



# Collectieve Intelligentie

SIMON PAUW

---

VECTOREN EN INPRODUCT







# Type algoritmes

---

## Regressie

Voorspellen continue waarden

Voorbeeld: ratings

## Classificatie

Voorspellen discrete categorieën

Voorbeeld: recommend/do not recommend



Matrix-  
factorizatie



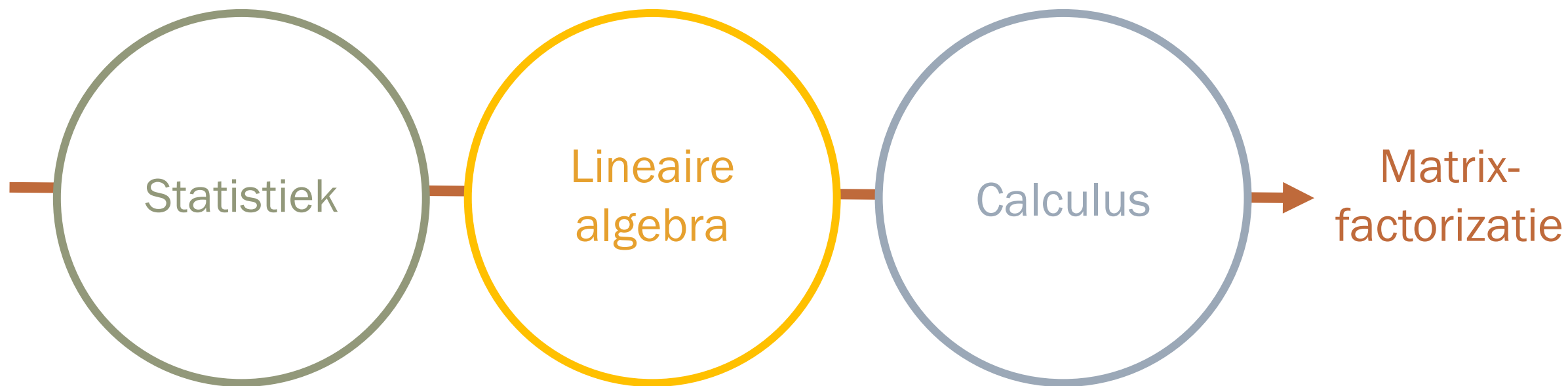
```
graph LR; A((Statistiek)) --> B[Matrix-factorizatie];
```

