NATURAL LANGUAGE UNDERSTANDING WITH MACHINE LEARNED ANNOTATORS & DEEP LEARNED ONTOLOGIES AT SCALE

David Talby
Ph.D., MBA, CTO @ Atigeo

SPARK SUMMIT EAST 2017

The problem

Who needs to be vaccinated?

Who is at risk for sepsis?

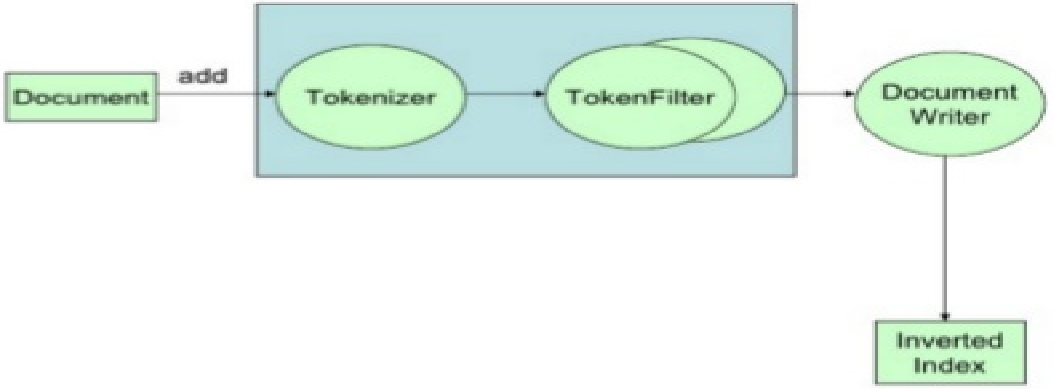
Who fits this clinical trial?

Who on this protocol did not have this side effect?

Who is getting meds they're allergic to?

At the beginning, there was search

- Query examples:
  - jazoon
  - jazoon AND java <=> +jazoon +java
  - jazoon OR java
  - jazoon NOT php <=> jazoon -php
  - conference AND (java OR j2ee)
  - “Java conference“
  - title:jazoon
  - j?zoon
  - jaz*
  - schmidt~ schmidt, schmit, schmitt
  - price:[000 TO 050]



Scalable & robust Indexing pipeline
Tokenizers & analyzers
Synonyms, spellers & Auto-suggest
File formats & header boosting
Rankers, link & reputation boosting

Then there was semantic search

"cheap red prom dresses"
"laptops under $500"
"italian restaurants near me that deliver"
"captain america civil war tonight"
"nba scores"

Dictionary Based Attribute Extraction

Dell - XPS 15.6 4K Ultra HD Touch-Screen Laptop - Intel Core i5 - 8GB Memory - 256GB Solid State Drive - Silver

Machine Learned Attribute Extraction

If you go for the ambience, you'll be disappointed. If you go for good, inexpensive and authentic Mexican food, then you're in the right place.

Then, you need to understand language

Prescribing sick days due to diagnosis of influenza.	*Positive*
Jane complains about flu-like symptoms.	*Speculative*
Jane may be experiencing some sort of flu episode.	*Possible*
Jane’s RIDT came back negative for influenza.	*Negative*
Jane is at high risk for flu if she’s not vaccinated.	*Conditional*
Jane’s older brother had the flu last month.	*Family history*
Jane had a severe case of flu last year.	*Patient history*

1.

Language gets complex and domain specific

Human language is wonderfully nuanced

Joe expressed concerns about the risks of bird flu.	*Nothing*
Joe shows no signs of stroke, except for numbness.	*Double Negative*
Nausea, vomiting and ankle swelling negative.	*Compound*

Patient denies alcohol abuse.	*Speculative*
Allergies: Penicillin, Dust, Sneezing.	*Compound*

(it gets worse – in reality a lot of text isn't valid English)

