Graph database approach to management and use of SNOMED CT encoded clinical data

W. Scott Campbell, Jay Pedersen, James R. Campbell
University of Nebraska Medical Center

University of Nebraska Medical Center

Overview

- Employment of graph database technology for SNOMED CT in context of clinical use
- Result of SIA Schema Project:
  - Manage post-coordinated expressions for surgical pathology
- Results and additional findings
- Sample queries

Initial Use Case

- 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

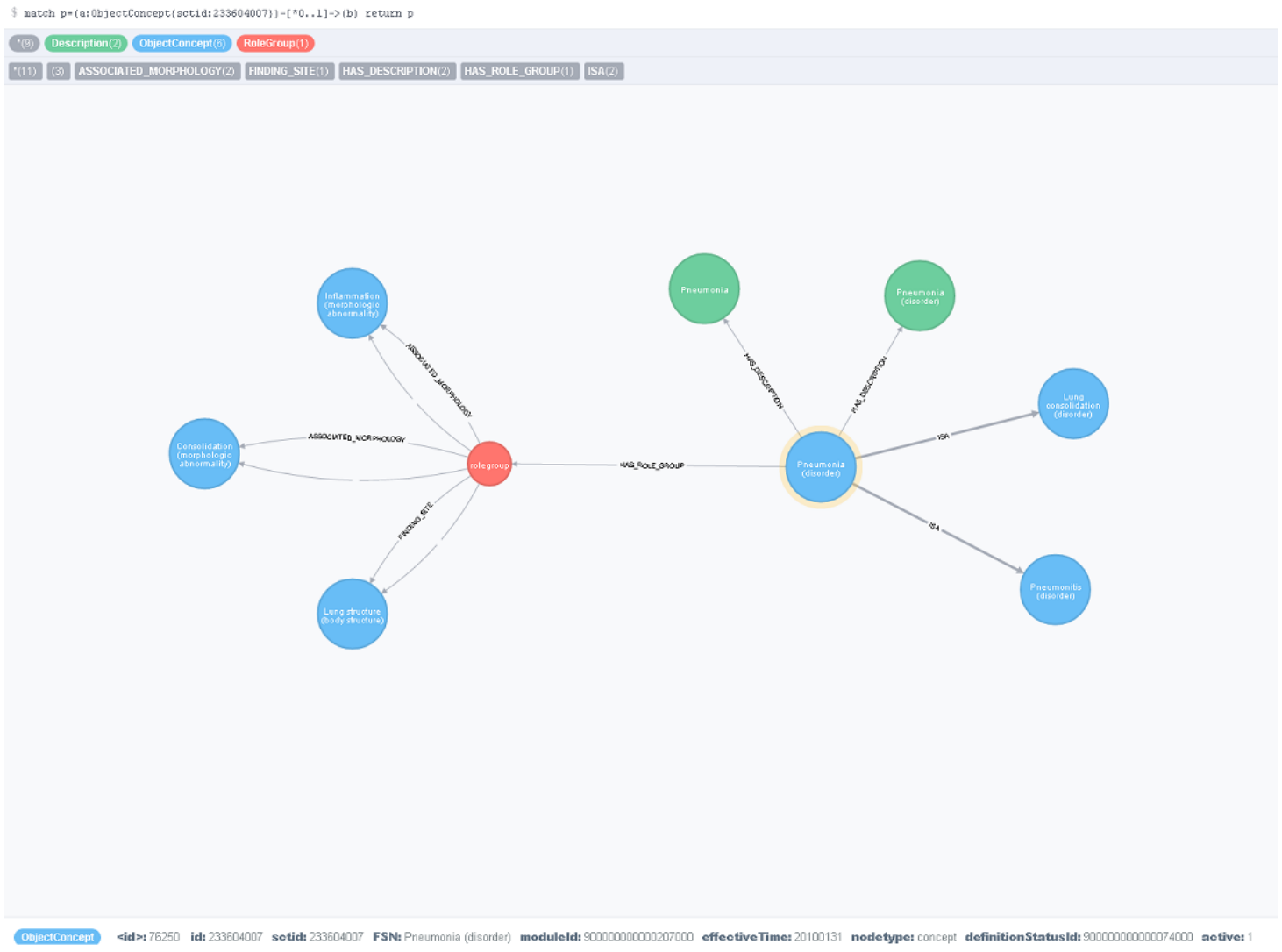
- Graph databases:
  - Class of NoSQL
  - Emphasize connectedness of data vs. rows/columns of data
  - Open world vs. closed world
    - Flexible
    - Transactionally ACID properties
  - SNOMED CT is a directed, acyclic graph

