

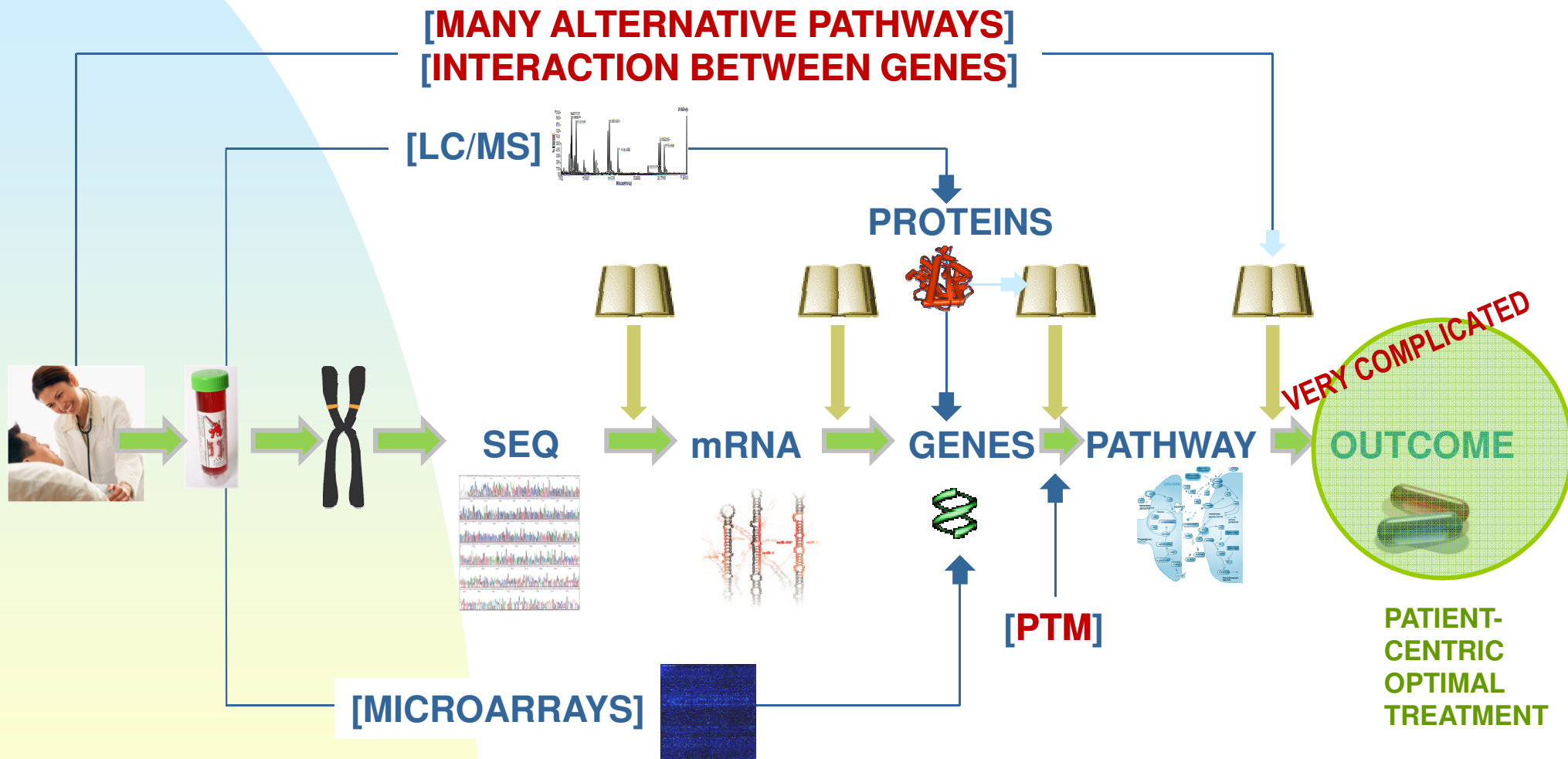
Analytic Knowledge Graph for Healthcare

by

Sheng-Chuan Wu, Ph.D.

Franz Inc.

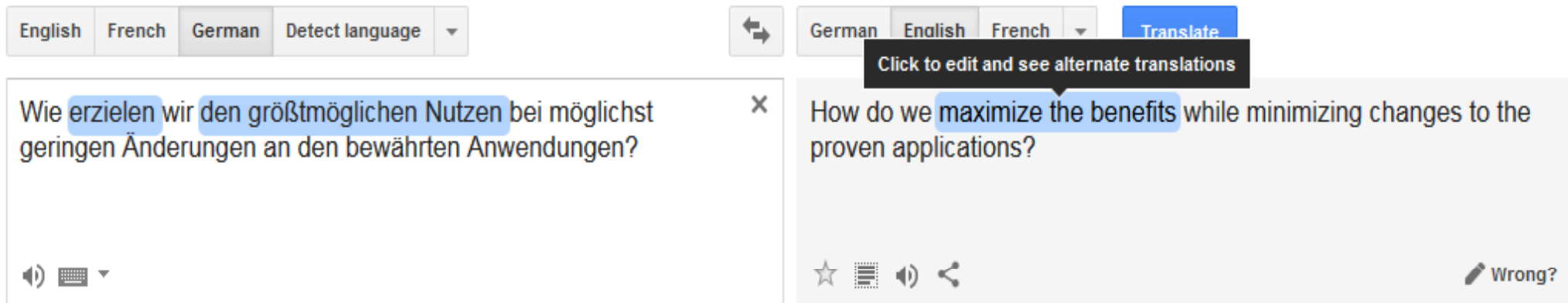
The Holy Grail of Targeted Medicine



The Elusive Grail for Targeted Medicine

- 1986 – observe breast tumor with over-expressed erbB2 gene has a much shorter remission period
- 1992 – confirm erbB2 over-expression links to inefficacy of standard chemo drug Tamoxifen.
- Testing of erbB2 gene prior to chemo becomes popular treatment practice
- 1998 – Genetech released Herceptin targeting erbB2 protein for chemo resistant tumor
- But, such success is very few and far in between
- Human biology is simply too complex to discern

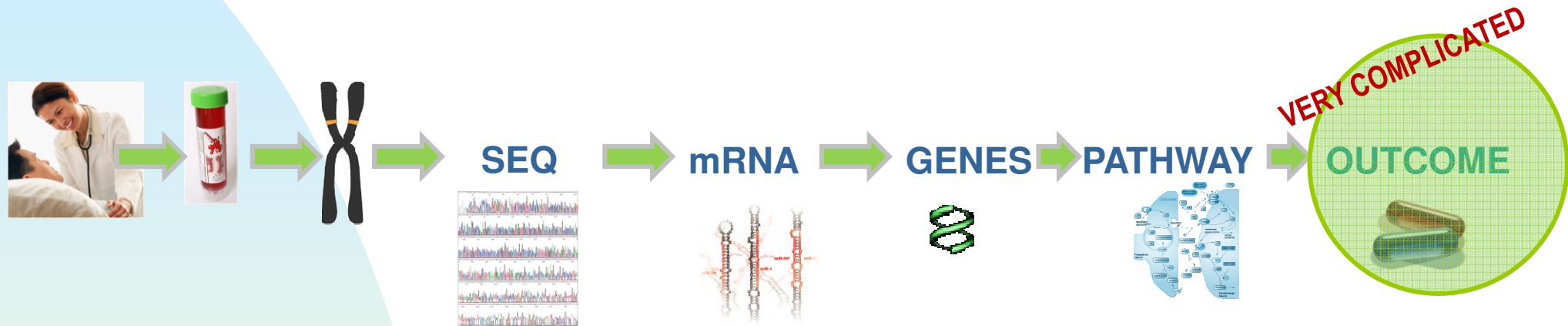
The Unreasonable Effectiveness of Data for Language Translation



“With enough data from the internet, you can reason statistically about what the next word in a sentence will be, right down to its conjugation, without necessarily knowing any grammatical rules or word meanings at all.”

-- Kevin Gold, interviewing Peter Norvig, Director of Research at Google

The Unreasonable Effectiveness of Data for Medical Science



“With enough data from the EMR, you can statistically reason about what diseases will likely happen without necessarily knowing any symptom in advance, or infer what the most effective treatment for a disease is without any knowledge of the biology behind it.”

-- paraphrase from Kevin Gold's interview with Peter Norvig

Example: Better Treatment for Ovarian Cancer Patients



- Ovarian cancer, one of the deadliest cancers with mean survival time of 3 years after diagnosis
- A study of over 1,400 women with ovarian cancer found patients with high blood pressure live, on average, 8 years after diagnosis¹, why?
- Common factor: patients with high BP who take non-selective beta blocker live much longer
- Why not give ovarian cancer patients non-selective beta blocker along with chemotherapy?
- Do we know the biology behind? Not yet!

¹ Cancer. 2015 Oct 1;121(19):3444-51.

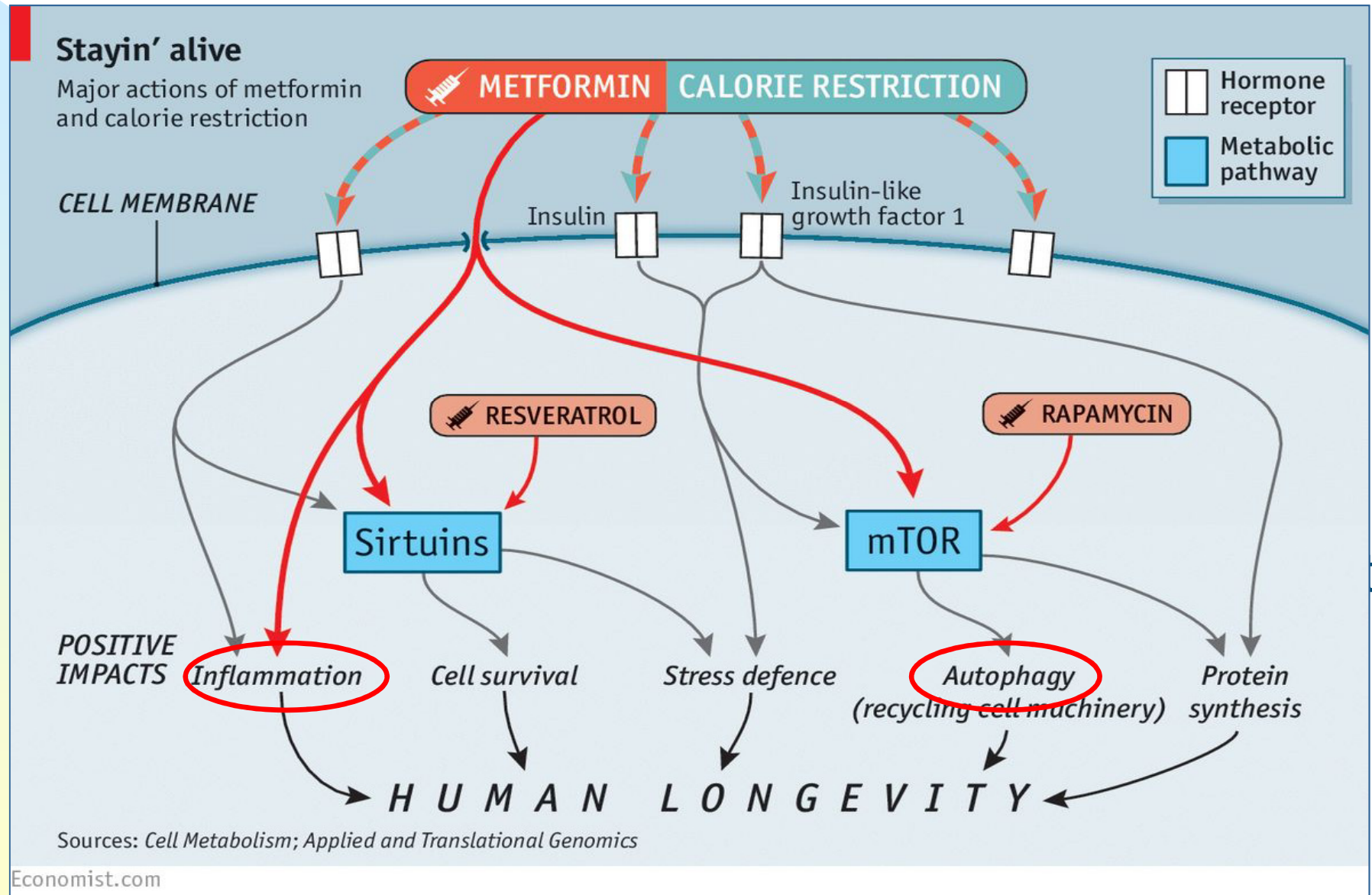
Diabetes Patients Live Longer

- A 2014 aging study of 90,000 elderly patients found those with type-2 diabetes survive much better than those non-diabetic patients
- Why?
- Type-2 diabetes patients typically take Metformin, which is shown to slow cell aging and decrease cancer risk in later clinical study
- Patients taking Metformin also have 62% reduced chance of getting pancreatic cancer
- Such is the unreasonable power of data

Turn Scientific Method Upside Down

- Traditional scientific method
 - develop hypothesis → design & conduct experiment
 - collect data → validate or disprove hypothesis
- New scientific method
 - collect (big) data → discover patterns → develop hypothesis → design & conduct clinical experiment
- From existing data we discern new insight
- Paradigm shift due to the power of big data

Turn Scientific Method Upside Down



The Miracle Drug from Data Analysis – Aspirin



- Daily use of baby aspirin results in
 - Substantial reduction in cardiovascular disease risk
 - 40 to 50% reduction in colorectal cancer¹
 - 75% reduction in esophageal cancer¹
- Do we know exactly why? Blood thinner? Anti-inflammation? We really not quite know!
- In 2009, US recommended that men over 45 and women over 55 take baby aspirin daily
- The unreasonable effectiveness of data

¹ The Lancet, Volume 376, No. 9754, p1741–1750, 20 November 2010

Where are All the Medical Data

- Existing patient Electronic Medical Records (EMR) are “live” clinical trial data bank
- Terabytes of EMR could be used to explore
 - What treatment most effective for what patients
 - Comparative effectiveness of different treatments
- Montefiore (research arm of Einstein Medical College) teams up with Franz to use EMR to support basic medical science, translational research, ACO, etc.

