Topics in Machine Learning Machine Learning for Healthcare

Rahul G. Krishnan
Assistant Professor
Computer science & Laboratory Medicine and Pathobiology

Slide credits: David Sontag, Clinical Machine Learning Lab, MIT

Outline

- Introduction to Machine Learning for Healthcare [MLHC]
  - Why should we care
  - Why do we have data
  - A brief history of MLHC
  - What does the future hold?
  - A word of caution
  - Key challenges in MLHC
- Logistics
  - Course staff
  - Course structure
  - Grading
  - Mandatory quiz
  - Lecture schedule

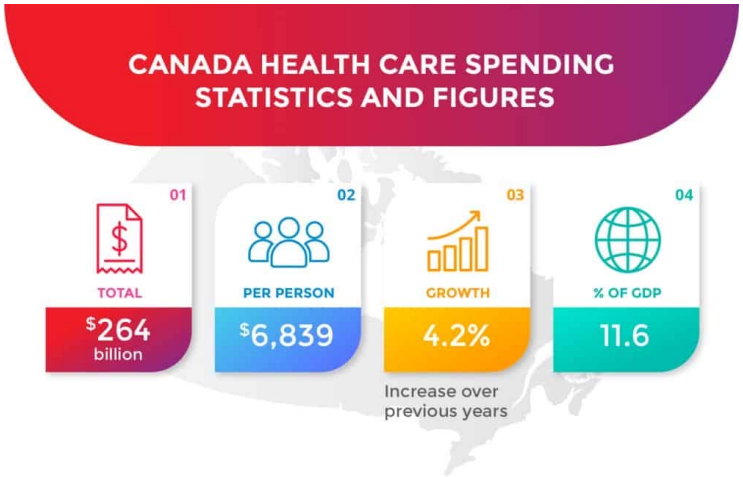
Course goals

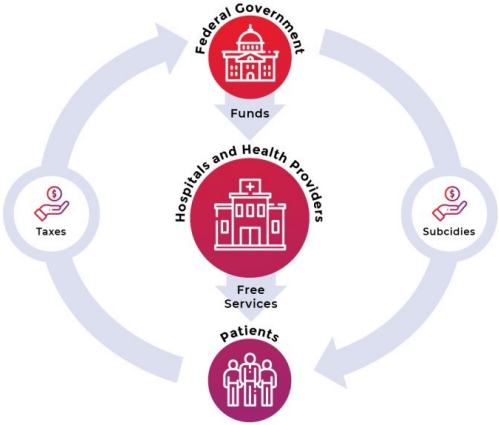
- Intuition for working with healthcare data
- Understand what problems are useful to solve, and what choices of models/learning algorithms to work with
- Appreciate subtleties in applying ML to healthcare problems
- Have fun working (and, if the course project goes well, publishing) in this space!

What is the problem with healthcare?

- Healthcare costs around the world are rising
  - People are living longer,
  - Chronic diseases are consuming clinician time,
- Canada:
  - Canadian Institute for Health Information (CIHI) publishes reports on healthcare costs
  - In 2019, Canada spent ~$264 billion on healthcare
- United States:
  - >65yo: Medicare
  - <65yo: Private/ public health insurance/ Medicare
  - Health expenditures exceed $3 trillion

Why should we care about costs?



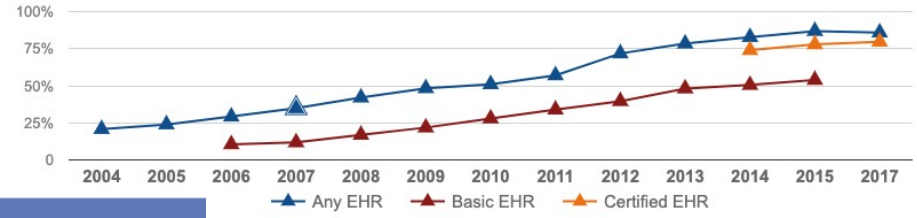


We’re paying more on healthcare and its increasing!

Image credit: [https://www.ephpp.ca/healthcare-funding-policy-in-canada/]

The opportunity with electronic medical records

The adoption of electronic medical records has dramatically increased in the last two decades!



