WHITEPAPER

The AI productivity challenge

JULY 2018

Artificial intelligence (AI), in particular the subfields of machine learning (ML) and deep learning (DL), have shown remarkable progress in recent times. There is a general consensus that over time every business will need to start using AI if they are to be competitive and successful. AI is not something for the future nor something only for the large technology titans but rather a technology that is increasingly entering the mainstream.

This is an introductory paper on how to drive productivity in the areas of ML and DL. The paper highlights some of the key barriers to getting good outcomes in these projects and then outlines some potential functions or capabilities organizations should consider as they look at technologies to help support their AI programs.

1/ The productivity challenge

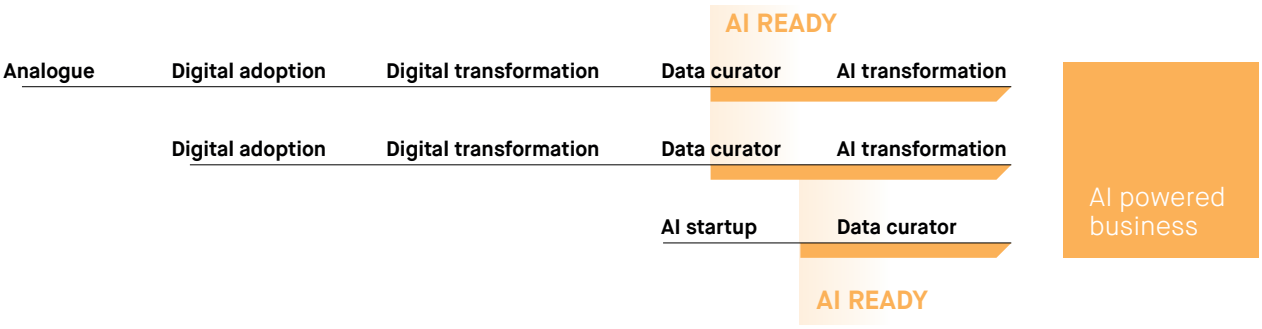
Productivity is an economic measure of output per unit of input. This paper will consider the productivity of AI and deep learning projects as well as the productivity of the data scientist and teams. Three points are relevant to improving productivity: identifying the impact of reducing input and costs as well as increasing output and value, and then analyzing the long-term impact in order to generate an ongoing productivity lift, not just a point-in-time boost. Businesses that only get one-off productivity boosts simply become less competitive over time.

AI, machine learning and deep learning have been in existence for over 50 years. Key triggers for the current momentum of AI are centered on data: much more actual data has been generated over the last decade, that data is more readily available for analysis and that data is able to be processed by increasingly powerful and sophisticated processors. Given that data is at the core of these projects, **the data scientist has emerged as an essential resource** for any company that wants to use data to improve their business and move into the more sophisticated AI realms.

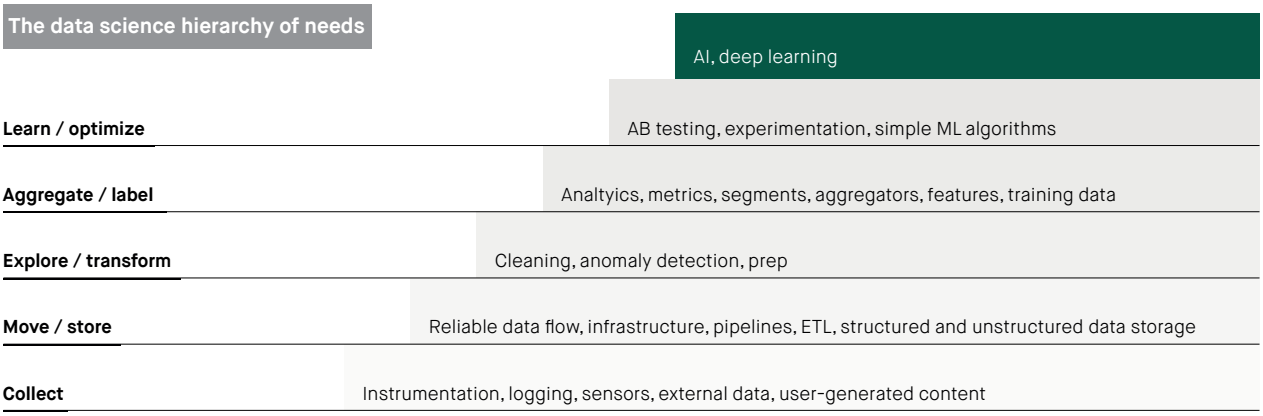
First of all, organizations should be realistic about their current competence and capability levels with data. For example, data-led organizations have built some competence and expertise through internal and external resourcing and have some experience with data, data-based projects and data scientists. There is a scale to determine the level of a data-centric or data-competent organization. Related to data capability is the digital transformation process and how progressive organizations are using data to drive digital initiatives across their internal and external processes. Companies that are advanced in their digital transformation efforts will typically have two attributes that are relevant to AI: plenty of source data and some level of competence in driving business activity from that data.

