DOMINO

WHITEPAPER

Introducing Model Management

A Framework to Build a Model-Driven Business

Contents

Executive Summary

This paper introduces Model Management, a new organizational capability for companies that want to put models at the core of business processes.

Models are the central output of data science, and they have tremendous power to transform companies, industries, and society. Amazon and Netflix — two of the most successful businesses of this decade — are just two examples of companies using models to create new products and drive ongoing operational improvement.

Despite the advantages of being model-driven, most companies are stuck trying to get there. A recent MIT Sloan study found only 5% of companies were extensively utilizing models in their business. Why is this happening? Companies are treating models like software when they are, in fact, very different — we call this the Model Myth. Even though models look like software and involve data, models have different input materials, different development processes, and different behavior.

To overcome the Model Myth, companies need to develop a new organizational capability called Model Management. Previously, model management referred to monitoring production models, but we believe it should encompass a much broader capability. Just as companies have built capabilities in sales, marketing, people management, finance, and so on, they need an equivalent capability in data science. Model Management is a new category of technologies and processes that work together to enable companies to reliably and securely develop, validate, deliver, and monitor models that create a competitive advantage.

Organizations that successfully build a Model Management capability will reap exponential rewards as more models drive better customer experiences and better margins. As models build on each other, more models also means more data and capacity for organizations to invest in new and better models. Those organizations will also better navigate common pitfalls that stymie the impact of models such as ethics and compliance risk. Ultimately, the haves and the have nots of this next era of computing will be determined by the quality of an organization's Model Management.

This paper is our effort to synthesize our learnings, distill the problem, and propose a path forward to achieve the full potential of data science. Part One describes what a model is and discusses how models drive business value. Part Two focuses on the essence of the problem — that models are different from anything built to date and it is a myth that organizations can manage them like other assets. Part Three dives into the details of a proposed framework for a new capability of Model Management which addresses the unique properties of models.

1 Models — What They Are and Why They Matter

In less than a decade, data science has risen from a niche function to a board-level focus. Why is that? What is it that data scientists do that is so valuable?

Beyond all the hype, all the buzzwords, and all the talk about AI and machine learning — at the heart of data science, the source of its power, is the model. Models are what data scientists make — they are where data scientists create their value.

A model is a special type of algorithm. In software, an algorithm is a hard-coded set of instructions to calculate a deterministic answer. Models are algorithms whose instructions are induced from a set of data and are then used to make predictions, recommendations, or prescribe an action based on a probabilistic assessment.

Many people have described data as the new oil. If that's the case, then models are the engines. Models make things happen. They initiate action. They can predict things before they happen more accurately than humans, such as catastrophic weather events or who is at risk of imminent death in a hospital. Models can build on each other. One model's output acts as the input to another, more complex model and then creates a living, connected, trainable army of decision makers. And for better or worse, models can do so autonomously, with a level of speed and sophistication that humans can't hope to match.

> Many people have described data as the new oil. If that's the case, then models are the engines.

Models started in finance and certain areas of risk management, but they are now proliferating to almost every industry. The forces of digital transformation are capturing more data about how businesses operate and are thereby creating more opportunities for data scientists to create models to improve how things are done.

How Models Create Value

Models dramatically lower the cost of prediction, similar to how semiconductors dramatically lowered the cost of arithmetic. This change makes models the new currency of competitive advantage, strategy, and growth. But how exactly do models translate into business success? There are two fundamental mechanisms by which models drive business value.

1. Models are the foundation for **breakthrough** products, killer features, or even entirely new revenue streams.

2. Models allow companies to create **operational efficiencies** that compound through **constant incremental improvement**.

The most successful companies are running on models

Many of the leading companies across industries have put models at the heart of their business, driving both net new products and customer experiences while constantly improving their core operations. Not surprisingly, some of the world's most transformative companies take a model-driven approach:

NETFLIX

The internet media giant famously developed a system of recommendation models which now drives more than 80% of content consumption, transforming the user experience. Recent estimates put the value of the recommendation models at more than $1 billion per year. Netflix also uses models to guide operational decisions such as which new shows are greenlit, subtly shifting the odds of success, which accumulates over time into more hits and fewer flops. These models are what they will depend upon to help them win the upcoming media battle with Hulu, Disney, and others.

amazon

One of the most iconic model-driven businesses, the e-commerce company exemplifies how to drive success via both the breakthrough and operational efficiency mechanisms. In his 2016 letter to shareholders, Jeff Bezos described their use of data science saying:

> "At Amazon, we've been engaged in the practical application of machine learning for many years now. Some of this work is highly visible: our autonomous Prime Air delivery drones; the Amazon Go convenience store that uses machine vision to eliminate checkout lines; and Alexa, our cloud-based AI assistant.
>
> But much of what we do with machine learning happens beneath the surface. Machine learning drives our algorithms for demand forecasting, product search ranking, product and deals recommendations, merchandising placements, fraud detection, translations, and much more."

