IN3120 week 13

Group 1

Teaching assistants next year

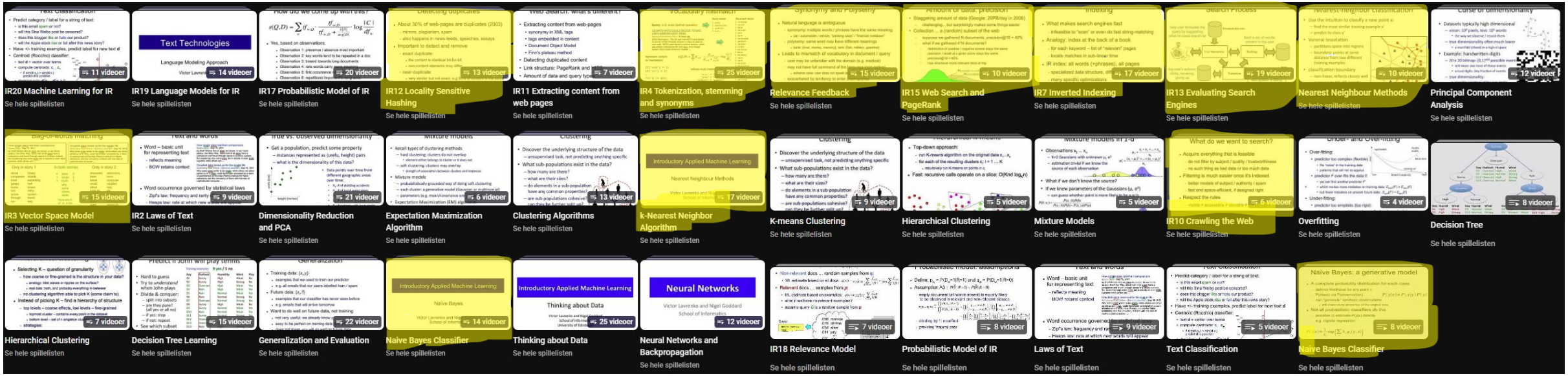
Apply, it’s fun!

Agenda

- Important topics
  - TAAT vs. DAAT
  - BSBI & SPIMI
- Double shoutout!!

Victor Lavrenko – The StatQuest of IR

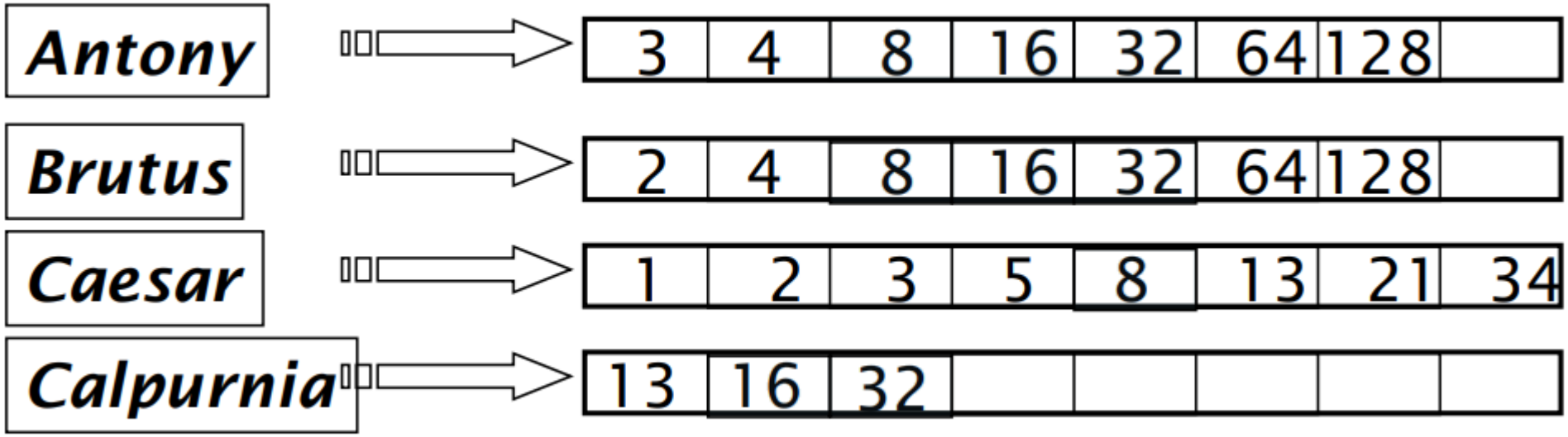
- Inverted index playlist (incl. DAAT, TAAT): https://www.youtube.com/playlist?list=PLBv09BD7ez_448q9kRfZRxYb3cbeEanRb



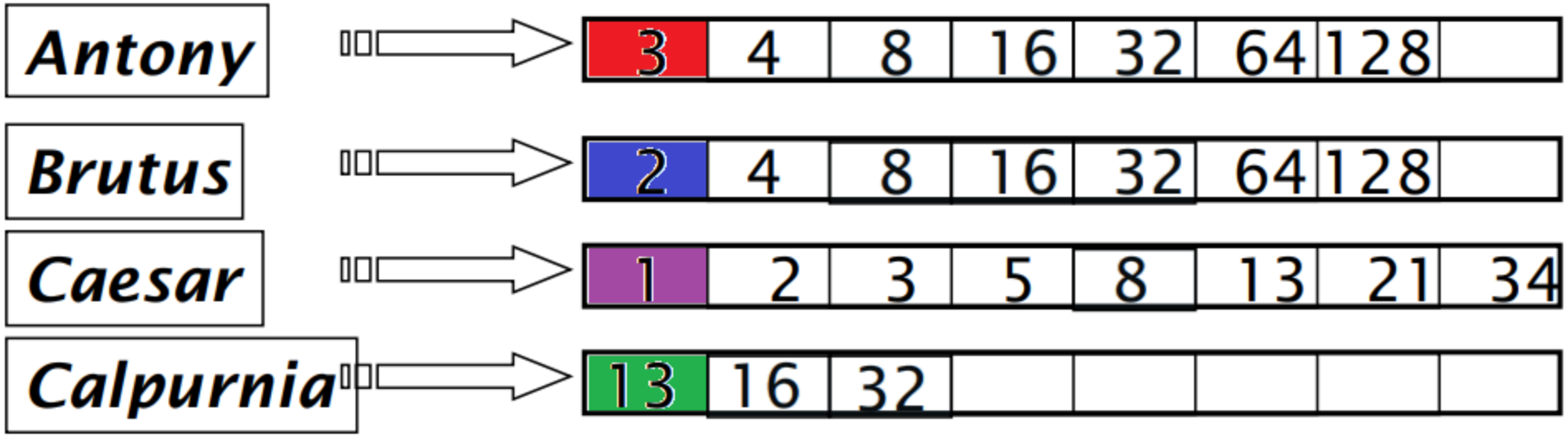
Document-at-a-time

- What we did in assignment C-1
- Compute the score for a document, then move on to new document
  - *But we keep pointers on many docs at the same time?*
- Documents with score 0 is not a part of the result set
  - *Why?*

