NLP

Feb 28, 2019
March 5, 2019

Outline

- Term spotting + handling negation, uncertainty
- ML to expand terms
- pre-NN ML to identify entities and relations
- language models
- Neural methods

Learning what features to use in term spotting

Electronic medical record phenotyping using the anchor and learn framework, using ED data

- Identify “anchors” using domain expertise
  - High PPV; not necessarily high sensitivity
  - Conditionally dependent only on phenotype
- Learn (using L2-regularized LR) to predict whether the anchor is present from the rest of the patient’s data
  - Binning continuous variables using breaks found in a decision tree
  - Narratives represented as bag-of-word + “significant bigrams” after negation detection
  - Odd trick: censor text within 3 words of anchor to avoid dependence
  - Estimate a calibration score
- Build phenotype estimators from the anchors + chosen predictors
  - Presence of anchor is assumed to indicate certain phenotype
  - Other predictors are scaled by their calibration score from predicting anchors
  - Supervision from judgments of ED docs

Halpern, Y., Choi, Y., Horng, S., & Sontag, D. (2014). Using Anchors to Estimate Clinical State without Labeled Data. Presented at the Proc. AMIA Symposium.

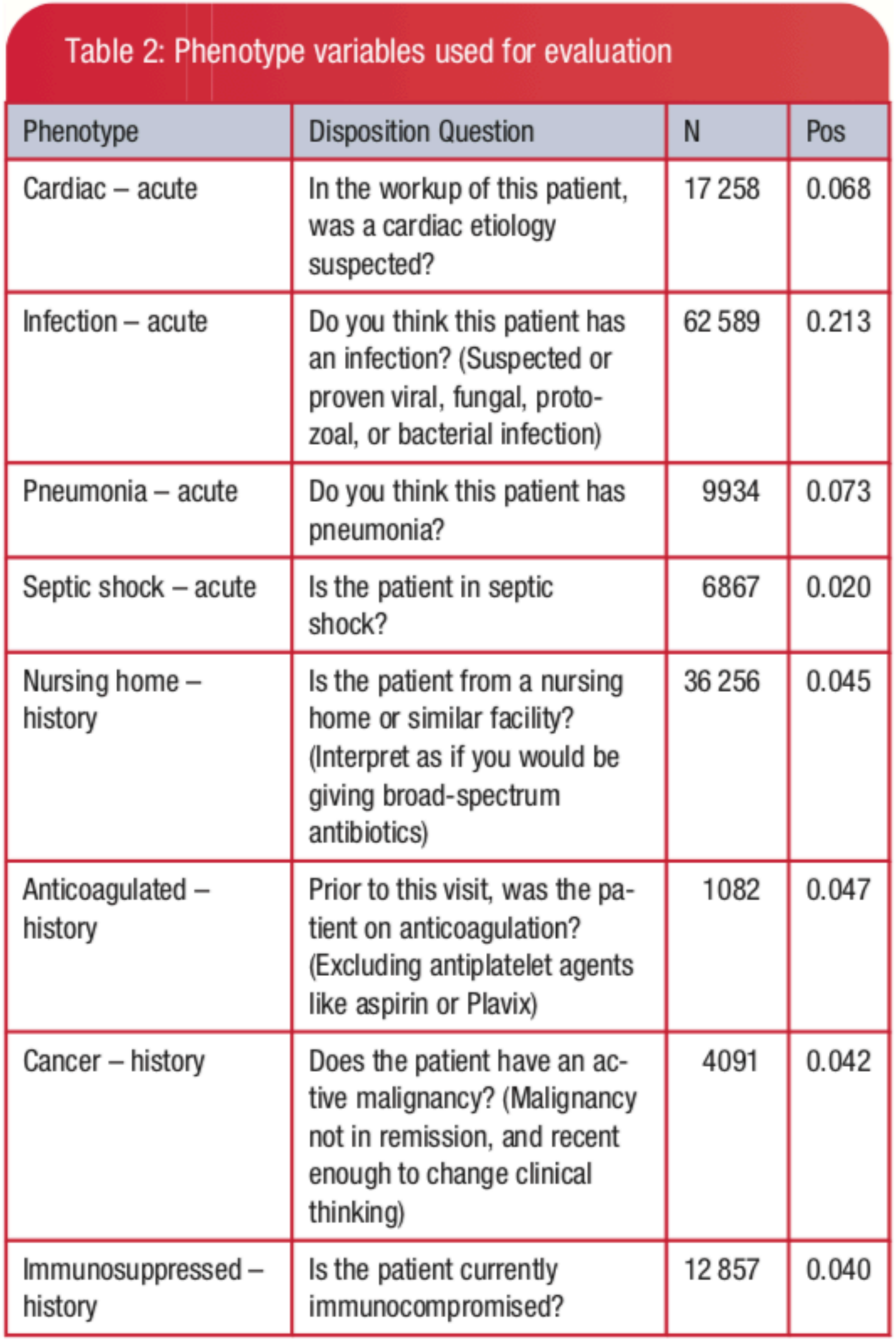
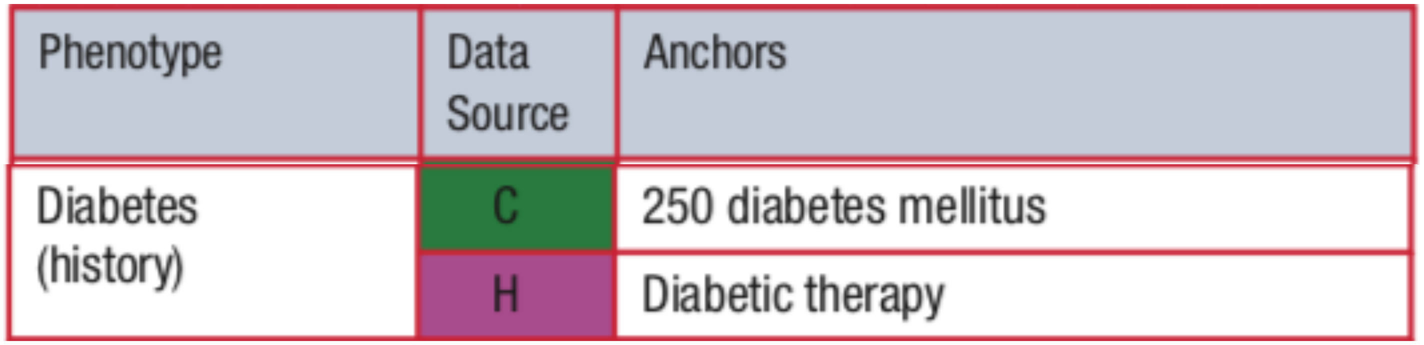


Table 2: Phenotype variables used for evaluation

Phenotype	Disposition Question	N	Pos	AUC
Cardiac – acute	In the workup of this patient, was a cardiac etiology suspected?	17 258	0.068	0.89
Infection – acute	Do you think this patient has an infection? (Suspected or proven viral, fungal, protozoal, or bacterial infection)	62 589	0.213	0.89
Pneumonia – acute	Do you think this patient has pneumonia?	9934	0.073	0,90
Septic shock – acute	Is the patient in septic shock?	6867	0.020	0.93
Nursing home – history	Is the patient from a nursing home or similar facility? (Interpret as if you would be giving broad-spectrum antibiotics)	36 256	0.045	0.87
Anticoagulated – history	Prior to this visit, was the patient on anticoagulation? (Excluding antiplatelet agents like aspirin or Plavix)	1082	0.047	0.83
Cancer – history	Does the patient have an active malignancy? (Malignancy not in remission, and recent enough to change clinical thinking)	4091	0.042	0.95
Immunosuppressed – history	Is the patient currently immunocompromised?	12 857	0.040	0.85

Anchors



Phenotype	Data Source	Anchors
Diabetes (history)	C	250 diabetes mellitus
	H	Diabetic therapy

Predictors of Phenotype

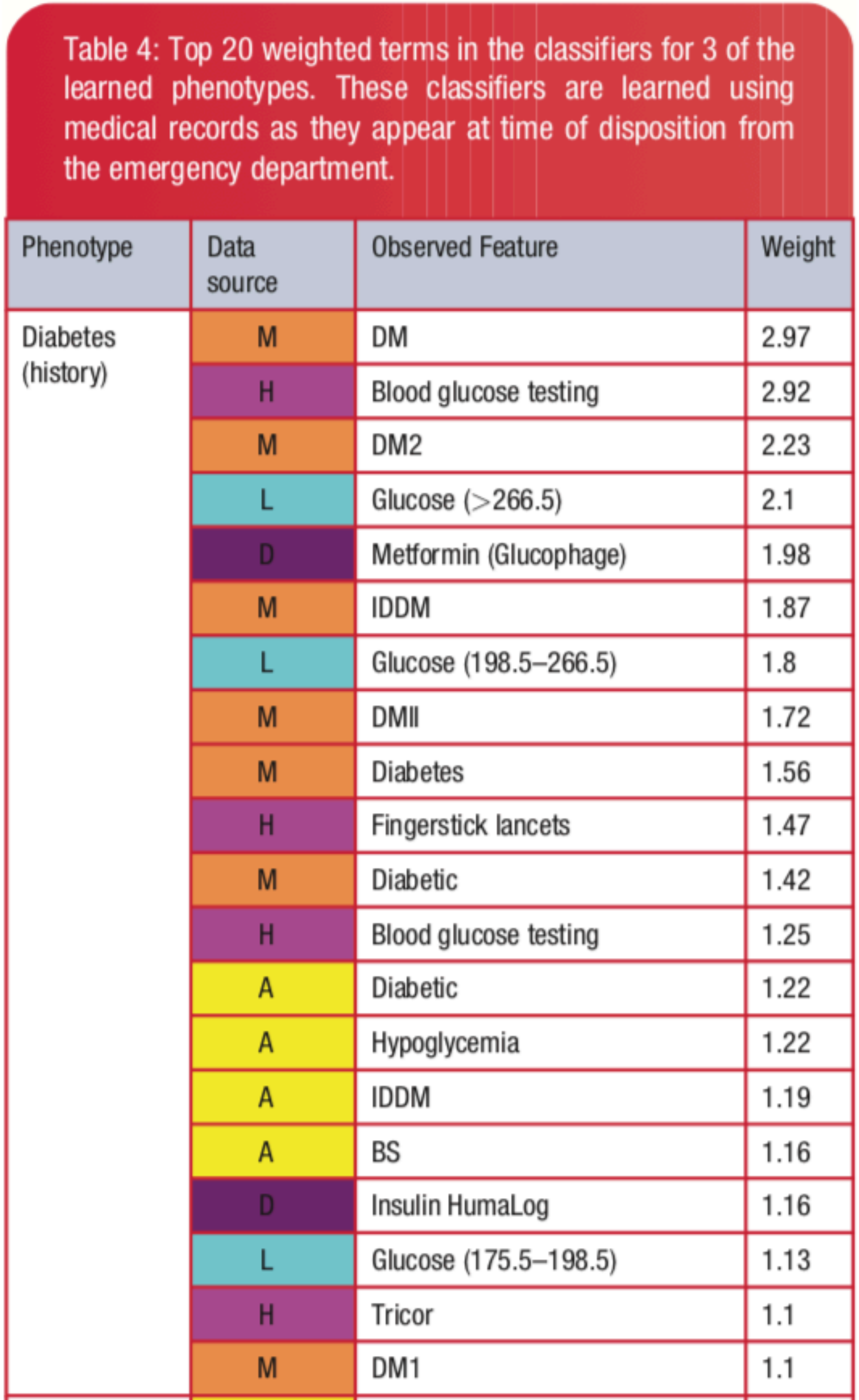
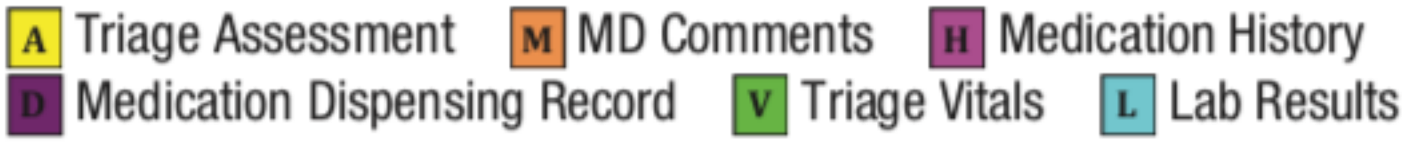


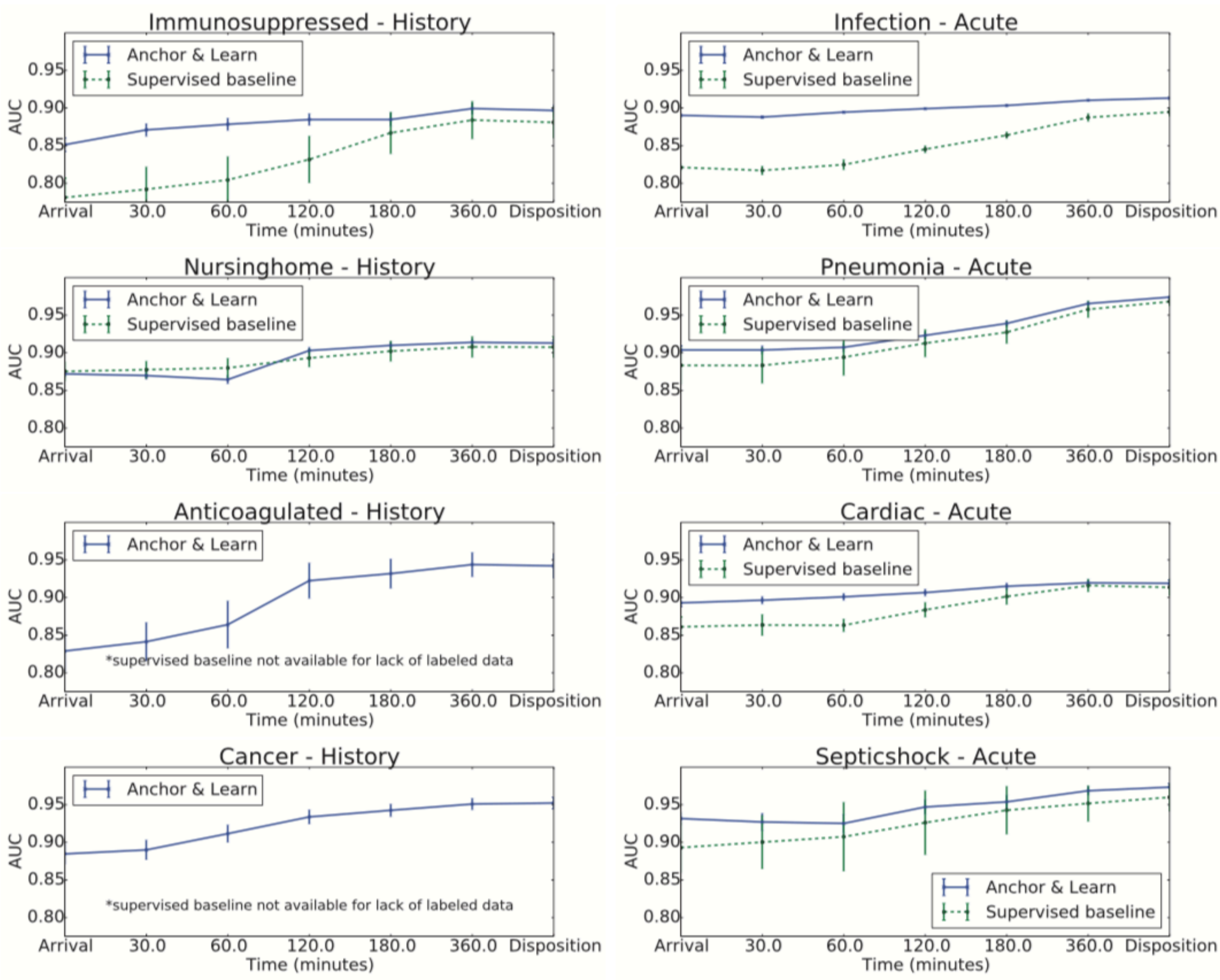
Table 4: Top 20 weighted terms in the classifiers for 3 of the learned phenotypes. These classifiers are learned using medical records as they appear at time of disposition from the emergency department.

Phenotype	Data source	Observed Feature	Weight
Diabetes (history)	M	DM	2.97
	H	Blood glucose testing	2.92
	M	DM2	2.23
	L	Glucose (>266.5)	2.1
	D	Metformin (Glucophage)	1.98
	M	IDDM	1.87
	L	Glucose (198.5–266.5)	1.8
	M	DMII	1.72
	M	Diabetes	1.56
	H	Fingerstick lancets	1.47
	M	Diabetic	1.42
	H	Blood glucose testing	1.25
	A	Diabetic	1.22
	A	Hypoglycemia	1.22
	A	IDDM	1.19
	A	BS	1.16
	D	Insulin HumaLog	1.16
	L	Glucose (175.5–198.5)	1.13
	H	Tricor	1.1
	M	DM1	1.1



A Triage Assessment M MD Comments H Medication History
D Medication Dispensing Record V Triage Vitals L Lab Results

Figure 1: Comparison of performance of phenotypes learned with 200 000 unlabeled patients using the semi-supervised anchor based method, and phenotypes learned with supervised classification using 5000 gold-standard labels. Error bars indicate 2 * standard error. For anticoagulated and cancer, there were not a sufficient number of gold-standard labels to learn with 5000 patients, so the fully supervised baseline is omitted.



The Importance of Context

- “Mr. Huntington was treated for Huntington’s Disease at Huntington Hospital, located on Huntington Avenue.”
    - Huntington
    - Huntington’s Disease
    - Mr. Huntington’s Disease
- “Atenalol was administered to Mr. Huntington.”
    - vs. “Atenalol was considered for control of heart rate.”
    - vs. “Atenalol was ineffective and therefore discontinued.”

Building Models

- Features of text from which models can be built
  - words, parts of speech, capitalization, punctuation
  - document section, conventional document structures
  - identified patterns and thesaurus terms
  - lexical context
    - ➡ all of the above, for n-tuples of words surrounding target
  - syntactic context
    - ➡ all of the above, for words syntactically related to target
    - E.g., “The lasix, started yesterday, reduced ascites ...”

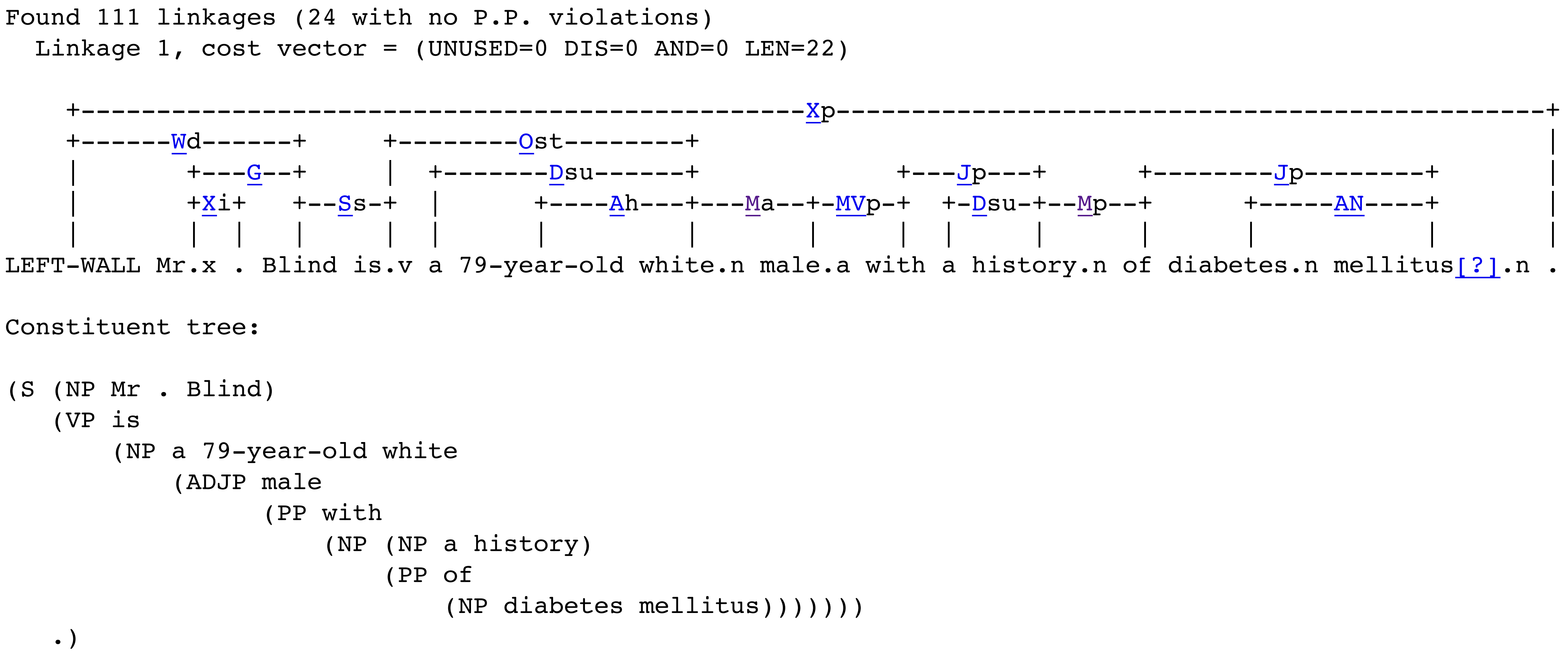
```
    +--------------------------------Xp---------------------------------+
    |          +-----------------Ss-----------------+                   |
    |          +----MXsp----+-------Xc-------+      |                   |
    +----Wd----+     +--Xd--+---MVpn---+     |      +-----Os-----+      |
    |          |     |      |          |     |      |            |      |
LEFT-WALL lasix[?].n , started.v-d yesterday , reduced.v-d ascites[?].n .
```

(Output from Link Grammar Parser, w/o special medical dictionary)

Uzuner, Ö., Sibanda, T. C., Luo, Y., & Szolovits, P. (2008). A de-identifier for medical discharge summaries. Artificial Intelligence in Medicine, 42(1), 13–35. http://doi.org/10.1016/j.artmed.2007.10.001

Parsing Can be Ambiguous

- Prepositional phrase attachment
- Part of speech
  - e.g., white.n vs. white.a
- Hope that there is enough redundancy to overcome such limitations

