Intro to Machine Learning for Health

CRTP Fall Term, Week 1

Matt Engelhard, Duke Biostatistics & Bioinformatics

Course Objectives

You might not be building models, but you'll use them. You should know:

1. What they can and can't do (*capabilities, limitations*)
2. When and how much to trust their predictions (*evaluation*)
3. What's inside the black box (*interpretation*, *understanding*)

Equip you to:

(a) build your own models (if you wish)

(b) design and manage data science research and/or QA/QI projects

(c) collaborate and communicate effectively with data scientists

(d) add rigor to model development and validation


ONE WAY
DO NOT
ENTER

Introductions!

your name

+ your role either just before CRTP or outside of CRTP

+ any other fact about yourself (e.g. a typical breakfast, last tv show you watched)

+ how you feel about the effects DS/ML/AI will have on health

  (excited, worried, both, something else)

Course Overview

We will learn about state-of-the-art data science techniques that have changed or will change clinical practice.

- How are these techniques different from what has come before?
- How are they the same?
- What do you need to know to use them effectively and responsibly?

I know that most of you are NOT going to be data scientists.

But you *will* work with data scientists, and you *will* have to make decisions about what models to use and how to use them. It is important to know enough to get in the weeds with the data scientists, because if applied/evaluated incorrectly, these models are certain to be unhelpful and *likely to be harmful*.

A Brief Tour of DS in 2023

General Application -> Biomedical Application

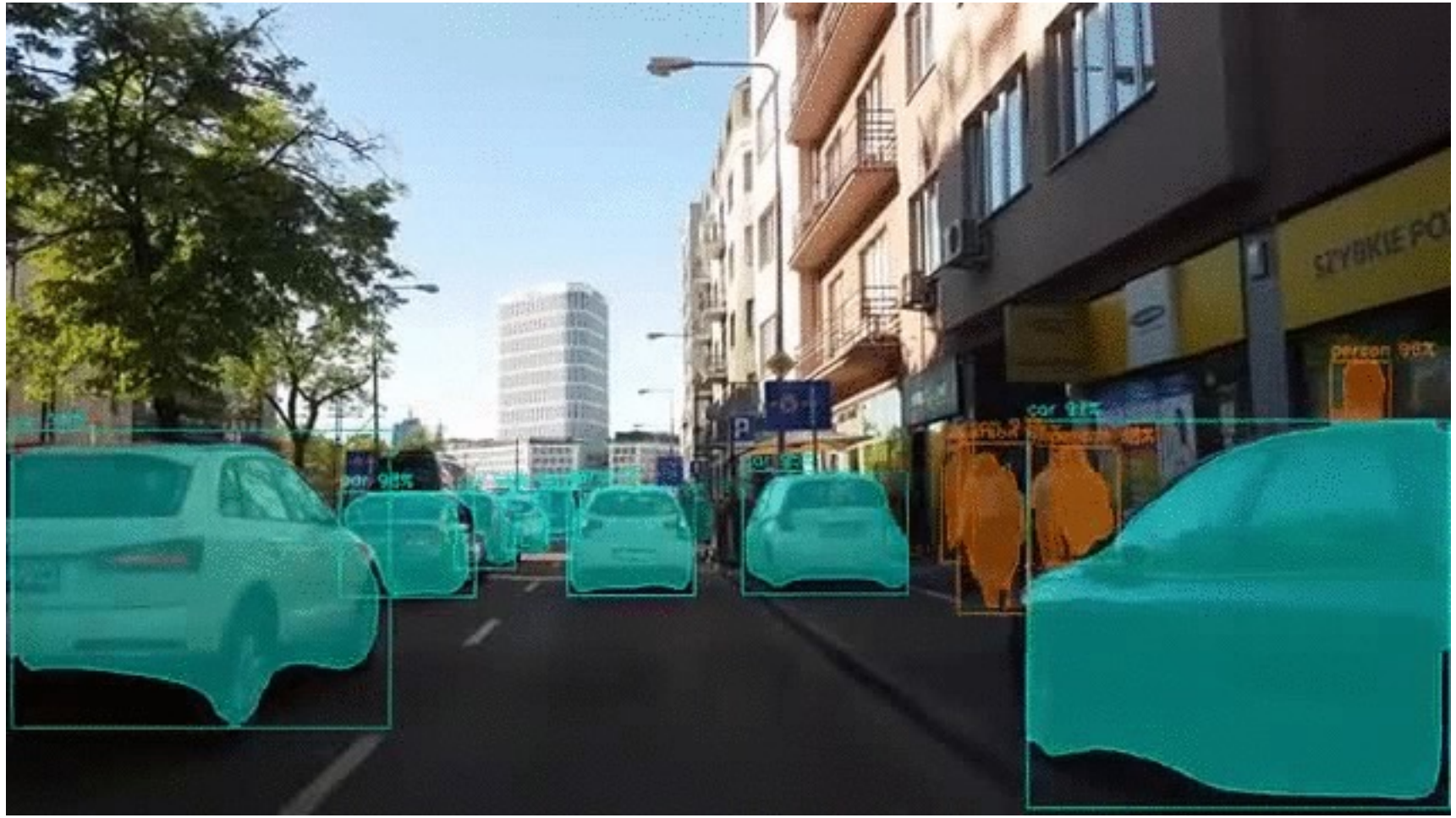
10,000-foot view of machine learning in 2023

