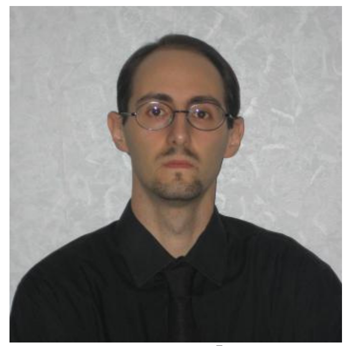
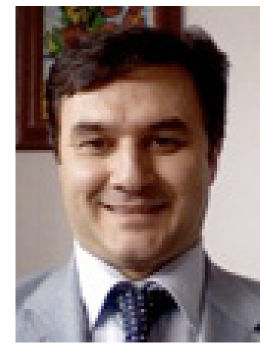
How to Effectively Use Topic Models for Software Engineering Tasks? An Approach Based on Genetic Algorithms





Annibale Panichella

Bogdan Dit

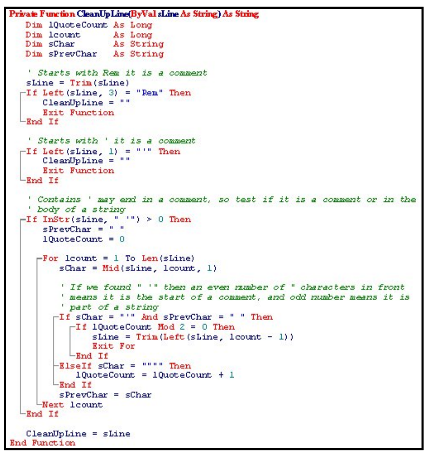
Rocco Oliveto

Massimiliano Di Penta

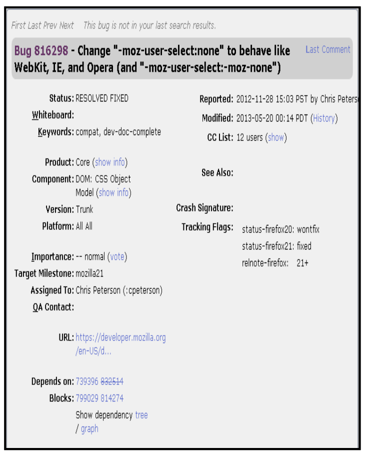
Denys Poshyvanyk

Andrea De Lucia

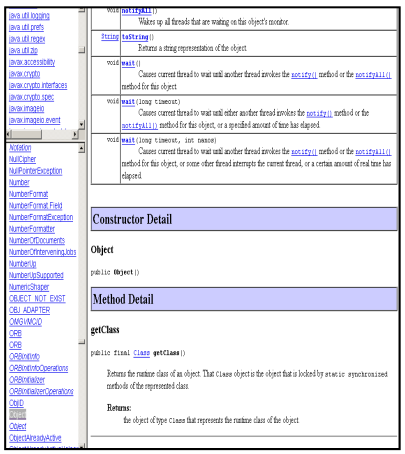
Source Code



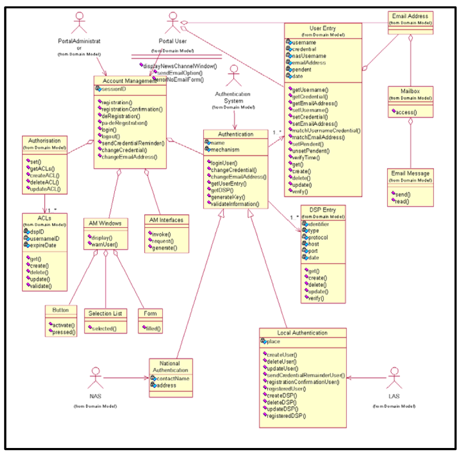
Bug Reports



Documentation



Design documents




Information Retrieval Techniques
Maintenance Tasks

Information Retrieval Techniques

Maintenance Tasks

Vector Space Model
Latent Semantic Indexing
Latent Dirichlet Allocation
Relational Topic Model

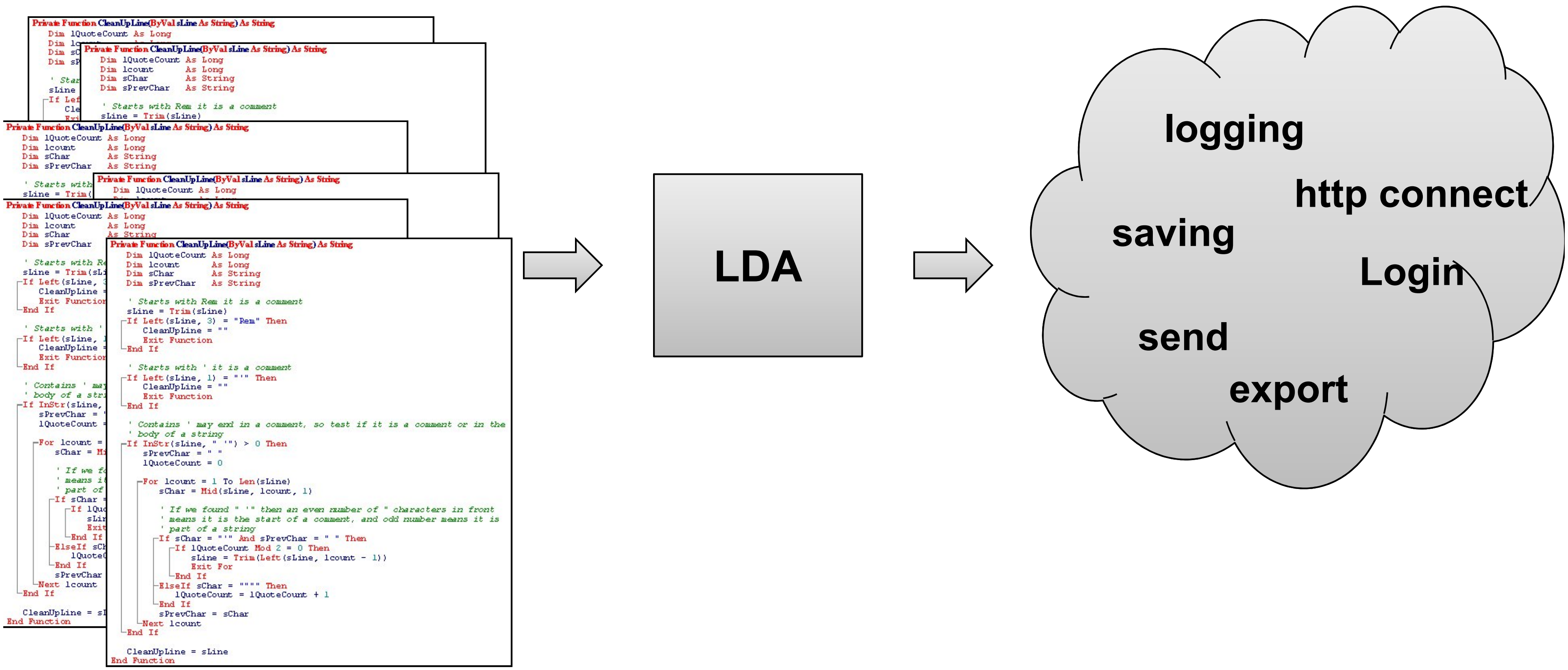
What is LDA?

