Isabel Metzger, Data Scientist

Data & Analytics

Leadership Series

About me

I have a thesis master's in deep learning from NYU. I graduated and deployed this work
Most of my research has focused on deep learning architectures for clinical text classification and adverse drug event detection from medical notes, social media text mining, e.g., dental affordability.

1 What is AI and NLP?

2 My journey

3 Designing an awesome thesis

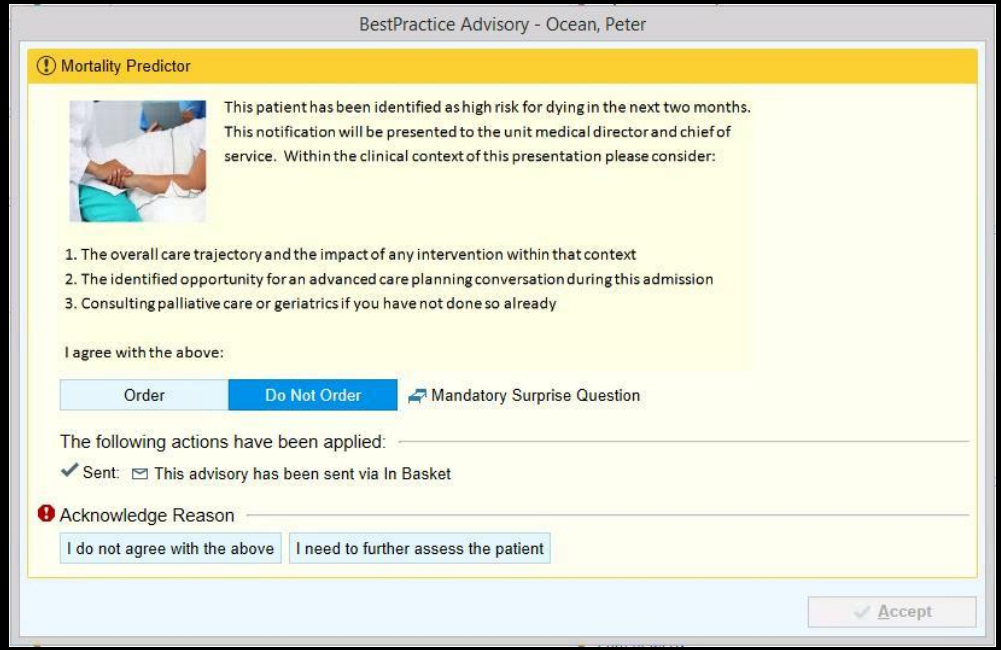
4 Building a great portfolio

5 Lessons Learned

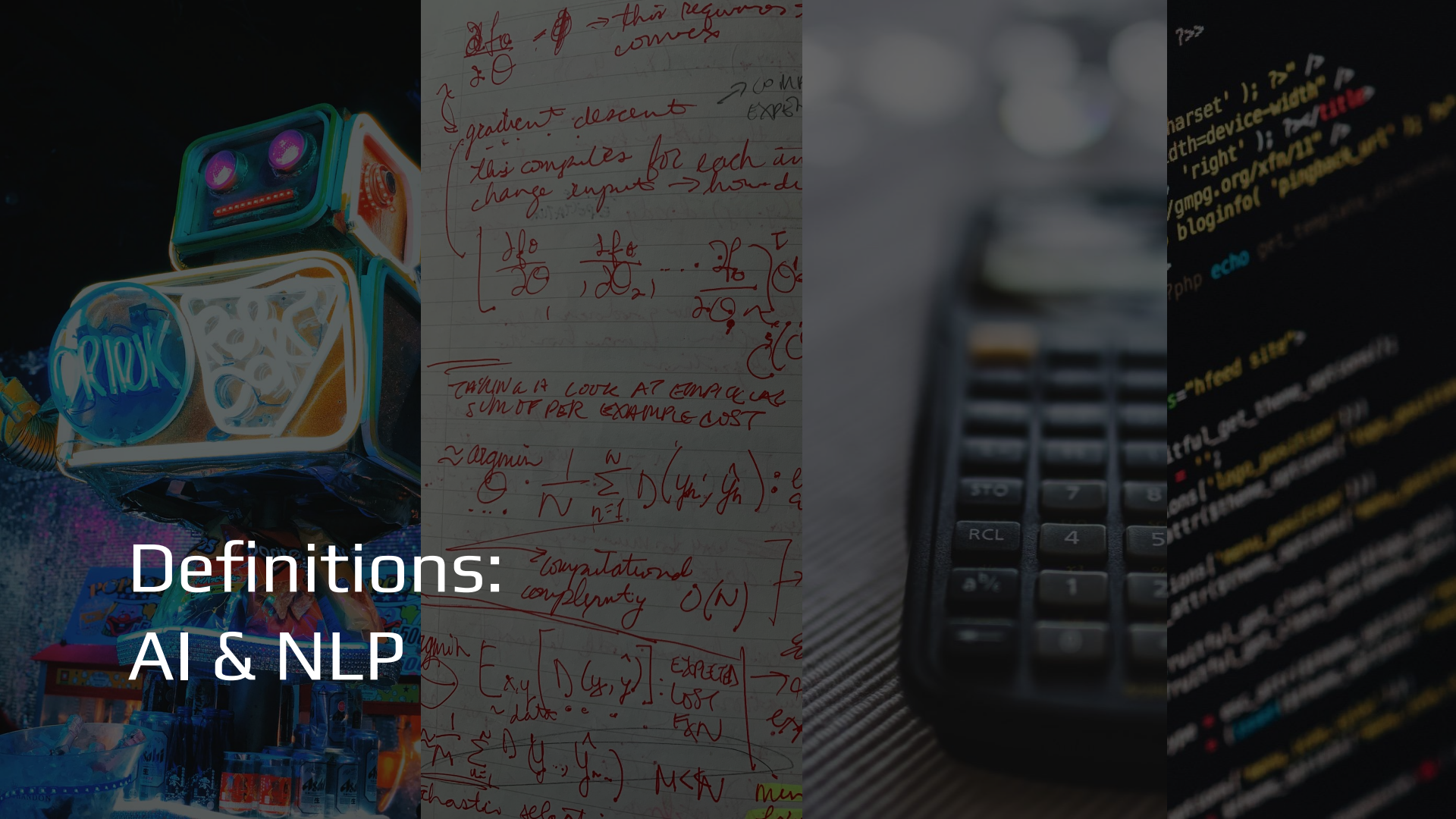
6 Q&A

the alert

- Presented to the attending of record
- Interruptive (cannot click away)
  - ***One chance to further assess***
- Prompts a MSQ (mandatory surprise question) order and suggests an ACP conversation and consulting Palliative Care/Geriatrics



Research shows a general under-estimation. In fact, a recent meta-analysis of MSQ answers reported very poor pooled precision results at ~ 0.37, where studies of non-cancer patients reported much worse discrimination (Downar et al. 2017)



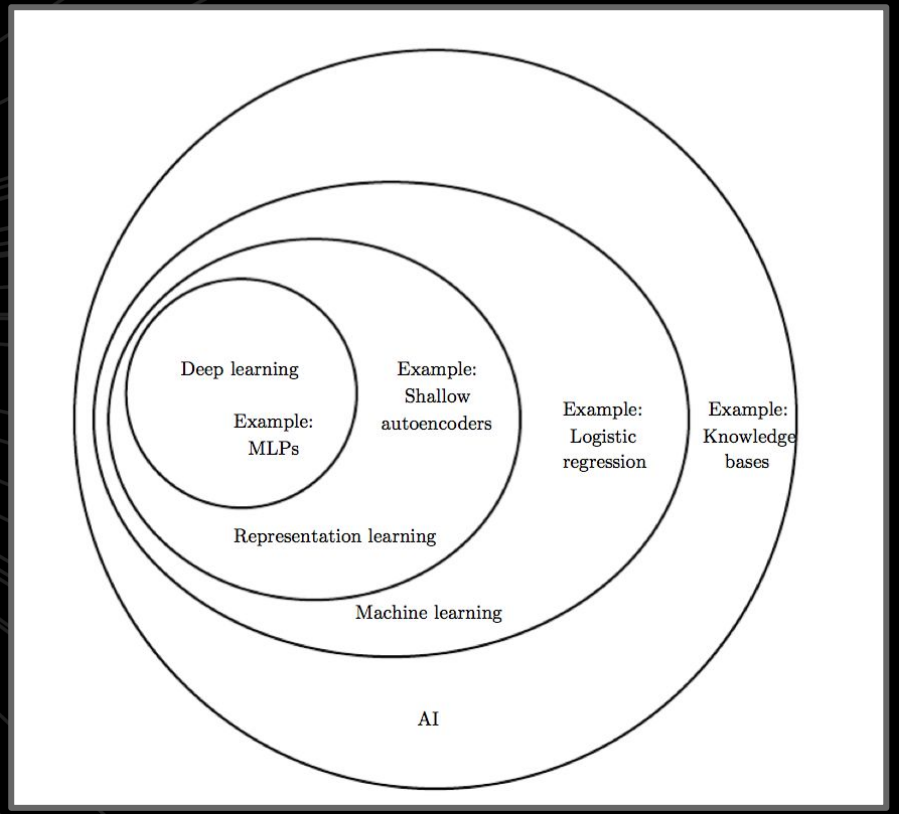
Definitions: AI & NLP

What is deep learning?