The data scientist is the key resource to getting value out of data and driving productivity from these projects. Regardless of the specific project objective and composition, the team needs to make sure the relevant data has been cleaned and structured in a way that allows further analysis in order to drive insight.

Organizational path to AI ubiquity



There is an increasing level of complexity in data science projects and more advanced analysis often depends on doing the more basic processes first. This starts with reliable data collection, storage and preparation. The dataset may need to be managed, cleaned, deduplicated and put into the right format with anomalies removed and gaps filled. Then comes a series of increasingly sophisticated processes to drive insight from the data, including basic structuring and classification to applying models through to the more advanced machine learning methodologies.



As of today, few companies have actually used machine learning or deep learning methods in meaningful ways. However, the technology is maturing and the data being generated lends itself to these kinds of methods. Businesses should be planning for these more advanced processes to become an integral part of their operations.

The value of any individual AI project may be great, but even more important is building deep data science expertise into a business in order to manage and run a broad suite of projects. Projects do come with inherent risks:

- Projects are often very complex
- Each project is an innovation project; the process requires experimentation and learning, and outcomes are not known at the start
- Project outcomes may present a challenge when linking to organizational goals (e.g., How do you turn real valuable insight into actual actions that drive impact/value?)

It is important to note that lots of **time and effort is expended by data scientists on necessary but lower value tasks** in these projects. There will be greater productivity improvements if:

- **Data scientists and teams are able to focus on higher value activities** like modeling, insight and outcomes rather than all the support processes. This is the 90/10 rule: **Data scientists spend up to 90% of their time on lower value support tasks.**[1] Every percentage point increase in higher value tasks would improve productivity.
- There is reduction in both effort or time to run these projects. There are a range of areas where **there are opportunities for "productising" some data science tasks.** Over time, more and more tasks can be productised.
- "Productising" tasks can allow more narrowly skilled or less experienced team members, for example, developers, **to start taking more involved and leading roles.**
- Data scientists and the organization are encouraged **to seek out ongoing improvements versus one-off gains.**

[1]Peltarion Research, March 2018

The productivity challenge leads us into **some of the specific barriers that arise in these projects**, and the criteria organizations need to consider for AI technology as they look to ways of **improving the productivity of their machine learning projects and related investments.**

2/ Challenges with AI projects

The following goes into some detail regarding the potential barriers to success and/or improved productivity. These challenges apply to the technology, people, resources and project processes themselves. Overcoming or reducing these barriers will significantly drive productivity improvements for AI projects.

