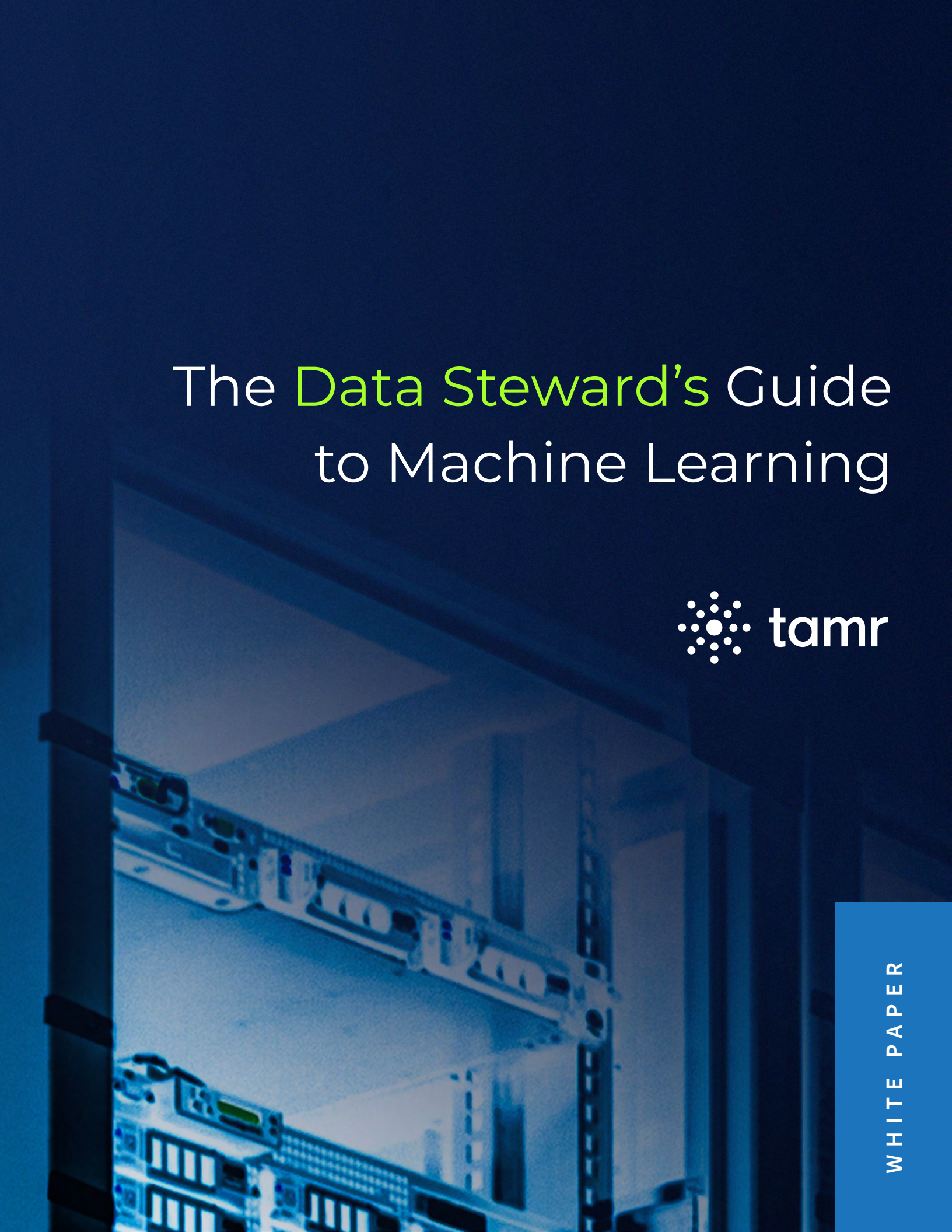
The Data Steward's Guide to Machine Learning



tamr

WHITE PAPER

Table of Contents

Introduction

Leading-edge consumer technology companies, such as Google, Amazon, and Netflix, have demonstrated the impact that machine learning can have on the customer experience. These brands have become some of the most valuable in the world by delivering experiences that feel magical to the end consumer by using machine learning to make helpful recommendations, tag pictures, and translate documents. They've also made machine learning top of mind among executives at enterprises across all industries who recognize the need to adopt it to avoid being disrupted.

