Case study 2 : Montefiore semantic data lake

1. Sources

1. The semantic datalake in healthcare, Webinar, Parsa Mirhaji, 2016.
2. **Montefiore Creates Data Analytics Platform to Advance Patient Care**
   Intel solution brief, posted as separate file.
3. **Semantic Big Data Lakes Can Support Better Population Health**
   news article, Jennifer Bresnik, 2015.
4. **Montefiore Semantic Data Lake Tackles Predictive Analytics**
   news article, Jennifer Bresnik, 2016.
5. **Semantic data lake architecture in healthcare and beyond**
   news article, G. Anadiotis, 2017.

The webinar (first +- 45 minutes) is the most important source (fairly strongly accented American English !), explaining data needs, organisation of the graph database and ontologies, and other features.

The short Intel brief is well done and provides succinct information on the needs and the solution.

The news articles can be skimmed through to glean information items on the context, status of progress at different points in time, and technology solution.

A large extract of these articles is provided in section 3 below. Some less relevant sections have been omitted, but the links to the full articles are provided for completeness.

2. Background vocabulary

ICD9 and ICD 10 : International Classification of Diseases codes.

HL7 : Health Level Seven : standards for the exchange, integration, sharing, and retrieval of electronic health information.

EHR : Electronic Health Record.

ACO : Accountable Care Organisation : an organization of health care practitioners that agrees to be accountable for the quality, cost, and overall care of Medicare beneficiaries; ties payments to quality metrics and the cost of care.

3. Article 1.

Semantic Big Data Lakes Can Support Better Population Health

Jennifer Bresnik, HeathITAnalytics, August 2015

https://healthitanalytics.com/news/semantic-big-data-lakes-can-support-better-population-health

ANALYTICS IN ACTION NEWS

Semantic Big Data Lakes Can Support Better Population Health

September 08, 2015 – [small section omitted].

As healthcare providers [few words omitted] try to predict how future regulations will shape their actions, the need to lay the groundwork for advanced population health management and accountable care is only becoming clearer.

No matter what the outcome of debates about the future course of the EHR Incentive Programs, one thing remains abundantly clear for organizations of all shapes and sizes: advancements in healthcare big data analytics will not be driven solely by rules and mandates, but by the pressing financial need to collect, corral, understand, and leverage information in order to refine and expand population health management techniques.

Developing the underlying architecture for value-based reimbursement, namely a strong framework for population health management, data governance, and big data analytics, is becoming a top priority for a growing number of providers looking to get a head start on the new realities of healthcare reform.

These organizations, like Montefiore Medical Center, are looking for cutting edge analytics tools which won't just help them meet the clinical and financial stresses of today's environment, but will also prepare them for the uncertain paths ahead.

In order to position Montefiore for success in the unpredictable future, Parsa Mirhaji, MD, PhD, Associate Professor of Systems and Computational Biology and the Director of Clinical Research Informatics at the Albert Einstein College of Medicine and Montefiore Medical Center-Institute for Clinical Translational Research, turned to one of the most promising developments in big data analytics: semantic data lakes.

A few steps beyond the traditional data warehouse, which requires a relatively narrow set of parameters in order to accept, compare, and retrieve information, semantic data lakes allow for unprecedented flexibility.

Data lakes, built using graph database technology, may soon allow clinicians to access sophisticated clinical decision support using natural language queries, thanks to their unique ability to synthesize and normalize disparate datasets and draw conclusions from seemingly unrelated pieces of information.

The potential for improving the quality of patient care is enormous, Mirhaji said in an interview with *HealthITAnalytics.com*, as the technology evolves to meet the full spectrum of as-yet-untold demands for detailed risk stratification, predictive analytics, and patient safety.

"From the standpoint of an accountable care organization (ACO), where we really need to cover the full spectrum of health data, we need to capture and represent everything that will give clinicians a precise understanding of an individual's care and wellness," Mirhaji said. "On one hand, that includes diving into clinical genetics, molecular medicine, and biomarkers."