- Computer vision (self-driving)
- Natural language processing (NLP) (ChatGPT)
- Reinforcement learning (AlphaGo)

- Each of these has high clinical relevance

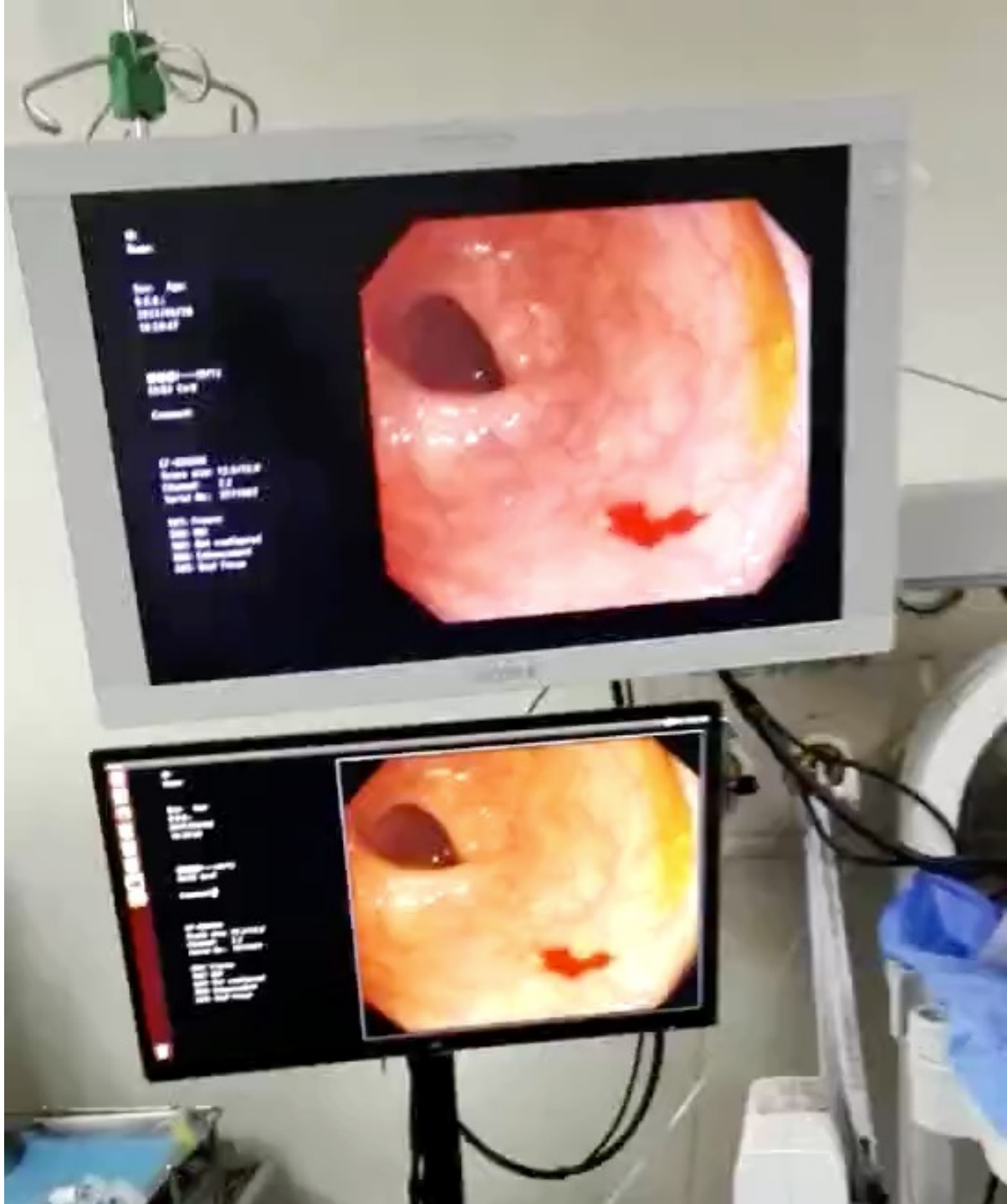
- ML for large, high-dimensional biomedical data
  - electronic health record
  - -omics
  - digital health

Computer Vision:
Real-Time Object Detection (& segmentation)

