CHAPTER 9

Case Studies by Industry

In this chapter, we will finally look at the most important aspect of data science: domain knowledge. You can think of this chapter as providing a kind of ladder, as you won't be a domain expert reading this chapter alone but maybe you will be able to get to where you want to go by asking the right questions. After all, data science is not about technology or code, but it's about the data and, more specifically, the domain knowledge and concepts that data represents. Each industry has unique problems that may not be well understood outside of that industry.

So what is data? Data in some sense is more general than even numbers since it can be both quantitative and qualitative. Numerical data is a type of data, while categorical or natural language-based data represent raw concepts. Domain and industry knowledge is really the "soul" behind the data, the elusive part of any data science project that gives the data meaning.

While many modeling problems are considered solved especially in supervised machine learning where basic classification and regression problems can be repurposed over and over again possibly with very little domain knowledge, to actually drive performance metrics and solve novel problems that will bring a competitive advantage to your industry and even to be able to identify which problems are actually "hard" will require domain knowledge.

D. Sorvisto, *MLOps Lifecycle Toolkit*, https://doi.org/10.1007/978-1-4842-9642-4_9

This is the last chapter in the book because unlike previous chapters, domain knowledge is not easy to learn; it has to be earned through years of experience. Mathematics can be learned and technology can be learned to some degree, but domain knowledge is purely experience based, knowing what to measure, how to measure it, what is noise, what features to throw away and what to keep, how to sample the data to avoid bias, how to treat missing values, how to compute features (code that encodes all of this business logic to eliminate bias in data can get very complex), and knowing what algorithms are currently used and why for the particular domain are all something that needs to be learned from experience. It also changes, and unless you're working in a domain, it can be hard to even get an understanding of what problems are important and what models are considered solved.

It's the goal of this chapter to discuss some of the problems across different industries and look at ways in which MLOps can improve the lives of domain experts in those areas or to provide some further information for data scientists that have experience in one industry and are looking to transfer that knowledge to another industry. We'll also take a look at how we can use this knowledge, store and share it across industries, and leverage it for strategic advantage for our organization by using the MLOps lifecycle.

Causal Data Science and Confounding Variables

One of the things that makes data science difficult is confounding variables and illustrates why we need to have domain knowledge to truly understand the dependence structure in our data. Without a solid understanding of our data, we won't be able to identify confounding variables, and we may introduce spurious correlations into our results, and our models won't be an accurate representation of reality.

What is a confounding variable? A confounding variable is a third variable that influences both the independent and dependent variables in a model. This is a causal concept meaning we can't just use correlations to identify confounders, but we need a solid understanding of the data domain and the causal factors that underlie the model; after all, correlation does not imply causation. A visual representation of a confounding variable is shown in Figure 9-1.

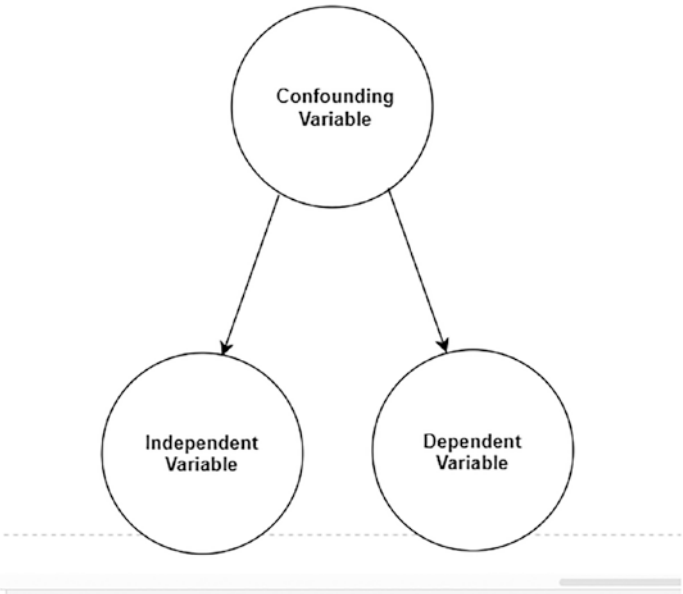

Figure 9-1. *A confounding variable influences both the independent and dependent variables*

For example, you might look at the impact of a variable like amount of alcohol consumed on a daily basis on mortality rate. However, there are many confounding variables like age that may be considered confounding variables since age can have an effect on both alcohol consumption and mortality rate. Another good example is the placebo effect where you might have a randomized experiment with two groups but one group was administered a placebo. Since the belief that the treatment is effective may

influence the outcome of the treatment (e.g., believing a treatment will give you more energy may cause you to feel less tired), the experiment needs to be controlled to account for the placebo effect by splitting participants into two randomized groups.

