intetics

Where software concepts come alive™

Is your Data ready?

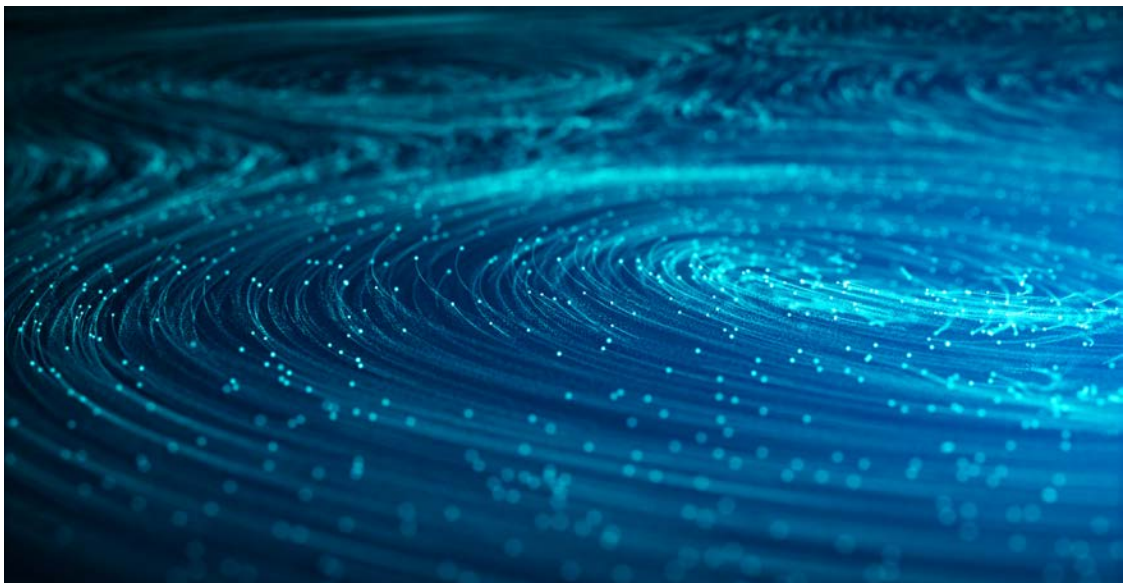
Data Science at your fingertips for your business.

White Paper

intetics
Where software concepts come alive™

Table of Content

Data Science Introduction



On any given day, 294 billion emails are delivered, 500 million Tweets are posted, and 4 petabytes of data are created via Facebook. And that's just the tip of the iceberg of the digital data society produces on a daily basis. We have an overwhelming amount of data at our fingertips; within it lie transformative insights that just need to be uncovered and polished.

The era of Data Science, big data, and analytics has driven significant industrial, governmental, and disciplinary interest and has shifted innovation and research. New developments in Data Science are creating exciting trends – but they also come with controversial pitfalls, which we'll examine.

Data Science combines math and statistics with scientific methodology, advanced analytics, specialized programming, AI, and even storytelling. Its purpose? To uncover and explain the business insights buried within massive amounts of data.

The multi-disciplinary approach to Data Science includes the preparation of data, the performance of advanced analysis, and the presentation of results and patterns to stakeholders in an intuitive, visual manner. This technology leads to a host of data-enabled opportunities in education, government, research, the economy, entertainment, and more.

In this whitepaper, we'll do a deep dive into the potential of Data Science, its main tech components, and the challenges of incorporating it into an analytics strategy.

How Data Science Helps Your Business Challenges

Data scientists don't solve data problems. Rather, they solve problems WITH data. Data Science is a mix of science and art; the process of solving business issues needs both a strong algorithm and creative problem-solving skills. There are two categorizations of problems that Data Science can be used to solve: high-level and general.

1

High-level problems

- Problems that require analytics of massive data
- Problems that are affected by the human bias
- Problems that require extreme precision
- Problems that are too expensive to be solved by humans

2

General problems

- Simplify product marketing and assist in making accurate sales forecasts
- Improve the precision of financial rules and models
- Easily detect spam without human oversight
- Facilitate accurate medical predictions and diagnoses
- Simplify time-intensive data entry
- Increase the efficiency of predictive maintenance in the manufacturing industry (automotive, transportation, building maintenance, energy, etc.)
- Provide better customer segmentation and accurate lifetime value predictions
- Recommend the right product to potential buyers
- Improve cybersecurity
- Automatic object recognition on videos/images/point clouds
- Reduce operating expenses

WE CAN ALSO BREAK IT DOWN BY INDUSTRY:

1. Manufacturing

- Predictive maintenance & condition monitoring
- Warranty reserve estimation
- Process optimization
- Telematics
- On-site safety monitoring

2. Retail

- Recommendation engines
- Propensity to buy
- Predictive inventory planning
- Market segmentation and targeting
- Upsell and cross-channel marketing
- Customer ROI and lifetime value

3. Healthcare

- Patient triage optimization
- Alerts and diagnostics from real-time patient data
- Disease identification and risk stratification
- Healthcare provider sentiment analysis
- Proactive health management

4. Finance

- Customer segmentation
- Risk analytics and regulation
- Sales and marketing campaign management
- Cross-selling and up-selling
- Creditworthiness evaluation

5. Energy, feedstock, and utilities

- Dynamic pricing
- Consumer feedback and interaction analysis
- Logistic scheduling
- Traffic patterns and congestion management
- Power usage analytics
- Smart grid management
- Energy demand and supply optimization
- Customer-specific pricing

6. Art

- Image generation
- Music generation
- Copyright violation recognition

Data Science also has vast applications outside the realm of business. In fact, big data and machine learning have the potential to tackle pressing societal problems – take, for instance, the 17 Sustainable Development Goals identified by the UN in 2015. These goals include eliminating poverty, supporting animal life on land and below water, reducing inequality between people groups, promoting responsible production and consumption, building sustainable cities and communities, and more.

So, how is data science poised to tackle these challenges? With modern technology – particularly data analysis tools – we now have access to the largest amount of data ever, as well as a deeper capacity to use it in creating services and products that meet fundamental human issues head-on.

When academics and researchers gather strong data sets on certain health, social, or environmental issues, they can get a better understanding of the issue's impact and severity. Working together, governments, academics, non-government organizations, and businesses can use their innovation, leadership, and entrepreneurial skills to make data-driven solutions.

Real example: Imperial College Business School, based in the UK, launched the Gandhi Centre for Inclusive Education in 2007. The Centre hosts a yearly Data Science for Social Good fellowship, in which Fellows develop data-powered solutions that address critical, real-world problems. In 2019's fellowship, attendees used Data Science to determine how traffic disruptions and policies impact London's road congestion – and, in turn, air pollution.

Prior to the fellowship, traffic statistics were obtained manually: workers stood on the side of the road and counted cars, but this was time-consuming and very costly. Furthermore, the statistics were compiled into yearly averages, which weren't detailed enough to provide insight on traffic initiatives. Furthermore, because there weren't accurate, real-time, junction-level data, manual prediction of pollution was determined to be underestimated by up to 30%.

So, the Centre's fellowship project analyzed London traffic videos from over 900 cameras across the city. The project created an algorithm that accurately generated a count of unique vehicles classified by type (bike, truck, car, bus, etc.) in real-time. And, more importantly, it extracted the number of stops and starts for each vehicle, which was the main factor in air pollution underestimation.

This kind of algorithm makes it possible to make improved estimates of London's air pollution, enabling the design of accurate emission zones, as well as the optimization of red lights and the planning of "green" routes.

And this is just one single application! Data Science can be used to tackle climate change, poverty, and many more societal challenges.

Yet, there are also challenges inherent to adopting big data and AI; we've identified 8 of the most pressing obstacles.

Obstacle 1: Data Preparation

Data scientists spend an astounding 80% of their time preparing and improving data quality before they analyze it. However, 57% of surveyed data scientists consider data preparation to be the most mundane part of their job. As part of this phase, they must sift through terabytes of data across many sources, platforms, formats, and functions – all the while keeping an activity log to avoid duplication.

To cut down on time spent on data preparation (and ease frustration), adopting AI-enabled Data Science technologies can help greatly. For instance, the Augmented Analytics approach automates data cleansing tasks, increasing the productivity of data scientists.

Obstacle 2: Multiple Data Sources

As organizations continue to generate multiple data formats with different kinds of apps and tools, data scientists will need to produce meaningful decisions based on an increased number of data sources. This could lead to manual data entry and lengthy searching, which leads to repetition, errors, and less-than-ideal decisions.

To eliminate this obstacle, organizations can use a centralized platform that integrates with several data sources so that data scientists can immediately retrieve information across multiple sources. Data can be gathered and controlled in real-time through this platform, saving data scientists a lot of time and effort.

Obstacle 3: Data Security

As organizations are moving to cloud data management, cyberattacks are becoming more commonplace. This causes two main issues: first, confidential data has become more vulnerable. Second, regulatory standards are evolving in response to heightened vulnerability, thus extending the data consent/utilization processes. The latter leads to lengthy audits and, potentially, hefty fines.
To address this challenge, organizations should use machine learning-enabled security platforms to safeguard data, supplemented by additional security checks.

Obstacle 4: Understanding the Business Problem

Before data scientists analyze data and build solutions, they must first have a thorough understanding of the business problem that's being addressed. In many cases, data scientists take a mechanical approach and begin data set analysis without a clearly defined objective.

To avoid any wastage of time and resources, data scientists should follow a workflow that is built after stakeholder collaboration, containing well-defined checklists.

