Health Care Breakout Session 5:
De-Identifying Patient Notes with RNNs

Matt Engelhard

Today

- “De-identification of patient notes with recurrent neural networks” (Dernoncourt et al., 2016)
- Brief survey of NLP in medicine
- Revisiting discussion & week in review

De-identification of patient notes with recurrent neural networks

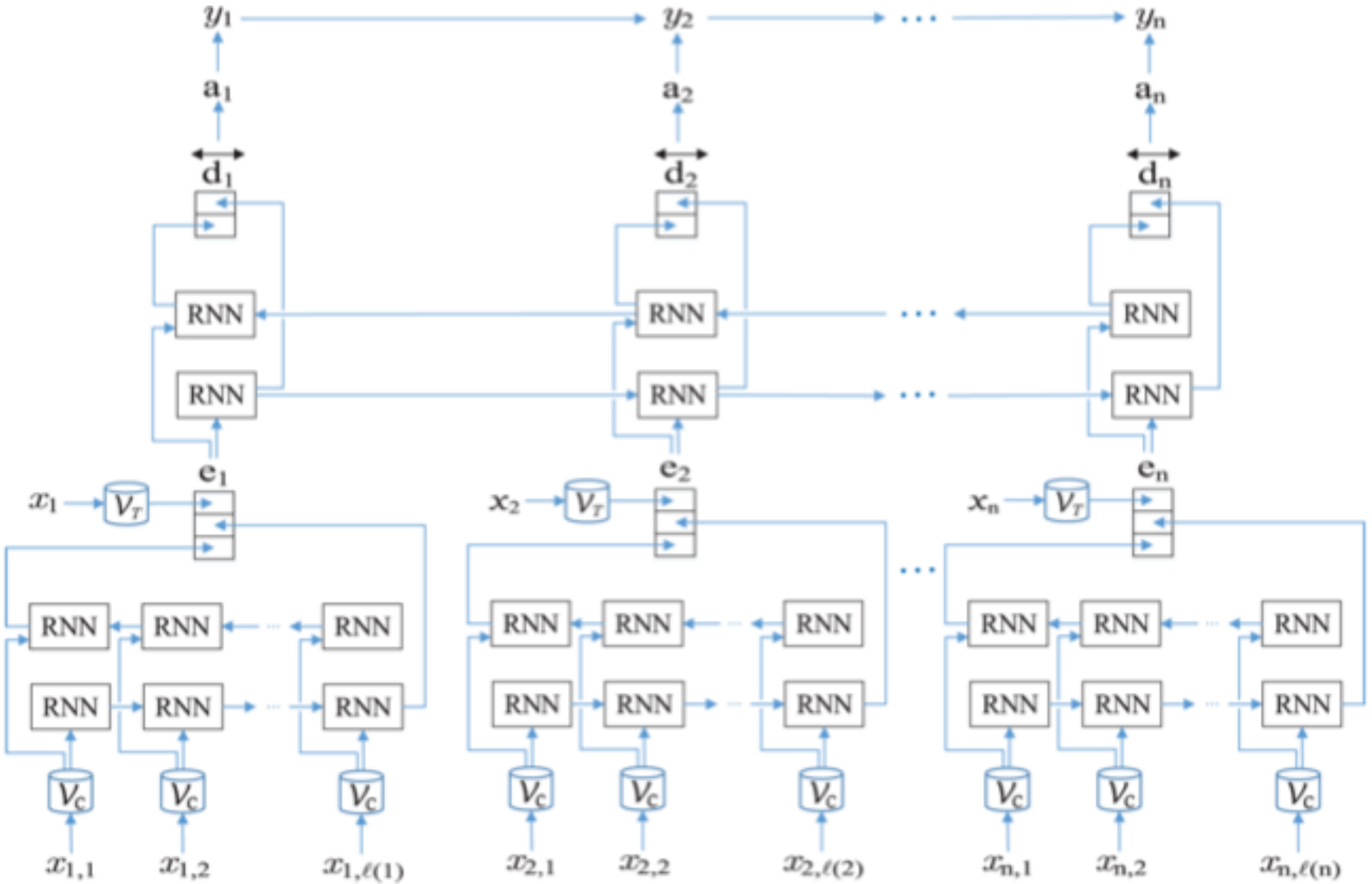
Dernoncourt F, Lee JY, Uzuner O, Szolovits P

JAMIA 24(3), 2017, 596–606

OVERVIEW:

- A bidirectional RNN is used to identify PHI (18 HIPAA fields) in patient notes
- Datasets:
  1. i2b2 2014 de-identification dataset (889 discharge summaries)
  2. MIMIC de-identification dataset (1635 discharge summaries) created for this work
  – 28,867 and 60,725 PHI instances, respectively
- State of the art recall and F1 metric (but not precision) on i2b2; and state of the art precision, recall, and F1 on MIMIC

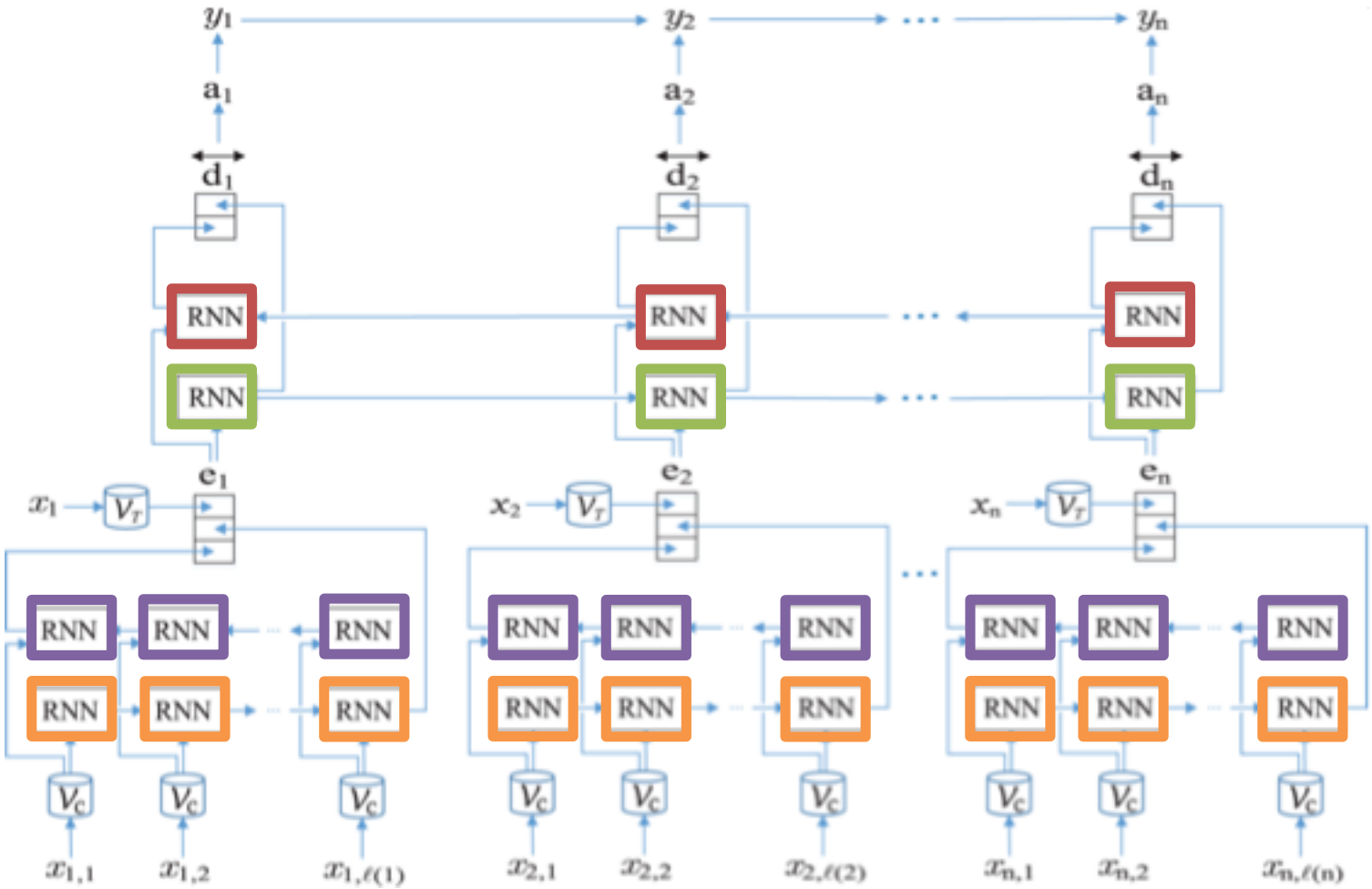
RNN MODEL



From: De-identification of patient notes with recurrent neural networks

J Am Med Inform Assoc. 2016;24(3):596-606. doi:10.1093/jamia/ocw156

Q: In Figure 1, all RNN blocks appear to be identical. While all blocks do use the same architecture – the LSTM specified on page 599 – not all have the same weights. How many distinct LSTMs are used in this model? How many parameters are found in each LSTM?



