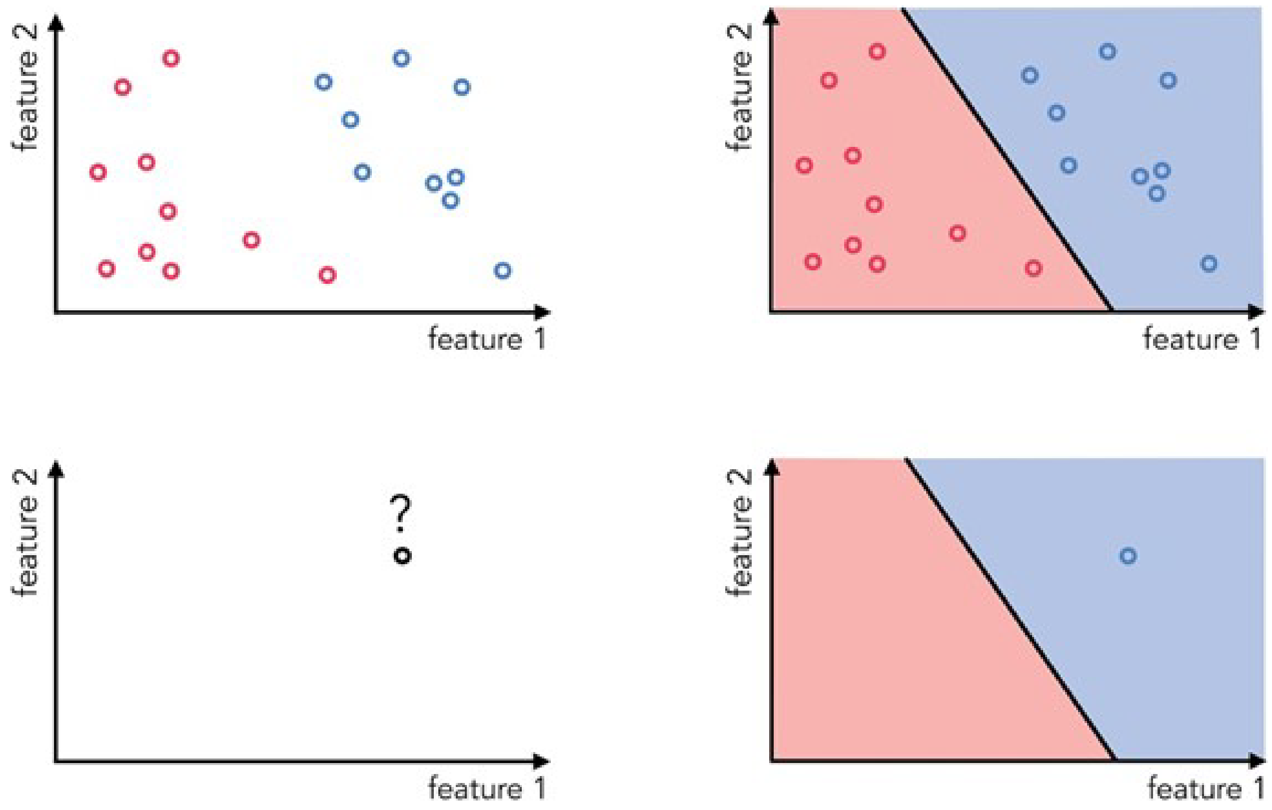
Machine learning primer

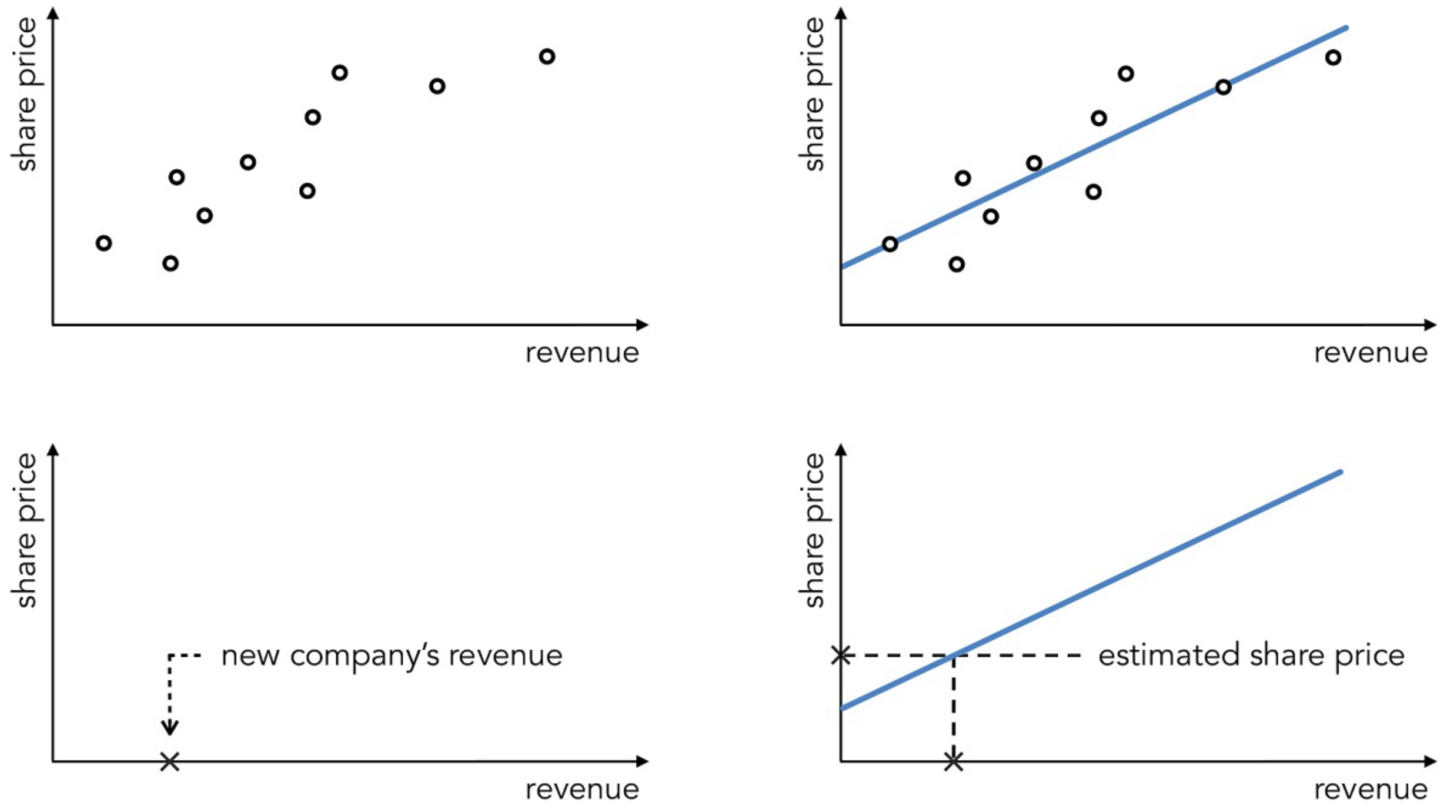
JOANNA BYSZUK & JEREMI OCHAB

DHSI 2024, “DIY COMPUTATIONAL TEXT ANALYSIS WITH R”

Task examples: classification



Task examples: regression



Terms and definitions

