

Machine Learning for Healthcare

6.871x

Foundations of Machine Learning



What is machine learning?

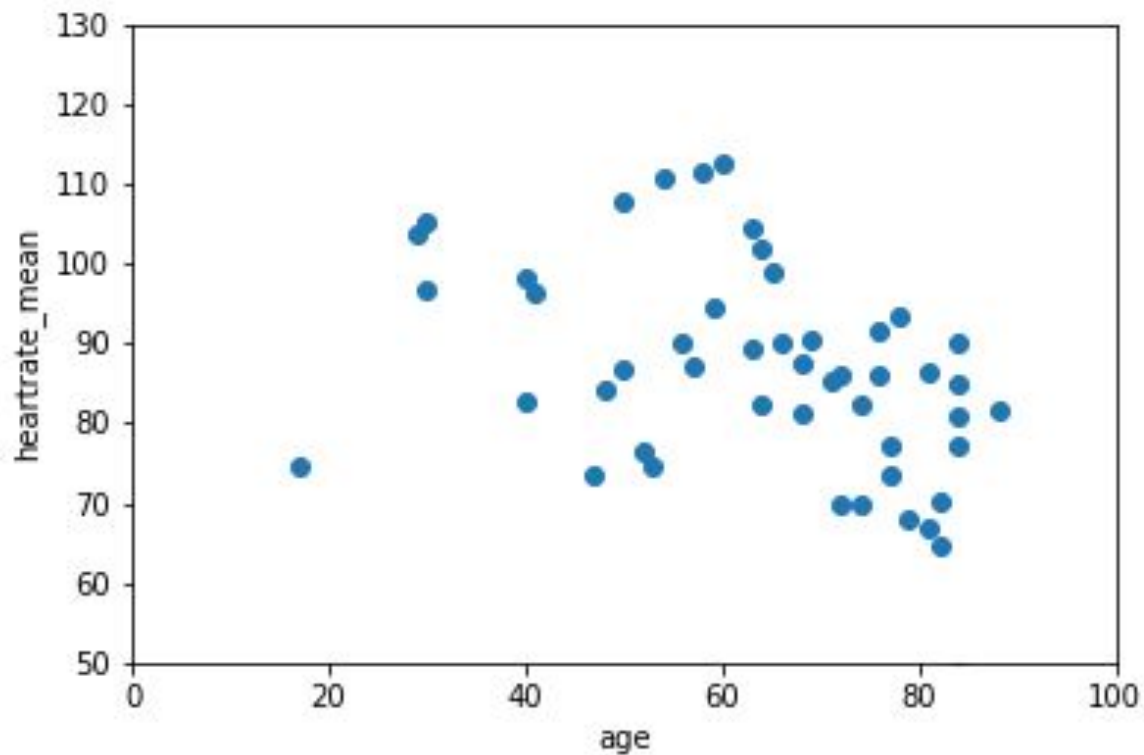
- Machine learning is the application of statistical, mathematical, and numerical techniques to derive some form of knowledge from data.
- This knowledge may afford us some summarization, visualization, grouping, or even predictive power over datasets.

Machine learning for Healthcare

What kinds of problems in healthcare can machine learning be used to solve?

