Intro to Natural Language Processing for Clinical Text

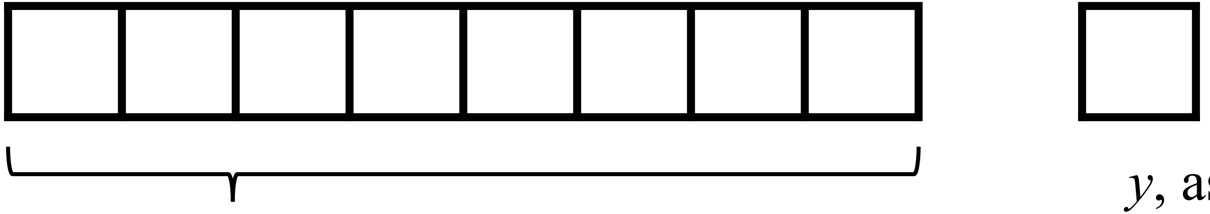
Matthew Engelhard

Today: NLP and Model Interpretability

- What can natural language processing (NLP) do in clinical medicine, and what is the role for *predictive* versus *generative* approaches?

- How does current NLP (i.e., large language models) work?
  - Foundations: count-based models
  - Foundations: word vectors
  - Modern LLM architectures (encoder, decoder, encoder-decoder)

Predictive models for tabular data



y, associated value or label

x, data/features for a subject or patient

End goal: predict y from x

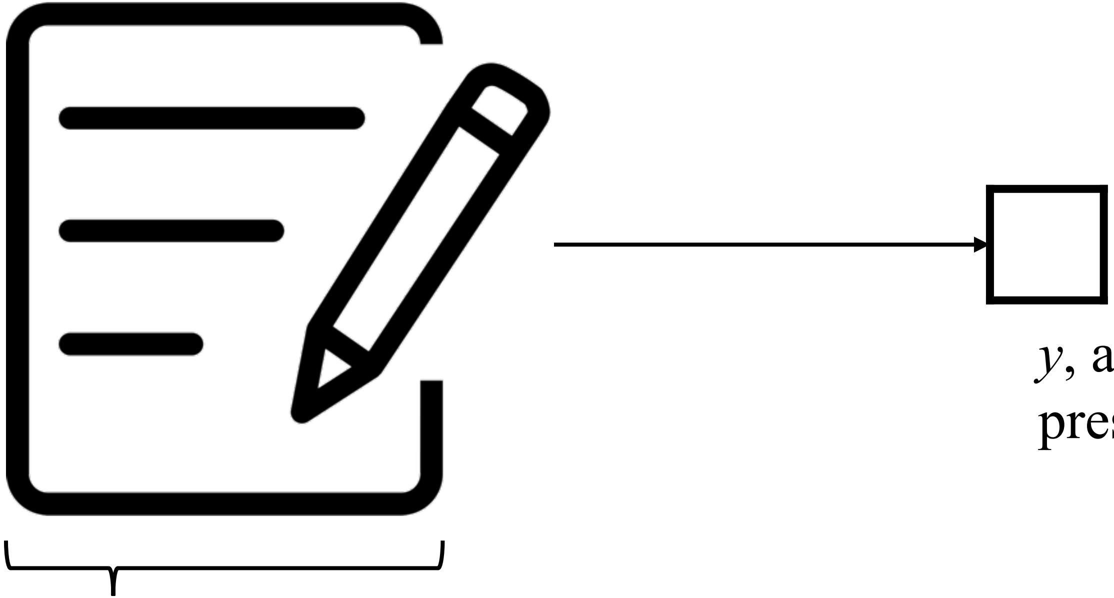
CNNs: predictive models for image data

x, retinal image

y, referrable diabetic retinopathy

End goal: predict y from x

NLP: predictive models for text data

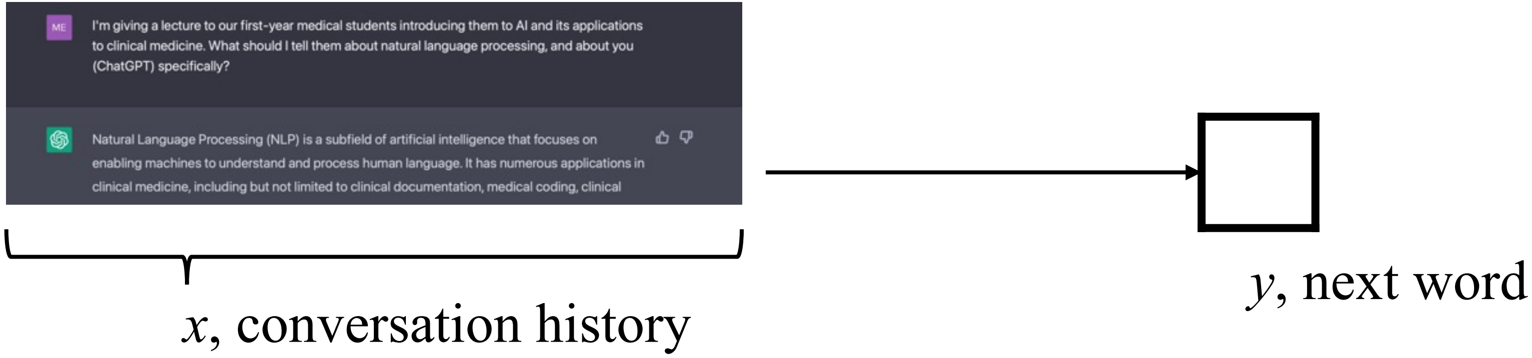


y, abnormality present

x, radiology report

End goal: predict y from x

Generative or predictive?



End goal: predict y from x

What can today’s NLP do?

And what is the emerging role of generative versus predictive approaches?

Clinical notes and other text contain key info not found elsewhere.

Journal of the American Medical Informatics Association, 26(11), 2019, 1189–1194
doi: 10.1093/jamia/ocz119
Advance Access Publication Date: 12 August 2019
Research and Applications

Research and Applications

Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies

Tina Hernandez-Boussard,[1,2,3] Keri L Monda,[4,5] Blai Coll Crespo,[4] and Dan Riskin[1,3,6]

Table 1. Cohort identification of diseases and procedures stratified by EHR-S and EHR-U data[a]

Cohort	Occurrence		EHR-S			EHR-U		
	Concept	Patient	Recall (%)	Precision (%)	F1-score (%)	Recall (%)	Precision (%)	F1-score (%)
Hyperlipidemia	2471	837	65.2	99.3	78.7	98.2	99.4	98.8
Hypercholesterolemia	1899	478	55.1	98.0	70.5	90.4	98.8	94.4
Coronary artery disease	1427	465	67.5	99.4	80.4	94.6	96.2	95.4
Diabetes mellitus	4502	1377	80.6	97.9	88.4	97.0	92.6	94.8
Myocardial infarction	523	282	29.8	86.2	44.2	90.4	76.5	82.9
Chronic kidney disease	640	101	40.8	97.6	57.6	92.9	97.9	95.3
Stroke	693	307	36.5	97.2	53.0	95.7	79.6	87.0
Dementia	317	103	62.1	100.0	76.6	93.1	90.0	91.5
Cataract	240	85	28.6	100.0	44.4	96.1	94.9	95.5
CABG[b]	194	73	32.2	100.0	48.7	96.6	95.0	95.8

[a]All comparisons were significant at $P < .0001$.
[b]Coronary artery bypass graft.

