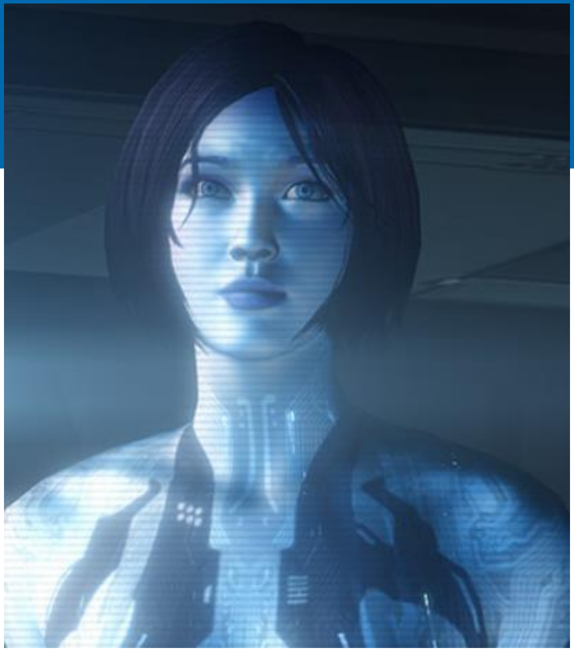
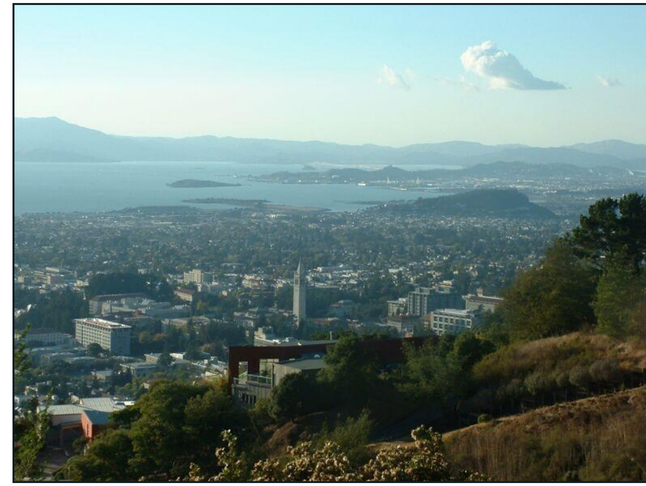
A scalable AI Knowledge Graph Solution for Healthcare (and many other industries)



Dr. Jans Aasman

FRANZ INC.

About Franz Inc.



- Privately held, Self-funded, Profitable since 1984
- Headquartered: **Oakland, CA**
- Flagship product: **AllegroGraph**
- Compiler History**:** **Common Lisp and Prolog** for **Artificial Intelligence and Complexity**
- Industries: **Healthcare, Intelligence/Defense, Pharma/Biotech**
- Key Customers: **Siemens, BlueCross/BlueShield, Novartis, Wells Fargo, Credit Suisse, AstraZeneca**
- Strategic Partners: **BAE, Intel, Cloudera, Montefiore Health Systems Mitre, Smart Logic, MongoDB, TopQuadrant, Expert Systems**

FRANZ INC.

Franz Clients - DoD and Intelligence

BOEING

NATIONAL SECURITY AGENCY · UNITED STATES OF AMERICA

OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE · UNITED STATES OF AMERICA

U.S.ARMY

U.S. AIR FORCE

SRA INTERNATIONAL, INC.

MACE

Multi-Agency Collaboration Environment

DARPA

ESS

NORTHROP GRUMMAN

Mission Systems

Modus Operandi

Raytheon

CTC Concurrent Technologies Corporation

IDA

AIR FORCE RESEARCH LABORATORY

LOCKHEED MARTIN

National Counterterrorism Center

United To Protect

MITRE

BAE SYSTEMS

NASA

MARINES

THE OFFICIAL WEBSITE OF THE UNITED STATES MARINE CORPS

ONR

Office of Naval Research

Elbit Systems

Australian Government

Department of Defence

Defence Science and Technology Organisation

SRI International®

DISA

APL The Johns Hopkins University Applied Physics Laboratory

JOINT INFORMATION OPERATIONS WARFARE CENTER

JOINT WARFARE ANALYSIS CENTER · JWAC

ManTech International Corporation

Booz | Allen | Hamilton

Knowledge Graphs in the intelligence world

- Police: advanced tracking of People of Interest.

- Recording and annotating meta data about satellite pictures

Franz Clients - Commercial

xerox
SIEMENS
Wolters Kluwer
University of Phoenix
TOHOKU UNIVERSITY
at&t
EPIM
University of California San Diego
AstraZeneca
gsk GlaxoSmithKline
Pfizer
The University of Texas MD Anderson Cancer Center
Making Cancer History
MERCK Be well
Lilly
South Carolina
The University of Texas Medical School at Houston
NOVARTIS
MAYO CLINIC
PALMETTO HEALTH
JOHNS HOPKINS UNIVERSITY
PayPal
citi
Bank of America
WELLS FARGO
J.P.Morgan
University of Yamanashi
TELEONTO
Kodak
Rensselaer
DBpedia Polska
AutoZone
Ford
GE Global Research
Modus Operandi
MIMOS
National Archives of Korea
Noah诺亚舟
Duke University
The Redlands Institute
cisco
KT
Adobe
amdocs
Booz | Allen | Hamilton
KILPATRICK TOWNSEND
OECD BETTER POLICIES FOR BETTER LIVES
Franz

A knowledge graph in healthcare

Addressing value-based healthcare with Intel® Xeon® processors and Franz AllegroGraph*

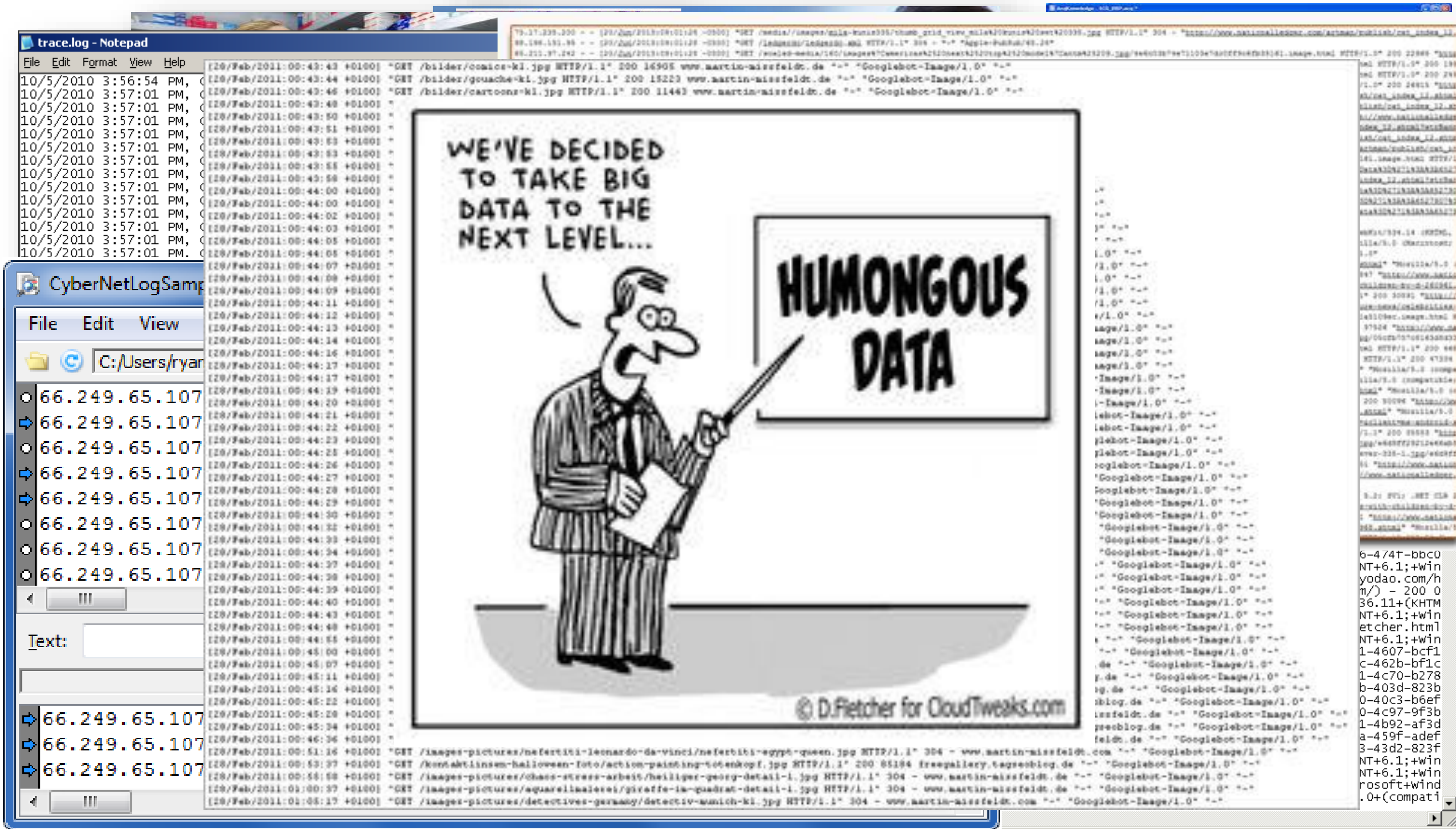
Challenge

Located in the Bronx, Montefiore Health System serves one of the most ethnically and socioeconomically diverse populations in the US. The complex includes the Montefiore Medical Center, the Albert Einstein College of Medicine, and a research facility. Unlike a pay-per-service model, as an accountable care organization Montefiore delivers value based on patients' long-term health—during their hospital or clinic visit and after they return to the community.

