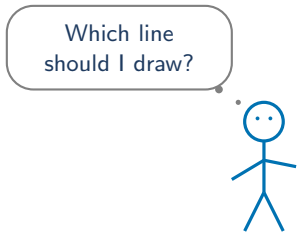
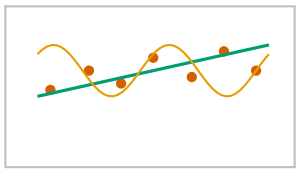




Two lines. Both fit. Which generalizes?

What Should You Be Able to Do After This Lecture?

By the end of this 30-minute mini-lecture, you will be able to:

1. **Analyze** why a model that fits training data perfectly usually fails on new data.
2. **Apply** the bias-variance decomposition to a concrete prediction problem.
3. **Evaluate** whether a given model is underfit, overfit, or appropriately tuned.
4. **Compare** Ridge and Lasso regularization to decide which fits a data situation.

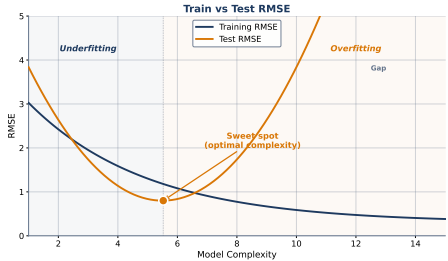
Retrieval: before we start – in one sentence, what does "generalization" mean to you?

What Do You Already Know About Prediction Error?

Before we begin – try to recall:

- What is the difference between “training error” and “test error”?
- What does it mean when a model’s test error is much higher than its training error?
- What is cross-validation used for?

Write one sentence for each before reading on.



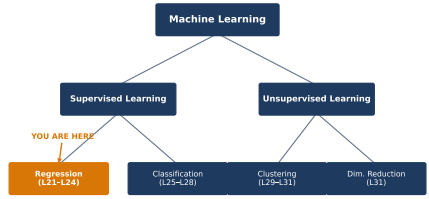
Retrieval: which error do you trust more as a measure of future performance?

Why Does Supervised Learning Dominate Modern AI?

Situation. Most AI you use daily is supervised learning:

- Credit card fraud detection (Visa, Mastercard)
- Email spam filters (Gmail)
- Credit scoring (FICO, alternative lenders)
- Medical imaging (skin cancer, diabetic retinopathy)
- Algorithmic trading signals (hedge funds)

All follow the same recipe:
features in $\rightarrow$ *labels out*.



Retrieval: name one real consequence when a production ML model predicts poorly on new data.

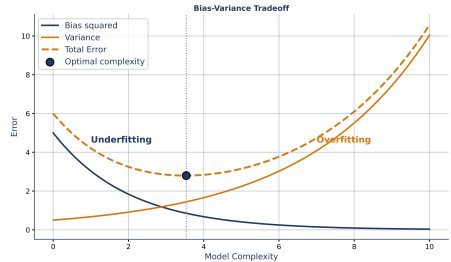
What Is the Generalization Paradox?

Complication. A hedge fund team trains a model on 5 years of historical stock prices. Training accuracy: 99%. In production: 52% – barely better than a coin flip.

The Generalization Paradox

The model that fits your training data best is usually not the one that predicts new data best.

Question. How can "fitting better" lead to "predicting worse"?



Retrieval: in your own words – why is 99% training accuracy with 52% test accuracy a WARNING, not a win?

Pause – 30 Seconds

Pause 30 seconds.

Write one sentence:

"Why might fitting a perfect curve through 10 training points be WORSE than fitting a rough line through those same 10 points?"

(Do not look ahead. Writing beats reading for memory.)

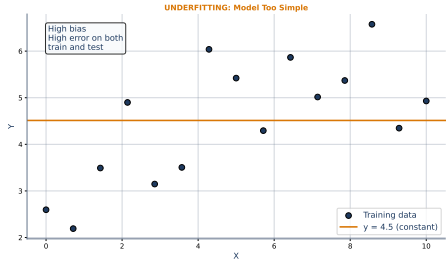
What Actually Is Bias-Variance?

Total expected error on a new point decomposes into three parts:

$$\text{Error} = \text{Bias}^2 + \text{Variance} + \text{Noise}$$

- **Bias.** How wrong on average? (Model is too simple.)
- **Variance.** How unstable between training sets? (Model is too flexible.)
- **Noise.** Irreducible. Not your fault.

The paradox reformulated: lowering bias usually raises variance, and vice versa.



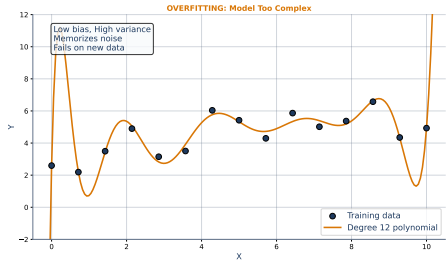
Retrieval: complete this sentence – "If I lower bias I usually raise . . . "

Can We See the Paradox in Stock Returns?

Case: Medallion Fund. Renaissance Technologies' Medallion Fund uses ML to trade equities.

- In-sample: tens of thousands of trades with Sharpe ≈ 3.0
- Out-of-sample ("new data"): only 1/3 of models survive walk-forward validation
- Production: strategies degrade in 6-18 months

Even Renaissance – arguably the world's most profitable quant fund – fights the paradox daily.



