INTRODUCTION TO MACHINE LEARNING

ASTR 324, SPRING 2021

STEPHEN PORTILLO

SUPPLEMENTAL READING

Ch 1, 5.1-5.3

Goodfellow, Bengio, and Courville (2016)

https://www.deeplearningbook.org/

Ch 18

Efron and Hastie (2016)

https://web.stanford.edu/~hastie/CASI_files/PDF/casi.pdf

WHAT IS MACHINE LEARNING?

"A computer program is said to learn[…] if its performance at tasks[…] as measured by [some performance measure], improves with experience."

Mitchell (1997)

RULE-BASED SYSTEMS

Does this tweet contain the string “dog”?

This task can be easily codified into a rule:

```
return “dog” in tweet
```

We often use computers to automate tasks that can be codified, like photometry

LIMITS OF RULE-BASED SYSTEMS

Does this image contain a dog?

This task is easy to perform for humans, but how can we get a computer to do it?

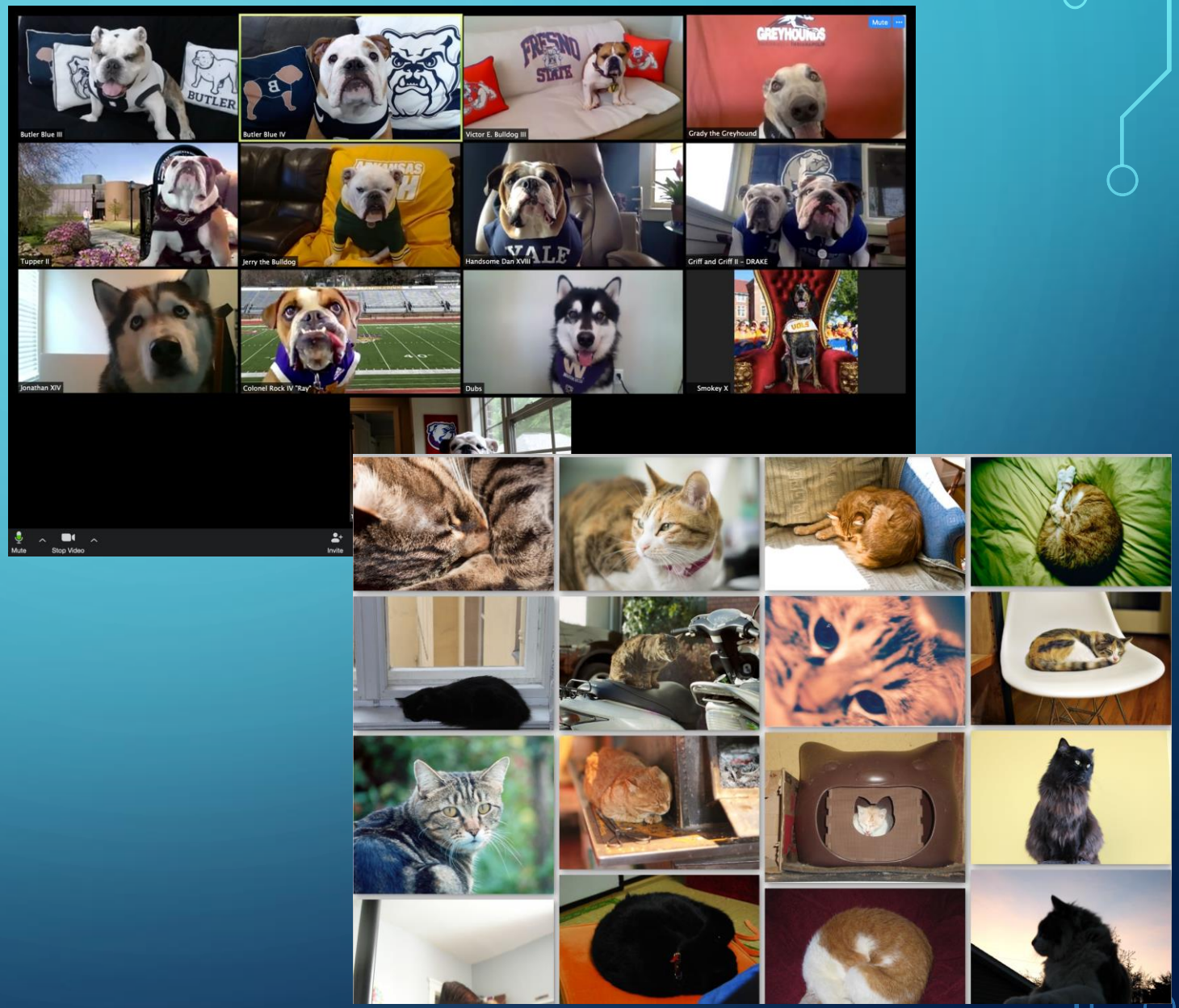
```
return dog in image #?!
```

Can we use computers to automate these messier tasks, like classifying spiral/elliptical galaxies?

MACHINE LEARNING

Instead, we can collect a **dataset** of images that do or do not contain dogs

Using machine learning, we can **train** a **model** that learns how to identify dogs



@UWDubs, Open Images Dataset V6

LIMITS & CONSEQUENCES OF MACHINE LEARNING

A machine learning model is only as good as the dataset it learns from

Biases in models can have real consequences

Models do not exist in isolation: they can create negative feedback loops

ProPublica: Machine Bias

Buolamwini: How I'm fighting bias in algorithms

Togootogtokh & Amartuvshin (2018)

Lum & Isaac (2016)