backward token LSTM

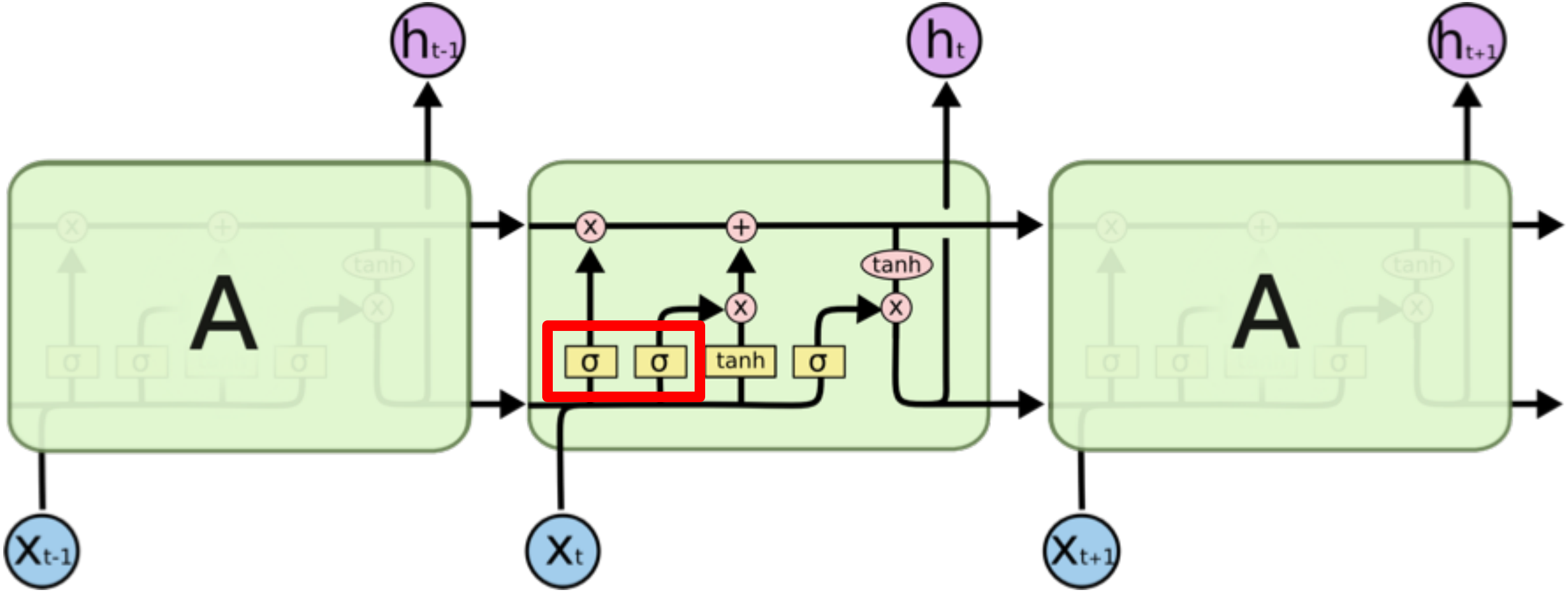
forward token LSTM

backward character LSTM

forward character LSTM

Inside an LSTM

http://colah.github.io/posts/2015-08-Understanding-LSTMs/



Neural Network Layer

Pointwise Operation

Vector Transfer

Concatenate

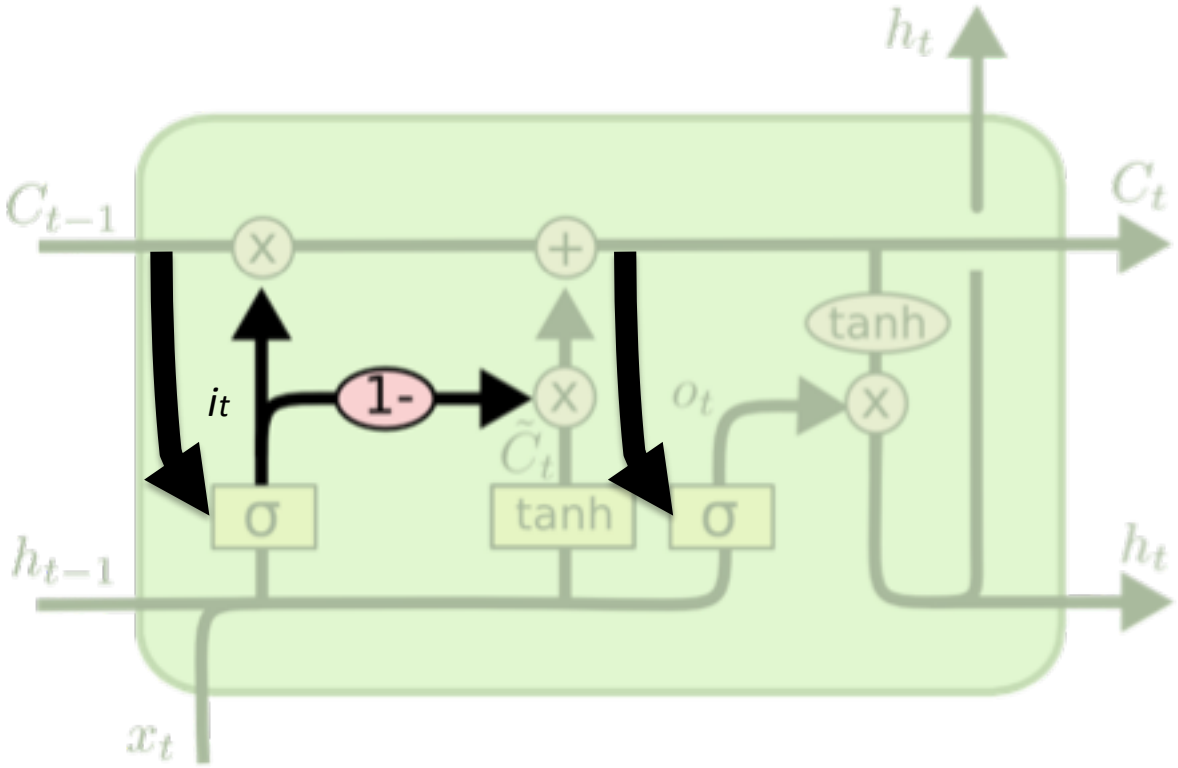
Copy

The repeating module in an LSTM contains four* interacting layers.

*The common layout shown above has four neural network layers, but the LSTM in this paper couples the first two sigmoid layers.

LSTM Parameters

$$\mathbf{i}_t = \sigma(W_i\,[\mathbf{x}_t;\,\mathbf{h}_{t-1};\,\mathbf{c}_{t-1}] + \mathbf{b}_i)$$
$$\tilde{\mathbf{c}}_t = \tanh(W_c\,[\mathbf{x}_t;\,\mathbf{h}_{t-1}] + \mathbf{b}_c)$$
$$\mathbf{c}_t = (1 - \mathbf{i}_t) \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t$$
$$\mathbf{o}_t = \sigma(W_o\,[\mathbf{x}_t;\,\mathbf{h}_{t-1};\,\mathbf{c}_t] + \mathbf{b}_o)$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)$$



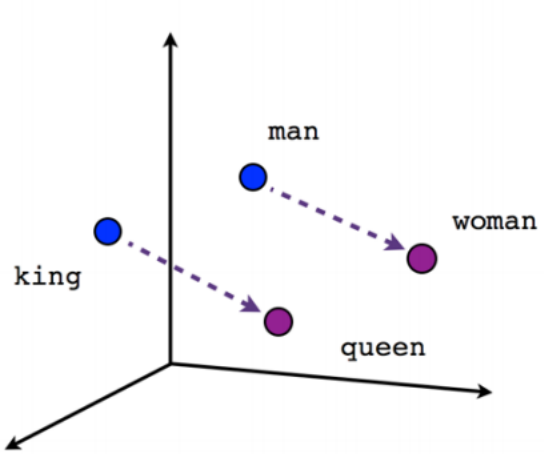
Token / Character Embeddings

Character embedding dimension: **25**

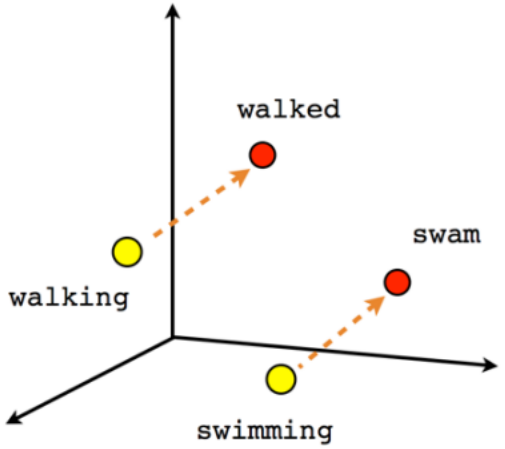
Character-based token-embedding LSTM dimension: **25**

Token embedding dimension: **100**

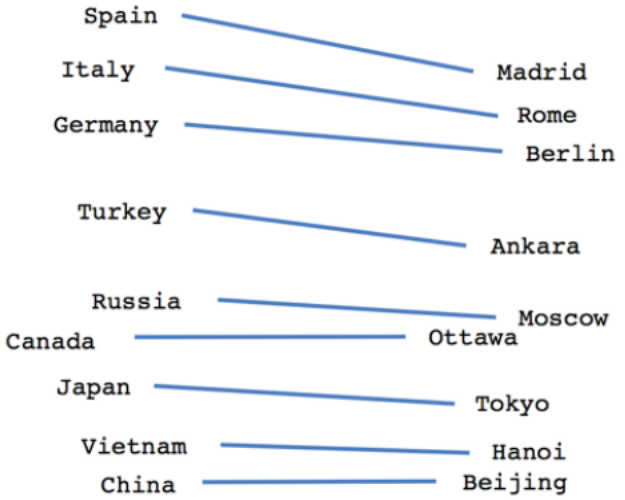
Label prediction LSTM dimension: **100**



Male-Female



Verb tense



Country-Capital

https://www.tensorflow.org/tutorials/word2vec

Explored pre-trained and custom embeddings using both GloVe and word2vec

Final choice: Wikipedia-trained GloVe

Q: This model is more complex than many RNNs because it uses character embeddings in addition to the more common token (i.e. word-level) embeddings. Why are token embeddings sufficient in many other applications, but not in this one?

EVALUATION STRATEGY

Evaluation Metrics

Precision, or positive predictive value:

$$\frac{\text{true positives}}{\text{all positive predictions}}$$

Recall, or sensitivity:

$$\frac{\text{true positives}}{\text{all condition positives}}$$

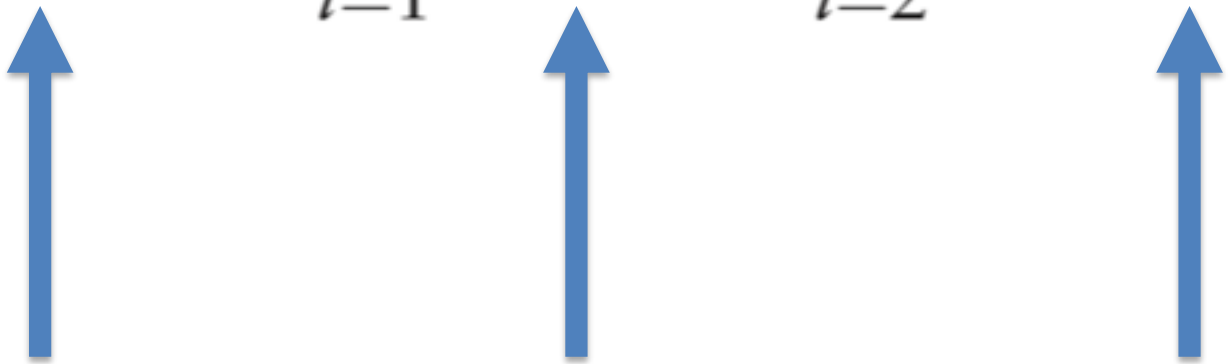
F1-score:

$$\frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

	Condition Positive	**Condition Negative**
Prediction Positive	True Positive	False Positive
Prediction Negative	False Negative	True Negative

Why No ROC Curve?

$$s(y_{1:n}) = \sum_{i=1}^{n} \mathbf{a}_i[y_i] + \sum_{i=2}^{n} T[y_{i-1}, y_i].$$



sequence score

label probabilities

transition probabilities

Train, Validation, Test

MIMIC:

80% train/validation

20% test

i2b2:

60% train/validation

40% test

"All results were computed using the official evaluation script from the i2b2 2014 de-identification challenge."

Table 3. Overview of the i2b2 and MIMIC datasets

Statistics	i2b2	MIMIC
Vocabulary size	46 803	69 525
Number of notes	1304	1635
Number of tokens	984 723	2 945 228
Number of PHI instances	28 867	60 725
Number of PHI tokens	41 355	78 633

RESULTS

An Example

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

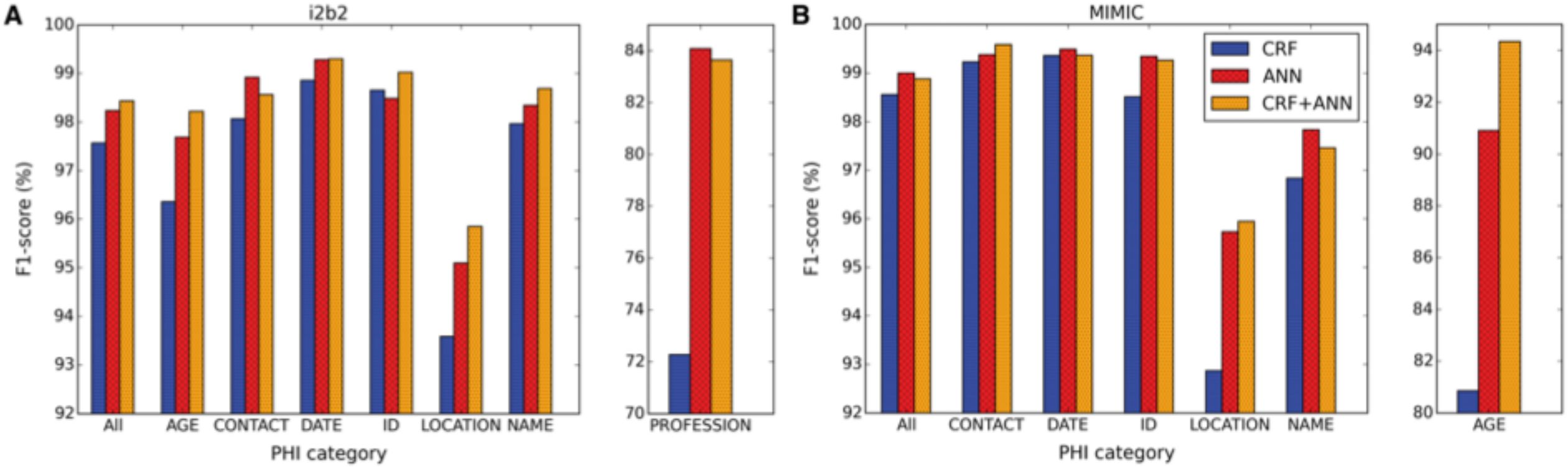
PHI category	ANN
AGE	Father had a stroke at **80** and died of?another stroke at age Personal data and overall health: Now **63**, despite his FH: Father: Died @ **52** from EtOH abuse (unclear exact etiology) Tobacco: smoked from age 7 to **15**, has not smoked since 15.
CONTACT	History of Present Illness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

ANN -> Highest Sensitivity and F1

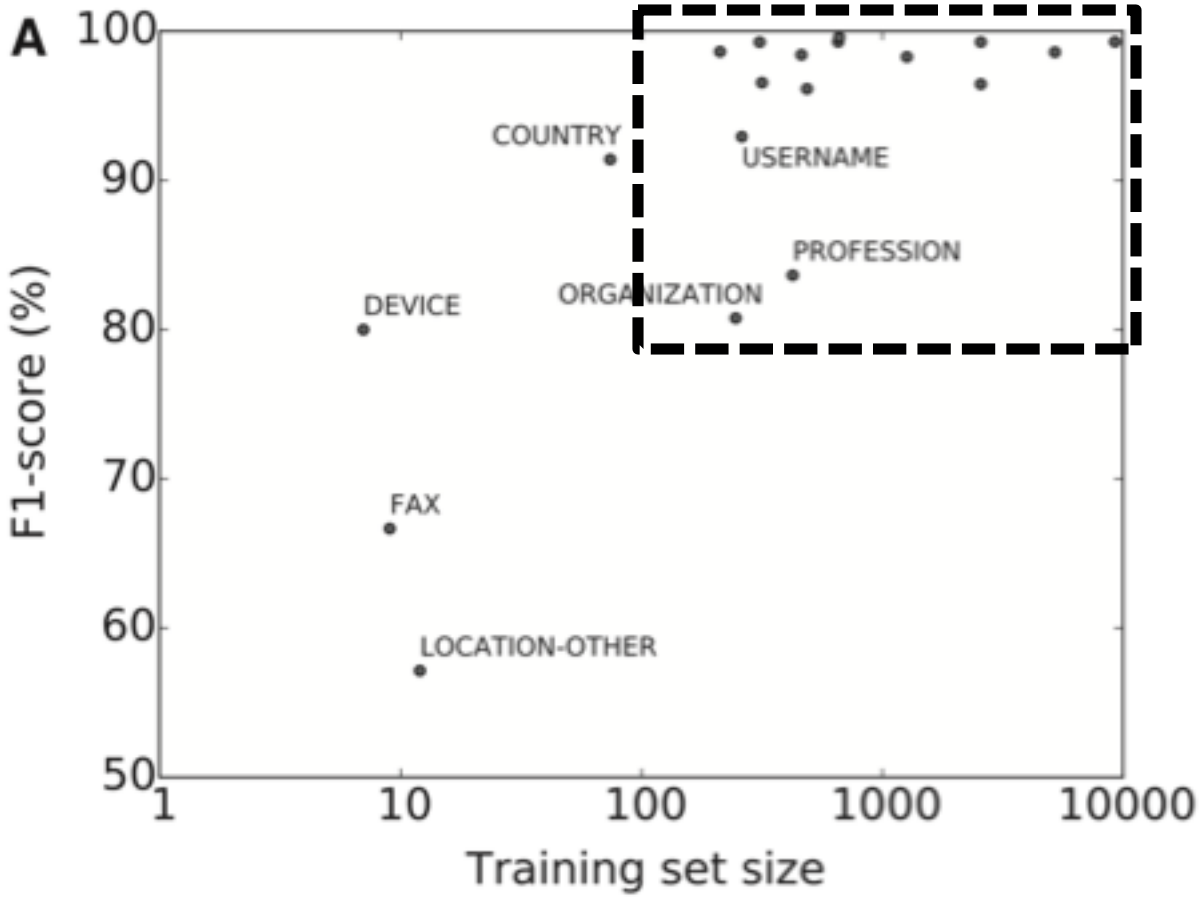
Table 4. Performance (%) on the PHI as defined in HIPAA

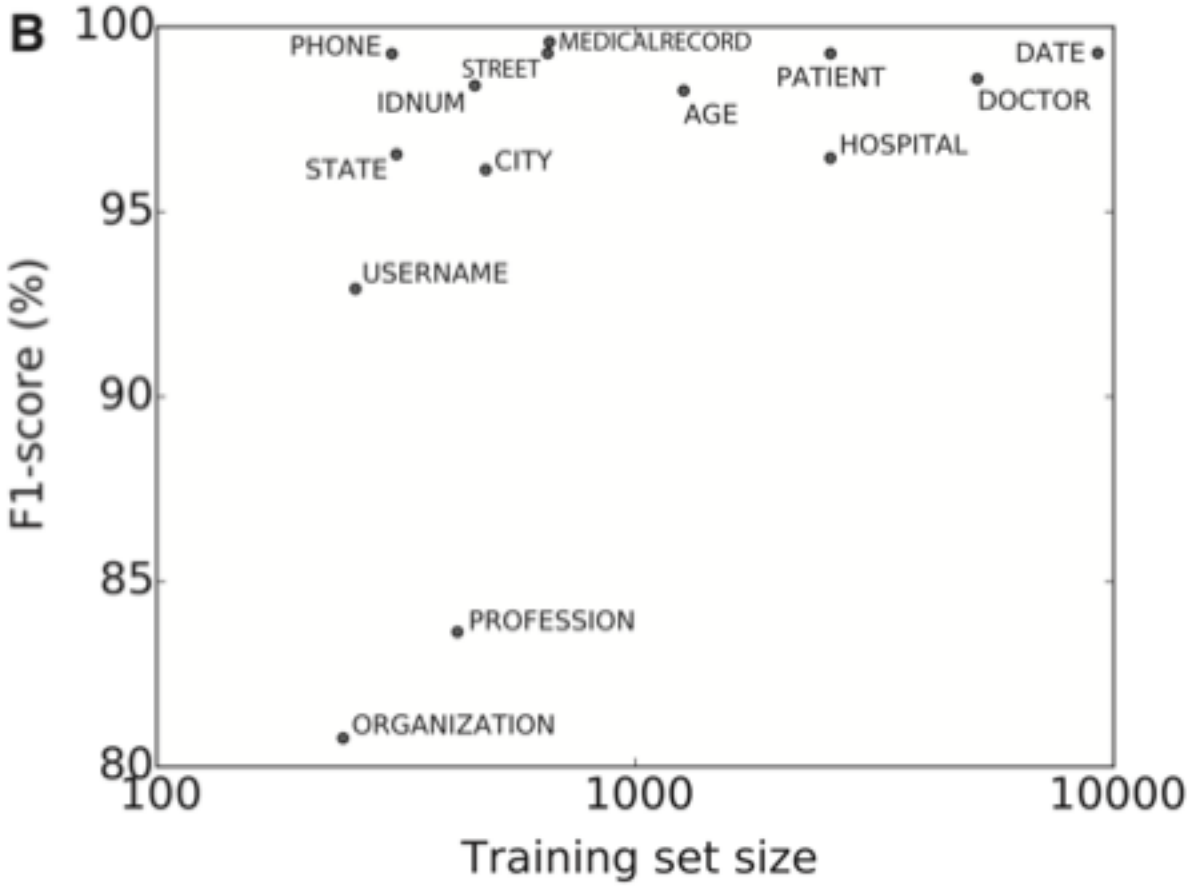
	i2b2			MIMIC		
Model	Precision	Recall	F1	Precision	Recall	F1
Nottingham	**99.000**	96.400	97.680	–	–	–
MIST	91.445	92.745	92.090	95.867	98.346	97.091
CRF	98.560	96.528	97.533	99.060	98.987	99.023
ANN	98.320	97.380	97.848	**99.208**	99.251	**99.229**
CRF + ANN	97.920	**97.835**	**97.877**	98.820	**99.398**	99.108

Not All PHI Categories are Equal

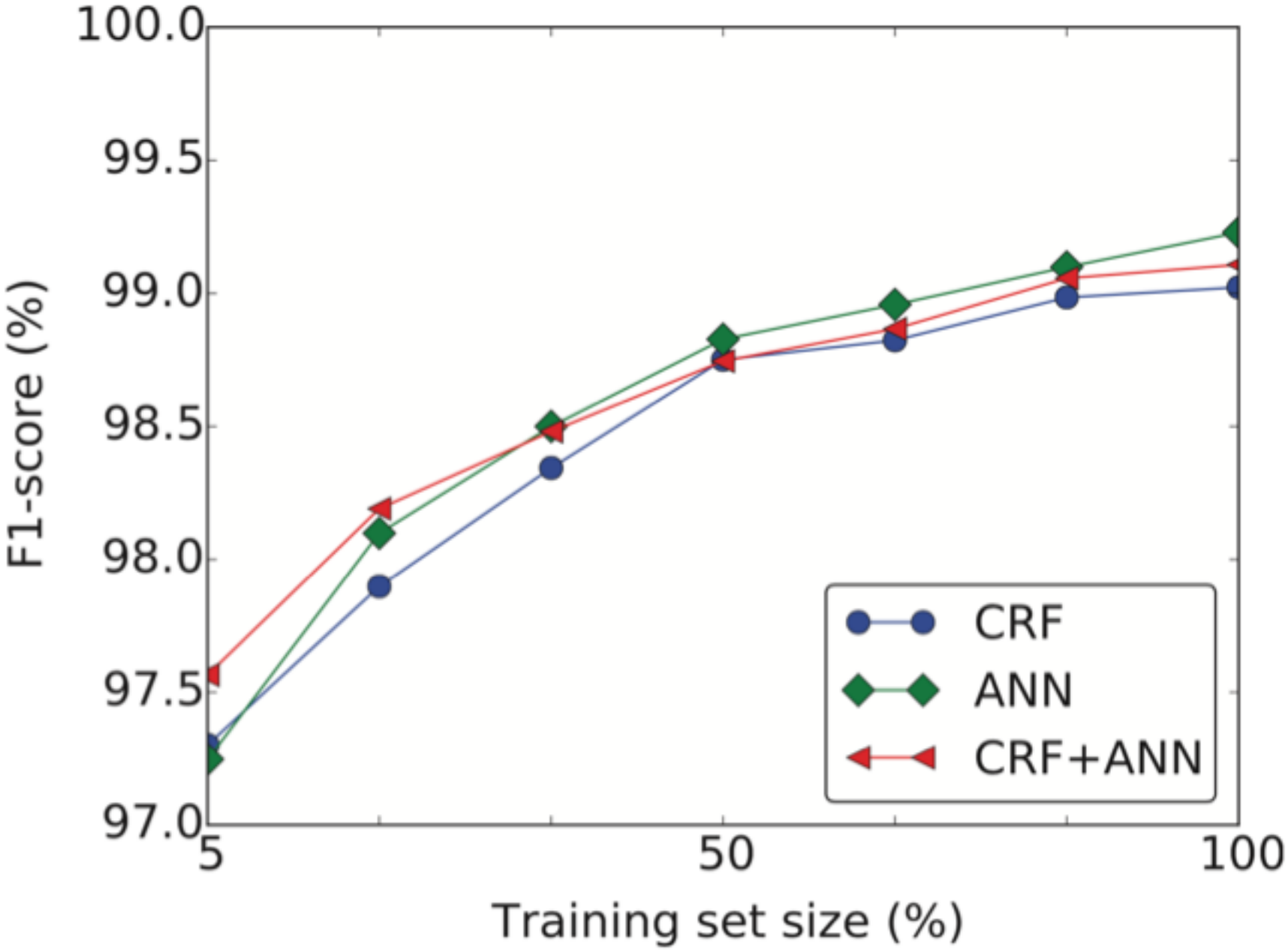


PHI Categories + Dataset Size





Benefit of More Data



Q: The authors emphasize that their model has higher recall (another term for sensitivity) compared to the baseline models. Why is recall/sensitivity particularly important in this application, and how do they combine the CRF and ANN models to improve it?

Q: Should 100% sensitivity be required before a system like this can be used for real de-identification tasks? Is there any guidance here?

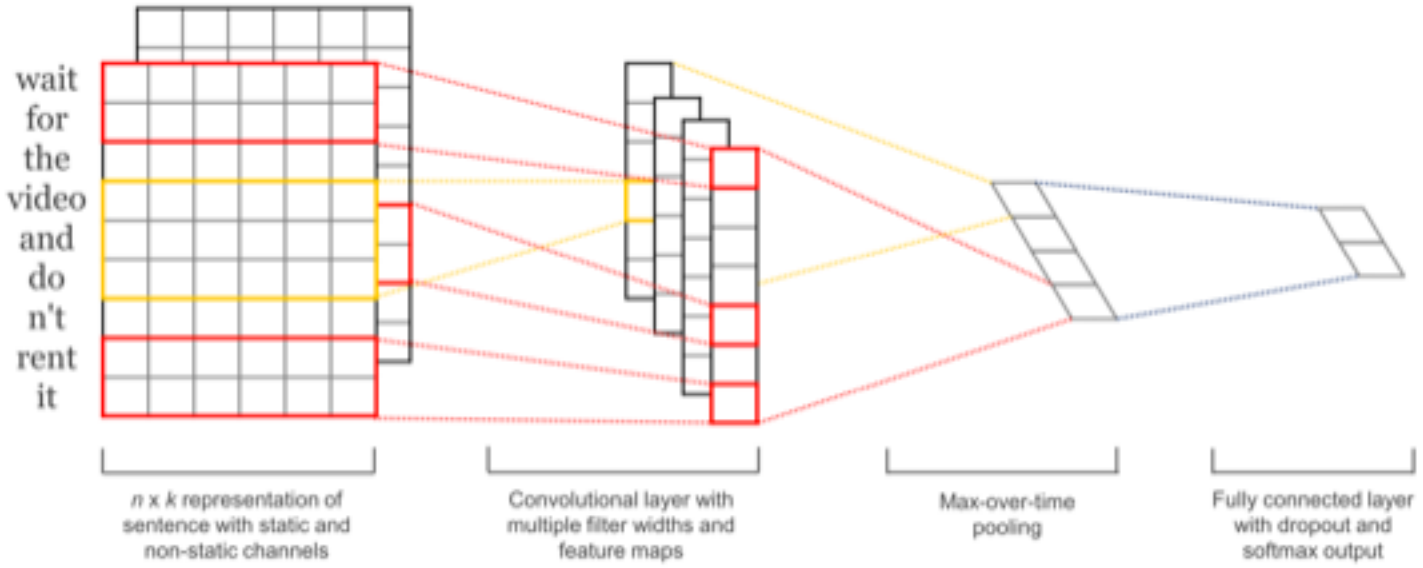
Natural Language Processing

SURVEY OF HEALTH CARE APPLICATIONS

Document Classification

Classification of radiology reports using neural attention models.

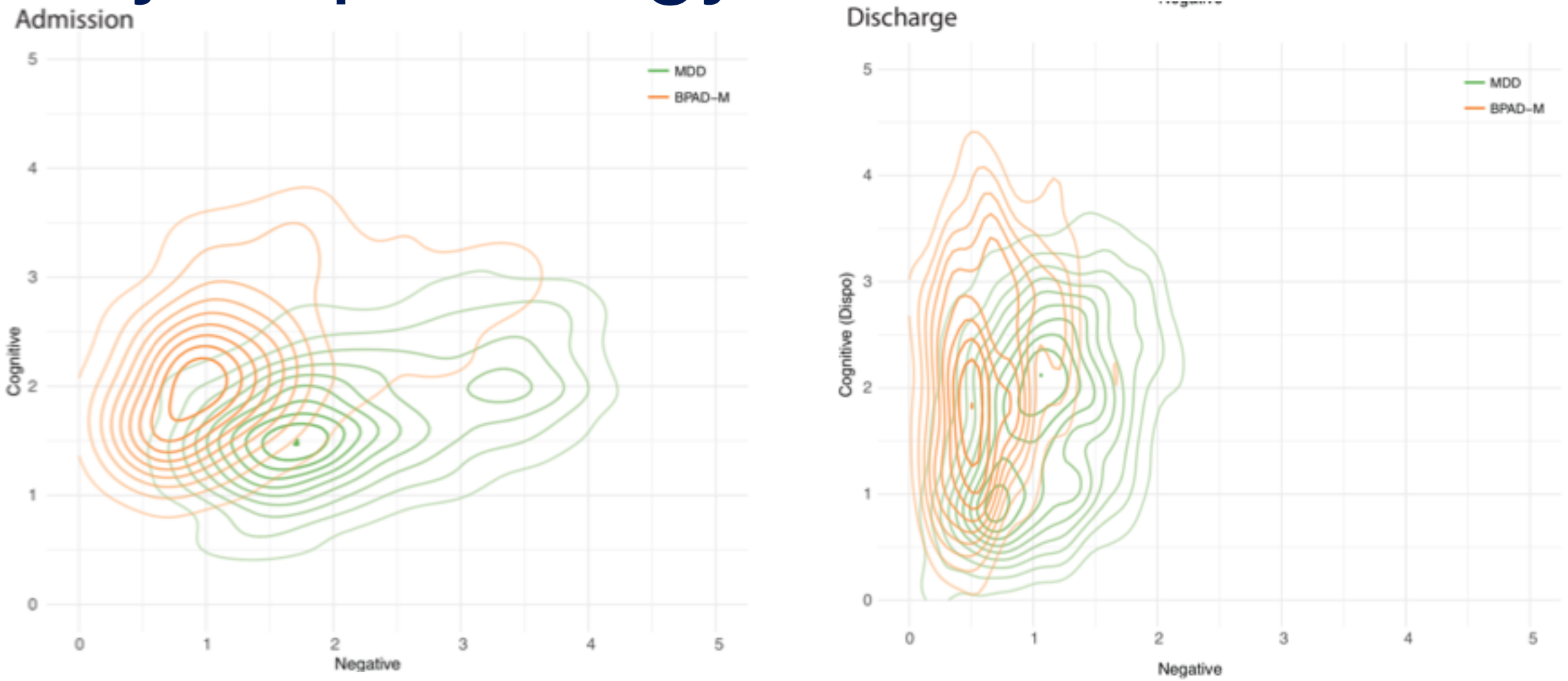
Shin B, Chokshi FH, Lee T, Choi JD.
In Neural Networks IJCNN 2017, pp. 4363-4370. IEEE.



From: http://www.wildml.com

- **CNN** architecture with **word2vec** embeddings
- **Five Prediction Tasks**, all on 0-2 scale:
  - Study severity
  - Acute IC bleed
  - Acute Mass Effect
  - Acute Stroke
  - Acute Hydrocephalus
- Accuracy ranging from **85% - 89.5%** and **superior to inter-annotator agreement on 3 of 5 tasks**
- Also incorporates an attention model, but this is beyond our current scope

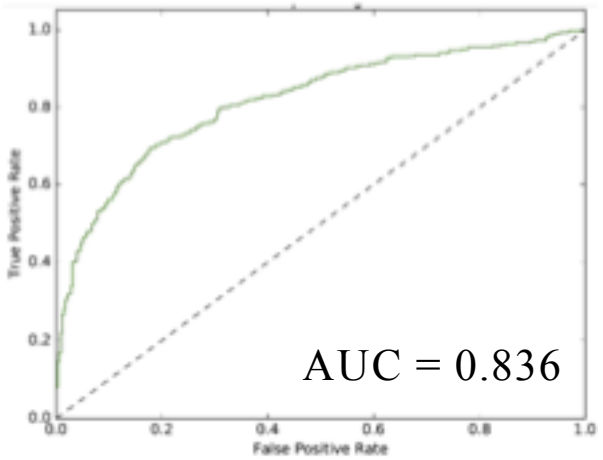
Psychopathology from Clinical Notes



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

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

Mental Health via Social Media



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

r/depression ? r/SuicideWatch

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

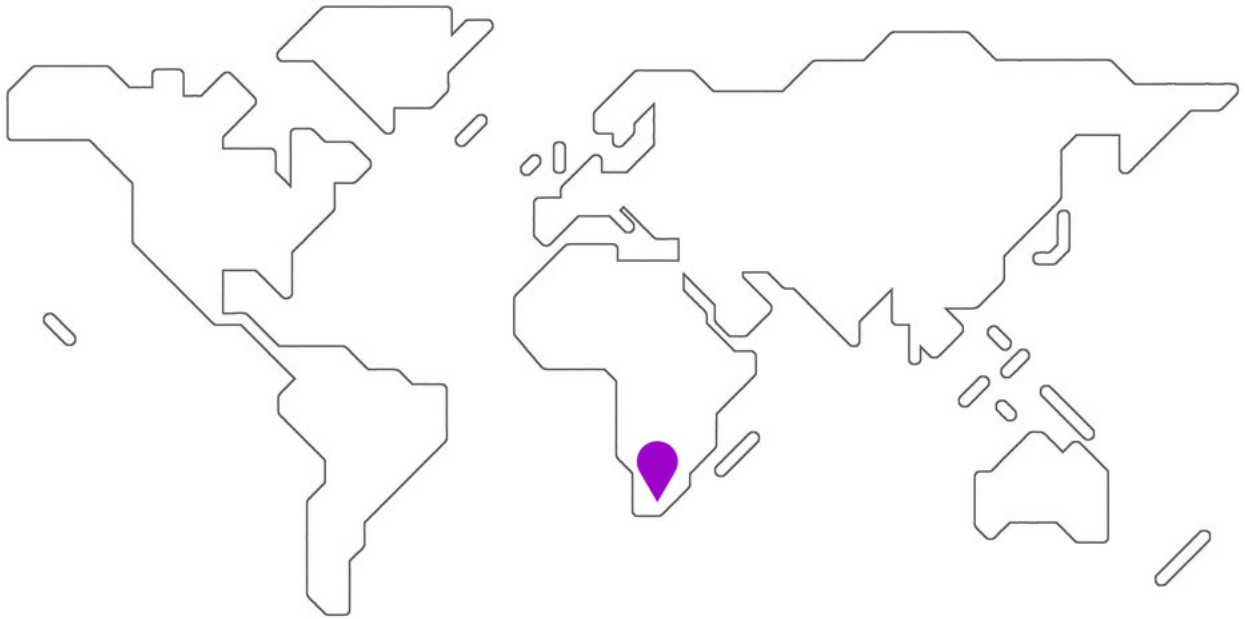
Q & A Games

Is it a person?	*No*
Is it on the shelves?	*No*
Is it on the floor?	*Yes*
Is it a cup?	*No*
Is it blue?	*No*
Is it wood?	*No*
Is it the bed?	*No*
Is it in the lower half of the image?	*Yes*
Is it in the lower left corner?	*No*
Is it near the blue tray with two cups?	*No*
Is it near the green and blue toy?	*No*
Is it near the boy laying on the floor?	*Yes*
Is it something he is touching?	*Yes*
Is it the remote?	*Yes*
Is it the one in his right hand (close to the wooden box)?	*No*

De Vries, Harm, et al. "GuessWhat?! Visual object discovery through multi-modal dialogue." *Proc. of CVPR*. 2017.

Global Maternal Health

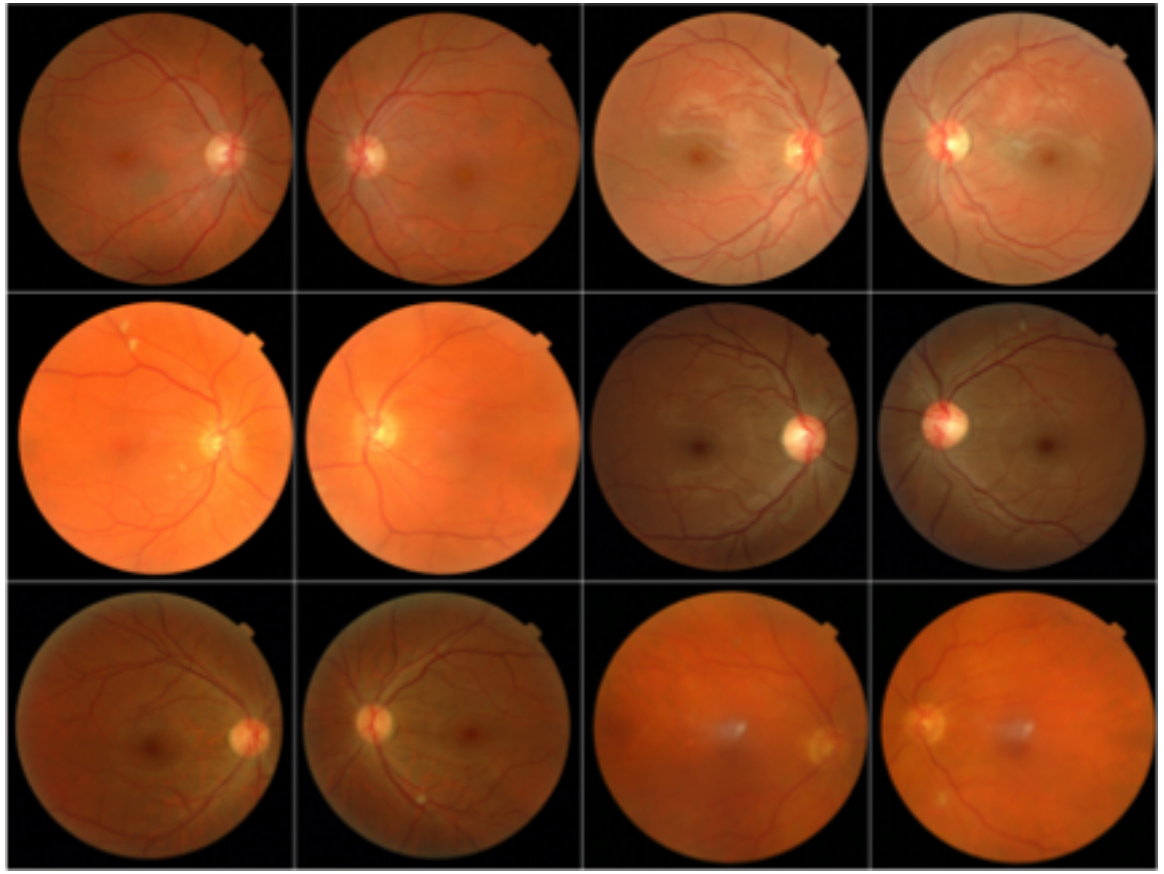
Streamlining and scaling up a national-level maternal health helpdesk with NLP



