

Topics in Machine Learning

Machine Learning for Healthcare

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



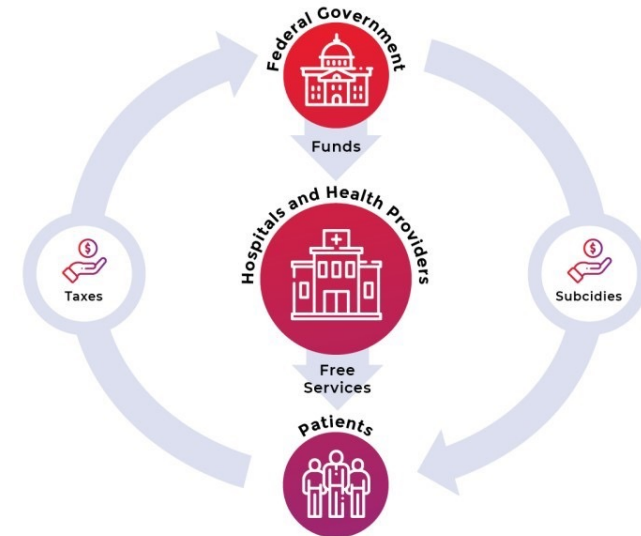
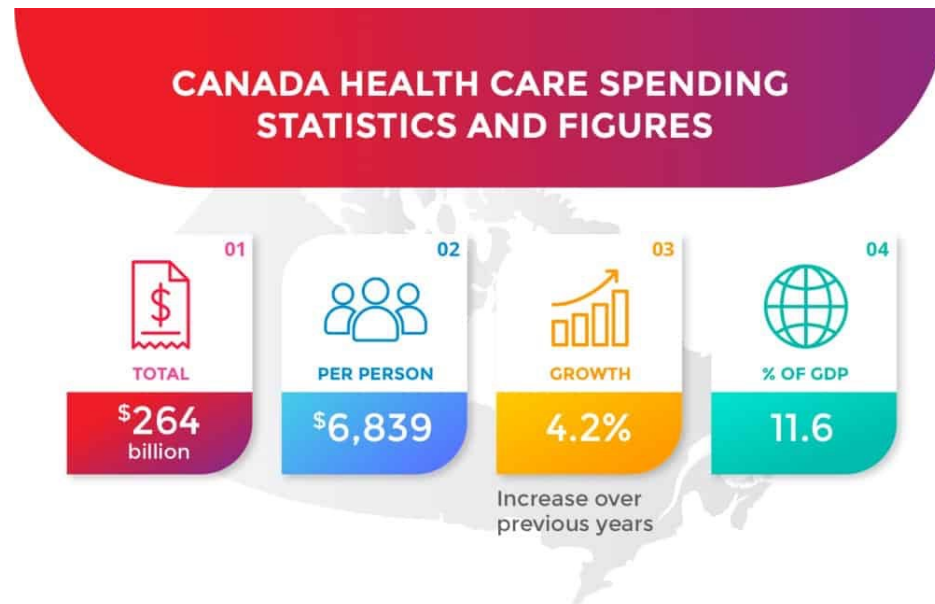
Outline

- 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

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

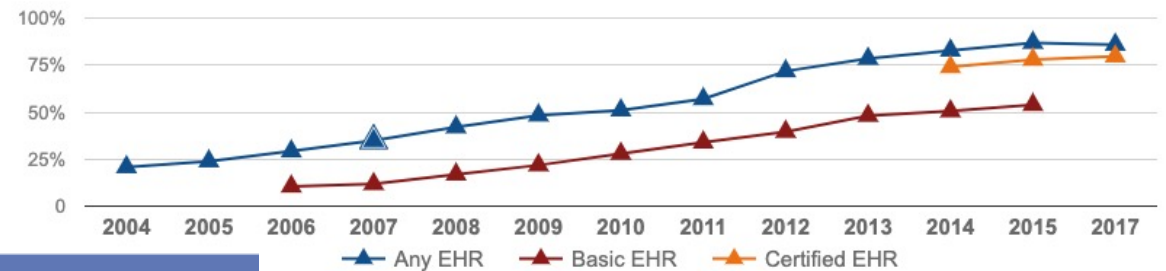
Why should we care about costs?



We're paying more on healthcare and its increasing!

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.

Hospitals, registries and clinics have an abundance
of data



Machine learning

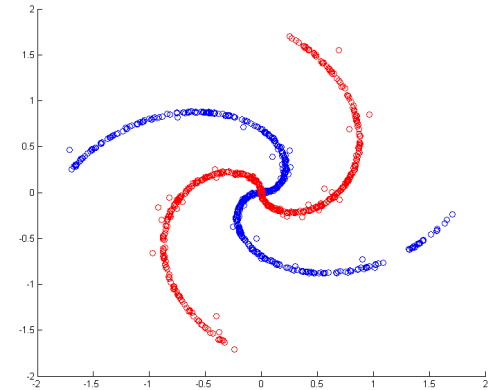
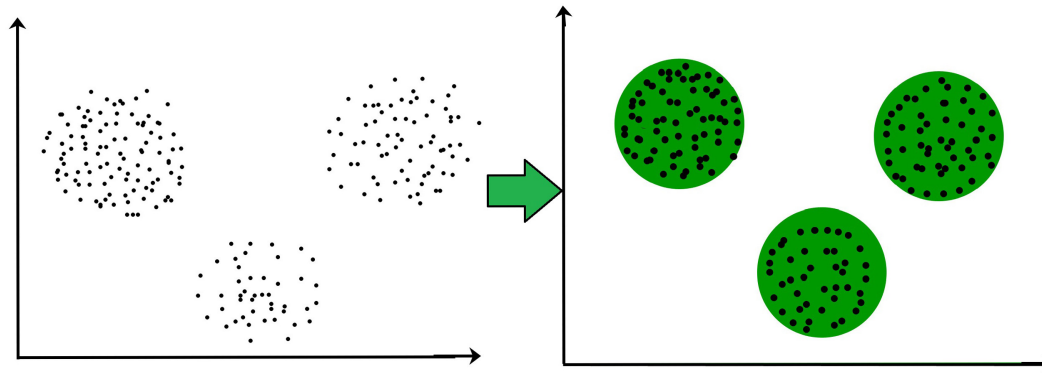
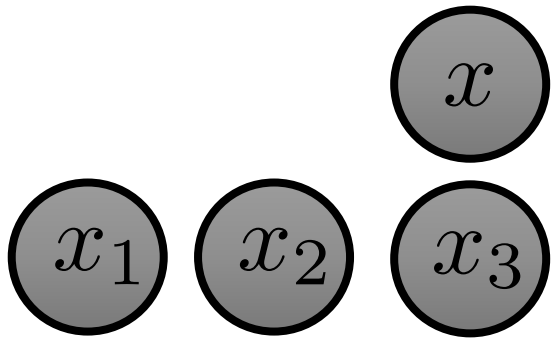
- **For the right tasks**, we could use the data to train machine learning algorithms.
- General recipe:
 - Identify a problem, which if automated, can reduce the cost of a process or help clinicians complete a task better/faster/with less error.
 - Program a model to automatically learn patterns from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

Supervised learning



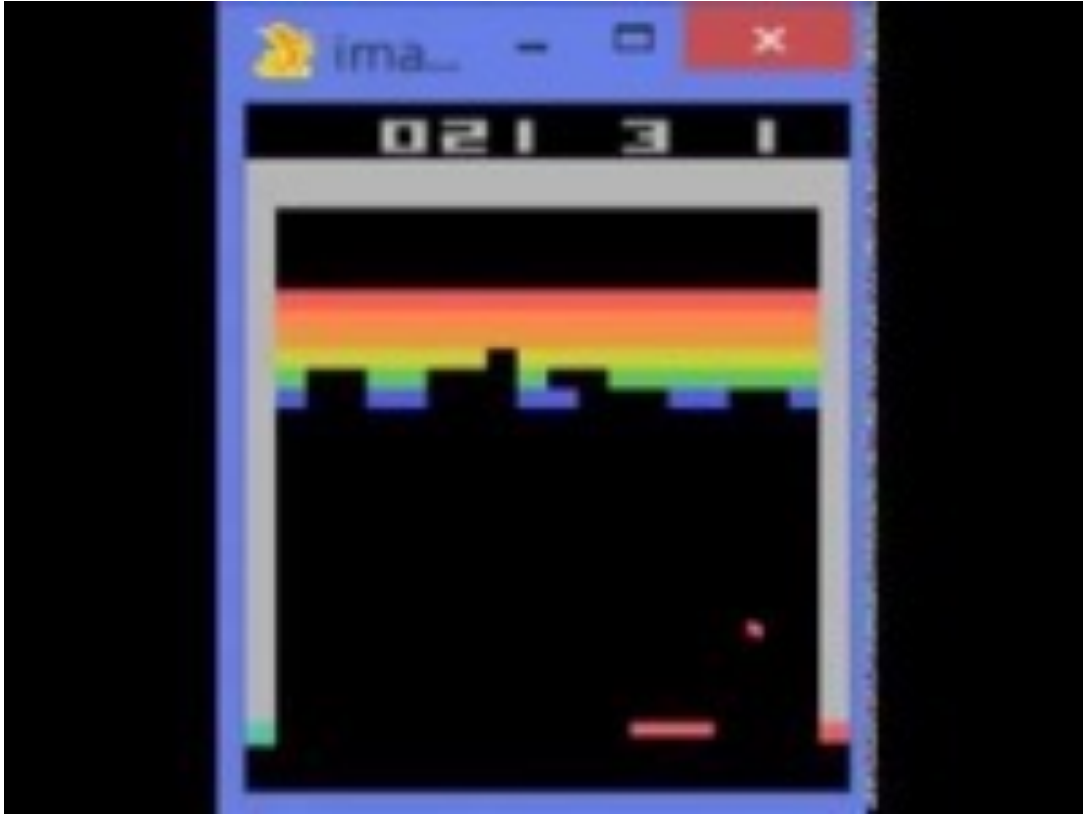
- Step 1: Collect a dataset or curate a subset of data with labels from an existing dataset
- Step 2: Learn the model using the dataset
- Step 3: Use the output of the model to build software to help clinicians reach better decisions, faster.
- **Examples:** Logistic regression, random forests, XGBoost,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

