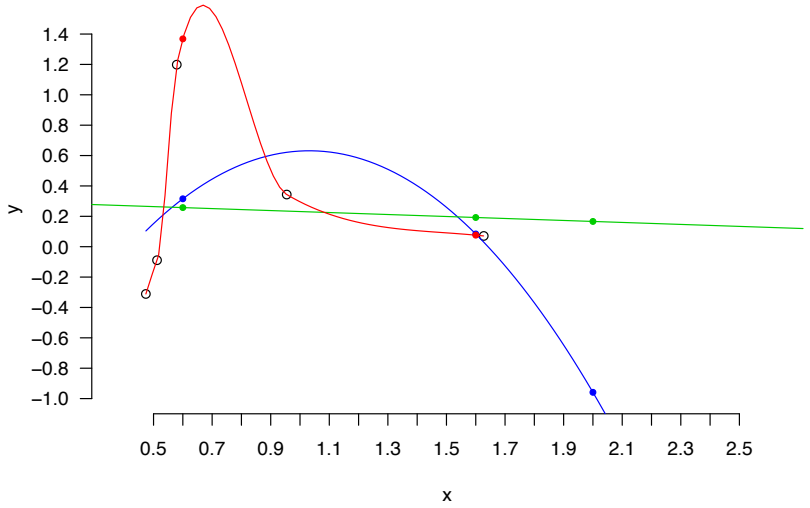
Data Wrangling and Data Analysis
Model evaluation

Daniel Oberski
Department of Methodology & Statistics
Utrecht University

VERSION: 2023-10-10



Model	$\hat{f}(0.6)$	$\hat{f}(1.6)$	$\hat{f}(2.0)$
Eyeballing	?	?	?
Linear regression	0.192	0.192	0.166
Linear regression w/ quadratic	0.315	0.084	-0.959
Nonparametric	1.368	0.076	-
THE REAL TRUTH!, $f(x)$	0.775	1.265	1.414

And the winner is…

Model	MSE	MSE (interpolation)
Eyeballing	?	?
Linear regression	0.992	0.709
Linear regression w/ quadratic	2.410	0.883
Nonparametric	-	0.883

Truth: $y = \sqrt{x} + \epsilon$, where $\epsilon \sim \text{Normal}(0,1)$.

What happened?

- There were few observations, relative to the complexity of most models (except perhaps linear regression);
- The observed data *were* a random sample from the true "data-generating process", $f(x) + \epsilon$;

BUT

- By chance, patterns appeared that are *not* in the true $f(x)$;
- The more flexible models $\hat{f}(x)$ **overfitted** these patterns.

Thought experiment

- Imagine we had sampled another 5 observations, re-trained all of our models, and predicted again.
- Each time we remember the predictions given.
- We do this a large number of (say, 1,000,000,000) times, and then take the average for the predictions over all samples

Questions

1. Which model(s) would give the right prediction **on average**?
2. Which model(s) would give wildly varying predictions?
3. Which model(s) would you *guess* have the lowest MSE overall?

Unbiased

Model that gives the correct prediction, on average over samples from the target population

- Unbiased in our example: nonparametric, square-root
- Biased in our example: all others

High variance:

Model that easily overfits accidental patterns.

- High variance in our example: nonparam., quadratic, sq-root
- Low variance in our example: linear regression

Bias and variance in our example

Some, possibly surprising, conclusions:

- The **best** model in one sense is the **worst** model in the other!
- The “correct” model is not (necessarily) the best model!

Bias-variance tradeoff

- Flexibility → less bias
- Flexibility → more variance

Bias and variance are implicitly linked because they are both affected by **model complexity** (“flexibility”, “capacity”)

What do you mean by “complexity”?

- Amount of information in data absorbed into model;
- Amount of compression performed on data by model;
- Number of effective parameters, relative to effective degrees of freedom in data.

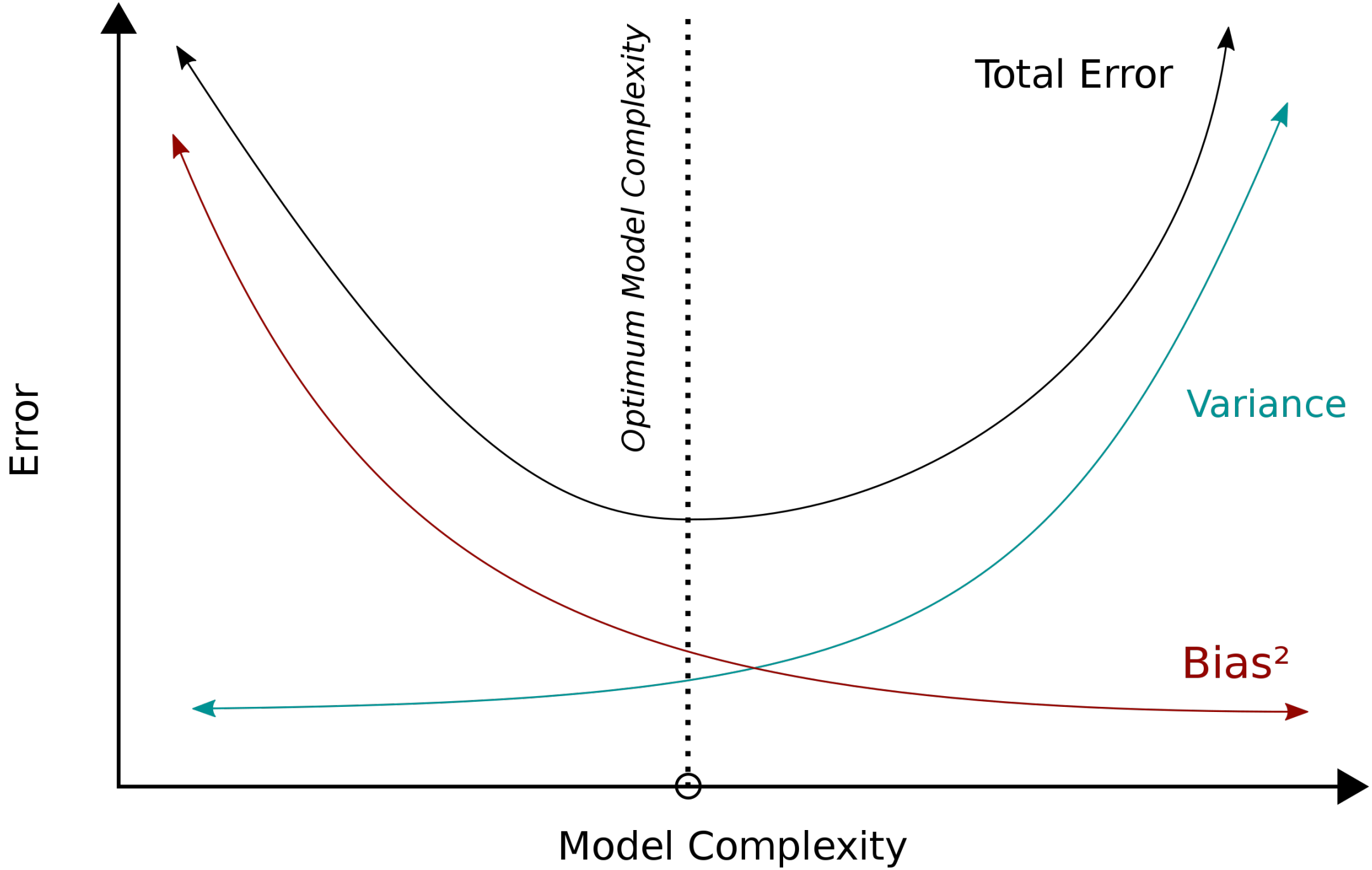
Examples of things that make model **more “complex”**:

- More predictors in linear regression
- Higher-order polynomial in linear regression (x^2, x^3, x^4, etc.);
- Smaller “neighborhood” in kNN
- ...

Question:

Does the bias-variance tradeoff occur with $n = 5$?