Electronic medical records give us a view into a patient's underlying physiological state.

Source: https://www.healthit.gov/data/quickstats/office-based-physician-electronic-health-record-adoption

Hospitals, registries and clinics have an abundance of data

Machine learning

- **For the right tasks**, we could use the data to train machine learning algorithms.
- General recipe:
  - Identify a problem, which if automated, can reduce the cost of a process or help clinicians complete a task better/faster/with less error.
  - Program a model to automatically learn patterns from data
  - Use the model to automate task
- Different kinds of machine learning strategies we can make use of

Supervised learning

Patient features

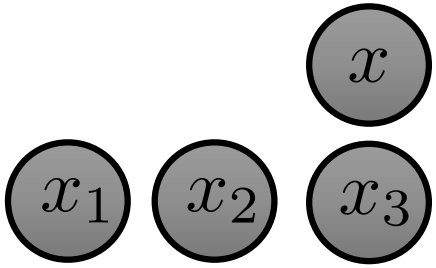
Chest X-ray image

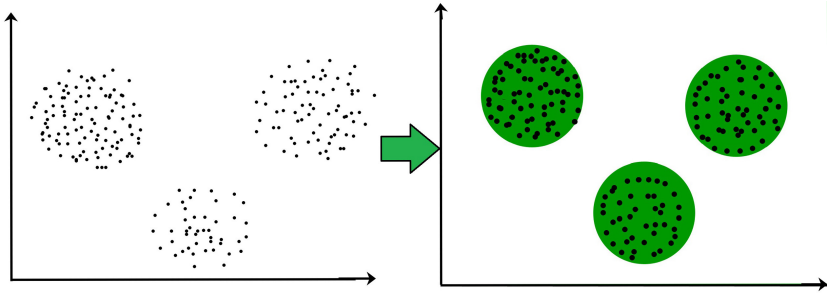
Time to readmission

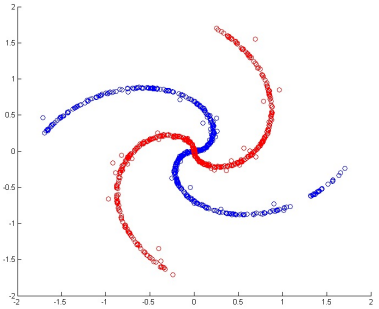
pneumonia/pneumothorax

- Step 1: Collect a dataset or curate a subset of data with labels from an existing dataset
- Step 2: Learn the model using the dataset
- Step 3: Use the output of the model to build software to help clinicians reach better decisions, faster.
- **Examples**: Logistic regression, random forests, XGBoost, Deep neural networks

Unsupervised learning

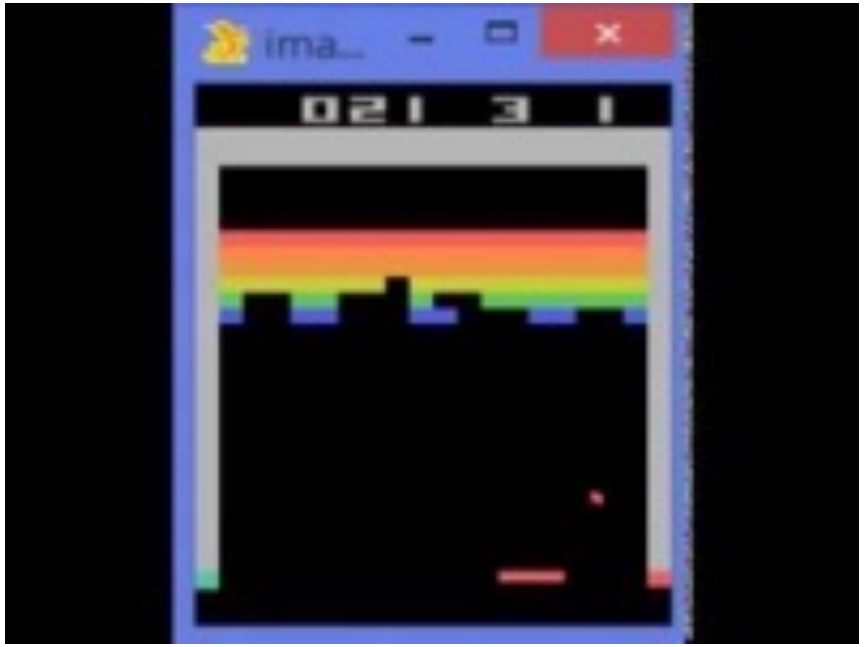






- Step 1: Collect a dataset or curate a subset of data with labels from an existing dataset
- Step 2: Learn the model using the dataset
- Step 3: Use parameters of the model uncover insights about the data and validate with domain experts
- **Examples**: Nearest neighbors, latent factor models, hidden markov models, variational autoencoders

