DOMINO

WHITEPAPER

Attention, CIOs:

Do You Know Where Your Data Scientists Are?

How to rein in the wild west of data science to make it a core enterprise capability.

Contents

Executive Summary

Data science represents the next era of analytics driving the enterprise. Firms that capitalize on its potential will outcompete their rivals, increase efficiency, and generate new revenue streams. Today's CIO is challenged to centralize data science infrastructure in a way that will increase governance without constraining data scientists' freedom and flexibility. Failure to act will result in a "wild west" of siloed, inconsistent technologies sprinkled across the enterprise, operating beyond IT's purview and hindering the business's opportunity to drive value from its data science investment.

Successful CIOs will help their organizations move data science from the business's periphery to its core with structure and discipline that provide unbridled access to the latest technologies, visibility and auditability, and close alignment with the business.

CIOs that implement the right platform will deliver a win-win-win:

- IT achieves better governance while enabling innovation that unlocks new business value.
- Data scientists gain self-service and agility.
- The business earns a bigger return from its investment in data science.

Data Science: What and Why

Data science represents the next frontier for the data-driven business, which has been evolving for decades.

The **1980s and 1990s** saw the dominance of data storage, data management and data warehousing technologies, teaching companies the value of capturing and storing data to improve business operations.

In the **late 1990s**, business intelligence (BI) technologies became prevalent, making the insights captured by data management technologies more consumable by the business.

The **2000s** saw the "big data" boom with the rise of NoSQL technologies like Hadoop, presenting an open source, low cost approach to data processing and storage that made it plausible to keep full fidelity data, indefinitely.

This evolution of data management and analytics paved the way for data science, a term popularized around 2010, sometimes also called "quantitative research" or "decision science." Data science encompasses machine learning (ML), the computational process of making predictions based on data inputs and continually improving those predictions as data changes. ML is just one type of weapon in the broad arsenal of data science.

Data science at large blends statistics with computer science to find patterns in big data and use those patterns to predict outcomes or to recommend actions or decisions.

To build predictive models, data scientists rely on statistical programming languages like:

The modern enterprise uses data science to:

- Predict and reduce employee retention and churn
- Predict customer lifetime value and churn
- Improve lead scoring
- Optimize logistics, operations, and supply chains
- Build predictive features (e.g., recommendations) into their products to improve customer experiences

Data science will separate winners from losers

For decades, organizations have aspired to become data-driven. It took years to develop technologies that make it possible to efficiently capture, store and manage data from the systems that are instrumenting today's world. Now that the data is available, it can benefit every person and every department across the enterprise, which is driving fast and furious adoption of analytics and data science.

Data science is widely recognized as a discipline that should become a core organizational capability, with the potential to drive new revenue streams, automate decisions, improve products and enhance customer experiences to increase a firm's competitive advantage. This potential is driving significant investment from executives.

IT organizations have an opportunity to help companies realize the full potential of this investment by providing the infrastructure that helps make data science a core organizational capability, rather than a collection of siloed people and tools.

> The revolution in data science will fundamentally change how we run our business.
>
> Marc Benioff, CEO of Salesforce [1]

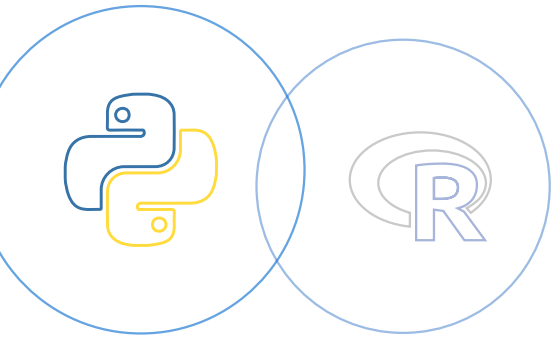
[1] Marc Benioff: "The Revolution in Data Science Will Fundamentally Change How We Run Our Business." Alexa Schirtzinger, Salesforce.com/blog

The CIO's Challenge

What's different about data science?

Previous generations of data technologies have involved centralized, monolithic components: a BI server, a database server, a data lake platform, for example. Data science work, in contrast, involves dozens of smaller tools and technologies, many of which are designed to be used locally on data scientists' workstations.

On top of that, these languages have rich ecosystems of "packages," which provide supplemental functionality for more specialized purposes. Many of these packages and tools are open source and available for download online, and data scientists regularly download dozens or hundreds of packages to use in their day-to-day work. And in the last several years, the open source ecosystem around these tools and packages has flourished, driving rapid innovation, frequent updates, and availability of entirely new packages every month. In other words, modern data science work lives across dozens or hundreds of clients, not in a centralized server.



