The bias-variance tradeoff

Readings for today

- Chapter 1: Introduction. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R (Vol. 6). New York: Springer.
- Chapter 2: Statistical learning. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R (Vol. 6). New York: Springer.

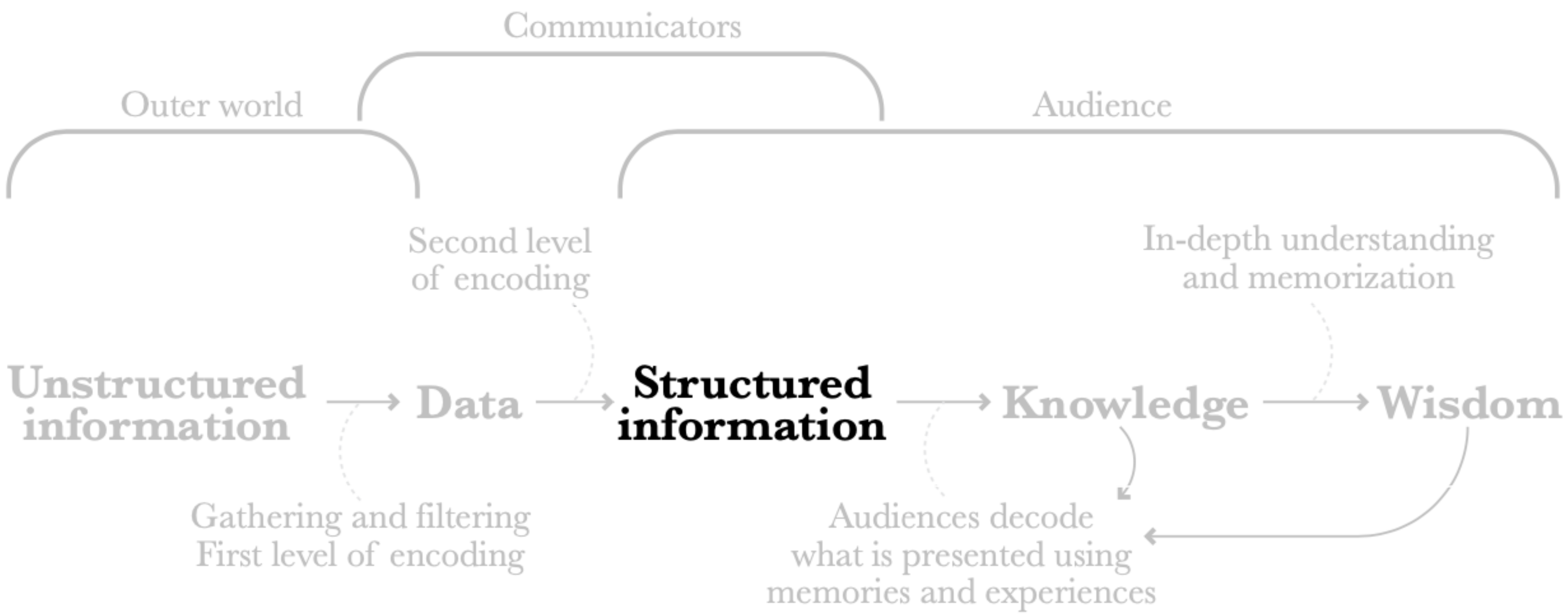
Topics

1. Fundamental form of statistical models

2. Bias-variance tradeoff

Fundamental form of statistical models

From data to wisdom



(Cairo 2012; Weinberg 2010)

Theory → Hypothesis → Statistical Model

Fundamental form of a theory

theory

$$c \leftarrow f(I) = O$$

capacity input output

(van Rooij & Baggio 2020)