So how can we identify confounding variables? There is no known algorithm for learning all of the cause and effect relationships and identifying confounding variables (although causal data science is an active area of research with promising techniques like Bayesian networks and counterfactual inference). We can't rely on correlation to help us here since correlation is not causation and you need to develop a mental model of the cause and effect relationships you're studying to truly avoid spurious results.

This understanding of cause and effect can only come from domain experience and is the main motivation for understanding the domain we're modeling. In the next section, we'll take a bottom-up approach and look at domain specific problems broken down by industry from energy, finance, manufacturing, healthcare, and more and try to get a better understanding of each industry's problem domains.

Energy Industry

One use of data science in the energy industry is in upstream oil and gas. Geostatistics, which takes into account spatial dependencies in data (data points that are close together on the Earth's surface are assumed to be similar), leads to important applications like Kriging and geospatial sampling.

Collecting data on well reservoirs is costly, and having techniques that can infer unknown data without directly measuring it is an important application. Another area is in midstream oil and gas, where energy needs to be physically transpired in pipelines or, in the case of utilities, physically transported over a distribution network. How do we detect anomalies and make this process more efficient?

Safety is also critical in this area, and using data science to identify leakages and other anomalies to reduce outages or, in the case of oil and gas, to prevent shutdowns is a focus of a lot of modeling applications.

Manufacturing

Manufacturing is becoming increasingly data-driven as organizations recognize the potential of data analytics to optimize their operations and increase operational efficiency. An increase in operational efficiency of just 0.1% can translate into large cost savings in absolute terms since that 0.1% is relative to EBITDA or operating income.

The use of data science and MLOps can help manufacturing executives gain insight into production processes, reduce waste, develop lean processes, improve quality control, and forecast equipment and component failures before they happen.

A concrete example is predictive maintenance or forecasting the time to failure or similar variable from sensor data. Sensor data may be from entire fleets of equipment or production equipment from manufacturers and can help predict which components are likely to fail. This may aid in scheduling maintenance, reducing downtime, and increasing output.

Statistical quality control is another area of operational research where data science can lead to innovative improvements. Identifying trends in product defects can help manufacturers analyze root causes of defects and adjust their processes to reduce future defects or to adjust inventory levels and lead times on the fly.

Transportation

Transportation and manufacturing have many overlaps in terms of data science use cases. Transportation itself is a huge industry that is best broken down into different subindustries that include railways, shipping,

aviation, and more. Striking the right balance between safety and efficiency is one of the drivers of using data science in transportation, and again predictive maintenance has many applications.

Edge devices that may be attached to entire fleets of vehicles emit various sensor readings like pressure and temperature (we looked at an example of this in our feature engineering lab) and can be analyzed to forecast time to failure and improve scheduling efficiency. It's important to note here that these sensor data sets are massive data sets especially if they come from entire fleets of vehicles and may include real-time data, so MLOps play a critical role in transportation data science especially in scheduling off-peak hours and minimizing scheduling conflicts and complex route planning.

For example, how do you determine the best routes to take when your loss function includes information on fuel consumption and other transportation costs and you have to minimize this loss over a massive data set of sensor readings? To make matters more complicated, the fleet of vehicles may extend across broad geographical boundaries and include different units that need to be normalized, and all of this data has to be processed in a way that takes into account operational safety as well as efficiency. Safety data itself can be multimodal coming from traffic cameras, sensors, and other sources to identify areas with high incidence rate.

Retail

Retail is an industry that has been completely transformed by data science in the last 20 years, from recommender systems and customer segmentation to pricing optimization and demand forecasting for new productions.

