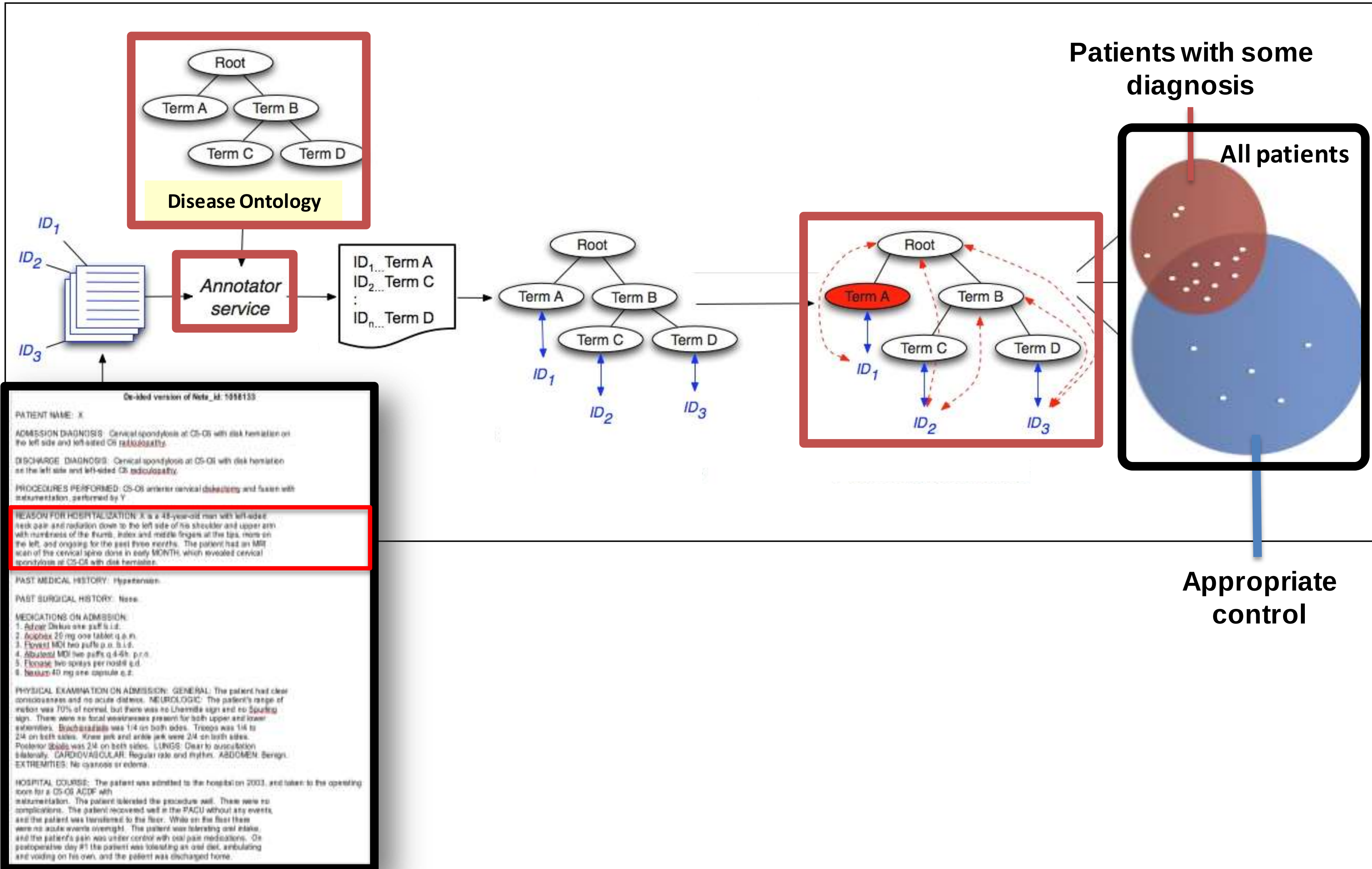
Using ontologies to mine unstructured data in medicine

Nigam Shah, MBBS, PhD

nigam@stanford.edu

STANFORD
SCHOOL OF MEDICINE

Profiling a patient set



Patients with some diagnosis

All patients

Appropriate control

De-ided version of Note_id: 1058133

PATIENT NAME: X

ADMISSION DIAGNOSIS: Cervical spondylosis at C5-C6 with disk herniation on the left side and left-sided C6 radiculopathy.

DISCHARGE DIAGNOSIS: Cervical spondylosis at C5-C6 with disk herniation on the left side and left-sided C6 radiculopathy.

PROCEDURES PERFORMED: C5-C6 anterior cervical diskectomy and fusion with instrumentation, performed by Y.

REASON FOR HOSPITALIZATION: X is a 48-year-old man with left-sided neck pain and radiation down to the left side of his shoulder and upper arm with numbness of the thumb, index and middle fingers at the tips, more on the left, and ongoing for the past three months. The patient had an MRI scan of the cervical spine done in early MONTH, which revealed cervical spondylosis at C5-C6 with disk herniation.

PAST MEDICAL HISTORY: Hypertension.

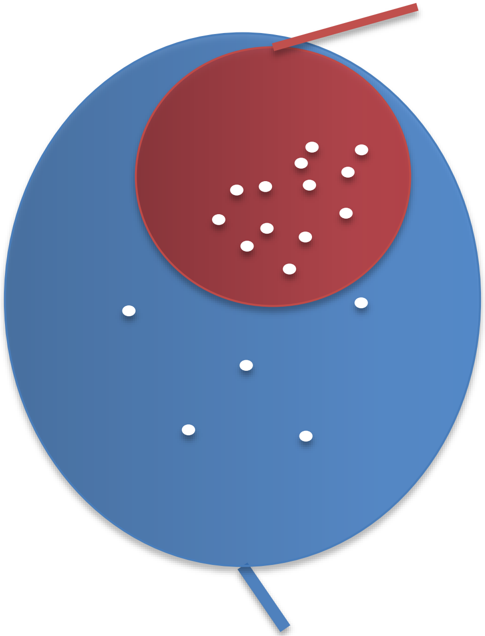
PAST SURGICAL HISTORY: None.

MEDICATIONS ON ADMISSION:
1. Advair Diskus one puff b.i.d.
2. Aciphex 20 mg one tablet q.a.m.
3. Flovent MDI two puffs p.o. b.i.d.
4. Albuterol MDI two puffs q.4-6h. p.r.n.
5. Flonase two sprays per nostril q.d.
6. Nexium 40 mg one capsule q.d.

