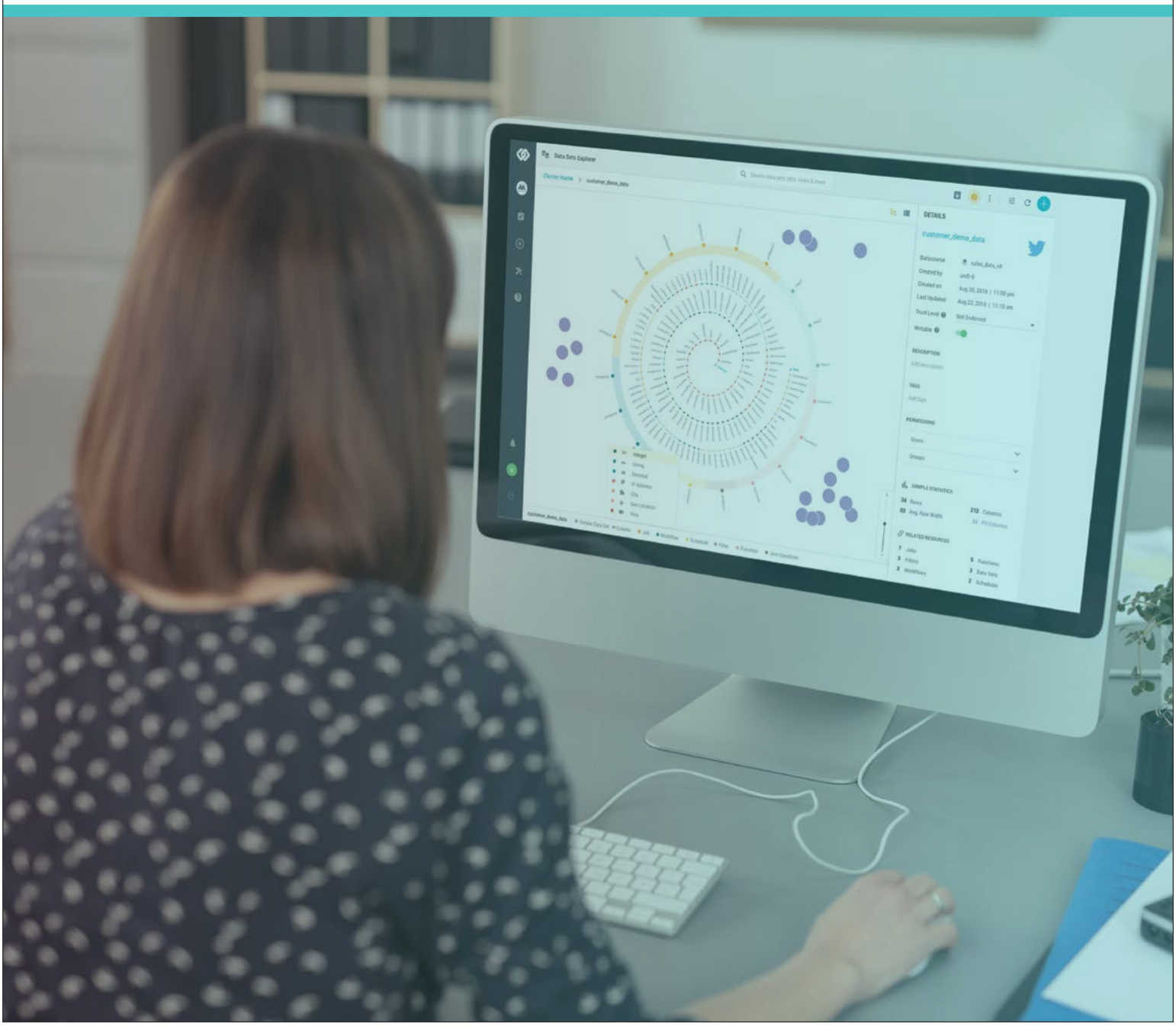
unifi

WHITEPAPER

Practical Examples of the Impact of AI in Data Management

By Neil Raden, CEO & Principal Analyst, Hired Brains Research



Overview

Data catalogs emerged as the must-have technology for dealing with vast collections of data files. But a data catalog alone does not solve the problems facing organizations of providing a simple discovery tool. A static catalog lacking an in-depth understanding of the variety of data formats only addresses a fraction of the problem. The application of AI to provide recommendations of mappings of data sources, and exposed through Natural Language Query, search and exploration in your own words with continuous update, makes this possible.

What impact can Unifi's employment of AI have on your business? The drive of organizations to be data-driven means a much wider audience of analysts and decision-makers need self-service discovery and use tools where AI replaces a user's lack of technical knowledge and does in minutes what would typically take hours to do.

What you will learn in this paper:

- How Unifi's application of AI provides access to data without the need for code
- How Unifi's embedded AI provides recommendations, not merely black-box solutions
- How Unifi understands similarity across datasets
- Explicitly what AI techniques Unifi applies to the different problems it solves
- How AI in Unifi benefits you

More Data Sources. More Data Volumes. More Data on the way - Every Day

The virtually unlimited amount of data, processing capacity, and tools to leverage it are a modern deluge without a doubt. The challenge today is not the volume of data, it's making sense of it, at scale, continuously. Consider this: most analysis and reporting in the past used data from one or two internal systems. Today's "digital" organizations may mobilize hundreds, thousands, or even millions of files to drive value.

Big data freed organizations to capture far more data sources, at higher levels of detail and vastly greater volumes which expose a massive semantic dissonance problem. Harkening to the call to be "data-driven" organizations are striving to ramp up skills in all manner of predictive modeling, machine learning, AI, or even deep learning. And of course, the existing analytics requirements cannot be left behind, so any solution has to satisfy those requirements too. Integrating data from your own ERP and CRM systems may be a chore, but for today's data-driven applications, the fabric of data rapidly became multi-colored. Sources for consideration exploded. For example:

- Data.gov contains over a quarter million datasets ranging from Coast Guard accidents to bird populations, demographics to Department of Commerce information. Social media platforms offer a wide variety of views of their data.
- Healthdata.gov https://www.healthdata.gov/ contains 125 years of US healthcare data, including claim-level Medicare data, epidemiology, and population statistics. These are just a few of thousands of external data sources.
- So-called secondary data such as structured interviews, transcripts from focus groups, email, query logs, published texts, literature reviews, and observation records present a challenging problem to understand. Records written and kept by individuals (such as diaries and journals) and accessed by other people are also secondary sources.

The primary issue is that enterprise data no longer exists solely in a data center or even a single cloud (or more than one, or combinations of both). Edge analytics for IoT, for example, capture, digest, curate and even pull data from other, different application platforms and live connections to partners, previously a snail-like process using obsolete processes like EDI or even batch ETL. Edge computing can be thought of as decentralized from on-premises networks, cellular networks, data center networks, or the cloud. All of these factors pose a risk of data originating in far-flung environments, where the data structures and semantics are not well understood or documented[1]. The risk of smoothly moving data from place to place or the complexity of moving the logic to the data while everything is in motion is too extreme for manual methods.

The problem arises because not one of these data sources is semantically compatible with the others. Only by drawing from multiple sources can useful analytics be derived.

The sudden surge in computing capacity drove demand for more data and more analytics, but also led to a temporary structural shortage of professional data scientists and statisticians. Exacerbating this insufficiency were productivity problems associated with the work, what became known as the 80/20 problem: 80% of the time spent managing data and only 20% doing actually quantitative investigation[2].

Better tools are needed to improve productivity (and job satisfaction) of highly skilled and compensated professionals, both IT and business analysts. In fact, even those performing analytics in organizations in more traditional ways can benefit from an intelligent and integrated product taking them from data discovery and profiling to navigating the data with an active, semantically-rich data catalog and applying it in unique and productive ways.

Spoiler Alert: Unifi Software does this. Unifi OneMind AI technology underlies the functionality from data prep and data catalog recommendations to the discovery of similar datasets, to Natural Language Query support that helps users directly get answers to data questions. Integrated governance and security help maintain compliance for self-service access to data. The Unifi Data Platform delivers advanced, OneMind-powered, self-service data preparation, and workflow automation functions to help users and organizational teams create repeatable reporting and analysis. Both the Unifi Data Catalog and Unifi Data Platform can be deployed on-premises, in private or public cloud, or hybrid environments.

Unifi offers automation across almost every capability within the application. The best use of automation / Machine Learning is when the user does not even know it is in use. Below are just a few examples of where these capabilities are employed.

Knowledge Graph – Powering Unifi Discovery & Recommendations

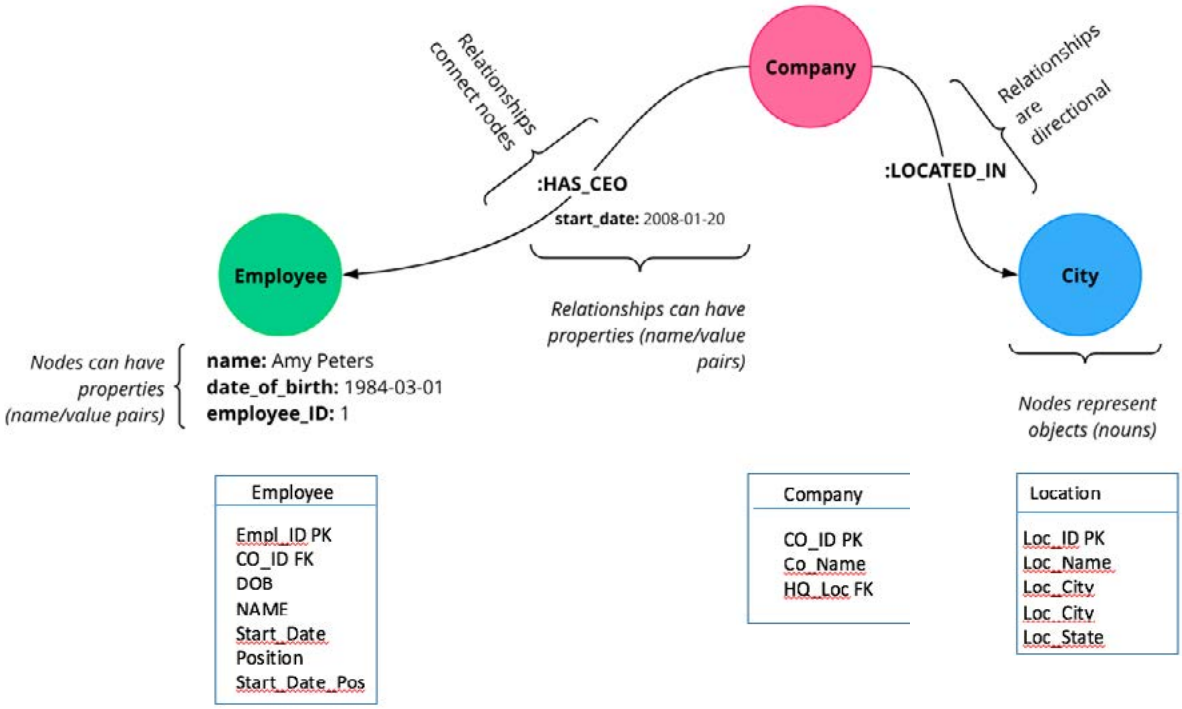
Graph Theory is a branch of mathematics that was first developed over three hundred years ago, but only in the twentieth century did it become an active area of research. Understanding how to apply graph theory does not require knowledge of the breadth of current research in graph theory, only a few simple concepts.

The schemas below contain roughly the same data, but their arrangements are entirely different. The relationships in the "relational" schema are only implied and are only materialized in the predicate of the SQL statement (the WHERE clause). Nevertheless, it is possible to get answers from either schema, but when there are thousands or hundreds of thousands of employees, the multiple joins required to satisfy the query can be quite costly (and remember, this is a trivial model). This is precisely where a graph model outperforms as the relationships are inherent in the model.

