TAUTOLOGY
INNOVATION
SCHOOL

TAUTOLOGY

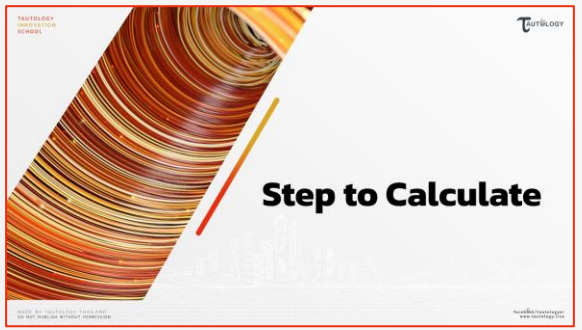
CROSS VALIDATION

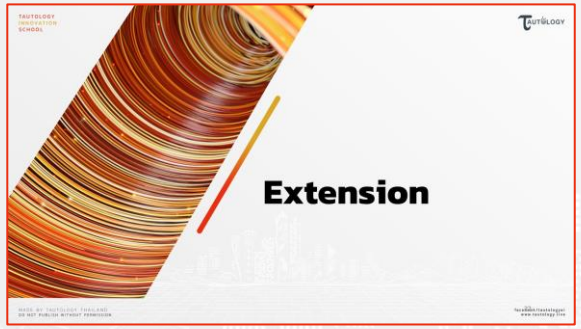
CROSS VALIDATION

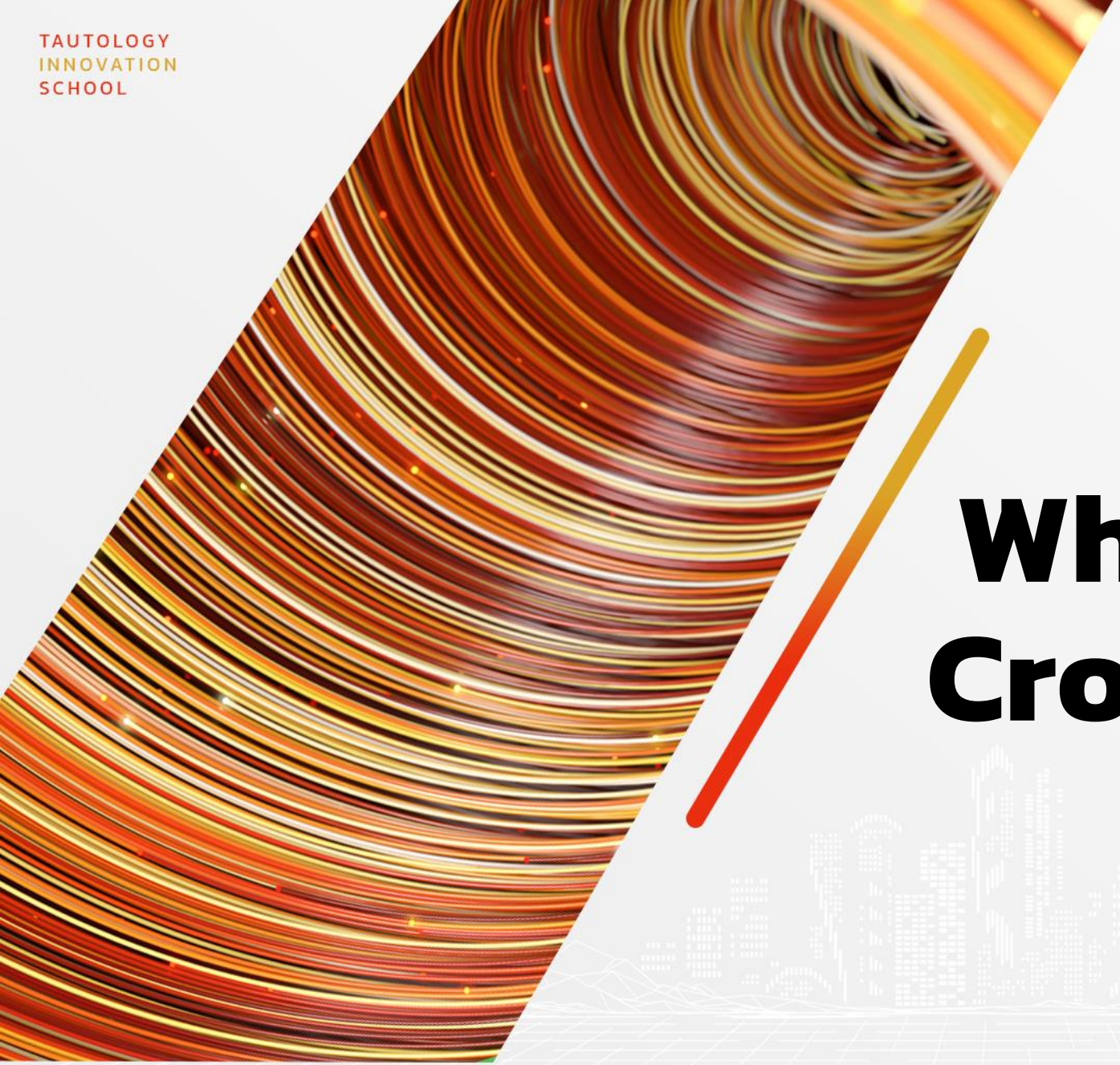
BY TAUTOLOGY

facebook/tautologyai
www.tautology.live

Cross Validation



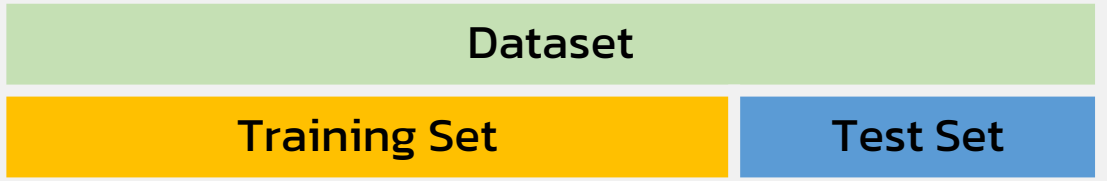




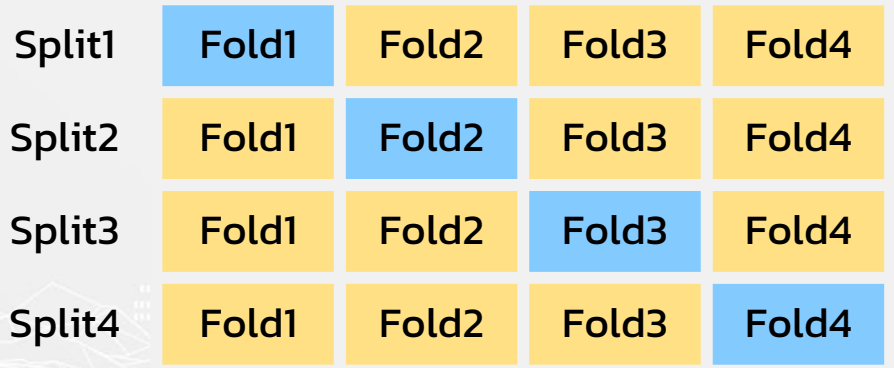
What is k-fold Cross Validation?

