Data Science

What is data science?

This question previously had details. They are now in a comment.

100+ Answers

Luis Otavio Martins, Mechatronics Technician, Engineer and Data Scientist
Updated Apr 12 · Upvoted by Vishal Gonjari, MTech Electrical Engineering & Data Science, Veermata Jijabai Technological Institute (2017)

(As of 22 July, 2016) I've just left an interview where they asked me the same question. After reading the other 41 answers, I'll try to address a simpler and correct one.

WHAT IS DATA SCIENCE?

The first time that I heard about the term "Data Science" was in Toronto, 2015. Back then, I really though that Machine Learning and Data Science were some kind of universal panacea for the world's biggest problems. OMG, I was so fuc&^*%% wrong.

Related Questions

How can I become a data scientist?

What is the future of data science?

Is data science difficult?

What is the difference between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, and Big Data?

Is data science overrated?

What is a data scientist's career path?

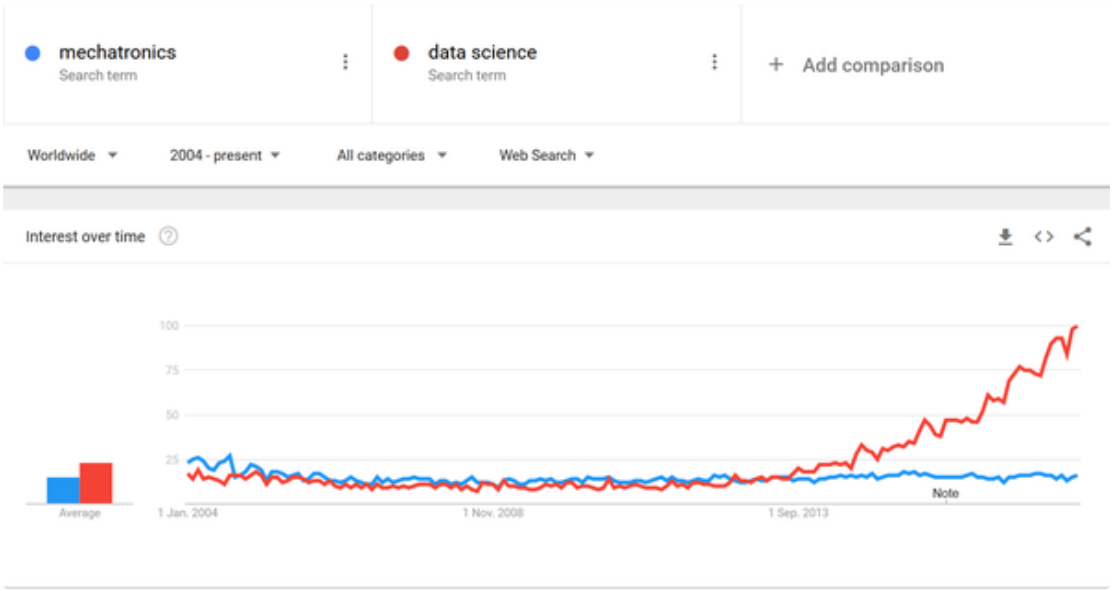
What are some good toy problems (can be done over a weekend by a single coder) in data science? I'm studying machine learning and statistics, ...

What are the best companies for a data scientist to work for?

+ Ask New Question

Our endeavor on this post will be to define and understand Data Science, so let's get some perspective. I have a background in mechatronics and mechanical engineering. Mechatronics is basically an attempt to understand every engineering piece, from electronics to robotics, from mechanical to computing and so on. You can imagine that I'm familiar to the struggles of Data Science. However, with Mechatronics I have a more intimate and older story. Everything started with my technical course on 2007. I thought that it was the universal panacea too. I'll leave up to you to guess if I was right or not.

DS vs Mecha (Nicknames = s2) Data Science versus Mechatronics



Ref.: Google trends

From this graph you can see in a glimpse that Data Science (red line) was not that famous on 2007, but surprisingly: "It was there! Wow!". Yeah, I know.

- **First conclusion**, Data Science on it's own it's older than what they are trying to make you believe.

