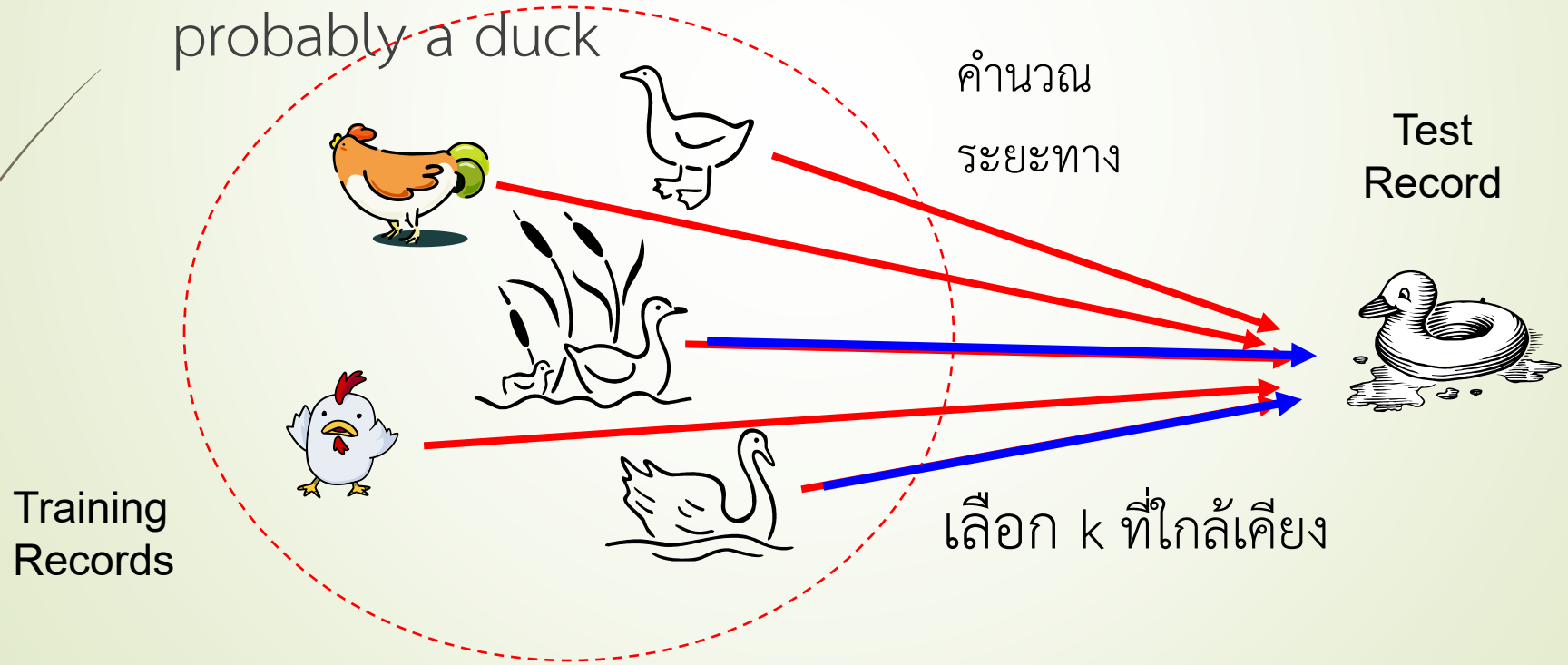
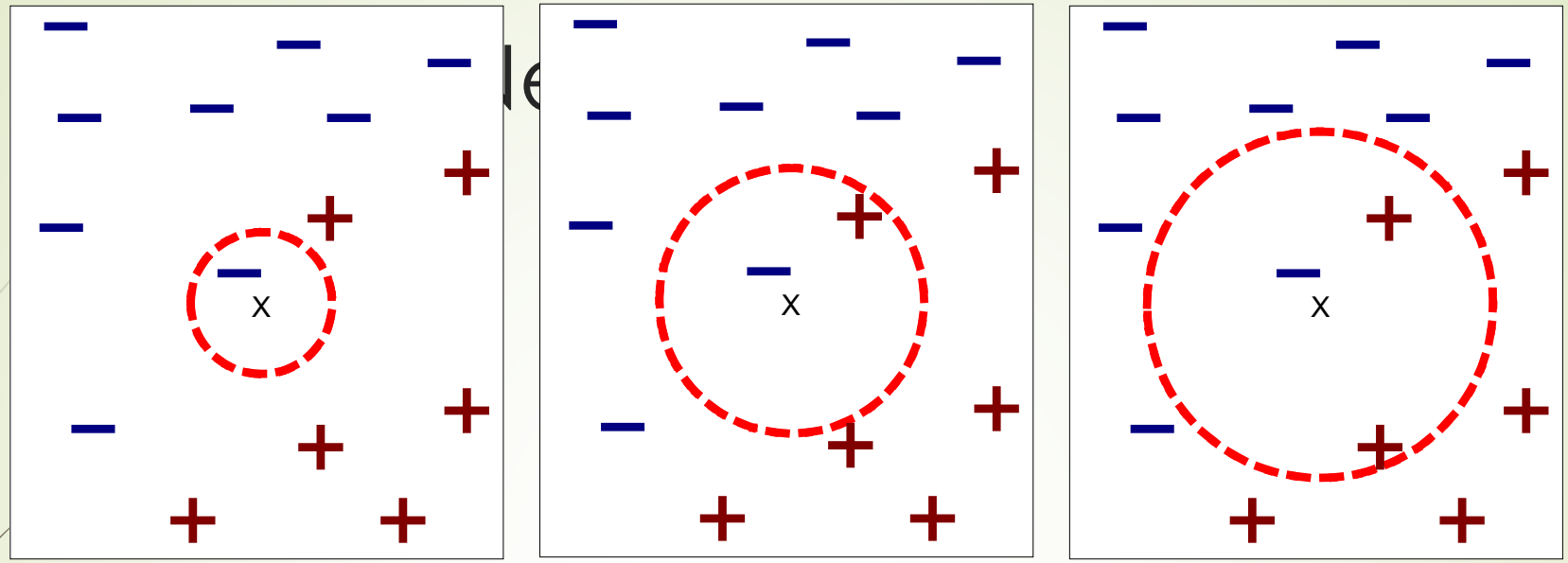
K-Nearest Neighbour Algorithm