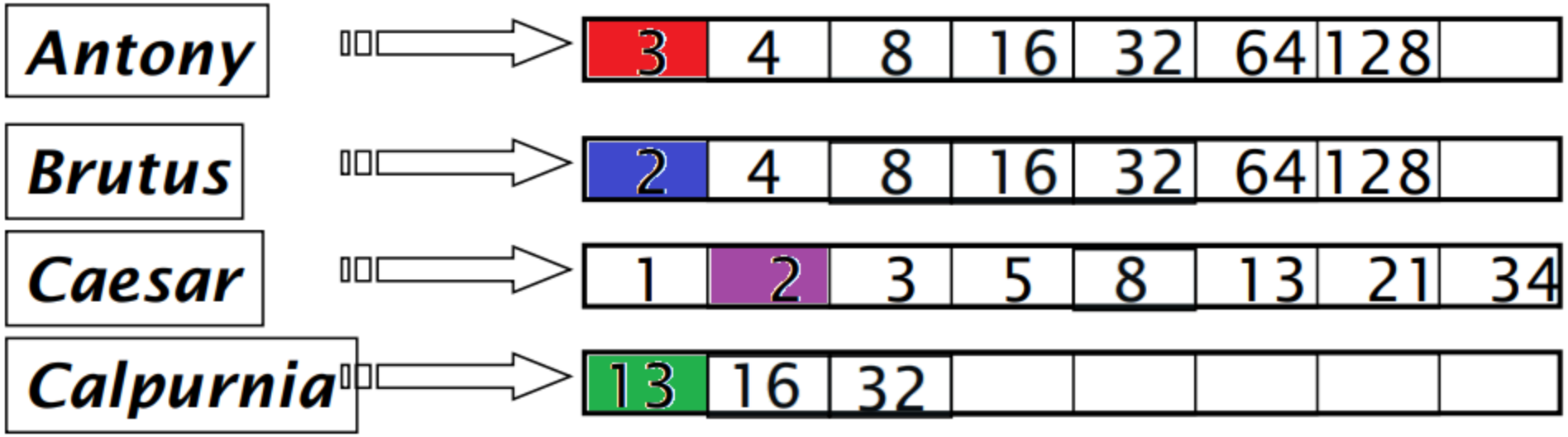
Eksempel-postinglister



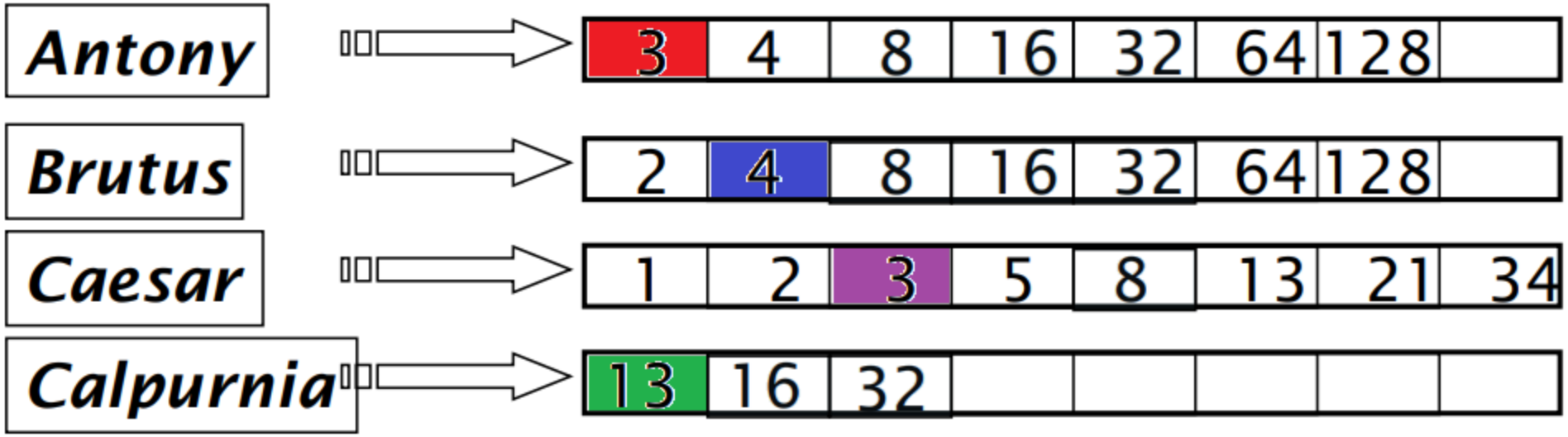
FRONTIER: 1


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

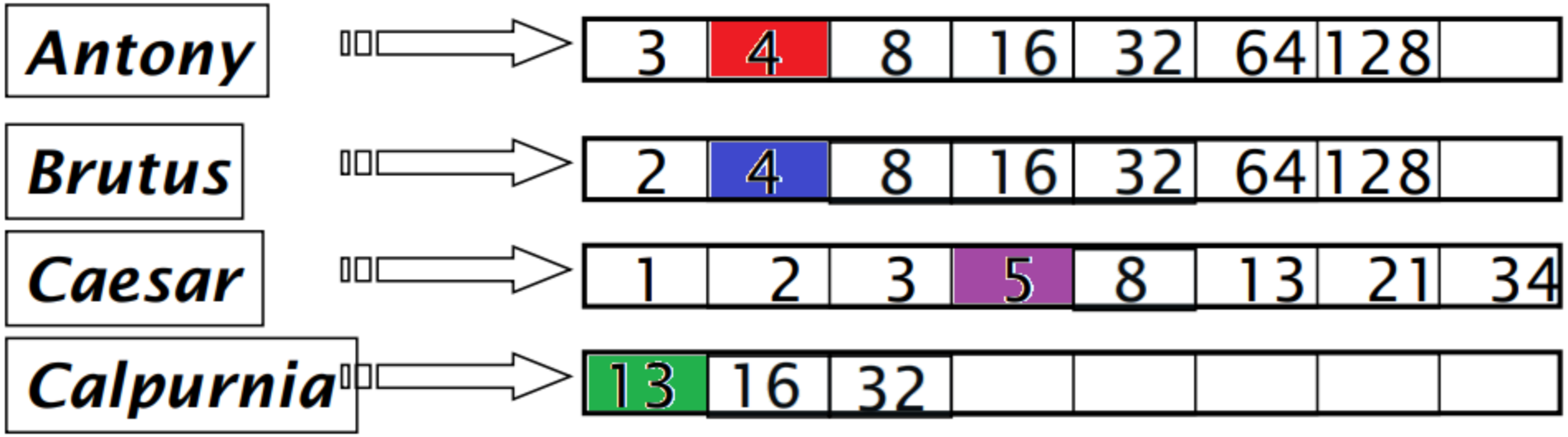
FRONTIER: 2


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

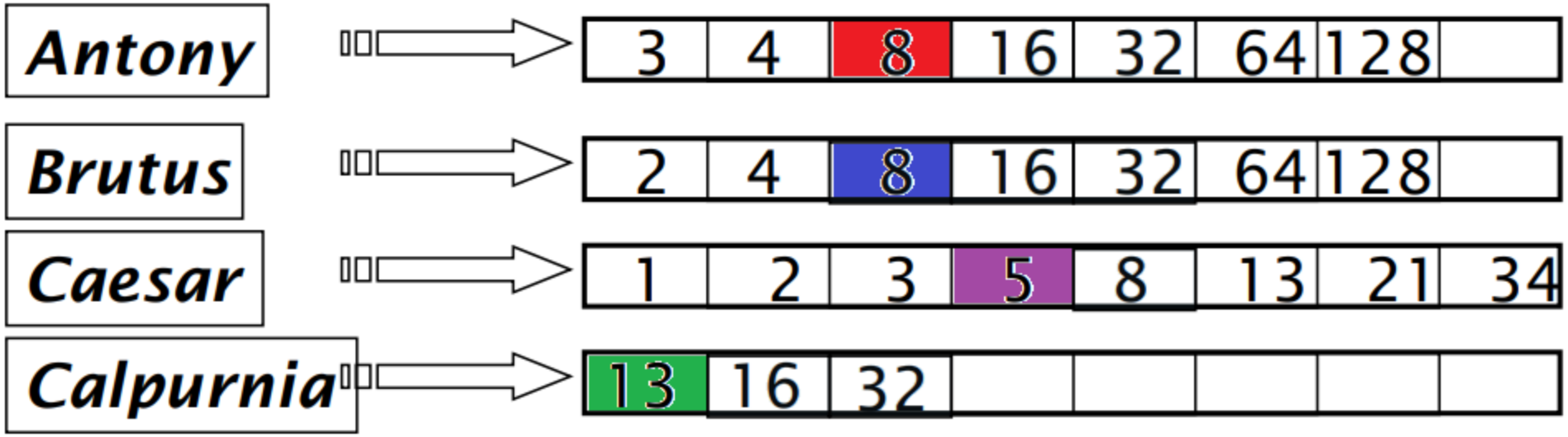
FRONTIER: 2


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

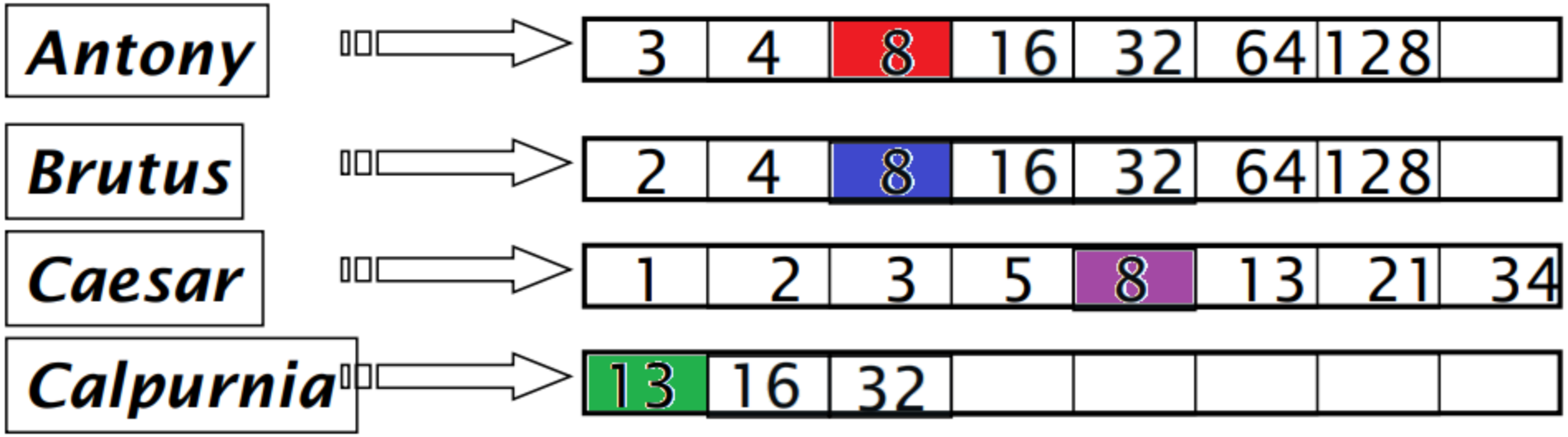
FRONTIER: 2


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

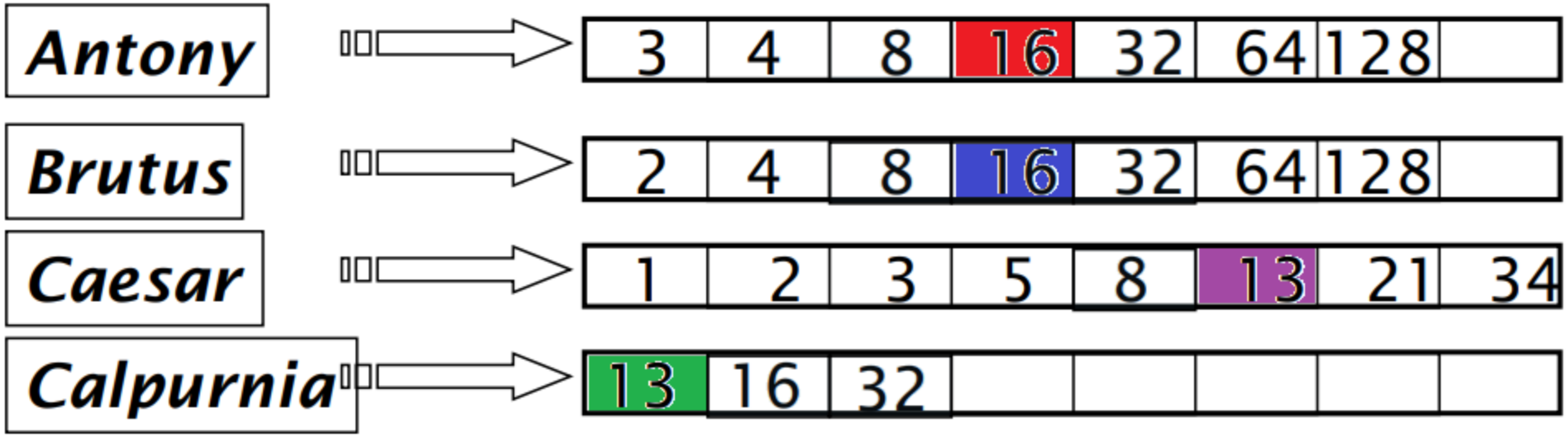
FRONTIER: 1


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

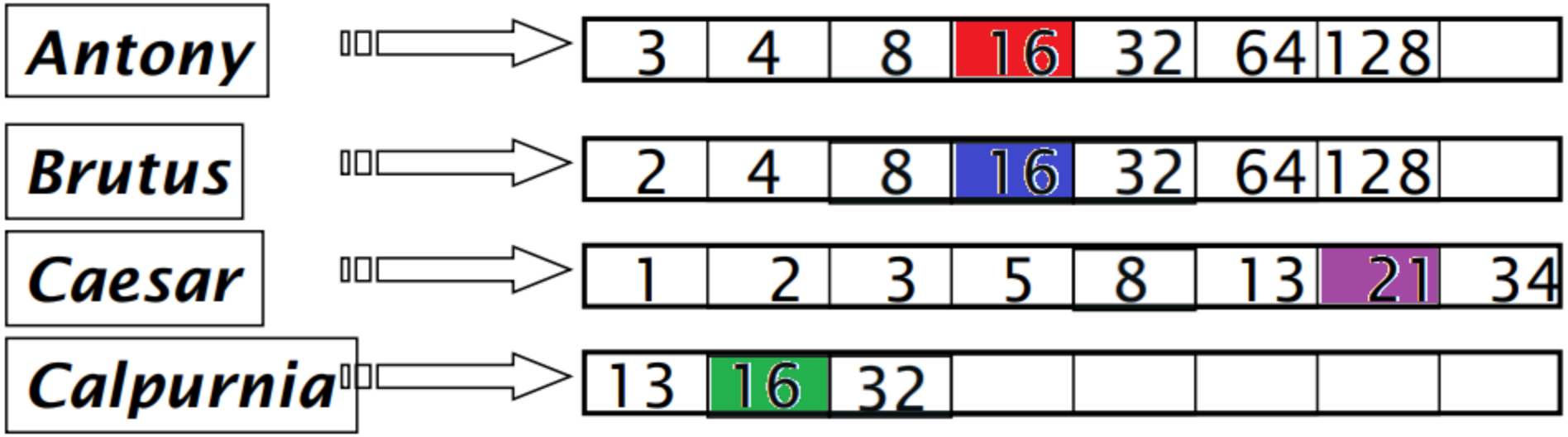
FRONTIER: 3


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

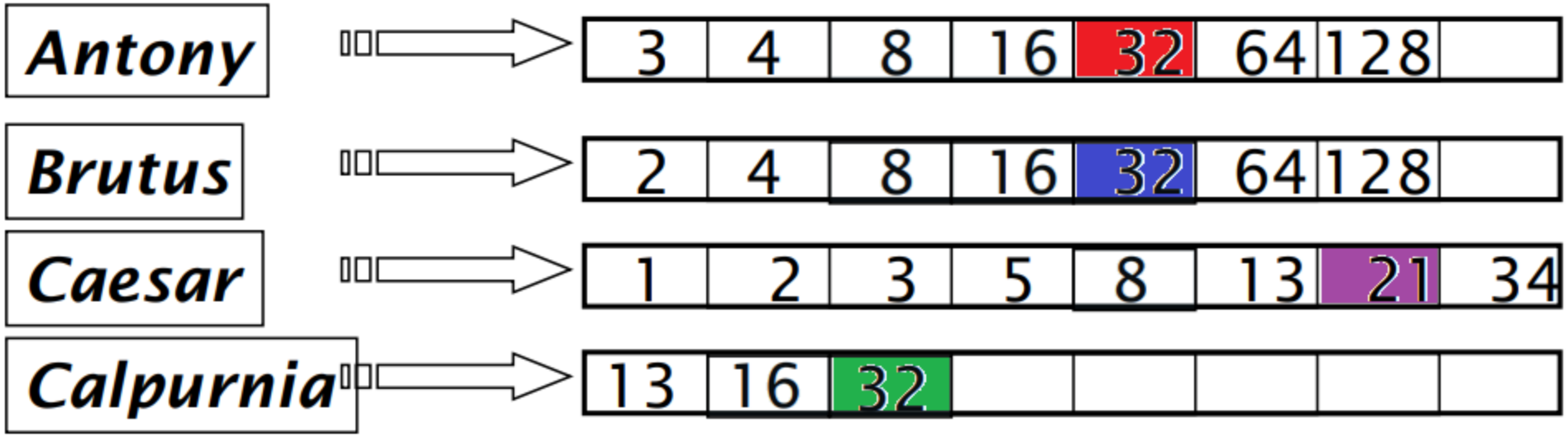
FRONTIER: 2


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

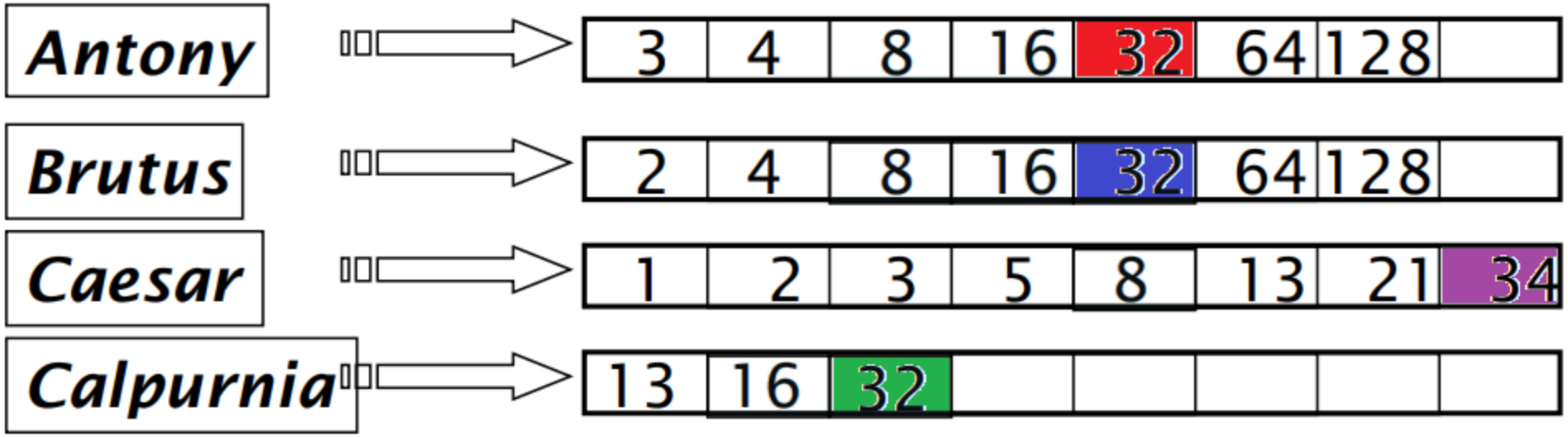
FRONTIER: 3


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

