Data Analysis & Machine Learning

Robert Bagheri

Department of Methodology, Statistics, & Data Science

Utrecht University

Today

- Who are we?
- Introduction to the course content
- Getting to know you a little
- Practicalities & course flow
- Introduction to Data Analysis and Machine Learning
- Regression & Evaluation

Who are we?

Javier Garcia Bernardo
j.garciabernardo@uu.nl

Robert A. Bagheri
a.bagheri@uu.nl

Daniel L. Oberski
d.l.oberski@uu.nl

Vira Dvoriak
v.dvoriak@uu.nl

Jonas Haslbeck
j.m.b.haslbeck@uu.nl

Elena Candellone
e.candellone@uu.nl

Gerko Vink
g.vink@uu.nl

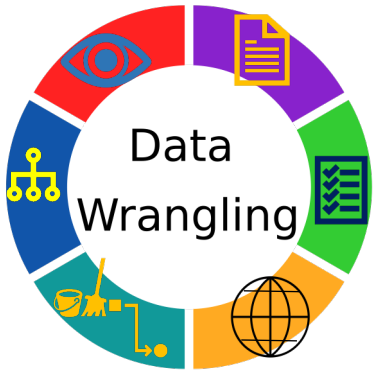
Özgür Togay
o.togay@uu.nl

Thom Volker
t.b.volker@uu.nl

Mahdi Shafiee Kamalabad
m.shafieekamalabad@uu.nl

Course Introduction

Data



Information

Regression: How do wages differ?

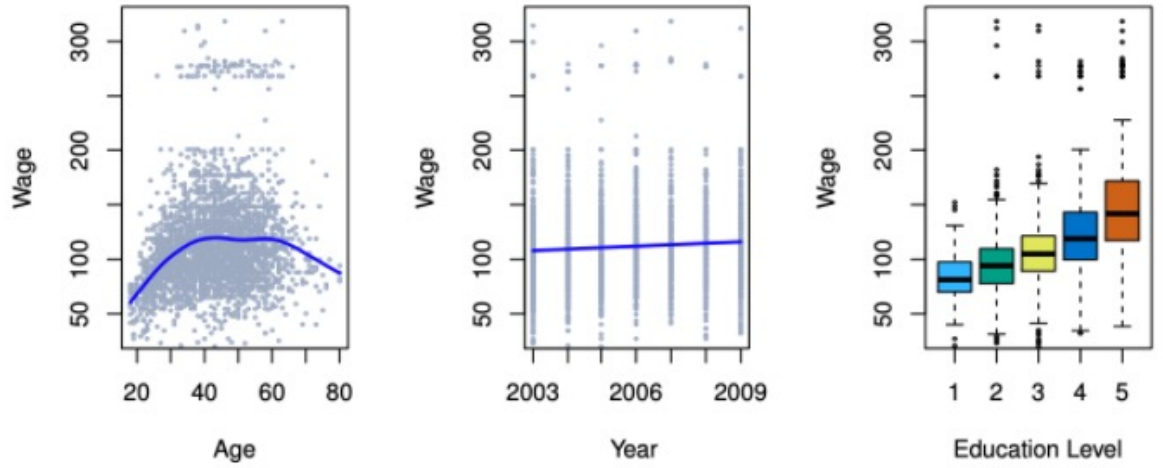
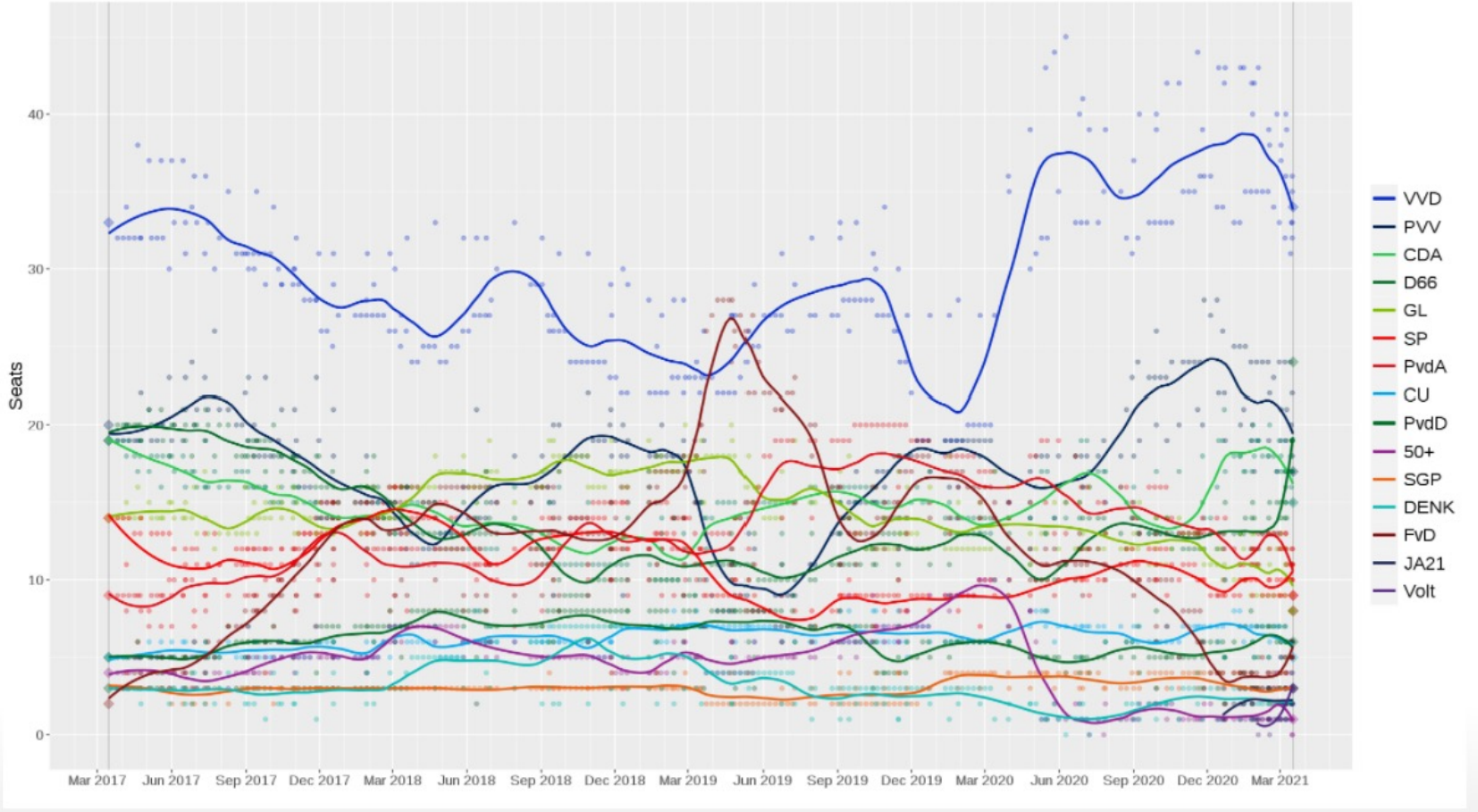


FIGURE 1.1. Wage *data, which contains income survey information for men from the central Atlantic region of the United States.* Left: wage *as a function of* age. *On average,* wage *increases with* age *until about* 60 *years of age, at which point it begins to decline.* Center: wage *as a function of* year. *There is a slow but steady increase of approximately* $10,000 *in the average* wage *between* 2003 *and* 2009. Right: *Boxplots displaying* wage *as a function of* education, *with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* wage *increases with the level of education.*

https://www.statlearning.com/

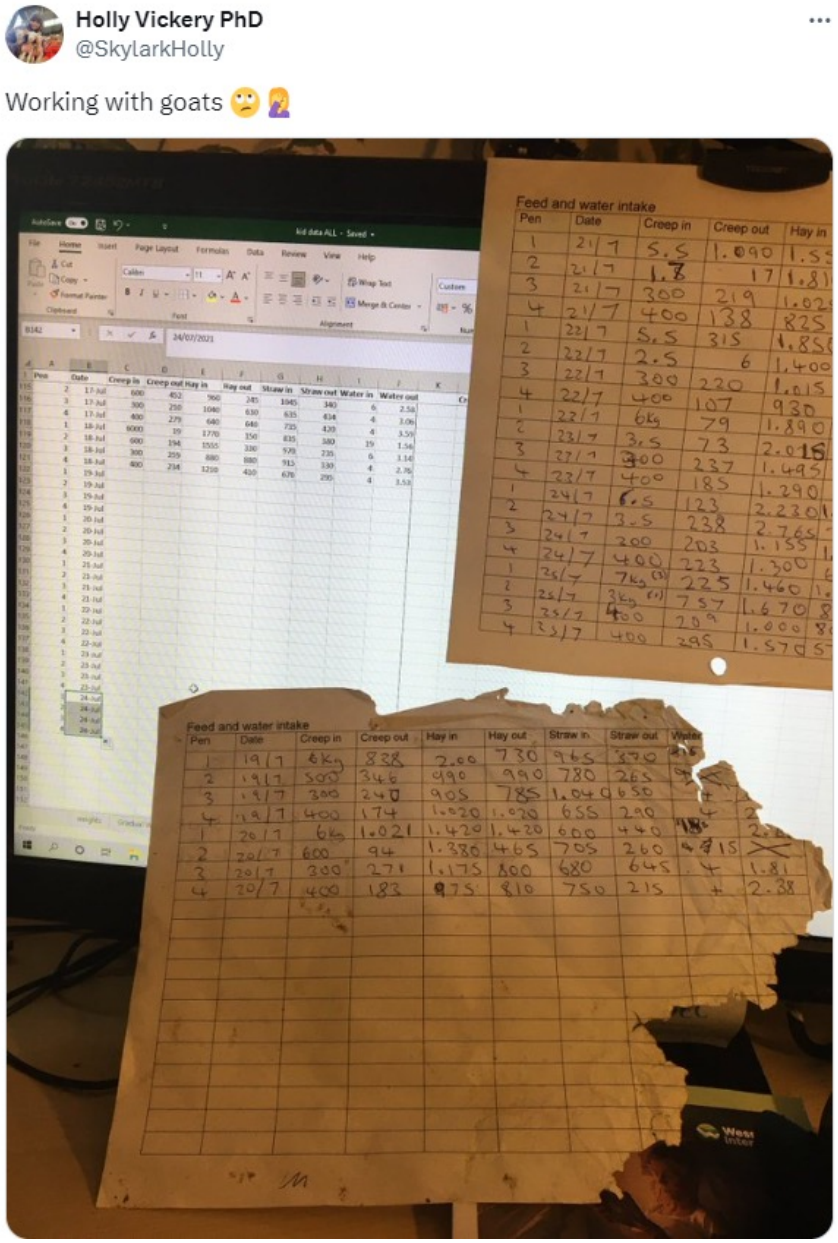
Classification & visualisation



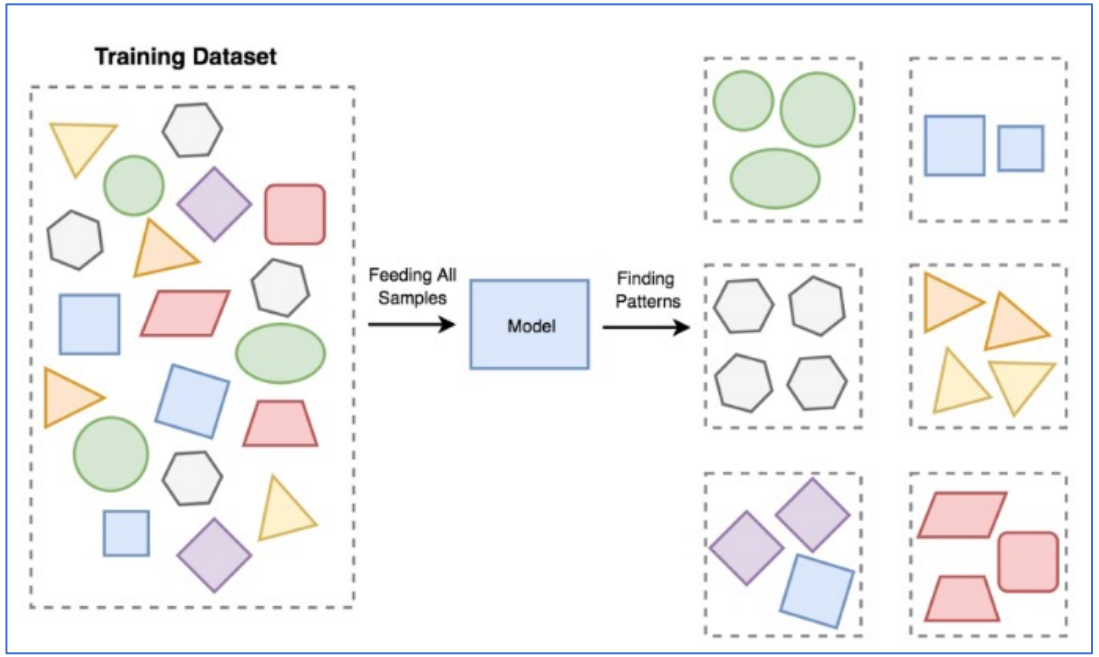
Predicting the outcome of the 2021 Dutch election:
https://gitlab.com/gbuvn1/opinion-polling-graph

Missing data

- Missing data mechanisms
- Imputation methods

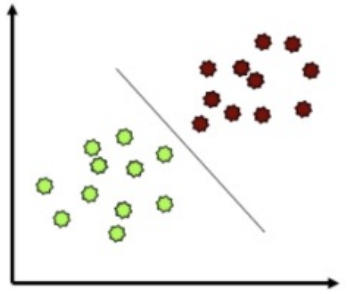


Clustering



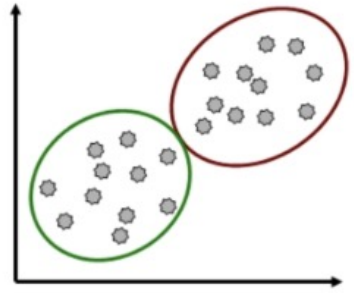
CLASSIFICATION

- Labeled data points
- Want a "rule" that assigns labels to new points
- Supervised learning