Obstacle 5: Communicating With Non-Technical Stakeholders

It's crucial for data scientists to communicate results effectively with business executives – yet, the latter party might not understand the work's technical jargon. However, "data storytelling" helps data scientists take a structured approach to developing a powerful narrative and understandable communication to their visualizations and analysis.

Obstacle 6: Collaborating With Data Engineers

Organizations typically have data engineers and data scientists work on the same projects – which means that to ensure the best output, there needs to be effective communication. Yet, the two parties usually have different workflows and priorities, which leads to misunderstanding and could potentially stifle knowledge-sharing.

To address this, management should actively enhance collaboration between data engineers and scientists. This can be done by setting up a real-time collaboration tool alongside a common coding language. Assigning a Chief Data Officer to supervise both departments can also go a long way towards improving collaboration.

Obstacle 7: Role Misconceptions

In some organizations, a data scientist is expected to wear many hats: they must clean and retrieve data, build models, and perform analysis. However, this is a lot to expect from one person. In an optimally-functioning data science team, tasks should be distributed among individuals rather than hoisted all onto one scientist. Along the same vein, Data Scientists must clearly understand their role when they begin working with an organization.

Obstacle 8: Undefined Metrics

When management teams have a misaligned understanding of Data Science, this can lead to unrealistic expectations placed upon the data team. To avoid this, businesses should have well-defined metrics to measure the data analysis's accuracy – as well as business KPIs to determine the business impact that the data analysis would have.

Despite all of these challenges, data scientists are among the most in-demand specialists in the IT market. Being a successful data scientist is about having relevant technical skills, clearly understanding business requirements, collaborating with stakeholders, and communicating effectively with business executives.

Data Science Overview

Data Science extracts actionable insights from huge volumes of data – the science encompasses data preparation, advanced data analysis, and presenting results to stakeholders.

The preparation phase may involve aggregating, cleaning, and manipulating data, so it's ready for different kinds of processing. The development and usage of analytic algorithms and AI models are necessary for the analysis phase. And this is driven by software that goes through massive amounts of data, finding patterns, transforming those patterns into predictions, and using those predictions to support business decisions. Of course, the predictions' accuracy needs to be validated through scientific tests. The end results are shared via data visualization tools, making it possible for stakeholders with less Data Science expertise to see the trends and understand the patterns.

As a result, Data Science practitioners require computer science – and general science – skills greater than those of an average data analyst. Such skills include:

- Applying scientific methodology, mathematics, and statistics
- Using many techniques and a wide range of tools for data evaluation and preparation, including data mining/data integration methods and SQL
- Using predictive analytics and AI (including deep learning and machine learning) to extract insights from data

- Writing applications that can automate calculations and data processing
- Transforming results into "stories" that make technical results more accessible to stakeholders and decision-makers of every level
- Explaining how results can be applied to solve business challenges

This skill combination is rare, which is why data scientists are in very high demand. According to IBM, data scientist job openings grow at more than 5% per year.

Exploring the Data Science Lifecycle

Data Science projects don't have well-defined, clean lifecycles like the software development process. In fact, the Data Science lifecycle can be comprised of 5 – 16 processes, depending on whom you ask. Although the workflow process for Data Science might not be clean, there is still a general standard workflow: acquisition, preparation, hypothesis/modeling, evaluation/interpretation, deployment, operations, and optimization. It's fairly similar to the Cross-Industry Standard Process for Data Mining (CRISP-DM), with modifications applied as per the specific Data Science project's requirements.

Step 1

Data Acquisition

In order to carry out Data Science procedures, you need data to work with. The first step in any Data Science project is to identify a person that knows which data to acquire – and when to acquire it. This person doesn't have to be a data scientist per se; rather, they have to understand the difference between available data sets and be capable of making weighty decisions about the organization's data strategy.

Various data sources can be identified for the project, including social media data, web server logs, web scraping, online repositories like US Census datasets, and more. Data acquisition essentially gathers data from all identified external and internal sources to help solve a business challenge.

One major challenge often encountered in this step is tracking the origins of each data slice and verifying whether it is up-to-date. This information must be tracked throughout the entire Data Science project, or else it might need to be re-acquired to run updated experiments and additional hypotheses.

Step 2

Data Preparation

The data preparation phase is also referred to as "data wrangling" or "data cleaning." As mentioned above, many data scientists identify this phase as the most time-consuming and monotonous task of the Data Science lifecycle. Why is it so necessary?

In many cases, the data acquired in the first phase of a Data Science project's lifecycle are not in a usable format. There may be inconsistencies, missing entries, or semantic errors. Thus, after the acquisition, data scientists must "clean" and reformat the data – this is done by writing code or manually editing the data in the spreadsheet.

Data preparation itself does not produce meaningful business insights. However, by cleaning data on a regular basis, data scientists can determine the weaknesses of the acquisition process, which assumptions should be made, and which models can be applied to generate analysis results.

Reformatted data can be converted to CSV, JSON, and other formats that are easily loaded into Data Science tools.

Exploratory data analysis also falls under the data preparation phase; data scientists summarize the clean data and identify patterns, outliers, and anomalies that can be referenced in later steps.

Step 3

Hypothesis and Modeling

In this phase, data scientists first reduce their data set's dimensionality – meaning they select relevant features and values that contribute to results' predictions. Next, it's time to determine which machine learning model will be best able to analyze and derive valuable business insights from the data. The model itself needs to be built; they are typically written in languages like MATLAB, Python, R, or Perl. Data analytics experts build the model, relying on tools and techniques like logistic regression, decision trees, and neural networks. Contending models are then trained with training data sets.

Step 4

Evaluation and Interpretation

There are different evaluation metrics depending on what the model aims to perform. For instance, if the designed machine learning model is supposed to predict daily stock, then the Root Mean Squared Error (RMSE) should be evaluated. If, instead, the model is supposed to classify spam emails, AUC, log loss, and average accuracy should be considered.

Data scientists must determine which dataset should be used to measure the machine learning model's performance. Evaluating the performance metrics of the trained dataset can be helpful, but it's not always right – the results may be overly optimistic since the model has already adapted to that training set.

Step 5

Deployment

Before deployment, machine learning models might have to be refactored – for instance, to enforce more strict security practices or add more transparent monitoring. After any necessary changes, machine learning models are deployed in a test or pre-production environment before becoming fully operational on the live servers.

Step 6

Operations and Maintenance

This phase involves the development and implementation of a long-term maintenance plan. The model's performance, as well as performance downgrade, is carefully monitored during this step. Data scientists can preserve knowledge from a specific project for shared learning and speed up the lifecycle of future projects.

Step 7

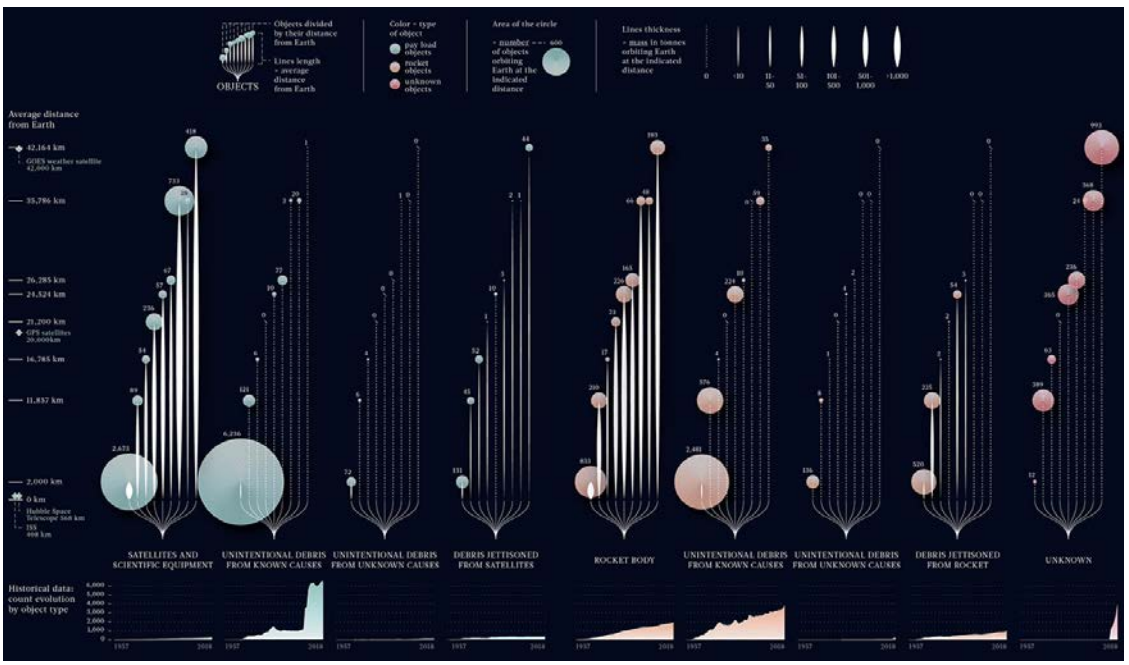
Optimization

The final phase of a Data Science project is the retraining of the machine learning model whenever new data sources come in.

Now, throughout this process, valuable insights extracted from the machine learning models are translated into digestible, visual presentations for stakeholders and business decision-makers. These rules can help you create successful data visualizations:

- Ensure the data answers a question
- Describe data precisely and in words; include headers, subheaders, and explanatory text
- Use a visual hierarchy; consider size and scale, color and contrast, typography, and perspective
- Make your data points easily comparable

Here's an example of data visualization that can easily communicate the results of a Data Science project to stakeholders:



Data Science Key Terms

Before delving deeper into the White Paper, we wanted to present you with some key words highly pertinent to Data Science:

Advanced analytics

Data analytics that could not be achieved by traditional data analysis means.

