

Chapter 6: K-Nearest Neighbor

CP363107 Data Science for marketing

สอนโดย
รศ.ดร.วรารัตน์ สงฆ์แป้น

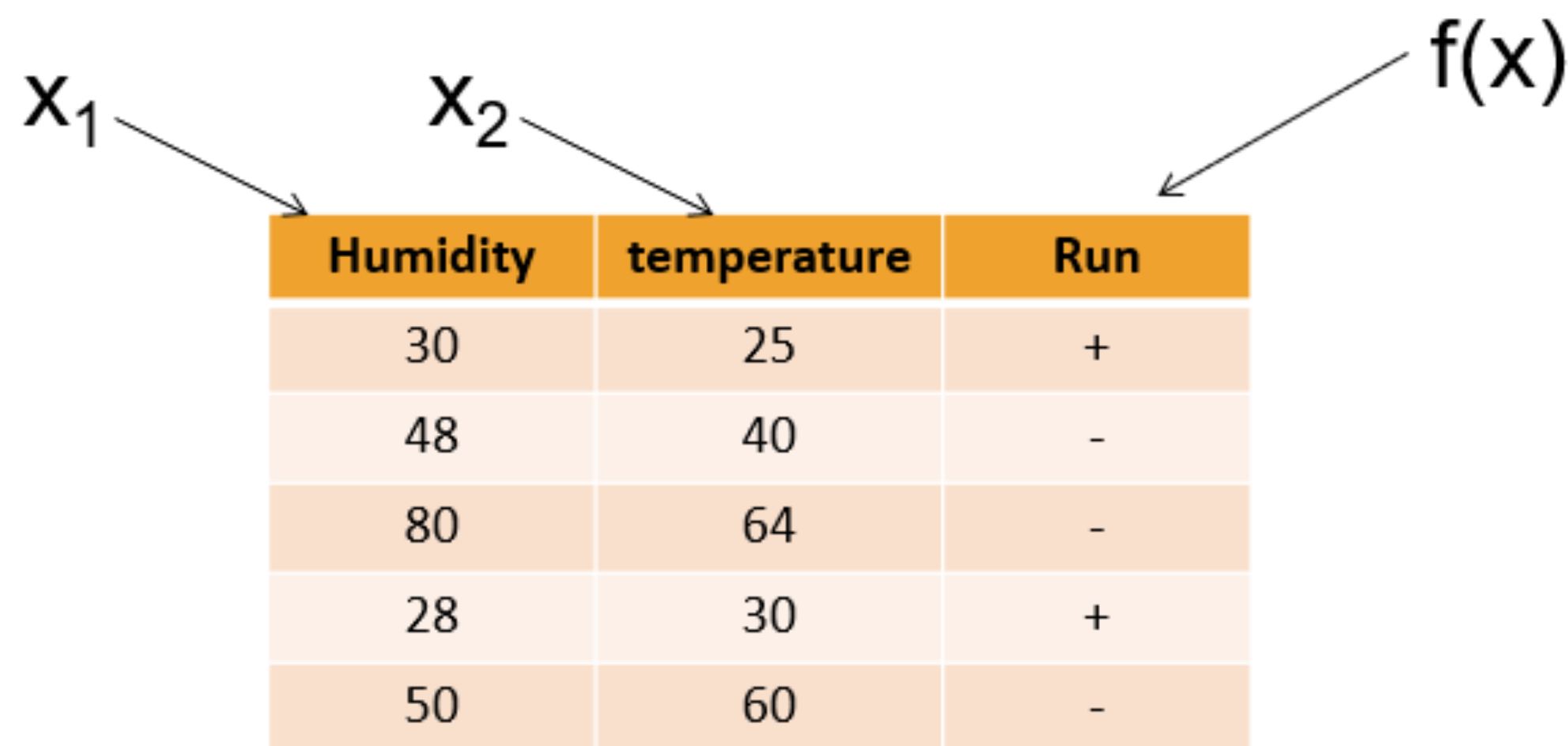
email: wararat@kku.ac.th

Department of Computer Science
College of Computing,
Khon Kaen University



K-Nearest Neighbor Classifier

Key idea: มีข้อมูลชุดการสอน (Training data) ให้เขียนอยู่ในรูป $\langle x_i, f(x_i) \rangle$ เช่น



The diagram shows a table with three columns: Humidity, temperature, and Run. Above the table, there are three labels: x_1 with an arrow pointing to the Humidity column, x_2 with an arrow pointing to the temperature column, and $f(x)$ with an arrow pointing to the Run column.

Humidity	temperature	Run
30	25	+
48	40	-
80	64	-
28	30	+
50	60	-

$\langle x_1, x_2, f(x) \rangle$ ตัวอย่างเช่น $\langle 30, 25, + \rangle$

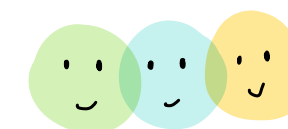
K-Nearest Neighbor Classifier

- ^{ค่าไม่ต่อเนื่อง} **Discrete-valued** หมายถึง ค่าป้ายบอกผลออกเป็นที่ยแบ่งประเภทชัดเจน เช่น วิ่ง หรือ ไม่วิ่ง ใช่ หรือ ไม่ใช่ เป็นต้น
 - ดังนั้นหาชุด x_q , ที่ใกล้เคียงที่สุดสำหรับชุดข้อมูลสอนมาเป็นตัวประมาณค่าสำหรับ x_n

