

IN[34]120 - Søketeknologi



2024-10-22 🎃 🦅

Tema: Klassifisering

- Poll om tidligere emner og veien videre
- Generelt om klassifisering
- Eksempler på klassifisering
- Oblighjelp



IN4120: Science fair deadlines

- ✓ Group self-assignment: 2024-10-21
- 📌 Topic selection: 2024-11-04
- Contact prof. Øhrn about this 🦅

ti. 22. okt.	14:15–16:00	Text classification.	<u>OJD</u> , <u>Datastue</u> <u>Chill</u>	<u>O. R. Jahren</u>	Hva er klassifisering? Bred introduksjon. Oblighjelp etter.
ti. 29. okt.	14:15–16:00	Naïve bayes.	<u>OJD</u> , <u>Datastue</u> <u>Chill</u>	<u>O. R. Jahren</u>	Repetisjon av Naïve bayes mtp å løse oblig E-1. Muligens gjennomgå LF for oblig C-1.
ti. 5. nov.	14:15–16:00	Gruppe 2	<u>OJD</u> , <u>Datastue</u> <u>Chill</u>	<u>O. R. Jahren</u>	Siste gruppetime før deadline for oblig E.
ti. 12. nov.	14:15–16:00	Gruppe 2	<u>OJD</u> , <u>Datastue</u> <u>Chill</u>	<u>O. R. Jahren</u>	Gjennomgå LF for D-1?
ti. 19. nov.	14:15–16:00	Eksamensforberedelser.	<u>OJD</u> , <u>Datastue</u> <u>Chill</u>	<u>O. R. Jahren</u>	Regner med vi løser <u>gamle eksamensoppgaver</u> i fellesskap.

Eksamen er 2024-11-29

Veien frem mot eksamen - hva vil dere?

Forslag til temaer

- Diskutere et paper
- Gå tilbake til temaer vi skippet
- Chatbot-workshop
- Gjennomgå prekode obligene ikke bruker

Minst gira til venstre, mest til høyre



Forslag til temaer for tiden frem mot eksamen (ranger hvor gira du er)

Diskutere et paper

5.1

Gå tilbake til temaer vi skippet

8.4

Chatbot-workshop

3.4

Gjennomgå prekode obligene ikke bruker

2.8

Lite gira

Veldig gira



10

Andre innspill til tema?




7



(Før jeg har sett svarene)

Ting vi trolig kommer til å dekke

- Naïve bayes 
- Bloom filters ✓
- PageRank ✓

[in3120-2024](#) / [seminars](#) / [gruppe2](#) / [temaer.md](#) 

 orjahren Fler temaer.

Preview

Code

Blame

12 lines (8 loc) · 281 Bytes

Temaer

Temaer timene til gruppe 2 ikke har dekket særlig som det trolig er lurt å ta en titt på før eksamen

Oppdateres fortløpende. Se tidspunkt for siste oppdatering over^

- MapReduce
- Compression
- ANN (approximate nearest neighbor)
- SVMer
- LTR (Learning to rank)

NB! Vi har fokusert mye på obligene.

