**What do you think being a data scientist is about?**

More generally, a data scientist is someone who knows how to extract meaning from and interpret data, which requires both tools and methods from statistics and machine learning, as well as being human. Data scientist spends a lot of time in the process of collecting, cleaning, and munging data, because data is never clean. This process requires persistence, statistics, and software engineering skills—skills that are also necessary for understanding biases in the data, and for debugging logging output from code.

**What do you see as the major duties and/or knowledge areas?**

Organizations today are grappling with how to make sense of an inordinate amount of disparate data. The ability to transform a sea of data into actionable insights can have a profound impact. By extrapolating and sharing these insights, data scientists help organizations to solve vexing problems. Combining computer science, modeling, statistics, analytics, and math skills—along with sound business sense—data scientists uncover the answers to major questions that help organizations make objective decisions.

**Data Scientist Role and Responsibilities:**

Data scientists work closely with business stakeholders to understand their goals and determine how data can be used to achieve those goals. They design data modeling processes, create algorithms and predictive models to extract the data the business needs, and help analyze the data and share insights with peers. While each project is different,  the process for gathering and analyzing data generally follows the below path:

1. Ask the right questions to begin the discovery process

2. Acquire data

3. Process and clean the data

4. Integrate and store data

5. Initial data investigation and exploratory data analysis

6. Choose one or more potential models and algorithms

7. Apply data science techniques, such as machine learning, statistical modeling, and artificial intelligence

8. Measure and improve results

9. Present final result to stakeholders

10. Make adjustments based on feedback

11. Repeat the process to solve a new problem

**Essential Data Science Skills:**

Most data scientists use the following core skills in their daily work:

·         Statistical analysis: Identify patterns in data. This includes having a keen sense of pattern detection and anomaly detection.

·         Machine learning: Implement algorithms and statistical models to enable a computer to automatically learn from data.

·         Computer science: Apply the principles of artificial intelligence, database systems, human/computer interaction, numerical analysis, and software engineering.

·         Programming: Write computer programs and analyze large datasets to uncover answers to complex problems. Data scientists need to be comfortable writing code working in a variety of languages such as Java, R, Python, and SQL.

·         Data storytelling:  Communicate actionable insights using data, often for a non-technical audience.

Data scientists play a key role in helping organizations make sound decisions. As such, they need “soft skills” in the following areas.

·         Business intuition: Connect with stakeholders to gain a full understanding of the problems they’re looking to solve.

·         Analytical thinking. Find analytical solutions to abstract business issues.

·         Critical thinking: Apply objective analysis of facts before coming to a conclusion.

·         Inquisitiveness: Look beyond what’s on the surface to discover patterns and solutions within the data.

·         Interpersonal skills: Communicate across a diverse audience across all levels of an organization.

What differences/similarities do you see between data scientists and statisticians?

|  |  |
| --- | --- |
| **Statistician** | **Data Scientist** |
| Data is given (i.e. test files) | Need to get data (i.e. database) |
| Focus on modeling | Focus on results |
| Bring data to model | Bring model to |
| Further away from product system | Embedded in production system |

Both statisticians and data scientists work closely with data. For typical traditional statisticians, the data set is usually well-formatted text files with numbers (i.e., numerical variables) and labels (i.e., categorical variables). The data’s size is typically small enough to be loaded in a PC’s memory or be saved in a PC’s hard disk. Comparing to statisticians, data scientists need to deal with more varieties of data:

* well-formatted data stored in a database system with a size much larger than a PC’s memory or hard disk.
* a huge amount of verbatim text, voice, image, and video;
* real-time streaming data and other types of records.

One unique power of statistics is to make statistical inferences based on a small set of data. Statisticians, especially in academia, usually spend most of their time developing models and don’t need to put too much effort into data cleaning. However, data becomes relatively abundant recently, and modeling is (often small) part of the overall effort. Due to open source communities’ active development, fitting standard models are not too far away from button-pushing. Data scientists in industry instead spend a lot of time preprocessing and wrangling the data before feeding them to the model.

Unlike statisticians, data scientists often focus on delivering actionable results and sometimes need to deploy the model to the production system. The data available for model training can be too large to be processed in a single computer. From the entire problem-solving cycle, statisticians are usually not well integrated with the production system where data is obtained in real-time, while data scientists are more embedded in the production system and closer to the data generation procedures. In summary, statisticians focus more on modeling and usually bring data to models, while data scientists focus more on data and usually bring models to data.

**How do you view yourself in relation to these two areas?**

According to me both Statistician and data Scientists are important roles. If someone wants to focus on significance, testing, experimental design, normality distribution, and diagnostic plotting, then become a Statistician. If someone wants to practice more software-engineering like coding and automation of machine learning models, then become a Data Scientist. I fell in love with mathematics when I was young pupil. It always amazes me that a series of numeric characters represent most of the problems in the world. I gained interest into Statistics when I was in 10th grade. So, with my love and affection towards Mathematica and statistics, I want to be a Statistician. However, I would like to explore Data Science knowledge areas with strong understanding of Statistics.