DATA ANALYTICS

Wanida Saetang Ph.D (candidate)
King Mongkut's University of Technology North Bangkok

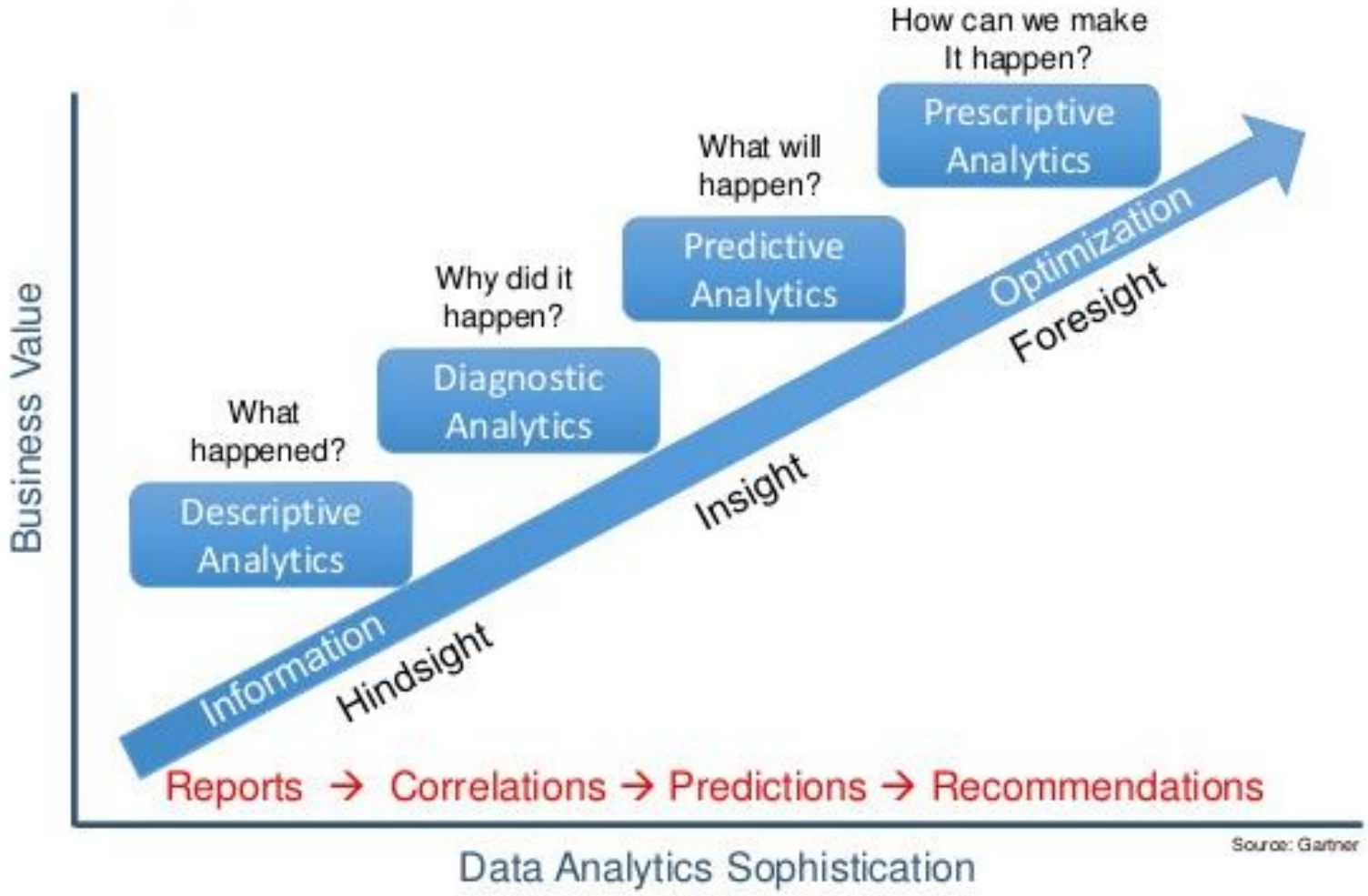
Agenda

- Data Analytics
  - Predictive analytics
- Data Mining Techniques
  - Decision Tree
  - K-means
- Apply Model & Validation Model
- Rapid Miner Studio
- Workshop

Data Analytics

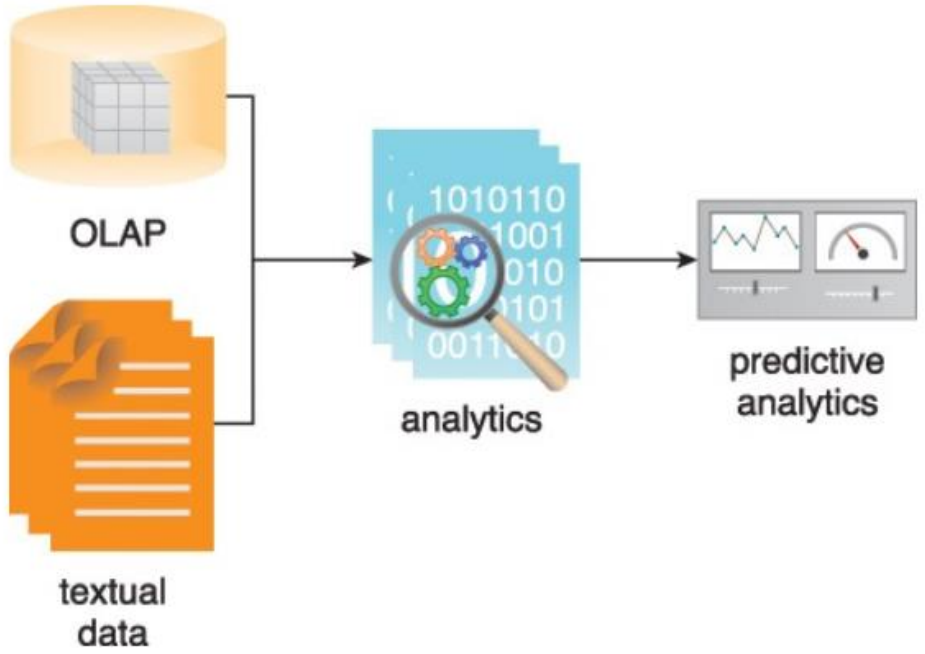
predictive analytics

The Progression of Analytics



Predictive Analytics

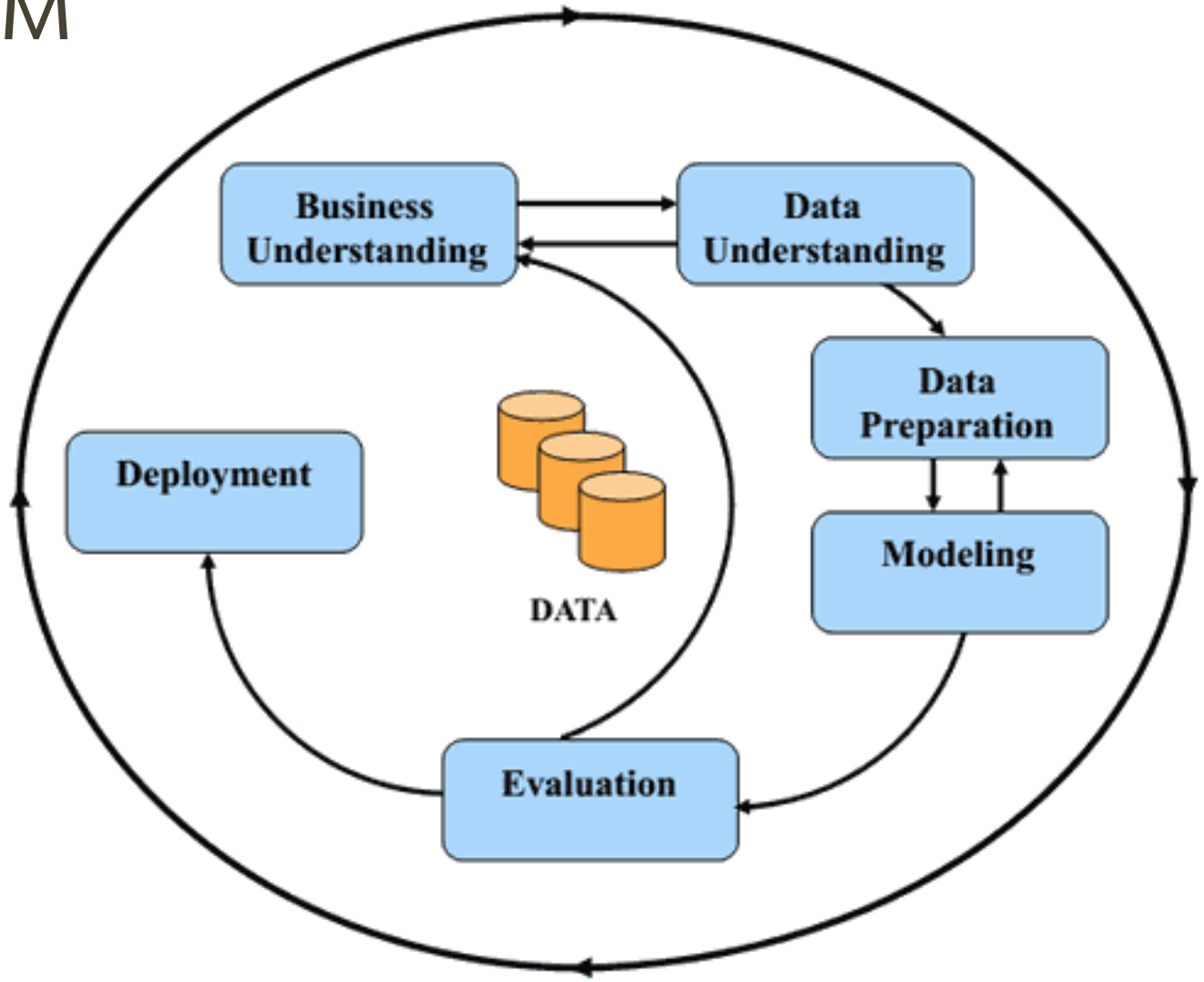
- carried out in an attempt to determine the outcome of **an event that might occur in the future**.
- the models used for predictive analytics have implicit **dependencies on the conditions under which the past events occurred**.