Predictive models remain highly relevant.

psychologist presenting problem NAME is a 3 year, 4 month old female who was referred for a neurodevelopmental assessment due to concerns regarding her overall development, behavior, and social emotional functioning and to

x, clinical note

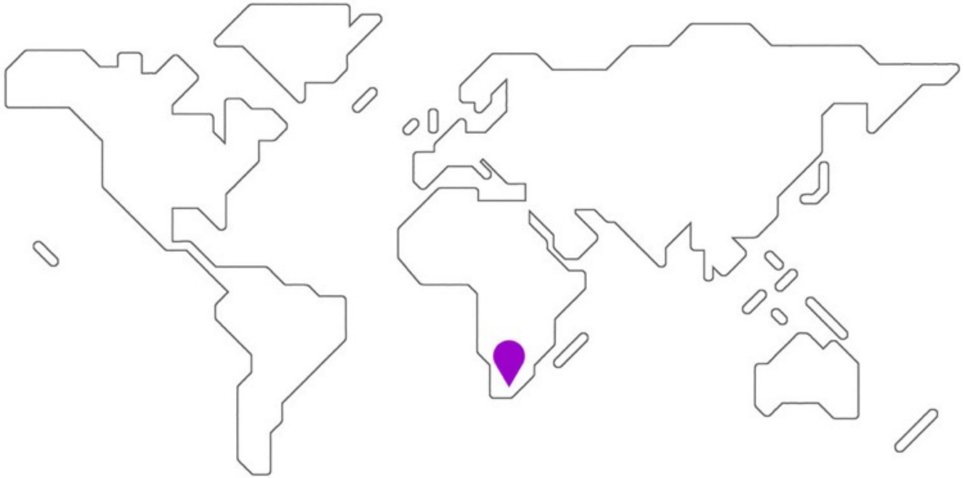
y, autism risk

End goal: predict y from x

Case Study: SMS Triage for 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)

Often predictive and generative models are complementary.

<u>Maternal health response system</u>:

- Speech to text (predictive)
- Translation (predictive)
- Identification of key concepts and topics (predictive)
- Triage (predictive)
- Generation of template and/or complete responses in specific cases (generative)

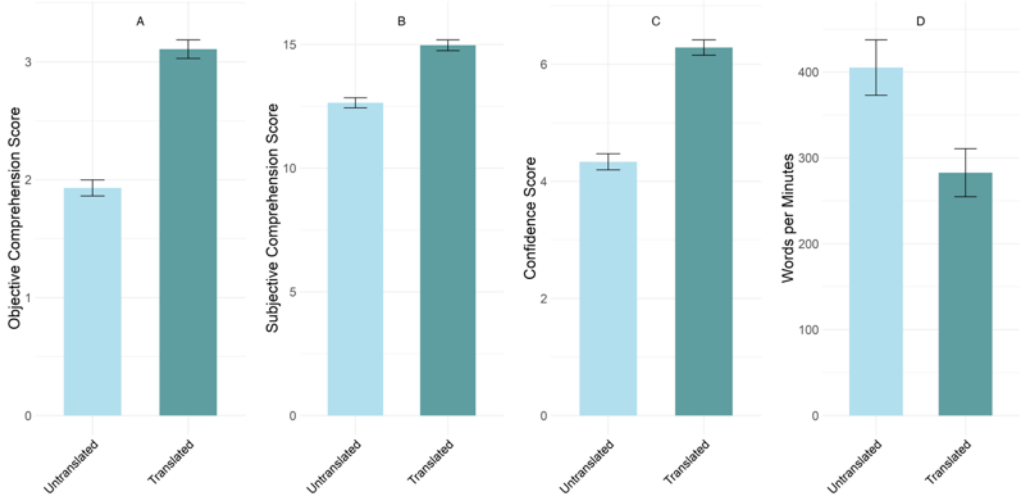
It is much easier to evaluate the performance of predictive models, and in turn to know when and how much to trust them.

In some cases, however, generative models have superseded predictive models.

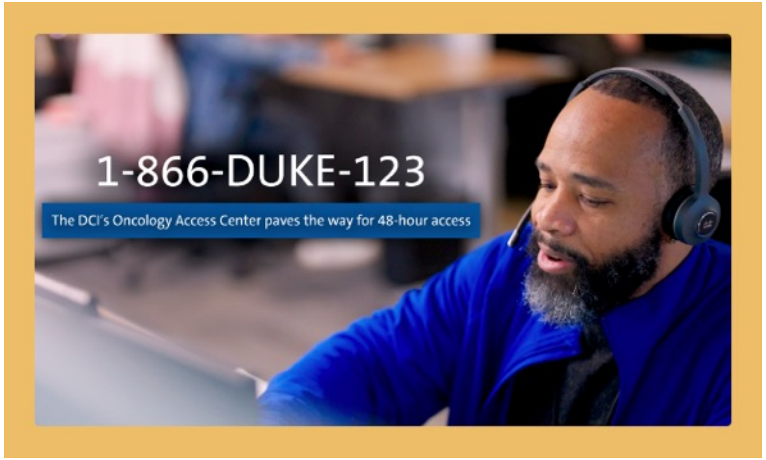
"Translation" of discharge notes into plain language using GPT4 substantially improves patients' ability to comprehend them.

Anivarya Kumar

Isabella Wang



Extracting specific info from documents: predictive or generative?



Data needed to determine scheduling needs
ICD-10
Stage
Age
Date of Dx
Hormone Receptors and molecular biomarkers
History of Treatment (med, surg, rad)
Undiagnosed patients only: Review of symptoms

While a predictive approach might be more trustworthy in principle, generative models can perform many tasks “zero-shot”, i.e. without task-specific training.

DIHI ‘24, led by Emily Norboge

Are all tasks special cases of text generation?

Model: GPT-4

ME

Please read the following echocardiogram report and tell me whether left ventricular hypertrophy was detected. Answer yes or no only.

Echo reading:
MILD LV SYSTOLIC DYSFUNCTION (See above)   WITH MILD LVH NORMAL RIGHT VENTRICULAR SYSTOLIC FUNCTION MODERATE VALVULAR REGURGITATION (See above) NO VALVULAR STENOSIS NO PERICARDIAL EFFUSION SEVERE BIATRIAL ENLARGEMENT AT LEAST MILD PULMONARY HTN ELEVATED RA PRESSURES NO PRIOR STUDY FOR COMPARISON

Yes

Are all tasks special cases of text generation?

ME

Please read the following echocardiogram report again and tell me whether left ventricular hypertrophy was detected. This time, please give me your confidence level as a percentage ranging from 0% (i.e. you are certain that it was NOT detected) to 100% (i.e. you are certain that it WAS detected). Answer with the percentage only.

Echo reading:
MILD LV SYSTOLIC DYSFUNCTION (See above) WITH MILD LVH NORMAL RIGHT VENTRICULAR SYSTOLIC FUNCTION MODERATE VALVULAR REGURGITATION (See above) NO VALVULAR STENOSIS NO PERICARDIAL EFFUSION SEVERE BIATRIAL ENLARGEMENT AT LEAST MILD PULMONARY HTN ELEVATED RA PRESSURES NO PRIOR STUDY FOR COMPARISON

100%

Are all tasks special cases of text generation?