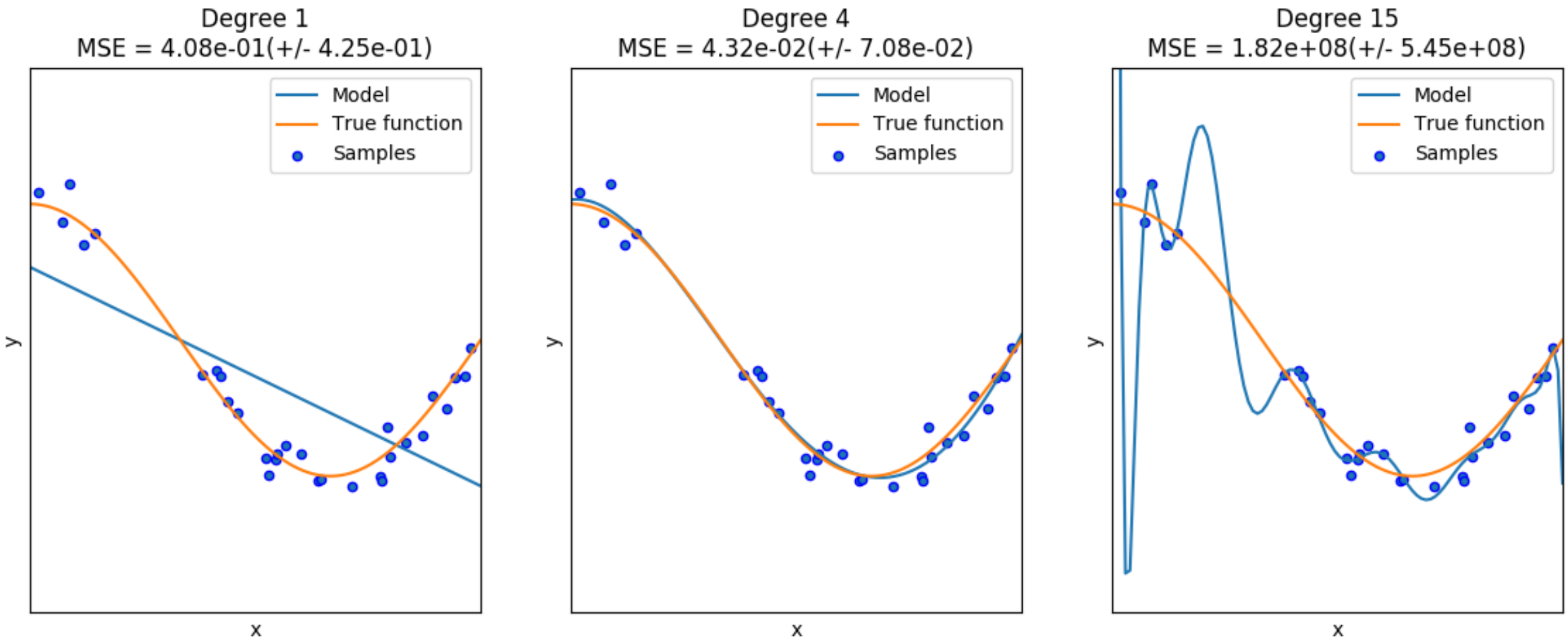
Does the bias-variance tradeoff occur with $n = 5{,}000{,}000{,}000$?


Error
Optimum Model Complexity
Total Error
Variance
Bias²
Model Complexity

$$E(\text{MSE}) = \text{Bias}^2 + \text{Variance} + \sigma$$

Population mean squared error is squared bias PLUS model variance PLUS irreducible variance.

(The E means “on average over samples from the target population”).


Degree 1
MSE = 4.08e-01(+/- 4.25e-01)
Model
True function
Samples
y
x
Degree 4
MSE = 4.32e-02(+/- 7.08e-02)
Model
True function
Samples
y
x
Degree 15
MSE = 1.82e+08(+/- 5.45e+08)
Model
True function
Samples
y
x

The train-val-test paradigm

Will my model succeed?

These factors should determine your success:

1. How doable the problem is in the first place: irreducible error;
2. How complex the model $\hat{f}(\boldsymbol{x})$ is;
3. How complex the true function $f(\boldsymbol{x})$ is;
4. The sample size.

All tricks of the trade attack one or more of these!

Problem	**Some ideas for plan of attack**	**Example**
Irreducible error	Get more features; Reduce measurement error	LIDAR on car; Multiple rating radiologists
Model complexity	Try models with range of complexity; Include prior knowledge in the model	Download pretrained model and use that as starting point
Task complexity	Choose something easier; Influence the process	Paint road signs for self-driving
Sample size	Get more examples	Why not use all of Wikipedia for NLP?

Question: What observations are we supposed to take the "average" over when calculating metrics for $\hat{f}(\boldsymbol{x})$?

A. Observations used to fit $\hat{f}(\boldsymbol{x})$.

B. New observations from the same source.

C. New observations from the intended prediction situation.

D. Other.

Back to reality!

Problem:

- Wait, we don’t actually have the population!
- And our training data were already used to train the model…

Solution:

- Instead, we will take a new, pristine, sample from population:
- The **test data**

Train/dev/test

Training data:

Observations used to train ("fit", "estimate") $\hat{f}(\boldsymbol{x})$

Validation data (or "dev" data):

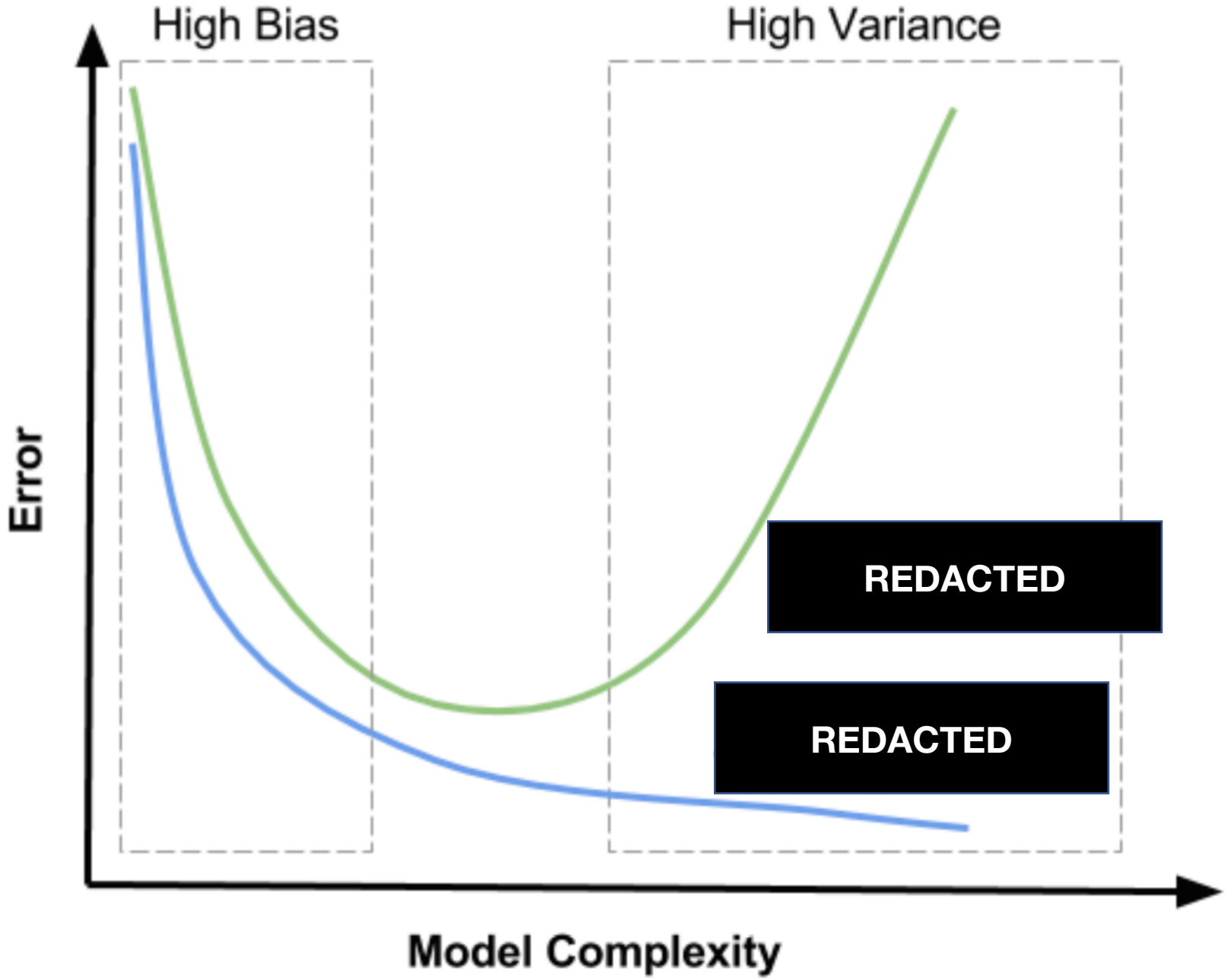
New observations from the same source as training data

Used several times to select model complexity)