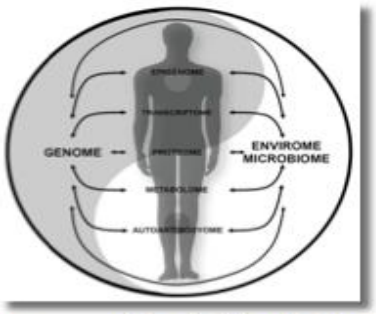
Montefiore/Franz built a Knowledge Graph for Healthcare

- A single data centric platform that can serve any type of analytic without building a new data mart for every new question.
- Currently 2.7 million patients with 10 years of data
- Fit for both transactional and analytical processing.

But first, we need to integrate all Electronic Medical Records (EMR)



One Knowledge Graph environment for all healthcare analytics



Personalized Medicine

Predictive Modeling

Translational Research

Fraud Detection

Risk Assessment

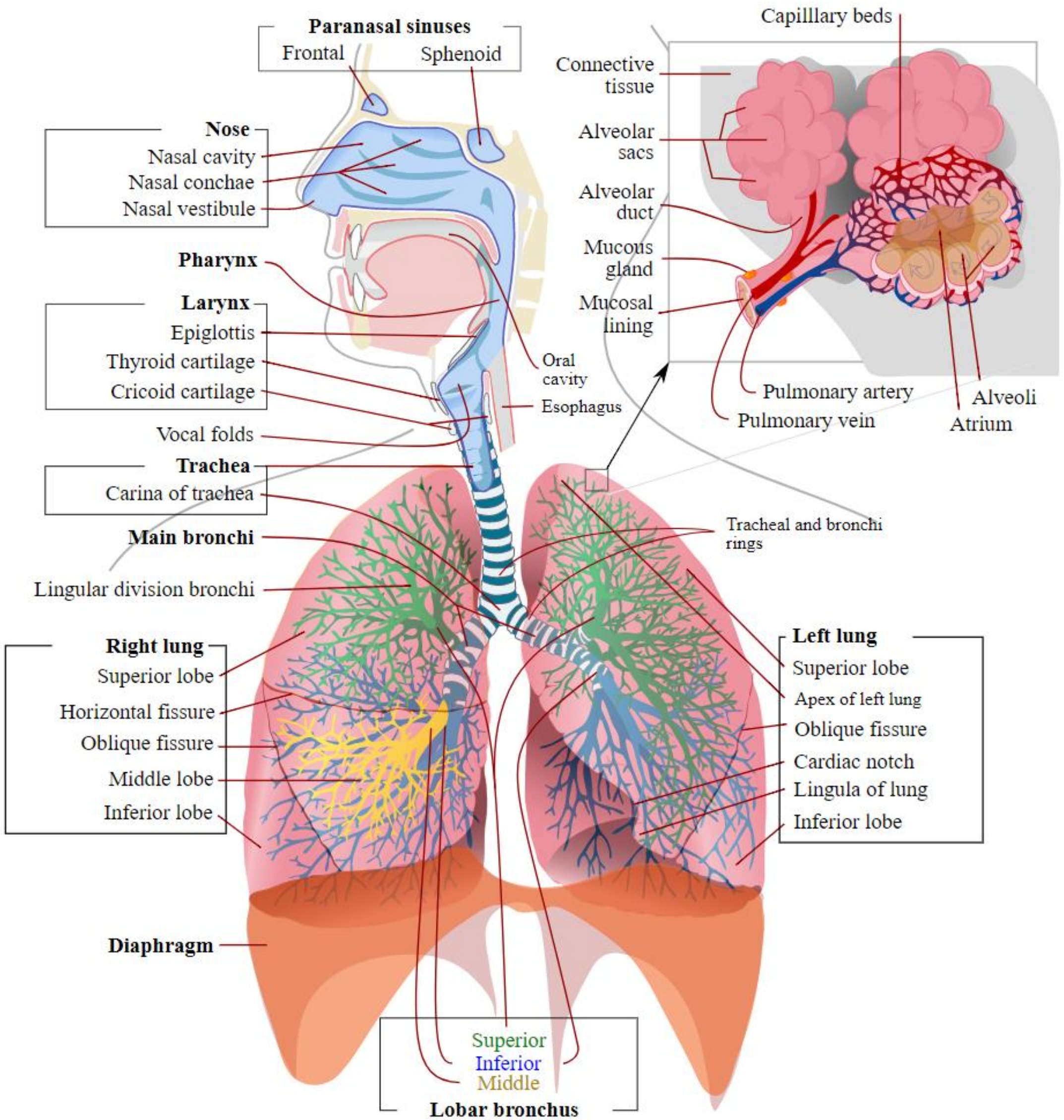
Business Intelligence

Public Health

Mobile Health

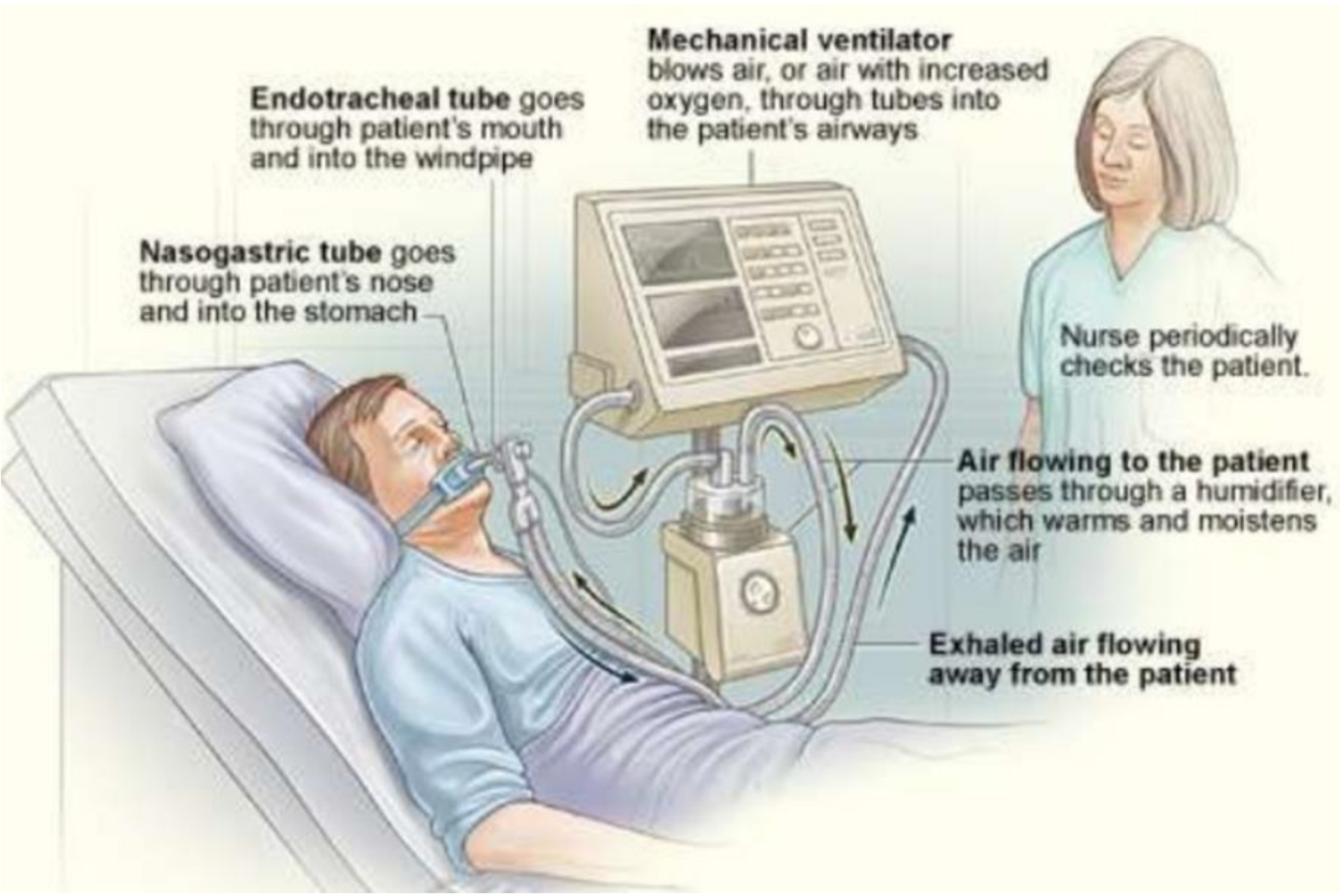
Decision Support

Respiratory failure is very dangerous. More than 40 % mortality even if it occurs in hospital