Fundamental form of a statistic

relational function

$$Y = f(X)$$

output input

$$f$$

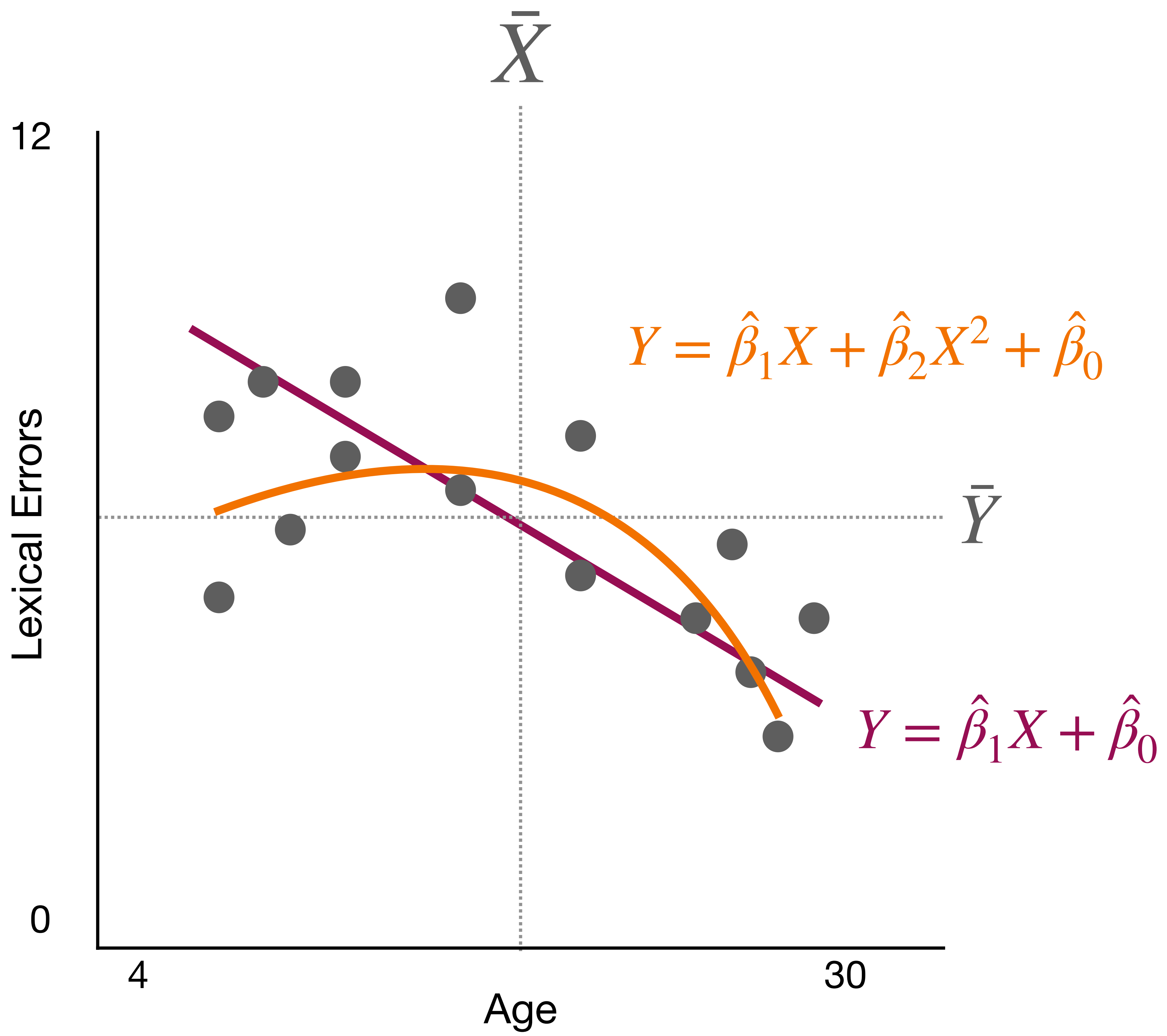
The form of a statistical test, f, is a quantitative description of a specific hypothesis being evaluated (whether or not a p-value is calculated)