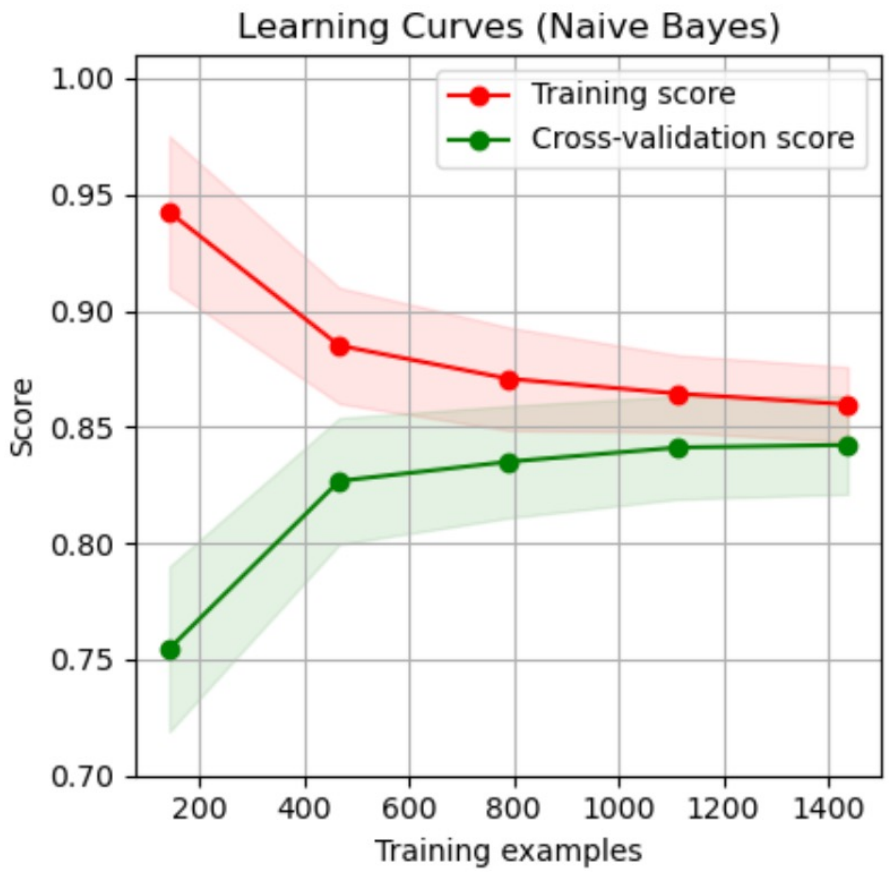
Test data:

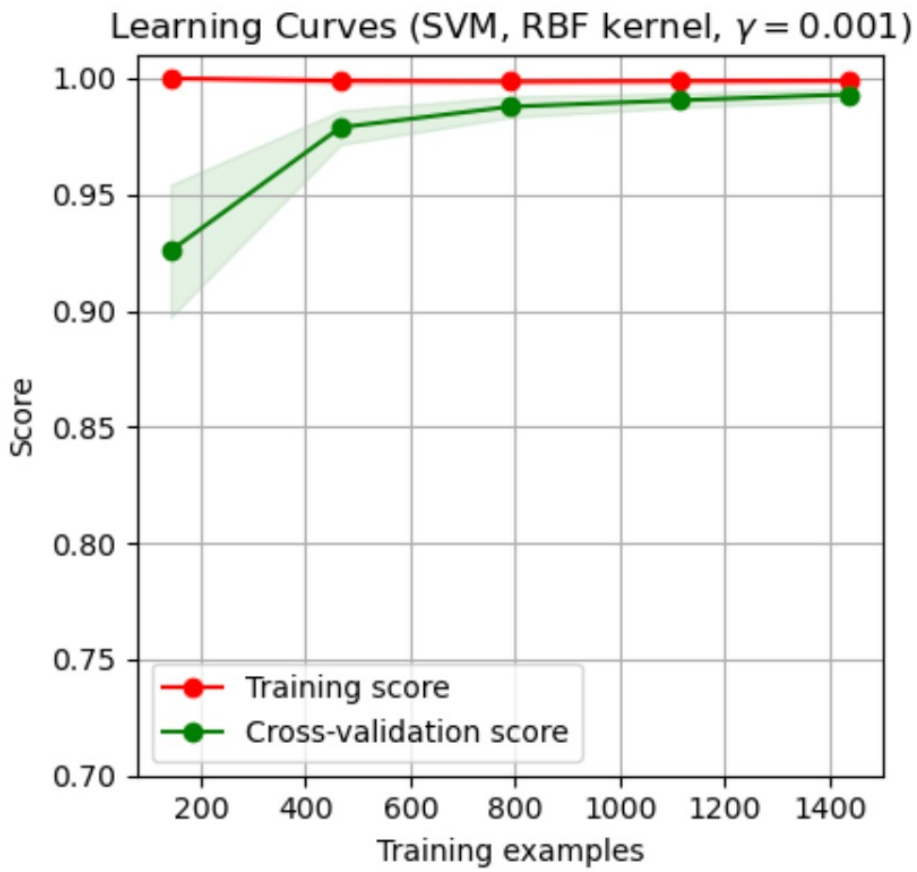
New observations from the intended prediction situation

Question: Why don't these give the same average MSE?

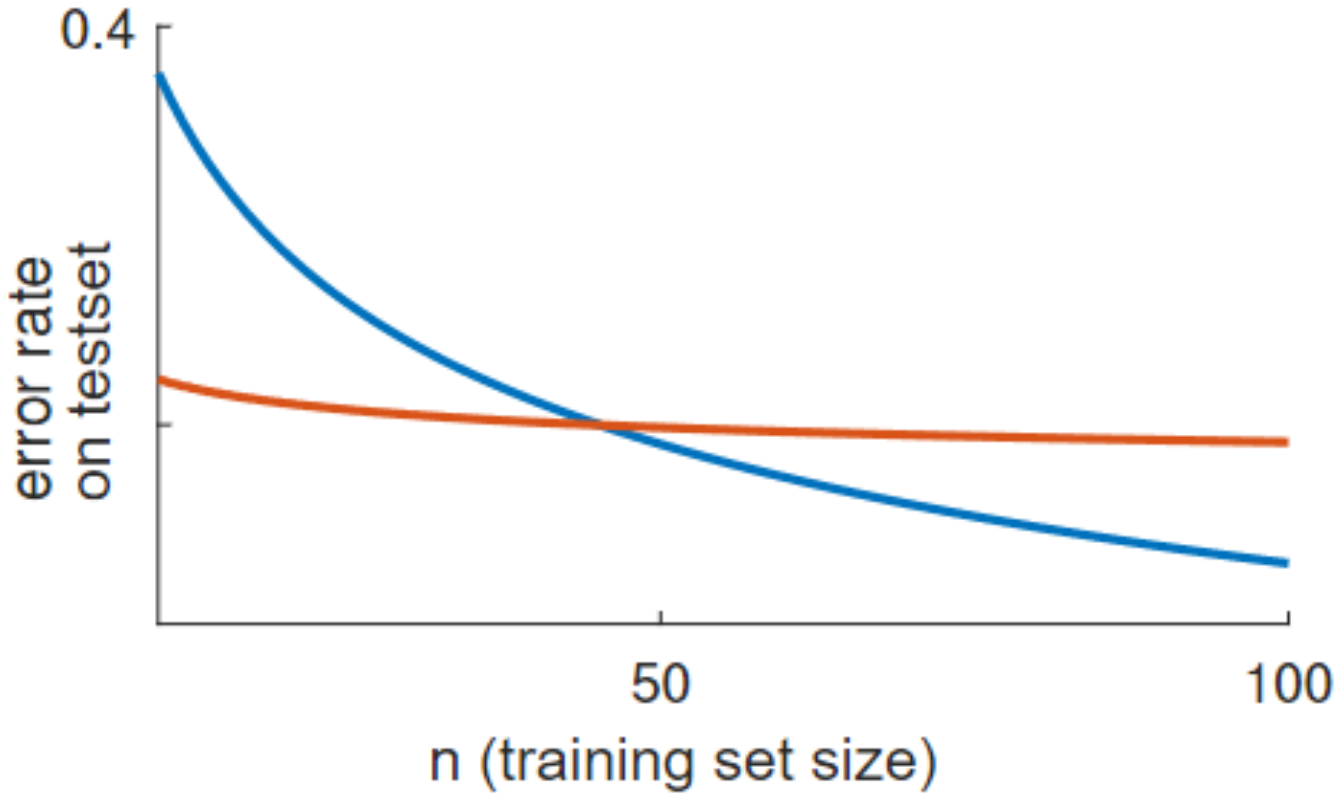

High Bias
High Variance
Error
REDACTED
REDACTED
Model Complexity

Learning curves: *n* vs. performance





Learning curves



Linear regression

Deep learning

Viering & Loog (2021). *The Shape of Learning Curves: a Review*. https://arxiv.org/pdf/2103.10948.pdf

The train-val-test paradigm

- The idea is that the average squared error in the test set MSE_{test} is a good estimate of the "Bayes error" E(MSE)
- This only holds when the test set is "like" the intended prediction situation!

