KPMG

Powering Data-Driven Innovations

Enterprise Data Science & Machine Learning Platforms

kpmg.ch

The adoption of data science & machine learning within the enterprises is a key success factor for driving client-centric innovations and boosting process efficiency. There have been significant changes in the last years and we see continuously that Open Source technologies empower enterprises to build and maintain fully integrated Data Science & Machine Learning (DSML) platforms aligned to their specific needs. The right platform strategy is important to engage employees into the enterprise AI journey.

I. Introduction

The complexity of Digital Transformation initiatives increases with the integration of business and technology platforms which requires end-to-end solutions. Achieving "data insights" and "unlocking the value of your own data" in the implementation stage relies on appropriate technological structures and platforms. It is about building a flexible and modular – but still governed – framework within an enterprise which allows to drive centrally data-driven initiatives for the business and avoid fragmented solutions from an IT point of view.

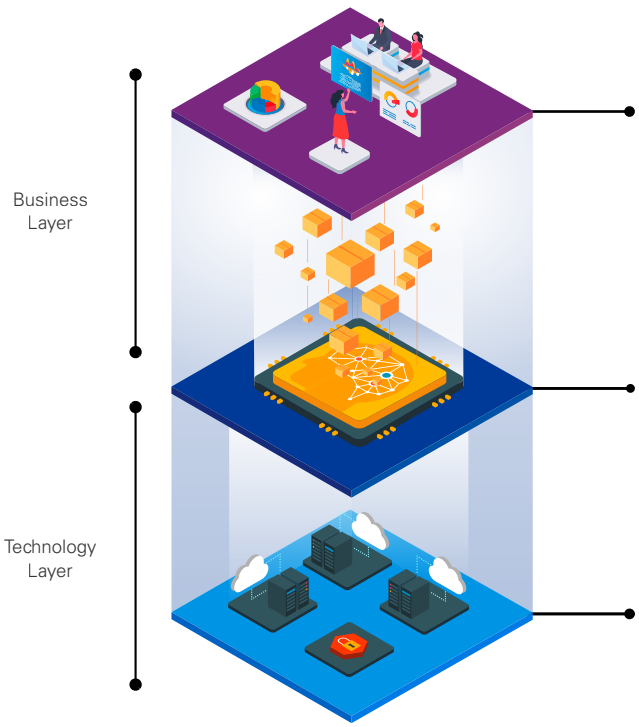
There are quite dynamic and different maturity levels within enterprises on Data Science & Machine Learning platforms. We believe that empowering enterprises to build, maintain and control their own enterprise DSML is a beneficial approach. It creates a different understanding of the own data within the enterprise and how to manage data-driven initiatives contributing the digital transformation journey. It is also important to balance appropriately the various platforms and modernization needs as core systems do not solve specific DSML requirements and vice versa.

Digital Transformation is delivered by people and enabled by technology

Team organization and agile culture are key factors to successfully adopt and embed AI capabilities and to integrate innovation and digital growth. As technologies and platforms are changing fast, the ability to adopt these changes, constantly gaining traction and advantages from a fast-changing environment is a fundamental organizational capability to stay competitive in the future.

From our experience the root cause for missing goals of innovation initiatives often lies in the organizational setup. It is important to link not only strategy, business and IT but also to involve impacted practitioners in the digital transformation process.

When defining teams and structures it is crucial to define a clear focus and to establish clear – and required – communication paths.



User Experience and Business Differentiation
- Create differentiating user experience to realize sustainable business outcomes
- Exploit new technologies to rethink business processes

Business and AI Applications
- Leverage external and enterprise-wide data to forge unique data insights in business processes
- Infuse Artificial Intelligence, Big Data and new technologies
- Compose re-usable building blocks to enable short iterations during solution development

Technology Platforms & Cloud
- Cloud-technology based flexible infrastructure orchestration
- Application Development platforms to enable agile solution development
- Data Science & Machine Learning Platforms to allow governed access to Data and enable standardized Data Science lifecycle

Fig 1. Capabilities for connected enterprises.