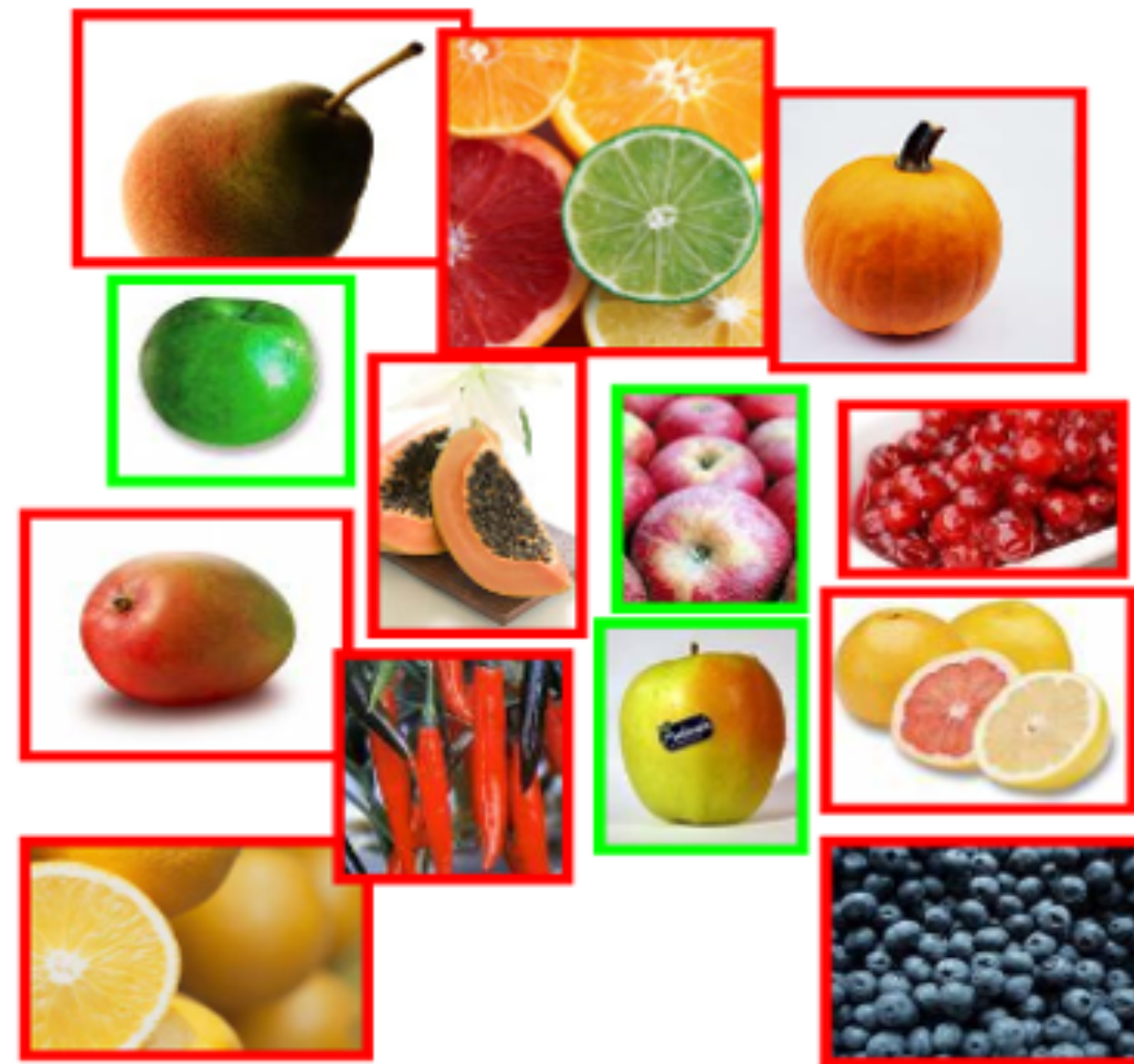
$$\hat{f}(x_q) \leftarrow f(x_n)$$

- **Real-valued** หมายถึง ค่าป้ายบอกผลออกเป็นตัวเลขทศนิยม เช่น การพยากรณ์ ปริมาณน้ำฝน อุณหภูมิ เป็นต้น

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^k f(x_i)}{k}$$



K-Nearest Neighbor Classifier

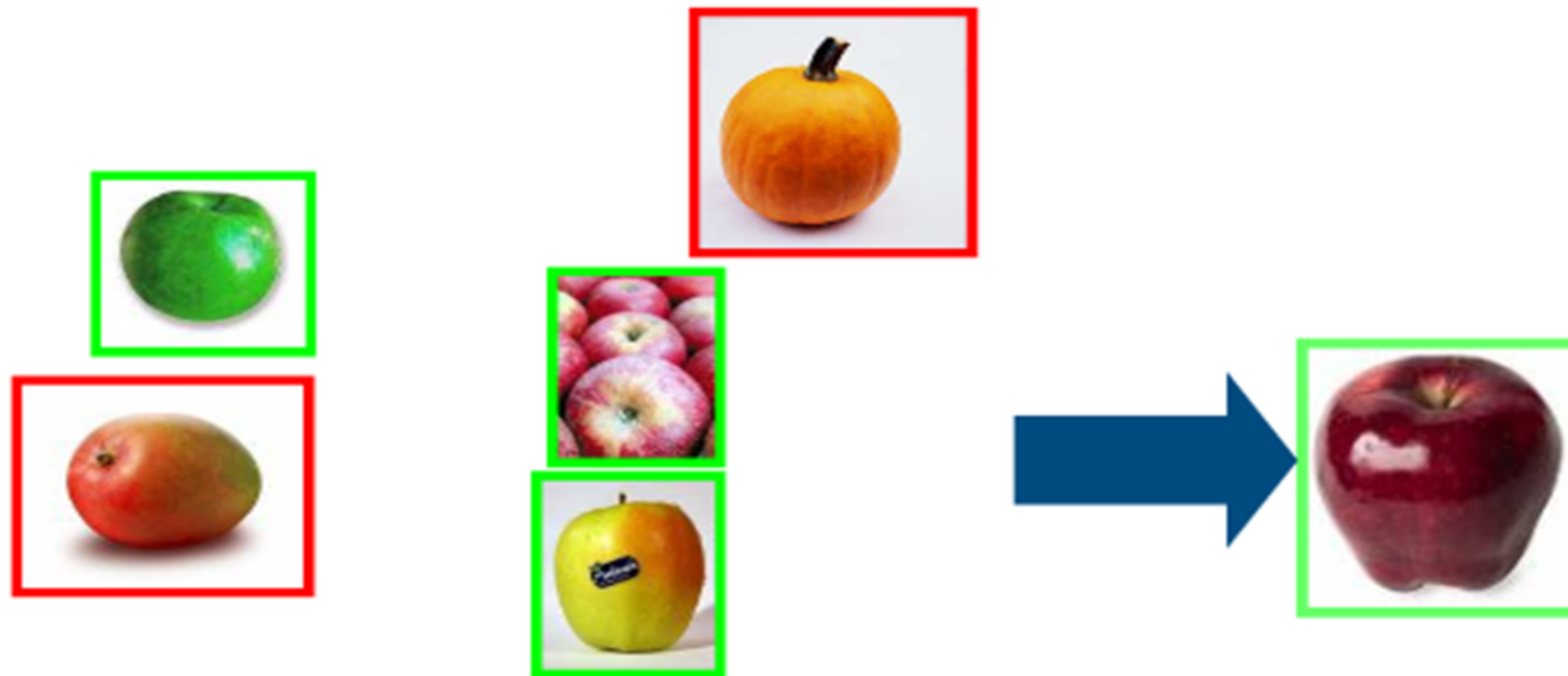


Is this an apple?



- ในการตัดสินใจผลไม้ที่ไม่เคยเห็นนั้นคือ แอปเปิ้ล นั่นคือเลือกพิจารณาภาพผลไม้จำนวน K ที่ใกล้เคียงมากที่สุด ดังนั้น การจำแนกประเภทผลไม้ที่ไม่ทราบจะใช้จำนวนผลโหวตของผลไม้แต่ละประเภทว่าเป็น แอปเปิ้ล

