## ****Data Science Tutorial: What is Data Science?****

The term Data Science has emerged recently with the evolution of mathematical statistics and data analysis. The journey has been amazing, we have accomplished so much today in the field of Data Science.

In the next few years, we will be able to predict the future as claimed by researchers from MIT. They already have reached a milestone in predicting the future, with their awesome research. They can now predict what will happen in the next scene of a movie, with their machine! How? Well it might be a little complex for you to understand as of now, but don’t worry by the end of this blog, you shall have an answer to that as well.

Coming back, we were talking about Data Science, it is also known as data driven science, which makes use of scientific methods, processes and systems to extract knowledge or insights from data in various forms, i.e either structured or unstructured.

What are these methods and processes, is what we are going to discuss in this Data Science Tutorial today.

Moving forward, who does all this brain storming, or who practices Data Science? A **Data Scientist**.

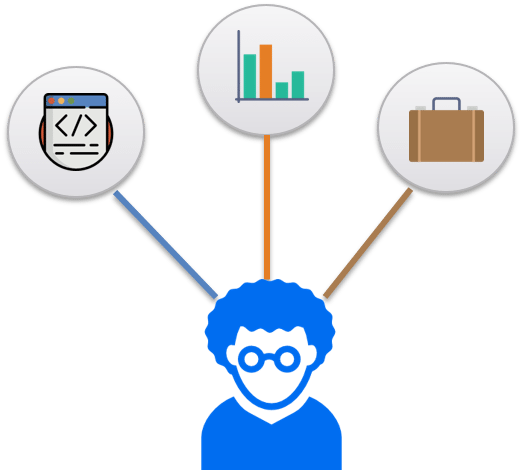
**What is Data Munging?**

The process of manual [data cleansing](https://www.trifacta.com/data-cleansing) prior to analysis is known as data munging. This process can be a laborious task without the right tools. The common interface used for data munging is often Excel, which lacks the sophistication for collaboration and automation to make the process efficient. 80% of the time spent on data analytics is allocated to data munging, where IT manually cleans the data to pass over to business users who perform analytics. Data munging is time consuming and disjointed process gets in the way of extracting true value and potential from data.

Instead of using data munging techniques to analyze your data, you should be [wrangling data](https://www.trifacta.com/data-wrangling/) with Trifacta. We’ve developed a 6 step guide to data wrangling using [Trifacta Wrangler’s](https://www.trifacta.com/products/) features for you.

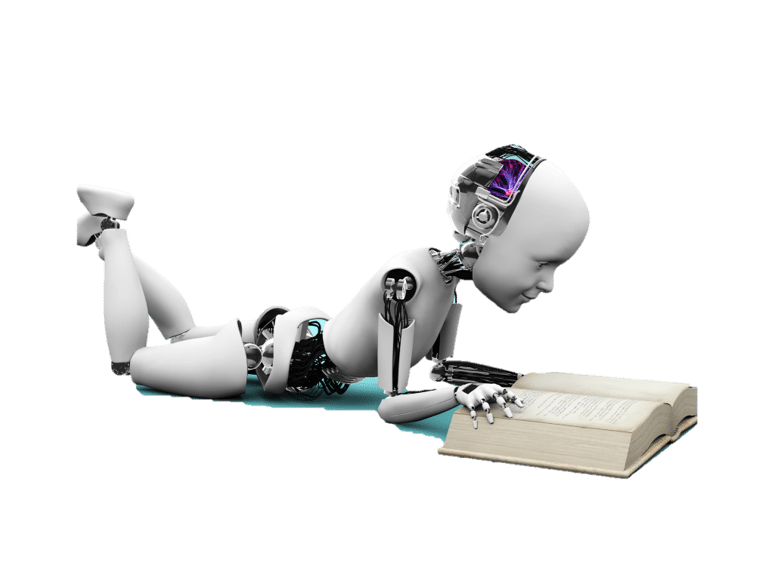
## ****Who is a Data Scientist?****

## Data Scientist - Data Science Tutorial - Edureka



As you can see in the image, a Data Scientist is the master of all trades! He should be proficient in maths, he should be acing the Business field, and should have great Computer Science skills as well. Scared? Don’t be. Though you need to be good in all these fields, but even if you aren’t, you’re not alone! There is no such thing as “a complete data scientist”. If we talk about working in a corporate environment, the work is distributed among teams, wherein each team has their own expertise. But the thing is, you should be proficient in atleast one of these fields. Also, even if these skills are new to you, chill! It may take time, but these skills can be developed, and believe me it would be worth the time you will be investing. Why? Well, let’s look at the job trends.

