Improving Doctor-Patient Interaction with AI-Enabled Medical Note Taking

Michael Brudno
Centre for Computational Medicine
Genetics and Genome Biology
Department of Computer Science

SickKids®
THE HOSPITAL FOR SICK CHILDREN

UNIVERSITY OF TORONTO

Medical data circa 2018

Pain of EHRs


The New York Times Magazine
THE HEALTH ISSUE
HOW TECH CAN TURN DOCTORS INTO CLERICAL WORKERS

Pain of EHRs

PERSPECTIVE NOV 08, 2018

Getting Rid of Stupid Stuff

Ashton M. | N Engl J Med 2018; 379:1789-1791

In an effort to reduce unintended burdens for clinicians, leaders at a health system in Hawaii asked all employees to look at their daily documentation experience and report anything in the EHR that they thought was poorly designed, unnecessary, or just plain stupid.

Pain of EHRs

The New York Times Magazine THE HEALTH ISSUE

HOW TECH CAN TURN DOCTORS INTO CLERICAL WORKERS

PERSPECTIVE NOV 08, 2018

Getting Rid of Stupid Stuff

Ashton M. | N Engl J Med 2018; 379:1789-1791

In an effort to reduce unintended burdens for clinicians, leaders at a health system in Hawaii asked all employees to look at their daily documentation experience and report anything in the EHR that they thought was poorly designed, unnecessary, or just plain stupid.

THE NEW YORKER

ANNALS OF MEDICINE NOVEMBER 12, 2018 ISSUE

WHY DOCTORS HATE THEIR COMPUTERS

Digitization promises to make medical care easier and more efficient. But are screens coming between doctors and patients?

By Atul Gawande

Electronic Health **Records**

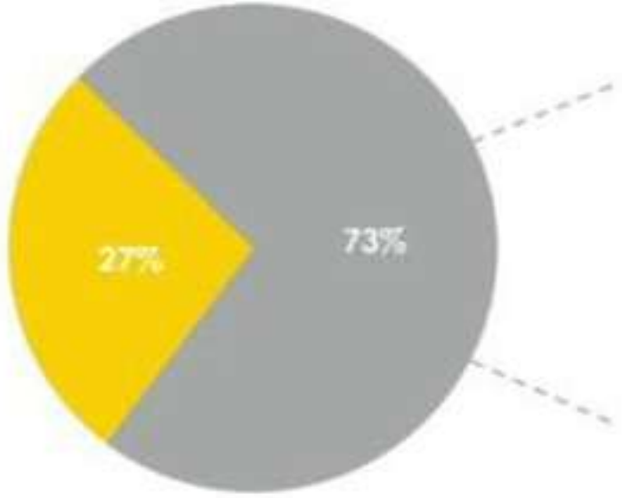
Structured data in EHR has little value to clinicians, and is centered around billing:

- What doctors did patients see
- What drugs were they given
- What tests were run
- Insurance codes

Doctors see the EHR as “a necessary evil”, and utilize it minimally.

Where do doctors spend their time?

Job Time Allocation in the Ambulatory Setting



Everything else

38% EHR[1] documentation and review

19% Administrative tasks related to insurance or billing

11% EHR review of tests, medications or other orders

6% Clinical planning and logistics

And this doesn't include:

- Self-reported **1 to 2 hours of evening** time spent on administrative tasks
- **39% of "patient-facing time"** spent in the EHR

1. Electronic health record

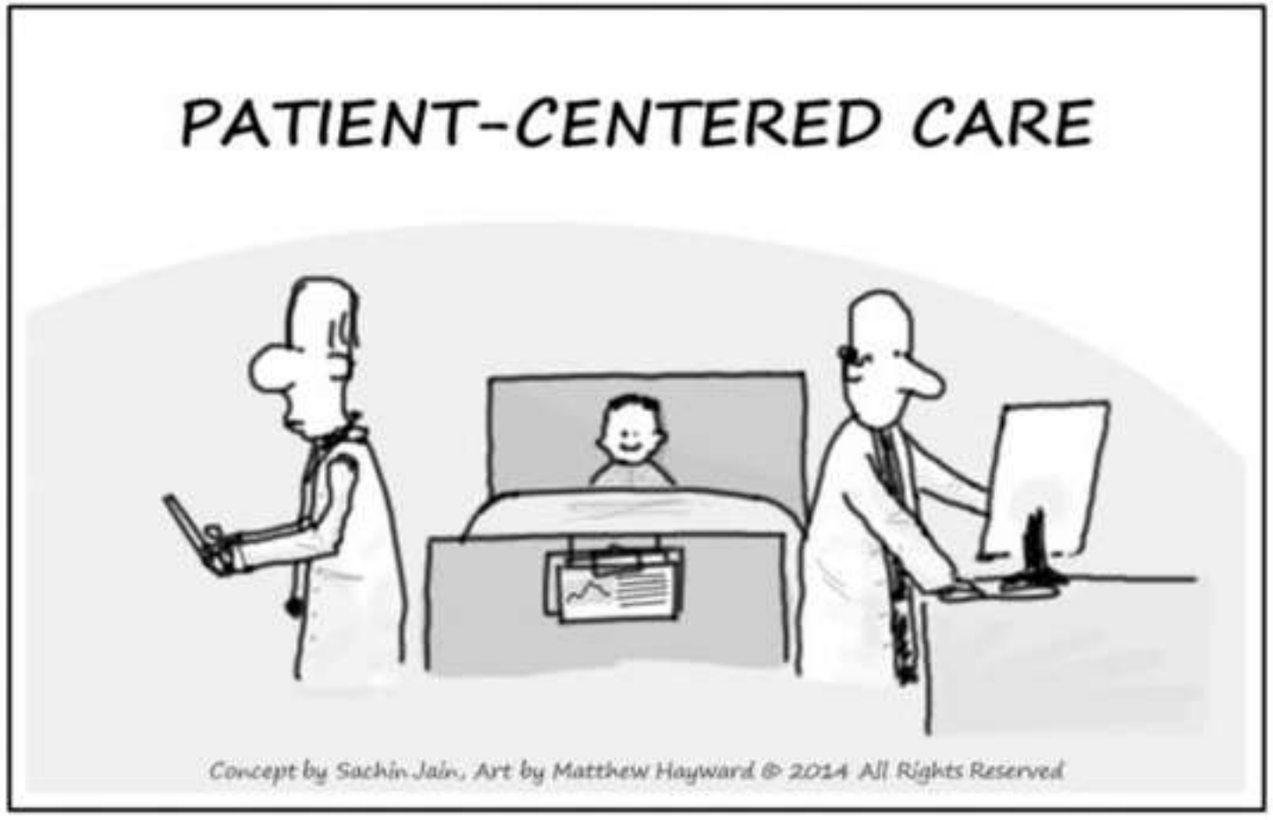
Source: Sinsky, Christine, MD; Lacey Colligan, MD; Ling Li, PhD; Sam Rivkin; Lindsey Garavalia; Johanna Westbrook, PhD; Michael Tutty, PhD; and George Miles, MD. "Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties." Annals of Internal Medicine 165.11: 753-60. 6 Dec. 2016. Web. 21 Feb. 2019. Old Healthcare Archive.

Patient-centric care circa 2018

Technology can be over and under-utilized simultaneously

- Computer, not patient, gets most attention
- Clinical notes are sporadic and inherently lossy
- Technology should be unobtrusive, but be available

Need for better technology, improved quality of captured data, and better patient and clinician experience



Paper or Computer?

Paper or Computer?

Paper pro:

- Intuitive, familiar workflow
- Ability to use signs, pictures and shorthand
- Cheap

Paper or Computer?

Paper pro:

- Intuitive, familiar workflow
- Ability to use signs, pictures and shorthand
- Cheap

Paper cons:

- Missing information
- Need extra time to transfer to EHR system
- Inconvenient to take photos or record videos
- Impossible to do computation

"Careful in there. The last doctor to go into Medical Records has still not been found."

Paper or Computer?