Lack of talent

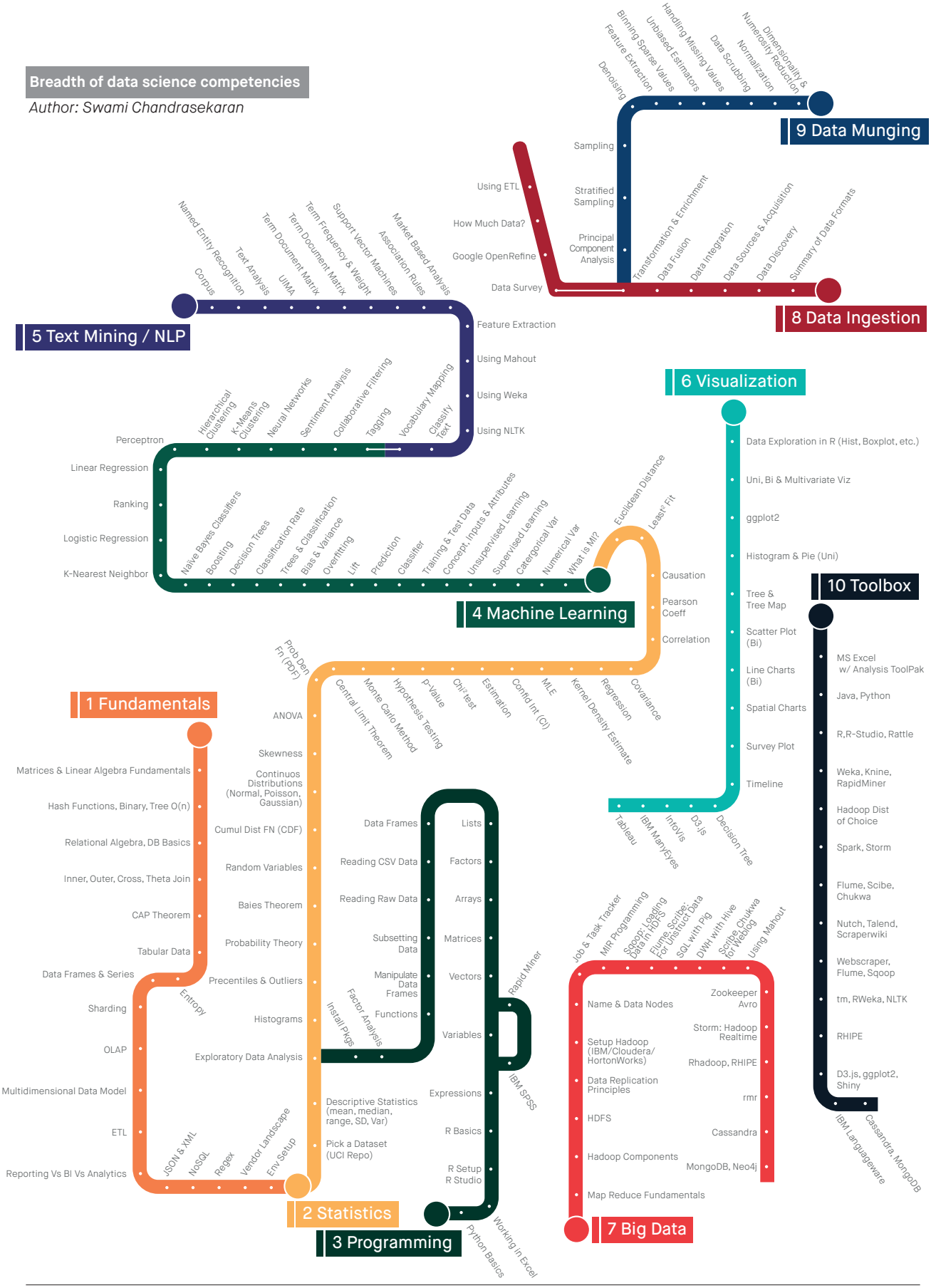
Many discussions on AI, machine learning or deep learning projects start with the need for talent. Historically, there have been few people working in the area and typically they have worked solely in academia. The explosion in demand as a result of the potential of AI, has created a talent gap, which is exacerbated for most organizations as the top talent is acquired by large companies, especially the large technology companies. For example, average salary costs at Google's AI laboratory DeepMind wash out at $USD 345,000 an employee.[2] Given there will be a mix of staff at DeepMind, the actual average salary of the data scientist will be much higher. If we layer in forecasted demand for these resources, the problem gets worse. For example, IBM predicts an increased demand for 700,000 more data scientists by 2020 in the U.S. alone.[3]

The talent gap will not improve in the short to medium term, and irrespective, a current necessary condition for a successful AI or deep learning project is skilled and capable data science resource(s), and they are hard to find and expensive. Projects also vary across business domains and business objectives, data types, relevant modeling approaches and target production environments. So the type of resources in a team needs to be mapped well to the specific project and objectives. Not only do you need good data scientists but you need the **right data scientists and teams mapped to the right projects.** As noted earlier, there is a very broad set of competencies for a data scientist, and what is needed changes per project. The next figure outlines skills and competencies needed in AI and doesn't even include the Deep Learning area.

