DARTMOUTH

Feature selection

Lecture 9 of “Mathematics and AI”

Outline

1. Bias-variance tradeoff (revisited)
2. Feature selection
3. Model selection
4. Data leakage

The Bias-Variance Tradeoff

Bias-variance tradeoff

How sensitive should our model be to our training data?

Expected mean squared error

$$E[\mathrm{MSE}] = E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \mathrm{Var}[\hat{f}(x_0)] + \left[\mathrm{Bias}[\hat{f}(x_0)]\right]^2 + \mathrm{Var}[\varepsilon]$$

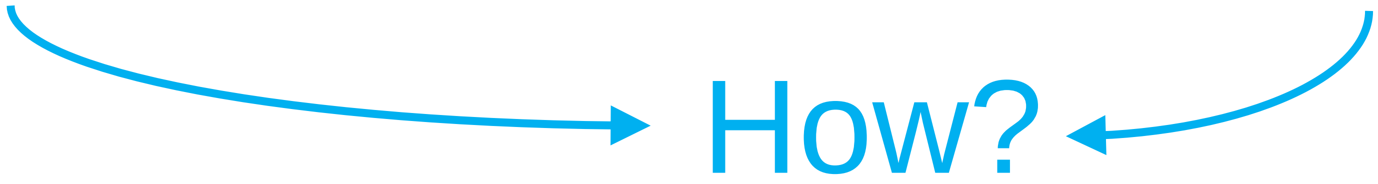
Bias-variance tradeoff

Large variance

- Model too sensitive to training data
- High training accuracy and low validation or test accuracy
- Reduce model complexity

Large bias

- Model not sensitive enough to training data
- Low training accuracy
- Increase model complexity



How?

Feature selection

Subset selection

- Best subset selection
  - Try all 2^p combinations of features (works only for small p)
- Stepwise forward selection
  - Start with 0 features, then add feature which gives best improvement, repeat
- Stepwise backward selection
  - Start with p features, then remove feature whose removal yields the smallest decrease of prediction accuracy, repeat

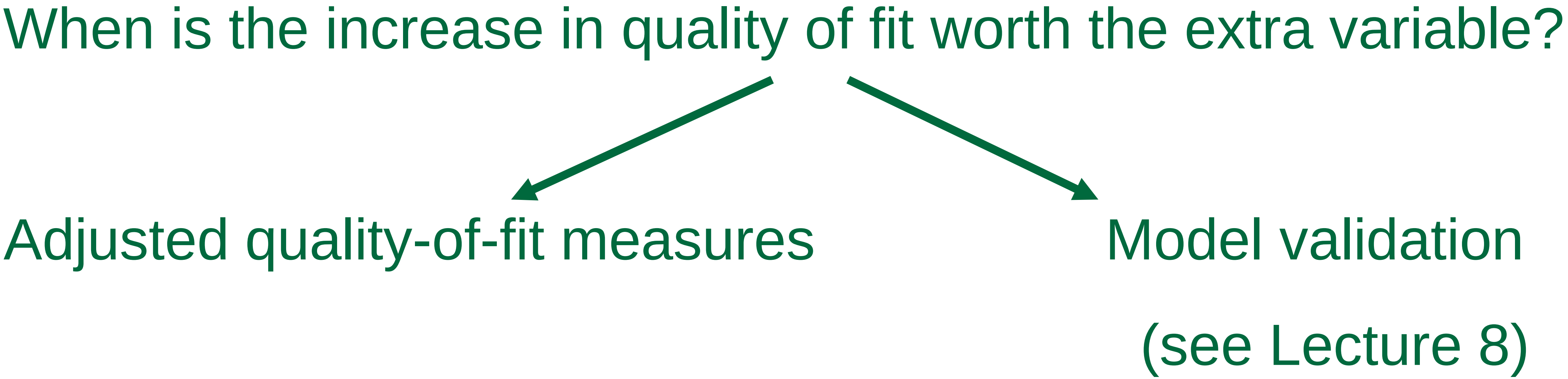
Subset selection

• Best subset selection	2^p candidate models
• Stepwise forward selection	p candidate models
• Stepwise backward selection	p candidate models
• Hybrid methods	p candidate models

Model selection

Model selection

Training error decreases with number of features.



When is the increase in quality of fit worth the extra variable?