Computer pro:

- Faster than handwriting*
- Easier to search
- More structured
- Easier to share

Paper or Computer?

Computer pro:

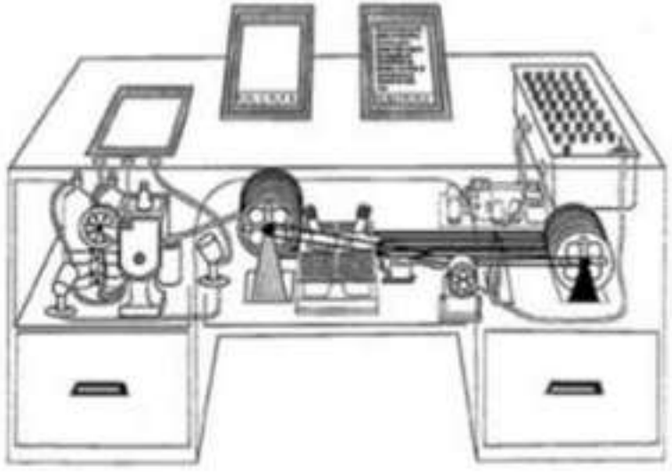
- Faster than handwriting*
- Easier to search
- More structured
- Easier to share

Computer cons:

- Missing information
- Slow* and distracting
- Impedes patient interaction
- Inconvenient to take photos or record videos

Future of note taking

"AS WE MAY THINK"
(1945)



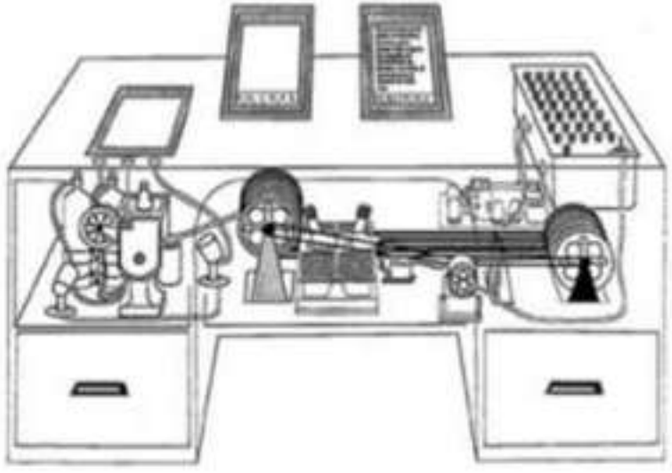
Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and to coin one at random, memex will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.

26

Vannevar Bush, 1945

Future of note taking

"AS WE MAY THINK"
(1945)



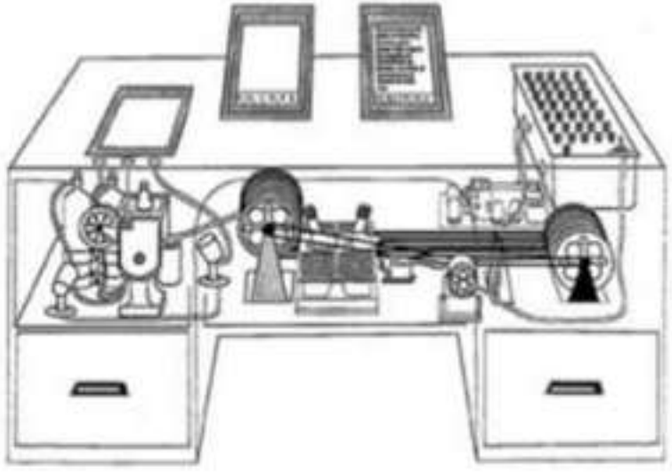
Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and to coin one at random, memex will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.

26

Vannevar Bush, 1945

Future of note taking

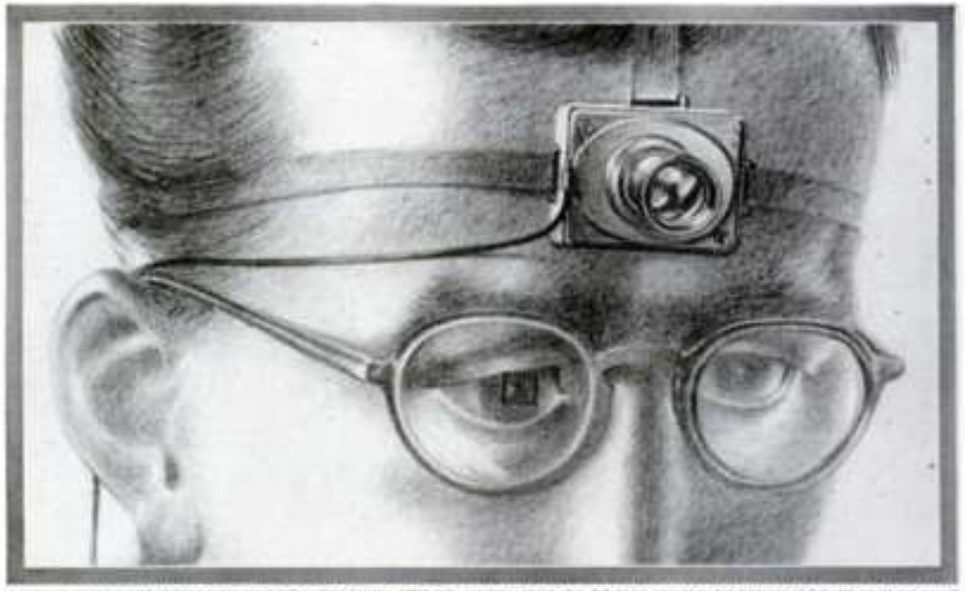
"AS WE MAY THINK"
(1945)



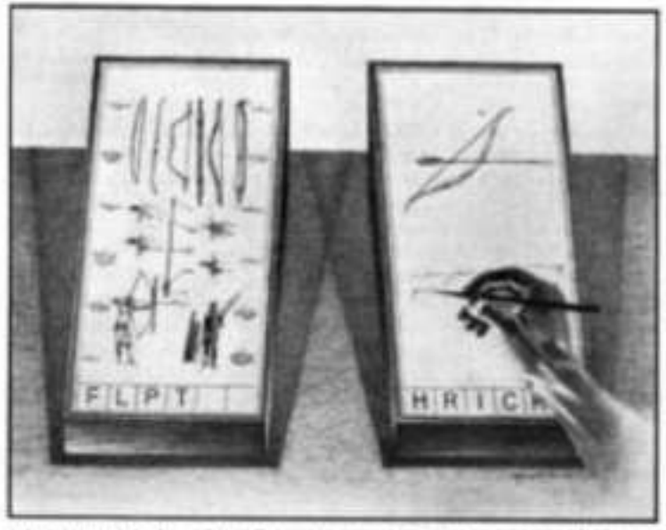
Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and to coin one at random, memex will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.

26

Vannevar Bush, 1945



A SCIENTIST OF THE FUTURE DESIGNS EXPERIMENTS WITH A TINY CAMERA FITTED WITH UNIVERSAL-FOCUS LENS. THE SMALL SQUARE IN THE EYEGLASS AT THE LEFT SHAPES THE OBJECT



MEMEX IN USE is shown here. On one transparent screen the operator of the future writes notes and commentary dealing with reference material which is projected on the screen at left. Insertion of the proper code symbols at the bottom of right-hand screen will tie the new item to the earlier one after notes are photographed on supersensitive film.

