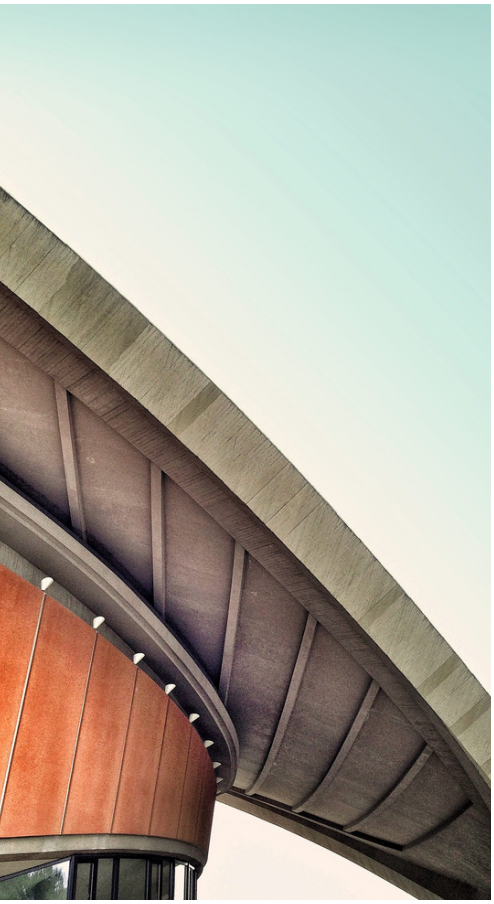
DATA SCIENCE

35 KEY DATA SCIENTIST SKILLS

YOU NEED TO SUCCEED

INTELLSPOT.COM



INTRODUCTION

The fact that big data science is one of the highest paid professional areas to get into, means you need a long list of data scientist qualifications and skills.

No matter if you live in the USA, Canada, UK, Australia, India or somewhere else, the minimum set of required skills (such as technical knowledge, software, math, and statistics) is not enough to truly succeed and to get above average median base salary – $110,000 per year.

You need a broad range of behavioral characteristics, traits, qualifications, knowledge, certificates, and understandings to be a professional who is able to bring accurate data insights for decision-making in an organization.

IN THIS WHITEPAPER:

35 key data science qualifications and skills

PROGRAMMING LANGUAGES

- R
- PYTHON
- JAVA
- SQL
- SAS
- MATLAB
- SCALA
- JULIA

When it comes to the fundamentals of data science, the programming skills are in the top place.

In their work, data scientists use several programming languages and software solutions to perform activities such as extracting, analyzing, cleaning, and visualizing data.

Here is a list of the 8 most popular programming languages and skills needed for data science:

1 R

R is a programming language – an excellent tool for unlocking the patterns in large datasets. It provides a great range of high-quality and open source packages that come with in-built statistical analysis methods and functions.

In addition, R allows you to manage matrix algebra very well. Also, good data visualization capabilities are a key characteristic of R.

2 PYTHON

Python is an absolute hit – a general purpose language, that is broadly used by the data science community.

Its wide range of purpose-built modules and active community support make Python a very popular and mainstream programming language.

Python is a very easy language to learn. This makes it a perfect first language for the newbies in programming.

With the boom in technologies like machine learning and artificial intelligence, the need for data scientists with a solid knowledge of Python skills is in a rising demand.

3 JAVA

This extremely popular general purpose language allows integrating data analysis methods directly into the codebase.

A large number of companies build their modern applications and system solutions upon a Java back-end.

Java is also relatively simple and easy-to-learn programming language. Java is suitable for writing intensive machine learning algorithms and ETL production code. Reliability, practicality, and compatibility are also some of the key Java benefits.

4 SQL

You might know that more than 50% of data scientist job listings on LinkedIn require expertise in SQL.

It is because SQL is efficient at querying and manipulating relational databases.

It lies at the heart of storing and retrieving data in an organization. SQL deals greatly with large databases, providing a fast processing time.

5 SAS

SAS is also one of the most popular programming languages in the data science world.

It provides a great range of statistical functions with a user-friendly GUI that helps you learn quickly.

As SAS is easy-to-learn language, it is very preferred from the beginners in the analytics industry.

6 MATLAB

MATLAB is a numerical computing language that is a fast, stable, works with solid algorithms for complex math and has a place in a lot of applications.

MATLAB is known as a hard-core language among mathematicians, data scientists, and scientists who deal with sophisticated systems.

7 SCALA

Scala is one more Java-based programming language and is becoming a preferred weapon for those doing machine learning at high-volume data sets and creating high-level algorithms.

The code written on Scala runs practically anywhere that Java runs.

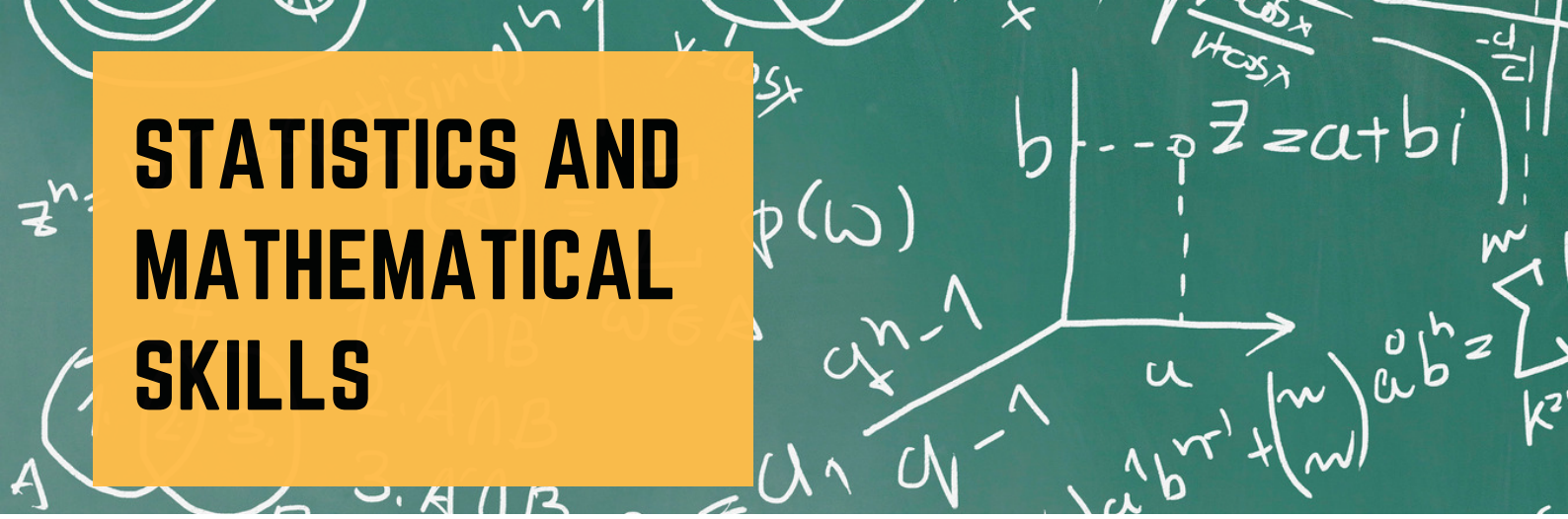
It makes Scala a vigorous general purpose language that is very good for data science.

8 JULIA

Julia is a kind of a newcomer to watch when it comes to modern data scientist qualifications and skills.

It is a high-level dynamic language for programming that aims to meet the needs of high-performance numerical analysis.

In fact, Julia is an impressive language that gains a great popularity amongst the data scientists and is adopted by some major companies including many operating in the finance industry.



STATISTICS AND MATHEMATICAL SKILLS

Discrete vs. continuous data

Binomial distribution

Regression

Hypothesis testing

Bayesian thinking & modeling

Machine Learning

Markov chains

Although the today software can fulfill all the necessary statistical activities, you still need math and statistical acumen and understandings to know which test to perform and how to interpret the results.

Here are some basic data scientist qualifications and skills required in the field of statistics.

9 Discrete vs. continuous data

Both discrete and continuous variables are the two types of quantitative data also called numerical data.

In practice, many data mining and statistical decisions depend on whether the basic data are discrete or continuous.

To know what is the difference, see our post discrete vs continuous data – with a comparison chart.

10 Binomial distribution

It is not too much to say that the path of mastering statistics and data science starts with probability. Solutions to many data science situations are often probabilistic in nature.

The binomial concept has its core role when it comes to defining the probability of success or failure in an experiment or other data science events.

Our post (binomial distribution examples) contains a lot of basic information.