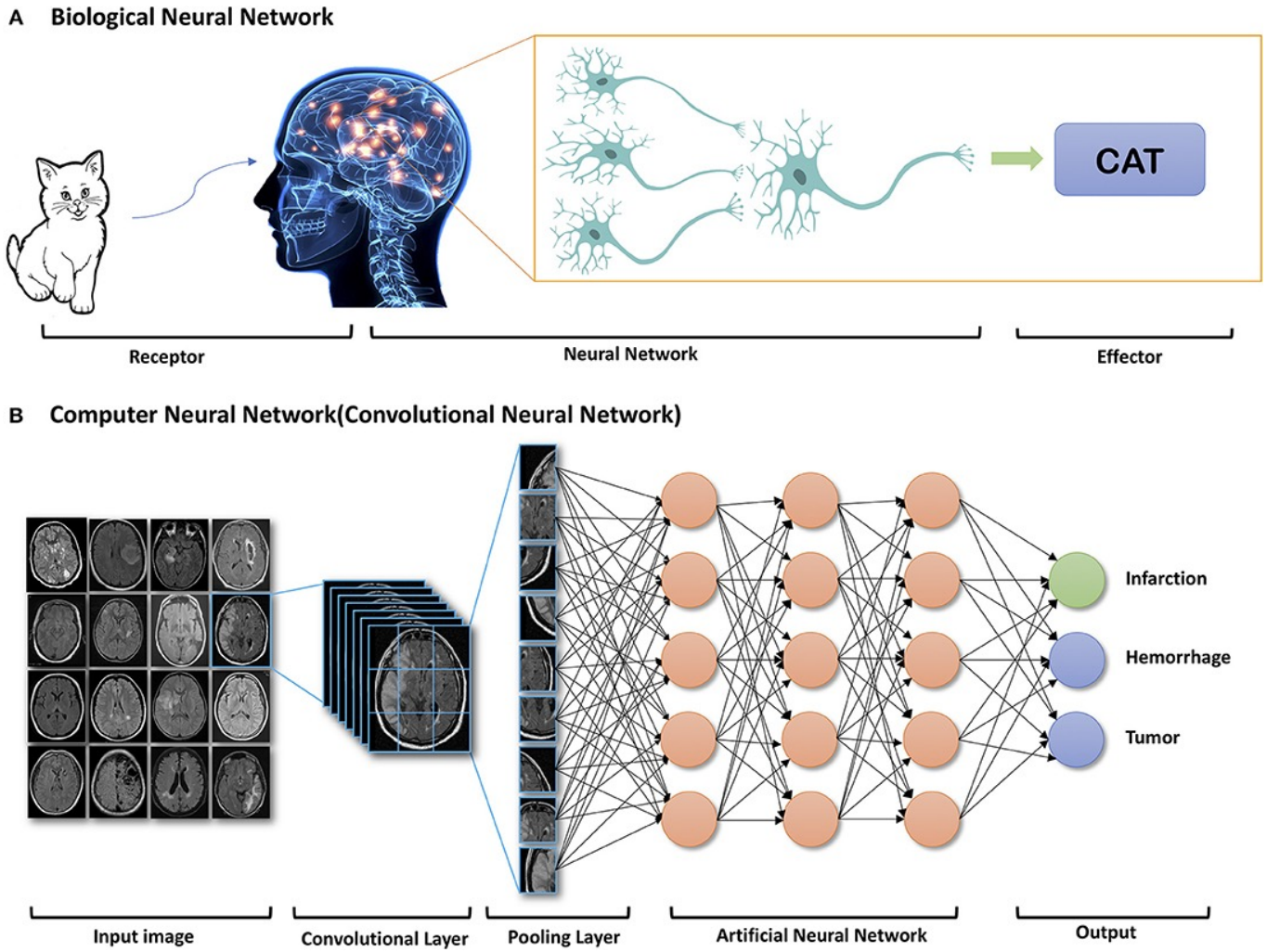


CLUSTERING

- Data is not labeled
- Group points that are "close" to each other
- Identify structure or patterns in data
- Unsupervised learning

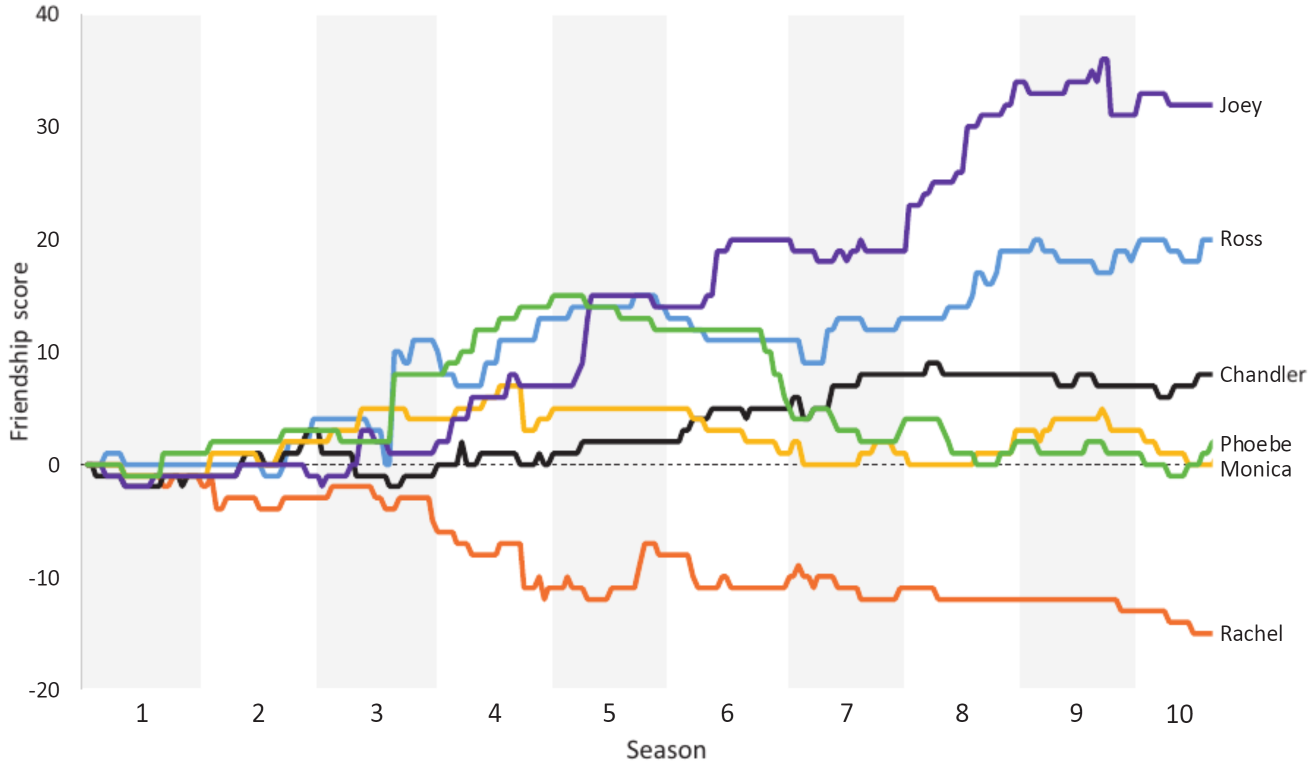


Deep learning



NLP & Text Mining: Who was the best Friend?

https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/1740-9713.01574



DAML Ice Breaker!

Waiting for participants ...

Practicalities

https://daml.robertab.nl

Practicalities

- Everything is on our course website **daml.robertab.nl**, except:
  - Room locations (on mytimetable.uu.nl)
  - Announcements (in your email and on uu.brightspace.com)
  - Assignment hand-in (on uu.brightspace.com)
  - Exam (on our digital assessment tool Remindo)
  - Lectures / Labs will be in person

🚧 **materials on the website are constantly under construction** 🏗️

Course flow

- This is a 7.5EC course
- Every week you will:

 Read the required readings
 Attend the lecture: Tuesday
 Attend and work on lab sessions: Thursday
 Work on group assignments
 Review materials diligently for the exam

Schedule & readings

- The syllabus contains a full schedule with required readings

  https://daml.robertab.nl/syllabus.html#required-readings

- Reading materials can be found online for free! 🕵️(you are resourceful students)

Lectures: Prepare!

1. Look up reading materials in syllabus
2. Look up time and location for the lecture

Labs

- Computer labs with practical exercises
- Put into practice what you learnt in the lectures
- We incorporate real-world data and use-cases: APPLIED data science!
- **Lab teachers are your main point of contact for questions!**
- Labs provide skills needed to do the assignments

Labs: prepare!

1. Quickly check what this week's lab is about
2. Look up time and location for the lab: mytimetable.uu.nl
3. Which parallel lab group are you assigned to? This should also be reflected in Brightspace

Assignments

- Two assignments (15% credit)
- Hand in on Brightspace
- Group work!
  - You can assign yourself in a group starting from tomorrow until early next week.
- **Plan ahead**

Assignments: prepare!

1. When is the deadline of the first assignment? https://daml.robertab.nl
2. Quickly check the content of the first assignment
3. Log in to Brightspace and find the INFOMDAML course
4. Find the group sign-up page there and sign up
5. If not, don't worry, it'll be there soon ☺


Read the syllabus

Student’s conference day!

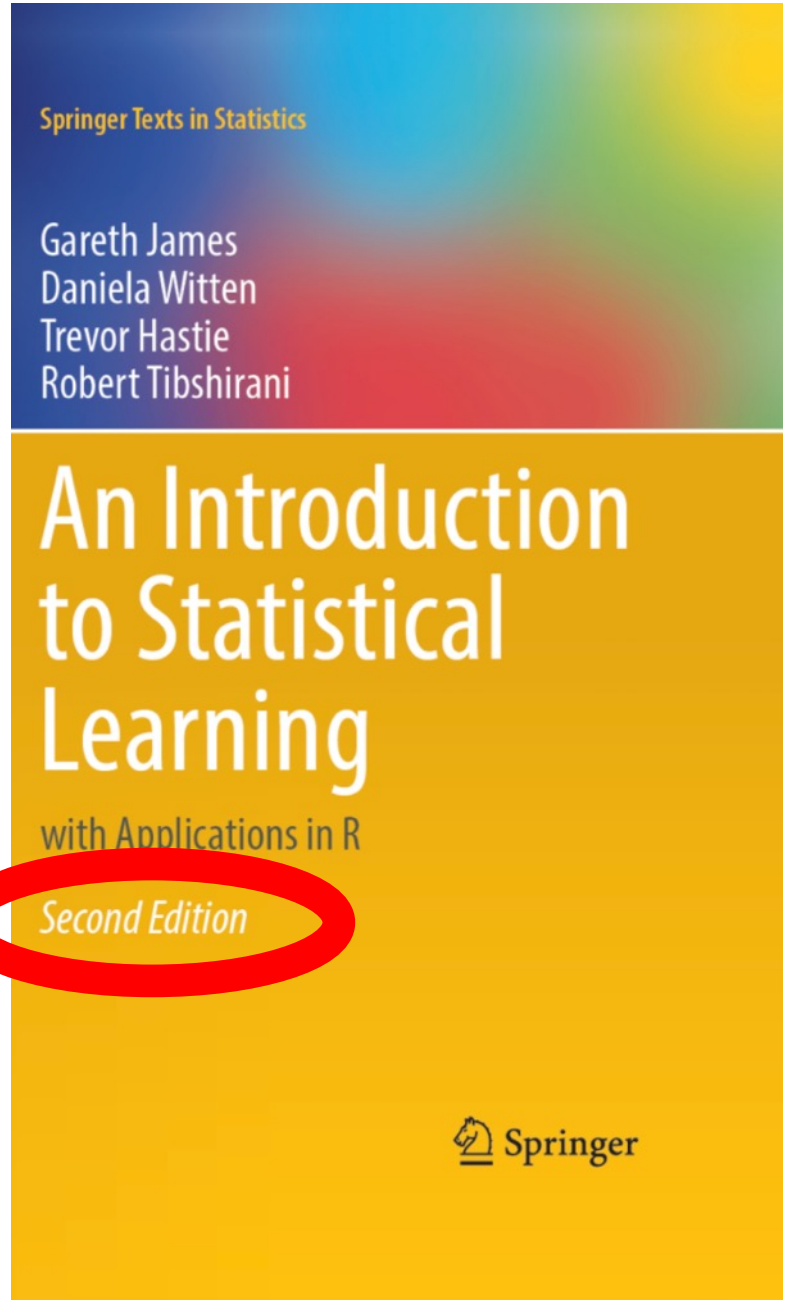
BREAK (5 minutes)

Introduction to Machine Learning

Robert Bagheri, Daniel Oberski

Department of Methodology, Statistics, and Data Science

Utrecht University



Online free version:
https://www.statlearning.com/

Bird’s eye (high-level) view of machine learning

ML definition (Samuel/Mitchell, 1959)

”A computer program is said to learn from **experience** *E* with respect to some class of **tasks** *T* and **performance measure** *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.”

Some *translation*

”A computer program is said to learn from **experience** *E* with respect to some class of **tasks** *T* and **performance measure** *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.”

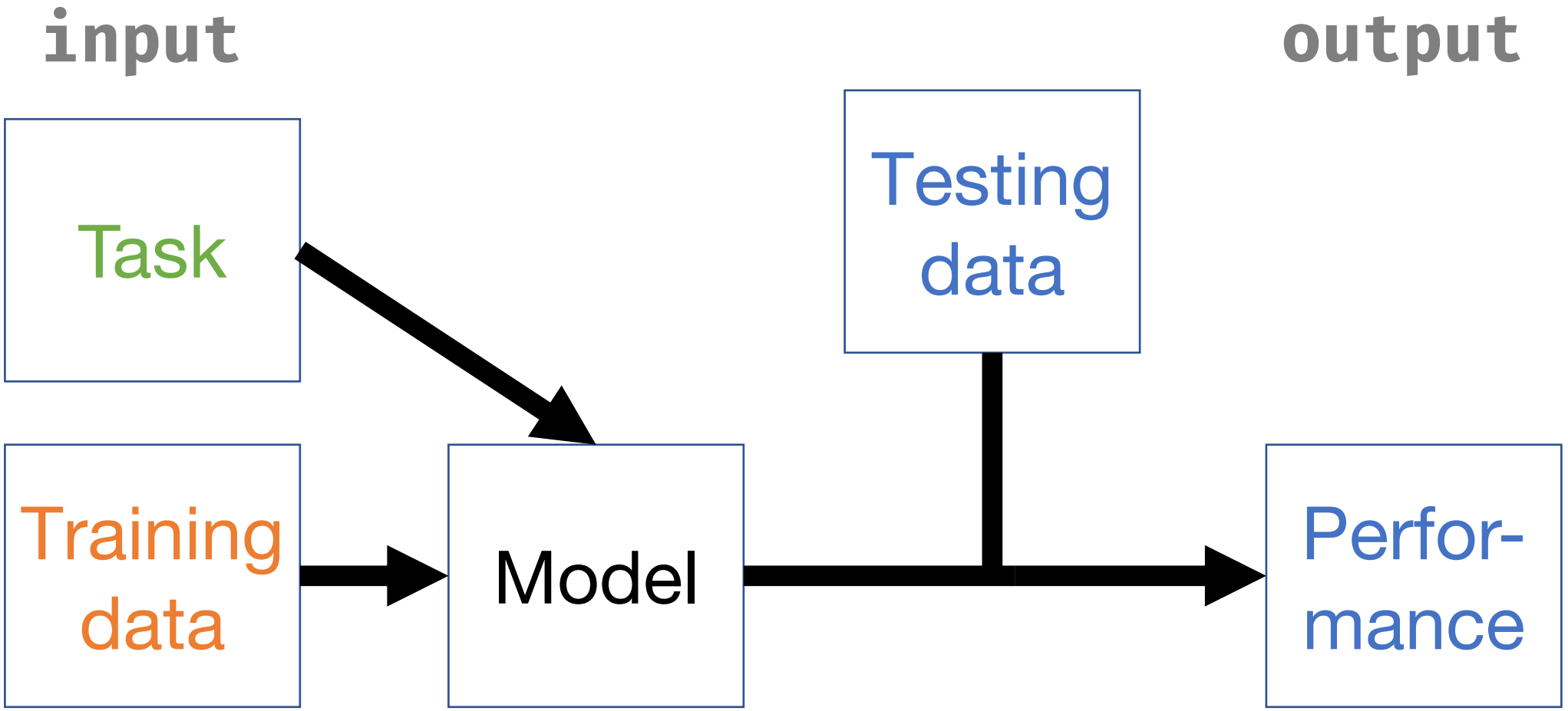
	(Loose) translations
Experience	Data; Example; Observation
Task	(Analysis) Purpose; Goal; Aim; Reason; Inference
Performance measure	Error; Loss; Cost; Risk;

Example tasks

1. Identifying customer segments and demographics to help build targeted advertising campaigns
2. Understanding sentiment of Twitter comments as either "positive" or "negative"
3. Identifying financial transactions that are potentially fraudulent
4. Ranking hotels you might like on hotel booking website
5. Recommending a book you might want to buy
6. Detecting brain lesions on an MRI
7. Predicting house prices based on house attributes such as number of bedrooms, location, and size

Utrecht University

What could be the **experience** *and the* **performance measure** *for each of these?*



good model. */*gʊd ˈmɒdl*/ *concept.* **1** (*statistics*) a model with perfect input. **2** (*machine learning*) a model with good output, as measured by the performance metric.

