Technical White Paper

Tamr lets organizations connect and enrich all of their data sources, both internal and from partners and third parties. It overcomes the traditional problems of manual data curation and allows for continuous, cost-effective and timely connectivity of hundreds and thousands of sources.

About Data Curation

Data curation is an end-to-end process consisting of:

- **Identifying** data sets of interest (whether from inside the enterprise or external),
- **Exploring** the data (to form an initial understanding),
- **Cleaning** the incoming data (for example, 99999 is not a valid zip code),
- **Transforming** the data (for example, to remove phone number formatting),
- **Integrating** it with other data of interest (into a composite whole), and
- **Deduplicating** the resulting composite.

There are four characteristics required of a data curation system that aims to connect all of the myriad data that exists throughout the enterprise. These are:

1. **Automation.** The effort involved in connecting hundreds or thousands of disparate data sets precludes any solution built around human effort. A system that works at this scale must automate most decision making, using human input only when absolutely necessary.

2. **Direct engagement of data experts.** When human input inevitably does become necessary, the people who are experts on the data must be directly engaged in making decisions, using a non-programmer interface that matches their skills. This is the only practical way to get the expert input required to make timely progress.

3. **Accommodate independently constructed data.** Though some enterprise data is carefully structured and conformed to standardized dimensions, most enterprise data is constructed partially or wholly independently of enterprise standards. The only practical

way to connect this large volume of independent data is to accommodate the data as-is, rather than first clean the data.

4. **Continuous, incremental operation.** New data is constantly being produced by and brought into the enterprise. This data contains an ever-changing variety of semantics and structure, so there is no point at which the task of describing this data is finished. To achieve the goal of comprehensive connection, data must be connected continuously and incrementally, as it arrives.

