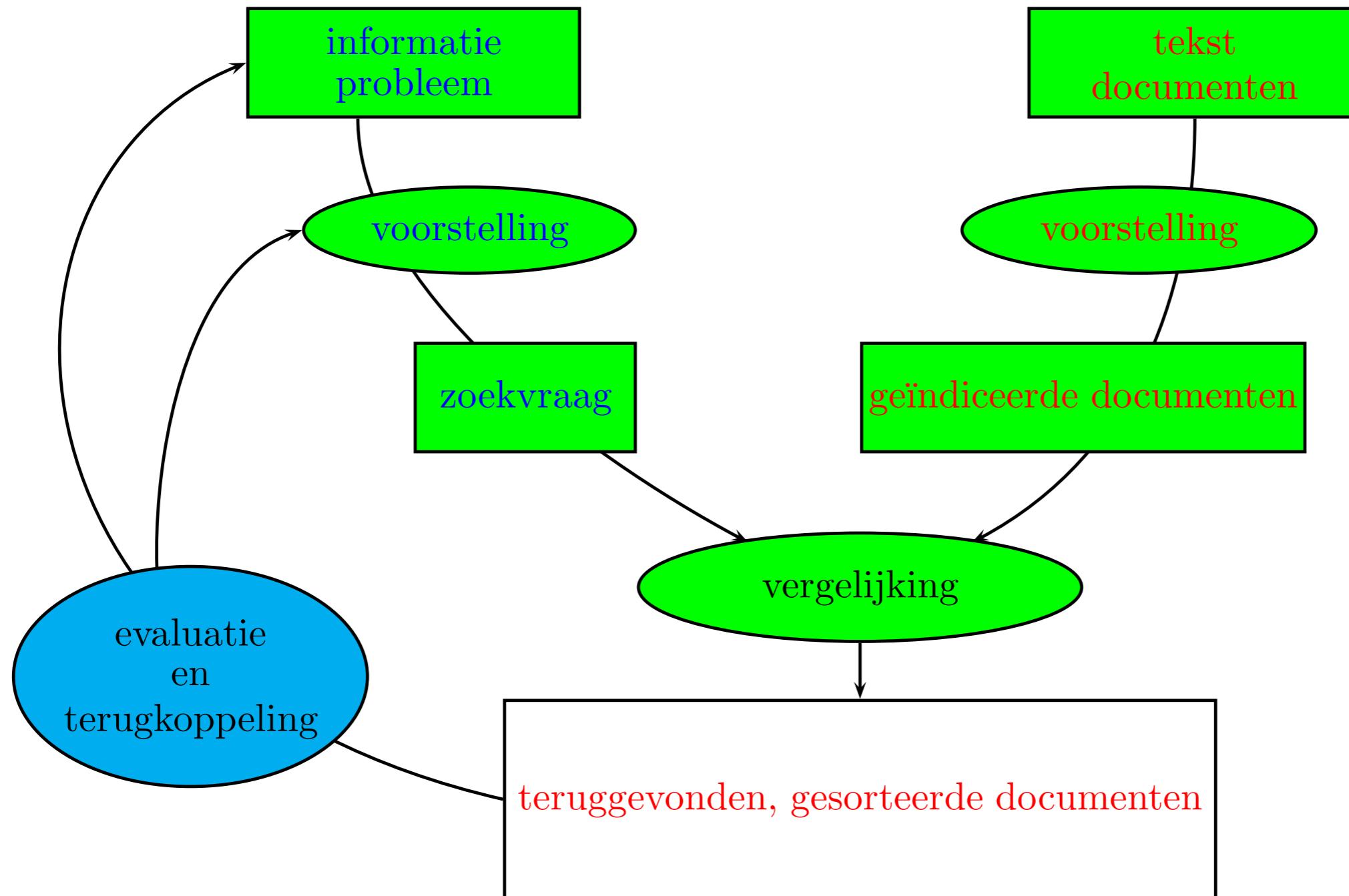


# Aanvullingen IR

# Herhaling



# Index

A word cloud visualization showing the frequency of various terms in a collection of documents. The size of each word indicates its relative importance or frequency. The most prominent term is 'document' (in dark red). Other significant terms include 'term', 'score', 'query', 'bananen', 'boek', 'from', 'gewicht', 'titel', 'gebruiker', and 'relevant'. Smaller terms provide context such as 'precision', 'recall', 'select', 'retrieve', 'contains', 'aantal', 'zoekvraag', 'index', 'zoektermen', 'woorden', 'verzameling', 'verwijderen', 'precision', 'IR', 'where', 'gevonden', 'verschillende', 'meerdere', 'slechts', and 'match'.

# Probleem

# Sommige woorden zijn gerelateerd: kijken naar betekenis



# Thesaurus

- Synoniemenwoordenboek
- Verschillende termen en relaties ertussen
- Doel: helpen bij indexeren en zoeken

# Termen

- Zelfstandige naamwoorden
- In standaardvorm
  - Meervoud voor telbare woorden
  - Zonder voorzetels, accenten, ...
- Niet te specifiek (meerdere documenten)
- Niet te algemeen (niet in alle documenten)
- Evt. met contextinformatie voor homografen

# Oorsprong van termen

- Woordenboeken, woordenlijsten
- Uit documenten
- Van domeinexperten

machine (ICT)

pc

desktop

personal  
computer

bak (ICT)



- Concept wordt beschreven door termen
- Kies één voorkeursterm
- Andere termen verwijzen naar voorkeursterm voor indexering enz.