Issues with EMR Data

- While modern EMR are coded, they
 - Reside in many different silos
 - Employ different taxonomies mostly

*“(CapGemini) finds the main obstacle to Big Data success is **scattered silos of data** – 79% of organizations have not fully integrated their data sources across the organization.”*

- Are difficult to query, let alone abstracting patterns and deriving value from them
- First, need to integrate data from all the silos

Data Integration of Mammoth Proportion

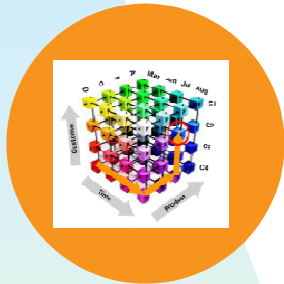


11/5/2010 3:58:07 PM, Category: General, Priority: 1, Severity: verbose, "MirrorJob.GetSettings Starting...", Machine: MOSS101
11/5/2010 3:58:07 PM, Category: General, Priority: 1, Severity: verbose, "MirrorJob.GetSettings Complete", Machine: MOSS101

No information integration, no true understanding and discovery, no true value

INTEGRATION IN DATA TAXONOMY, DATA SCHEMA AND DATA REPRESENTATION

Current Data Integration Practice - Data Warehouse



Multidimensional



Relational, Structured



Governance
Security



Used by 80% Enterprise
Analytics BI Users



Limited Data Types
No Data Semantic
(mostly character strings)
Rigid, Slow, Expensive

New Data Integration Practice - Data Lake

Single Ingest Native Values

Source of ALL data

No Upfront Data Prep

Enables Multiple views on Input
(Schema-on-read)

Staging Area

EVOLVING

Governance

Data Inventory Management,
Curation

Meta Data Management, MDM

Analytics

Access Subsets of Data

As required for specific
analyses

Need **ETL** to fit target schema
every time – Data Mart

Analytic Specific - **No Reuse**

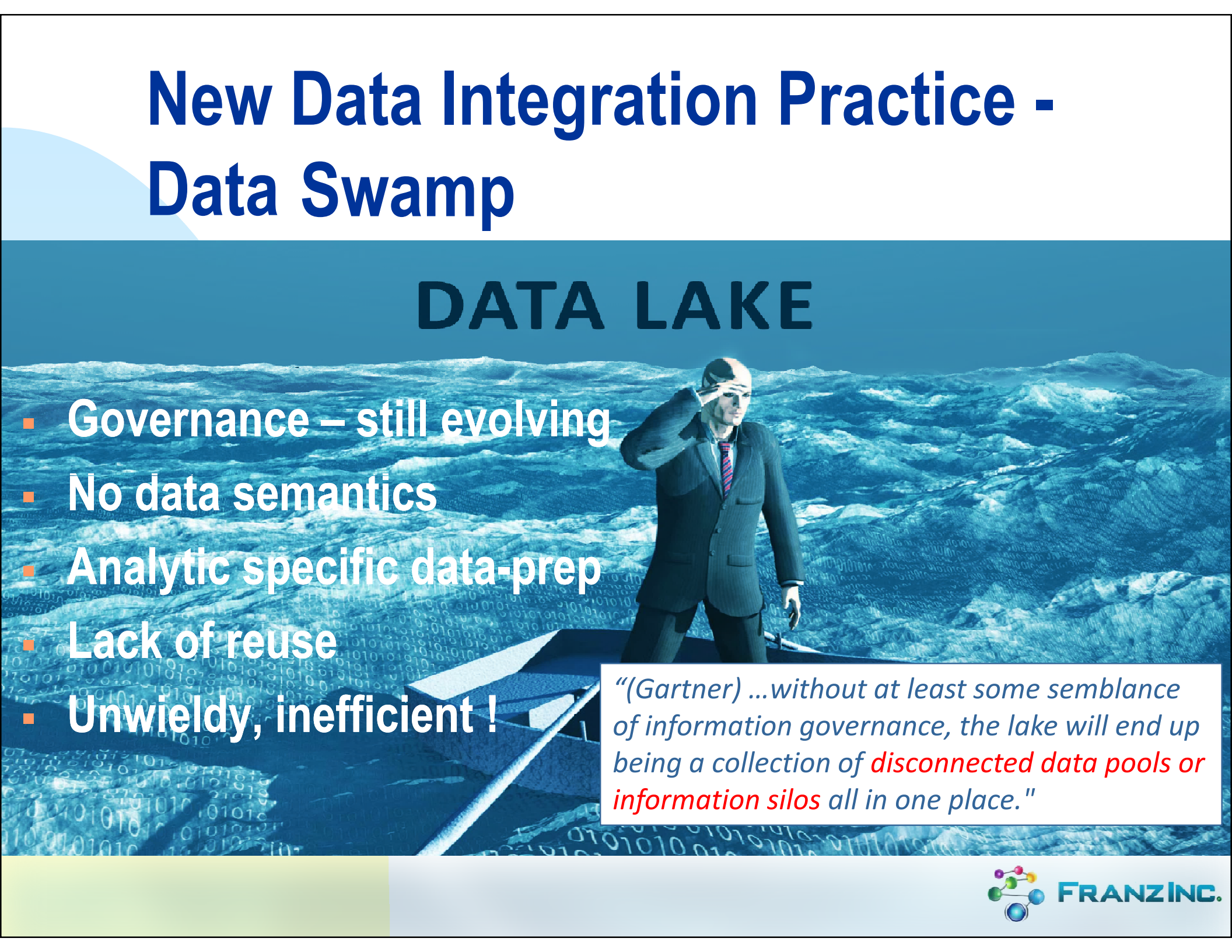


Source: Gartner
Copyright LeadSemantics® 2016

New Data Integration Practice - Data Swamp

DATA LAKE

- Governance – still evolving
- No data semantics
- Analytic specific data-prep
- Lack of reuse
- Unwieldy, inefficient !



*"(Gartner) ...without at least some semblance of information governance, the lake will end up being a collection of **disconnected data pools or information silos** all in one place."*

Medical Analytic Practice Today

- Researchers query physician(s) for patients with certain disease and diagnosis, or scour EMR database for such patients
- Select clinical parameters for the study, based on prior works or own clinical experience
- Manually go thru EMR for each patient
 - Enter a patient number into various EMR DB silos
 - Look through records and documents for relevant parameters
 - Enter those values into an Excel sheet for all patients
- Perform analysis on the spreadsheet table
- An extremely laborious and time-consuming process

There ought to be a better way for medical analytics!

Facts about Information Integration

- Information represents relations between entities; it is a graph, not joined tables
- Information (data) integration is not possible without a shared reference
- All information must speak the same language for computer to integrate them automatically
- That's what semantic data model is all about!

Same Word for Very Different Things

– “I own a Nokia”



Different Words for the Same Thing



中國銀行

BANK OF CHINA

中银

中国银行

Bank of China

Computer cannot help without clear data semantics

惠普

HP

H-P

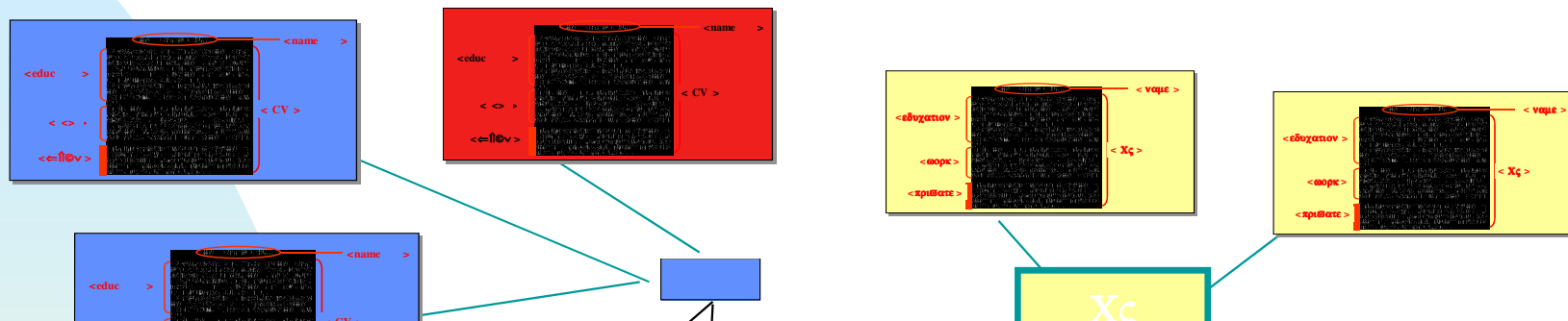
Hewlett-Packard

Hewlet-Packard



i n v e n t

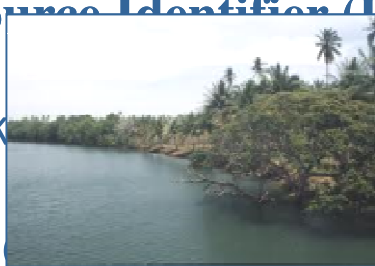
URI Remove Ambiguity



■ Uniform Resource Identifier (URI) as common

<http://meta.wik>

<http://www.fed>



Bank?



■ Ontology

Provide common references, structure and property for individual terms

■ Semantic languages (RDF)

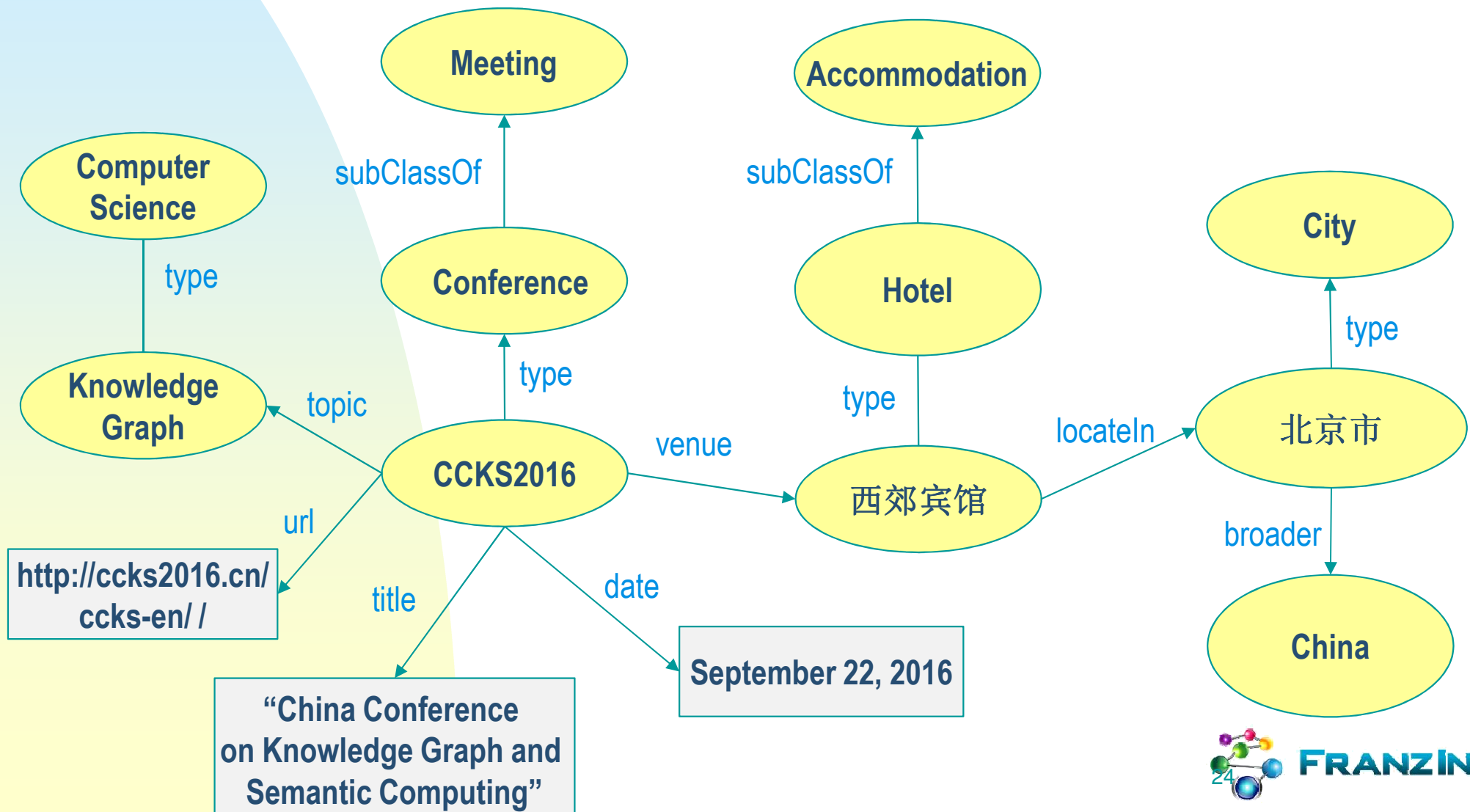
To describe mappings, structures and properties of data