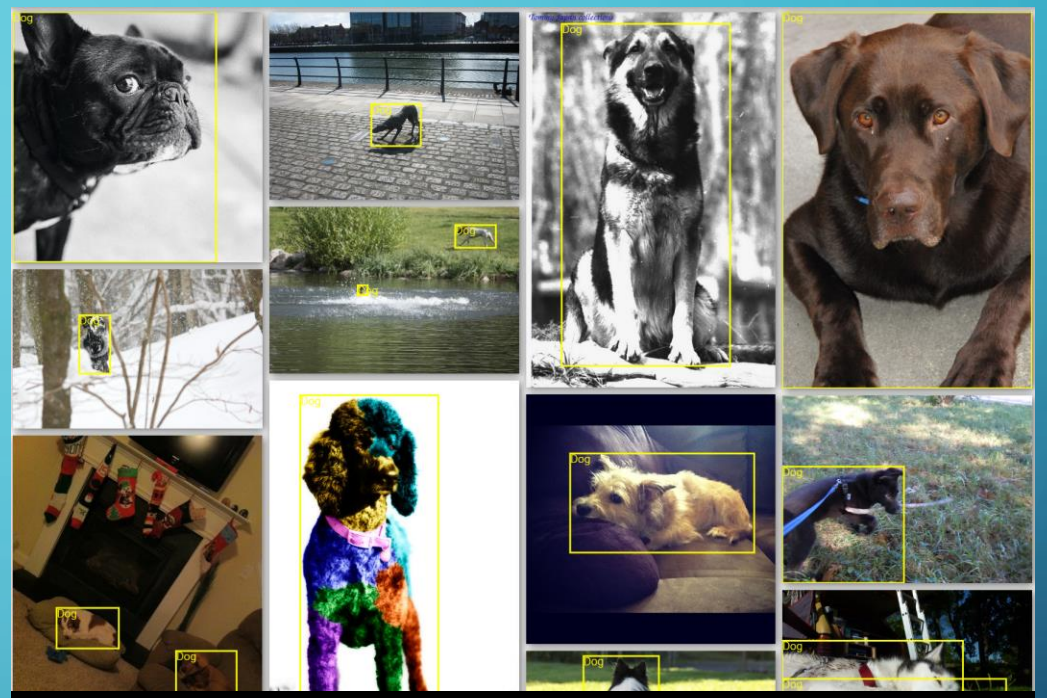
SUPERVISED LEARNING

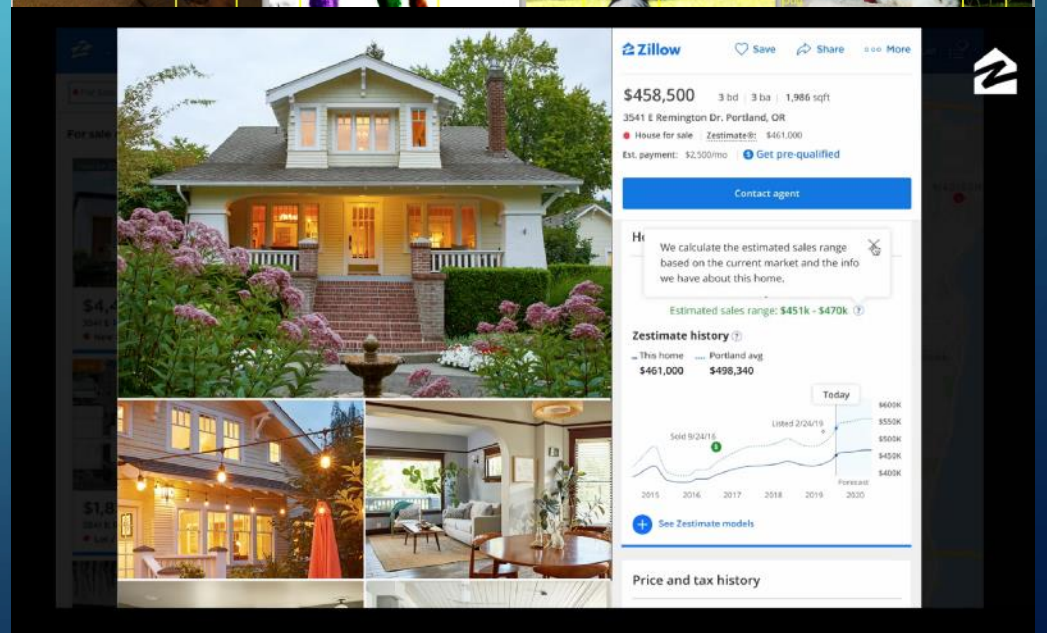
Dataset has **features** and a **label/target**

We want the model to use the features to predict the target

Classification: target is a category

Regression: target is a number





Open Images Dataset V6, Zillow

UNSUPERVISED LEARNING

Dataset has features and the model learns something useful about the structure of the dataset

Generative Adversarial Networks (GANs) trained on a dataset of faces can generate new faces

Supervised/unsupervised is a spectrum

GPT-2 is trained to predict only the next word in text, but learns to mimic writing

SYSTEM PROMPT (HUMAN-WRITTEN)	*Legolas and Gimli advanced on the orcs, raising their weapons with a harrowing war cry.*
MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)	The orcs' response was a deafening onslaught of claws, claws, and claws; even Elrond was forced to retreat. "You are in good hands, dwarf," said Gimli, who had been among the first to charge at the orcs; it took only two words before their opponents were reduced to a blood-soaked quagmire, and the dwarf took his first kill of the night. The battle lasted for hours until two of the largest Orcs attempted to overwhelm Aragorn. When they finally stopped, they lay defeated and lifeless for miles and miles.

