altamira

Where AI is today and where it is headed tomorrow

Experts define best practices and pitfalls

Altamira
www.altamira.ai
Phone number: **+421 948 656 863**
Email: **hello@altamira.ai**

Intro

No other topic captures the imagination like Artifical Intelligence.

But what exactly is it and where are the opportunities for us as technology leaders? Some see new opportunities to optimize and improve existing IT systems and processes, but we also see many startups implementing and integrating AI/ML technologies right at the very heart of their product ideas period.

Tech juggernauts such as IBM, Google, Apple, Microsoft, and others are deploying and investing substantial resources into developing new software and hardware solutions, thus accelerating the value and impact of technology implementation.

This focus leads to solutions such as the GPT-4 algorithm model developed by OpenAI. Today, we are witnessing how the computing power of modern NLP models is becoming comparable with human intellect. There are now more reusable solutions in the marketplace than ever before, leading to questions about whether tech leaders should invest in re-using existing capabilities or focus on their unique data and insight differentiators.

Earlier, the Covid-19 pandemic highlighted the importance of alternative data and the ability to process high-frequency data in an ultra-dynamic marketplace, highlighting the need for technical skills to adapt to changing customer needs.

In addition, faster CPUs (Central Processing Units) are coming out every year, surpassing the performance of their predecessors. One application for chips is using them in specialised hardware solutions for neural networks, which can prove even more efficient than the usual TPUs (Tensor Processing Units), which deliver higher-quality results with fewer data points.

A great deal is happening in the field right now, and technology leaders must interpret these developments to effectively support business strategy and direction.

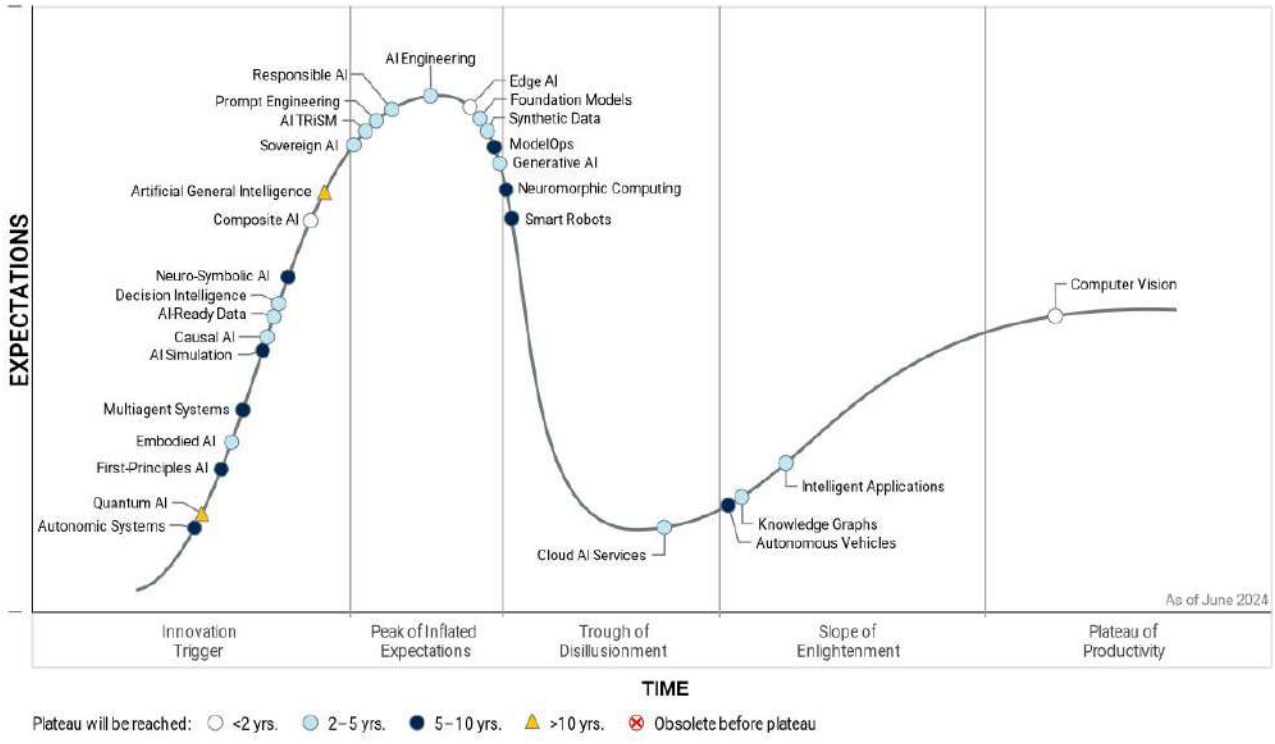
The key questions are:

