

# KDnuggets

[Subscribe to KDnuggets News](#)[Contact](#)

search KDnuggets

Search



- [SOFTWARE](#)
- [News/Blog](#)
- [Top stories](#)
- [Opinions](#)
- [Tutorials](#)
- [JOBS](#)
- [Companies](#)
- [Courses](#)
- [Datasets](#)
- [EDUCATION](#)
- [Certificates](#)
- [Meetings](#)
- [Webinars](#)

[KDnuggets Home](#) » [News](#) » [2018](#) » [Jun](#) » [Tutorials, Overviews](#) » Data Science Predicting The Future  
( [18:n24](#) )

## Data Science Predicting The Future

[◀ Previous post](#)[Next post ▶](#)

Like 20



Share 19

Tags: [Data Science](#), [Forecasting](#), [Machine Learning](#), [Programming Languages](#), [Regression](#)

In this article we will expand on the knowledge learnt from the last article - The What, Where and How of Data for Data Science - and consider how data science is applied to predict the future.

---

By [Iliya Valchanov](#), 365 Data Science.

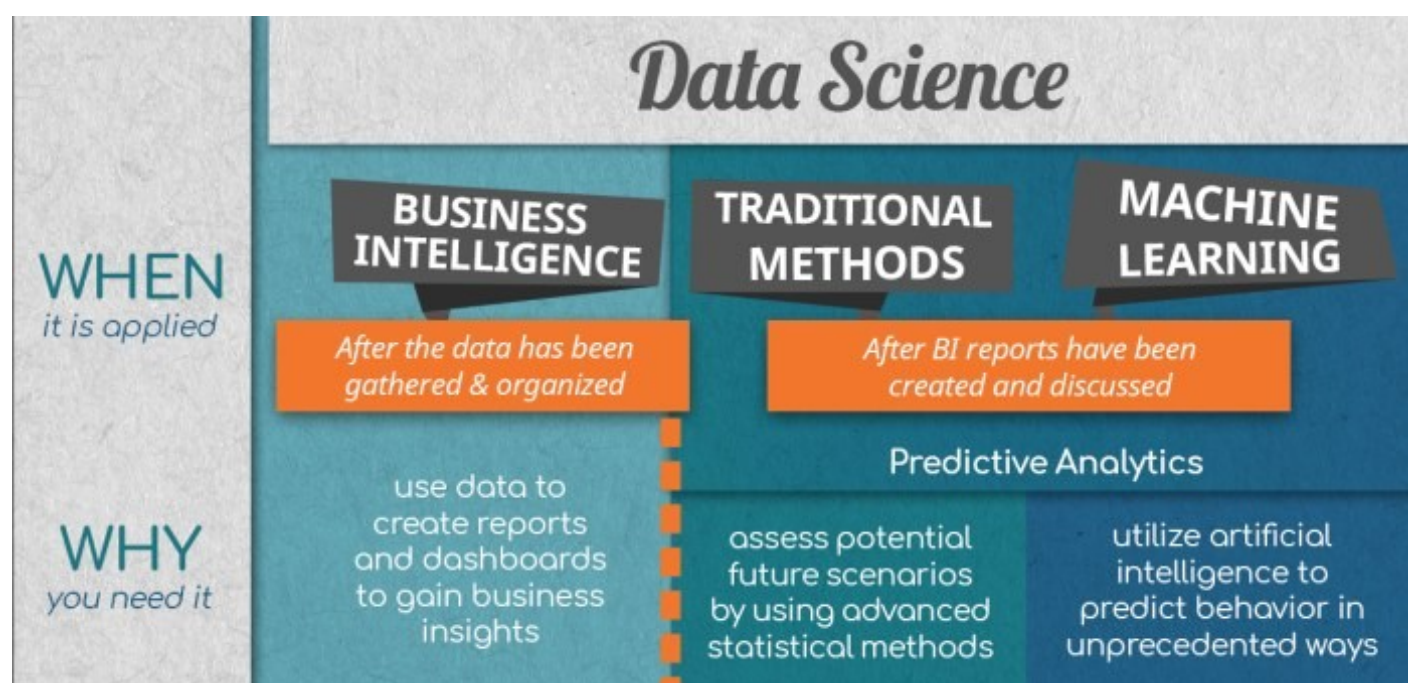
[comments](#)

Predictive analytics in data science rest on the shoulders of explanatory data analysis, which is precisely what we were discussing in our previous article – [The What, Where and How of Data for Data Science](#). We talked about data in data science, and how business intelligence (BI) analysts use it to explain the past.