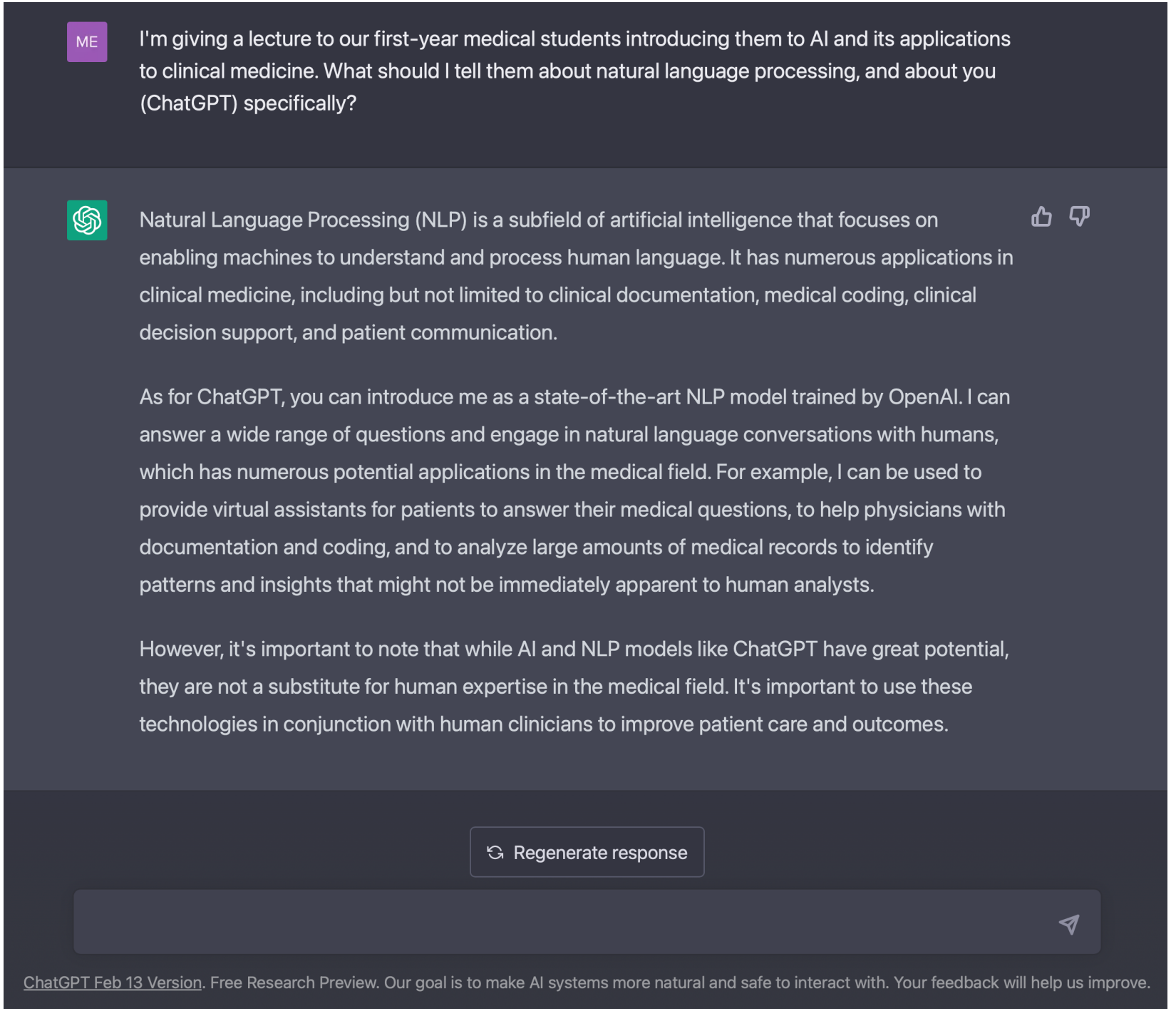


Real-time polyp detection:

a second set of eyes during colonoscopy



NLP: Text Generation



Text Generation via GPT-3 (ChatGPT precursor)

SYSTEM PROMPT (HUMAN-WRITTEN)

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

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials. The Nuclear Regulatory Commission did not immediately release any information.

According to the release, the U.S. Department of Energy’s Office of Nuclear Material Safety and Security is leading that team’s investigation.

“The safety of people, the environment and the nation’s nuclear stockpile is our highest priority,” Hicks said. “We will get to the bottom of this and make no excuses.