- **How applicable are AI and ML technologies in business at the moment?**
- **Where is the line between hype and real business value?**

Part 1: AI, today and the near future, and how to start benefitting

Unlike many other Hype Cycles, this one features numerous innovations with transformational or high benefits, none with moderate benefits, and only one with a low benefit. Gartner predicts that composite AI will become the standard methodology for AI system development within two years and see widespread adoption. Another transformational innovation, computer vision, is already experiencing mass consumer adoption through smart devices.

HYPE CYCLE FOR ARTIFICIAL INTELLIGENCE, 2024



Notable innovations expected to reach mainstream adoption in two to five years include decision intelligence, embodied AI, foundation models, generative AI (GenAI), intelligent applications, and responsible AI. Early adoption of these technologies will provide significant competitive advantages and help address challenges related to integrating AI models into business processes.

At the same time, AI/ML is not a specific technology but rather a whole set of related technologies that appeared at different times and often grew in entirely different conditions. Therefore, they are in different stages on the hype scale and add various risks and opportunities to consider when implementing new technology and product solutions.

Many technologies in the second hype wave, which are on the rise, are directly or indirectly related to AI/ML and will invariably impact business within 2 to 5 years.

In other words, we are currently going through a stage when these technologies are reaching their peak popularity and still have significant promise and potential, which will one day certainly be fulfilled. One challenge is that AI/ML technology is fragmented, and keeping track of all the potential can be hard. In these areas, technology leaders should invest time, money, and effort to realise business value.

Where is AI being applied today?

COMPOSITE AI

To put it simply, composite AI refers to the integration of various AI techniques to improve learning efficiency and expand knowledge representation. This approach broadens AI abstraction mechanisms, providing a universal platform for solving a wide range of business problems.

Acknowledging that no single AI technique is a cure-all, Composite AI combines "connectionist" approaches like machine learning and deep learning with "symbolic" methods such as rule-based reasoning, graph analysis, and optimisation techniques. This fusion aims to create AI solutions that require less data and energy, embodying more abstract mechanisms.

At the same time, composite AI extends capabilities to organisations lacking large amounts of historical or labelled data but possessing substantial human expertise. It improves the scope and quality of AI applications, addressing more diverse reasoning challenges. Additional benefits include improved interpretability, embedded resilience, and support for augmented intelligence.

Limited data availability drives organisations to merge multiple AI techniques, using methods like knowledge graphs and generative adversarial networks (GANs) to generate synthetic data.

As you know, integrating AI techniques is more effective than solely relying on heuristics or a purely data-driven approach. For example, a heuristic or rule-based method can complement a deep learning model in predictive maintenance. Expert rules or physical/engineering model analysis might identify sensor readings indicating inefficient operations, which can be used to train a neural network for asset health assessment, incorporating causal AI capabilities.

The expansion of computer vision and NLP solutions helps in identifying or categorising people or objects in images. This output can enrich or generate a graph representing the entities and their relationships.