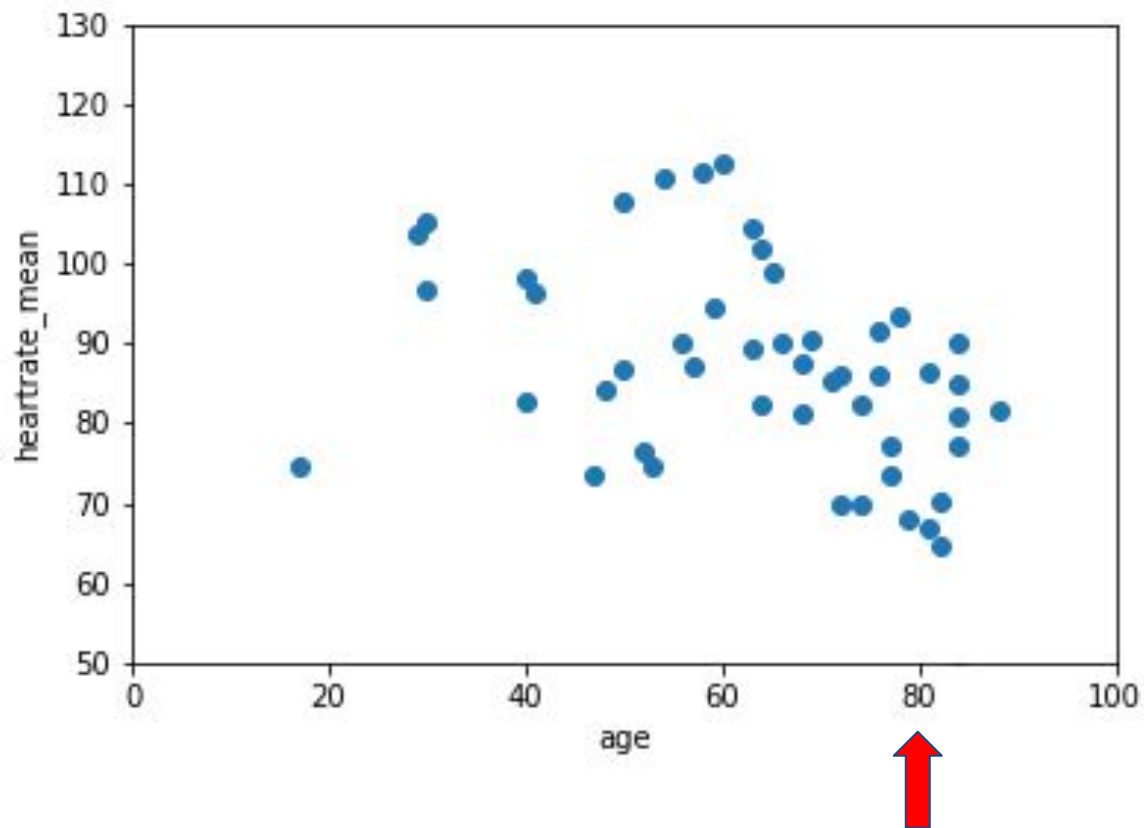
Example: Regression

predicting a numeric value



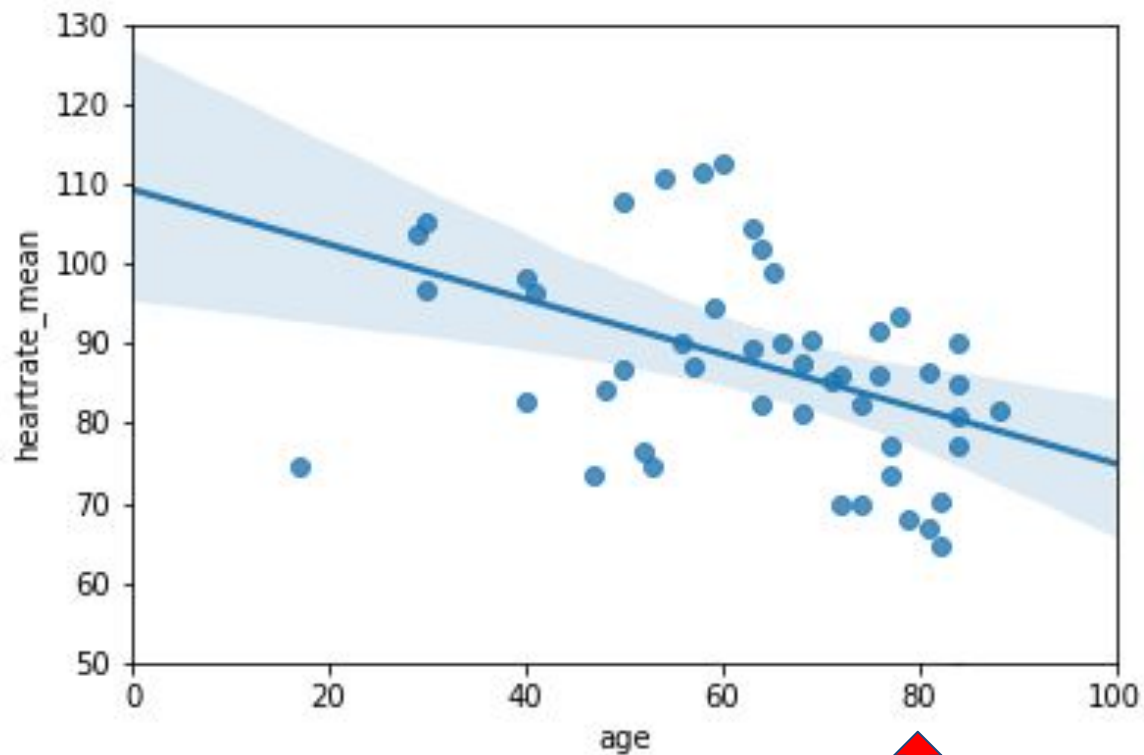
Example: Regression

predicting a numeric value



Example: Regression

predicting a numeric value

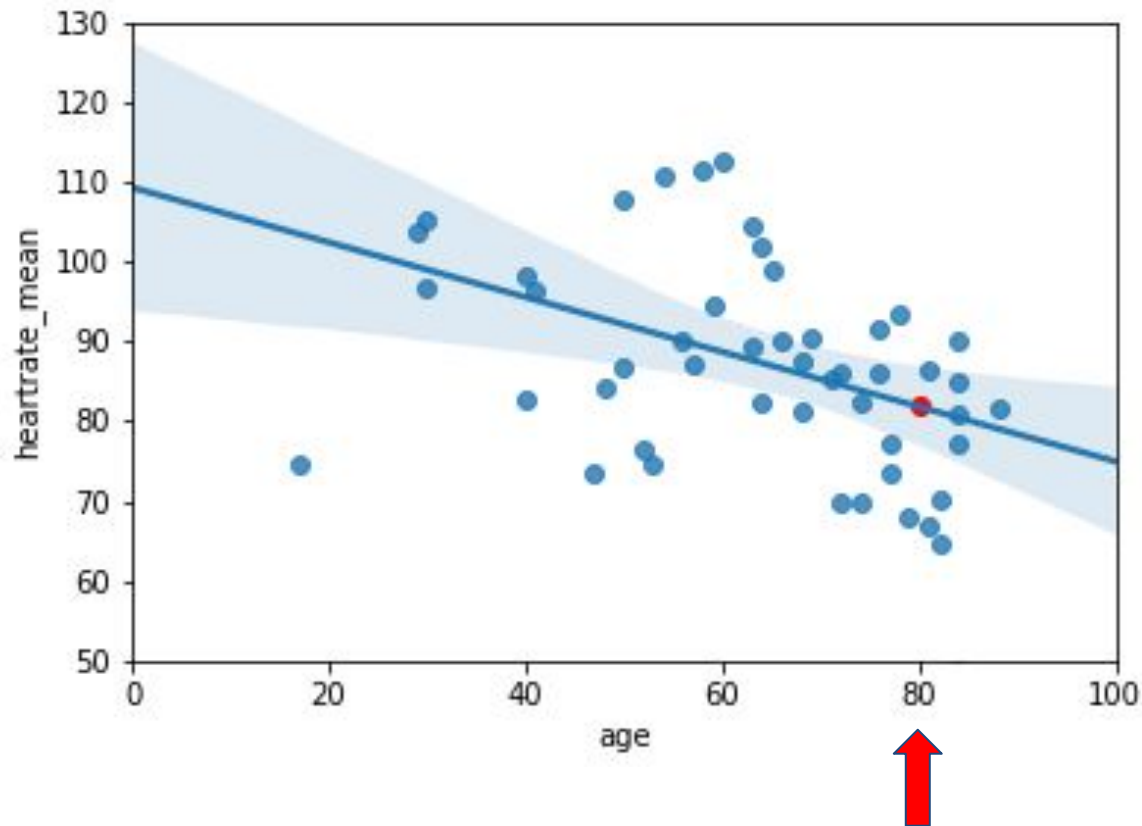


Example: Regression