Build teams with clear focus areas to achieve continuous successful results

A DSML allows value stream-aligned teams to build and maintain data-driven solutions while ensuring IT quality standards and compliance with data privacy regulations. Such platforms allow data scientists to easily access data, work with data and the required infrastructure to build, integrate and maintain solutions. To foster collaboration between data scientists, application teams, data owners and data stewards, a platform is underpinned by provided workflows and real-time views on dataflows and data usage.

Thus, any initiative towards a central platform for DSML should essentially stick to one central goal:

Making data usable, while governing compliance and risks at the same time are the major goals for an enterprise Data Science & Machine Learning Platform

The term platform is trending in the discussions around enterprise grade data science. The market landscape is highly fragmented and hard to understand as focus areas and approaches of platform products highly differ from each other because they are originating from different angles. Traditional Analytics and Business Intelligence providers try to include end-to-end machine learning (ML) capabilities in their products. Technology and Cloud providers on the other hand extend their portfolio into the direction of analytics while providing strong capabilities for ML infrastructure and integration into DevOps processes and tooling.

A Data Science and Machine Learning Platform is not a single big black box

It is very unlikely to find a standard product that really covers the whole range of the Data Science & Machine Learning Lifecycle. The integration into an existing IT landscape and existing data governance processes lacks behind. We recommend assembling a platform based on open source components, adapted to – and growing with – the needs of the enterprise. This also ensures the engagement and trust of employees which work on and with the platform.

A platform is not required to be one big application. What is more important is that everything, which is covered by the platform, is accessible via common interfaces and tools (e.g.Web UIs, Command Line Interfaces, Software Development Kits). Thus, the platform is more likely an integration layer which orchestrates dataflows and infrastructure while ensuring defined data governance processes. As such, a platform can evolve with an enterprise AI strategy and with technical trends without being coupled to a product road map provided by a specific vendor. All knowledge and technologies are free and publicly available. State-of-the-art results of AI models can be achieved with open source knowledge and components without exception.

A proper platform consists of various components across the whole IT landscape

A DSML consist of various tools and components distributed across the data scientists' workspace, application development environments and the production environment. The following components can be identified:

- Sandboxes: isolated environments
- Runtimes: managed infrastructure
- Low-Code components to create UIs and dynamic reports
- Experiment tracking, collaboration and project-sharing tools
- Data Shop: a central hub to find and access data
- MLOps CI/CD integration
- Automated re-training toolchains
- Fast data processing integration building blocks
- AI in Control – full lifecycle monitoring for AI solutions to ensure integrity, explainability, fairness and resilience.

Fig 2. A DSML embeds and connects various elements across the data science, application development and production environment.

II. Platform elements within the Data Science & Machine Learning Lifecycle

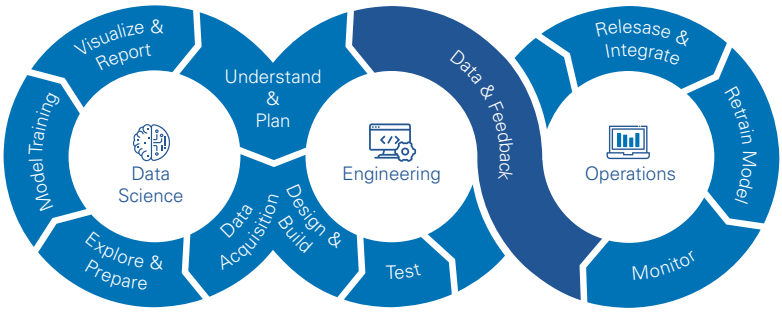


Fig 3. The Data Science Lifecycle as an extension to the prevailing DevOps loop.

Mature data science practices should be embedded in existing application development and operation processes. Thus, the prevailing DevOps Cycle can be adopted to integrate data science processes. An important difference to a DevOps Cycle is the propagation of data from production environments back to the data science experimentation environment. Additionally, to traditional monitoring & application analytics measures, data-driven solutions require dedicated mechanisms to return feedback from end-users or expert-users back to data scientists.

In the following we outline the phases of the lifecycle.

Understand & Plan as foundation stone for dedicated focus on user centricity

This first phase focusing on user centricity launches the overall implementation road map. It depicts the business challenge to be solved not from a technological point of view but from an end-user view. User centricity is at the core of all succeeding steps and serves as guideline to continuously reflect on the 'why' of the solution. To articulate the need co-creative techniques are leveraged and the end-users are actively involved to maximize the adoption rate.