What is k-fold Cross Validation?

k-fold cross validation คือ วิธีการวัดประสิทธิภาพของ model บนข้อมูล k กลุ่มที่แตกต่างกัน



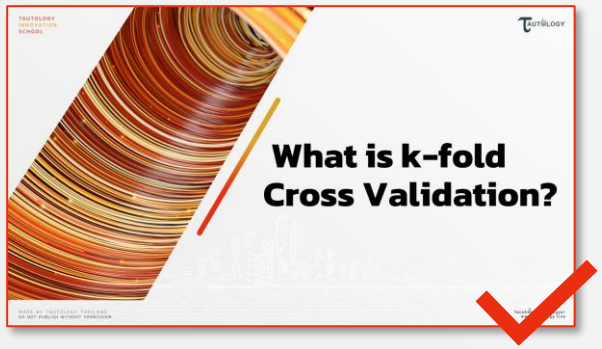
K = 4

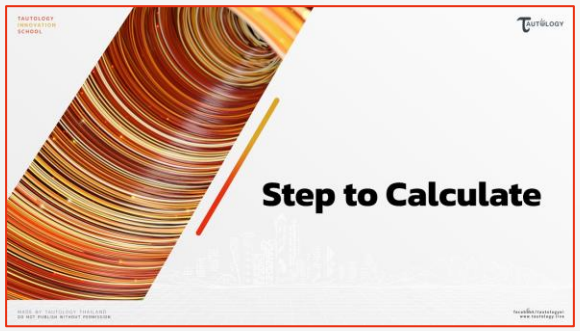


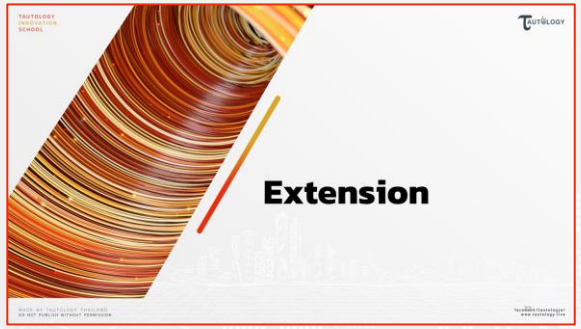
Training data

Validation data

Cross Validation







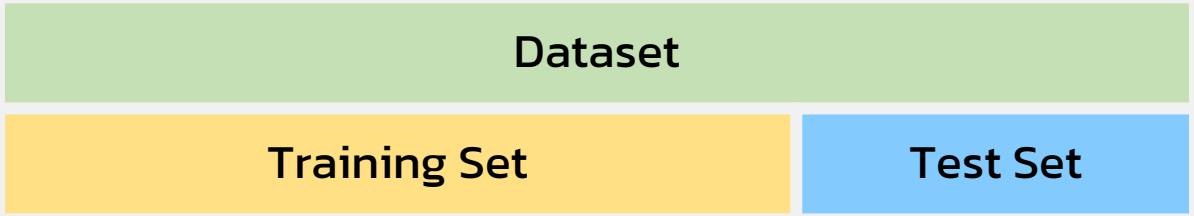
Step to Calculate

Step to Calculate

1. แบ่งข้อมูลใน dataset ออกเป็น training set และ test set
2. แบ่งข้อมูลใน training set ออกเป็น k กลุ่ม (k folds)
3. สร้างชุดข้อมูล k ชุด (k splits) จากข้อมูล k กลุ่ม (k folds)
4. ในแต่ละ split เรากำหนดให้มี 1 fold เป็น validation set และ fold ที่เหลือเป็น training set
5. สำหรับแต่ละ split ให้สร้าง model จาก training set และวัดประสิทธิภาพบน validation set
6. พิจารณาประสิทธิภาพบน validation set ของทุก split

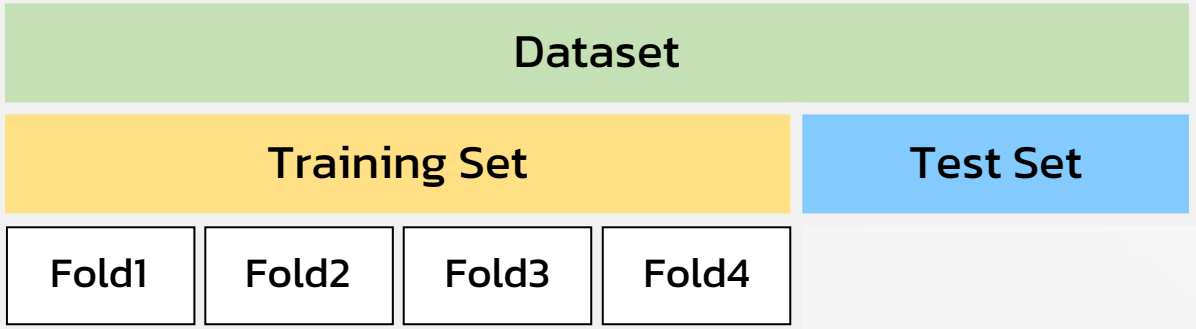
Step to Calculate

1. แบ่งข้อมูลใน dataset ออกเป็น training set และ test set



Step to Calculate

2. แบ่งข้อมูลใน training set ออกเป็น k กลุ่ม (k folds)

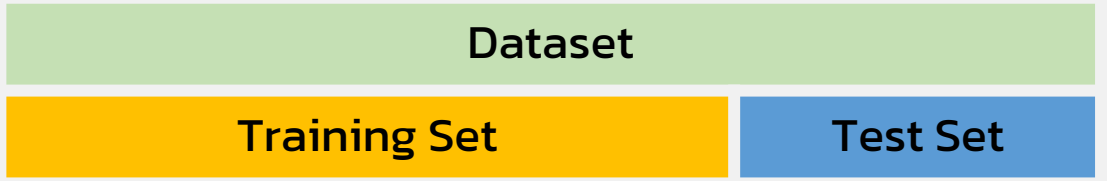


Step to Calculate

split = # fold (Q1)

ไม่ได้ copy ขึ้นมาจริง ? (Q2)

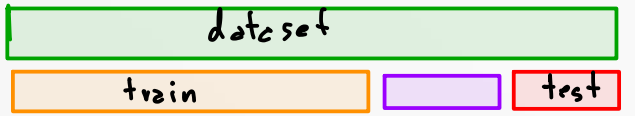
3. สร้างชุดข้อมูล k ชุด (k splits) จากข้อมูล k กลุ่ม (k folds)



Split1	Fold1	Fold2	Fold3	Fold4
Split2	Fold1	Fold2	Fold3	Fold4
Split3	Fold1	Fold2	Fold3	Fold4
Split4	Fold1	Fold2	Fold3	Fold4