- 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

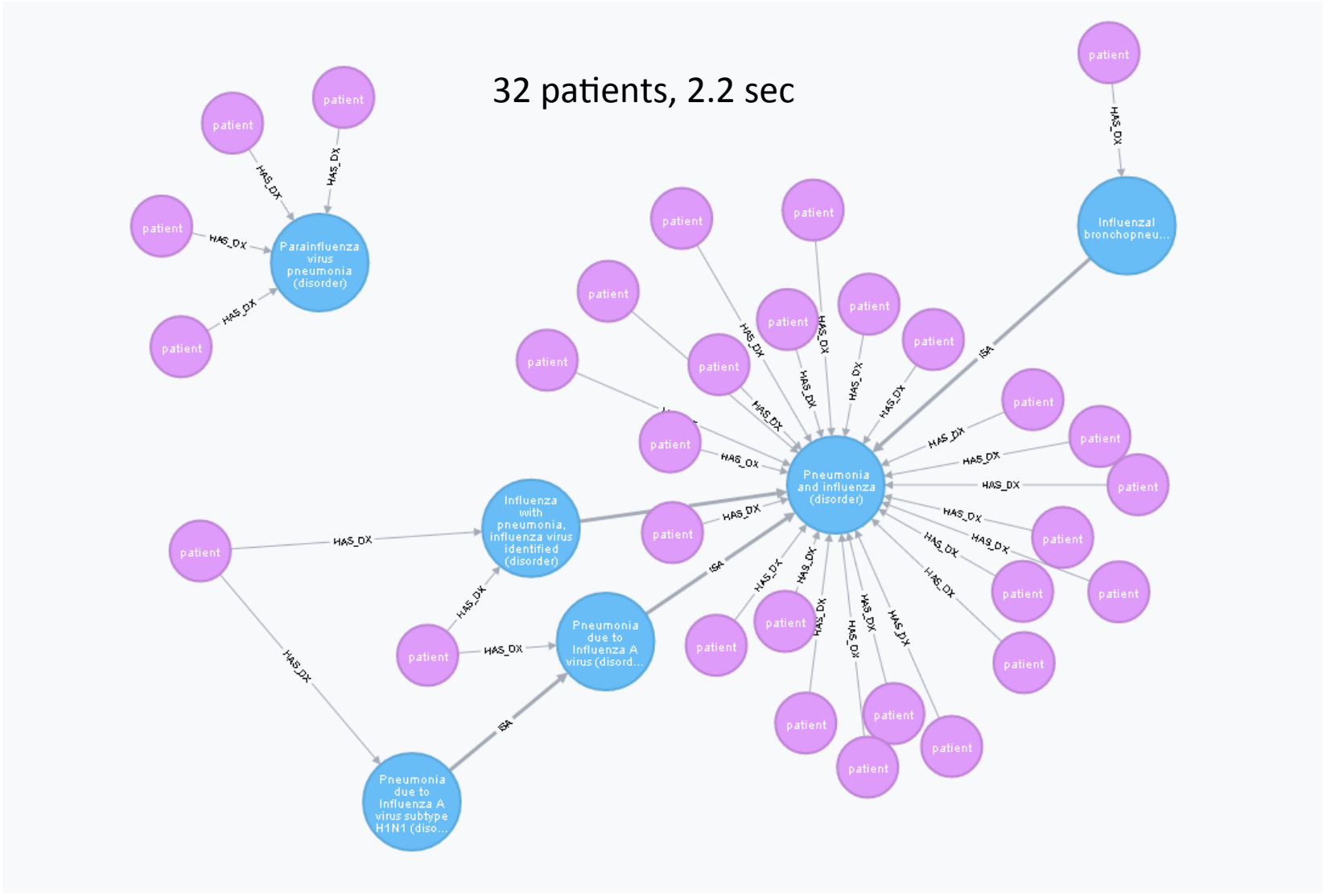
Example: Pneumonia



Add Patient Data

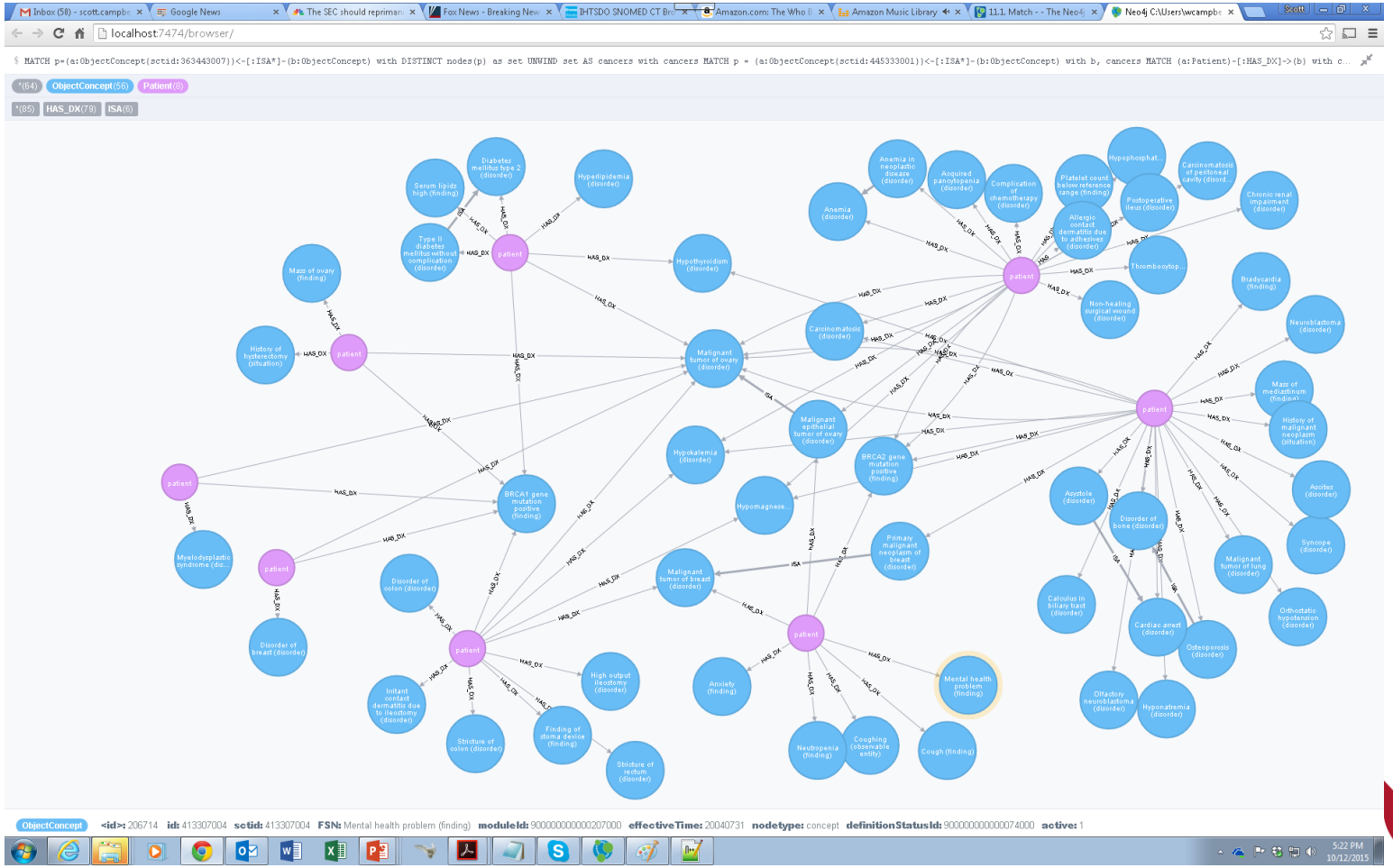
- Import patient records from de-identified clinical data warehouse
- Approximately 465,000 patients
- Import patient problem lists (All SNOMED CT encoded)
  - Up to 20 years of data
  - 2,770,000 diagnoses in total
  - Properties:
    - Date of diagnosis (start and end dates)
    - Active, inactive 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



Queries of undefined depth

- Find all patients with positive BRCA1 or BRCA2 gene mutation who have an ovarian cancer diagnosis.
- Return all shared diagnoses.



How about Historicity?

- Graph database calculation of transitive closure table – FAST

- David Markwell's challenge
    - Can the database produce TC tables for multiple years AND a Delta TC table between and two release dates?

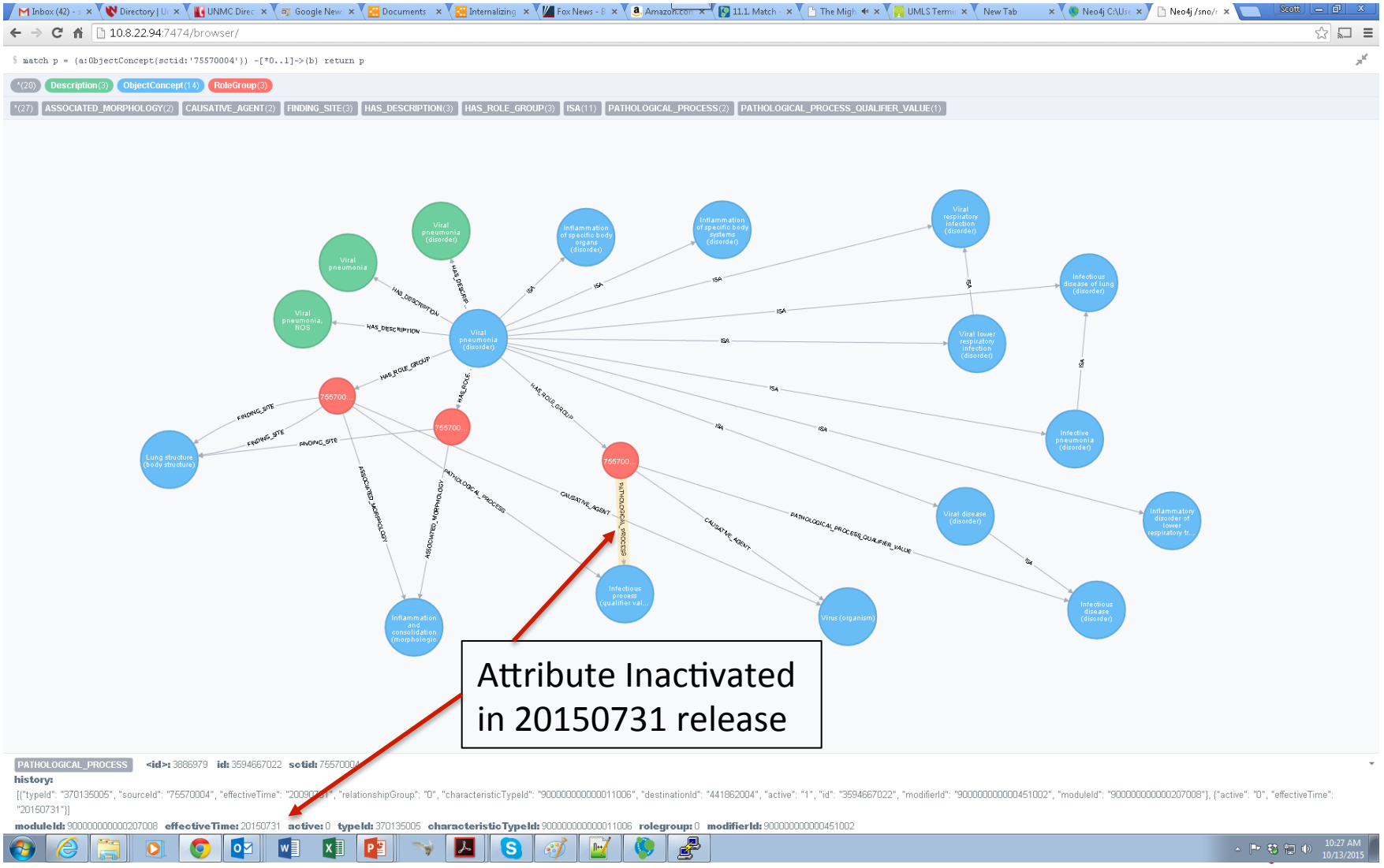
    - Beneficial for SNOMED CT sites to assess of effects of terminology updates on implementations

•Challenge accepted!