As consumers, we're primarily aware of how machine learning impacts the 'last mile' aspects of the customer experience. But this technology is also readily applied to all areas of business operations. Data stewardship, an often ill-understood but vital part of DataOps, is one such area. Let's first define what we mean by "data stewardship" then discuss how machine learning can be used to increase the effectiveness of data stewards.

Data Stewardship Defined

As a data steward, you sit between raw data sources and data consumers, which include data scientists, data analysts, and business professionals. You are ultimately responsible for ensuring that data is well-managed and well-understood. This includes creating data dictionaries, monitoring and improving data quality, establishing governance, and defining the procedures required to meet security & privacy requirements.

Organizations with a formal DataOps function may have people with the explicit title of "data steward". The majority of companies, however, have people (or teams) operating in this capacity without the formal title. Regardless of who assumes these responsibilities, enterprises are required to address the associated challenges to effectively make use of their data.

Areas Where Machine Learning Can Help

There are a range of applications for machine learning but at its core, it works great for pattern recognition. The best machine learning problems are those where enough data exists for patterns to emerge. Data volumes don't need to be massive—machine learning can be applied on hundreds of records for simple problems—but they do need to be large enough for patterns to exist.

The best machine learning problems also have a clear outcome. You should not expect machine learning to answer questions that aren't being asked, but you should expect it to identify patterns and provide insights that are not readily apparent. Sample applications of machine learning include:

- **Classification:** Bucketing items into defined categories (e.g., "what type of data is this?", "what category of product was sold?")
- **Prediction:** Predicting a future outcome based on historical data (e.g., "who will win the NBA Finals?")
- **Optimization:** Determining the best allocation of a scarce resource to optimize a specific outcome (e.g., route optimization to minimize fuel costs, pricing optimization to maximize profits)
- **Clustering:** Grouping together similar data points (e.g., customer segmentation, recommendation systems)
- **Anomaly detection:** The opposite of clustering—identifying data points that fall outside of expectations (e.g., fraudulent transactions, malware)

Applying Machine Learning to Data Stewardship

The ratio between data consumers and data stewards is often significant. Large enterprises may have thousands of data consumers for every data steward. As a result, data stewards spend a significant amount of their time identifying patterns within data to determine how they should prioritize their time and what fixes they should implement.

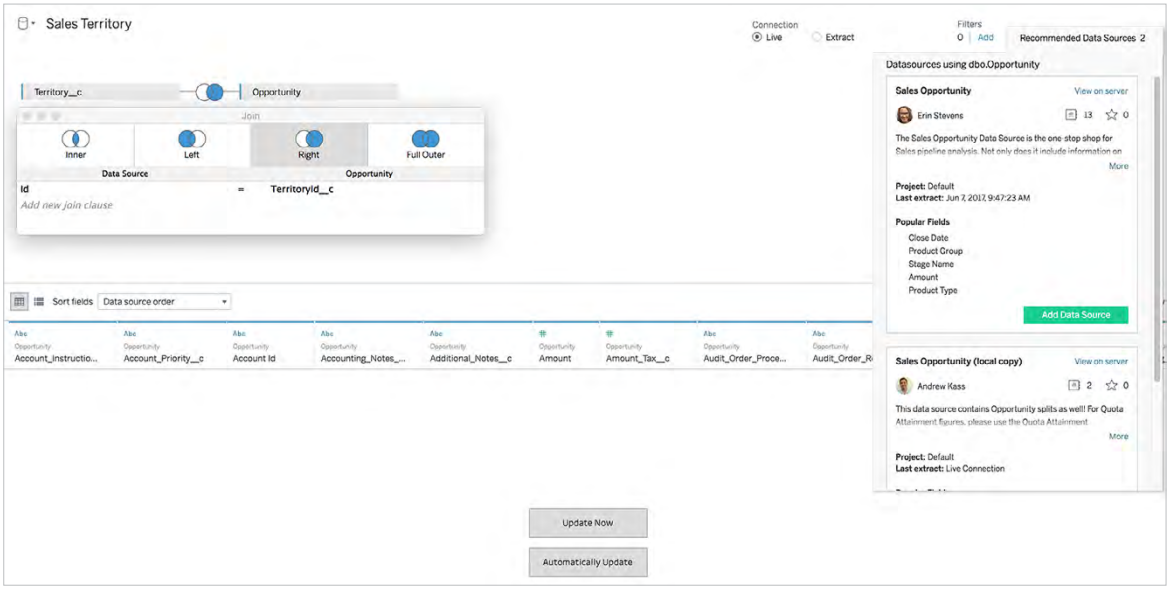
This is what makes data stewardship such a ripe area for machine learning. The amount of data available to data stewards is significant, and it is impossible for them to keep up with the demands of their data consumer counterparts.