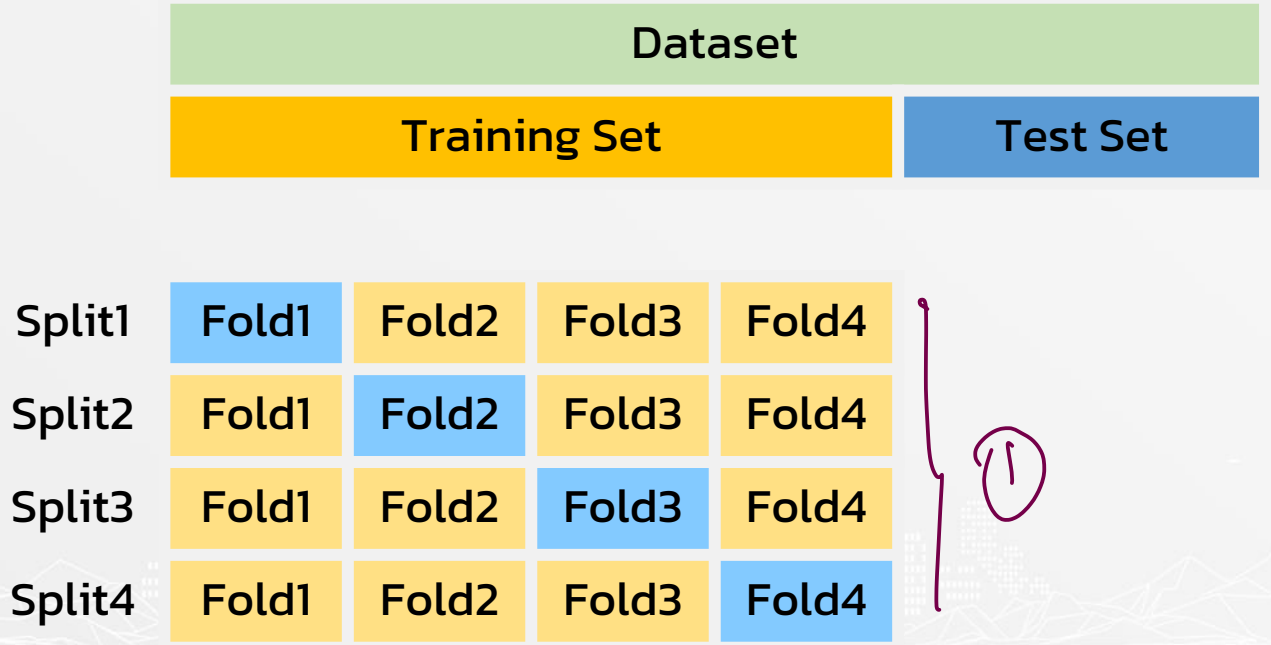
Step to Calculate

เป็นค่าใช้เรียก unseen data ที่อยู่บน training set



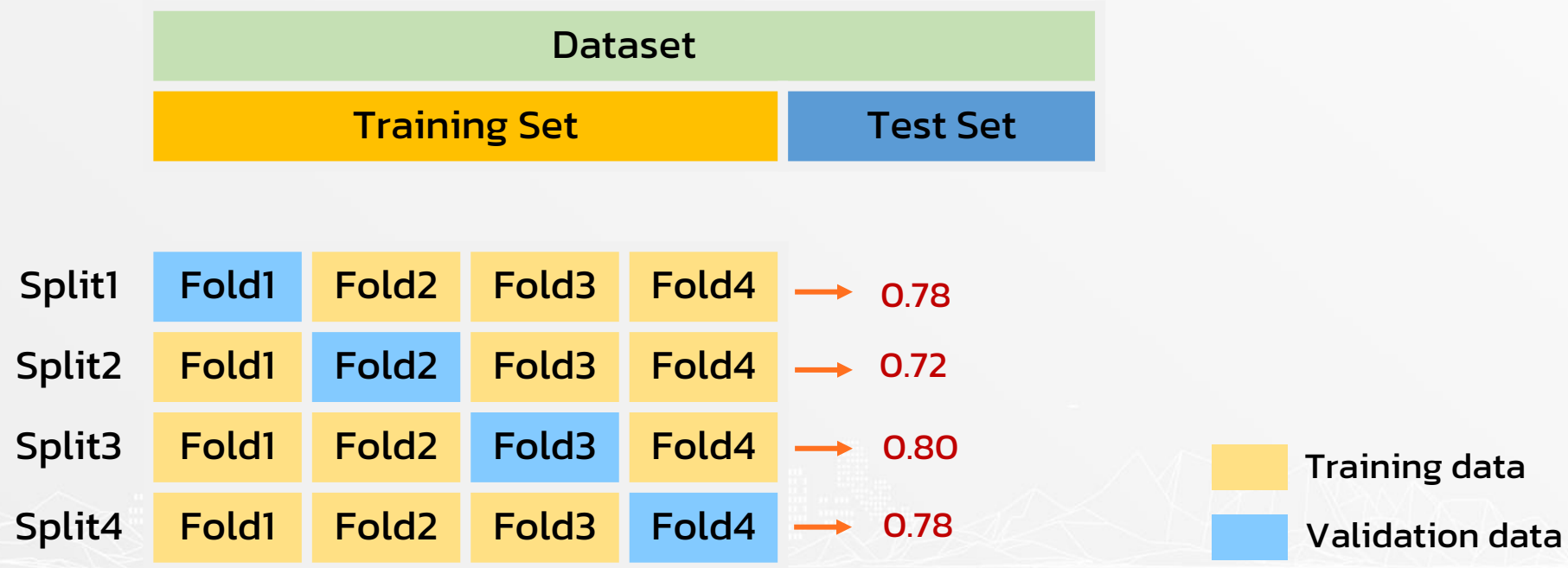
(มีไว้เพื่อเลือก hyper parameter ก่อนที่จะเอาไปทดสอบกับ test)

4. ในแต่ละ split เรากำหนดให้มี 1 fold เป็น validation set และ fold ที่เหลือเป็น training set



Step to Calculate

5. สำหรับแต่ละ split ให้สร้าง model จาก training set และวัดประสิทธิภาพบน validation set



Step to Calculate

5, 10

6. พิจารณาประสิทธิภาพบน validation set ของทุก split

เทคนิคของสำนัก : พิจารณา candidate ที่ได้ variance ไม่สูงเกินไป (filter)
→ ค่อยเลือกตัวที่ performance ดีที่สุด

ค. แปรปรวนของ performance

	Fold				Mean	Variance
	1	2	3	4		
$\lambda = 0.01$	0.78	0.72	0.80	0.78	0.77	0.0009
$\lambda = 0.02$	0.99	0.98	0.99	0.24	0.80	

