Data Wrangling and Data Analysis
Machine learning!
#2

Daniel Oberski

Department of Methodology & Statistics

Utrecht University

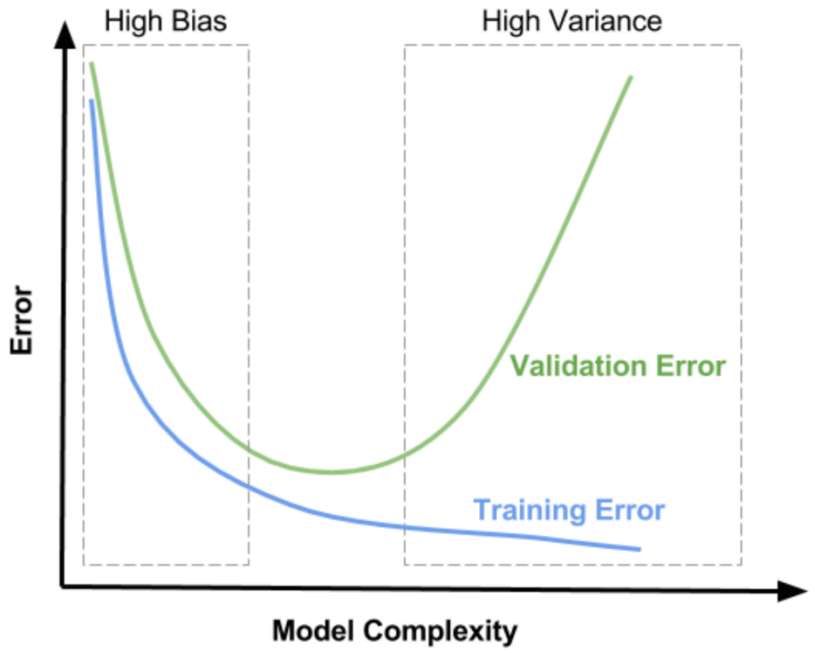
VERSION: 2025-10-06

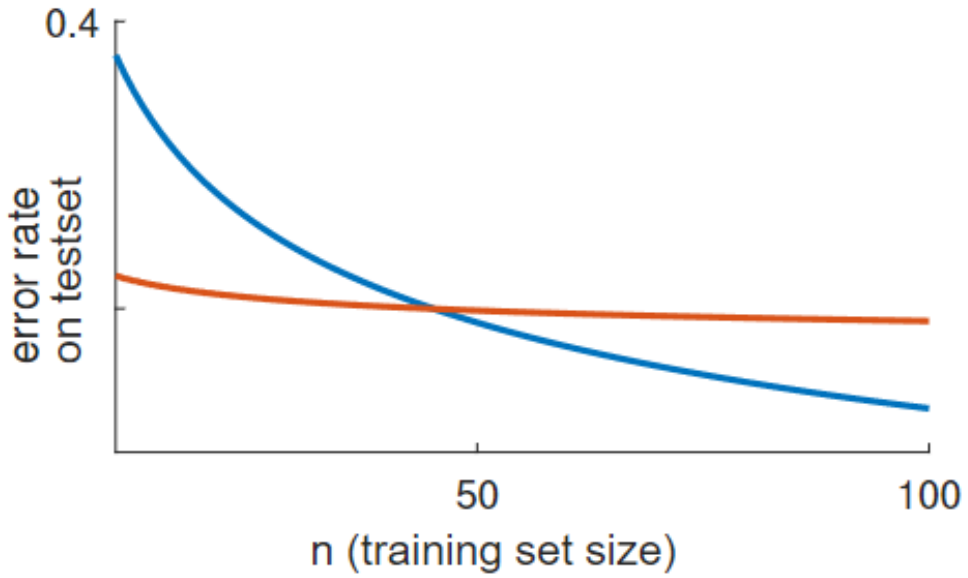
Supervised machine learning

1. Regression (predicting continuous outcomes)
2. **Regression: evaluation**
3. Classification (predicting discrete outcomes)

Recap: **regression**

$$y = f(\boldsymbol{x}) + \epsilon$$





Linear regression

Deep learning

Recap: regression

- Even with true model $f(\boldsymbol{x})$, there is irreducible error
- *Regression* tries to find a "good" $\hat{f}(\boldsymbol{x})$
- Bias-variance tradeoff (capacity vs. stability)
- With higher-capacity model, need more data to learn stable patterns
- Cannot use training performance to evaluate

Will my model succeed?

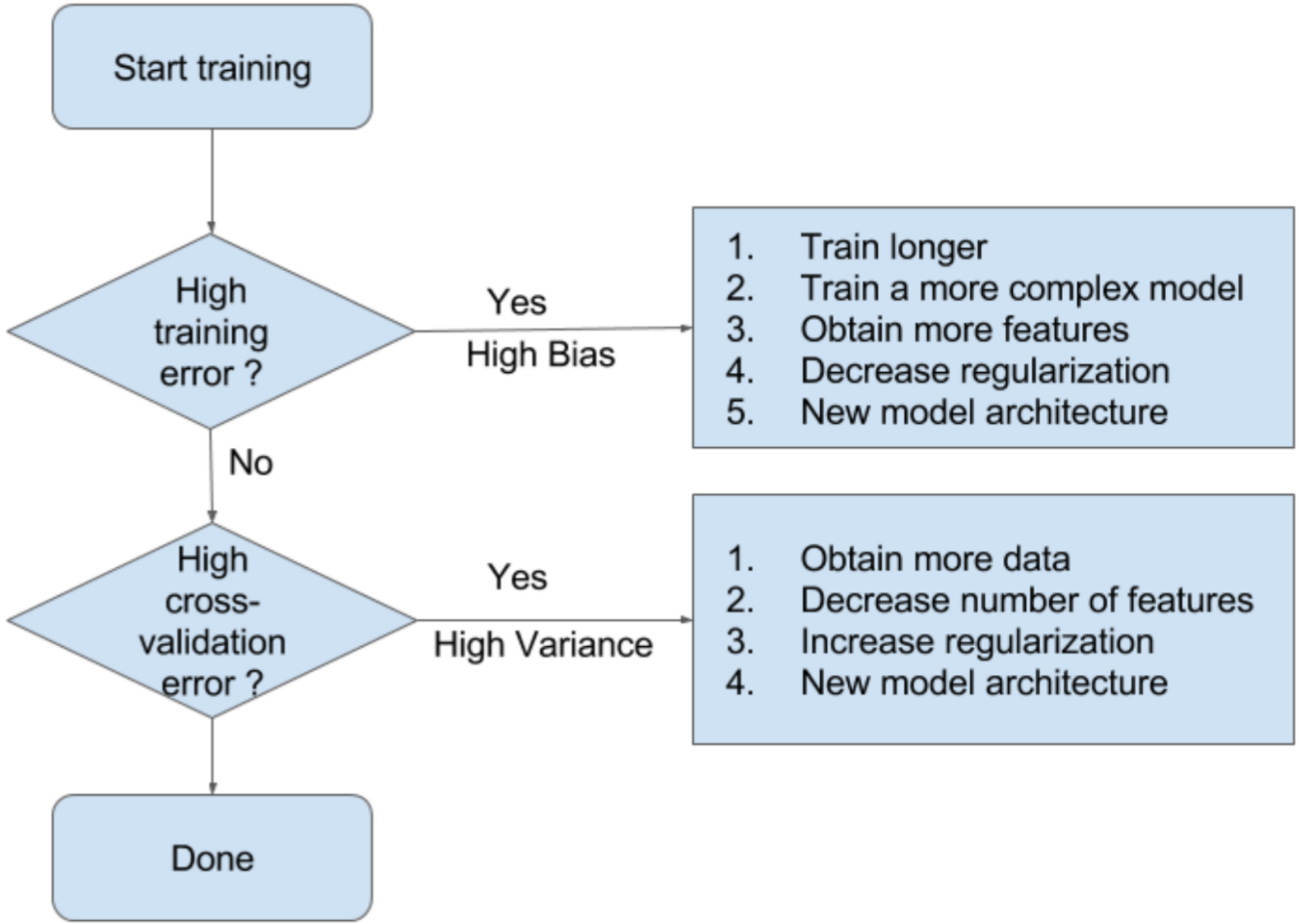
These factors should determine your success:

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 capacity	Try models with range of capacities; 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	Illegally download all digital books and pay fine later.

Bias-Variance Flowchart (Andrew Ng, Coursera)



LLM #parameters and dataset size

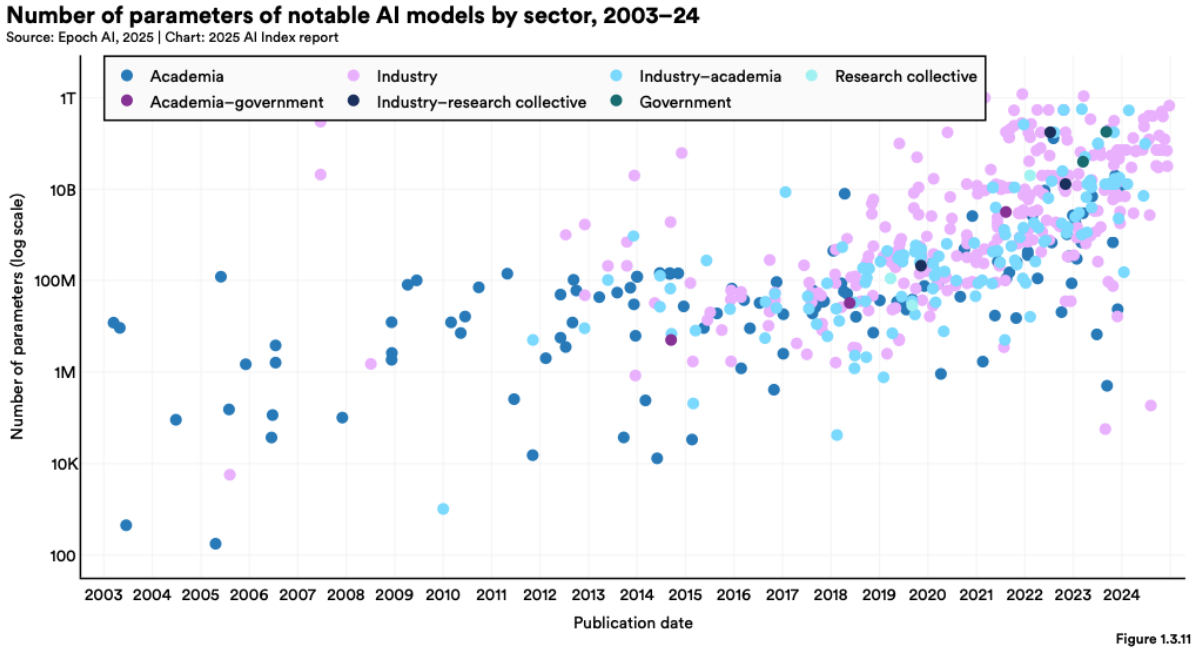
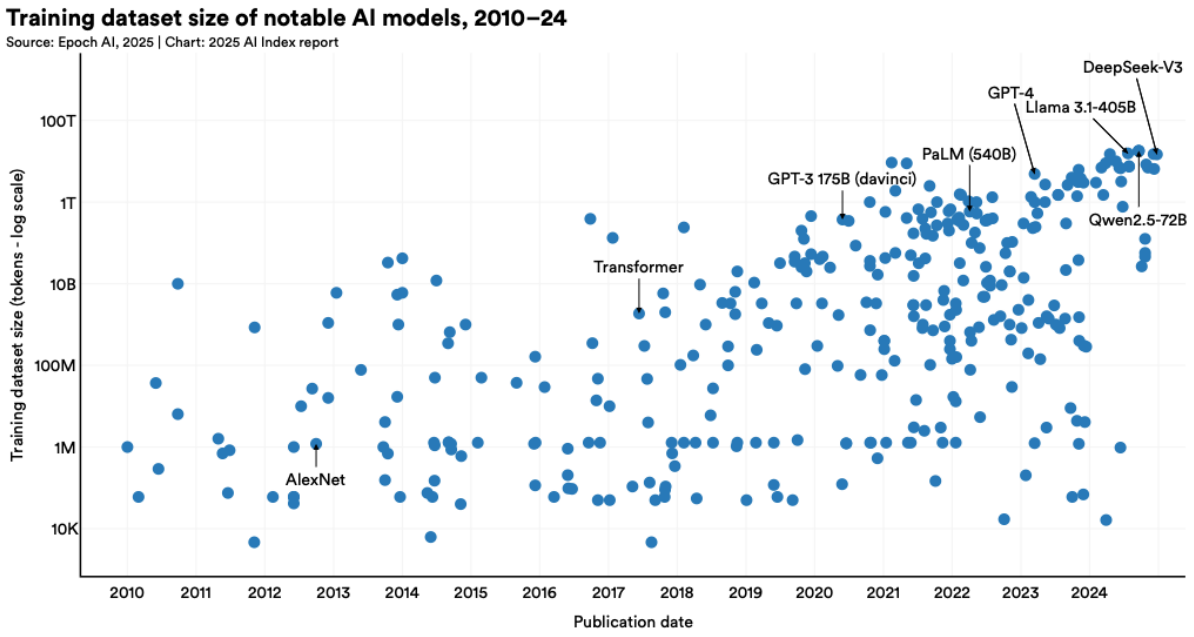