Artificial Intelligence

Handwriting recognition

Speech recognition

Speaker diarization

Natural language processing

Clinical decision support

Artificial Intelligence

Handwriting recognition

Speech recognition

Speaker diarization

Natural language processing

Clinical decision support

Artificial Intelligence

Handwriting recognition

Speech recognition

Speaker diarization

Natural language processing

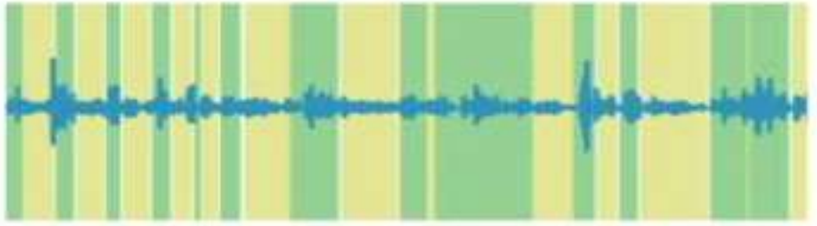
Clinical decision support

Artificial Intelligence

Handwriting recognition

Speech recognition

Speaker diarization



Natural language processing

Clinical decision support

Artificial Intelligence

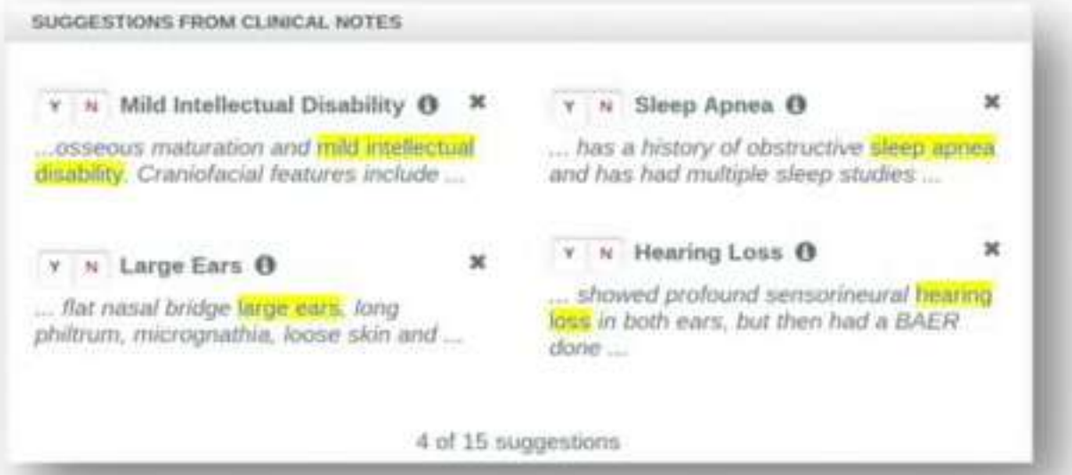
Handwriting recognition

Speech recognition

Speaker diarization

Natural language processing

Clinical decision support



Artificial Intelligence

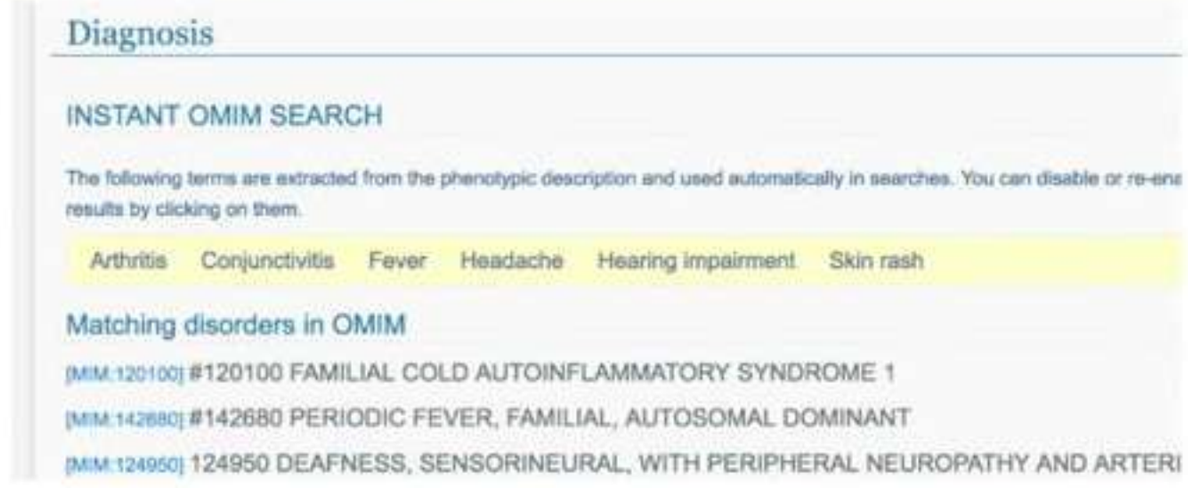
Handwriting recognition

Speech recognition

Speaker diarization

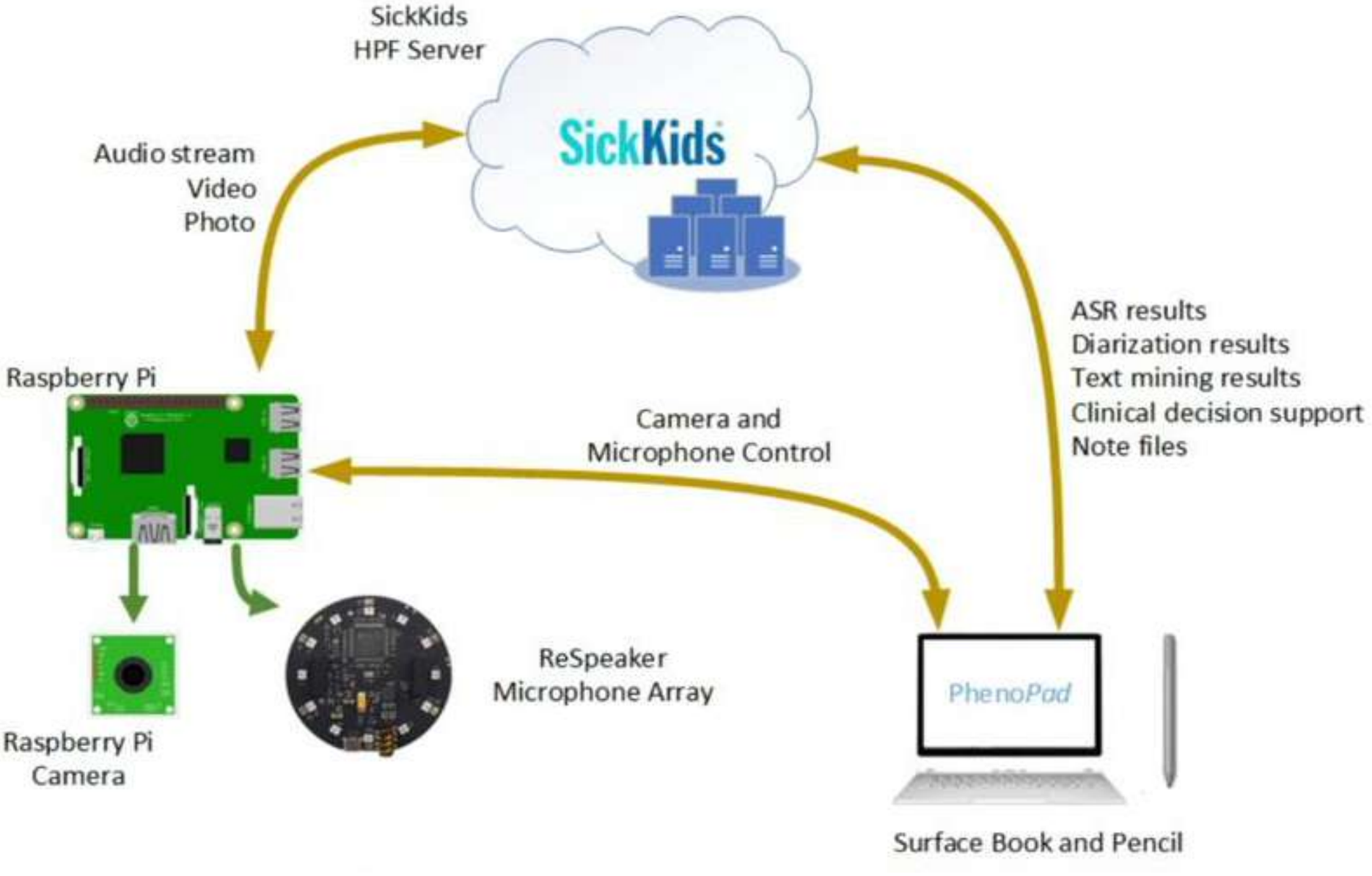
Natural language processing

Clinical decision support



PhenoPad: a new way of taking notes (for clinical genetics)

