Multileaved Comparisons for Fast Online Evaluation

Anne Schuth, Floor Sietsma, Shimon Whiteson,
Damien Lefortier, Maarten de Rijke

CIKM'14
November 4, 2014, Shanghai, China

Motivation

- Search engines constantly evolve
- Engineers and researchers develop new rankers, potential improvements
- Goal: improving over production ranker
- Tool: comparisons *between* experimental rankers

Motivation cont.

- Information for engineers and researchers

	R1	R2	…	Rn
P	.59	.47	…	.51

Motivation cont.

- Information for engineers and researchers

	R1	R2	…	Rn
P	.59	.47	…	.51
R1	.50	**.78**	**…**	**.62**
R2	.22	.50	**…**	**.48**
…	…	…	.50	**…**
Rn	.38	.52	…	.50

Motivation cont.

- Information for engineers and researchers

	R1	R2	…	Rn
P	.59	.47	…	.51
R1	.50	**.78**	**…**	**.62**
R2	.22	.50	**…**	**.48**
…	…	…	.50	**…**
Rn	.38	.52	…	.50

- **Interleaving**: 0.5 * N * (N - 1) comparisons

Motivation cont.

- Information for engineers and researchers

	R1	R2	…	Rn
P	.59	.47	…	.51
R1	.50	**.78**	**…**	**.62**
R2	.22	.50	**…**	**.48**
…	…	…	.50	**…**
Rn	.38	.52	…	.50

- **Interleaving**: 0.5 * N * (N - 1) comparisons

- **Multileaving**: 1 comparison

Comparison Methods

- **Team Draft Interleave (TD)**
  F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM ’08. ACM Press, 2008.

- **Team Draft Multileave (TDM)**
  our multileave extension of TD

- **Optimized Interleave (OI)**
  F. Radlinski and N. Craswell. Optimized interleaving for online retrieval evaluation. In WSDM ’13. ACM Press, 2013.

- **Optimized Multileave (OM)**
  our multileave extension of OI

Comparison Methods

- **Team Draft Interleave (TD)**
  F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM ’08. ACM Press, 2008.

- **Team Draft Multileave (TDM)**
  our multileave extension of TD

- **Optimized Interleave (OI)**
  F. Radlinski and N. Craswell. Optimized interleaving for online retrieval evaluation. In WSDM ’13. ACM Press, 2013.

- **Optimized Multileave (OM)**
  our multileave extension of OI

Team Draft Interleave (TD)

A	B
doc 1	doc 2
doc 2	doc 4
doc 3	doc 7
doc 4	doc 1
doc 5	doc 3

Team Draft Interleave (TD)

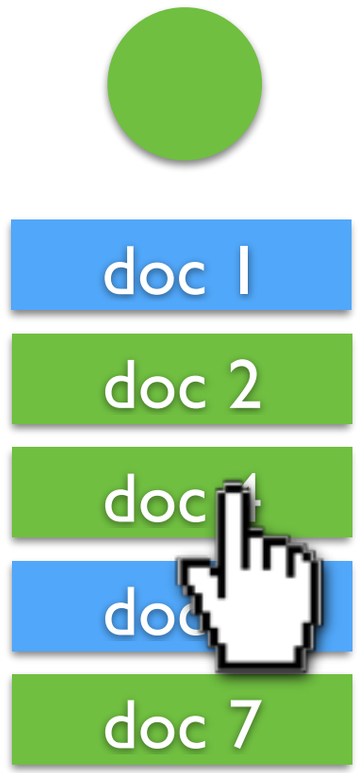
A	B
doc 1	doc 2
doc 2	doc 4
doc 3	doc 7
doc 4	doc 1
doc 5	doc 3

Team Draft Interleave (TD)

A B

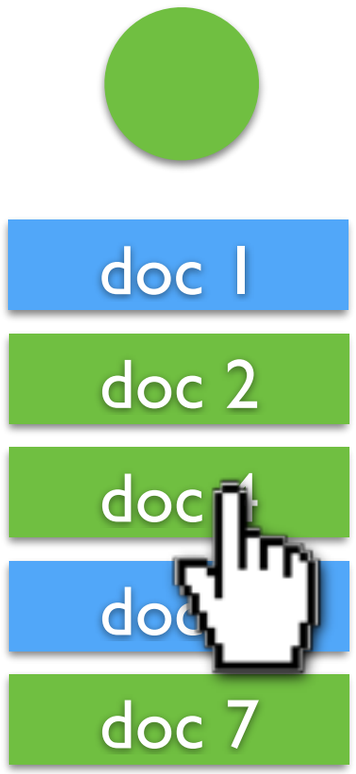
Team Draft Interleave (TD)

A B



Team Draft Interleave (TD)

A B



Inference:

B > A

Comparison Methods

- **Team Draft Interleave (TD)**
  F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM ’08. ACM Press, 2008.

- **Team Draft Multileave (TDM)**
  our multileave extension of TD

- **Optimized Interleave (OI)**
  F. Radlinski and N. Craswell. Optimized interleaving for online retrieval evaluation. In WSDM ’13. ACM Press, 2013.

- **Optimized Multileave (OM)**
  our multileave extension of OI

Comparison Methods

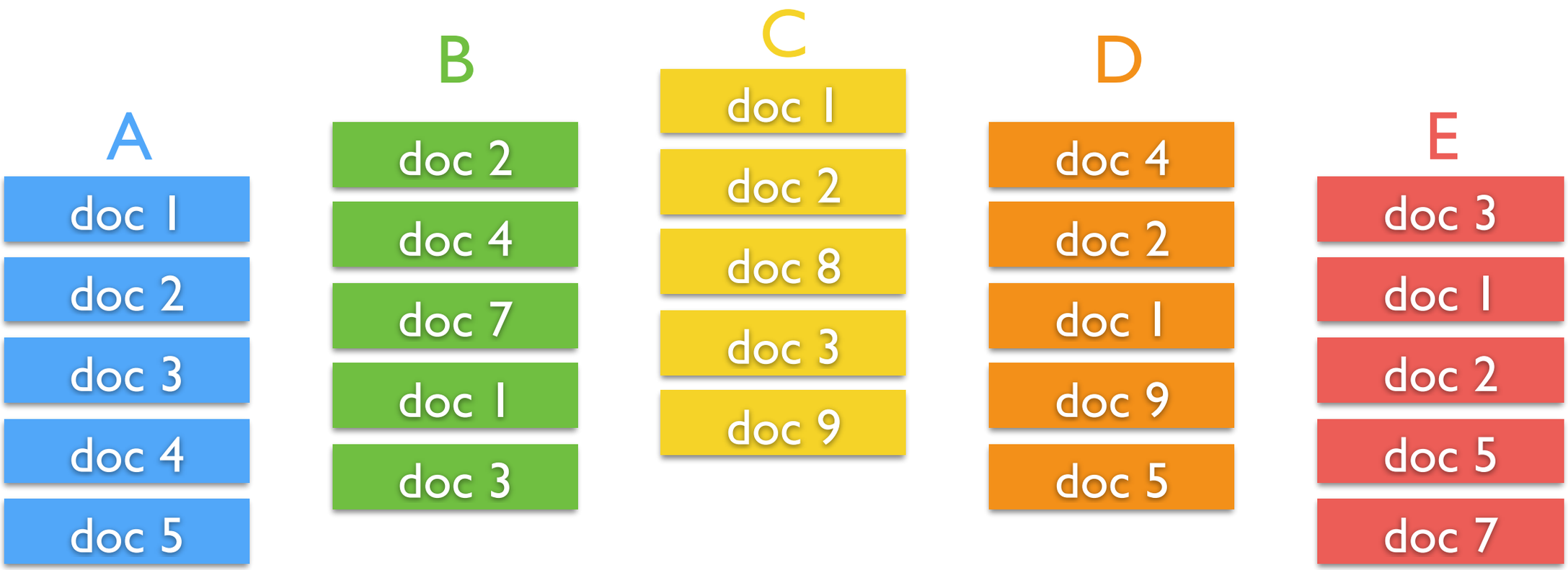
- **Team Draft Interleave (TD)**
  F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM ’08. ACM Press, 2008.

- **Team Draft Multileave (TDM)**
  our multileave extension of TD

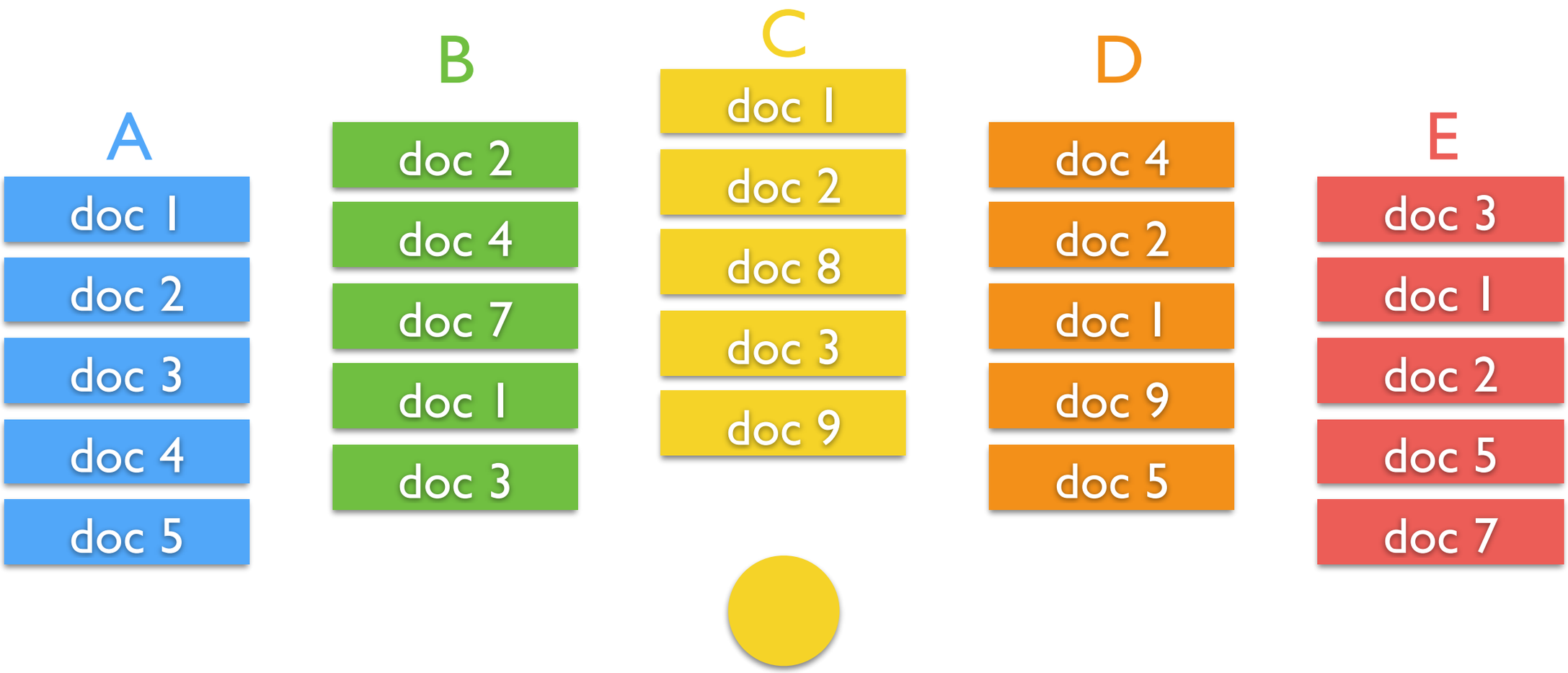
- **Optimized Interleave (OI)**
  F. Radlinski and N. Craswell. Optimized interleaving for online retrieval evaluation. In WSDM ’13. ACM Press, 2013.

