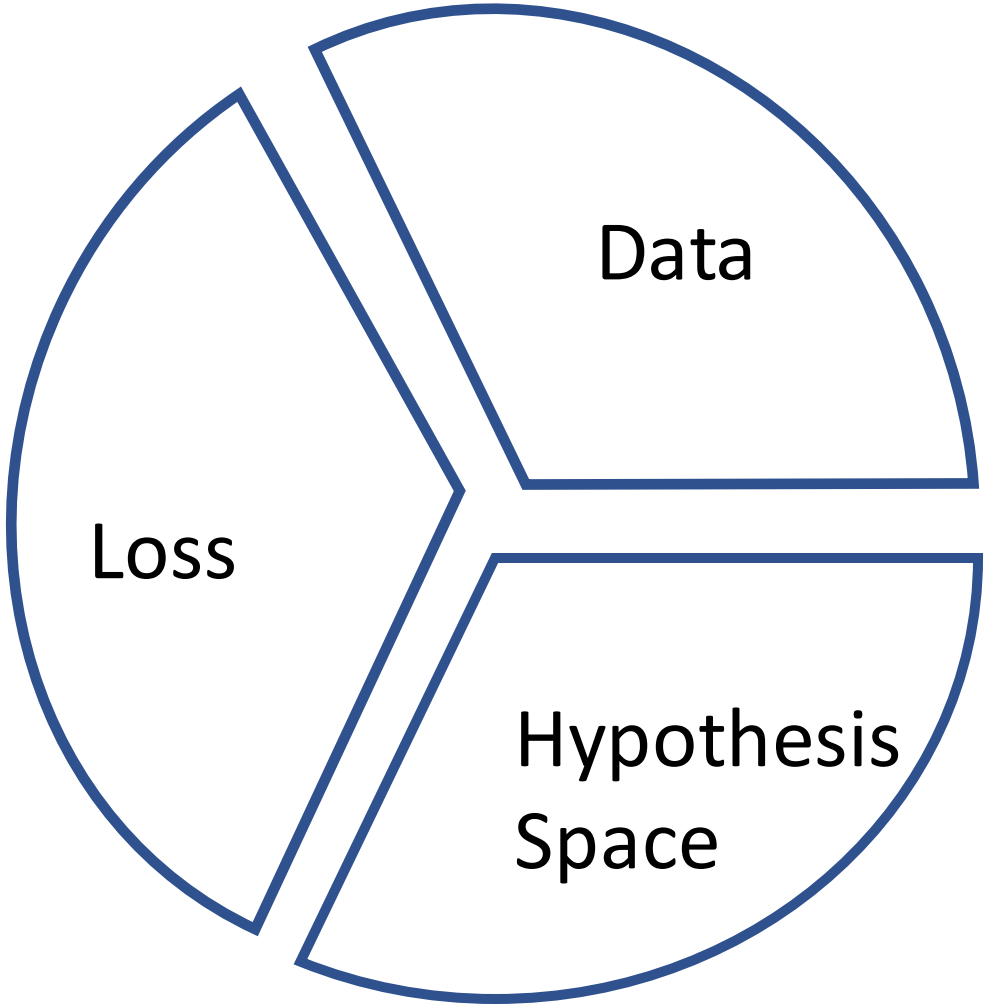
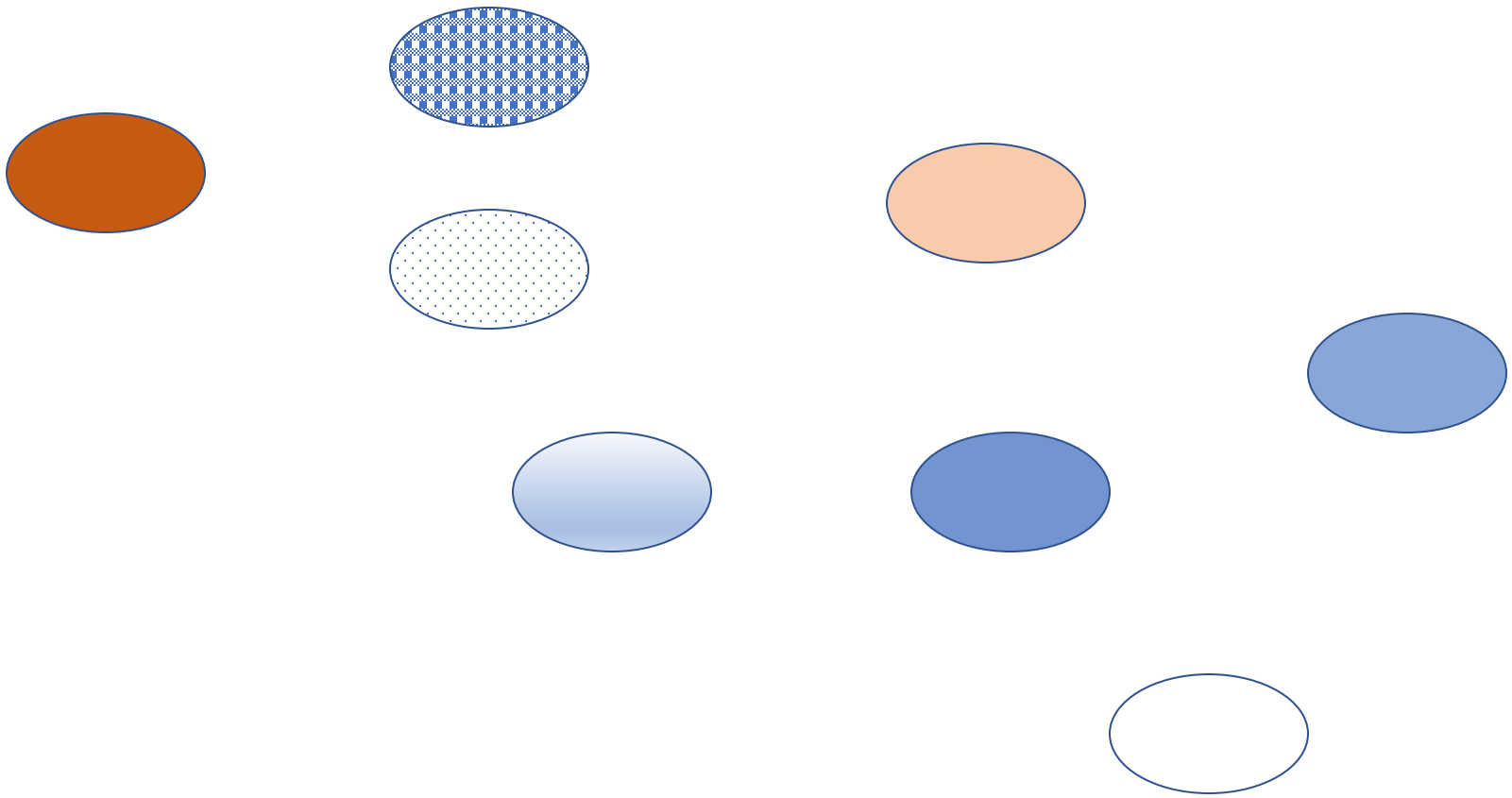
Three Components of Machine Learning

