Chapter 2

Application of Data Science in the Sustainable Energy Industry

Chapter details

Chapter DOI:

https://doi.org/10.4322/978-65-86503-88-3.c02

Chapter suggested citation / reference style:

Marques, André L. F. (2022). "Application of Data Science in the Sustainable Energy Industry". *In* Jorge, Ariosto B., et al. (Eds.) *Uncertainty Modeling: Fundamental Concepts and Models*, Vol. III, UnB, Brasilia, DF, Brazil, pp. 19–53. Book series in Discrete Models, Inverse Methods, & Uncertainty Modeling in Structural Integrity.

P.S.: DOI may be included at the end of citation, for completeness.

Book details

Book: Uncertainty Modeling: Fundamental Concepts and Models

Edited by: Jorge, Ariosto B., Anflor, Carla T. M., Gomes, Guilherme F., & Carneiro, Sergio H. S.

Volume III of Book Series in:

Discrete Models, Inverse Methods, & Uncertainty Modeling in Structural Integrity

Published by: UnB City: Brasilia, DF, Brazil Year: 2022

DOI: https://doi.org/10.4322/978-65-86503-88-3

Application of Data Science in the Sustainable Energy Industry

André Luís Ferreira Marques

Naval & Nuclear Engineer - Rear Admiral (Naval Constructor) (retired) Brazilian Navy. Affiliations: University of São Paulo (USP), Electrical & Computer Engineering - PhD candidate. Paulista University (UNIP), Electrical Engineering (ICET) - Professor. Green Yellow of Brazil - Data Scientist. Brazil. E-mail: aferreiramarques@uol.com.br

Abstract

The application of Data Science (DS) in engineering fields is presented from a broad approach, taking the basics and historical motivation to apply statistical tools into practical needs, from Health Sciences to structural monitoring, the latter linked to the sustainable energy context, like solar, wind, geothermal and nuclear. Moreover, brief bullets about the DS techniques and models are also covered, most in the Machine Learning branch, but with some indications of Artificial Intelligence applications. Math techniques like Kringing, Support Vector Machines, Cross-Validation, Data Regularization, Surrogate Models, for instance, also make part of the text, where the practical issues are evaluated in a brief technical manner. The chapter covers the structural reliability and its DS application too.

Keywords: Data Science, Machine Learning, sustainable energy, uncertainty modeling, regularization techniques.

1 Introduction

The analysis of data has improved in several dimensions along the evolution of the digital systems in the last three decades, with more data volume been processed faster and faster every year. The Big Data environment has used tools in the analysis of phenomena and scenarios regarding the day routine of people, services, and countries. Data analysis has been a key instrument on the evolution of expansion campaigns of enterprises, marketing operations, public planning among others [1].

One example of how complex a data analysis can be, take the examples from war battles or campaigns, where data from different subjects come together, such as logistics and systems operational performance, to build up complex scenarios, to ease decision on the maximization of the combat capability. For instance, some data have origins from different time zones and production rates, and must be integrated into databanks for further processing, which normally considers the extraction, treatment and launching tasks. Nowadays, the military decision makers have data sources such as satellites, fixed

radar stations, mobile communication units, drones and even from cell phones, piling up amounts of data with different standards, languages, types, and volume. In the end, the analysis of data shall produce a valuable information to support the decision processes and culture in a short and long term as well [2, 3]

Earth environment sciences have been another technical subject with a high profit from the application of DS, manipulating data from satellites, fixed and mobile instruments aboard of planes, ships and voitures [4]. Equally important, people and animals can also provide data to be collected by research centers and universities linked to the environment surveys. The digital hardware has increased its reliability and portability with the development of electric materials and associated software. The data gathered go to databanks, data lakes etc., where can be managed to feed up analytical methods to compose data analysis about air pollution, water quality, land contamination over time, as few examples.

This chapter deals with some applications of Data Science (DS) into engineering fields, focusing current renewable energies options. The scope covers to show and understand the application of DS, throughout some examples, to address technical needs and outcomes, centered in data exploration, data classification and prediction, within dynamic space and time boundaries [5].

2 Development

2.1 Data Science: what's for?

In a few words, DS can be outlined by the combination of mathematics and computing disciplines and tools, including numeric calculus, analytic geometry, statistics, probability, to mention some of them, to handle a set of data, or a dataset, to provide some understanding, or even a better recognition, of characteristics and behaviors that might be inside the dataset. The whole pack of mathematical techniques have been integrated into computer schemes, accelerating the outcomes towards a more data centered conclusions. Figure 1 shows an overall view of the DS scope [6].

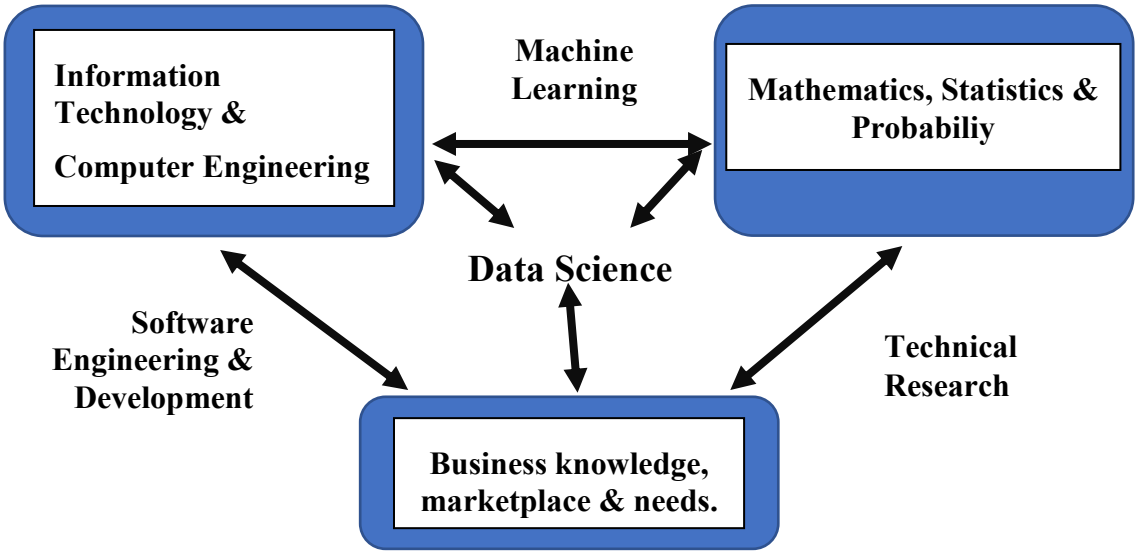

Figure 1: Scope of Data Science and other IT knowledge fields [6].

Imagine an enterprise received a large amount of data, from different sources, like sensors, human observations and so on. The enterprise has some points of interest to check and evaluate whether the data collected may present some insights or ideas never devised. Thus, the enterprise needs a set of criteria and procedures to go over the data and try to grab the outputs foreseen. Nonetheless, the insights may not be present in the way earlier planned and the data may require a cleanup or organization not so easy to detect throughout a simple visual inspection. That is the scenario where DS applies with different approaches.

Just organizing the data in a certain format can provide conclusions or deeper research directions, saving money and work hours from analysis not needed in the end, or more wisely to aim the effort into what really matters for the purpose of the enterprise or investigation. Take the example of the issues dealing with diseases, by checking the data about the occurrences in space and time. Mapping the cases and their characteristics can lead to know about a specific pattern and, thus, to preview future events or trends, which allow the people involved to get ready to dim the intensity of the consequences or even to stop the whole cause or the disease. As a remark, “mapping” means to organize the data into matrices, with dimensions related to the problem, which demands a deep knowledge of the details. That is the contribution of Florence Nightingale, during the Crimea War, which devised a way to organize data, from soldier casualties and deaths, using a bidimensional map, linking the data in terms of time [7]. This tool helped to know more about on how to decrease the harsh time of patient´s recovery and to avoid worse cases as well. Figure 2 shows her drawings, forecasting what it has been known today as “Data Analytics”.

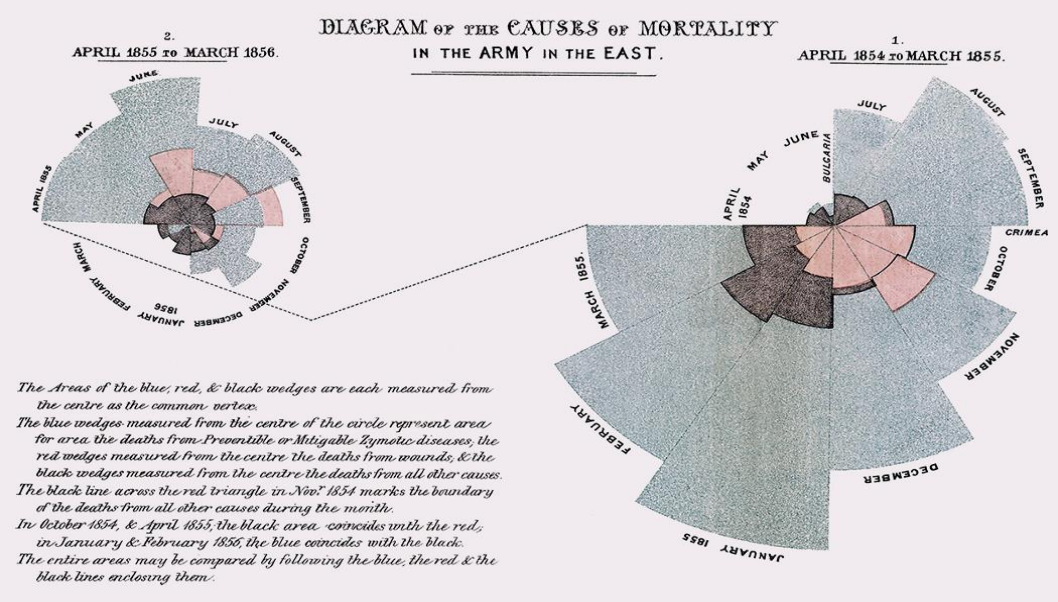

Figure 2: Florence Nightingale maps over soldier casualties and time [7]

With the development of the digital systems and the internet, the uses of math algorithms have spread to almost all areas of knowledge, improving the power and quality of observations, testing, and the accumulated experience [8]. Weather forecast, traffic control and disease control are one of the examples, managing large amounts of data, with different spectra and range of variables, only possible with digital computers and the algorithms tailored tunned. To interpret this huge set of data and build up a conclusion from it, the statistical algorithms have been used in large scale and faster, providing chances for insights not foreseen before. The higher the computing capability, the larger can be the dataset to be handle, and therefore more chances for better information to be checked and transmitted, and to allow more focused decisions to be made.

One fair example of the data handling linking odd parameters like cholera outbreaks and geographic locations, comes from medical doctor Snow, in London (1854), he was not a cartographer though [9]. In the end, but the association of the disease people data and where they used to live and the housing conditions, Dr. Snow was able to take conclusions and to control the situation. In fact, he used something closer today what is known as Data Analytics, well done. To dive into something closer to DS, the team of Dr. Snow could check a math model to classify the people according to the money income, age, water and food supply, education level, the intensity of the fever or other symptoms, for how long they had survived and others. Thus, the output of the model would be an index or number, grading the cholera case in terms of time and position, allowing some guessing how the disease would spread to other parts of London.

This text deals with the DS subject in a broad or horizontal view, with some applications, not covering coding examples, nor detailed or specific math procedures, steps, or models, by their equations. Rather, straight explanations and references will be used.

2.1.1 Brief history of DS

The DS history can be outlined as a progressive crescent timeline, with different increasing rates though. Taking the developments centered to handle information, the set up to data into tables, graphics and maps may be considered one of the first moves of DS, in parallel with description texts about phenomena and human actions. The main reason

for that also may come from religion and political issues, as the first motion springs to act in the societies. This feature may also be made without the scientific reason known as today and, therefore, reviewing historic pieces of plans, rule registers and description texts it is possible to identify the application basics of DS, when the use of rational steps to make conclusions based on collected data, mainly after the XVII and XVIII centuries in Europe.

It is worth noting the data processing period at that time can be taken as a couple of months or weeks. Hour and minutes could not be achieved due to the technology obstacles when dealing with the energy processing. The spread of electric hardware changed this frame after the middle of the XIX Century.