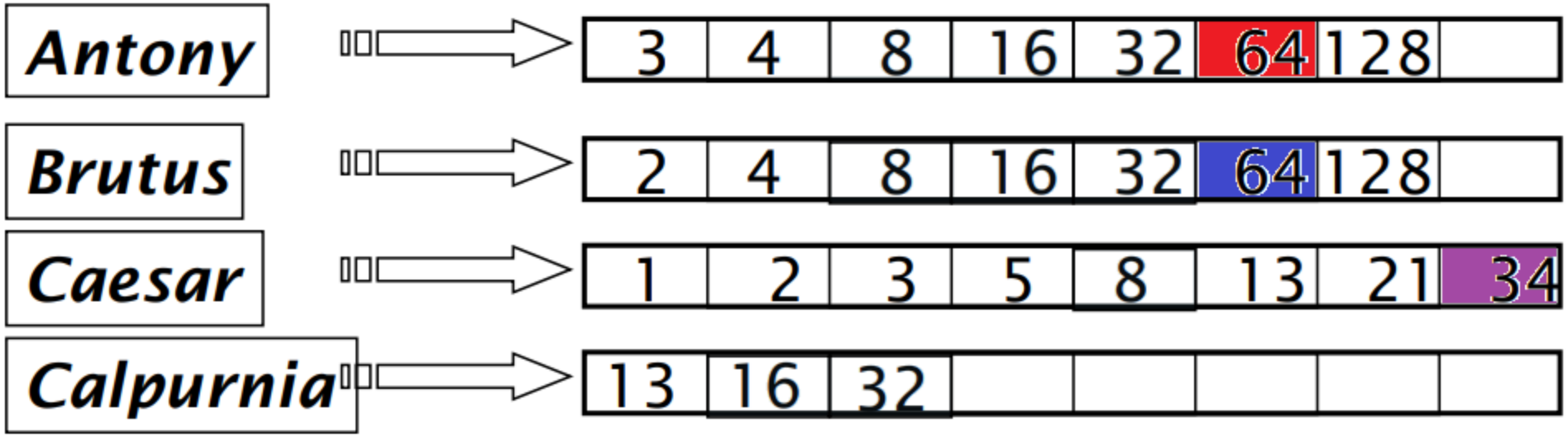
FRONTIER: 1


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

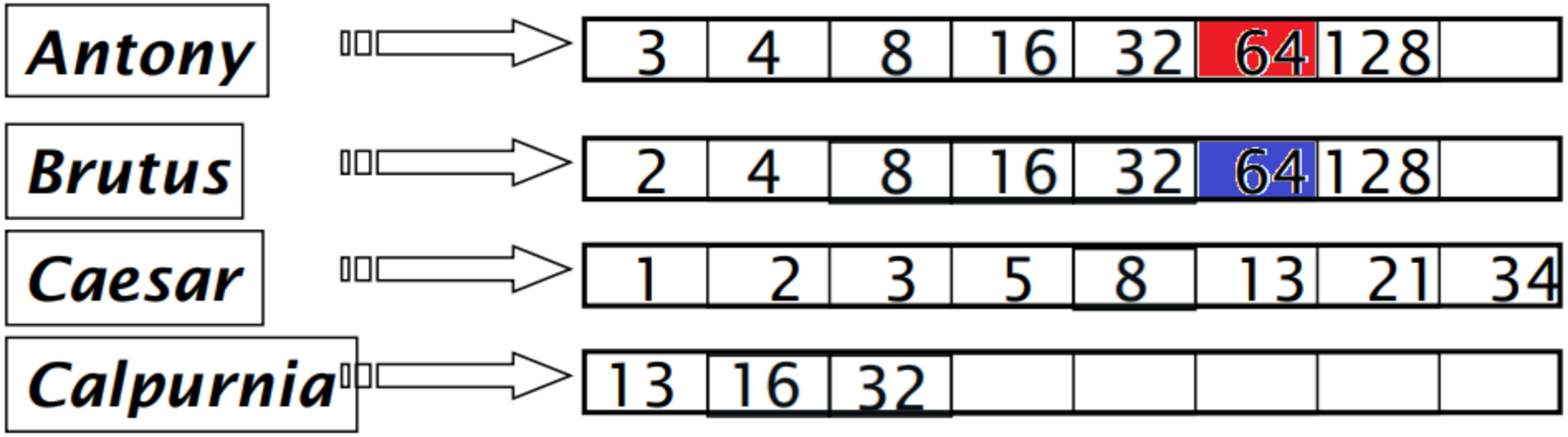
FRONTIER: 3


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



Term-at-a-time

- Incrementally compute the score for all documents
- Instead of keeping a pointer at the frontier of the posting lists, keep a structure of the current scores of all documents and iterate over the query