- It is a subset of machine learning but works **vastly differently** from most traditional machine learning models, i.e., what were they actually learning?
- It turns out not very much
- Human has to define the features weights, then machine perform numerical optimization

Deep Learning

- Multiple layers of learned representations



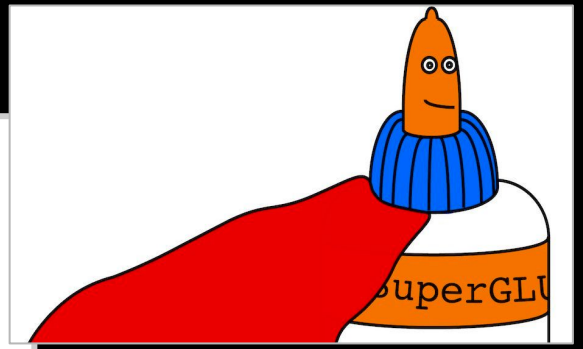
What is Natural Language Processing?

- **An intersection of Artificial Intelligence, Computer Science, and Computational Linguistics**
  - Tasks such as automation and classification (e.g., email spam or not spam)
  - More exciting:
    - Tasks include predicting mortality, machine understanding of emotion, caption generation, intelligent conversational bots, question/answering and much more!

Why is NLP hard?

Language is "human", and full of interesting signals such (tones, gestures, speech)

- Large vocabularies (e.g., other other types of vocabularies such as scientific vocabulary) that is continuously expanding (e.g., emojis)
- domain-specific language, dialects, low-resources, etc.



THE
BEGINNING

Pharmacy

YEAR 2016

NIH Endowment Scholarship

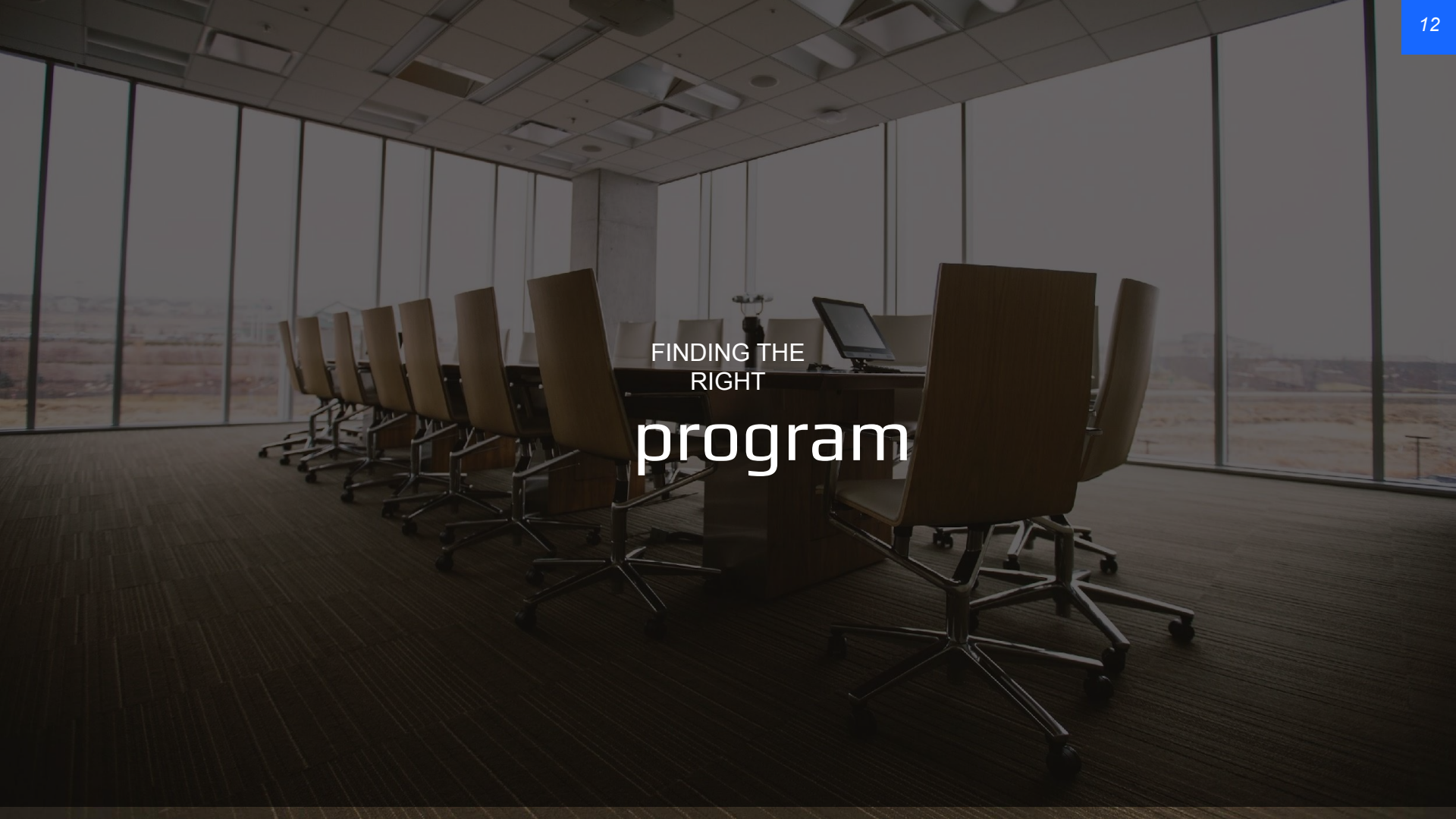
The purpose of the scholarship is to provide financial support to high achieving, qualified students from socially or economically disadvantaged groups as defined by the National Institutions of Health (NIH).

YEAR 2016

NIH Endowment Scholarship

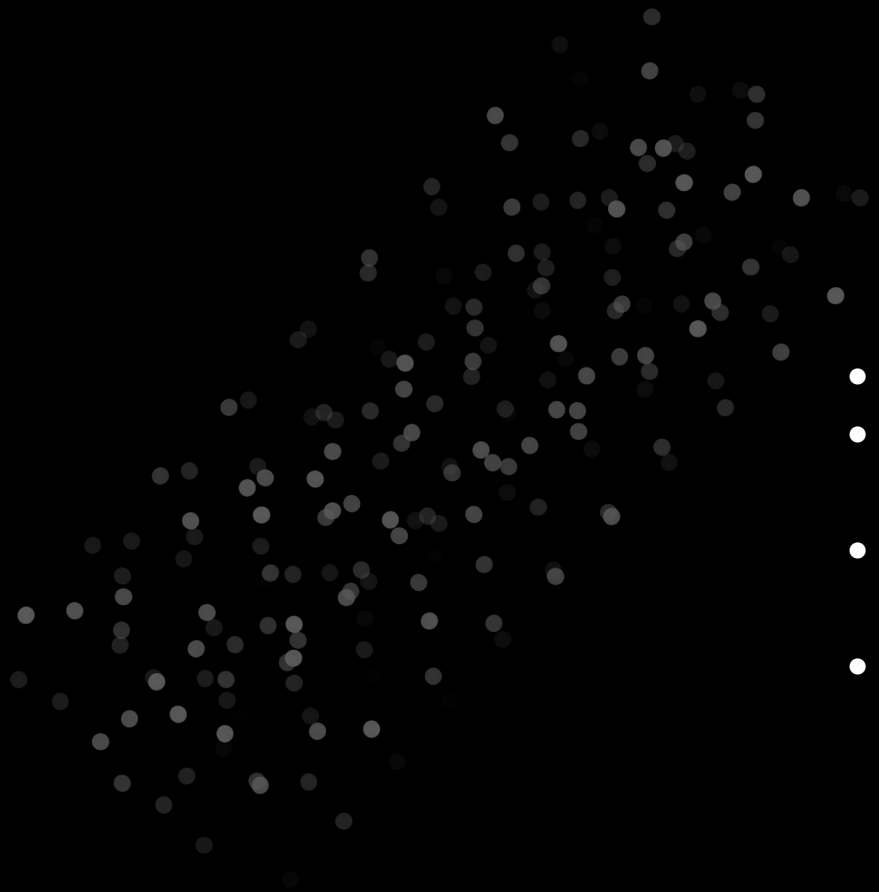
The purpose of the scholarship is to provide financial support to high achieving, qualified students from socially or economically disadvantaged groups as defined by the National Institutions of Health (NIH).