A half century ago almost, the paper "A Mathematical Theory of Communications" (1948) written by Claude Shannon, presented the term 'bit, coined by John W. Tukey in the prior year, starting the information process path of today, at the time the digital computer evolved. Almost 30 year later, the book "Concise Survey of Computer Methods" [9], written by Peter Naur, synthetized the data processing methods used in several practical subjects, and the definition of DS appeared as: "…the science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences". In a brief review, it resumes the main issues in DS. In 1989, the workshop "Knowledge Discovery in Databases (KDD)" was organized by Gregory P-Shapiro and can be seen as definitive move towards a stronger link between data analysis and statistics.

As private companies have handled more data related to marketing campaigns, the notion of a more rational use of math, statistics, and computer science to check what messages and relations may be present inside a data file, now collected more extensively as the communications webs enlarged exponentially. Thus, in 1996 two marks happened to clarify the DS scope: the book "From Data Mining to Knowledge Discovery in Databases" written by Usama Fayyad, Gregory P-Shapiro and Padhraic Smyth, and the conference "Data Science, classification, and related methods", held by the International Federation of Classification Societies (IFCS), both presented the first definitions and applications of all practices so far, focusing the classification problems mostly, which can be seen as the first natural field of DS. Leon Breiman pushed the envelope in 2001 with the book "Statistical Modelling: The Two Cultures", enhancing the importance of statistics application when analyzing data: 'If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools". This was said because he realized the use of statistics approaches had led to questionable conclusions and that the algorithmic modeling had been used outside statistics, dealing with different sizes of datasets [10].

The evolution of DS has all to do with the development of the digital technology because, now than ever, the capacity to perform hard calculations has been feasible in shorter time intervals (order of milliseconds on average, connecting billions of machines. Electronics and materials have increased their performance in an exponential scale, making data analysis more available to a broader set of researchers, enterprises, and governments. One example of this development is the 'smartphone', which is no longer only a 'mobile phone', but also a versatile 'data source' and 'data processor'.

Looking ahead, Figure 3 presents a distribution of subjects in terms of time and areas, regarding potential expansion applications, migrating from 'things' to thoughts. DS

works among all the sectors taking different signals transformed into binary numbers and checking what information can be draw from them. Currently, optical character processing has gained focus due to applications related to asset security and document automatic processing, saving money and time. A potential field of interest is the natural language processing (NLP) and the 'wearable hardware', coming closer and closer to human ordinary tasks.

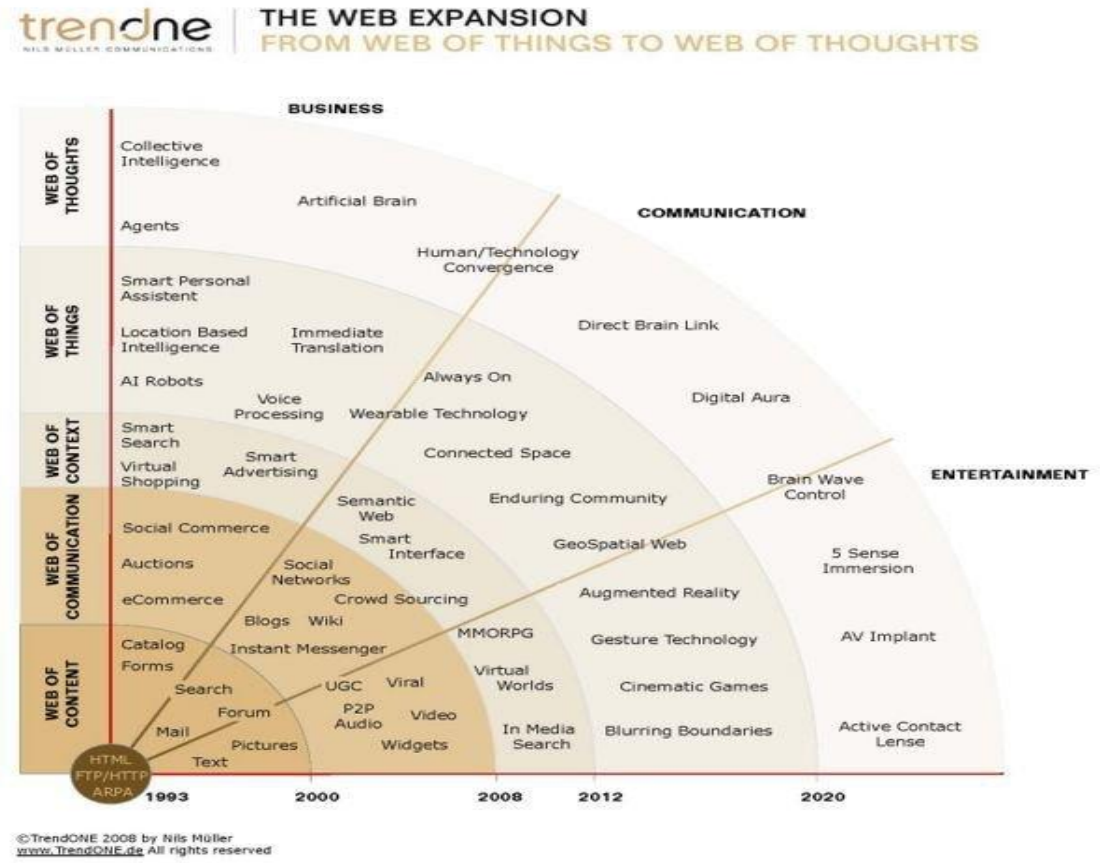

Figure 3: A short view of the web expansion and distribution [11]

In the last twenty years, journals, workshops, and thematic congress have been organized and the application of DS has focused the education alternatives and the market need fulfillment in many countries. As today, the DS has key presence in our routines, some not easy to be detected, and "Data Scientists" belong to the professionals needed for the XXI Century and beyond.

2.2 Sustainable & renewable energies: the main types and frameworks

In the last fifty years, the atmospheric phenomena have taken the world attention, due to the higher intensity of the climatic events, such as flooding, forest fires, ocean level increase, draughts span and others. One of the causes of this complex scenario is the emission of greenhouse gases, where Carbon Dioxide (CO2) and Methane (CH4) represents the most important ones for instance [12].

Several solutions have been under evaluation to dim the greenhouse gas emission, such as to reduce the use of fossil fuels, using less environment aggressive energy sources, like

the solar, wind and nuclear options. Each one has a trade-off, and their selection depends on economics, geopolitical and logistics issues. The recent evaluations using data grabbed from satellites, fixed instrument stations and even from people, lead to conclusions data driven, thinking on the balance of the Carbon generation and absorption, in terms of its presence in the atmosphere.

The planning of the ways to go to handle the Greenhouse gases require data management, because the wide source of gases sources and sinks, which may vary in different clusters of time, say long term (more than 10 years) or short term (less than 1 year), depending on the geographic region and human activity present. Among the data sources, the satellites deserve a special attention once they are not land intrusive and nor exposed to the atmospheric harsh events more frequent so far [13]. As a result, the satellites present data in different profiles in time, position, and substance of interest, with the development of new materials and data processing techniques.

With the whole set of data, the evaluation of the gas emission in a specific region can be checked against a potential source of sustainable energy, while dealing with the regional energy matrix. The consensus goes towards a more diverse matrix, combining manifold sources of energy, which may consider the overall Carbon emission reduction within a certain time frame. In this more diverse energy matrix, the sustainable options may be optimized taking their best outputs and operational profile. For instance, based on the data from satellite, the government planners can have elements to plan a smooth energy source transition, migrating from the sources burning oil towards options with low carbon signature, such as wind and solar units.

Figure 4 shows the expansion trend of sustainable energy sources, to provide more flexibility to the energy matrices. It is an example of Data Analytics, with an extract of many energy sources, linking colors to its concept. The first four are represented in grey tones, probably related to not renewable explanations. The other sources have colors such as light green, yellow and blue, in a psychological approach, which makes sense with the message to be sent. It is also clear their increasing rates, mainly after 2025, regardless geopolitical boundary conditions, which can vary in a different manner as considered when collecting and analyzing the data. The option biomass appears as a natural candidate with smaller challenges to be overcome. On the other hand, the option ‘wind-offshore’ can increase but the technical burdens will demand more resources than foreseen today, which may be checked later [14].

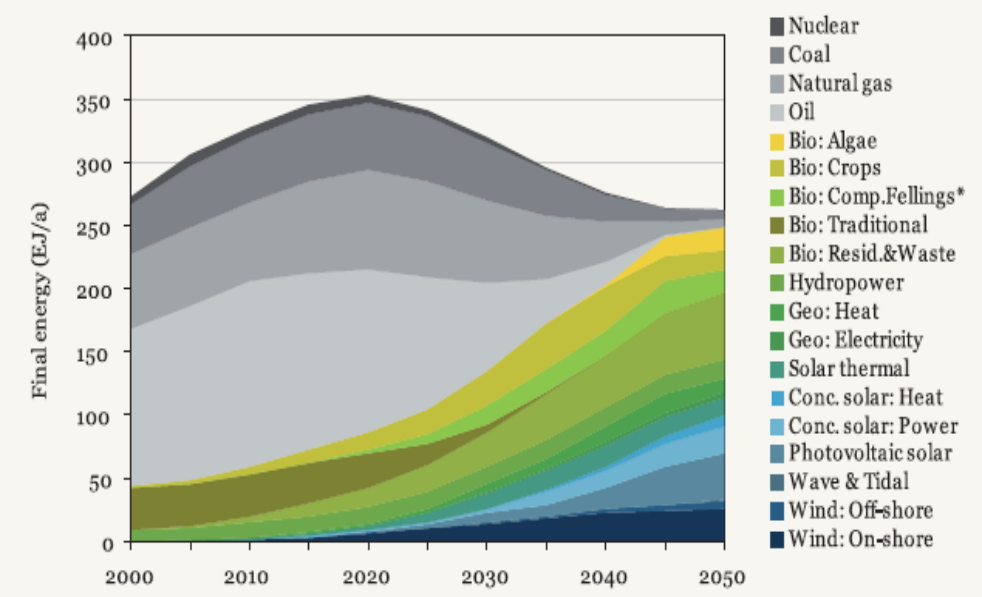

Figure 4: Sustainable energy trends

2.3 The Data Science cycle

In a DS project or experiment, there are steps that can be seen in Figure 5, in a loop cycle, with some internal feedback as well.

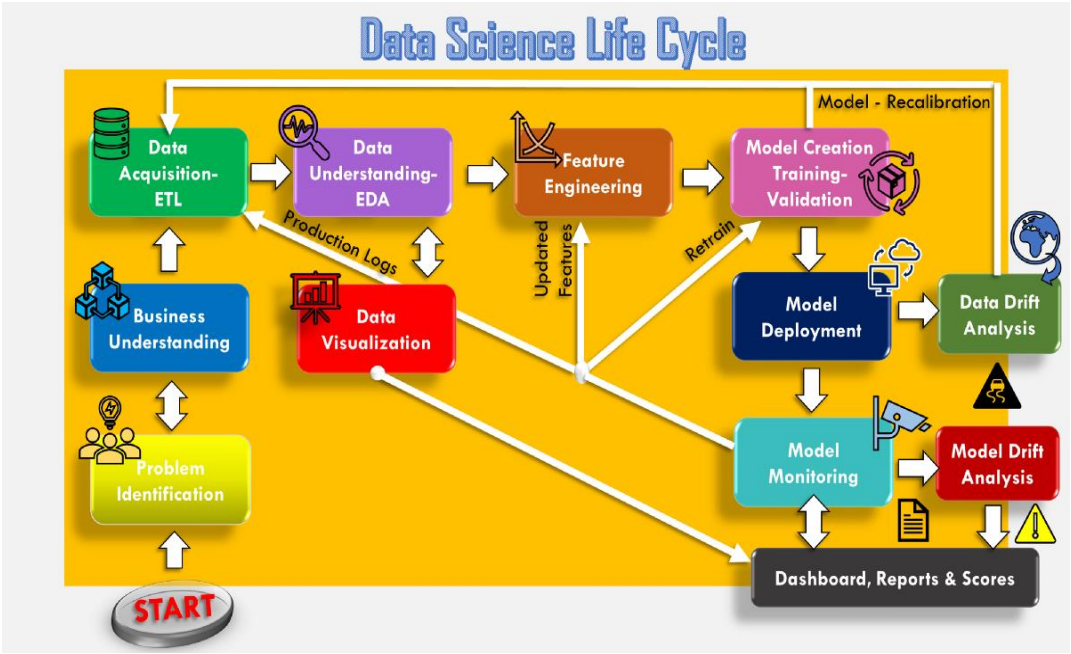

Figure 5: The general Data Science cycle [15]

As seen in other knowledge fields, the definition of the problem is the beginning of the cycle, after all it defines the needs to be fulfilled and the results to be reached. Make note the problem definition represents the hardest part of the cycle, in most of the cases, due to the unknow variables and details not considered initially. This first step has a strong

link with the business understanding because the effort will be done to enlarge or sustain business boundaries, orbiting the need to improve profit or deal conditions in long terms.