“Better Language Models and Their Implications”
2/14/19
OPENAI

Natural Language Processing: Find Predictive Text in Notes

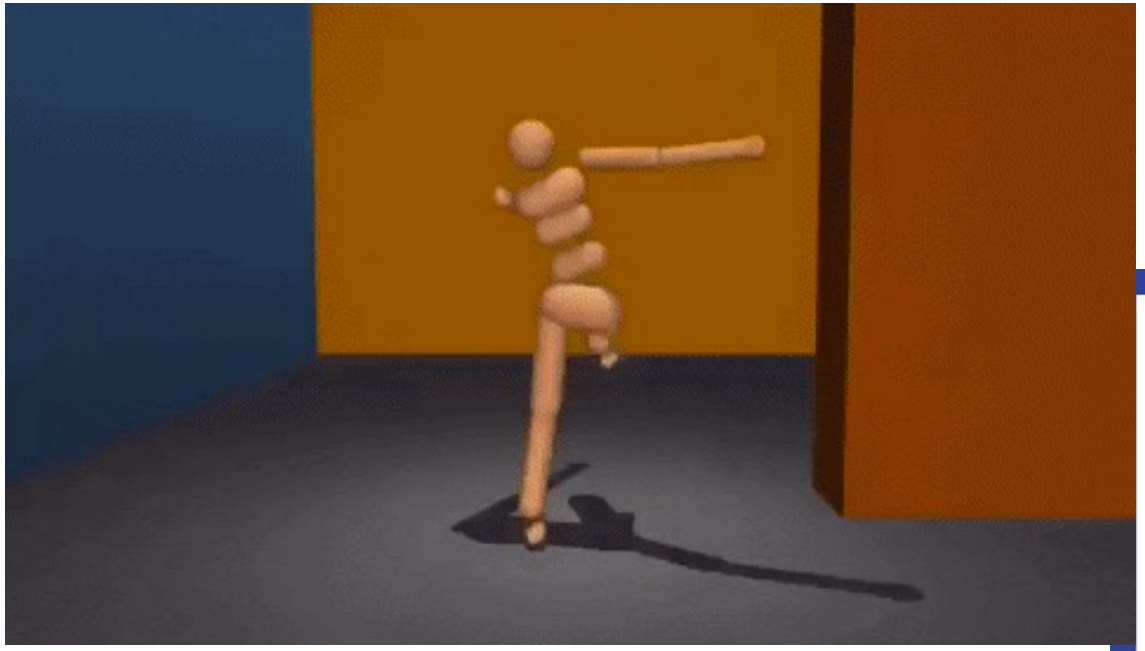
Passage (from note)	Change in predicted autism dx log-odds
subjective intake chief complaint problems with sleep, inattention, and behavioral concerns both in the home and school setting. DATE, recently more anger and recent tic like behavior	+6.95
psychologist presenting problem NAME is a 3 year, 4 month old female who was referred for a neurodevelopmental assessment due to concerns regarding her overall development, behavior, and social emotional functioning and to assess for autism spectrum disorder	+6.82
problem list diagnosis • disruptive behavior disorder • impaired speech articulation • daytime enuresis • other subjective visual disturbances • hypermetropia of both eyes • adhd attention deficit	+6.81
problem list diagnosis • anemia of prematurity • history of colitis • meconium tox for thc • extreme immaturity of newborn, 27 completed weeks • nasal congestion of newborn • presumed	+6.78
motor delay DATE • hypotonia DATE • clasped thumb DATE • polydactyly DATE • developmental	+6.74
therapy NAME was seen for developmental support during rop eye exam today. the	+6.65

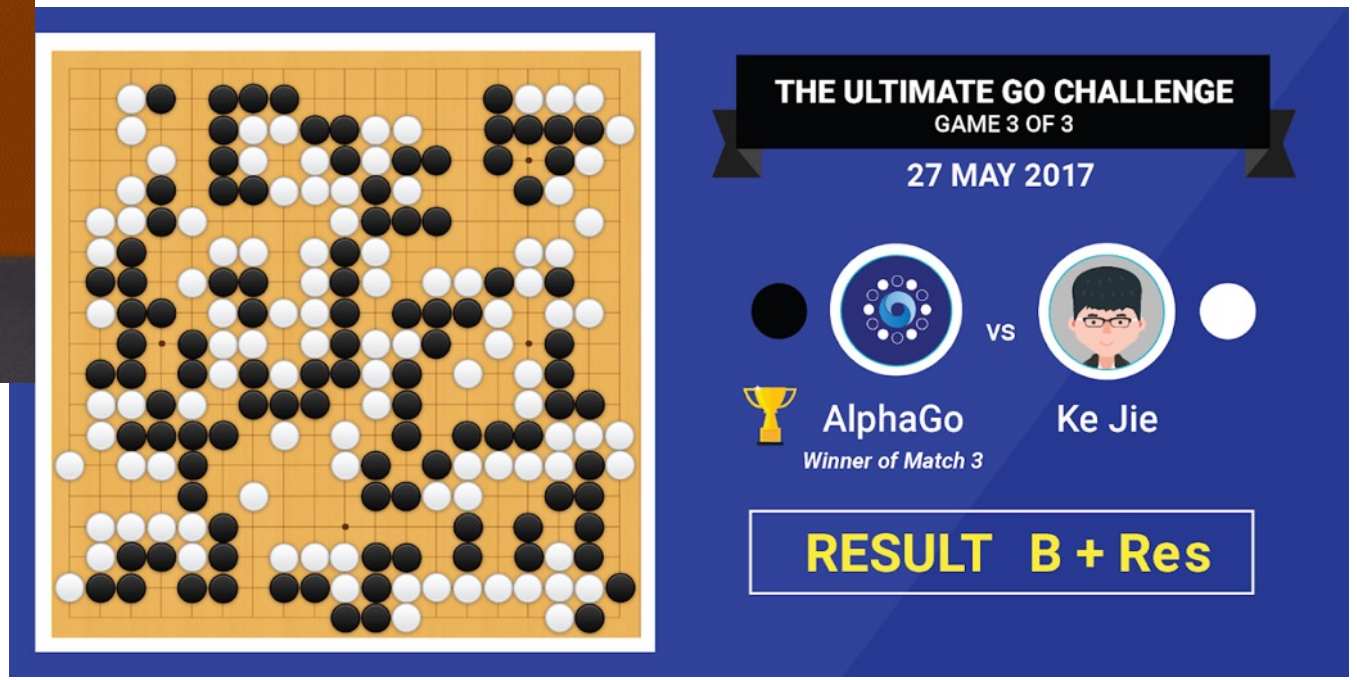
Developmental and behavioral concerns are highly predictive

Premature birth and perinatal complications are also highly predictive

Subramanian V, Engelhard MM, Berchuck SI, Chen L, Carin L. SpanPredict: Extraction of Predictive Document Spans with Neural Attention. Submitted to NAACL.