It's this use of models that has made Amazon one of the most successful companies in the world.

Technology companies are not the only companies taking a model-driven approach. Others companies from various industries are also running their businesses on models:

The modern agriculture company increasingly uses precision farming models to improve crop yields. For example, in an effort to produce more resilient crops, Monsanto uses rainfall and soil chemistry models based on satellite imagery to predict the optimal locations for planting. Unsurprisingly, the power of models has promoted an increase in the industry's M&A activity, with Monsanto buying the Climate Corporation in 2013 for approximately $1 billion, and Bayer now looking to acquire Monsanto itself.

The drink manufacturer uses a model called the the Black Book to manufacture orange juice efficiently and consistently despite a limited harvest window in a single geography. The Black Book contains data on more than 600 flavors that make up an orange—including details on acidity, sweetness and pulp—as well as consumer preferences for each. It considers weather patterns, expected crop yields, and cost estimates on a real-time basis and prescribes an orange juice recipe optimized for customer satisfaction and profitability.

Models are a matter of existential risk

Companies like Allstate, Netflix, Amazon, and other early movers among model-driven businesses create flywheel effects. They are able to build a few models, gather additional data, improve those models, spread learnings from one area to another, increase the probability of breakthroughs, and drive ever-more efficiency. While flywheel effects started in digital-first businesses with rapid model feedback loops (e.g., online retail, digital advertising, and finance), the underlying forces of digital transformation give nearly every business the opportunity to be model-driven.

Similarly, it means no business is immune to the threat from a model-driven competitor. A recent McKinsey study showed model-driven leaders sustain operating margins 7% above the industry average. Non-adopters had margins 2% below the industry average. The nine percentage point difference today is magnified over time as model effects compound to push winners further ahead while laggards fall further, and eventually, impossibly far behind.

Some businesses may comfort themselves that they have strong competitive advantages in the form of their infrastructure, people, or data. However, the competitive advantages many organizations have previously relied upon are already eroding:

- Proprietary algorithms and infrastructure are increasingly threatened by open and cheap competitive offerings due to the rise of open source and cloud computing.
- Unicorn hires are increasingly transitive, with the median tenure of a data scientist less than 2 years.
- Proprietary data is increasingly accessible through sharing arrangements, aggregators, or regulated by rules like GDPR.

However, what remains a competitive differentiator is an organization's ability to develop, validate, deliver, and

> Model-driven organizations sustain operating margins 7% above the industry average. Non-adopters had margins 2.5% below the industry average.

2 The Model Myth Holds Us Back

As the previous analysis shows, there is tremendous power when a company is model-driven. Unfortunately, few companies are able to pull it off. According to analysis by MIT Sloan and BCG, only 5% of companies are making extensive use of models. Why is that? If being model-driven is so important, why aren't companies making more progress?

Many companies seem stuck in their journey to become more model driven. Said differently, it's not that companies aren't trying to become model driven; it's that their approach (for some reason) isn't working. Here are just a few of the things we have heard in our interactions with people at many companies:

- Data scientists face day-to-day issues with accessing tooling. One organization described how data scientists covertly brought their personal laptops to work because it took months to get new Python packages installed.
- Data science managers struggle with tracking the institutional knowledge generated during model development. One data science manager mentioned, "It's so bad that everyday is like my first day."
- Both IT systems owners and data science managers struggle to quantify project costs, right-size project resources, calculate ROI to justify work, and get promising models into production. One firm said they built their North American headquarters faster than they could get a model into production.
- Decision makers who use data science-generated insights seldom understand the assumptions and background of those insights. This lack of understanding leads to decision makers ignoring or misinterpreting the insights.