Alexander Jung


Data
Loss
Hypothesis
Space

Data

Data = (Large) Set of “Data Points”



data points are different objects but which are of similar “type”

Data Point = Atomic Unit of Information

- highly abstract concept
- data points can represent persons
- data points can represent realizations for random variables
- data points can represent sets of data points
- ML methods require sufficient amount of data points

Features and Labels

- data points have many properties
- features=properties that can be measured/computed easily
- labels=properties that require human experts
- labels are higher-level facts or quantities of interest
- we like to learn predicting the label based on features

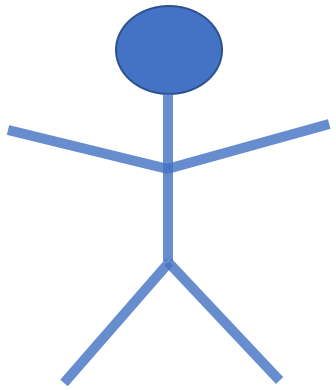
Data Point = “Some Photo”

features could include:

- red, green, blue intensities of pixels
- timestamp of photo
- location of photo shot
- identify to photographer

label could be:
hiking duration to the mountain peak

Data Point = “Some Person”



features:

- name
- healthcare records
- credit card transactions
- social media posts
- genetic fingerprint
- fingerprint
- travel history

label:

- how likely will person need intensive care next week

Data Point = “Some Dataset”

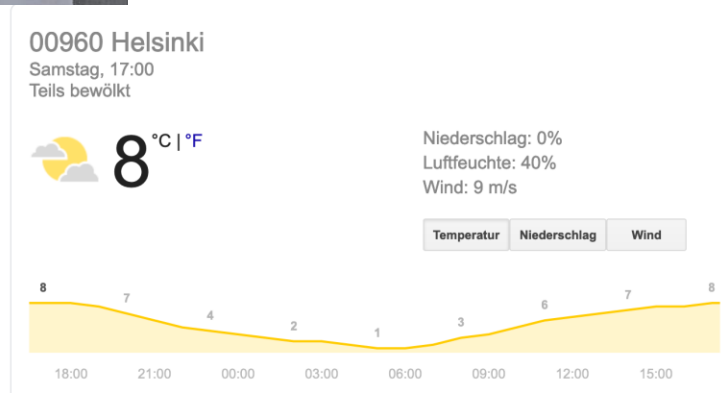
features:

- number of data points
- what type of features are used for data points
- what label is used for data points

label:

- accuracy with which label can be predicted based on features

Data Point = “Some Ski-Day Ahead”



features:

- snapshot in the morning
- morning temperature
- weather forecast

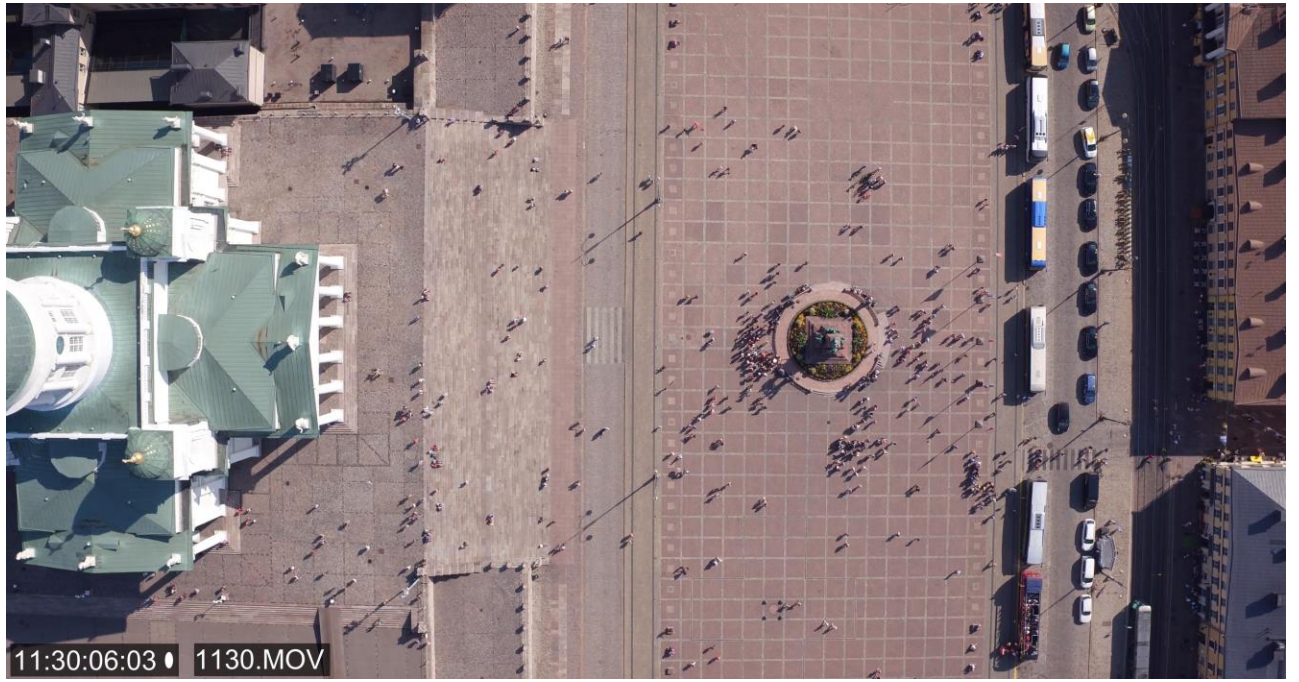
label:

- maximum daytime temperature (important for ski waxing)

Data Point = “Somewhere in Helsinki”

features:

- coordinates of place
- city building maps
- current traffic statistics
- CET time
- drone video



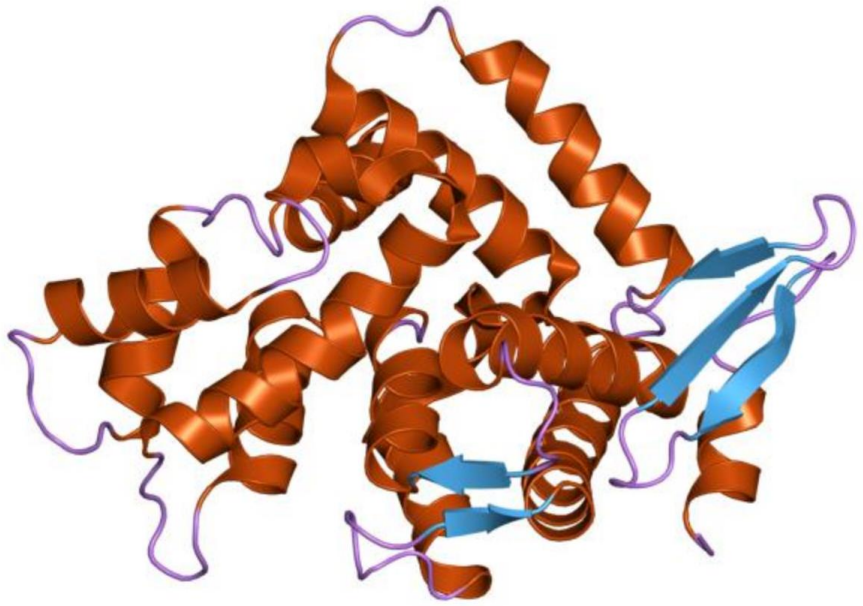
label:

how many people are currently at this place ?

Data Point = “Some Protein”

features:

- protein structure
- physical measurements
- scientific papers on this protein



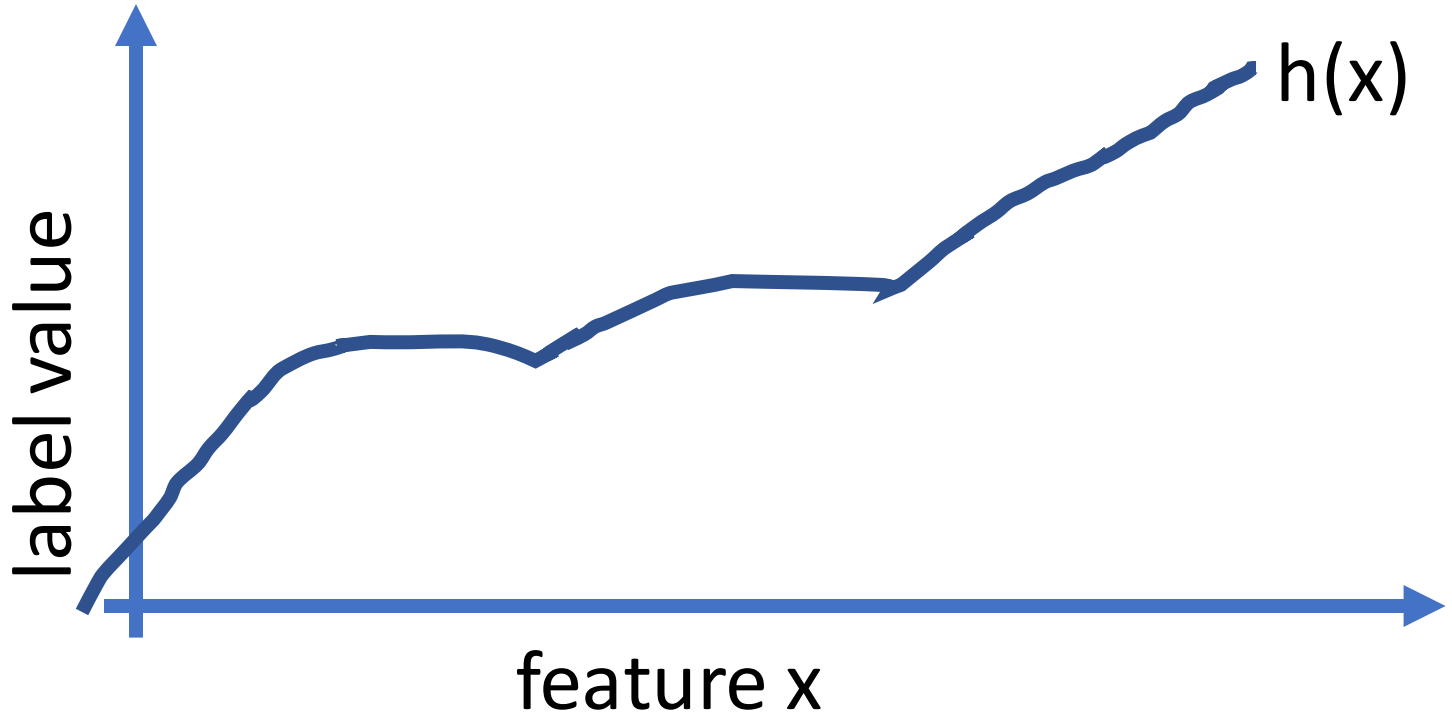
label:

is the protein toxic?

Hypothesis Space

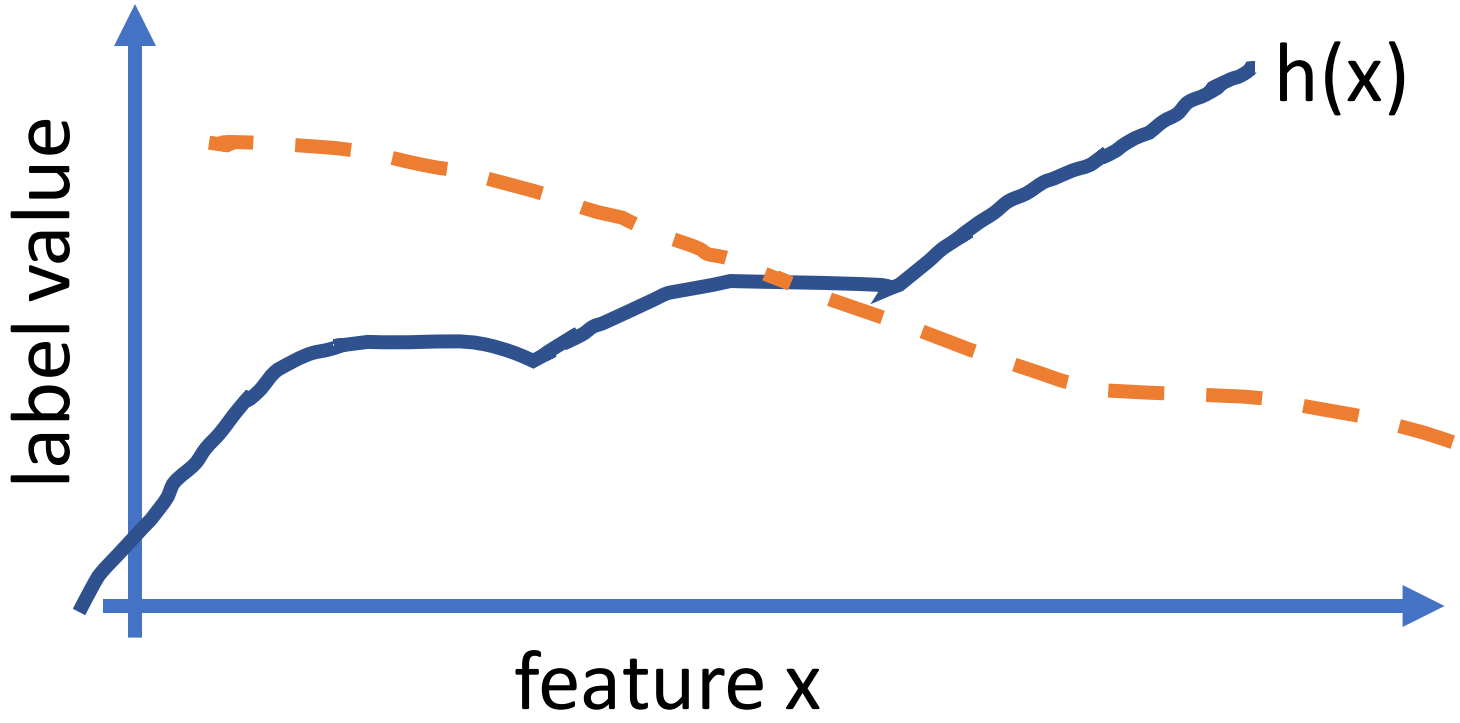
Predictors

- consider some data point with features x and label y
- we would like to predict y based solely on features x
- we use a predictor map h(x) such that $h(x) \approx y$



Hypothesis Space of Predictor Maps

- ML is about finding or learning a good predictor
- we do not have time to search over all possible maps
- there are simply too many of them
- must restrict to a subset of predictor maps