- **Optimized Multileave (OM)**
  our multileave extension of OI

Team Draft Multileave (TDM)

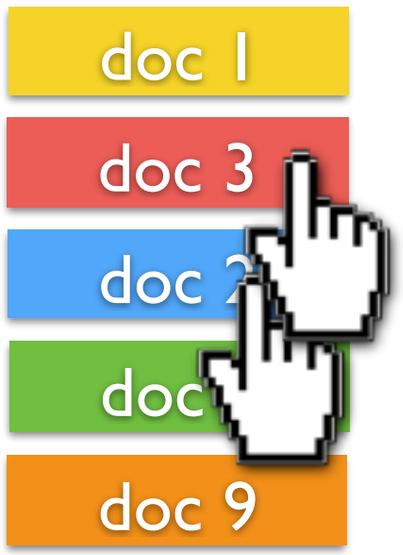


Team Draft Multileave (TDM)

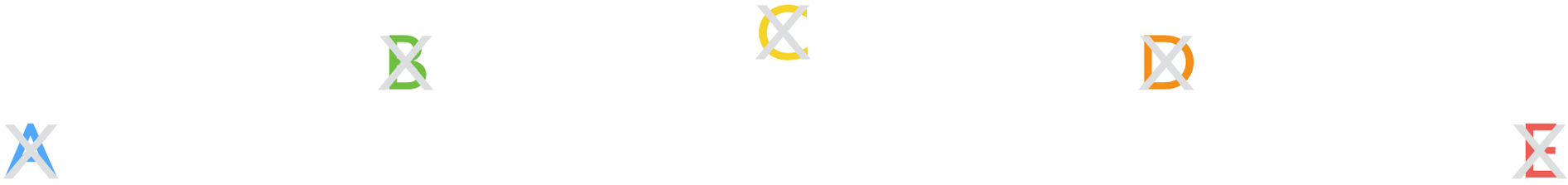


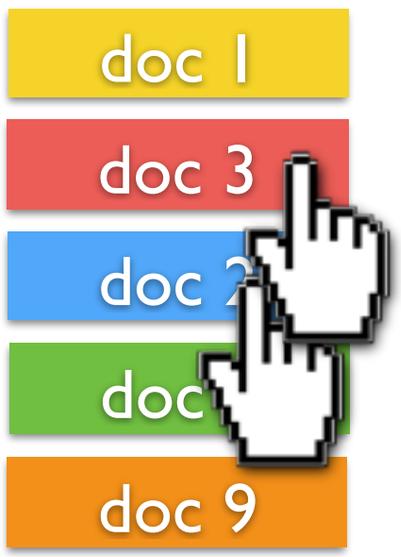
Team Draft Multileave (TDM)

Team Draft Multileave (TDM)



Team Draft Multileave (TDM)





Inference:

A & E > B & C & D

Potential Problems with TDM

- SERP length
  - limits the number of rankers that can be compared
  - never more rankers then slots in the SERP
- Solution: Optimized Interleave

Comparison Methods

- **Team Draft Interleave (TD)**
  F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM ’08. ACM Press, 2008.

- **Team Draft Multileave (TDM)**
  our multileave extension of TD

- **Optimized Interleave (OI)**
  F. Radlinski and N. Craswell. Optimized interleaving for online retrieval evaluation. In WSDM ’13. ACM Press, 2013.

- **Optimized Multileave (OM)**
  our multileave extension of OI

Comparison Methods

- **Team Draft Interleave (TD)**
  F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM ’08. ACM Press, 2008.

- **Team Draft Multileave (TDM)**
  our multileave extension of TD

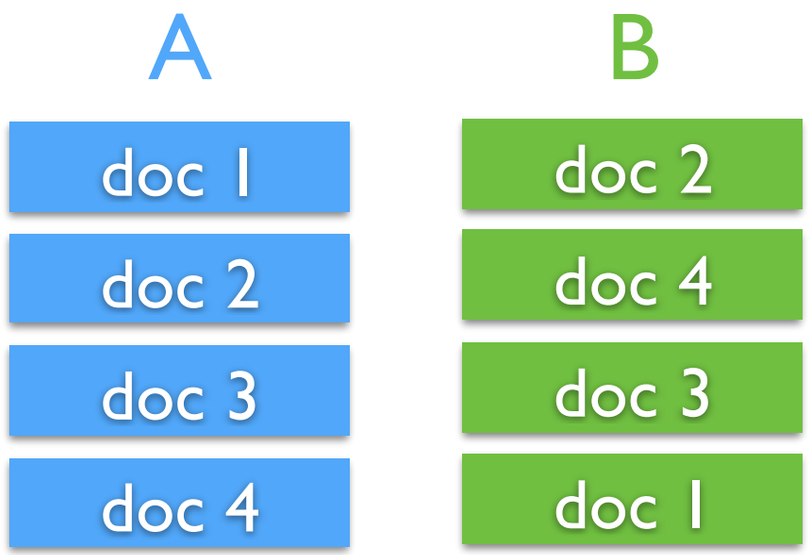
- **Optimized Interleave (OI)**
  F. Radlinski and N. Craswell. Optimized interleaving for online retrieval evaluation. In WSDM ’13. ACM Press, 2013.

- **Optimized Multileave (OM)**
  our multileave extension of OI

Optimized Interleave (OI)

Constraints:

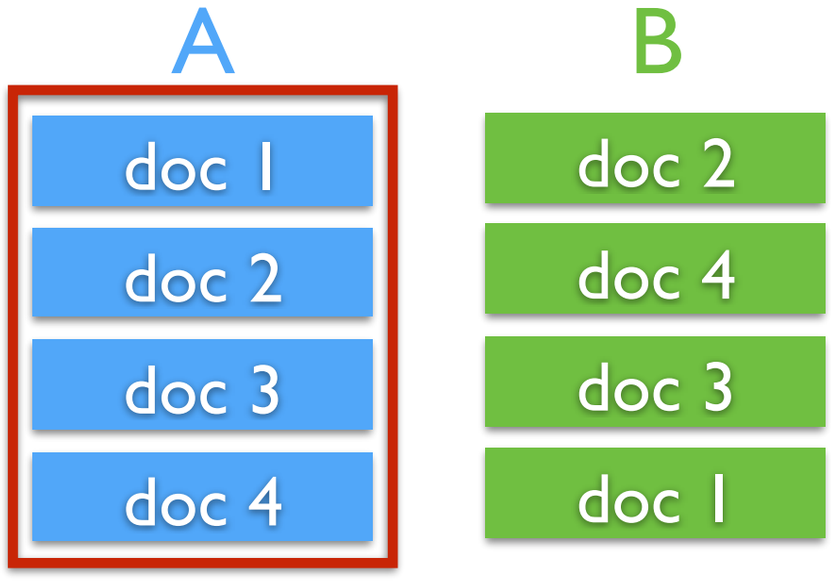
1. Prefix



Optimized Interleave (OI)

Constraints:

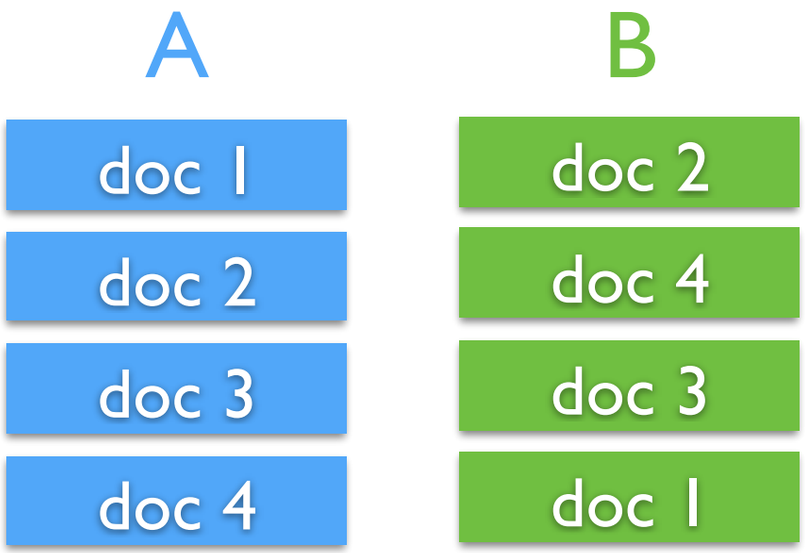
1. Prefix



Optimized Interleave (OI)

Constraints:

1. Prefix



doc 1

doc 2

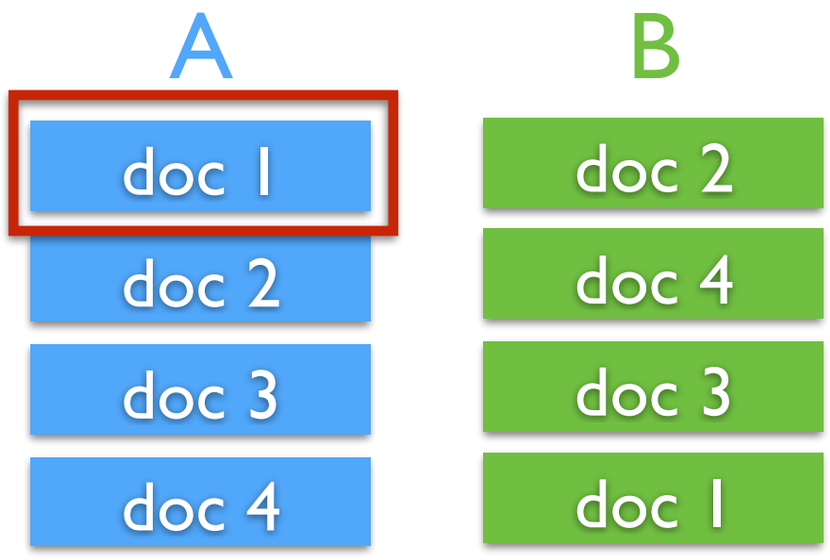
doc 3

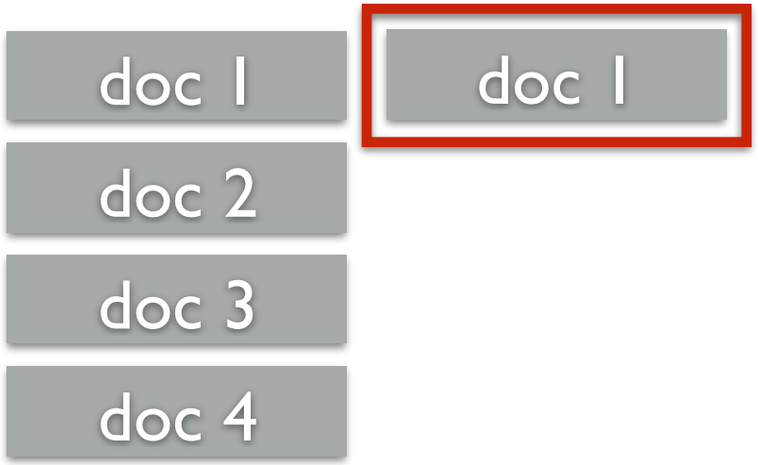
doc 4

Optimized Interleave (OI)

Constraints:

1. Prefix

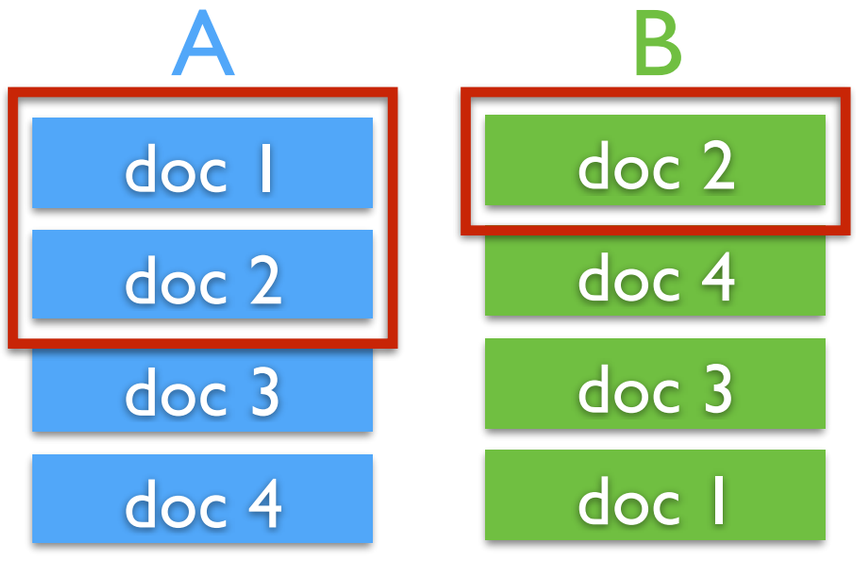




Optimized Interleave (OI)

Constraints:

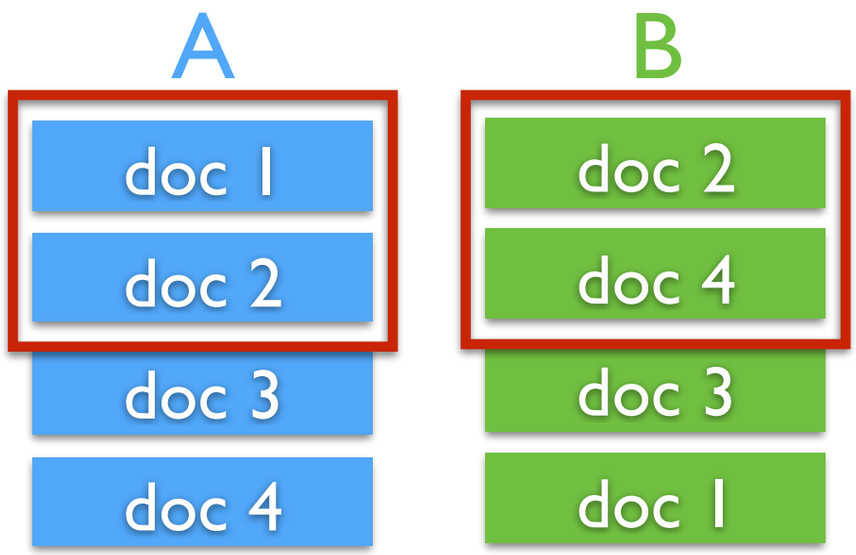
1. Prefix



Optimized Interleave (OI)

Constraints:

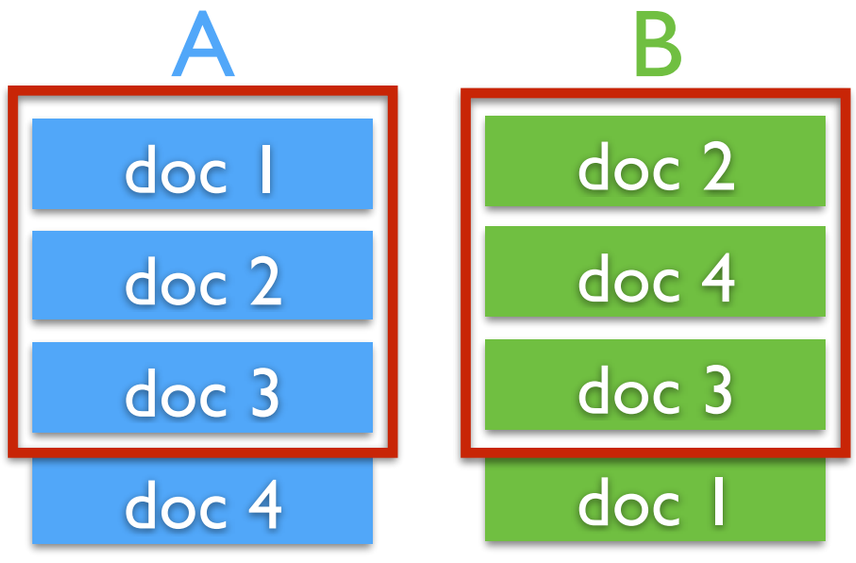
1. Prefix

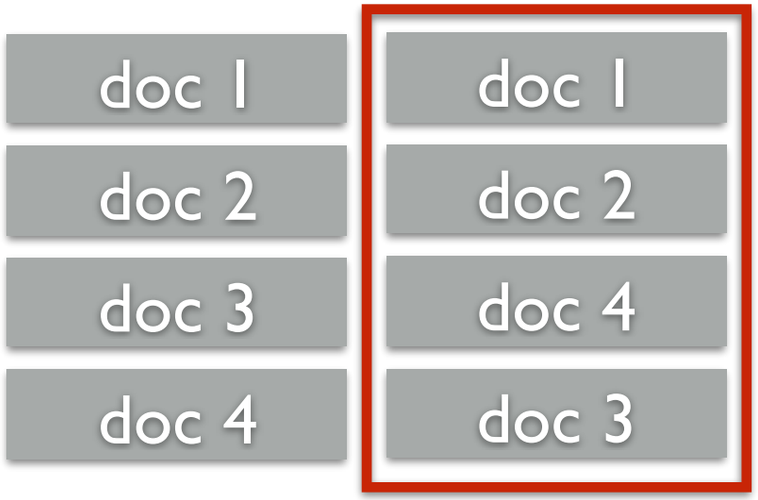


Optimized Interleave (OI)

Constraints:

1. Prefix





Optimized Interleave (OI)

Constraints:

1. Prefix

A	B
doc 1	doc 2
doc 2	doc 4
doc 3	doc 3
doc 4	doc 1

doc 1	doc 1
doc 2	doc 2
doc 3	doc 4
doc 4	doc 3

Optimized Interleave (OI)

Constraints:

1. Prefix

A	B
doc 1	doc 2
doc 2	doc 4
doc 3	doc 3
doc 4	doc 1

doc 1	doc 1	doc 2	doc 2	doc 2	doc 2
doc 2	doc 2	doc 1	doc 1	doc 4	doc 4
doc 3	doc 4	doc 3	doc 4	doc 1	doc 3
doc 4	doc 3	doc 4	doc 3	doc 3	doc 1

Optimized Interleave (OI)

Constraints:

1. Prefix
2. Unbiased

A	B
doc 1	doc 2
doc 2	doc 4
doc 3	doc 3
doc 4	doc 1

doc 1	doc 1	doc 2	doc 2	doc 2	doc 2
doc 2	doc 2	doc 1	doc 1	doc 4	doc 4
doc 3	doc 4	doc 3	doc 4	doc 1	doc 3
doc 4	doc 3	doc 4	doc 3	doc 3	doc 1

Optimized Interleave (OI)

Constraints:

1. Prefix
2. Unbiased

	A	B				
	doc 1	doc 2				
	doc 2	doc 4				
	doc 3	doc 3				
	doc 4	doc 1				
3	doc 1	doc 1	doc 2	doc 2	doc 2	doc 2
	doc 2	doc 2	doc 1	doc 1	doc 4	doc 4
	doc 3	doc 4	doc 3	doc 4	doc 1	doc 3
	doc 4	doc 3	doc 4	doc 3	doc 3	doc 1

Optimized Interleave (OI)

Constraints:

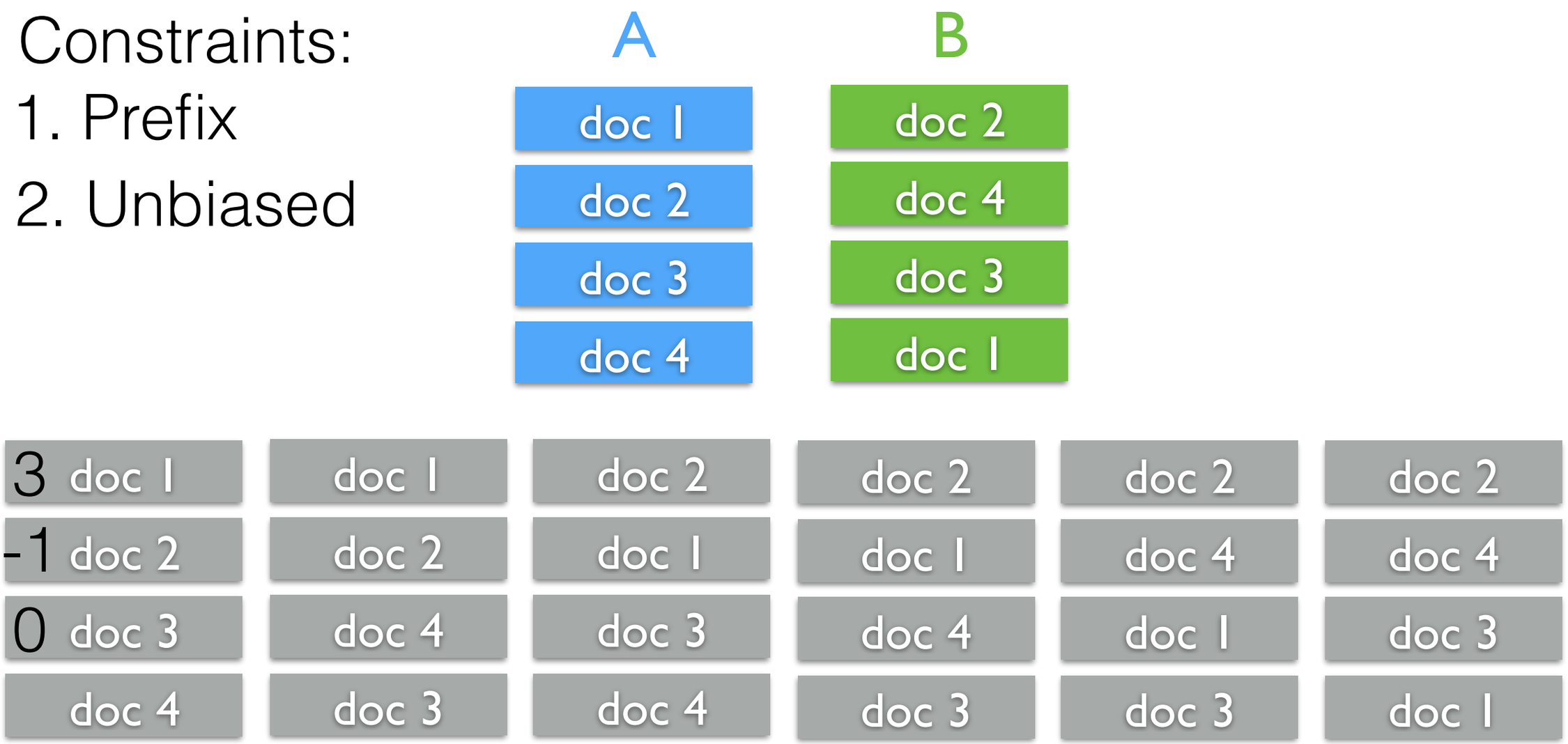
1. Prefix
2. Unbiased

	A	B				
	doc 1	doc 2				
	doc 2	doc 4				
	doc 3	doc 3				
	doc 4	doc 1				
3	doc 1	doc 1	doc 2	doc 2	doc 2	doc 2
-1	doc 2	doc 2	doc 1	doc 1	doc 4	doc 4
	doc 3	doc 4	doc 3	doc 4	doc 1	doc 3
	doc 4	doc 3	doc 4	doc 3	doc 3	doc 1

Optimized Interleave (OI)

Constraints:

1. Prefix
2. Unbiased



Optimized Interleave (OI)

Constraints:

1. Prefix
2. Unbiased

A	B
doc 1	doc 2
doc 2	doc 4
doc 3	doc 3
doc 4	doc 1

3 doc 1	3 doc 1	-1 doc 2	-1 doc 2	-1 doc 2	-1 doc 2
-1 doc 2	-1 doc 2	3 doc 1	3 doc 1	-2 doc 4	-2 doc 4
0 doc 3	-2 doc 4	0 doc 3	-2 doc 4	3 doc 1	0 doc 3
-2 doc 4	0 doc 3	-2 doc 4	0 doc 3	0 doc 3	3 doc 1

Optimized Interleave (OI)

Constraints:

1. Prefix
2. Unbiased

A	B
doc 1	doc 2
doc 2	doc 4
doc 3	doc 3
doc 4	doc 1

3 * p1 +	3 * p2 +	-1 * p3 +	-1 * p4 +	-1 * p5 +	-1 * p6	= 0
-1 doc 2	-1 doc 2	3 doc 1	3 doc 1	-2 doc 4	-2 doc 4	
0 doc 3	-2 doc 4	0 doc 3	-2 doc 4	3 doc 1	0 doc 3	
-2 doc 4	0 doc 3	-2 doc 4	0 doc 3	0 doc 3	3 doc 1	

Optimized Interleave (OI)

Constraints:

1. Prefix
2. Unbiased

A	B
doc 1	doc 2
doc 2	doc 4
doc 3	doc 3
doc 4	doc 1

$$
\begin{aligned}
3\,p_1 + 3\,p_2 - 1\,p_3 - 1\,p_4 - 1\,p_5 - 1\,p_6 &= 0 \\
-1\,p_1 - 1\,p_2 + 3\,p_3 + 3\,p_4 - 2\,p_5 - 2\,p_6 &= 0 \\
0\,p_1 - 2\,p_2 + 0\,p_3 - 2\,p_4 + 3\,p_5 + 0\,p_6 &= 0 \\
-2\,p_1 + 0\,p_2 - 2\,p_3 + 0\,p_4 + 0\,p_5 + 3\,p_6 &= 0
\end{aligned}
$$

Optimized Interleave (OI)

Constraints:

1. Prefix
2. Unbiased

A	B
doc 1	doc 2
doc 2	doc 4
doc 3	doc 3
doc 4	doc 1

3 * p1 +	3 * p2 +	-1 * p3 +	-1 * p4 +	-1 * p5 +	-1 * p6	= 0
-1 * p1 +	-1 * p2 +	3 * p3 +	3 * p4 +	-2 * p5 +	-2 * p6	= 0
0 * p1 +	-2 * p2 +	0 * p3 +	-2 * p4 +	3 * p5 +	0 * p6	= 0
-2 * p1 +	0 * p2 +	-2 * p3 +	0 * p4 +	0 * p5 +	3 * p6	= 0
	p2=.25		p4=.35	p5=.40		

Optimized Interleave (OI)

Constraints:

1. Prefix
2. Unbiased

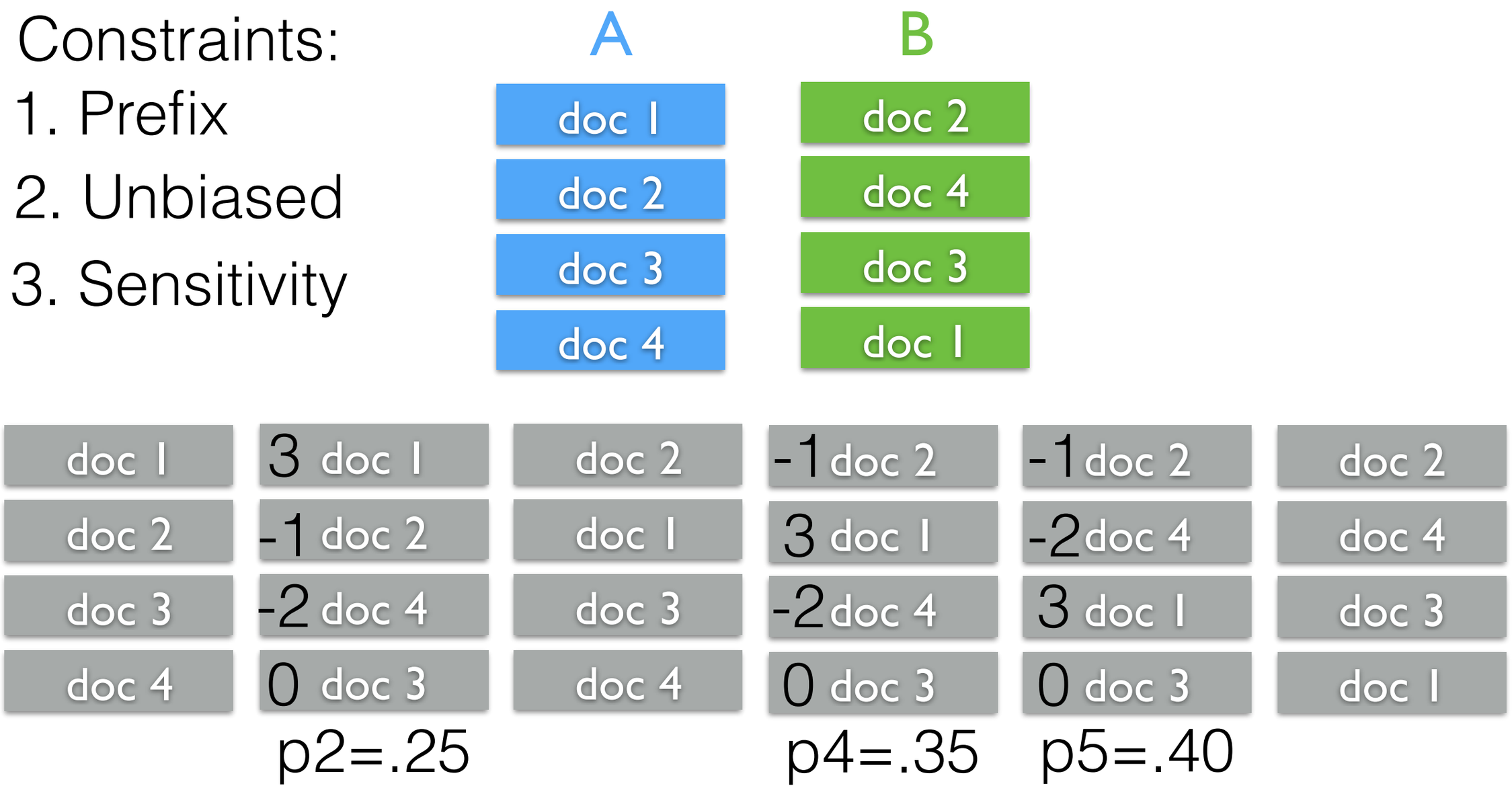
A	B
doc 1	doc 2
doc 2	doc 4
doc 3	doc 3
doc 4	doc 1

doc 1	3 doc 1	doc 2	-1 doc 2	-1 doc 2	doc 2
doc 2	-1 doc 2	doc 1	3 doc 1	-2 doc 4	doc 4
doc 3	-2 doc 4	doc 3	-2 doc 4	3 doc 1	doc 3
doc 4	0 doc 3	doc 4	0 doc 3	0 doc 3	doc 1
	p2=.25		p4=.35	p5=.40	

Optimized Interleave (OI)

Constraints:

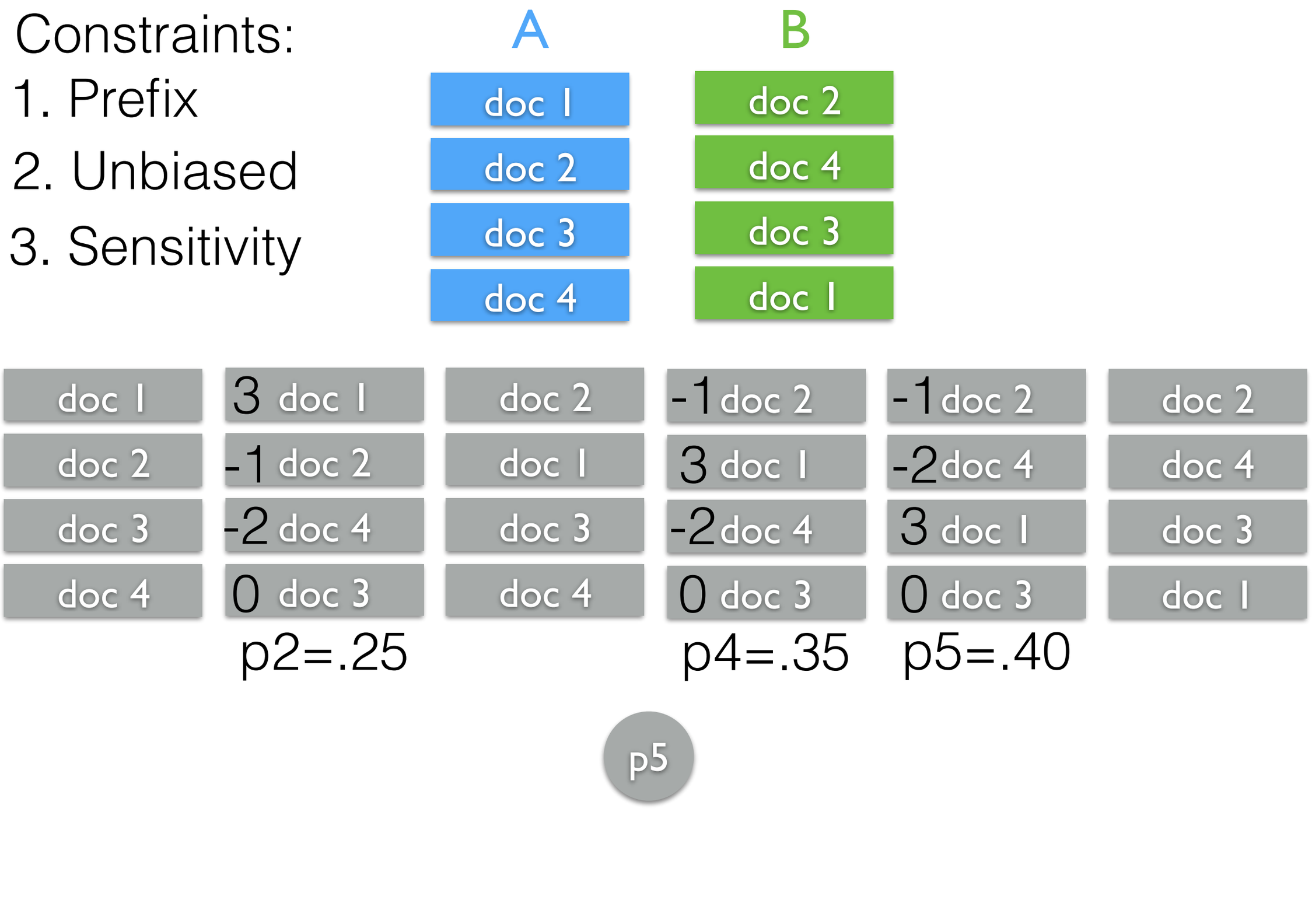
1. Prefix
2. Unbiased
3. Sensitivity



Optimized Interleave (OI)

Constraints:

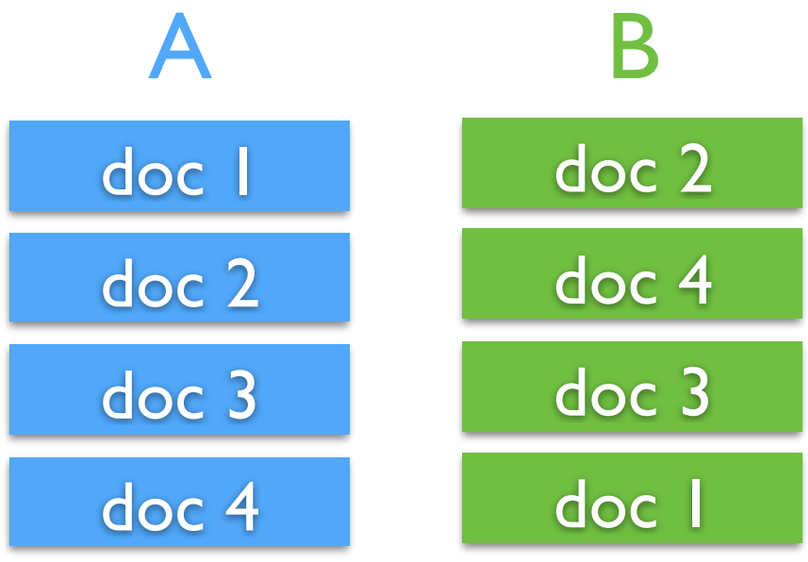
1. Prefix
2. Unbiased
3. Sensitivity



Optimized Interleave (OI)

Constraints:

1. Prefix
2. Unbiased
3. Sensitivity



-1 doc 2

-2 doc 4

3 doc 1

0 doc 3

Optimized Interleave (OI)

Constraints:

1. Prefix
2. Unbiased
3. Sensitivity

A	B
doc 1	doc 2
doc 2	doc 4
doc 3	doc 3
doc 4	doc 1

-1 doc 2

-2 doc 4

3 doc 1

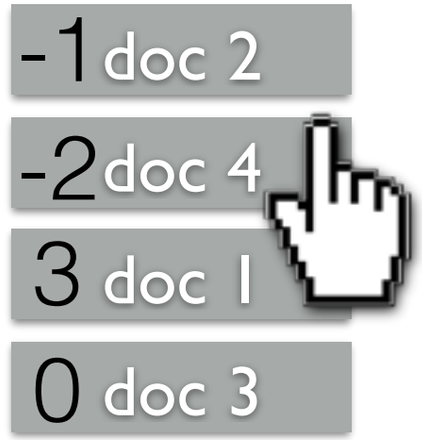
0 doc 3

Optimized Interleave (OI)

Constraints:

1. Prefix
2. Unbiased
3. Sensitivity

A	B
doc 1	doc 2
doc 2	doc 4
doc 3	doc 3
doc 4	doc 1



Inference:

B > A

Comparison Methods

- **Team Draft Interleave (TD)**
  F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM ’08. ACM Press, 2008.

- **Team Draft Multileave (TDM)**
  our multileave extension of TD

- **Optimized Interleave (OI)**
  F. Radlinski and N. Craswell. Optimized interleaving for online retrieval evaluation. In WSDM ’13. ACM Press, 2013.

- **Optimized Multileave (OM)**
  our multileave extension of OI

Comparison Methods

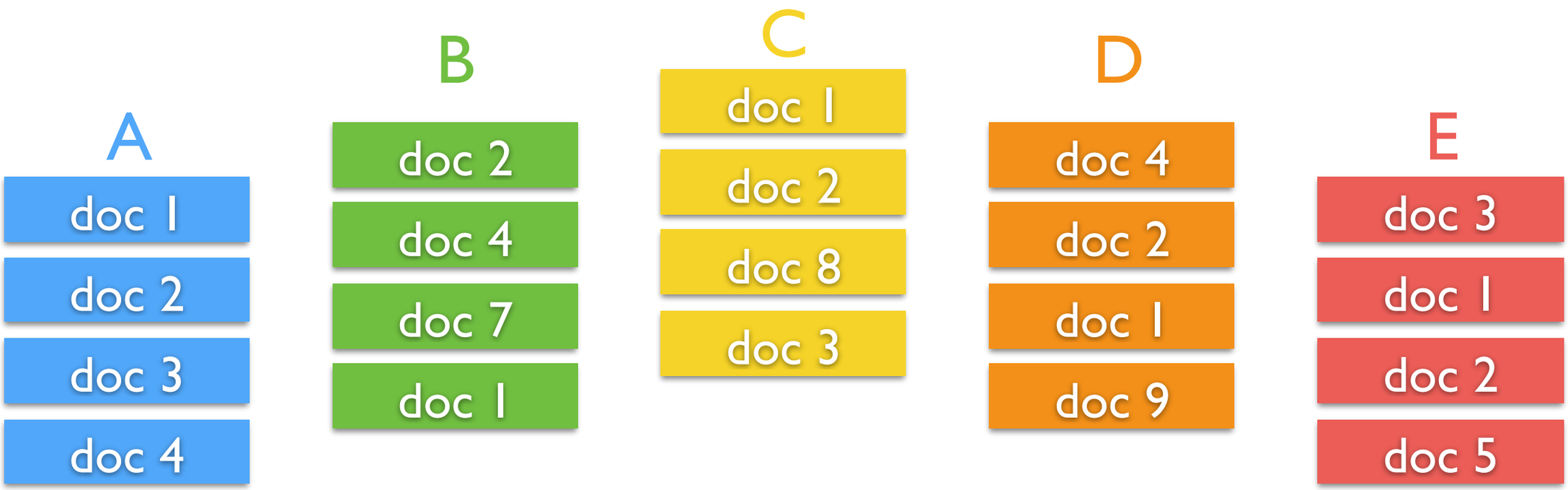
- **Team Draft Interleave (TD)**
  F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In CIKM ’08. ACM Press, 2008.