PHYSICAL EXAMINATION ON ADMISSION: GENERAL: The patient had clear consciousness and no acute distress. NEUROLOGIC: The patient's range of motion was 70% of normal, but there was no Lhermitte sign and no Spurling sign. There were no focal weaknesses present for both upper and lower extremities. Brachioradialis was 1/4 on both sides. Triceps was 1/4 to 2/4 on both sides. Knee jerk and ankle jerk were 2/4 on both sides. Posterior tibialis was 2/4 on both sides. LUNGS: Clear to auscultation bilaterally. CARDIOVASCULAR: Regular rate and rhythm. ABDOMEN: Benign. EXTREMITIES: No cyanosis or edema.

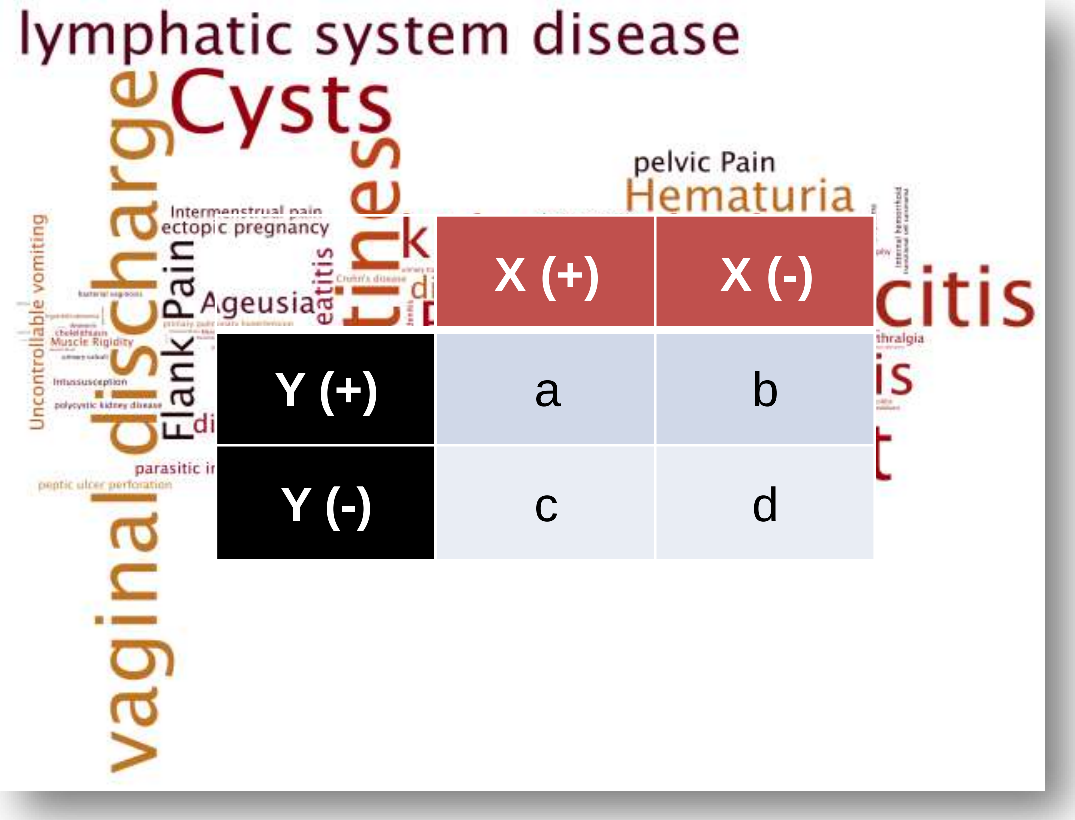
HOSPITAL COURSE: The patient was admitted to the hospital on 2003, and taken to the operating room for a C5-C6 ACDF with instrumentation. The patient tolerated the procedure well. There were no complications. The patient recovered well in the PACU without any events, and the patient was transferred to the floor. While on the floor there were no acute events overnight. The patient was tolerating oral intake, and the patient's pain was under control with oral pain medications. On postoperative day #1 the patient was tolerating an oral diet, ambulating and voiding on his own, and the patient was discharged home.

Profiling patient sets



ICD9 789.00
(*Abdominal pain, unspecified site*)