Data Mining Techniques

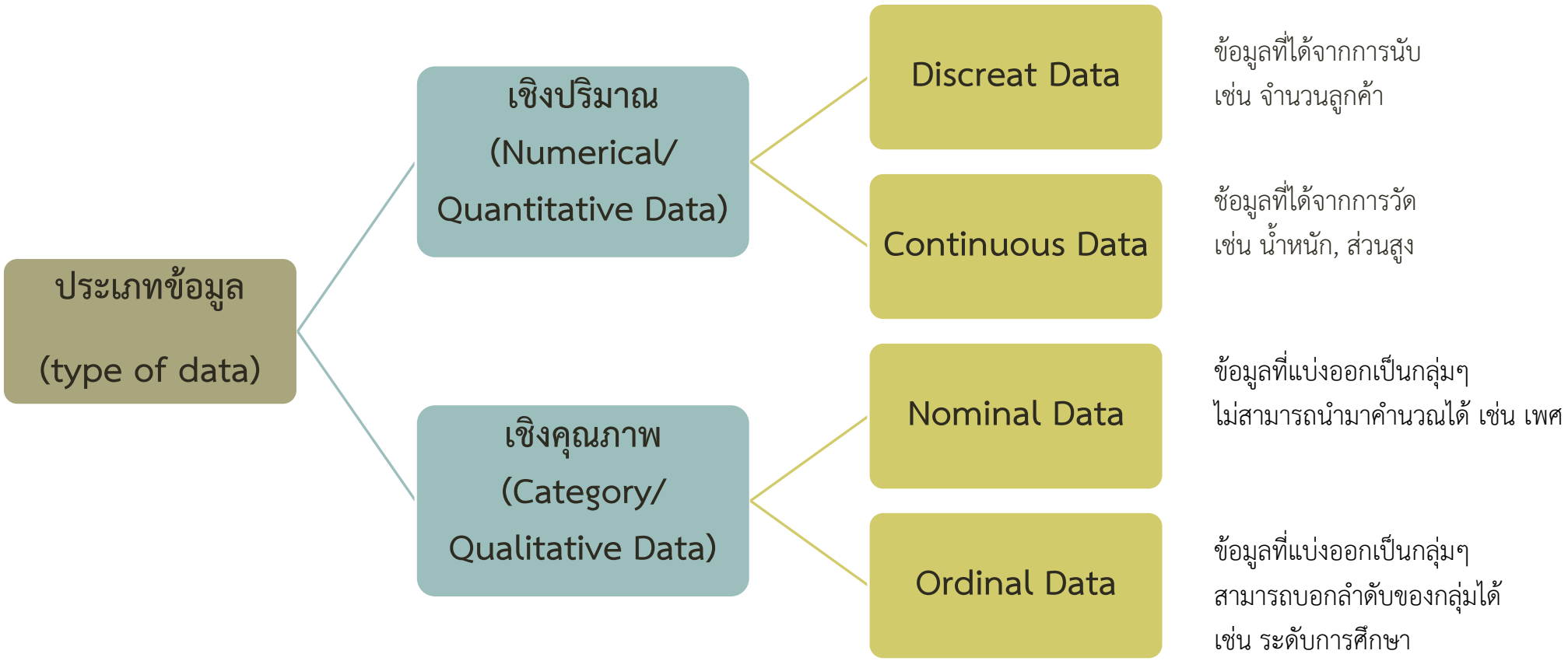
Decision tree
K-Means

CRISP-DM

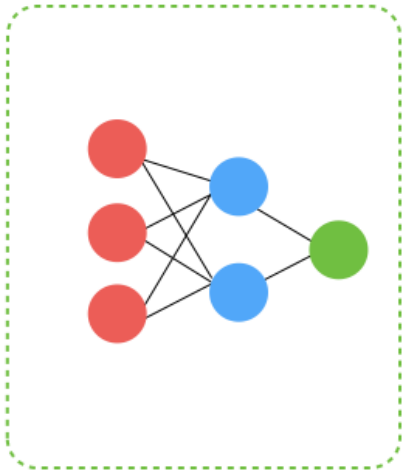


http://mlwiki.org/index.php/CRISP-DM

ประเภทของข้อมูล

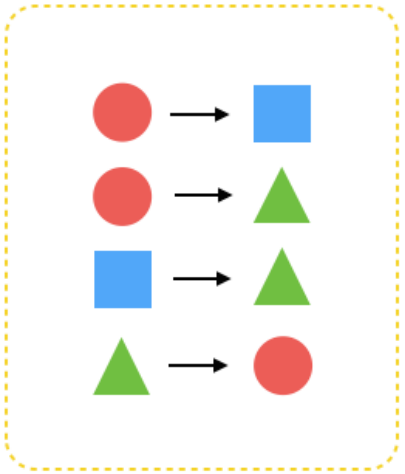


เทคนิคเหมืองข้อมูล



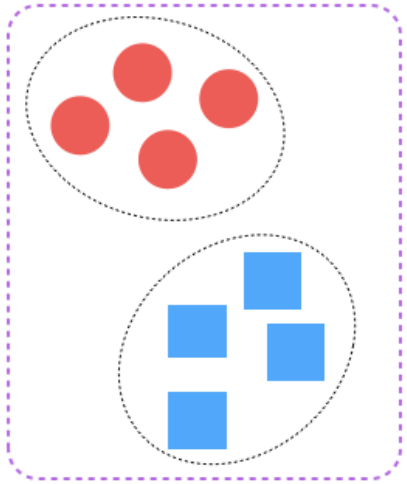
Classification

- Decision Tree
- Naive Bayes
- Neural Network
- Support Vector Machines (SVM)



Association Rules

- Apriori algorithm
- Eclat algorithm
- FP-growth algorithm



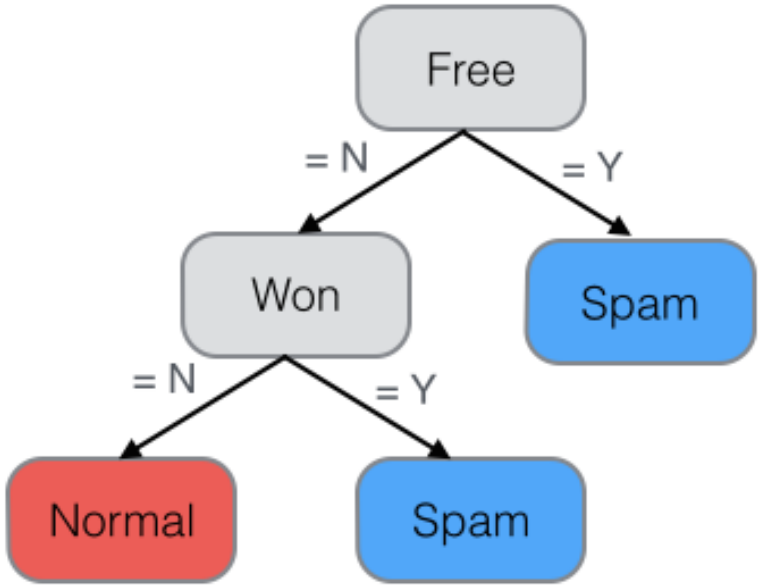
Clustering

- K-Means
- DBSCAN
- EM Clustering using GMMs.
- Agglomerative Hierarchical

Classification Decision Tree

Decision Tree

- ต้นไม้ตัดสินใจ (decision tree) เป็นการ**จำแนกกลุ่ม**โดยที่**ทราบจำนวนกลุ่ม**ปลายทาง
- เป้าหมายของการจำแนก คือ ทำนายค่า หรือตัวแปรเป้าหมาย (class/label)
- ต้นไม้ตัดสินใจเป็นเหมือนกราฟ หรือแผนผัง มีลักษณะเป็นต้นไม้กลับหัว
- ประกอบด้วย **Node** (โหนด) โดยแต่ละโหนด แทนตัวแปรอินพุต (input attribute) ต่าง ๆ ในชุดข้อมูล และ **Edge** (เส้นเชื่อม) แทนค่าของตัวแปร (numerical attributes)
- โหนดบนสุดเรียกว่า root node หรือโหนดราก และแตกกิ่งออกมาเป็น leaf node หรือโหนดใบ



Decision Tree: Information Gain

ขั้นตอนการสร้าง decision tree จะทำการคำนวณเลือกแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสมาใช้งาน

- ค่า Information Gain สามารถคำนวณได้จากสมการ ด้านล่างนี้

$$\text{Information Gain} = \text{Entropy(initial)} - [P(c_1) \times \text{Entropy}(c_1) + P(c_2) \times \text{Entropy}(c_2) + \ldots]$$