Figure 1.3.11



Model evaluation

Reminder: **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?

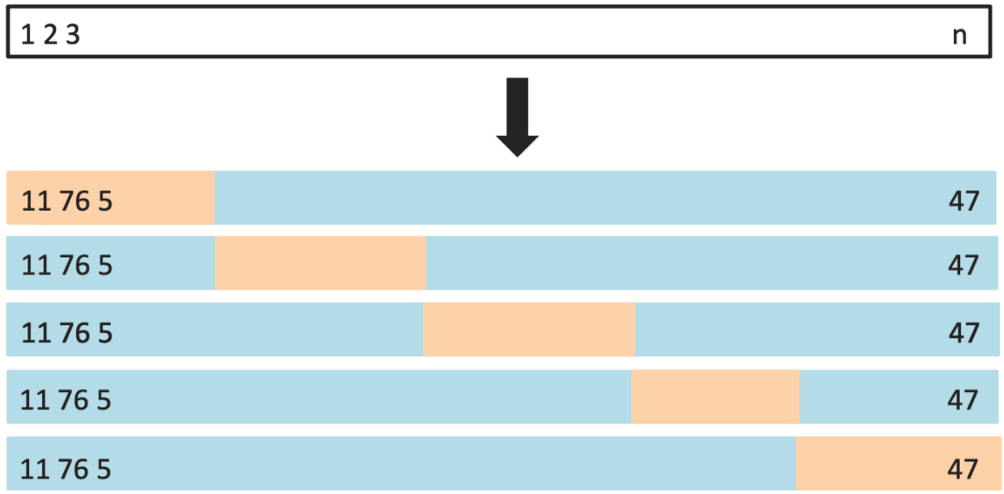
Drawbacks of train/dev/test

- The validation estimate of the test error is based on a single sample and can therefore be **highly variable**
- Only a subset of observations are used to train the model (model could have learned more)
- This suggests that the validation set error may tend to **overestimate the test error** for the model fit on the entire data set.

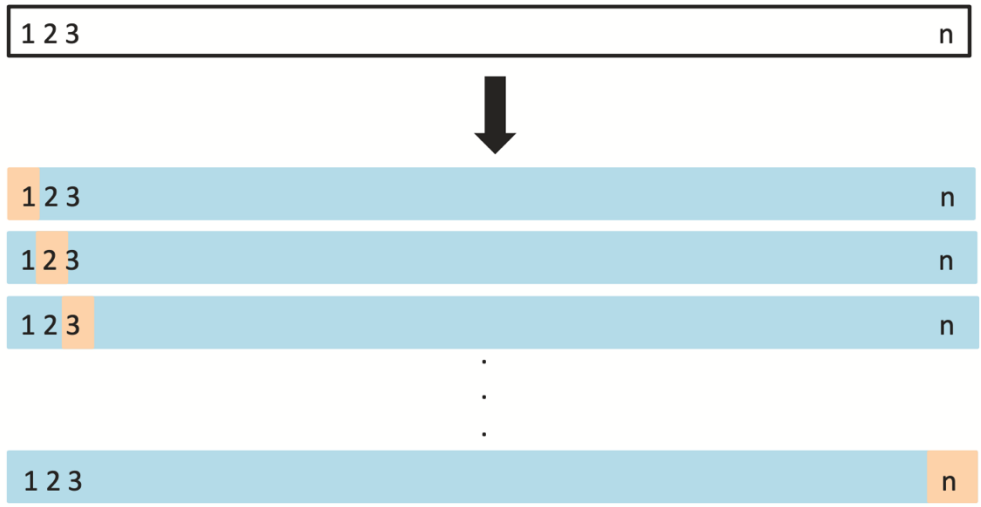
K-fold crossvalidation

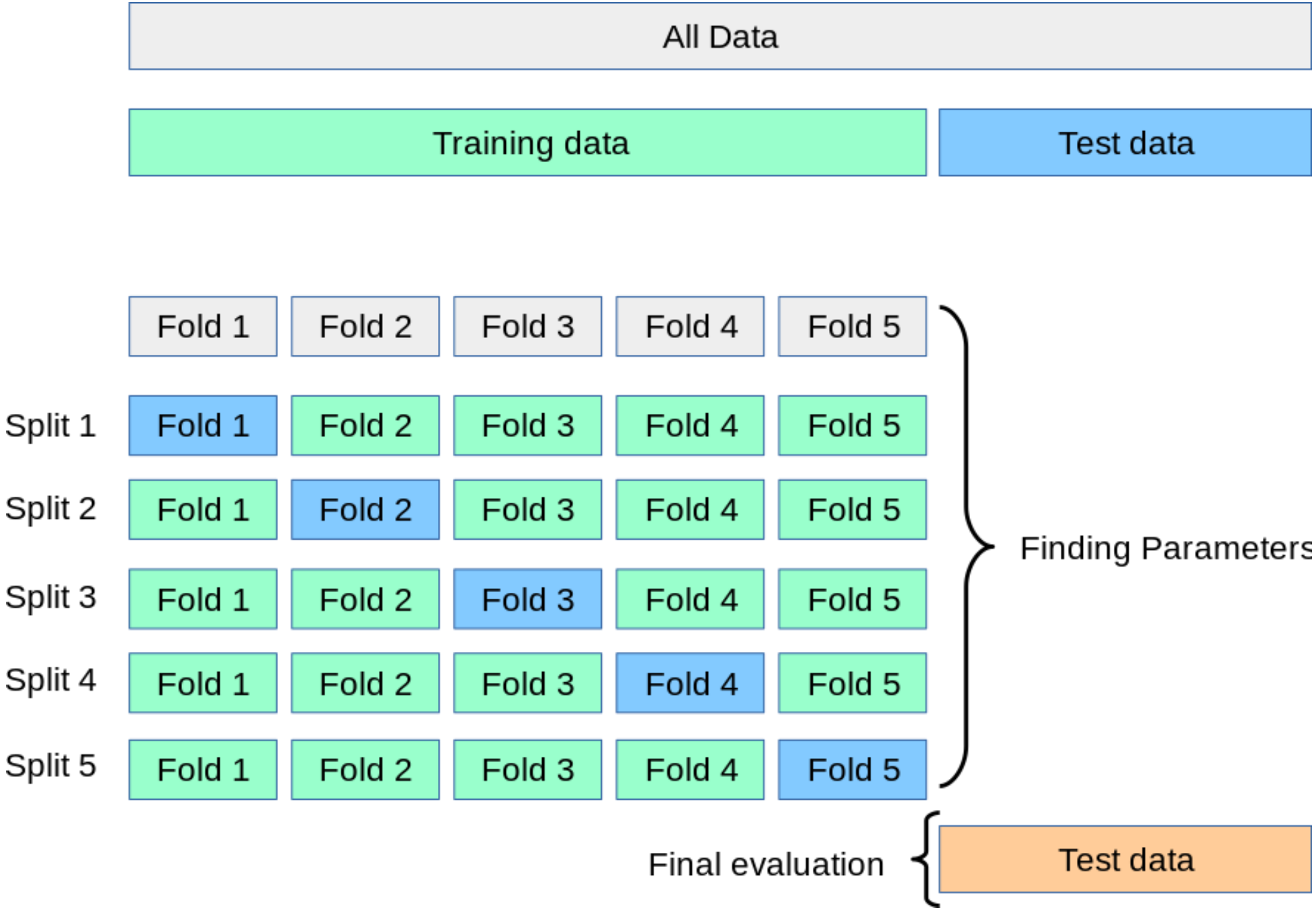
- “Cross-validation” replaces single dev set approach;
- Perform the train/dev split several times, average the result.
- Usually K = 5 or K = 10

Here, K = 5 as is more usual



When K = n, “leave-one-out”




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!

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.

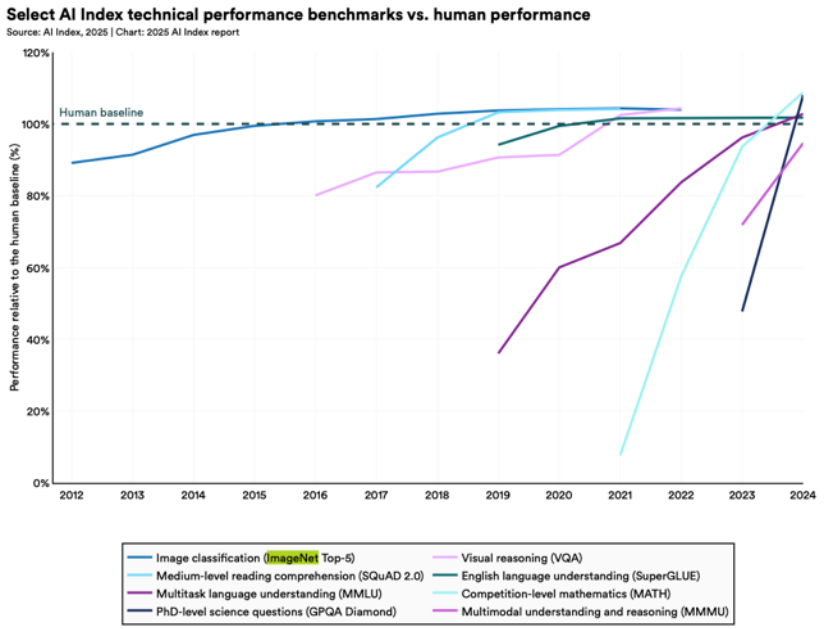
(c) A **scoring referee**, to which competitors can submit their prediction rule.