According to a 2017 study by KDnuggets, the most popular languages for data science are Python and R. [2]

A "wild west" of data science

Data scientists, eager to stay on the cutting edge and utilize the latest techniques, experiment liberally with a variety of tools and packages. That pace of experimentation is increasing as the open source ecosystem innovates more rapidly. The combination of client-based work, a large number of easily accessible technologies, and a desire for rapid experimentation has created a "wild west" of data science tooling in most organizations. Inconsistent technologies are spread across disparate parts of the organization without governance or transparency around any of them.

Worse, in many organizations, "shadow IT" is cropping up to support these systems. For example, a small team might install RStudio or Jupyter (both free downloads) on a shared server to use for their group, without considering support requirements or consistency with other parts of the organization.

New Leader, Trends, and Surprises in Analytics, Data Science, Machine Learning Software Poll, Gregory Piatetsky, KDnuggets

Beyond the obvious problems, this "wild west" of siloed data science work creates several other issues:

- **Important business processes become dependent on unreliable infrastructure.** Data scientists will often set up scheduled jobs to run on their own local machines, or operate shared servers as "lab" or "dev" machines. One Fortune 10 bank had a critical business process that depended on a model a data scientist had been running nightly on his laptop — only to be discovered when he left and the laptop was decommissioned.
- **Compute costs can become excessive and uncontrolled.** Unlike BI, data science involves computationally intensive techniques, which demand high-powered machines and specialized resources like GPUs. Especially in a cloud environment, data scientists in the wild west can unintentionally burn thousands of dollars a month by leaving expensive machines running unnecessarily.
- **High-value intellectual property is improperly secured.** Predictive models and analyses can encapsulate insights key to competitive advantage, and that work is often scattered throughout network drives, wikis, or Sharepoint sites.
- **Data scientists waste time on DevOps work.** Data scientists are precious, highly paid people, yet they often must spend 25% of their time dealing with DevOps tasks like installing packages and moving files between machines.
- **Data scientists waste time duplicating effort and reinventing the wheel.** Beyond individual data scientists wasting time on DevOps, entire teams can waste time pursuing projects that reinvent the wheel or don't build upon past organizational knowledge, because that past work was siloed and undiscoverable.

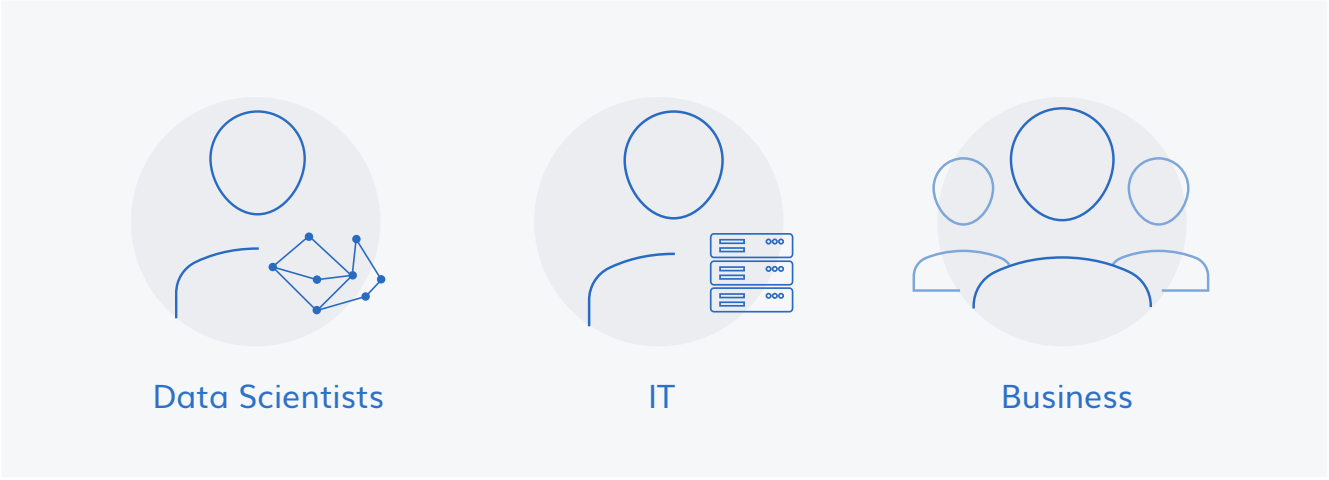
Keep an ear out for mention of these tools

People or teams using them may be creating a siloed ecosystem outside of any standards or governance. Sometimes they can grow completely under the radar of an IT organization: one large insurance organization hadn't realized they had 30 people in their HR department building models in R!

