Alternative database architectures for clinical data management

James R. Campbell MD
W. Scott Campbell PhD MBA

Nebraska Medicine
University of Nebraska Medical Center
Omaha, NE

Nebraska Medicine
SERIOUS MEDICINE. EXTRAORDINARY CARE.

Overview

- Employment of graph database technology for SNOMED CT in context of clinical use
- Initial experiments and results
- Current Use
- Future work

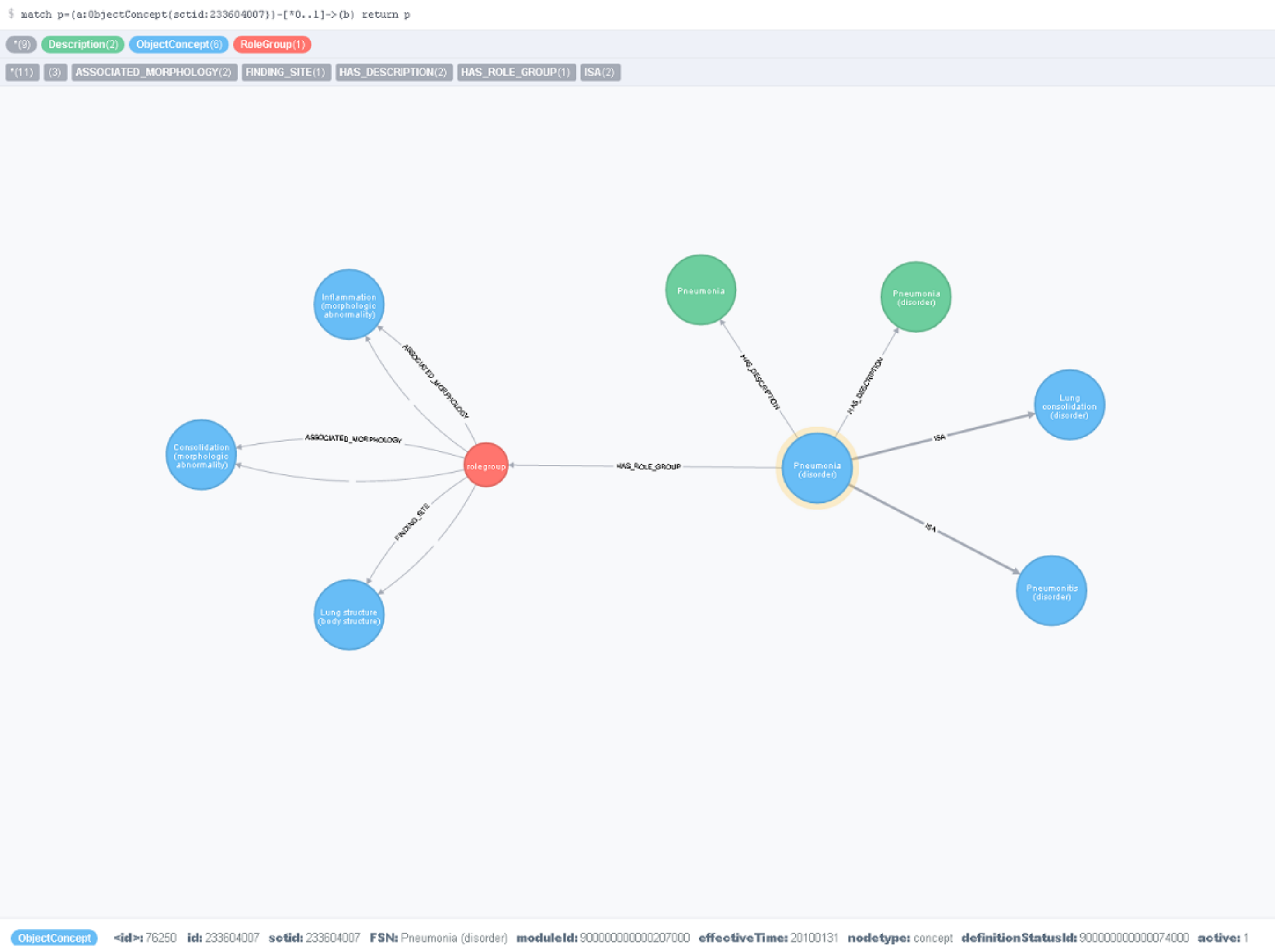
Initial Use Case – circa 2015

- Instantiate a data base with numerous, real-time post-coordinated expressions of surgical pathology findings.
- Relational database designs resulted in HUGE join tables
  - Suggested a use case for a triple-store database (RDF?)
  - Investigation of NoSQL options suggested graphDB’s
- Graph databases:
  - Class of NoSQL
  - Emphasize connectedness of data vs. rows/columns of data
  - Open world vs. closed world
    - Flexible
    - Transactionally ACID properties
  - SNOMED CT is a directed, acyclic graph
- Used Neo4j (San Mateo, CA), open sourced, java based

