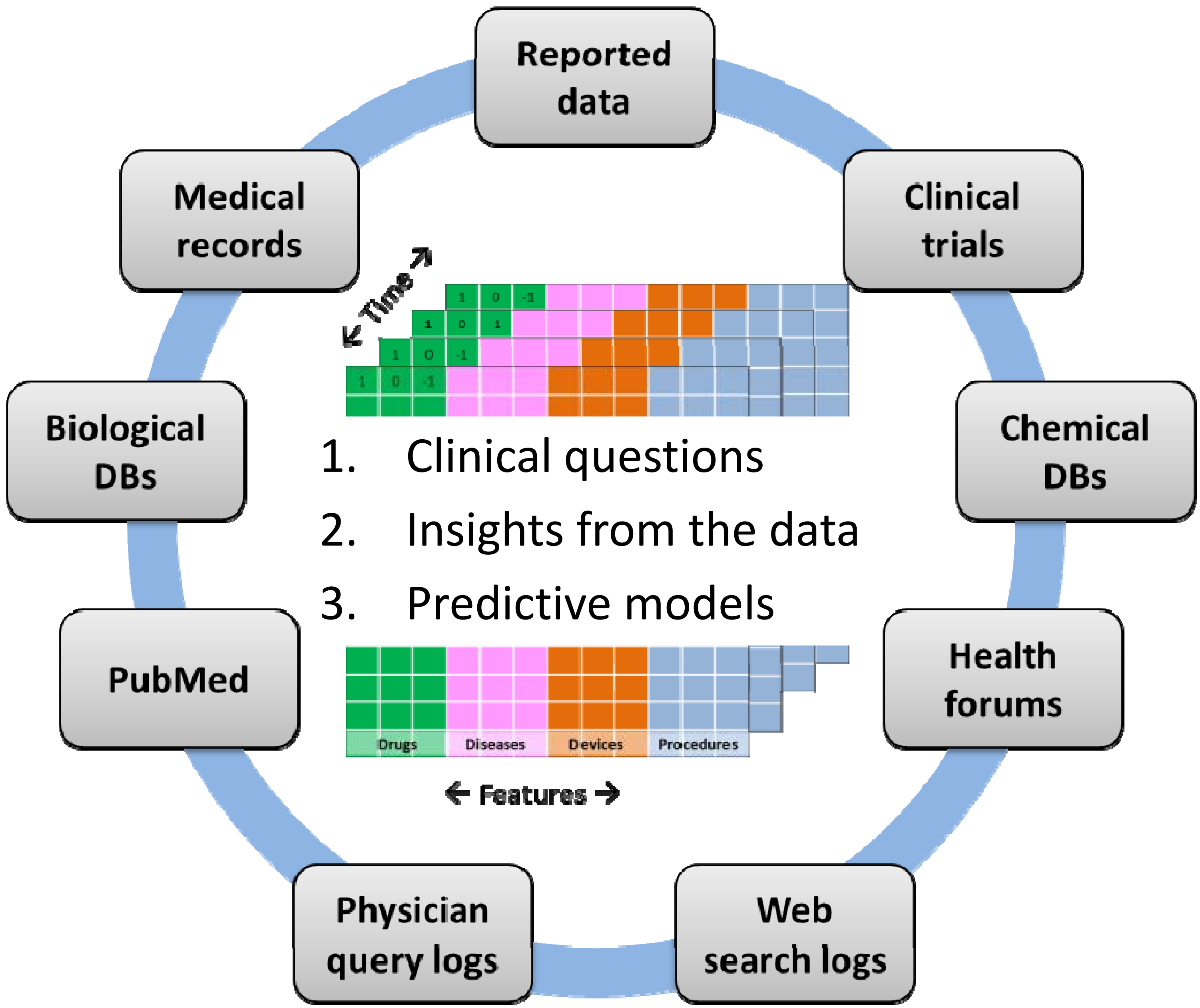
Mining the EHR to understand diseases, drugs, and adverse events

Nigam Shah
nigam@stanford.edu

STANFORD
SCHOOL OF MEDICINE



Reported data
Clinical trials
Chemical DBs
Health forums
Web search logs
Physician query logs
PubMed
Biological DBs
Medical records
Time
1. Clinical questions
2. Insights from the data
3. Predictive models
Drugs
Diseases
Devices
Procedures
Features

Structured vs. Unstructured Data

row ID	patient ID	hospital ID	order	ICD9 code
1960	156	199280	12	2930
2627	214	197273	7	2930
2764	223	105694	17	2930
5902	502	116367	15	2930
5957	505	116719	1	2930
6463	546	127873	14	2930
6807	588	170452	16	2930
8475	735	140547	5	2930
9379	807	121760	13	2930
10811	923	151107	5	2930
12122	1040	118695	6	2930
12135	1041	159277	13	2930
12140	1042	130732	5	2930
15121	1332	161256	10	2930
14487	1269	172465	6	2930
14526	1271	151184	13	2930
13630	1178	132240	7	2930
14976	1321	139634	3	2930
16311	1439	120530	13	2930
16414	1446	150030	11	2930
15526	1354	144830	5	2930
17424	1551	131207	7	2930
17478	1559	147179	5	2930
29983	2658	126865	5	2930
30936	2750	120236	9	2930
31006	2759	107939	4	2930
31982	2842	124867	25	2930
30087	2667	149534	10	2930
18125	1610	145005	16	2930
19005	1701	167723	9	2930
18272	1621	102458	22	2930
33490	3002	185653	20	2930
33699	3024	139229	14	2930

~25%

IMPRESSION (ACC 6075491) :

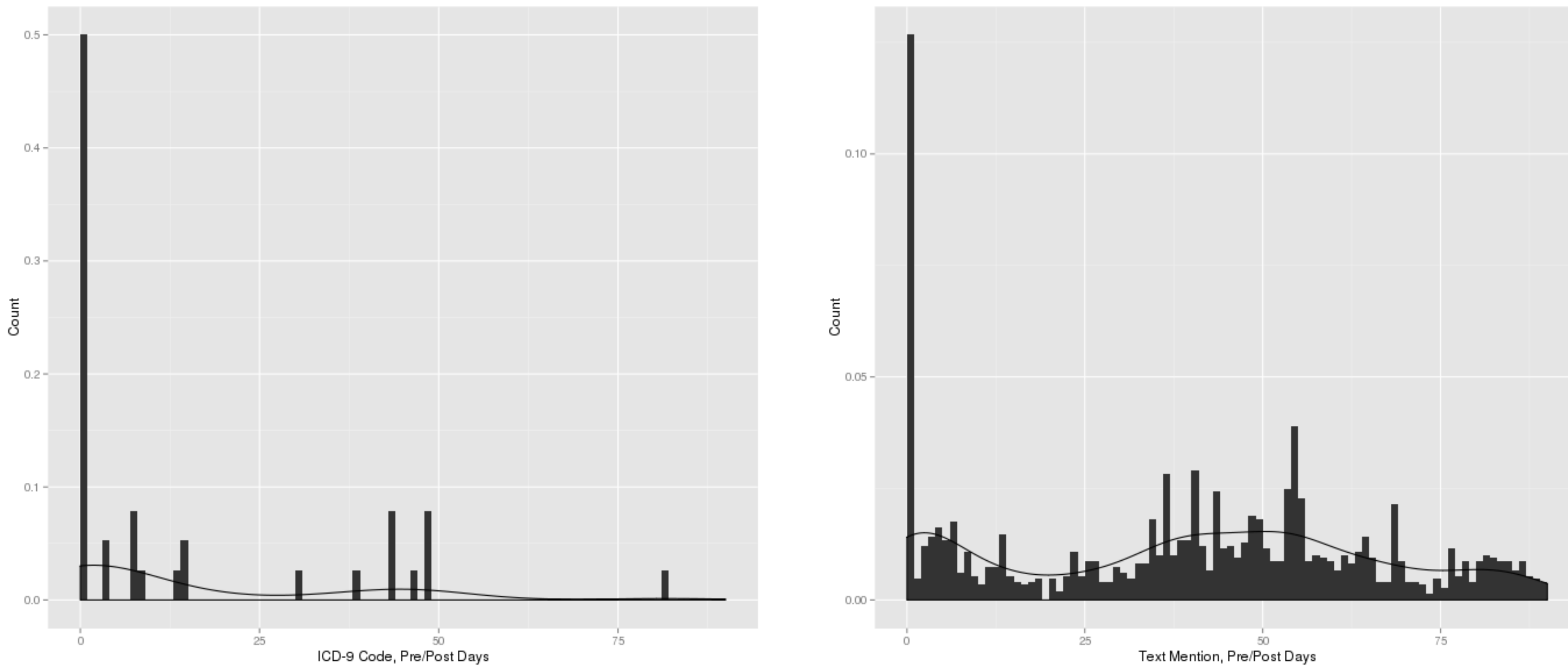
addendum beginsexam association only. addendum endsbilateral diagnostic digital mammogram with computer-aided detection 3/31/2011 8:14 amright axillary ultrasound 3/31/2011 8:14 am indication: female, 73 years old, right breast lateral tenderness, no discrete mass. history:post-menopausal patient. comparison: 3/7/2006 (stanford hospital), 7/24/2009 (advanced medicine center) technique: full-field digital mammograms were obtained with computer-aided detection to assist in interpretation of the study, including bilateral craniocaudal and mediolateral oblique views coma with an additional right lateral view. real-time breast ultrasound was then performed targeted to