More Related Questions

In other languages

En español: ¿Qué es la ciencia de datos?

Question Stats

905 Public Followers

268,137 Views

Last Asked Feb 21

24 Merged Questions

Edits

Stay with my train of thought towards our second conclusion.

Now you know that the term “Data Science” is coined before 2004. (unfortunately I just have Google Trends data after 2004 and it’s a post on its own to discover when happened the first mention or meaningful number of searches).

Now you may be wondering:

> How could people live 30 years ago without *Data Science* ?

The answer is siple: *they lived*.

And here it goes:

- **Second conclusion**, Data Science was *always* there but without all the buzz, officially coined term and rebranding. (I say *always* because we’re not talking about cavemen here, ok?)

After these two main conclusions. Let’s get something more dynamic. A picture is worth a thousand words (think in Data Visualizations).

What about a video? Take one minute of your life and watch the video below (I promise that will worth it).

https://www.youtube.com/watch?v=...
Yann LeCun was recognizing hand written digits in 1993 (I was 2 years old by then lol). This was not only about Machine Learning , it was required to have a specific database (MNIST) to train the model (not going to be technical here, more at: LeCun’s Demos).

Nowadays, if you need to create MNIST again from the NIST Special Database and a model to recognize digits, prototype the whole process into a product and

present to C-suite executives, you would be probably considering to hire a Data Scientist.

Data Scientists work with data. That's why we have Data Science. Simple enough.

What is Data Science?

1. It is a little bit of a *misnomer* and a *buzz word* that media is using to describe everything. However, it's good to have this dicussion to come into an agreement;
2. The question is about Data science. So I will not talk about Data Scientists. Go to What is a data scientist? if you are interested in that answer;
3. The biggest error that I found in most of the answers was some sort of:
    a. *"Data Science is when you are dealing with Big Data, large ammounts of data"*.
    b. That is not true, Data Science can be applied to a data set with one thousand lines, there is no problem with this.
4. Another misconception:
    a. *"Data Science only exists coupled with Machine Learning"*.
    b. That is not true. What about decades of statistical and mathematical modeling? It's true that Machine Learning empowers and automate Data Science, but it's not 100% necessary on 100% of use cases.
5. If we are going to call it "science" we need to consider the Science and Scientific Method definition. According to this, Data Science is not

only about the practical or empirical methods, it needs s...

(more)

Upvote · 423 Share · 1

Add a comment... Recommended All

Pronojit Saha, Data Aficionado.

Answered Jul 5, 2014 · Upvoted by Jalem Raj Rohit, Sr. Data Scientist at Episource

As this is a very open ended generic question, so would like to state a very broad answer for the same from my blog here: What is Data Science? by Pronojit Saha on Journey to planet Datum & Beyond

Data Science is the **practice** of:

1. Asking questions (formulating hypothesis), answers to which solve known problems or unearth unknown solutions that in turn drive business value,
2. Defining the data needed or working with an existing data set and employing tools (computer science based) to collect, store and explore such data generally in huge volume & variety (often more than 1 TB and 1000s of dimensions),
3. Identifying the type of analysis to be done to get to the answers and performing such analysis by implementing various algorithms/tools

(statistics based), often in a distributed and parallel architecture,

4. Communicating the insights gathered from the analysis in the form of simple stories/visualizations/dashboards (the Data Product) that a non-data scientist can understand and build conversation out of it. (It should be kept in mind that a product can also be an piece of code that is internal to a company and is used by various departments. The presentation, maintenance, scalability, etc of the code are then the product features, which is often not practiced in many organizations)
5. Building a higher level abstraction that does steps 2-3-4 in an autonomous way, analyzing & taking actions on new data as they are fed to the system.

45.4k Views · View Upvoters

Saurabh Singh, works at Jindal Steel & Power Limited

Answered May 31

Data science is the study of where information comes from, what it represents and how it can be turned into a valuable resource in the creation of business and IT strategies. Mining large amounts of structured and unstructured data to identify patterns can help an organization rein in costs, increase efficiencies, recognize new market opportunities and increase the organization's competitive advantage.