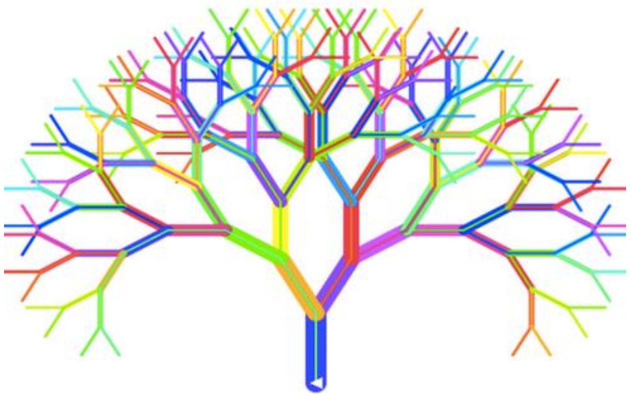


FRANZ INC.

Needs instant intubation if at risk; but when?



Montefiore Hospital developed a predictive model for respiratory failure on top our Knowledge Graph



- A random forest model detects respiratory failure up to 48 hours before the event. Faster than doctors and nurses
- The model uses 46 complex variables, a doctor in general looks at 3 to 4 variables.
  - A mix of lab results, (real time) measurements, diagnostics, other
  - Complex: give me largest difference between the value for SerumCalciumLevel and the midpoint 9.5 in the range of 0 to 20 for the last 24 hours.
- Greatly reduce unnecessary intubation with same clinical outcome
- Save ~$4 million per year

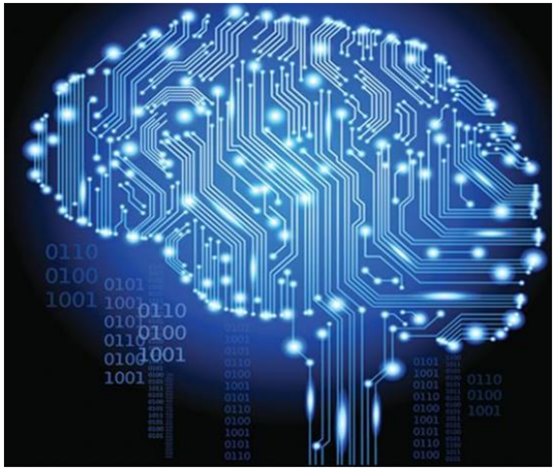
FRANZ INC.

Ok: so now let’s talk about

- Artificial Intelligence and

  Knowledge Graphs

  for Enterprise Data Warehouses



FRANZ INC.

Some people say the end of EDW is near

It's the End of the Data Warehouse as We Know It

According to TDWI survey data, about half of all enterprises expect to replace their data warehouse systems -- in some cases, their analytics tools, too -- over the next three years. What should they replace them with?

By Steve Swoyer
January 11, 2017

According to TDWI survey data, about half of all enterprises expect to replace their data warehouse systems -- and in some cases, their analytics tools, too -- over the next three years.

These systems could (and probably should) be replaced by modern data warehouse systems that -- like the database equivalent of a Swiss Army knife -- integrate multiple fit-for-purpose analytics engines. These systems could (but probably should not) be replaced by Hadoop and other NoSQL platforms, which are no less Swiss Army-like.

It is misleading to frame this as a question of rip and replacement, however; the issue isn't a zero-sum one. It's more complicated than that.

[1] Data sprawled all over heterogeneous silos

Imagine you do this in your EDW:

```
Give me every bit of information on this [mobile phone user, patient,
product] since her first occurrence in our EDW sorted on time and
returned as a structured object.

Find the top 100 patients that are most similar to this patient, based
on tests, genetic make up, diagnostics, procedures, medications,
treatments, and outcomes. Take into account temporal unfolding of
these events.
```

- Data about entities and events might be in more than 4000 different tables and 20,000 columns.
- Just the many ways the start and end times are ‘named’ in the tables is mind blowing (250 in one database we know of)
- Less than a handful of people know the entire EDW structure, sometimes even NO one.
  it takes months to train a person on how things fit together.
- Query 1 would take tens of pages of SQL and many seconds to minutes to execute.

FRANZ INC.

[2] Vocabularies and Taxonomies NOT built into EDW

Imagine you do this in your EDW

```
Find all patients that have some form of lung cancer and use a
medication that contains some form of opiate.

Which type anti wrinkle cream increased the most in price on the major
shopping portals in the last year.

Mobile telephony: what is the average time between a google search for
a particular restaurant and the time for a response (a call or a
open-table interaction, or a website visit booking?)
```

- Note that there are many types of lung cancer that even don’t have the string ‘lung’ or ‘cancer’, think Adenocarcinoma.
- Note that there at least three major taxonomies for ecommerce to classify products
- Taxonomies can be deeply nested hierarchies, an EDW is not built for deeply nested hierarchies and recursive queries over hierarchies.

FRANZ INC.

[3] Machine Learning results NOT integrated into EDW

- Imagine you do this in your EDW

```
Set grace period of one week to all the customers that are in the
category 'low payment risk' that are too late paying their bill but
haven't exceeded their average 'late time'

Given all the diagnostics for this patient, what do we expect to be the
next disease for this person?

Find highly similar "posey facial anti wrinkle creams" that use
completely different UPC codes (ASIN, GTIN).
```

Ideally one would

- Enrich a customer with their risk factor for payments
- Have a complete statistical co-occurrence, oddsratio or association graph for any combination of diagnostics
- Take the similarity between on-line products, based on more than 10 features, and create a full similarity graph in the database that contains all ecommerce

FRANZINC.

Challenge

Can you have a solution

that keeps your EDW as the ‘truth’

but fixes all these problems?

Yes,

our knowledge graph solution

that we developed with our customers.

Solution [1] from Data Sprawl to very concentrated data

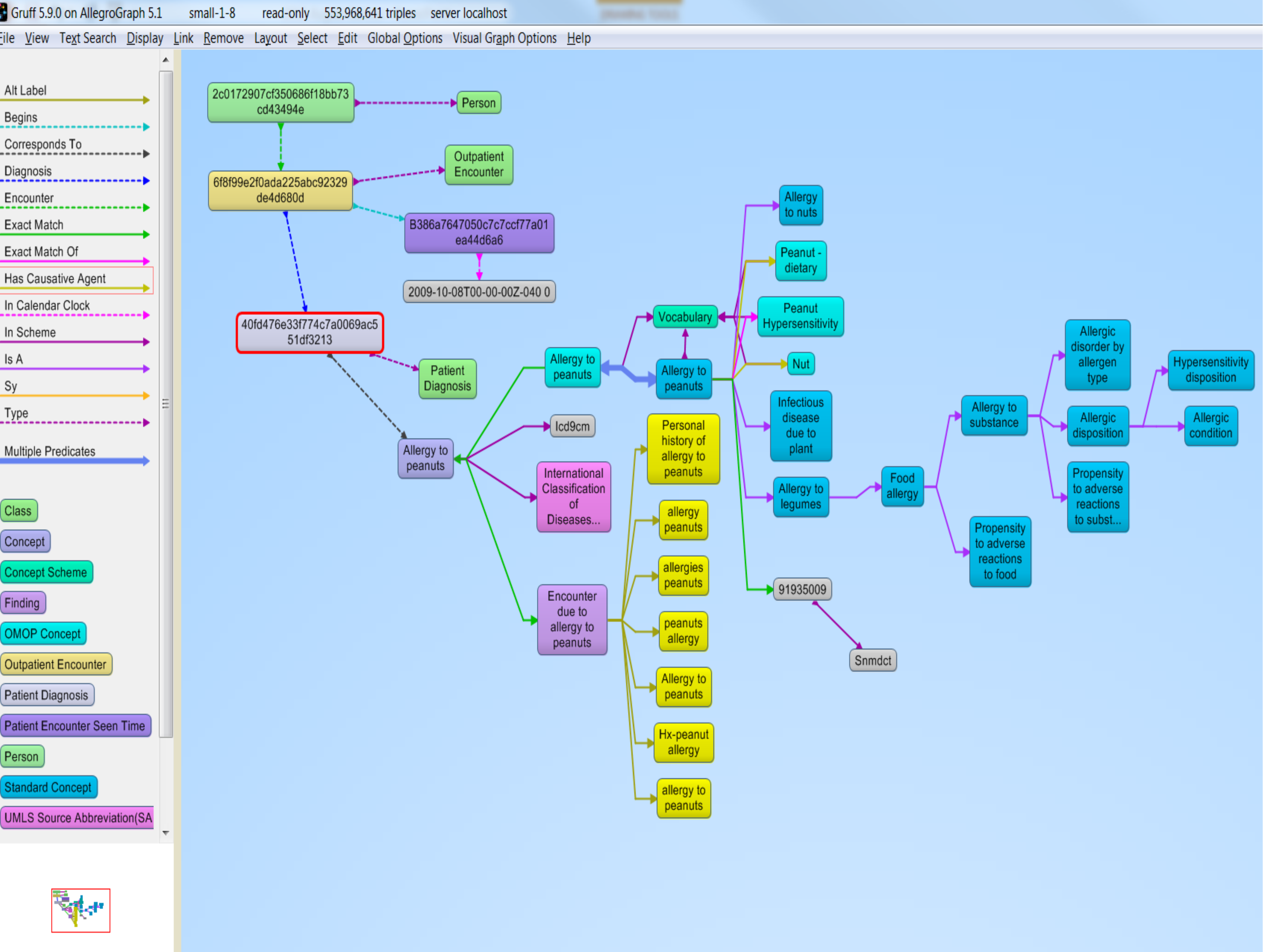
we turn everything into an EVENT

From thousands of tables to one event table (well: event graph)