โดยที่ $\text{Entropy}(c_1) = -P(c_1) \log_2 P(c_1)$

และ $P(c_1)$ คือ ค่าความน่าจะเป็น (probability) ของ c_1

Decision Tree: spam e-mail classification

ID	Text	Type
1	Please call our customer service representative on FREE PHONE 0808 145 4742 between 9am-11pm as you have WON a guaranteed £1000 cash	spam
2	You have won $1,000 cash or a $2,000 prize! To claim, call 09050000327	spam
3	I'm gonna be home soon and I don't want to talk about this stuff anymore tonight	normal
4	Is that seriously how you spell his name?	normal
5	Double mins and txts 4 6months FREE Bluetooth on Orange. Available on Sony, Nokia Motorola phones.	spam

keywords

ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal
4	N	N	N	normal
5	Y	N	N	spam
6	Y	N	N	spam
7	N	N	N	normal
8	N	Y	N	spam
9	N	N	N	normal
10	N	N	N	normal

Decision Tree: Information Gain

ID	Type
1	spam
2	spam
5	spam
6	spam
8	spam
3	normal
4	normal
7	normal
9	normal
10	normal

P(spam) = 5/10 = 0.5

P(normal) = 5/10 = 0.5

Entropy (initial) = - [P(spam) × $\log_2$ P(spam) + P(normal) × $\log_2$ P(normal)]

Entropy(initial) = - [0.5 x $\log_2$ (0.5) + 0.5 x $\log_2$ (0.5)]

= - [0.5 x (-1) + 0.5 x (-1)]

= 1

Decision Tree: Information Gain

ID	Free	Type
1	Y	spam
5	Y	spam
6	Y	spam
2	N	spam
3	N	normal
4	N	normal
7	N	normal
8	N	spam
9	N	normal
10	N	normal

P(spam) = 3/3 = 1.0

P(normal) = 0/3 = 0.0

Entropy(***Free = Y***) = -[1.0 x $\log_2$ (1.0) + 0.0 x $\log_2$ (0.0)]

= -[1.0 x 0 + 0.0 x 0]

= 0

P(spam) = 2/7 = 0.29

P(normal) = 5/7 = 0.71

Entropy(***Free = N***) = -[0.29 x $\log_2$ (0.29) + 0.71 x $\log_2$ (0.71)]

= -[0.29 x (-1.79) + 0.71 x (-0.49)]

= 0.87

Information Gain (***Free***) = Entropy(***initial***) – [*P* (***Free = Y***) × Entropy(***Free = Y***)

+ *P*(***Free =*** N) × Entropy(***Free = N***)]

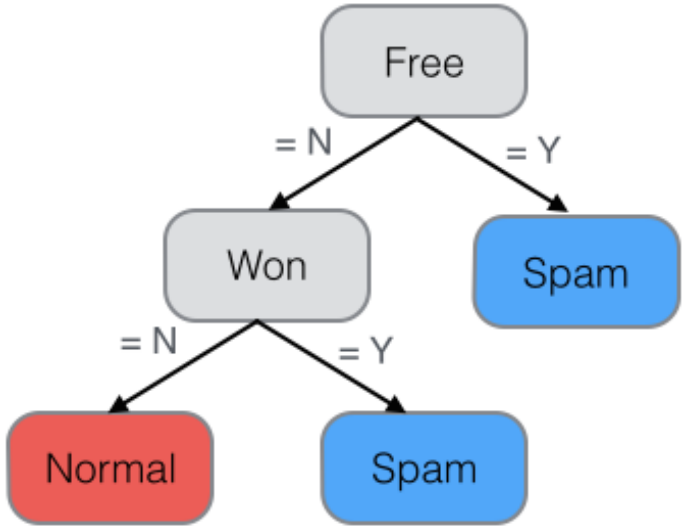
= 1 – [0.3 × 0 + 0.7 × 0.87]

= 0.39

สร้างโมเดล (Classification model)

attribute | label

ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal
4	N	N	N	normal
5	Y	N	N	spam
6	Y	N	N	spam
7	N	N	N	normal
8	N	Y	N	spam
9	N	N	N	normal
10	N	N	N	normal



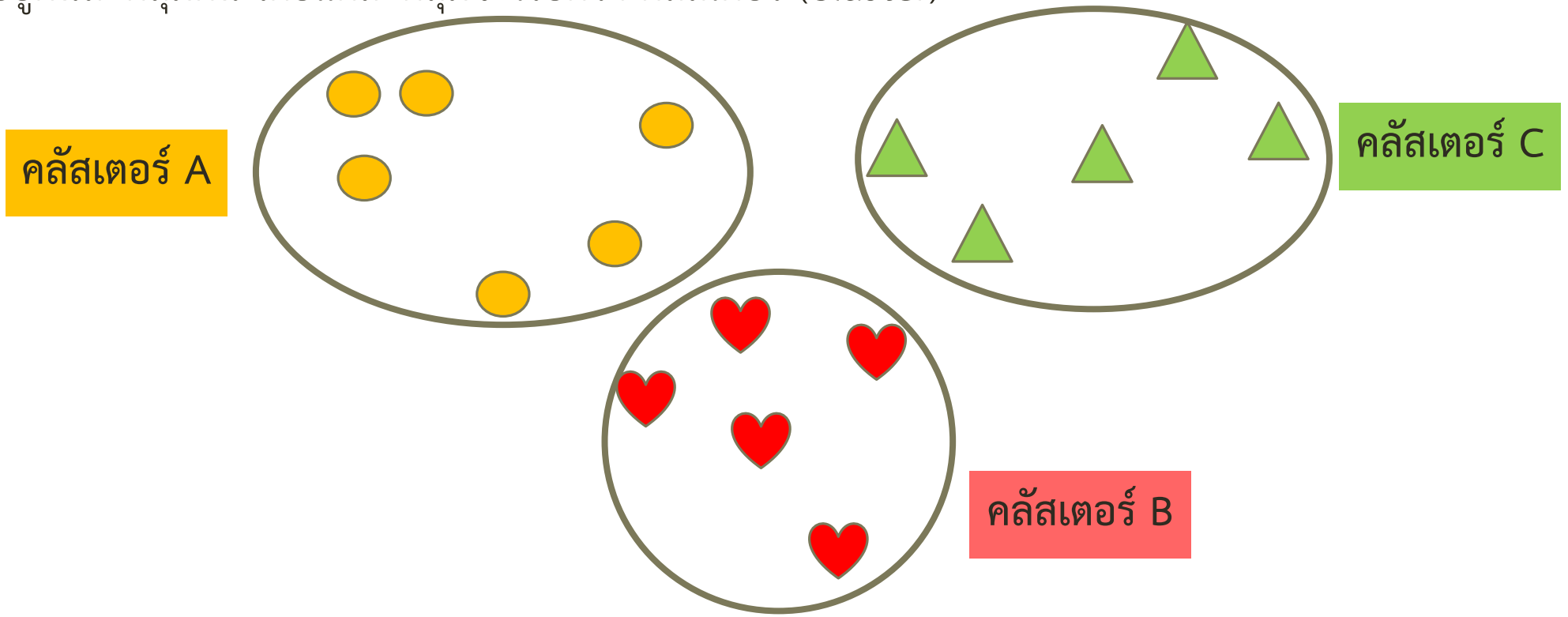
classification model

http://dataminingtrend.com/2014/data-mining-techniques/ensemble-model/

Clustering
K-means

Clustering

- การทำ Clustering คือ การ**แบ่งกลุ่ม**หรือ**จัดกลุ่ม**ข้อมูล โดย**ไม่ทราบจำนวน**กลุ่มปลายทาง
- ข้อมูลที่มีลักษณะคล้าย ๆ กัน จะอยู่กลุ่มเดียวกัน ข้อมูลที่มีลักษณะที่แตกต่างกันมาก ๆ จะถูกจัดให้อยู่คนละกลุ่มกัน โดยแต่ละกลุ่มจะเรียกว่า คลัสเตอร์ (cluster)

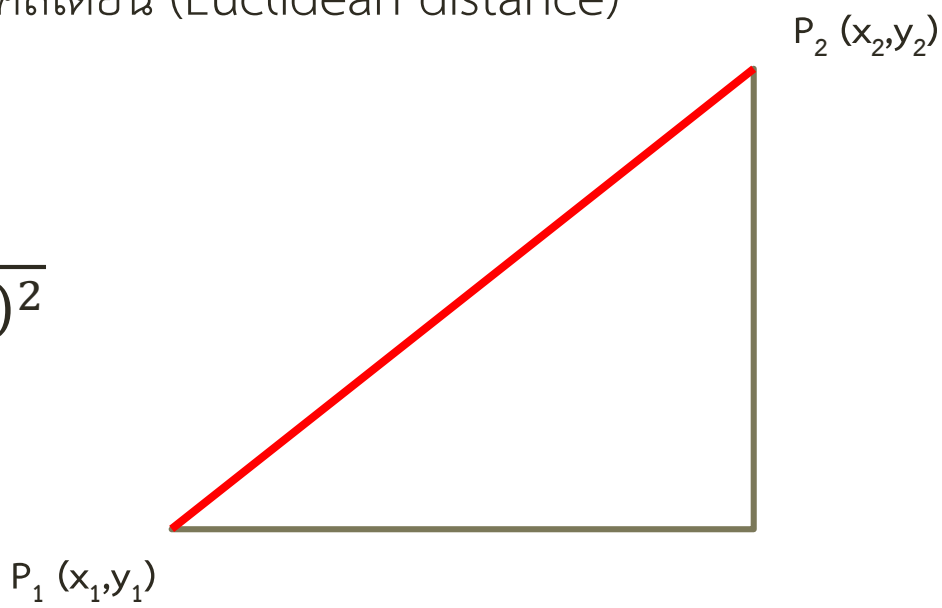


Clustering

การจัดข้อมูลให้อยู่ในกลุ่มต่าง ๆ จะต้องมีการวัดค่าความคล้ายคลึง (similarity) หรือค่าระยะห่าง (distance) ระหว่างข้อมูลแต่ละตัว (example)