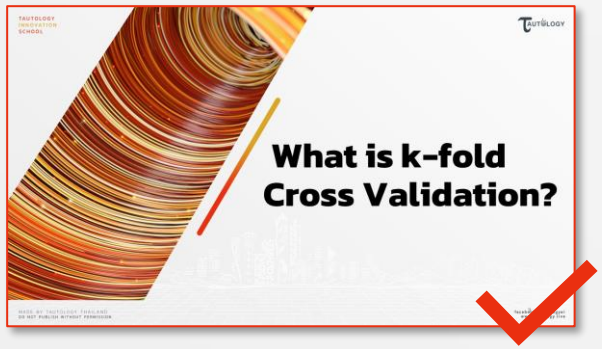
$\frac{0.78 + 0.72 + 0.80 + 0.78}{4}$

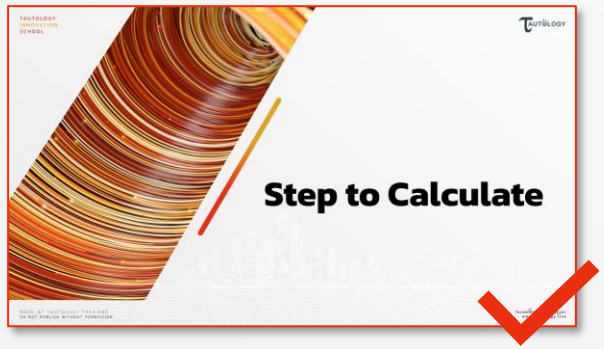
mean ของ performance

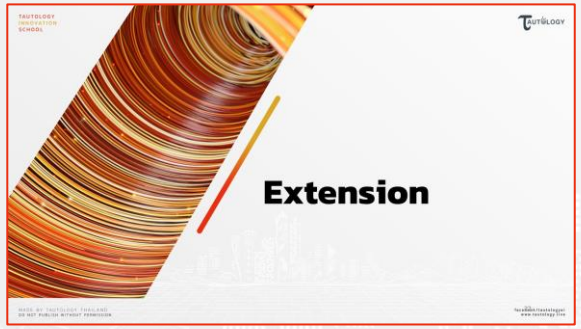
$\frac{\Sigma(x - \bar{x})^2}{n}$

R^2, mse, mae, mape

Cross Validation







Advantage of k-fold Cross Validation

Advantage of k-fold CV

Close to Real Performance

Model Selection

Close to Real Performance

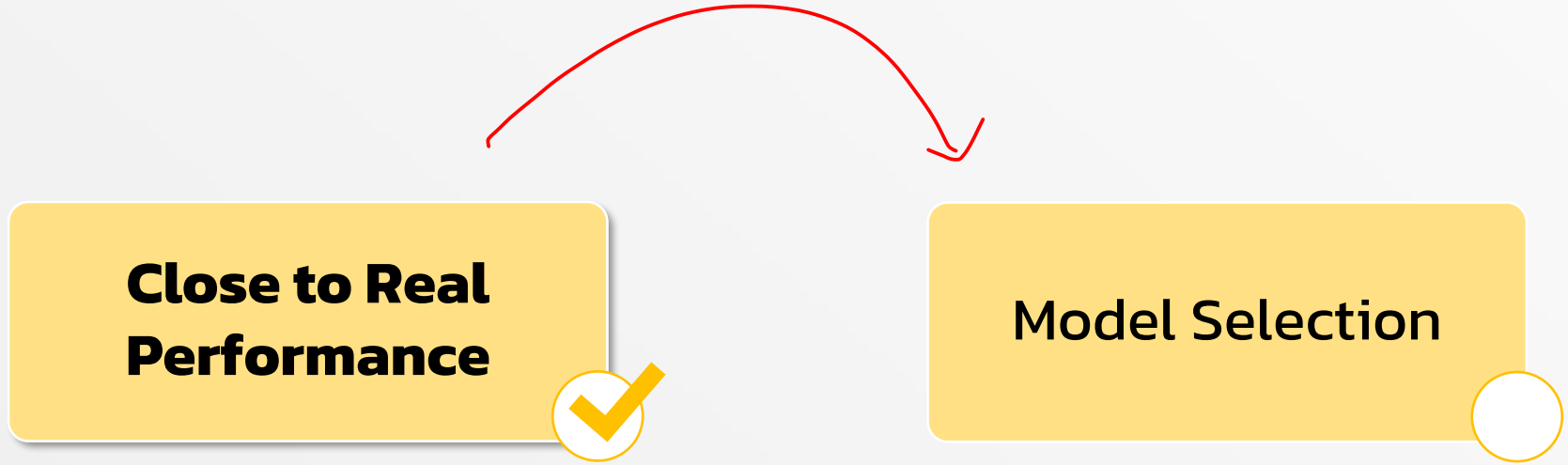
การวัดประสิทธิภาพของ model บน validation set ที่แตกต่างกันหลายชุด ทำให้เราได้ประสิทธิภาพที่ใกล้เคียงประสิทธิภาพที่แท้จริงของ model

Law of Large Number

ชุดข้อมูล

Fold				Mean	Variance
1	2	3	4		
0.78	0.72	0.80	0.78	0.77	0.0009

Advantage of k-fold CV



Model Selection

การที่เรารู้ประสิทธิภาพที่ใกล้เคียงจริง ทำให้เราสามารถเปรียบเทียบประสิทธิภาพของ model หลาย ๆ ตัว เพื่อเลือก model ที่เหมาะสมกับการใช้งานที่สุดได้

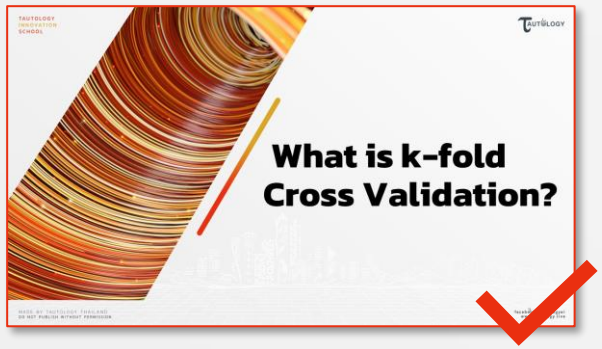
Model	Fold				Mean	Variance
	1	2	3	4		
Model 1	0.78	0.72	0.80	0.78	0.77	0.0009
Model 2	0.80	0.80	0.72	0.80	0.78	0.0012
Model 3	0.82	0.74	0.80	0.74	0.775	0.001275
Model 4	0.79	0.74	0.78	0.76	0.7675	0.000369
Model 5	0.74	0.72	0.76	0.78	0.75	0.0005