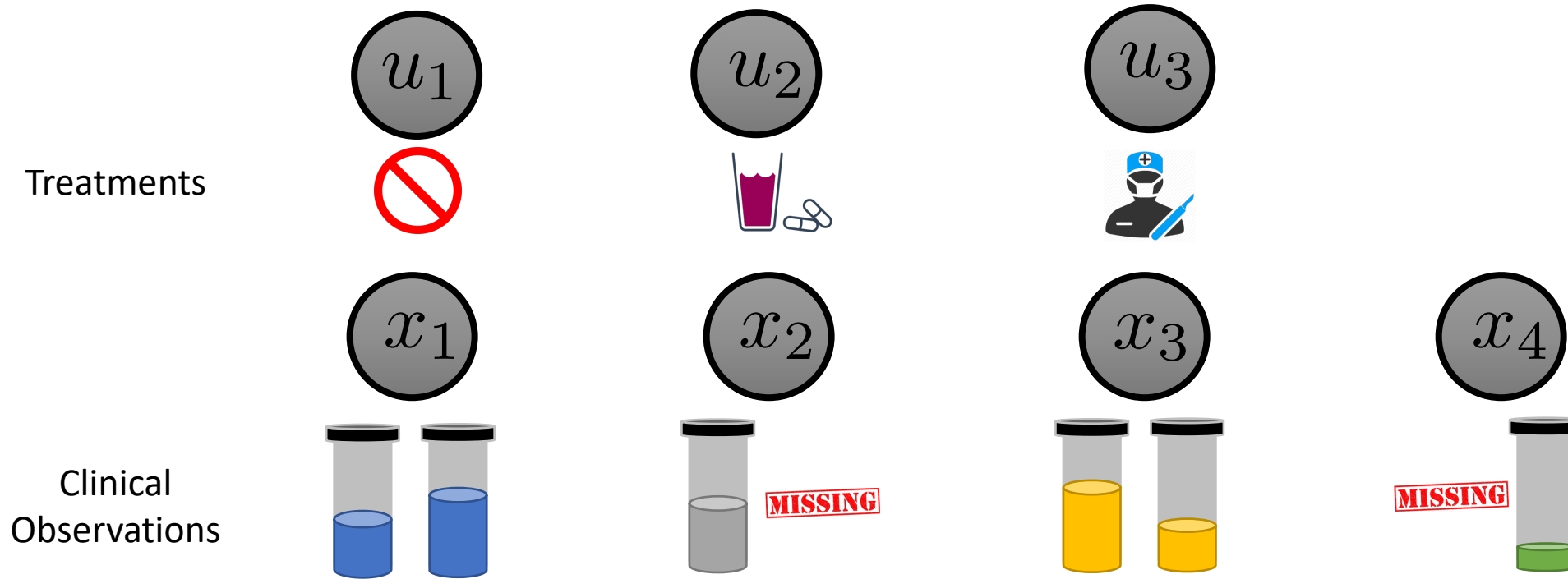
RL in healthcare



- 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

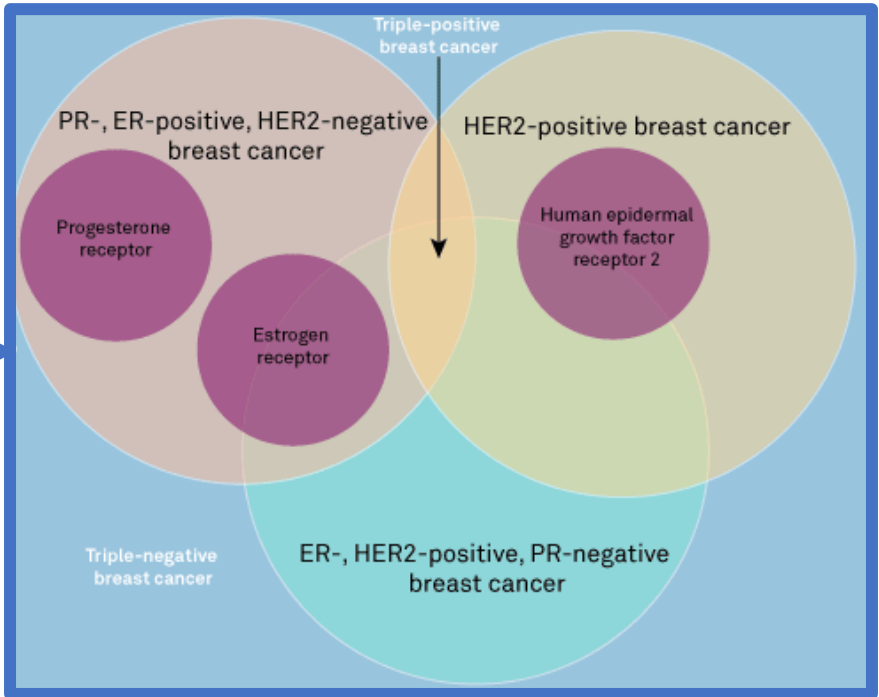


- Step 1: Collect a longitudinal dataset of patient states and clinician actions
- Step 2: Learn an off-policy model using the observational dataset
- Step 3: Use the model to suggest what action to take for a new patient state
- **Examples:** [Marginalized Off-Policy Evaluation](#)

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?



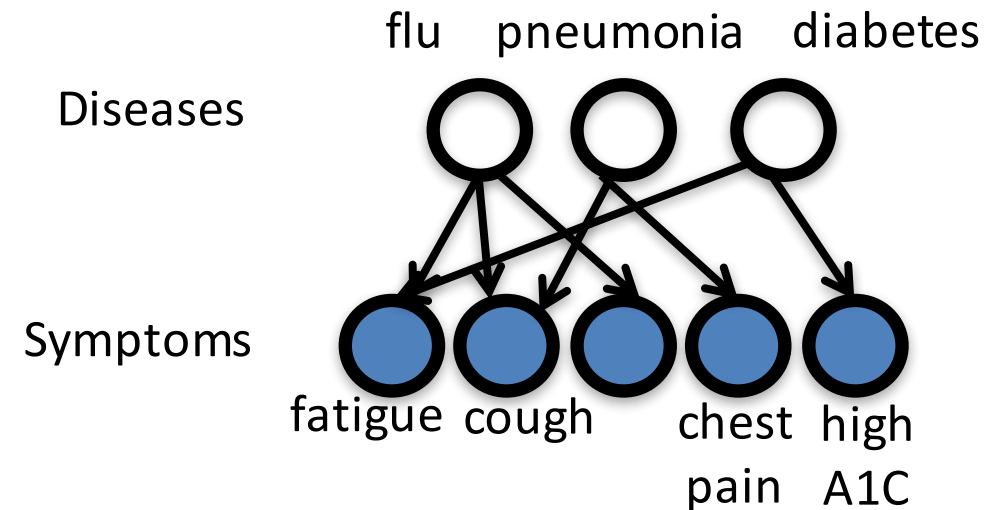
1978: Mycin expert system at Stanford

- 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

- Used very few features to make predictions with
- Data collected by chart review

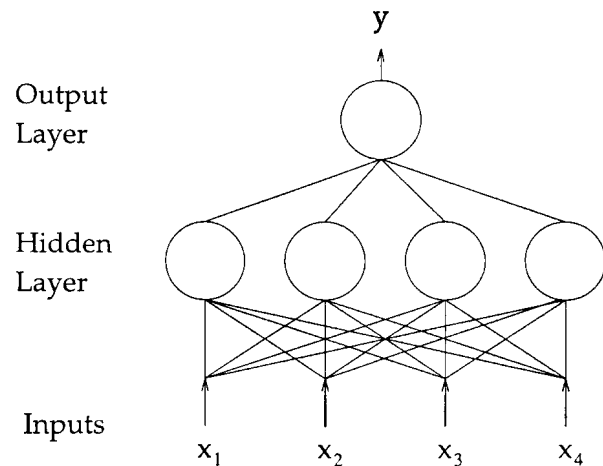


FIGURE 2. A multilayer perceptron. This is a two-layer perceptron with four inputs, four hidden units, and one output unit.

Did not generalize well to new places and difficult to fit into clinical workflow

So why now – better data?

- Large datasets
 - Truven [bought by IBM] has data collected on 230 million patients since 1995
 - All of Us 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
 - Midjourney can create synthetic looking videos

Staging diseases

Using machine learning to uncover stages of disease progression

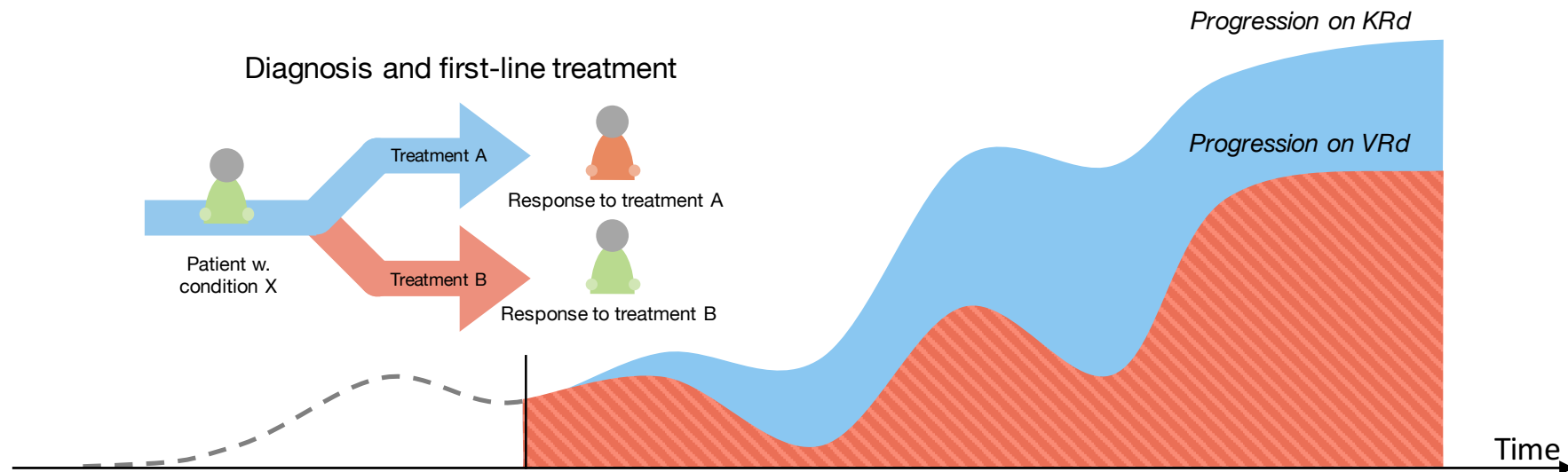
PROGRESSION OF CHRONIC KIDNEY DISEASE (CKD)