THE DEFINING
MOMENT



FINDING THE
RIGHT

program



Defining your program criteria

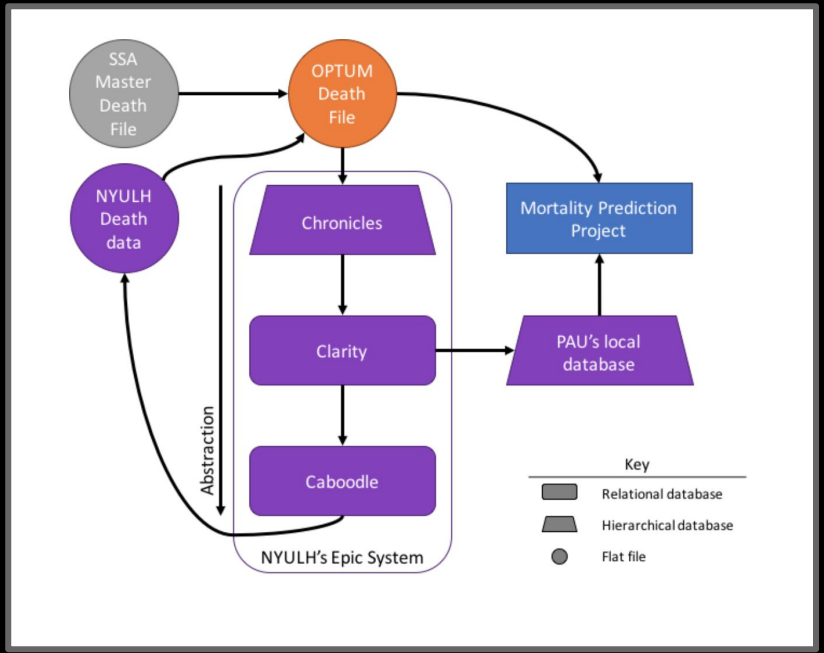
- Ability to work with electronic health record data (EHR)
- Opportunity to focus on deep learning
  - Computational resources (HPC)
- Ability to take classes in various departments
  - e.g., Engineering and math departments
- Rigorous thesis requirements and practicum
  - e.g., Expected to spend ~15 hours/week on practicum work throughout the Spring semester and ~40 hours/week during Summer II

TIPS

Designing an Awesome thesis

(in my perspective)

- DATA AVAILABILITY
- OPPORTUNITY
- POTENTIAL TO CUT COSTS
- YOU ARE PASSIONATE ABOUT THE TOPIC
- SUBJECT MATTER EXPERTS AVAILABLE
- **IMPLEMENTATION:** *IS THERE AN OPPORTUNITY TO PRODUCTIONALIZE YOUR MODEL?*



other tips!

- Don’t overlook the importance of mentors, advocates and sponsors
  - Ask questions!! Don’t be afraid to feel or look stupid
- **Deliver excellence always**
- Don’t put up with being made to feel invisible
  - You have a lot to bring to the table
- Ask for feedback

ZERO START
Deep Learning Models to Predict Mortality from Clinical Text

ZERO START PROBLEM refers to when there are no previous records about a patient who visits us and thus we cannot use any data besides the note made upon admission to the facility. EPIC alert system will not be triggered in these situations.

- **the aim: Can we predict 2-month mortality from unstructured data (clinical text), more specifically the H&P note?**

Deep learning is expensive and resource heavy (e.g., XLNET & BERT)

• Successfully applying deep learning requires more than just a good knowledge of what algorithms exist and the principles that explain how they work.

• We also need to know how

- to choose an algorithm for a particular application
- to monitor and respond to feedback obtained from experiments in order to improve a machine learning system

• During development of deep learning systems, we need to decide:

- whether to gather more data
- increase or decrease model capacity
- add or remove regularizing features
- debug the software implementation of the model
- It is important to be able to determine the right course of action rather than blindly guessing

• Understand what task you are solving and what model architecture you should use

- e.g., machine translation → seq2seq, named entity recognition → biLSTM-CRF or even a CNN, image caption generator → deep CNN + RNN

FOCUS ON PALLIATIVE CARE

what is important to terminally ill patients?

“Not to be kept alive on life support when there is little hope for a meaningful recovery” (55.7%),

“That information about your disease be communicated to you by your doctor in an honest manner” (44.1%) and

“To complete things and prepare for life's end — life review, resolving conflicts, saying goodbye” (43.9%).

Motivation - Patient

Advanced care planning (ACP) is known to

- improve adherence to goals of care (GOC) / end-of-life (EOL) wishes;
- improve patient and family satisfaction with care; and
- decrease family stress, anxiety and depression (Detering et al. 2010).

But, when do we start the process?

Motivation – Prognosis in Practice

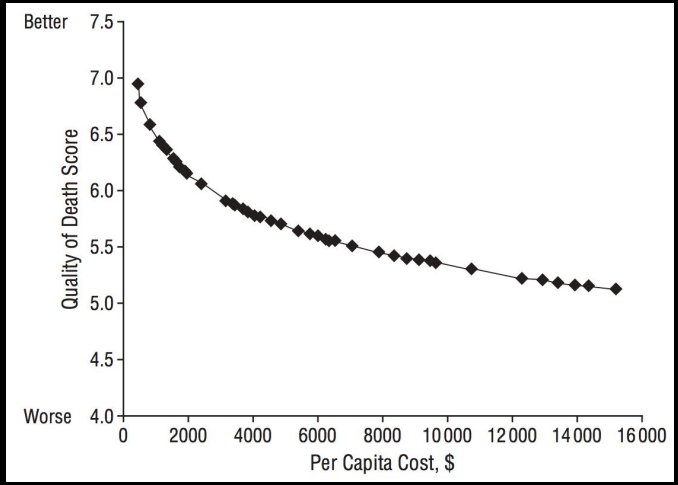
- Estimating when a patient will die is difficult.
  - Physicians are known to overestimate time until death.
  - Optimism can lead to continued treatment that can diminish quality of life.

Motivation – Costs

- Health care costs are soaring and efforts are focusing on reducing waste and low-value care (Rumball-Smith, Shekelle, and Bates 2017).

In one study, higher costs were negatively associated with caregiver reported quality of death, and EOL conversations reduced total cost by 36% ($2780 vs. $1925).

Additional life-sustaining care is expensive and may be detrimental to perceived quality.



(Zhang et al. 2009)

“Unstructured EHR” such as medical notes may provide unique insight and possibly more information than “structured”

A recent publication using the MIMIC III medical notes to predict sepsis within 24 hours found that the unstructured text data performed better than the structured tables. (Culliton, 2017)

why natural language processing?

P.H., a 68-yo is being admitted to the ED after experiencing an episode of sustained CP while mowing his yard. After waiting 1 hour, he called 911 and was transported to the ED. P.H.'s chest pain radiates to his left arm and jaw, and he describes the pain as "crushing" and "like an elephant sitting on my chest." He rates it as a "10/10" in intensity. Thus far, his pain has not responded to five SL nitroglycerin NTG tablets at home and three more in ambulance. His ECG reveals a 3-mm ST segment elevation and Q waves in leads I and V2 to V4. Based on his health & physical examination, P.H. is diagnosed with an AMI.

What symptoms and signs are consistent with an anterior infarction?

(this example is taken from a GPM question from my first year at University of Maryland School of Pharmacy)

why natural language processing?

P.H., a 68-yo is being admitted to the ED after experiencing an episode of sustained CP while mowing his yard. After waiting 1 hour, he called 911 and was transported to the ED. P.H.'s chest pain radiates to his left arm and jaw, and he describes the pain as "crushing" and "like an elephant sitting on my chest." He rates it as a "10/10" in intensity. Thus far, his pain has not responded to five SL nitroglycerin NTG tablets at home and three more in ambulance. His ECG reveals a 3-mm ST segment elevation and Q waves in leads I and V2 to V4. Based on his health & physical examination, P.H. is diagnosed with an AMI.

What symptoms and signs are consistent with an anterior infarction?

The answer lies in the **"patient's story"** then in the ECG and then the labs.