วิธีการคำนวณค่าระยะห่างที่นิยมใช้ เช่น ระยะห่างยูคลิเดียน (Euclidean distance)

$$C = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$



Clustering

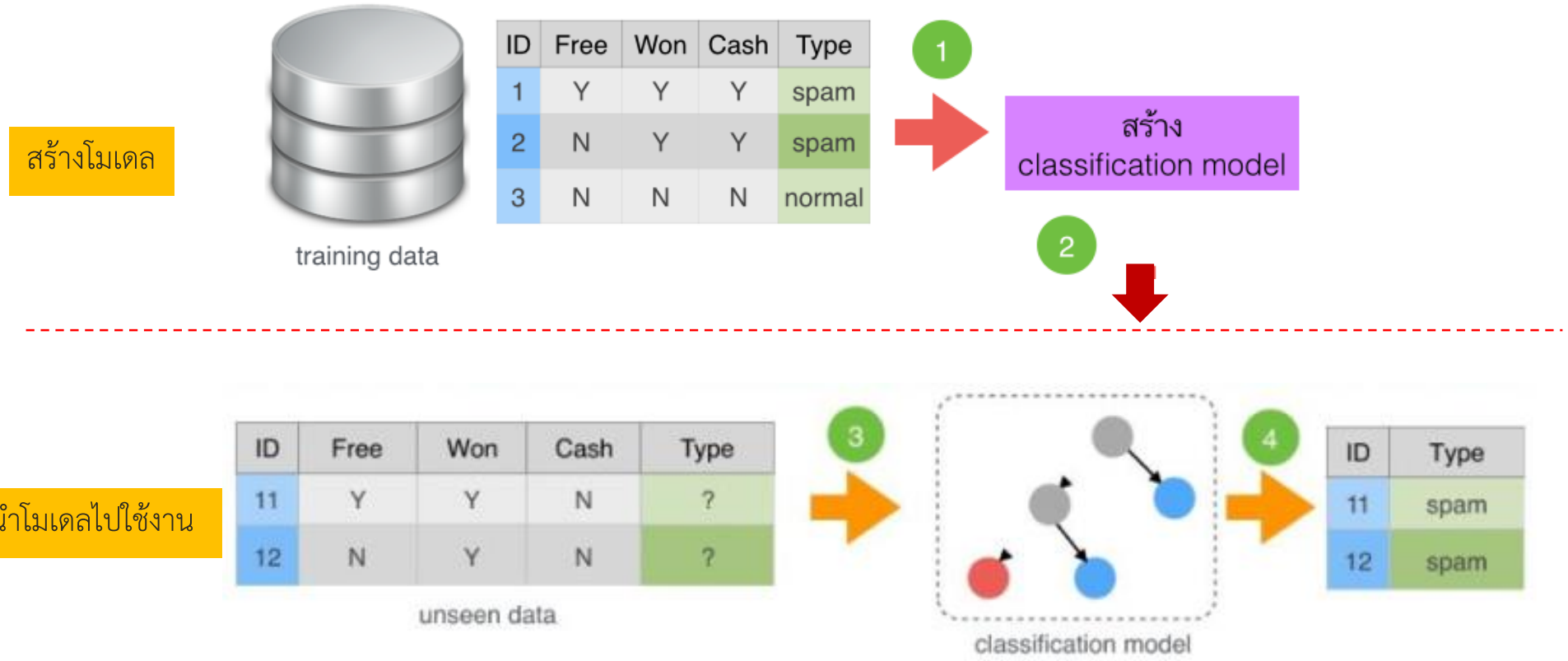
ในการทำ Clustering มีพารามิเตอร์ที่ต้องกำหนด คือ จำนวนกลุ่มที่ต้องการแบ่ง หรือจำนวนคลัสเตอร์ แทนด้วยตัวแปร K

ขั้นตอนการทำงาน

1. เลือกจำนวนของคลัสเตอร์ (K)
2. สุ่มเลือกจุดศูนย์กลาง (centroid) ขึ้นมาตามจำนวนคลัสเตอร์
3. กำหนดให้ข้อมูลอยู่ในคลัสเตอร์ที่ใกล้ที่สุด
4. คำนวณหาจุดศุนย์กลางแต่ละคลัสเตอร์ใหม่
5. ทำซ้ำข้อ 3 และ 4 ซ้ำ จนกระทั่ง centroid ไม่มีการเปลี่ยนแปลง

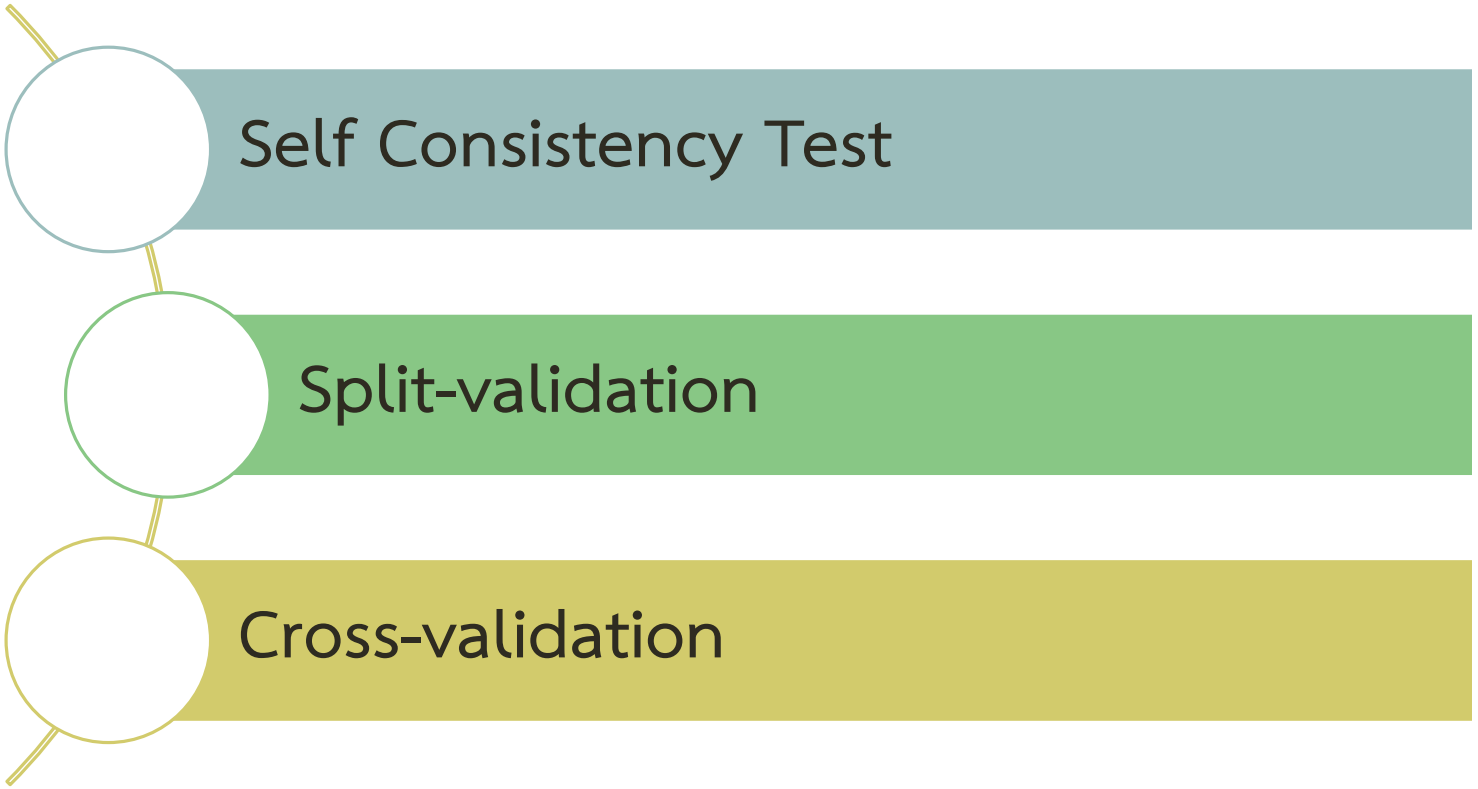
Apply Model
Validation Model

การประยุกต์ใช้โมเดล (Apply model)