findings: mammogram: the breast tissue is largely fatty. there is a skin bb marker over a palpable abnormality in the right axillary region. there are no features to suggest malignancy. ultrasound: targeted ultrasound reveals a normal appearing lymph node in the 11 o'clock position 10 cm from the nipple in the right axillary region 9x 6 x 4 mm. otherwise no discrete solid or cystic masses identified.

impression: 1. right breast: bi-rads 1, negative. left breast: bi-rads 1, negative. recommend the finding prompting ultrasound should be followed on a clinical basis alone. assuming clinical stability, recommend annual screening mammography.

~75%

The utility of looking into notes

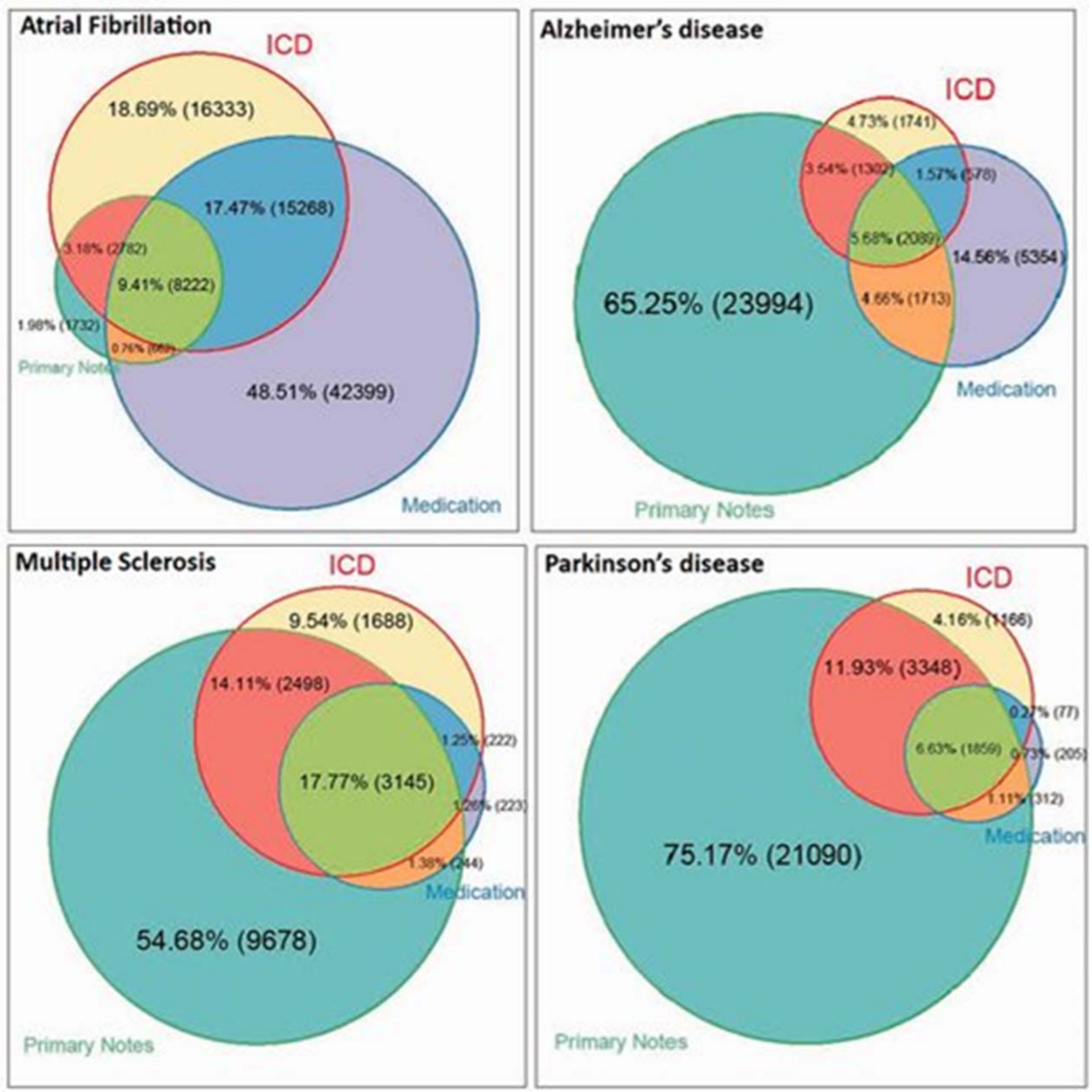


TYPE OF EHR INFORMATION	POSITIVE DOCUMENTATION OF URINARY INCONTINENCE	NEGATIVE DOCUMENTATION OF URINARY INCONTINENCE*	ABSENCE OF DOCUMENTATION OF URINARY INCONTINENCE
Text	450	1035	3868
ICD-9	4	n/a	5349