The second step into the data field is the data acquisition, with 'ETL' operations: extraction, treatment and launching, which will be explained as follows. Extractions deals with to identify the data sources: sensors, collectors, documents, transactions logs and others. Each source may have a proper standard to handle the data: word length, number format, sampling frequency, time scale and so on. Inside the jungle of details and fashion, the data treatment will take care of the standard needed for the data project, cleaning and even reshaping the data set: math transformations, null or incoherent data drop, and data scale adjustment, to mention some of them. In the end, the data treated will represent the useful available set to be worked, and once so, the data can be transmitted or launched to the processing area. The documentation of this treated data receives a key attention because it has details and definitions, which will be key along the data project. The documentation receives the title of 'Metadata', combining explanations such as: data origins; how they were gathered; references; time span; accuracy; analytical methods used to validate them; name of persons involved and how to contact them; among others.

Figure 6 shows a short view of the ETL process, and almost not further word is needed to understand its message.

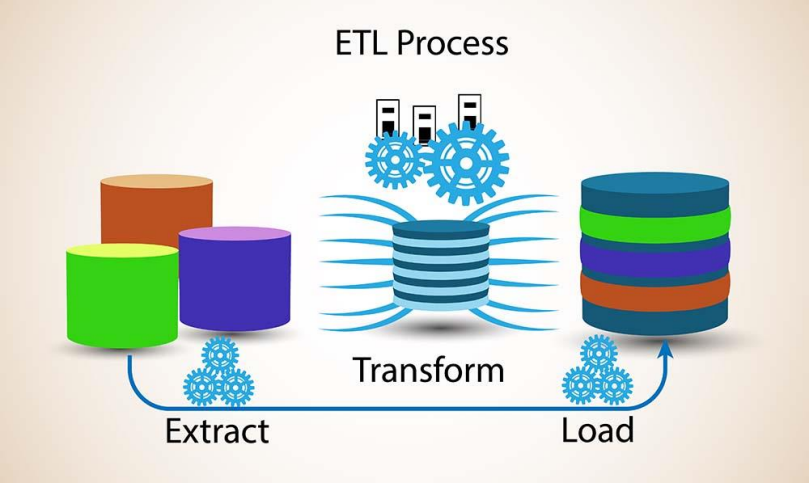

Figure 6: the ETL process in pictures [16]

The next step of the DS cycle is the exploratory data analysis, where the overall boundaries will be assessed, or in other words: how the data is spread, according to the main parameters or variables; which are the maximum and minimum values found; what kind of correlation can be detect at a first glance, although there is no guarantee the correlation makes physical or practical sense; what kind of initial clusters can be spotted; what area the statistics found: mean, standard deviation, mode, median, etc. This exploratory data analysis helps to evaluate whether the data gotten will make part of the solution of the problem to be solved, or whether the data can provide some initial insights. Several tools contribute for this exploratory analysis, and most of them deal with graphics and tables, touching what is also called 'Data Analytics" nowadays.

Following, the feature engineering will provide an overview of what parameters matter most, which saves times and resources in the next steps. Feature engineering also employs a set of math tools to understand how significant a certain parameter can be to explain the

data related. Equally important, the feature engineering presents the order of priority of the parameters in that explanation, which indicates a scaling of value linked to each parameter. Sometimes, to save computing resources, taking the feature engineering order of priority will ease any further work. For instance, in a certain job, the feature engineering check every one of the twenty parameters involved, and issued that only four can explain the data with more than 70% of importance. Therefore, the project may consider just these four parameters and drop the rest, saving time and money in the end.

Figure 7 shows some the overall wrap up of the feature engineering, when dealing with a set of variables or parameters of a job analysis problem, where the data about the candidate matters.

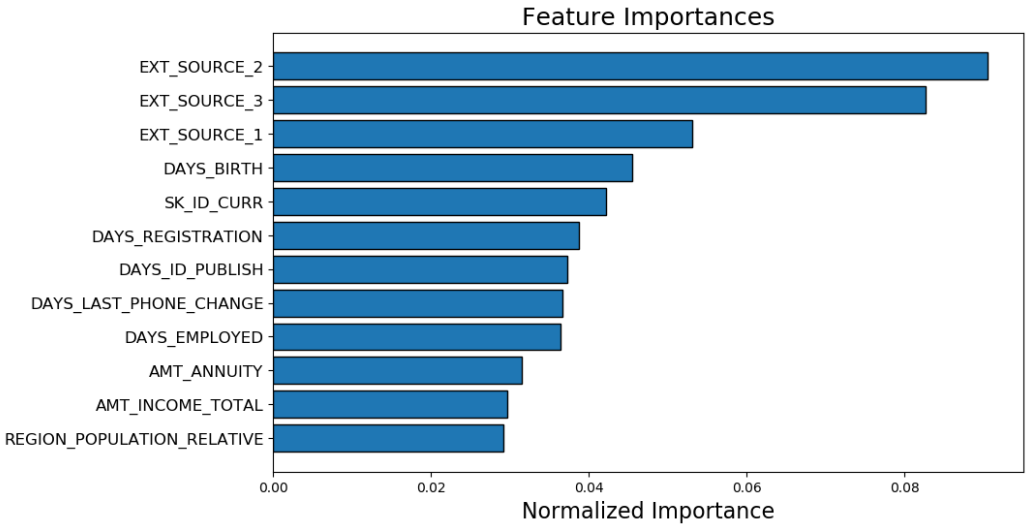

Figure 7: feature engineering relative numbers [17]

The model creation can be seen at the core of the DS cycle because it is the step where most of the math models are choose and set to aid the solution of the initial problem. Using the prior steps, the first idea of which methods may apply normally appears easily, or the ones not to be considered either. If the available data shows a linear correlation, this might be an indication that the Linear Regression approach can make some sense. On the other hand, if the exploratory data analysis does not show this kind of alignment, other models fit better, such as the Logistic Regression or the methods based on Decision Trees, or even a combination of other methos, normally called Ensemble.

Figure 4 shows the suggestion of nine important models for this step, where SVM stands for supported vector machines, KNN for k-nearest neighbors.

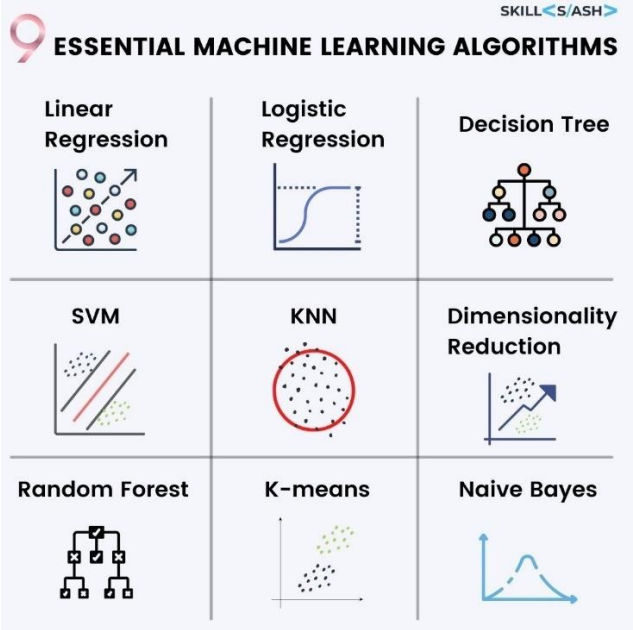

Figure 8: nine first models to be considered in DS [18]

These models are the simplest and the first choice in terms of cost and computing time, although the accuracy and performance may vary. Surprisingly, in some problems, they have a better performance when compared to more complex options as the ensemble ones. Here, the text indicates to get into them for those not so familiar with the DS subjects. Equally important, data verification and validation can be carried out during this phase, and they have methods and strategies that can be seen in Reference [19].

The term 'model deployment' in DS means the use of the developed model (classification, prediction etc.) with a new set of data to produce the insights, conclusions and decisions foreseen. In synthesis, it is the reason DS is for. A model can be defined as an equation and its coefficients, or a set of rules that produce the result with the input data. The deployment may consider an interactive process, handling the data pipeline in a continuous way, which demands to reshape the model based on the data time variations. In a first interaction, the best model could be the Random Forest, but as data is input sequentially, in a second round the best model could be then the Adaptative Boos (AdaBoost). Thus, the deploy will require a dynamic tunning, which quite frequent in the credit card and bank operations. Reference [20] has details of the use of the Predict Model Markup Language (PMML), which allows to build up the predict model to be deployed to compliant applications, considering 'cloud' options as well.

Monitoring makes part of all quality or action control step and deserves a special attention to achieve what is planned. After all, the use of real data into the models will pay back the investment, if the goals fulfill or overfill the expectancies. The only way to reach this conclusion comes from the model monitoring. The classical way of model monitoring is the performance metrics surveillance, during the software production phase. In some cases, it will be necessary to re-run the model, tunning it again, due to the new input data profile. If the performance decays more than a certain tolerance, then a whole process of re-training, re-defining and re-tunning will be carried out, and will be sent back to the DS cycle.

Reference [21] makes a clear resume of the model monitoring, which is near to the end of the DS cycle, but it can also be set anywhere in the middle, because it deals with the data presentation in two or more dimensions, easing to foresee any relation between data. It has a close relation with the prior step, model monitoring, also sharing some of the technical features: graph or trend line tools, classification plots etc. It may be part of the output package of application with DS.

The data collection starts the whole work and followed by the data treatment, when most of the mishaps and errors present in the dataset can be identified and corrected.

2.3.1 Uncertainty modeling in DS

In logistic chains, with several connections, it seems quite hard to predict all the parameters with a stunning precision and with no errors along the time. For instance, production plans considering a market expansion forecast demand to know or to grab data from different stages. This action will provide data to check each logistic step seems coherent. For instance, linear production rates do not stay for a long period, because outside factors change the logistic parameters in different frequencies and intensities. Therefore, the production rate needs to vary in time to avoid material shortage or overproduction, affecting the business plan.

Many sources indicate two benefits while attacking the data uncertainty: room for improvement and transparency. The more knowledge about the input data, the better can be the entire DS cycle, leaving space to drop data wisely, saving computing resources. Thus, treating the uncertainty remains in the field of 'additional improvement' way. More related to the current corporate practices, transparency is an advantage to the client, who will have a clear idea of the boundaries of his data and what may come about from the DS cycle and models. The use of DS models in the Health and Medicine applications demands transparency to aid doctors and researchers to run into acceptable conclusions.

Some authors define the uncertainty as the work with incomplete or incorrect data or even information. The uncertainty management can be done with math tools, such as probability and the Fourier Transforms, catching the time variance more accurately than visual observation or linear techniques. The three main sources of uncertainties can be nominated as:

- Imperfect or incomplete models, due to lack of experience or business knowledge.
- Noise, because the associated data without a reasonable explanation.
- incomplete data coverage, restricting the work field and bounding it wrongly.

There are two classical ways to cluster the uncertainty: aleatory or epistemic. The first one deals with the inherent noise associated with the data or environment where the data is captured. The second one aims the uncertainty of the model parameters, or how unable a model can the data relations. As seen in human activities, further training tasks can dim this uncertainty, according to the move towards the test phase. In the opposite, no additional training will help to reduce the aleatory uncertainty, only better measuring clearance will have a chance to reduce it. Suppose a system has cranky sensors but the model can explain more than 60% of the data. This situation will have more aleatory uncertainty than epistemic, and, by chance, has allowed this 60% score. While changing with better sensors, the aleatory branch can be reduced and now, the epistemic will be suitable to try the more training tasks, for instance, to reach a fairer result.

In a broad view, the application of probability and statistics aid to quantify the expected value and its variability, based on the data domain. Although both math fields were not devised for data variation strictly, they can be adjusted to detect data variance and its implications. Naïve-Bayes models consider the use of probability and account for the possible variations in a set of data, learnt from a train-test split. Reference [22] shows some of the math techniques bound for this issue, within the scope of Big Data.