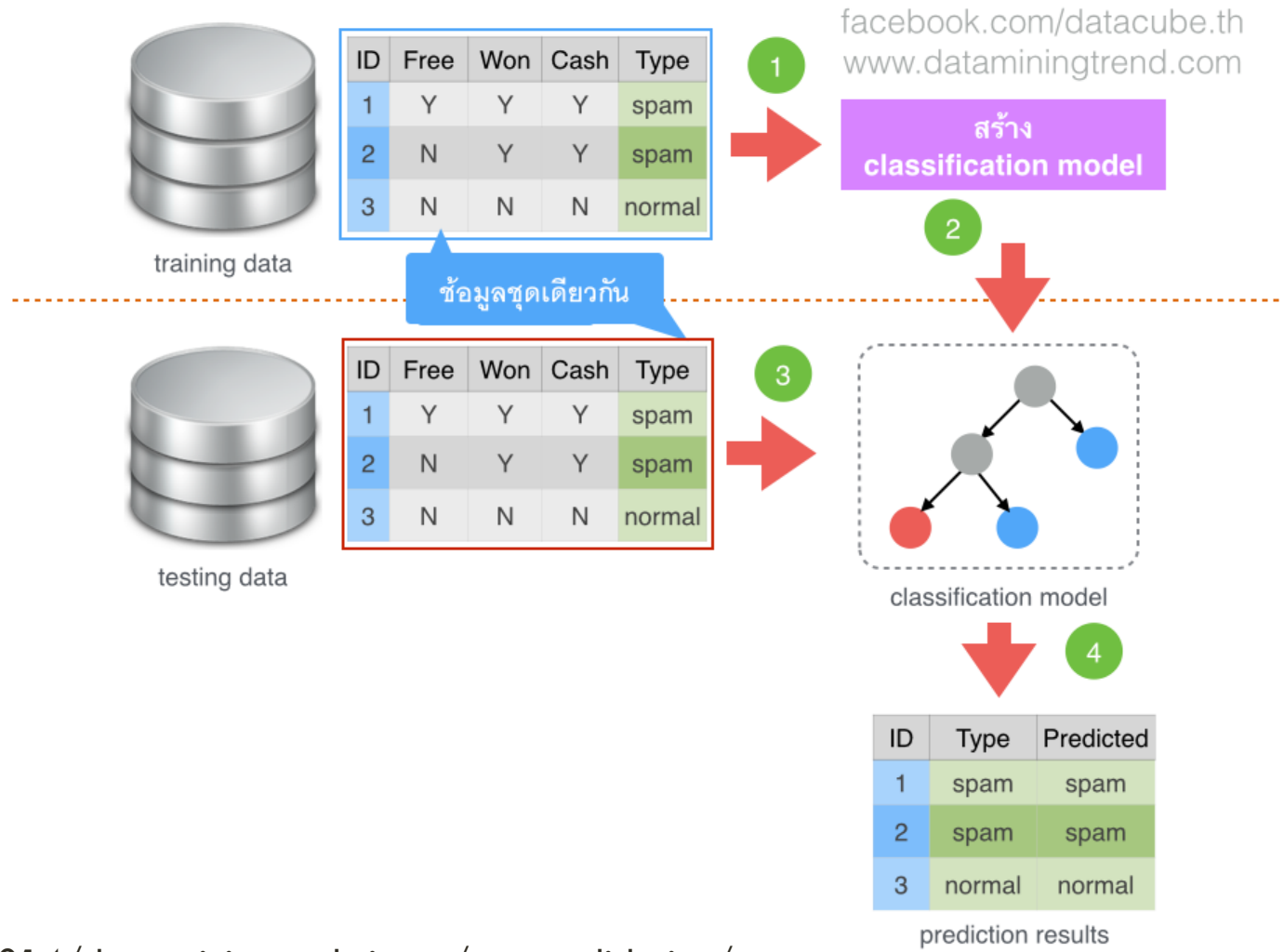
http://dataminingtrend.com/2014/data-mining-techniques/ensemble-model/

Validate Model



Self-consistency Test

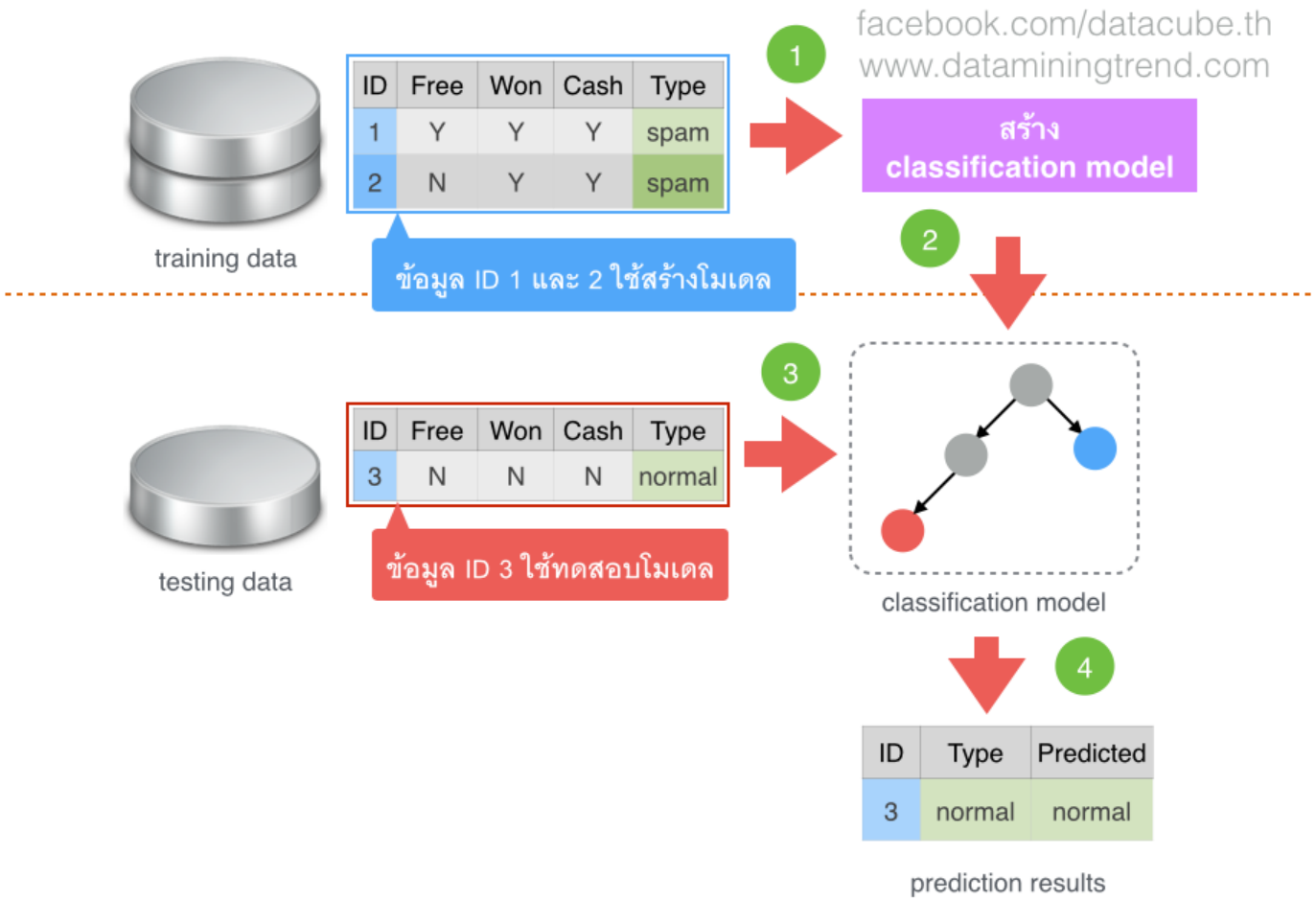
- ใช้ข้อมูล training ในการทดสอบประสิทธิภาพของโมเดล



http://dataminingtrend.com/2014/data-mining-techniques/cross-validation/

Split Test

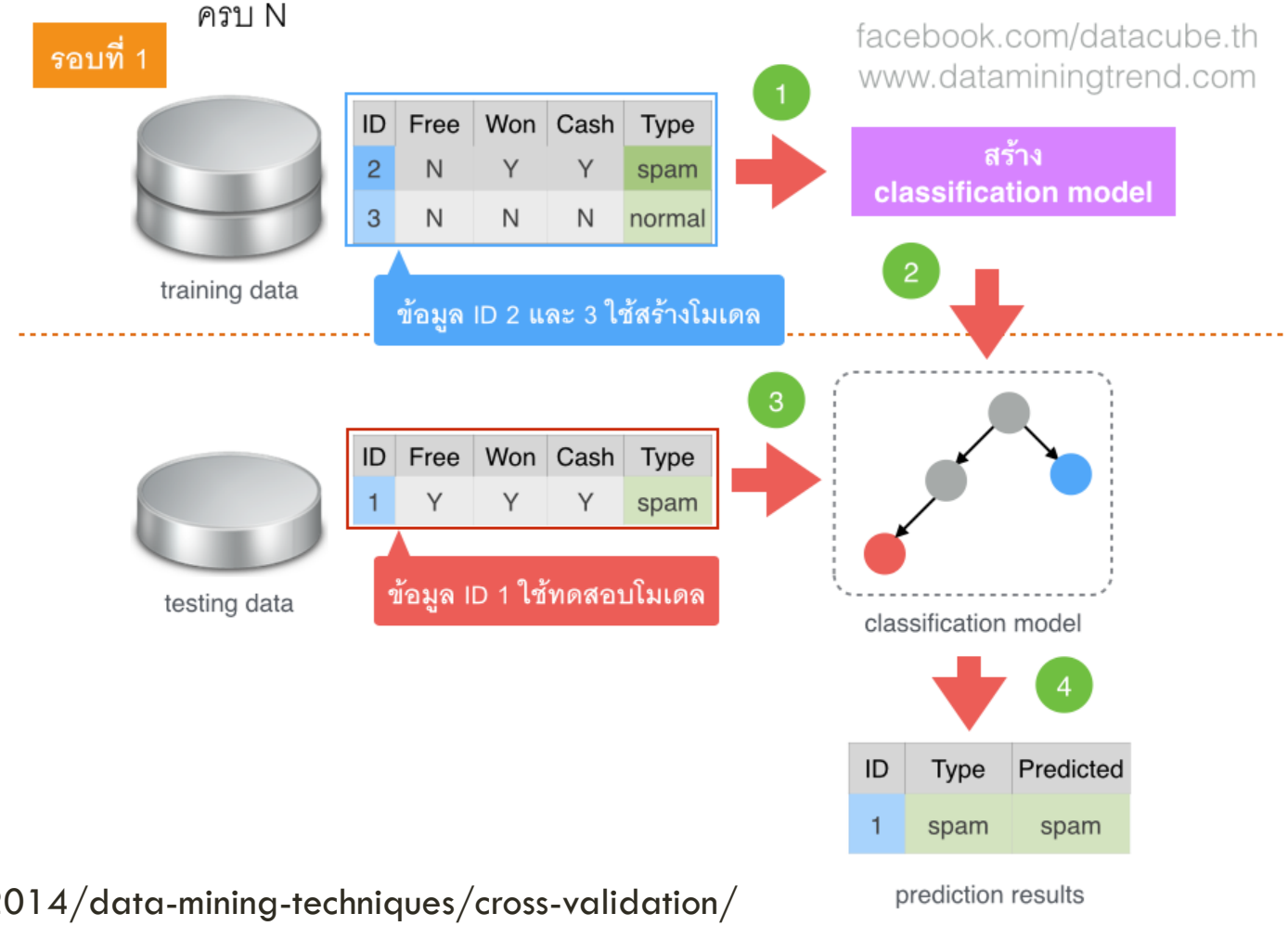
- แบ่งข้อมูลออกเป็น 2 ชุด
  - training data สำหรับสร้างโมเดล และ testing data สำหรับทดสอบ



