CHAPTER 1

Data Science in the Modern Enterprise

Data science is the hottest trend in IT, and it is not showing signs of cooling down anytime soon. It is often used as a catchall phrase for all latest innovation, such as machine learning (ML), artificial intelligence (AI), and Internet of Things (IoT). This is not an inaccurate representation since data science is after all the foundation for ML, AI, and IoT.

Data science is also responsible for creating new processes and methodologies that impact how businesses should be operated today. As a result, data science has led to the spawning of initiatives such as digital transformation and data-driven decision making. Such initiatives change the way organizations solve problems, budget their IT investments, and change the way they interact with customers and citizens.

However, the true definition of data science is that it is a field of study that uses scientific methods, processes, algorithms, and software to extract knowledge and insights from any type of data (structured, unstructured, or semi-structured). Data science is a combination of statistics and computer science applied to relevant domain knowledge of data in order to predict outcomes. And as the saying goes, the enterprise that can most accurately predict the outcomes will have the utmost advantage over its competitors.
Mindset of the modern enterprise

The modern enterprise is one that can quickly respond to the changing environment. Any organization that has an initiative to leverage the use of data it owns and expand on the (ethical) collection of data it does not have, but would be valuable, is considered a modern enterprise. Most organizations recognize that this is an important transformation they must undertake. The digital transformation we are seeing today involves organizations making data the center of their decision-making processes.

It is useful to evaluate the maturity of an organization as it relates to data. If an organization is spending most of its resources addressing tactical issues instead of being strategic, then the organization is lower on the data maturity model.

Commercial entities

Commercial entities are the earliest adopters of data science. For example, retail and customer data allow commercial entities to mine the information to improve sales. Customer sentiment, taste, and even political and social media data can be used as enriching features to help drive sales strategies.

Using manufacturing as another example, data science is used to optimize manufacturing processes by identifying anomalies and defects, thereby reducing or eliminating factors leading to manufacturing defects.

In healthcare, data science is being used to study cell behavior and diseases to develop cures. During the COVID-19 pandemic, researchers also used data science to analyze social media data in order to identify symptoms in patients, which would enable medical professionals to better diagnose patients who may have the virus.

Reference: https://medicalxpress.com/news/2020-08-social-media-covid-long-haulers.html
Government entities

Even though public sector entities are traditionally more conservative and will generally lag the commercial sector in adopting modern technologies, data science is one area where the gap is much narrower, at least in the United States and most developed countries.

This is because government organizations traditionally collect a lot of data, and the data is extremely diverse in nature. Different departments in every state have information of its citizens through all the services the state provides. Multiply that by 50 states plus the information collected by the federal government, that is a lot of data to mine and a lot of knowledge and insight that can be gleaned through data science.

Consumer and personal

The consumer sector is probably the most prolific when it comes to data growth, thanks to IoT devices associated with connected homes, wearable technology, and social media. All these technologies fall under the umbrella of Big Data, which we will discuss in detail in Chapter 6. For now, we just want you to appreciate the vast source of rich data that is being collected on a daily basis that is growing at an exponential rate.

One important aspect to also recognize is the blurring of the lines between all the different domains mentioned earlier. For example, with the use of personal assistants (e.g., Siri and Alexa), information and data from each domain are now more blended than ever as we rely on personal assistants to carry out tasks.
Ethics in data science

As seen, data proliferation is occurring in every aspect of our lives – at work, where we shop, what we buy, how we interact with others, our political and social views, our governments, and our homes.

This type of digital revolution happens only once in a while, the last being equivalent to the birth of the Internet, and the birth of personal computing before that. As such, new social and ethical challenges arise, and we must acknowledge and address them accordingly.

Ethics in data science is more commonly known as ethical AI. But before we look at ethical AI, let us first see how some things that were once considered science fiction are no longer fiction. We can get an idea on the ethical challenges based on story lines in science fiction.

Science fiction and reality – a social convergence

Hollywood science fiction movies have played their part in introducing intriguing effects of AI on society in movies like *AI Artificial Intelligence* (2001) and *Minority Report* (2002).

Both movies touch on extreme themes. In *AI Artificial Intelligence*, audiences are left to ponder whether an artificial life form is worthy of the same protections as humans when an artificial child is built to possess emotions. In *Minority Report*, predictions made by AI are enforceable before the actual crime is committed.

As AI makes robots and digital assistants more human, the line between fiction and reality is indeed getting blurred. For example, contributing to Japan's issue of aging population is the generation of
otakus,² men who are in relationships with virtual girlfriends – a concept highly promoted in the anime culture and gaming.

