Department of Veterans Affairs
Veteran Health Administration
Knowledge Based Systems
Informatics Architecture Support Services

Creating a Framework for Extracting Unique Identifiers from Relevant Medical Text

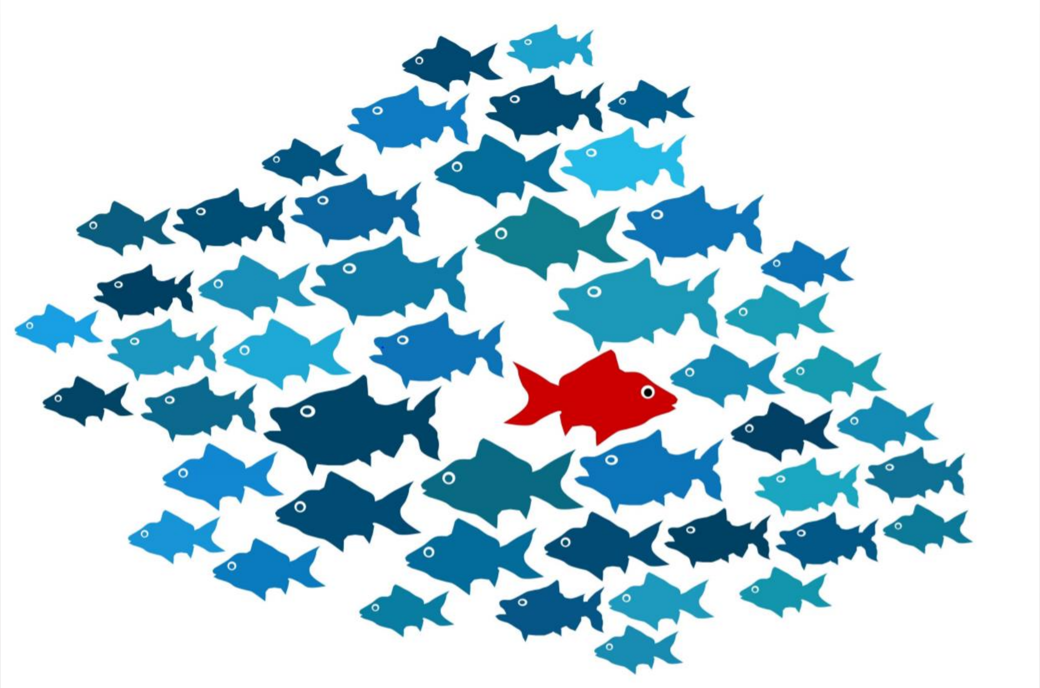
Sravan K. Elineni, Data Scientist, Lambda Squared

Veterans Health Administration, Office of Informatics & Analytics and Health Informatics
Informatics Architecture Support Services Contract VA701-16-C-0039
CLIN 0012, 5.4.1A. Bi Weekly Web Based Workshops

DATE: 11/21/2017

www.BookZurman.com

MACHINE LEARNING



- Machine learning is a field of science & engineering where computer systems adapt to learn from example data and past experience.
- Machine Learning is useful where human coders cannot explicitly code rules using a programming language.

image reference: http://www.tatvic.com/blog/wp-content/uploads/2017/01/fetured.jpg

Introduction(What?)

- Build a machine learning based framework to assist knowledge workers to add/remove or modify knowledge artifacts into SOLOR
- Subtask to accomplish above task is to extract relations, cause and effects, unique identifiers etc from general medical free form text such as clinical notes, medical journals, CMS quality frameworks etc

Why?

- Most of the medical relations and semantic networks can be uncovered from medical records.
- New relations can be uncovered with machine learning.
- Such extracted information can be used to uncover missing pieces of knowledge artifacts in SOLOR.

How?