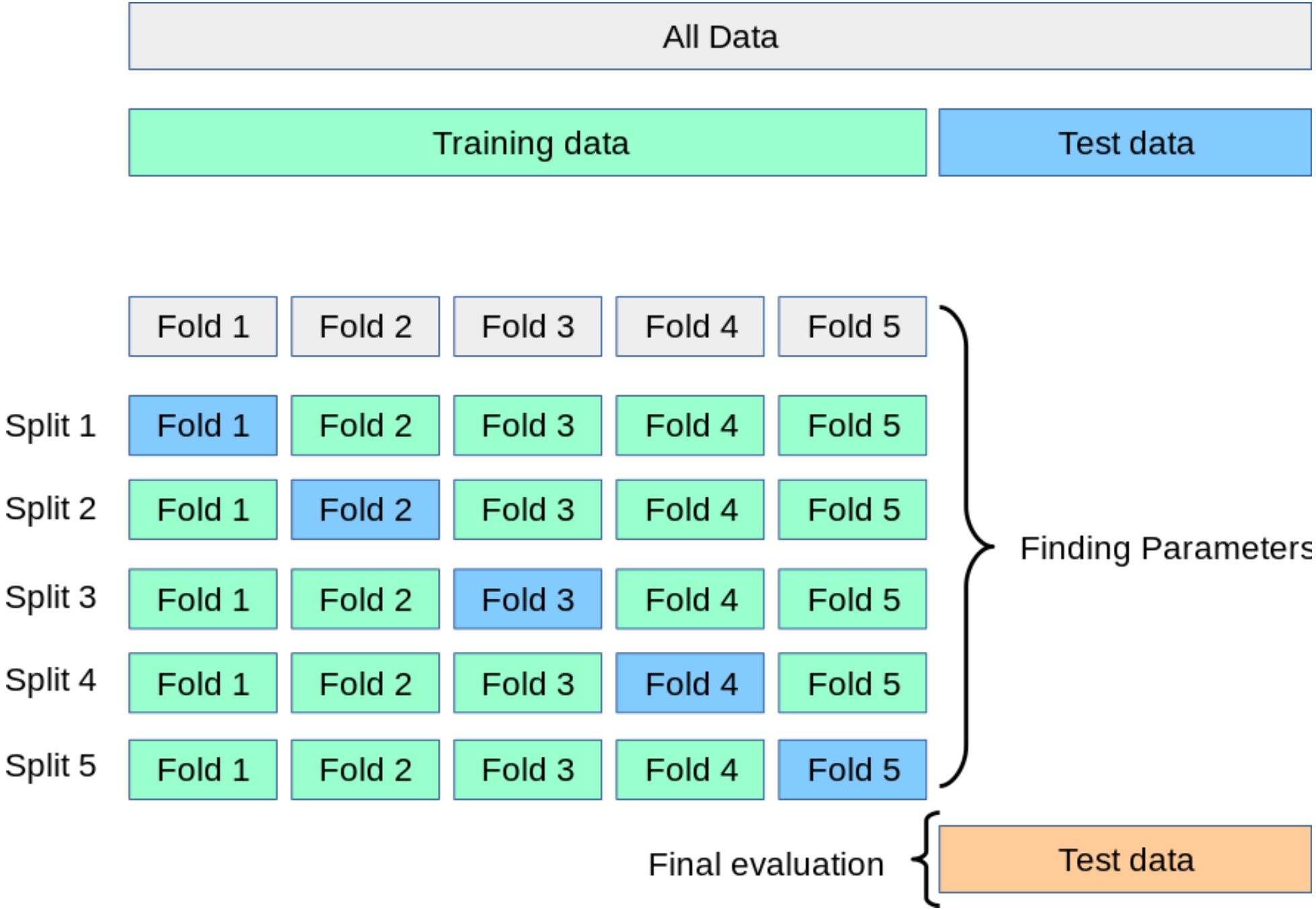
Drawbacks of train/dev/test

- the validation estimate of the test error can be **highly variable**, depending on precisely which observations are included in the training set and which observations are included in the validation set.
- In the validation approach, only a subset of the observations — those that are included in the training set rather than in the validation set — are used to fit the model.
- This suggests that the validation set error may tend to **overestimate the test error** for the model fit on the entire data set.

From `https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning`

K-fold crossvalidation

- “Cross-validation” often used to replace single dev set approach;
- Perform the train/dev split several times, and average the result.
- When K = n, “leave-one-out”;
- Usually K = 5 or K = 10


All Data
Training data
Test data
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Split 1
Split 2
Split 3
Split 4
Split 5
Finding Parameters
Final evaluation
Test data

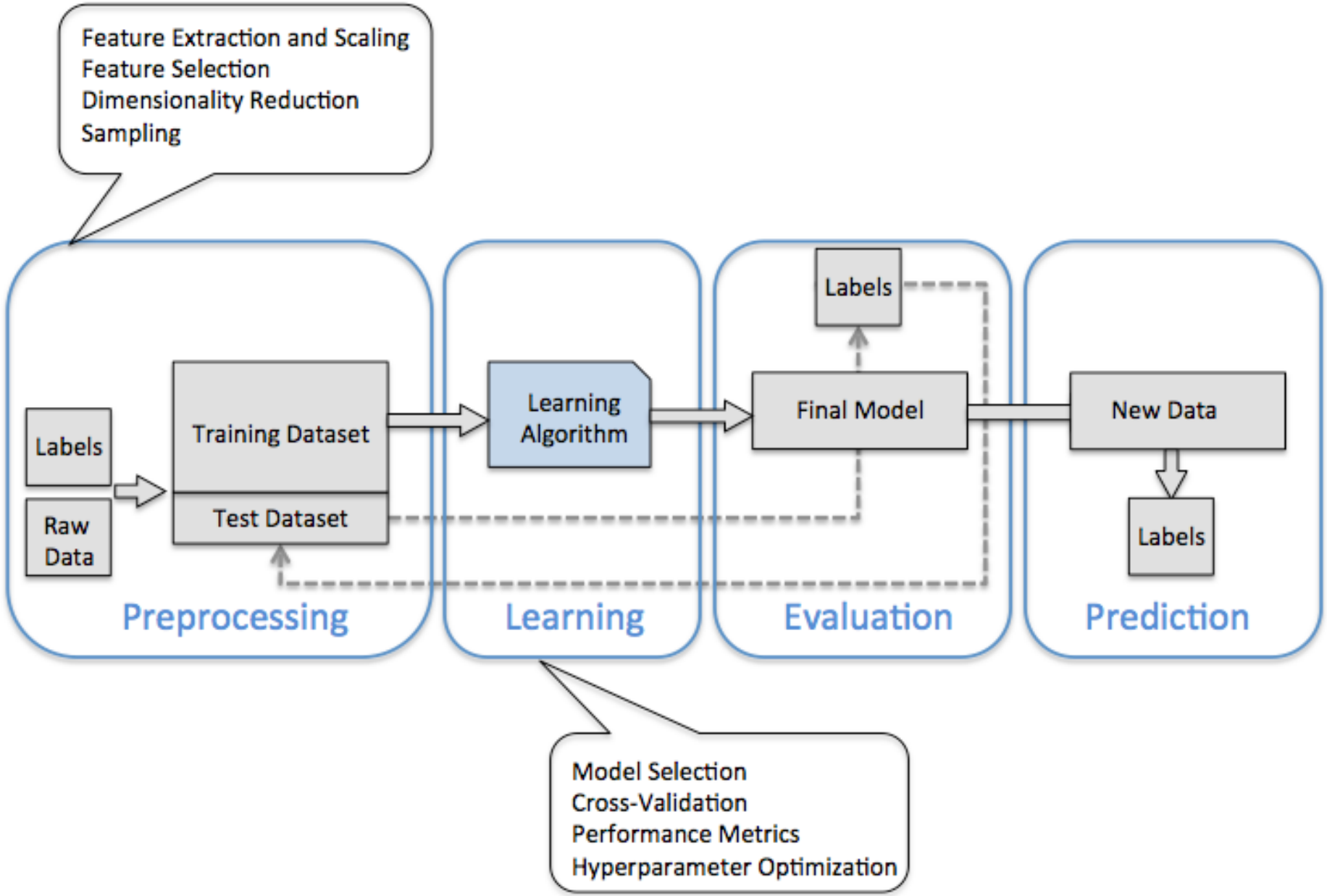
Consider a simple regression used to predict an outcome:

1. Starting with 5000 predictors and 500 cases, find the 100 predictors having the largest correlation with the outcome;
2. We then fit a linear regression, using only these 100 predictors.