In fact, everything is connected. Once the BI reports and dashboards have been prepared and insights – extracted from them – this information becomes the basis for predicting future values. And the accuracy of these predictions lies in the methods used.

**Recall the distinction between traditional data and big data in data science or refer to our first article on the What-Where-How of Data science.**

We can make a similar distinction regarding predictive analytics and their methods: traditional data science methods vs. Machine Learning. One deals primarily with traditional data, and the other – with big data.

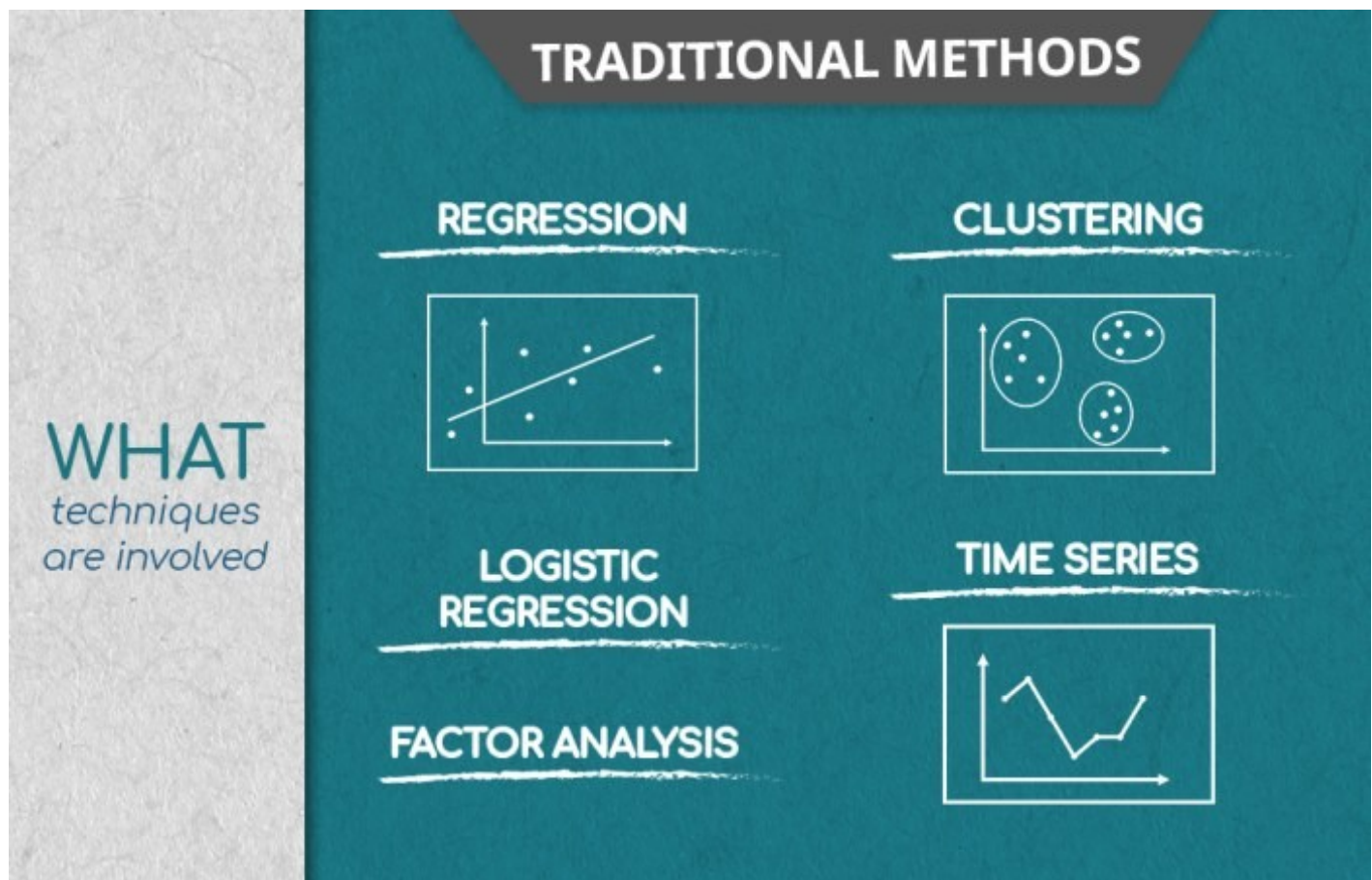


## Traditional forecasting methods in Data Science: What are they?

Traditional forecasting methods comprise the classical statistical methods for forecasting – linear regression analysis, logistic regression analysis, clustering, factor analysis, and time series. The output of each of these feeds into the more sophisticated machine learning analytics, but let's first review them individually.

A quick side-note. Some in the data science industry refer to several of these methods as machine learning too, but in this article machine learning refers to newer, smarter, better methods, such as deep learning.

SHARES



### Linear regression

In data science, the linear regression model is used for quantifying causal relationships among the different variables included in the analysis. Like the relationship between house prices, the size of the house, the neighborhood, and the year built. The model calculates coefficients with which you can predict the price of a new house, if you have the relevant information available.

### Logistic regression

Since it's not possible to express all relationships between variables as linear, data science makes use of methods like the logistic regression to create non-linear models. Logistic regression operates with 0s and 1s. Companies apply logistic regression algorithms to filter job candidates during their screening process, for instance. If the algorithm estimates that the probability that a prospective candidate will perform well in the company within a year is above 50%, it would predict 1, or a successful application. Otherwise, it will predict 0.

### Cluster analysis

This exploratory data science technique is applied when the observations in the data form groups according to some criteria. Cluster analysis takes into account that some observations exhibit similarities, and facilitates the discovery of new significant predictors, ones that were not part of the original conceptualization of the data.

### Factor analysis

If clustering is about grouping observations together, factor analysis is about grouping features together. Data science resorts to using factor analysis to reduce the dimensionality of a problem. For example, if in a 100-item questionnaire each 10 questions pertain to a single general attitude, factor analysis will identify these 10

SHARES

