



# NLP

---

Feb 28, 2019  
March 5, 2019



**Massachusetts  
Institute of  
Technology**

# Outline

---

- **Value of the data in clinical text**
- Hyper-simplified linguistics
- Term spotting + handling negation, uncertainty
- ML to expand terms
- pre-NN ML to identify entities and relations
- language models
- Neural methods

# Bulk of Valuable Data are in Narrative Text

---

orange=demographics

blue=patient condition, diseases, etc.

brown=procedures, tests

magenta=results of measurements

purple=time

Mr. Blind is a 79-year-old white male with a history of diabetes mellitus, inferior myocardial infarction, who underwent open repair of his increased diverticulum November 13th at Sephsandpot Center.

The patient developed hematemesis November 15th and was intubated for respiratory distress. He was transferred to the Valtawnprinceel Community Memorial Hospital for endoscopy and esophagoscopy on the 16th of November which showed a 2 cm linear tear of the esophagus at 30 to 32 cm. The patient's hematocrit was stable and he was given no further intervention.

The patient attempted a gastrografin swallow on the 21st, but was unable to cooperate with probable aspiration. The patient also had been receiving generous intravenous hydration during the period for which he was NPO for his esophageal tear and intravenous Lasix for a question of pulmonary congestion.

On the morning of the 22nd the patient developed tachypnea with a chest X-ray showing a question of congestive heart failure. A medical consult was obtained at the Valtawnprinceel Community Memorial Hospital. The patient was given intravenous Lasix.

# Selection of Rheumatoid Arthritis Cohort

**Table 4. Comparison of performance characteristics from validation of the complete classification algorithm (narrative and codified) with algorithms containing codified-only and narrative-only data\***

Model	RA by algorithm or criteria, no.	PPV (95% CI), %	Sensitivity (95% CI), %	Difference in PPV (95% CI), %†
Algorithms				
Narrative and codified (complete)	3,585	94 (91–96)	63 (51–75)	Reference
Codified only	3,046	88 (84–92)	51 (42–60)	6 (2–9)‡
NLP only	3,341	89 (86–93)	56 (46–66)	5 (1–8)‡
Published administrative codified criteria				
≥3 ICD-9 RA codes	7,960	56 (47–64)	80 (72–88)	38 (29–47)‡
≥1 ICD-9 RA codes plus ≥1 DMARD	7,799	45 (37–53)	66 (57–76)	49 (40–57)‡

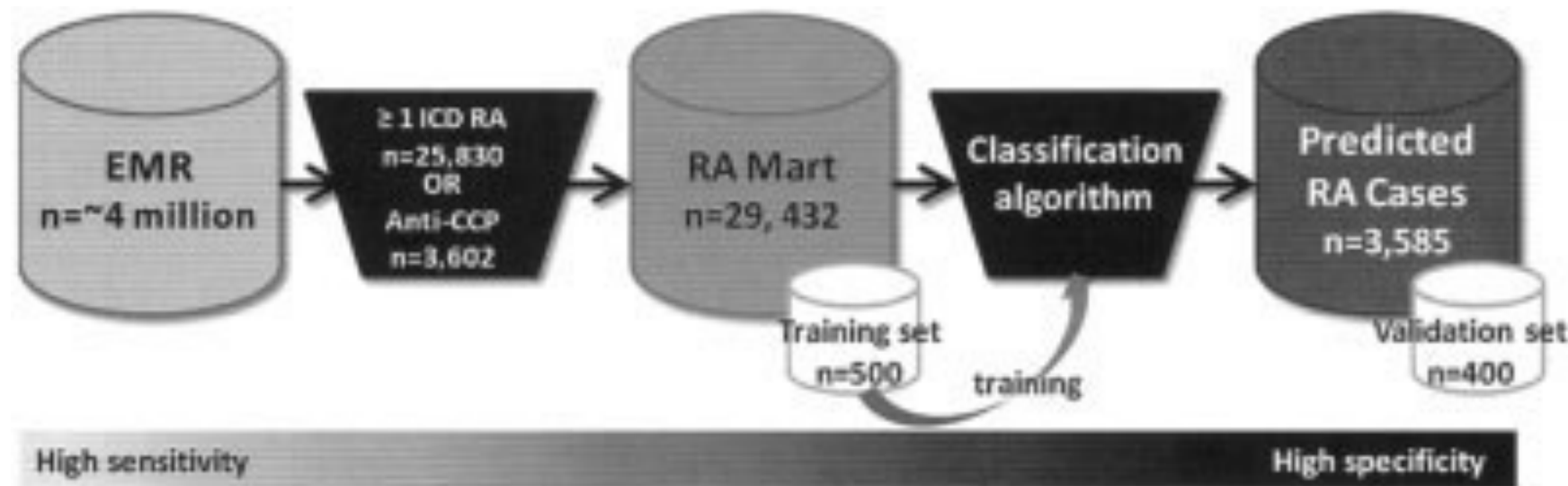
\* The complete classification algorithm was also compared with criteria for RA used in published administrative database studies. RA = rheumatoid arthritis; PPV = positive predictive value; 95% CI = 95% confidence interval; NLP = natural language processing; ICD-9 = International Classification of Diseases, Ninth Revision; DMARD = disease-modifying antirheumatic drug.

† Difference in PPV = PPV of complete algorithm – comparison algorithm or criteria.

‡ Significant difference in PPV compared with the complete algorithm.

Liao, K. P., Cai, T., Gainer, V., Goryachev, S., Zeng-Treitler, Q., Raychaudhuri, S., Szolovits, P., Churchill, S., Murphy, S., Kohane, I., Karlson, E., Plenge, R. (2010). Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care & Research*, 62(8), 1120–1127. <http://doi.org/10.1002/acr.20184>

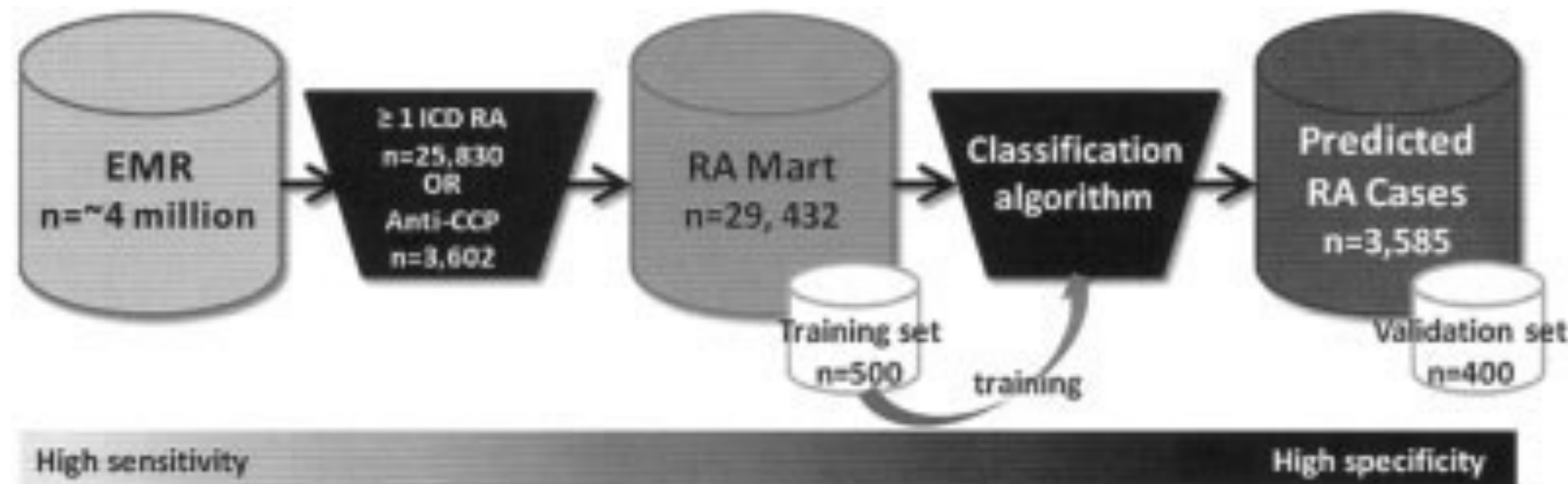
# Finding a Cohort of Rheumatoid Arthritis Cases



- Coded data:
  - ICD-9 codes, including RA and related diseases
    - ignore codes within 1 week of previous code
  - electronic prescriptions for
    - DMARDs: methotrexate, azathioprine, leflunomide, sulfasalazine, hydroxychloroquine, penicillamine, cyclosporine, and gold
    - Biologic agents: anti-TNF agents infliximab and etanercept, and abatacept, rituximab, anakinra, etc.
  - anti-cyclic citrullinated peptide (anti-CCP) & rheumatoid factor (RF) labs
  - total number of “facts” in the EMR



# Finding a Cohort of Rheumatoid Arthritis Cases



Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006;6:30.

- Narrative text data (processed by HITEx)
  - From health care provider notes, radiology reports, pathology reports, discharge summaries, and operative reports
  - Extracted disease diagnoses (RA, SLE, PsA, and JRA)
  - medications (same as from prescriptions, with the addition of adalimumab)
  - laboratory data (RF, anti-CCP, and the term “seropositive”)
  - radiology findings of erosions on radiographs
- Hand-made lists of equivalent terms
- Negation detection, including special terms, e.g., “RF-”

**Table 3. Variables selected for the complete algorithm (narrative and codified EMR data) from the logistic regression in order of predictive value\***

Variable	Standardized regression coefficient	Standard error
Positive predictors		
NLP RA	1.11	0.48
NLP seropositive	0.74	0.26
ICD-9 RA normalized†	0.71	0.23
ICD-9 RA	0.66	0.44
NLP erosions	0.46	0.29
Codified RF negative	0.36	0.36
NLP methotrexate	0.3	0.34
Codified anti-TNF‡	0.29	0.3
NLP anti-CCP positive	0.27	0.25
NLP anti-TNF§	0.2	0.36
NLP other DMARDs	0.13	0.34
Negative predictors		
ICD-9 JRA	−0.98	0.9
ICD-9 SLE	−0.57	1.09
NLP PsA	−0.51	0.74

\* EMR = electronic medical record; NLP = natural language processing; RA = rheumatoid arthritis; ICD-9 = International Classification of Diseases, Ninth Revision; RF = rheumatoid factor; anti-TNF = anti-tumor necrosis factor; anti-CCP = anti-cyclic citrullinated peptide; DMARDs = disease-modifying antirheumatic drugs; JRA = juvenile rheumatoid arthritis; SLE = systemic lupus erythematosus; PsA = psoriatic arthritis.

† ICD-9 RA normalized =  $\ln$  (no. of ICD-9 RA codes per subject  $\geq 1$  week apart).

‡ Codified anti-TNF = etanercept and infliximab (adalimumab was not available in our EMR).

§ NLP anti-TNF = adalimumab, etanercept, and infliximab.

# Algorithm for RA was Portable (!)

---

- Study replicated at Vanderbilt and Northwestern

	Partners	Northwestern	Vanderbilt
EHR	Local	Epic (inpatient) Cerner (outpatient)	Local
# Patients	4M	2.2M	1.7M
Meds	Structured meds entries (in- and outpatient) and text queries	Structured outpatient meds entries and in- and outpatient text queries	NLP (MedEx) for outpatient medications and structured inpatient records
NLP Queries	Custom RegEx	Custom RegEx from Partners	Generic UMLS concepts, derived from KnowledgeMap web interface

Carroll, R. J., Thompson, W. K., Eyler, A. E., Mandelin, A. M., Cai, T., Zink, R. M., et al. (2012). Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*, 19(e1), e162–9. <http://doi.org/10.1136/amiajnl-2011-000583>



**Table 3** Model performance

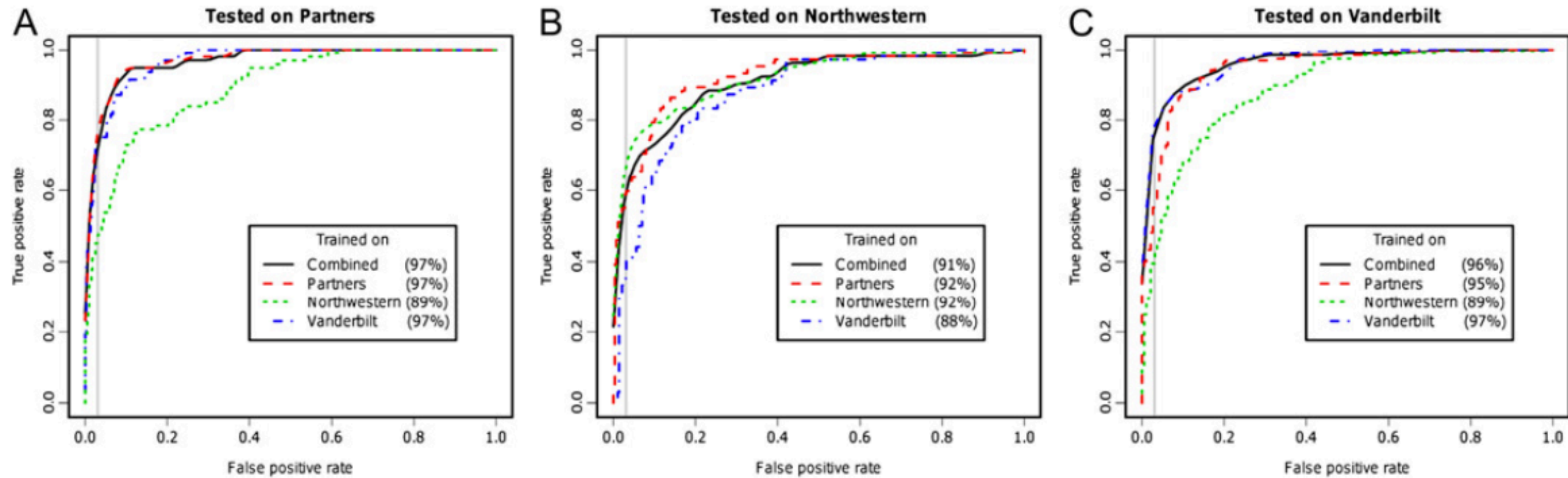
Algorithm	Testing set											
	Partners			Northwestern			Vanderbilt			Average		
	PPV	Sensitivity	AUC	PPV	Sensitivity	AUC	PPV	Sensitivity	AUC	PPV	Sensitivity	AUC
Published algorithm	88%*	79%*	97%*	87%	60%	92%	95%	57%	95%	90%	65%	95%
Retrained with												
Northwestern	79%	47%	89%	87%	73%	92%	93%	43%	89%	86%	54%	90%
Vanderbilt	85%	74%	97%	82%	40%	88%	97%	81%	97%	88%	65%	94%
Combined	86%	71%	97%	86%	65%	91%	97%	82%	96%	90%	72%	95%
ICD-9 only†												
≥1 RA code	22%	97%	N/A	26%	100%	N/A	49%	100%	N/A	33%	99%	N/A
≥3 RA code	55%	81%	N/A	42%	87%	N/A	73%	98%	N/A	57%	89%	N/A
97% Specificity	80%	49%	88%	80%	36%	84%	93%	43%	93%	84%	43%	88%
Code count for 97% specificity	53			29			48			43.3		

The PPV and sensitivity values reported represent model performance with a specificity set at 97% for logistic regression models.

\*These results are from a fivefold cross-validation on the Partners training set. The PPV and sensitivity as published in Liao *et al* was calculated from a separate Partners validation set (PPV 94%, sensitivity 63%).

†ICD-9 cut-off used the count of 714.\* codes, excluding codes for juvenile RA (714.3\*).

AUC, area under the receiver operating characteristic curve; ICD-9, International Classification of Diseases, version 9 CM; PPV, positive predictive value; RA, rheumatoid arthritis.



**Figure 3** Receiver operating characteristic curves for each test set. The vertical line represents the 97% specificity cut-off used in this study. The test performance at Partners, Northwestern, and Vanderbilt are found in (a), (b), and (c), respectively.

# Warning: Telegraphic Language

(Barrows00)

---

3/11/98 IPN	
SOB & DOE ↓	
VSS, AF	
CXR ⊕ LLL ASD no Δ	
WBC 11K	
S/B Cx ⊕ GPC c/w PC, no GNR	
D/C Cef → PCN IV	

# Telegraphic Language

---

3/11/98 IPN	(date of) Intern Progress Note,
SOB & DOE ↓	the patient's shortness of breath and dyspnea on exertion are decreased,
VSS, AF	the patient's vital signs are stable and the patient is afebrile,
CXR ⊕ LLL ASD no Δ	a recent new chest xray shows a left lower lobe air space density that is unchanged from the previous radiograph,
WBC 11K	a recent new white blood cell count is 11,000 cells per cubic milliliter,
S/B Cx ⊕ GPC c/w PC, no GNR	the patient's sputum and blood cultures are positive for gram positive cocci consistent with pneumococcus, no gram negative rods have grown,
D/C Cef →PCN IV	so the plan is to discontinue the cefazolin and then begin penicillin treatment intravenously.

# Typical Goals of MNLP

---

- for any word or phrase, assign it a meaning (or null) from some taxonomy/ontology/terminology;
  - e.g., “rheumatoid arthritis” ==> 714.0 (ICD9)
- for any word or phrase, determine whether it represents protected health information;
  - e.g., “Mr. Huntington suffers from Huntington’s Disease”
- determine aspects of each entity: time, location, certainty, ...
- having identified two meaningful phrases in a sentence, determine the relationship (or null) between them;
  - e.g., precedes, causes, treats, prevents, indicates, ...
  - note: we also need a taxonomy of relationships
- in a larger document, identify the sentences or fragments most relevant to answering a specific medical question;
  - e.g., where is the patient’s exercise regimen discussed?
- summarization
  - as data sets balloon in size, how to provide a meaningful overview



# Two Types of Tasks

---

- Every word counts
  - De-identification
  - Extraction of all
    - entities
    - time
    - certainty
    - causation and association
- Aggregate judgment
  - E.g., “smoking” challenge
    - Most text may be irrelevant to specific result
  - Cohort selection—does a patient satisfy some set of inclusion and exclusion criteria
    - Often definite presence of a disease, complication, ...

# Outline

---

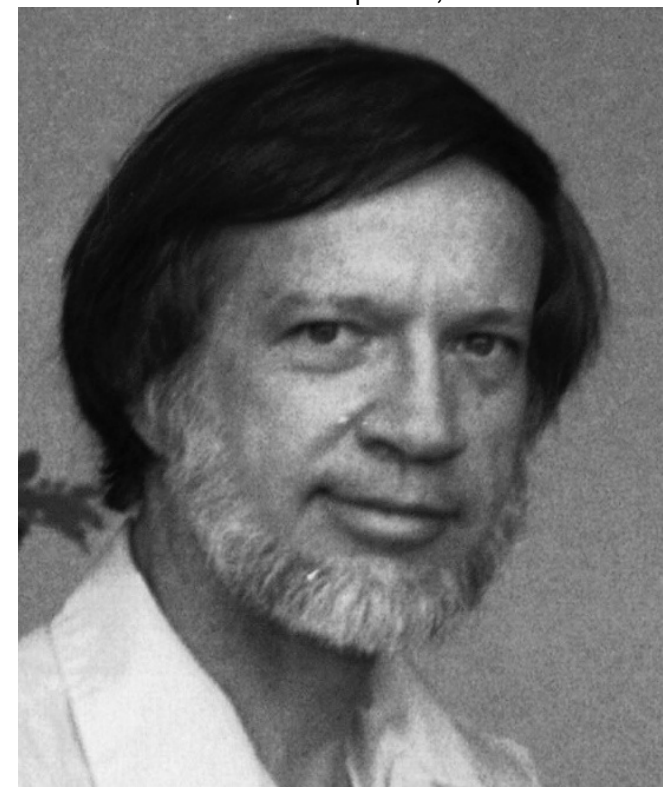
- Value of the data in clinical text
- **Hyper-simplified linguistics**
- Term spotting + handling negation, uncertainty
- ML to expand terms
- pre-NN ML to identify entities and relations
- language models
- Neural methods