K-Nearest Neighbor Classifiers (เพื่อนบ้านใกล้สุด k ตัว)

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck





(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

ตัวอย่าง KNN animation

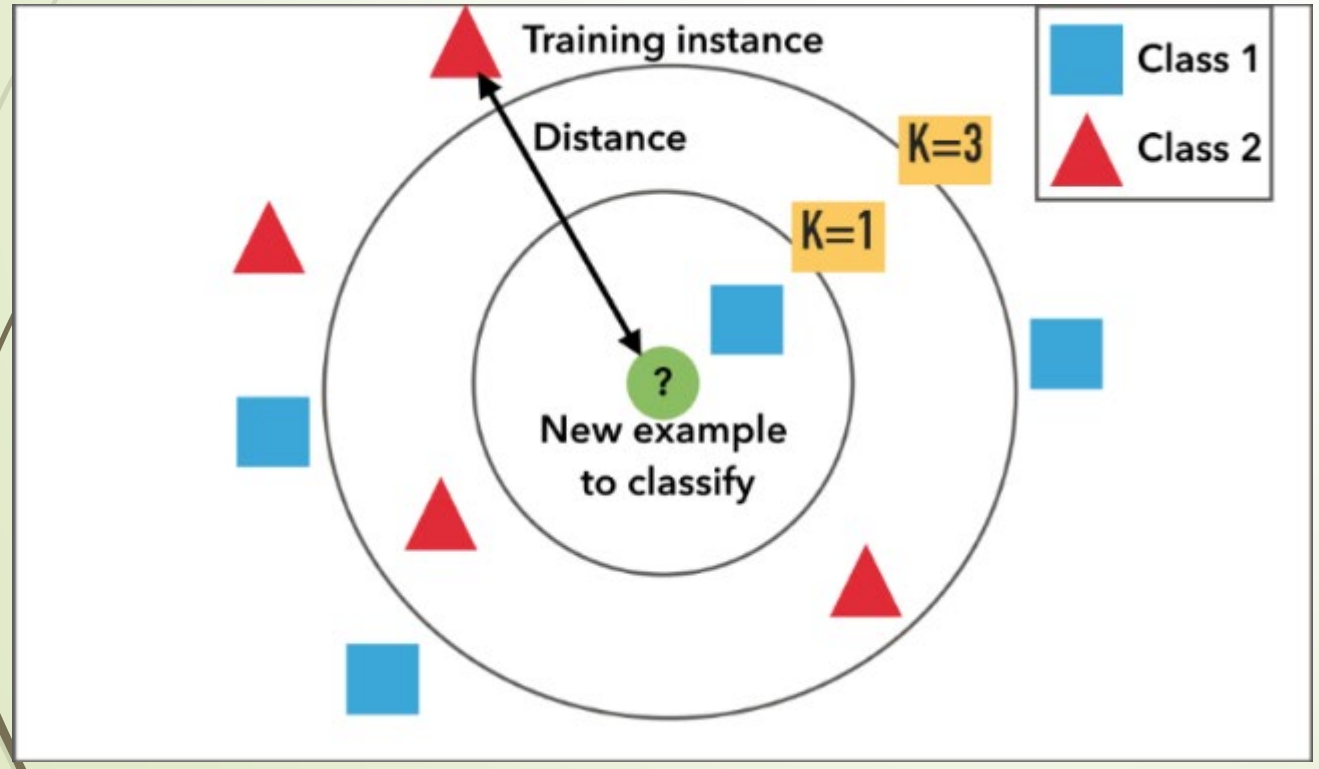
https://www.tomasbeuzen.com/animated-data/content/supervised-learning/k-nearest-neighbours.html