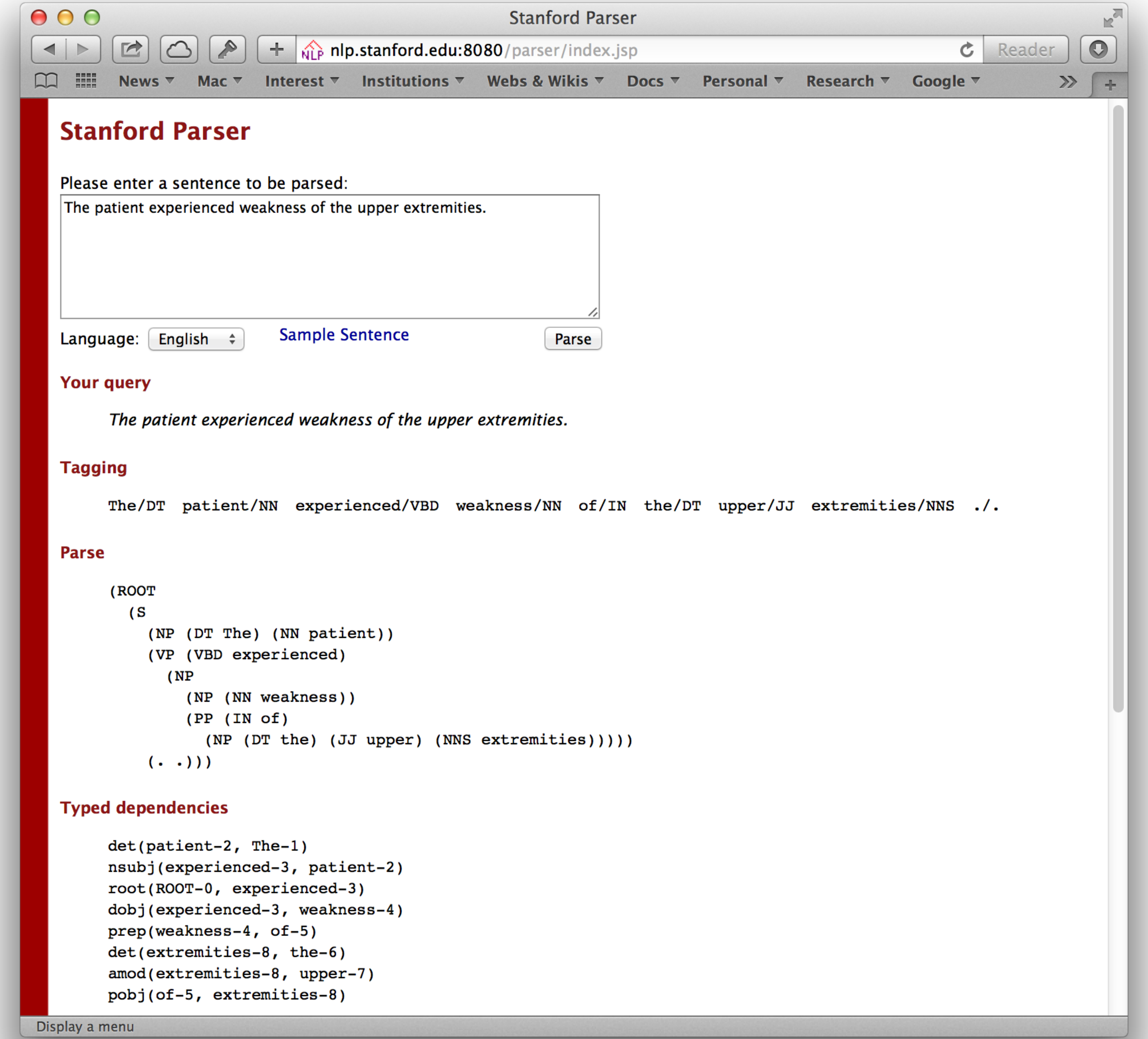


```
Found 111 linkages (24 with no P.P. violations)
  Linkage 1, cost vector = (UNUSED=0 DIS=0 AND=0 LEN=22)

    +------------------------------------------------Xp-----------------------------------------------+
    +------Wd------+     +--------Ost--------+                                                        |
    |       +---G--+     |  +-------Dsu------+             +---Jp---+      +--------Jp--------+       |
    |       +Xi+   +--Ss-+  |      +----Ah---+---Ma--+-MVp-+  +-Dsu-+--Mp--+      +-----AN----+       |
    |       |  |   |     |  |      |         |       |     |  |     |      |      |           |       |
LEFT-WALL Mr.x . Blind is.v a 79-year-old white.n male.a with a history.n of diabetes.n mellitus[?].n .

Constituent tree:

(S (NP Mr . Blind)
   (VP is
       (NP a 79-year-old white
           (ADJP male
                 (PP with
                     (NP (NP a history)
                         (PP of
                             (NP diabetes mellitus)))))))
   .)
```



Stanford Parser

Please enter a sentence to be parsed:

The patient experienced weakness of the upper extremities.

Language: English

Sample Sentence

Parse

Your query

The patient experienced weakness of the upper extremities.

Tagging

```
The/DT  patient/NN  experienced/VBD  weakness/NN  of/IN  the/DT  upper/JJ  extremities/NNS  ./.
```

Parse

```
(ROOT
  (S
    (NP (DT The) (NN patient))
    (VP (VBD experienced)
      (NP
        (NP (NN weakness))
        (PP (IN of)
          (NP (DT the) (JJ upper) (NNS extremities)))))
    (. .)))
```

Typed dependencies

```
det(patient-2, The-1)
nsubj(experienced-3, patient-2)
root(ROOT-0, experienced-3)
dobj(experienced-3, weakness-4)
prep(weakness-4, of-5)
det(extremities-8, the-6)
amod(extremities-8, upper-7)
pobj(of-5, extremities-8)
```

Example of Features Available for Model

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

```
263 266 "Mr."
   TUI: T060,T083,T047,T048,T116,T192,T081,T028,T078,T077; SP-POS: noun; SEM: _modifier,_disease,_procparam;
   CUI: C0024487,C0024943,C0025235,C0025362,C0026266,C0066563,C0311284,C0475209,C1384671,
      C1413973,C1417835,C1996908,C2347167,C2349188;   lptok: 6;
   MeSH: C07.465.466,C10.292.300.800,C10.597.606.643,C14.280.484.461,C23.888.592.604.646,D12.776.826.750.530,
   D12.776.930.682.530,E05.196.867.519,F01.700.687,F03.550.600,Z01.058.290.190.520;
267 468 "Blind is a 79-year-old white white...hsandpot Center." sent: nil;
267 272 "Blind"
   TUI: T062,T047,T170; SP-POS: verb,adj,noun; SEM: _disease; CUI: C0150108,C0456909,C1561605,C1561606;
   lptok: 1; MeSH: C10.597.751.941.162,C11.966.075,C23.888.592.763.941.162;
273 277 "is a" TUI: T185,T169,T078; SEM: _modifier; CUI: C1278569,C1292718,C1705423;
273 275 "is" SP-POS: aux,noun,adj; lptok: 2;
276 277 "a" SP-POS: det,noun,adj; lptok: 3;
278 289 "79-year-old" lptok: 4;
290 295 "white" TUI: T098,T080; SP-POS: noun,adj; SEM: _modifier; CUI: C0007457,C0043157,C0220938; lptok: 5;
296 301 "white" TUI: T098,T080; SP-POS: noun,adj; SEM: _modifier; CUI: C0007457,C0043157,C0220938; lptok: 6;
302 306 "male"
   TUI: T032,T098,T080; SP-POS: adj,noun; SEM: _modifier,_bodyparam;
   CUI: C0024554,C0086582,C1706180,C1706428,C1706429; lptok: 7;
307 311 "with" SP-POS: prep,conj; lptok: 8;
312 313 "a" SP-POS: det,noun,adj; lptok: 9;
314 342 "history of diabetes mellitus" TUI: T033; SEM: _finding; CUI: C0455488;
```

```
314 321 "history"  TUI: T090,T170,T032,T033,T080,T077; SP-POS: noun; SEM: _modifier,_finding,_bodyparam;   CUI:
C0019664,C0019665,C0262512,C0262926,C0332119,C1705255,C2004062; Iptok: 10; MeSH: K01.400,Y27;
322 324 "of" SP-POS: prep; Iptok: 11;
325 333 "diabetes"   TUI: T047; SP-POS: noun; SEM: _disease; CUI: C0011847,C0011849,C0011860; Iptok: 12;   MeSH:
C18.452.394.750,C18.452.394.750.149,C19.246,C19.246.300;
334 342 "mellitus" Iptok: 13;
342 343 "," Iptok: 14;
344 374 "inferior myocardial infarction" TUI: T047; SEM: _disease; CUI: C0340305;
344 352 "inferior" TUI: T082,T054; SP-POS: noun,adj; SEM: _modifier; CUI: C0542339,C0678975; Iptok: 15;
353 374 "myocardial infarction" TUI: T047; SEM: _disease; CUI: C0027051; MeSH: C14.280.647.500,C14.907.585.500;
353 363 "myocardial"   TUI: T024,T082; SP-POS: adj; SEM: _modifier; CUI: C0027061,C1522564; Iptok: 16;   MeSH:
A02.633.580,A07.541.704,A10.690.552.750;
364 374 "infarction"   TUI: T046; SP-POS: noun; SEM: _disease; CUI: C0021308; Iptok: 17; MeSH:
C23.550.513.355,C23.550.717.489;
374 375 "," Iptok: 18;
376 379 "who" SP-POS: pron; Iptok: 19;
380 389 "underwent" SP-POS: verb; Iptok: 20;
390 401 "open repair" TUI: T061; SEM: _procedure; CUI: C0441613;
390 394 "open" TUI: T082; SP-POS: adj,verb,adv; SEM: _modifier; CUI: C0175566,C1882151; Iptok: 21;
395 401 "repair"   TUI: T040,T169,T061,T052,T201; SP-POS: noun,verb; SEM: _finding,_procedure,_modifier,_bodyparam;
CUI: C0043240,C0205340,C0374711,C1705181,C2359963; Iptok: 22; MeSH: G16.100.856.891;
402 404 "of" SP-POS: prep; Iptok: 23;
405 408 "his" SP-POS: noun,pron; Iptok: 24;
409 418 "increased" TUI: T081,T169; SP-POS: verb,adj; SEM: _modifier; CUI: C0205217,C0442805,C0442808; Iptok: 25;
419 431 "diverticulum"   TUI: T190,T170; SP-POS: noun; SEM: _disease; CUI: C0012817,C1546602; Iptok: 26; MeSH:
C23.300.415;
```

11,146 annotations for this document of 1,518 tokens

Learning Models

- Given a target classification, build a machine learning model predicting that class
  - support vector machines (SVM)
  - classification trees
  - naive Bayes or Bayesian networks
  - artificial neural networks
  - ...
- class(word) = function (feature$_1$, feature$_2$, feature$_3$, ...)
  - sometimes, astronomically large (binary) feature set; SVM can deal with it
    - f$_1$ ... f$_{100,000}$: whether the word is “a”, “aback”, “abacus”, ..., “zymotic”
    - f$_{100,001}$ ...: whether word’s POS is “noun”, “verb”, “adj”, ...
    - f$_{100,100}$ ...: whether the word maps to CUI “C0000001”, “C0000002”, ...
    - f$_{3,000,000}$ ...: same as above, but for 1st, 2nd, 3rd word to right/left
    - f$_{6,000,000}$ ...: {lp-link, word} for 1st, 2nd, 3rd link in parse to right/left
    - ...

Using this model for de-identification

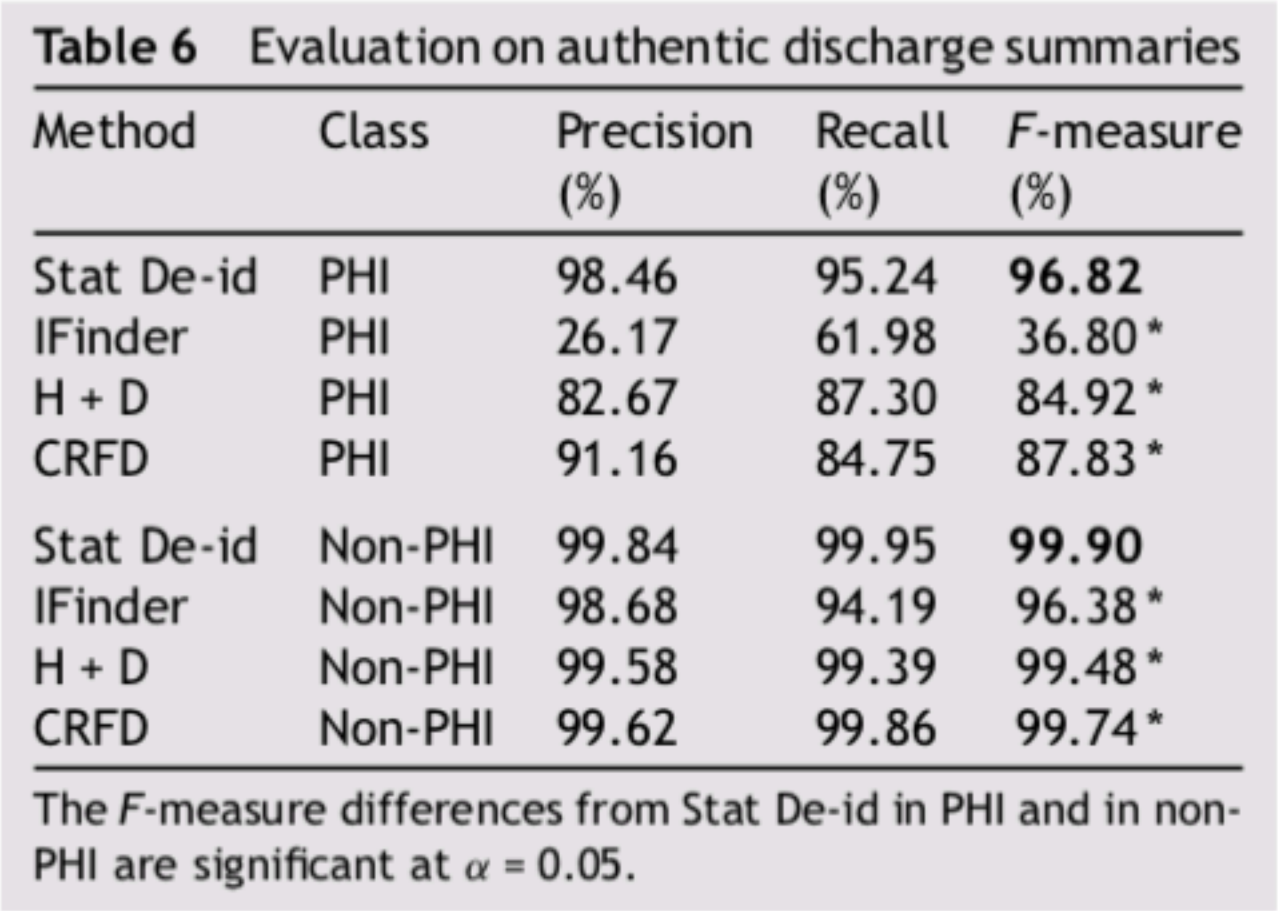


Table 6 Evaluation on authentic discharge summaries

Method	Class	Precision (%)	Recall (%)	*F*-measure (%)
Stat De-id	PHI	98.46	95.24	**96.82**
IFinder	PHI	26.17	61.98	36.80*
H + D	PHI	82.67	87.30	84.92*
CRFD	PHI	91.16	84.75	87.83*
Stat De-id	Non-PHI	99.84	99.95	**99.90**
IFinder	Non-PHI	98.68	94.19	96.38*
H + D	Non-PHI	99.58	99.39	99.48*
CRFD	Non-PHI	99.62	99.86	99.74*

The *F*-measure differences from Stat De-id in PHI and in non-PHI are significant at $\alpha = 0.05$.

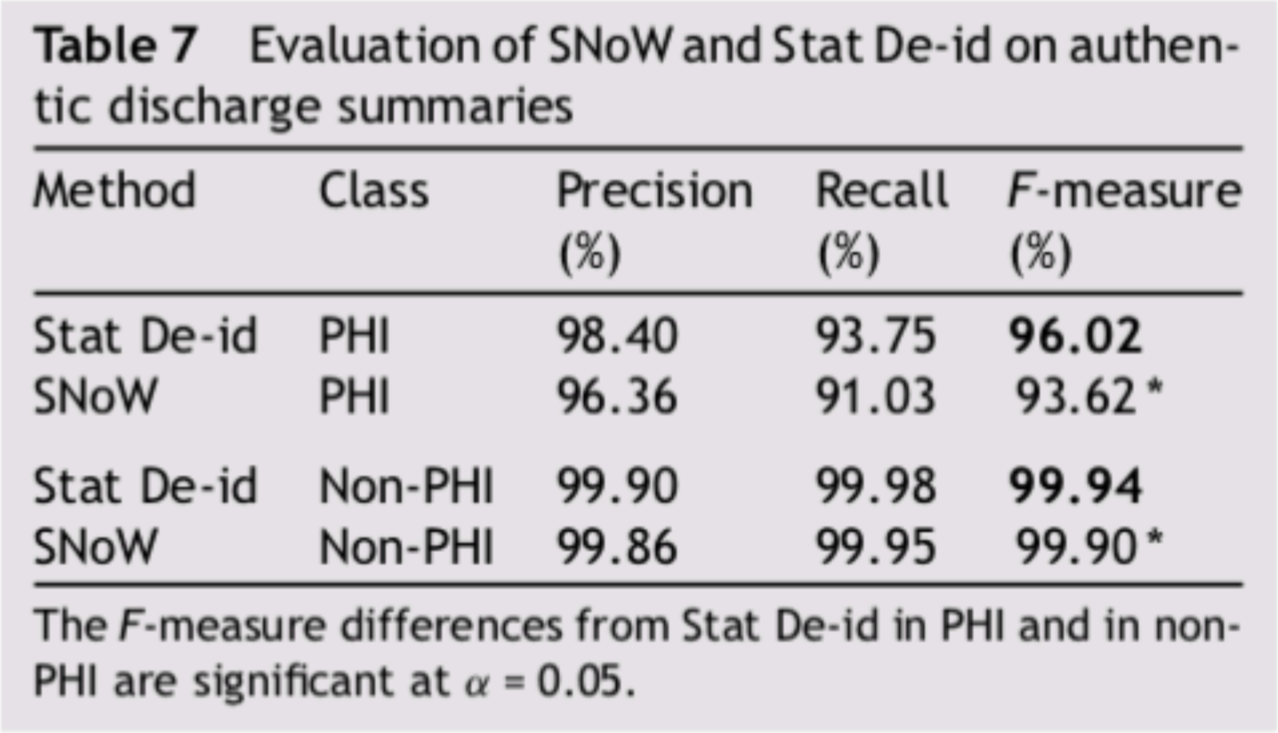


Table 7 Evaluation of SNoW and Stat De-id on authentic discharge summaries

Method	Class	Precision (%)	Recall (%)	*F*-measure (%)
Stat De-id	PHI	98.40	93.75	**96.02**
SNoW	PHI	96.36	91.03	93.62*
Stat De-id	Non-PHI	99.90	99.98	**99.94**
SNoW	Non-PHI	99.86	99.95	99.90*

The *F*-measure differences from Stat De-id in PHI and in non-PHI are significant at $\alpha = 0.05$.