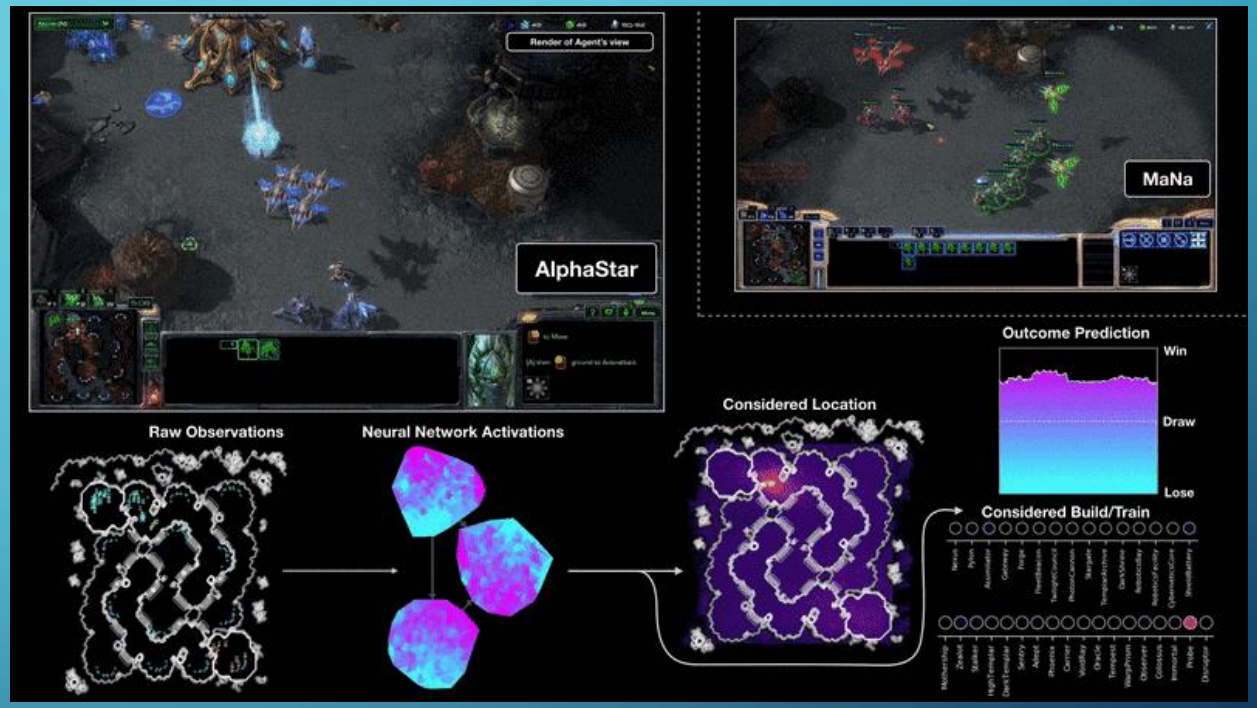
https://www.thispersondoesnotexist.com/

Karras et al. (2018), Radford et al. (2019)

REINFORCEMENT LEARNING

The dataset is not fixed – instead the model interacts with an environment

AlphaStar achieved Grandmaster level in Starcraft II by learning from human games and self-play



Vinyals et al. (2019)
https://youtu.be/nbiVbd_CEIA

BUILDING A MACHINE LEARNING MODEL

Dataset

Loss Function

Model

Optimization Procedure

LINEAR REGRESSION

Dataset: measurements of the redshift z and distance modulus μ of Type Ia supernovae

z	μ
0.4686	41.97
0.7455	43.10
0.0294	35.69
0.3832	41.10
0.2622	40.95
0.2116	39.92

Loss Function: we want a prediction $\hat{\mu}(z)$ – let's use mean squared error

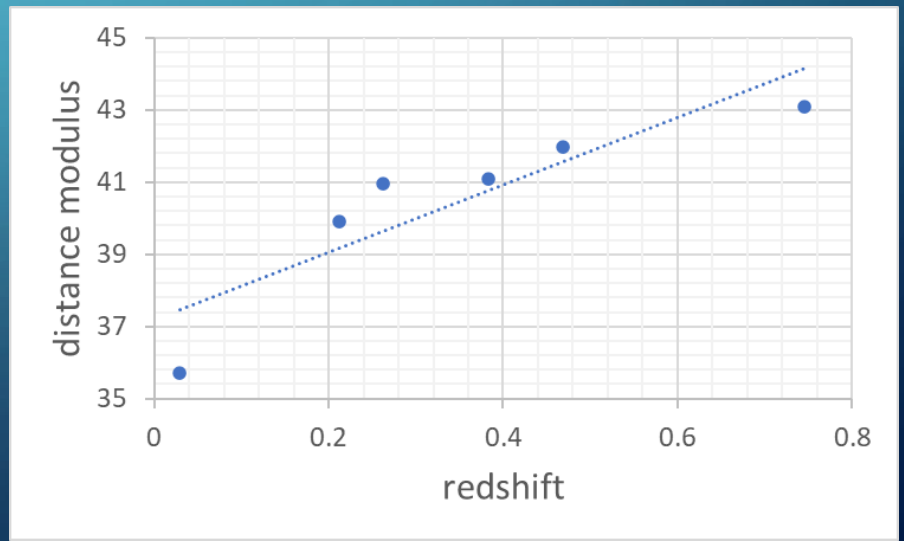
$$L = \frac{1}{N}\sum\left(\mu - \hat{\mu}(z)\right)^2$$