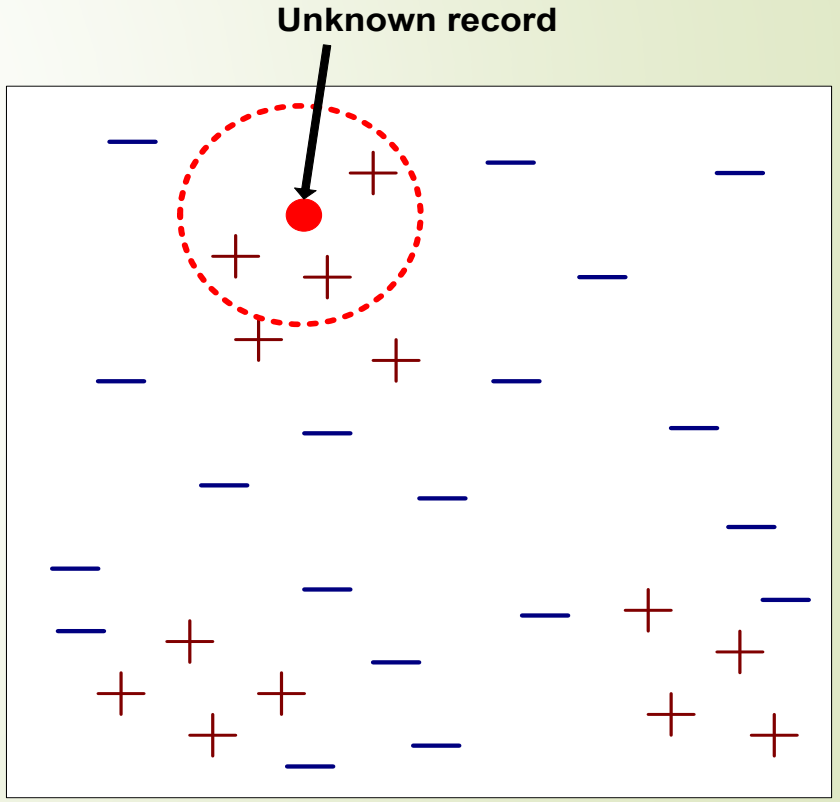
Introduction

- KNN เป็น Instance-based Learning (Memory-Based Learning) หาคำตอบของข้อมูลชุดใหม่โดยมาเปรียบเทียบกับตัวอย่างที่อยู่ใน training set
- ใช้หลักการเปรียบเทียบข้อมูลที่สนใจกับข้อมูลใน training set ว่ามีความคล้ายคลึงมากน้อยเพียงใด
- หากข้อมูลที่กำลังสนใจ อยู่ใกล้ข้อมูลใดมากที่สุด ระบบจะให้คำตอบเป็นเหมือนคำตอบของข้อมูลที่อยู่ใกล้ที่สุดนั้น
- หาผลรวม (Count Up) ของจำนวนเงื่อนไข หรือกรณีต่างๆสำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด
- เทคนิคของ K-NN ไปใช้ในนั้นเป็นการหาวิธีการวัดระยะห่างระหว่างแต่ละAttributeในข้อมูลให้ได้ และจากนั้นคำนวณค่าออกมา
- วิธีนี้จะเหมาะสำหรับข้อมูลแบบตัวเลข แต่ตัวแปรที่เป็นค่าแบบไม่ต่อเนื่องนั้นก็สามารถทำได้ เพียงแต่ต้องการการจัดการแบบพิเศษเพิ่มขึ้น อย่างเช่น ถ้าเป็นเรื่องของสี จะใช้อะไรวัดความแตกต่างระหว่างสีน้ำเงินกับสีเขียว ต่อจากนั้นต้องมีวิธีในการรวมค่าระยะห่างของ Attribute ทุกค่าที่วัดมาได้

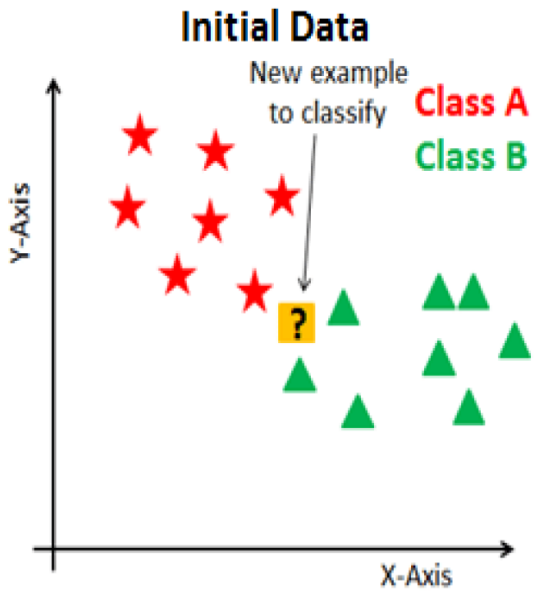
Introduction

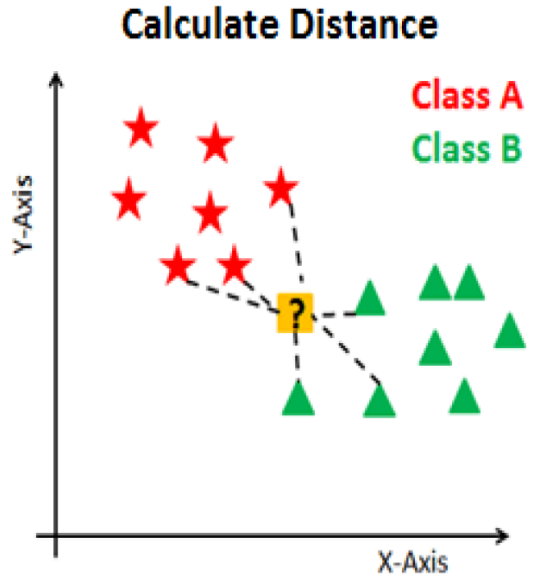
- หลังจากนั้นคำนวณระยะห่างระหว่างเงื่อนไขหรือกรณีต่างๆ ได้จากนั้นจะเลือกชุดของเงื่อนไข ที่ใช้จัดคลาสมาเป็นฐานสำหรับการจัดคลาสในเงื่อนไขใหม่ๆ
- การตัดสินได้ว่าขอบเขตของจุดข้างเคียงที่ควรเป็นนั้น ควรมีขนาดใหญ่เท่าไร และอาจตัดสินใจได้ด้วยว่าจะนับจำนวนจุดข้างเคียงตัวมันได้อย่างไร

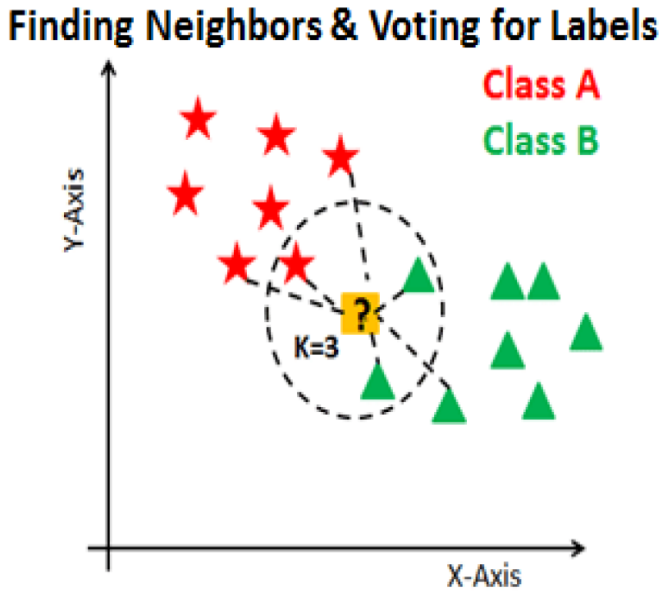




How k-NN works?

