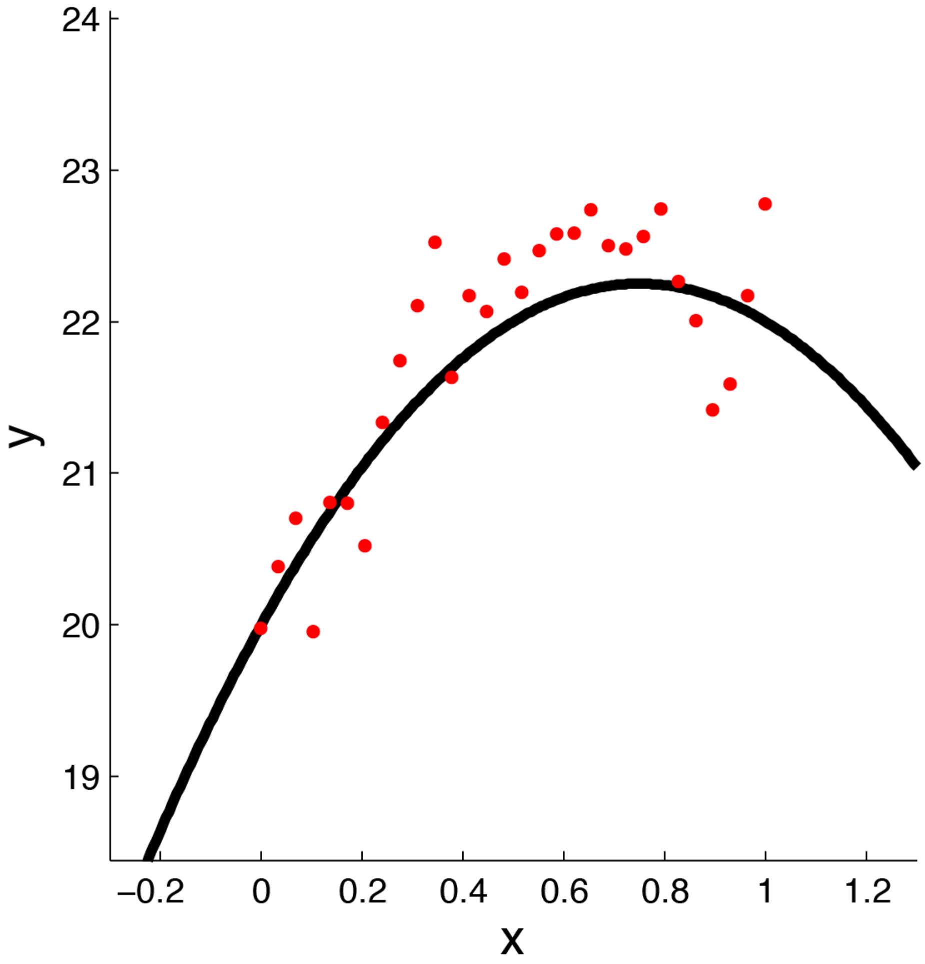
# Statistics and Data Analysis in MATLAB

# Lecture 5:
# Model accuracy

Kendrick Kay
Washington University in St. Louis
March 21, 2014

# Quantifying model accuracy



Squared error = 5.4
(dependent on units, hard to interpret)

