MAKE BIG DATA WORK WITH MACHINE LEARNING

By Serge Haziyev and Iurii Milovanov

softserve

The Current State of Big Data

There is no doubt that many advanced technology companies have secured compelling competitive advantages thanks to big data. For the last few years its adoption has moved significantly from visionaries to pragmatists, so today it's difficult to find any enterprise organization that isn't investing in storing, processing, and extracting value from their mountain of growing data.

However, once companies have become capable of collecting and processing massive data, the question then becomes: "How can I get the maximum value out of it?"

Every company is at a different point in its big data journey, so maximizing its business value means moving big data adoption to the next stage. Although there is not a single standard or metric to measure an organization's current state, the concept of "maturity models" is a simplified but efficient way to assess something's potential relative to other organizations.

Big Data Analytics Maturity Model

The graphic below illustrates the Big Data Analytics Maturity Model, which characterizes the path of an organization from its beginning stage, “Data,” to “Wisdom,” which serves as both a pinnacle and a limitless destination in terms of its business possibilities.

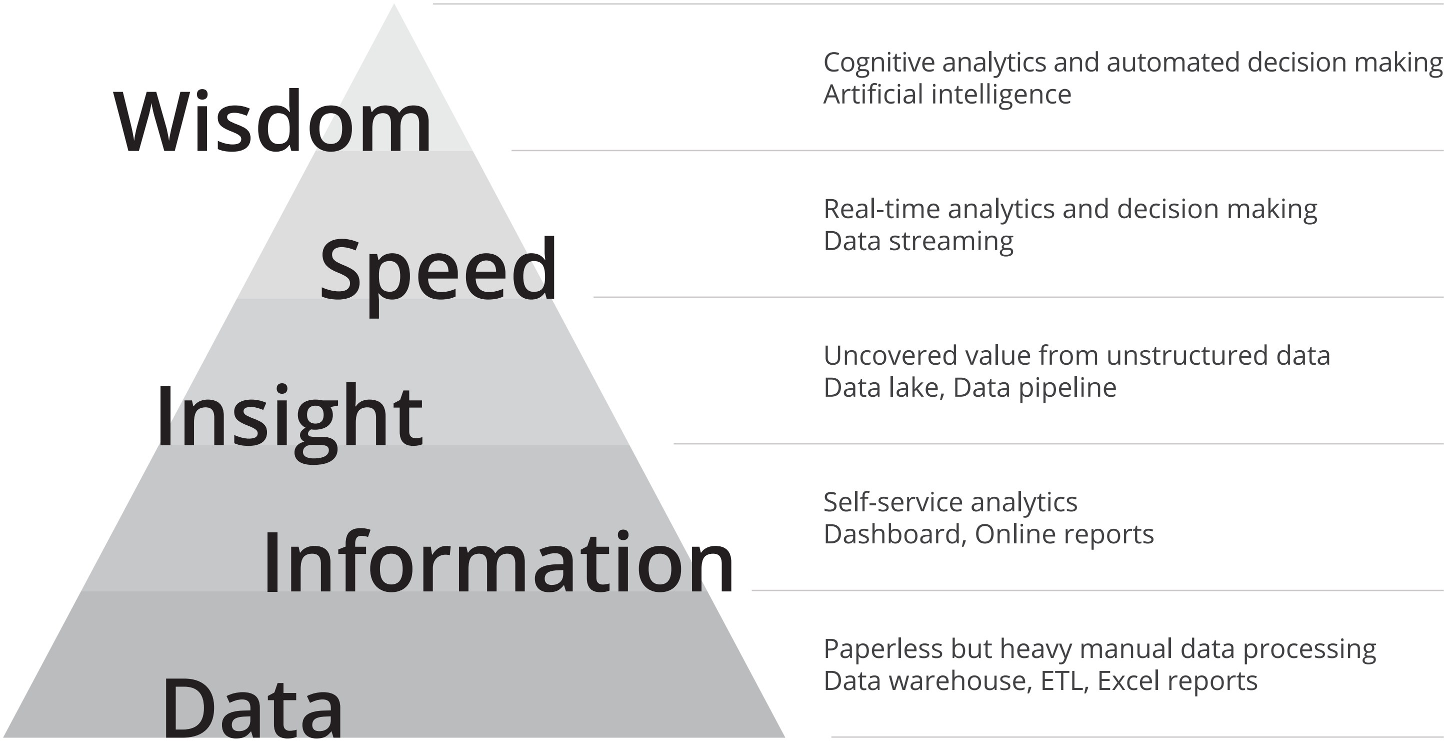
Wisdom
Cognitive analytics and automated decision making
Artificial intelligence
Speed
Real-time analytics and decision making
Data streaming
Insight
Uncovered value from unstructured data
Data lake, Data pipeline
Information
Self-service analytics
Dashboard, Online reports
Data
Paperless but heavy manual data processing
Data warehouse, ETL, Excel reports