# Historical Thought ...

---

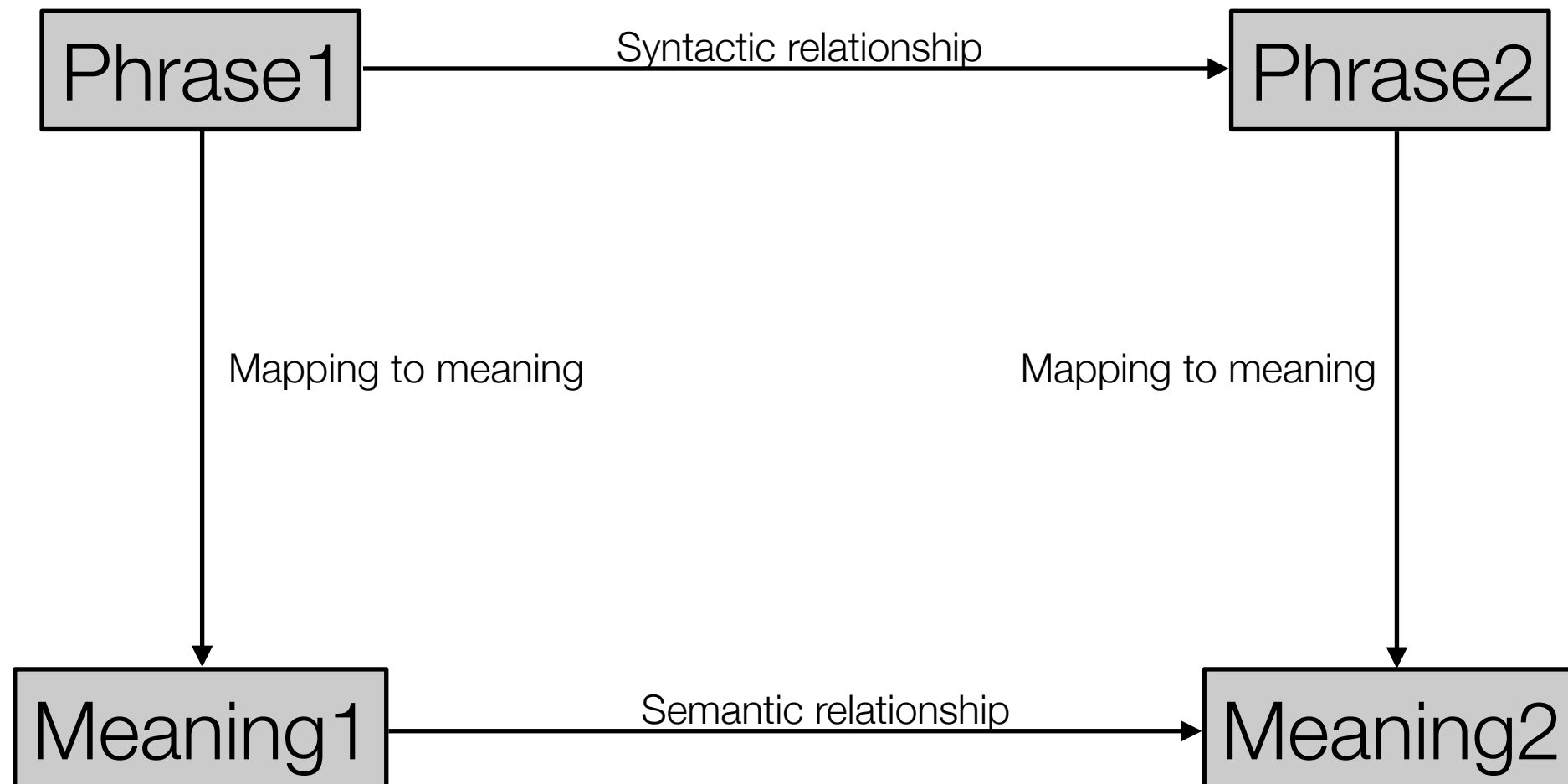
- Frederick B. Thompson, “English for the Computer.” *Proceedings of the Fall Joint Computer Conference* (1966) pp. 349-356
- Grammar defined by context-sensitive production rules + transformations
- Semantics defined by mappings:
  - Each grammar rule matches a semantic function
  - Terminal symbols are *referents* or *functions*
  - An environment is (in modern terms) a semantic network of complex interrelationships
  - Meaning is compositional, in terms of the semantic functions
- *Minor* 😏 remaining question: how to represent the “real world”?

Fred Thompson, ~1973



# Proposed relationship between syntax and semantics

---



# Formal language semantics

---

- SRI's DIAMOND/DIAGRAM system (~1980)
- each passage is expressed as a proposition or a conjunction of propositions:
  - a particular procedure for the prevention of hepatitis B could have associated with it the proposition "immunize(GAMMA-GLOBULIN,HEPATITIS-B)"
  - a passage concerned with the etiology of the disease could have the proposition "transmit(TRANSFUSION,HEPATITIS-B)"
  - synonym and hyponym relations
  - ... *a language of primitives for the domain*
- French Remède system
  - “medical documentary language using current medical terms and few syntactic rules”
  - taught to doctors to write notes
  - ... *not popular*

Walker, D. E., Hobbs, J. R., 1981. Natural Language Access to Medical Text\*. (pp. 269–273). Presented at the Proc Annu Symp Comput Appl Med Care.