Real use cases examples:

- **Healthcare:** Integrating image recognition with ML models and knowledge graphs to diagnose diseases from medical images and patient records.
- **Supply chain optimisation:** Combining ML models for demand forecasting with optimisation algorithms and business rules to manage inventory levels and optimise supply chain operations.
- **Fraud detection:** Using ML models to detect anomalous patterns combined with rule-based systems and graph analytics to identify fraudulent activities.
- **Customer service automation:** Combining collaborative filtering techniques with knowledge graphs and business rules to provide personalised recommendations to users.

When adopting composite AI, it is wise to focus on projects where a purely data-driven, ML-only approach is inefficient or unsuitable, such as when data is insufficient, or patterns cannot be captured by existing ML models.

Also, remember to merge ML, image recognition, or NLP with graph analytics to incorporate higher-level, symbolic, and relational intelligence into your solutions.

Where is AI being applied today?

NLP & Chatbots

NLP is a common example of the application of AI embedded in many products and websites. The technology is beyond the peak of the hype wave and is already in the production cycle and well entrenched in the market.

Large corporations such as IBM, Google, Microsoft, and Facebook have introduced a lot of new turnkey solutions and APIs.

Many ready-made models were compiled and can be found in the resource https://huggingface.co/. They are distributed and accessible under a free license.

While there are many exemplary implementations of simple chatbots, much research and development are still required for an interface that allows for quality dialogue between humans and technology. The technology is at an augmented level of customer support services and simple interfaces, but we are still quite some ways off before humans are replaced entirely.

From real use cases, it can be noted:

- **eCommerce:** Consultant bots allow you to advise clients when choosing a product based on their tastes and preferences at a level that will enable them to replace people's work. This approach cuts costs by reducing the number of jobs and enhances the user experience through the "knowledge" gathered during user interactions.
- **Education:** Large companies use chatbots to provide simplified access to databases, provide a streamlined tool for managing complex systems, and even solutions that make it possible to diagnose a node more accurately and with significant time and resource optimisation.
- **Enterprise:** Large companies use chatbots to provide simplified access to databases, provide a streamlined tool for managing complex systems, and even solutions that make it possible to diagnose a node more accurately and with significant time and resource optimisation.

Where is AI being applied today?

Computer Vision (CV)

Computer vision has been around for some time, and there are many applications of the technology. It is now used extensively in many areas, from self-driving cars to space exploration, medical diagnostics, and interviewing technology. It is also used in many supply chain, warehouse, and logistics processes.

When working on real projects, we follow a strict rule: if an object is recognisable by a person, then a machine can also recognise it. It's just a matter of training a dataset or reusing existing datasets.

With every passing year, technology improves. In some cases, computers handle pattern recognition even better than the average person and have a lower margin of error.

Again, the main players are large companies that have a lot of data to train the models and that can afford to deploy substantial resources to create models, which then act as a basis for more specifically used and sophisticated models.

For example, CV is now very actively used wherever you can rely on visuals:

- **diagnostics of defects in production**
- **unmanned vehicle system**
- **medicine (e.g., automatic analysis of X-rays)**

In one of the projects, we implemented a solution that allows you to monitor the instruments used during surgery, improve quality control, and reduce the required number of personnel in the operating room, which has become very important due to the limitations and restrictions imposed by the COVID-19 epidemic and further precautions.

What interesting CV trends do we see in 2024?

Emotion AI

The algorithm that determines a person's emotions by their facial expressions and behaviour is a fairly new technology that is increasingly growing in popularity. Realeyes invested $12.4M to help brands such as AT&T, Coca-Cola, Mars, and Hershey's identify emotions from facial images to rate the effectiveness of different promotional materials. Facebook is also working on creating its product using this technology.

CV and Emotion AIs are being used in people-to-people communication systems. We see new services pop up in areas such as coaching, hiring, and management focused on identifying critical dynamics of human emotion and understanding/interpreting what is being communicated in the language to build a rich understanding of interactions.