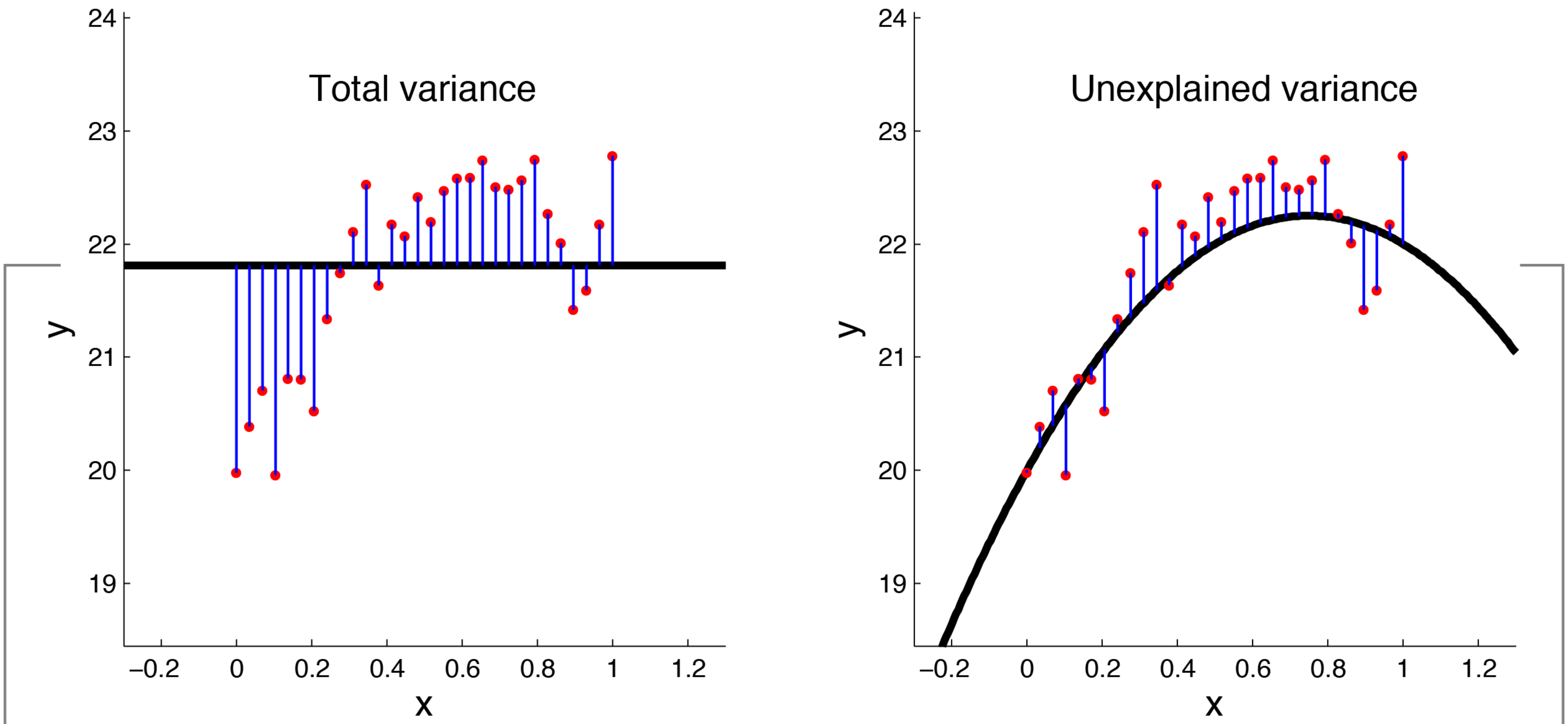
$R^2$ = 75%
(independent of units, easy to interpret)

Kendrick Kay, Washington Univ. in St. Louis

# Variance

$$\text{variance} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

# Coefficient of determination ($R^2$)



$$R^2 = 100 \times \left( 1 - \frac{\text{unexplained variance}}{\text{total variance}} \right)$$

