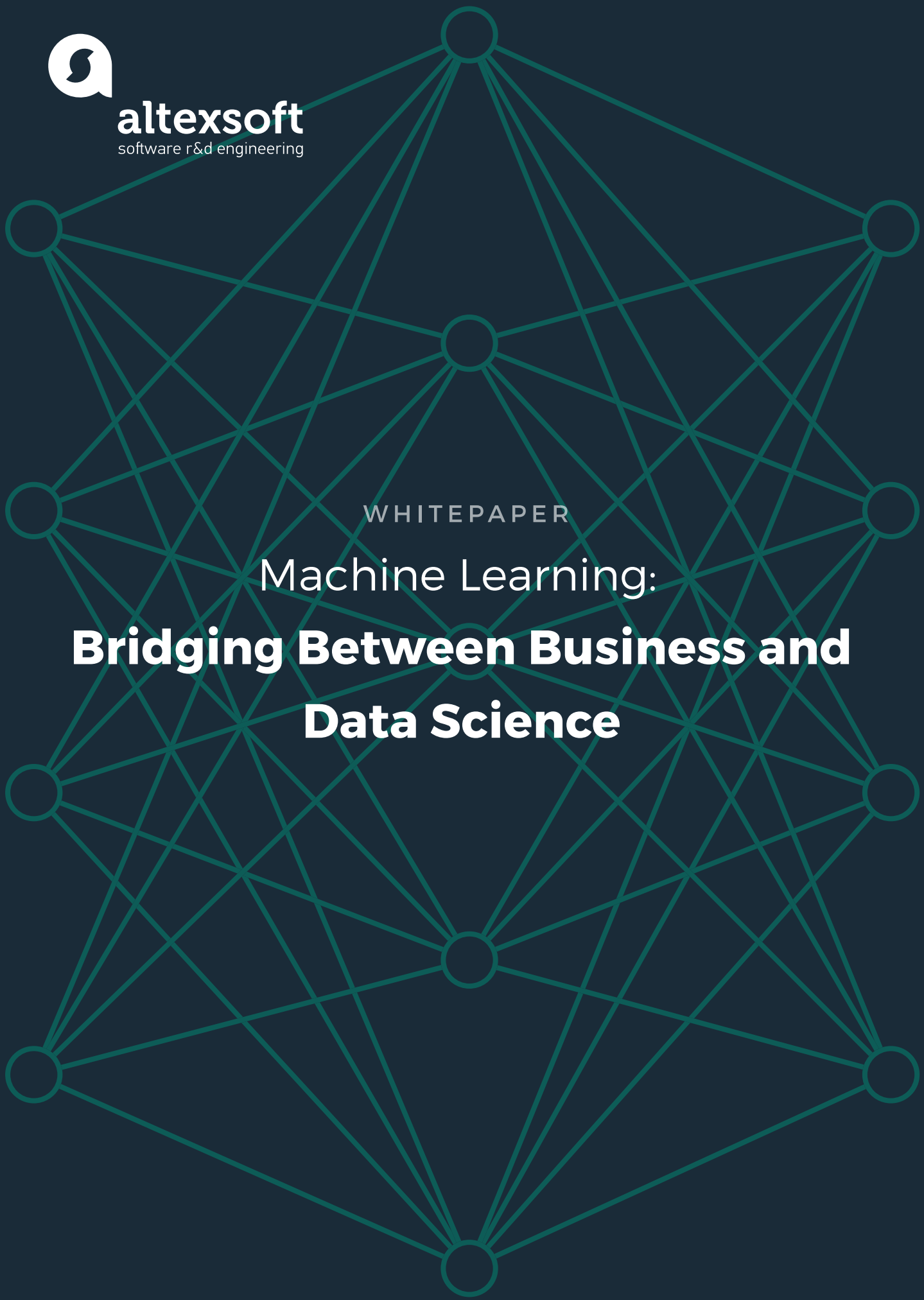


altexsoft
software r&d engineering

WHITEPAPER

Machine Learning:
Bridging Between Business and Data Science

Foreword

If the past few years hasn't found you living on a desert island without electricity or communication with the outside world, you've likely heard about machine learning (ML). It's hard to miss the trend. Every time we talk about self-driving cars, chatbots, AlphaGo, or predictive analytics, we're discussing some implementation of machine learning techniques. While success stories and evangelists abound, machine learning hasn't become the obligatory for business yet. In the public's perception, algorithms that are applied in ML are close to science fiction, and rolling out a concrete plan for ML adoption is still a high hurdle.

Hence, this whitepaper is aimed at answering practical questions instead of setting the vision and evangelizing the trend. This is about an umbrella term data science and how its subfields interact, the main problems that machine learning can solve, and how these problems can be translated into the language of business. We will also contemplate the main decisions to make concerning talent acquisition and pinpoint the challenges to be considered in advance. Because we've covered data science's potential in articles dedicated to the travel and industries, we will only touch on it briefly today.

