# Lecture 2: Electronic Medical Records

CSCI6410/EPAH6410/CSCI4148

Finlay Maguire (finlay.maguire@dal.ca)

#### **Learning Outcomes**

- Describe electronic medical/health record systems and the types of data they typically contain
- Distinguish structured, semi-structured, unstructured text data
- Describe approaches to searching text
- Outline key steps in preparing text for analysis
- Explain the general concept of learnt word embeddings
- Explain how embeddings can be tuned/customised
- Identify differences between named entity recognition, parts of speech tagging, and dependency parsing

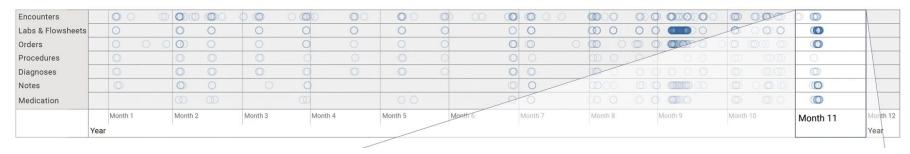
Not covered: fuzzy search and text indexing

## What is an Electronic {Medical,Health} Record?

#### EMR are digital patient charts

- Repository of patient information over time
- Prone to fragmentation between primary / hospital care
- Ideally contains all of a given patient's details on:
  - Every encounter with health professionals (e.g., admitted to hospital)
  - Details and results of diagnostic testing and vitals (e.g., blood test, urine cultures etc.)
  - Diagnostic/therapeutic orders (e.g., Nil per os/NPO,
  - Procedures performed (e.g., appendectomy, PET-scan)
  - Medical note (e.g., primary physician, consult information)
  - Medication (e.g., antibiotics)

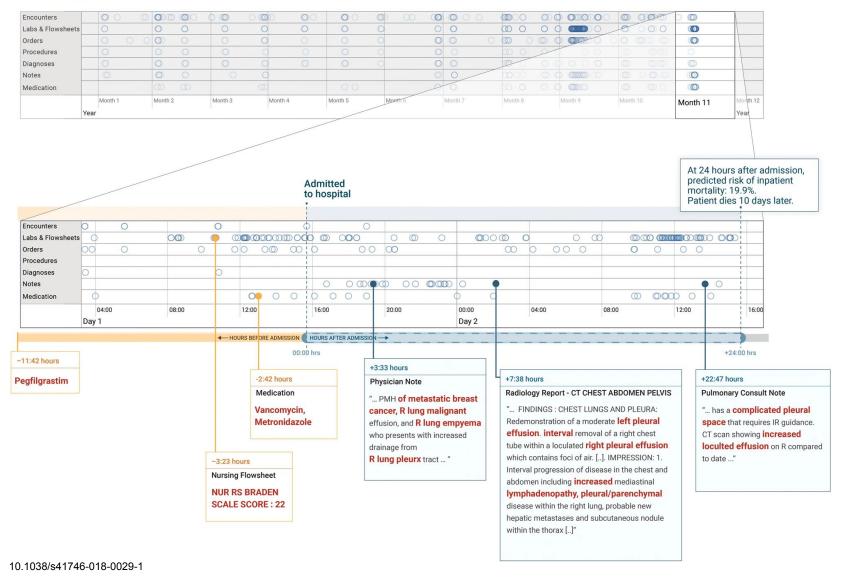
#### **Patient Timeline**



10.1038/s41746-018-0029-1

#### EMR data varies in structure

#### **Patient Timeline**



### EMRs have a surprisingly long history

Paper medical records come into steady use only in 1900-1920. Medical records are kept on paper and filed manually until the 1960s.

Lockheed develops the first EHR, also known as "Clinical Information System".

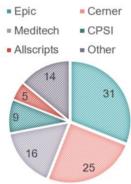


The federal government begins using EHR. The department of Veteran Affairs implements VistA, originally known as Decentralized **Hospital Computer Program** (DHCP).



An Institute of Medicine report argues the case for using EHR, and recommends the use of computerbased patient record systems in both the public and private sectors.

A 2004 random sample of healthcare facilities from across the U.S. finds that 13% of respondents have an EHR system fully implemented while 10% do not have or do not plan to have an EHR system.



Source: Becker's Health IT

1960

Academic medical centers start

developing their own systems

1970

1980 1990

In the late 1980's and early 1990's, web-based EHRs become popular, with the rise of compact and affordable personal computers, local area networks and the internet.

**University of Utah** develops one the first clinical decision support systems, Health **Evaluation** through Logical Processing (HELP).

Massachusetts General Hospital develops the Computer Stored Ambulatory Record (COSTAR)

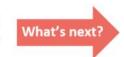


The Institute of Medicine publishes a study of medical errors, To Err is Human, concluding that healthcare would be safer with such systems as computerized physician order entry in place.

2000

2010 2020

> A 2016 survey finds that 96 percent of United States hospitals use certified EHR technology.



President Obama incorporates EHR into his American Recovery and Reinvestment Act of 2009 as part of the Health Information Technology for Economic and Clinical Health Act (HITECH).

Public and private healthcare providers and other eligible professionals (EP) are required to adopt and demonstrate "meaningful use" of electronic medical records (EMR) by January 1, 2014, in order to maintain their existing Medicaid and Medicare reimbursement levels.

-	deliver size I be recover to the	
	United States	-
	Edfre .	
F	THE PROPERTY AND PARTY AND	*
-	Texton Control of the last	
	TOMOS-	



https://mayaberlerner.medium.com/a-brief-history-of-ehrs-c51a2125a247

#### EMRs are common and increasing in use in Canada

- Use of EMRs is common and increasing in Canada (2017-2024: 82% to 95%)
- Primary care is main users (97%) compared to specialists (93%)
- Atlantic regions have lowest use rate (86%)
- Main features used:
  - Ordering diagonstic tests / accessing results (72%)
  - Prescription system with automatic warning of adverse drug interactions (60%)
  - Communication of discharge/consultation notes (40%)
- Not used:
  - Clinical decision support (<27%)
  - Appointment scheduling (<25%)

