The Semantic Data Lake in Healthcare

Parsa Mirhaji MD, PhD
Director, Clinical Research Informatics, Montefiore Health System
CTO, NYC-Clinical Data Research Network
Analytic Maturity Model

Level 8: Personalized Medicine & Prescriptive Analytics
Level 7: Clinical Risk Intervention & Predictive Analytics
Level 6: Population Health Management & Suggestive Analytics
Level 5: Waste & Care Variability Reduction
Level 4: Automated External Reporting
Level 3: Automated Internal Reporting
Level 2: Standardized Vocabulary & Patient Registries
Level 1: Enterprise Data Warehouse
Level 0: Fragmented Point Solutions

Evidence Based Learning System
ACO (Managed care, Quality driven, Cost-efficient, Population based)

ROI (EMR)
Ubiquitous Data Access
Regulatory Compliance
Data Culture Emergence
Collaborations and Partnerships

Reporting
(Dashboards, Data Marts)
Semantic Data Lake (Argos)

- Platform for Big Data Analytics and Cognitive Computing
- Based on principles of HDFS/Spark/GraphX, and Semantic Web
- Knowledge Based (metadata, terminology, ontologies)
  - Enables linking Data -> Knowledge -> Linked Open Data
- URI level security, authorization, provenance
- Optimized for integration of heterogeneous health and well-ness data
  - Structured and non-structured data
- Sophisticated temporal and geospatial analytics, Machine Learning
- To support basic science, translational research, operational care coordination, population health management (ACO)
- A Montefiore, Franz, Intel, Cisco Collaboration
Our Evidence Generation Paradigm:

- ClinicalTrials
- CER
- PubMed

HCPC → Procedure Code
OMIM, GONG → Genetic Profile
Disease Classification → Diagnosis Codes
PharmKGB → Drug Codes
DrugBank → Drug Classification
Manufacturers

Analytic Tapestry (closed loop analytics)

Time
Population, Community
Pt.
Census 2010
Socioeconomic/Environmental
The Knowledgebase

- OMOP
- OMIM
- Geospatial & Census
- GO
- NCI Thesaurus
- ClinicalTrials
- PubMED
- DrugBank
- PharmKGB
- LOD
- SKOS
All UMLS Source Vocabularies
(Linked to OMOP Terminology)
Diabetes Mellitus
Taxonomic and Semantic Relations
CONSTRUCT { 
?code skos:exactMatch mth:C0011849. 
?code ndfrt:may_be_treated_by ?treat. 
}
WHERE 
{ 
?code skos:exactMatch mth:C0011849. 
?code ndfrt:may_be_treated_by ?treat. 
?cui a semnet:Clinical_Drug. 
}
The Data Lake
Event driven model

MPI

Record Locator

Health Data

Web/LOD Ontologies

Contextualization Ontologies

Temporal Entity

Description (SKOS)

UMLS (SKOS)

AnatomicConcept (SKOS)

Demographics

Source

Documentation

Consent

Modifiers

Unit

History

Device

Gender

Race

Qualitative

Quantitative

Certainty

Negation

Laterality

Normality

Scale

Size

Severity
Semantic Data Lake:
Patient having an Emergency Encounter with Balanoposthitis
Metadata Management
Lineage, Provenance, Quality of Data
Analytic Tapestry

Closed loop analytics
Evidence Generation in Big-Data Environments

- Evidence generation frequently happens within fragmented and siloed analytic data-marts outside of enterprise resources, and are not traceable.
- No formal and consistent method of tracking and linking the value added information to the original source data repositories (e.g. EDW).
- This is an impediment to evidence based delivery of care, reproducibility of research findings, security and compliance, and results in loss of value-added information, and enterprise intellectual property and assets, and unnecessary duplication of efforts.
The Analytic Tapestry
A Generic Method for Persisting Analytics in Big-Data Environments

• *Transforming a big-data repository to a system of insight*

• Systematically binds the results of statistical and analytic processes, and their computational byproducts to the following:
  – The underlying data and the context within which the analytics are done
  – The description of *the analytic processes and session metadata*
  – The description of the results and analytic byproducts

• Enables construction of higher level analytics based on byproducts of previous findings (meta-analysis)

• Enables systematically and longitudinally linking all analytic byproducts related to a patient to the patient’s profile
  – Key component of enterprise patient engagement strategy, by enabling access to research findings and analytics that pertain to an individual, within the enterprise

• Systematic enterprise wide view of all analytic processes on data at all times.
Analytic Tapestry
A Graphical Query
The Semantic Data Lake

Event-Time Ontology

- OMOP
- OMIM
- PharmKGB
- PubMED
- DrugBank
- NCI Thesaurus
- ClinicalTrials
- Geospatial & Census
- GO

Registries
- Financial
- Free Text
- PRO
- Devices
- Bio Bank
- EMR
- LIMS
- HL7 Feed
- Web Services
- Legacy
- Omics
- CTMS
- Claims

PharmKGB

Time

Pt.
The Annotation Engine
The Semantic Data Lake

Registries, Financial, Free Text, PRO, Devices, Bio Bank
EMR, LIMS, HL7 Feed, Web Services, Legacy
Omics, CTMS
Claims

SDL Loader

HDFS
Allegro Graph
Spark
ML-Lib/R
Prolog
Java API

SPARQL