Customer segmentation (looking at the customer along dimensions like geography, psychology, demographic, buying patterns, purchasing preferences, and other traits) helps to personalize messages and product offerings and can be used in combination with a recommender system.

Demand forecasting can help retailers analyze historical sales and transactional data to study variables such as weather, promotions, holidays, and macroeconomic data to predict demand for new productions and decide where to allocate resources and marketing efforts. Demand forecasting is primarily used for reducing inventory levels and increasing sales by anticipating spikes in sales volumes. Demand forecasting can also be applied to optimize supply chains by adapting to spikes and decreases in demand.

Related to sales, price optimization is a classical use case for retail data science. Data science can help optimize pricing strategies based on current market trends like inflation and interest rates. Customer demand forecasts can also be fed into these models with competitor pricing to maximize profit margins year over year. Developing pricing strategy is necessary in competitive markets like retail where pricing may be a key differentiator of the product.

Recommender systems can be created by querying models to recommend new products and services to customer segments based on past purchases (when available) or other data like browsing history for online retailers.

Agritech

Agricultural communities developed over 10,000 years ago, so it's not every day agriculture gets a major overhaul, but data science has tremendous potential in the intersection of agriculture and technology called agritech to improve agricultural processes and increase yields and overall efficiency in precision farming.

Precision farming in particular uses data science to collect data from sensors, drones, and other sources to optimize crop yields and reduce use of fertilizers and pesticides that can harm the environment and reduce yield. The main factors that influence crop yield include weather, agricultural land, water, and harvest frequency, and data collected from sensors can be used to maximize yield.

Finance Industry

Finance is one of the most interesting industries for data science applications. Fraud detection (a kind of anomaly detection problem) seeks to detect fraudulent transactions and prevent financial crime. Fraud detection is possible in part because of our ability to process vast amounts of data and to measure a baseline behavior in the transactions to detect patterns of fraud even if they're less than 1% of the entire sample.

Risk management is another area where data scientists build predictive models, to quantify risk and the likelihood and frequency of occurrence. Predicting which customers might default on a loan, for instance, is an important problem in predictive modeling. MLOps can help to streamline risk management problems by bringing transparency and explainability into the modeling process, introducing mathematical methods like SHAP or LIME to report on which attributes went into a particular loan decision. Model explainability and fairness are particularly important in credit risk scoring where demographic features (income, geographic location), payment history, and other personal information are fed into the model in hope of getting a more accurate picture of someone's credit risk at the current time.

Metrics like net promoter score and customer lifetime value are frequently used in modeling. Industry standards are extremely important in the finance industry especially when it comes to data science. Risk modeling, for instance, is important because if we can calculate risk of

churn or risk of default, we can concentrate resources around preventing those customers from churning provided they have sufficient customer lifetime value.

However, how you approach the risk model in finance may be different than other industries. Continuous features are often discretized, meaning the features are placed into buckets. One of the reasons for this is so we can create a scorecard at the end since there are often laws and regulations around reporting credit risk and the models need to be interpretable by someone without advanced knowledge of the model. There's also the assumption of monotonicity with credit risk. This is difficult to include in some models.

Healthcare Industry

We can look at applications of data science in healthcare and predict where MLOps can impact healthcare. One area that is very active is in medical image analysis and, more generally, preventative medicine.

Preventative medicine uses X-rays, CT scans, MRI scans, and healthcare data to detect abnormalities and diagnose disease and malities faster than a human doctor or even a traditional lab test could. Imagine you could diagnose disease years in advance and treat them before they become a problem that threatens the health of the patient.

While X-rays, CT scans, and other types of medical imaging would require computer vision models such as convolutional neural networks, preventative medicine may also look at the entire history of the patient to summarize it for medical professionals (e.g., autoencoders or topic analysis algorithms) require natural language processing and domain knowledge of medicine. These will be vast data sets and require infrastructure to support big data as well as require security and data privacy safeguards to protect patient data (e.g., this data may be regulated by HIPAA or similar regulation).