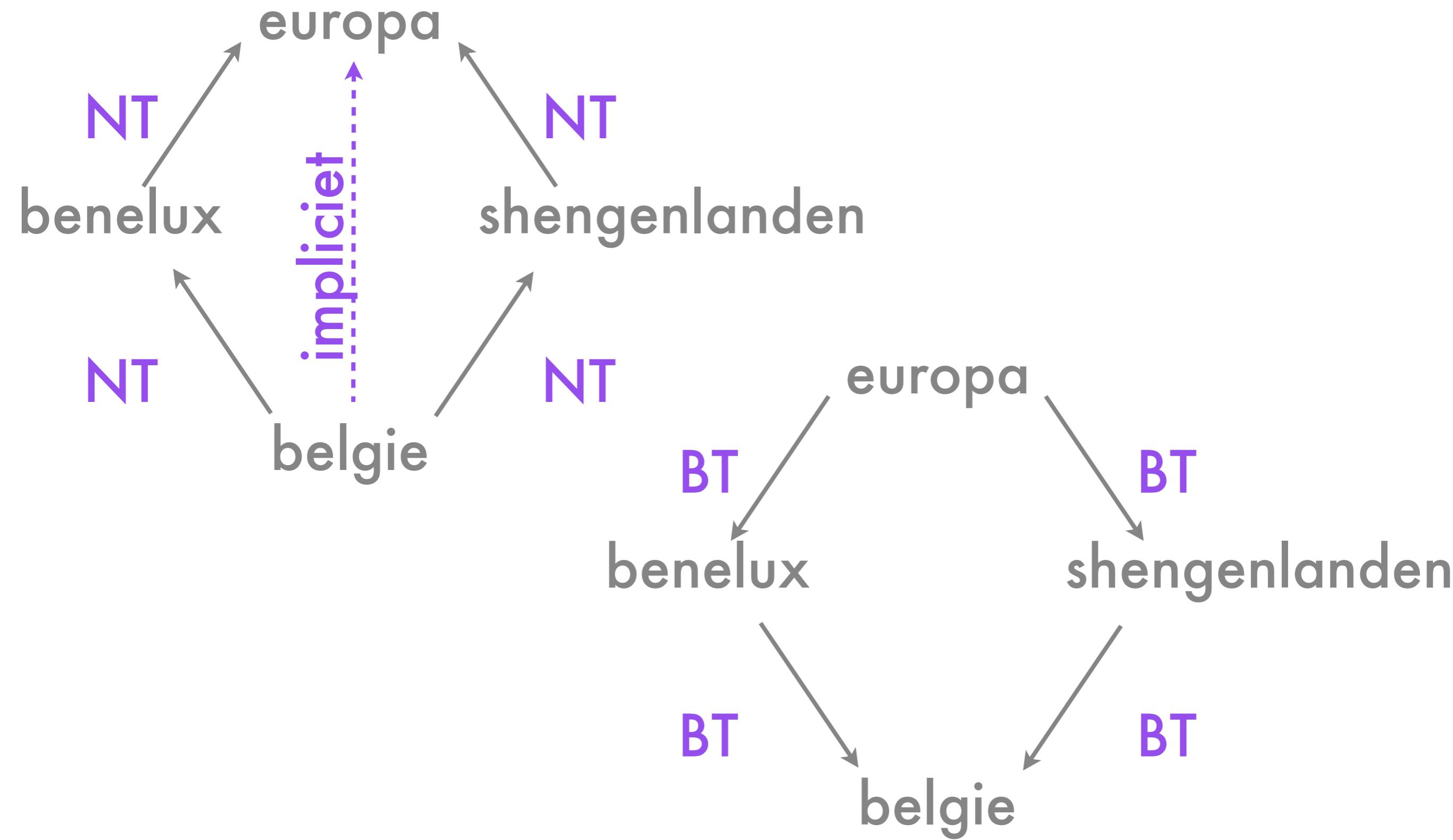
# U(sed)F(or) en USE

- Niet-voorkeur USE voorkeur
- Voorkeur UF niet-voorkeur
  - desktop **UF** bak (ICT)
  - bak (ICT) **USE** desktop
- Meerdere termen
  - pc **USE** windows + desktop
  - windows **UF+** pc

# Semantische relaties

- Hierarchie tussen bredere term (**BT**) en een engere term (**NT**)
- Beide zijn hetzelfde soort van ding en  $BT \supseteq NT$
- Soort: poedel **BT** hond
- Instantiatie: lessius mechelen **BT** hogeschool
- Geheel/onderdeel:
  - onderzoeksgroep **BT** departement
  - belgie **BT** europa

# Structuur: polyhiërarchisch



# Gerelateerde termen

- Geen synoniemen (UF,USE) of hierarchische relatie (BT,NT)
- Maar als je zoekt op het een, kan het ander ook nuttig zijn
- term **RT** term
  - de nayer **RT** sint katelijne waver
- RT is zijn eigen inverse

# Hoe vinden?

tijd	vrijetijdslectuur RT vrijetijd
product	scheepsbouw RT schip
toepassing	computers RT tekstverwerking
onderdeel	voertuigen RT wielen

plaats	vreemde talen RT taallabos
oorzaak	vandalisme RT vijandigheid
apparaat	schilderen RT borstels
complement	ouders RT kinderen

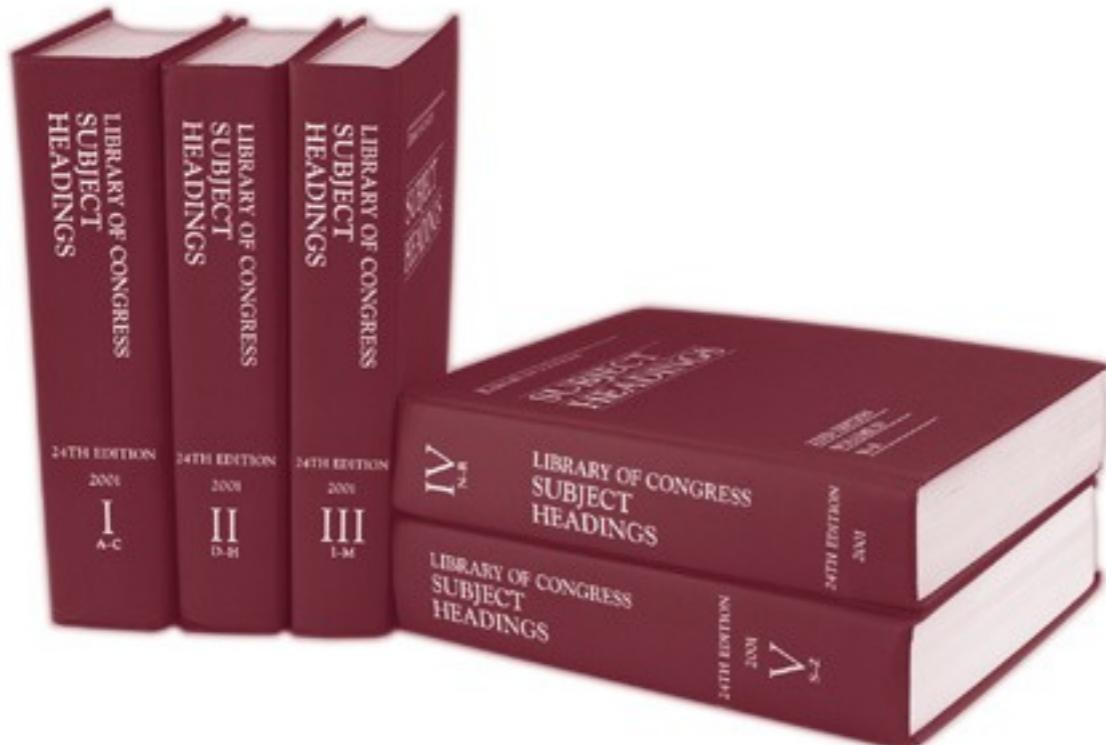
# Scope notes

- Meer informatie geef bij term
  - Definitie
  - Waarschuwing of hints over gebruik
- hogeschool **SN** aanbieder van niet-universitair hoger onderwijs
- mechelen lessius **SN** bevat tegenwoordig ook Campus De Nayer

# Library of Congress



# Library of Congress Subject Headings



<b>Ergoloid mesylates</b> ( <i>May Subd Geog</i> ) <i>[RM666.E78]</i>
<b>UF</b> Dihydroergotoxine
Dihydrogenated ergot alkaloids
<b>BT</b> Ergot alkaloids
Psychotropic drugs
<b>Ergometer</b>
<b>USE</b> Dynamometer
<b>Ergonomics</b>
<b>USE</b> Work measurement
<b>Ergonomics</b>
<b>USE</b> Human engineering
<b>Ergosterol</b>
<b>BT</b> Sterols
<b>NT</b> Ergocalciferol
<b>Ergot</b>
<i>[RSI65.E7 (Pharmacy)]</i>
<b>UF</b> Rye smut
Spurred rye
<b>BT</b> Adrenergic alpha blockers
Dopamine—Agonists
<b>NT</b> Ergot alkaloids

online versie

# In Oracle

## Inladen uit tekstbestand met ctxload

term

SYN synoniem

NT engere term

BT bredere term

RT gerelateerde term

USE geprefereerde term

SN tekst

term

NT1 engere term

NT2 engere term

NT3 engere term

NT2 engere term

# Gebruiken

```
declare termen varchar2(200);
begin
    termen := ctx_thes.bt('bananen',3);
    dbms_output.put_line(' bredere termen voor bananen: ' || termen );
end;

select titel , score(1) from boek
where contains(tekst, 'bt(bananen)', 1) > 0
```