RL in healthcare



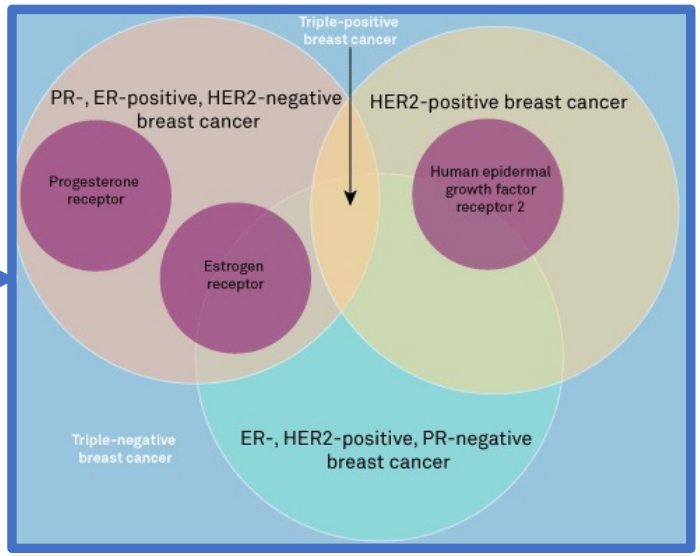
- On the left is an example of Deepmind's ATARI RL agent that learns to move the paddle at the bottom
- Can we use similar techniques for problems in healthcare such as developing strategies to treat people?
- **Challenge**: Difficult to build good simulators of how the human body will react to drugs

Reinforcement learning from observational data

- Step 1: Collect a longitudinal dataset of patient states and clinician actions
- Step 2: Learn an off-policy model using the observational dataset
- Step 3: Use the model to suggest what action to take for a new patient state
- **Examples**: Marginalized Off-Policy Evaluation

What can we do with data?

Subtype discovery



Build clinical tools

A brief history of machine learning/AI and medicine

This seems like an obvious idea – hasn't this been done before?

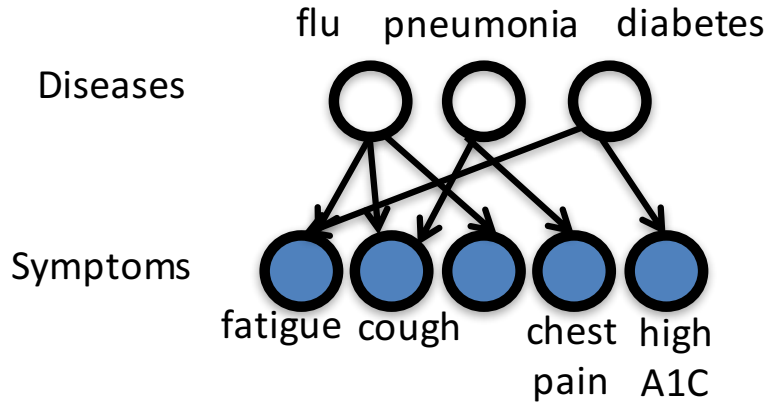
1978: Mycin expert system at Stanford

- Early expert system/AI to diagnose patients based on symptoms and test results
- Used >500 prediction rules:
  - If A & B then predict pneumonia
- Worked better than specialists in blood infections and better than general practitioners

1986 : INTERNIST-1/QUICK MEDICAL REFERENCE (QMR) Project

- Automated diagnosis for internal medicine
- Probabilistic model:
  - hundreds of disease variables,
  - thousands of symptom variables
  - >40000 directed edges between them

The creation of this model led to several advancements in probabilistic inference!



