**PSYCH 201B**

*Statistical Intuitions for Social Scientists*

# Hypothesis testing as Model Comparison

01/21/2026

# Today's Plan

1. Recap

2. Bias-Variance Trade-off

3. Hypothesis testing as Model Comparison

# Last time…

# What is a model?


**"A model is a logical story expressed in *math* (code)"**

~ Andrew Gelman

*one of the most famous statisticians; inventor of Stan; Prof at Columbia*

# What is a model?

A **theory** of how **observed data** were **generated**

$$\texttt{Data} = \texttt{Model} + \texttt{Error}$$

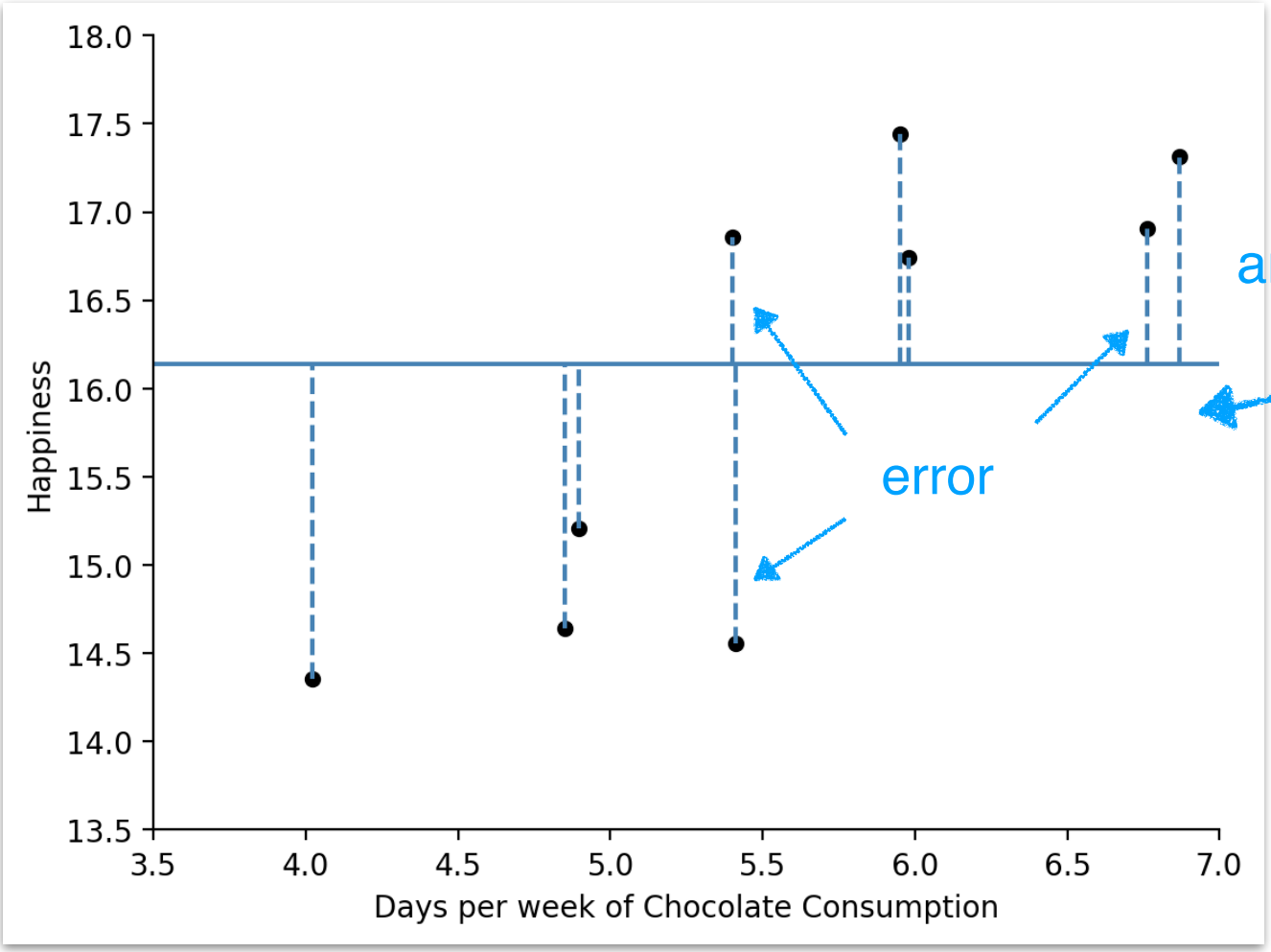
how shall we define this?

**Residual**: the part that's left over after we have used the model to predict/explain the data

$$\text{Error} = \text{Data} - \text{Model}$$

*typically **SSE***
*(sum-of-squared-errors)*

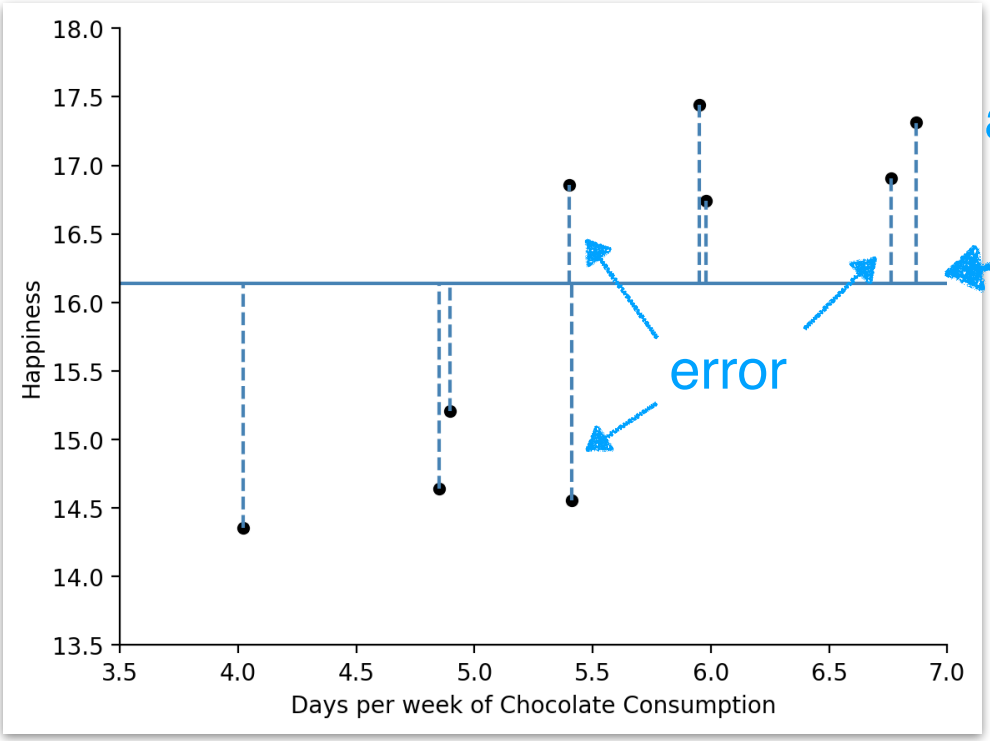
# Error = Data - Model




why squared?

- positive and negative prediction errors don't cancel out
- larger errors are weighted more

# Error = Data - Model




if we **average** them
(divide sum by N)
we get **variance** aka **MSE**

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N} = \frac{\text{SSE}}{N} = MSE$$

# Quick aside on math notation

this is how we write a **for loop** in math

```
for i in range(n):
for obs in data:
```

*(more pythonic)*

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$$

# Quick aside on math notation

```
for obs in data:
```

*(more pythonic)*

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$$

this is **body** of the loop

```
results.append(obs - mean(data)**2)
```

# Quick aside on math notation

this tells us what to do with the results = **sum** them

```
for obs in data:
```

*(more pythonic)*

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$$

```
    results.append(obs - mean(data)**2)

sum(results)
```

# Quick aside on math notation

```
for obs in data:
```

*(more pythonic)*

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$$

divide by the number of observations

```
results.append(obs - mean(data)**2)
```

```
sum(results) / len(data)
```

# Quick aside on math notation

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$$

```
results = []
for obs in data:
    results.append(obs - mean(data)**2)
variance = sum(results) / len(data)
```

# Questions?

# Last time we said…

The mean is the best 1 parameter model

…assuming squared-errors

The median is the best 1 parameter model

…assuming absolute-errors

## lets formalize this…

# The 4 fundamental intuitions

- Aggregation
  - Gives us a model
- **Sampling**
  - Tells us where it applies
- **Uncertainty**
  - Keeps us honest
- **Learning**
  - Forces iteration

# We aggregate by choosing estimators —which are defined by the losses they minimize

## Estimator

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{Median} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd} \\ \dfrac{x_{n/2} + x_{n/2+1}}{2}, & \text{if } n \text{ is even} \end{cases}$$

no other single summary statistic will make this smaller!

## Loss-function

$$\sum_{i=1}^{n} (x_i - \mu)^2 = SSE$$

$$\sum_{i=1}^{n} |x_i - \mu| = SAE$$

# Remember the big ideas?

## 1) Law of Large Numbers

As we average more independent observations our estimator stabilizes

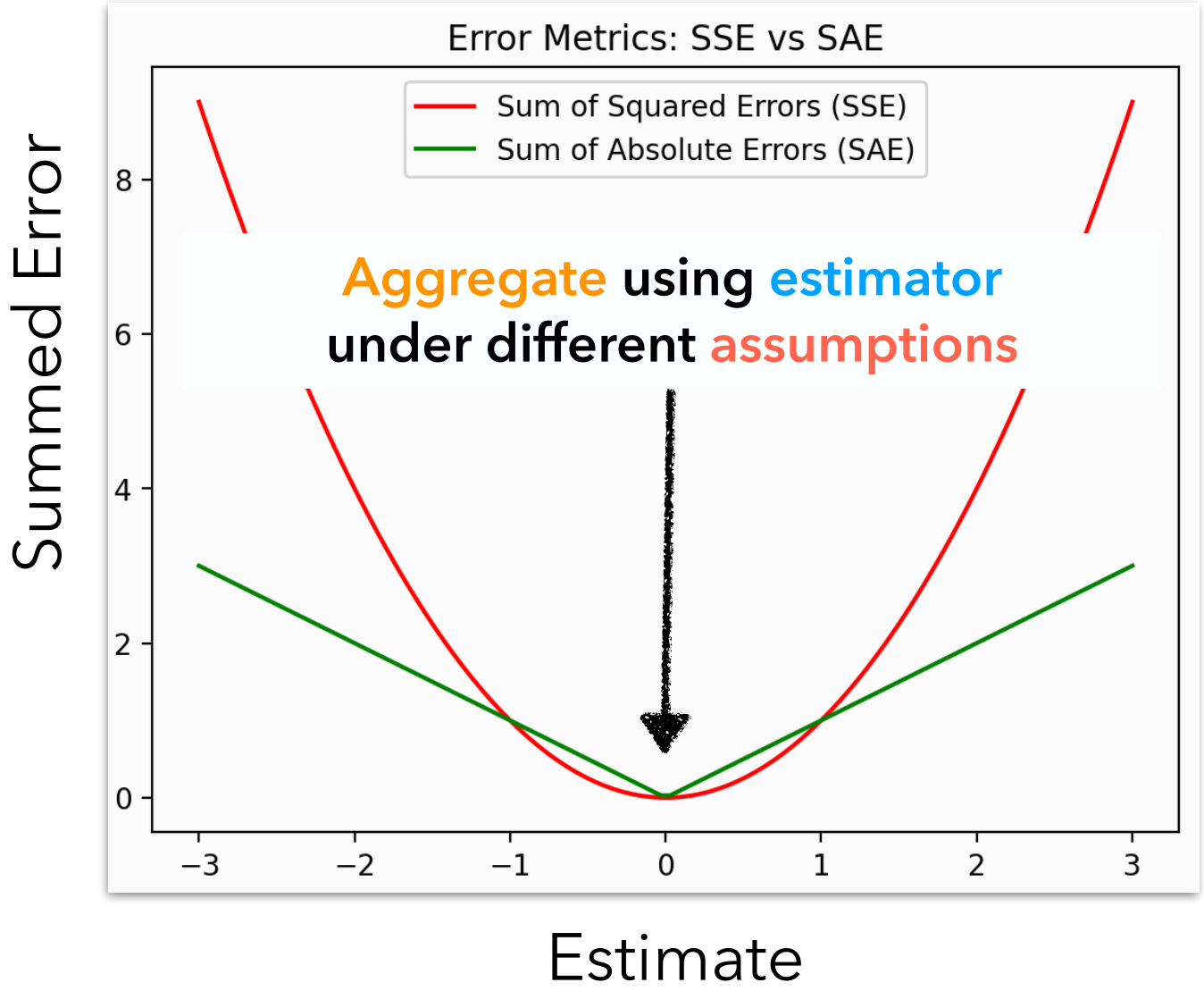
## 2) Central Limit Theorem

Distribution of estimator converges to normal distribution even if data distribution are not normal