## Time series analysis

Time series is a popular method for following the development of specific values over time. Experts in economics and finance use it because their subject matter is stock prices and sales volume – variables that are typically plotted against time.

## Where does data science find application for traditional forecasting methods?

The application of the corresponding techniques is extremely broad; data science is finding a way into an increasingly large number of industries. That said, two prominent fields deserve to be part of the discussion.

### User experience (UX) and data science

When companies launch a new product, they often design surveys that measure the attitudes of customers towards that product. Analysing the results after the BI team has generated their dashboards includes grouping the observations into segments (e.g. regions), and then analyzing each segment separately to extract meaningful predictive coefficients. The results of these operations often corroborate the conclusion that the product needs slight but significantly different adjustments in each segment in order to maximize customer satisfaction.

### Forecasting sales volume

This is the type of analysis where time series comes into play. Sales data has been gathered until a certain date, and the data scientist wants to know what is likely to happen in the next sales period, or a year ahead. They apply mathematical and statistical models and run multiple simulations; these simulations provide the analyst with future scenarios. This is at the core of data science, because based on these scenarios, the company can make better predictions and implement adequate strategies.

## Who uses traditional forecasting methods?

The data scientist. But bear in mind that this title also applies to the person who employs machine learning techniques for analytics, too. A lot of the work spills from one methodology to the other.

The data analyst, on the other hand, is the person who prepares advanced types of analyses that explain the patterns in the data that have already emerged and overlooks the basic part of the predictive analytics.

## Machine Learning and Data Science

Machine learning is the state-of-the-art approach to data science. And rightly so.

The main advantage machine learning has over any of the traditional data science techniques is the fact that at its core resides the algorithm. These are the directions a computer uses to find a model that fits the data as well as possible. The difference between machine learning and traditional data science methods is that we do not give the computer instructions on how to find the model; it takes the algorithm and uses its directions to learn on its own how to find said model. Unlike in traditional data science, machine learning needs little human involvement. In fact, machine learning, especially deep learning algorithms are so complicated, that humans cannot genuinely understand what is happening “inside”.