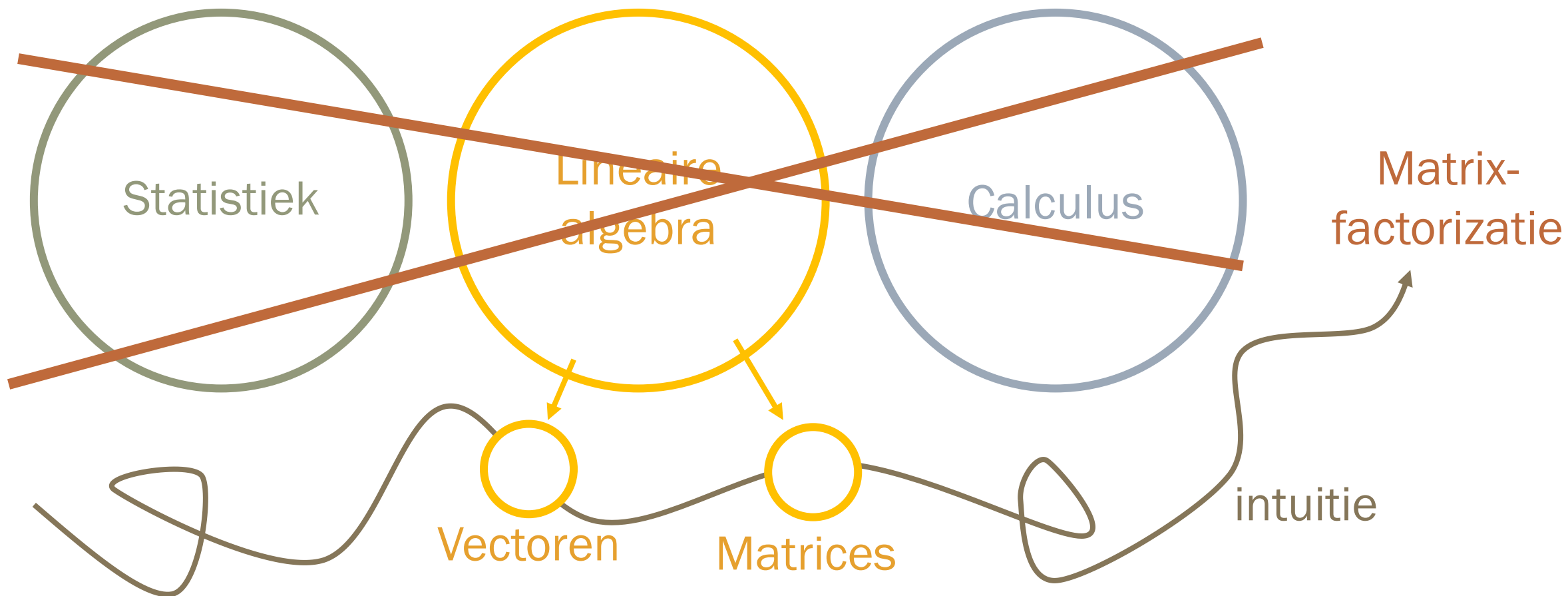
Statistiek

Matrix-  
factorizatie









# Cosine similarity

---

$$\cos(a, b) = \frac{\sum_{i=1}^n a_i \cdot b_i}{\sqrt{\sum_{i=1}^n a_i^2} \cdot \sqrt{\sum_{i=1}^n b_i^2}}$$

# Cosine similarity

---

	Meike	Lars	Rudolf	...
<i>Titanic</i>	0.9	0.5	0.9	
<i>Luca</i>	0.7	0.8	0.7	
<i>The Great Dictator</i>	1.0	0.3	0.8	
...				

$$\cos(a, b) = \frac{\sum_{i=1}^n a_i \cdot b_i}{\sqrt{\sum_{i=1}^n a_i^2} \cdot \sqrt{\sum_{i=1}^n b_i^2}}$$

$$\cos(\text{Luca}, \text{The Great Dictator}) = \frac{0.7 \cdot 1.0 + 0.8 \cdot 0.3 + 0.7 \cdot 0.8}{\sqrt{0.7 \cdot 0.7 + 0.8 \cdot 0.8 + 0.7 \cdot 0.7} \cdot \sqrt{1.0 \cdot 1.0 + 0.3 \cdot 0.3 + 0.8 \cdot 0.8}}$$

# Implementatie [notebook]

---

# Vektoren

---

	Meike	Lars	Rudolf	...
<i>Titanic</i>	0.9	0.5	0.9	
<i>Luca</i>	0.7	0.8	0.7	
<i>The Great Dictator</i>	1.0	0.3	0.8	
...				

$$l = \begin{pmatrix} 0.7 \\ 0.8 \\ 0.7 \end{pmatrix} \quad d = \begin{pmatrix} 1.0 \\ 0.3 \\ 0.8 \end{pmatrix}$$

# Operaties op vectoren

---

Som

	Meike	Lars	Rudolf	...
<i>Titanic</i>	0.9	0.5	0.9	
<i>Luca</i>	0.7	0.8	0.7	
<i>The Great Dictator</i>	1.0	0.3	0.8	
...				

$$l = \begin{pmatrix} 0.7 \\ 0.8 \\ 0.7 \end{pmatrix} \quad d = \begin{pmatrix} 1.0 \\ 0.3 \\ 0.8 \end{pmatrix}$$

$$l + d = ?$$

# Operaties op vectoren

---

Som

	Meike	Lars	Rudolf	...
<i>Titanic</i>	0.9	0.5	0.9	
<i>Luca</i>	0.7	0.8	0.7	
<i>The Great Dictator</i>	1.0	0.3	0.8	
...				