1. Clarifying the terms in data science

The concept of machine learning was first introduced back in the 1950s by people who were the remarkable AI pioneers of that time. In 1950, Alan Turing published the "Computing Machinery and Intelligence" paper that suggested a famous AI-evaluation test that we know today as the Turing Test. In 1959, Arthur Lee Samuel coined the term "machine learning." Many theoretical discoveries that we use were made at that time. But why are we talking so much about machine learning and data science today?

Perhaps, the most important difference is the computational powers and the amount of data we can collect and analyze compared to previous decades. A smartphone that easily fits in the palm of the hand today can store and process more data than a mainframe computer of the '60s, which occupied several rooms. Instead of relying on thoroughly curated and small datasets, we can use large and unorganized data with thousands of parameters to train algorithms and draw predictions. The amount and quality of data are what also differentiates modern machine learning techniques from statistics. While statistics usually rely on a few variables to capture a pattern, machine learning can be effectively utilized with thousands of data characteristics.

In this section, we'll discuss several fields of data science and how they are connected with each other.

1.1 Data Science

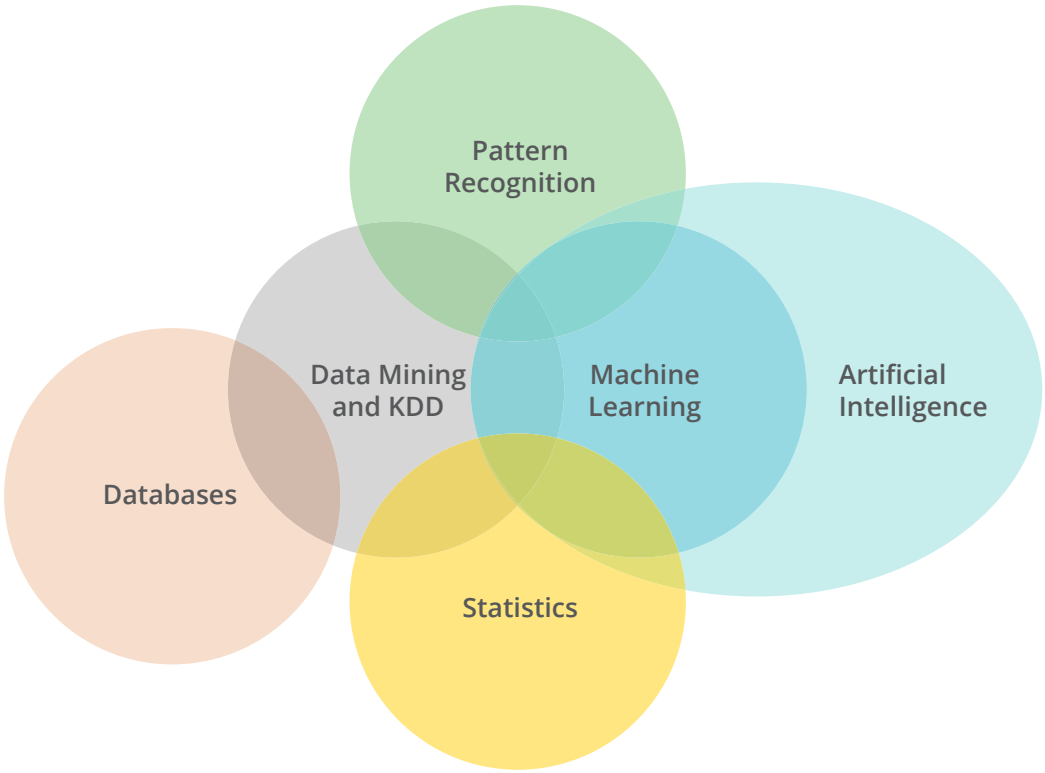
The term data science was conceived back in the 1960s. While there are many definitions of it, the one which is business-centric was articulated by John W. Foreman, the Chief Data Scientist for MailChimp:

> ***"Data science is the transformation of data using mathematics and statistics into valuable insights, decisions, and products"***

As data science evolves and gains new "instruments" over time, the core business goal remains focused on finding useful patterns and yielding valuable insights from data. Today, data science is employed across a broad range

of industries and aids in various analytical problems. For example, in marketing, exploring customer age, gender, location, and behavior allows for making highly targeted campaigns, evaluating how prone customers are to make a purchase or leave. In banking, finding outlying client actions aids in detecting fraud. In healthcare, analyzing patients' medical records can show the probability of having diseases, etc.