We've listed 5 ways that machine learning can have a big impact on data stewardship, but many more exist. Our recommendation is to get started in one or two areas to gain comfort with the technology and deliver quick wins so that your organization will buy into adopting it more broadly.

1. Identifying Data Sources

The data sources used by an individual consumer are often largely driven by their function. For example, people in sales and marketing are likely heavy users of CRMs and marketing automation tools. These patterns can be inferred by monitoring the data sources often used together.

Applications like Tableau have realized this and created "Data Source Recommendation" features that point users in the direction of other data sources they may find useful. These features are also becoming common in data catalog tools. Data stewards can leverage these applications to better understand consumer behaviors and unify data sources as needed to ensure seamless usage of disparate datasets.

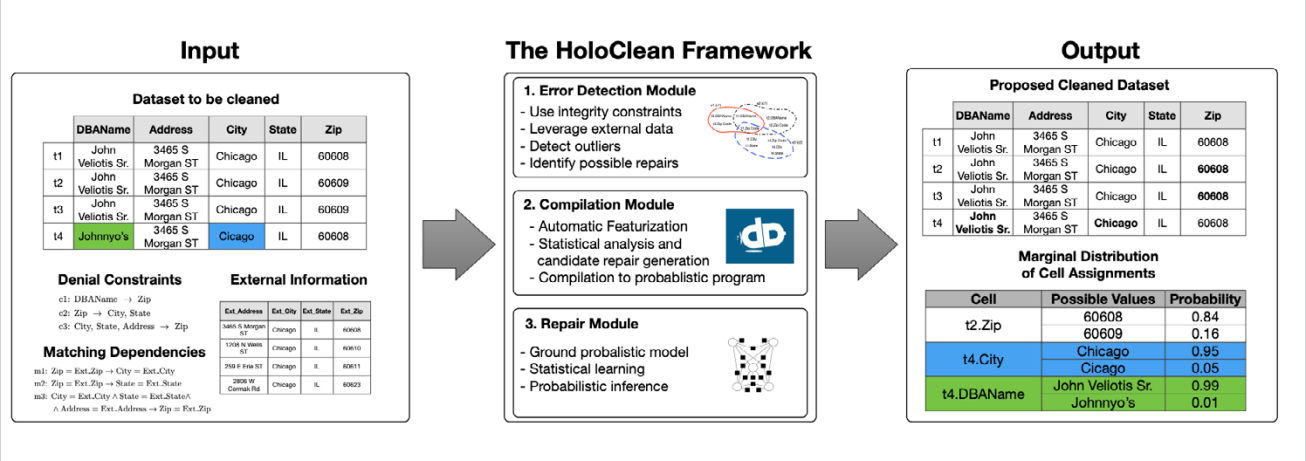


Tableau's data source recommendation feature
Source: Tableau

2. Fixing Data Quality Issues

Many data quality issues—such as missing values or erroneous information—can be corrected through anomaly detection. For example, if 10 records about “Jane Jones” have her address listed as “123 Maple Lane”, but an 11th record has her address listed as “123 Mpl Ln”, we can infer that this is a mistake and the address should be corrected.

New tools, such as HoloClean, are now available and leverage anomaly detection to repair datasets and fix data quality issues. This is a useful component of a large scale data pipeline, where companies will often have monitoring in place but lack the “auto-healing” capabilities necessary to resolve the overwhelming number of issues that arise during daily operations.



HoloClean framework for using machine learning to fix data quality issues
Source: HoloClean

3. Mapping Datasets to a Schema

Acquiring new data sources is a necessary part of doing business now. Often, these new data sources come from external sources and loosely governed internal sources, such as Excel, introducing a variety of new schemas. This isn’t a problem when you’re acquiring one or two new sources every year, but for businesses acquiring dozens of new sources every month, manually mapping all of these sources to a fixed schema quickly becomes an expensive endeavor.

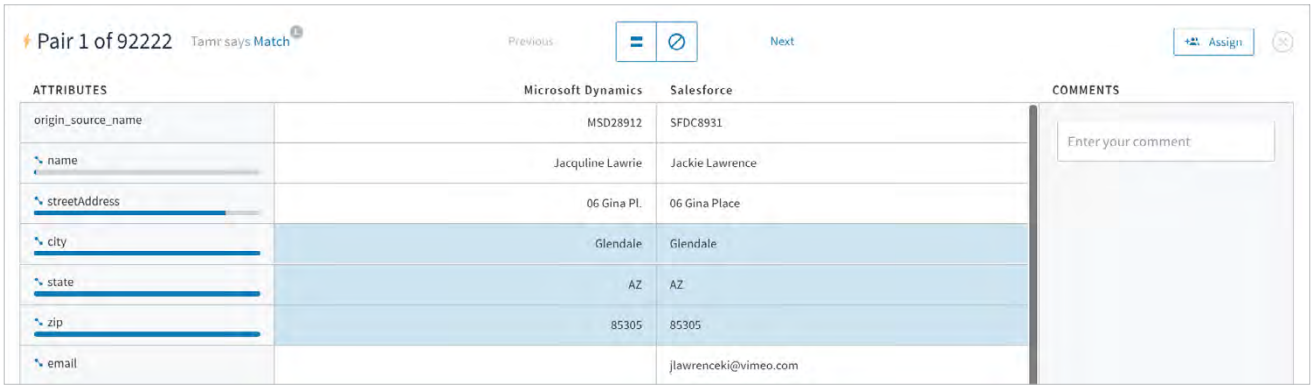
Luckily, this is a great problem for machine learning. A model can be trained on how to classify attributes based on the profile of the underlying data and the attribute headers themselves. This model can be applied to new or unmapped sources, providing recommendations for how these sources should map to a target schema.

4. Clustering Similar Records Together

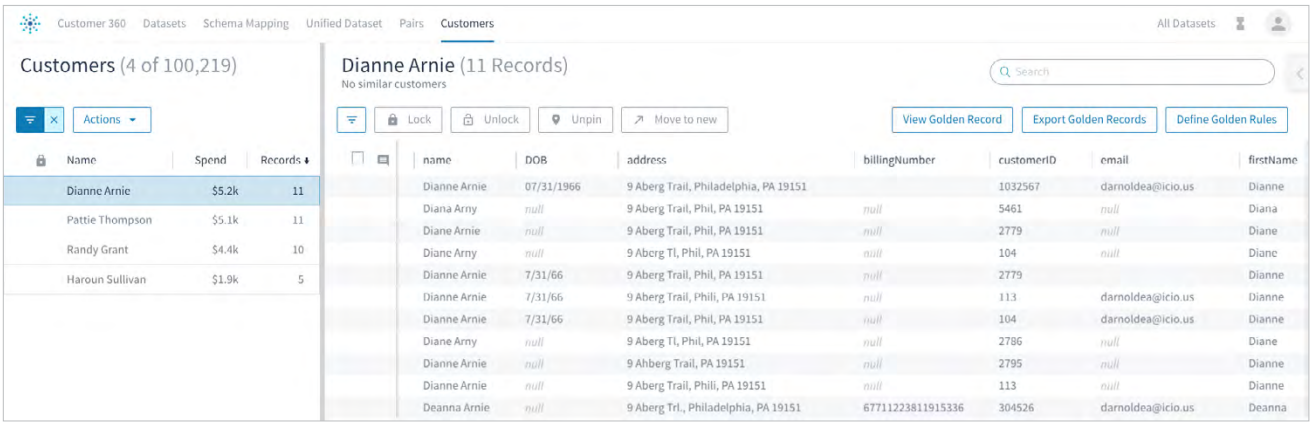
Clustering is one of the areas where we've seen machine learning have the greatest impact. Machine learning does an excellent job of taking a random sample of feedback or training data and extrapolating that out across an entire dataset. This works extremely well when the goal is identifying similar or duplicate records, such as all of the instances of a single customer or all the records that belong to people living in the same household.