predicting a numeric value



$$\text{heartrate_mean} = -0.34 * \text{age} + 109.3$$

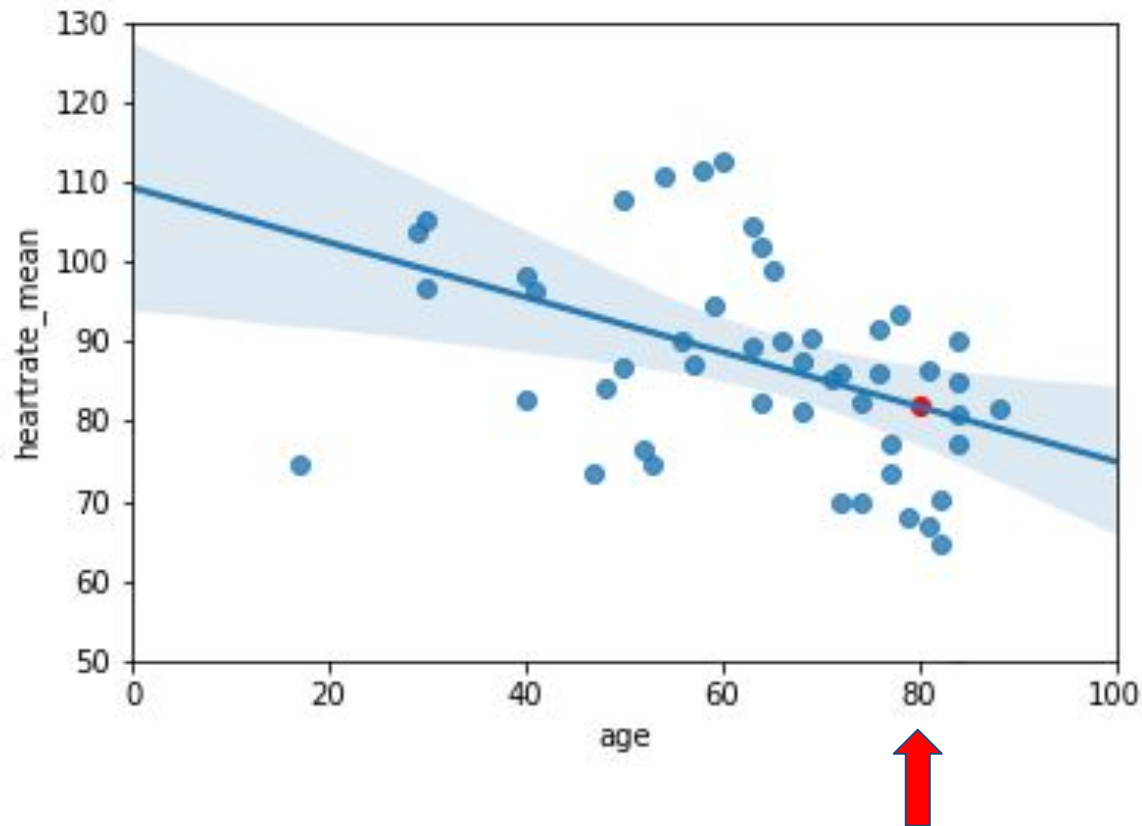


Example: Regression

predicting a numeric value



$$\text{heartrate_mean} = -0.34 * \text{age} + 109.3$$

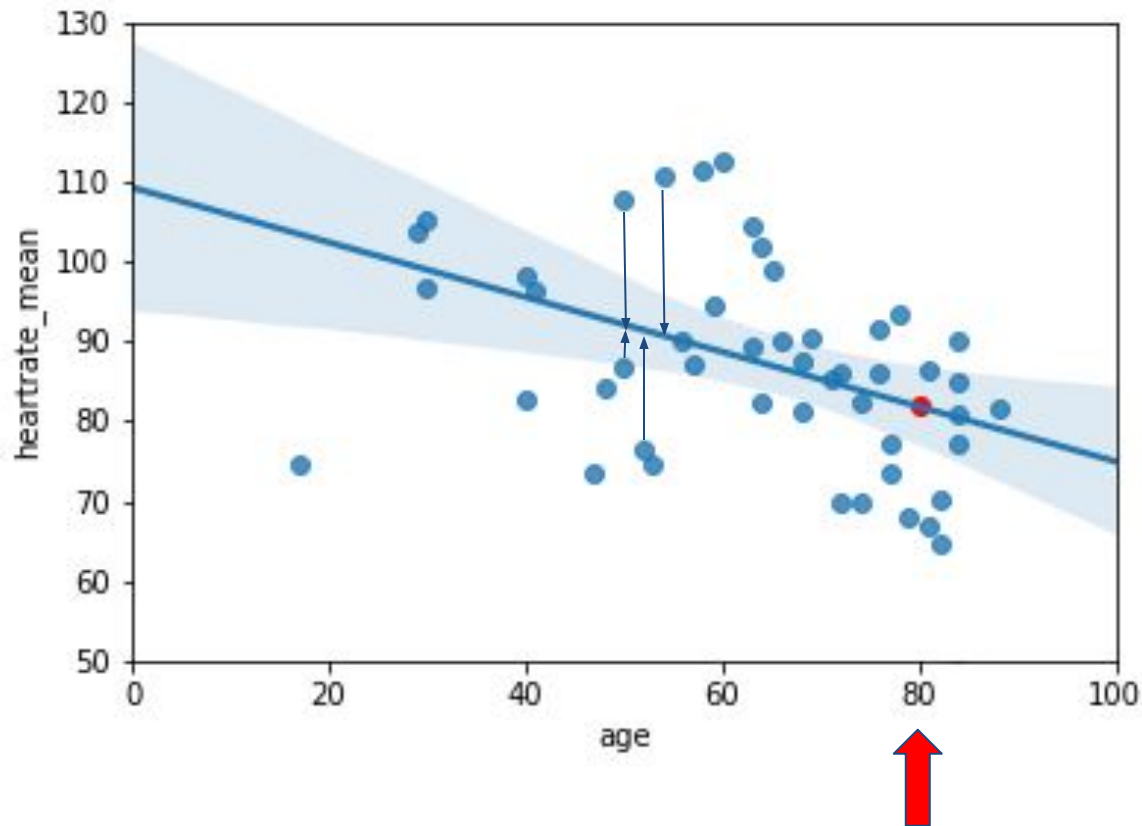


Example: Regression

