Introduction to Machine Learning

Loss Functions, Bias and Variance

Loss functions

- 0/1 loss
  - Perceptron

Any hope?

- Consider modifying function to add a regularizer
- R(w,b)

- Optimize trade-off between low training error and simplicity

Relationship between margin and loss

- Small changes to w,b have large impact on objective function

Relationship between margin and loss

- Sigmoid function

Surrogate loss functions

- 0/1
- Hinge
- Logistic
- Exponential
- Squared

$$\ell^{(0/1)}(y, \hat{y}) = \mathbf{1}[y\hat{y} \leq 0]$$

$$\ell^{(\text{hin})}(y, \hat{y}) = \max\{0, 1 - y\hat{y}\}$$

$$\ell^{(\log)}(y, \hat{y}) = \frac{1}{\log 2} \log\left(1 + \exp[-y\hat{y}]\right)$$

$$\ell^{(\exp)}(y, \hat{y}) = \exp[-y\hat{y}]$$

$$\ell^{(\text{sqr})}(y, \hat{y}) = (y - \hat{y})^2$$

Surrogate loss functions

- 0/1
- Hinge
- Logistic
- Exponential
- Squared

$$\ell^{(0/1)}(y, \hat{y}) = \mathbf{1}[y\hat{y} \leq 0]$$

$$\ell^{(\text{hin})}(y, \hat{y}) = \max\{0, 1 - y\hat{y}\}$$

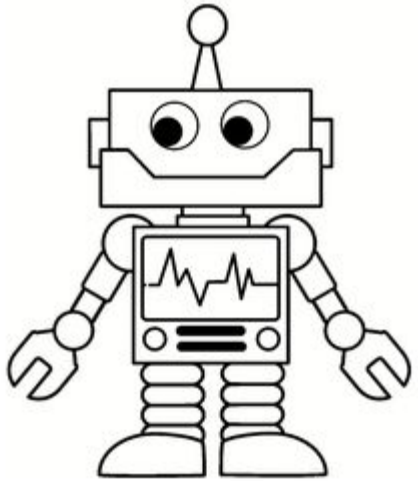
$$\ell^{(\log)}(y, \hat{y}) = \frac{1}{\log 2} \log\left(1 + \exp[-y\hat{y}]\right)$$

$$\ell^{(\exp)}(y, \hat{y}) = \exp[-y\hat{y}]$$

$$\ell^{(\text{sqr})}(y, \hat{y}) = (y - \hat{y})^2$$

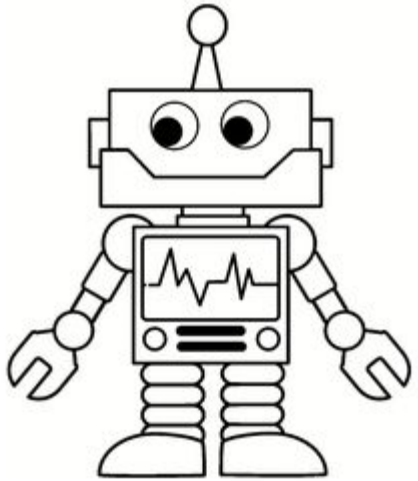
Regularizer

- Helps to prevent overfit
- Perceptron $\min_{w,b} \sum_n \mathbf{1}[y_n(\boldsymbol{w} \cdot \boldsymbol{x}_n + b) \leq 0] + \lambda R(\boldsymbol{w}, b)$
- What are good choices of R(w,b) for hyperplanes?



Regularizer

- Helps to prevent overfit
- Perceptron $\min_{w,b} \sum_n \mathbf{1}[y_n(\boldsymbol{w} \cdot \boldsymbol{x}_n + b) \leq 0] + \lambda R(\boldsymbol{w}, b)$
- What are good choices of R(w,b) for hyperplanes?



Norm

- Popular choice
- Keeps function (weights) from changing too quickly
- $R^{(\text{norm})}(\boldsymbol{w}, b) = ||\boldsymbol{w}|| = \sqrt{\sum_d w_d^2}$
  - Convex
  - Smooth
- l-norm (p-norm)

$$||\boldsymbol{w}||_p = \left(\sum_d |w_d|^p \right)^{\frac{1}{p}}$$

P-norm

$$||\boldsymbol{w}||_p = \left(\sum_d |w_d|^p \right)^{\frac{1}{p}}$$

Bias and fairness

Bias and Fairness

Business

Apple Card algorithm sparks gender bias allegations against Goldman Sachs

COMPUTING

Racial Bias Found in a Major Health Care Risk Algorithm

When a Computer Program Keeps You in Jail

What types of bias can exist?

- Adaptation

Fairness and data bias

- Labeling
- Sample selection
- Task
- Feedback loops

Error

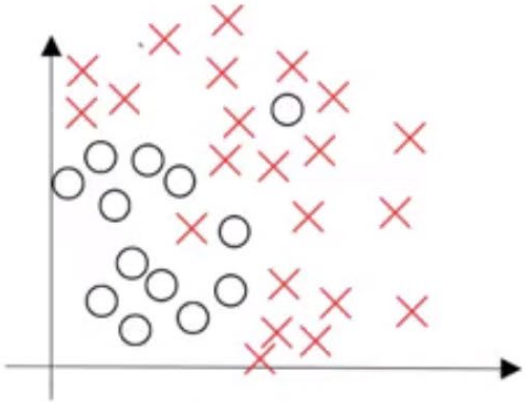
$$\text{error}(f) = \underbrace{\left[\text{error}(f) - \min_{f^* \in \mathcal{F}} \text{error}(f^*)\right]}_{\text{estimation error}} + \underbrace{\left[\min_{f^* \in \mathcal{F}} \text{error}(f)\right]}_{\text{approximation error}}$$