86k patient Reports



	X (+)	X (-)
Y (+)	a	b
Y (-)	c	d

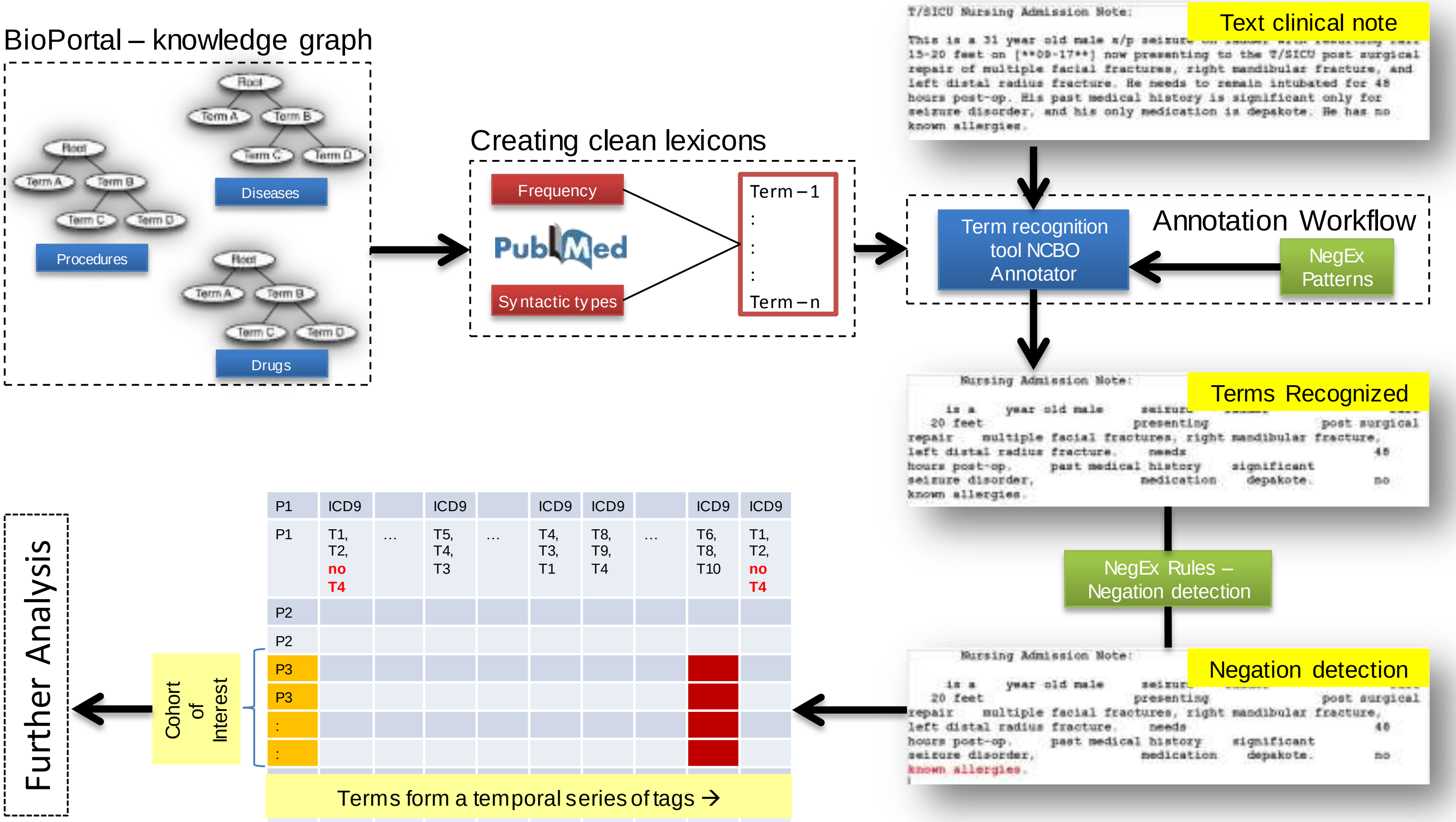
Patient records processed from U. Pittsburg NLP Repository with IRB approval.

Associations and outcomes

	Gene	Disease	Drug	Device	Procedure	Environment
Gene		Gene Enrichment				
Disease			Off-label Indications			
Drug		Side effects				
Device						
Procedure						
Environment						

What associations can we find?

Generation of annotated data at scale



BioPortal – knowledge graph

Diseases

Procedures

Drugs

Creating clean lexicons

Frequency

PubMed

Syntactic types

Term – 1
:
:
:
Term – n

Text clinical note

T/SICU Nursing Admission Note:

This is a 31 year old male s/p seizure on ladder with resulting fall 15-20 feet on [**09-17**] now presenting to the T/SICU post surgical repair of multiple facial fractures, right mandibular fracture, and left distal radius fracture. He needs to remain intubated for 48 hours post-op. His past medical history is significant only for seizure disorder, and his only medication is depakote. He has no known allergies.

Annotation Workflow

Term recognition tool NCBO Annotator

NegEx Patterns

Terms Recognized

Nursing Admission Note:

is a year old male seizure
20 feet presenting post surgical
repair multiple facial fractures, right mandibular fracture,
left distal radius fracture. needs 48
hours post-op. past medical history significant
seizure disorder, medication depakote. no
known allergies.

NegEx Rules – Negation detection

Negation detection

Nursing Admission Note:

is a year old male seizure
20 feet presenting post surgical
repair multiple facial fractures, right mandibular fracture,
left distal radius fracture. needs 48
hours post-op. past medical history significant
seizure disorder, medication depakote. no
known allergies.

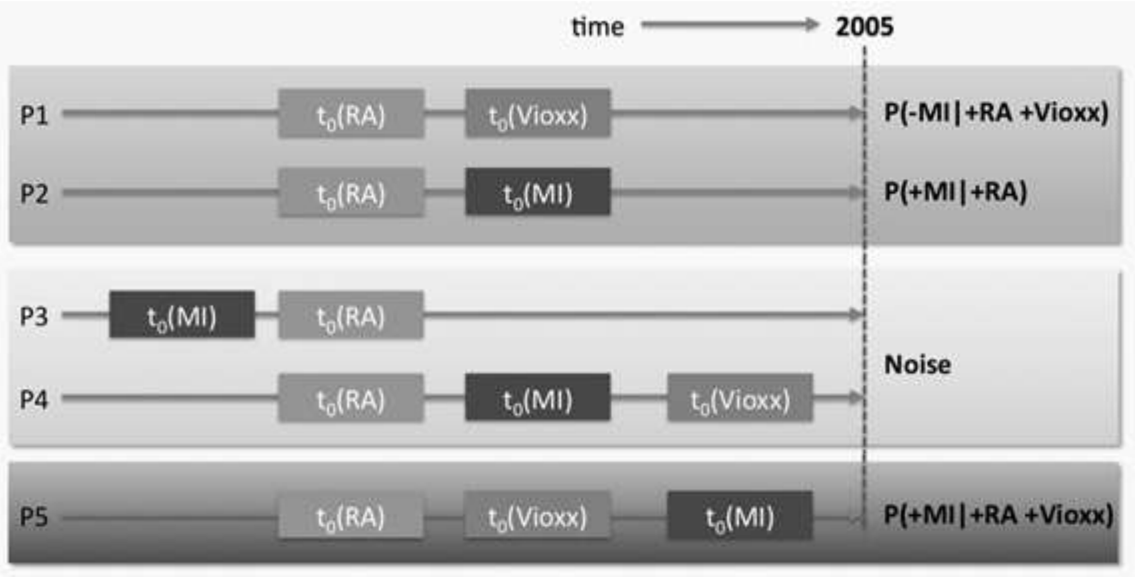
P1	ICD9		ICD9		ICD9	ICD9		ICD9	ICD9
P1	T1, T2, **no** **T4**	…	T5, T4, T3	…	T4, T3, T1	T8, T9, T4	…	T6, T8, T10	T1, T2, **no** **T4**
P2									
P2									
P3								🟥	
P3								🟥	
:								🟥	
:								🟥	