---

[1] A trucking company may have more than twenty separate telematics providers in the cab, each with its own protocols for applications that require the trucking company to absorb and react to in near-real-time

[2] "Data scientists spend 80% of their time preparing and cleaning their data. They spend the other 20% of their time complaining about preparing and cleaning their data." Data Science Central, April 22, 2018



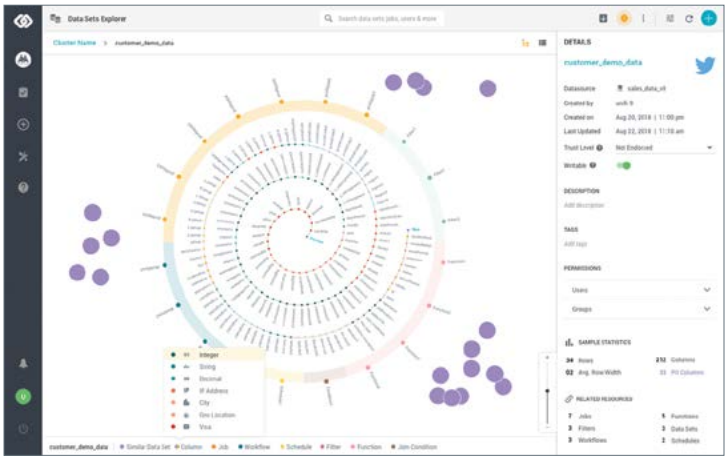
Simple Relational Schema

Modeling this information in relational databases serves the purposes of transactional integrity, archival and general referential uses, but also causes a high number of many-to-many relationships which in turn leads to a succession of very costly JOIN operations when querying the data. Graph structures are ideally suited to use-cases such as the one above. They are not simple aggregate solutions (such as Key-Value Stores, Column-Family Stores, or Document Databases). Because the graph query can traverse relationships, instead of joining large tables, it can efficiently tease out inferences that can only be drawn from SQL by, essentially, asking about them.

Using the Knowledge Graph: An Example

An organization was an early adopter of a data lake and proceeded without proper effort to maintain metadata, lineage, and provenance. As a result, the environment was difficult to secure and administer. The data lake was full of files, mostly from within the organization's data centers, but there were also files from desktop/laptop applications, email, memoranda, among others, and files pulled from external sources and scraped from social media. There were typical structured formats as well as many others, data from applications, logs. Additionally, an active Machine Learning group generated hundreds of new files every day. Because many of the files ingested into the data lake also existed outside of it, spread around the organization, it was impossible to know which ones were current, to assess data quality, or to apply any rational way to avoid duplication.

Unfortunately, finding the right data for data scientists, data engineers, and AI engineers was done by word of mouth, passing notes back and forth and pure manual examination-much of the insight was derived from tribal knowledge. The data lake was never considered strategic because projects never progressed beyond the pilot phase, never addressing the big questions, never productized or operationalized. It was an endless repository with little value, but substantial cost.



The Unifi Knowledge Graph Navigator provides an alternative, visually intuitive method of exploring their data.

Unifi's solution was to build a Knowledge Graph. Unifi can identify 80 different data formats including JSON, AVRO, Parquet, XML, RDP including semi-structured data via its OneParse Data Parsing technology. Unifi "crawls" through the universe of data and uses AI techniques to create an Enterprise Knowledge Graph that encodes relationships between files and depicts all of the relationships in the entire corpus of data. The Knowledge Graph underlies the catalog and powerfully distills complexity to something understandable. Connections between nodes form a directed graph which is the bedrock of the catalog and background processing. Unifi employs parts of a suite of AI tools named OneMind™ to understand the data. AI tools employed for constructing and using the Knowledge Graph are:

- Recurrent Convolutional Neural Networks RCNN
- Semi-Structured Data Parsing: Hidden Markov Model and Gene Sequencing algorithms

Graphs are essential because it is impossible to navigate through the ocean of unlike data available for modeling and analysis without some tools to illuminate the process. Graphs are about relationships and provide the ability to traverse far away relations easily and quickly, something for which relational databases are quite limited. The Important part of managing unalike data is finding relationships. Manual methods for finding relationships in unalike data is too limited to be effective. Technical metadata, such as column names is useful, but the magic is understanding the data to determine what it is. Without robust relationship analysis in extensive collections of data, error and bias are unavoidable.