To be clear, here we must note that machine learning methods STEP ON traditional ones. Supervised learning, for example, has two subtypes – regression and classification (e.g. multinomial logistic regression). Naturally, many traditional methods also fall under the ‘machine learning’ umbrella term. That is logical because linear

SHARES

Clustering and PCA, on the other hand, are unsupervised learning algorithms (with PCA the debate is even fiercer than with regressions).

Either way, the distinction between traditional methods and ML is more or less subjective. Some draw a line, others don't. In our framework, the simplicity (which is also elegance in a way) of traditional methods is the main reason for the distinction. An interesting point of view on that issue can be explored here:







<https://www.kdnuggets.com/2017/06/regression-analysis-really-machine-learning.html>

Finally, deep learning is very computationally expensive compared to traditional methods. To give you some context, I've seen works where linear regressions were worked out on paper, by hand.

So, for me, the line is drawn at: can you create a CNN and work it out on paper in some rational time? Not really, so, that's something I'd label machine learning.

## What is machine learning in data science?

A machine learning algorithm is like a trial-and-error process, but the special thing about it is that each consecutive trial is at least as good as the previous one. But bear in mind that in order to learn well, the machine has to go through hundreds of thousands of trial-and-errors, with the frequency of errors decreasing throughout.

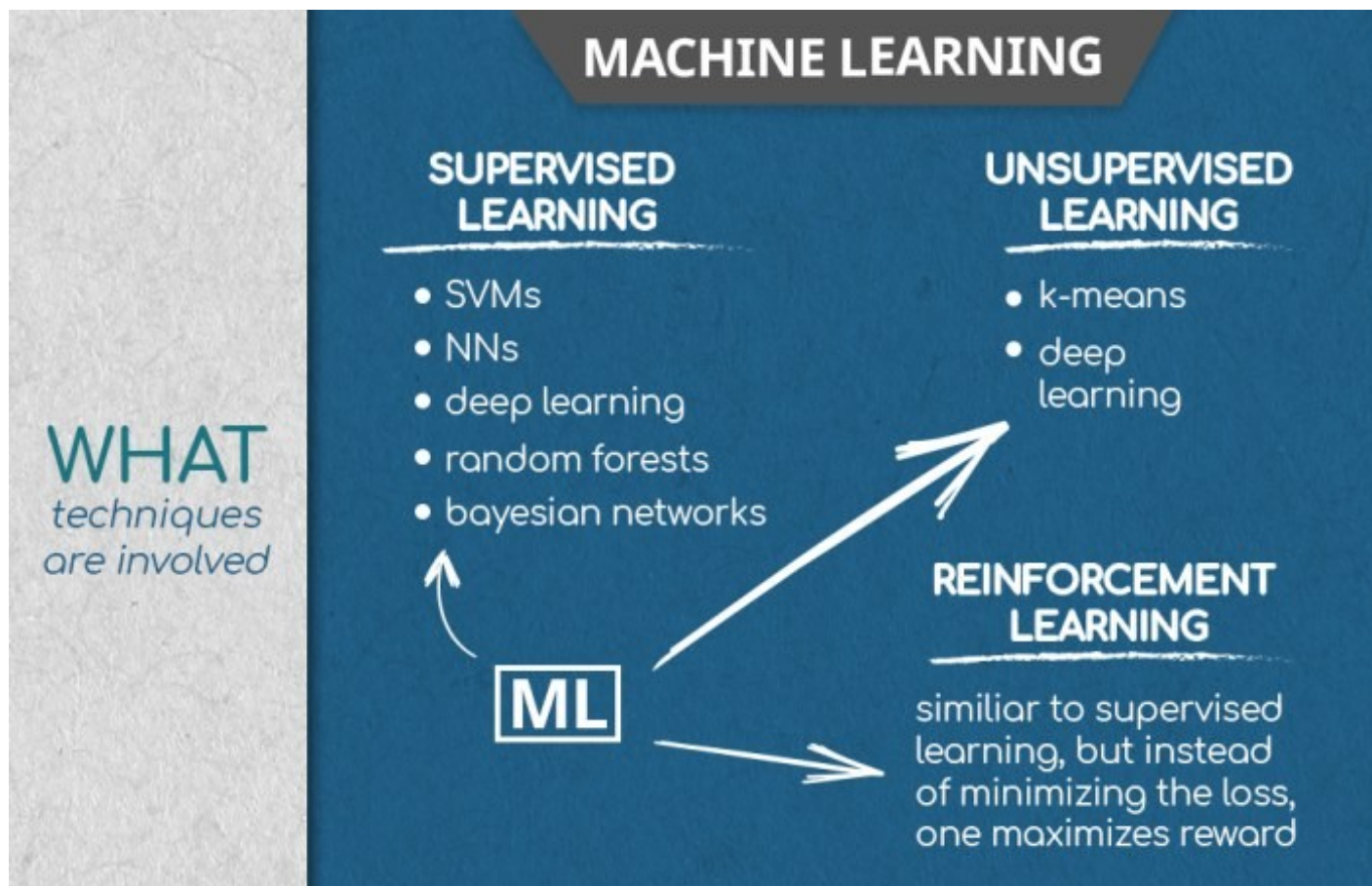
OUTPUT	CORRECT VALUE	OBJECTIVE FUN.	VALUE
		Far from reality	200
		Closer	100
		Very close	0