Breaking down 'Data Science': Data is drawn from different sectors and platforms including cell phones, social media, e-commerce sites, healthcare surveys, internet searches, etc. The increase in the amount of data available opened the door to a new field of study called Big Data or the extremely large data sets that can help produce better operational tools in all sectors. The continually increasing sets of and easy access to data are made possible by a collaboration of companies known as fintech, which use technology to innovate and enhance traditional financial products and services. The data produced creates even more data which is easily shared across entities thanks to emergent fintech products like cloud computing and storage. However, the interpretation of vast amounts of unstructured data for effective decision making may prove too complex and time consuming for companies, hence the emergence of data science.

How Data Science works: Data science incorporates tools from multi disciplines to gather a data set, process and derive insights from the data set, extract meaningful data from the set, and interpret it for decision-making purposes. The disciplinary areas that make up the data science field include mining, statistics, machine learning, analytics, and some programming. Data mining applies algorithms in the complex data set to reveal patterns that are then used to extract usable and relevant data from the set. Statistical measures

like predictive analytics utilize this extracted data to gauge events that are likely to happen in the future based on what the data shows happened in the past. Machine learning is an artificial intelligence tool that processes mass quantities of data that a human would be unable to process in a lifetime. Machine learning perfects the decision model presented under predictive analytics by matching the likelihood of an event happening to what actually happened at the predicted time.

So its very prominent that Data Science has a very promising future and has a lot of scope. There is a massive shortage of human resources in this field, and especially in India; it is estimated that by 2019, there will be a shortfall of 1.5 million data scientist. Bearing this in mind, both students and professionals are all able to have an edge over all other applicants if they leverage their degree or a certification on the same. Some of the courses worth mentioning are:

1. **Coursera-Data Science Specialization** : This Specialization covers the concepts and tools you'll need throughout the entire data science pipeline, from asking the right kinds of questions to making inferences and publishing results. In the final Capstone Project, you'll apply the skills learned by building a data product using real-world data. At completion, students will have a portfolio demonstrating their mastery of the material.
2. **Great Learning-Data Science & Engineering** : Great Lakes PGP-DSE (Full Time) is a 5-month full time program for fresh graduates and early career professionals looking to build their career in data science & analytics and move to roles such as business analysts, data analysts, data engineer, analytics engineer etc. by learning relevant data science techniques, tools and technologies, and hands-on application through industry case studies.

a. **World Class Faculty:** You gain from the decades of experience and expertise brought to the table by Great Lakes faculty in their chosen domains. Our faculty comes from leading international and national schools such as Harvard, Stanford, Kellogg, University of Chicago, IIMs, and IITs.

b. **Industry Mentors:** Current industry knowledge and insights from industry practitioners in knowledge sharing sessions and guest lectures allow you to stay ahead in the industry.

c. **Hands on learning:** Candidates get hands-on exposure to data science tools & techniques such as R, SAS, Python, Tableau, MySQL, Hadoop, Spark. Industry case studies aid in effective application.

d. **Capstone project:** An industry oriented 2 month project gives candidates the much needed practical exposure they require before sitting for placements.

e. **Placement assistance:** In this program, candidates who clear the program are also offered placement assistance. Great Lakes, being one of India's top 10 business schools and a leading institution when it comes to analytics education, would use its corporate network to help candidates in the program make the transition to careers in analytics. All qualifying candidates will be extended Placement assistance until 3 months post completion of program.

3. **Microsoft- Professional Program for Data Science** : Microsoft consulted data scientists and the companies that employ them to identify the core skills they need to be successful. This informed the curriculum used to teach key functional and technical skills, combining

highly rated online courses with hands-on labs, concluding in a final capstone project.

4. **edX- Data Science Course** : Multiple course programs exist to get you on a path to a job as a data scientist. The Micro Master's program teaches you essential Python programming needed to perform data tasks and explores machine learning and big data analytics using Spark. And completing a Micro Masters can jump start a data science degree or data science masters. The Programs feature multi-course tracks designed to give you in-depth knowledge and training.

I would personally recommend **Great Learning's Post Graduate Program in Data Science & Engineering** considering Great Lakes being the no. 1 analytics school in the country with great industry connects and the highest no. of transitions of alumni into Analytics.