ME

OK. Now, please read the following text message and tell me whether you believe the sender was angry when sending it. Please give me your confidence level as a percentage ranging from 0% (i.e. you are certain that they were NOT angry) to 100% (i.e. you are certain that they WERE angry). Answer with the percentage only.

> Hey, you didn't show up today. What's the deal?

60%

Current directions: RAG vs long context

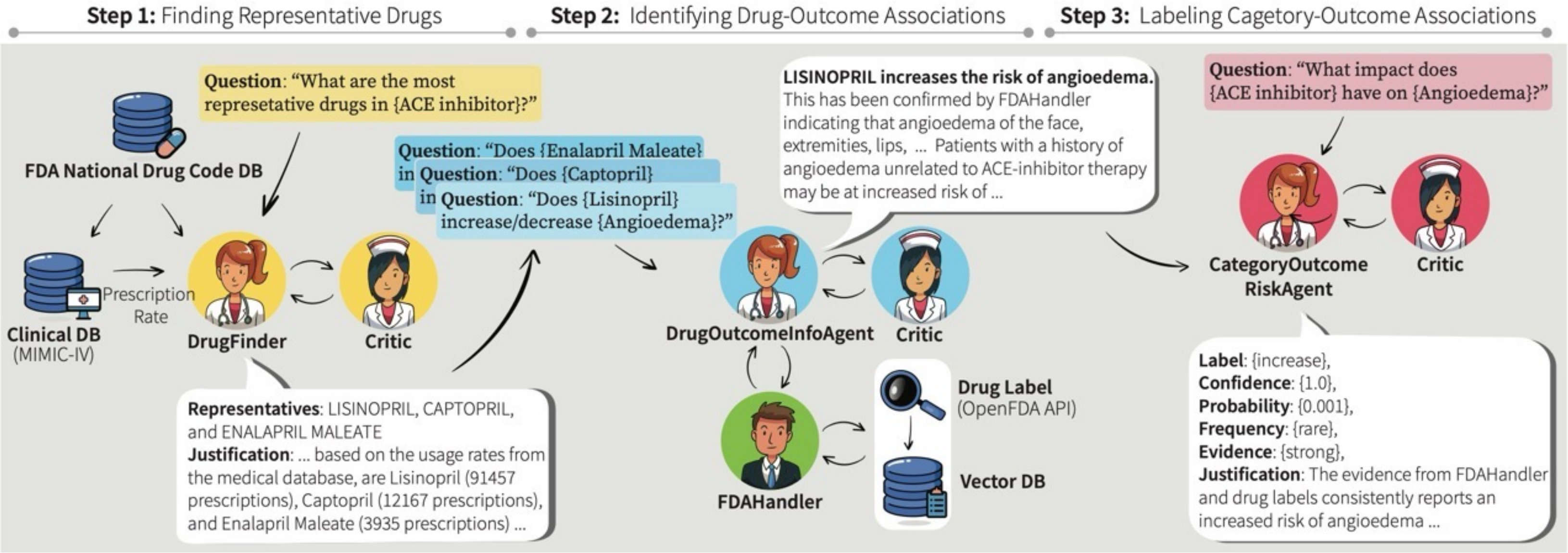
Suppose we wish to answer questions or extract information from a large collection of documents (e.g., all clinical notes for a given patient)

How do we use an LLM to do this?

- Option 1: *very* long context window
- Option 2: retrieval augmented generation (RAG)

Currently, RAG is commonly used because it is effective and scalable, and it (partly) addresses LLM pitfalls, including hallucinations.

Current directions: agentic AI



<u>Multi-LLM-agent systems including</u>:

- An orchestrator agent
- Task-specific agents (e.g. DocChat, SQLchat
- Critic agents

Choi, Palumbo, et al.; submitted to *Machine Learning for Healthcare*

Current directions: evaluation of generative models

<u>Open questions include</u>:

- Can we trust models to evaluate other models?

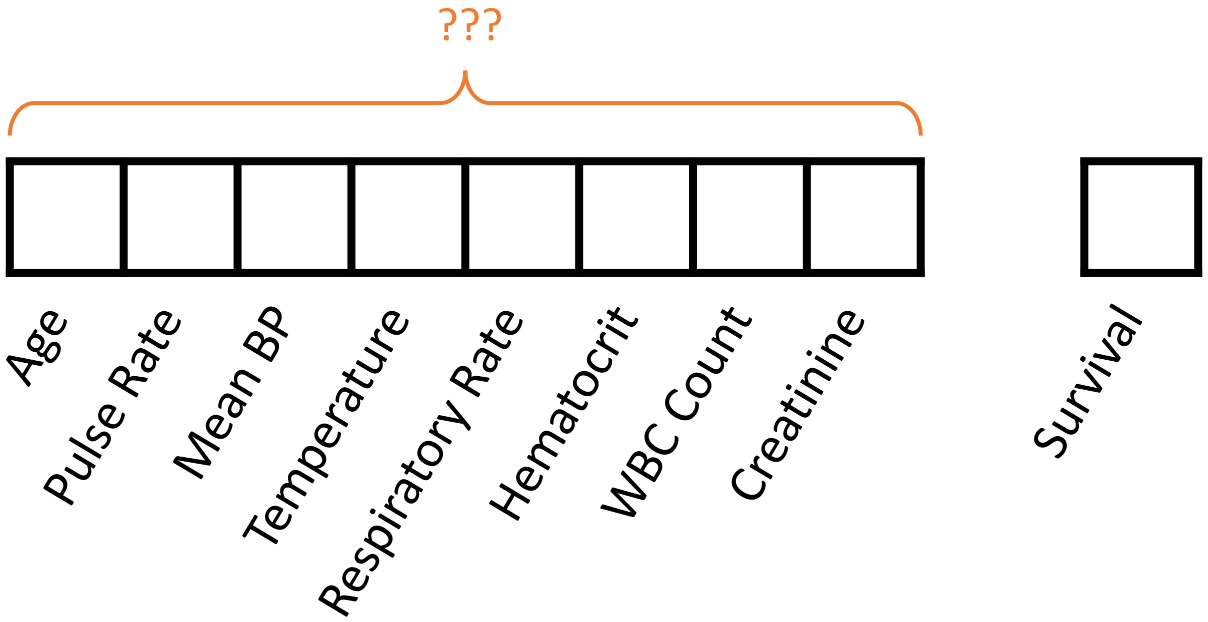
- How can models *learn when to defer* to human experts?

- How can we identify LLM biases? (more on this in PIONEER sessions)

How does NLP work?