Example: item, instance of the data used.

Features: attributes associated to an item, often represented as a vector (e.g., word counts).

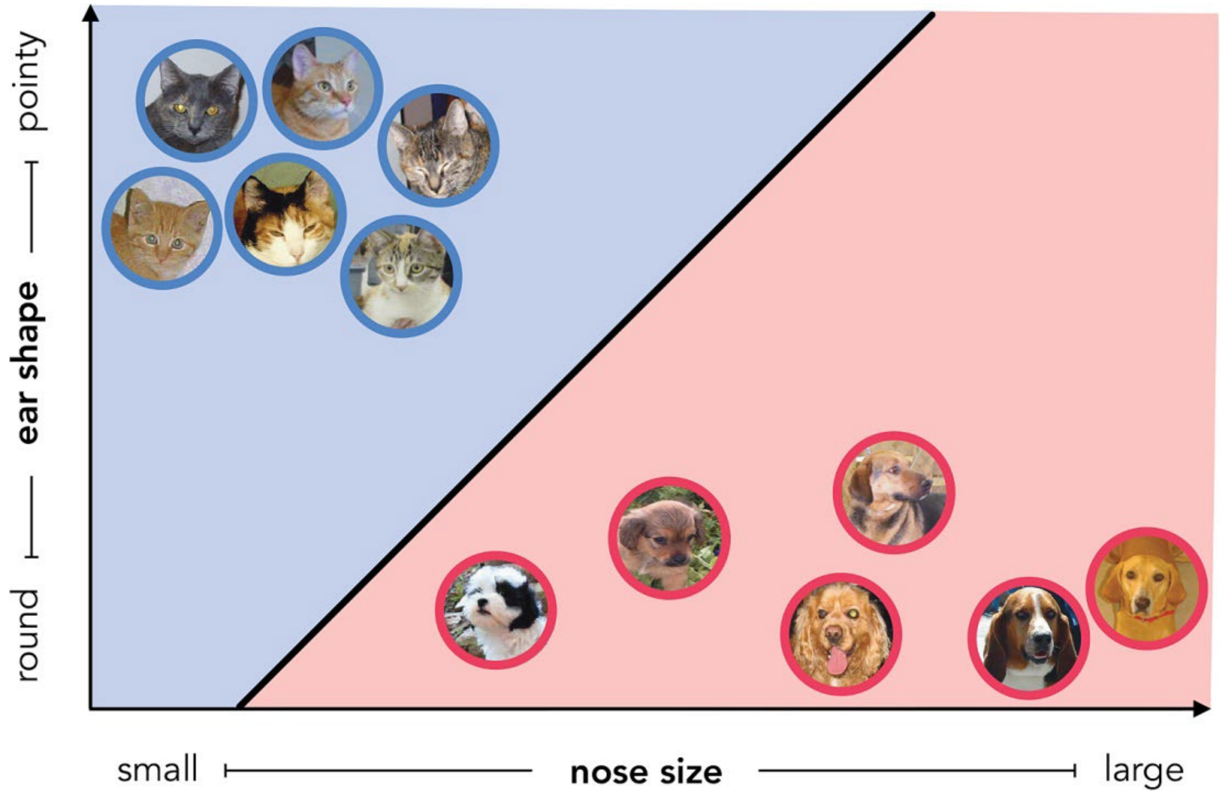
Labels: category (classification) or real value (regression) associated to an item.

Data:

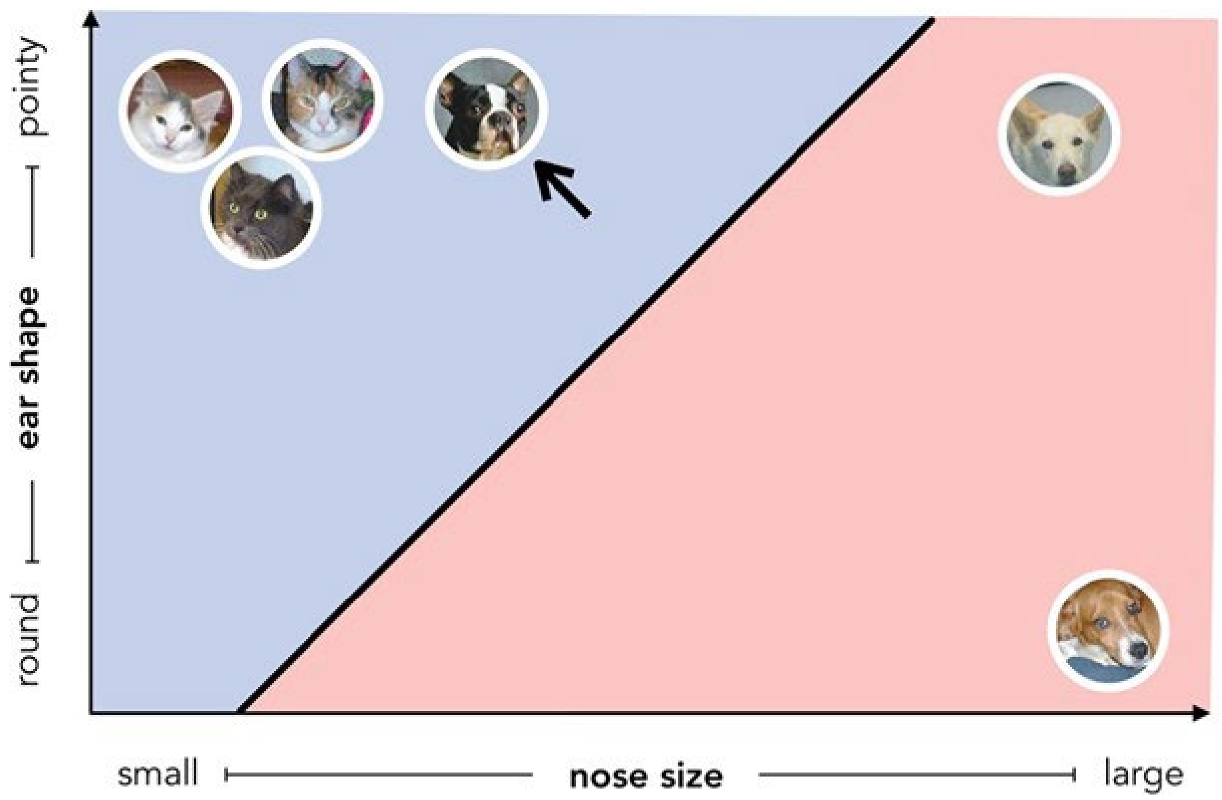
- training data (typically labeled).
- test data (labeled but labels not seen).
- validation data (labeled, for tuning parameters).

Learning stages

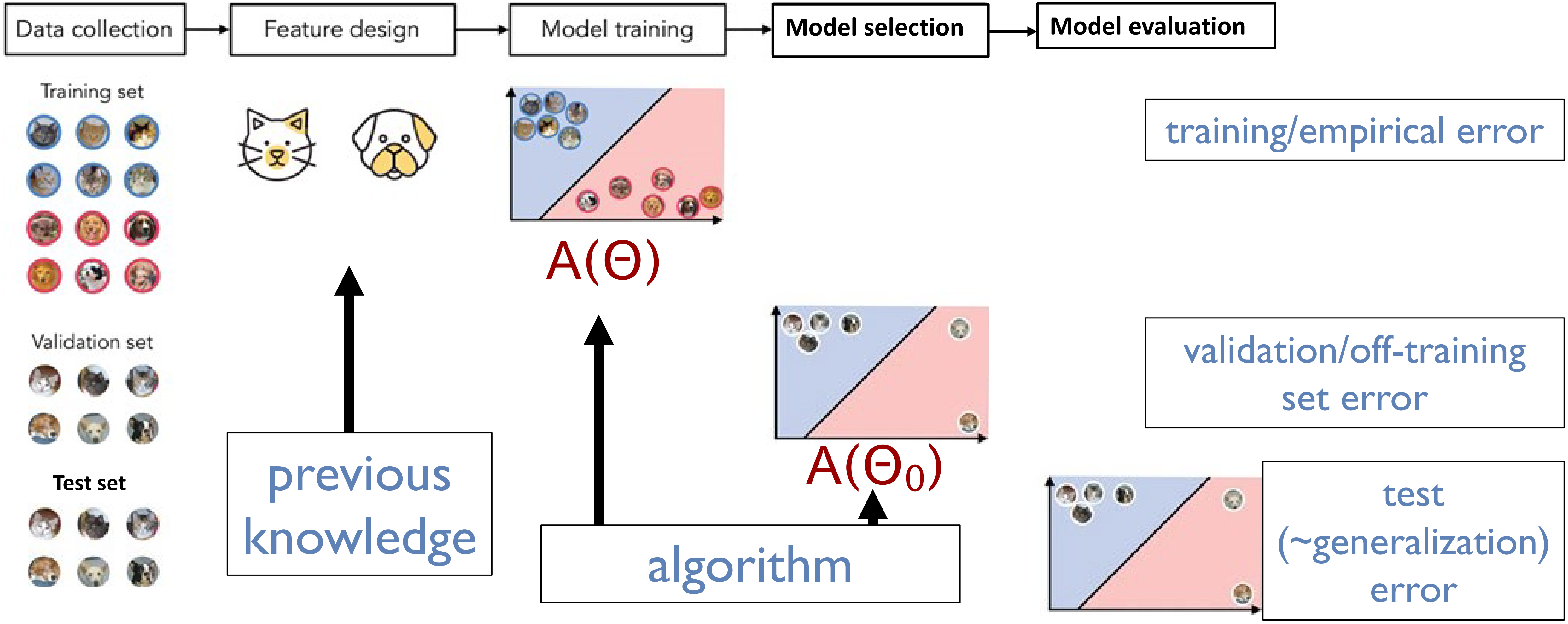
Learning stages



Learning stages



Learning stages



J. Watt, R. Borhani, A.K. Katsaggelos. *Machine Learning Refined*. © Cambridge University Press 2020

Bias-variance tradeoff

Bias-variance tradeoff

Najłatwiej na przykładzie dopasowania krzywej regresji:

- *F*(**x**) – a true (but unknown) function with continuous valued output with noise
- *D* – set of *n* training samples generated by F(**x**)
- *g*(**x**; *D*) – the estimated regressions function F (depends on the set *D*!)
- Estimator effectiveness: average (over all sets *D* of size *n*) mean-square deviation:

- bias – the difference between the true (but unknown) value and our expectations [=estimation accuracy]
- variance – the instability of the estimate due to the variability of the training set

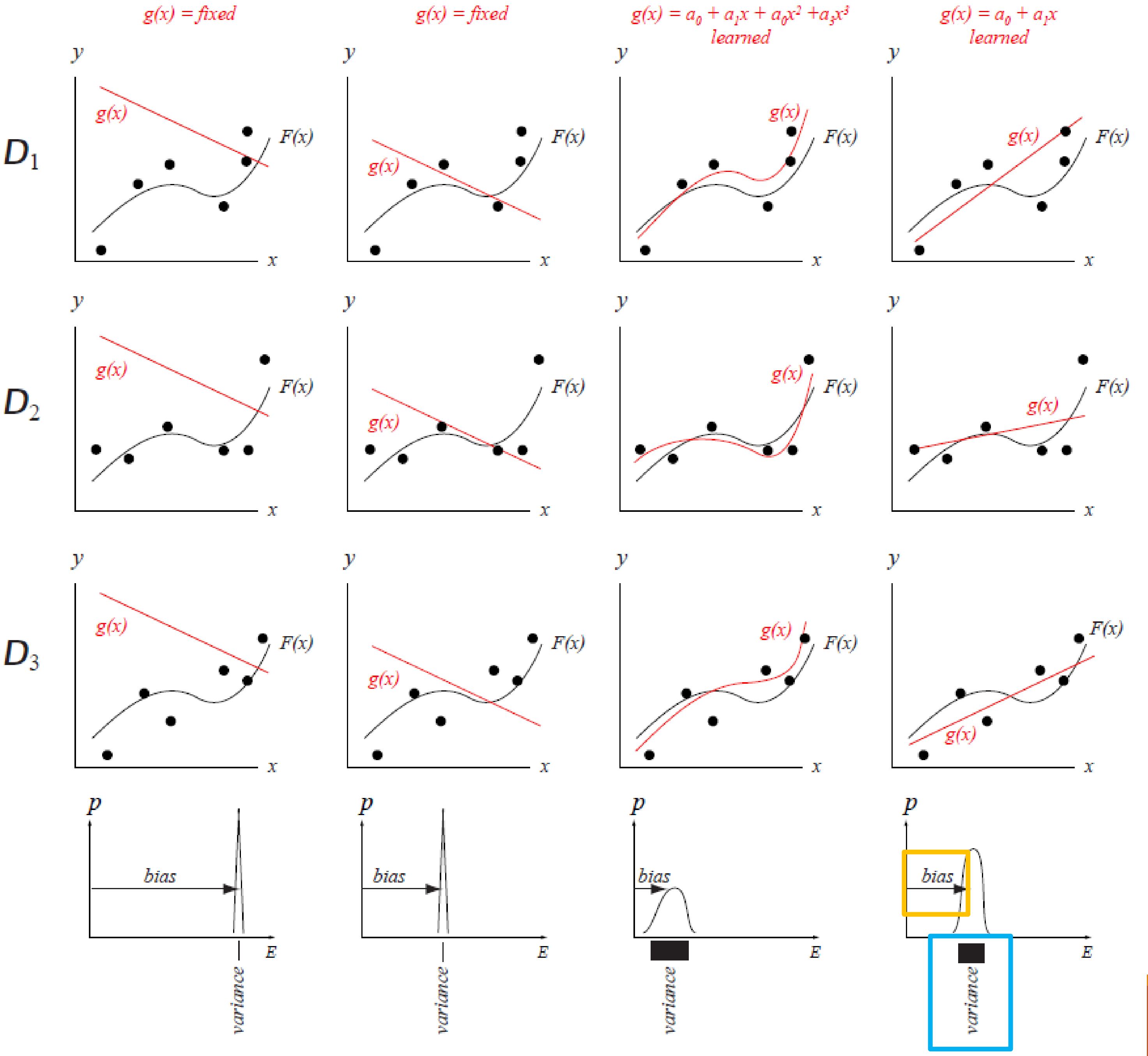
$$\mathcal{E}_{\mathcal{D}}\left[(g(\mathbf{x};\,\mathcal{D}) - F(\mathbf{x}))^2\right]$$

$$= \boxed{\underbrace{(\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\,\mathcal{D}) - F(\mathbf{x})])^2}_{bias^2}} + \boxed{\underbrace{\mathcal{E}_{\mathcal{D}}\left[(g(\mathbf{x};\,\mathcal{D}) - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\,\mathcal{D})])^2\right]}_{variance}}.$$

Duda, Hart, Stork. *Pattern Classification, 2nd ed*. Wiley (2000).