Uzuner, Ö., Sibanda, T. C., Luo, Y., & Szolovits, P. (2008). A de-identifier for medical discharge summaries. Artificial Intelligence in Medicine, 42(1), 13–35. http://doi.org/10.1016/j.artmed.2007.10.001

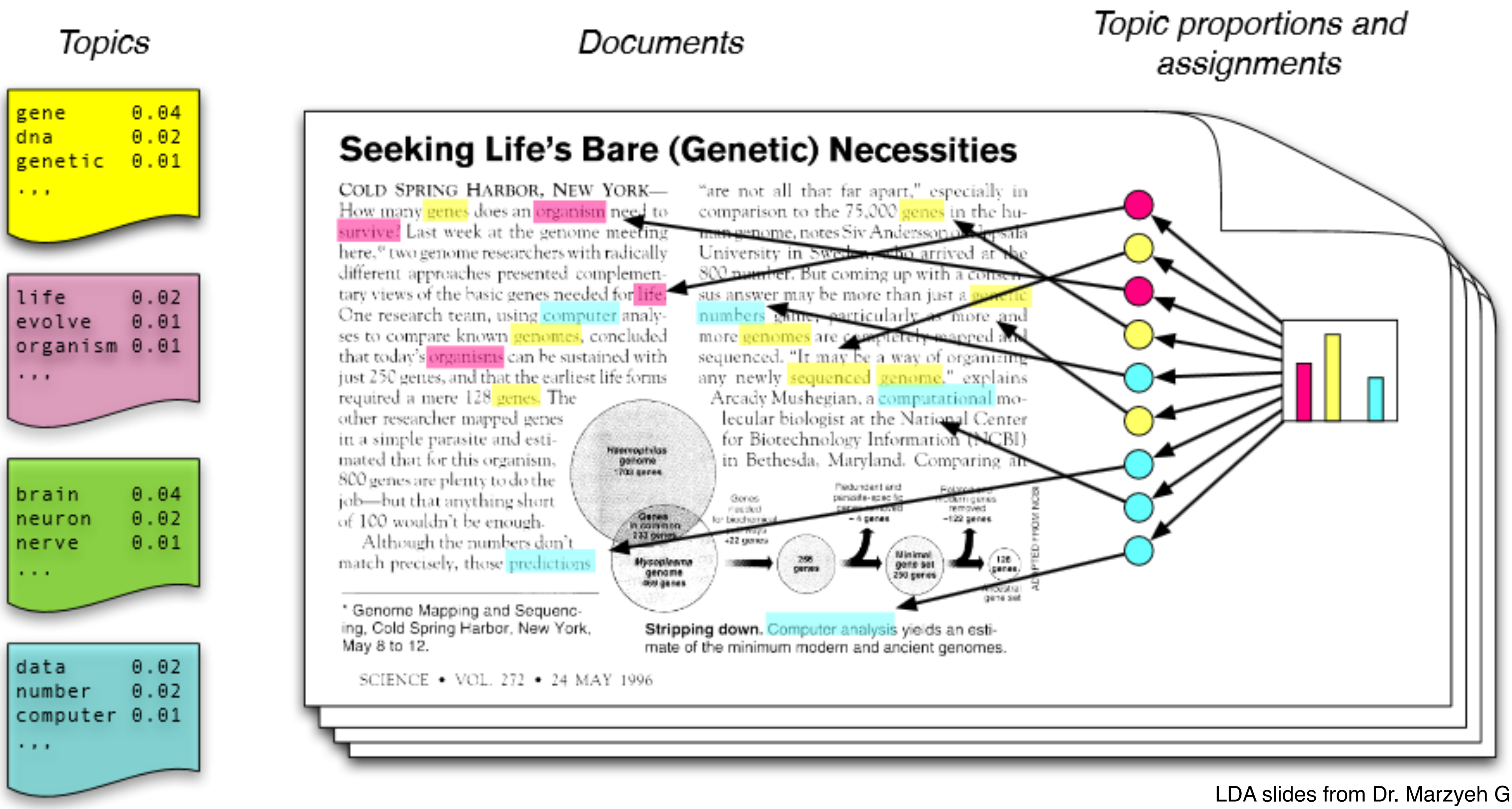
Predicting early psychiatric readmission by LDA

- Can we predict 30-day psych readmission?
- Cohort: patients admitted to a psych inpatient ward between 1994-2012 with a principal diagnosis of major depression
  - 470 of 4687 were readmitted within 30 days with a psych diagnosis; 2977 additionally were readmitted in 30 days with other diagnoses; 1240 not readmitted
- Compare predictive models built using SVM from
  - baseline clinical features
    - age, gender, public health insurance, Charlson comorbidity index
  - + common words from notes
    - 1–1000 most informative words per patient, by TF-IDF
    - top-1 used 3013 unique words, top-10 used 18 173, top-1000 use almost entire vocabulary (66 429/66 451 words)
  - + 75 topics from LDA on notes

Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V. M., McCoy, T. H., & Perlis, R. H. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Translational Psychiatry, 6(10), e921–5. http://doi.org/10.1038/tp.2015.182

Intuition: Documents are made of Topics

- Every document is a mixture of topics
- Every topic is a distribution over words
- Every word is a draw from a topic


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 468 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Functionally equivalent genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

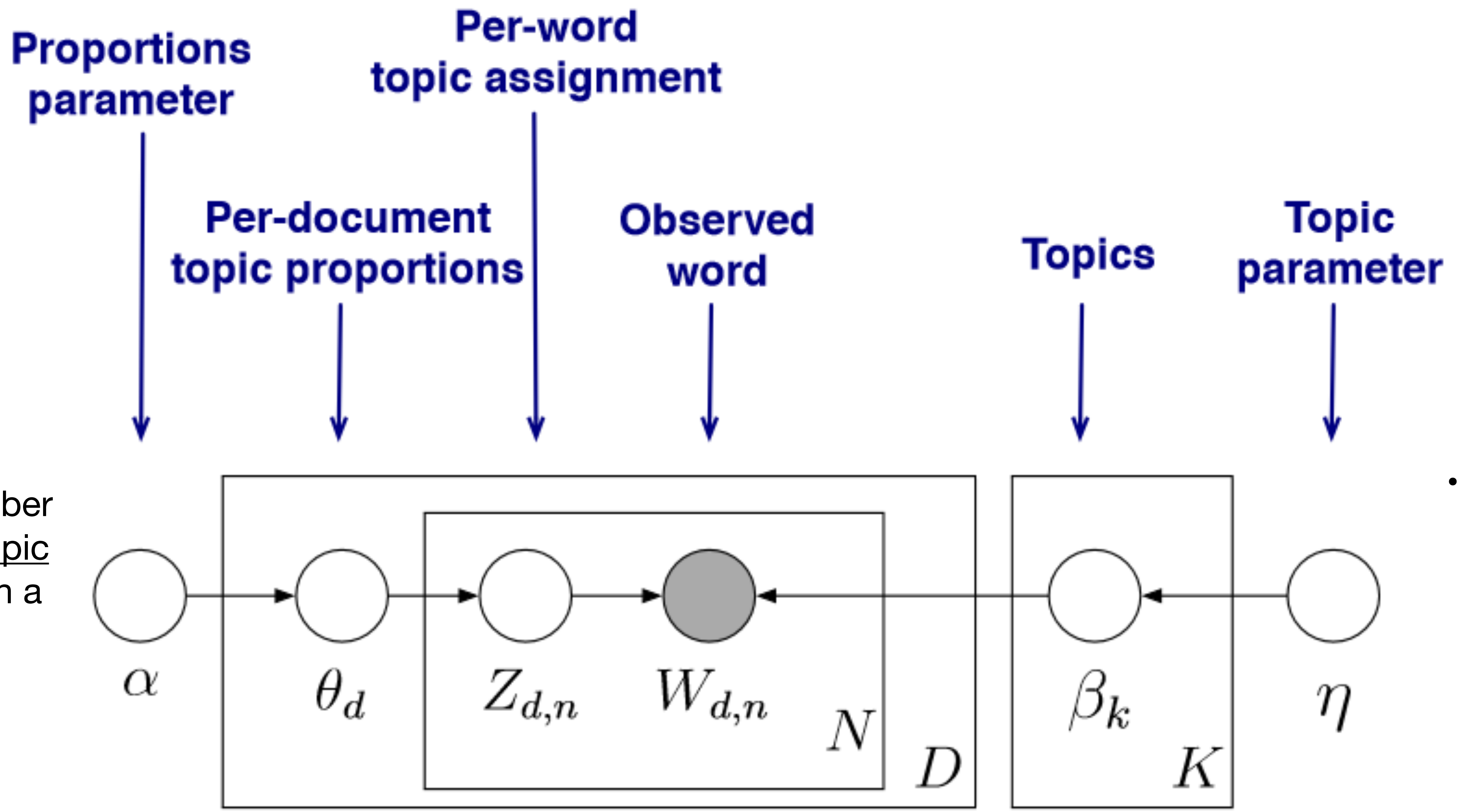
SCIENCE • VOL. 272 • 24 MAY 1996

LDA slides from Dr. Marzyeh Ghassemi

LDA – Latent Dirichlet Allocation

- We observe words, we infer everything else, with our assumed structure

- *α* is the number of times a topic is sampled in a document (prior)



- *η* is the number of times words are sampled from a topic (prior)

$$\prod_{i=1}^{K} p(\beta_i|\eta)\prod_{d=1}^{D} p(\theta_d|\alpha)\left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{1:K},z_{d,n})\right)$$

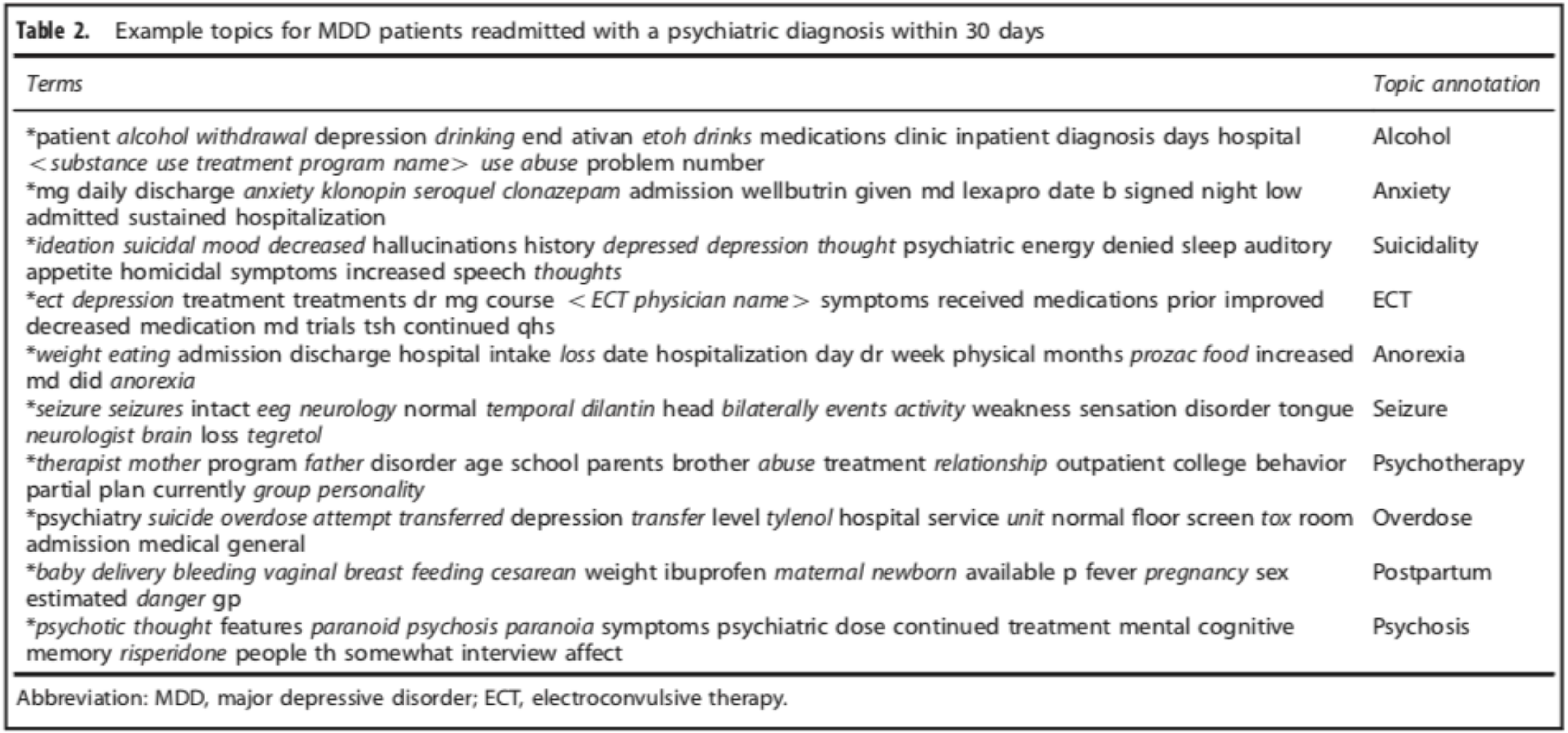


Table 2. Example topics for MDD patients readmitted with a psychiatric diagnosis within 30 days

Terms	*Topic annotation*
*patient *alcohol withdrawal* depression *drinking* end ativan *etoh drinks* medications clinic inpatient diagnosis days hospital *<substance use treatment program name> use abuse* problem number	Alcohol
*mg daily discharge *anxiety klonopin seroquel clonazepam* admission wellbutrin given md lexapro date b signed night low admitted sustained hospitalization	Anxiety
**ideation suicidal mood decreased* hallucinations history *depressed depression thought* psychiatric energy denied sleep auditory appetite homicidal symptoms increased speech *thoughts*	Suicidality
**ect depression* treatment treatments dr mg course *<ECT physician name>* symptoms received medications prior improved decreased medication md trials tsh continued qhs	ECT
**weight eating* admission discharge hospital intake *loss* date hospitalization day dr week physical months *prozac food* increased md did *anorexia*	Anorexia
**seizure seizures* intact *eeg neurology* normal *temporal dilantin* head *bilaterally events activity* weakness sensation disorder tongue *neurologist brain* loss *tegretol*	Seizure
**therapist mother* program *father* disorder age school parents brother *abuse* treatment *relationship* outpatient college behavior partial plan currently *group personality*	Psychotherapy
*psychiatry *suicide overdose attempt transferred* depression *transfer* level *tylenol* hospital service *unit* normal floor screen *tox* room admission medical general	Overdose
**baby delivery bleeding vaginal breast feeding cesarean* weight ibuprofen *maternal newborn* available p fever *pregnancy* sex estimated *danger* gp	Postpartum
**psychotic thought* features *paranoid psychosis paranoia* symptoms psychiatric dose continued treatment mental cognitive memory *risperidone* people th somewhat interview affect	Psychosis

Abbreviation: MDD, major depressive disorder; ECT, electroconvulsive therapy.

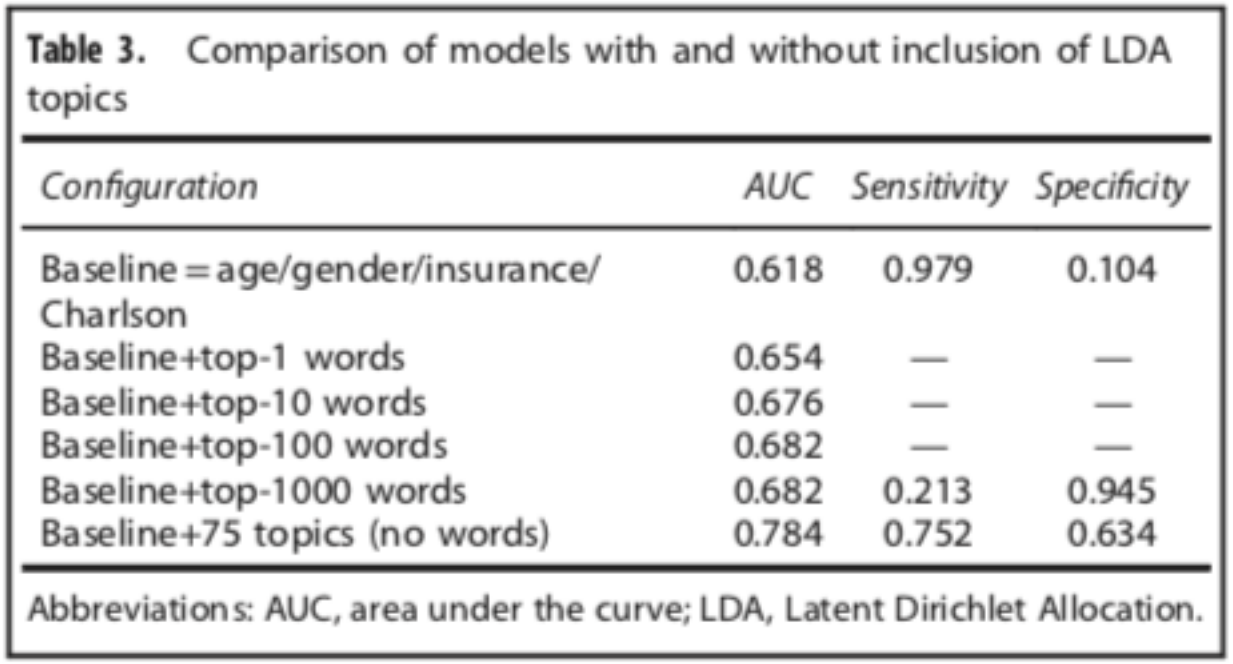


Table 3. Comparison of models with and without inclusion of LDA topics

Configuration	*AUC*	*Sensitivity*	*Specificity*
Baseline = age/gender/insurance/ Charlson	0.618	0.979	0.104
Baseline+top-1 words	0.654	—	—
Baseline+top-10 words	0.676	—	—
Baseline+top-100 words	0.682	—	—
Baseline+top-1000 words	0.682	0.213	0.945
Baseline+75 topics (no words)	0.784	0.752	0.634

Abbreviations: AUC, area under the curve; LDA, Latent Dirichlet Allocation.

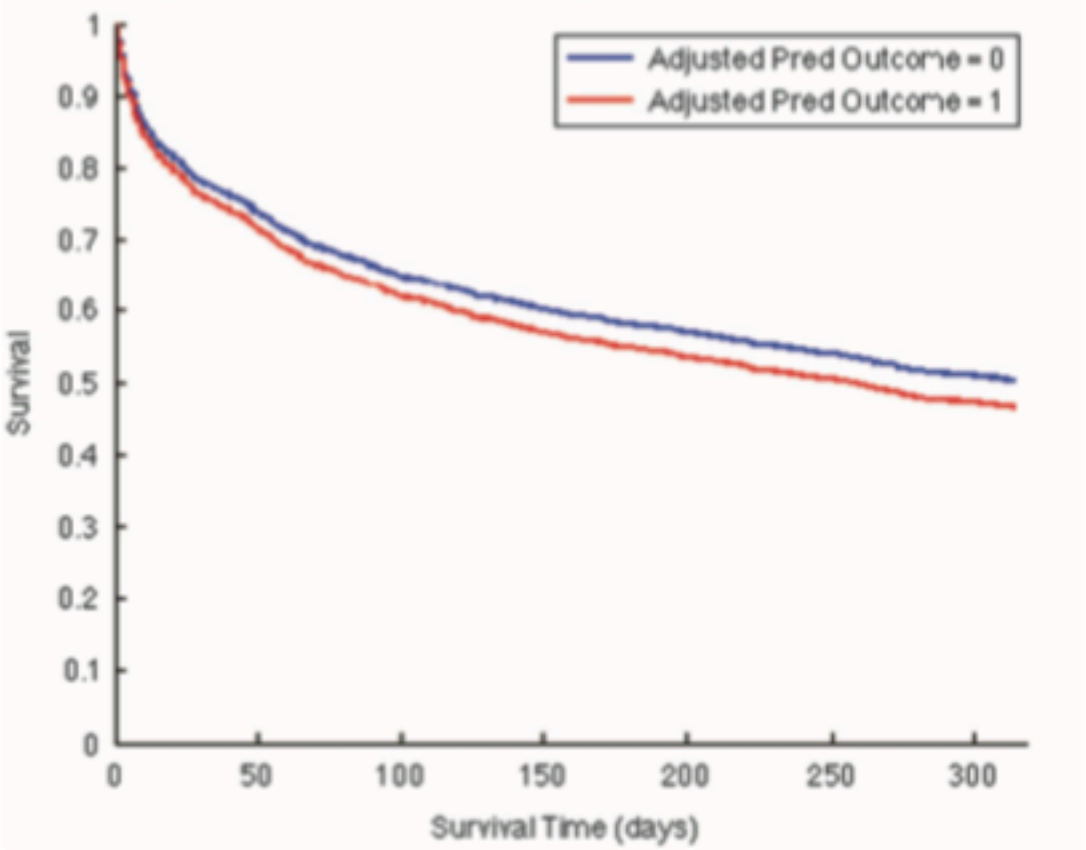


Figure 1. Kaplan–Meier survival curve for time to psychiatric hospital readmission, for a model built using baseline sociodemographic and clinical variables only. Patients are plotted separately for two groups identified by the support vector machine model as: (1) likely psychiatric readmissions in red; and (2) unlikely psychiatric readmissions in blue.

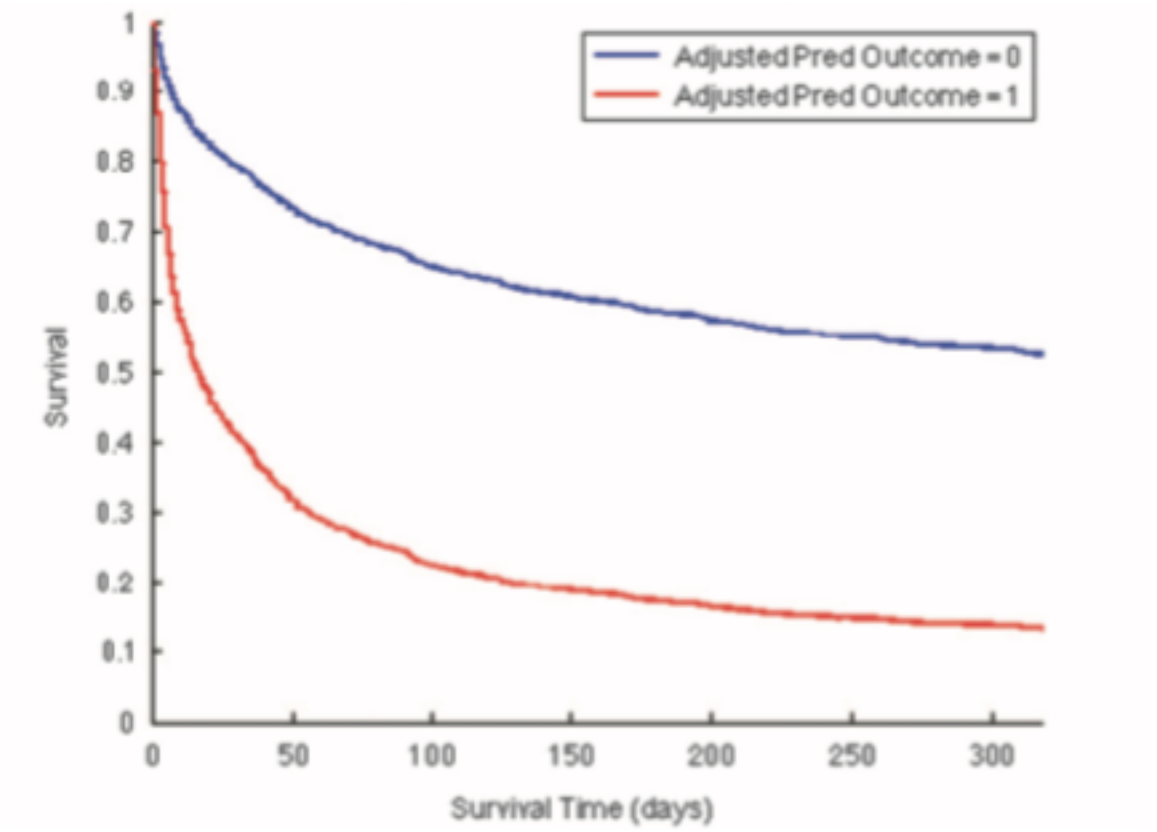


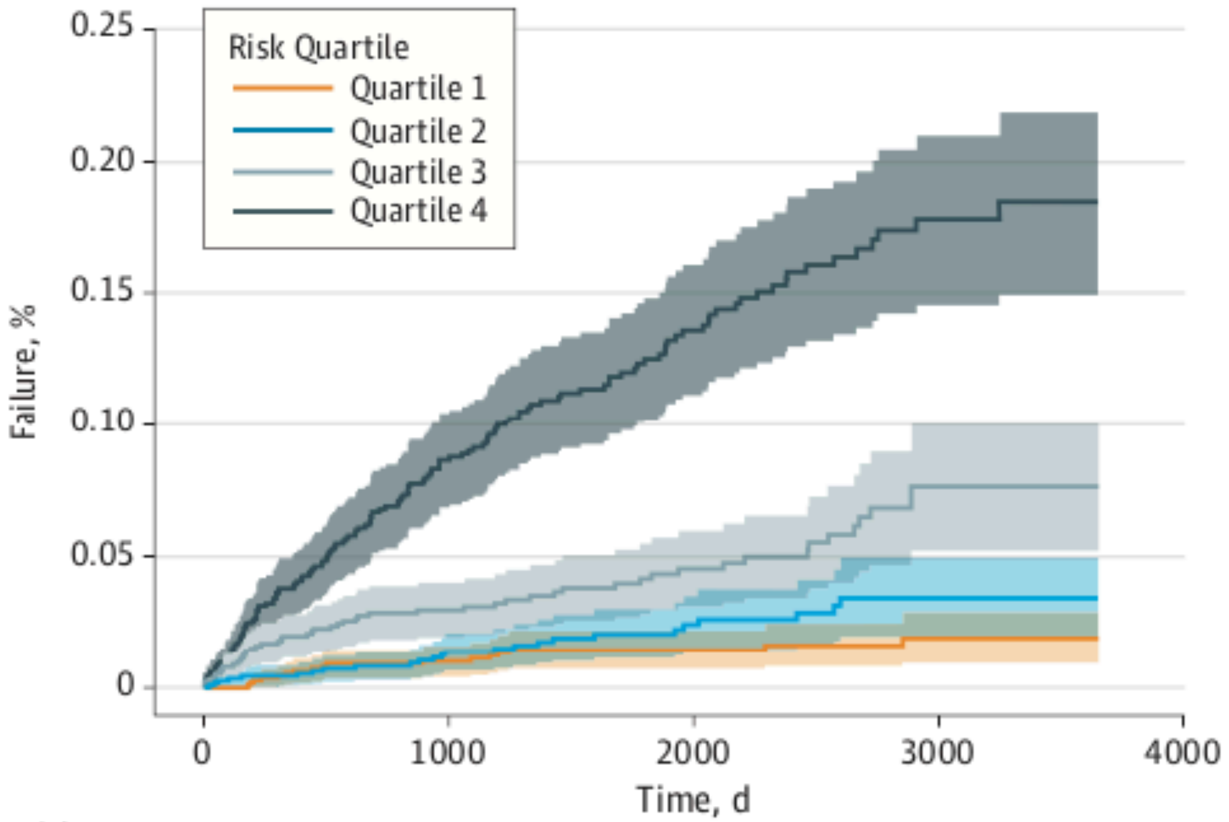
Figure 2. Kaplan–Meier survival curve for time to psychiatric hospital readmission, for a model built using the baseline variables and 75 topics. Patients are plotted separately for two groups identified by the support vector machine model as: (1) likely psychiatric readmissions in red; and (2) unlikely psychiatric readmissions in blue.