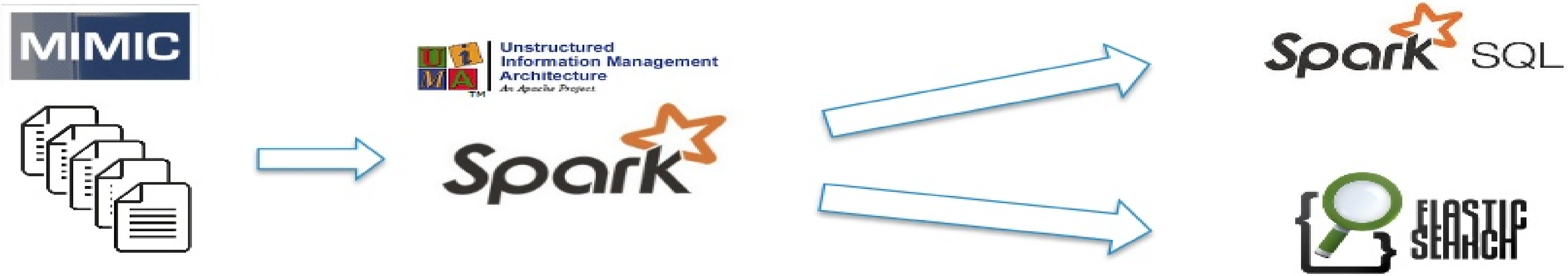
Let’s build this!

The input (patient records)

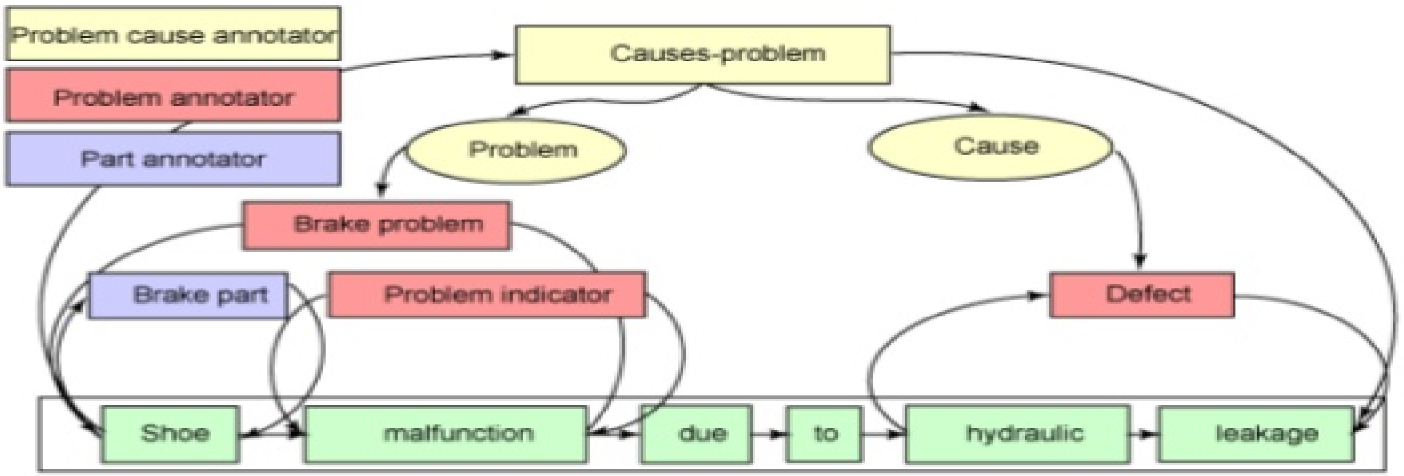
The processing framework

The output

The query engines



Unstructured Information Management Architecture
An Apache Project



SENTENCE DETECTION | TOKENIZER | LEMMATIZER

SECTION DETECTION | CONDITIONAL SCOPE

SPECULATIVE SCOPE | NEGATION DETECTION

DATE | NUMBER | UNIT | QUANITITY

STOPWORD REMOVAL | CONCEPT EXTRACTION

File Edit View Insert Cell Kernel Help

Heading 3 Cell Toolbar: None

Convert the table produced by XLP in Parquet format

In [2]:
```
hc.sql('DROP TABLE strata.mimic2_p')
hc.sql('SET spark.sql.parquet.cacheMetadata = true')
hc.sql('SET spark.sql.parquet.compression.codec = snappy')
hc.sql('CREATE TABLE strata.mimic2_p STORED AS PARQUET AS SELECT * FROM strata.annotations_mimic2_strata')
hc.sql('USE strata')
#hc.sql('CACHE TABLE mimic2_p')
```