Machine learning, a subset of AI, uses algorithms to build mathematical models based on sample data for training and learning, with the goal of making the most accurate predictions and decisions possible. How confident are we with machine-based decisions? Look around you at the Teslas, IoT devices, and smart homes. The US Navy and Boeing have a highly classified program (code-named CLAWS) to develop AI-powered Orca class fully autonomous submarines that are outfitted with torpedoes, meaning these submarines may be able to make the decision to kill indiscriminately. The obvious ethical question: Should a machine be able to kill without human intervention? The increase in concern regarding the need, or potentially lack thereof, of human intervention in AI decision making has led to the alternative and growing popularity of not calling the acronym AI Artificial Intelligence, but rather Augmented Intelligence.

---

**Note** Another sign of the times is the explosion of fake news. Advancements in AI and machine learning are making it easier to create extremely convincing fake videos, known as deepfake. Experts fear that deepfake videos will create an even larger trust issue in journalism and these videos are also being employed in politics.

---

²In 2012, a 35-year-old otaku, who is a school administrator, married his virtual girlfriend named Hatsune Miku. [www.techspot.com/news/77385-japanese-man-marries-anime-hologram-hatsune-miku.html](http://www.techspot.com/news/77385-japanese-man-marries-anime-hologram-hatsune-miku.html)
**What is ethical AI?**

Ethical AI is a study and governance surrounding the use of AI technologies in a way that does not violate human rights and endanger lives. We are not exactly at the point of having to address the rights of robots, so that part is still the domain of science fiction.

The impact of AI on life is so important, and as we rapidly meet and surpass human parity in many areas, religious organizations have played a leading role to proactively address this issue. In February 2020, the Vatican in Rome released a guideline for ethical AI, called “Rome Call for AI Ethics,” to which Microsoft is one of the first to sign. Listen to Microsoft President Brad Smith’s speech at [https://romecall.org/](https://romecall.org/).

There are many discussions and representations of what ethical AI means, but the three principles in Rome Call for AI Ethics very succinctly summarize the core concepts – ethics, education, and rights.

Other organizations involved in the design and development of AI are also incorporating the important aspect of ethics. DeepMind is another example of an organization that is not only comprised of scientists, engineers, and researchers in the field of AI but also brings together philosophers and policy experts to form truly interdisciplinary teams to address the broad impact that AI has on society. It is indisputable that there is a need for accountability for the development and use of AI, perhaps more so than any emerging technology in recent history. With backing from large corporations like Google, organizations like DeepMind provides a forum to engage citizens, governments, and corporations to agree on fundamental concepts involving ethics. Find out more at [https://deepmind.com/about/ethics-and-society](https://deepmind.com/about/ethics-and-society).
**Note** “AI-based technology must never be used to exploit people in any way, especially those who are most vulnerable. Instead, it must be used to help people develop their abilities (empowerment/enablement) and to support the planet.”

—The Vatican

### Microsoft AI principles

Microsoft has a site dedicated to the company’s AI principles at [www.microsoft.com/en-us/ai/responsible-ai](http://www.microsoft.com/en-us/ai/responsible-ai) and is a leader in the charge to ensure ethics are deeply engrained in the development of AI technologies. The most important aspect of Microsoft’s AI principles is that it applies to AI at scale since the technology is built into every service and shared across its entire portfolio. For example, threat patterns gleaned by AI in the Xbox service are shared across other services like Office 365, Azure, and Dynamics in order to better protect the entire cloud services portfolio. Even though Microsoft has an Office of Responsible AI tasked with putting the company’s principles into practice, common ethical AI design principles are followed across all groups in the organization, from engineering to marketing.

**Note** “An important hallmark of our approach to responsible AI is having [the] ecosystem to operationalize responsible AI across the company, rather than a single organization or individual leading this work.”

—Microsoft
Microsoft’s AI principles revolve around the six key concepts:

- **Fairness**: AI systems must treat all people fairly and with no bias to age, culture, gender, or national origin.
- **Inclusiveness**: AI systems must engage and empower everyone, with no bias to age, culture, gender, or national origin.
- **Reliability and safety**: AI systems should perform reliably and safely.
- **Transparency**: AI systems should be understandable.
- **Privacy and security**: AI systems should be secure and protect and respect privacy.
- **Accountability**: People should be accountable for AI systems.

We dedicated a generous amount of time discussing ethical AI because it is a very important one as the technology evolves. For the rest of this chapter, we will introduce the different AI technologies from Microsoft and look at some use cases. However, where applicable, we will bring up specific ethical and legal considerations as it relates to the services.

**Azure Machine Learning**

Machine learning is a subset of AI and is the other topic we will cover in detail in Chapters 4 and 5. The advancement in all the cognitive services mentioned in the previous section is based on machine learning and the vast amount of training data that are available.

Machine learning is used to train models to not only recognize or detect but also to predict outcomes and thus help drive decision making. Machine learning is fundamentally based on data science and statistical
methods. The amount of data and statistical methods often requires a lot of memory and compute, which is an ideal fit for Azure.

