Intro to NLP in Medicine: Bag of Words Models

July 10, 2020

Applied Data Science
MMCi Term 4

Matthew Engelhard

Lecture Outline

- What is natural language processing?
- What can NLP do?
- What can NLP do in medicine?
- How does it work? (version 1)
- Next time: How does it work (version 2)

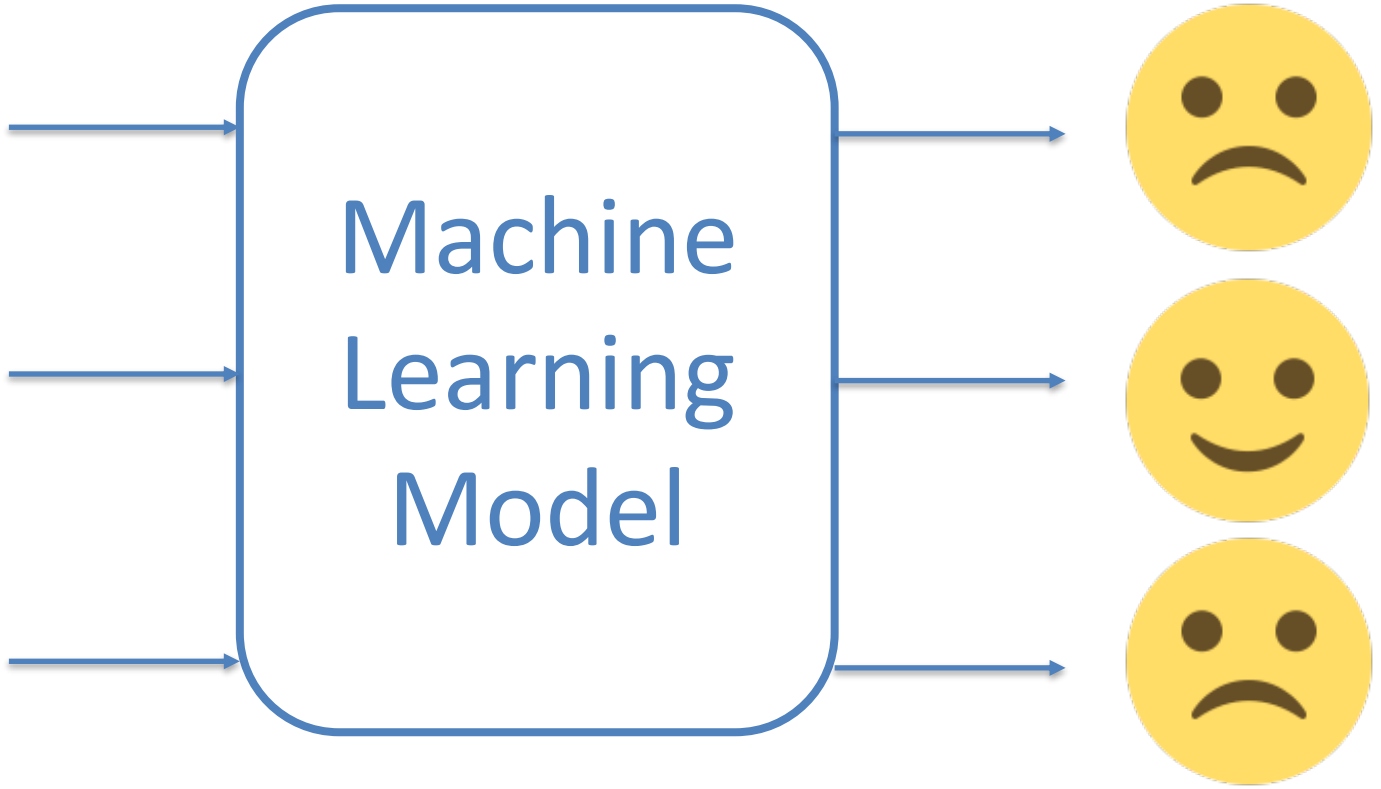
WHAT IS NATURAL LANGUAGE PROCESSING?

Sentiment Analysis

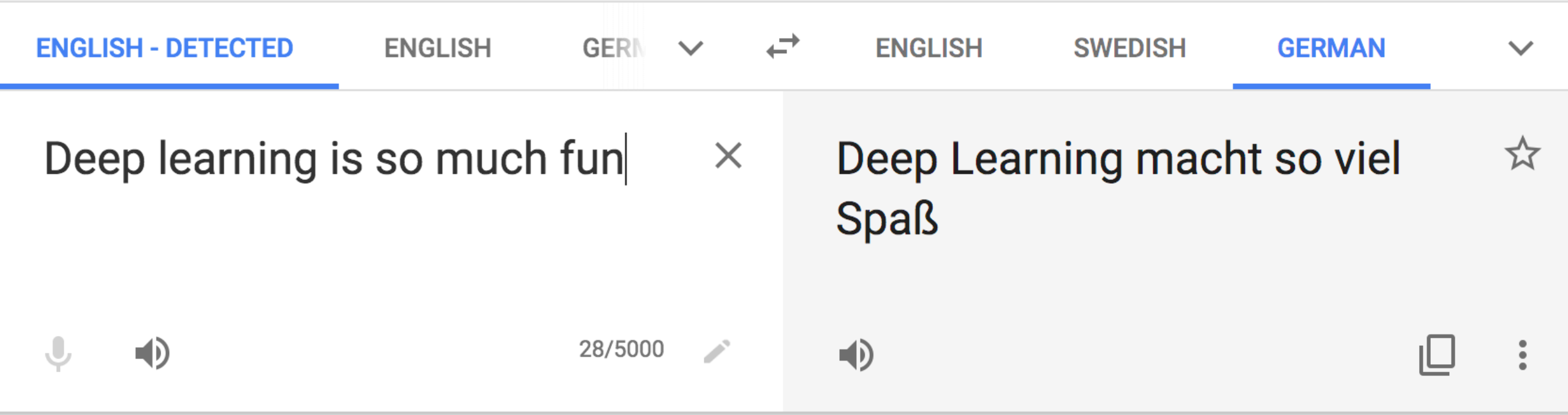
"That movie was terrible!"

"This is the best jacket that I've ever owned"

"Not really that good."



Translation



translate.google.com

Question Answering

Microorganisms or toxins that successfully enter an organism encounter the cells and mechanisms of the innate immune system. The innate response is usually triggered when microbes are identified by pattern recognition receptors, which recognize components that are conserve microorganisms, or when damaged, injur signals, many of which (but not all) are r those that recognize pathogens. Innate i meaning these systems respond to path not confer long-lasting immunity against is the dominant system of host defense

What part of the innate immune system identifies microbes and triggers immune response?

Ground Truth Answers: pattern recognition receptors | receptors | cells | tors

inant system of defense?

e system | innate immune

n

ize components present in broad

icroorganisms

in a generic way, meaning it is

non-specific | non-specific

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	**86.673**	**89.147**
2 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	85.150	87.715

Text Generation

SYSTEM PROMPT (HUMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

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.

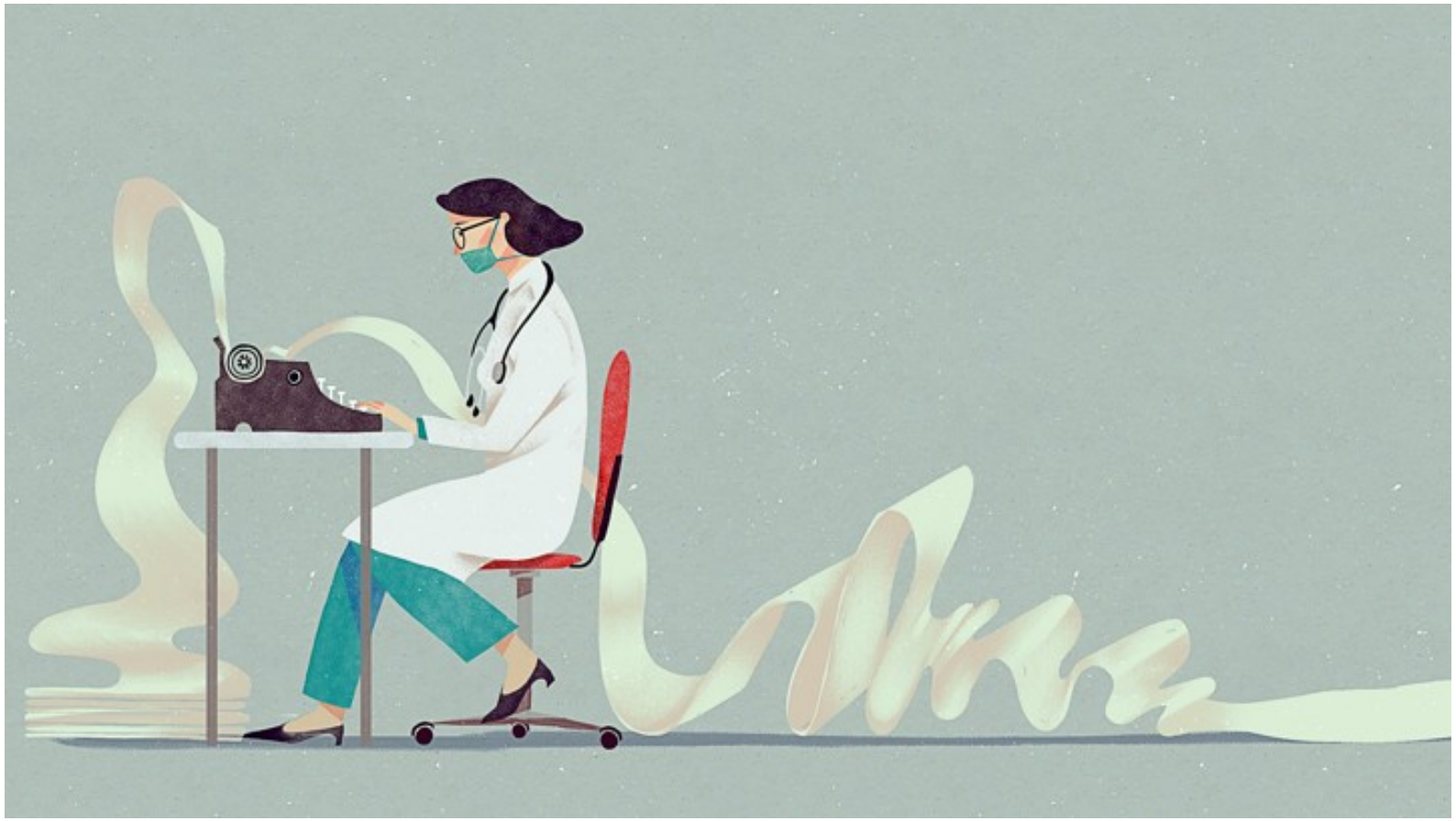
“The safety of people, the environment and the nation’s nuclear stockpile is our highest priority,” Hicks said. “We will get to the bottom of this and make no excuses.