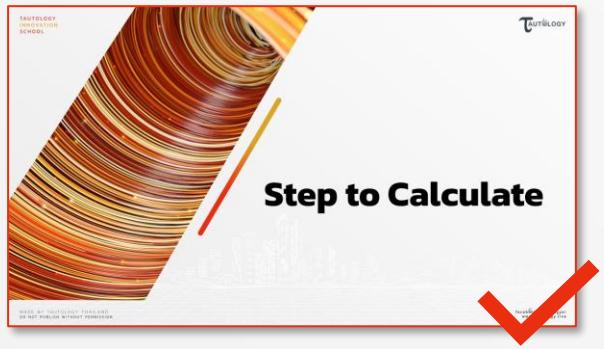
Advantage of k-fold CV

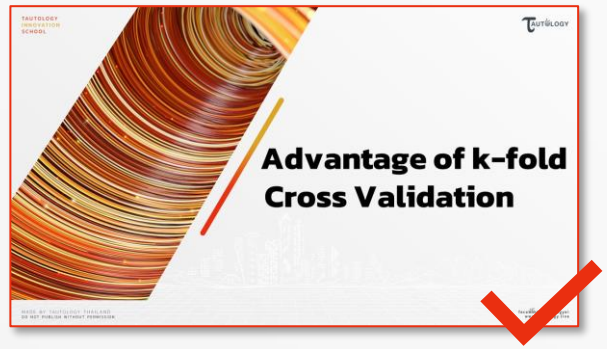
Close to Real Performance

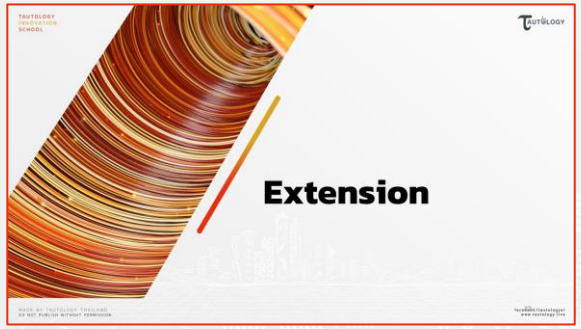
Model Selection

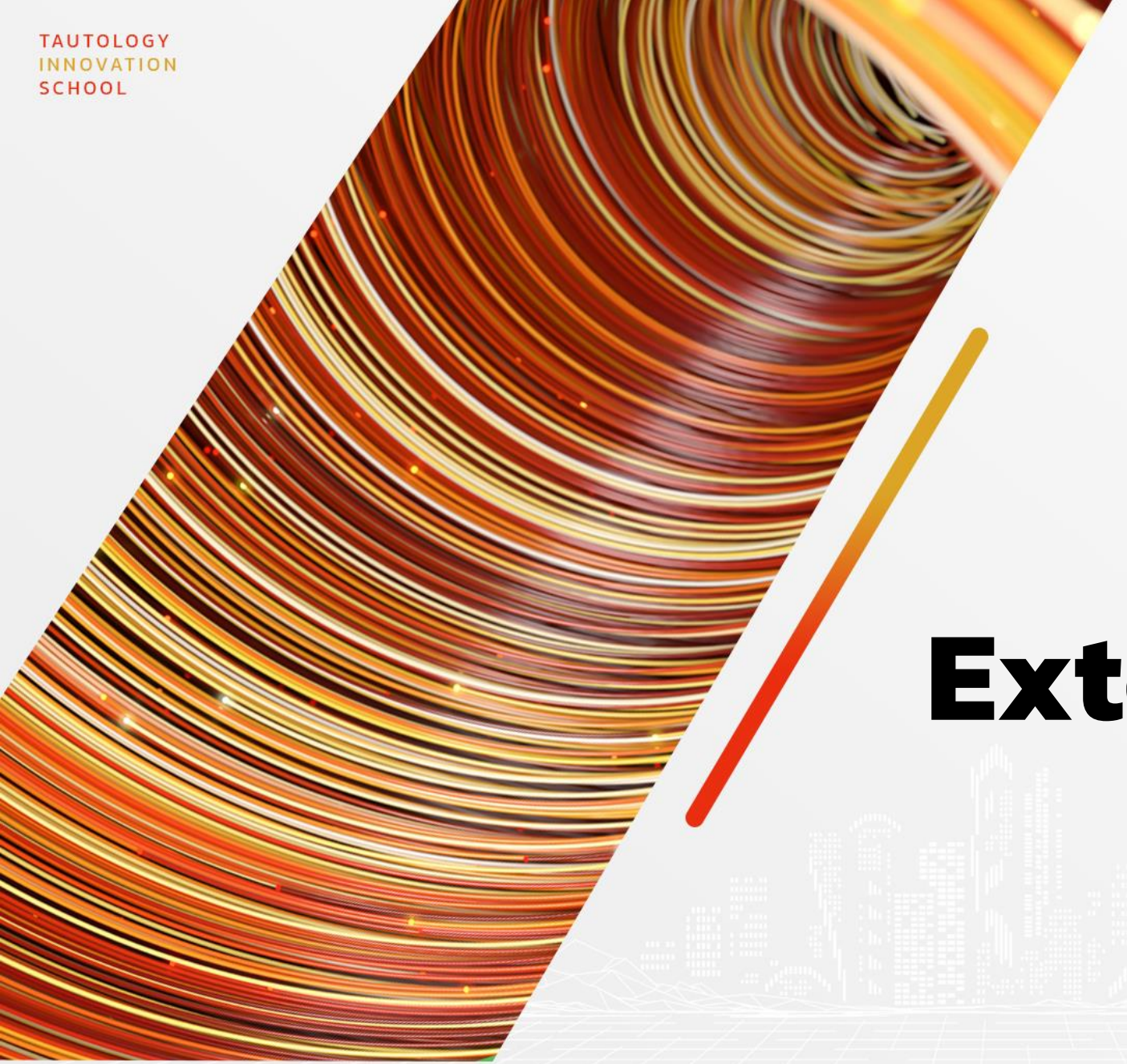
Cross Validation











Extension

Extension (Special Topic)

How to select k?

Cross Validation for Time Series

How to select k?

สมมติว่า training มี n sample
สมัย n split => วัด perf
(คุณ. perfectionist => เปลือง cost มาก!)

มาจากการทดลอง

"การเลือก k = 5 หรือ k = 10 นั้นเพียงพอแล้วที่จะแสดงประสิทธิภาพที่แท้จริงของ model"

-หนังสือ An introduction to statistical learning, James et al. (2013)-

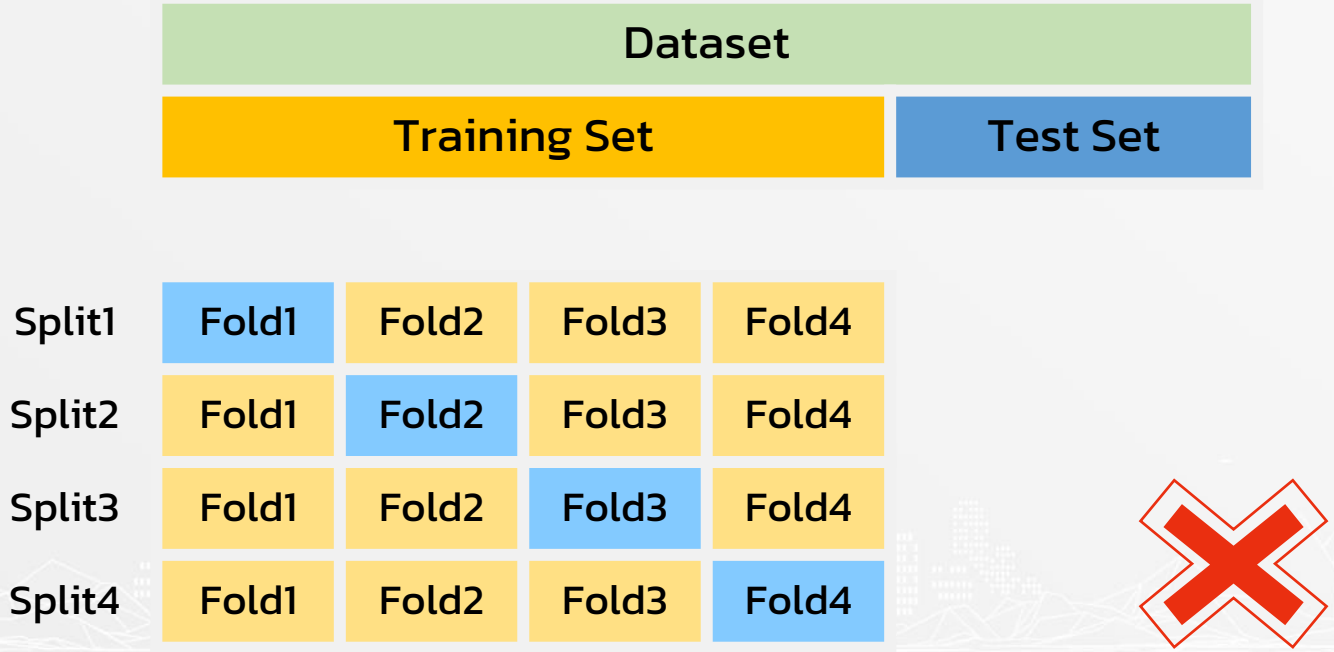
Extension

How to select k?