Demystify TF-IDF in Indexing and Ranking

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

TF-IDF

Term x within document y

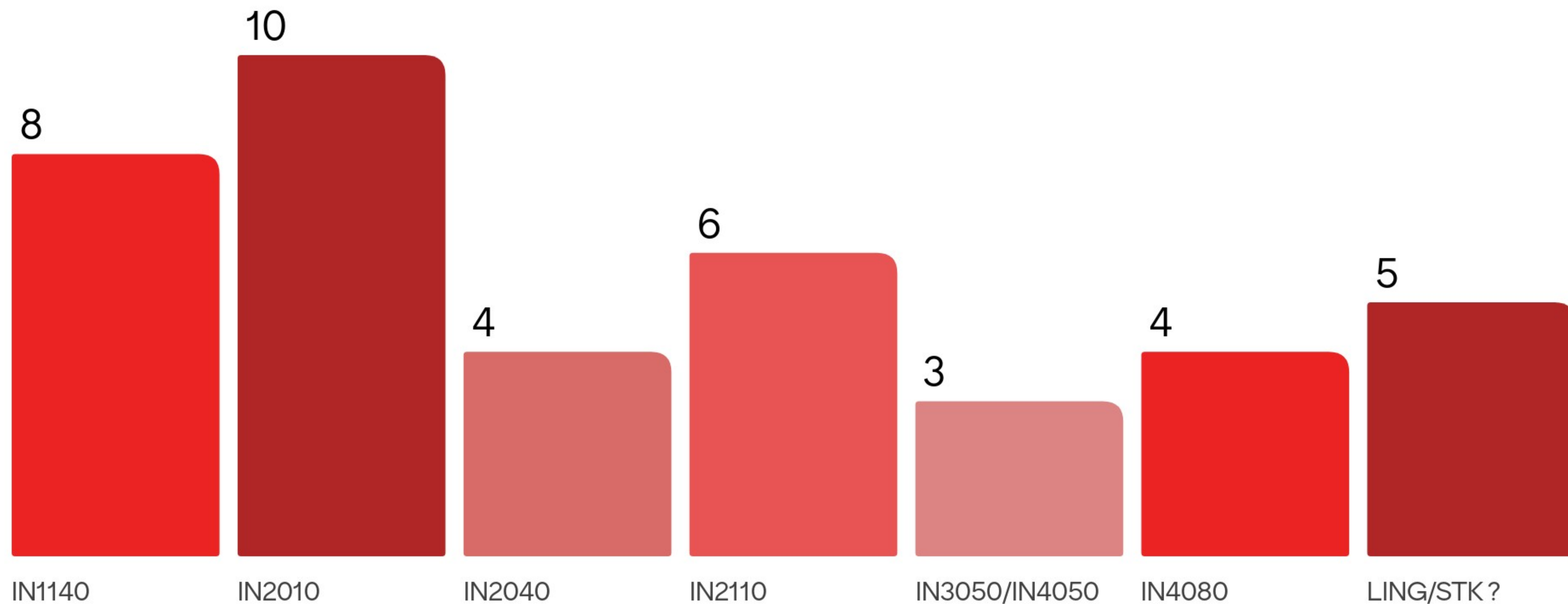
$tf_{x,y}$ = frequency of x in y

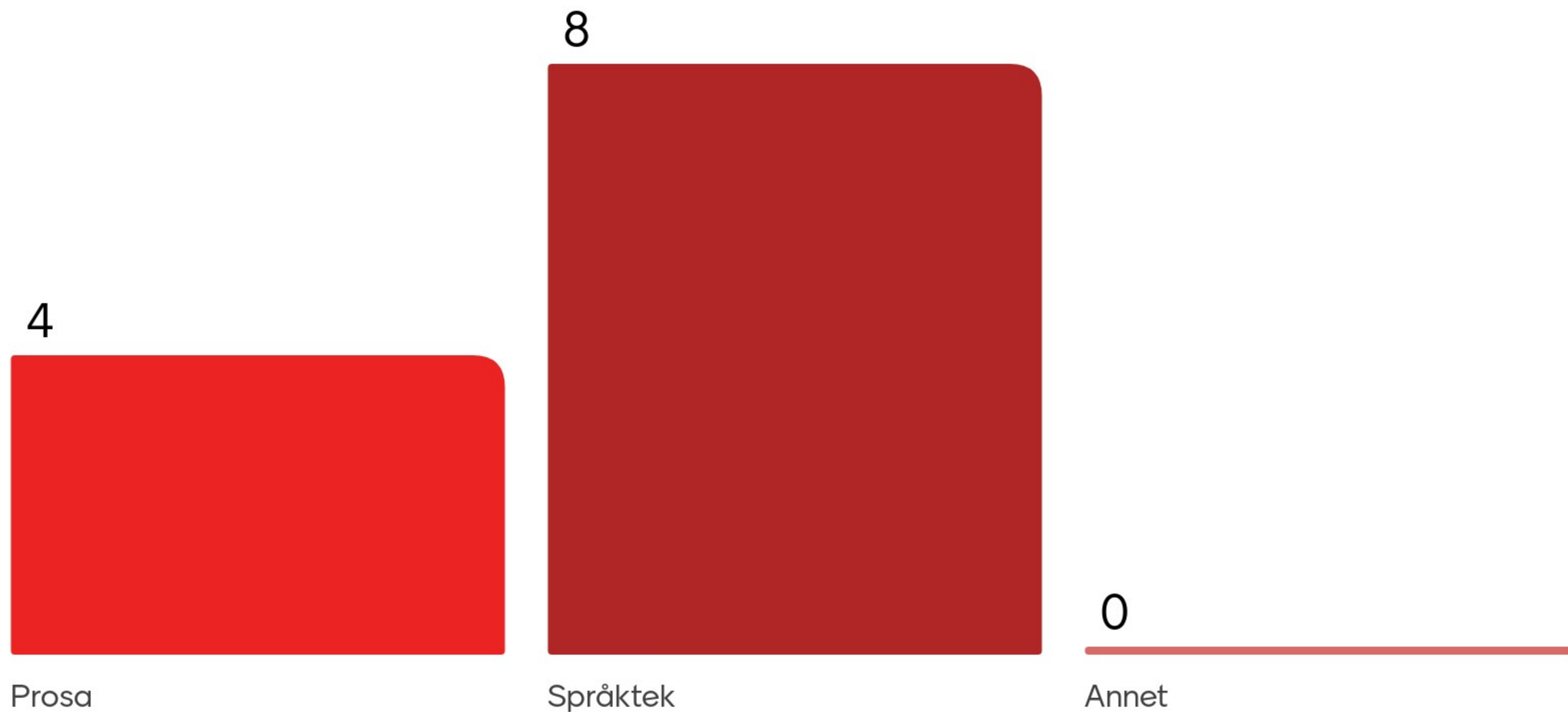
df_x = number of documents containing x

N = total number of documents

Husk: Oblig D frist på fredag

Hvilke emner har du tatt tidligere?





Klassifisering

Hva??

Hva er klassifisering?



Kategorisere

automatisk tilordning av
klasser

putte noe i en eller flere
kategorier

klassifisere docs i ulike
klasser

Automatisk tildeling av
klasse til elementer

Tillegne klasse til et nytt
datapunkt

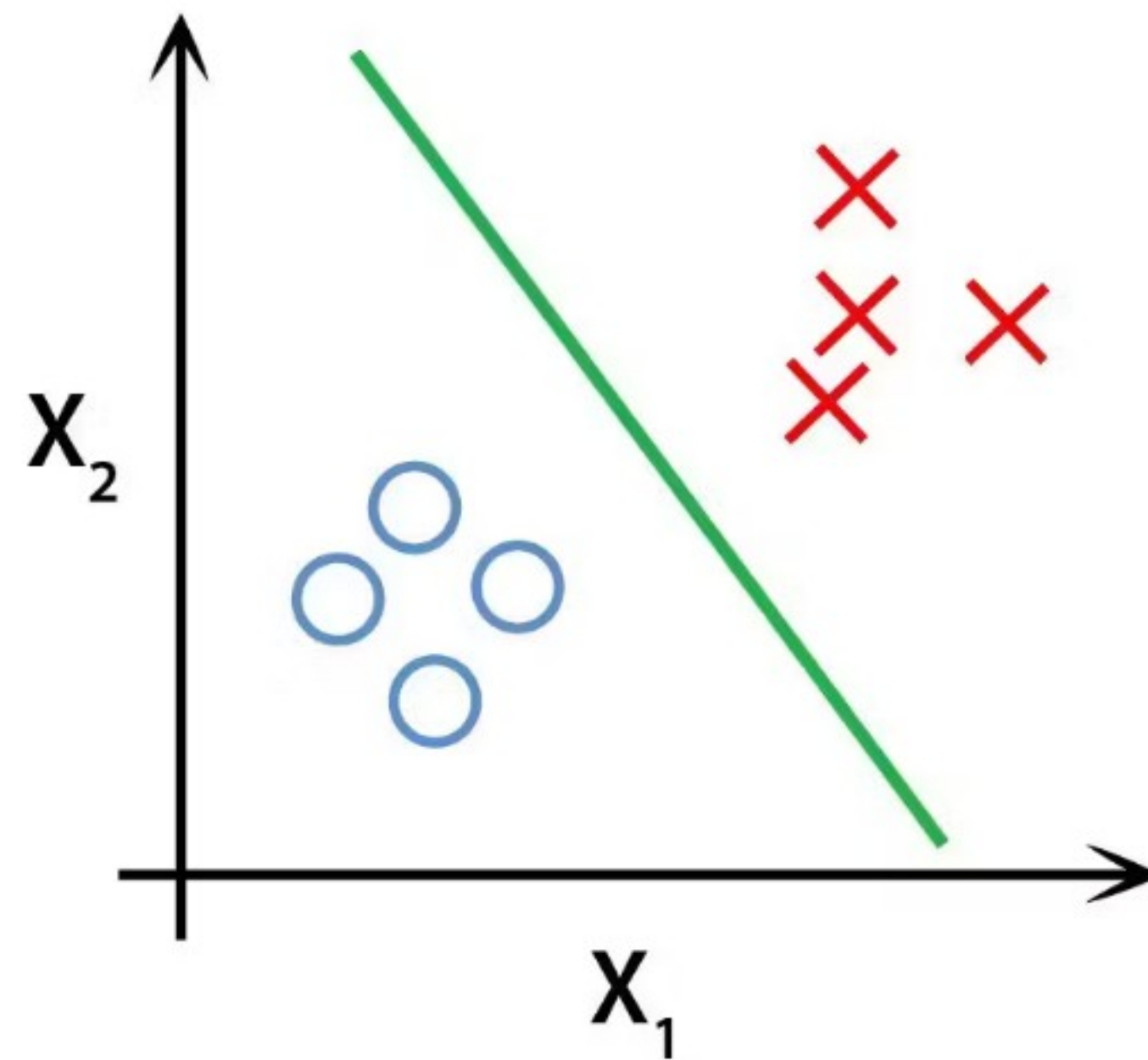
Kategorisering. Tilordne
element til riktig klasE