Precision oncology

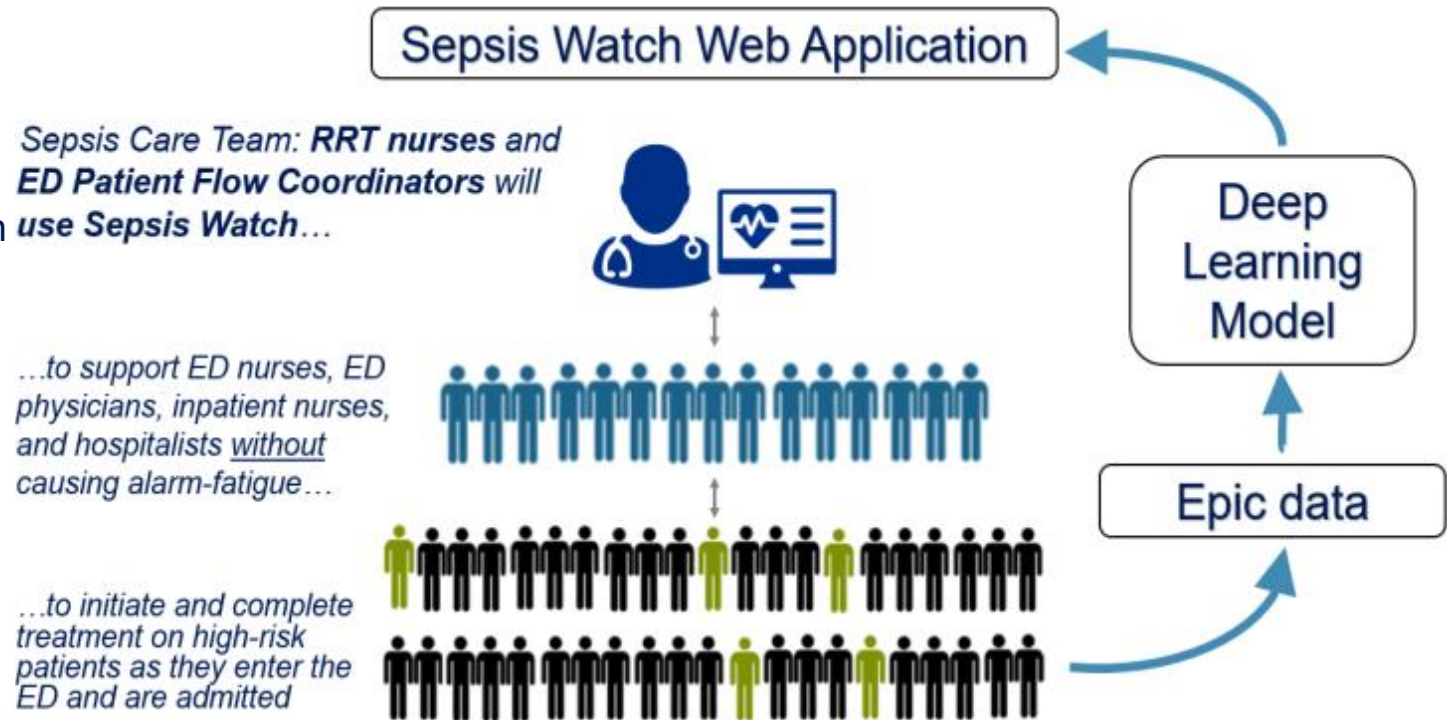
Using machine learning to guide treatment decisions for cancers therapy

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



Predicting sepsis in patients admitted to the hospital

- **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).



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

KERMIT,F [69 / M]

Temp 99 HR 102 BP 150/70 RR 24 O2sat 99%

69 y/o M Patient with severe intermittent RUQ pain. Began soon after eating.
Also is a heavy drinker.

Chief Complaints:

RUQ abdominal pain
Allergic reaction
L Knee pain
Rectal pain
Right sided abdominal pain

Transfer
MCI

Enter Cancel

Triage note

Predicted chief complaints

KERMIT,F [69 / M]

Temp 99 HR 102 BP 150/70 RR 24 O2sat 99%

69 y/o M Patient with severe intermittent RUQ pain. Began soon after eating.
Also is a heavy drinker.

Chief Complaints: a

RIGHT UPPER QUADRANT PAIN
RUQ ABDOMINAL PAIN
RUQ PAIN
ALLERGIC REACTION
L KNEE PAIN
RECTAL PAIN
RIGHT SIDED ABD PAIN
RIGHT SIDED ABDOMINAL PAIN
L WRIST PAIN
RIGHT SIDED CHEST PAIN
TESTICULAR PAIN
KNEE PAIN
ELBOW PAIN
RIB PAIN
L ELBOW PAIN
HAND PAIN

Enter Cancel

Contextual auto-complete

Many more applications

- Drug discovery for faster, cheaper drug development pipelines
- Automating polyp detection in gastrointestinal 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?



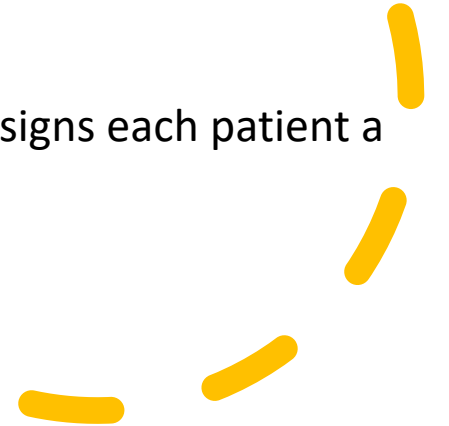
HOLD YOUR HORSES
SLOW DOWN AND THINK



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
 - Not all solutions are necessary, need to talk to stakeholders to find the ones that are worth solving
 - They may not be the problems you want to solve!

Challenges for machine learning in healthcare

- Patient populations are different:
 - **Why:** Everyone 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:** Countries have regulations (such as HIPAA) that require patient data to be kept private
 - What do we need:
 - New ideas in federated learning for institutions not comfortable with data-sharing
 - Automated methods for de-identification
- Deploying ML software in the clinic
 - **Why:** Machine learning models can stop working after a period of time
 - What do we need:
 - New techniques for lifelong learning
 - Ways to handle domain shift/covariate shift

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!



Course map

Course staff



Website: <https://csc2541-2023.github.io/schedule>
Important course announcements will be posted here + Quercus

A thematic review of machine learning for healthcare

- Week 1&2: Supervised learning and survival analysis
- Week 3: Project Planning
- Week 4: Time Series Modeling and Disease Progression
- Week 5: Clinical NLP/LLMs
- Week 6: Medical Imaging & Self Supervised Learning
- Week 7: Deployment
- Week 8: Fairness
- Week 9: Causality 1
- Week 11: Causality 2

Thematic readings

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

Using GPT4 as an aid

- Graduate seminar classes an excellent way to accelerate your learning by combining them with generative tools.
- Some ideas:
 - Use it to better understand background concepts in papers that you read,
 - Use it to accelerate development of software code – very useful for building dataloaders and helping accelerate model development,
 - Use it to understand research papers – useful for summarizing as well as paraphrasing ideas in terms that you might be more familiar with.

Grades

- Individual
 - 5% class participation (attendance and engagement)
 - 20% Paper summary 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'
- Groups of 3
 - 15% project proposal
 - 20% project presentation
 - 40% course project report

Course project

- Undertake a course project where the 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).
 - 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 **at least two** and **no more than four** people
- The grading rubric will not depend on the number of people contributing to a project
- Project planning lecture in two weeks will introduce you to freely available clinical datasets that you can use to brainstorm projects
- The goal is to tackle a meaningful problem using healthcare data:

TA and Instructor office hours

- Help brainstorm questions about the course project,
- Questions about machine learning methods in readings,
- Come chat about research, grad school coursework, industry in machine learning for healthcare

Course timeline

- Week 1-3
 - Lectures by myself
 - Week 3 is a project planning class, we'll have a lecture dedicated to potential ideas that one could work on during the course project.
- Week 4-11
 - Hour 1 – Lecture
 - Hour 2 - Guest lectures from researchers working on a diverse array of applications of machine learning to healthcare
- Week 12, 13, 14
 - Student project presentations

TODOs (for now and during the course)

- **TODO:**
 - Complete [CITI training for Physionet](#) (source for many healthcare datasets)
 - Join Discord! <https://discord.gg/C7kBerC5Aj>
 - Starting reaching out to 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



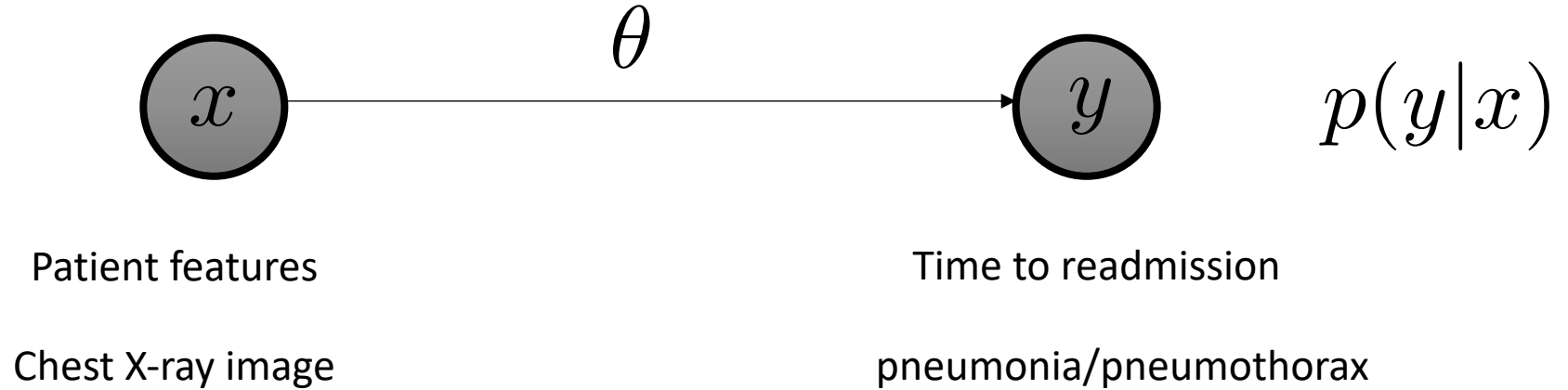
Outline

- Risk stratification: [35 minutes]
 - Stratification as a prediction problem
 - **Case study:** Predicting the onset of diabetes
- Summary and sneak peek of next week [7 mins]