- We use specific medical ontology to convert unstructured data and relations to structured data and relations.
- Later basic graph techniques like association rules algorithm can be used to create a recommendation engine with proposed modifications to SOLOR.
- Since the extracted information using NLP is from a pre defined ontology, any missing representation and relationships among medical diagnosis etc can be inserted into SOLOR using refsets 1.1-1.3

- Types of data extraction
  - Semantic similarity search
  - Approximate dictionary search
  - Entity extraction using statistical techniques

Semantic Similarity Search

- Semantic similarity Search uses contextual meaning to identify and match text to corresponding keywords.
- Needs a defined ontology to function.
- Categorize the results over semtypes by computing weighted links(uses hierarical and taxonomy depth to compute)

https://arxiv.org/ftp/arxiv/papers/1310/1310.8059.pdf

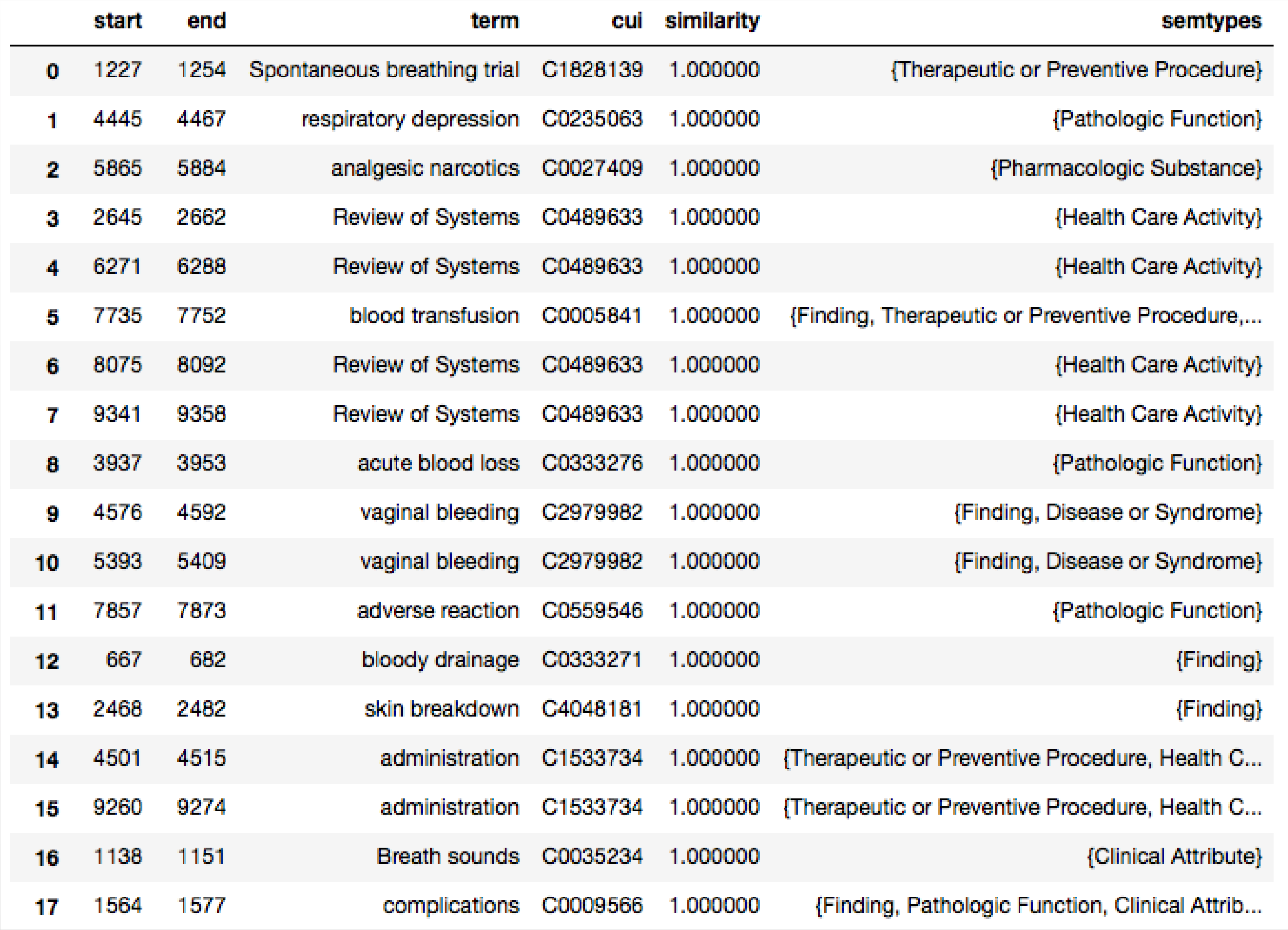
Sorry, no content matched your criteria
data data data data data data

https://ontotext.com/semantic-search-the-paradigm-shift-from-results-to-relationships/

Approximate dictionary Search

- Approximate dictionary match is a technique of find a string match in approximate proximity rather direct match.
- The central idea in using this technique is that the free form text (medical journals, provider notes etc) is matched to text from ontologies such as UMLS to their respective unique identifiers.
- Such identifiers are then grouped into the respective graphs of semtypes (refer next slide)
- For this study a package quickUMLS is used to extract semtypes and cui from free form text. Modifications to this package are required to match IDs in SOLOR system

•Luca Soldaini and Nazli Goharian. "*QuickUMLS: a fast, unsupervised approach for medical concept extraction.*" MedIR Workshop, SIGIR 2016.