predicting a numeric value

$$\text{heartrate_mean} = -0.34 * \text{age} + 109.3$$

$$\frac{1}{N} \sum_{i=1}^N (y_i - f(x_i))^2$$



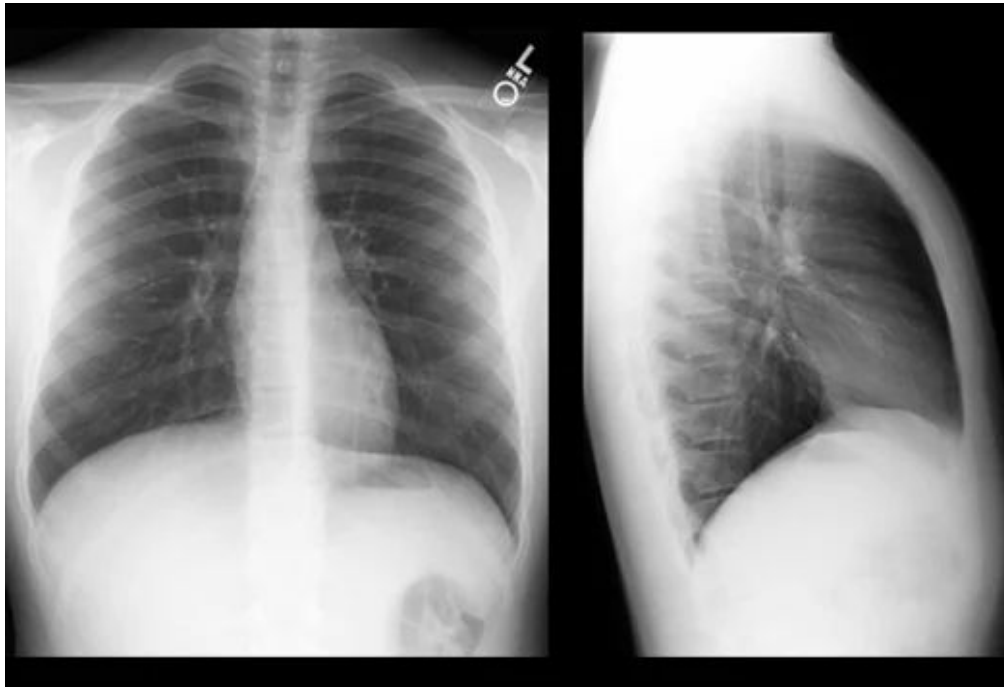
Machine learning for Healthcare

What kinds of problems in healthcare can machine learning be used to solve?