Announcements

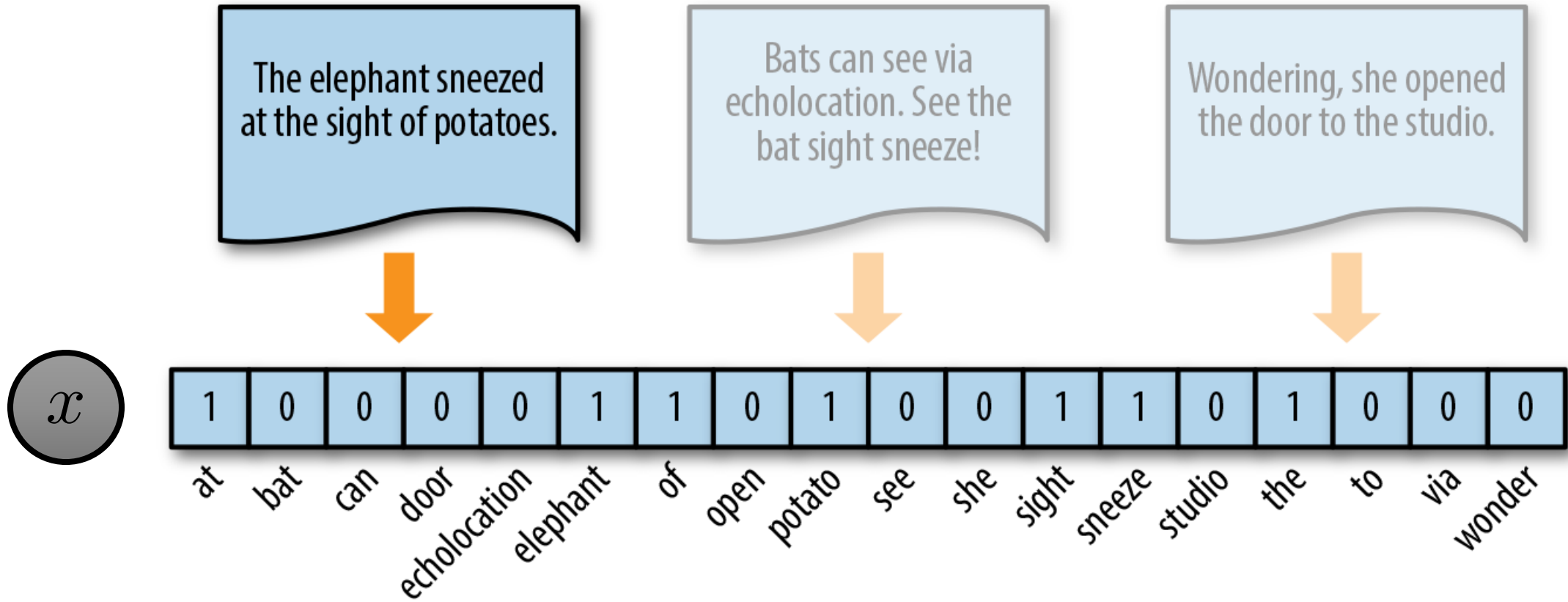
- Reading list for the first six weeks of class is now available
 - Please do the readings for the class ahead of the week, they will introduce you to research areas and the lectures during the week will provide more context for the theme
- Complete the quiz if you have not done so already

Supervised learning – (1)

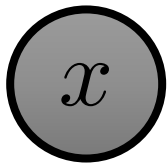


- Step 1: Collect a dataset or curate a subset of data with labels from an existing dataset
- Step 2: Learn the model using the dataset
- Step 3: Use the output of the model to build software to help clinicians reach better decisions, faster.
- **Examples:** Logistic regression, random forests, XGBoost,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

Vectorization – Text data



Vectorization – Clinical variables in tabular data



ICD10 - Diabetes	NDC code Metformin	CPT – Surgery, Aortic Valve
2	4	1

Vectorization – Image data

x



Pixel position [0,0]	...	Pixel position [32,15]	...	Pixel position [255,255]
1	...	0	...	1



Defining labels is challenging

- Examples:
 - Binary: Does the patient die or not
 - Real-valued: When does the patient die
 - Set-valued: The set of complications that a patient has
- Unlike domains such as computer vision, NLP, the true labels in healthcare can be **very** noisy
- We will discuss several kinds of noise in the upcoming lectures that require careful attention to detail

Supervised learning – (2)

- x : random variables
- y : outcome random variable
- θ : model parameters
- x typically high-dimensional [medical images, clinical variables]
- y (typically) low-dimensional [outcomes of interest]
- Model parameters depend on the functional class used:
 - Logistic regression: vector of weights
 - Decision tree: tree where each node is a feature to select and the value to threshold the feature on
 - Random Forest: collection of decision tree parameters

Supervised learning – (2)

x_1 y_1

x_2 y_2

x_3 y_3

Dataset (N=3)

- Given a dataset, the model parameters are learned via **maximum likelihood estimation**

$$\mathcal{L}(y, x) = \log p(y|x; \theta)$$

Score function (high is good, low is bad)

$$\theta = \arg \max_{\theta} \sum_{i=1}^N \mathcal{L}(y_i, x_i)$$

Solve this optimization problem to **learn** the model. Often formulated as a minimization of the negative of the log-likelihood function



Coarse-grained control

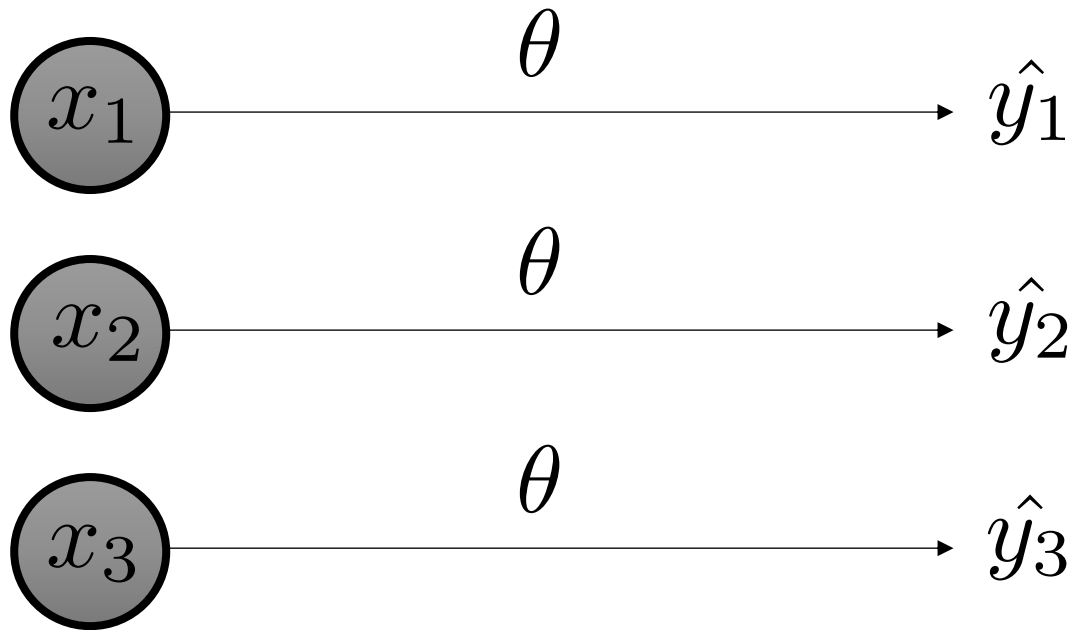


Fine-grained control



Supervised machine learning -- (3)

- The goal of a supervised model is good **generalization**
 - Predict well on data that it has not observed before



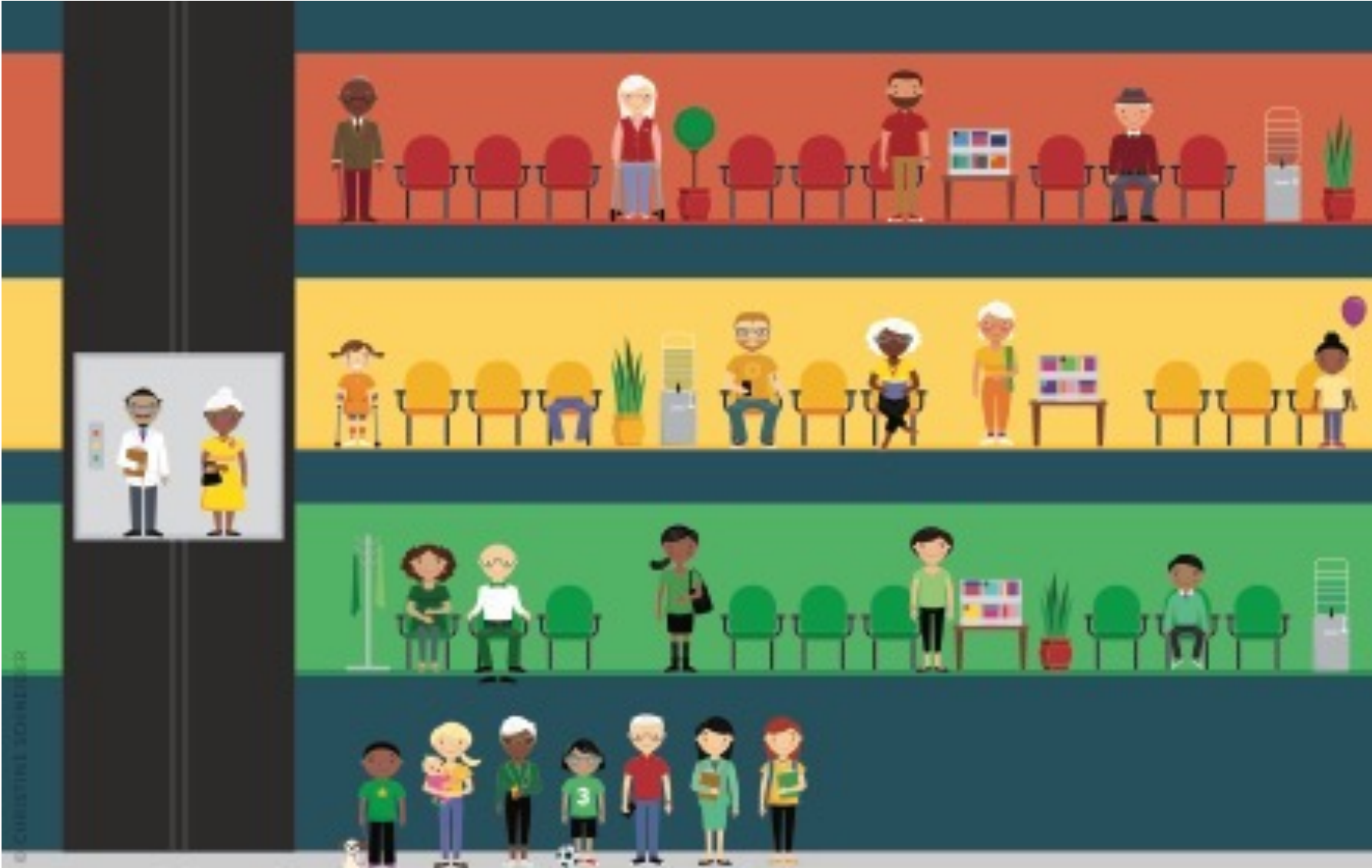
Questions?