“On the other hand, patients interact with their environments and with each other in a community setting, which makes it very important to look at population health management and community care at the other end of the spectrum. For an ACO, it’s all about the coordination of care within different communities.”

But healthcare providers cannot look to coordinate care in the community if they do not have an organized method for keeping their own house in order. From EHRs to research results to financial data and patient demographics, big data is everywhere in the typical healthcare organization, and each type of data may be locked into its own individualized analytics architecture.

Not only is it expensive and time-consuming to craft separate infrastructures for each category of information, but it prevents data scientists from drawing actionable insights from cross-pollinated datasets.

“We don't have the time and resources to build silos and specialized systems for specific needs,” Mirhaji said. “So we did a rigorous analysis of the use cases we need to cover, starting from personalized precision medicine and moving all the way up to population health management.”

“We asked ourselves what functional competencies and technical competencies we need in order to support all of these. Where do we need to make investments, and what are the properties of the technologies that we need to support more than one use-case at a time? A lot of our problems could only be solved by graph databases. Other technologies we looked into were not really prepared to address this long spectrum of requirements.”

A standard relational database can help many organizations meet a number of their goals, Mirhaji says, but they include some inherent limitations. “Relational databases require a very fine structure that you have to plan out before you can use it - you have to frame your problems in a very specific way,” he explained. Within that frame, you can do wonderful things, but you have to pre-coordinate your schema before you start investing in application development and data management.”

“The problem with that is that you have to predict all future-use cases,” he continued. “And the costs of changing your mind or your requirements are huge. And that's why you end up with these data silos. You end up with different architectures for different problems, because you have to box the problem before you begin.”

In contrast, flexibility and adaptability are built into the fabric of graph databases, which use cognitive computing techniques to help draw connections across datasets that may be vastly different in size, detail, or scope.

“You don’t have to predict the future,” says Mirhaji. “You can start from where you are, from exactly where you are, based on the kinds of needs that you have right now with the confidence that it will grow into the dimensions and directions as your organization wants to grow.”

What allows graph databases to operate with such a high level of fluidity? It’s the way that data points are identified, codified, and linked within the system, explained Dr. Jans Aasman, CEO of Franz, Inc., which has worked with Montefiore to develop its big data capabilities.

Semantic data lakes are structured in a fundamentally different way than relational databases. In a semantic graph system, each element is given a standardized, unique identifier, which allows the database to link separate concepts and generate complex insights the way a human brain does.

[section omitted explaining concept of triple]

Jans Aasman

Parsa Mirhaji

But that is only the beginning of what this type of technology can do, Aasman says. “The important thing about semantic graphs is that every node is not just a simple word, but it’s actually a Unique Resource Identifier (URI) that can globally identify and contain data on a whole concept.”

“For example, when we talk about aspirin, we have a standardized URI for the concept of ‘aspirin.’ We might want to use a clinical trial database that talks about aspirin, and a disease database, and a side effect database – we will all use the same URI for the same concept across all those datasets.”

“That means we can load up all these databases into a single location, and suddenly we can start making connections across disparate sources because they all took care to describe ‘aspirin’ as a standardized element shared between them.”

[section omitted]

These capabilities can be especially important for organizations that may handle very complex cases and rare diseases, or simply experience something out of the ordinary that might initially give a physician pause. When it comes to patient safety issues involving medications, graph databases can help providers take a precision approach to solving an unusual adverse reaction or forestalling a large-scale event originating from a specific manufacturer or dosage problem.

“You might have a code that tells you what medication was given to a specific patient,” Mirhaji conjectured. “That code can also be mapped to the National Drug Codes, which will tell you who made that medication and the specific formulation that went into it.”

“And then there are pharmaceutical knowledge bases that have the chemical structure of those medications. And then there are databases at the FDA that have data on known complications or the ingredients in those medications. Then there are clinical trials that are actually doing observations and trials on the effects of those medications or ingredients on different diseases and clinical problems.”

“Now, imagine that all of these datasets are different on their own,” he said. “Each is its own type of tree, as it were. Different people with different backgrounds are building them. You can’t predict how each developer is going to make changes.”

