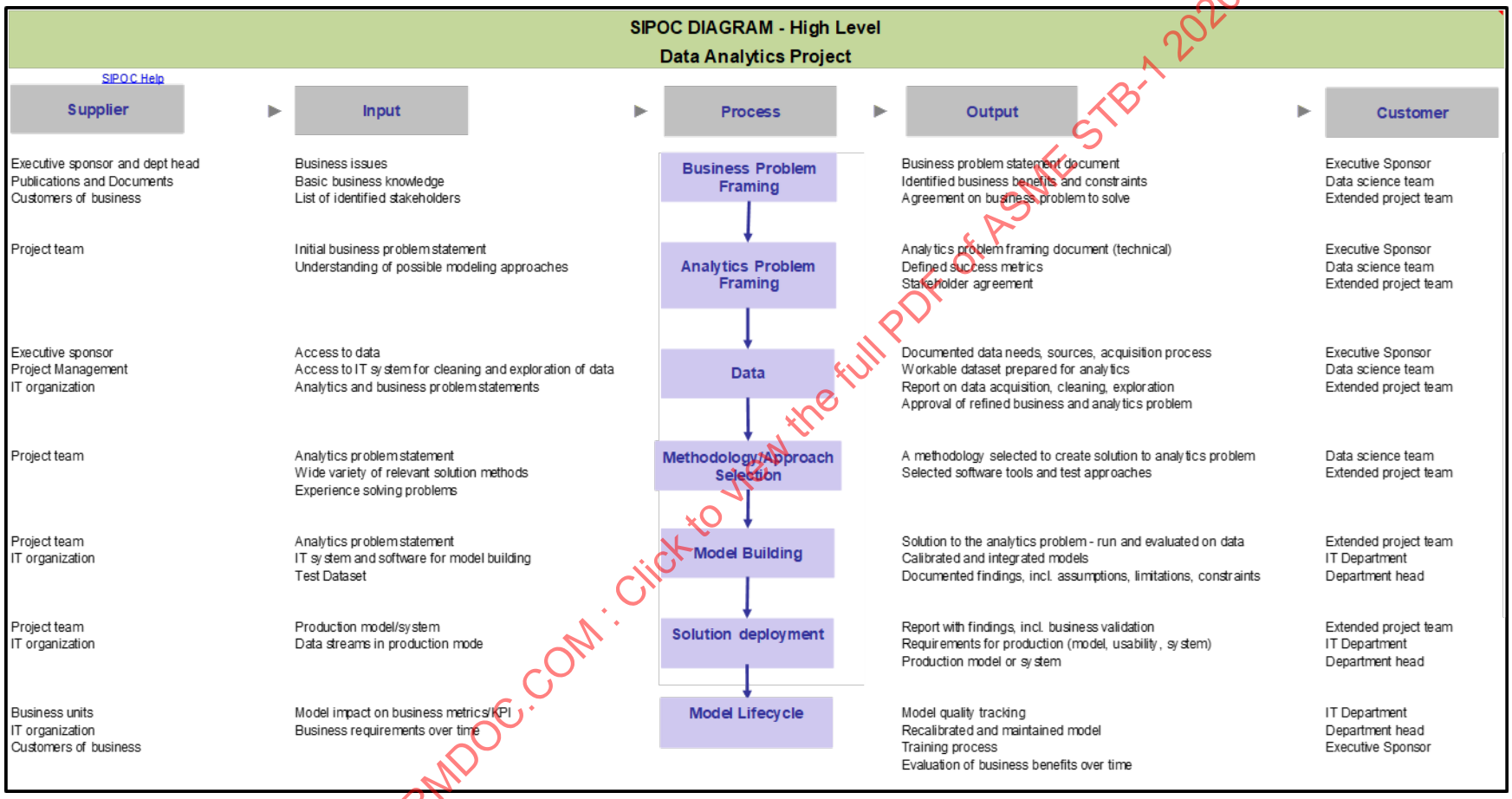
MANDATORY APPENDIX II

II-1 Detailed Data Journey



Detailed Data Journey, Reprinted by Permission Swami Chandrasekaran

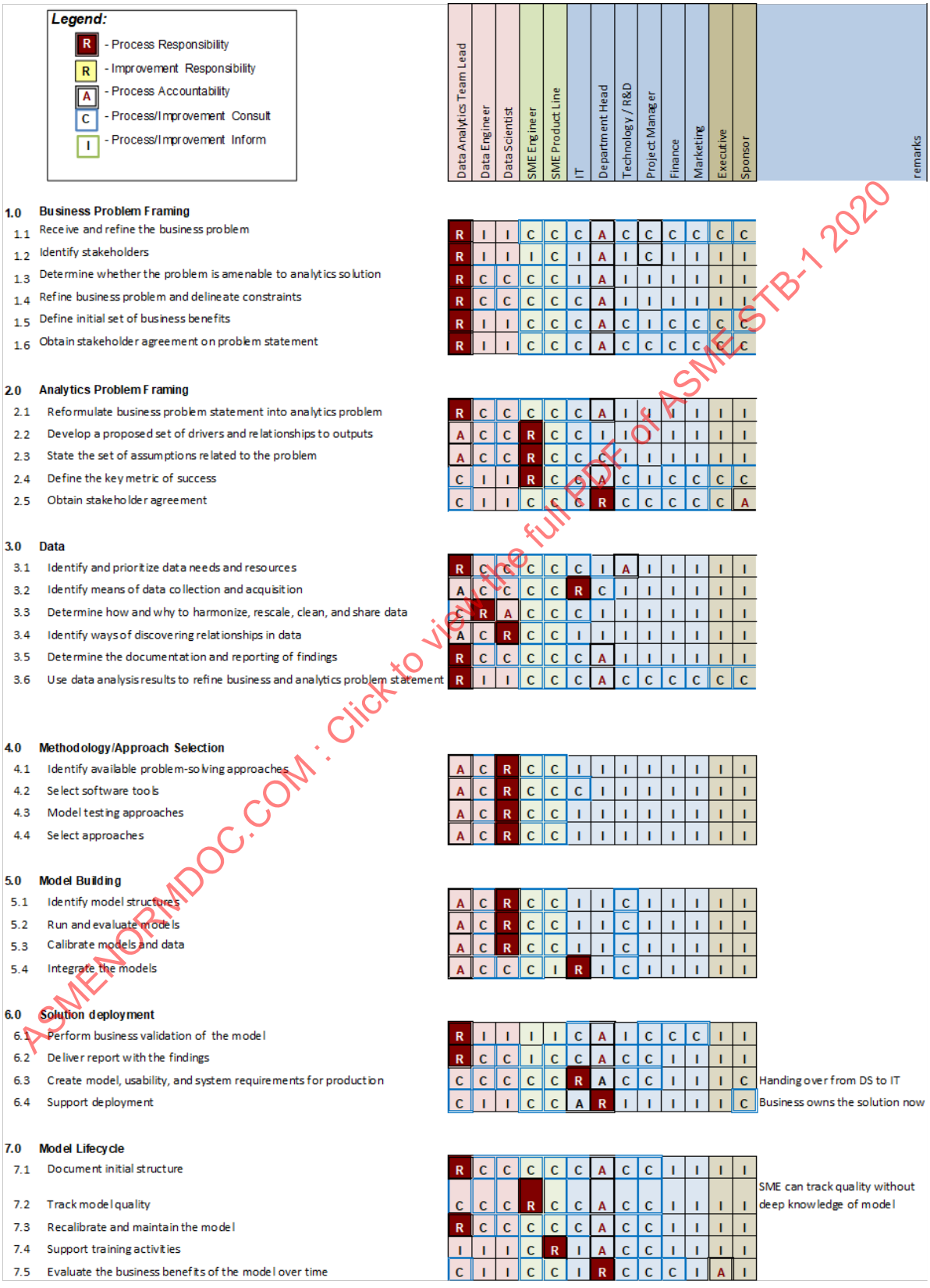
II-2 SIPOC Chart



II-3 Job Function Descriptions

Title	Job Description	General Role in Data Analytics Projects
Data Analytics Lead	Leader of a team of SMEs related to data collection, cleaning, modeling, deployment	Is the main connection in the team between the business challenge and how to structure data analytics processes
Data Engineer	Expert in collecting, managing, securing, and distributing data	Curates the data for the Data Analytics team and cleanses, prepares for analysis
Data Scientist	Expert in modeling and interpreting data	Under the supervision of the Data Analytics Lead, prepares and tests various analytical models and visualizations
SME Engineer	Expert in a specific field or process	Is the main source of information with respect to a business's product or service
SME Product Line	Expert in the particular product or service that requires business analytics	Guides the team with respect to the particulars of the product or service
IT	Leader of information systems, data management, data lakes, security	Supports all information requirements, software, and hardware systems
Department Head	Leader of the Department for Data Analytics to include Lead, Engineers and Analysts	Provides adequate resources to project team to execute project
Technology/ R&D	Domain experts in new product development	Interfaces with Data Analytics on integrating new products into digital frameworks
Project Manager	Leader of any Project within a company that requires resources and planning	With information and support of team, determines budgets, roles, responsibilities, delivery schedule
Finance	Leader of Financial systems	Provides financial information to team
Marketing	Responsible for messaging activities external to the company	Works with team to promote data solution to business problem external to the company
Executive	Business Leader with responsibility for P&L, Operations, Business Development	Has a business problem or challenge that wants to use data analytics to find a solution
Sponsor	Member of management team assigned to oversee a critical activity	Provides a connection from Executive to Project Team and solves existential problems