Big data

Data that is too large or complex to be effectively handled by traditional data-related tools.

Data analysis

Processing data with traditional theories, technologies, and tools, with the purpose of obtaining useful information with a practical application.

Data analytics

Theories, tools, technologies, and processes that facilitate the discovery and in-depth understanding of big data, deriving actionable insights. Data analytics comprises descriptive analytics, predictive analytics, and prescriptive analytics.

Descriptive analytics

Data analytics that uses statistics to describe the data used to get information or for other practical purposes.

Predictive analytics

Data analytics that makes predictions about future events and explains the reasons behind those predictions, typically by using advanced analytics.

Prescriptive analytics

Data analytics that optimizes indications and prescribes actions for smart decision-making.

Explicit analytics

Descriptive analytics done by reporting, descriptive analysis, alerting, and forecasting.

Implicit analytics

Deep analytics done by predictive modeling, optimization, prescriptive analytics, and actionable knowledge delivery

Deep analytics

Data analytics that gains an in-depth understanding of why and how things have happened (or will happen), which cannot be determined by descriptive analytics.

Brief History

Modern "Data Science" was described by William S. Cleveland in 2001, in his book, Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. Around a year later, the CODATA Data Science Journal began its yearly publication, and things really started booming. However, even though Data Science as we know it was defined in the early 2000s, we can trace its roots back to the 1960s.

In 1962, John Tukey, a famous American mathematician, published The Future of Data Analysis, which covered a shift in the statistics world. Essentially, Tukey drew attention to the converging of computers and statistics, explaining how measurable outcomes could be introduced far more quickly with a computer than when done by hand.

In 1974, Peter Naur published Concise survey of Computer Methods, in which the term "Data Science" – however, the term's meaning was very vague and not aligned with its modern definition. In Naur's words, Data Science is "The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."

In the late 1970s, the Data Science industry experienced a switch from conceptualization to deeper research. The International Association of Statistical Computing was founded in 1977 – and, that same year, John Tukey published Exploratory Data Analysis. This scientific paper argued that there should be a heavy emphasis on using data to suggest, test, and verify a hypothesis. This was the birth of two intertwined practices, confirmatory and exploratory data analysis.

> ***Exploratory data analysis is where you are figuring out what to glean from data: you're establishing which questions you want to ask, determining how to frame those questions, and figuring out how you'll present and manipulate data to extract necessary insights. It involves establishing key variables, finding anomalies, checking assumptions, and so on. On the other hand, Confirmatory analysis involves testing your hypothesis, producing specific estimates, and doing regression and variance analysis. In other words, it's where you use confirmatory analysis to put your data findings on trial.***

In 1989, the first Data Science workshop was established: Knowledge Discovery in Databases. It still exists today, under a different name – it's the annual ACM SIGKDD Conference on Knowledge Discovery and Data Mining.

Data science really began picking up steam **in the 1990s**, demonstrated by Business Week's cover story in September 1994, Database Marketing. Businesses worldwide began to understand the importance of gathering and applying data – so, they started stockpiling huge amounts of data. The problem, though, was that they weren't sure what to do with this data at that point in time.

In 1996, data mining was defined in the publication, From Data Mining to Knowledge Discovery in Databases. Therein, data mining was stated as being "the application of specific algorithms for extracting patterns from data."

In 2001, we started seeing huge developments in the world of Data Science, largely due to the aforementioned book by William S. Cleveland. He theorized that Data Science practitioners must be knowledgeable in six key areas: models and methods of data, multi-disciplinary investigations, data computing, tool evaluation, pedagogy, and theory.

Over the next decade, "Data Science" became a hot buzzword, and companies were slowly learning how to use their stockpiled data to gain insights. **In 2013**, IBM announced that 90% of the world's data were created in just the previous two years.

Major tech giants gave credit to Data Science for increased profits – for instance, Amazon announced that it sold more Kindle eBooks than ever before, and Apple also gave Big Data credit for increased sales. As these tech giants put their public support behind the industry of data science, companies that were dragging their feet began to follow suit and gather data from all kinds of resources.

Nowadays, Data Science is a lucrative, in-demand industry and is being used to solve all kinds of problems – and every organization is trying to get in on the action. But, as you can see, Data Science didn't have a prestigious beginning. It was largely ignored for the first couple of decades and really only started to boom once business people realized it could boost their profits.

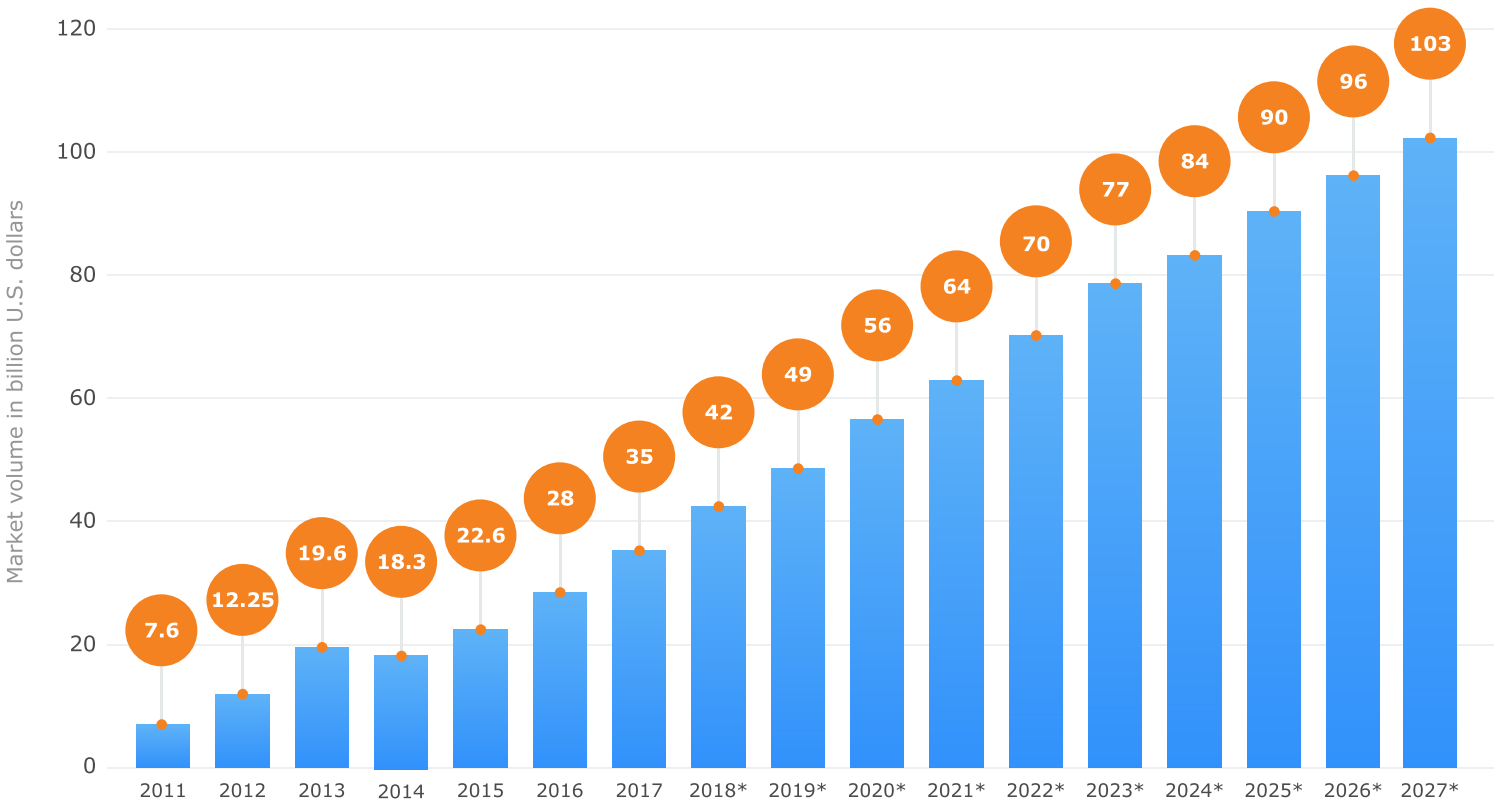
Total Market Volume

While Data Science has experienced rapid growth, it's not even close to slowing down. We've gathered several key statistics showing the scope of the Data Science market.