> Data scientists covertly brought their personal laptops to work because it took months to get new Python packages installed.
>
> One firm said they built their North American headquarters faster than they could get a model into production.

Models Are Different

But what's the deeper issue here? Given the importance of what's at stake and how much companies are investing in these capabilities, they're certainly not *trying* to get it wrong. Our view is that data science leaders and other executives at these companies are treating models like things they've seen before — software, data, or business intelligence — rather than recognizing that models are fundamentally different.

When figuring out how to do something, the common first instinct is to look to other capabilities organizations already know well — in this case, software engineering or data management — and apply similar principles. The misguided hope is that if data scientists can just act like a well-oiled software organization, the problems described above will fade away and transformative impact will be realized. They cite the proliferation of Agile and DevOps which helped dramatically improve development efficiency. However, they are making a critical mistake.

We call this mistake the Model Myth — the misconception that because models involve code and data, you can treat them the same way you treat software or data. Models are different and failing to treat them as such results in companies getting stuck on their path to be more model-driven.

Why models are different

Models involve code, but they aren't software. Models use data, but they aren't data assets. Models may generate visualizations but they are not business intelligence dashboards. They are a new species of asset, a new type of digital life. The most successful model-driven businesses recognize models are different, and act accordingly.

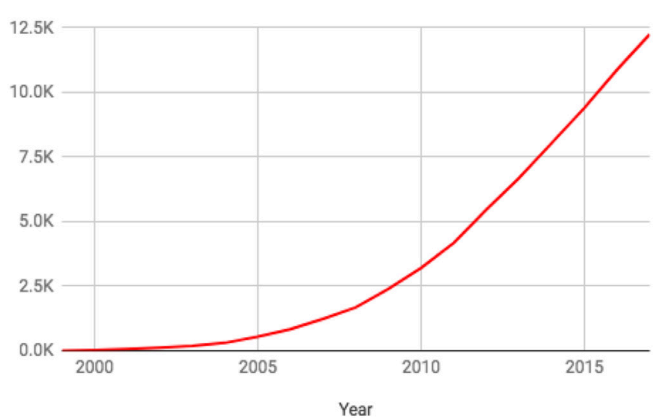
Models differ from other business assets in three ways:

1 Models use different materials

Models involve code, but they use different techniques and different tools from software engineering. They use more computationally-intensive algorithms, so they benefit from scalable compute and specialized hardware like GPUs. They leverage packages from a vibrant ecosystem that's innovating every day. As a result, data scientists need extremely agile technology infrastructure to accelerate research. Software developers and BI analysts generally have a standard development environment, while a data scientist's stack evolves constantly and is unique to each data scientist. To illustrate this, the chart below shows the exponential growth in the number of packages in CRAN, the canonical repo for the R language.