“Better Language Models and Their Implications”
2/14/19
OPENAI

NLP IS CENTRAL TO AI-ASSISTED MEDICINE

Reducing Burden and Restoring Patient-Provider Interaction

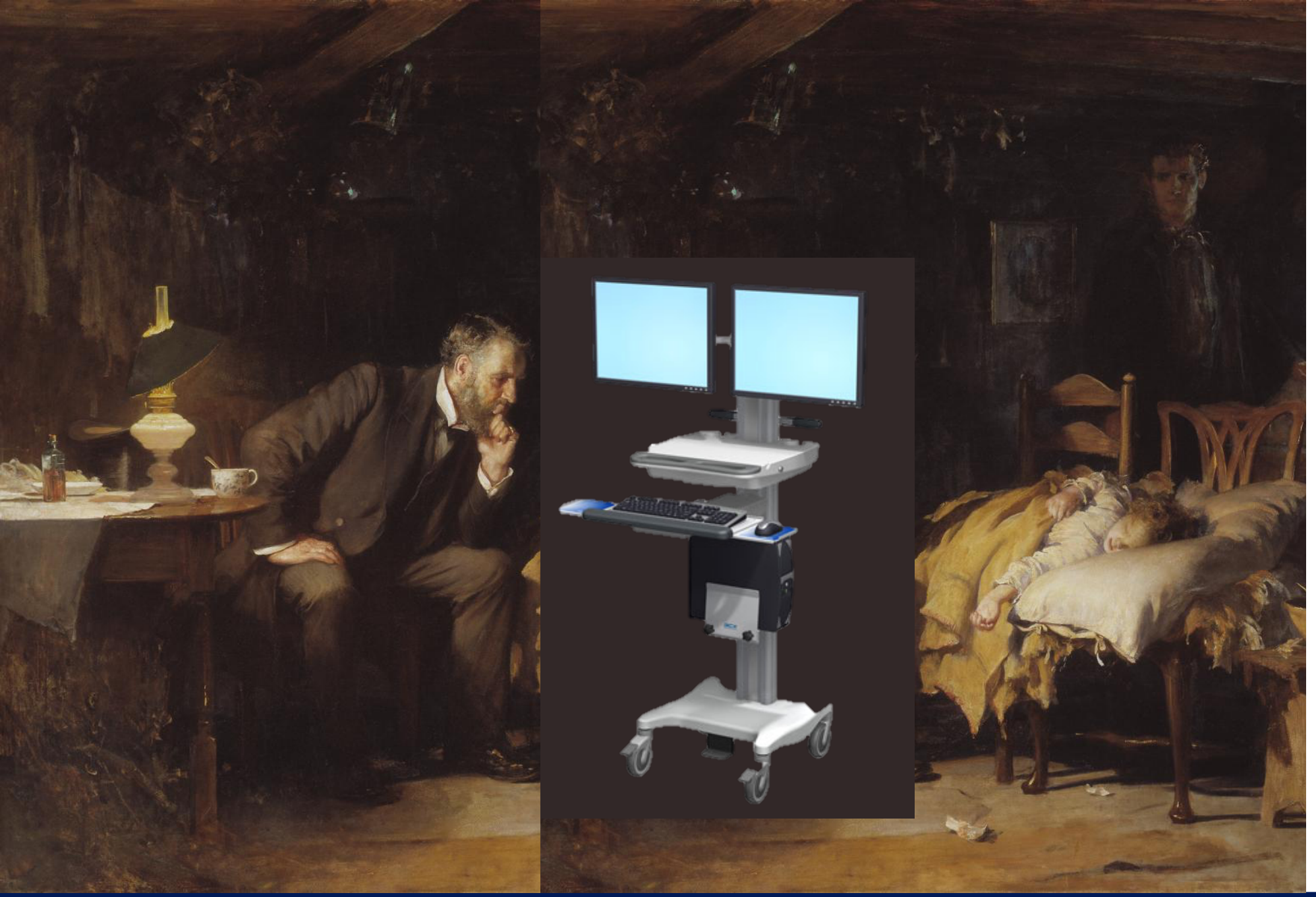
The Burnout Crisis in American Medicine
Rena Xu
The Atlantic
May 11, 2018



The Doctor

(Luke Fildes, 1891)

Inspired by MLHC keynote by Abraham Verghese, MD, MACP, Stanford University



The Doctor,

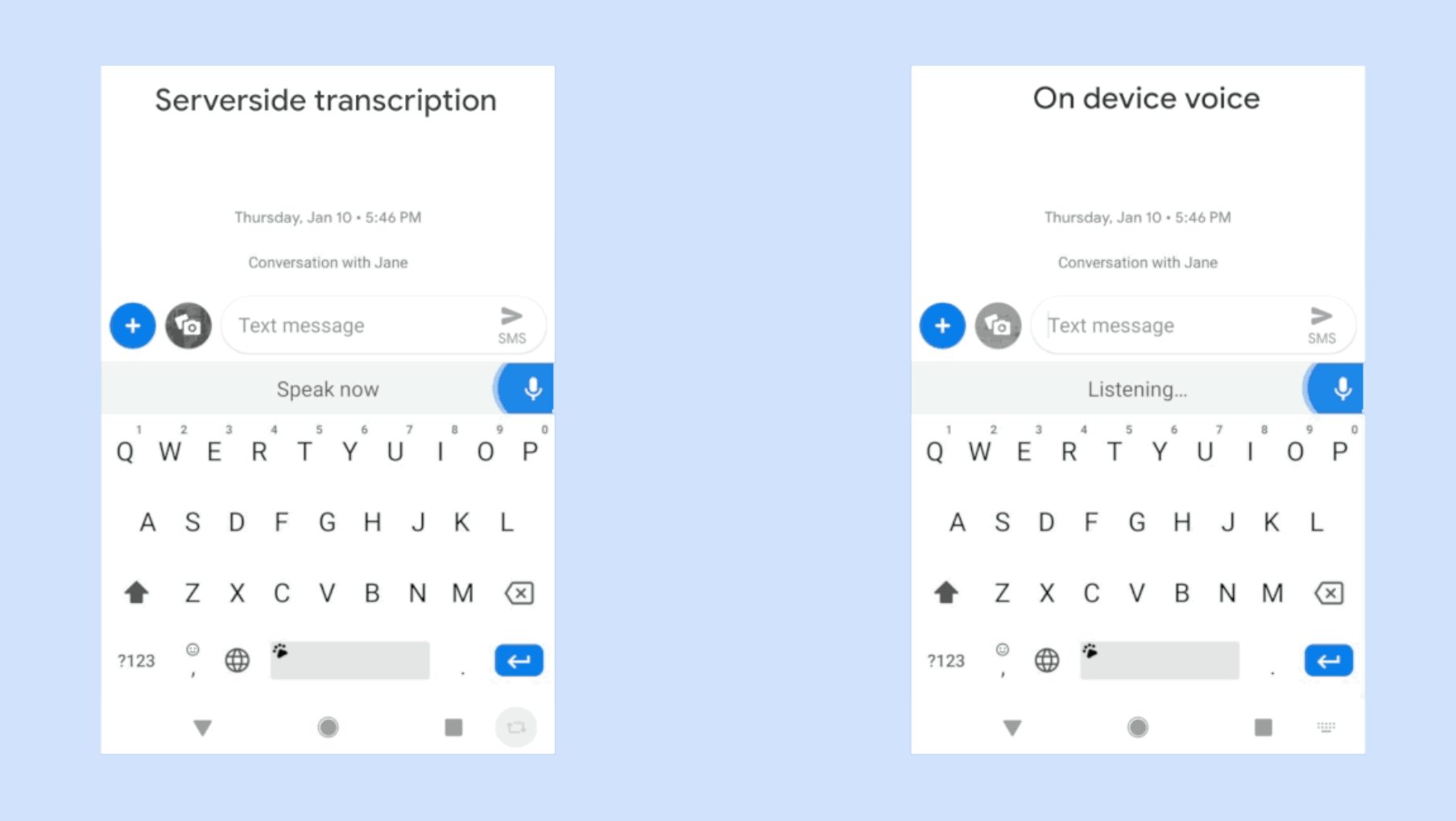
circa 2018

Inspired by MLHC keynote by Abraham Verghese, MD, MACP, Stanford University

The Doctor,

circa 2030

Speech Processing is Largely Solved



https://ai.googleblog.com/2019/03/an-all-neural-on-device-speech.html

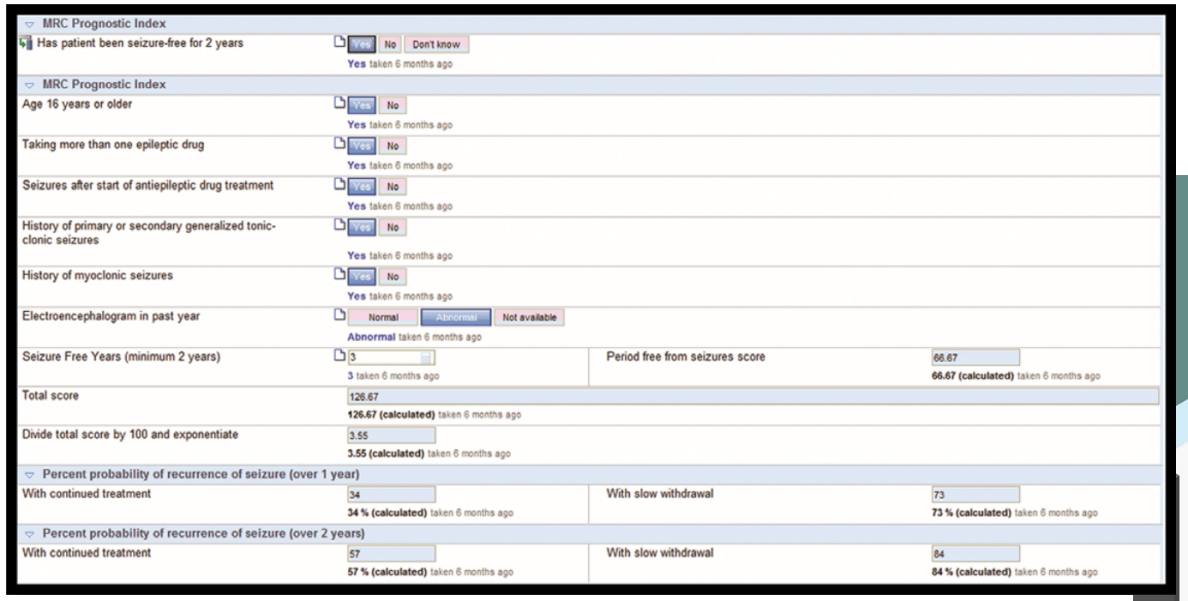
The Doctor,

circa 2030

Use NLP to:

- **Populate standardized notes and fields**

Populating Standardized Forms



MRC Prognostic Index

Has patient been seizure-free for 2 years	Yes / No / Don't know Yes taken 6 months ago		
MRC Prognostic Index			
Age 16 years or older	Yes / No Yes taken 6 months ago		
Taking more than one epileptic drug	Yes / No Yes taken 6 months ago		
Seizures after start of antiepileptic drug treatment	Yes / No Yes taken 6 months ago		
History of primary or secondary generalized tonic-clonic seizures	Yes / No Yes taken 6 months ago		
History of myoclonic seizures	Yes / No Yes taken 6 months ago		
Electroencephalogram in past year	Normal / Abnormal / Not available Abnormal taken 6 months ago		
Seizure Free Years (minimum 2 years)	3 3 taken 6 months ago	Period free from seizures score	66.67 66.67 (calculated) taken 6 months ago
Total score	126.67 126.67 (calculated) taken 6 months ago		
Divide total score by 100 and exponentiate	3.55 3.55 (calculated) taken 6 months ago		
Percent probability of recurrence of seizure (over 1 year)			
With continued treatment	34 34 % (calculated) taken 6 months ago	With slow withdrawal	73 73 % (calculated) taken 6 months ago
Percent probability of recurrence of seizure (over 2 years)			
With continued treatment	57 57 % (calculated) taken 6 months ago	With slow withdrawal	84 84 % (calculated) taken 6 months ago

Narayanan et al, *Epilepsia* (2017)

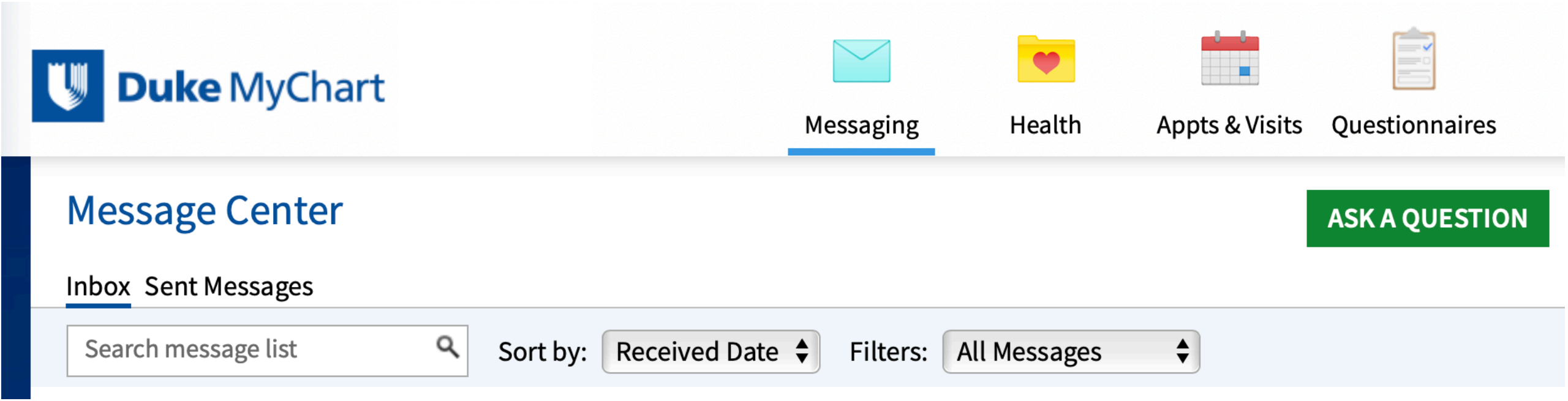
The Doctor,

circa 2030

Use NLP to:

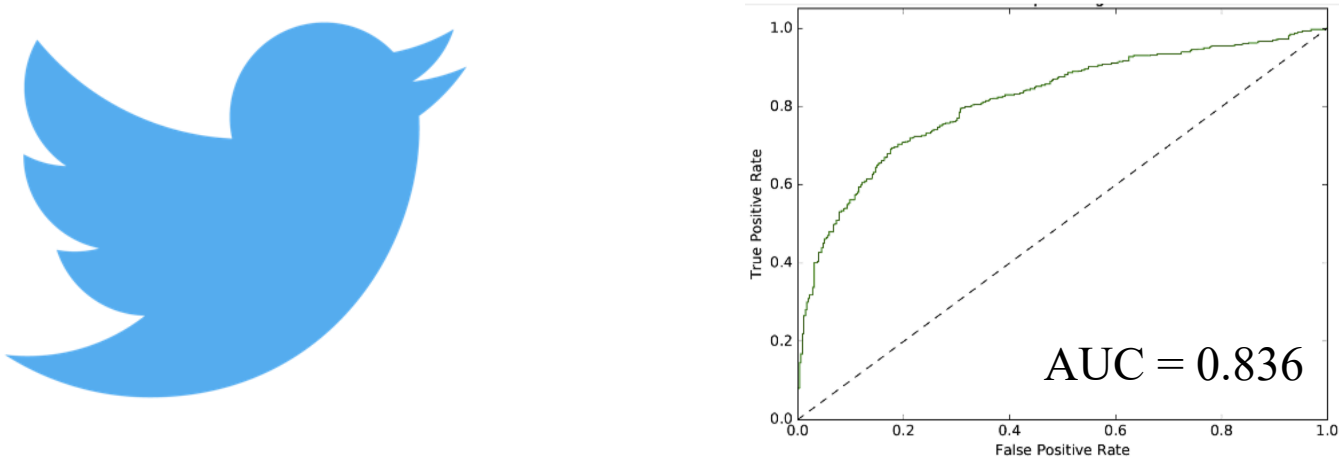
- Populate standardized notes and fields
- **Answer new questions from transcripts as needed**

Suggested Email Responses



OUR FOCUS: CLASSIFICATION

Mental Health via Social Media



Guntuku, Sharath Chandra, et al. "**Language of ADHD in Adults on Social Media**." *Journal of attention disorders* (2017): 1087054717738083.

Binary Classification: ADHD or not ADHD

Mental Health via Social Media

r/depression → ? → r/SuicideWatch

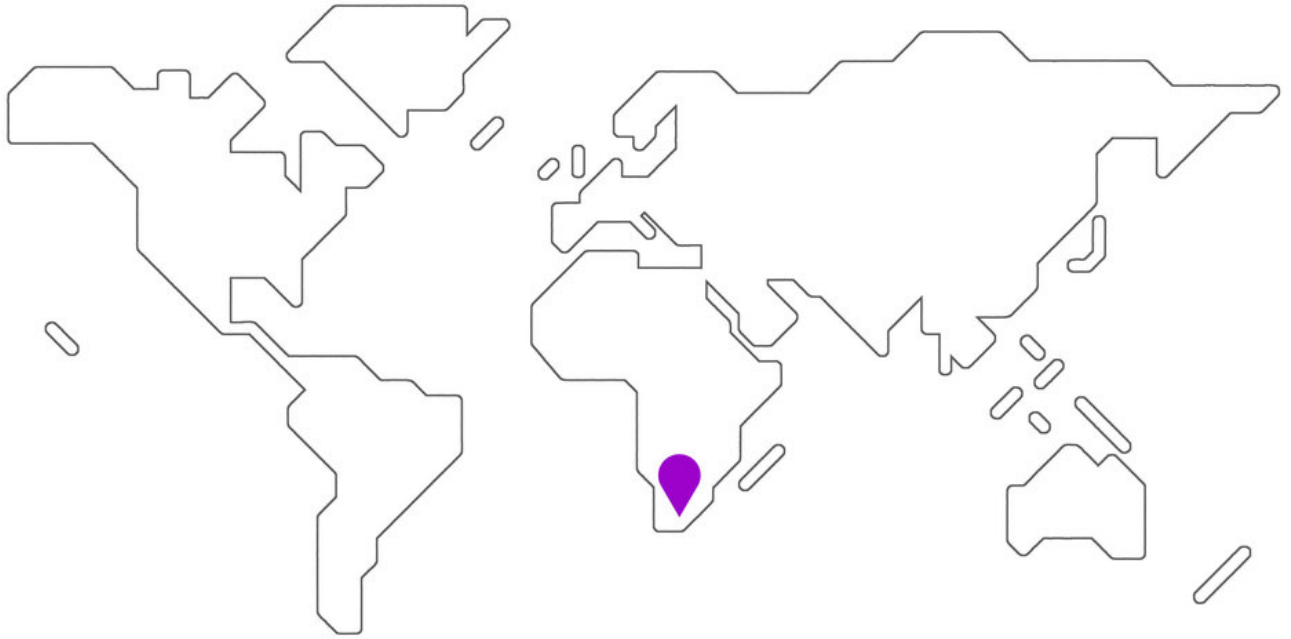
De Choudhury, Munmun, et al. "**Discovering shifts to suicidal ideation from mental health content in social media**." *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 2016.

Binary Classification: Suicidal Ideation? (Yes/No)

Global Maternal Health

Maternal Health HelpDesk:

2 million women connected to NDoH staff via SMS

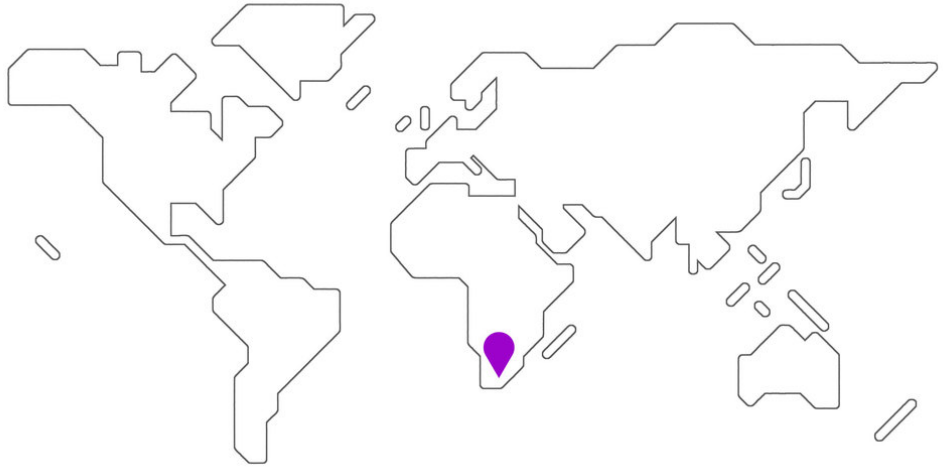


https://www.praekelt.org

Binary Classification: Urgent Message? (Yes/No)

HOW DOES THIS FIT IN OUR PREVIOUS FRAMEWORK?