Figure: Big Data Analytics Maturity Model

The pyramid itself is an adaptation of the DIKW (data–information–knowledge–wisdom) hierarchy, which is one of the fundamental and widely recognized concepts regarding the data value chain. We use this model since it is easy to understand and its graphical illustration speaks for itself.

Now, let’s connect the pyramid levels to the business needs of a modern enterprise:

Data

What does it mean if an organization is at the Data level? Usually it means the organization has paperless but heavy manual processes and a wide variety of data, from documents and transactional records to media files, machine generated logs, and sensor data. Data might have been collecting for decades from different sources and stored in a data warehouse or a raw format, but the analysis of that data is underdeveloped.

In this case, the organization’s teams dedicate a great deal of resources to extracting factual information in a non-automated fashion, typically to build reports in Microsoft Excel or Adobe PDF files.

It’s obvious but worth stating: being on this level means an organization’s data processes – or lack thereof – are holding it back from unlocking its potential, both in operational efficiency and in its ability to innovate new products or services.

Information

When data is being collected, it should be processed to be useful and answer who, what, where and when questions. This is a descriptive state of information, which is often characterized in the form of online reports, dashboards, and self-service capabilities that retrieve information of interest without requesting the support of IT or other departments.

From a technology perspective, typically this means an organization has a “business intelligence (BI) platform” that provides access to corporate data. Although this is a significant step in maturity from the data stage, and many organizations can be found at this level, a set of questions still remain unanswered for business users – how, why, and what if?

Insight

This level is more complex to explain than the data and information levels of the maturity model, as it moves closer to a manifestation of the human mind. For the sake of conciseness, we will define it as taking advantage of actionable information that conveys understanding, accumulated learning, and expertise. In other words, this level increases an individual’s capacity to take effective action. Technologically speaking, it includes the synthesis of multiple sources of information and the ability to uncover hidden patterns and correlations.

Consider the example of a doctor who makes the right diagnosis based on comprehensive patient information, including test results, patient interviews, and the patient’s medical history. He can then select the most effective treatment plan supported by his or her knowledge of industry studies and professional experience.

There are several elements in an enterprise architecture landscape that enable actionable decision-making at this level, including a data lake that collects a wide variety of data from structured, semi-structured, and unstructured data sources and a data processing pipeline that transforms that raw data into ready-to-consume holistic representations.

Organizations that have achieved this level of data maturity typically boast significant improvements to operational efficiency and/or increased revenue from products or services.

What could be better? An acceleration of decision-making, which is the next level.

Speed

This level refers to an organization's speed of decision-making. Supported by a solid foundation of insight (or actionable information), this level generates numerous possibilities by making real-time or near-realtime access to information a competitive advantage.
Data streaming, mobile computing, and IoT (Internet of Things) technologies make these advantages possible from a technology perspective.

Greater speed can lead to increased productivity and additional revenue sources from new products or services. It can also support the optimization of processes that can dynamically adopt themselves to environments that are constantly changing, due to factors like market conditions or even the weather.

Today, there are not many companies that can claim a solid footing at this stage or that they are reaping all of the benefits from this level, though many industry visionaries are investing heavily to get there. In fact, many might still consider the "speed" level the ultimate goal, not realizing what is possible when a company moves beyond speed to what we call "wisdom."

Wisdom

Some may argue that this is a rather elusive concept and has more to do with human intuition, understanding, interpretation, and actions than it does with enterprise systems or technology.

In our definition, wisdom is the highest level of the data value chain; the state where digital solutions can achieve par-human or even superhuman performance when solving difficult tasks intractable for traditional computing techniques. These tasks often imply complex situations that are characterized by ambiguity, uncertainty, and conflicting evidence.