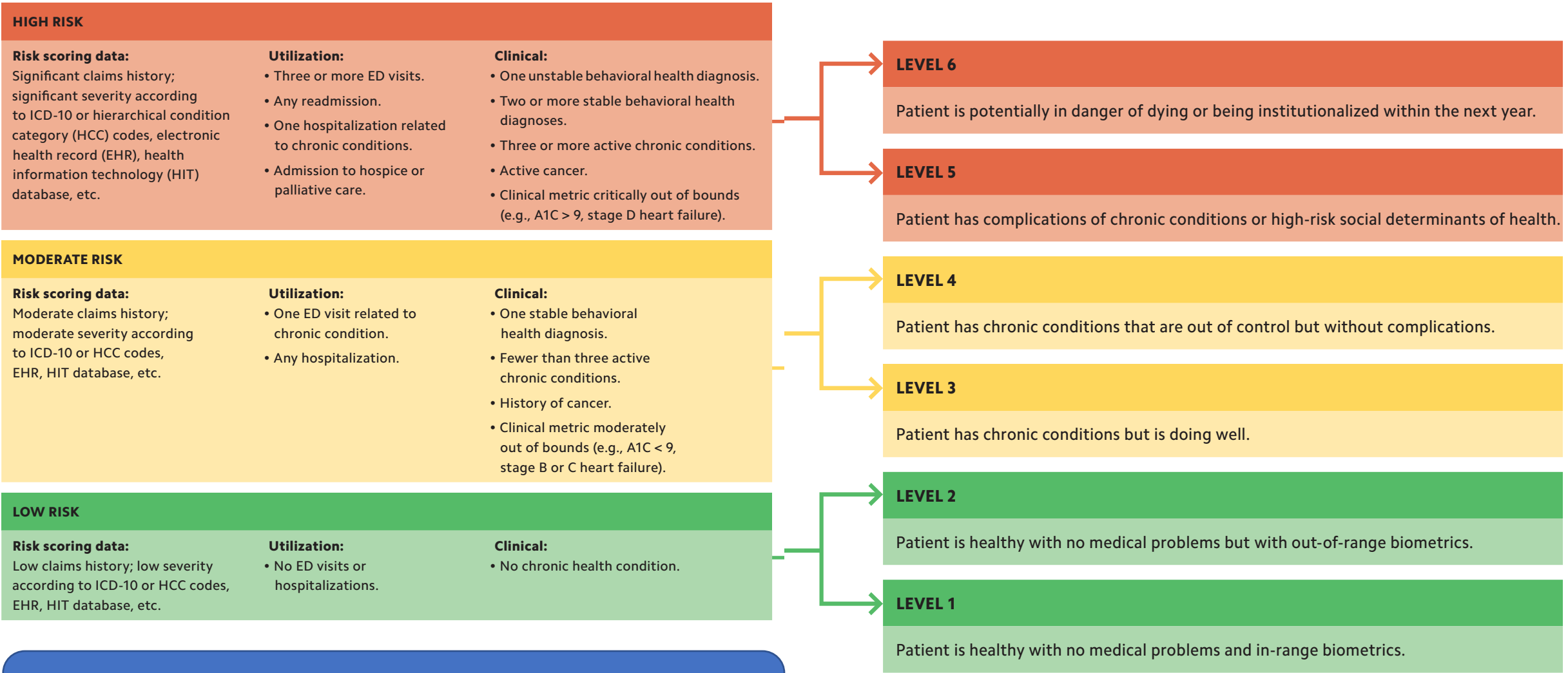
There are no bad questions or questions with obvious answers.

Machine learning for risk stratification

- Clinical task whose goal is to separate patients into high-risk and low-risk of some outcome
- **What do you do with risk:**
 - Choice of interventions prescribed to the patients will vary based on risk
- **Why do this:**
 - Coarse-grained form of personalization
 - Direct clinician attention to patients who need it more

More efficient use of resources



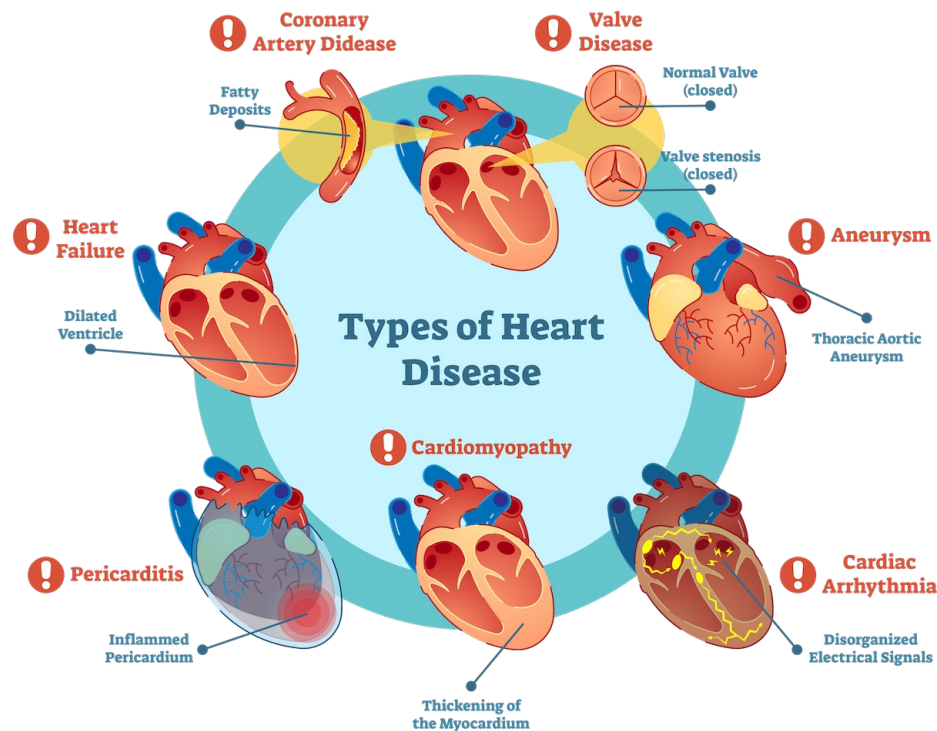


Risk groups via discussion or patients assigned to risk manually (e.g. during rounds)

Some risk scores are easy to estimate

- Heart disease risk score:

<https://www.mayoclinichealthsystem.org/locations/cannon-falls/services-and-treatments/cardiology/heart-disease-risk-calculator>



Heart Disease Risk Calculator

Your 30 year risk of cardiovascular disease

13%*

Your 30 year risk represents the chance that you'll have cardiovascular disease at any point in the next 30 years.

[Take action](#)

[Risk factors](#)

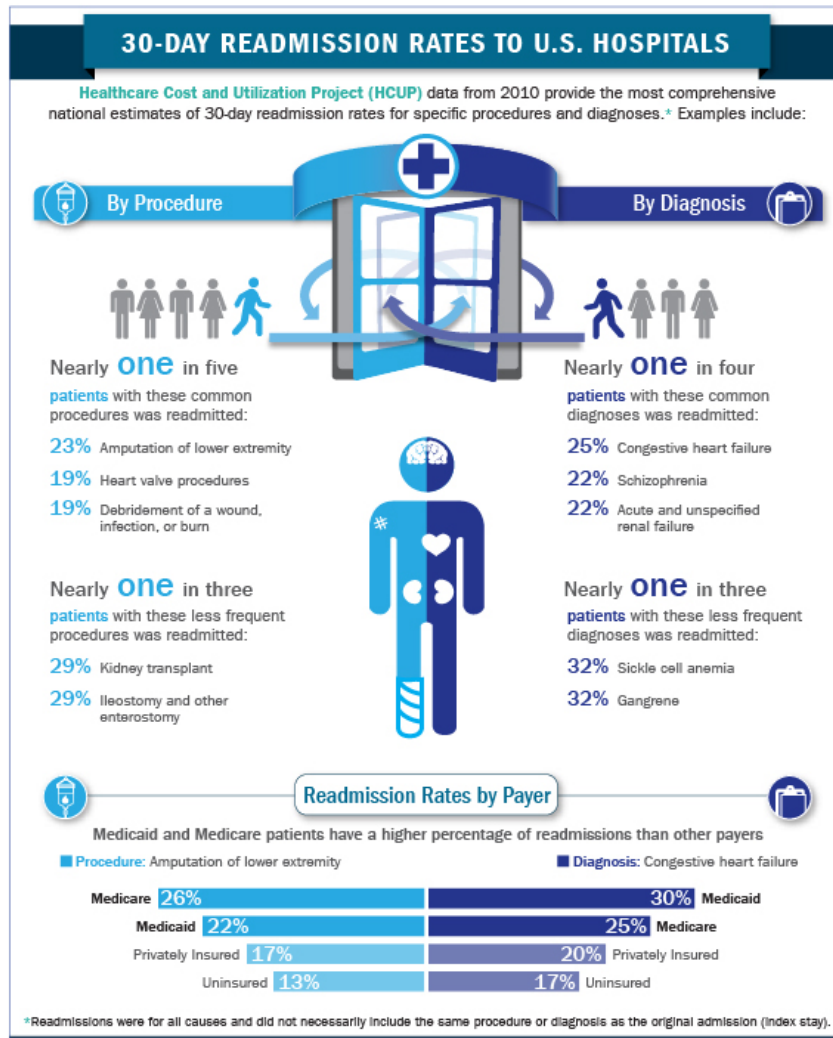
If you were to control your risk factors for cardiovascular disease to acceptable levels, then your 30 year risk would be:

8%*

Before increasing your physical activity level, check with your doctor to make sure it's safe for you to proceed.