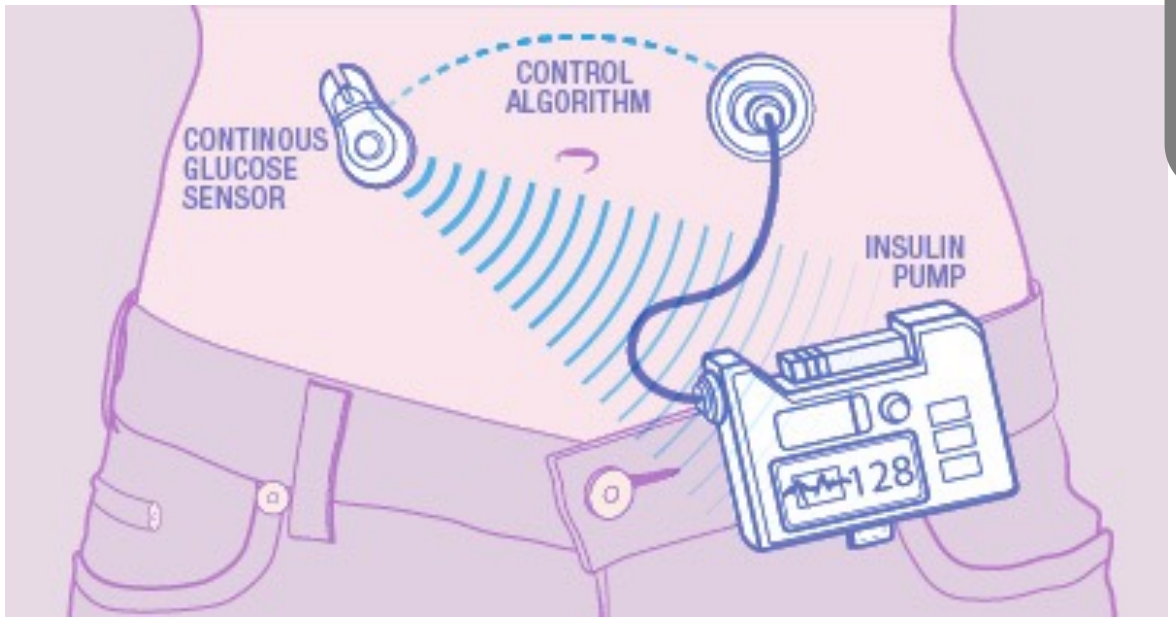
Reinforcement Learning:
goal-directed sequential decision-making