11 Regression

Non-linear and linear regression models as the oldest and widely used supervised machine learning algorithms for predictive analysis have a lot to do with your professional development.

Linear regression modeling and formula have a range of applications in the business data science.

For example, they are used to evaluate business trends and make forecasts and estimates.

They can also be used to analyze the result of price changes on the consumer behavior.

12 Hypothesis testing

Hypothesis-driven thinking in data science is something you should be familiar with.

The statistical process for doing a hypothesis test is to set out hypotheses and to perform an appropriate statistical test to reject or accept the hypothesis.

13 Bayesian thinking & modeling

Bayesian modeling is an extremely powerful suite of tools for modeling any random variable, such as a business KPI and demographic statistic.

The popular Bayes theorem defines the probability of an event to occur based on the prior information about the conditions that might be related to that event.

Data scientists provide their understanding of a particular problem and some data, and as a result get a quantitative measure of certainty of a specific fact.

14 Markov chains

Markov chains are simple methods to model random processes in a statistical way.

They are a popular way of learning data science techniques and probabilistic modeling.

This was a main run-down of some core areas that can help a data scientist beginner have a better understanding of what statistical knowledge is expected of him/her.

15 Machine Learning

Machine Learning is a constantly growing area that is used when credit scoring, placing ads, stock trading and for many other purposes.

Machine learning consists of developing, testing, and applying algorithms for predicting future outcomes using data.

For more information, see our post supervised vs unsupervised algorithms.

Supervised and unsupervised learning represent the two key methods in which the machines can automatically learn and improve from the experience.

Nowadays, there are many easy, fast, affordable or free ways to learn Machine Learning for beginners and advanced learners.

SOFTWARE AND ANALYTICAL TOOLS

- KNIME
- Weka
- RapidMiner
- Apache Hadoop
- Apache Spark
- Data Melt
- Apache Storm
- TensorFlow

Since the data science area is enormous right now, what software tools the different companies use vary significantly.

Some data scientists provide data cleaning services.

Other data scientists do specific researches.

Nevertheless, there are some essential software tools – an integral part of the list of data scientist qualifications and skills.

16 KNIME

KNIME is a popular software company that offers an open source analytics platform (KNIME Analytics Platform) for data reporting, data mining, and predictive analysis.

KNIME Analytics Platform helps data scientist all over the world discover the potential hidden in the data, gain fresh insights, or predict new futures.

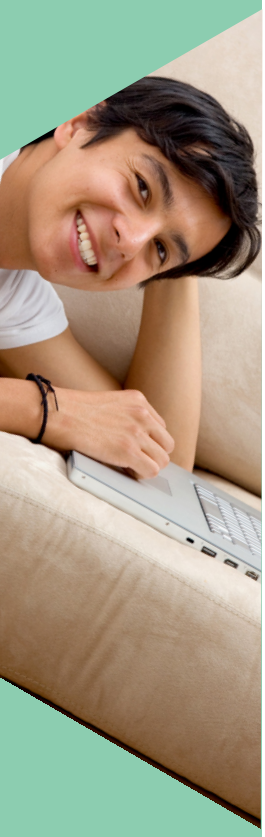
With more than 2000 modules, a great range of integrated tools, and a big list of advanced algorithms available, KNIME Analytics Platform is the perfect toolbox for any data scientist.

17 WEKA

Weka is machine learning software provided by The University of Waikato.

Weka combines machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Weka is open source software issued under the GNU General Public License.



18 RapidMiner

RapidMiner is a software platform for data scientists to help you build predictive models faster. The tool unites data prep, machine learning, and predictive model deployment.

The base of the platform (RapidMiner Studio) is a free and open-source platform. There also are enterprise-level solutions priced by the number of logical processors, the amount of data used, and productivity features.

19 Apache Hadoop

Apache Hadoop software library (one of the best big data tools) is a framework, written in Java, for processing large and complex datasets.

Hadoop is a highly scalable platform. It can store and easily distribute very large data sets over hundreds of servers. Hadoop services have it all: data storage, data processing, governance, operations, and security.

20 Apache Spark

Apache Spark is a unified analytics engine for large-scale data processing, a cluster-computing framework for data analysis. Fast, flexible, and data scientist-friendly, Apache Spark is a leading platform for large-scale SQL, batch, and streaming data processing, and machine learning.