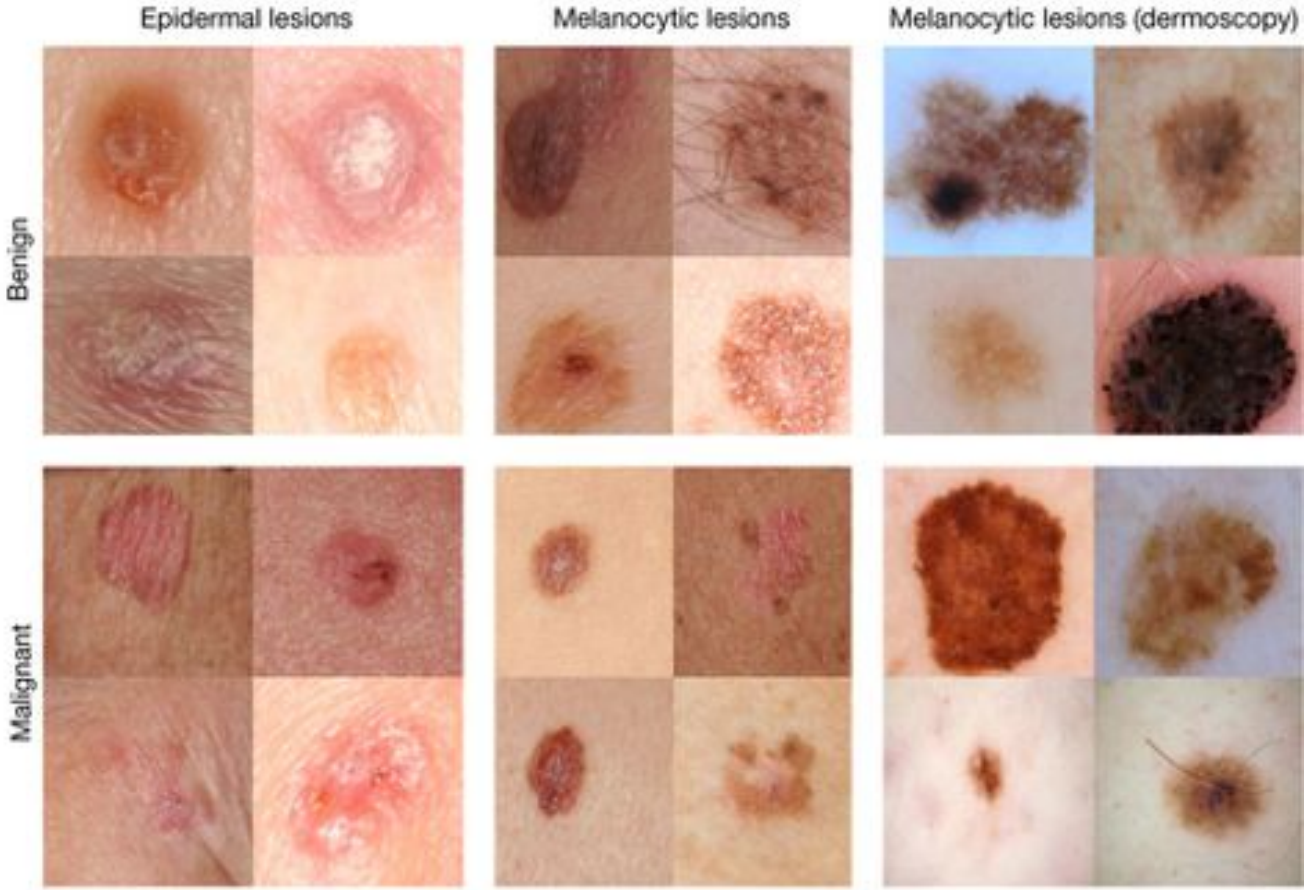
2 m
Registered users
500 k
Messages processed by helpdesk
95%
Clinics in SA participating

WEEK IN REVIEW

Convolutional Neural Networks



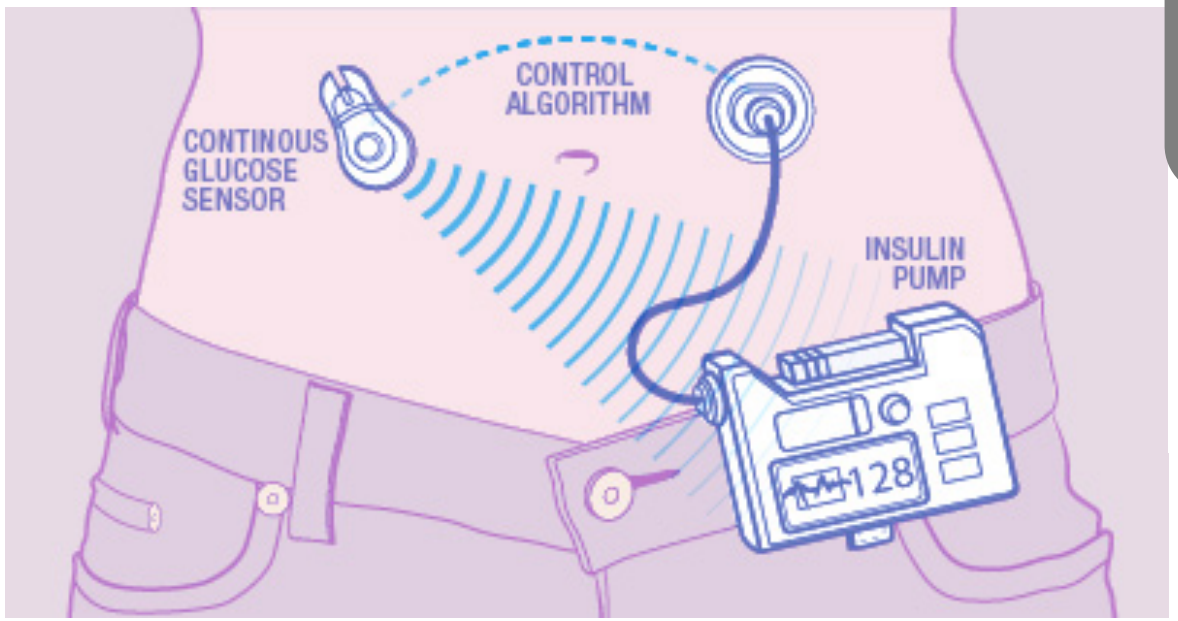
Improved Automated Detection of Diabetic Retinopathy
Invest. Ophthalmol. Vis. Sci.. 2016;57(13):5200-5206. doi:10.1167/iovs.16-19964