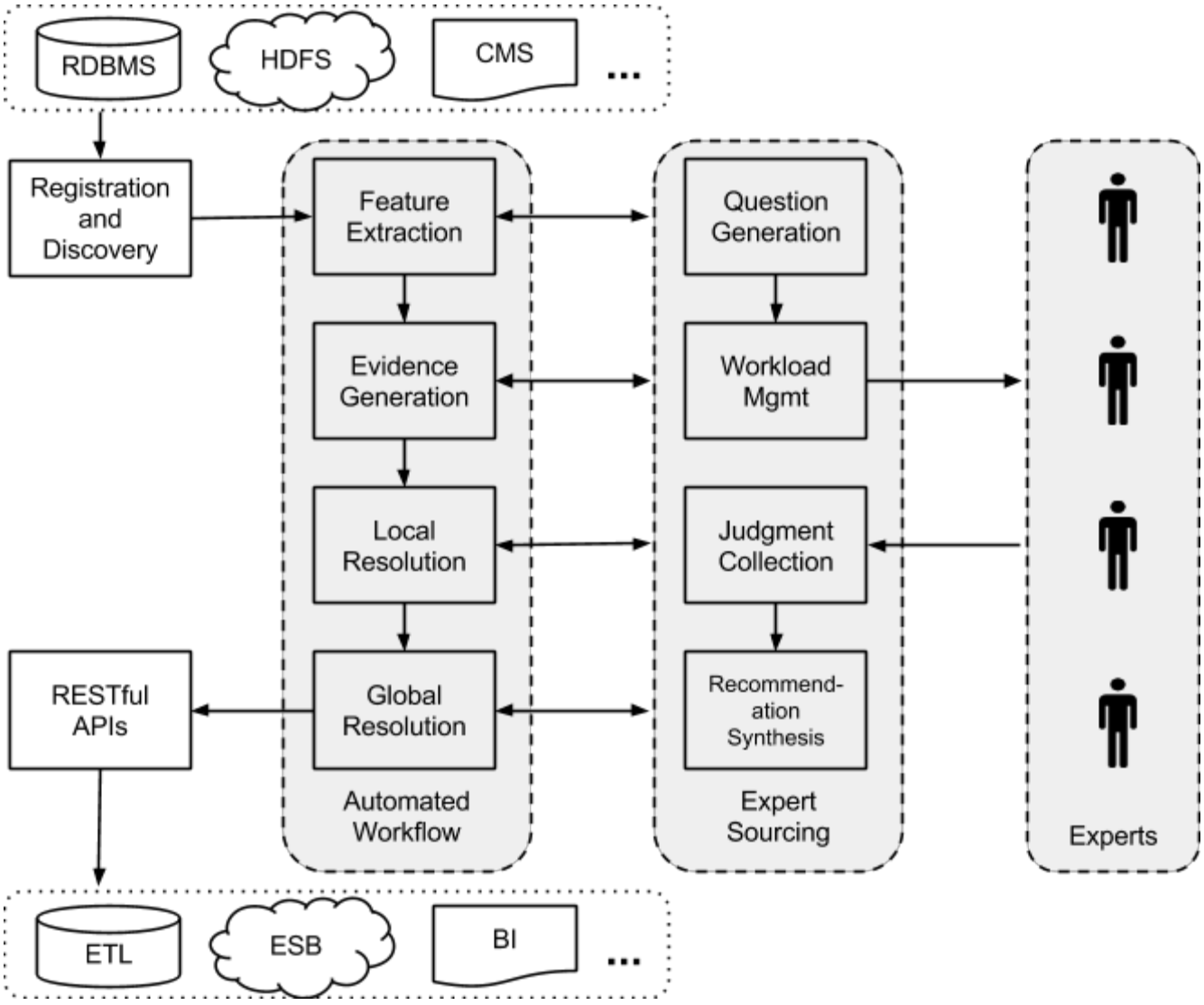

Figure 1: Tamr Architecture

Tamr's architecture (Figure 1) delivers each of these characteristics. The core of the Tamr solution is the pairing of a flexible, automated workflow, which automatically performs most of the curation effort, with an expert involvement module, which engages the line-of-business

people best able to provide necessary input. In the following sections, we describe how the Tamr system accomplishes this crucial blend of being machine driven, yet human guided.

Machine Driven

The driving force behind the Tamr system is an automated data curation workflow. Within this context, decisions are made by a self-contained, modular machine learning system that is specifically tailored to data integration and curation. This system starts with data, then, through a series of refinements, ultimately produces global resolution -- either at the attribute level or at the record level, depending on the context in which it is run. The stages of refinement are outlined in Figure 2 and explained in more detail in the following sections.

Figure 2: The Tamr Machine Learning Information Stack

Data

A system can only provide assistance with data it can actually accommodate, so it is critical that the Tamr system be built on a flexible repository for sampled data and metadata. This repository is able to accommodate arbitrary content, in any shape. This data store needs to accommodate a wide variety of independently constructed data, as well as schema-less data, dirty or missing data, data with hierarchical or irregular structure, poorly formatted data, and data with a high degree of redundancy. This flexible data store allows the very first stages of curation to occur within the system, rather than pushing them out to other data preparation systems. From this potentially very messy data store, the process of machine-driven curation can begin.

Features

In the domain of Information Retrieval, “Feature Extraction” refers to the process of identifying possibly obscured features in data, and separating them from other data, thereby making them available to downstream processing. Tamr implements an automated feature extraction system to enhance both the precision and recall of our machine learning algorithms. Tamr’s feature extraction system is modular and extensible, allowing different features to be used in different contexts. Our schema mapping algorithms take advantage of metadata features such as attribute name, data type, and constraints; as well as data features such as statistical profiles of values or value length, or TF*IDF-weighted tokenized terms. Our record deduplication algorithms take advantage of such data features as unweighted or weighted attribute values, TF*IDF-weighted tokenized terms, or geographic location. The modular structure of the system makes it straightforward to extend with additional features.

For those cases where the automatic feature extraction system does not have an existing method of identifying and extracting a feature, Tamr also supports manual transformation, allowing users to create new attributes that expose information that would otherwise be difficult to access. Transformations include things such as name or address standardization, extracting keywords from descriptive text, removing flag values from measurements, and parsing separate values out of composite fields. Tamr’s transformation system is built to work with the flexible data repository, and retains detailed provenance information to ensure that values can be traced back to their source.