II-4 RACI Chart



Legend:

- R - Process Responsibility
- R - Improvement Responsibility
- A - Process Accountability
- C - Process/Improvement Consult
- I - Process/Improvement Inform

		Data Analytics Team Lead	Data Engineer	Data Scientist	SME Engineer	SME Product Line	IT	Department Head	Technology / R&D	Project Manager	Finance	Marketing	Executive	Sponsor	remarks
1.0	**Business Problem Framing**														
1.1	Receive and refine the business problem	R	I	I	C	C	C	A	C	C	C	C	C	C	
1.2	Identify stakeholders	R	I	I	I	C	I	A	I	C	I	I	I	I	
1.3	Determine whether the problem is amenable to analytics solution	R	C	C	C	C	I	A	I	I	I	I	I	I	
1.4	Refine business problem and delineate constraints	R	C	C	C	C	C	A	I	I	I	I	I	I	
1.5	Define initial set of business benefits	R	I	I	C	C	C	A	C	I	C	C	C	C	
1.6	Obtain stakeholder agreement on problem statement	R	I	I	C	C	C	A	C	C	C	C	C	C	
2.0	**Analytics Problem Framing**														
2.1	Reformulate business problem statement into analytics problem	R	C	C	C	C	C	A	I	I	I	I	I	I	
2.2	Develop a proposed set of drivers and relationships to outputs	A	C	C	R	C	C	I	I	C	I	I	I	I	
2.3	State the set of assumptions related to the problem	A	C	C	R	C	C	C	I	I	I	I	I	I	
2.4	Define the key metric of success	C	I	I	R	C	C	A	C	I	C	C	C	C	
2.5	Obtain stakeholder agreement	C	I	I	C	C	C	R	C	C	C	C	C	A	
3.0	**Data**														
3.1	Identify and prioritize data needs and resources	R	C	C	C	C	C	I	A	I	I	I	I	I	
3.2	Identify means of data collection and acquisition	A	C	C	C	C	R	C	I	I	I	I	I	I	
3.3	Determine how and why to harmonize, rescale, clean, and share data	C	R	A	C	C	C	I	I	I	I	I	I	I	
3.4	Identify ways of discovering relationships in data	A	C	R	C	C	I	I	I	I	I	I	I	I	
3.5	Determine the documentation and reporting of findings	R	C	C	C	C	C	A	I	I	I	I	I	I	
3.6	Use data analysis results to refine business and analytics problem statement	R	I	I	C	C	C	A	C	C	C	C	C	C	
4.0	**Methodology/Approach Selection**														
4.1	Identify available problem-solving approaches	A	C	R	C	C	I	I	I	I	I	I	I	I	
4.2	Select software tools	A	C	R	C	C	C	I	I	I	I	I	I	I	
4.3	Model testing approaches	A	C	R	C	C	I	I	I	I	I	I	I	I	
4.4	Select approaches	A	C	R	C	C	I	I	I	I	I	I	I	I	
5.0	**Model Building**														
5.1	Identify model structures	A	C	R	C	C	I	I	C	I	I	I	I	I	
5.2	Run and evaluate models	A	C	R	C	C	I	I	C	I	I	I	I	I	
5.3	Calibrate models and data	A	C	R	C	C	I	I	C	I	I	I	I	I	
5.4	Integrate the models	A	C	C	C	I	R	I	C	I	I	I	I	I	
6.0	**Solution deployment**														
6.1	Perform business validation of the model	R	I	I	I	I	C	A	I	C	C	C	I	I	
6.2	Deliver report with the findings	R	C	C	I	C	C	A	C	C	I	I	I	I	
6.3	Create model, usability, and system requirements for production	C	C	C	C	C	R	A	C	C	I	I	I	C	Handing over from DS to IT
6.4	Support deployment	C	I	I	C	C	A	R	I	I	I	I	I	C	Business owns the solution now
7.0	**Model Lifecycle**														
7.1	Document initial structure	R	C	C	C	C	C	A	C	C	I	I	I	I	
7.2	Track model quality	C	C	C	R	C	C	A	C	C	I	I	I	I	SME can track quality without deep knowledge of model
7.3	Recalibrate and maintain the model	R	C	C	C	C	C	A	C	C	I	I	I	I	
7.4	Support training activities	I	I	I	C	R	I	A	C	C	I	I	I	I	
7.5	Evaluate the business benefits of the model over time	C	I	I	C	C	I	R	C	C	C	I	A	I	