- Healthcare: everything that can happen to a patient is a time based (sub) event: Check In, Check Out, Test, Diagnosis, Procedure, Medication administration, Medication order, Sensor reading for vital signs, Invoice, Bill payment, Non-bill payment, all insurance interactions. Yes: doctor's notes too.
- Telecom: everything that happens with a telco user is a time based (sub) event: telephone call, sms, whatsapp, web site visit, location record, crm call, bill pay, non-bill pay. Yes: CRM agent notes too.

Solution [2] Vocabularies & Taxonomies built into the Knowledge Graph

- Our terminology system is a combination of 180+ medical and life science vocabularies, taxonomies, ontologies and thesauri into one terminology system/knowledge base
- Ultimately linked to the structured events data in our Unified Clinical Event Ontology via think ICD9 and ICD10, CPT, LOINC, NDC.

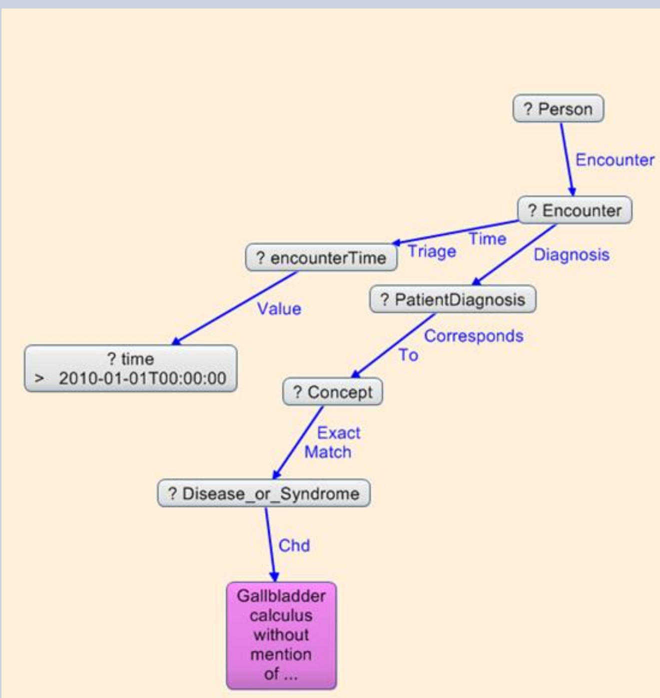

Gruff 5.9.0 on AllegroGraph 5.1
small-1-8
read-only
553,968,641 triples
server localhost
File
View
Text Search
Display
Link
Remove
Layout
Select
Edit
Global Options
Visual Graph Options
Help
Alt Label
Begins
Corresponds To
Diagnosis
Encounter
Exact Match
Exact Match Of
Has Causative Agent
In Calendar Clock
In Scheme
Is A
Sy
Type
Multiple Predicates
Class
Concept
Concept Scheme
Finding
OMOP Concept
Outpatient Encounter
Patient Diagnosis
Patient Encounter Seen Time
Person
Standard Concept
UMLS Source Abbreviation(SA
2c0172907cf350686f18bb73 cd43494e
Person
6f8f99e2f0ada225abc92329 de4d680d
Outpatient Encounter
B386a7647050c7c7ccf77a01 ea44d6a6
2009-10-08T00-00-00Z-040 0
40fd476e33f774c7a0069ac5 51df3213
Patient Diagnosis
Allergy to peanuts
Allergy to peanuts
Icd9cm
International Classification of Diseases...
Encounter due to allergy to peanuts
Vocabulary
Allergy to peanuts
Personal history of allergy to peanuts
allergy peanuts
allergies peanuts
peanuts allergy
Allergy to peanuts
Hx-peanut allergy
allergy to peanuts
Allergy to nuts
Peanut - dietary
Peanut Hypersensitivity
Nut
Infectious disease due to plant
Allergy to legumes
91935009
Snmdct
Food allergy
Allergy to substance
Propensity to adverse reactions to food
Allergic disorder by allergen type
Allergic disposition
Propensity to adverse reactions to subst...
Hypersensitivity disposition
Allergic condition

Queries are dramatically simplified

How many patients had 'gallbladder calculus' in 2010 and later

SPARQL

Select (count(distinct ?pt) as ?count) where {
?pt cdm:encounter ?enc .
?enc timew3c:begins/upper:value ?date ;
cdm:diagnosis/upper:correspondsTo/skos:exactMatch ?cui.
filter (?date > '2010-01-01T00:00:00'^^xsd:dateTime)
?cui mth:chd* mth:C0497327.}



SQL

(Part 1 of 3)

Select dx.dxCode, dx.dxCount, cl.dxClassName, cl.ClassCount
From (select distinct c.person_source_value as person,
c.condition_source_concept_id as dxCode, f.concept_name as
dxName,
a.concept_id_2 as dxClass, e.concept_name as dxClassName from
omopv5.concept_relationship a
join cdrn.condition_occurrence c on
a.concept_id_1=c.condition_source_concept_id join
omopv5.concept_ancestor d
on d.descendant_concept_id=a.concept_id_2 join omopv5.concept e
on e.concept_id=d.ancestor_concept_id
join omopv5.concept f on f.concept_id=d.descendant_concept_id
where a.relationship_id='Maps to' and
d.min_levels_of_separation=2 and e.standard_concept='S') dx JOIN
(select distinct c.person_source_value as person,
c.condition_source_concept_id as dxCode, f.concept_name as
dxName, a.concept_id_2 as dxClass, e.concept_name as
dxClassName from omopv5.concept_relationship a join
cdrn.condition_occurrence c on
a.concept_id_1=c.condition_source_concept_id join
omopv5.concept_ancestor d on
d.descendant_concept_id=a.concept_id_2
join omopv5.concept e on e.concept_id=d.ancestor_concept_id join
omopv5.concept f on f.concept_id=d.descendant_concept_id
where a.relationship_id='Maps to' and d.min_levels_of_separation=2
and e.standard_concept='S') cl on dx.dxClass=cl.dxClass
order by cl.dxClass desc

Solution [3] Output of Machine Learning directly Integrated into Database

- By adding results of machine learning back into the database we create learning systems.

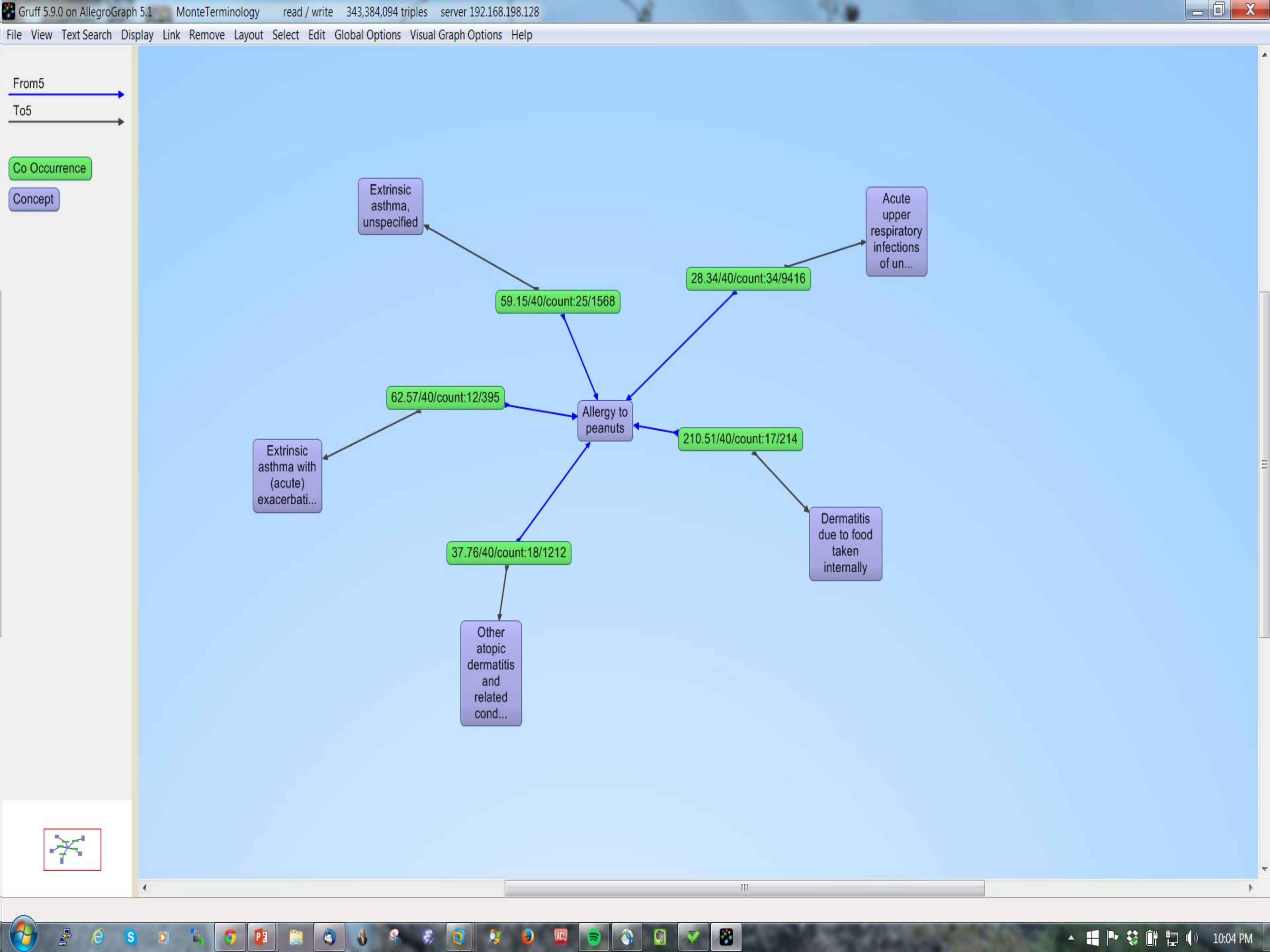
Odds ratio

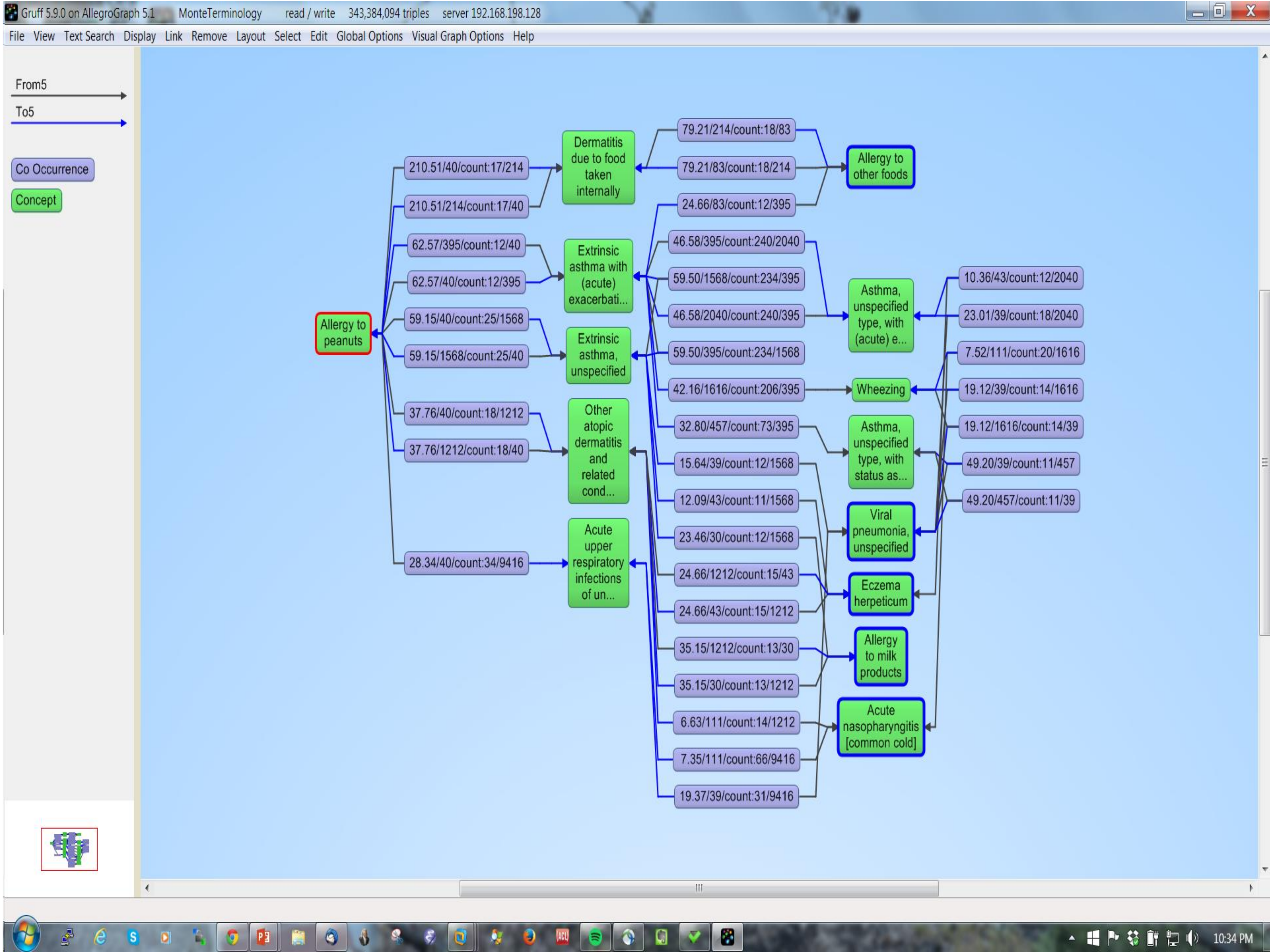
The Odds Ratio is a way to quantify how strongly the presence or absence of property A is associated with the presence or absence of property B in a given population

Patient Population **1,802,464**

		Ingestion Dermatitis		
		TO+	TO-	
Peanut Allergy	FROM+	544 (5)	736	**1,280**
	FROM-	6304	1,795,424	
		6,848		

Odds Ratio	210.51
95% CI Lower	187.91
95% CI Upper	235.82


Gruff 5.9.0 on AllegroGraph 5.1
MonteTerminology
read / write
343,384,094 triples
server 192.168.198.128
File View Text Search Display Link Remove Layout Select Edit Global Options Visual Graph Options Help
From5
To5
Co Occurrence
Concept
Extrinsic asthma, unspecified
Acute upper respiratory infections of un...
28.34/40/count:34/9416
59.15/40/count:25/1568
62.57/40/count:12/395
Allergy to peanuts
210.51/40/count:17/214
Extrinsic asthma with (acute) exacerbati...
Dermatitis due to food taken internally
37.76/40/count:18/1212
Other atopic dermatitis and related cond...
10:04 PM



Gruff 5.9.0 on AllegroGraph 5.1
MonteTerminology
read / write
343,384,094 triples
server 192.168.198.128
File View Text Search Display Link Remove Layout Select Edit Global Options Visual Graph Options Help
From5
To5
Co Occurrence
Concept
Allergy to peanuts
210.51/40/count:17/214
210.51/214/count:17/40
62.57/395/count:12/40
62.57/40/count:12/395
59.15/40/count:25/1568
59.15/1568/count:25/40
37.76/40/count:18/1212
37.76/1212/count:18/40
28.34/40/count:34/9416
Dermatitis due to food taken internally
Extrinsic asthma with (acute) exacerbati...
Extrinsic asthma, unspecified
Other atopic dermatitis and related cond...
Acute upper respiratory infections of un...
79.21/214/count:18/83
79.21/83/count:18/214
24.66/83/count:12/395
46.58/395/count:240/2040
59.50/1568/count:234/395
46.58/2040/count:240/395
59.50/395/count:234/1568
42.16/1616/count:206/395
32.80/457/count:73/395
15.64/39/count:12/1568
12.09/43/count:11/1568
23.46/30/count:12/1568
24.66/1212/count:15/43
24.66/43/count:15/1212
35.15/1212/count:13/30
35.15/30/count:13/1212
6.63/111/count:14/1212
7.35/111/count:66/9416
19.37/39/count:31/9416
Allergy to other foods
Asthma, unspecified type, with (acute) e...
Wheezing
Asthma, unspecified type, with status as...
Viral pneumonia, unspecified
Eczema herpeticum
Allergy to milk products
Acute nasopharyngitis [common cold]
10.36/43/count:12/2040
23.01/39/count:18/2040
7.52/111/count:20/1616
19.12/39/count:14/1616
19.12/1616/count:14/39
49.20/39/count:11/457
49.20/457/count:11/39
10:34 PM