## 3) No Free Lunch Theorem

Any estimator that performs well somewhere — must perform poorly elsewhere

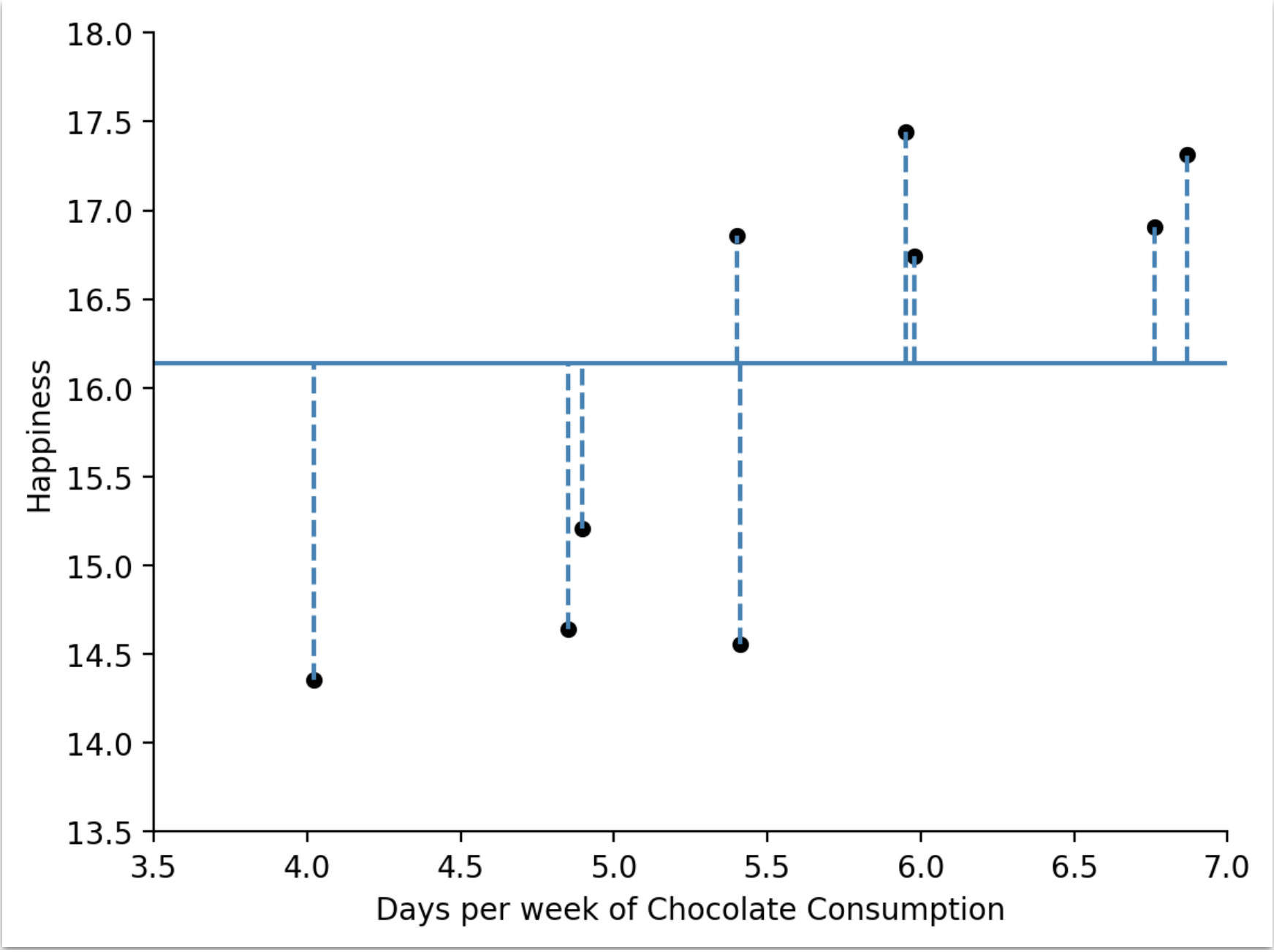
# We aggregate by choosing estimators —which are defined by the losses the minimize



# Can we do *better*?

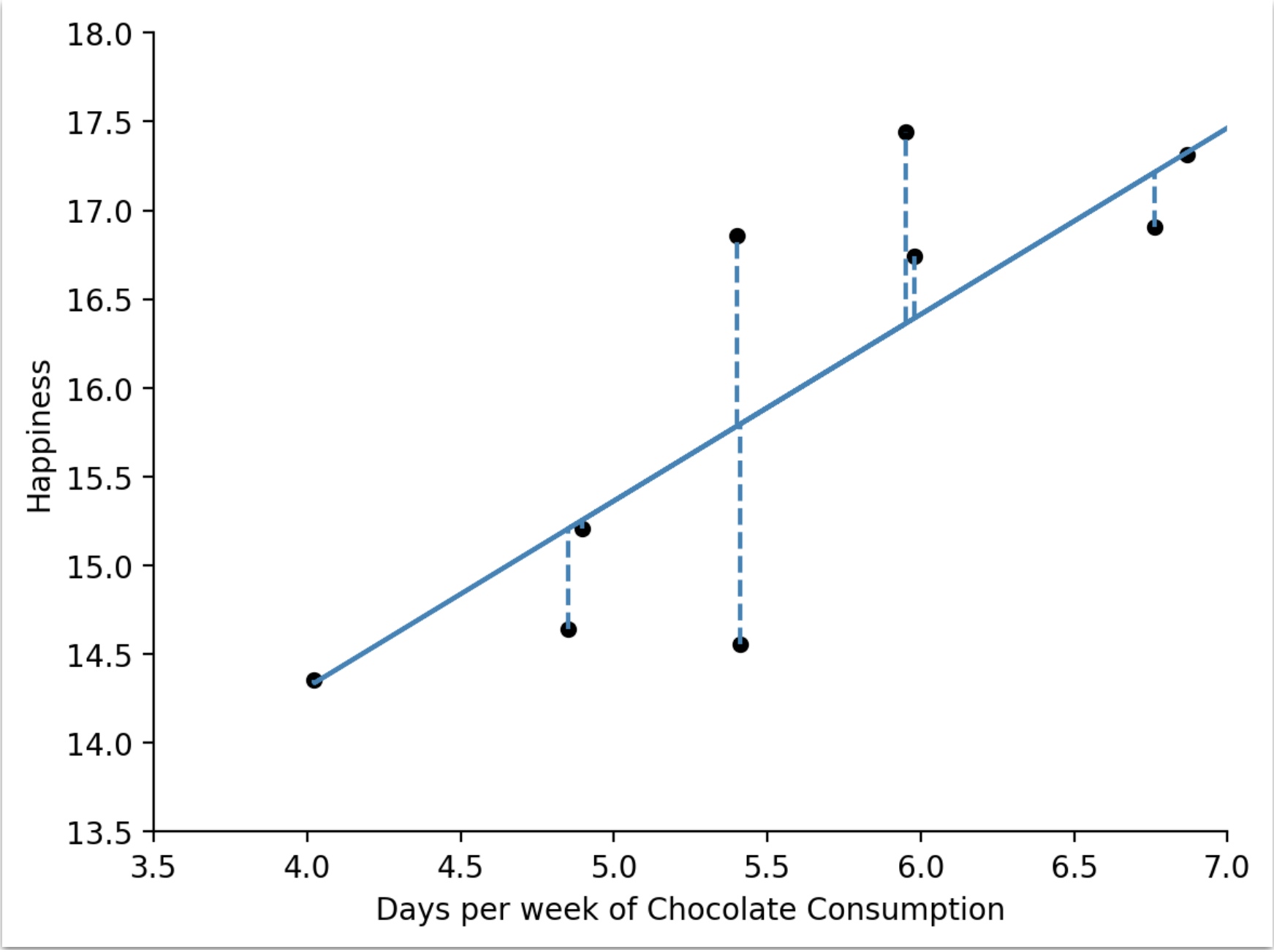
# The 1 parameter model of happiness

(mean)



# The 2 parameter model of happiness

(mean + change in chocolate consumption)



# Adding Model Parameters

$$\texttt{Data} = \texttt{Model} + \texttt{Error}$$

$\text{happiness}_{\text{prediction}} = \text{mean}_{\text{happiness}} + \text{error}$

$$Y_i = \beta_0 + \epsilon_i$$

the model has a single parameter

$\text{happiness}_{\text{prediction}} = \text{mean}_{\text{happiness}} + \text{slope}_{\text{chocolate}} + \text{error}$

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

the model is a linear combination of parameters

# Adding Model Parameters

$$\texttt{Data} = \texttt{Model} + \texttt{Error}$$

$\texttt{happiness}_{\texttt{prediction}} = \texttt{mean}_{\texttt{happiness}} + \texttt{error}$

$$Y_i = \beta_0 + \epsilon_i$$

but how do we know it's worth it?

$\texttt{happiness}_{\texttt{prediction}} = \texttt{mean}_{\texttt{happiness}} + \texttt{slope}_{\texttt{chocolate}} + \texttt{error}$

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

# Hypothesis testing as model comparison

- We build a model with parameters and estimate their values by minimizing error given data
- Adding additional parameters will **always improve model fit** (reduce error further)
- So there's a fundamental trade-off between complexity and accuracy = worth it?

$$Error = Data - Model$$

To reduce error we can:

improve this (201a)

improve this… but what criteria?

# Hypothesis testing as model comparison

```
Error = Data - Model
```

We can actually decompose this...

A "hat" means estimated not measured

$$Error = bias + variance + \epsilon$$

$$\mathbb{E}(\hat{y}) - y$$

average model prediction

true data

# The Bias-Variance Tradeoff

- **Bias**
  - Error from simplifying assumptions
  - *How systematic are mis-predictions?*

$$Error = bias + variance + \epsilon$$

$$\mathbb{E}(\hat{y}) - y$$

average model prediction

true data

# The Bias-Variance Tradeoff

- **Variance**
  - Error from data sensitivity
  - *How much do predictions bounce around?*