Prediction of Suicide and Accidental Death After Discharge

- Very large cohort: 845 417 discharges from two medical centers, 2005–2013
  - 458 053 unique individuals
- Imbalanced: 235 suicides, but all-cause mortality was 18% during 9 years
- Censoring: median follow-up was 5.2 years
- “Positive Valence” assessed using *curated list of 3000 terms* found in discharge summaries
  - “Valence, as used in psychology, especially in discussing emotions, means the intrinsic attractiveness/"good"-ness (positive valence) or averseness/"bad"-ness (negative valence) of an event, object, or situation.[1] The term also characterizes and categorizes specific emotions. For example, emotions popularly referred to as "negative", such as anger and fear, have negative valence. Joy has positive valence.” —Wikipedia

McCoy, T. H., Jr, Castro, V. M., Roberson, A. M., Snapper, L. A., & Perlis, R. H. (2016). Improving Prediction of Suicide and Accidental Death After Discharge From General Hospitals With Natural Language Processing. JAMA Psychiatry, 73(10), 1064–8. http://doi.org/10.1001/jamapsychiatry.2016.2172

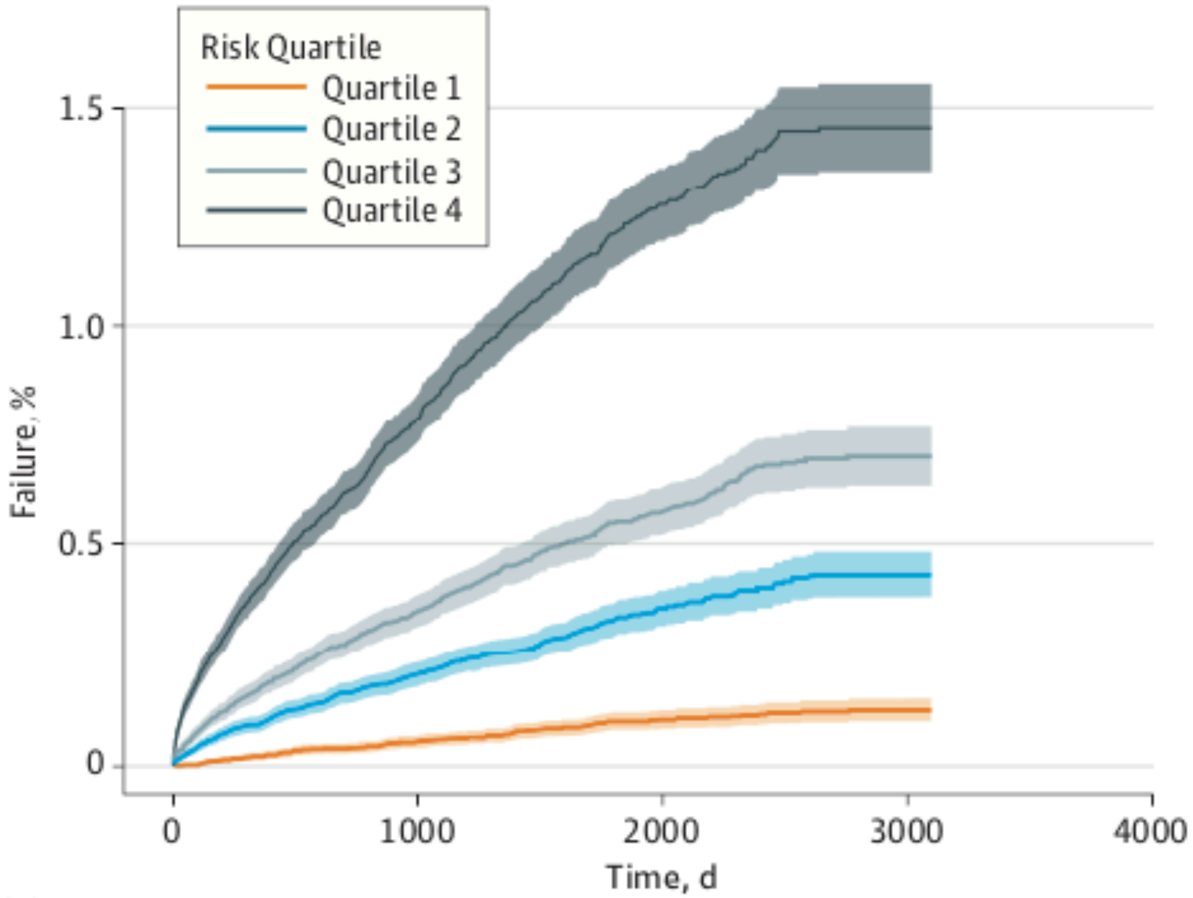
Figure 1. Kaplan-Meier Curves for Time to Death by Suicide Among 458 053 Individuals With at Least 1 Hospital Discharge by Predicted Risk Quartile



No. at risk				
Quartile 1	114514	93698	63289	31025
Quartile 2	114513	85693	52292	22697
Quartile 3	114513	82810	49258	21580
Quartile 4	114513	85746	51707	21541

The axes are rescaled inside the figure to improve interpretability.

Figure 2. Kaplan-Meier Curves for Time to Death by Suicide or Accidental Death Among 458 053 Individuals With at Least 1 Hospital Discharge by Predicted Risk Quartile



No. at risk				
Quartile 1	114514	99448	68260	33935
Quartile 2	114513	89270	55871	25263
Quartile 3	114513	84465	50944	21768
Quartile 4	114513	74764	41471	15877

The axes are rescaled inside the figure to improve interpretability.

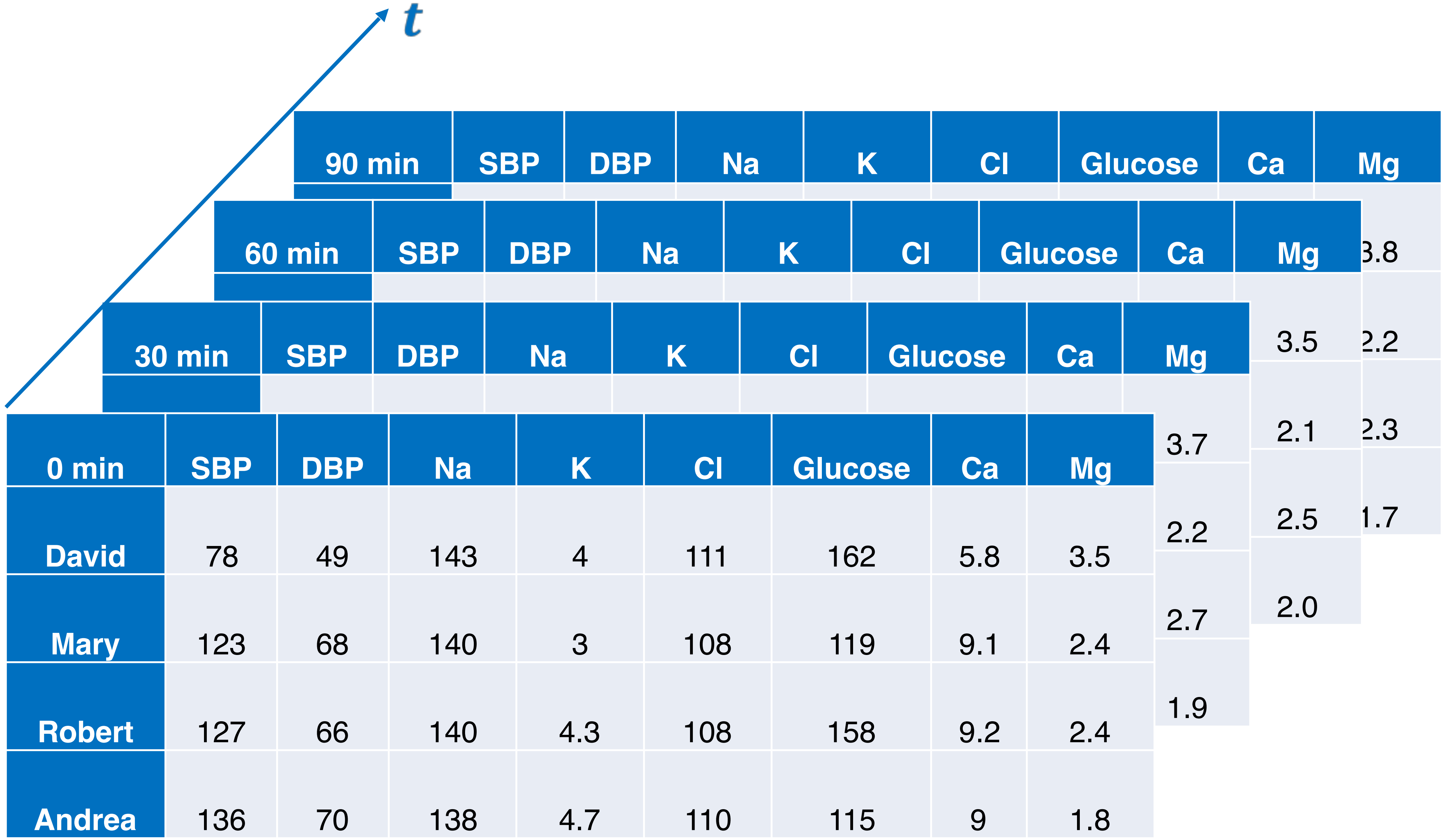
Tensor Factorization for Unsupervised Exploitation of Text

- Goals:
  - Identify patients with subtypes of lymphoma by analysis of their pathology notes
- Unsupervised approach
  - Do the core “clusters” of patient descriptions correspond to known lymphoma types?
  - Can we use these to help refine out understanding of the types?

Luo, Y., Sohani, A. R., Hochberg, E. P., & Szolovits, P. (2014). Automatic lymphoma classification with sentence subgraph mining from pathology reports. Journal of the American Medical Informatics Association, 21(5), amiajnl–2013–002443–832. http://doi.org/10.1136/amiajnl-2013-002443

Generalizing Matrix to Tensor

- *N*-dimensional data structure (N ≥ 3)
- Example: patient and timed physiological measurements



90 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
								3.8
								2.2
								2.3
								1.7

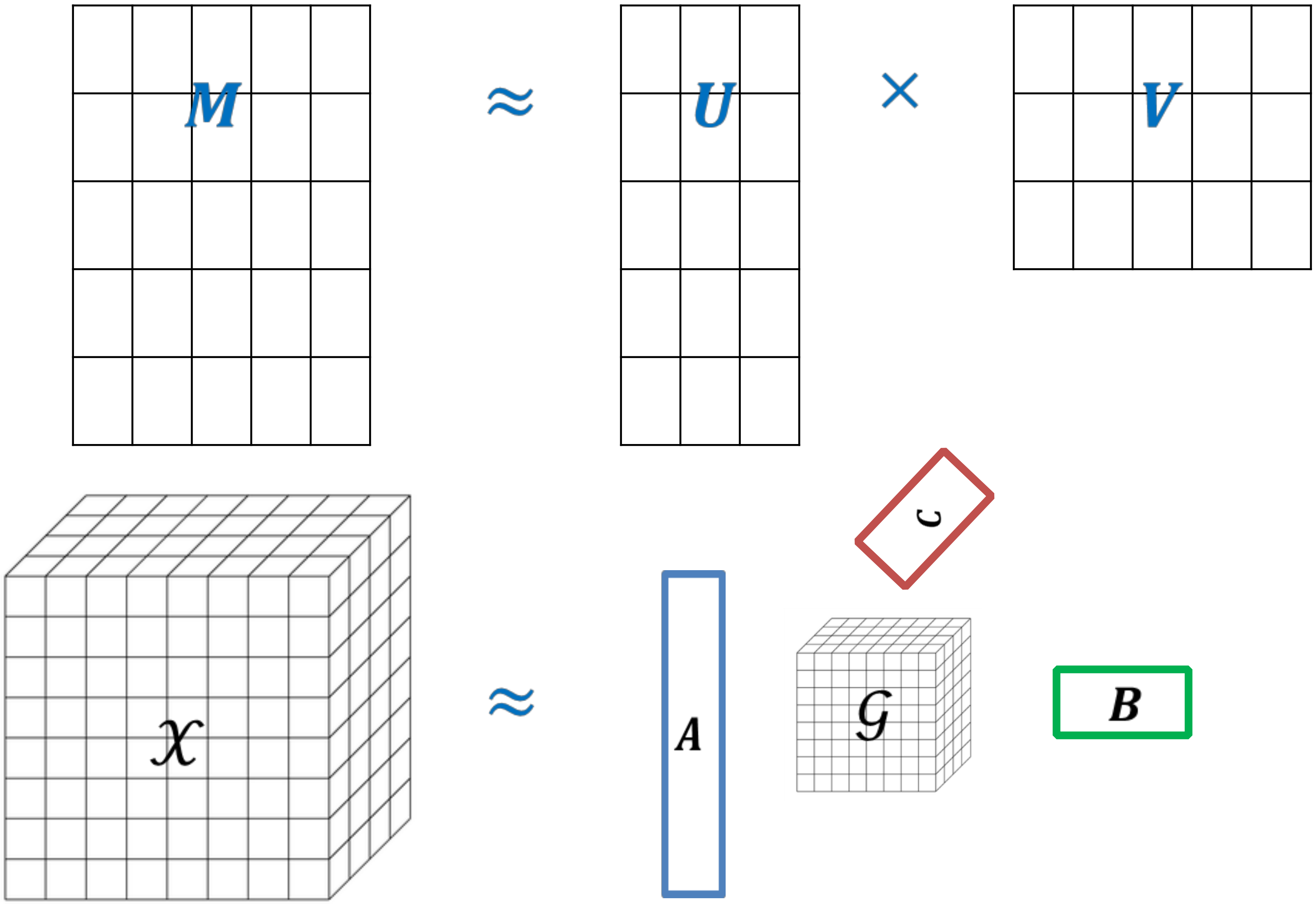
60 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
								3.5
								2.1
								2.5
								2.0

30 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
								3.7
								2.2
								2.7
								1.9

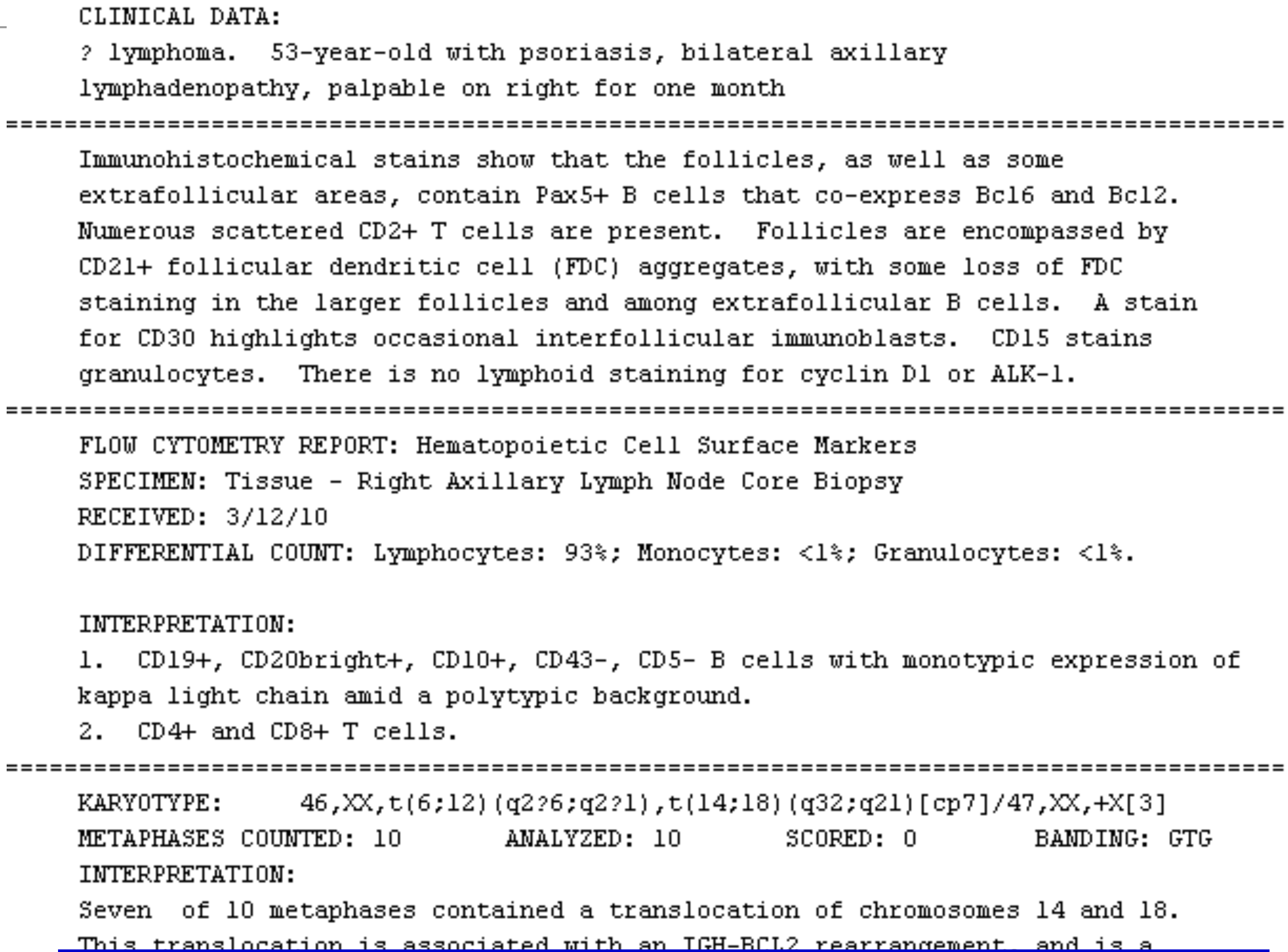
0 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	78	49	143	4	111	162	5.8	3.5
Mary	123	68	140	3	108	119	9.1	2.4
Robert	127	66	140	4.3	108	158	9.2	2.4
Andrea	136	70	138	4.7	110	115	9	1.8

Non-Negative Tensor Factorization

- NMF extension to tensors of arbitrary order
- Tucker model, a generalized form of spectral modeling



Representation of Narrative Sentences



```
     CLINICAL DATA:
     ? lymphoma.  53-year-old with psoriasis, bilateral axillary
     lymphadenopathy, palpable on right for one month
=======================================================================================
     Immunohistochemical stains show that the follicles, as well as some
     extrafollicular areas, contain Pax5+ B cells that co-express Bcl6 and Bcl2.
     Numerous scattered CD2+ T cells are present.  Follicles are encompassed by
     CD21+ follicular dendritic cell (FDC) aggregates, with some loss of FDC
     staining in the larger follicles and among extrafollicular B cells.  A stain
     for CD30 highlights occasional interfollicular immunoblasts.  CD15 stains
     granulocytes.  There is no lymphoid staining for cyclin D1 or ALK-1.
=======================================================================================
     FLOW CYTOMETRY REPORT: Hematopoietic Cell Surface Markers
     SPECIMEN: Tissue - Right Axillary Lymph Node Core Biopsy
     RECEIVED: 3/12/10
     DIFFERENTIAL COUNT: Lymphocytes: 93%; Monocytes: <1%; Granulocytes: <1%.

     INTERPRETATION:
     1.  CD19+, CD20bright+, CD10+, CD43-, CD5- B cells with monotypic expression of
     kappa light chain amid a polytypic background.
     2.  CD4+ and CD8+ T cells.
=======================================================================================
     KARYOTYPE:     46,XX,t(6;12)(q2?6;q2?1),t(14;18)(q32;q21)[cp7]/47,XX,+X[3]
     METAPHASES COUNTED: 10       ANALYZED: 10       SCORED: 0       BANDING: GTG
     INTERPRETATION:
     Seven  of 10 metaphases contained a translocation of chromosomes 14 and 18.
     This translocation is associated with an IGH-BCL2 rearrangement, and is a
```