Model: Let's start simple – linear regression

$$\hat{\mu}(z) = a + b\,z$$

Optimization procedure:

$$\frac{\partial L}{\partial a} = 0 \quad \frac{\partial L}{\partial b} = 0$$



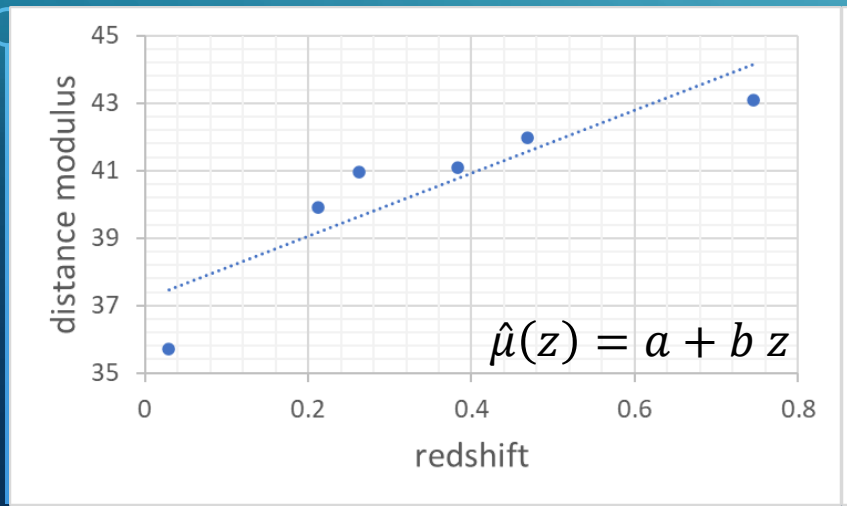
CAPACITY

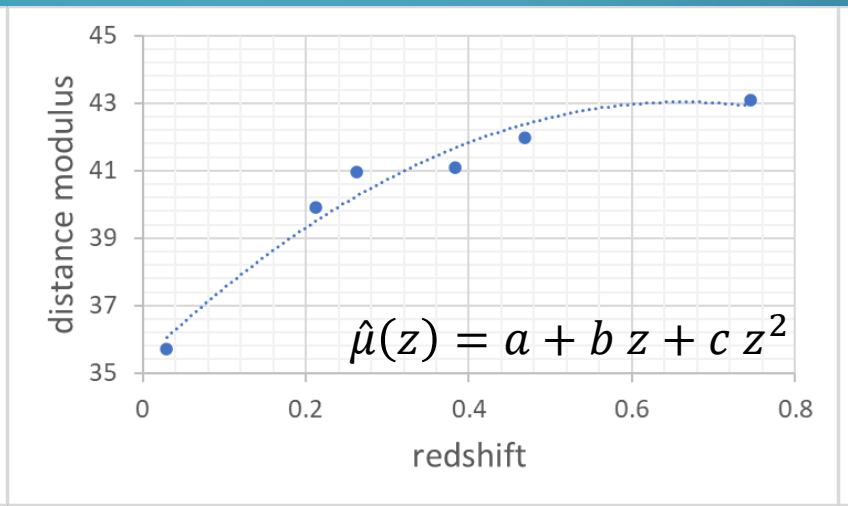
We can get a better fit to the dataset by making the model more flexible, increasing its **capacity**

We can increase capacity by using higher order polynomials

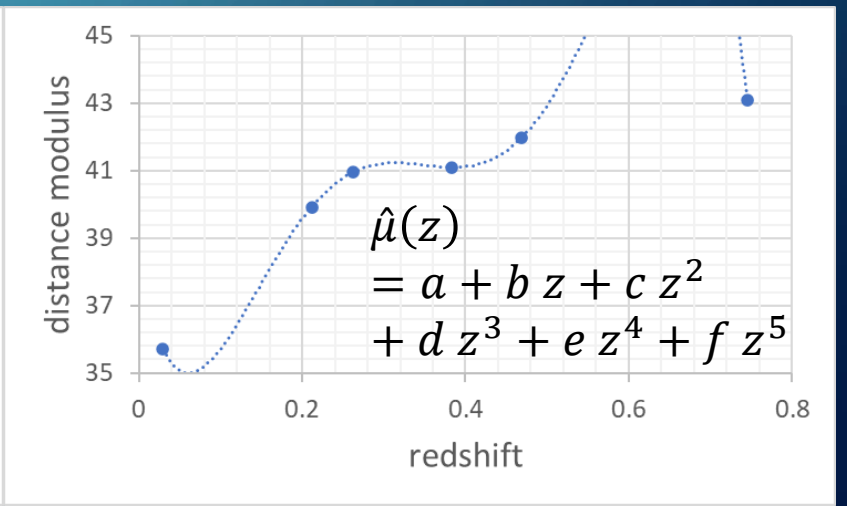
In fact, using a 5th order polynomial as the model gives us zero loss