Error

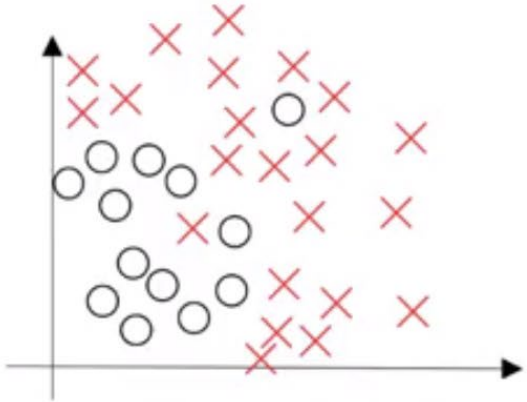
- Bias error
  - Simplifying assumptions that make target function easier to learn
- Variance error
  - Amount that estimated target function will change if change training data

Bias-Variance tradeoff

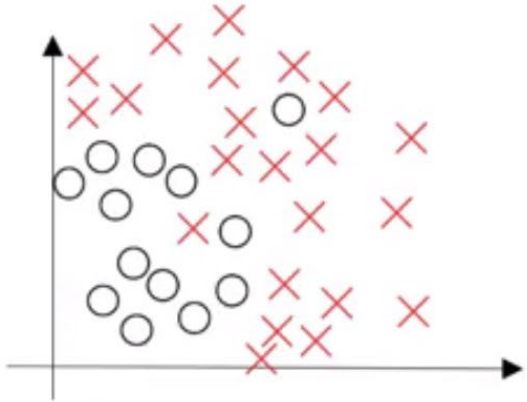
- Goal: low bias and low variance
- As we increase complexity
  - Bias decreases (better fit to data)
  - Variance increases (fit varies more with data)



high bias



"just right"



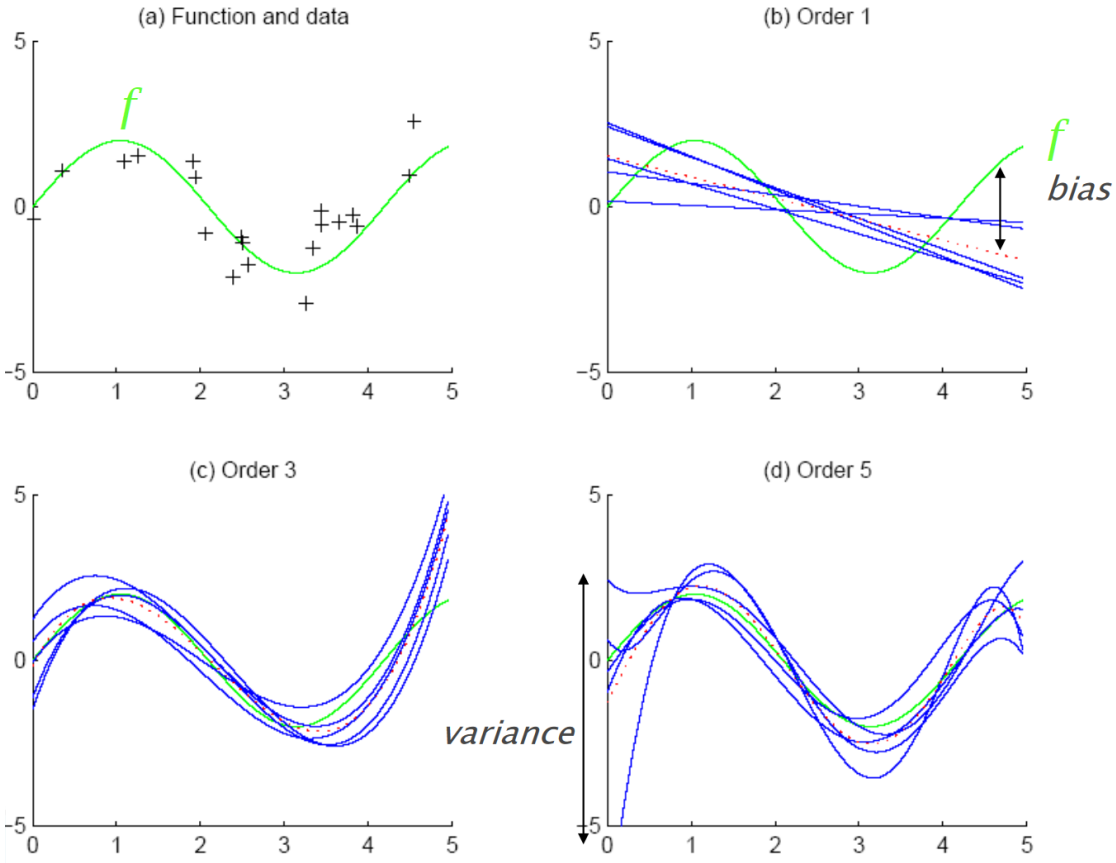
high variance

Bias and variance

Cat classification

Example

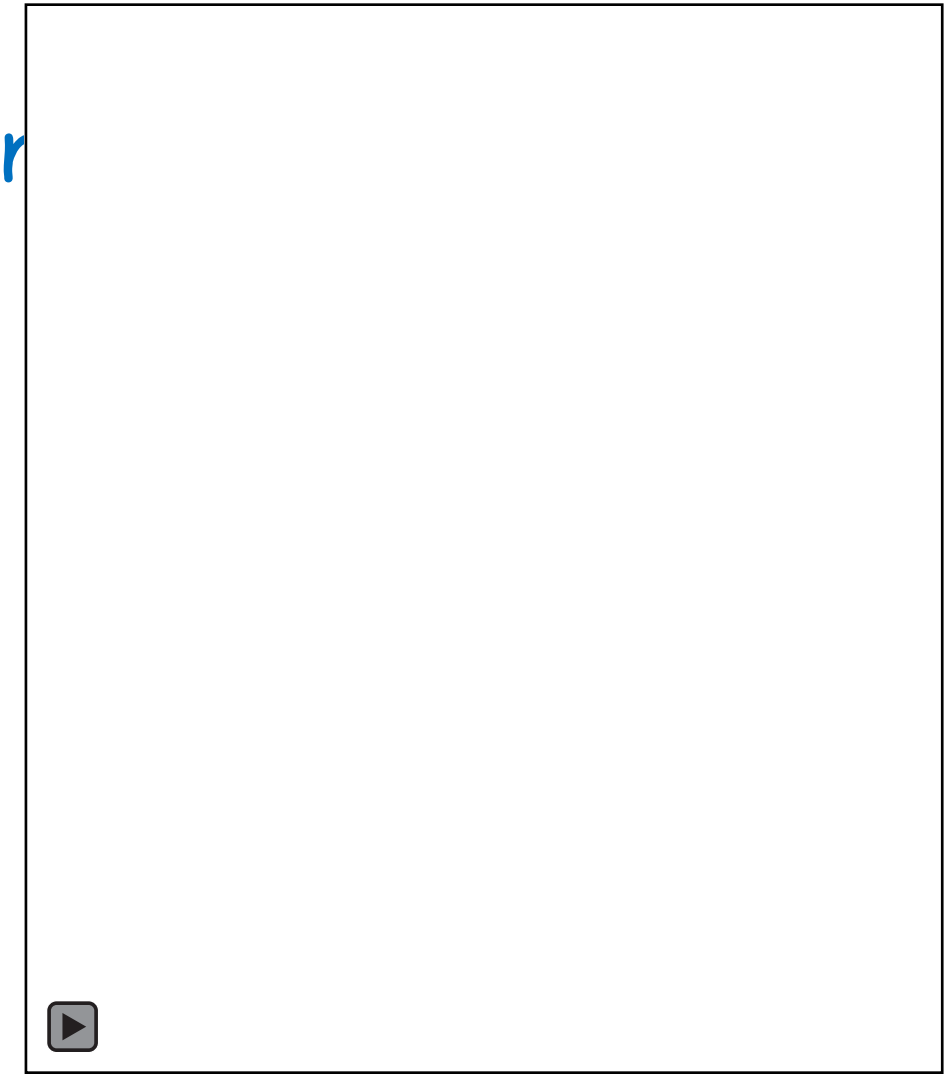
- f(x) = 2 sin(1.5x)