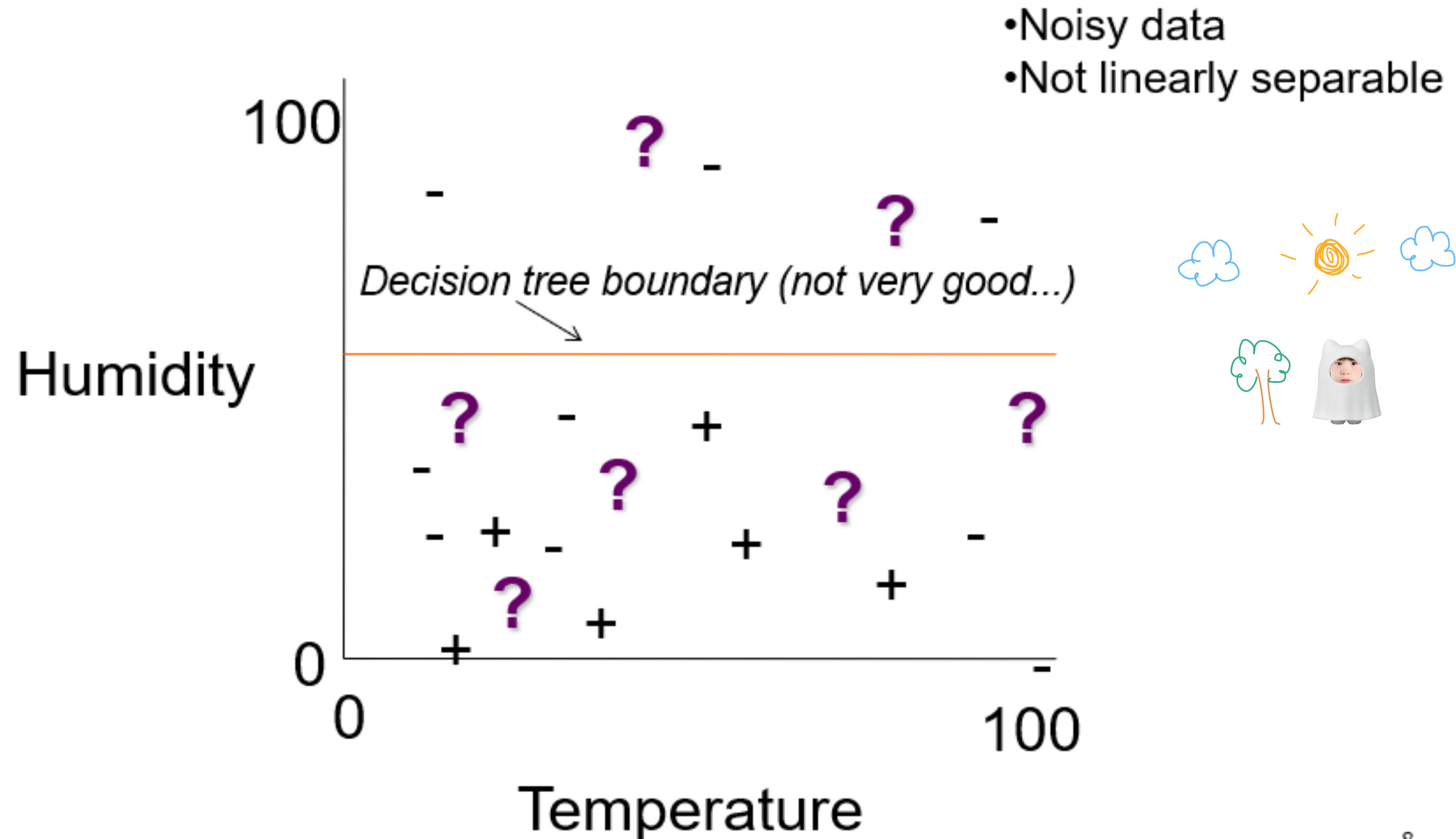
K-Nearest Neighbor Classifier



- ถ้า $k=5$, นั้นหมายถึงเลือกภาพผลไม้ 5 ภาพที่ใกล้เคียงมากที่สุด เพื่อบ่งบอกประเภทของต้นไม้ที่ต้องการแบ่งกลุ่ม
- ดังนั้นจากภาพจะเห็นได้ว่า ผลไม้ทั้ง 5 ภาพส่วนใหญ่เป็นภาพของ แอปเปิ้ล ดังนั้นจึงตอบผลไม้ว่าเป็น แอปเปิ้ล

ข้อเสียของ Decision Tree Classifier

20.00



การปรับค่า K มีผลต่อคำตอบ

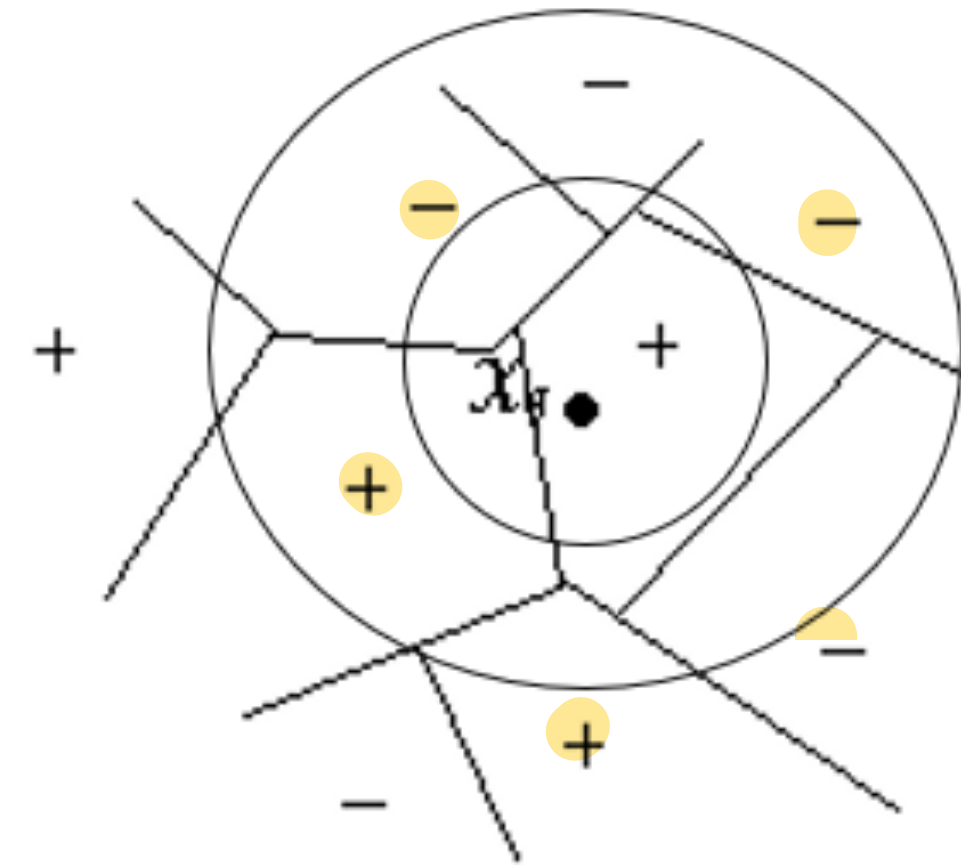
(Discrete-valued)