$$Error = bias + variance + \epsilon$$

$$\mathbb{E}(\hat{y}) - y \qquad \mathbb{E}[\hat{y} - \mathbb{E}(\hat{y})^2]$$

single prediction

average model prediction

# The Bias-Variance Tradeoff

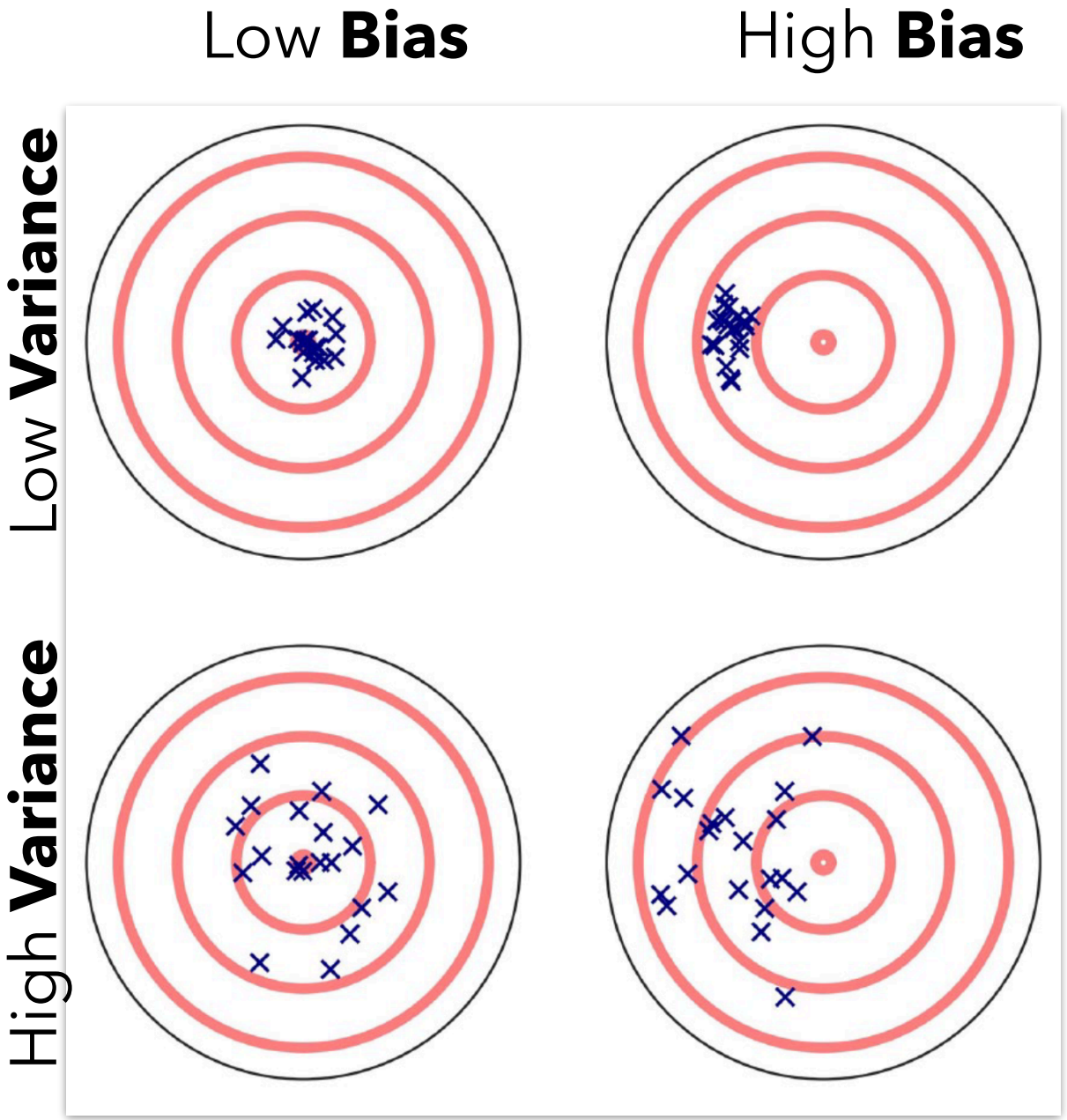
- **Bias**
  - Error from simplifying assumptions
  - How much a predictions are wrong *on average*

$$\text{Bias}(\hat{Y}) = E(\hat{Y}) - Y$$

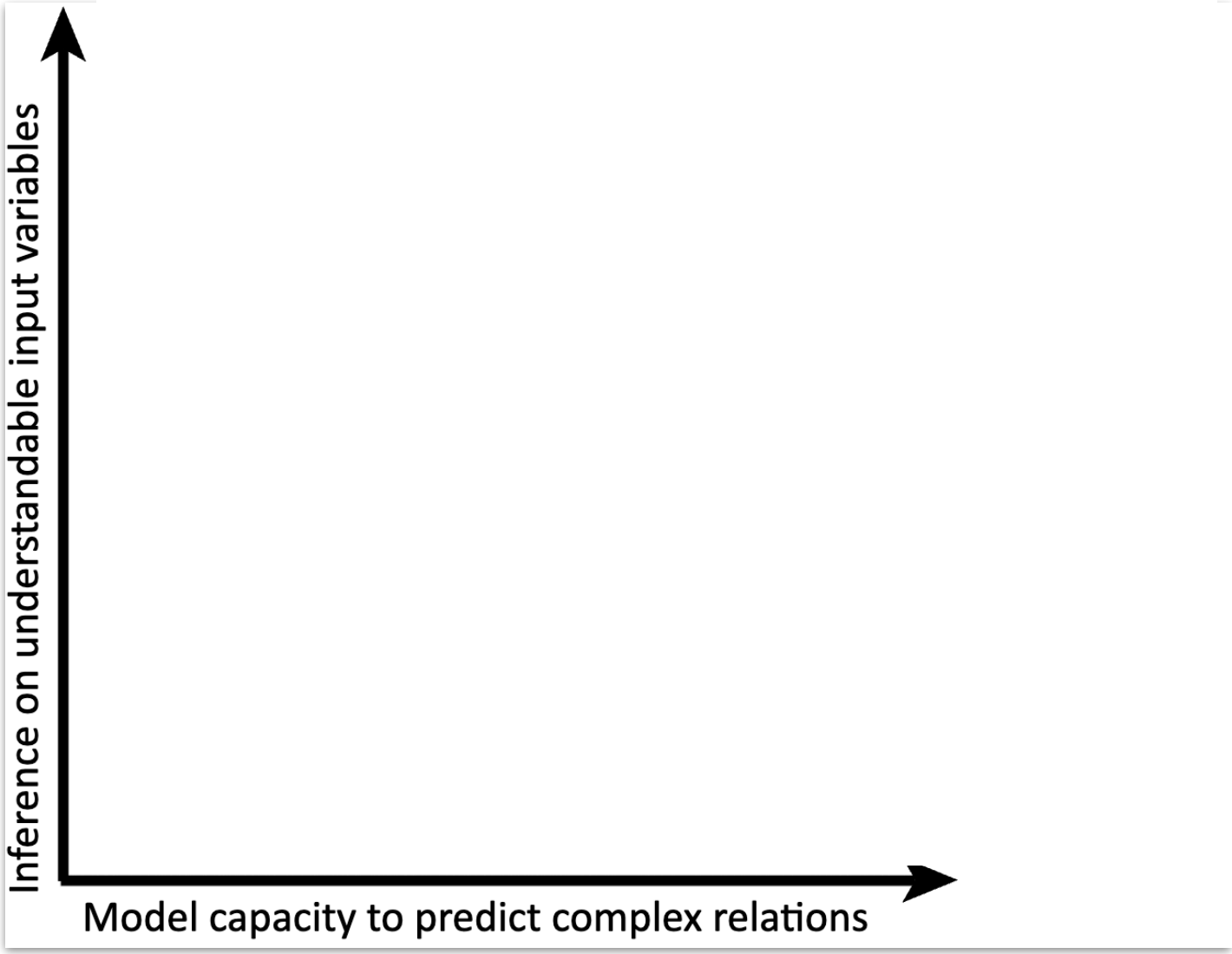
- **Variance**
  - Error from data sensitivity
  - How much predictions change across samples