Note: *Negative Documentation refers to patients reporting that they are not suffering from urinary incontinence

http://repository.edm-forum.org/egems/vol4/iss3/1/

The utility of looking into notes

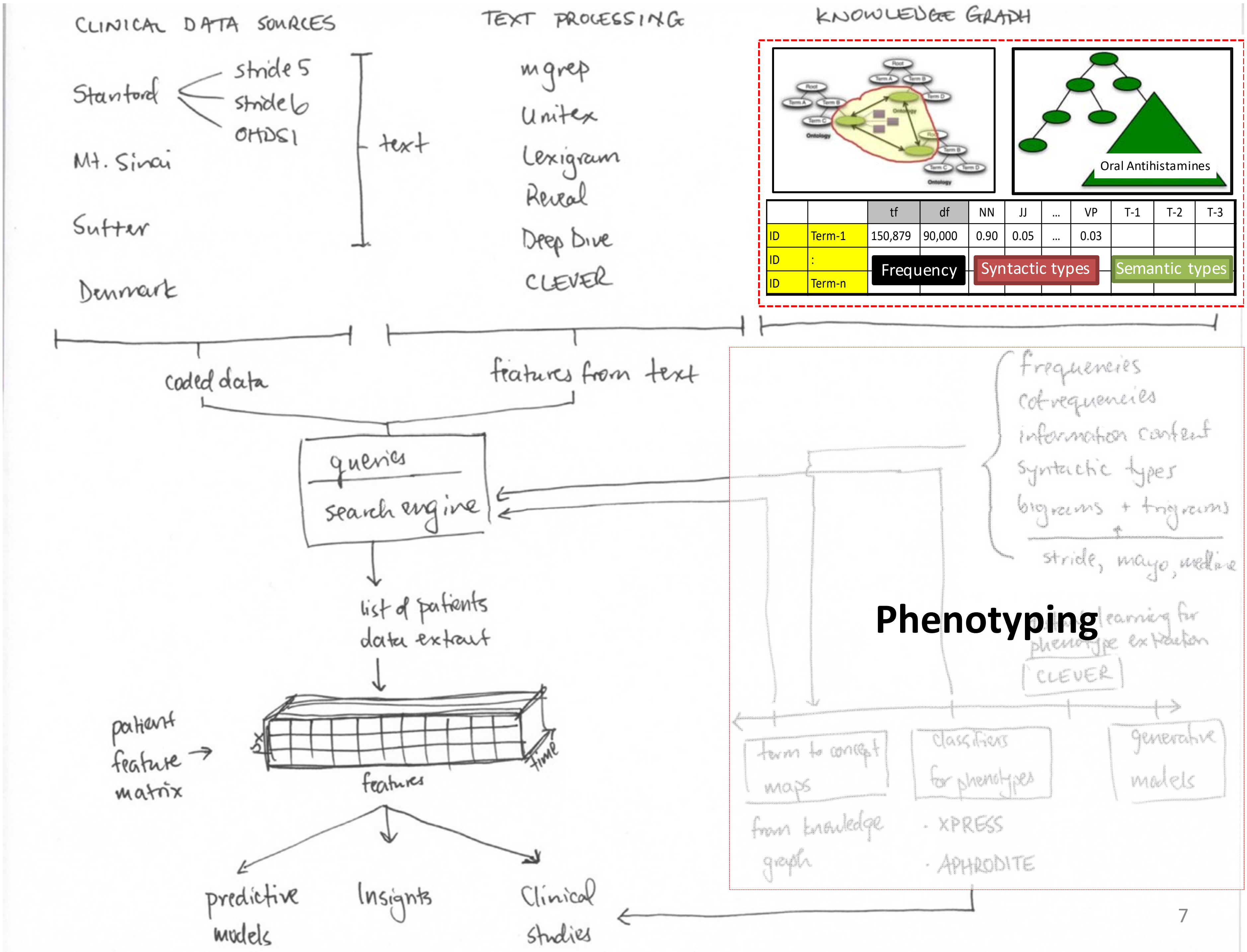


Wei WQ, et al 2015 JAMIA

NLP or Text-mining

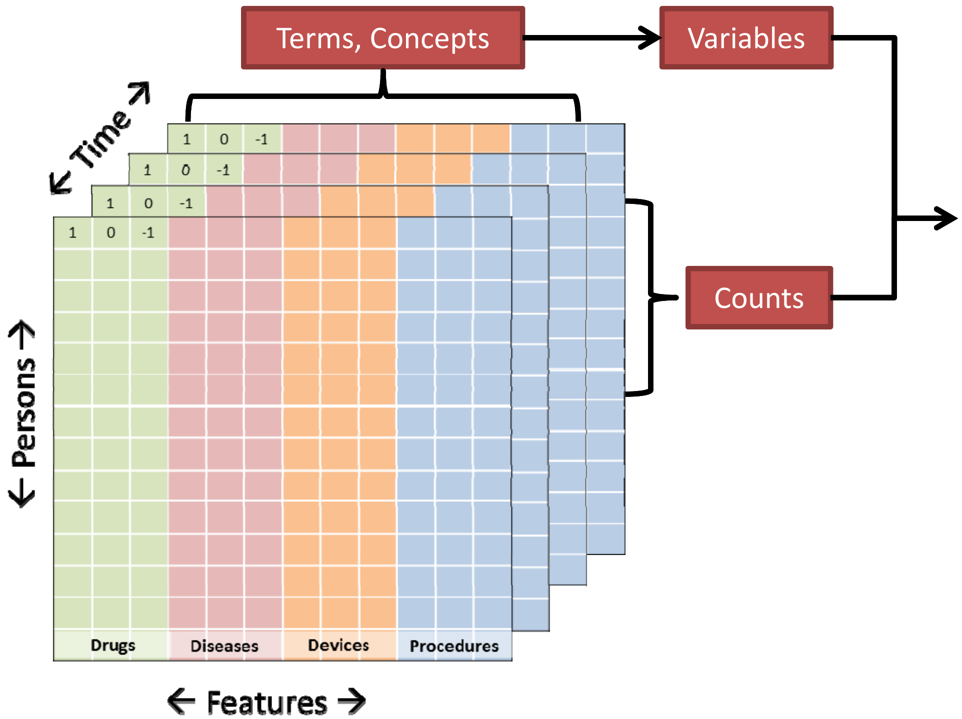
Natural language processing (NLP) is a discipline which attempts to understand human (natural) languages using computers

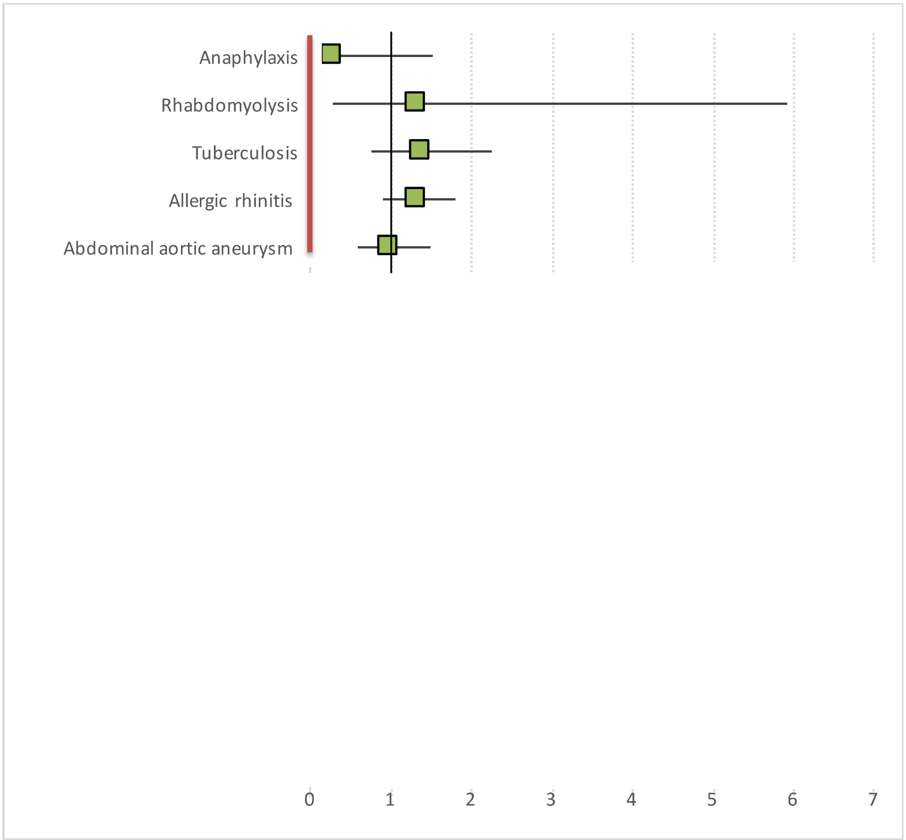
Text mining is the process of discovering and extracting knowledge from unstructured data