NONMANDATORY APPENDIX A: CASE STUDY

DATA DRIVEN PERMIAN BASIN PRODUCTION FORECASTING USING MACHINE LEARNING

Yaz Meridji March 27th , 2020
Project Coach: Eric Ziegel

EXECUTIVE SUMMARY

- Permian Basin has 180 Billion Barrels of technically recoverable resource
- Permian Basin has the most sought after acreage
- Completion, technology, area and operator drive oil production
- Random Forest is the most suited to identify key production drivers
- Both existing & future well performance can be predicted using this methodology
- Technique allows us to evaluate what-if scenarios (operational limitations)

2

BUSINESS IMPACT

- Use ML to identify key controllable well parameters that drive production
- Improve operational efficiency, reduce cost & enhance well deliverability
- Forecast productivity from existing & future wells
- Identify AOI (Area Of Interest) true economic potential
- Use the model findings to optimize field development plans
- Identify outliers & possible re-completion candidates
- Sweet spot mapping
- Identify key players in the basin & why are they successful ?

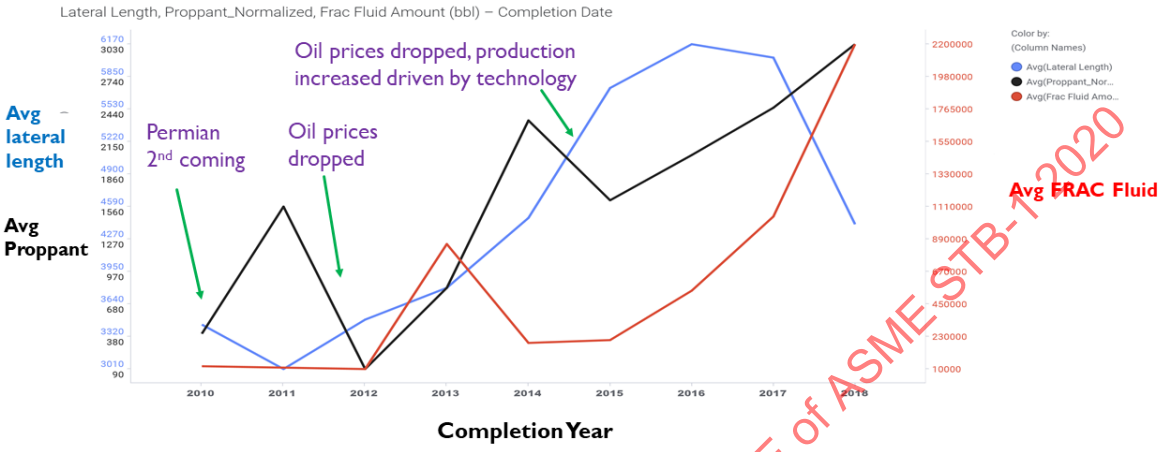
3

PERMIAN BASIN – SHALE REVOLUTION & GROWTH KEY DRIVERS

- In 2018 US became the #1 oil producer most contribution from Permian Basin
- Technology was key to improved well performance
- Permian has nearly 3,000 ft of good quality shales & vast aerial extent 75,000 square miles
- Improved economics
- Existing infrastructure
- Better recovery factors 10-15%