Although anterior infarction can have onset at rest, the pain described is usually a sensation, and chest pain is often described as *"the worst pain I've ever felt"*

SCOPE OF WORK

CENTER FOR HEALTHCARE INNOVATION AND DELIVERY SCIENCE (CHIDS) of NYU LANGONE HEALTH is contracting with Isabel Metzger to build 2 month mortality machine learning based models from initial history and physical notes of admitted patients. Accurate mortality prediction with initial notes would allow models to be built for patients that do not have prior data. These models help providers and patients coordinate in delivering supportive care that align with patient wishes. Specifically we will

(1) integrate a model into our scalable text classification infrastructure,
(2) emit model classifications at a performance threshold daily to an inpatient team for feedback,
(3) deliver a publication ready writeup of the model and results.

Categories requested by physicians

Term	Definition
High Risk	Predicted to die within 2 months.
Appropriate	Expected to die within 6 monhts. GOC/ ACP warranted.
Inappropriate	Neither of the above.
Helpful	High Risk or Appropriate

Feedback from Physicians

email examples - inappropriate fire

Rank	Date	Hosp	Physician Summary
2	--	XXX	**_num_ yo F hx of HTN, HLD, hypothyroidism, stage IIIa lung adenocarcinoma s/p VATS lobectomy m/yyyy s/p cis/alimta completed in December now with radiologic resolution of tumors as of m/yyyy, renal cell carcinoma s/p cryoablation m/yyyy inactive, and PSC (dx yyyy)requiring multiple dilatations and biliary stent placements, with recent presentation in m/yyyy for cholangitis requiring stent placement,necessitating biliary stent exchange, now removed m/d/yy, and now presenting with 3 days of fevers to 105.0F at home and productive cough. Excellent baseline functional status.** **This is an inappropriate fire of the trigger.**
9	--	XXX	_num_ F who presented in labor now s/p c-section. She has no risk factors. **This is an inappropriate trigger**. Patient and baby are doing well.

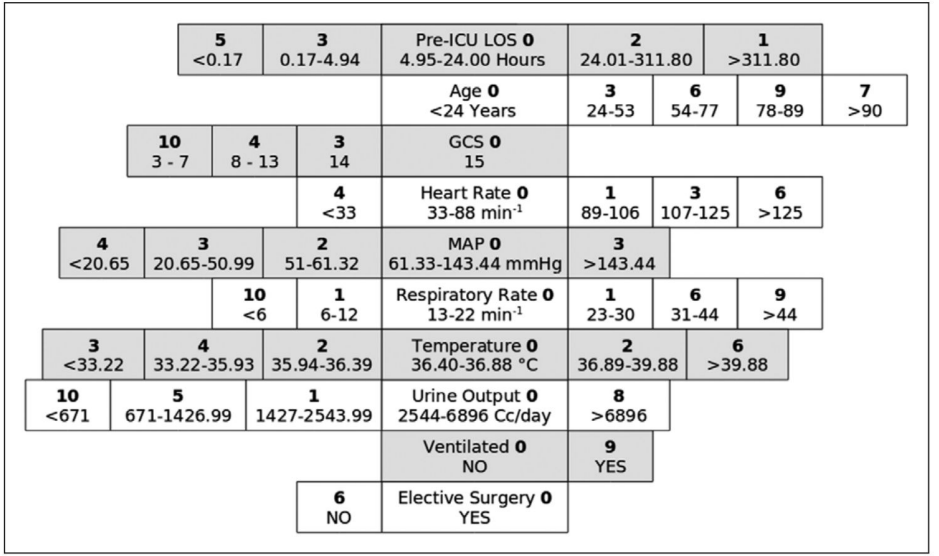
looking back to see real ground truth + meaningful feedback

"One case (I can provide the details if that would be helpful), at the beginning of the admission to me would not have triggered that pt was an end of life patient, but as the admission unfolded this became clear as pt became more acutely ill. Since you are running this tool retrospectively I am curious if this patient would have flagged at the start of pt admission."

Background – *Score Based Metrics*

The problem:

- Points-based metrics discretize real world physiology into coarse, weighted bins.
  - Makes them easy to use and interpretable
  - Restricted to integer weights, a small number of variables, a small number of bins
    - limits the estimator's freedom to represent complex diseases
      - thus, generalizability and performance.



5 <0.17	**3** 0.17-4.94	Pre-ICU LOS **0** 4.95-24.00 Hours	**2** 24.01-311.80	**1** >311.80		
		Age **0** <24 Years	**3** 24-53	**6** 54-77	**9** 78-89	**7** >90
10 3 - 7	**4** 8 - 13	**3** 14	GCS **0** 15			
		4 <33	Heart Rate **0** 33-88 min^{-1}	**1** 89-106	**3** 107-125	**6** >125
4 <20.65	**3** 20.65-50.99	**2** 51-61.32	MAP **0** 61.33-143.44 mmHg	**3** >143.44		
	10 <6	**1** 6-12	Respiratory Rate **0** 13-22 min^{-1}	**1** 23-30	**6** 31-44	**9** >44
3 <33.22	**4** 33.22-35.93	**2** 35.94-36.39	Temperature **0** 36.40-36.88 °C	**2** 36.89-39.88	**6** >39.88	
10 <671	**5** 671-1426.99	**1** 1427-2543.99	Urine Output **0** 2544-6896 Cc/day	**8** >6896		
			Ventilated **0** NO	**9** YES		
		6 NO	Elective Surgery **0** YES			

Figure 1. Component weights and bins for the Oxford Acute Severity of Illness Score (OASIS). The *bold values* are the individual scores assigned to an associated range of measured values. For each variable, the worst score across the first day should be used to tabulate OASIS. The final OASIS score is the sum of all the component weights. LOS = length of stay, GCS = Glasgow Coma Score, MAP = mean arterial pressure.

CONCLUSIONS & LIMITATIONS

The results of my thesis illustrate a potential path to identifying instances where intervention is beneficial when we have patients with no other data besides their admission note (clinical text).

The top ten leading causes of death as reported by the CDC include Heart disease, Cancer, Chronic lower respiratory diseases, Accidents (unintentional injuries), Stroke (cerebrovascular diseases), Alzheimer's disease, Diabetes, Influenza and Pneumonia, Nephritis, nephrotic syndrome and nephrosis,and Intentional self-harm (suicide) (CDC 2016).

Not all of these would be recommended for palliative care referral but they can be used with some sort of intervention.