Approach

- Graphs consist of Nodes and Relationships (edges) that connect Nodes
  - Nodes and Edges can have properties (“Property Graph”)
- Used Snapshot, RF2 release of SNOMED CT International release (classified version)
- All SNOMED CT concepts represented as nodes
  - All RF2 metadata represented as properties of nodes
  - Active, module ID, definition status ID, effective time
- All SNOMED CT attributes represented as edges
  - RF2 Metadata as properties
- All names set as nodes with relationship to SNOMED CT expression node
- Result: A graph database with 100% of SNOMED CT content
- Fast! – Transitive Closure Calculation time < 60 sec on laptop

Example: Pneumonia



```
$ match p=(a:ObjectConcept{sctid:233604007})-[*0..1]->(b) return p
```

*(9) Description(2) ObjectConcept(6) RoleGroup(1)

*(11) (3) ASSOCIATED_MORPHOLOGY(2) FINDING_SITE(1) HAS_DESCRIPTION(2) HAS_ROLE_GROUP(1) ISA(2)

ObjectConcept **<id>:** 76250 **id:** 233604007 **sctid:** 233604007 **FSN:** Pneumonia (disorder) **moduleId:** 900000000000207000 **effectiveTime:** 20100131 **nodetype:** concept **definitionStatusId:** 900000000000074000 **active:** 1

Add Patient Data

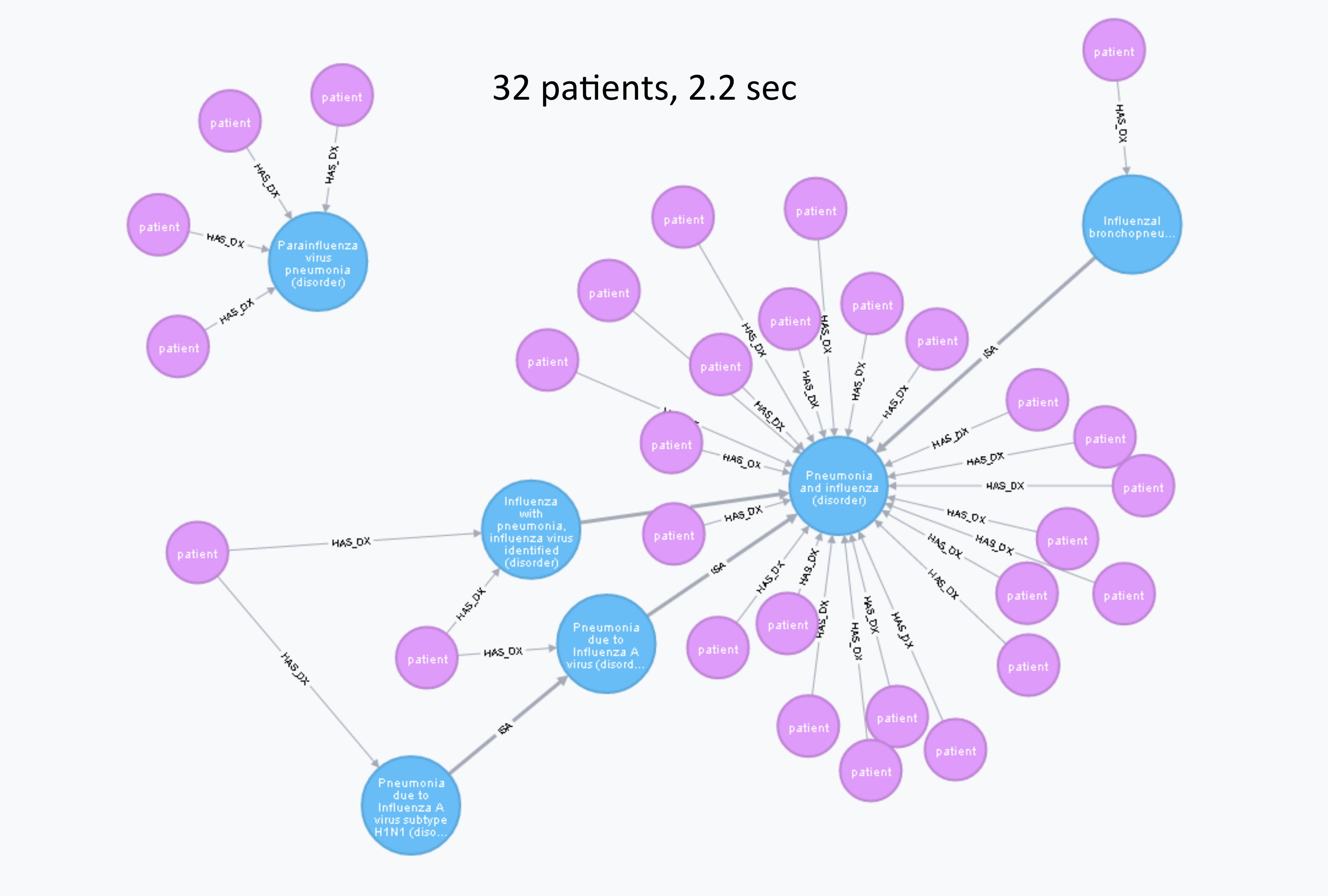
- Import patient records from de-identified clinical data warehouse