de Heaulme M, Tainturier C, Thomas D. [Computer treatment of medical reports: example of the "Remède" system (author's transl)]. Nouv Presse Med. 1979 Oct 22;8(40):3223-6. French. PubMed PMID: 534182



# Outline

---

- Value of the data in clinical text
- Hyper-simplified linguistics
- **Term spotting + handling negation, uncertainty**
- ML to expand terms
- pre-NN ML to identify entities and relations
- language models
- Neural methods

# Term Spotting

---

- Traditionally, lists of coded items, narrative terms and patterns hand-crafted by researcher
- Negation and uncertainty handled by somewhat ad-hoc methods
  - NegEx is widely used,  $\exists$  many more sophisticated variants
- Generalize terms
  - Manually or automatically identify high-certainty “anchors”
  - Learn related terms to augment the set of terms
    - From knowledge bases such as UMLS
    - From co-occurrence in EMR data
    - From co-occurrence in publications

# Negation

---

- “Identifying pertinent negatives, then, involves identifying a proposition ascribing a clinical condition to a person and determining whether the proposition is denied or negated in the text.”
- Simpler than general problem of negation in NLP because negation applies mostly to noun phrases indicating diseases, tests, drugs, findings, ...
- NegEx
  - Find all UMLS terms in each sentence of a discharge summary
    - “The patient denied experiencing chest pain on exertion”  $\Rightarrow$  “The patient denied experiencing S1459038 on exertion”
  - Find patterns
    - <negation phrase>  $\ast\{0,5\}$  <UMLS term>
      - "no signs of", "ruled out unlikely", "absence of", "not demonstrated", "denies", "no sign of", "no evidence of", "no", "denied", "without", "negative for", "not", "doubt", "versus"
    - <UMLS term>  $\ast\{0,5\}$  <negation phrase>
      - “declined”, “unlikely”
  - Pseudo-negation: "gram negative", "no further", "not able to be", "not certain if", "not certain whether", "not necessarily", "not rule out", "without any further", "without difficulty", "without further"

# NegEx results

---

- Baseline:
  - <negation phrase> \* <UMLS term>
    - "no", "denies", "not", "without", "\*n't", "ruled out", "denied"

	Baseline			NegEx		
	Group 1 sentences (i.e. containing NegEx negation phrases)	Group 2 sentences (i.e., not containing NegEx negation phrases)	All sentences	Group 1 sentences (i.e. containing NegEx negation phrases)	Group 2 sentences (i.e., not containing NegEx negation phrases)	All sentences
n	500	500	1000	500	500	1000
Sensitivity	88.27	0.00	<b>88.27</b>	82.31	0.00	77.84
Specificity	52.69	100.00	85.27	82.50	100.00	<b>94.51</b>
PPV	68.42	—	68.42	84.49	—	<b>84.49</b>
NPV	79.46	96.99	<b>93.01</b>	80.21	96.99	91.73

- Extremely simplistic schemes (kind of) work

# Generalize Terms

---

- Use synonymous terms as well as the starting ones
- Take advantage of others related terms
  - hypo- or hypernyms
  - other associated terms
    - e.g., common symptoms or treatments of a disease
- Recursive ML problem: learn how best to identify cases associated with a term
  - “phenotyping”



# Available Classification Thesauri

## Most Available through UMLS

---

- Unified Medical Language Systems project of NLM; since ~1985
- *Metathesaurus* now (2018ab version) includes 161 source vocabularies
  - MeSH, SNOMED, ICD-9, ICD-10, LOINC, RxNORM, CPT, GO, DXPLAIN, OMIM, ...
- Synonym mappings across vocabularies;
  - e.g., “heart attack” = “acute myocardial infarct” = “myocardial infarction” ...
  - 3,773,462 distinct concepts, represented by concept unique identifier (CUI)
- Jumbled compendium of every hierarchy drawn from every source
- *Semantic Network*
  - Hierarchy of
    - 54 relations
    - 127 types
  - Every CUI assigned  $\geq 1$  semantic type

# Wealth of UMLS Concepts of Various Types

```
mysql> select tui,sty,count(*) c from mrsty group by sty
order by c desc;
```

tui	sty	c
T061	Therapeutic or Preventive Procedure	260914
T033	Finding	233579
T200	Clinical Drug	172069
T109	Organic Chemical	157901
T121	Pharmacologic Substance	124844
T116	Amino Acid, Peptide, or Protein	117508
T009	Invertebrate	111044
T007	Bacterium	110065
T002	Plant	95017
T047	Disease or Syndrome	79370
T023	Body Part, Organ, or Organ Component	73402
T201	Clinical Attribute	60998
T123	Biologically Active Substance	55741
T074	Medical Device	51708
T028	Gene or Genome	49960
T004	Fungus	47291
T060	Diagnostic Procedure	46106
T037	Injury or Poisoning	43924
T191	Neoplastic Process	33539
T044	Molecular Function	31369
T126	Enzyme	25766
T129	Immunologic Factor	25025
T059	Laboratory Procedure	24511
T058	Health Care Activity	19552
T029	Body Location or Region	16470
T013	Fish	16059
T046	Pathologic Function	13562
T184	Sign or Symptom	13299
T130	Indicator, Reagent, or Diagnostic Aid	12809
T170	Intellectual Product	12544
T118	Carbohydrate	10722
T110	Steroid	10363
T012	Bird	9908
T043	Cell Function	9758