Lessons Learned

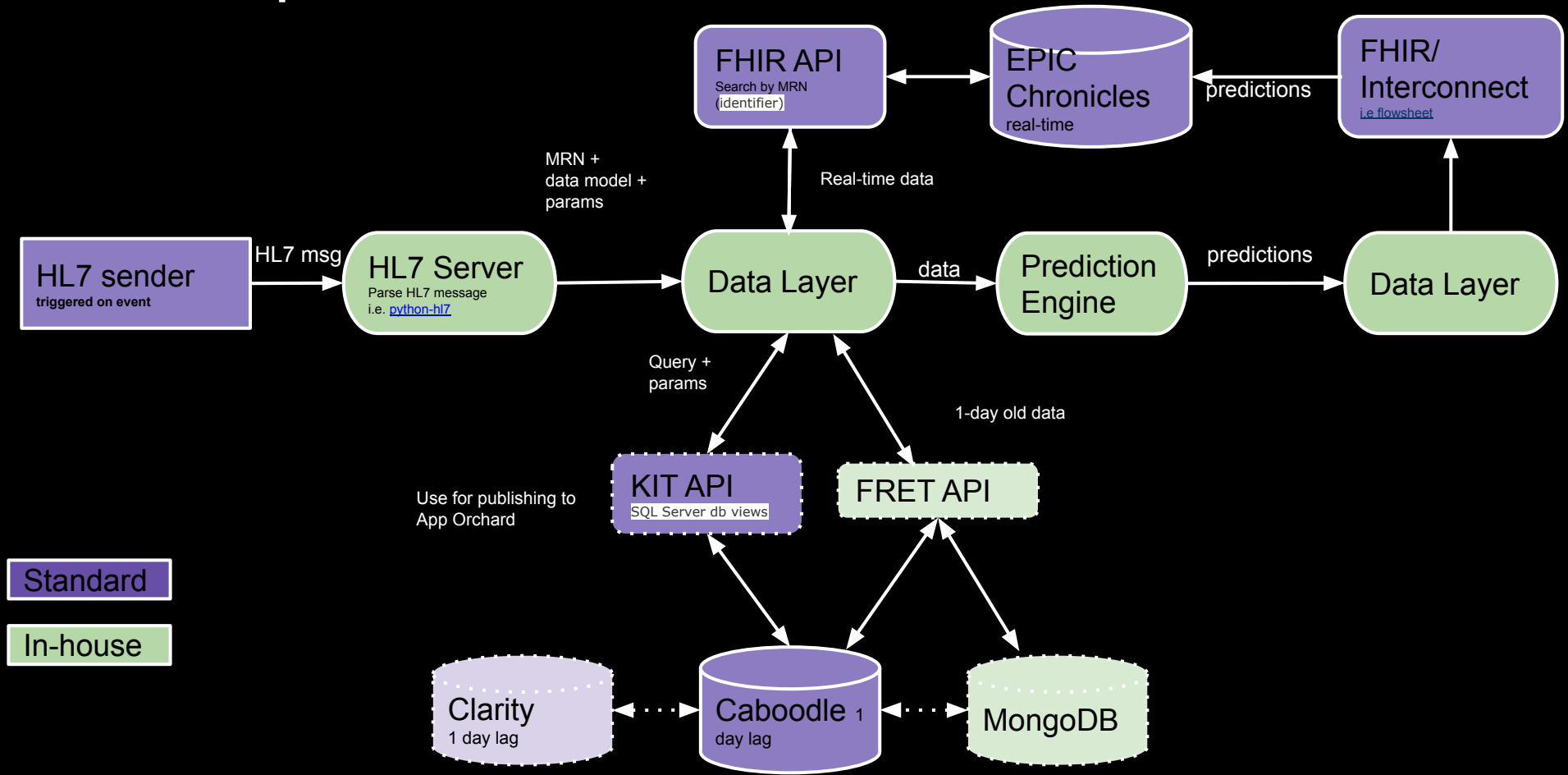
MISTAKE 1

While it is good to be proactive, recognize that you do not have to say yes to any project when trying to connect with mentors

Avoid politically risky projects, especially when you are in the start of your program!

CASE STUDY: “Value Opportunity Tool Project”

PAU Operational / Data Flow



Project Goal

Align physicians with corporate strategy while improving the effectiveness of care for our patients

A potential path to success:

Peer match physicians across 8 financial and 4 quality dimensions, normalized by length of stay and CMI

-by Division/Dept, DRG Procedure

Targeted outcomes:

Foster discussions among physicians

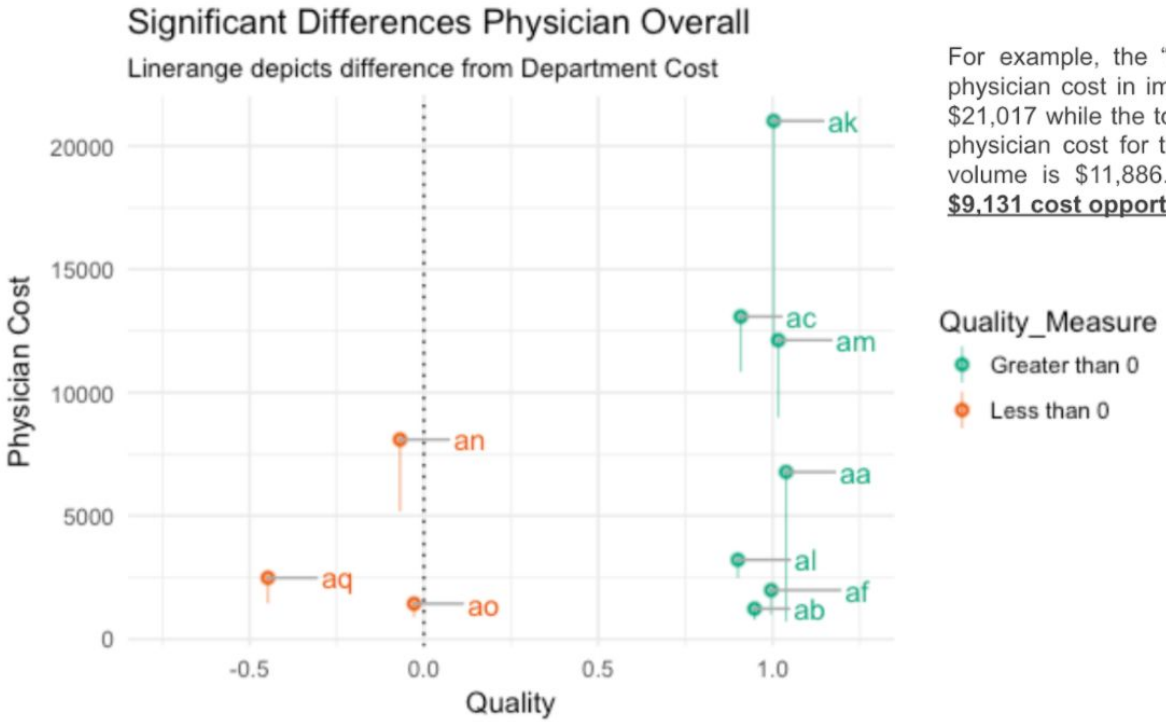
Identify best practices

Create a better environment for both physicians and patients

--> Improve patient care while managing costs for the hospital

highlighting physicians and expense categories with the **most** opportunity within a department/division

Summary across all DRGs and physicians unadjusted for length of stay



For example, the "ak" total physician cost in implants is $21,017 while the total other physician cost for the same volume is $11,886. Thus a **$9,131 cost opportunity.**

INITIAL MODELING

- **Observations:** 17,437
  Variables: 121
- 17,437 encounters (visits), 646 unique DRG procedures, 621 attending physicians
- Variables include Length of Stay, Admission Type, descriptions, Expected LOS, cost by multiple sub-categories, and etc..

Modeling Assumptions & Notes

Cost Medians:

- We compare medians in a DRG between one physician and every other physician using the Wilcoxon rank sum test.

Windowed Analysis:

- We consider only patients seen between October 2015 and February 2016.

Minimum Patient Count::

- Physicians must see a minimum of 5 patients in a DRG for comparison to this peers.

Not an Average::

- Not an average. Both the cost and quality axis are not averages. Thus if a physician does poorly in some DRGs but well in others, this will not be represented. Currently, only the DRGs that the physician performs worse will be shown.

template example

Dear Dr. S364,

In the time period from July 2015 to Dec 2015, in your medicare patient population, you saw **79 patients** and your colleagues saw **198 patients** for **total hip replacement**.

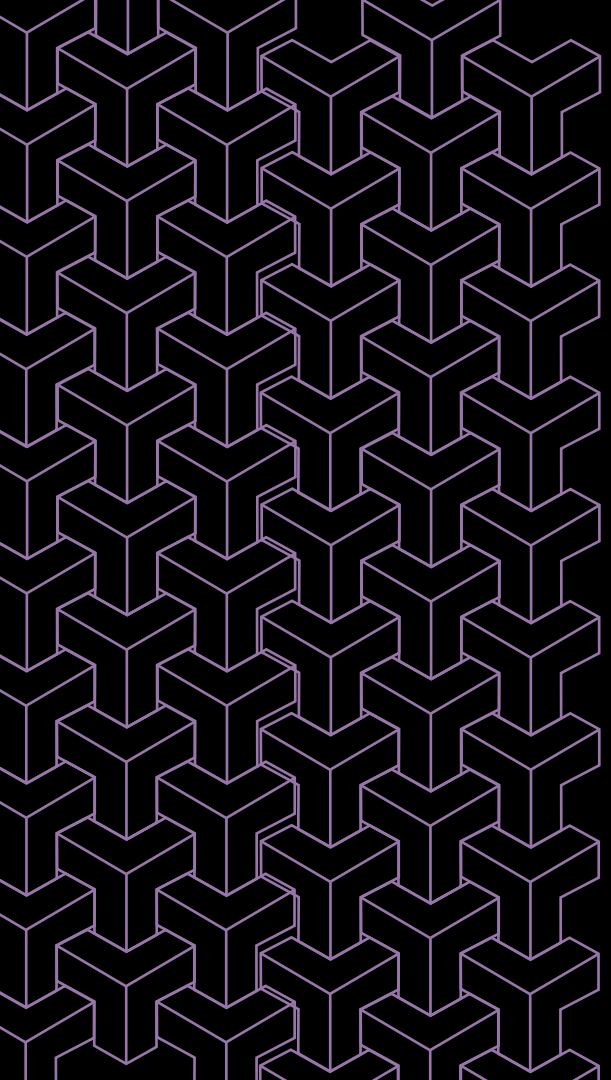
There are several opportunities for decreasing cost. After adjusting for length of stay and CMI severity of the patients you saw, you have two statistically significant opportunities to decrease costs.

Radiology: You spent $110.00 more median adjusted dollars per case than your colleagues for this procedure. Optimistically if you were able to reduce the cost of all your patients to the median of all your colleagues, you could save $17k.