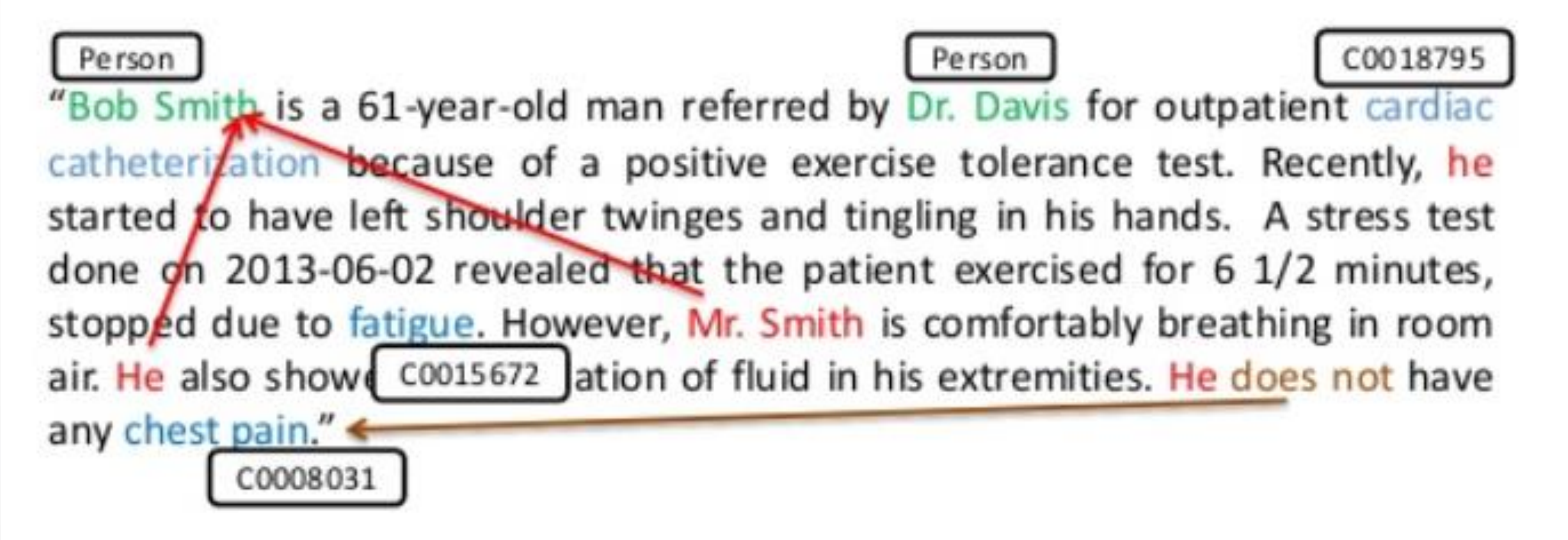
	start	end	term	cui	similarity	semtypes
0	1227	1254	Spontaneous breathing trial	C1828139	1.000000	{Therapeutic or Preventive Procedure}
1	4445	4467	respiratory depression	C0235063	1.000000	{Pathologic Function}
2	5865	5884	analgesic narcotics	C0027409	1.000000	{Pharmacologic Substance}
3	2645	2662	Review of Systems	C0489633	1.000000	{Health Care Activity}
4	6271	6288	Review of Systems	C0489633	1.000000	{Health Care Activity}
5	7735	7752	blood transfusion	C0005841	1.000000	{Finding, Therapeutic or Preventive Procedure,...
6	8075	8092	Review of Systems	C0489633	1.000000	{Health Care Activity}
7	9341	9358	Review of Systems	C0489633	1.000000	{Health Care Activity}
8	3937	3953	acute blood loss	C0333276	1.000000	{Pathologic Function}
9	4576	4592	vaginal bleeding	C2979982	1.000000	{Finding, Disease or Syndrome}
10	5393	5409	vaginal bleeding	C2979982	1.000000	{Finding, Disease or Syndrome}
11	7857	7873	adverse reaction	C0559546	1.000000	{Pathologic Function}
12	667	682	bloody drainage	C0333271	1.000000	{Finding}
13	2468	2482	skin breakdown	C4048181	1.000000	{Finding}
14	4501	4515	administration	C1533734	1.000000	{Therapeutic or Preventive Procedure, Health C...
15	9260	9274	administration	C1533734	1.000000	{Therapeutic or Preventive Procedure, Health C...
16	1138	1151	Breath sounds	C0035234	1.000000	{Clinical Attribute}
17	1564	1577	complications	C0009566	1.000000	{Finding, Pathologic Function, Clinical Attrib...