$k=1$

1-NN classifies x_q as + ✓

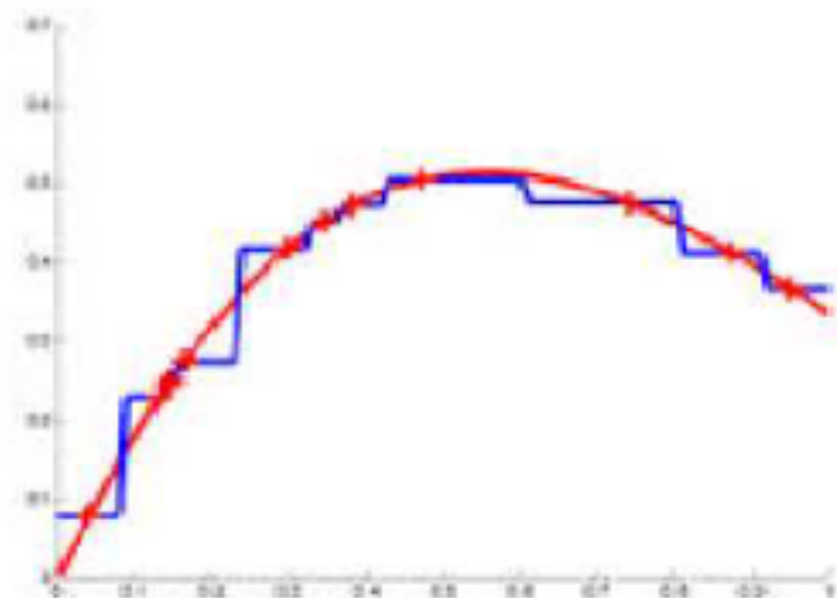
$k=5$

5-NN classifies x_q as -

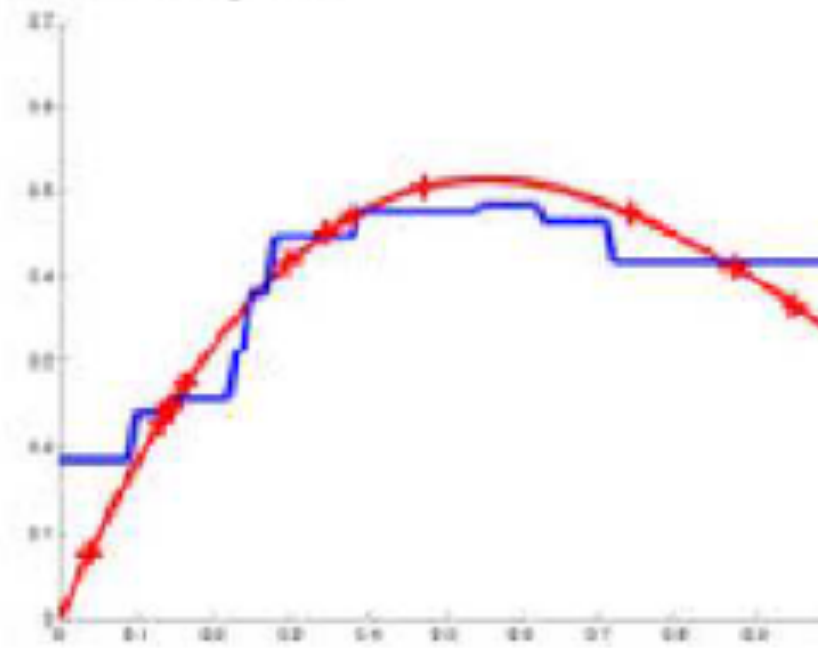


(real-valued target function)

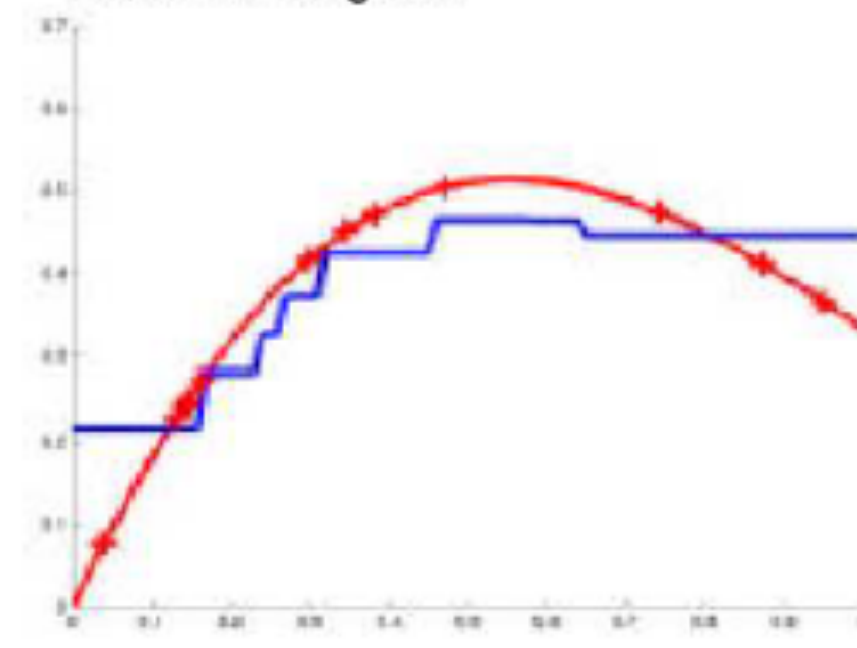
1-nearest neighbor



3-nearest neighbor



5-nearest neighbor



การปรับค่า K มีผลต่อคำตอบ

- ไม่ควรเลือก K เล็กเกินไป เพราะจะทำให้เบี่ยงเบนสูง
- ไม่ควรเลือก K ใหญ่เกินไป เพราะจะทำให้ข้อมูลเกินความจำเป็น
- เพราะฉะนั้นการเลือกค่า K ขึ้นอยู่กับข้อมูล ต้องมีการปรับค่าการประเมิน เช่น

Cross-validation

- ระยะทางที่ใช้วัด คือ
 - ถ้า \mathbf{x} ประกอบไปด้วย Attribute $\langle a_1(x), a_2(x), \dots, a_n(x) \rangle$ ดังนั้น $a_r(x)$ ดังกล่าว
จึงแทนด้วยค่าในด้วย \mathbf{x}
ค่าระยะทางที่ใช้ เรียกว่า Euclidean Distance

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^n (a_r(x_i) - a_r(x_j))^2}$$

K-NN Classifier

39 29

Discrete values

Humidity	temperature	Run
30	25	+
48	40	-
80	64	-
28	30	+
50	60	-

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^n (a_r(x_i) - a_r(x_j))^2}$$

$x = \langle \text{humidity, temperature} \rangle$

New instance $x_q = \langle 40, 30, \text{run}=? \rangle$ We can run inside(+) or outside (-)

9.11.11

$$d(x_q, x_1) = \sqrt{(40 - 30)^2 + (30 - 25)^2} = 11.18$$

$$d(x_q, x_2) = \sqrt{(40 - 48)^2 + (30 - 40)^2} = 12.80$$

$$d(x_q, x_3) = \sqrt{(40 - 80)^2 + (30 - 64)^2} = 52.5$$

$$d(x_q, x_4) = \sqrt{(40 - 28)^2 + (30 - 30)^2} = 12$$

$$d(x_q, x_5) = \sqrt{(40 - 50)^2 + (30 - 60)^2} = 31.62$$

1-NN (x_1)
Answer run inside(+)

2-NN (x_1, x_4)
Answer run inside(+)

3-NN (x_1, x_2, x_4)
Answer run inside (+)

4-NN (x_1, x_2, x_4, x_5)
Answer run inside (+)

5-NN
Answer run inside(-)

ผลในข้อสอบ

K-NN Regressor

Real values

Humidity	temperature	Rainfall
30	25	5.1
48	40	15.5
80	64	20.2
28	30	3.2
50	60	12.0

$x = \langle \text{humidity, temperature} \rangle$
 New instance $x_q = \langle 40, 30, \text{Rainfall} = ?? \rangle$

$$d(x_q, x_1) = \sqrt{(40 - 30)^2 + (30 - 25)^2} = 11.18$$

$$d(x_q, x_2) = \sqrt{(40 - 48)^2 + (30 - 40)^2} = 12.80$$

$$d(x_q, x_3) = \sqrt{(40 - 80)^2 + (30 - 64)^2} = 52.5$$

$$d(x_q, x_4) = \sqrt{(40 - 28)^2 + (30 - 30)^2} = 12$$

$$d(x_q, x_5) = \sqrt{(40 - 50)^2 + (30 - 60)^2} = 31.62$$

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^n (a_r(x_i) - a_r(x_j))^2}$$

1-NN (x_1)
 Rainfall = 5.1

2-NN (x_1, x_4)
 Rainfall = $(5.1 + 3.2)/2 = 4.15$

3-NN (x_1, x_2, x_4)
 Rainfall = $(5.1 + 15.5 + 3.2)/3 = 7.9$

4-NN (x_1, x_2, x_4, x_5)
 Rainfall = $(5.1 + 15.5 + 3.2 + 12.0)/4 = 8.95$

5-NN (x_1, x_2, x_3, x_4, x_5)
 Rainfall = $(5.1 + 15.5 + 3.2 + 20.2 + 12.0)/5 = 11.2$