Feature representation is the key to both interpretability and generalizability

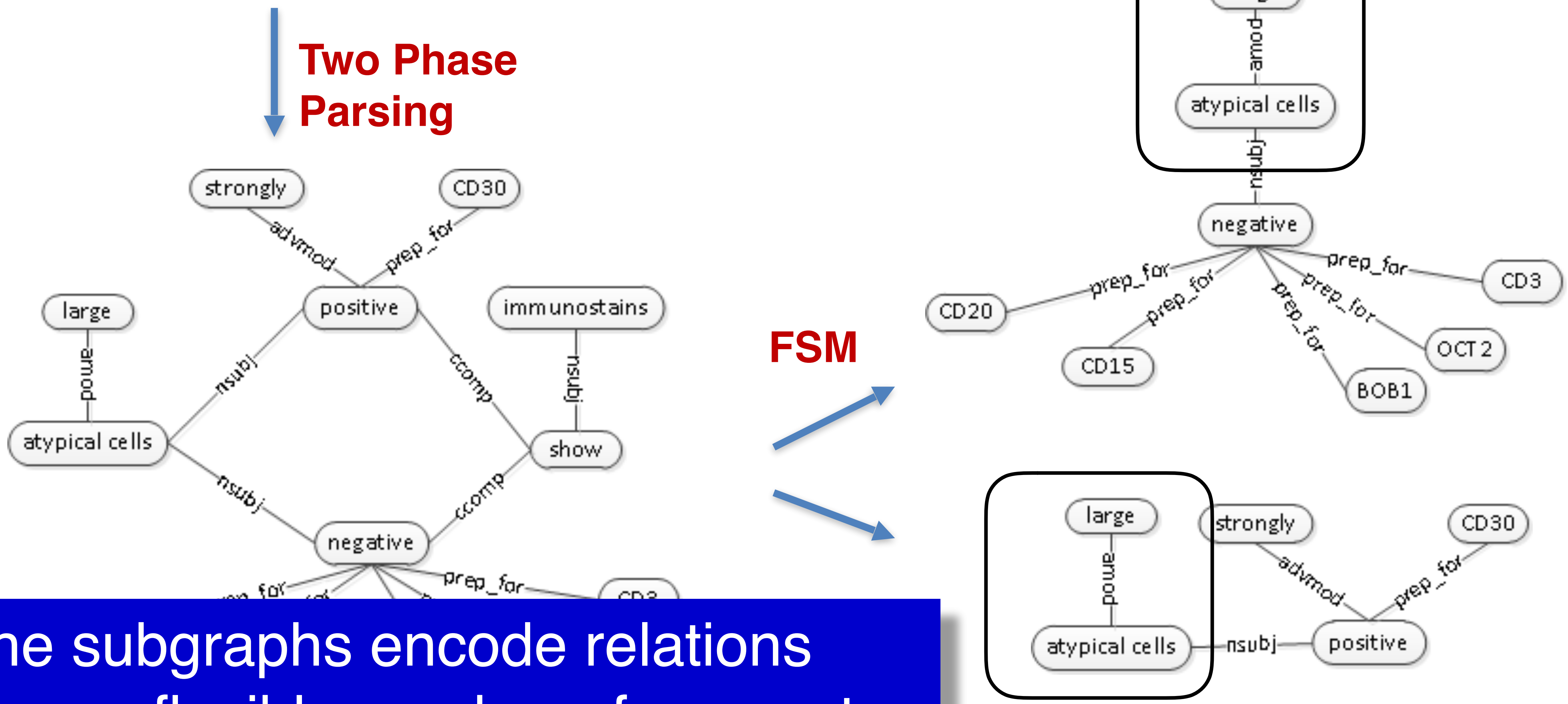
Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."
- The sentence tells relationships among procedures, cells, and immunologic factors
- Feature choices
  - Words
  - UMLS (Unified Medical Language System) concepts, e.g. LCA and CD45
- Can we do better? Relations?

Graph representation is the universal language for modeling relationships among flexible number of concepts

Representation of Narrative Sentences

- “Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3.”



Two Phase Parsing

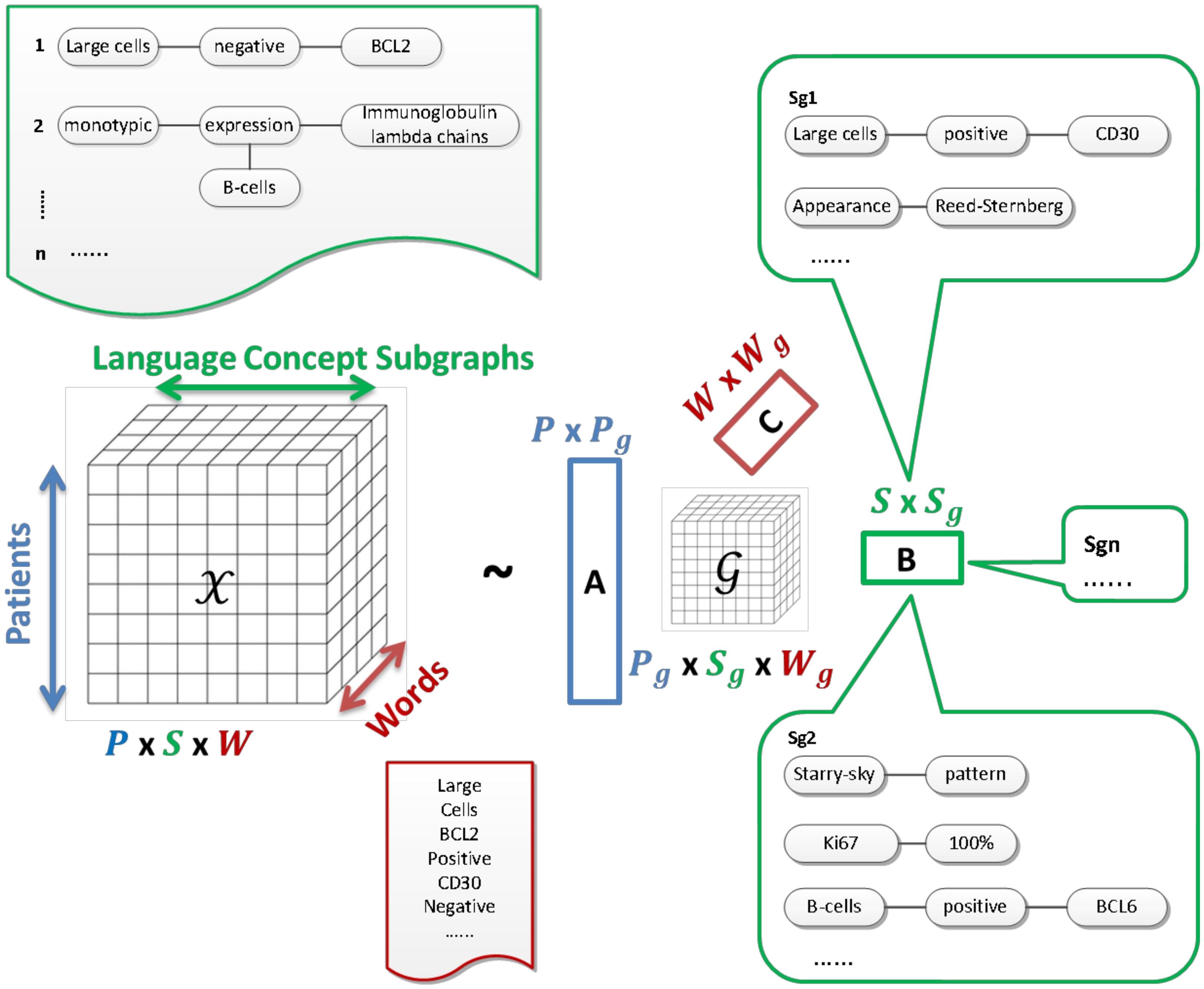
FSM

The subgraphs encode relations among flexible number of concepts

(Luo et al. 2013a)

Multi-Mode Learning

SANTF schematic view



Unsupervised Learning – Clustering Results

- Non-negative matrix factorization as baseline
  - Traditional two-dimensional view
  - Three matrix formulation baselines
    - Patient by word
    - Patient by subgraph
    - Patient by subgraph and word
- SANTF as target (Luo et al. 2014b)
  - Patient by subgraph by word

Clinical Narrative Text			
Lymphoma	All	Train	Test
DLBCL	589	305	284
Follicular	184	101	83
Hodgkin	124	65	59

Metrics Methods	Macro Average			Micro Average		
	Precision	Recall	F-measure	Precision	Recall	F-measure
(1) NMF pt × wd	0.492	0.495	0.428	0.626	0.626	0.626
(2) NMF pt × sg	0.621	0.765	0.601	0.605	0.605	0.605
(3) NMF pt × [sg wd]	0.637	0.787	0.615	0.614	0.614	0.614
(4) SANTF pt × sg × wd	**0.720**^{1,2,3}	**0.849**^{1,2,3}	**0.743**^{1,2,3}	**0.751**^{1,2,3}	**0.751**^{1,2,3}	**0.751**^{1,2,3}

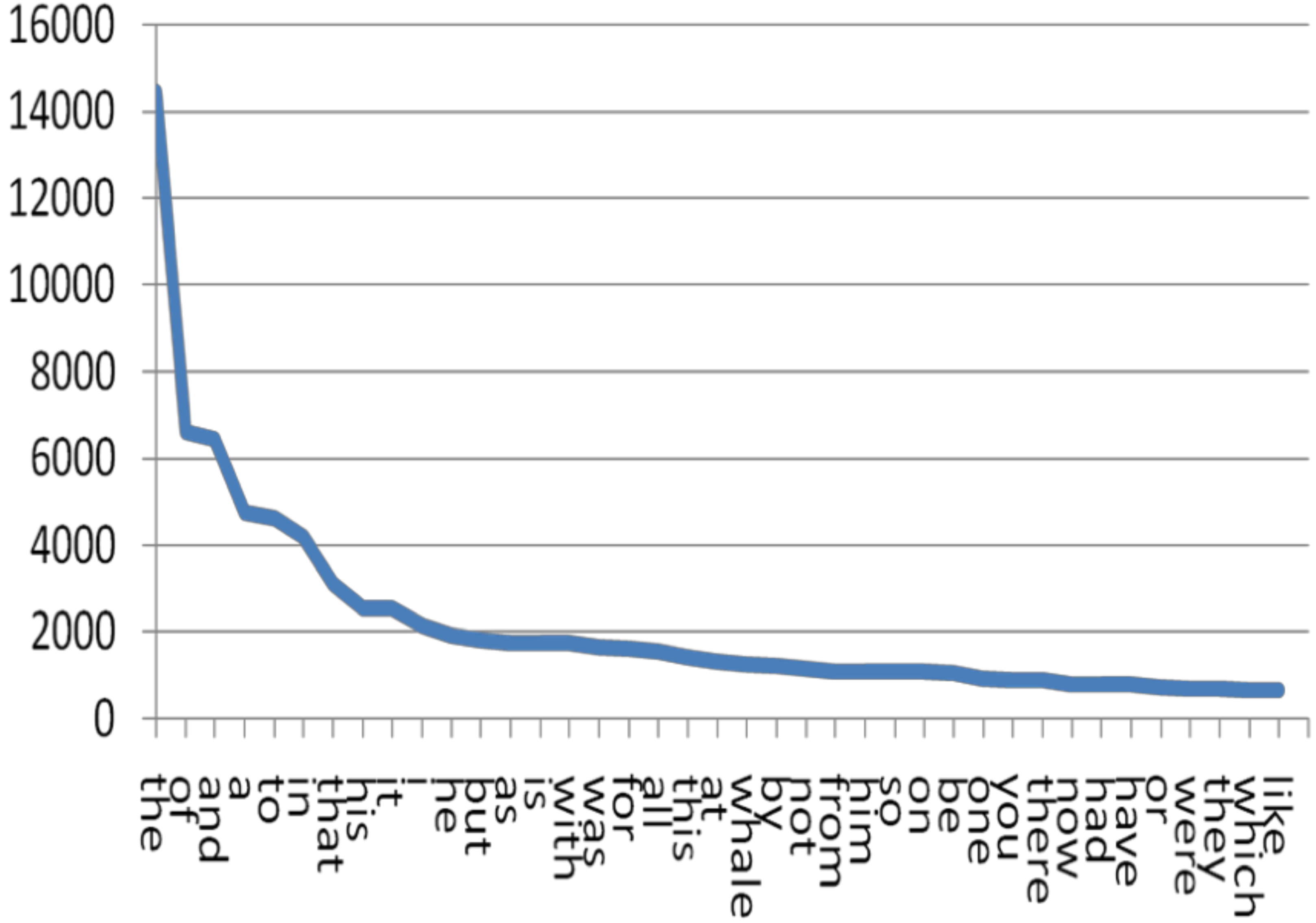
Language Modeling

- Predict the next token given the ones before it
  - In unigram model, P(token) is just estimated from frequency in corpus
- Markov assumption simplifies model so
  - P(token | stuff before) = P(token | previous token) [bigram model]
  - $P(t_k$ | stuff before) = P(tk | t_{k-1}, …, t_{k-n}) [n-gram models]

- Perplexity is an aggregate measure of the complexity of a corpus
  - $2^{H(p)}$ where H(p) is the entropy of the probability distribution
  - intuitively, the number of likely ways to continue a text
    - a perplexity of *k* means that you are as surprised on average as you would have been if you had to guess between *k* equiprobable choices at each step
  - For example, we compared perplexity of dictated doctors' notes (8.8) vs. that of doctor-patient conversations (73.1)
    - What does that tell you about the difficulty of accurately transcribing speech for these applications?

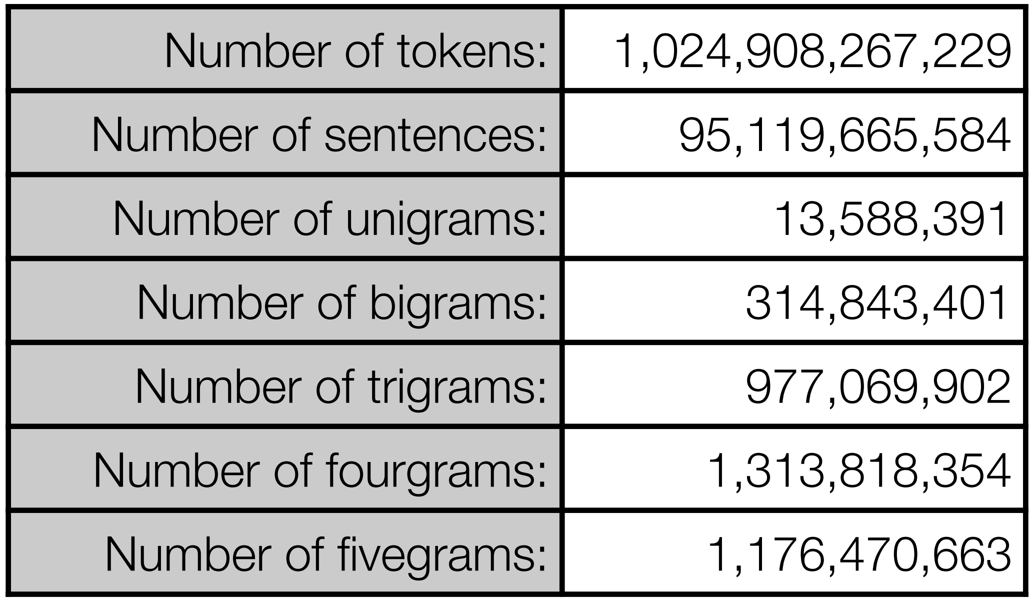
Statistical Models of Language
Zipf's law

- There are very few very frequent words
- Most words have very low frequencies
- The frequency of a word is inversely proportional to its rank
- In the Brown corpus, the 10 top-ranked words make up 23% of total corpus size (Baroni, 2007)
-



N-gram models

- Shakespeare as a Corpus
  - N=884,647 tokens, V=29,066
  - Shakespeare produced 300,000 bigram types out of V2= 844 million possible bigrams...
    - So, 99.96% of the possible bigrams were never seen
- Google released corpus of 1,024,980,267,229 (i.e., ~1T) words in 2006
  - 13.6M unique words occurring at least 200 times
  - 1.2B five-word sequences that occur at least 40 times



Number of tokens:	1,024,908,267,229
Number of sentences:	95,119,665,584
Number of unigrams:	13,588,391
Number of bigrams:	314,843,401
Number of trigrams:	977,069,902
Number of fourgrams:	1,313,818,354
Number of fivegrams:	1,176,470,663

https://ai.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html

Example Google 3-grams

ceramics	collectables	collectibles	55
ceramics	collectables	fine	130
ceramics	collected	by	52
ceramics	collectible	pottery	50
ceramics	collectibles	cooking	45
ceramics	collection	,	144
ceramics	collection	.	247
ceramics	collection	</S>	120
ceramics	collection	and	43
ceramics	collection	at	52
ceramics	collection	is	68
ceramics	collection	of	76
ceramics	collection	|	59
ceramics	collections	,	66
ceramics	collections	.	60
ceramics	combined	with	46
ceramics	come	from	69
ceramics	comes	from	660
ceramics	community	,	109
ceramics	community	.	210
ceramics	community	for	61
ceramics	companies	.	53
ceramics	companies	cpnsultants	173

Example Google 4-grams

serve	as	the	incoming	92
serve	as	the	incubator	99
serve	as	the	independent	79
serve	as	the	index	223
serve	as	the	indication	72
serve	as	the	indicator	120
serve	as	the	indicators	45
serve	as	the	indispensable	111
serve	as	the	indispensible	40
serve	as	the	individual	234
serve	as	the	industrial	52
serve	as	the	industry	607
serve	as	the	info	42
serve	as	the	informal	102
serve	as	the	information	838
serve	as	the	informational	41
serve	as	the	infrastructure	500
serve	as	the	initial	5331
serve	as	the	initiating	125
serve	as	the	initiation	63
serve	as	the	initiator	81
serve	as	the	injector	56
serve	as	the	inlet	41

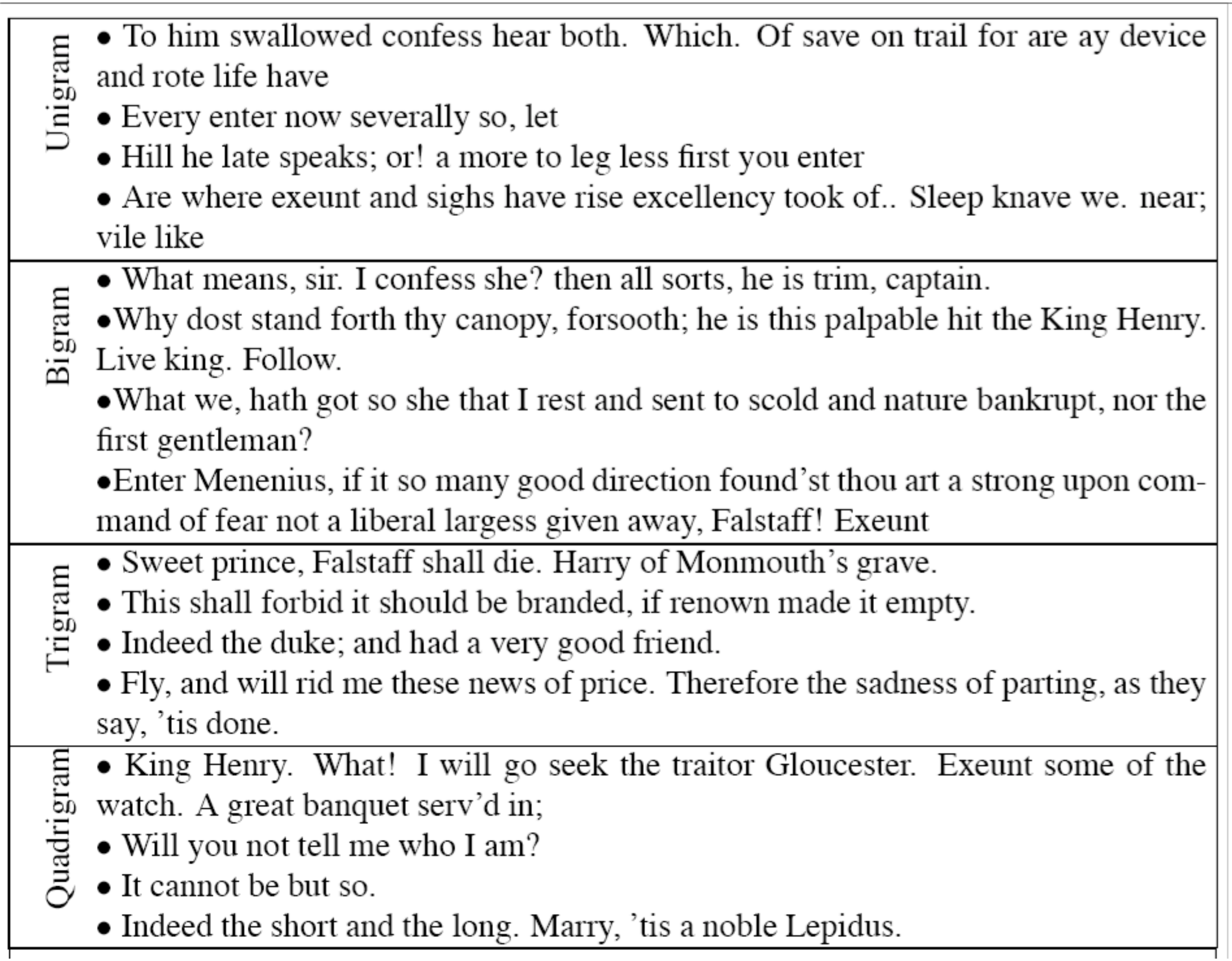
Generating Sequences

- This model can be turned around to generate random sentences that are like the sentences from which the model was derived.
- Generally attributed to Claude Shannon.
    - Sample a random bigram (<s>, w) according to its probability
    - Now sample a random bigram (w, x) according to its probability
    - Where the prefix w matches the suffix of the first.
    - And so on until we randomly choose a (y, </s>)
- Then string the words together

```
<s> I
    I want
      want to
           to get
              get Chinese
                  Chinese food
                          food </s>
```