Fundamental form of a statistical model

output $\downarrow$ input $\downarrow$ "noise" $\downarrow$

$$Y = f(X) + \epsilon$$

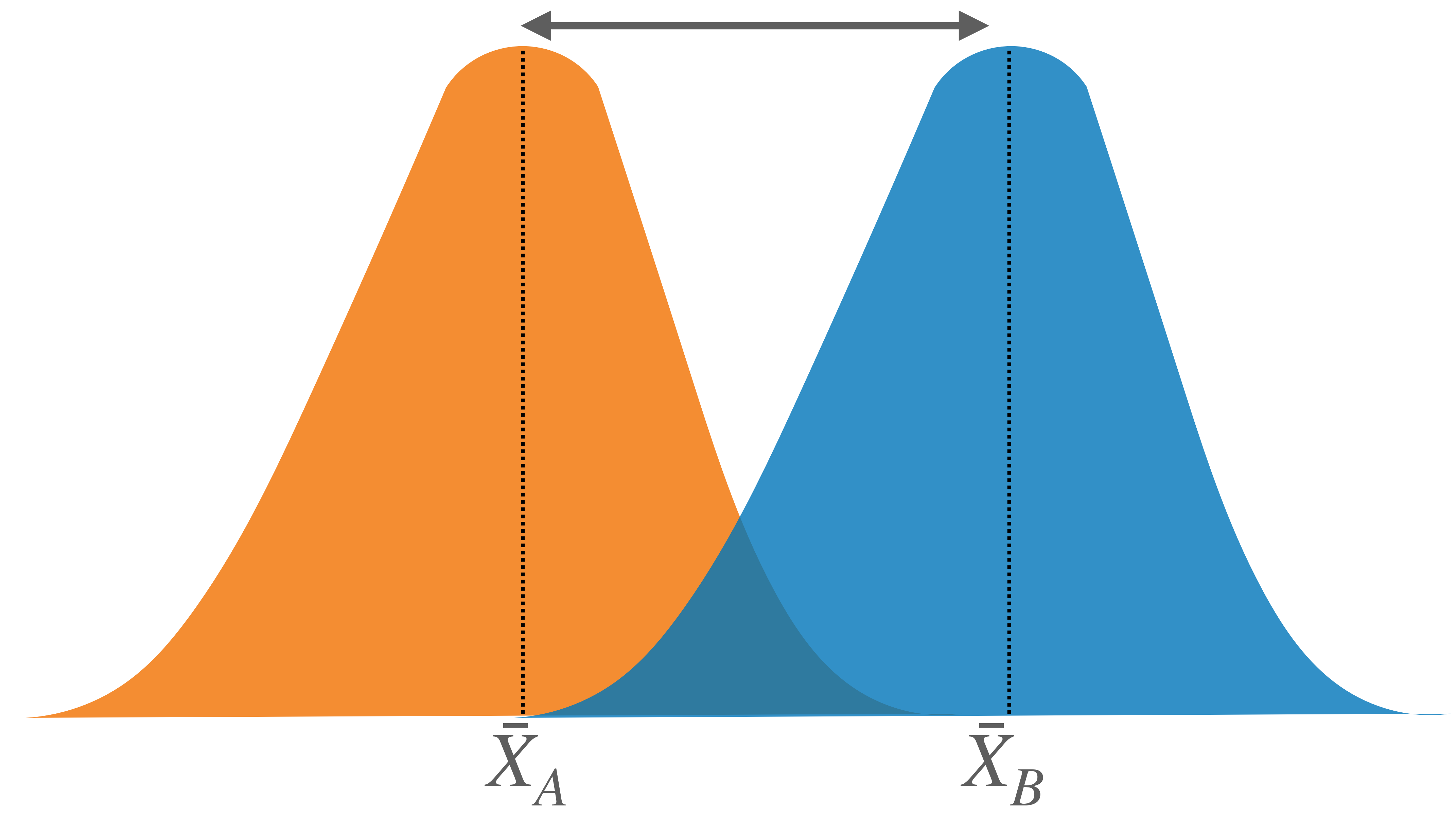
$$\begin{pmatrix} y_1 \\ \cdots \\ y_n \end{pmatrix} \quad \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \cdots & & \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$



(James et al. 2013)

Example: t-test

<u>T-Test</u>: $t = \dfrac{\bar{X}_A - \bar{X}_B}{\sigma_{A,B}} = \dfrac{E(X_A) - E(X_B)}{\sigma_{A,B}}$



(James et al. 2013)

Example: t-test

T-Test: $t = \frac{\bar{X}_A - \bar{X}_B}{\sigma_{A,B}} = \frac{E(X_A) - E(X_B)}{\sigma_{A,B}}$

If group A, otherwise 0

$$\begin{pmatrix} 100 \\ 232 \\ \dots \\ 452 \end{pmatrix} = Y = f(X) = f(\begin{pmatrix} 1 \\ 0 \\ \dots \\ 1 \end{pmatrix}) = \beta_1 X + \beta_0$$

(James et al. 2013)