Cross Validation
for Time Series

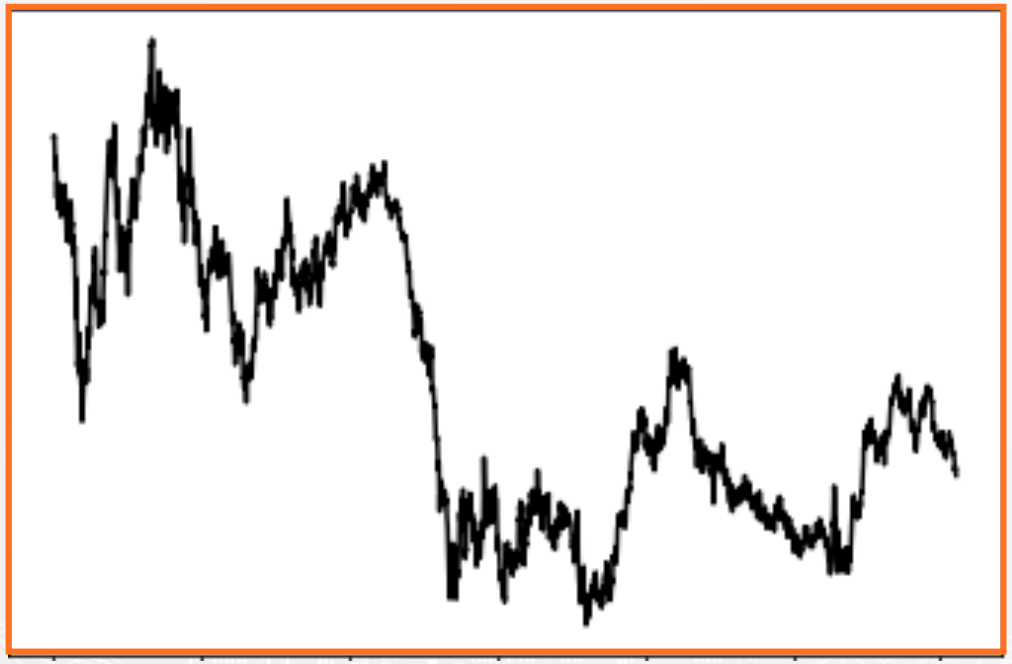
Cross Validation for Time Series

สำหรับข้อมูลที่อยู่ในรูปแบบของ time series นั้น เราไม่สามารถทำ k-fold CV แบบปกติได้



Cross Validation for Time Series

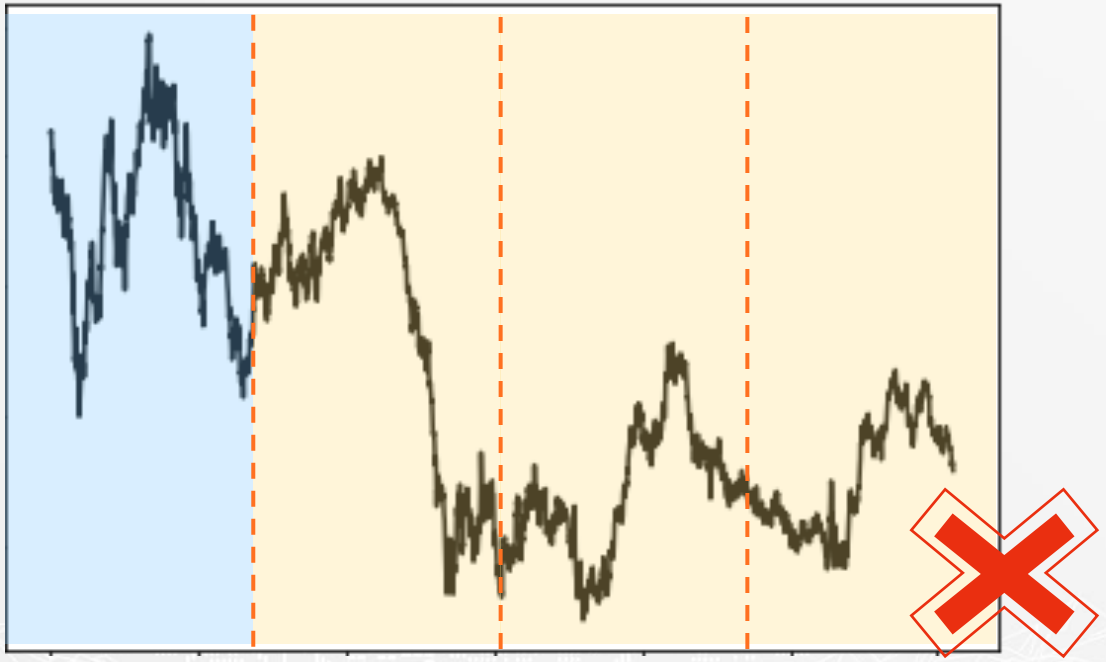
เนื่องจากข้อมูลที่อยู่ในรูปแบบของ time series เป็นข้อมูลที่ขึ้นกับเวลา ลำดับการเกิดขึ้นก่อนหลังของข้อมูลจึงมีความสำคัญ



กราฟแสดงข้อมูลระหว่างราคาสูงสุดของ EURUSD ในแต่ละวัน

Cross Validation for Time Series

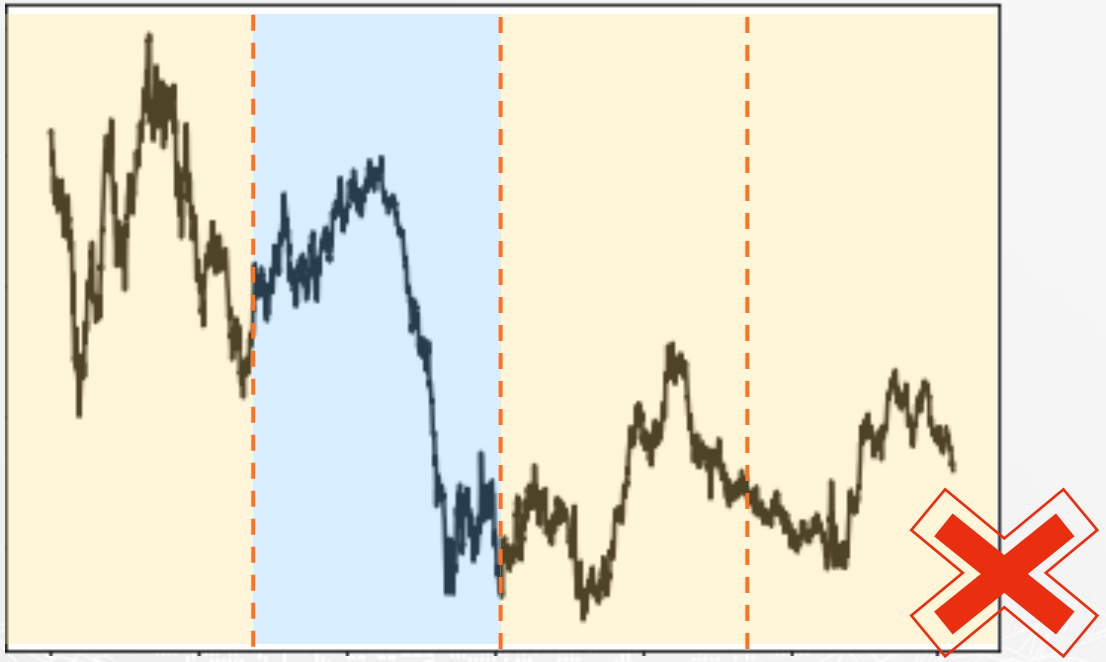
การนำข้อมูลที่เกิดขึ้นทีหลังมาใช้เป็น training set และนำข้อมูลที่เกิดขึ้นก่อนมาใช้เป็น validation set จะทำให้ประสิทธิภาพที่วัดได้ ไม่สื่อถึงประสิทธิภาพที่แท้จริง