$$l = \begin{pmatrix} 0.7 \\ 0.8 \\ 0.7 \end{pmatrix} \quad d = \begin{pmatrix} 1.0 \\ 0.3 \\ 0.8 \end{pmatrix}$$

$$l + d = \begin{pmatrix} 0.7 + 1.0 \\ 0.8 + 0.3 \\ 0.7 + 0.8 \end{pmatrix} = \begin{pmatrix} 1.7 \\ 1.1 \\ 1.5 \end{pmatrix}$$



# Operaties op vectoren

---

Som

	Meike	Lars	Rudolf	...
<i>Titanic</i>	0.9	0.5	0.9	
<i>Luca</i>	0.7	0.8	0.7	
<i>The Great Dictator</i>	1.0	0.3	0.8	
...				

$$l = \begin{pmatrix} 0.7 \\ 0.8 \\ 0.7 \end{pmatrix} \quad d = \begin{pmatrix} 1.0 \\ 0.3 \\ 0.8 \end{pmatrix}$$

$$l \cdot d = ?$$

# Inproduct

---

Inproduct (*dot product*)

	Meike	Lars	Rudolf	...
<i>Titanic</i>	0.9	0.5	0.9	
<i>Luca</i>	0.7	0.8	0.7	
<i>The Great Dictator</i>	1.0	0.3	0.8	
...				

$$l = \begin{pmatrix} 0.7 \\ 0.8 \\ 0.7 \end{pmatrix} \quad d = \begin{pmatrix} 1.0 \\ 0.3 \\ 0.8 \end{pmatrix}$$

$$l \cdot d = \begin{matrix} 0.7 \cdot 1.0 \\ + \\ 0.8 \cdot 0.3 \\ + \\ 0.7 \cdot 0.8 \end{matrix}$$

# Inproduct

---

Inproduct (*dot product*)

	Meike	Lars	Rudolf	...
<i>Titanic</i>	0.9	0.5	0.9	
<i>Luca</i>	0.7	0.8	0.7	
<i>The Great Dictator</i>	1.0	0.3	0.8	
...				

$$l = \begin{pmatrix} 0.7 \\ 0.8 \\ 0.7 \end{pmatrix} \quad d = \begin{pmatrix} 1.0 \\ 0.3 \\ 0.8 \end{pmatrix}$$

$$l \cdot d = 0.7 \cdot 1.0 + 0.8 \cdot 0.3 + 0.7 \cdot 0.8 = 0.87$$

# Inproduct

---

Inproduct (*dot product*)

...			
$a$	$a_1$	$a_2$	...
$b$	$b_1$	$b_2$	...
...			

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \end{pmatrix} \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \end{pmatrix}$$

$$a \cdot b = a_1 \cdot b_1 + a_2 \cdot b_2 + \dots$$

# Inproduct

---

$$a = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \quad b = \begin{pmatrix} 100 \\ 10 \\ 1 \end{pmatrix}$$

$$a \cdot b = ?$$



# Inproduct

---

$$a = \begin{pmatrix} 2 \\ 2 \\ 3 \\ 5 \end{pmatrix} \quad b = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$


$$a \cdot b = ?$$



# Inproduct in Pandas: @

---

```
1 s1 = pd.Series([1, 2, 3])
2 s2 = pd.Series([100, 10, 1])
3
4 inproduct = s1 @ s2
5 print(inproduct)
```



123

# Inproduct in Pandas

---

```
1 df1 = pd.DataFrame([[1,1,1], [1,2,3],  
2                      [100,10,1], [0,1,1]])  
3 display(df1)
```

	0	1	2
0	1	1	1
1	1	2	3
2	100	10	1
3	0	1	1

```
1 s1 = df1.loc[2]  
2 s2 = df1.loc[3]  
3 inproduct = s1 @ s2  
4 print(inproduct)
```



# Cosine similarity

---

	Meike	Lars	Rudolf	...
<i>Titanic</i>	0.9	0.5	0.9	
<i>Luca</i>	0.7	0.8	0.7	
<i>The Great Dictator</i>	1.0	0.3	0.8	
...				