Goals of learning f

Control

1. Prediction: predict a future observation in Y from X.

2. Inference: understand how changes in X associate with changes in Y.

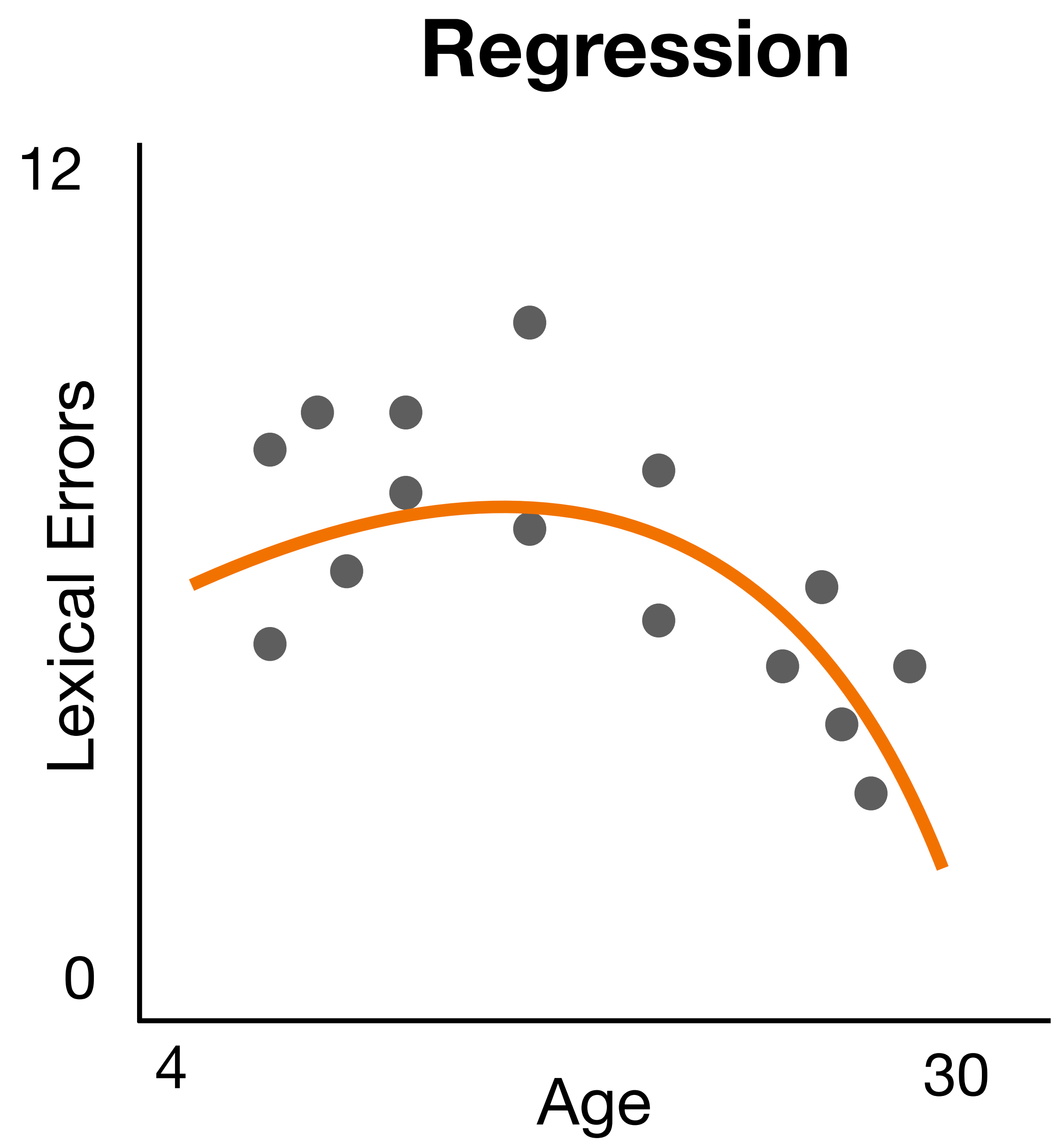
Intuition

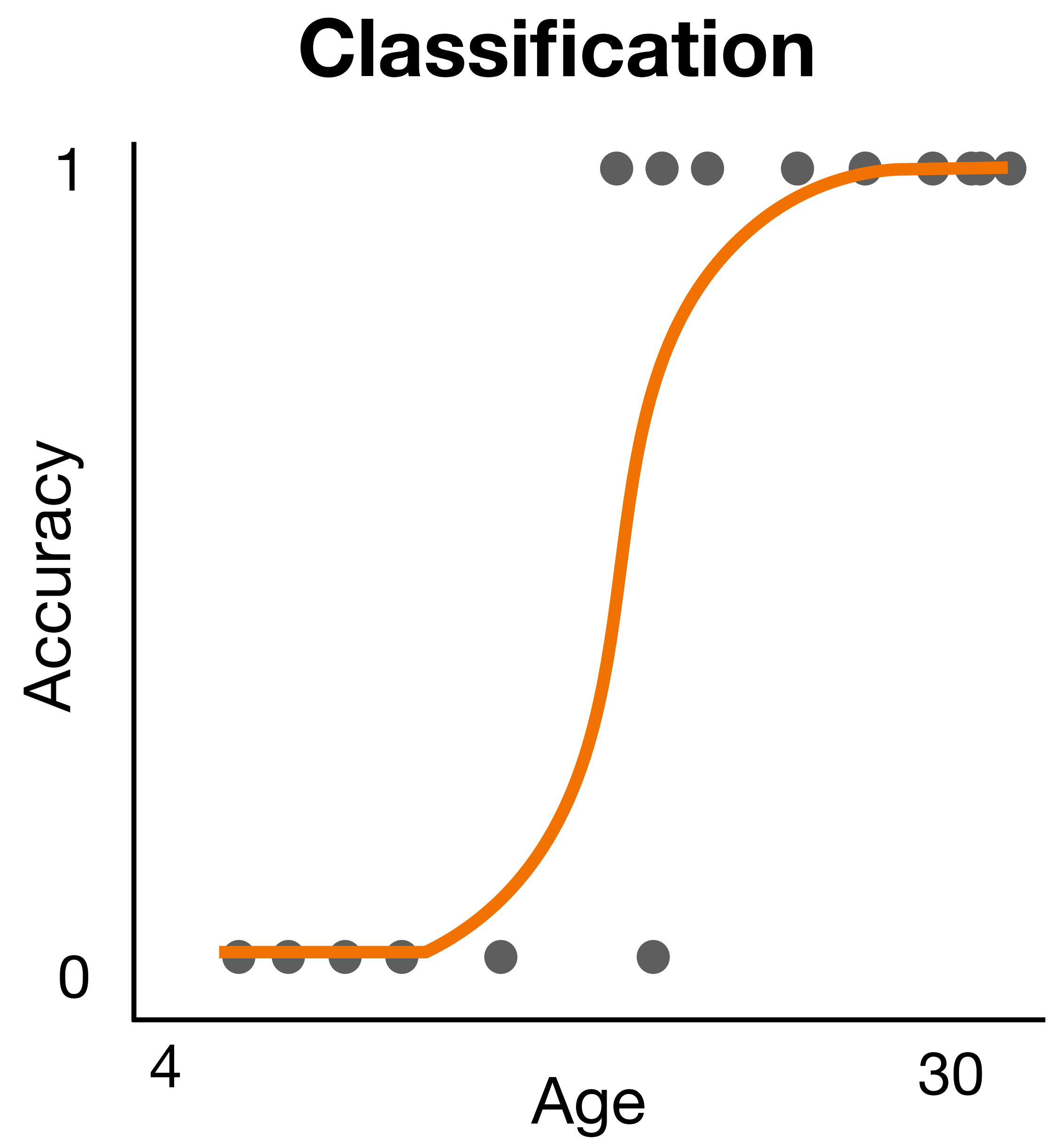
(James et al. 2013)