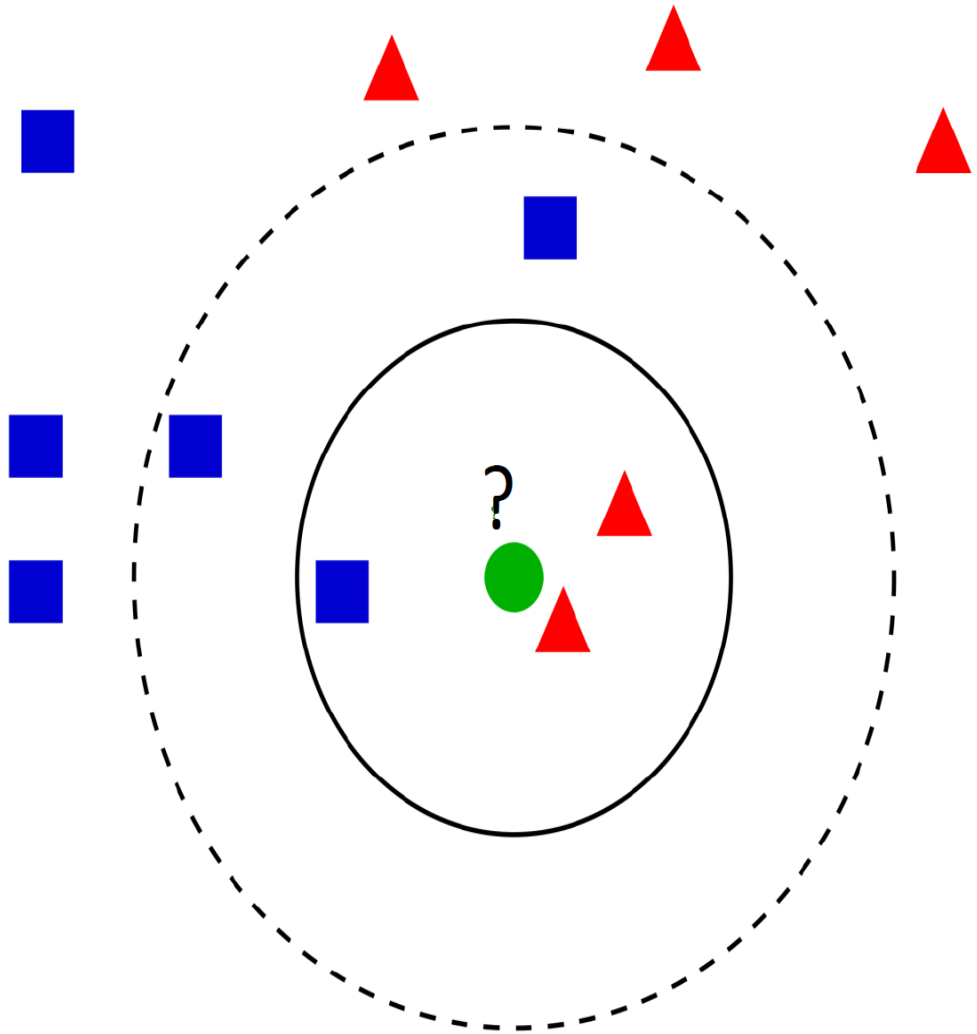






https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

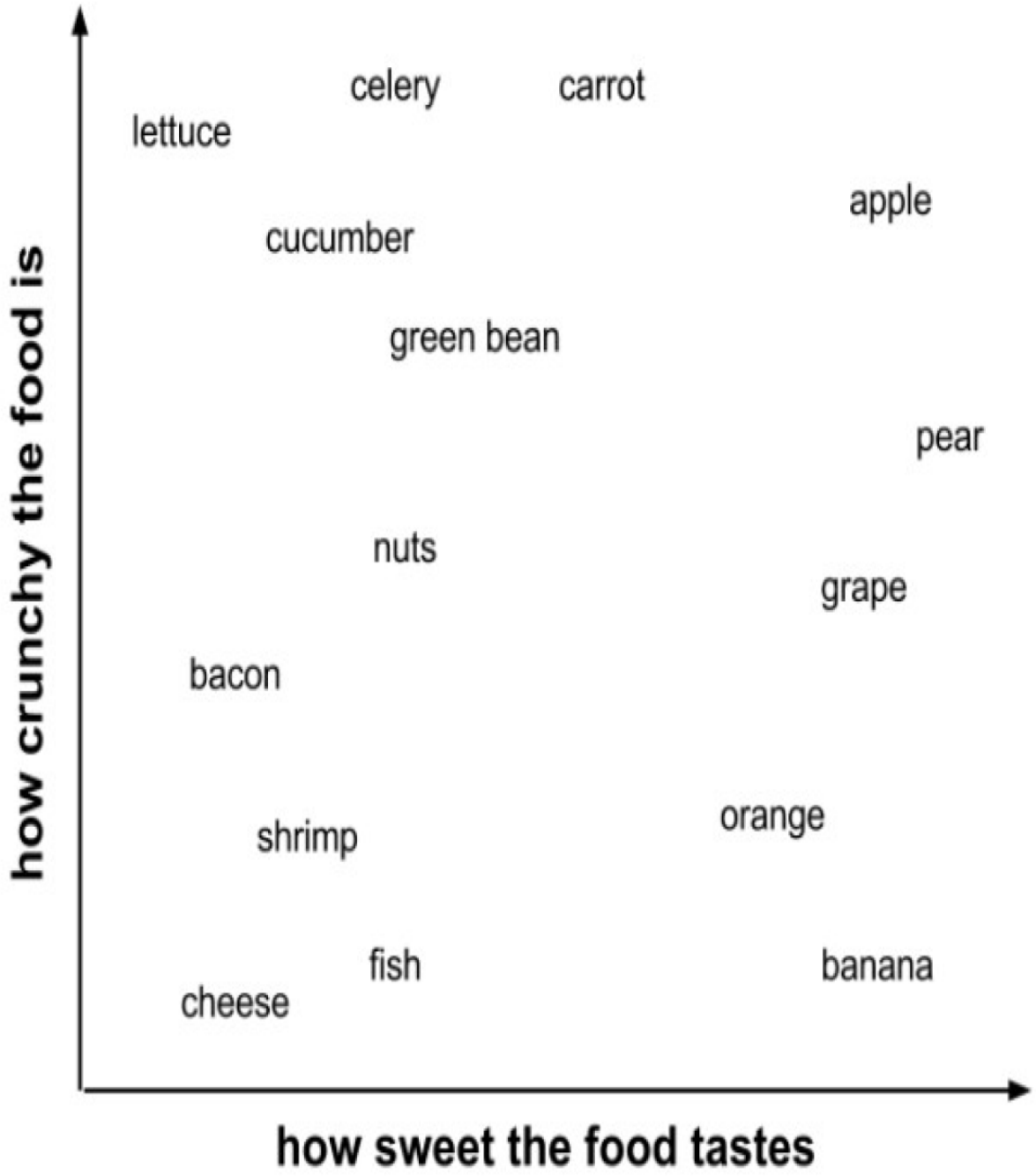
k-NN example (3-NN vs 5-NN)