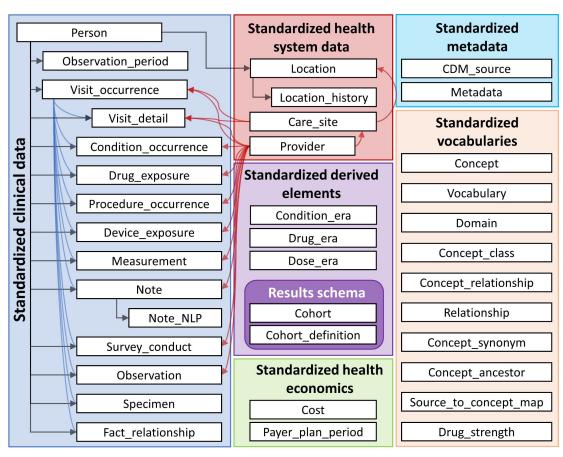
EMR System	National	Nova Scotia
Accuro - QHR	15.1%	32.0%
Med Access - Telus	13.1%	38.5%
PS Suite - Telus	12.8%	0%
Epic	8.0%	0%
Secure EHR - Citrix	5.6%	0%
Meditech	5.6%	20.3%
Oracle Cerner	1.5%	0%
Other	11.7%	9.2%

### One Person Patient One Record Experience Record

- Mix of Meditech, SHARE, individual silos, 80+ systems (HCWs use 5 on average per patient)
- One Person One Record
  - 2014 December: Strategy Approval
  - 2015 July:
    - NSHA CIO appointed to lead project
    - RFP solutions hired to monitor procurement process to lead OPOR
  - 2016 March: Meeting moratorium
  - 2016 May/December: Gevity/Allscripts NSHA meetings
  - 2017 January: Request for Supplier Qualifications
  - 2017 February: Submissions from 4 big firms (**Epic**, Allscripts, Cerner, **Meditech**) and 2 small ones (Evident, Harris Healthcare Group)
  - 2017 June: Allscripts and Cerner named as only finalists based on 50 page RFSQ
  - 2017 August: Evident complaint
  - ....
  - ???
  - ...
  - 2022: rename to One Patient One Experience
  - 2023: rename to One Person One Record & 10-year \$365M contract with Oracle Health (formerly Cerner)
  - 2025 February: Planned partial roll-out
  - 2025 May: Delay to planned roll-out & new physician lead hiring...
  - 2025 December: New rollout planned date

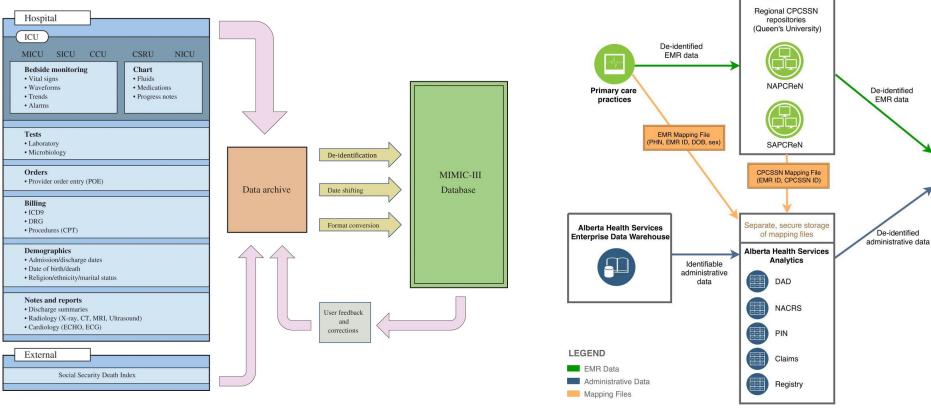
## Reality: fragmented EMRs requiring difficult linkage

- Interoperability is a competitive disadvantage
- Standarised format and vocabulary for EMR data: Observational Medical Outcomes Partnership (OMOP) common data model
- Observational Health Data Sciences and Informatics (OHDSI) tooling
- Fast healthcare interoperability resources (FHIR)



https://ohdsi.github.io/TheBookOfOhdsi/images/CommonDataModel/cdmDiagram.png

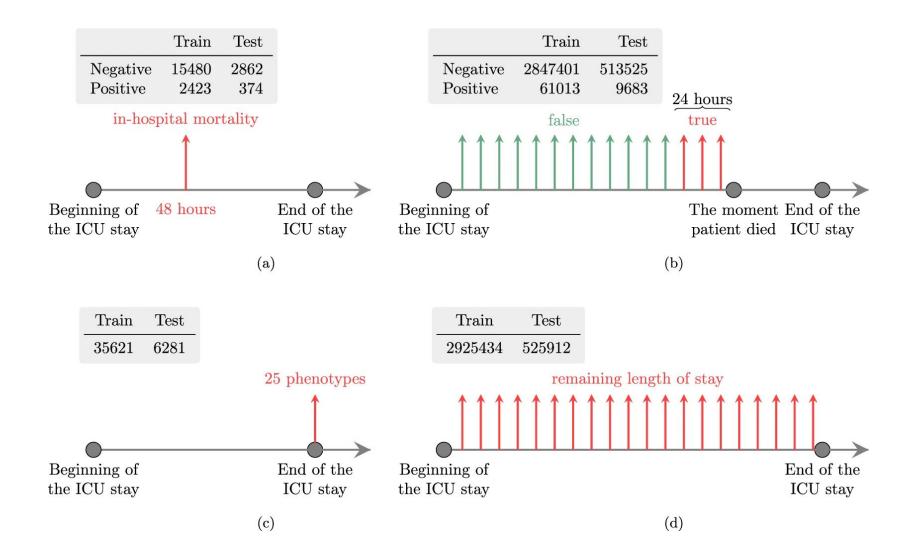
## EMR datasets: MIMIC-.\* / Canadian Primary Care Sentinel Surveillance Network / STARR



10 1038/sdata 2016 35

https://informatics.bmj.com/content/27/3/e100161

#### EMR allow us to ask complex questions



What kind of data is in an EMR?

#### Many types of data in EMRs

- Discrete Physiological Parameters e.g., blood test metric measures
- Diagnostic Imaging Data e.g., MRI image data
- Physiological Sensor Data e.g., EKG/EEG signal data
- Ordinal scale assessments e.g., frailty index

#### Text:

- Structured text e.g., CPT/ICD-10 codes (V89.2XXA, S06.0)
- <u>Semi-structured Text</u> e.g., {"Symptoms": "Head pain, dizziness, emesis", "Cause": "Car crash", "Diagnosis": "Likely concussion"....}
- <u>Unstructured Text</u>: "Patient was involved in a car crash and presented to the ER with pain, vomiting, and mild dizziness. Most likely they are concussed but should follow up with a head CT to confirm no other brain injuries"