“But you want to be able to combine them to ask a complicated question, such as, ‘How many patients currently admitted to the hospital have been given a medication that contains ingredients tested by a clinical trial that may produce this specific complication, and what company is marketing the formulation used by those patients?’”

“So now we have combined five different databases together to produce a query that resembles very much like how a clinician thinks,” added Mirhaji. “As a clinician, I may think that I have given my patient a medication that might explain why he has a new rash on his face; and I may wonder if this has a precedence in a clinical trial. I may want to know if all formulations of the medicine in the market have this complication or they happen more or less in products of a certain company.”

[Image omitted – same image in article 3]

Researchers have not yet refined the ability to input such a query into a simple text-based interface linked to a clinical decision support window in an EHR. But when the capability is fully developed, clinicians will have a powerful ability to get answers to on-the-fly questions that require little in the way of specialized knowledge about how the database works.

“As a clinician, I probably don’t want to know the schema behind this data,” Mirhaji says. “I just want to know if there is a clinical trial somewhere that addresses the potential impact of some ingredient in a medication that might be producing an adverse effect.”

“I want a machine to do it for me. I don’t want my clinicians to have to know and think about whether the data exists and in what format and how exactly if it can be connected to some other piece of data, before they can even start asking a question. I want my clinicians to think freely and get the answers they need.”

Before that can happen, however, graph database technology needs to address some of its biggest challenges. Standardizing elements across different healthcare terminologies, such as ICD-10, LOINC, CPT, and SNOMED, is problematic for most health IT applications, even the most advanced EHR systems. The same basic interoperability concerns apply to graph databases, Aasman and Mirhaji acknowledged.

Additionally, in order to use a semantic data lake for meaningful population health management, users must also be able to get answers to queries that may include multiple events that take place at different times, in specific locations, or in certain sequences.

The system must be able to incorporate temporal reasoning that arranges events in relation to one another, Mirhaji explains. "Some events happen within the other, one after the other, or overlap in duration of the other. There are meaningful inferences to be made if you know how exactly these events are temporally arranged."

"Geospatial relationships are another thing that require a specialized approach," he added. "Combining time and space is especially important for community-based population health management. Behavioral data, mobility questions, and health disparities all require tracking where and when an event takes place. We cultivate our data in such a way that it knows how to account for the temporal and geospatial relationships between events."

[section omitted]

4. Article 2

Montefiore Semantic Data Lake Tackles Predictive Analytics

Jennifer Bresnik, HeathITAnalytics, May 2016

https://healthitanalytics.com/news/montefiore-semantic-data-lake-tackles-predictive-analytics

May 31, 2016 - Semantic computing is becoming a hot topic in the healthcare industry as the first wave of big data analytics leaders looks to move beyond the basics of population health management, predictive analytics, and risk stratification.

This new approach to analytics eschews the rigid, limited capabilities of the traditional relational database and instead focuses on creating a fluid pool of standardized data elements that can be mixed and matched on the fly to answer a large number of unique queries.

Montefiore Medical Center, in partnership with Franz Inc., is among the first healthcare organizations to invest in a robust semantic data lake as the foundation for advanced clinical decision support and predictive analytics capabilities.

Six months after introducing the concept to readers at *HealthITAnalytics.com*, Parsa Mirhaji, MD, PhD, has provided an update on Montefiore's progress with a sophisticated, potentially revolutionary predictive analytics pilot program.

"The Semantic Data Lake is up and running, and it's doing well," said Mirhaji, Associate Professor of Systems and Computational Biology and the Director of Clinical Research Informatics at the Albert Einstein College of Medicine and Montefiore Medical Center-Institute for Clinical Translational Research.

"Right now, we are still in the middle of a pilot program that uses predictive analytics to flag any patient hospitalized at Montefiore Health System locations, who is at risk of death or of the need for intubation within the following 48 hours, which is the window of opportunity to complete an effective intervention for the course of events."