Predictive analytics can ultimately be used to reduce hospital readmissions and reduce patient risk factors over the long term, but moving these metrics requires reporting, monitoring, and ability to feed patient outcomes back into the model for retraining.

Two emerging areas of research in healthcare are drug discovery and clinical decision support systems. Causal inference can be applied to discover new combinations of drugs and build new treatments or even speed up the clinical trials or augment data sources that are too expensive to collect.

Monitoring infrastructure can be set up to monitor patients in real time and provide healthcare practitioners with real-time data on patients that can be used to make better healthcare decisions resulting in better patient outcomes and reduced hospital visits. The potential to increase the efficiency and optimize resource allocation in the healthcare space will be one of the most important applications of MLOps in the twenty-first century.

Insurance Industry

If you're a data scientist in this field, then there's a lot of opportunity for innovation. One unique example is preventative maintenance. We might not think of preventative maintenance having applications in something like insurance, but what if insurance companies could use sensor data to predict when vehicles need to be maintained, preventing breakdowns before they happen and keeping claims at a minimum. This would benefit both the drive of the vehicle and the insurance company.

Data science is playing an increasingly vital role within the insurance industry, enabling insurers to make more accurate assessments of risk and personalize policies. Most people know there are large volumes of customer data available to predict the likelihood of claims behind made, for example, insurers could use customer demographic data, credit scores, and historical claim data and develop personalized risk strategy models.

Fraudulent claims are expensive to insurers and lead to bottlenecks and inefficiencies in the process as insurers seek to eliminate fake claims with strict policy rules and procedures for underwriting. However, analyzing patterns in customer data, we can use anomaly detection to identify fraudulent claims without the additional cost.

Customer experience is another area within the insurance industry that could use some improvements. Although we don't usually think of insurance companies as being in the customer service industry, with increasing competition in this space, using data science to fine-tune policy to customer needs would lead to new business opportunities. All of these types of models require big data, and MLOps can make the insurance industry much more operationally efficient, to automate the underwriting process and make personalized policy recommendations based on customer risk profiles and other factors.

Product Data Science

Each of the industries mentioned continue to be disrupted by innovative technology companies that are increasingly becoming data and analytics companies that leverage data to improve traditional business processes. A great example is in healthcare where machine learning is being applied to preventative medicine to diagnose disease and infections before they become advanced or untreatable. By developing new diagnostic techniques with machine learning, countless lives can be saved.

Another area ripe for disruption is in the financial industry where customers that would not traditionally qualify for a loan may be considered because there's data available to evaluate the risk of default.

While product data science is different in the sense that you need to understand the product end to end rather than building a model to make an existing process more efficient, your model needs to have product-market fit. Understanding the customer or end user of the group across

various dimensions such as demographic, psychological, behavioral, and geographic data sets can help to segment customers and provide insight into what kind of model may best meet the needs of each customer segment.

Customer segmentation may be an invaluable approach to product data scientists and, also, the ability to ask questions to establish and uncover novel ways of modeling a problem since, unlike in industry, the modeling problem itself may not be a solved problem. This is why advanced knowledge of statistics and experience with research are required to be an effective product data scientist.

Research and Development Data Science: The Last Frontier

Data science at the edge is a rugged landscape, a mixture of many different disciplines that are constantly evolving. In fact, some industries may not even be invented yet. You might wonder how data science might look in 50 years. While predicting something like how data science will evolve 50 years out is clearly not possible, if we want to predict how we might better position ourselves to understand the massive amount of change in this field, we might want to look at research and development and the kind of impact data science has had on scientific research and business innovation.

In particular, we can look at areas from applied research to commercialization to new lines of business in various industries. Data science has increasingly become important in science and engineering, and although we can't predict the future, we can look at fields like genomics, neuroscience, environmental science, physics, mathematics, and biomedical research to gain an understanding of some global trends. We summarize these trends in the following.

