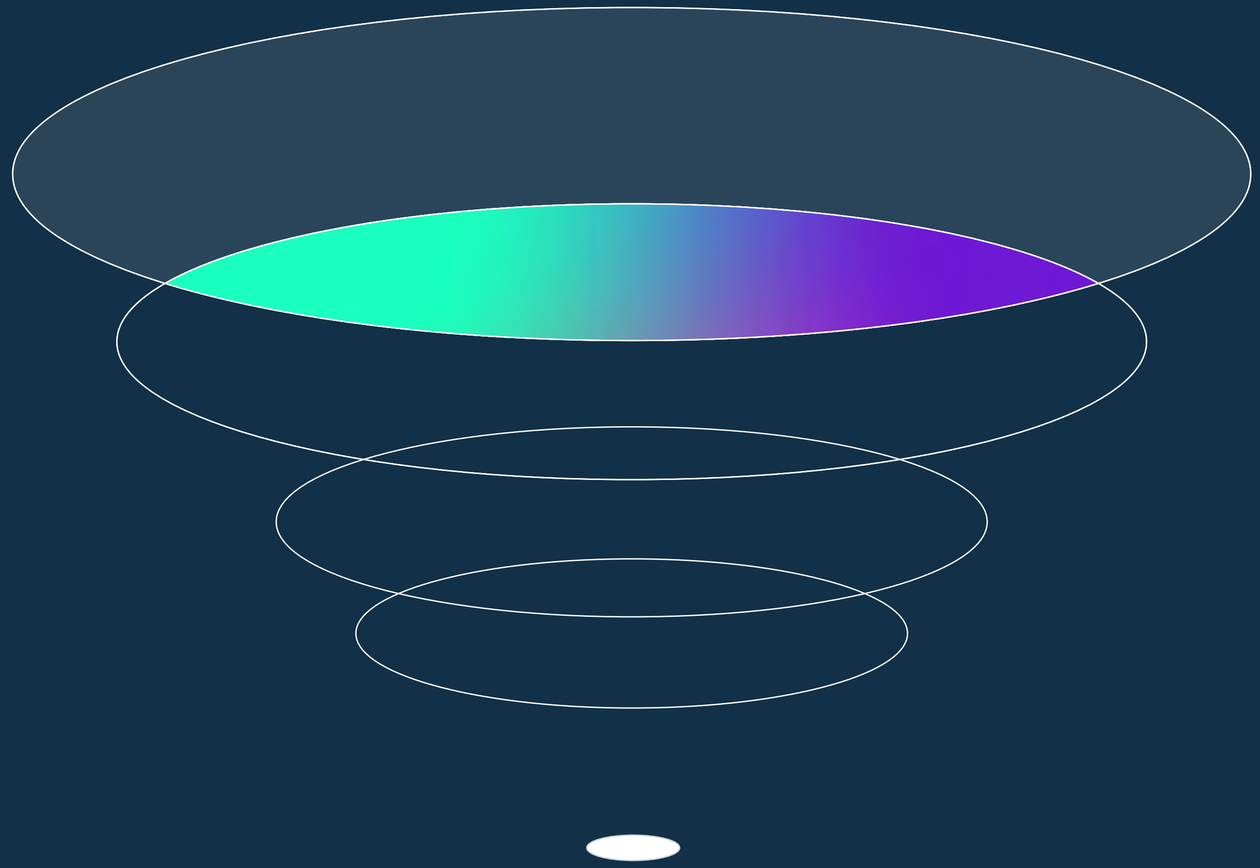
SQREAM

Dig deeper

The Case for the ML Use Cases

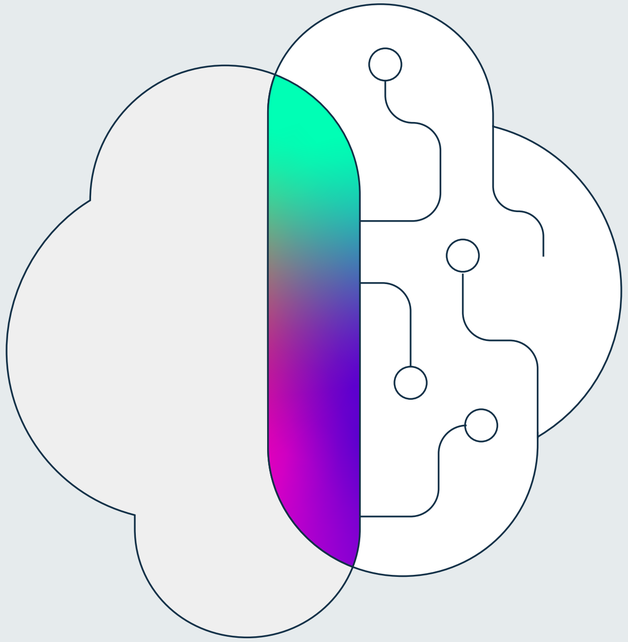


The answer's got to be in there somewhere. In your buyer behavior data. Your location-based data. Your IoT data. You've got petabytes to work with, but you're struggling to connect all the dots inside this data. Sure, you can run some basic queries to get some basic answers. But basic answers won't help you make big, transformative decisions – the ones everyone's counting on your department to deliver.

Dig deeper means uncovering more meaningful and impactful insights because you have the ability, augmented by **AI/ML**, to execute more complex data queries.

SQREAM

Introduction: The Evolution of Machine Learning



Data has been growing exponentially for years, and people have become more dependent on internet services. Additionally, the advent of the Internet of Things brought billions of bytes of data as devices proliferated.

As a result of its growth, data has become an organization's most valuable asset. It's integral to a successful business, yet achieving the full value from data presents challenges. Organizations may own their data, but many are still limited in their ability to ingest, prepare, and analyze it. This means that organizations miss out on potentially valuable critical insights that could be used to make better business decisions.

Machine learning (ML) allows companies to turn their exabytes of data into revenue-generating, actionable commands. **But how does ML magically derive insights from data?**

SQREAM

What Is ML and Why Use It?

ML is a field within AI that enables emergent insights on an organization's data through training the machine on a massive amount of data. This enormous dataset means that, as data practitioners are using ML, they can extract increasingly accurate insights from data through the use of statistical principles, linear algebra, and calculus embedded into the ML algorithms.