จำนวน ๗ ตัว

ค่า

K-NN แบบ Distance Weight

ถ้าต้องการให้มีการประมาณค่าได้รายละเอียดมากขึ้น ดังนั้นจึงต้องคำนวณ
ค่าน้ำหนักสำหรับการแบ่งประเภท ดังต่อไปนี้

where

$$w_i \equiv \frac{1}{d(x_q, x_i)^2}$$

and $d(x_q, x_i)$ is distance between x_q and x_i

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^k w_i f(x_i)}{\sum_{i=1}^k w_i}$$

K-NN แบบ Distance Weight

ถ้าต้องการให้มีการประมาณค่าได้รายละเอียดมากขึ้น ดังนั้นจึงต้องคำนวณ
ค่าน้ำหนักสำหรับการแบ่งประเภท ดังต่อไปนี้

where

$$w_i \equiv \frac{1}{d(x_q, x_i)^2}$$

and $d(x_q, x_i)$ is distance between x_q and x_i

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^k w_i f(x_i)}{\sum_{i=1}^k w_i}$$

K-NN and Distance Weight

Humidity	temperature	Run
30	25	1
48	40	0
80	64	0
28	30	1
50	60	0

$x = \langle \text{humidity, temperature} \rangle$

New instance $x_q = \langle 40, 30 \rangle$ We can run inside(+) or outside (-) , by $k=3$

$$\hat{f}(x_q) = \frac{w_1 f(x_1) + w_2 f(x_2) + w_3 f(x_3)}{w_1 + w_2 + w_3}$$

$$\hat{f}(x_q) = \frac{\left(\frac{1}{11.18^2}\right) * (1) + \left(\frac{1}{12.8^2}\right) * (0) + \left(\frac{1}{12^2}\right) * (1)}{\left(\frac{1}{11.18^2}\right) + \left(\frac{1}{12.8^2}\right) + \left(\frac{1}{12^2}\right)} = 0.68 \approx 1$$

K-NN and Distance Weight

Humidity	temperature	Run
30	25	1
48	40	0
80	64	0
28	30	1
50	60	0

$x = \langle \text{humidity, temperature} \rangle$

New instance $x_q = \langle 40, 30 \rangle$ We can run inside(+) or outside (-) , by $k=3$

$$\hat{f}(x_q) = \frac{w_1 f(x_1) + w_2 f(x_2) + w_3 f(x_3)}{w_1 + w_2 + w_3}$$

$$\hat{f}(x_q) = \frac{\left(\frac{1}{11.18^2}\right) * (1) + \left(\frac{1}{12.8^2}\right) * (0) + \left(\frac{1}{12^2}\right) * (1)}{\left(\frac{1}{11.18^2}\right) + \left(\frac{1}{12.8^2}\right) + \left(\frac{1}{12^2}\right)} = 0.68 \approx 1$$

Quiz พฤติกรรมคนใช้ Facebook

Sex	Occupation	Checkin	TypeOfPost	ClickAds
Male	Student	Education	Text	Camera
Female	Student	Bar	Video	Cosmetic
Female	Student	Bar	Text	Cosmetic
Male	Programmer	Bar	Text	Camera
Male	Student	Travel	Live	Camera

New instance $x_q = \langle \text{Female, Programmer, Travel, Live, ClickAds=?} \rangle$

เช่น ฟอนกัน = 9 กลั = 0
ไม่เหมือน = 1 กลั = 1

Use $k=1, 3$ or 5

จาก K-NN ทำไมใช้กับระบบ Recommendation

- **ช่วยคัดเลือกของที่คิดว่าผู้ใช้จะสนใจ**

สินค้าในระบบออนไลน์ต่างๆ หากมีมากมายมหาศาล การจะให้ผู้ใช้เลือกเองทั้งหมดนั้นก็จะเป็นการเปลืองเวลาและเปลือง resource ไปมาก ซึ่งเมื่อผู้ใช้เข้ามาในระบบเราแล้วค้นหาของไม่เจอ หรือของเยอะเกินไป สุดท้ายก็อาจจะเกิดความรู้สึกไม่อยากใช้เว็บเรา ดังนั้นระบบ recommendation จึงช่วยคัดเลือกสินค้าที่คิดว่าผู้ใช้จะสนใจ แทนที่จะนำเสนอทั้งหมด

- **ช่วยแนะนำสินค้าอื่นๆมาให้ผู้ใช้**

ตัวอย่างเช่น อย่างนี้ที่การที่เราแนะนำเสนอของอื่นๆที่ยังเกี่ยวข้องกับผู้ใช้อยู่ ซึ่งก็จะเป็นการเปิดโอกาสให้ผู้ใช้那儿ได้รู้จักกับสินค้าใหม่ๆ จากระบบ recommendation ด้วย