- 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.

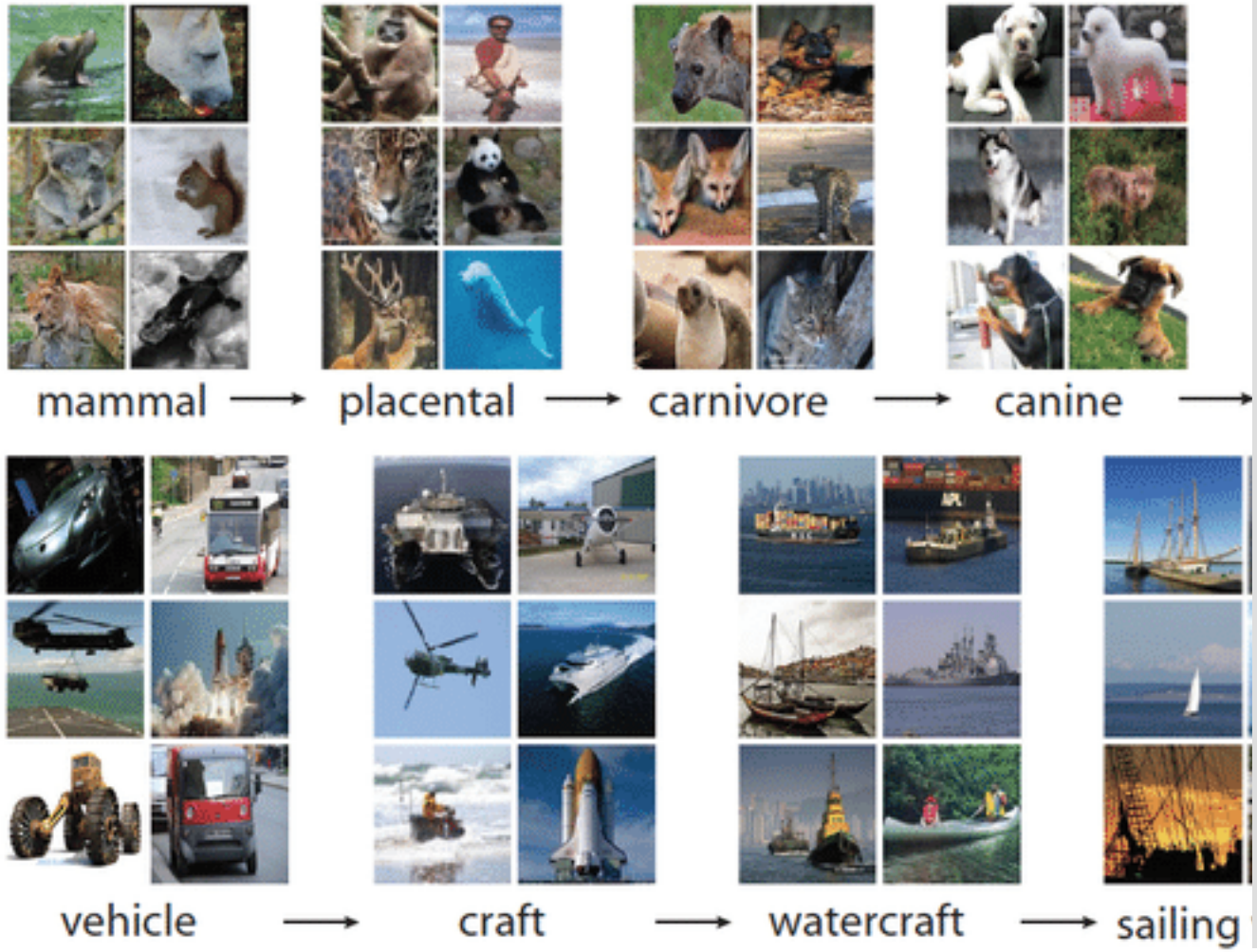
Donoho (2017). *50 years of data science.* J Comp Graphical Stat.

CTF/benchmarking advantages

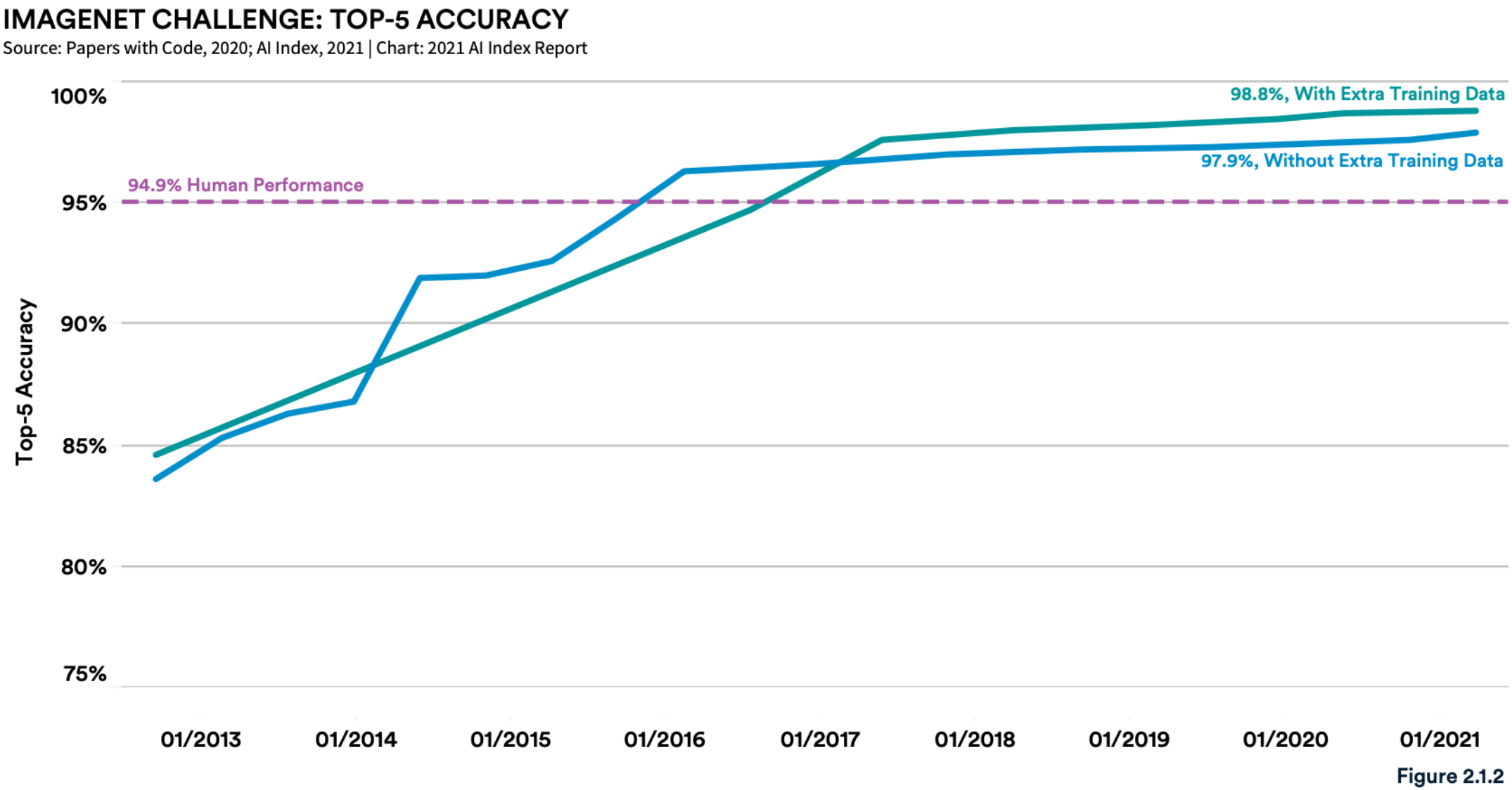
1. Error rates decline by a fixed percentage each year, to an asymptote depending on task and data quality.
2. Progress usually comes from many small improvements; a change of 1% can be a reason to break out the champagne.
3. Shared data plays a crucial role.