Terms form a temporal series of tags →

Cohort of Interest

Further Analysis

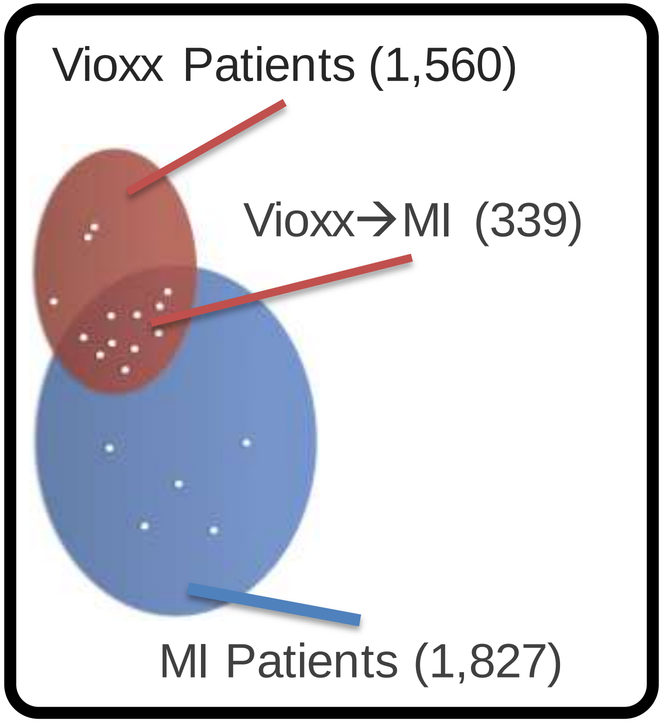
Detecting the Vioxx Risk Signal



ROR of 2.058, CI of [1.804, 2.349]
The X^2 statistic has p-value $< 10^{-7}$

ROR=1.524, CI=[0.872, 2.666] X^2 p-value = 0.06816.

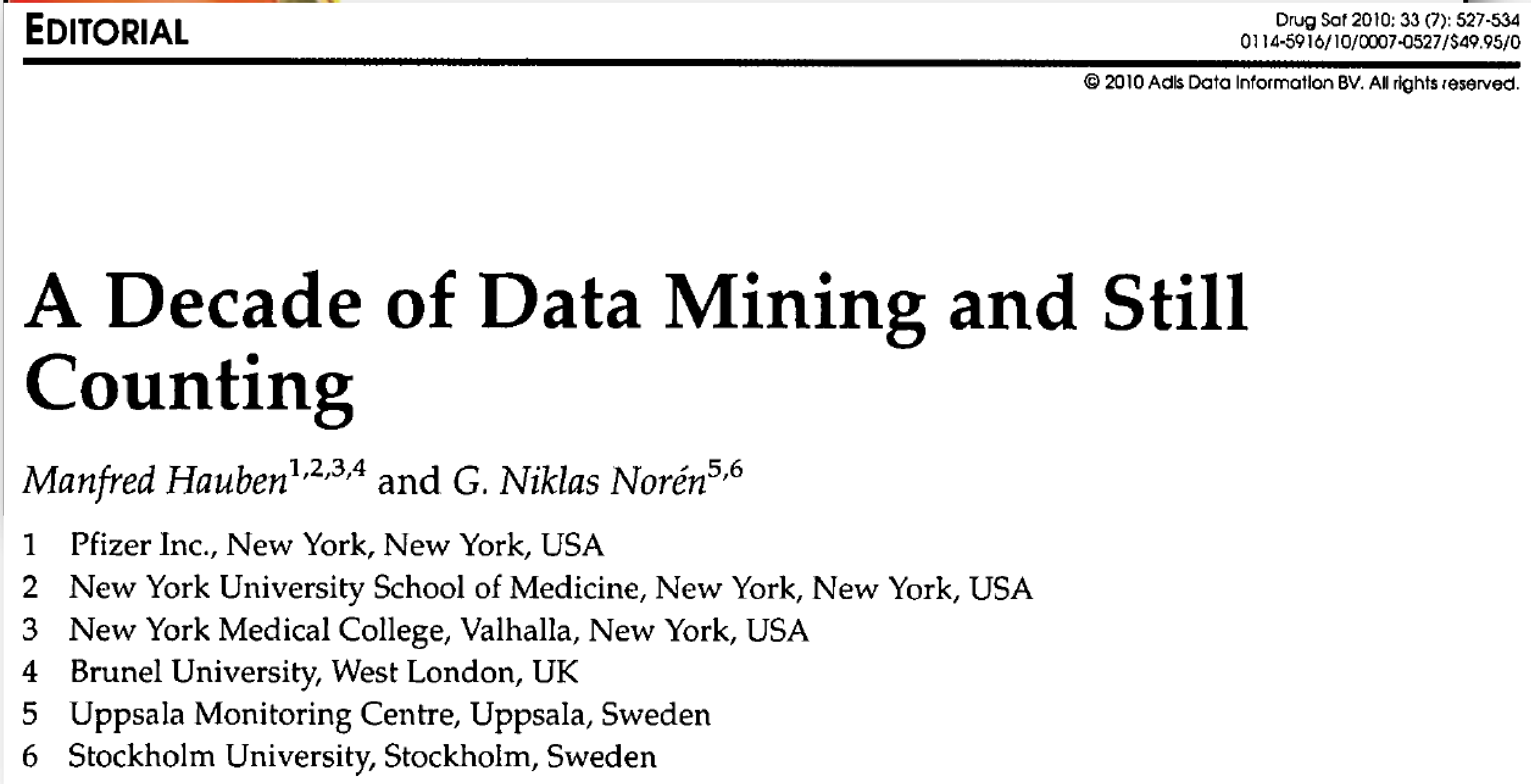
	MI	No MI
Vioxx	a = 339	b = 1221
No Vioxx	c = 1488	d = 11031



RA Patients (14,079)

p-value $< 1.3 \times 10^{-24}$

We should stop acting as if our goal is to author extremely elegant theories, […] and make use of the best ally we have: the unreasonable effectiveness of data.