All the best and I hope this answer was useful.

1.6k Views · View Upvoters

Add a comment... Recommended All

Rahul Agarwal, Senior Statistical Analyst at Walmart Labs
Updated Dec 24, 2015

Data science is a form of **Rebranding** that was needed.

Before the Data Science trend, people who used to work with data were called Business Analysts. Statisticians. Computer Scientists. Data analysts. _____.____(Put More data titles here).

Now they are called **Data Scientists.** That's it.

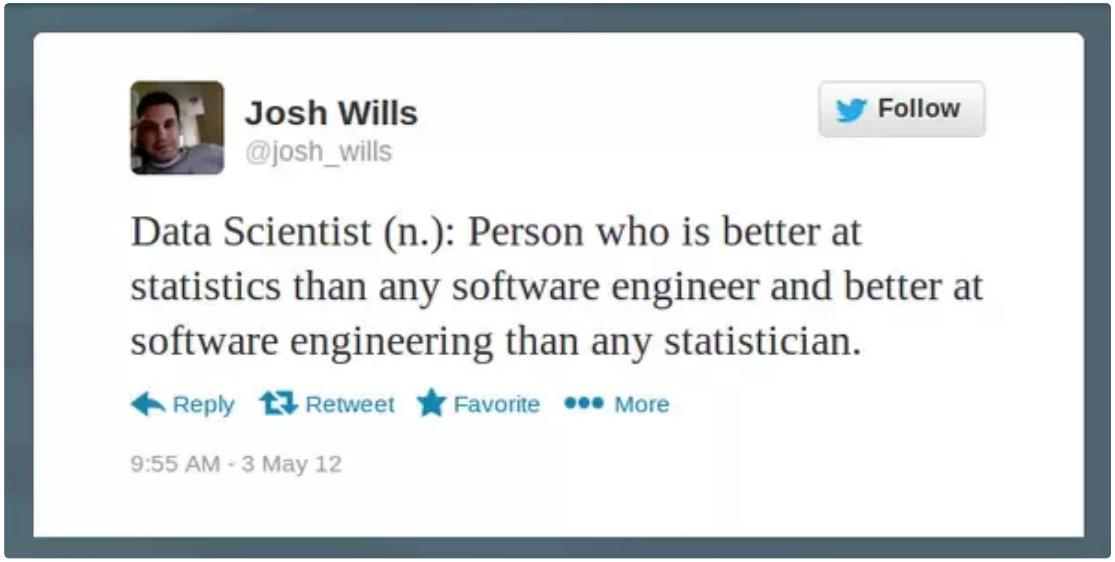
The Need of the Rebranding:

Somewhere along the way we recognized the power data could bring. The volume, velocity and variety of data needed people who could handle such sort of data.

Now a basic problem arised when people who were statisticians were not able to handle these copious large amounts of data. While people who were Computer

scientists were not able to formulate proper analysis on the data to get good value out.

This led to the term of Data Scientist, which people have explained in many different ways. One I like the most is:



Now where do you get such people? You need to create such Job Titles which demand these skills. And hence the Rebranding took its roots.

The Skillset Required:

Skillset required for Data Science position may vary a lot based on different organizations. I like to bifurcate it into Fundamental knowledge and Tool based knowledge.

1. Fundamental Requirements:

a) Linear Algebra: To understand various algorithms.

b) Multivariable Calculus: To understand various algorithms.

c) Probability and Statistics: To understand various algorithms.

d) Coding in at least one Language preferably Python/R: Scripting and Development.

e) SQL and Excel: Probably the two tools you will use the most.

f) Knowledge of machine Learning algorithms:

Rahul Agarwal's answer to What are the top algorithms that every data scientist should have in their toolbox?

Rahul Agarwal's answer to How smart do you need to understand the algorithms and maths behind machine learning?

2. Tool Based Requirements:

a) Scikit-Learn/Caret for Machine Learning: Rahul Agarwal's answer to What are the best data science MOOCs?

b) Hadoop/Spark for Big Data: Rahul Agarwal's answer to How do you learn big data?

c) I am recently seeing job openings that demand Javascript and knowledge of Web frameworks for Data Science jobs.

