Deployment and Monitoring

It’s showtime!

In the summer of 2019, Apple launched its credit card, which involved a partnership with Goldman Sachs. It was designed to emphasize simplicity, privacy, and security (the card did not have a number, CVV security code, expiration date, or signature line). There were also no fees, the interest rates were competitive, and there was a good rewards program.

What about AI? It was a big part of the Apple Card. It was not only useful to help customers understand their spending but also to optimize the credit limits.

Cool, right? Definitely. But unfortunately, the AI had some glaring issues. The algorithms actually were biased. When Apple made an offer for credit lines, there were generally smaller ones for women vs. men! When customers realized this, they went straight to Twitter and made their displeasure known. Even legendary computer entrepreneur Steve Wosniak spoke out. His wife got a smaller credit limit even though her credit rating was higher than his.
Apple’s support team was caught off guard and stumbled badly. When handling incoming calls and texts, the explanation was an unsatisfactory: “It’s just the algorithm.”¹

Goldman Sachs then said that the algorithms were not biased because they did not use gender as a factor in the model. The firm also noted that it used a third party to test the Apple Card for bias.²

But none of this was enough for customers. The plain fact was that the Apple Card was skewed. If anything, it was a demonstration that avoiding the use of data like gender does not necessarily matter. Other types of data may correlate with this anyway (this is known as a data proxy). For example, if you use a person’s job description, this could bias the data for a certain gender.

This case study highlights how difficult deployment can be when it comes to AI. It also shows that some of the top companies can easily get things very wrong.

So in this chapter, we’ll take a deeper look at how to be successful with deployment in the AI process.

Types of AI Deployments

At a high level, there are two main ways to deploy AI. The first one is called an analytical model, which is similar to a BI (business intelligence) tool. This means that employees use the system to get reports, find trends, and get insights. Note that the analytical model is the most common.

Next, there is an operational model, and this is the far more complicated approach. This is where AI is embedded in a product, such as Uber. This requires access to real-time information and sophisticated systems to update the models. Before an operational model is launched, there needs to be lots of testing and quality assurance. Even a small glitch can be damaging to a company’s brand (as seen with the Apple Card example).

By and large, for those organizations that are starting with AI, the best approach is to focus on analytical models.

¹www.forbes.com/sites/tomtaulli/2019/11/11/apple-card--did-ai-run-amok/#6837a2802e8b
²www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/
MLOps

The category that involves handling the deployment and monitoring of AI is called MLOps (this is short for “machine learning operations”). It is similar to DevOps, which is where there is a combination of software development and operational functions.

To help with MLOps, you can use an AI platform (we covered a variety of them in the prior chapter). Such a system will have guardrails to prevent noncompliance issues, monitoring of performance, and functions to help with the handing off to different teams. The result is that the process is streamlined and less prone to error. There is also more time for the data scientists to focus on building better models and other value-added activities.

As these AI platforms get more sophisticated, they are handling more of the workload and providing much-needed automation. But MLOps still requires putting together organizational structures and workflows for the teams. There must be coordination across many different types of functions like version control, testing, analytics/tracking, continuous integration, model score validation, debugging, configuration, data management, and so on. Putting this all together does take time and a commitment from management. If not, then even a successful AI model could fail because it will not be deployed properly.

For example, suppose it takes a couple months to deploy a model. Well, during this period of time, the accuracy may decay as the data goes stale. This is actually a very common problem with AI deployment.

Keep in mind that MLOps is still in the early stages and there is no standard approach. Because of this, Cloudera has announced its Cloudera Machine Learning (CML) MLOps organization, which involves numerous partners to develop open standards and best practices for governance.

Another key element for MLOps is the need to have an assessment of the current IT environment. What legacy systems are in place? Do some of them need to be replaced or modified? This is absolutely critical because there needs to be a solid foundation for AI deployment.

Testing

Quality assurance or the testing of an AI model is part of the role of the data scientist. As seen in the last chapter, they evaluate the results by looking at accuracy, precision, recall, and so on.

Yet a data scientist can easily miss issues. The fact is that AI is fairly open ended, which can make it challenging to think through the many possibilities.
This is why it’s a good idea to have a QA person to focus solely on testing. Such a person may spot subtle problems but may also help with suggestions on the UI. Having an engaging user experience will go a long way in getting adoption across an organization or with customers.

An approach to consider for this is test-driven development or TDD. With it, you design tests for the functions of an AI application and, interestingly enough, you want them to fail (Figure 7-1 shows the process). In other words, you want to think in terms of cracking the system. The next step is to create code that makes sure that the test will pass. After this, there is often some refactoring or improvements to the code.