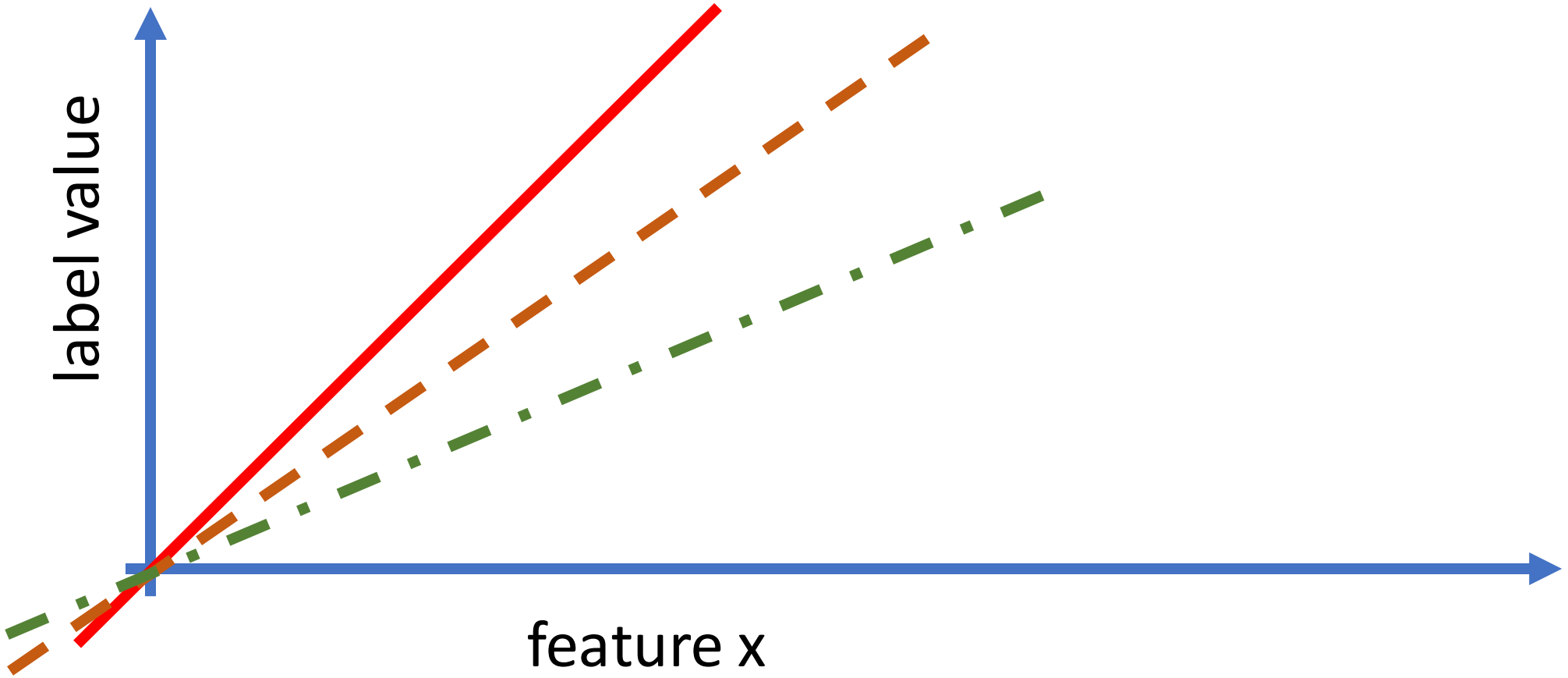


hypothesis space consisting of two predictor maps h(x) and h'(x)

Hypothesis Space of Linear Predictors

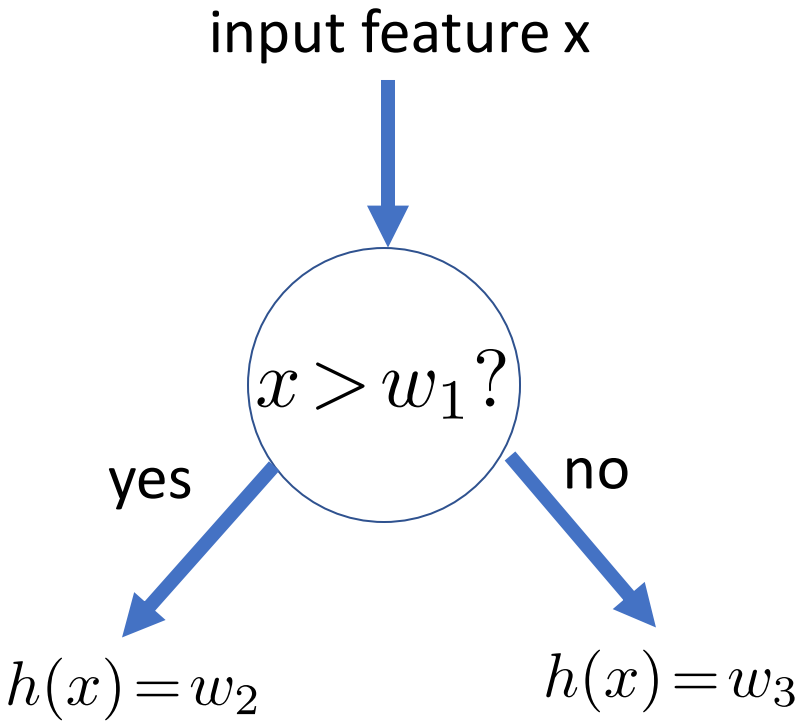
- data points with feature vectors $\mathbf{x} = \left(x_1, \ldots, x_n\right)^T \in \mathbb{R}^n$
- numeric label $y \in \mathbb{R}$
- linear predictors $h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} = \sum_{i=1}^{n} w_i x_i$
- each predictor defined by weights $\mathbf{w} = \left(w_1, \ldots, w_n\right)^T$