Latent **D**irichlet **A**llocation (LDA)

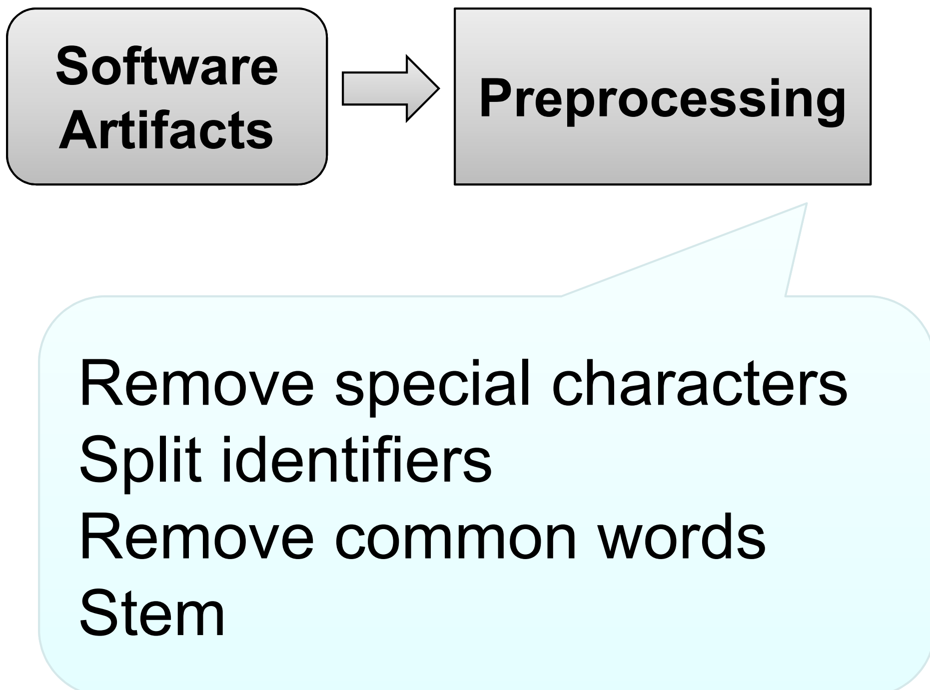
- Topic model that generates the distribution of latent topics from textual documents

Latent Dirichlet Allocation (LDA)

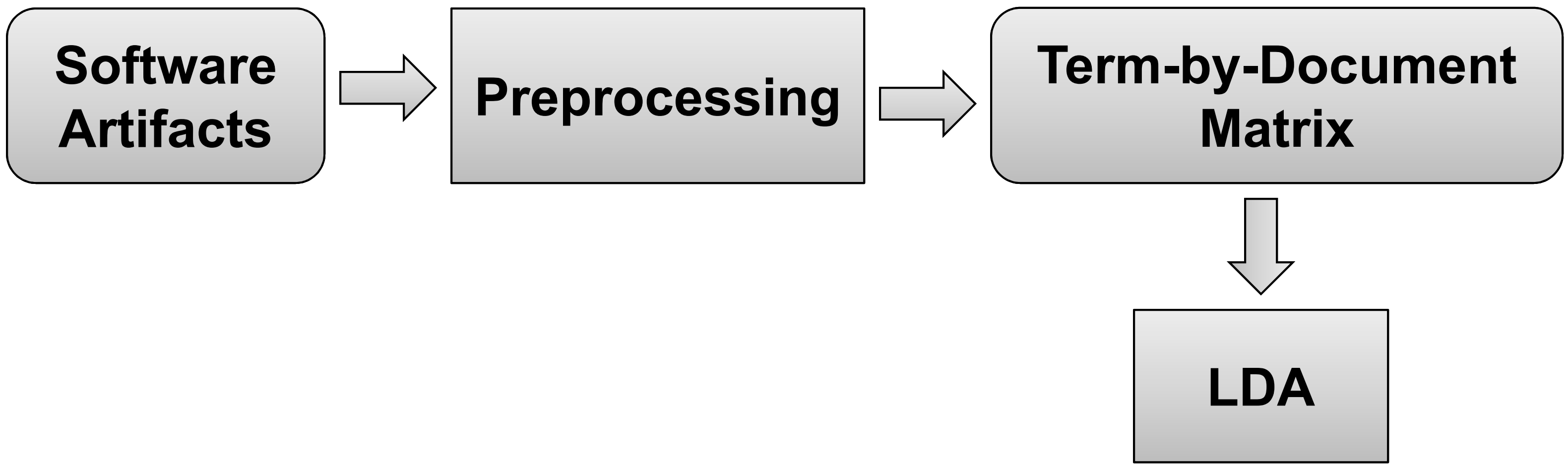
- Topic model that generates the distribution of latent topics from textual documents