Sounds kind of tedious? It is. But the TDD approach really does lead to higher quality code. There is also much more emphasis on it being modular, and this makes it easier for modifications.

There are certainly a myriad of tools that can help with the testing process. They can provide assistance with areas like data analysis systems to detect problems like bias and outliers and with debugging tools to find overfitting and underfitting.

![Figure 7-1](image-url). This shows the process for test-driven development.
The UI

In 1996, Microsoft introduced Clippy as part of its Office platform. Clippy was an AI assistant that was in the form of a cartoon character (it was a paper clip with eyes!). But unfortunately, the user response was brutal.

This was despite the fact Microsoft had spent six months working on in-depth research. The company’s designers actually came up with over 250 potential characters. Yet Clippy turned out to be the one that was most trustworthy and engaging, according to surveys.

But it didn’t matter. When Clippy was deployed at scale, the results were downright awful.

This case certainly holds some valuable lessons for AI deployment. And yes, this means that UI really needs to be a priority.

So then, what are some of the takeaways with Clippy? How is it possible to avoid the problems?

Let’s take a look:

- **Little things matter**: A big problem with Clippy was its eyes. Some people thought they were … creepy.

- **Diversity**: The team that created Clippy was predominantly male and white. This likely influenced the characteristics and gestures of the figure. Consider that many women saw Clippy as off-putting.

- **First use**: Clippy would pop up when it looked like you were writing a letter (such as when you started with “Dear”). This was helpful. But when Clippy kept showing up for the same things, the experience became annoying.

- **Capabilities**: When Clippy was released, PCs were still constrained in terms of memory and capabilities. This meant that the AI was basic and this was a key reason why failure was common. For the most part, Microsoft was setting up unrealistic expectations.

Coming up with compelling UIs is tough and takes experimentation. But it is well worth the effort.

“The results of the data science project will be offered to the final users, usually non-data science experts,” said Rosaria Silipo, who is the principal data scientist at KNIME. “How much should the application GUI hide the complexity, and how much should be visible for customization by the end user? How important is the final GUI to allow the data science application to
communicate with the end user? Well, the application GUI must be informative and yet at the same time easy to use. Just a few clicks to move across the application but plenty of interactivity when exploring the results.\(^3\)

**Monitoring**

After an AI app has been launched—that is, put into production—there needs to be a focus on the operationalization. This means that the technology must be monitored and tracked to ensure the results are meeting your KPIs (key performance indicators).

After all, AI is evolving as it takes in new data. The goal is that it will get smarter and more effective over time. But then again, the reverse can happen as well. This is because the underlying circumstances of the model may change. This could lead to deteriorating results. This is known as model drift.

Let’s take an example of this. When the COVID-19 pandemic hit, various AI models simply could not handle the impact. Keep in mind that they were based on a history of data that did not include sudden spikes in demand for toilet paper, hand sanitizer, paper towels, and N95 masks. Such items generally see fairly stable patterns. The bottom line: There was extensive disruption in supply chains, which took time to fix.

But the COVID-19 pandemic had other adverse consequences. One was for mapping. Since traffic fell to unprecedented levels, the AI systems were unable to effectively optimize the routes.

True, these are extreme examples. Yet they do highlight the fact that models make various assumptions, which often change.

AI tools can help deal with such things, such as by providing alerts. But there should be ongoing testing of the models in terms of the accuracy rates. If there is slippage, then the model probably needs to be updated.

Having people in the loop is definitely crucial for this. Take a look at Phrasee, which develops AI software to automate the creation of copy for email campaigns and other marketing efforts. When the COVID-19 pandemic hit, the company had to act swiftly to change its models. To this end, the data scientists avoided the use of phrases like “going viral,” which would only alienate users. This even included avoiding the use of certain emojis that provoked fear.\(^4\)

---

\(^3\)From the author’s interview with Rosaria Silipo on June 12, 2020.

\(^4\)www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/
Note  Here’s another AI joke. An AI walks into a bar and the bartender asks, “What are you having?” The AI says, “What did the last 10,000 people have?”

Security

Besides the need for monitoring of the models, there must also be an emphasis on security. AI models often include sensitive data. Any breach could result in substantial fines and reputational damage. Thus, when creating an AI application, there needs to be an ongoing evaluation of security measures. This means reaching out to IT and the legal department to make sure the organization’s policies are in place.

