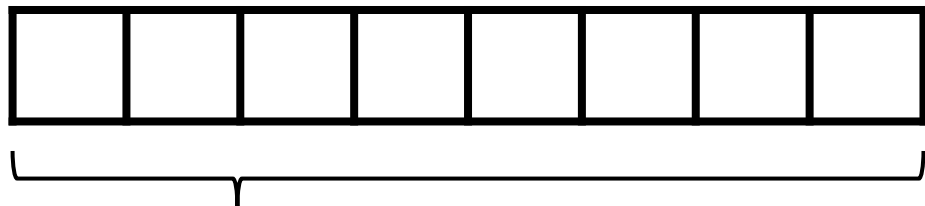


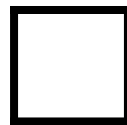
# Intro to Natural Language Processing for Clinical Text

Matthew Engelhard

# Predictive models for tabular data



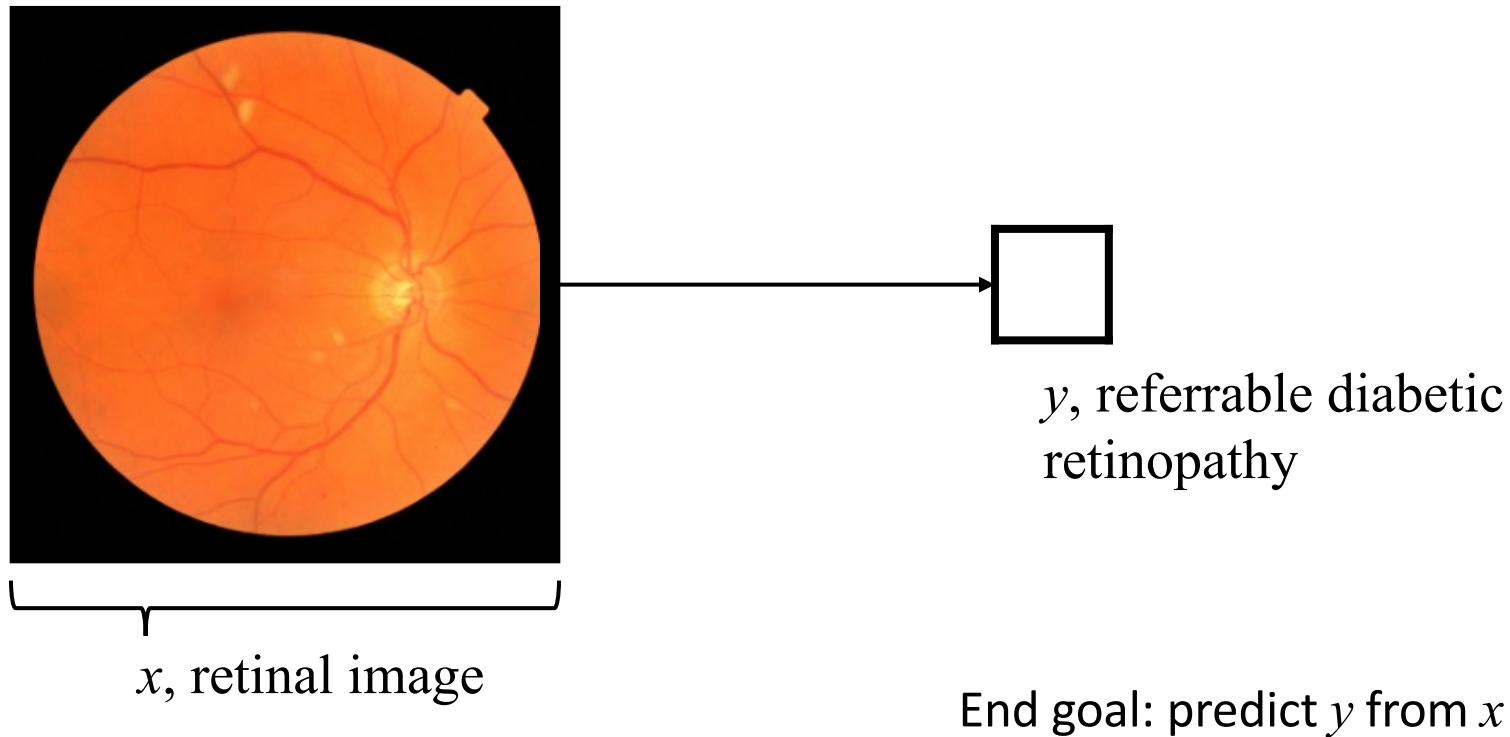
$x$ , data/features for  
a subject or patient