This is exactly the goal of **cognitive computing**, which in its turn is based on **machine learning** techniques that mimic the way the human brain works. One of the latest techniques, deep learning, accomplished more in just three years than what has been done within the preceding 25 years in the field of AI (artificial intelligence).

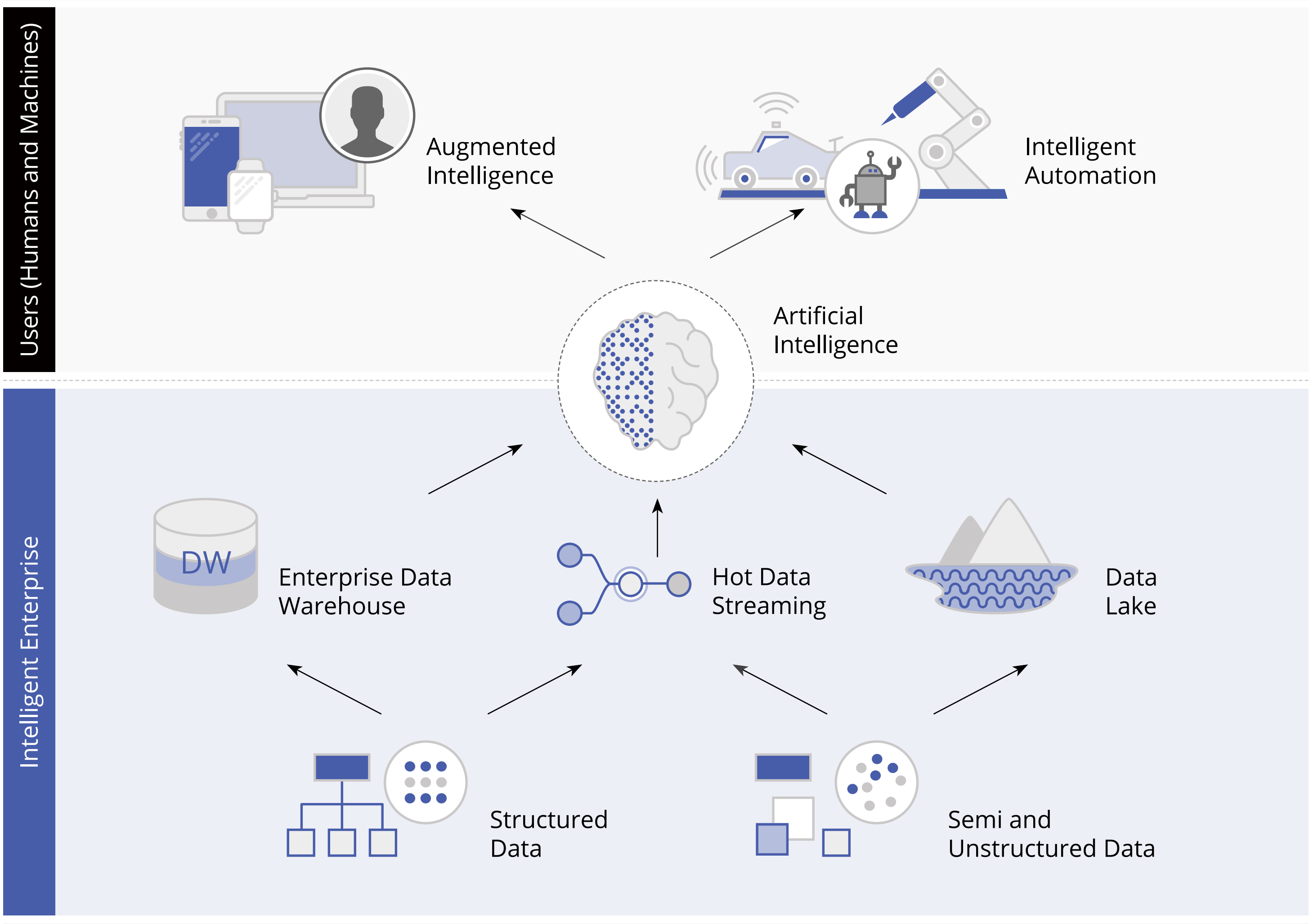

Figure: Simplified example of an enterprise big data architecture at the Wisdom level

It is worth mentioning that not all company departments can be at the same level of big data adoption maturity. In fact, it’s rather an exception across an industry. But what is most common among departments is that they can elevate their maturity level and get more value from data with machine learning in spite of a corporate function.