[2]New York Times, Oct 2017
[3]Forbes, May 2017, "IBM Predicts Demand For Data Scientists Will Soar 28% By 2020"

Breadth of data science competencies

Author: Swami Chandrasekaran



Organizational culture

These projects work best when there is an organization investing time and resources into the AI project or program. Therefore, aligning the organization and the AI team is important. The promise of AI is to make significant impact on the business. In assessing the potential of a project, follow the money (e.g., revenue, cost or margin), which is where the greatest impact can be for an organization. For nonprofits, the focus on impact applies in a similar manner; however, the measures veer away from money or profit. There may also be a tendency for organizations to pick short-term marginal improvement projects that are deemed safer. The ideal scenario is a mix of projects with some big impact AI projects sitting in a portfolio with others that are faster to implement, lower risk and have smaller impact. Treating these projects as a portfolio is a useful way of managing expectations so that there is no excessive focus on a single project (which may fail), and a wider set of more "diversified" projects may also reduce overall risk.

Migrating to production

In any project, there are challenges to migrating to production. Even if the data science team comes up with a model that works, it will need redevelopment and adjustment to work in a live production environment. Moving from an initial prototype or from an academic implementation into a production-ready version in a live environment can create a significant amount of work and investment in the infrastructure surrounding the AI system. The code needed to support the systems can be referred to as "plumbing" and may be significantly larger than the core code itself (Note: Using the terminology of Lin 2013 scaling). The typical process involves the data scientists handing over the prototype project to the main IT organization to re-implement in a way that fits the requirements of the production environment. This naturally introduces both new effort, as the data and the model need to fit into the more complex live production environment, and new bugs, as the production environment is different which means the effort needed is increased even more.

Maintaining the model

Over time companies will end up with large sets of models that have been put into production. Keeping track of those models' dependencies and making sure they are updated to remain stable is a big effort in itself. Not only do the most popular software packages for AI update very often with breaking changes, hardware also updates regularly with breaking changes as well. On top of that, any model needs to be monitored continuously to ensure it is still valid. Since the reality that the model captures changes over time, it is necessary to retrain the model to keep it up to date.

Effort estimation

Although the goals of an AI project can be well defined, there is no way of guaranteeing when a model will achieve the desired goal. It is very difficult to predict the model structure and parameter settings that will lead to good results. Therefore, the data scientist will run a series of experiments until the solution is good enough. This is the nature of any research project. In addition, it is normally not possible to either adjust/decrease scope or run the project in a time-based manner with a predefined delivery date.

It will always be hard to estimate the effort needed in a project. If it were possible to do more experiments per time unit the likelihood of getting new projects approved would increase. In the productivity challenge section of this paper, it was noted that running data science projects have inherent risks, and if you run a portfolio of projects it can reduce the focus on any individual project. Further, it can be useful to overlay specific innovation processes, and implementing decision benchmarks can help guard against projects being shut down despite promising intermediate results.

Infrastructure setup and resource limitations

Working with deep learning projects requires having access to specialized hardware. Regardless of whether it is on premises with dedicated servers or utilizing cloud providers to get access to GPU machines, there is a lot that needs to be configured to be able to start running the training code.

The tools and libraries for setting up and working with GPUs are getting better and more user-friendly, but it is still a challenging task even experienced data scientists will often find difficult. The GPUs themselves are improving and there is also entirely new hardware being developed that is specially designed for deep learning calculations which can require thousands or even millions of matrix calculations. Once this type of hardware is readily available it will be orders of magnitude faster than current general-purpose hardware. However, it is likely that the software will need major reworking because it will need to use an entirely different library to benefit from the new speedups.

Training a model often requires large amounts of data. The project needs to be able to fit the model and a training batch of data in the GPU memory, and even though memory capacity is increasing there can still be bottlenecks and constraints. Therefore the training model may be required to be distributed across multiple GPUs and machines. Working with data on a distributed system adds another layer of complexity compared to single machine solutions. There are open source solutions that help, but these solutions introduce complexity themselves. Finally, another significant challenge introduced by making use of distributed systems is the increased difficulty in troubleshooting problems which will be covered in more detail below.