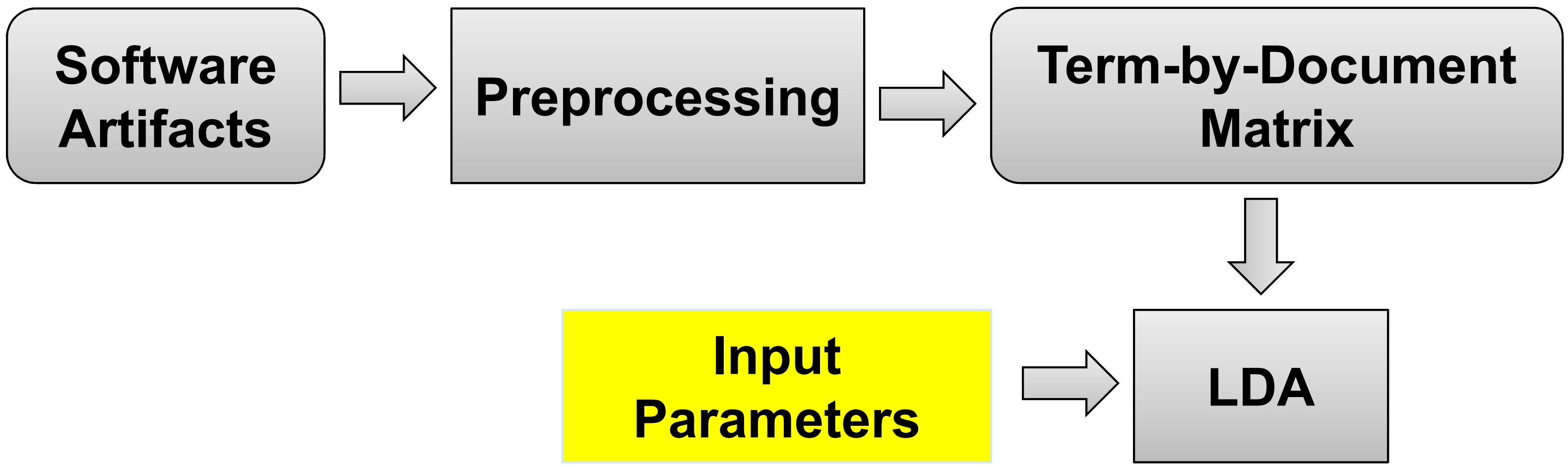


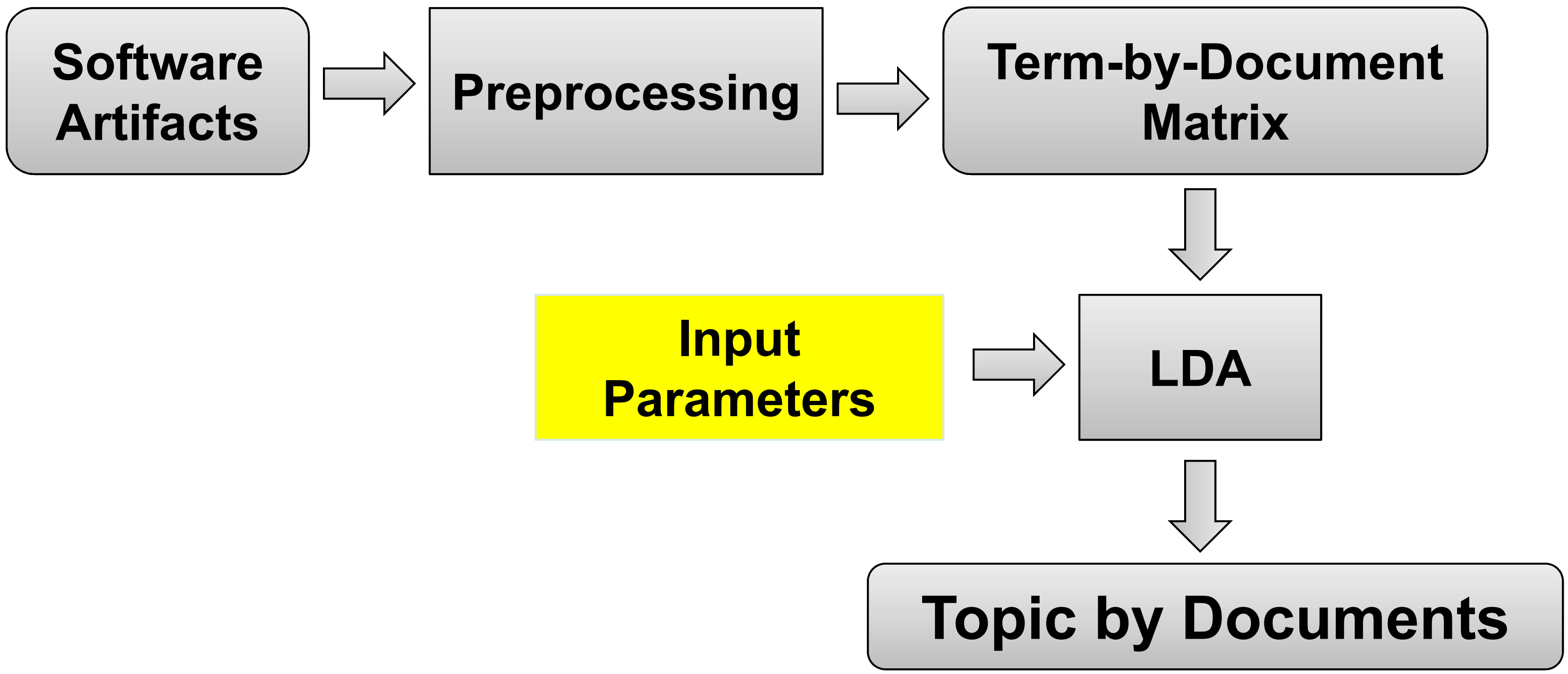
Software
Artifacts

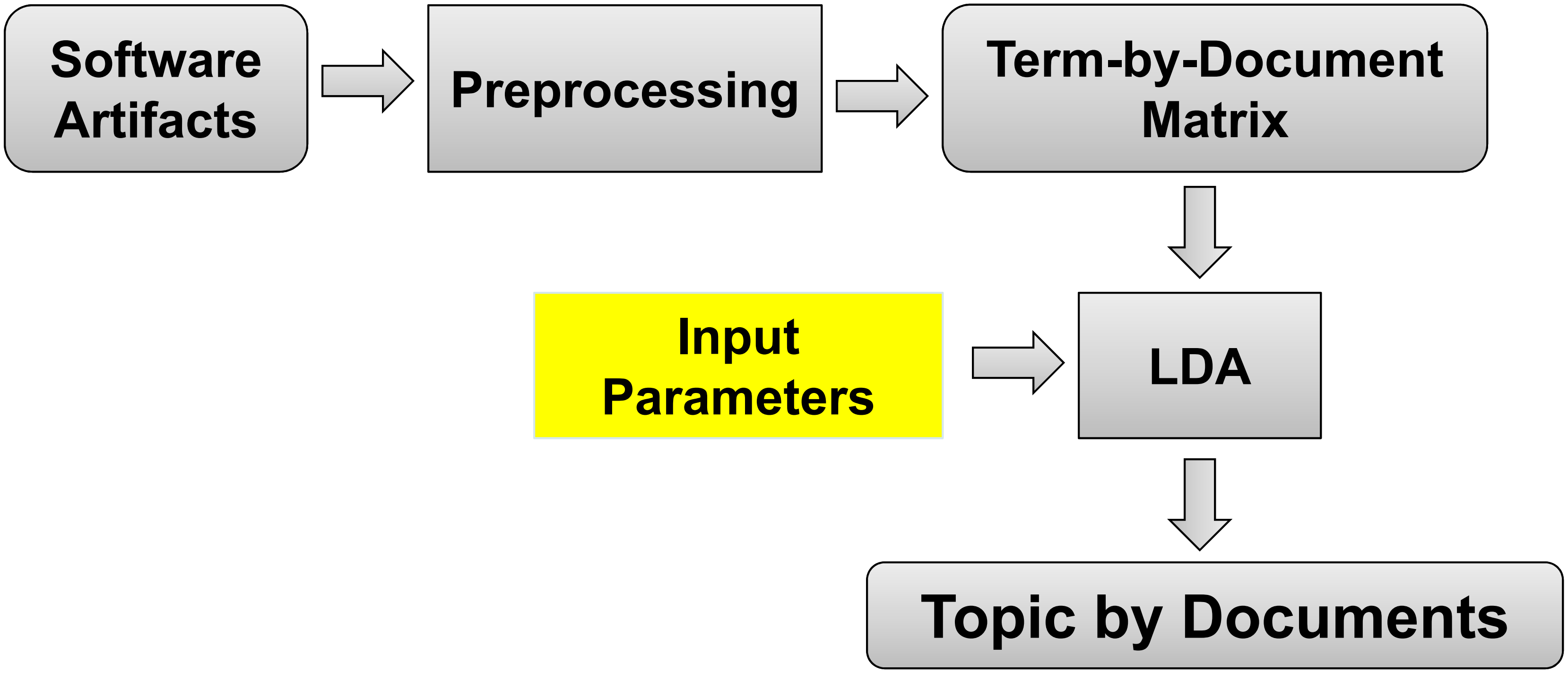

Software Artifacts
Preprocessing
Remove special characters
Split identifiers
Remove common words
Stem