ML vs. Traditional Analytics

ML may already sound intriguing, but what makes it so valuable? To understand why ML is the "oil of our generation," we should compare it with traditional analytics: business intelligence + fixed policy automation. Look at the comparisons in the following chart:

Trait	Business Intelligence	Fixed Policy Automation	ML-Supported Automation
Known Trends	Strong	Strong	Strong
Large Amount of Data	Weak	Medium	Weak
Emergent Insights	Medium	Weak	Strong
Data Anomalies	Weak	Medium	Strong

Business intelligence involves people using their domain knowledge to mine data. There is merit here; people can be extremely insightful and creative, drawing insights with their data in a variety of ways. Businesses have layered fixed policy automation on top of that, applying known rules of their business or industry to process their multitude of data.

SQREAM

The Traits

Now let’s take a closer look at the traits of each type of analytics.

Known Trends

Fixed policy automation is strong at this. The rules are known and thus relatively easy to codify into automation. ML automation is also good at this. Known trends are often the most obvious clustering of data, easy for a model to pinpoint.

Large Amount of Data

Business intelligence tends to fall flat here; the human brain is limited in its capacity to easily process large amounts of data. Fixed policies are a step up because, with the right infrastructure, one can process large amounts of data through the same rules. This is where ML actually thrives. Large data sets that reach a certain threshold of quality increase the ability of ML to find insights in data.

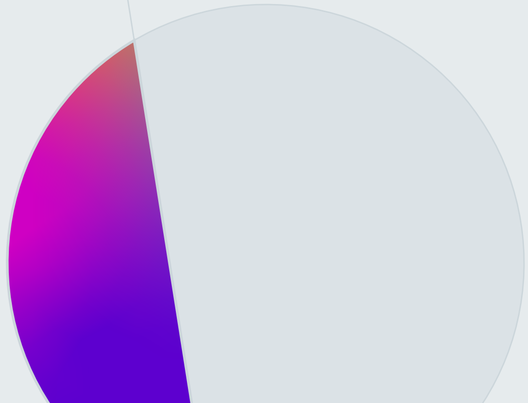
Emergent Insights

This is the ability to find things that at first glance may seem unrelated—for example, when Target realized it could accurately predict whether a woman was pregnant based on the type of lotion she purchased. Business intelligence is not well geared for this, but our collective creative genius can occasionally find these correlations. Fixed policy can’t find them by its very nature, as it is too rigid and can only operate mostly on the known. But here ML shines once again - it’s able to repeatedly find surprising results.

Data Anomalies

This is one area where business intelligence falls behind. The human brain tends to ignore information that is not tied to its biases. Fixed policy doesn’t do much better, as hard-coded rules struggle with varied-edge cases. But yet again ML triumphs, as its models are quite resilient to anomalies. They are not immune, but they tend to withstand a lot.

You may notice a trend in the chart: ML is an evolution of previous data analytics and is not even the end. The more we deploy ML-supported automation, the closer we come to autonomous decision-making.



ML Applications

As we learned above, ML can find emerging insights from large amounts of data with variation very well – even better and often faster than previous techniques. This makes it useful for a variety of business sectors:

Intelligence and Cybersecurity

More and more tools leverage ML to smartly identify malicious users and threats to digital infrastructure. On the other hand, ML can detect red flags that can reveal when people may be threats to national security or financial institutes or likely to commit violent crimes.

Ad-tech

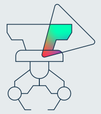
With Ad-tech, we can mine emergent insights to target ads more specifically to customers. In the case of our earlier example, Target leveraged its knowledge of lotion purchasing to advertise to women with baby registries.

Facial Recognition

There are numerous applications for facial recognition. One of its more prominent uses is to improve personal security through face identification instead of traditional passwords, as Apple intends to leverage.

Financial

Credit card history and other financial transactions are a veritable gold mine of insightful data. How people buy things and how much debt they accrue are very predictive of what habits they may have in their life and what interests they have.



Manufacturing

The ability to detect anomalies in data makes electronic manufacturing, and specifically semiconductors, a perfect application for ML. When collecting sensor logs from the production floor for fault detection and classification, the ability to analyze the massive data in crucial time is only possible using machine learning.

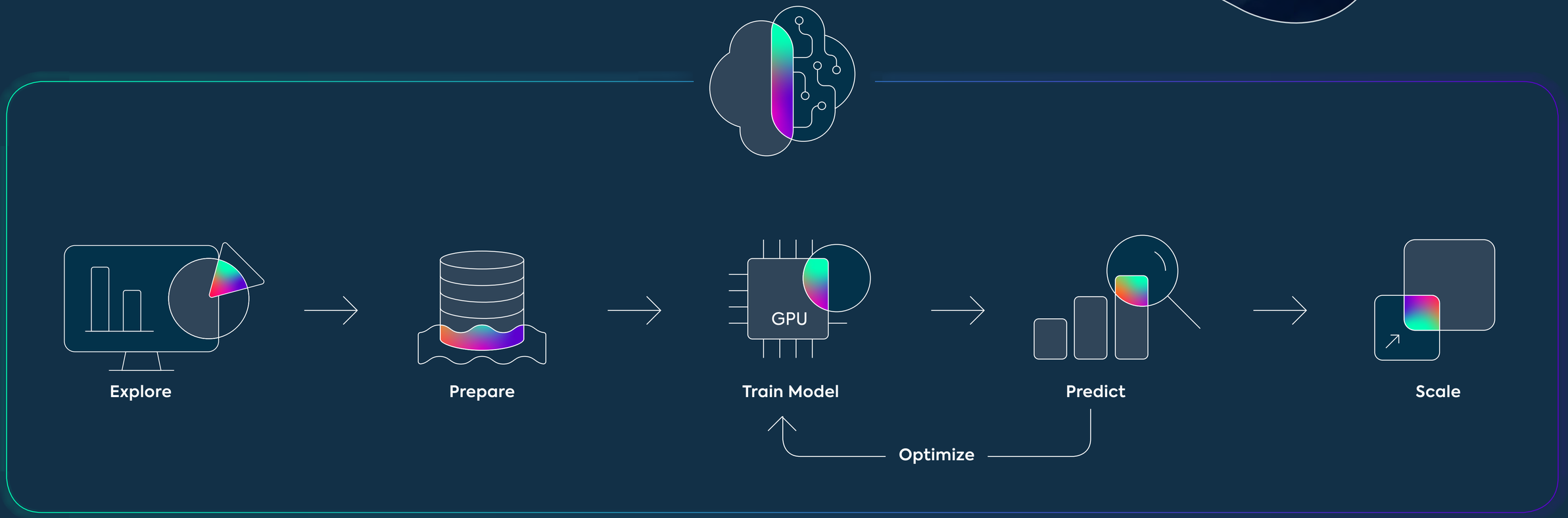
Healthcare

ML can potentially save lives, identifying behaviors that increase the risk for illness or analyzing a set of symptoms for surprisingly accurate diagnoses.

SQREAM

ML Pipeline

ML usefulness in business applications requires a fairly common set of steps, no matter the business context or the type of data you have. We call these steps an ML pipeline.



SQREAM

Data Exploration/Visualization

One key to a strong ML model is to choose the right algorithm and apply the correct variables for that algorithm. To start this process, a data scientist often needs to do some initial exploration of the data.

Sometimes obvious correlations pop up immediately. Perhaps this step could even supplant ML and produce a valuable algorithm at the outset. ML is not needed for every scenario, after all.

It can greatly help this step to create visualizations of the data. A picture (or, more likely, a fancy graph) is worth a thousand words, right? Your data is trying to tell you a story, and the ways to visualize this story are quickly evolving.

Data Cleaning

ML models live and die by the quality of their data, which could be its one weakness. Loads and loads of bad data can make a model useless, and bad or inaccurate data can be very easy to accidentally collect. Bad data can take many forms. It may be completely missing —maybe we know a demographic's gender but not their age, or unclear values, such as numbers with no unit of measure. Whatever the case, you need to carefully prepare the data.

Preparing the data is a process of trial and error. It often requires increasing amounts of data, so there may be back and forth with data engineers. As a matter of fact, based on a survey conducted by Anaconda, data scientists spend 26% of their time on data cleaning, which is more than the amount of time they spend building an ML model.

Data cleaning is a crucial step to ML models that is often worth the time you invest.

Data Modeling

This is the heart of ML - where data scientists fit an algorithm to the data. Good exploration and visualization can give an idea of what algorithms may work, but later steps will validate the fit more thoroughly.

To fit the algorithm, the scientist usually trains the model with a sample of the data. Training can introduce biases, as you are only dealing with a subset of the data and thus its characteristics. Such biases must be controlled while sampling the data. For instance, a random shuffle can be used while sampling (this cannot be used for time series data). However, in the real-world case, samples will always have some bias.

Despite the existence of biases, well-trained models tend to be quite accurate.

SQREAM

Testing and Using the Model in Real-World Applications

No matter how much testing is done, almost all good software algorithms will need to prove themselves in the field of battle: production. The real world is still the undisputed champion of testing things in ways that cannot be reproduced in a test or training environment. ML models are no different.

But how do we real-world test our models? Let's look at the example of fraud detection. As soon as the transaction is made, you need to enact that model. If the transaction is deemed fraudulent, then the institution would have 24 hours to stop or revert the charges.

And once you build a model to 95% or more accuracy, how do you connect it to the environment? The ability to connect this model to the environment is driven by architecture and needs to be part of the data pipeline.

Regularization/Optimization

Just because your model is seeing real-world use does not mean you should rest on your laurels. Most software needs performance tweaking, bug fixing, and other maintenance efforts after it is released. ML models are no different. This usually means improving the sensitivity of the model, allowing it to react accurately to a more diverse set of data.