$$\text{Variance} = E[(\hat{Y} - E[\hat{Y}]^2]$$

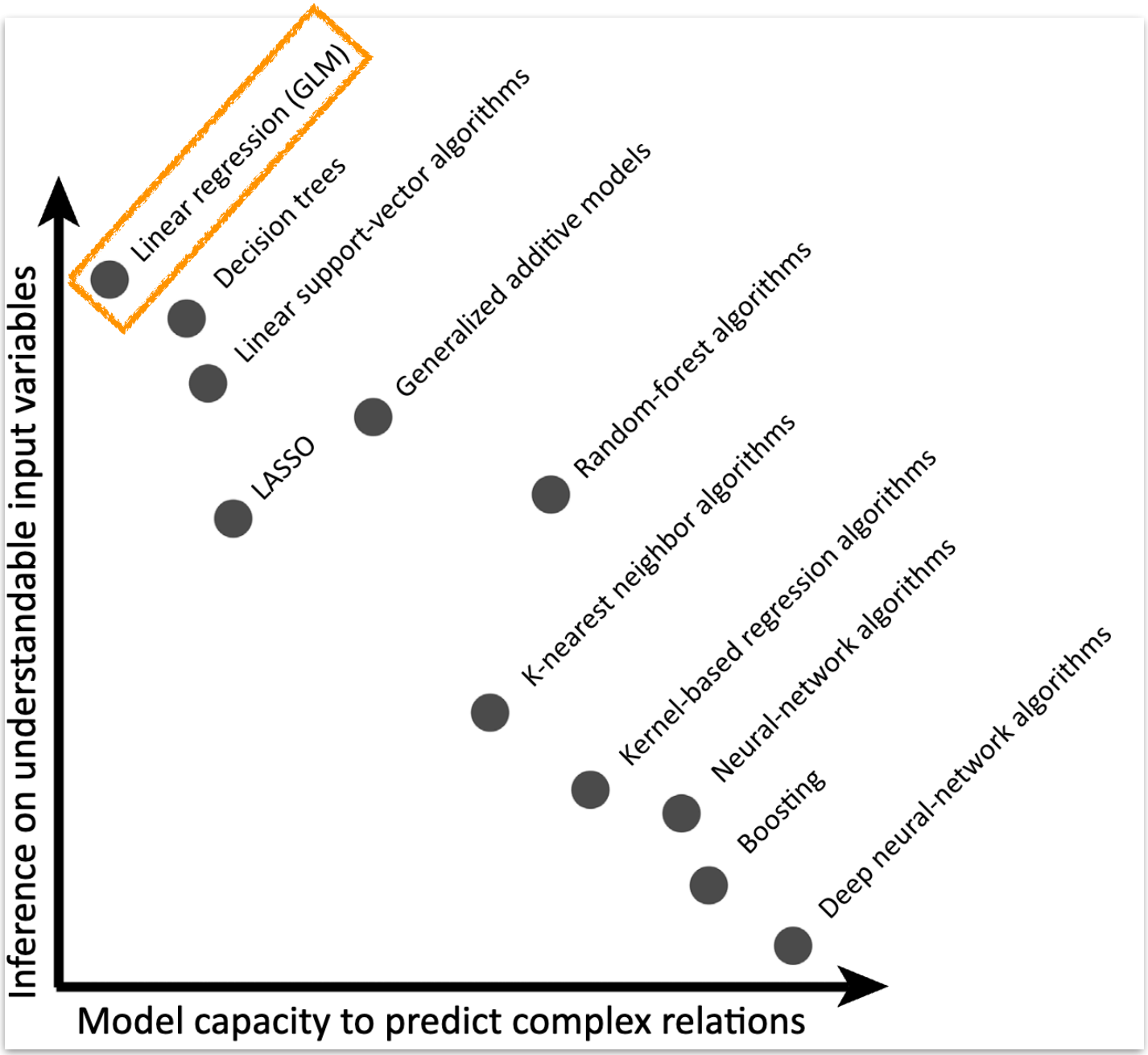
# The Bias-Variance Tradeoff



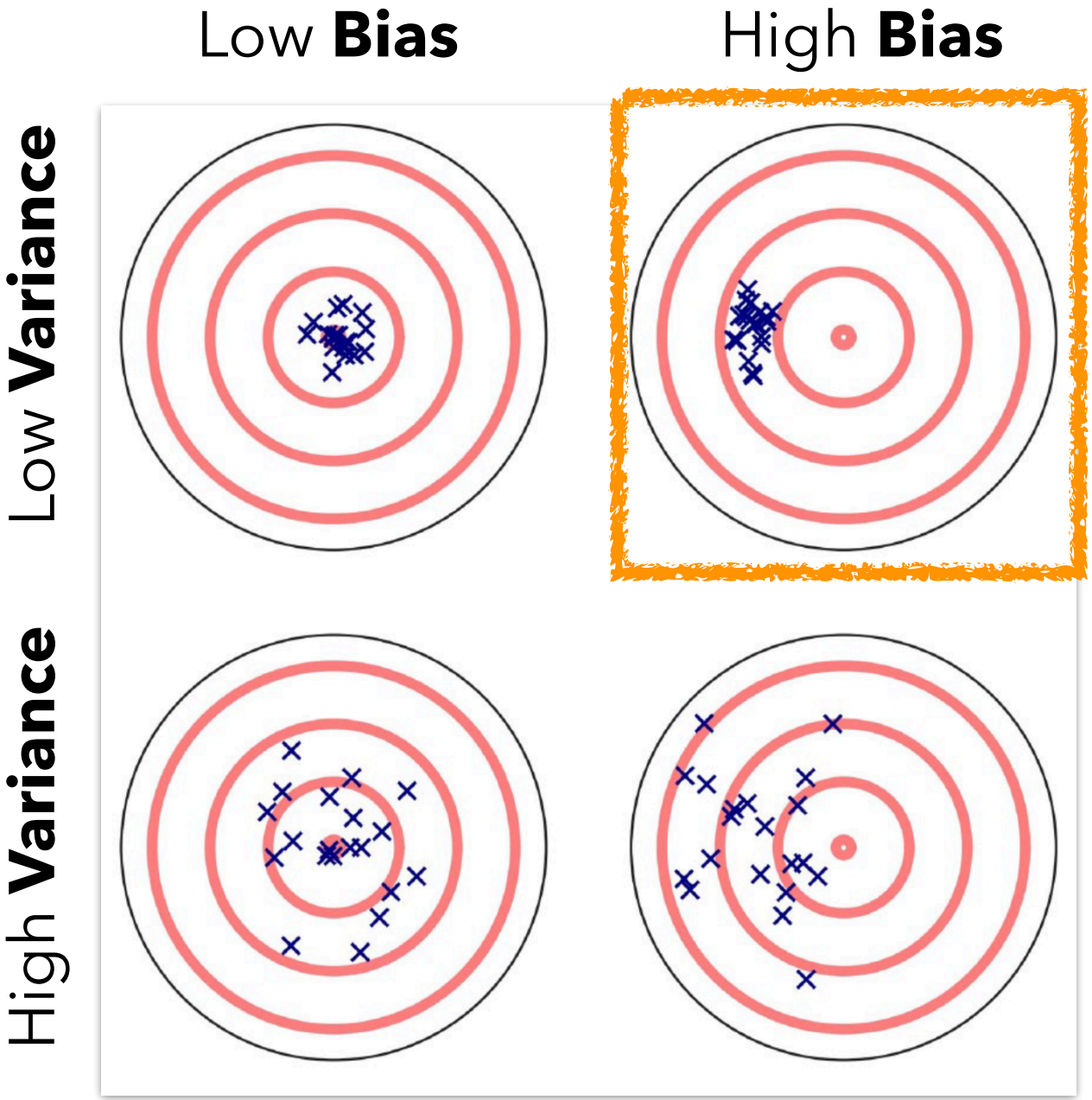
# The Bias-Variance Tradeoff



# The Bias-Variance Tradeoff

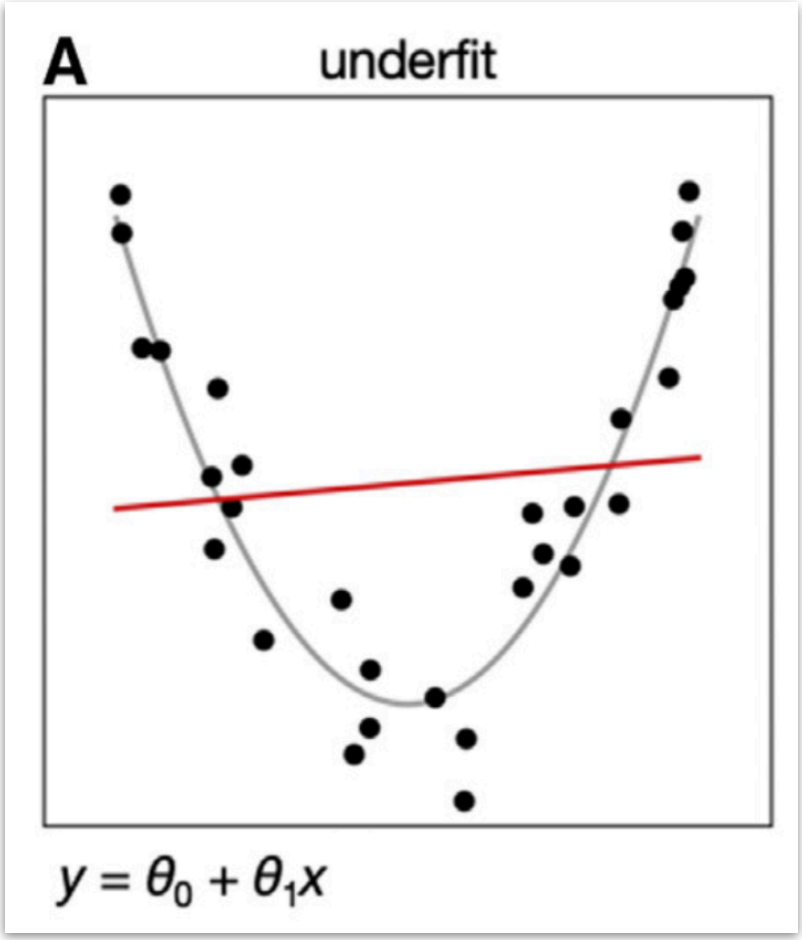


# The Bias-Variance Tradeoff

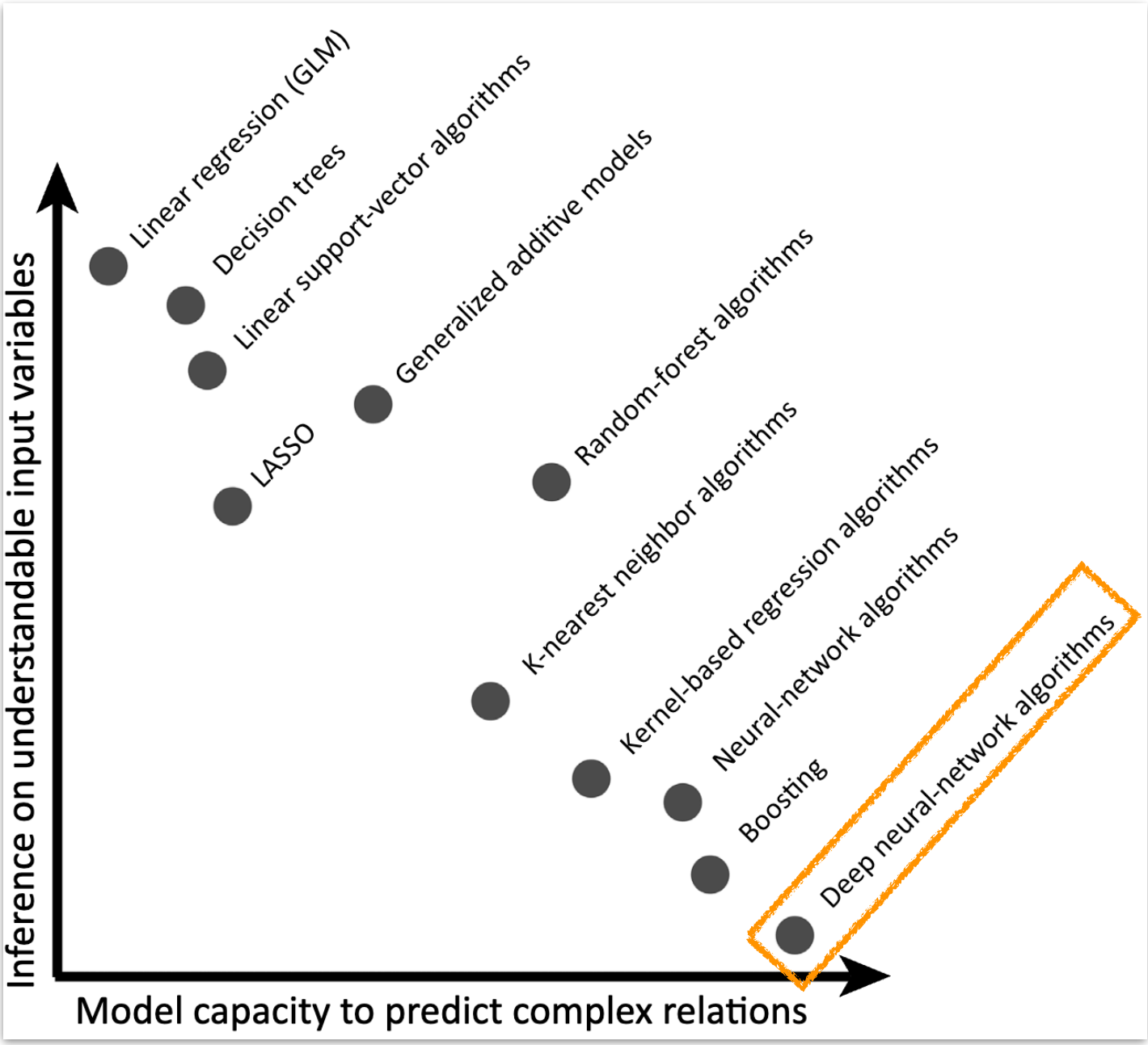


# The Bias-Variance Tradeoff

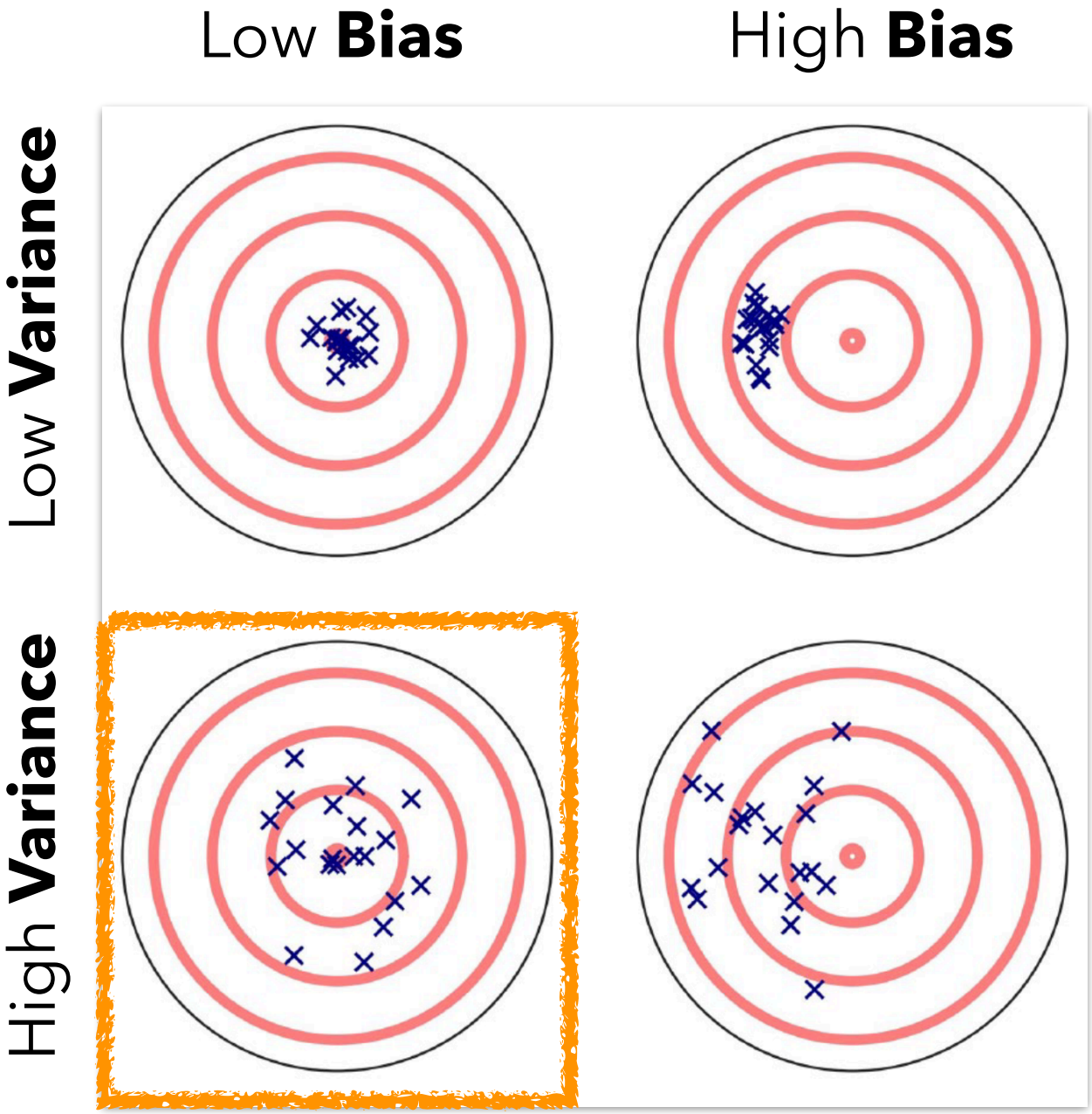
- **High Bias - Low Variance**
  - Behaves very **consistently**
  - By **under-fitting** the data (generalizes poorly to new data)