Hypothesis Space of Linear Predictors

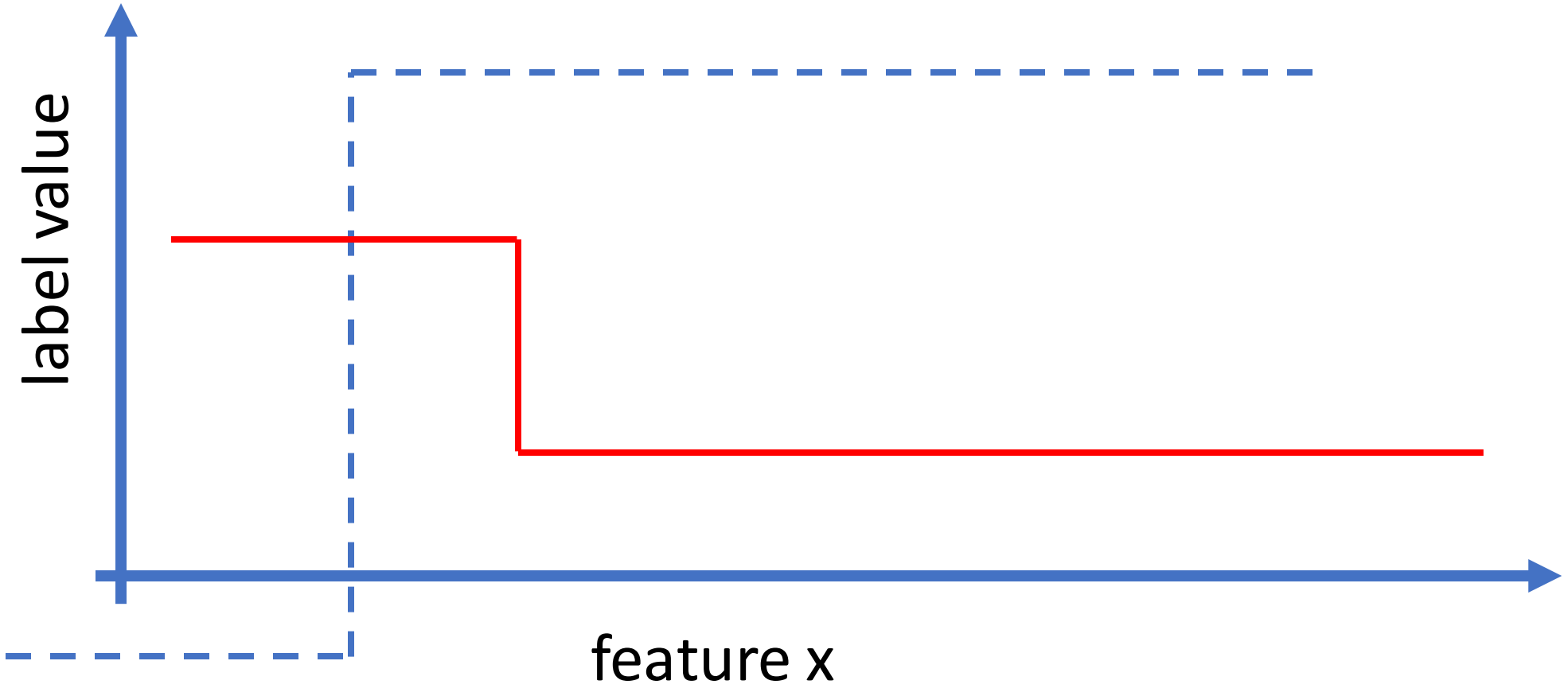


Hypothesis Space of Decision Trees

- represent predictor map by flow chart ("decision tree")