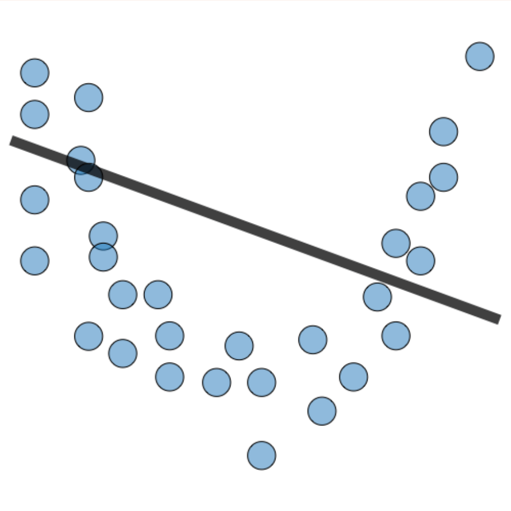
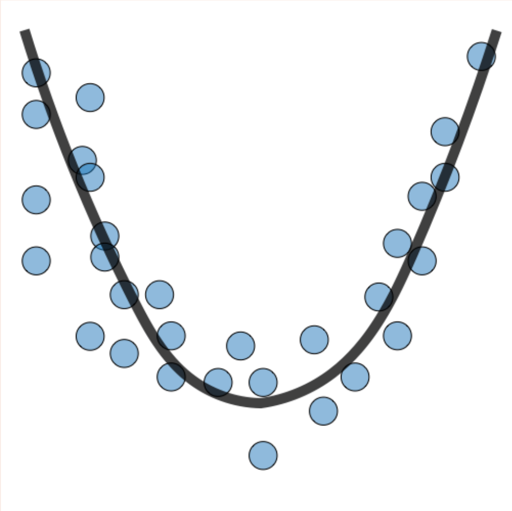
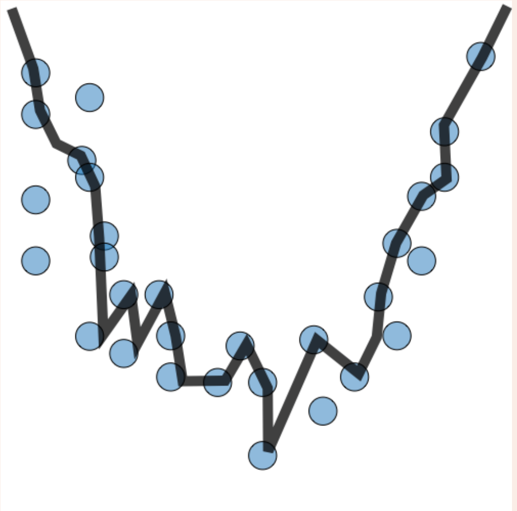
E – mean squared error

p – probability that we will randomly get D with the error E



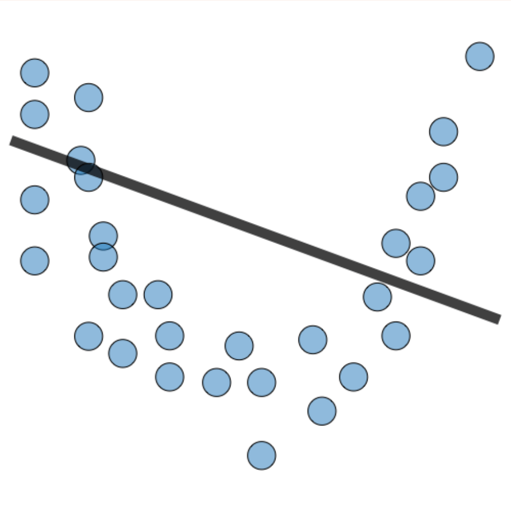
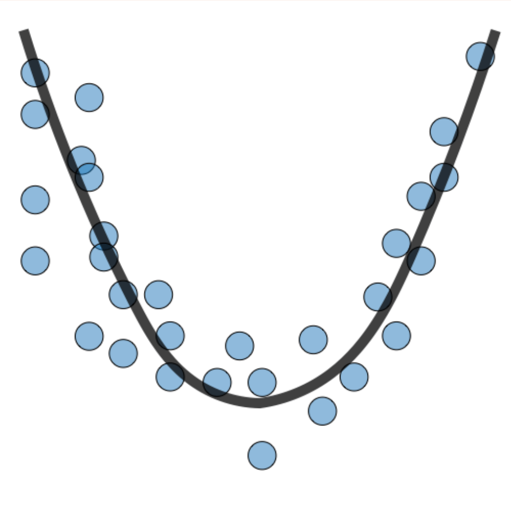
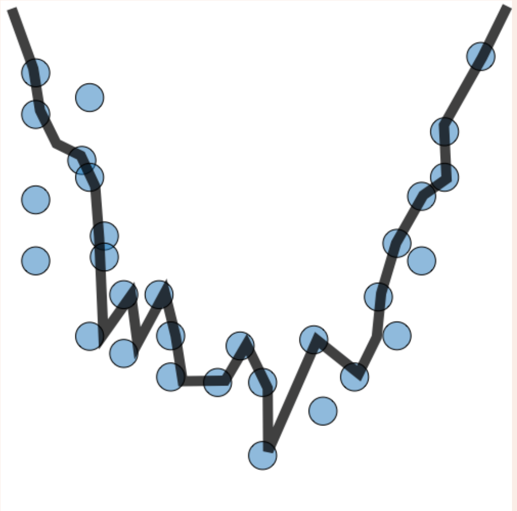
Duda, Hart, Stork. *Pattern Classification, 2nd ed*. Wiley (2000).

Bias-variance tradeoff

Regression			

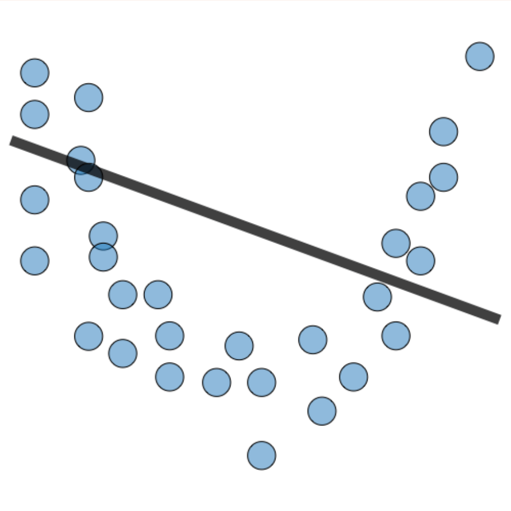
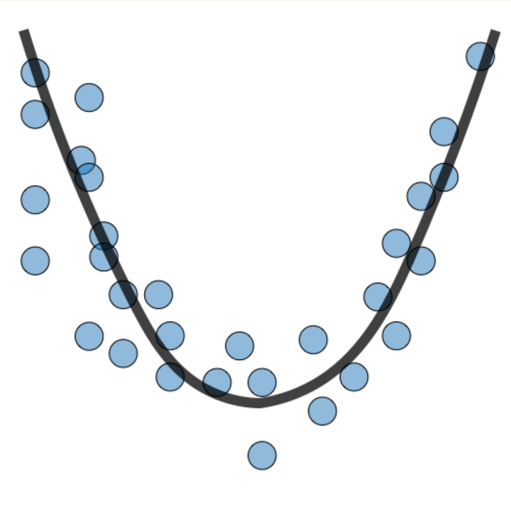
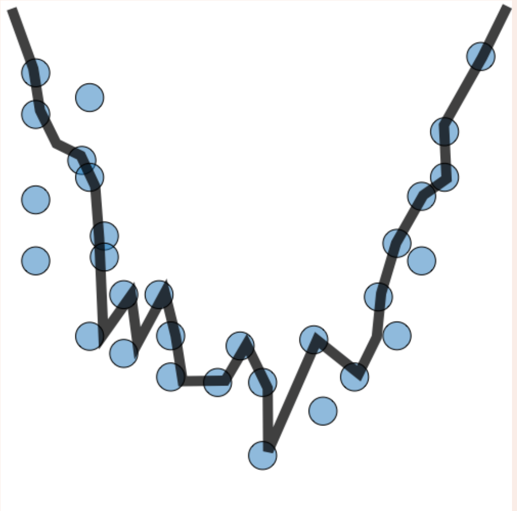
Afshine Amidi, Shervine Amidi. *Machine Learning cheatsheets for Stanford's CS 229*. (2018).