In this section, we will cover the options available in Azure that help data scientists and researchers do their jobs. There are primarily two options – the initial implementation known as Azure Machine Learning Studio (classic) and the more recently released Azure Machine Learning. Figure 1-1 summarizes the differences between the two offerings.

| Feature                        | ML Studio (classic) | Azure Machine Learning |
|--------------------------------|---------------------|------------------------|
| Drag and drop interface       | Classic experience  | Updated experience - Azure Machine Learning designer (preview) (Requires Enterprise workspace) |
| Code SDKs                      | Unsupported         | Fully integrated with Azure Machine Learning Python and R SDKs |
| Experiment                     | Scalable (10-GB training data limit) | Scale with compute target |
| Training compute targets       | Proprietary compute target, CPU support only | Wide range of customizable training compute targets, includes GPU and CPU support |
| Deployment compute targets     | Proprietary web service format, not customizable | Wide range of customizable deployment compute targets, includes GPU and CPU support |
| ML Pipeline                    | Not supported       | Build flexible, modular pipelines to automate workflows |
| MLOps                          | Basic model management and deployment | Entity versioning (model, data, workflows), workflow automation, integration with CED tooling, and more |
| Model format                   | Proprietary format, Studio (classic) only | Multiple supported formats depending on training job type |
| Automated model training and hyperparameter tuning | Not supported | Supported, Code-first and no-code options |
| Data drift detection           | Not supported       | Supported |
| Data labeling projects         | Not supported       | Supported |

**Figure 1-1.** Comparing Azure Machine Learning and Azure Machine Learning Studio (classic). Source = https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-machine-learning-studio
## Azure Machine Learning

Azure Machine Learning is a cloud-based environment for data scientists to train, deploy, automate, manage, and track ML models. It is often used in conjunction with cloud storage such as Azure Blob Storage and Azure Data Lake Storage because of the potentially large amount of data being used in the ML process.

Azure Machine Learning supports all the tools and languages used by data scientists, such as R, Python, and Scala. It also has tools such as Jupyter notebooks and supports open source add-ons such as PyTorch, TensorFlow, scikit-learn, and YOLO. It comes with a web interface called the Azure Machine Learning designer.

Azure Machine Learning supports classical ML, deep learning, as well as supervised and unsupervised learning. Azure Machine Learning also integrates with other Azure services, such as Azure DevOps, to help secure the data scientists’ work through source control.

## Machine Learning Studio (classic)

Azure Machine Learning Studio (classic) is the initial implementation and ML offering in Azure. It is still available and should not be confused with Azure Machine Learning. See the differences in Figure 1-1 again. Download the Azure Machine Learning Studio (classic) architecture to see all its capabilities at [https://download.microsoft.com/download/C/4/6/C4606116-522F-428A-BE04-B6D3213E9E52/ml_studio_overview_v1.1.pdf](https://download.microsoft.com/download/C/4/6/C4606116-522F-428A-BE04-B6D3213E9E52/ml_studio_overview_v1.1.pdf).

## Azure Databricks (Chapters 6 and 7)

Azure Databricks is the native implementation of Databricks in Azure and is provided as a PaaS offering. Azure Databricks provides all the tools and resources that data scientists and researchers need and is jointly managed by Databricks and Microsoft engineers.
Azure Databricks are provisioned as workspaces that contain the customizable compute clusters, the notebooks, datasets, and storage. Users can use Databricks File System (DBFS) or mount external storage like Azure Blob Storage or Azure Data Lake Storage to access data for projects. Azure Databricks notebooks allow users with different skill sets to collaborate because the notebooks support different languages like Python, R, Scala, and Spark SQL. All notebooks can be source controlled using GitHub or Azure DevOps.

To optimize resources and reduce cost, compute clusters in Azure Databricks can be automatically shut down due to inactivity. The default is 120 minutes, but this can be configured as needed. Compute clusters can be customized with the desired libraries needed, such as PyTorch, scikit-learn, and TensorFlow. Compute clusters can be pinned so the customization will be preserved even after the cluster is shut down due to inactivity.

Use cases for Azure Databricks

Azure Databricks is often used to prepare raw data for downstream systems and analysis. It is usually front ended by low-latency ingestion solutions such as IoT hubs or event hubs and other types of storage serving non-streaming data. Figure 1-2 depicts a typical architecture of an enterprise data pipeline architecture with data inputs from different sources (streaming and non-streaming) with Azure Databricks as the data preparation environment before sending the prepared data to downstream datastores.
Figure 1-2. Data ingestion and data manipulation pipeline leveraging Azure Databricks. (Source = https://cloudarchitected.com/2019/03/event-based-analytical-data-processing-with-azure-databricks/)

You can also read up on two other use cases located at

- https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/ai/batch-scoring-databricks
- https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/data/stream-processing-databricks
Summary

In this chapter, we made the case as to why data science is an inevitable undertaking that all modern organizations need to address in order to stay competitive. We also spent some time in understanding new ethical challenges because of AI and ML, which are a result of advancements in data science. Lastly, we provided a quick overview of the modern technologies and tools in Azure that will help data engineers and data scientists conduct their work in a more efficient manner.