$$\cos(a, b) = \frac{\sum_{i=1}^n a_i \cdot b_i}{\sqrt{\sum_{i=1}^n a_i^2} \cdot \sqrt{\sum_{i=1}^n b_i^2}}$$

$$\cos(\text{Luca}, \text{The Great Dictator}) = \frac{0.7 \cdot 1.0 + 0.8 \cdot 0.3 + 0.7 \cdot 0.8}{\sqrt{0.7 \cdot 0.7 + 0.8 \cdot 0.8 + 0.7 \cdot 0.7} \cdot \sqrt{1.0 \cdot 1.0 + 0.3 \cdot 0.3 + 0.8 \cdot 0.8}}$$

# Cosine similarity

---

	Meike	Lars	Rudolf
<i>Titanic</i>	0.9	0.5	0.9
<i>Luca</i>	0.7	0.8	0.7
<i>The Great Dictator</i>	1.0	0.3	0.8

$$\cos(l, d) = \frac{\sum_{i=1}^n l_i \cdot d_i}{\sqrt{\sum_{i=1}^n l_i^2} \cdot \sqrt{\sum_{i=1}^n d_i^2}}$$

$$\cos(\text{Luca}, \text{The Great Dictator}) = \frac{0.7 \cdot 1.0 + 0.8 \cdot 0.3 + 0.7 \cdot 0.8}{\sqrt{0.7 \cdot 0.7 + 0.8 \cdot 0.8 + 0.7 \cdot 0.7} \cdot \sqrt{1.0 \cdot 1.0 + 0.3 \cdot 0.3 + 0.8 \cdot 0.8}}$$

$$l = \begin{pmatrix} 0.7 \\ 0.8 \\ 0.7 \end{pmatrix}$$

$$d = \begin{pmatrix} 1.0 \\ 0.3 \\ 0.8 \end{pmatrix}$$

$$\cos(l, d) = ?$$

# Cosine similarity

---

	Meike	Lars	Rudolf
<i>Titanic</i>	0.9	0.5	0.9
<i>Luca</i>	0.7	0.8	0.7
<i>The Great Dictator</i>	1.0	0.3	0.8

$$\cos(l, d) = \frac{\sum_{i=1}^n l_i \cdot d_i}{\sqrt{\sum_{i=1}^n l_i^2} \cdot \sqrt{\sum_{i=1}^n d_i^2}}$$

$$\cos(\text{Luca}, \text{The Great Dictator}) = \frac{0.7 \cdot 1.0 + 0.8 \cdot 0.3 + 0.7 \cdot 0.8}{\sqrt{0.7 \cdot 0.7 + 0.8 \cdot 0.8 + 0.7 \cdot 0.7} \cdot \sqrt{1.0 \cdot 1.0 + 0.3 \cdot 0.3 + 0.8 \cdot 0.8}}$$

$$l = \begin{pmatrix} 0.7 \\ 0.8 \\ 0.7 \end{pmatrix}$$

$$d = \begin{pmatrix} 1.0 \\ 0.3 \\ 0.8 \end{pmatrix}$$

$$\cos(l, d) = \frac{l \cdot d}{\sqrt{l \cdot l} \cdot \sqrt{d \cdot d}}$$



# Deel 2: Vectoren voor content based filtering

---

# CF: Ratings

---

	Anna	Karel	Marie	...
Inception	0.8		0.3	
Frozen	0.6	0.5		
...				



Vectoren → Similarities → Voorspelde ratings (knn) → Aanbevelingen

# CBF: Genres

---

Inception	Action	Adventure	Sci-Fi	Thriller
Frozen	Adventure	Comedy	Fantasy	Musical
Blade Runner	Action	Thriller	Darma	Sci-Fi
...				



Vectoren → Similarities → Voorspelde ratings (knn) → Aanbevelingen

# CBF: Genres

---


Inception	Action	Adventure	Sci-Fi	Thriller
Frozen	Adventure	Comedy	Fantasy	Musical
Blade Runner	Action	Thriller	Darma	Sci-Fi
...				