# The Bias-Variance Tradeoff

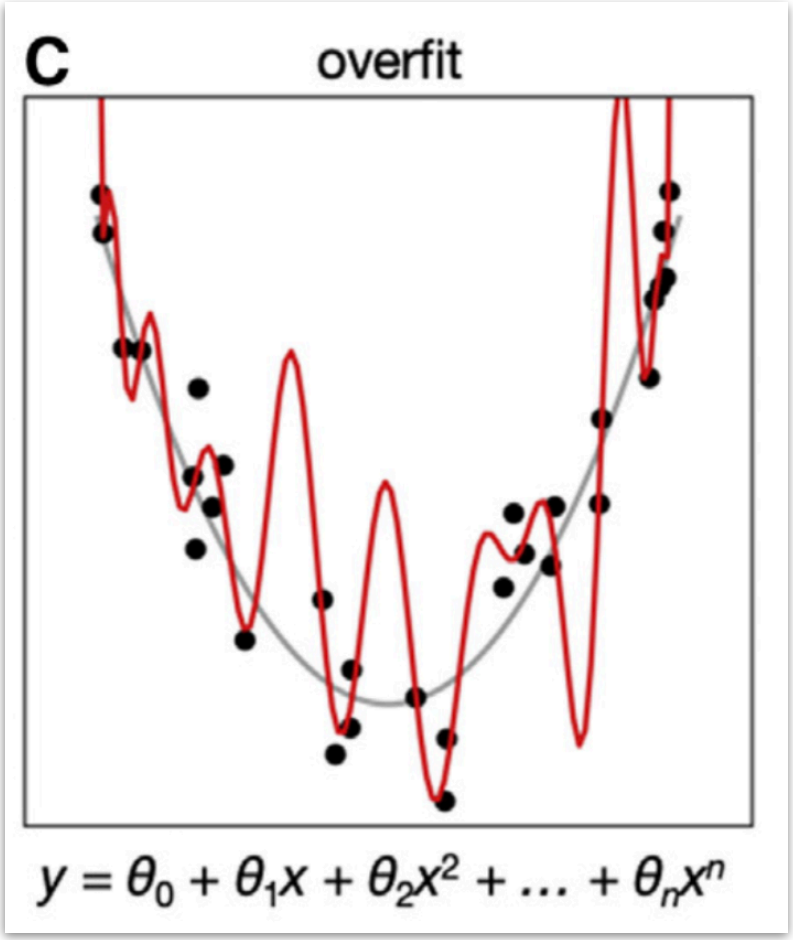


# The Bias-Variance Tradeoff

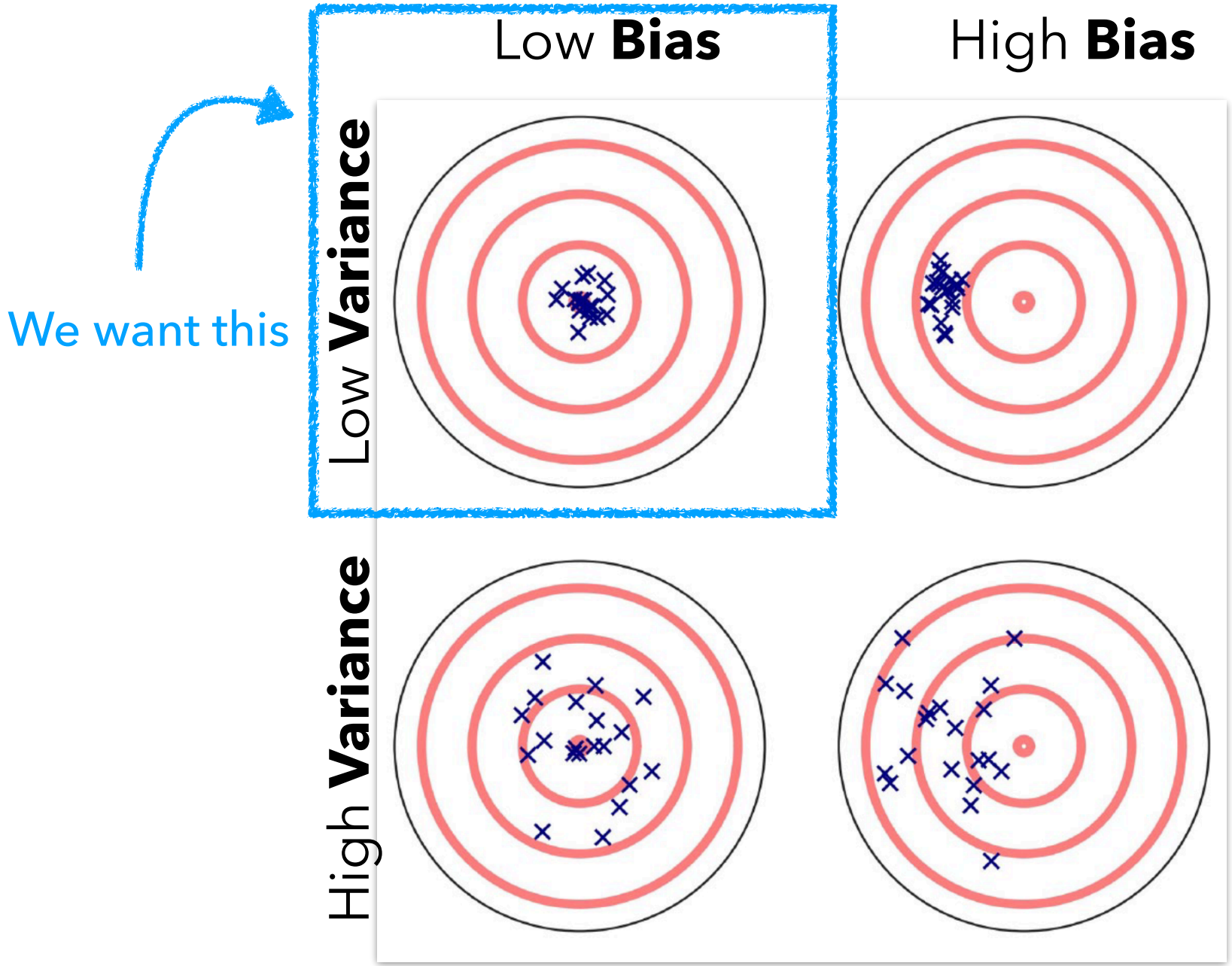


# The Bias-Variance Tradeoff

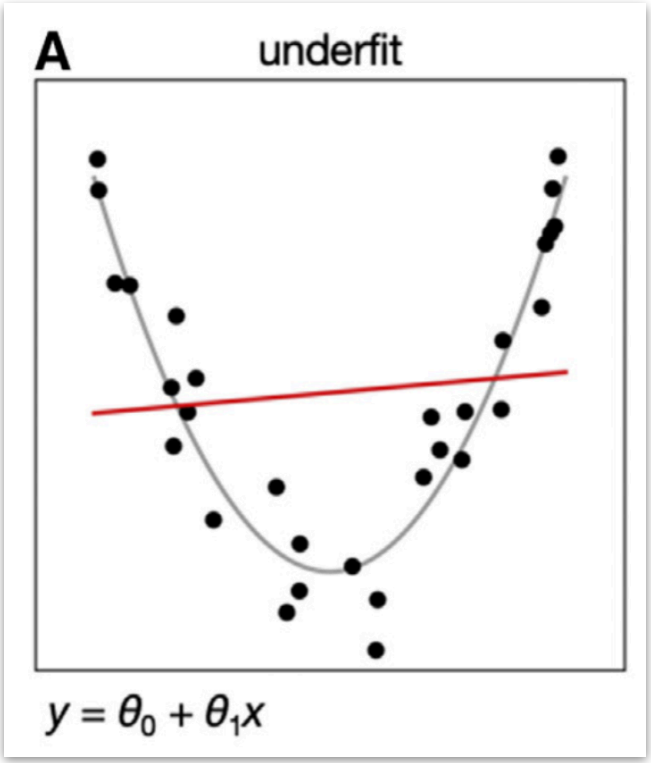
- **Low Bias - High Variance**
  - Over-fitting data (memorize real patterns & noise)
  - Highly variable predictions (too sensitive to data)