Machine Learning Business Use-Cases

What is most important for any business is staying competitive and not falling victim to disruption or displacement. That being said, it is crucial to look at big data, and machine learning in particular, not as a technology challenge to be solved by an organization's IT department, but as a business leadership opportunity. Meaning that the business strategy should incorporate machine learning to get maximum value out of data and to use it as a powerful vehicle of digital transformation.

Many organizations have difficulties at this stage selecting business use-cases that will bring the most business value and start technology implementations. Indeed, the variety of possibilities and potential use-cases of machine learning is overwhelming; there is no such distinctness in its adoption as in the adoption of mobile or cloud technologies. That's why we often hear questions: "What can big data and machine learning do for our business? What are examples of successful business use-cases in our industry?"

Although every business is unique in many respects, most business usecases can be derived from the universal monetary drivers, such as business growth and business operations:

Business growth

- New revenue creation (G1)
- Existing revenue increase (G2)

Business operations

- Productivity improvements, cost optimization (O1)
- Risk reduction, minimization of loses (O2)

The table below depicts examples of 20 use-cases for retail/digital commerce, financial services, and healthcare/life sciences mapped to the growth optimization drivers.

Driver	Business Use-case	Input	Output	Method
	Retail, Digital Commerce			
G2	**Product recommendations**	Customer profile and purchase history, specific product	Likelihood that the customer will buy a given product	Collaborative filtering, cluster analysis, deep learning
G2, O2	**Demand prediction**	Market situation, sales history, specific product category and time range	Forecast on the number of products in a given category to be sold in a given time range	Time-series analysis, cluster analysis, deep learning, decision trees
G2, O1	**Sales trending**	Multi-dimensional historical sales data	List of significant long- and short-term sales trends in business line dimensions and their combinations	Time-series and statistical analysis
G1	**Voice-enabled digital commerce**	Audio stream, customer profile, sales history	Recognized commands matched to products to sell, recommendations for cross-sale	Deep learning, natural language processing
G2, O1	**Merchandise display optimization**	Raw images of product shelves, sales history	Optimal product placement patterns for a given type of products	Deep learning, time-series analytics, cluster analysis
	Financial Services			
O2	**Fraud detection**	Customer financial and social profiles, specific transaction	Likelihood that a given transaction is fraudulent	Statistical analysis, deep learning, natural language processing
O2	**Risk management**	Customer personal and social profiles, transaction history	Size of the credit line that minimizes the risks or a credit score	Deep learning, time-series analysis, natural language processing, graph analytics
O1	**Claims processing automation**	Scanned images, PDF files, emails, faxes	Structured claim information, ready for inserting into a database or sending to an external system	Fuzzy matching, graph matching, heuristic-based search, decision trees
G1, G2	**Personalized car insurance**	Vehicle model and telemetry data, customer profile	Individualized optimal car-insurance plan	Digital signal processing, cluster analysis, deep learning, decision trees
G1, G2, O1	**Asset management**	News events, stock data, investment portfolio	Recommendations for investing or restructuring portfolio	Deep learning, decision trees, probabilistic modelling