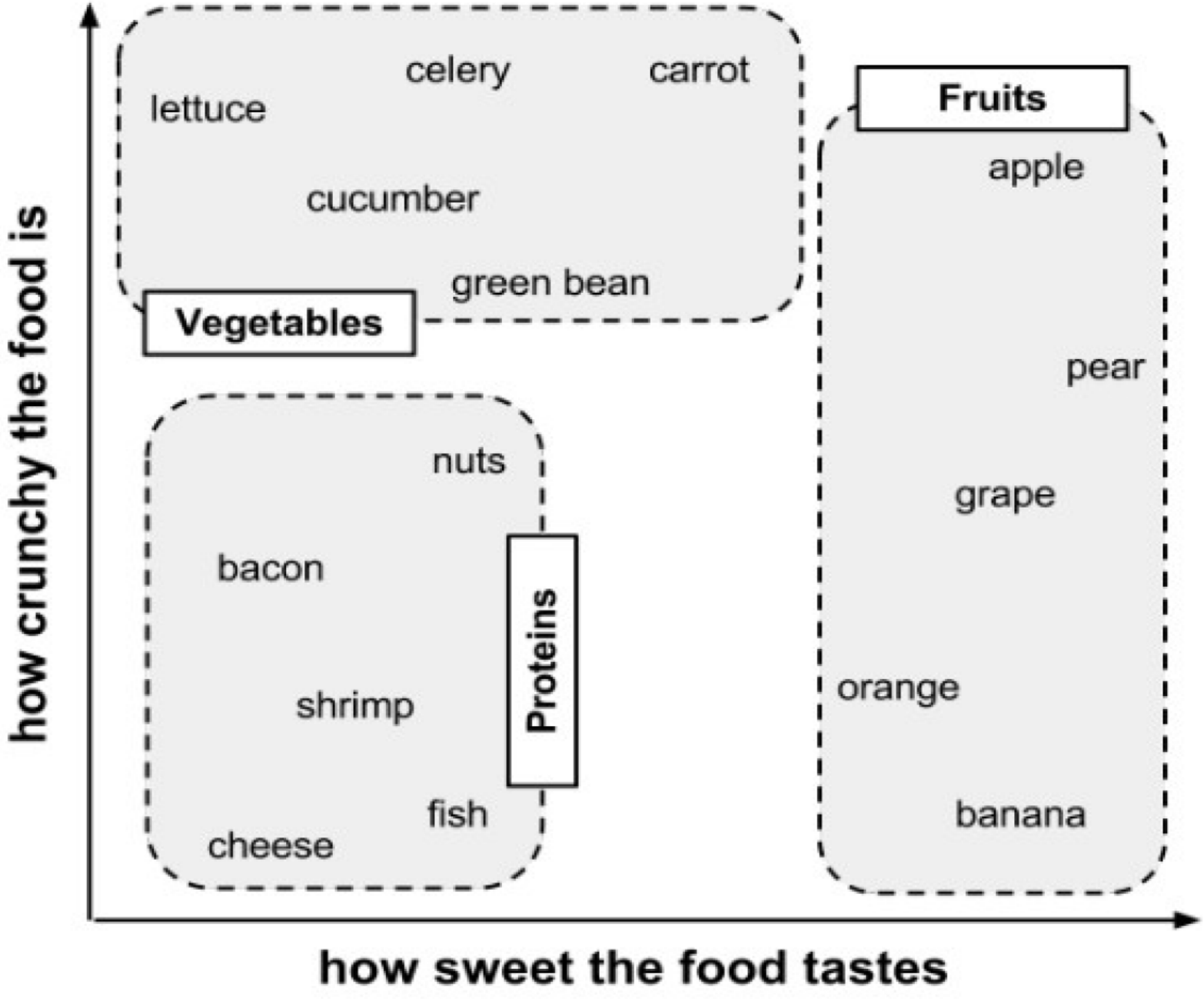


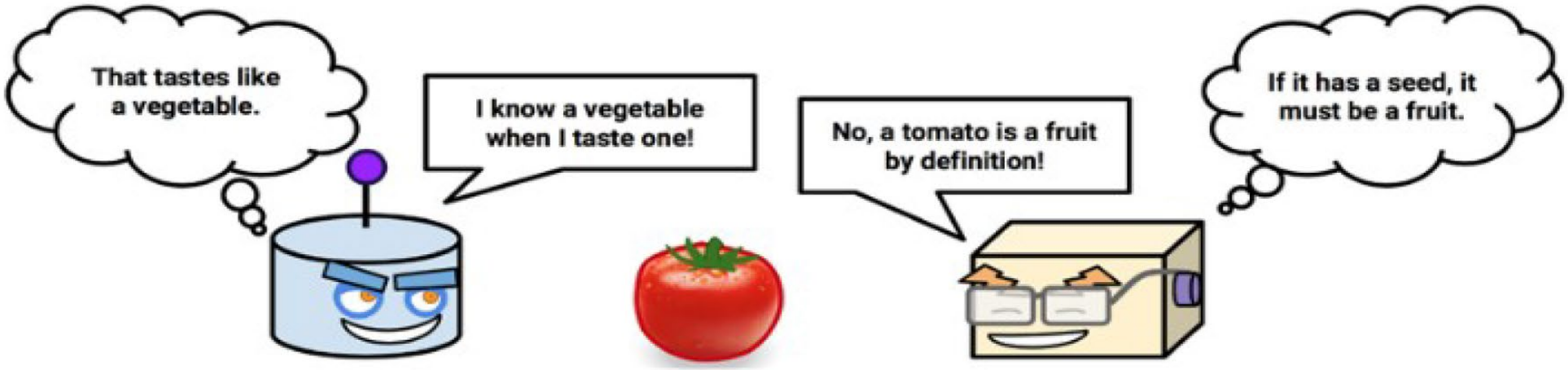
https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

ingredient	sweetness	crunchiness	food type
apple	10	9	fruit
bacon	1	4	protein
banana	10	1	fruit
carrot	7	10	vegetable
celery	3	10	vegetable
cheese	1	1	protein