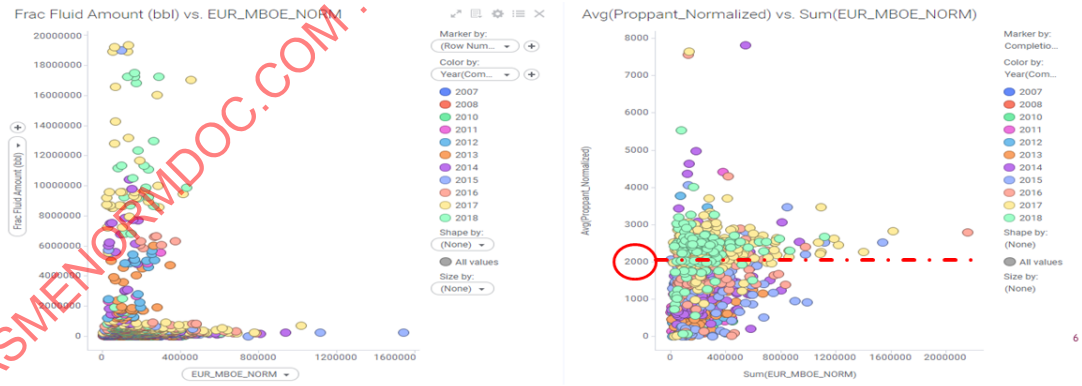
4

TECHNOLOGY DRIVING GROWTH IN THE PERMIAN



AFTER 2018 TECHNOLOGY NOT HELPING ANYMORE

- 20 Mil Barrels is the max number
- More fluid → not necessarily better wells
- More proppant → more oil "to a certain limit"
- 2000 LB/ ft is the industry standard



PROBLEM & HYPOTHESIS

- **Problem:**
- What is driving production in Permian Basin ?
- Is it Localized ? Why are certain counties better than others ?
- Operators / Basin analysis

- **Hypothesis:**
- Investigate the theory that lateral length & proppant amount are what drives production.

- **Process**
- Clean up data
- Integrate all completion & production data into the model.
- Confirm whether or not the assumption that lateral length and proppant are the two key drivers.

7

DATA SOURCE

- Obtained the data from Tomlinson Geophysical Services and will be using n=1887 wells in the Delaware basin
- Same vintage data all wells drilled & completed 2007-2018
- 14 dependent variables were used to predict EUR (Estimated Ultimate Recovery)
- Several missing data points were removed, some replaced with the mean values
- Used n=1512 wells post data clean up
- Focused on top two reservoirs: Wolfcamp & Bonespring (most prolific)
- All wells are in the Delaware Basin (West Permian Basin)

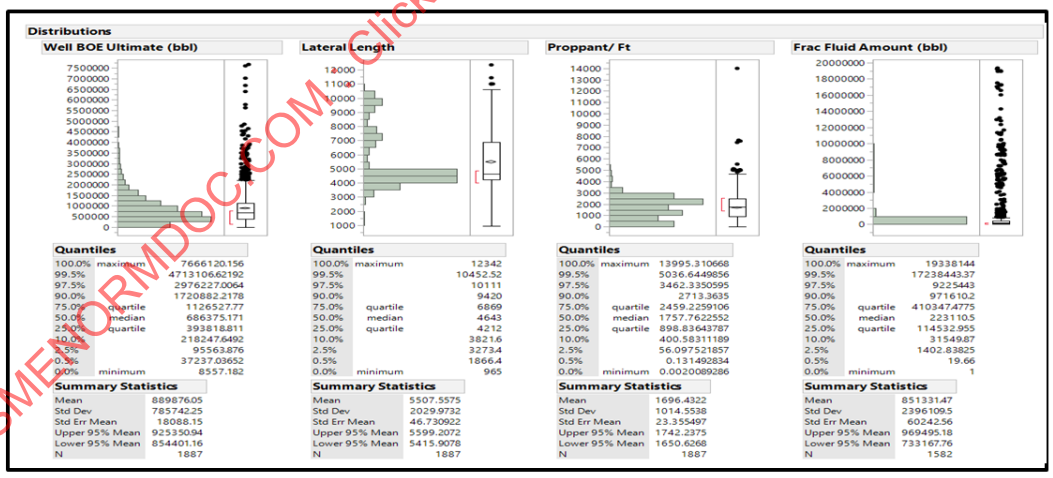
8

DATA DICTIONARY

Completion Variable	Definition
Drilling days	number of days to drill the well
Lateral length	well horizontal section length
Number of stages	total number of frac stages
frac fluid type	frac fluid type (gel, slick water..)
frac fluid amount (bbl)	amount of liquid pumped
frac fluid amount/ ft (bbl)	amount of liquid pumped / ft
proppant type	frac proppant type
proppant amount lbs	total proppant pumped
proppant amount/ ft (lbs)	total proppant pumped / ft
EUR / 1000 ft (BOE)	estimated ultimate recover from every 1000 ft of the lateral (normalized)
BOE max/ 1000 ft (BOE)	Barrels of oil equivalent / 100 ft of lateral (normalized)
Max treatment rate	Max surface pump rate
Max treatment pressure	Max surface treatment pressure
Treatment remarks	A description of the Frac job