Software Artifacts
Preprocessing
Term-by-Document Matrix

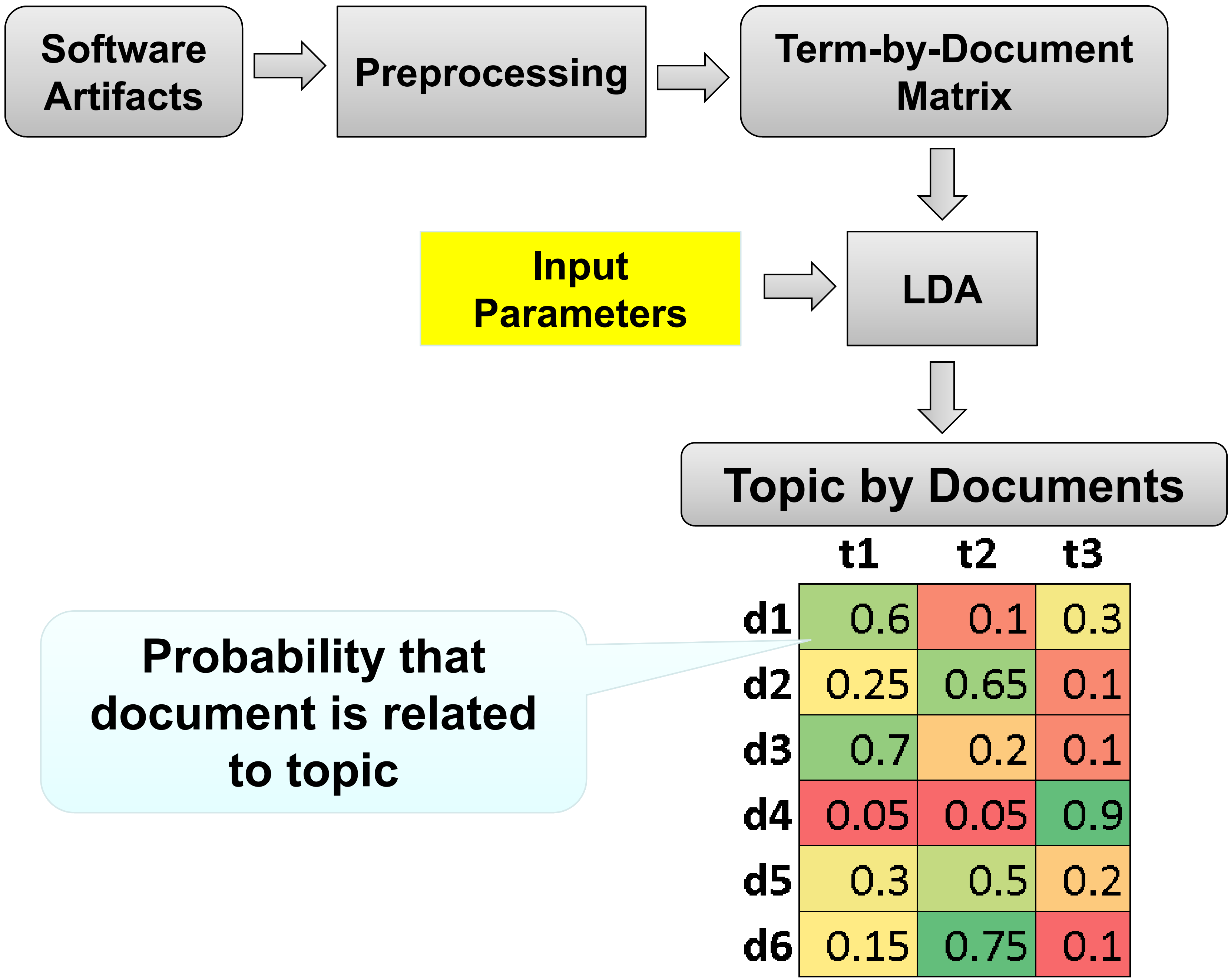

Software Artifacts
Preprocessing
Term-by-Document Matrix
LDA


Software Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA

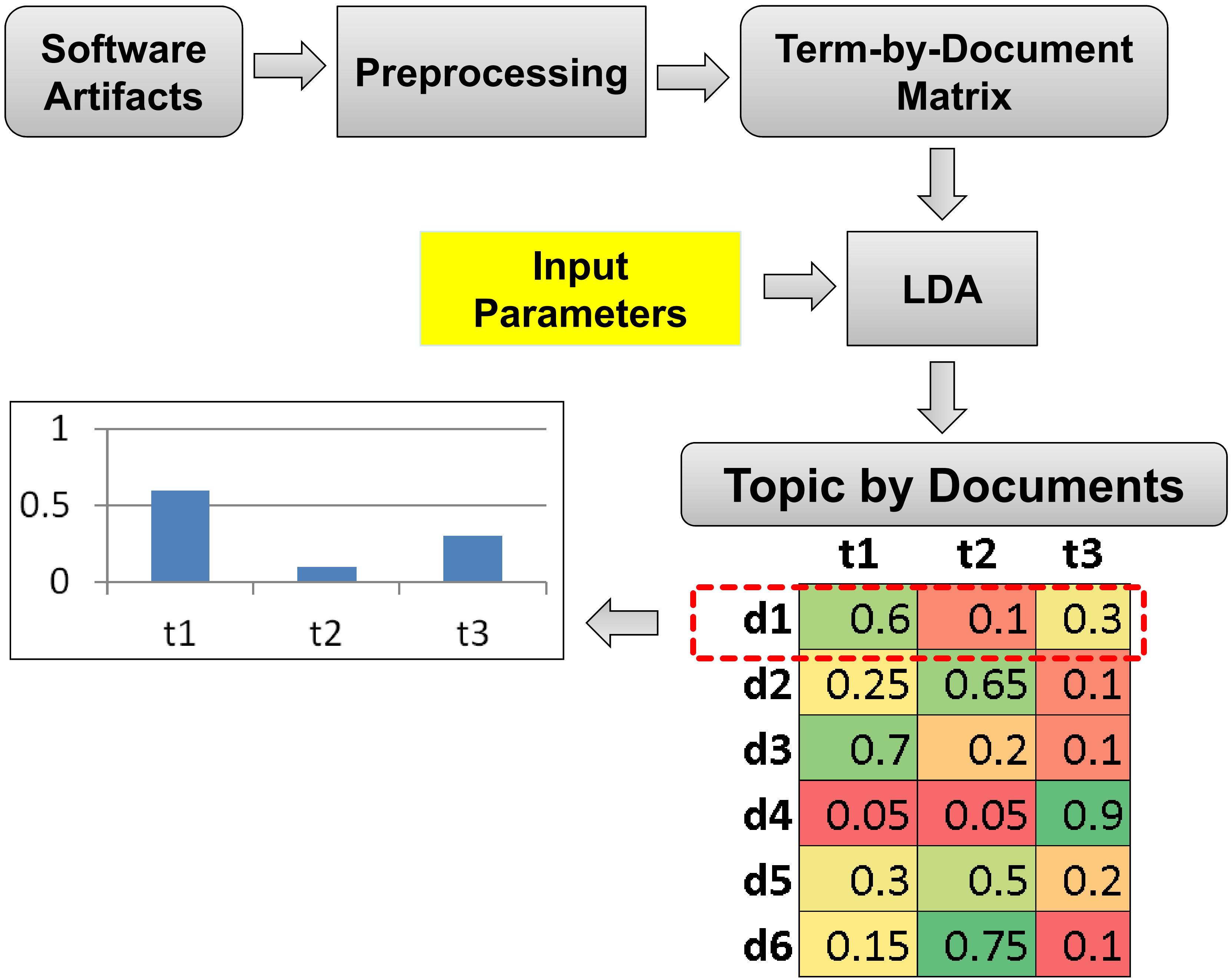

Software Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
Topic by Documents



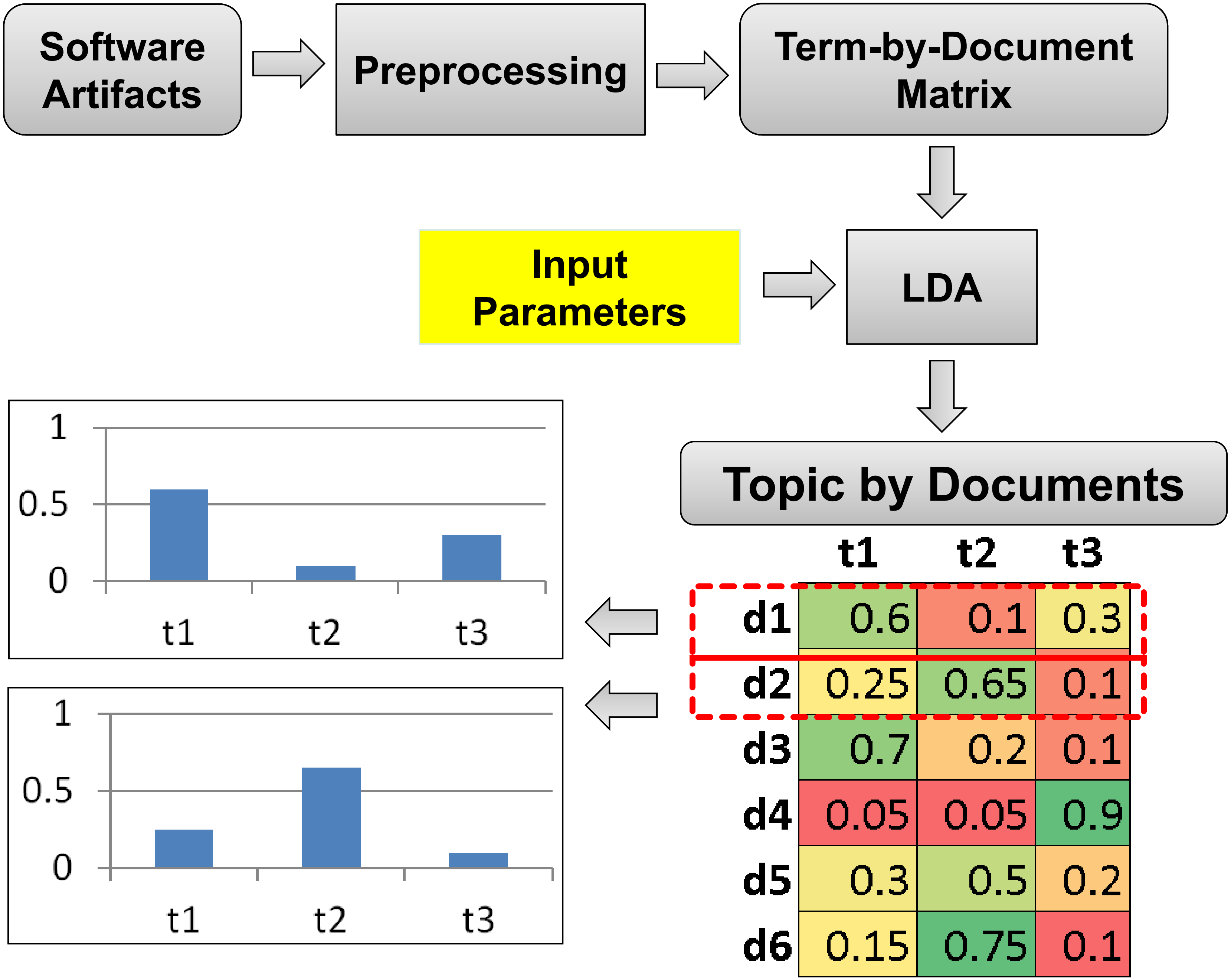
	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1