CAT SCANS			•
ABDOMEN		ICD-10	DESCRIPTION
Abdaman/a aantmat 74150	ormations and	<b>Chromosomal</b>	Abnormalities (Including Down's Syndrome)
Abdomen w/o contrast	EC	Q64.79	Other congenital malformations of bladder and urethra
Abdomen w/o & w/ contrast		Q90.9	Down syndrome, unspecified
	10S	Q05.4	Unspecified spina bifida with hydrocephalus
CHEST/THORAX		Q05.8	Sacral spina bifida without hydrocephalus
Chest/Thorax w/o contrast71250	nitourinary Sy	stem Diseases (	Including Incontinence and UTI)
Chest/Thorax w/ contrast/4150	0S	N17.9	Acute kidney failure, unspecified
Chest/Thorax w/o & w/ contrast71270	20-80-64	5-100 000	

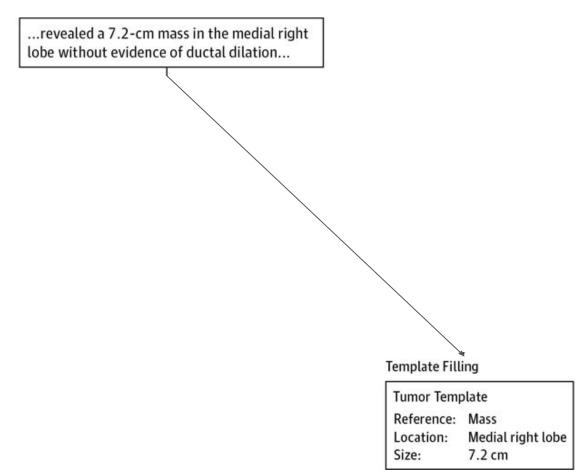
#### Medicine loves unstructured text

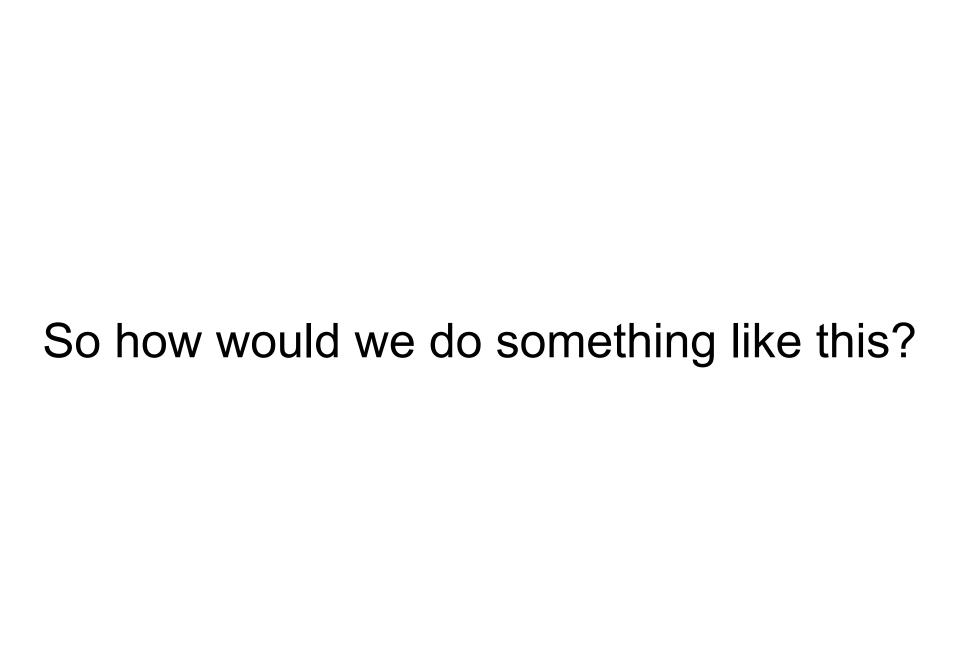
- Unstructured text is and will likely forever remain the primary form of communication in medical clinical settings.
- Highly flexible, efficient, and expressive across a range of communication contexts for medicine.
- Mainstay of charts, notes, consults, discharge summaries, procedure/operative logs.

S of U (whole exampled with 00 T $\chi_{\rm U}$ (pulphone following) of T $_{\rm U}$ (subphone agent) of E $_{\rm U}$ (small following)	Date:   N.T. To To W. E		OFFICE WEITS	PADES COUNTS
	1000		CONTRACTOR N	The state of the s
Physical Assessment	*D	*ND		3 III.4.A
Temperature pm	10000	2000000	ADDRESS TO THEM	Y FORZ PHONE (HOME) (MON)
Palse pm	100			taken + 000 19/4 05-05 Me 37
Respiration pm	0		DMPLOYER E-SL S	OF CHARLES AND ADDRESS OF THE PARTY OF THE P
Blood Prossure pm			10000000000000000000000000000000000000	CAROTE STATE OF THE STATE OF TH
Height pm			DATE/VITAL SIGNS	May - 2 - 2004
Weight pm		1		
Chest sounds q visit			1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	+ po (F) Oc Guylar fighters in the of
		_	THE WAS IN SE	
02 Sut (cf available)	1000	100	10 30	To as half the of much whereas & chie
Asthma Control:	. 0	ND	1/1/	an deal was coult refreshing degree
Cough, wherea or chest tightness (<4ds/wh)	11 45500	11000	Persont in white the	relational & I want the symptom shirted
Wake Shight (~DWO	-	-	LEW POWER CANADA CONTRACTOR	The state of the s
Physical activity limited due to asthma			thoughout lest \$10	statusty and now a constant a month ag
Needs Reliever with exercise	-		- Ward So h S. 1 (14	of allowed a for within exercise hadron.
Needs Retiever (~4/e/c)				4 4 4 6
Poscerbations (hospital admit, ER visit, Walk-in Clinic) since last visit			awate however to	glat from executly distributed for the 14
	1.5		the se nows	- 11 January 17 17 18 18
School-work absunce since but visit			a box & ban S	wis an Salbatawalow way 11/dags
Spirometry	D	ND		while he not totally, she has alway to
PEV Lore	-		- de	15th toursed, smoking also approvable but 5
FEV I post	100		13	I by the teach a space of the additionage one of
FEV1 %charge			6.0	bushand small but of a Le the knope
PEF me			17	
PEF post	-		200	il lune, and som Sougha 315 years a
PEF% charge	11 210		THE WINDS IN SEC. AS	
Review	. 0	ND		od court local such sympleses in the past
Definition of Authors	- 4	100	Text 1872 time sent	THE MOUNT SWEAT , BUT BUT BUT BUT GOLDEN, There ,
		_		an some phan with sugh thin it out
Action Plan.—(ACP)		_		
Action Plan - (norbal)		_		
Med Admin Technique			der lu	ed spriously by the result als equiviral
Warring signs	-	_		To come text also also 1861 then Gorm
Trigger factors	- 1		SCI.	THE THE THE REST OF THE GALL
Environmental control			Sprimetry C	BC & RMI' WAL.
Coping Writington			150 may 1 150	II was a TED strange of It and was con TE .
Medications			1. 1 10 11 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
Current			Mad Salutenil De	I year Since the thirty her we sign of 78 .
			a.D @	Miller and Art Last No.
			- Flower Tro Kee	cashic Spainted dysfunctioning
Prescribed			The second secon	LV. Con HX: HEL II Children, her post page
			- OCP 10	water complicated with progressy induced by
				what 6 was copper 5 plan bys.
Monitor potential side effects	D	ND	10	Ax Fether will a 65 NO because at mon ship
height outerprovisis, etc.	S 2725	10000		APPLY and DAY.
Referrals	D	ND	156	mid the & smooth & great, over it his child
Authora Education Program		100	The	without in a supportive hometop the days allow,
Authors Support Group	-		1.7	me in a southerest destroy in milk willias !
Specialist			4.7	~~~

### Unstructured text is challenging

- English language especially has many synonymous and highly flexible grammatical structure.
- Medical english has many synonyms and similar seeming non-synonyms:
  - Bilateral salpingectomy
  - Salpingectomy
  - Fallopian Transection
  - Fallopian Tubectomy
  - Fallopian Tubal Ligation
  - Tubal ligation
  - Tubal sterilisation
  - Tubal
  - CPT58600
- Now add typos and transcription errors!
- Difficult to search
- Difficult to summarize
- Difficult to analyze

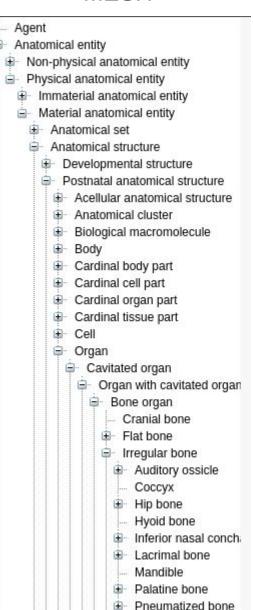




#### **MESH**

### **Natural Language Processing!**

- NLP is any computer-based method that handles/augments/transforms natural language so that it can be represented for computation.
- Approximate synonyms: text mining, text processing, computational linguistics
- Example problem:
  - 1. "Find every medical note in the EMR related to the spine"
  - 2. Identify key search terms e.g., "back", "spine", "vertebra", "lumbar", "neck", "cervical", "thoracic", "sacrum", "coccyx" (expertise, ontology/vocabulary)
  - 3. Search for EMR for these terms



Vertebra

Let's start simple: searching text

#### Searching for exact matches: Ctrl-F

Many exact match algorithms with varied properties (typically ctrl-F will mix and match them in a context-dependent way).

- Scan over all text and look for things that exactly match your query
- Make things more efficient: Boyer-Moore/Knuth-Morris-Pratt/Rabin-Karp etc.

```
P: word

T: There would have been a time for such a word

word

U doesn't occur in P, so skip next two alignments

P: word

T: There would have been a time for such a word

word skip!

word skip!

word skip!

word
```

#### More flexible searches for keywords: Regular Expressions

- Need to find "spine" and "spinal" = spin(a|e)/?
- Can also be used to capture words/before after: \w+\sspin(a|e)/?\s\w+
- Builds on lots of well-developed CS theory

You have a problem,
 you use regex, you
 now have 2 problems

Character	Description	Example
[]	A set of characters	"[a-m]"
\	Signals a special sequence (can also be used to escape special characters)	"\d"
	Any character (except newline character)	"heo"
^	Starts with	"^hello"
\$	Ends with	"world\$"
*	Zero or more occurrences	"aix*"
+	One or more occurrences	"aix+"
{}	Exactly the specified number of occurrences	"al{2}"
I	Either or	"falls stays"
()	Capture and group	

#### Regular expressions can get very complicated!

#### RCF5322 Email validation regex:

```
 (?:[a-z0-9!\#\$\%\&'^*+/=?^__`\{]\}\sim-]+(?:\.[a-z0-9!\#\$\%\&'^*+/=?^__`\{]\}\sim-]+)^*|"(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f]|\(\x01-\x09\x0b\x0c\x0e-\x7f])^*")@(?:(?:(?:(2.5[a-z0-9])?)^*]-[a-z0-9])?\(?:[a-z0-9])?\(?:[a-z0-9])?\(?:[a-z0-9])?\(?:(2.5[a-z0-9]))\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(?:[a-z0-9])\(
```

## Can we make the text easier to search instead?

#### Most NLP methods start with text normalisation

- 1. Tokenisation
- 2. Normalising word formats
- 3. Segmenting sentences

### Splitting text into words: Segmentation/Tokenizing

- Breaking text into individual units (letters/morpheme/words/sentences/paragraphs)
   can make it much easier to handle.
- Process is known as tokenisation (a subset of segmentation)
- Units you break into are known as tokens:

"Indication of significant spinal contusions."

-> "Indication" "of" "significant" "spinal" "contusions."

- Easy approach:
  - split on spaces
  - Has to be fast (finite state automata)
- Challenges: punctuation can matter (e.g., 01/02/22), not all languages use spaces, may want to treat multiword expressions (MWE) as tokens e.g., "New York", "bilateral salpingo-oophorectomy", "ice box"/"ice-box"/"icebox"

#### Simplifying language: word normalisation

- "Ph.D.", "PhD", "phd" probably shouldn't be counted differently
- Case folding: collapse everything to lowercase (although case can often be informative: "US" vs "us")
- Lemmatization: identifying words with common root (lemma) e.g., "operation" and "operations" -> "operation"; "am", "are", "is" -> "be"
  - "Surgeon is performing surgical procedures" -> "Surgeon be perform surgical procedure"
- Requires morphological parsing splitting <u>stems</u> (central morpheme) from <u>affixes</u> (modifying/adidtional meaning)

- Lemmatization is difficult : alternative = stemming
  - Remove final affixes e.g., remove "-ing, -s, -ational, -sses"
  - "This was not the correct operation" -> "Thi wa not the correct operat"

### Splitting sentences: sentence segmentation

- Sentences are delineated on punctuation: ".", "?", "!"