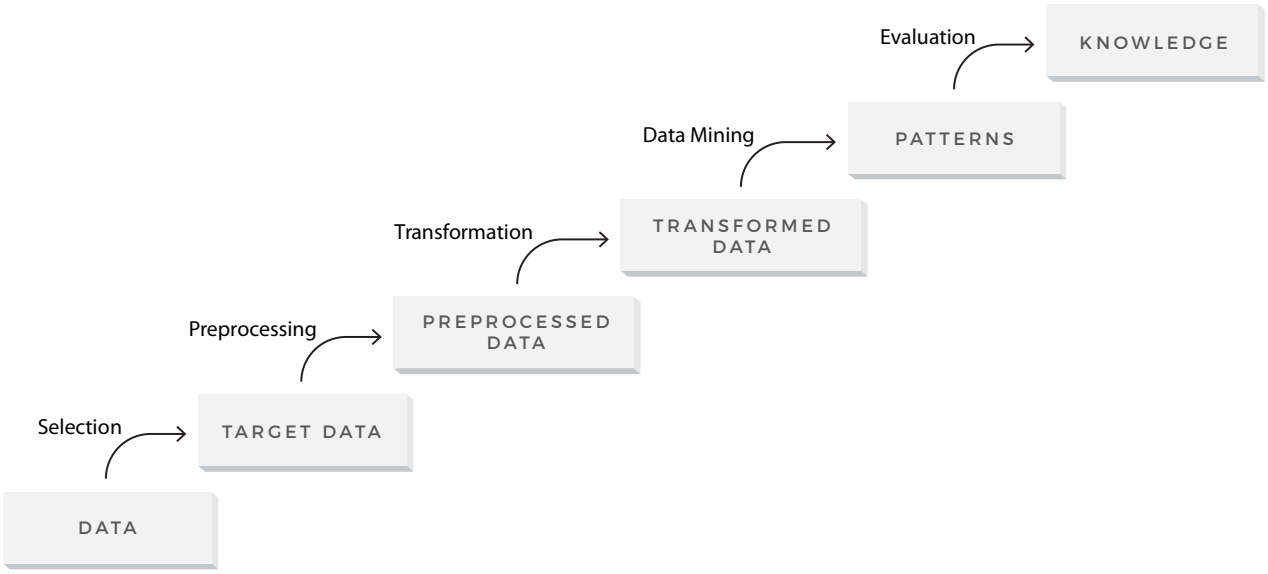
The data science landscape encompasses multiple interconnected fields that leverage different techniques and tools.



Data Science Disciplines

1.2 Data Mining and Knowledge Discovery in Databases

As you see from the diagram, all data science fields are connected with data mining as it constitutes the core set of practices within data science. The term data mining is a misnomer and doesn't portray what it stands for. Instead of mining data itself, the discipline is about creating algorithms to extract valuable insights from large and possibly unstructured data. The basic problem of data mining is to map available data and convert it into digestible patterns. Data mining is considered to be a part of a broader process called Knowledge Discovery in Databases (KDD) which was introduced in 1984 by Gregory Piatetsky-Shapiro.



Knowledge Discovery in Databases

While it seems that data mining and KDD solely address the main problem of data science, machine learning adds business efficiency to it.

1.3 Machine Learning

There's a difference between data mining and machine learning. Machine learning is about creating algorithms to extract valuable insights. It's heavily focused on continuous use in dynamically changing environments and emphasizes adjustments, retraining, and updating algorithms based on previous experiences. The goal of machine learning is to constantly adapt to new data and discover new patterns or rules in it. Sometimes it can be realized without human guidance and explicit reprogramming.

Machine learning is the most dynamically developing field of data science today due to a number of recent theoretical and technological breakthroughs. They led to natural language processing, image recognition, or even the generation of new images, music, and texts by machines. Machine learning remains the main instrument of building artificial intelligence.

1.4 Artificial Intelligence

Artificial intelligence (AI) is perhaps the least understood field of data science. It also stands distinctly apart from the rest. The main idea behind building AI is to use pattern recognition and machine learning to build an agent able to think and reason as humans do (or approach this ability). However, with this term so widely used, we haven't yet agreed on interpreting the I in AI. Intelligence is hard to delineate, and ways to define it are numerous. In business language, AI can be interpreted as the ability to solve new problems. Effectively, solving new problems is the outcome of perception, generalizing, reasoning, and judging.

In the public view, AI is usually conceived as the ability of machines to solve problems related to many fields of knowledge. This would make them somewhat similar to humans. However, the concept of artificial general intelligence (AGI) remains in the realm of science fiction and doesn't yet match existing state-of-the-art advancements. Such famous systems as AlphaGo, IBM Watson, or Libratus, which has recently beaten humans in Texas Hold'em, are representative of artificial narrow intelligence (ANI). They specialize in one area and can perform tasks based on similar techniques to process data. So, scaling from ANI to AGI is the bridge that data science is yet to cross, and this breakthrough isn't likely to happen for several decades. While the growing fear that machines may take over many jobs is not unreasonable, the scenario in which machines dominate the world is.