A false dilemma

Data scientists will err on the side of innovation, driven by a desire to use the latest technology and largest machines to develop better models faster than competitors. Unlikely to perceive the medium- and long-term consequences of a lack of standardization and governance, like water flowing around rocks in a river, they will find the path of least resistance. If IT isn't offering them what they need, they will find workarounds and unintentionally put the organization at risk over the long run.

It's natural, but overly simplistic, to view the situation as a trade off between innovation and safety/security. That framing binds the CIO between stifling business progress and competitiveness, or endorsing chaos and risk. This framing is a false dilemma and misses an opportunity to align the goals and incentives of stakeholders across the business.



The CIO's Opportunity

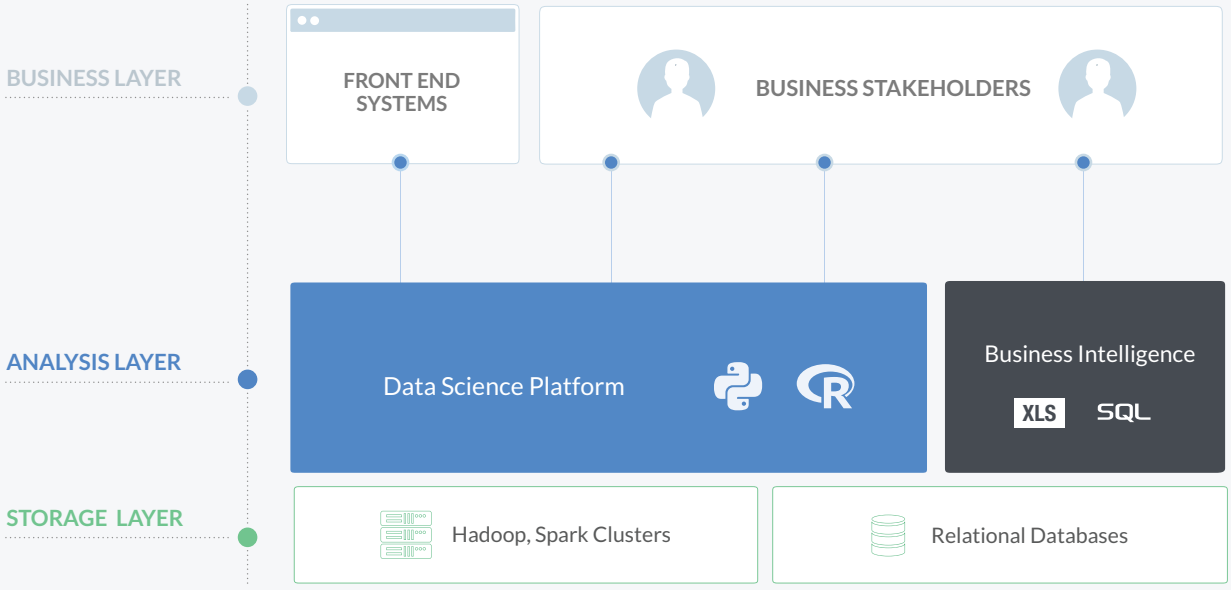
Within the challenges above lies a tremendous opportunity to bring order to chaos while enabling a critical business transformation. It's a pivotal point in many organizations' journey toward becoming truly data driven, and if built correctly, an effective data science function will transform every business.

Creating centralization and governance

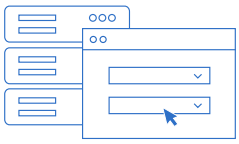
If databases and data lakes were the central architectural components of incumbent generations, the foundational technology for the data science era is the data science platform. Unlike a database, a data science platform doesn't house your data—instead, it houses the artifacts and work product associated with data science workflows.

Just as sales organizations use a CRM to create maturity and scalability, and engineering organizations use version control, enterprises are deploying data science platforms to create more maturity and discipline around data science work.