RDF – A Uniform Way for Information Representation

- Knowledge representation standard for the semantic web by W3C, using XML syntax
- Meta data for content (web pages, multimedia contents, versioning) allows machines to help people search information and organize their lives.
- Quickly became standard for non-web meta data
- **It's a way to serialize semantic networks in XML**
<subject predicate object>
China located_in Asia
- **RDF triples – classic knowledge representation technique from AI research**

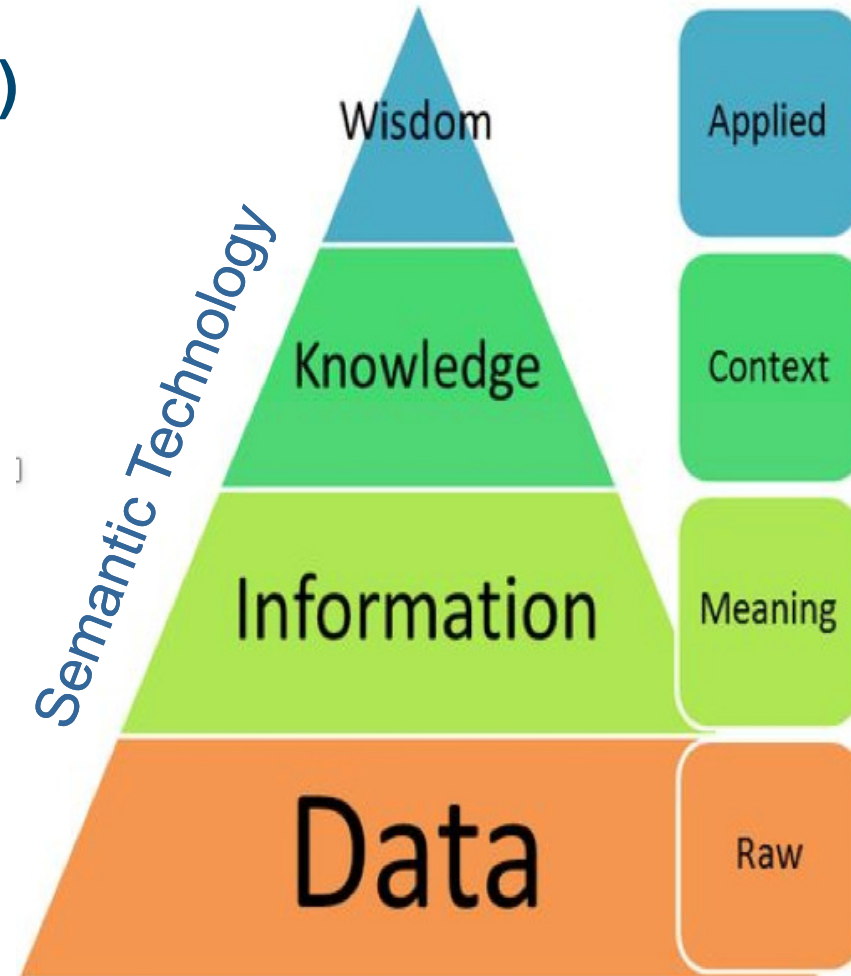
URI/RDF Representing “Data” in Semantic Graphs

Graph = nodes linked by labeled edges



Semantic Data Model

- **Consistent Encoding (URI/RDF/XML)**
- **W3C Standards**
 - RDF – ENCODING USING URIs
 - RDFS/ OWL – SEMANTIC MODEL
 - SPARQL – QUERY
- **Successful Implementations**
 - Knowledgebase Systems, Intelligence, Defense, Life Sciences & Healthcare, Financial Services, Research, Digital Learning..



Data Lake with Semantic Model

Single Ingest
One-time ETL

Staging Area
Access to ALL

Analytics
Ad Hoc Querying

Data from ALL Sources

ALL Analytics Use Cases

SPARQL

RDF/XML - Mimics single source, without loss of *context* and meaning, structure, format, type, lineage for traceability to source

RDF Graph

Schema-on-read

Access to ALL

Semantic Model – Clarifies Context and Relationships with RDF Graph, Maintains Semantic Consistency, Enables Curation

Governance, Security, Provenance via Layered Ontologies

Enhanced Data Discovery & Exploration

Extract any ad hoc features or dimension for analytics directly – no more ETL each time

Self Service Analytics – no more need for Data Mart

Semantic Data Lake

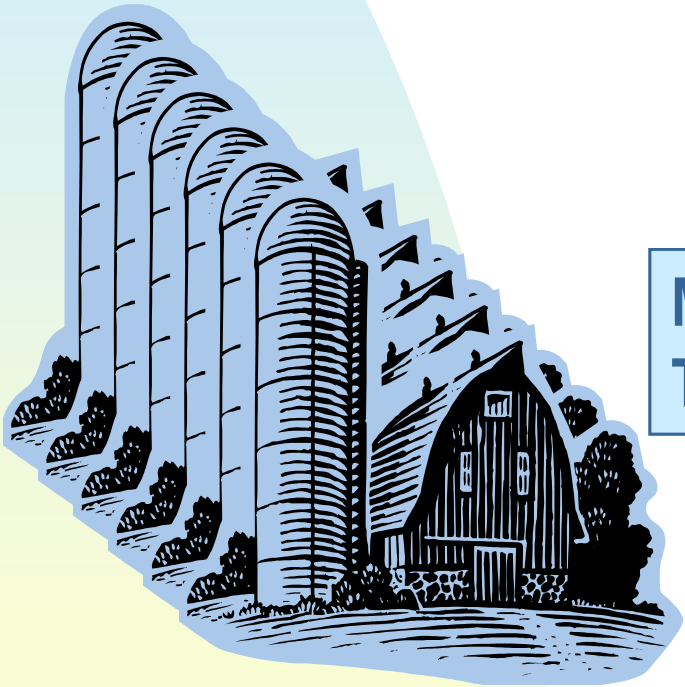
SMART BIG DATA LAKE

Gartner Group

*“Meeting the needs of wider audiences require curated repositories with governance, **semantic consistency** and access controls.”*

Integrate EMR Data Silos with URI and RDF

EMR Silos

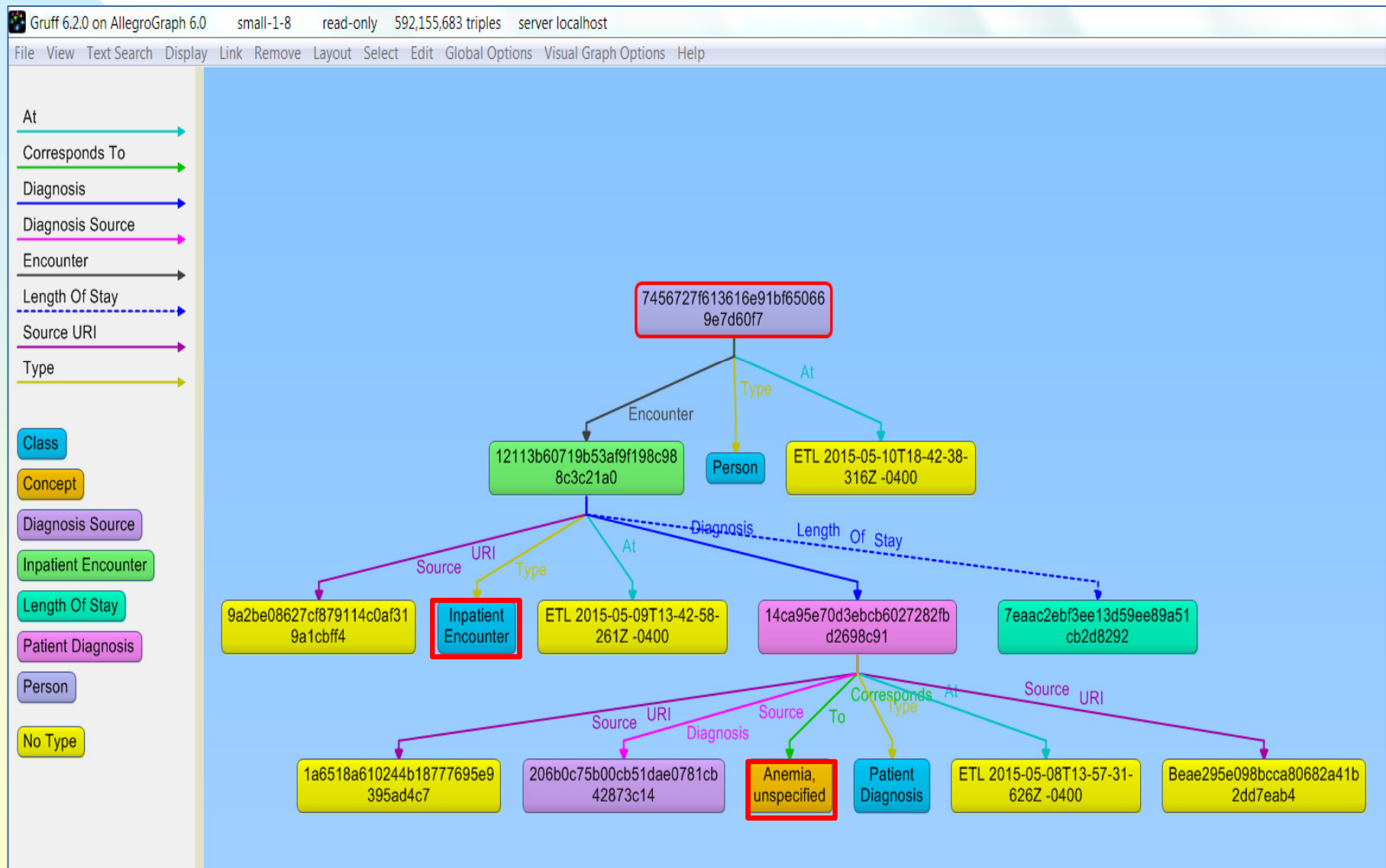


Map -> URI
Transform -> RDF

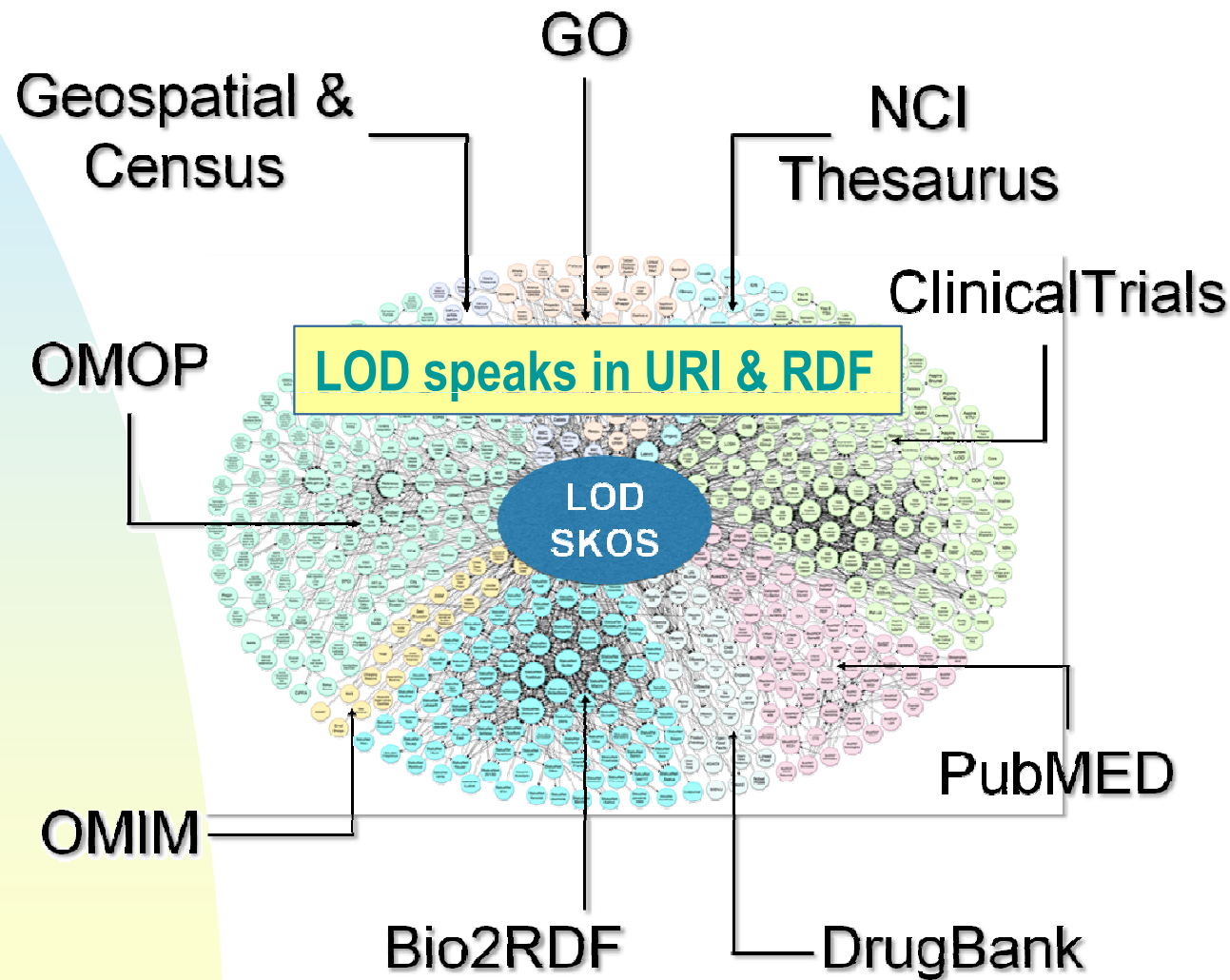
EMR
Semantic
database



An InPatient Encounter with Anemia in Semantic Database



Most Medical Knowledgebase in URI and RDF Too



Abdominal Pain (C0000737)

The screenshot displays a Semantic Web editor interface with two main panes. The left pane shows the ontology structure, and the right pane shows the details for the concept 'Abdominal Pain (C0000737)'.