1.5 Big data

Big data is also an overly hyped and misunderstood concept. The growth of digital transformation in business allowed for gathering increasingly large datasets that contain various, usually unstructured, records about customers, employees, and corporate assets. These relate to demographics, interactions and behaviors, endpoint devices, and literally everything that can be tracked by digital means or input manually. However, these unstructured datasets aren't yet big data.

> ***"Collecting doesn't mean discovering."***
>
> Sean McClure, Ph.D.Director, Data Science at Space-Time Insight

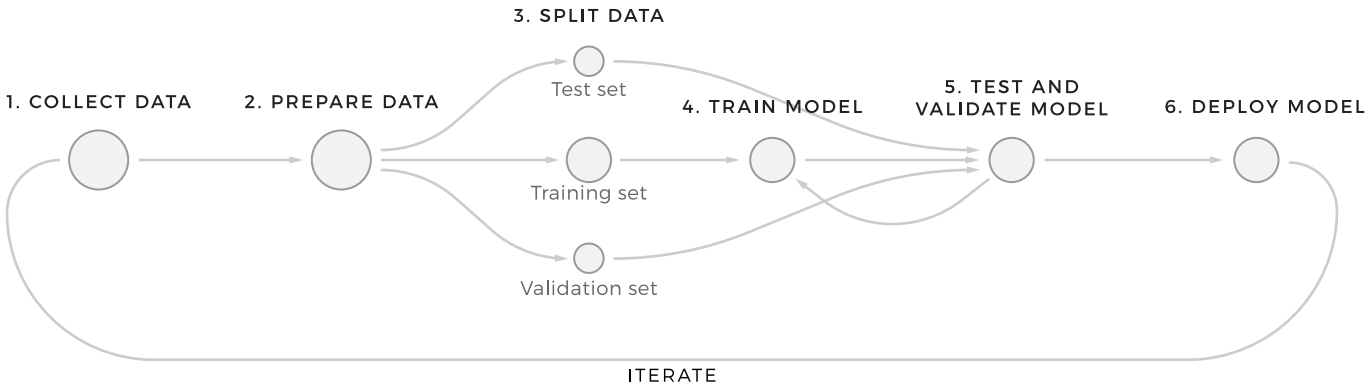
Collecting large quantities of data doesn't necessarily equate with the discovery of insightful patterns in it. The concept of big data implies discovering patterns in large datasets using the techniques of data mining and machine learning. Why is there so much emphasis on big data today? The popularity of big data among technology evangelists stems from the recent advancements in computational power. Instead of using limited subsets of data to discover and extrapolate the results to the entire subject field, we can process all raw data, achieve higher accuracy, and find more hidden dependencies. This requires building high-end infrastructure capable of computing increasingly large sets of unstructured data, then acquiring the tools and expertise to properly visualize the data and yield the insights contained in it.

2. Machine learning workflow by steps

So how do we make algorithms find useful patterns in data? The main difference between machine learning and conventionally programmed algorithms is the ability to process data without being explicitly programmed. This means that an engineer isn't required to provide elaborate instructions to a machine on how to treat each type of data record. Instead, a machine defines these rules itself relying on input data. Regardless of a unique machine learning application, the general workflow remains the same and iteratively repeats once the results become dated or need higher accuracy. This section is focused on introducing the basic concepts that constitute machine learning
workflow.

The core artifact of any machine learning execution is a mathematical **model**, which describes how an algorithm processes new data after being trained with a subset of historic data. The goal of **training** is to develop a model capable of formulating a **target value (attribute)**, some unknown value of each data object. While this sounds complicated, it really isn't.

For example, you need to predict whether customers of your eCommerce store will make a purchase or leave. These predictions buy or leave are the target attributes that we are looking for. To train a model in doing this type of prediction, you "feed" an algorithm with a **dataset** that stores different records of customer behaviors and the results (whether customers left or made a purchase). By learning from this historic data, a model will be able to make predictions on future data.



Machine Learning Workflow

Generally, the workflow follows these steps:

1. **Collect data**. Use your digital infrastructure and other sources to gather as many useful records as possible and unite them as a dataset.
2. **Prepare data**. Prepare the data to be processed in the best possible way. Data preprocessing and cleaning procedures can be quite sophisticated, but they usually aim at filling the missing values and correcting other flaws in the data, like different representations of the same values in a column (e.g. December 14, 2016 and 12.14.2016 won't be treated the same by the algorithm).
3. **Split data**. Separate subsets of data to train a model and further evaluate how it performs against new data.
4. **Train a model**. Use a subset of historic data to enable the algorithm recognize the patterns in it.
5. **Test and validate a model**. Evaluate the performance of a model using testing and validation subsets of historic data to understand how accurate the prediction is.
6. **Deploy a model**. Embed the tested model into your decision-making framework as a part of an analytics solution or let users leverage its capabilities (e.g. better target your product recommendations).
7. **Iterate**. Collect new data after using the model to incrementally improve it.