Example

- Given the query `"Antony Brutus Caesar Calpurnia"`
- And the inverted index

```
index = {
    "Antony" : [{"docId" : 1, "tf" : 3}, {"docId" : 5, "tf" : 1}],
    "Brutus" : [{"docId" : 4, "tf" : 10}, {"docId" : 5, "tf" : 1}],
    "Caesar" : [{"docId" : 1, "tf" : 1}, {"docId" : 2, "tf" : 1}, {"docId" : 3, "tf" : 1}],
    "Calpurnia" : [{"docId" : 2, "tf" : 4}, {"docId" : 6, "tf" : 8}]
    ...
}
```

Antony

- `"Antony" : [{"docId" : 1, "tf" : 3}, {"docId" : 5, "tf" : 1}]`

	1	2	3	4	5	6
Antony	3	0	0	0	1	0
Brutus						
Caesar						
Calpurnia						

Brutus

- `"Brutus" : [{"docId" : 4, "tf" : 10}, {"docId" : 5, "tf" : 1}]`

	1	2	3	4	5	6
Antony	3	0	0	0	1	0
Brutus	3	0	0	10	2	0
Caesar						
Calpurnia						

Caesar

- "Caesar" : [{"docId" : 1, "tf" : 1}, {"docId" : 2, "tf" : 1}, {"docId" : 5, "tf" : 1}]

	1	2	3	4	5	6
Antony	3	0	0	0	1	0
Brutus	3	0	0	10	2	0
Caesar	4	1	0	10	3	0
Calpurnia						

Calpurnia

- `"Calpurnia" : [{"docId" : 2, "tf" : 4}, {"docId" : 6, "tf" : 8}]`

	1	2	3	4	5	6
Antony	3	0	0	0	1	0
Brutus	3	0	0	10	2	0
Caesar	4	1	0	10	3	0
Calpurnia	4	5	0	10	3	8

Final scores

1	2	3	4	5	6
3	0	0	0	1	0
3	0	0	10	2	0
4	1	0	10	3	0
4	**5**	**0**	**10**	**3**	**8**

Agenda

- Some random topics
  - TAAT vs. DAAT
  - BSBI & SPIMI
- Double shoutout!!

BSBI –Blocked sort-based indexing

1. Split disk into blocks fitting in memory
2. Tokenize documents in block and create postings
3. Sort lexicographically
4. Repeat 1-3 for entire disk
5. Read and merge from multiple blocks, write back to disk

Blocked sort-based indexing

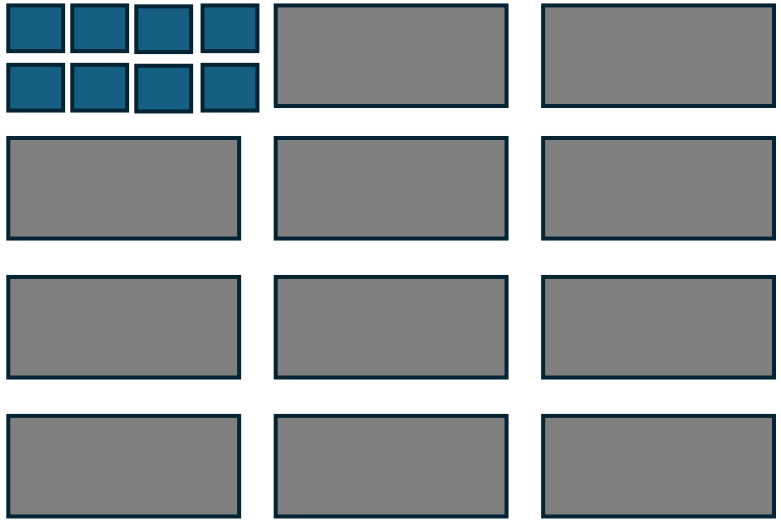
- Break corpus into **blocks** which can approximately fit in main memory

Blocked sort-based indexing

- Read **one block at a time**

Blocked sort-based indexing

- **Tokenise** the documents



Blocked sort-based indexing

- Create **postings** from all terms, **sort** alphabetically

Term1, doc: 1	Term2, doc: 1	Term3, doc: 1	Term4, doc: 1
Term5, doc: 1	Term6, doc: 2	Term7, doc: 2	Term8, doc: 3

Blocked sort-based indexing

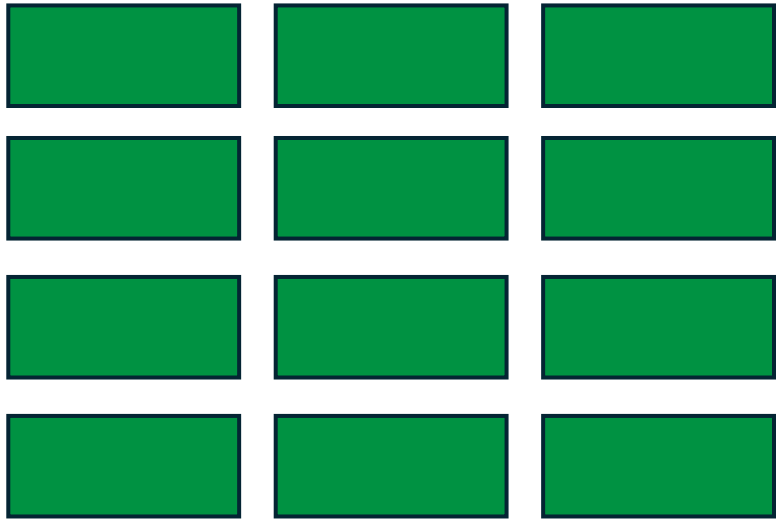
- Write it **back** to disk

Blocked sort-based indexing

- Start over with a **new** block

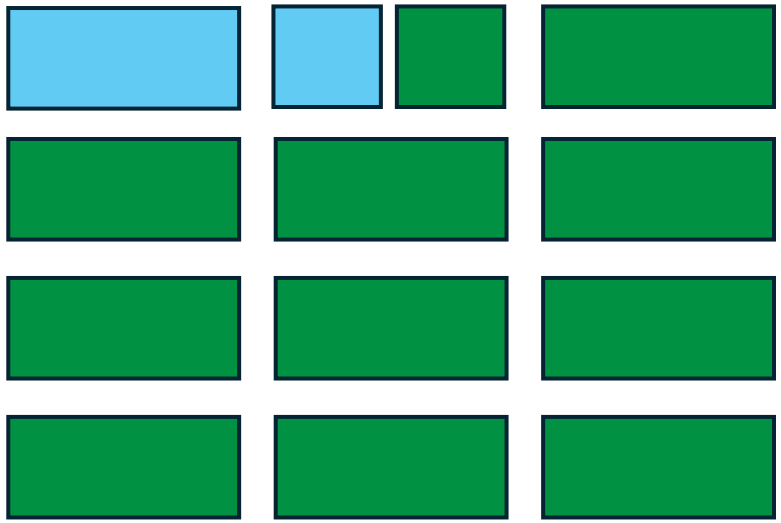
Blocked sort-based indexing

- Continue until **all** blocks are converted



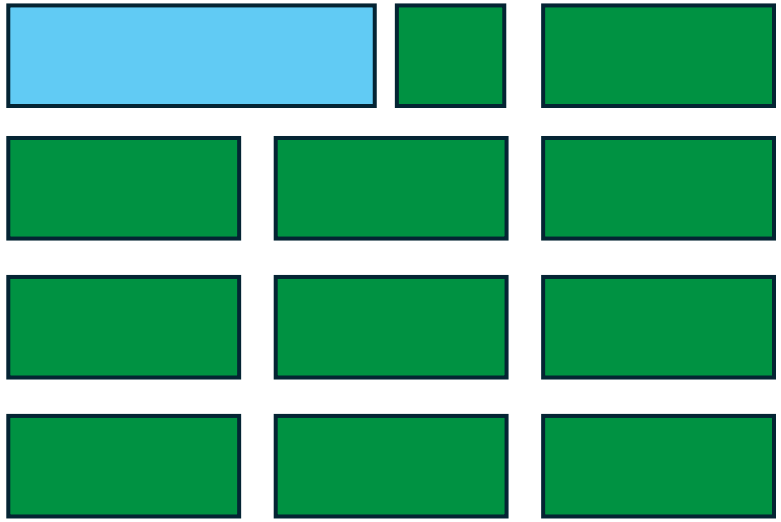
Blocked sort-based indexing

- Read **parts** of blocks from disk



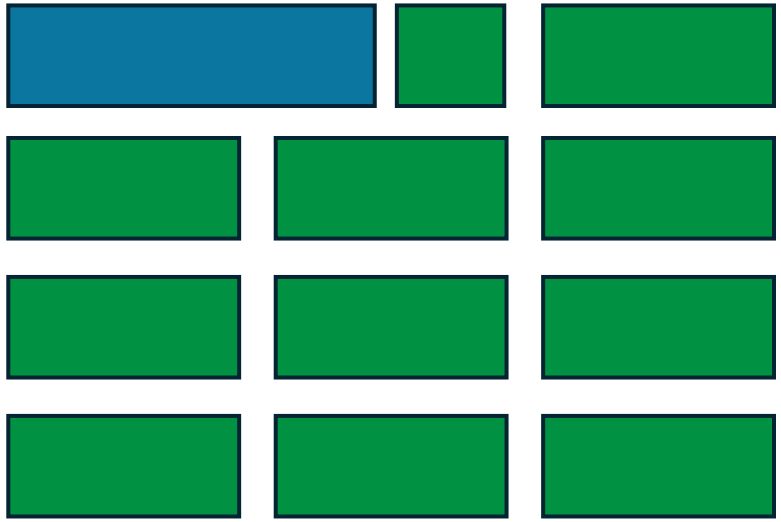
Blocked sort-based indexing

- **Merge** them



Blocked sort-based indexing

- Write them **back** to disk



Blocked sort-based indexing

- Continue until **all** the postings are one big inverted index

SPIMI – Single pass in memory indexing

- What if we can’t keep the dictionary in memory?
- 2 key points
  1. Create separate dictionaries for each block
  2. Accumulate postings in posting lists as they occur. Don’t sort
- Each block will be a complete inverted index
- The separate indexes can be merged into one larger index