EDITORIAL

Drug Saf 2010; 33 (7): 527-534
0114-5916/10/0007-0527/$49.95/0

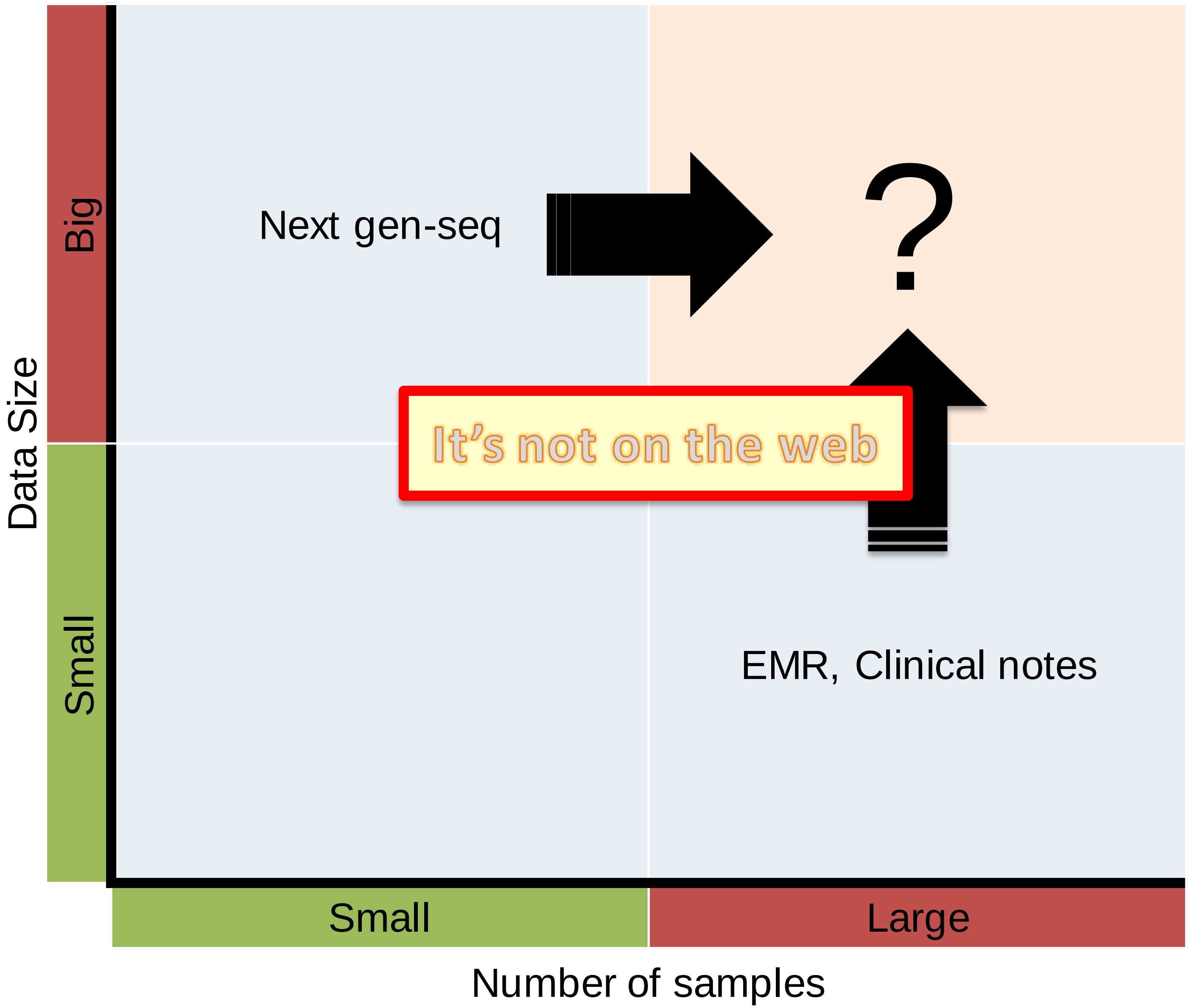
© 2010 Adis Data Information BV. All rights reserved.

A Decade of Data Mining and Still Counting

Manfred Hauben[1,2,3,4] and G. *Niklas Norén*[5,6]

1 Pfizer Inc., New York, New York, USA
2 New York University School of Medicine, New York, New York, USA
3 New York Medical College, Valhalla, New York, USA
4 Brunel University, West London, UK
5 Uppsala Monitoring Centre, Uppsala, Sweden
6 Stockholm University, Stockholm, Sweden

Big Data in biomedicine



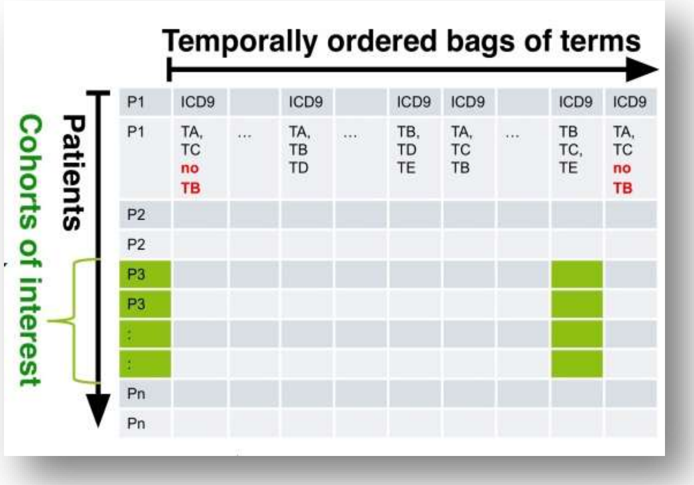
The problem

	On-label	Off-label
Indication	What Pharma companies get approval for	Whatever else the doctor prescribes for
Side effect / Adverse effect	Found during the pre-marketing phase	Goal of drug-safety surveillance

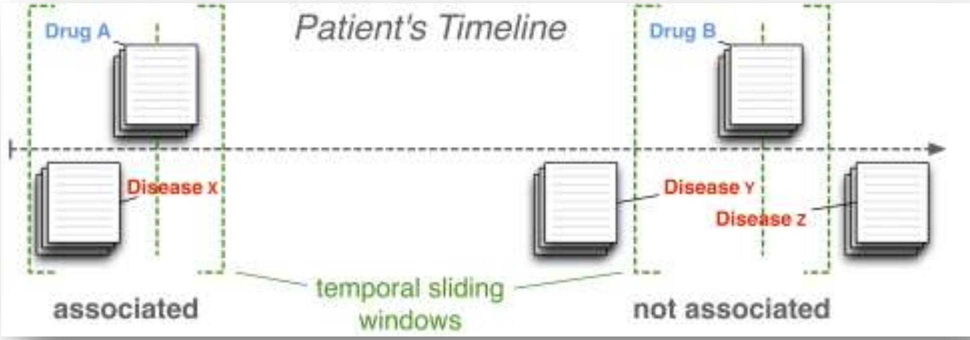
- Ambulatory: 100,000 deaths and $177 billion annually
- In patient: estimated that roughly 30% of hospital stays have an adverse drug event