3. Five groups of task that machine learning solves

In business terms, machine learning addresses a broad spectrum of tasks, but on higher levels, the tasks that algorithms solve fall into five major groups: classification, cluster analysis, regression, ranking, and generation.

3.1 Classification

Classification algorithms define which category the objects from the dataset belong to. Thus, categories are usually referred to as **classes**. By solving classification problems, you can address a variety of questions:

Binary classification problems

- *Will this lead convert or not?*
- *Is this email spam or not?*
- *Is this transaction fraudulent or not?*

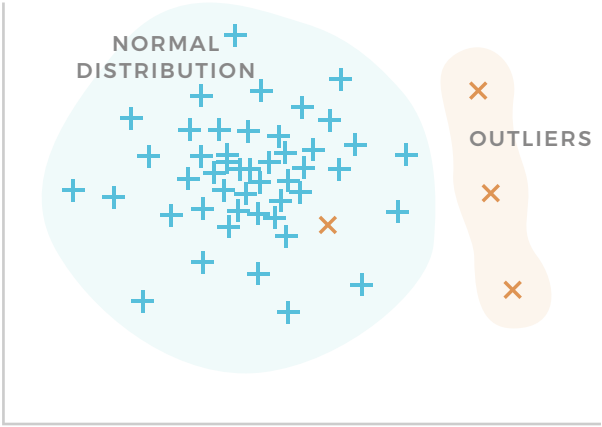
And, multiclass problems

- *Is this apartment in New York, San Francisco, or Boston?*
- *What is pictured: a cat, a dog, or a bird?*
- *Which type of product is this customer more likely to buy: a laptop, a desktop, or a smartphone?*

Binary classification

Another highly specific type of classification task is **anomaly detection**. It's usually recognized as the one-class classification because the goal of anomaly detection is to find **outliers**, unusual objects in data that don't appear in its normal distribution. It can solve these types of problems:

- *Are there any untypical customers in our dataset?*
- *Can we spot unusual behaviors among our bank clients?*
- *Does this patient deviate from the rest according to the records?*

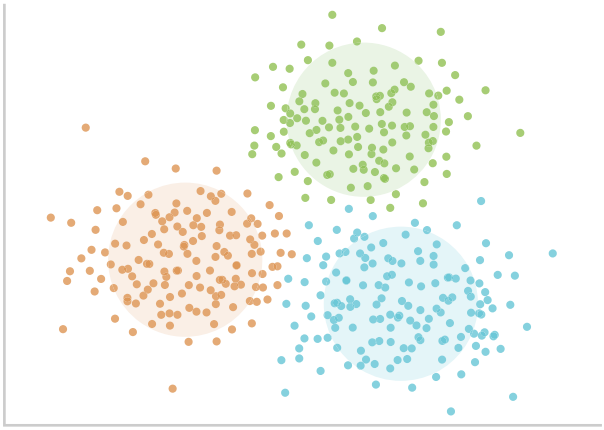


Anomaly detection

3.2 Cluster analysis

The main difference between regular classification and clustering is that the algorithm is challenged to group items in clusters without predefined classes. In other words, it should decide the principles of the division itself without human guidance. Cluster analysis is usually realized within the unsupervised learning style, which we will talk about in a minute. Clustering can solve the following problems:

- *What are the main segments of customers we have considering their demographics and behaviors?*
- *Is there any relationship between default risks of some bank clients and their behaviors?*
- *How can we classify the keywords that people use to reach our website?*

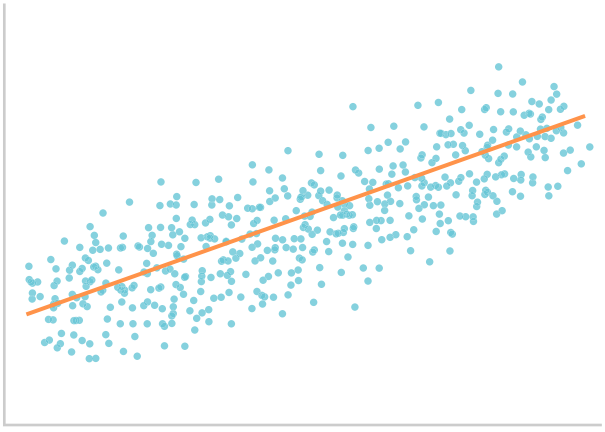


Cluster analysis (estimated number of clusters: 3)

3.3 Regression

Regression algorithms define numeric target values instead of classes. By estimating numeric variables, these algorithms are used in predicting product demand, sales figures, marketing returns, etc. For example:

- *How many items of this product will we be able to sell next month?*
- *What's will the airfare be for this destination?*
- *What's going to be the rental price for this house?*