Kendrick Kay, Washington Univ. in St. Louis

# Coefficient of determination ($R^2$)

$$R^2 = \text{percent explained variance}$$

$$R^2 = 100 \times \left(\text{fraction explained variance}\right)$$

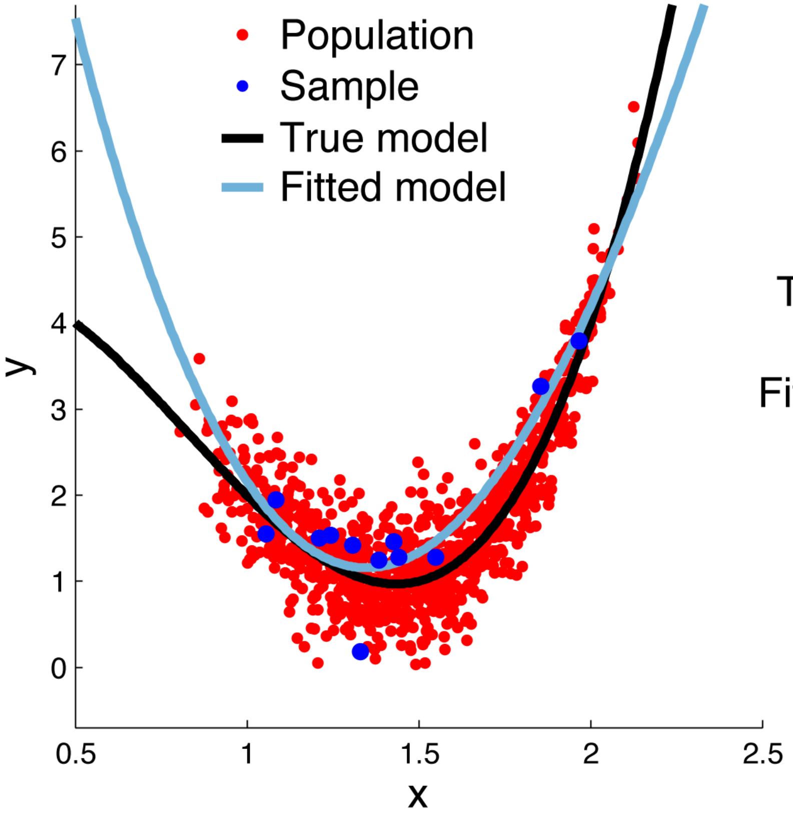
$$R^2 = 100 \times \left(1 - \frac{\text{unexplained variance}}{\text{total variance}}\right)$$

$$R^2 = 100 \times \left(1 - \frac{\dfrac{\sum_{i=1}^{n}(d_i - m_i)^2}{n-1}}{\dfrac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}\right)$$

$$R^2 = 100 \times \left(1 - \frac{\sum_{i=1}^{n}(d_i - m_i)^2}{\sum_{i=1}^{n}(d_i - \bar{d})^2}\right)$$

*Kendrick Kay, Washington Univ. in St. Louis*

# Direct calculation of *$R^2$* overestimates model accuracy



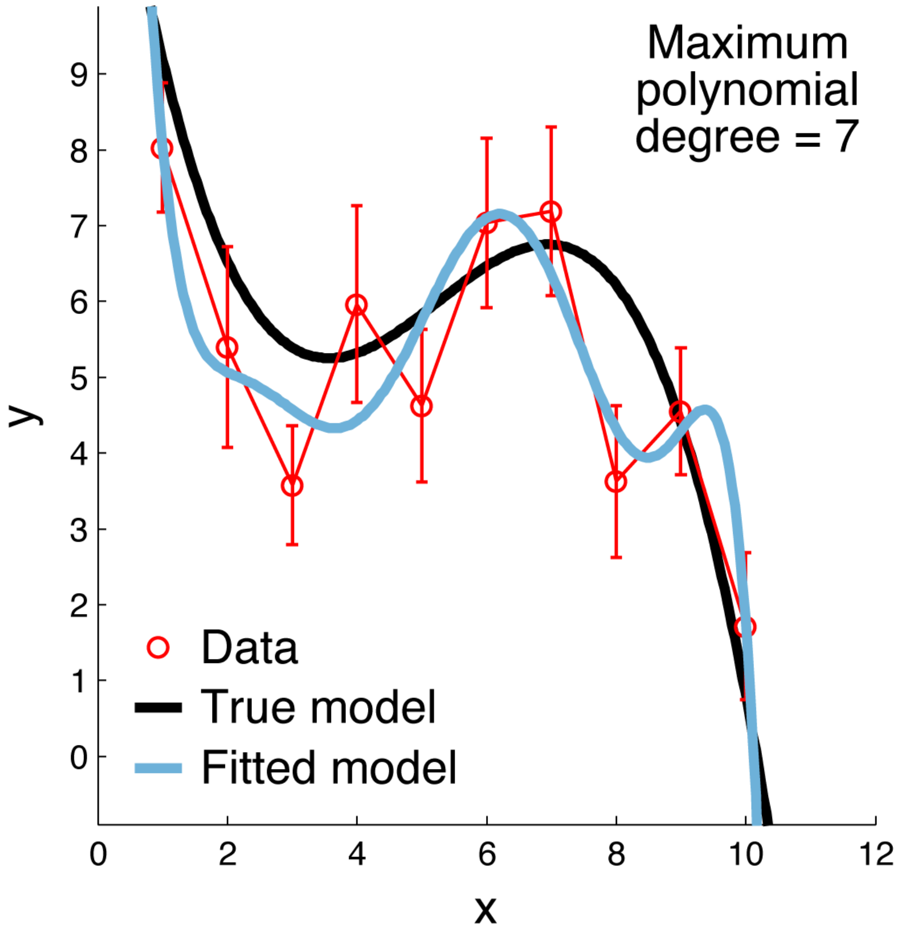
| | Population | Sample |
|---|---|---|
| True model | high (80%) | high (79%) |
| Fitted model | low (69%) | very high (85%) |