In most cases, using more data will result in better accuracy, one of the reasons ML thrives on large data sets. Having enough data while training the model can greatly impact its performance.

A thing to note about this pipeline: ML does not replace human analysis. It actually accelerates it. ML shifts the data scientist from having to directly analyze data to curating the model that does the analysis. In a sense, the model abstracts and commoditizes data analysis, which in turn opens the door to innovations built on top of the commodity.

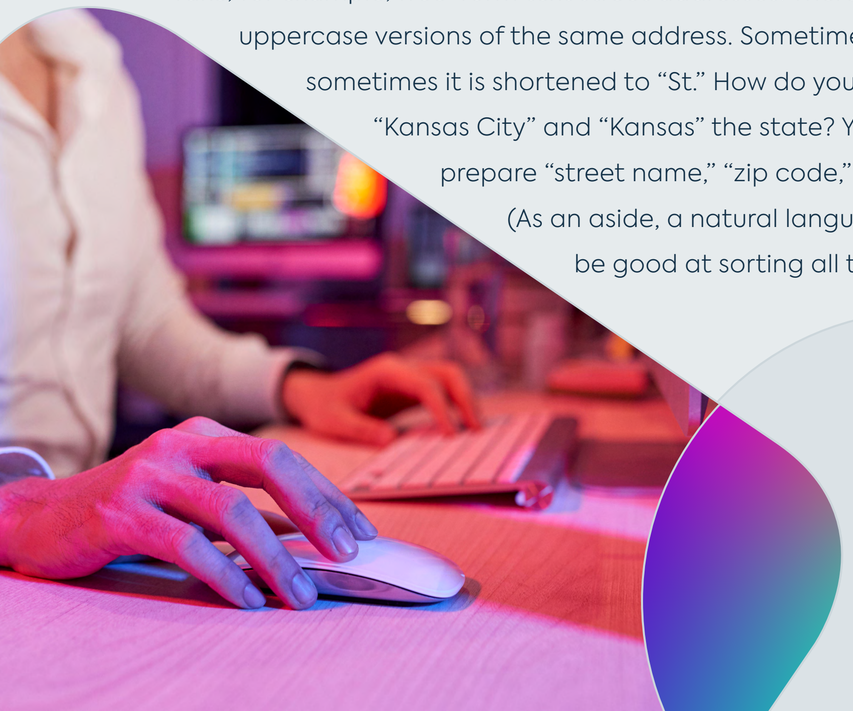
Real-World Challenges in ML Pipelines

Like everything else in life, ML pipelines come with their own challenges. And these challenges are far from solved across the industry. SQream, however, is a hybrid analytics platform that mitigates and deals with many of these challenges, as we will see below.

Tedious Process

The first challenge we encounter with an ML pipeline is that the data can be tedious and expensive, and just preparing it for a model can be a slog. ML data can easily be unrelated and unstructured sets. This requires preparation to get the data in some consistent structure to ingest.

Take, for example, free-form text street addresses. There could be lowercase and uppercase versions of the same address. Sometimes “street” is spelled out, sometimes it is shortened to “St.” How do you tell the difference between “Kansas City” and “Kansas” the state? You have to figure out a way to prepare “street name,” “zip code,” “building #,” etc. for ingestion. (As an aside, a natural language ML model would probably be good at sorting all that out.)



This also speaks to data quality. The dirtier the data coming in, the more cleaning you need to do. If you are doing this manually, each quality issue can compound and slow your ingestion by orders of magnitude. If you have automated cleaning, each quality issue means another set of business rules that needs to run at a massive scale. As stated earlier, ML models live and die by the quality of their data.

In our free-form address example, what if a user confused it for a notes field and typed in “at the white fence?” We have to clear those out too.

Inability to Train

The inability to train a model can stop it dead in its tracks. This is because not enough data or low-quality data can make your ML model feel like trying to adjust blinds that keep seesawing from one end to the other.