Probability theories take the randomness and generally handle the statistical characteristics of the input data. Within this field, the Bayesian theory considers the interpretation of the probability, based on prior knowledge, which depends on the user background, and taken as an expression of a rational degrees of belief about uncertain values. Other tool is the "Belief function theory", which aggregates imperfect data through an information fusion process. In addition, when dealing with classification tasks, the Classification entropy measures ambiguity among classes throughout the index of confidence, which varies between 0 and 1, and its values closer to 0 point out complete classification in only a single class. In the opposite, values closer to one indicate links among several classes.

The concepts related to the Fuzzy theory are also used, to measure uncertainty in classes, notably in natural language examples. Fuzzy logic then handles the uncertainty associated with human perception by creating a reasoning mechanism, taking a more human or subjective approach. Other tool focuses the Shannon's entropy, which quantifies the amount of information in a variable, coming from the application of theory of communication and transmission of information. It provides a method of information quantification when it is not feasible to quantify weights. To model inaccurate or incomplete data, the Probabilistic theory and Shannon's entropy are often employed. For ambiguous data, due its nature, the Fuzzy theory is more recommended.

Another example is the Monte Carlo dropout technique, which is linked to Artificial Intelligence networks. In short words, this method drops some intermediate layers/neurons to produce a model away from overfitting, in a model regularization fashion. The dropout layers are taken away randomly after the training phase. For the test phase, the dropout layers will be available and, to reduce the uncertainty, this procedure shall be used more than three times, dropping out the layers randomly at each time, while predicting the results. Thus, the prediction will be tunned with less uncertainty as much as the process is repeated.

Now, think other way around, instead repeating the same procedure, even with a randomly process in the middle, or aleatory steps, take multiple process in parallel, doing the quite same tasks, at the same time. When taking multiple choices, the process has a label of 'ensemble', but each choice taking some layers randomly. This feature also produces different model parameters trying to reduce the uncertainty. As a general recommendation, take more than 3 different choices to work in parallel, which is equivalent of the dropout of at least 3 different intermediate layers of the Monte Carlo method above.

2.3.2 Regularization techniques

In many applications of DS, a common trouble has the designation of 'overfitting', which means the classification or prediction model tries to fit all the data, no matter it is an outlier, or a mistake not dropped out after the exploratory data analysis (EDA). The

aftermath of the overfitting is a model unable to predict or classify extra intake data, showing poor efficiency. This situation happens with a mathematical model that cannot capture the underlying data structure, taken more parameters that can be explained by the data itself. On the other hand, underfitting means a math model lacking parameters to explain all the data. It is quite easy to be detected when using a lower order model trying to fit a data distribution with a higher order of complexity, such as a second order model (i.e., parabolic) trying to shape a third or higher order data profile (i.e. hyperbolic). The prediction performance with both overfitting or underfitting is poor in the end, and these situations need to be skipped.

It is not advisable to operate with an overfitted model. The Figure 9 presents the comparison of the training and validation phases of a prediction model. It comes from the development of a neural network where the number of epochs is represented in the horizontal axis. The math accuracy is shown in the vertical axel.

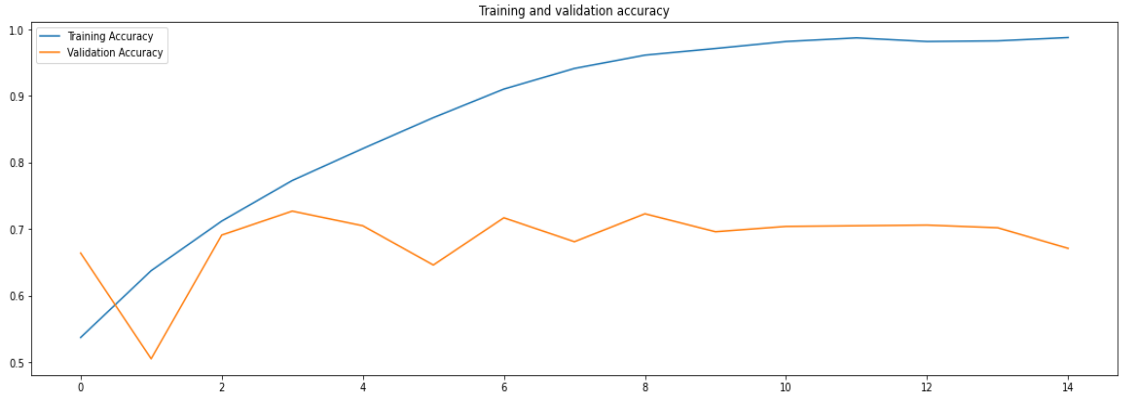

Figure 9: Comparison of model accuracy for the training and validation phases - I

When the accuracy of the training phase is higher than in the validation stage, it indicates the overfitting occurrence. The more trained is the model, the higher its accuracy. Nonetheless, the validation accuracy does not increase as the number of epochs increase, allowing the conclusion the model is not recommended. Depending on the input data and the split of training and validation, the suitable prediction model will show a different pattern. The best predict model shall have a better validation accuracy than in the training phase, because working like that shows to be prepared to produce fair prediction with data not known yet or fresh data. The expected shape said above is presented by figure 10, and regard the upward rate of the validation phase, as the accuracy increases.

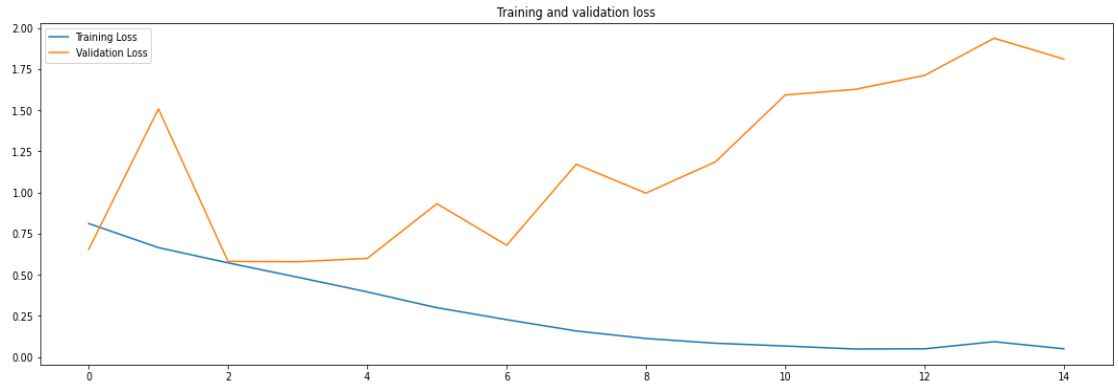

Figure 10: comparison of model accuracy for the training and validation phases -- II

Regularization is a set of techniques to avoid the overfitting, using a compensation function that penalizes the parameters responsible for the fluctuations to fit all the data, without any key explanation. Among these techniques, the 'least absolute shrinkage and selection operator', or Lasso – L1, the Ridge – L2 and a combination of both or Elastic net are the most famous used, three options in the end. The Lasso model operates into the sum of the module of the parameters and a penalty coefficient to take the sum towards 0. The parameters without significance, or which are redundant with others, will receive weights equal to 0. In the Ridge approach, it works with the sum of each square of the parameters and a penalty coefficient that reduces the sum, but without assigning it to 0. The L1 tends to provide a sparse solution mostly, being indicated to select prediction models or to carry out the feature importance determination, but it is biased towards the non-null parameters. The L1 works well when there many parameters, in comparison to the number of data or observations. The L2 approach leads to a quite complete result model, although it is not recommended to select models, because some least germane parameter will stay in the model.

2.3.3 Missing data fill in with Bootstrap

Incomplete data reveals to be more common than expected, in the normal DS life, requiring an evaluation and decision on how to handle them. The first shoot could be to drop out these data or positions, but it brings burdens not easy to cope with sometimes. Thus, a classical way to treat the missing data is to fill in them with reasonable guessing or estimation. One at hand is to use the nearest data, or its mean, trying to maintain such a coherence, although some anomaly may be present at that point, and it will be missed unfortunately. Another quick act is to fill in with the mode, solving some statistics issue in advance, but again introducing some external bias by choice. In all of them, it normal to consider filling in with a single value, no matter its origin.

Filling the missing data may consider the following:

a) Deleting the rows or columns where the missing data is detected, in a tradeoff between to have a better algorithm accuracy or performance against to drop other key data;
b) Imputing values, which may be continuous or categorical, in a tradeoff between to make something to work or to input some bias;

In the first option, which may be seen as the easiest, the burdens remain with the Data Scientist, which shall take the overall boundary and initial conditions to solve the stated problem. The Exploratory Data Analysis can provide a first assessment about the data quality and whether to drop the missing or weird values will represent an odd procedure.

The second choice to fill in the missing data can choose from some options as below:

a) the median or mean of the data of the same parameter, which harms the covariance of the data.
b) the last or closest observation, which can also be used the very right next one. Either way, some bias will be shown, small though.
c) the use of other algorithms, as a predictor tool, from Machine Learning or even Deep Learning. In a first glance, linear interpolation appears a reasonable candidate method, but this decision requires a better assessment.

Reference [23] presents another way to consider this data placement, with a combination of prediction methods and, in the end, many values will be available to fill in, depending on fair judgments, because the missing value may be at random, partially, or even not at all.

For this text, the meaning of bootstrap refers to a method of inferring results, for a population, from outcomes taken from a set of smaller random samples of that population, with the replacement during the sampling process, to get advantage of the aleatory process to enhance the quality of the result, demanding some computer power to carry on. Bootstrap is advised for problems with a large data population and tight schedules to reach a solution. The overall scheme is presented in Figure 11. Make note the way the sampling is carried out in many different combinations, but with the same pattern, say the number of original data at each sampling. The order does not matter at all.

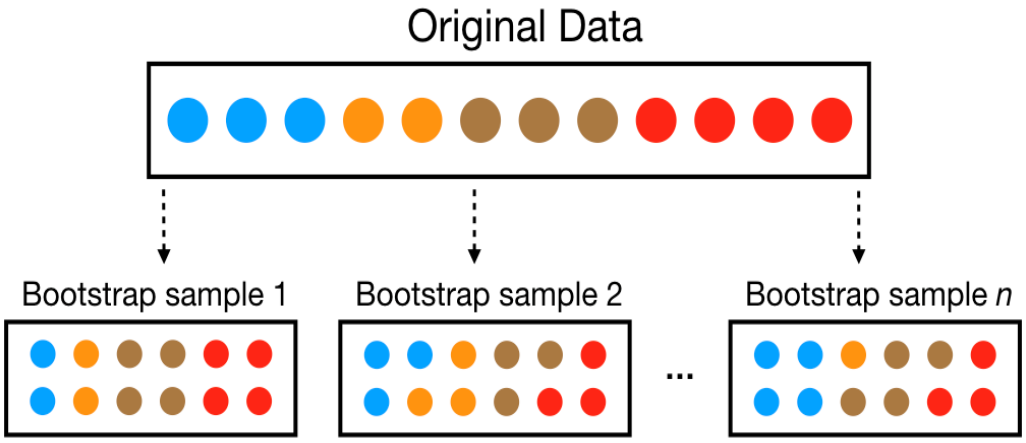

Figure 11: an overview of the bootstrap approach [23]

As a clear picture of the combined technique to fill in the missing value, using more than one single attempt, Figure 12 shows how the bootstrap can propose an answer for this problem. The simplest and common procedure would be to take out the bootstrap phase and proceed in a single line. Nonetheless, the combined sampling technique reveals a more complex process, thus demanding more computer resources and the use of a mean guessed value in the end. The term ‘multiple imputation’ can be the use of different models to predict the value, such as linear regression, decision trees and others. In the end, each option will pose a value, and the one to be chosen can come from the mean between all multiple imputation trials. The common start is the incomplete raw dataset. The ‘decision tree’ following could be any other math model to handle the problem definition and son.

