# Refined Fundamental Trade-Off

- Let $E_{best}$ be the irreducible error (lowest possible error for *any* model).
  - For example, irreducible error for predicting coin flips is 0.5.
- Some learning theory results use $E_{best}$ to futher decompose $E_{test}$:

$$E_{test} = \underbrace{(E_{test} - E_{train})}_{\text{"variance"}} + \underbrace{(E_{train} - E_{best})}_{\text{"bias"}} + \underbrace{E_{best}}_{\text{"noise"}}$$

- This is similar to the bias-variance decomposition:
  - Term 1: measure of variance (how sensitive we are to training data).
  - Term 2: measure of bias (how low can we make the training error).
  - Term 3: measure of noise (how low can any model make test error).

# Refined Fundamental Trade-Off

- Decision tree with high depth:
  - Very likely to fit data well, so bias is low.
  - But model changes a lot if you change the data, so variance is high.
- Decision tree with low depth:
  - Less likely to fit data well, so bias is high.
  - But model doesn't change much you change data, so variance is low.

- And degree does not affect irreducible error.
  - Irreducible error comes from the best possible model.

# Bias-Variance Decomposition

- You may have seen "bias-variance decomposition" in other classes:
  - Assumes $\tilde{y}_i = \bar{y}_i + \varepsilon$, where $\varepsilon$ has mean 0 and variance $\sigma^2$.
  - Assumes we have a "learner" that can take 'n' training examples and use these to make predictions $\hat{y}_i$.
- Expected squared test error in this setting is

$$\underbrace{\mathbb{E}[(\tilde{y}_i - \hat{y}_i)^2]}_{\text{"test squared error"}} = \underbrace{\mathbb{E}[(\hat{y}_i - \bar{y}_i)]^2}_{\text{"bias"}} + \underbrace{(\mathbb{E}[\hat{y}_i^2] - \mathbb{E}[\hat{y}_i]^2)}_{\text{"variance"}} + \underbrace{\sigma^2}_{\text{"noise"}}$$

  - Where expectations are taken over possible training sets of 'n' examples.
  - Bias is expected error due to having wrong model.
  - Variance is expected error due to sensitivity to the training set.
  - Noise (irreducible error) is the best can hope for given the noise ($E_{best}$).

# Bias-Variance vs. Fundamental Trade-Off

- Both decompositions serve the same purpose:
  - Trying to evaluate how different factors affect test error.

- They both lead to the same 3 conclusions:
  1. Simple models can have high $E_{train}$/bias, low $E_{approx}$/variance.
  2. Complex models can have low $E_{train}$/bias, high $E_{approx}$/variance.
  3. As you increase ‘n’, $E_{approx}$/variance goes down (for fixed complexity).

# A Theoretical Answer to “How Much Data?”

- Assume we have a source of IID examples and a fixed class of parametric models.
    - Like “all depth-5 decision trees”.
- Under some nasty assumptions, with ‘n’ training examples it holds that:

  E[test error of best model on training set] – (best test error in class) = O(1/n).

- You rarely know the constant factor, but this gives some guidelines:
  – Adding more data helps more on small datasets than on large datasets.
    - Going from 10 training examples to 20, difference with best possible error gets cut in half.
      – If the best possible error is 15% you might go from 20% to 17.5% (this does **not** mean 20% to 10%).
    - Going from 110 training examples to 120, error only goes down by ~10%.
    - Going from 1M training examples to 1M+10, you won’t notice a change.

  – Doubling the data size cuts the error in half:
    - Going from 1M training to 2M training examples, error gets cut in half.
    - If you double the data size and your test error doesn’t improve, more data might not help.