# Classificatie (van documenten)

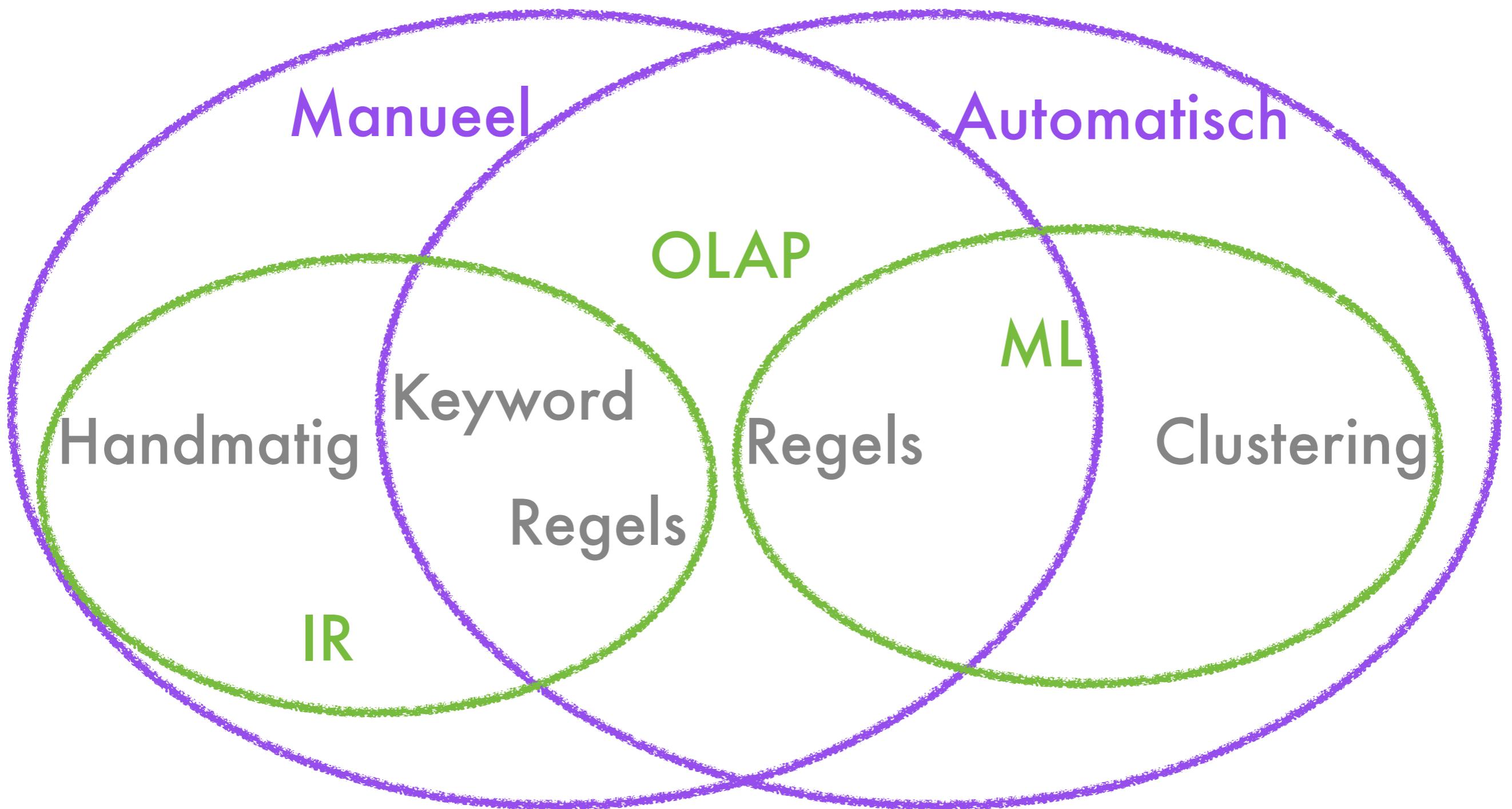
# Classificatie

- Grote verzameling documenten
- Opdelen in 2 delen, bv.
  - Interessant
  - Niet interessant
- IR is classificatie taak

# Andere voorbeelden

- Produkten in supermarkt
  - Succesvol
  - Onsuccesvol
- Moleculen
  - Actief
  - Niet actief
- Financiële transacties
  - Verdacht
  - Niet verdacht
- Studenten
  - Extra begeleiding
  - Niet nodig

# Technieken



# Technieken

- Handmatig elk document classificeren (~ web directories)
- Zoeken naar keywords (~ search engines)
- Gebruik maken van **classificatieregels**: interessant IF iPhone | iPad | iPod & touch
  - Handmatig schrijven
  - Automatisch ontdekken

# Automatisch ontdekken

- Gesuperviseerd machinaal leren
- Gegeven: training data
  - Positieve voorbeelden
  - Negatieve voorbeelden
- Vind: regels die zo goed mogelijk training data vatten

# Voorbeeld

Een iPhone is cool

Windows 7 is beter dan Windows Vista

Een iPad heeft een touchscreen

Op een iPhone draait geen Windows bro!

Machine Learning

interessant IF (iPhone | iPad) & ! Windows

# Clustering (unsupervised)

Een iPhone is cool

Windows 7 is beter dan Windows Vista

Een iPad heeft een touchscreen

Op een iPhone draait geen Windows bro!

Taak: deel op in 2 groepen

# Clustering

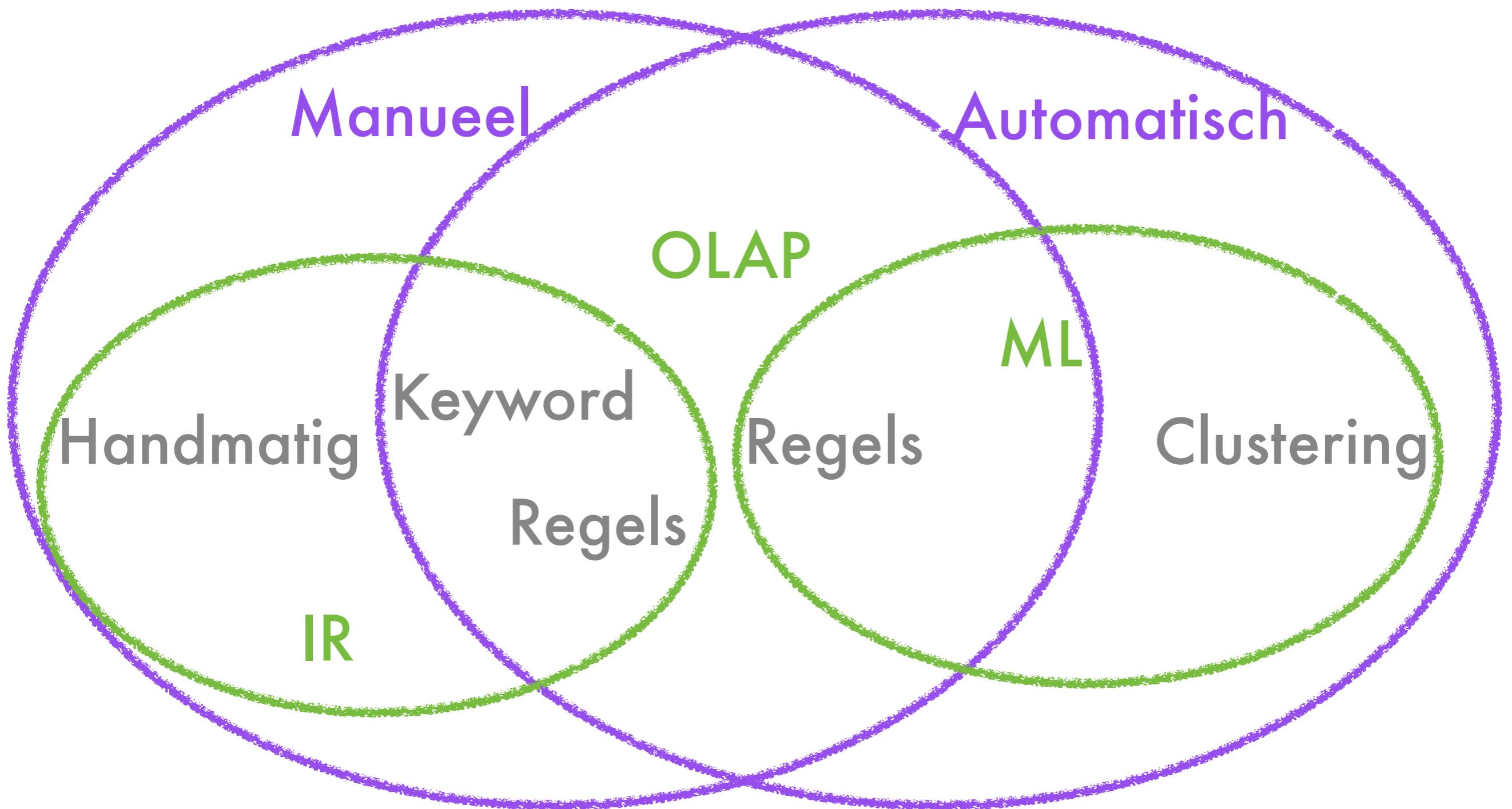
Een iPhone is cool

Een iPad heeft een touchscreen

Op een iPhone draait geen Windows bro!