ASIDE: Medallion Fund has returned 39% annualized since 1988 – but only to employees; outside investors cannot access it.

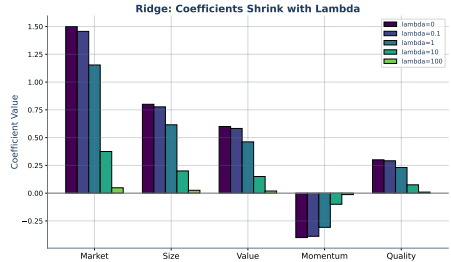
How Do We Escape the Paradox?

Regularization. Add a penalty that restrains model complexity.

- **Ridge (L2).** Shrinks all coefficients toward zero smoothly.
- **Lasso (L1).** Shrinks some coefficients to exactly zero (feature selection).
- **ElasticNet.** A hybrid.

One parameter, λ (lambda), controls how much you restrain the model.

Choose λ by cross-validation, not by eye.



Retrieval: if you suspect only 5 of 100 features truly matter, which regularization would you reach for first?

Pause – 30 Seconds

Pause 30 seconds.

Write your answer:

"You have 100 features and suspect only 5 matter. Ridge or Lasso? Why?"

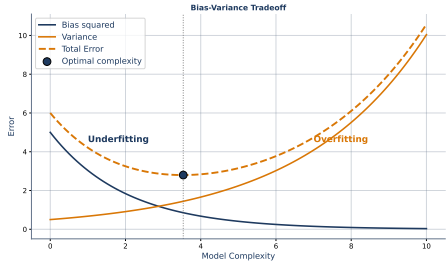
(Commit to one before you read on.)

What Happens When We Ignore the Paradox?

Three common silent killers:

1. **Data leakage.** Feature accidentally contains the label (e.g., "next-day return" in training).
2. **Snooping bias.** Testing many models, picking the best test score. The winner is overfit to the test set.
3. **Lookahead bias.** Using future information at training time (e.g., fiscal-year revisions).

All three make training error artificially low and production failure inevitable.



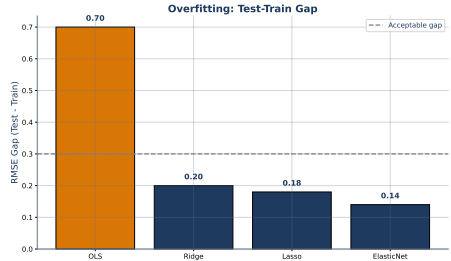
Retrieval: name one concrete form of data leakage you might encounter in finance.

Where Does the Paradox Show Up?

The paradox is universal – but shows up differently:

- **Regression (continuous y).** Overfit curve weaves through training points.
- **Classification (discrete y).** Overfit boundary carves data into tiny cells.
- **Parametric (OLS, logistic).** Bias tends to dominate.
- **Nonparametric (trees, KNN).** Variance tends to dominate.

Every method has a bias-variance story. None escapes.



Retrieval: give one example of a regression task AND one of a classification task from your own life.

Pause – 30 Seconds

Pause 30 seconds.

Write your answer:

"Name one production ML system you've used today. Regression or classification? Parametric or nonparametric? Would the paradox bite?"

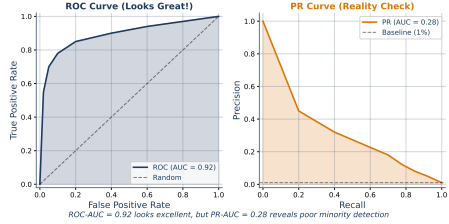
(Yes, even your email spam filter.)

Who Pays When the Paradox Wins?

Real consequences of ignored generalization:

- **Zillow Offers, 2021.** Algorithmic home-flipping lost $881M. Overfit to COVID-era price patterns.
- **Basel III ML models.** Regulators reject ~30% of bank ML credit models for overfitting.
- **Medical diagnostic AI.** FDA warning letters cite "insufficient validation" as leading cause.

The cost of the paradox is paid by customers, shareholders, and patients – not by the modeler.



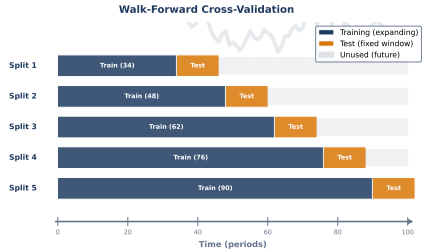
ASIDE: Zillow's Q3 2021 loss alone exceeded the total R&D budget of most AI safety labs that year.

What Should You Remember?

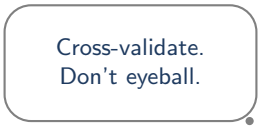
The Generalization Paradox revisited:
The best model is not the one that fits best – it is the one that generalizes best.

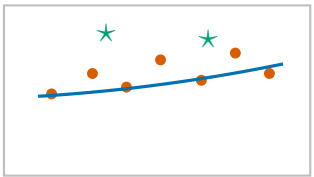
Three habits this lecture should install:

1. Always hold out a test set you never touch until the end.
2. Tune hyperparameters by cross-validation, not training error.
3. Treat training accuracy above 95% as suspicion, not success.



Retrieval: explain the paradox in one sentence WITHOUT using the word "data."





• Training points ⋆ Held-out test

One question to act on: "What is my test set – and have I peeked?"