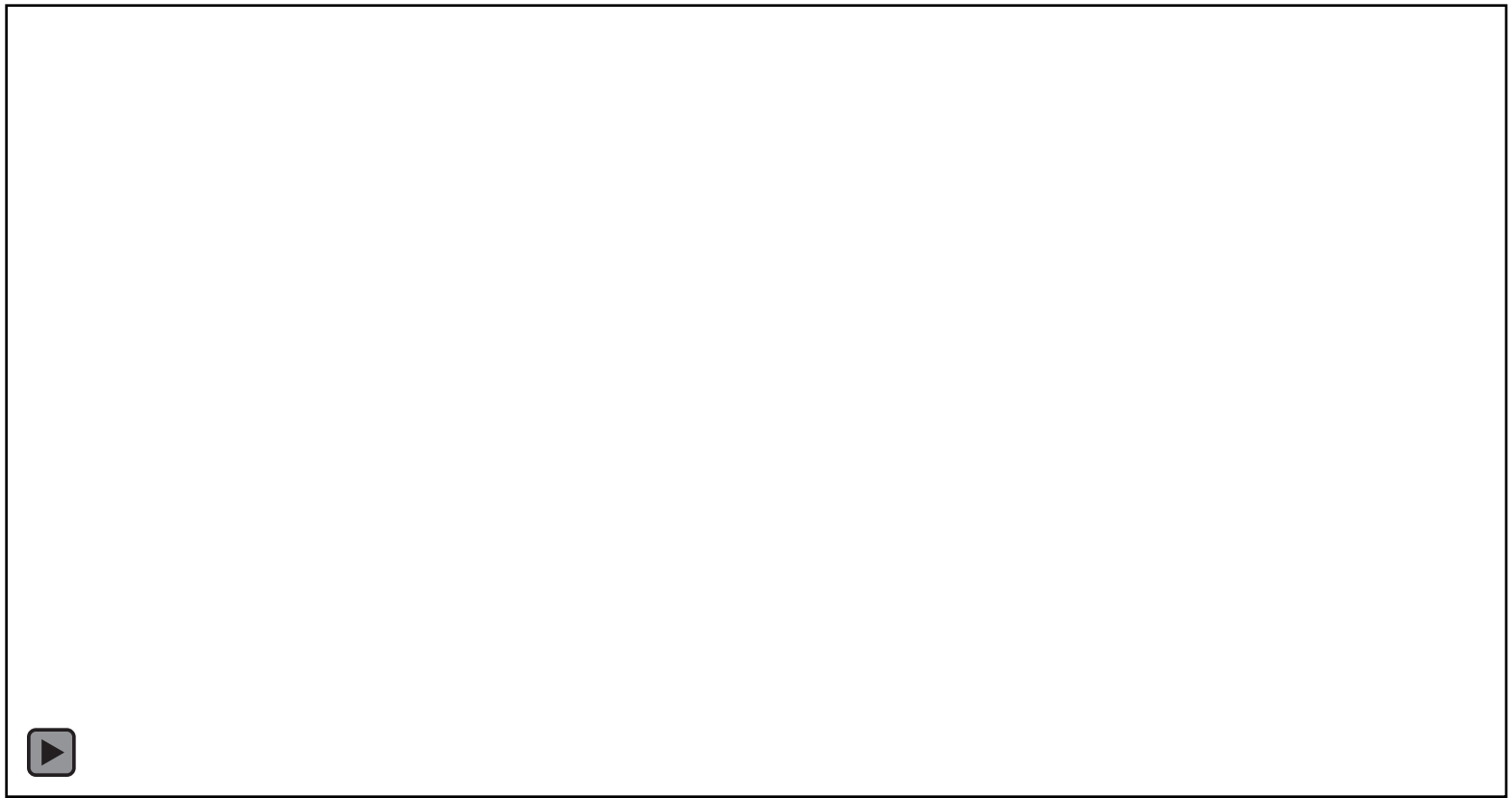
Regularization

- Path to bias-variance trade-off

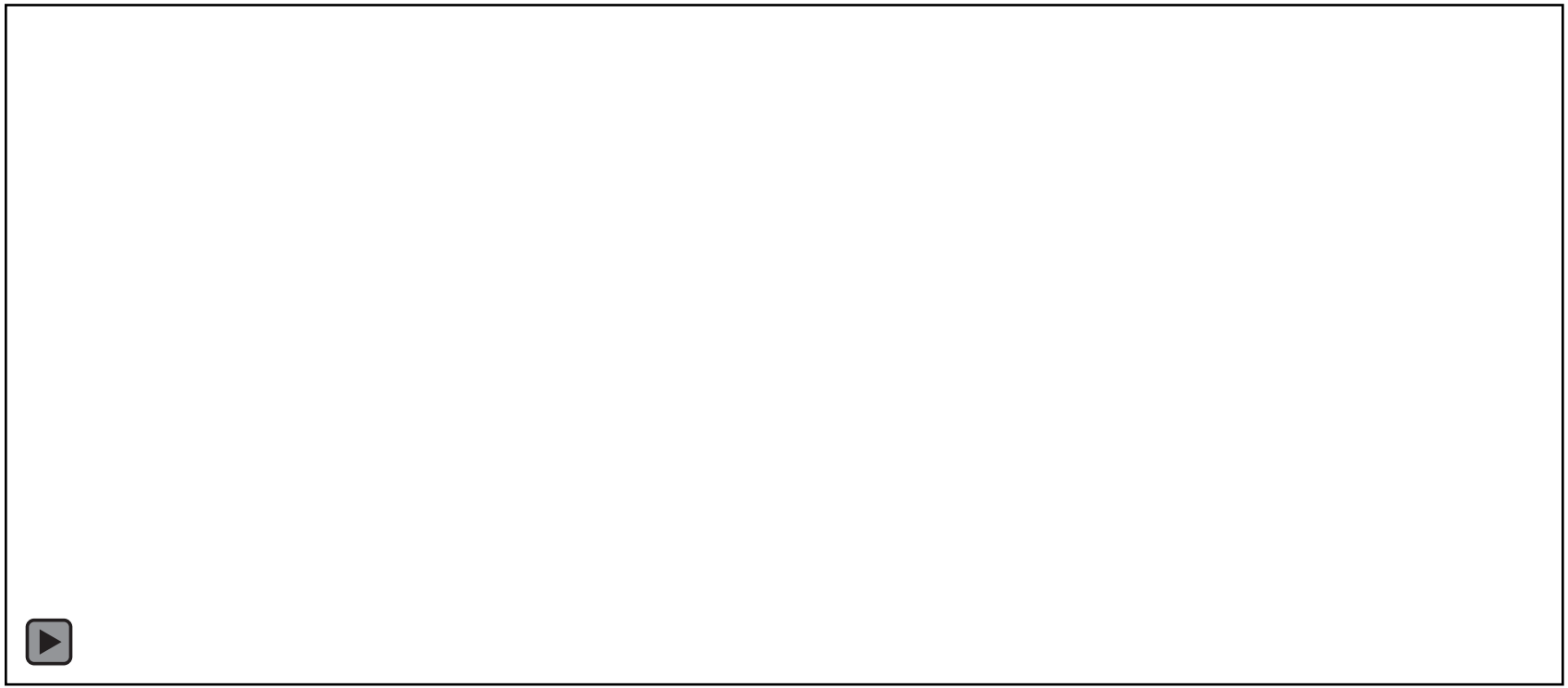
