Søketek uke 7

Gruppe 1 & 2

Science fair

- Frist for å lage gruppe: 21.10
- Tilfeldige grupper for de som ikke sender inn
- Kan også være 1 eller 3
- For alle som stresser: Jeg går gjennom min egen science fair 23.10

Hei!

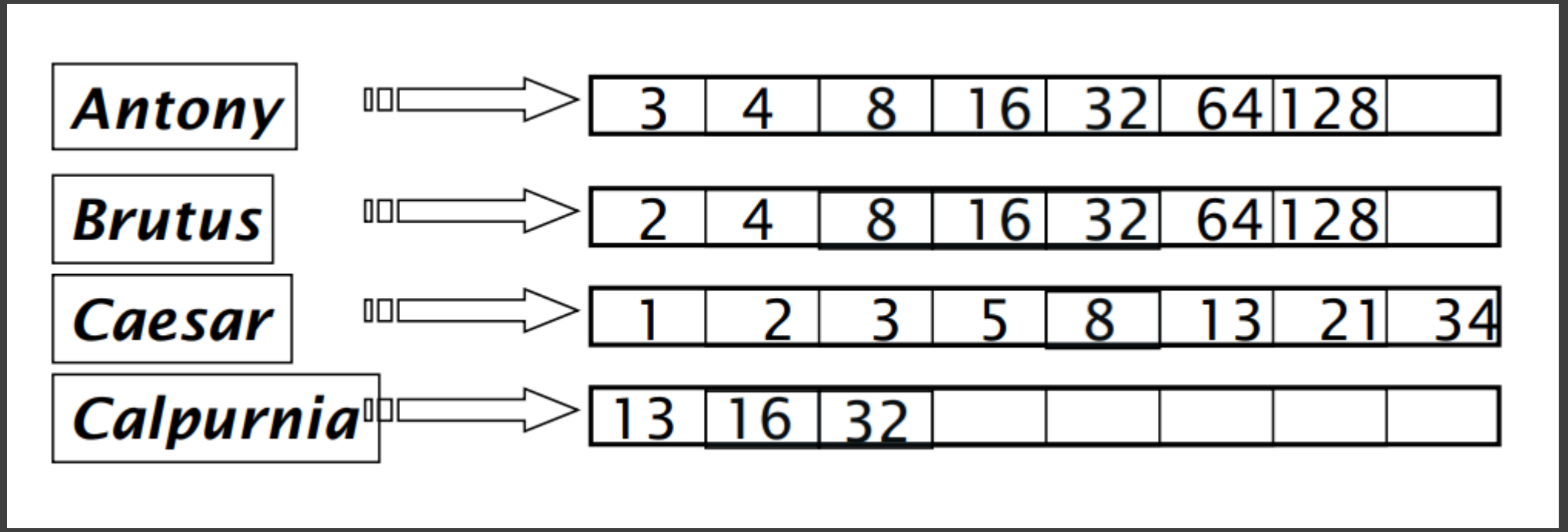
Jeg og Torkild Engen Finne (torkilef) har tenkt å jobbe sammen med Science Fair.

Mvh
Truls Hestetræet

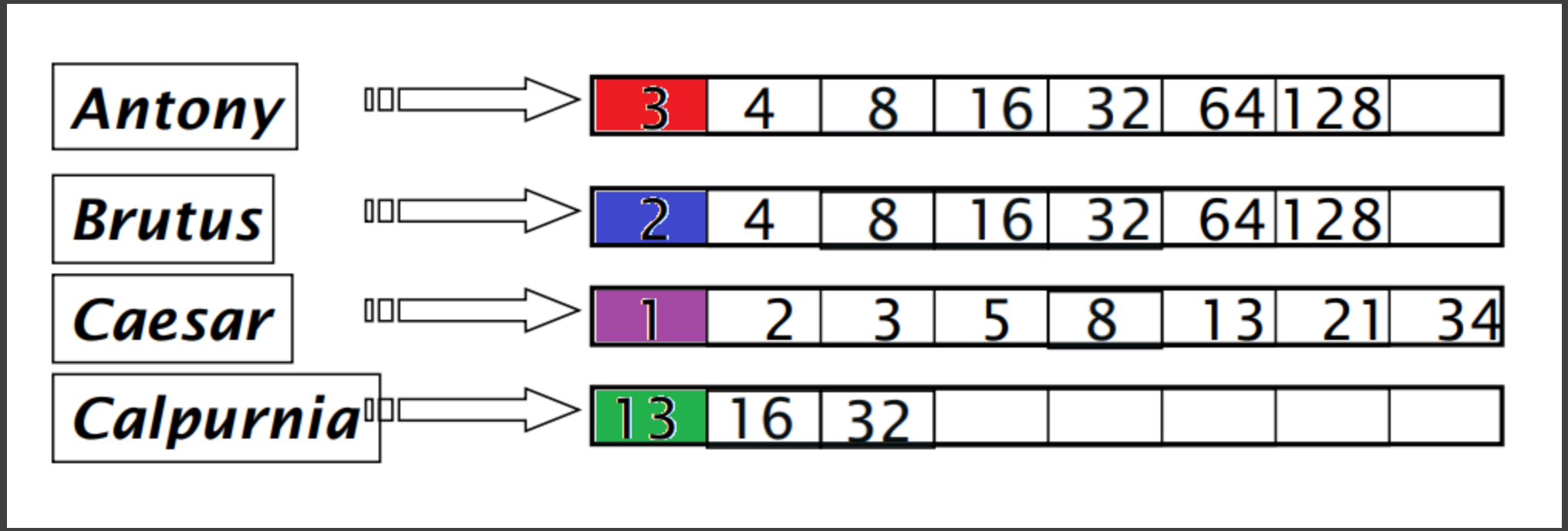
Agenda

- Eksempel på n-of-m-matching i oblig c-1
- Repetisjon
  - Precision & recall
  - Precision@k
  - Mean Average Precision (MAP)
  - Kendall Tau distance
  - Normalized Discounted Cumulative Gain (NDCG)
- Ukas shoutout
- Selvstendig arbeid

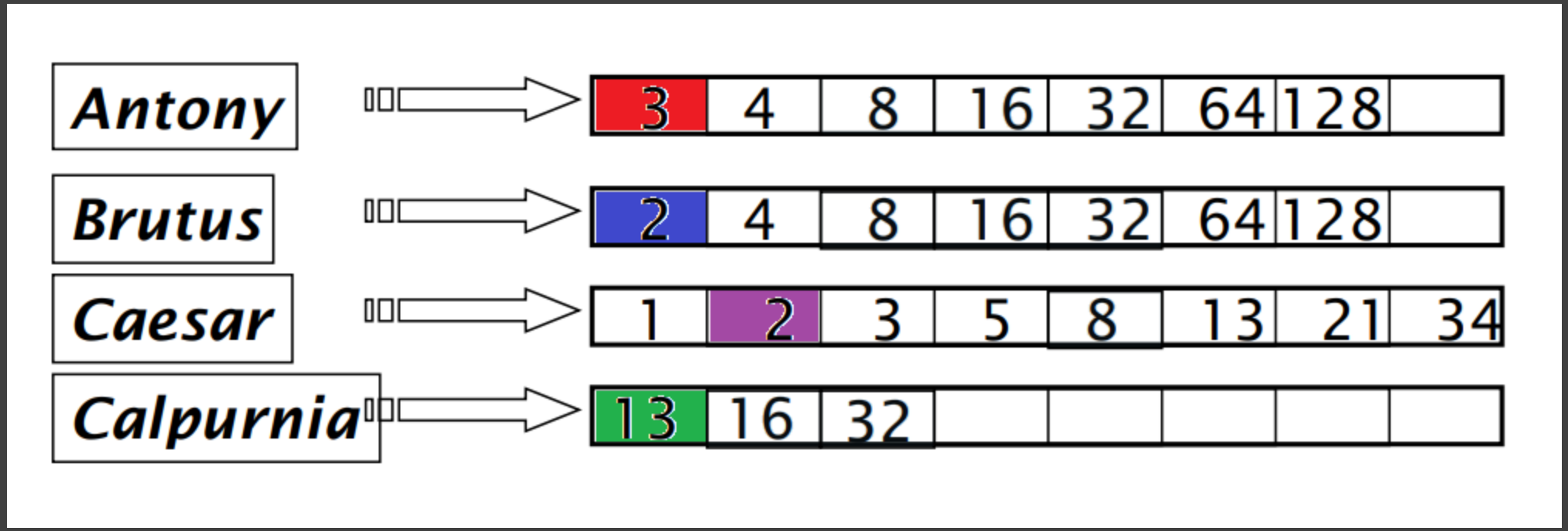
Eksempel-postinglister



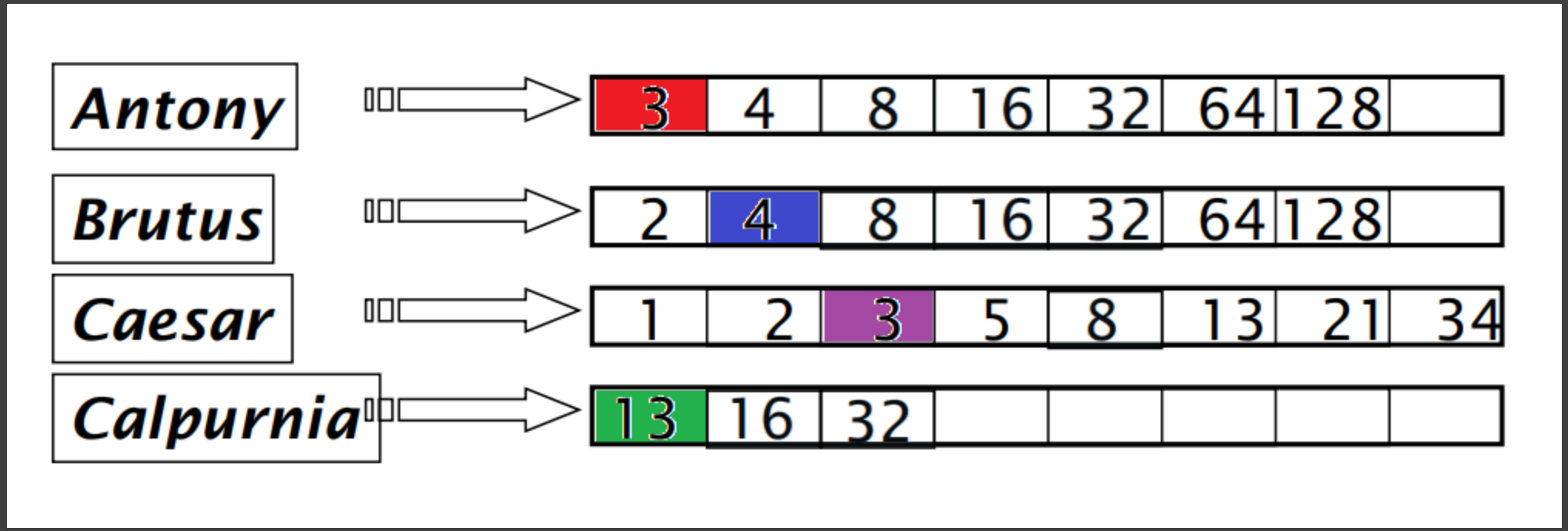
FRONTIER: 1


Antony
3 4 8 16 32 64 128
Brutus
2 4 8 16 32 64 128
Caesar
1 2 3 5 8 13 21 34
Calpurnia
13 16 32

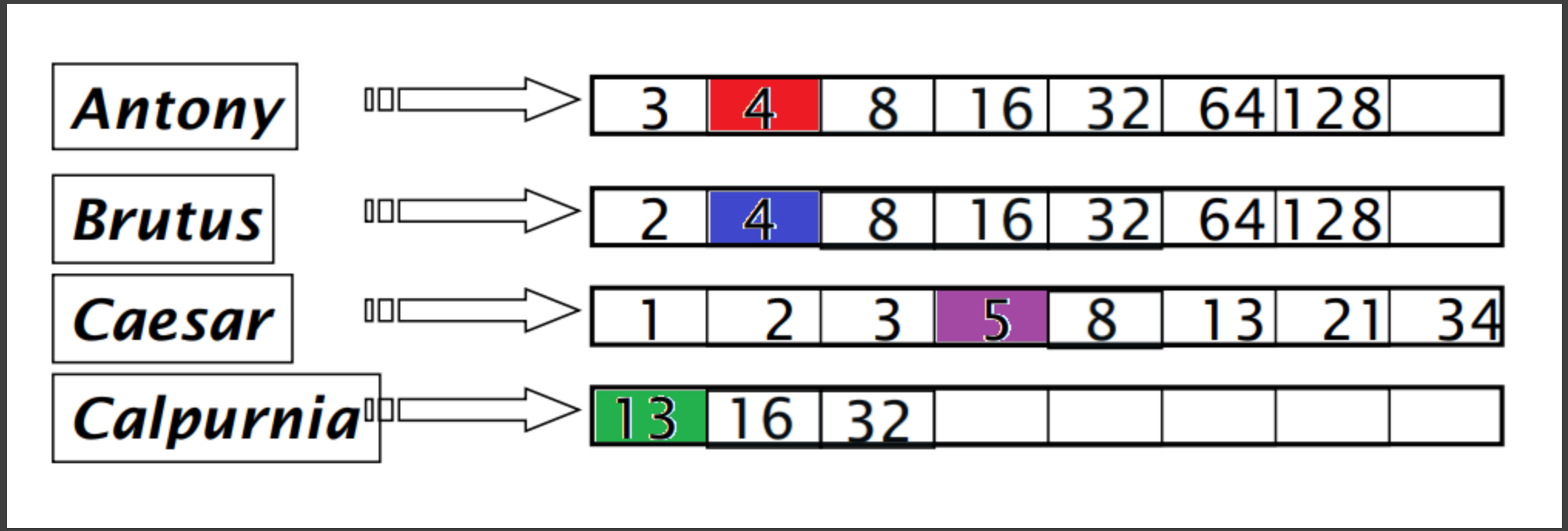
FRONTIER: 2


Antony
3 4 8 16 32 64 128
Brutus
2 4 8 16 32 64 128
Caesar
1 2 3 5 8 13 21 34
Calpurnia
13 16 32

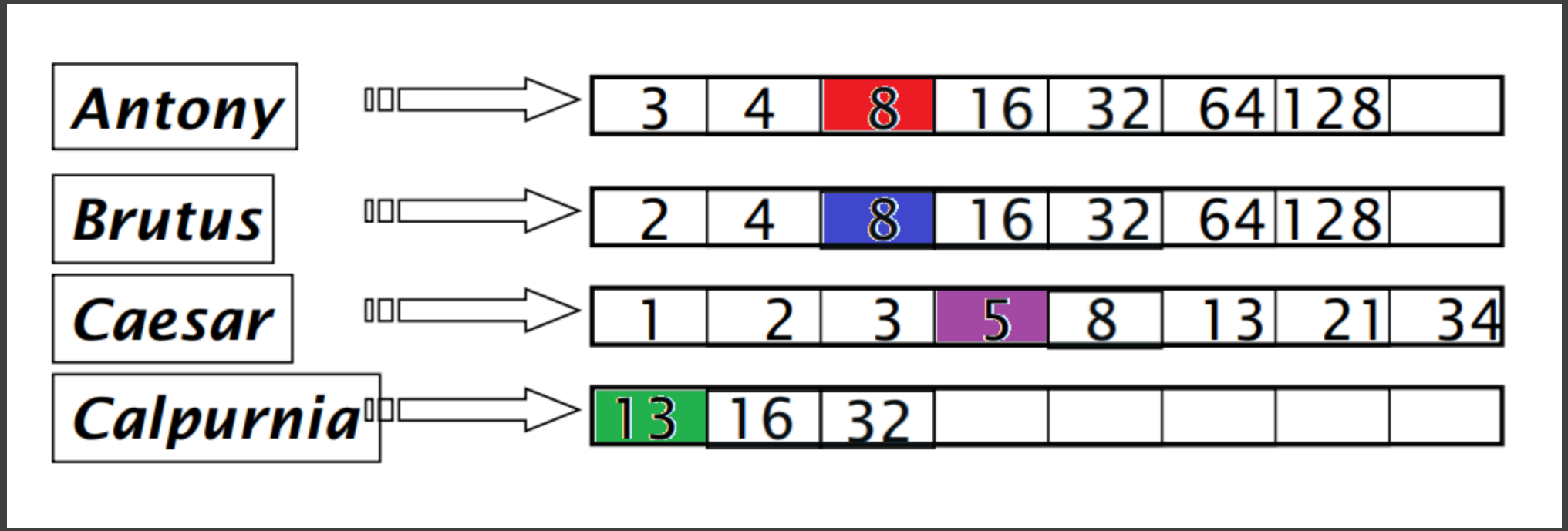
FRONTIER: 2


Antony
3 4 8 16 32 64 128
Brutus
2 4 8 16 32 64 128
Caesar
1 2 3 5 8 13 21 34
Calpurnia
13 16 32

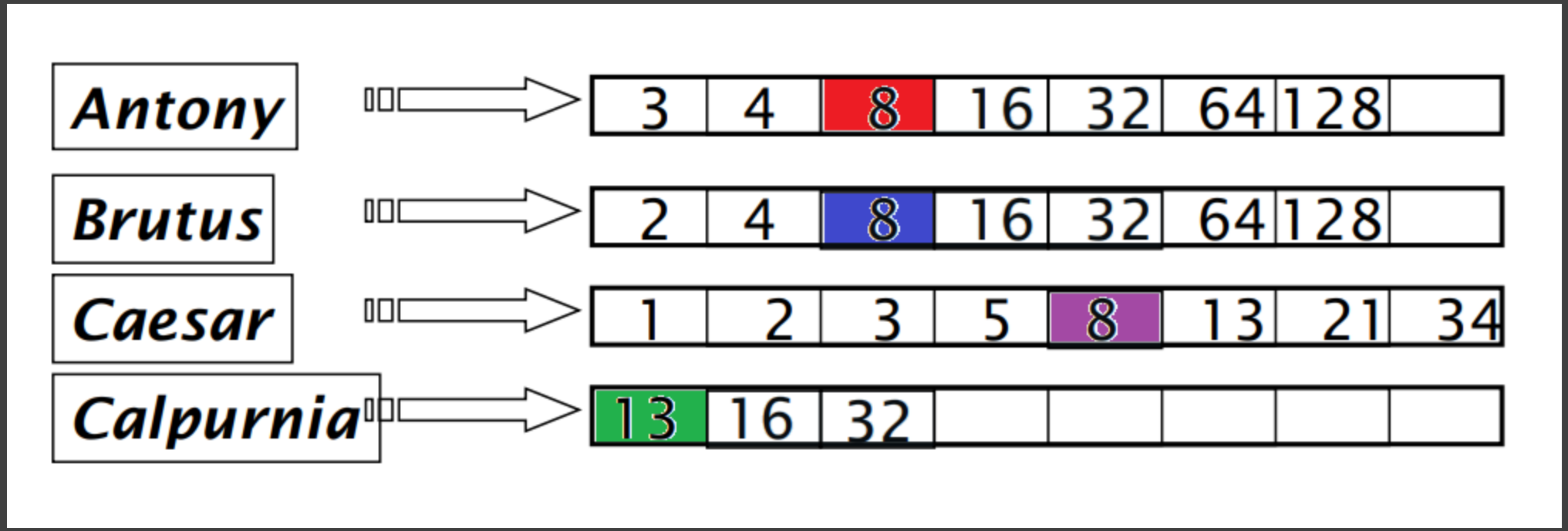
FRONTIER: 2


Antony
3 4 8 16 32 64 128
Brutus
2 4 8 16 32 64 128
Caesar
1 2 3 5 8 13 21 34
Calpurnia
13 16 32

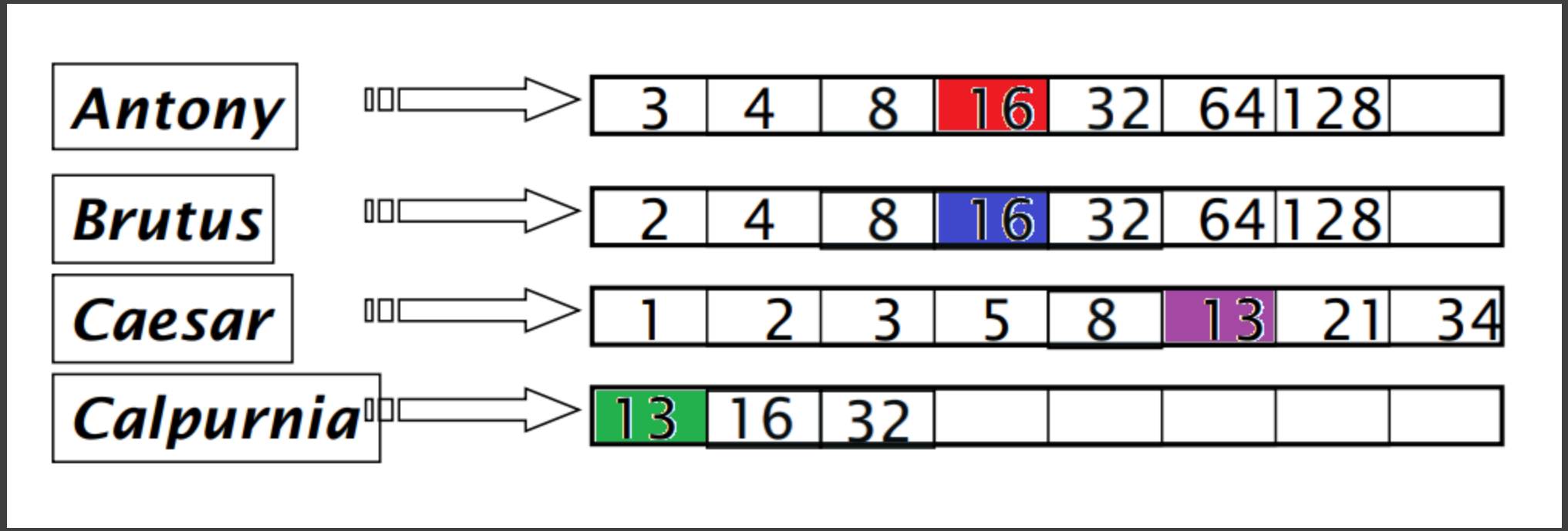
FRONTIER: 1


Antony
3 4 8 16 32 64 128
Brutus
2 4 8 16 32 64 128
Caesar
1 2 3 5 8 13 21 34
Calpurnia
13 16 32

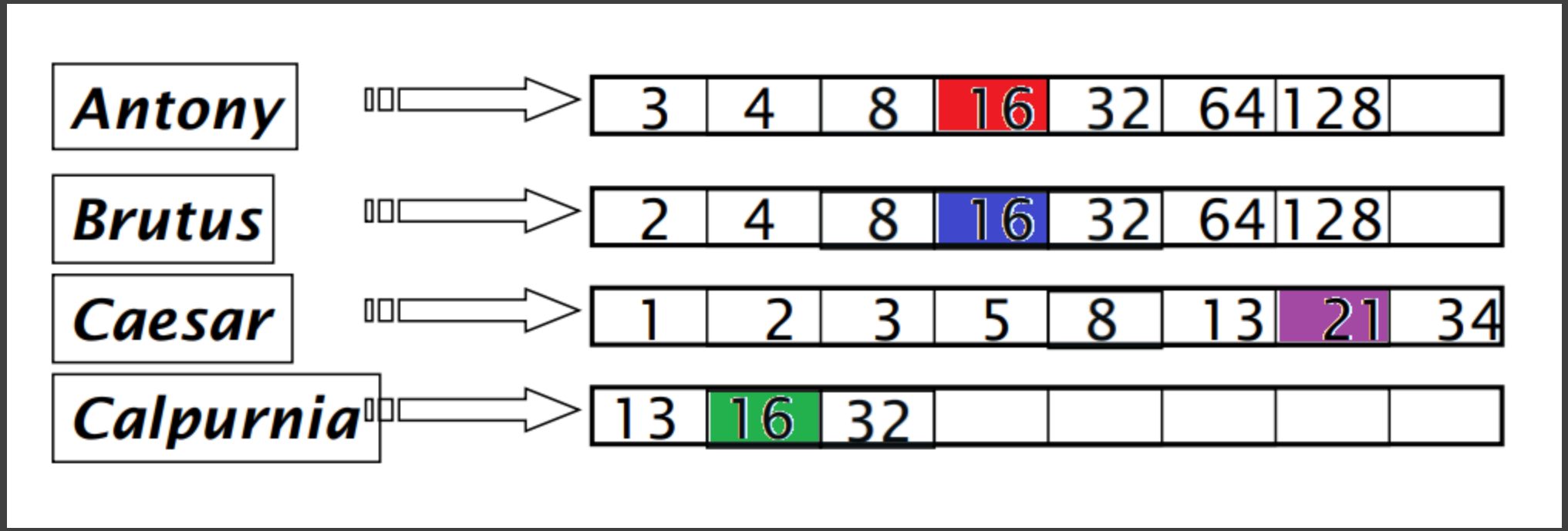
FRONTIER: 3


Antony
3 4 8 16 32 64 128
Brutus
2 4 8 16 32 64 128
Caesar
1 2 3 5 8 13 21 34
Calpurnia
13 16 32

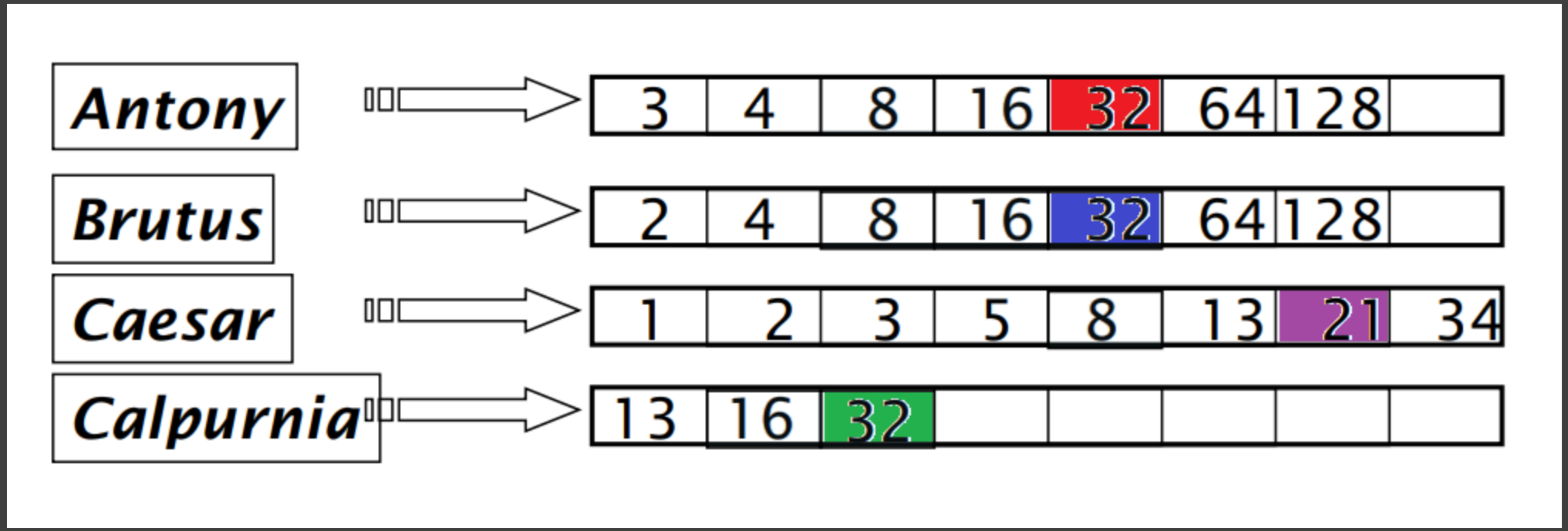
FRONTIER: 2


Antony
3 4 8 16 32 64 128
Brutus
2 4 8 16 32 64 128
Caesar
1 2 3 5 8 13 21 34
Calpurnia
13 16 32

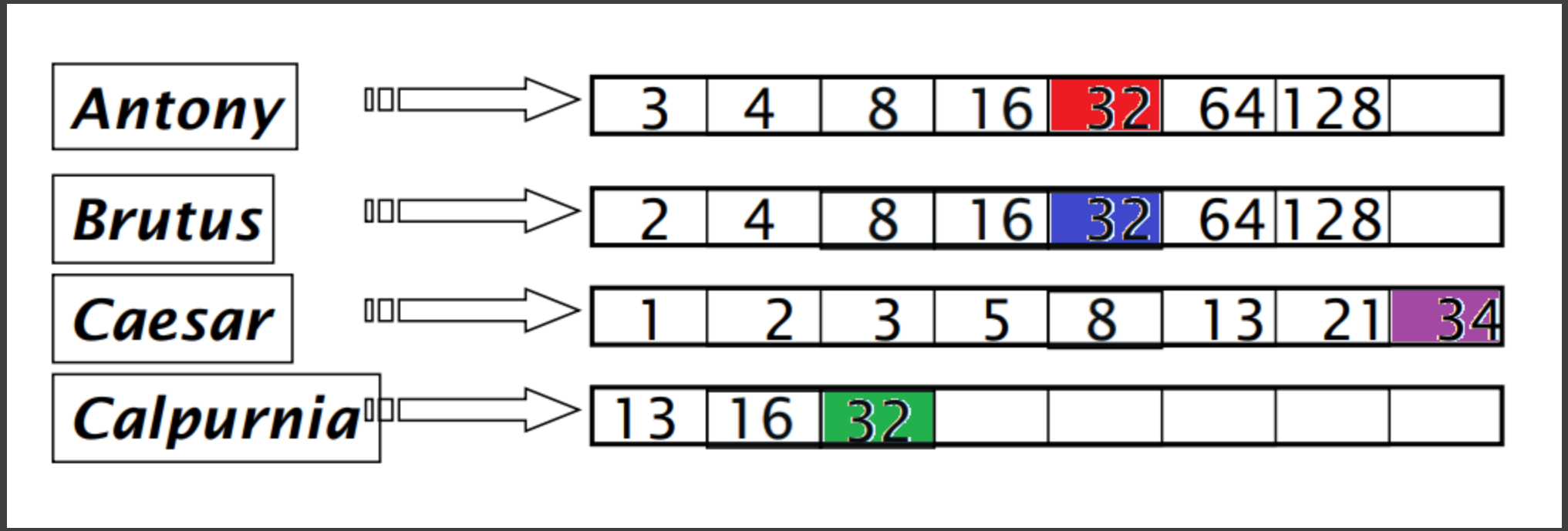
FRONTIER: 3


Antony
3 4 8 16 32 64 128
Brutus
2 4 8 16 32 64 128
Caesar
1 2 3 5 8 13 21 34
Calpurnia
13 16 32

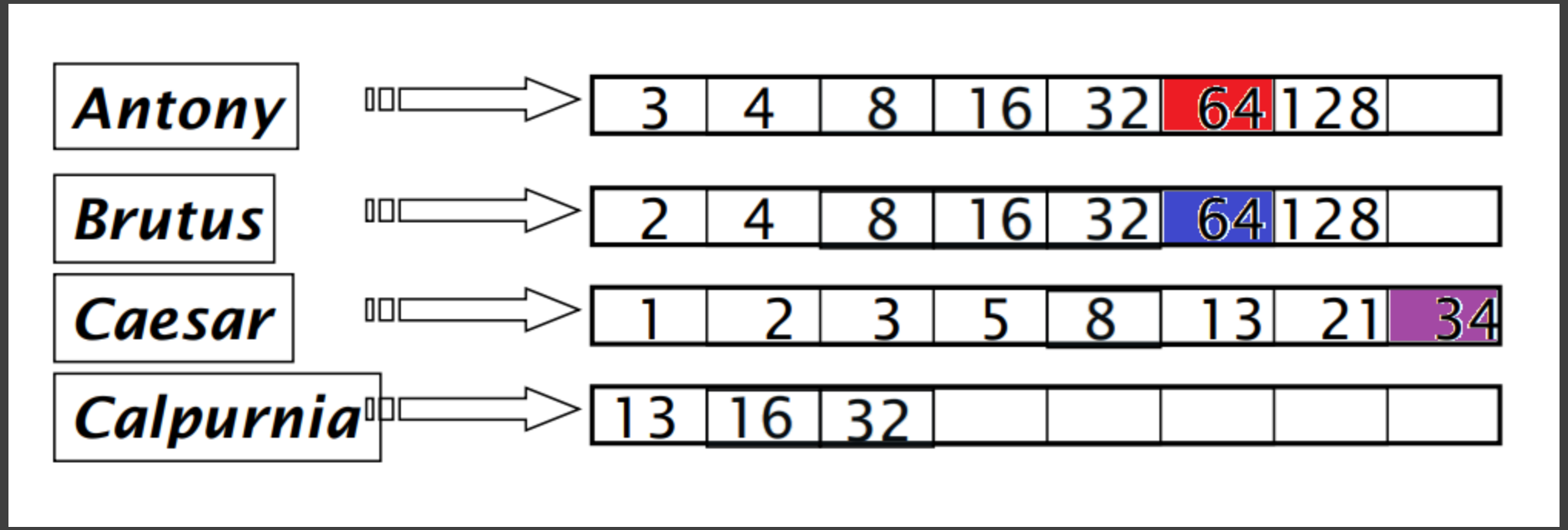
FRONTIER: 1


Antony
3 4 8 16 32 64 128
Brutus
2 4 8 16 32 64 128
Caesar
1 2 3 5 8 13 21 34
Calpurnia
13 16 32

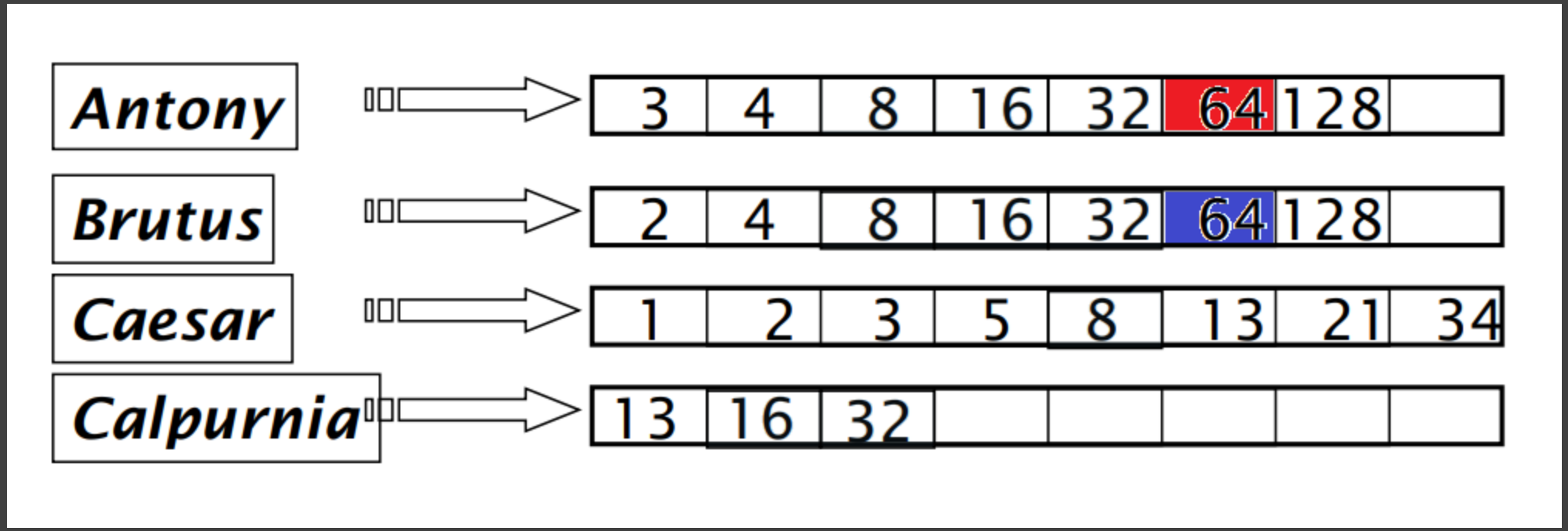
FRONTIER: 3


Antony
3 4 8 16 32 64 128
Brutus
2 4 8 16 32 64 128
Caesar
1 2 3 5 8 13 21 34
Calpurnia
13 16 32

FRONTIER: 1


Antony
3 4 8 16 32 64 128
Brutus
2 4 8 16 32 64 128
Caesar
1 2 3 5 8 13 21 34
Calpurnia
13 16 32

LEN(REMAINING_CURSORS) < N



Agenda

- Eksempel på n-of-m-matching i oblig c-1
- Repetisjon
  - Precision & recall
  - Precision@k
  - Mean Average Precision (MAP)
  - Kendall Tau distance
  - Normalized Discounted Cumulative Gain (NDCG)
- Ukas shoutout
- Selvstendig arbeid

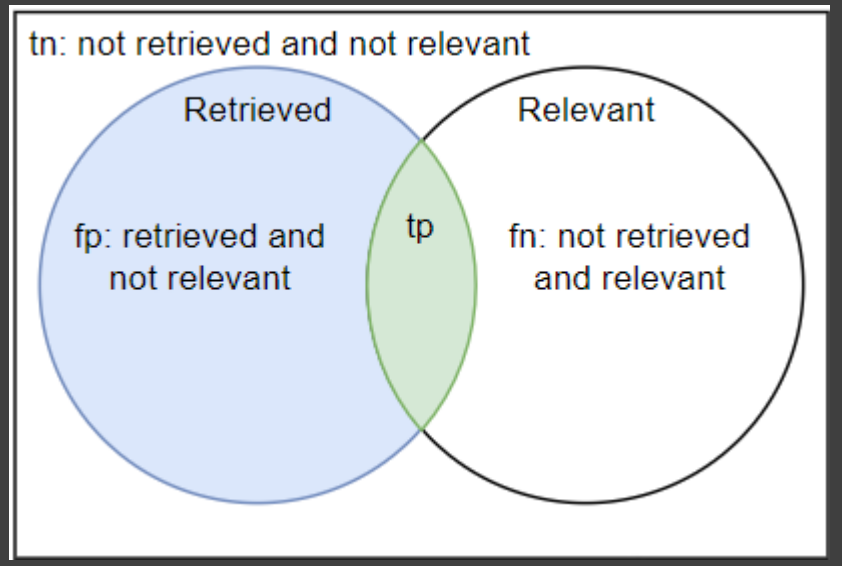
Hva skal vi egentlig gjøre?

- Vi skal måle relevans (retrieval evaluation)
  - Hvor bra er de dokumentene vi hentet?
  - Hvor høy er «kvaliteten» på søkemotoren vår?
- Hvor bra stemmer dokumentene med *behovet* i queryen?
- Denne timen: Måter å måle relevans for query-resultater

Precision

- Av alle dokumenter jeg hentet, hvor mange er relevante?
- Viktig når vi trenger noe relevant, men ikke alt som er relevant
  - Youtube-videoer: Trenger ikke se *alle* relevante videoer, men de jeg ser må være relevante

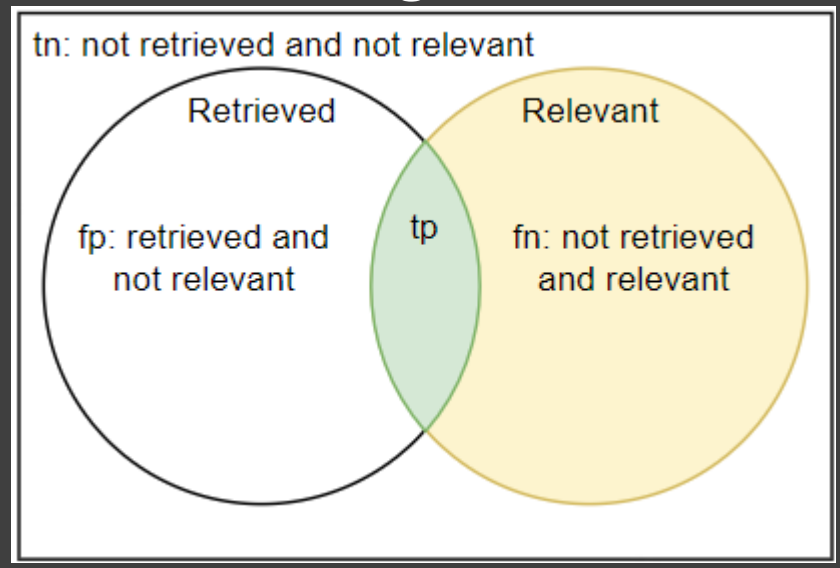
$$\frac{tp}{tp + fp}$$



Recall

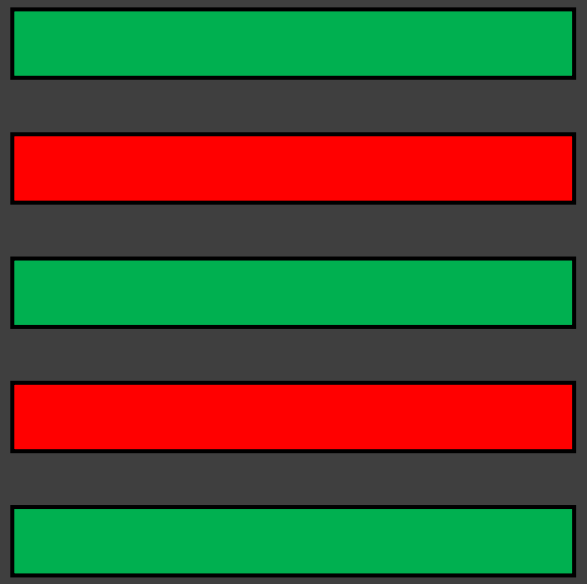
- Av alle relevante dokumenter, hvor mange hentet jeg?
- Dette brukes når vi trenger alle relevante dokumenter, selv om noen irrelevante kanskje også blir med
  - Research til en PhD: Vi vil lese alle relevante dokumenter og kan filtrere bort de som ikke er relevante selv

$$\frac{tp}{tp + fn}$$



Precision@k

- Av de k høyest rangerte dokumentene, hvor mange er relevante?