Classes of models

Y

	Quantitative	Qualitative
Regression	✓	
Classification		✓



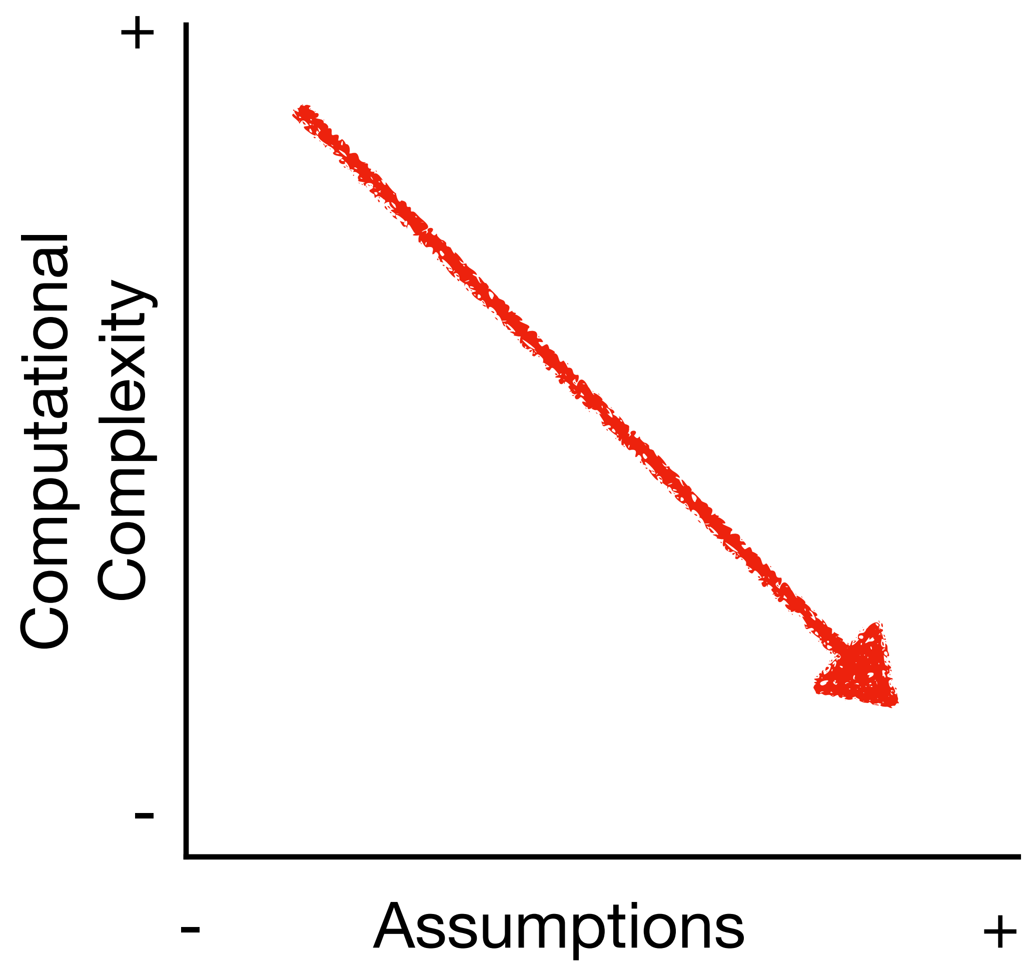


(James et al. 2013)

Forms of f

1. <u>Parametric</u>: assume a functional form of f that can be described by a small number of parameters.

2. <u>Nonparametric</u>: estimate f as close to the data as possible.



Linear regression

$$f(y_i \mid x_i, \beta_1, \beta_0, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}}$$

kNN regression

$$f(y_i) = \frac{1}{K} \sum_{i \in K} y_i$$

(James et al. 2013)

Bias-variance tradeoff

Dimensions

Dimensions of a model

n: number of observations (i.e., rows)

p: number of features/independent variables (i.e., columns)

Dimensionality of a model: n x p

As n $\rightarrow$p, dimensionality increases

$$\begin{pmatrix} x_{1,1} \\ \cdots \\ x_{n,1} \end{pmatrix} \rightarrow \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,15} \\ \cdots & & & \\ x_{n,1} & x_{n,2} & \cdots & x_{n,15} \end{pmatrix}$$

(James et al. 2013)

Learning f

1. Supervised: find a model that minimizes a loss function on X and Y.

2. Unsupervised: find f independent of Y.

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

$$\arg\min_{S} = \sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu_i \|^2$$

(James et al. 2013)

Types of error

$$Y = f(X) + \epsilon$$

<u>Error:</u>

$$E[(Y - \hat{Y})^2] = E[(Y - \hat{f}(X))^2]$$

$$= E[(f(X) + \epsilon - \hat{f}(X))^2]$$

$$= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{Var(\epsilon)}_{\text{Irreducible}}$$

1. <u>Reducible Error</u>: Error that can be explained by $\hat{f}$.

2. <u>Irreducible Error</u>: Error that you have no control over (i.e., noise).