Embedded vision

Thanks to the development of portable devices, it has become possible to use neural networks and computer vision models autonomously, allowing to create and operate high-tech solutions that do not require connection to the Cloud. This allows for a fast response time and complete autonomy, which is very important for high-risk solutions – unmanned vehicles or devices, quality control systems in the field, and other portable equipment.

CV as a Service (CVAAS)

On the other hand, using computer vision as a service opens up great business opportunities in terms of the speed of integration and quick delivery of high-quality results. For example, Google provides the service in the SaaS mode, which enables almost instant access to the following functionality:

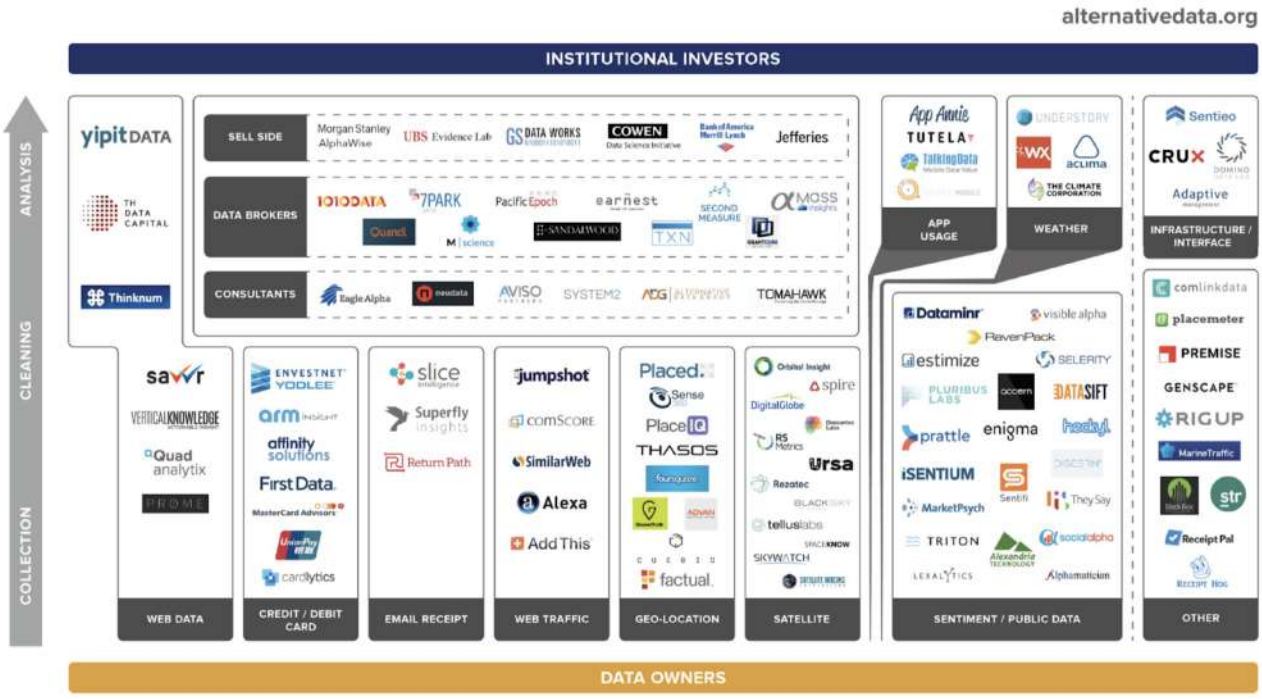
- **Object detection in images**
- **Text recognition**
- **Locate faces in images**
- **Identify popular locations and company logos**
- **Identify and flag explicit content** (violence, adult scenes, foul language, etc.)

This solution is also being built into Google, which makes it possible to continuously train models on existing data with minimal resource outlay, thereby increasing their accuracy.

Summarising the above, one can conclude that CV is one of the most stable and dependable solutions that open up new business opportunities. Simultaneously, the entry barrier is being reduced every year due to the constant work and development of turnkey solutions in the field.