1/1

1/2

2/3

2/4

3/5

Mean average precision

- Litt vanskelig navn (average average precision)

1. **Precision**: precision@k på relevante dokumenter fram til k
2. **Average**: gjennomsnitt av (1)
3. **Mean**: gjennomsnittet av (2) for flere queries

Average precision

Average precision

Average precision

1/1 = 1

2/3 = 0.67

3/5 = 0.6

Average precision

1.0 + 0.67 + 0.6 = 2.27

2.27 / 3 = 0.76

Mean Average Precision

Query 1	(1.0 + 0.67 + 0.6) / 3 = 0.76
Query 2	(1.0 + 0.67 + 0.75 + 0.8) / 4 = 0.80
MAP	(0.76 + 0.80) / 2 = 0.78

Kendall Tau

- Et dokument er enten mer eller mindre relevant enn andre dokumenter
- Hvor enig er vår søkemotor med denne rangeringen?

Kendall Tau

- Parvis rangering: {(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)}
- 1 > 2 2 > 3 3 > 4
- 1 > 3 2 > 4
- 1 > 4

Kendall Tau

- X: Antall enigheter (agreements)
- Y: Antall uenigheter

$$KT = \frac{X - Y}{X + Y}$$

Eksempler

- [1, 3, 2, 4]: (5 – 1) / (5 + 1) = 0,67
- [4, 3, 2, 1]: (0 – 6) / (0 + 6) = -1.0
- [1, 4, 3, 2]: (3 – 3) / (3 + 3) = 0

Discounted cumulative gain

- En score for hvor relevante resultatene våre er
- Gain: relevanse-score for hvert dokument (f.eks: 0, 1, 2)
  - 0 = helt irrelevant, 2 = veldig relevant
- Vi må ta høyde for at lavere rangerte dokumenter mindre sannsynlig blir valgt. Vi bryr oss mest om de høyest rangerte dokumentene

Eksempel

- Vi har gain: [0, 1, 2] og ett dokument i hver klasse
- Ideell rekkefølge er 2, 1, 0
- Hvert dokument får tildelt en score/gain

Discounted cumulative gain

- Formelen fra forelesning: summen av gain delt på **$\log_2$(i)** fra og med i = 2
- For vårt eksempel: **2**
  + 0 / $\log_2$(2)
  + 1 / $\log_2$(3)

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

Discounted cumulative gain

- 2 + 0 / $\log_2(2)$ + 1 / $\log_2(3)$
- 2 + 0 + 0,63
- 2,63

Normalisert DCG

- Hvordan ser vår DCG ut relativt til beste mulige?
- Høyeste mulige DCG: $2 + \frac{1}{\log2(2)} + \frac{0}{\log2(3)} = 2 + 1 + 0 = 3$
- Vår DCG: $2 + \frac{0}{\log2(2)} + \frac{1}{\log2(3)} = 2 + 0 + 0{,}63 = 2{,}63$
- NDCG: $\frac{2{,}63}{3} = 0{,}876$

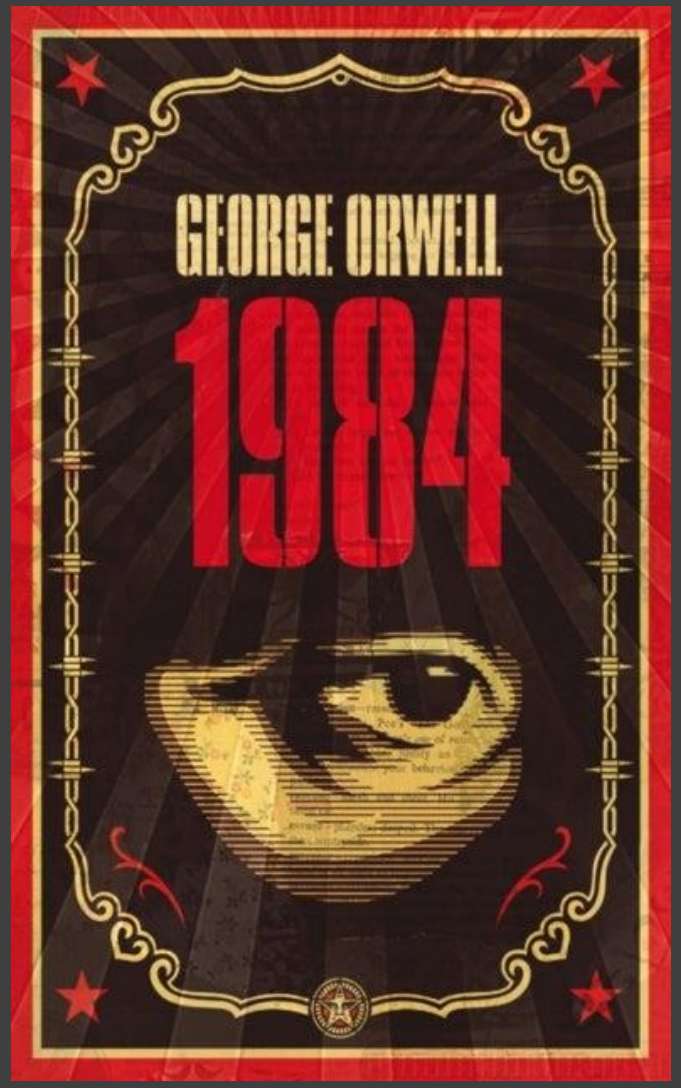
Begrepene forklart

- **GAIN** Score tildelt hvert dokument
  - F.eks.: 0, 1, 2
- **CUMULATIVE GAIN:** summen av gain
  - Vårt eksempel: 2 + 0 + 1 = 3
- **DISCOUNTED CUMULATIVE GAIN:** Legge til en økende straff jo lavere dokumentet blir rangert
  - Intuisjonen bak: Brukeren ser bare de første dokumentene
- **NORMALIZED DISCOUNTED CUMULATIVE GAIN**: DCG delt på beste mulige DCG-verdi. Normaliserer til mellom 0 og 1

Agenda

- Eksempel på n-of-m-matching i oblig c-1
- Repetisjon
  - Precision & recall
  - Precision@k
  - Mean Average Precision (MAP)
  - Kendall Tau distance
  - Normalized Discounted Cumulative Gain (NDCG)
- Ukas shoutout
- Selvstendig arbeid

Ukas shoutout: 1984



15 min pause

Selvstendig jobbing resten av tiden