The major Data Science players include, among others:

Google IBM ORACLE Microsoft amazon

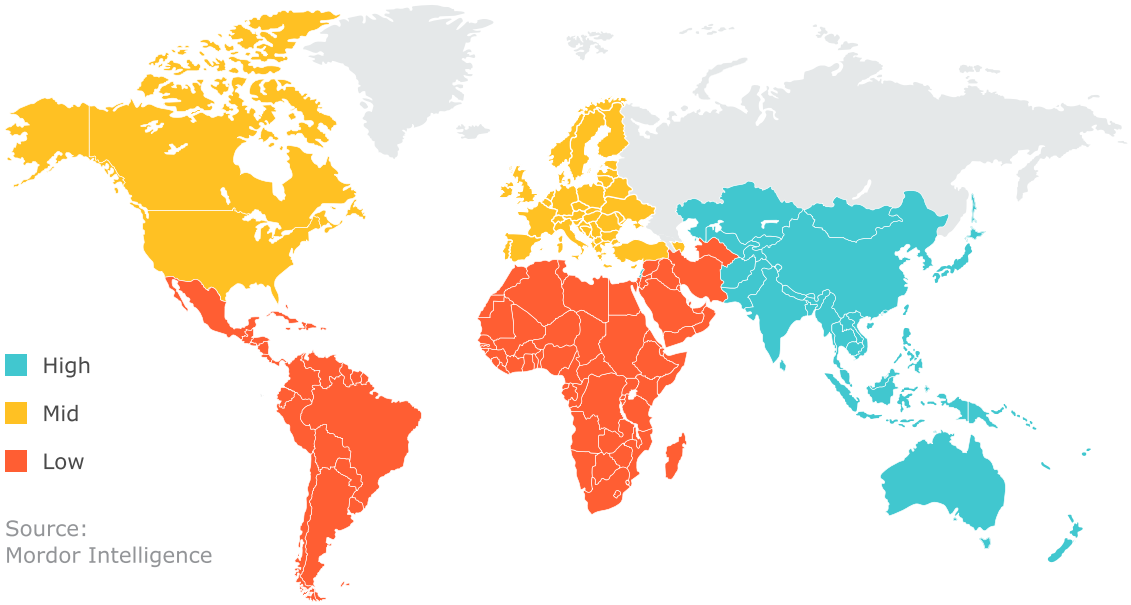
BIG DATA MARKET SIZE REVENUE FORECAST WORLDWIDE FROM 2011 TO 2027 (in billion U.S. dollars)



Source: Statista

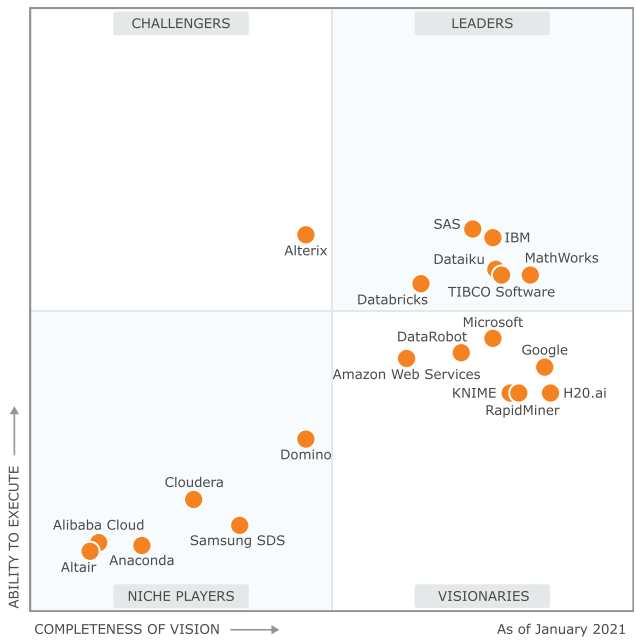
According to a study from Mordor Intelligence, **the highest geographical growth areas** are located in the Asia-Pacific, while North America and Western Europe are experiencing a mid-growth rate. South America and Africa are growing slowly. However, even though North America is not growing at the fastest rate, it currently holds the largest share of the Data Science market, according to Market Watch.

DATA SCIENCE PLATFORM MARKET – GROWTH RATE BY GEOGRAPHY (2020-2025)



Source:
Mordor Intelligence

MAGIC QUADRANT FOR DATA SCIENCE AND MACHINE LEARNING PLATFORMS



Source: Gartner (March 2021)

Data Science Market Forecast

Grandview Research has released a Data Science Market Forecast for 2020-2027. Their report explains that global Data Science was worth $3.93 billion in 2019 and $4.89 billion in 2020. It is anticipated to hold an annual growth rate of 26.9% until 2027. Other key findings in the Forecast include:

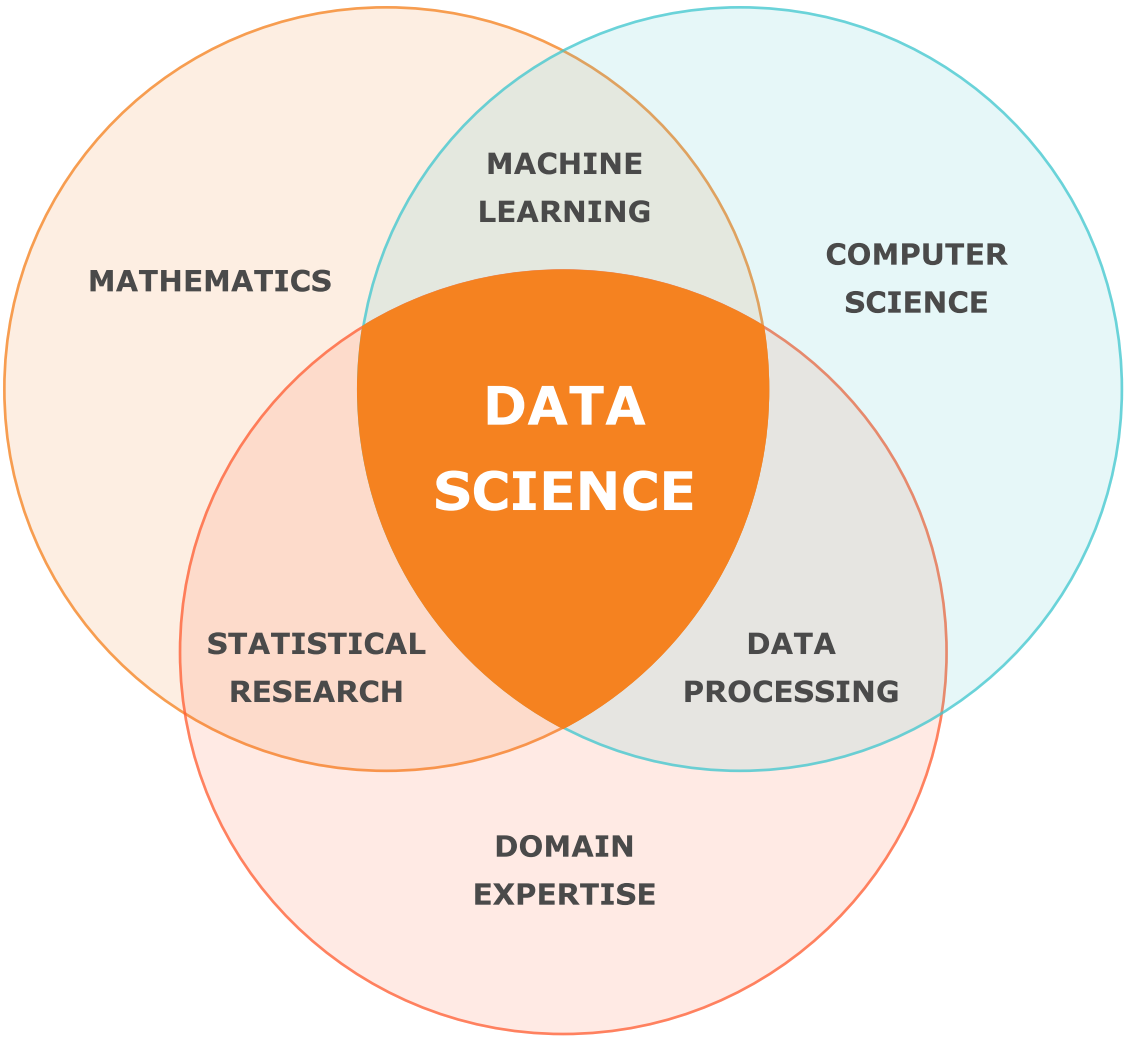
- Driving factors of the Data Science market are advancements in AI, the Internet of Things, and machine learning.
- Because global data volumes are increasing every day, advanced platforms and tools for handling data have a substantial positive impact on business growth.
- COVID-19 has influenced the Data Science industry; previous forecasting and segmentation models failed due to changes in online shopping patterns and interruptions in the supply chain. So, companies are having to revise their assumptions made during data analysis.
- Healthcare applications of Data Science are expected to grow significantly during the forecast period.
- The Asia-Pacific region is anticipated to experience the highest Compound Annual Growth Rate during the forecast period.

Technical Side of Data Science

The main components of Data Science are as follows:

- **Statistics** — This is a way of collecting and analyzing large quantities of numerical data, all while gleaning meaningful insights.
- **Domain Expertise** — This is what binds Data Science together – a specialized skill set or knowledge of a particular area. Areas of domain expertise include knowledge of business and operational dynamics, industry segments, the competitive environment, industry-specific IT solutions, and best practices within the domain.
- **Data Engineering** — It involves acquiring, storing, retrieving, processing, and transforming data – as well as managing metadata.
- **Visualization** — Data are represented in a visual context so that stakeholders can easily understand their components.
- **Software Development** — This involves designing, writing, debugging, and maintaining a computer program's source code.
- **Mathematics** — This critical portion of Data Science depends on the usage of complex mathematical algorithms that allow for the production of desired results in a reasonable time on limited hardware resources.
- **Machine learning** — It is the backbone of Data Science. Essentially, it trains a machine to identify various patterns in data. Over time, its capabilities and accuracy improve without being explicitly programmed to do so. Various machine learning algorithms are applied in Data Science to solve business challenges.