Association rules

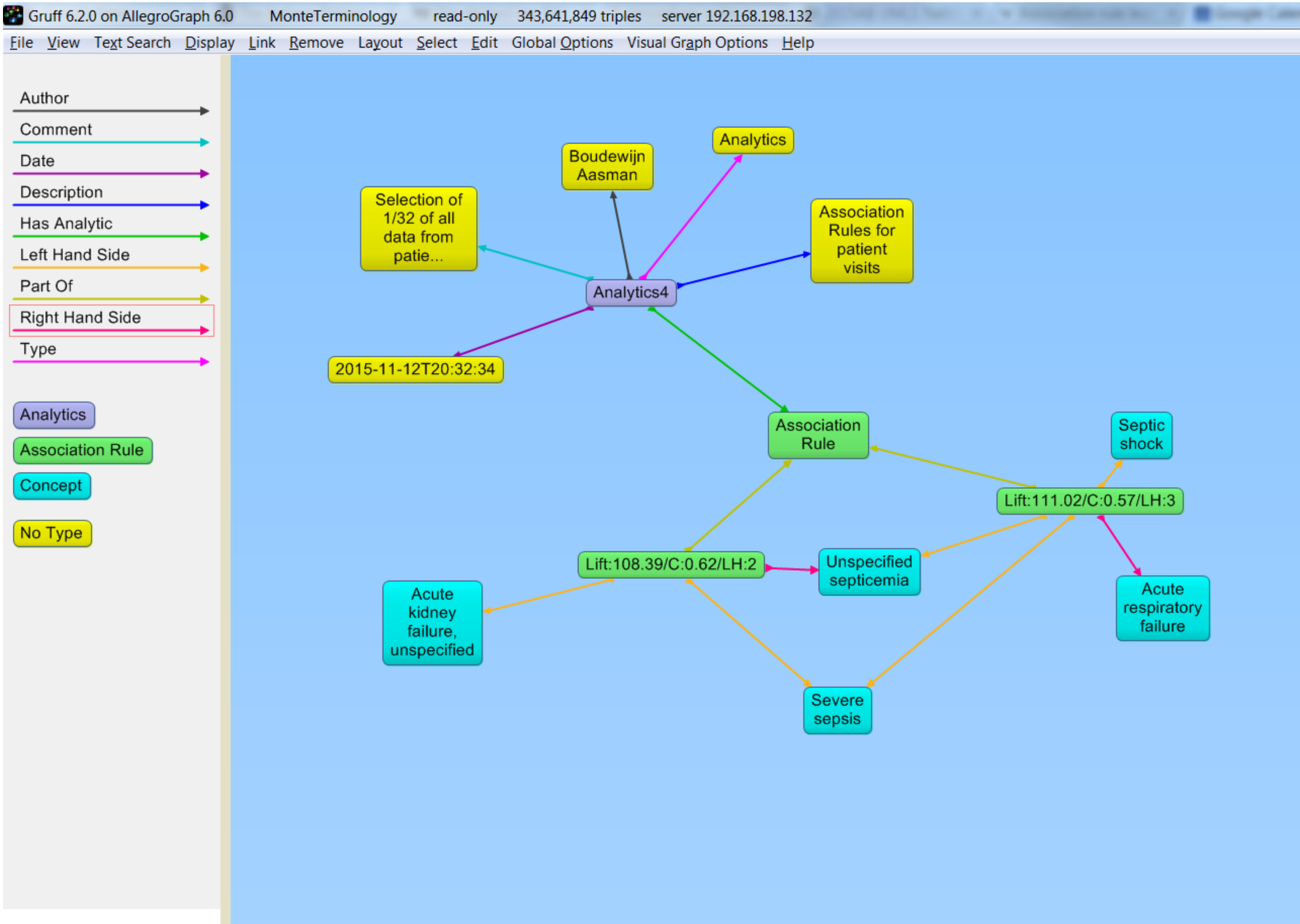
Association rule learning is a method for discovering interesting relations between variables in large databases.

$$Rule: X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$



Gruff 6.2.0 on AllegroGraph 6.0 MonteTerminology read-only 343,641,849 triples server 192.168.198.132
File View Text Search Display Link Remove Layout Select Edit Global Options Visual Graph Options Help
Author
Comment
Date
Description
Has Analytic
Left Hand Side
Part Of
Right Hand Side
Type
Analytics
Association Rule
Concept
No Type
Boudewijn Aasman
Analytics
Selection of 1/32 of all data from patie...
Association Rules for patient visits
Analytics4
2015-11-12T20:32:34
Association Rule
Septic shock
Lift:111.02/C:0.57/LH:3
Lift:108.39/C:0.62/LH:2
Unspecified septicemia
Acute kidney failure, unspecified
Acute respiratory failure
Severe sepsis

A summary of how we solved the EDW problem with an AI based Knowledge Graph

Data sprawled all over the place	**Unified entity and event model, can be organically grown, no more repetitive ETL**
Taxonomies and vocabularies NOT built in	**Fixed!**
Machine learning NOT integrated	**Fixed!**
Domain knowledge and rules NOT built in	Linked Open Data integrated
There is a graph in your EDW	Ready for any graph algorithm!

Or let us say this in a positive way

The future of the smart enterprise is Knowledge Graphs

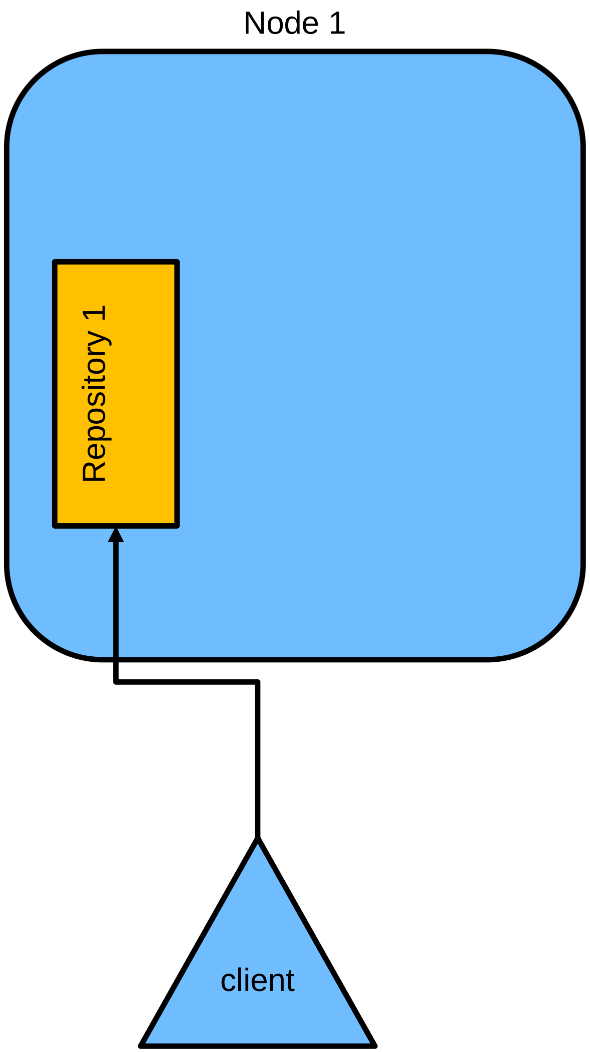
- Deeply simplified schema centered around entities
- Vocabularies and taxonomies built in
- Knowledge bases built in
- Artificial Intelligence built in from the ground up.
    - Symbolic: Inferencing, Prolog, Rules, ETC
    - Statistical: Machine Learning, Deep Learning, Predictive Analytics, etc.
- Graph analytics built in from the ground up.

And this is all BIG DATA

But what if you have a 100 Billion triples?