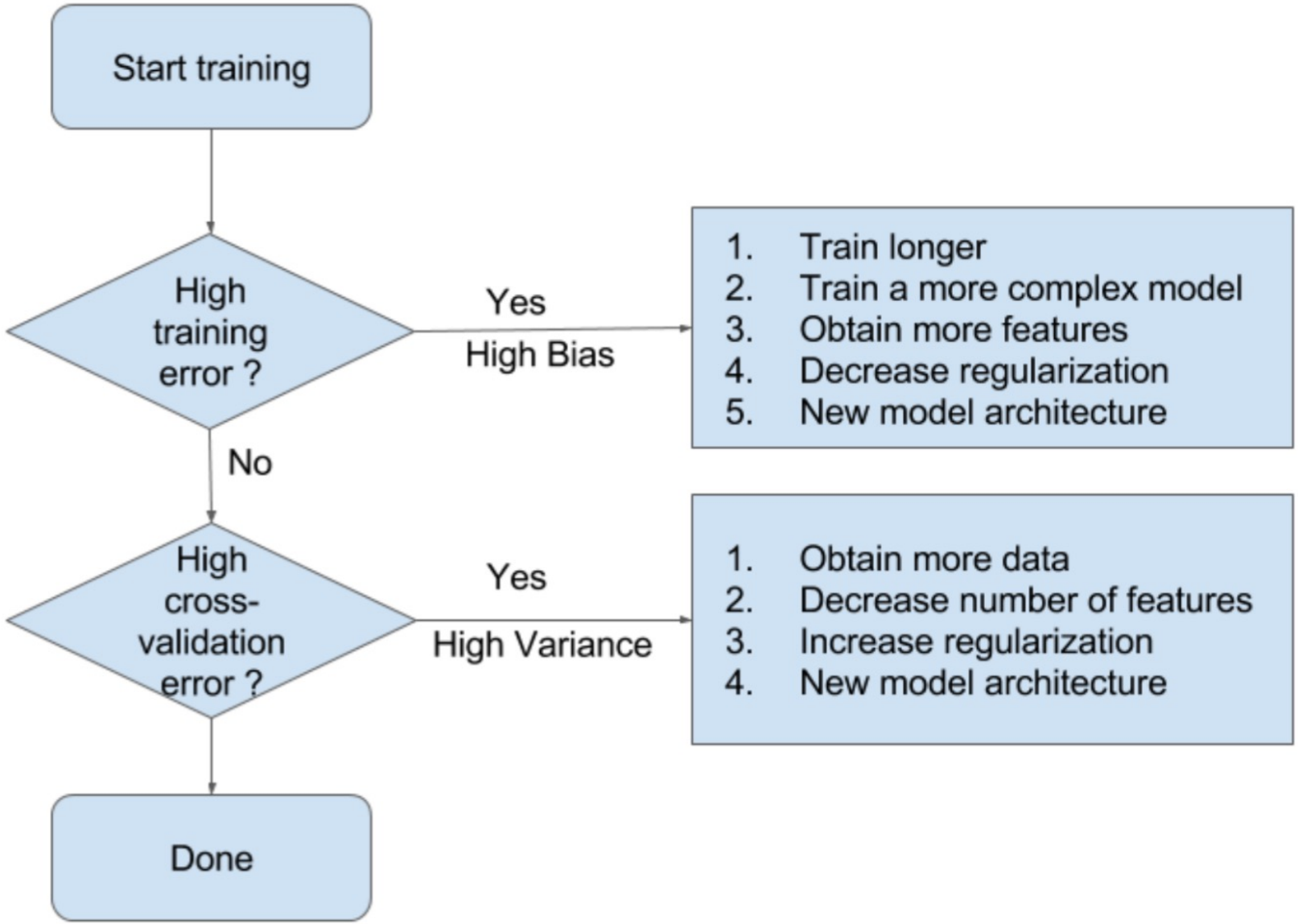
Class exercise:

- How do we estimate the test set performance of this classifier?
- Can we apply cross-validation in step 2, forgetting about step 1?

Answer: In Step 1, the procedure has already seen the labels of the training data, and made use of them. This is a form of training and must be included in the validation process!


Feature Extraction and Scaling
Feature Selection
Dimensionality Reduction
Sampling
Labels
Raw Data
Training Dataset
Test Dataset
Learning Algorithm
Labels
Final Model
New Data
Labels
Preprocessing
Learning
Evaluation
Prediction
Model Selection
Cross-Validation
Performance Metrics
Hyperparameter Optimization

Bias-Variance Flowchart (Andrew Ng, Coursera)



Common task framework (CTF)

a.k.a. “benchmarking”

The Common Task Framework

(a) A **publicly available training dataset**

(b) A set of **enrolled competitors** whose common task is to infer a class prediction rule from the training data.

(c) A **scoring referee**, to which competitors can submit their prediction rule.

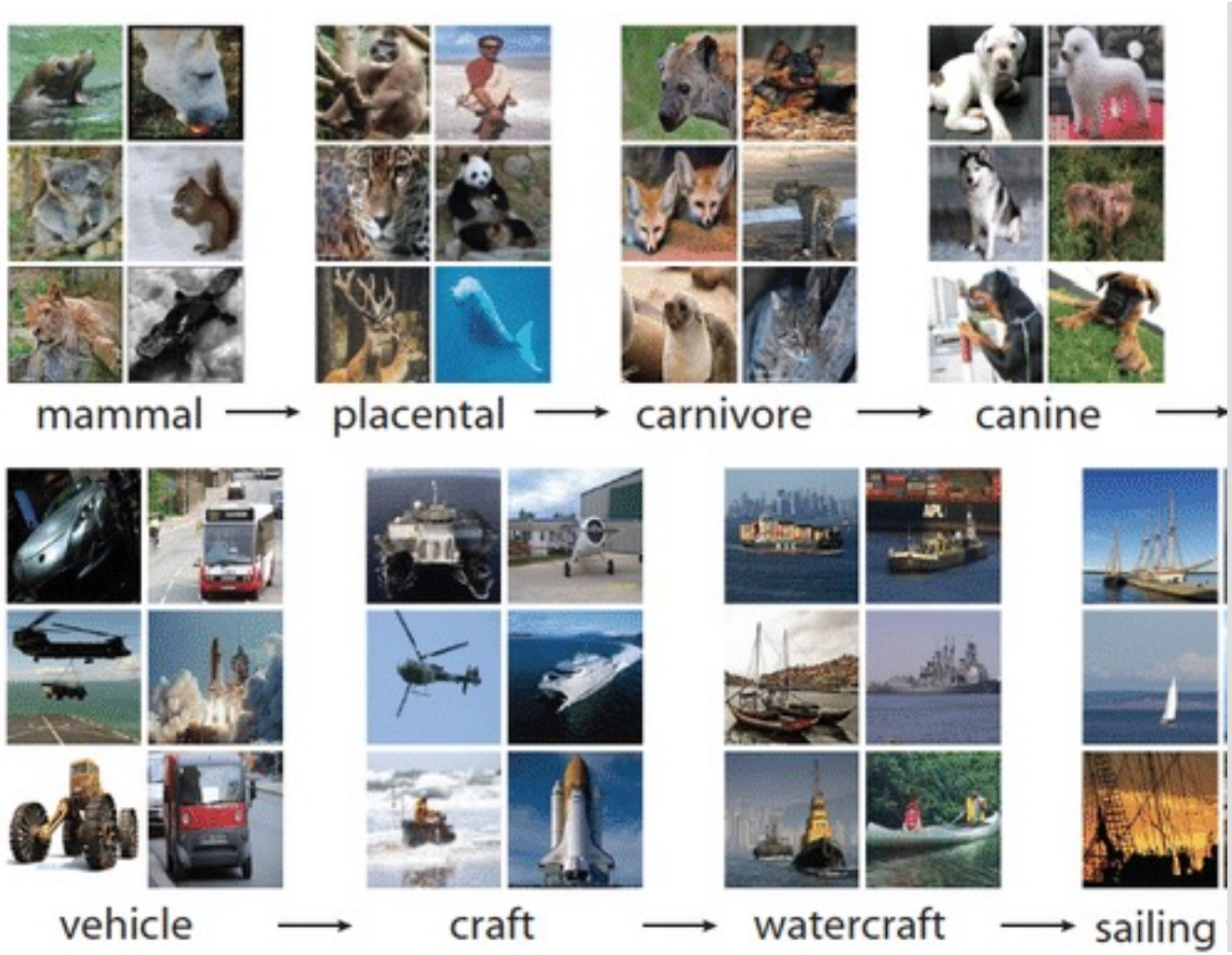
- The referee runs the prediction rule against a testing dataset, which is sequestered behind a Chinese wall.
- The referee objectively and automatically reports the score achieved by the submitted rule.