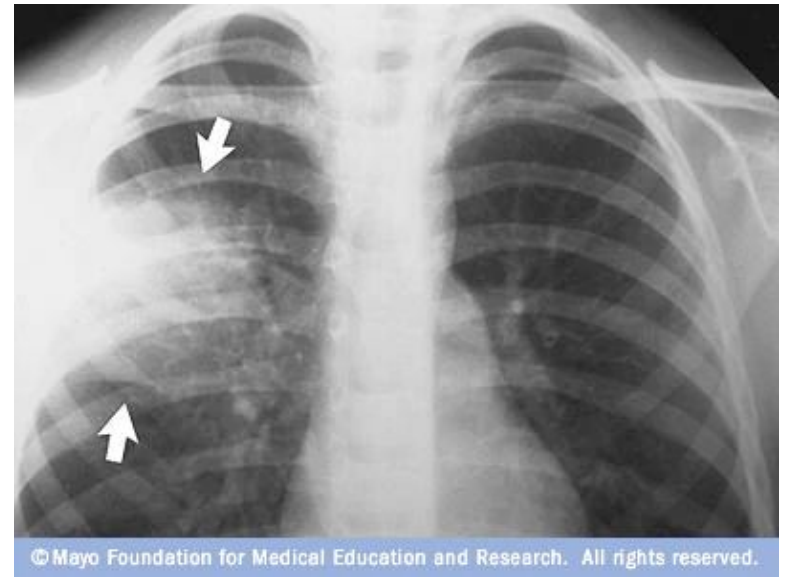
- Supervised learning: Learn to predict labeled outputs of interest.

Example: Classification

from data to discrete classes







label: pneumonia **negative**



label: pneumonia **positive**

Example: Classification

from data to discrete classes

	classified negative	classified positive
true negatives	 ✓	 false positive ✗
true positives	 false negative ✗	 ✓

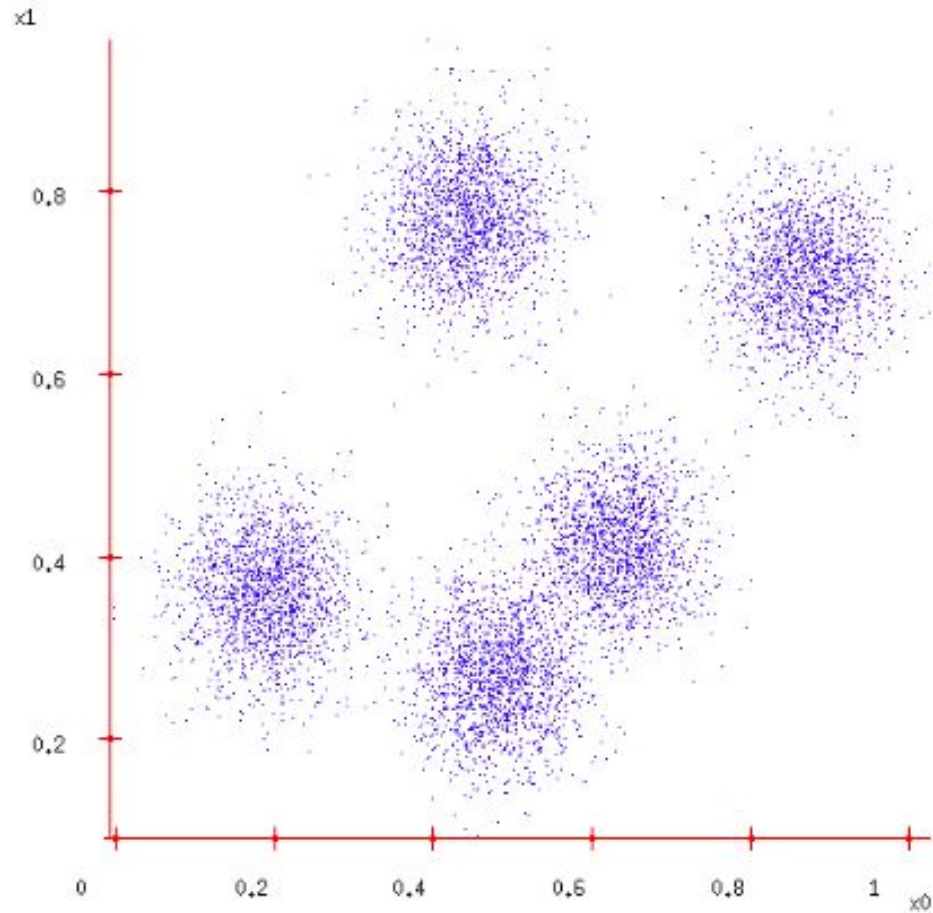
Machine learning for Healthcare

What kinds of problems in healthcare can machine learning be used to solve?

- Supervised learning: Learn to predict labeled outputs of interest.
- Unsupervised learning: Learn patterns and groupings from a dataset with no pre-existing labels.

Example: Clustering

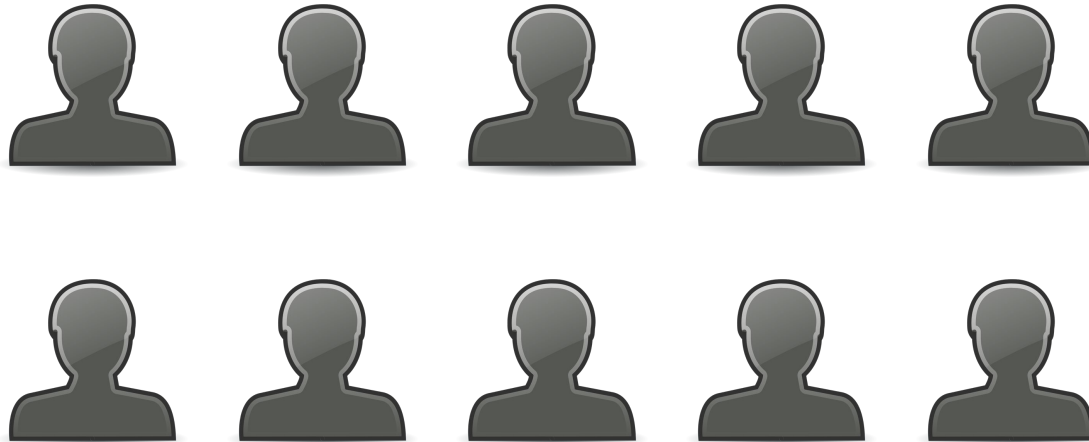
discovering structure in data



Example: Clustering

discovering structure in data

“It is now recognized that there are **distinct asthma phenotypes**, and that these distinct phenotypes respond differently to therapy.”



[Haldar et al. 2008]

Example: Clustering

discovering structure in data

“It is now recognized that there are **distinct asthma phenotypes**, and that these distinct phenotypes respond differently to therapy.”



average age of onset: 14
BMI: 26

severe asthma exacerbations
in past 12mo: 1.9

Early-Onset Atopic Asthma



average age of onset: 35
BMI: 36

severe asthma exacerbations
in past 12mo: 1

Obese Noneosinophilic



average age of onset: 28
BMI: 26

severe asthma exacerbations
in past 12mo: 0.3

Benign Asthma

[Haldar et al. 2008]

Machine learning for Healthcare

What kinds of problems in healthcare can machine learning be used to solve?

- Supervised learning: Learn to predict labeled outputs of interest.
- Unsupervised learning: Learn patterns and groupings from a dataset with no pre-existing labels.
- Reinforcement learning: Learn strategies on how to take actions to maximize rewards.

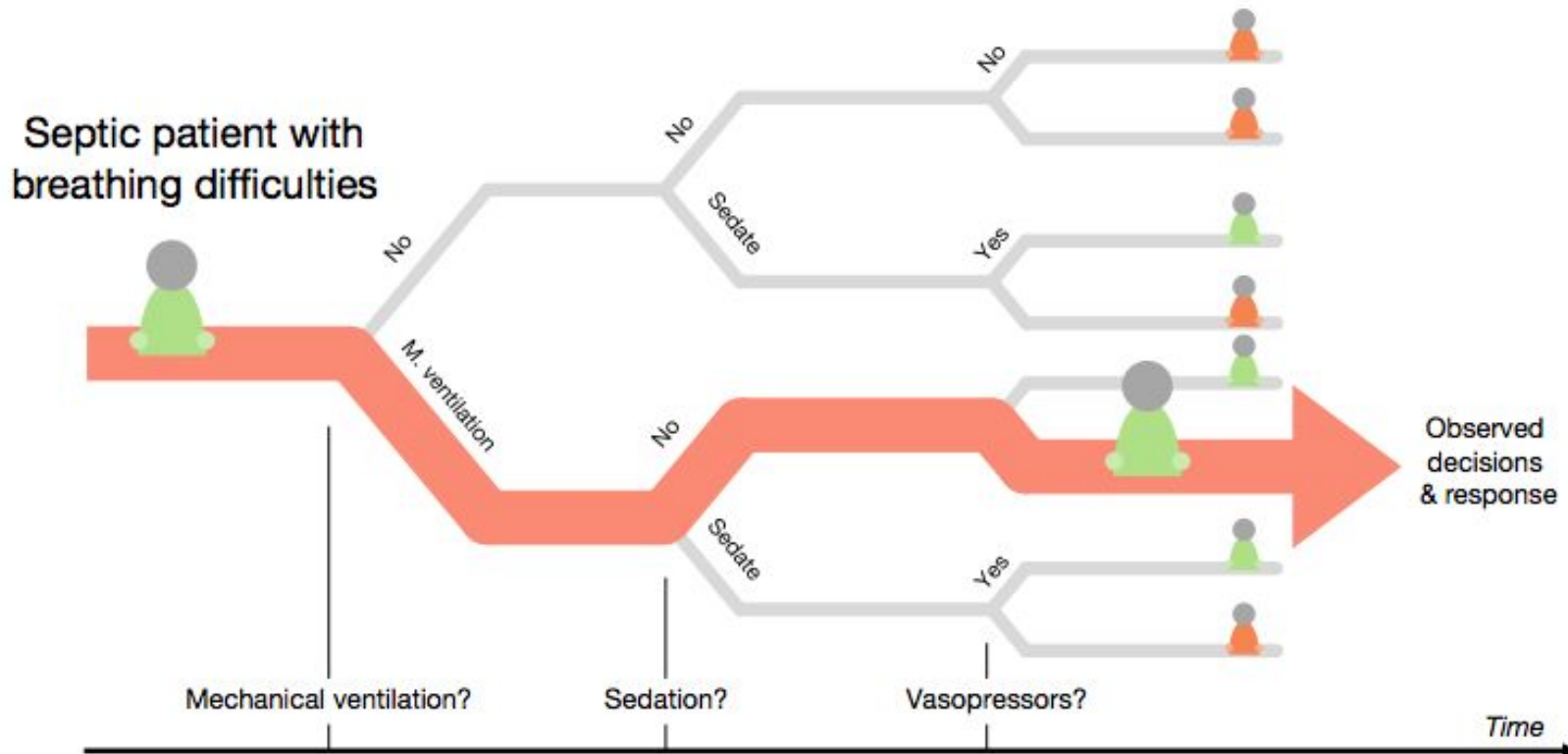
Example: Reinforcement Learning

learning which actions to take



Example: Reinforcement Learning

learning which actions to take



Summary

- There are 3 kinds of machine learning problems that we will focus on in this course: supervised, unsupervised, and reinforcement learning.
- There are several different types (subcategories) of models within each of these 3 groups!
- When considering which kind of model to use, ask yourself:
 - What, if anything, are you trying to “predict”?
 - Do you have labeled outputs of interest?
 - Does the scenario you wish to model involve taking multiple actions in sequence?

Machine learning for Healthcare

So how can we make a machine learning model?

Machine learning for Healthcare

So how can we make a machine learning model?

1. Problem definition: Define your objective and examine any underlying assumptions. Use these to choose the kind of model you'd like to learn.

Machine learning for Healthcare

1. Problem Definition

“I want to be able to predict if a patient will develop diabetes.”

Machine learning for Healthcare

1. Problem Definition

“I want to be able to predict if a patient will develop diabetes.”

- What are my prior **beliefs**?

What do I believe will be influential in predicting diabetes? Are these relationships linear, or are they more complex?

- What should our model take as **input**?

Should we use the patient's entire history of data, or just their data at 1 particular point in time?

- What should our model **output**?

Is our goal to predict if the patient will develop diabetes tomorrow? One year from now? Should our model output “yes” or “no”, or return the probability (between 0 and 1) that the patient will develop diabetes?

- What does it mean for our model to be **accurate**?

Do we care more about some types of error than others?

Machine learning for Healthcare

1. Problem Definition

“I want to be able to predict when a patient will develop diabetes.”

- What are my prior **beliefs**?

There is a linear relationship between demographics, vitals, and labs and time of diabetes onset. -> use a **linear model**.

- What should our model take as **input**?

We will compute features that summarize the patient's past 1 year of clinical data.

- What should our model **output**?

Our goal is to predict whether the patient will develop diabetes in 2 years (**classification**).

- What does it mean for our model to be **accurate**?

We hope to prevent both false positives and false negatives.

Machine learning for Healthcare

So how can we make a machine learning model?

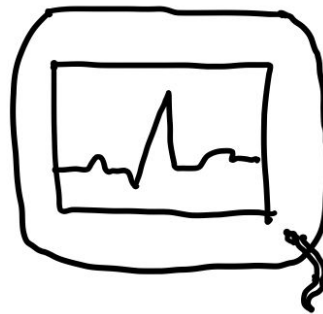
1. Problem definition: Define your objective and examine any underlying assumptions. Use these to choose the kind of model you'd like to learn.
2. Data collection: Measure or obtain the features that you need from the population.

Machine learning for Healthcare

2. Data collection



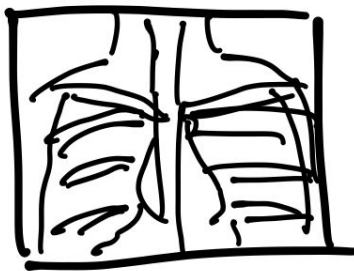
Health insurance billing



Labs and vitals



Nursing notes



Imaging



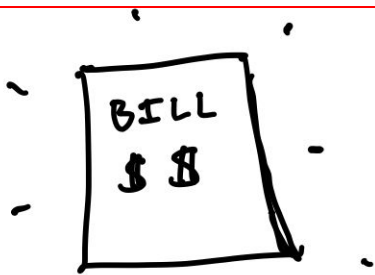
Devices



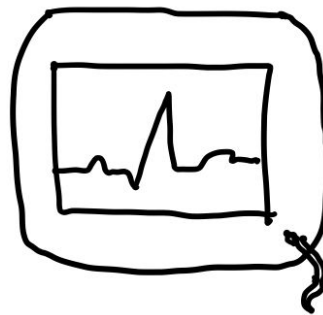
Genomics

Machine learning for Healthcare

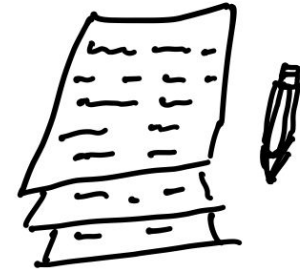
2. Data collection



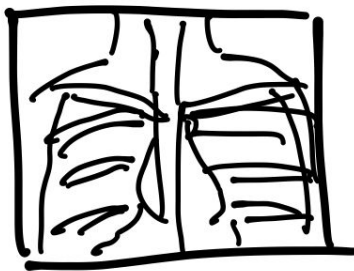
Health insurance billing



Labs and vitals



Nursing notes



Imaging



Devices



Genomics

Machine learning for Healthcare

So how can we make a machine learning model?

1. Problem definition: Define your objective and examine any underlying assumptions. Use these to choose the kind of model you'd like to learn.
2. Data collection: Measure or obtain the features that you need from the population.
3. Training: Train the model on data you already have.

Machine learning for Healthcare

3. Training

Data: $\{X_i = (x_{i,1}, x_{i,2}, \dots, x_{i,m}), y_i\}_{i=1}^N$

	age	eth_black	eth_hispanic	eth_other	eth_white	heartrate_mean	sysbp_mean	diabetes	
X_1	$x_{1,1}$ 65.0	$x_{1,2}$ 0	$x_{1,3}$ 0	\dots 0	1	84.791667	152.631579	y_1	1
X_2	$x_{2,1}$ 41.0	$x_{2,2}$ 0	$x_{2,3}$ 0	\dots 1	0	87.644737	154.355263	y_2	-1
X_3	$x_{3,1}$ 72.0	$x_{3,2}$ 0	$x_{3,3}$ 0	\dots 0	1	83.208333	143.500000	y_3	-1
X_4	$x_{4,1}$ 39.0	0	0	0	1	99.149254	129.430769	y_4	-1
X_5	$x_{5,1}$ 47.0	0	0	0	1	82.400000	102.131148	y_5	1