- Eat a healthy diet that emphasizes:
 - Fruits, vegetables and whole grains
 - Low-fat dairy products and low-fat proteins, such as poultry, fish and legumes

Other risk scores are much harder



Re-admissions are costly to the hospital and to the patient but difficult to detect.


How do I know what is particularly relevant for an increased risk of readmission?

Figure source:
<https://www.air.org/project/revolving-door-u-s-hospital-readmissions-diagnosis-and-procedure>

Risk stratification as supervised learning

- **Key idea:**

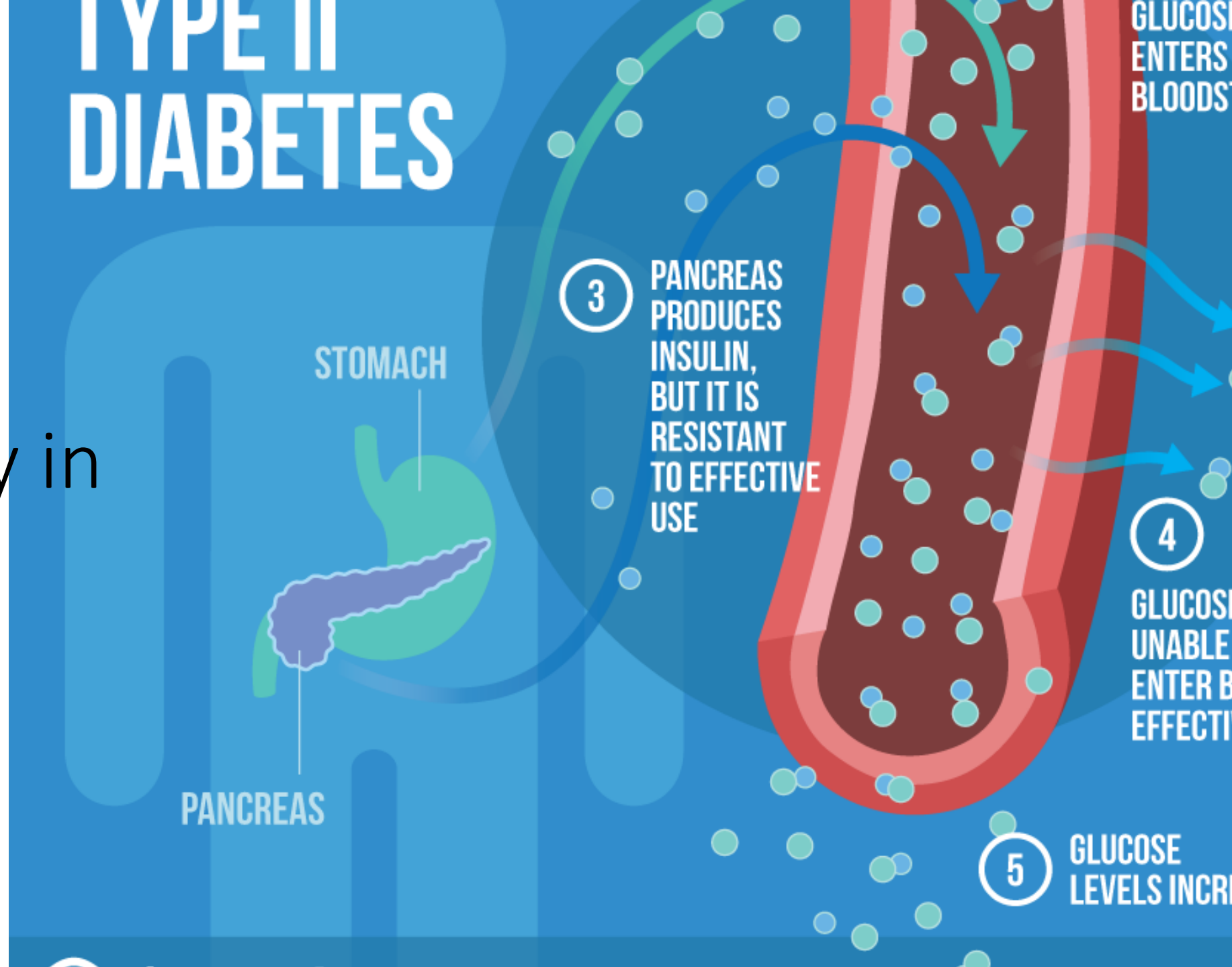
- Outcome [y] needs to be an adverse outcome (or strongly correlated with it)
- Train a predictive model to predict \mathbf{y} from clinical variable \mathbf{x}
- Threshold/bin the predicted outcome of the model to assess risk


$$0 \leq p(y|x; \theta) \leq 0.3$$

$$0.7 \leq p(y|x; \theta) \leq 1$$

Questions?

A case study in diabetes



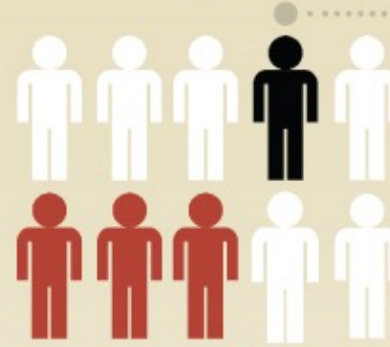
Every **3 minutes** another Canadian is diagnosed with diabetes.

29% of Canadians are currently **living with diabetes or prediabetes.**

This will rise to **33%** by **2025** if current trends continue.

TODAY 3.4 million Canadians are estimated to be living with diabetes.

Diabetes is costing the country
\$14 billion per year



At least
1 in 10

deaths in Canadian adults was attributable to diabetes in 2008/09.

2025 That number is expected to reach more than **5 million** people in the next 10 years.

In 10 years it will cost approximately
\$17.5 billion per year

diabetes.ca | 1-800-BANTING (226-8464)

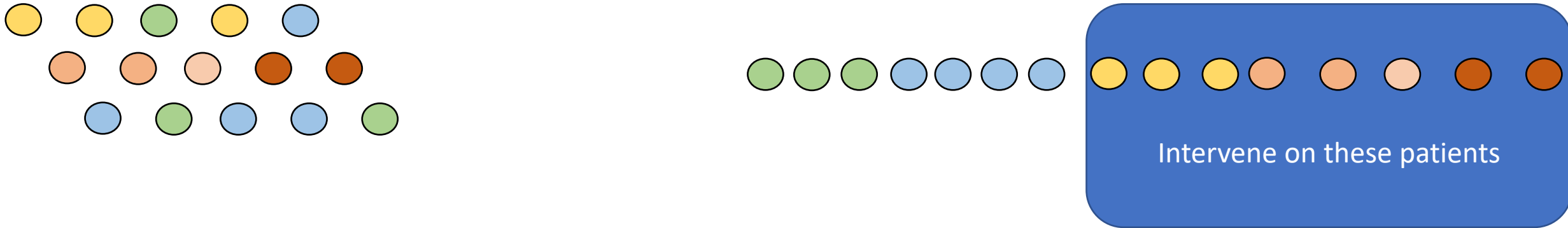
 Canadian
Diabetes
Association

The costs of diabetes

Early detection of T2diabetes

- Early detection of undiagnosed diabetes mellitus: a US perspective, Harris et. al, 2000
- “There is a latent phase before diagnosis of Type 2 diabetes..... risk factors for diabetic micro- and macrovascular complications are markedly elevated and diabetic complications are developing.”
- “define a group of individuals with significant hyperglycemia who also have a high frequency of risk factors for micro- and macrovascular disease.”
- “treating hyperglycemia to prevent complications is more effective than treating these complications after they have developed”

What is that paper saying?



- If you have a patient population and can predict those at high-risk
- You can intervene on those high-risk patients **early**
- Effects of early intervention
 - “preventive interventions should start as early as possible in order to allow a wide variety of relatively low- and moderate-intensity programs”
 - Source: https://care.diabetesjournals.org/content/39/Supplement_2/S115.full-text.pdf

Traditional risk prediction models

- Risk prediction models:
 - ARIC [Atherosclerosis **Risk** in Communities]
 - FRAMINGHAM [Coronary heart disease]
- Easy to ask questions, or measure in clinics when patients come in
- Simple model (typically rule based list or equations)

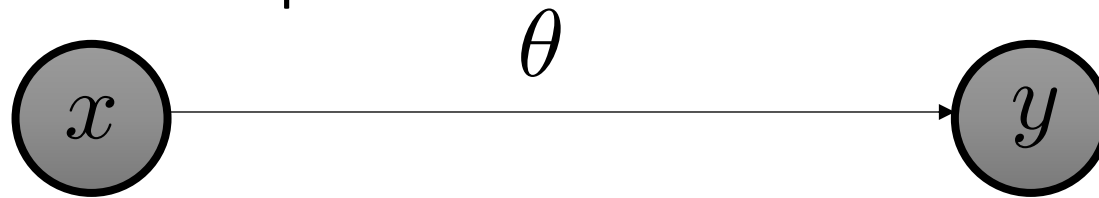
Challenges of traditional risk models

- Difficult to scale
 - Requires patient to come in to know they are high-risk
 - Can be time-consuming
- Existing risk scores do not work well if all values in risk calculator are not observed

Automated population-level risk scoring

- **Key idea:**

- The early detection of progression onto diabetes gives clinicians opportunities for early intervention.
- Develop a predictive model from population level data
- Use the predictive model to estimate risk
- Scale up to millions of patients



If $y > \text{threshold}$:
The patient and the clinician
have a conversation about
how to reduce downstream
complications

y : probability of contracting diabetes in the future



Who cares about this?