## ****What is Machine Learning?****



It is a type of Artificial Intelligence that makes the computers capable of learning on their own i.e without explicitly being programmed. With machine learning, machines can update their own code, whenever they come across a new situation.

Concluding in this Data Science Tutorial, we now know Data Science is backed by Machine Learning and its algorithms for its analysis. How we do the analysis, where do we do it. Data Science further has some components which aids us in addressing all these questions.

Before that let me answer how MIT can predict the future, because I think you guys might be able to relate it now. So, researchers in MIT trained their model with movies and the computers learnt how humans respond, or how do they act before doing an action.

For example, when you are about shake hands with someone you take your hand out of your pocket, or maybe lean in on the person. Basically there is a “pre action” attached to every thing we do. The computer with the help of movies was trained on these “pre actions”. And by observing more and more movies, their computers were then able to predict what the character’s next action could be.

Easy ain’t it? Let me throw one more question at you then in this Data Science Tutorial! Which algorithm of Machine Learning they must have implemented in this?

**Data Wrangling in Practice: What to Expect**

There are typically six iterative steps that make up the [data wrangling](https://www.trifacta.com/products/) process.

1. **Discovering:** Before you can dive deeply, you must better understand what is in your data, which will inform how you want to analyze it. How you wrangle customer data, for example, may be informed by where they are located, what they bought, or what promotions they received.
2. **Structuring:** This means organizing the data, which is necessary because raw data comes in many different shapes and sizes. A single column may turn into several rows for easier analysis. One column may become two. Movement of data is made for easier computation and analysis.
3. **Cleaning:** What happens when errors and outliers skew your data?  You clean the data. What happens when state data is entered as CA or California or Calif.? You clean the data. Null values are changed and standard formatting implemented, ultimately increasing data quality.
4. **Enriching:** Here you take stock in your data and strategize about how other additional data might augment it. Questions asked during this data wrangling step might be: what new types of data can I derive from what I already have or what other information would better inform my decision making about this current data?
5. **Validating:** Validation rules are repetitive programming sequences that verify data consistency, quality, and security. Examples of validation include ensuring uniform distribution of attributes that should be distributed normally (e.g. birth dates) or confirming accuracy of fields through a check across data.
6. **Publishing:**  Analysts prepare the wrangled data for use downstream – whether by a particular user or software – and document any particular steps taken or logic used to wrangle said data. Data wrangling gurus understand that implementation of insights relies upon the ease with which it can be accessed and utilized by others.

## What is Data Wrangling

Data is transforming the world every day. However, it is said that Data Scientists spend nearly 70 percent of the time cleaning and preparing data because not all data out there can be useful in their raw format. In addition, Data Wrangling meaning is often misunderstood. Let me give you a clear definition first.

Data Wrangling is the process of converting and mapping data from its *raw* form to another format with the purpose of making it more valuable and appropriate for advance tasks such as Data Analytics and Machine Learning.

#### The Goals of Data Wrangling

* It should provide precise and actionable data to Business Analysts in a *timely matter.*
* Reduce the time which is being spent on collecting and arranging data
* Enable Data Scientist to focus mainly on analysis rather than wrangling of data
* Drive better decisions based on data in short time span

You have a basic idea about what Data Wrangling is, now let’s look into key steps in Data Wrangling process with basic examples to get you started.

### 1— Acquiring Data

The first and most important step is of course, acquiring and sorting data. Or we can say that, finding your data to investigate it further might be the most crucial step towards reaching your goal of answering your questions. However, before finding data, you must know the following properties and you must be okay with that, because this is just a start of a tedious process.

### Not All Data Is Created Equal

Albeit we would like to believe in the truthfulness and quality of data we see, not all data will measure up to our expectations. When first exploring data, you must ask yourself small set of questions:

* Is the author of source reachable if I have any questions or concerns?
* Does the data appear to be regularly updated?
* Does it come with information as to how It was acquired and what kind of samples were used in its acquisition?
* Is there any other source where you can verify the data?

If your answers to three or more question is *yes*than you are on right track, whereas if answer to one or more question is *no* than you have to dig a little more into it.