PhenoPad System Architecture



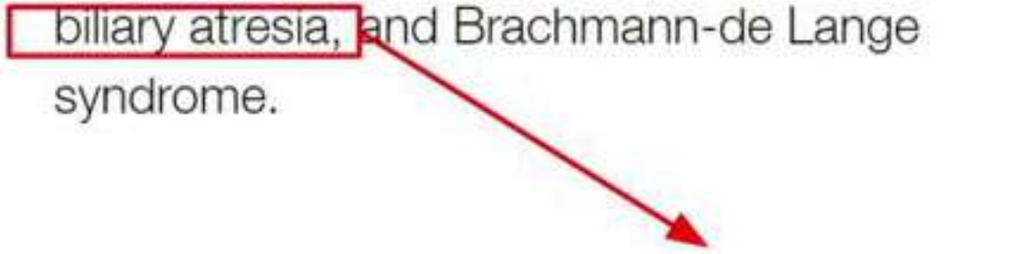
Demo

More details on the annotation example

There was marked variability in the severity and nature of manifestations with 2 having severe hand and foot involvement in addition to craniofacial changes compatible with a diagnosis of Freeman-Sheldon syndrome. Other apparently unrelated hereditary disorders in the family included ectrodactyly, biliary atresia, and Brachmann-de Lange syndrome.

More details on the annotation example

There was marked variability in the severity and nature of manifestations with 2 having severe hand and foot involvement in addition to craniofacial changes compatible with a diagnosis of Freeman-Sheldon syndrome. Other apparently unrelated hereditary disorders in the family included ectrodactyly, biliary atresia, and Brachmann-de Lange syndrome.



HP:0005912
Biliary atresia

More details on the annotation example

There was marked variability in the severity and nature of manifestations with 2 having severe hand and foot involvement in addition to craniofacial changes compatible with a diagnosis of Freeman-Sheldon syndrome. Other apparently unrelated hereditary disorders in the family included ectrodactyly, biliary atresia, and Brachmann-de Lange syndrome.

HP:0002260
Abnormal facial shape

HP:0005912
Biliary atresia

More details on the annotation example

There was marked variability in the severity and nature of manifestations with 2 having severe hand and foot involvement in addition to craniofacial changes compatible with a diagnosis of Freeman-Sheldon syndrome. Other apparently unrelated hereditary disorders in the family included ectrodactyly, biliary atresia, and Brachmann-de Lange syndrome.

HP:0002260
Abnormal facial shape

HP:0005912
Biliary atresia

***Synonyms for* HP:0002260**

Facial dysmorphism

Malformation of face

Unusual facial appearance

Unusual facies

Facial Dysmorphism

Funny looking face

Dysmorphic facial features

Deformity of face

Dysmorphic facies

Distinctive facies

Distortion of face

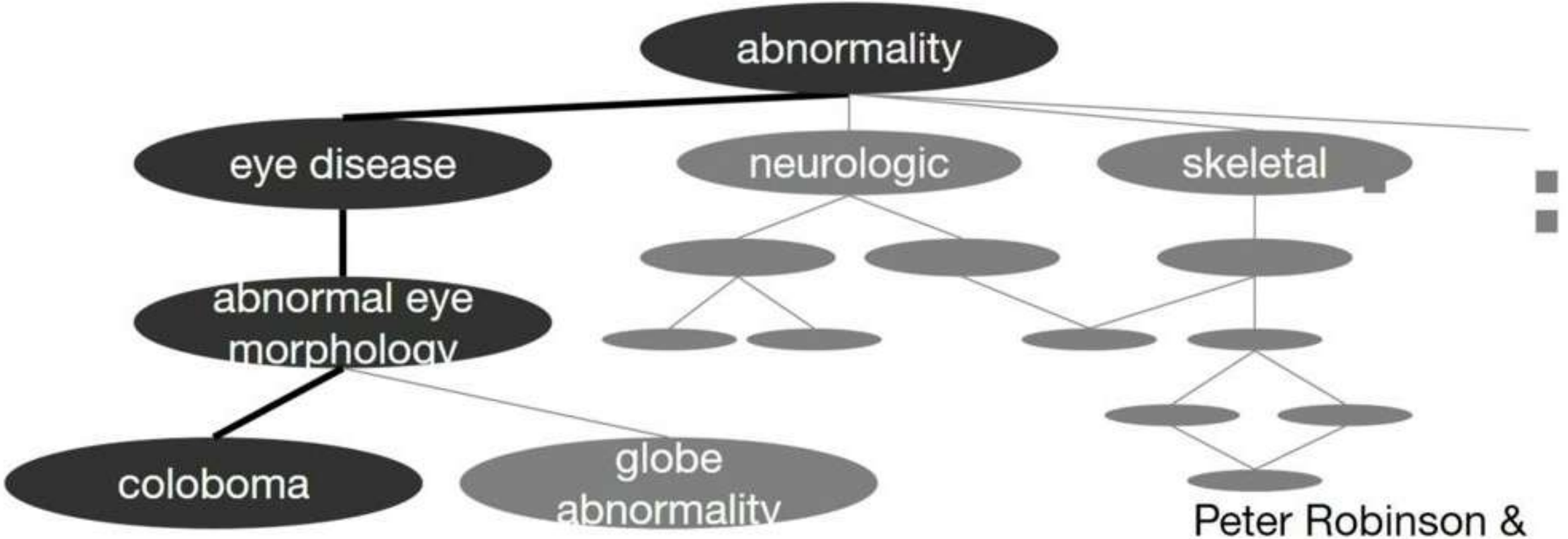
Challenges and opportunities

- Large number of concepts in medical ontologies
  - More than 12k terms in HPO
  - Named entity recognition methods (e.g. LSTM-CRF, Lampel et. al 2016) work well for few classes (<10)
- Popular named entity recognizers need labeled data
- Labeled datasets do NOT cover most concepts
  - Expensive to annotate medical text by experts
- + side: Medical ontologies provide a taxonomy of concepts

Human Phenotype Ontology

12,000+ terms

100,000+ links to 5,000+ OMIM/ORDO Disorders



Peter Robinson &
Monarch Consortium

More details on the annotation example

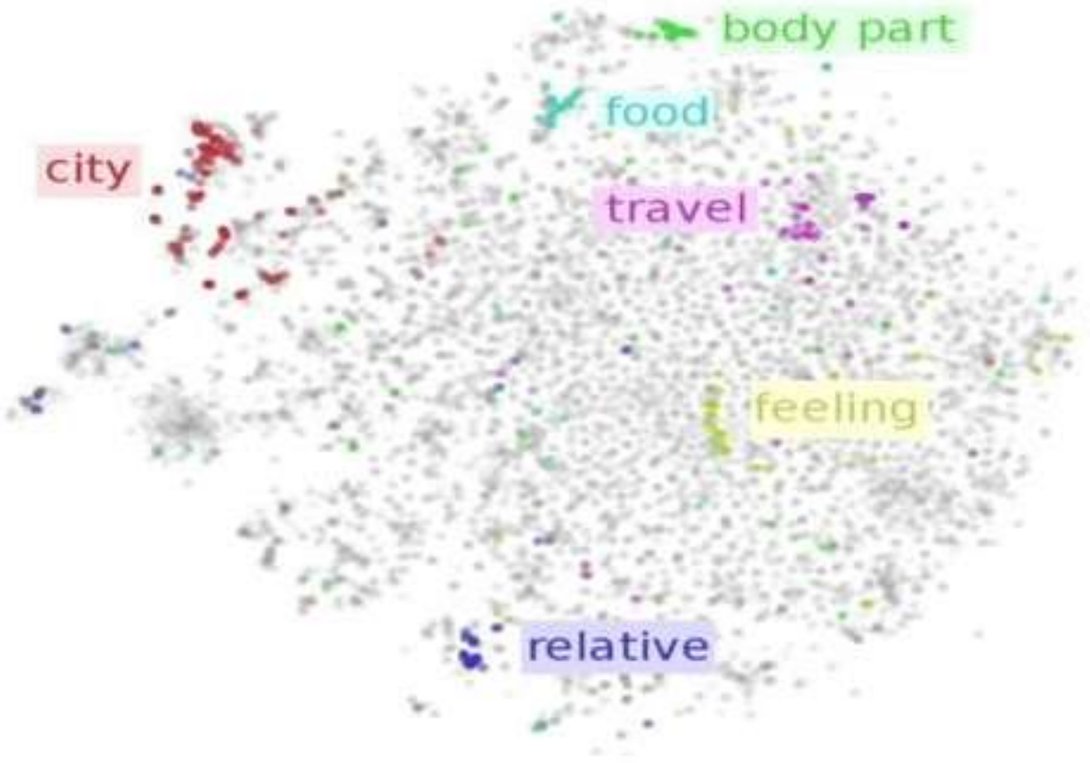
There was marked variability in the severity and nature of manifestations with 2 having severe hand and foot involvement in addition to craniofacial changes compatible with a diagnosis of Freeman-Sheldon syndrome. Other apparently unrelated hereditary disorders in the family included ectrodactyly, biliary atresia, and Brachmann-de Lange syndrome.

