Building a data science team at the California Department of Tax and Fee Administration: A modern approach to solving data-rich analysis problems

Monica G. Bobra[1], Reid Johnsen[2], Jared Choy[2], Sarah Letson[1], Rijin Baby[2], Kimberly Hicks[1], Genti Droboniku[2]

[1] California Office of Data and Innovation, 401 I Street, Ste 200, Sacramento CA 95814
[2] California Department of Tax and Fee Administration, May Lee State Office Complex, 651 Bannon Street, Suite 100, Sacramento, CA 95811-0299

Introduction

Departments across the State of California rely on extracting insights from increasingly large datasets to address their business needs. This is particularly true at the California Department of Tax and Fee Administration (CDTFA). Researchers at CDTFA build statistical models to efficiently and effectively collect revenue, assess the economic environment, and forecast the impact of policy changes on various populations and communities throughout the state. In late 2025, CDTFA made a once-in-a-generation investment in a data science team to help with these efforts. This paper describes how the Office of Data and Innovation (ODI) worked with CDTFA to build the strategic case for this team.

Data science in the State of California

The field of statistical modeling changed considerably 10 years ago, with the rise of large datasets, open source scientific software, and access to on-demand computing. Together, these three developments enabled a new field of research called data science (Donoho, 2017). Data scientists, when enabled with the appropriate resources, can quickly build and deploy models designed to derive insights from large datasets.

Some California state departments, however, struggled to capitalize on these developments. The 2026-2027 California Statewide Data Strategy, which is led by ODI, outlines a few reasons for this struggle: legacy systems and outdated tools, a lack of common data standards, few data sharing agreements at scale, workforce gaps for skilled data scientists, data engineers, and data policy experts, and no place for data practitioners to find a sense of community and share lessons learned. ODI's mission aims to address these struggles for departments across the state through accelerator services, designed to help departments rapidly prototype models using modern infrastructure, as well as advising, training, and community development.

How ODI and CDTFA built a strategic case for a data science team

In order to prove the value of a data science team, ODI and CDTFA developed and executed a five-step plan. The goal was not only to secure executive buy-in, but to build trust and a culture of collaboration across the organization.

1. **Clearly identify the problem.** CDTFA could solve a wide variety of problems using advanced statistical methods like machine learning. One example of an ad-hoc analysis includes identifying and correcting customer registration errors. CDTFA also builds processes that are re-run on a schedule, such as identifying potential taxpayer noncompliance. Building a robust model that runs efficiently and routinely requires data pipelines that can reliably process large quantities of high-quality data and models that are easy to understand, test, and maintain. Without investing in data pipelines and software, organizations risk acting on erroneous predictions and accumulating technical debt (Avgeriou, 2016). Furthermore, to scale projects effectively, organizations must improve their tools and infrastructure, rather than limiting their ambitions to what their current systems can support.
2. **Clearly identify the stakeholders.** Changes in a process or organizational structure affect multiple stakeholders. ODI and CDTFA together identified our stakeholders early and listened to their concerns. This process helps with building trust across the organization. The executive team, the IT team, and the Research and Statistics section are three critical stakeholders for data science work at CDTFA.

3. **Clearly communicate the problem.** Data science blends technical skill with the art of storytelling. ODI and CDTFA together reached out to multiple audiences and openly communicated the problems with our current approach, using their language and acknowledging their constraints. This method helped build a culture of collaboration. For example, CDTFA secured buy-in from the resource-constrained IT department by aligning with their incentives. Best practices in code development and data engineering increases compliance with state policies, security, and privacy. This reduces IT workload and, ultimately, makes their lives easier.
4. **Build a proof-of-concept.** ODI and CDTFA together identified a critical and timely problem to solve as a proof-of-concept using current resources and infrastructure. Together, we solved this problem using the data science life cycle (Yu & Barter, 2024) by identifying the business need, conducting user research, building a data pipeline, prototyping and validating interpretable models, and generating model outputs best suited to actionable decision-making.
5. **Attach metrics to the proof-of-concept at scale.** ODI and CDTFA created a process map that clearly identifies each step and stakeholder in the current process. To motivate a change, we compared the current process to our new one and identified simple, measurable outcomes. Some particularly resonant outcomes included saving staff time, making more timely decisions, and generating more revenue.