CLINICAL DATA SOURCES
TEXT PROCESSING
KNOWLEDGE GRAPH
Stanford
stride 5
stride 6
OHDSI
Mt. Sinai
Sutter
Denmark
text
mgrep
Unitex
Lexigram
Reveal
Deep Dive
CLEVER
Oral Antihistamines
tf
df
NN
JJ
...
VP
T-1
T-2
T-3
ID
Term-1
150,879
90,000
0.90
0.05
...
0.03
ID
:
ID
Term-n
Frequency
Syntactic types
Semantic types
coded data
features from text
queries
search engine
list of patients
data extract
patient feature matrix
features
time
predictive models
Insights
Clinical Studies
frequencies
co-frequencies
information content
Syntactic types
bigrams + trigrams
stride, mayo, medline
Phenotyping
learning for phenotype extraction
CLEVER
term to concept maps
from knowledge graph
classifiers for phenotypes
· XPRESS
· APHRODITE
generative models

Clinical questions

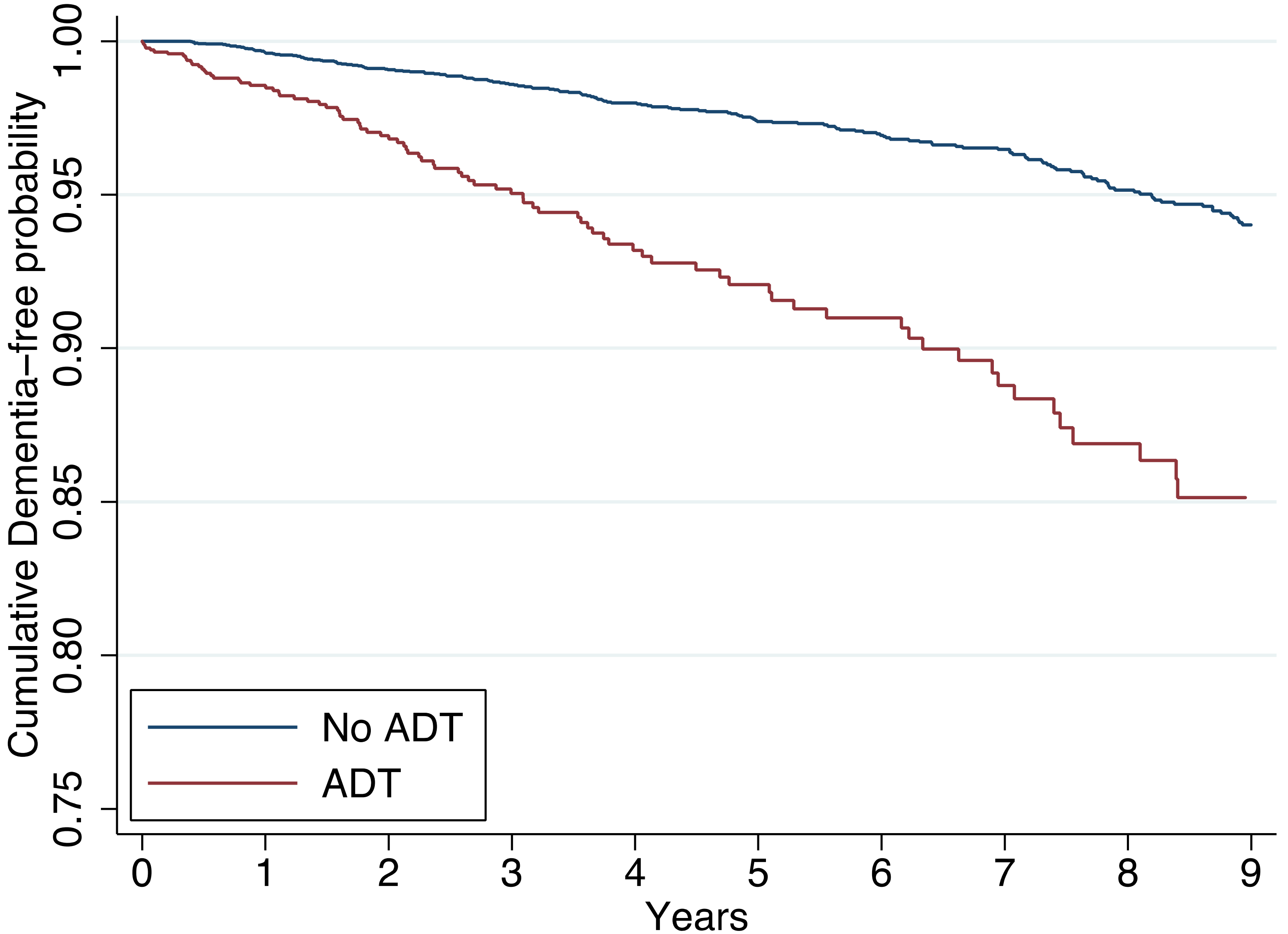
Androgen deprivation & Alzheimer's risk





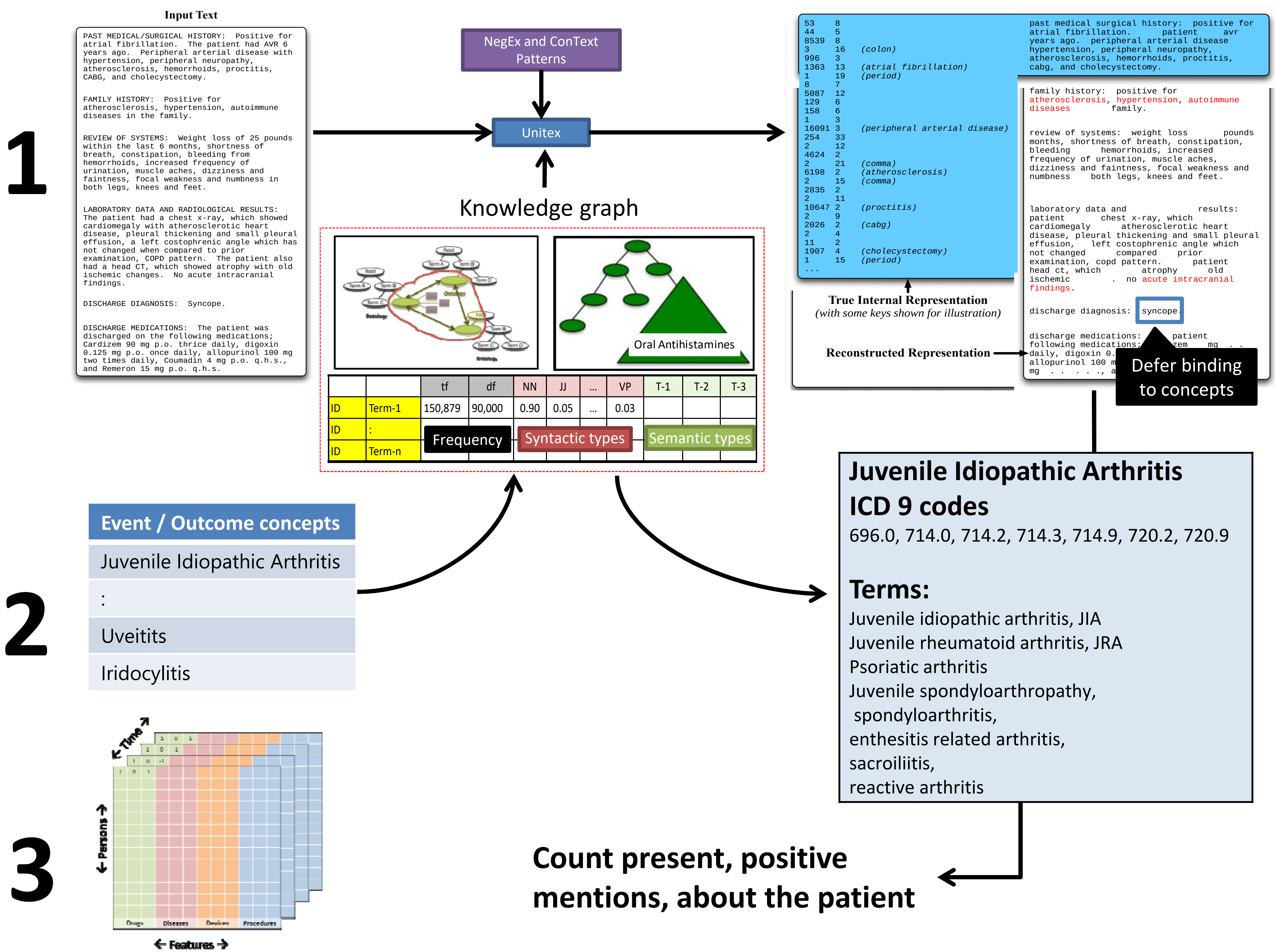
www.tinyurl.com/JCO-ADT

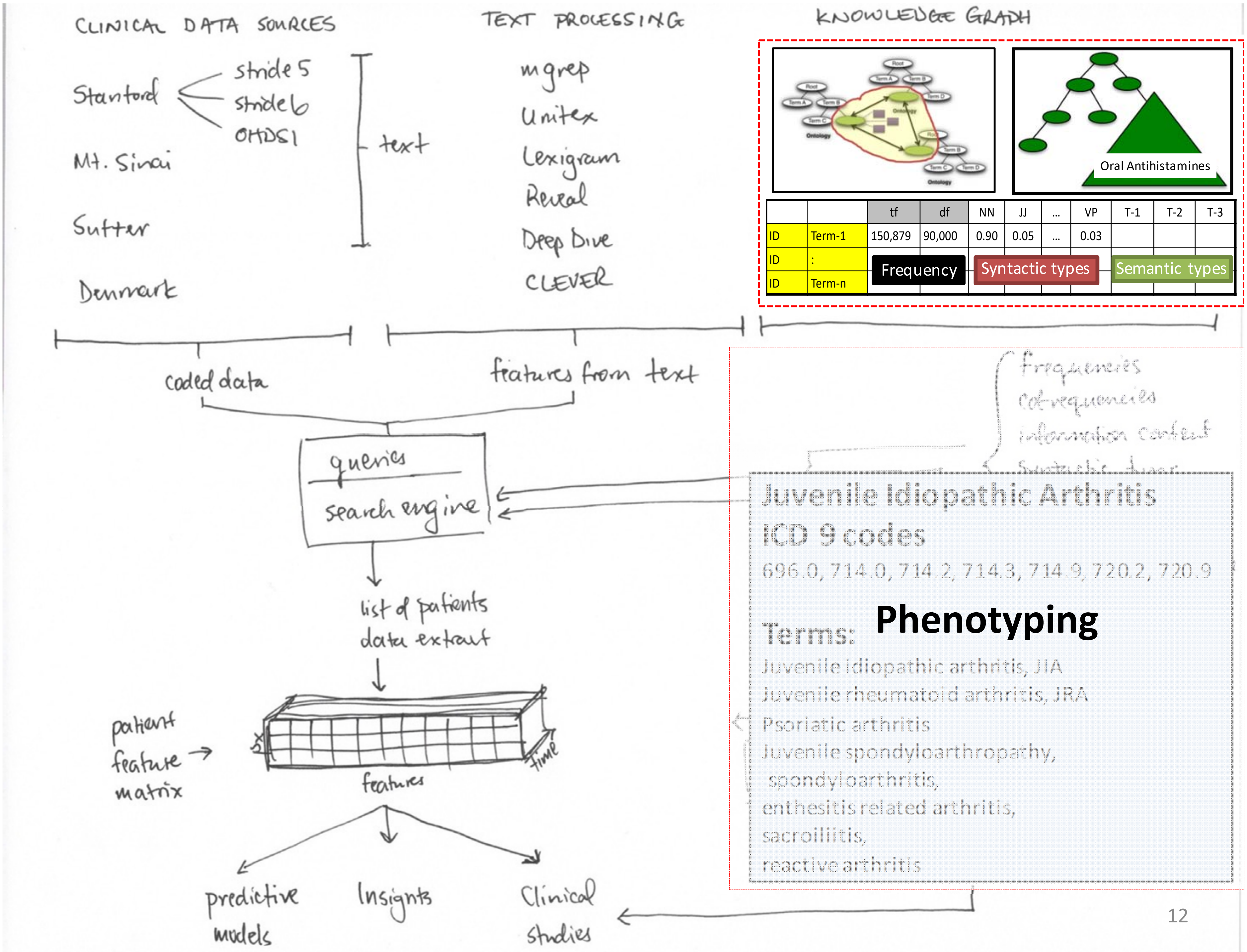
Androgen deprivation & Dementia risk



Number at risk

No ADT	7446	5448	4587	3889	3231	2708	2233	1833	1518	1214
ADT	1826	1177	872	645	470	367	282	211	160	124


1
Input Text
NegEx and ConText Patterns
Unitex
Knowledge graph
Oral Antihistamines
Frequency
Syntactic types
Semantic types
True Internal Representation
(with some keys shown for illustration)
Reconstructed Representation
Defer binding to concepts