Following a structured approach allows for:

- rapidly understanding of the users' current situation and the corresponding pain points
- Ideating potential solutions
- Creating quickly tangible results for the end-users and the stakeholders
- Capturing feedback and continuously improve the potential solution
- Scaling the solution across the enterprise

To avoid only short-time experimentation and move to sustainable technology-driven innovation projects collaboration is essential. Thus, emphasis is on leveraging interdisciplinary teams consisting of Design, Business, Technical, and Data Specialists. During this process design, SME play a particular important role in order to create clickable prototypes within few days to visualize the future user experience. This creates alignment across all stakeholders, which is a key success factor. From client-projects-driven experience a funneled focus on desirability, feasibility and viability also helps to identify solutions with scalability potential.

To reach a production-ready state from a PoC the transition into agile delivery is seamless, while the continuous improvement is still maintained. The features defined in previous phases will be prioritized in a product backlog which acts as a repository for future sprint planning.

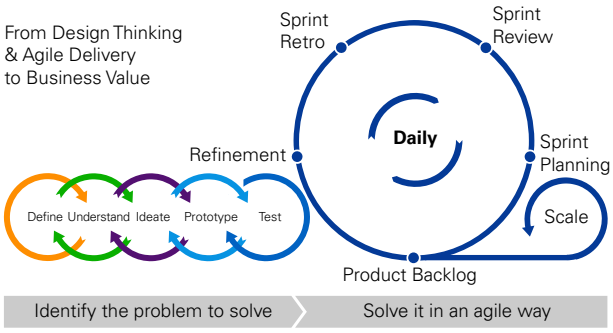


Fig 4. The convergent journey from a user-centric approach into an agile base delivery approach.

Start with acquiring real world, up-to-date data

The first request asked by the enterprises' data scientists after understanding the problems and needs of the users will be: "Provide us with the data. We want to try it out!" After fruitful discussions around the use cases and problems to solve, this can be a more problematic task than one might think. The following questions must be answered:

- Who owns the data?
- How to access the data?
- Does it contain PSI (Personal Sensitive Information)?
- Who can give you access to the data?
- In what format is the data structured?
- Is it clean or does it contain a lot of noise?
- Is the data source documented or does it need further explanation?

Most of these questions can be answered with proper Data Governance in place and this must be supported by tools and platforms.

One essential part of a DSML is a place where internal and external data can be stored:

- in a centralized, normalized and explorable form
- in a convenient accessible way
- according to data governance rules and legal restrictions
- by the data owners
- in near-real time
- under the supervision of the data stewards
- for the data scientists and business users

We call this place the Data Shop.

While the idea of offering productive data in a shop initially sounds frightening from a data governance point of view, one need to think about the current reality. Internal data is exported and copied via CSVs, separate database user accounts need to be given to data scientists etc. – getting the data wastes a lot of valuable capacity.

Explore and experiment early

As mentioned above, data scientists want and should get their hands on the data as quickly and conveniently as possible. This might be due to highly motivated people but for sure due to the element of uncertainty in using solutions that involve machine learning. Data scientists often are specialized in domain specific data or know scientific procedures to explore and work with data in a new way. For example, reading in research papers or trying ideas heard on conferences. And in almost all enterprise use cases both have never been tried with the exact same data set or with the exact same algorithms. Therefore: You only know if it works, if you have tried it out!

It is essential for all successful data science projects and use cases to decrease the time between the initial idea and the first experiment and make the results reproduceable and reusable.

A DSML needs to support the collaboration process of the company by:

- making the data sources searchable for permitted users
- storing the insights for productive use and enhancements in following projects and use cases
- creating re-useable building blocks from your experiments

While data indexing and searching capabilities can be achieved by already established database solutions or for example Apache Lucene based search engines, the latter two requirements demand new procedures and concepts.

From storing data to storing knowledge and indications

Most enterprises have established or are in the building process of a centralized Data Warehouse, however they are still struggling with harnessing the information to generate real knowledge. Remarkable insights generated by smart people from good data are often encapsulated in isolated solutions and departments. This knowledge often remains in this information silo because the story of how it was gleaned never gets told let alone documented.