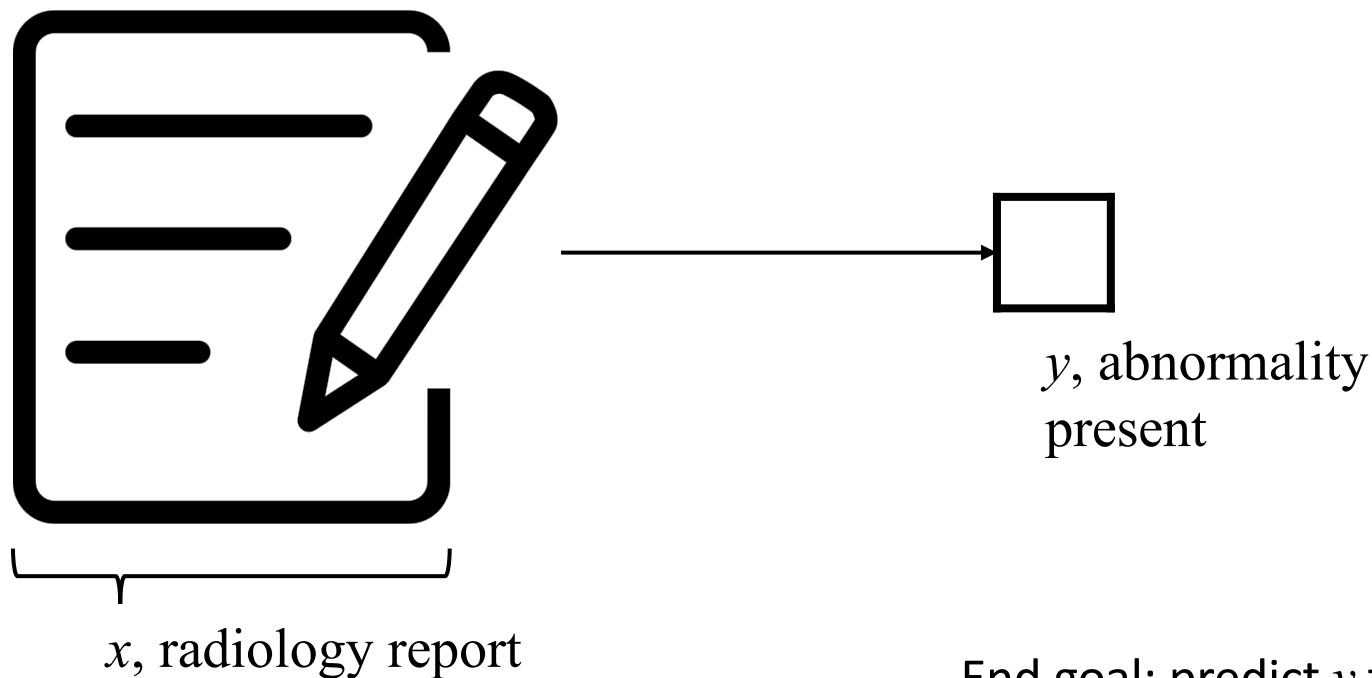
$y$ , associated  
value or label

End goal: predict  $y$  from  $x$

# CNNs: predictive models for image data



# NLP: predictive models for text data



End goal: predict  $y$  from  $x$

# Today: NLP and Model Interpretability

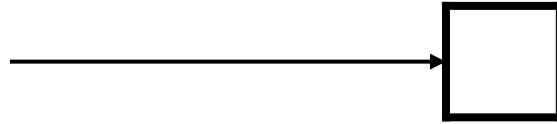
- What can natural language processing (NLP) do, and how is it relevant to clinical medicine?
- How does NLP work?
  - Methods based on word *counts* (“bag of words”)
  - Methods based on word *vectors* (including deep neural networks)

What can today's NLP do?