2
Event / Outcome concepts
Juvenile Idiopathic Arthritis
:
Uveitits
Iridocylitis
Juvenile Idiopathic Arthritis
ICD 9 codes
696.0, 714.0, 714.2, 714.3, 714.9, 720.2, 720.9
Terms:
Juvenile idiopathic arthritis, JIA
Juvenile rheumatoid arthritis, JRA
Psoriatic arthritis
Juvenile spondyloarthropathy,
spondyloarthritis,
enthesitis related arthritis,
sacroiliitis,
reactive arthritis
3
Time
Persons
Features
Drugs
Diseases
Devices
Procedures
Count present, positive mentions, about the patient


CLINICAL DATA SOURCES
TEXT PROCESSING
KNOWLEDGE GRAPH
Stanford
stride 5
stride 6
OHDSI
Mt. Sinai
Sutter
Denmark
text
mgrep
Unitex
Lexigram
Reveal
Deep Dive
CLEVER
Oral Antihistamines
tf | df | NN | JJ | ... | VP | T-1 | T-2 | T-3
ID | Term-1 | 150,879 | 90,000 | 0.90 | 0.05 | ... | 0.03
ID | :
ID | Term-n
Frequency
Syntactic types
Semantic types
coded data
features from text
frequencies
cofrequencies
information content
queries
search engine
Juvenile Idiopathic Arthritis
ICD 9 codes
696.0, 714.0, 714.2, 714.3, 714.9, 720.2, 720.9
Terms:
Phenotyping
Juvenile idiopathic arthritis, JIA
Juvenile rheumatoid arthritis, JRA
Psoriatic arthritis
Juvenile spondyloarthropathy,
spondyloarthritis,
enthesitis related arthritis,
sacroiliitis,
reactive arthritis
list of patients
data extract
patient feature matrix
features
time
predictive models
Insights
Clinical studies

