Pump Up Data Science Productivity with a Modern Workbench

WHITEPAPER

Table of Contents

Executive Summary

Predictive modeling is transforming the nature of how businesses are run. Models provide insights that let leaders reliably manage for the future instead of using indicators that only show what happened in the past. A model-driven enterprise relies on data science – particularly, upon data scientists who possess the technical skills to execute on the promise of predictive modeling. For most organizations, early forays into modeling often yield quick results on small trial projects, but efforts to scale data science for enterprise production usually fall flat.

The culprit is the lack of scalable and flexible tooling and workflows that allow large teams of data scientists to systematically experiment and collaborate on projects that are unlike typical software or product development. Without the freedom and ability to try new tools, algorithms, and infrastructure (e.g., GPUs and distributed compute, productivity of data scientists is often spotty, with massive efforts yielding minimal results. The answer is a specialized data science workbench – a self-service software platform that instantly spins up any tool, package or compute resource needed by teams to do their work quickly and effectively, without requiring IT administration for each request.

This white paper describes three pillars of productivity for a re-imagined data science workbench and how they address scalable requirements for a model-driven enterprise with consistency in using any tool or process required for a model; context for research, collaboration, and knowledge acceleration; and coordination of projects to solve complex business problems.

Challenges of Scaling Data Science for the Enterprise

There is a huge lure of transforming modern business with data science. The legacy use of data dwells on backward-looking indicators to answer “What happened?” Predictive machine learning models for operational use at scale are helping businesses to understand “What should we do next?” Data scientists are combining data, code, and parameters to produce models that generate results -- i.e., predictive insights to make the business more profitable. Things like: What will customers want to buy next? How much inventory should be kept on hand? How can we produce the new products in the most cost-effective manner? With data science, a business no longer needs to plan for now; it can productively plan for next and get continuous improvements in the models and their results.

Capturing the benefits of data science at scale, however, is not easy. In fact, being a productive, model-driven business is hard! An organization's first venture into data science is usually a small experiment with a few experts working on an exciting problem. Success is often quick to achieve, easy to see and relish. But when the organization tries to scale data science, it usually slams into three obstacles to productivity:

Access to tools and infrastructure. Data science requires access to powerful compute resources, high-value and sensitive data, and the latest open-source tools and packages. This flexibility is essential both to making data scientists productive and to innovation. If companies want to realize the promise of data science, they must empower data science teams with compute and tool agility to support diverse experiments. When data science teams cannot get access to the infrastructure they need, they resort to ad-hoc workarounds that involve building and maintaining their own local infrastructure (e.g., unsecured laptops, local servers, unmanaged cloud environments). Productivity becomes impossible. The multiple stacks of data science tools and bespoke hardware for each team slows data scientists down, frustrates IT professionals, creates significant support costs, and increases operational and security risks.

Breaking down the silos in which data scientists typically work. Productivity is unachievable because data scientists often work independently with a variety of different tools, so there are no standard ways of working, which compromises governance to enable auditability, reproducibility, security, and so forth. The lack of visibility into in-flight projects also enables the loss of intellectual property and institutional knowledge when data scientists move on. Teams cannot find and reuse previous work and cannot effectively collaborate with one another. New employees struggle to onboard and contribute in meaningful ways. Everybody winds up doing lots of repeat work, which wastes time and effort.

Operationalizing models so they have actual business impact. The best model in the world means nothing if it is not moved into production so it can impact actual business processes. This delays business value and increases risk. Every time you want to put a model into production, you are reinventing processes and adding more complexity – the opposite result of productivity! When you finally do get a model into production, there are inconsistent (or nonexistent) monitoring practices that create additional risks.

To meet these challenges, data scientists need a specialized workbench, which is a self-service software platform that enables data science for enterprise-scale production.

Three Pillars of a Data Science Workbench

Traditional software engineering principles are insufficient to effectively manage the materials, process, and behavior of a data science model. New tools and frameworks are required to meet unique needs for doing data science. Unfortunately, their piecemeal acquisition for specific tasks leaves IT and data science teams with a disconnected hodge-podge of tooling. They inhibit productive collaboration. They are not equipped to handle a portfolio of models in production. Nor do they capture the full range of artifacts needed to reproduce model results on demand. What's missing is the means to help data scientists to leverage the ecosystem of task-specific tools and frameworks in a collaborative manner for enterprise scale deployment. A data science workbench is the modern mechanism for accomplishing this scale. It supports three pillars of productivity that are essential for the unique attributes of doing data science: Consistency, Context, and Coordination.