Unifi's catalog energizes the data lake, turning it into an invaluable source of data. "Data of Interest" may be found within hundreds, thousands or even millions of data files. It would be impossible for a human or a whole team of humans to create a catalog of all of that data, not once, but continuously. The catalog would contain all of the typical technical data (metadata) but also many other useful pieces of information such as names of columns (not all files have columns), profiling of the data and, as we'll see in the next section, Unifi's semantic search and Natural Language Query, recommendation, transformation and a host of other invaluable functions.

Profiling data using Classifying techniques is another example of OneMind AI in action. Perhaps most valuably, in light of an ever-changing regulatory and compliance environment, auto-detection of PII (Personally Identifiable Information) in your data. When a data scientist is working with data sources that are familiar, this is not much of a problem, but introducing new sources can inadvertently reveal PII or, even more insidious, provide enough non-PII data to that the machine learning algorithm is able to de-anonymize the data though the latent values in the model, potentially causing ethical risk and even harm. Dataset and attribute classification is Unifi's ability to classify both the dataset itself as well as all attributes of a dataset, Unifi can auto mask, auto tag, and auto authorize and is serviced by a series AI tools:

- Classification is a technique to categorize data into a desired, and distinct number of classes and then assign a label to each class. Logistic Regression, K-Nearest Neighbor, ANN, Support Vector Machine
- Parsing Expression Grammar (PEG's) algorithms
- ML Regular Expression Framework

Business Value: The Unifi Data Catalog enables data to automatically be cataloged where it sits, preventing users from having to move data to a central repository or data lake, or engage in this time-consuming process manually. Metadata is collected and cataloged, no matter where original data sources currently exist. For an organization with a dysfunctional data lake, which seems to be prevalent, the Knowledge Graph powers features like the Data Catalog and Neural Language Processing (described in the next example). Powering an army of analysts to develop the applications that drive measurable results, like, embedding inference into a dynamic sourcing system, screening resumes of both candidates and employees for best fit or in financial institutions, fine-tuning reserves more efficiently. The catalog and associated features of Unifi can energize a latent data lake by providing data efficiently to those who need it and keeping the catalog fresh behind the scenes.

Natural Language Processing (NLP), Natural Language Query Example

In addition to a nicely-thought-out and intuitive user interface for interacting with the catalog, and support for API's to do so programmatically, Unifi also provides a Natural Language interface to the catalog where one can ask questions in their language about discovered objects and many other kinds of queries supported. For

example, NLP for the catalog can answer questions about:

- Metadata & Quality (show me a description of something)
- Statistics for datasets and columns (what's the average value of something)
- Governance (give me datasets that have PII)
- Permissions (show me users who have access to dataset <customer>)

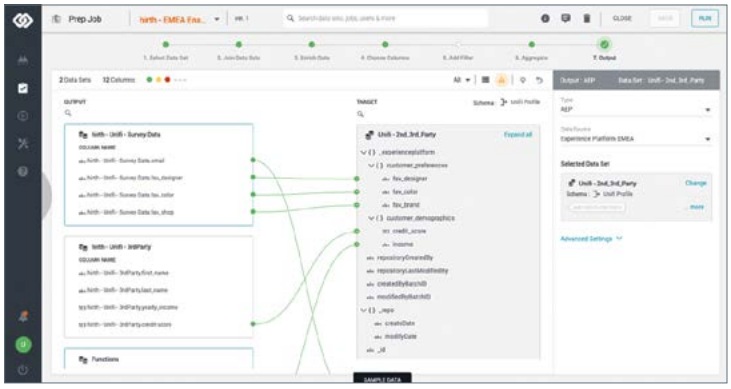
For example, an ordinary question can be asked and understood and serviced through all of the underlying complexity of the data. Keep in mind; no data is moved. The answers are derived from the data where it currently rests:

- Who accessed the PII data in the last three days?
- Who has access to these customer records?
- What are the PII columns that don't have masking defined for them?
- Which users have the highest anomaly in their login behavior in the past month?
- Who accessed this table in the past week, and what subsequent tables were derived from it?
- What attributes are most frequently used to develop a dynamic sourcing model?
- BI/Tableau (show me Tableau views for <revenue>)
- Glossary (show me glossary with <word> in title/ description)

It would be a mistake to assume that a static catalog, even one replete with valuable information, would be an adequate solution. What if your catalog could make recommendations for you, such as:

- People who searched for this also searched for this?
- Pointing out to you datasets that are subsets and supersets of each other?
- Showing you where redundant data sets exist?
- Suggesting that your search for auto recalls also suggested other datasets with related information such as accidents, recall costs, etc.?

'Who accessed PII data in the last three days?', 'Who has access to these customer records?', 'What are the PII columns that don't have masking defined for them?' or 'Which users have the highest anomaly in their login behavior in the past month?'. AI is not only used in finding patterns associated with the risk but also in making NLQ interface stronger and deeper.



AI-powered auto-mapping of attributes will substantially reduce a manual task fraught with the potential for human error. Unifi determines which attributes are an exact match, likely match and manual match. Intuitive drag and drop confirmation confirms the AI results.

The lift from that kind of technology can be measured in "time to value" – when companies set up a data catalog often the effort to populate the data dictionary and map data sets to the catalog can take months. The Unifi Knowledge Graph technology and the recommendations engine (described in the next section) compress the time it takes to establish business value and reduce data discovery time measured in hours rather than weeks or months.

The underlying structure of the data, derived from the Knowledge Graph described earlier, makes for more accurate recommendations from day one. If Unifi has determined that the customer data is related to the service history data by virtue of a unique customer identifier, when on day one someone in the business searches for "customer" they will automatically be recommended that they might be interested in the customer support history. By clicking on the recommendation their action reinforces the AI-driven recommendation and the algorithm is updated accordingly with an accuracy store. Without such an underlying method of recommendations the "learning" aspect of recommendations will take significantly longer as early use will be uninformed browsing which leads to false positives on human-augmented model learning.

Auto-Attribute Mapping – Recommendations

An organization has grown rapidly through acquisitions. As a result, there are many data science and AI development groups spread across not only geography but on operational focus. They are operating

independently. Data Science and ML use and generate a lot of data files. Starting with source files, they prune and merge and decorate until they have a file they can use for their model. Then they separate the file into training and result files, the latter to evaluate how well the model is predicting the outcomes from the training set. This is typically an iterative process, with many versions of similar files.

The datasets are typically "labeled," a laborious manual process, which gives the algorithm direction as it proceeds to its maximizing or minimizing its cost function. For example, if the model is being tested for its accuracy in finding a dog in a picture, the training set will label each instance with a dog. This way, the algorithm is "trained" to spot a dog. A significant part of the 80/20 phenomenon is this labeling process.

To take this a step further, suppose four of the six AI groups are using a copy of the same master dataset, and three of those groups are spending days and weeks labeling a billion rows in the file. Once the Unifi Catalog comes up, this duplication of effort is recognized, and all groups begin the properly labeled dataset.

Unifi can determine if datasets are related in several ways, but in particular, when one is a subset of another or a superset, and a perfect match or even a partial subset. Suppose an engineer needed twenty attributes for a model. He typically uses the same file which has 300 attributes. Unifi could recommend a file with 16 of the attributes and mapping the other four from the superset file. Once OneMind begins to understand the subject areas of interest to the developer, it can recommend this approach without being asked.

When you have hundreds, thousands, even millions of datasets to consider, it is clearly beyond human capability to map files for investigations. Doing it manually generally leads to repetitive use of the same sources and stale investigations.

How does AI provide recommendations? It begins to capture mappings that people already use because recommendations are governed and based on previous mappings (refer to earlier recommendation example).