Once the training is complete, the machine will be able to apply the complex computational model it has learned to novel data still to the result of highly reliable predictions.

There are three major types of machine learning: supervised, unsupervised, and reinforcement learning.

SHARES





### Supervised learning

Supervised learning rests on using labeled data. The machine gets data that is associated with a correct answer; if the machine's performance does not get that correct answer, an optimization algorithm adjusts the computational process, and the computer does another trial. Bear in mind that, typically, the machine does this on 1000 data points at once.

Support vector machines, neural networks, deep learning, random forest models, and Bayesian networks are all instances of supervised learning.

### Unsupervised learning

When the data is too big, or the data scientist is under too much pressure for resources to label the data, or they do not know what the labels are at all, data science resorts to using unsupervised learning. This consists of giving the machine unlabeled data and asking it to extract insights from it. This often results in the data being divided in a certain way according to its properties. In other words, it is clustered.

Unsupervised learning is extremely effective for discovering patterns in data, especially things that humans using traditional analysis techniques would miss.

Data science often makes use of supervised and unsupervised learning together, with unsupervised learning labelling the data, and supervised learning finding the best model to fit the data. One instance of this is semi-supervised learning.

### Reinforcement learning

SHARES

sub-optimally, the optimization algorithms do not adjust the computation. Think of a puppy learning commands. If it follows the command, it gets a treat; if it doesn't follow the command, the treat doesn't come. Because treats are tasty, the dog will gradually improve in following commands. That said, instead of minimizing an error, reinforcement learning maximizes a reward.

## Where is Machine Learning applied in the world of data science & business?

### Fraud detection

With machine learning, specifically supervised learning, banks can take past data, label the transactions as legitimate, or fraudulent, and train models to detect fraudulent activity. When these models detect even the slightest probability of theft, they flag the transactions, and prevent the fraud in real time.

### Client retention

With machine learning algorithms, corporate organizations can know which customers may purchase goods from them. This means the store can offer discounts and a 'personal touch' in an efficient way, minimizing marketing costs and maximizing profits. A couple of prominent names come to mind: Google, and Amazon.