Bias-variance tradeoff

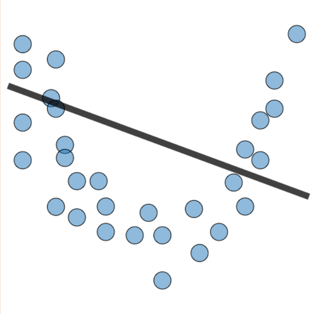
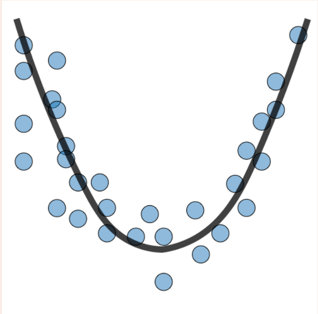
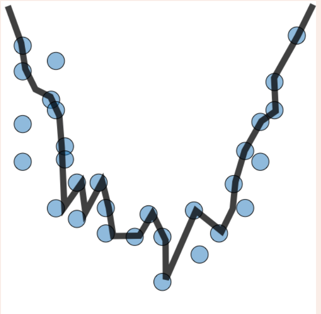
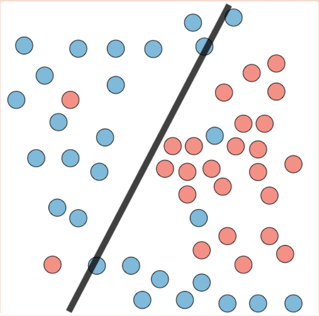
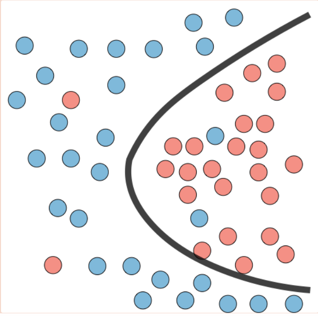
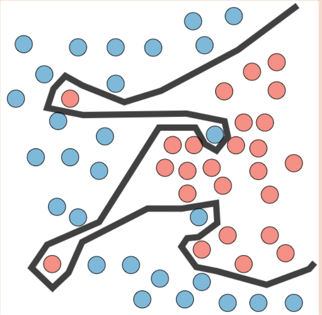
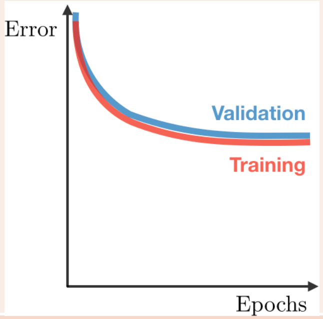
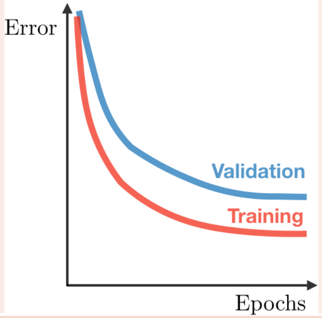
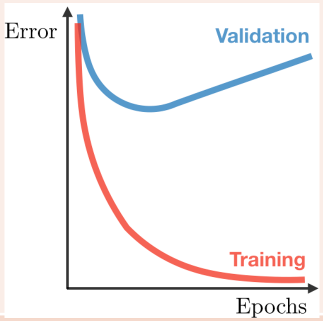
	Underfitting	Just right	Overfitting
Regression			

Afshine Amidi, Shervine Amidi. *Machine Learning cheatsheets for Stanford's CS 229*. (2018).

Bias-variance tradeoff

	Underfitting	Just right	Overfitting
Symptoms	▪ High training error ▪ Training error close to test error ▪ High bias	▪ Training error slightly lower than test error	▪ Low training error ▪ Training error much lower than test error ▪ High variance
Regression			

Afshine Amidi, Shervine Amidi. *Machine Learning cheatsheets for Stanford's CS 229*. (2018).