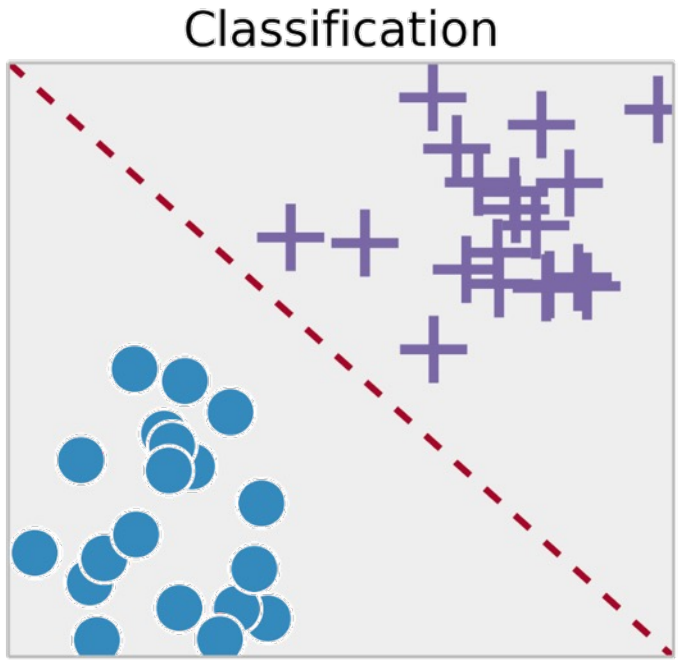
Types of machine learning

We often distinguish **3 types** of machine learning:

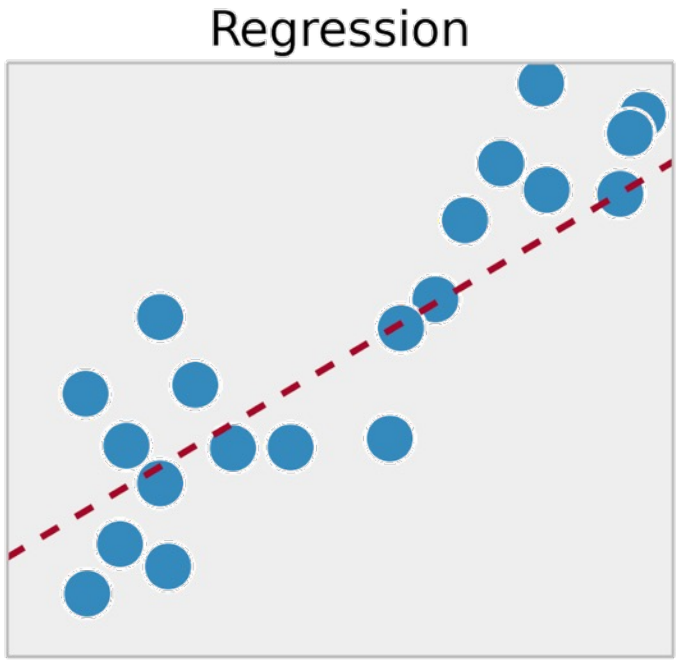
- **Supervised Learning**: learn a model from labeled training data, then make predictions
- **Unsupervised Learning**: explore the structure of the data to extract meaningful information
- **Reinforcement Learning**: develop an agent that improves its performance based on interactions with the environment

Supervised classification vs. regression

Classification: predict a class label (discrete category)

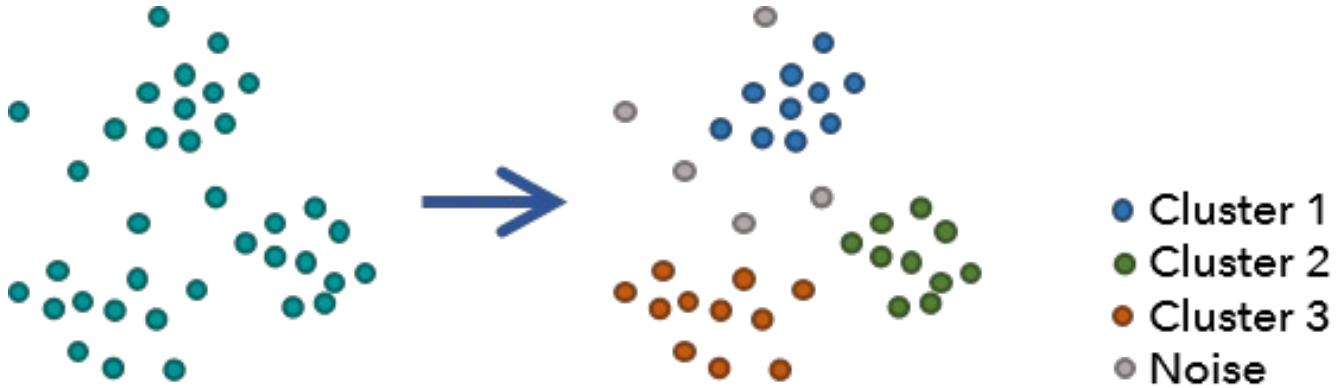


Regression:
predict a continuous value



Unsupervised learning

Learn a model from **unlabeled** training data



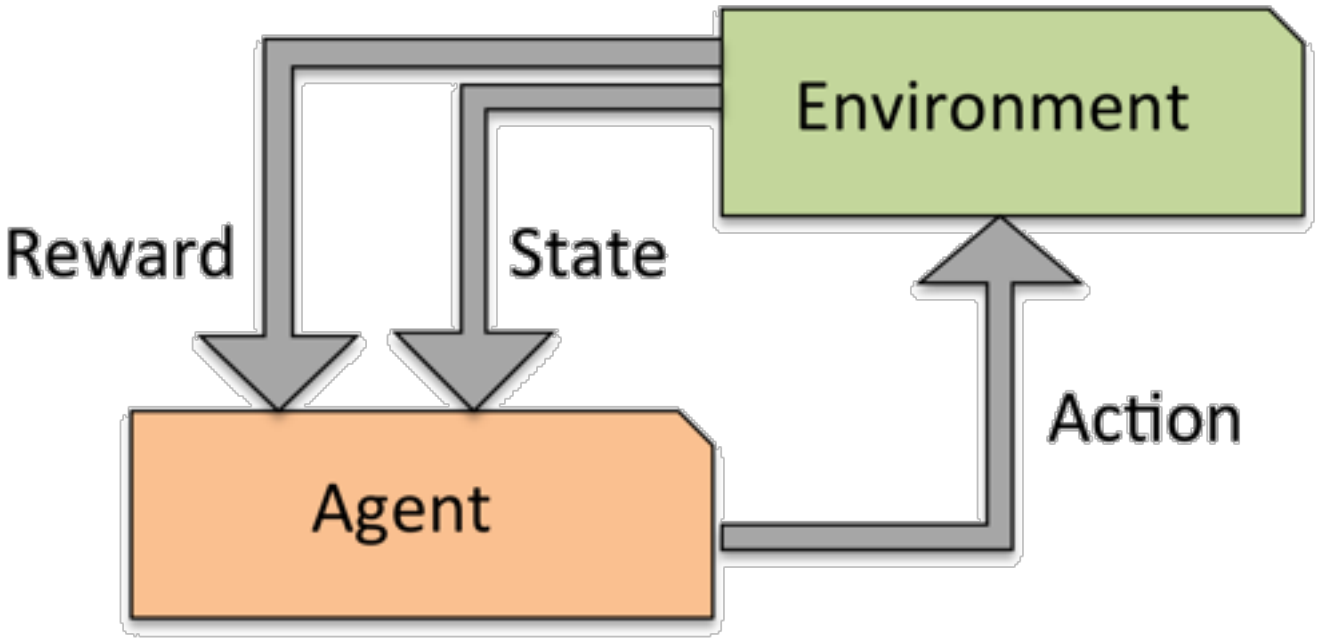
Predict “labels” from unlabeled data, or predict new data

E.g.: PCA, clustering, VAE, GANs, …

Generate a photo of students in the Master Applied Data Science at Utrecht University.

Reinforcement learning
(not in this course)

- Develop agent that improves performance based on interactions with environment
- Example: Chess, Go, ¿healthcare?…
- Reward function defines how well actions work
- Learn actions that maximize reward through exploration



Regression

Regression

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

There are usually a bunch of x's. We keep notation legible by saying **x** might be a vector of *p* predictors.

Regression

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

y : Observed outcome;

$\boldsymbol{x}$: Observed predictor(s);

$f(\boldsymbol{x})$: *True** prediction function, to be estimated (so **unknown**);

ϵ : Unobserved residuals, just *defined* as the "irreducible error",

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

The higher the variance of the irreducible error, var(ϵ) = σ, the less we can hope to explain with the data ($\boldsymbol{x}$) at hand.

Different goals of regression

Prediction:

- Given $\boldsymbol{x}$ and y, work out $f(\boldsymbol{x})$ as closely as possible.

Inference:

- "Is $\boldsymbol{x}$ related to y?"
- "How is $\boldsymbol{x}$ related to y?"
- "How precise are parameters of $f(\boldsymbol{x})$ estimated from the data?"

Galit Schmueli (2010). *To Explain or to Predict?*. Statistical Science.

Example regression techniques

Least-squares based:

- Linear regression $f(x_i) = b_0 + b_1 x_i$
- Polynomial regression $f(x_i) = b_0 + b_1 x_i + b_2 x_i^2 + b_3 x_i^3$
- Basis functions, Splines $f(x_i) = b_0 + b_1 x_i + b_2 x_i^2 + b_3 x_i^3 + b_4 (x - \xi)_+^3$

Tree-based:

- Regression trees
- Random forests
- Gradient boosting

Other:

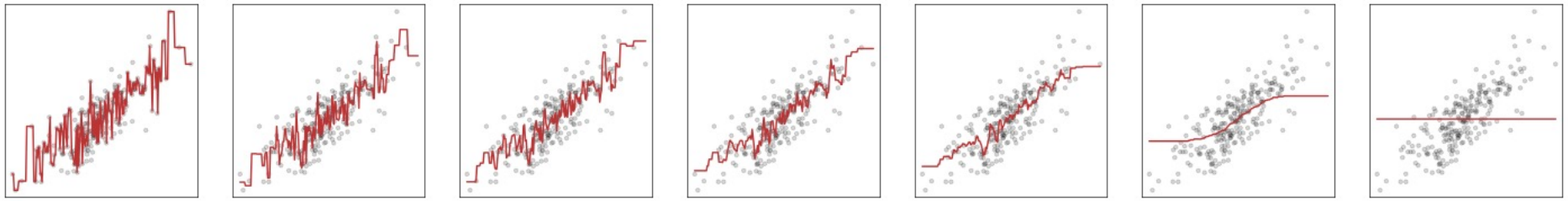
- Neural networks, deep learning, k-NN

“Nonparametric” regression

- Most modern regression techniques are ”nonparametric”
- **Nonparametric** means the *effective number of parameters increases with the sample size:*
- Model capacity grows with the amount of data
- Simplest example: k-nearest neighbors (ISLR Section 3.5)

- *Effective number of parameters for k-NN is* $\frac{n}{k}$

$k = 1$ $k = 2$ $k = 3$ $k = 5$ $k = 10$ $k = 100$ $k = n$



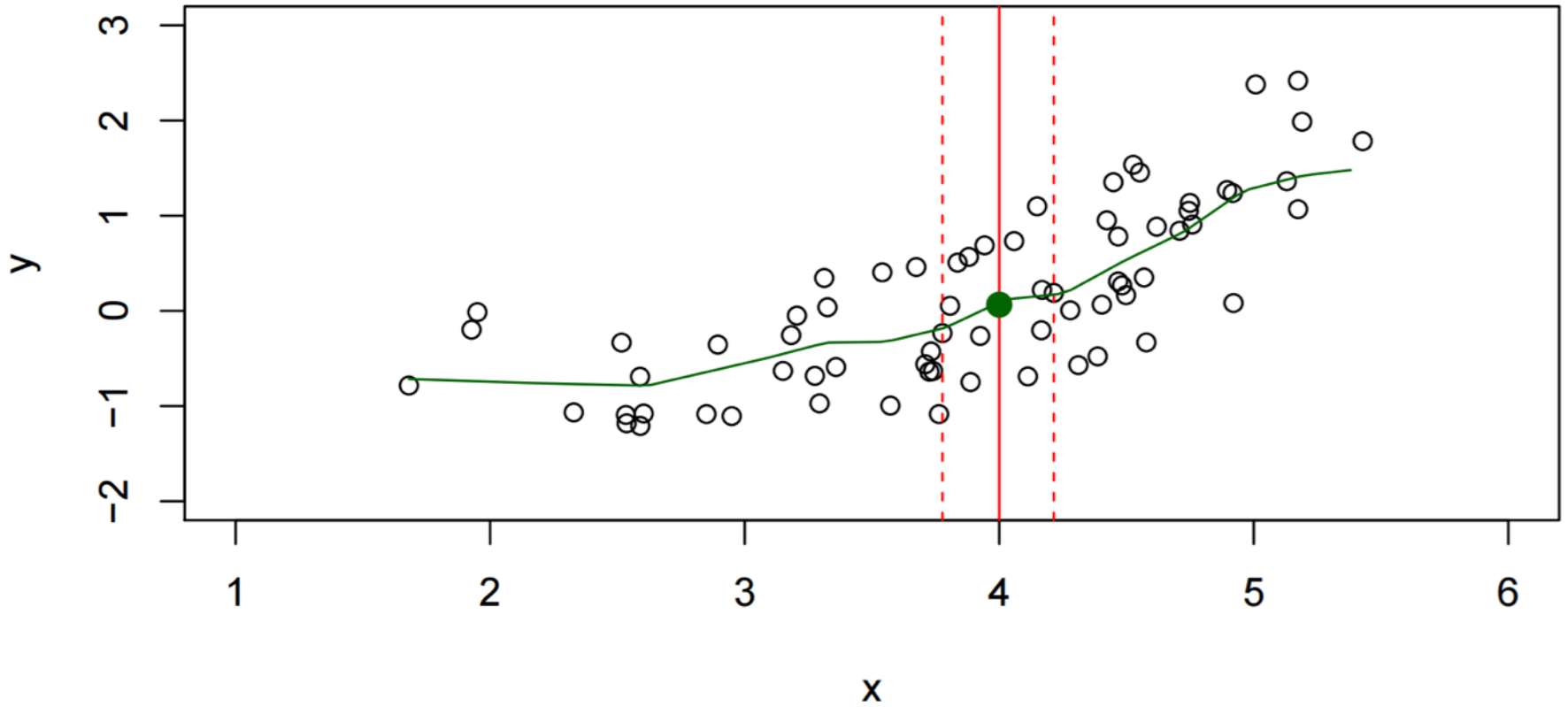
More nonparametrics.

- k-NN is not so great at controlling complexity when the number of features is more than ~5.
- Hence the rest of machine learning.

“Bottom-up”: complexity increases with data	**“Top-down”: complexity decreases down to data**
K-nearest neighbors	Random forests
Kernel regression	Deep learning / Neural nets / Transformers
Support vector regression	
Boosting	
Gaussian process regression	

Estimating $f(\boldsymbol{x})$ with k-nearest neighbors

- Typically we have no data points with $x = 4$ exactly.
- Take "neighborhood" of points around 4 and predict average:



Model performance

Model performance

- The predictions $\hat{y} = \hat{f}(x)$ differ from the true $f\ (x)$;
- We can evaluate how much this happens “on average”.