Insights

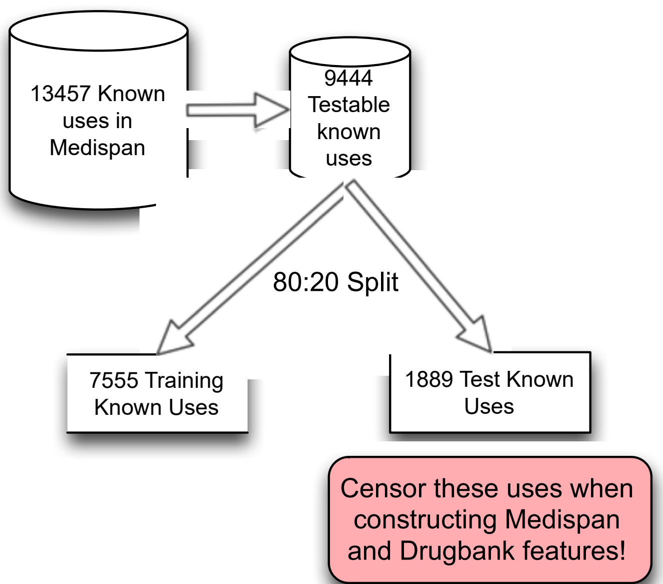
Detecting drug-treats-disease relations

- I2b2 2010 Challenge on information extraction from free text
- Goal: find NER, assertions, and relationships between entities.
  - Relationships included <Drug used to treat Indication>
- Best performance: F1 ~ 0.75 for relations
- Problem solved?

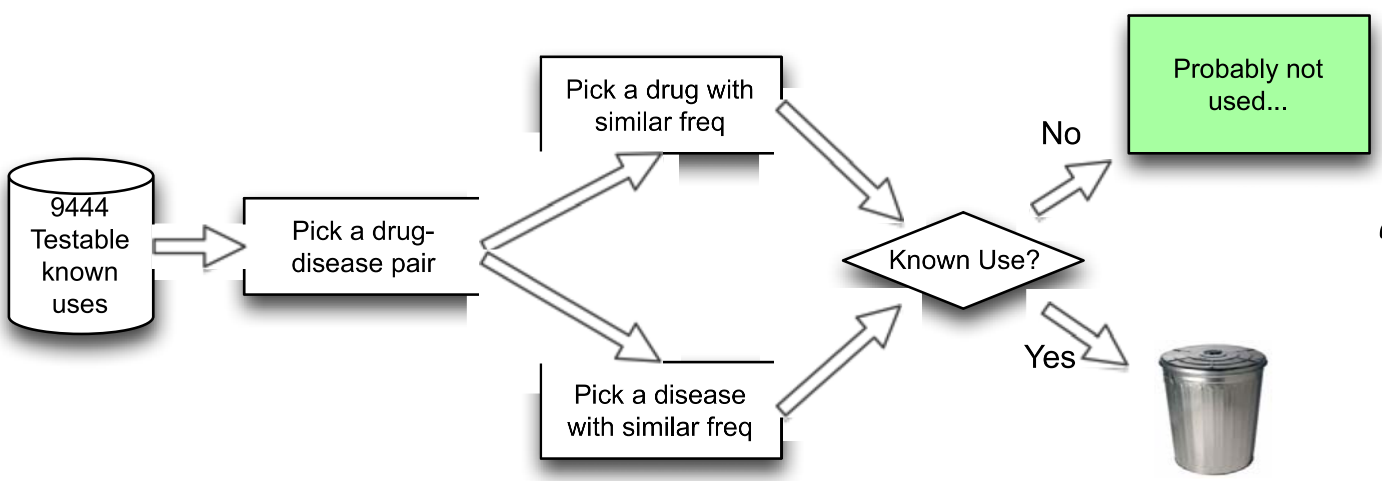
Off-label use via machine learning

- We don’t care about note-level accuracy.
- Can we detect useful signals from aggregate statistics based on noisy low level features?
  - For given drug-disorder pair, is drug used to treat the disorder?
  - Off-label if it is not approved.