Left Pane: Ontology Structure

- rdfs:Resource
 - dcam:VocabularyEncodingScheme
 - owl:DataRange
 - owl:Thing
 - owl:Nothing
 - skos:Collection
 - skos:Concept (246)
 - SN:Entity
 - SN:Conceptual_Entity
 - SN:Finding (7)
 - SN:Laboratory_or_Test_Result
 - SN:Sign_or_Symptom (4) **(Selected)**
 - SN:Group
 - SN:Group_Attribute
 - SN:Idea_or_Concept
 - SN:Intellectual_Product (2)
 - SN:Language
 - SN:Occupation_or_Discipline
 - SN:Organism_Attribute
 - SN:Organization
 - SN:Physical_Object
 - SN:Event
 - skos:ConceptScheme (1)
 - skos-xl:Label

Right Pane: Concept Details (Name: umls:C0000737)

Annotations

- skos:altLabel
 - ABDMNAL PAIN UNSPCF SITE
 - ABDOMINAL PAIN
 - ABDOMINAL PAIN NOS
 - AP - Abdominal pain
 - Abdomen pain
 - Abdominal Pains
 - Abdominal pain
 - Abdominal pain (finding)
 - Abdominal pain, unspecified site
 - Bellyache
 - PAIN ABDOMINAL
 - Pain, Abdominal
 - Pains, Abdominal
 - [D]Abdominal pain
 - [D]Abdominal pain (context-dependent category)
 - [D]Abdominal pain (situation)
 - [D]Abdominal pain NOS
 - [D]Abdominal pain NOS (context-dependent category)
 - [D]Abdominal pain NOS (situation)
 - abdominal pain
- skos:definition
 - Sensation of discomfort, distress, or agony in the abdominal region; generally associated with functional disorders, tissue injuries, or diseases.
- skos:prefLabel
 - Abdominal Pain

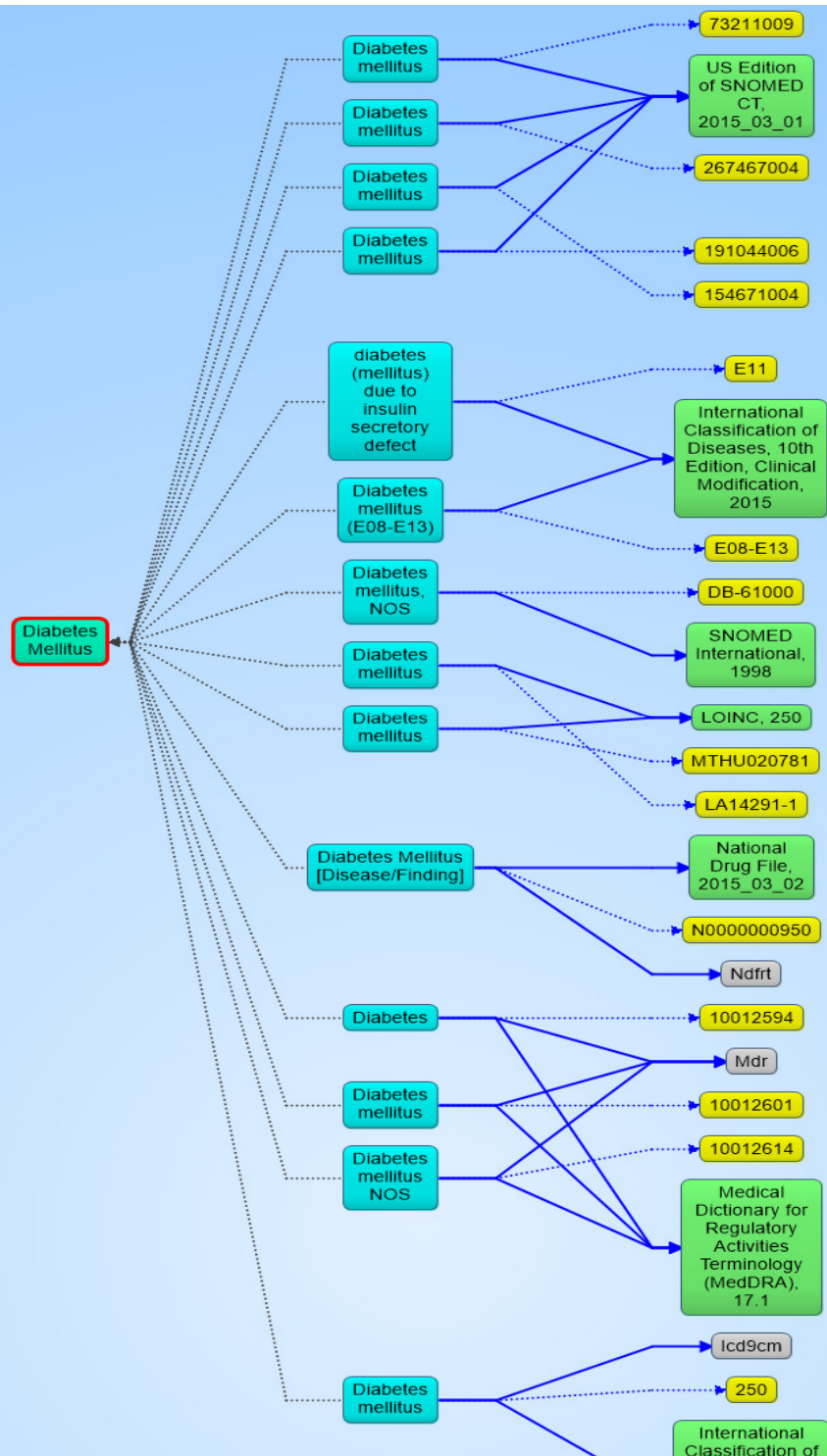
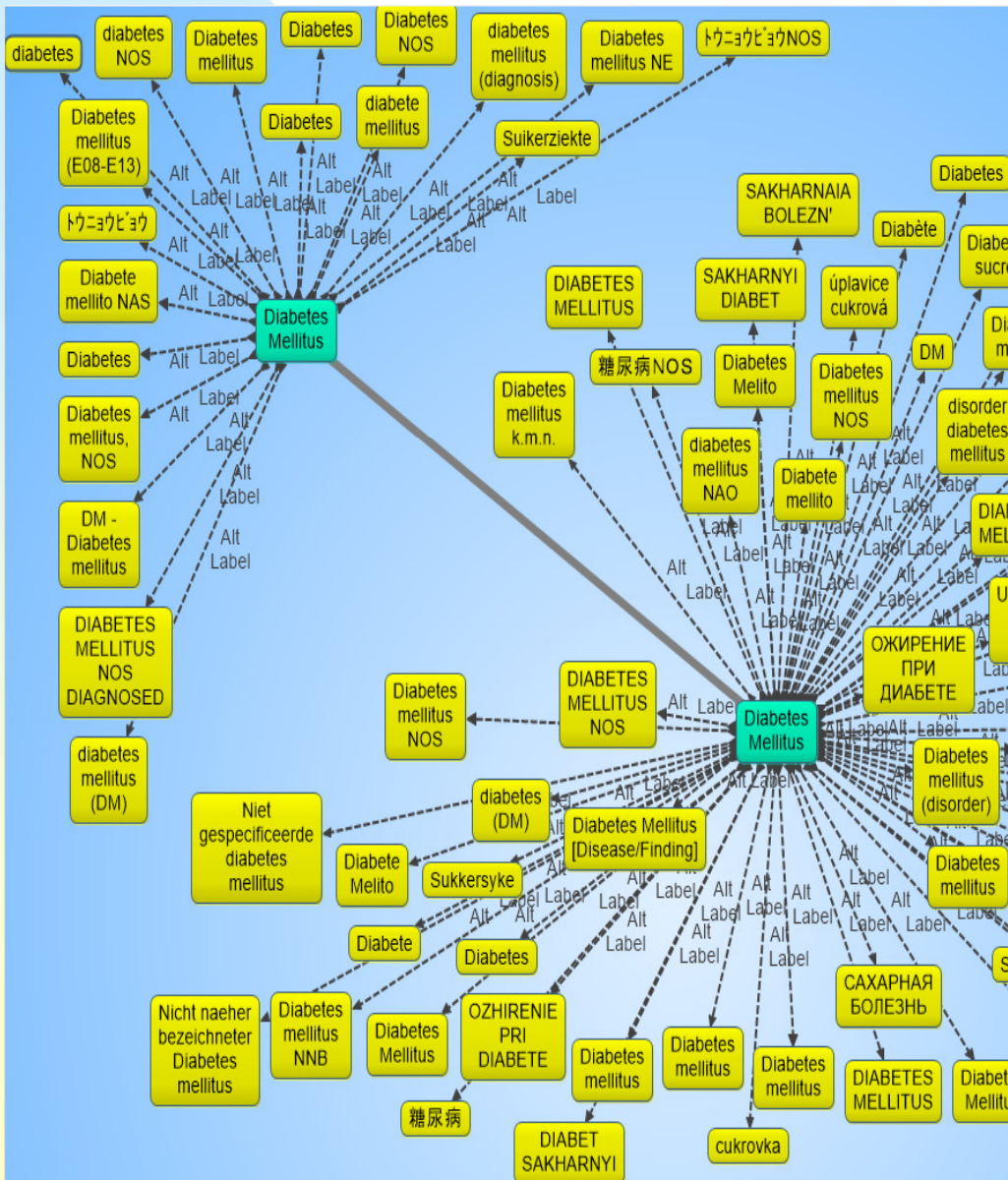
Other Properties

- umls:correspondsTo
 - SNOMEDCT_139313005
 - SNOMEDCT_158498003
 - SNOMEDCT_158512003

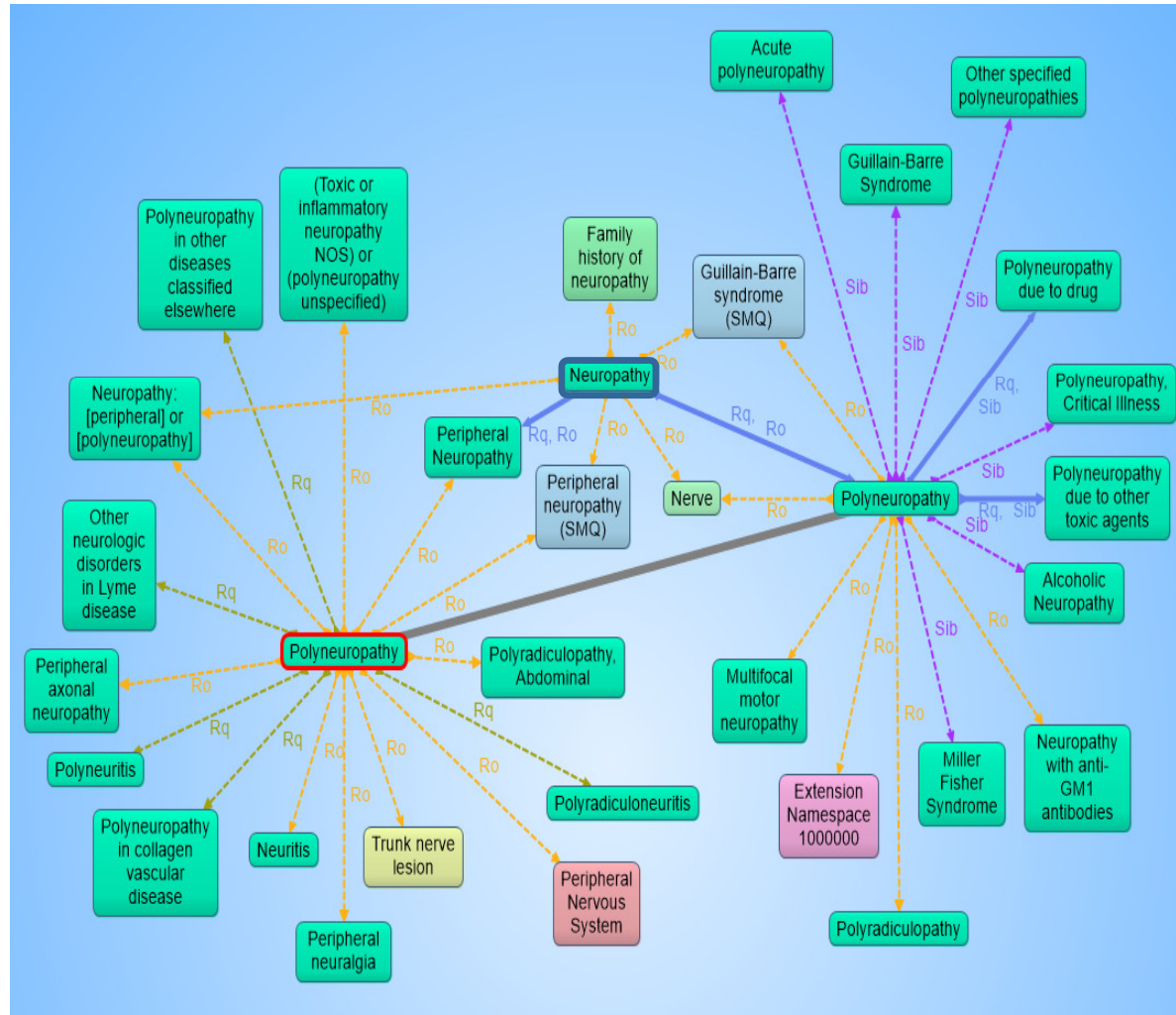
Instances

[Resource]	rdf:type	rdfs:label	rdfs:comment
umls:C00007 SN:Sign_or_Symp			
umls:C00007 SN:Sign_or_Symp			
umls:C00007 SN:Sign_or_Symp			
umls:C00008 SN:Sign_or_Symp			