What Data Science is Not:

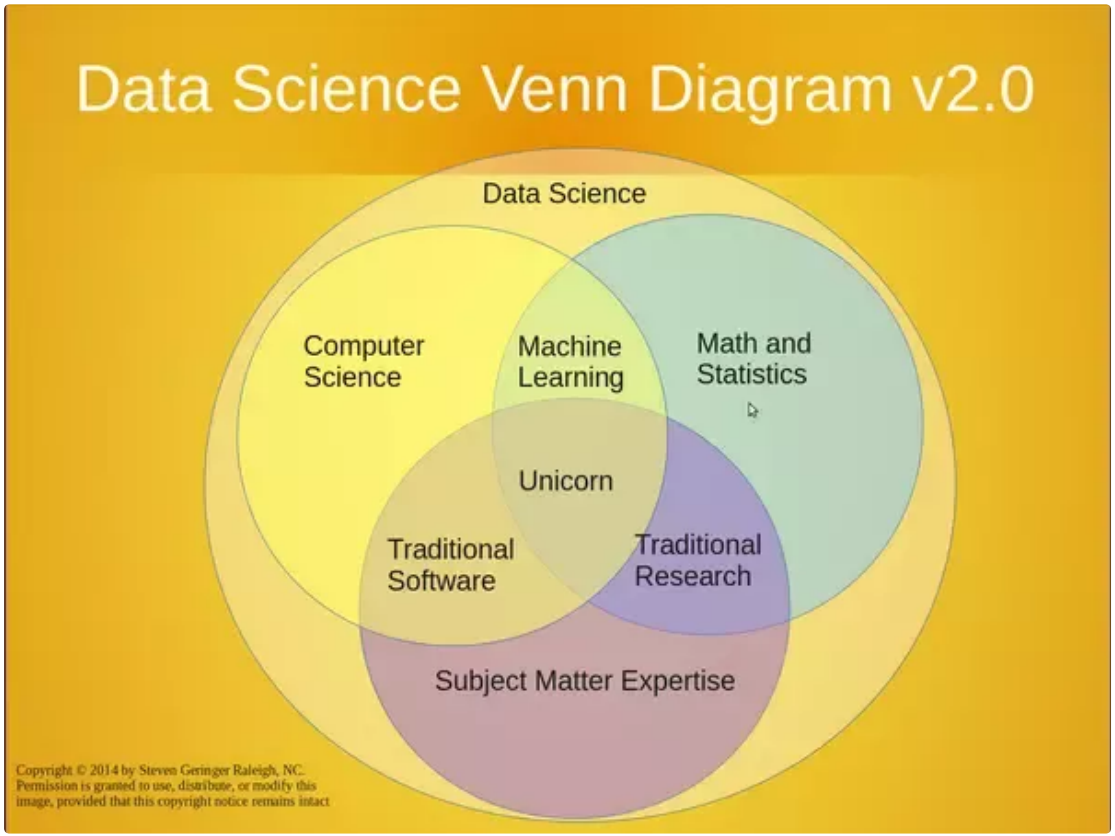
1. Data Science is **not Magic:** although it might feel like that to some. You don't get predictions using a crystal ball you get predictions using data. And wherever

data is involved, no magic is involved. Data Science is just a way to take Data driven Decisions. **Data Science alone won't solve all of your problems.**

2. Data Science is **not Easy**: There are a lot of additions happening nearly everyday to the field that would require people to read ad learn a lot everyday. You need to learn about old algorithms, **you need to learn about new algorithms and then you would have to continue working side by side.** This is not to discourage people, actually it is one of the things that attracts me a lot towards this field. I feel fascinated about the kind of learning opportunities and scope in this domain.

3. Data Science is **not a Fad**: Rahul Agarwal's answer to Is data science a fad? How long will it continue to grow as a field?

4. Data Science is **a Unicorn:** Actually nobody quite knows what the exact definition of data Science is. To give you a perspective here is **Drew Conway 2.0** for you(not made by Drew Conway). Everyone has his own opinion of data science.