Challenges and opportunities

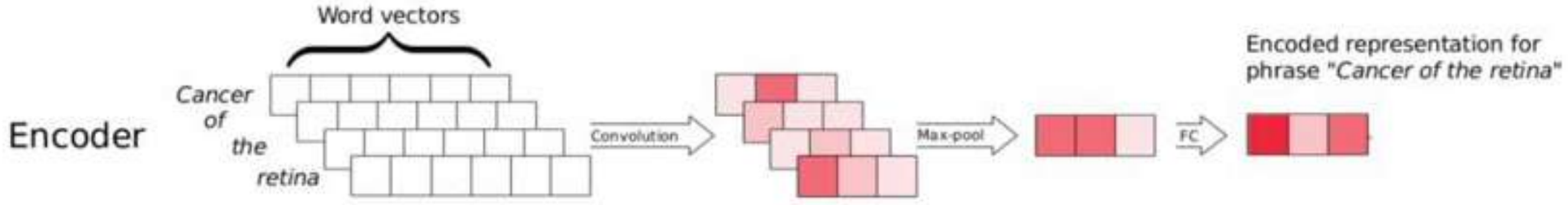
- Large number of concepts in medical ontologies
  - More than 12k terms in HPO
  - Named entity recognition methods (e.g. LSTM-CRF, Lampel et. al 2016) work well for few classes (<10)
- Popular named entity recognizers need labeled data
- Labeled datasets do NOT cover most concepts
  - Expensive to annotate medical text by experts
- + side: Medical ontologies provide a taxonomy of concepts

Word embeddings

- Each word is mapped to a vector in high-dimensional space
- The vector representations are learned by training on large corpora
- Words that occur in similar contexts tend to have closer vectors



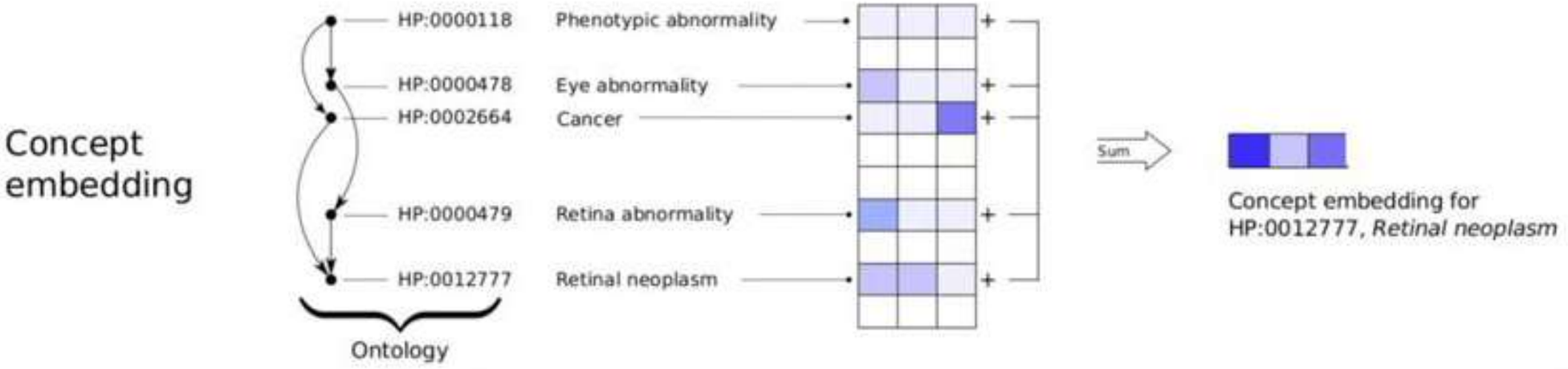
Phrase Embedding



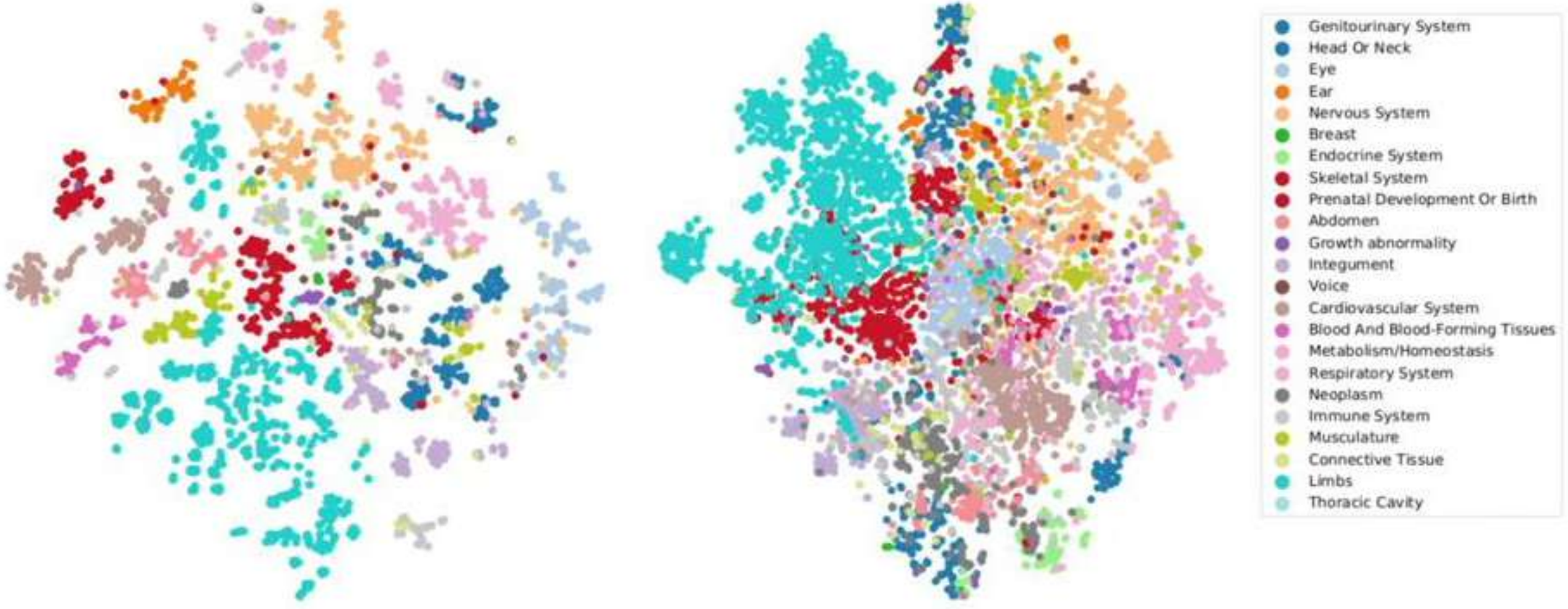
Hierarchical Concept Embedding

- Instead of directly learning embedding H_c for concept c, we learn *"what it adds"* to its parent as $\tilde{H}_c$: $H_c = \tilde{H}_c + \sum_{c' \in ancestors(c)} \tilde{H}_{c'}$