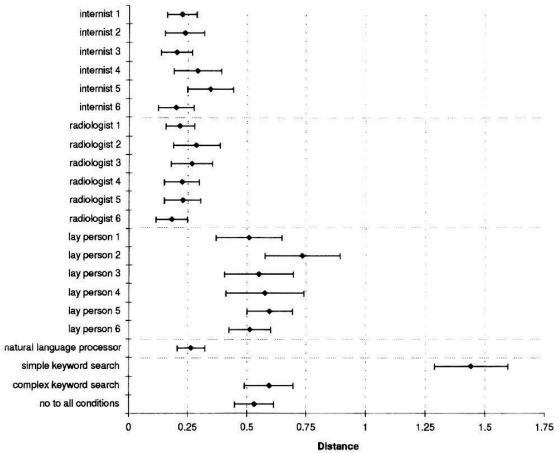
- Often we want to segment phrases/clauses, more challenging:
  - "Patient presented to ER with pain/confusion, most likely as a sequelae of a head injury" ->
    - ["Patient", "presented", "to" "ER", "with", "pain", "confusion"]
    - ["most", "likely", "as", "a", "sequelae", "of", "a", "head", "injury"]

#### Hash-based text search

- Can find exact matches very efficiently
- Tokenize/lemmatise/normalise words in each note, then hash:
  - Note 1: ["spine", "car", "head"] -> [a11, a92, a53]
  - Note 2: ["car", "tree", "CT"] -> [a92, a57, a99]
- Hash query words "back", "spine", "vertebra", "lumbar", "neck", "cervical", "thoracic", "sacrum", "coccyx":
  - a55, a11, ...
- See if query hashes are present in note hash sets
  - a55 in Note1 = No, a55 in Note2 = No
  - a11 in Note1 = Yes, a11 in Note 2 = No

#### Can use these to create manual rules

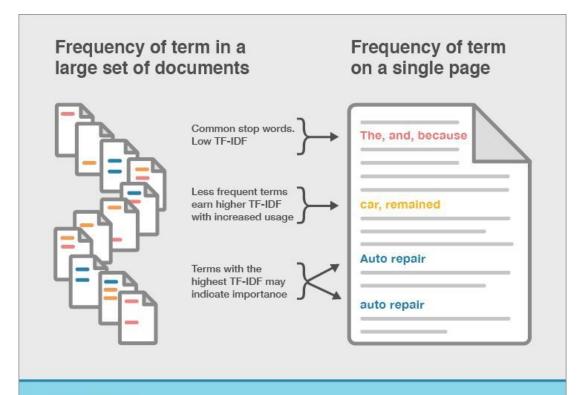
```
if finding is in ("pneumothorax"; "hydropneumothorax")
and certainty-modifier is not in
("no"; "rule out"; "cannot evaluate")
and status-modifier is not in
("resolved")
then
conclude true;
endif;
```



https://doi.org/10.7326/0003-4819-122-9-199505010-00007

## What if we don't know the query terms in advance?

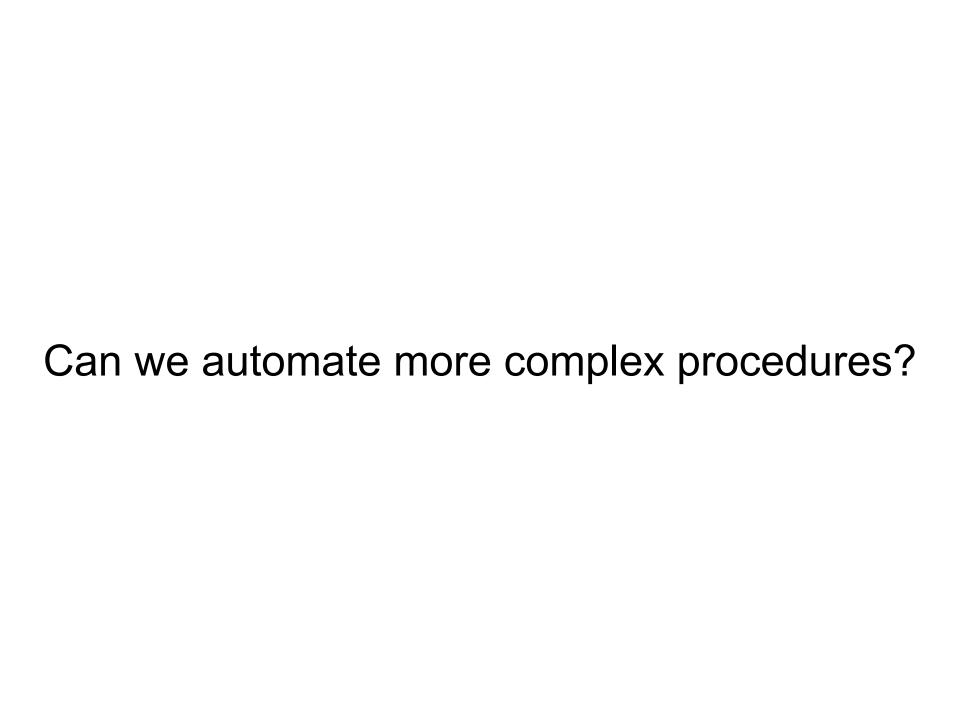
#### Identify frequently used terms

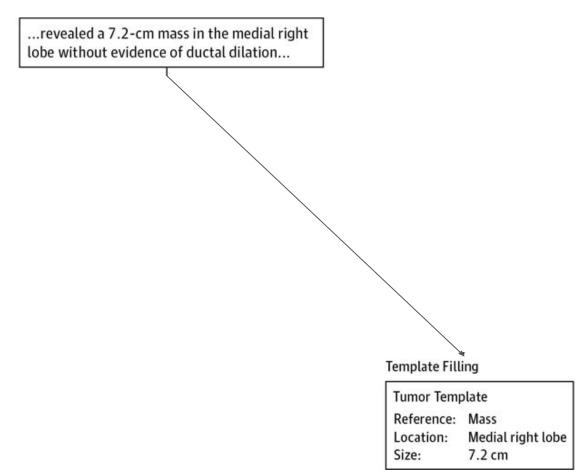


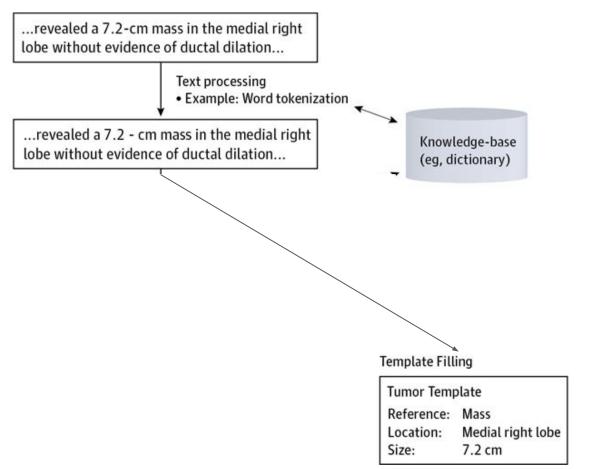
TF-IDF

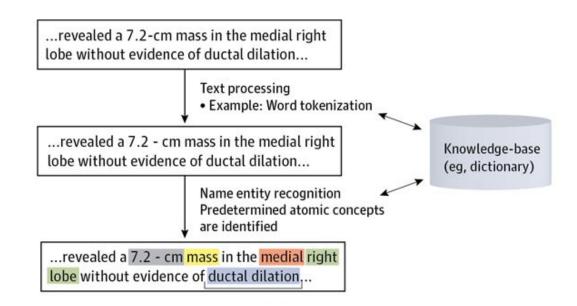
Term frequency-inverse document frequency (TF-IDF) measures the importance of a keyword phrase by comparing it to the frequency of the term in a large set of documents. Many advanced textual analysis techniques use a version of TF-IDF as a base.