Consider that AI is increasingly being used to pull off cyberattacks. The first known case of this came in 2007, with the CyberLover chatbot. It used NLP (natural language processing) to provide personalized responses to users that resulted in higher conversions of links to steal personal information.5

In some cases, the impact has been on a wide scale. Take the example of TaskRabbit, an online marketplace for freelancers. In 2018, it suffered an attack that impacted 3.75 million user accounts, which resulted in the exposure of Social Security numbers and bank account information.6

There has also emerged something called AML or adversarial machine learning. This is where someone creates malware with the intention of destabilizing an AI model (it’s also known as “model poisoning”). This usually involves flooding the system with new data, such as false positives, that distorts the training set. For the most part, such an attack can be difficult to detect.

Another vulnerability is with sensors, like with IoT (Internet of Things) systems. They can be hijacked and the data manipulated.

No doubt, the risks are scary and they are getting more sophisticated. This is why it is important to have a solid security strategy. It is usually a good idea to bring in a consultant who can secure the data and models.

Return on Investment

While monitoring the AI app, there should be a periodic assessment of the ROI (return on investment). And yes, there are various approaches to calculating this, although, given that AI provides long-term benefits, there probably should be some latitude. The early investments will likely be

5www.darkreading.com/vulnerabilities---threats/malicious-use-of-ai-poses-a-real-cybersecurity-threat/a/d-id/1337690
6www.infoq.com/articles/ai-cyber-attacks/
significant and there will be failed projects, which will depress the ROI in the early stages. This is why senior management needs to take a long-term view with AI. It is about a strategic opportunity that can potentially transform a business.

OK then, so when looking at ROI, let's first see the cost side of the equation. Some of the components include

- Computer power and storage
- Consulting fees, such as for the creation of the model, UI design, and integration with the existing IT infrastructure. There may also be a need to hire a subject matter expert (SME).
- Data acquisition
- Training the users of the AI application
- Costs for maintenance and support
- Compensation for data scientists, data engineers, AI solution architects, AI testers, and machine learning engineers
- Licenses or subscriptions for AI software platforms

AI also has unique aspects that make it different from a typical software project when it comes to the costs. How so? Well, consider an analysis from venture capitalists at Andreessen Horowitz that has the following interesting insights:

- Training an AI model can be expensive in terms of computer power. The potential costs could be in the six figures. But these are usually not one-time costs since data and models change over time.
- Model inference, which is about coming up with predictions while the AI app is in production, is much more complex than traditional software. The reason is the series of matrix multiplications.
- Storage costs can be significant since many AI applications use images, video, and audio.
- Even with the various AI tools, the data preparation process is still a highly manual process.
- AI applications can be challenging to scale in the cloud because of latency issues. As a result, a company may need to have trained models spread across different regions.
According to the post: “Taken together, these forces contribute to the 25% or more of revenue that AI companies often spend on cloud resources. In extreme cases, startups tackling particularly complex tasks have actually found manual data processing cheaper than executing a trained model.”  

This is not an argument against AI. Andreessen Horowitz remains a major investor in the space and is quite bullish on the prospects. However, an AI project does need a conservative approach with the budget and it may take some time to get efficiencies.

Note that there is another cost that is often overlooked: the cost of doing nothing. In other words, this could mean being put at a competitive disadvantage to competitors, which could lead to lower growth or even extinction.

So now, let’s take a look at the “return” part of the ROI calculation. The most common metric is to look at time saved. This could allow for employees to spend more time on value-add activities.

The calculation for this is straightforward. Example: Suppose you are using AI to automate the processing of invoices. In your finance department, employees spend 200 hours a week on this activity. Assuming the average pay is $20 per hour, then the annual savings would come to $192,000. If the cost of the deployment of the AI system is $130,000, then the ROI would be 48%.

But there are intangible benefits for an AI project. It could be, for example, an improvement in the user experience of a mobile app or an increase in the NPS (Net Promoter Score), which is an effective measure of customer loyalty. For various industries, AI can even be used to help improve the safety of employees, such as with offshore oil rigs or mines.

In some cases, there may even be revenue opportunities. To this end, an AI system could use a recommendation system to cross-sell existing customers or there could be an optimization of the pricing. In fact, the ROI will often include several metrics.

**Change Management**

Change management involves the preparation and support of an organization to adopt something new. No doubt, AI needs this for success. But change management is particularly challenging and is a big reason for the failure of any new technology implementation.

---

7 [https://a16z.com/2020/02/16/the-new-business-of-ai-and-how-its-different-from-traditional-software/]
AI may even more troublesome. Let’s face it, there is often much fear from employees that they will lose their jobs or be transitioned to another role, which may have fewer opportunities. There are also the problems with handling legacy systems.