Machine learning for Healthcare

3. Training

Data: $\{X_i = (x_{i,1}, x_{i,2}, \dots, x_{i,m}), y_i\}_{i=1}^N$

$$f(x_i) = \text{sign}(w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \dots + w_m x_{i,m})$$

Our goal: find weights w that minimize the following loss function:

$$L(w) = \frac{1}{N} \sum_{i=1}^N \max(-y_i \cdot w^T X_i, 0)$$

Machine learning for Healthcare

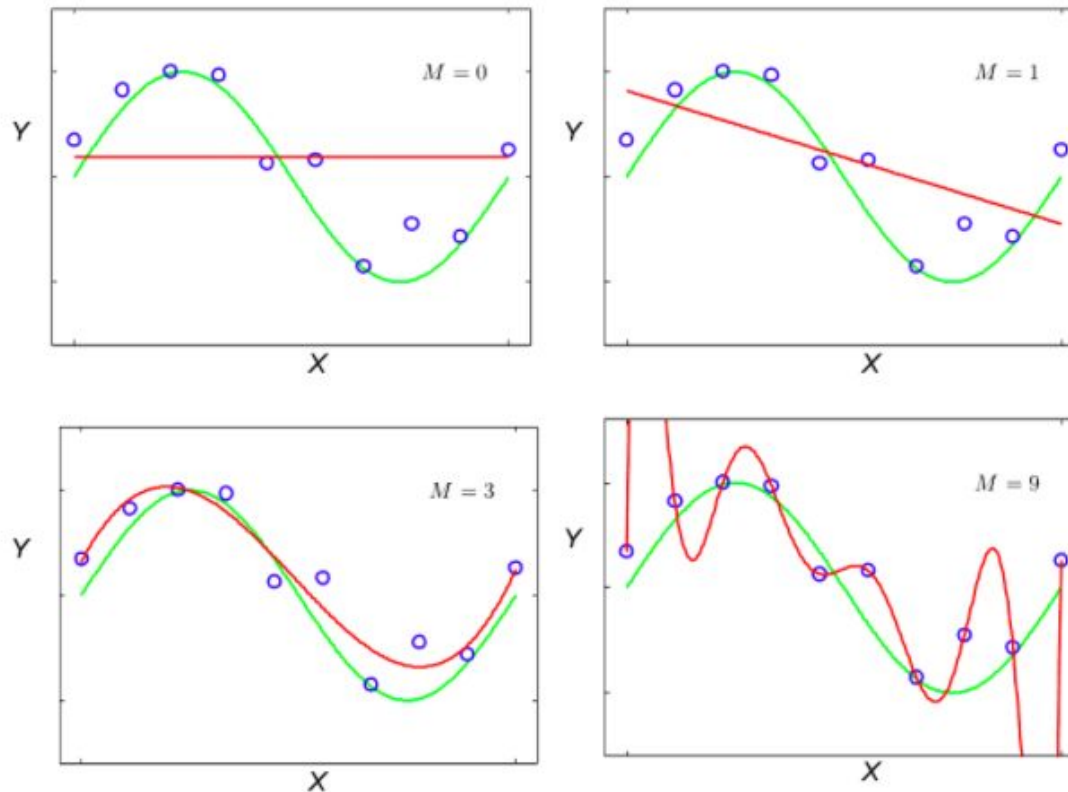
So how can we make a machine learning model?

1. Problem definition: Define your objective and examine any underlying assumptions. Use these to choose the kind of model you'd like to learn.
2. Data collection: Measure or obtain the features that you need from the population.
3. Training: Train the model on data you already have.
4. Validation: Validate the model on other data that you already have, that you didn't use to train on.

Machine learning for Healthcare

4. Validation

Which one is best?



Machine learning for Healthcare

4. Validation

In machine learning, our goal is to move from data to knowledge. In most cases, we hope that our models are **generalizable**: we wish to use them to make predictions in the future using **new data** that we don't have available to us right now!

Machine learning for Healthcare

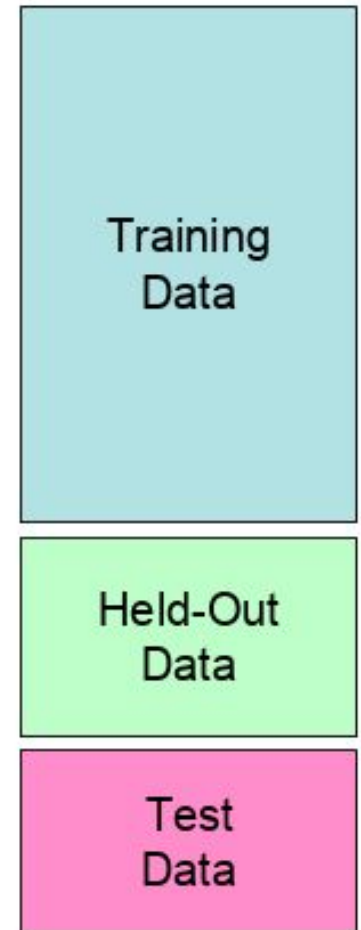
4. Validation

Question: How well will this model do on new data?

One solution: Treat some data that you do have as the “new data” - and evaluate on it!

Split the data you do have into separate datasets for training, validation, and testing (evaluation).

Do not train on the evaluation set.



Summary

- These four steps typically happen chronologically, but it's very common to go back to earlier steps and revise throughout the process!
- At each step of the process, it's important to **examine underlying assumptions!**