Evidence

Comparing two data objects starts with comparing their features; this generates evidence supporting a connection between the two. Different features require different comparisons, so this subsystem is also modular and extensible. For example, a Z-test is a good way to compare statistical profiles of values, whereas Jaccard or cosine similarity are good ways to compare tokenized values, and geographic distance is good for comparing locations. Tamr's collection of comparators evolves along with the collection of features.

Traditional analysis methods use blocking to partition data into discrete regions in order to reduce the number of comparisons that must be performed. We have found that traditional blocking methods don't work well, both because their expressiveness is insufficient to find good blocking criteria on large, heterogeneous data, and because their decoupling from machine learning training prevents them from adapting to an ever-changing data landscape. Instead, we have developed a closely-coupled, two-stage machine learning system, where the first stage provides an automatic, adaptive alternative to blocking that allows comparisons to be performed on-demand.

Local Resolution

Within a given business domain, the Tamr system learns how to accumulate all the evidence supporting connections between data objects into local resolution decisions - decisions that consider just the objects being compared. The typical way to accomplish this task is to build a classifier using training data specific to the domain. Tamr uses a single set of training data to build a two-stage classifier. The first stage is designed to generate candidate connections, delivering the best possible recall while aggressively eliminating comparisons that are guaranteed to be unnecessary. The second stage is a more traditional classifier aimed at achieving the desired precision and recall; this second stage only needs to consider the candidates retrieved by the first stage. Each stage uses techniques such as kernel density estimation to handle different data distributions, discretization to identify uniform data regions, and prefix filtering to prune the search space.

There are many techniques for training classifiers, spanning unsupervised, supervised and active learning. We use a combination of unsupervised and active learning techniques to rapidly converge on a very high-quality classification model without requiring a technician to oversee the training process. The training process takes advantage of such information as externally-defined rules, bulk training data, and external knowledge of attribute weights, but does not require them. To ensure that a robust classifier is developed, the training process must ensure that the training data represents a complete and unbiased sample from the overall population; therefore, even if bulk training data is provided, the active learning process will generate training questions to be answered by humans. These questions are generated using stratified sampling to ensure good coverage of data variety while minimizing human effort, and our expert sourcing subsystem is used to engage data experts directly in answering these questions. The Tamr system continually

monitors the quality of the classifier, and will perform additional training to re-tune it when necessary.

Global Resolution

Local resolution decisions may contain conflicts and uncertainty, so further refinement is necessary to turn it into global resolution of objects and entities. Tamr is able to automatically resolve many types of conflict and uncertainty using clustering analysis. We have found that a greedy clustering algorithm based on weighted network correlation is both extremely efficient and extremely effective in forming very high quality clusters. We have developed an incremental global entity resolution methodology, such that changes in local resolution can be accommodated with minimal re-calculation.

No automated machine learning system is 100% accurate, so there must be a way to incorporate expert review into global resolution decisions. The Tamr system engages humans both to address the areas where the automated system can not reach a high-confidence decision, and to spot-check high-confidence results to ensure that the final result meets the desired quality objectives. Our expert sourcing subsystem is applied in both cases to engage multiple data experts directly in each of these tasks. This coupling of automated processing with expert input provides a practical way to to achieve high-quality, global resolution at scale.

Human Guided

The machine-driven nature of the Tamr system allows it to make constant progress, but human guidance is required to ensure that it is making progress in the right direction. The core of Tamr's method for engaging humans is an Expert Sourcing system that enables the automated system to determine when and how to engage human experts. This system starts by generating questions from a specific context, then reaches out to multiple human experts to make judgments about those questions, aggregating those judgments into recommendations, which can be used by a human or the automated system to make decisions. These stages are outlined in Figure 3, and explained in more detail in the following sections.

Figure 3: The Tamr Expert Sourcing Information Stack

Questions

Engaging human experts starts with asking the right questions. For each of the different contexts in which Tamr may want to engage human experts, it employs a question generator to determine the best question to ask. The goodness of a question depends on two things: how well an expert will be able to answer it, and how much benefit the system will get out of the answer.