- *Genomics:* Data science is used to analyze genomic and proteomic data to identify patterns, sequences, and mutations in genes and proteins. Deep learning systems like AlphaFold can accurately predict 3D models of protein structures and are accelerating research in this area.
- *Neuroscience:* Data science is increasingly being combined with brain imaging such as fMRI and EEG to unlock structure and function in the brain and provide new treatments for brain diseases.
- *Environmental science:* Data science is being used to analyze climate data, satellite imagery, and oceanographic and seismic data to understand how human activities impact our environment and to create new climate adaption technologies.
- *Physics:* Data science is used in physics to analyze big data sets and identify complex patterns in data from particle accelerators, telescopes, and scans of the universe. This information can be used to find new star systems and planets (data-driven astronomy) and to even develop new theories and models of the universe.
- *Mathematics:* Most of the focus on large language models has been on training these models to understand natural language and not formal languages like mathematics. While AI may not replace mathematicians completely, generative AI may be used to generate proofs, while formal verification may be used to validate these proofs. Building a system that uses both generative AI and formal verification systems

like automated theorem provers as components will lead to groundbreaking results and the first proofs completely generated by AI mathematicians.

- *Biomedical research:* Data science is used to analyze clinical trial data, biomedical data, and data from drug trials to develop new treatments and interventions for debilitating diseases. Causal data science is an emerging area within data science that has tremendous potential to expedite biomedical research.

Although we can list many active areas of research where data science has had an impact or will have an impact in the future, this relationship goes in both directions. While branches of mathematics like statistics and linear algebra have had the biggest impact on data science so far, other areas of mathematics like topology continue to find its way into mainstream data science through manifold learning techniques like t-SNE belonging to the emerging field of topological data analysis.

Other areas of mathematics are slowly making their way into data science, and people find new ways to apply old mathematical techniques to data processing. One interesting area is in algebraic data analysis where age-old techniques like Fourier transforms and wavelets are being used to change the way we analyze data. I mentioned this in Chapter 2, but if you take the Fourier transform of a probability distribution, you get the characteristic function of that distribution. Characteristic functions are a kind of algebraic object, and they've been applied in many proofs in mathematical statistics like proof of the central limit theorem. While other applications of Fourier transforms like wavelet signal processing are being used in some areas of data science, there are many mathematical techniques that will eventually find their way into mainstream data science.

In the next few sections, we'll pivot back to more concrete use cases of data in industry and how you can apply them in your own organization.

Building a Data Moat for Your Organization

A data moat is a competitive advantage where data itself is treated as a business asset. By leveraging data as an asset, businesses can create barriers to entry for competition and use data as a strategic advantage. The key to building a data moat is using data in a way that cannot easily be replicated. As a data scientist, you know what data is valuable, but as an MLOps practitioner, you can use this knowledge to build a data moat.

The first step would be to collect as much data as possible but to implement quality gates to safeguard the quality of the collection. This may require an investment in IT systems and tools to collect and process data effectively to determine what should be kept and what is noise.

The next phase is to identify what data cannot be replicated. This is the most valuable asset for a business and might be customer data, industry data, or data that was extremely difficult to collect.

Once you have identified enough quality data sources that cannot be easily replicated, you can analyze this data to leverage it in your operations. The full MLOps lifecycle applies at this stage, and you may start with a single data science project and slowly, iteratively build toward becoming a data-driven organization where you can offer new innovative service lines and products from this data.

Finally, after you've integrated a certain level of MLOps maturity meaning you're able to create feedback into your data collecting process to create more data, insights, services, and products, you need to safeguard the data and protect it like any other highly valuable asset by implementing proper data governance policies. This entire process may happen over a number of years.

One of the difficulties in building your organization's data moat is lack of domain expertise and capturing domain expertise in your MLOps process. In the next section, we'll look at the history of domain experts and how organizations have attempted to capture domain expertise when building their data moats.

The Changing Role of the Domain Expert Through History

Throughout history, there have been many AI winters and many attempts to capture domain expertise and store it. Expert systems were formally around as early as the 1960s and were designed to mimic expert decision-making ability in a specific domain. The first commercial expert system called Dendral was invented to help organic chemists identify unknown organic molecules by analyzing mass spectra. This and subsequent expert systems were rule based, and by the 1980s, they were able to make use of some simple machine learning algorithms. In the 1990s, expert systems were used in industries ranging from finance to healthcare and manufacturing to provide specialist support for complex tasks, but there was a problem: These expert systems were difficult to maintain, requiring human experts to update the knowledge base and rules.