Imagenet



Future Nobel (?) Fei-Fei Li



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

A sample question from MMLU

Source: Hendrycks et al., 2021

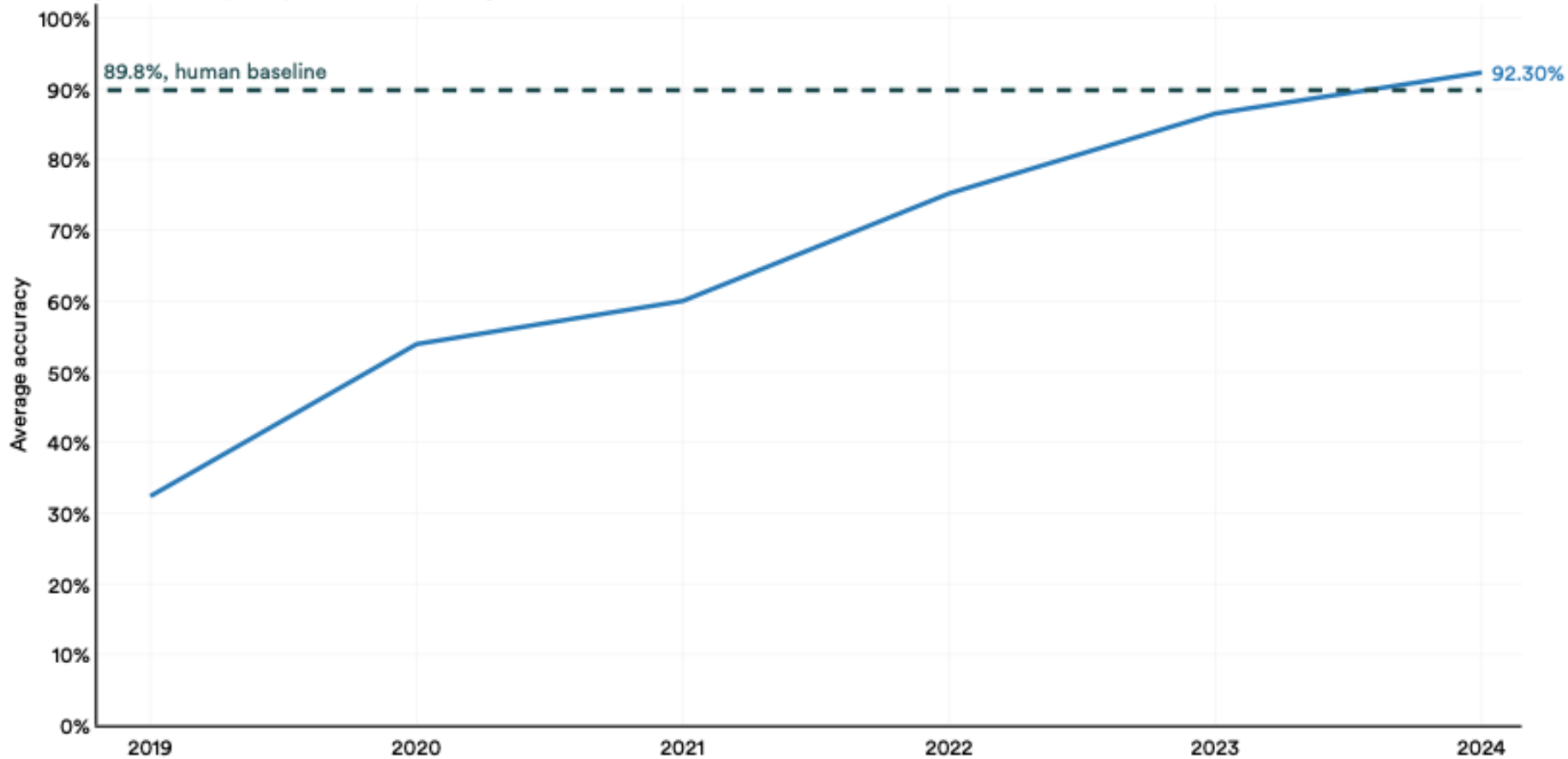
Microeconomics

One of the reasons that the government discourages and regulates monopolies is that
(A) producer surplus is lost and consumer surplus is gained. ✗
(B) monopoly prices ensure productive efficiency but cost society allocative efficiency. ✗
(C) monopoly firms do not engage in significant research and development. ✗
(D) consumer surplus is lost with higher prices and lower levels of output. ✓

Figure 2.2.3

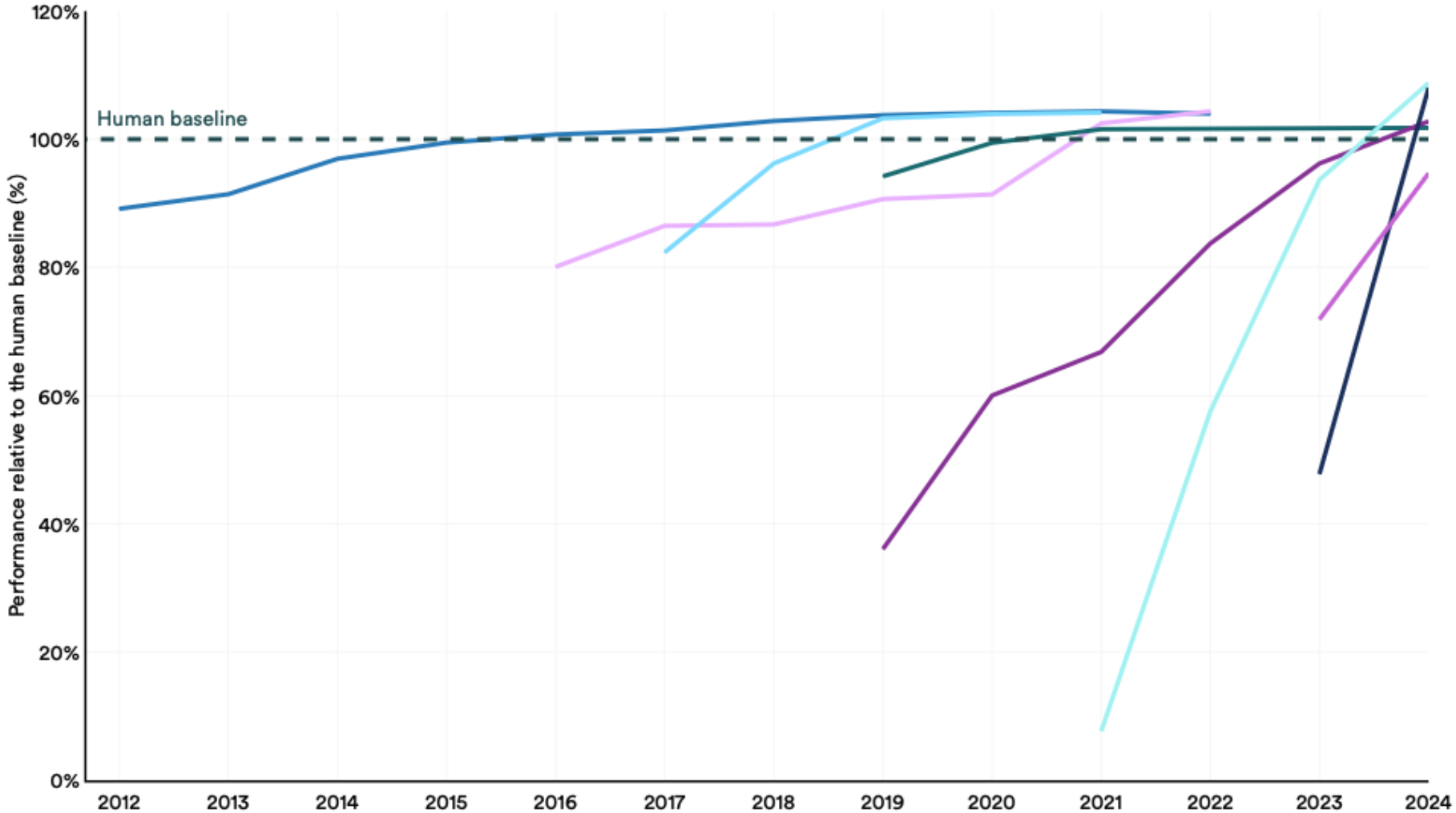
MMLU: average accuracy

Source: Papers With Code, 2025 | Chart: 2025 AI Index report



Select AI Index technical performance benchmarks vs. human performance

Source: AI Index, 2025 | Chart: 2025 AI Index report



“Humanity’s last exam”

Source: Phan et al., 2025

Classics

Question:

Here is a representation of a Roman inscription, orginally found on a tombstone. Provide a translation for the Palmyrene script.
A transliteration of the text is provided: RGYNˁ BT HRY BR ˁTˀ HBL

Henry T
Merton College, Oxford

Ecology

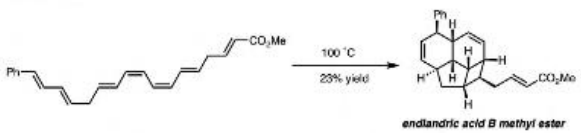
Question:

Hummingbirds within Apodiformes uniquely have a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of m. depressor caudae. How many paired tendons are supported by this sesamoid bone? Answer with a number.

Edward V
Massachusetts Institute of Technology

Chemistry

Question:



endiandric acid B methyl ester

The reaction shown is a thermal pericyclic cascade that converts the starting heptaene into endiandric acid B methyl ester. The cascade involves three steps: two electrocyclizations followed by a cycloaddition. What types of electrocyclizations are involved in step 1 and step 2, and what type of cycloaddition is involved in step 3?

Provide your answer for the electrocyclizations in the form of [nπ]-con or [nπ]-dis (where n is the number of π electrons involved, and whether it is conrotatory or disrotatory), and your answer for the cycloaddition in the form of [m+n] (where m and n are the number of atoms on each component).

Noah B
Stanford University

Linguistics

Question:

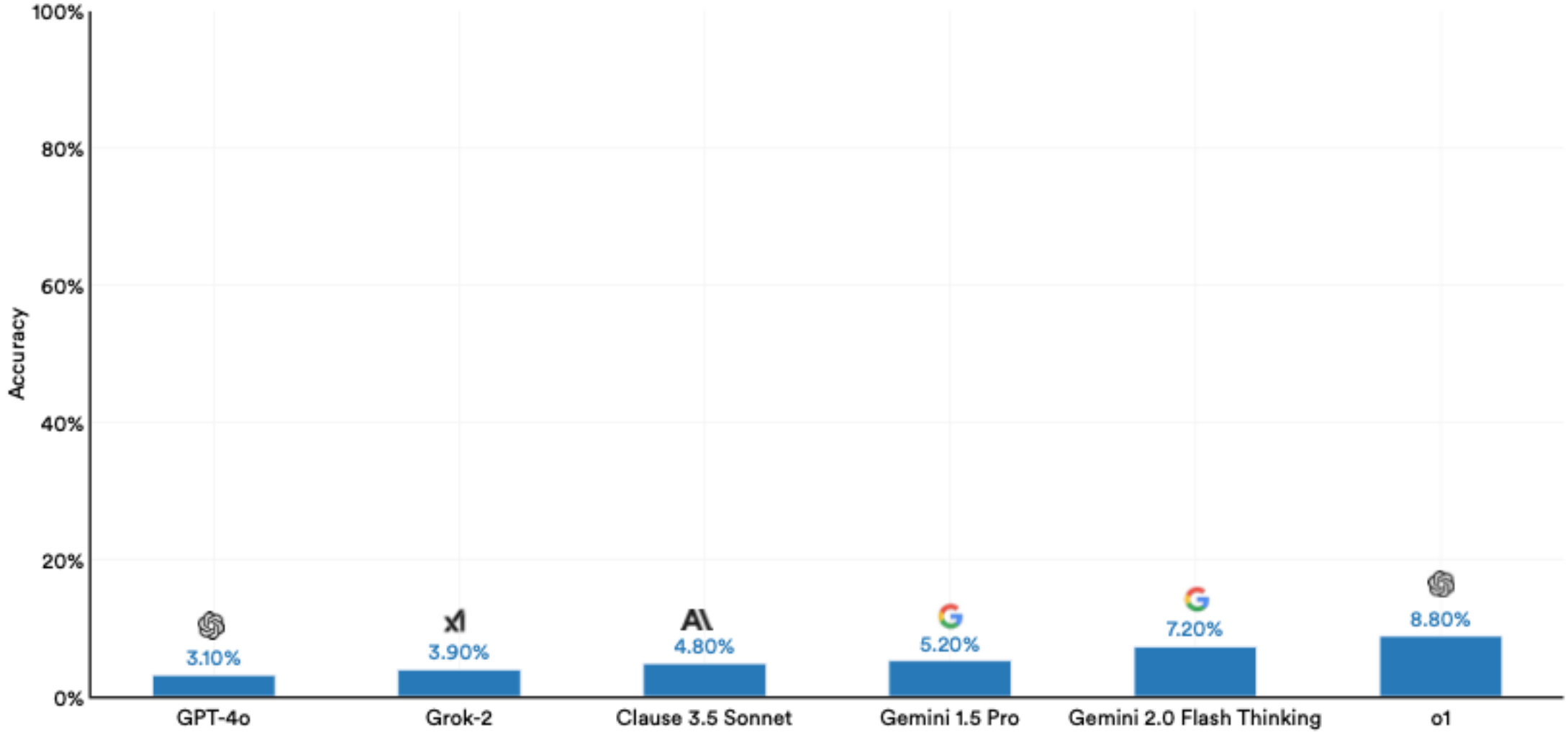
I am providing the standardized Biblical Hebrew source text from the Biblia Hebraica Stuttgartensia (Psalms 104:7). Your task is to distinguish between closed and open syllables. Please identify and list all closed syllables (ending in a consonant sound) based on the latest research on the Tiberian pronunciation tradition of Biblical Hebrew by scholars such as Geoffrey Khan, Aaron D. Hornkohl, Kim Phillips, and Benjamin Suchard. Medieval sources, such as the Karaite transcription manuscripts, have enabled modern researchers to better understand specific aspects of Biblical Hebrew pronunciation in the Tiberian tradition, including the qualities and functions of the shewa and which letters were pronounced as consonants at the ends of syllables.

