What is Data Scientist? How to become Data Scientist?

Data Scientist is the most sexy profession of the 21st century, according to Harvard Business Review. With intensive skillset and spread across many fields, Data Scientist is also considered 'rare as unicorn'.

Data Scientist is the most sexy profession of the 21st century, according to Harvard Business Review. With intensive skillset and spread across many fields, Data Scientist is also considered 'rare as unicorn'.

Here are the share of Nguyen Hoan about Data Scientist career to better understand.

1. What is Data Scientist? Their specific work?
2. What qualities and skills are needed?
3. What to learn to become a Data Scientist?

Biography : Nguyen Hoan graduated with a Bachelor degree from Ho Chi Minh City University of Natural Sciences, majoring in Software Engineering. Later, he studied Master of Data Mining at University of Trento, Italy. In 2013, Hoan returned home, started working at Sentifi with Data Scientist position.

Currently, Mr. Hoan is living in France and working remotely with a position of Data Scientist for Xomad company based in LA, USA.

According to him, what is Data Scientist?

Data Scientist is the creator of the data value, with two main tasks:

1. **Collect and process data to find valuable insights.**

For example, based on the information gathered from post / comment / status on social networks, Data Scientist can find out: every nearly to Valentine's Day, the frequency of the ABC brand is much higher.

This is a valuable insight that the Marketing department can use for Valentine season advertising campaigns.

1. **Explain, present those insights to stakeholders, to transform insight into action.**

For example, when finding insights worth from data, you need to make reports / presentations, or visualization to perform, explain to stakeholders:

1. What is that insight, what does it mean?
2. How specific applications can be used to benefit businesses / products / users.

However, Data Scientist is a very new profession, so its definition is quite vague and ambiguous (even in the world). Therefore, depending on the company that describes the job, skillset requirements, even job title may differ slightly.

What is the difference between Data Analyst and Data Scientist?

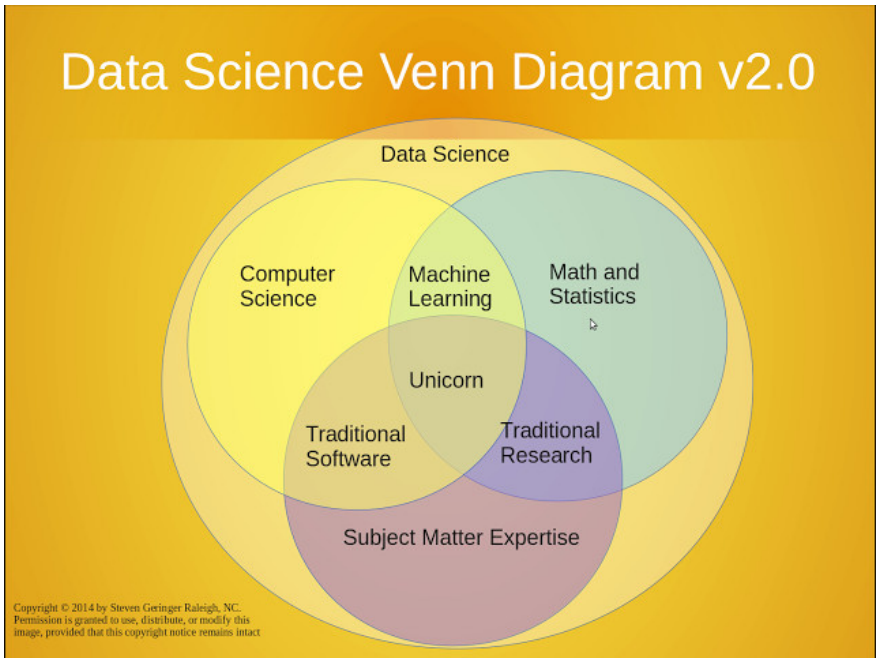
It is true that these two jobs have similar responsibilities. In some companies, Data Scientist may also be Data Analyst, or even be confused with both Machine Learning Engineer and Data Engineer.

Personally, I think Data Scientist divides into 2 main forms, temporarily calling A (Analysis) branch and Branch B (Building), specifically:

1. **Data Scientist A** (Analysis) **branch** is the thinker. Their main task is to analyze data using statistical methods to find value insight. Data Scientist branch A can also be called Data Analyst.
2. **Data Scientist branch B (** Building) is more powerful about software engineering. They are responsible for processing / storing data, writing code / algorithms for company data products.

If you need a narrow and specific definition for Data Scientist, then the Data Scientist branch B job description will be more accurate.

I myself belong to Data Scientist branch B, so all sharing will revolve around this branch.



What is the biggest difference between Data Scientist's A and B branches?