Since Tamr seeks to engage data experts, the questions must be about data, not about data processing. We have found that an effective structure for questions is to present one or several possible decisions - e.g. whether two objects should be connected, or whether a transformed view of the data is correct - and ask which, if any, is correct. The questions presented to experts

must show that the system respects their time and effort; to achieve this, the generators use the results of the automated system to ensure that the decisions around which the questions are composed are likely but non-obvious, or make it clear that the expert is being asked to verify an automated decision. To aid the experts in addressing questions, they are composed with rich contextual information to minimize the need to consult other systems.

To minimize the number of questions asked, the question generators search for a minimal set of questions that will maximize the impact on the automated system, for example, using stratified sampling to ensure good coverage of data variety. Finally, since the questions are being asked of humans, they must be presented in a form that is carefully constructed to counteract known sources of bias, such as the serial position effect or confirmation bias.

Judgments

Once a question or set of questions has been generated, it needs to be sent to one or more experts so that they can render judgment on it. There are many varieties of expert that can effectively address Tamr's questions, such as source owners, data engineers, data stewards, data curators, data architects and data scientists. These are all people who know the data and the business context in which the it is used. Given that there may be many experts able to answer a given question, the system performs intelligent load-leveling to balance the overall workload fairly across the entire pool of experts. This load leveling takes into account each expert's current and recent workload, recent responsiveness, and efficiency when answering a series of closely related questions.

Human experts are not perfect, and the Tamr system tracks this using levels of expertise - a model of the accuracy of an expert's judgments on questions sent out by the system. Furthermore, a given expert can not judge all questions with the same accuracy; for this, the Tamr system uses knowledge domains. Within a single knowledge domain, an expert has a single level of expertise, but can have different levels of expertise in other domains. This allows expert responses to questions to contain uncertainty; for a given question, the Tamr system weighs a judgment by the expert's level of expertise in that question's knowledge domain.

Recognizing that an expert's level of expertise in a single domain will vary over time, the Tamr system uses adaptive expertise - expertise that changes over time in response to observed performance. To continually assess an expert's level of expertise, the Tamr system will intersperse assessment questions with known answers into an expert's workload. We can then use windowing and Bayesian inference to model that expert's changing level of expertise over time.

Recommendations

If judgment rendered by a single expert does not provide sufficient confidence for the automated system to take action on it, the system can present the question to multiple experts, and use their potentially conflicting judgments to build a recommendation with a higher aggregate

confidence. This recommendation is for a decision - e.g. whether data objects should be linked, or whether a transformed view of data is correct.

To determine how many experts to send the question to, the system models the expected confidence of their combined responses. The system can select groups optimized for different metrics, such as the cost to achieve the desired confidence, or the number of experts consulted. This, in turn, is used as input to the load-leveling mechanism. The experts' responses are integrated into an overall recommendation using Bayesian inference, which also produces a confidence score.

Decisions

A recommendation coming from the expert sourcing system can be acted upon either by a human data steward or by the automated curation process. Acting on a recommendation is making a decision - examples of decisions are that data should or should not be linked, or that a transformation should or should not be made, or that two objects do or do not have a particular relationship. All human decisions feed forward into the automated process, to ensure that it continues in the right direction. Expert feedback gathered as part of a training process has a natural place in the automated process, but that gathered outside the context of training needs to be carefully weighed before feeding into the automated process, as it is part of a biased sample - it is biased towards those areas where existing models are uncertain. Data steward decisions can also be used as "gold standard" responses to feed into the system that manages adaptive expertise.

In all cases, the Tamr system retains a complete audit trail, allowing the provenance of every decision to be examined in detail. Decisions can be reviewed, commented upon, revised, and retracted, and these follow-on decisions can feed forward in the same way as the original.

Conclusion

The machine driven but human guided nature of the Tamr system delivers a high degree of automation, while directly engaging data experts. It readily accommodates independently constructed data into a continuous, incremental curation process. This closely-coupled design enables practical, end-to-end curation at the scale of of hundreds to thousands of disparate data sets. It is the only system designed from the bottom up to scale across the entire enterprise.