1. Consistency in Using any Tool or Process Required for a Model

This pillar is about the consistent application of tools and processes used for conducting data science. Consistency helps ensure logic, accuracy, and fairness in the result, which is creation and deployment of a model. Consistent patterns and practices drive productivity and cost savings. They also enable trust in the models backed by on-demand reproducibility of the output.

The requirement of consistency is important because doing data science entails many moving parts. For example, data science practitioners require flexibility to use whatever tool or process is required for experimentation with model development.

Examples of flexible options may include integrated development environments (IDEs); languages; libraries and frameworks for machine learning and deep learning; data storage; and compute infrastructure such as distributed clusters. Other options essential for flexibility include support for the entire data science lifecycle: management, development, deployment, and monitoring.

Finally, practitioners also need strong integration capabilities to ensure process consistency, such as Git integrations, data integrations, security integrations, and so forth. A solo practitioner or a small team of data scientists can use whatever they want (or are able to persuade the IT department to provision for them) without impacting the larger organization.

Domino Accelerates Onboarding to About 20 Minutes

"The biggest benefit we've seen is to auto-onboard somebody into our data science analytics environment in about 20 minutes of a quick run through a tool, versus weeks it would take previously for somebody to get configured on their own."

-Senior Director of Decision Sciences, Software Services Industry

Flexibility and consistency are much bigger issues for large teams of data scientists who are simultaneously working on an array of complex problems aimed at enterprise deployment. For this scenario, flexibility can trigger pain points for data science leaders and IT professionals. Quality may suffer, projects may be delayed, support costs may increase, reproducibility of experiment results may be harder or even impossible, and other issues can thwart getting optimum results from data science. For example:

- Data scientists working on market problems may prefer SAS and have a library of models already available, while an operations team may prefer Python.
- Visibility across a portfolio of model research using multiple tools and processes becomes opaque and time-consuming when leaders look under the hood.
- Data scientists often spend too much time on DevOps work and not enough time on modeling research.
- Projects requiring significant computing resources such as a deep learning model may demand long delays for training and tuning. The lack of computing power may tempt the sacrifice of tuning to meet production deadlines, which risks model quality. Practitioners may have to wait days or weeks for provisioning.

All these issues are addressed by the emphasis on consistency in a next-generation data science workbench. With this capability, the workbench helps to reduce support costs; ease management with consistent steps along the data science lifecycle; enable repeatability of experiments to increase the potential for new breakthroughs, better quality, and discipline; and reduce the onboarding time for new data scientists.

Domino's Open Tooling Brings Flexibility to Top 10 Global Bank

"You can't hire a highly skilled data scientist without providing a state-of-the-art working environment. Otherwise, I would not have been able to set up a highly skilled team – nor would I have been able to set up the global watchtower of what is done in data science worldwide."

-Technical Leader, Top 10 Global Bank

Some organizations focus on consistency, but in the wrong way: standardizing on a single or reduced set of tools and forcing all machine learning operations (MLOps) processes to use only the approved resources. This approach is a big mistake! For example:

- The IT staff's solution for publishing models may only build in the use of R, which prevents the ability to deploy models built with Python.
- The organization may be unable to recruit data scientists who prefer other tools.
- Vendor lock-in may prevent the organization from tapping new innovations in the data science community.
- Data scientists may adopt rogue tools and processes, or corporate acquisitions may hit turbulence because a newcomer does not use what the acquirer sanctions.
- Infrastructure may be over- or under-utilized for one-size-fits-all environments.

For these reasons, the pillar of consistency must not compromise flexibility. Otherwise, the organization will risk creating new problems and technical debt and frustration may grow. When your company can provide consistency with the flexibility of a modern workbench, stakeholders can focus on creating business value while helping the organization to become model driven.