If we assess the maturity of the data-storing capability of an organization applying data science and artificial intelligence, we do it in five levels as depicted in figure 5.

The KPMG Signals Repository[1] as an example of a an advanced data store, is a collection of sources (data) that is harnessed to interpret the impact of internal and external factors (signals) on a company's execution, to derive insights from the patterns (indications), and to accelerate and affect meaningful decision-making on a continuous basis.

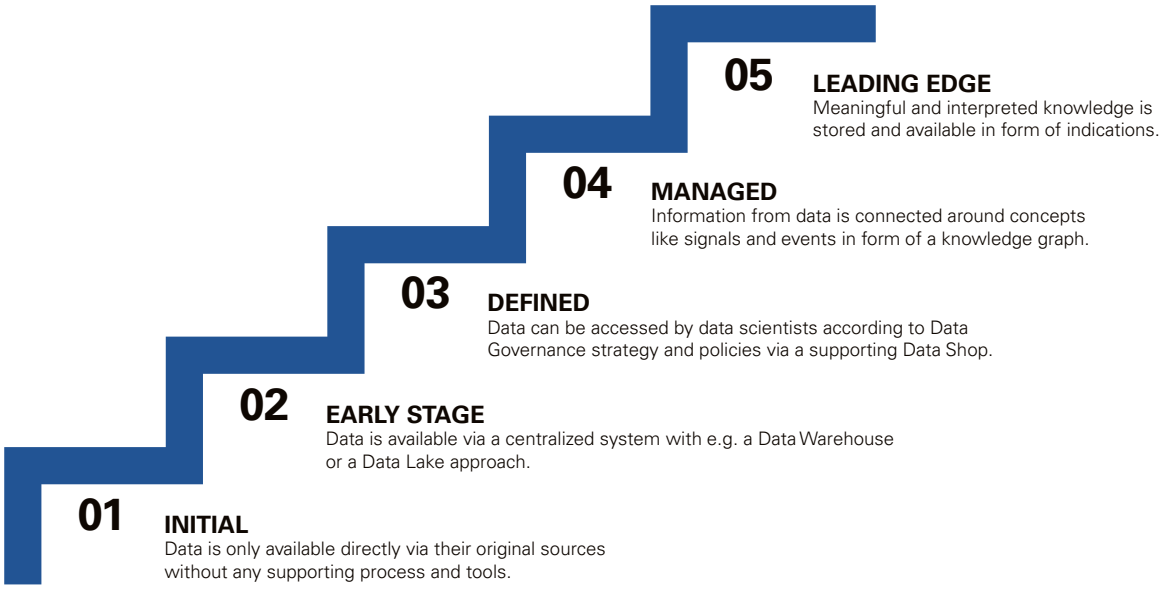


Fig 5. The five maturity levels of a data science supporting data store.

[1] For more information, please visit: https://advisory.kpmg.us/services/dataanalytics/ lighthouse/signals-repository.html