# CBF: Genres (one-hot encoding)

1	Action	Adventure	Comedy	Drama	Fantasy	Musical	Sci-Fi	Thriller
Inception	1	1	0	0	0	0	1	1
Frozen	0	1	1	0	1	1	0	0
Blade Runner	1	0	0	1	0	0	1	1

  
**Vectoren**  
one-hot enc

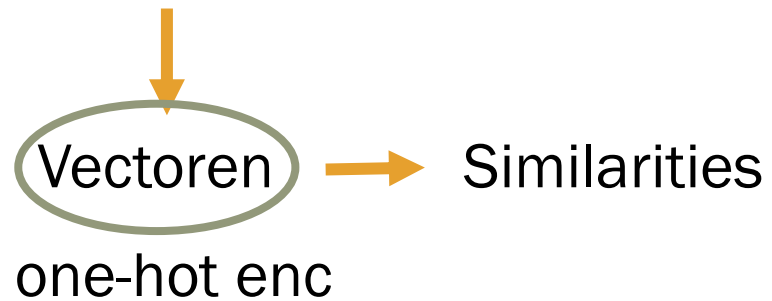
$$I = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

$$F = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

# CBF: Genres (one-hot encoding)

1	Action	Adventure	Comedy	Drama	Fantasy	Musical	Sci-Fi	Thriller
Inception	1	1	0	0	0	0	1	1
Frozen	0	1	1	0	1	1	0	0
Blade Runner	1	0	0	1	0	0	1	1



$$\cos(a, b) = \frac{a \cdot b}{\sqrt{a \cdot a} \cdot \sqrt{b \cdot b}}$$

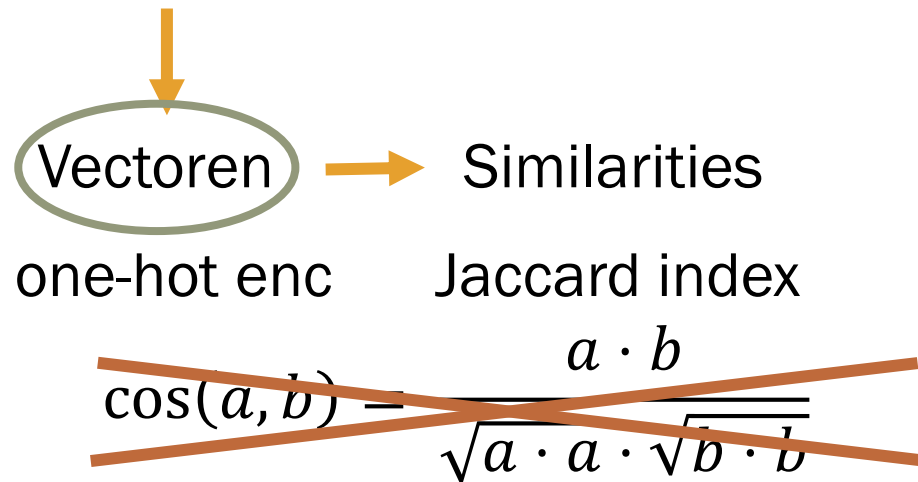
$$I = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

$$F = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

# CBF: Genres (one-hot encoding)

1	Action	Adventure	Comedy	Drama	Fantasy	Musical	Sci-Fi	Thriller
Inception	1	1	0	0	0	0	1	1
Frozen	0	1	1	0	1	1	0	0
Blade Runner	1	0	0	1	0	0	1	1



$$I = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

$$F = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

$$B = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

# CBF: Genres

1	Action	Adventure	Comedy	Drama	Fantasy	Musical	Sci-Fi	Thriller
Inception	1	1	0	0	0	0	1	1
Frozen	0	1	1	0	1	1	0	0
Blade Runner	1	0	0	1	0	0	1	1



# CBF: Text

## Webshop

- beschrijving product
- naam product
- reviews

## Films

- synopsis
- ondertitels
- Reviews

## Boeken

- Inhoud
- ...