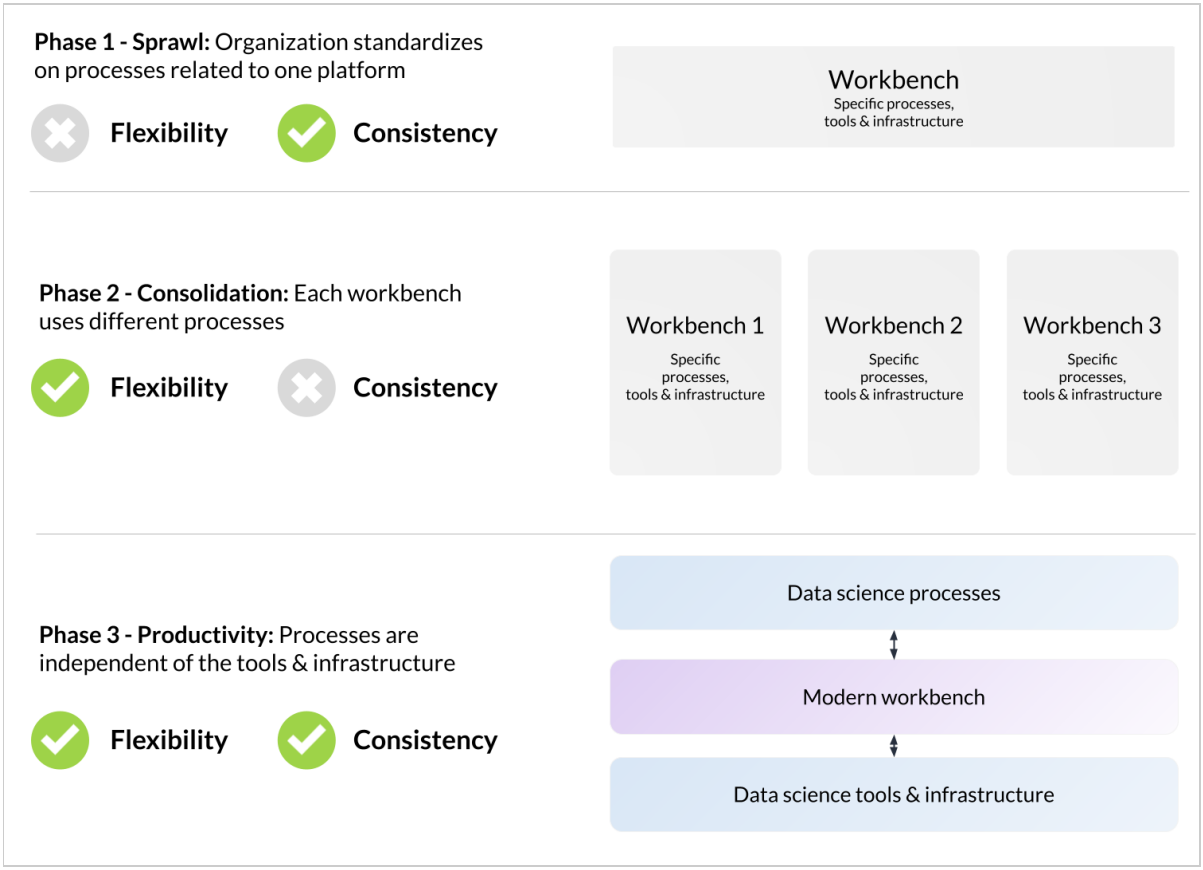

Figure 1. Enterprise options for scaling data science.

Workbench Features for Better Consistency

- Data connection management
- Extensible to new tools/infrastructure
- Flexible, browser-based workspaces
- Support for multiple languages and IDEs
- Support for analytics variety such as deep learning, machine learning, or extract/transform/load (ETL) data integration
- Kubernetes-based automatic selection and scale of compute with security and user/team guardrails
- Git integration
- Best practice documentation (processes, project templates, and pre-written code snippets)

2. Context for Collaboration and Knowledge Acceleration

Collaboration is a vital element of data science, where practitioners work with each other in conducting experiments and developing models. Data scientists collaborate because it can accelerate progress as hurdles are overcome faster. Different approaches to solving problems can trigger better problem-solving. Sharing knowledge can bring new and innovative ideas as practitioners learn from each other.

A caveat is that many tools for collaboration help with just one format: working together at the same time. This limitation is unhelpful for typical complex model development, which involves large teams of data scientists probing details of hundreds, even thousands of experiments conducted over many years – sometimes in different organizations and by researchers who long ago left for jobs with other institutions.