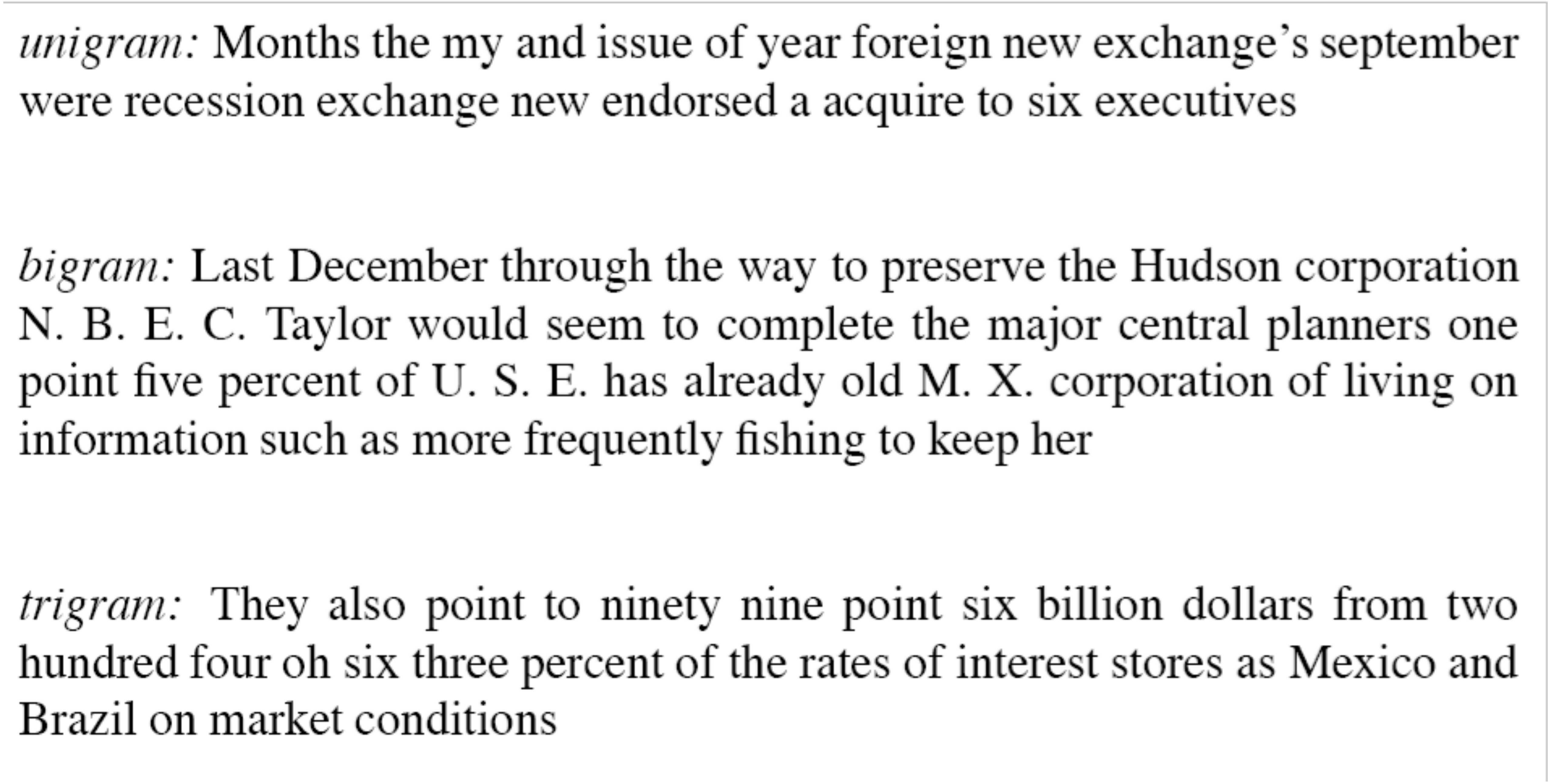
Slide adapted from Anna Rumshisky

Generating Shakespeare



Unigram	• To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have • Every enter now severally so, let • Hill he late speaks; or! a more to leg less first you enter • Are where exeunt and sighs have rise excellency took of.. Sleep knave we. near; vile like
Bigram	• What means, sir. I confess she? then all sorts, he is trim, captain. •Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow. •What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman? •Enter Menenius, if it so many good direction found'st thou art a strong upon command of fear not a liberal largess given away, Falstaff! Exeunt
Trigram	• Sweet prince, Falstaff shall die. Harry of Monmouth's grave. • This shall forbid it should be branded, if renown made it empty. • Indeed the duke; and had a very good friend. • Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.
Quadrigram	• King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in; • Will you not tell me who I am? • It cannot be but so. • Indeed the short and the long. Marry, 'tis a noble Lepidus.

Generating the *Wall Street Journal*



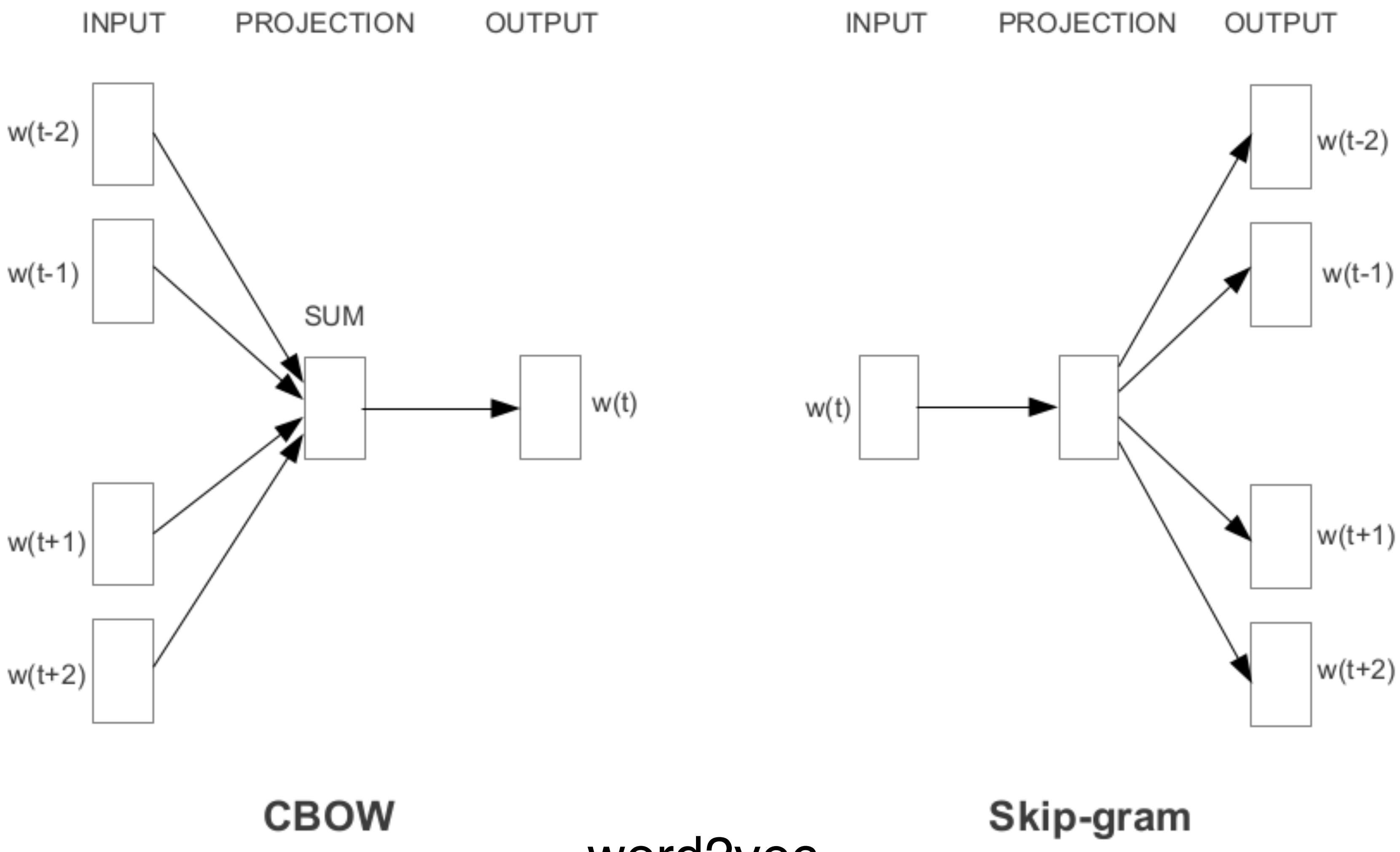
unigram: Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives

bigram: Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

trigram: They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

Distributional Semantics

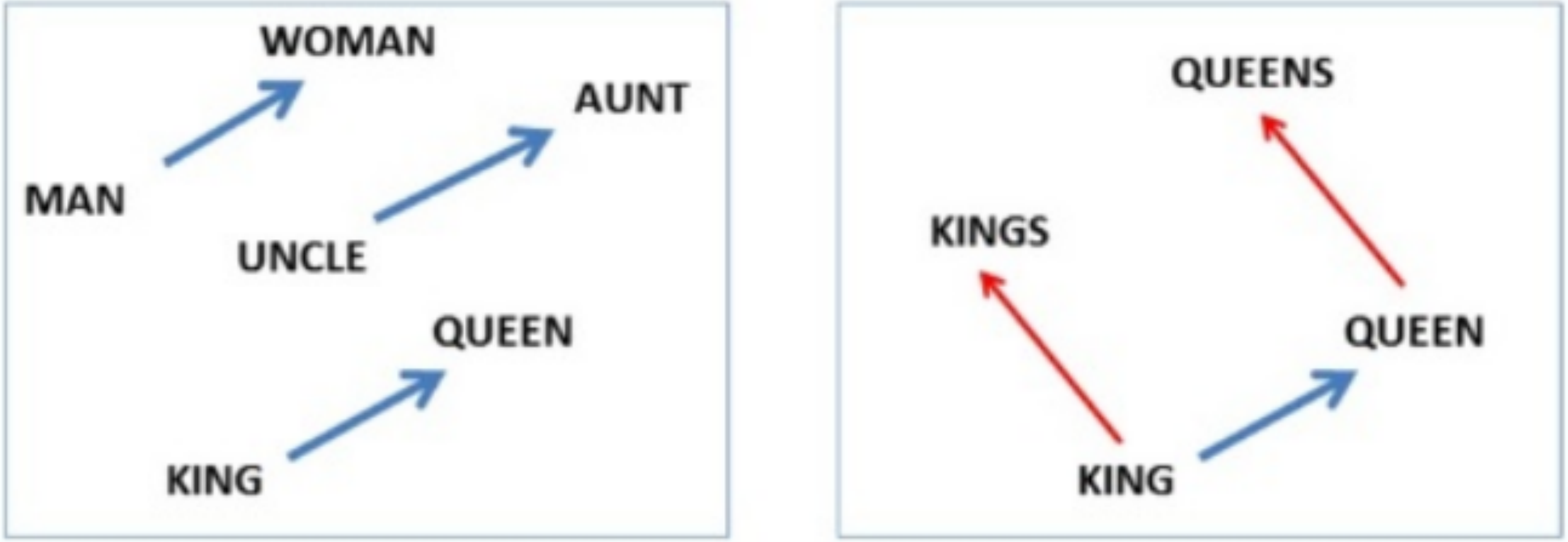
- Terms that appear in the same context of other words are (probably) semantically related
- Every term is mapped to a high-dimensional vector (the embedding space)
- Ever more sophisticated versions of embeddings, equivalent to matrix factorization
    - Word2Vec
    - GloVe
    - Elmo
    - Bert
    - GPT



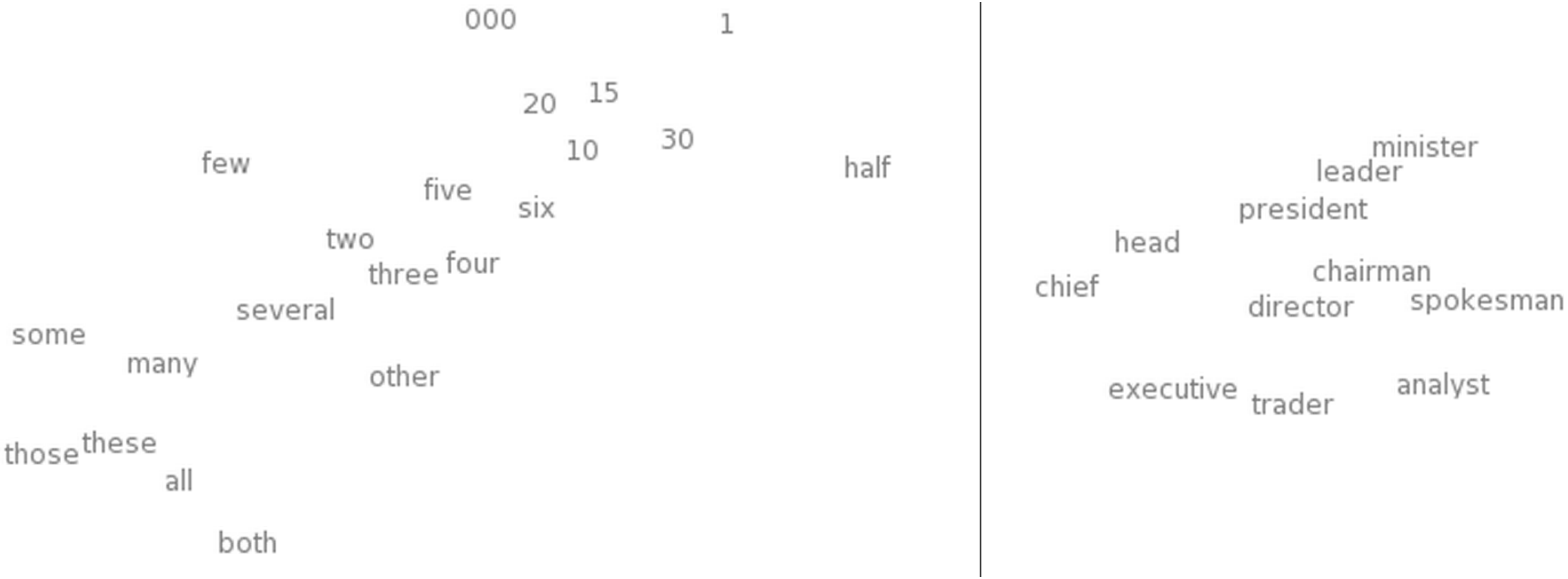
word2vec

Plausibility of semantic claims

vec("man") – vec("king") + vec("woman") = vec("queen")



t-Distributed Stochastic Neighbor Embedding



van der Maaten, L., Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2579–2605.

Feature extraction for phenotyping from semantic and knowledge resources (SEDFE)

- Goal: “fully automated and robust unsupervised feature selection method that leverages only publicly available medical knowledge sources, instead of EHR data”
  - Surrogate features derived from knowledge sources
- Method:
  - Build a word2vec skipgram model from . 5M Springer articles (2006-08) to yield 500-D vectors for each word
  - Sum vectors for each word in the defining strings for UMLS Concepts, weighted by IDF
  - For each disease in Wikipedia, Medscape eMedicine, Merck Manuals Professional Edition, Mayo Clinic Diseases and Conditions, and MedlinePlus Medical Encyclopedia use NER to find all concepts related to the phenotype

- Retain only concepts that occur in at least 3 of 5 knowledge sources
- Choose top *k* concepts whose embedding vectors are closest (by cos distance) to the embedding of the phenotype
- Define the phenotype as a linear combination of its related concepts, learn weights by least squares, and choose *k* to minimize BIC

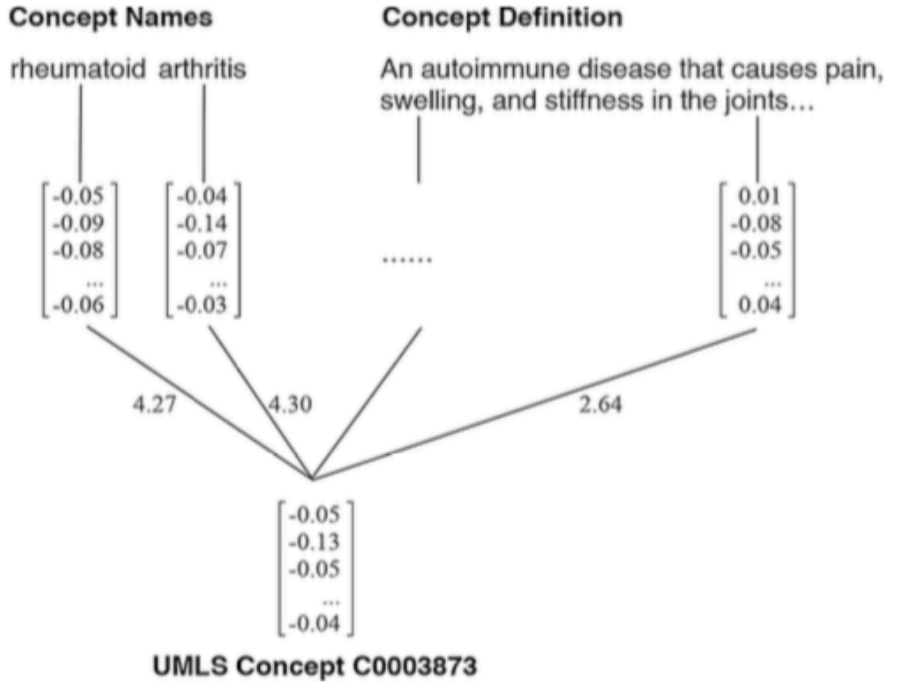
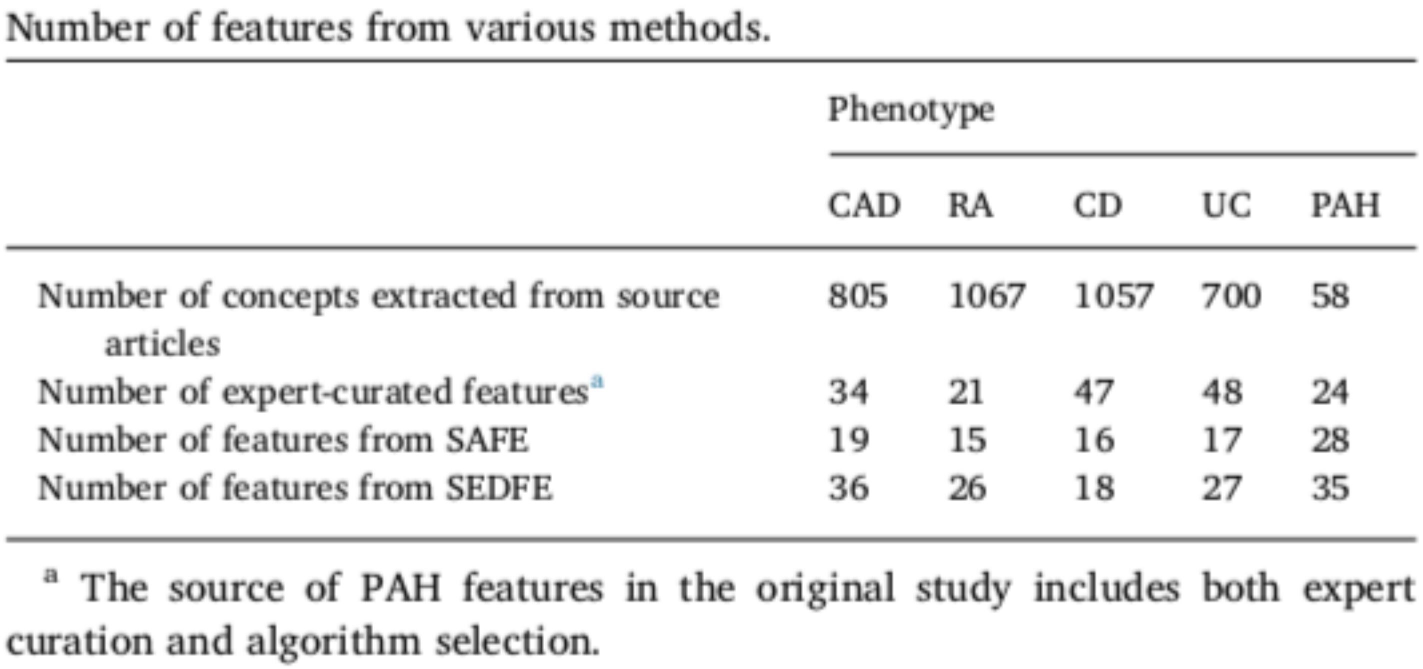


Fig. 1. Generating concept vector representations from word vectors in the paraphrase.