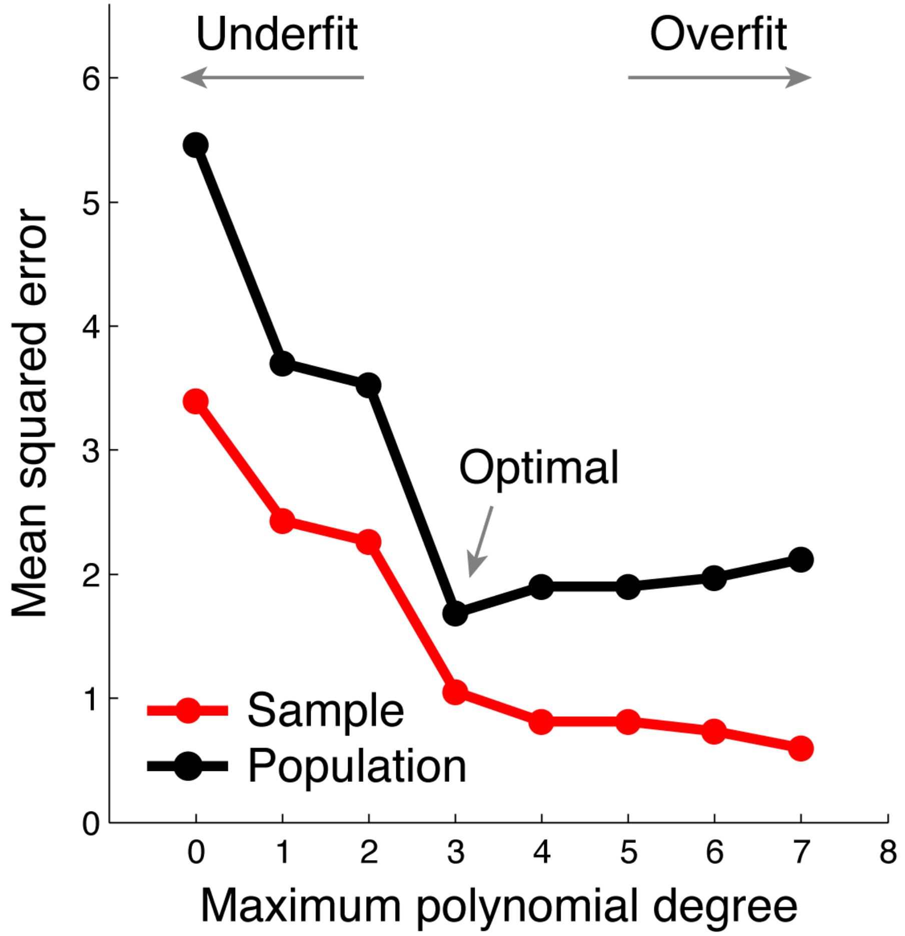
Accuracy of fitted model on sample overestimates true accuracy of fitted model

*Kendrick Kay, Washington Univ. in St. Louis*

# Overfitting

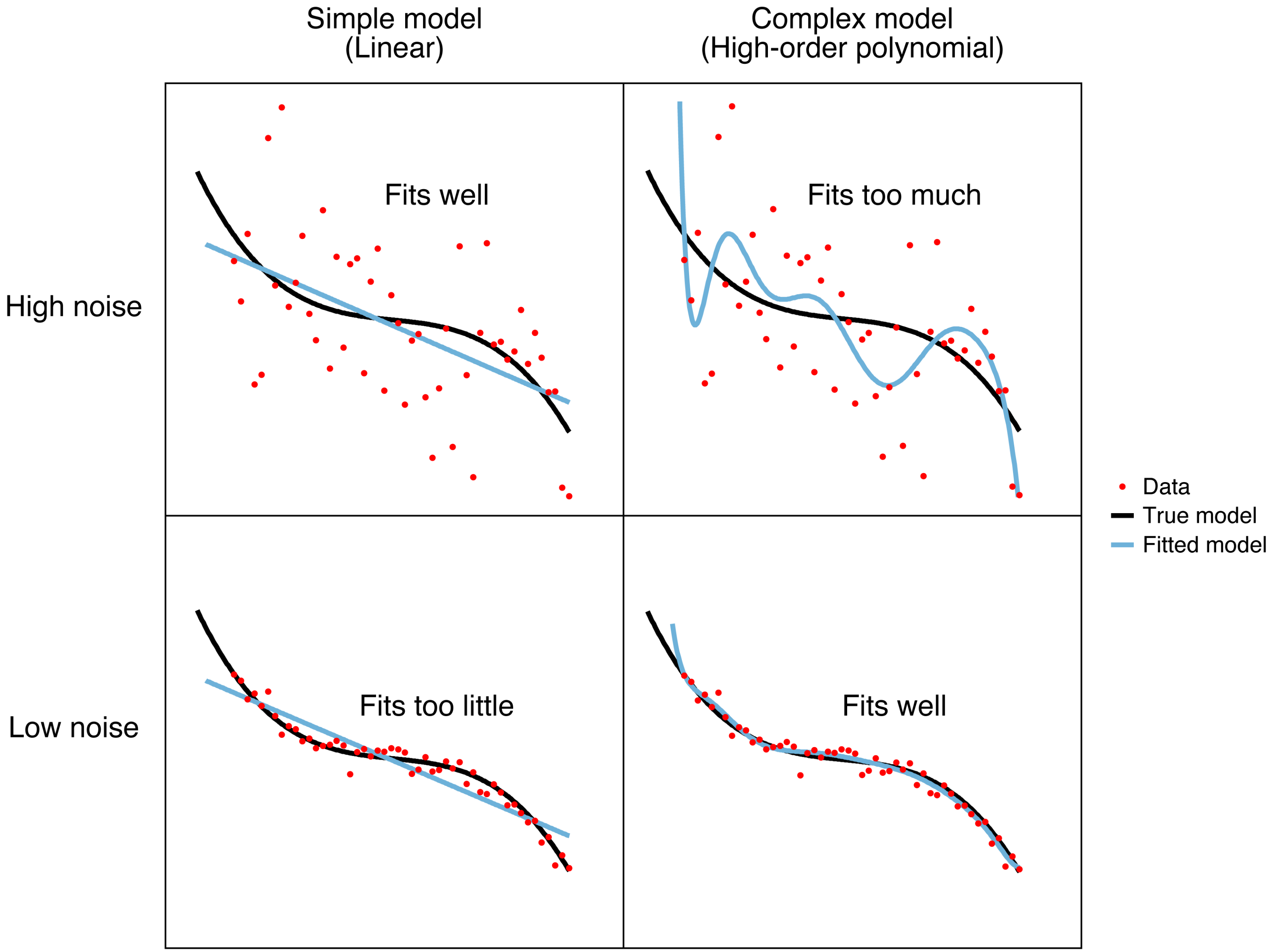
$$y = ax^7 + bx^6 + cx^5 + dx^4 + ex^3 + fx^2 + gx + h$$