In these common scenarios, collaboration capability in a data science workbench must include the pillar of context, which injects complete understanding of all elements of every experiment and associated data, no matter how long ago or by whom the work was done.

For example, consider an analogy of a life sciences research lab with multiple teams of researchers working toward a common goal of developing an important vaccine. Think about how the individuals on those teams collaborate, capture, and share knowledge. Systematic processes will be in place to learn from failures and document even partial discoveries. Many efforts leverage research done years ago. Other teams, working on projects with some overlap, will have access to the research of their peers. Data science should act in the same manner, so a data science workbench must facilitate this kind of collaboration by adding context to everyone's work results.

The reuse of prior work for current model development is especially important to data science productivity. It shortens time-to-value by reusing prior code or prepared data sets. Reuse compounds value by combining ideas, such as merging the clustering of times series sales data with a sales forecasting model to improve accuracy. Reuse accelerates learning, such as looking at multiple approaches to image classification while formulating an approach for a new business problem. It lets practitioners leverage a failure, experiment, or partial discovery from a past project and apply those results to a new project. Reuse allows teams to benefit from past work started by employees who are no longer with the company. It also enables the elimination of duplicate efforts to keep researchers on track for faster results.

The value of reusing past work cannot be overemphasized. And, it is a productivity requirement not typically supported by workbenches due to the lack of built-in context.

Reuse of old work depends on the ability of a workbench to provide reproducibility without putting a damper on the productivity of data scientists. Reproducibility in data science looks slightly different than it does in other sciences. The process includes five elements: data, code, experiments, software, and hardware.

Collaboration Enabled by Domino: "Incredibly Useful"

"Whether you are a seasoned veteran or learning about new data and data science approaches in a complex R&D space, recognizing that there are hundreds of like-minded professionals who have dealt with the same issues is incredibly useful."

-Data Science Leader, Top Pharma Company

Data. Imagine a team receives an old dataset without background for understanding how each column is defined or whether any values were imputed. Is the dataset usable? Maybe. But without proper context, the team may need to create a new dataset for ensuring model accuracy.

Code. Data and code are closely related. It's critical to capture the data preparation and model building details. Code versioning and integration with code management tools like Git are critical workbench capabilities.

Experiments. Unlike legacy manual efforts, a workbench keeps track of everything done in an experiment including data, model types, and model hyperparameters that control the learning process. These elements may start simply enough, but when conducting hundreds of test permutations, the sheer number of combinations of inputs can quickly become staggering. Automating documentation with a workbench enables productivity: a researcher can fail faster, avoid losing time from work duplication, and find the optimal solution as quickly as possible.

Software environments used by data scientists are prone to package dependency issues, which can hamper or prevent reproducibility. Consider Python and R environments where typically there are hundreds of packages installed at the same time by research teams. Packages often have interdependencies; a new version of one package can trigger errors in others. Issues often arise

with the operation system, drivers, and frameworks. Consequently, versioning, tracking, and managing software environments is complex without a workbench. These software technical issues often cause data scientists to create multiple snapshots of environments (typically Docker environments).

Hardware. Reproducibility also extends to the compute environment. Would results be different by using a different number of cores, or a CPU instead of a GPU? Given the ability of many data science algorithms to use multithreading, and the importance of speed in model training, the probability of getting different results is high. Hardware specifications must also be documented to ensure full reproducibility.

A modern data science workbench is the solution for enabling next-level collaboration with the pillar of context. The workbench is a consolidated system of record for all data science model production and research. Teams should be able to easily see and find past and current work within your company. A workbench documents related artifacts for every experiment and iteration. Merely tracking revisions of code is not enough for collaboration. A modern workbench should provide all related artifacts, data, code, APIs, and other content to enable teams to easily understand context for every iteration of every project. A workbench also provides activity management and goal tracking for collaboration. It provides all stakeholders with full visibility into the goals, decisions, and assumptions made during any data science project

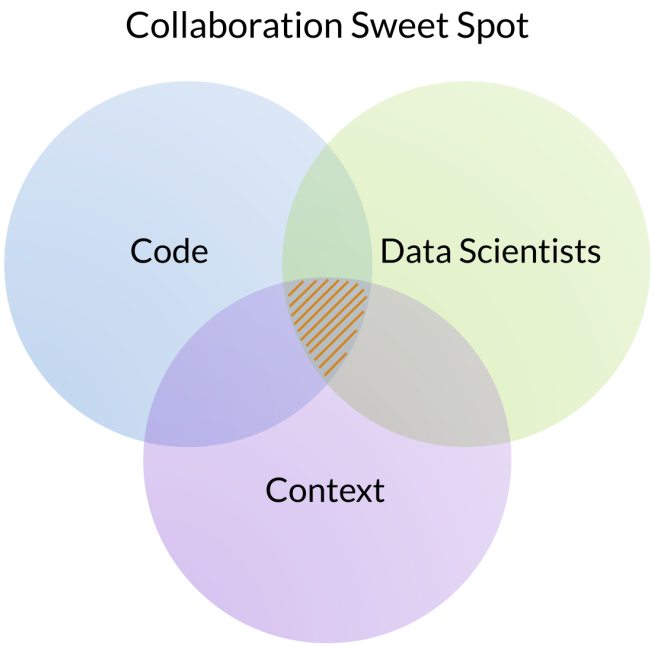

Figure 2. Data science productivity is enabled by three points of collaboration.

Workbench Features for Better Context

- Reproducibility - versioning, traceability of all related artifacts
- Searchable and shareable knowledge management across code, projects, models, etc.
- Environment creation, management, and sharing
- Context and history of work
- Collaboration on code and in conversations
- Experiment management
- Centralized tracking of key assets

3. Coordination of Projects to Solve Complex Business Problems

We have seen how consistency and context in a data science workbench are essential to scaling the work of data science teams. But far too often organizations fail to finish the swing on data science projects and the model never makes it into production. According to Gartner, "Through 2021, 75% of AI projects will remain at the prototype level as AI experts and organizational functions cannot engage in a productive dialogue."

The biggest culprits in this failure are a lack of coordination with the business and an inability to govern a large portfolio of projects as initiatives scale.

Project management. For the administration of data science projects, coordination of models with the business is top-of-mind in a model-driven business. Successful entities are using special-skilled people to help bridge the gap between data science and those who need to use its resulting models. As data science scales, the insights that a translator brings enable all stakeholders collaborating on actual work to see it at the project level and in "plain English." A workbench can capture rich metadata essential to project context. In addition to descriptive text and searchable tags, a workbench can allow the tracking of a data science project through stages with formal goals that act as gates between stages. The rich context can then be linked to completion of goals for better documentation.

Portfolio governance is becoming an important issue for data science used at scale. For example, the CEO of a major investment and insurance company recently asked the VP of Analytics a simple question, "How many data science projects do we have going on right now?" The VP did not know the exact answer but told the CEO it was "about 15." The CEO requested a specific number. As the VP conducted a manual survey, teams reported the company had more than 50 projects inflight. Oops!