Out[2]:
```
DataFrame[result: string]
```

View the data coming out of XLP

In [3]:
```
df = d.execute_query('SELECT * FROM strata.mimic2_p LIMIT 100')
df.head(10)
```

Out[3]:

	mimic2_p.noteid	mimic2_p.annotationtype	mimic2_p.featurejson
0	hdfs://10.0.2.85:8020/user/ubuntu/datasets/dem...	DocumentAnnotation	{"language":"x-unspecified","begin":0,"end":26...
1	hdfs://10.0.2.85:8020/user/ubuntu/datasets/dem...	Note	{"doctorName":null,"patientName":null,"visitId...
2	hdfs://10.0.2.85:8020/user/ubuntu/datasets/dem...	Section	{"normalizedSectionName":"NURSING_NOTE","origi...
3	hdfs://10.0.2.85:8020/user/ubuntu/datasets/dem...	Sentence	{"begin":0,"end":112}
4	hdfs://10.0.2.85:8020/user/ubuntu/datasets/dem...	SectionHeader	{"normalizedSectionType":"NURSING_NOTE","begin...
5	hdfs://10.0.2.85:8020/user/ubuntu/datasets/dem...	Token	{"begin":0,"end":11}
6	hdfs://10.0.2.85:8020/user/ubuntu/datasets/dem...	Assertion	{"id":"0","originalSection":"Neonatology - NNP...
7	hdfs://10.0.2.85:8020/user/ubuntu/datasets/dem...	NormalizedToken	{"normalizedText":"neonatology","begin":0,"end...
8	hdfs://10.0.2.85:8020/user/ubuntu/datasets/dem...	ConceptAnnotation	{"cui":"C0027621","codes":"U003151","sources":...
9	hdfs://10.0.2.85:8020/user/ubuntu/datasets/dem...	Token	{"begin":12,"end":13}

Out of all annotations, select the ones where annotatioType = Assertion (annotations with various atributes)

In [4]:
```
hc.sql('DROP TABLE strata.tmp_assertions')
hc.sql("CREATE TABLE strata.tmp_assertions STORED AS PARQUET AS \
       SELECT noteid, \
```

File Edit View Insert Cell Kernel Help

Heading 3 · Cell Toolbar: None

Analyze the values of various attributes of the Assertions extracted out of 100 MIMIC II notes

In [ ]:
```
df = d.execute_query('SELECT polarity,count(*) as cnt FROM strata.tmp_assertions group by polarity order by cnt desc')
df.head(30)
```

Out[7]:

	polarity	cnt
0	POSITIVE	461266
1	SPECULATIVE	24281
2	POSSIBLE	5192
3	CONDITIONAL	2914

In [8]:
```
df = d.execute_query('SELECT subject,count(*) as cnt FROM strata.tmp_assertions group by subject order by cnt desc')
df.head(30)
```

Out[8]:

	subject	cnt
0	PATIENT	387766
1	PRESENT_CONDITION	56677
2	PATIENT_HISTORY	32228
3	HISTORY_PRESENT_CONDITION	12922
4	FAMILY_HISTORY	2188
5	OTHER	1572

File Edit View Insert Cell Kernel Help

Heading 3 · Cell Toolbar: None

In [ ]:
```
df = d.execute_query('SELECT conceptType,count(*) as cnt FROM strata.tmp_assertions group by conceptType order by cnt desc')
df.head(20)
```

Out[9]:

	concepttype	cnt
0	CONCEPT	151204
1	NON_MEDICAL	139147
2	BODY_PART	32447
3	ABNORMALITY	16170
4	ANATOMICAL	16014
5	SOCIAL	15545
6	LAB	14550
7	CHEMICAL	12538
8	DEVICE	12517
9	PREVENTIVE_PROCEDURE	11239
10	DRUG	10739
11	PROCEDURAL	9718
12	DISEASE	9575
13	BIOLOGIC_FUNCTION	9088
14	BODY_SYSTEM	8587
15	DIAGNOSTIC_PROCEDURE	7183
16	PSYCHOLOGICAL	6961
17	PHENOMENON	3012
18	INJURY	2338
19	DISCIPLINE	2110

2.

you’ll need

machine learning early

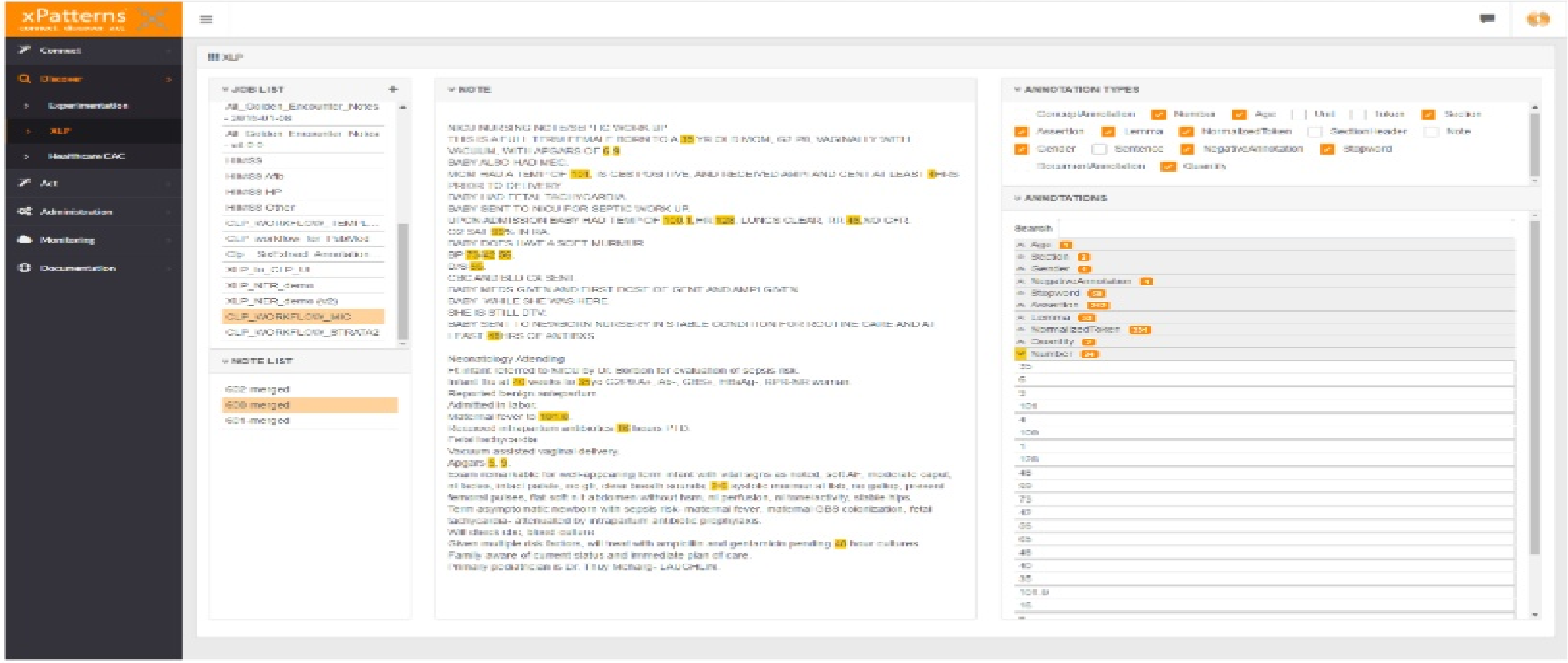
Machine learned annotators

Sometimes, it's easier to just code an annotation's business logic

Grammatical Patterns	Direct Inferences	Lookups
If … then …	Age < 18 ==> Child	RIDT (lab test)

But sometimes it's easier to learn it from examples:

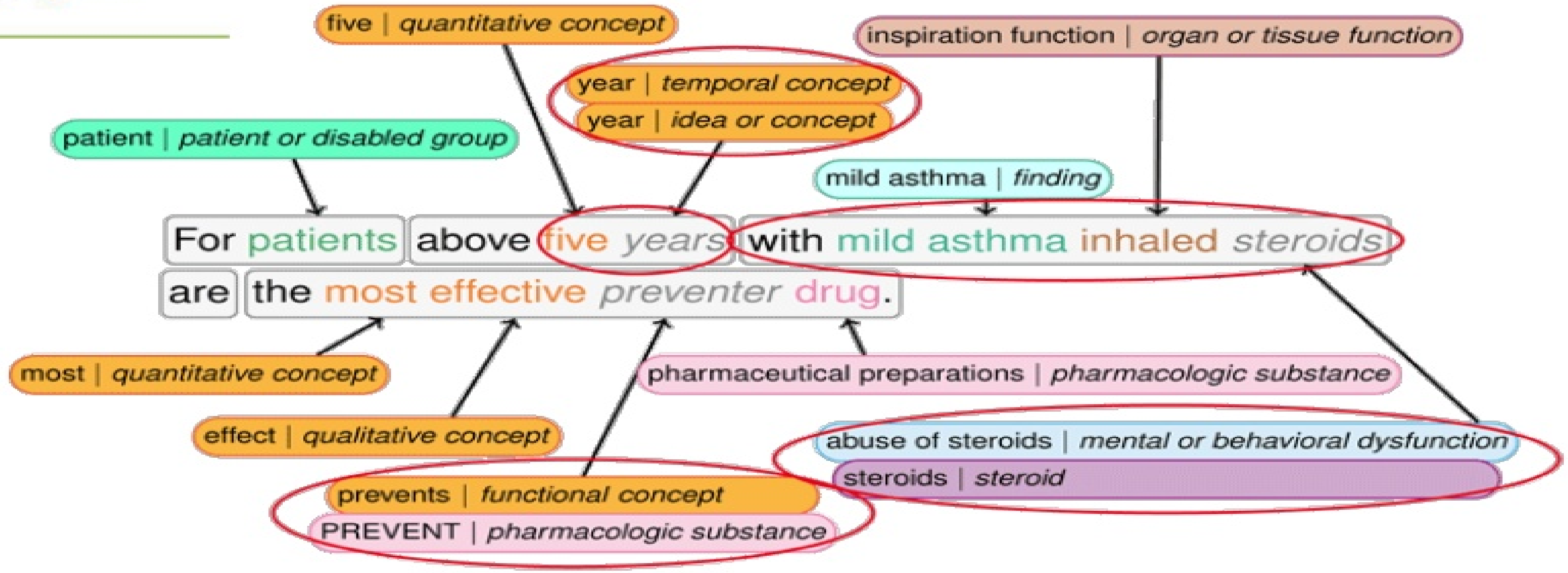
Under-diagnosed conditions	Implied by Context
Flu Depression	relevant labs normal


xPatterns
connect. discover. act.
Connect
Discover
Experimentation
XLP
Healthcare CAC
Act
Administration
Monitoring
Documentation
JOB LIST
NOTE LIST
NOTE
ANNOTATION TYPES
ANNOTATIONS

3.