Fordele data til ulike
kategorier

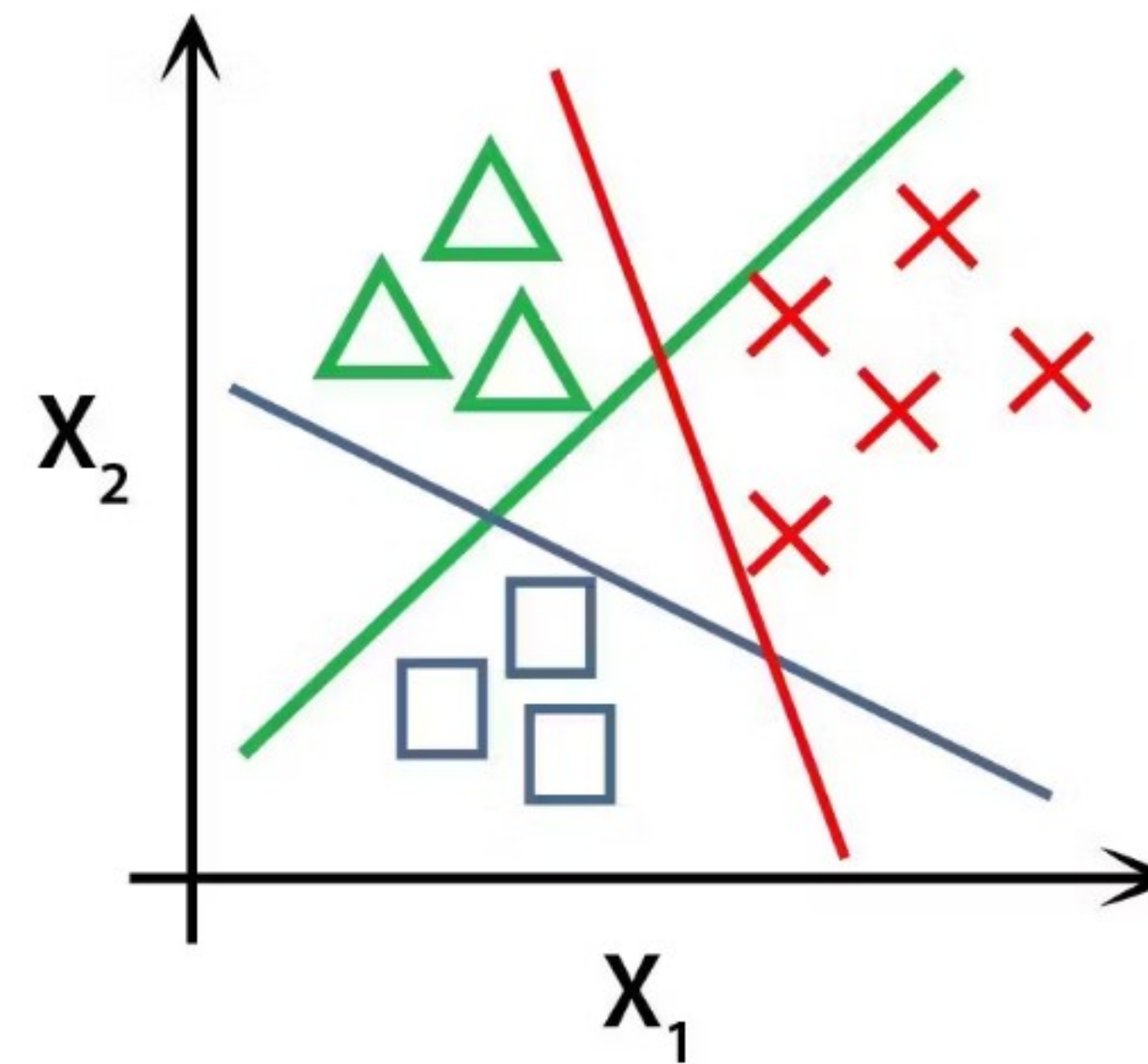
Klassifisering 101

- Gitt en ting, gi den en kategori
- "Kategori", "klasse", "*label*"

BINARY CLASSIFICATION



MULTI-CLASS CLASSIFICATION



Visuell klassifisering

Kategorier i klassifisering

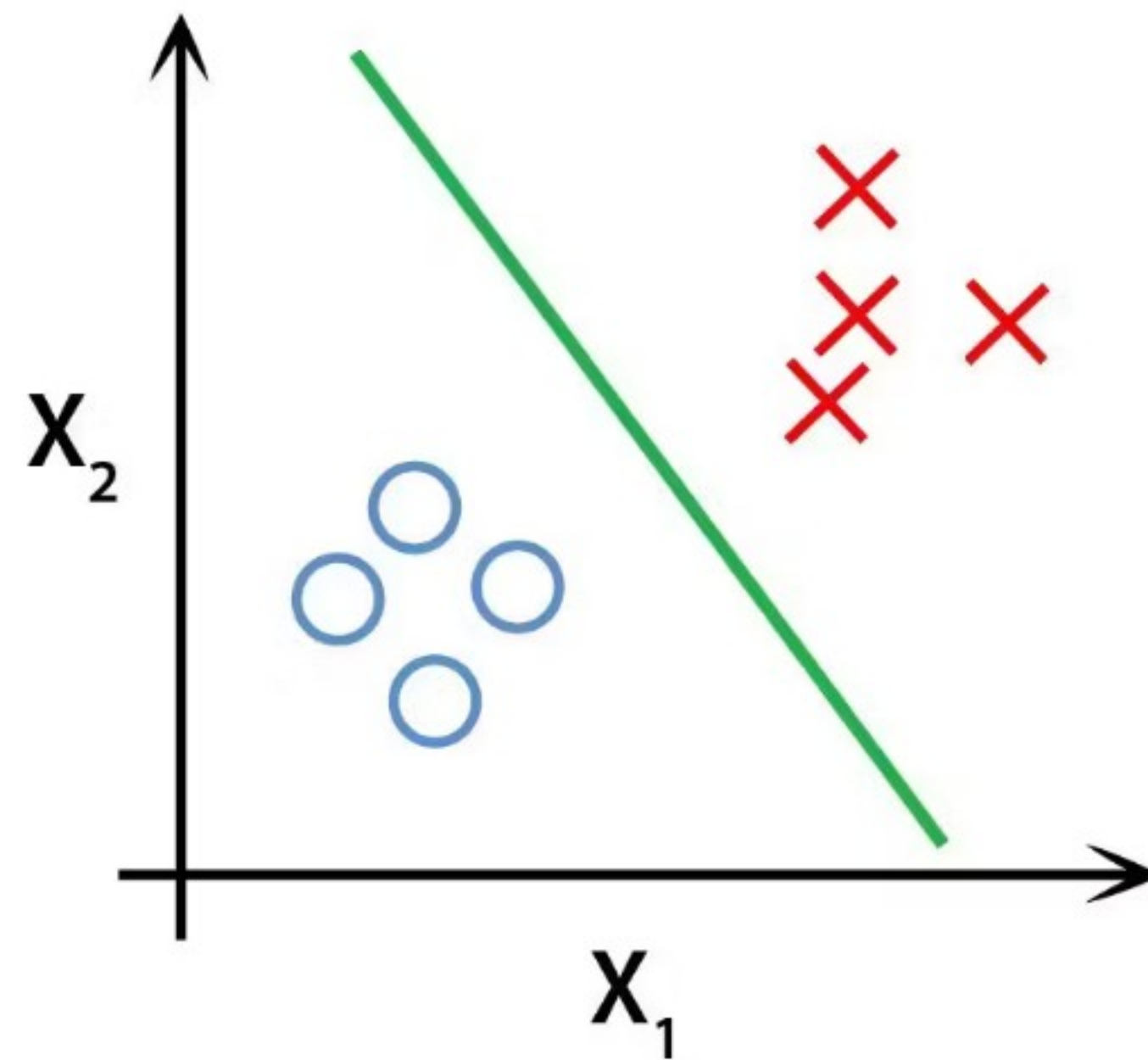
- Tema for innhold
- Språk (oblig E-1)
- *Noe helt annet*