- Find highest TF-IDF terms
- Filter them manually for new search terms
- Apply prior search
   approaches (or any of the
   fuzzy matching approaches)
- Among other unsupervised approaches (e.g., following material)
- Matrix encoding using tf-IDF and/or co-occurrence









#### Template Filling

**Tumor Template** 

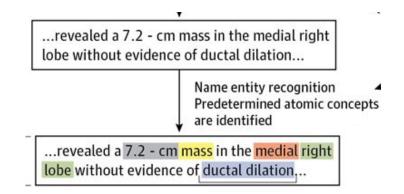
Reference: Mass

Location: Medial right lobe

Size: 7.2 cm

## Identifying tokens referring to things: Named Entity Recognition

- Identify specific categories of entities e.g., places, times, anatomy
- Lots of pre-trained approaches/vocabularies
- Text classification problem => requires some way to encode text to a numerical vector

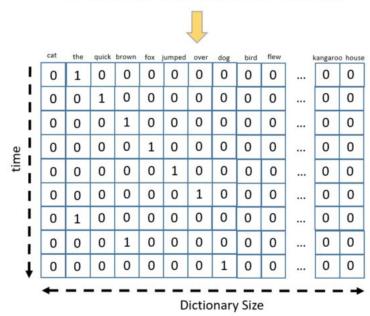


## Encoding text as a vectors

- One-Hot encoding
- TF-IDF frequency based encodings
- Very large vectors with even moderate vocabulary size
- Very sparse vectors (lots of 0s)

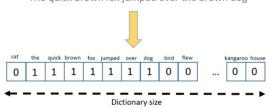
#### One-Hot Encoding

The quick brown fox jumped over the brown dog



**Document Vectorization** 

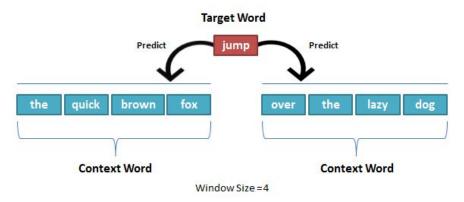
The quick brown fox jumped over the brown dog



Sum over columns for each note to get a vector representation of the document instead (TF-IDF is a normalisation of this representation)

## Reducing the dimensionality of these vectors

- Standard dimensionality reduction methods struggle
- Text has semantic (meaning) AND syntactic (grammar) components
- We want to find a lower dimensional embedding that captures these aspects
- "You shall know a word by the company it keeps" (Firth, J. R. 1957:11)
- "The meaning of a word is its use in the language" (Wittgenstein)
- Can we use the CONTEXT of a given word to find a meaningful vector representation?
- Word2vec:
  - Skip-Gram:
    - Predict context from word
    - Better for rare words
  - Continuous Bag of Words:
    - Predict word from context
    - Better for common words



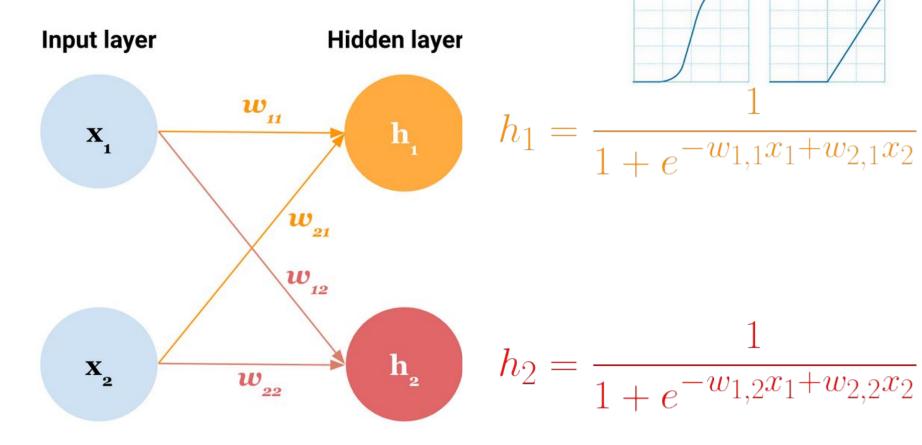
#### **Neural Networks -**

- Need: efficient models that capture high-dimensional non-linear relationships
- Solution: stack many simple models with a non-linearity (e.g., logistic / ReLU)