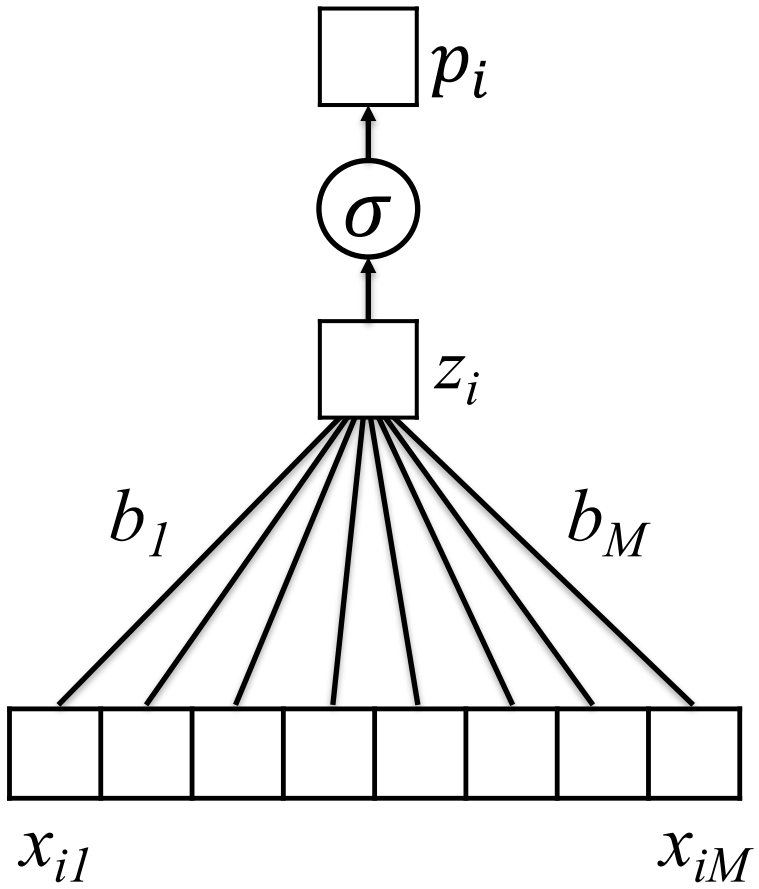
Text Classification



https://www.praekelt.org

Can we use logistic regression to solve this problem?

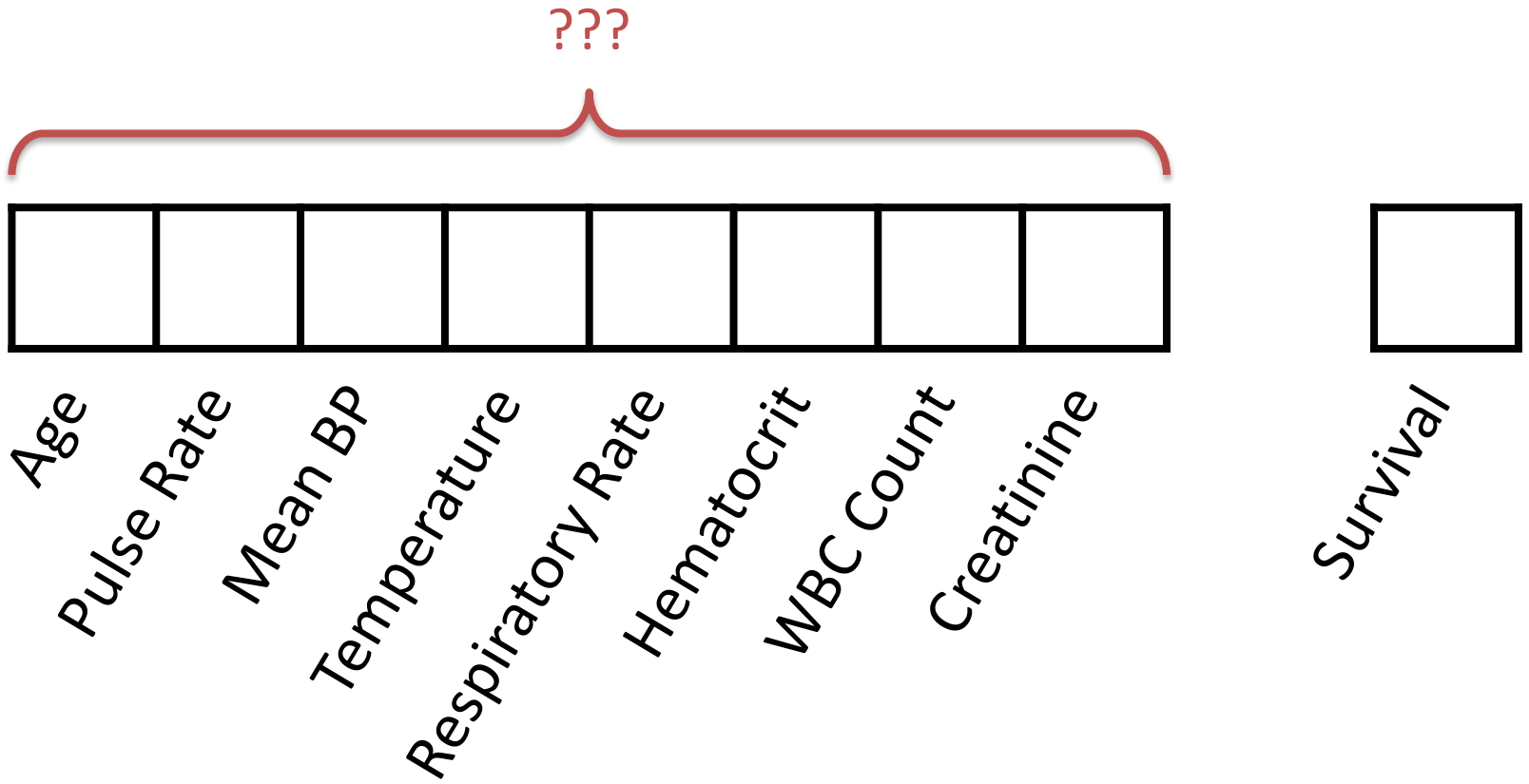
How can we use logistic regression for text classification?



This looks good…

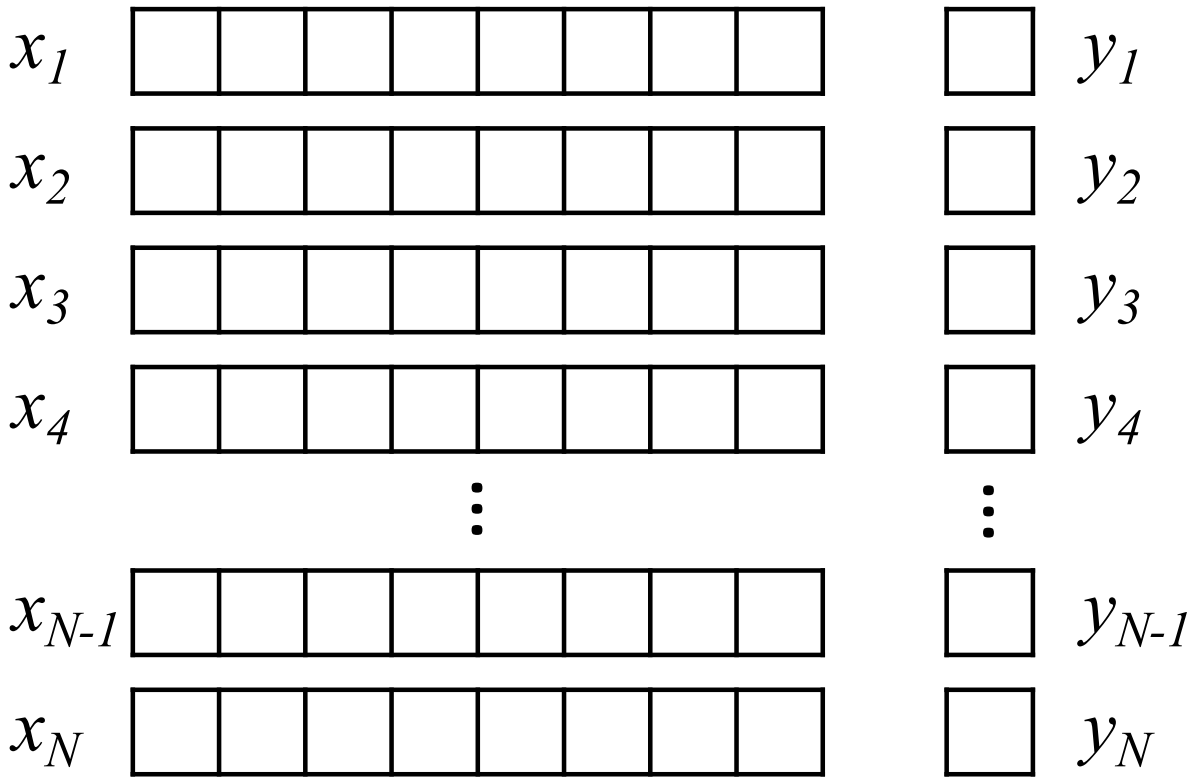
But what are our features??

