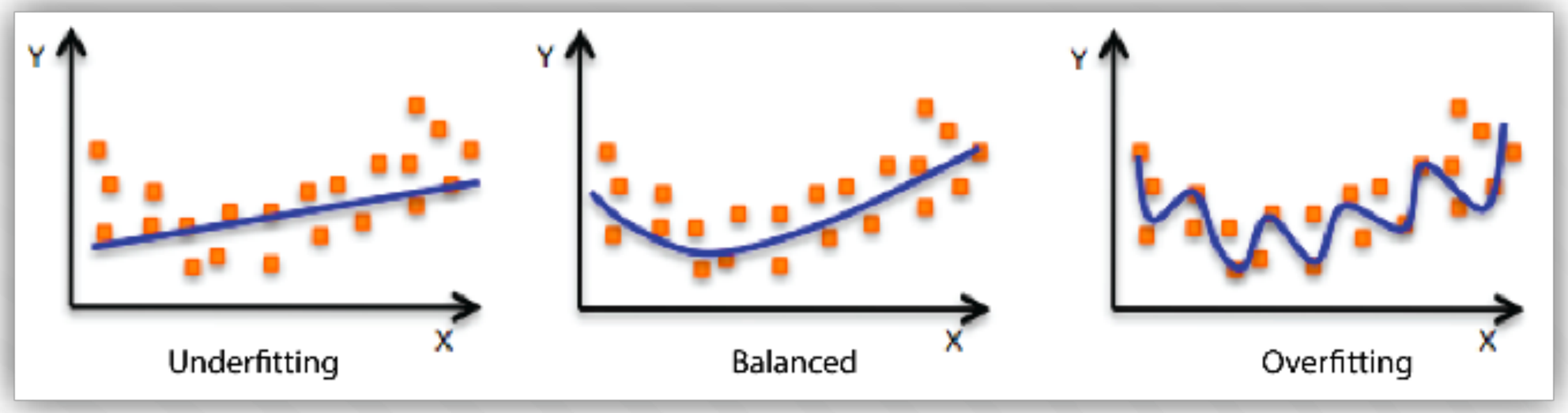
أكاديمية طويق
TUWAIQ ACADEMY

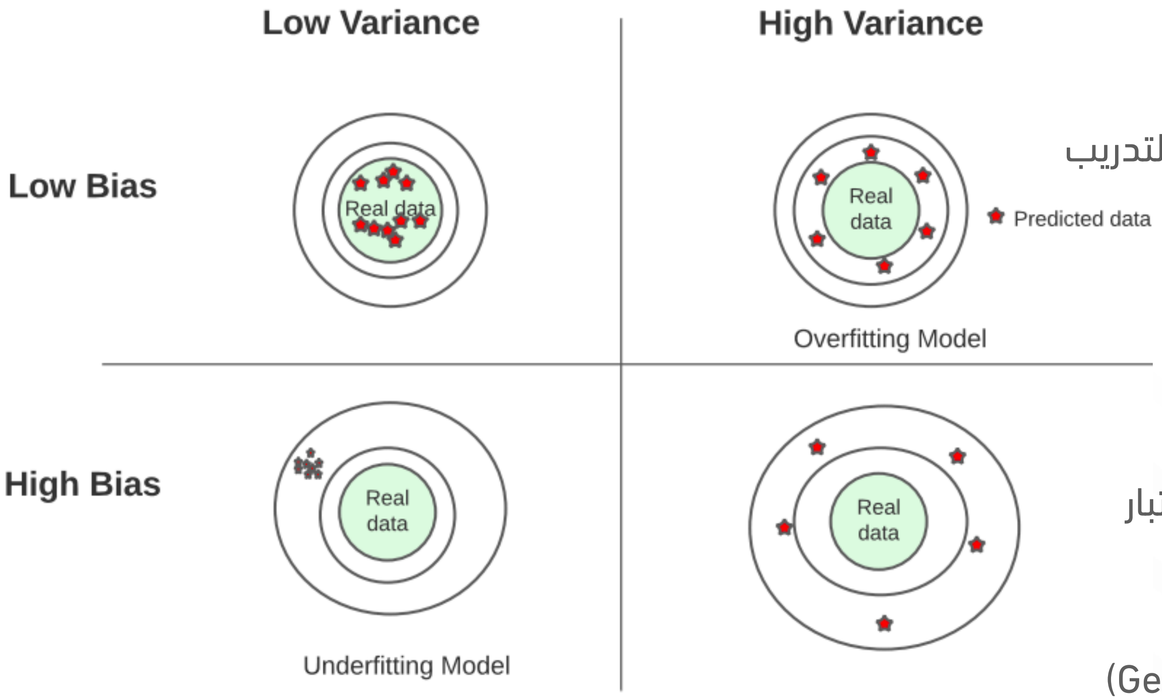
مقدمة في تعلم الآلة

مفهوم Overfitting - Underfitting

- في Machine Leaning يتم تقييم أداء النموذج على أساس:
  - الدقة Accuracy و تعني مدى دقة النموذج في توقع القيم الصحيحة.
  - التعميم Generalization ويعني مدى قدرة النموذج في توقع البيانات الجديدة بشكل دقيق كما في البيانات التي تعلمها سابقا.
- يتم تقسيم البيانات لمجموعات تدريب ومجموعات اختبار ويتم اعتبار النموذج دقيق إذا كانت دقته متقاربة في كلا المجموعتين.