# Classification and Regression

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



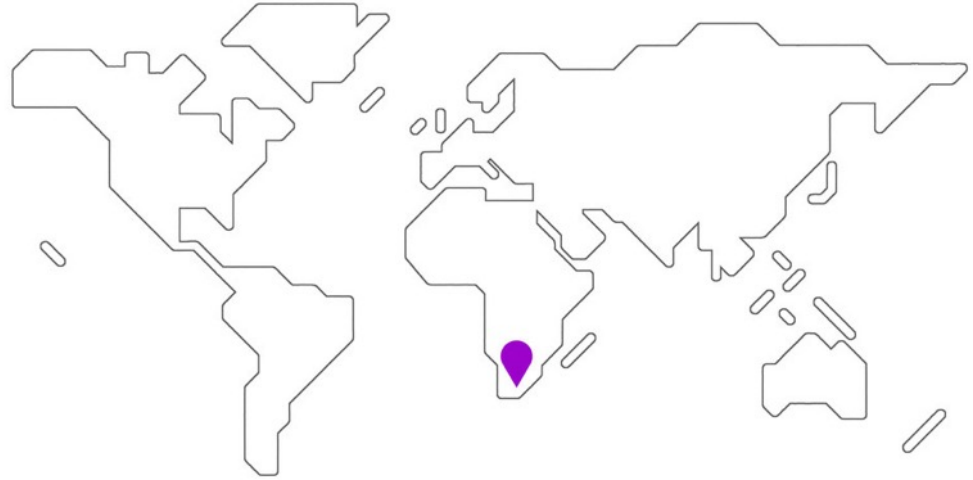
$y$ , autism risk

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)**



# Why use clinical notes?

Notes >> structured data for  
identifying diseases, procedures

## Research and Applications

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

Tina Hernandez-Boussard,<sup>1,2,3</sup> Keri L Monda,<sup>4,5</sup> Blai Coll Crespo,<sup>4</sup> and Dan Riskin<sup>1,3,6</sup>

**Table 1.** Cohort identification of diseases and procedures stratified by EHR-S and EHR-U data<sup>a</sup>

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 <sup>b</sup>	194	73	32.2	100.0	48.7	96.6	95.0	95.8

<sup>a</sup>All comparisons were significant at  $P < .0001$ .

<sup>b</sup>Coronary artery bypass graft.

# Text Translation

ENGLISH - DETECTED

ENGLISH

GERM



ENGLISH

SWEDISH

GERMAN



Deep learning is so much fun| ×



28/5000



Deep Learning macht so viel  
Spaß



[Send feedback](#)

# Text Translation --> De-Identify Notes

**Table 5.** Examples of correctly detected PHI instances (in bold) by the ANN

PHI category	ANN
AGE	Father had a stroke at <u>80</u> and died of?another stroke at age Personal data and overall health: Now <u>63</u> , despite his FH: Father: Died @ <u>52</u> from EtOH abuse (unclear exact etiology) Tobacco: smoked from age 7 to <u>15</u> , has not smoked since 15.
CONTACT	History of Present Illness <u>86F</u> reports worsening b/l leg pain. by phone, Dr. Ivan Guy. Call w/ questions <u>86383</u> . Keith Gilbert, H/O paroxysmal afib VNA <u>171-311-7974</u> ===== Medications
DATE	During his <u>May</u> hospitalization he had dysphagia Social history: divorced, quit smoking in <u>08</u> , sober x 10 yrs, She is to see him on the <u>29th</u> of this month at 1:00 p.m. He did have a renal biopsy in teh late <u>60s</u> adn thus will look for results, Results <u>02/20/2087</u> NA 135, K 3.2 (L), CL 96 (L), CO2 30.6, BUN 1 Jose Church, M.D. /ray DD: 01/18/20 DT: <u>01/19/0</u> DV: 01/18/20

# 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 conserved among microorganisms, or when damaged, injury signals, many of which (but not all) are related to those that recognize pathogens. Innate immunity, meaning these systems respond to pathogens, but does not confer long-lasting immunity against them. Innate immunity 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

## 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 <i>Stanford University</i> (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) <i>Google AI Language</i> <a href="https://github.com/google-research/bert">https://github.com/google-research/bert</a>	86.673	89.147
2 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) <i>Google AI Language</i> <a href="https://github.com/google-research/bert">https://github.com/google-research/bert</a>	85.150	87.715

tors

### Dominant system of defense?

e system innate immune

m

## Identify components present in broad

microorganisms

s in a generic way, meaning it is

non-specific   non-specific

# Question Answering --> Query the EHR

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 conserved among broad groups of microorganisms, or when damaged, injured or stressed cells send out alarm signals, many of which (but not all) are recognized by the same receptors as those that recognize pathogens. Innate immune defenses are non-specific, meaning these systems respond to pathogens in a generic way. This system does not confer long-lasting immunity against a pathogen. The innate immune system is the dominant system of host defense in most organisms.