ICU Mortality: APACHE III



End goal: predict odds of hospital mortality

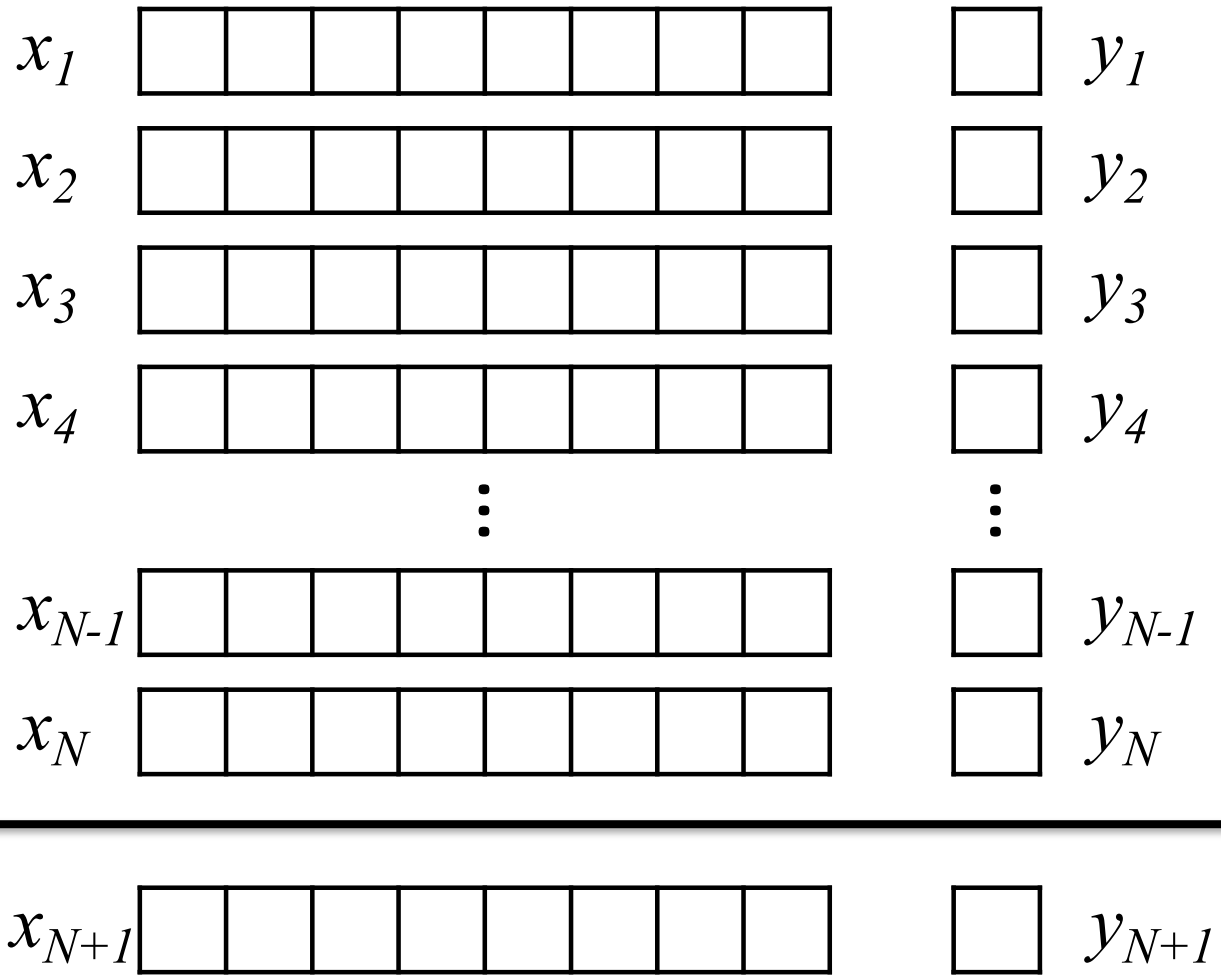
Training Set (Historical Data)



Find an equation that predicts y based on x across the training set

We'll begin by supposing y is binary (i.e. $y \in \{0, 1\}$)

Making Predictions for New x



Find an equation that predicts y based on x across the training set

We'll begin by supposing y is binary (i.e. $y \in \{0, 1\}$)

<- Learn to predict new y

This time, our training data is text

x_1	What helps with morning sickness?	☐	y_1
x_2	How many months should I breastfeed?	☐	y_2
x_3	I passed out and Mom said I was shaking	☐	y_3
x_4	Where is the nearest clinic?	☐	y_4
⋮		⋮	
x_{N-1}	I am having heavy bleeding, what should I do?	☐	y_{N-1}
x_N	What foods should I eat while pregnant?	☐	y_N

y_i: Urgent or Not Urgent?

---

x_{N+1}	My heart is racing and I can't catch my breath	☐	y_{N+1}

<- Learn to predict new y

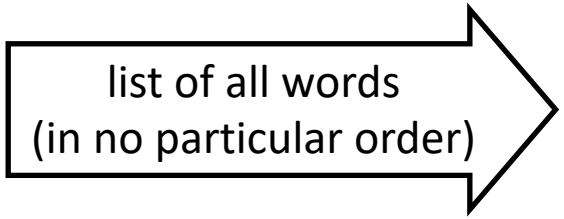
We need numbers, not words

- **Can we convert our text to a vector or sequence of numbers?**

- If yes, we can start using our previous methodology!

First try: count words in each SMS
Step 1: Define a **vocabulary** of words

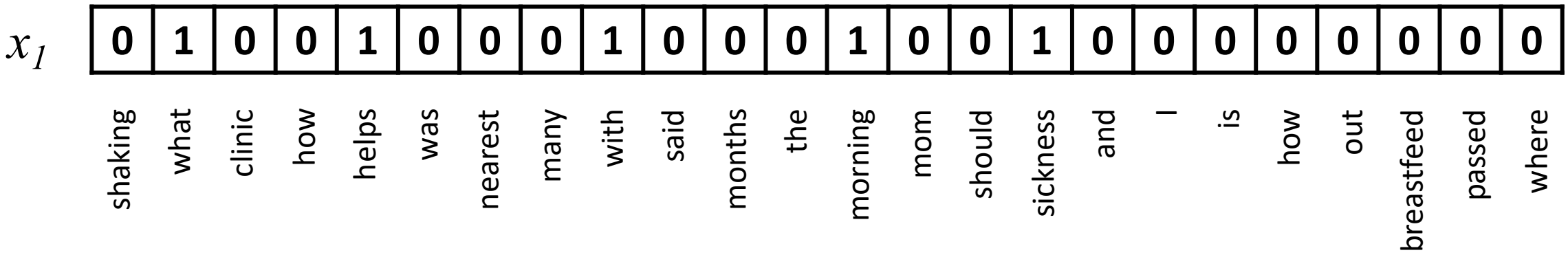
x_1	What helps with morning sickness?
x_2	How many months should I breastfeed?
x_3	I passed out and Mom said I was shaking
x_4	Where is the nearest clinic?



shaking	with	and
what	said	I
clinic	months	is
how	the	how
helps	morning	out
was	mom	breastfeed
nearest	should	passed
many	sickness	where

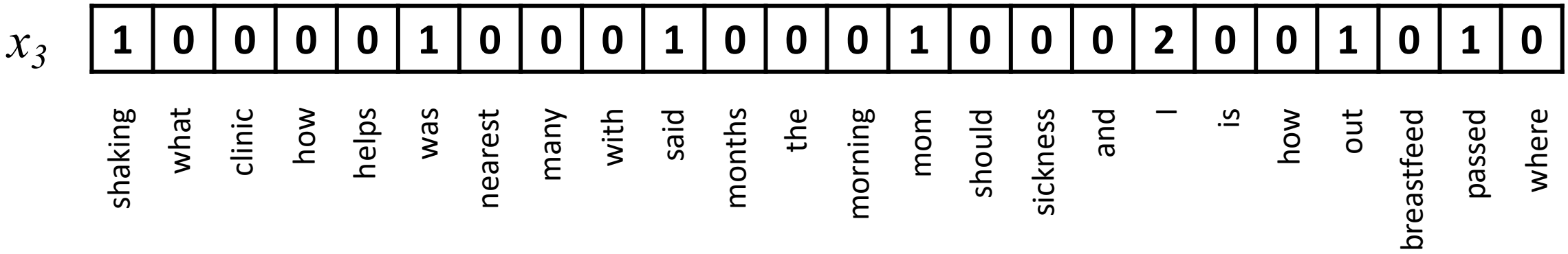
Step 2: count how many times each vocabulary word appears in a given SMS

What helps with morning sickness?



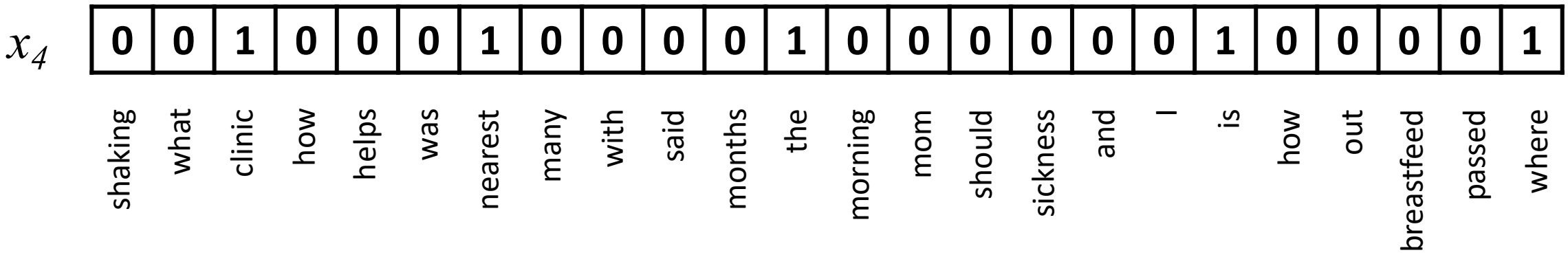
Step 2: count how many times each vocabulary word appears in a given SMS

I passed out and Mom said I was shaking



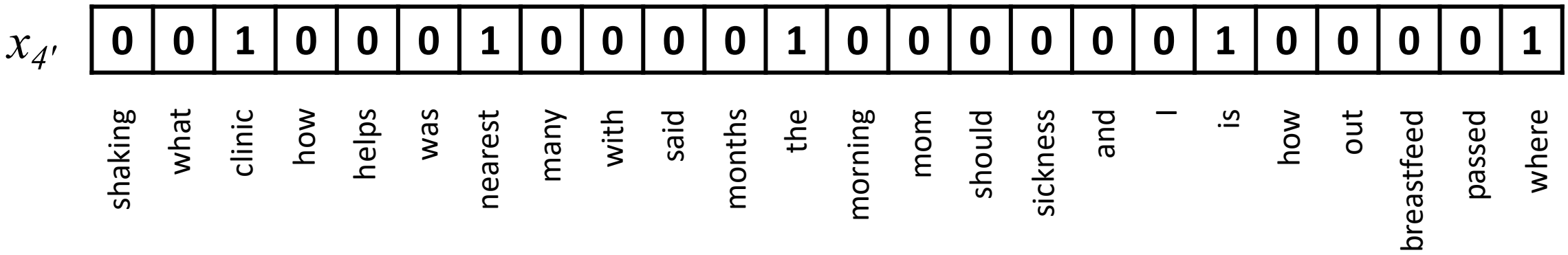
Step 2: count how many times each vocabulary word appears in a given SMS

Where is the nearest clinic?