There are a variety strategies and frameworks for change management. But there is one that has proven to be quite effective and it is John Kotter’s eight-step process. He came up with it during the mid-1990s (and it’s the basis of his best-selling book, *Leading Change*). It’s based on research on over 100 companies like GM, British Airways, and Bristol-Myers Squibb.

Here’s a brief overview of his approach:

- **Create a sense of urgency**: This needs to come from senior management. It is about crafting a vision about the importance of the initiative. An example of this is what Mark Zuckerberg did in 2012. At the time, Facebook was lagging in its mobile phone efforts. But Zuckerberg took swift action and led by example. He greatly increased the usage of his own phone and even wrote the company’s annual shareholder report with it.

- **Build a guiding coalition**: The senior management team is not the only influential group within an organization. There are always other people who have tremendous respect and authority. Thus, it is important to get their buy-in for the vision.

- **Form a change vision**: The vision needs a clear-cut destination that must be realistic, flexible, and attainable. This should be communicated in a brief manner, say in five minutes or less.

- **Communicate the vision**: Leaders in the organization must constantly talk about the vision. The fact is that it takes time for employees to internalize a new message.

- **Empower broad-based action**: Make sure that people are acknowledged for their efforts. And this should happen even if an initiative did not live up to expectations or even failed. The idea is that experimentation should be encouraged.

- **Generate short-term wins**: This helps to gin up momentum and enthusiasm. If not, there will likely be resistance to the vision.

- **Consolidate gains and produce more change**: There should not be the perception that the celebration of a quick win
is the end of the journey. Change management needs to strike a balance between the short run and the long run.

- **Anchor new approaches in the culture**: Sustaining the change is challenging. It might be the toughest part of the process. This is why the leaders in the organization need to have constant communication of the vision.

With these principles, it is certainly possible to create an AI-driven organization. The idea is that there will be a focus on analyzing data, not just having hunches about certain things.

A company that has been a leader in this is Stitch Fix. Here’s how the platform works: a user takes a survey to explain their fashion interests (the average client provides 90 data points) and then an AI system makes personalized clothing choices. As time goes by, Stitch Fix collects more data, which means that the results get better and better.

The company has 200 software engineers and 125 data scientists. According to its 10-K filing: “We use data science throughout our business, including to style our clients, predict purchase behavior, forecast demand, optimize inventory, and design new apparel.”

As a sign of the company’s strategic focus on AI, it has a Chief Algorithms Officer, who is Eric Colson. Prior to joining Stitch Fix, he was the VP of Data Science and Engineering at Netflix.

In an article for the Harvard Business Review, Colson set forth some of his key principles for creating an AI-driven culture. They include the following:

- Data science must be its own entity, not a subset of another department like product development or marketing. It should have a direct report to the CEO. Granted, the data science team will need to collaborate with other departments but this should be done on an equal basis.

- Data scientists need the resources required for success. This includes access to data, the ability to make hiring decisions, the authority to purchase sophisticated AI platforms, and the necessary computing systems. If there is an onerous budget process, this will stunt innovation.

---

8 [www.sec.gov/ix?doc=/Archives/edgar/data/1576942/000157694219000013/stitchfix10kfy2019.htm](http://www.sec.gov/ix?doc=/Archives/edgar/data/1576942/000157694219000013/stitchfix10kfy2019.htm)

9 [https://hbr.org/2018/11/curiosity-driven-data-science](https://hbr.org/2018/11/curiosity-driven-data-science)
Experimentation needs to be encouraged and rewarded. There also needs to be ongoing investment in training, not just for data scientists but the whole organization. Achieving data and AI literacy should be a strategic priority.

---

**Note** In a study from Deloitte, of the 37% of the companies in the survey that exhibited strong analytics cultures, about 48% of them greatly exceeded their business goals for the past year.¹⁰

---

**Deployment Case Study: Lemonade**

In 2015, Daniel Schreiber and Shai Wininger launched Lemonade. The goal was to disrupt the traditional insurance industry by leveraging the power of AI. Both founders had strong backgrounds in the tech world. Schreiber was the former CEO of Powermat Technologies (a wireless charging company) and he started an Internet company during the late 1990s. Wininger was also a serial entrepreneur; his biggest venture was Fiverr International (NYSE:FVRR), an online marketplace for freelancers. The company has a market value of over $2 billion.

Consider that the insurance industry is perhaps the first industry that used data. It was during the 1600s that John Graunt created modern statistics and used probabilities to determine life expectancies. All this would result in one of the largest companies in England: Lloyd’s of London.