Open Source Innovation Accelerating Exponentially

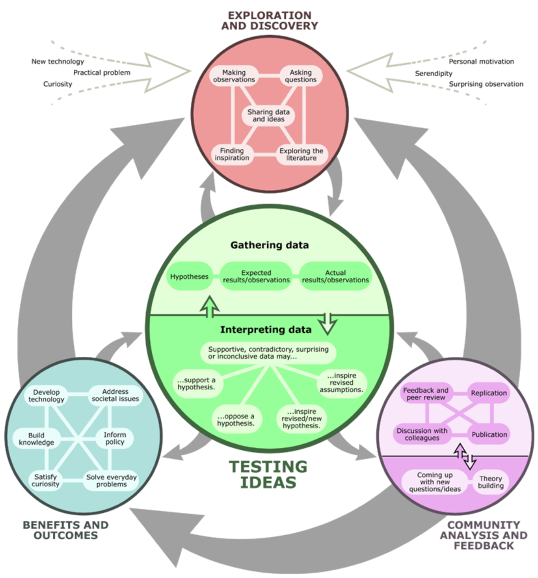
Cumulative CRAN packages released

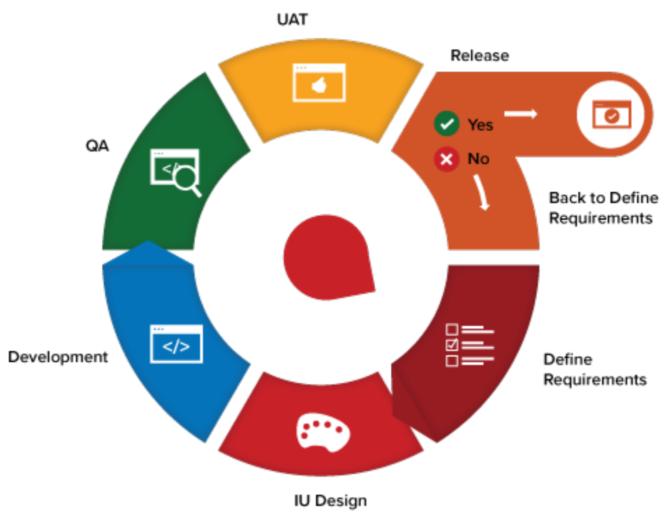


2 Models are built differently

The process to develop models is different. Data science is research — it's experimental, iterative, and exploratory. You might try dozens or hundreds of ideas before getting something that works. In software development, such false starts and dead ends are not preserved. In software, when you make a mistake, it's a bug. In data science, a failure can be the genesis of the next breakthrough.

One data science leader described the consequences of this difference well: "Every breakthrough we've ever had has come from one person picking up someone else's work and taking it in an entirely new direction." This has implications for the tools and processes data scientists need to accelerate research of data science. The charts below highlight the difference in workflow of data science and software development. The chart on the left is from UC Berkeley's scientific process diagram while the chart on the right shows an agile software process. While both are iterative, the scientific process is much more fluid, benefiting from cross-pollination and the ability to retrace your steps before finding a new path forward.



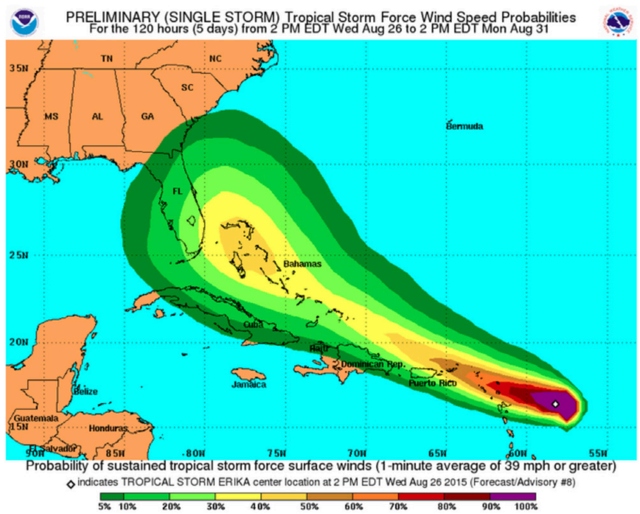


Sources:
1. https://cdn.arstechnica.net/Science/2009-3-16/science_process_diagram_big.gif
2. https://www.smartsheet.com/understanding-agile-software-development-lifecycle-and-process-workflow

3 The behavior of a model is different

Unlike software, which implements a specification, models prescribe action based on a probabilistic assessment. Statistician George Box captured the difference well saying, “All models are wrong but some are useful.” Models have no “correct” answer — they can just have better or worse answers once they’re live in the real world. And while nobody needs to “retrain” software, models can change as the world changes around them. They can also spawn self-reinforcing feedback loops or be self-canceling. Therefore, organizations need different ways to review, quality control, and monitor them.

Most people and organizations are accustomed to the binary behavior of software and are less familiar with the uncertainty of a model’s behavior. A model’s probabilistic behavior conflicts with many of our existing cognitive biases, such as overconfidence and recency bias. This conflict leads to misinterpretation or limited adoption of models. The image below shows an example of this behavior with a hurricane forecast model. The range of possible outcomes complicates how residents and downstream systems (such as emergency resource coordination) should respond.



In fact, models are so different that when companies fall victim to the Model Myth, they get stuck in a morass of problems. Their data scientists aren’t equipped properly (e.g., “I don’t have the compute infrastructure I need”). Data scientists are siloed in their work (e.g., “I have no idea what anyone has done on this problem before me”). Projects aren’t framed correctly for executives (e.g., “Why are so many of your projects failures?”). Models cannot get deployed quickly (e.g., “First we have to reimplement in another language”). And then, as a result of all of this, companies don’t become model driven. Solving the Model Myth requires a different approach.

You’re living the Model Myth when...