THE MAIN COMPONENTS OF DATA SCIENCE

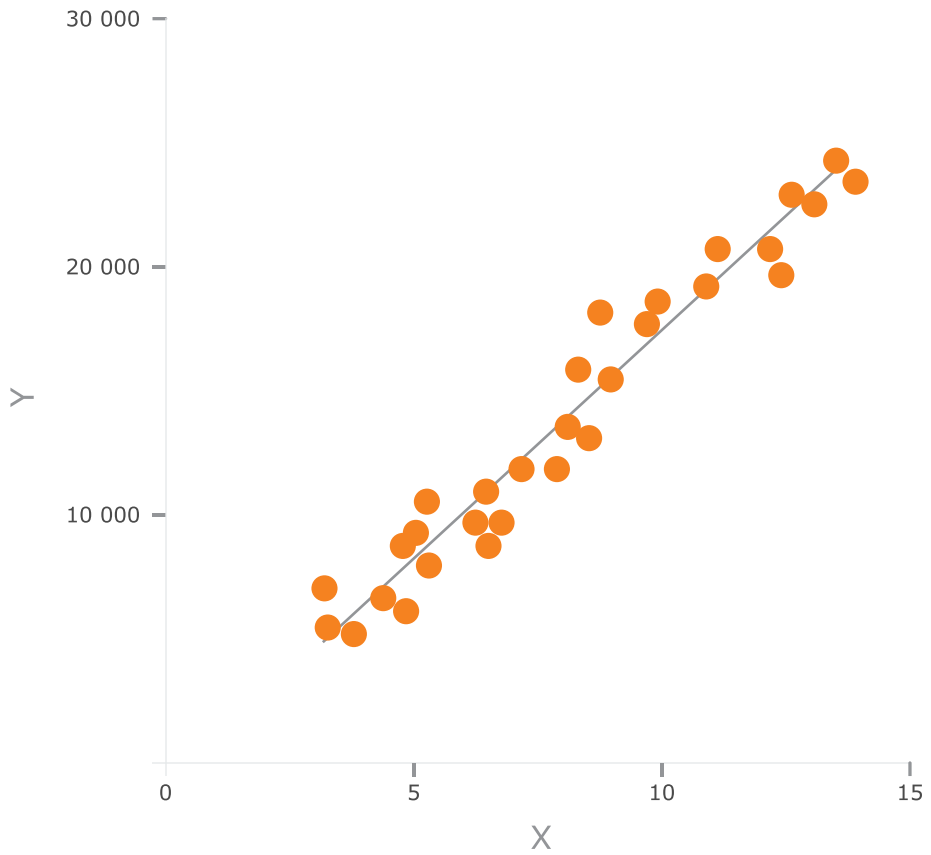


Machine Learning in Data Science

It is critical to be aware of Data Science's machine learning algorithms; we'll provide a basic introduction to three of the most broadly used ones.

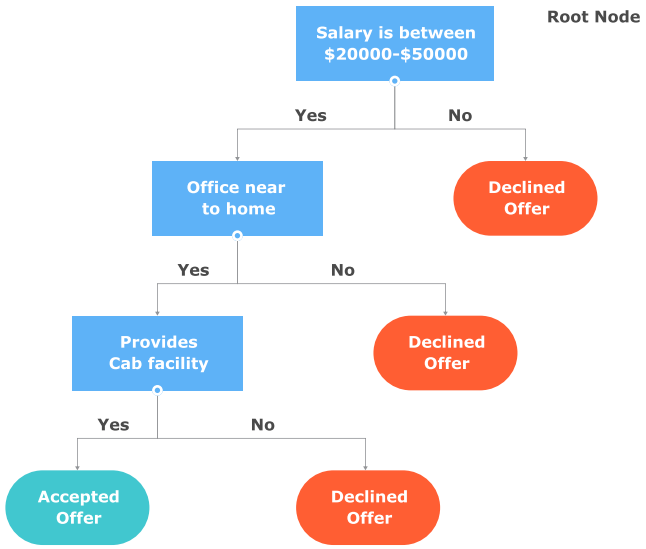
1. Linear Regression Algorithm

This popular machine learning algorithm is based on supervised learning. It uses regression, which models target values based on independent variables. The visualization is in the form of a linear equation, which has a relationship between predictive output and set inputs. Linear regression is mostly used in predictions and forecasting applications.



2. Decision Tree

A decision tree is another supervised algorithm, and it can be used for regression and classification problems. With this algorithm, the visualization is in a "tree" format, meaning that each node is a feature, every branch is a decision, and the "leaves" are outcomes. In a decision tree, we start at the tree's root and compares the root attribute values with those of the record attribute. We follow the branch according to the value and move on to the next node. We continue on like this, comparing values until we get to the leaf node with predicate class value.



3. K-Means Clustering

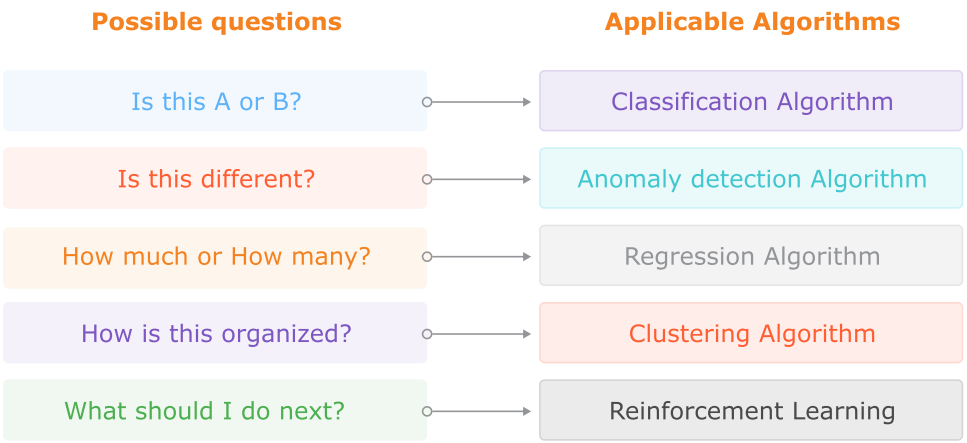
K means clustering is a kind of unsupervised learning, in which you have unlabeled data. According to Oracle, this algorithm aims to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of the K-groups based on the features that are provided. Data points are clustered based on feature similarity. Results include:

- K-cluster centroids are made and used to label new data
- Each data point is assigned to a single cluster

Besides these algorithms, principal component analysis, Naïve Bayes, support vector machine, artificial neural networks, and Apriori are also used.

Solving Data Science Problems With Machine Learning Algorithms

See the diagram below for applicable kinds of algorithms for Data Science questions:



QUESTION	MEANING
Is this A or B?	Has two fixed solutions (Yes or No; May or May Not, etc.)
Is this different?	Is a piece of data an anomaly from a pattern?
How much or how many?	Asks for numerical figures or values like the temperature or time
What should I do next?	Small decisions need to be made without human guidance

ALGORITHM	MEANING
Classification	Predictive calculations that analyze sets of data and assign them to preset categories
Anomaly Detection	Identifies rare events, items, and observations that are significantly different from the rest of the data
Regression	Data fed into the system, input features are detected, and output values are predicted
Clustering	Involves automatically discovering natural data groups
Reinforcement Learning	Finds the best possible path to find a reward (e.g., the most reward with the least hurdles)

Main Tech Architectures, Tools, Stacks Used

A Data Science tech stack doesn't just contain the runtime for inference jobs or the model framework – it extends to business intelligence tools, how models are deployed, and the entire data engineering pipeline. We'd like to bring your attention to critical areas that must be considered when creating your Data Science tech stack.

Data Warehouse

The data warehouse choice depends on whether you need a cloud-based or on-premise solution. Cloud-based software's main draw is that you don't have to perform maintenance; you can focus on the core analytics challenge without distractions or wasting money on hardware. Popular cloud-based data warehouse solutions include Azure, BigQuery, and Redshift.

On the other hand, on-premise solutions allow you to retain complete control over your data. A popular on-premise data warehouse solution is to combine Spark, Tez, or execution engine with a querying layer like Presto or Hive.

ETL Tool

Any machine learning model or analytics module is only as good as its input features. And a great ETL (Extract, Transform, Load) tool creates quality input features. Spark QQL in Python and Scala or Spark-based transformation functions with custom code are popular choices for on-premise solutions. In these cases, you must build your own schedulers and frameworks to be sure that the feature-building process is reliable. You could also use data integration from Pentaho or another open-source tool – but this isn't as flexible as a custom solution.

If you're looking for a cloud-based solution instead, AWS Glue, Azure Data Bricks, and Google Cloud Dataflow are great SaaS options. Each of these supports native Data Science modeling and automatic code generation based on visual interfaces. However, the main disadvantage is that each solution is optimized for its own stack – for instance, Glue is better used with an AWS stack.

Business Intelligence and Visualization Tools

Business intelligence and visualization tools play a crucial role in exploratory data analysis. Some popular on-premise solutions include Microsoft Power BI and Tableau. If you want custom code-based solutions, Seaborn, Matplotlib, and similar Python libraries are good data visualization options.

Some SaaS alternatives in this section include Azure Data Explorer, Google Data Studio, and AWS Quicksight. AWS's solution uses machine learning capabilities to forecast values, detect anomalies, and create automatic dashboards. Using SaaS makes sense if you're already on the provider's stack.

Machine Learning and Analytics Implementation Frameworks

Python has been the go-to for custom code-based ML and analytics implementation. Scikit-learn and stats-model are popular choices for statistical analysis and modeling. R is also a good choice for statistical models due to its rich set of functions and deployment in production.

MXNet, Pytorch, and TensorFlow can be used for deep learning. And if you prefer Java, Deeplearning4j library is a solid choice. When choosing a deep learning solution, you need to consider community support because, in most cases, the developer will need to do a lot of research before finalizing the model pipeline.

