|  |  |
| --- | --- |
|  | acuitykp.com |
|  |  |
|  | |
| Data Science | |
| **A step by step guide on how to become a successful Data Scientist** | |
| October 2021 | |

Contents

[**Table Identification** 3](#_Toc74148395)

[**Deep Learning model Architecture** 4](#_Toc74148396)

[**Evaluation Metrics** 5](#_Toc74148397)

[**Dataset** 6](#_Toc74148398)

[**Data annotation** 6](#_Toc74148399)

[**Training Dataset** 6](#_Toc74148400)

[**Validation Dataset** 6](#_Toc74148401)

[**Data Directory** 6](#_Toc74148402)

[**Training of the table detection model** 6](#_Toc74148403)

[**Training results** 7](#_Toc74148404)

[**Testing of the trained table detection model** 10](#_Toc74148405)

[**Test results** 10](#_Toc74148406)

[**Inference using the detection model** 12](#_Toc74148407)

[**References** 14](#_Toc74148408)

[**Appendices** 14](#_Toc74148409)

[**Financial Spreading Model Details** 14](#_Toc74148410)

[**Model files** 15](#_Toc74148411)

# **What is Data Science?**

## Data science defined

Data science combines multiple fields, including statistics, scientific methods, artificial intelligence (AI), and data analysis, to extract value from data. Those who practice data science are called data scientists, and they combine a range of skills to analyze data collected from the web, smartphones, customers, sensors, and other sources to derive actionable insights.

Data science encompasses preparing data for analysis, including cleansing, aggregating, and manipulating the data to perform advanced data analysis. Analytic applications and data scientists can then review the results to uncover patterns and enable business leaders to draw informed insights.

## Data science: An untapped resource for machine learning

Data science is one of the most exciting fields out there today. But why is it so important?

Because companies are sitting on a treasure trove of data. As modern technology has enabled the creation and storage of increasing amounts of information, data volumes have exploded. It’s estimated that 90 percent of the data in the world was created in the last two years. For example, Facebook users upload 10 million photos every hour.

But this data is often just sitting in databases and data lakes, mostly untouched.

The wealth of data being collected and stored by these technologies can bring transformative benefits to organizations and societies around the world—but only if we can interpret it. That’s where data science comes in.

Data science reveals trends and produces insights that businesses can use to make better decisions and create more innovative products and services. Perhaps most importantly, it enables machine learning (ML) models to learn from the vast amounts of data being fed to them, rather than mainly relying upon business analysts to see what they can discover from the data.

Data is the bedrock of innovation, but its value comes from the information data scientists can glean from it, and then act upon.

Three common data preprocessing steps are formatting, cleaning and sampling:

* **Formatting**: The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file.
* **Cleaning**: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonymized or removed from the data entirely.
* **Sampling**: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

## Step 3: Transform Data

The final step is to transform the process data. The specific algorithm you are working with and the knowledge of the problem domain will influence this step and you will very likely have to revisit different transformations of your preprocessed data as you work on your problem.

Three common data transformations are scaling, attribute decompositions and attribute aggregations. This step is also referred to as feature engineering.

* **Scaling**: The preprocessed data may contain attributes with a mixtures of scales for various quantities such as dollars, kilograms and sales volume. Many machine learning methods like data attributes to have the same scale such as between 0 and 1 for the smallest and largest value for a given feature. Consider any feature scaling you may need to perform.
* **Decomposition**: There may be features that represent a complex concept that may be more useful to a machine learning method when split into the constituent parts. An example is a date that may have day and time components that in turn could be split out further. Perhaps only the hour of day is relevant to the problem being solved. consider what feature decompositions you can perform.
* **Aggregation**: There may be features that can be aggregated into a single feature that would be more meaningful to the problem you are trying to solve. For example, there may be a data instances for each time a customer logged into a system that could be aggregated into a count for the number of logins allowing the additional instances to be discarded. Consider what type of feature aggregations could perform.

You can spend a lot of time engineering features from your data and it can be very beneficial to the performance of an algorithm. Start small and build on the skills you learn.

How to select features and what are Benefits of performing feature selection before modeling your data?

· Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.

· Improves Accuracy: Less misleading data means modeling accuracy improves.

· Reduces Training Time: fewer data points reduce algorithm complexity and algorithms train faster.

* **Feature Selection**: Select a subset of input features from the dataset.
  + **Unsupervised**: Do not use the target variable (e.g. remove redundant variables).
    - Correlation
  + **Supervised**: Use the target variable (e.g. remove irrelevant variables).
    - **Wrapper**: Search for well-performing subsets of features.
      * RFE
    - **Filter**: Select subsets of features based on their relationship with the target.
      * Statistical Methods
      * Feature Importance Methods
    - **Intrinsic**: Algorithms that perform automatic feature selection during training.
      * Decision Trees
* **Dimensionality Reduction**: Project input data into a lower-dimensional feature space.