Driver	Business Use-case	Input	Output	Method
	Healthcare, Life Sciences			
O1	**Medical diagnosis, radiography**	X-Ray image	Localization of broken or fractured bones	Probabilistic modeling, deep learning
G2, O2	**Public acceptance of drugs**	Social networks, online forums	Usage patterns, popularity, side-effects, comparison with competitors	Natural language processing
G1, O1	**Drug trials and testing**	Molecular and clinical data	New potential candidate drug formulas	Deep learning, genetic algorithms, probabilistic modelling
G1, G2	**Personal healthcare assistant**	Care plan, patient medical and social profiles, mobile data	Optimal healthcare schedule and motivation strategy	Deep learning, natural language processing, GEO analytics, behavioral modelling
G1	**Diabetes monitoring**	Patient medical and social profiles, blood glucose meter data, mobile data	Real-time healthcare statistics, abnormal pattern detection	Deep learning, statistical and time-series analysis
	Cross-domain			
O2	**Customer churn reduction**	Customer profile, customer transactions and engagement	Likelihood that a customer may decide to discontinue an ongoing contract	Deep learning, probabilistic modelling, time-series analysis
O1	**Supply-chain optimization**	Supply chain topology and historical data, customer sentiment data, geo-location data	Bottlenecks and low-performance supply chain segments, key performance drivers	Natural language processing, graph analysis, statistical and probabilistic modeling
G2	**Price optimization**	Product features, sales history, price constraints, market features	Optimal price range for a given product category and market segment	Time-series analysis, statistical modeling, cluster analysis, deep learning
O2	**Predictive maintenance**	Structured and un-structured monitoring measures (sensors, raw images and audio)	Likelihood that the given part or detail will fail in near future	Probabilistic modeling, deep learning
G2, O1	**Customer support automation**	Customer profile and engagement history, call center and service desk history	Machine-generated support feedback or optimal ticket assignment	Deep learning, natural language processing

While these industry-proven use-cases can be replicated in most organizations, there are much more specific use-cases which can become organizations' unique differentiatiors. In the next section, we will illuminate how to select significant business use-cases and achieve results in a predictable manner, in spite of the associated technological challenges.

Machine Learning Projects Best Practices

Over almost 30 years, machine learning has been developing techniques and algorithms, gradually increasing its accuracy. For many years, performance of the majority of AI applications remained sub-human, i.e. worse than average human performance. But since 2012 the situation has started to change, and the changes are quite drastic in terms of technology breakthrough and market demand for skilled professionals in this discipline.

Typical questions that business leaders are seeking answers to are: How can we pick proper business use-cases? How do we initiate the project? Are there any methods and processes that can help us achieve results in a predictive manner? We will address these questions in this paper by describing the three key best practices – Ideation Workshop, Multidisciplinary Team, and Rapid Prototyping.

Ideation Workshop

This is the first important exercise to prioritize business use-cases through ideation of potential business opportunities. The ideation workshop includes both business and technology stakeholders along with big data and machine learning experts to generate the use-cases and to estimate their preliminary business value and ease of implementation. The prioritization method shown below is based on the best combination of these two factors and serves as a roadmap for further execution.