Today, chatbots use a different approach: generative AI creating new data from old data that are far less brittle than expert systems. However, currently there is no way to update these chatbots in real time (requiring layers of reinforcement learning), and if data used to train these models is insufficient, the knowledge will be inaccurate. These models are also expensive to train with a fixed cost per token, and you have to train one model per domain; there is little to no transfer between domains leading to knowledge silos. This leads to an interesting question: What is the role of the domain expert in data science in the face of this change?

A challenging problem is there are many different kinds of data scientists not just differentiated by role and skill but domain expertise: knowing what tools are useful and what problems have been solved before and being able to communicate that knowledge.

Mathematics is a universal language. Data visualization can be used to communicate results of data analysis but hides the details of how you arrived at that problem. To make things worse, each industry has its own

vocabulary, standards, and ways of measuring. That being said, there are still a couple ways to store domain knowledge and maybe share that knowledge across industry and teams.

- *Documentation:* This is a straightforward way to store domain knowledge and share it across teams and industries. This may include books, technical manuals, research blogs from leading companies, academic journals, and trade journals.
- *Knowledge graphs:* Knowledge graphs are a way to organize domain knowledge into relationships between concepts. For domain knowledge that is highly relationship driven like social networks, this may be a good tool to represent knowledge.
- *Expert systems:* Expert systems were an attempt to represent expertise in a rule based system but have many limitations.
- *Ontologies:* Ontologies are a formal way to store academic domain knowledge by representing knowledge as a set of concepts and relationships between concepts. This differs from knowledge graphs in that ontologies are full semantic models for an entire domain while knowledge graphs are specific to a task.
- *Generative AI:* Large language models are increasingly being used to store domain knowledge. At this time, training large language models on custom data is an expensive process, but as the cost per token decreases over time, generative AI may become the standard way to share domain knowledge.

- *Code:* An example of this is the toolkit we created, but open source projects are good way to share knowledge across domains, for example, developing an R library to solve a problem in one industry and sharing it on CRAN so it can be applied in another industry.
- *Metadata:* Defining standard vocabulary for your industry and developing a metadata dictionary, for example, to annotate features in a feature store.

While we have many ways to share domain expertise from simple documentation to more formal methods to represent entire domains, sharing knowledge is only one piece of the puzzle. Data and the knowledge it represents grow over time and need to be processed not just stored. The situation where data outpaces processing capabilities may become a limiting factor in some domains.

Will Data Outpace Processing Capabilities?

IoT data is increasing at an alarming rate. This scenario is often called data deluge and happens when data grows faster than our ability to process it? While exascale computing promises to provide hardware capable of 10^{18} IEEE 754 Double Precision (64-bit) operations (multiplications and/or additions) per second, data is much easier to produce than it is to process. So-called dark data is produced when organizations have collected data but do not have the throughput to process it. While MLOps can help in unlocking some of this dark data, new systems, technologies, and hardware may have to be incorporated into the MLOps lifecycle to handle increasing volumes of data.

The MLOps Lifecycle Toolkit

The reader of this book is encouraged to use the MLOps lifecycle toolkit that is provided with the code for this chapter. I have added MLFlow and Jupyter lab components that use containers to the Infrastructure folder and added the model fairness code to the fairness folder so you can use it in your own projects. The accompanying software ("the toolkit") suit their own needs. The idea for an agnostic toolkit that can be used as a starter project or accelerator for MLOps can facilitate data science in your organization in combination with this book that serves as documentation for the toolkit.

Building a toolkit that is agnostic that includes tools for containerization, model deployment, feature engineering, and model development in a cookie cutter template means it's highly customizable to the needs of your particular industry and project as a way to share knowledge and provide a foundation for domain experts doing data science.

The field of data science is very fast-paced encompassing not only machine learning but nonparametric algorithms and statistical techniques, big data, and most importantly domain knowledge that shapes the field as a whole. As domain knowledge changes, the toolkit may evolve, but the invariants like mathematical knowledge, algorithmic thinking, and principles for engineering large-scale systems will only be transformed and applied to new problems. Figure 9-2 shows the end-to-end MLOps lifecycle components used in the toolkit as an architectural diagram.