*Wrapper methods evaluate multiple models using procedures that add and/or remove predictors to find the optimal combination that maximizes model performance.*

## What’s the difference between data science, artificial intelligence, and machine learning?

To better understand data science—and how you can harness it—it’s equally important to know other terms related to the field, such as artificial intelligence (AI) and machine learning. Often, you’ll find that these terms are used interchangeably, but there are nuances.

Here’s a simple breakdown:

* **AI** means getting a computer to mimic human behavior in some way.
* **Data science** is a [subset of AI](https://www.oracle.com/in/data-science/cloud-infrastructure-data-science.html), and it refers more to the overlapping areas of statistics, scientific methods, and data analysis—all of which are used to extract meaning and insights from data..
* [**Machine learning**](https://www.oracle.com/in/data-science/machine-learning/what-is-machine-learning/) is another subset of AI, and it consists of the techniques that enable computers to figure things out from the data and deliver AI applications.  
  And for good measure, we’ll throw in another definition.
* [**Deep learning**](https://www.oracle.com/in/data-science/machine-learning/what-is-deep-learning/) which is a subset of machine learning that enables computers to solve more complex problems.

## How data science is transforming business

Organizations are using data science to turn data into a competitive advantage by refining products and services. Data science and machine learning use cases include:

* Determine customer churn by analyzing data collected from call centers, so marketing can take action to retain them
* Improve efficiency by analyzing traffic patterns, weather conditions, and other factors so logistics companies can improve delivery speeds and reduce costs
* Improve patient diagnoses by analyzing medical test data and reported symptoms so doctors can diagnose diseases earlier and treat them more effectively
* Optimize the supply chain by predicting when equipment will break down
* Detect fraud in financial services by recognizing suspicious behaviors and anomalous actions
* Improve sales by creating recommendations for customers based upon previous purchases

Many companies have made data science a priority and are investing in it heavily. [In Gartner’s recent survey](https://www.gartner.com/en/newsroom/press-releases/2018-01-25-gartner-says-self-service-analytics-and-bi-users-will-produce-more-analysis-than-data-scientists-will-by-2019) of more than 3,000 CIOs, respondents ranked analytics and business intelligence as the top differentiating technology for their organizations. The CIOs surveyed see these technologies as the most strategic for their companies, and are investing accordingly.

## How data science is conducted

The process of analyzing and acting upon data is iterative rather than linear, but this is how the data science lifecycle typically flows for a data modeling project:

**Planning:**  Define a project and its potential outputs.

**Building a data model:**  Data scientists often use a variety of open source libraries or in-database tools to build machine learning models. Often, users will want APIs to help with data ingestion, data profiling and visualization, or feature engineering. They will need the right tools as well as access to the right data and other resources, such as compute power.

**Evaluating a model:**  Data scientists must achieve a high percent of accuracy for their models before they can feel confident deploying it. Model evaluation will typically generate a comprehensive suite of evaluation metrics and visualizations to measure model performance against new data, and also rank them over time to enable optimal behavior in production. Model evaluation goes beyond raw performance to take into account expected baseline behavior.

**Explaining models:**  Being able to explain the internal mechanics of the results of machine learning models in human terms has not always been possible—but it is becoming increasingly important. Data scientists want automated explanations of the relative weighting and importance of factors that go into generating a prediction, and model-specific explanatory details on model predictions.

**Deploying a model:**  Taking a trained, machine learning model and getting it into the right systems is often a difficult and laborious process. This can be made easier by operationalizing models as scalable and secure APIs, or by using in-database machine learning models.

**Monitoring models:**  Unfortunately, deploying a model isn’t the end of it. Models must always be monitored after deployment to ensure that they are working properly. The data the model was trained on may no longer be relevant for future predictions after a period of time. For example, in fraud detection, criminals are always coming up with new ways to hack accounts.

## Tools for data science

Building, evaluating, deploying, and monitoring machine learning models can be a complex process. That’s why there’s been an increase in the number of data science tools. Data scientists use many types of tools, but one of the most common is open source notebooks, which are web applications for writing and running code, visualizing data, and seeing the results—all in the same environment.

Some of the most popular notebooks are Jupyter, RStudio, and Zeppelin. Notebooks are very useful for conducting analysis, but have their limitations when data scientists need to work as a team. Data science platforms were built to solve this problem.

To determine which data science tool is right for you, it’s important to ask the following questions: What kind of languages do your data scientists use? What kind of working methods do they prefer? What kind of data sources are they using?

For example, some users prefer to have a data source-agnostic service that uses open source libraries. Others prefer the speed of in-database, machine learning algorithms.