5. Data Science is **not Sexy:** Data science has been tagged particularly as the sexiest job of the 21st century by Hal Varian. I disagree. **Data Science is tedious.** A lot of time is used up in data wrangling, tuning and other mundane stuff.
HBR: The Sexiest Job of the 21st Century is Tedious, and that Needs to Change

Although I would still call it the **most experimental/ interesting** job out there. The possibilities of what you could do with the data are still endless: Rahul Agarwal's answer to What are some best practices in Feature Engineering?

6. Data Science itself is **not predictable:** Getting insights from data needs creating hypotheses and then work towards disproving them. And most of them

actually end up getting disproved. There would be issues with your data and it might take time to resolve them. You may go days without getting a valuable valid hypothesis. What this means is that It would take time to gain value from your data and you cannot delve on deadlines. One of the profs of CS109, a course I highly recommend said: "under promise and over deliver" is one of the basic trait of a data scientist.

Roland Kofler
you start with the Rebranding argument and then you explain that its something total...

Sandeep Dayananda, Data Science & Machine Learning Enthusiast
Answered Dec 6, 2017

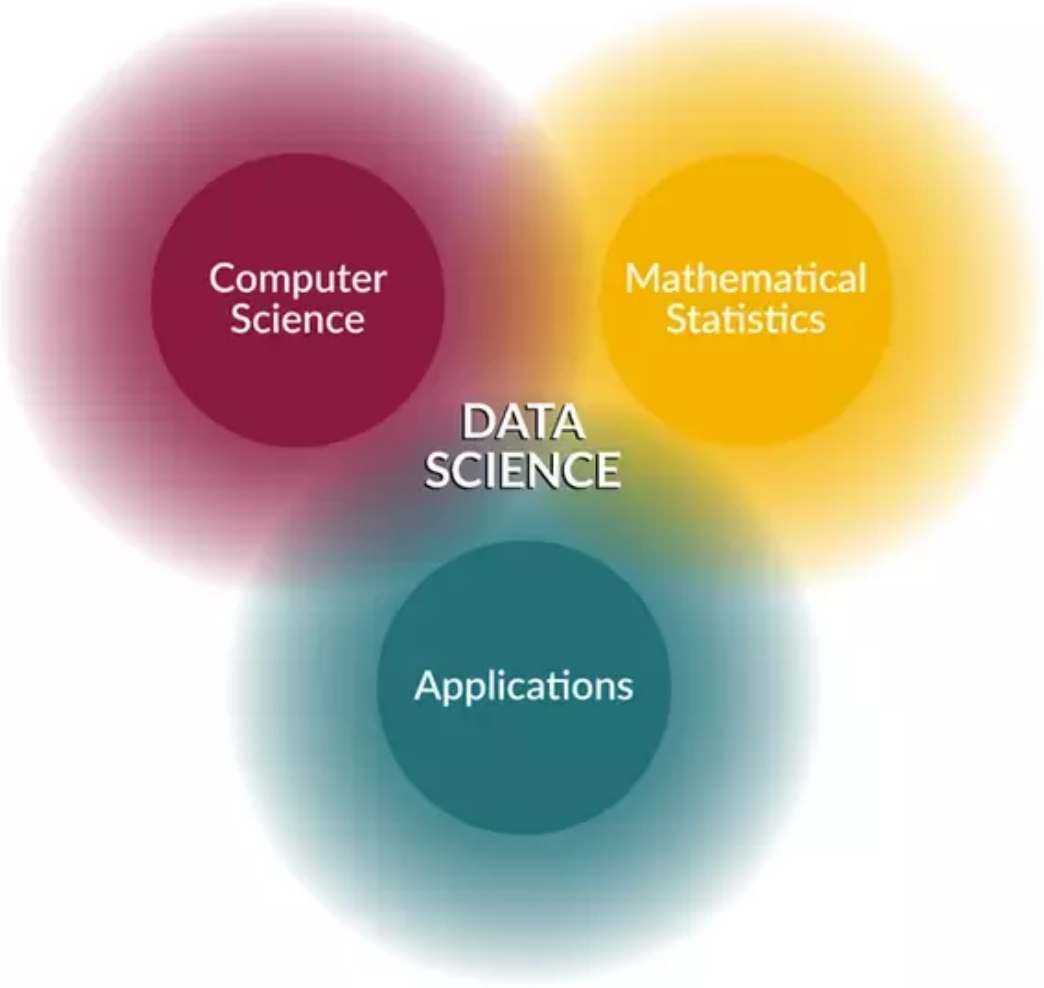
I am happy to share my knowledge on Data Science. I have got to work on Data Science through projects in R and Python Programming. To help you out with Data Science, let me give you a quick guide into what Data Science is.