A few model evaluation metrics

- Mean squared error (MSE):

$$\text{MSE} = n^{-1} \sum_{i=1}^{n} (y - \hat{y})^2$$

- Root mean squared error RMSE $= \sqrt{\text{MSE}}$
- Mean absolute error (MAE:)

$$\text{MAE} = n^{-1} \sum_{i=1}^{n} |y - \hat{y}|$$

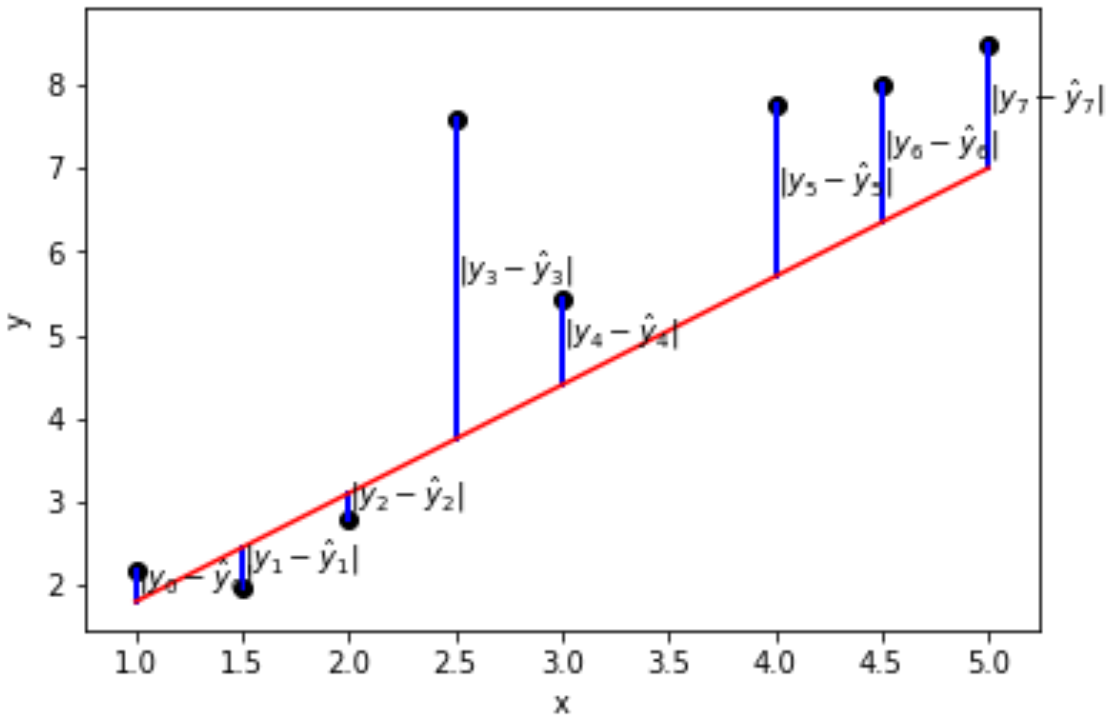
- Median absolute error (mAE):

$$\text{mAE} = \text{median}|y - \hat{y}|$$

- Proportion of variance explained: (R^2)

$$R^2 = \text{correlation}(y, \hat{y})^2$$

- etc...



Bias-variance tradeoff

Unbiased

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

High variance:

Model that easily overfits accidental patterns.

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

What do you mean by “complexity”?

- Amount of information in data absorbed into model;
- Amount of compression performed on data by model;
- Effective number of parameters

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

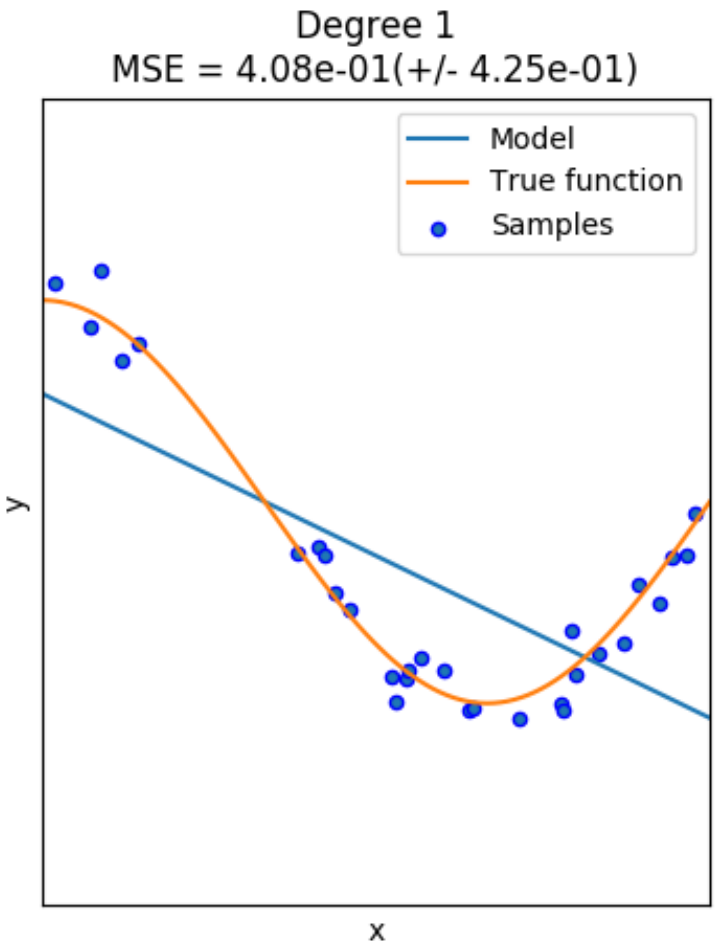
- More predictors/features in linear regression
- Higher-order polynomial in linear regression (x^2, x^3, x^4, etc.);
- Smaller k in kNN
- …

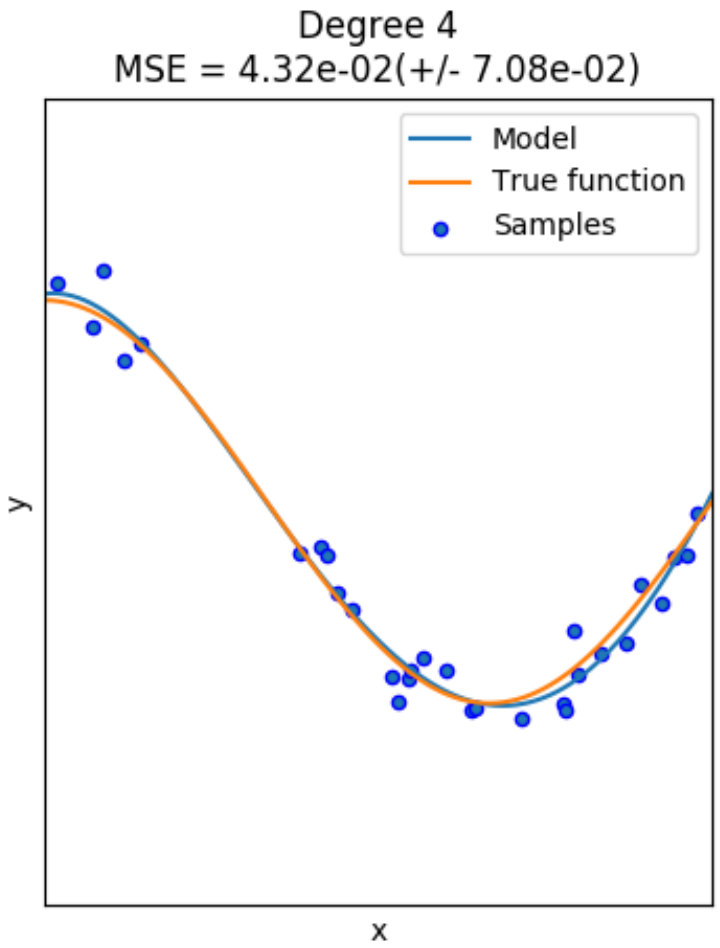
$$E(\text{MSE}) = \text{Bias}^2 + \text{Variance} + \sigma$$

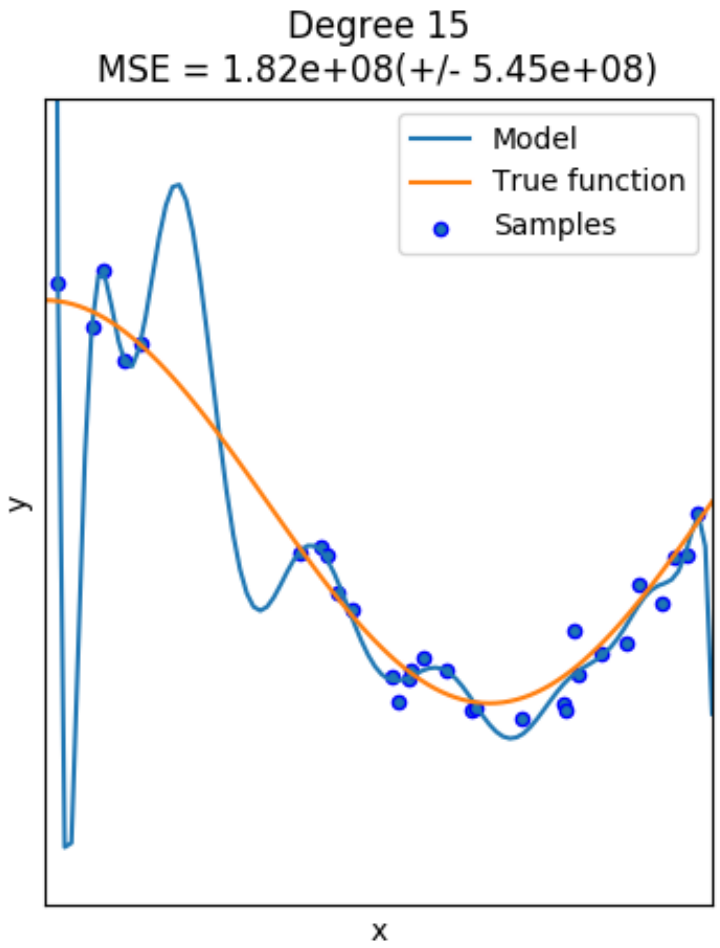
Population mean squared error is squared bias PLUS model variance PLUS irreducible variance.

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

Overfitting vs Underfitting







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**

The train-val-test paradigm

Train/dev/test

Training data:

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

Validation data (or "dev" data):

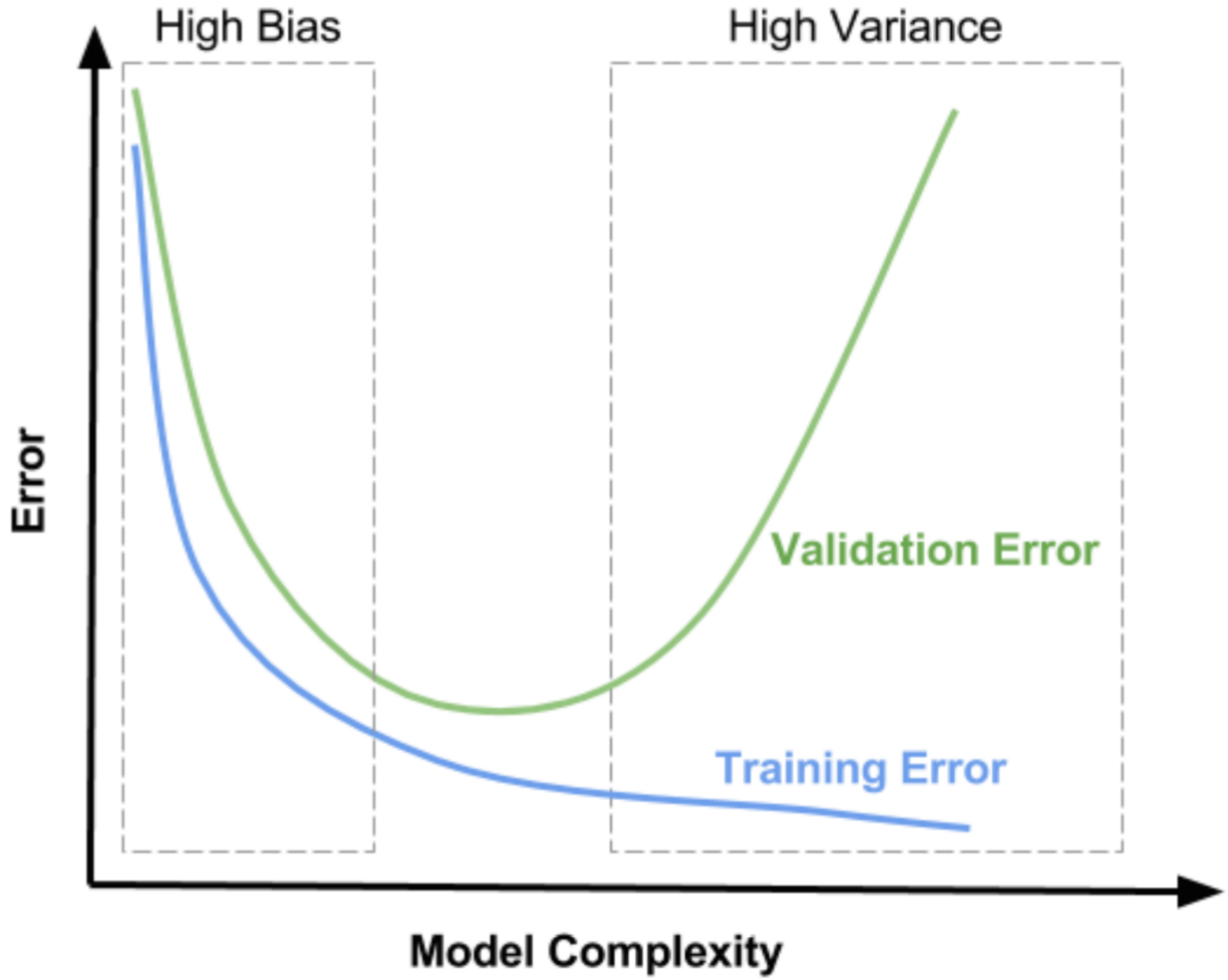
New observations from the same source as training data

Used several times to select model complexity)

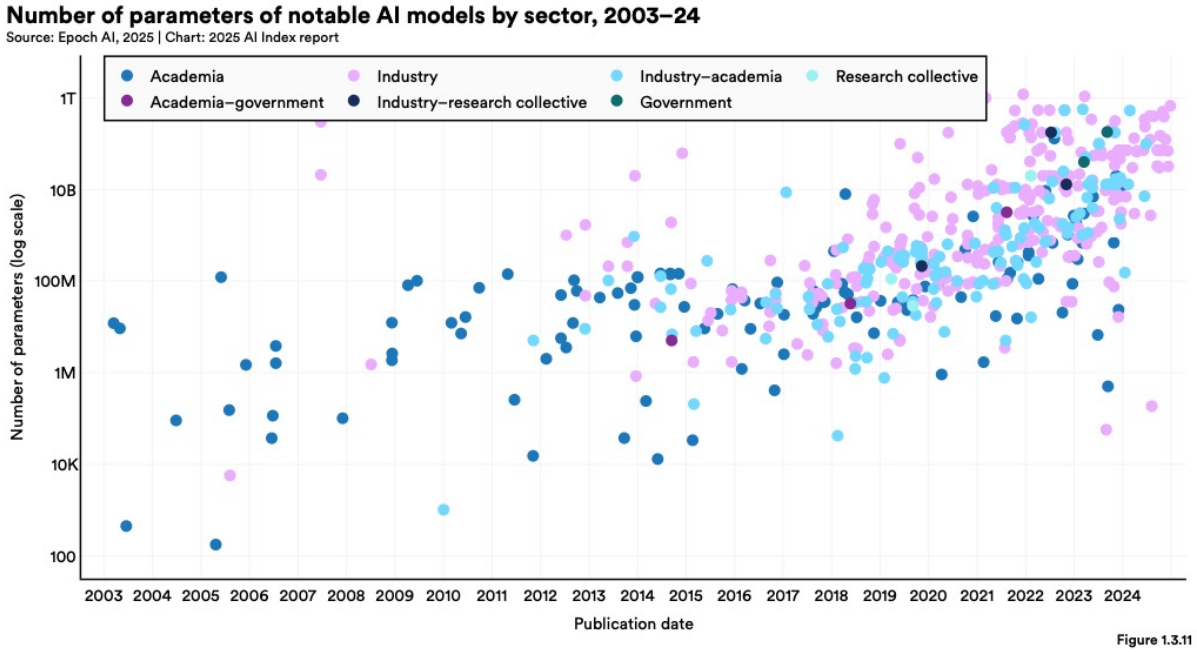
Test data:

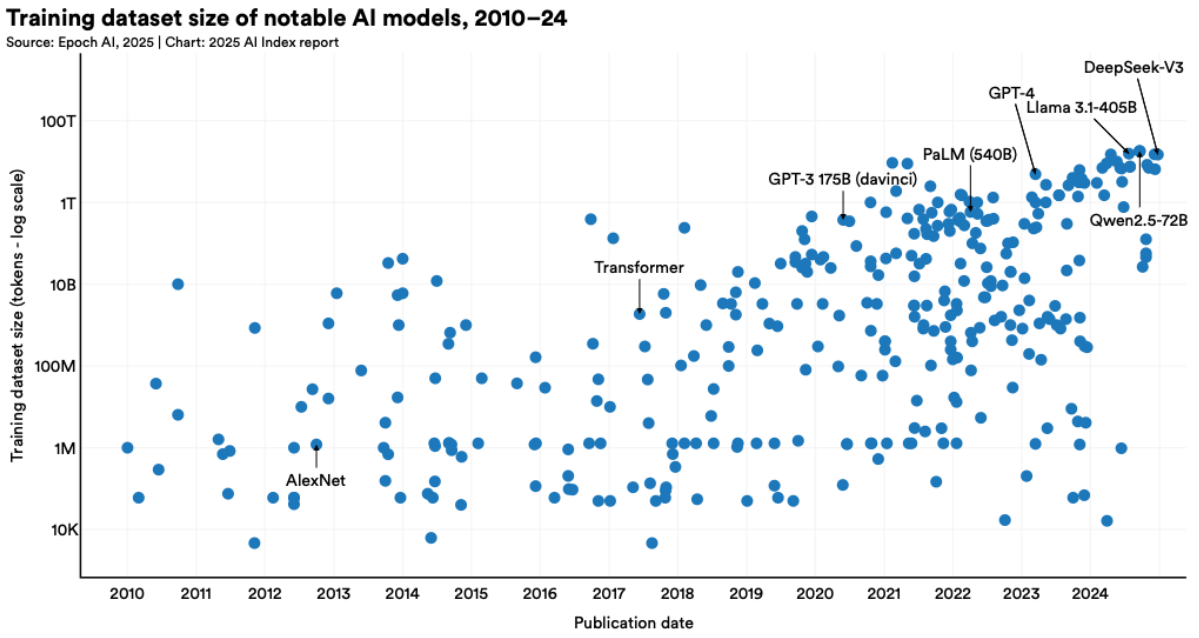
New observations from the intended prediction situation

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


High Bias
High Variance
Error
Validation Error
Training Error
Model Complexity

LLM #parameters and dataset size



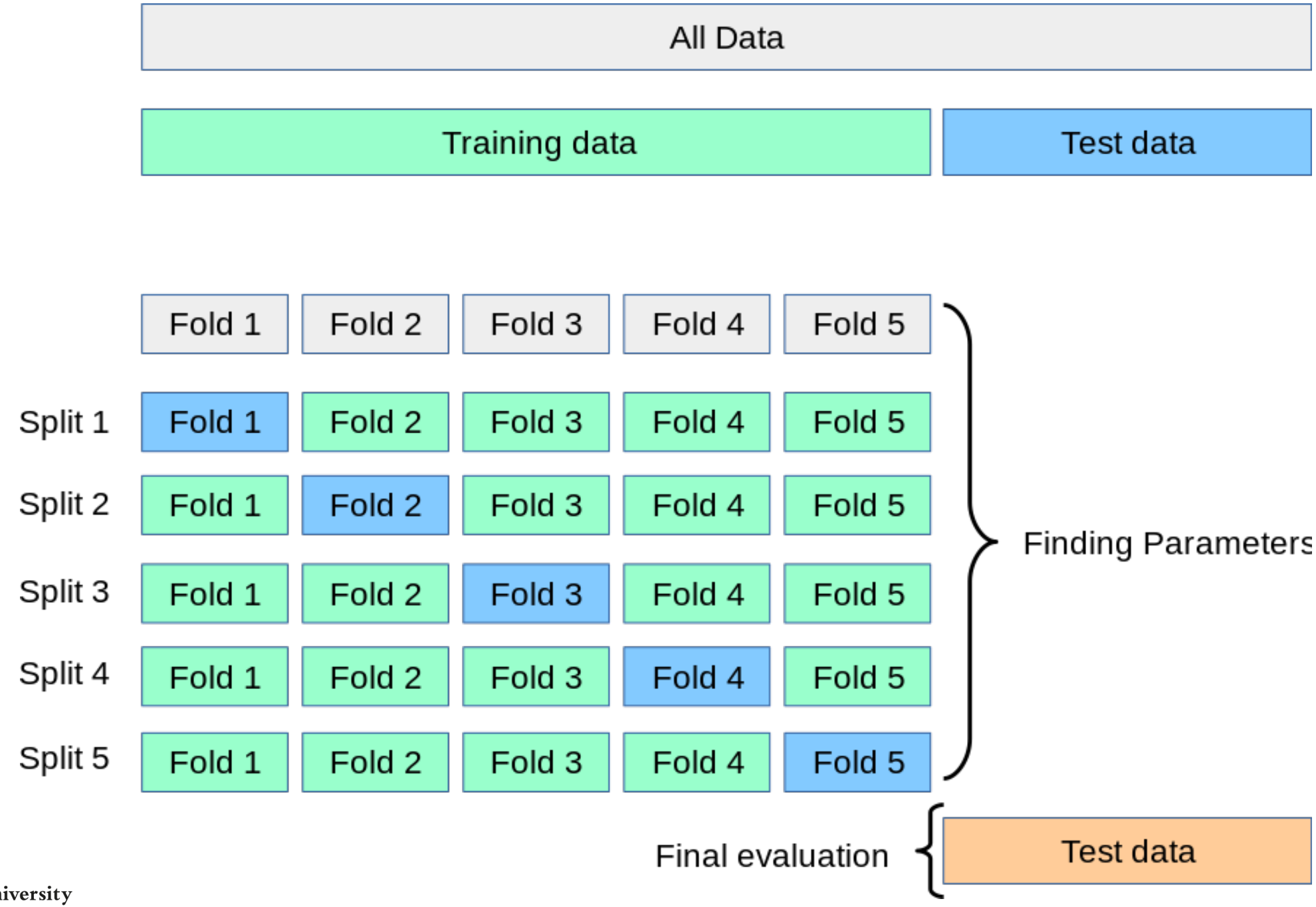


Drawbacks of train/dev/test

- 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 cross validation

- “Cross-validation” replaces single dev set approach;
- Perform the train/dev split several times, average the result.
- 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

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.

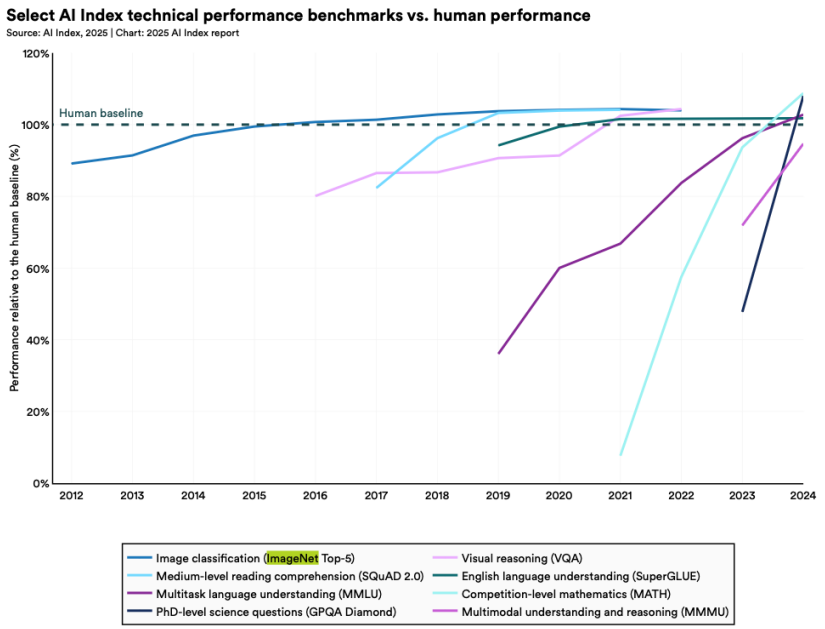
(c) A **scoring referee**, to which competitors can submit their prediction rule.

- 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.

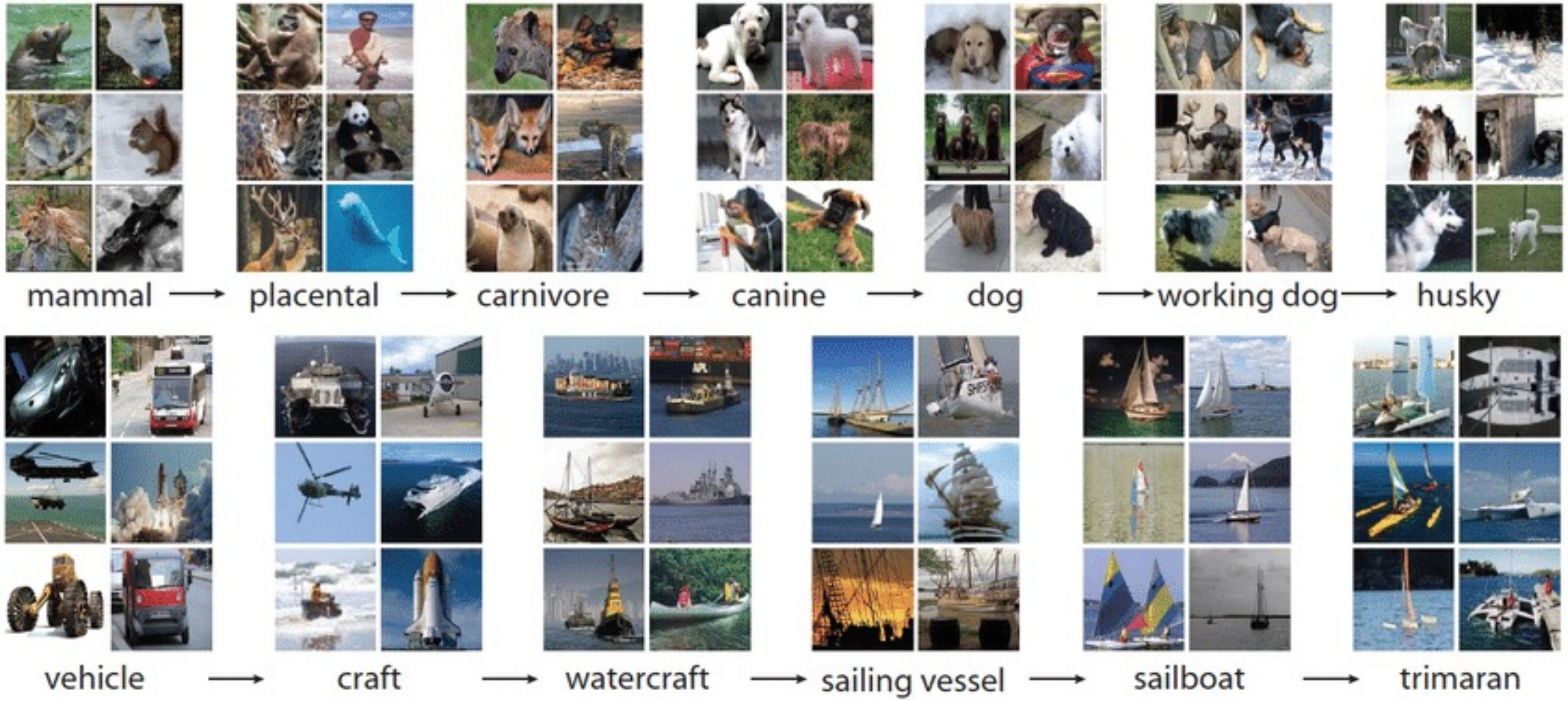
Donoho (2017). *50 years of data science.* J Comp Graphical Stat.

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


mammal
placental
carnivore
canine
dog
working dog
husky
vehicle
craft
watercraft
sailing vessel
sailboat
trimaran