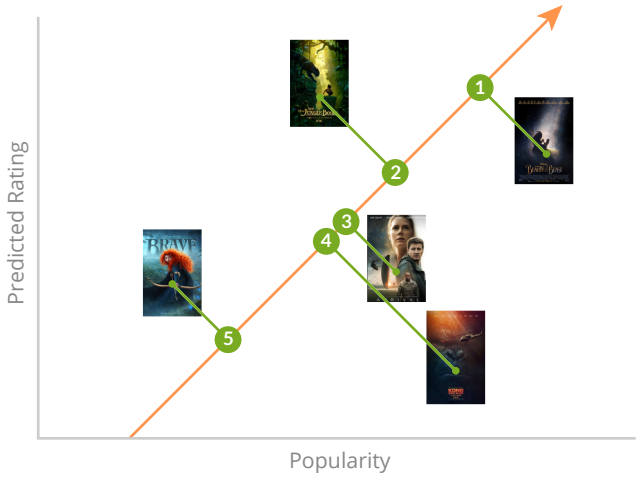


Linear regression

3.4 Ranking

Ranking algorithms decide the relative importance of objects (or items) as related to other objects. The most well-known example is PageRank, which is heavily used by Google to rank pages on the search engine results page. Ranking algorithms are also applied by Facebook to define which posts in a news feed are more engaging to users than others. What other problems can ranking address?

- *Which movies this user will enjoy the most?*
- *What hotels will be on the most-recommended list for this customer?*
- *How should we rank products on a search page of an eCommerce store?*



Movie recommendation ranking

3.5 Generation

Generation algorithms are applied to generate text, images, or music. Today they are used in such applications as Prisma that converts photos to artwork-style images, or WaveNet by DeepMind that can mimic human speech or create musical compositions. Generative tasks are more common for mass consumer applications, rather than predictive analytics solutions. That's why this type of machine learning has big potential for entertainment software. What type of tasks are in the realm of generative algorithms?

Image converted to artwork using "The Great Wave off Kanagawa" piece of art

- *Turn photos into specific style of painting.*
- *Create text-to-speech applications for mobile voice assistants (e.g. the Google assistant).*
- *Create music samples of one style or that are reminiscent of a particular musician.*

To meet these tasks, different model training approaches (or training styles) are used. Training is a procedure to develop a specific mathematical model that is tailored to dependencies among values in historic data. A trained model will be able to recognize these dependencies in future data and predict the values that you look for. So, there are three styles of model training.

4. Three Model Training Styles

Choosing training styles depends on whether you know the target values that should be found. In other words, you can have training datasets where the target values are already mapped and you just want the algorithm to predict these exact values in future data. Or your goal may be to figure out hidden connections among values. In the latter case, target values are unknown both for historic data and future data. This difference in goals impacts the training style choice and defines which algorithms you choose.

4.1 Supervised learning

Supervised learning algorithms operate with historic data that already has target values. Mapping these target values in training datasets is called **labeling**. In other words, humans tell the algorithm what values to look for and which decisions are right or wrong. By looking at a label as an example of a successful prediction, the algorithm learns to find these target values in future data. Today, supervised machine learning is actively used both with classification and regression problems as target values are usually available in training datasets.

This makes supervised learning the most popular approach utilized in business. For example, if you choose binary classification to predict the likelihood of lead conversion, you know which leads converted and which didn't. You can label the target values (converted/not converted or 0/1) and further train a model. Supervised learning algorithms are also used in recognizing objects on pictures, in defining the mood of social media posts, and predicting numeric values as temperature, prices, etc.

4.2 Unsupervised learning

Unsupervised learning is aimed at organizing data without labeled target values. The goal of machine learning, in this case, is to define patterns in values and structure the objects by similarities or differences. In classification tasks area, unsupervised learning is usually applied with clustering algorithms and anomaly detection. These models are useful in finding hidden relations among items, solving segmentation problems, etc.

For example, a bank can use unsupervised learning to split clients into multiple groups. This will help to develop specific instructions for dealing with each group. Unsupervised learning techniques are also employed in

ranking algorithms to provide individualized recommendations and in generative tasks.

4.3 Reinforcement learning

Reinforcement learning is perhaps the most sophisticated style of machine learning and is inspired by game theory and behaviorist psychology. An agent (an algorithm) must make decisions based on input data and then be "awarded" or "punished," depending on how successful these decisions were. By iteratively facing awards and punishments, the agent alters its decisions and gradually learns to achieve better results.

Reinforcement learning techniques today are actively used in robotics and AI development. A well-known AlphaGo algorithm by DeepMind used reinforcement learning to estimate the most productive moves in the ancient game of Go instead of enumerating all possible board combinations. Allegedly, reinforcement learning is employed by the Tesla autopilot along with supervised learning techniques. The style is utilized when the autopilot is on and a driver corrects its decisions.