As mentioned above, Data Scientist branch B is more powerful about software engineering. Therefore, their main job responsibility is to build data products for the company.

1. Data products are also a software technology product, but are built on data. For example, Amazon's recommendation feature is a data product. It is built on the data base that Amazon has accumulated before. (What items did this user buy, what are the characteristics, similar items, items that should be purchased, items that other users have similar behaviors bought, etc.)

1. Data products can be a separate product, or part of a larger product. For example, the recommendation feature is a large product data product, the Amazon.com website.

1. Data products include many components, but there is always a core model (data model) developed by machine learning.

Can you explain in more detail the data model (model)?

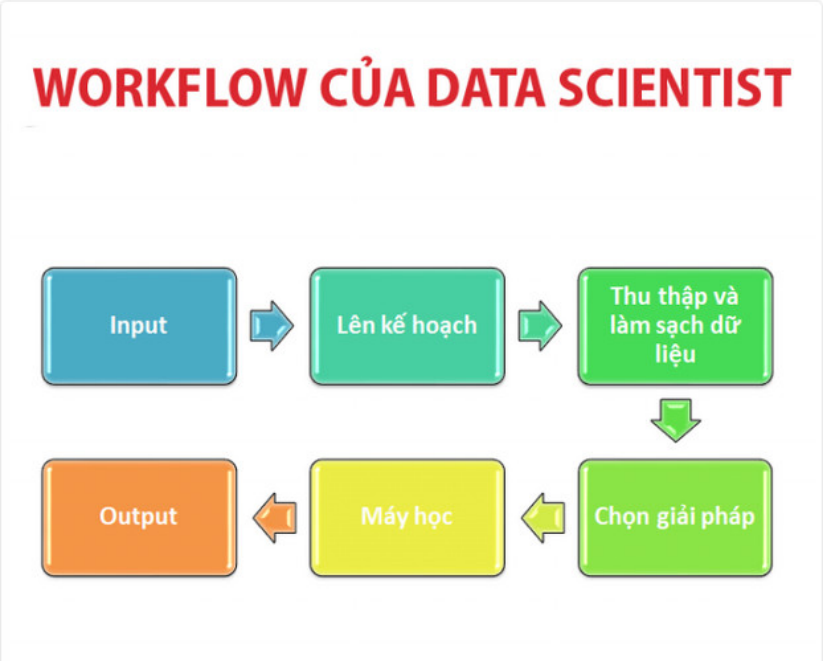
I talk about machine learning first!

For example, imagine the 'machine' here is a black box. You want to use this black box to distinguish the dog image from the cat. Then:

1. You have to find lots of pictures of dogs, and pictures of cats.
2. Then let the black box read these images.
3. Then teach the black box: which features on the picture will indicate that it is a dog image, and what other characteristics will indicate it is a cat image.
4. Finally, you give two new images. The black box will identify you where the dog is, and what the cat image is based on what it was learned.

This whole process is called machine learning. And the black box is a data model. Machine learning is an area of ??artificial intelligence, in which computer algorithms are used to self-learn based on input data without having to be specifically programmed.

What is Data Scientist's Workflow?



Step 1 - Input:

Data Scientist's Workflow starts with a need / task. For example, Google's image search needs: give the camera a picture, the results will return similar images.

This need may stem from:

1. Because the business department collects user feedback, and offers more ABC features.
2. Or, Data Scientist itself when working with data, studying product / company properties as well as the type / amount of data available . comes up with an initiative to develop the XYZ feature.

Step 2 - Plan:

After identifying the needs / tasks, Data Scientist will meet and discuss with the business department as well as the stakeholders to consider:

1. Is this feature feasible?
2. What kind of data will be needed? Where to find? How much is enough? How to get data ?. etc .
3. How many resources (manpower, time .)?
4. Where will this feature be embedded in the final product of the company, which will help users .

Step 3 - Collect and clean data:

To teach the machine how to distinguish the dog from the cat, for example, it must be given as many pictures as possible. Should have to gather data.

The collected data will still be very messy and junk, then we have to clean the data. Or if the data is not enough, you must make more.

For example: If there are pictures I don't need, remove them. The image I need but is blurry makes it clearer. Or a rough image (unassigned), label it. You can also find additional data sources that are open source and have

already been labeled.

Then must synchronize data.

For example, the image collected has many different sizes, it must be put in the same size or format, depending on the model you choose.

Step 4 - Choose solution:

1. If the problem is available solution

Then choose / combine the solutions again (eg choose the ABC or XYZ algorithm), run the test, check which test is the best and why, then choose which solution to further develop. etc .

1. If the problem is not available yet

We need to do research: find out before us, has anyone ever done this, what is their solution, is it feasible, is there any better solution, etc.