1990s: Neural networks in clinical medicine

- Few features to make predictions with
- Data collected by chart review
- Did not generalize well to new places and difficult to fit into clinical workflow

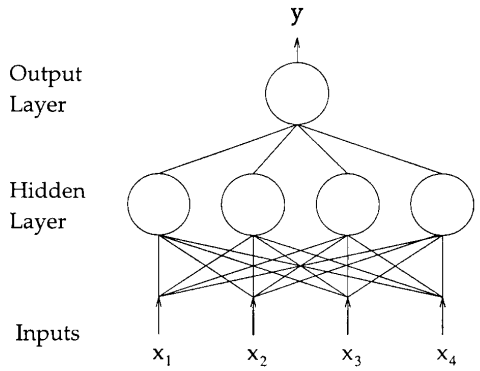


FIGURE 2. A multilayer perceptron. This is a two-layer perceptron with four inputs, four hidden units, and one output unit.

So why now?

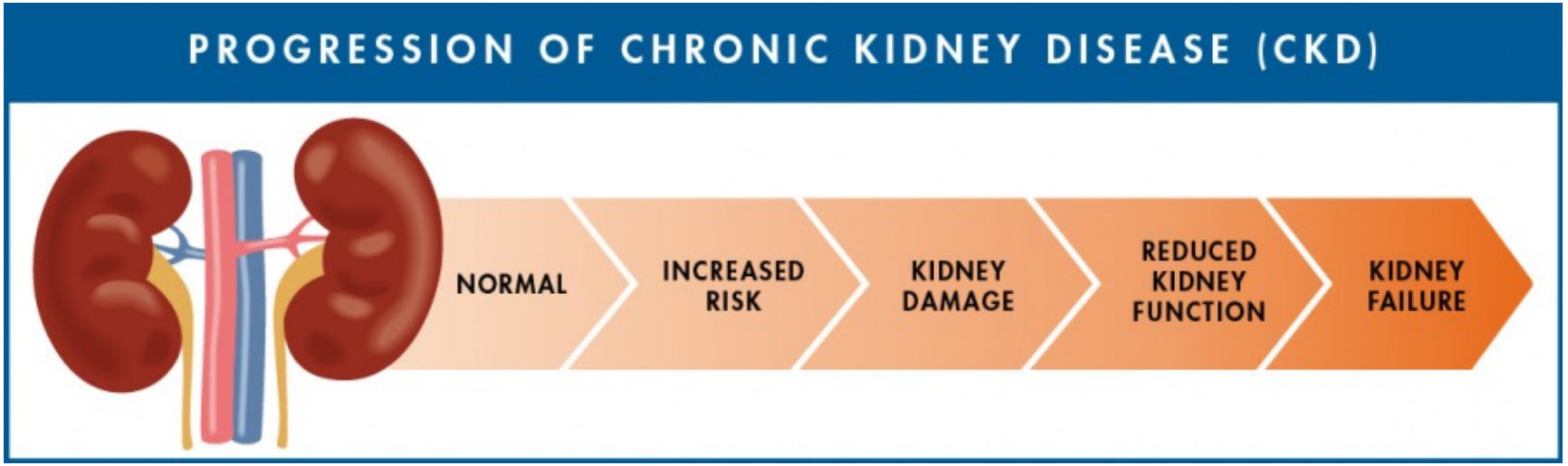
- Large datasets
  - Truven [bought by IBM] has data collected on 230 million patients since 1995
  - AllofUs precision medicine initiative: deep phenotyping of 1 million people in the US
  - GEMINI dataset
- Data standardization
  - FHIR, OHDSI
- Digital health funding
  - ~7B in venture funding in 2018
- Industry interest from Microsoft, Google, IBM

So why now – advances in machine learning!