	Underfitting	W sam raz	Overfitting
Symptoms	▪ High training error ▪ Training error close to test error ▪ High bias	▪ Training error slightly lower than test error	▪ Low training error ▪ Training error much lower than test error ▪ High variance
Regression			
Classification			
Deep learning			
Remedies?	▪ complexify model ▪ Add more features ▪ Train longer		▪ Regularise ▪ Get more data

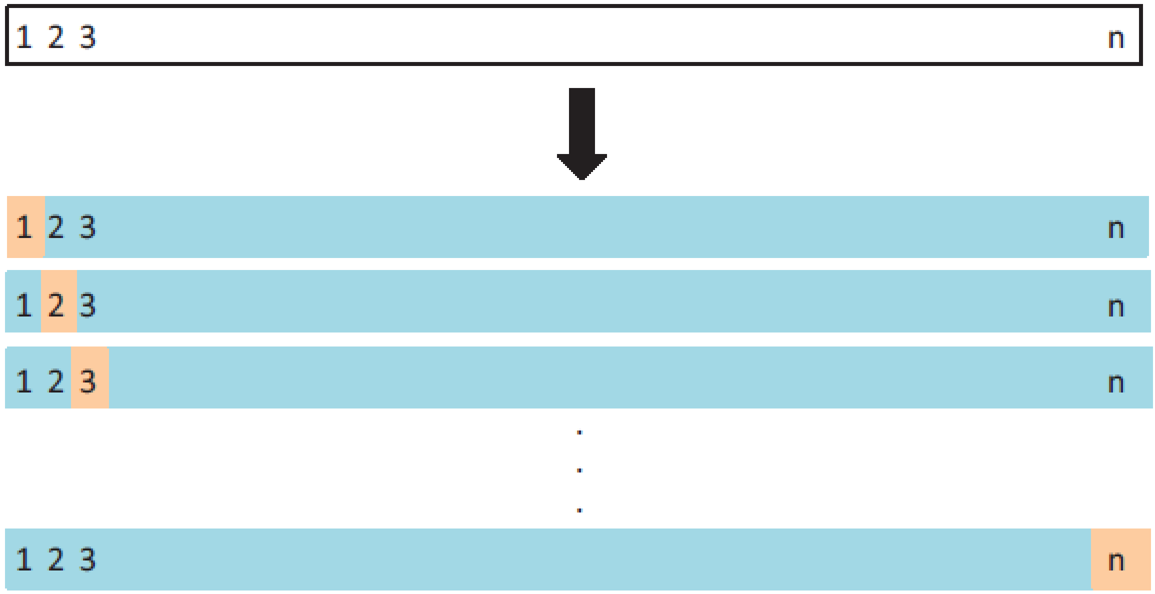
Generalisation

OBSERVATIONS

- the best hypothesis on the sample may not be the best overall.
- generalization is not memorization.
- complex rules (very complex separation surfaces) can be poor predictors.
- trade-off: complexity of hypothesis set vs sample size (underfitting/overfitting).

Cross-validation

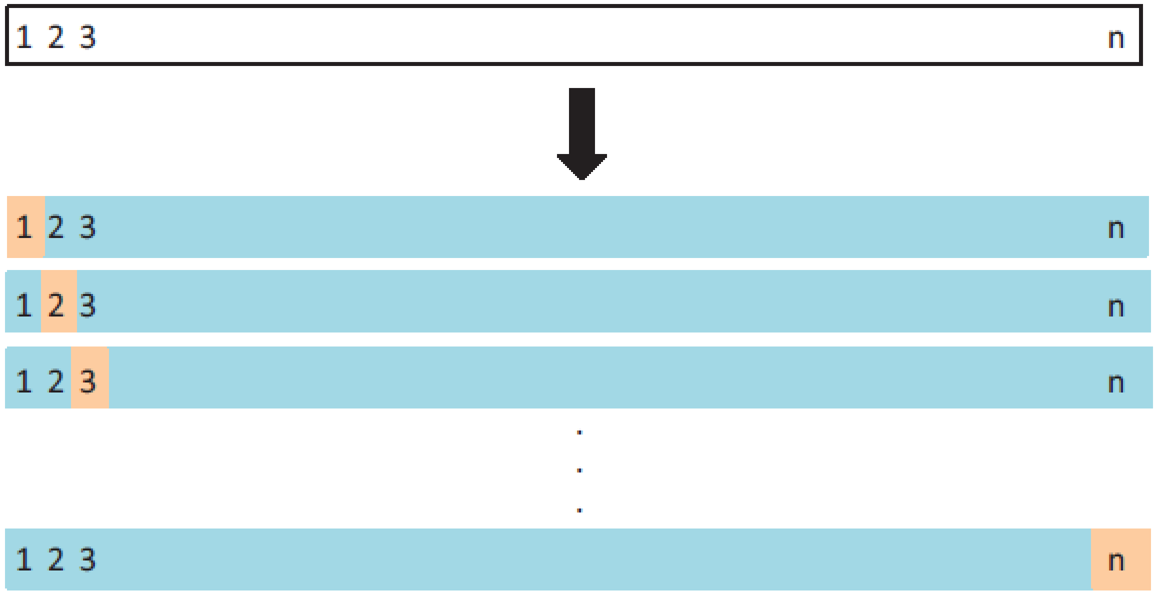
LEAVE-ONE-OUT CROSS-VALIDATION [JACK-KNIFE]



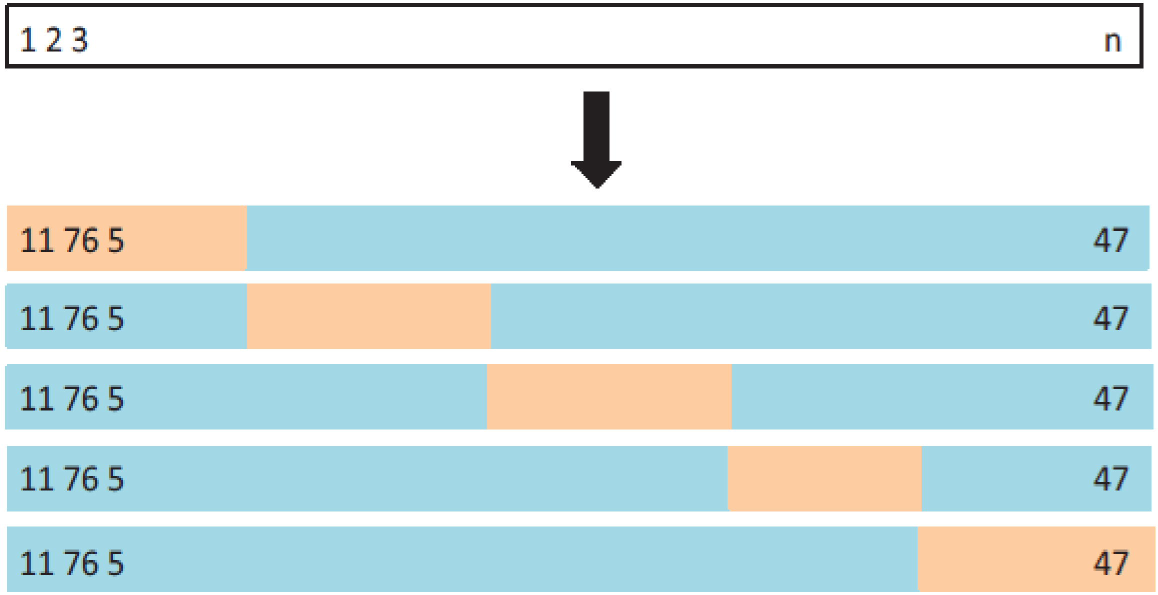
James, Witten, Hastie, Tibshirani, *An Introduction to Statistical Learning,* Springer (2013).