Entity Extraction

Entity extraction is a process of labelling free form text with known entity definitions and can be fast and scalable if Apache SPARK is used. Current open source platforms are very slow.
Entity extraction can be used in two ways

1) A set semtype is defined as an entity and the web of semtypes can be researched as a collective unit.

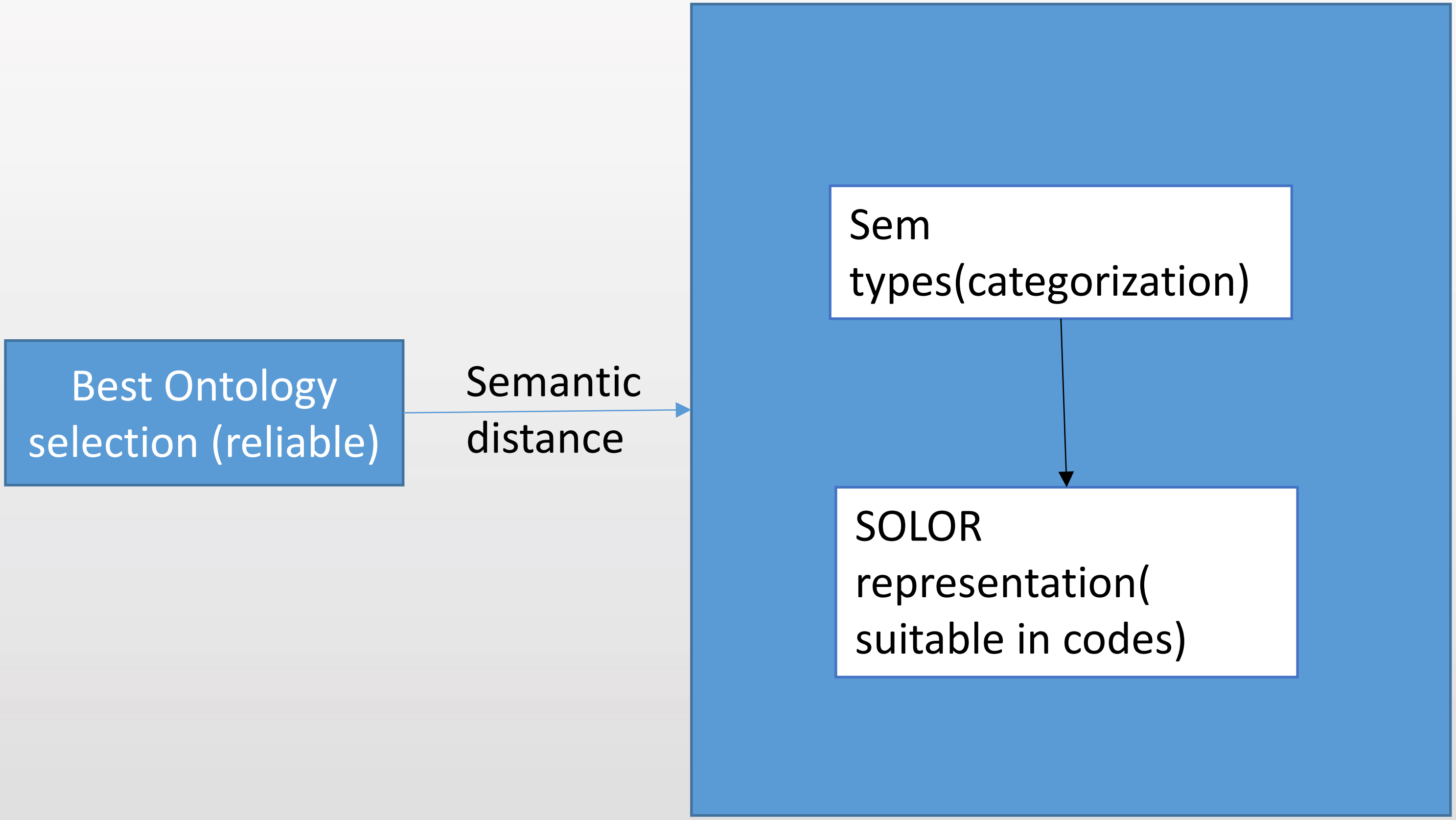
2) Alternatively ontologies such as UMLS metathesaurus can be used as key:value definitions and such keys can be assigned to free text using statistical extraction technique. A hybrid approach can also be utilised. Please see next slide