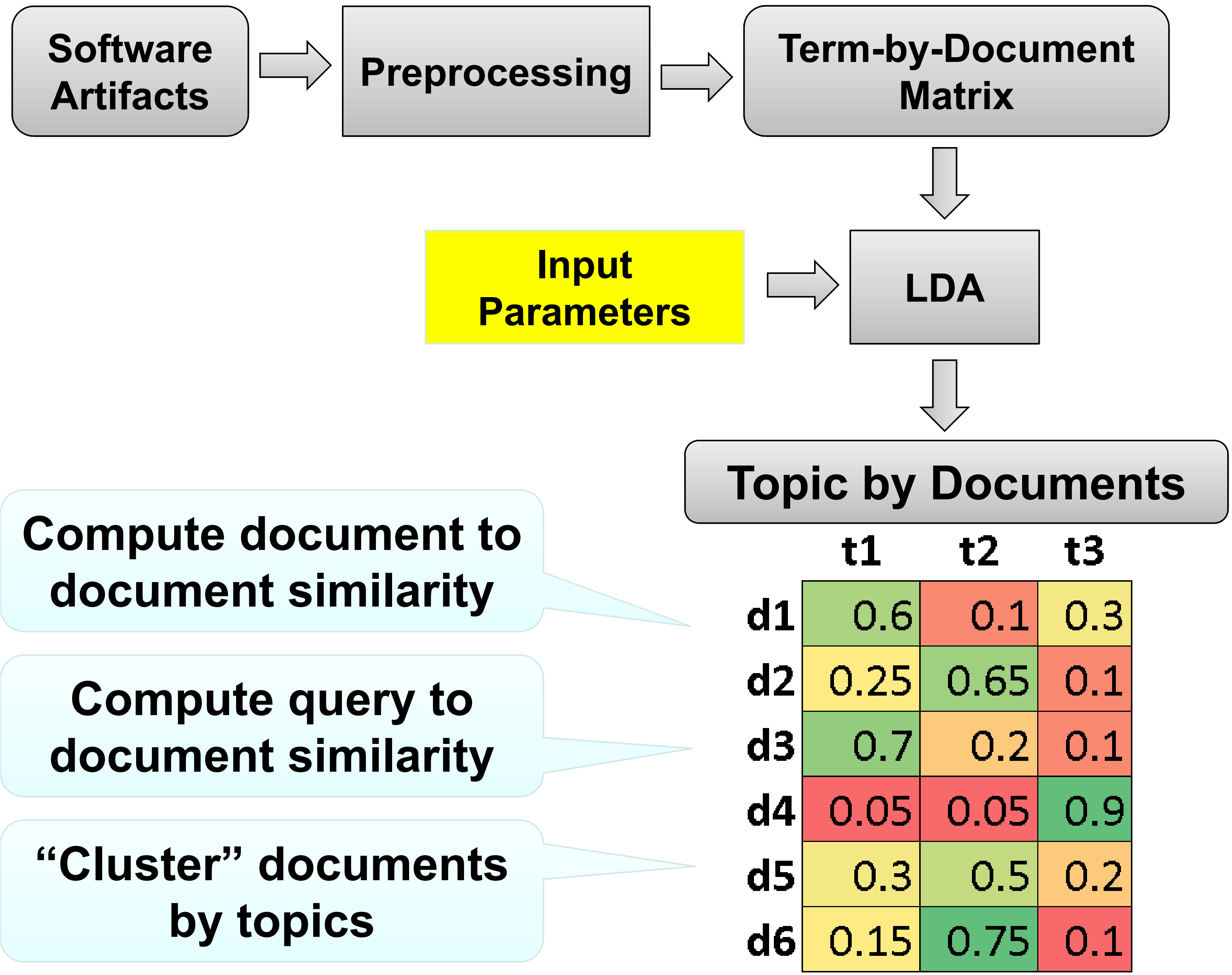
	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1



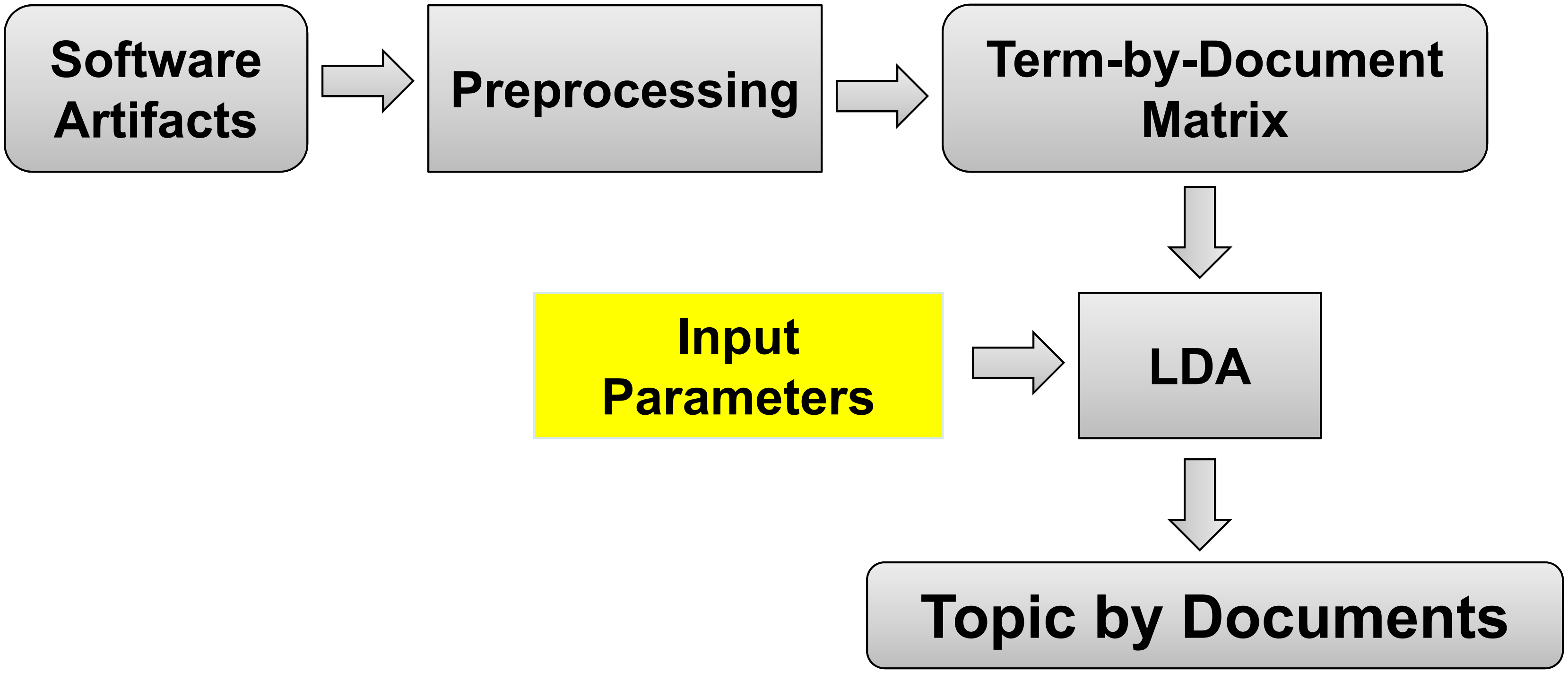
Software Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
Topic by Documents
t1 t2 t3
d1 0.6 0.1 0.3
d2 0.25 0.65 0.1
d3 0.7 0.2 0.1
d4 0.05 0.05 0.9
d5 0.3 0.5 0.2
d6 0.15 0.75 0.1
1
0.5
0
t1
t2
t3



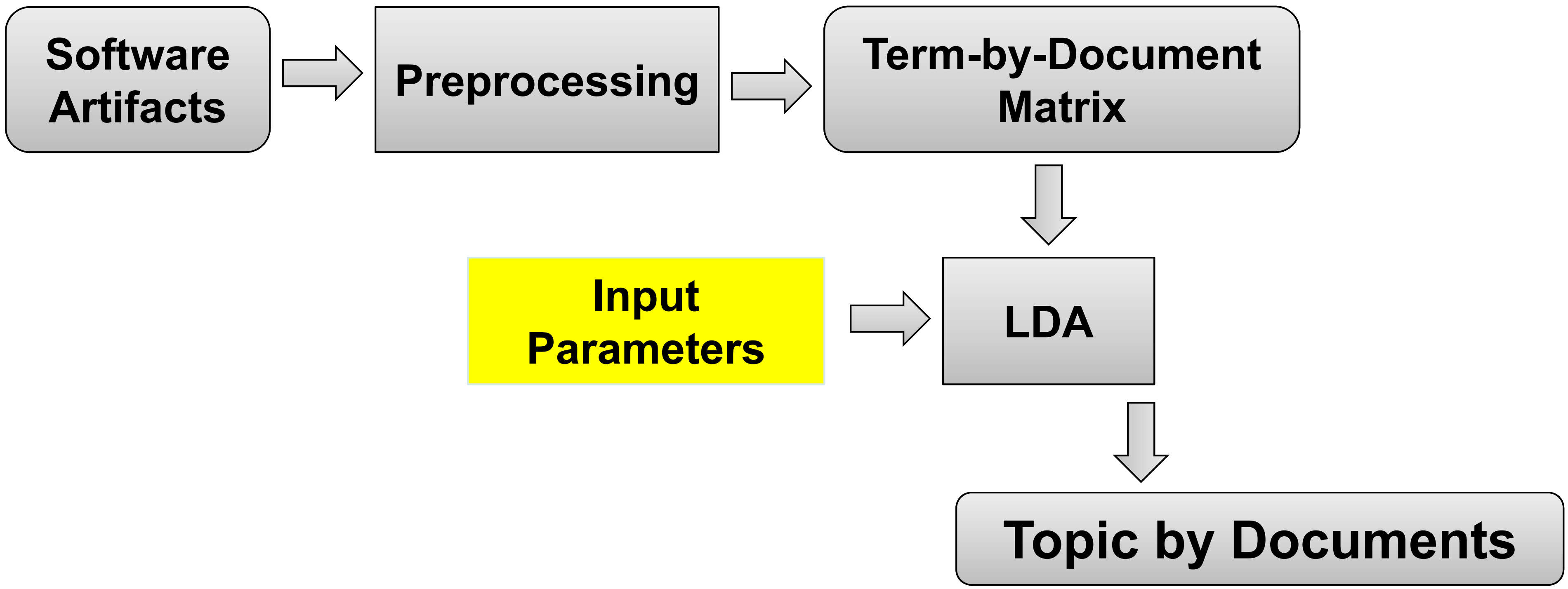
Software Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
Topic by Documents
t1 t2 t3
d1 0.6 0.1 0.3
d2 0.25 0.65 0.1
d3 0.7 0.2 0.1
d4 0.05 0.05 0.9
d5 0.3 0.5 0.2
d6 0.15 0.75 0.1
1
0.5
0
t1 t2 t3
1
0.5
0
t1 t2 t3



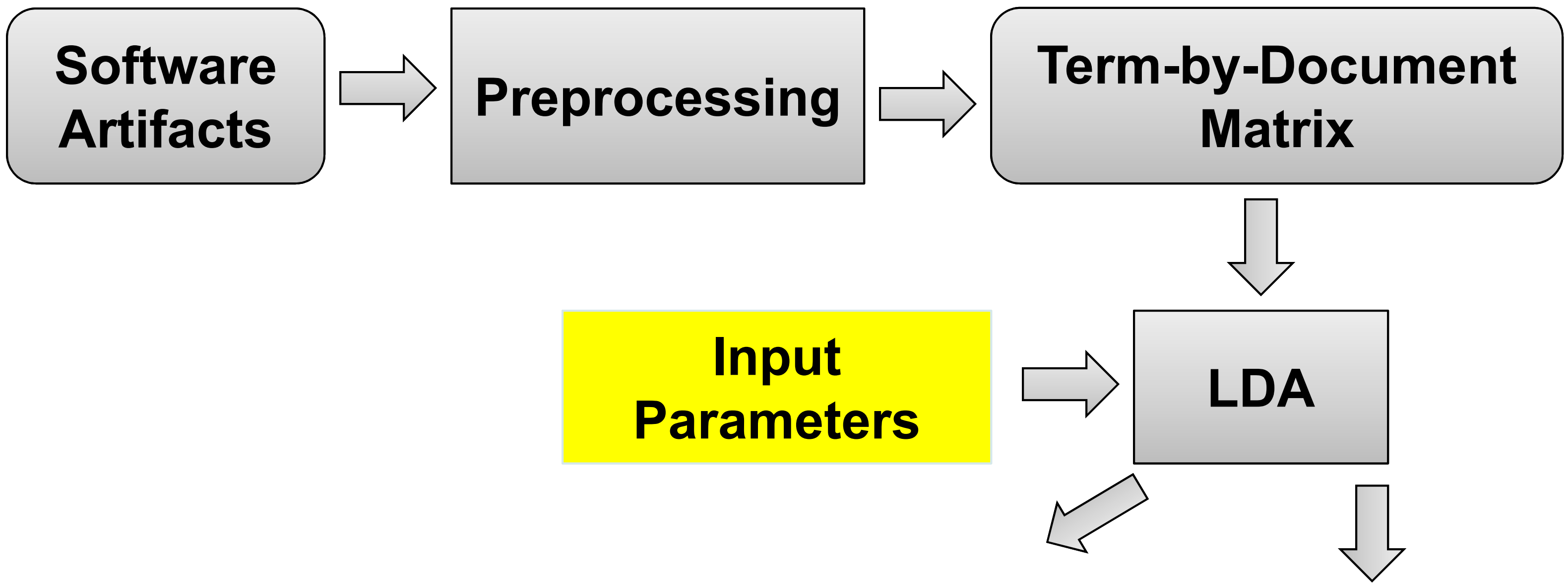
	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1