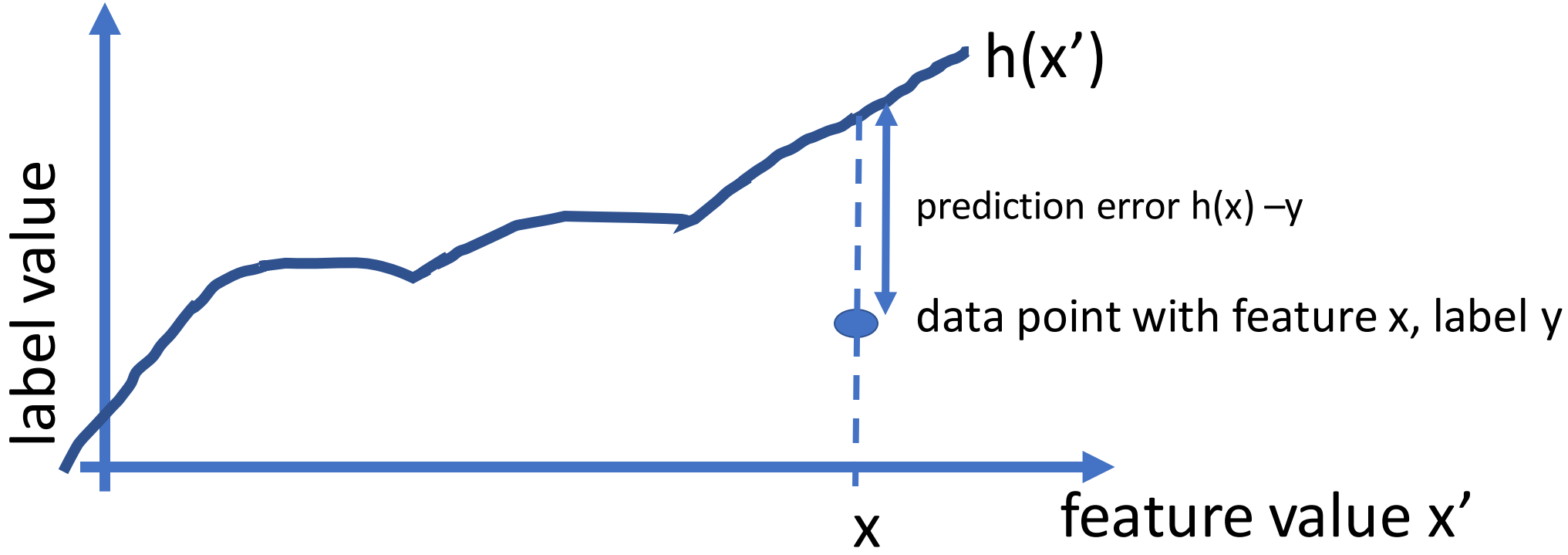
Hypothesis Space of Decision Trees



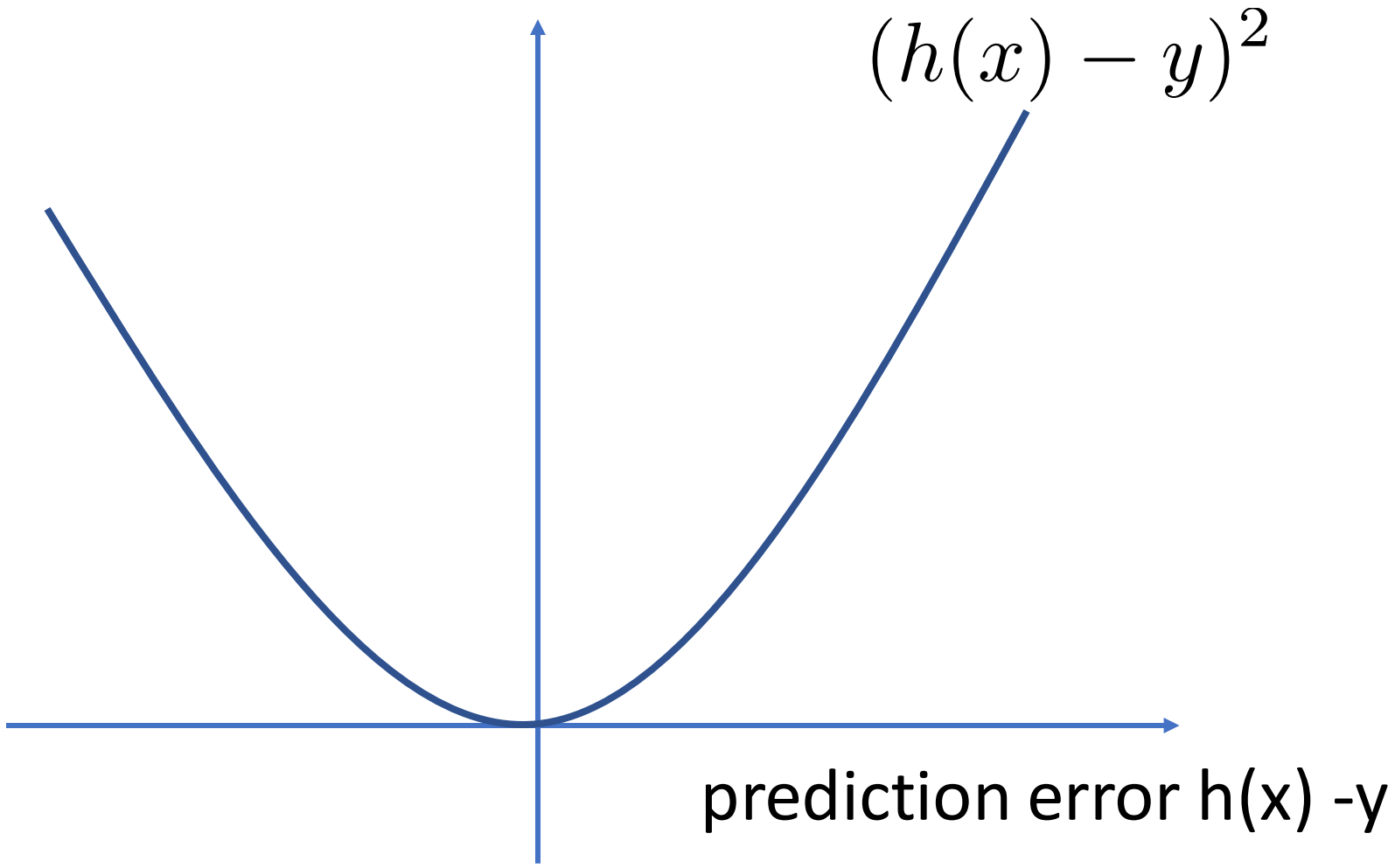
Loss

How Good is a Predictor?