- The only embedding parameters are $\tilde{H}_c$ and H_c is completely conceptual, computed based on $\tilde{H}_c$



T-SNE representations of classes



(a) Hierarchy used for training

(b) Hierarchy NOT used for training

Other methods

- **BioLarK:** Rule based concept recognizer customized for HPO
- **OBO**: Rule based tool mainly customized for HPO
- **NCBO**: Rule based tool with access to hundreds of biomedical ontologies

Experiments: NCR trained on HPO

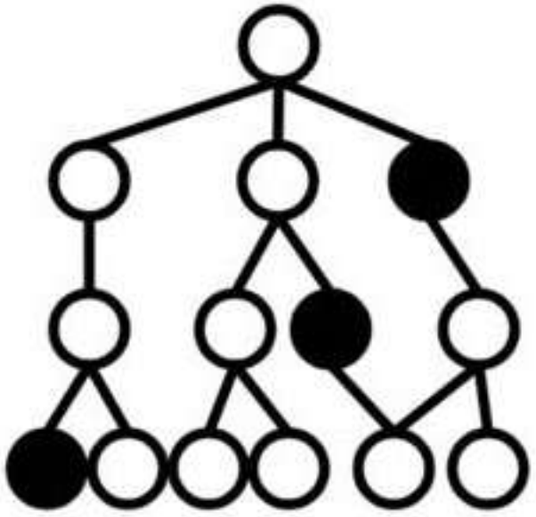
- 228 annotated PubMed abstracts (Groza et. al 2015)
  - Manually annotated with HPO concepts
  - Used 40 as validation to tune calling threshold
  - Tested on 188 remaining ones

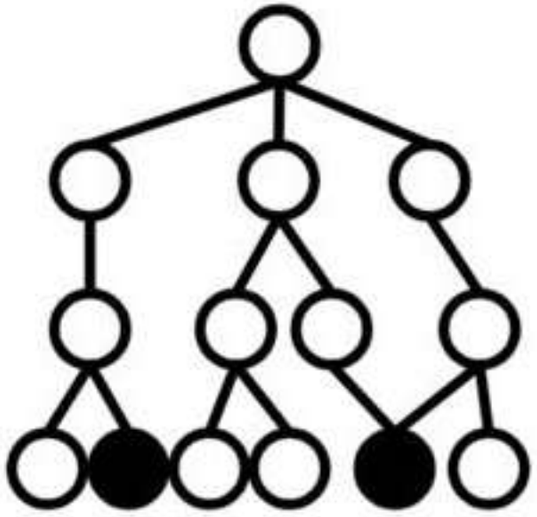
- 39 Clinical reports from Undiagnosed Disease Program (UDP)
  - HPO concepts loosely annotated for each patient
  - Used the threshold tuned for PubMed experiments

Results on annotated PubMed abstracts

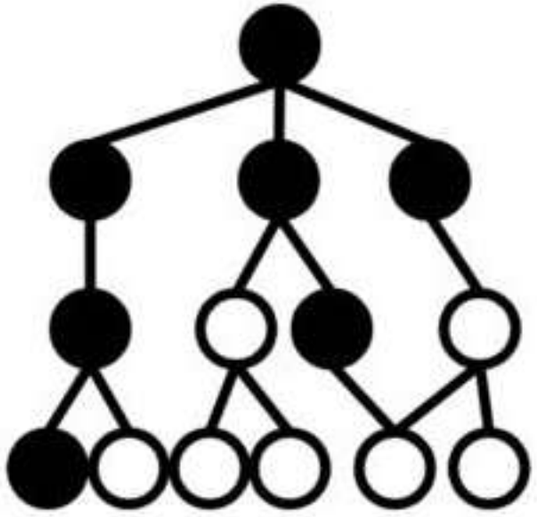
	Micro (%)			Macro (%)		
Method	Precision	Recall	F1	Precision	Recall	F1
BioLarK	78.5	60.5	68.3	76.6	66.0	70.9
cTAKES	72.2	55.6	62.8	74.0	61.4	67.1
OBO	78.3	53.7	63.7	79.5	58.6	67.5
NCBO	**81.6**	44.0	57.2	79.5	48.7	60.4
NCR	80.3	62.4	**70.2**	**80.5**	68.2	**73.9**
NCR -H	74.4	61.5	67.3	72.2	67.1	69.6
NCR -N	78.1	**62.5**	69.4	76.6	**68.3**	72.2
NCR -HN	77.1	57.2	65.7	76.5	63.4	69.3