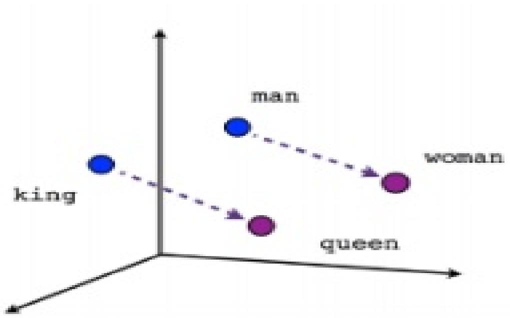
bootstrap and then expand your vocabulary

Unified Medical Language System

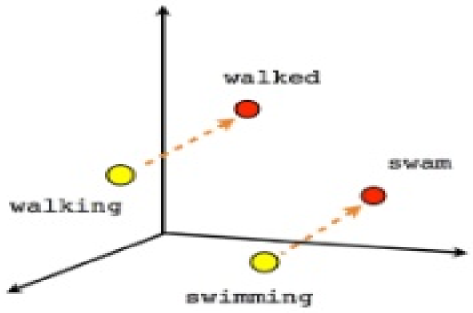


Expanding & updating ontologies

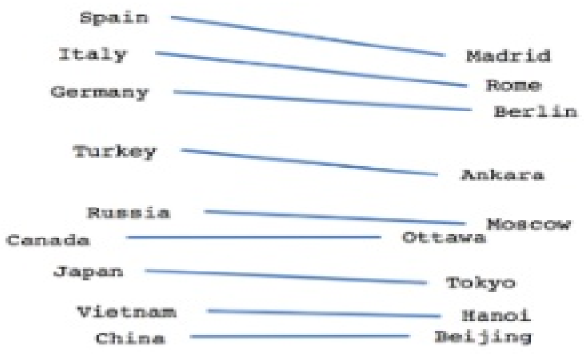
$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=-k}^{j=k}\log p(w_{t+j}|w_t) \qquad p(w_i|w_j) = \frac{\exp(u_{w_i}^{\top} v_{w_j})}{\sum_{l=1}^{V}\exp(u_l^{\top} v_{w_j})}$$