### Fact Checking

Fact checking your data, although most of the time annoying, is paramount to the validity of your reporting. If you have access to some of the tools such as LexisNexis, Cornell University’s arXiv Project, Google’s Scholar search, and recently introduced Google’s Data Search, you can study what others have studied and used of a project or research. Once you have validated and fact checked your data, it will be easier to determine its validity in future.

### Where to Find data

It is obvious that you are not going to ring everyone’s telephone to collect data. Just like there are multiple source to validate your data, there are enormous number of sources from where you can collect your data. Which includes Government data, Data from NGOs, Educational or University Data, Medical or Scientific Data, Crowd sourced Data and so on. Know the [best places to find datasets for Data Science Projects](https://www.digitalvidya.com/blog/best-places-to-find-data-sets-for-data-science-projects/).

Now let’s jump to our main step, which is Data Cleaning.

### 2— Data Cleaning

Cleaning up data is not more of a glamorous task but it is the essential part of Data Wrangling. To become a Data Cleaning expert you must have precision, knowledge of the particular field, and on top of that patience. Yes, Patience.

Moving towards technical side, Python can help you clean your data easily. Assuming that you have basic knowledge of Python, in this chapter we will look at some Data Wrangling with Python.

### Data Clean up basics

To perform operations, we need data. Here, we will use data-set of UNICEF related to child labor. Let me give you a little insight into the data. In the initial data-sets there are *Multiple Indicator Cluster Surveys (MICS)*. These surveys are household-level surveys performed by UNICEF workers and volunteers to help research the living conditions of women and children throughout the world. In looking through the latest surveys, we pulled some data from Zimbabwe’s latest MICS to analyse. You will find the updated .csv [here](https://github.com/jackiekazil/data-wrangling/tree/master/data/unicef).

Now, to the six activities!

1. Discovering is something of an umbrella term for the entire process; in it, you learn what is in your data and what might be the best approach for productive analytic explorations. For example, if you have a customer data set, and you learn that most of your shoppers are from a single part of the country, you’re going to keep that in mind as you proceed with your data [](http://2s7gjr373w3x22jf92z99mgm5w.wpengine.netdna-cdn.com/wp-content/uploads/2015/09/data-discovery.jpg)work. If all of your users are in the Northeast, for example, that will suggest that you might be interested in winter weather patterns if you want to properly understand their behavior. In the world of food, discovering is akin to learning how to leverage eggs, or figuring out that your garlic is more intense than you’re used to.

2. Structuring is needed because data comes in all shapes and sizes. For example, you might have a transaction log where each entry might have one or more items associated with it (think shopping basket). To conduct an inventory analysis, you will likely need to expand each transaction into individual records for each purchased item. Alternatively, you might want to analyze which products are often bought together. In this case, expanding each transaction into every pair of purchased items might be appropriate. With food, structuring is often about cutting (chopping, dicing, julienning, etc.). But it also involves actions like blending, emulsifying, wrapping or infusing.

3. Cleaning involves taking out data that might distort the analysis. A null value, for example, might bring an analytic package to a screeching halt; you [](http://2s7gjr373w3x22jf92z99mgm5w.wpengine.netdna-cdn.com/wp-content/uploads/2015/09/data-cleansing.png)may well want to replace it with a zero or an empty string. You might want to standardize a particular field, replacing the many different ways that a state might be written out — such as CA, Cal and Calif — with a single standard format. In broad strokes, cleaning requires knowledge about data quality and consistency – knowing how various data values might impact your final analysis. Similarly, cleaning food is often about removing the bits that you don’t want or need – e.g., de-seeding a pepper to cut the spiciness or rinsing pickled ingredients to soften their pungency or trimming the fat from a cut of meat to balance the texture.

4. Enriching allows you to take advantage of the wrangling you have already done to ask yourself: “Now that I have a sense of my data, what other data might be useful in this analysis?” Or, “What new kinds of data can I derive from the data I already have?” In other words, enrichment is often about joins and complex derivations (like convolutions and converting a time stamp to a day of week). Purchase transaction data, for example, might benefit from data associated with each customer’s profile or historical purchase patterns. Similarly, a car insurance underwriter might want to know crime rates in the neighborhoods of the customers they insure to better estimate risk. This new information is sometimes available in in-house databases, but, and increasingly so, may be sourced from marketplaces for third-party data. The quintessential example of enrichment in the food world is the addition of spices – salt, pepper, turmeric, saffron, etc. The intent is to complement what you have to achieve a better final result.