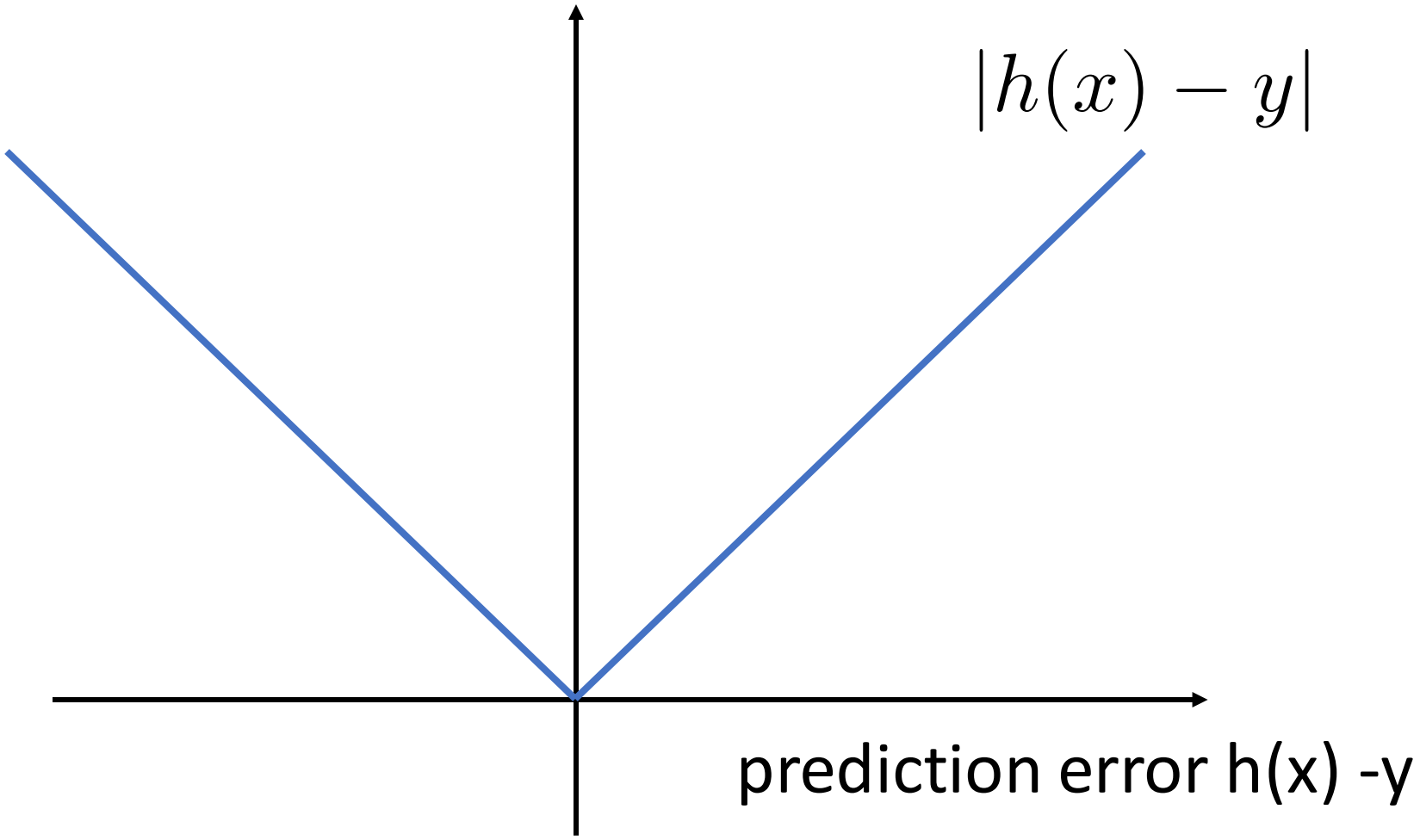
- we use h(x) to predict (estimate) the label y based on features x
- in general, prediction error h(x)-y is not zero
- a loss function measures the "size" of prediction error



The Squared Error Loss



The Absolute Error Loss



Putting Together the Pieces

Design Choices

- what features and labels to use is a design choices
- what hypothesis space to use is a design choice
- what loss function to use is a design choice
- choices must meet statistical and computational requirements

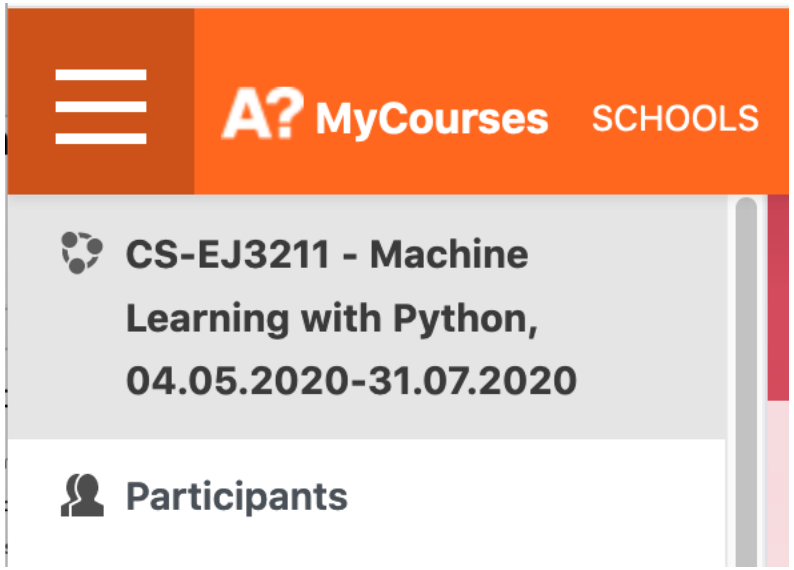
Choosing Loss Function

- squared loss function can be minimized easily
- absolute error loss more difficult to minimize
- squared error loss sensitive to outliers
- absolute error loss robust to outliers

Example: Linear Regression

- features = real numbers
- labels = real number
- hypothesis space = linear predictor maps
- loss = squared error loss

Online Course @ FiTech and Aalto University



https://fitech.io/en/studies/machine-learning-with-python/

FITech
NETWORK UNIVERSITY