Donoho (2017). *50 years of data science.* J Comp Graphical Stat.

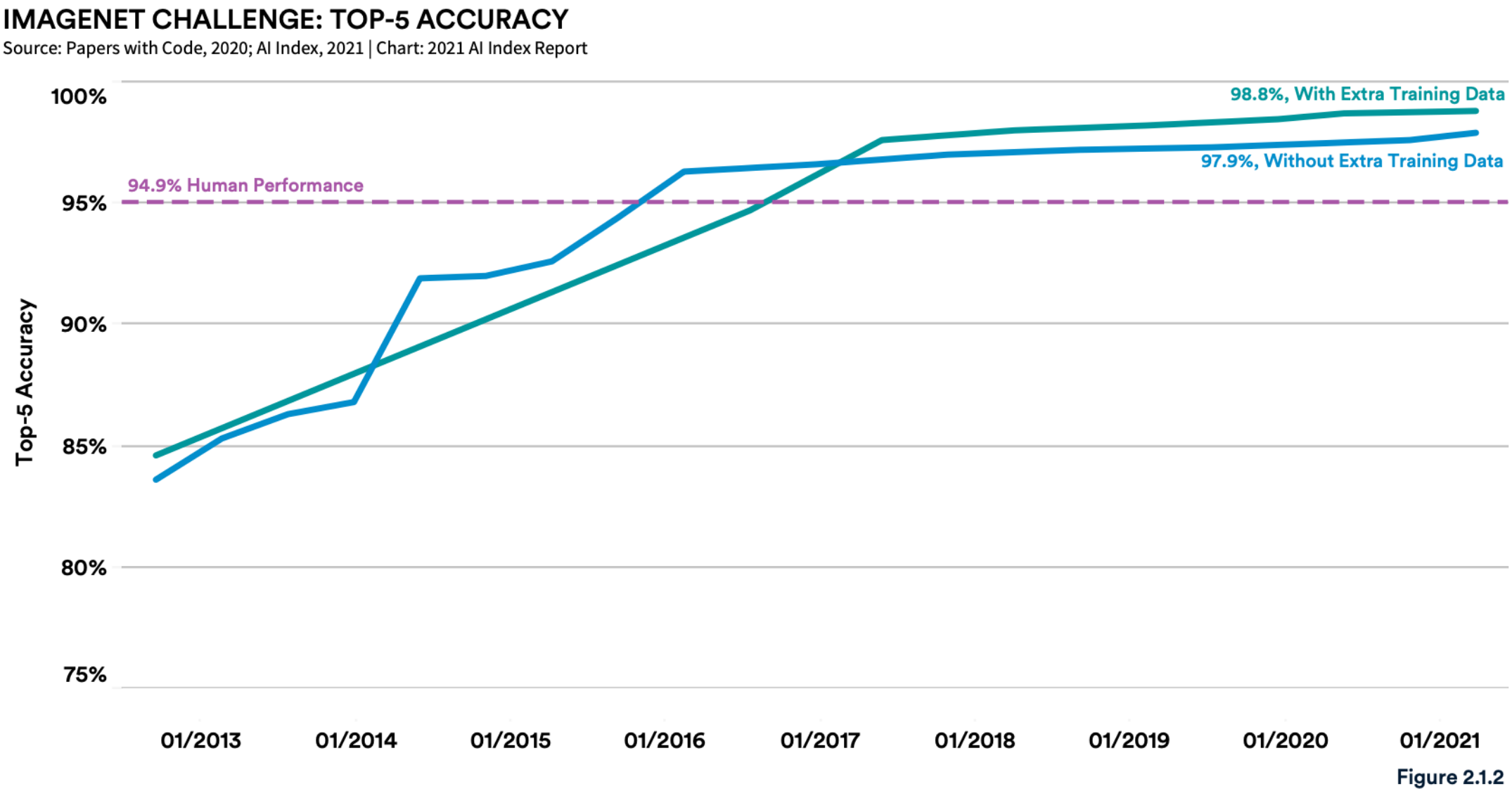
CTF/benchmarking advantages

1. Error rates decline by a fixed percentage each year, to an asymptote depending on task and data quality.
2. Progress usually comes from many small improvements; a change of 1% can be a reason to break out the champagne.
3. Shared data plays a crucial role—and is reused in unexpected ways.

Imagenet



Future Nobel (?) Fei-Fei Li (李飞飞)



https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report_Master.pdf

Scottish_Parliament

The Stanford Question Answering Dataset

Following a referendum in 1997, in which the Scottish electorate voted for devolution, the current Parliament was convened by the Scotland Act 1998, which sets out its powers as a devolved legislature. The Act delineates the legislative competence of the Parliament – the areas in which it can make laws – by explicitly specifying powers that are "reserved" to the Parliament of the United Kingdom. The Scottish Parliament has the power to legislate in all areas that are not explicitly reserved to Westminster. The British Parliament retains the ability to amend the terms of reference of the Scottish Parliament, and can extend or reduce the areas in which it can make laws. The first meeting of the new Parliament took place on 12 May 1999.

When was the current parliament of Scotland convened?
Ground Truth Answers: Following a referendum in 1997 | 1998 | 1998
Prediction: 1998

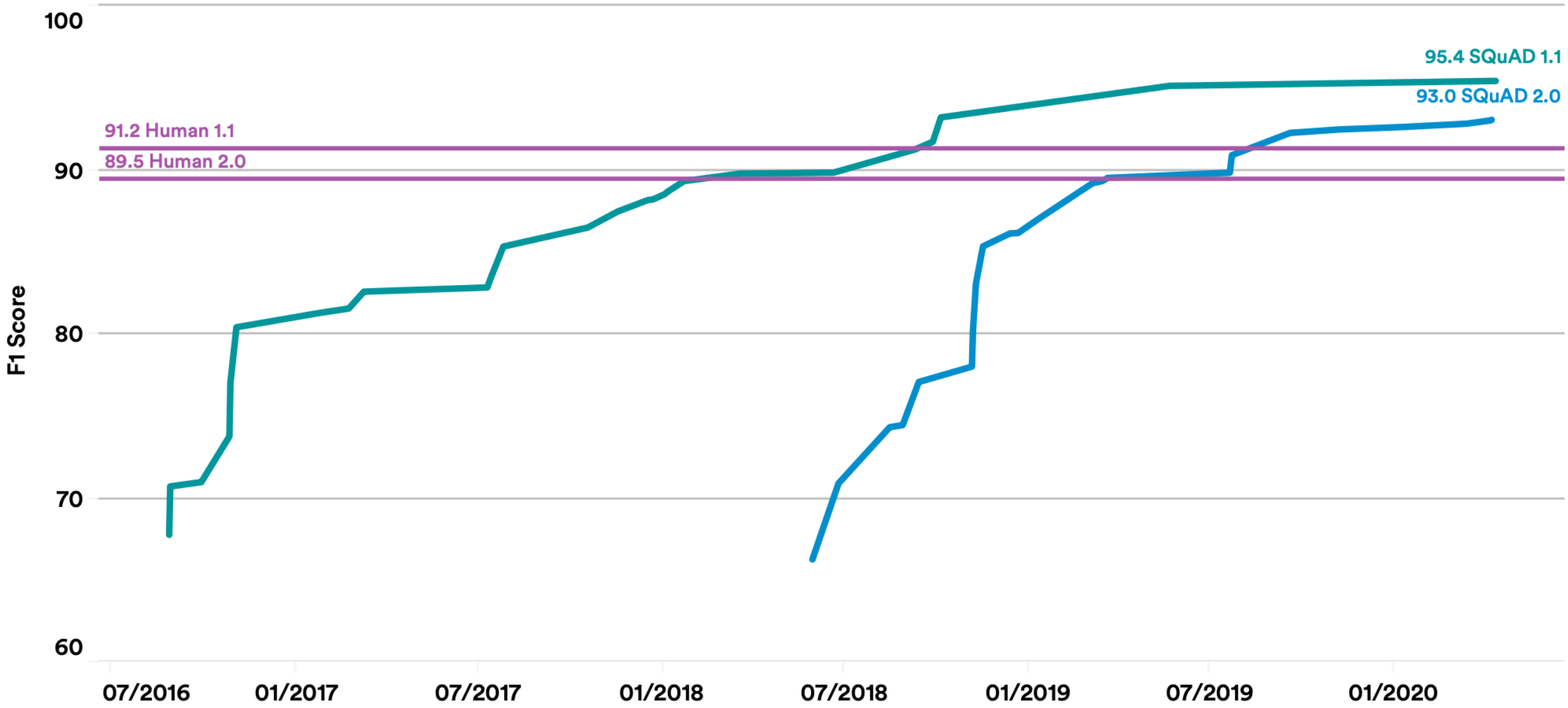
What act set out the Parliament's powers as a devolved legislature?
Ground Truth Answers: Scotland Act 1998 | Scotland Act 1998 | Scotland Act
Prediction: Scotland Act 1998

The legislative competence of the Parliament species what areas?
Ground Truth Answers: in which it can make laws | the areas in which it can make laws | powers that are "reserved" to the Parliament of the United Kingdom
Prediction: the areas in which it can make laws

To what body are certain powers explicitly specified as being reserved for?
Ground Truth Answers: Parliament of the United Kingdom | Parliament of the United Kingdom | The British Parliament
Prediction: Parliament of the United Kingdom

SQUAD 1.1 and SQUAD 2.0: F1 SCORE

Source: CodaLab Worksheets, 2020 | Chart: 2021 AI Index Report



Protein folding

Every protein is made up of a sequence of amino acids bonded together

These amino acids interact locally to form shapes like helices and sheets

These shapes fold up on larger scales to form the full three-dimensional protein structure

Proteins can interact with other proteins, performing functions such as signalling and transcribing DNA

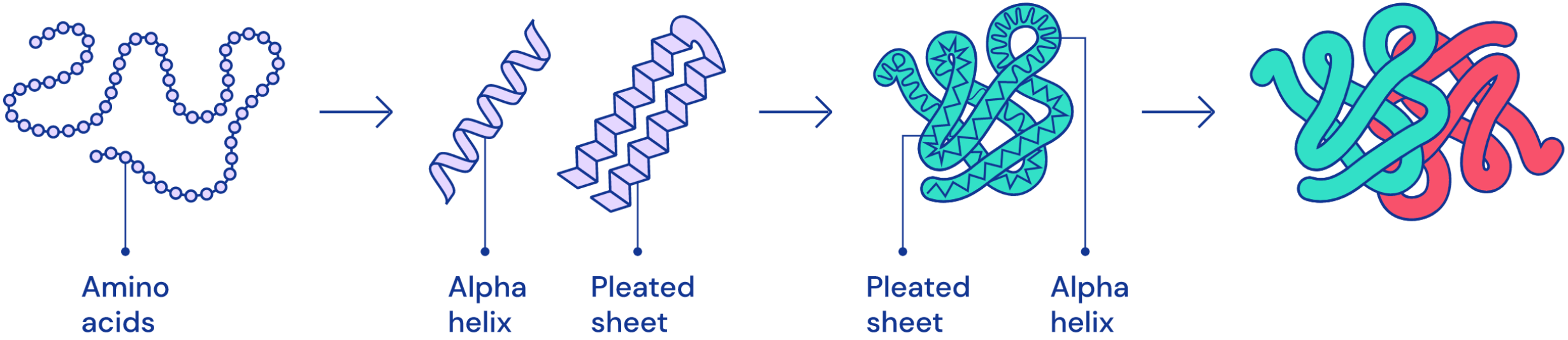
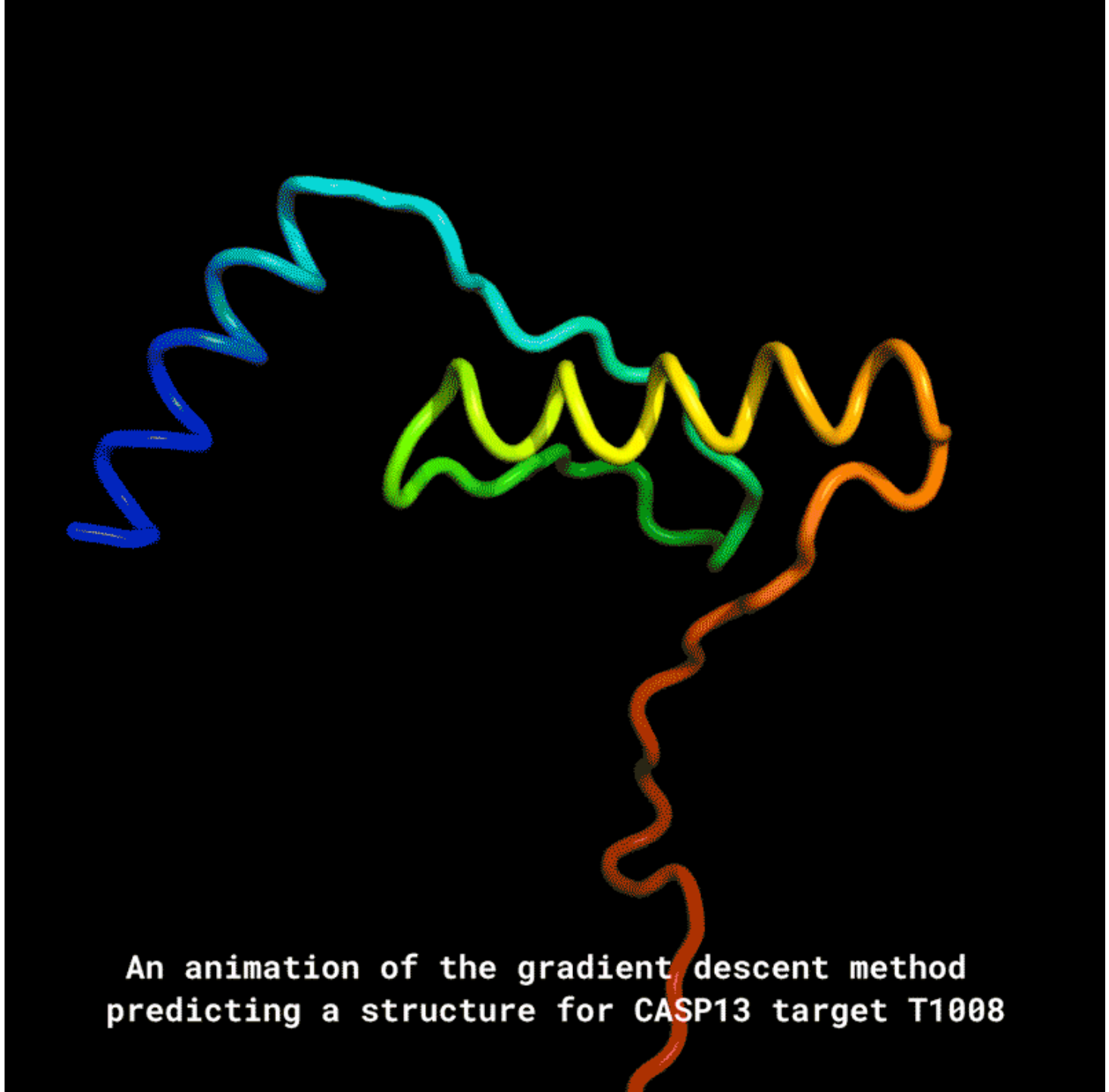
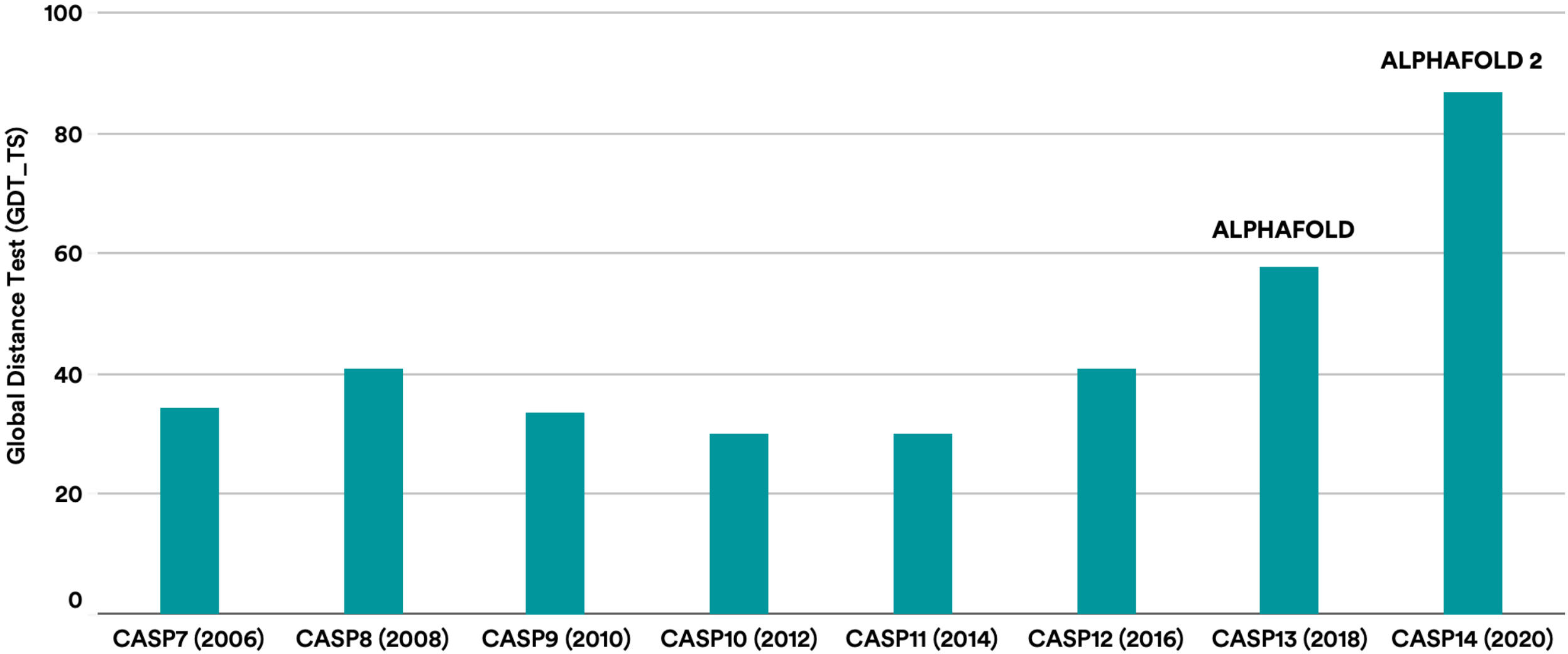


Figure 1: Complex 3D shapes emerge from a string of amino acids.


An animation of the gradient descent method
predicting a structure for CASP13 target T1008

CASP: MEDIAN ACCURACY of PREDICTIONS in FREE-MODELING by THE BEST TEAM, 2006-20

Source: DeepMind, 2020 | Chart: 2021 AI Index Report



Not great for the environment?

Stubel et al (2019)

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

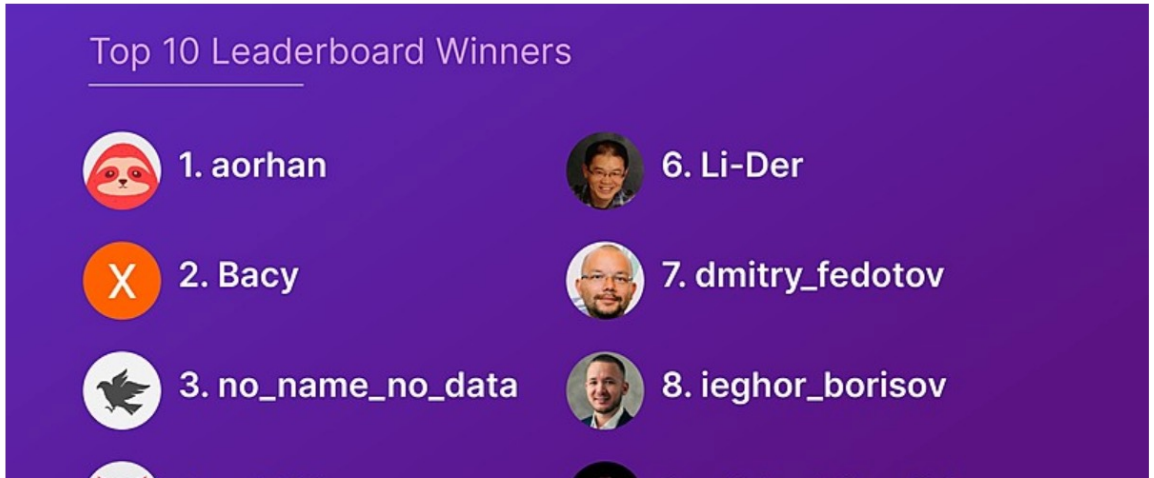
ADDI ALZHEIMER'S DETECTION CHALLENGE

$50,000 CASH PRIZES | 3 PS 5 | 1 XBOX SERIES X | 5 DJI MAVIC MINI 2 | 5 OCULUS QUEST 2

By ADDI | 43.1k | 1390 | 103 | 7071 | 81 | Follow

Overview Leaderboard Notebooks Discussion Insights Resources Submissio

Congratulations to all the winners! 🎉



https://www.aicrowd.com/challenges/addi-alzheimers-detection-challenge

Notebooks Discussion Insights Resources Submissions Winners Rules

Prizes will be awarded for best scores and Contest community contributions. There four (4) cash prizes and 14 non-cash prizes:

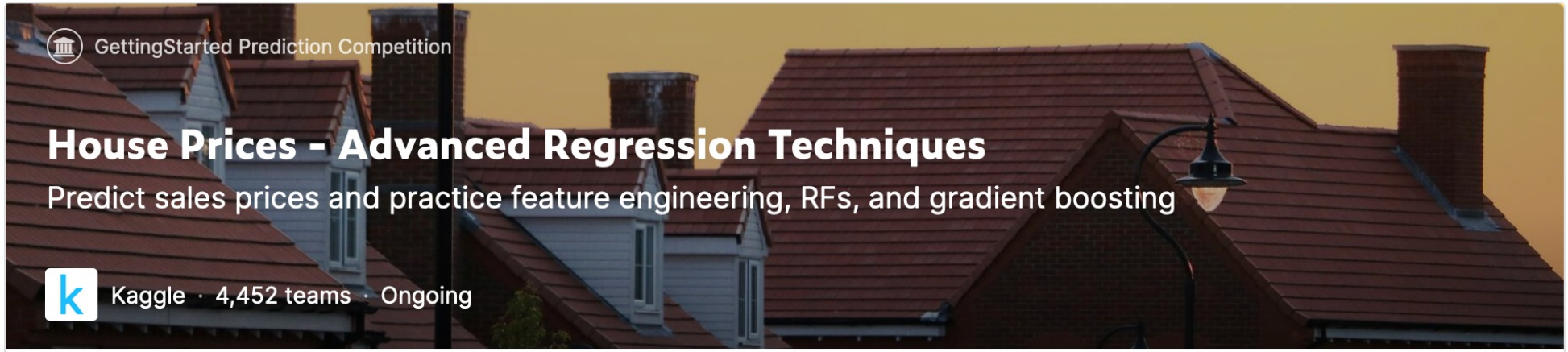
Score-based Prizes:

- Rank #1 $20,000 USD
- Rank #2 $15,000 USD
- Rank #3 $10,000 USD
- Rank #4 $5,000 USD
- Rank #5 1 x Sony PlayStation 5
- Rank #6 1 x Sony PlayStation 5
- Rank #7 DJI Mavic Mini 2
- Rank #8 DJI Mavic Mini 2
- Rank #9 Oculus Quest 2
- Rank #10 Oculus Quest 2

Contest Community Contribution Prizes (8 total):

- 1 x Sony PlayStation 5
- 1 x X-Box Series X
- 3 x DJI Manic Mini 2

Let’s take a look at an existing challenge we can actually do



Data Description

File descriptions

- **train.csv** - the training set
- **test.csv** - the test set
- **data_description.txt** - full description of each column, originally prepared by Dean De Cock but lightly edited to match the column names used here
- **sample_submission.csv** - a benchmark submission from a linear regression on year and month of sale, lot square footage, and number of bedrooms

Public Leaderboard Private Leaderboard

This leaderboard is calculated with all of the test data.

Raw Data Refresh

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Xavier Casanoves García			0.00000	7	2mo
2	fedesoriano			0.00000	2	14d
3	Javad Khiabani			0.00044	7	11d
4	Shivam Chhetry			0.00044	4	1mo
5	Doug LaMaster			0.00044	1	2mo