Random forest algorithms work particularly well here, learning from simple feedback on a subset of data (e.g., "Is John Walker the same person as Jonathan Walker?") to be able to make good decisions on all of the data. Taking this approach, we've seen companies achieve results such as mastering / de-duplicating a list of 25M customers, representing all parts of Europe, by reviewing a sample of fewer than 2,000 records.

No data stewardship program can be considered viable without a meaningful solution to mastering, and for any large enterprise, this means getting serious about machine learning. This was one of the primary problems Tamr Unify was developed to solve, along with attribute mapping and record classification.



Training a model to cluster records in Tamr Unify
Source: Tamr



Clustered records in Tamr Unify
Source: Tamr

5. Classifying Records

We previously discussed how machine learning can be used to map datasets to a target schema; i.e., classify dataset attributes. Those same techniques of using a machine to learn the relative significance of individual or groups of words to a classification category can be applied when classifying individual records as well. Anyone who uses a corporate expense reporting tool, such as Expensify, has likely interacted with these types of classifiers.

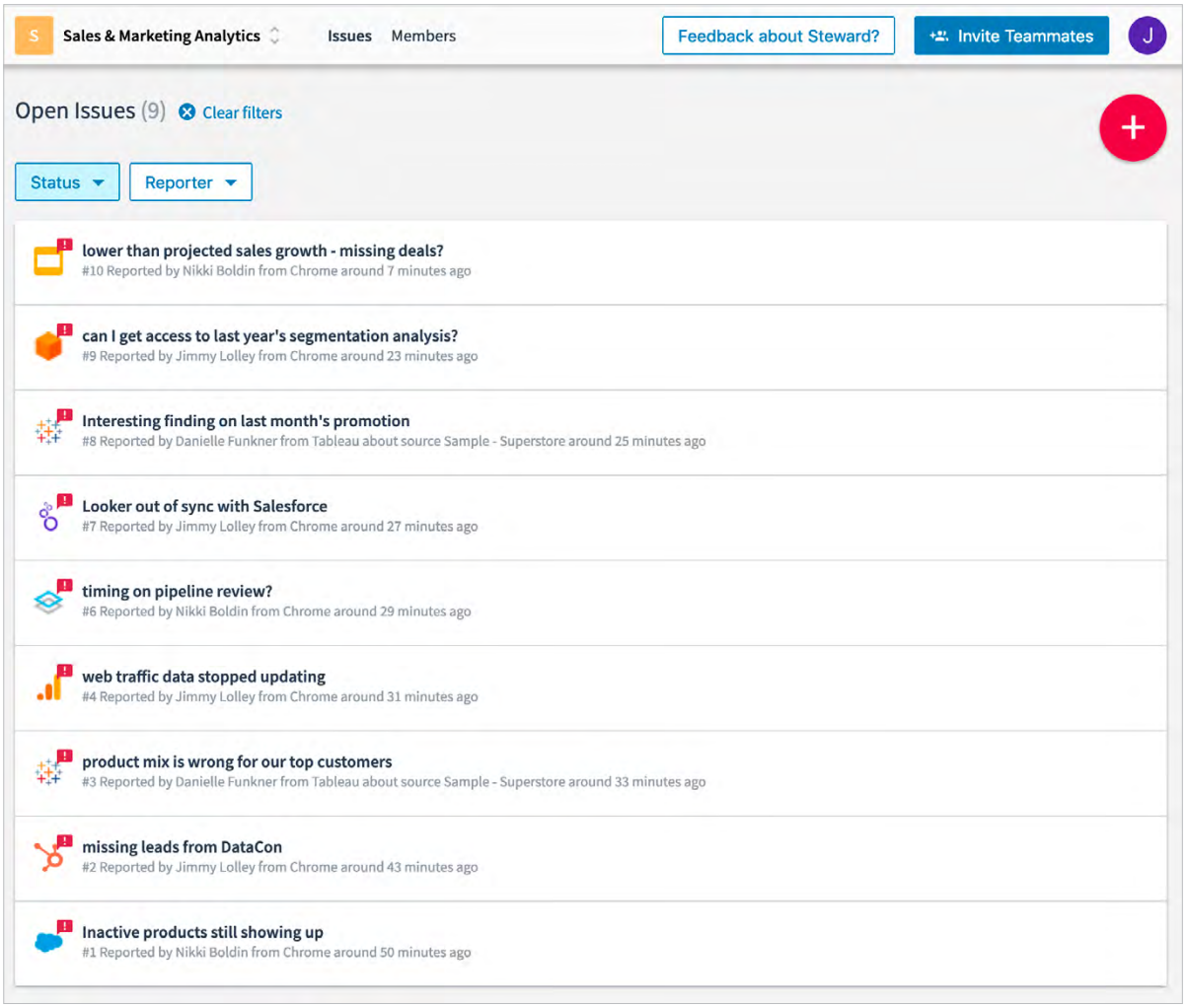
Tools like Tamr Unify and MonkeyLearn help with this problem, along with many open source machine learning libraries such as Scikit-learn, a machine learning tool for Python. It's important to consider how often your taxonomy and data will be changing when choosing a tool for this problem. If your taxonomy and data are subject to constant change, you should select a tool that allows you to easily introduce new categories and retrain the model.

