O'REILLY®

Agile Data Mastering

Andy Oram
Foreword by Tom Davenport

Agile Data Mastering

Andy Oram
Foreword by Tom Davenport

Beijing · Boston · Farnham · Sebastopol · Tokyo
O'REILLY®

Agile Data Mastering
by Andy Oram

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com/safari*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Marie Beaugureau
Production Editor: Melanie Yarbrough
Copyeditor: Amanda Kersey

Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

November 2017: First Edition

Revision History for the First Edition
2017-11-30: First Release

While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-492-02845-1

[LSI]

Table of Contents

Foreword by Tom Davenport

Thomas H. Davenport
Distinguished Professor, Babson College
Research Fellow, MIT Initiative on the Digital Economy
Senior Advisor, Deloitte Analytics and Cognitive Practices
Member of Tamr's Board of Advisors

My focus for the last several decades has been on how organizations get value from their data through analytics and artificial intelligence. But the dirty little secret of analytics and AI is that the people who do this work—many of them highly skilled in quantitative and technical fields—spend most of their time wrestling with dirty, poorly integrated data. They end up trying to fix the data by a variety of labor-intensive means, from writing special programs to using "global replace" functions in text editors. They don't like doing this type of work, and it greatly diminishes their productivity as quantitative analysts or data scientists. Who knows how much they could accomplish if they could actually spend their time analyzing data?

This is particularly true within large companies and organizations, where data environments are especially problematic. They may have the resources to expend on data engineering, but their problems are often severe. Many have accumulated multiple systems and databases through business unit autonomy, mergers and acquisitions, and poor data management. For example, I recently worked with a large manufacturer that had over 200 instances of an ERP system. That means over 200 sources of key data on critical business entities like customers, products, and suppliers. Even where there are greater levels of data integration within companies, it is hardly unusual to find many versions of these data elements. I have heard the term “multiple versions of the truth” mentioned in almost every company of any size I have ever worked with.

Addressing this problem has been prohibitive in terms of time and expense thus far. As pointed out later in this report, companies have primarily attempted to solve it through the collection of techniques known as “master data management,” or MDM. One objective of MDM is to unite disparate data sources to achieve a single view of a critical business entity. But the ability to accomplish this objective is often limited.

As this report will describe, rule engines are one approach to uniting data sources. Most vendors of MDM technology offer them as a key component to their technology. But just as in other areas of business, rule engines don’t scale well. This 1980s technology has some virtues—rules are easy to construct and are often interpretable by amateurs. However, dealing with large amounts of data and a variety of disparate systems—attributes of multiple-source data in large organizations—are not among those virtues. In this as in most aspects of enterprise artificial intelligence, rule engines have been superseded by other technologies like machine learning.

Machine learning is, of course, a set of statistical approaches to using data to teach models how to predict or categorize. It has proven remarkably powerful in accomplishing a wide variety of analytical objectives—from predicting the likelihood that a customer will buy a specific product, to identifying potentially fraudulent credit transactions in real time, and even to identifying photos on the internet. Much of the enthusiasm in the current rebirth of artificial intelligence is being fueled by machine learning. It’s great that we can now apply this powerful tool to one of our most persistent problems—inconsistent, overlapping data across an organization.

Unifying diverse data may not be one of the most exciting applications of machine learning, but it is one of the most beneficial and financially valuable. The technology allows systems like Tamr’s to identify “probabilistic matches” of multiple data records that are likely to be the same entity, even if they have slightly different attributes. This recent development makes a very labor intensive and expensive data mastering initiative into one that is much faster and more feasible. Projects that would have taken years without machine learning can be done in a few months.

Of course, as with other applications of AI, there is still some occasional need for human intervention in the process. If the probability of a match is below a certain level, the system can refer the doubtful data records to a human expert using workflow technology. But it's far better for those experts to deal with a small subset of weak matches than an entire dataset of them.