Single-pass in-memory indexing

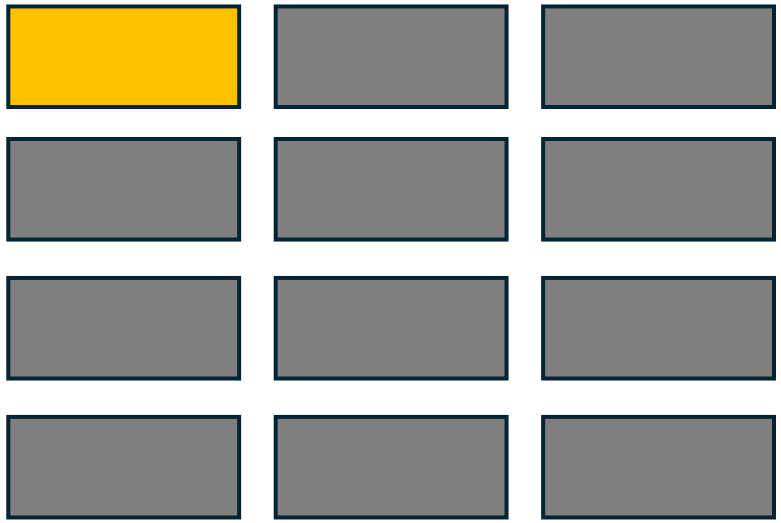
- Given a huge corpus

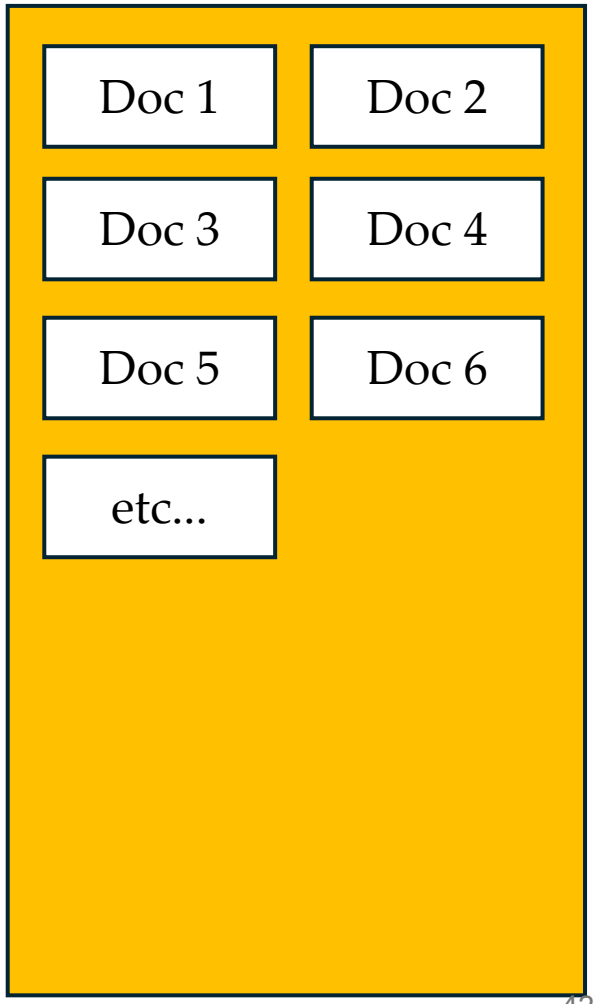
Single-pass in-memory indexing

- Break corpus into blocks which can approximately fit in main memory

Single-pass in-memory indexing

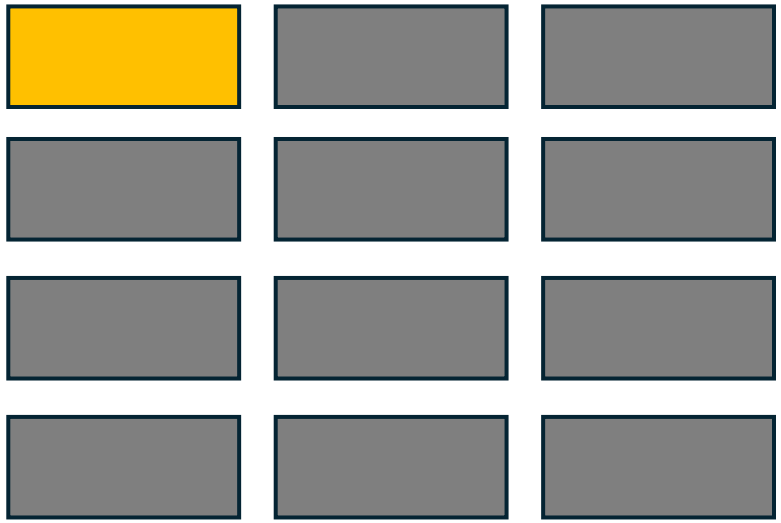
- Read one block at a time

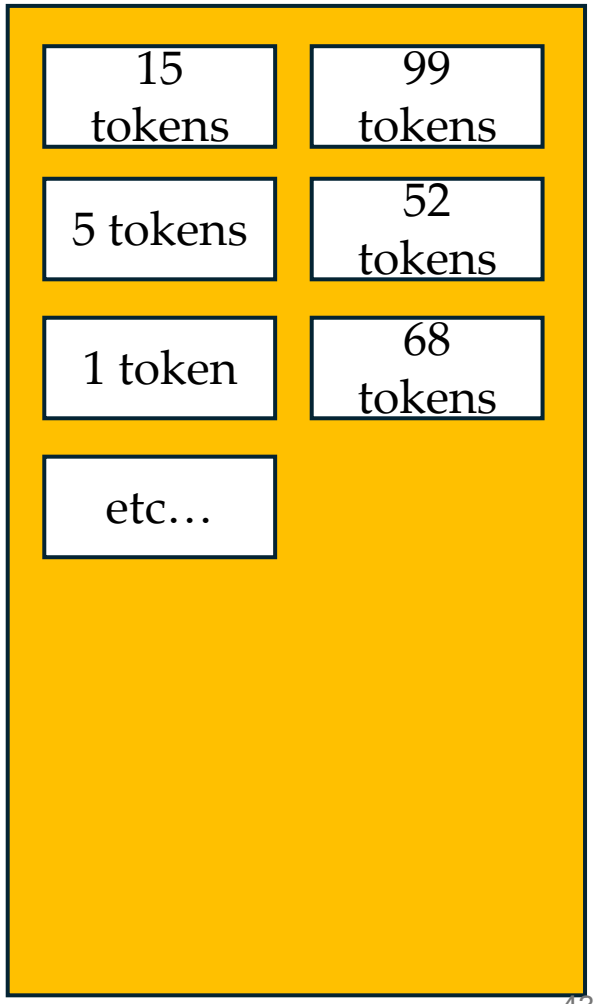




Single-pass in-memory indexing

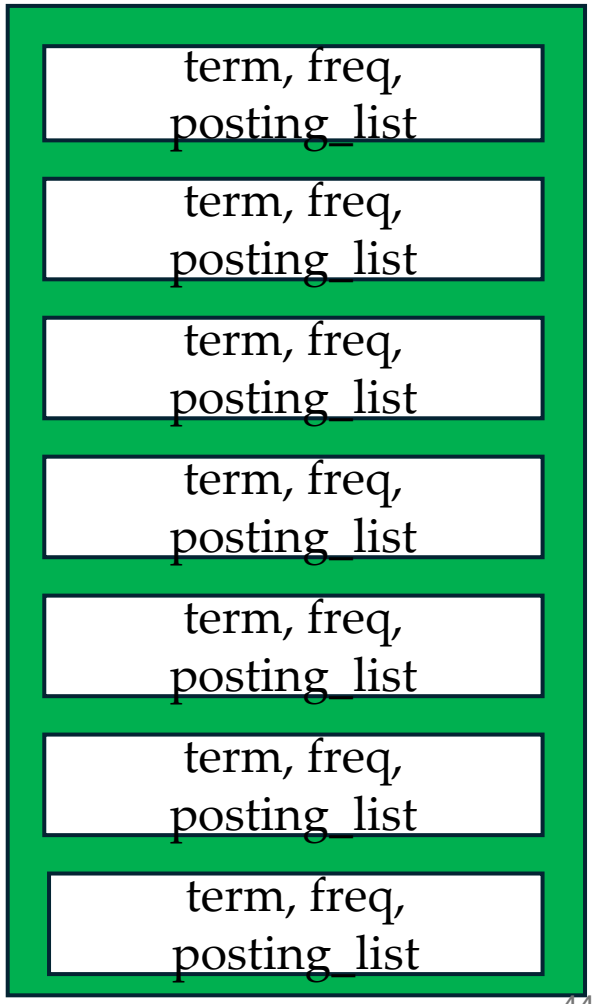
- Tokenise the documents





Single-pass in-memory indexing

- Create an inverted index for the block

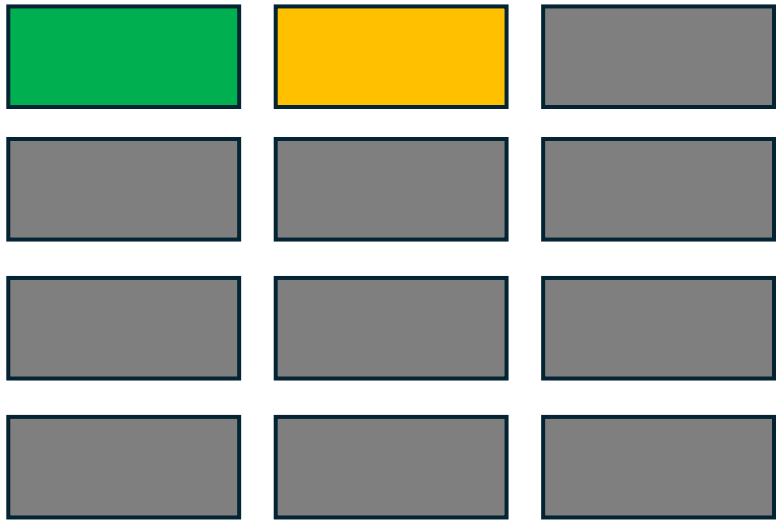


Single-pass in-memory indexing

- Write the block/mini inverted index to disk

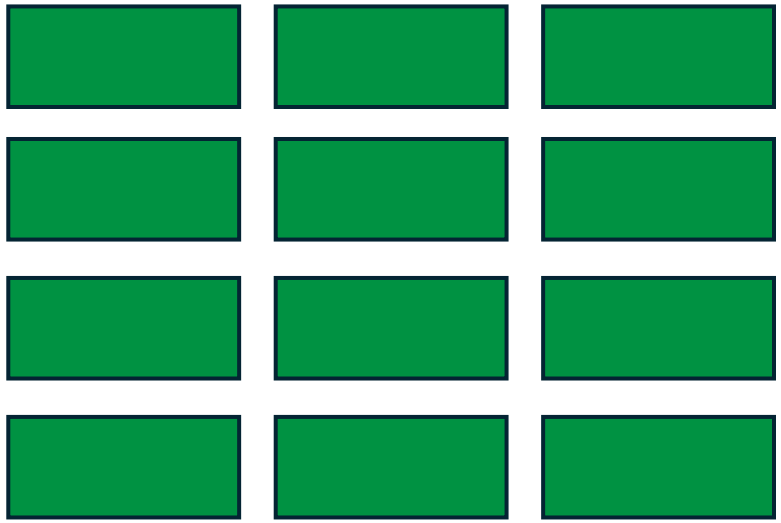
Single-pass in-memory indexing

- Start over with a new block



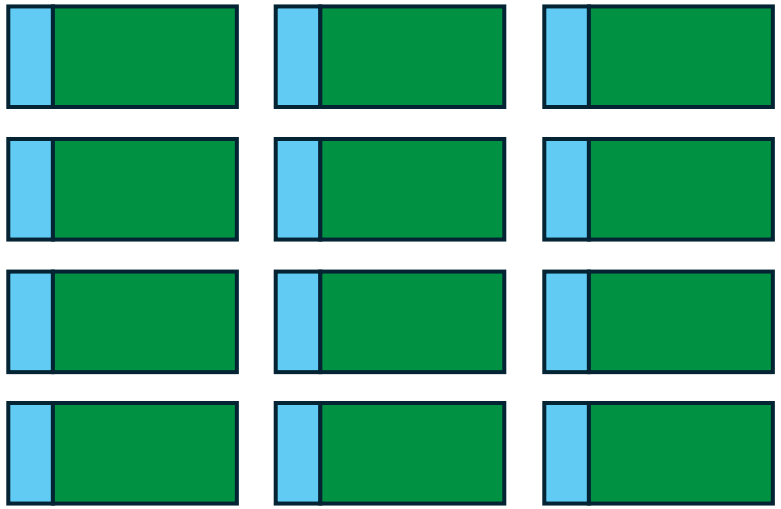
Single-pass in-memory indexing

- Continue until all blocks are converted



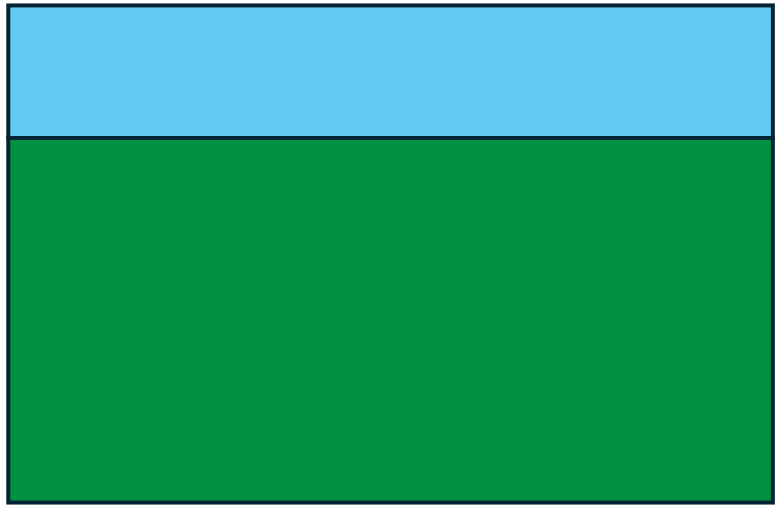
Single-pass in-memory indexing

- Using a k-way merge, merge the blocks to a single inverted index



Single-pass in-memory indexing

- We keep some of it in memory and store the rest on disk



BSBI

- Create postings of global index, merge postings togetherin this index