The data science team

CDTFA successfully hired its first data science team to improve organizational efficiency and decision-making. Recognizing the challenges of creating a new unit in an organization, CDTFA housed the team within the Research and Statistics section. Buy-in from Research and Statistics was essential to quickly establishing the data science team.

The CDTFA data science team now consists of five employees:

- A principal data science lead, who leads the team,
- A data scientist, who is responsible for advanced analysis work,
- A data engineer, who is responsible for building data pipelines,
- An analytics engineer, who is responsible for collecting business requirements, define successful analytics outcomes, and supporting data scientists, and

- A software engineer, who is responsible for delivering production-grade code that adheres to best practices.

How the CDTFA data science team operates

The CDTFA data science team adheres to four foundational principles designed to facilitate collaboration, build trust within the organization, and enable robust, reliable, and reproducible decision-making.

1. **Intentional design.** The team set standards and processes for software development, cloud computing, and model validation. This requires high-level thinking beyond addressing an immediate need. The team also created a Developer's Guide, designed to formalize these practices within the team and onboard new members, and a living roadmap to clearly communicate the team's long-term vision to executives.
2. **Continuous development.** Model architectures can improve, and more data almost always leads to better results. Instead of creating and deploying a product once, the team strives to continuously improve products. The progress-oriented mindset helps keep work grounded and gives projects room to breathe. The team embraces the California Agile approach to project management, where the results of one phase of the project inform the next, rather than a traditional waterfall approach.
3. **Trustworthy, interpretable, and reproducible models.** A trustworthy model requires much more than strong predictive performance. It requires reliable data inputs, appropriate validation, documented assumptions, and a clear understanding of model limitations. Transparency and reproducibility throughout the modeling process allows others to review, understand, and improve upon model results. When model outputs inform operational decisions, they should complement subject matter expertise and human judgment rather than replace them. Interpretable models help us, as government workers, explain and defend our results to many stakeholders, including the general public.
4. **Personal development.** The team closely follows the continuously changing landscape of advanced statistics, including machine learning and artificial intelligence, by attending conferences and workshops and connecting with data scientists, data engineers, and data policy experts across the state through the Office of Data and Innovation's Data Community of Practice.

Conclusion

Together, ODI and CDTFA developed a successful model for building a flexible data science team at CDTFA capable of using large and rich datasets to solve complex problems. Developing this five-person team required a strategic approach and some endurance. Together, we found that investing in a robust foundation at the outset enables better results during times of need. We can likely generalize our lessons to other organizations, but we recognize this is one of many ways to build a team. We're grateful to our teams at CDTFA and ODI and many other individuals for the time, support, and energy they invested into building this team.

> "One of the biggest lessons we learned from building our data science team was the importance of starting small and establishing trust. Setting clear, achievable goals helped our team demonstrate value early, while at the same time creating the foundation for larger, more impactful work. The technology matters, but it is the people, the relationships, and the shared sense of purpose that make lasting change possible."
>
> — Genti Droboniku, Chief Deputy Director, CDTFA

References

Avgeriou, P., Kruchten, P., Ozkaya, I., & Seaman, C. (2016). Managing technical debt in software engineering (dagstuhl seminar 16162). Dagstuhl reports, 6(4), 110-138, DOI: 10.4230/DagRep.6.4.110

Donoho, D. (2017) 50 Years of Data Science, Journal of Computational and Graphical Statistics, 26:4, 745-766, DOI: 10.1080/10618600.2017.1384734

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5), 206-215, DOI: 10.1038/s42256-019-0048-x

Yu, B., & Barter, R. L. (2024). *Veridical data science: The practice of responsible data analysis and decision making.* MIT Press.