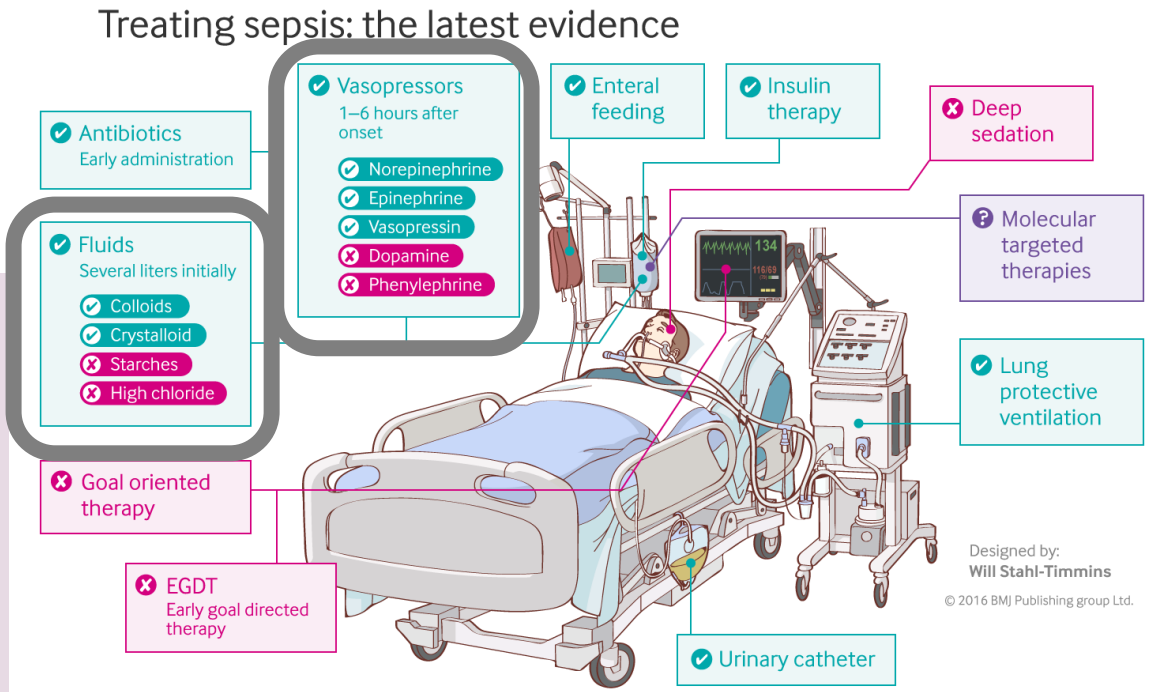


Reinforcement Learning in Medicine

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



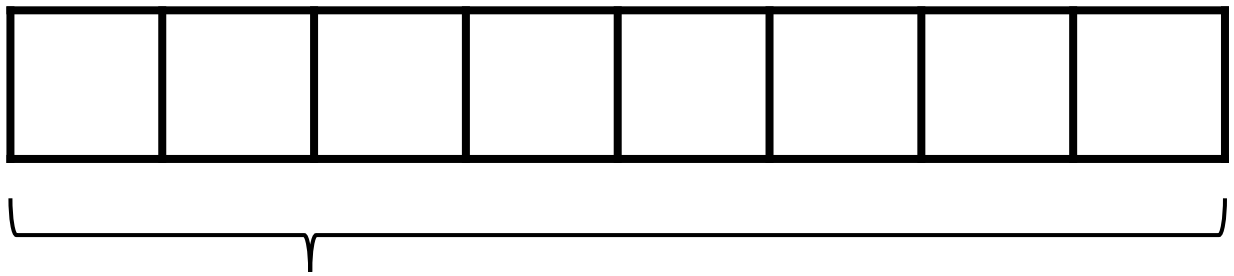
https://www.mayo.edu/research/labs/artificial-pancreas/overview



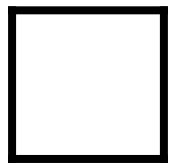
Fluid and vasopressor administration for sepsis treatment

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

All of these have, at their core, a predictive model



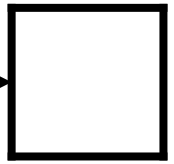
x, data/features for
a subject or patient