For example, in one of our projects, we integrated a CV system into monitoring surgery. Our approach decreased risks related to human factors in operations and reduced the number of personnel required to be present in the operating room. This new approach to surgery was essential during the pandemic and led to much more efficient operations in the future.

ALTERNATIVE DATA STACK



The diagram above shows the most popular Alternative Data sources collected from https://alternativedata.org/.

Alternative data sources have been used in the financial sector for quite some time to obtain information from indirect sources, which may, for example, indicate that a significant investment manoeuvre is taking place. Because such data is generally not structured and voluminous, it was challenging to work with it before the advent of AI / ML. Still, technologies allow using many sources of alternative data at once and the processing of large volumes of it in real time.

Thanks to this, it is possible, for example, to assess the development success of a young company at an early stage and to predict investment risks, taking into account current trends.

Machine Learning (ML)

ML plays a big role in tracking changes and forecasting the economic situation.

"Most of the current economic value gained from ML is based on supervised learning use cases," says Saniye Alaybeyi, Senior Director Analyst at Gartner.

The application of ML is so extensive and spread out that it is difficult to pinpoint all the niches where it has been bringing real, lasting benefits for some time. The main takeaway, the overarching principle of the technology, can be summarised as the ability to find connections where it is difficult or impossible to establish a relationship algorithmically.

A striking example of ML application, where it is indispensable and irreplaceable, is various anti-fraud solutions. For example, a startup https://scalarr.io/ uses ML to determine how to wind up advertising counters, saving millions of dollars a year for companies that have been the victims of fraud.

ML is being broadly deployed as an anti-fraud agent on a larger scale, helping to identify fraudulent banking transactions, detect defects in complex systems, and even speed up vaccine trials.

How to get started with AI: AI development flow

First of all, you need to find the area where AI can carry the maximum value specifically in your business.

This could be:

- **automation of some processes**
- **business analytics**

Next, you need a person who will have the clearest understanding of how your product and the person (or team) of data science work. The goal of this collaboration is to find a balance between technology and real business value.

This is important, because in the case of AI, we can observe two extremes:

- When a business wants an unrealistic solution (for example, the development of a chatbot, which is expected to completely replace human communication in the entertainment industry)

- When high-flying technologies captivate the human mind so much that a company begins to spend colossal resources on implementing a product whose benefits may turn out to be questionable or irrelevant in practice.

Machine Learning (ML)

Next, it is important to draw up a roadmap that will include a mandatory PoC stage and minimal iterations.

It is better to stick to the practice of “doing little, but often.” This will allow you to receive feedback from the business and quickly influence the development course.

Yes, this approach is more typical for startups than for enterprise companies, but it is fast agility that makes it possible to use modern technologies to the full extent.

It is worth noting that at the PoC stage, you may not get good accuracy, which is normal for AI / ML (more about methods and evaluation criteria is said in Part 2).

Accuracy indicators in the range of 60-70% are sufficient to talk about the continuation of development and usually indicate movement in the right direction.

When PoC is ready, it will be important to choose the shortest path before starting to integrate the solution into production in test mode. Yet, we have to note that it is not so much about using the prediction function or analytic results to make some decisions as about collecting the necessary data.

In AI, the amount of data is more important than the quality of the algorithm.

As soon as your product begins to give the first positive results, do not rush to immediately introduce it into full operating mode. A sufficient testing phase is important at this stage.

We often come across overtrained models that behave very well on the data we are used to. But if some factors change (these may be completely unexpected and, at first glance, insignificant things), how the work of the model can give unpredictable results. Typically, after the solution appears to be working, the testing phase takes an additional 1 to 3 months (or more) in observer mode to ensure the solution is fully operational.

That is, the product performs all functions except to somehow influence the situation in a hands-off, automatic mode.

The product has to include a capability for users to make adjustments and tweaks to it. If someone is creating the model for you, they have to be instructed to consider this important element when training the algorithm.