- Approximately 465,000 patients

- Import patient problem lists (All SNOMED CT encoded)
  - Up to 20 years of data
  - 2,770,000 diagnoses in total
  - Properties:
    - Date of diagnosis (start and end dates)
    - Active, inactive or deleted status

- Result
  - Patient identification by SNOMED CT codes/subsumption same as RDBMS based clinical data warehouse.
  - Queries were fast! Desktop on par with enterprise class server.
  - Unintended finding: Queries of negation, disjunction, depth

Example: Find all patients with Pneumonia due to some influenza virus or some parainfluenza virus

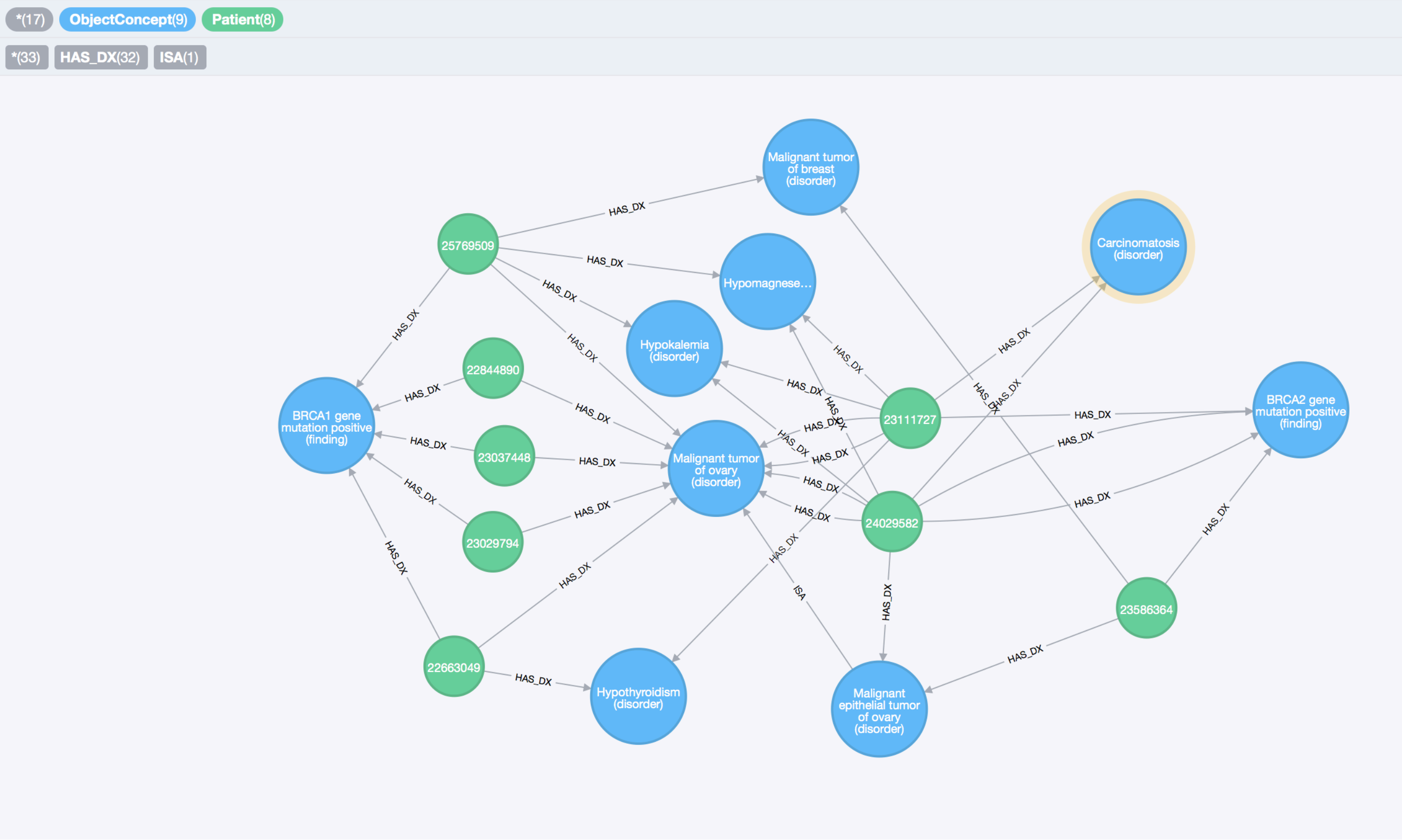


Parainfluenza virus pneumonia (disorder)
Influenzal bronchopneu...
Pneumonia and influenza (disorder)
Influenza with pneumonia, influenza virus identified (disorder)
Pneumonia due to Influenza A virus (disord...
Pneumonia due to Influenza A virus subtype H1N1 (diso...
patient
HAS_DX
ISA

32 patients, 2.2 sec

Queries of undefined depth

- Find all patients with positive BRCA1 or BRCA2 gene mutation who have an


*(17)
ObjectConcept(9)
Patient(8)
*(33)
HAS_DX(32)
ISA(1)
Malignant tumor of breast (disorder)
Carcinomatosis (disorder)
Hypomagnese…
Hypokalemia (disorder)
BRCA1 gene mutation positive (finding)
Malignant tumor of ovary (disorder)
BRCA2 gene mutation positive (finding)
Hypothyroidism (disorder)
Malignant epithelial tumor of ovary (disorder)
25769509
22844890
23037448
23029794
22663049
23111727
24029582
23586364
HAS_DX
ISA

How about Historicity?

- Graph database calculation of transitive closure table – FAST
  - Can the database produce TC tables for multiple years AND a Delta TC table between and two release dates?
  - Beneficial for SNOMED CT sites to assess of effects of terminology updates on implementations

Result

- Following methods used previously for a single release
- Added property to maintain historical representations of SNOMED CT concepts and relationships
- Instantiated graph DB with classified, full RF2 release
  - 6 GB, ~425K concepts and ~6.9M relationships
- TC calculations created using the graph model by year match TC tables created for any single release year.
- Creation of delta TC table between any two years < 4 min
  - TC table year 1 < 30 sec
  - TC table year 2 < 30 sec
  - Delta TC table calculation and write to file – 2.5 min

What about patient data?

- Added same patient data used in Snapshot graph DB
  - 465,000 patients
  - ~2.77 million associated problems/clinical findings (20140901 US extension)
  - GraphDB build on 20150901 US extension)

- Queried all SNOMED CT expressions with existing relationship to any patient AND Active status = ‘0’ (Inactive concept)
  - Return – 79 inactive concepts
  - Affected – 6134 distinct patients
  - All concept changes due to changes in 20150131 International release

ID all patients with active diagnosis linked to inactive SNOMED CT concept

SCTID	Fully Specified Name	Patients
23346002	Sunburn (disorder)	7
91340006	Extrinsic asthma with status asthmaticus (disorder)	6
601000119109	History of bee sting allergy (situation)	1
71275003	Pseudoprimary aldosteronism (disorder)	17
431347008	Lipodystrophy associated with Human immunodeficiency virus infection (disorder)	4
312403005	Legionnaire's disease (disorder)	6
367530008	Spondyloepiphyseal dysplasia congenita (disorder)	3
440181000	Apparent life-threatening event (finding)	19
44008002	Somatotropin deficiency (disorder)	131
395657006	Pallister-Killian syndrome (disorder)	1
429081000124107	History of extracorporeal membrane oxygenation (situation)	15

What we learned

- Information model and design places semantic terminology/ concept model at core of database
- Patient data is built upon the semantics initially vs. terminology as an afterthought
- Queries start with the full semantic model (SNOMED CT)
  - Real-time subsumption queries without logical abstraction (transitive closure)
  - Semantic queries using defining attribute edges vs. ISA-only at run time
- Persistent and query-able representation of data over time in BOTH current and past SNOMED CT representations

Current Use Case (Nebraska CARES)