http://dataminingtrend.com/2014/data-mining-techniques/cross-validation/

Cross-validation

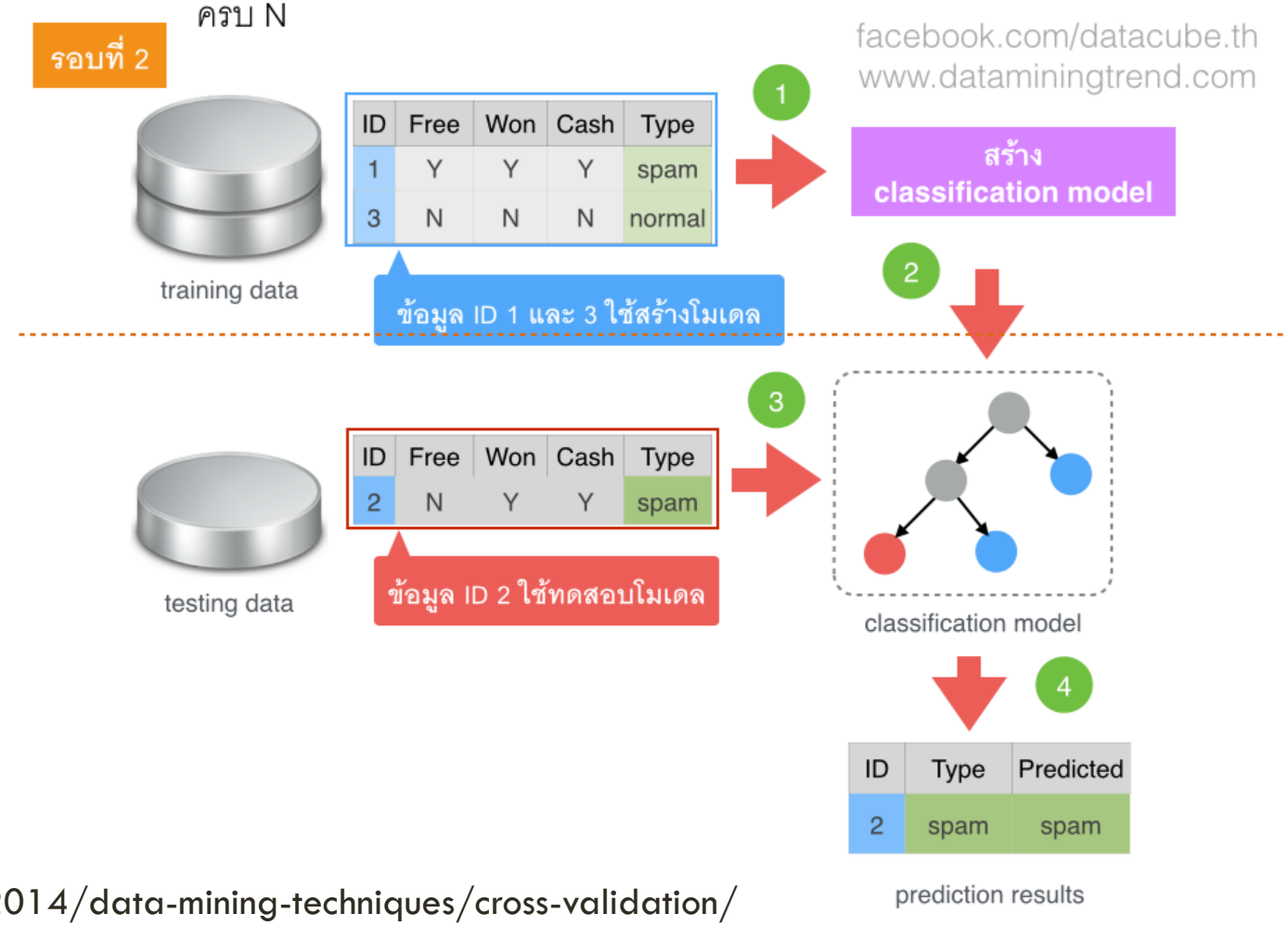
- แบ่งข้อมูลออกเป็น N ชุด เช่น N = 5 หรือ 10
  - ข้อมูล N-1 ชุดสำหรับสร้างโมเดล และ ข้อมูลส่วนที่เหลือสำหรับทดสอบ วนทำจนครบ N



http://dataminingtrend.com/2014/data-mining-techniques/cross-validation/

Cross-validation

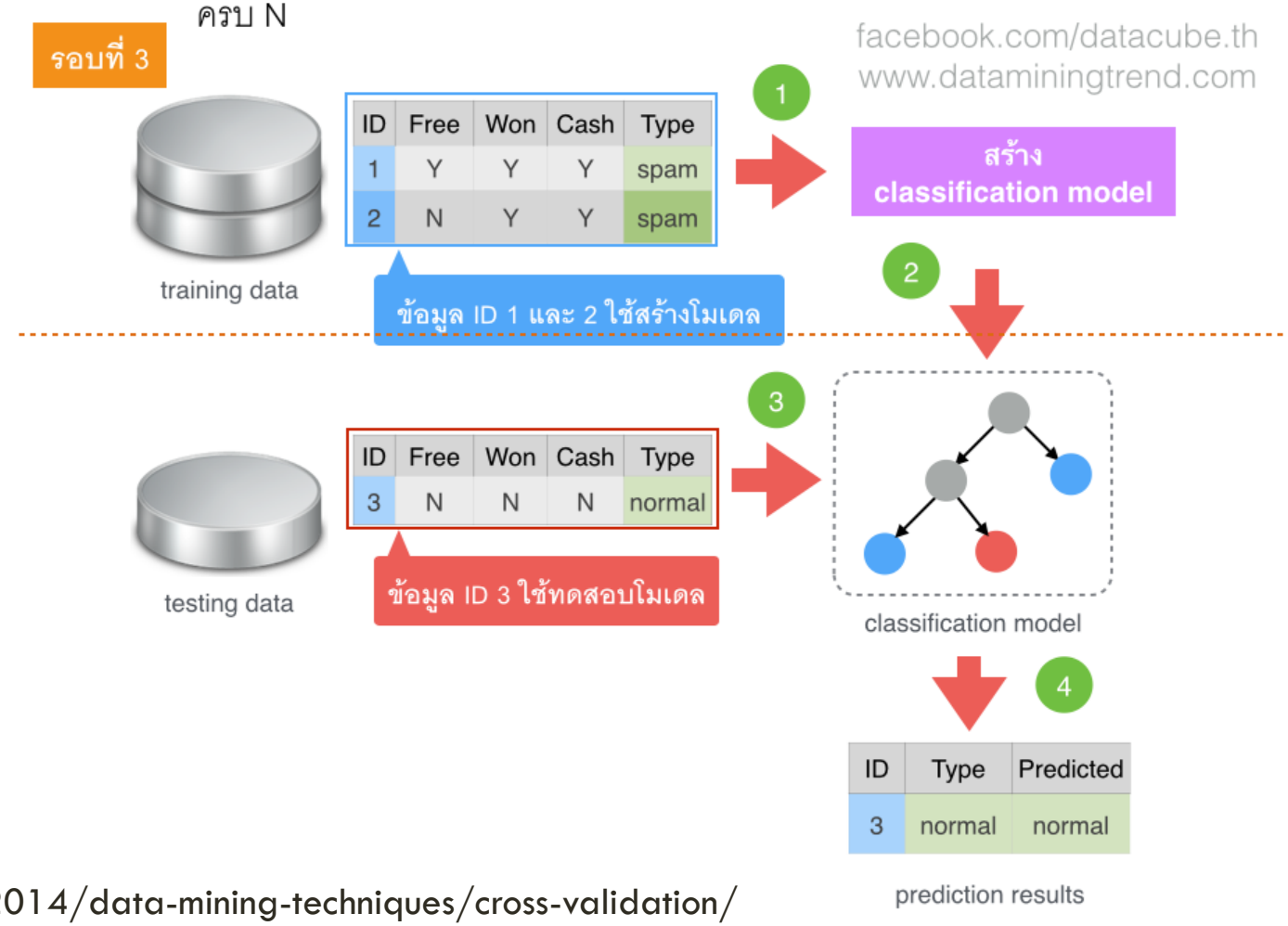
- แบ่งข้อมูลออกเป็น N ชุด เช่น N = 5 หรือ 10
  - ข้อมูล N-1 ชุดสำหรับสร้างโมเดล และ ข้อมูลส่วนที่เหลือสำหรับทดสอบ วนทำจนครบ N



http://dataminingtrend.com/2014/data-mining-techniques/cross-validation/

Cross-validation

- แบ่งข้อมูลออกเป็น N ชุด เช่น N = 5 หรือ 10
  - ข้อมูล N-1 ชุดสำหรับสร้างโมเดล และ ข้อมูลส่วนที่เหลือสำหรับทดสอบ วนทำจนครบ N