- Patients (us!) & disease registries:
 - Better outcomes for patients
- Provincial governments that pay for clinician time
 - Frees up clinician time
- Taxpayers (us!)
 - Lower costs for healthcare

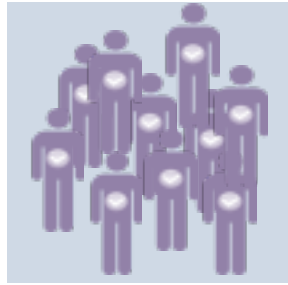
Using administrative claims data from the United States to predict diabetic onset

[Population level prediction of T2 diabetes from health claims and analysis of risk factors, Razavian, Blecker, Schmidt, Smith-McLallen, Nigam, Sontag. Big Data. '16]

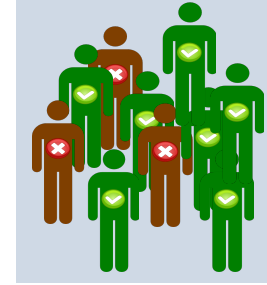
[Early detection of diabetes from health claims, Krishnan, N Razavian, Y Choi, S Nigam, S Blecker, A Schmidt, D. Sontag, Neurips Workshop 2013]

Learning from retrospective data

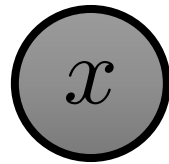
5 years ago



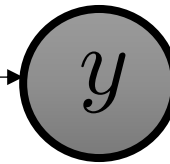
Now



- **Idea:** retrospective data to build predictive models that we can use right now?