מִן־גַּעֲרָתְךָ֣ יְנוּסָ֑ון מִן־ק֥וֹל רַ֝עַמְךָ֗ יֵחָפֵזֽוּון (Psalms 104:7) ?

Lina B
University of Cambridge

Humanity's Last Exam (HLE): accuracy

Source: Phan et al., 2025 | Chart: 2025 AI Index report



Congratulations to all the winners! 🎉

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

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

Score-based Prizes:

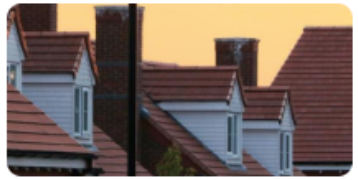
- Rank #1 $20,000 USD
- Rank #2 $15,000 USD
- Rank #3 $10,000 USD
- Rank #4 $5,000 USD
- Rank #5 1 x Sony PlayStation 5
- Rank #6 1 x Sony PlayStation 5
- Rank #7 DJI Mavic Mini 2
- Rank #8 DJI Mavic Mini 2
- Rank #9 Oculus Quest 2
- Rank #10 Oculus Quest 2

Contest Community Contribution Prizes (8 total):

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

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting



Overview

This competition runs indefinitely with a rolling leaderboard. Learn more

Description

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

💡Getting Started Notebook

To get started quickly, feel free to take advantage of this starter notebook.

Competition Description

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

Competition Host
Kaggle

Prizes & Awards
Knowledge
Does not award Points or Medals

Participation
909,547 Entrants
4,419 Participants
4,377 Teams
18,733 Submissions

Tags

Regression | Tabular | Root Mean Squared Logarithmic Error

Table of Contents

Description
Evaluation
Tutorials
Frequently Asked Q...
Citation

Some models

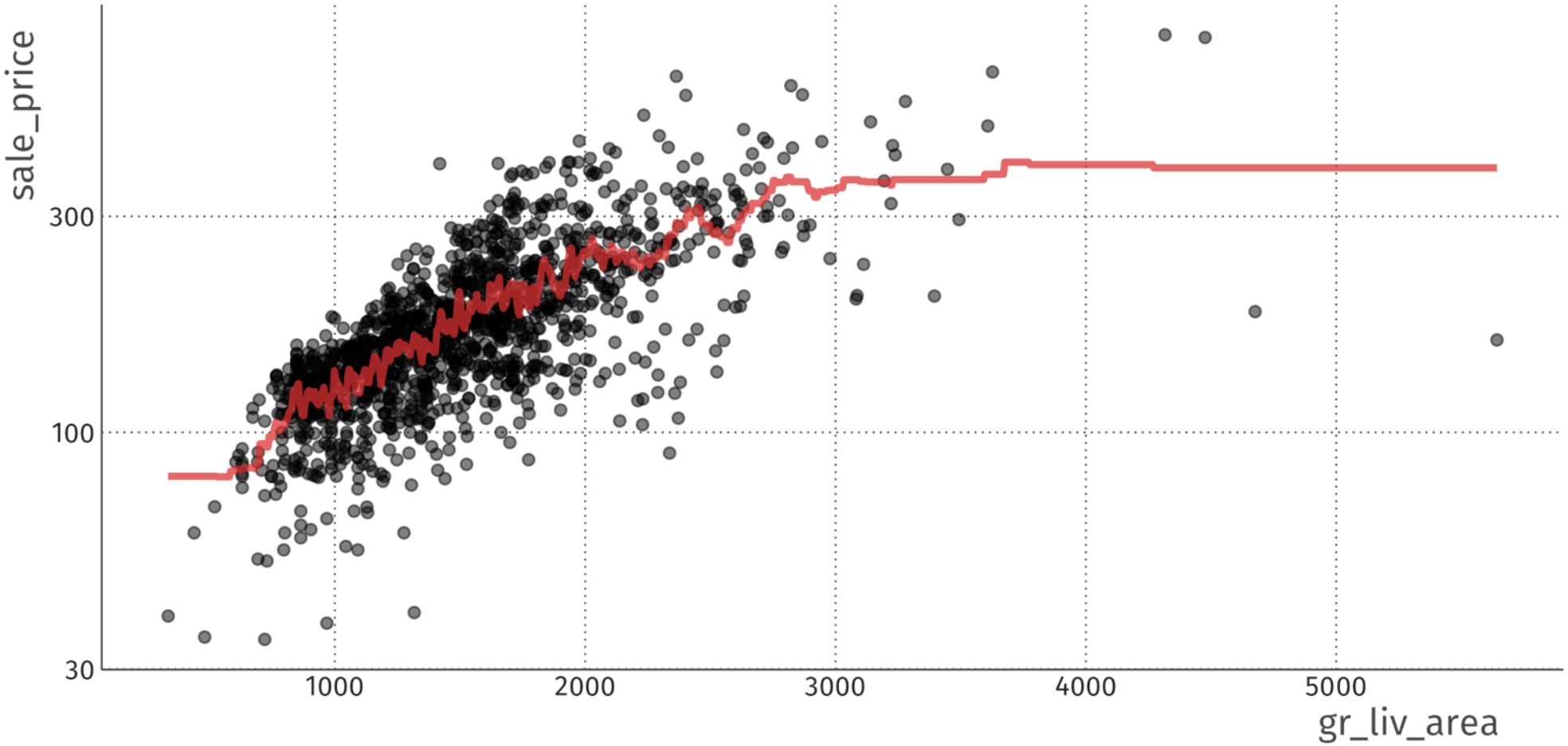
Reminder

- **Goal**: given a new, ***unseen***, vector data point $\mathbf{x}_i$ (with potentially many dimensions), predict a scalar (single value) y_i ;
- To achieve this task, we will learn a function $\hat{f}(\boldsymbol{x})$ (model) using an algorithm (estimator, learner) from labeled training data points ($\mathbf{x}$, y);
- When asked to perform the task on new, unseen, data, we will output the prediction function learned earlier, $\hat{y} = \hat{f}(\boldsymbol{x})$.
- We will now look at some of these functions (models) and how they can be learned (estimated)

Some models

- Linear regression
- kNN
- Trees
- Random forests
- Boosting (e.g. `xgboost`)
- Support vector regression (SVR)
- Neural networks / Deep learning

kNN regression (k = 30)



Regression tree

Prediction for:

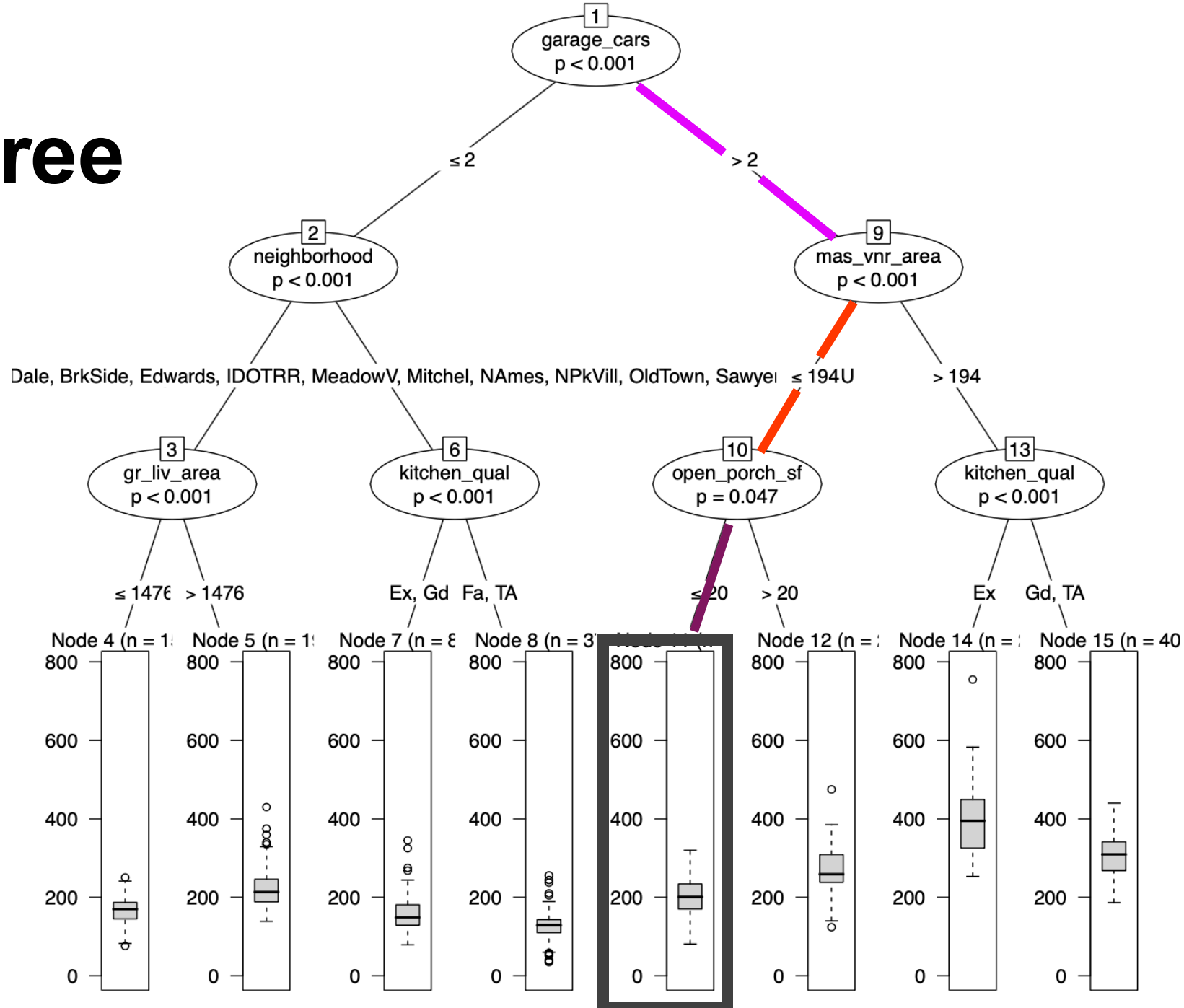
- `garage_cars` = 3
- `mas_vnr_area` = 104
- `open_porch_sf` = 10

$$\hat{y} = 200$$

What about for:

- `garage_cars` = 4
- `mas_vnr_area` = 200
- `kitchen_qual` = Ex
- `open_porch_sf` = 10

$$\hat{y} = ?$$



Regression tree

Prediction for:

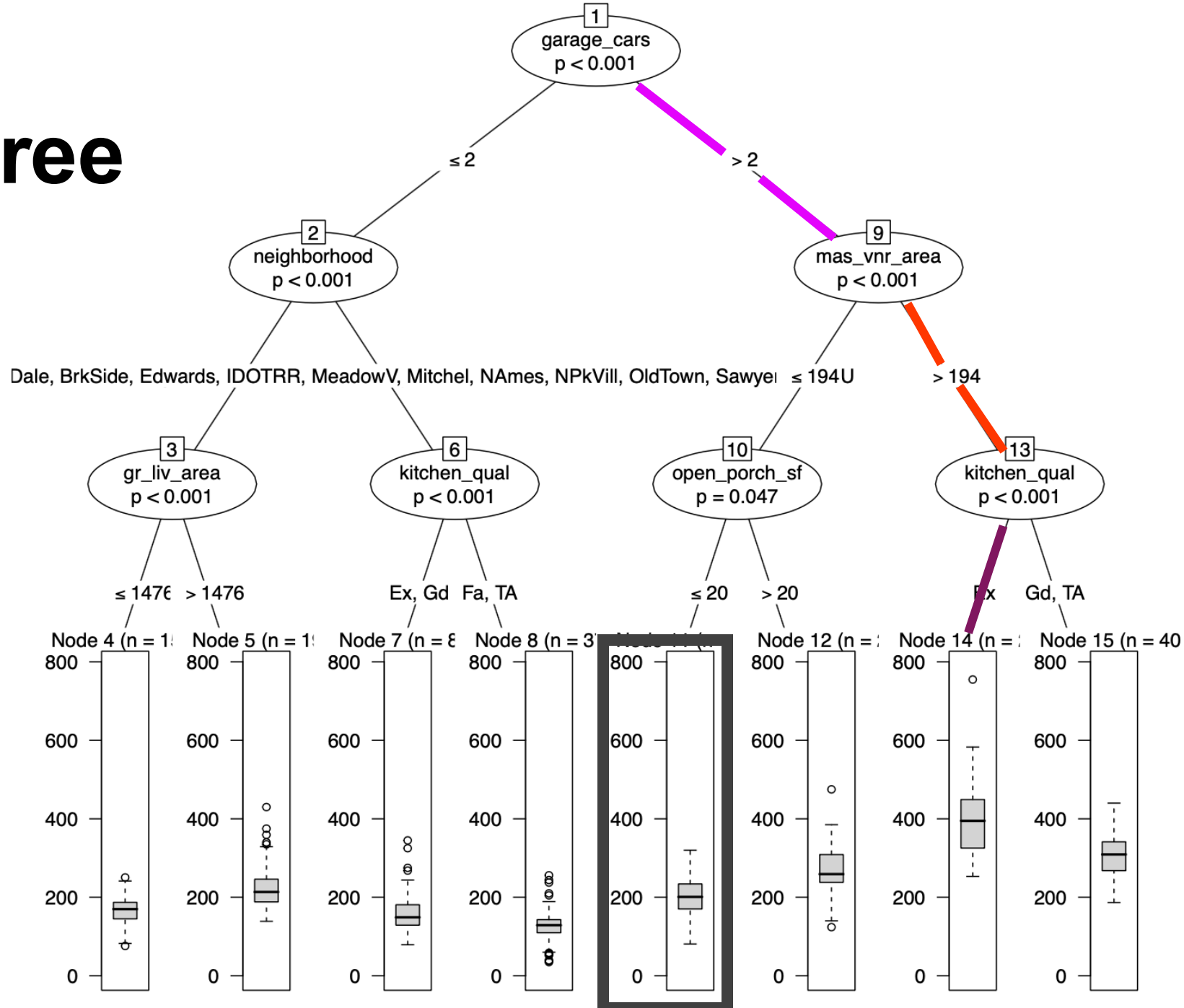
- `garage_cars` = 3
- `mas_vnr_area` = 104
- `open_porch_sf` = 10

$$\hat{y} = 200$$

What about for:

- `garage_cars` = 4
- `mas_vnr_area` = 200
- `kitchen_qual` = Ex
- `open_porch_sf` = 10

$$\hat{y} = 400$$



The problem with trees

- Trees are not a bad idea, but in practice they tend to overfit

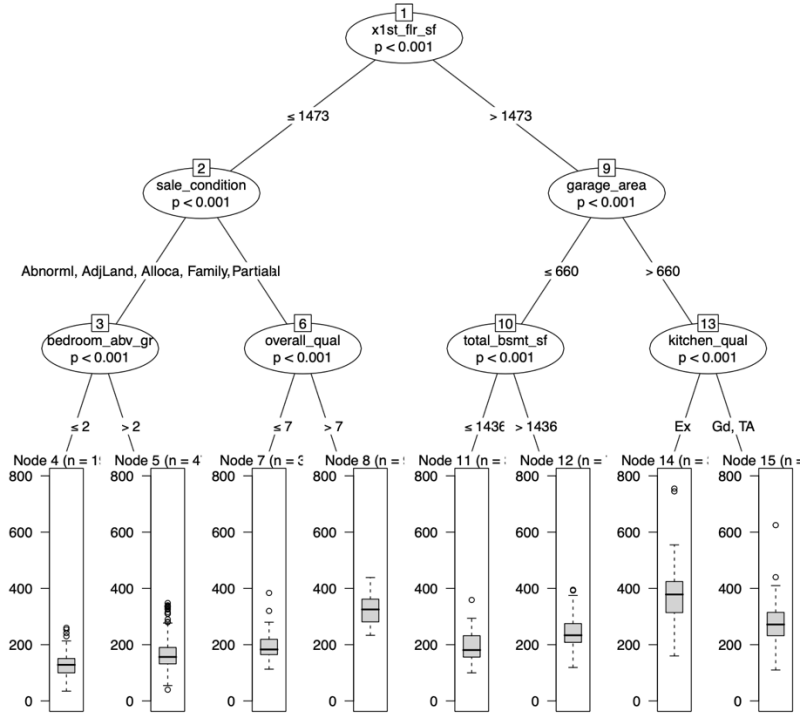
→ Use them as **basic building block** for **ensembles**

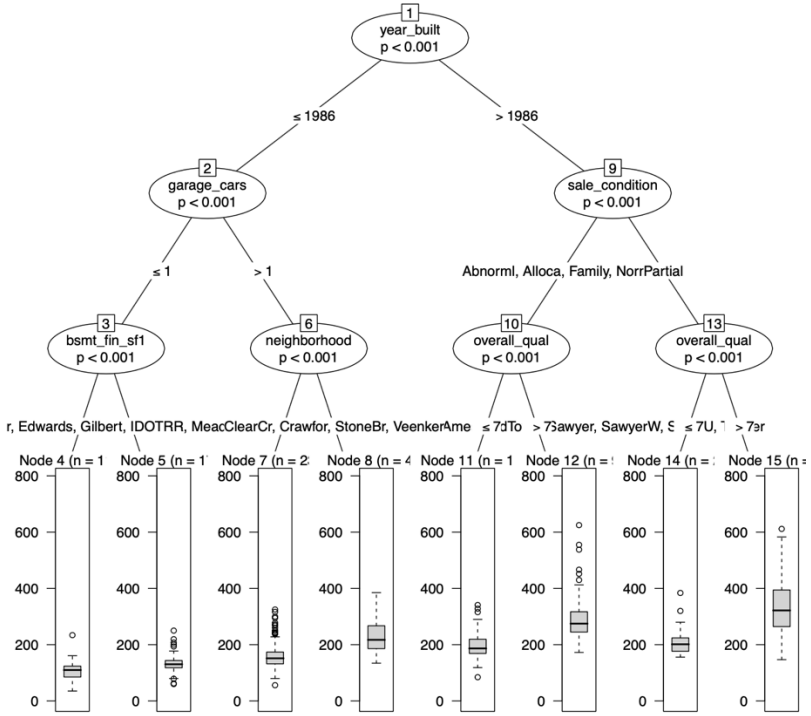
- **Random forests**: “bagged trees with feature sampling”
  - Make trees that are too complex (low bias, high variance);
  - Average over bootstrapped samples to cancel out the overfitting parts.

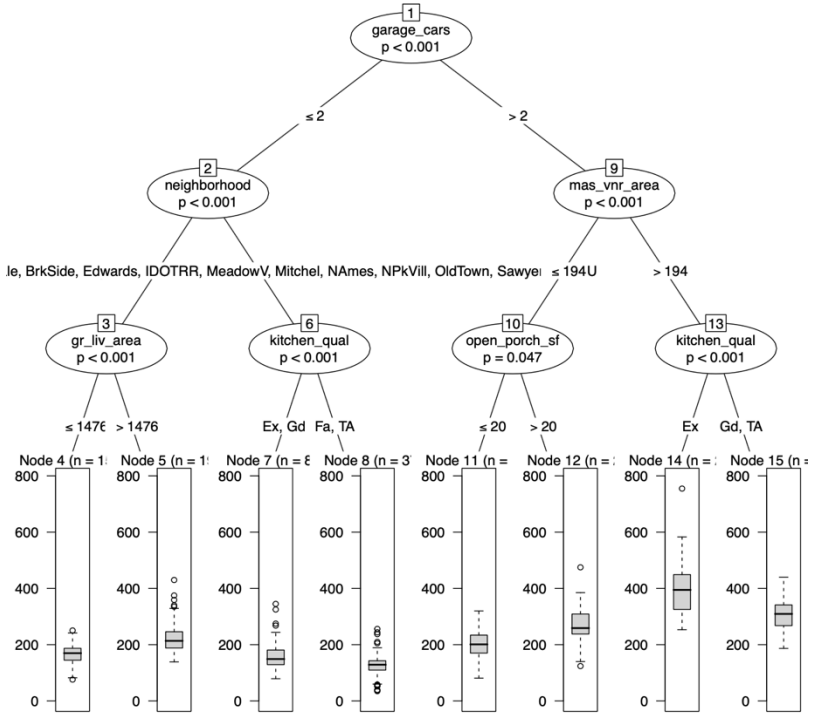
- **Boosting**: “ensemble of weak learners”
  - Make trees that are too simple (high bias, low variance);
  - Make more of them for observations with big residuals;
  - Average them.

More about random forests and boosting next time!