Then, choose one or more methods to test the same as above.

Step 5 - Machine learning:

After you have chosen a solution, you need to spend time with the machine.

Depending on what the model is, what tools are used, what systems the company already has, etc., that we will let the model run through the program, then adjust to control the output performance of that model.

When training a model, imagine like you have a control panel with a lot of buttons. Try tweaking this button a little bit, see the result a little better then keep it, then try adjusting the other button.

Just like that, until you get the best results.

For example, there are many factors to distinguish dogs from cats. It is up to you to adjust the machine to focus on which signs are more (the muzzle / areas that appear the muzzle, the color of the hair, etc.) It will prioritize those signs to better identify them.

Step 6 - Output:

The output of Data Scientist is a model as described above. Then, usually, this model will be attached to a large product.

For example: model to suggest purchase of Amazon website.

Sometimes, if the model is a new solution / innovation, your company's Data Science department will be responsible for writing articles or organizing scientific conferences to publish research results.

However, only a few large companies such as Facebook and Google . have specialized research departments on Data Science.

And in fact, it is very rare for innovations to be practical. Because many times, you create a good model, accurate but run too slowly, too costly resources, it can not be put to use.

What is the essential factor to become Data Scientist?

1. Be patient

This quality is extremely important, because Data Scientist has to spend most of the time collecting data and cleaning them.

For example, you want to make a model that predicts house prices.

You will have to collect data from home from various sources.

Each of these sources stores data in a separate structure. Then you have to attribute them to a common structure.

Then you clean up by removing inappropriate data, like:

1. Data missing: there are number of rooms without area.
2. Spam data: area of ??10m2 which cost 200 billion.

2. Good communication

Data Scientist's job requires a lot of communication, specifically:

1. Communicate with team business: To better understand product and requirements, find out valuable insights.

1. Communicate with team engineer: To apply your model to the system, or to ask them to organize their data / systems for use.

1. Present / explain insights to relevant parties to understand: To find ways to put into practice applications.

3. Like to learn and try new things

The Data Scientist profession is new and uses a lot of interdisciplinary knowledge. In particular, each industry has its own advance or new technology.

Therefore, you need to like to learn and try new things, to be able to update knowledge continuously.

What are the skills needed for Data Scientist?

The Data Scientist profession requires a lot of general knowledge and skills, including (but not limited to):

1. Machine learning: to learn data, thereby creating predictive models.
2. Database: helps store and retrieve data, as well as perform some calculations.
3. Programming languages: write code to apply the models learned above to specific products, or to manipulate the database.vv .
4. Visualization: help better understand data (eg how to distribute data), or to present analysis results.

Wow! Many qualities and skills need to be trained! So in it, what are the 3 most important factors?

According to me, the 3 most basic skills are also 3 basic factors to follow the profession of Data Scientist:

1. Mathematical knowledge

If you want to follow this industry, you need to know about math. Reason:

1. Machine learning is a combination of mathematical models running below. For example, when giving a machine a photo to distinguish the dog from the cat. Then that picture will be divided into several areas corresponding to 100 squares. Then you teach the machine that, in the picture, the cell in the cluster on the left corner has a lot of black, in combination with the cell in the cluster on the right corner, there is a lot of white, that is the trait of recognizing the dog.
2. When processing / working with data, you will need to use a lot of math, probability, statistics, etc. knowledge.
3. Math thinking will help you absorb and learn other skills faster.

2. Software programming ability

Data Scientist's work is very close to the work of a software engineer. Therefore, 'hard code' is an important requirement.

For example, making a tool to get data about . etc . **3. Sensitivity**

When looking at data, you need to be sensitive enough to guess: for this type of data, what can be done with it, how to estimate it, etc.

For example, with Amazon's data type, it is possible to build purchase suggestions for users.

This acumen is a quality, but can also accumulate over time and work experience.

What are the factors that determine your career? Data Scientist is what?

In addition to the above factors, you need to ask yourself more:

1. Do you like working with data (every day)?
2. Can you read the science paper without feeling it is a great barrier?
3. Do you like machine learning? (because things that seem interesting will often use machine learning to do it)

If the answer is 'yes', then you can pursue Data Scientist.

What is the knowledge in school to learn to be a Data Scientist?

A list of skills and knowledge to learn to become a Data Scientist is listed in detail at datasciencemasters.org, so you should consult.

In the university environment in Vietnam, I think you should study:

1. Linear algebra and statistical probability.
2. Derivative integral

You finished reading the article "**What is Data Scientist? How to become Data Scientist?**" edited by the **TipsMake** team. We hope this article has provided you with many useful tech tips and tricks. You can search for similar articles on tips and guides. Thank you for reading and for following us regularly.

© 2019 TipsMake.com