Avoiding Pitfalls

One of the biggest roadblocks we see to customers being successful in applying machine learning to data stewardship is that they don't know what problems to prioritize. There is never a shortage of anecdotal evidence about the problems that exist within a company's data, but rarely can someone quantify the data challenges that are impacting the most number of consumers or the highest value activities.

A key reason for this is that data quality assessments are still largely top-down activities, such as data profiling, done outside of the context of data usage or value. We recommend data stewards get closer to data consumers by instrumenting ticketing and feedback systems that allow consumers to raise a flag when they have a question about a piece of data or identify a data quality issue. Tamr Steward, Jira, and Asana are all applications that we see being used broadly in this capacity.

This can not only give data stewards insight into specific issues but also paint an accurate picture of who is consuming what data and how they are using it. Armed with this information, the data steward can make a better assessment of what actions need to be taken to improve the value of their data.

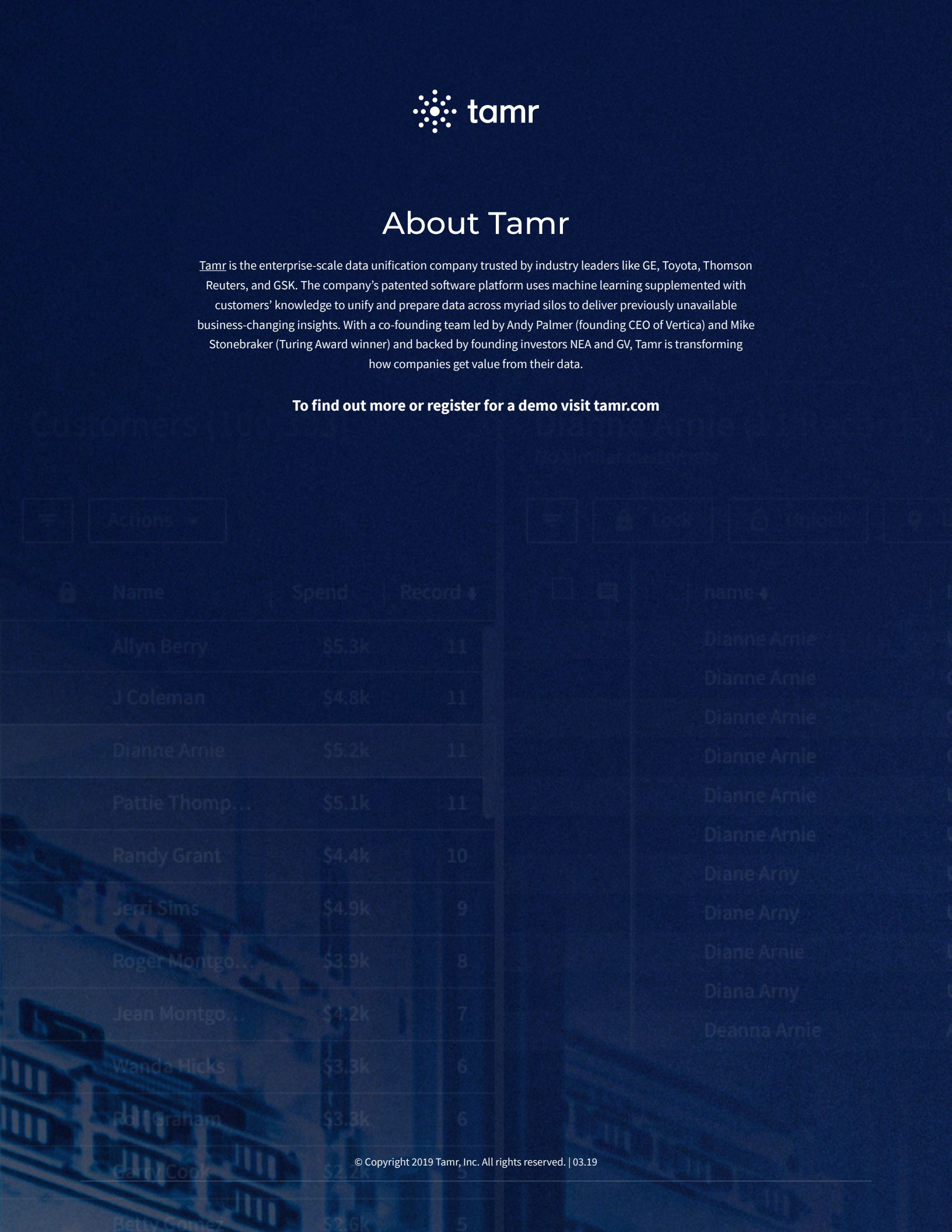


Overview of open tickets within Tamr Steward
Source: Tamr

Getting Started: Think Small