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

*Ground Truth Answers:* pattern recognition receptors receptors cells

*Prediction:* pattern recognition receptors

**For most organisms, what is the dominant system of defense?**

*Ground Truth Answers:* innate immune system innate immune system The innate immune

*Prediction:* The innate immune system

**Pattern recognition receptors recognize components present in broad groups of what?**

*Ground Truth Answers:*

microorganisms microorganisms microorganisms

*Prediction:* microorganisms

**The innate immune system responds in a generic way, meaning it is what?**

*Ground Truth Answers:* non-specific non-specific non-specific

*Prediction:* non-specific

# Automatic Image Captioning



a cow is standing in front of a store



a group of elephants standing next to each other



a table that has wooden spoons on it



a cat is eating some kind of food



a bunch of bananas are sitting on a table



a motorcycle is parked next to a window

# Automatic Image Captioning --> annotate imaging

Classification of radiology reports using  
neural attention models, *IJCNN 2017*



a cow is standing in front  
of a store



a group of elephants  
standing next to each other



a table that has wooden  
spoons on it



a cat is eating some kind of  
food



a bunch of bananas are  
sitting on a table



a motorcycle is parked next  
to a window



# Text Generation

## SYSTEM PROMPT (HUMAN-WRITTEN)

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

## **“Better Language Models and Their Implications”**

2/14/19

OPENAI

## 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.



# Text Generation -> Suggested Email Responses



Messaging



Health



Appts & Visits



Questionnaires

## Message Center

ASK A QUESTION

Inbox Sent Messages

Search message list



Sort by:

Received Date



Filters:


All Messages



# 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?



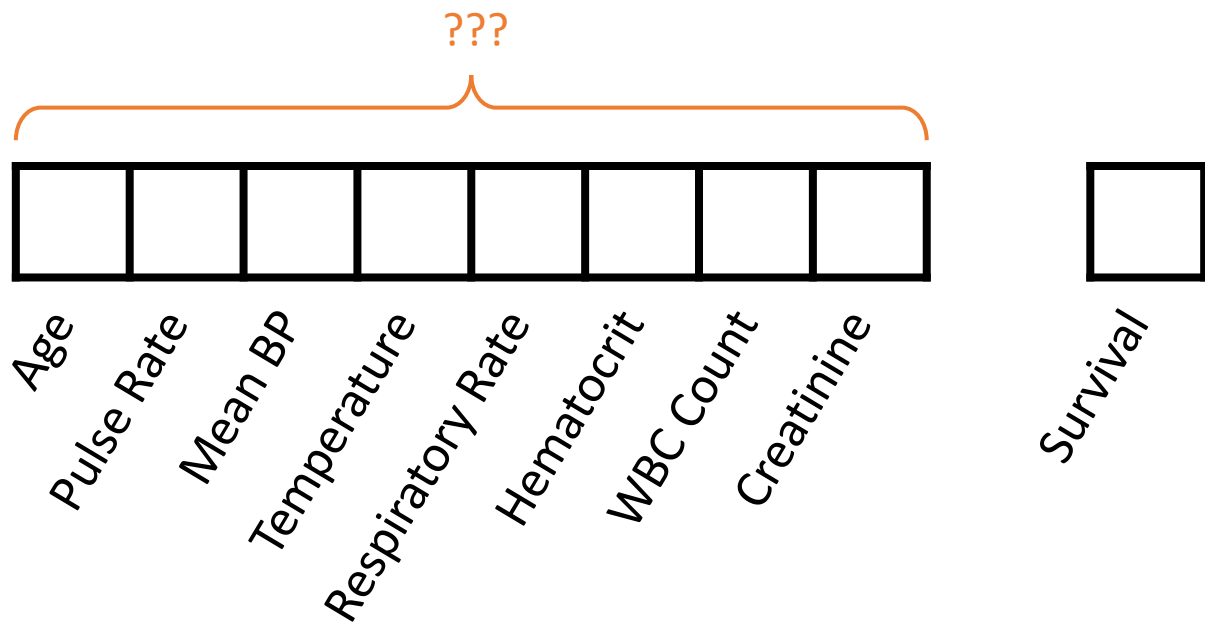
60%



# How do we make predictions from text?

Answer 1: word counts

# A Simple Predictive Model: ICU Mortality



End goal: predict odds of hospital mortality

# Training Set (Historical Data)

$x_1$	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	$y_1$
$x_2$	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	$y_2$
$x_3$	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	$y_3$
$x_4$	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	$y_4$
	$\vdots$									$\vdots$
$x_{N-1}$	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	$y_{N-1}$
$x_N$	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	$y_N$

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

# Making Predictions for New $x$

$x_1$	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	$y_1$
$x_2$	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	$y_2$
$x_3$	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	$y_3$
$x_4$	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	$y_4$
$x_{N-1}$	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	$y_{N-1}$
$x_N$	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	$y_N$

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

---

$x_{N+1}$	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	$y_{N+1}$
-----------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	-----------

<- 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?    ☐     $y_4$

$\vdots$

$\vdots$

$x_{N-1}$     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 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  
what  
clinic  
how  
helps  
was  
nearest  
many

with  
said  
months  
the  
morning  
mom  
should  
sickness

and  
I  
is  
how  
out  
breastfeed  
passed  
where

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

What helps with morning sickness?

$x_I$

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

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

I passed out and Mom said I was shaking

$x_3$

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

Note that word order does not matter!

clinic is where nearest the

$x_4$

0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	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

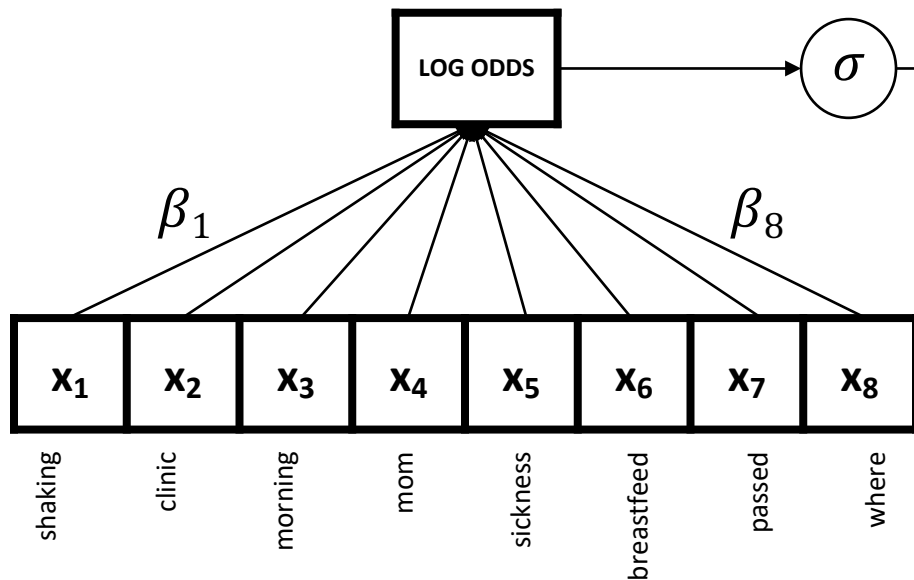
A “bag of words”





Now we can use logistic regression.

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



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

# Variations on count-based feature extraction

- 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

# Summary

- A central challenge of NLP lies in converting text documents into feature vectors that can be used in a predictive model
- Bag of words models solve this challenge by constructing a feature vector based on counts of each word of interest
- Even though they ignore word order and semantic relationships, these models are very powerful

# Other text processing details

N-grams, tf-idf, pre-processing steps

# Count 1- and 2-grams in each SMS (i.e. extend vocabulary to include 2-word phrases)

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	

$x_1$

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?

2-grams

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

# 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**



# 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."



"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

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

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

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

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

# Preprocessing

- remove punctuation

I passed out, and Mom said I was shaking.

- to lowercase

I passed out and Mom said I was shaking

- “tokenization”

i passed out and mom said i was shaking

- “stemming”

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

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