### Quick Mill 810 en 820 Rood

€ 669,00

Deze Quick Mill espresso machine is meer dan 35 jaar geproduceerd. Quick Mill geheim schuilt in de Dé oplossing tegen kalkaanslag. De koffie temperatuur, dus bij elke (espresso)koffie genieten. Het is want binnen 10 seconden een cappuccino en latte machi-

## AVATAR: THE WAY OF WATER REVIEWS

All Critics Top Critics All Audience Verified Audience



**Matt Brunson**  
Film Frenzy



I'll say this for James Cameron: At this point, he can slap his name on an old print of Plan 9 From Outer Space, re-release it as Avatar 3: The Way of Outer Space, and incessantly hype it until it crosses the billion-dollar mark and racks up the awards.

[Full Review](#) | Original Score: 2/4 | Apr 18, 2023

stant example where the visual spectacle (it can dazzle) swallows up an unexceptional story.



## Alice's Adventures in Wonderland

Lewis Carroll, Chris Riddell



Chris Riddell's brilliant new sumptuous hardback and j much-loved favourite class

# CBF: Text

---

Inception	Cobb and Arthur are "extractors"; they perform corporate espionage using experimental dream-sharing technology to infiltrate their targets' subconscious and extract information...
Frozen	Princess Elsa of Arendelle possesses magical powers allowing her to control ice and snow, often using them to play with her younger sister Anna....
...	

 Vectoren → Similarities → Voorspelde ratings (knn) → Aanbevelingen

# Vectorisatie text

---

Hoe kunnen we teksten omzetten in een lijst met getallen?

	brutus	de	fietser	hond	kat	over	steekt	straat
<i>De hond steekt de straat over</i>	0	2	0	1	0	1	1	1
<i>De fietser steekt de straat over</i>	0	2	1	0	0	1	1	1
<i>De kat steekt de straat over</i>	0	2	0	0	1	1	1	1
<i>Brutus steekt de straat over</i>	1	1	0	0	0	1	1	1

# Vectorisatie text

---

Probleem 1. Langere teksten -> hogere scores (grotere kans ergens op te lijken).

	brutus	de	fietser	hond	kat	over	steekt	straat
<i>De hond steekt de straat over naar de fietser met de kat</i>	0	3	1	1	1	1	1	1
<i>De hond steekt de straat over</i>	0	2	1	0	0	1	1	1
....								



# Vectorisatie text

---

Oplossing 1. Term Frequency (TF):  $\# \text{voorkomens} / \# \text{lengte tekst}$

	brutus	de	fietser	hond	kat	over	steekt	straat
<i>De hond steekt de straat over naar de fietser met de kat</i>	0	3/12	1/12	1/12	1/12	1/12	1/12	1/12
<i>De hond steekt de straat over</i>	0	2/6	1/6	0	0	1/6	1/6	1/6
....								

# Vectorisatie text

---

Oplossing 1. Term Frequency (TF):  $\frac{\text{\#voorkomens}}{\text{\#lengte tekst}}$

	brutus	de	fietser	hond	kat	over	steekt	straat
<i>De hond steekt de straat over naar de fietser met de kat</i>	0	0.25	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
<i>De hond steekt de straat over</i>	0	0.333	0.1667	0	0	0.1667	0.1667	0.1667
....								

# Vectorisatie text

---

Oplossing 1. Term Frequency (TF):  $\frac{\text{\#voorkomens}}{\text{\#lengte tekst}}$

	brutus	de	fietser	hond	kat	over	steekt	straat
<i>De hond steekt de straat over</i>	0	0.333	0	0.167	0	0.167	0.167	0.167
<i>De fietser steekt de straat over</i>	0	0.333	0.167	0	0	0.167	0.167	0.167
<i>De kat steekt de straat over</i>	0	0.333	0	0	0.167	0.167	0.167	0.167
<i>Brutus steekt de straat over</i>	0.2	0.4	0	0	0	0.2	0.2	0.2