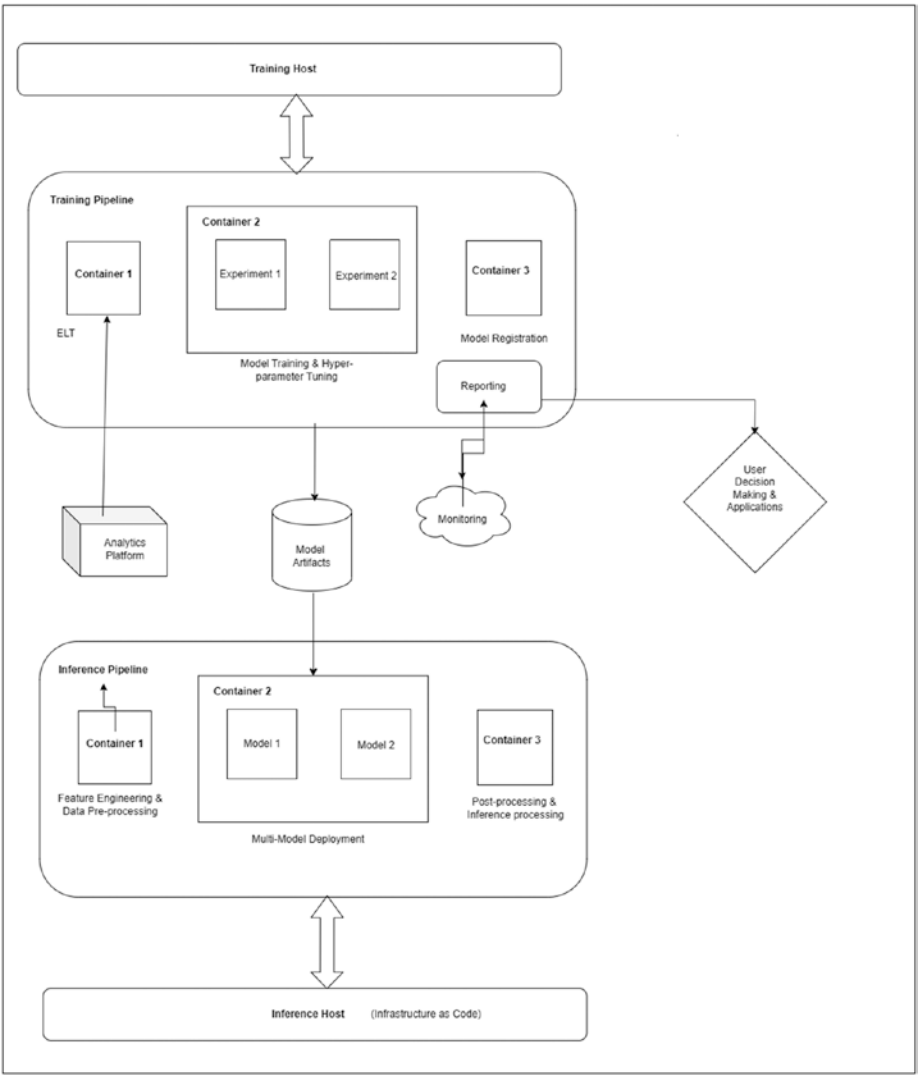

Figure 9-2. *Relationships between MLOps lifecycle components*

Summary

In this chapter, we looked at various applications of data science by industry. Some of the industries we discussed include the following:

- Energy Industry
- Finance Industry
- Healthcare Industry
- Insurance Industry

We concentrated on broad problems within each industry and emphasized the importance of industry standard techniques, vocabulary, and domain knowledge in data science. We looked at the changing role of the domain expert throughout history including attempts to capture knowledge and store it as expert systems and in generative AI. We discussed the hypothetical point where our ability to produce data may outpace our ability to process data and how this may impact the MLOps lifecycle in the future. Finally, we talked about contributing to the MLOps toolkit, the accompanying piece of software that comes with this book, providing the final version in the code that comes with this chapter with all the previous labs and infrastructure components used in previous chapters.