DL Bias and Variar



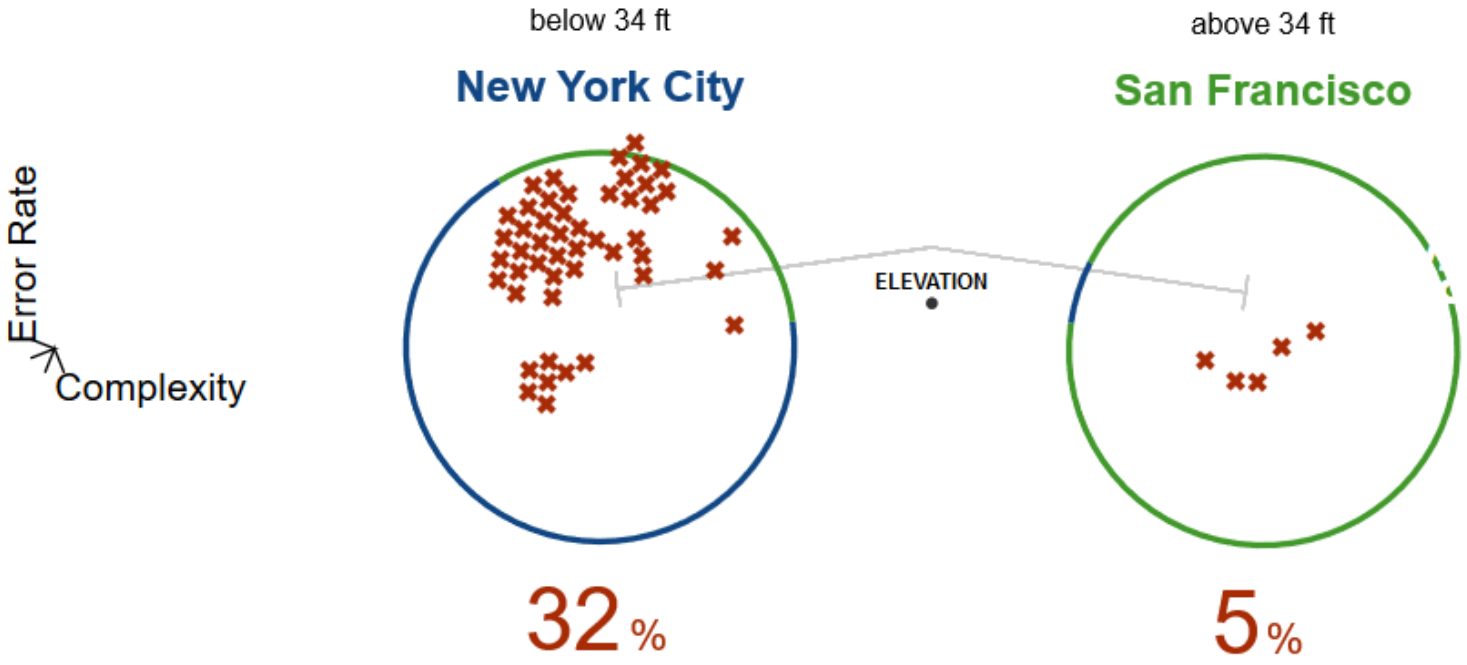
Decision stump



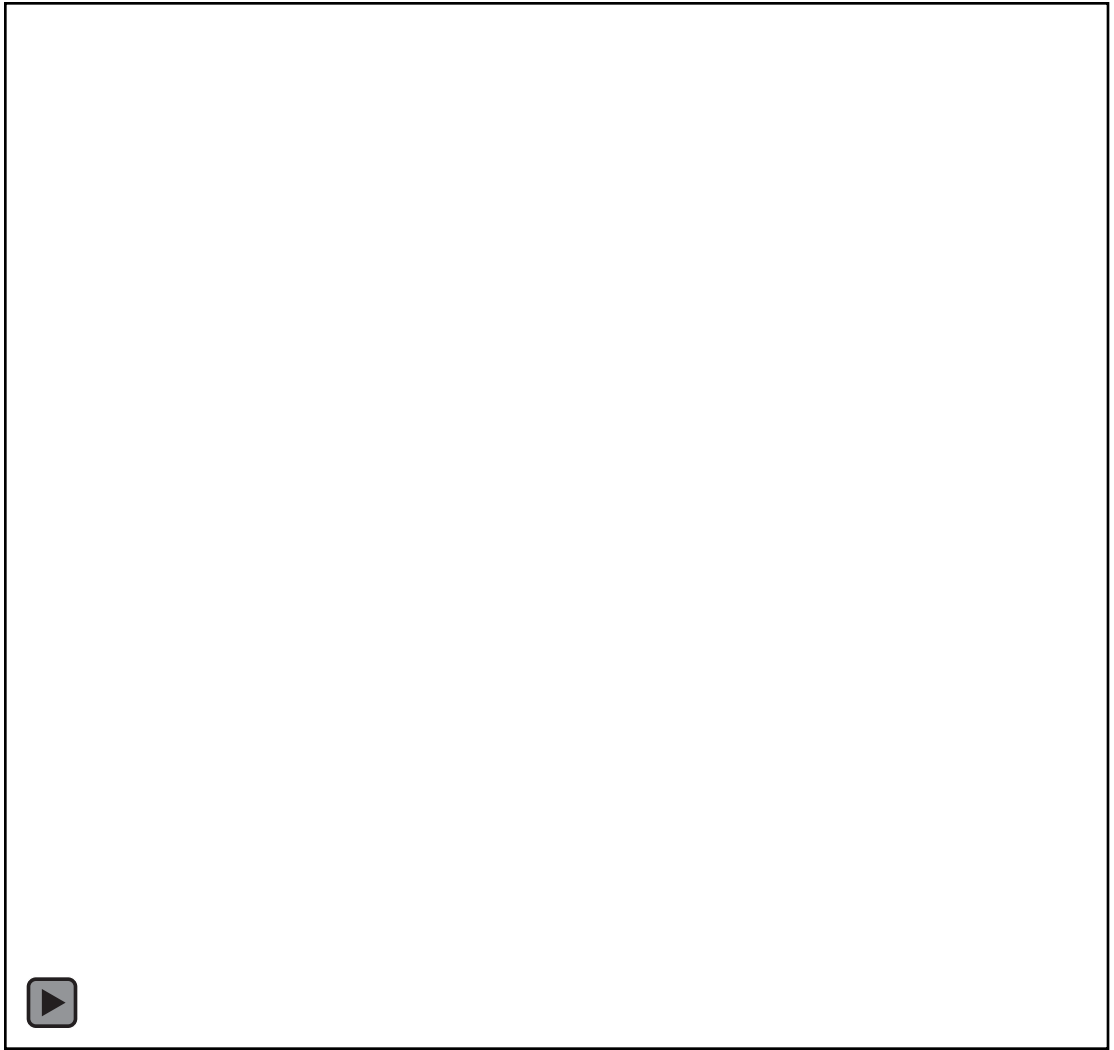
Decision stump



Errors due to bias

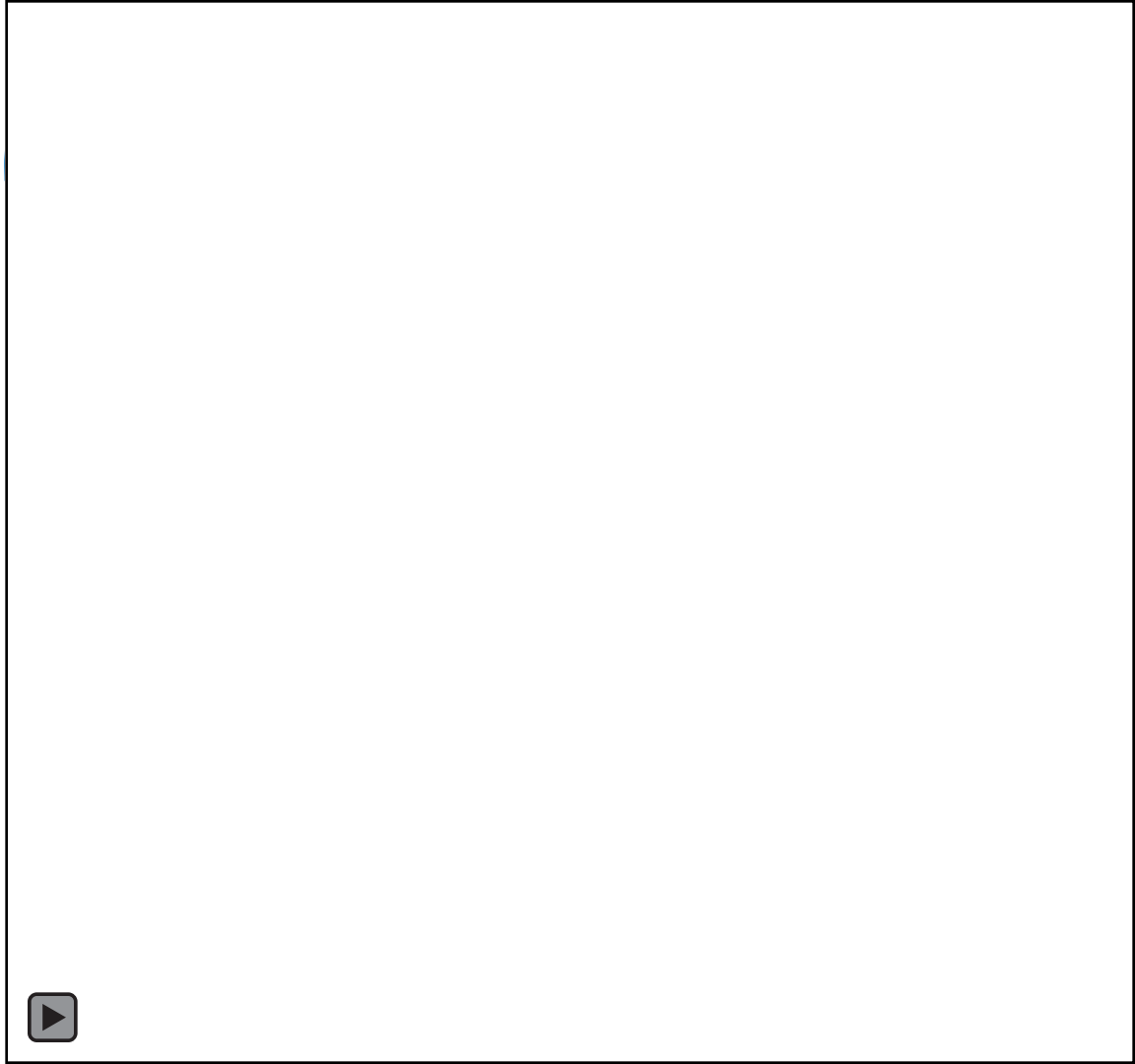


Decrease error due to bias

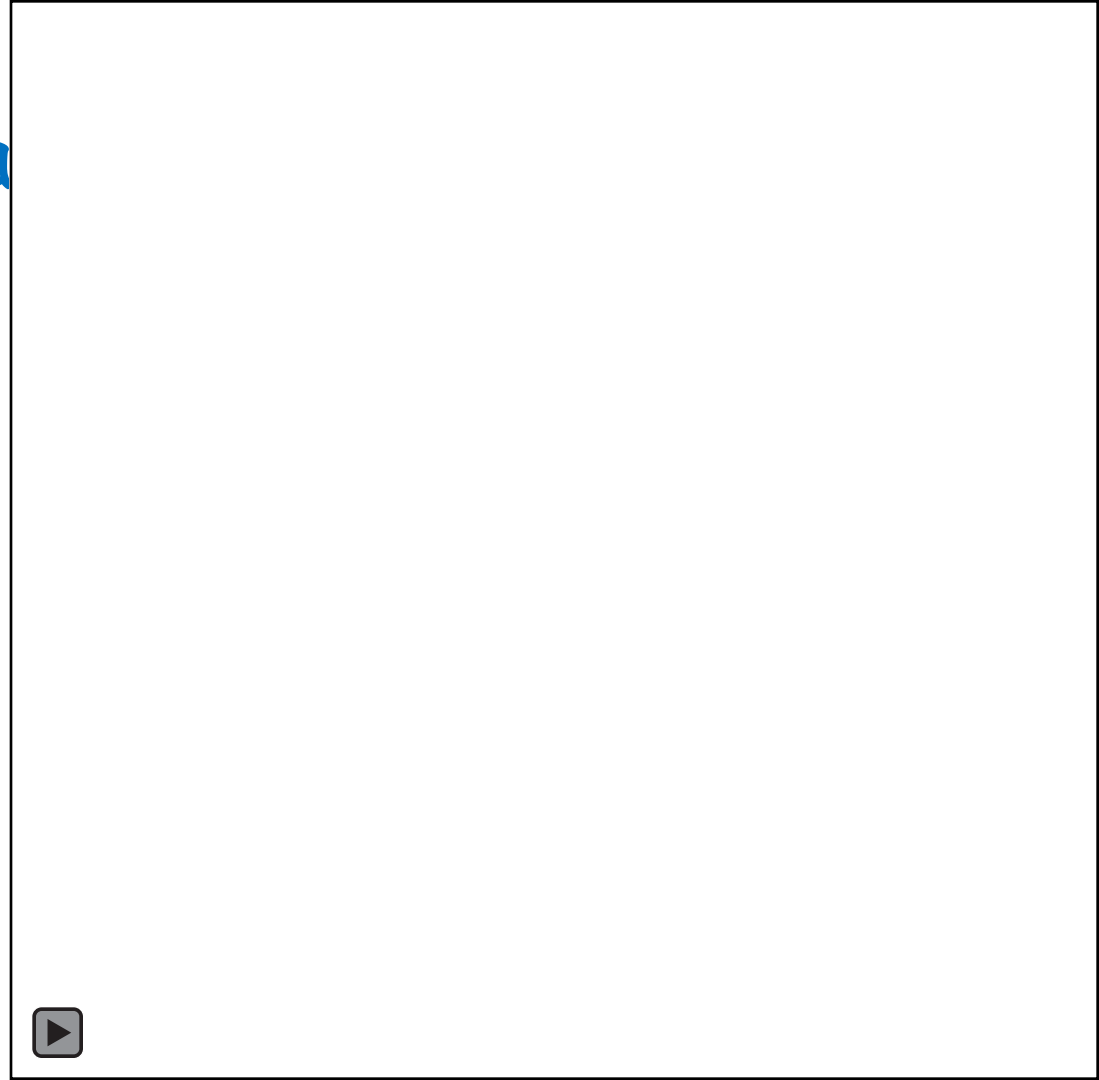