Diabetes Mellitus



Taxonomic and Semantic Relations

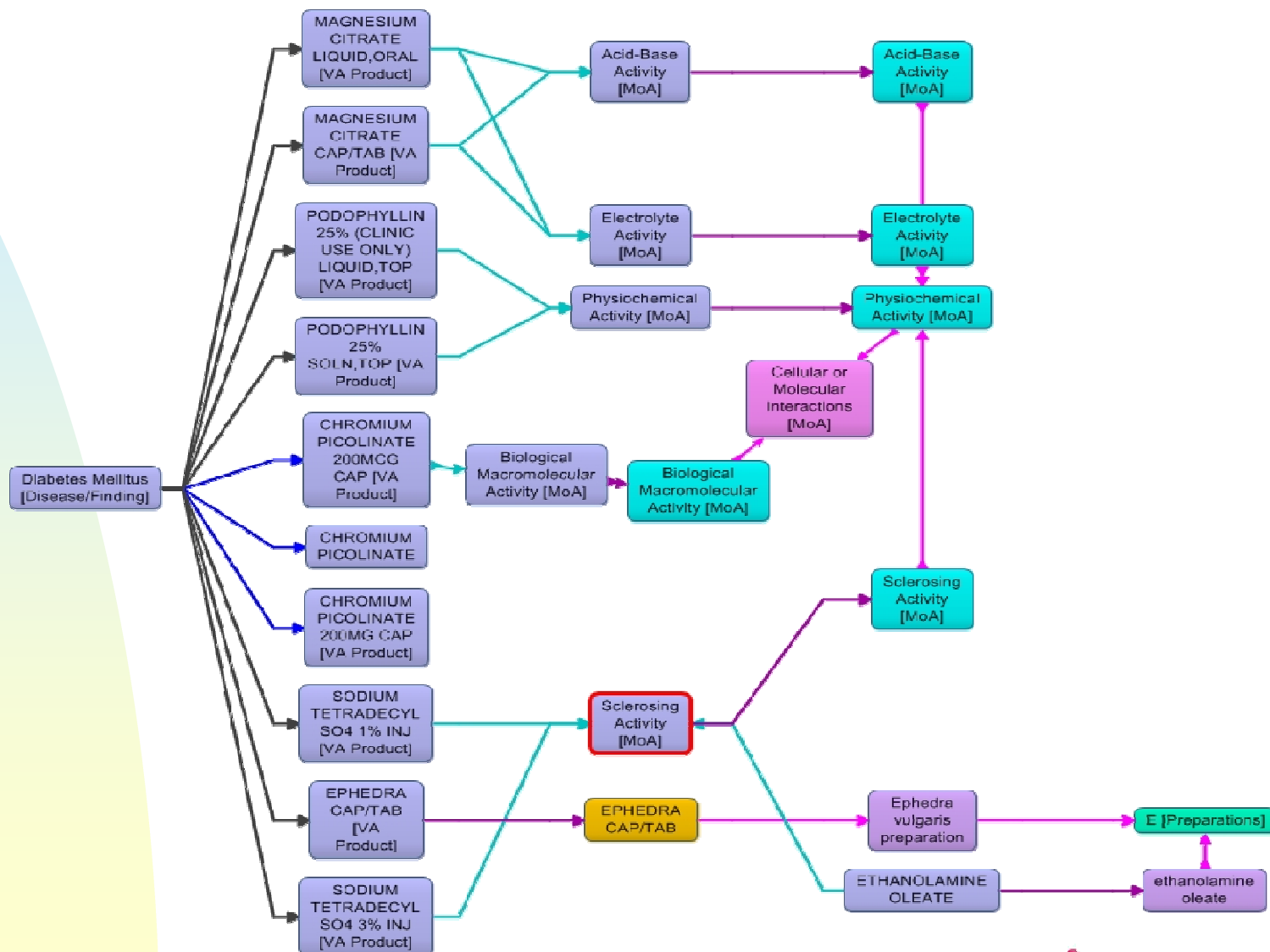


Demo – Path Finding

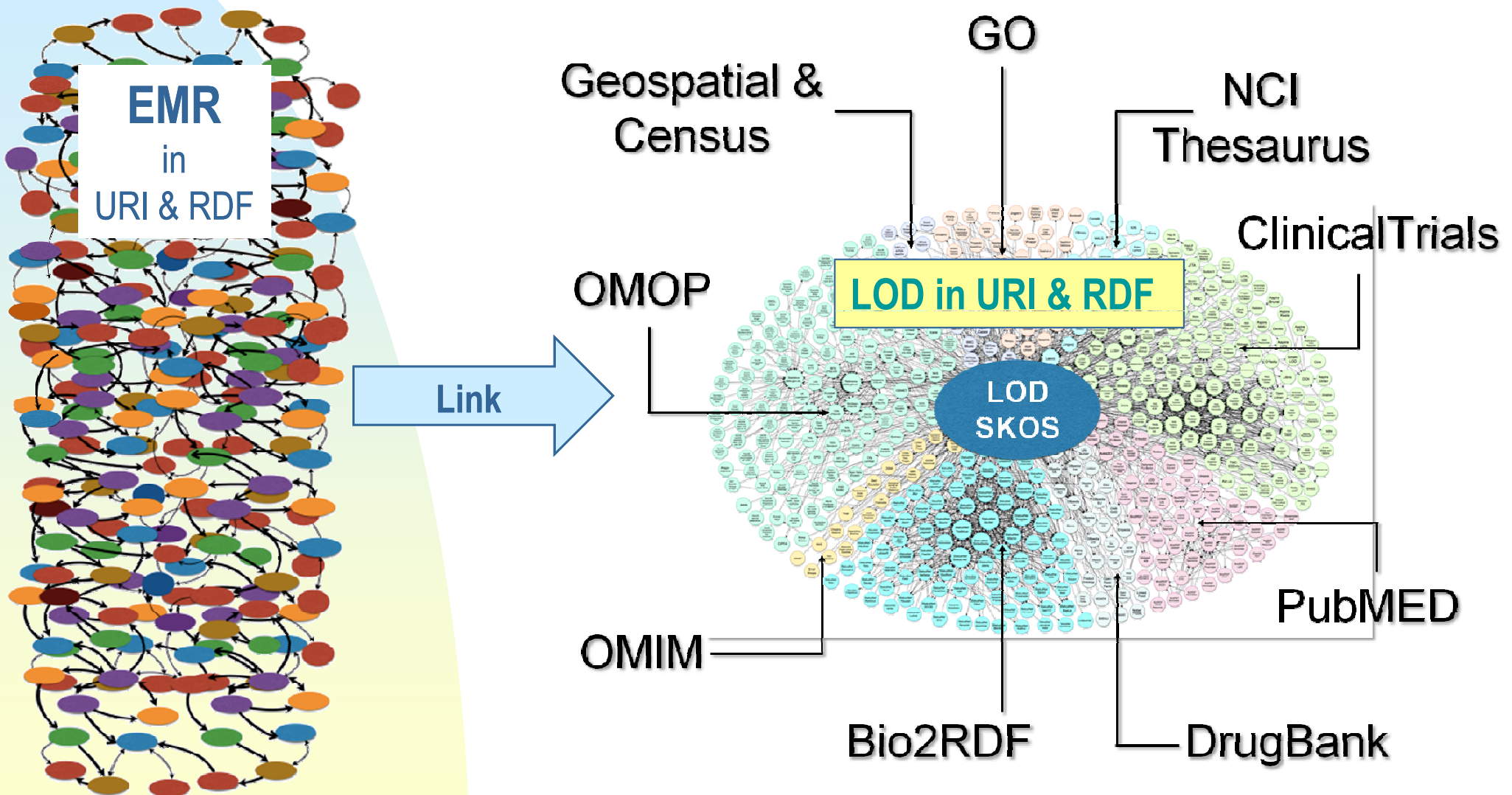
Example – Path Finding

- Chd →
- Exact Match →
- Has contraindicated dru →
- Has mechanism of action →
- May be treated by →

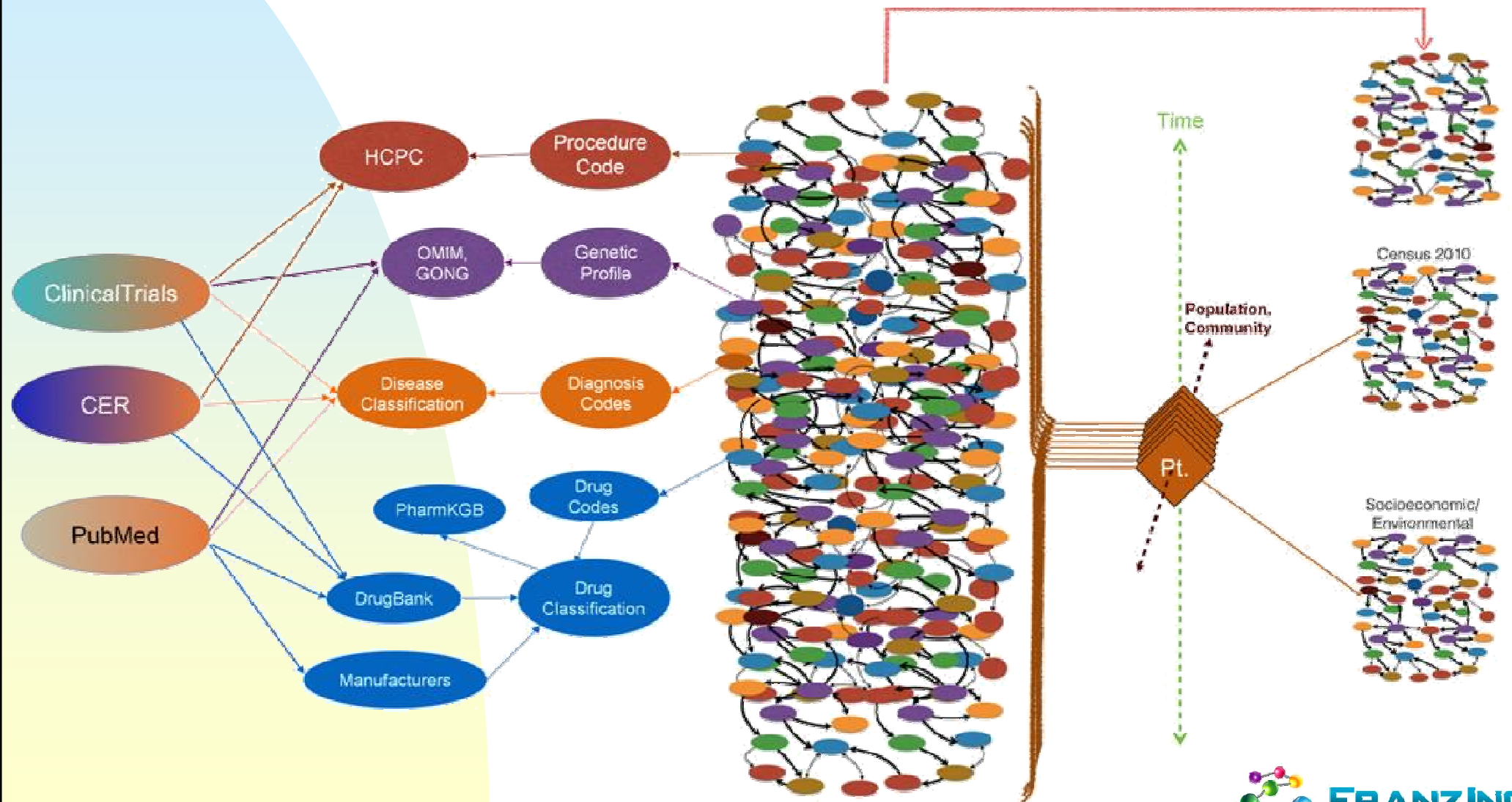
- Cell Function
- Classification
- Clinical Drug
- Concept
- Molecular Function
- Organic Chemical