จะแนะนำได้ต้องดู Feedback

- ประเภทของ Feedback มาจาก Rating แบ่งเป็น 2 ประเภท คือ

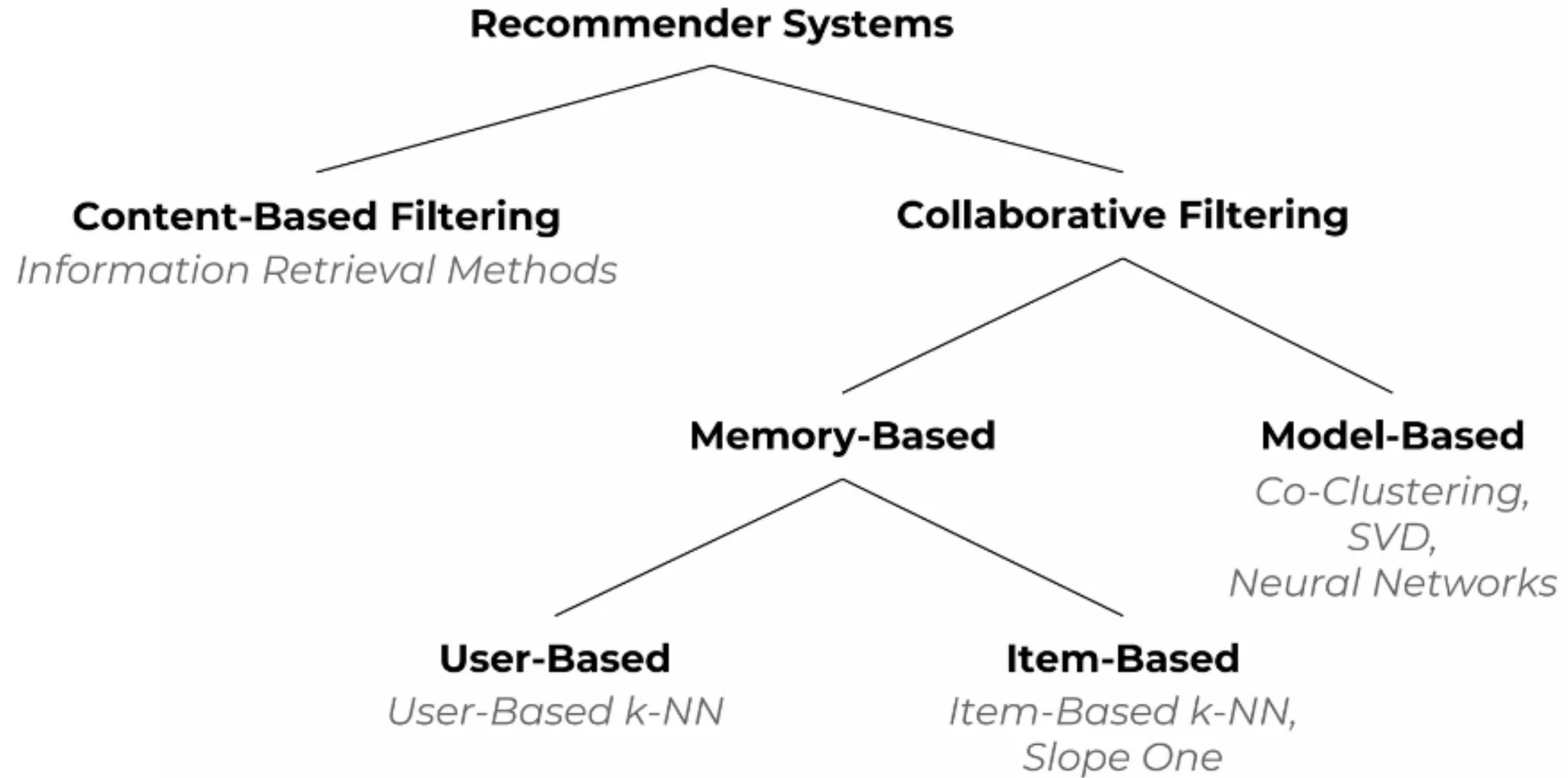
1. Explicit Rating

การให้คะแนนรีวิวต่างๆ เช่น ผู้ใช้ A อาจจะให้คะแนนรีวิวสินค้า B ด้วยคะแนน 8/10 คะแนน เราก็จะสามารถรู้ได้ว่าผู้ใช้ A นั้นค่อนข้างชอบสินค้า B หรือ อาจจะให้คะแนนสินค้า C แค่ 1/10 คะแนน เราก็จะพอทราบได้ว่าผู้ใช้ A ไม่ชอบสินค้า C เท่าไหร่ เป็นต้น

2. Implicit Rating

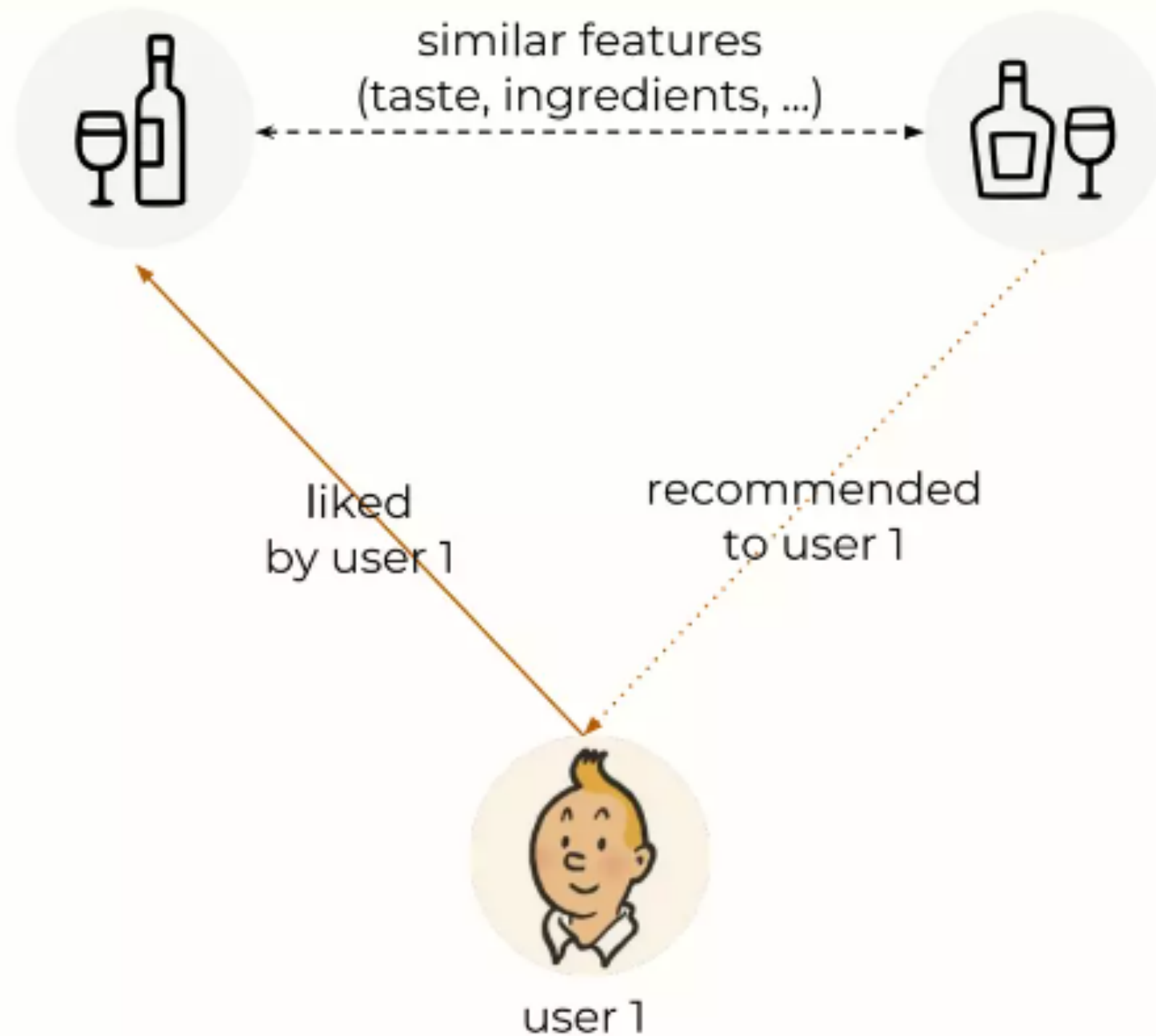
ช่องทางอื่นอีกที่เราจะพอสามารถเดาๆได้ว่าผู้ใช้ชอบหรือมีความสนใจในสินค้าของเราหรือไม่ ตัวอย่างเช่น การคลิกเข้าไปดูสินค้านั้น หรือการตัดสินใจที่จะซื้อสินค้านั้น หรือการที่ผู้ใช้กด favorite เป็นต้น ซึ่ง rating ประเภทนี้นั้น จะมีปริมาณข้อมูลที่เยอะมากกว่า explicit rating มาก รวมไปถึงเมื่อมีผู้ใช้ใหม่เข้ามา เราก็สามารถเก็บข้อมูลนี้ได้เลยทันทีเมื่อเค้าเริ่มคลิก แต่ feedback ที่ได้นั้นก็จะไม่ชัดเจนเท่า explicit rating