The benefits of this activity can be enormous. How valuable is it, for example, to avoid bothering a customer with multiple marketing messages, or to be able to focus marketing and sales activities on an organization's best customers with speed and clarity? How important is it to know that many different functions and business units within your company are buying from the same supplier? And would it be useful to know that you have more than you need in inventory of an expensive component of your products? All of these business benefits are possible with agile data mastering fueled by machine learning. And a side benefit is that the employees of your organization won't have to spend countless hours trying to figure out whose data is correct or creating a limitless number of rules.

Even with this powerful technology, it still requires resolve, effort, and resources to unify and master your data. And after you've done it successfully, you still need effective governance to limit ongoing proliferation of key data. But now it is a reasonable proposition to think about a set of "golden records" that can provide long-term benefits for your organization. One version of the truth is in sight, and that is an enormously valuable business resource.

Executive Summary

In the Big Data era, the vision of virtually every large enterprise is to maximize the use of their information assets to build competitive advantage. Much of the focus to date has been around the use of storage technologies and analytical tools to accomplish this goal. However, a frequently overlooked piece of the puzzle is leveraging new methods for managing the data that connects the storage systems with downstream uses such as analytics. Without complete and clean data, the analytics become incomplete, inaccurate, and even misleading. This report focuses on the importance of creating golden, master records of critical organizational entities (e.g., customers, suppliers, and products) and why leveraging machine learning to make the process more agile is critical to success.

Master records are the fuel for organizational analytics; they represent a complete view of unique entities across the distributed, messy data environments of large organizations. Analytic tools rely on such records to ensure the data being pulled is relevant to the entity being analyzed and that all of the data is captured, ultimately ensuring completeness and trust in the result. The traditional methods for building these master records, like master data management (MDM) software platforms, have been effective at a small scale, but are struggling to keep pace in the current environment. Typical MDM tools rely heavily on manual programming of rules that match and merge records to create a single golden record. When a data environment grows too large and too diverse, however, this becomes an unscalable practice. Unsustainable amounts of time and expense are required to keep pace with the amount of data being captured and, most often, the initiative will not deliver the return on investment that is needed.

Now is the time when organizations need to evaluate a new approach to mastering their data, an approach that cost-effectively delivers these golden records at speed and scale, across domains, and with the ability to classify them so organizations can fully realize the benefits of their analytic endeavors. This approach is called agile data mastering, and it revolves around the use of human-guided machine learning to match, merge, and classify core organizational entities like customers, suppliers, and products. Machine learning algorithms employ probabilistic models that attempt to master raw data records while an internal expert validates the results, which tunes the algorithms, delivering the aforementioned benefits while also ensuring an underlying accuracy and trust in the results. This report dives deeper into the elements of agile data mastering and the methods that power it so companies across any industry and with any type of data environment can manage their data to support their digital transformation goals and maintain their relevance.

Agile Data Mastering

Introduction

Organizations across all industries are attempting to capitalize on the promise of Big Data by using their information assets as a source of competitive advantage. In doing so, they are investing heavily in areas such as analytic tools and new storage capabilities. However, they often neglect the data management layer of the equation: it's not simply about finding an optimal way to store or analyze the data; it's also vital to prepare and manage the data for consumption. After all, if the data is inaccurate or incomplete, it can undermine confidence of the people who rely on that data and lead to poor decision-making. Whether the organizations are processing customer records, supplier records, or information about other entities, they are dealing with datasets containing errors and duplicates. Typically, organizations expend a lot of time on manual data cleaning and vetting to create *master records*—a single, trusted view of an organizational entity such as a customer or supplier—and this is often the area where most help is needed.

Machine learning can be immensely powerful in the creation of the master data record. This report will describe the importance of the master record to an organization, discuss the different methods for creating a master data record, and articulate the significant benefits of applying machine learning to the data mastering process, ultimately creating more complete and accurate master records in a fraction of the time of traditional means.

Importance of the Master Record

The difficulties of collecting and managing Big Data are hinted at in one of Big Data's famous Vs: *variety*. This V covers several types of data diversity, all potentially problematic for creating a valid master record:

- In the simplest form, variety can refer to records derived from different databases. You may obtain these databases when acquiring businesses and when licensing data sets on customers or other entities from third-party vendors. Many organizations have to deal with different databases internally as well. All of these can have inconsistent names, units, and values for fields, as well as gaps and outdated information.
- Data in your official databases can be incorrect, perhaps because it is entered manually, because a programming error stored it incorrectly, or because it is derived from imperfect devices in the field. Arbitrary abbreviations, misspellings, and omissions are all too common.
- Some data is naturally messy. A lot of organizations, for example, are doing sentiment analysis on product reviews, social media postings, and news reports. Natural language processing produces uncertain results and requires judgment to resolve inconsistencies. When products are mentioned, for instance, reviewers may not indicate the exact product they're talking about and in the same way.

Thus, you may end up with half a dozen records for a single entity. When collecting data on customers, for instance, you may find the records like those shown in Table 1-1 scattered among a large collection of different databases.

Table 1-1. Multiple records about the same person

Family name	Given name	Gender	Street address	City	State
Pei	Jin-Li	Female	380 Michigan Avenue	Chicago	IL
Pei	Jin-Li	Female	380 Michigan Avenue	Chicago	Michigan
Pei	Julie	Female	380 Michigan Avenue	Chicago	Illinois
Pei	Jin-Li		380 Michigan Ave	Chicago	IL
Pei	Julie	F	31 Park Place	Chicago	Illinois

It is fairly easy for a human observer to guess that all these records refer to the same person, but collectively they present confusing differences. Julie may be a common English name that Jin-Li has chosen to fit in better in the United States. The state is spelled out in some records while being specified as a two-letter abbreviation in others, and is actually incorrect in one entry, perhaps because the street name confused a data entry clerk. One entry has a different address, perhaps because Jin-Li moved. And there are other minor differences that might make it hard for a computer to match and harmonize these records.

This small example illustrates some of the reasons for diverse and messy data. The biggest differences, and the ones most amenable to fixing through machine learning, are caused by merging data from different people and organizations.

Data sets with entries like Table 1-1 present your organization with several tasks. First, out of millions of records, you have to recognize which ones describe a particular entity. Given fields of different names, perhaps measured in different units or with different naming conventions, you have to create a consistent master record. And in the presence of conflicting values, you have to determine which ones are correct.

Another V of Big Data, *volume*, exacerbates the problems of variety. If you took in a few dozen records with messy data each day, you might be able to assign staff to resolve the differences and create reliable master records. But if 50,000 records come in each day, manual fixes are hopeless.

Inconsistent or poorly classified data is dragging down many organizations and preventing their expansion. Either it takes them a long time to create master records, or they do an incomplete job, thus missing opportunities. These organizations recognize that they can pull ahead of competitors by quickly producing a complete master record: they can present enticing deals to customers, cut down on inefficiencies in production, or identify new directions for their business.

But without a robust master record, they fail to reap the benefits of analytics. For instance, if it takes you six months to produce a master data record on Pei Jin-Li, she may no longer be interested in the product that your analytics suggest you sell her. Duplicates left in the master data could lead to the same person receiving multiple communications—or even more embarrassing interactions. For instance, you may know that Pei bought a car from you, so you send her an offer to finance her car with you as well, not realizing that she already financed her car with you—or perhaps tried to do so and was declined.

Reasons for Data Mastering

Corporations often find themselves seeking to use master data records for two purposes. *Operational mastering* supports everyday activities and current tasks, whereas *analytical mastering* looks toward new areas for the corporation to profitably grow.

Operational mastering can answer questions such as: when a bank customer logs in, does he see all his accounts? When a customer searches for "loan," does she see all the products we offer under that category (and no others)? To support operations, data mastering ensures that the master data records are accurate and that the classification is up to date.

Analytical mastering can provide the data to answer questions such as: What product shall we offer to a customer? Where is the cheapest source of a part we need to order?

This section lays out a few applications of master data in an organization:

- Generating organizational analytics
- Data mining
- Creating classifications

Generating Organizational Analytics

Data resides in many places within large companies. One division of a company could sell clothing while another sells sporting goods—and each capture customer data differently. Combining the datasets of these two divisions could help the organization as a whole understand their customers better. This can improve both large-scale planning (what will a lot of people want to buy next season?) and small-scale marketing (what does Andy want for his exercise routine?).

As a simple example, if customers are spending more on sporting goods than expected, they may be open to spending more on clothing as well, especially what they wear during their sporting events. A more sophisticated example comes from sourcing products: analytics may reveal that two functionally interchangeable products cost different amounts, and might lead a company to save money by substituting a cheaper material for a more expensive one.

One Fortune 100 manufacturer uses analytics on its master data of suppliers and parts to control supplier costs, such as by finding the lowest-cost supplier for each part. Using machine learning to master and classify tens of millions of records over a relatively short period of time saved the company hundreds of millions of dollars. On another set of multiple CRM systems, they reduced more than one billion customer records to derive a few hundred thousand master records.

Data Mining

Datasets can also be used to create a *mosaic effect*, which refers to laying out a wide variety of data elements that say relatively little in isolation but reveal a pattern when combined. For instance, researchers have shown that combining de-identified data from different sets can turn up relationships that re-identify a person.

A legitimate use of the mosaic effect is in law enforcement. Homeland Security may notice that a resident of Los Angeles named Sam Smith was recently fired from his job for inappropriate behavior, and that a resident of Long Beach named "Spiffy" Smith recently bought a semi-automatic weapon. Are Sam Smith and "Spiffy" Smith the same person, and has his behavior changed recently in suspicious ways? Analyzing information from data sets that have been combined and mastered may yield the answer.

Creating Classifications

A classification, or taxonomy, is a way of understanding the world by grouping and categorizing. Should I consider a watch that tracks my footsteps to be an item of clothing, a piece of sporting equipment, or jewelry? It can be multiple things, but I should be able to find it by searching a website in ways that are intuitive to me.

Most classifications keep things simple and arrange products in hierarchies. But as the watch example shows, different taxonomies can classify the same entity in different ways. Master records with different classifications must be harmonized in order to produce a consistent classification.

The classification problem becomes even harder for an organization in the event of a merger or acquisition. Naturally, classifications that come from different companies during a merger or acquisition are likely going to be very different. The organization must start by deciding which classification to adopt. All the records of the other company must be fit into the chosen classification, although the chosen one can be expanded. But the organization must determine how the thousands or millions of new records match to the chosen classification.

Traditional Approaches to Master Data Management

Master data management (MDM) is a large and thriving industry trying to solve data mastering challenges in order to quickly produce robust master records. Currently, it is largely rule-based and explicit. In other words, a programmer writes large, complex programs that search for values in common, generally using pattern matching such as regular expressions: the expression "Ave*" matches both Ave and Avenue. (Matching Boulevard, Boul, and Blvd is harder but still possible.) Large, predefined tables match F to Female in the gender field and Ave to Avenue in the address field. Programs also contain specific coding to decide whether records are close enough to be considered a match.

But machine learning can go much deeper into data handling, creating probabilistic models to consolidate, merge, and classify data. There is no need to hand-craft a table specifying that Boulevard, Boul, and Blvd are all the same moniker; machine learning can figure it out statistically. Rigid, rules-based data mastering is slow and misses potential matches. An agile, machine learning–based approach is faster, more accurate, comprehensive, and flexible.

What Is Agile Data Mastering?

Thus, the use of machine learning to create master data records is called *agile data mastering*. It includes the following functionality:

Matching

Also known as clustering (see "Clustering" on page 7), this task assembles all the records pertaining to single entity, such as Pei Jin-Li of Chicago. A number of statistical techniques can be employed to match relevant records.

Merging

Once the matching process has turned up the records related to a single entity, the correct values must be extracted to create a single trustworthy record. Data values must also be normalized.

Classifying

Most organizations create hierarchies or taxonomies to aid in searches. For instance, the top tier of a taxonomy may consist of terms such as Hardware and Machinery. Hardware, in turn, may consist of a second tier such as Nuts and Bolts. Bolts can also have subtiers to refer to quarter-inch copper round-head bolts, etc. The same techniques used for matching and merging can also place entities into these classifications.

We'll turn now to the advantages of agile data mastering and the specific machine learning–based techniques it uses.

Clustering

This is one of the areas being researched most intensely in artificial intelligence. The goal of clustering is to find entities in data that are similar. The different characteristics used to compare records—such as the size and color of a garment—are called *features*. The more features that are the same (or close to being the same) among two entities, the more similar the entities are. If the relationships are graphed, the similar entities will be bunched together while other features will be further away. The algorithms can indicate which features are most likely to group entities, and highlight these features by assigning them high *weights*.

A large number of techniques are available for clustering, and more are being discovered on a regular basis. Some techniques ask the user to specify features in advance, whereas others develop the features through comparisons of different data items.

In MDM, clustering finds records that are likely to refer to the same entity. In other applications, clustering may find customers who share interests—so a user may determine that a movie that is popular among some of the customers is likely to be enjoyed by others. Thus, clustering underlies recommendation systems.

Another use for clustering is to find correlations. Machine learning here possess a power not available in classical statistics. Historically, statisticians could compare two populations (such as through a T-test) or even multiple populations (through an ANOVA test) to find out whether there is a strong correlation between them for a particular feature—but the statistician must specify that feature. Machine learning can check thousands of features and find the ones that have the strongest correlations. Clusters reveal those correlations. For instance, biochemists use machine learning to find genes associated with particular diseases or the success of particular treatments.

The Advantages of Agile Data Mastering

In this section, we look at processes based on machine learning that can create master data records and related outputs, such as classifications. This section is based on the operation of Tamr, a relatively young company that applies human-guided machine learning to data mastering.

Traditional techniques of determining if records match has always involved the application of rules. As we have seen, current MDM usually requires people to define the rules and embed them in programs. One rule, for instance, might accept a match if the text in two different records is sufficiently similar, while another might reject a match if two birthdates are sufficiently far apart in time. Rules can also involve sophisticated processing, such as setting the weight of different fields. Statisticians may determine that matching a name is much more important than matching certain other fields that are more likely to change, be entered incorrectly, or be less important. So the name fields will have a higher weight when trying to determine whether two records refer to the same entity. Similar techniques for matching records are used for merging them into master data records and for classifying them into taxonomies.

With agile data mastering, finding the same words in different records (one classified and another not classified) allows Tamr to determine whether records are referring to the same entity and/or where they should be placed in a taxonomy. And if matches are not exact, machine learning can determine that the fields are close enough to classify new records properly.

Tamr does not start out with rules; it finds all the relationships, weights, and probabilities through machine learning. There are many advantages to generating rules dynamically:

- The system adapts to the data presented by each organization, generating record matches that are uniquely appropriate for the domain of interest (customers, suppliers, products, etc.). This is important for organizations in different industries or areas of research, with arcane terminology and relationships.
- The resulting rules are more complex than a human being could design or understand, allowing for more accuracy in matching and classification.
- The system is fast, generating results in a few hours.
- Cultural and language differences are handled automatically.
- The system can be run regularly to pick up changes in data, keeping the rules in sync with the environment.

Machine learning is essentially the balancing of probabilities. Thus, in the case of Pei Jin-Li, the two records in Table 1-2 are very similar and therefore are highly likely to refer to the same person.

Table 1-2. Two records that are likely to be the same person

Family name	Given name	Gender	Street address	City	State
Pei	Jin-Li	Female	380 Michigan Avenue	Chicago	IL
Pei	Jin-Li		380 Michigan Ave	Chicago	IL

On the other hand, the records in Table 1-3 are less similar, and less likely to be the same person.

Table 1-3. Two records that are less likely to be the same person

Family name	Given name	Gender	Street address	City	State
Pei	Jin-Li	Female	380 Michigan Avenue	Chicago	IL
Pei	Julie	F	31 Park Place	Chicago	Illinois

Because it involves probabilities, machine learning also involves risk. Algorithms might calculate, based on all the records it finds about Pei in Chicago, that the previous two records have a 90%

chance of being the same person. If the algorithm accepts 1,000 matches with a 90% probability of being right, your master data will contain 100 people who are wrongly associated with another person (100 false positives, or type I errors). But this may be better than rejecting those one thousand matches, because then you'll have 900 duplicates (false negatives, or type II errors). There is always some balance between false positives and false negatives; decreasing one risk leads to an increase in the other. In the field of information retrieval, this balance can also be viewed as a balance between precision (did we get each match right?) and recall (did we get all available matches?).

Many machine learning algorithms tell you the risk of false positives and negatives, and allow you to tune the algorithm. A false positive for customers, as we've seen, may lead you to market a product to Jin-Li that she's not interested in, which is not a particularly bad outcome. In contrast, false positives may have much more dangerous consequences if you are matching parts from vendors. You might end up placing 10,000 orders for a part that you think performs the same role as a current part, but is cheaper. You may then discover that the cheap part differs from the current part in some crucial way that renders it unfit for your use.

Part of the implementation of an algorithm, therefore, involves handling ambiguous cases where there is a risk of a false positive or negative. We'll see how Tamr handles these cases as we look at its operation.

Starting Elements

Tamr does not need pre-existing tables of common matches, because relationships are uncovered through statistical analytics. Tamr starts with a training set of data prepared by the client, a standard practice in machine learning. The number of records needed for training is proportional to the total amount of records in a data set, and in practice Tamr usually takes a couple hundred records for training data.

Additionally, to help develop a classification, Tamr accepts a preexisting taxonomy from the organization. If the organization doesn't have a taxonomy already—most do—Tamr can start with a common default such as the United Nations Standard Products and Services Code. The reason for starting with a template, instead of trying to develop a template from scratch through machine learning, is that the right template depends on the interests and needs of the user—it requires humans to make value judgments.

As a somewhat ludicrous example, let's look at a clothing store and ask what template would serve it. You may be looking for a blue blouse and end up with a green or purple blouse. But while looking for a blouse, you are not going to buy a winter overcoat just because it is blue. The store has to organize clothing by garment type and by size. Color is far down in the classification. But left to its own devices, machine learning may decide that color is a top-level tier and that everything in the store should be arranged primarily by color. Statistically, the choice would be perfectly sensible—it's up to humans to decide what makes sense for them.

Human intervention is also required during the processes of matching, merging, and classification described in "Importance of the Master Record" on page 2. Let's suppose that Tamr can match 80% of its input records with confidence. It will take some of the ambiguous ones and show them to a user, who can then indicate whether or not they are matches—for instance, whether Jin-Li and Julie are the same person. A couple dozen matches by the user can help Tamr handle a large number of the ambiguous cases and greatly increase its successful classification rate. A user can also flag a match as incorrect, again helping Tamr improve its algorithm for the entire data set.

Goals of Training

We all learn from experience, what logicians call inductive reasoning. This kind of learning—deriving general principles from a large set of specific examples—also characterizes the training that is a critical part of machine learning. The process starts with data that has been checked by humans and is guaranteed to be accurate; after processing this data, the machine learning can start to understand record matching principles that help it process new, raw data.

Below are examples of intelligence that machine learning systems can develop by running statistics on large amounts of data:

Outliers

These are entries that differ wildly from the others in some way. For instance, if data indicates that a doctor performed 200 surgeries in one day, machine learning will probably flag him as someone to examine. The reason may be an error in programming or data entry. In other circumstances, it can uncover fraud. Or it may simply occur because one doctor is billing for all the surgeries performed in his department.

Outliers can be recognized without applying explicit rules such as, "A doctor can perform at most three surgeries each day." Instead, the outliers are discovered during clustering. The outlier may have some features (data fields) in common with other records, but at least one of the features is so different that it ends up far away from the other records. If one were to graph the data, one might see several groups of records that are related, and a few outliers that lie off in their own regions of the graph with large gaps between them and the rest of the data.

Typos and other inaccuracies

These are particular types of outliers. Unlike the doctor performing 200 surgeries, a typo is not wildly different from other values. Instead, it is only slightly different: "Avnue" instead of "Avenue." Machine learning can identify "Avnue" as an outlier because it occurs rarely, is very similar to the correct spelling, and appears in a context that makes the similarity clear—for instance, the rest of the address is the same as in the fields that say "Avenue."

Ranges and limits

These identify other kinds of outliers. Machine learning will discover that an age of -1 or 200 for a person is an error.