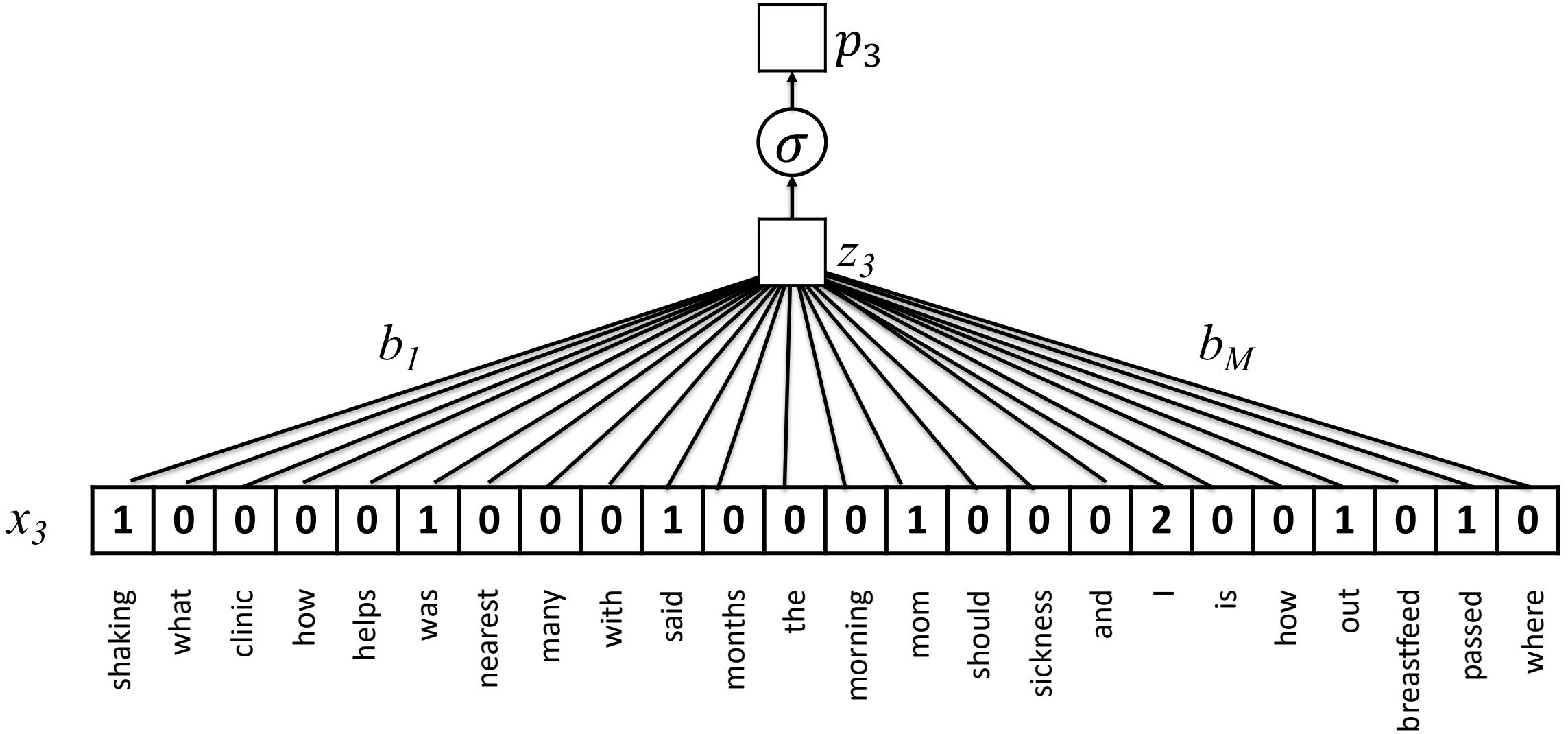
Note that word order does not matter!

clinic is where nearest the



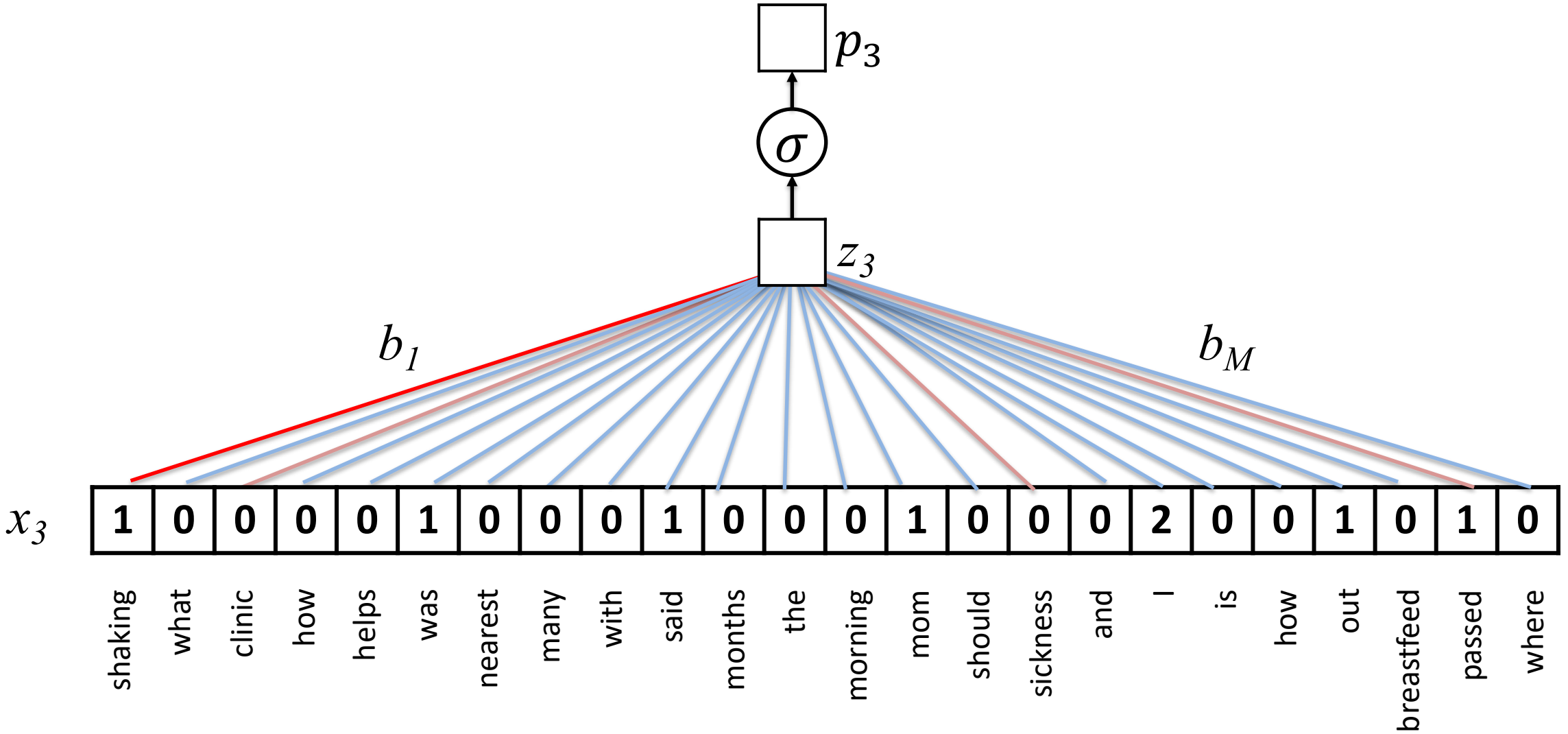
A “bag of words”

Logistic Regression for Text Classification



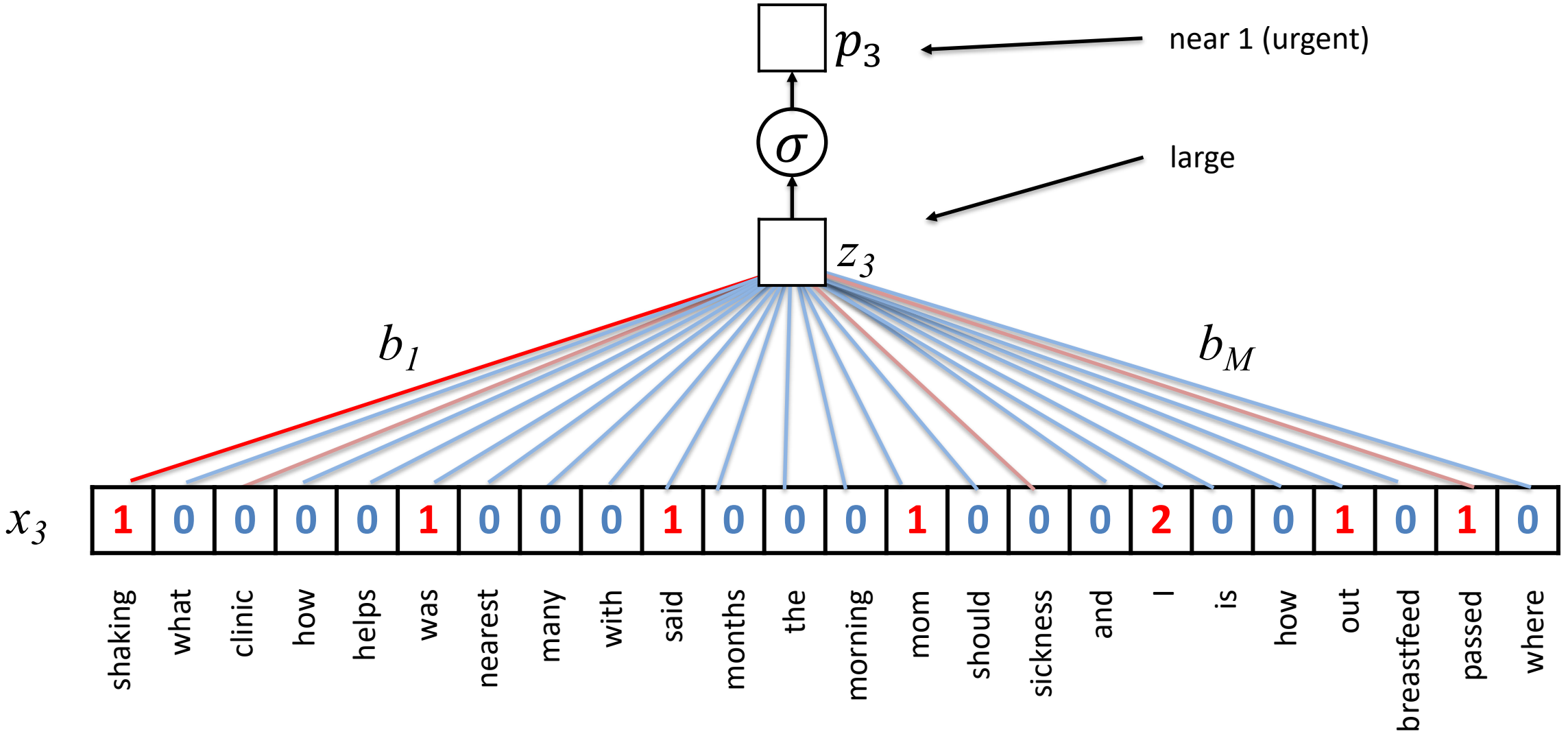
I passed out and Mom said I was shaking

Logistic Regression for Text Classification



I passed out and Mom said I was shaking

Logistic Regression for Text Classification



I passed out and Mom said I was shaking

Strengths and Weaknesses

- (+) This approach is simple and works surprisingly well in practice
- (+) Often the best approach with small datasets
- (-) Does not capture word order
- (-) Does not group synonyms together or understand semantic relationships between words

2nd try: count 1- and 2-grams in each SMS (i.e. extend vocabulary to include 2-word phrases)

x_1	What helps with morning sickness?
x_2	How many months should I breastfeed?
x_3	I passed out and Mom said I was shaking
x_4	Where is the nearest clinic?