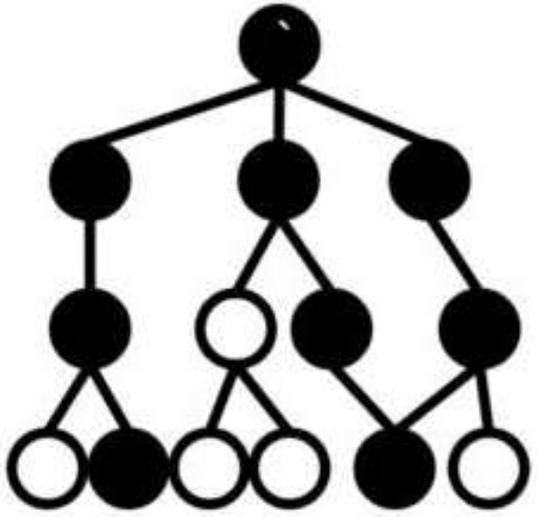
How better to evaluate accuracy



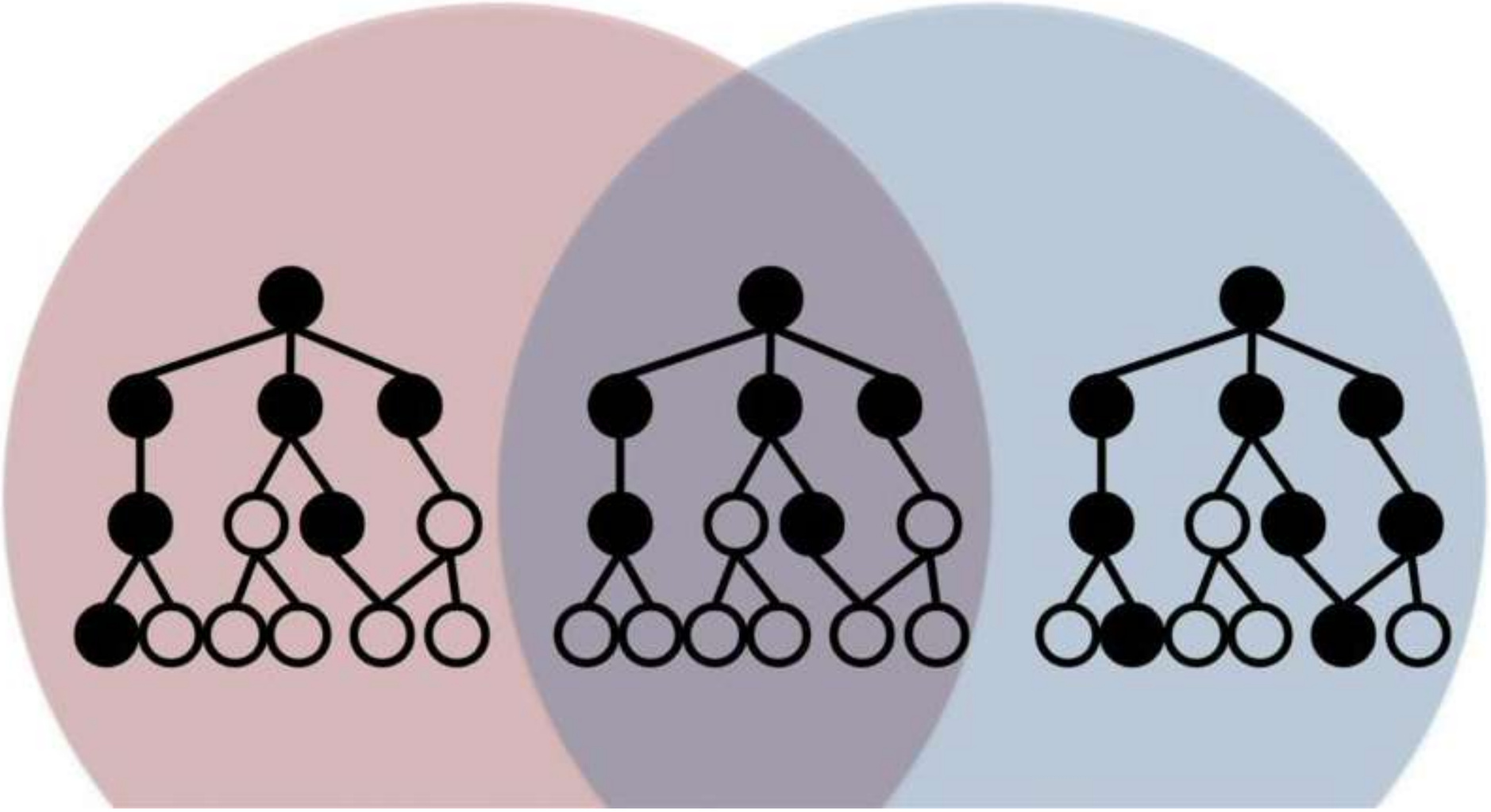


How better to evaluate accuracy





How better to evaluate accuracy



Results on annotated PubMed abstracts (Cont.)

	Extended (%)			Jaccard (%)
Method	Precision	Recall	F1	
BioLarK	91.5	80.8	85.8	76.9
cTAKES	95.6	73.9	83.3	72.1
OBO	92.4	77.9	84.5	74.4
NCBO	**95.8**	65.4	77.7	64.3
NCR	93.3	82.1	**87.3**	**79.1**
NCR -H	86.5	**83.8**	85.1	76.7
NCR -N	90.6	83.1	86.7	78.2
NCR -HN	89.7	78.9	83.9	73.2

Results on UDP Clinical reports

	Extended (%)			Jaccard (%)
Method	Precision	Recall	F1	
BioLarK	58.9	42.6	49.5	29.5
cTAKES	68.5	36.7	47.8	27.3
OBO	59.2	46.4	52.0	31.3
NCBO	**69.8**	37.2	48.5	27.2
NCR	57.1	49.4	**53.0**	**31.5**
NCR -H	54.0	49.4	51.6	30.5
NCR -N	54.7	**50.5**	52.5	31.4
NCR -HN	56.5	49.0	52.5	31.3

Results on UDP Clinical reports

	Extended (%)			Jaccard (%)
Method	Precision	Recall	F1	
BioLarK	58.9	42.6	49.5	29.5
cTAKES	68.5	36.7	47.8	27.3
OBO	59.2	46.4	52.0	31.3
NCBO	**69.8**	37.2	48.5	27.2
NCR	57.1	49.4	**53.0**	**31.5**
NCR -H	54.0	49.4	51.6	30.5
NCR -N	54.7	**50.5**	52.5	31.4
NCR -HN	56.5	49.0	52.5	31.3

Expert analysis of false positives for 3 UDP cases

- 73 unique false positives
- 47.9% correctly added more information
- 8.2% reported a more specific concept
- 16.4% were mentioned as negations
- 6.8% were reported but not confidently diagnosed
- 20.7% “true” false positives

Conclusion

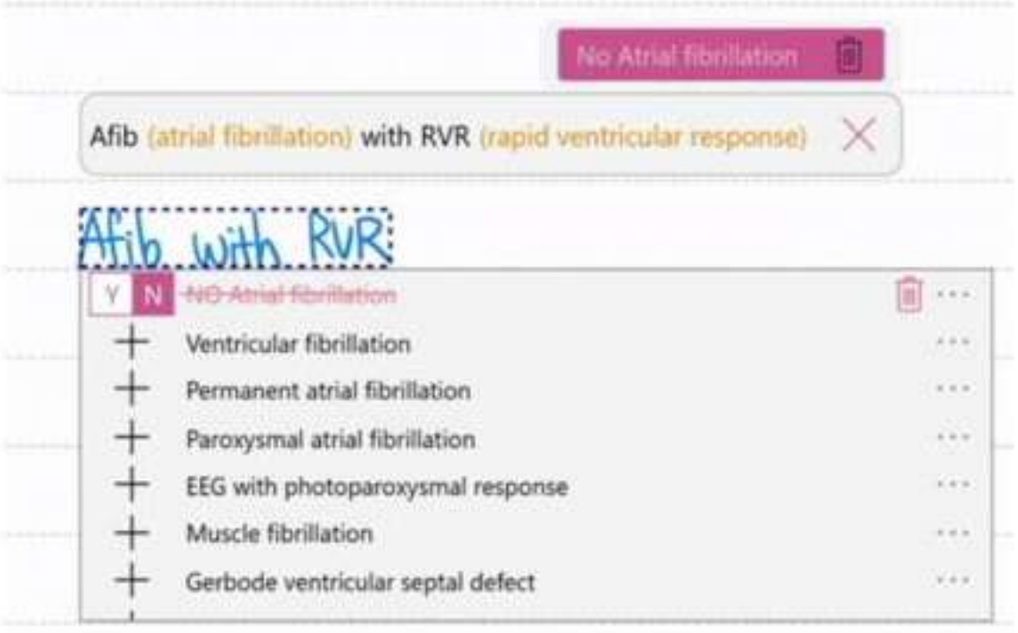
- We introduced a new method for ontology concept matching
- We used a novel way for sharing information between concepts based on their taxonomy
- By convolving our concept matching model we can mine concepts from longer text
- In future work we will use NCR for other knowledge extraction tasks such as relation extraction

Current Use & Availability

- PhenoTips (integrated)
- PhenoLines (Glueck et al 2018)
  - Used to identify HPO terms in clinical records for subsequent visualization
- Phenopad.ai
  - Modern tablet-based UI for clinical note taking
- Foundation29
  - Global Commission to End Diagnostic Odyssey for RD (Microsoft)
- https://ncr.ccm.sickkids.ca
  - BSD License, code on Github