- 1990s – AI winter, but a productive one!
  - Markov Chain Monte Carlo
  - Variational Inference
  - Convolutional neural networks
  - Reinforcement learning
- 2000s – Vision and NLP started adopting ML models
- 2013: Imagenet – watershed moment for deep learning
- 2018-now:
  - Photorealistic GANs
  - GPT-3 can simulate text indistinguishable from text written by humans

Staging diseases

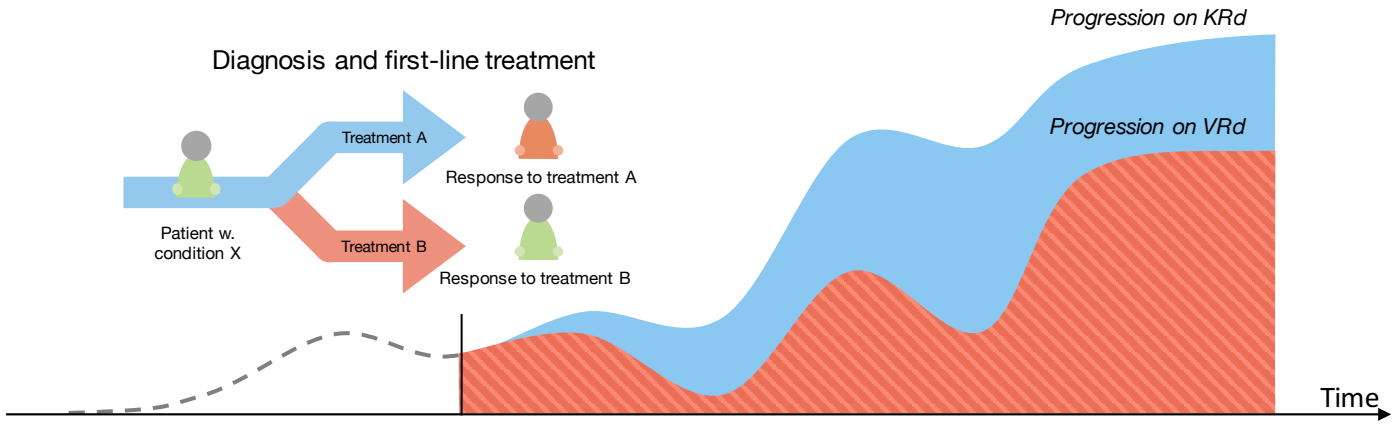
Using machine learning to uncover stages of disease progression



Precision oncology

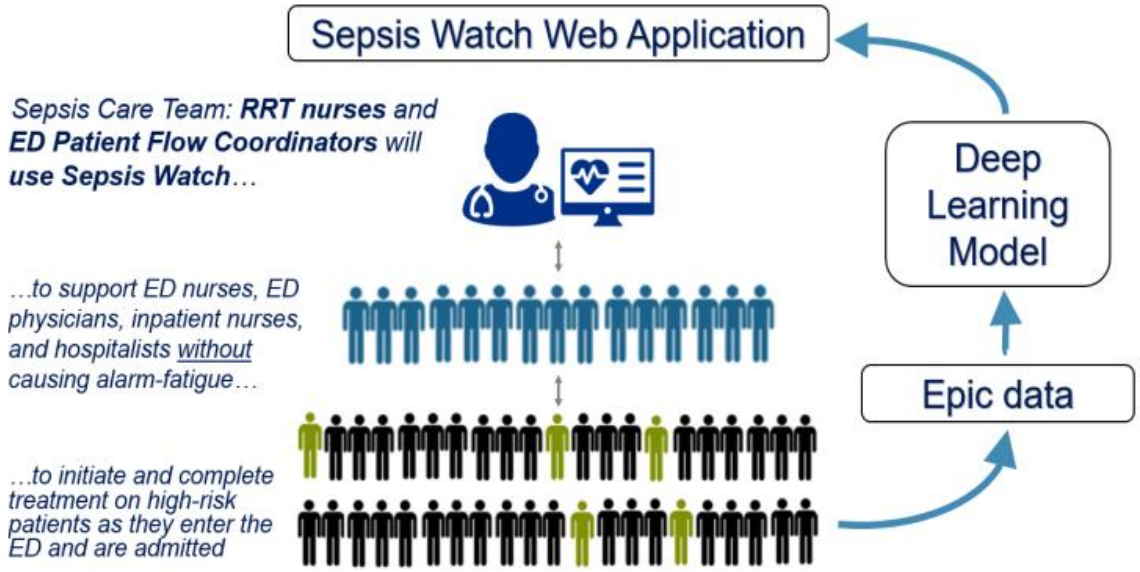
Using machine learning to guide treatment decisions for cancers therapy