*Kendrick Kay, Washington Univ. in St. Louis*

# Simple models vs. complex models



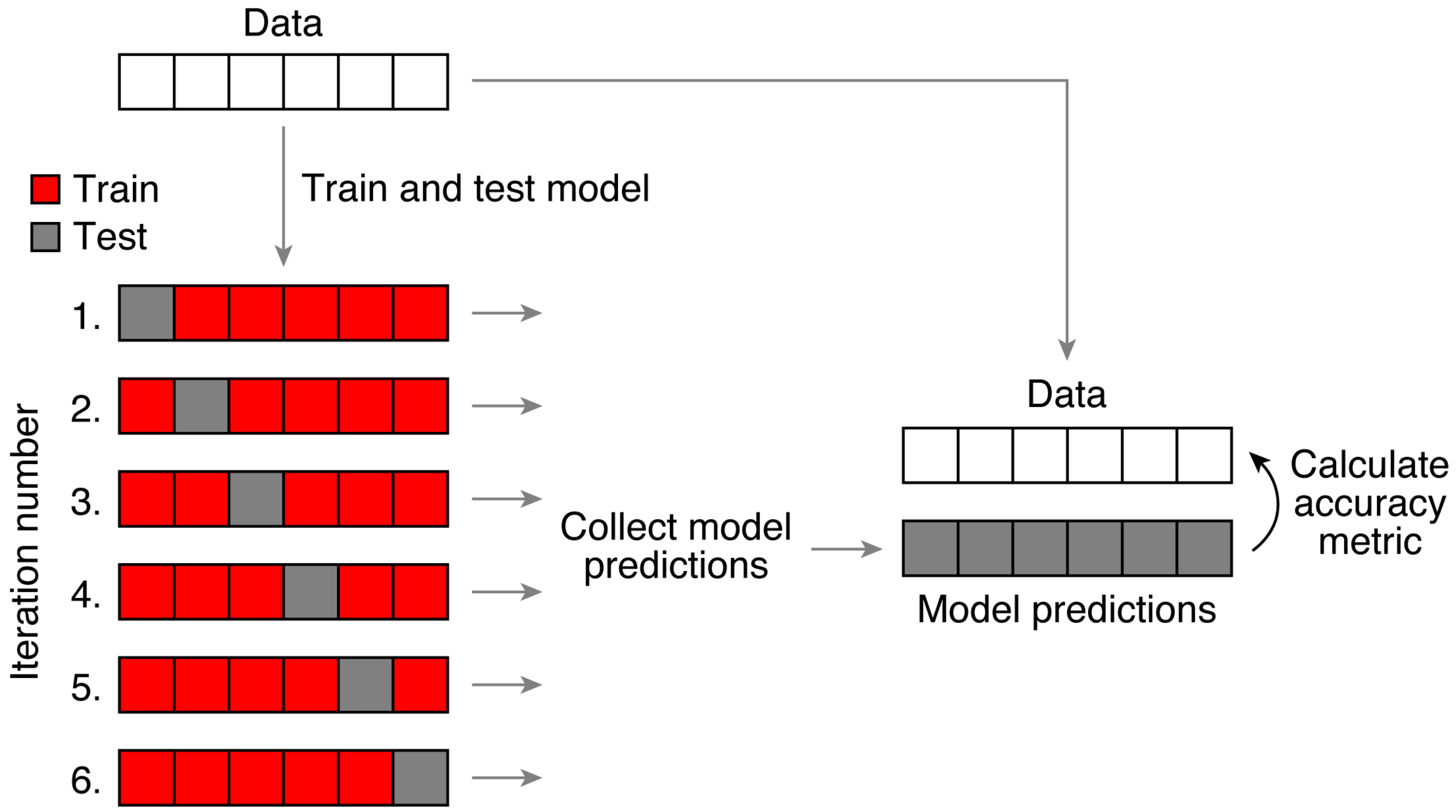
*Kendrick Kay, Washington Univ. in St. Louis*

## Cross-validation

- Goal: estimate true accuracy of a model
- Approach:

  Leave some data out<br>
  Fit model<br>
  Evaluate model on left-out data

# Leave-one-out cross-validation



Data

🟥 Train

■ Test

Train and test model

Iteration number

1.

2.

3.

4.

5.

6.

Collect model predictions

Data

Model predictions

Calculate accuracy metric

*Kendrick Kay, Washington Univ. in St. Louis*