Abbreviation Disambiguation with Global Context & Medical Knowledge

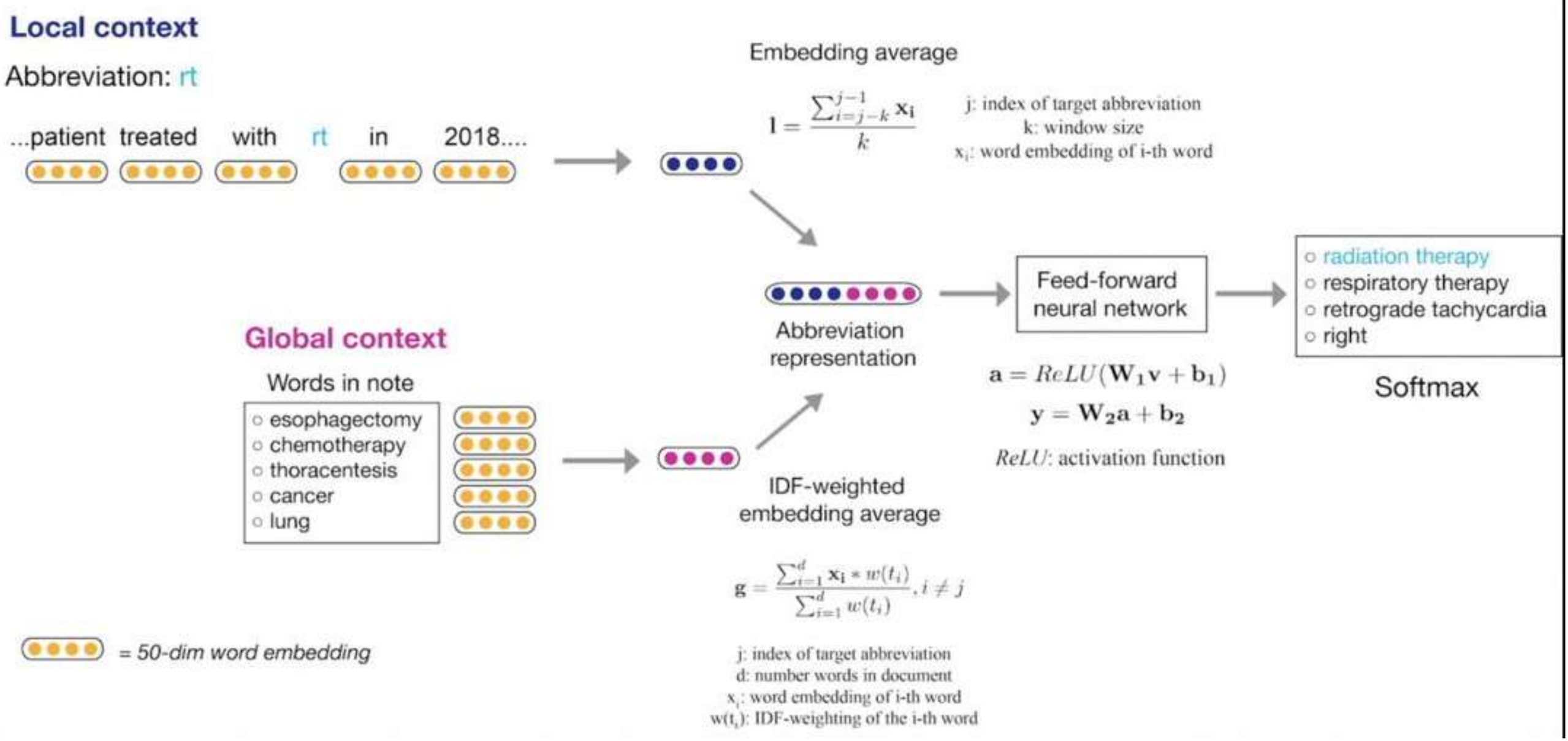
Abbreviation Disambiguation with Global Context & Medical Knowledge



Local v Global context

- Huang et. al showed that word sense disambiguation (WSD) can be improved by representing words jointly by their local and global contexts
- Local context: all words directly surrounding the target word within a window of size *k*
- Global context keywords in a note can identify the relevant domain and help disambiguation
  - words with the highest IDF weights within a collection of documents
  - Ex: *ca* is more likely to expand to *cancer* and not *cardiac arrest* if note also contains words such as tumour, benign, or chemotherapy
- Better abbreviation representations can reduce the amount of training data needed and improve the generalizability of our models

Model overview



Results: Comparison of best 10-fold Cross-Validation scores on 50 abbreviations from CASI with other publications

- We compared the performance of our model to models trained on the same 50 abbreviations
- Baseline: SVM using bag-of-word features (BoW)

Model	**Accuracy (%)**
Naive-Bayes [1]	93.72
SVM [1]	93.85
BSC-WSD [1]	94.55
CNN [2]	95.14
SVM baseline [2]	92.27
Our SVM baseline	92.41
Our best model	**97.17**

[1] S. Moon et al., 2013.
[2] V. Joopudi et al., 2018.

Results: Comparison of best 10-fold Cross-Validation scores on 50 abbreviations from CASI with other publications

- We compared the performance of our model to models trained on the same 50 abbreviations
- Baseline: SVM using bag-of-word features (BoW)

Model	Accuracy (%)
Naive-Bayes [1]	93.72
SVM [1]	93.85
BSC-WSD [1]	94.55
CNN [2]	95.14
SVM baseline [2]	92.27
Our SVM baseline	92.41
Our best model	**97.17**

[1] S. Moon et al., 2013.
[2] V. Joopudi et al., 2018.

Results: Comparison of best 10-fold Cross-Validation scores on 50 abbreviations from CASI with other publications

- We compared the performance of our model to models trained on the same 50 abbreviations
- Baseline: SVM using bag-of-word features (BoW)

Model	**Accuracy (%)**
Naive-Bayes [1]	93.72
SVM [1]	93.85
BSC-WSD [1]	94.55
CNN [2]	95.14
SVM baseline [2]	92.27
Our SVM baseline	92.41
Our best model	**97.17**

[1] S. Moon et al., 2013.
[2] V. Joopudi et al., 2018.

Are we done?

- Supervised models with hand-labelled abbreviations perform well, but they cannot generalize to abbreviations from other sources and obtaining data is expensive
  - CASI has 67 acronyms. AllAcronyms has ~80,000 medical abbreviations
- Reverse substitution (RS) to eliminate hand-labelling can auto-generate training data
  - E.g. "Patient was given intravenous fluid" becomes "Patient was given ivf", and the label for the abbreviation is "intravenous fluid".
- Problem: RS generates imbalanced datasets
  - E.g. In MIMIC-III, “intravenous fluid” occurs 2,503 while “in vitro fertilization” does not occur at all.

Datasets

(1) MIMIC-III: large corpus of clinical notes from MIT used from training word phrase embeddings

(2) AllAcronyms: crowd-sourced database with 80k+ medical abbreviations
(3) UMLS: ontology of medical concepts

(4) CASI: clinical notes with 67 hand-labelled abbreviations (~500 examples for each) from the University of Minnesota; used as test set

Using related medical concepts to reduce false priors

Train FastText word embedding model on MIMIC-III

pros of FastText: can impute embeddings of missing words! Useful for concepts not in training corpus

Augment training set by sampling related medical concepts (with replacement)

Compute relatedness by taking Euclidean distance of FastText word embeddings between target concept and all medical concepts in UMLS & AllAcronyms

Collect training samples from MIMIC-III using RS

Train CNN to perform classification task

Method comparison

Data sampling technique	MIMIC TEST		CASI TEST	
	Micro Accuracy	Macro Accuracy	Micro Accuracy	Macro Accuracy
M1: control	0.906	**0.868**	0.622	0.624
M2: sampling with replacement	**0.906**	0.856	0.644	0.646
M3: + relatives	0.891	0.842	**0.727***	**0.726**

* $p < 0.01$, Wilcoxon signed rank test

Summary

AI-enabled Note Taking

AI-enabled Note Taking

- Use technology to assist patient interaction
- Capture everything: notes, speech, video

AI-enabled Note Taking

- Use technology to assist patient interaction
- Capture everything: notes, speech, video
- Generate recordings that are easy to browse

AI-enabled Note Taking

- Use technology to assist patient interaction
- Capture everything: notes, speech, video
- Generate recordings that are easy to browse
- Locate important information automatically

AI-enabled Note Taking

- Use technology to assist patient interaction
- Capture everything: notes, speech, video
- Generate recordings that are easy to browse
- Locate important information automatically
- Clinical decision support on the fly

Electronic Health **Systems**

An Electronic Health System should be an active participant in the care of the patient. It should

- Record clinical data (non-invasively)
- Help clinician navigate previous notes and laboratory tests
- Structure data for computation
- Be built with data sharing in mind

Acknowledgements

Jixuan Wang
Jingbo Yang
Helen Lu
Haochi Zhang

Marta Skreta
Aryan Arbabi
Jixuan Wang
Brittney Johnstone

Aryan Arbabi
David Adams (NIH)
Sanja Fidler

GenomeCanada

NSERC CRSNG

CARE forRARE

CanSHARE

SickKids®

Method comparison

Data sampling technique	MIMIC TEST		CASI TEST	
	Micro Accuracy	Macro Accuracy	Micro Accuracy	Macro Accuracy
M1: control	0.906	**0.868**	0.622	0.624
M2: sampling with replacement	**0.906**	0.856	0.644	0.646
M3: + relatives	0.891	0.842	**0.727***	**0.726**

* $p < 0.01$, Wilcoxon signed rank test