21 Data Melt

DataMelt is a free graphing software for scientists, engineers, and students. It can be used for numeric computation, statistics, symbolic calculations, data analysis and data visualization.

Additionally, it provides advanced mathematical calculations, statistical analysis, and data mining capabilities.

22 Apache Storm

Apache Storm is a free and open source distributed platform for real-time analytics. Storm makes it easy to reliably process unbounded streams of data, doing for real-time processing what Hadoop did for batch processing. Storm is simple, can be used with any programming language, and is a lot of fun to use!

23 TensorFlow

TensorFlow is an open source machine learning framework for everyone – from students and researchers to data science professionals and innovators. It is a software library for high-performance numerical computation. It allows you to access the big power of deep learning without even understanding the complicated principles behind it.

DATA VISUALIZATION SKILLS

TABLEAU

GOOGLE CHART

D3.JS

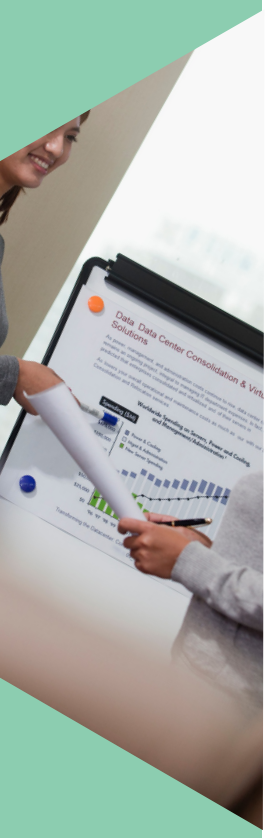
JUPYTER

ORANGE

It is impossible to develop data scientist qualifications and skills without including data visualization. In the end, data scientists are those who help others make data-driven decisions.

Pictures can communicate much more effectively than words, so it is a must a data scientist to present data in a visually compelling way.

This means you not only have to use data visualization tools but also understand the principles of visualizing data effectively.



24 Tableau

Tableau is one of the most popular data visualization and dashboarding tools out there.

Tableau is a business intelligence software that helps people see and understand their data.

It allows you to connect and visualize your data in minutes, combine multiple views of data to get richer insight, and share dashboards on the web.

25 Google chart

Google chart is a powerful, simple to use, and free solution for visualization of big data.

It is totally free and has a great support from Google.

Google chart provides a wide variety of charts.

From simple scatter plot to hierarchical treemap and multi-dimensional interactive matrixes, you can find many professional and cool ways to show data.

26 D3.JS

D3.js (stands for Data-Driven Document) is a JavaScript library for manipulating documents based on data. It is an open source library for interactive big data visualization.

D3.js is a very powerful solution, but to use it you need a solid understanding of Javascript to operate with the data and present it in an appropriate form.

27 JUPYTER

JupyteR is an open-source project allows you to analyze, visualize and real-time collaborate on software development across a variety of programming languages. Project Jupyter is a non-profit project born to support interactive data science and scientific computing across all programming languages.

28 ORANGE

Orange is an open source data visualization and machine learning solution for everyone – from newbies to experts. It makes data science fun and interactive. Orange allows you to perform clever data visualization, explore statistical distributions, box and whisker plot and scatter plots, decision trees, hierarchical clustering, heatmaps, linear projections and many more.

NON-TECHNICAL SKILLS

- COMMUNICATION SKILLS
- DATA-DRIVEN DECISION MAKING
- DOMAIN KNOWLEDGE AND BUSINESS ACUMEN
- TEAMWORK
- INTELLECTUAL CURIOSITY AND PASSION FOR WORK
- GOOD DATA INTUITION
- PROJECT MANAGEMENT SKILLS

Guess what! Some of the best data scientists are not engineers, not statisticians, not programmers, and even not computer specialists. Many of them have a degree in other areas such as Psychology, Marketing, Economic, and etc.

Although, data science is a technical field and you need strong technical data scientist qualifications and skills, there are many other “non-technical” abilities that can set you apart from a regular data workers.

Here are some key non-technical skills you surely need as a data scientist.

29 Communication Skills

Working as a data scientist means working with other members of a team – for example, working with product managers, engineers, designers, stakeholders, and more.