Random forests

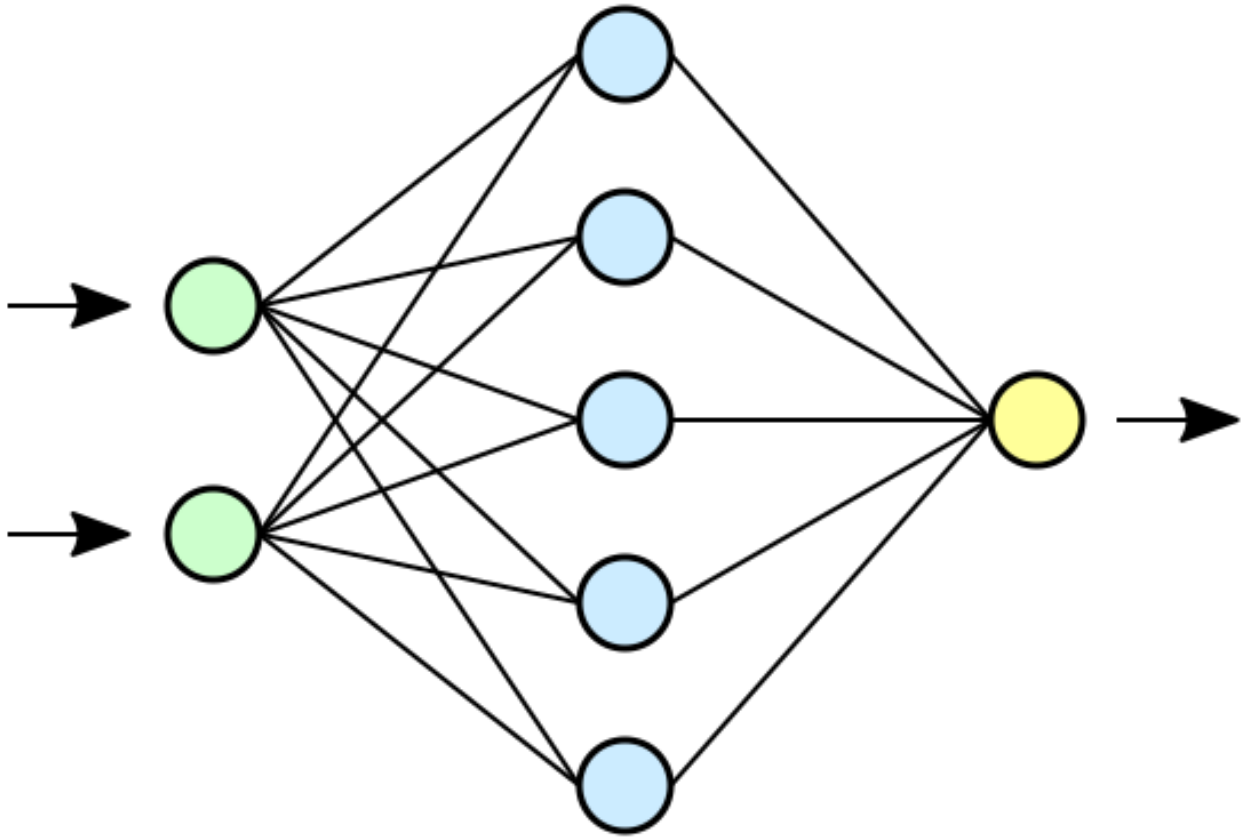




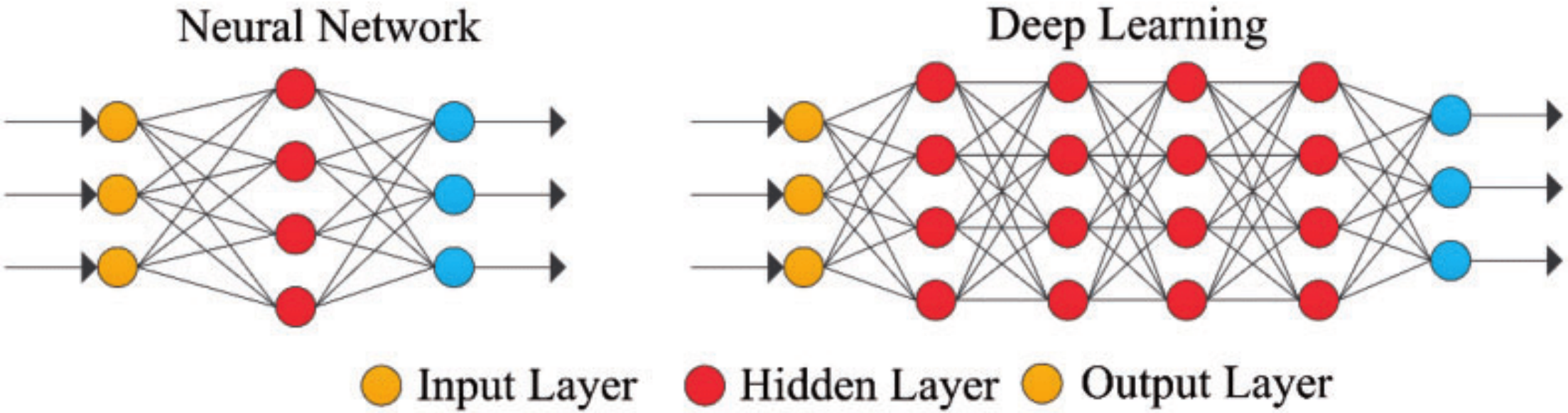


- Fit many trees, each to a different **random sample of the training data**,
- and **randomly sampling** a small set of **features** at each node.
- The final random forest prediction is the **average of the individual tree predictions**.

Neural network



What makes a neural net “deep”?

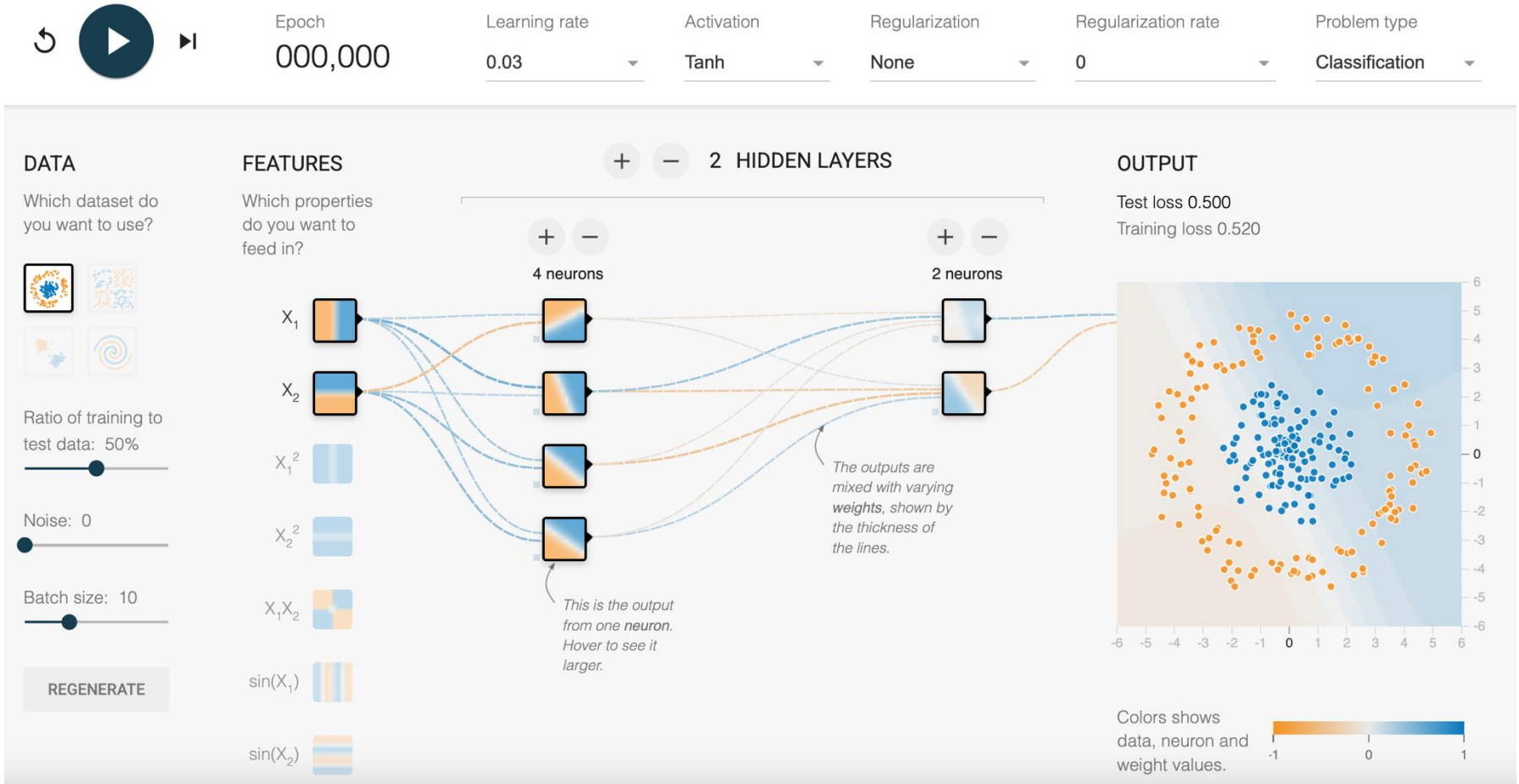


Deep learning

- Output of each hidden layer is input to subsequent one
- Allow representation learning by building complex features out of simpler ones
- Aggressive parameterization + aggressive regularization
  (“top-down” nonparametric approach)
- Compositional: efficient parametrization
- Learn relevant features: “End-to-end”

More about deep learning later!