- **Team Draft Multileave (TDM)**
  our multileave extension of TD

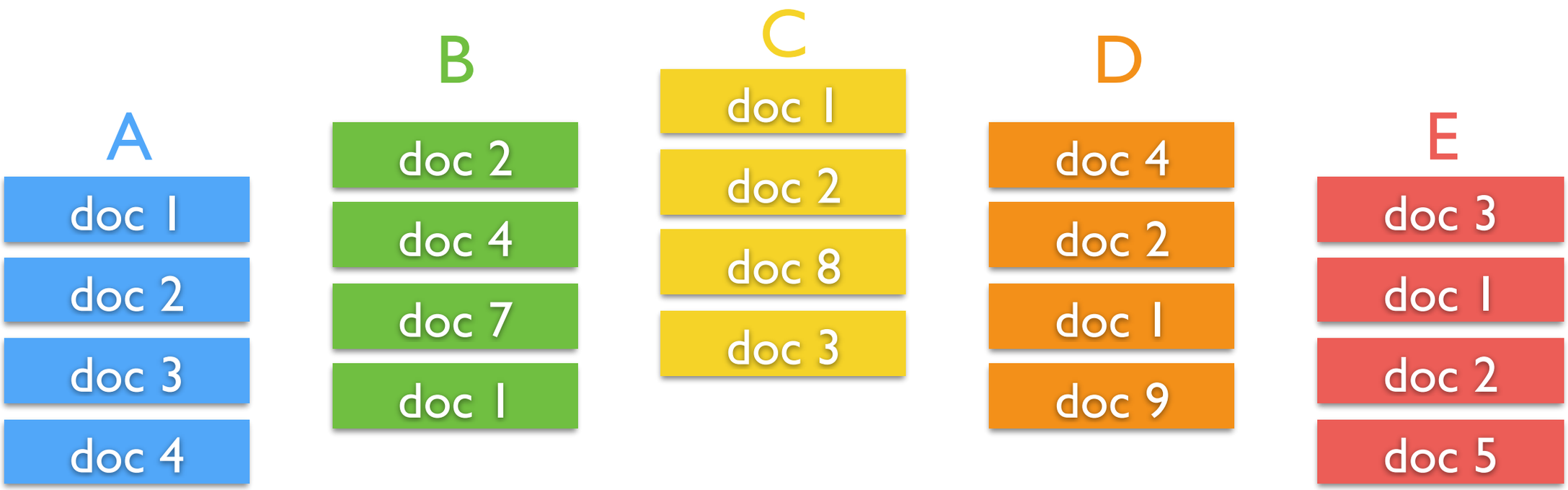
- **Optimized Interleave (OI)**
  F. Radlinski and N. Craswell. Optimized interleaving for online retrieval evaluation. In WSDM ’13. ACM Press, 2013.

- **Optimized Multileave (OM)**
  our multileave extension of OI

Optimized Multileave (OM)

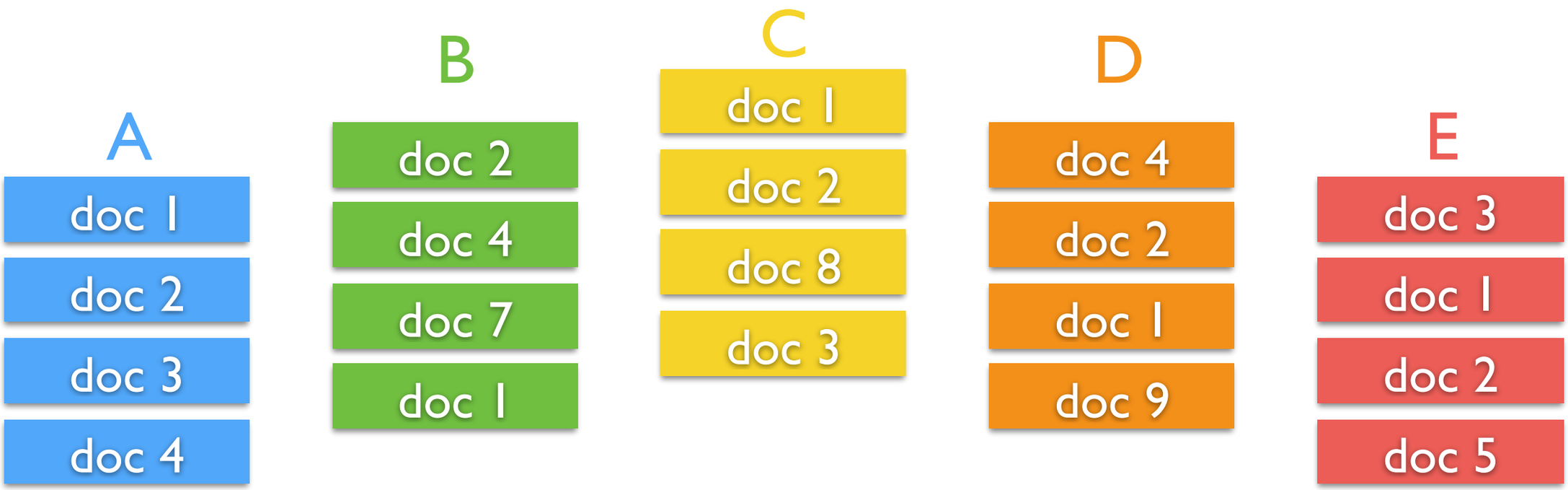


Optimized Multileave (OM)



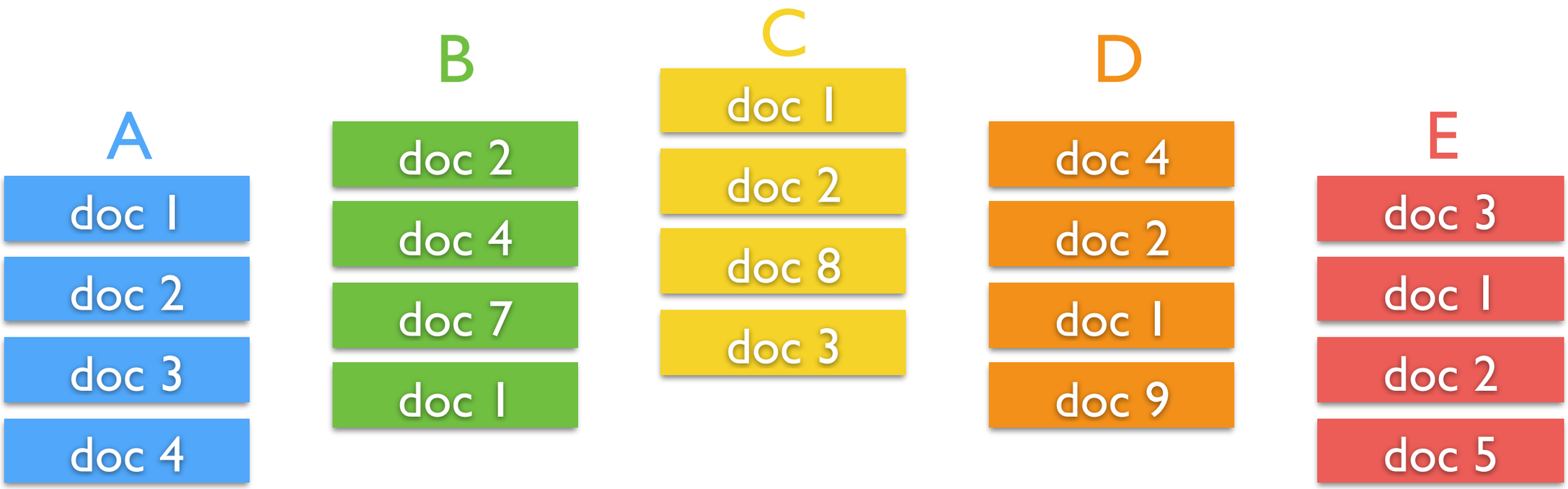
- Prefix constraint: too many multileavings

Optimized Multileave (OM)



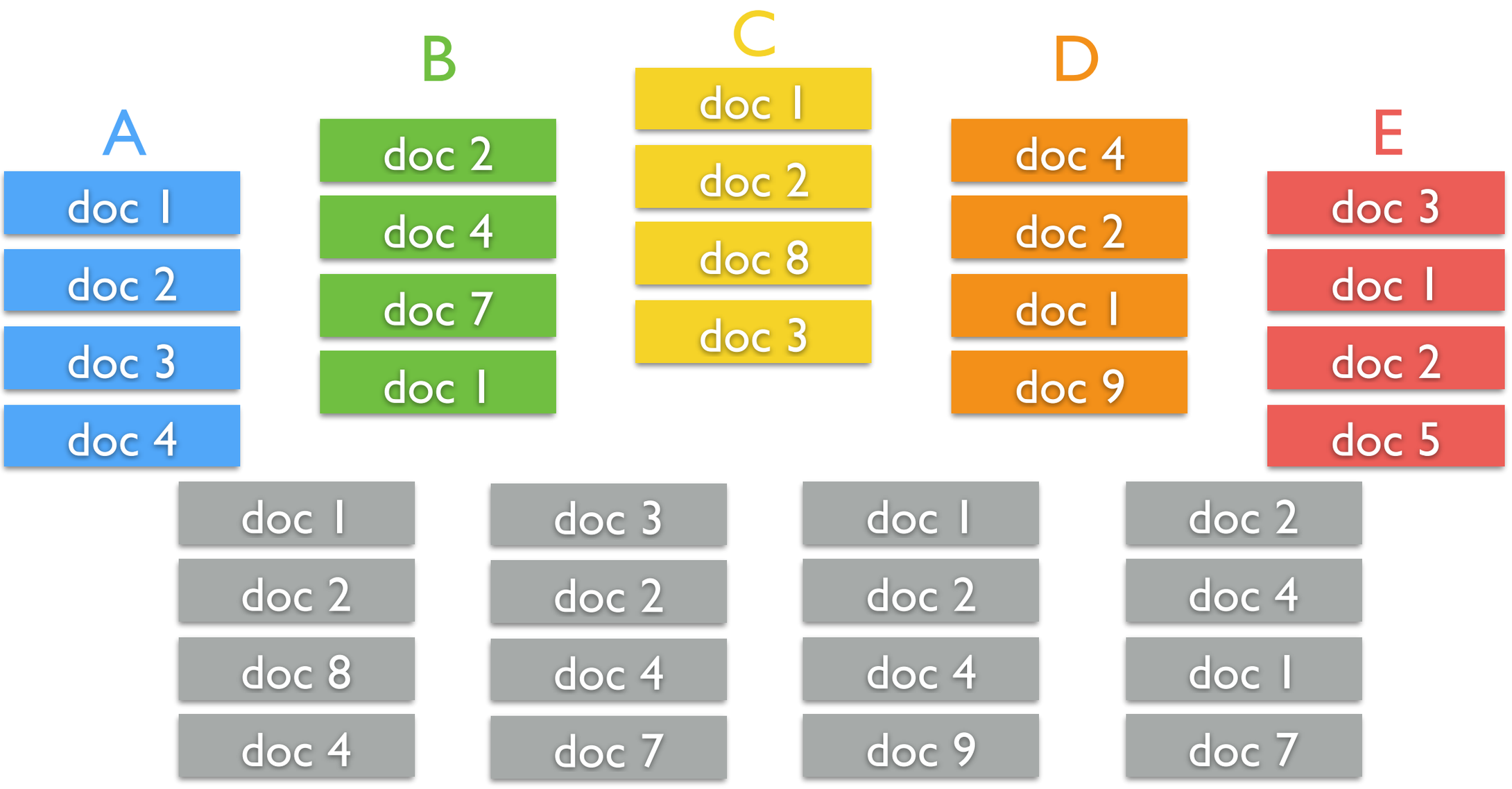
- Prefix constraint: too many multileavings
  - Sampling

Optimized Multileave (OM)