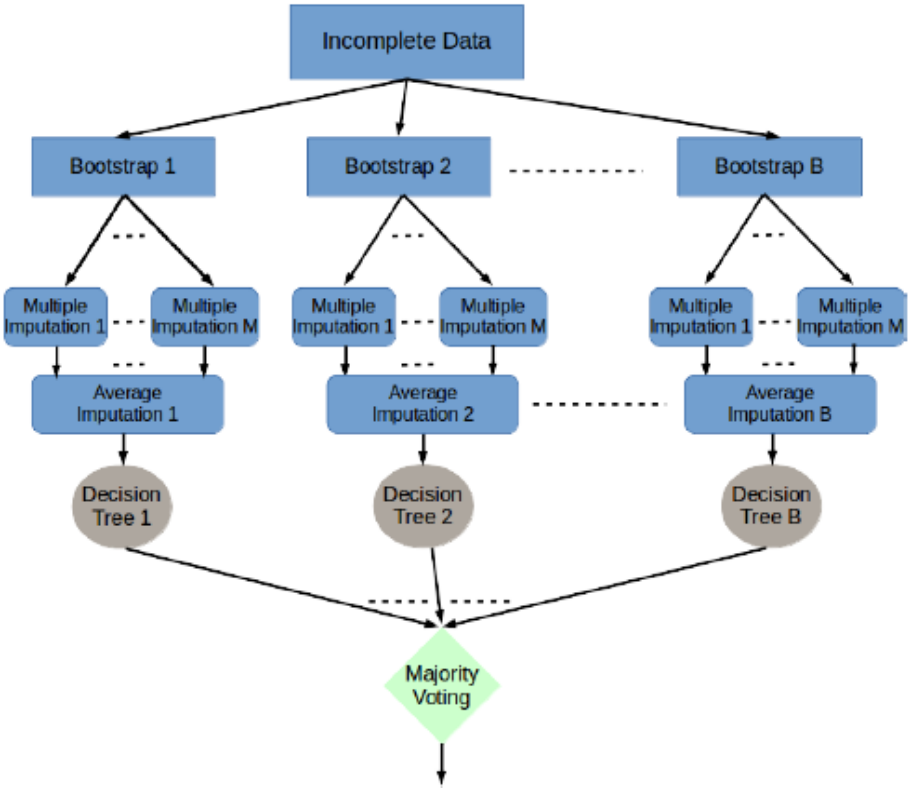

Figure 12: Bootstrap operations to full in a missing value [23]

It is recommended to note the 'average imputation' phases help to work on the uncertainties of the raw data, because it manages the variations randomly, preventing any strong bias to prevail. The use of decision trees also represents a pro factor because it can explain the reasons for a certain indication, even taking scarce or too diverse fields of data distribution. The majority voting stage may introduce some decision criteria based on the experience of the researchers or any specialist team. For instance, in the nuclear industry, normally the choice of a value will require a voting process 4 to 1, or take 4 options to select 1, based on how dissimilar, or not, the choice is shown against the other 3.

2.3.4 Surrogate models

The meaning of the word 'surrogate' is representative, substitute or equivalent for the scope of this text. Taking the physics of fluid aerodynamics, the concept of drag or pressure drop is expressed in terms of the quotient between force and area, derived from phenomena observation, modelling and the natural simplification, coming from dimension analysis, because it is simple and straight. Nonetheless, there is another way to comprehend the same fluid dynamics, now accepting a 'surrogate model', simpler and less accurate either, in other words is an approximation model.

In DS, for instance, the Machine Learning tools are a kind of surrogate models, because they are based on observations & measurements with plentiful of data, normally, belonging to complex phenomena, hard to see any correlation or direct explanation of the output at the beginning. While using surrogate models, the computing costs can be managed more suitably and the whole work can operate with more protective means to avoid leakage of germane data or information.

The spline of the surrogate models lays on the data input, which can come from natural observations or aleatory simulation. The more diverse and numerous the data, the better the surrogate model can provide the output needed. The initial Data Exploratory Analysis has a key value to decipher any quick correlation inside the input data, allowing a better tunning of the surrogate model as expected. After that, the next step is to construct the parameter features that will make part of the model, also known as feature engineering. Some parameters come with the input data, but other may be lined up to increase the chance of a fair result. These featured engineering parameters may be a combination of the natural parameters, or an implementation of other parameters based on the study and knowledge related to the data, requiring previous skills to carry on wisely. For instance, one input data transformation needed sometime is the "log transformation", to reshape the input numbers in a better way to understand the phenomena or natural correlation, caring to make the inverse transform when checking the output for the answer of the problem.

The surrogate will be a set of equations or business rules that take the input data and can provide an answer, based on the method reasoning to do so. The equations have coefficients which after combination with the input features provide the suggested answer (prediction or classification).

As a practical example of surrogate models, reference [24] works with surrogate models to predict the net present value from different oil extraction options, using data from oil wells already in use. They try to maximize the NPV by choosing the number of oil wells and their spatial coordinates according to the maximum oil production. The input data is complex because it takes manifold parameters from Geology, Chemistry, Logistics, Environmental rules and so on, a situation well suited to apply the Machine Learning algorithms. The reference considered the following regression (prediction) methods: support vector machine (SVM), gradient tree boosting (GTB), genetic algorithm (GA), multi-layer perceptron (MLP), elastic net (EN) which takes care of the order regularizations L1 and L2 altogether, and the K nearest neighbor (KNN), which was recommend in the end to predict the NPV.

2.3.5 Lower order models

As a general approach in any scientific field, it is better to work with models or problems with the lowest orders of complexity or magnitude. One benefit is the lower processing or computing costs, besides a faster answer and agility in the decision phase. In DS, the complexity of a math model deals with the number of parameters, for instance, to be handled. Nevertheless, the order reduction shall provide the quite same result as the high order models.

The common procedure is to take the measurements or data and derive the prediction models, which will have a high order for instance. Secondly, these models will be transformed into the low order models, by math transforms, such as projection, optimal Hankel norm approximation, or Krylov methods. More recently, a new suggestion has emerged as 'data driven order reduction', which does not need to go over a high dimension model firstly, as it can be seen with more details in reference [25].

Specifically in the DS field, the Principal Component Analysis (PCA) is a key technique to evaluate the dimensions that can be scrapped to reduce the order of the problem and related math model. The concept is to produce components by a combination of the

original parameters, in a manner that the components made will not covary among themselves, as they were orthogonal as possible or kind of. In terms of analytical geometry, the principal component will be a direction that explains the maximum amount of variance of the parameters. Therefore, the new direction or axis takes most of the information from the parameters and will be like axels where the difference between parameters will be detected easily. After the PCA, the variables or parameters will not have the original one, but a wise combination of them. The key trade-off to manage is some decrease of accuracy when reducing the problem order.

PCA requires a set of operations, normally:

a) Standardization – the initial variables will have their values into a standard range, normally between 0 and 1, or -1 and 1, taking care with the diversity of range sets, using the standard deviation and the mean value of the variables or parameters.
b) Matrix of covariance – this step is needed because variables may a high covariance and have redundant information as well. The covariance matrix is a numerical indication to check this dual condition.
c) Eigenvectors and eigenvalues of the covariance matrix – these concepts are useful to determine the principal components, which will be not correlated among themselves, based on the eigenvectors and eigenvalues. The first principal component will retain most of the information from the data, and the second the most important information also remaining and so on. The principal components shall behave like orthogonal ones. The higher the eigenvalue, the more principal the component is.
d) Determination of the features of the vectors – the dimensions of a matrix made of the vectors will be the principal components to be chosen, a decision that depends on the order of magnitude to accept. The columns of the matrix will be the vectors.
e) Redistribution of the original data into the new axis or components – this step is needed to shape the data into the new dimension axis or references, with a smaller order of magnitude and most of the original information retained.

In a simple graph explanation, Figure 13 shows the concepts above in a 2D approach. The initial variables were draw in black, referring to the original parameters. Searching for new orientations axis, following the steps right above, the new dimensions will be the red and blue options, where the red has retained more information than the blue, in a way to reduce the order of the model. In this example, the order remains the same, but it is easier to follow the math steps.

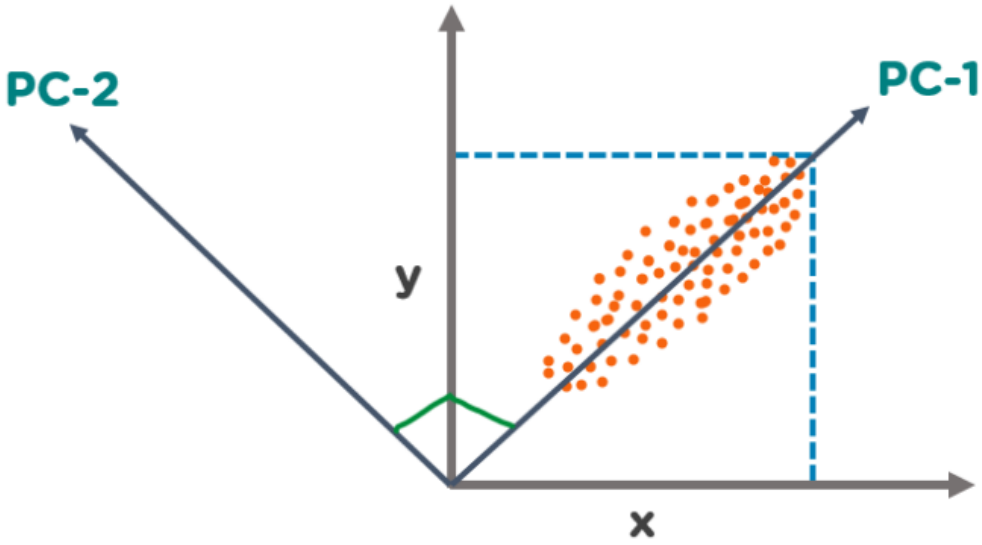

Figure 13: 2D example of Principal Component Analysis [26]

Note that the new variables orientation (red and blue) will be a combination of the black ones (x and y), requiring attention when analyzing the final prediction results, for instance. It may be needed a recovery transformation to stick to the original problem definition.

2.3.6 Kringing techniques for interpolation

As an example of a practical use of DS interpolation or data fitting, the term 'Kriging' refers to a specific math tool related to geostatistical procedure to adjust a surface to a 3D pattern: two localization variables (x, y, latitude, longitude etc.) and a third variable related to an interest value (z, outcome of function etc.). The technique is also known as Wiener-Kolgomorov prediction. It takes a limited set of observed data to make inference of a large data space. It condenses complex math operations to fit data into a functional pattern, with direct applications in Environmental and Health Sciences, Chemical Engineering and others, which deals with spatial distributions, particle dispersion inside variant media, among others, demanding a high computer capacity to work with data varying in time and space. The original motivation for the development of this technique was the gold dispersion prediction in South Africa as an oddity.

The main concept remains to predict a value based on average-data in the surroundings of the region of interest. An interesting way to see this technique is like a subset of the Bayesian inference. It commences with a distribution over functions. In geostatistical models, data comes from random events, although it may not be in this way really. This basic assumption refers to a spatial inference for locations without measured data, computing also the uncertainty associated to the prediction. The first task comprises to build up a random process capable to describe the measured data. The solution in the geostatistical framework focuses the assumption of diverse degrees of stationarity related to the random function, to infer a set of statistic values.

While comparing to Linear Regression or Gaussian decays, the Kringing technique uses spatial correlation between sampled data to predict the values in the completely spatial field, considering now the empirical measurements, rather than on a preconceived spatial

distribution. As a bonus, the Kriging technique produces uncertainty estimates of the neighborhoods of each predicted value. The closer the position to the interpolation point, the greater the weight for the spatial distribution. The farther positions receive smaller weights consequently. To prevent bias in the prediction phase, cluster points also receive smaller weights, because they do not add much information as single points. The Kringing technique tries to minimize the prediction error for a specific point, making its prediction at the measured points equals to the measured values.

The Kringing technique uses variograms extensively, which are a visual taking of the covariance between sample points. For each pair of sample points, the semi-variance, or the half mean-squared difference of their values, is associated to the distance between them. The actual measured points make part of a variogram labeled as 'experimental', or the 'reality'. The 'model variogram' is the distributional model that best adjust to the data, considering a set of math functions allowed previously, such as linear, spherical or power models, which includes the exponential too. Figures 14 and 15 show variograms and the concepts above. Note the vertical axel represents the statistical parameter (variance, semi-variance), as the horizontal axis is the distance between the points.