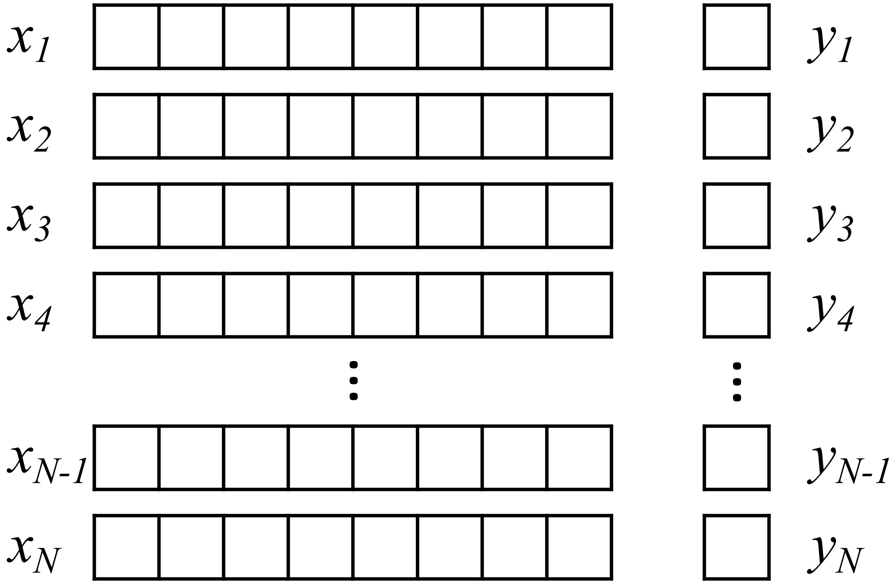
Key problem: how do we make predictions from text?

A Simple Predictive Model: ICU Mortality



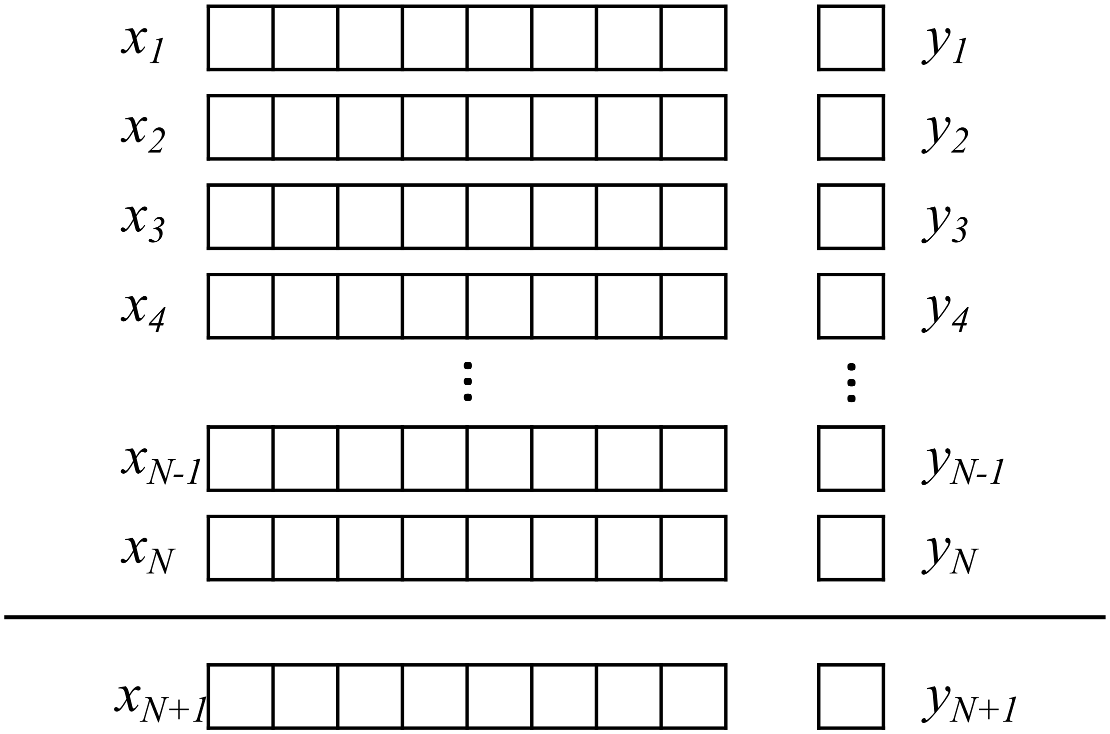
End goal: predict odds of hospital mortality

Training Set (Historical Data)



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

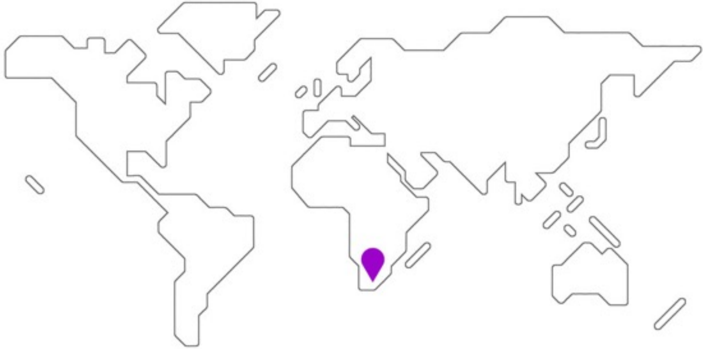
Making Predictions for New x



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

<- Learn to predict new y

Case Study: SMS Triage for Global Maternal Health



https://www.praekelt.org

Can we use a standard predictive model setup to solve this problem?

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

y_i: Urgent or Not Urgent?

<- 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 use logistic regression (or any other predictive model)!

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?

list of all words
(in no particular order)

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?

x_1

shaking	what	clinic	how	helps	was	nearest	many	with	said	months	the	morning	mom	should	sickness	and	I	is	how	out	breastfeed	passed	where
0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0

Step 2: <u>count how many times each vocabulary word appears in a given SMS</u>

I passed out and Mom said I was shaking

	shaking	what	clinic	how	helps	was	nearest	many	with	said	months	the	morning	mom	should	sickness	and	I	is	how	out	breastfeed	passed	where
x_3	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	2	0	0	1	0	1	0

Note that word order does not matter!

clinic is where nearest the

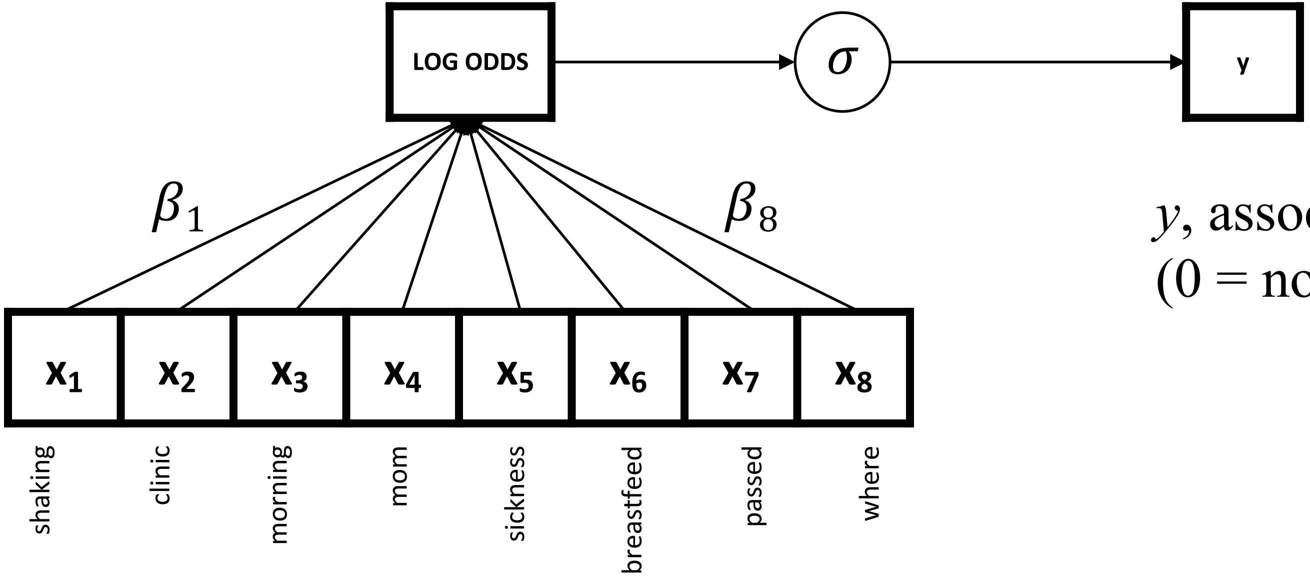
x_4

shaking	what	clinic	how	helps	was	nearest	many	with	said	months	the	morning	mom	should	sickness	and	I	is	how	out	breastfeed	passed	where
0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1