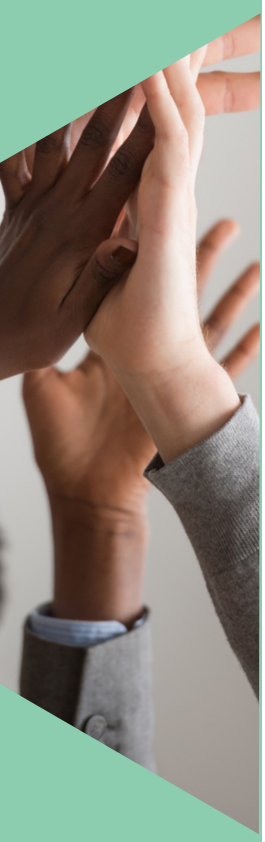
Communication skills can help you build trust and understanding, which is incredibly important for those being stewards of the data.

In addition, you must be able to report your technical findings to non-technical colleagues such as those from the marketing department.

Actually, one of the critical skills to have is effective business communication.

A data scientist needs to be very persuasive, especially when reporting his/her findings.

All of the statistical and mathematical computation can be useless if the data scientist can't communicate insights properly.



30 Domain knowledge & business acumen

As data scientist aims to make a positive difference in business, expertise in developing algorithms and building statistical models is not enough to achieve this.

You must know the business relevance of the algorithms and statistical models at hand.

Understanding the fundamentals of the industry you are operating in and objectives of your company only can help to achieve the expected results, making a difference in the today's competitive market. The technical skills without understanding the business context are unthinkable.

31 Teamwork

Every data scientist should be a team player. They are deeply involved in a company at different levels.

As a data scientist, you need to collaborate with the team members to understand their needs, requirements, feedback in order to achieve the best use of data insights that you draw for the needs of various departments.

In today's world of connected people, strong teamwork skills are an essential centerpiece of digital change.

32 Intellectual Curiosity and Passion for Work

If you wish to become a data scientist, you need to understand data, and to know why and how to eliminate the present mess in the big data – from both structured and unstructured data.

You need to come out with insights and solutions data that makes sense which is essential for preparing valuable business reports.

It means you not only need knowledge, but also a passion for your work and for finding patterns and answers to business problems.

Sometimes, you might not even have a clear problem or situation to work on it, just signals that there is something wrong.

This is where your intellectual curiosity should help you observe new or complicated areas to find what is going on.

33 Good data intuition

Data intuition is perhaps one of the most valuable non-technical data scientist skills that you can possess.

It involves seeing patterns where none are observable on the appearance. Data intuition can make you much more efficient in your career. However, this is a skill which comes with experience and years working in the world of the data science.

You need to look at a vast number of data sets, play with it, catch what it is going on, and understand the kinds of the problems you need to resolve.

Once you've exposed yourself to enough data work, you'll develop your data intuition to know the right questions and find the right answers.

34 Data-driven decision making

There are so many questions and problems, a data scientist needs to answer and decide. For example, what tools and qualitative data analysis methods to use, how to visualize and communicate data in the best way.

Today, the whole management world talks about how to create successful data-driven decision-making process and models in business to improve results.

35 Project Management Skills

Most data analytics work is project based and have to be handled effectively.

You might know that there is a high degree of uncertainty in each project.

For example, the uncertainty might relate to what the purpose of the project is, whether enough data is available to resolve the problem, methods that need to be applied to achieve the purpose and etc.

All this means that it is so hard to create the right plan at the beginning of the project.

That is why project management, is a key skill in data analytics.

Unfortunately, project management skills are very underrated and most data scientists lack them.

Conclusion

The data scientist qualifications and skills can be grouped in different ways. In fact, the data scientist job consists of a variety of positions, which require very different abilities and understandings.

> The above-listed skills are essential and form the fundamentals of your development as a professional.

The first two groups of skills: programming and statistical skills are perhaps what most people first think about when they consider the data scientist role and position.

While those are truly important and build the technical background of your knowledge, it is crucial to note that not-technical skills are also so important and required.

Every decade has its top job opportunities and hottest science fields. Today, it is more than clear that big data, artificial intelligence, and machine learning are a key success factor that can define whether businesses are successful or not.

Hopefully, this round-up will help you create a clearer picture for you and help you understand the core skill set that employers require today.

DATA SCIENCE

THANK YOU

AND BEST WISHES ON YOUR OWN PATH TO DEVELOP YOUR DATA SCIENTIST QUALIFICATIONS AND SKILLS.

INTELLSPOT.COM