However, in business computing, reinforcement learning is still hard to apply as most algorithms can successfully learn only within the unchanging framework of rules, goals, and world circumstances. That's why today's many modern reinforcement learning advancements are tethered to games like Go or old Atari titles where these three parameters are stable. Another problem of reinforcement learning is the longevity of learning cycles. In games, the time between the first decision and achieved points is relatively short, while in real-life circumstances the time to estimate how successful the decision was may take weeks.

5. Embarking on machine learning

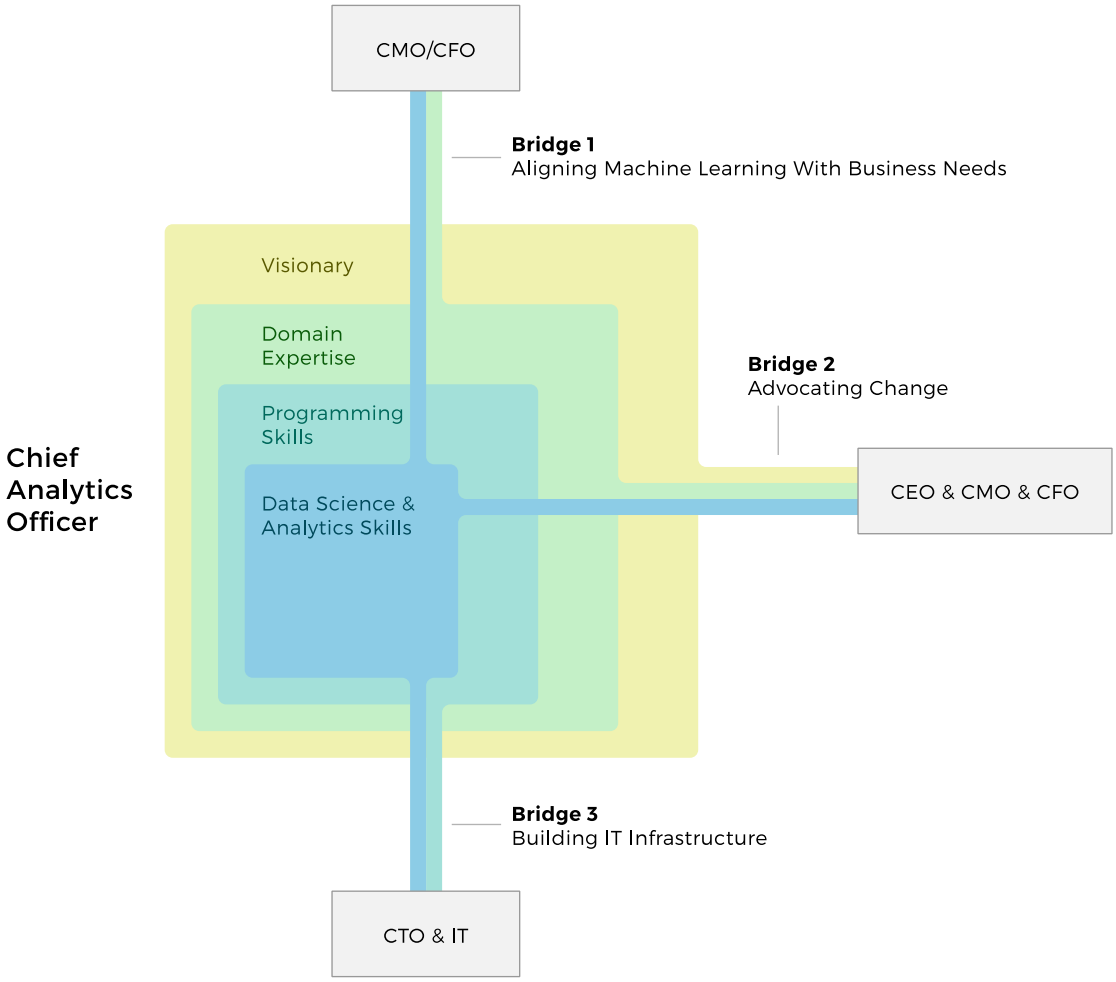
Predictive analytics and machine learning are still terra incognita for most businesses. Although the evolution of machine learning tools seems impressive, capturing the business value is still challenging. Companies stumble over talent acquisition barriers, internal leadership difficulties, and, last but not least, the rigidity of overregulated corporate culture. It's relatively easy to theorize about the great potential of big data–which looms large in the media–but the reality is that the number of companies planning to invest in big data sank from 31 to 25 percent in 2016. On the other hand, the investment in big data is generally up thanks to big players. This means that the competitive gap only increased for smaller or less flexible businesses.

In this section, we'll talk about the most critical decisions that should be made on executive level to overcome these barriers and align with competition.