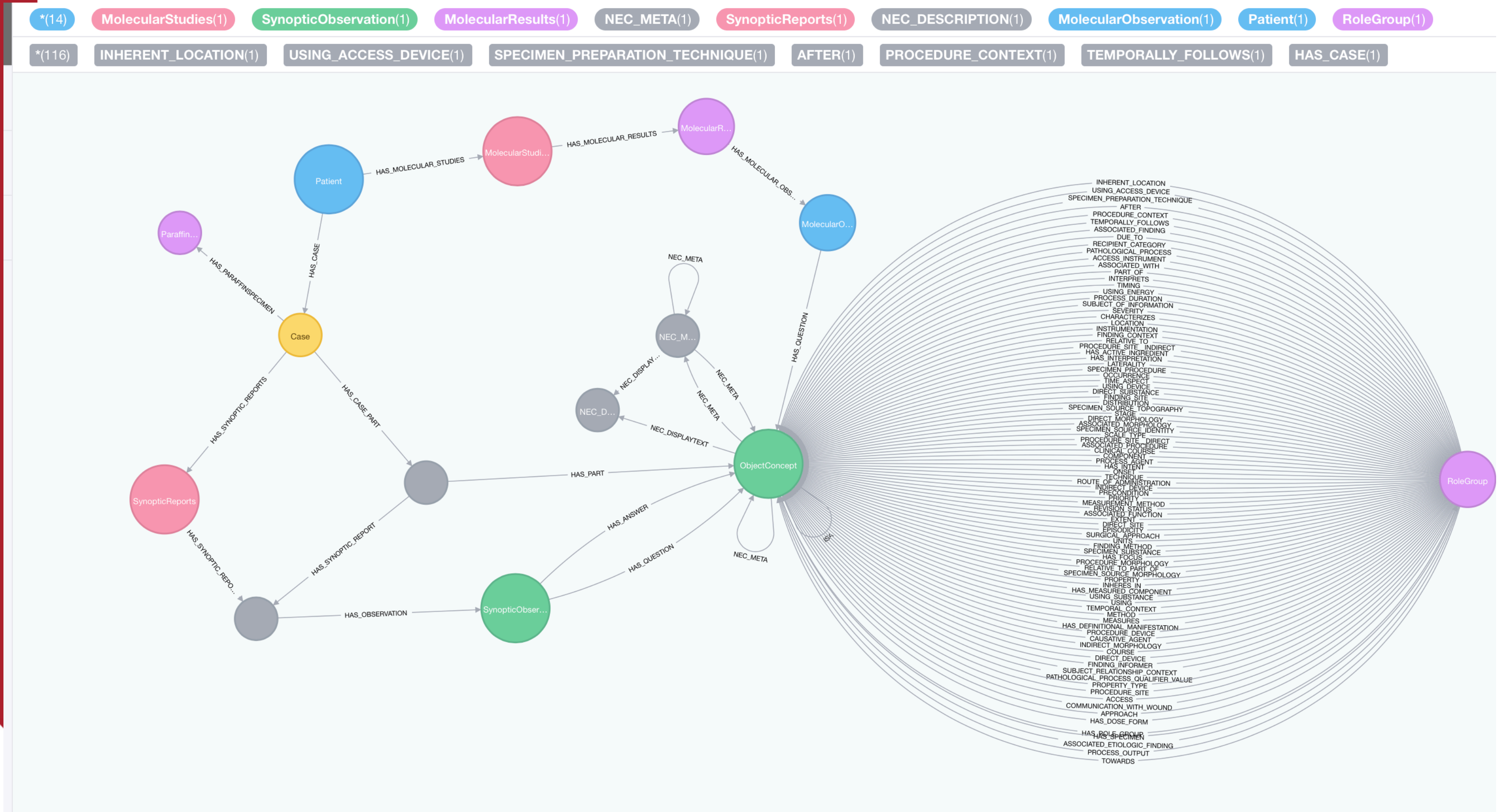
- Deploy graph model with SNOMED semantic core
  - Cancer registry integrated into operational ecosystem
  - Biorepository and inventory management

- Expose to general users for de-identified exploration of tissue availability by characteristics