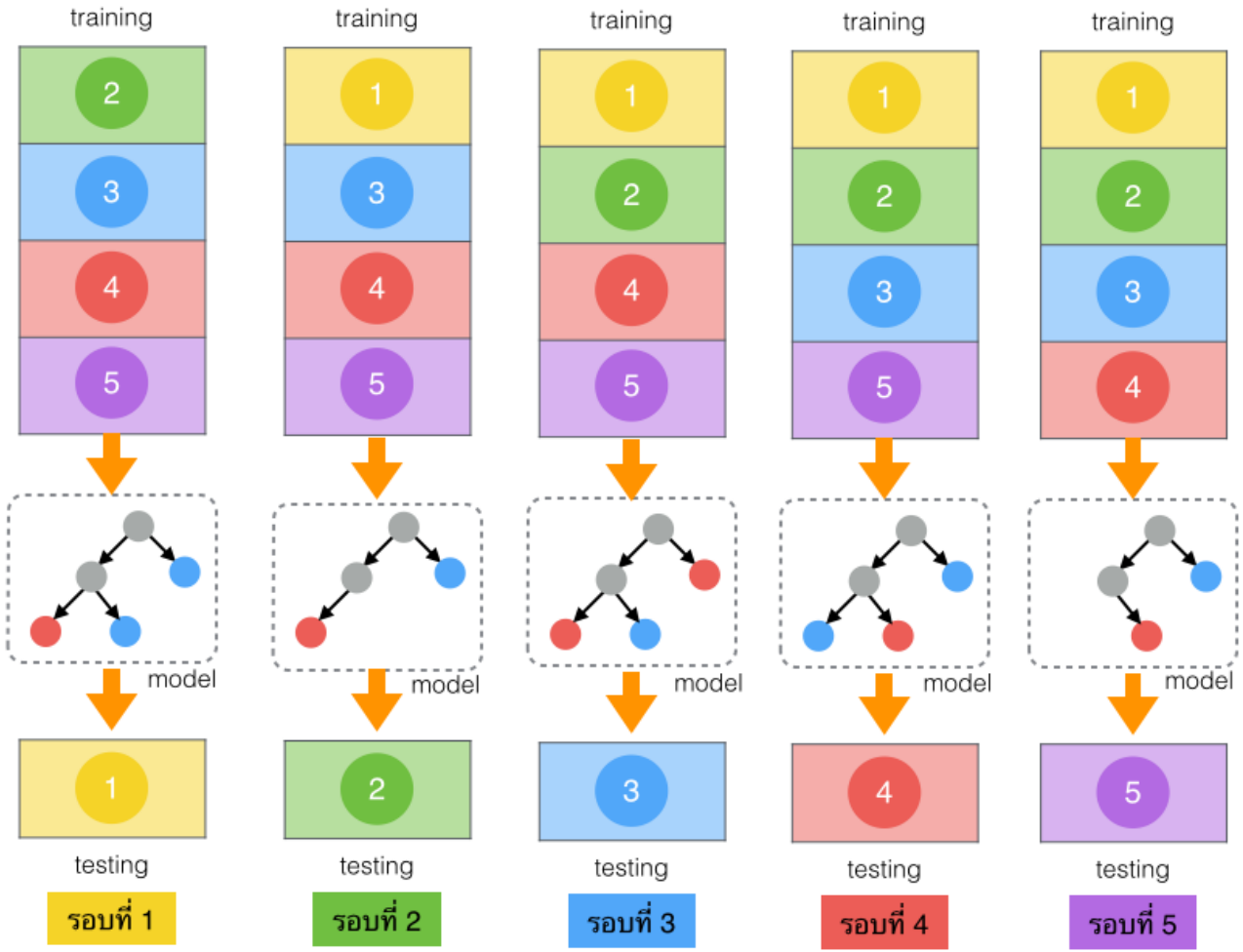


http://dataminingtrend.com/2014/data-mining-techniques/cross-validation/

Cross-validation

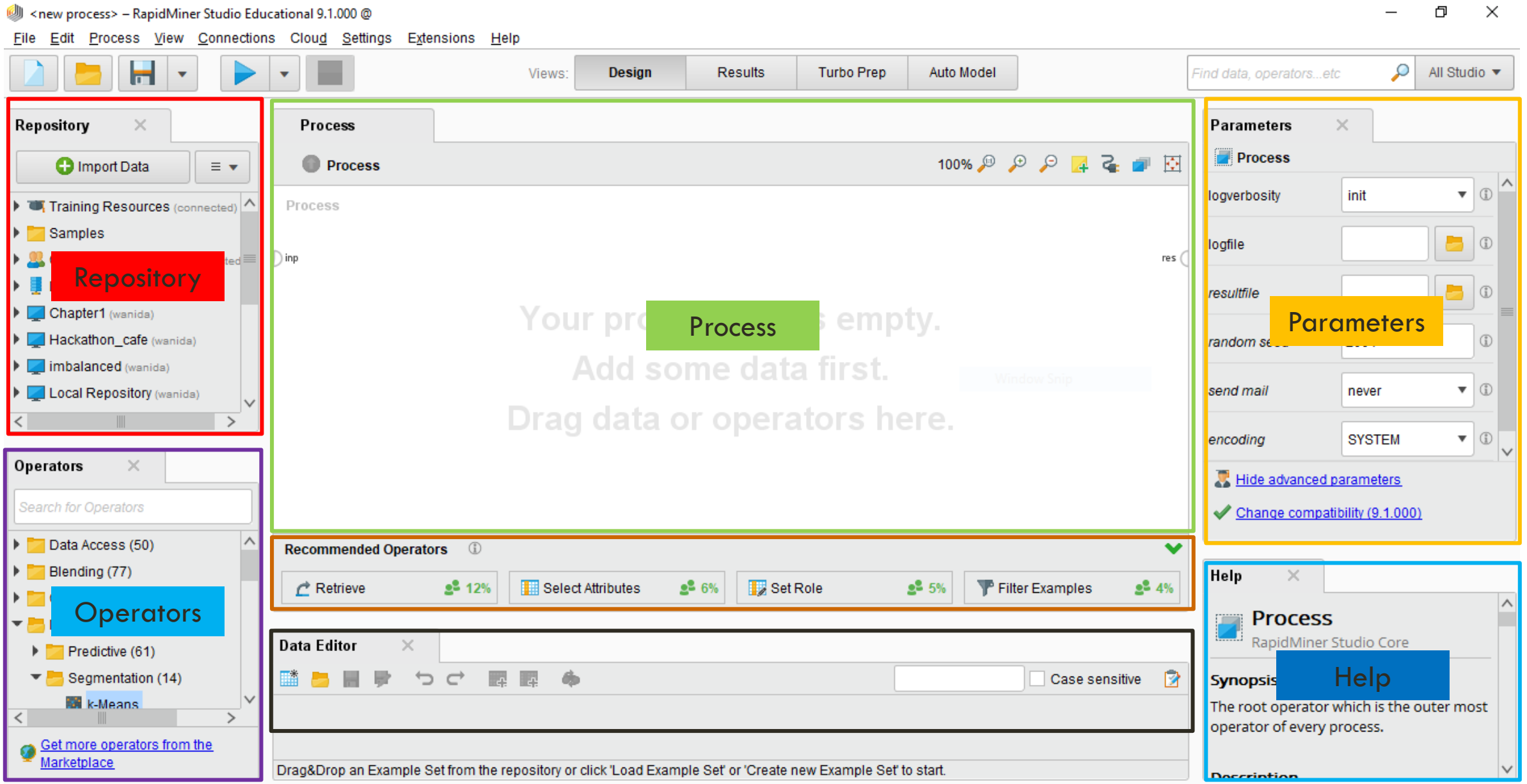
(data)^3
base | warehouse | mining

- ตัวอย่าง 5-fold cross-validation



http://dataminingtrend.com/2014/data-mining-techniques/cross-validation/

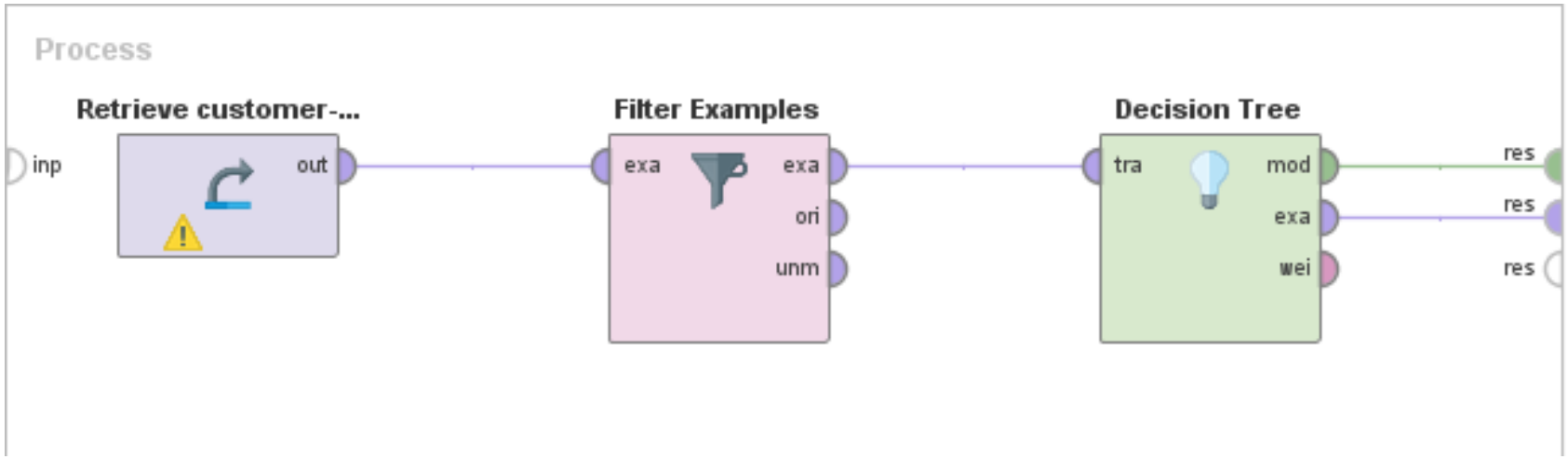
rapidminer


<new process> – RapidMiner Studio Educational 9.1.000 @
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Samples
Chapter1 (wanida)
Hackathon_cafe (wanida)
imbalanced (wanida)
Local Repository (wanida)
Repository
Process
Process
100%
inp
res
Process
Add some data first.
Drag data or operators here.
Parameters
Process
logverbosity
init
logfile
resultfile
send mail
never
encoding
SYSTEM
Hide advanced parameters
Change compatibility (9.1.000)
Operators
Search for Operators
Data Access (50)
Blending (77)
Predictive (61)
Segmentation (14)
k-Means
Get more operators from the Marketplace
Operators
Recommended Operators
Retrieve 12%
Select Attributes 6%
Set Role 5%
Filter Examples 4%
Data Editor