Training and test sets

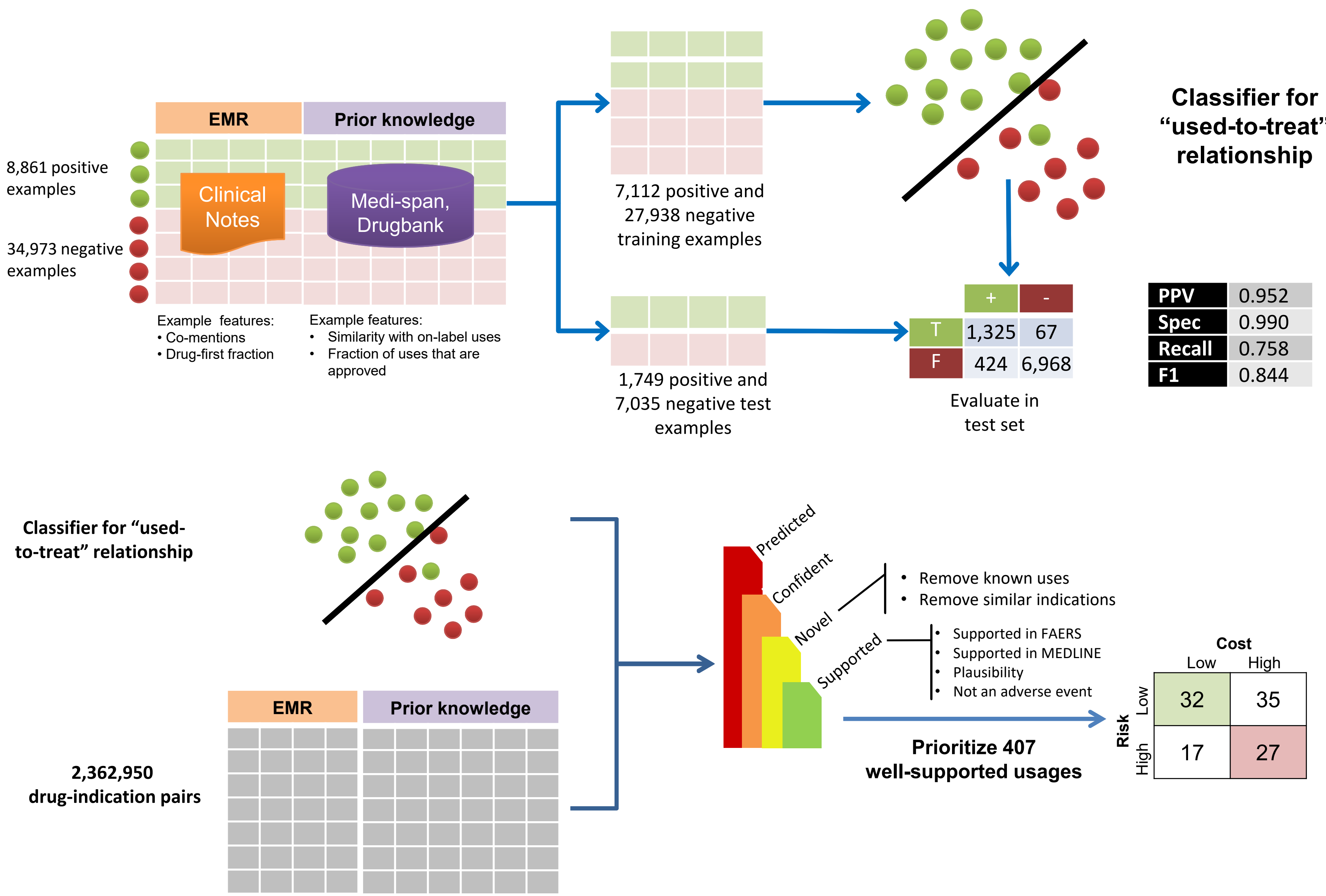


“known” use



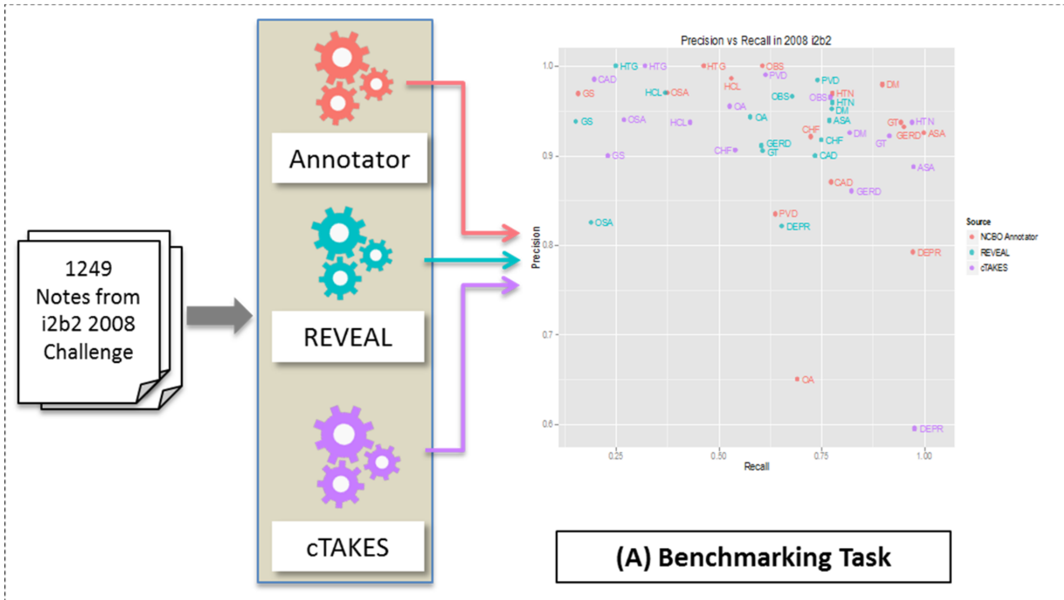
“not used”