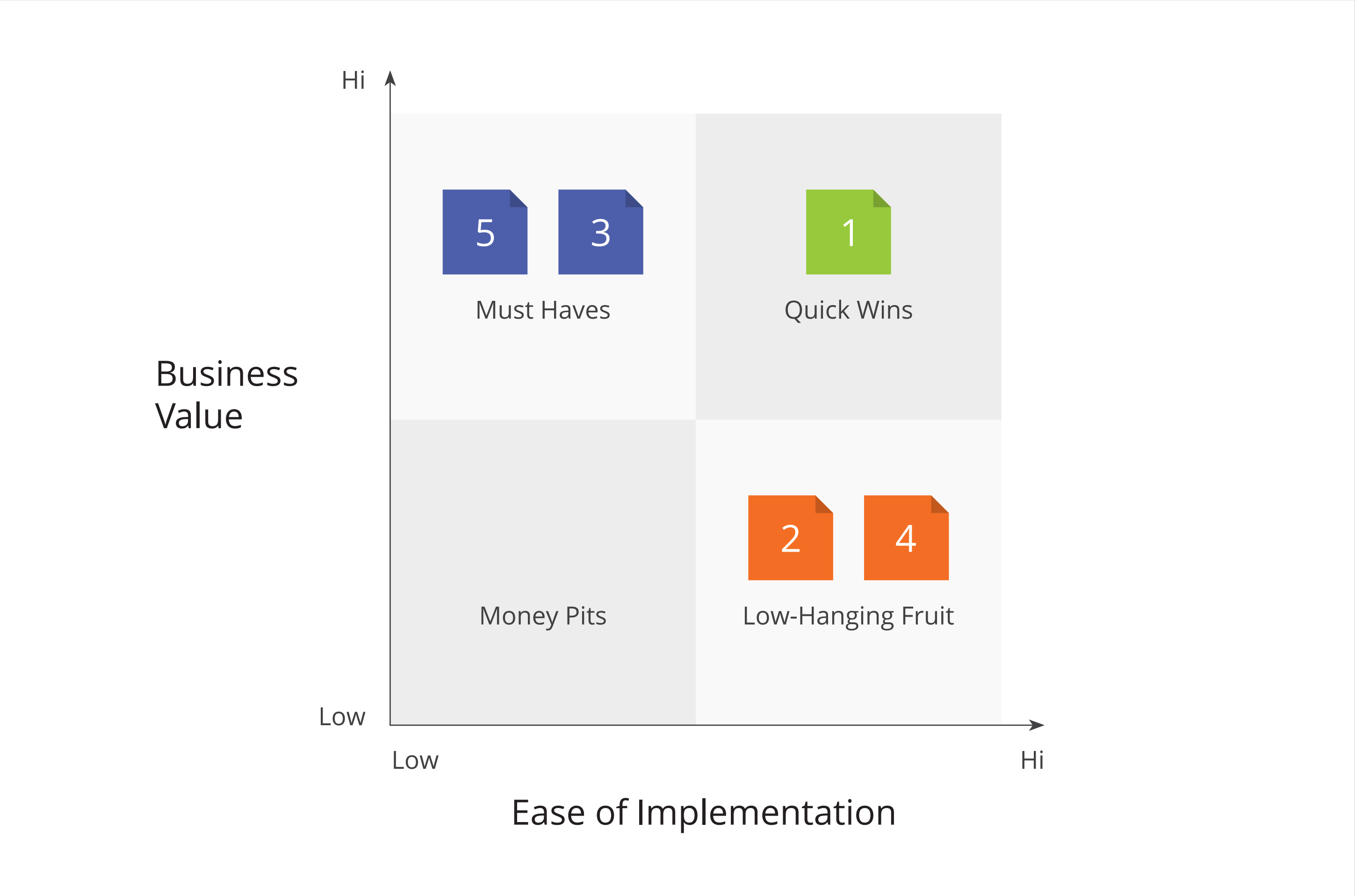

Figure: Sample of prioritized business use-cases

Multidisciplinary Team

When business use-cases are defined, the project scope and complexity often require the involvement of a team of professionals with different skillset. It would be a fallacy to wait for the super-hero data scientist to come along. Modern business and technology challenges, in order to be solved successfully, require tight cooperation of multiple disciplines.

Data science, big data engineering, experience design, and subject matter expertise are competencies that form a project core team. Design thinking, and in particular the design sprint process, help set up an interdisciplinary collaboration and align the team towards a common target.

Rapid Prototyping

Almost every machine learning project is more research-driven than most software engineering projects developed with traditional computation techniques. Business use-cases are based on hypotheses rather than facts, so they must be validated through experiments and tests to make them real. In Data Science, such experiments are performed with help of throwaway prototypes, also known as proofs of concept (PoC) using highly iterative and agile cycles.

For this purpose, data scientists use a set of tools, which noticeably reduce a PoC cycle to a few weeks or even days. Once the hypothesis is tested and the model's accuracy is satisfactory, the research phase can be considered as done and use-case validation complete.