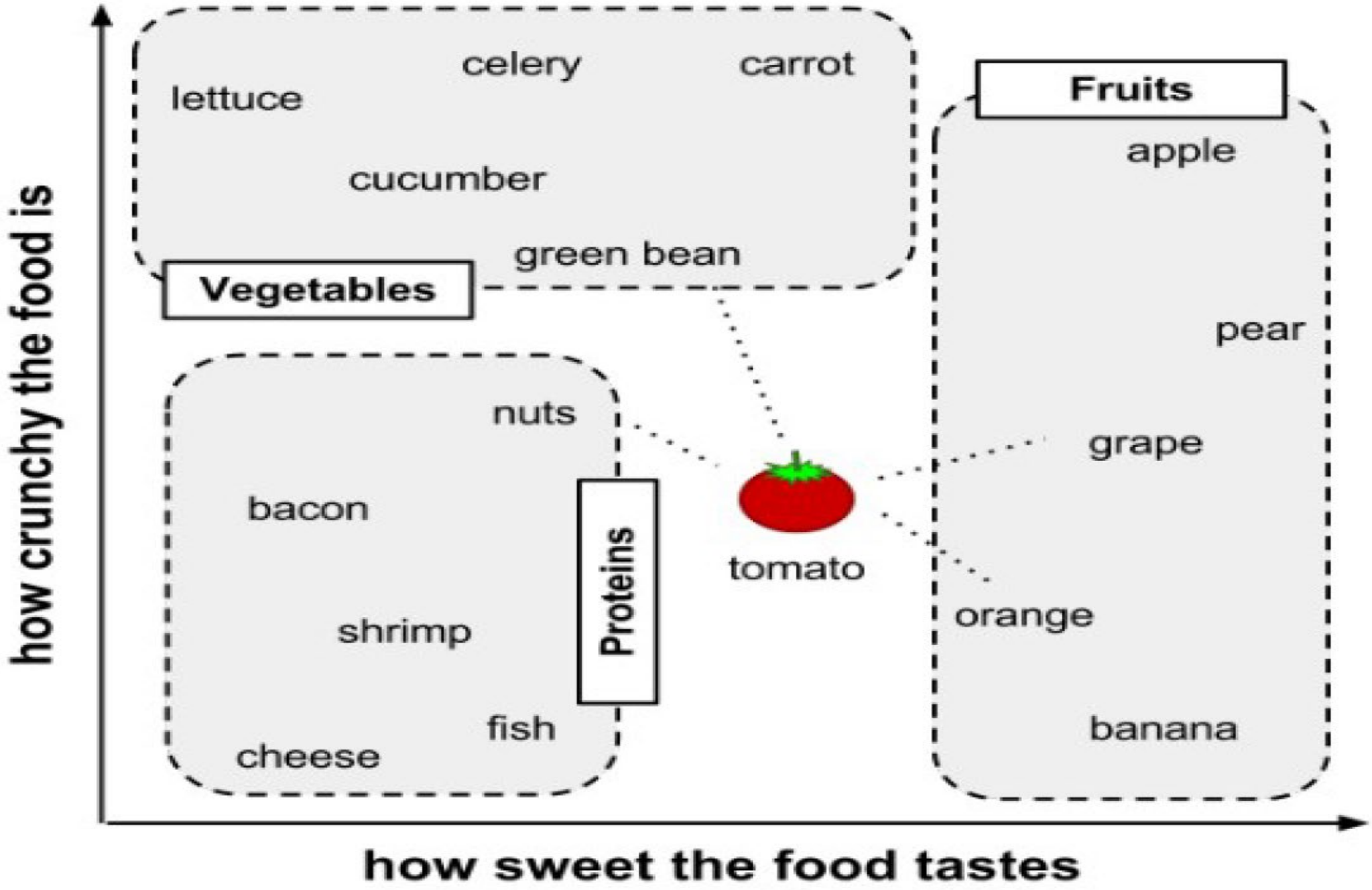
Reference: Machine Learning with R, Brett Lantz, Packt Publishing


how crunchy the food is
how sweet the food tastes
celery
carrot
lettuce
apple
cucumber
green bean
pear
nuts
grape
bacon
shrimp
orange
fish
banana
cheese