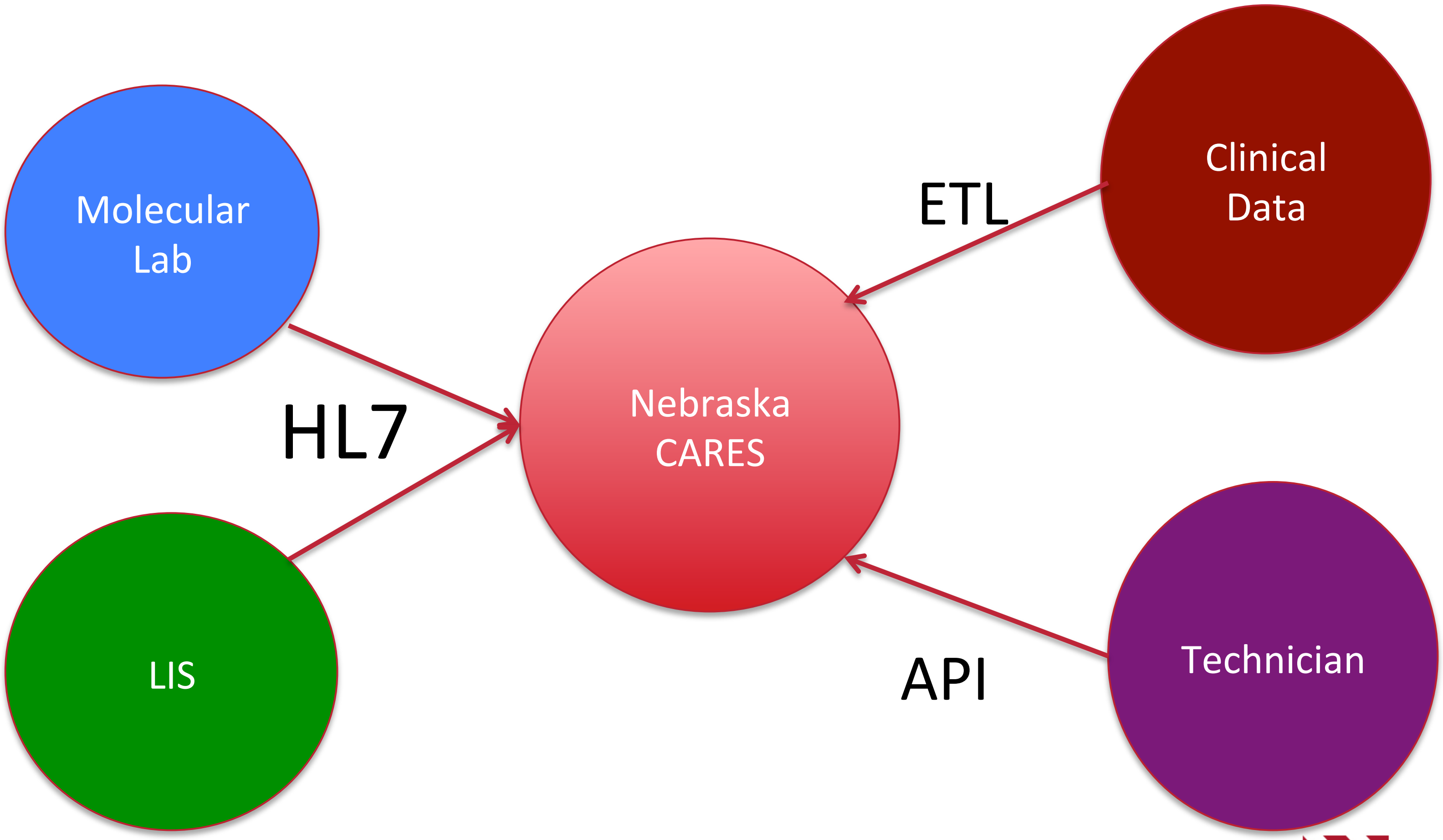
Information modeling

- ➢ Nodes used for SNOMED CT concepts
- ➢ Edges used for SNOMED CT attributes and relationships
- ➢ Nodes used for:
  - – Patients, Cases, Tissue Specimens, synoptic cancer reporting values, sequence results
  - – Edges for all type of relationships
- ➢ Sparseness of data (i.e. normal form)