y, associated
value or label

End goal: predict y from x

Computer Vision:
prediction via convolutional neural network

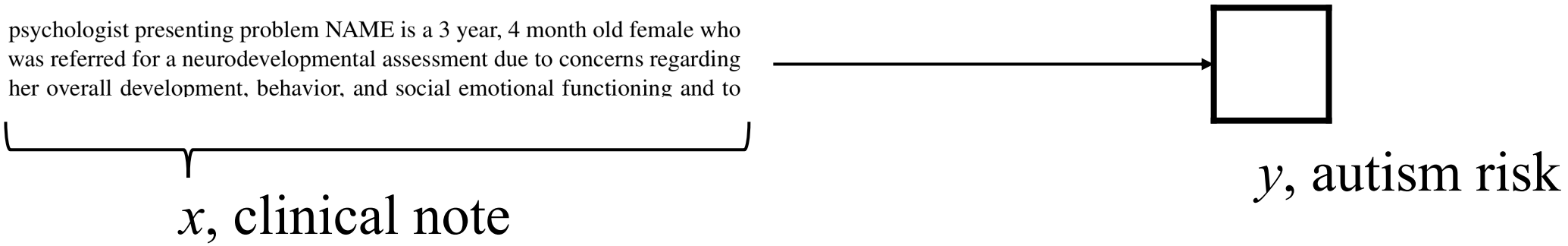


y, referrable diabetic retinopathy

x, retinal image

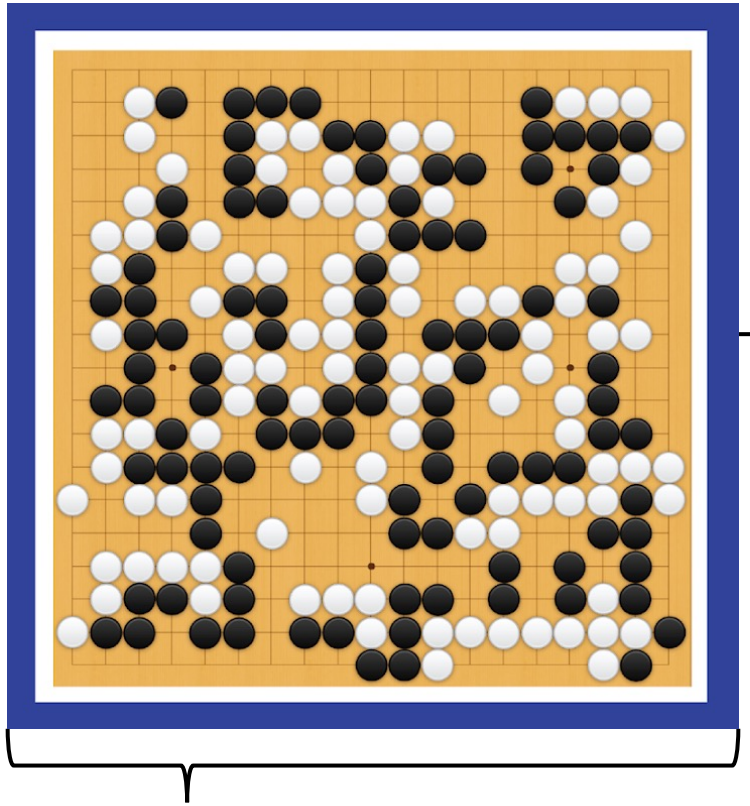
End goal: predict y from x

Natural Language Processing: prediction via stacked attention networks

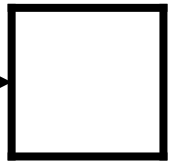


End goal: predict y from x

Reinforcement Learning:
state to action predictions via multilayer perceptron



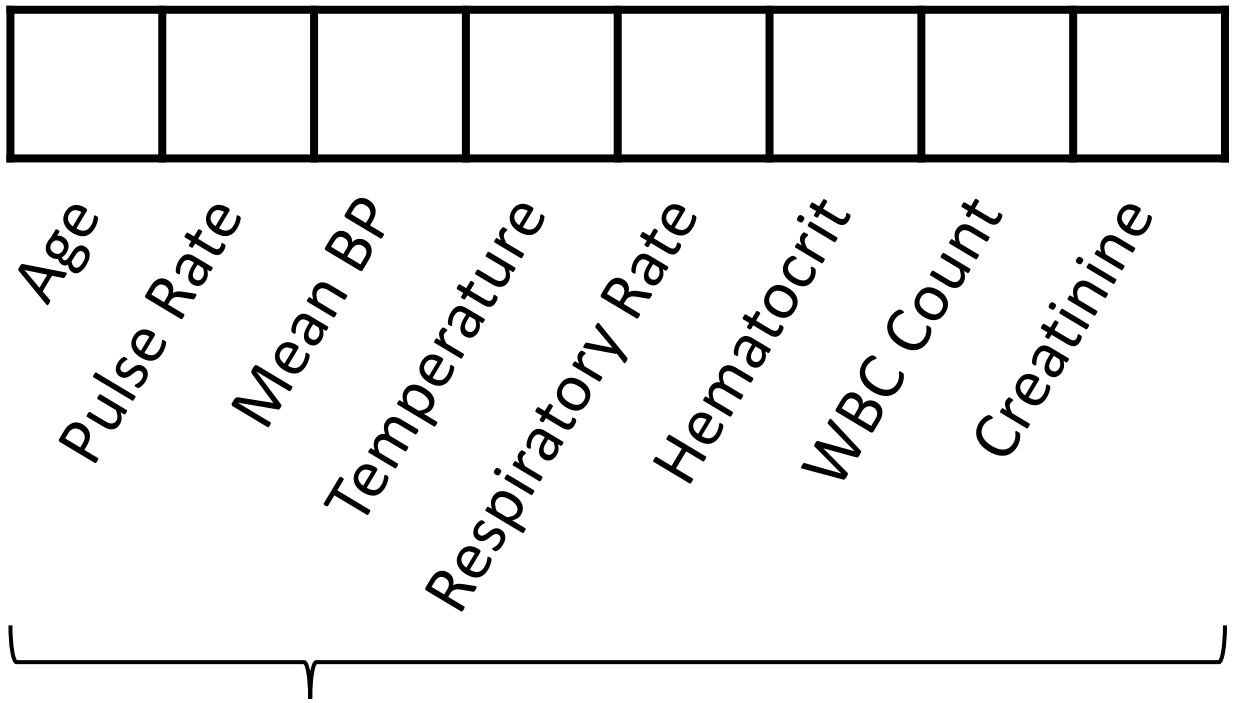
x, Go board state



y, next move

End goal: predict y from x

APACHE III: A model for predicting ICU mortality

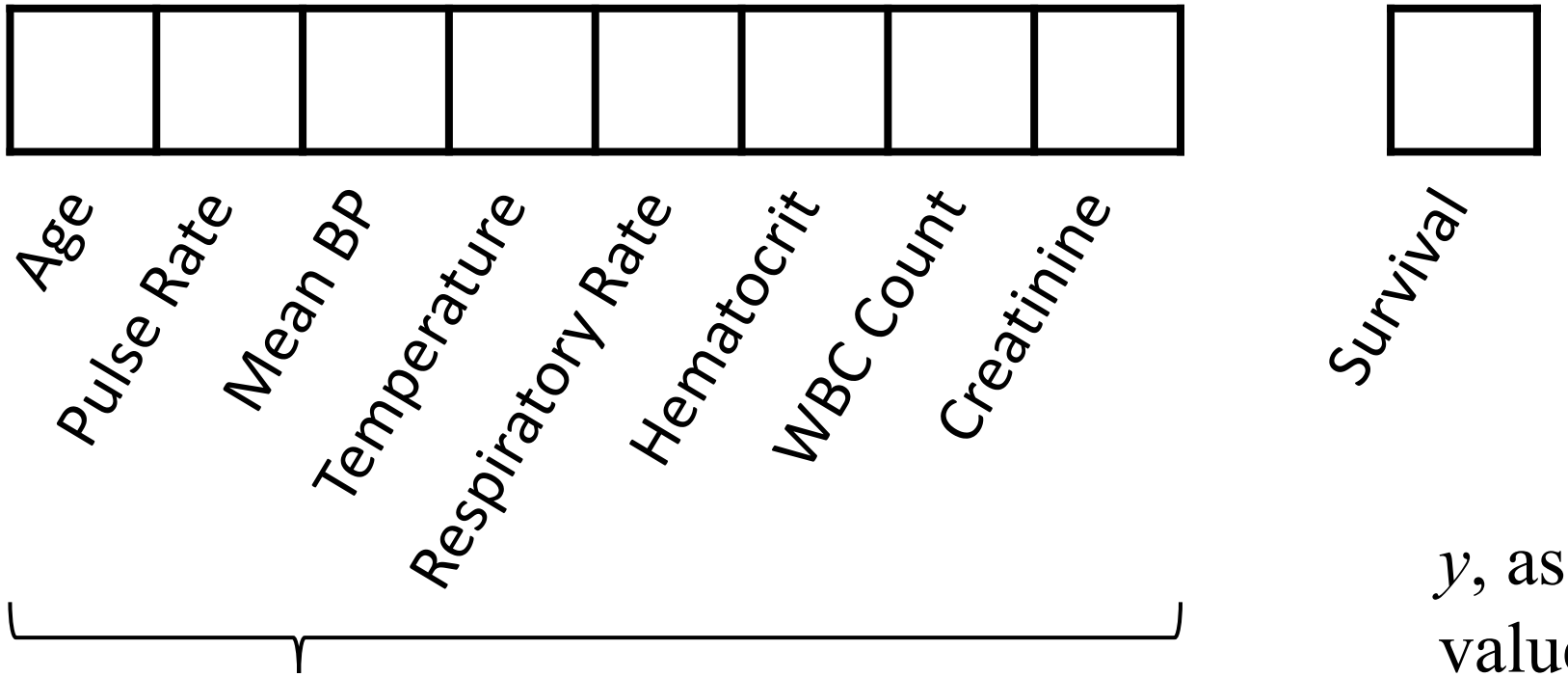


x, data/features for a subject or patient