Patterns and arbitrary rules

For instance, a state called Cal is probably California, and Bob is an alternative name for Robert. As with typos, these variants can be discovered through context. As we saw with Pei Jin-Li, machine learning can compare several similar records to suggest that Julie is an alternative for Jin-Li. The match would be an impossible stretch if we saw the two names in isolation, but in the context of the other fields, the match seems probably correct.

The Training Process

Tamr has a unique training process that involves human input. Initial training usually requires a few hours.

The process starts with known matches, provided by the customer, and derives as much information as it can from the input data on its own. It can compare most records quickly, because most records are very different. When Tamr traverses columns from different data sets and compares fields, it determines how close two strings are through common algorithms such as Levenshtein distance. (That's just one classic formula for checking whether strings are similar; there are many others.) The word "Avenue" is close to the misspelled "Avnue," creating what is called a *strong signal* in machine learning, whereas Avenue is less close to Ave. As we saw in the Pei Jin-Li example, machine learning can determine that Avenue and Ave refer to the same thing by comparing large numbers of records where other strings within the field, or other fields, are strongly matched.

Certain fields will predict matches better than other fields, so Tamr can assign a weight to each field during training to determine how much confidence to invest in that field later when processing raw data. The result of applying traditional clustering techniques is a huge set of very small clusters. For instance, the Pei Jin-Li shown in Table 1-1 contained only five records. There could be tens of thousands of these clusters, one for each customer.

If confidence is high that a row belongs in a cluster, it is associated with the master data record and may supply some of the data values for that record. Each cluster results in a single row in the master data table.

Inevitably, some records will be ambiguous. They are so similar that they might be considered a match, yet different enough to be suspect. Machine learning can make confident decisions on many records by comparing all the fields of the records, assigning more confidence to higher-weighted fields, and by finding patterns that appear frequently in the data. For instance, if Bob and Robert have been shown to refer to the same person in a thousand records, the machine learning algorithms place more confidence that Bob equals Robert when comparing two records that differ in other ways as well. When the decision that they're the same is not clear enough, Tamr can present the records to a human. Having humans mark the pair as correct or incorrect through very simple yes or no questions helps Tamr tune its algorithms.

To improve its classification process, Tamr follows a similar process: using its algorithms to attempt to match records to a taxonomy, and then asking a human user a simple yes/no question for validation if necessary. For instance, Tamr may start with records associated with a taxonomy that contains branches for nuts and bolts. Records from a recently acquired company are then input, some describing half-inch steel screws, some describing quarter-inch copper screws, and so on. Because the original taxonomy contains bolts that are half-inch steel or quarter-inch copper, Tamr can recognize that the screws are the same as bolts—but it will ask for human intervention if it is not sure.

Thus, the user helps to train the algorithm with just a few simple choices. With each iteration of training the algorithm, it can classify more and more records with less and less human help. One media company achieved both precision and recall rates of over 95% with Tamr.

Machine learning requires retraining over time. Tamr will continue to refine both the match/merge algorithm and the classification algorithm for an organization's data by processing new data on a regular basis.

Conclusion

Industry observers like to emphasize the importance of areas like data processing and analytics when discussing Big Data, but the centrality of the master data record is rarely noted. Data about the critical elements of your business is of little value if it remains scattered across countless, disparate data silos. Master data management is crucial to extracting full value from your data. When input records reach into the millions or billions, as they increasingly do for more and more organizations, creating master data records requires more sophisticated tools than the simple application of predefined rules. Machine learning can provide the necessary insight. It can recognize new patterns, pull together related records, and master the data to prepare it for analysis. In this way, matching, merging, and classifying operations can all benefit from machine learning.

The world has only begun to discover the potential applications for artificial intelligence. The success of companies like Tamr shows that machine learning can reach into numerous areas of human endeavor. And in doing so, we can expect new applications for these machine learning-based technologies to continue emerging in places where no one expected them.

About the Author

Andy Oram is an editor at O’Reilly Media. An employee of the company since 1992, Andy specializes in programming topics. His work for O’Reilly includes the first books published commercially in the United States on Linux, and the 2001 title *Peer-to-Peer*.