Supervised/unsupervised learning

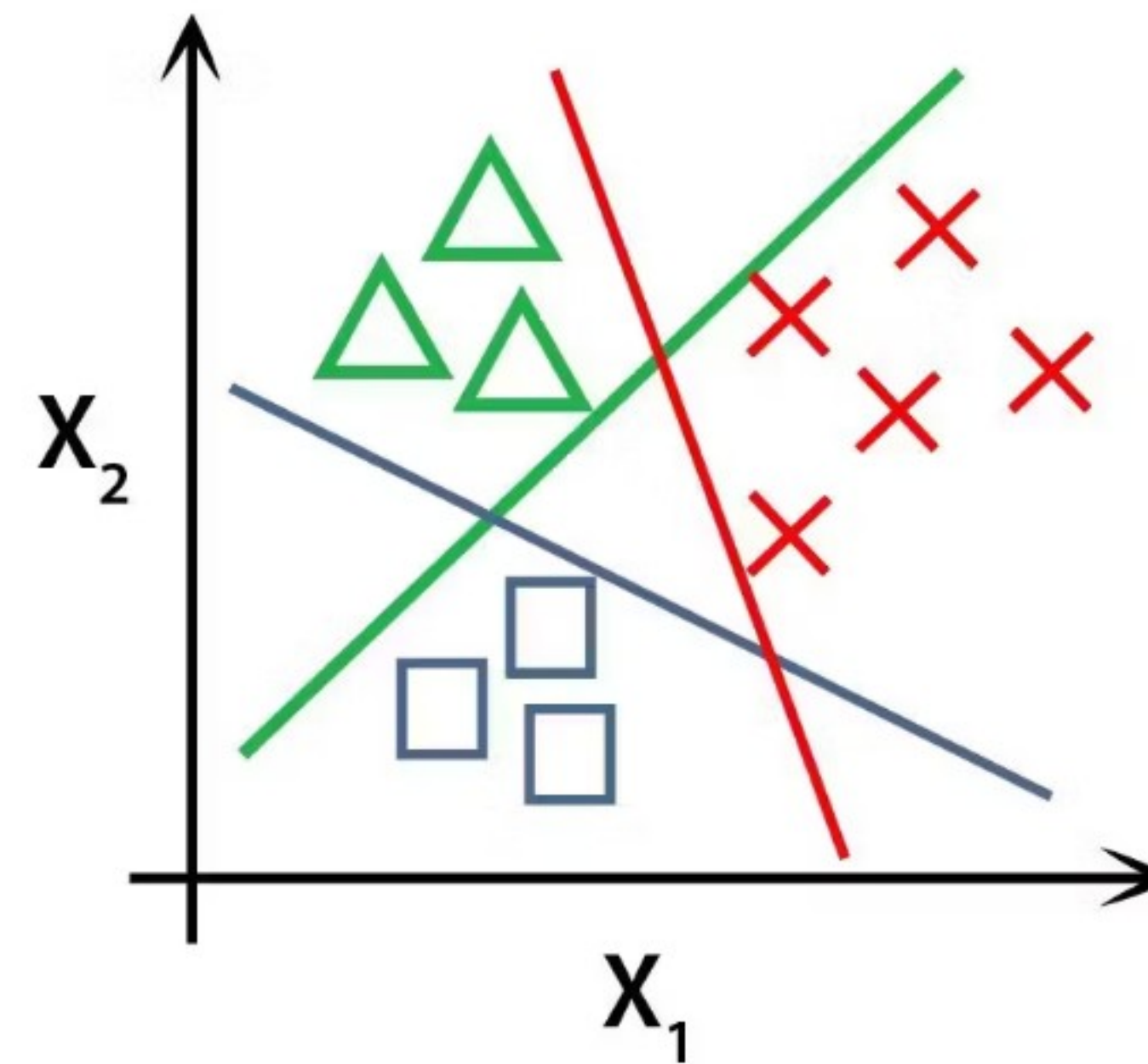
Supervised learning

- Du vet på forhånd hvilke kategorier
- Lær å kjenne igjen kategoriene dine

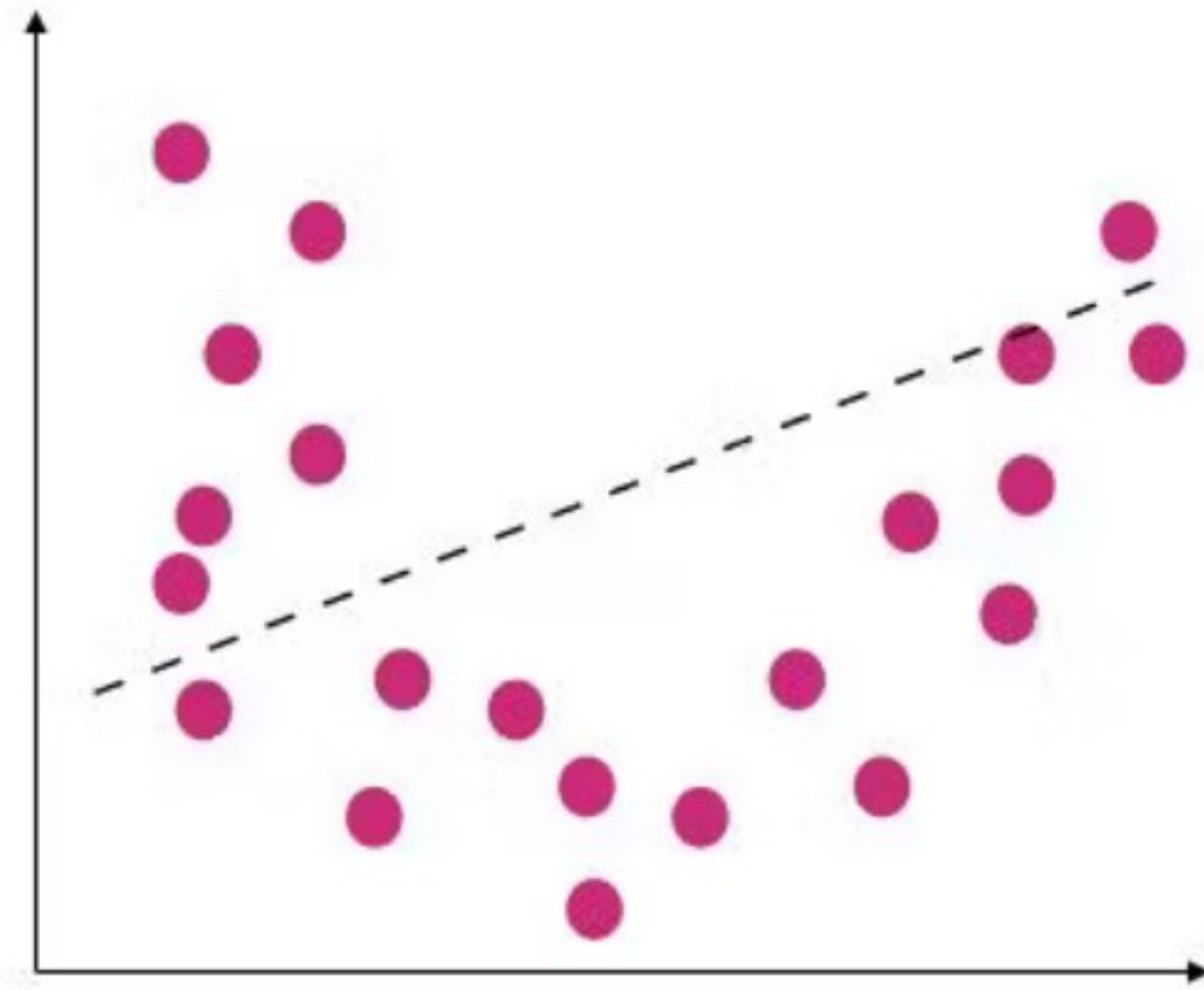
BINARY CLASSIFICATION