As part of a collaboration with the Mayo Clinic, Montefiore is in the process of refining a predictive algorithm founded on retrospective data from more than 68,000 patients across the two institutions. The data lake delivers real-time data for perspective surveillance on real patients, Mirhaji says, using actionable clinical data.

"It creates risk scores based on the patient's likelihood of a major event within 48 hours," he explained. "Then there's another engine that kicks in based on those risk scores and other factors to determine what we can do for that particular patient to avoid the crisis. It can send a personalized checklist of proposed interventions to the practitioner in charge of that case."

At the moment, the system is still in its pre-clinical validation stage. The algorithm is working in parallel with the traditional care delivery process to test its capabilities, but clinicians are not currently receiving notifications for their patients.

Instead, results are being sent to a group of clinical investigators who are comparing the predictive analytics with real-life patient care procedures to see how well the system is working.

"We are very happy with what we're seeing right now, as the information is very sensitive and very specific," Mirhaji added. "We can find almost all the high-risk patients in our population with only a one percent error, which is a very good result."

With such impressive early progress, go-live is slated for July of 2016, he said.

[section omitted]

Montefiore's semantic computing infrastructure may be able to do much more in the future than flag crisis patients, he added. In conjunction with several clinical partners, the New York-based health system is looking into how the database could aid diabetes management and provide support for patients with sleep disorders, such as apnea.

“Additionally, we are investigating a way to predict behavioral health needs for Montefiore patients to see if there is a relationship between behavioral or mental health issues and outcomes of care,” said Mirhaji.

“That will help us better manage patients with these needs and make improvements to the care delivery system to taking these variabilities into account.”

Medication reconciliation and discharge education are also on the horizon, he said, as well as other use cases involving continuous monitoring and multiple streams of data.

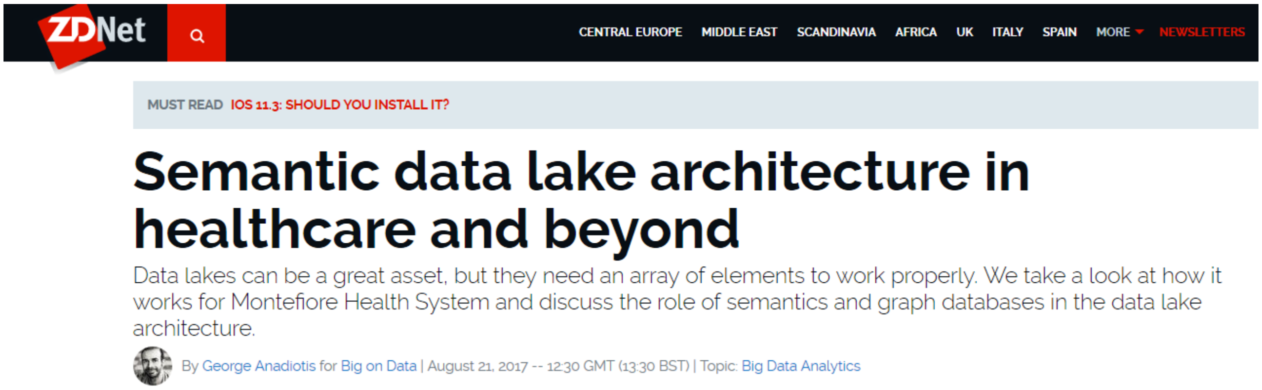
[section omitted]

5. Article 3

Semantic data lake architecture in healthcare and beyond

George Anadiotis, ZSNet Big on data, 2017

http://www.zdnet.com/article/semantic-data-lakes-architecture-in-healthcare-and-beyond/



MUST READ IOS 11.3: SHOULD YOU INSTALL IT?

Semantic data lake architecture in healthcare and beyond

Data lakes can be a great asset, but they need an array of elements to work properly. We take a look at how it works for Montefiore Health System and discuss the role of semantics and graph databases in the data lake architecture.

By George Anadiotis for Big on Data | August 21, 2017 -- 12:30 GMT (13:30 BST) | Topic: Big Data Analytics

Data lakes stink. That's because lots of them turn to data swamps, and swamps stink. What's the difference between a data lake and a data swamp?