y, associated value or label

End goal: predict odds of hospital mortality

Traditional models (e.g. logistic regression) tend to work well for data like this.

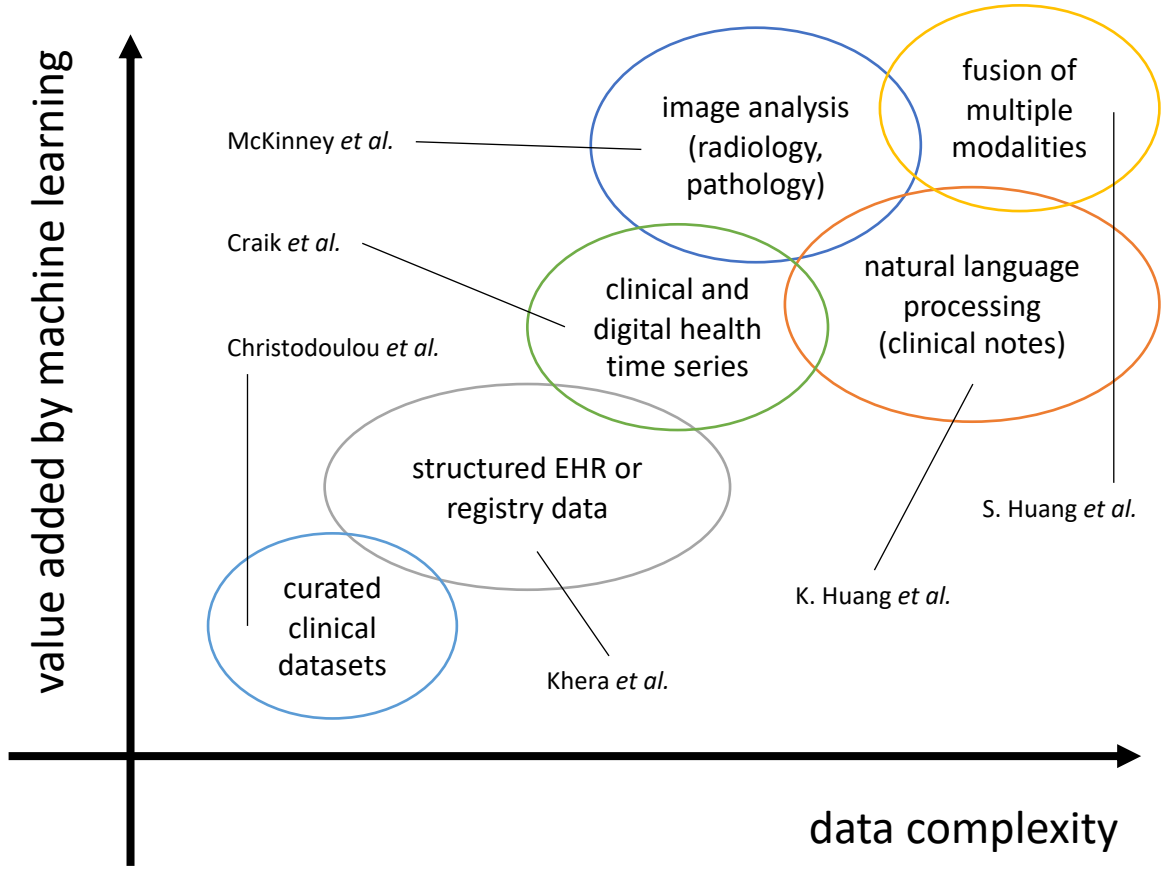


x, data/features for a subject or patient

y, associated value or label

End goal: predict odds of hospital mortality

As the complexity of the underlying data increases, so too does the probable value added by machine learning



Simple models based on clinical variables remain highly relevant
Use the simplest model that gets the job done

Course Logistics

- Let's take a look at the website
- Questions & discussion about course requirements, materials, or activities
- Contact me: m.engelhard@duke.edu