Windows 7 is beter dan Windows Vista

# Technieken



# ML algoritmes

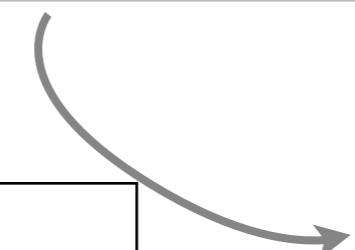
Een iPhone is cool

Windows 7 is beter dan Windows Vista

Een iPad heeft een touchscreen

Op een iPhone draait geen Windows bro!

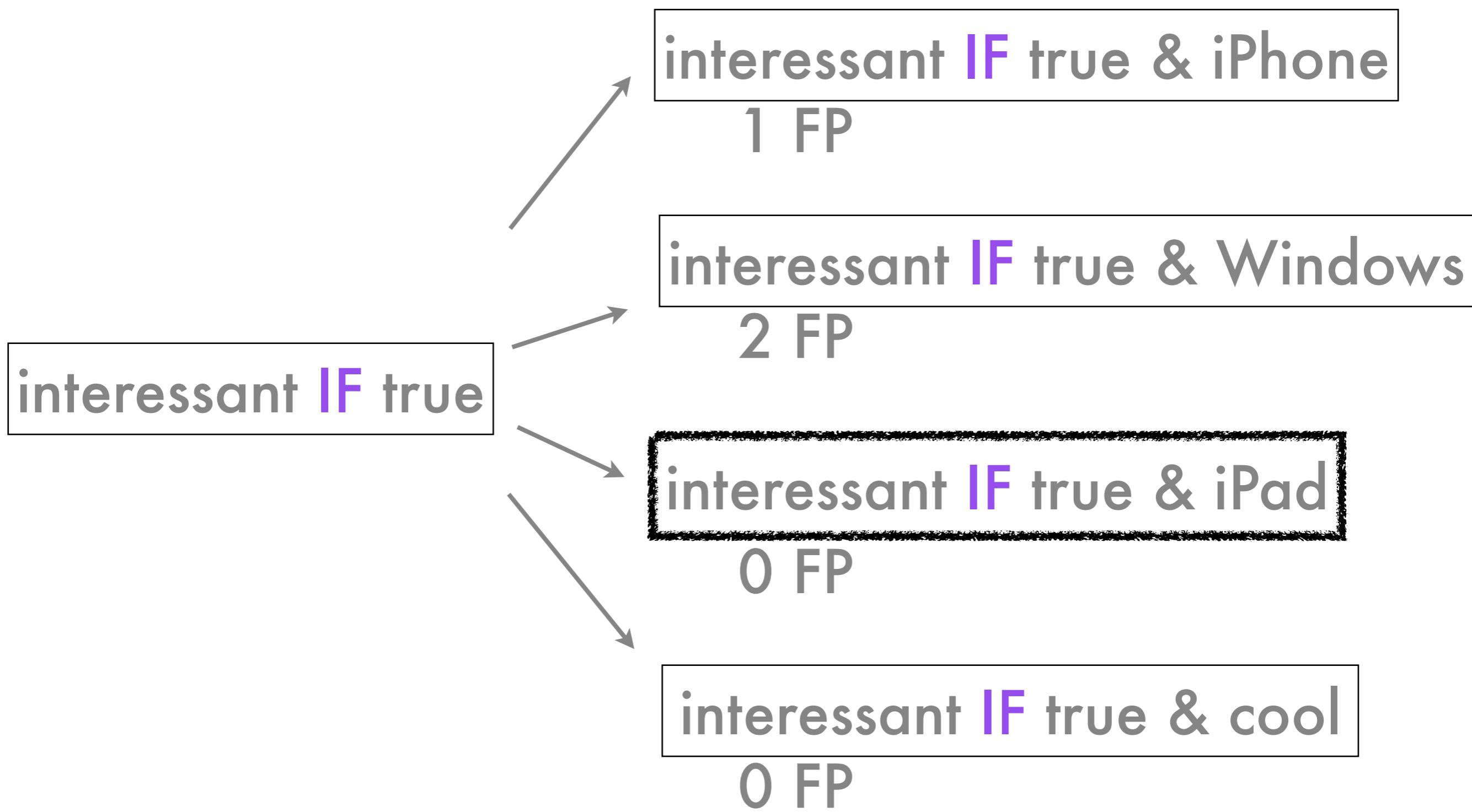
interessant **IF** true



fouten: 2 false positives

dus verstrekken

# ML algoritmes



# ML algoritmes

Een iPhone is cool

Windows 7 is beter dan Windows Vista

Een iPad heeft een touchscreen

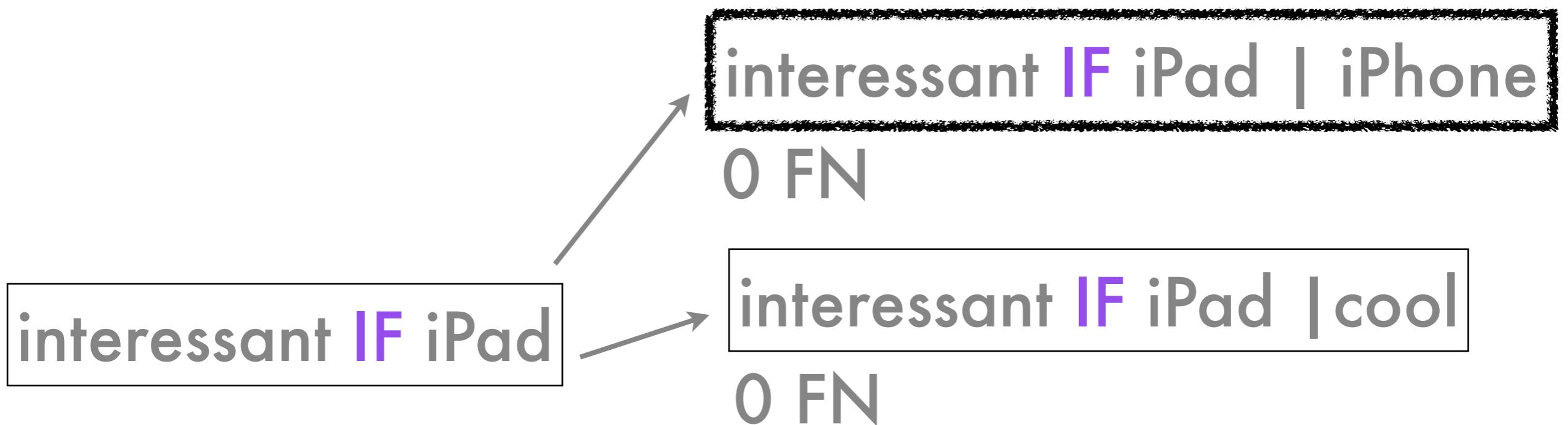
Op een iPhone draait geen Windows bro!

interessant IF iPad

fouten: 1 false negative

dus versoepelen

# ML algoritmes



# ML algoritmes

Een iPhone is cool

Windows 7 is beter dan Windows Vista

Een iPad heeft een touchscreen

Op een iPhone draait geen Windows bro!