1-grams

shaking	was	months	sickness	out
what	nearest	the	and	breastfeed
clinic	many	morning	I	passed
how	with	mom	is	where
helps	said	should	how	

2-grams

what helps	I breastfeed	was shaking
helps with	I passed	where is
with morning	passed out	is the
morning sickness	out and	the nearest
how many	and mom	nearest clinic
many months	mom said	
months should	said I	
should I	I was	

-gram[1]

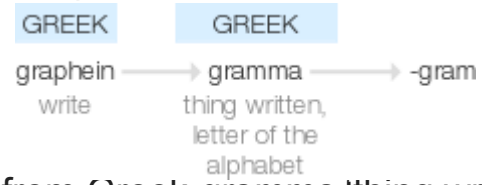
combining form

suffix: **-gram**

1.in nouns denoting something written or recorded (especially in a certain way).

"cryptogram"

Origin

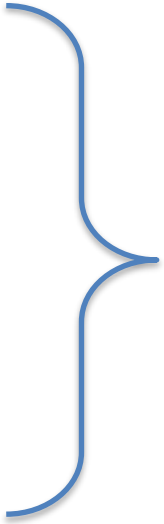


from Greek *gramma* ‘thing written, letter of the alphabet’, from *graphein* ‘write’.

n-grams can be very helpful!

I am not sick and feel great

I am not great and feel sick

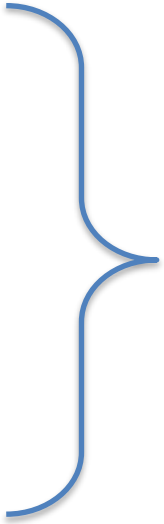


Bag of 1-grams: no difference between these sentences

n-grams can be very helpful!

I am not sick and feel great

I am not great and feel sick



Bag of 1- and 2-grams:

not sick, feel great

versus

not great, feel sick

3rd try: more powerful methods to work with…

- (a) word meaning: assign words to vectors that encode their meaning numerically
- (b) words in context: neural network architectures that act on *sequences* of words (rather than a bag of words)
- More on this next lecture

Bag of Words

MISCELLANEOUS TEXT PROCESSING DETAILS

Variations on counting: term frequency

term count: ‘times’

2

“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.”

1

“And the first one now
Will later be last
For the times they are a-changin'.”

Variations on counting: term frequency

term frequency: ‘times’

2/119

“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.”

1/16

“And the first one now
Will later be last
For the times they are a-changin'.”

-> better measure of the importance of the term within a given text sample

Variations on counting: inverse document frequency

2/2

document frequency: ‘times’

“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.”

“And the first one now
Will later be last
For the times they are a-changin'.”

Variations on counting: inverse document frequency

1/2

document frequency: ‘evil’

“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.”

X

“And the first one now
Will later be last
For the times they are a-changin'.”

term frequency-inverse document frequency (tf-idf)

- What helps with morning sickness?
- How many months should I breastfeed?
- I passed out and Mom said I was shaking
- Where is the nearest clinic?
- I am having heavy bleeding, what should I do?
- What foods should I eat while pregnant?
- My heart is racing and I can't catch my breath

$$\frac{\text{term frequency}}{\text{document frequency}}$$ for 'shaking'

$$\frac{1/9}{1/7} = .78$$

$$\frac{\text{term frequency}}{\text{document frequency}}$$ for 'I'

$$\frac{2/9}{5/7} = .31$$

Preprocessing

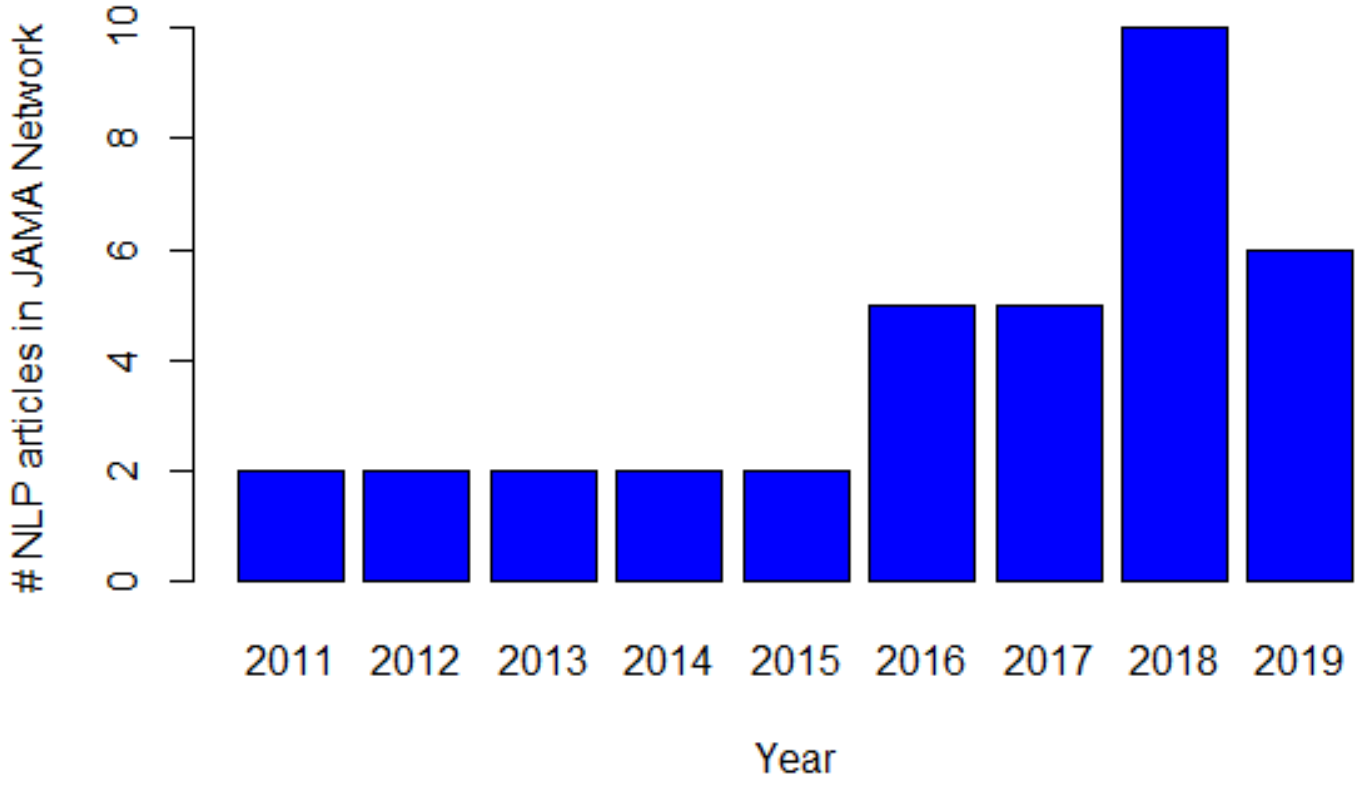
- remove punctuation
- to lowercase
- “tokenization”
- “stemming”

```
I passed out, and Mom said I was shaking.

I passed out and Mom said I was shaking

i passed out and mom said i was shaking

[i, passed, out, and, mom, said, i, was, shaking]

[i, pass, out, and, mom, said, i, wa, shake]
```

(mostly bag of words)

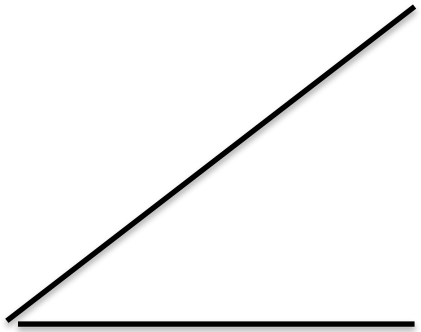
NLP IN MEDICINE SO FAR

A Survey of NLP in JAMA…



A Survey of NLP in JAMA…

28 research articles



1 uses word vectors

28 use word counts or frequencies

- The majority of these search only for a limited number of keywords or expressions

A Survey of NLP in JAMA…

28 research articles

4 utilize EHR text features in predictive models (e.g. mortality)

20 identify specific diagnoses or medical events from the EHR

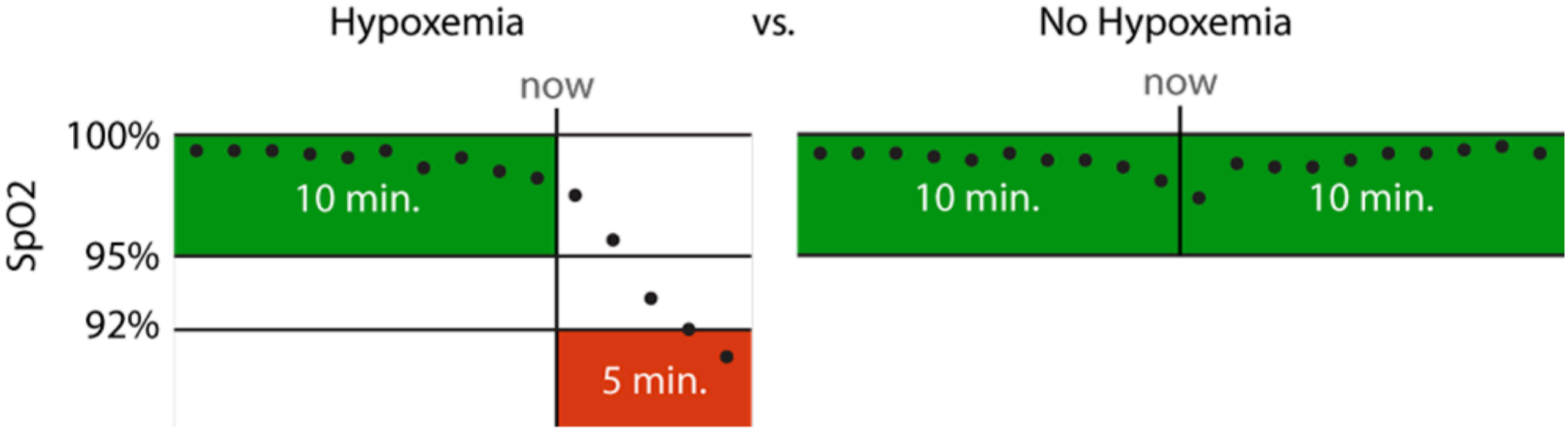
4 identify other care-related information in the EHR