Pharmacy: You spent $33.00 more median adjusted dollars per case than your colleagues for this procedure. Optimistically if you were able to reduce the cost of all your patients to the median of all your colleagues, you could save $8,767.

Both recommendations adjust for the statistically significant difference between your case mix index between you (at 2.91) and your colleagues (at 3.11). We look forward to your comments or thoughts on how to validate and incorporate these insights into your practice.

Thank you,
Your Chairman



KEY TAKEAWAYS

1. DON’T FEEL OBLIGATED TO IMMEDIATELY TAKE ON THE PROJECT
2. ASSESS YOUR PROJECT NOT FROM PURE ROI. WILL YOUR PROJECT GET BUY IN?
3. DO YOU HAVE A STAKEHOLDER RELATIONSHIP AND THE EXPERTISE TO SEE YOUR PROJECT SUCCEED

MISTAKE 2

FEAR & PERFECTIONISM

- Examples of ambitious project ideas:
  - Scrape Genius.com data, and use it to automatically explain what a song lyric means.
  - Crowdsource a list of antonyms (like hot–cold), and use them to help improve *word embedding* models.
  - Build a semantic parser to answer questions about NYC's public database of past taxi trips.
  - See if multiscale recurrent neural networks learn to identify noun phrase boundaries when they're trained on text understanding tasks.
  - Write an extension to MacCartney's Natural Logic that allows it to reason about questions.
- We'll talk more about the project on Feb 21 and April 4, but start thinking about what interests you.

How many patients with cystic fibrosis were given TPN?

NER:

```
How many patients with PROBLEM@1 were
given TREATMENT@1 {PROBLEM@1: cystic
fibrosis, TREATMENT@1: TPN}
```

Seq2Seq:

```
SELECT count(DISTINCT hadm_id)
FROM DIAGNOSES_ICD
WHERE icd9_code IN
(SELECT DISTINCT icd9_code
FROM D_ICD_DIAGNOSES
WHERE long_title LIKE '%PROBLEM@1%') AND
hadm_id IN
(SELECT DISTINCT hadm_id
FROM PRESCRIPTIONS
WHERE drug LIKE '%TREATMENT@1%')
```

PROJECT GOAL

to build an interface across the electronic health record (EHR) that would translate questions from “natural language” to SQL queries, while accounting for domain-specific abbreviations, arises from realizing the need for a tool to aid clinicians in obtaining information for critical patient care decisions faster

NAMED ENTITY RESOLUTION WITH NER

We handle entities in the utterances and SQL by replacing them with their types, using incremental numbering to model multiple entities of the same type.

- Bi-directional Long Short Term Memory Cell with Conditional Random Forest (BiLSTM-CRF) initialized on pre-trained word embeddings learned on texts from Wikipedia, PubMed, and MEDLINE.
- BiLSTM-CRF trained on 2012 i2b2 golden-standard corpus tagged with entities: *problem, test, treatment, department, evidential, occurrence*
- Additional entities *religion, insurance, ethnicity* and *numeric* take a limited set of values which can be recognized using typical string matching algorithms.

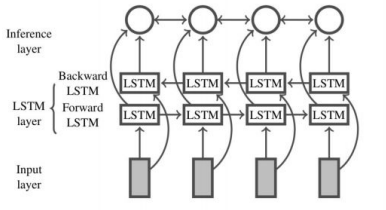


Figure 1. Architecture of NER

TRANSLATION TO SQL

We use an encoder-decoder model with global attention.

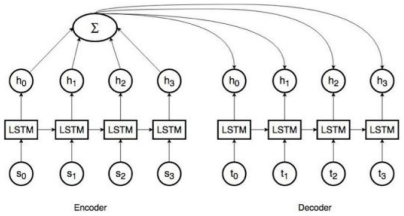


Figure 2. Architecture of Seq2Seq with Attention

- Dong and Lapata(2016) proposed Seq2Seq models with attention for semantic parsing
- Comparatively similar results on GEO(~500) and JOBS(~600) question to SQL datasets

examples from reviewer comments

This writeup reports development of a question-answering (QA) system over electronic health records (EHRs) using recent technology. The work seems to be solid and of interest, but the paper ends abruptly in the middle of the method section, without results or discussion.

This paper presented an interesting natural language interface for EHR by combining various tasks/approaches from the literature. However, no evaluation has been conducted such that one can understand the effectiveness of the proposed system. I would suggest the authors to conduct a thorough evaluation of each component of the system and finally, do an overall end-to-end evaluation to show the impact of the proposed system on the clinician's workflow.

Although the work itself was novel and the methodology was sound, I felt unsatisfied with the quality of the submission itself and decided to withdraw the paper.

Withdrawing the paper was a missed opportunity

- *a year later a similar work was published in JAMIA*

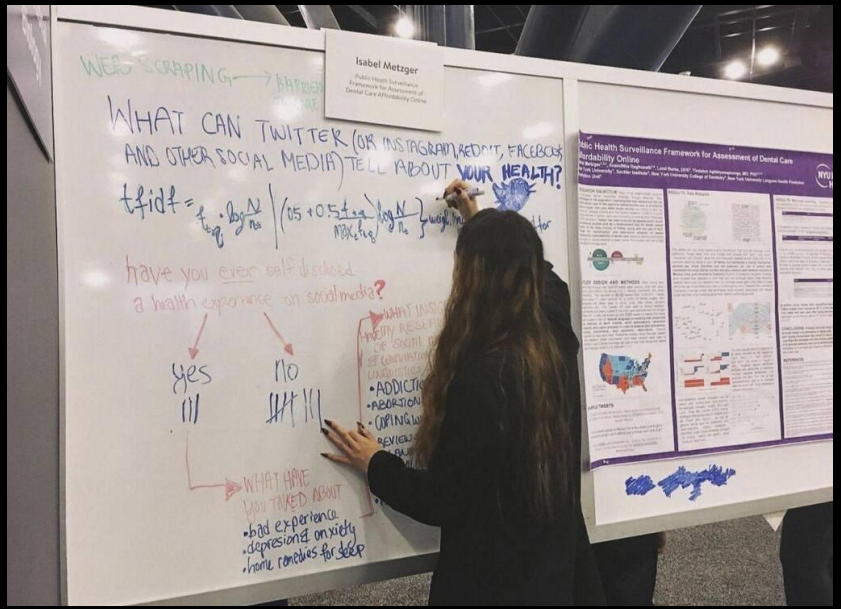
KEY TAKEAWAYS

1. Delivery and presentation
2. SUBMIT YOUR WORK!
3. If accepted, publish!! Don’t let perfect be the enemy of good!

BUILDING
YOUR

DATASCIENCE / AI PORTFOLIO

1. PUBLICATIONS (white papers, research papers, abstracts, etc.)
2. CONFERENCE PRESENTATIONS
3. WEB PRESENCE/WEBSITE & GITHUB
4. MISC, e.g., DESIGNING TUTORIALS (e.g., corporate lunch & learns, data science course curriculums, etc.), participating in datathons, innovation challenges, and more!



Present at conferences and at open data weeks!

PQA
HEALTHCARE QUALITY
INNOVATION CHALLENGE

WHO WE ARE

- PQA's mission is *optimizing health by advancing the quality of medication use.*
- PQA has 4 cross-cutting roles: measure developer, researcher, healthcare quality educator, and convener.
- Established in 2006, the Pharmacy Quality Alliance (PQA) is a 501(c)3 designated non-profit alliance with over 200 member organizations including health plans, pharmacies, pharmacy benefit managers, pharmaceutical manufacturers, academic institutions, healthcare technology developers, associations, patient advocacy groups, and state and federal agencies.

WHAT IS THE PQA HEALTHCARE QUALITY INNOVATION CHALLENGE?

- A team-based student competition designed to promote student engagement in leveraging technology to create realistic solutions that improve healthcare quality measurement
- Teams will be provided 3 prompts and will choose 1 to respond to with a business summary (no more than 12 pages).
- The top 3 teams will pitch their solution during the PQA Annual Meeting in May 2018.

WHY SHOULD I PARTICIPATE?

- Recognition and awards
- Showcase innovative and creative ideas that leverage technology
- Use business and entrepreneurial skills
- Collaborate with other students
- Expand your knowledge of healthcare quality and quality measurement
- Various networking opportunities for finalists

Competition sponsored by:

PQS
PHARMACY QUALITY SOLUTIONS

other types of competitions

(not just coding competitions!)

PROMPT: Patient Engagement with the Medicare Prescription Drug Program: Develop a technology-enabled solution to help Medicare beneficiaries report on their experiences with their Part D plan more easily or to help plan sponsors interact more directly with beneficiaries to improve upon their CMS Star Ratings.