A “bag of words”

Now we can use logistic regression.

$$\text{URGENCY LOG ODDS} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$



y, associated label:
(0 = not urgent, 1 = urgent)

Count-based feature extraction is still useful!

- Entirely data-driven
  - Vocabulary of words we care about is derived from the training data
  - We then represent text as counts of each vocabulary word
  - We can also count 2- and 3-word phrases; this helps with negation and context

- Knowledge-driven extraction of key words or concepts
  - Rather than creating a vocabulary from the data, we can identify words we (or content experts) believe are important for a given task
  - Concept extraction systems (e.g. cTakes) will identify many alternative phrasings for the same clinical concept (e.g. diagnosis) and group them together as a single feature

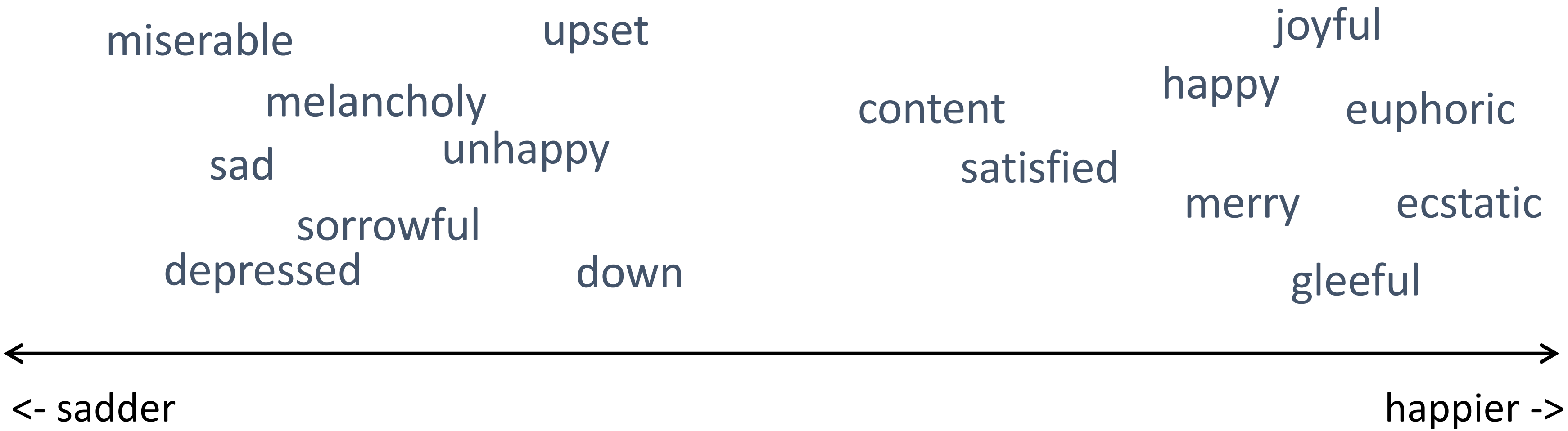
Strengths and Weaknesses

- (+) Count-based approaches are simple and work 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

How does *current* NLP work?

Key problem: understanding nuances of meaning and context

Word vectors: a numeric representation of words that encodes their meaning



Numeric value indicating whether the word is happy or sad

Training a robot to buy groceries

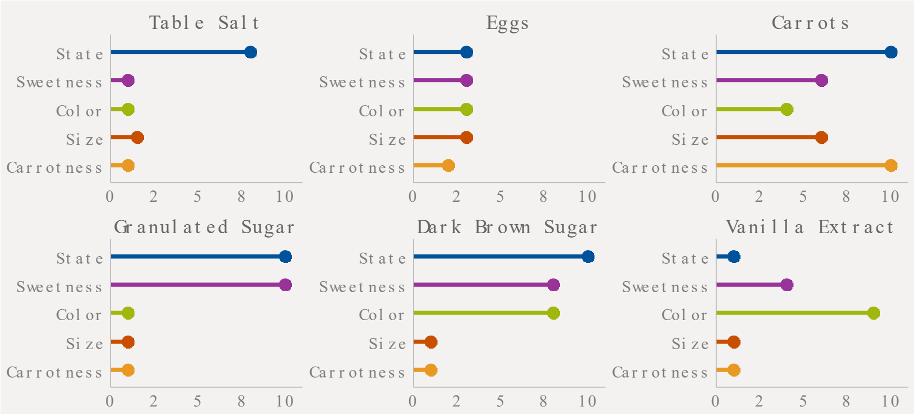
Example from Anand Chowdhury, MMCi 2019

Grocery List

- ❑ granulated sugar
- ❑ vanilla extract
- ❑ dark brown sugar
- ❑ carrots
- ❑ table salt
- ❑ eggs

Identify items by their attributes
(including previously unseen items)

Dimension	1	10
State	Liquid	Solid
Sweetness	Bland	Sweet
Color	Light	Dark
Size	Small	Large
Carrotness	Not really	Platonic essence of carrot



Why does this help us?

- The model can make sense of words it hasn't seen before (weren't used in training)
- Similar words (e.g. synonyms) will have similar attributes, and therefore will have similar effect on model predictions
- (more complicated) Now we can convert text to a sequence of vectors; and we were already very good at making predictions from sequences of vectors

How do we learn these attributes?

-> In brief, for now, but there's an additional, optional lecture on this

KEY IDEA: words are *defined* by the context in which they appear

A man strolls down the street

A woman strolls down the street

A child strolls down the street

A crocodile strolls down the street

A banana strolls down the street

A concept strolls down the street

How do we learn these attributes?