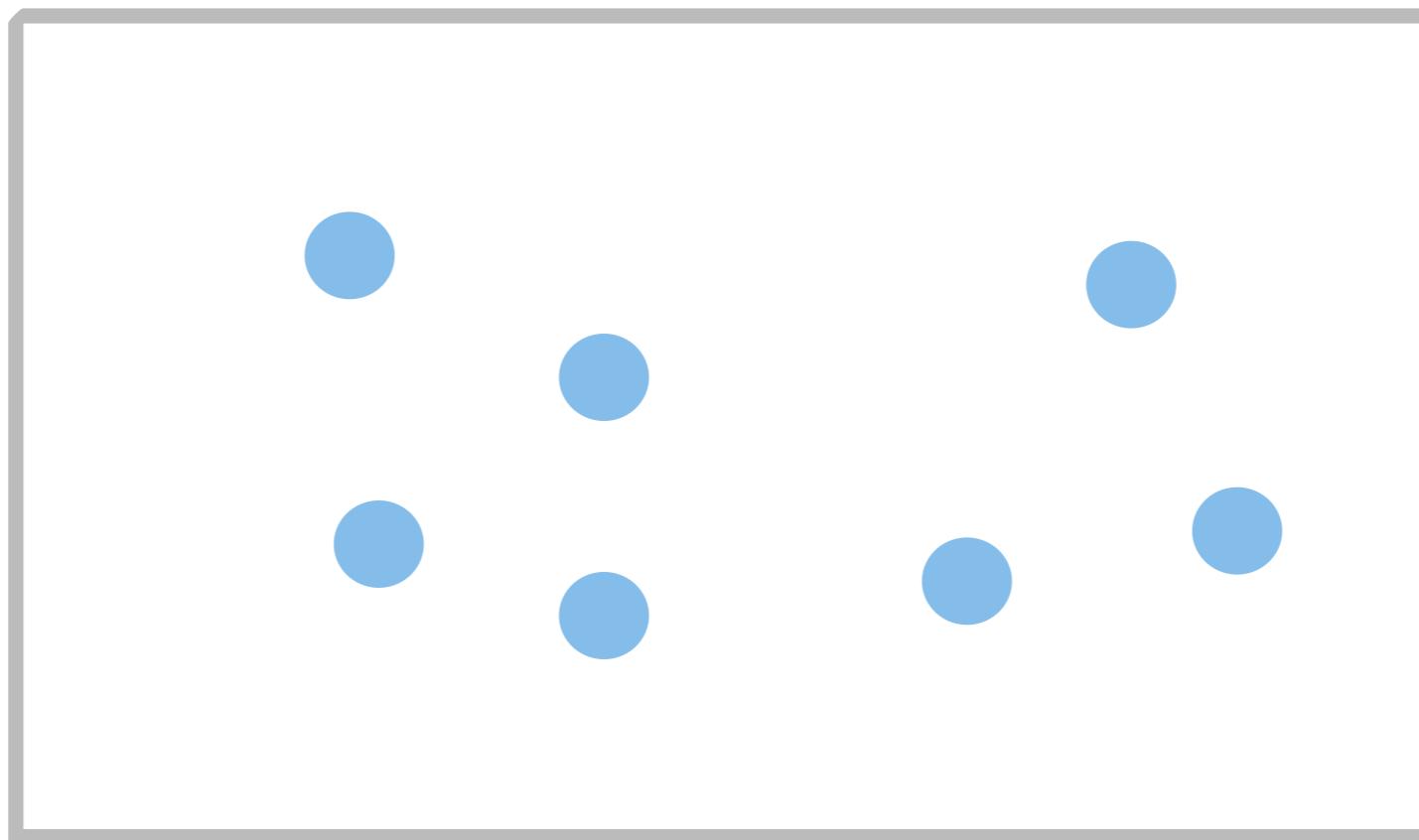


fouten: 1 false positive

interessant IF iPad | iPhone

dus verstrekken

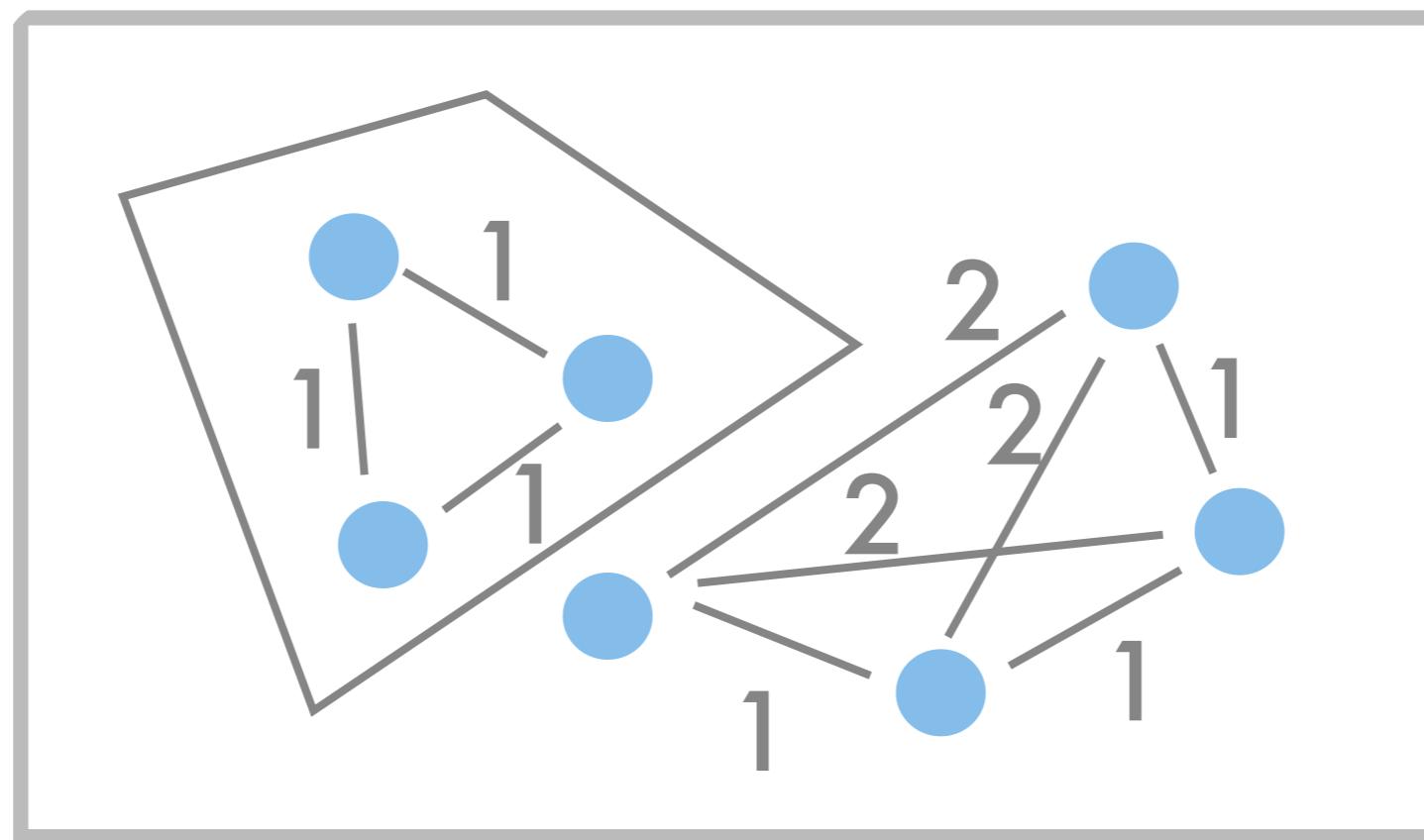
# k-means clustering



Interne afstand in cluster C:  $\sum_{x,y \in C} (x - y)^2$

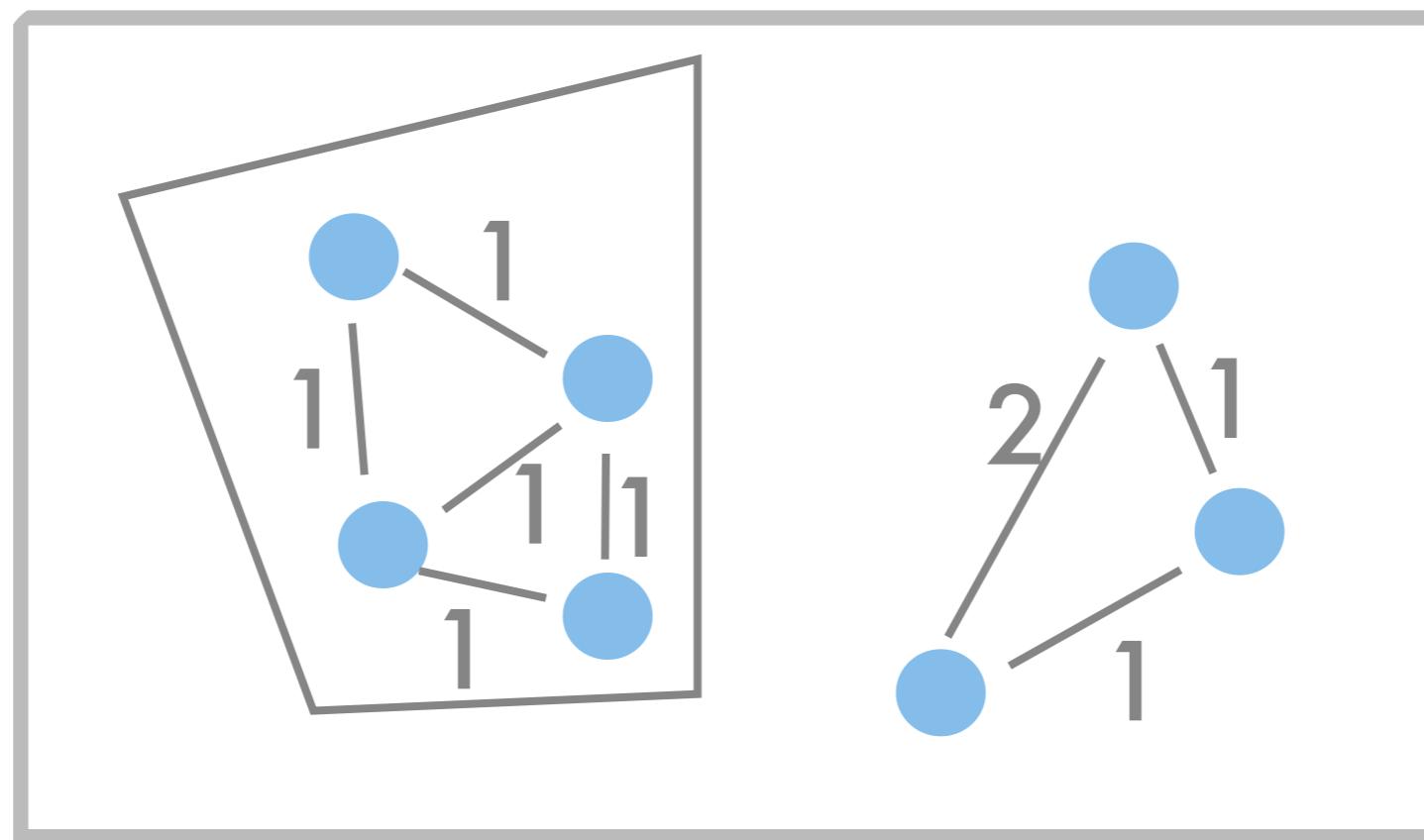
Totale afstand van clustering:  $\sum_{C \in \mathcal{C}} \sum_{x,y \in C} (x - y)^2$