TITLE: Utilizing Social Media and Machine Learning to Engage Medicare Beneficiaries and Capture their Experiences with the Part D Program.

Final Thoughts

- It is super cool to get to work on problems using **diverse** data to support decision making
  - structured fields
  - free text

- Important to know the state-of-the-art in machine learning in various domains
- Important to closely collaborate with the stakeholders and subject matter experts to address the *right* problem
- Collaboration is key to seeing data projects succeed
- Know your limitations (e.g., stakeholder communication and presentation of work)
- **Visualization and interpretability are important**
- Important to be in a place where you can grow and learn new skills

“there’s a lot more to creating useful data projects than just training an accurate model!”

“Talent wins games, but teamwork and intelligence win championships.” - MJ

Special Thanks

Any Questions

Isabel Metzger
isabel.metzger@sumitovant.com

references

MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available from: http://www.nature.com/articles/sdata201635.

Sun W, Rumshisky A, Uzuner Ö. (2013). "Annotating temporal information in clinical narratives". J Biomed Inform. 2013 Dec;46 Suppl:S5-12. doi: 10.1016/j.jbi.2013.07.004. Epub 2013 Jul 19. http://www.ncbi.nlm.nih.gov/pubmed/23872518.

Sun W, Rumshisky A, Uzuner Ö. (2013). "Evaluating temporal relations in clinical text: 2012 i2b2 Challenge". J Am Med Inform Assoc. 2013 Sep-Oct;20(5):806-13. doi: 10.1136/amiajnl-2013-001628. Epub 2013 Apr 5. http://www.ncbi.nlm.nih.gov/pubmed/23564629.

Avati A, Jung K, Harman S, et al. (2017) Improving Palliative Care with Deep Learning. arXiv [cs.CY]. Available from: http://arxiv.org/abs/1711.06402.

Detsky ME, Harhay MO, Bayard DF, et al. (2017) Discriminative Accuracy of Physician and Nurse Predictions for Survival and Functional Outcomes 6 Months After an ICU Admission. JAMA: the journal of the American Medical Association 317(21): 2187–2195.

Ghassemi MM, Richter SE, Eche IM, et al. (2014) A data-driven approach to optimized medication dosing: a focus on heparin. Intensive care medicine 40(9): 1332–1339.

Li J, Chen X, Hovy E, et al. (2016) Visualizing and Understanding Neural Models in NLP. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 681–691.

Ghassemi MM, Richter SE, Eche IM, et al. (2014) A data-driven approach to optimized medication dosing: a focus on heparin. Intensive care medicine 40(9): 1332–1339.

Huang J, Osorio C, Sy LW. (2018) An Empirical Evaluation of Deep Learning for ICD-9 Code Assignment using MIMIC-III Clinical Notes. arXiv [cs.CL]. Available from: https://arxiv.org/abs/1802.02311.

references continued

Sun W, Rumshisky A, Uzuner Ö. (2013). "Annotating temporal information in clinical narratives". *J Biomed Inform.* 2013 Dec;46 Suppl:S5-12. doi: 10.1016/j.jbi.2013.07.004. Epub 2013 Jul 19. http://www.ncbi.nlm.nih.gov/pubmed/23872518.

Sun W, Rumshisky A, Uzuner Ö. (2013). "Evaluating temporal relations in clinical text: 2012 i2b2 Challenge". *J Am Med Inform Assoc.* 2013 Sep-Oct;20(5):806-13. doi: 10.1136/amiajnl-2013-001628. Epub 2013 Apr 5. http://www.ncbi.nlm.nih.gov/pubmed/23564629.

Aczon, M., D. Ledbetter, L. Ho, A. Gunny, A. Flynn, J. Williams, and R. Wetzel. 2017. “Dynamic Mortality Risk Predictions in Pediatric Critical Care Using Recurrent Neural Networks.” arXiv [stat.ML]. arXiv. http://arxiv.org/abs/1701.06675.

Cai, Xiongcai, Oscar Perez-Concha, Enrico Coiera, Fernando Martin-Sanchez, Richard Day, David Roffe, and Blanca Gallego. 2016. “Real-Time Prediction of Mortality, Readmission, and Length of Stay Using Electronic Health Record Data.” Journal of the American Medical Informatics Association: JAMIA 23 (3): 553–61.

Ustun, Berk, and Cynthia Rudin. 2016a. “Learning Optimized Risk Scores on Large-Scale Datasets.” arXiv Preprint arXiv:1610. 00168. https://arxiv.org/abs/1610.00168.

Gagne, Joshua J., Robert J. Glynn, Jerry Avorn, Raisa Levin, and Sebastian Schneeweiss. 2011. “A Combined Comorbidity Score Predicted Mortality in Elderly Patients Better than Existing Scores.” Journal of Clinical Epidemiology 64 (7): 749–59.

Glare, Paul, Kiran Virik, Mark Jones, Malcolm Hudson, Steffen Eychmuller, John Simes, and Nicholas Christakis. 2003. “A Systematic Review of Physicians’ Survival Predictions in Terminally Ill Cancer Patients.” BMJ 327 (7408): 195–98.

back up slides

Deep learning is expensive and resource heavy (e.g., XLNET & BERT)

• Successfully applying deep learning requires more than just a good knowledge of what algorithms exist and the principles that explain how they work.

• We also need to know how

- to choose an algorithm for a particular application
- to monitor and respond to feedback obtained from experiments in order to improve a machine learning system

• During development of deep learning systems, we need to decide:

- whether to gather more data
- increase or decrease model capacity
- add or remove regularizing features
- improve the optimization of a model
- improve approximate inference in a model
- debug the software implementation of the model
- It is important to be able to determine the right course of action rather than blindly guessing

• Understand what task you are solving and what model architecture you should use

- e.g., machine translation → seq2seq, named entity recognition → biLSTM-CRF or even a CNN, image caption generator → deep CNN + RNN

Background – *Deep Learning and achine Learning Approaches*

Cohort	Method	AUROC	Comments	
Advanced cancer in palliative care	23 independently predictive variables combined. Predict three cases: days, weeks, and months+	0.79–0.86	Likely won't generalize to non-cancer	Gwilliam et al. (2015)
Elderly Medicare patients	Split prior ICD codes into two sets, recent and baseline. Random forest model to predict 6 month	0.83	RF cannot couple the two variables	Makar et al. (2015)
Cancer patients starting chemotherapy	Split data into two sets, recent and baseline. XGBoost model to predict 30-day	0.94	Similar to above, better performance via more data	Elfiky et al. (2017)

DETERMINE ERROR METRIC AND TARGET VALUE

- examples, F1 score of a particular class or 80% PPV at 20% sensitivity, PRAUC, etc.

- Establish a working end-to-end pipeline **AS SOON AS POSSIBLE**
- **THE INPUTS AND OUTPUTS OF THE MODEL**
- **THE PRE-PROCESSING/POST-PROCESSING**

DIAGNOSE POOR PERFORMANCE

- identify where your model is performing poorly, e.g., whether it is due to
  - Overfitting
  - Underfitting
  - defect in data or your implementation
  - issue with how you compiled and saved your model

- Make incremental changes in terms of incorporating new data, optimization, and changing algorithms
  - if the problem you're researching is a well-studied problem, follow the leading architectures

(I have a bunch of other tips if anyone is interested/ best practices -- deep learning is computationally expensive and if you don't know what you are doing, it can be incredibly wasteful)

network architectures

CNN Convolutional layer has kernel size 1,2, and 3, 256 hidden units and ReLU activation. Dropout with ratio of .3 and batch normalization are applied at all levels of the network. CNN + dense Same structure as CNN with an additional dense hidden layer with 256 units before output layer. BiLSTM 1 layer BiLSTM with 256 hidden units in each direction. We do not use dropout or batch normalization in this model. BiLSTM + dense Same structure as BiLSTM with an additional dense hidden layer with 256 hidden units before output layer.

“Pipeline Jungles. As a special case of glue code, pipeline jungles often appear in data preparation. These can evolve organically, as new signals are identified and new information sources added incrementally. Without care, the resulting system for preparing data in an ML-friendly format may become a jungle of scrapes, joins, and sampling steps, often with intermediate files output. Managing these pipelines, detecting errors and recovering from failures are all difficult and costly.

Testing such pipelines often requires expensive end-to-end integration tests. All of this adds to technical debt of a system and makes further innovation more costly.

The clean-slate approach of scrapping a pipeline jungle and redesigning from the ground up is indeed a major investment of engineering effort, but one that can dramatically reduce ongoing costs and speed further innovation.