Glue code

One property of machine learning systems, and especially deep learning systems, is that only a small part of the system deals with the actual model. In a production-ready system, around 5% of the code may deal with the model, while the rest is "glue code" that interacts with supporting systems and glues all libraries and systems together [Sculley 2015 hidden]. Also, as the number and efficiency of different cloud providers continue to increase, a greater percentage of the code interacts with external systems that are outside the control of the designer.

Keeping the glue code up to date, and keeping up with changes in cloud services, requires effort and can introduce unexpected challenges. At the same time, making use of cloud services means not having to build and manage all services yourself, leading to significant reductions in time to production.

Troubleshooting

Developing an AI system is fundamentally different from a standard software project in that you leave much of the responsibility for finding a working solution to the computer. The lack of model transparency and reproducibility makes it hard to identify issues and chase bugs, which in turn makes the effort required to complete the project even more unpredictable. As noted, the sheer complexity of the process means this can create huge effort to chase and then resolve the bugs without them breaking another part of the model.

Experiment management

During the development of models, a large number of experiments are usually performed to identify the optimal model. Experiments can differ in a number of ways, and in order to have reproducible results, it may be necessary to know the exact version of components such as:

- Hardware (e.g., GPU models primarily)
- Platform (e.g., operating system and installed packages)
- Source code (e.g., model training and pre-processing)
- Configuration (e.g., model configuration and pre-processing settings)
- Data sources (e.g., input signals and target values)
- Training state (e.g., versions of trained model)

Furthermore, different versions of every model are created during training, each with different parameters and metrics that need to be properly measured and tracked. With the addition of data dependencies and a high degree of configuration parameters, it can be very challenging to properly maintain the systems in the long run. Also, you want to perform hyperparameter optimization of models whereby you generate hundreds of versions of the same data and model but with different configuration parameters [Cite Golovin 2017 Google].

Although there are some basic tools and open source libraries that make it easier to keep track of your experiments, many data scientists still rely on manually keeping track of details in a spreadsheet. This is an intensely manual and recurring process which requires significant effort from the data scientist or data science team and is often error prone.

Model transparency

Deep learning models lack transparency which makes it difficult to understand and be able to explain how results are reached. For example, in highly regulated industries such as banking or healthcare, it is very important to be able to explain model decisions. Often, convincing decision-makers in an organization requires some ability to explain the "model." The inability to explain in detail may lead to less efficient methods being used because it is easier to explain how it works. In circumstances where the model can be explained, the effort required to actually explain it will be significant.

3/ Dealing with those challenges

Productivity improvements will come when **you free up both the data scientist and teams to focus on high value activities** like modeling, insight and the business domain areas and business outcomes rather than all the support processes. This represents the 90/10 rule referred to earlier.

In the preceding section a number of barriers to productivity were identified. Some of the areas involve the time of the data science team. Others involve challenges like effort estimation that will never be totally addressed, and a few others were noted where putting best practices in place will help. The overall message is to aim to effectively utilize the time of the data science team where it matters most like understanding the business and business problems.

For AI and deep learning to really match its promise, all companies should have access to more effective technology. Improving the technology is a necessary condition for improving productivity.

There are a range of platform/product capabilities that could help positively impact productivity. For example, operating an on-demand model would improve deployment, usage and scalability. In addition, working with prepared datasets and an intuitive graphical interface for developing and quickly iterating models, training them, evaluating and comparing results would also have a big impact.

It takes an experienced AI data scientist some time to get even a simple proof-of-concept model training. Running a "platform-based model" makes it simple to get started from scratch and create bespoke AI models using your own data. Models should be able to perform well on the dataset whether image-based, tabular, written text or combinations of all of these. The ability to effectively combine very different types of data out-of-the-box is one of the most attractive properties of DL models.

If teams could use GUIs and graphical representation of data and models, it would take minutes to import your data and create a fitted deep learning model, getting guidance on how to design your model based on the data you use and the problem you are trying to solve. Once you start training the first model in the background you can try your next idea in parallel and check in on the progress of the ongoing trainings as needed. For complicated models, it might take some time to train but it should have early results which enable you to start working closer to the people in the business who are eagerly awaiting the results you will provide.