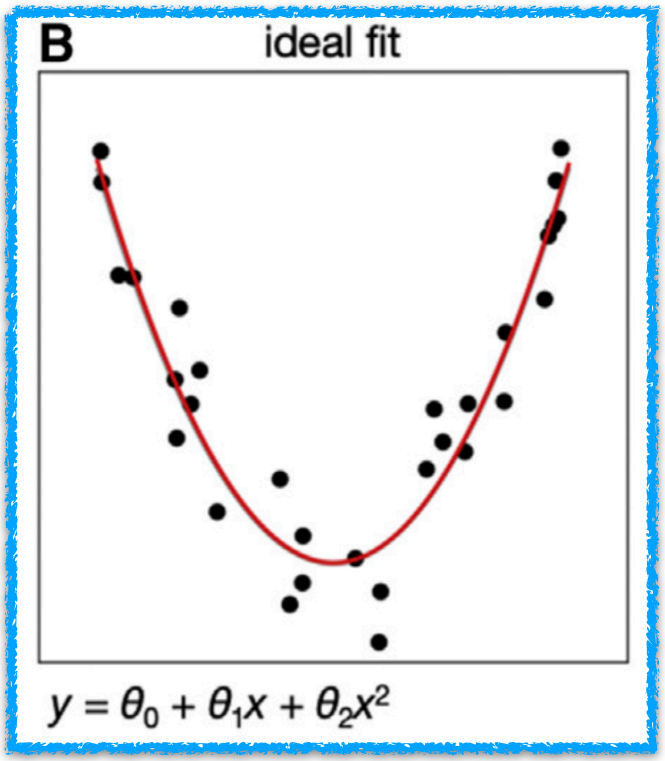


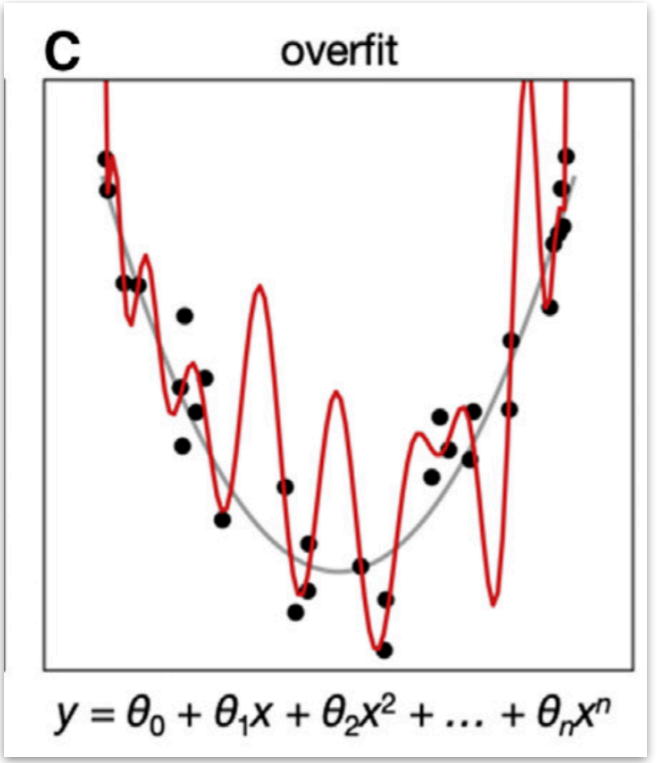
# The Bias-Variance Tradeoff



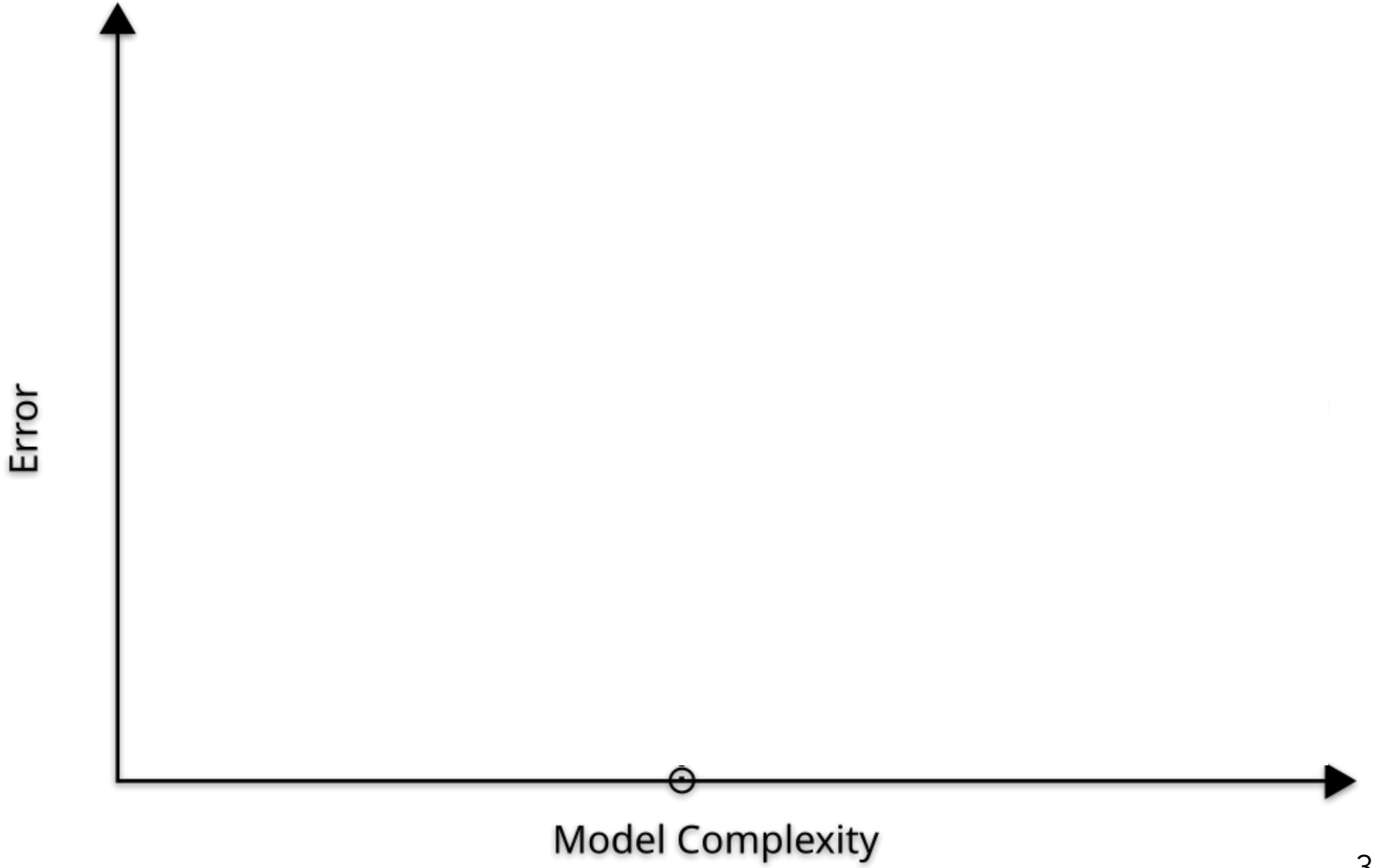
# The Bias-Variance Tradeoff



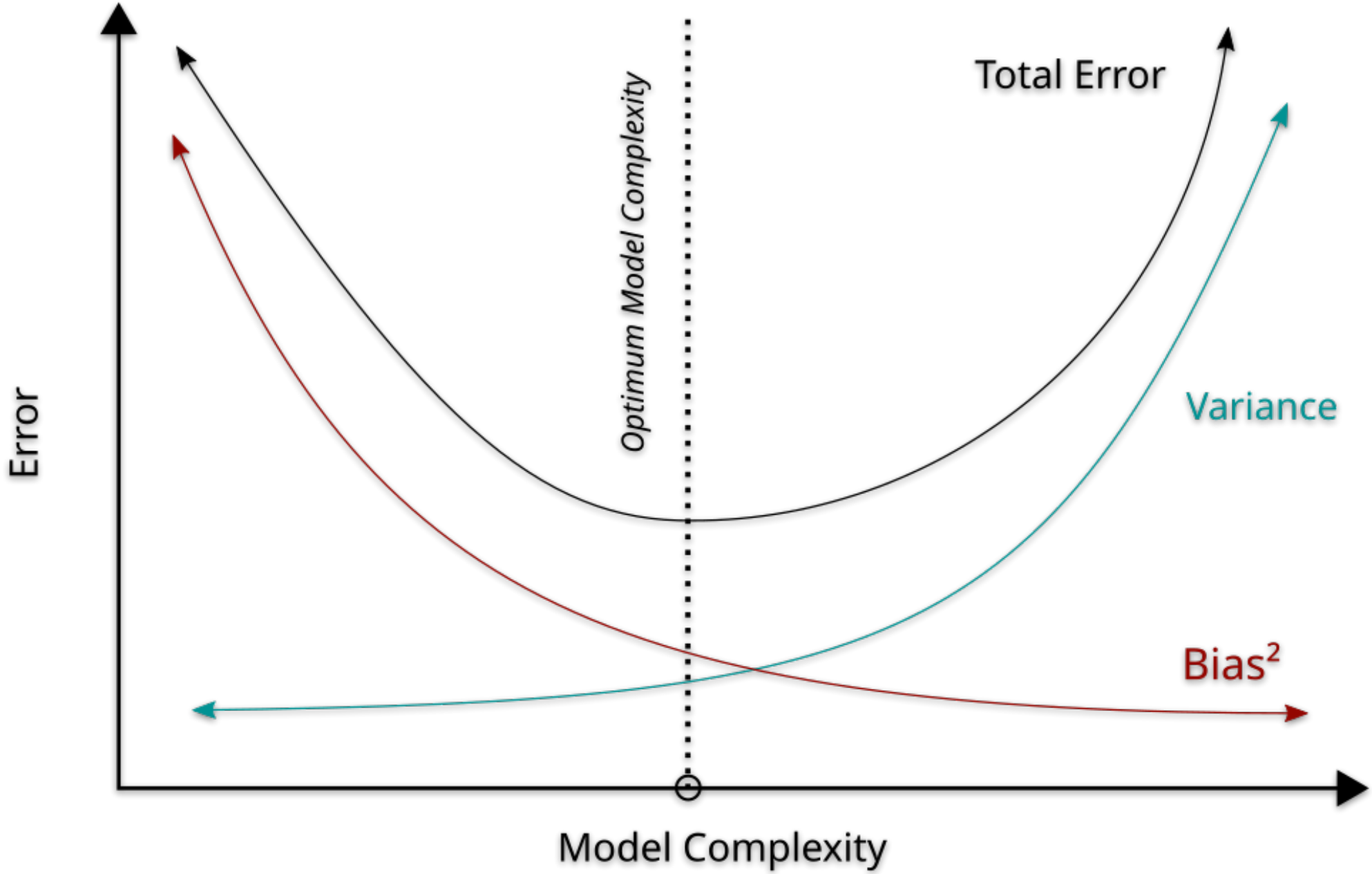



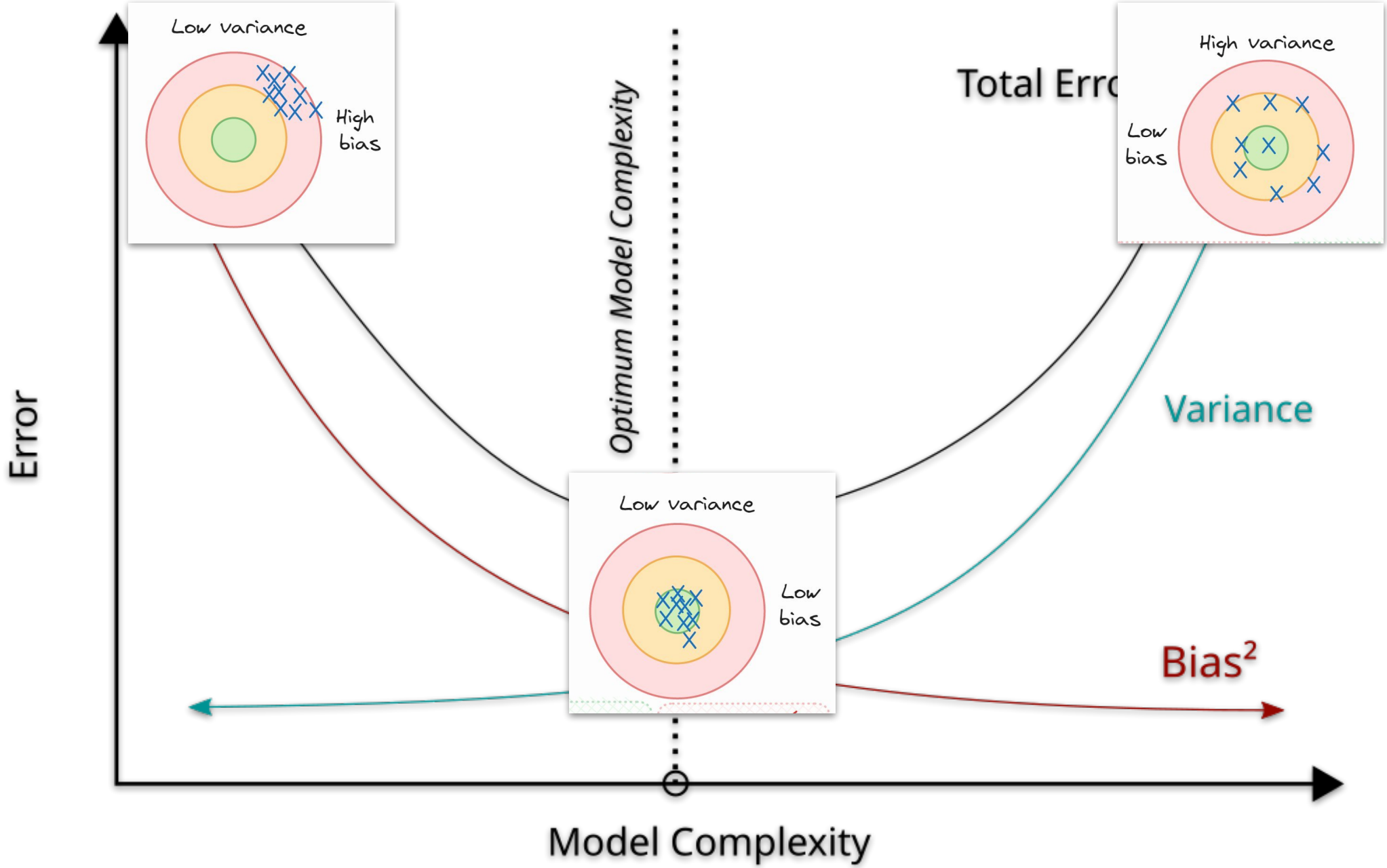
# The Bias-Variance Tradeoff



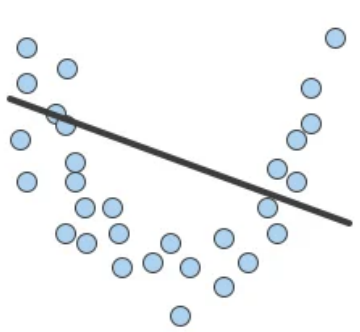
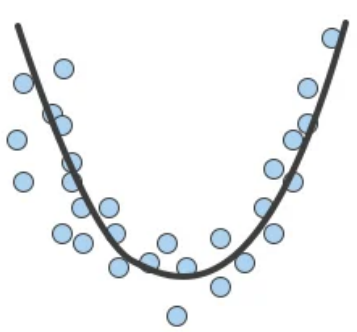
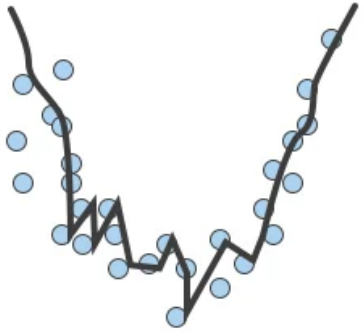
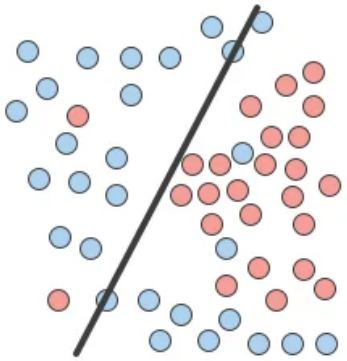
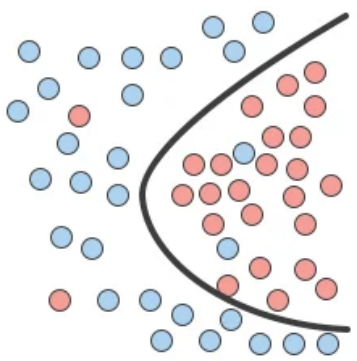
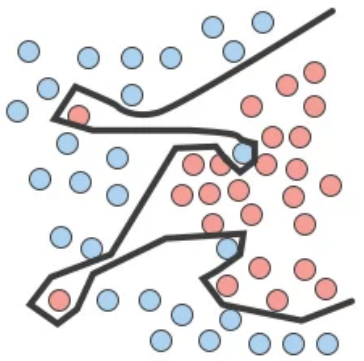
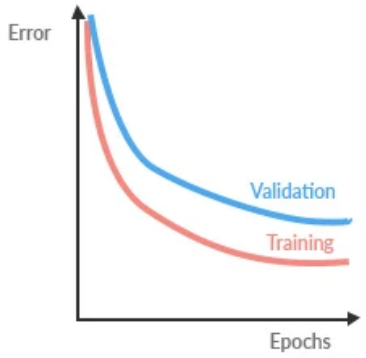
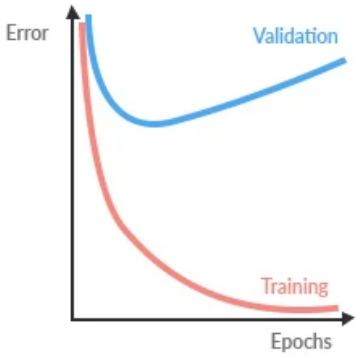
# The Bias-Variance Tradeoff



# The Bias-Variance Tradeoff



# The Bias-Variance Tradeoff

| | Underfitting | Just right | Overfitting |
|---|---|---|---|
| **Symptoms** | • High training error<br>• Training error close to test error<br>• High bias | • Training error slightly lower than test error | • Very low training error<br>• Training error much lower than test error<br>• High variance |
| **Regression illustration** |  |  |  |
| **Classification illustration** |  |  |  |
| **Deep learning illustration** |  |  |  |

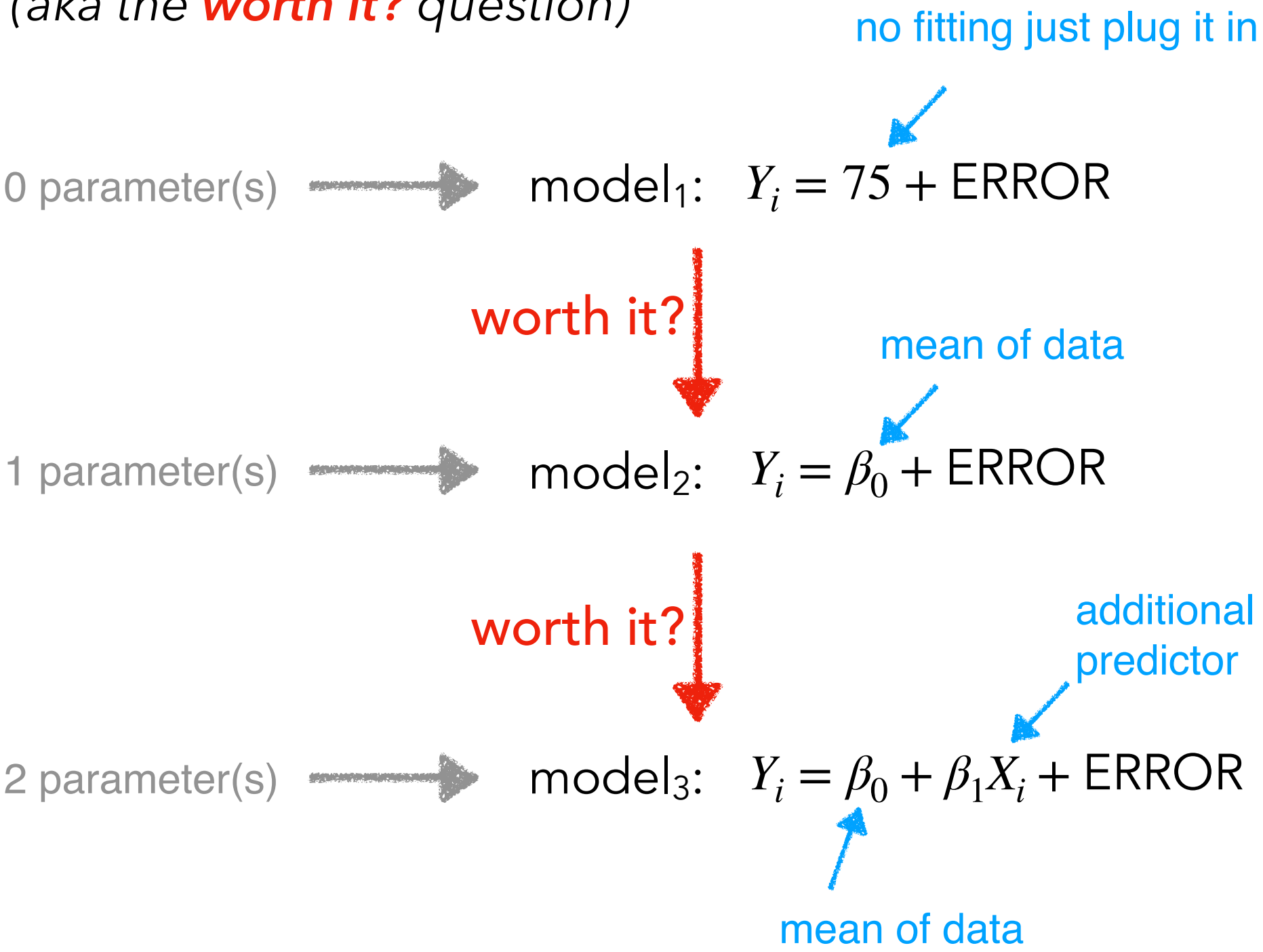
# The Bias-Variance Tradeoff: Summary

- We can separately think about the error of an estimator along two dimensions:
  - **Bias**: simplifying assumptions, under-fitting
  - **Variance**: memorizing signal + noise, over-fitting
- So there's a fundamental trade-off: **is adding additional complexity (more parameters)** worth it?
- This gives us a framework for **hypothesis testing:**

**Model Comparison**

# Hypothesis testing as model comparison

*(aka the **worth it?** question)*



# Hypothesis testing as model comparison

*(aka the **worth it?** question)*

1 parameter(s)

model$_2$: $Y_i = \beta_0 + \text{ERROR}$

2 parameter(s)

model$_3$: $Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

# Hypothesis testing as model comparison

*(aka the **worth it?** question)*

## Compact Model

1 parameter(s)

$\text{model}_C$: $Y_i = \beta_0 + \text{ERROR}$