ประเภทโมเดล Recommendation

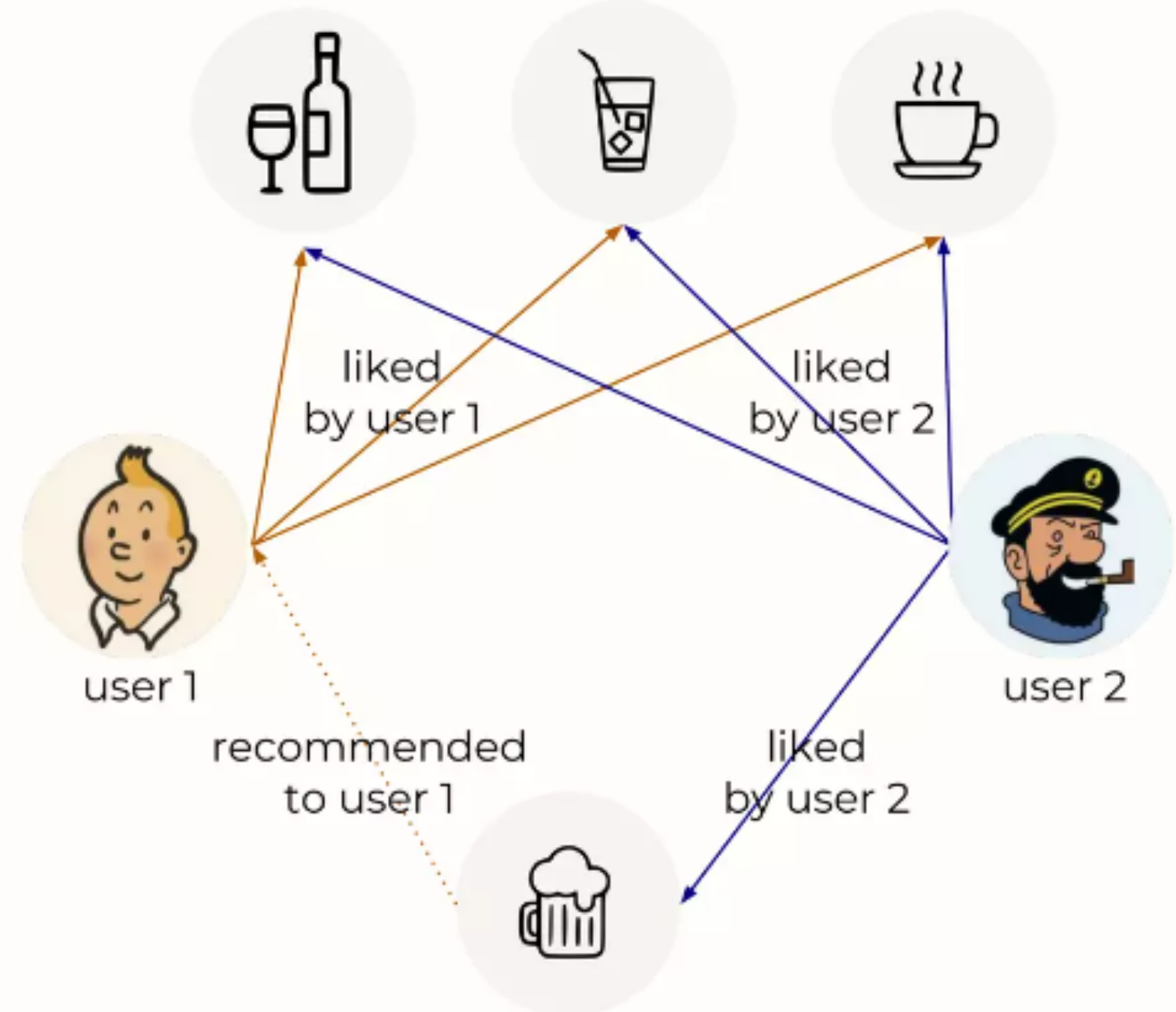


Content Based VS Collaborative Filtering

Content-Based Filtering

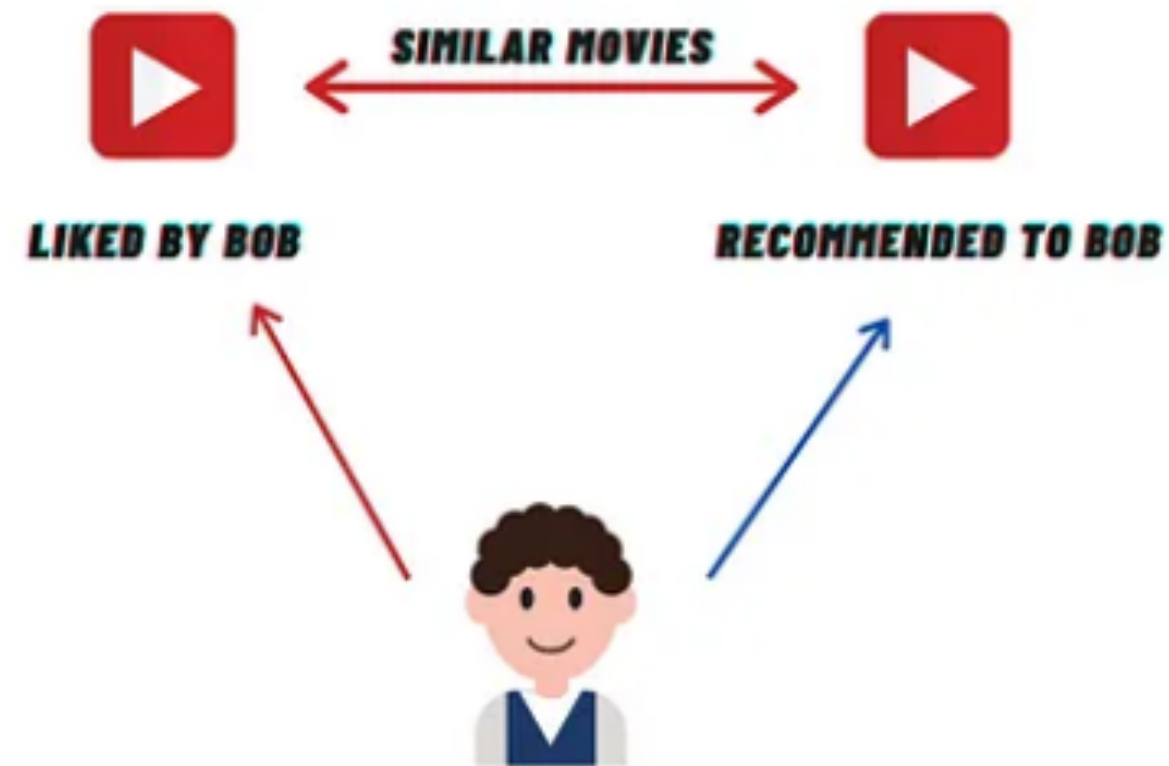


Collaborative Filtering

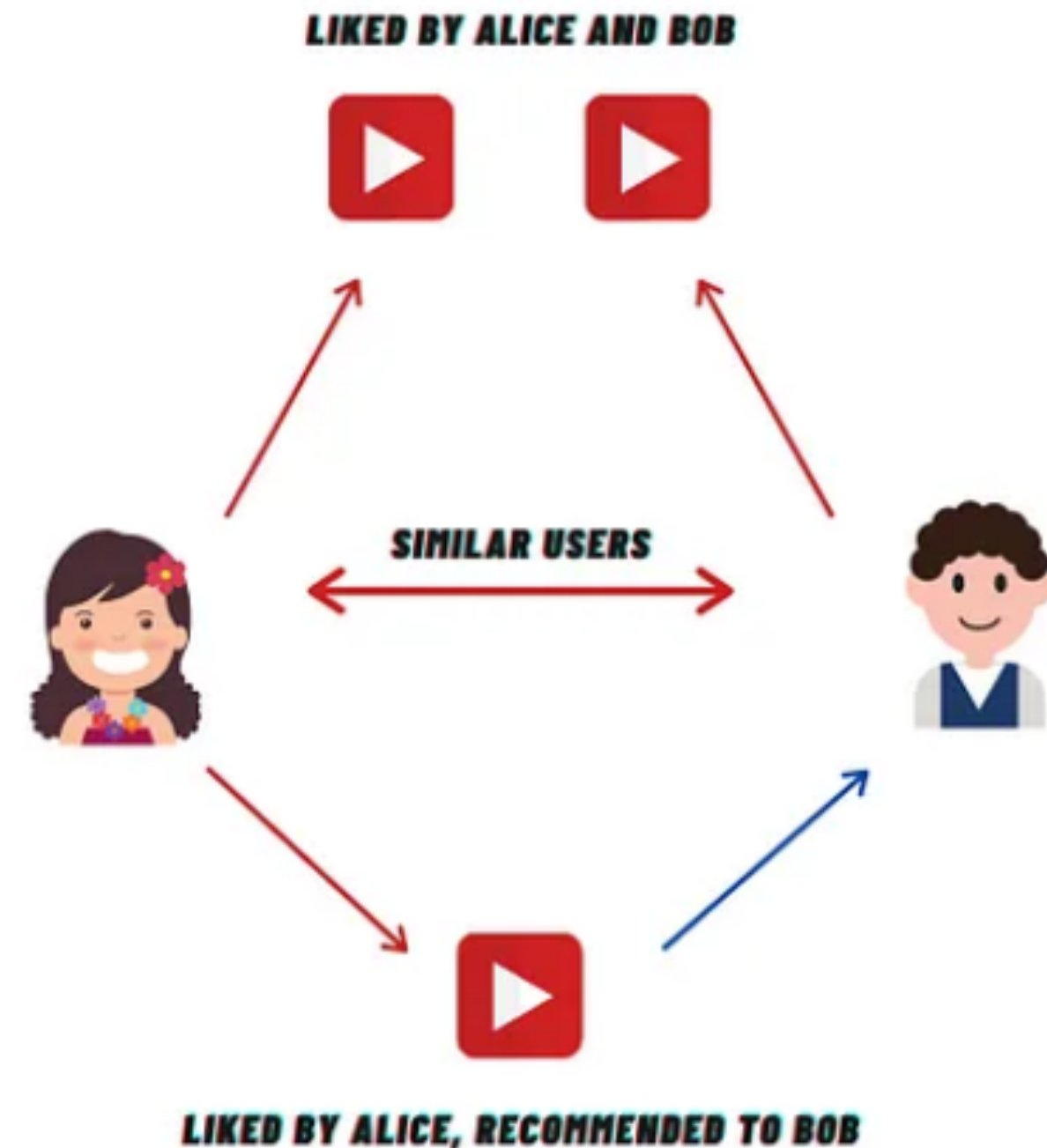


Content Based VS Collaborative Filtering

CONTENT-BASED FILTERING



COLLABORATIVE FILTERING



Collaborative Filtering

เป็นหนึ่งใน Algorithm ที่ใช้ในการสร้าง Patterns หรือ Rules ในการแนะนำสินค้า ซึ่งแบ่งได้ง่ายๆ เป็น 2 แบบ