WHAT IS DATA SCIENCE?

Data Science involves using automated methods to analyze massive amounts of data and to extract knowledge from them.

There are 3 important sciences which are form Data Science. These are:

1. Computer Science
2. Mathematical Statistics
3. Applications



It is the combination of all the 3 sciences and every Data Science project involves using them to attain the results required. By combining aspects of statistics, computer science, applied mathematics and visualization, data science can turn the vast amounts of data the digital age generates into new insights and new knowledge.

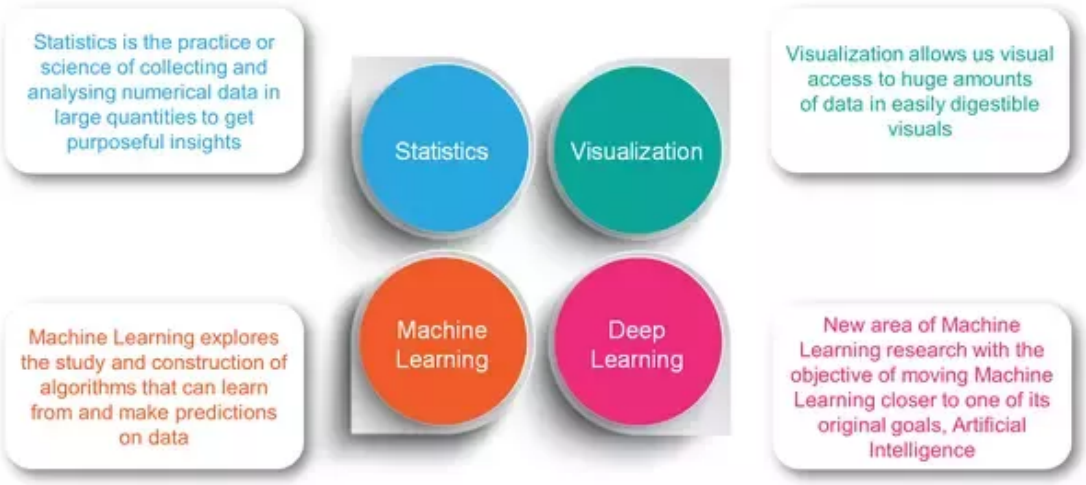
*Go through our blog on **What is Data Science** to completely understand the important concepts involved.* I would like to recommend Edureka's blogs on Data Science. Check out these blogs which have all the basic content required to clear a Data Science Interview.

1. Top 35 Data Science Interview Questions
2. Data Science Tutorial
3. Edureka's Data Science Blogs

*So after you go through these, you can also check out our **Data Science Training** YouTube playlist.*

COMPONENTS OF DATA SCIENCE

Data Science has the following components:



1 . STATISTICS

1. Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation and organization of data.

2. Statistics began in the ancient civilization, going back at least to the 5th century BC, but it was not until the 18th century that it started to draw more heavily from calculus and probability theory.

Figure: *Concepts of Statistics*

2 . VISUALIZATION

Visualization is when we display the results of Data Science analysis in a simpler way using diagrams, charts and graphs.

It improves decision making, sense of work, customer relationship and financial performance.

3. MACHINE LEARNING

1. Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data.
2. Closely related to computational statistics.
3. Used to devise complex models and algorithms that lend themselves to a prediction which in commercial use is known as predictive analytics.

4. DEEP LEARNING

Deep learning is one of the only methods by which we can circumvent the challenges of feature extraction in machine learning. This is because deep learning models are capable of learning to focus on the right features by themselves, requiring little guidance from the programmer.