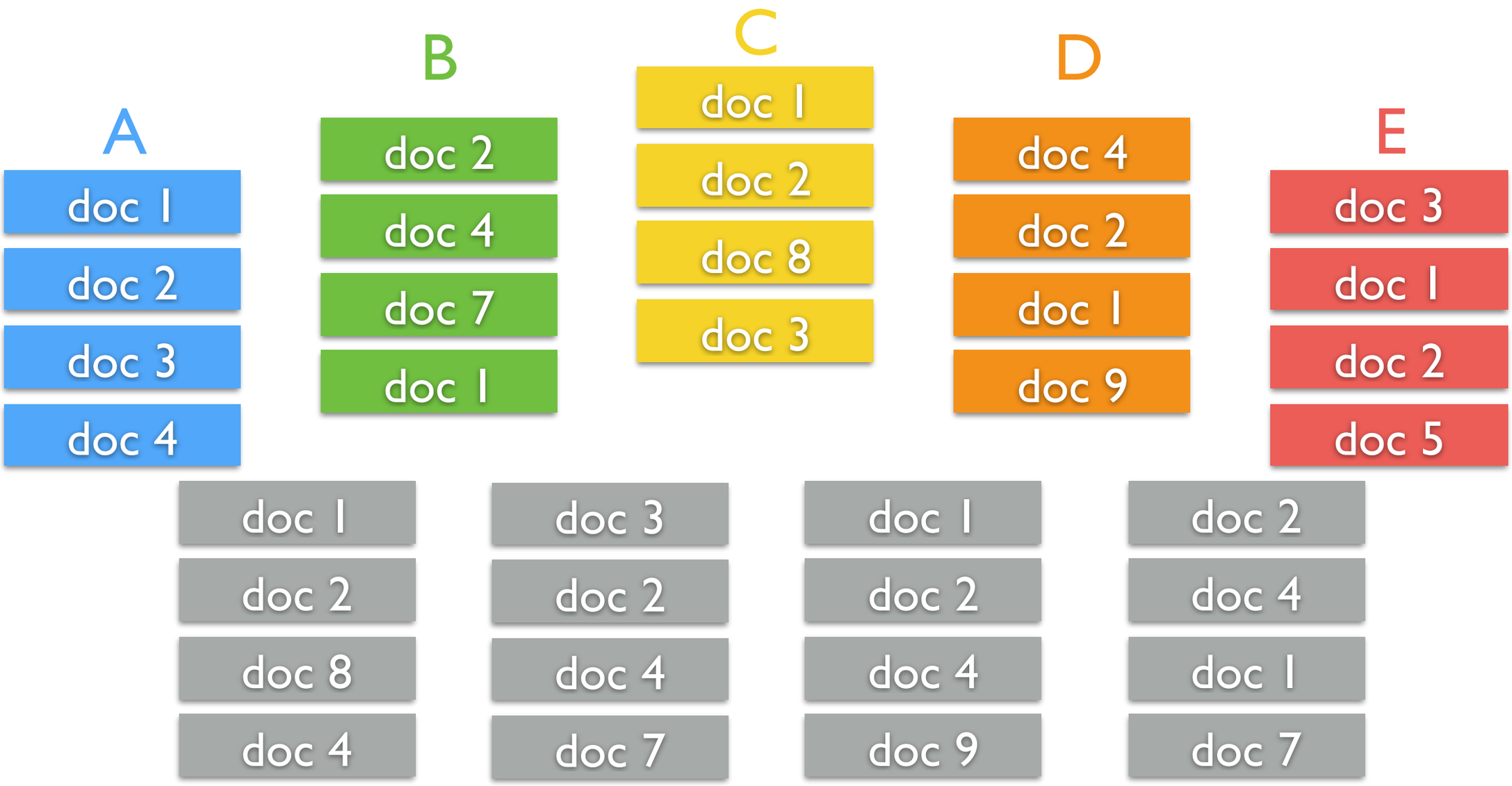


- Prefix constraint: too many multileavings
  - Sampling
  - In expectation unbiased

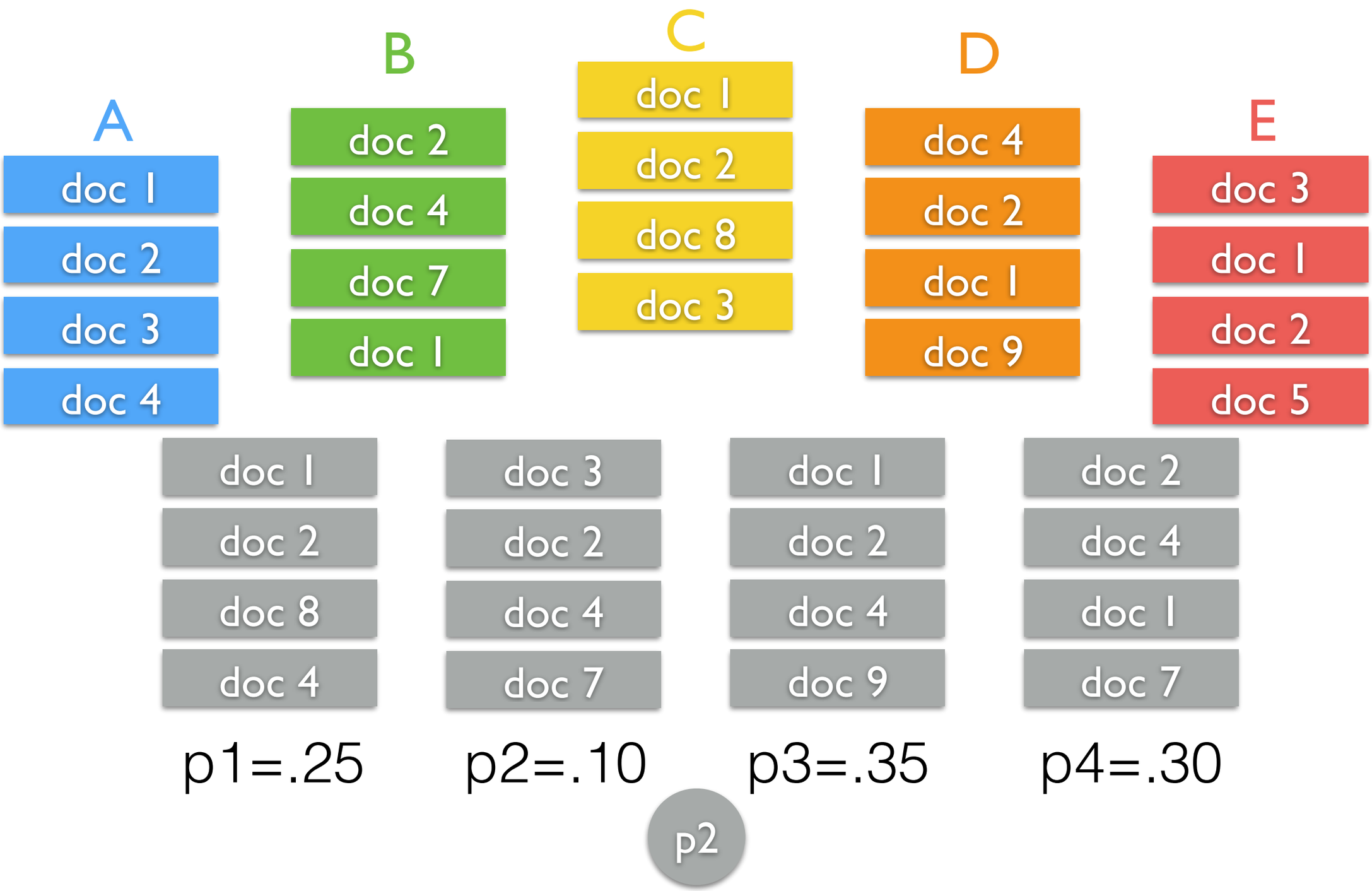
Optimized Multileave (OM)



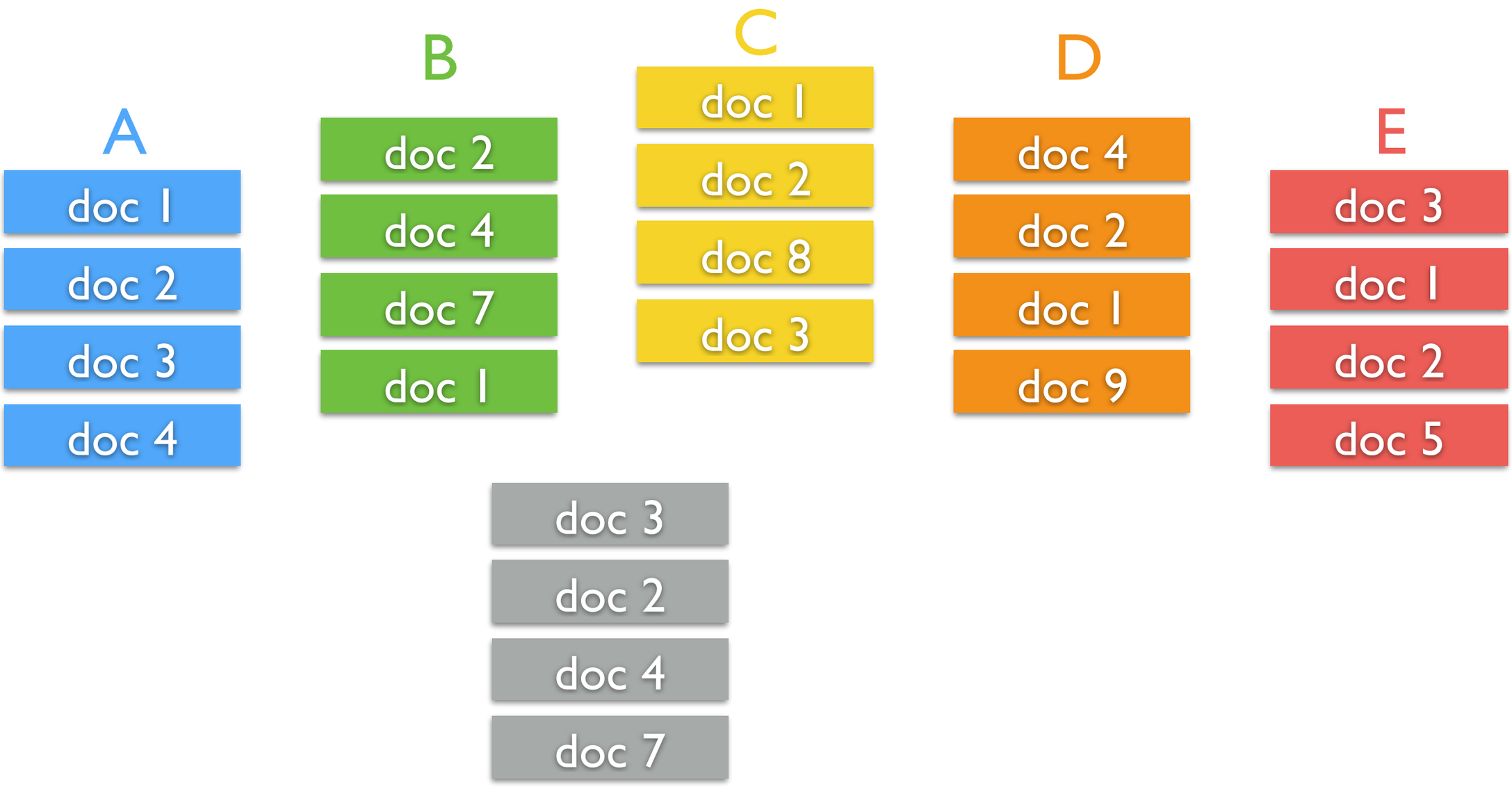
Optimized Multileave (OM)



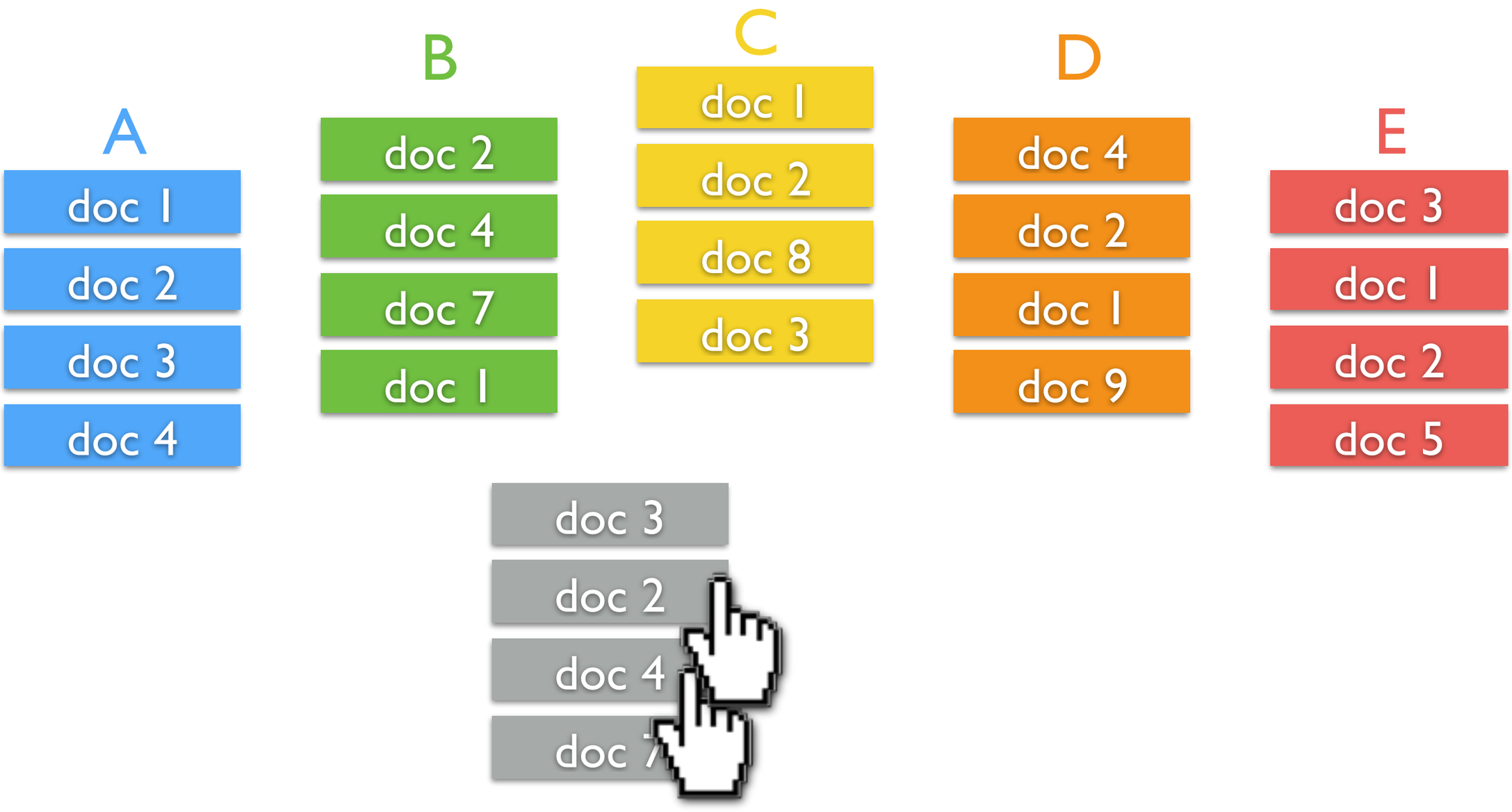
Optimized Multileave (OM)



Optimized Multileave (OM)



Optimized Multileave (OM)



Optimized Multileave (OM)

A

doc 1

doc 2

doc 3

doc 4

1/2 + 1/4

B

doc 2

doc 4

doc 7

doc 1

1/1 + 1/2

C

doc 1

doc 2

doc 8

doc 3

1/2 + 1/5

D

doc 4

doc 2

doc 1

doc 9

1/2 + 1/1

E

doc 3

doc 1

doc 2

doc 5

1/3 + 1/5

doc 3

doc 2

doc 4

doc 7

Optimized Multileave (OM)

A

doc 1
doc 2
doc 3
doc 4

1/2 + 1/4

B

doc 2
doc 4
doc 7
doc 1

1/1 + 1/2

C

doc 1
doc 2
doc 8
doc 3

1/2 + 1/5

D

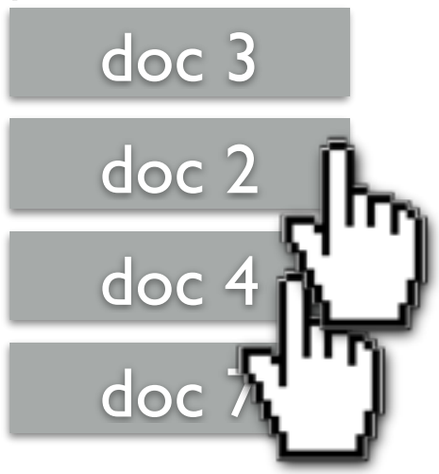
doc 4
doc 2
doc 1
doc 9

1/2 + 1/1

E

doc 3
doc 1
doc 2
doc 5

1/3 + 1/5



Inference:

B & D > A & E & C

A > E & C

C > E

Research Questions

Research Questions

1. Can **multileaved** comparison methods identify preferences between rankers **faster than interleaved** comparison methods?

Research Questions

1. Can **multileaved** comparison methods identify preferences between rankers **faster than interleaved** comparison methods?

2. Does OM **scale** better with **the number of rankers** than TDM?

Research Questions

1. Can **multileaved** comparison methods identify preferences between rankers **faster than interleaved** comparison methods?

2. Does OM **scale** better with **the number of rankers** than TDM?

3. How does the **sensitivity** of multileaving methods compare to that of interleaving methods?

Experimental Setup

- LETOR Data (queries, documents represented by features, relevance judgments)
- A ranker is a single feature (BM25, Pagerank, …)
- Simulate clicks using *cascade click model*
- Measure error

$E_{bin} =$ Fraction of incorrect preferences

with: ground truth from NDCG preferences

Faster?

Can **multileaved** comparison methods identify preferences between rankers **faster than interleaved** comparison methods?

Faster?

- 5 rankers

Can **multileaved** comparison methods identify preferences between rankers **faster than interleaved** comparison methods?

Faster?

- 5 rankers

$$P_{ij} =$$

	R1	R2	R3	R4	R5
R1	0	**+1**	**-1**	**+1**	**+1**
R2	-1	0	**+1**	**-1**	**+1**
R3	+1	-1	0	**+1**	**+1**
R4	-1	+1	-1	0	**+1**
R5	-1	-1	-1	-1	0

Can **multileaved** comparison methods identify preferences between rankers **faster than interleaved** comparison methods?

Faster?

- 5 rankers
- 5k queries

$$P_{ij} =$$

	R1	R2	R3	R4	R5
R1	0	**+1**	**-1**	**+1**	**+1**
R2	-1	0	**+1**	**-1**	**+1**
R3	+1	-1	0	**+1**	**+1**
R4	-1	+1	-1	0	**+1**
R5	-1	-1	-1	-1	0

Can **multileaved** comparison methods identify preferences between rankers **faster than interleaved** comparison methods?

Faster?

- 5 rankers
- 5k queries
- Updates:

$$P_{ij} =$$

	R1	R2	R3	R4	R5
R1	0	**+1**	**-1**	**+1**	**+1**
R2	-1	0	**+1**	**-1**	**+1**
R3	+1	-1	0	**+1**	**+1**
R4	-1	+1	-1	0	**+1**
R5	-1	-1	-1	-1	0

Can **multileaved** comparison methods identify preferences between rankers **faster than interleaved** comparison methods?

Faster?

Can **multileaved** comparison methods identify preferences between rankers **faster than interleaved** comparison methods?

- 5 rankers
- 5k queries
- Updates:
- Interleaving (TD, OI):

$$P_{ij} =$$

	R1	R2	R3	R4	R5
R1	0	+1	-1	+1	+1
R2	-1	0	+1	-1	+1
R3	+1	-1	0	+1	+1
R4	-1	+1	-1	0	+1
R5	-1	-1	-1	-1	0

Faster?

Can **multileaved** comparison methods identify preferences between rankers **faster than interleaved** comparison methods?

- 5 rankers
- 5k queries
- Updates:
- Interleaving (TD, OI):
  - 10 queries for the whole matrix

$$P_{ij} =$$

	R1	R2			
R1	0	+1	-1	+1	+1
R2	-1	0	+1	-1	+1
R3	+1	-1	0	+1	+1
R4	-1	+1	-1	0	+1
R5	-1	-1	-1	-1	0

Faster?

Can **multileaved** comparison methods identify preferences between rankers **faster than interleaved** comparison methods?

- 5 rankers
- 5k queries
- Updates:
- Interleaving (TD, OI):
  - 10 queries for the whole matrix
- Multileaving (TDM, OM):

$$P_{ij} =$$

	R1	R2	R3	R4	R5
R1	0	+1	-1	+1	+1
R2	-1	0	+1	-1	+1
R3	+1	-1	0	+1	+1
R4	-1	+1	-1	0	+1
R5	-1	-1	-1	-1	0

Faster?

Can **multileaved** comparison methods identify preferences between rankers **faster than interleaved** comparison methods?

- 5 rankers
- 5k queries
- Updates:
- Interleaving (TD, OI):
  - 10 queries for the whole matrix
- Multileaving (TDM, OM):
  - 1 queries for the whole matrix

$$P_{ij} =$$

	R1	R2	R3	R4	R5
R1	0	+1	-1	+1	+1
R2	-1	0	+1	-1	+1
R3	+1	-1	0	+1	+1
R4	-1	+1	-1	0	+1
R5	-1	-1	-1	-1	0

Faster?

Can **multileaved** comparison methods identify preferences between rankers **faster than interleaved** comparison methods?

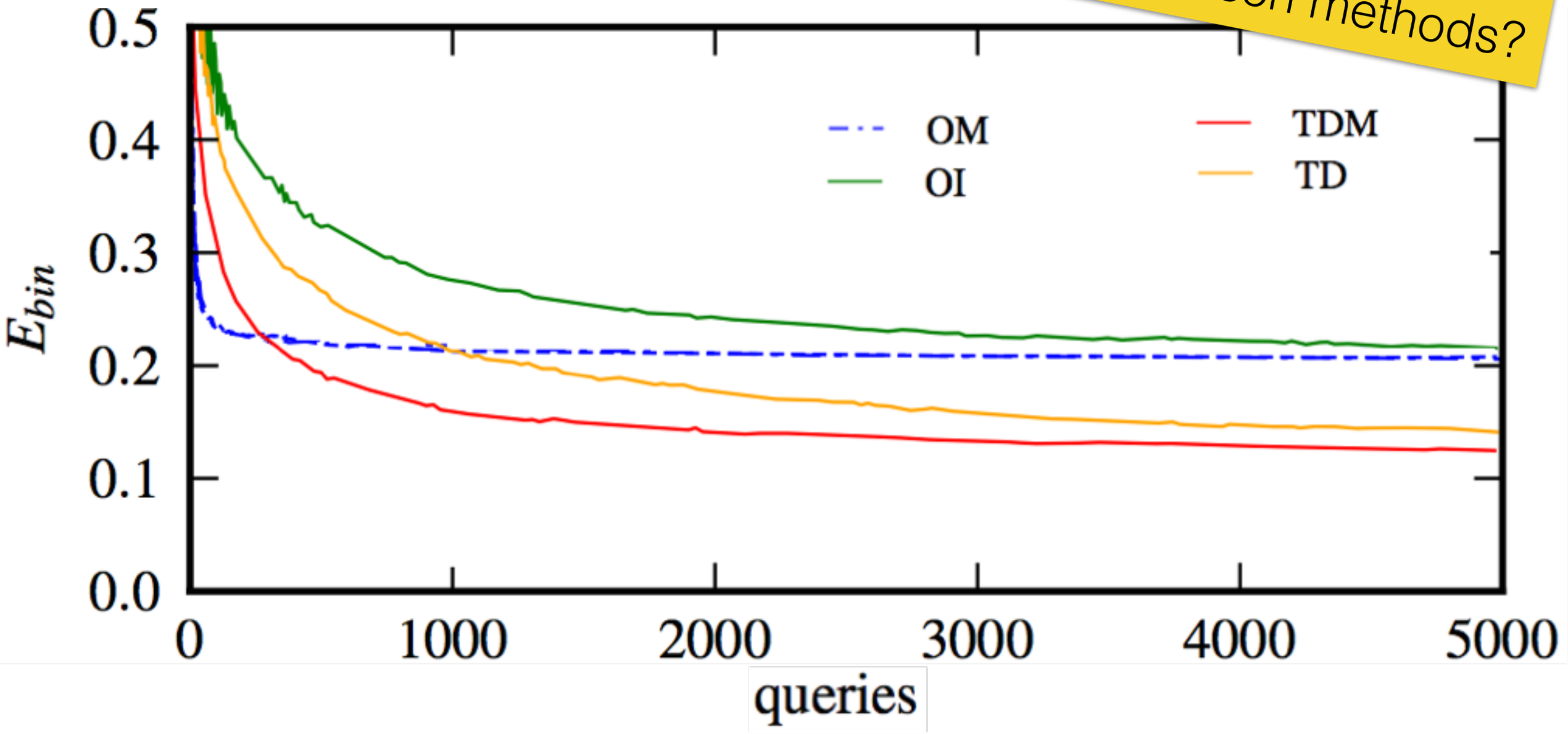
- 5 rankers
- 5k queries
- Updates:
- Interleaving (TD, OI):
  - 10 queries for the whole matrix
- Multileaving (TDM, OM):
  - 1 queries for the whole matrix
  - 10 times faster?

$$P_{ij} =$$

	R1	R2			
R1	0	+1	-1	+1	+1
R2	-1	0	+1	-1	+1
R3	+1	-1	0	+1	+1
R4	-1	+1	-1	0	+1
R5	-1	-1	-1	-1	0

Faster?

Can **multileaved** comparison methods identify preferences between rankers **faster** **than interleaved** comparison methods?



Scaling

Does OM **scale** better with **the number of rankers** than TDM?