low capacity





high capacity



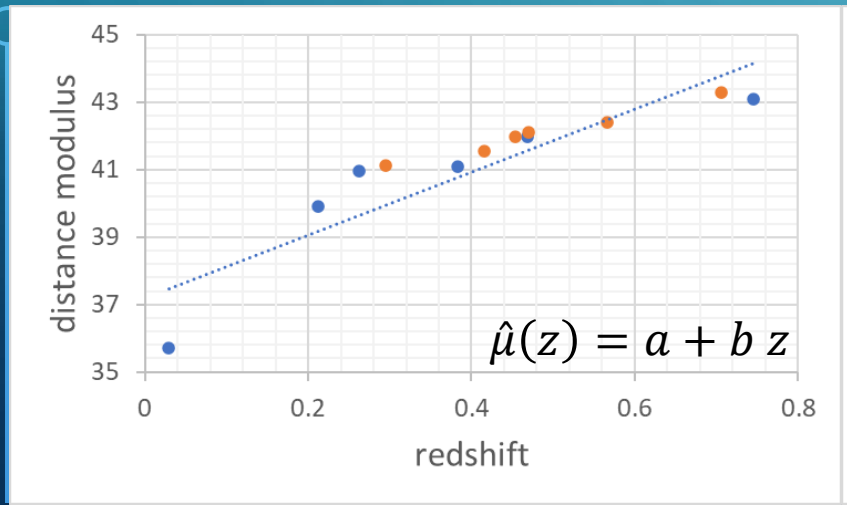
GENERALIZATION ERROR

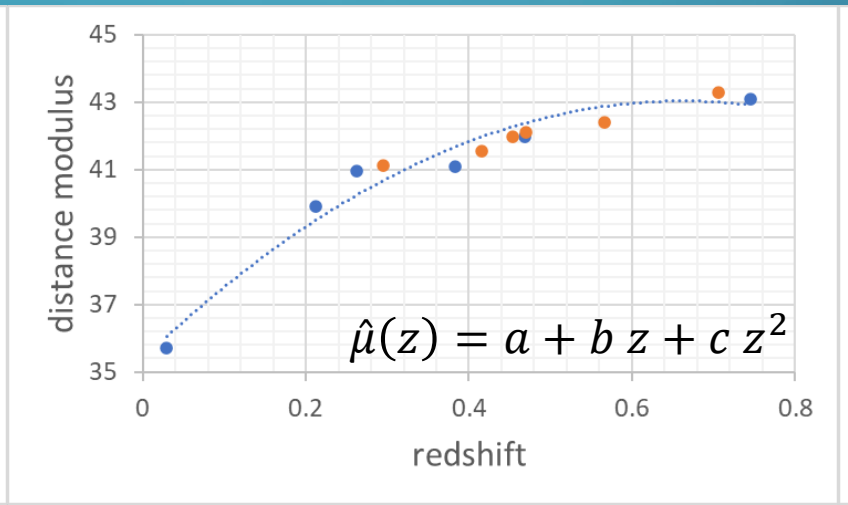
But we want our model to perform well on data it was not trained on

Use a separate **validation set** to see how well the models generalize

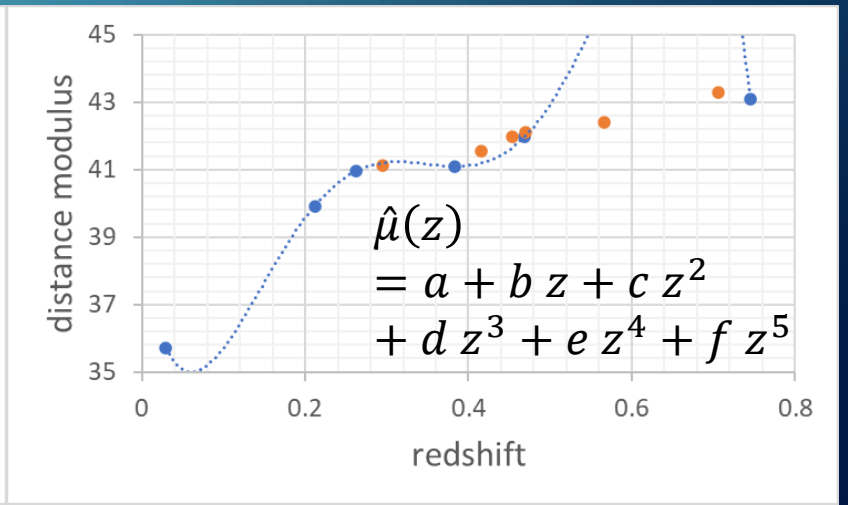
Underfitting occurs when capacity is too low, **overfitting** when it is too high