# k-means clustering



Totale afstand van clustering:  
 $(1+1+1) + (2+2+2+1+1+1) = 12$

# k-means clustering

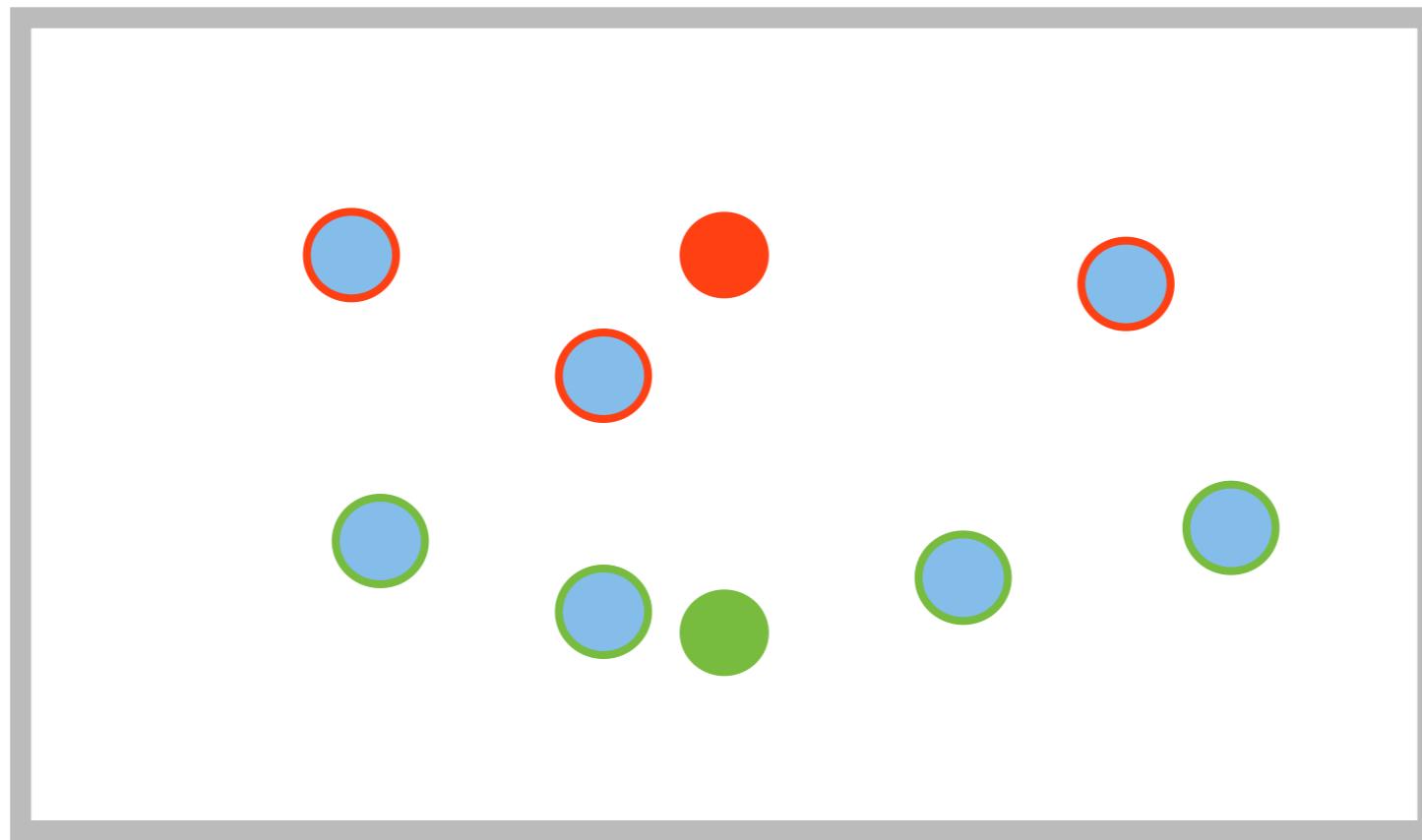


Totale afstand van clustering:  
 $(1+1+1+1+1) + (2+1+1) = 9$

# k-means clustering

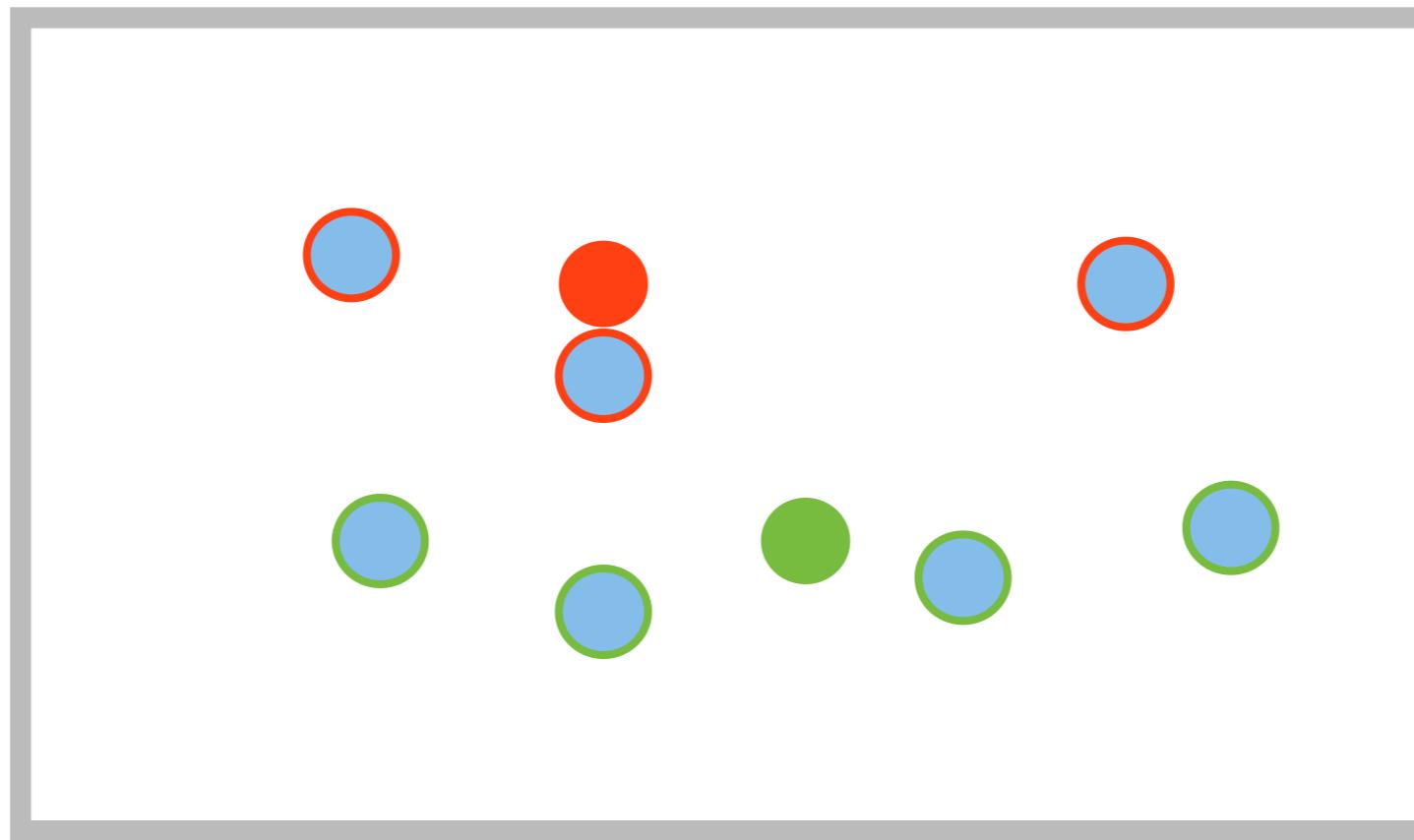
- Doel:  $\operatorname{argmin}_{\mathcal{C}} \sum_{C \in \mathcal{C}} \sum_{x, y \in C} (x - y)^2$