- You need the latest version of a Python package but your IT department makes you wait months to run it through its change management process.
- You can’t reproduce the results of an old experiment because Github and your wiki don’t track the data sets you used as inputs.
- It takes months to deploy a model because someone has to reimplement in Java so it runs in the production environment, and nobody has considered that this approach will make it nearly impossible to retrain at the necessary frequency.
- You have to explain that many of your team’s projects fail, and that’s okay, because data science is research.

3 The Solution—Model Management

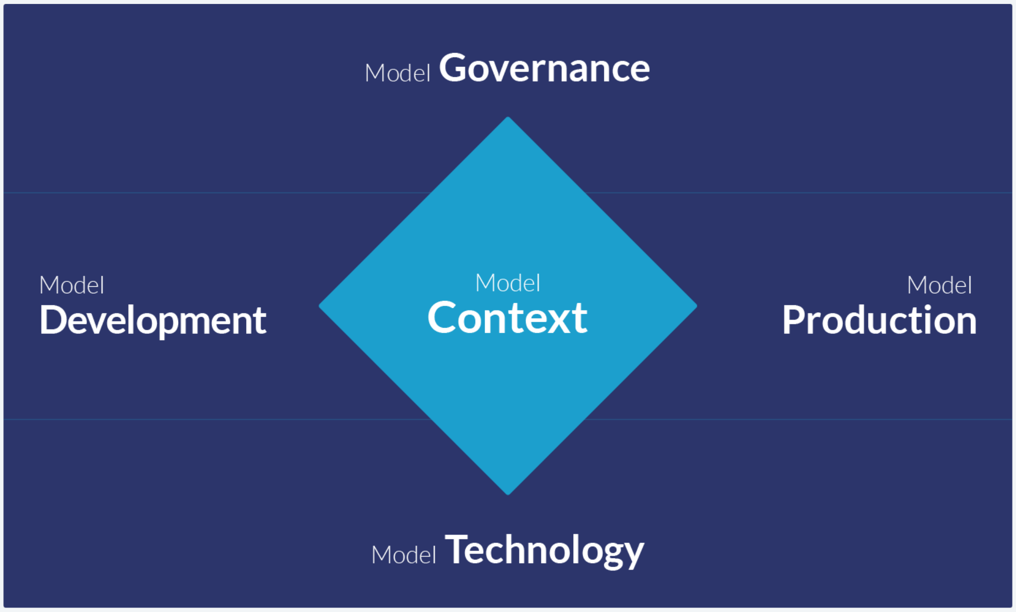
Businesses have developed organizational capabilities for all essential activities: sales, marketing, people management, software engineering, among others. Organizational capabilities allow a business to perform a key function reliably at scale and move that function beyond a dependence on individual heroics or silver bullet tools. Model Management is the name of that organizational capability when applied to data science.

Model Management is a new capability comprised of technologies and processes that work together to enable companies to reliably and securely develop, validate, deliver, and monitor models that create a competitive advantage. We believe it is the solution to the Model Myth. It is the way to unlock the transformative potential of models.

In the past, “model management” has been used as a technical term of art for tracking and monitoring models running in production. We mean it as something much broader. In this section, we offer a framework for this new organizational capability and a vision for how it can change the way we do data science.

The Pillars of Model Management

There are five pillars to Model Management. The diagram below shows the blueprint of Model Management.



Most organizations today have invested in building out a capability in two or maybe three of these pillars. Typically, they have focused on Model Development (the ability to build models), Model Technology (the ability to equip data scientists with technology), and, hopefully, Model Production (a way to productionize those models). However, the leaders in this space recognize those alone are incomplete and unreliable as a team's size and expectations scale. We believe that the full potential of models can only be unlocked with a complete solution that unifies all of these capabilities— a solution that embraces the unique traits of models, including the ability to govern the entire system and manage the knowledge generated throughout. The remainder of this section describes each pillar and how it will change the way we do data science.

Model Technology

Model Technology encompasses the compute infrastructure and software tooling that gives data scientists the agility they need to develop and deploy innovative models. The input materials for models are different from software or BI. Models are constantly evolving, with unprecedented innovation in both the open-source and commercial ecosystems. Models also use more computationally-intensive algorithms, so they benefit from scalable compute and specialized hardware like GPUs. Model Technology provides the backbone of the complete data science system, allowing organizations to use the appropriate tools at each step, from early data exploration to ongoing monitoring of mission-critical model products.

How does your world change?

- Data scientists in large enterprises today often endure months-long package acquisition and approval processes. This forces data scientists to create their own unsanctioned shadow IT if they want to use modern, industry-standard tools. With Model Technology, data science teams can seamlessly choose and swap out the right tools and libraries for each step of the experimental process.

- It is common for models to overwhelm local workstations, prompting endless IT resource requests, and hours wasted waiting for training jobs to error or complete. With Model Technology, teams have one-click access to scalable compute, parallelized training, and specialized hardware like GPUs without unduly burdening their IT partners.

- Managers of data science teams and the IT teams that support them frequently stress about shaky infrastructure supporting running models, such as local Cron jobs driving mission-critical workloads and failing silently. With Model Technology, enterprise-grade infrastructure underlies every delivered model product, scaling effortlessly to meet the needs of internal stakeholders and customers.

Model Development

Model Development allows data scientists to rapidly develop models, experiment, collaborate, and drive breakthrough research. The process to build a model is much more experimental than software. Data scientists explore data interactively and programmatically test many different permutations of features and algorithm types. Also, data scientists share insights with colleagues and stakeholders as well as store insights for later use. Often the insights gained in model development are just as important to the success of an organization as the model itself and should be categorized and retained. Given the increased significance of models in companies, validating models also requires organizations to be able to run rigorous testing, consult subject matter experts, and ensure a robust promote-to-production process.

How does your world change?

- Manual hyperparameter tuning elongates the data science lifecycle by days, even weeks, and limits the scale of problems that are feasible to tackle. With Model Development, teams can try many experiments in parallel to search and prune large problem spaces with ease to find the optimal model.

- Data scientists today try in vain to remember the steps they took to finish a model, so they preserve only the end state, and even then it is often incomplete. With Model Development, data science teams document and preserve each step of the open-ended experimental process and any insights unearthed for collaboration with other colleagues or their future selves.

- Today, stakeholder collaboration means a dreaded bi-weekly stakeholder PowerPoint update where charts are taken out of context and feedback becomes strewn across many email chains. With Model Development, there is real-time, contextualized collaboration and feedback from subject matter experts and IT engineers who will ultimately consume the model.

- Model validation is a bottleneck because even cursory validation and code review take months given the impossibility of reproducing results, and even then do little to mitigate risk. With Model Development, teams can progress through rigorous validation processes and minimize risk of methodological error or compliance issues.

Model Production

Model Production is how a data science team's work is operationalized. It is how it goes from an innovative project to a live model product integrated into business processes, affecting decisions and driving value. Since a model behaves differently than software, Model Production enables organizations to consume models in ways that allow humans to easily apply necessary judgment or provide feedback. Model Production also tracks how a model performs, how it is used, and ensures a closed loop to drive iterations and improvements.

How does your world change?

- Data scientists often throw models "over the fence" to downstream stakeholders with little hope or expectation of feedback. With Model Production, stakeholders can easily discover all model products and have open lines of communication to the model product creator.

- It is not uncommon for data scientists today to dumb down promising models into a hard-coded set of coefficients for manual re-coding in Java if they hope to deploy into production. With Model Production, data scientists and application developers work hand-in-hand to create model products that seamlessly integrate into downstream systems and processes via APIs and web apps.

- Data science leaders struggle to know if the models their teams ships have any measurable impact because of a lack of testing infrastructure and culture. With Model Production, data science teams can see how downstream consumers and end users are engaging with model products and know their impact on the business with techniques like A/B testing, multi-armed bandits, and universal holdout groups.

- IT teams responsible for a production model will often let it drift rather than hunt down the original creator and run the gauntlet of validation and deployment again to re-train. With Model Production, you can instantly detect models drifting, retrain and deploy new versions, or shut down and revert to a prior version to mitigate risk.

Model Governance

Model Governance is how a company can keep a finger on the pulse of the activity, cost, and impact of data science work across its organization as well as understand what's going on with projects, production models, and the underlying infrastructure supporting those. Governance of the whole model system is far more complex than other systems because of the confluence of previously described unique traits of models: rapidly evolving toolkits, research-based development, and probabilistic purpose. While governance sounds antithetical to the scientific ideals of data science, it is critical to delivering business value and mitigating risk that can undermine the transformative potential of models.

How does your world change?