A) KRd: carfilzomib-lenalidomide-dexamethasone, **B) VRd:** bortezomib-lenalidomide-dexamethasone



Predicting adverse outcomes in critical care patients

- **42,000+ inpatient encounters analyzed** at Duke Hospital over 14 months, **21.3%** with a sepsis event.
- **32+ million data points incorporated**: 25 million vital sign measurements, 2 million med admins, 5.2 million labs.
- **34** physiological variables (5 vitals, 29 labs).
  - At least one value for each vital in 99% of encounters.
  - Some labs rarely measured (2-4%), most measured 20-80% of the time.
- **35** baseline covariates (e.g. age, transfer status, comorbidities).
- **10** medication classes (antibiotics, opioids, heparins).

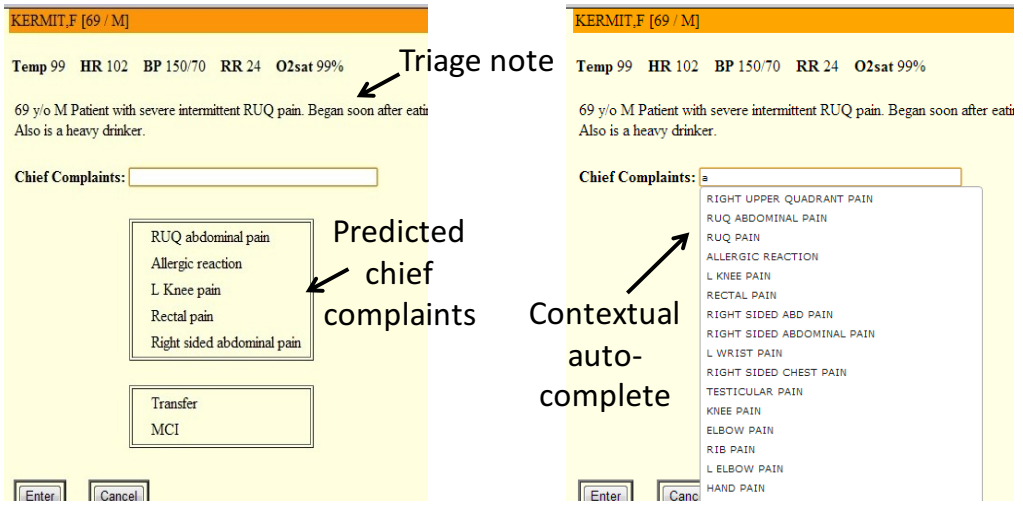


Source: https://dihi.org/wp-content/uploads/2020/02/Sepsis-Watch-One-Pager.pdf

A smart EHR system

The Burden and Burnout in Documenting Patient Care: An Integrative Literature Review

The surge of EHRs has had an unintended consequence : an increase in physician administrative load



Many more applications

- Drug discovery for faster, cheaper drug development pipelines
- Automating polyp detection in gasterointestinal diseases

- New and upcoming places for machine learning to have an impact in healthcare:
  - Microbiome
  - Liquid biopsies for cancer detection and tracking

Should we all be doing this?

HOLD YOUR HORSES

SLOW DOWN AND THINK

MAIA SZALAVITZ BACKCHANNEL 08.11.2021 06:00 AM

The Pain Was Unbearable. So Why Did Doctors Turn Her Away?

A sweeping drug addiction risk algorithm has become central to how the US handles the opioid crisis. It may only be making the crisis worse.

- "NarxCare also offers states access to a complex machine-learning product that automatically assigns each patient a unique, comprehensive Overdose Risk Score"
- Source: https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/

HEALTH

AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind

Machine learning has the potential to save thousands of people from skin cancer each year—while putting others at greater risk.