5. Validating is the activity that surfaces data quality and consistency issues, or verifies that they have been properly addressed by applied transformations. Validations should be conducted along multiple dimensions. At a minimum, assessing whether the values of an attribute/field adhere to syntactic constraints (e.g., boolean fields encoded as ‘true’/‘false’ as opposed to ‘1’/’0’ or ‘T’/’F’) as well as distributional constraints (e.g., birth dates in a customer database should be fairly uniformly distributed over months of the year). Additional validations might involve cross-attribute/field checks like ensuring all negative bank transactions have the appropriate transaction type (e.g., ‘withdrawal’, ‘bill pay’, or ‘check’). Food evaluations are similarly multi-dimensional – checking things like temperature, taste, appearance and texture.

6. Publishing refers to planning for and delivering the output of your data wrangling efforts for downstream project needs (like loading the data in a particular analysis package) or for future project needs (like documenting and archiving transformation logic). Some data software tools have dramatic [](http://2s7gjr373w3x22jf92z99mgm5w.wpengine.netdna-cdn.com/wp-content/uploads/2015/09/publish-newspapers.jpg)performance increases when they encounter data structured in a certain fashion. Efficient data analysts will know this, and will wrangle the data to match its format for the data’s eventual target. Across projects, it often makes sense to replicate a set of data wrangling steps/practices for re-use on other datasets. Experienced data analysts maintain a library (often personal, sometimes shared) of common transformation logic that they can leverage new projects. In food preparation, there are actions that can be taken to speed up cooking times (like de-boning a chicken or pre-soaking ingredients) or to improve the flavor or texture development of a final dish (e.g., peeling the skin of vegetables).

So that is how we at Trifacta think about data wrangling. Two points are worth repeating. First, that you can, and often will, move back and forth among the different activities until you think you’ve nailed your data preparation. Second, modern data wrangling software itself does a lot of the grunt work for you, leveraging sophisticated algorithms and built-in knowledge of downstream constraints to guide users into good wrangling actions.

**Data Munging**

The next phase of a machine learning project involves a process called “data munging.” It is often the case where the data imported into the

R environment is inconvenient or incompatible with machine learning algorithms, so with data munging (also known as data transformation) the data can be massaged into a more hospitable form. Data munging cannot be taken lightly as many times it can consume up to 80% of the entire machine learning project. The amount of time needed for a particular project depends on the health of the data: how clean, how complete, how many missing elements, etc. Open source R has many mechanisms and packages to facilitate data transformations and cleaning, e.g. dplyr, reshape2, lubridate, etc. The specific tasks and their sequence should be recorded carefully so you can replicate the process. This process becomes part of your data pipeline. Here is a short list of typical data munging tasks, but there potentially are many more depending on the data:

* Data sampling
* Create new variables
* Discretize quantitative variables
* Date handling (e.g. changing dates stored as integers to R date objects)
* Merge, order, reshape data sets
* Other data manipulations such as changing categorical variables to multiple binary variables
* Handling missing data
* Feature scaling
* Dimensionality reduction

Alternatively, RRE has big data functions for data transformations like rxSplit() which can be used to minimize the number of passes through a large data set. It can efficiently split a large data set into pieces in order to distribute it across the nodes of a cluster. You might also want to split your data into training and test data so that you can fit a model using the training data and validate it using the test data. RRE also can perform sorting with rxSort(), merging with rxMerge(), and missing value handling with big data features such as the removeMissing argument for the rxDTree() algorithm.

Exploratory Data Analysis

Once you have clean, transformed data inside the R environment, the next step for machine learning projects is to become intimately familiar with the data using exploratory data analysis (EDA). The way to gain this level of familiarity is to utilize the many features of the R statistical environment that support this effort — numeric summaries, plots, aggregations, distributions, densities, reviewing all the levels of factor variables and applying general statistical methods. A clear understanding of the data provides the foundation for model selection, i.e. choosing the correct machine learning algorithm to solve your problem.

Open source R has many mechanisms for EDA including hist() for histograms, boxplot() for boxplots, barplot() for barplots, plot() for scatterplots, heatmap() for heatmaps, etc. Using these tools allows for a deep understanding of the data being employed for machine learning. This understanding serves the purpose of feature engineering.