Ning, W., Chan, S., Beam, A., Yu, M., Geva, A., Liao, K., et al. (2019). Feature extraction for phenotyping from semantic and knowledge resources. *Journal of Biomedical Informatics*, *91*, 103122. http://doi.org/10.1016/j.jbi.2019.103122

Evaluating SEDFE

- Used to create phenotypes for coronary artery disease (CAD), rheumatoid arthritis (RA), Crohn's disease (CD), ulcerative colitis (UC), and pediatric pulmonary arterial hypertension (PAH)

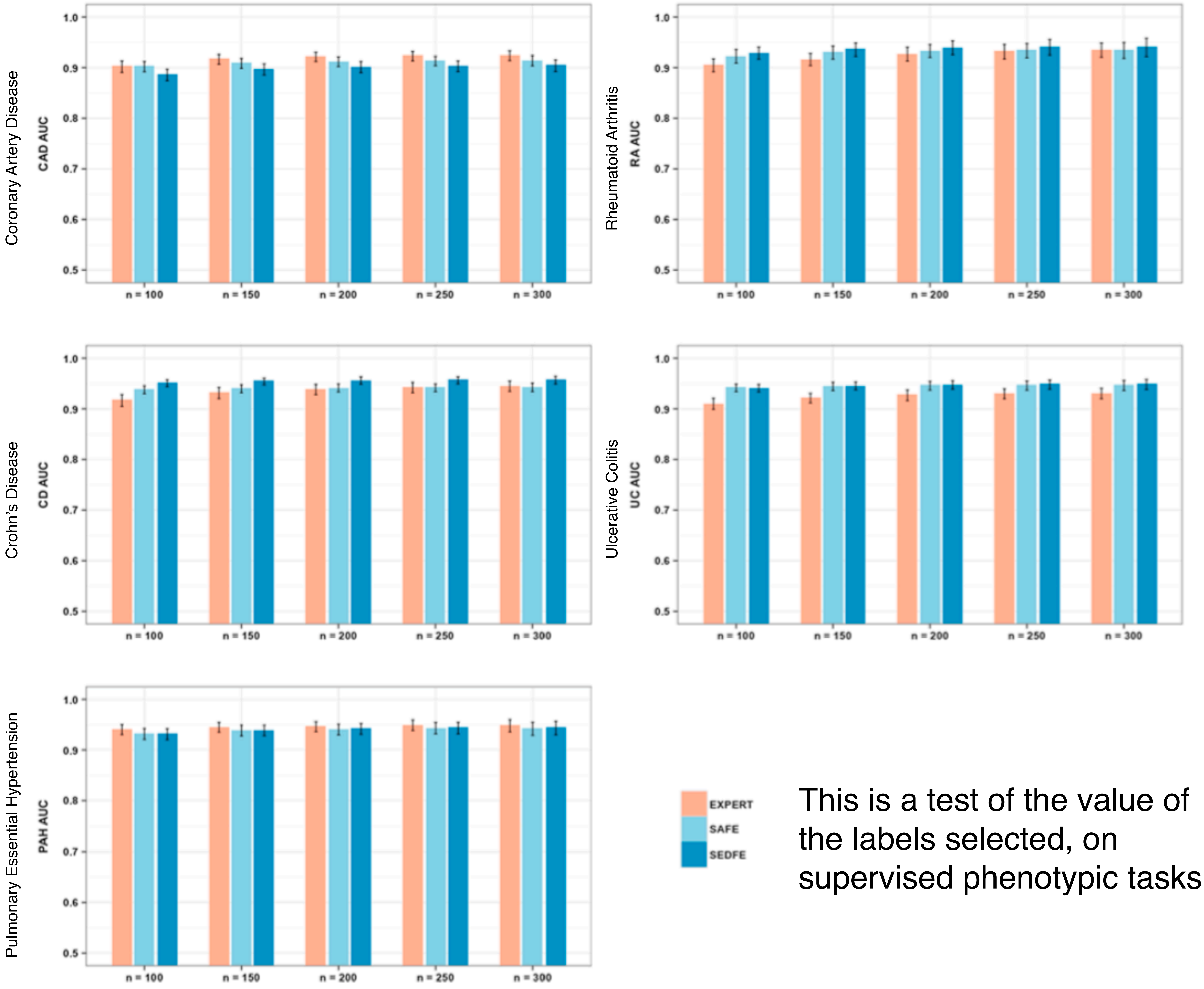


Number of features from various methods.

	Phenotype				
	CAD	RA	CD	UC	PAH
Number of concepts extracted from source articles	805	1067	1057	700	58
Number of expert-curated features[a]	34	21	47	48	24
Number of features from SAFE	19	15	16	17	28
Number of features from SEDFE	36	26	18	27	35

[a] The source of PAH features in the original study includes both expert curation and algorithm selection.

	AFEP	SAFE	SEDFE
Commonality	Applies NER to online articles about the target phenotype to find an initial list of clinical concepts as candidate features		
Feature selection method	Frequency control, then threshold by rank correlation with the NLP feature representing the target phenotype	Frequency control, majority voting, then use sparse regression to predict the silver-standard labels derived from surrogate features	Majority voting; Use concept embedding to determine feature relatedness; Use semantic combination and the BIC to determine the number of needed features
Data requirement	EHR data (hospital dependent and not sharable)	EHR data (hospital dependent and not sharable)	A biomedical corpus for training word embedding (usually sharable)
Tuning parameters	Threshold for the rank correlation	(1) Upper and lower thresholds of the surrogate features for creating the silver standard labels, which are affected by the distribution of the features, and therefore phenotype dependent; (2) The number of patients to sample, which affects the number of selected features	The word embedding parameters, which are not overly sensitive. The embedding is done only once for all phenotypes



This is a test of the value of the labels selected, on supervised phenotypic tasks.

Fig. 3. AUC of supervised algorithms trained with features selected by EXPERT, SAFE, and SEDFE.

ANN model for de-identification

- Label-sequence optimization layer

$$s(y_{1:n}) = \sum_{i=1}^{n} \mathbf{a}_i[y_i] + \sum_{i=2}^{n} T[y_{i-1}, y_i]$$

- Label prediction layer
- Character-enhanced token-embedding layer

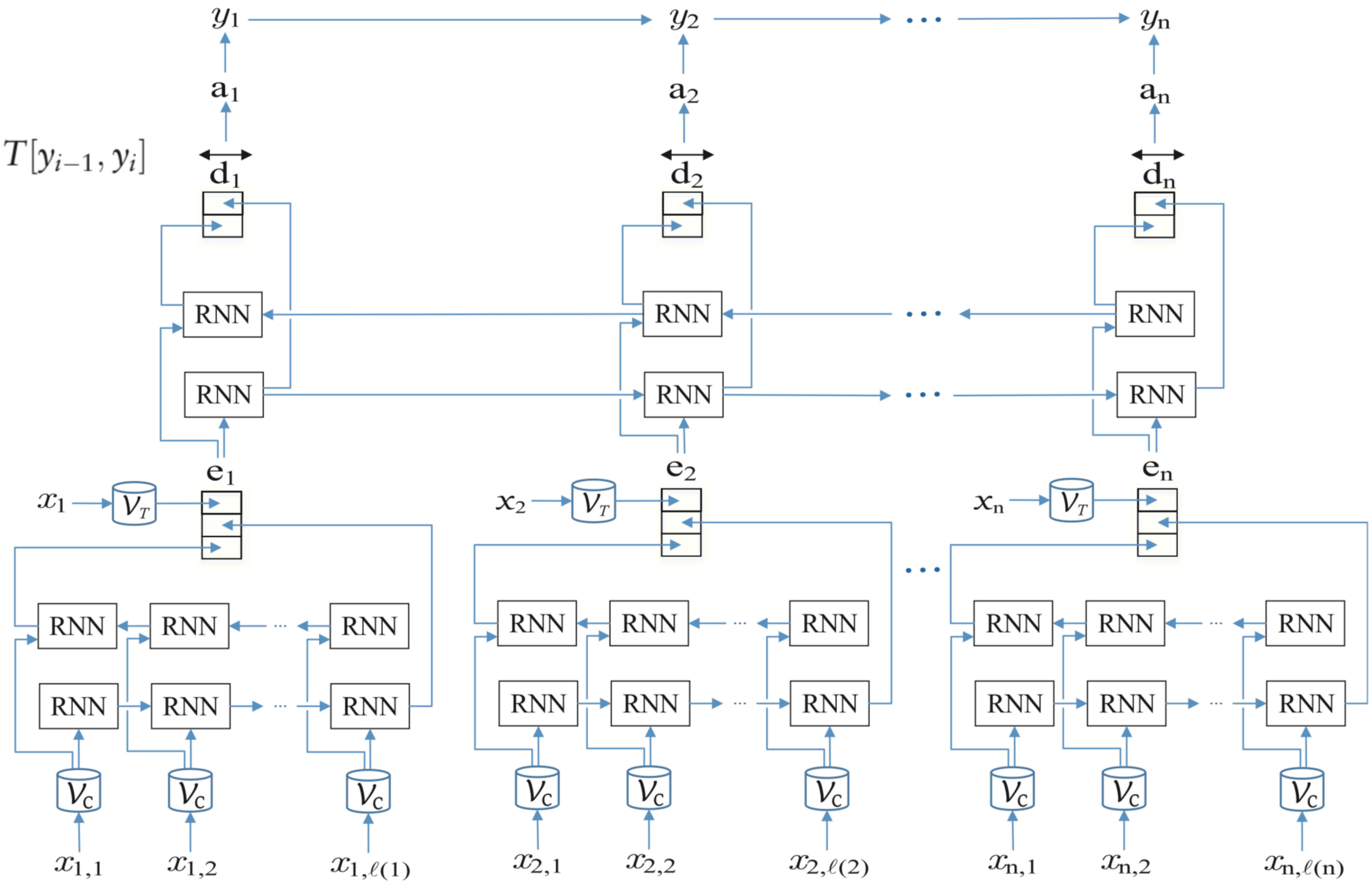
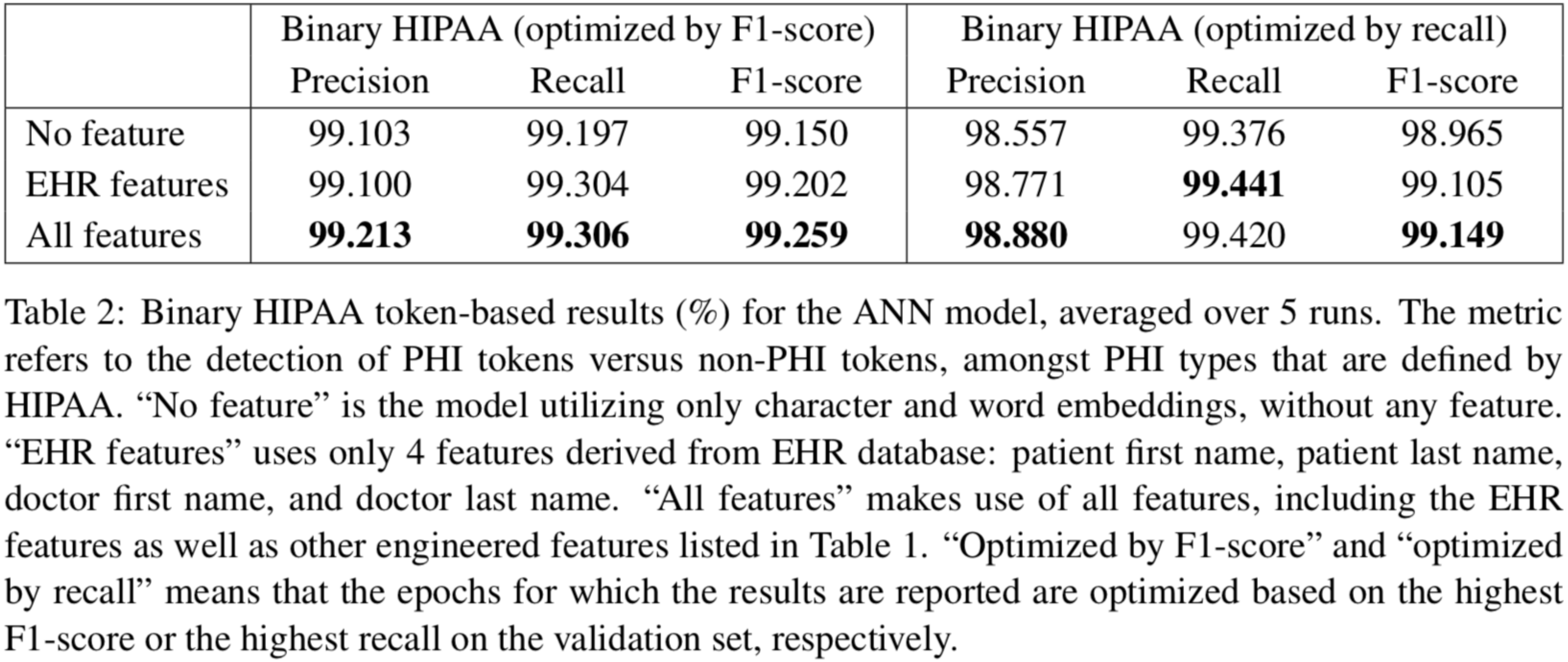


Figure 1. Architecture of the artificial neural network (ANN) model. (RNN, recurrent neural network.) The type of RNN used in this model is long short-term memory (LSTM). n is the number of tokens, and x_i is the i^{th} token. $\mathcal{V}_T$ is the mapping from tokens to token embeddings. $\ell(i)$ is the number of characters and $x_{i,j}$ is the j^{th} character in the i^{th} token. $\mathcal{V}_C$ is the mapping from characters to character embeddings. e_i is the character-enhanced token embeddings of the i^{th} token. $\overleftrightarrow{\mathrm{d}}_i$ is the output of the LSTM of the label prediction layer, a_i is the probability vector over labels, y_i is the predicted label of the i^{th} token.

Dernoncourt, F., Lee, J. Y., Uzuner, Ö., & Szolovits, P. (2016). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, ocw156. http://doi.org/10.1093/jamia/ocw156

De-Identifier performance



	Binary HIPAA (optimized by F1-score)			Binary HIPAA (optimized by recall)		
	Precision	Recall	F1-score	Precision	Recall	F1-score
No feature	99.103	99.197	99.150	98.557	99.376	98.965
EHR features	99.100	99.304	99.202	98.771	**99.441**	99.105
All features	**99.213**	**99.306**	**99.259**	**98.880**	99.420	**99.149**

Table 2: Binary HIPAA token-based results (%) for the ANN model, averaged over 5 runs. The metric refers to the detection of PHI tokens versus non-PHI tokens, amongst PHI types that are defined by HIPAA. “No feature” is the model utilizing only character and word embeddings, without any feature. “EHR features” uses only 4 features derived from EHR database: patient first name, patient last name, doctor first name, and doctor last name. “All features” makes use of all features, including the EHR features as well as other engineered features listed in Table 1. “Optimized by F1-score” and “optimized by recall” means that the epochs for which the results are reported are optimized based on the highest F1-score or the highest recall on the validation set, respectively.

“Revolutionary Advances” in Embeddings

- The year 2018 has been an inflection point for machine learning models handling text (or more accurately, Natural Language Processing or NLP for short). Our conceptual understanding of how best to represent words and sentences in a way that best captures underlying meanings and relationships is rapidly evolving.
  —Jay Alammar (http://jalammar.github.io/illustrated-bert/ — *good tutorial*)

- Bidirectional LSTM applied to learn context-specific embeddings (ELMo)
- Transformer architecture — focus on attention mechanism
- Bidirectional Encoder Representations from Transformers (BERT)
- Generative Pre-Training (GPT-2) — transformer with multi-task training, huge corpus, huge model

Sequence-to-Sequence models

- Natural application: machine translation
  - But also usable for question-answer problems
  - Equivalence and natural implication problems
  - Conversion from text to some formal representation
- One of a variety of RNN models

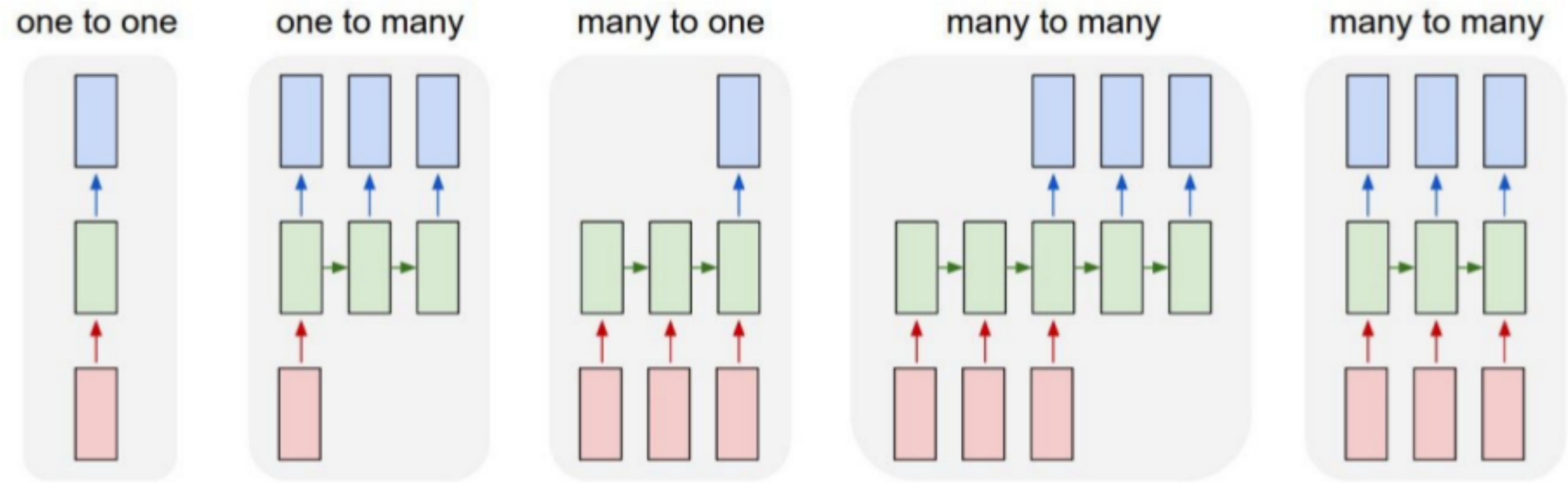


Image Captioning

Translation

Vanilla NN

Sentence Classification

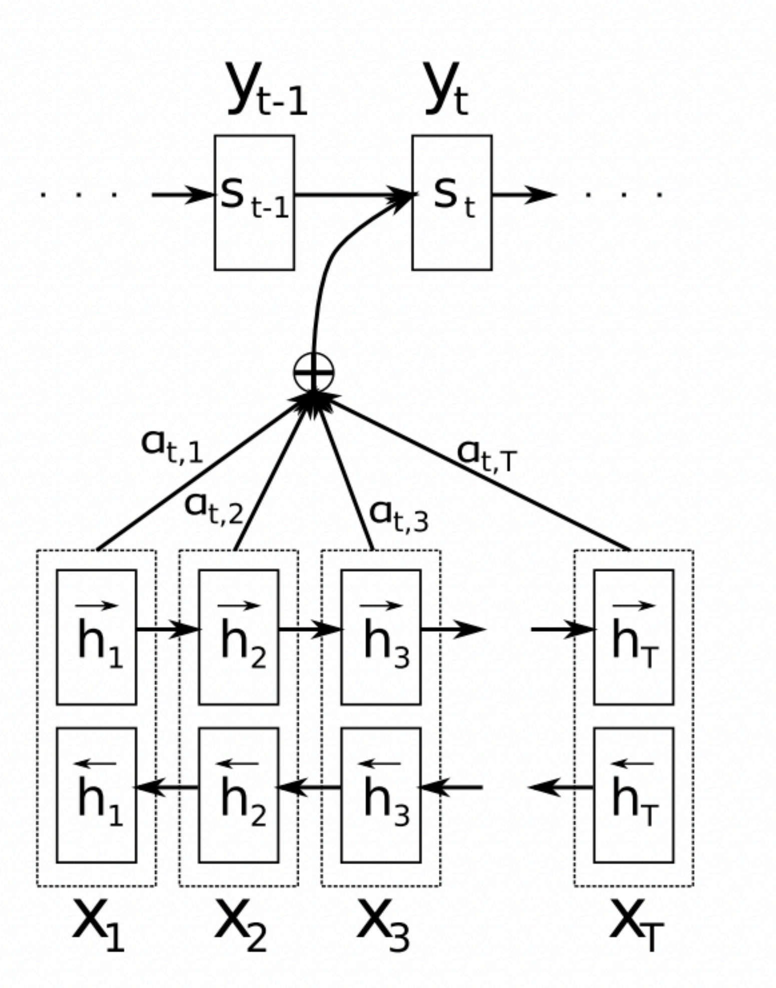
Sequence Classification

- For translation, odd to encode entire meaning of source into one state!

Slide adapted from Anna Rumshisky

Attention tells where in the source to focus

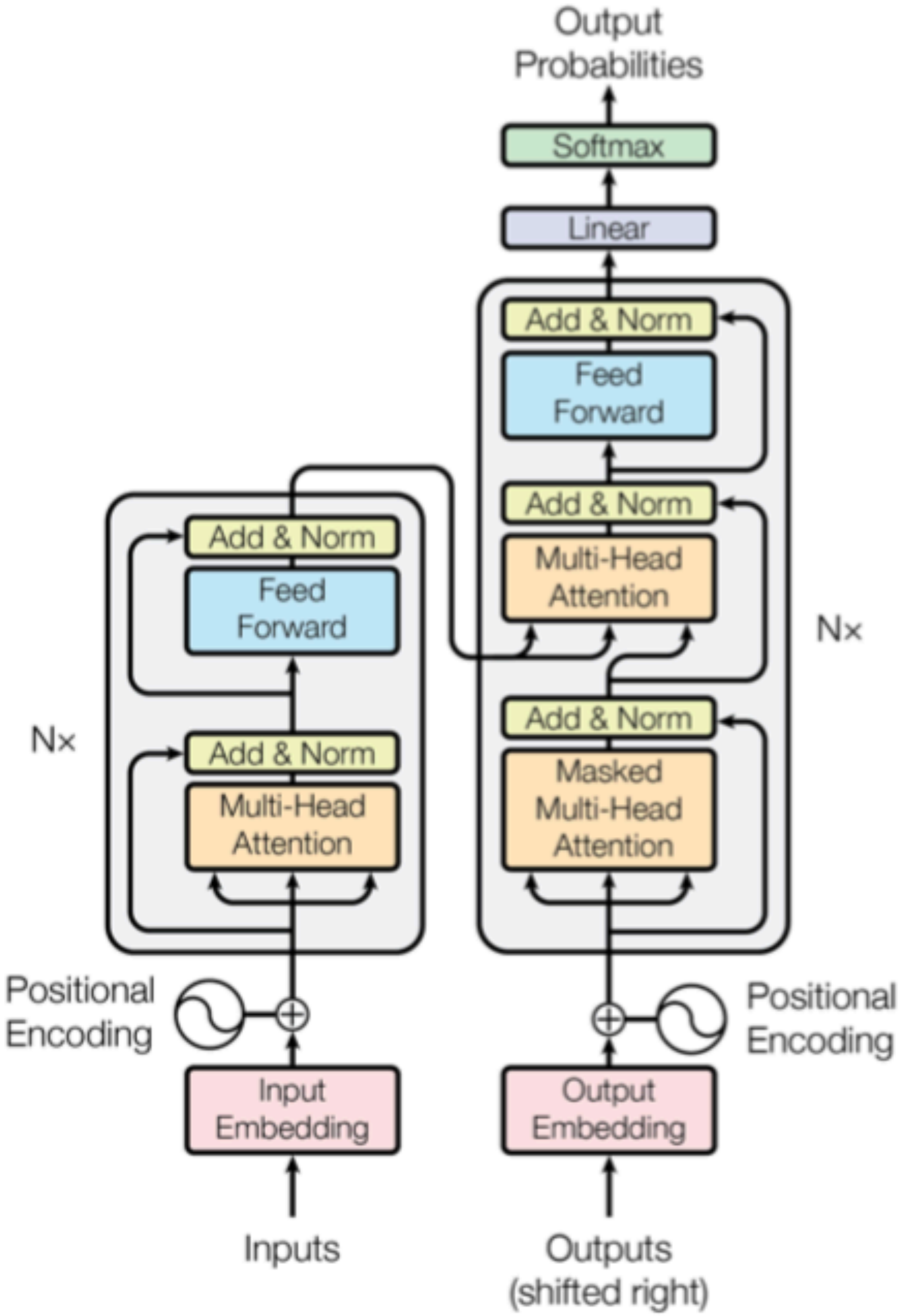
- Each decoder output word y_t now depends on a weighted combination of all the input states, not just the last state.
- The α's are weights that define how much of each input state should be considered for each output.
- Application: Automatic "alignment" of source and target languages in MT



Bahdanau, D., Cho, K., & Bengio, Y. (2014, September 1). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv*.