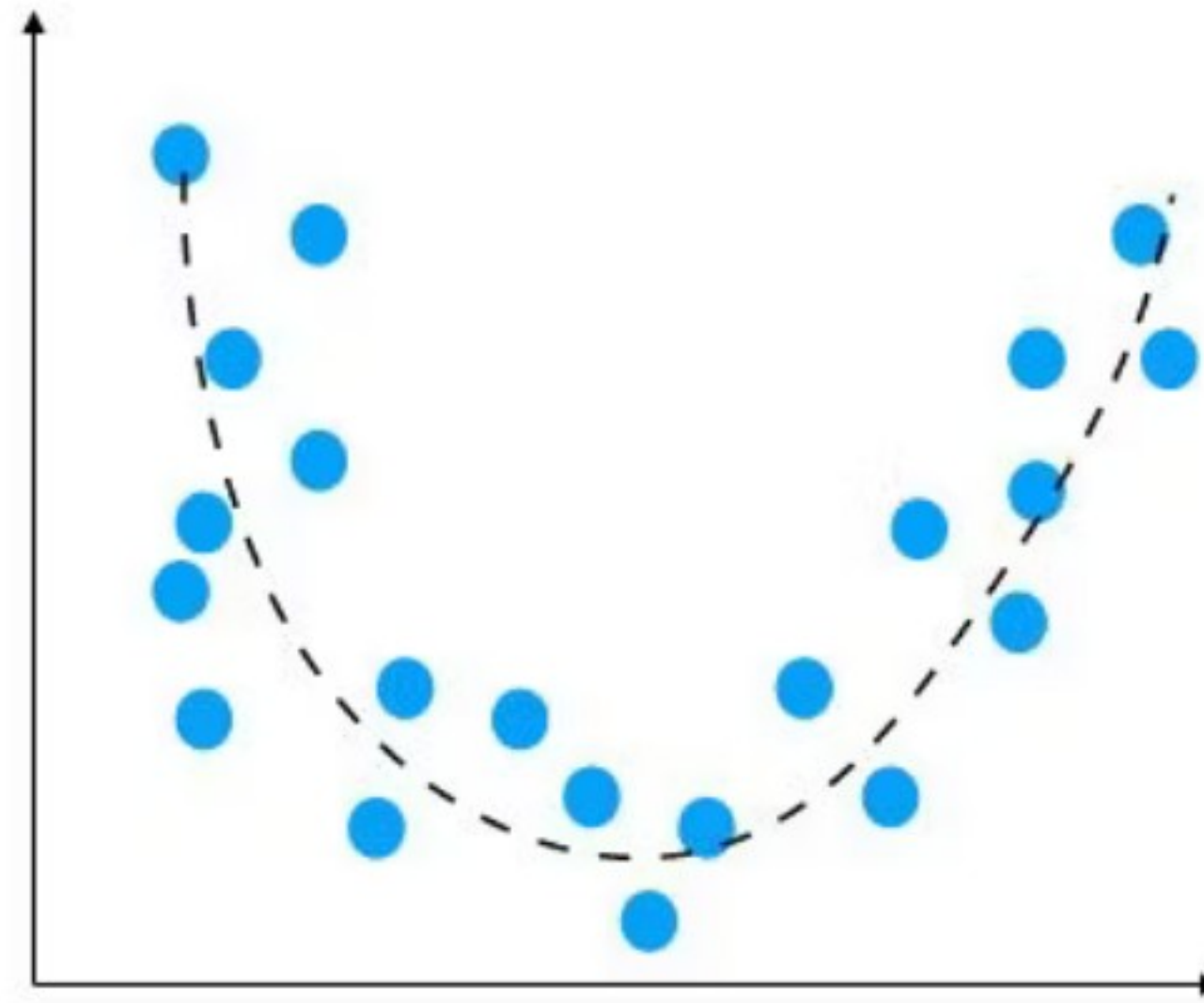
MULTI-CLASS CLASSIFICATION



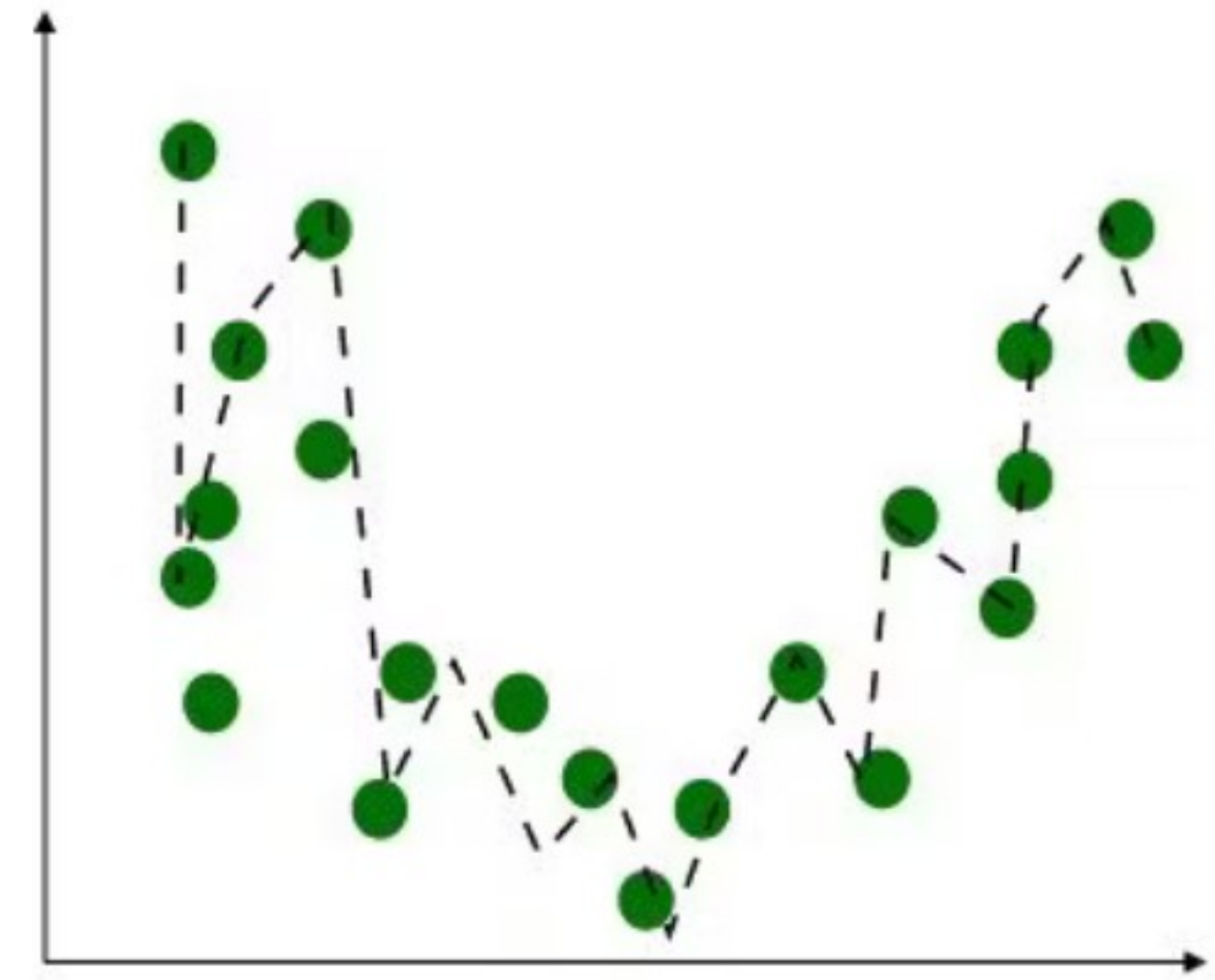
Visuell klassifisering



Under-fitting



Good Fit

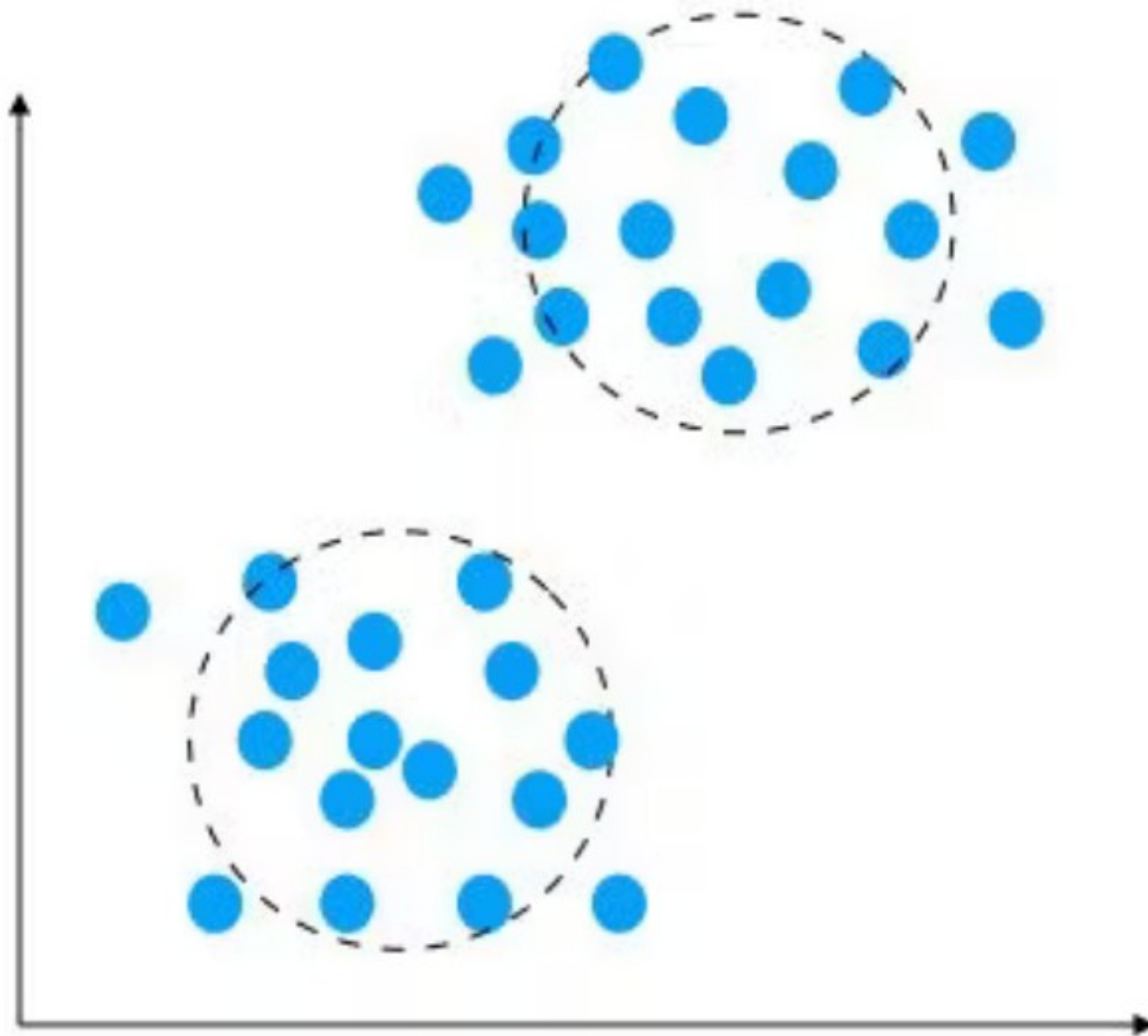


Over-fitting

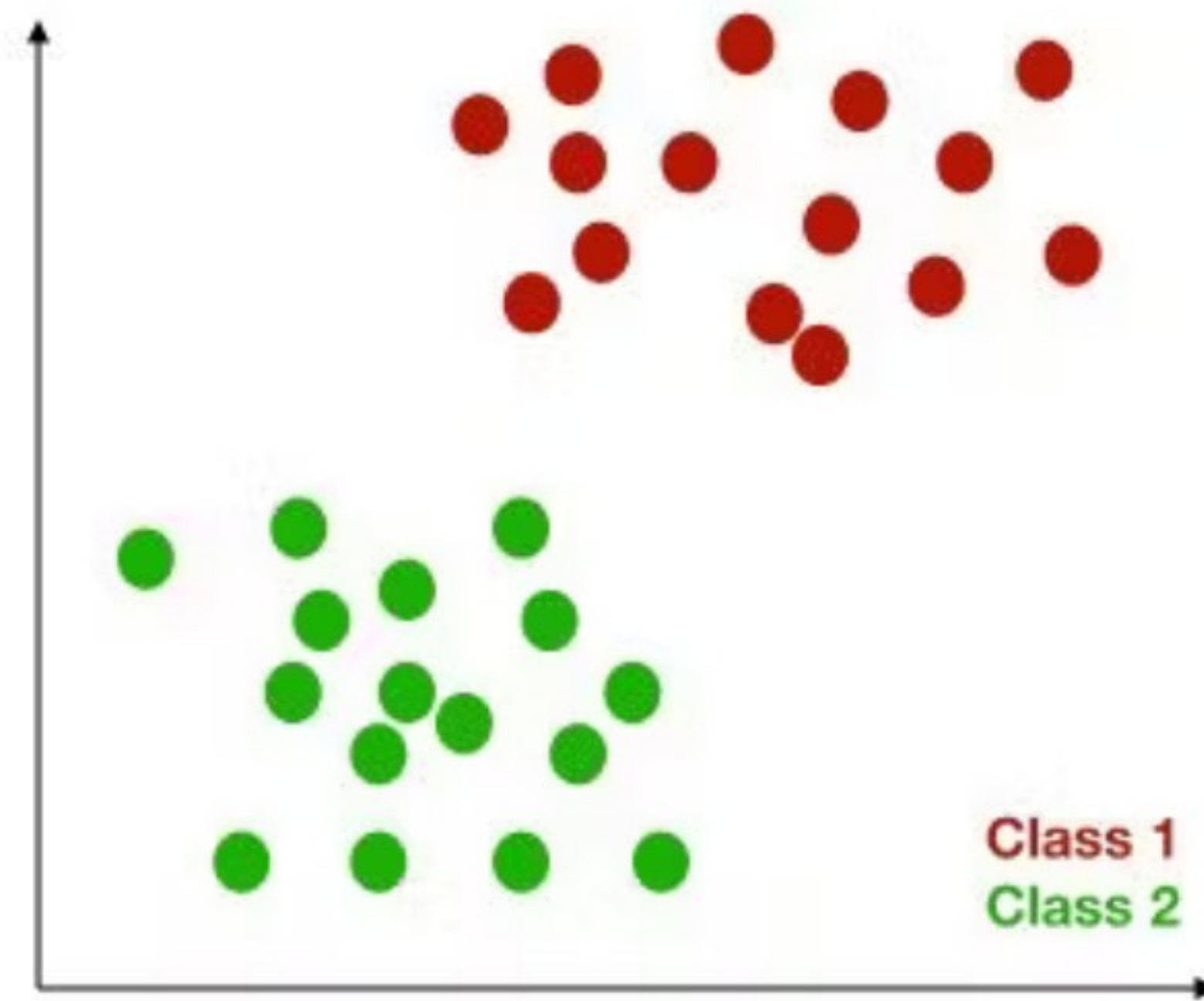