## Augmented Model

2 parameter(s)

$\text{model}_A$: $Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

**PRE**

Proportional Reduction in Error

$$= \frac{\text{ERROR}(C) - \text{ERROR}(A)}{\text{ERROR}(C)}$$

$$= 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)}$$

# Hypothesis testing as model comparison

*(aka the **worth it?** question)*

## Compact Model

1 parameter(s)

model$_C$: $Y_i = \beta_0 + \text{ERROR}$

$$\text{ERROR}(C) = 50$$

## Augmented Model

2 parameter(s)

model$_A$: $Y_i = \beta_0 + \beta_1 X_i + \text{ERROR}$

$$\text{ERROR}(A) = 30$$

Increasing the complexity of the model by 1 parameter reduced the error by 40%.

$$= 1 - \frac{\text{ERROR}(A)}{\text{ERROR}(C)}$$

***But how do we know if this is good?***

$$= 1 - \frac{30}{50} = .40$$

# Hypothesis testing as model comparison

*(aka the **worth it?** question)*

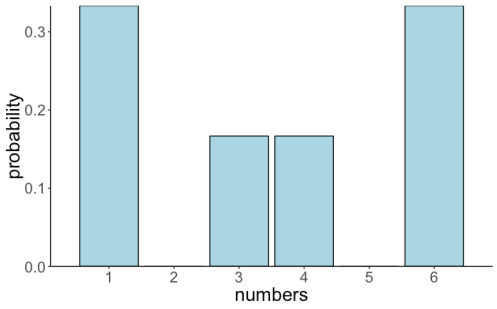
**PRE** is the estimate of an **unknown** true reduction of error $\eta^2$

We need a **sampling distribution** of PRE:

- A distribution of what PRE *would* look like if we repeated our experiment (collected new samples)

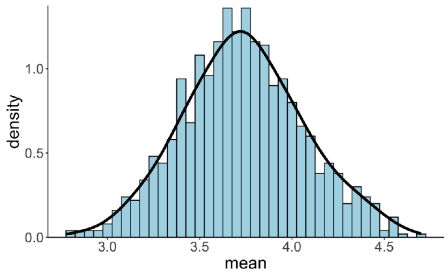
Then we can **compare our observed** PRE to that **distribution!**

# Reminder: 3 main distributions we work with



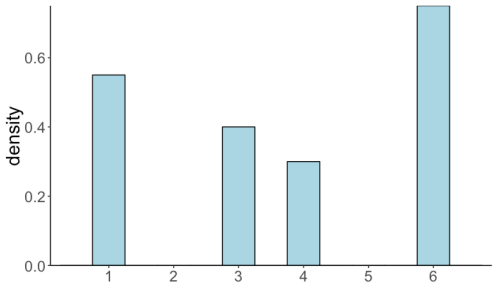

Population distribution

- Unknown - our target for inference & generalization
- If we knew it / could measure it we wouldn't need stats!
- Could have any shape




Sampling distribution

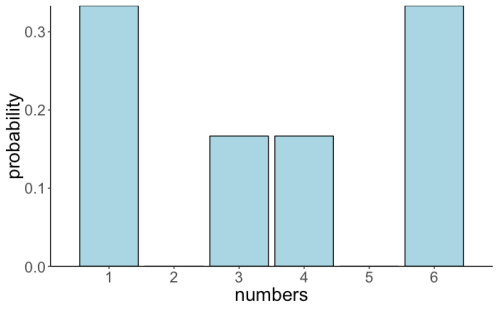
- Key bridge between sample and population
- Shows how estimator varies between samples
- LLN + CLT = Normally distributed
- Derived mathematically - asymptotic distributions
- Or via resampling (bootstrap, permutation)




Sample distribution

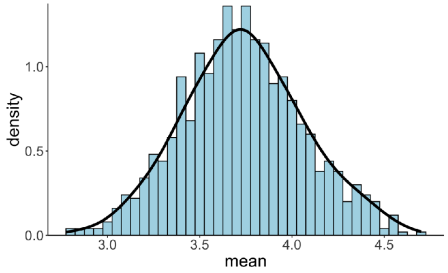
- The **data we actually collect and observe**
- Estimators & statistics of interest we calculate (mean, variance, correlation, ...)
- Use it + sampling distribution to make population inference