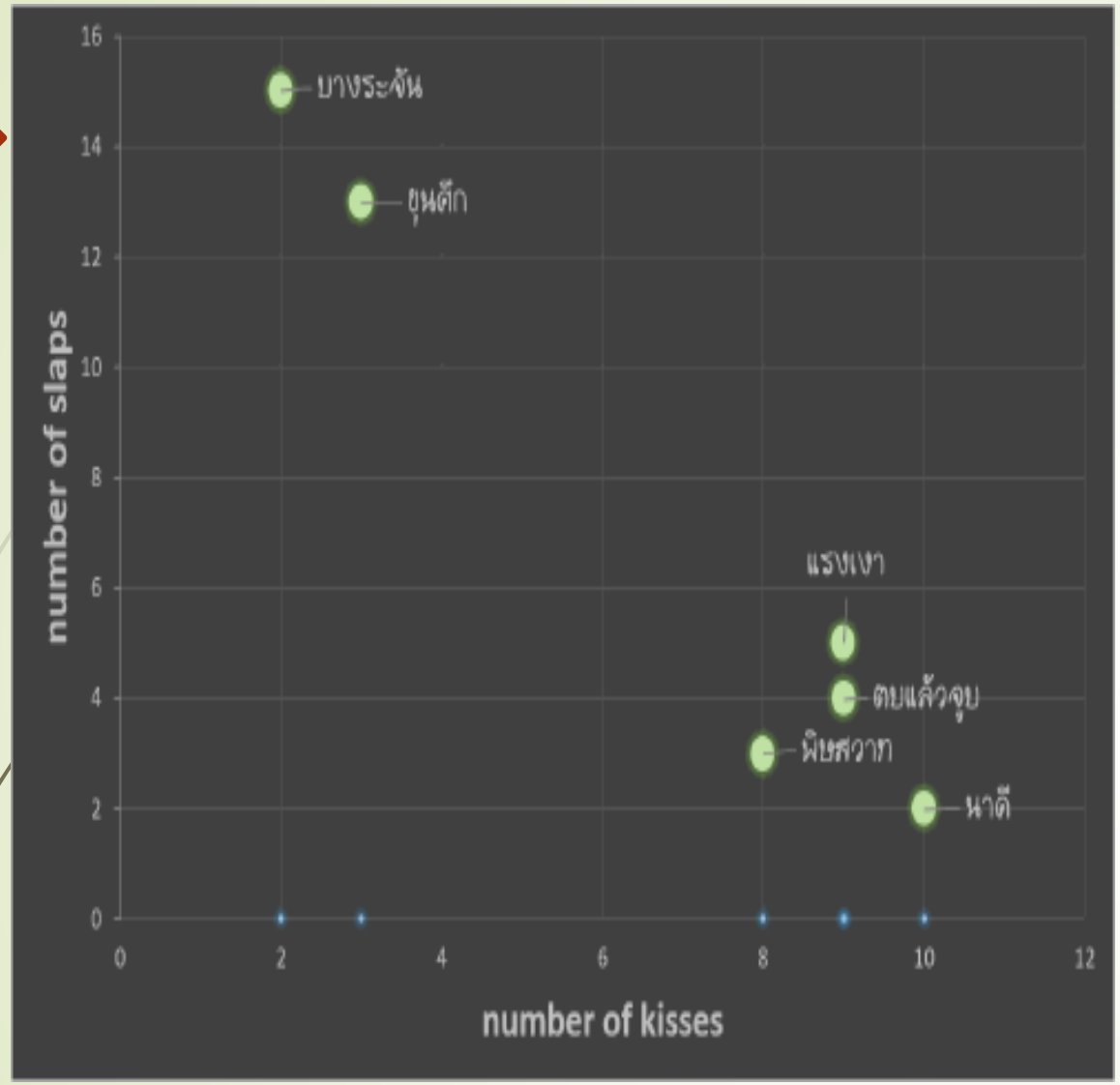

how crunchy the food is
how sweet the food tastes
celery
carrot
lettuce
cucumber
green bean
Vegetables
Fruits
apple
pear
grape
orange
banana
nuts
bacon
shrimp
fish
cheese
Proteins

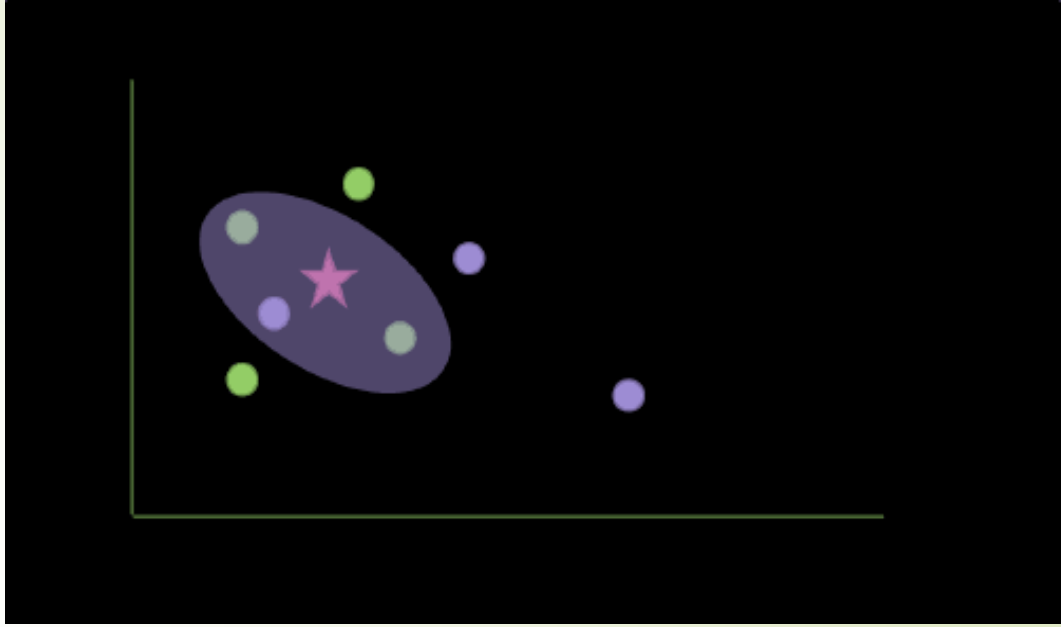


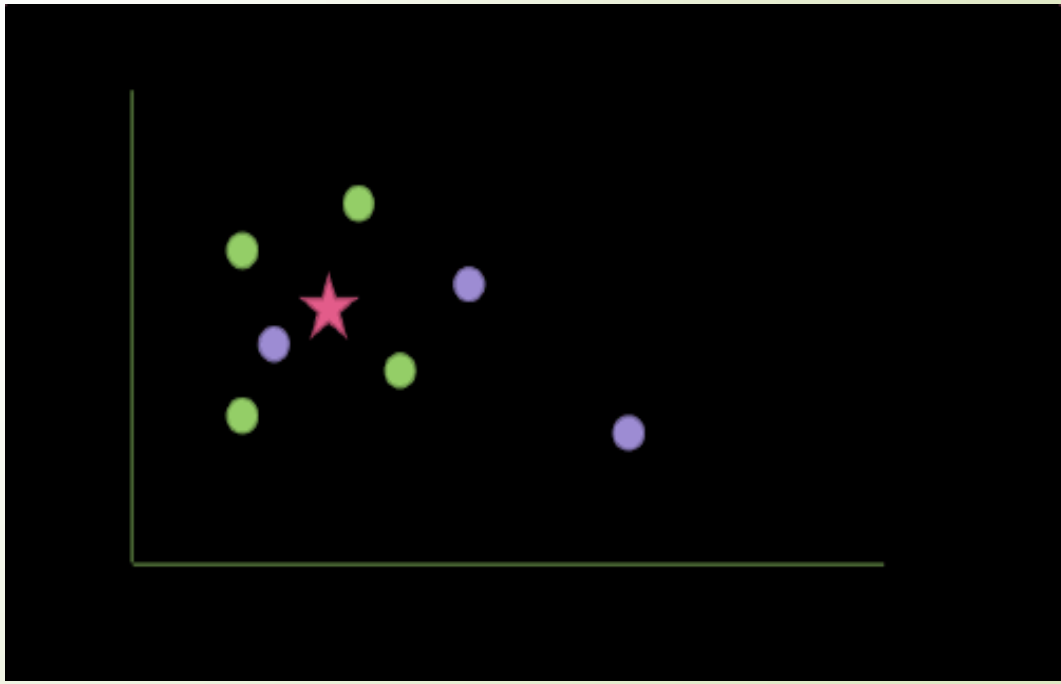
มะเขือเทศ sweet = 6, crunchiness = 4


how crunchy the food is
how sweet the food tastes
Vegetables
lettuce
celery
carrot
cucumber
green bean
Fruits
apple
pear
grape
orange
banana
Proteins
nuts
bacon
shrimp
fish
cheese
tomato

เรื่อง	จำนวนฉากจูบ	จำนวนฉากตบ	ประเภท
นาคี	10	2	โรแมนติก
พิษสวาท	8	3	โรแมนติก
แรงเงา	9	5	โรแมนติก
บางระจัน	2	15	แอคชัน
ขุนศึก	3	13	แอคชัน
ตบแล้วจูบ	9	4	?