Example of var

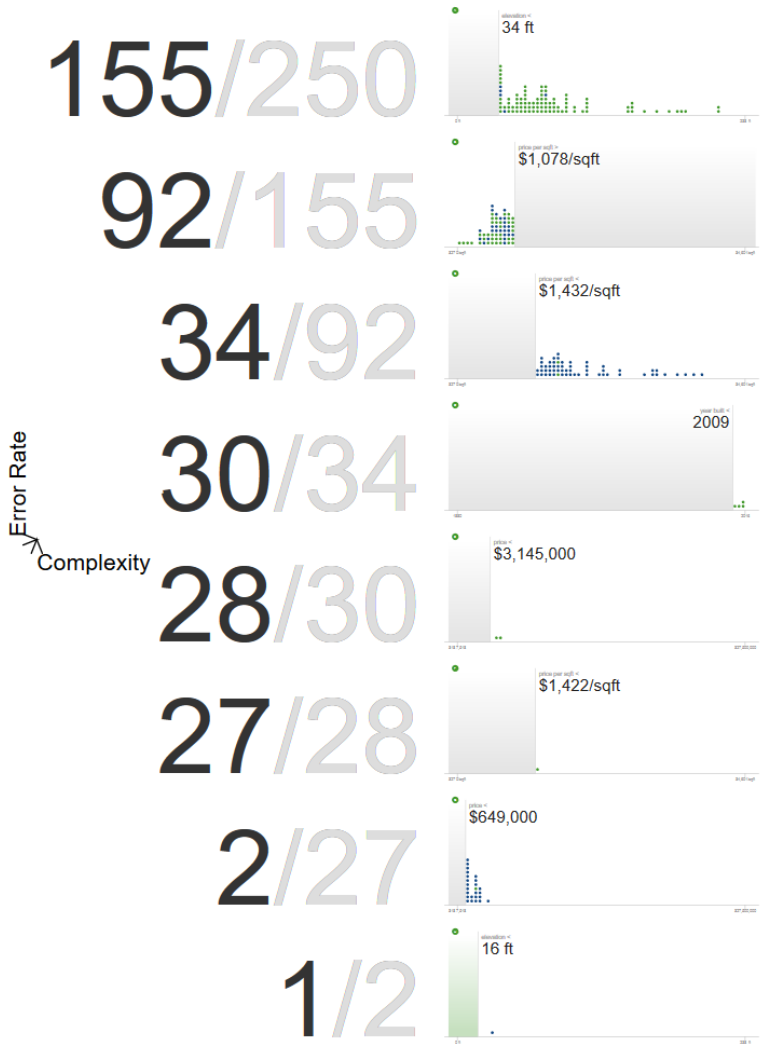
Example of var



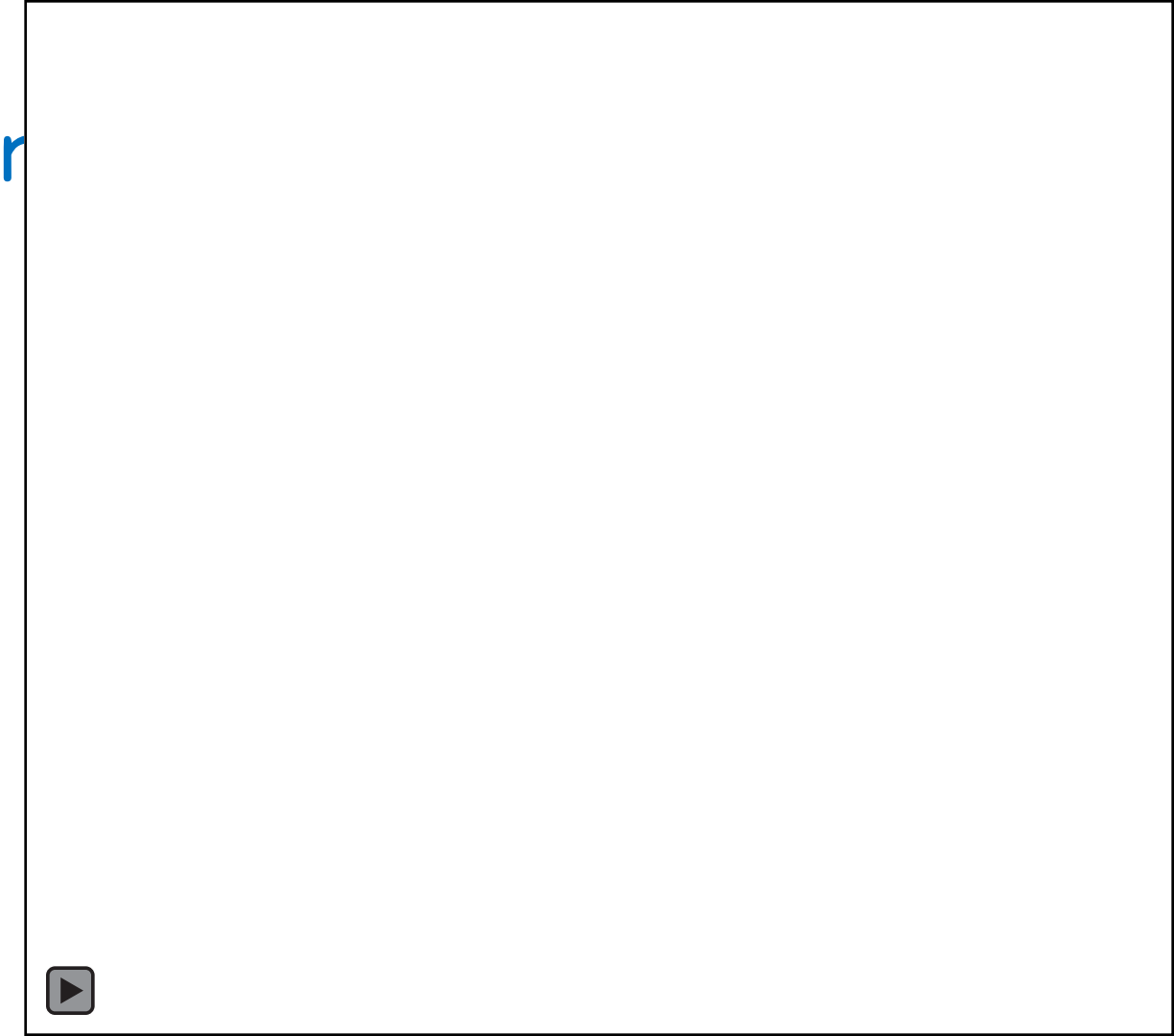
Example of varia