ER Diagram


*(14)
MolecularStudies(1)
SynopticObservation(1)
MolecularResults(1)
NEC_META(1)
SynopticReports(1)
NEC_DESCRIPTION(1)
MolecularObservation(1)
Patient(1)
RoleGroup(1)
*(116)
INHERENT_LOCATION(1)
USING_ACCESS_DEVICE(1)
SPECIMEN_PREPARATION_TECHNIQUE(1)
AFTER(1)
PROCEDURE_CONTEXT(1)
TEMPORALLY_FOLLOWS(1)
HAS_CASE(1)
Patient
HAS_MOLECULAR_STUDIES
MolecularStudi...
HAS_MOLECULAR_RESULTS
MolecularR...
HAS_MOLECULAR_OBS...
MolecularO...
Paraffin...
HAS_PARAFFINSPECIMEN
HAS_CASE
Case
HAS_SYNOPTIC_REPORTS
HAS_CASE_PART
SynopticReports
HAS_SYNOPTIC_REPO...
HAS_SYNOPTIC_REPORT
HAS_PART
HAS_OBSERVATION
SynopticObser...
HAS_ANSWER
HAS_QUESTION
HAS_QUESTION
NEC_META
NEC_M...
NEC_DISPLAY...
NEC_D...
NEC_META
NEC_META
NEC_DISPLAYTEXT
ObjectConcept
NEC_META
ISA
RoleGroup
INHERENT_LOCATION
USING_ACCESS_DEVICE
SPECIMEN_PREPARATION_TECHNIQUE
AFTER
PROCEDURE_CONTEXT
TEMPORALLY_FOLLOWS
ASSOCIATED_FINDING
DUE_TO
RECIPIENT_CATEGORY
PATHOLOGICAL_PROCESS
ACCESS_INSTRUMENT
ASSOCIATED_WITH
PART_OF
INTERPRETS
TIMING
USING_ENERGY
PROCESS_DURATION
SUBJECT_OF_INFORMATION
SEVERITY
CHARACTERIZES
LOCATION
INSTRUMENTATION
FINDING_CONTEXT
RELATIVE_TO
PROCEDURE_SITE__INDIRECT
HAS_ACTIVE_INGREDIENT
HAS_INTERPRETATION
LATERALITY
SPECIMEN_PROCEDURE
OCCURRENCE
TIME_ASPECT
USING_DEVICE
DIRECT_SUBSTANCE
FINDING_SITE
DISTRIBUTION
SPECIMEN_SOURCE_TOPOGRAPHY
STAGE
DIRECT_MORPHOLOGY
ASSOCIATED_MORPHOLOGY
SPECIMEN_SOURCE_IDENTITY
SCALE_TYPE
PROCEDURE_SITE__DIRECT
ASSOCIATED_PROCEDURE
CLINICAL_COURSE
COMPONENT
PROCESS_AGENT
HAS_INTENT
ONSET
TECHNIQUE
ROUTE_OF_ADMINISTRATION
INDIRECT_DEVICE
PRECONDITION
PRIORITY
MEASUREMENT_METHOD
REVISION_STATUS
ASSOCIATED_FUNCTION
EXTENT
DIRECT_SITE
EPISODICITY
SURGICAL_APPROACH
UNITS
FINDING_METHOD
SPECIMEN_SUBSTANCE
HAS_FOCUS
PROCEDURE_MORPHOLOGY
RELATIVE_TO_PART_OF
SPECIMEN_SOURCE_MORPHOLOGY
PROPERTY
INHERES_IN
HAS_MEASURED_COMPONENT
USING_SUBSTANCE
USING
TEMPORAL_CONTEXT
METHOD
MEASURES
HAS_DEFINITIONAL_MANIFESTATION
PROCEDURE_DEVICE
CAUSATIVE_AGENT
INDIRECT_MORPHOLOGY
COURSE
DIRECT_DEVICE
FINDING_INFORMER
SUBJECT_RELATIONSHIP_CONTEXT
PATHOLOGICAL_PROCESS_QUALIFIER_VALUE
PROPERTY_TYPE
PROCEDURE_SITE
ACCESS
COMMUNICATION_WITH_WOUND
APPROACH
HAS_DOSE_FORM
HAS_ROLE_GROUP
HAS_SPECIMEN
ASSOCIATED_ETIOLOGIC_FINDING
PROCESS_OUTPUT
TOWARDS