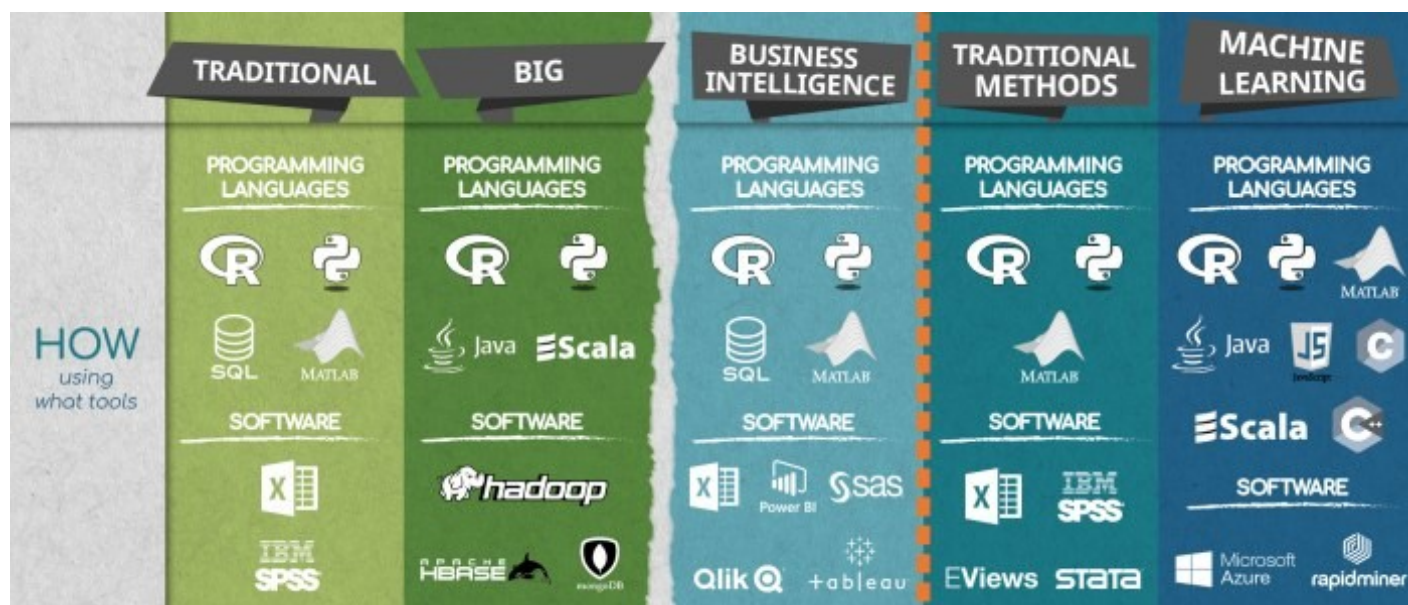
## Who uses machine learning in data science?

As mentioned above, the data scientist is deeply involved in designing machine algorithms, but there is another star on this stage.

The machine learning engineer. This is the specialist who is looking for ways to apply state-of-the-art computational models developed in the field of machine learning into solving complex problems such as business tasks, data science tasks, computer vision, self-driving cars, robotics, and so on.

## Programming languages and Software in data science

Two main categories of tools are necessary to work with data and data science: programming languages and software.



SHARES

Knowing a programming language enables the data scientist to devise programs that can execute specific operations. The biggest advantage programming languages have is that we can reuse the programs created to execute the same action multiple times.

R, Python, and MATLAB, combined with SQL, cover most of the tools used when working with traditional data, BI, and conventional data science.

R and Python are the two most popular tools across all data science sub-disciplines. Their biggest advantage is that they can manipulate data and are integrated within multiple data and data science software platforms. They are not just suitable for mathematical and statistical computations; they are adaptable.

SQL is king, however, when it comes to working with relational database management systems, because it was specifically created for that purpose. SQL is at its most advantageous when working with traditional, historical data, for example when preparing a BI analysis.

MATLAB is the fourth most indispensable tool for data science. It is ideal for working with mathematical functions or matrix manipulations.

Big data in data science is handled with the help of R and Python, of course, but people working in this area are often proficient in other languages like Java or Scala. These two are very useful when combining data from multiple sources.

JavaScript, C, and C++, in addition to the ones mentioned above, are often employed when the branch of data science the specialist is working in involves machine learning. They are faster than R and Python and provide greater freedom.

### **Software in data science**

In data science, the software or, software solutions, are tools adjusted for specific business needs.

Excel is a tool applicable to more than one category—traditional data, BI, and Data Science. Similarly, SPSS is a very famous tool for working with traditional data and applying statistical analysis.

Apache Hadoop, Apache Hbase, and Mongo DB, on the other hand, are software designed for working with big data.