A data lake is built on top of cost efficient infrastructure. More often than not these days this is Hadoop, leveraging two of its most alluring properties: Lots of storage for cheap and schema-on-read. That means you can store all your data and more now and worry about it later.

And that's exactly what many organizations end up doing, resulting in a data swamp. A data swamp is a data lake where data goes to die: Without descriptive metadata and a mechanism to maintain it, you get a big pile of data that is effectively unusable.

A part of this has to do with Hadoop, as support for metadata and data governance has been one of its sore points. The situation there is improving, but there still are a couple of issues.

The first one is obvious: Even the greatest tools are no use if you don't put them to use. So the fact that there is the option to add metadata to your data does not mean that everyone does it.

The second one is that not all metadata are created equal. When we start talking about descriptive metadata, the need for semantics quickly becomes pronounced. So what do we get when we add semantics on top of data lakes? Semantic data lakes.

Healthcare challenges, Semantic Data Lake solutions

Montefiore Health System has implemented a semantic data lake (SDL), and we discuss with Franz Inc., the provider of the semantic element, about overall architecture and the role of semantics.

Located in the Bronx, Montefiore Health System serves one of the most ethnically and socioeconomically diverse populations in the US. The complex includes, but is not limited to, Montefiore Medical Center, Albert Einstein College of Medicine, and a research facility.

Like all healthcare organizations, Montefiore faces many data-related challenges. As Dr. Andrew D. Racine, system senior vice president and chief medical officer at Montefiore puts it:

"The challenge where you've got hundreds of thousands of patients impacting the institution at any given point is to have the appropriate information about each one of those patients at the fingertips of the therapist who's interacting with them at the time of that interaction."

Montefiore is using its varied and vast amounts of raw data for deeper analysis to flag patients who are at risk or help clinicians identify optimal treatment plans. In order to be able to build such advanced analytics solutions, Montefiore has deployed a Semantic Data Lake (SDL) utilizing an array of technologies and components.

The SDL solution provides capabilities that include:

- Predictive analytics at scale to anticipate and account for various patient outcomes in timeframes in which treatment can be administered to affect care.
- Machine learning algorithms to integrate the results of previous outcomes that significantly impact the analysis and effects of future patient objectives.
- Disposable data marts to quickly provision project-specific environments to manipulate data and analytics results without duplication or redundancy.
- Ontological pipeline to rapidly integrate new data sources and requirements into existing models, and validate the clinical process for highly targeted patient subsets.

An ontological data pipeline

An ontological data pipeline sounds fancy, but what is it exactly and why should you care? It's a data pipeline in which incoming data is annotated with metadata using an ontology. An ontology is arguably the most advanced form of schema around in terms of its ability to capture semantics, hence the semantic aspect of the data lake.

We discussed the approach and architecture with Dr. Jans Aasman, CEO of Franz, Inc. Franz Inc. is the vendor behind AllegroGraph, the RDF graph database that handles the descriptive metadata/ontological pipeline aspect of the solution.

Aasman explains that the SDL supports both fast real time input (for example HL7 streams) and large, batch oriented bulk inserts from ETL (Extract Transform Load) processes.

But the million-dollar question is how does the semantic annotation happen. Are all data that enter the lake already annotated upon ingestion, or is there further annotation required? How is it performed -- automatically, semi-automatically, manually? Are there tools for this?

Aasman says they use a visual ETL tool to draw a mapping between data in the EDW or HL7 streams to a healthcare ontology that covers everything that could ever happen to a patient in the hospital life cycle.

[figure omitted]

"This creates a declarative mapping that is read in by a Java program that automatically transforms (mostly) relational data into a graph representation (aka triples). Every element in the graph is annotated by the table and column it came from and the ETL date.

"In addition, we annotate every triple with what we call 'triple attributes' that enable us to selectively make data available for users in their different roles. This is a spectacular new feature in AllegroGraph that we will be publicly announcing soon.