(James et al. 2013)

Flexibility & Generalizability

$$E[(Y - \hat{f}(X))^2] = Var(\hat{f}(X)) + [Bias(\hat{f}(X))]^2 + Var(\epsilon)$$

expected error

how $\hat{f}$ changes with different X & Y

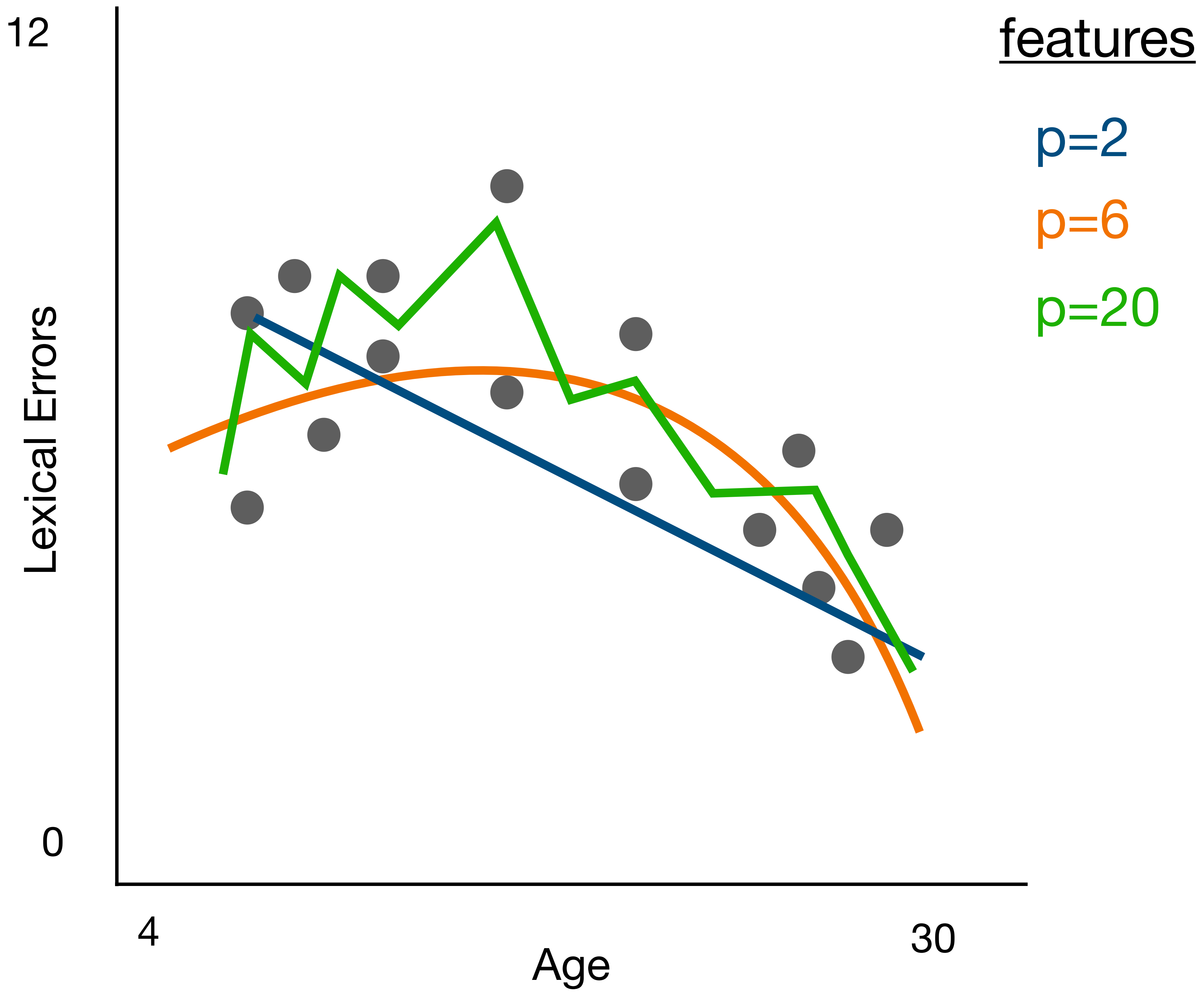
how well $\hat{f}$ generalizes to a new sample of Y

the unexplainable

Bias-variance tradeoff

(James et al. 2013)