If your organization isn't planning on developing custom models or hiring ML experts, most cloud service providers also offer automated model building as a service. Google Cloud AI, Azure Machine Learning, and AWS Machine Learning Services enable you to build models without too much coding. Azure ML Studio, AWS Sagemaker, and Google Data Lab are good platforms for Data Science development; however, remember that any machine learning implementation is limited by its input features, so your ETL tool is of the utmost importance here.

Deployment Stack

After the models are built, you must deploy them for batch or real-time inferences. With an on-premise setup, the common choice is to wrap models in Django, Flask, or a similar web framework and run it inside Docker containers. They can be horizontally scaled via a load balancer or a container orchestration framework.

The deciding factor for a solution here is the amount of effort involved and the necessary expertise. Inference modules can be highly complex and batching, threading, and other concepts must be carefully applied to achieve the best performance. TensorFlow, Pytorch, and Apache MXNet all have their own deployment functions, so you don't have to reinvent the wheel.

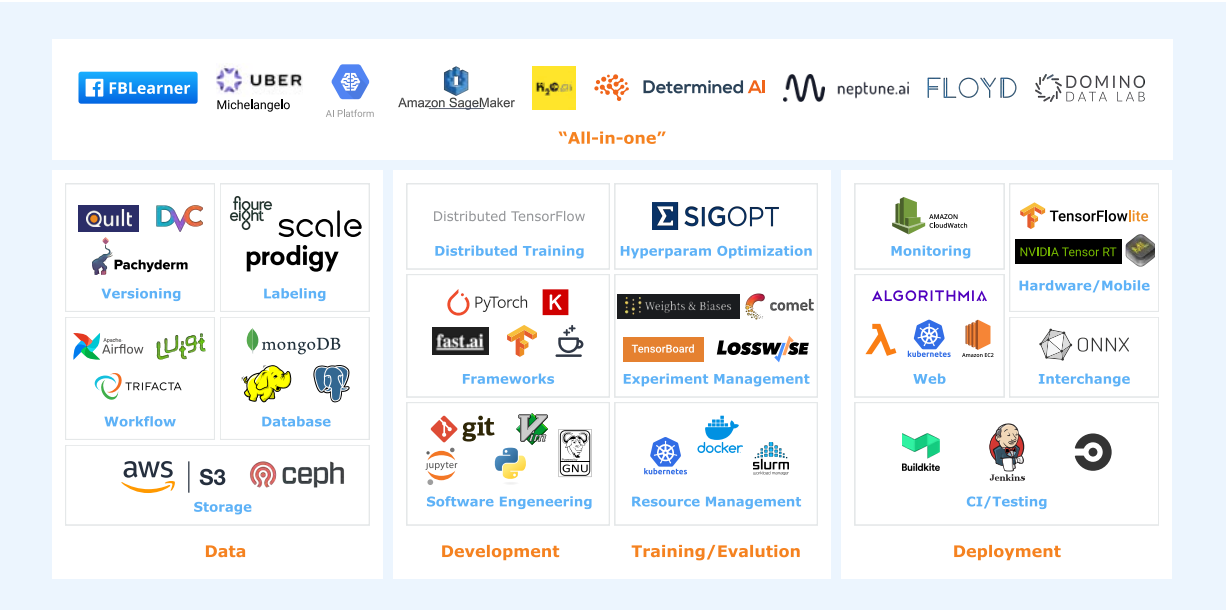
And, of course, to avoid a complicated deployment process, you can use a cloud provider's ML-serving option. GCP, Azure, and AWS all have built-in deployment mechanisms as part of their machine learning services. They also allow the deployment of custom models that were created outside of their systems. The prime advantage here is that scaling is totally automated with these solutions.

Stack Conclusion

Choosing your Data Science stack isn't simple – there are several questions you must ask yourself, including:

- Do you need cloud-based or on-premise services?
- Are you already working with one of the cloud service providers?
- Are you capable of creating your own analytics functions and models?
- Do you need real-time data ingestion and analytics?

For your reference, here are examples for "all-in-one" Data Science stacks:



intetics
Where software concepts come alive™

Case studies

We've gathered a few real-life data case studies that we have worked on, so you can see first-hand how companies leverage Data Science to boost profits, productivity, and more.

Building Footprint Detection

Objective

> To build a solution for feature extraction from aerial photos.

Challenges

- Extract building footprints from aerial photos
- Edge angles of the building footprint should be close to 90 degrees
- Consecutive angles of the building footprint should not be very sharp

Technologies

> Python, PyTorch, OpenC

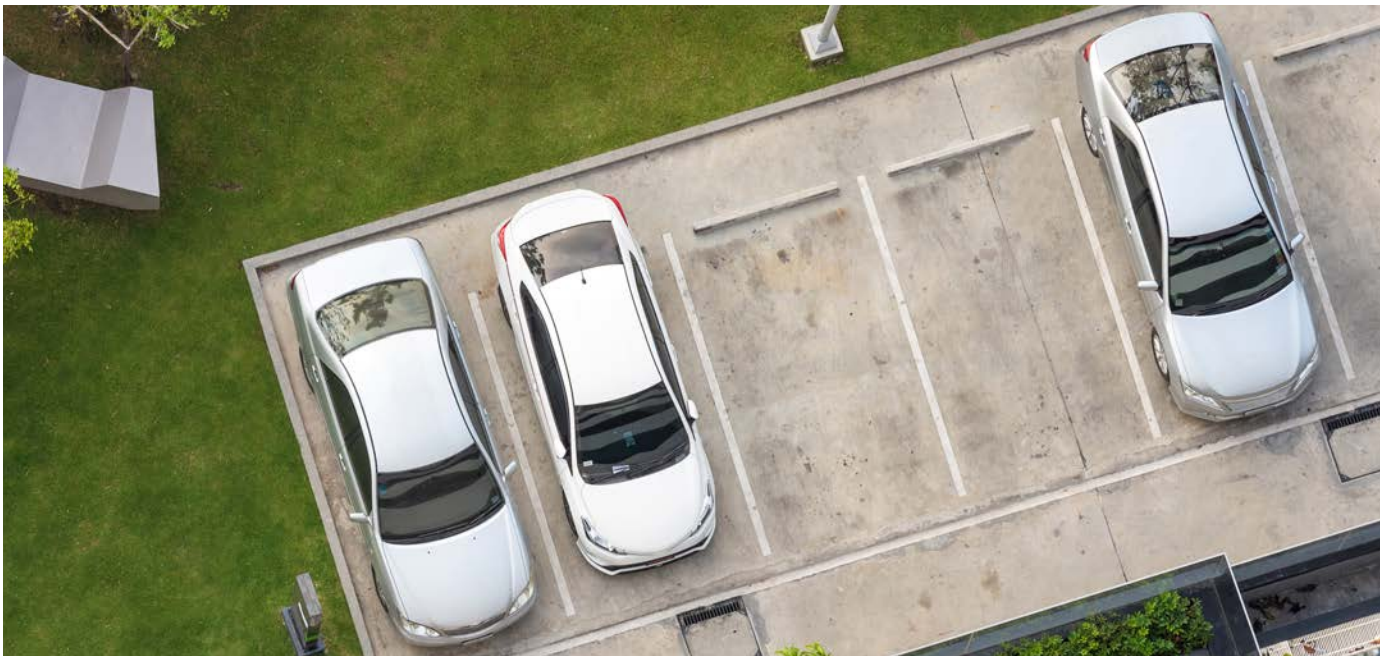
Solution

- We developed a working solution allowing the Client to benefit from the sale of not only aerial photos but also extracted data
- Footprint detection was developed using a semantic segmentation approach and pixel-mask polygonization
- Building polygons were received by approximation of the predicted pixels with the use of the Douglas-Peucker algorithm

Benefits

- ★ The application allowed the Client to get building polygons from high-resolution aerial imagery
- ★ Hundreds of hours of manual labeling have been optimized

Parking Detection From Satellite Imagery



Objective

To detect parking from satellite imagery

Challenges

- Split moving and static objects (cars) to understand if an object is a parked vehicle or a road
- Extract parking from satellite imagery

Technologies

Python, PyTorch, OpenCV

Solution

We built and trained a model that recognizes on-street and off-street parking in US cities.

Benefits

★ Saving resources on manual online parking detection and the work of field teams

Bus Predictive Maintenance

Objective

To monitor the following metrics:

- Battery-pack-level diagnostics
- Wear of powertrain components and bearing systems
- When oil needs to be refilled
- The health of engines, motors, and electric generators (in case of electric or hybrid vehicles)

Technologies

Python, Pandas, Sklearn, XGboost

Solution Steps

1. The set of IoT device with required sensors was defined
2. Initial data was collected and labeled
3. The model parameters were adjusted to suit the training dataset
4. The model was tested on real-world data to check its performance quality
5. The working model was deployed to the customer's server

Benefits

- ★ Predict breakdowns, real-time alerts for routine maintenance
- ★ Limits the time vehicles are in maintenance shops, reducing operational costs
- ★ Optimizes idle time
- ★ Avoid additional maintenance costs related to any vehicle component described above
- ★ Real-time telemetry dashboard and vehicle location tracking

Standards in Use

As the role of Data Science in business is expanding rapidly, the demand for Data Scientists is likewise growing. However, as nearly every business has its own way of assigning data analytics titles and defining job roles, this has resulted in a chaotic market, confusing to Data Science professionals, employers, and training institutions alike.

Therefore, various initiatives have been formed to develop Data Science standards. For instance, the Initiative for Analytics and Data Science Standards (IADSS) aims to build a framework for crucial Data Science skills and support the creation of measurement and assessment methodologies.