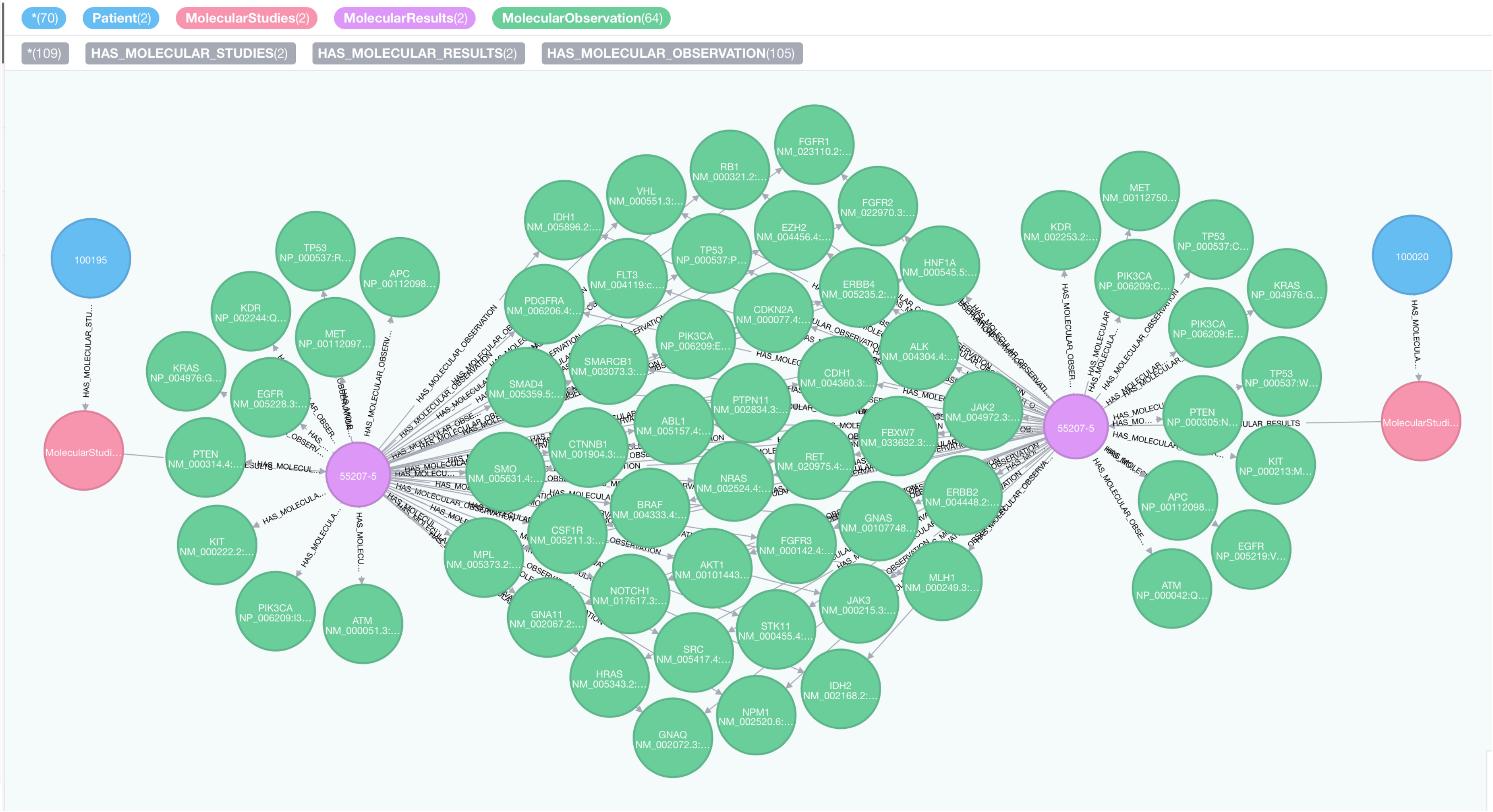
Data flow


Molecular Lab
LIS
HL7
Nebraska CARES
ETL
Clinical Data
API
Technician

Query Sample

- All patients with Colorectal Cancer
- Histology type of Mucinous carcinoma
- All NGS results in 50 gene panel
  - Those that are in common
  - Those that are different

Visual Result



Long term objectives

- Goal: create entire clinical data warehouse using the graph model

- Compare performance to existing RDBMS models

- Desired benefits
  - Queries of undefined depth (tractable)?
  - Pattern identification
    - Health and disease are patterns
    - New relationship identification – correlation/causation

Challenges and Learnings

- Information modeling
  - This is not your parents’ RDBMS
  - Requires changes to modeling
    - Sparseness of data
    - Edges are key
- It’s in the Java
  - Neo4J (i.e., Neo for Java)
  - Plug-ins access graph algorithms not directly available in Cypher

References

- 1. Lee D, Cornet R, Lau F, de Keizer N. A survey of SNOMED CT implementations. J Biomed Inform. 2013 Feb;46(1):87-96.

- 2. Campbell WS, Campbell JR, West WW, McClay JC, Hinrichs SH. Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings. J Am Med Inform Assoc. 2014 Sep;21(5):885-92.

- 3. Campbell WS, Pedersen J, McClay JC, Rao P, Bastola D, Campbell JR. An alternative database approach for management of SNOMED CT and improved patient data queries. J Biomed Inform. 2015 Aug 21.

- 4. Campbell JR, Campbell WS, Hickman H, Pedersen J, McClay J. Employing complex polyhierarchical ontologies and promoting interoperability of i2b2 data systems. *[Accepted Proc AMIA Symp. 2015]*

Questions?

James R. Campbell
campbell@unmc.edu

W. Scott Campbell
wcampbel@unmc.edu

Nebraska Medicine

SERIOUS MEDICINE. EXTRAORDINARY CARE.®