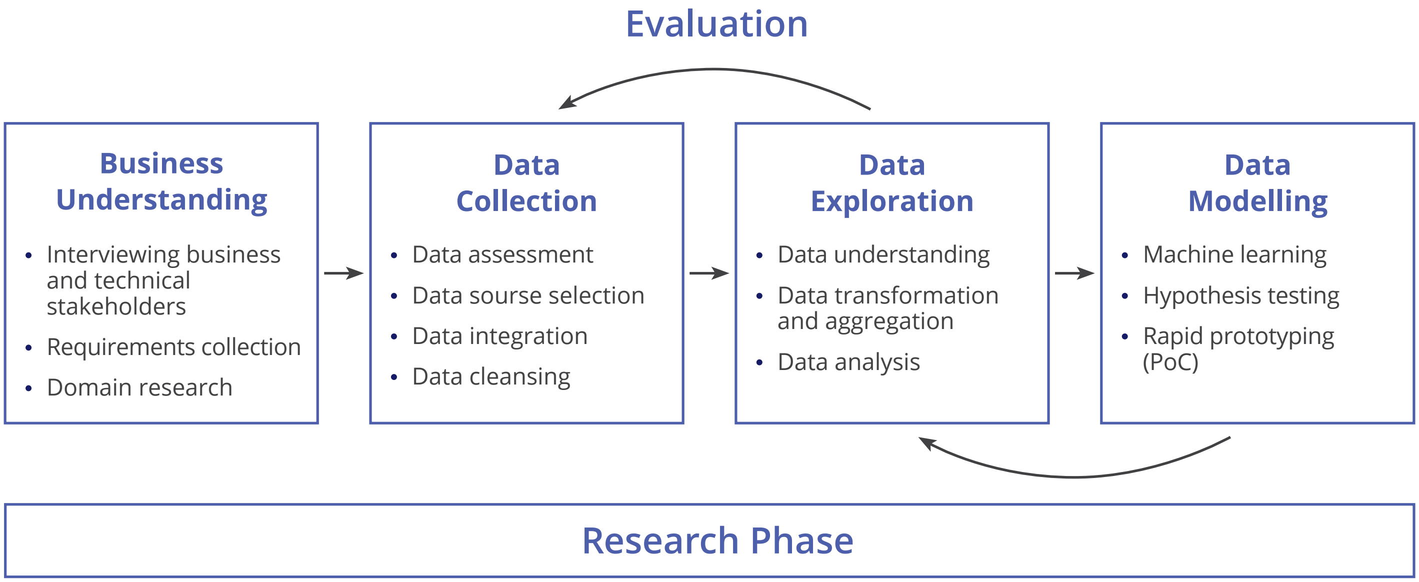

Figure: Machine learning project research phase process

In the meantime, the big data engineering team works on connecting data sources, a data processing pipeline, and other possible ingredients of a big data solution to be ready for deployment of the machine learning model in production. The **Strategic Prototyping method** can be helpful in bringing predictability into big data/machine learning projects and in avoiding cost, schedule and quality risks.

Summary

In this paper we provided answers to a number of frequent and important questions for any business that seeks to incorporate big data and machine learning into a business digital transformation strategy, including:

- How to get the maximum value out of data
- What big data and machine learning can do for our business
- How to pick the most promising business use-cases
- How to start your project

Once business leadership is ready for creative and breakthrough projects, anything is possible; even the most radical ideas get implementation with latest bleeding edge technologies. Perhaps this is a beauty of the digital era we live in: to make the wonders of science fiction into the realities of our lives.

References

- Geoffrey Moore. *"Why Crossing The Chasm is Still Relevant"*, Forbes, 2013
- Jennifer Rowley. *"The wisdom hierarchy: representations of the DIKW hierarchy"*. Bangor Business School, University of Wales, Bangor, UK. May 2006
- Serge Haziyev, Yuriy Milovanov. *"Cognitive Computing: How to Transform Digital Systems to the Next Level of Intelligence"*. Dataconomy.com. Jan 2017
- Alexander Linden, Tom Austin, Svetlana Sicular. *"Innovation Insight for Deep Learning"*. Gartner. Jan 2017
- Bill Schmarzo. *"Organizational Analytics Adoption: A Generation Away?"*. infocus.emc.com, Feb 2017
- Hong-Mei Chen, Rick Kazman, Serge Haziyev. *"Strategic Prototyping for Developing Big Data Systems"*. IEEE Software. Mar-Apr 2016

ABOUT US

SoftServe is a global digital authority and consulting company, operating at the cutting edge of technology. We reveal, transform, accelerate, and optimise the way large enterprises and software companies do business. With expertise across healthcare, retail, media, financial services, software, and more, we implement end-to-end solutions to deliver the innovation, quality, and speed that our clients' users expect.

SoftServe delivers open innovation – from generating compelling new ideas, to developing and implementing transformational products and services. Our work and client experience is built on a foundation of empathetic, human-focused experience design that ensures continuity from concept to release.

Ultimately, we empower businesses to re-identify their differentiation, accelerate market position, and vigorously compete in today's digital, global economy.

Visit our **website**, **blog**, **Facebook**, **Twitter**, and **LinkedIn** pages.

USA HQ

201 W 5TH STREET, SUITE 1550
AUSTIN, TX 75703
+1 866 687 3588

EUROPEAN HQ

One Canada Square
Canary Wharf
London E14 5AB
+44 (0)800 302 9436

info@softserveinc.com
www.softserveinc.com

softserve