Case sensitive
Drag&Drop an Example Set from the repository or click 'Load Example Set' or 'Create new Example Set' to start.
Help
Process
RapidMiner Studio Core
Synopsis
The root operator which is the outer most operator of every process.
Help

Rapid Miner Studio

Input Ports (*inp*)

- example set *(exa)*
- training set *(tra)*



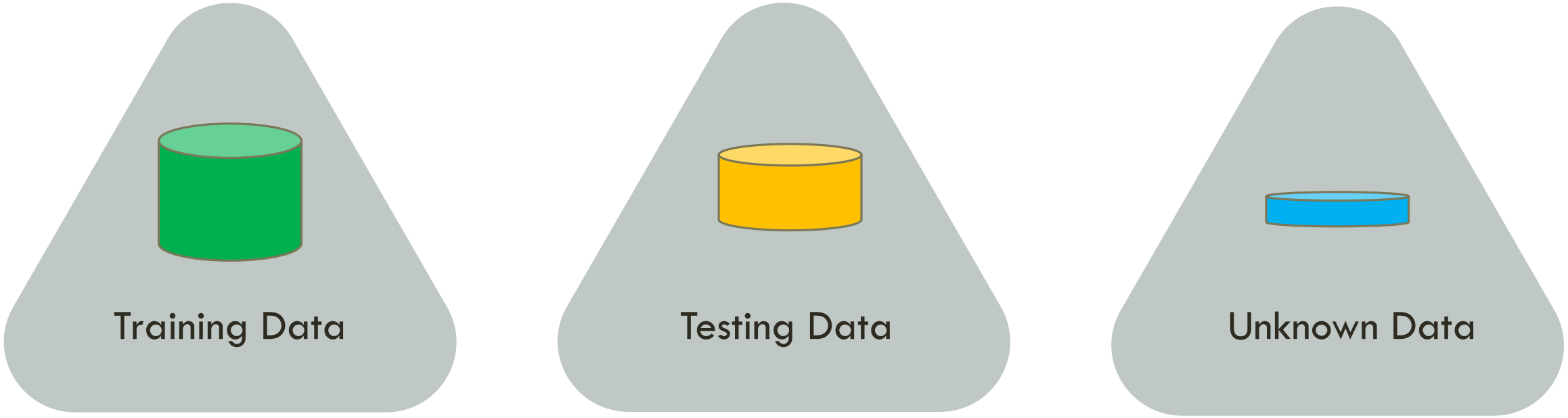
Output Ports (*res*)

- Output (out)
- model *(mod)*
- example set *(exa)*

Workshop

Decision tree
K-Means

การเตรียมข้อมูล



Modeling Process

Blending
Ex. Select attributes (เลือกคอลัมน์)
Filter examples (เลือกแถว)

Cleansing
Ex. Replace missing values (เติมข้อมูลที่เป็น missing values ด้วยค่าอื่น)

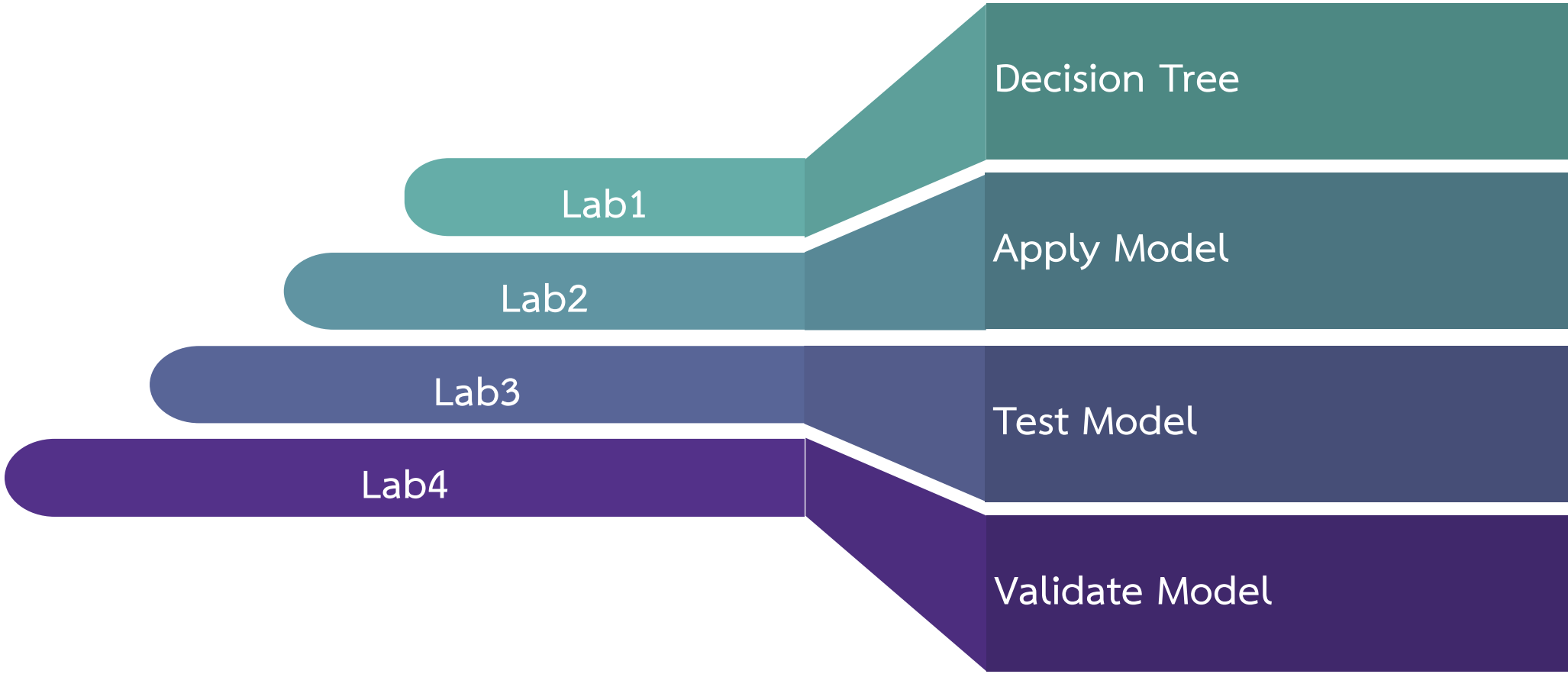
Modeling
Ex. Decision tree, Random Forest, k-means, Rules Induction, Deep learning

Scoring
Ex. Apply model

Validation
Ex. Cross validation, Split validation

Validation
Ex. Performance classification, Cluster Distance Performance

Workshop 1 Decision Tree



Workshop 2 K-means

- Lab1 K-means
- Lab2 Apply Model
- Lab3 Test Model
- Lab4 Validate Model