# Vectorisatie text

---

Probleem 2. Veelvoorkomende woorden scoren hoog

	brutus	de	fietser	hond	kat	over	steekt	straat
<i>De hond steekt de straat over</i>	0	0.333	0	0.167	0	0.167	0.167	0.167
<i>De fietser steekt de straat over</i>	0	0.333	0.167	0	0	0.167	0.167	0.167
<i>De kat steekt de straat over</i>	0	0.333	0	0	0.167	0.167	0.167	0.167
<i>Brutus steekt de straat over</i>	0.2	0.4	0	0	0	0.2	0.2	0.2



# Vectorisatie text

---

Oplossing 2. Inverse Document Frequency (IDF):

$\ln(\text{\#aantal documenten} / \text{\#aantal documenten met het betreffende word})$

brutus	de	fietser	hond	kat	over	steekt	straat
$\ln(4/1)$	$\ln(4/4)$	$\ln(4/1)$	$\ln(4/1)$	$\ln(4/1)$	$\ln(4/4)$	$\ln(4/4)$	$\ln(4/4)$

# Vectorisatie text

---

Oplossing 2. Inverse Document Frequency (IDF):

$\ln(\text{\#aantal documenten} / \text{\#aantal documenten met het betreffende word})$

brutus	de	fietser	hond	kat	over	steekt	straat
1,386	0	1,386	1,386	1,386	0	0	0

# Vectorisatie text

---

Oplossing 2. Term Frequency (TF-IDF):  $TF * IDF$

TF					
fietser	hond	kat	over	steekt	straat
0	0.167	0	0.167	0.167	0.167
0.167	0	0	0.167	0.167	0.167
0	0	0.167	0.167	0.167	0.167
0	0	0	0.2	0.2	0.2

IDF					
brutus	de	fietser	hond	kat	o
1,386	0	1,386	1,386	1,386	0

# Vectorisatie text

---

Oplossing 2. Term Frequency (TF-IDF):  $TF * IDF$

	brutus	de	fietser	hond	kat	over	steekt	straat
<i>De hond steekt de straat over</i>	$0 * 1,39$	$0.33 * 0$	$0 * 1,39$	$0.17 * 1,39$	$0 * 1,39$	$0.17 * 0$	$0.17 * 0$	$0.17 * 0$
<i>De fietser steekt de straat over</i>	$0 * 1.39$	$0.33 * 0$	$0.17 * 1,39$	$0 * 1,39$	$0 * 1,39$	$0.17 * 0$	$0.17 * 0$	$0.17 * 0$
<i>De kat steekt de straat over</i>	$0 * 1.39$	$0.33 * 0$	$0 * 1,39$	$0 * 1,39$	$0.17 * 1,39$	$0.17 * 0$	$0.17 * 0$	$0.17 * 0$
<i>Brutus steekt de straat over</i>	$0.2 * 1,39$	$0.4 * 0$	$0 * 1,39$	$0 * 1,39$	$0 * 1,39$	$0.2 * 0$	$0.2 * 0$	$0.2 * 0$



# Vectorisatie text

---

Probleem 2. Veelvoorkomende woorden scoren hoog

	brutus	de	fietser	hond	kat	over	steekt	straat
<i>De hond steekt de straat over</i>	0	0	0	0.236	0	0	0	0
<i>De fietser steekt de straat over</i>	0	0	0.236	0	0	0	0	0
<i>De kat steekt de straat over</i>	0	0	0	0	0.236	0	0	0
<i>Brutus steekt de straat over</i>	0.278	0	0	0	0	0	0	0



# Vectorisatie text

---

## Probleem 3. Geen semantiek:

- Een kat is meer als een hond dan een fiets?
- Brutus kan de naam van een hond zijn?
- Een bank (voor geld) is hetzelfde als een bank (voor zitten).

## Probleem 4. Geen gramatica:

- Woordvolgorde wordt genegeerd.  
“Meike eet een taart” is hetzelfde als “Een taart eet Meike” (surrealistische horrorfilm?)

# Vectorisatie text

---

Geen TF-IDF gebruiken maar bijvoorbeeld word2vec

# Vragen?

---