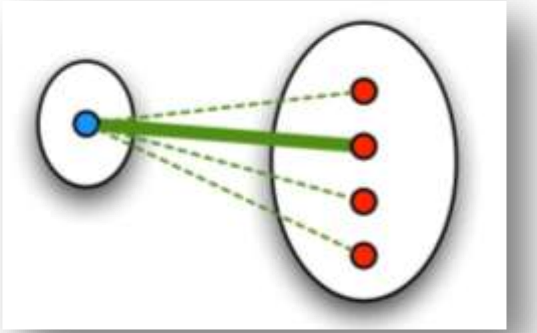
- 21% of prescriptions
- 73% with very little evidence

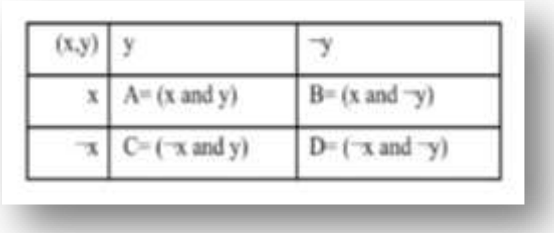
Detecting Off-label use



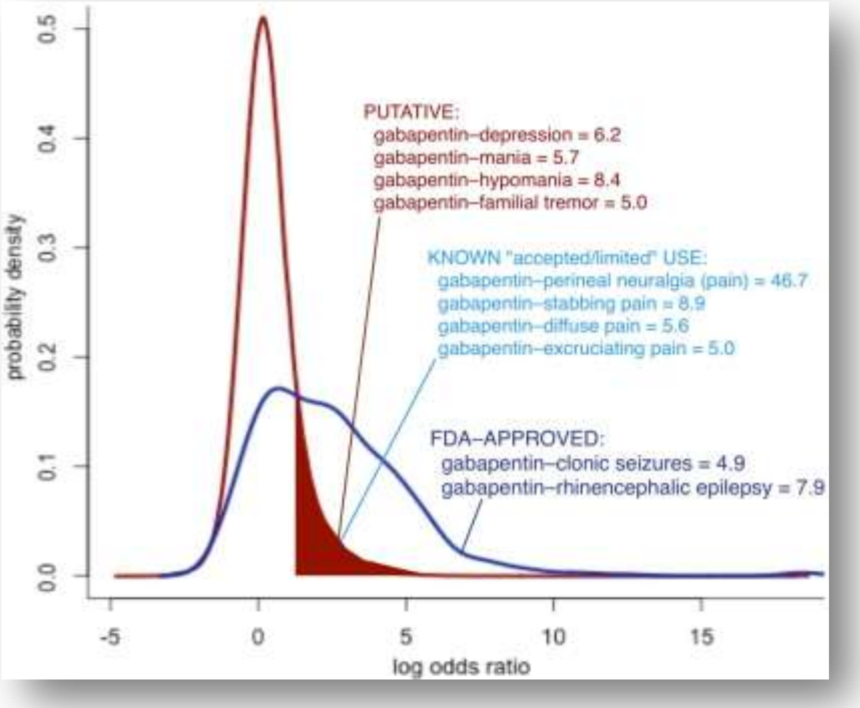
P1	ICD9		ICD9		ICD9	ICD9		ICD9	ICD9
P1	TA, TC no TB	…	TA, TB TD	…	TB, TD TE	TA, TC TB	…	TB TC, TE	TA, TC no TB
P2									
P2									
P3									
P3									
:									
:									
Pn									
Pn									







(x,y)	y	¬y
x	A= (x and y)	B= (x and ¬y)
¬x	C= (¬x and y)	D= (¬x and ¬y)

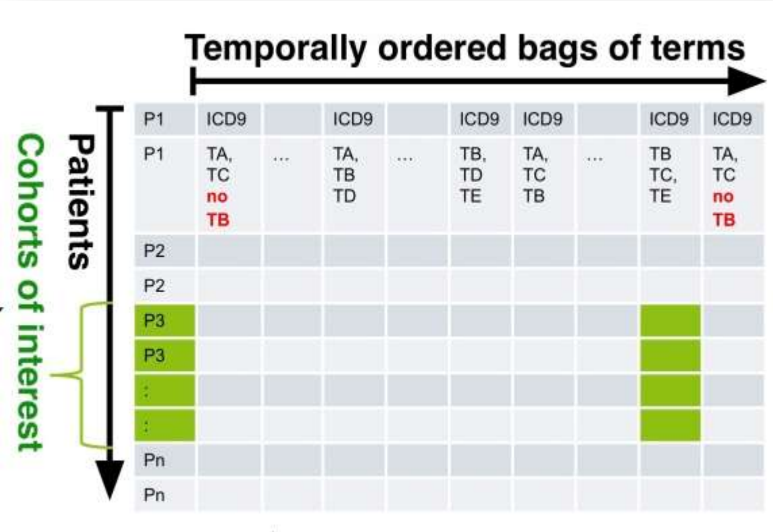


PUTATIVE:
gabapentin–depression = 6.2
gabapentin–mania = 5.7
gabapentin–hypomania = 8.4
gabapentin–familial tremor = 5.0

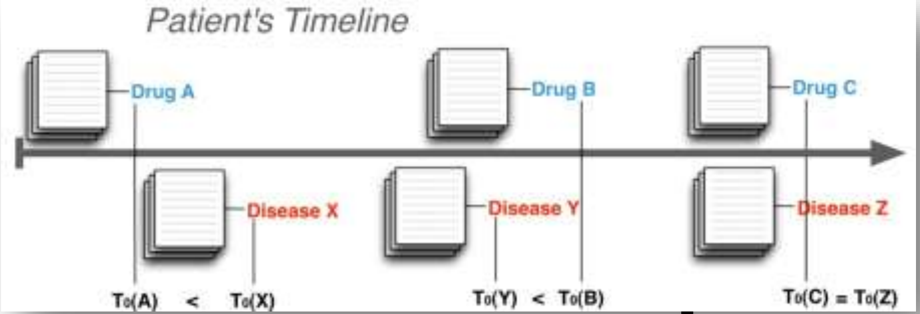
KNOWN "accepted/limited" USE:
gabapentin–perineal neuralgia (pain) = 46.7
gabapentin–stabbing pain = 8.9
gabapentin–diffuse pain = 5.6
gabapentin–excruciating pain = 5.0