NB: Overfitting

Unsupervised learning

- Du vet ikke kategoriene selv
- 🤖 finner det ut "automatisk"
- Kan gi ny innsikt



Unsupervised



Supervised

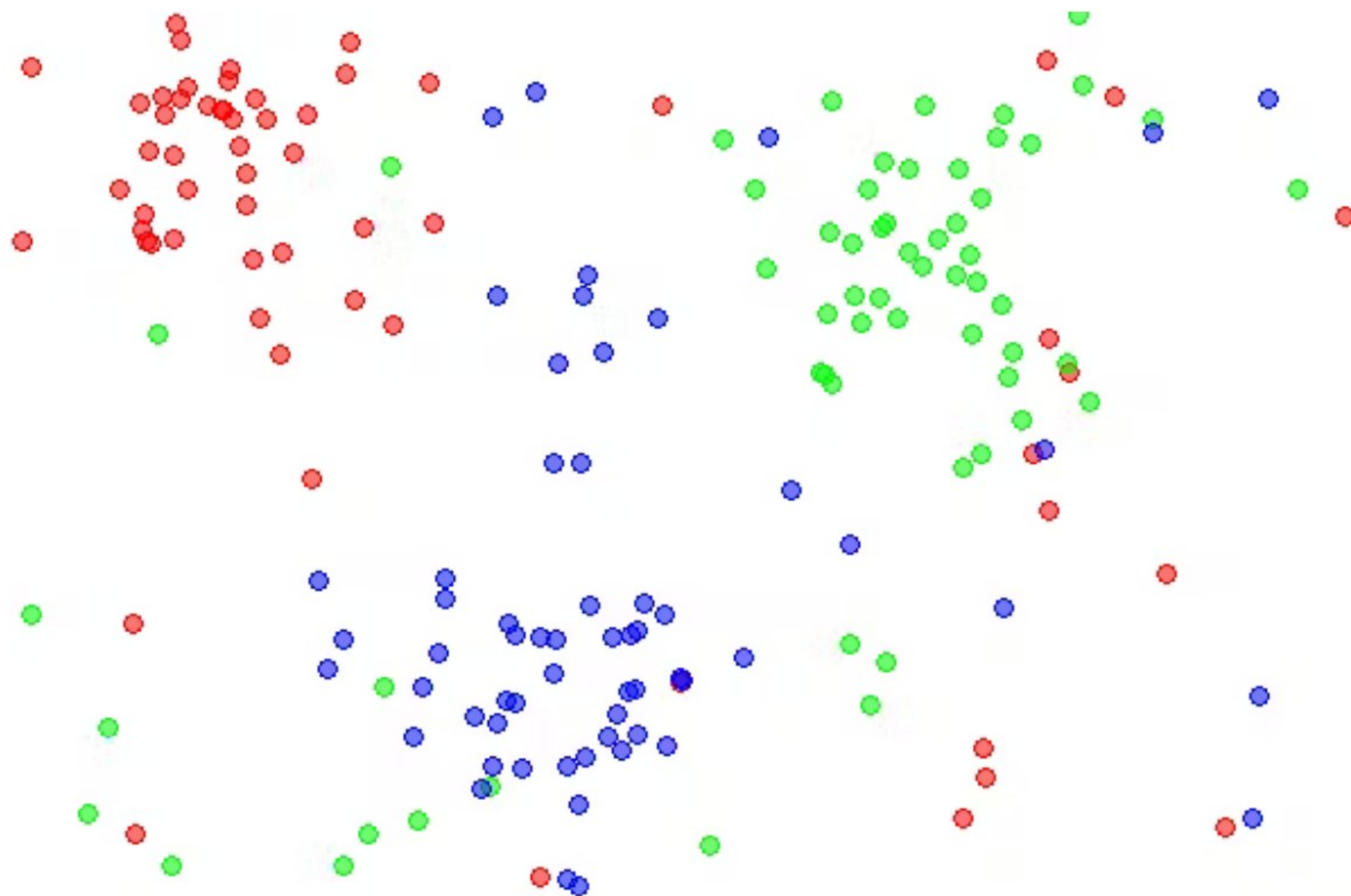
Eksempler på klassifisering

Supervised.