underfitting





overfitting

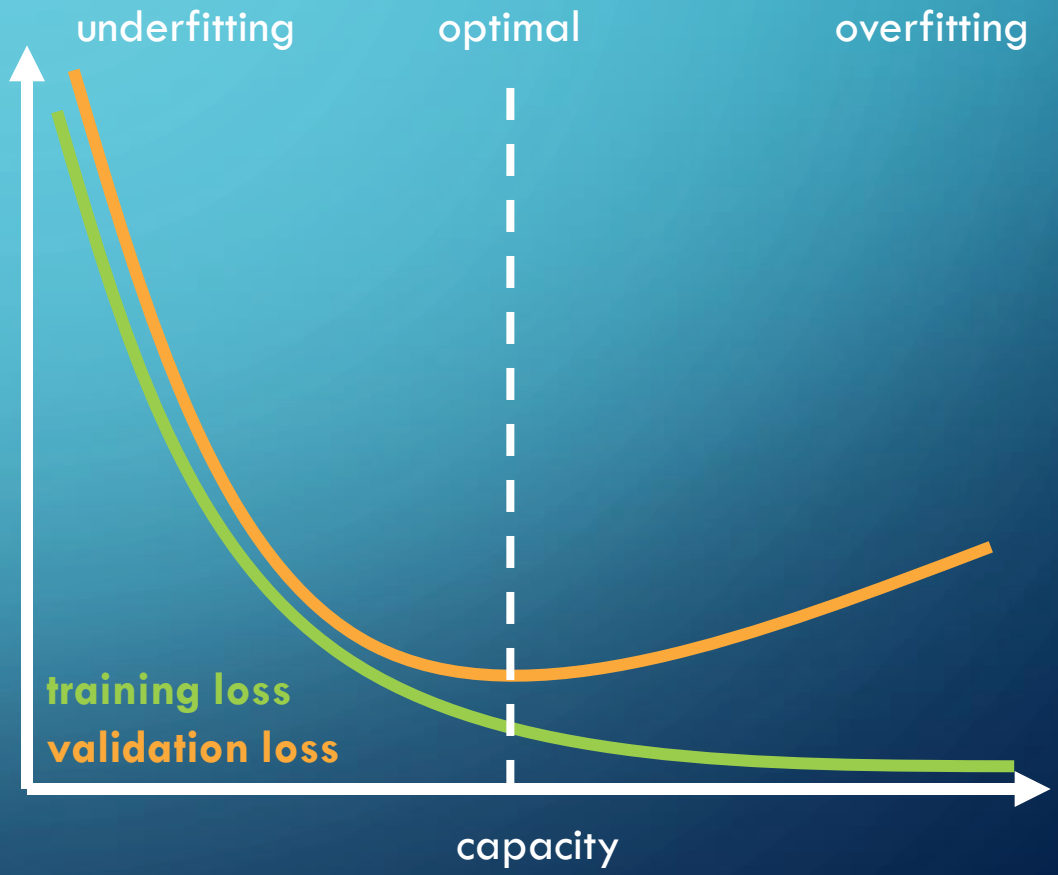


VALIDATION AND TEST SETS

Model choices (like polynomial order) are often called **hyperparameters**

Choose hyperparameters that minimize validation loss

When comparing to others' work, compare the loss on a separate **test set**



ISN'T THIS JUST STATISTICS?

Larry Wasserman:

"They are both concerned with the same question: how do we learn from data?"

My take:

Statistics – you have a model you might actually believe

$$H^2 = H_0^2(\Omega_m(1+z)^3 + \Omega_\Lambda)$$

Machine Learning – optimize a really flexible model using a lot of data

scikit-learn algorithm cheat-sheet