KEY IDEA: words are *defined* by the <u>context</u> in which they appear

-> if words are always exchangeable, they must have very similar meaning

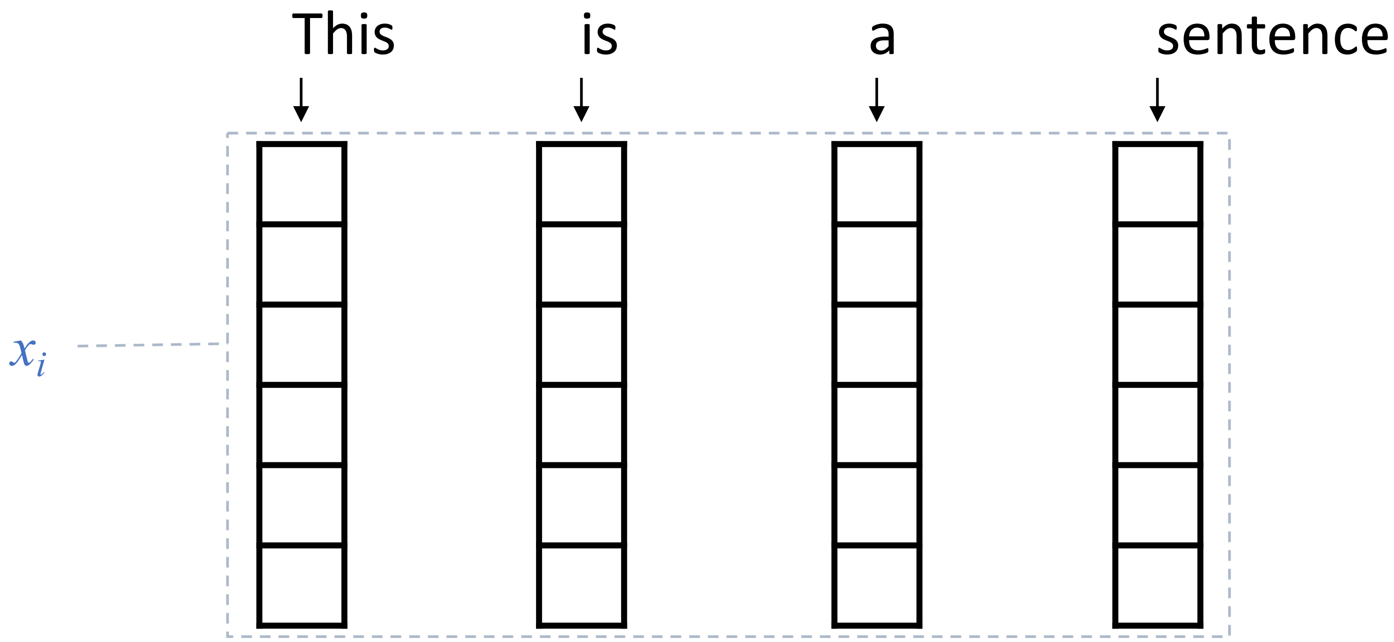
learn word meaning like an adult:
explicit definitions

https://www.parenting.com/activities/baby/teach-baby-to-talk/

learn word meaning like an child:
<u>implicit definitions from context</u>

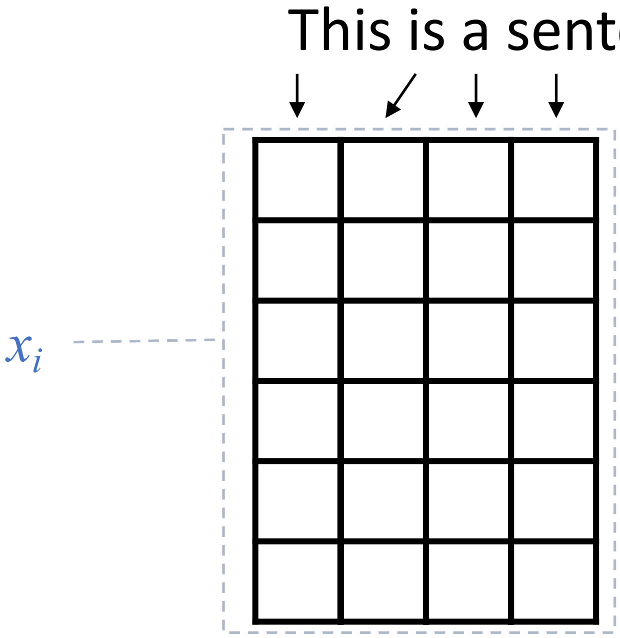
What happens when we embed all words in a sentence?

- Look up words individually to obtain their vectors
- Construct a sequence of vectors



What happens when we embed all words in a sentence?

- Look up words individually to obtain their vectors
- Construct a sequence of vectors



Now we have a grid of numbers

Similar in many ways to an image

Now we can use deep learning…

…to learn to extract increasingly complex aspects of meaning

Now we can use deep learning to build our hierarchy of features.

pixels

low-level motifs → high-level motifs → dog

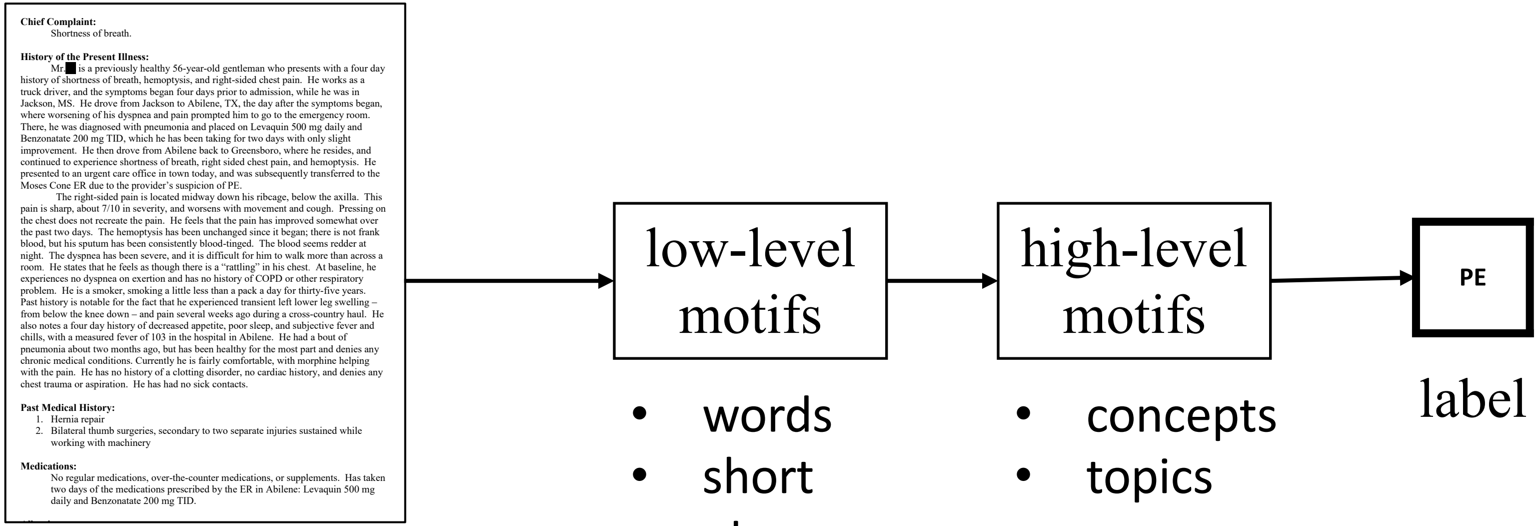
label

- edges
- shapes
- textures

- eyes
- ears
- paws

End goal: predict *dog* from *pixels*

Now we can use deep learning to build our hierarchy of *semantic* features.


Chief Complaint:
Shortness of breath.