- Leaders of major data science organizations frequently wonder exactly how many models are in-flight and bemoan perpetually outdated model inventories. With Model Governance, leadership has real-time transparency into the aggregate model portfolio.

- Due to a lack of visibility, junior data scientists can waste three weeks down a rabbit hole over-optimizing a model before their manager can step in to help. With Model Governance, managers can spot problems and dive into any project for rapid course correction and quality coaching.

- Data science and IT clash because of struggles to forecast compute spend accurately, leading to missed budgets or wasted resources. With Model Governance, both groups have granular knowledge of where key resources are used and how they can be used more efficiently.

- Infrastructure teams deal with CACE (change anything, change everything) problems and the unknown risk to downstream models and systems. With Model Governance, teams have real-time mapping of their model graph that encompasses all of the dependencies and linkages across critical system artifacts.

Model Context

Model Context is all the knowledge, insights, and artifacts that are generated while building or using models. It represents the complete provenance of a model which can be modularized (features, datasets, environments, code, subject matter experts, validation checks, monitoring plans) and reassembled into new models. This is often a company's most valuable intellectual property, and the ability to find, reuse, and build upon it is critical to driving rapid innovation. Like in science, bottoms-up cross-pollination of knowledge is the most effective and sustainable to scale the impact of models. Model Context is also essential to proactively addresses concerns around compliance and auditability of models.

How does your world change?

- Data science teams depend on tribal knowledge and fear what will be lost when someone leaves the team. With Model Context, leaders have confidence knowing the organization is resilient to turnover.

- Even in organizations with years of track record, most projects start from scratch because it is easier than digging up old work that wouldn't run anyway. With Model Context, most projects have a headstart as they are a recombination of existing insights and artifacts.

- Stakeholders want to know how critical models are built, but they typically find a black box which undermines their trust and willingness to change their workflows. With Model Context, all stakeholders have transparency into the complete provenance of a model including dependencies, judgment calls, and known risks for compliance, audit, and ethics purposes.

- Today, an escalating **reproducibility crisis** undermines confidence in data science output, stalling progress across companies and academic fields. With Model Context, there is reproducibility of past results and the confidence to use those results as the basis for more ambitious and complex projects.

- In companies of all sizes, data scientists hunt for elusive subject matter experts and often end up discovering them too late, or not at all. With Model Context, there is a mapping of an organization's expertise on technical and business domains that accelerates research and feedback.

These pillars form the capability of Model Management. While not a step-by-step how-to, this framework has helped the leading organizations we work with understand the shape their eventual solution must take.

4 Conclusion

History offers an important lesson for the type of change that organizations now need to drive. What the data science industry is currently experiencing today may feel novel, but it bears a striking resemblance to the computing transformation from the hardware-focused era to the software-focused era roughly 20 years ago. At the time, Waterfall was the accepted methodology for building hardware. It had its roots back millenia to Roman construction best practices. But when companies started to build software at scale in the 1990s using Waterfall, engineers got stuck. The Agile movement was the answer. It reflected a more conscious recognition that building software was fundamentally different from physical engineering. Software can be fixed faster and how it is used changes faster.

Today companies are building a new thing, like software was new then, but falling prey to the same trap. Data science is as different from software development as software is from hardware. By understanding what a model is and why it is different, organizations can unlock the full potential of data science, just as Agile unlocked the potential of software development.

By extension, Model Management is the next great paradigm shift for businesses. The organizations that become model-driven will continue to distance themselves from laggards. Much of the frustration and disappointment to date stems from shoehorning models into existing organizational paradigms like software development or data management. By recognizing what is unique about models and building an organizational capability that takes the learnings of other capabilities and embraces these differences, businesses can generate more breakthroughs and consistently improve their core operations. This is the key to remaining competitive in this next era.

> By recognizing what is unique about models and building an organizational capability that embraces these differences, businesses can generate more breakthroughs and consistently improve their core operations.

DOMINO

Domino Data Lab provides an open data science platform to help companies run their business on models. Model-driven companies like Allstate, Instacart, Dell, and Monsanto use Domino to accelerate breakthrough research, increase collaboration, and rapidly deliver high-impact models. Founded in 2013 and based in San Francisco, Domino is backed by Sequoia Capital, Bloomberg Beta, and Zetta Venture Partners.
To learn more, visit dominodatalab.com