Overfitting - Underfitting مفهوم



أسباب ظهور Underfitting - (low variance, high bias)

- عندما تكون دقة النموذج ضعيفة في بيانات التدريب والاختبار
- بساطة النموذج وعدم قدرته على تعلم complex pattern والعلاقات من بيانات التدريب
- قلة عدد features
- قلة عدد البيانات

يتم حل المشكلة عن طريق:

إضافة features عن طريق Feature Selection

أسباب ظهور Overfitting - (low bias, high variance)

- يحدث عندما تكون دقة النموذج عالية في بيانات التدريب و ضعيفة في بيانات الاختبار
- عند بناء نموذج معقد جدا بحيث يتم المبالغة بتعليمة إلى أن يحفظ جميع pattern الموجودة في بيانات التدريب
- عندما تتعلم الآلة بطريقة حفظ البيانات وبالتالي لاتستطيع التعميم (Generalization)
- عندما تحفظ الخوارزمية البيانات الشاذة (captured the noise of the data)

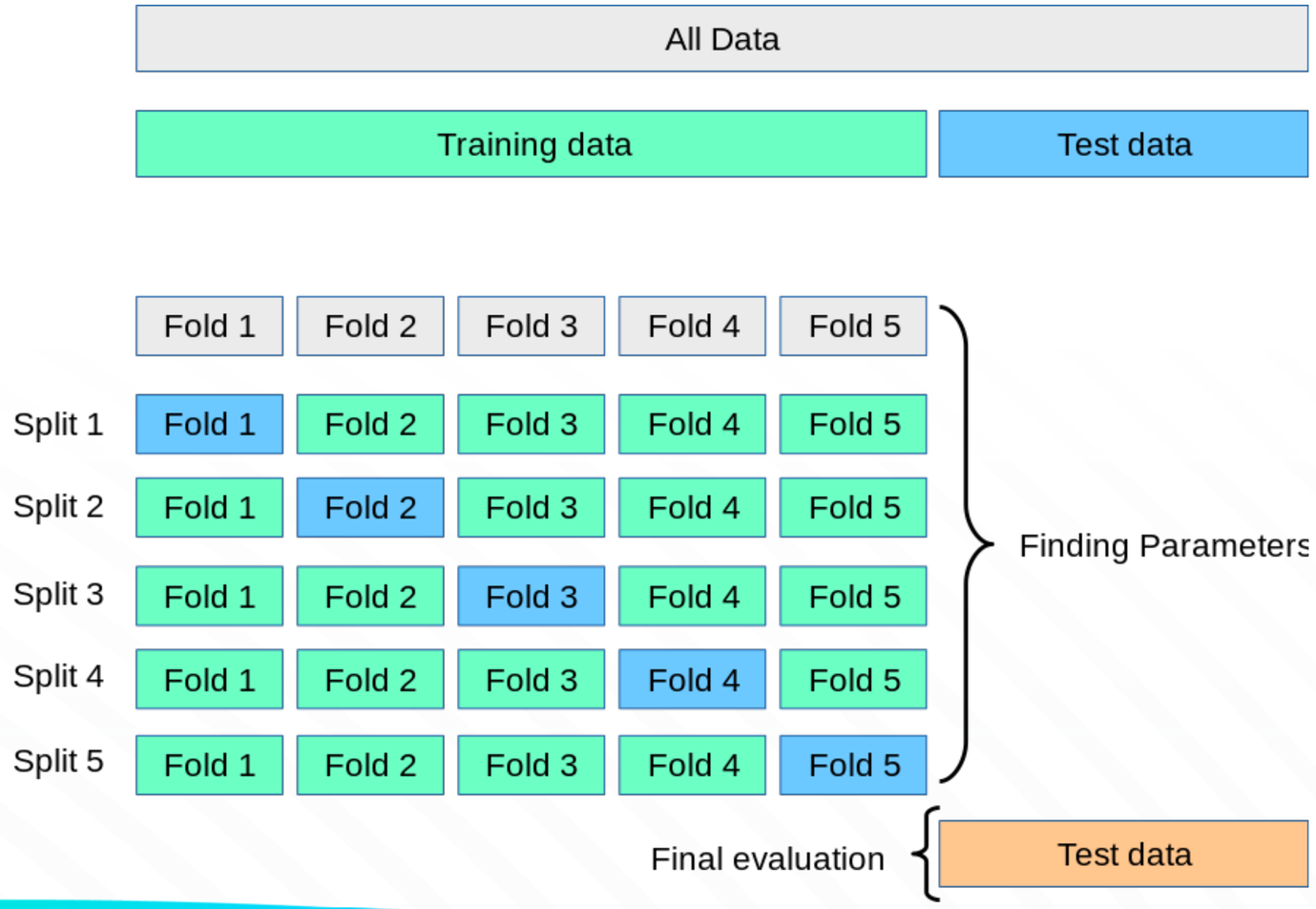
يتم حل المشكلة عن طريق:

استخدام k-fold cross validation

تقسيم البيانات (Cross-Validation)

- تقسيم البيانات بشكل عشوائى إلى عدة مجموعات متساوية تسمى folds ونرمز لعدد folds ب k
- مثال: عندما يتم تقسيم البيانات إلى k = 5 فهذا يعني وجود خمسة مجموعات بحيث يتم استخدام أربعة مجموعات منها في مرحلة التدريب والمجموعة الخامسة في مرحلة الاختبار.

تقسيم البيانات (Cross-Validation)



Resources

- Introduction to Data Science [https://link.springer.com/book/10.1007/978-3-319-50017-1].
- Data Mining: Concepts and Techniques [https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques].