Introduction (Cont.)

Petchara Pattarakijwanich

Introduction to Data Science, 26 August 2022

Types of Data Science and Machine Learning

- Supervised Learning

  ต้องรู้ทายอะไร ทายอย่างไร (รู้เป้าหมาย)

  (training data สร้างโมเดล ⇒ ใช้ test data ทดสอบโมเดล ⇒ ใช้งาน)

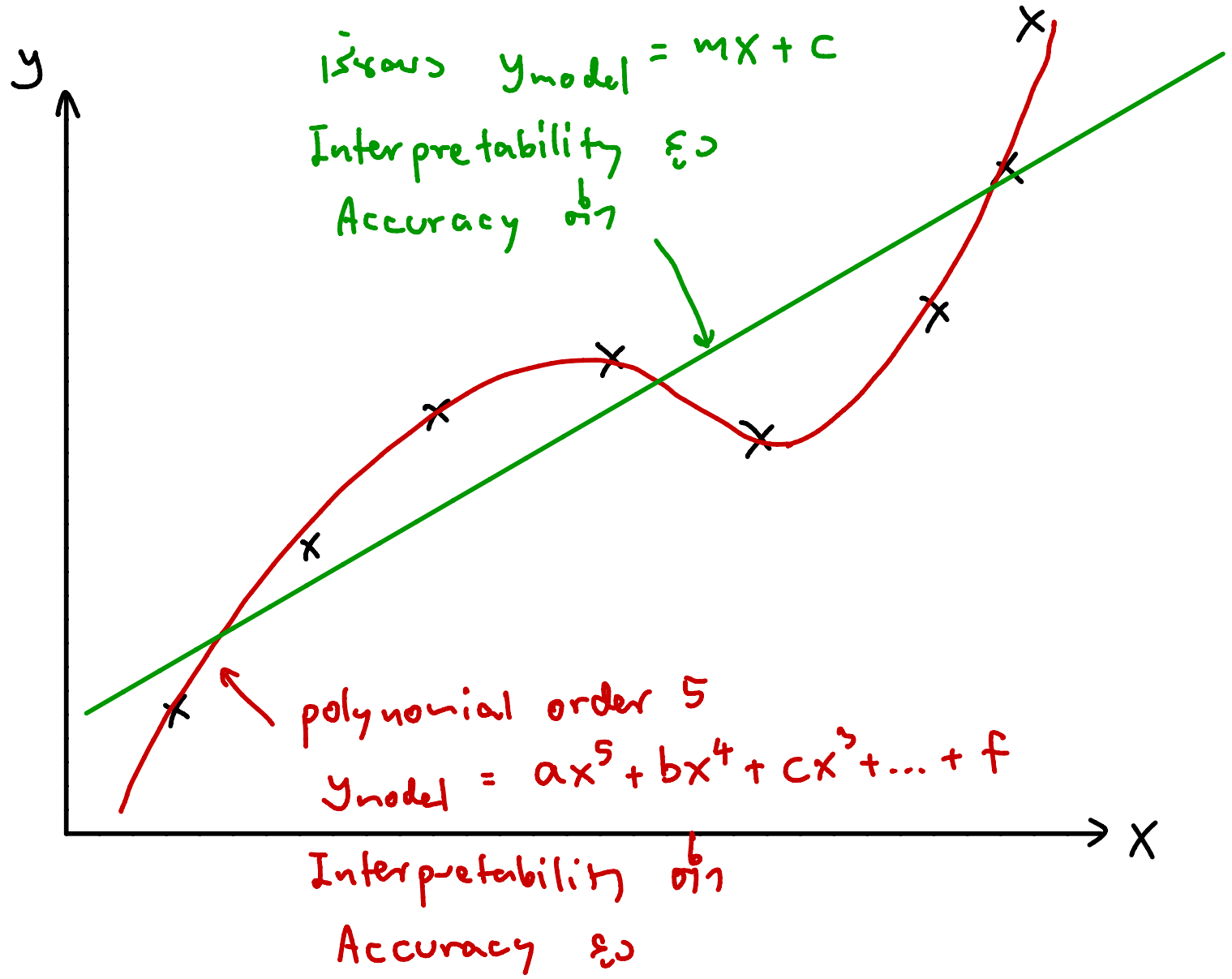
  - Regression ถ้าสิ่งที่ต้องการทาย เป็นตัวแปรต่อเนื่อง (เป็นตัวเลข) ↖ Quantitative
  - Classification ถ้าสิ่งที่ต้องการทาย เป็นตัวแปรไม่ต่อเนื่อง (เป็นกลุ่มๆ) ↖ Qualitative

- Unsupervised Learning

  [Data Exploration]

  - Visualization
  - Correlation
  - Clustering

Interpretability vs Accuracy



Methodology of Data Science

- Training Data

  รู้ input รู้ output ก็รู้ความสัมพันธ์ของโมเดล

- Test (Validation) Data

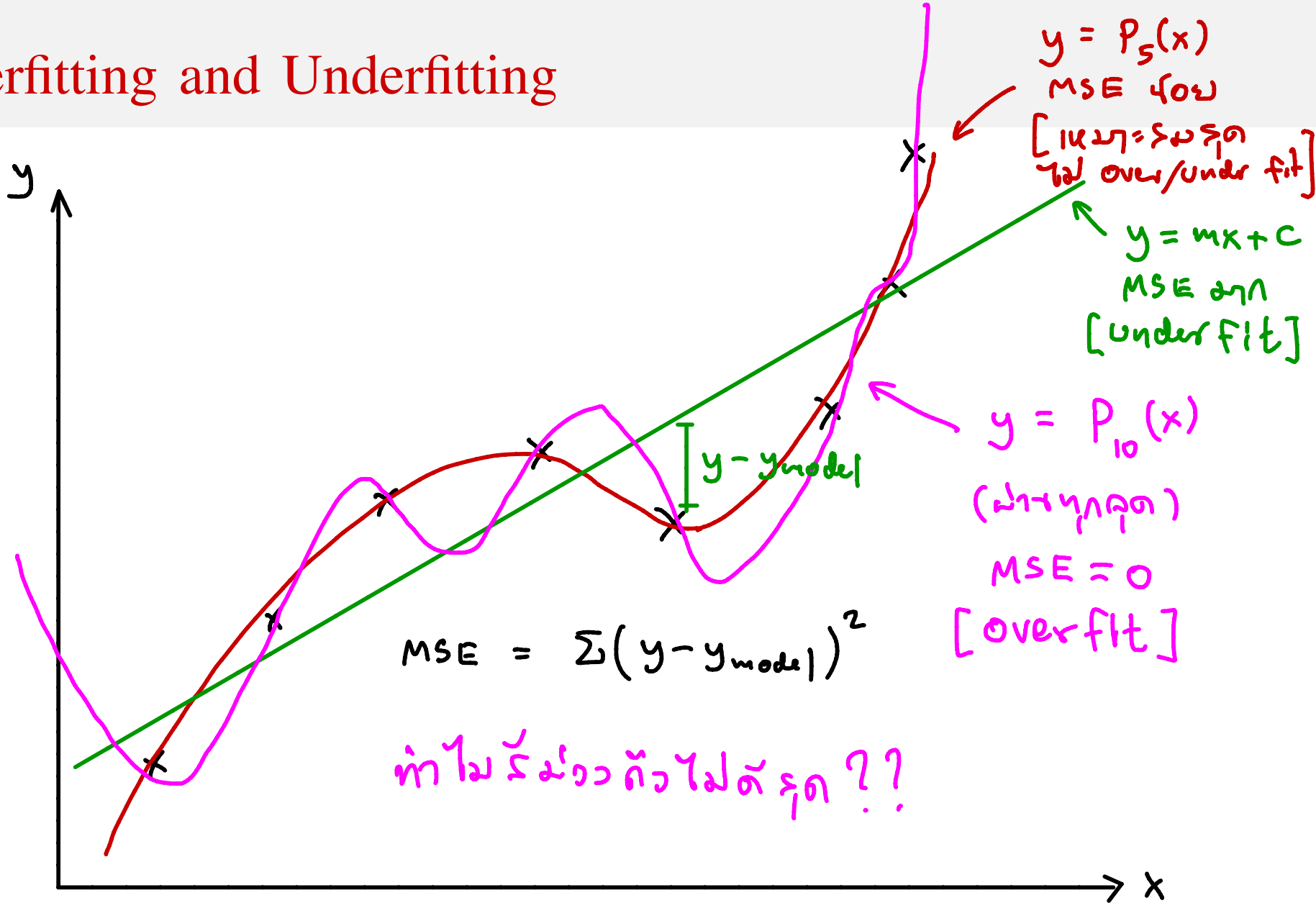
  รู้ input รู้ output ก็รู้ผลของโมเดล

  [ถ้าโมเดลทาย output → เทียบกับคำตอบที่ถูก]

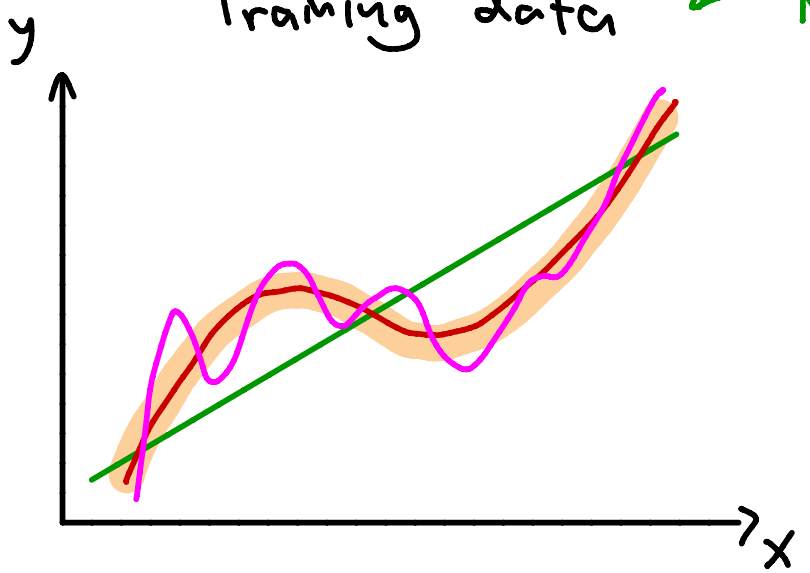
- Real-world use

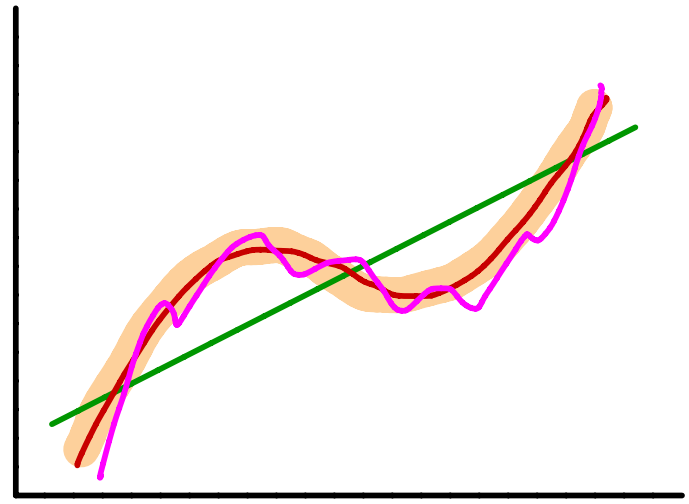
  รู้ input ไม่รู้ output ก็รู้โมเดลทาย output ⇒ ก็รู้ประโยชน์

Overfitting and Underfitting



Overfitting and Underfitting

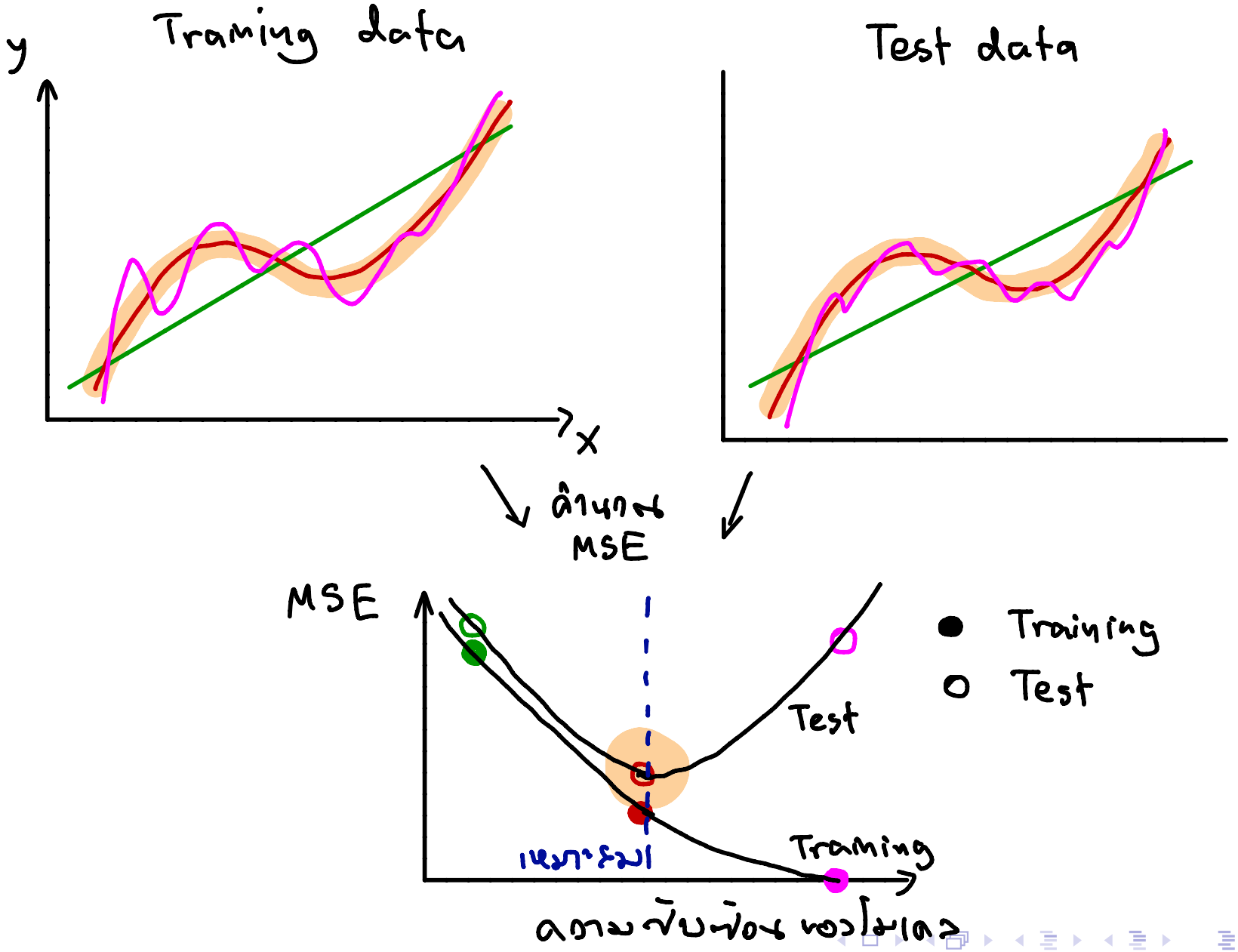




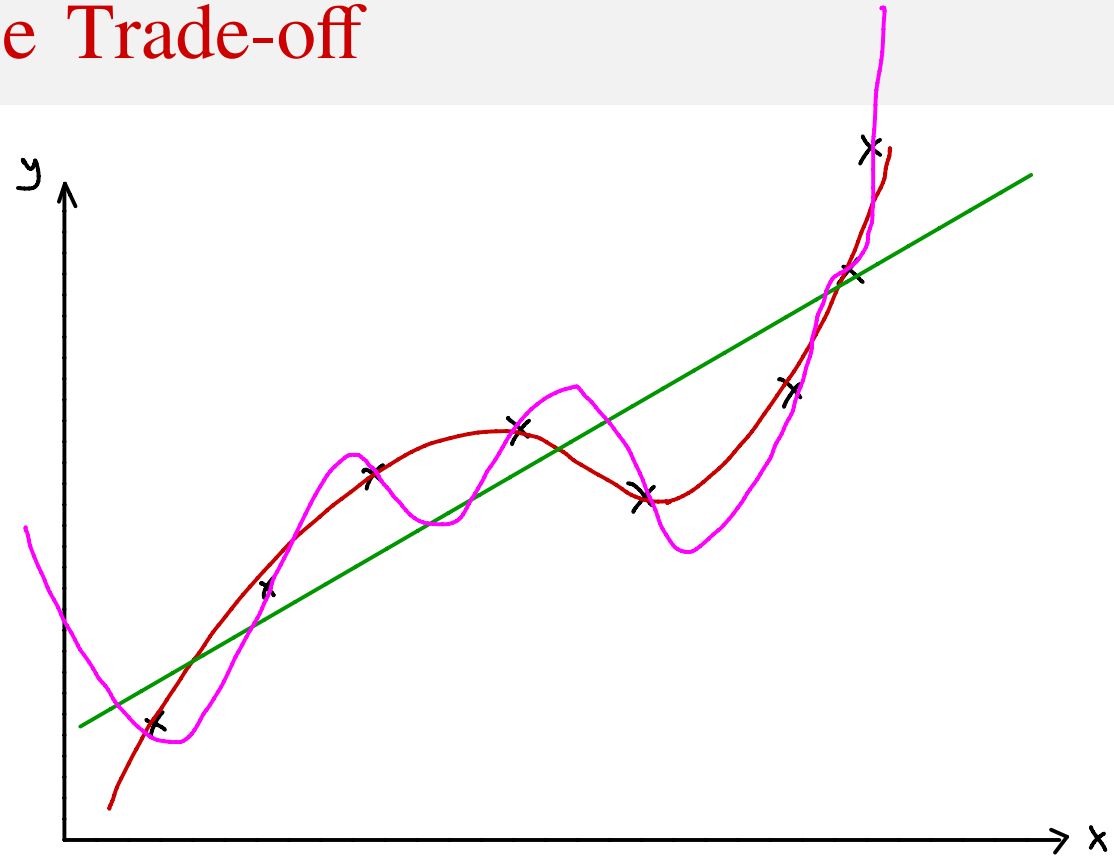
- ผิดกับ training data ดีมาก
- ผิดกับ traing data ดีระดับหนึ่ง
- ผิดกับ training data แย่

- ผิดกับ test data ไม่แย่
- ผิดกับ test data ดีระดับหนึ่ง
- ผิดกับ test data แย่มาก

Overfitting and Underfitting



Bias-Variance Trade-off



Bias โมเดลอ่อนไหวกับข้อมูลแต่ละชุด
โมเดลยืดหยุ่นแต่ไหน

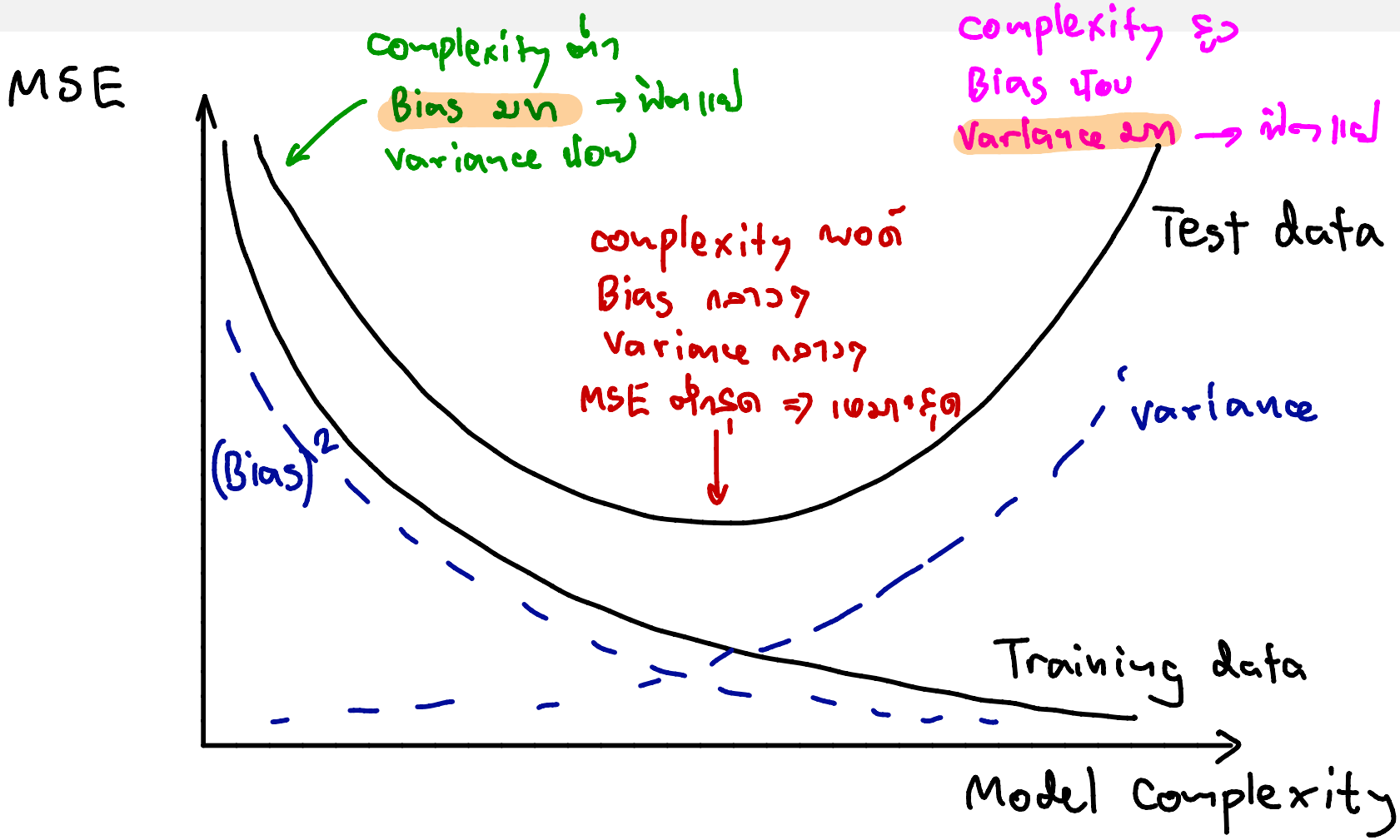
มาก
น้อย
น้อย ?????

Bias-Variance Trade-off

Variance โมเดล sensitive กับข้อมูลแต่ละชุด

น้อย
มาก
มาก ?????

Bias-Variance Trade-off



$$\text{MSE}_{\text{test}} = \text{Noise} + (\text{Bias})^2 + \text{Variance}$$

Bias-Variance Trade-off

- The *bias* error is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
- The *variance* is an error from sensitivity to small fluctuations in the training set. High variance may result from an algorithm modeling the random noise in the training data (overfitting).