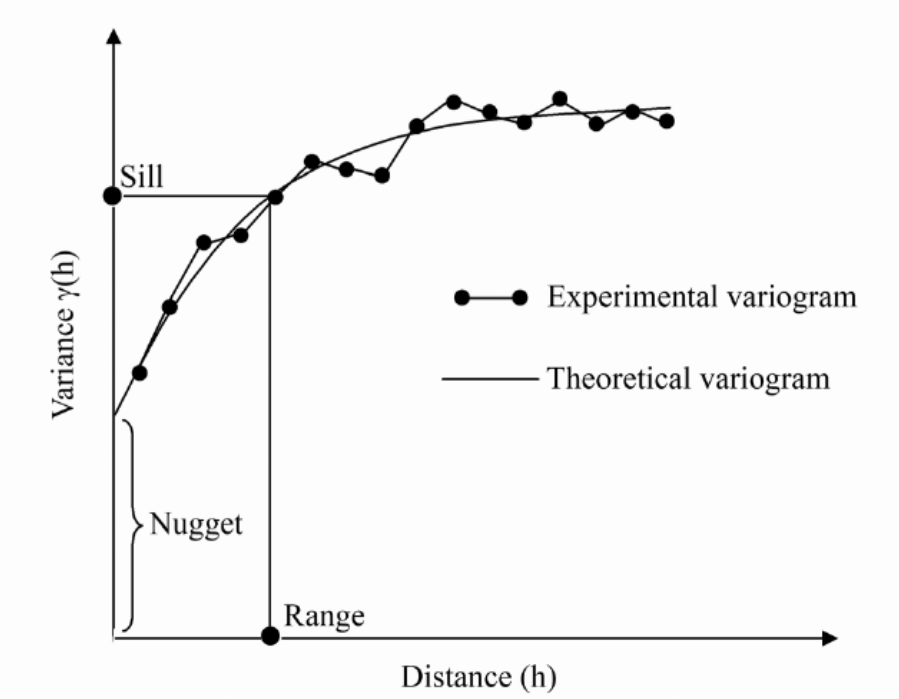

Figure 14: 2D variogram linked to the Kringing technique [27]

As shown right above, the experimental variogram deals with the measured points and are linked by a straight line, showing a discrete profile, and the theoretical variogram is a continuous line, generated by math functions prior selected. The term 'Nugget' means the initial predicted value for the distance equals to 0 or the 'small-scale' variability of the data. If the variance is low, the correlation between the points is high, and the opposite is true, checked at the upper part of the variogram, where the continuous line becomes almost horizontal. The Sill means the maximum variability between pairs of points. Conceptually, the Kringing model has the values of the measured points, although it has worked with a sample set of these points.

As seen below, different math models produce different theoretical models, which tries to reproduce the experimental model in the best way. The spherical and circular equations tend to occupy the upper left part of the graph, while the linear and exponential stay closer to the natural diagonal, or inclination of 45º. The Gaussian model has an odd shape, with an inflection point somewhere the middle of the Distance range, depending on the problem and data available.

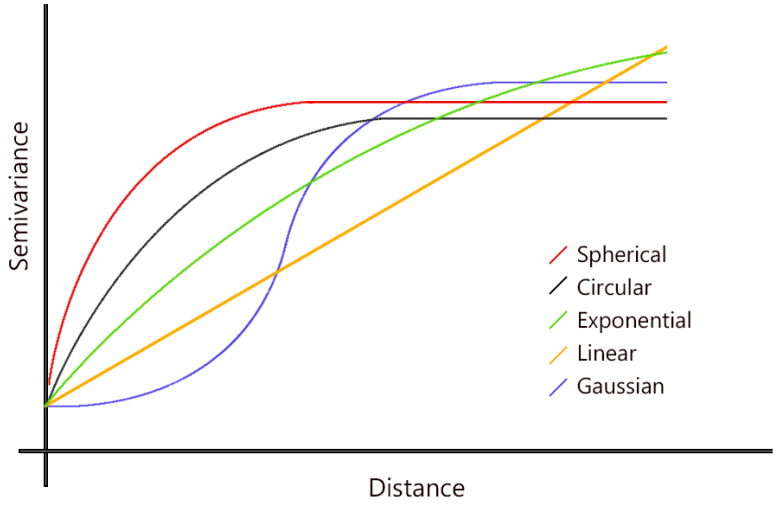

Figure 15: Different math models allowed in the Kriging model [27]

2.3.7 Bayesian techniques for data pipelines

As already mentioned in along this text, a prediction or classification model may require a constant update when dealing with a data pipeline, which is quite common with logistic and energy efficiency applications. Data comes from several sources or measuring instruments, following the sample strategies, and stored into databanks or data lakes to fill in DS tools. This update can be done by several techniques and schemes, depending on the experience of the DS team and their prior experiences, recent data recorded, and the information derived from them. Thus, the data background has a strong influence on the work to carry out, throughout the learning involved from different directions.

The first natural choice to a data update focuses the frequentist approach, which in a short essence deals with the question: Which model, or set of parameter values, are we most certain is the true model? Frequentist statistics is based on hypothetically resampling the data from an underlying true model, using the concept of likelihood, straightforward [28]. Nonetheless, this approach does not consider any previous experience or wise observation of hidden details eventually. Thus, another way to update the DS model may take concepts from the Bayes Theorem and its aftermaths. In this sense, our belief is taken into consideration, which sticks to the most common behavior, although it represents a hard technical step, due to the variability involved. For instance, referring to past data, the rate of sick people in a town may vary with the local temperature and season, with figures of 100 sick people for every 5ºC decrease in the winter. These numbers may not be so precise as stated, requiring a distribution approach to be more consistent. Let take a Gaussian

distribution (bell shape) to set the prior knowledge, setting an average and standard deviation. It is a straight application of the probability distribution functions or 'pdf'. Using a high mean and small standard deviation means the prior knowledge reflects a strong certainty about the data or situation in case. Being humbler, however, knowing that recent quick changes in the application of vaccines, let take a smaller average and larger standard deviation, reshaping the bell outline of the Gaussian distribution assigned to this case or problem. After all, the parameters can be seen as random variables and the pdf concept applies. The propose above to use Gaussian distribution on the parameter's estimation represents a way to regularize the input data, taking care of outliers as an example. Thus, it is the application of the Bayes technique, setting something based on the previous knowledge or evidence of an occurrence related to problem. Equally important, it is a iterative process, allowing permanent reviews of guessing and conclusions, based on the outcomes, in a closed or feedback loop.

Hierarchical models are a standard Bayes approach for data updating, using a scheme of data clustering following a knowledge path. The advantage of this way is to consider different data sources, with high or low data quantities, and dealing wisely with the outliers, because the Gaussian distributions shapes will take that math.

Another type of subject inside the data update is the measurement of the accuracy. Sometimes a set of conditions arise [28]:

a) There is no data enough to carry on the training phase, typical in problems of shortage of data. This scenario gets worse with the need to use cross validation.
b) The time involved in the training phase may be too long.
c) When using an unsupervised technique, like k-means, in classification for instance, the accuracy may not be easy to compute, because there is no labeled data to detect the learning improvement.

As a result, the application of the Bayesian method in the data update will be made to the point where the model accuracy does not change significantly. Also, the Bayesian method handles the errors linked to the measurement of the data with a probabilistic way, more suitable to be treated mathematically.

The Bayesian parameters are reviewed based on the new data input, changing the model set previously, in a continuous way. The update of the model aims to reduce the distance between the forecast results and the new measured data, and it is worth noting the model update time interval must be shorter than the data input process, otherwise the model will be always not fitted to handle the problem. This gap time will require a piece of money in the whole DS cycle.

2.4 The application of Data Science in engineering

Many DS applications have been closer to the common understanding recently, with the diffusion of algorithms linked to the prediction of human behavior, sales campaigns, and logistic planning. In this chapter, the use of DS will focus the renewable energy field. For instance, the application of the math algorithms and digital computing into the analysis of materials behavior provide more insights about the interactions between particles, in different size scales, which affects the capacity of the material to transfer heat, to conduct electricity or to recover from strain loads.

The two large application areas of DS in engineering are classification and prediction. The first area deals with to recognize in what kind of groups, patterns, or clusters a set of data can be assigned. In this task, the group identification comes from the alignment of several parameters, as close as possible, of common characteristics among the data sampled.

2.4.1 DS and solar and wind power sources

In this section, some practical applications of DS covering types of sustainable energy will be addressed. With a top-down approach, Reference [29] presents a scheme on how DS can be applied to propose or indicate a set of sustainable energy sources can be on the radar of energy planners or policy makers. It starts with data from the sources, towards a set of parameters, engineered or not, in a big data scale. Then, depending to the class of the problem, clustering, classification, or regression, or both at the same time, can be done using the vast DS toolbox, providing more than straight presentation from Data Analytics.

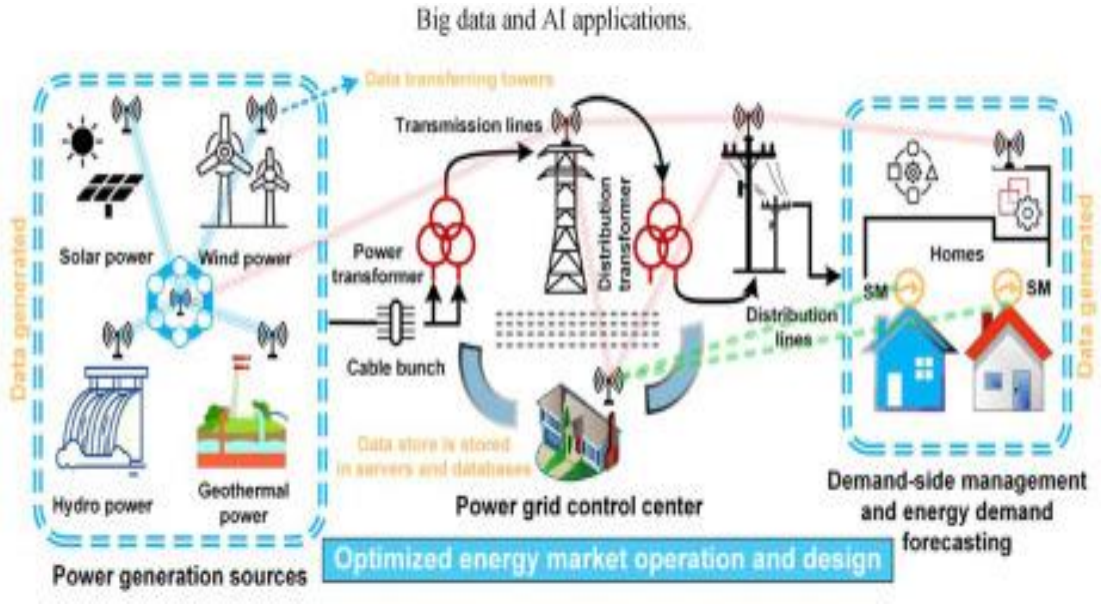

Figure 16: General scheme for DS application into renewable/sustainable energy field. [29]

As the sustainable energies make more common to our routines nowadays, the combined use of different sources makes all the sense to optimize the investment. This is not new in the energy market, but the difference calls because the decision process comes closer to the users than it used to be, with large generator and distributor groups taking the wheel for most of the time.

Reference [30] worked on a combined model with solar and wind energy generators, quite easy to have in most of the houses and buildings near the shorelines. Both exhibits energy production transients based on the local and climate boundary conditions, such as landscape details, latitude, and others. For instance, solar radiance varies along time and latitude, due to astronomical known reasons, but the efficiency of the solar panels depends also on the local temperature. Above 40 ^{o}C, the photovoltaic efficiency of the current solar panels drops drastically, in factors 5:1 or worse than that, depending on installation

and maintenance practices. Figure 17 shows the technical hardware involved, where the solar option is labeled by salmon color and the wind option by green tone.

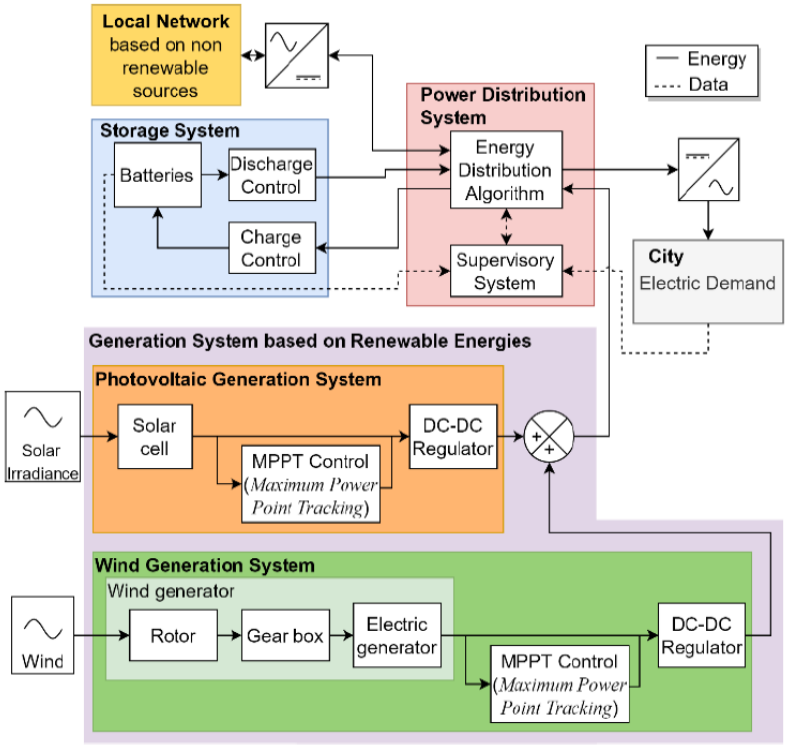

Figure 17: Wind and solar power sources and DS [30]

Both energy inputs are variable, signed by the wave form on the left bottom. The thinking core of the proposed system is the Energy Distribution Algorithm (purple box), which will compute a set of data to decide how to do with the energy stored, taking at least two possible decisions: supply or not the local network, with has equivalent sustainable energy sources, or supply or not the city, which is a bigger energy sink than the first one. Data comes along the energy to be managed, and that is the key difference from the conventional energy grid available in most cities. With DS, it is possible to use its tools to bound the decision along different time intervals, to maximize profits or cut unreasonable costs, based on the electric efficiency of each option or the combination of both [31].

2.4.2 DS and geothermal energy

Geothermal energy has gained the screens more recently, although it is well known for centuries, because it uses the natural hear source from the Earth´s core and can be linked to the factors bound with the occurrence of earthquakes. This energy source does not require sophisticated technology to be used, long pipelines, heat exchangers and industrial instrumentation, available from the oil industry mostly. Depending on the use, it can be set anywhere, but the economics play a key decision factor as usual. That is the point where DS comes in.

Reference [32] refers to a practical case to identify the best locations to build up geothermal stations, based on earth big data, from different sources in time and location. The general result of the research is presented in Figure 18.

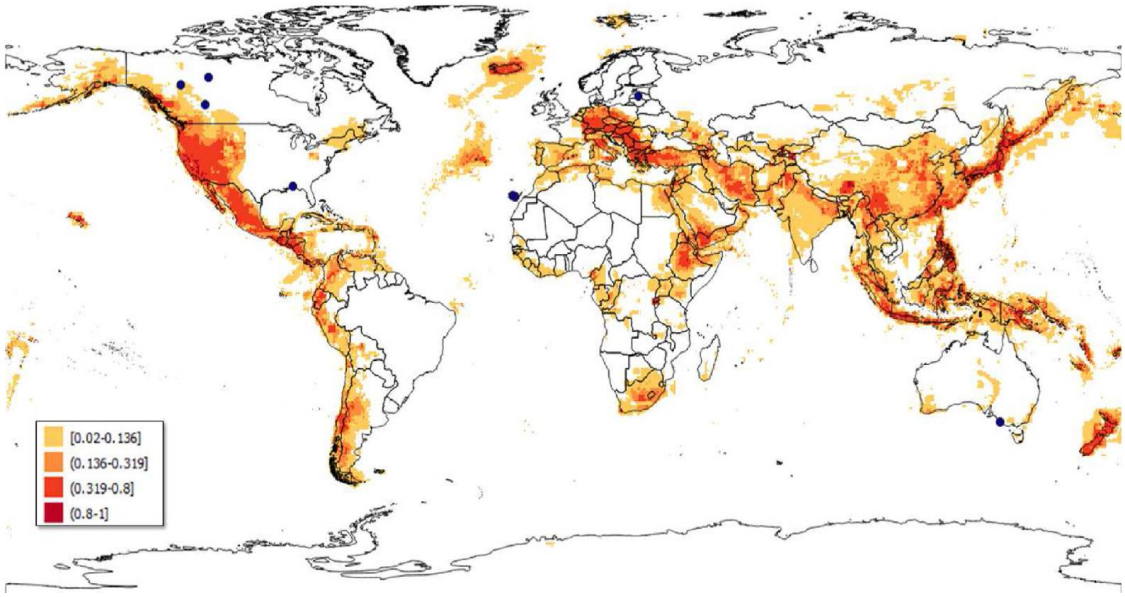

Figure 18: Optimal geothermal distribution for energy management [32]

The higher the number or reddish, the better potential to establish a geothermal energy station. This survey came from the application of DS, taking data from geological sources, time and geographic coordinates, as well as data from satellites of CO_2 distribution, earthquakes density, depth, magnitudes, global heat flow, groundwater sources, surfaced air temperatures, distance from land lines, among others. The model used to process the data sources, after engineered features, was the Maximum Entropy, which states the best way to explain a dataset will be the one with the highest entropy according to constraints with a prior knowledge. For instance, the current geothermal plants are a well-known constraint, also the new geological anomalies produced by recent volcano eruptions etc.

Taking the data pipelines from survey installations or instrumented field systems, or even mobile, the evaluation models can be updated constantly, refreshing the model parameters as discussed in the items above.

2.4.3 DS and nuclear energy

In the nuclear industry, the application of safeguards has taken more importance to prevent proliferation of nuclear materials and technology. It is necessary to avoid the use of weapons of mass destruction, such as thermonuclear or even dirty bombs, which can leave long time consequences for people and their generations.

One of the safeguards instruments is the frontier inspection of goods being transported. Some industrial nuclear hardware has its profile very close to ordinary ones, such as pumps, valves, or heat exchangers, although constructed under restricted materials and refined techniques. Thus, the customs agents must have a special training to be able to recognize whether an industrial good may be controlled by the safeguard regimes adopted. One way to help those agents considers the developing of an intelligent system, trained with certified data, and based on AI, to provide a first guidance whether the good should be controlled or not. In brief words, it is the application of NPL, in this case digital images.

Reference [33] presents a simple option focusing the most important nuclear hardware associated both to the fuel cycle and for nuclear power reactors. The more the AI systems is trained, the better will be the chances for a fair alarm to the customs agents, paying attention for the overfitting indicated the items above.

Another application of DS in the nuclear field deals with data filling regarding the operational performance, as pointed out by Reference [34], helping to detect problems in the operation of a nuclear system or plant. The target here are the operational parameters, which may have missing values in their records, due to a faulty instrumentation, lack of communications connections or even human errors, and all combined, as it was detected in the Chernobyl´s accident in 1986. Unfortunately, at that time, the DS was a kind of starting and would help to prevent the horrible outcomes. In their studies, the authors considered the options of the least squares support vector machines, the K-means clustering algorithm based on a noise algorithm. In the end, the reference uses wavelet analysis and the Mahalanobis distance, for data work, when dealing with analog sources for instance. The algorithm proposed by the authors works better than the mean procedure for instance, because it treats the oddities with more detailed scrutiny.

2.4.4 DS and wind power: structural monitoring application

The occurrence of winds on Earth comes from the heat transfer in the atmosphere, influenced by the incident heat flux from the Sun mostly. Thus, the prediction of wind can be based on physical models of heat transfer of large masses of gases, assuming the air as a classical gas in most of the situations.

However, the number of variables and their crossing influences make this problem complex to be solved, even with numeric procedures and computers, although feasible with some accuracy and many hypotheses. The other way to predict the wind occurrence is data driven, based on large quantities of accumulated data from real observations and following the DS cycle presented at the beginning of this text.

One of the applications of the data drive approach aims the structural reliability of wind towers used in the wind energy sector. In a few words, the structural tower is a vertical beam, with a concentrated load on the top and with a connection on the ground without any freedom degree. The dynamic load on the structure comes from the wind itself on the propeller blades, on the tower and from the ground to the tower, in a first definition, together with the static loads, such as weight and ground reactions. The loads will stress the structure material and strain will be detected, providing a data source on how the dislocations and other geometry changes occur in the material lattice.

This strain field is the data input for DS applications. Figure 19 shows the material distortion or singularity will be digitally treated to be a set of numbers, as an example of NLP application. Particularly, the arrangement has a standard AI framework to produce the dataset [35].

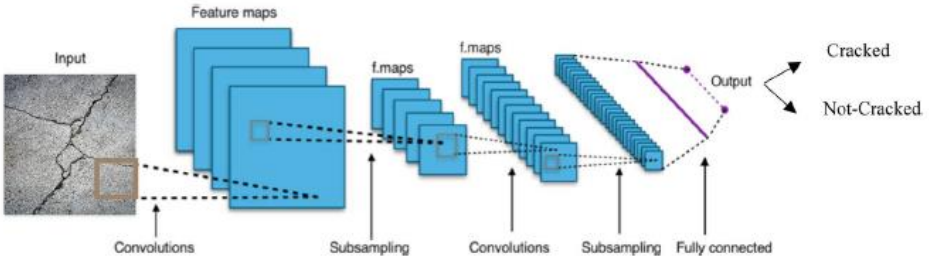

Figure 19: digital matrix data treatment for DS applications [35]

With the structural sensors, checking the strain regime of the tower, mainly at positions with stress concentration (distortion or change of geometries etc.), another set of sensors work to measure the wind characteristics, such as speed, direction, and variation rates of them, among others associated with the electric power production and transmission. Combining the data, the tower has a dataset associated to structure behavior, which must be within a safety standard envelope, associated to the concept of structural reliability.

Reference [36] presents an example of the use of AI to handle the situation when a blade of a wind turbine turns to be unbalanced, indicating the points with faults, suitable to be also useful in predictive maintenance programs. It is the same kind of application found in the aerospace industry more recently, improving the quality assurance in most cases.

Before the WWII, the term 'reliability' was associated to any system that had produced the same output for a time interval, close to the concept 'repeatability'. Nowadays, with the statistics advances, the concept of reliability has incorporated probability distribution functions (pdf) and can be framed as the probability of a system to perform its functions during a defined time interval. Connecting DS into this field, the prediction of structural status, according to dynamic loads, materials degradation, wind variations, can use manifold ML methods, starting with the clustering of the engineering factors and without using any specific instructions. ML looks like a surrogate model of the wind tower structural response, submitted to load conditions (stochastic or deterministic).

Classification can be set to indicate whether a set of data will associate to a safe state. On the other hand, ML prediction can provide a parameter future value, or index, of a structural component under a certain function analysis. Both cases can manage the ML techniques to do so, using different definitions of "distance", as the Manhattan distance for clustering.

As outlined above, a data drive method using ML demands a dataset with input parameters and one output at least, and with this arrangement, algorithms can be set to make predictions, which can be supervised or non-supervised. Also relevant, DL can also be set to perform the same tasks, with perceptrons and learning layers for instance. Structurally speaking, the prediction aims to forecast a limit state or boundary, throughout the concept of Limit State Functions (LSF), which may consider physical models or not. As already known from other structure applications, the use of ML in the wind power structures can be challenging, due to the need to have data from reliable measuring systems and from lab tests. Summing up, the use of ML may face a rare event, with few available data, nearing the situation of high dimensionality, or more parameters than occurrences, requiring special math techniques. The selection of the parameters demands a closer analysis in the problem definition as well, which will deal with a time-series in most of the cases.

The ML models used for the application focused here are shown in Figure 20. Noteworthy that many can be used together, in different phases of the DS cycle, say in classification and prediction, in this order, to provide an insight about the need. References and summarize them for the structural applications highlighted.

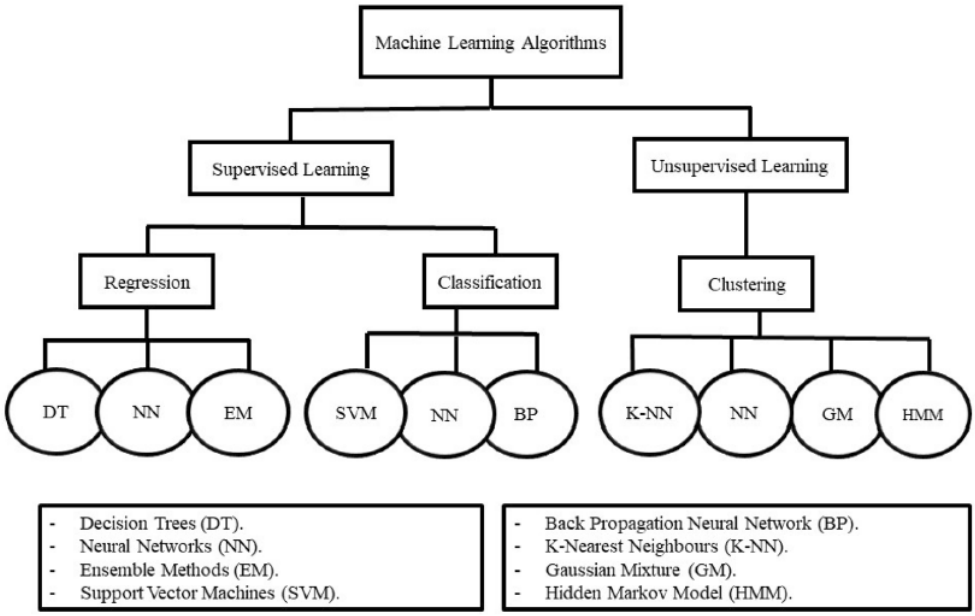

Figure 20: ML models aimed for structural monitoring application [36]

In terms of AI, the feedforward Multi-Layer Perceptron is commonly used in structural reliability studies, with four layers initially, which may vary based on the problem of interest, but as a matter of fact, the number of layers does not need to be large indeed. The Adaptative Sampling Methods can be used with MLP, tackling the data training phase, refining the AI model till a specific accuracy is achieved. The selection of the training points is based on the training selected previously and the associated performance index, leading to a lower computing cost. Reference [37] has a fair literature review of the application of AI in the structural reliability.

Looking into the wind power practical applications, Figures 21 and 22 show the overall instrumentation used with DS to monitor the performance of a wind power electric generator in an offshore application, that is more complex than on land, and the parallel with a structural bridge. Those systems have similarities on providing field data to check the structural reliability of their own cases [38].

Figure 21: wind power monitoring system linked to DS [38]

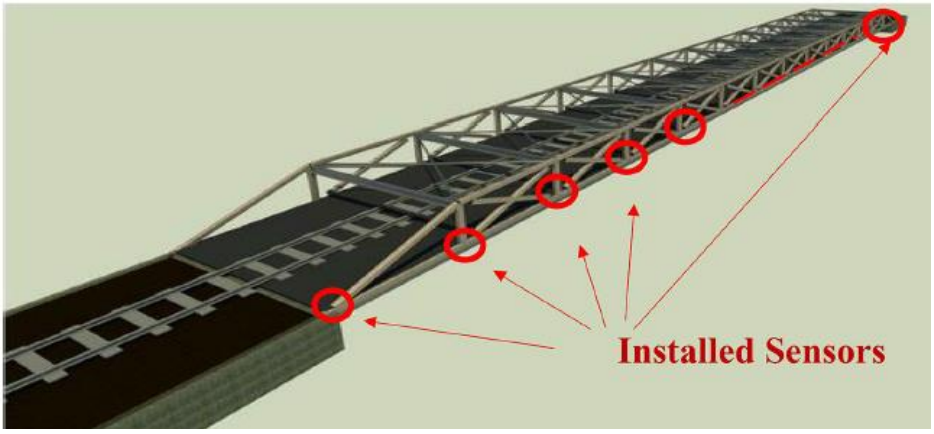

Figure 22: Structural bridge monitoring system linked to DS [38]

In both cases, data is gathered from a set of sensors, with a sampling frequency, into a data bus or pipeline, building up a timeline or time series, to be treated with DS algorithms. As seen in other engineering applications, this scheme will help other assessments and to feed up the industry with feedback from field applications, associated to the design, testing and improvement of new advanced materials and construction techniques, managing fuzzy phenomena as delamination of carbon fiber structures under dynamic loads and solar radiation together. The DS cycle will provide insights for a data driven decision on the choice of the resins, for instance, or the way the carbon fiber thread will be designed and carried out, nearer to the statistical control and structure health monitoring application, as shown in Figure 23.

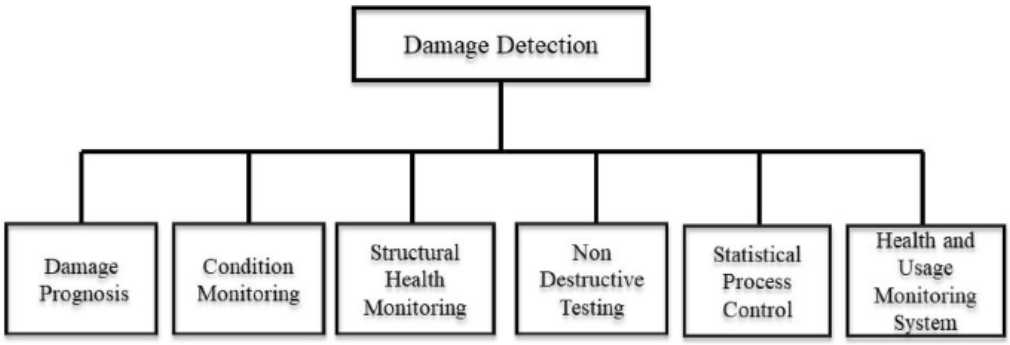

Figure 23: Damage detection fields related to advanced structures [38]

The DS application will require the data confirmation, data reliability and other data quality features, at a high level, because the data driven decision may have impact in the cost and performance evaluation of the wind power structure and, moreover, also have links with disputes among constructors and operators in case of incidents or even accidents and damages.

3 Concluding Remarks

In this chapter, a general view of the ship design was presented, covering its phases and relevant technical issues. The characterization of the ship dimensions and systems, along design drawings and materials application, was described in a progressive way. As the sections of the chapter were presented, the engineering materials field had been expanded in terms of technical figures and details. The focus was to provide a spark to the reader about the variety of topics around the ship materials selection and its evolution, without the pretense explore all the fields in deep. The data presented came from open sources, intending to provide an order of magnitude of the reasoning discussed along the text, and they may be seen as mean values. All topics here shown are under continuous development, mainly the nuclear examples, and more recently the application of Machine Learning and Artificial Intelligence. As a high potential subject for technical improvement, the environmental impact reduction with new fuels and propulsion systems has its place to catch the general attention.

Acknowledgements

The author thanks University of São Paulo (USP) and Paulista University (UNIP) for all their academic and technical support.

References

[1] https://www.analyticsinsight.net/top-10-big-data-disruption-coming-in-in-2022/, accessed on June, 30th 2022.

[2] Desclaux, Gilles. (2018). Big Data & Artificial Intelligence for Military Decision Making.

[3] https://www.iai.it/sites/default/files/978195445000.pdf , accessed on June, 30th 2022.

[4] Ferster, Colin & Coops, Nicholas. (2015). Integrating volunteered smartphone data with multispectral remote sensing to estimate forest fuels. International Journal of Digital Earth. 9. 1-26. 10.1080/17538947.2014.100286.

[5] https://people.smp.uq.edu.au/DirkKroese/DSML/DSML.pdf, accessed on June, 20th 2022.

[6] Chen, Jiangping & Ayala,, Brenda & Alsmadi, Duha. (2018). Fundamentals of Data Science for Future Data Scientists. 10.1201/9781315209555-6.

[7] Mcdonald, Lynn. (2013). Florence Nightingale, statistics and the Crimean War. Journal of the Royal Statistical Society: Series A (Statistics in Society). 177. 10.1111/rssa.12026.

[8]https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1063041/ddat-playbook-march-2022.pdf, accessed on June 20th, 2022.

[9] https://doi.org/10.1002/asi.4630270213, accessed on June, 20th 2022.

[10] http://www2.math.uu.se/~thulin/mm/breiman.pdf, accessed on June, 17th 2022.

[11] Dineva, Snejana & Nedeva, Veselina. (2018). Expanding Web and Innovation Skills for 21st Century.

[12] Kosse, Pascal & Blomberg, Kati & Mikola, Anna & Heinonen, Mari & Kuokkanen, Anna & Lange, Ruben-Laurids & Lübken, Manfred & Wichern, Marc. (2016). Climate change and green- house gas emissions within the context of urban waste water management. WaterSolutions. 89-94.

[13]https://www.ipcc.ch/site/assets/uploads/2018/05/ipcc_wg_I_1992_suppl_report_section_a1.pdf, accessed on June, 16th 2022.

[14] https://www.eia.gov/todayinenergy/detail.php?id=41533, accessed on June, 17th 2022.

[15] https://morioh.com/p/9ba3ed05fad1, accessed on June, 15th 2022.

[16] https://www.dreamstime.com/illustration/etl-process.html, accessed on June, 15th 2022.

[17] Zheng, Huan, and Yanghui Wu. 2019. "A XGBoost Model with Weather Similarity Analysis and Feature Engineering for Short-Term Wind Power Forecasting" *Applied Sciences* 9, no. 15: 3019. https://doi.org/10.3390/app9153019.

[18] https://twitter.com/datasciencedojo/status/1404167692173692931, accessed on June, 30th 2022.

[19] https://www.analyticsvidhya.com/blog/2021/03/data-validation-and-data-verification-from-dictionary-to-machine-learning/, accessed on June, 15th 2022.

[20] https://www.ibm.com/docs/en/db2/11.1?topic=analytics-pmml-markup-language-data-mining, accessed on June, 20th 2022.

[21] https://www.oracle.com/a/ocom/docs/data-science-lifecycle-ebook.pdf, accessed on June, 15th 2022.

[22]https://www.webpages.uidaho.edu/~stevel/517/The%20Data%20Science%20Design%20Manual.pdf, accessed on June 30th 2022.

[23] Aljuaid, Tahani & Sasi, Sreela. (2016). Proper imputation techniques for missing values in data sets. 1-5. 10.1109/ICDSE.2016.7823957.

[24] Yetunde M. Aladeitan, Akeem O. Arinkoola, Okhiria D. Udebhulu & David O. Ogbe | Wenjun Xu (Reviewing editor) (2019) Surrogate modelling approach: A solution to oil rim production optimization, Cogent Engineering, 6:1, DOI: 10.1080/23311916.2019.1631009

[25] Brunton, Steven & Kutz, J.. (2020). 7 Data-driven methods for reduced-order modeling. 10.1515/9783110671490-007.

[26]https://www.simplilearn.com/tutorials/machine-learning-tutorial/principal-component-analysis, accessed on June 30th 2022.

[27] https://doi.org/10.1016/B978-0-12-410437-2.00008-4, accessed on June, 20th 2022.

[28] Redman, Thomas. (2005). Measuring data accuracy: A framework and review. Information Quality. 21-36.

[29] https://doi.org/10.1016/j.jclepro.2021.125834, accessed on June, 20th 2022.

[30] C. F. Camilo, J. M. Rosário, M. M. Jefry and D. Dumur, "Data Science-based Sizing Approach for Renewable Energy Systems," 2020 IX International Congress of Mechatronics Engineering and Automation (CIIMA), 2020, pp. 1-6, doi: 10.1109/CIIMA50553.2020.9290303.

[31] P. Li, R. Dargaville, F. Liu, J. Xia and Y. -D. Song, "Data-Based Statistical Property Analyzing and Storage Sizing for Hybrid Renewable Energy Systems," in IEEE Transactions on Industrial Electronics, vol. 62, no. 11, pp. 6996-7008, Nov. 2015, doi: 10.1109/TIE.2015.2438052.

[32] https://doi.org/10.1016/j.jclepro.2020.121874, accessed on June, 20th 2022.

[33] Marques, André Luis Ferreira (2021). Image captions generation for non-proliferation control with Artificial Intelligence. Brazil: ABEN. INAC-2021.

[34] https://www.hindawi.com/journals/stni/2022/4172622/, accessed on June, 30th 2022.

[35] Garrett, Joseph & Mei, Hanfei & Giurgiutiu, Victor. (2022). An Artificial Intelligence Approach to Fatigue Crack Length Estimation from Acoustic Emission Waves in Thin Metallic Plates. Applied Sciences. 12. 1372. 10.3390/app12031372.

[36] Saraygord Afshari, S., Enayatollahi, F., Xu, X., & Liang, X. (2022). Machine learning-based methods in structural reliability analysis: A review. In Reliability Engineering & System Safety (Vol. 219, p. 108223). Elsevier BV. https://doi.org/10.1016/j.ress.2021.108223.

[37] https://doi.org/10.1016/j.ress.2021.108223, accessed on June, 30th 2022.

[38] Flah, Majdi & Vargas, Itzel & Ben Chaabene, Wassim & Nehdi, Moncef. (2020). Machine Learning Algorithms in Civil Structural Health Monitoring: A Systematic Review. Archives of Computational Methods in Engineering. 28. 10.1007/s11831-020-09471-9.

Suggested Reading

A - Energies 2022, 15(2), 578; https://doi.org/10.3390/en15020578

B - https://hdsr.mitpress.mit.edu/pub/d9j96ne4/release/3

C - https://www.hindawi.com/journals/sp/2020/8616039/

D - https://www.sciencedirect.com/science/article/pii/S0925753519308835