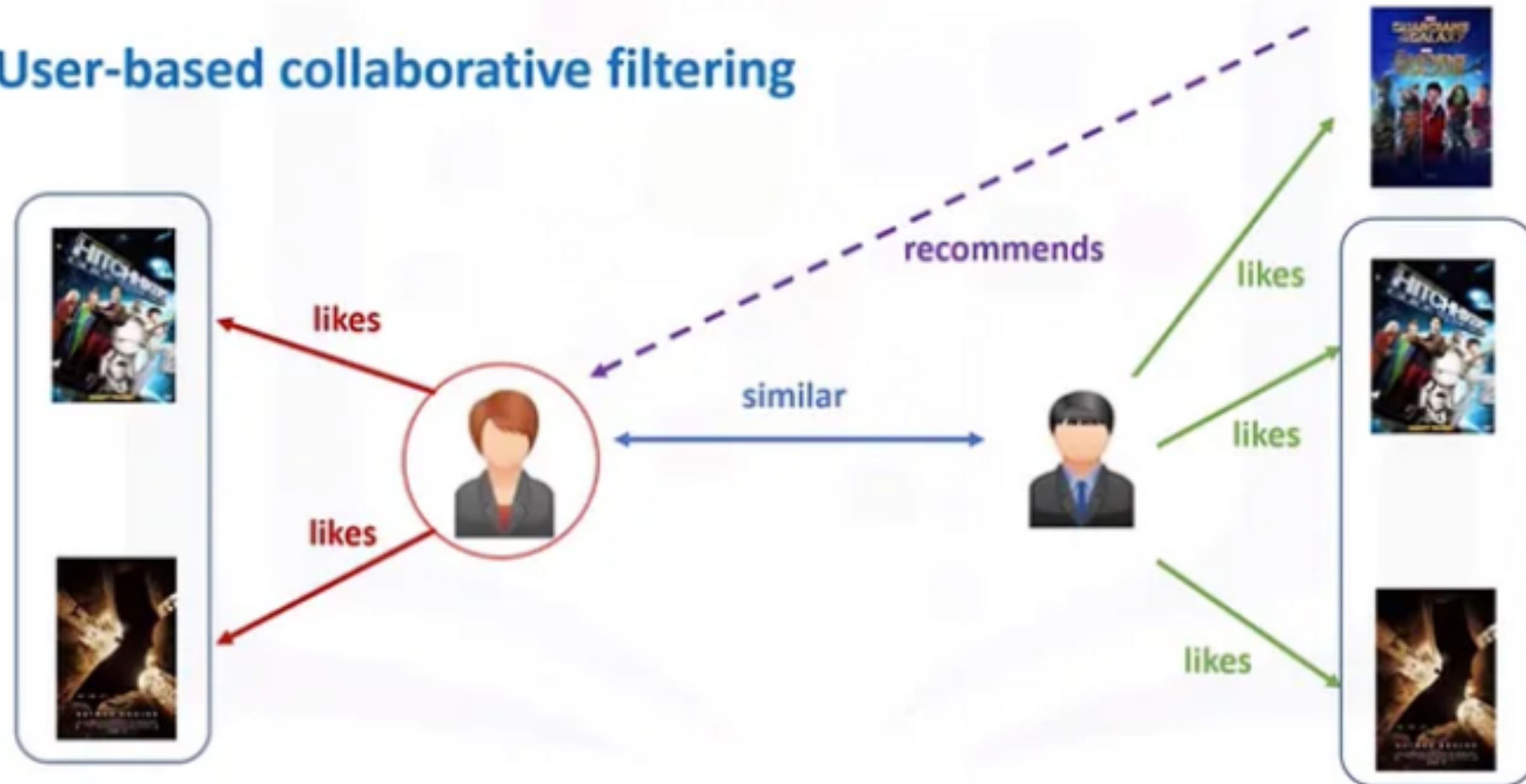
แบบที่ 1: User-Based Filtering เป็นการแนะนำโดยหา **ลูกค้าที่พฤติกรรมเหมือนกัน**

แบบที่ 2: Item-Based Filtering เป็นการแนะนำโดยหา **สินค้าที่ถูกซื้อด้วยลูกค้ากลุ่มเดียวกัน**

User-based filtering

Collaborative filtering

- User-based collaborative filtering



User-based filtering

Learning the similarity weights

					
	9	6	8	4	
	2	10	6		8
	5	9		10	7
	?	10	7	8	?

User-based filtering

Learning the similarity weights



Item-based filtering