Glue code and pipeline jungles are symptomatic of integration issues that may have a root cause in overly separated “research” and “engineering” roles. A hybrid research approach where engineers and researchers are embedded together on the same teams (and indeed, are often the same people) can help reduce this source of friction significantly.”

I strongly believe this. My most successful deep learning models have been in situations where I worked closely with infrastructure & devops at NYU and also at Express Scripts Software Engineering *(I actually had my own dedicated devops engineer)*.

specifics

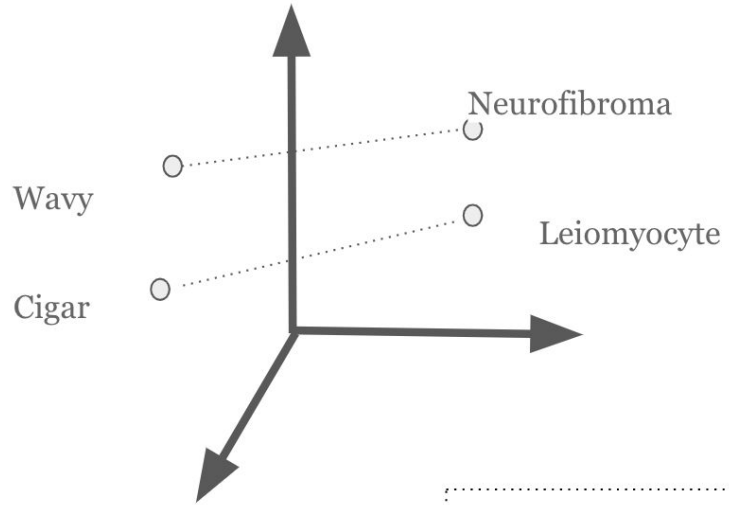
vanilla CNN, CNN with attention (CAML), LSTM, bi-LSTM, GRU, bi-GRU, for clinical note classification. Similar to [33] we used BCEWithLogitsLoss with early stopping with Adam optimizer (CNN family: input length =2000, epoch 200, number of filters = 500, activation function =relu/tanh, number of epochs = 100, batch size = 512, dropout=0.2, lr: 1e-3/3e-3. RNN family: input length =2000, vocab size = 20000, hidden size = 512, number of epoch = 200, batch = 64., lr: 3e-3.)

“Unstructured EHR” such as medical notes may provide unique insight and possibly more information than “structured”

A recent publication using the MIMIC III medical notes to predict sepsis within 24 hours found that the unstructured text data performed better than the structured tables. (Culliton, 2017)

qualitative exploration of word embeddings

A glance at nearest neighbors with domain specific acronyms



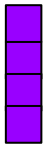
wiki+medical notes **VS** **wikipedia**

Query word? dnr
dni 0.942678
hcp 0.749637
resuscitate 0.711382
code 0.65287
intubate 0.638306
cmo 0.637459
dnri 0.612248
reintubate 0.606106
hcps 0.603236
hospice 0.58464

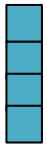
Query word? dnr
dnl 0.954677
dnssec 0.940462
cwp 0.936214
dpb 0.934438
hvdc 0.929971
bvu 0.927638
dnq 0.927279
tpb 0.925606
pkc 0.924558
hvd 0.923596

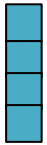
"dnr" stands for do not resuscitate, and using nearest neighbors on the unsupervised model we created on the large amount of medical notes plus a little bit of wikipedia, we see "dni", "hcp", "code", "cmo"--these are domain jargon we often see referring to end of life "dni stands for do not intubate, "hcp" means health care provider, "code" is one part of "full code" usually referring to death, "cmo" means comfort measures only

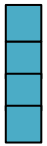
how to convolve on text

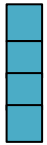


Patient with



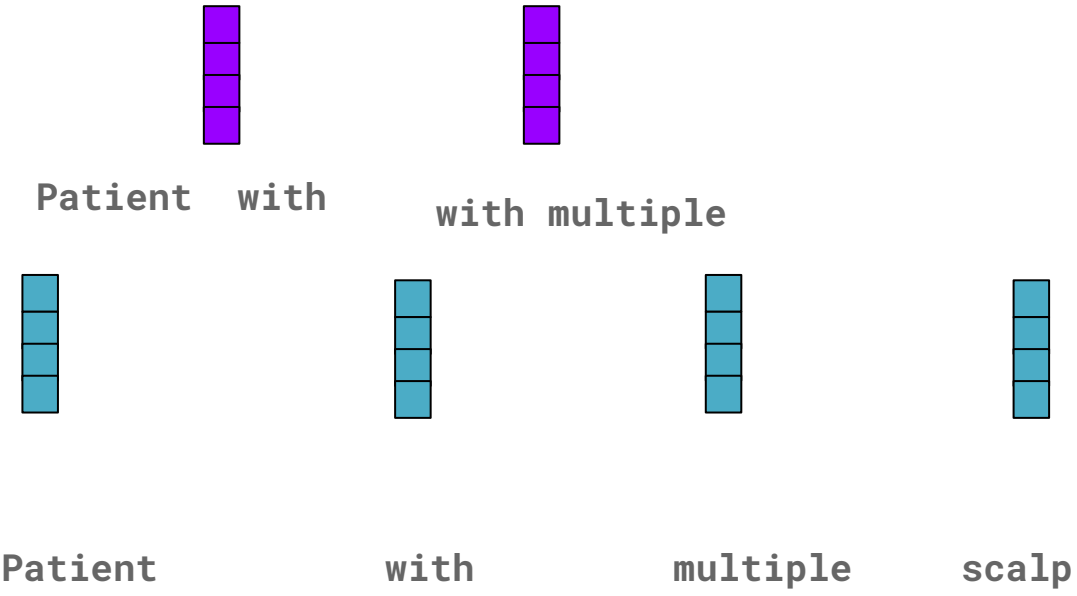




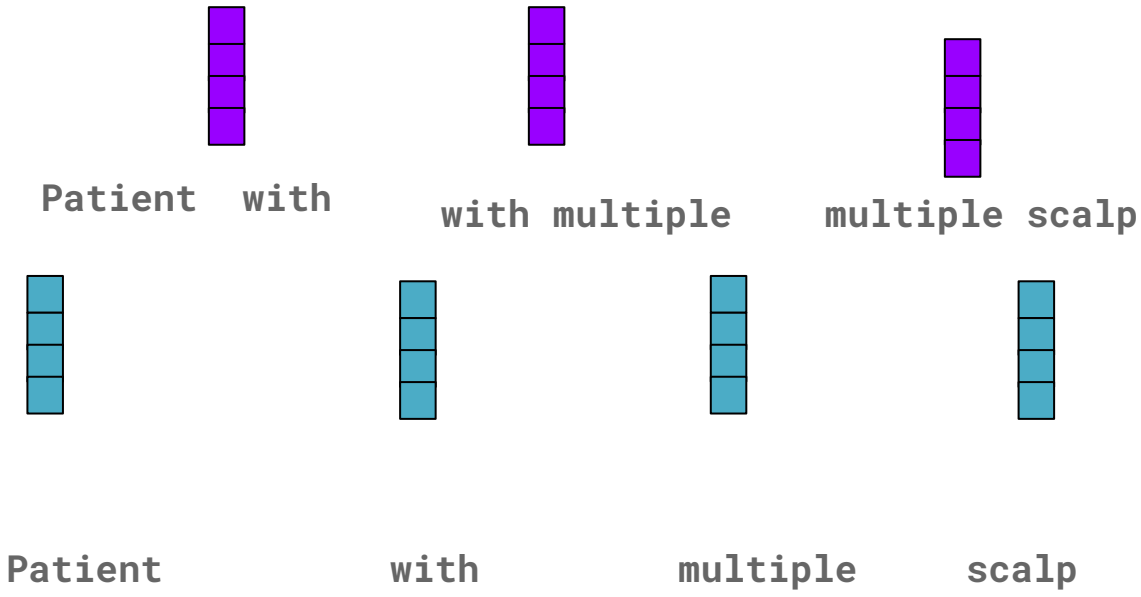


Patient with multiple scalp

how to convolve on text



how to convolve on text



What goes into deploying AI?

- **Packaging of models into containers with dependencies**
- Diagnostics, issue resolution and redeployment (dynamic process)
- Container cluster management and scheduling (Kubernetes)
- Distributed, event-driven architecture, messaging middleware (Kafka)
- Distributed logging / monitoring infrastructure
- Databases (**MongoDB**, PostgreSQL)
- Multiple stacks (R, **Python**, Matlab)
- Evaluation (A/B tests, other designs?) of **multiple models, in different envs**
- Using ops techniques such as decoys, canary models
- Monitoring of models in production

Multiple Projects & Workflows