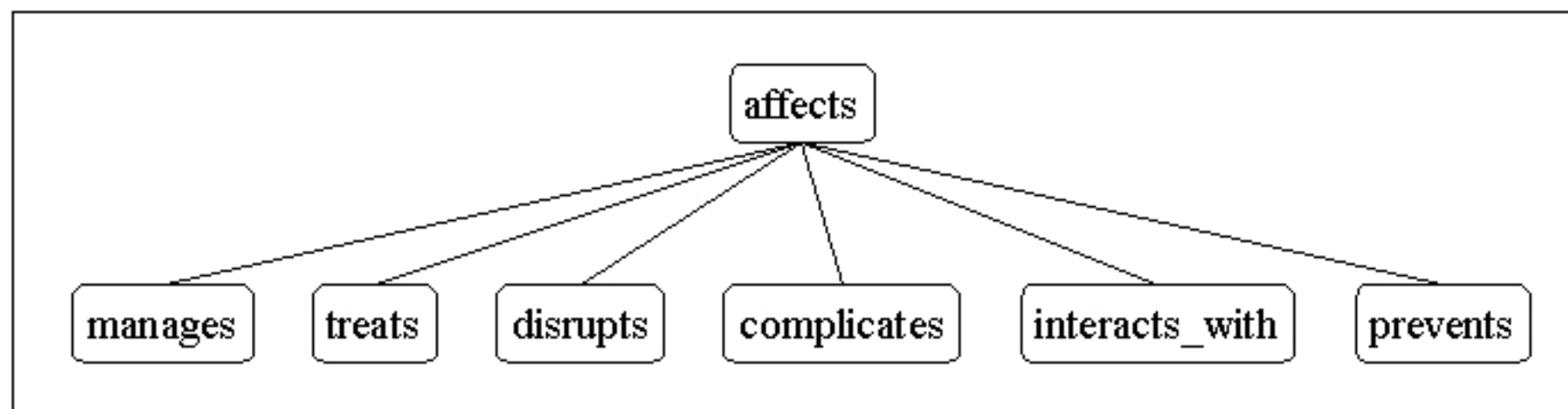
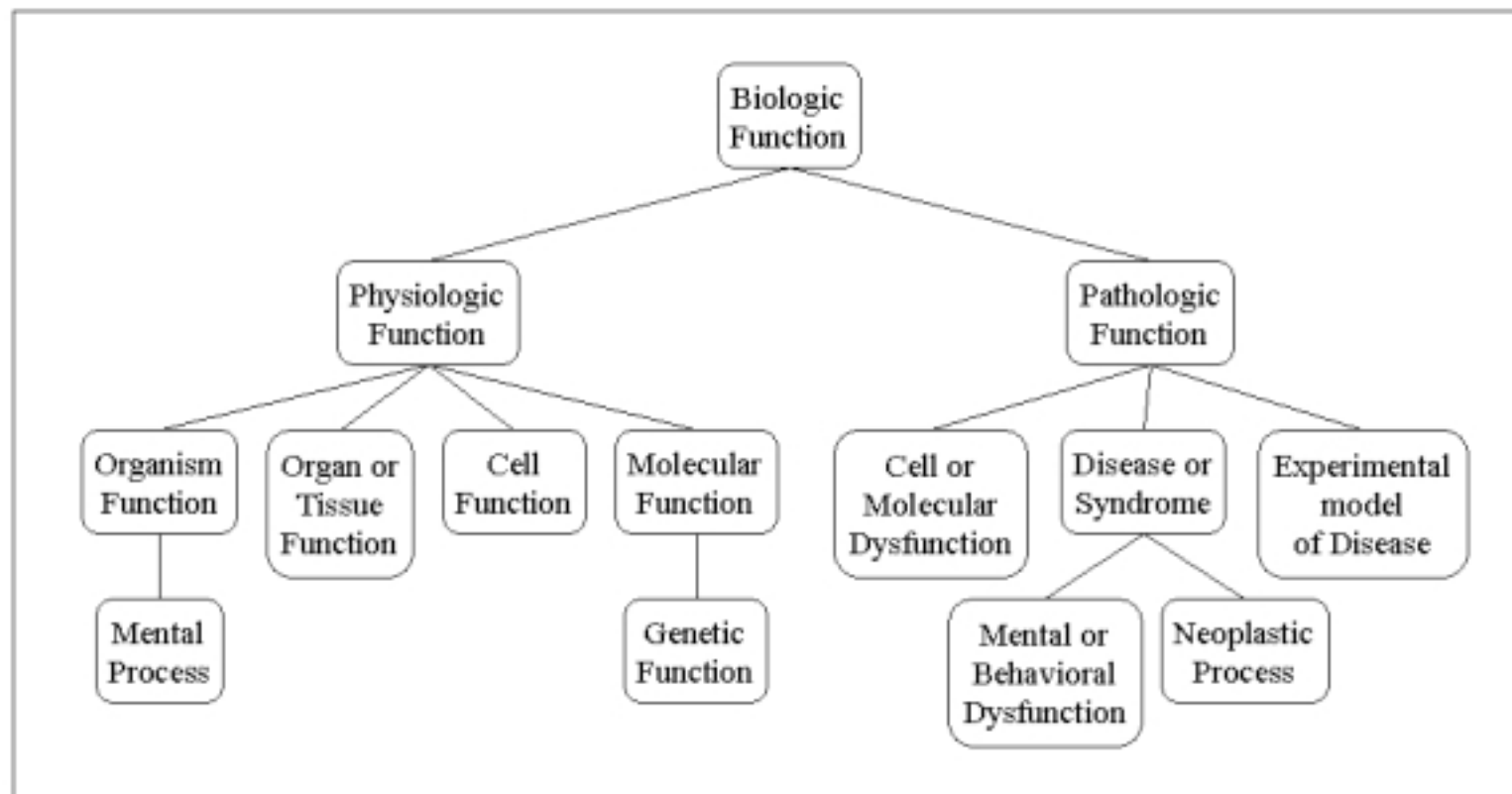
...

```
select c.cui,c.str from mrconso c join mrsty s on c.cui=s.cui
where c.TS='P' and c.STT='PF' and c.ISPREF='Y' and
c.LAT='ENG' and s.tui='T047';
```