sigmoid

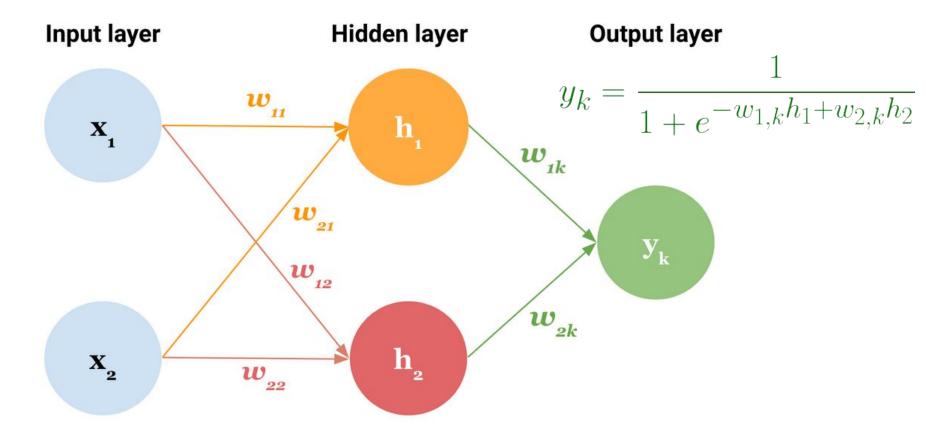
ReLU





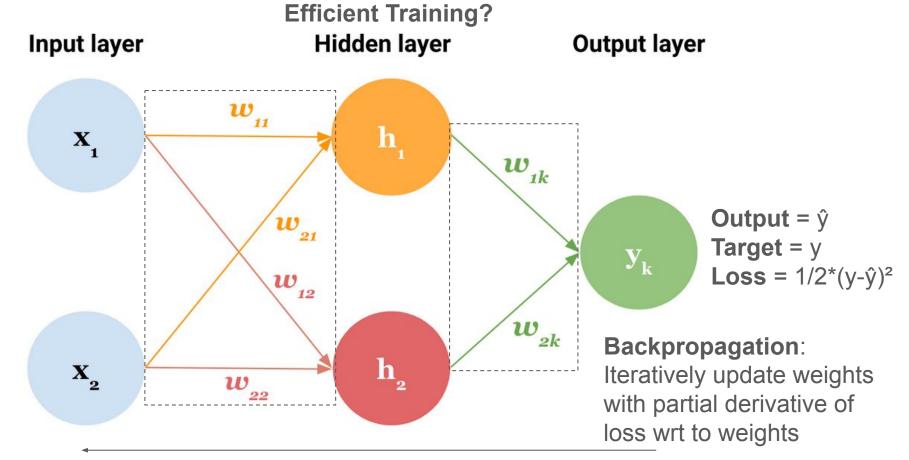
#### **Neural Networks -**

- Need: efficient models that capture high-dimensional non-linear relationships
- Solution: stack many simple models with a non-linearity (e.g., logistic / ReLU)
- Neural Networks:

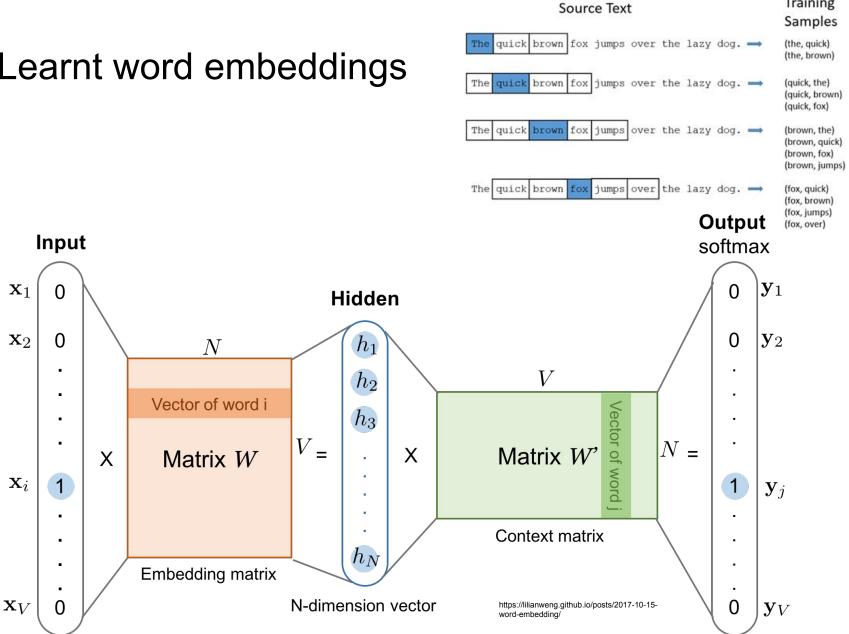


#### **Neural Networks**

- Need: efficient models that capture high-dimensional non-linear relationships
- Solution: stack many simple models with a non-linearity (e.g., logistic / ReLU)
- Neural Networks:



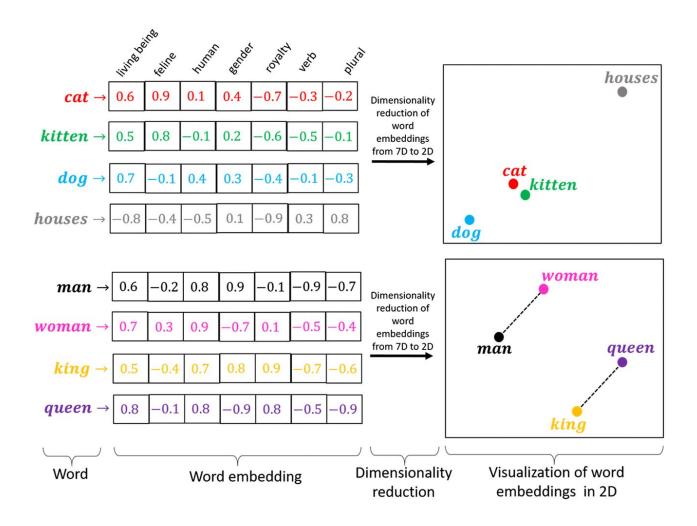
### Learnt word embeddings



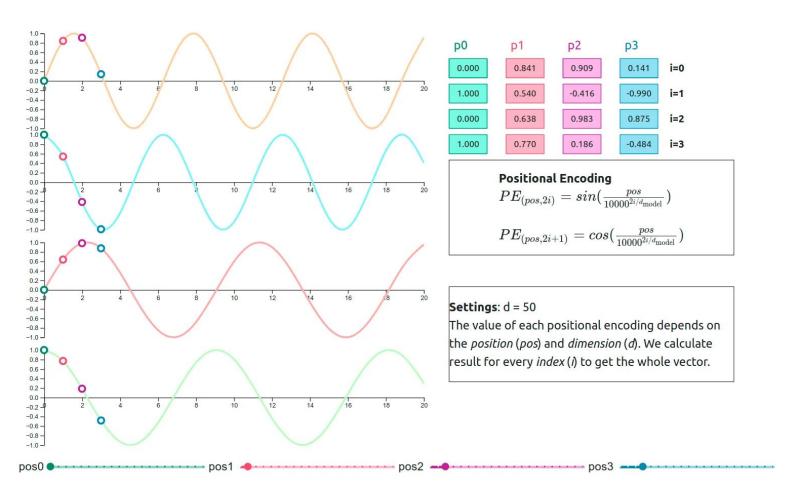
**Training** 

Word2Vec/GloVe/FastText -> ELMo -> BERT -> ERNIE -> GPT-3/Megatron/T5 -> GPT-4/Llama3