A
- doc 1
- doc 2
- doc 3
- doc 4

B
- doc 2
- doc 4
- doc 7
- doc 1

C
- doc 1
- doc 2
- doc 8
- doc 3

D
- doc 4
- doc 2
- doc 1
- doc 9

E
- doc 3
- doc 1
- doc 2
- doc 5

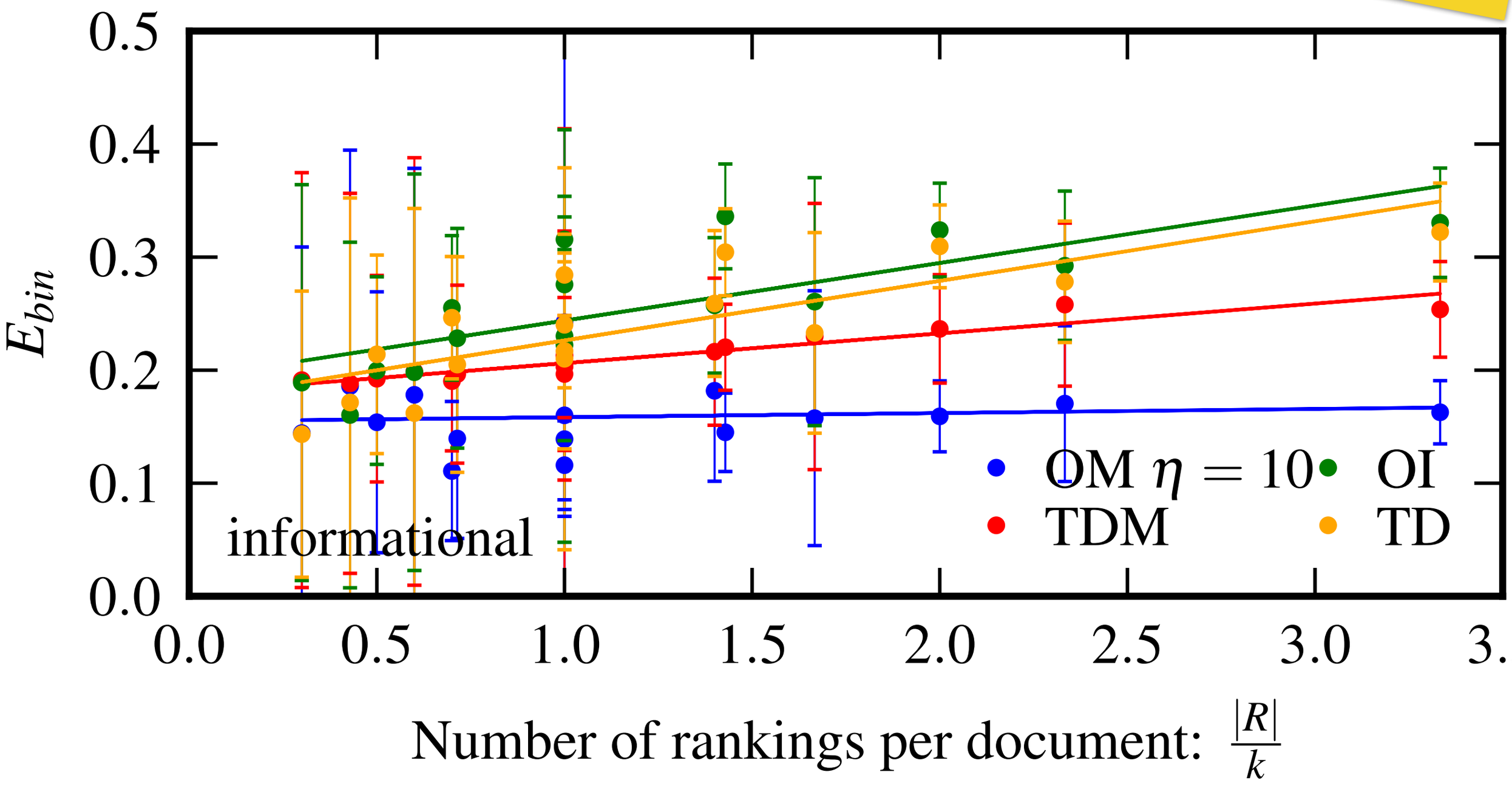
- TDM: not all rankings can be represented
- Rankers/slots (5/3)

Scaling

Does OM **scale** better with **the number of rankers** than TDM?

Scaling

Does OM **scale** better with **the number of rankers** than TDM?



Sensitivity

How does the **sensitivity** of multileaving methods compare to that of interleaving methods?

- Can small differences be detected?

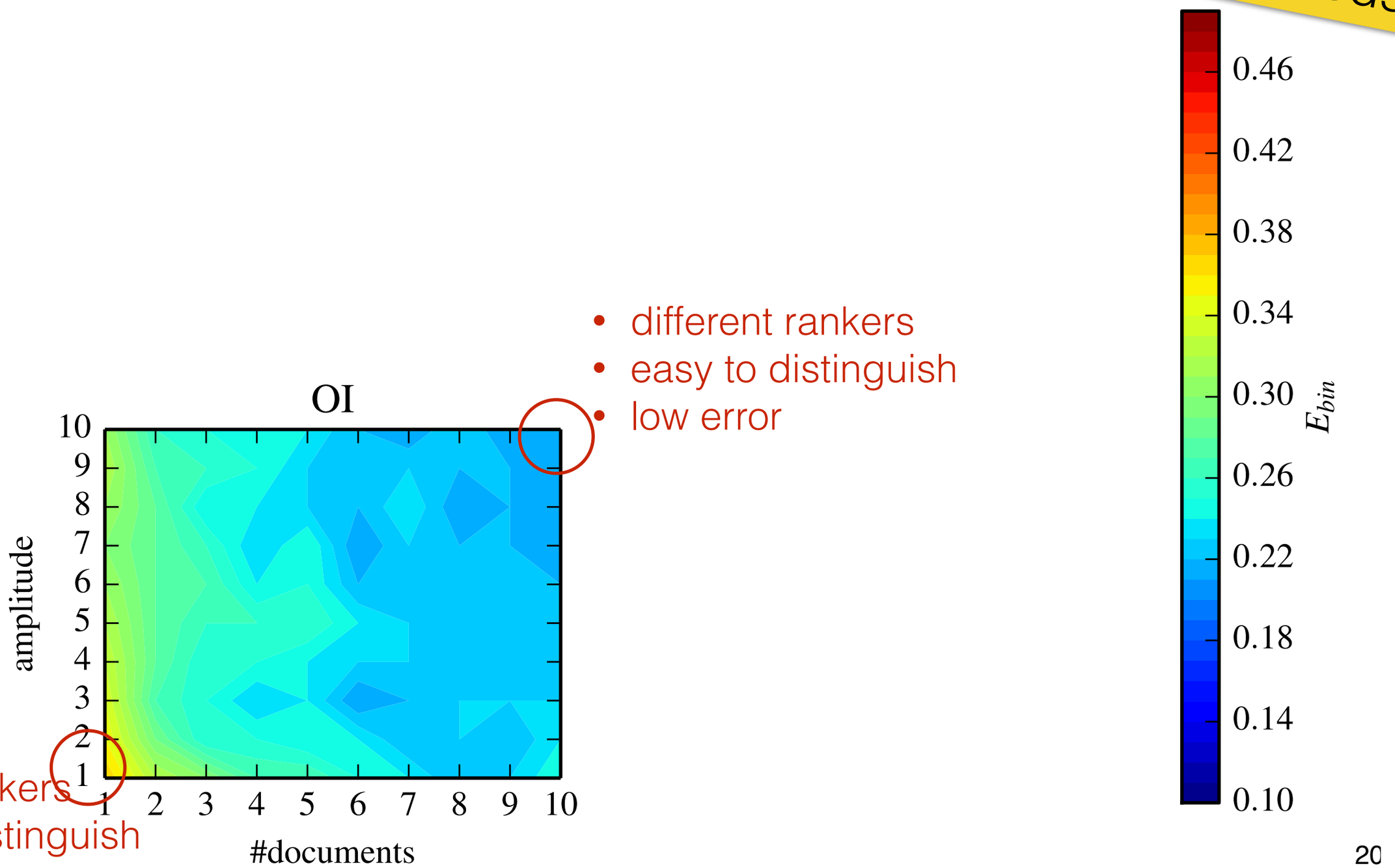
- Control variation among compared rankers
  - Number of changed documents
  - Amount of change

Sensitivity

How does the **sensitivity** of multileaving methods compare to that of interleaving methods?

Sensitivity

How does the **sensitivity** of multileaving methods compare to that of interleaving methods?

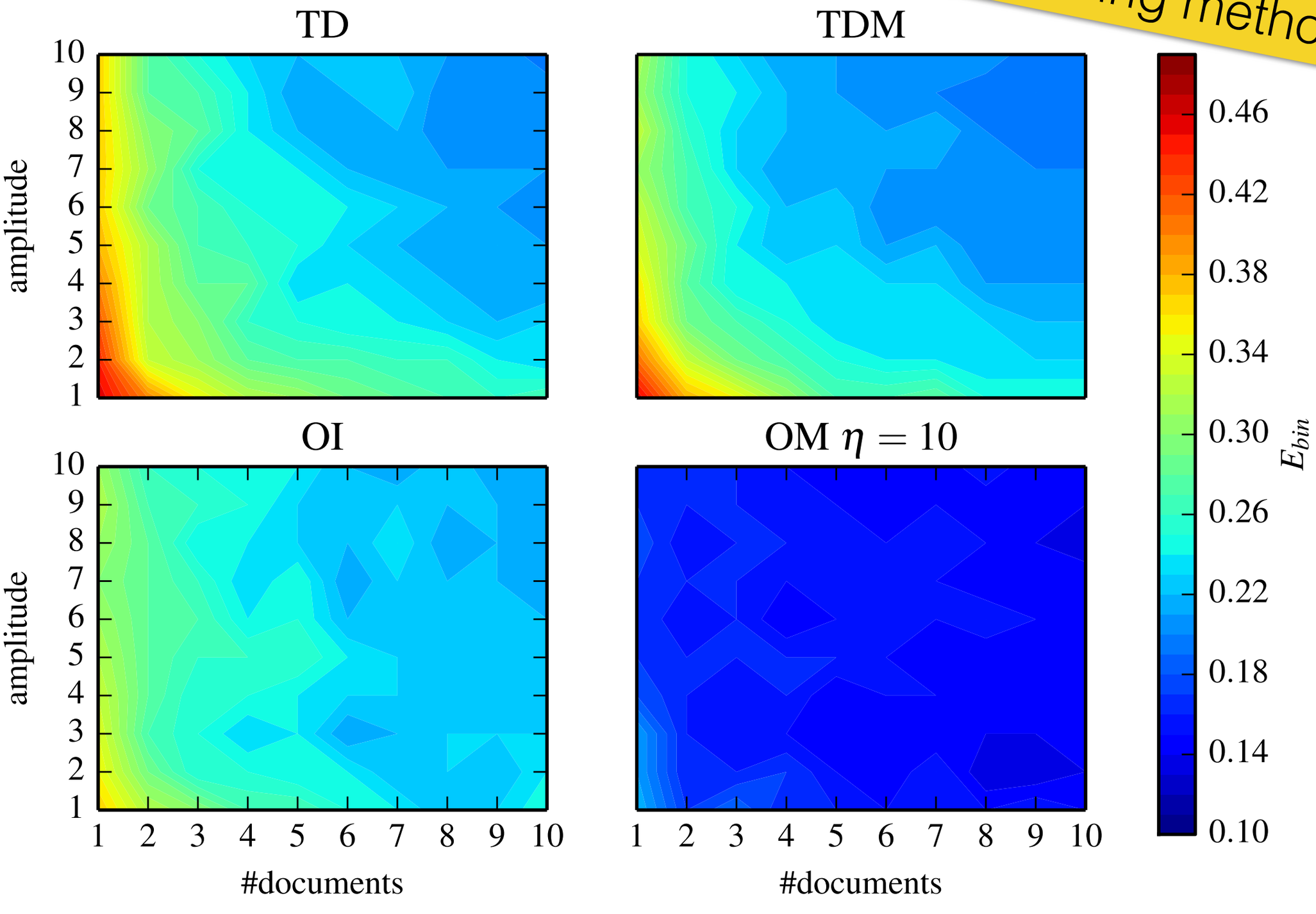


- different rankers
- easy to distinguish
- low error

- similar rankers
- hard to distinguish
- high error

Sensitivity

How does the **sensitivity** of multileaving methods compare to that of interleaving methods?



Conclusions

Conclusions

- Multileave: new online evaluation paradigm

Conclusions

- Multileave: new online evaluation paradigm
- New algorithms:

Conclusions

- Multileave: new online evaluation paradigm
- New algorithms:
  - Team Draft Multileave (TDM) and Optimized Multileave (OM)

Conclusions

- Multileave: new online evaluation paradigm
- New algorithms:
  - Team Draft Multileave (TDM) and Optimized Multileave (OM)
- Experimental results:

Conclusions

- Multileave: new online evaluation paradigm
- New algorithms:
  - Team Draft Multileave (TDM) and Optimized Multileave (OM)
- Experimental results:
  - TDM/OM are faster

Conclusions

- Multileave: new online evaluation paradigm
- New algorithms:
  - Team Draft Multileave (TDM) and Optimized Multileave (OM)
- Experimental results:
  - TDM/OM are faster
  - OM scales better than TDM

Conclusions

- Multileave: new online evaluation paradigm
- New algorithms:
  - Team Draft Multileave (TDM) and Optimized Multileave (OM)
- Experimental results:
  - TDM/OM are faster
  - OM scales better than TDM
  - TDM/OM are more sensitive

Conclusions

- Multileave: new online evaluation paradigm
- New algorithms:
  - Team Draft Multileave (TDM) and Optimized Multileave (OM)
- Experimental results:
  - TDM/OM are faster
  - OM scales better than TDM
  - TDM/OM are more sensitive
- Future Work

Conclusions

- Multileave: new online evaluation paradigm
- New algorithms:
  - Team Draft Multileave (TDM) and Optimized Multileave (OM)
- Experimental results:
  - TDM/OM are faster
  - OM scales better than TDM
  - TDM/OM are more sensitive
- Future Work
  - Online Learning with Multileave Feedback

thank you

- Lerot: Online Learning to Rank Framework
  - Interleaving/Multileaving
  - Simulations
  - Learning methods

by Katja Hofmann and Anne Schuth

bitbucket.org/ilps/lerot