Where a data science platform fits in your infrastructure

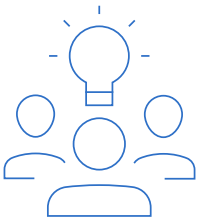


Data science platforms allow IT organizations to rein in the wild west of data science tools, assets and infrastructure spread across the organization. Instead of working in disparate local environments, data scientists do their work in one central place. In order to support the range of use cases involved in data science work, an effective data science platform will provide:



Self-service infrastructure, so data scientists can do exploratory data analysis and model development without configuring and using their own compute resources. The data science platform encompasses compute resources—as well as the languages, packages and tools necessary for modern data science work—with controls and reporting around resource usage to administer or attribute costs.

Ways to deploy, productionize or operationalize finished models, instead of driving data scientists to set up shadow systems. This includes deploying models to power scheduled jobs, reports, APIs or dashboards in one place. The data science platform also provides a consistent baseline of non-functional requirements (security, HA, etc.) and a catalog that offers transparency into assets and utilization across the enterprise.



Governance, collaboration and knowledge management around all the artifacts created in the process of the research and deployment work described above.

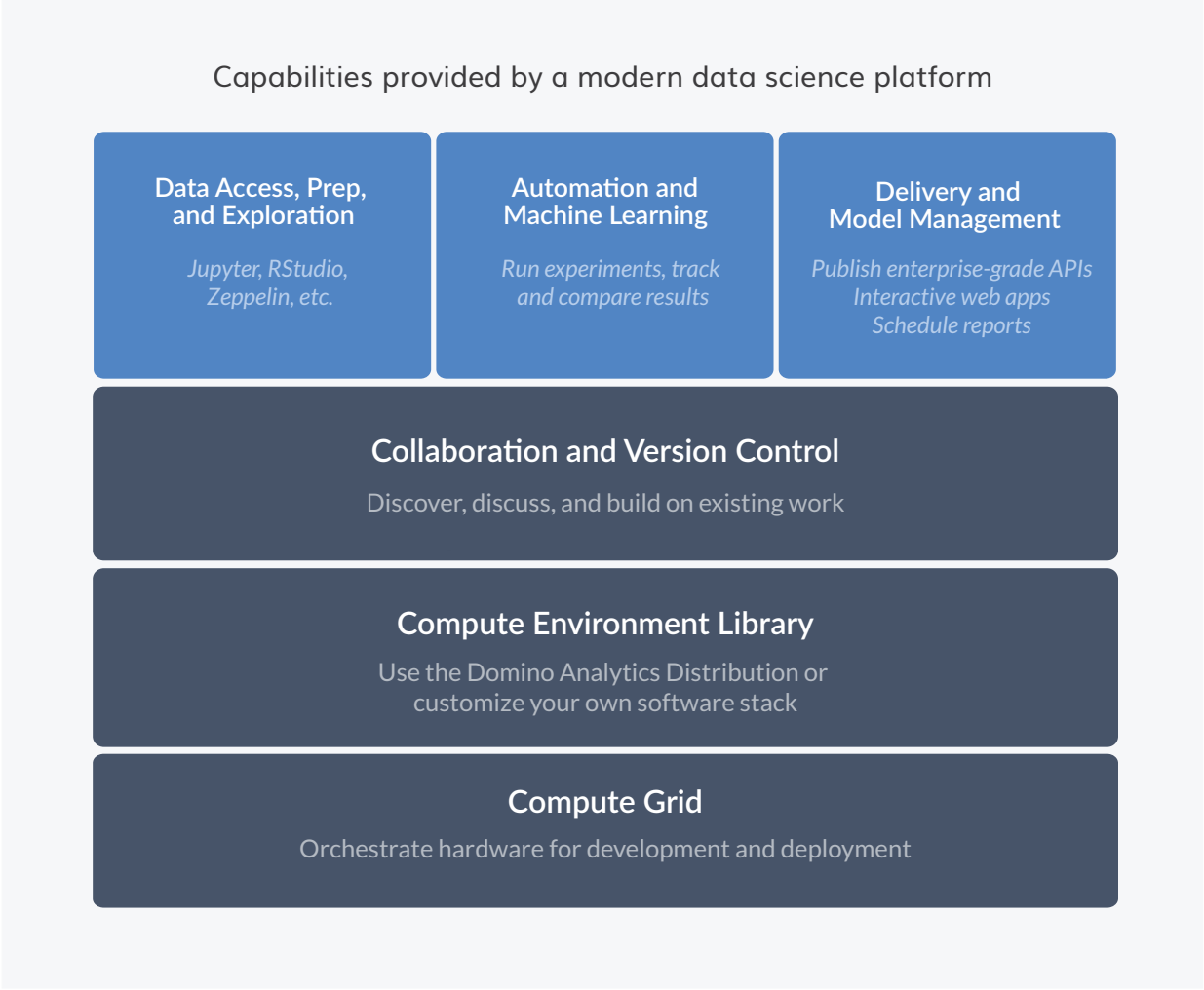
Winning in the cloud

Data science workflows are ideally suited for the cloud, because they benefit from burst compute and specialized resources like GPUs.