กราฟแสดงข้อมูลระหว่างราคาสูงสุดของ EURUSD ในแต่ละวัน

Cross Validation for Time Series

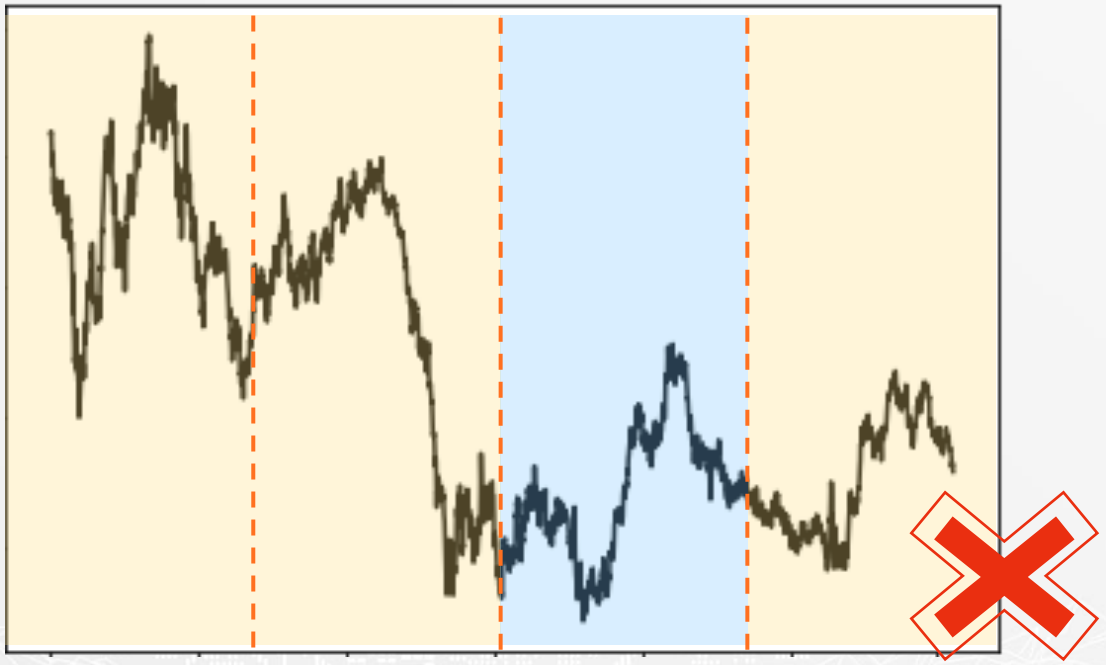
การนำข้อมูลที่เกิดขึ้นทีหลังมาใช้เป็น training set และนำข้อมูลที่เกิดขึ้นก่อนมาใช้เป็น validation set จะทำให้ประสิทธิภาพที่วัดได้ ไม่สื่อถึงประสิทธิภาพที่แท้จริง



กราฟแสดงข้อมูลระหว่างราคาสูงสุดของ EURUSD ในแต่ละวัน

Cross Validation for Time Series

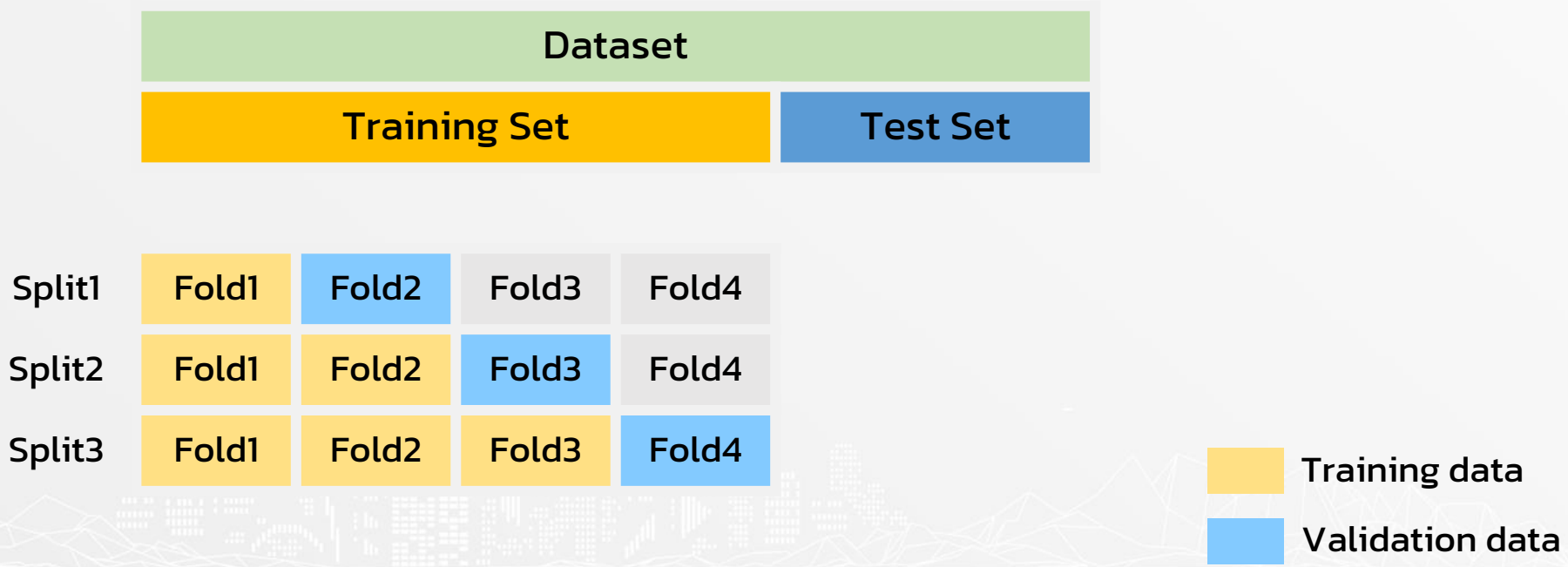
การนำข้อมูลที่เกิดขึ้นทีหลังมาใช้เป็น training set และนำข้อมูลที่เกิดขึ้นก่อนมาใช้เป็น validation set จะทำให้ประสิทธิภาพที่วัดได้ ไม่สื่อถึงประสิทธิภาพที่แท้จริง