The more the product is used at a customer, the more the product learns about that customer's usage patterns and, through observation, the more contextual information it gets about how different data objects and entities are related. Further recommendations are then provided to help the analyst perform such tasks as joining data sets (utilizing "OneClick Joins"), enriching the data (with "OneClick Functions"), choosing columns, adding filters and aggregating the data. Then, the algorithms convert the mapping recommendation problem into a machine translation problem.

AI approaches used:

- Encoder-Decoder architecture[3] for primitive one-to-one mappings
- Then using maximal grouping
- An Attention Neural Network (ANN) is used to resolve the recommendation.

Another feature of OneMind is the OneClick Function. With OneMind many of the steps a user must take to cleanse, enrich, parse, normalize, transform, filter, and format data prior to visualization are now entirely automated. When Unifi predicts how a user will want to join multiple datasets together for business insights, it recommends specific tasks for data cleansing and normalizing. A user can opt to have Unifi proceed automatically or manually direct each process.

For example, if a call center needs an auto dialer list to inform customers of a special offer or problem with the service (such as school closure), the list of phone numbers must be formatted so the auto-dialer can read them – Unifi's AI will recognize that an attribute is a phone number and recommend to the user of that data set that they apply a OneClick Function to normalize the number in the format required for downstream processing – in this case the auto-dialer.

Another powerful feature is AutoMapping, a recommendation that identifies attributes that are an exact match, possible match and not sure but these are still only recommendations, you must confirm your recommendation or change them manually then run the transformation.

[3] The encoder-decoder models in the context of recurrent neural networks (RNNs) are sequence-to-sequence mapping models. An RNN encoder-decoder takes a sequence as input and generates another sequence as output.

... The encoded representation is then used by the decoder network to generate an output sequence.

Mapping Recommendations – An Altogether Unique Experience

The intricate workings of ANN's (Attention in Neural Networks) are beyond the scope of this paper. If you don't mind a little code, this is the best description of how they work I've found.[4] It is a handy and powerful feature and is also used in Unifi for the NLP system for understanding searches/questions - because mapping is very like a language translation problem.

When you have hundreds, thousands, even millions of datasets to consider, it is clearly beyond human capability to map files for investigations. Doing it manually generally leads to repetitive use of the same sources and stale investigations.

How does AI provide recommendations? Recommendations are governed and based on previous mappings. Then, the algorithms convert the mapping recommendation problem into a machine translation problem using ANN techniques to resolve the recommendation.

Perhaps one of the most visually and powerful auto-attribute mapping examples manifests itself in the integration of Unifi with the Adobe Experience Platform. Designed by Adobe to capture and process thousands of attributes related to a user so as to profile that user more accurately or deliver compelling user experiences, the Adobe Experience Platform is destined to revolutionize how both B2C and B2B companies interact with their customers.

In just one example; a major hotel chain is using the Adobe Experience Platform to deliver unique customer experiences. Loyalty card customers can add their streaming music and video accounts to their profile. Now when they enter their room the smart speaker is playing their favorite music station or playlist and their TV screen is showing their favorite streaming services already logged in and ready to pick up the latest binge watching session from where they left off. This is just the start of the unique experience; coming soon home thermostat preferences set to adjust in-room temperature, in-room scent dispenser ready to emit the favorite fragrance right before the guest enters the room. Prefer a high floor, or a room away from the elevators, loyalty card premium customers will be able to choose their room just like they chose a seat on the plane. Their electronic key will be transmitted to their mobile device which is used to unlock the chosen room or presented with a last minute room upgrade proactively pushed to the mobile app.

For frequent travelers you can appreciate the value that these services will bring in personalizing your stay. For data geeks amongst the readers you'll appreciate the Herculean task of matching attributes from dozens of first and second party systems into a unified (no pun intended) customer profile containing all of these attributes. Then mapping the profile to the hotel's bespoke CRM system that handles the reservations and room allocations.

Drone On! Discovering Similar Data Sets

The FAA reported that in 2018 there were more than 122,000 commercial drone pilots and 878,000 hobbyists. Thus, there were more than 1 million total drone registrations last year. If every one of those drone pilots made just one flight per week that would correspond to over 50 million flights annually. Now add commercial and private pilot flights, plus military flight operations and the U.S. airspace is getting really crowded.