"Bob Smith is a 61-year-old man referred by Dr. Davis for outpatient cardiac catheterization because of a positive exercise tolerance test. Recently, he started to have left shoulder twinges and tingling in his hands. A stress test done on 2013-06-02 revealed that the patient exercised for 6 1/2 minutes, stopped due to fatigue. However, Mr. Smith is comfortably breathing in room air. He also showe ation of fluid in his extremities. He does not have any chest pain."

https://www.slideshare.net/sujanucsc/implicit-entity-recognition-in-clinical-documents-62614168

Process in detail



Best Ontology selection (reliable)

Semantic distance

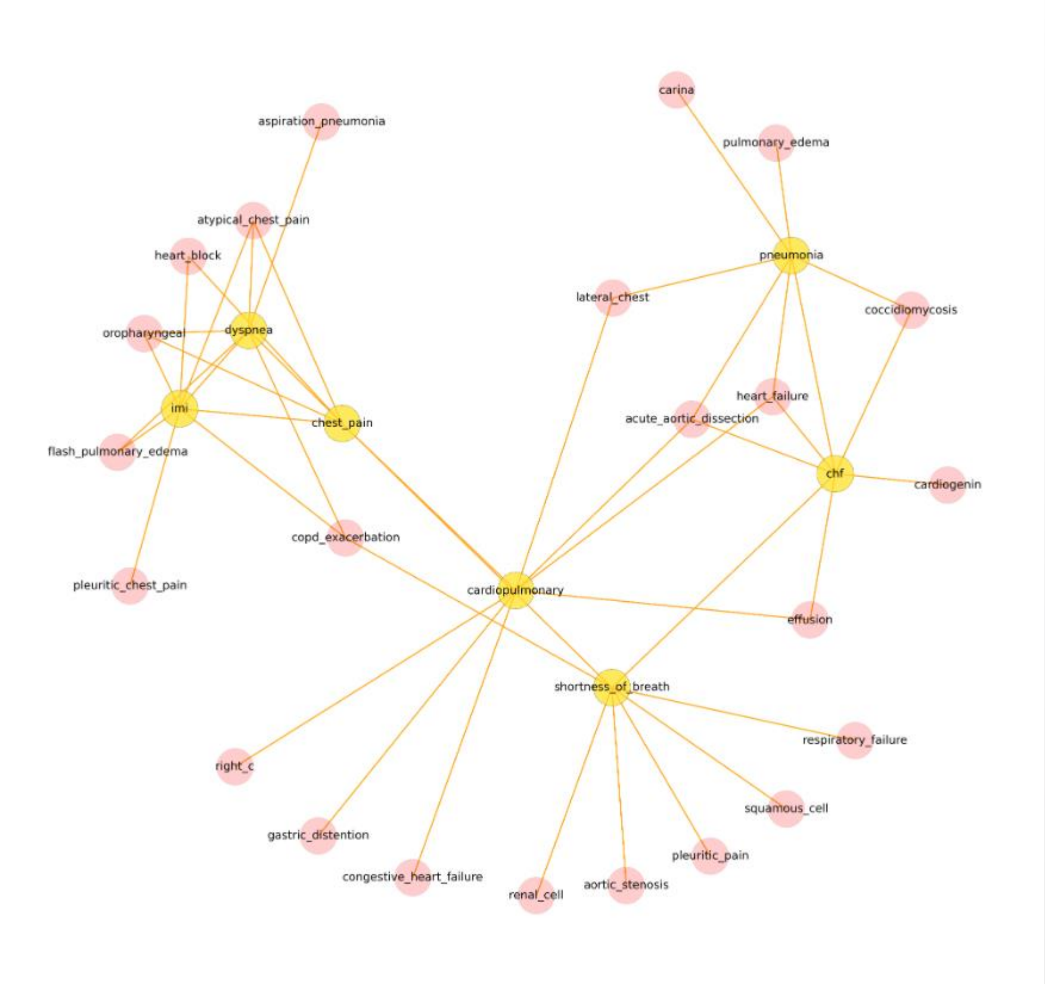
Sem types(categorization)

SOLOR representation(suitable in codes)

- Reliability of knowledge reference sets depends on the ontological selection.
- For knowledge workers to update SOLOR system, reliable information graphs should be presented.
- For example if a particular relation from entity extraction is identified and could not be mapped to any relation in SOLOR, a report has to be generated to be used by knowledge worker. Such report is powered by machine learning algorithm.

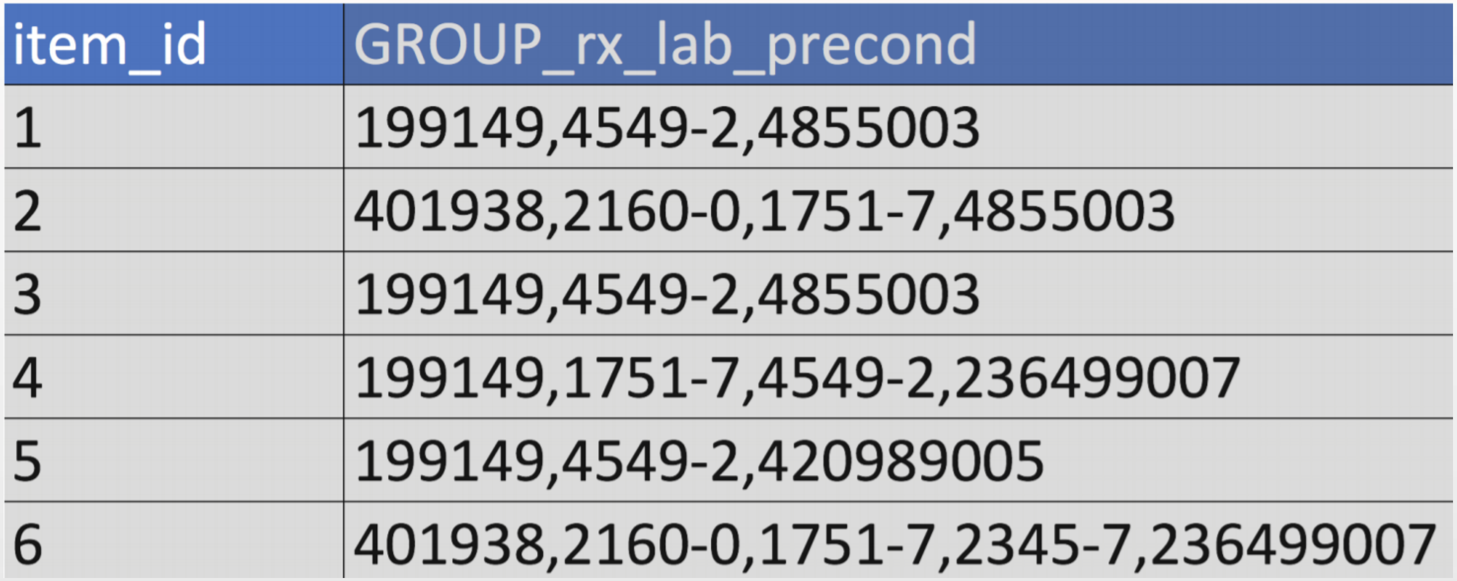
How can it be connected?

- Networkx package was used to demonstrate the possibility
- Each concept and semtypes are networked based on occurrence in text



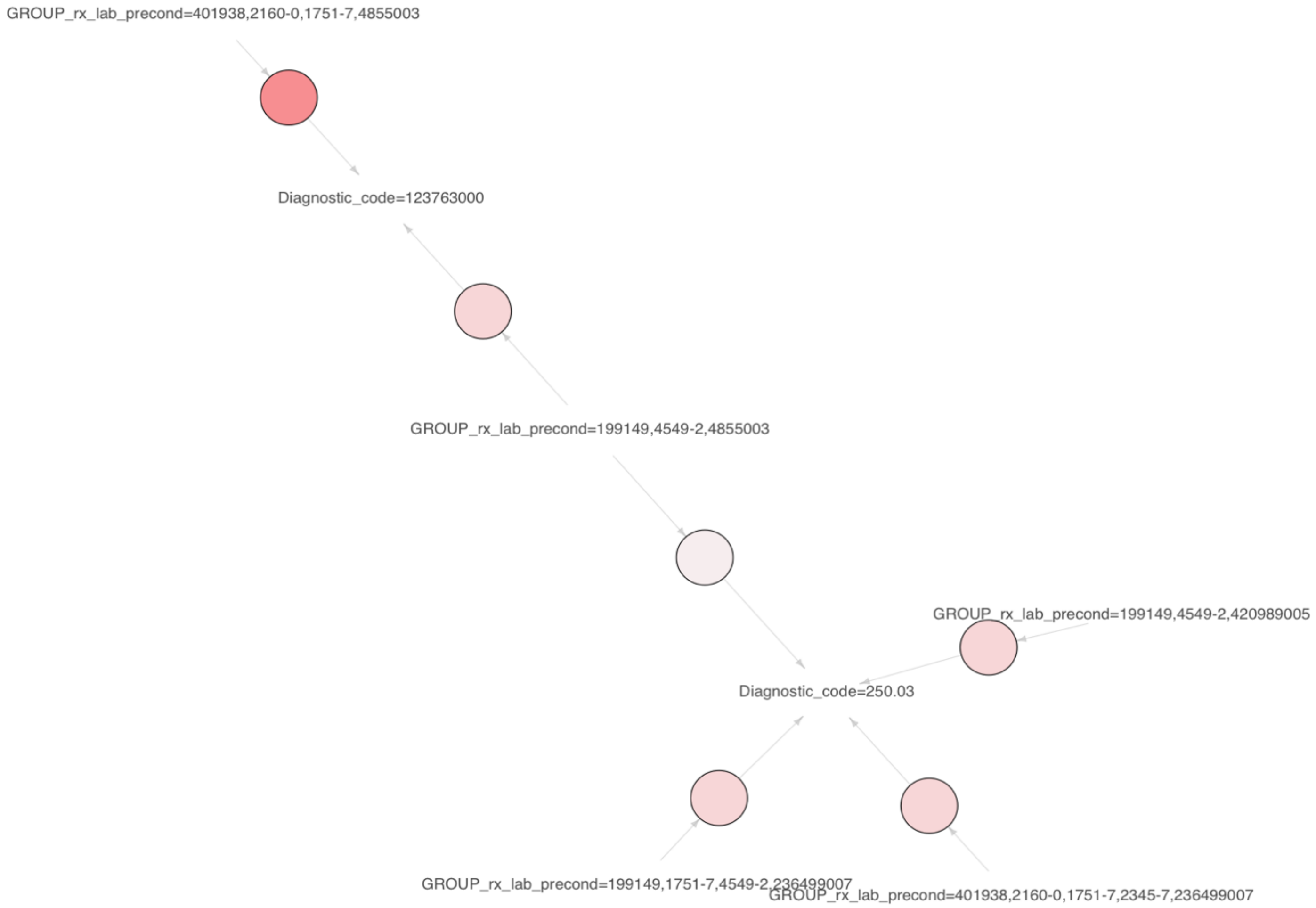
Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

ASSOCIATION RULES ALGORITHM



item_id	GROUP_rx_lab_precond
1	199149,4549-2,4855003
2	401938,2160-0,1751-7,4855003
3	199149,4549-2,4855003
4	199149,1751-7,4549-2,236499007
5	199149,4549-2,420989005
6	401938,2160-0,1751-7,2345-7,236499007

- Basket analysis: P(Y | X) probability that a lab X prescribed to patient also was prescribed medicine Y where X and Y are medical care related items
- Example: P(401938 | 2160-0) = 0.8