Challenge Met

- 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
- Creation of terminology update process with active patient data
  - Terminology updates “on the fly”

Example: Viral pneumonia



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

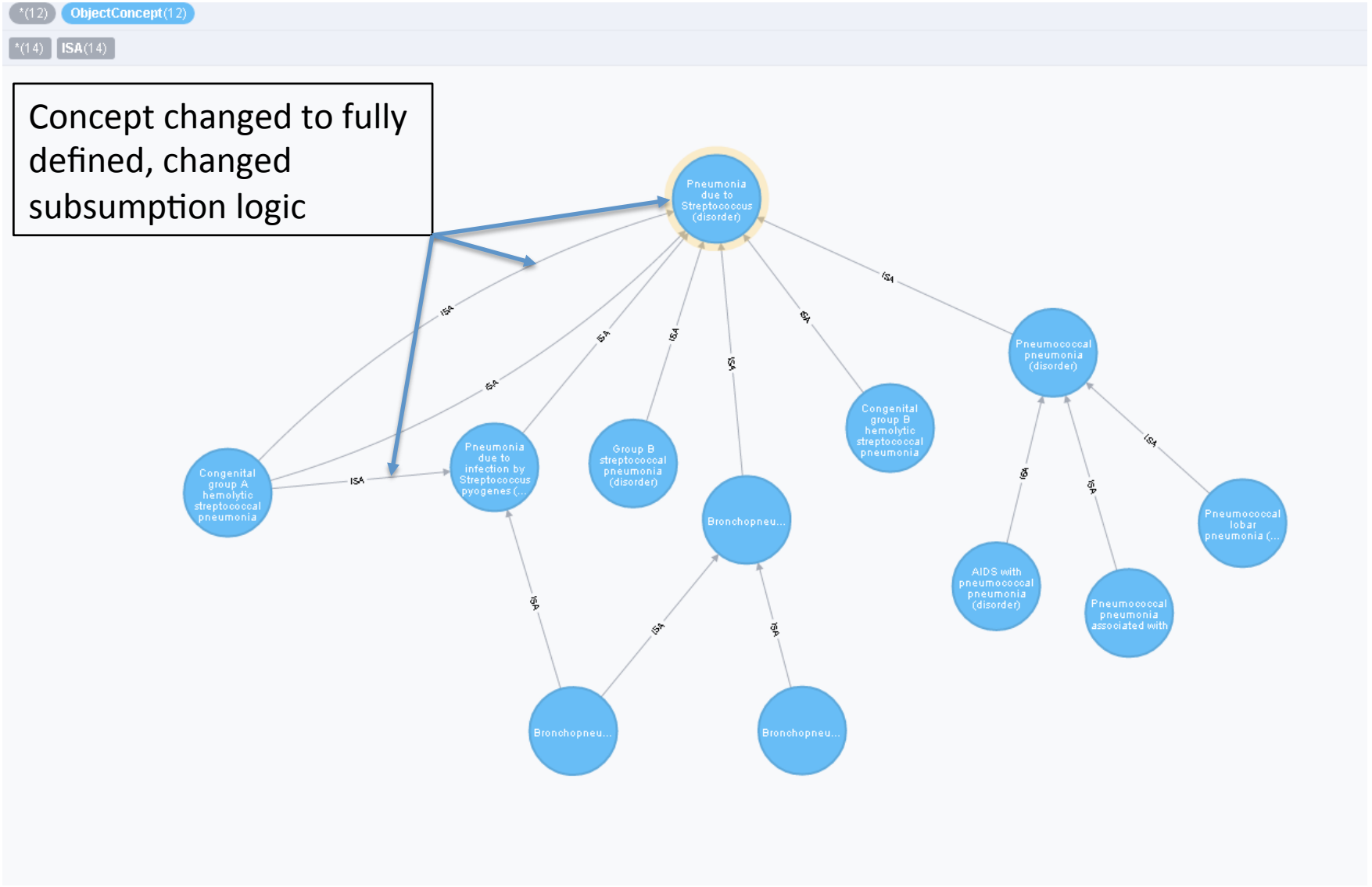
ID all patients with active diagnosis linked to inactive SNOMED CT concept

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 are the implications?