This is not solely related to data quality or quantity - you may have picked an algorithm that is a bad fit for your data. It’s OK to admit defeat and start over with a different algorithm. In the end, picking an algorithm often ends up being more art than science..

SQREAM

Dealing with Biases

When comparing ML to traditional analysis, we mentioned how business intelligence could be plagued with cognitive biases. We also mentioned in the Data Modeling section that ML models have their own biases. This is, at least to our knowledge, inescapable. Unless you can crunch all data in the universe into your model, biases will emerge.

For example, imagine you have country-wide statistics on the average steps taken per day per person in the US. In order to train your model, you only take statistics from city populations. How might those differ from steps taken in a suburb? Or in rural counties? Or perhaps we only train with statistics from the West Coast. Is it possible those in the Midwest, East Coast, or elsewhere have different walking habits? Even if we avoid these biases, will our model work in Europe? Are their walking habits similar enough? We would be surprised if they were.

Biases won't often break a model, but you should keep an eye on them as they can mess with your accuracy in the real world.

Architectural Needs

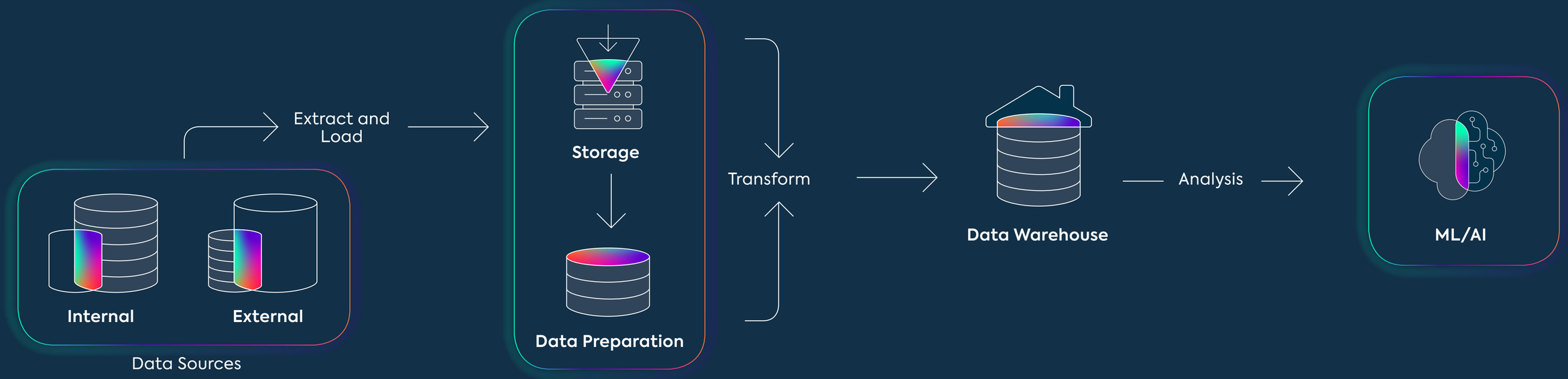
Once you have a model, you should have a plan for how to graft it into your computer systems – an architecture enables change. Ideally, you have a way to quickly react to accuracy issues. You probably don't want to spend a week coding in the model only to find out its accuracy is lower than expected, then waiting a week for a developer to remove it from the codebase.

Techniques like feature toggling and canary releasing can be a core part of this architecture. Another advanced practice is shadow analysis, where you run your model and get a result but report it to your monitoring team instead of acting on it. This lets you compare accuracy without directly risking customer impact.

SQREAM

Ask bigger – SQream Capabilities and Benefits

SQream heavily leverages GPUs and parallel processing for fast ingestion, enabling rapid total time to insight. Queries are also fast due to SQream's ability to take a string of events and connect it to a batch of data. SQream has found success dealing with machine learning models in the following ways.



Data Visualization

Once a scientist is ready to present their insights inside their business intelligence tool, using SQream means that the behind-the-scenes querying is completed faster. SQream also does a good job connecting training models to the business processes.