"In this setting, vocabulary management is extremely important. Healthcare has more than 180 vocabularies, taxonomies and terminology systems, such as Mesh, Snomed, UMLS, LOINC, RxNorm, etc."

Data integration is one of the strong points of ontological modeling, and Aasman says that these taxonomies are all interconnected and linked to important 'real life' concepts like ICD9 and ICD10, procedure codes and NDC for medications:

"This combined and integrated terminology system (the healthcare ontology) is at the heart of the ETL process, and is incredibly important for queries and analytics," he says.

SPARQL over Spark

Ontologies and graph databases sound great and all, but there's more to the SDL solution. Where and how exactly does ontological modeling and AllegroGraph fit in the big picture?

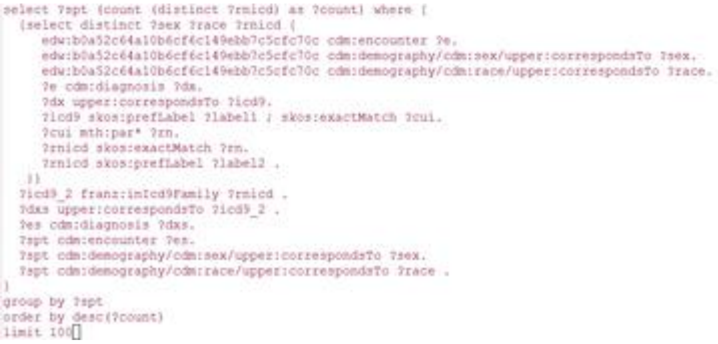
Aasman explains: "We run distributed AllegroGraph on a Cloudera cluster. We can read/write from HDFS and we can run Spark on top and use MLlib for our analytics. Distributed AllegroGraph, the database underneath the SDL architecture, provides all the features of a Lambda architecture."

That's an unusual choice, one which means for example that instead of SQL, SPARQL is used as the query language. Why go for it? And how well does it perform compared to more conventional solutions?

"Relational databases do great when your data fits in relatively simple schema, there is no network in your data and you do big aggregate queries. Graph databases do better when you do graph algorithms where it is unpredictable how deep your graph algorithm will go.

"In addition, graph databases perform far better when you have a lot of ad hoc queries or when your data is ridiculously complex or if your application will benefit from reasoning," Aasman says.

What about query complexity? Aasman says that as a vendor they see queries ranging from one line to 1,500 lines of code, and provided a typical SPARQL query from the Montefiore project for good measure:



```
select ?spt (count (distinct ?rnicd) as ?count) where {
  {select distinct ?sex ?race ?rnicd {
     edw:b0a52c64a10b6cf6c149ebb7c5cfc70c cdm:encounter ?e.
     edw:b0a52c64a10b6cf6c149ebb7c5cfc70c cdm:demography/cdm:sex/upper:correspondsTo ?sex.
     edw:b0a52c64a10b6cf6c149ebb7c5cfc70c cdm:demography/cdm:race/upper:correspondsTo ?race.
     ?e cdm:diagnosis ?dx.
     ?dx upper:correspondsTo ?icd9.
     ?icd9 skos:prefLabel ?label1 ; skos:exactMatch ?cui.
     ?cui mth:par* ?rn.
     ?rnicd skos:exactMatch ?rn.
     ?rnicd skos:prefLabel ?label2 .
   }}
  ?icd9_2 franz:inIcd9Family ?rnicd .
  ?dxs upper:correspondsTo ?icd9_2 .
  ?es cdm:diagnosis ?dxs.
  ?spt cdm:encounter ?es.
  ?spt cdm:demography/cdm:sex/upper:correspondsTo ?sex.
  ?spt cdm:demography/cdm:race/upper:correspondsTo ?race .
}
group by ?spt
order by desc(?count)
limit 100
```

A real-world SPARQL query from the Montefiore use case. (Image: Franz Inc.)

"This query finds the top 100 patients that are most similar to one particular patient from a set of 2.7 million patients. The first subquery finds for a particular patient his or her gender and race and all the icd9 codes.