Power BI, SaS, Qlik, and especially Tableau are top-notch examples of software designed for business intelligence visualizations.

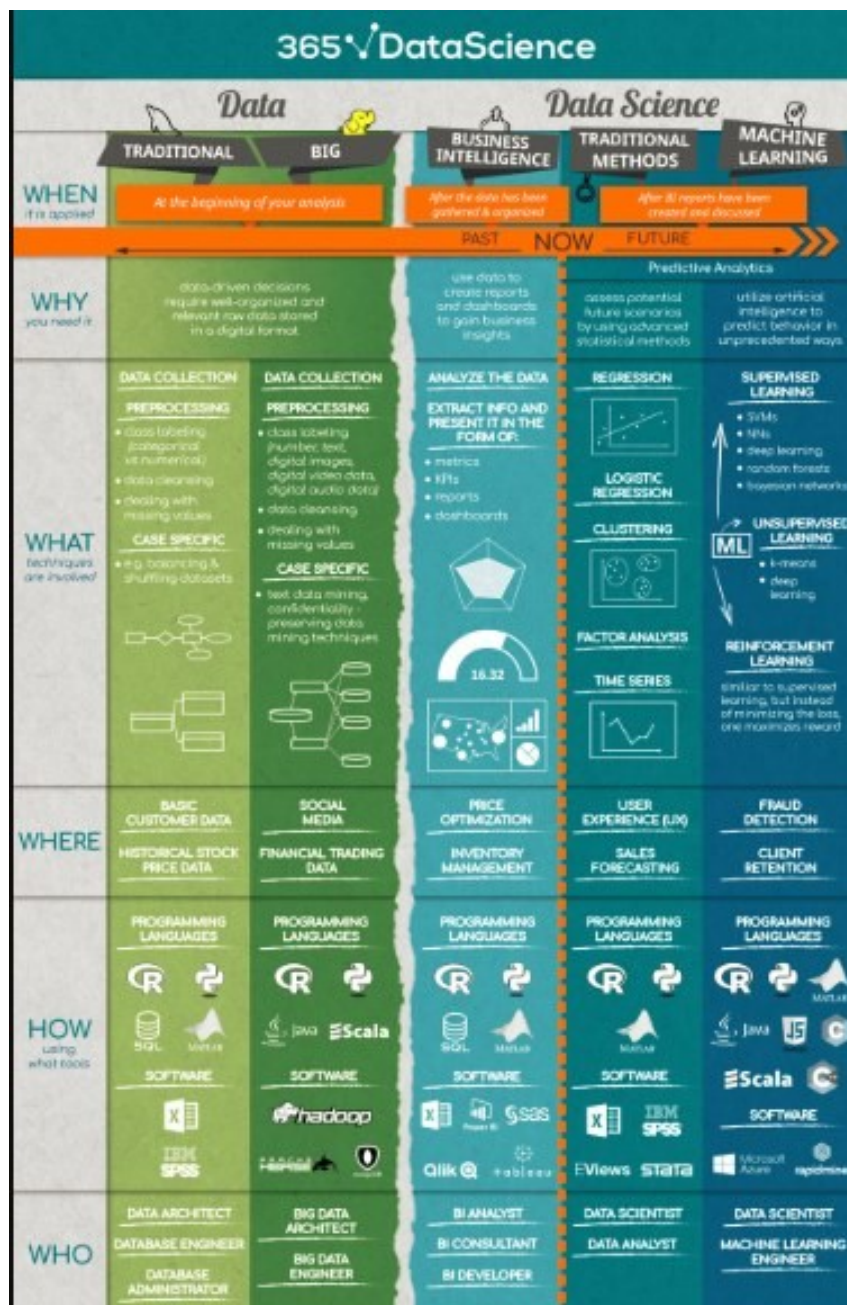
In terms of predictive analytics, EViews is mostly used for working with econometric time-series models, and Stata—for academic statistical and econometric research, where techniques like regression, cluster, and factor analysis are constantly applied.

## **This is Data Science**

Data science is a slippery term that encompasses everything from handling data – traditional or big – to explain patterns and predict behavior. Data science is done through traditional methods like regression and cluster analysis or through unorthodox machine learning techniques.