0 interpret words *in context*

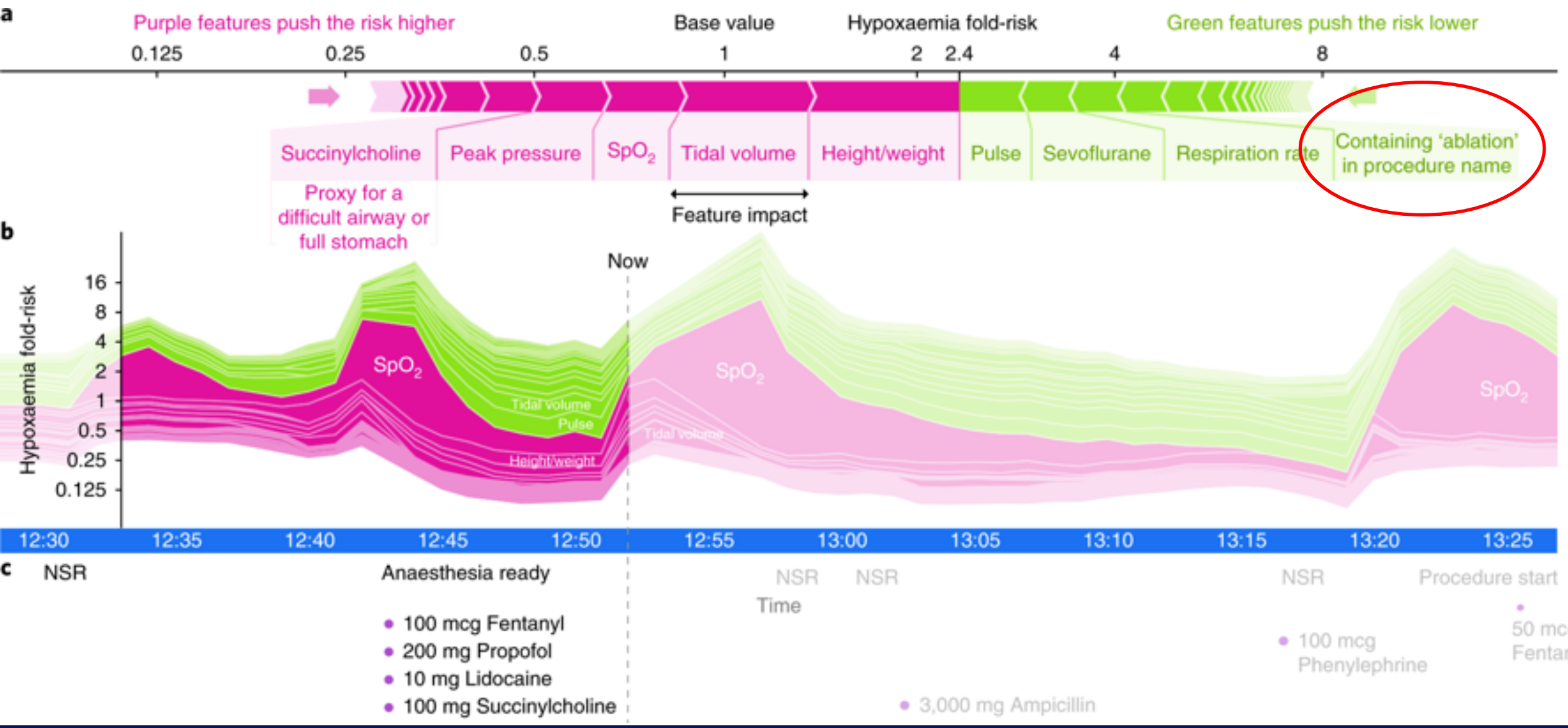
Hypoxemia Prediction during Surgery

Real-time Prediction Task:

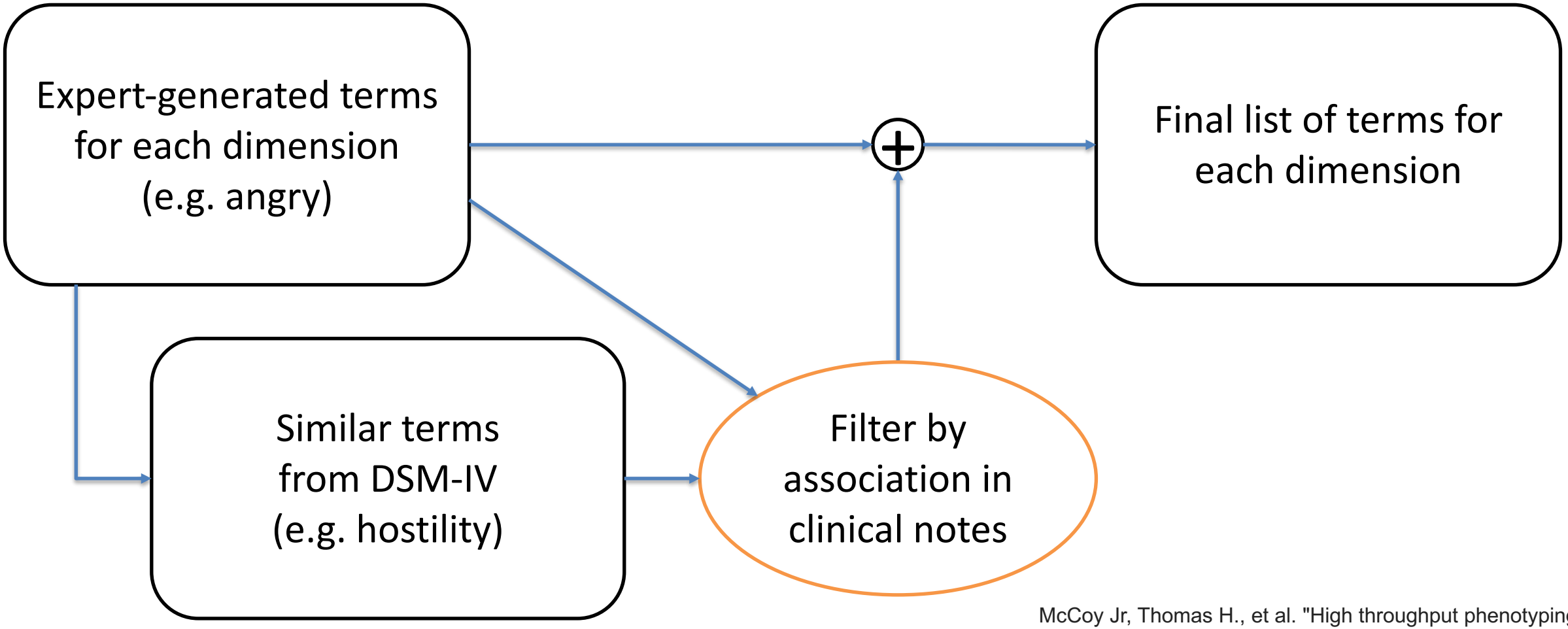
- hypoxemia (yes/no) in the next 5 minutes
- based on data from the Anesthesia Information Management System
- static features + real-time features collected up to that time point



A majority of features are keyword counts

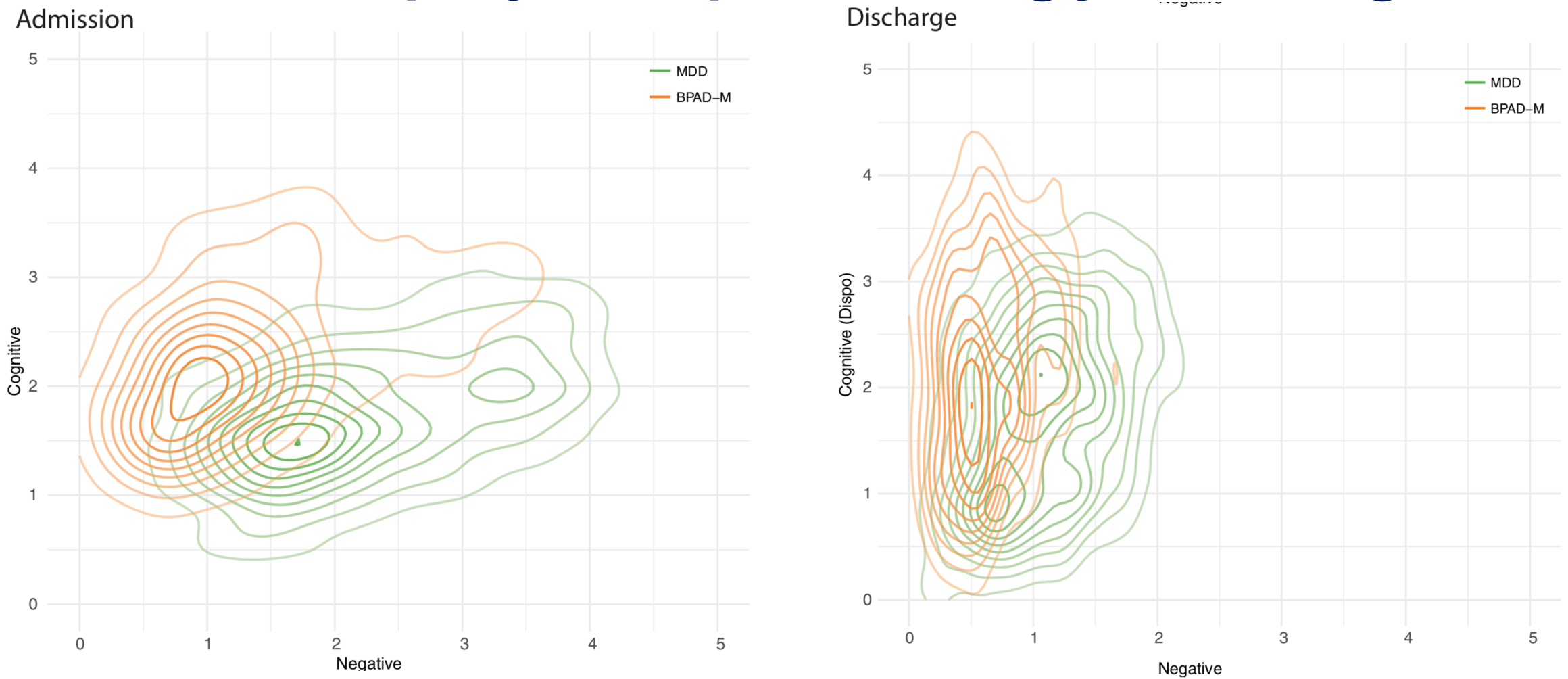


Scoring Dimensional Psychopathology



McCoy Jr, Thomas H., et al. "High throughput phenotyping for dimensional psychopathology in electronic health records." *Biological psychiatry* 83.12 (2018): 997-1004.

Trends in psychopathology during stay



McCoy Jr, Thomas H., et al. "High throughput phenotyping for dimensional psychopathology in electronic health records." *Biological psychiatry* 83.12 (2018): 997-1004.

Figure 1. Domain comparison contour plots showing change between admission (top) and discharge (bottom). BPAD-M, bipolar disorder–mania; MDD, major depressive disorder.

Conclusions

- NLP is approaching human performance on benchmark tasks like question answering
- Text data are central to clinical medicine, so the potential for NLP impact is high
- We can use word counts to turn text samples into vectors that we already know how to work with
- The techniques we have discussed already go beyond the majority of “NLP” found in the medical literature