"Because these icd9 codes are very specific, we link the icd9 codes to concepts in our knowledge base and we go up the terminology ladder recursive way and then down again to find all family members of that icd9 code.

"Once we have those we find all the other patients that have the highest overlap in icd9 codes (well, the super members) with our start patient. This is another example of the compactness of SPARQL.

"We can also use Spark to do a SPARQL query against distributed AllegroGraph. We use Spark for analytics and then we can save the results of analytics back into AllegroGraph as newly learned information," he says.

The SDL supports both fast real time input and large, batch oriented bulk inserts from ETL processes. AllegroGraph is an append only graph database, explains Aasman, so new data are appended to the existing indices:

"There are continuous background optimization processes that merge all the chunks of data into one linearly sorted index space, but the reality is that if data is streaming 24/7 the indices are never perfectly sorted so the query engine has to look both in the existing indices and appended new chunks."

Graph browsers, time machines and machine learning

Aasman adds that Gruff, AllegroGraph's graph browser, allows users to visually create a query and then generate SPARQL (or Prolog) query code. Franz Inc just released a new version of Gruff, adding what they call "Time Machine" capabilities to it.

Many use cases for graph databases involve temporal events. Events are modeled as objects that have a start time, end time, a type, some actors and a geospatial location.

Aasman says Gruff v7.0's new time slider feature enables users to visually demonstrate how graphs comprised of temporal events are constructed over time, allowing time machine like exploration of your data.

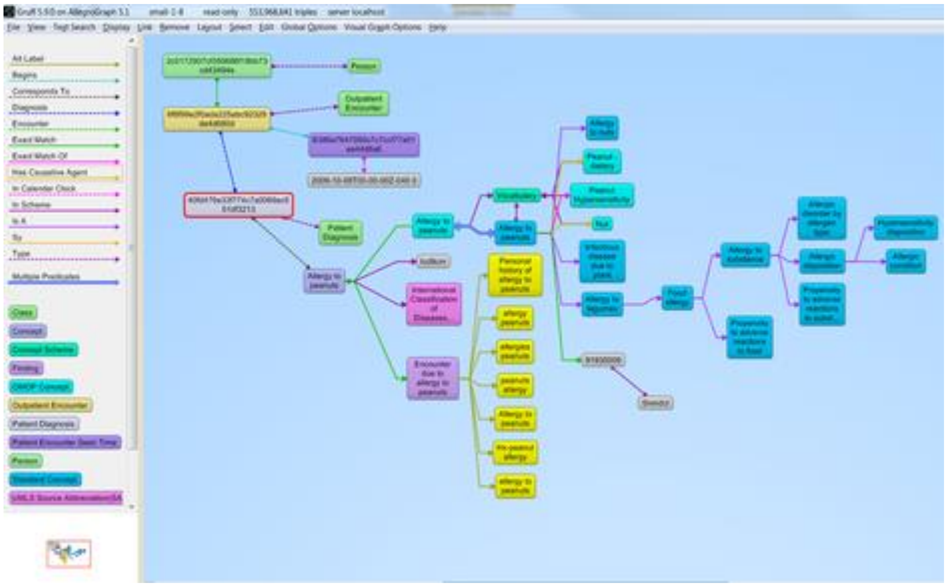
Last but not least, the Machine Learning part. This is not something graph databases typically offer, so how does it work for AllegroGraph?

Data scientists don't really care what they do their analytics against, claims Aasman, as long as they can get their feature sets from the underlying data store as a csv file, or even better, as a (panda) data frame.

"To make life more simple for data scientists that want to work with AllegroGraph we currently have an open source R interface and an open source AllegroGraph - Python interface that is directly installable via Anaconda.

"However, we have an even better integration point and that is that we put all the results of analytics back in AllegroGraph as triples and then make that navigable via Gruff.

"See an example below. Not only do we store all the results, but also the metadata about the results, such as: who did the analysis, when, what scripts were used, what data sets were used, etc," he says.



Rich metadata is one of the benefits of semantic data lakes. (Image: Franz Inc.)

[Remaining part of article omitted]