5.1 "Business translator" and visionary

Proper analytics and data science leadership is the greatest barrier to achieving data-driven culture. According to the McKinsey Global Institute survey, 45 percent of companies are struggling to set the right vision and strategy for data and machine learning. Consider this–the challenges of talent acquisition are well-known: Data science talent is scarce and expensive both in terms of compensation and retention. While finding a data scientist is hard, finding an analytics leader is even more difficult, according to the survey. Ironically, this role is critical to manage efficient data processes. Although you can introduce some machine learning implementations and compensate for a few missing links without an analytics leader, this approach is destined to remain responsive rather than proactive.

The skillset of this "business translator," or chief analytics officer (CAO), is a multidisciplinary bridge between business values and data science capabilities. The person should take the lead and reconcile the efforts of the information technology department, data science, marketing, finance, and stakeholders to build and develop a data strategy.



Chief Analytics Officer Engagement Field

Another important mission of an analytics leader is a visionary one. It implies foreseeing business application potential in new data science research works before they are widely adopted. Most of the machine learning techniques that have met business demands lately have been known in data science for decades. And many–like reinforcement learning–have yet to find their implementations beyond prominent labs like DeepMind. By capturing these advancements early and finding ways to convert them into business use a business translator can keep the organization ahead of the competition.

However, analytics specialist acquisition won't be simple. The current mismatch between the demand for senior analytics positions and talent supply stands at 5:1. And if recruitment fails, this division implies finding and training an analytics expert internally. The best-fit opportunity, in this case, is to engage a person who has both technical and domain business background. Sometimes, this role can be obtained by a chief technology officer, a data scientist who transitions into management, or even a chief executive officer. That decision depends on the organization size.

5.2 Data-driven organization

A data scientist alone can only be effective within a fertile corporate environment. Introducing a machine learning initiative should be supported and understood on all organizational levels. With each new technology coming, not only training is required, but also immense effort in evangelizing change. If you plan to use machine learning as a support to decision-making or as a lever to make important decisions, most likely this way of thinking is going to face reasonable resistance. People are used to making decisions based on their intuition and experience, which made them professionals prior to predictive analytics bursting on the scene. The role of an analytics leader (or CAO) and other C-level executives is to educate employees and foster the innovation. This is the reason why communication and presentation skills are preferred qualities for a data scientist.

Siloed data. The siloed structure of departments is another barrier to building a data-driven organization. Access to data can be either overregulated or warily guarded by departments that may want to keep the data they collect to themselves. By combating this behavior you can achieve much better results in acquiring more useful data.

Anonymized data. Sometimes regulations are imposed legally in such businesses as banking or insurance and data can't be easily shared. In this case, all values in data can be turned into anonymized numbers at the data preparation stage. Thus sensitive business or customer details won't be revealed.

Conclusion

This paper isn't intended to be exhaustive and shouldn't be considered as a playbook for your emerging machine learning initiative. While there is much to explore, we rather suggest using this white paper as a guide to evaluate your strategy.

The bottom line problem for business today is to understand how and when this strategy is going to be realized to keep up with the pace of change that machine learning and predictive analytics can provide. The modern era of business decisions will put ahead of the competition those who can make the best use of data they collect.

Further Reading

1. Data Smart: Using Data Science to Transform Information into Insight, Per John W. Foreman - https://books.google.cat/books?id=CfjpAQAAQBAJ&pg=PR14#v=onepage&q&f=false
2. From Data Mining to Knowledge Discovery in Databases, Fayyad, Piatetsky-Shapiro & Smyth, 1996 - http://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131
3. A Collection of Definitions of Intelligence, Shane Legg, Marcus Hutter, 2007 - https://arxiv.org/pdf/0706.3639.pdf
4. https://www.linkedin.com/pulse/data-science-big-two-very-different-beasts-sean-mcclure-ph-d-?trk=mp-reader-card
5. http://www.gartner.com/newsroom/id/3466117
6. The Age of Analytics: Competing in a Data-Driven World, 2016 - http://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world
7. http://www.forbes.com/sites/adigaskell/2016/10/21/do-organizations-need-a-chief-analytics-officer/#785f6edf5f71
8. http://www.kdnuggets.com/2016/05/10-must-have-skills-data-scientist.html
9. https://www.oreilly.com/ideas/2015-data-science-salary-survey

About AltexSoft

AltexSoft is a Technology & Solution Consulting company co-building technology products to help companies accelerate growth. The AltexSoft team achieves this by leveraging their technical, process and domain expertise and access to the best price-for-value Eastern European engineers. Over 100 US-based and 200 worldwide businesses have chosen the company as their Technology Consulting Partner.

US Sales HQ

701 Palomar Airport Road,

Suit 300, Carlsbad, CA 92011

+1 (877) 777-9097

Global HQ

32 Pushkinskaya Str.,

Kharkiv, Ukraine 61057

+38 (057) 714-1537