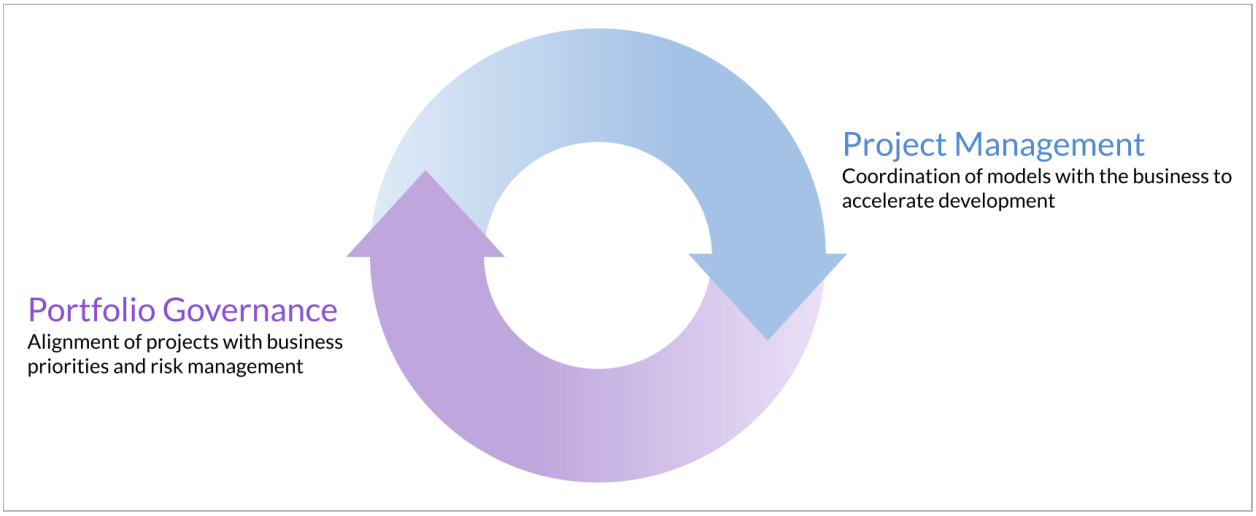

Figure 3. Data science workbench enables project management and portfolio governance.

What this VP needed was a view of the entire project portfolio. This workbench feature would have provided an aggregate snapshot of all the vital context captured for each project such as stages, owner, collaborators, number of projection assets, and so forth. Leaders could better coordinate people resources based on workload and project status.

A natural extension of a project portfolio view would be a similar snapshot of production assets – not just models, but any form of pipeline or application that results from a data science project. Tracking and monitoring these production assets by a workbench with strong linkage back to the original project provides a reproducible audit trail of data science work. As an additional benefit, the imposition of future regulations would be answered with the audit trail to document work for compliance. In a similar way, IT and data science leaders can track hardware usage to ensure meeting project cost control targets.

Workbench Features for Better Coordination

- Project portfolio overview
- Asset portfolio management
- Goals & validation flows
- Activity feed
- Project metadata, stages, and blockers
- Jira or other task management integration
- Resource and cost controls

Big Four Bank Auto-documents Data Science with Domino

A big four American bank outsources the prediction of macro market moves for its investment banking business. It requires the services firm to use Domino, which is the same governed, centralized data science workbench used by its in-house practitioners. In addition to code collaboration across time zones, Domino provides a way to test and validate models. The bank gets an auto-documented, step-by-step log of how the services firm built the assets the bank is deploying into production.

Conclusion

A model-driven business is dependent on a foundation of tools and processes that enable teams of data scientists to efficiently create and tune these engines of transformation. To meet this productivity challenge, data science practitioners need help with easing the use of any tool they need to do the job – and to align the efforts of teams such that business leaders can understand and rely on the results of predictive modeling.

Workbench Productivity: Top Feature Checklist

Consistency	✔ **Tooling** – flexible use of any tool or language
	✔ **Provisioning** – instant spin up of required hardware and software
	✔ **Data** – internal versioned stores and easy access to external sources
Context	✔ **Reproducibility** – versioning of data, code, experiments, software, and hardware
	✔ **Discovery** – enterprise search across metadata, project summaries, code, and deployed assets
	✔ **Collaboration** – provide teams with common access and communication tools
Coordination	✔ **Visibility** – portfolio-level tracking of projects and deployed assets
	✔ **Progress** – short-term recent activity feeds and long-term milestone management
	✔ **Governance** – usage, cost, and permission controls

The Domino Enterprise MLOps platform provides organizations with a modern data science workbench for efficient, rapid development of models. As a modern platform, Domino delivers on the three pillars required for data science at scale: consistency, context, and coordination. More than 20 percent of the Fortune 100 already rely on Domino to unify and accelerate complex data

science production initiatives. We invite your company to investigate how Domino can meet this productivity challenge without having to spend years of time and resources building these capabilities in-house.

A good place to start is the Forrester Research study, The Total Economic Impact of the Domino Enterprise MLOps Platform. Forrester found that Domino provides a 542 percent return on investment over three years, and pays for itself in under six months. To learn more about the benefits of Domino, please visit our website at DominoDataLab.com.

About Domino

Domino powers model-driven businesses with its leading Enterprise MLOps platform that accelerates the development and deployment of data science work while increasing collaboration and governance. More than 20 percent of the Fortune 100 count on Domino to help scale data science, turning it into a competitive advantage. Founded in 2013, Domino is backed by Sequoia Capital and other leading investors. For more information, visit dominodatalab.com.