- Each block contains postings sorted lexicographically
- Needs global term-termID mapping: need dictionary to fit in memory

SPIMI

- Create multiple local indexes, merge these together to make global index
- Each block is a separate index
- A block has term-termID mappings for its own index: Only need to know the terms in one block

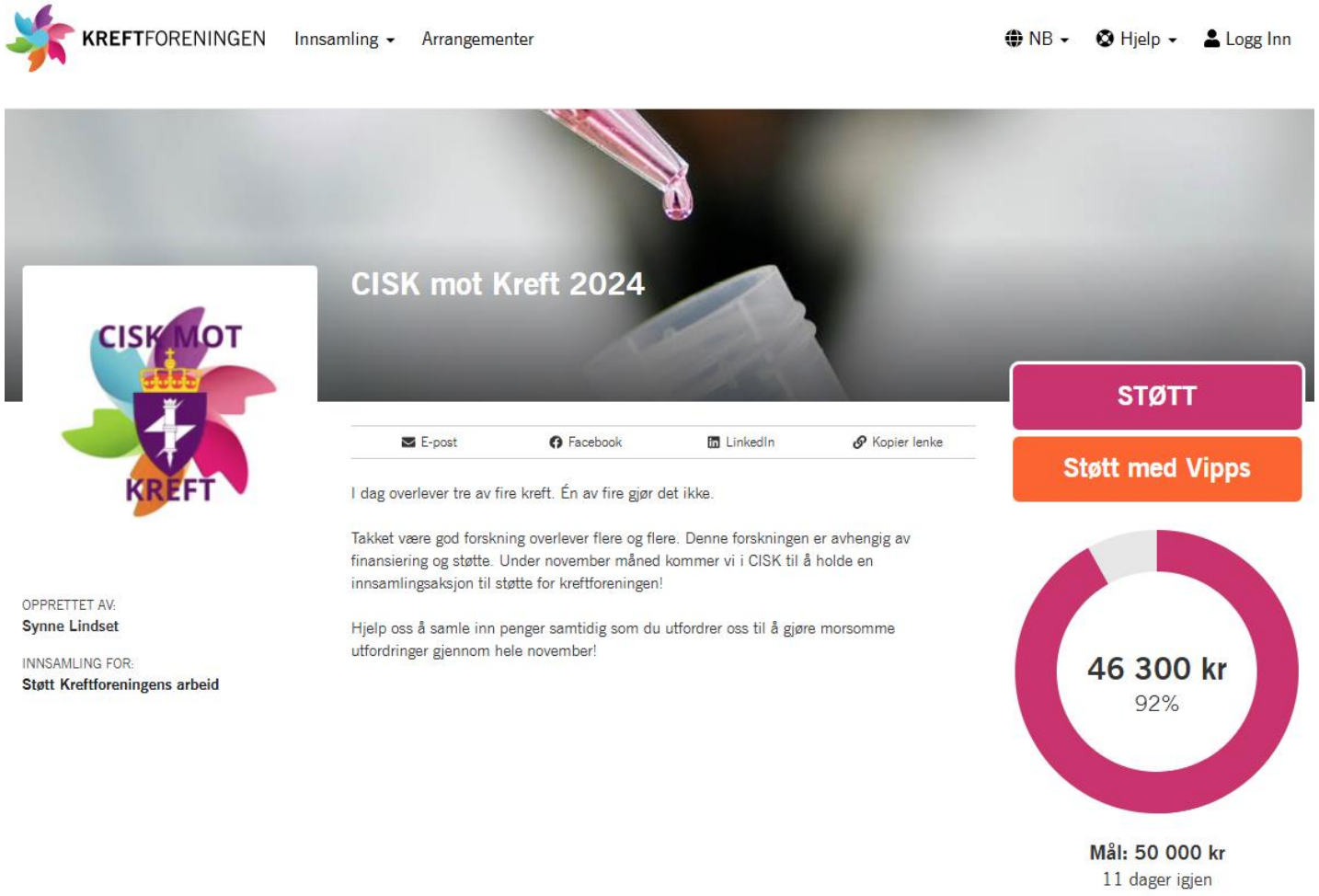
Agenda

- Some random topics
  - TAAT vs. DAAT
  - BSBI & SPIMI
- Double shoutout!!

Shoutout #1

Netcompany

Shoutout #2


KREFTFORENINGEN
Innsamling
Arrangementer
NB
Hjelp
Logg Inn
CISK mot Kreft 2024
CISK MOT
KREFT
E-post
Facebook
LinkedIn
Kopier lenke
I dag overlever tre av fire kreft. Én av fire gjør det ikke.
Takket være god forskning overlever flere og flere. Denne forskningen er avhengig av finansiering og støtte. Under november måned kommer vi i CISK til å holde en innsamlingsaksjon til støtte for kreftforeningen!
Hjelp oss å samle inn penger samtidig som du utfordrer oss til å gjøre morsomme utfordringer gjennom hele november!
OPPRETTET AV:
Synne Lindset
INNSAMLING FOR:
Støtt Kreftforeningens arbeid
STØTT
Støtt med Vipps
46 300 kr
92%
Mål: 50 000 kr
11 dager igjen

15 min break