You don’t need to boil the ocean to start seeing value from applying machine learning to data stewardship. Pick one domain (e.g., customer data) or one application (e.g., Salesforce) and start collecting feedback on its data. After hearing from consumers for 1 - 2 months, you should have more confidence in what problem to solve. Once you do that, you can craft a small POC, isolated from day-to-day operations, to see how well a machine learning-based approach solves the problem.

Getting a small, quick win is the key to being able to launch a broader machine learning initiative. Google, Amazon, and Netflix were experimenting with machine learning long before it became a pervasive part of the consumer experience. Transforming your data stewardship program into a differentiating force starts with gaining hands-on familiarity with how machine learning can make your data consumers more successful. The inherent scalability of the technology means that it won’t be long after realizing those quick wins that your data stewardship program becomes a competitive advantage.



About Tamr

Tamr is the enterprise-scale data unification company trusted by industry leaders like GE, Toyota, Thomson Reuters, and GSK. The company's patented software platform uses machine learning supplemented with customers' knowledge to unify and prepare data across myriad silos to deliver previously unavailable business-changing insights. With a co-founding team led by Andy Palmer (founding CEO of Vertica) and Mike Stonebraker (Turing Award winner) and backed by founding investors NEA and GV, Tamr is transforming how companies get value from their data.

To find out more or register for a demo visit tamr.com