Link EMR to LOD Medical Knowledgebase



Linking Data with LOD for Analytic Knowledge Graph

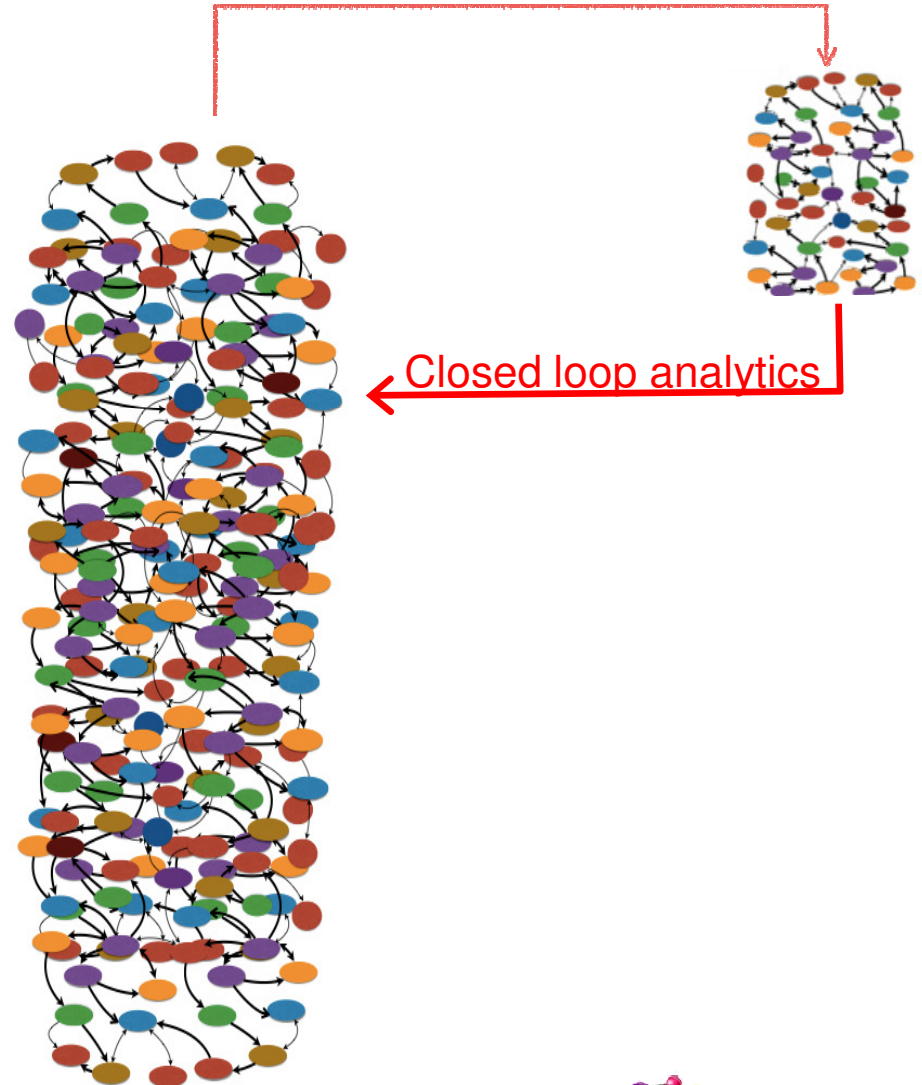


Data Scientist Dream Sandbox

- With such analytic knowledge graph, data scientists can easily perform any analytics
 - Prediction of next possible disease
 - The chance of a patient readmitting within 30 days
 - What treatment most effective for what patients
 - Comparative effectiveness of different treatments
 - The effectiveness and efficiency of a provider (doctor, nurse, department, etc)
- Output hard to read (and use) by medical researchers and doctors

Self-Enhancing Analytic Knowledge Graph

- Feed the analytics output back into the semantic database
- Knowledge abstracted from data analysis becomes part of the “graph”, making it easier to query and navigate
- It's an analytic tapestry



A Probability Knowledge Graph

- Extract all diagnoses (ICD9) from all patients and create ICD→ICD matrix

	Stroke	Cystic Fibrosis	Abdominal Aortic Aneurysm	Eating Disorders
Stroke				
Cystic Fibrosis	4666			
Abdominal Aortic Aneurysm	1331	1343		
Eating Disorders	2789	1999	6915	
Parkinson's Disease	6671	7165	5553	6929
Myocardial Infarction	2981	7744	131	5232
Melanoma	184	6002	467	1249
Dementia	4673	5072	6765	3654
Lupus	7914	5205	5479	1169
Rheumatoid Arthritis	6541	2249	3313	228
Colorectal Cancer	5764	6316	1620	434
Prostate Cancer	4400	4845	7174	66
ADHD	6860	7144	5983	110
Lung Cancer	746	4431	4422	730
Breast Cancer	878	3533	6055	72
Atrial Fibrillation	6314	1388	7995	47
Thrombosis	3521	495	3679	370
Epilepsy	4576	4724	5445	705
Congestive Heart Failure	132	1265	1042	388
Asthma	6133	6520	813	284
COPD	893	1024	977	176

	Stroke	Cystic Fibrosis	Abdominal Aortic Aneurysm	Eating Disorders	ma	COPD
Stroke						
Cystic Fibrosis	4666					
Abdominal Aortic Aneurysm	1331	1343				
Eating Disorders	2789	1999	6915			
Parkinson's Disease	6671	7165	5553	6929		
Myocardial Infarction	2981	7744	131	5232		
Melanoma	184	6002	467	1249		
Dementia	4673	5072	6765	3654		
Lupus	7914	5205	5479	1169		
					2474	

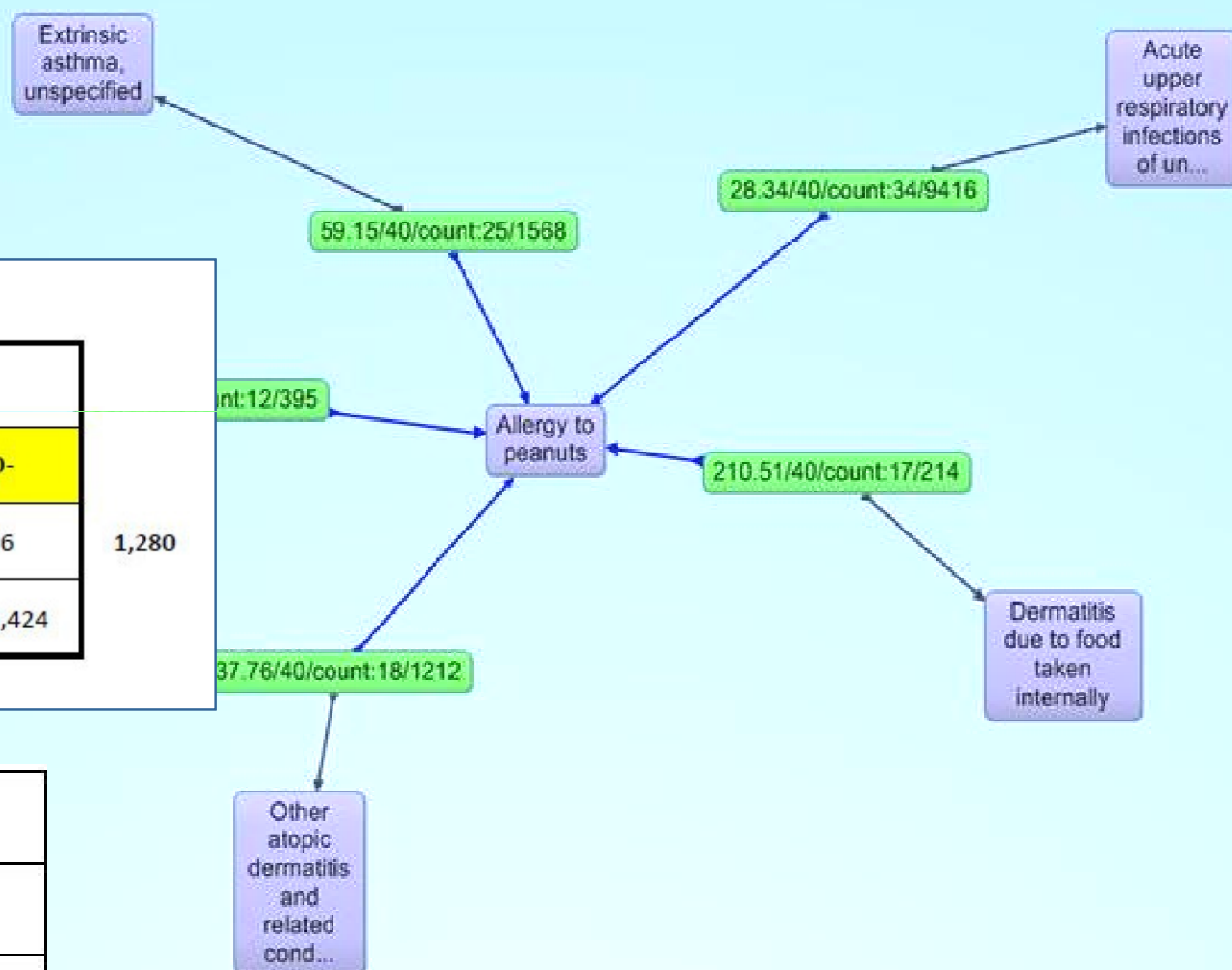
Patient population: 1,802,464

Example: Peanut Allergy Odds Ratio

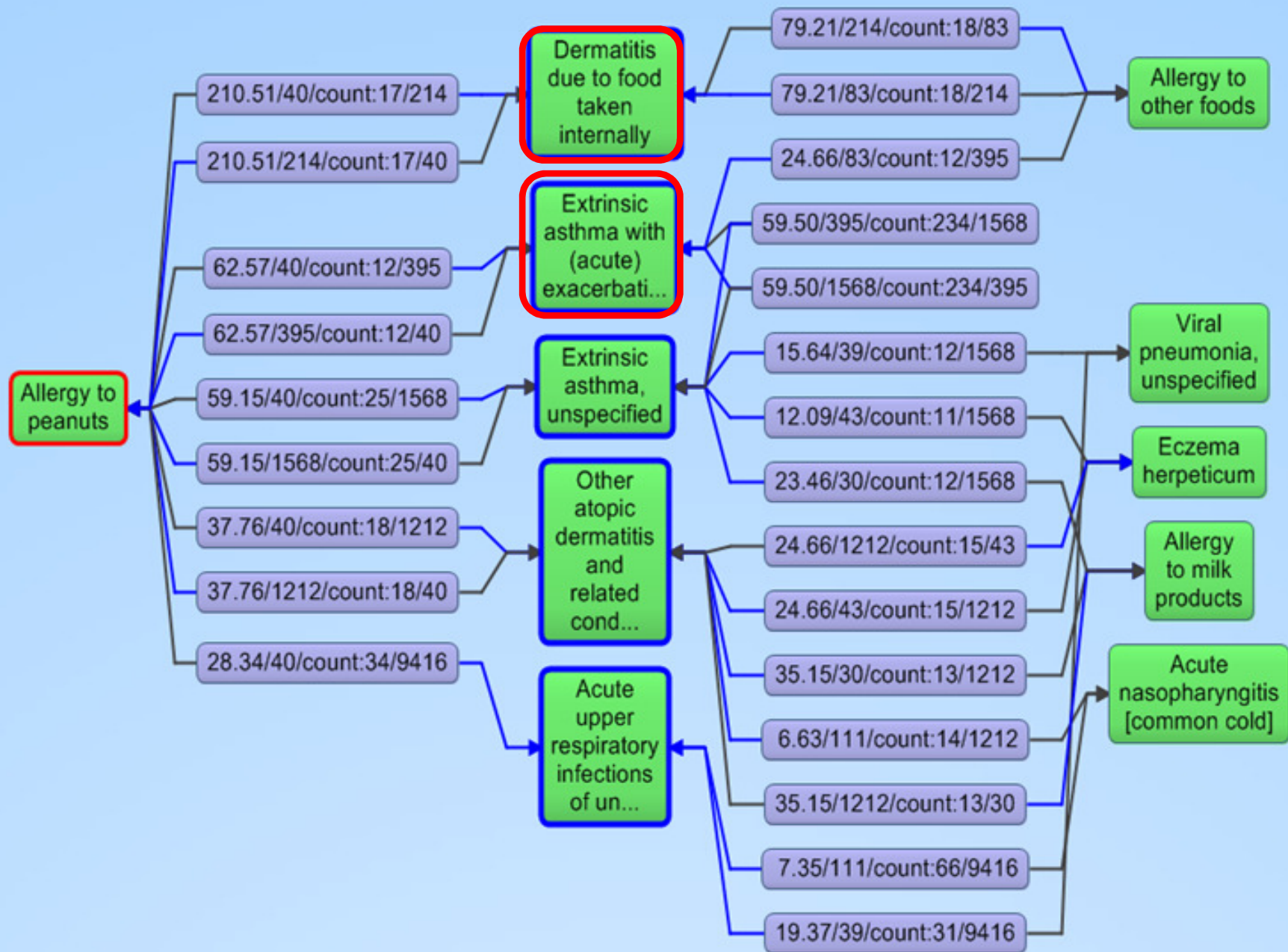
Patient Population		1,802,464	
		Ingestion Dermatitis	
		TO+	TO-
Peanut Allergy	FROM+	544	736
	FROM-	6304	1,795,424
		6,848	