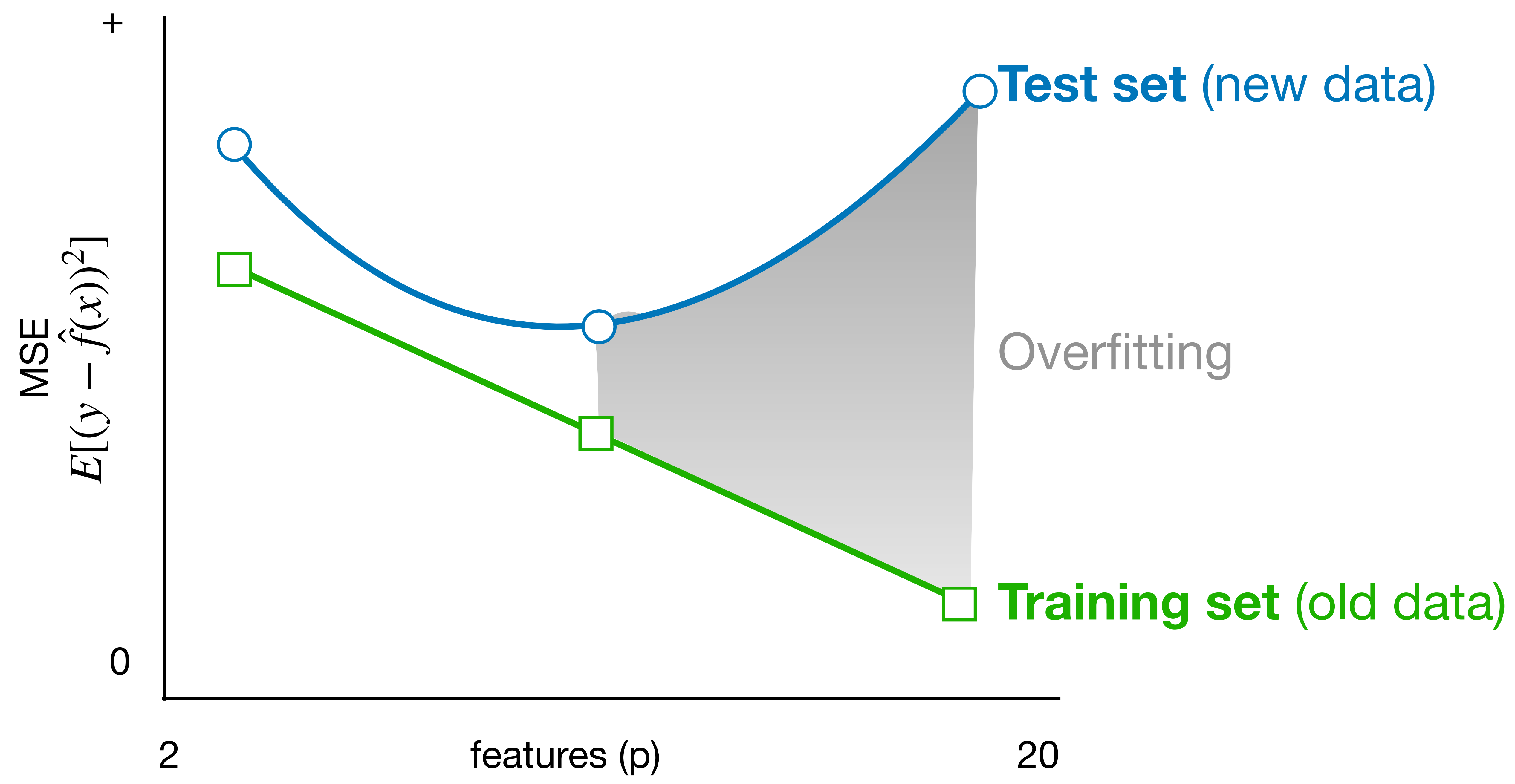
Bias-variance tradeoff



↑ flexibility = ↓ bias

(James et al. 2013)

Bias-variance tradeoff



<u>Goal:</u> Find the right complexity that balances the flexibility (variance) of a trained model with its bias.

(James et al. 2013)

Take home message

- The fundamental form of all relationships in statistics is $Y = f(X) + \epsilon$.
- The best models account for reducible error in a way that explains the most variance (flexibility) while retaining the ability to generalize to new data (bias).