ภาพนี้นำมาจาก https://www.babelcoder.com/blog/posts/k-nearest-neighbors

ฟังก์ชั่นการดำเนินการในอัลกอริทึม k-NN

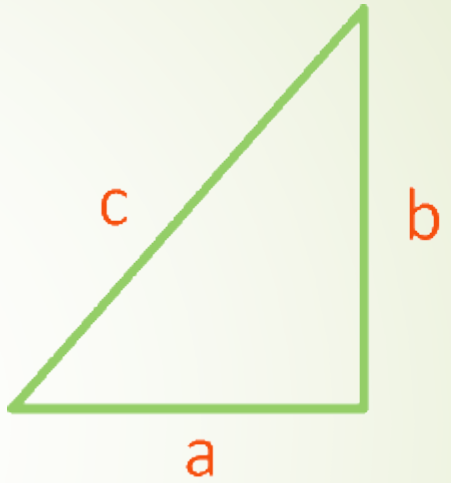
การดำเนินการของอัลกอริทึมแบบ k-NN ประกอบไปการทำงานของ 2 ฟังก์ชั่น

- ฟังก์ชั่นระยะทาง (Distance Function)
  - เป็นการคำนวณค่าระยะห่างระหว่างสองเรคคอร์ดข้อมูล เพื่อที่จะมาวัดความคล้ายคลึงกันของข้อมูล โดยมีเงื่อนไขคือ
  - ค่าระยะทาง(ความห่าง)ที่คำนวณได้ต้องไม่ติดลบ
  - ถ้าตำแหน่งเดียวกันฟังก์ชันต้องเป็นศูนย์(ค่าเหมือนกัน)
  - การคำนวณวัดระยะทางไปกลับต้องเท่ากัน
- การดำเนินการการหาระยะทางระหว่าง จุด A และ B ใด ๆ ทำได้โดย
  - ใส่ค่าสัมบูรณ์ (Absolute) ให้กับค่าระยะทาง: $|A-B|$
  - ยกกำลังสองให้กับค่าระยะทาง : $(A-B)^2$

Nearest Neighbor Classification

- Compute distance between two points:
  - Euclidean distance

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$



$c^2 = a^2 + b^2$

  - Manhattan

Manhattan $\sum_{i=1}^{k} |x_i - y_i|$

ตัวอย่างที่ 1

No.	Gender	Age	Salary	inactive
1	F	27	19,000	No
2	M	51	64,000	Yes
3	M	52	105,000	Yes
4	F	33	55,000	Yes
5	M	45	45,000	No
new	F	45	100,000	???

ตัวอย่างที่ 1

	Identical Gender	|A-B| Age	|A-B| Salary	Distance dsum
d(1,n)	0	18	81,000	81018
d(2,n)	1	6	36,000	36007
d(3,n)	1	7	5,000	5008
d(4,n)	0	12	45,000	45012
d(5,n)	1	0	55,000	55001

>>> เรียงdsumจากน้อยไปหามากจะได้เป็น 3 2 4 5 1 – Y,Y,Y,N,N

ตัวอย่างที่ 2

Discrete values

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

x = <humidity, temperature>

New instance x_q = <40, 30, run=?? > We can run inside(+) or outside (-)

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

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

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

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

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

1-NN (x1)
Answer run inside(+)

2-NN (x1,x4)
Answer run inside(+)

3-NN (x1,x2,x4)
Answer run inside (+)

4-NN (x1,x2,x4,x5)
Answer run inside (+) or (-)

5-NN
Answer run inside(-)

ตัวอย่างที่ 3

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 = <humidity, temperature>
New instance x_q = <40, 30, Rainfall =?? >

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

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

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

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

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

1-NN (x1)
Rainfall = 5.1

2-NN (x1,x4)
Rainfall = (5.1+3.2)/2 = 4.15

3-NN (x1,x2,x4)
Rainfall = (5.1+15.5+3.2)/3
= 7.9

4-NN (x1,x2,x4,x5)
Rainfall = (5.1+15.5+3.2+12.0)/4 = 8.95

5-NN (x1,x2,x3, x4,x5)
Rainfall = (5.1+15.5+3.2+20.2+12.0)/5 =
11.2

ข้อดี-ข้อเสียของ KNN

- ข้อดี
  - หากเงื่อนไขการตัดสินใจมีความซับซ้อนวิธีนี้สามารถสร้างโมเดลที่มีประสิทธิภาพได้
  - ทำนายข้อมูลใหม่โดยอาศัยการเปรียบเทียบกับข้อมูลเรียนรู้จำนวน K ตัวที่อยู่ใกล้ที่สุด
  - พิจารณาคำตอบจาก
    - คำตอบของข้อมูลเรียนรู้ที่อยู่ใกล้ที่สุด K ตัวที่พบมากที่สุดเป็น คำตอบ หรือ
    - ให้ค่าน้ำหนักโดยการพิจารณาระยะห่างระหว่างข้อมูลที่สนใจกับข้อมูลที่อยู่ใกล้สุด K ตัวร่วมด้วย
- ข้อเสีย
  - ใช้ระยะเวลาในการคำนวณนาน
  - ถ้าAttributeมีจำนวนมากจะเกิดปัญหาในการคำนวณค่า
  - ทำนายได้เฉพาะข้อมูลที่เป็นแบบประเภท (nominal) เท่านั้น เช่น หญิง หรือ ชาย ฯลฯ