9

HIGHLY SKEWED DISTRIBUTIONS- LOG TRANSFORM REQUIRED



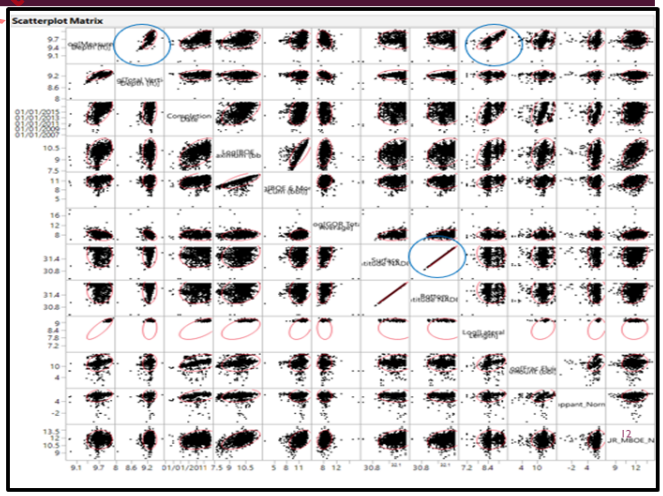
10

WHO'S INVOLVED ?

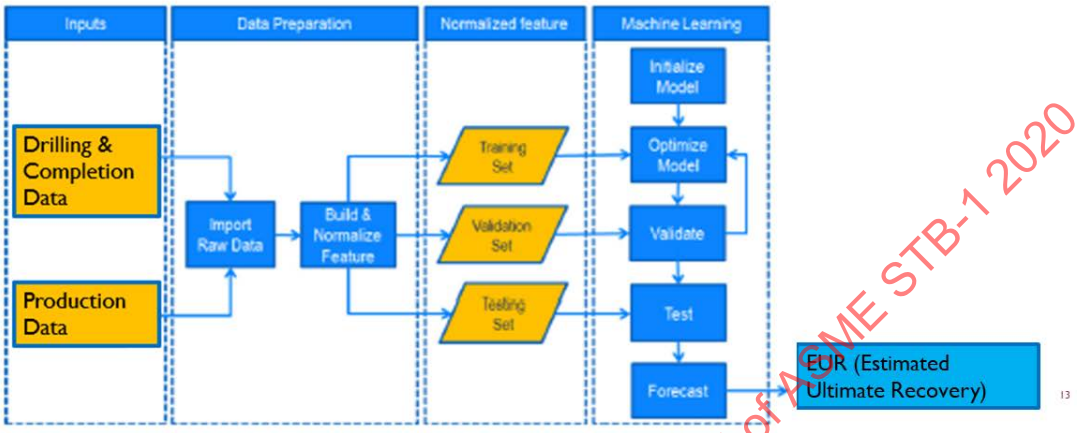
- **Reservoir engineer:** Provide the facts, information and knowledge necessary to control operations to obtain the maximum possible recovery from a reservoir at the least possible cost.
- **Completion engineer:** Accountable for managing financial controls, financial targets, and long-term completion strategic planning.
- **Geologist:** Gain a better understanding of multiple reservoir horizons which are proven or prospective within the asset area.
- **Data scientist(Me):** Use ML to identify key controllable well parameters that drive production, improve operational efficiency, reduce cost & enhance well deliverability

CORRELATION MATRIX

- Remove all correlated variables
- Removed all dates except Completion date
- Used Surface Lat & Bottom Lat (removed Longitude)
- Total Depth & Lateral length & TVD show correlation but were kept in the model (blue circles)



WORKFLOW



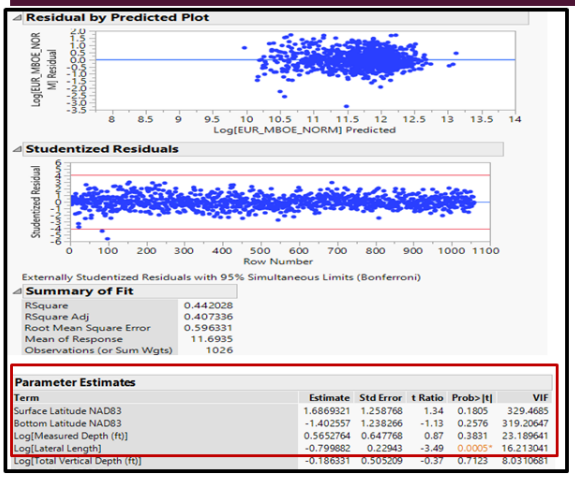
Modified from SPE-180984-MS

BUILDING THE MODELS

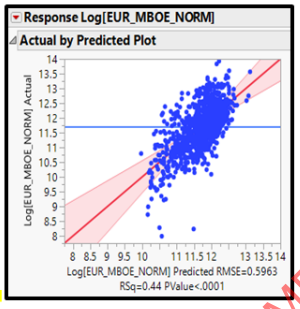
- Normalize parameters
- Log transform
- Split the data into training 70% n=1058 wells, validation 30% n=454 wells
- Use training set to build models & run RF, Decision Trees, Neural Networks & Stepwise Regressions
- Select best model based on validation set best R-square & lowest RMSE
- Ensure no over/ under fitting

14

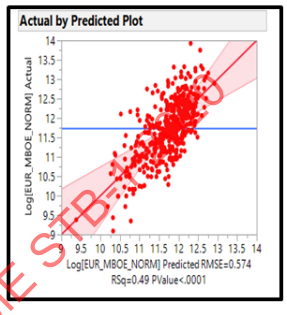
MODEL#1: LINEAR REGRESSION



Training



Validation



- Studentized residuals look random– few outliers
- VIF < 10 mostly, first four are higher by design

15

MODEL#1: CONCLUSIONS

Effect Tests

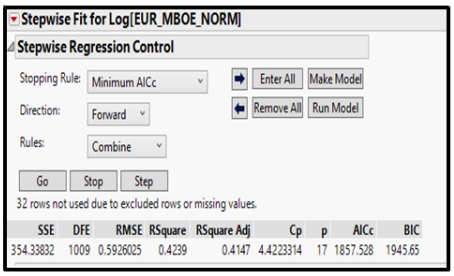
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Log[BOE Maximum (bbl)]	1	1	59.504251	167.2898	<.0001*
Current Operator	43	43	31.120308	2.0352	0.0001*
Log[Lateral Length]	1	1	4.322427	12.1549	0.0005*
Log[BOE 6 Month Cum (bbl)]	1	1	1.642328	4.6183	0.0319*
Log[Proppant_Normalized]	1	1	1.507066	4.2380	0.0398*
Well Formations	1	1	1.115127	3.1358	0.0769
County	5	5	2.990047	1.6816	0.1363
Surface Latitude NAD83	1	1	0.638673	1.7960	0.1805
Bottom Latitude NAD83	1	1	0.456234	1.2830	0.2576
Log[Measured Depth (ft)]	1	1	0.270806	0.7615	0.3831
Log[GOR Total Average]	1	1	0.149412	0.4202	0.5170
Completion Date	1	1	0.058647	0.1649	0.6848
Log[Total Vertical Depth (ft)]	1	1	0.048373	0.1360	0.7123
Log[Frac Fluid Amount (bbl)]	1	1	0.000609	0.0017	0.9670

- Significant variables: Max BOE (Barrels Oil Equivalent), operator, lateral length, BOE 6 months, proppant amount
- Lateral length & proppant amount can be controlled
- Few Outliers these are usually new exploratory wells where companies are trying new ideas

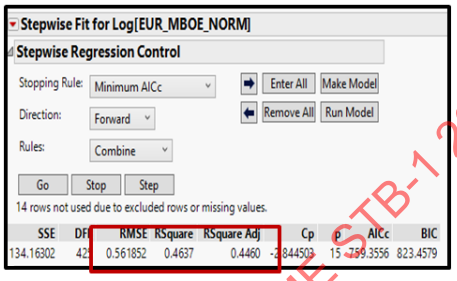
16

MODEL#2: STEPWISE FORWARD REGRESSION

Training



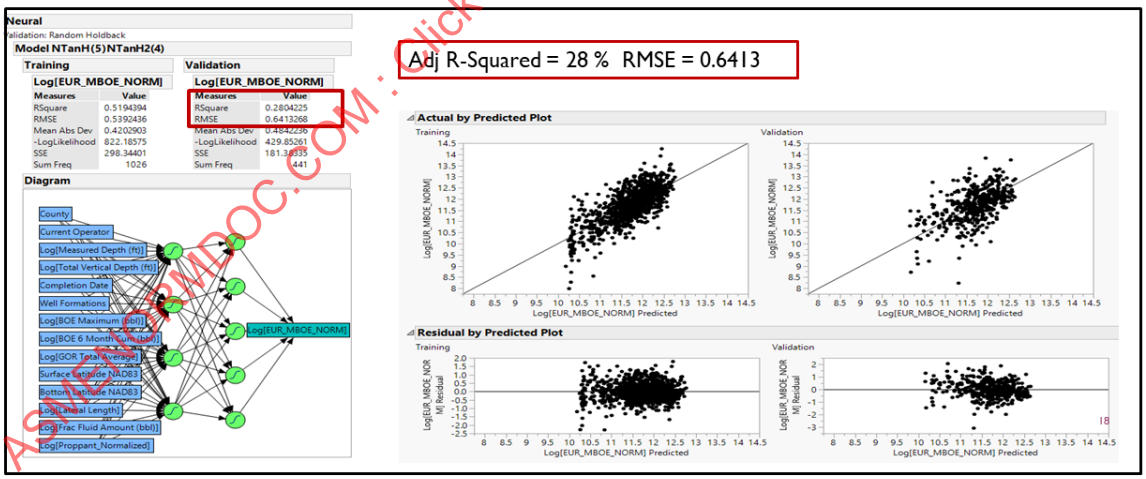
Validation



Validation Adj R-Squared = 44.6 % RMSE = 0.4637 - Based on the lowest AIC

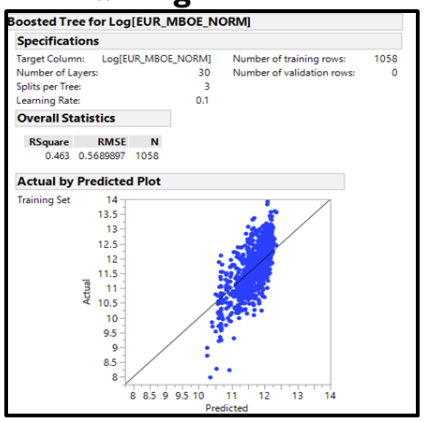
17

MODEL#3: NEURAL NETWORKS 2 HIDDEN LAYERS & 9 NEURONS

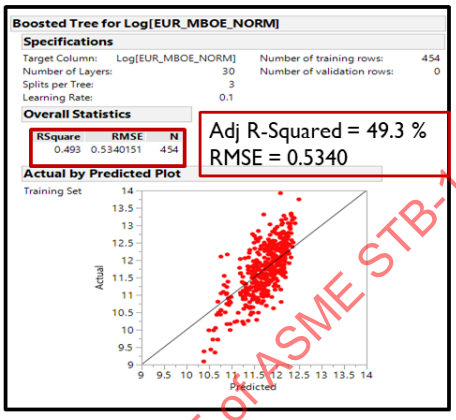


MODEL#4: DECISION TREES MAX DEPTH = 30

Training



Validation

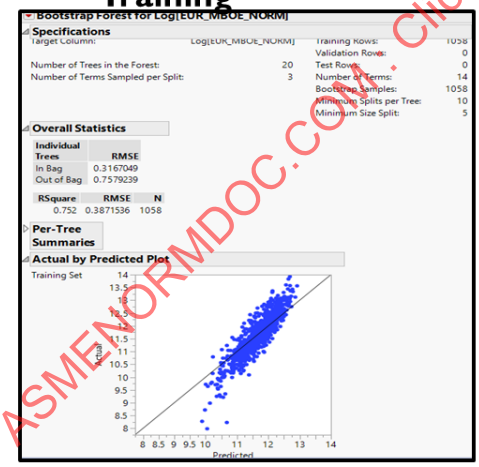


Adj R-Squared = 49.3 %
RMSE = 0.5340

19

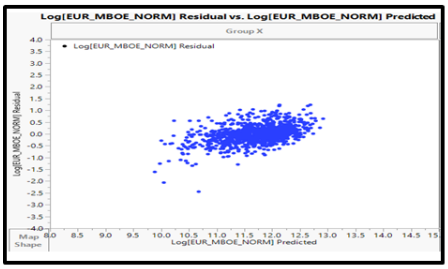
MODEL#5: BEST MODEL RANDOM FOREST MAX DEPTH = 20

Training



- R-Square = 75.2 %
- RMSE = 0.3167

Residuals



20