History of the Present Illness:
Mr.[redacted] is a previously healthy 56-year-old gentleman who presents with a four day history of shortness of breath, hemoptysis, and right-sided chest pain. He works as a truck driver, and the symptoms began four days prior to admission, while he was in Jackson, MS. He drove from Jackson to Abilene, TX, the day after the symptoms began, where worsening of his dyspnea and pain prompted him to go to the emergency room. There, he was diagnosed with pneumonia and placed on Levaquin 500 mg daily and Benzonatate 200 mg TID, which he has been taking for two days with only slight improvement. He then drove from Abilene back to Greensboro, where he resides, and continued to experience shortness of breath, right sided chest pain, and hemoptysis. He presented to an urgent care office in town today, and was subsequently transferred to the Moses Cone ER due to the provider's suspicion of PE.

The right-sided pain is located midway down his ribcage, below the axilla. This pain is sharp, about 7/10 in severity, and worsens with movement and cough. Pressing on the chest does not recreate the pain. He feels that the pain has improved somewhat over the past two days. The hemoptysis has been unchanged since it began; there is not frank blood, but his sputum has been consistently blood-tinged. The blood seems redder at night. The dyspnea has been severe, and it is difficult for him to walk more than across a room. He states that he feels as though there is a "rattling" in his chest. At baseline, he experiences no dyspnea on exertion and has no history of COPD or other respiratory problem. He is a smoker, smoking a little less than a pack a day for thirty-five years. Past history is notable for the fact that he experienced transient left lower leg swelling – from below the knee down – and pain several weeks ago during a cross-country haul. He also notes a four day history of decreased appetite, poor sleep, and subjective fever and chills, with a measured fever of 103 in the hospital in Abilene. He had a bout of pneumonia about two months ago, but has been healthy for the most part and denies any chronic medical conditions. Currently he is fairly comfortable, with morphine helping with the pain. He has no history of a clotting disorder, no cardiac history, and denies any chest trauma or aspiration. He has had no sick contacts.

Past Medical History:
1. Hernia repair
2. Bilateral thumb surgeries, secondary to two separate injuries sustained while working with machinery

Medications:
No regular medications, over-the-counter medications, or supplements. Has taken two days of the medications prescribed by the ER in Abilene: Levaquin 500 mg daily and Benzonatate 200 mg TID.

grid of semantic attributes

- words
- short phrases

- concepts
- topics

label

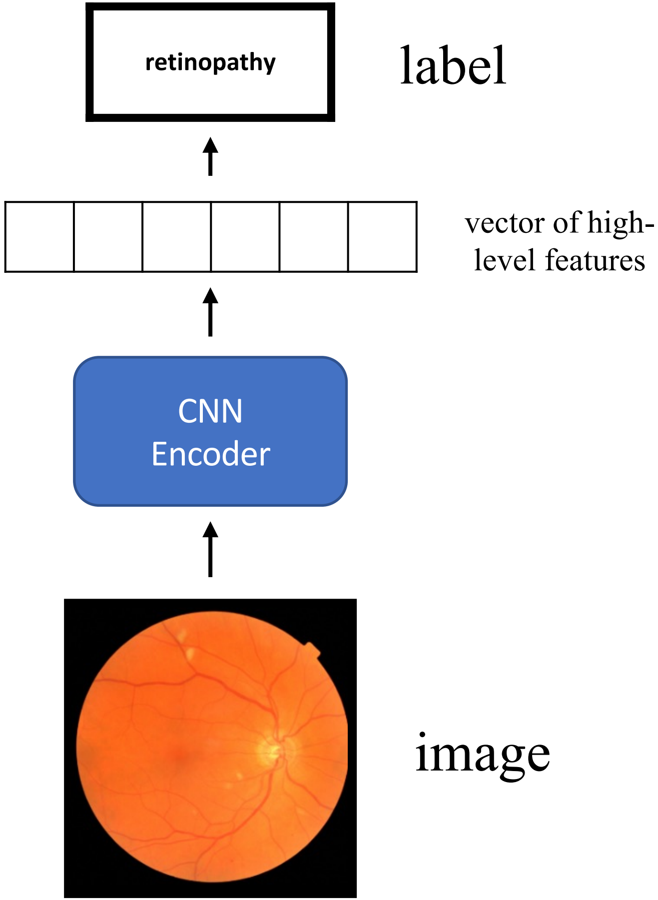
End goal: predict *pulmonary embolism* from *text*

Recall: in image processing, we start with a pre-trained *encoder*

1. A CNN *image encoder* that converts the raw image to a vector of high-level motifs / features.

2. A *final layer*, or *prediction head* – this is a logistic regression model – that makes predictions about the label from these high-level features.

- We will reuse the encoder but replace the prediction head, since it is specific to the previous (non-medical) task.

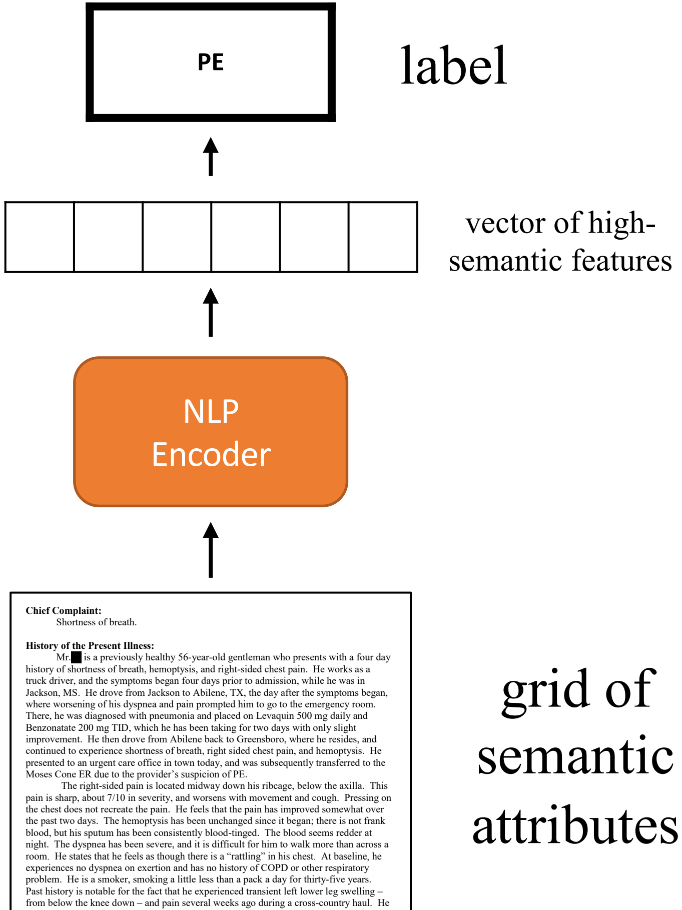


In modern (deep) NLP, we also start with a pre-trained *encoder*

1. A transformer network *image encoder* that converts the raw semantic attributes to a vector of high-level motifs / features.

2. A *final layer*, or *prediction head* – this is a logistic regression model – that makes predictions about the label from these high-level features.

- We will reuse the encoder but replace the prediction head, since it is specific to the previous task.



Chief Complaint:
Shortness of breath.