- Niet exact te berekenen
- Dus iteratief benaderen
  - Cluster voorstellen door zijn zwaartepunt
  - Begin: kies willekeurig  $k$  zwaartepunten

# k-means clustering



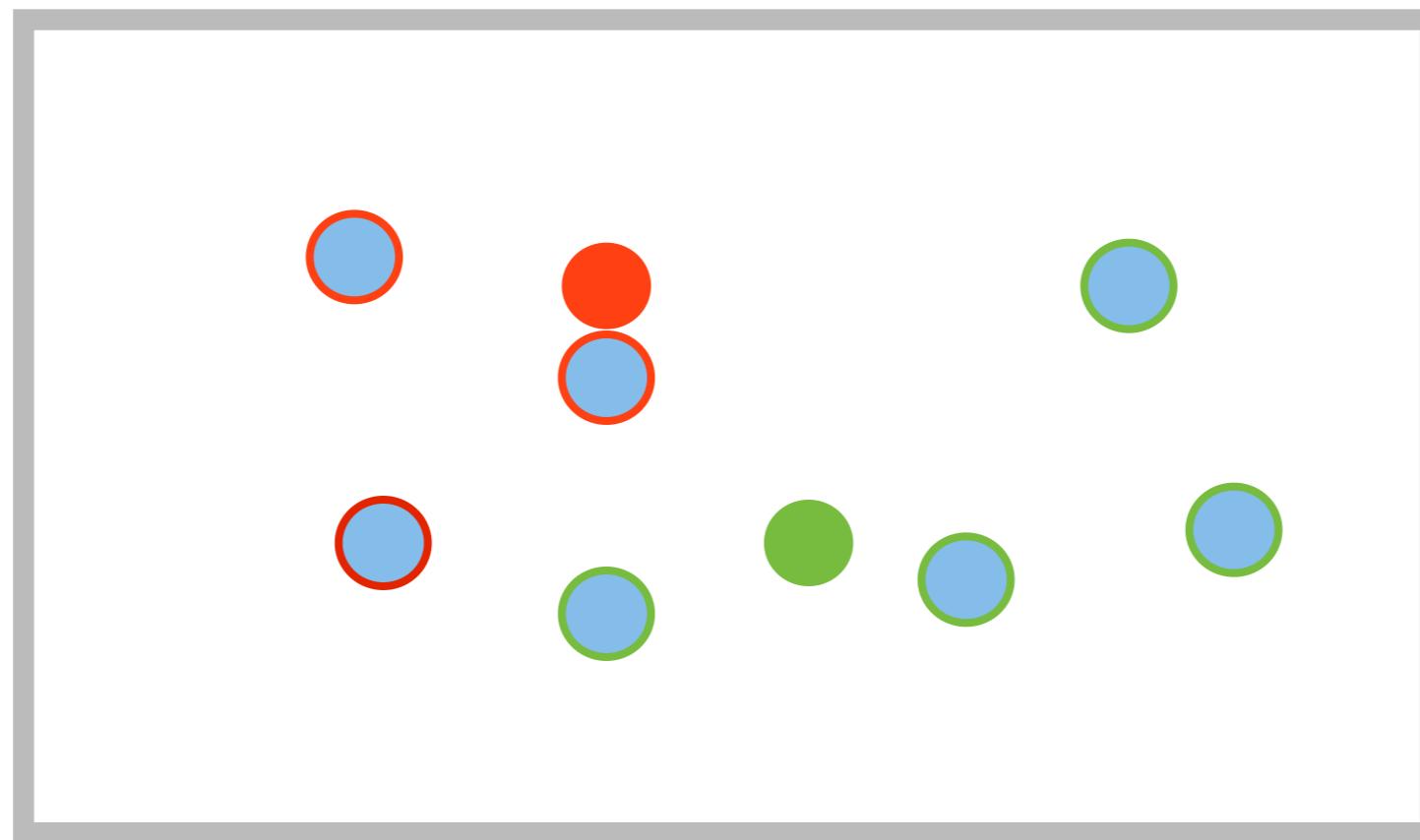
- Maak clusters: elk punt gaat naar cluster van dichtstbijzijnde zwaartepunt
- Volgende stap: verschuif zwaartepunten