	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1



	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1

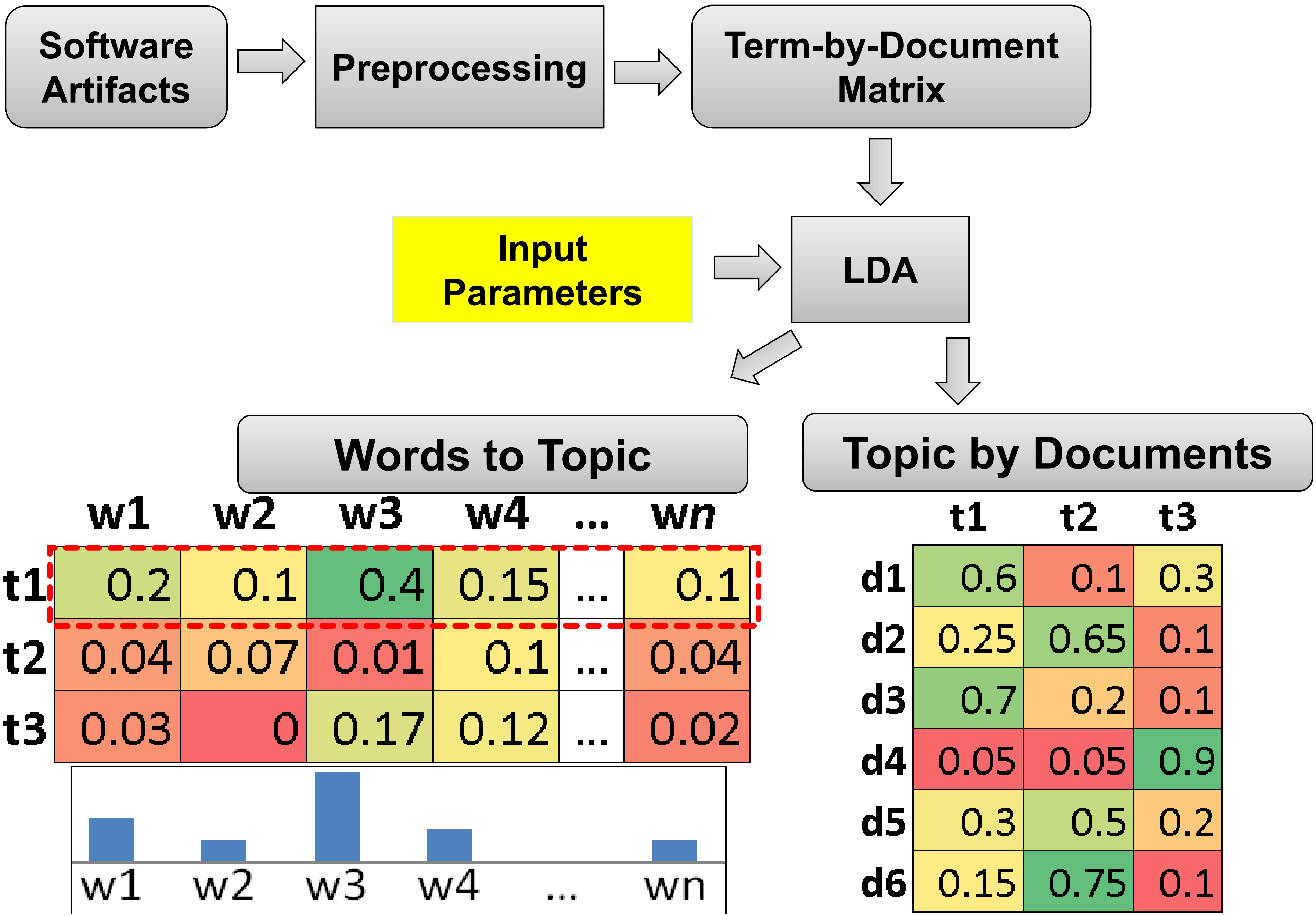


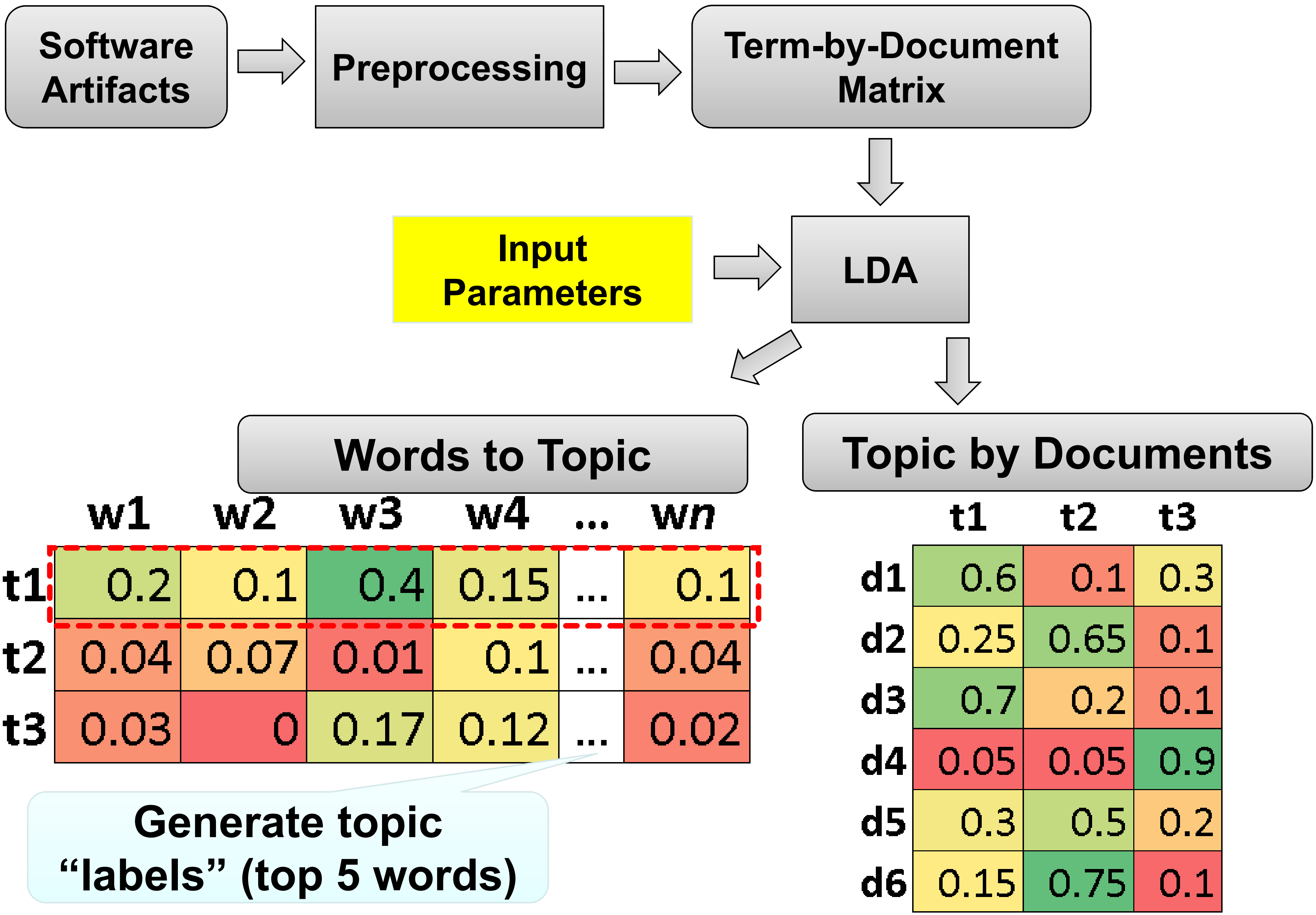
Words to Topic

	w1	w2	w3	w4	...	wn
t1	0.2	0.1	0.4	0.15	...	0.1
t2	0.04	0.07	0.01	0.1	...	0.04
t3	0.03	0	0.17	0.12	...	0.02