The National Institute of Standards and Technology (NIST) has launched its own Data Science Research Program, aiming to accelerate research on analytic data methods.

Lastly, the Data Science Association sponsors the Data Science Standards Committee, which is committed to developing standards for various specialty domains, including AI, NLP, neural networks, business analytics, and the mining, management, storage, processing, sharing, and visualization of big data.

Data Science Professional Communities to Join

- **Harvard Data Science Review:**
  This provides open access to the Harvard Data Science Initiative, featuring research milestones, foundational thinking, major applications, and educational innovations of Data Science. Its main emphasis is on replicability, readability, and reproducibility of content. This premier research journal serves as the crossroads between Data Science research and applications that are societally important.

- **Data Science Association:**
  It is a non-profit association for data scientists; it aims to eliminate bias in the field, improve the profession, and advance ethical Data Science worldwide.

- **Data Science Council of America:**
  Here, members can earn Data Science certifications and subscribe for insights about Big Data.

- **IBM Data Science Community:**
  Interact with other data scientists, read monthly newsletters on the latest industry news, and join specialized groups, such as their Data and AI Learning Group.

Data Science Authorities to Follow

People

Randy Lao is a Data Science mentor at Data Science Dream Job – an online educational platform that helps people get jobs in the Data Science field.

Kyle McKiou is the founder of the abovementioned Data Science Dream Job; he also regularly shares his insights and Data Science experience on Linkedin.

Favio Vásquez hosts informative webinars on the YouTube channel, Data Science Office Hours, along with other industry leaders.

Kirill Eremenko is the CEO and founder of SuperDataScience, another online educational platform for data scientists. His platform features numerous analytic courses covering Machine Learning A-Z, R Programming, Tableau, Python, and more.

Platforms

SAS offers a suite of Data Science products and advanced analytics. Users can gain access to any format of data from all sources, automated data preparation, and model/data lineage management through the platform. SAS's Machine Learning and Visual Data Mining generate automatic insights for models' common variables. Project summaries can also be created with natural language generation.

IBM Watson Studio allows scalable AI models to be built, run, and managed across any cloud. Watson Studio is offered as part of IBM's Cloud Pak for Data. With this solution, users can prepare and build models visually, use one-click integration to deploy and run said models, and use explainable AI for monitoring and management. The solution enables users to use open-source frameworks, including TensorFlow, PyTorch, and scikit-learn.

Daitaku has a unified framework, enabling organizations to immediately access features necessary for designing data tools from scratch. Then, users can apply Data Science and machine learning techniques to build and run predictive data flows.

Databricks has a unified Apache Spark and cloud-based platform, which combines Data Science and data engineering functionality. Its Data Science Workspace lets users collaborate on exploring data and building models. Users can also get one-click access to preconfigured machine learning environments using popular frameworks.

Available Certifications for Practitioners

Data Science is one of the most in-demand sectors for IT jobs, as companies are becoming increasingly reliant on data. If you are interested in getting into this lucrative field or want to get an edge over your competition, certifications may be the solution. We've compiled several top Data Science certifications:

- **Certified Analytics Professional (CAP):** This vendor-neutral certification verifies that you can transform complex data into valuable insights and express them to stakeholders. To qualify for this exam, you will need five years of experience with a related Bachelor's degree or three years of experience with a Master's degree. Without any related degree, you will need seven years of experience. The standard cost for the CAP exam is $695, and the certification is valid for three years.

- **Senior Data Scientist Certification from Data Science Council of America:** This certification program is made for professionals who have 5+ years of experience in research and analytics. Students should have knowledge of spreadsheets, databases, statistical analytics, quantitative methods, R, SPSS/SAS, and the basics of object-oriented programming. The certification costs $650, and it is valid for five years.

- **Principal Data Scientist Certification from Data Science Council of America:** This certification is for professionals who have 10+ years of big data experience. The exam includes data business strategies, machine learning, big data best practices, scholastic modeling, natural language processing, and more. Depending on the track chosen, certification costs range from $300 - $950. The certification does not expire.

- **Dell EMC Data Scientist Track:** This certification includes two programs: Associate level and Specialist level. The former covers the foundations of big data analytics and Data Science, while the latter covers natural language processing, visualization methods, Pig, Hive, Hadoop, HBase, and advanced analytical methods. Each exam costs $230, and the certifications do not expire.

- **IBM Data Science Professional Certificate:** This certification program contains nine courses: open-source tools, Data science, methodologies, databases/SQL, Python, data analysis, visualization, machine learning, and an applied capstone project course. It can be taken for free via Coursera, although you would need to pay Coursera's standard certification fee upon completing the program. The credentials for this certification do not expire.

Healthcheck

With all of the business insights that Data Science can bring, it might seem wise to hire a Data Scientist as soon as possible. However, this is not always ideal. Very young startups will often only have basic data infrastructure in place, and they aren't ready to use precious resources on more advanced data products and analytics. You've got to focus on keeping the website backend running, tracking how users use your products, and keeping signups flowing.

You'll also need storage for central transactional analytics, which should scale for the next year or so. Unless you have a good reason, opt for relational databases; they might not be flashy, but they're a safe choice. Centralize your data, and create better ETLs for data marshaling. You'll want to hire a data engineer for these processes, which is significantly cheaper than a data scientist.

So, after hiring a data engineer and strengthening your data infrastructure, is it time to hire a data scientist? Not quite. Next, you should hire a data analyst. A growing company is most likely still developing its business model, feeling things out, figuring out where to get funding, and trying to strategize on where to go. Traditional analytics can help answer these questions – a good data analyst will go a long way towards making the right business decisions. And if they can do some modeling in R or know SQL, even better. And, just like with a data engineer, a data analyst is much cheaper than a data scientist.

Now, after following the above steps, you have decent data infrastructure, reasonably solid data quality processes, and the startup founders' basic data needs are met. And, even now, it still might not be time to hire a data scientist. Truly, it depends on whether Data Science is a central part of your business model. Let's say you could only hire one more person on your team – in that case, what role would give the biggest return?

If your business's primary offering is a data product or a data-science-driven process (like a recommendation engine), then it might indeed be a good time to hire a data scientist. But, if not, perhaps it would be better for you to hire another analyst.

To sum up: don't just hire a data scientist for the sake of having one on board. You need to have a good idea of why you need to add them to the team. Be able to justify the costs – because making a poor team model decision can be the bane of startups.

Another option is to outsource your Data Science . This is often more affordable than hiring your own data scientist, and it allows for access to a wider talent pool, greater flexibility, and easy scaling.

Further Reading

Interested in learning more about Data Science? We recommend checking out the following resources:

- **Towards Data Science:** This is Medium's biggest publication on Data Science topics. Read in-depth descriptions of specific ML algorithms, latest industry news, summaries of scientific papers, and beginner-friendly tutorials with code.

- **PyData:** This is NumFOCUS's education program, which promotes open practices in data, research, and scientific computing. You can view their talks to get examples of real-life Data Science cases, introductions to new libraries, and general Python best practices.

- **Papers With Code:** This is a free, open-resource pool of ML papers, code, and evaluation tables. It is great for when you would like to experiment and apply innovative approaches to your dataset without writing the code yourself.

- **Naked Statistics: Stripping the Dread From Data:** Learn how statistical concepts apply in real life with interesting, often humorous examples.

Interesting to Know

Data Science Quotes From Thought Leaders

"Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician."

—Josh Wills, Director of Data Engineering at Slack

"What sort of personality makes for an effective data scientist? Definitely curiosity.... The biggest question in Data Science is 'Why?' Why is this happening?"

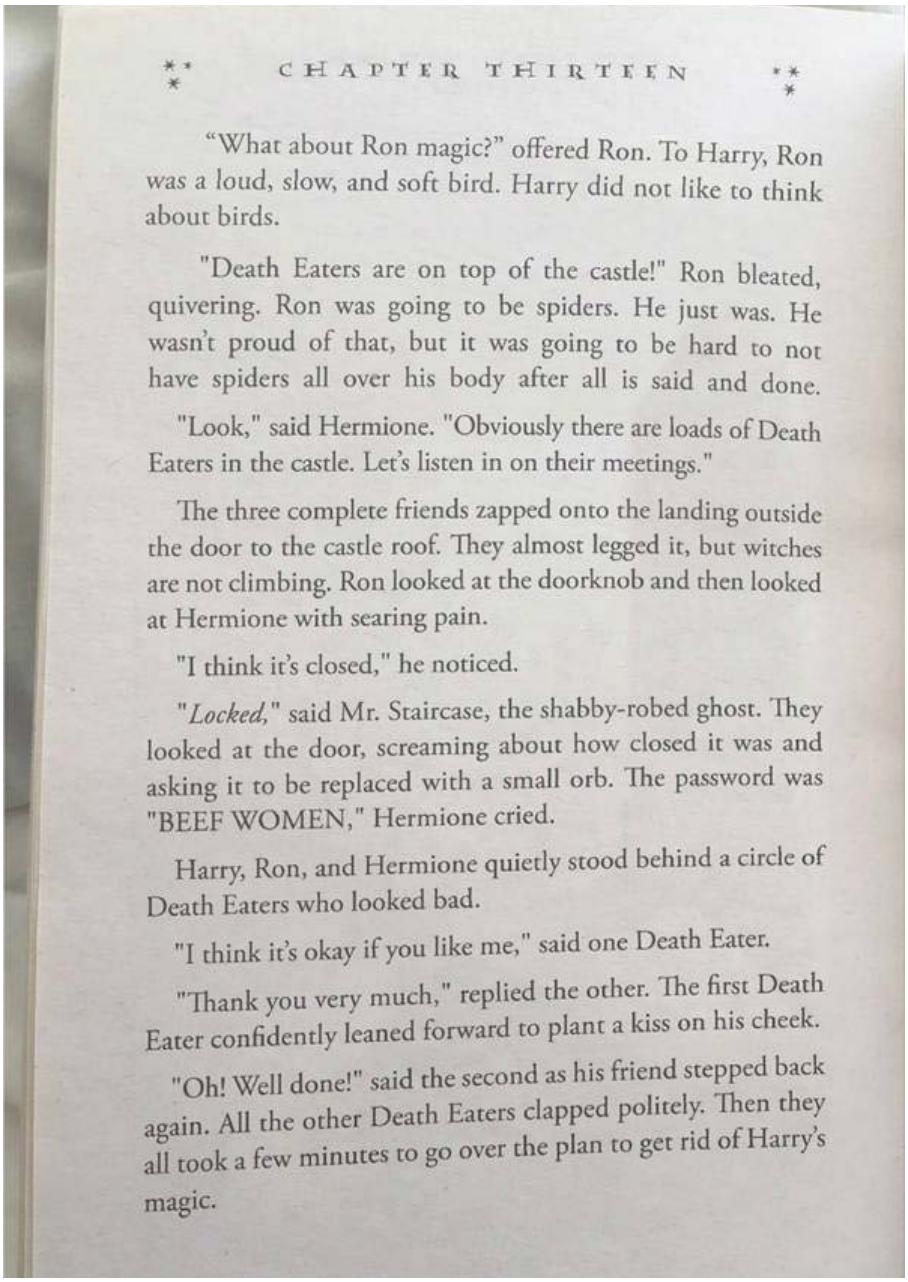
—Carla Gentry, Data Scientist at Talent Analytics