Buyer's guide

Many of the challenges identified for these projects can be impacted by technology. The market always responds to challenges and opportunities. Solutions will come to market to support making data scientists and teams more productive. These may evolve to solve specific problems across areas such as experiment
management or infrastructure setup and so on. The current market state is that many point solutions are coming to market as well as some end-to-end platforms which address the complete workflow. Currently, the end-to end platform offers are superficial and have limited capabilities but expect these platforms and capabilities to deepen.

As we return to the productivity challenge, we are looking at a process to impact inputs like cost, outputs (like business KPIs) and drive improvement over time. Stating the obvious, if 90% of data scientist tasks are focusing on support tasks then there is significant ability to both free up time for more value impacting tasks, and in effect "upskill" the talent pool. Freeing up time for value added tasks directly impacts all. There is also the potential for more junior resources to take on more challenging projects, and other groups like developers can migrate into these projects.

Moving from point solutions and software to a more integrated platform environment where the workflow can naturally move from one task to another would be a big step forward. Further, as these platforms evolve more capabilities will be available to help improve AI/DL processes and activities. The evolution of capabilities typically will impact the productivity of these projects.

With regard to the solution space, some buyer guidance may help to ensure important requirements are considered. Whether taking the first step toward integrating AI solutions into a company or a company who already works with AI to a large extent but wants to do so more productively, these are features or capabilities that should be covered in prospective solutions.

Generic buyer guidance

General

Cloud and data agnostic: Regardless of data location, you should be able to work with your models without moving everything to a specific cloud.

Zero setup: You shouldn't need to install hardware/software to get started.

AI simplified: Easy-to-use drag and drop editor to help you build advanced deep learning models without coding or sacrificing flexibility when coding directly in Keras, Pytorch or similar.

Flexible performance level: Able to pick the compute resources that fit your data/problem to avoid overspending.

Track record in machine learning/deep learning and compelling vision: Solutions are backed by organizations that intimately understand machine learning and have a strong grasp on solving current and future machine learning/deep learning challenges.

Security

Security: Flexible role access management for groups of users, assured privacy of sensitive datasets, models and projects.

Modeling

Assisted experimentation: Automatic data type detection and assistance to select the best model design to reach your goal.

Pre-trained models: Get started with pre-trained models. Simply select or import a model and start building your solution around it.

Model and experiment versioning: Models should be automatically saved after each epoch during training. Users can easily access previous model versions.

Assessment

Model comparison: Compare model performance across versions by viewing results in a consolidated graphical display.

Parallelized training: Distributed multi-machine training to increase turnaround time.

Project

Collaboration: Work together in projects using a nonlinear, parallel interactive workflow, while maintaining comparability and traceability across all tasks.

Project mechanics: Tools to manage the resources, status and outcome for projects.

Deploy

Zero-effort deploy: Whether you want to create a simple REST API or you want to create a library to integrate into your app or business process it should be no more than a couple of clicks away.

Support and maintenance: In a live environment, the model and software needs to be supported, maintained and updatable for both changes in the surrounding software and hardware environment and also for updating the model itself.

Future proof: Solutions which abstract up and don't get locked into a hardware or software stack that will get outdated before the end of the project life cycle.

Conclusion

Businesses looking to the future know that artificial intelligence will become central to achieving a competitive edge. This paper summarizes some of the key strategies to drive productivity in the AI and deep learning space. And while there are a number of barriers to achieving successful outcomes in these projects, some specific to AI or deep learning and others to complex software development projects, AI can be accessible to all. Going forward, Peltarion will continue to provide more insight to help all organizations succeed in AI. The Peltarion operational AI platform forms part of achieving better outcomes in AI projects.

About Peltarion

Peltarion's operational AI platform is a collaborative, graphical cloud platform for building, managing and deploying machine and deep learning models at scale. Peltarion makes AI technology more attainable and affordable by eliminating the software engineering overhead related to AI. Peltarion helps companies to focus on the actual value driving side of machine learning, instead of the infrastructure. Cutting down time to market and significantly improving the value of AI projects.

Contact us

To request early access to our brand-new platform and speak with an expert at Peltarion, contact us today.