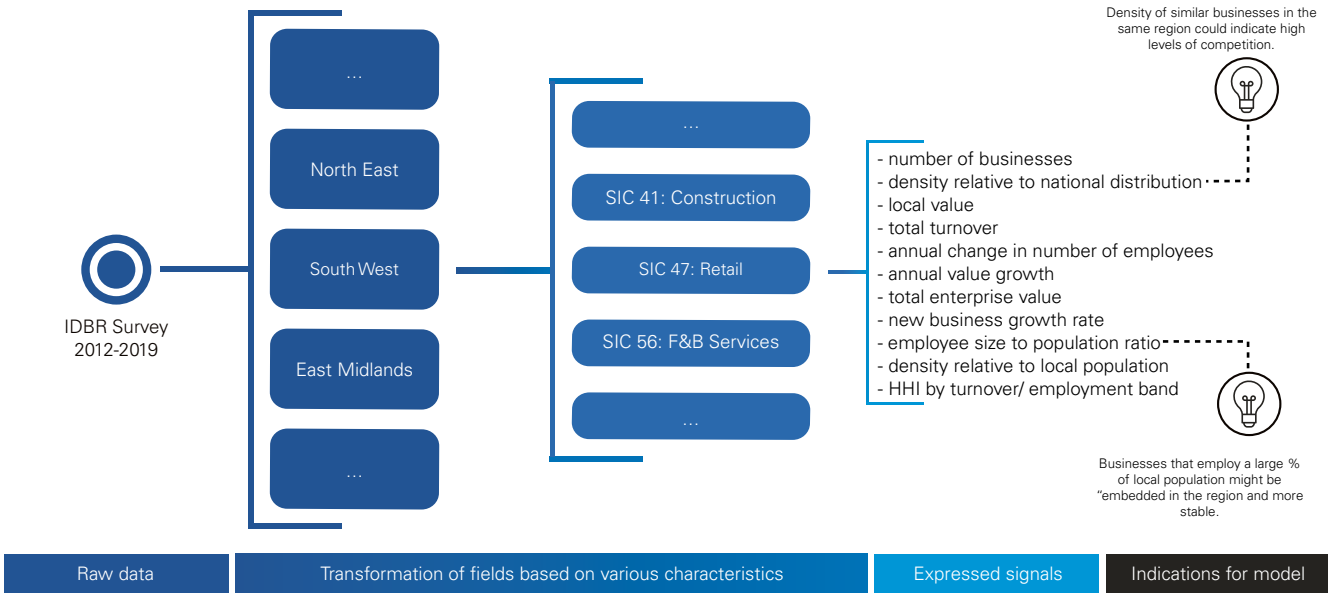


Fig 6. From raw data to meaningful indication.

Leveraging KPMG Signals Repository, it is easy to "listen" to the tens of thousands of signals around us and then use machine learning to make sense of it all. In figure 6 you can see examples of various signals which can be expressed by transforming the raw data contained in the UK Inter-Departmental Business Register (IDBR) Survey. Regions and the Standard Industrial Classification (SIC) codes could be used to extract signals such as the density of similar businesses in a region which could be an indication of high levels of good consumption.

To create such data stores, open source technologies like PostgreSQL can be easily adopted and maintained with the skill set of your data administrators due to their relational database roots. The additional advantage lies in the natural support of JSON data objects, common as communication backbone in any platform and the advanced support for full text searches. Even more suitable for natural language processing use cases is the usage of an Elasticsearch store. And to save real signals and context-based indications the implementation of a graph-based storage with the help of Neo4j or JanusGraph is inevitable.

State-of-the-art algorithms demand state-of-the-art processing runtimes

Another way of increasing the efficiency of your data science workers is by reducing their waiting time on training the models. While classic regression-based algorithms are in general quite fast on regular CPUs, more modern deep learning based algorithms work better on the parallel processing power of GPUs.

An experiment in integrating the relatively new BERT Model into a classic NLP classification task with nearly 3000 documents in 45 classes lead to the processing times shown in table 1.

	Computing Resource	Model Training Time
"Classic" ML Approach	8 Core CPU	1 min 30 sec
BERT Model – CPU	8 Core CPU	5 days
BERT Model – GPU	2 Tesla V100 GPUs	40 min

Tab 1: Experimental processing times of ML models on CPUs and GPUs.

This creates the need for giving your data scientists access to shared GPU runtimes over your platform.

Additionally, the research development in quantum computing should be monitored closely, as its capability of simulating multiple algorithmic states at the exact same time, will increase the speed of training deep learning models tremendously. However, it is unrealistic to have such a machine in your data center or even under your desk in the predictable future. Again, the solution lies in offering access to these runtimes on a shared basis.

A model repository is more than an artifact store

A model repository stores trained models and its related artifacts so that these models can be referenced and used by applications and services. Comparable to package repositories like Nexus or Artifactory. But a model repository needs to be more than this, because a model is not just a set of binaries, it also can contain huge amounts of data – depending on the used machine learning algorithm.

In addition, some models require – more or less – large datasets like language models to function. The volume of this data is too large to be stored together with the remaining artifacts in a package manager.

Thus, a full-blown model repository requires a store for such data and capabilities to integrate (authorized) data download from the scoring code during runtime – actually like some well-known libraries like NLTK or spaCy do. This functionality might be offered by a Data Repository or by some plain Object Storage which is part of the DSML platform.

Data-driven solutions often require big data, fast streaming integration applications

The integration of machine learning models into processes and systems often requires processing huge volumes of data or processing data in near-real-time. Technologies like Apache Kafka, Apache Spark or Apache Flink are a good foundation to build such integration solutions.