## Challenges of implementing data science projects

Despite the promise of data science and huge investments in data science teams, many companies are not realizing the full value of their data. In their race to hire talent and create data science programs, some companies have experienced inefficient team workflows, with different people using different tools and processes that don’t work well together. Without more disciplined, centralized management, executives might not see a full return on their investments.

This chaotic environment presents many challenges.

**Data scientists can’t work efficiently.** Because access to data must be granted by an IT administrator, data scientists often have long waits for data and the resources they need to analyze it. Once they have access, the data science team might analyze the data using different—and possibly incompatible—tools. For example, a scientist might develop a model using the R language, but the application it will be used in is written in a different language. Which is why it can take weeks—or even months—to deploy the models into useful applications.

**Application developers can’t access usable machine learning.** Sometimes the machine learning models that developers receive are not ready to be deployed in applications. And because access points can be inflexible, models can’t be deployed in all scenarios and scalability is left to the application developer.

**IT administrators spend too much time on support.** Because of the proliferation of open source tools, IT can have an ever-growing list of tools to support. A data scientist in marketing, for example, might be using different tools than a data scientist in finance. Teams might also have different workflows, which means that IT must continually rebuild and update environments.

**Business managers are too removed from data science.** Data science workflows are not always integrated into business decision-making processes and systems, making it difficult for business managers to collaborate knowledgeably with data scientists. Without better integration, business managers find it difficult to understand why it takes so long to go from prototype to production—and they are less likely to back the investment in projects they perceive as too slow.

[Learn about the data science lifecycle (PDF)](https://www.oracle.com/in/a/ocom/docs/data-science-lifecycle-ebook.pdf)

## The data science platform delivers new capabilities

Many companies realized that without an integrated platform, data science work was inefficient, unsecure, and difficult to scale. This realization led to the development of [data science platforms](https://www.oracle.com/in/data-science/). These platforms are software hubs around which all data science work takes place. A good platform alleviates many of the challenges of implementing data science, and helps businesses turn their data into insights faster and more efficiently.

With a centralized, machine learning platform, data scientists can work in a collaborative environment using their favorite open source tools, with all their work synced by a version control system.

## The benefits of a data science platform

A data science platform reduces redundancy and drives innovation by enabling teams to share code, results, and reports. It removes bottlenecks in the flow of work by simplifying management and incorporating best practices.

In general, the best data science platforms aim to:

* Make data scientists more productive by helping them accelerate and deliver models faster, and with less error
* Make it easier for data scientists to work with large volumes and varieties of data
* Deliver trusted, enterprise-grade artificial intelligence that’s bias-free, auditable, and reproducible

Data science platforms are built for collaboration by a range of users including expert data scientists, [citizen data scientists](https://blogs.gartner.com/carlie-idoine/2018/05/13/citizen-data-scientists-and-why-they-matter/), [data engineers](https://blogs.oracle.com/datascience/ai-influencer-blog-series-why-data-engineering-is-crucial-to-scalable-data-science-v2), and machine learning engineers or specialists. For example, a data science platform might allow data scientists to deploy models as APIs, making it easy to integrate them into different applications. Data scientists can access tools, data, and infrastructure without having to wait for IT.

The demand for data science platforms has exploded in the market. In fact, [the platform market is expected to grow](https://www.analyticsinsight.net/top-10-data-science-platforms-that-cash-the-analytics-code/) at a compounded annual rate of more than 39 percent over the next few years and is projected to reach US$385 billion by 2025.

## What a data scientist needs in a platform

If you’re ready to explore the capabilities of data science platforms, there are some key capabilities to consider:

**Choose a project-based UI that encourages collaboration.** The platform should empower people to work together on a model, from conception to final development. It should give each team member self-service access to data and resources.

**Prioritize integration and flexibility.** Make sure the platform includes support for the latest open source tools, common version-control providers, such as GitHub, GitLab, and Bitbucket, and tight integration with other resources.

**Include enterprise-grade capabilities.** Ensure the platform can scale with your business as your team grows. The platform should be highly available, have robust access controls, and support a large number of concurrent users.

**Make data science more self-service.** Look for a platform that takes the burden off of IT and engineering, and makes it easy for data scientists to spin up environments instantly, track all of their work, and easily deploy models into production.

**Ensure easier model deployment.** Model deployment and operationalization is one of the most important steps of the machine learning lifecycle, but it’s often disregarded. Make sure that the service you choose makes it easier to operationalize models, whether it’s providing APIs or ensuring that users build models in a way that allows for easy integration.

## When a data science platform is the right move

Your organization could be ready for a data science platform, if you’ve noticed that:

* Productivity and collaboration are showing signs of strain
* Machine learning models can’t be audited or reproduced
* Models never make it into production

A data science platform can deliver real value to your business. Oracle’s [data science platform](https://www.oracle.com/in/data-science/) includes a wide range of services that provide a comprehensive, end-to-end experience designed to accelerate model deployment and improve data science results.

AR PURPOSE.