Despite this, the insurance industry has generally lagged with cutting-edge technologies. But the founders of Lemonade saw this as a great opportunity. They built an extensive AI platform that includes the following:

- **AI Maya**: This is a bot that uses NLP to handle the collection of information, the creation of quotes, and the facilitation of payments. This has shown to greatly reduce onboarding times for new customers.

- **AI Jim**: This is the bot that manages insurance claims. AI Jim can handle the entire process and has been able to resolve matters in a third of the cases. And for those claims that cannot be processed by the bot, they are sent to the right person. Oh, and the resolution time is usually

¹⁰[www2.deloitte.com/us/en/insights/topics/analytics/insight-driven-organization.html](http://www2.deloitte.com/us/en/insights/topics/analytics/insight-driven-organization.html)
quite low because AI Jim has done much of the heavy lifting. As should be no surprise, the claims process is the most delicate for an insurance company. So the fact that AI Jim is able to manage this effectively is a testament to the power of AI.

- **CX.AI**: This bot resolves customer requests and has been able to handle about a third of the cases. CX.AI can help with coverage questions, pre/post purchase inquiries, changes to policies, and payments. By using this technology, Lemonade saw an 87% drop in the number of tickets processed by employees.

- **Forensic Graph**: This AI system uses behavioral economics and big data to predict, deter, and block fraud. Fraud accounts for $40 billion in waste for the insurance industry in the US.

- **Cooper**: This is an internal bot that runs various parts of the company. Just some of the processes it manages include processing paper checks and the running of thousands of tests for software releases. In fact, Cooper even analyzes the spectrometry imaging from NASA’s satellites!

Of course, at the core of all this is a massive data repository. So Lemonade built the Customer Cortex to manage all this.

According to the company’s SEC filing: “The power of our system goes beyond the sheer tonnage of data it generates, as we are able to put data to work in ways we believe, based on the experience of our management in the insurance industry, that legacy systems cannot. Our systems are entirely integrated, so data generated in a customer support interaction can inform the claims process, while claims data routinely impacts marketing campaigns, and so forth. Likewise, our bots do not merely collect data, but also adapt in real time in response to the data they collect.”

All in all, this is the cutting-edge of AI deployment and is likely to be a major competitive advantage, especially as Lemonade battles some of the world’s largest companies.

---

[11](https://www.sec.gov/Archives/edgar/data/1691421/000104746920003943/a2242013z424b4.htm#cu40510_our_lemonade_stand_a_letter_from_our_co-founders)
Conclusion

MLOps is something that will grow significantly in the years ahead. We will see new tools emerge and best practices evolve. The result will be much more powerful AI. Actually, as seen with companies like Lemonade, there can be the creation of breakout companies that transform industries.

In the next chapter, we’ll take a look at responsible AI.

Key Takeaways

- There are two main types of AI deployment. First, there is the analytical model approach, which is similar to a BI tool. In other words, the AI generates reports to provide insights, say about customers. The analytical model is the most common.

- Then there is operational AI. This is where the AI is embedded in a product (an example is Uber). By far, this a much more complicated form of deployment and requires significant testing.

- MLOps is an abbreviation for “machine learning operations” and is about the development, deployment, and monitoring of AI. There are various tools to help with this. But having a strong organizational structure is also important.

- Even though a data scientist will help with testing AI models, it is usually a good idea to have a QA person. They can find those issues that may be missed but also help with non-technical problems, such as with the UI.

- Test-driven development is where you develop software by setting up tests to fail and then find ways to fix them. Essentially, it is a proactive approach in finding problems with the application.

- Monitoring of an AI system can be done with software tools. But there should still be ongoing tracking from data scientists and other employees. It’s common for models to change over time and get less accurate.

- Security is another key part of the monitoring process since AI models use valuable data. This is why there should be coordination with IT.
Implementing AI Systems

- Adversarial machine learning is malware that destabilizes AI models.
- Periodically, the AI should be assessed for the ROI.
- In calculating ROI, the costs include factors like compensation for data scientists and other AI personnel, computer power/storage, consulting fees, data acquisition, maintenance and support, and tools.
- As for the “return” component of ROI, the most common is the reduction in the number of hours for a certain process. But there are certainly other benefits like improved customer experiences. AI may even result in higher revenues.
- For true adoption of AI, there needs to be change management. This involves techniques and strategies to galvanize an organization to take on a major effort. AI may be even more challenging with change management because there is often fear of the technology from employees.