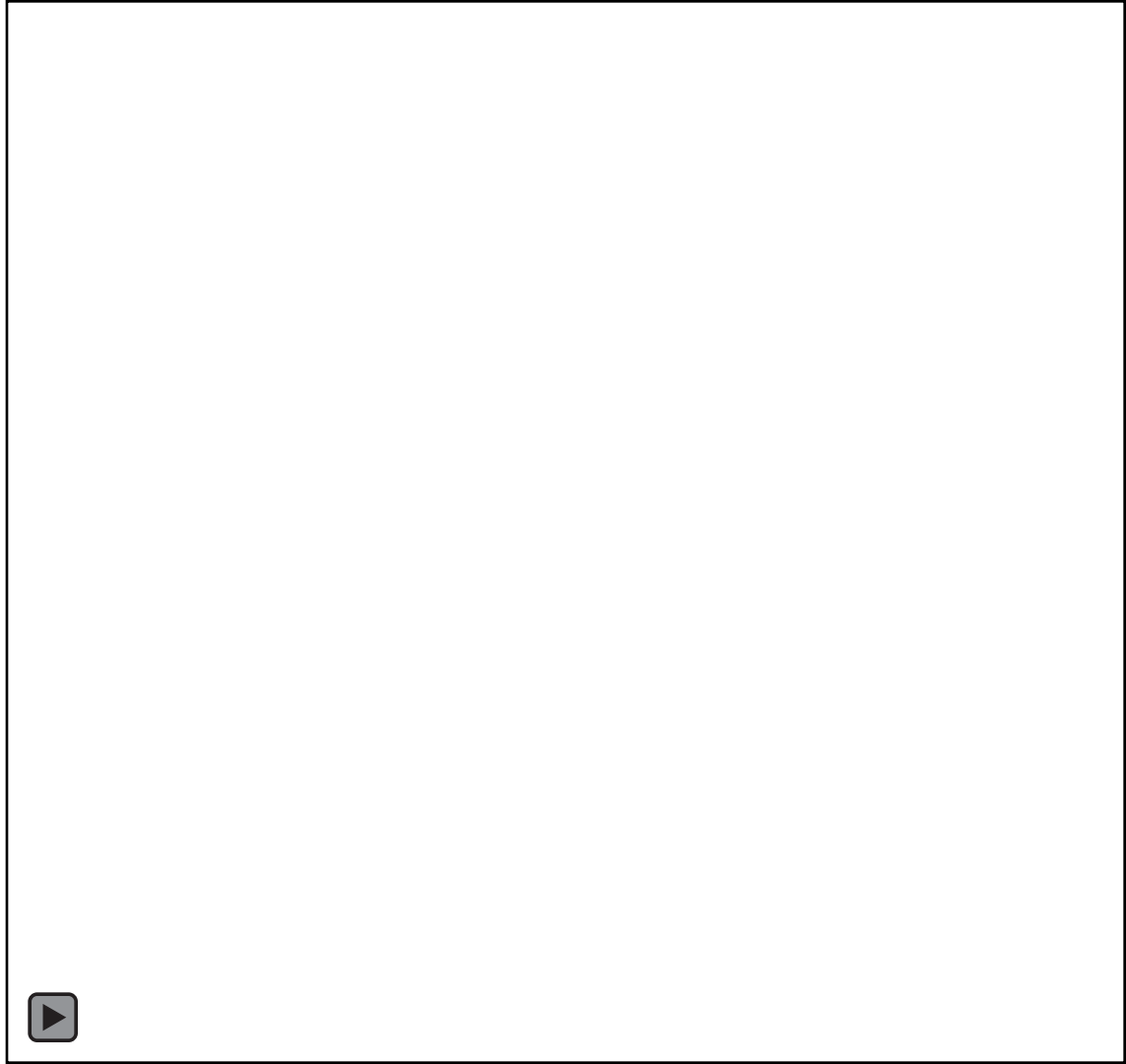
Example of variance



Flukes are nor



Bias-variance tradeoff



Bias-va
tradeof

Bias-var
tradeof