Finally, even after integrating the AI solution, it is necessary to regularly update the models and automate this process as much as possible.

Read more technical details in Part 2: *How to design & develop production-ready AI application*

Part 2: How to design & develop production-ready AI application

How can businesses benefit from AI? Where is the best place to start, and how can realistic expectations be set and risks managed?

Let's explore this topic by analysing hands-on cases.

The entire process can essentially be divided into several stages:

- Drafting a list of requirements
- Analyzing data
- Describing the logic
- Selecting the appropriate architecture
- Prototyping (iterative repetition, until the desired result is attained)
  - Normalising the data
  - Training the model
  - Assessing the quality of the model at the end 'fine-tuning'
- Expanding the volume and scale of the data
- Setting up continuous training

If we focus on the first objective—developing an accurate list of topics and training models on dialogue data—we aim to teach the system to deliberately avoid certain questions by redirecting them with appropriate responses. Additionally, another strategy is to identify dialogues with a high probability of being troublesome and transfer those conversations to a human operator. While transferring communication to a human operator shows promise, it has proven to be more challenging and problematic to implement effectively.

This is because, for starters, there are no solutions so far that confidently pass the Turing test, and chatbots are not created to replace a person (although there are startups that ambitiously declare that chatbots can now replace psychiatrists). In other words, the requirements should be as clear, specific, and measurable as possible.

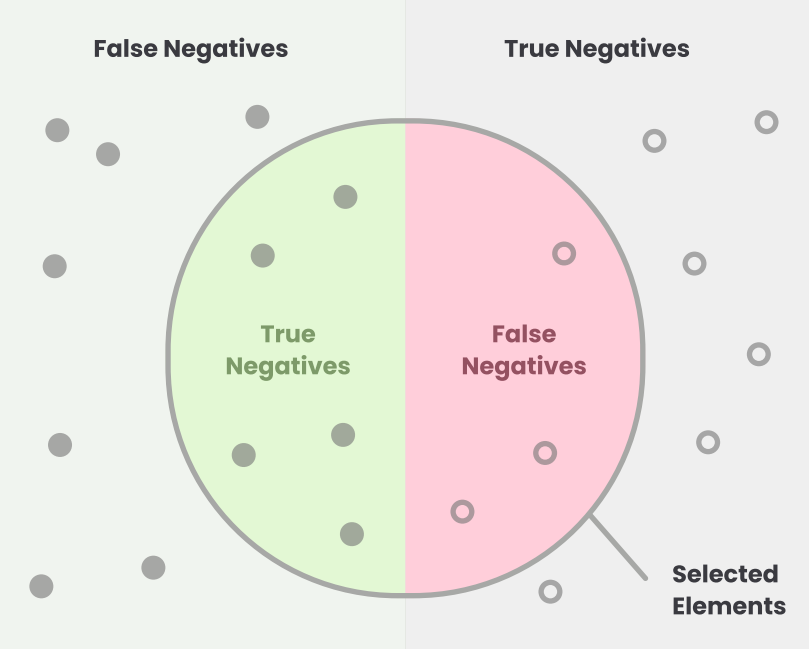
There are generally accepted metrics used to evaluate the quality of a model.

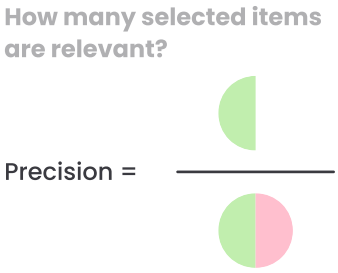
Accuracy is the share of correct answers by the algorithm, calculated by the formula

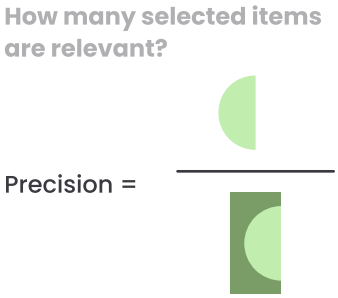
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The problem is that this metric will not be indicative in a model with unequal classes. Therefore, **Precision** and **Recall metrics** are used more often.

Precision can be interpreted as the proportion of objects named as positive by the classifier, which is a true positive. Recall shows what proportion of objects of a positive class out of all existing objects of a positive class are found by the algorithm.







Using business metrics to select the appropriate values and draft expectations is also a good way, but it is advisable to already use metrics and values in technical terms when drawing up requirements specifications. At this stage, it is very important to bring in a technical expert who can sync up the business requirements with the technical terms and validate these requirements against technical feasibility and context specificity.

Analyzing data

Sometimes, this stage is closely related to and even dependent on the previous one; in some cases, it serves as the starting point. If AI is the heart of a project, then data is its blood. The project's final result will depend on the quality and, most importantly, the volume of data.

At this stage, sample data is typically collected to get a sense of what lies ahead. Data homogeneity and “purity” are assessed. In addition, rough estimates are made for the volume of data processing capacity to understand the estimated costs associated with processing and storing this information.

Resources are still quite expensive, especially GPU/TPU. Therefore, at this stage, it is important to make a realistic estimate of the maximum financial outlay required to complete the models’ training and maintenance stage to ensure that the spending fits into the allocated budget.

Due to the extensive text data and massive models such as GPT-2, the training stage can be so expensive that looking for other ways of solving such business problems will prove more feasible.

Describing the logic

AI can do what humans can, only faster and at great volume and scale. Before assigning tasks to a technical team, it is necessary to have a person try operating the flow where AI is to be introduced. The data needs to be in a state where a human can understand it and make good decisions. If it works for a person, then it is possible within the bounds of a real neural network.

When dealing with big data, a human may be unable to process large volumes of information quickly enough. Therefore, the task may need to be scaled down to smaller volumes. However, it's important to ensure that if we scale the flow back up for the big data algorithm, the process will not change conceptually.

This stage should result in a clear and understandable flow, with data examples, which will be handed over to the technical team and implemented in AI.

Selecting the appropriate architecture

Depending on the written logic, the volume and nature of the data, the requirements for the speed of work, and other inputs, the optimal solution architecture includes the following:

- **Methods of collecting, normalising, and storing information;**
- **Underlying data model** (in some cases, the best approach would be to use a pre-trained model as the base layer; solutions give the best results when data-based models are built from scratch)**;**
- **Model training and hosting tools.**

The result of his stage should be architecture diagrams and their textual descriptions, which could serve as a starting point for implementing the first version, which will test the concept.

Lightweight prototype (iterative repetition until the desired result is attained)

The goal of this stage is to verify or refute the concept with minimal resource outlay. Usually, at this stage, you use tools such as Python, Notebook, and SaaS solutions such as Google Colab (if necessary, you can connect external GPU/TPU instances to expand computing power).

The goal is to continually refine prototypes to develop a proof of concept that can achieve satisfactory accuracy rates and be embedded in the user experience you aim for.

The prototype implementation moves iteratively through a cycle:

- **Normalising the data**
- **Training the model**
- **Assessing the quality of the model**
- **Fine-tuning**

Each of these stages can have an impact on the outcome. Usually, at this stage, the resource consumption indicators have little importance when compared to the achievement of an accuracy result that is acceptable. After the prototype successfully demonstrates a proof of concept for how it might fit within your system or product, one can estimate the resources needed to bring the solution model to production.

A helpful rule of thumb we have found in our work to turn a prototype into a full production version is that 20% of the time and budget are spent on the prototype and 80% on developing a scalable production version.

Expanding the volume and scale of the data

After the prototype is implemented at a small scale and the basic version of the model fulfils its task, it's time to test the model on a realistic data set of greater scale.