It is a vast field, and we hope you are one step closer to understanding how all-encompassing and intertwined with human life it is.





**Bio:** [Iliya Valchanov](#) is a Co-founder at 365 Data Science.

### Related:

- [Sales forecasting using Machine Learning](#)
- [Step Forward Feature Selection: A Practical Example in Python](#)
- [How should I organize a larger data science team?](#)

SHARES

0 Comments KDNuggets

1 Login ▾

Recommend Share

Sort by Best ▾



Start the discussion...

LOG IN WITH



OR SIGN UP WITH DISQUS (?)

Name

Be the first to comment.

[Subscribe](#) [Add Disqus to your site](#) [Add Disqus](#) [Disqus' Privacy Policy](#) [Privacy Policy](#) [Privacy Policy](#)[◀ Previous post](#)[Next post ▶](#)

## Top Stories Past 30 Days

### Most Popular

1. [Python eats away at R: Top Software for Analytics, Data Science, Machine Learning in 2018: Trends and Analysis](#)
2. [10 More Free Must-Read Books for Machine Learning and Data Science](#)
3. [A Beginners Guide to the Data Science Pipeline](#)
4. [Data Science vs Machine Learning vs Data Analytics vs Business Analytics](#)
5. [Why so many data scientists are leaving their jobs](#)
6. [The 6 components of Open-Source Data Science/ Machine Learning Ecosystem:](#)

### Most Shared

1. [Python eats away at R: Top Software for Analytics, Data Science, Machine Learning in 2018: Trends and Analysis](#)
2. [10 More Free Must-Read Books for Machine Learning and Data Science](#)
3. [Top 20 R Libraries for Data Science in 2018](#)
4. [Data Lake – the evolution of data processing](#)
5. [A Beginners Guide to the Data Science Pipeline](#)
6. [ETL vs ELT: Considering the Advancement of Data Warehouses](#)

SHARES

## 7. [DIY Deep Learning Projects](#)

## [Science/ Machine Learning Ecosystem; Did Python declare victory over R?](#)

### [Latest News](#)

- [Why the Data Lake Matters](#)
- [7 Simple Data Visualizations You Should Know in R](#)
- [Simple Tips for PostgreSQL Query Optimization](#)
- [UCD Dublin: Research Fellow](#)
- [What is it like to be a machine learning engineer in 2018?](#)
- [An Intuitive Introduction to Gradient Descent](#)

### More Recent Stories

- [An Intuitive Introduction to Gradient Descent](#)
- [Detecting Sarcasm with Deep Convolutional Neural Networks](#)
- [Deep Learning Best Practices – Weight Initialization](#)
- [Top tweets, Jun 6–19: #MachineLearning predicts #WorldCup...](#)
- [Technical Content Personalization](#)
- [The 5 Clustering Algorithms Data Scientists Need to Know](#)
- [KDNuggets 18:n24, Jun 20: Data Lakes – The evolution ...](#)
- [Get Packt Skill Up Developer Skills Report](#)
- [5 Key Takeaways from Strata London 2018](#)
- [Data Science Predicting The Future](#)
- [Choosing the Right Metric for Evaluating Machine Learning Mode...](#)
- [Natural Language Processing Nuggets: Getting Started with NLP](#)
- [Kent State University: Faculty Tenure-Track – Management...](#)
- [Institute for Defense Analyses \(IDA\): Research Analyst –...](#)
- [Drexel Online MS in Data Science](#)
- [ZS Associates: Data Science Associate Consultant](#)
- [Every time someone runs a correlation coefficient on two time ...](#)
- [Top Stories, Jun 11-17: Data Lake – the evolution of data pr...](#)
- [Step Forward Feature Selection: A Practical Example in Python](#)
- [CVS: Application Developer, Rebates Forecasting](#)

[KDNuggets Home](#) » [News](#) » [2018](#) » [Jun](#) » [Tutorials, Overviews](#) » Data Science Predicting The Future  
( [18:n24](#) )

© 2018 KDNuggets. [About KDNuggets](#). [Privacy policy](#). [Terms of Service](#)

Subscribe: Email  Your email  Name  your name

SHARES

X

---

SHARES