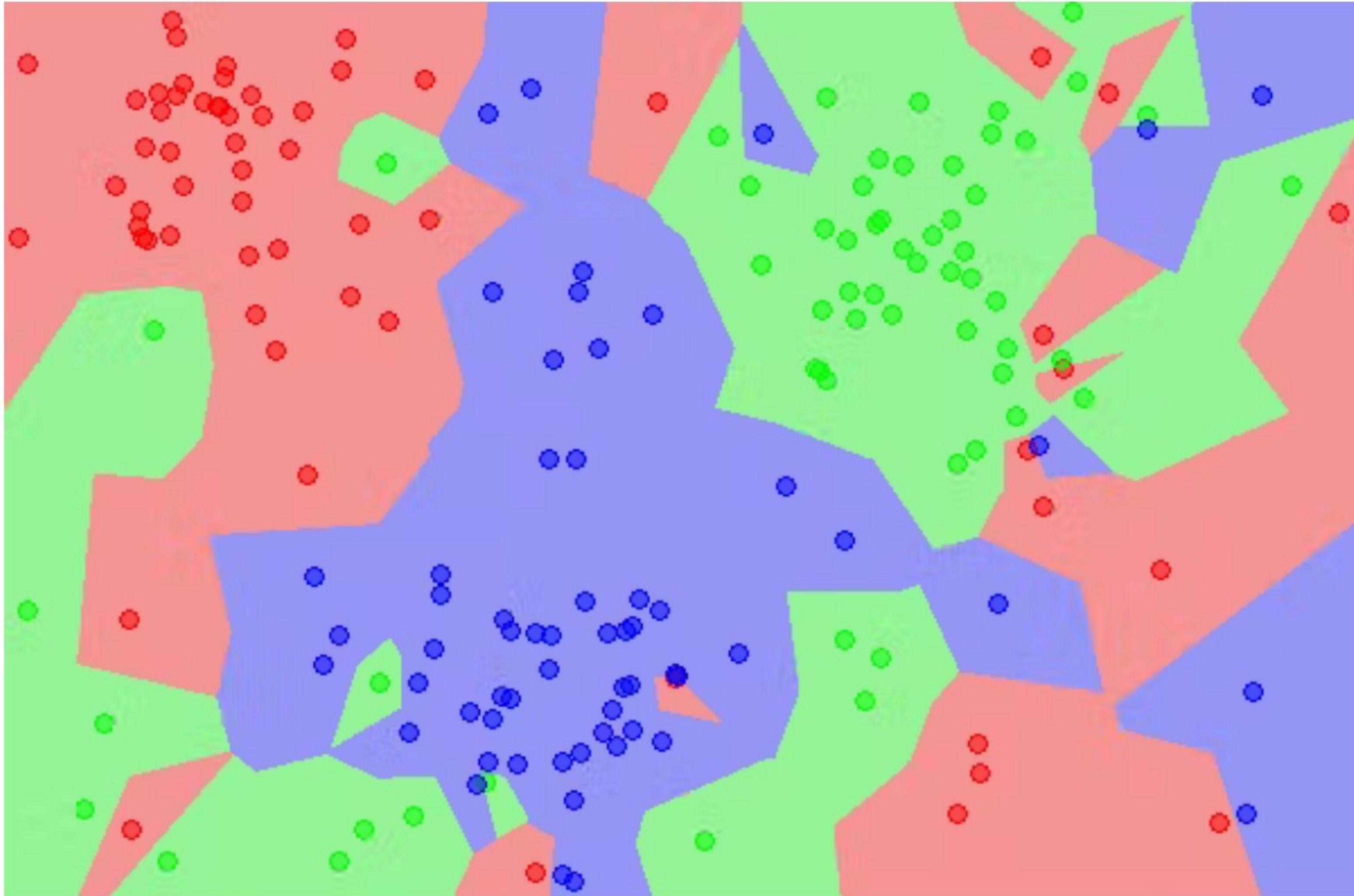
kNN og Rocchio.

kNN: "k nærmeste naboer"

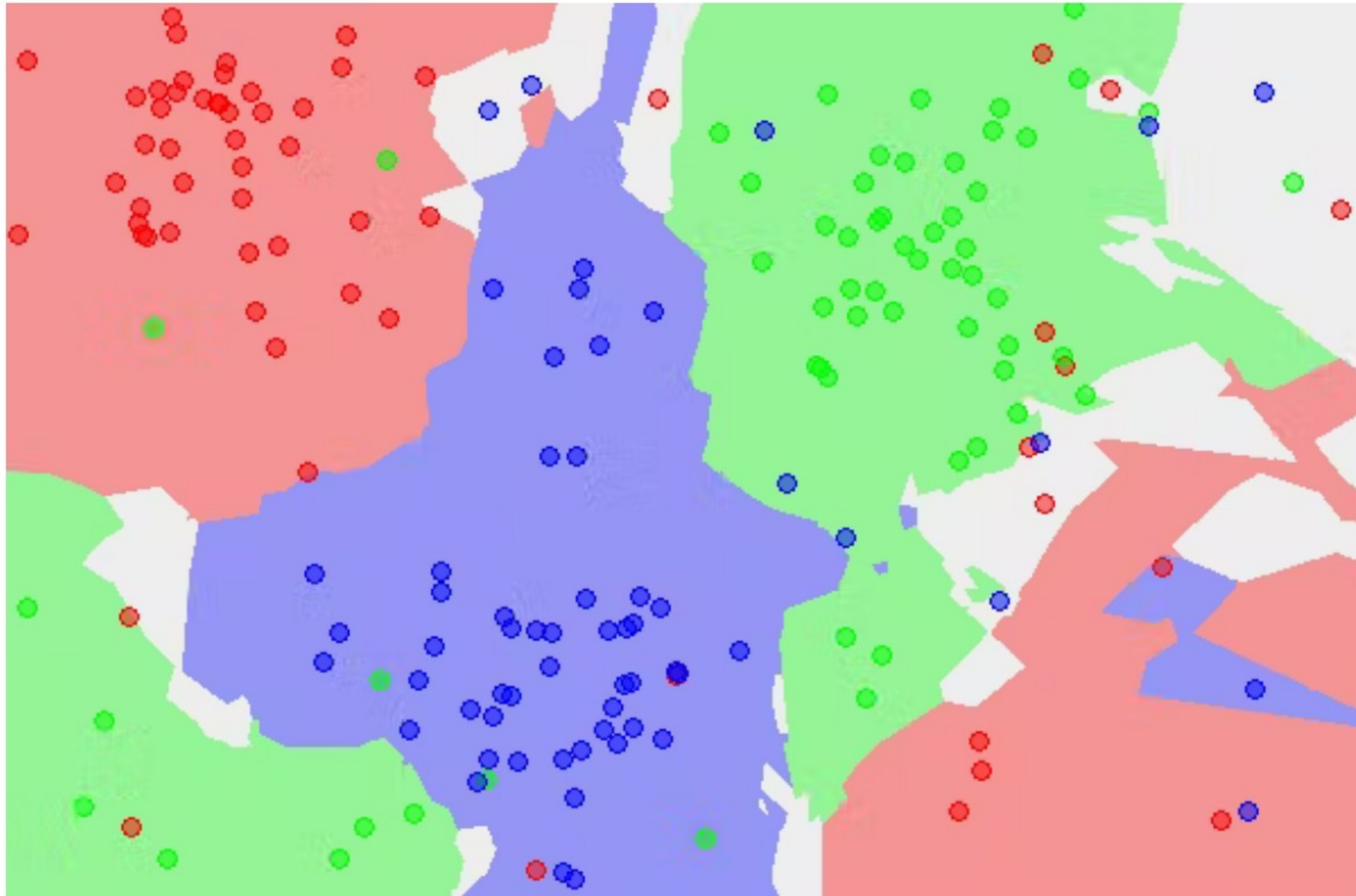
→ Gitt en ting, hva ligger den nærmest?



Starter med treningsdata.