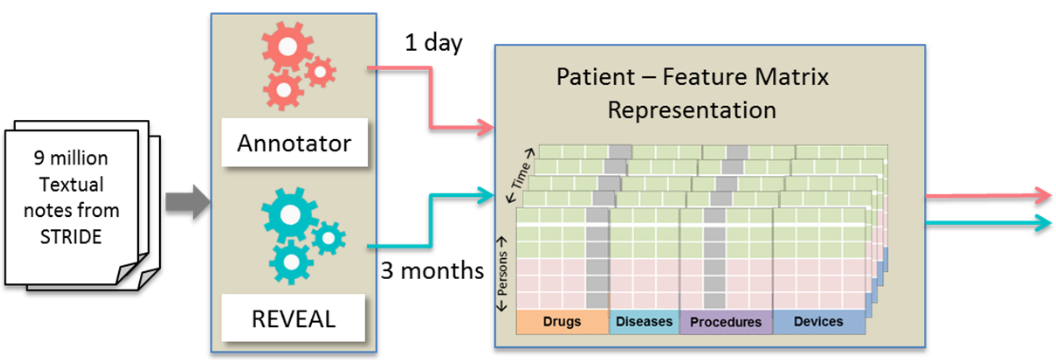
Off-label drug use

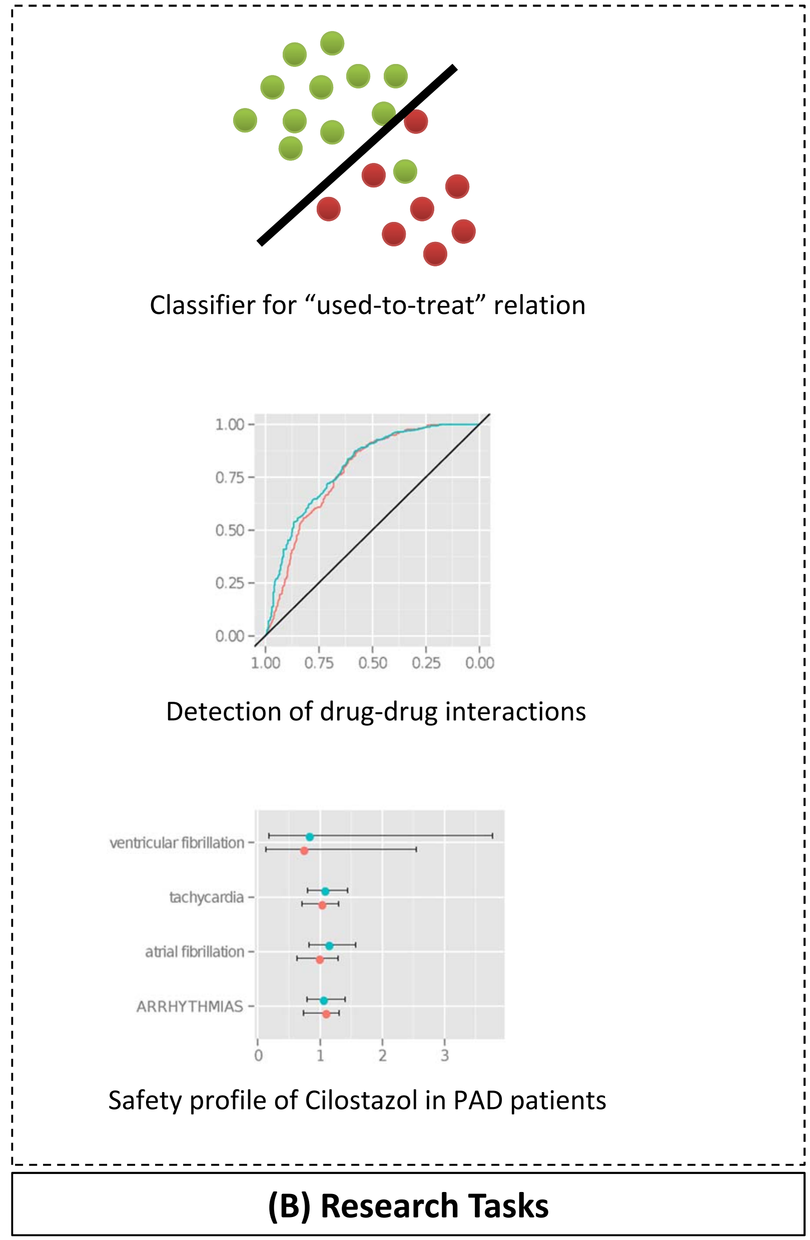


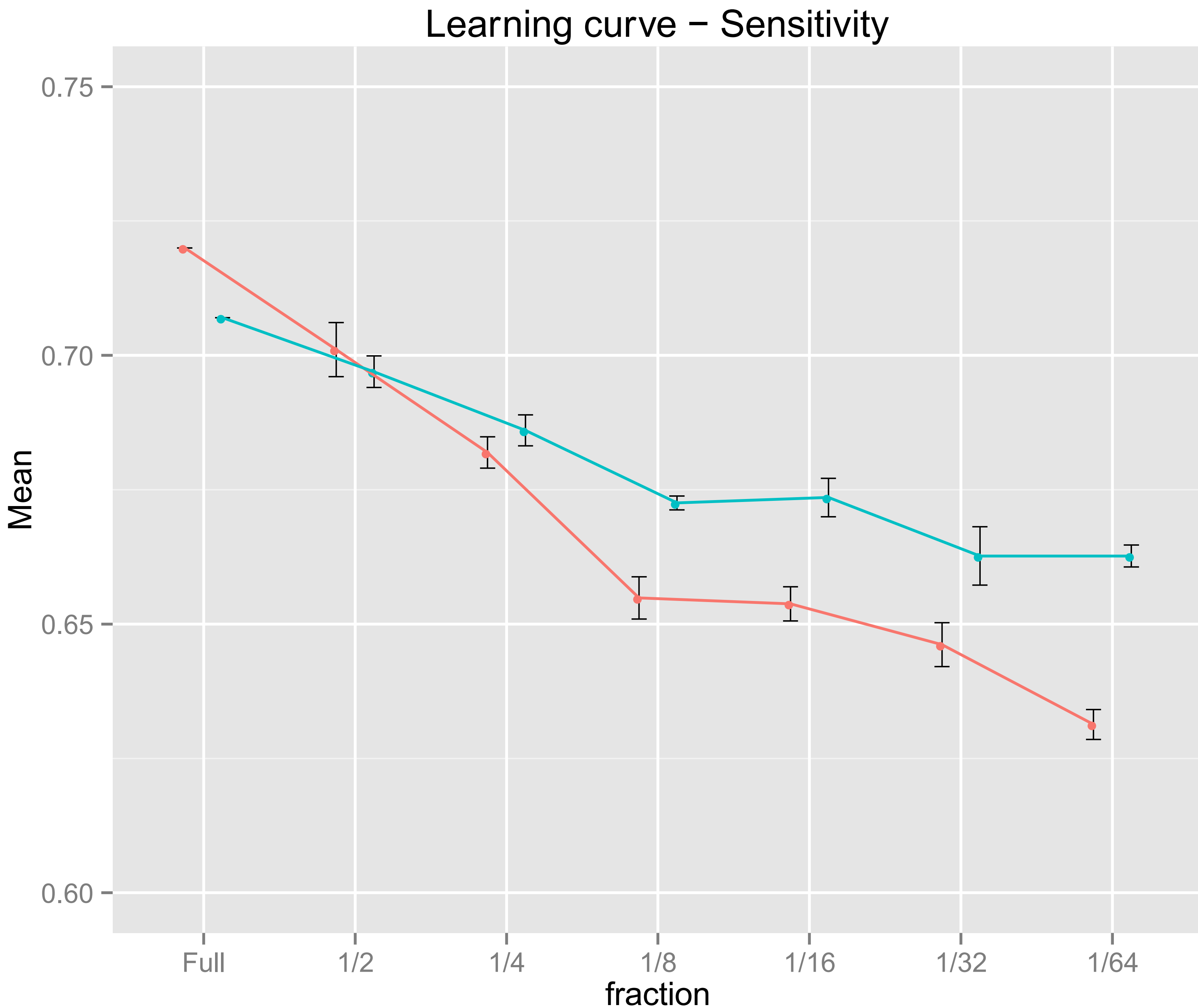
‘Just enough’ text mining

Trade-off: simple or advanced [text-processing]









Learning curve − Sensitivity
0.75
0.70
0.65
0.60
Mean
Full
1/2
1/4
1/8
1/16
1/32
1/64
fraction

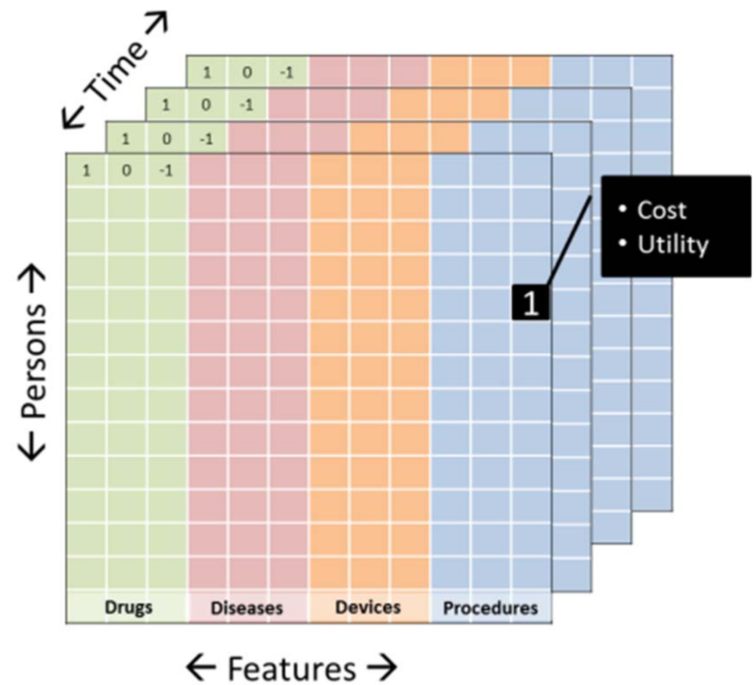
A note on evaluating your “NLP”

1. **String matching**: can you grep, can you grep with typos, can you find the right term, span etc.
2. **Knowledge graph handling**: can you use a knowledge graph to infer that Simvastin is a type of statin
3. **Context and negation**: can you differentiate mentions that are about patient vs. other, negated, historical vs present
4. **Intuitive**: can you infer things that are not mentioned
   - e.g. 5 feet tall, 200 lbs —> obesity
5. **Phenotype:** can you recognize [known] phenotypes correctly
   - e.g. exposure to drug + ALP >= 2x ULN + normal lab measurements prior to exposure to drug --> drug induced liver injury
6. **Functional**: how accurately can the output of the processing be used to accomplish a research task, such as detect adverse drug events
7. **Utility:** if the method was used to generate results, would it change practice
   - e.g. we give pneumovax to a 100 more patients, because text-mining told us that they had a splenectomy

Ask: about the cost-utility trade-off

- EHR mining is a process
- Text is one of the many sources
- Time needs special handling

- Machine learning is used in many places
  - Sorting the documents that contain the text of interest
  - In the processing of the text to extract features and facts (”NLP”)
  - In the processing of time to extract features
  - Finding associations among the extracted features



Acknowledgements

Group Members:

- **Scientists:** Ken Jung, Alison Callahan, Juan Banda
- **Fellows**: Rohit Vashisht, Azadeh Nikferjam, Katie Quinn
- **Engineer:** Vladimir Polony
- **BMI Students:** Sarah Poole, Alejandro Schuler, Vibhu Agarwal
- **Med Students:** Mehr Kashyap, Jassi Pannu

Alums: Anna Bauer-Mehren (Roche), Srini Iyer (Facebook), Amogh Vasekar (Citrix), Sandy Huang (Berkeley), Paea LePendu (Lexigram), Rave Harpaz (Oracle), Tyler Cole (Barrow Inst.), Sam Finlayson (Harvard), Will Chen (Yale), Yen Low (Netflix), Elsie Gyang (Fellowship in Surgery), Suzanne Tamang (Instructor)

Collaborators: Purvesh Khatri, Tina Hernandez-Boussard, Winn Haynes, Kevin Nead, Nick Leeper

Funding:

- NIH – NLM, NIGMS, NHGRI, NINDS, NCI, FDA
- Stanford Internal – Dept. of Medicine, Population Health Sciences, Clinical Excellence Research Center, Dean's office
- Fellowships – Med Scholars, Siebel Scholars Foundation, Stanford Graduate Fellowship
- Industry – Apixio, CollabRx, Healogics, Janssen R&D, Oracle, Baidu USA, Amgen

IT: Alex Skrenchuk, SCCI team