The task of capturing and analyzing all this flight data falls to the transportation group of the federal government. To help capture the data most commercial drone manufacturers are voluntarily submitting anonymized flight plans to the organization, one drone vendor has already provided over 50,000 separate CSV files – unfortunately the schema chosen by each manufacturer and for each model of drone varies. Plus the FAA already has a preferred flight plan schema used by commercial and private pilots to register IFR (instrument flight rules) and VFR (visual flight rules) flight plans.

The sheer volume of data makes the schema definition a nightmare. If we start with the FAA flight plan schema

[4] http://akosiorek.github.io/ml/2017/10/14/visual-attention.html

we must be able to determine which of the submitted flight plans is:

A. An exact match to the FAA flight plan schema

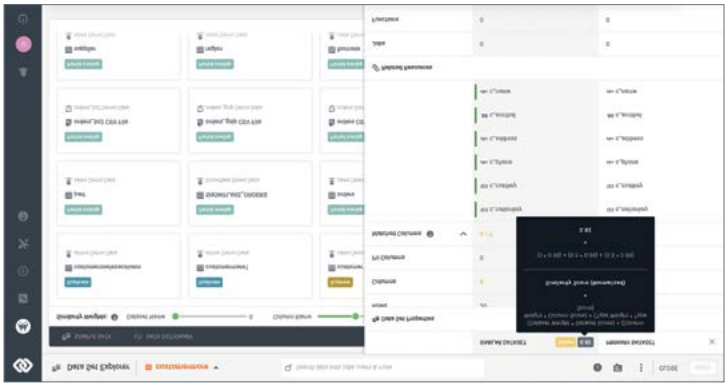
B. Is a superset of the schema

C. Is a subset of the schema

D. Is an overlap of the schema

E. Is one of the above with the same attributes given different names (i.e. semantic issues)

F. Is one of the above but the schema is in a different order



Similar datasets discovery can eliminate storage waste through duplication removal and drive a more curated, trusted data environment.

This is a perfect use case for OneMind AI and the Similar Data Set technology!

Conclusion: How Does All This AI Actually Help Me?

It is common in the industry today to proclaim that a software product is AI-enabled or AI-assisted. Some of these claims are pretty flimsy. Challenges today raise higher expectations on AI from management to be "data-driven" or a "digital enterprise," but it takes more than will. Our investigation revealed that AI practices are pervasive in the Unifi products, and can be enhanced and maintained easily by intelligently putting all of the AI methods into one module that can be applied to any other part of the product set. In addition to the "number crunching, curve fitting" types of models, Unifi charged ahead with exotic neural networks, Natural Language Processing (NLP), so you can ask questions in your own language, and perhaps most of all, a way for the sometimes mysterious operations of AI, a kind of "Cooperative AI", where the decisions of people using the system are fed back to the algorithms.

Benefit to You in Your Daily Operations

Reducing the number of files by pruning duplicates and artifacts no longer used can save some cost in storage, but with storage costs asymptotically approaching zero, it is not a compelling benefit. On the other hand, cutting the workload of your data scientists and AI engineers, allowing them to spend more time developing models, is incalculable. These professionals are in short supply and the best ones in extremely short supply. They will gravitate to environments where they can be challenged and be most effective.

About the Author

Neil Raden, based in Santa Fe, NM, is an active industry analyst, consultant and widely published author and speaker and also the founder of Hired Brains Research. Hired Brains provides thought leadership, context and advisory consulting and implementation services in Information Management, Analytics/ Data Science, AI, AI Ethics, and IoT Edge Analytics. Hired Brains also provides consulting, market research, product marketing, and advisory services to the software industry. Neil is Principal Investigator and author of "Ethical Use of Artificial Intelligence for Actuaries," sponsored by the Society of Actuaries and is the co-author of Smart (Enough) Systems: How to Deliver Competitive Advantage by Automating Hidden Decisions, Prentice-Hall. His articles can be found at **www.diginomica.com/author/neil-raden**. He welcomes your comments at **nraden@hiredbrains.com**.