กราฟแสดงข้อมูลระหว่างราคาสูงสุดของ EURUSD ในแต่ละวัน

Cross Validation for Time Series

รูปแบบการทำ k-fold CV สำหรับ time series เป็นดังนี้

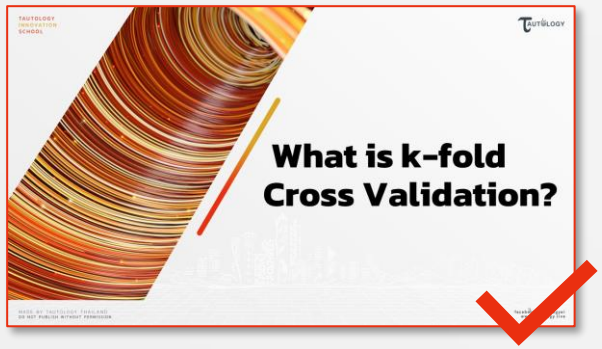


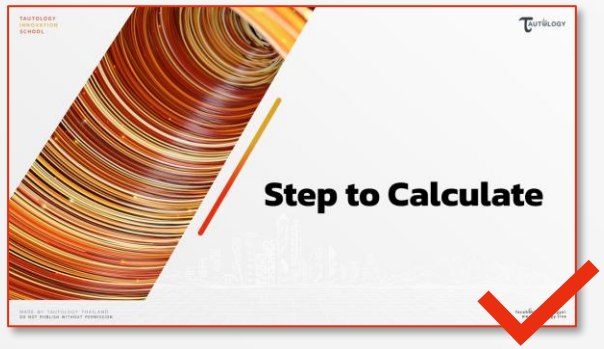
Extension

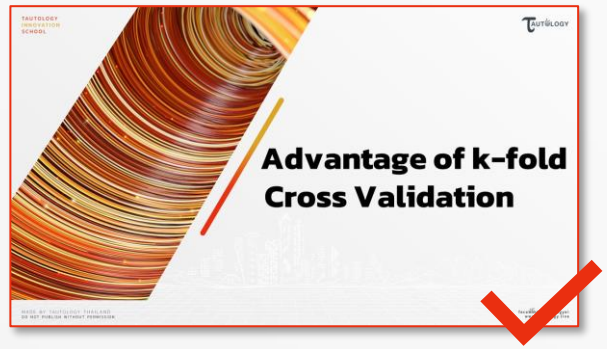
How to select k?

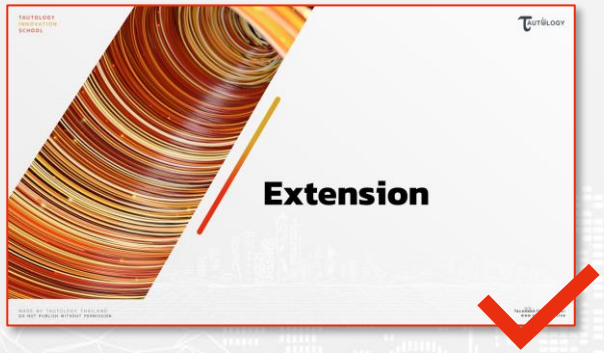
Cross Validation for Time Series

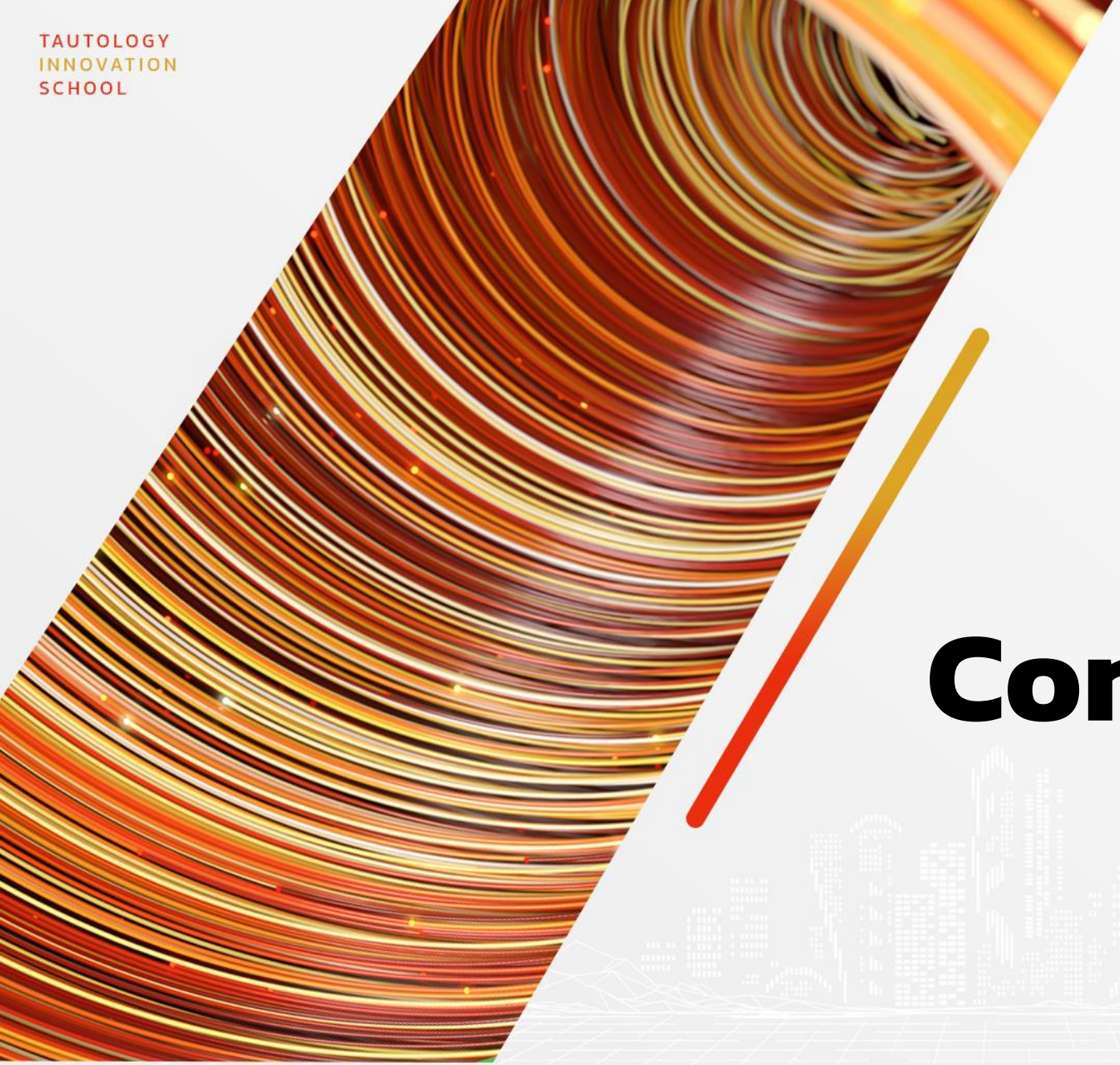
Cross Validation











Conclusion

Conclusion

- k-fold cross validation คือ การวัดประสิทธิภาพของ model บนข้อมูล k กลุ่มที่แตกต่างกัน
- ประสิทธิภาพของ model ที่วัดได้จาก k-fold CV จะใกล้เคียงประสิทธิภาพที่แท้จริงของ model
- สามารถนำไปต่อยอดเพื่อทำ model selection
- k ที่เหมาะสมคือ 5 folds หรือ 10 folds
- สำหรับข้อมูลที่เป็น time series ต้องคำนึงถึงลำดับก่อนหลังของข้อมูลด้วย

Cross Validation