https://playground.tensorflow.org

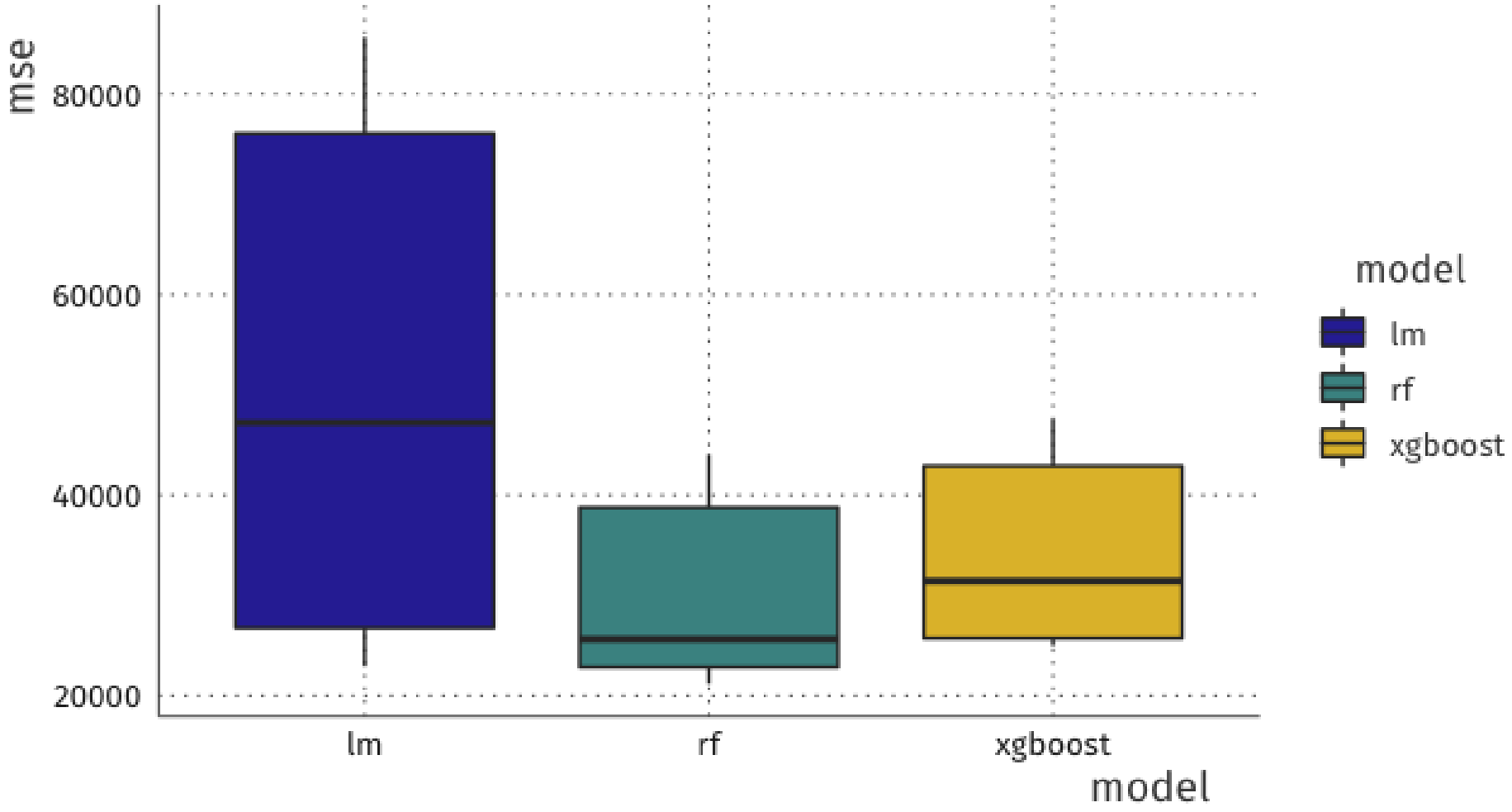


No free lunch

”Any two optimization algorithms are equivalent when their performance is averaged across all possible problems”

(Wolpert & MacReady)

Our housing example



Tuning

- Learners (models) have “hyperparameters”
- Example: number of trees in RF, depth of neural network
- General idea is to use **cross-validation** to select hyperparameters
- Some models more sensitive to good choice of hyperparamters
- Examples are boosting and neural nets

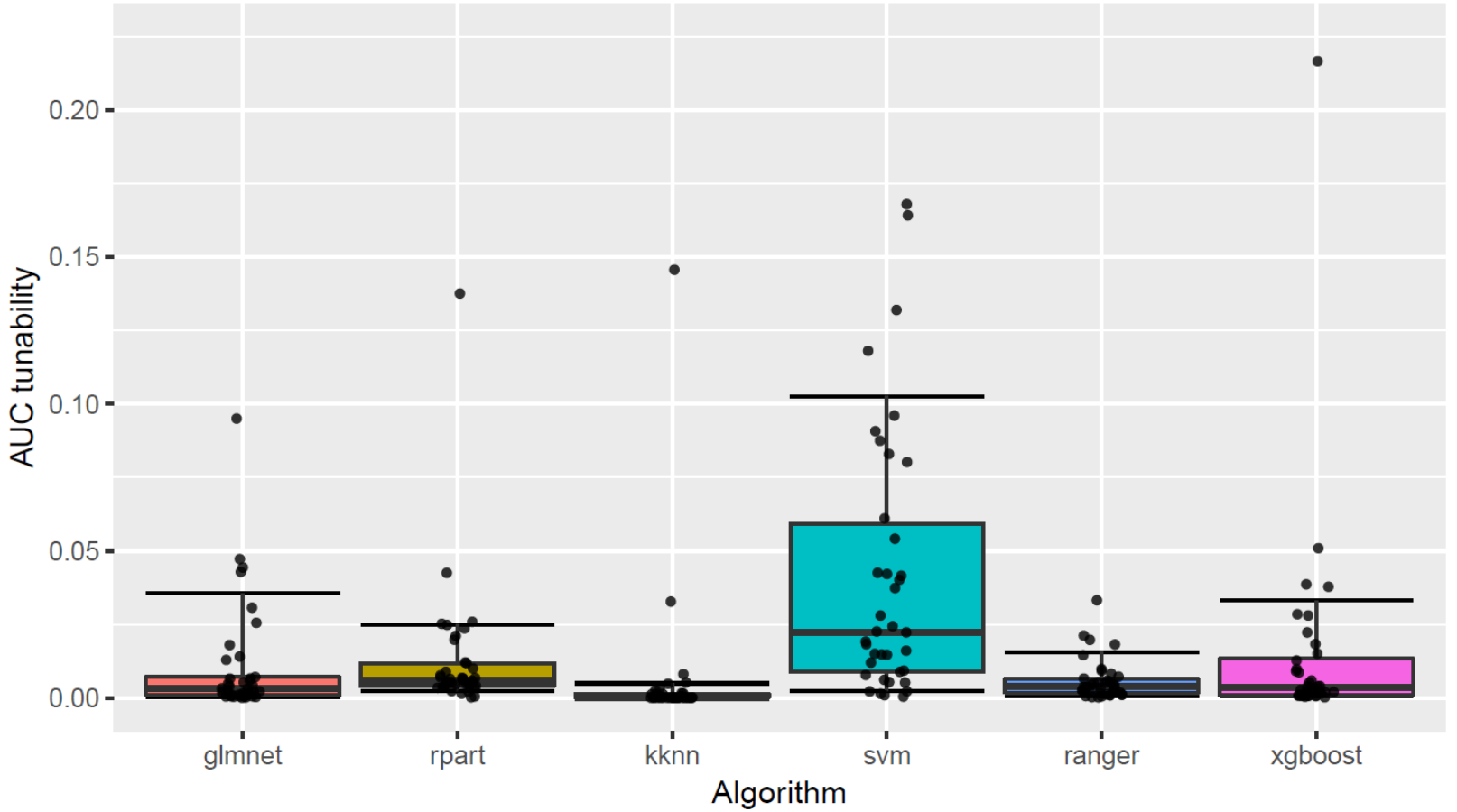
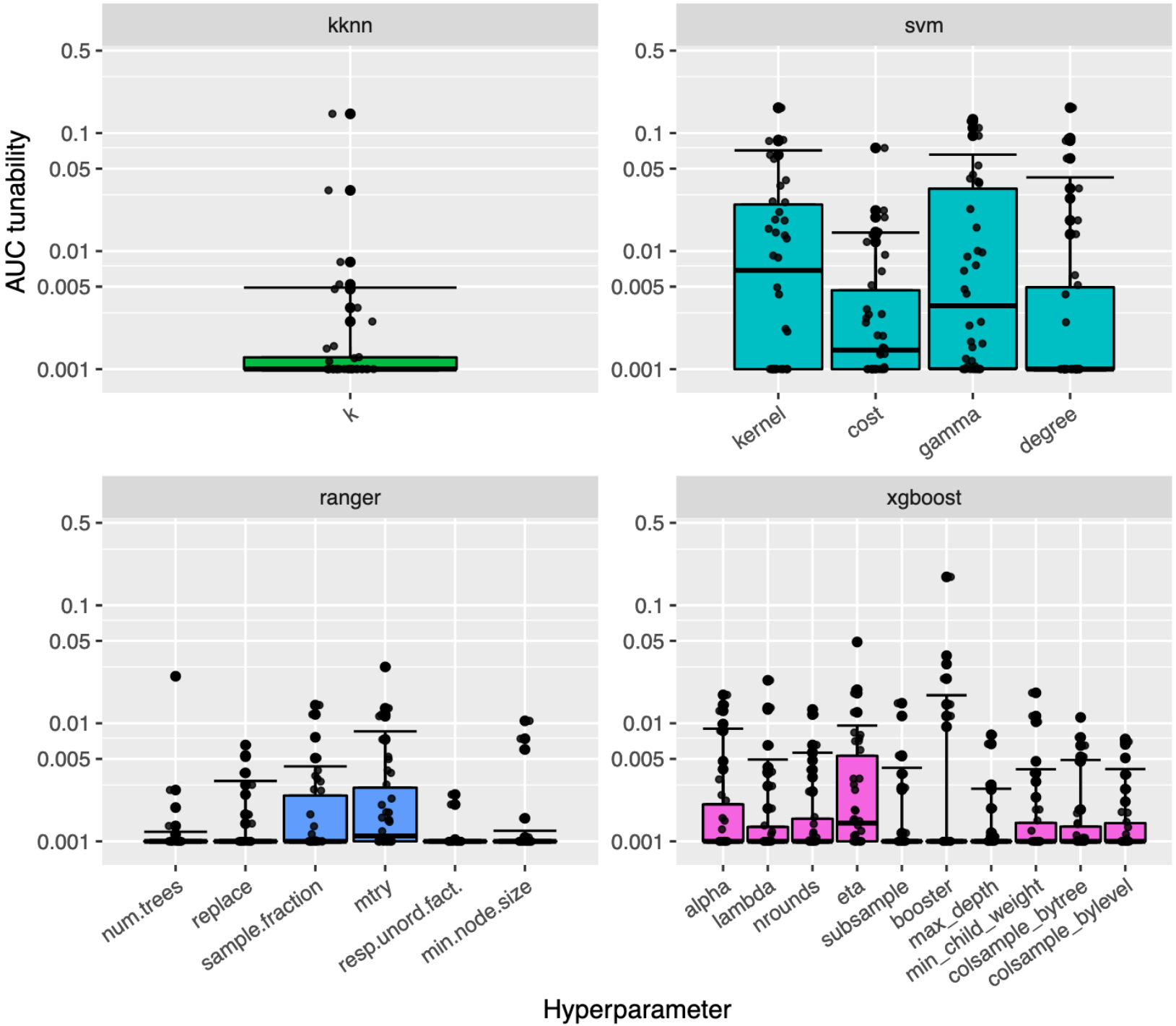


Figure 2: Boxplots of the tunabilities (AUC) of the different algorithms with respect to optimal defaults. The upper and lower whiskers (upper and lower line of the boxplot rectangle) are in our case defined as the 0.1 and 0.9 quantiles of the tunability scores. The 0.9 quantile indicates how much performance improvement can be expected on at least 10% of datasets. One outlier of `glmnet` (value 0.5) is not shown.

Probst et al. JMLR



Probst et al. (2019). JMLR

AutoML

- The idea is to perform all the steps of comparing different models using cross-validation, choosing hyperparameters etc automatically.
- Very “black box”
- Many cloud platforms offer some autoML option nowadays

AutoML for our example using h2o

```
fit_aml <- h2o.automl(
     y = "SalePrice",
     training_frame = dat_h2o_train)
```

Achieves about 18,000 on held out test data, better than RF

Selects very highly tuned boosting model

Conclusion

- A new model is being invented, literally every day
- The most important skill you should practice these weeks is

  How to learn about a model you did not know about yet
- The practical should help you do this