# Reminder: 3 main distributions we work with




Population distribution

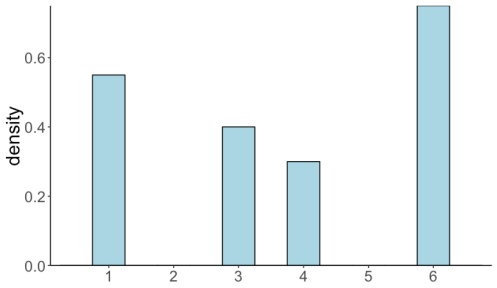
So we can make a claim about this




Sampling distribution

But we need to know this…

*if only there was a distribution that already does this…*




Sample distribution

We calculated PRE using this

# F-distribution = sampling distribution of PRE!

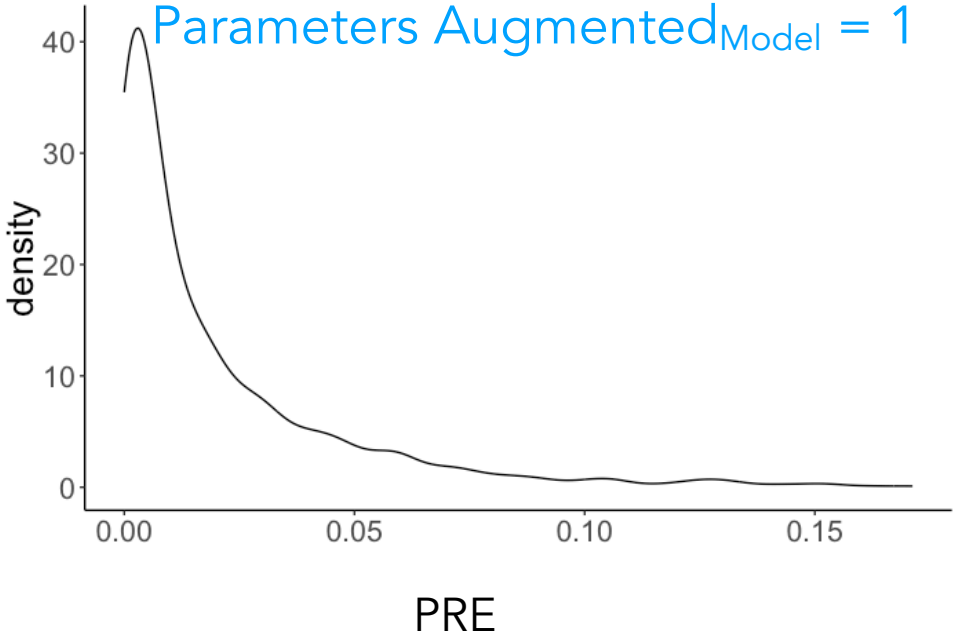
deterministic mapping

*(not in most textbooks!)*

**sampling distribution of PRE**

Parameters $Compact_{Model}$ = 0

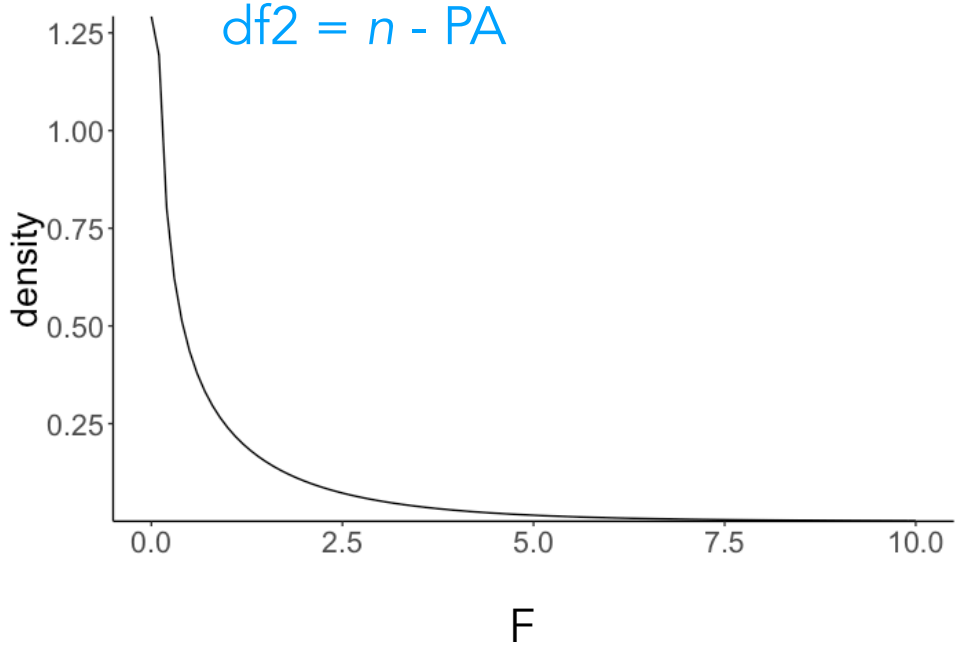
Parameters $Augmented_{Model}$ = 1




***F*(df1, df2) distribution**

df1 = PA - PC

df2 = *n* - PA



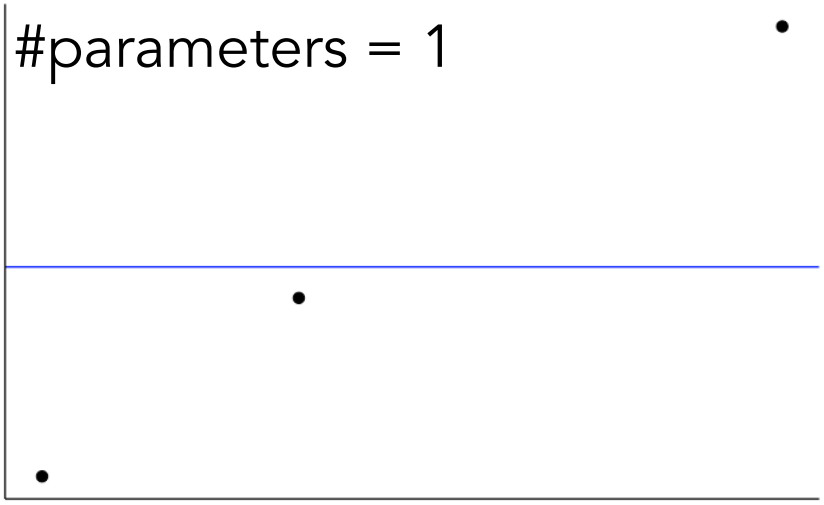

F-statistic: **ratio** of variances

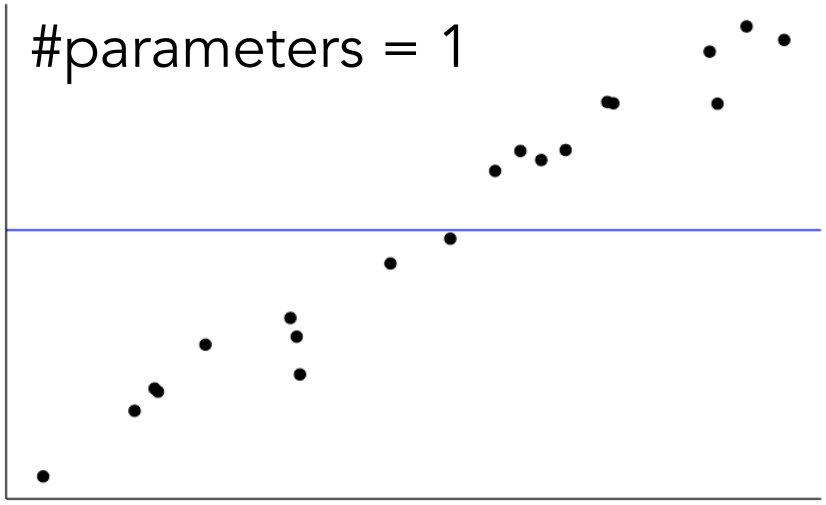
Aka PRE: **ratio** of model performances

# Hypothesis testing as Model Comparison

- We can **formalize any hypothesis** a comparison between models
  - *Does the addition of model parameters justify PRE given the data we have?*
- The answer is more likely to be yes when:
  - **PRE** is high
  - The number of additional parameters is small compared to N

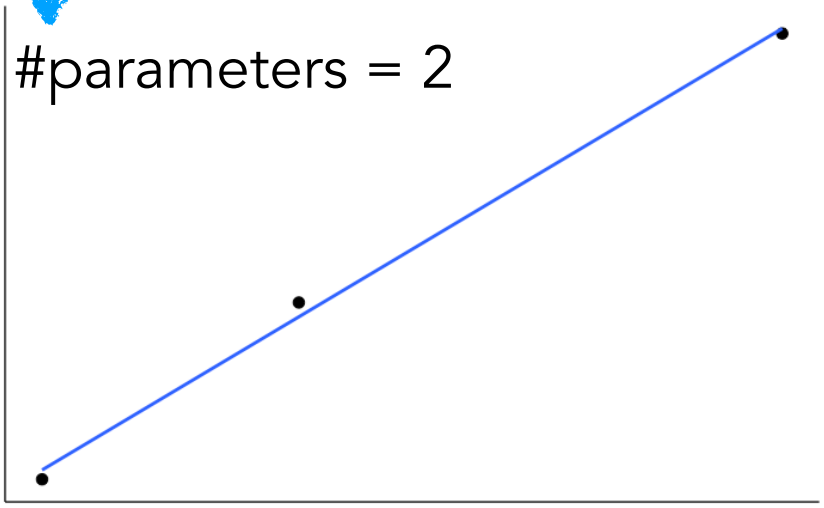
# PRE per parameter for different N

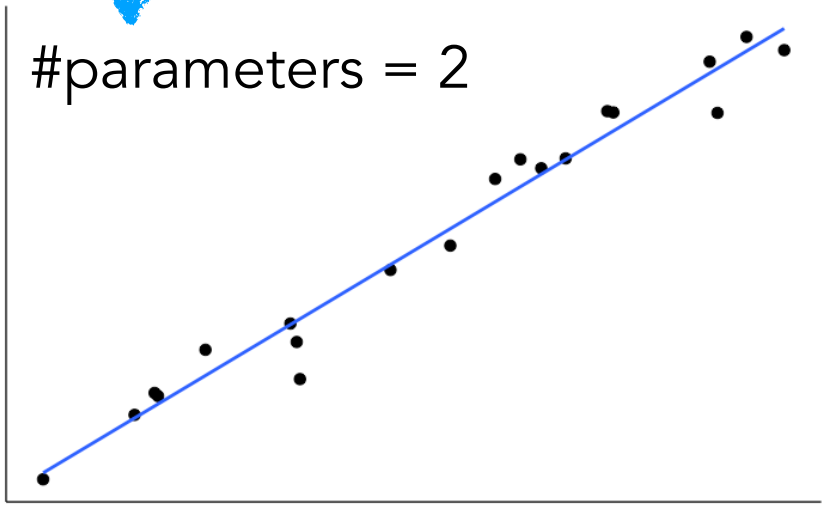



only 2 parameters needed
predict 3 data points
ok…

only 2 parameters needed
predict *many* data points
impressive!



# Hypothesis testing as Model Comparison

(the recipe)

1. Start with research question
2. Formulate hypothesis as a comparison between compact and augmented model
3. Fit parameters for each model using data (next week)
4. Compare models using proportional reduction of error (PRE)
5. Use sampling distribution, (F-distribution or resampling) to decide if worth it

# Hypothesis testing as Model Comparison

(the recipe)

**Restated in frequentist statistics lingo:**

- $model_C$ = compact model = $H_0$ (null hypothesis )
- $model_A$ = augmented model = $H_1$ (alternative hypothesis)

**Hypothesis test:**

- $H_0$: all parameters in $model_A$ but not $model_C$ = 0
- $H_1$: all parameters in $model_A$ but not $model_C$ != 0

# Wrap-Up

- **Statistical models** allow us specify a theory for how **data** were **generated**
- Good models **balance** *complexity* - number of parameters - with *accuracy* - minimizing error of predictions - known as the **bias-variance tradeoff**
- We can formulate hypotheses as **model comparison** - proportional reduction in error relative to addition of more parameters


**Please check out this interactive explorer!**

https://mlu-explain.github.io/bias-variance/