IMAGENET CHALLENGE: TOP-5 ACCURACY

Source: Papers with Code, 2020; AI Index, 2021 | Chart: 2021 AI Index Report

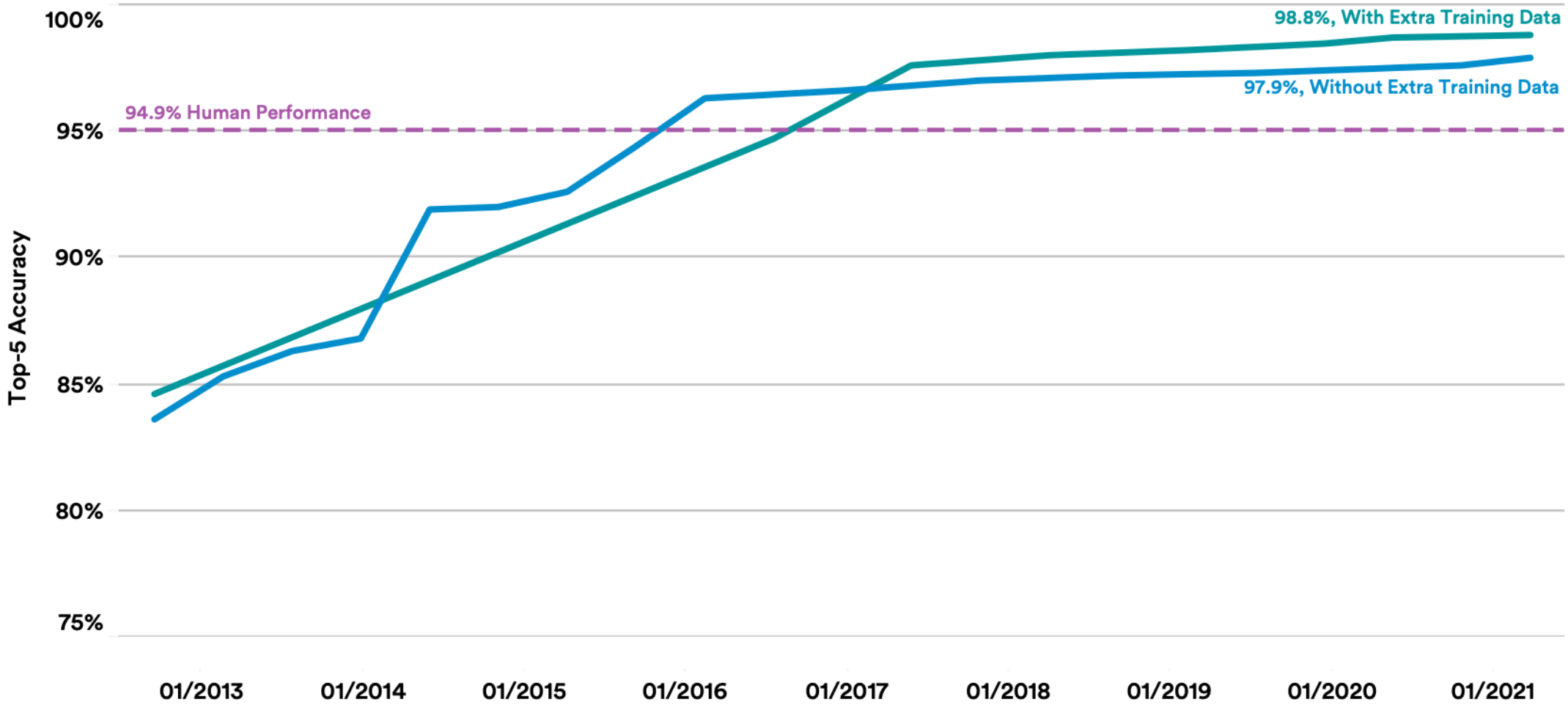


Figure 2.1.2

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

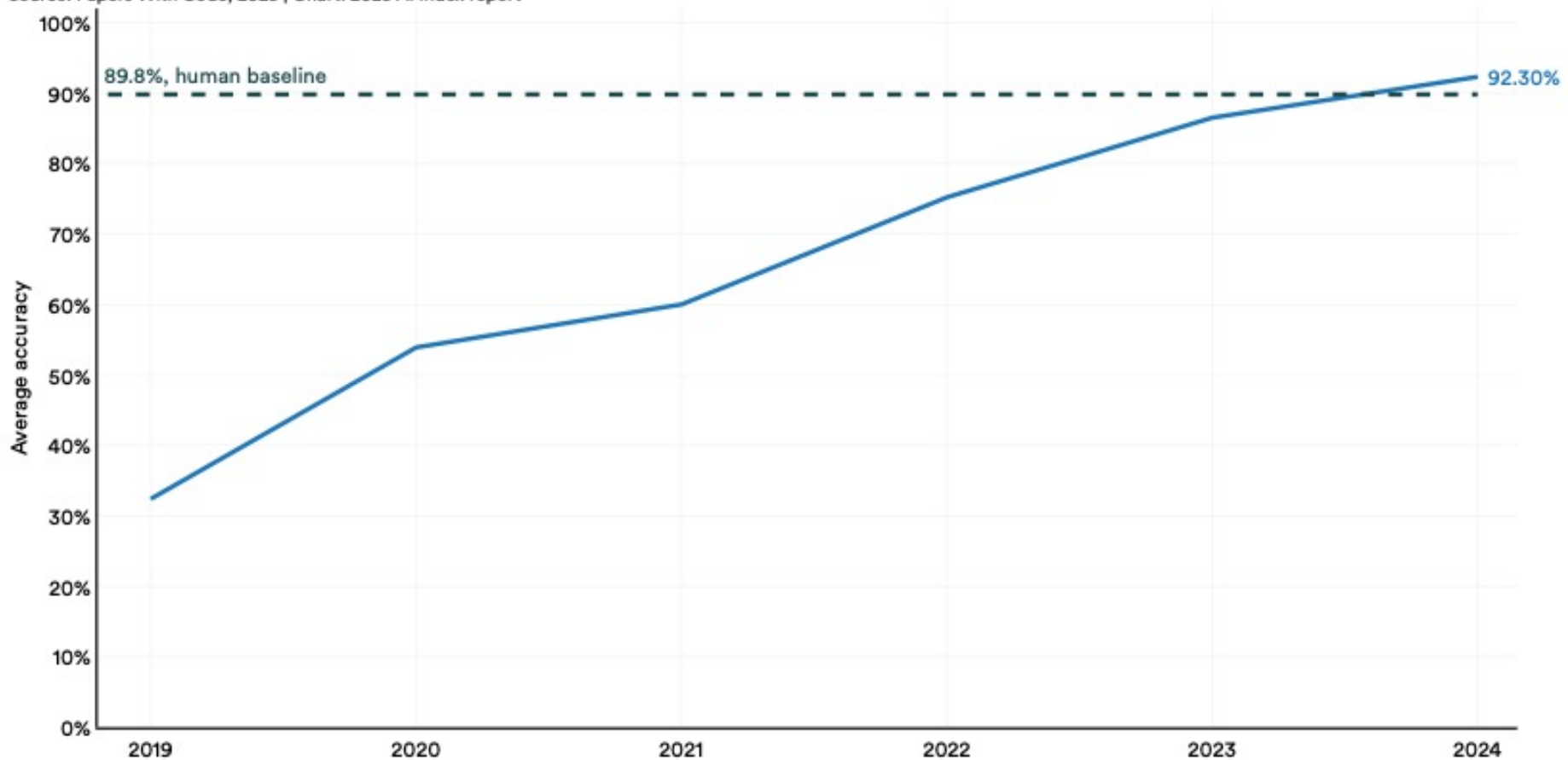
Microeconomics

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

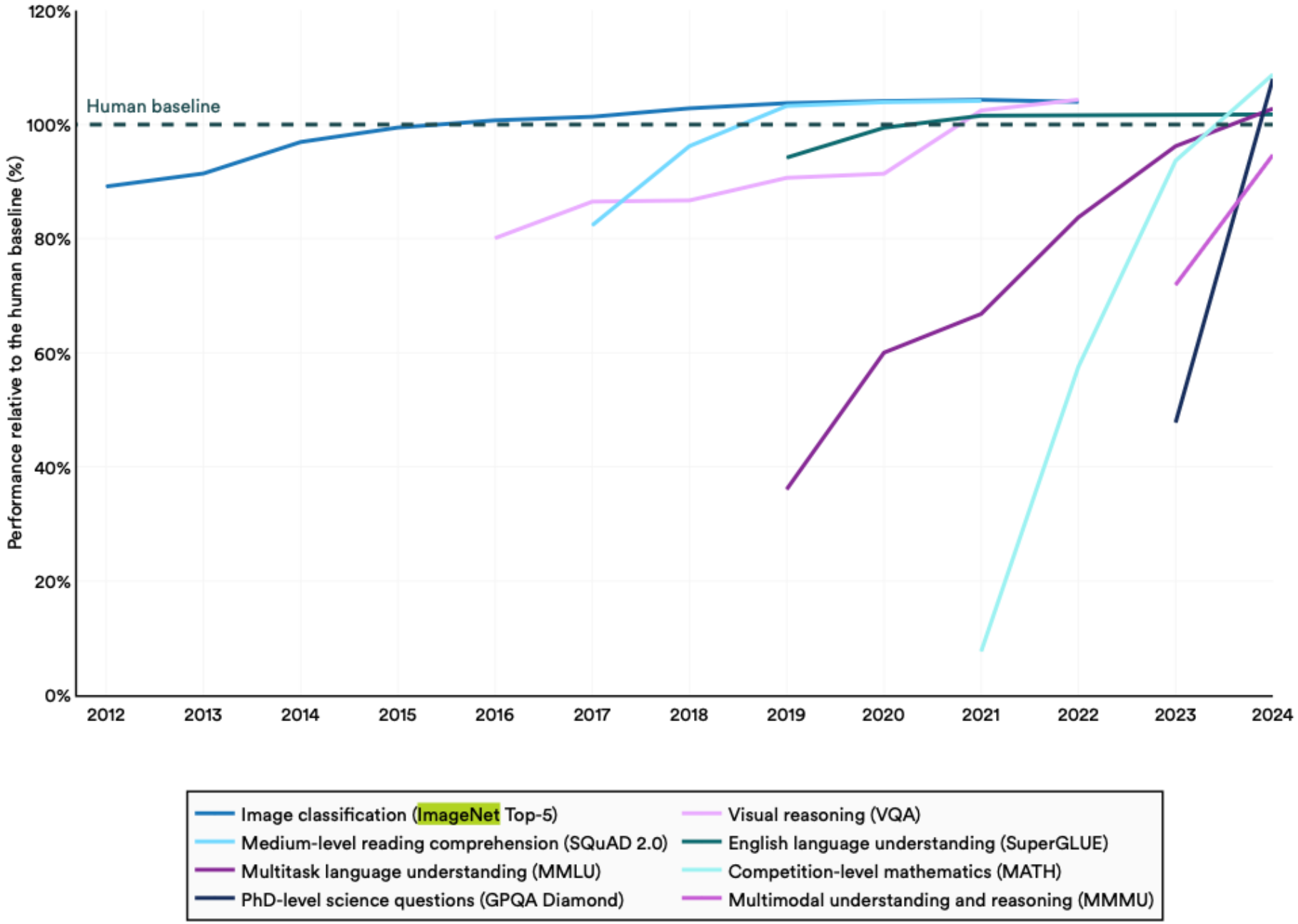
MMLU: average accuracy

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

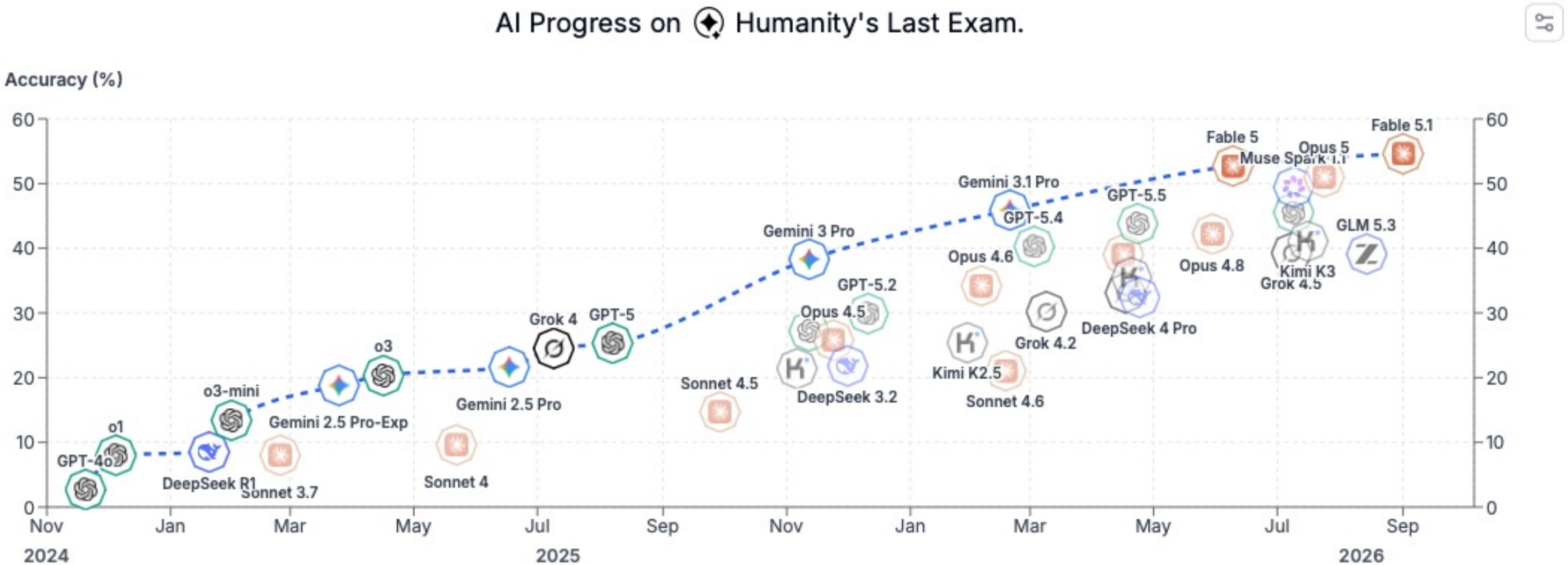


“Humanity’s Last Exam”

- Humanity's Last Exam (HLE) is a global collaborative effort, with questions from nearly 1,000 subject expert contributors affiliated with over 500 institutions across 50 countries – comprised mostly of professors, researchers, and graduate degree holders.

https://agi.safe.ai

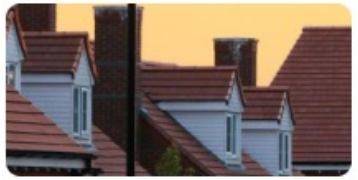
Source: Phan et al., 2025


AI Progress on Humanity's Last Exam.
Accuracy (%)
60
50
40
30
20
10
0
Nov
2024
Jan
Mar
May
Jul
2025
Sep
Nov
Jan
Mar
May
Jul
2026
Sep
GPT-4o
o1
DeepSeek R1
o3-mini
Sonnet 3.7
Gemini 2.5 Pro-Exp
o3
Sonnet 4
Gemini 2.5 Pro
Grok 4
GPT-5
Sonnet 4.5
Gemini 3 Pro
Opus 4.5
GPT-5.2
DeepSeek 3.2
Opus 4.6
Gemini 3.1 Pro
GPT-5.4
Kimi K2.5
Sonnet 4.6
Grok 4.2
GPT-5.5
DeepSeek 4 Pro
Fable 5
Opus 4.8
Muse Spark 1.1
Opus 5
Kimi K3
Grok 4.5
GLM 5.3
Fable 5.1

Example - CTF

House Prices - Advanced Regression Techniques

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



Overview Data Code Models Discussion Leaderboard Rules

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

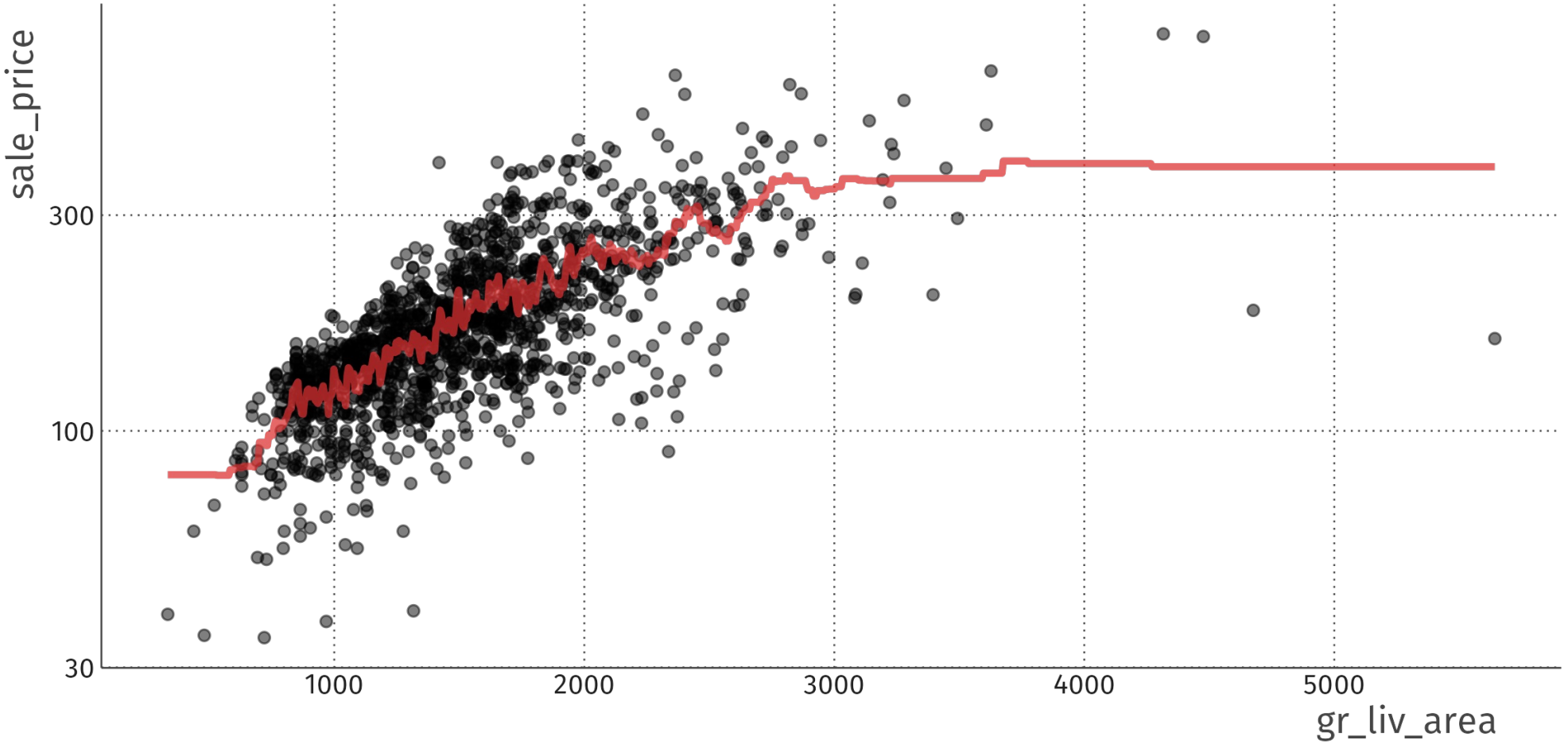
Reminder

- **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:

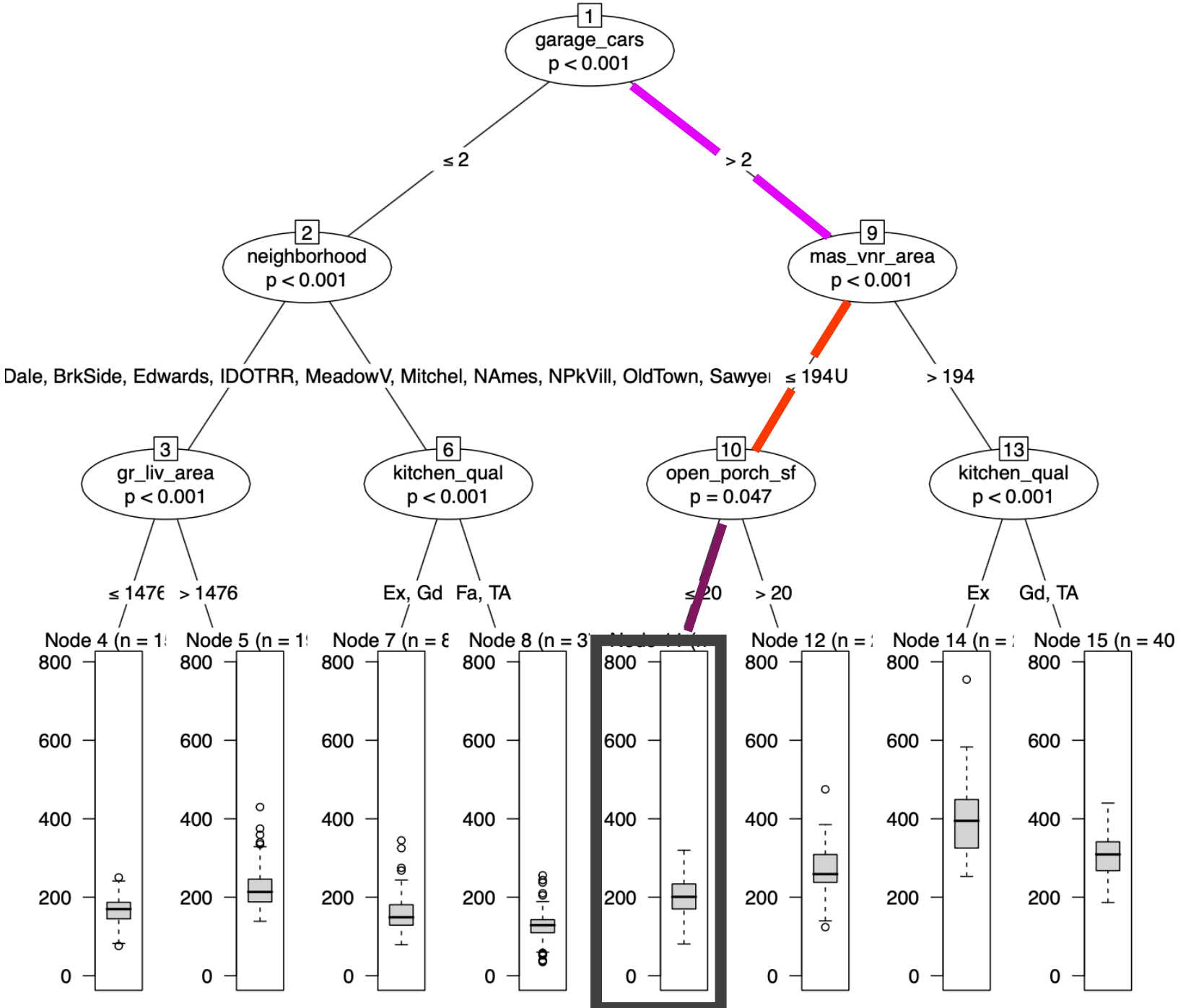
- `garage_cars = 3`
- `mas_vnr_area = 104`
- `open_porch_sf = 10`

$$\hat{y} = 200$$

What about for:

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

$$\hat{y} = ?$$



Regression tree

Prediction for:

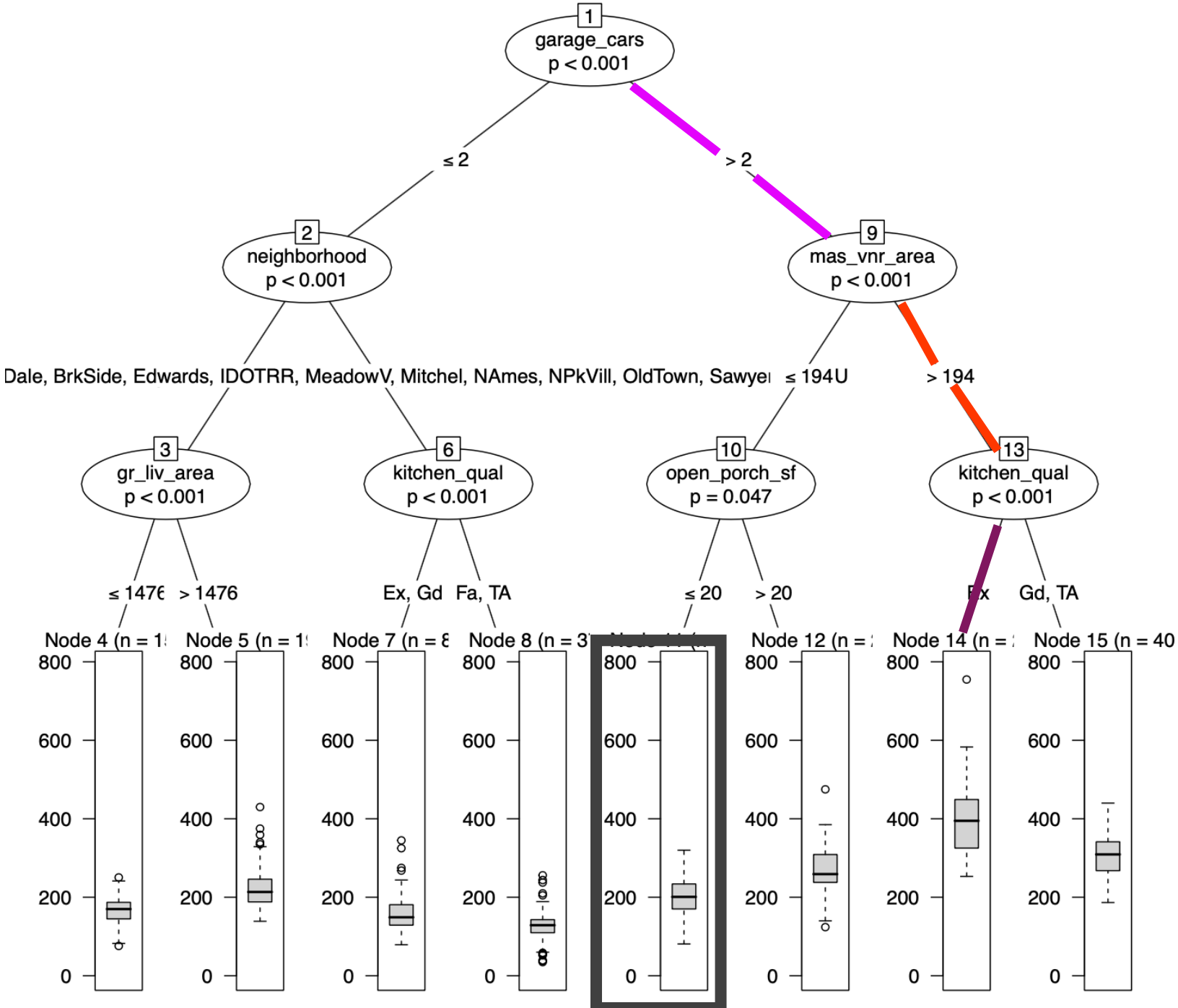
- garage_cars = 3
- mas_vnr_area = 104
- open_porch_sf = 10

$$\hat{y} = 200$$

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

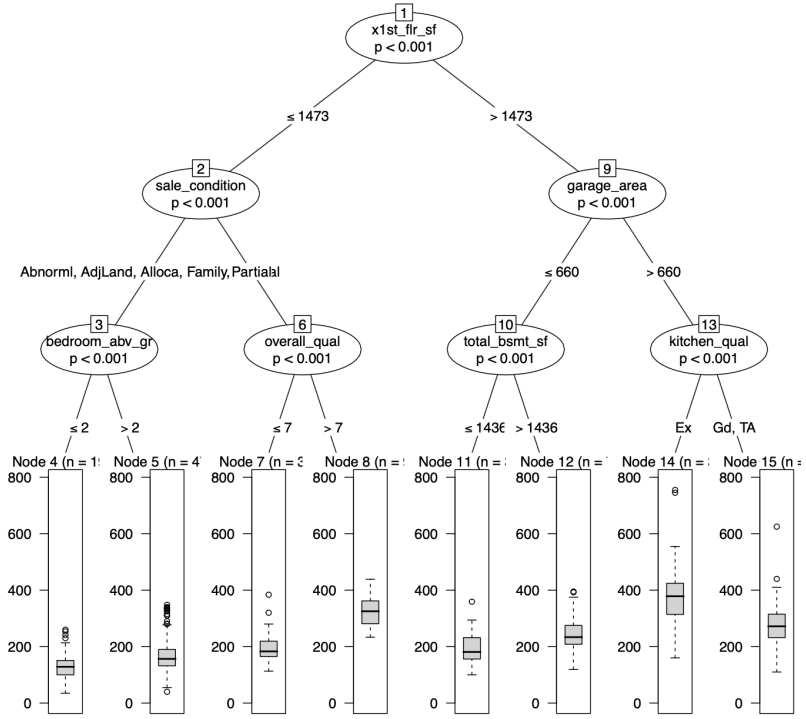
→ Use them as **basic building block** for **ensembles**

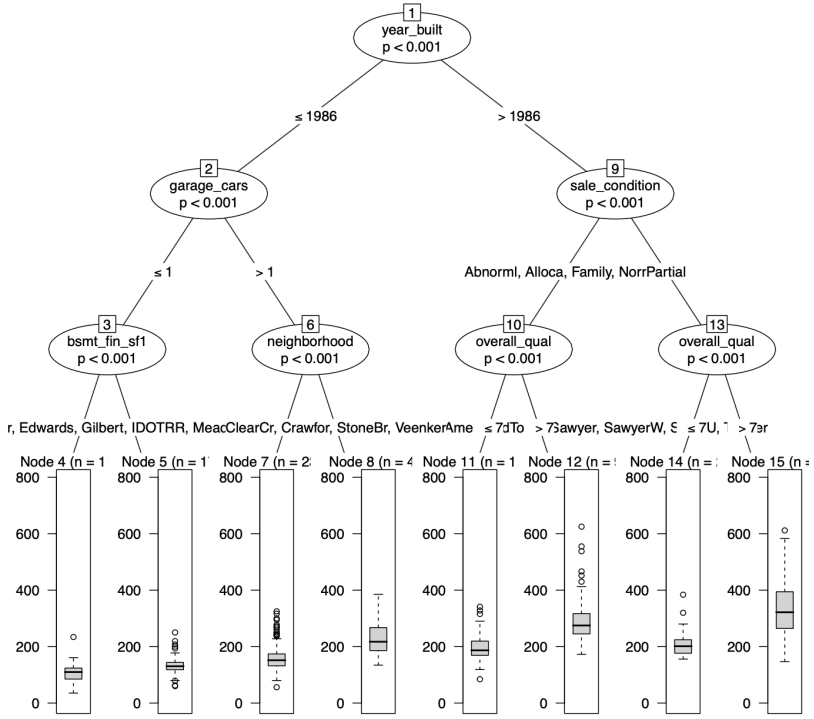
- **Random forests**: “bagged trees with feature sampling”
  - Make trees that are too complex (low bias, high variance);
  - Average over bootstrapped samples to cancel out the overfitting parts.

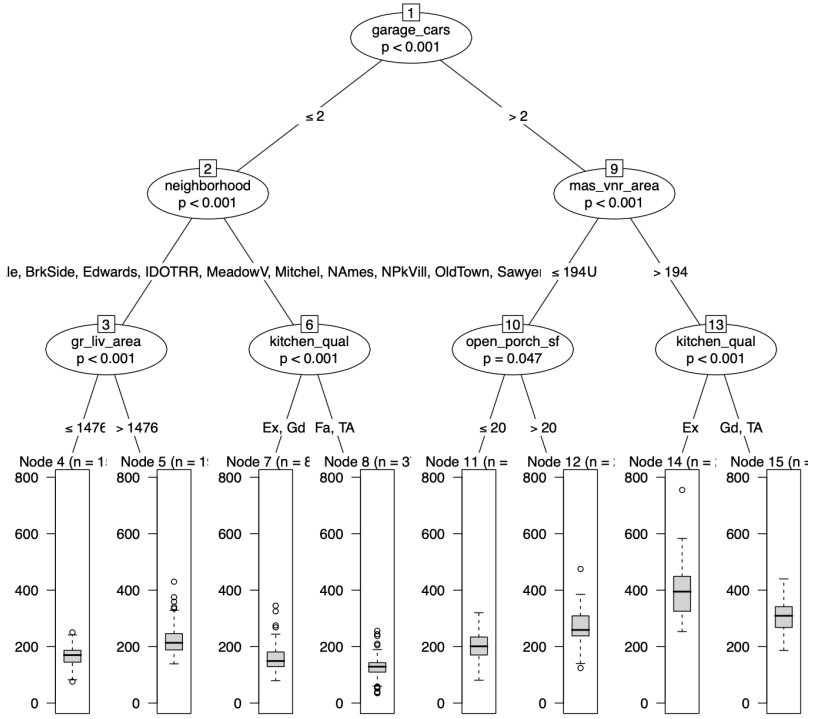
- **Boosting**: “ensemble of weak learners”
  - Make trees that are too simple (high bias, low variance);
  - Make more of them for observations with big residuals;
  - Average them.

More about random forests and boosting next time!