1NN: "1 nærmeste naboer"

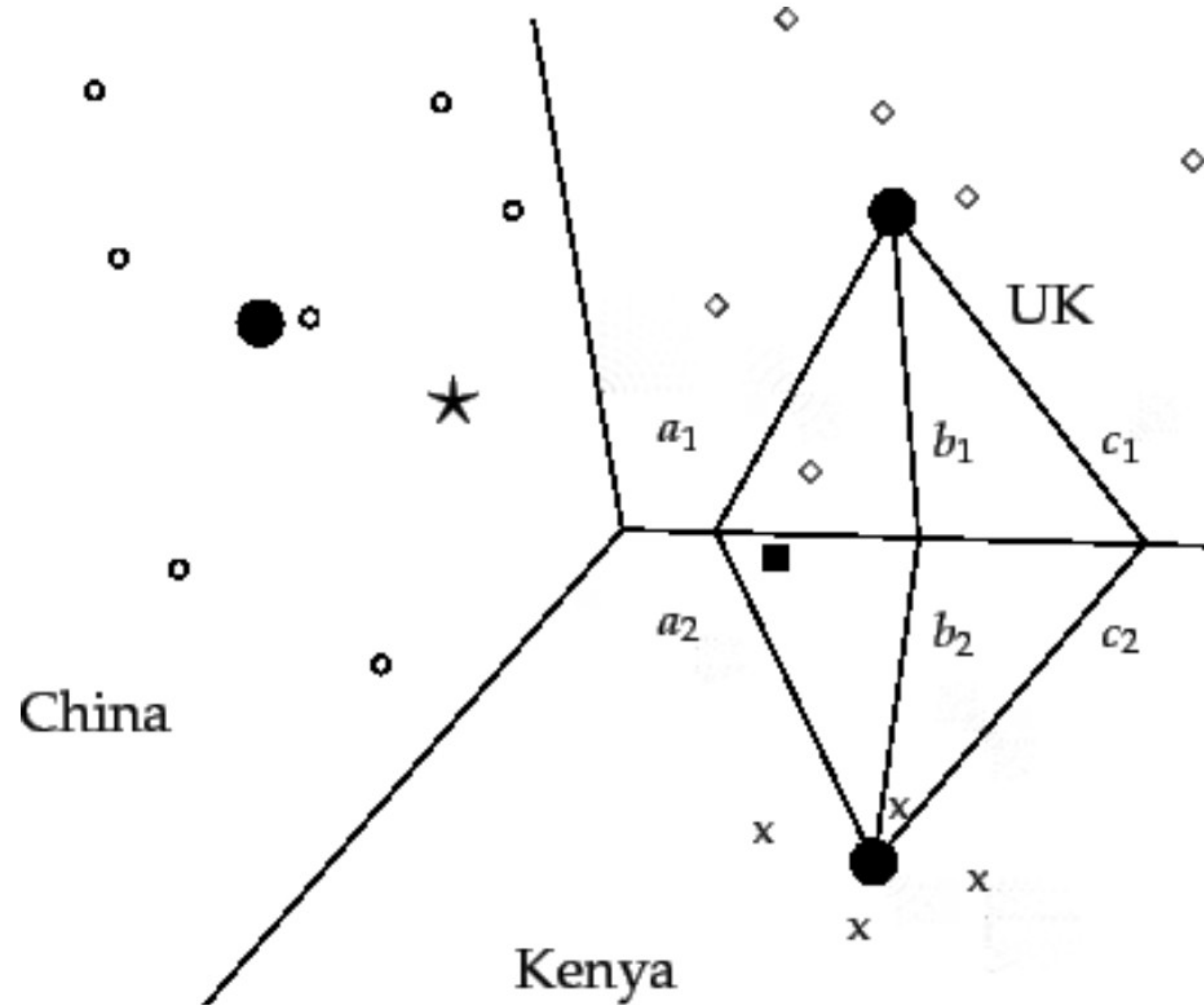


Merk outliers

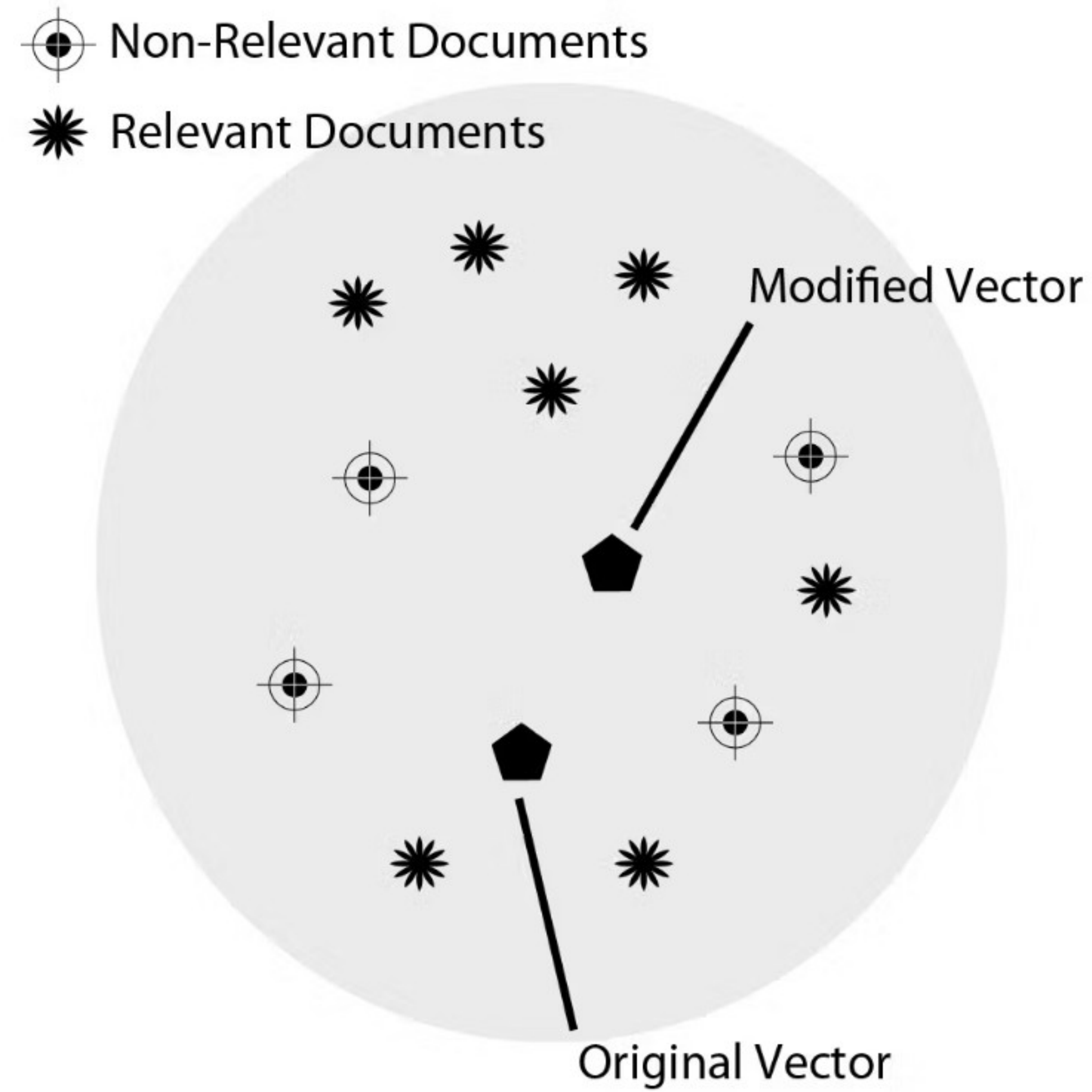
5NN: "5 nærmeste naboer"

Fra kNN til Rocchio

- Rocchio lagrer *centroider* for hver klasse
- Klassifiserer basert på likhet mot centroiden
- En centriode er snittet av klassen



Rocchio classification



Rocchio query expansion

Further reading:

- Stanford IR kap 13
- Assorterte papers (spør ved interesse)
- https://maelfabien.github.io/machinelearning/ml_base/
- <https://github.com/aohrn/in3120-2024/tree/main/seminars/gruppe1/uke08>
- IN2110, IN4050, IN4080, IN5550 (språktek-emner)

Neste gang:

- Naïve bayes
- Introdusere oblig E
- Live-progge C-1?



Assignment workshop. Write something here and we'll discuss it towards the end.





Break until 15:15 😊