Male-Female



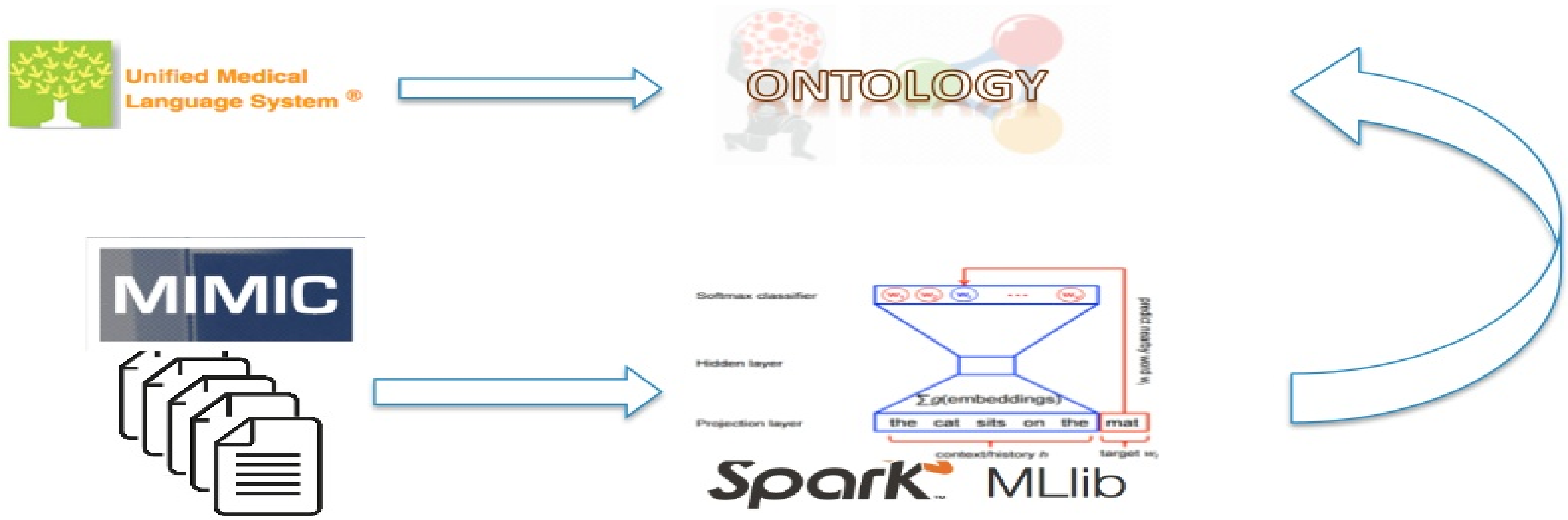
Verb tense

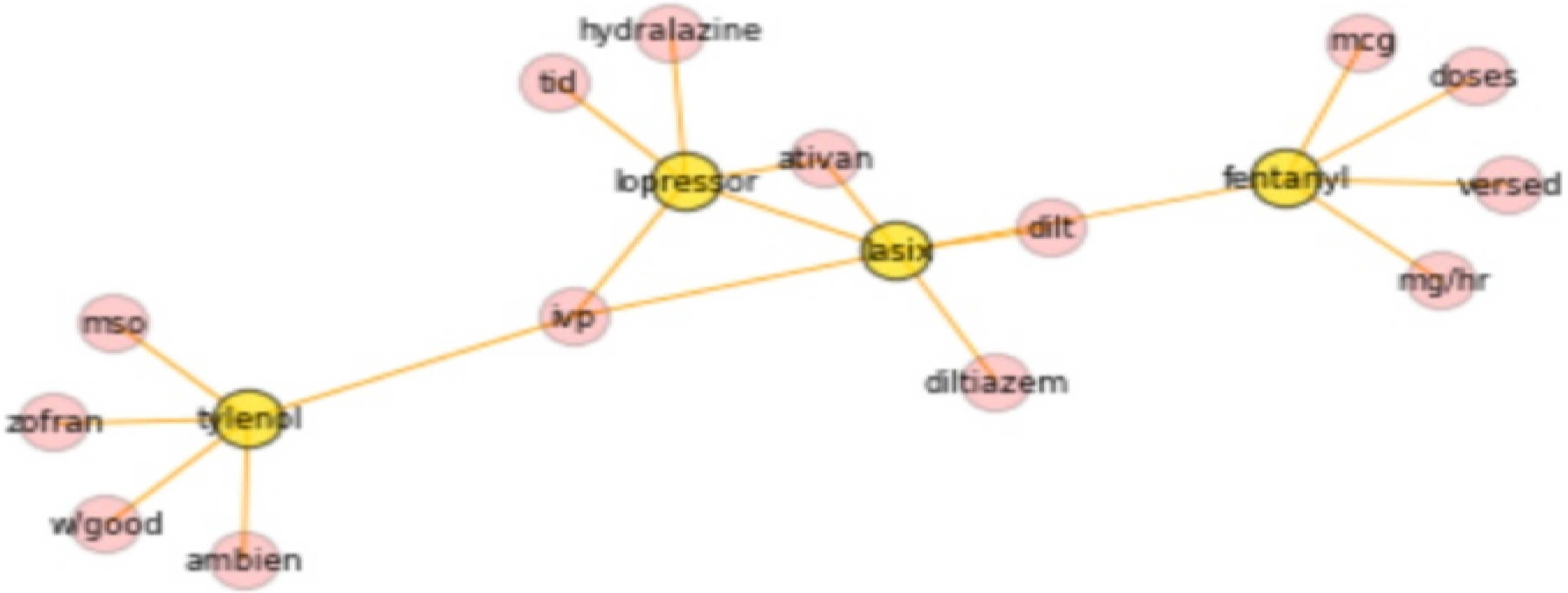


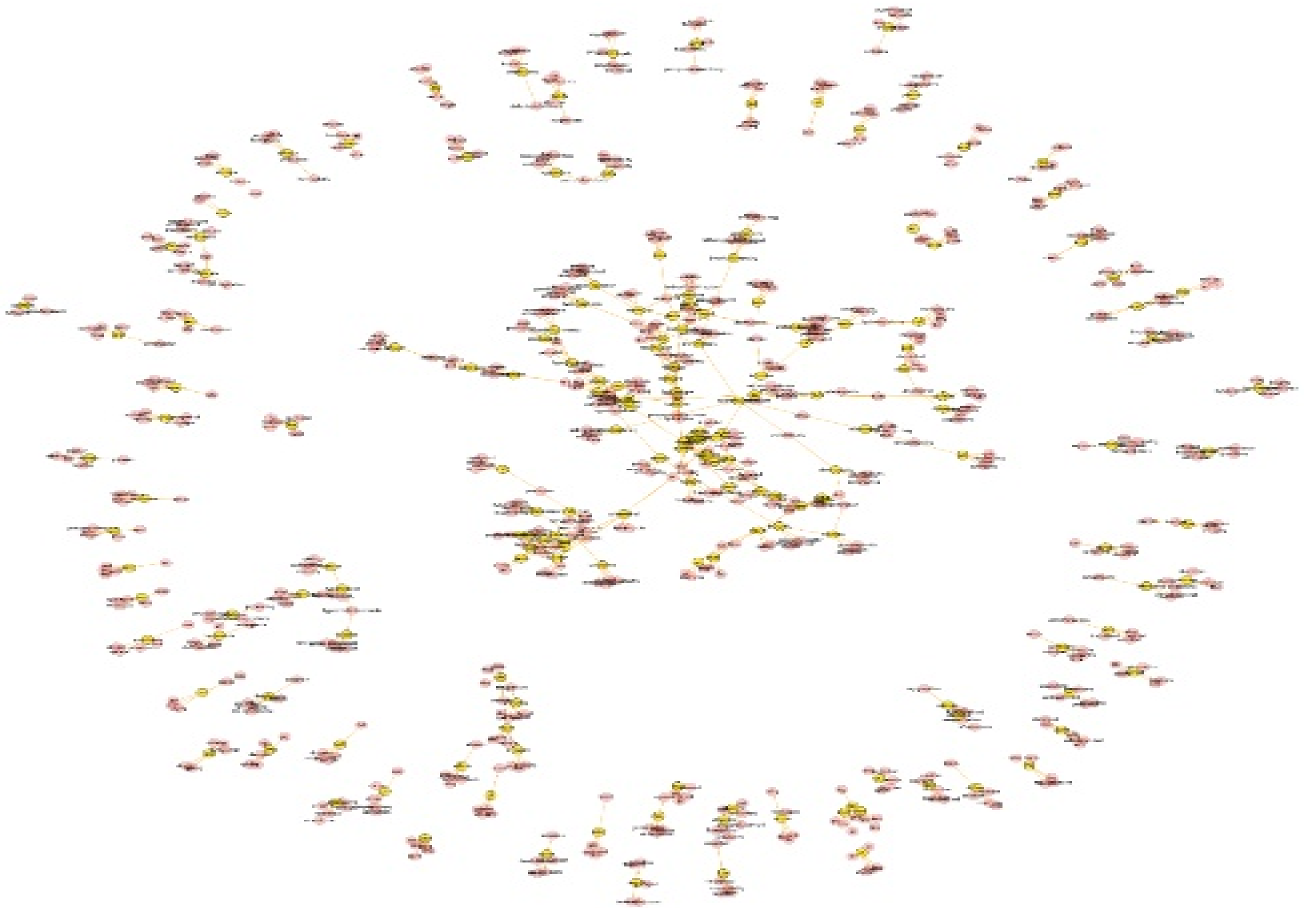
Country-Capital

Word2Vec

Let's build this too!




hydralazine
tid
lopressor
ativan
lasix
dilt
fentanyl
mcg
doses
versed
mg/hr
diltiazem
ivp
mso
zofran
tylenol
w/good
ambien



Summary: How

1. Language gets complex and domain specific
2. You'll need machine learning early
3. Bootstrap & then expand your vocabulary

Summary: Why

Who is at risk for sepsis?

Who needs to be vaccinated?

Who fits this clinical trial?

Thank You!

github.com/atigeo/nlp_demo

@davidtalby