cui	str
C0000744	Abetalipoproteinemia
C0000774	Gastrin secretion abnormality NOS
C0000786	Spontaneous abortion
C0000809	Abortion, Habitual
C0000814	Missed abortion
C0000821	Threatened abortion
C0000822	Abortion, Tubal
C0000823	Abortion, Veterinary
C0000832	Abruptio Placentae
C0000880	Acanthamoeba Keratitis
C0000889	Acanthosis Nigricans
C0001080	Achondroplasia
C0001083	Achromia parasitica
C0001125	Acidosis, Lactic
C0001126	Renal tubular acidosis
C0001127	Acidosis, Respiratory
C0001139	Acinetobacter Infections
C0001142	Acladiosis
C0001144	Acne Vulgaris
C0001145	Acne Keloid
C0001163	Vestibulocochlear Nerve Diseases
C0001168	Complete obstruction
C0001169	Acquired coagulation factor deficiency NOS
C0001175	Acquired Immunodeficiency Syndrome
C0001197	Acrodermatitis
C0001202	Acrokeratosis
C0001206	Acromegaly
C0001207	Hypersomatotropic gigantism
C0001231	ACTH Syndrome, Ectopic
C0001247	Actinobacillosis

...

# Hierarchy of UMLS Semantic Network Types and Relations



# Lexical Variant Generation (LVG) Tools

(from National Library of Medicine)

---

- Normalized words and phrases used as index to UMLS
- Lemmatization of words
  - stripping typical prefixes, suffixes
    - plurals, in-word negation, gerunds
- Discarding “noise” words, punctuation
- Lower-casing
- Alphabetic order of all remaining words

Mr. Huntington was admitted to Huntington Memorial Hospital for acute chest pain in March.

Mr. Huntington was admitted to Huntington Memorial Hospital for acute chest pain in March.|acute admit be chest hospital huntington huntington march memorial mr pain

Mr. Huntington was admitted to Huntington Memorial Hospital for acute chest pain in March.|acute admit chest hospital huntington huntington march memorial mr pain was

Mr. Huntington was admitted to Huntington Memorial Hospital for acute chest pain in March.|acute admitted be chest hospital huntington huntington march memorial mr pain

Mr. Huntington was admitted to Huntington Memorial Hospital for acute chest pain in March.|acute admitted chest hospital huntington huntington march memorial mr pain was

Weakness of the upper extremities

Weakness of the upper extremities|extremity upper weakness



# UMLS Terminology Services

## Metathesaurus Browser

Welcome back,  
pszolovits

SearchTreeRecent Searches

☒ Term ☐ CUI ☐ Code

?

Go

Release:

2014AA

Search Type:

Word

Source:

All Sources

AIR

ALT

AOD

AOT

Search Results (1)

[C2750237](#) Proximal weakness, upper extremities (1 pat

Basic View

Report View

Raw View

?

Print

Link

+

Concept: [C2750237] Proximal weakness, upper extremities (1 patient)

-

Semantic Types

Finding [T033]

-

Atoms (1) string [AUI / RSAB / TTY / Code]

-

Proximal weakness, upper extremities (1 patient) [A17467681/OMIM/PTCS/MTHU025233]

-

Relations (2) REL | RELA | RSAB [SType1 - SType2] STypeId | String | CUI

CHD | | OMIM [ATOM - ATOM] A11965599 | Peripheral nervous system | [C0206417](#)

RO | manifestation\_of | OMIM [ATOM - ATOM] A11922722 | HEART-HAND SYNDROME

-

Contexts (1)

-

OMIM/PTCS/Proximal weakness, upper extremities (1 patient) (1)

-

MTHU025233/Proximal weakness, upper extremities (1 patient) [Context 1]

-

Ancestors

-

[Online Mendelian Inheritance in Man](#)


-

[NEUROLOGIC](#)

[Peripheral nervous system](#)

-

Siblings (750)

[ : 1 - 10 : 'Onion bulb' formation on nerve biopsy

['Onion bulb' formations](#)

['Onion bulb' formations \(rare\)](#)