These technologies can be combined with Reactive Streams toolkits such as Akka Streams and connector frameworks like Kafka Connect or Alpakka to build resilient, scalable integration solutions. For application management and monitoring it is inevitable to integrate measures such es throughput, latency, memory and CPU usage into a monitoring and alerting solution. To enable fast development of such applications we suggest providing a customized toolkit like Alpakkeer which includes required building blocks for stream-processing, error handling, scheduling, monitoring and alerting – adjusted to the target enterprise landscape. Such a toolkit can hide complexities of mature big data, fast streaming applications to developers.

Continuously improve your models with live data and feedback

Automatic model retraining is required if a model should continuously be re-trained on regular basis to involve new or updated data. E.g. a model which is trained based on current housing prices which are published yearly by local government. Ideally the whole process from fetching the updated data, storing the data in a Data Repository, re-training the model, model validation, model publishing to the model repository and finally model deployment should run completely without human interaction.

Human interaction would only be required if the validation process identified issues with the model quality.

Such process can be realized with tools like Kubeflow or Apache Airflow. Ideally the platform integrates one of these tools automatically based on a definition which is stored in the model's repository. Applications which have a dependency to the model can act upon the update depending on their needs or re-deploy a new version with an updated dependency to the latest model version, realized with traditional CI/CD capabilities.

To close the data science life cycle, you start over again, because like your employees, also your AI Infused solutions should never stop learning from your experts. To enable this, you need to always include a process to re-introduce feedback data into your automated training cycle. Your solution can actively query your users for new labels, which is then called Active Learning.

There is not (yet) a single best Open Source tool to implement Active Learning into your AI solutions. It is a combined effort:

- between your data scientists on how the chosen algorithm can profit from new labeled data
- between your data engineers on how to feed this data back into the training process
- between the product owners on how labelling can be integrated smoothly into the regular usage of the solution

III. How to move on

The previous sections outlined how a DSML platform supports activities to implement and run data-driven solutions. We also named various Open Source technologies which are a good basis for implementing a platform.

To get more details and move on with your journey, we suggest the following steps:

- Visualize a day in the life of your data scientists, envision how it should look like to fasten the development of solutions
- Review your plans to align platform strategy with AI strategy – develop both with clear use-case focus
- Assess the maturity of your processes and toolsets

Contact us to learn more

To get more details on how Open Source components can be used to implement a DSML which supports the whole Data Science Lifecycle, request additional material. Find out how KPMG can help your organization on your journey to uncover the full potential of your data.

Why KPMG?

- Extensive know-how in the development of solutions in the digital transformation area.
- Profound business expertise.
- Technological knowledge to implement end-to-end solutions.
- Support in development of data-driven solutions from ideation to implementation.
- Strong capabilities in Tech Risk Auditing and Data Management.
- Our team consists of experts in Digital Experience, Data Science and Solution Architecture.
- Extensive capabilities to create strategy, envision architectures and implement platforms, including trainings and support.
- Many client engagements in the financial services industry that involve building successful Open Source DSML platforms.

Contacts

KPMG AG
Räffelstrasse 28
PO Box
8036 Zurich

Bobby Zarkov
Partner, Financial Services
Digital Innovation

+41 79 402 14 89
bzarkov@kpmg.com

Thierry Kellerhals
Director, Financial Services
Digital Innovation

+41 79 281 22 50
tkellerhals@kpmg.com

Johannes Forster
Senior Manager, Financial Services
Digital Innovation

+41 76 345 84 18
johannesforster@kpmg.com

Michael Wellner
Senior Manager, Financial Services
Digital Innovation

+41 76 804 68 84
michaelwellner@kpmg.com

kpmg.ch/fs

The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavor to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received, or that it will continue to be accurate in the future. No one should act on such information without appropriate professional advice after a thorough examination of the particular situation. The scope of any potential collaboration with audit clients is defined by regulatory requirements governing auditor independence. If you would like to know more about how KPMG AG processes personal data, please read our Privacy Policy, which you can find on our homepage at www.kpmg.ch.

© 2021 KPMG AG, a Swiss corporation, is a subsidiary of KPMG Holding AG, which is a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Data Science & Machine Learning Platforms January 2021