- Elastic compute and access to GPUs aligns with the lumpy workloads of model development cycles.
- Giving data scientists self-service cloud access via a data science platform alleviates DevOps work and enables automatic elastic compute, which they'll love.
- A data science platform in the cloud provides IT with cost controls, resource tracking and reporting.

Moving data science work onto a centralized platform will ensure that:

- Any model or analysis involved in a business process is centrally persisted and monitored, even if the original creator leaves the organization.
- Data scientists work from consistent, standardized tools, reducing support burden and operational risk.
- All data science assets are permissioned, and those permissions are auditable.



Aligning with stakeholders across the business

Implementing a data science platform to centralize data science work will reduce risk and support burden for IT organizations. But getting buy-in from other parts of the organization—especially data scientists who are likely to balk at talk of "governance"—will be critical.

A key part of the CIO's challenge is delivering effective, tailored communications to different stakeholders; rallying the troops to align behind a shared goal for successful data science. Doing so requires empathy to understand the unique motivations and perspectives of different constituents. Fortunately, there are a wide variety of benefits that can be communicated to align interests.

Data Scientists

Priority is to innovate as quickly as possible by taking advantage of the best and newest tools in a self-service environment:

- Promote the benefits of self-service environments for data science that will allow them to independently provision infrastructure, spin up workspaces with their tools of choice (e.g. Jupyter, RStudio) and safely experiment with new packages and tools. They won't waste time doing their own DevOps work and they won't need IT support.
- They can run experiments faster and collaborate with others in the same place they're doing development work, saving time that would otherwise be wasted reinventing the wheel.

Buy vs. build

Inevitably, enterprises will consider building their own data science platform, either because they believe it will be cheaper, or because they believe their environment is so unique that it requires a custom solution.

Before heading down this path, consider several costs associated with a homegrown solution:

1. **Opportunity cost and comparative advantage.** Your models are your differentiation, not the platform you use to develop them. What could you build instead with your engineering resources?
2. **It's harder than you think.** A data science platform combines infrastructure orchestration, sophisticated workflow and UX, and capabilities for production-grade deployment. Many companies have spent over a year trying to build their own before giving up to pursue a third-party solution.
3. **You'll be making a permanent commitment of resources to ongoing support and maintenance.**

You haven't built the CRM system that your sales team uses, or the version control system that your engineers use—a data science platform is no different.

Executives

Priority is to derive return from investments in data science by quickly integrating insights to improve business processes:

- Promote the concept of a data science "system of record", akin to the function a CRM fulfills for a sales organization. It centralizes all workstreams and communications between data scientists and other business stakeholders in Engineering, IT and Compliance, facilitating a more mature, predictable, scalable way for data science teams to deliver value.
- Faster experimentation will lead to more data science projects and research breakthroughs completed faster.
- Easier ways to operationalize or deploy models will reduce the time from insight to impact, turning data science work into realized business value at a faster pace.
- The flexibility to accommodate modern tools and technology to data scientists will help to recruit top talent in a competitive field.
- Automatically maintaining a complete audit log of every model's development will reduce operational and regulatory risk of algorithmic decision-making.

IT Organization

Priority is to control infrastructure costs and maintain a single, integrated environment:

- Promote the idea of an infrastructure orchestration platform that integrates with existing systems and tools, offering real-time scoring, batch scoring and app hosting options.
- Risks and issues can be proactively identified by tracking hardware, tools usage and changes to production models.
- Usage of expensive compute resources (especially in a cloud environment) can be more easily monitored, limited and attributed.

IT

> Centralization of data science work increases governance and reduces risk.

Business

> Data science is delivering more value faster, fulfilling the promise of our investment.

> IT has been a strategic partner enabling us to integrate data science as a core business capability.

Data Science Platform

Data Scientists

> Self-service infrastructure and tool agility let us deliver projects faster and stay on the cutting edge.

By successfully navigating each internal stakeholder's concerns and deploying a data science platform, everyone wins: IT management successfully mitigates risk through governance and centralization, while delivering productivity gains for data scientists. Establishing a data science platform leaves IT poised for success, and the business is equipped to drive faster innovation.