History of the Present Illness:
Mr.[redacted] is a previously healthy 56-year-old gentleman who presents with a four day history of shortness of breath, hemoptysis, and right-sided chest pain. He works as a truck driver, and the symptoms began four days prior to admission, while he was in Jackson, MS. He drove from Jackson to Abilene, TX, the day after the symptoms began, where worsening of his dyspnea and pain prompted him to go to the emergency room. There, he was diagnosed with pneumonia and placed on Levaquin 500 mg daily and Benzonatate 200 mg TID, which he has been taking for two days with only slight improvement. He then drove from Abilene back to Greensboro, where he resides, and continued to experience shortness of breath, right sided chest pain, and hemoptysis. He presented to an urgent care office in town today, and was subsequently transferred to the Moses Cone ER due to the provider's suspicion of PE.

The right-sided pain is located midway down his ribcage, below the axilla. This pain is sharp, about 7/10 in severity, and worsens with movement and cough. Pressing on the chest does not recreate the pain. He feels that the pain has improved somewhat over the past two days. The hemoptysis has been unchanged since it began; there is not frank blood, but his sputum has been consistently blood-tinged. The blood seems redder at night. The dyspnea has been severe, and it is difficult for him to walk more than across a room. He states that he feels as though there is a "rattling" in his chest. At baseline, he experiences no dyspnea on exertion and has no history of COPD or other respiratory problem. He is a smoker, smoking a little less than a pack a day for thirty-five years. Past history is notable for the fact that he experienced transient left lower leg swelling – from below the knee down – and pain several weeks ago during a cross-country haul. He

Pre-training on biomedical corpora is becoming less important with current LLMs.

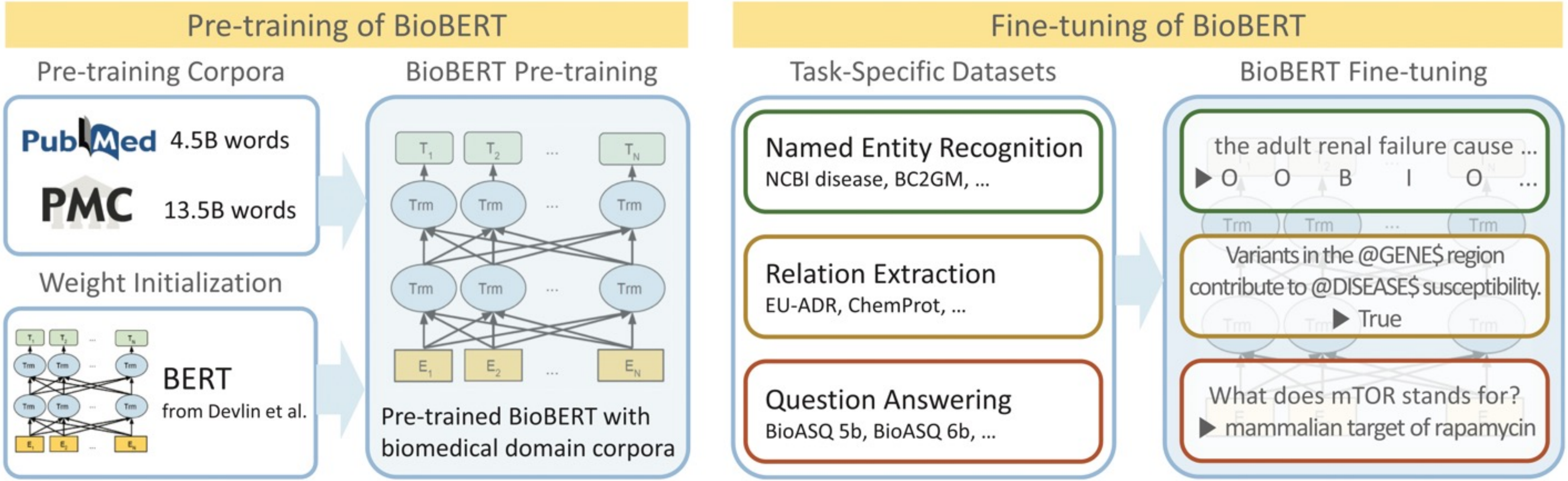


Fig. 1. Overview of the pre-training and fine-tuning of BioBERT

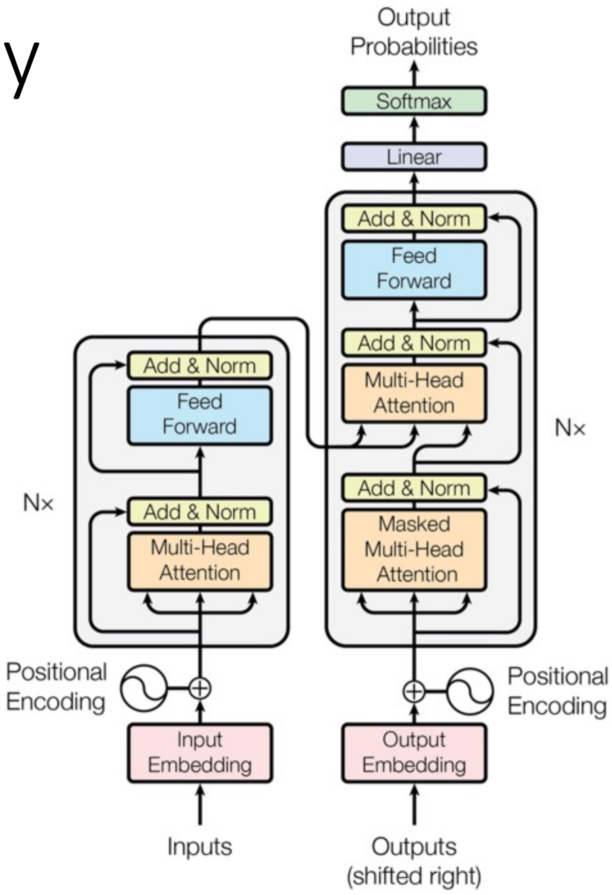
Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234-40.

Pre-training on biomedical corpora is becoming less important with current LLMs.

- Common LLMs (e.g. BERT, GPT4) have millions or billions of parameters (up to 1T)
- However, the principles remain the same: neural networks performing hierarchical feature extraction
- Different tasks require slightly different final modifications to the architecture
- Deep NLP is becoming more accessible (and common in the clinical literature) as tools to acquire and use these models continue to improve

A Brief Tour of LLM Terminology

- Encoder, Decoder
- Autoregressive
- Multi-head attention
- Masked or next token prediction



Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

How to build a large language model (LLM)

Step 0: invent word embeddings, transformer architecture, and other building blocks

Step 1: train to predict the next/missing word or sentence across a huge collection of documents

- Generalist models: Wikipedia, common crawl, twitter (i.e., the internet)
- Biomedical models: PubMed, MIMIC notes

Step 2: refine and align the models by having humans rate their outputs (i.e., reinforcement learning from human feedback)

- Many variations on this, some of them closely kept
- Possible role of critic models and *learning to defer*

Conclusions**

- Text data are central to clinical medicine, so the potential for NLP impact is high (but *not yet realized*)

- Simple, count-based NLP models are surprisingly effective in most clinical applications.

- Complex, deep learning NLP models have exceeded human performance. In these models, words are converted to vectors of semantic attributes, and increasingly complex, heirarchical semantic features are then extracted.

- Similar to image processing, we can take advantage of complex NLP models by repurposing them for a specific clinical task via fine-tuning of parameters.