GROUP_rx_lab_precond=401938,2160-0,1751-7,4855003

Diagnostic_code=123763000

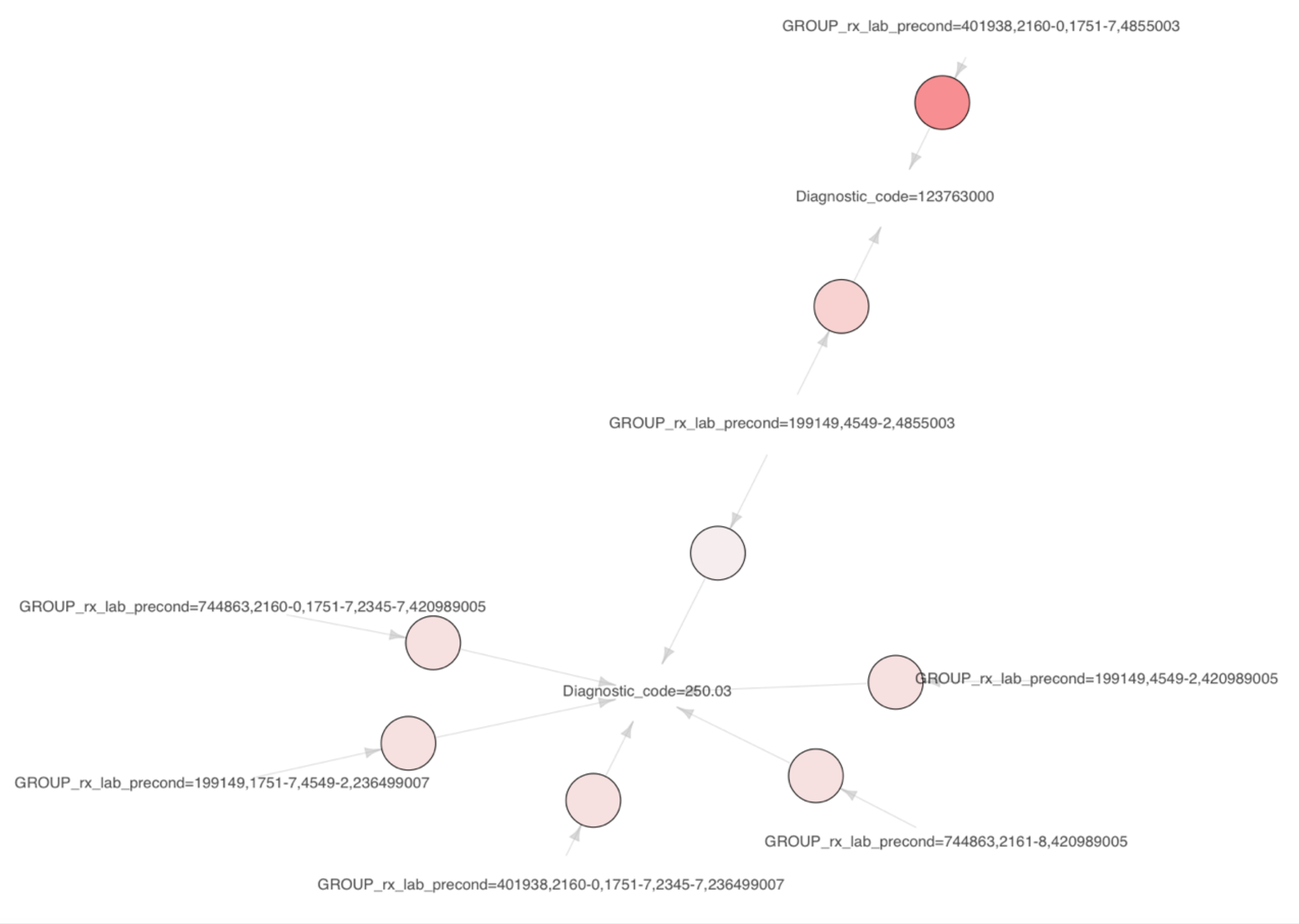
GROUP_rx_lab_precond=199149,4549-2,4855003

GROUP_rx_lab_precond=199149,4549-2,420989005

Diagnostic_code=250.03

GROUP_rx_lab_precond=199149,1751-7,4549-2,236499007

GROUP_rx_lab_precond=401938,2160-0,1751-7,2345-7,236499007



GROUP_rx_lab_precond=401938,2160-0,1751-7,4855003

Diagnostic_code=123763000

GROUP_rx_lab_precond=199149,4549-2,4855003

GROUP_rx_lab_precond=744863,2160-0,1751-7,2345-7,420989005

Diagnostic_code=250.03

GROUP_rx_lab_precond=199149,4549-2,420989005

GROUP_rx_lab_precond=199149,1751-7,4549-2,236499007

GROUP_rx_lab_precond=744863,2161-8,420989005

GROUP_rx_lab_precond=401938,2160-0,1751-7,2345-7,236499007

THANK YOU