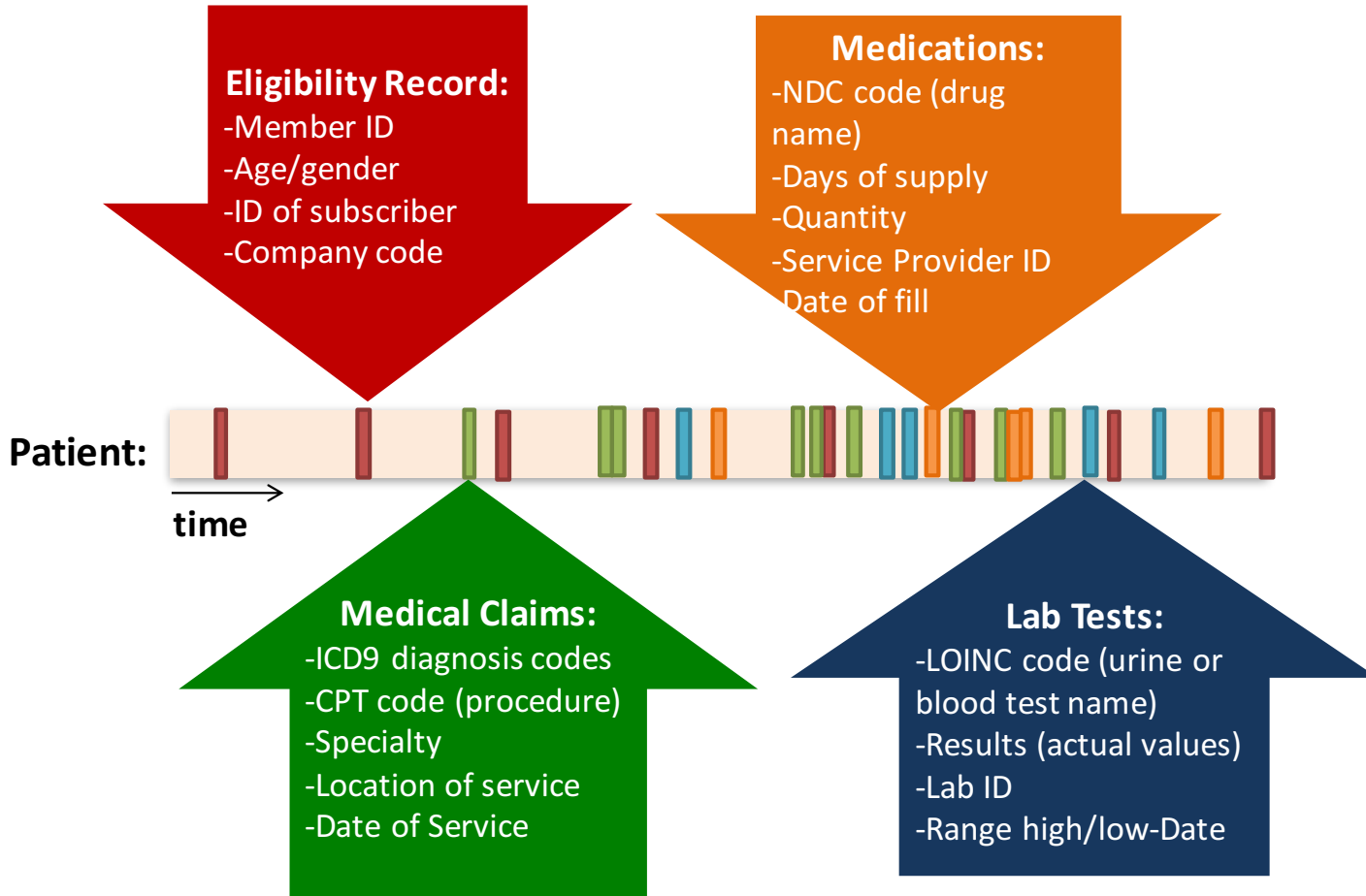


Clinical features



Diabetic status in the future

Administrative claims data

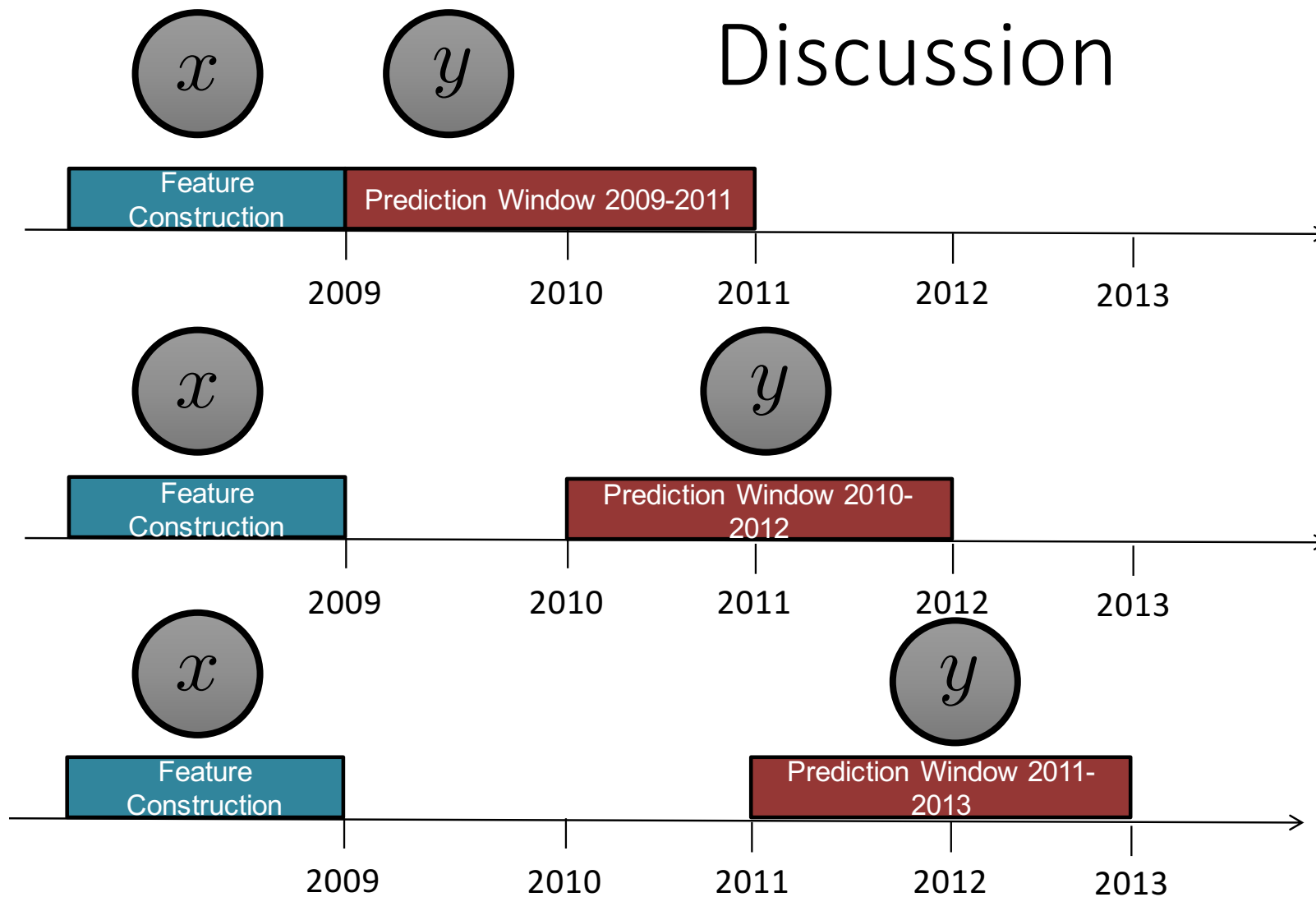


x

y

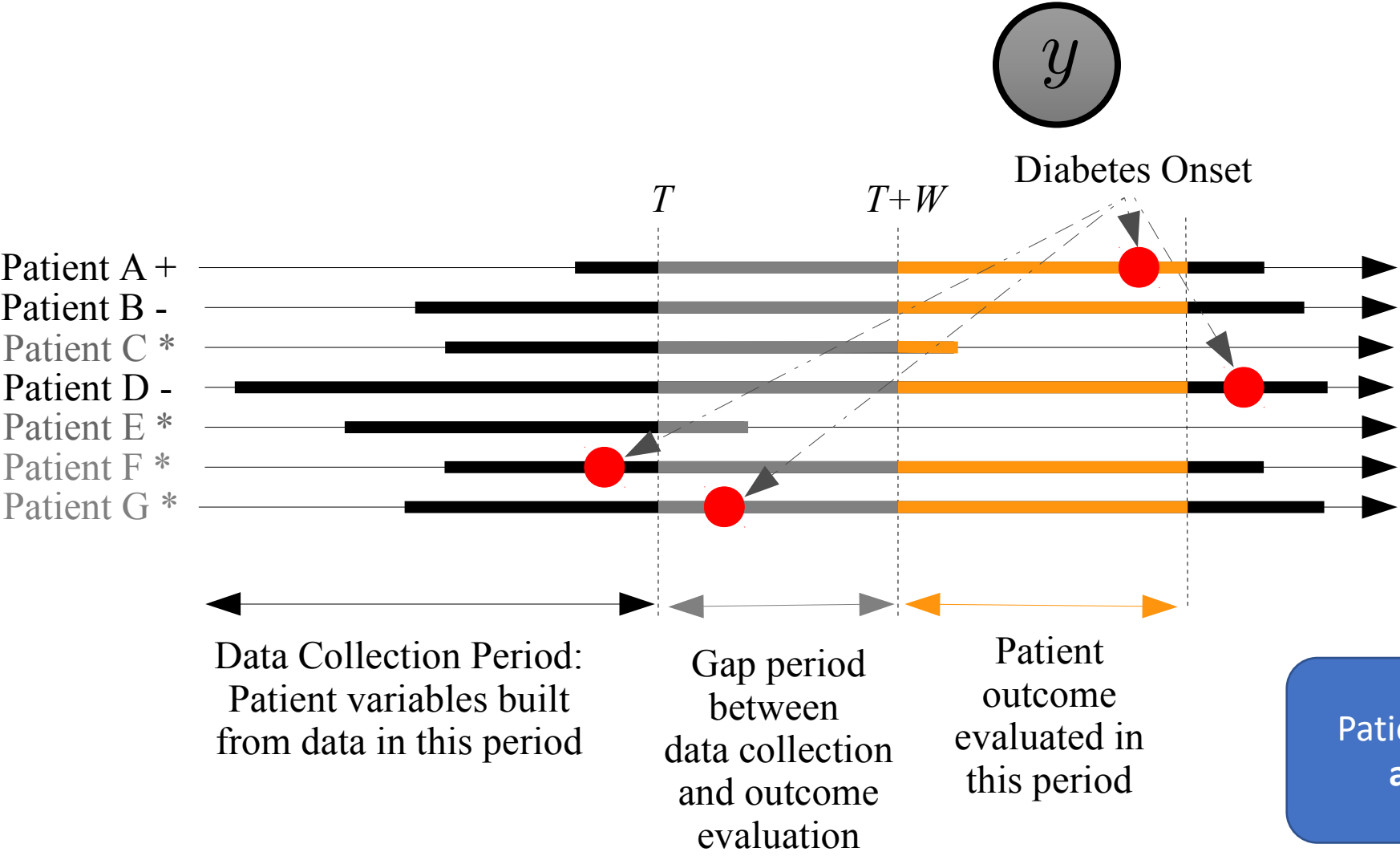
This data is used to define both x and y !

Discussion



Q: Which of these is preferable and why?

Reduction to binary classification



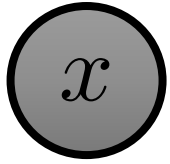
Patient alignment by absolute time

Patient alignment

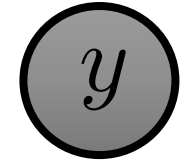
- Absolute time: Collect features based on their status as of 2008
- Relative time:
 - Collect features based on their first visit to their family doctor
 - Collect features by aligning based on each patient's first major comorbidity
- The choice of alignment will depend on how you want to use your model.

Best practices for creating clinical cohorts

- **Cohort design:** For patients that have more than one datapoint, make sure they appear either in the train, the validate or test set
- **Label leakage:**
 - Work with clinicians to understand their practice,
 - There are often subtle, easy to miss signs of disease indication
 - Errors in coding
 - Prescription of a drug
- **Selection bias:**
 - Ensure that the cohort is representative of the population you want to test it on.



Methods



- X: patient features, Y: did the patient have diabetes in the window of time
- L1 regularized logistic regression:
 - Optimize for predictive performance while
 - Doing aggressive feature selection
- Penalize the L1 norm of the weight vector.

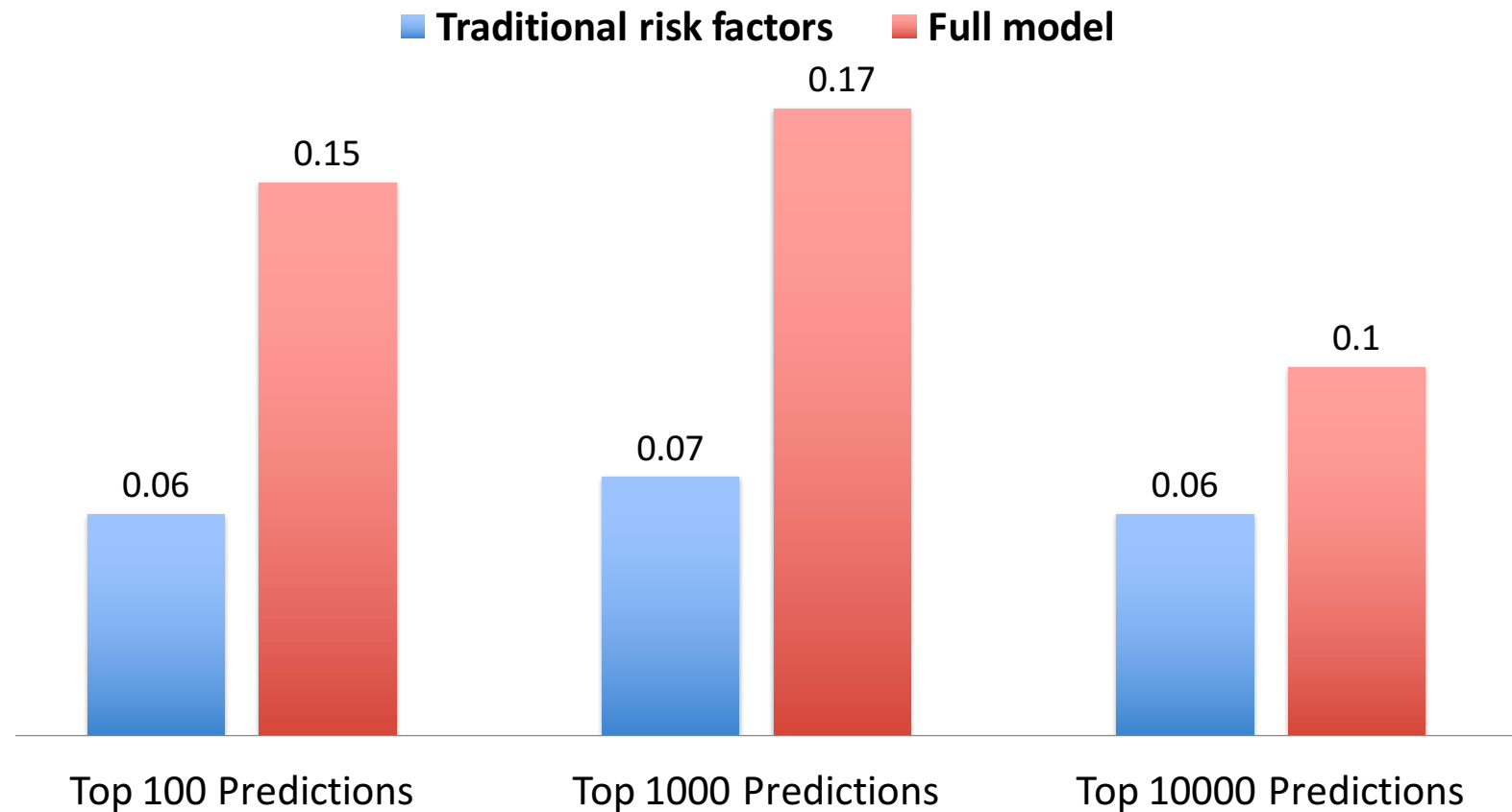
L1 regularized regression

$$\min_{\theta} \sum_{i=1}^N -\mathcal{L}(y_i, x_i) + \lambda \|\theta\|_1 \quad \|\theta\|_1 = \sum_d |w_d|$$

- X: patient features, Y: did the patient have diabetes in the window of time
- L1 regularized logistic regression:
 - Optimize for predictive performance while
 - Doing aggressive feature selection
- Penalize the L1 norm of the weight vector.
 - d: dimension of feature vector

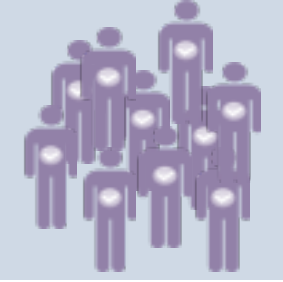
Highlights of results

- Total number of features in model: 42000

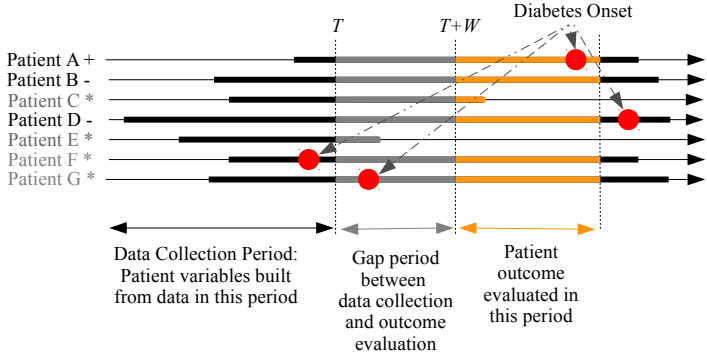
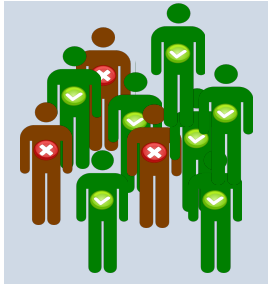


Recap and summary

5 years ago



Now



Clinical features Diabetic status in the future

Discussion : What are the limitations of this approach?

Feedback welcome!

- Key advantage of this style of class – feedback!
- Are there interesting topics you'd like to learn more about?
- Send me & course staff an email.

TODOs

- In two weeks, we will have a project planning session. It will involve a discussion with Alistair Johnson (lead creators of the MIMIC dataset)
 - Please watch the following posted videos **to get a sense of what you can do with MIMIC**
- [Introducion to MIMIC](#)
- [MIMIC analysis tutorial](#)
- [MIMIC data tutorial](#)