Adjusted quality-of-fit measures

Model validation

(see Lecture 8)

Adjusted quality of fit measures

- Mallow's C_p
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- Adjusted R^2

Adjusted quality of fit measures

Measure of quality of fit	Calculation for linear regression	Calc. for logistic regression
Unadjusted quality of fit	$\text{MSE} = \frac{1}{n}\text{RSS}$ and $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$	$\hat{L}$ (estimated maximum likelihood)
Akaike information criterion (AIC)	$\text{AIC} = \frac{1}{n}(\text{RSS} + 2d\hat{\sigma}^2)$	$\text{AIC} = -2\log\hat{L} + 2d$
Bayesian information criterion (BIC)	$\text{BIC} = \frac{1}{n}(\text{RSS} + \log(n)\, d\hat{\sigma}^2)$	$\text{BIC} = -2\log\hat{L} + \log(n)\, d$
Adjusted R^2	$\text{Adj } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$	n/a

Model selection

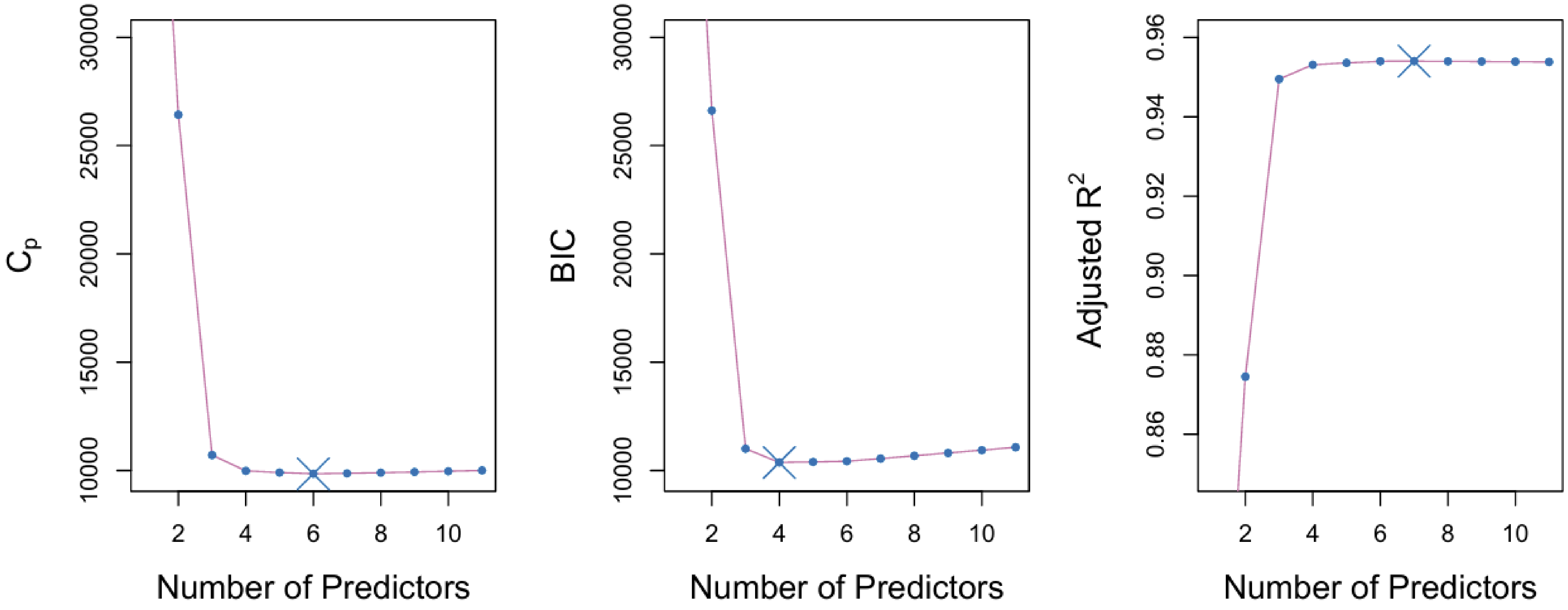


Image source: Introduction to Statistical Learning in Python

Model selection

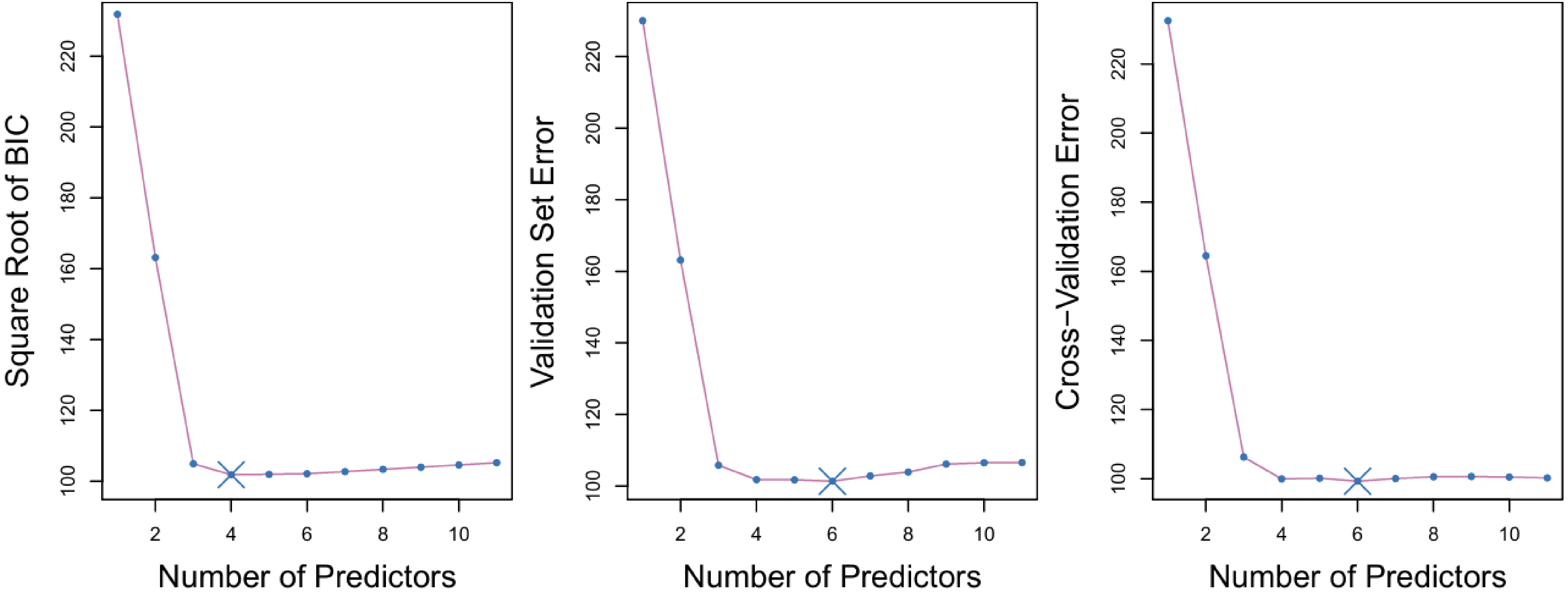
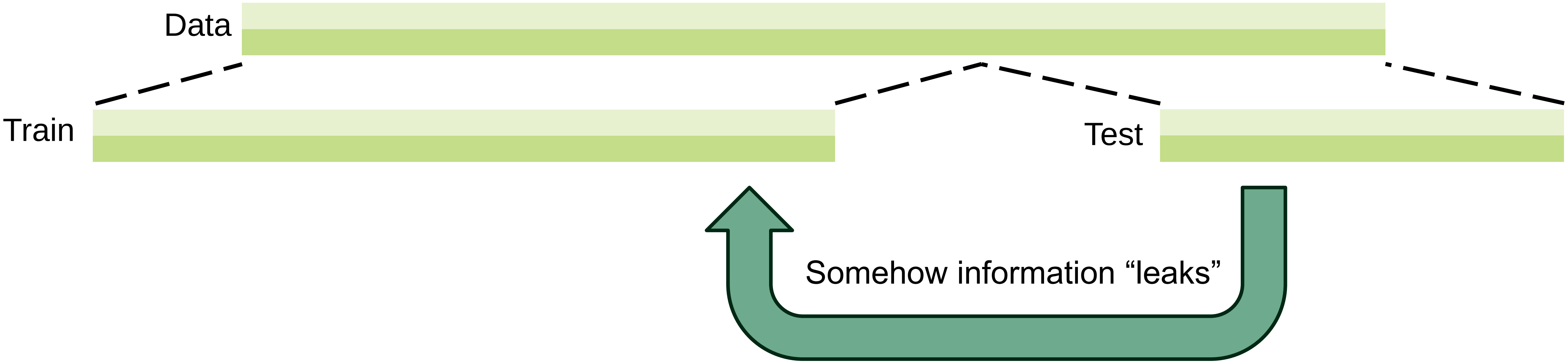


Image source: Introduction to Statistical Learning in Python

Data leakage

“Data leakage can be [a] multi-million dollar mistake in many data science applications.”

Dan Becker (Kaggle Instructor)



Examples:

- Observations of the same test subject in train and test set (compare "eigenfaces" example)
- Conducted feature selection or data imputation on the full data set