Cross-validation

LEAVE-ONE-OUT CROSS-VALIDATION [JACK-KNIFE]



K-FOLD CROSS-VALIDATION



James, Witten, Hastie, Tibshirani, *An Introduction to Statistical Learning,* Springer (2013).

Cross-validation

K-FOLD CROSS-VALIDATION

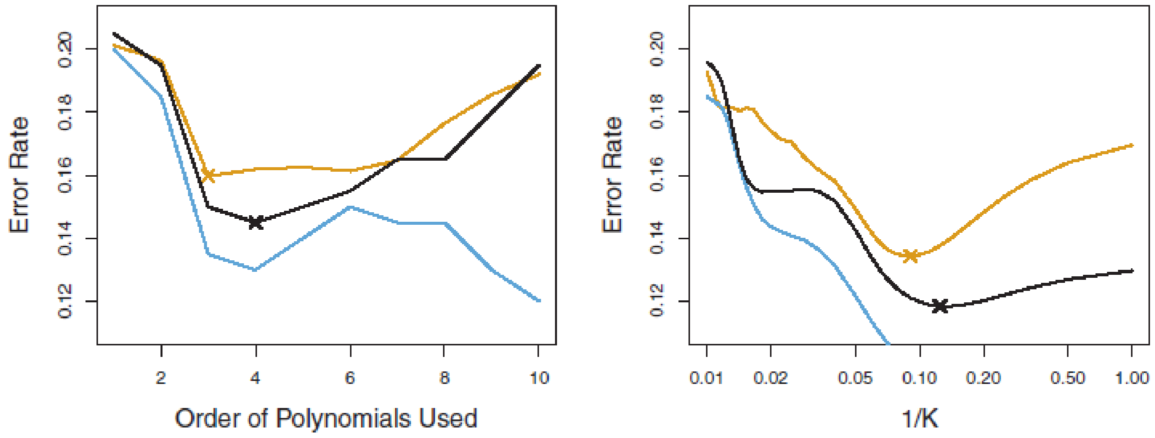
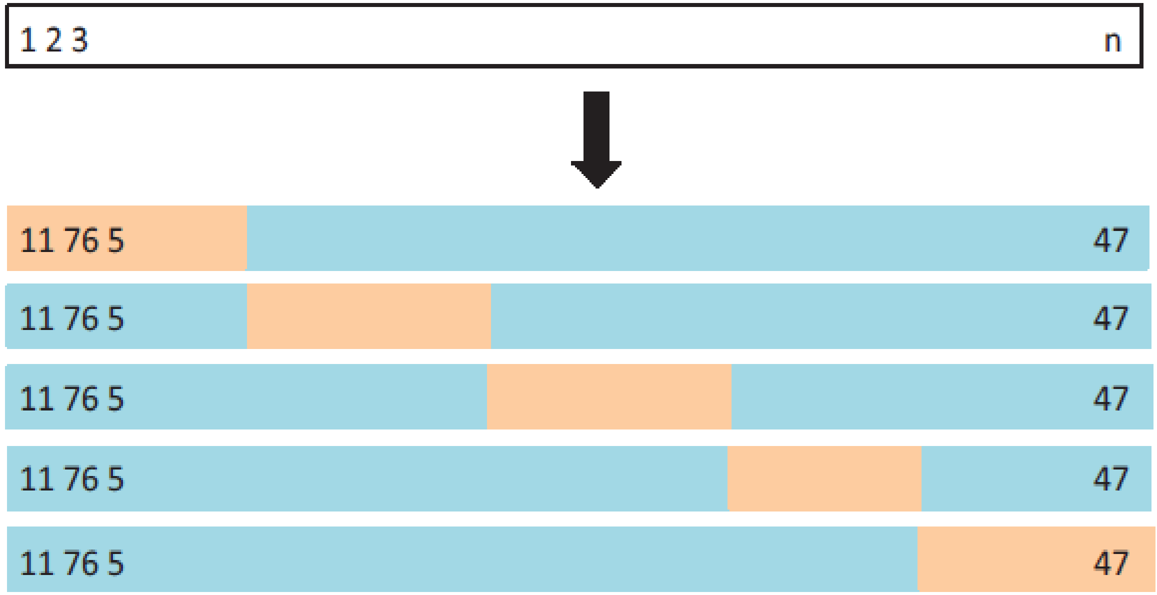


FIGURE 5.8. *Test error (brown), training error (blue), and 10-fold CV error (black) on the two-dimensional classification data displayed in Figure 5.7.* Left: *Logistic regression using polynomial functions of the predictors. The order of the polynomials used is displayed on the x-axis.* Right: *The KNN classifier with different values of K, the number of neighbors used in the KNN classifier.*



James, Witten, Hastie, Tibshirani, *An Introduction to Statistical Learning,* Springer (2013).

Cross-validation

MORE TYPES

- Stratified CV
- Group CV

https://scikit-learn.org/stable/modules/cross_validation.html#k-fold

stratified