# k-means clustering



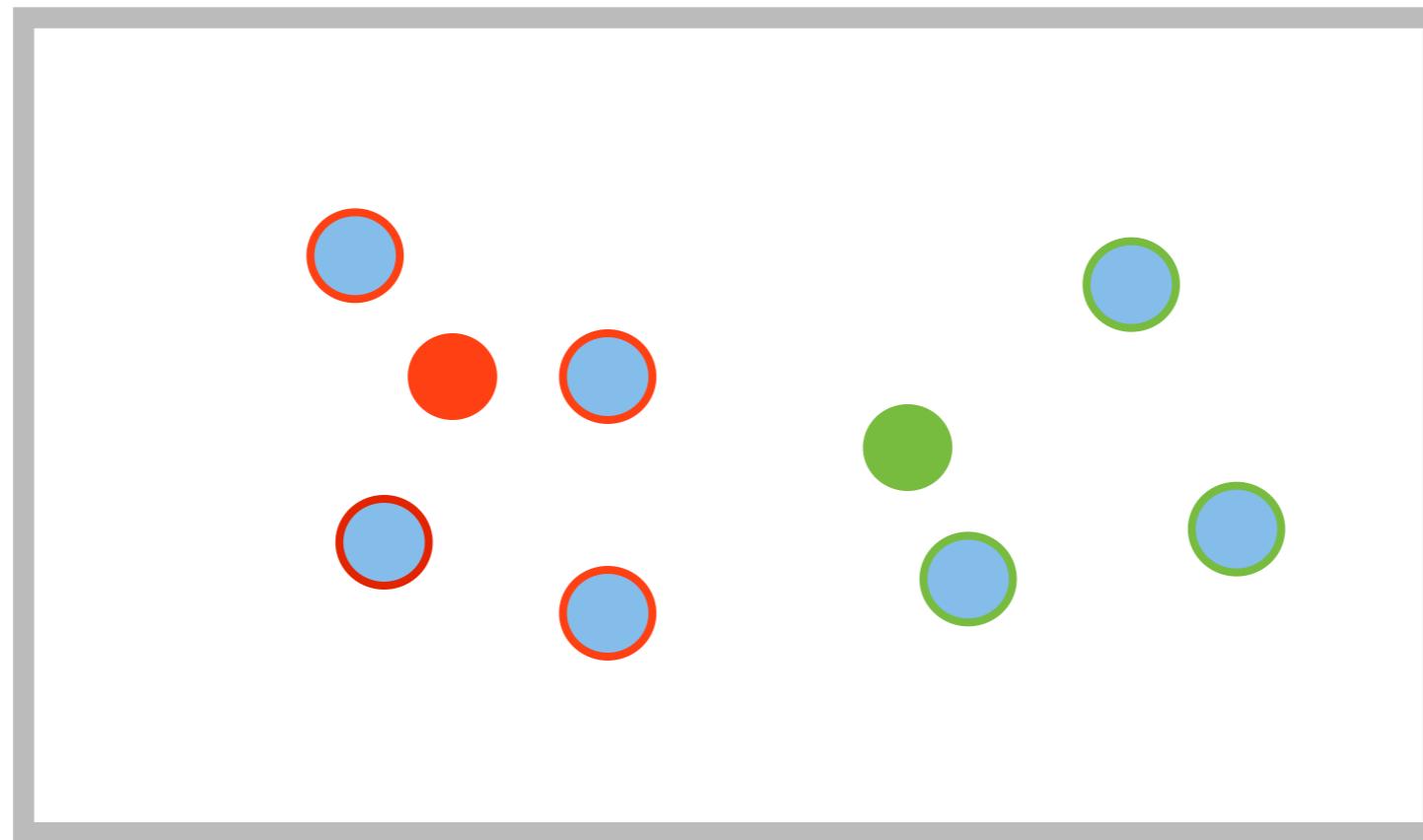
- Volgende stap: pas clusterindeling aan

# k-means clustering



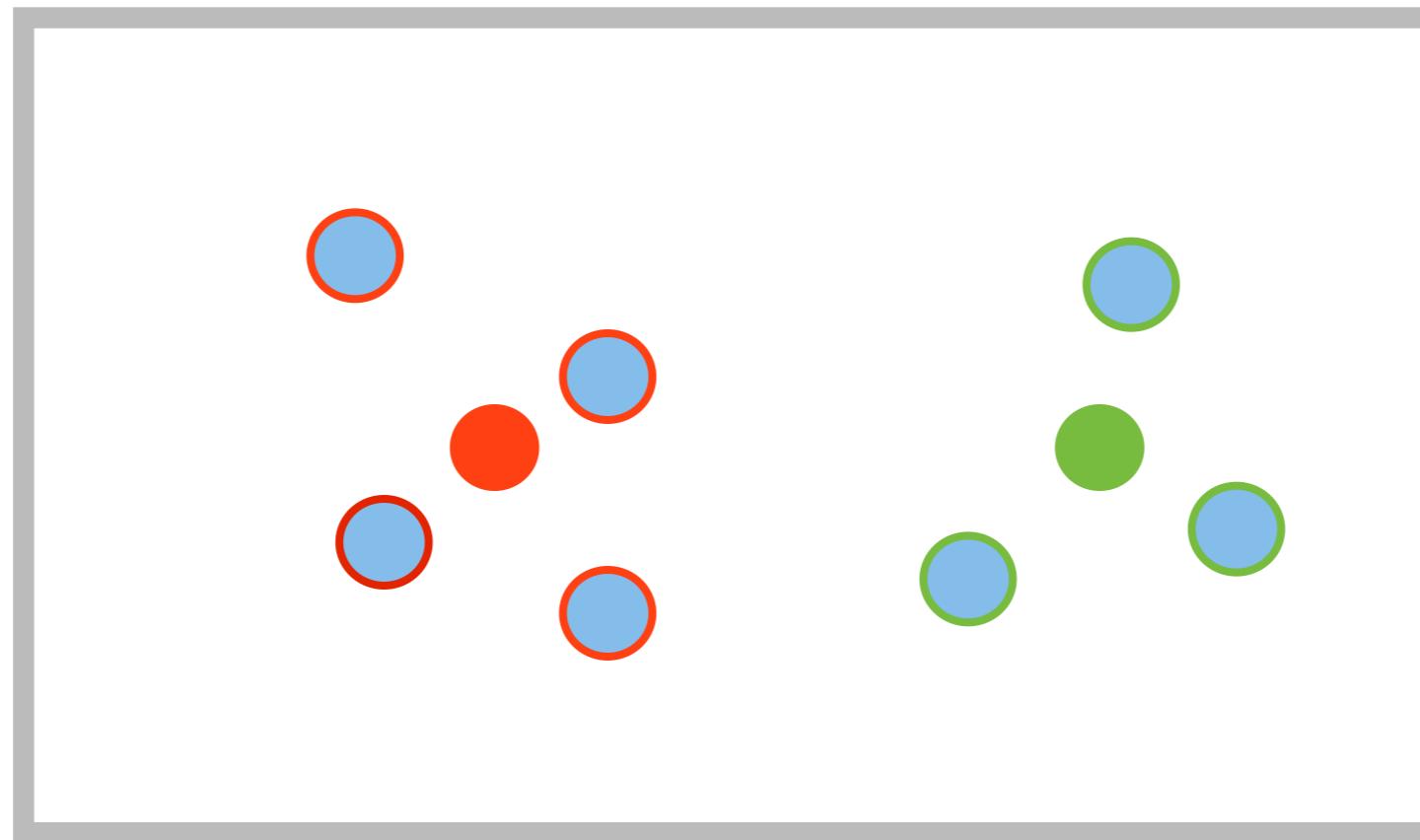
- Herhaal tot er niets meer verandert

# k-means clustering



- Herhaal tot er niets meer verandert

# k-means clustering



- Herhaal tot er niets meer verandert