- Persistent and query-able representation of patient data over time in BOTH current and past SNOMED CT representations
- Quality metrics reporting based on any referenced point in time?
- Historical identification of patients with diagnoses that have been refined?
  - Perspective on medical knowledge/understanding
  - Epidemiologic study
- SNOMED CT quality reviews in real-time/impact on patient queries
  - Identification of changes and potential classifier issues in:
    - Concept deprecation
    - Subsumption changes
    - Defining relationships changes
    - Primitive vs. fully-defined

Example – Congenital B strep pneumonia



Where next?

- Dynamic environment to retain developing SNOMED CT content in anatomic pathology data (cancer synoptic data)
- Updated expressions accommodated as released and on demand
- Existing data with new SNOMED CT representation easily identified and modified.
  - No loss of historicity of data!
- Environment to test and prove out patient data in other domains
  - Microbiology?
  - Drug to bug to patient rapid identification?

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.

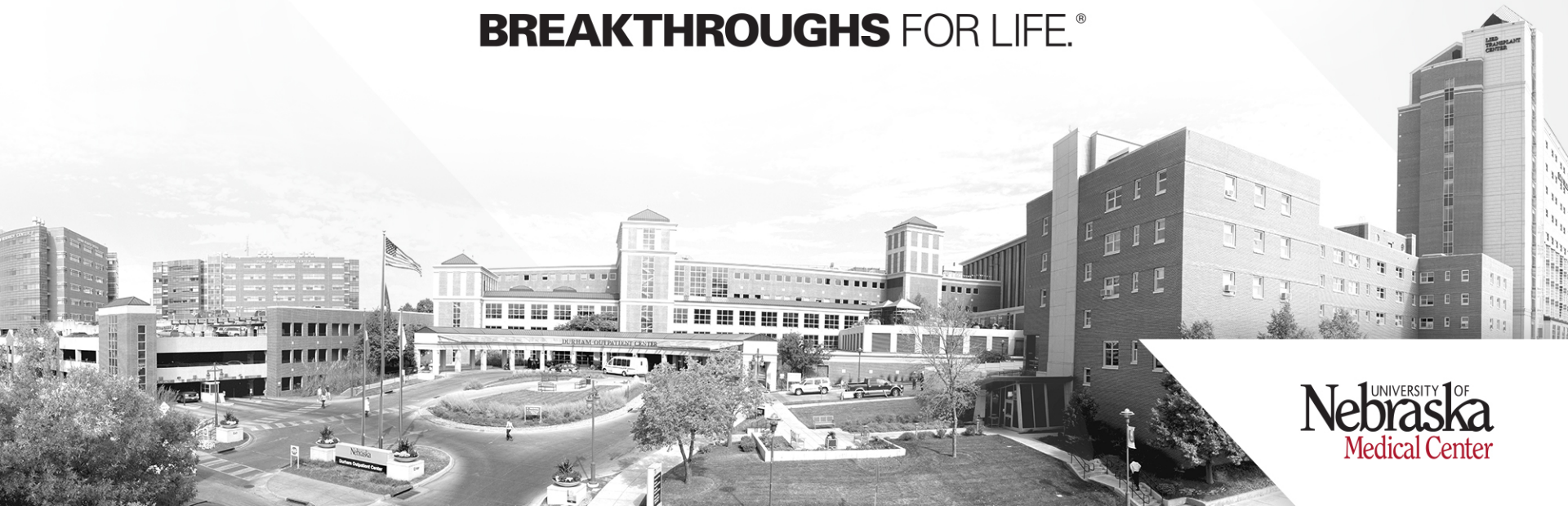
2. Campbell WS, Campbell JR, West WW, McClay JC, Hinrichs SH. Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings. J Am Med Inform Assoc. 2014 Sep;21(5):885-92.

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]*

University of Nebraska Medical Center

BREAKTHROUGHS FOR LIFE.®



University of Nebraska Medical Center