1,280

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

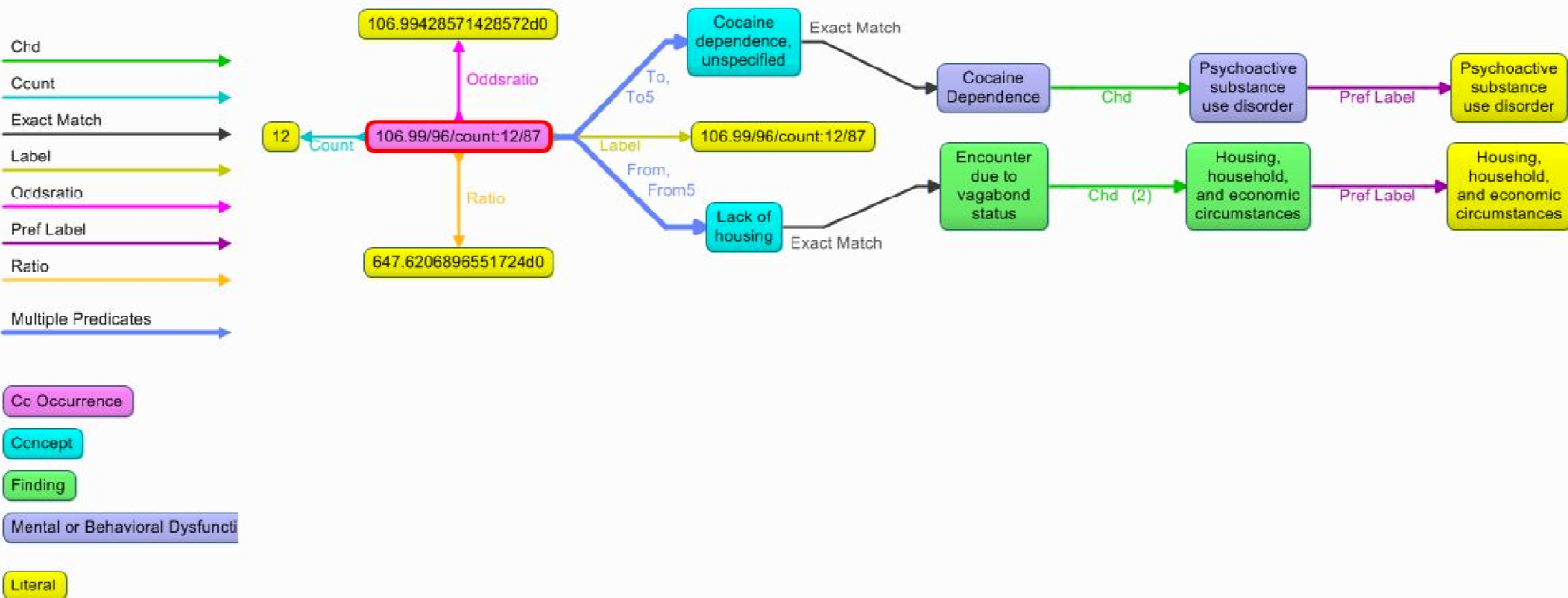


Demo

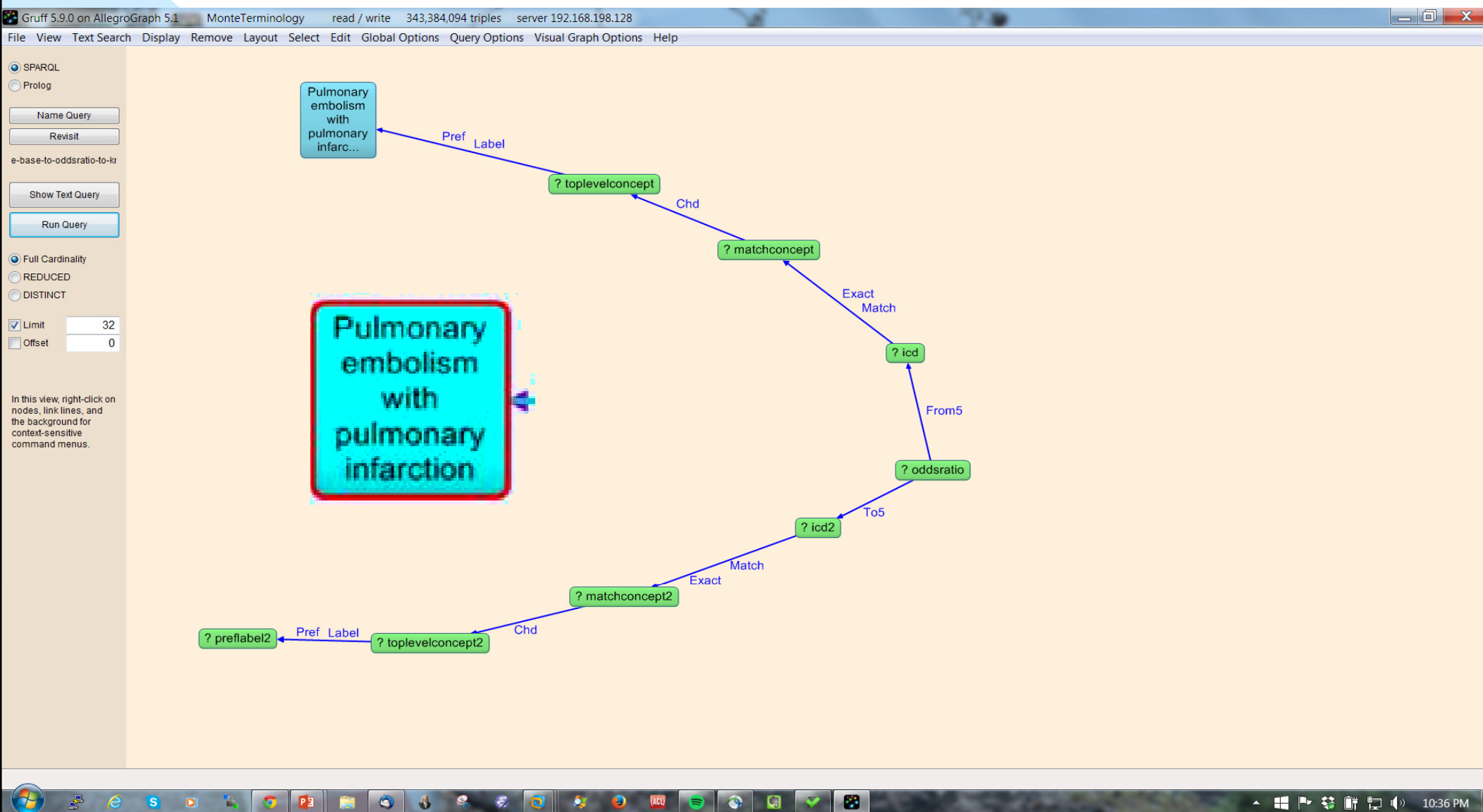


Example: Lack of Housing

- Homeless likely becomes drug addict (OR 107)



Example: Pulmonary Embolism



Chd
Exact Match
From5
Pref Label
To5

Co Occurrence
Concept
Disease or Syndrome
Pathologic Function
Therapeutic or Preventi
Literal

Pulmonary embolism with pulmonary infarction

OR 1,302

Chronic disease of respiratory system

chronische ademhalingsaandoening

Chronic disease of respiratory system

Chronische Erkrankung der Luftwege

Maladie respiratoire chronique

Malattia respiratoria cronica

Idült légzési betegség

pulmonary infarc...

longembolie en -infarct

iatrogenic pulmonary embolism and infarction

iatrogenic pulmonary embolism and infarction

27.92/32/count:11/1048

260.42/261/count:17/32

Other specified surgical operations and ...

O spe sur oper an

Surgical operation and other surgical pr...

처치 당시에는 재난의 언급이 없었으나 환자 에게 이상반응 이나 후에 합병 증을...

Operatie en overige heekundige verricht...

cirugia y otro procedimiento quirúrgico ...

Surgical operation and other surgical pr...

respiratory system

Chronische Erkrankung der Luftwege

Maladie respiratoire chronique

Malattia respiratoria cronica

Idült légzési betegség

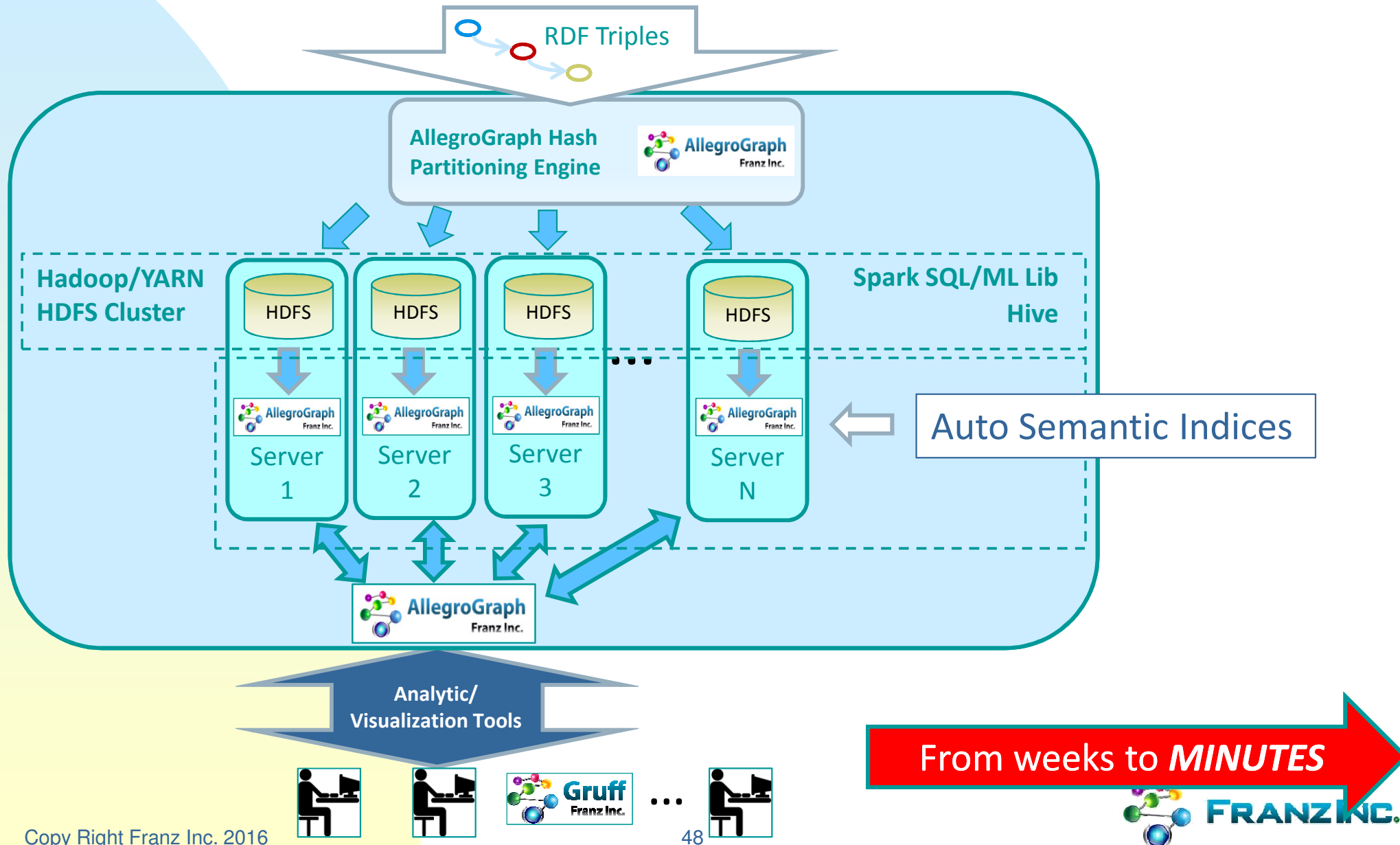
Chronic pulmonary embolism

Chronic disease of respiratory system

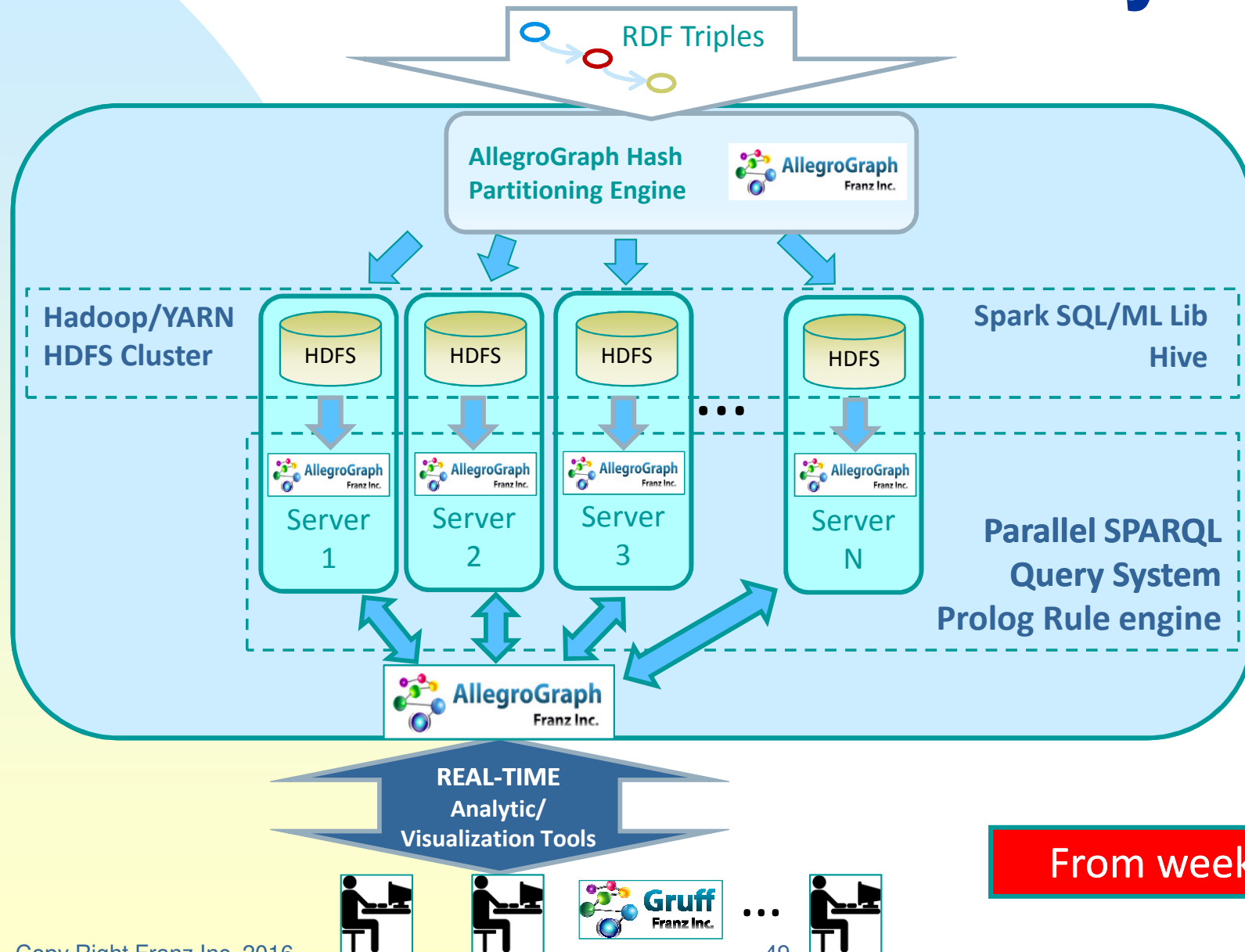
Information sharing and integration is NOT
practical without scalability

SCALABILITY IN DATA SIZE, DATA AVAILABILITY AND QUERY PERFORMANCE

A New Semantic DB Architecture – Semantic Data Lake in Hadoop



Semantic Data Lake – Parallel Distributed Query



Hadoop/Hive/Spark

- Distributed repository
- Distributed processing
- High Availability
- Machine learning
- Unlimited scale

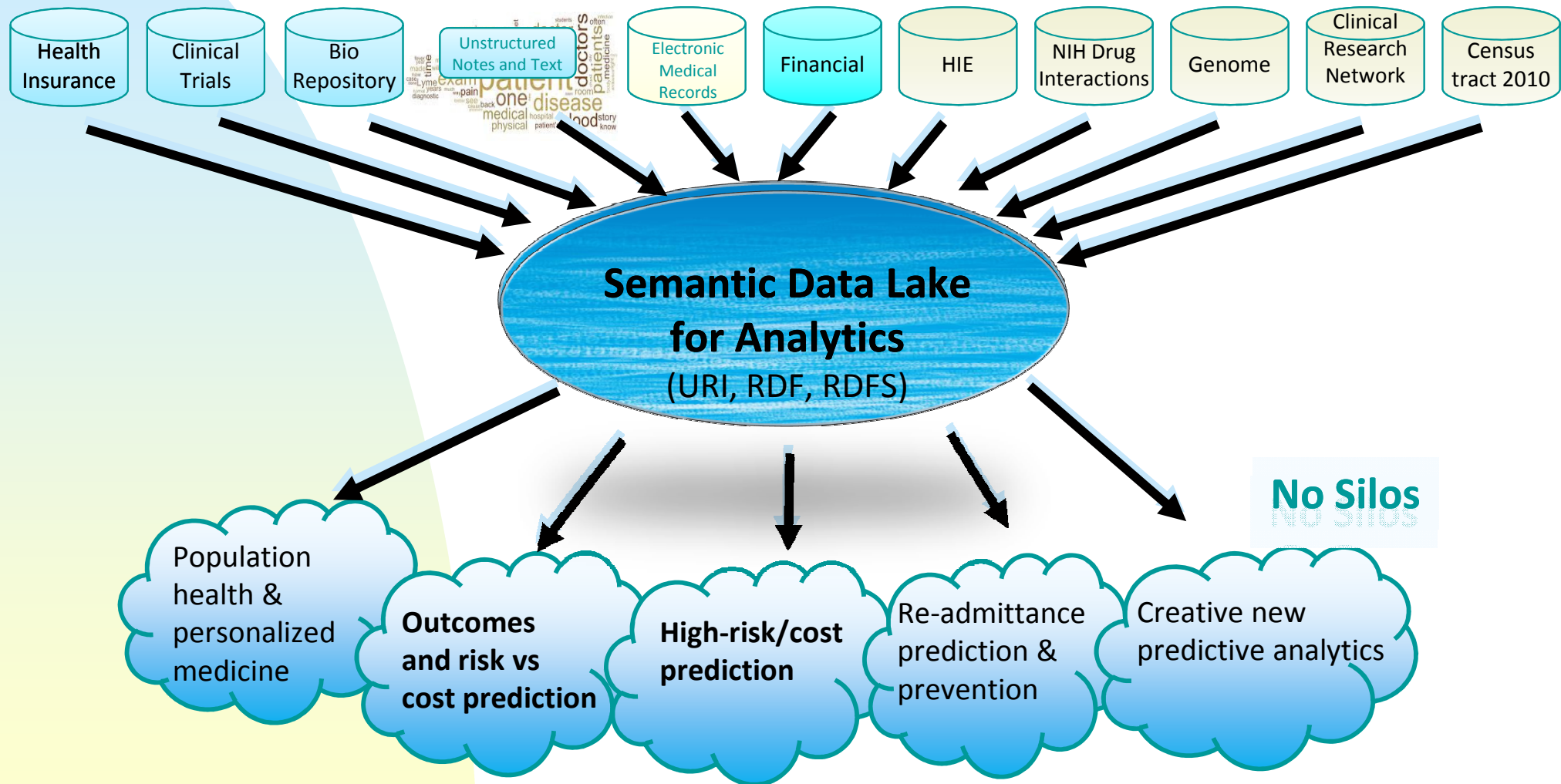
AllegroGraph Semantic Graph index

- Distributed index
- Distributed query
- Semantics index
- Reasoning & rules
- Ontology enforcement
- Unlimited scale

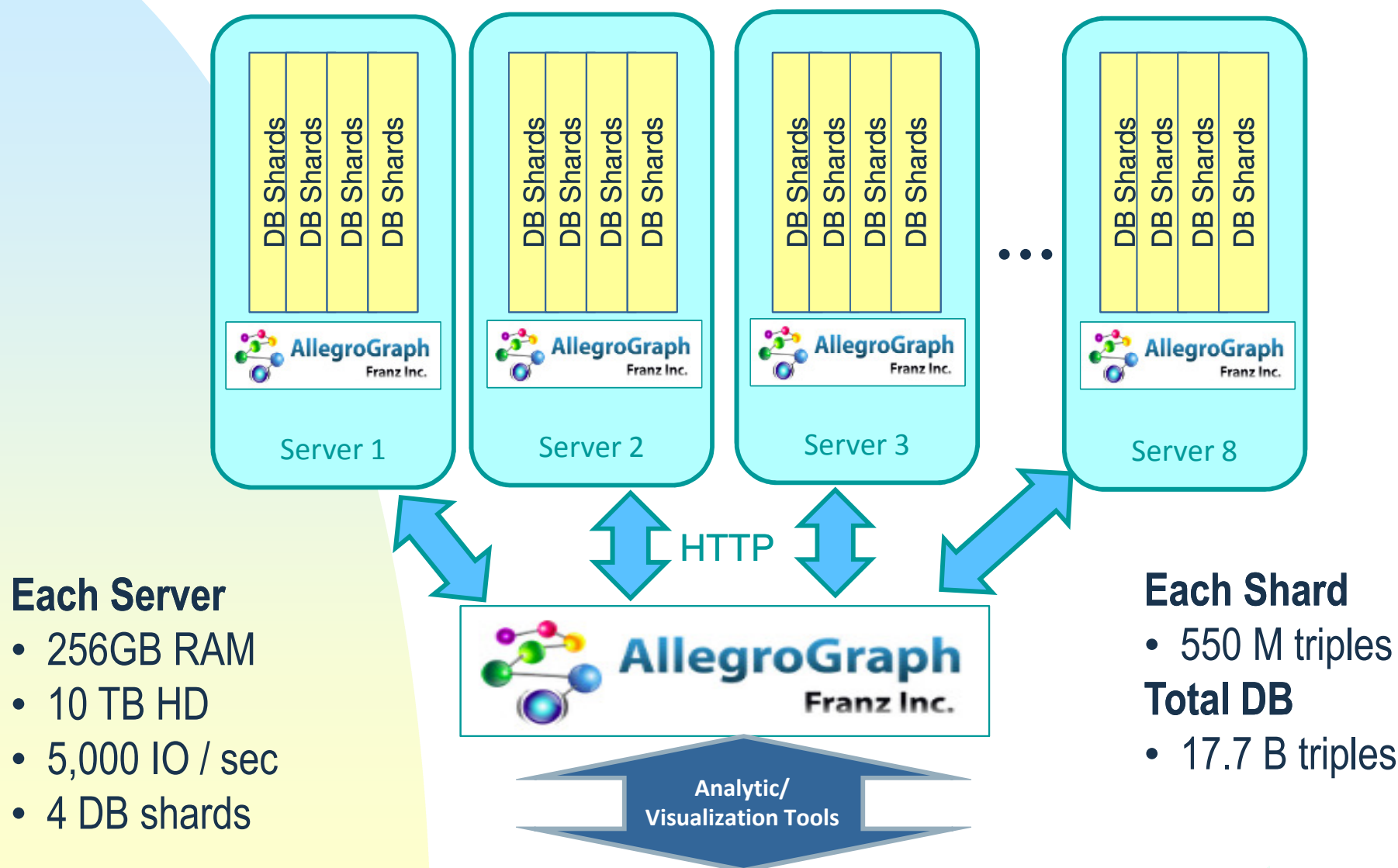
From weeks to **MINUTES**



Semantic Data Lake – A Distributed Semantic Database



Benchmark Cluster Configuration



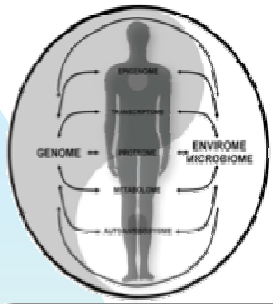
Benchmark Queries

- **Attrition:** how many patients we lose (for any reason) from the year 2010 to the year 2011. Medium complexity as it contains some interesting unions and joins and filters. (CPU Intensive)
- **Demographics:** query some demographics for all patients
- **Diagnostics:** query every diagnostic in the database for inpatient/outpatient encounters. (Data intensive)
- **CDRN:** an example of an ETL query used for the CDRN project, high complexity

Benchmark Results – ~Constant Horizontal Scalability

	Attrition (sec)	Demographics (sec)	Diagnostics (sec)	CDRN (sec)
Machine 1, Shard 1 (1/32)	4.2	9.8	222	426
Machine 1, 4 Shards (1/8)	5.8	12	226	516
Cluster	7.2	17.6	418	519
Small (1/32 data on Bignew)	5.5	18.4	254	543
Bignew (0.5TB RAM, in-memory store)	609	622	9065	not completed)
Rows of bindings returned	1	1,600,000	69,000,000	67,000,000
Query characteristics	CPU Intensive		Data Intensive	
Performance ratio cluster/bignew	84.58	35.34	21.69	N/A
Performance ratio cluster/1shard	1.71	1.80	1.88	1.22

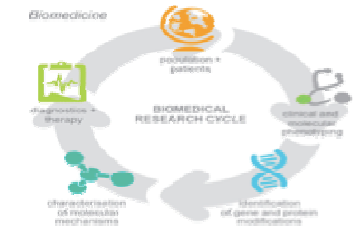
The Value Propositions



Personalized Medicine



Predictive Modeling



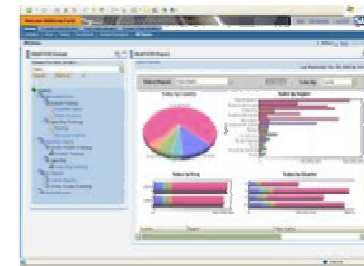
Translational Research



Fraud Detection



Risk Assessment



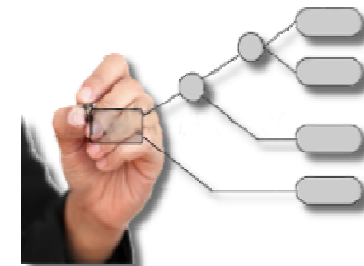
Business Intelligence



Public Health



Mobile Health



Decision Support

Summary

- The unreasonable effectiveness of (big) data, applying machine learning on big data
- Data offer great healthcare values, if we can
 - Integrate EMR data silos
 - Leverage existing LOD medical knowledge
 - Achieve scalability in data size and performance
- That's what Semantic Data Lake is all about – the Analytic Knowledge Graph for Healthcare!

Thank You

scw@franz.com

*Analytics Knowledge Graph
that Works for Healthcare*