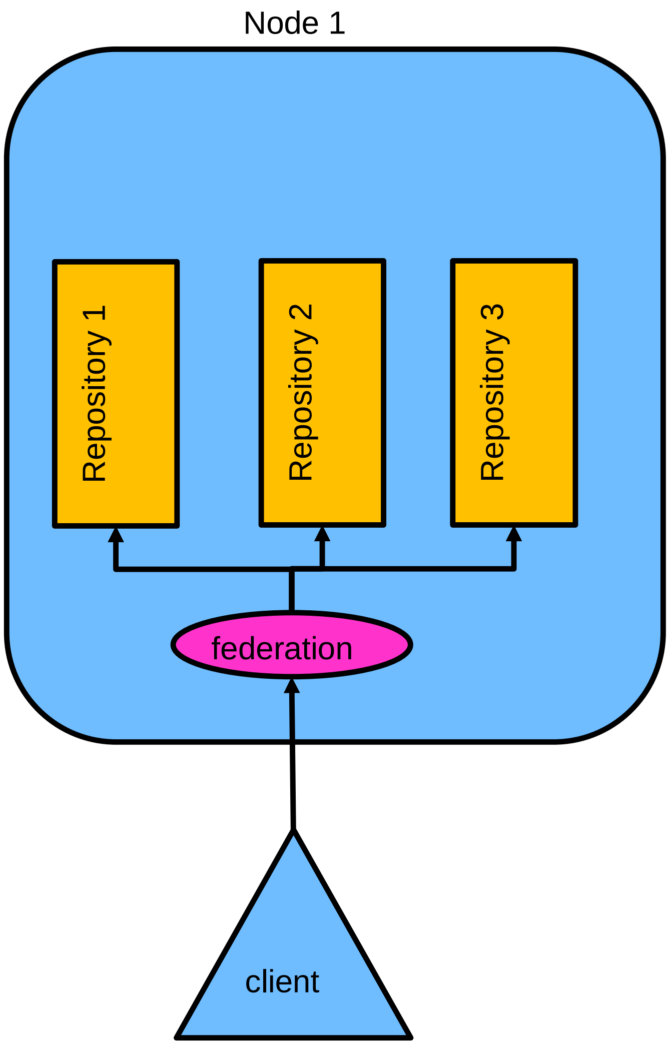
- Distributed AllegroGraph to the rescue

Normal operation



- Normal operation like any triple store
- A client communicates with a repository in a server in using the SPARQL Endpoint protocol.

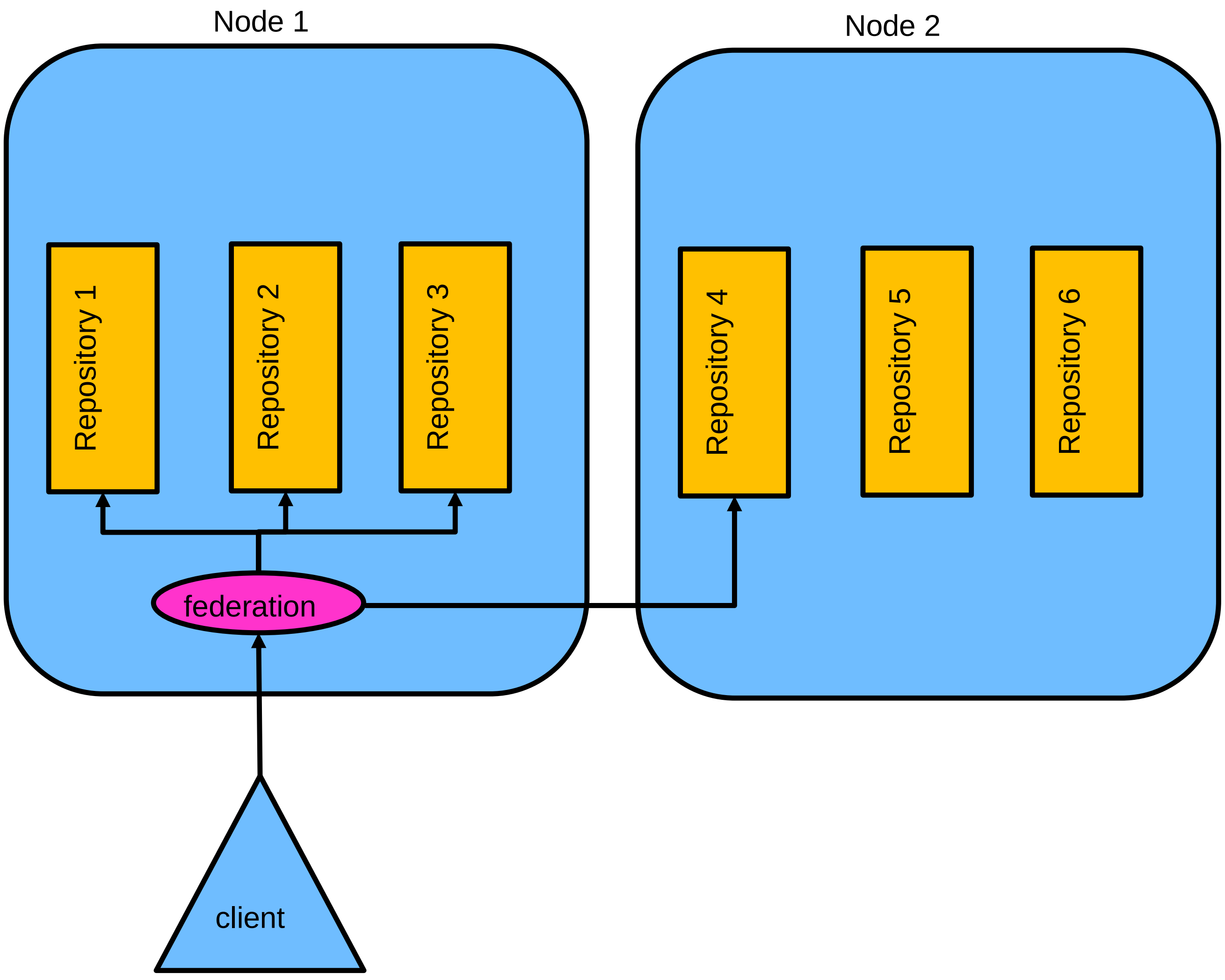
AllegroGraph Federation (same machine)



- A federation object is a virtual semantic graph database.
- Repositories on same machine (or node) are in the same memory space
- SPARQL runs in the federation object over a set of cursors from each repository.
- A client still communicates with the federation object using the SPARQL endpoint protocol.
- Performance is close to a single store database that combines all repositories.

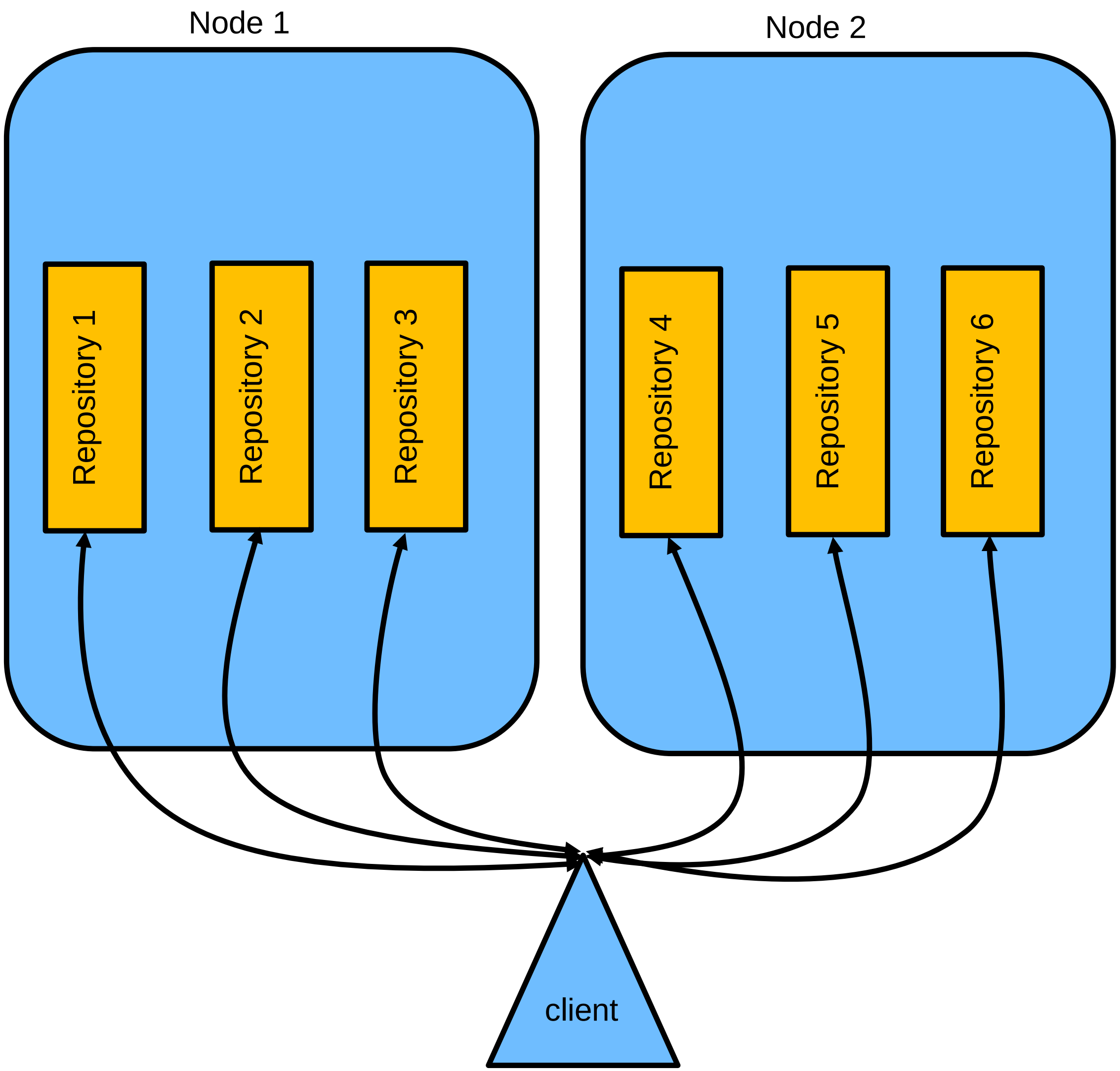
FRANZ INC.

AllegroGraph Federation (multiple nodes)



- We can federate to remote nodes
- But not as efficient as same machine federation

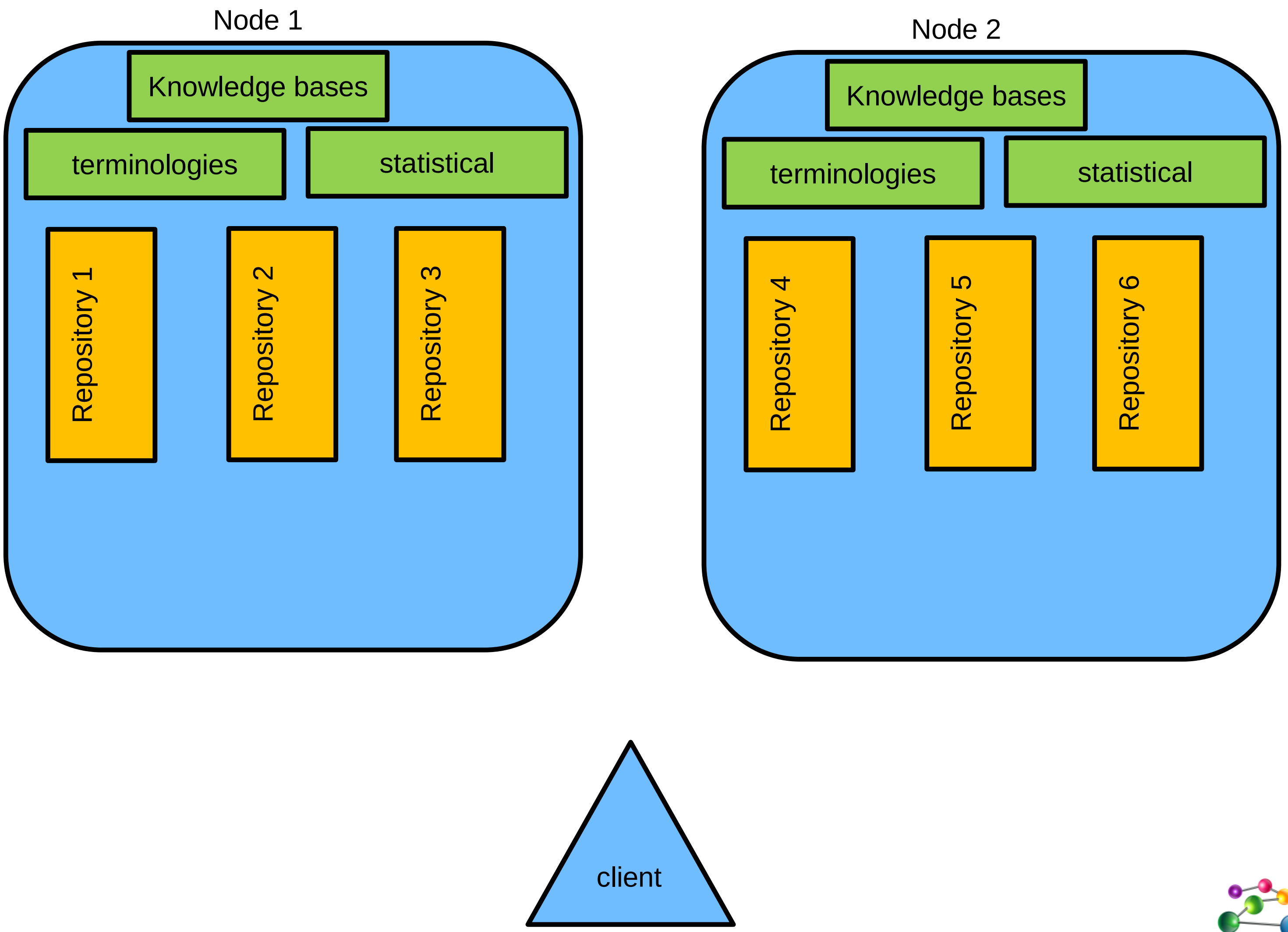
Parallel SPARQL (new in 7.0)



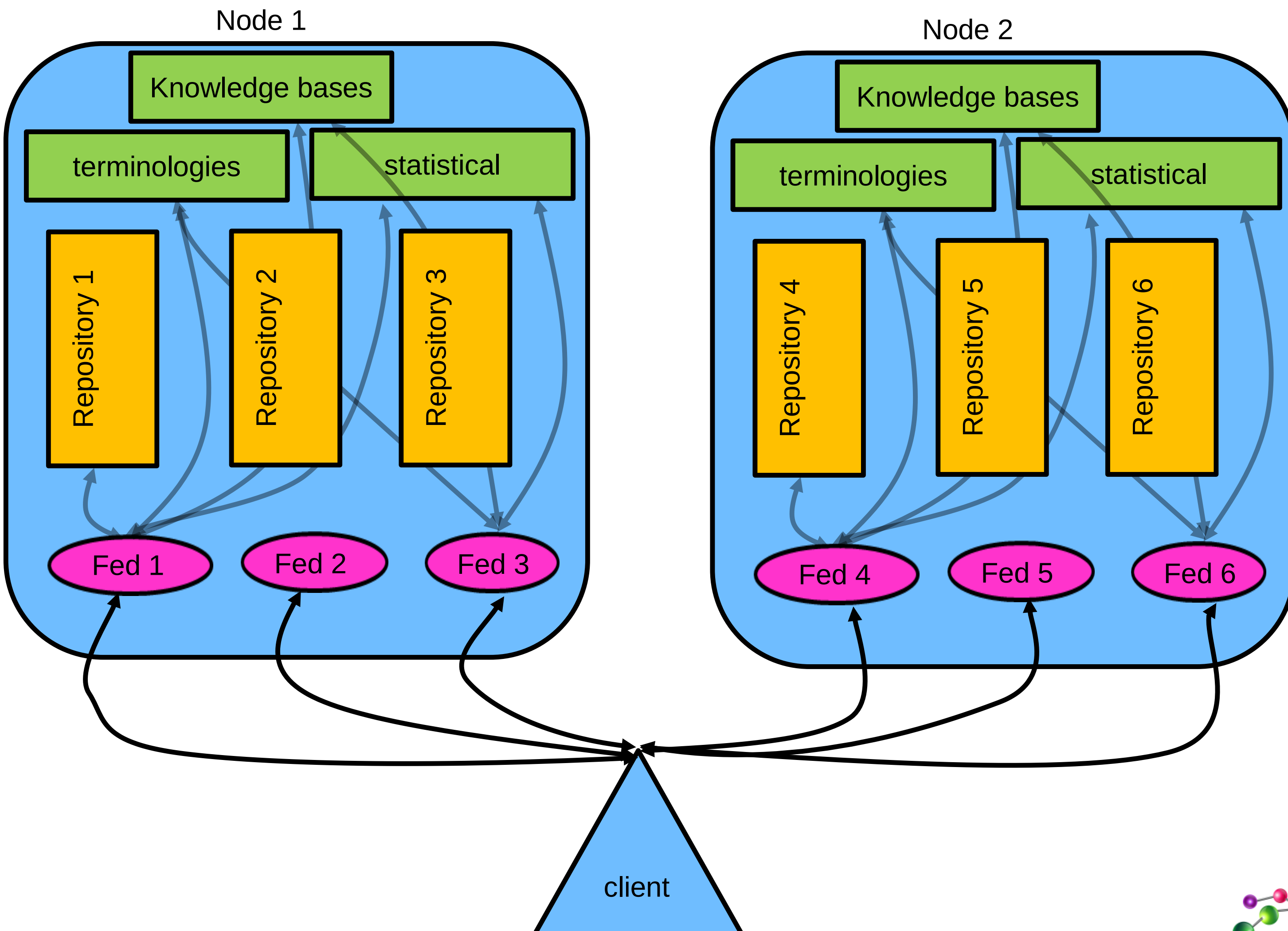
- For data that allows for logical partitions
- We created a PARALLEL SPARQL
- Now we can use all resources on all nodes and achieve maximum efficiency

- But what if you have un-partitionable knowledge bases and linked open data and you need that to be accessible in each partition?

We store a replica of all knowledge bases on each node



A hybrid parallel/federated SPARQL



The End

- Contact me for questions at ja@franz.com
- Website at http://franz.com