Therefore, we can say that Deep Learning is:

1. A collection of statistical machine learning techniques
2. Used to learn feature hierarchies
3. Often based on artificial neural networks

JOB ROLES IN DATA SCIENCE

There are 8 major job profiles available for anyone who is interested to work in Data Science. They are the following:

I think by this, you have learnt about the fundamentals of Data Science. *You can go through my blog on* ***Data Science Interview Questions*** *to go through the top interview questions you will face in an interview.*

If you are more convenient watching a video, check out our YouTube tutorial on Data Science below:

I hope my answer helps!

Do check out our Data Science Certification Training to learn more about Data Science and Machine Learning algorithms.

2k Views · View Upvoters

Upvote · 6 Share

Add a comment... Recommended All

Michael Hochster, Director of Data Science at Stitch Fix
Answered Jan 17, 2014 · Upvoted by Alex Blocker, Staff Statistician at GRAIL, Xoogler, statistics PhD and William Chen, Data Science Manager at Quora

Data Scientists are people with some mix of coding and statistical skills who work on making data useful in various ways. In my world, there are two main types:

Type A Data Scientist: The A is for Analysis. This type is primarily concerned with making sense of data or working with it in a fairly static way.The Type A

Data Scientist is very similar to a statistician (and may be one) but knows all the practical details of working with data that aren't taught in the statistics curriculum: data cleaning, methods for dealing with very large data sets, visualization, deep knowledge of a particular domain, writing well about data, and so on.

The Type A Data Scientist can code well enough to work with data but is not necessarily an expert. The Type A data scientist may be an expert in experimental design, forecasting, modeling, statistical inference, or other things typically taught in statistics departments. Generally speaking though, the work product of a data scientist is not "p-values and confidence intervals" as academic statistics sometimes seems to suggest (and as it sometimes is for traditional statisticians working in the pharmaceutical industry, for example). At Google, Type A Data Scientists are known variously as Statistician, Quantitative Analyst, Decision Support Engineering Analyst, or Data Scientist, and probably a few more.

Type B Data Scientist: The B is for Building. Type B Data Scientists share some statistical background with Type A, but they are also very strong coders and may be trained software engineers. The Type B Data Scientist is mainly interested in using data "in production." They build models which interact with users, often serving recommendations (products, people you may know, ads, movies, search results).

At Google, a Type B Data Scientist would typically be called a Software Engineer. Type B Data Scientists may use the term Data Scientist to refer just to themselves, and since the definition of the field is very much in flux, they may be right. But I see the term being used most often in the general way I am

proposing here.

This categorization is crude. Many Data Scientists are some mix of A and B. But this answer is long enough already.

81.7k Views · View Upvoters · Answer requested by William Chen

Upvote · 794 Share · 2

Alex Blocker, Staff Statistician at GRAIL, Xoogler, statistics PhD
I like this classification, but I would emphasize "Many Data Scientists are some mix of...

Md Akash, worked at Digital Marketing
Answered May 20

Data science is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining. Data science is focused on extracting insights from data of any form or shape using a multitude of disciplines and technologies for the purpose of creating new data products and services or improving the existing ones. There are three key terms here:

Data, not just structured data and not just big data.

Multitude of disciplines and technologies including predictive analytics (PA), machine learning (ML), AI, math/statistics, pattern recognition, programming, databases, algorithms, and big data techniques.

Data products/ services highlights the ultimate goal of data science .

Data Science as a practice is nothing new and has been practiced in the last three decades under a variety of names including data mining, advanced analytics, knowledge discovery, and predictive analytics to name a few. Only the

term itself is considered new and the increased level of interest in it in the recent years due to explosion of data.

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.

When Harvard Business Review called it "The Sexiest Job of the 21st Century" the term became a buzzword, and is now often applied to business analytics, or even arbitrary use of data, or used as a sexed-up term for statistics. While many university programs now offer a data science degree, there exists no consensus on a definition or curriculum contents. Because of the current popularity of this term, there are many "advocacy efforts" surrounding it

Top Stories from Your Feed