"There is no bottleneck for data scientists... The bottleneck is very often for companies who don't have a culture of working with data to actually cut down the process into the right steps."

—Lutz Finger, Director of Data Science at Snap

Weird Data Science: Harry Potter and the Portrait That Looked Like a Large Pile of Ash!

Botnik Studio, a community of writers, used natural language processing to generate a Harry Potter chapter! Members fed a predictive algorithm with text from all 7 Harry Potter books. In just a few moments, the algorithm determined a pattern of words and began making suggestions. The results turned out to be pretty funny!



CHAPTER THIRTEEN

"What about Ron magic?" offered Ron. To Harry, Ron was a loud, slow, and soft bird. Harry did not like to think about birds.

"Death Eaters are on top of the castle!" Ron bleated, quivering. Ron was going to be spiders. He just was. He wasn't proud of that, but it was going to be hard to not have spiders all over his body after all is said and done.

"Look," said Hermione. "Obviously there are loads of Death Eaters in the castle. Let's listen in on their meetings."

The three complete friends zapped onto the landing outside the door to the castle roof. They almost legged it, but witches are not climbing. Ron looked at the doorknob and then looked at Hermione with searing pain.

"I think it's closed," he noticed.

"*Locked,*" said Mr. Staircase, the shabby-robed ghost. They looked at the door, screaming about how closed it was and asking it to be replaced with a small orb. The password was "BEEF WOMEN," Hermione cried.

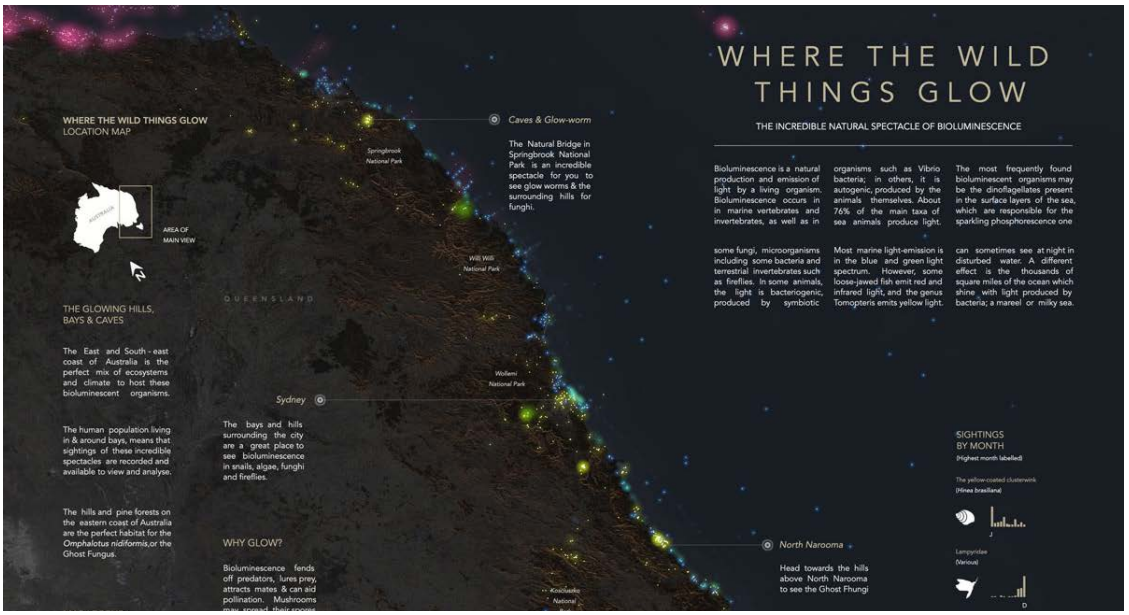
Harry, Ron, and Hermione quietly stood behind a circle of Death Eaters who looked bad.

"I think it's okay if you like me," said one Death Eater.

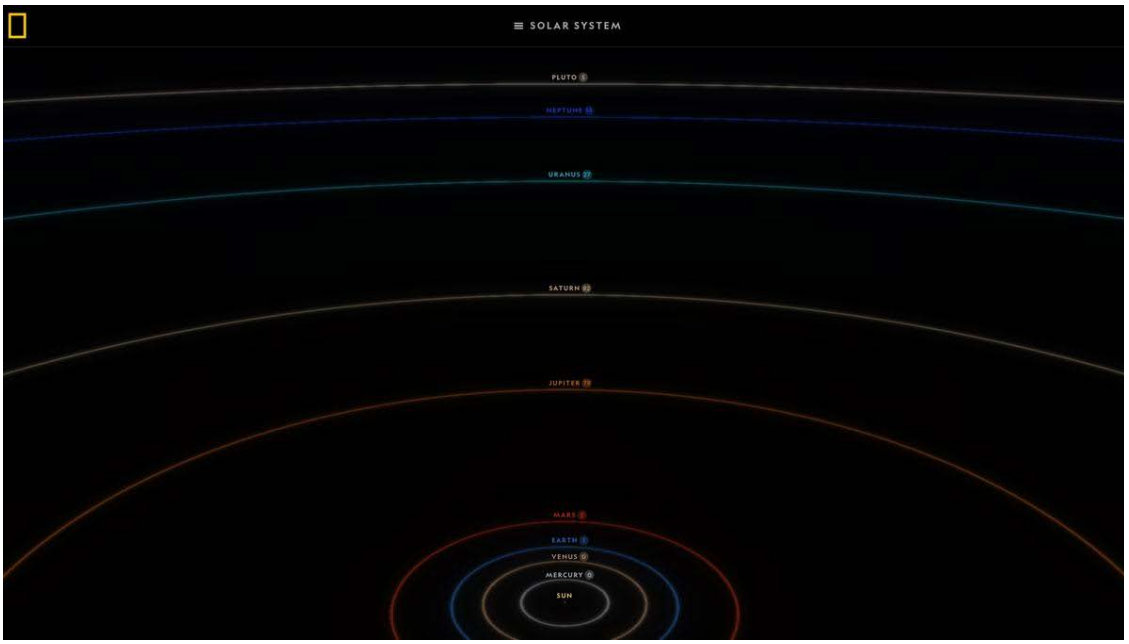
"Thank you very much," replied the other. The first Death Eater confidently leaned forward to plant a kiss on his cheek.

"Oh! Well done!" said the second as his friend stepped back again. All the other Death Eaters clapped politely. Then they all took a few minutes to go over the plan to get rid of Harry's magic.

Best Data Visualizations of 2020



We love this "Where the Wild Things Grow" visualization; it was created for National Geographic and showed the presence of bioluminescence along Australia's Southeastern coat.



This is another excellent data visualization from National Geographic, titled "The Atlas of the Moons." It's a scrollable visualization that takes the viewer on a journey through space, starting with our own moon.

We love this "Where the Wild Things Grow" visualization; it was created for National Geographic and showed the presence of bioluminescence along Australia's Southeastern coat.

Clousing Thoughts

The biggest advantage of Data Science is that it can draw valuable conclusions from huge piles of seemingly unrelated data. It is a great way to utilize the overwhelming amount of information. 90% of all data in history has been created in the last decade – and data production will only continue to grow from here. This suggests that companies that adopt Data Science analysis can enjoy increased revenue and business expansion. Data Science's scope grows with each passing year, especially in the healthcare, transport, and e-commerce sectors. And as organizations are turning towards ML, AI, and big data, the job market for data scientists is skyrocketing – now is a great time to get in on the action.

intetics
Where software concepts come alive™

INTETICS MEANS YOUR SUCCESS

Toll Free: +1 (877) SOFTDEV
US: +1 (239) 217-4907
DE: +49 (211) 3878-9350
UK: +44 (20) 3514-1416
Email: intetics@intetics

www.intetics.com