By Angela Lashbrook

Source: https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/

Challenges for machine learning in healthcare

- Challenging risk/reward ratios
  - Why: In healthcare, clinicians make life or death decisions
  - What do we need:
    - Algorithm development should proceed with caution and care
    - Need **robust** algorithms with checks and balances
    - Algorithms need to be **fair** and **accountable**
- Labelled data is scarce
  - Why: Clinician time is expensive
  - What do we need:
    - New methods for unsupervised and semi-supervised learning

Challenges for machine learning in healthcare

- Patient populations are different:
  - **Why**: Each individual is unique and people from Mumbai display different clinical phenotypes than those in Toronto
  - What do we need:
    - New methods for transfer learning so that models generalize well across different hospitals

- Missingness
  - **Why:** We only go to the doctor/clinician/hospital when we are sick; hospital administrators may forget to annotate data, records can go missing
  - What do we need:
    - Machine learning models that can make robust predictions even when data is missing

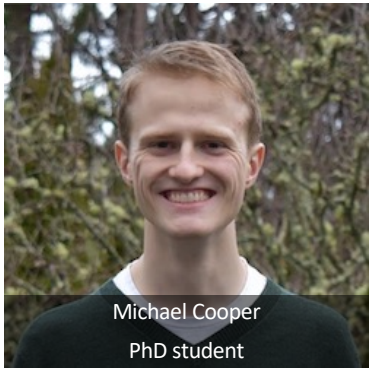
Challenges for machine learning in healthcare

- Data silos
  - **Why:** Patient data
  - What do we need:
- Deploying ML software in the clinic
  - **Why:**
  - What do we need:

Course map

Course staff

Office hours forthcoming



A thematic review of machine learning for healthcare

- Week 2: Supervised learning and survival analysis
- Week 3: Time series modeling
- Week 4: Disease progression
- Week 5: Clinical Natural Language Processing
- Week 6: Medical Imaging
- Week 7: Interpretability
- Week 8: Fairness
- Week 9: Clinical trials
- Week 11: Missingness and causality

Thematic readings

- Each week will have several readings in the course schedule for methodological and applied papers
- **Prerequisite**: Reading and digesting the papers for this class will require a strong foundation in probability, statistics, probabilistic graphical models and a variety of topics in machine learning.
  - Unless you have gotten prior approval from me, please make sure you are comfortable with advanced topics in machine learning
  - Do the readings early in the week, they will make your life easier
- **TODO:** The

Course announcements

- The course announcements will be posted to Quercus
- The course website will be your one-stop shop for information on the course: https://csc2541hf-2021.github.io/

Grades

- Individual
  - 5% class participation (attendance and engagement)
  - 15% assignment
    - Paper deconstruction: Summarize four papers of your choosing: highlight the key ideas, what makes them tick, why you think they work and how they could be improved
  - 15% paper presentation: Present (in pairs) one of the papers from the theme of the week, papers will be assigned on a first come first serve basis,
- Groups
  - 10% project proposal
  - 15% project presentation
  - 40% course project report

Course project

- Undertake a course project where to goal is to create a workshop abstract by the end of the semester
- You are free to use your own healthcare data (should you have access to it). In the next few weeks we will go through and describe several different publicly available datasets that you can apply for access to for your project.
- **IMPORTANT: Form groups and apply for access to projects early! Getting access to healthcare data can take a few weeks and it is important to get started on this now!**

Ethics training

- It is vital to understand and respect clinical data!
- This is data that may look like numbers and figures to you, but always remember that behind them is a real human being, respect their choice to share it and treat the data with care,
- Do not share the data with anyone who is not credentialed to have access to it.
- Never try to re-identify de-identified data

- **CITI training**: https://physionet.org/about/citi-course/

Course project

- Groups of **atleast two** and **no more than four** people
- The grading rubric will not depend on the number of people contributing to a project
- Friday lectures in the next few weeks will introduce you to freely available clinical datasets

Course timeline

- Week 1-3
  - Lectures
- Week 4-8
  - Guest lectures from researchers working on a diverse array of applications of machine learning to healthcare
  - Student presentations
- Week 9, 11, 12, 13, 14
  - Student presentations, project presentations

Key times for deliverables

- **TODO:**
  - **Complete CITI training for Physionet**
  - **Complete prerequisite quiz**
  - **Form project groups and partners for paper presentations**
  - **Start brainstorming ideas of problems you may want to explore**
- Week 4: Project proposal due
- Week 8: Paper assignment due
- Week 13: Group presentations begin
- Week 14: Project report due

Lecture in a slide

- Why take this course?
- What are the key challenges in machine learning – review this as you begin to think about your course project