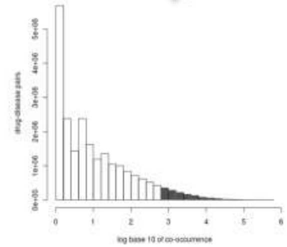
FDA–APPROVED:
gabapentin–clonic seizures = 4.9
gabapentin–rhinencephalic epilepsy = 7.9

Detecting Adverse Events

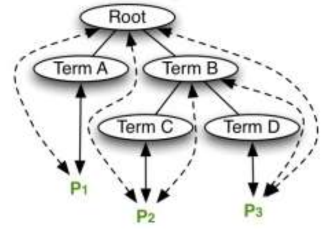


P1	ICD9		ICD9		ICD9	ICD9		ICD9	ICD9
P1	TA, TC **no** **TB**	…	TA, TB TD	…	TB, TD TE	TA, TC TB	…	TB TC, TE	TA, TC **no** **TB**
P2									
P2									
P3								🟩	
P3								🟩	
:								🟩	
:								🟩	
Pn									
Pn									

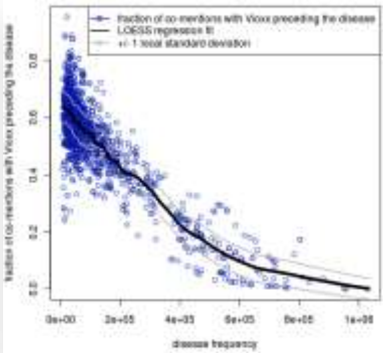


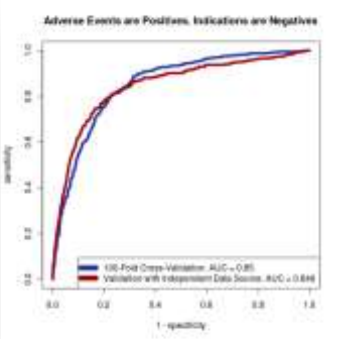


492 thousand pairs



959 thousand pairs





Patterns worth testing (off-label usage, which is risky)

✓ **Identify off-label use**

- Find drug-indication pairs that “look like” indications

✓ **Identify which use “may be risky”**

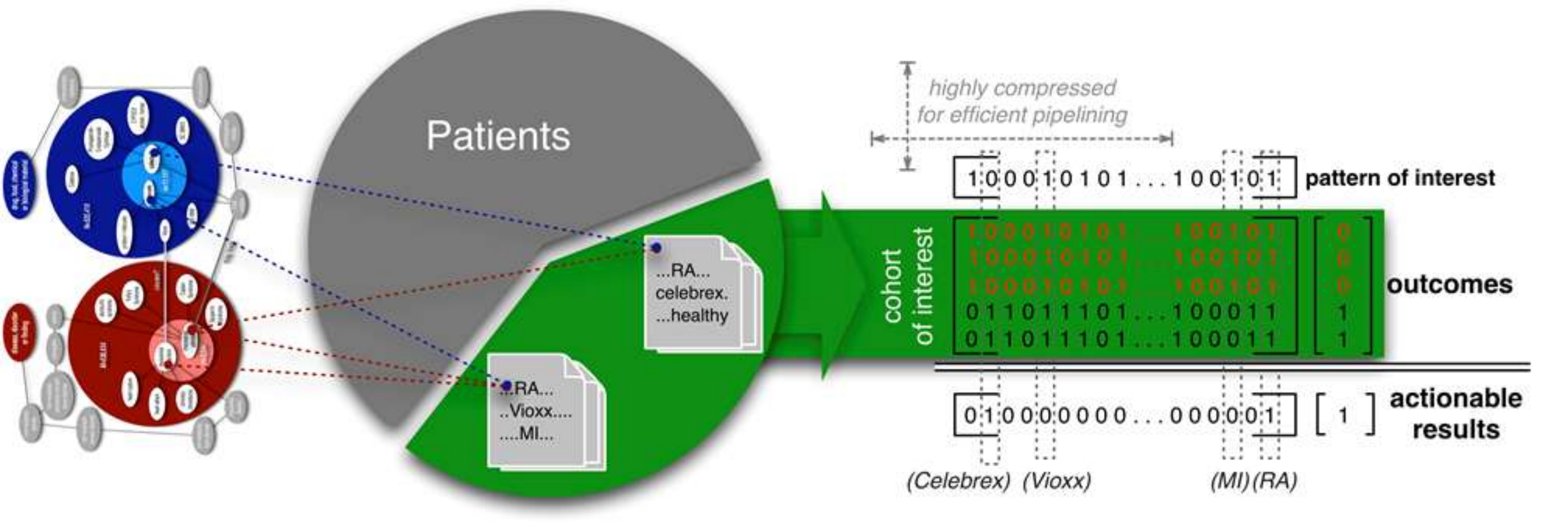
- Use existing, known side effect databases
- Learn drug-disease associations that look like side effects

✓ **Assemble I-D-A triplets**

- Indication – Drug – Adverse effect. e.g. RA – Vioxx – MI

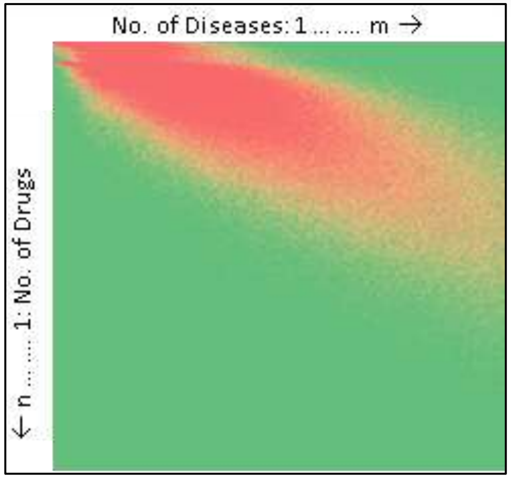
✓ **Test on unstructured data**

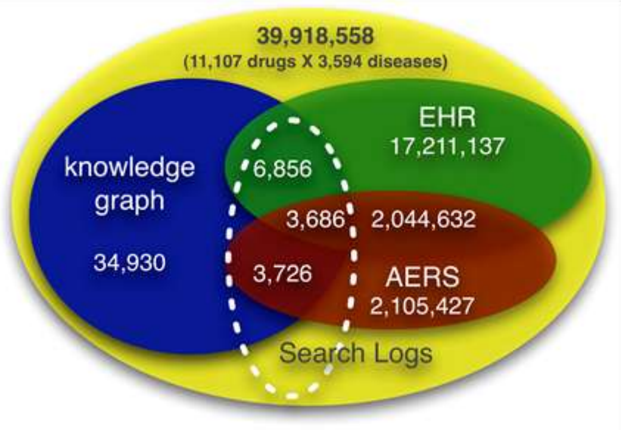
Testing ‘interesting patterns’