### Learnt embeddings are powerful



## Beyond word embeddings: encoding position



## Attention mechanisms (massive topic!)

- Self-similarity vs similarity to other words
- Auto-regressive

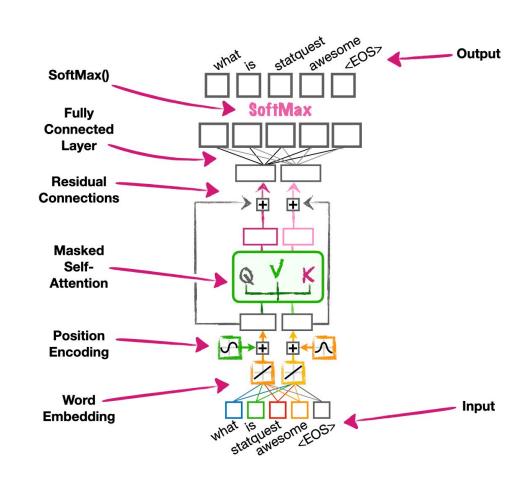
   (mask self-attention)
   if only prior words

```
The FBI is chasing a criminal on the run.
The FBI is chasing a criminal on the run.
     FBI is chasing a criminal on the run.
The
     FBI is chasing a criminal on the run.
The
     FBI
               chasing a
                           criminal on the run.
The
               chasing
                           criminal on
                                         the run.
The
     FBI
               chasing
                           criminal
                                          the
                        a
                                     on
                                              run .
```

https://shorturl.at/KCfx1

## Attention mechanisms (massive topic!)

- Self-similarity vs similarity to other words
- Auto-regressive
   (mask self-attention)
   if only prior words
- Combining all these mechanisms with a lot of data gives you transformer models (e.g., GPT1-4)

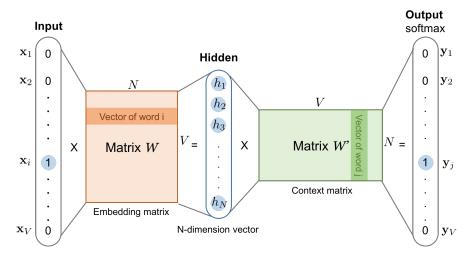


https://shorturl.at/J1ff6

## Custom embeddings and fine-tuning

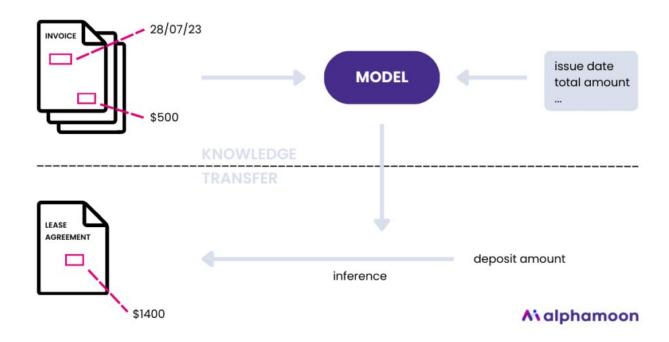
- Same approach can be used beyond just words:
- Med2Vec
- EHR2Vec
- BioALBERT

- Corpora used to create embedding may not be a good fit for specialised text (i.e., EMRs aren't representative of the internet at large... we hope).
- Repeat training on your data but initialise with pre-trained weights



### Multimedia/multimodal embeddings

- This approach can be extended to joint embeddings of multiple data types (e.g., "multimodal" CLIP embeddings/Diffusion in Module 3)
- Zero-shot: using model trained on your unrelated problem with no fine-tuning

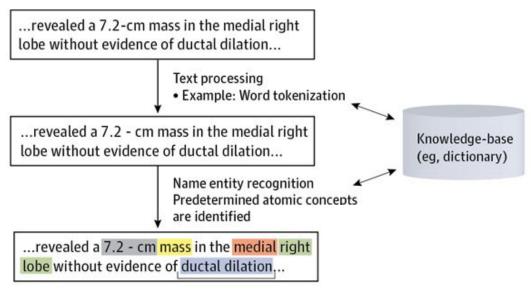


May require exponential data (especially if multimodal)

https://arxiv.org/pdf/2404.04125

# With embeddings we can build/use models for more complex problems

## Train classifier on labelled medical text (e.g., ontology) = Named Entity Recognition



#### Template Filling

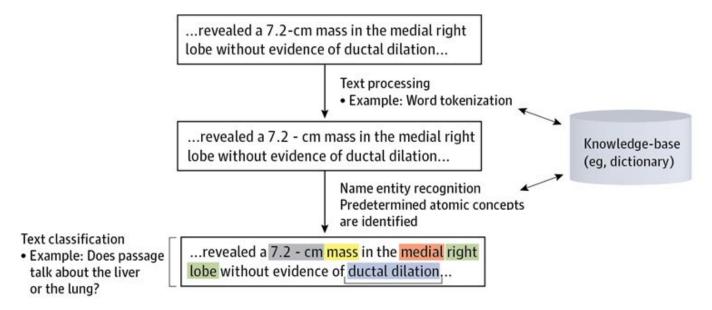
**Tumor Template** 

Reference: Mass

Location: Medial right lobe

Size: 7.2 cm

## Train document classifier on EMR notes labelled by organ => Text classification



#### Template Filling

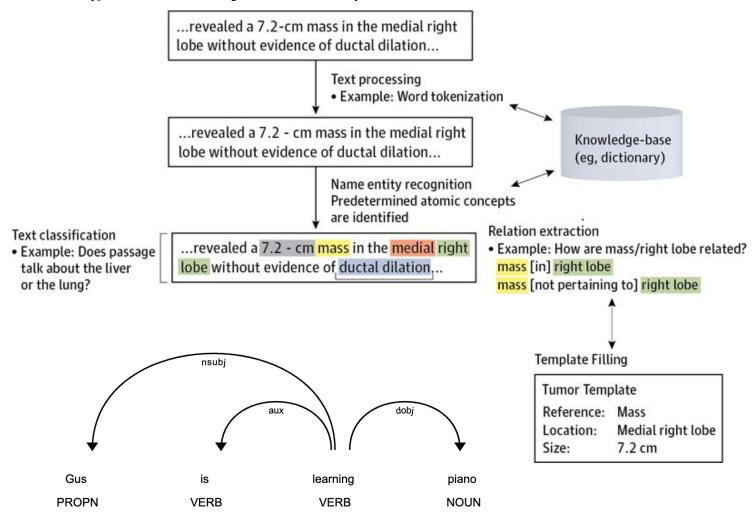
Tumor Template

Reference: Mass

Location: Medial right lobe

Size: 7.2 cm

## Use classifier trained to identify parts of speech and their relations (previously HMMs)



#### Overview

- Describe electronic medical/health record systems and the types of data they typically contain
- Distinguish structured, semi-structured, unstructured text data
- Describe approaches to searching text
- Outline key steps in preparing text for analysis
- Explain the general concept of learnt word embeddings
- Explain how embeddings can be tuned/customised
- Identify differences between named entity recognition, parts of speech tagging, and dependency parsing

Not covered: fuzzy search and text indexing