Data Ingestion and Preparation

The first steps in ML operations are data ingestion and preparation. SQream's ability to ingest and prepare massive, peta-scale datasets in minimized time to insight, makes it the perfect query engine for MLops.

Native Model Training

Using peta-scale datasets for training is crucial for effective and accurate training. Only with SQream

Use Cases

Even with this information, finding your own ML use case can be daunting. We can take a look at an existing use case and then an example use case to see how it all fits together.

Customer Experience Case Study

First up is a case study in which Thailand's leading mobile network provider, AIS, strove for high customer experience in a very competitive market, but had trouble finding a system to handle its billions of records. SQream was able to process their massive amount of records in seconds and allow AIS to plug into Tableau for data visualization. They were also able to drill down into device-level data for deep insights.

Example: Fraud Detection ML

Fraud detection is a great business process to apply ML. Fraudulent credit card users tend to have their own habits versus users with integrity, but these habits may not be obvious at first glance. Let's run through the steps we may take:

Data Exploration/Visualization

Firstly, let's choose our data source: credit card transactions. We can throw some user demographics in there as well, along with information about the locations where they use their card. We also may visualize these trends a bit to show that many users seem to provide some initial insights, such as what hours of the day most purchases are made.

Data Cleaning

We may choose to strip out outliers, such as validating transactions that are performed when someone adds a new card to a website. Or perhaps we want to remove transactions where we do not have clear location information.

Data Modeling

After exploring the data, we realize there are many categories: location, credit versus debit, amount, time of day, etc. Then, we decide that a random forest of decision trees would be the best fit for the data we have. Finally, we train the model with our cleaned data and find promising results: 90%+ accuracy.

Testing the Model in the Real World

We then coordinate with our payment processing software team to extract an interface for incoming transactions that matches the existing process. After that, we create a new implementation for that interface that uses the ML model. Then we code a feature toggle that routes to the new implementation and emits telemetry to our monitoring system on its results.

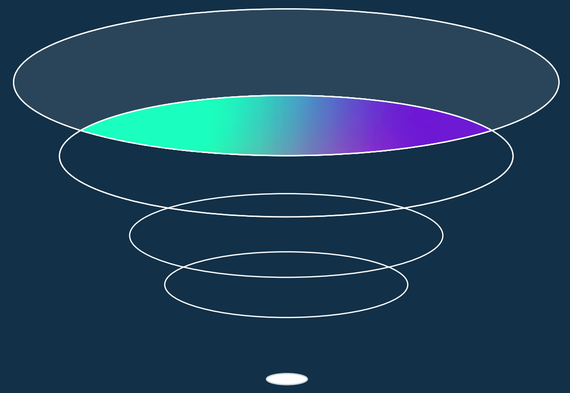
Regularization/Optimization

We find that the fraud detection struggles when people go on vacation, dropping from 98% to 89% accuracy. So we focus the model's sensitivity around vacationing data, enhancing our location decision tree with more decision points around whether the location is a hotel, a beach house, etc.

SQREAM

Conclusion

ML is more than a trendy topic; it is an evolution of data analytics that finds itself at home solving numerous business challenges. It enables you, as a data scientist, to focus less on direct analysis and more on curating your models and innovating your organization with new insights. Once you understand the steps, you can create your own use case. And tools like SQream help to mitigate or even eliminate many of the challenges you would otherwise encounter along the way.



Go faster and dig more deeper - The future of machine learning

The coexistence concept of end-to-end machine learning built around the GPU is the future of Machine learning. By coexistence, we mean being able to run all data preparations, machine learning models nativity, and inference together on the same GPU.

The results of the coexistence is 10X enhanced performance.

Think of the possibilities

Data scientists will ask much bigger questions, query more of their data, and increase the efficiency of their models. Businesses could dig deeper in their data and uncover more meaningful and impactful insights, saving tremendous time and costs of transformation, preparations, and storage.

From much faster and less costly fraud or anomaly detection to improved, always-on marketing that will increase incomes, new opportunities and improved use cases across all of the business units will pop in every industry.

SQREAM