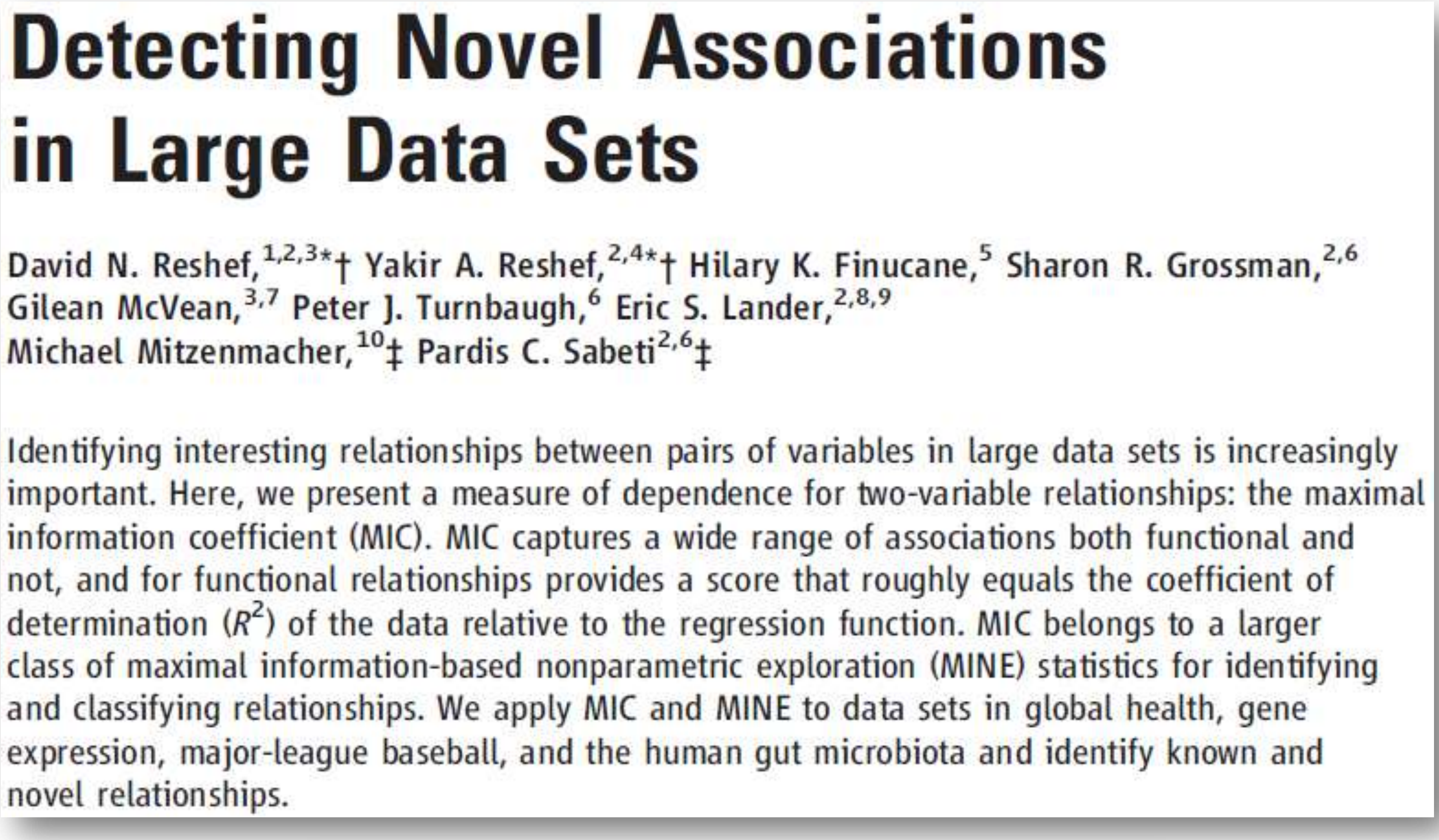


1 0 0 0 1 0 1 0 1 . . . 1 0 0 1 0 1 pattern of interest

1 0 0 0 1 0 1 0 1 . . . 1 0 0 1 0 1	0
1 0 0 0 1 0 1 0 1 . . . 1 0 0 1 0 1	0
1 0 0 0 1 0 1 0 1 . . . 1 0 0 1 0 1	0
0 1 1 0 1 1 1 0 1 . . . 1 0 0 0 1 1	1
0 1 1 0 1 1 1 0 1 . . . 1 0 0 0 1 1	1
0 1 0 0 0 0 0 0 0 . . . 0 0 0 0 0 1	1







Detecting Novel Associations in Large Data Sets

David N. Reshef,[1,2,3*†] Yakir A. Reshef,[2,4*†] Hilary K. Finucane,[5] Sharon R. Grossman,[2,6] Gilean McVean,[3,7] Peter J. Turnbaugh,[6] Eric S. Lander,[2,8,9] Michael Mitzenmacher,[10‡] Pardis C. Sabeti[2,6‡]

Identifying interesting relationships between pairs of variables in large data sets is increasingly important. Here, we present a measure of dependence for two-variable relationships: the maximal information coefficient (MIC). MIC captures a wide range of associations both functional and not, and for functional relationships provides a score that roughly equals the coefficient of determination (R^2) of the data relative to the regression function. MIC belongs to a larger class of maximal information-based nonparametric exploration (MINE) statistics for identifying and classifying relationships. We apply MIC and MINE to data sets in global health, gene expression, major-league baseball, and the human gut microbiota and identify known and novel relationships.

Science **334**, 1518 (2011);
DOI: 10.1126/science.1205438

The team @
www.bioontology.org/project-team

NIH Roadmap grant U54 HG004028

STANFORD
SCHOOL OF MEDICINE