Random forests

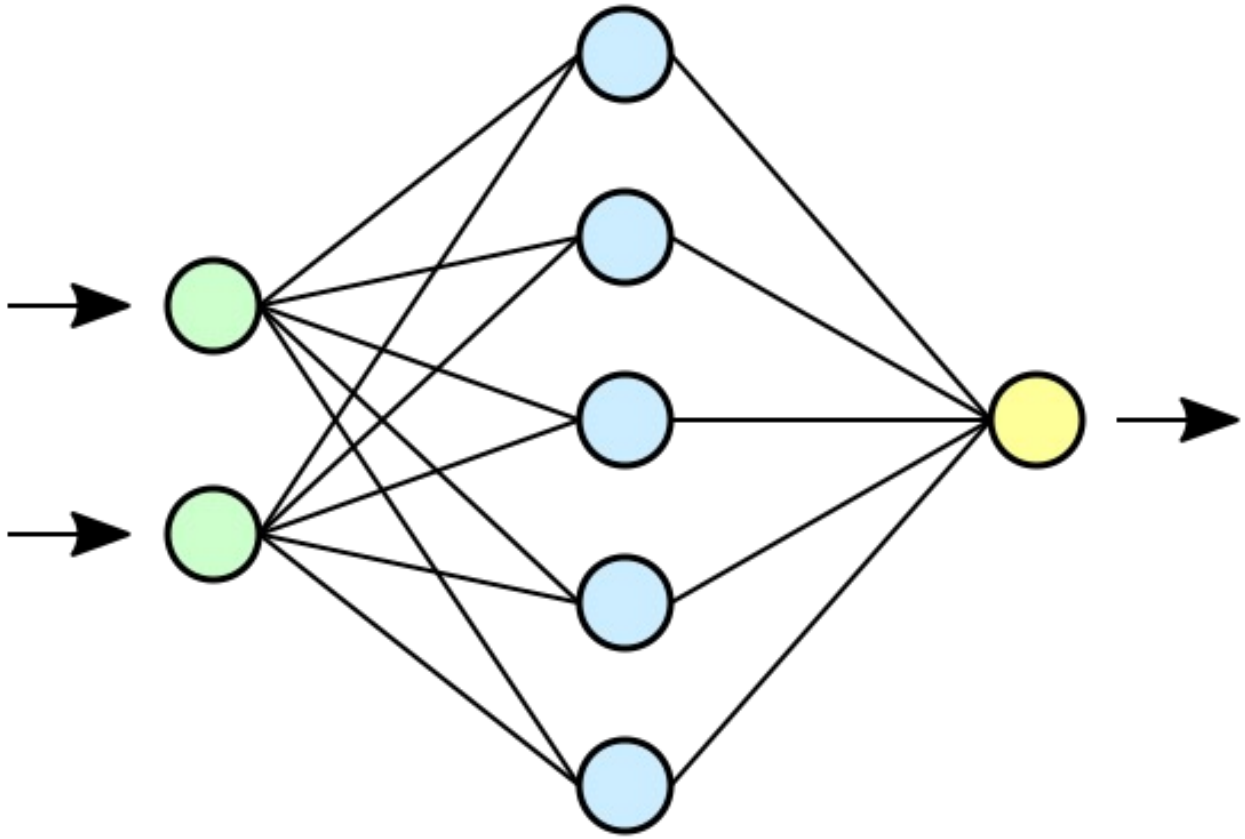




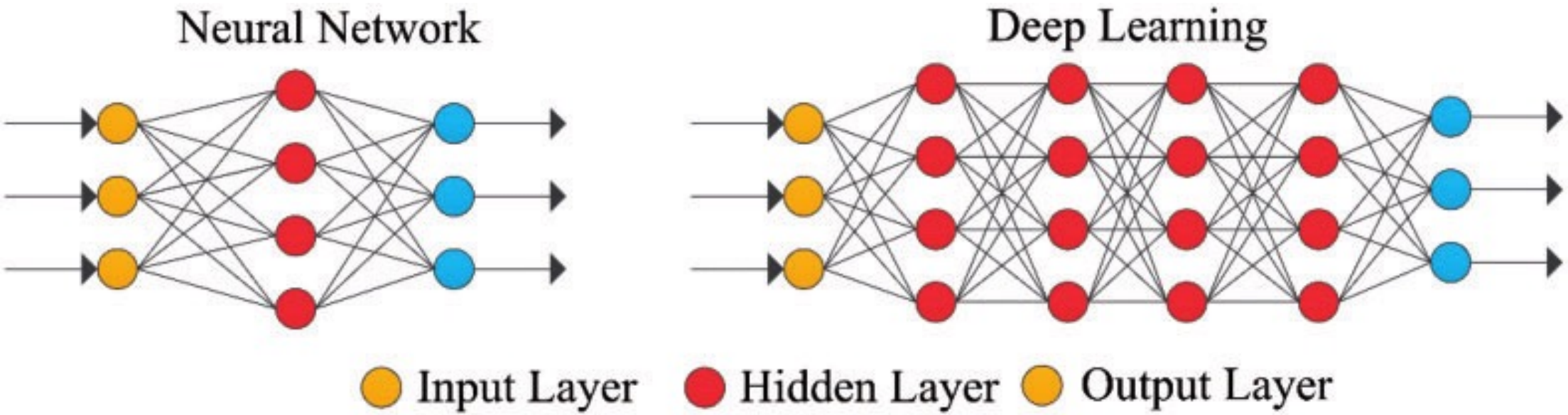


- Fit many trees, each to a different **random sample of the training data**,
- and **randomly sampling** a small set of **features** at each node.
- The final random forest prediction is the **average of the individual tree predictions**.

Neural network



What makes a neural net “deep”?



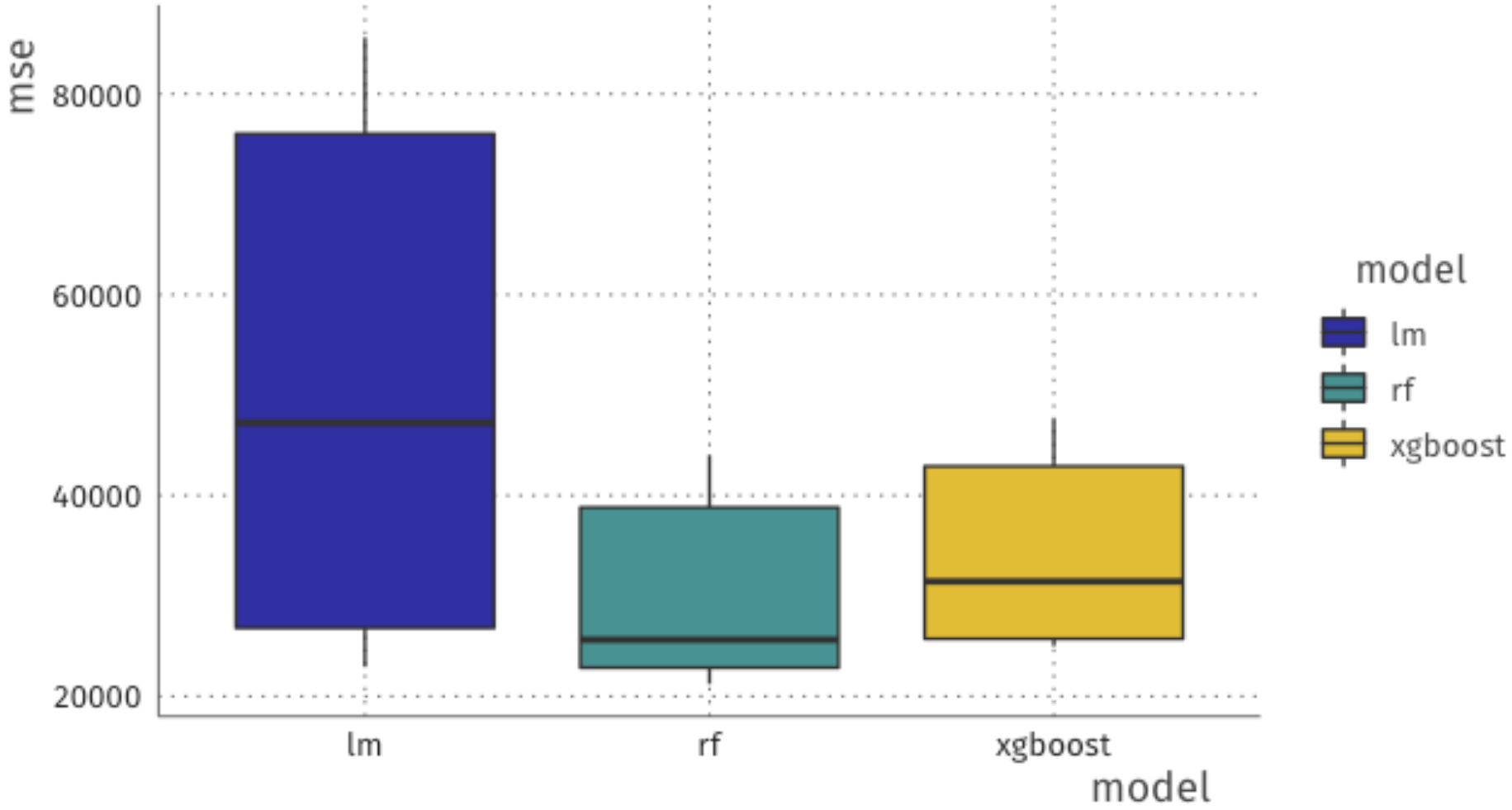
Deep learning

- 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!

Our housing example



Tuning / Hyperparameter optimization

- 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

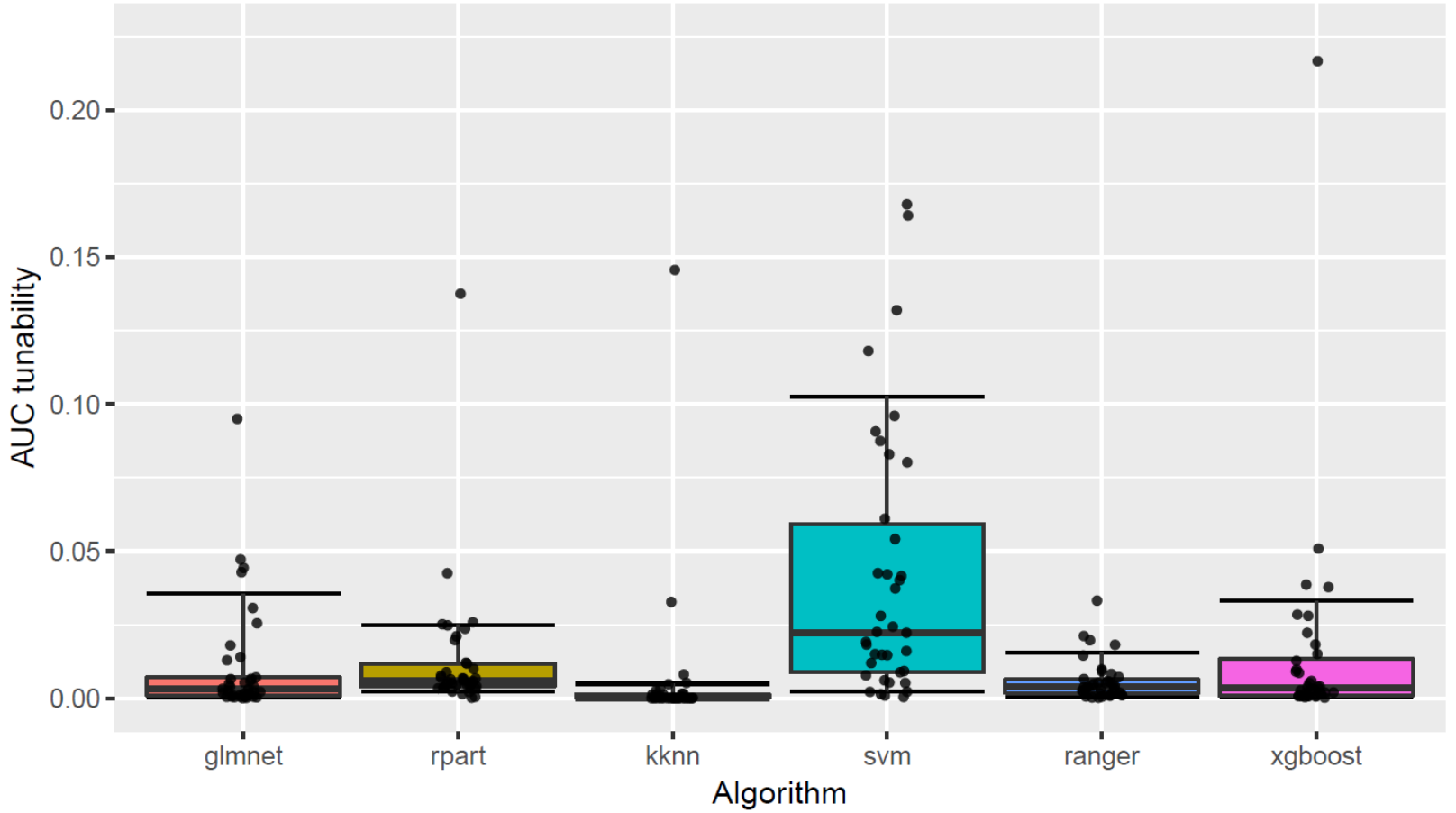
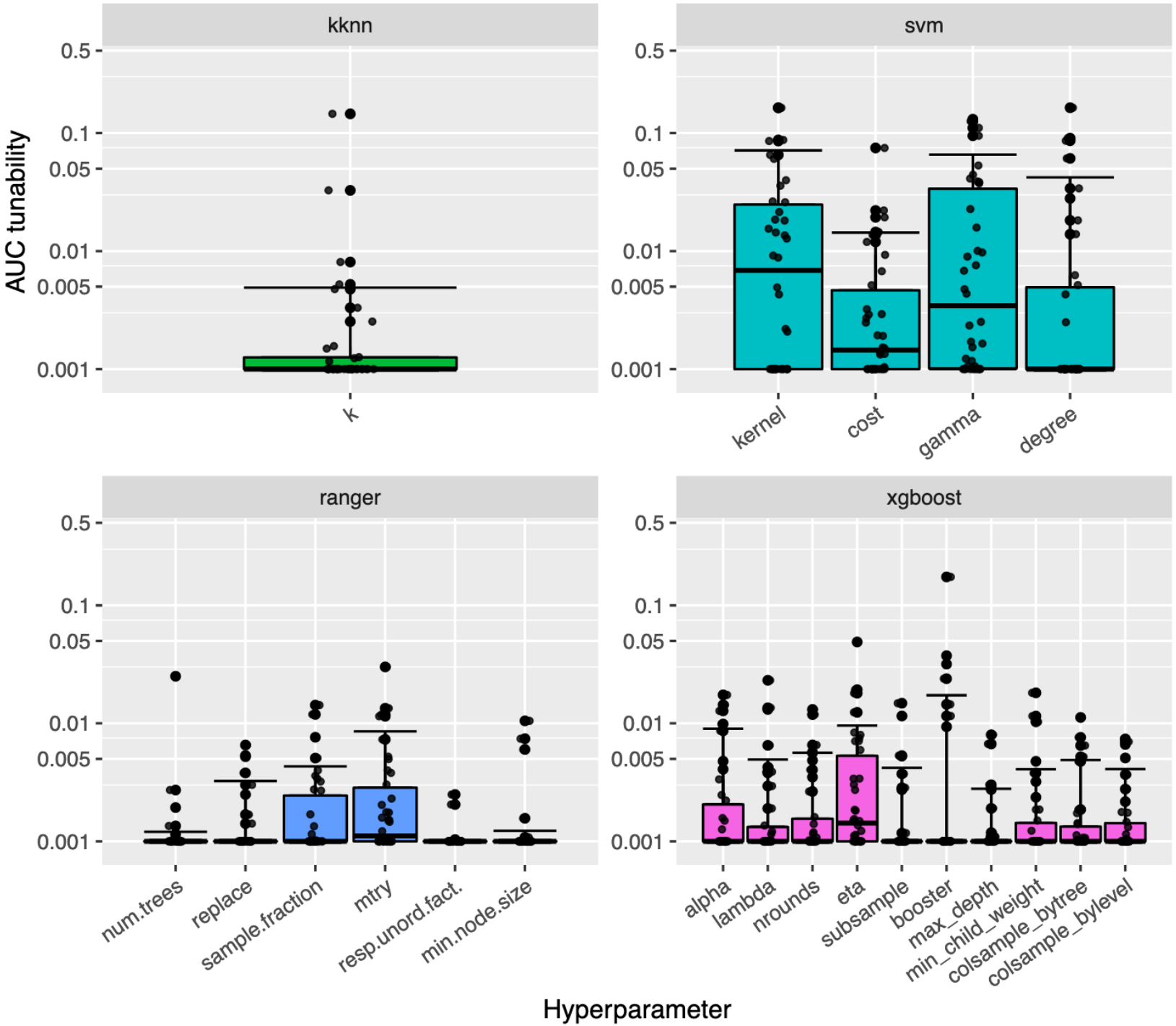


Figure 2: Boxplots of the tunabilities (AUC) of the different algorithms with respect to optimal defaults. The upper and lower whiskers (upper and lower line of the boxplot rectangle) are in our case defined as the 0.1 and 0.9 quantiles of the tunability scores. The 0.9 quantile indicates how much performance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



Probst et al. (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 1

- Choose your **data**, **goal**, and **performance metric** with care
- Bias and variance trade off, in **theory** and **practice**;
- We try to estimate this using **train/dev/test** paradigm;
- Getting good test data is difficult problem;
- **Cross-validation** is a useful alternative to separate dev set;
- Beware that *any* procedure that makes decisions based on the data requires validation!
- **Machine learning is not (only) about models/algorithms. It is just as much about data, the reality of your task, etc.**

Conclusion 2

- 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 lab session on Thursday should help you do this