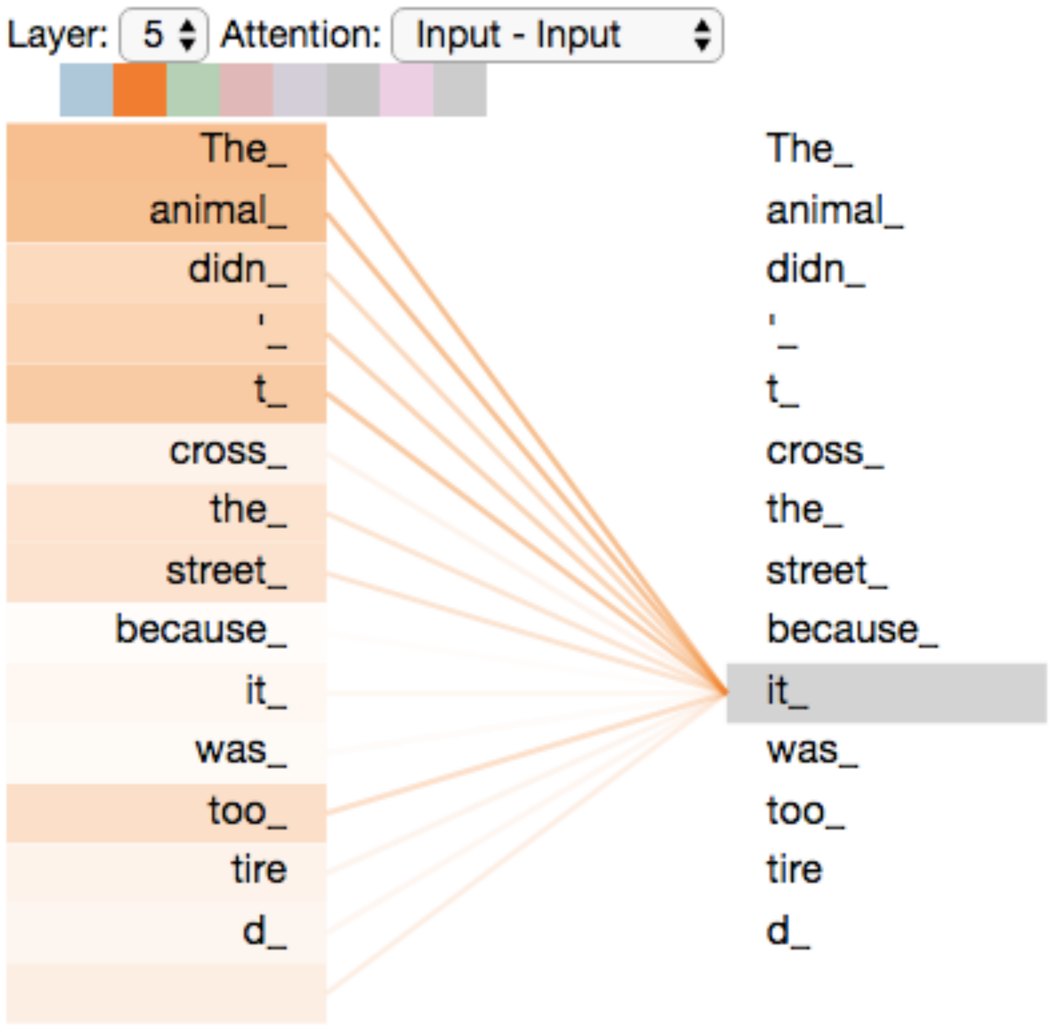
Transformer architecture

- Details well explained at https://jalammar.github.io/illustrated-transformer/
- Self-attention — vaguely reminiscent of CNNs
- Multi-headed attention — like multiple convolution kernels in CNN
- Key-value pairs passed from encoder to decoder
- Positional encoding
- Only look to left in decoder
- Scaling

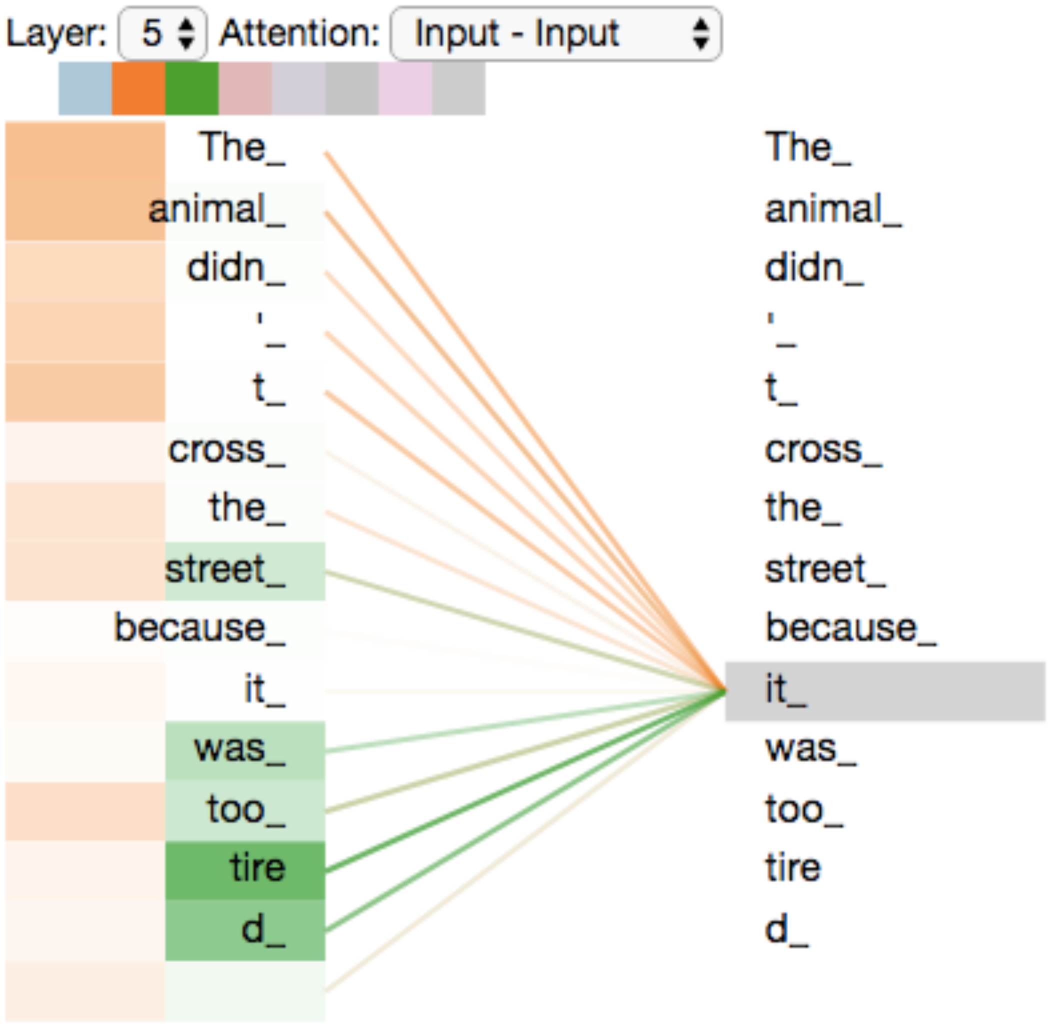


Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017, June 12). Attention Is All You Need. Lrec 2018.

Multi-headed attention



Layer: 5 Attention: Input - Input

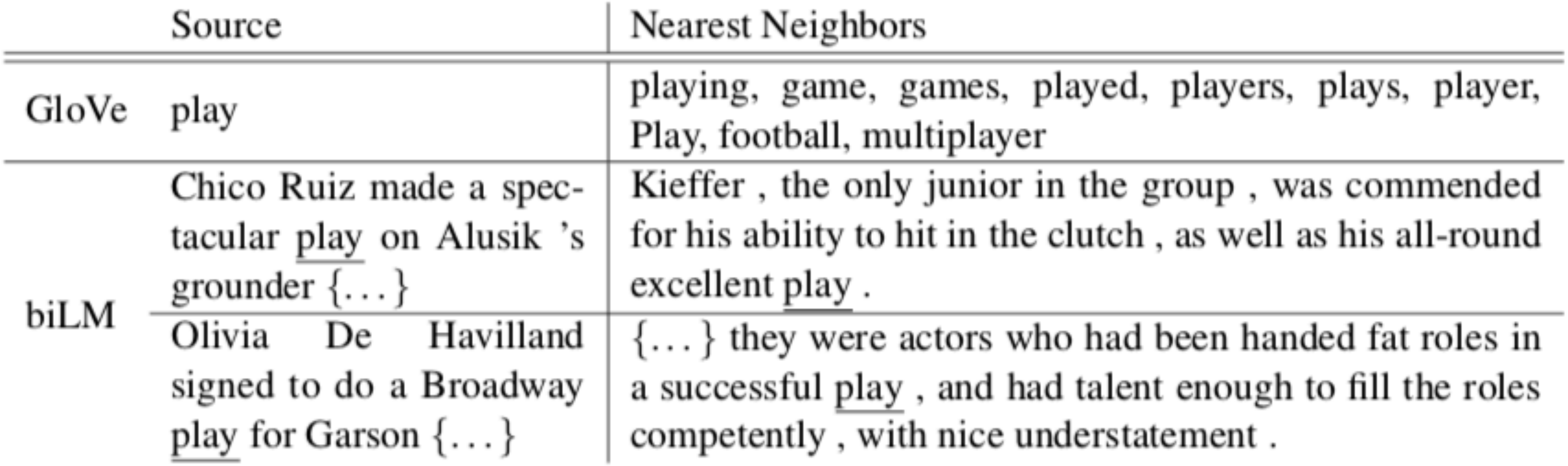


Layer: 5 Attention: Input - Input

https://jalammar.github.io/illustrated-transformer/

ELMo—Embeddings from Language Models

- Bidirectional LSTM
- Builds models for every *token*, not just for every *type*
  - i.e., different embeddings for the same word in different contexts
  - basis for word-sense disambiguation
- Significantly improves performance on nearly all NLP tasks



	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Table 4: Nearest neighbors to "play" using GloVe and the context embeddings from a biLM.

Peters, M. E., Neumann, M., Iyyer, M., 0001, M. G., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Naacl-Hlt*.

BERT
Bidirectional Encoder Representations from Transformers

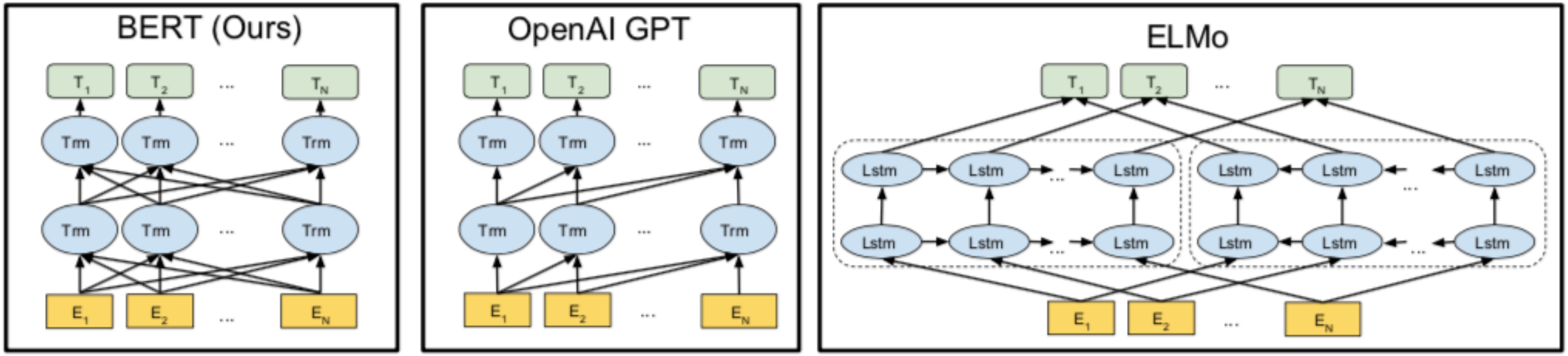


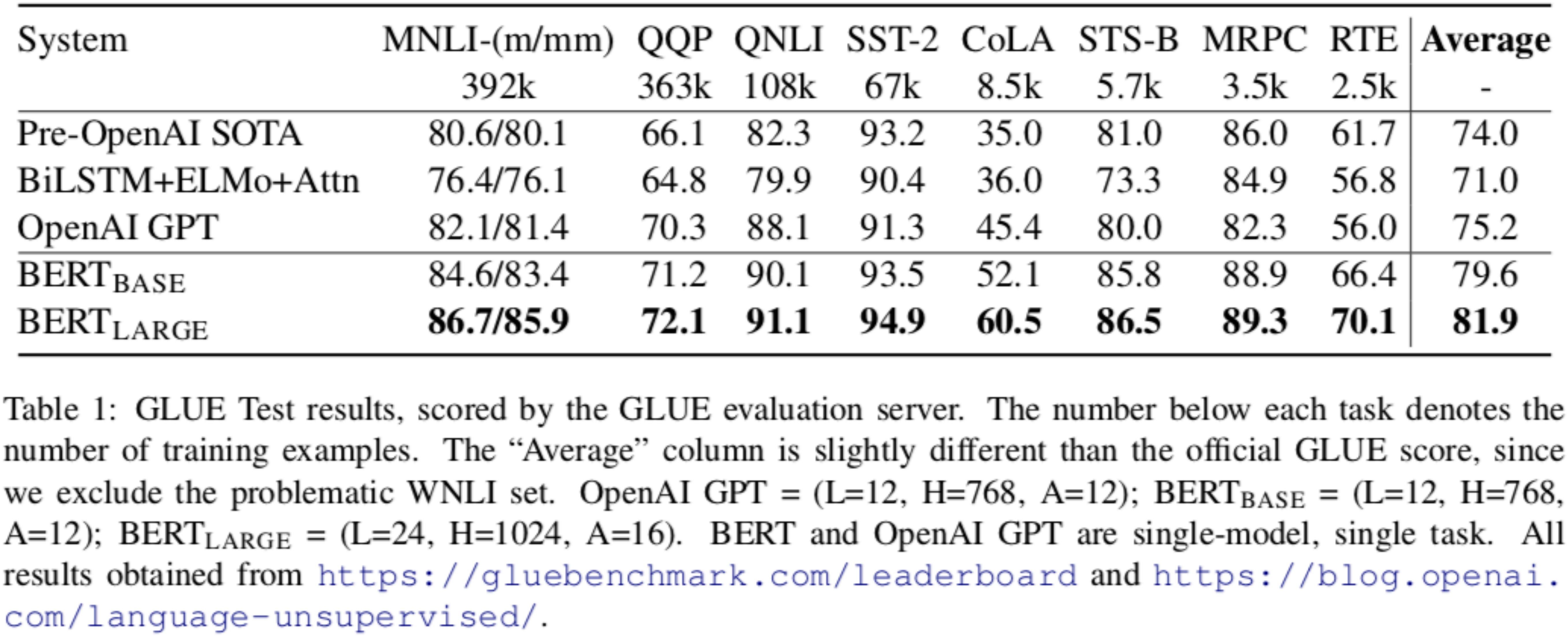
Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

- Word-piece tokens
- Predict masked tokens (~15%)
- Predict next sentence
- Trained on 800M word Books, 2,500M word Wikipedia corpus

- Large performance improvement on many tasks

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, October 10). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*.

BERT Performance Improvements



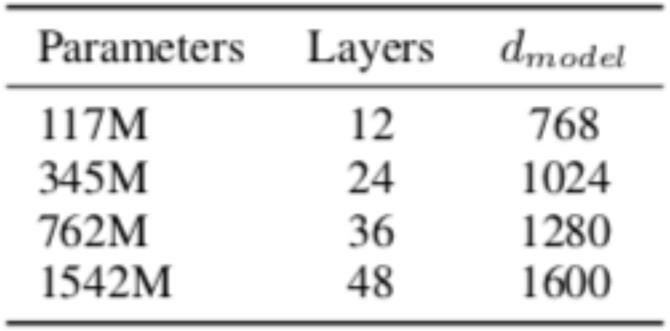
System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

Table 1: GLUE Test results, scored by the GLUE evaluation server. The number below each task denotes the number of training examples. The "Average" column is slightly different than the official GLUE score, since we exclude the problematic WNLI set. OpenAI GPT = (L=12, H=768, A=12); $\text{BERT}_{\text{BASE}}$ = (L=12, H=768, A=12); $\text{BERT}_{\text{LARGE}}$ = (L=24, H=1024, A=16). BERT and OpenAI GPT are single-model, single task. All results obtained from https://gluebenchmark.com/leaderboard and https://blog.openai.com/language-unsupervised/.

- MNLI Multi-Genre Natural Language Inference
- QQP Quora Question Pairs
- QNLI Question Natural Language Inference
- SST-2 The Stanford Sentiment Treebank
- CoLA The Corpus of Linguistic Acceptability
- STS-B The Semantic Textual Similarity Benchmark
- MRPC Microsoft Research Paraphrase Corpus
- RTE Recognizing Textual Entailment

GPT-2 (Generative Pre-Training)

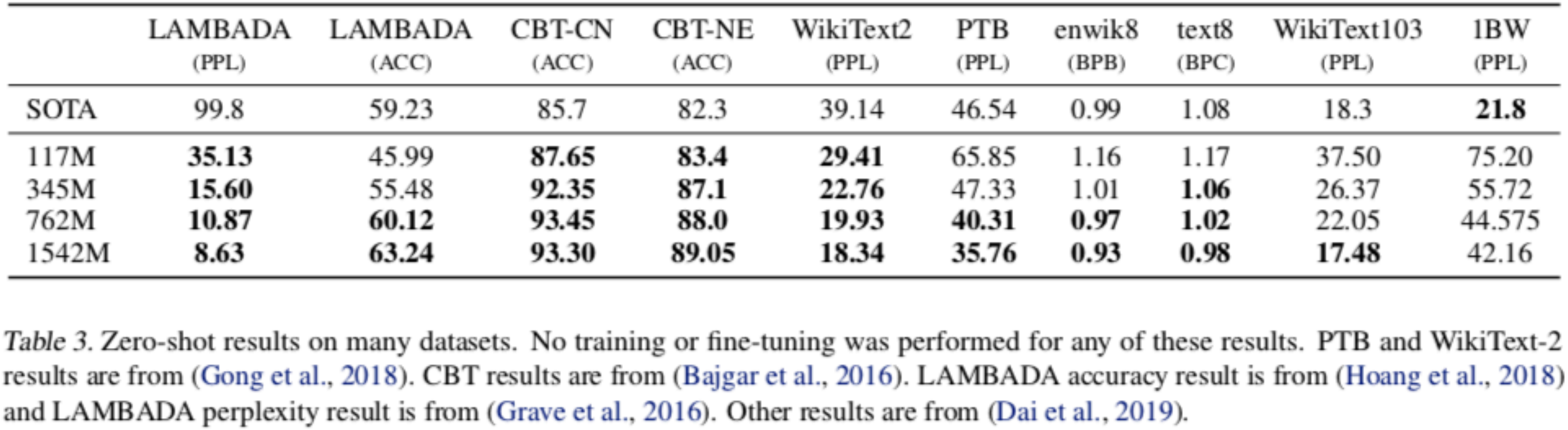
see https://blog.openai.com/better-language-models/



Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

- Unified transformer-based architecture for many tasks
- Task is itself given as a sequence of tokens, e.g.,
  - (translate to french, english text, french text)
  - (answer the question, document, question, answer)

- WebText: scrape of 8M documents linked from Reddit with “karma” ≥ 3
  - ~40GB
- Byte-level language models
- PPL = perplexity
- ACC = accuracy on a choose best next word out of 10



	LAMBADA	LAMBADA	CBT-CN	CBT-NE	WikiText2	PTB	enwik8	text8	WikiText103	1BW
	(PPL)	(ACC)	(ACC)	(ACC)	(PPL)	(PPL)	(BPB)	(BPC)	(PPL)	(PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019, February 14). Language Models are Unsupervised Multitask Learners.

Open-AI's GPT-2 model
Generative Pre-trained Transformer; ~1.5B parameters

prompt completion

- A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.
- The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

  In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

  "The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation," said Tom Hicks, the U.S. Energy Secretary, in a statement. "Our top priority is to secure the theft and ensure it doesn't happen again."

  The stolen material was taken from the University of Cincinnati's Research Triangle Park nuclear research site, according to a news release from Department officials.

  The Nuclear Regulatory Commission did not immediately release any information.

  According to the release, the U.S. Department of Energy's Office of Nuclear Material Safety and Security is leading that team's investigation.

  "The safety of people, the environment and the nation's nuclear stockpile is our highest priority," Hicks said. "We will get to the bottom of this and make no excuses."

https://blog.openai.com/better-language-models/#sample2