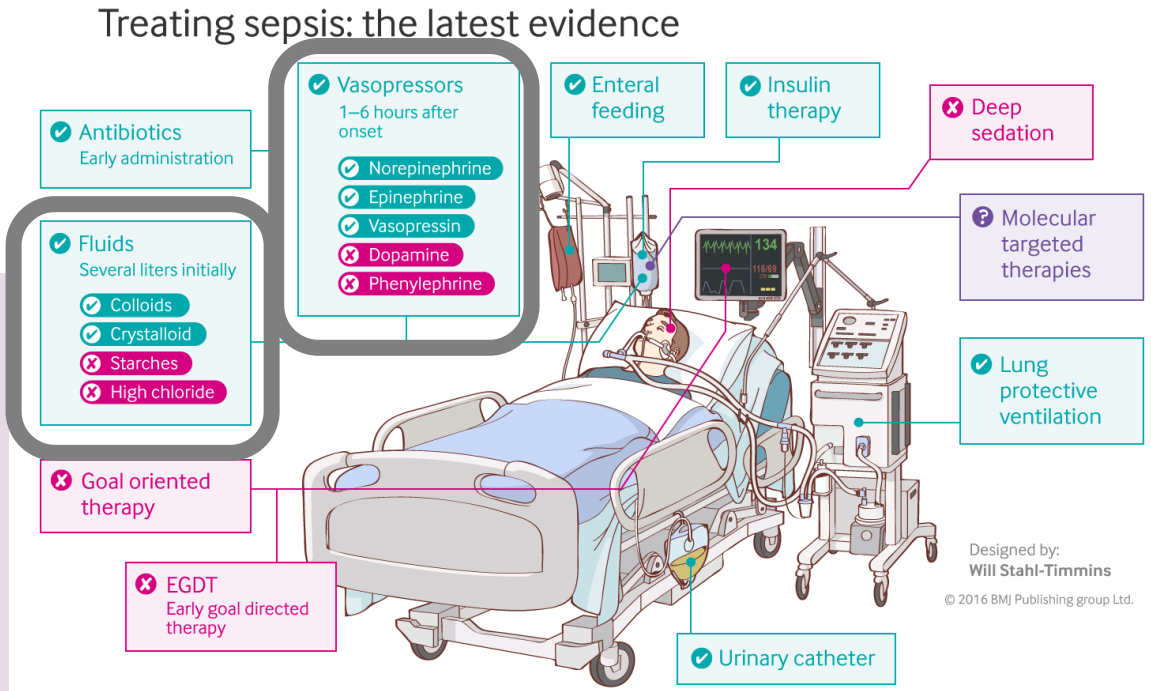
Dermatologist-level classification of skin cancer
Nature volume 542, pages 115–118 (02 February 2017)

Reinforcement Learning

Closed-loop blood glucose control ("artificial pancreas")



https://www.mayo.edu/research/labs/artificial-pancreas/overview



Fluid and vasopressor administration for sepsis treatment

Gotts JE, Matthay MA. Sepsis: pathophysiology and clinical management. bmj. 2016 May 23;353(i1585).

Adversarial Learning

Adversarial Time-to-Event Modeling
Chapfuwa P, Tao C, Li C, Page C, Goldstein B, Carin L, Henao R
ICML 2018

$$p(t|\boldsymbol{x})$$

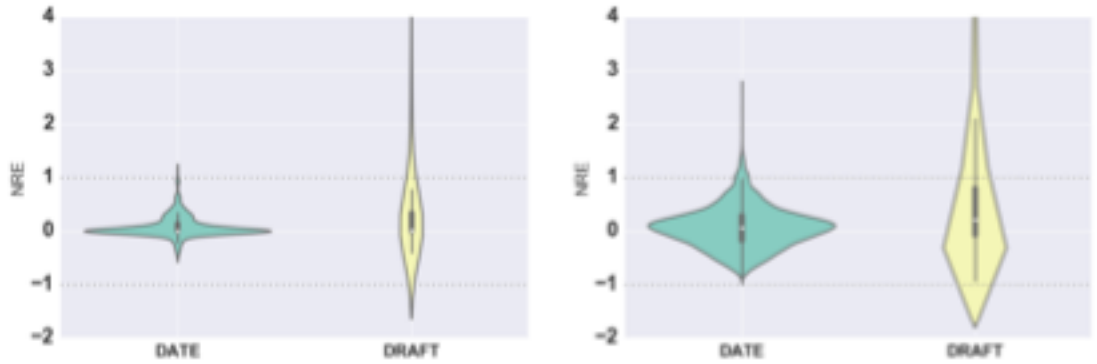
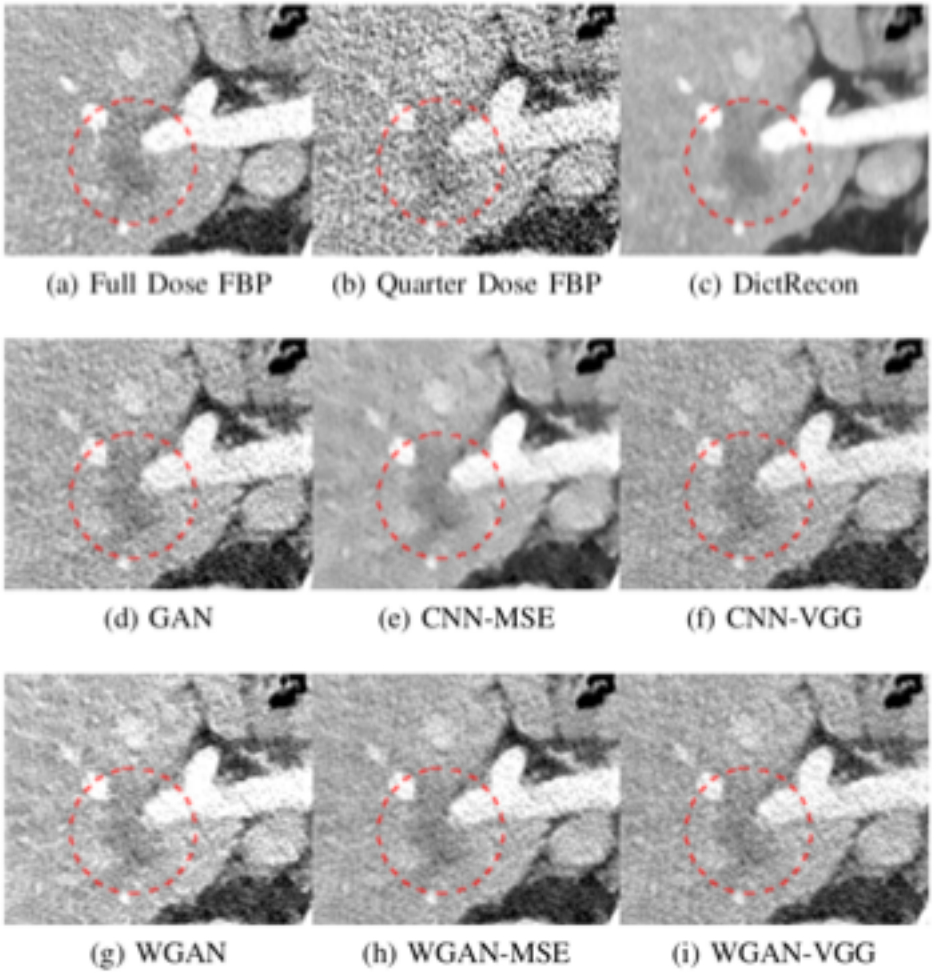


Figure 3. Normalized Relative Error (NRE) distribution for SUPPORT (top) and EHR (bottom), test-set non-censored events. The horizontal dashed lines represent the range of the events, $t_{\max} = 120$ months and $t_{\max} = 365$ days, respectively.



Yang, Qingsong, et al. "Low dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss." *IEEE transactions on medical imaging*(2018).

Natural Language Processing

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

Mass effect from extradural hemorrhage
https://radiopaedia.org

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

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

De-identification of patient notes with recurrent neural networks
JAMIA 24(3), 2017, 596–606

Q: Which of these areas impact the practice of medicine most? Which will impact practice first?

Q: If our goal is prediction and we rigorously evaluate predictive performance, is the “fishing expedition” still a concern?

Traditional Hypothesis:
measure M predicts outcome O

ML Hypothesis:
data D predicts outcome O

THANK YOU!

Questions or ideas? Please contact me at m.engelhard@duke.edu