At this stage, simple scripts used in the prototype will likely be struggling for collection and normalisation. Tools like Apache Airflow, Apache Nifi, or Hadoop/Spark will be needed to operationalise these processes for production scale. A well-built architecture and the choice of appropriate tools will initially allow to effectively scale up the process without slowing down the development by changing the technologies on the fly.

We usually recommend that our customers use SaaS solutions such as GCP BigQuery/AWS Redshift to store data and use GCP Pub/Sub or AWS Kinesis for transporting data. In case the company's privacy policies do not allow access to Cloud Providers, there are alternative solutions. To store data on BareMetal, you can use the Cloudera stack or its more modern counterparts.

Practice shows that expanding data volumes increases accuracy with minimal fine-tuning of the model. In cases where the reverse occurs, the issue is most likely in the normalisation of the data.

You can always go back to the beginning and go through all the steps – from normalisation to the model evaluation using different arrays of information. It is important to collect metrics and document any changes to expose any tendencies that change the results. A conscious approach can save you a lot of time and resources by the time you reach the implementation of the final production version; just think about how time-consuming each new training session is.

Depending on the algorithm and area you are considering, you might be okay with a one-off process of training a model. However, many algorithms need re-training over time as you learn more about the processes and data. Our experience shows that many models functioning well today may not remain that effective even after a short period. Therefore, as you develop your solution, it is important to completely automate the solution's training process for continued improvement and automate the collection of the model's efficiency metrics and deployment process.

Solutions such as https://www.kubeflow.org/ help automate this process both at the stage of development as well as when models are operational.

This area is often overlooked in projects as people rush to quickly embed AI in products and systems. However, it is an area of huge value creation for businesses and can make a difference in creating successful AI/ML capabilities.

Conclusions

AI is already being successfully applied by many companies in myriad ways.

Now, we see an actively developing trend in using edge technologies for AI / ML. They allow performing AI / ML operations on edge devices as quickly as possible and without loading centralised systems, which opens up a new horizon of possibilities. Such technologies are already used in video surveillance systems and other security systems (as a rule, solving the face recognition problem at the edge and transmitting the information already received to the cloud), advertising, and even healthcare.

The part of the AI stack dealing with language models, computer vision, and machine learning proves immensely efficient and is used not only by startups but also by extremely large companies.

The market capitalisation of AI solutions is growing, and investments in the development of hardware solutions and ever more new and in-depth research are growing along with it.

There are plenty of examples of AI integration into existing businesses that have helped solve problems such as quality control, process automation, and forecasting based on historical data.

User behaviour analysis allows for a better understanding of the customer and to test new theories and hypotheses in a safe environment. It saves a great deal of resources at the stage of preparing innovations and reduces risks associated with launching new products.

There is no denying the fact that the world has already irreversibly changed. AI has become deeply ingrained in the list of tools to be widely applied with each passing year. Although we have yet to achieve this, we can say with certainty that AI can solve certain tasks at the level of a human being or even better, and its targeted use can not only cut expenses and streamline processes but also bring the business to a fundamentally new level of quality.

About Altamira

Altamira is a global digital transformation partner, specialising in helping organizations scale faster and more sustainably than anyone else.

With market-leading capabilities across all aspects of product and technology development, we help you find the optimal way of implementing AI/ML solutions to achieve business goals.

- Understanding the business goals and assessing the possibility of achieving them with AI/ML solutions
- Identification of possible use cases of AI/ML models and their implementations within your business environment
- Feasibility validation and understanding the quality of your data and their usage by the proper AI/ML models
- Identification of the right pre-built AI/ML models to be implemented within your projects

Let's turn your AI concepts into tangible realities through Artificial Intelligence and Machine Learning.

HAVE A QUESTION? LET'S SCHEDULE A CALL.
WE LIKE TO HELP OUR CLIENTS BE SUCCESSFUL!

Phone number: **+421 948 656 863**
Email: **hello@altamira.ai**
www.altamira.ai