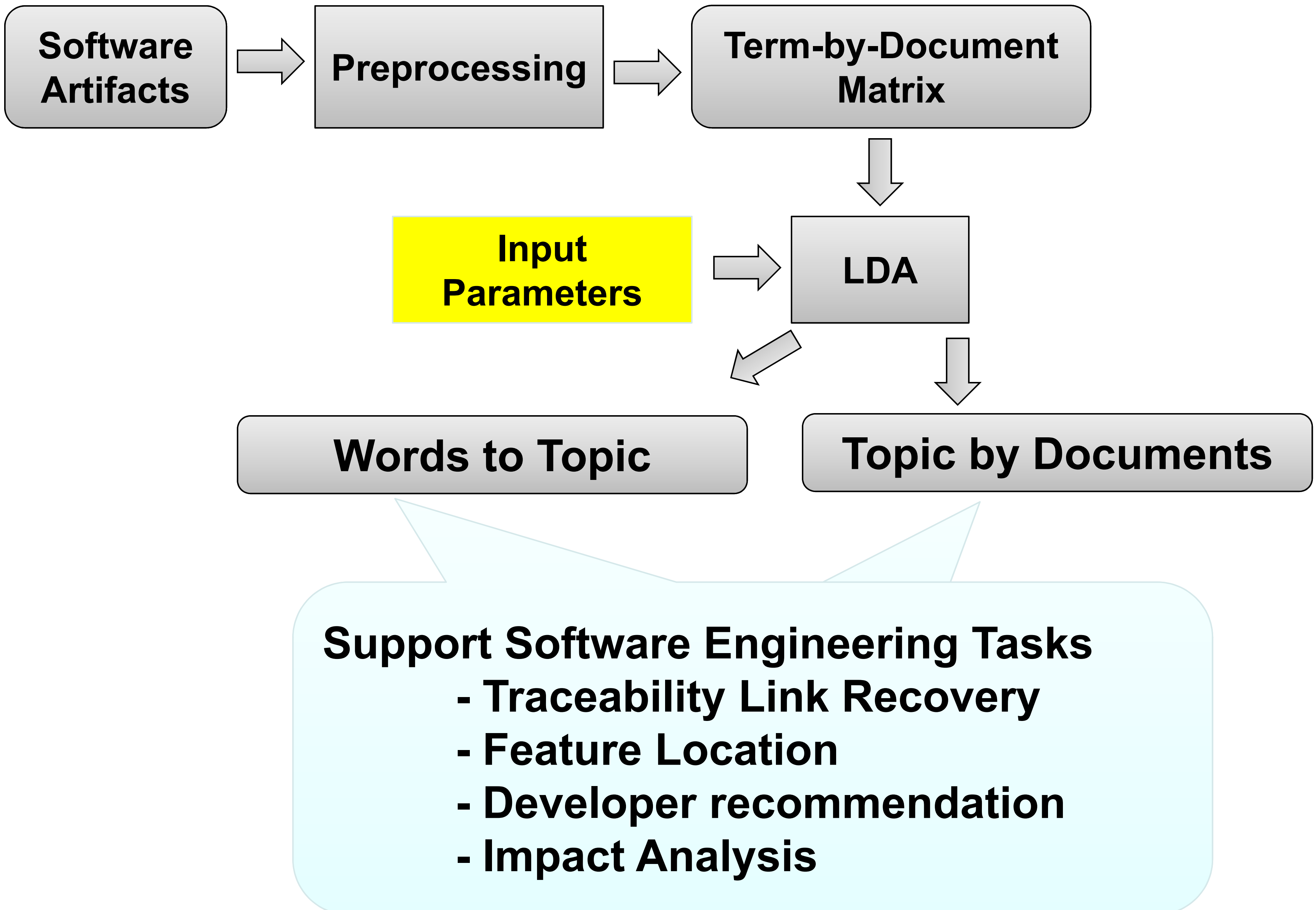
Topic by Documents

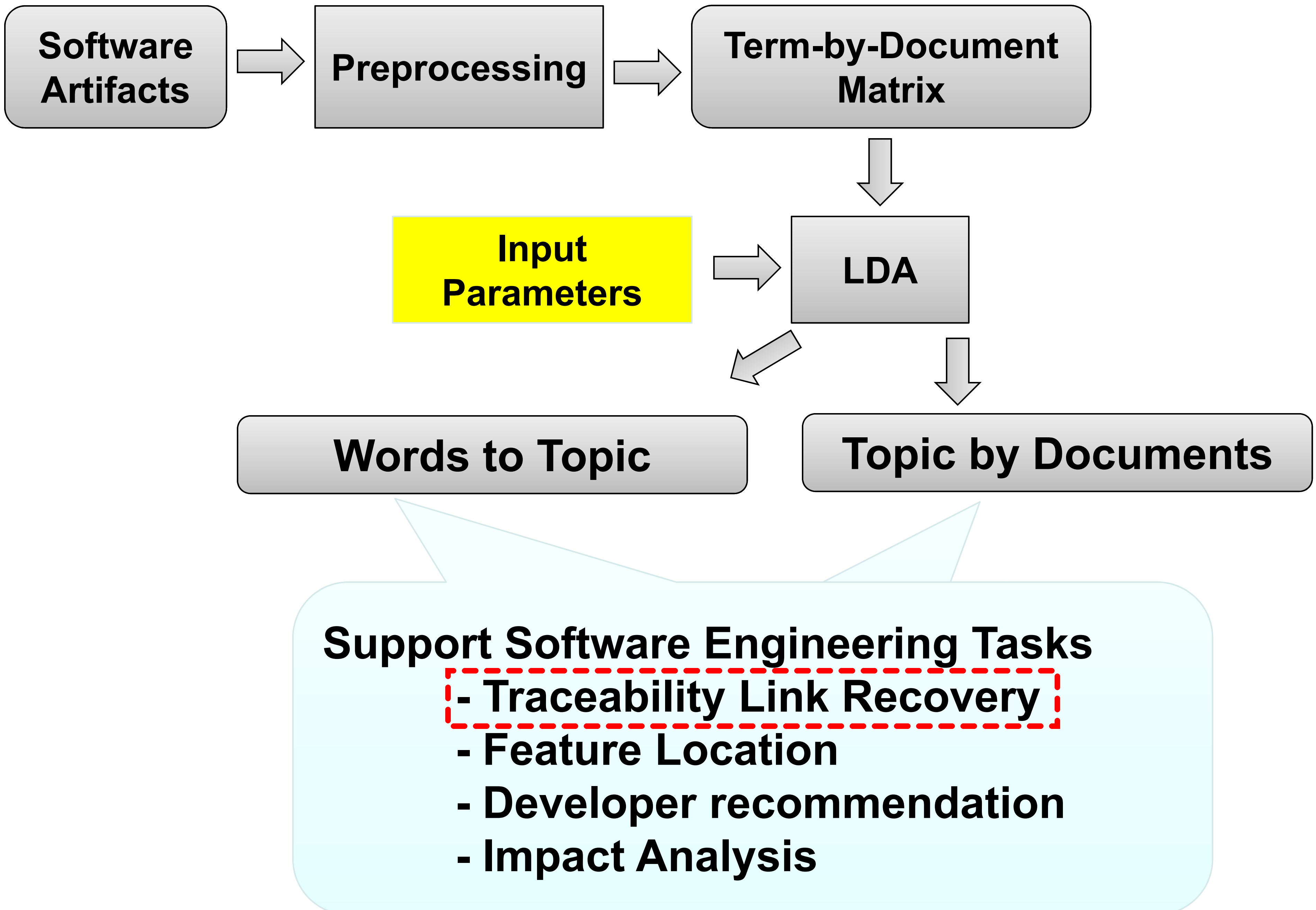
	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1

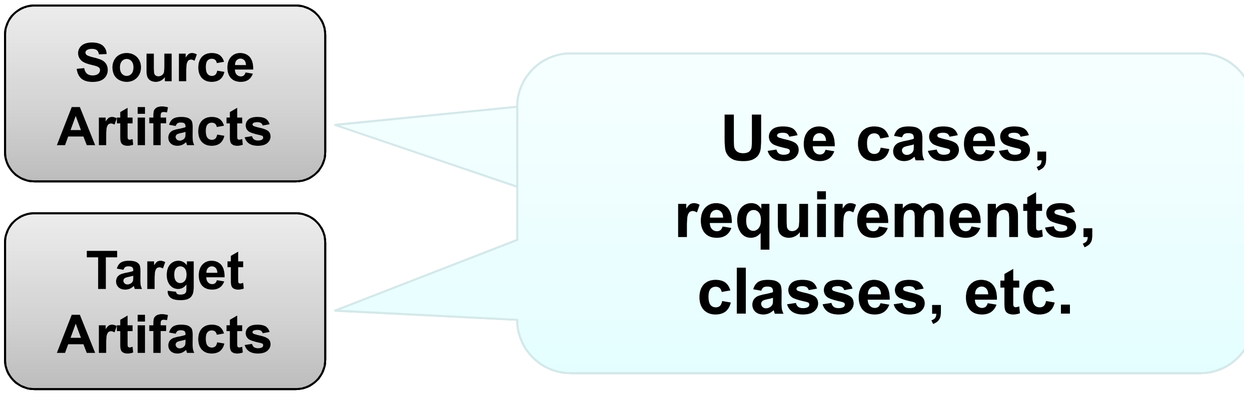

Software Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
Words to Topic
Topic by Documents
w1 w2 w3 w4 ... wn
t1 0.2 0.1 0.4 0.15 ... 0.1
t2 0.04 0.07 0.01 0.1 ... 0.04
t3 0.03 0 0.17 0.12 ... 0.02
w1 w2 w3 w4 ... wn
t1 t2 t3
d1 0.6 0.1 0.3
d2 0.25 0.65 0.1
d3 0.7 0.2 0.1
d4 0.05 0.05 0.9
d5 0.3 0.5 0.2
d6 0.15 0.75 0.1

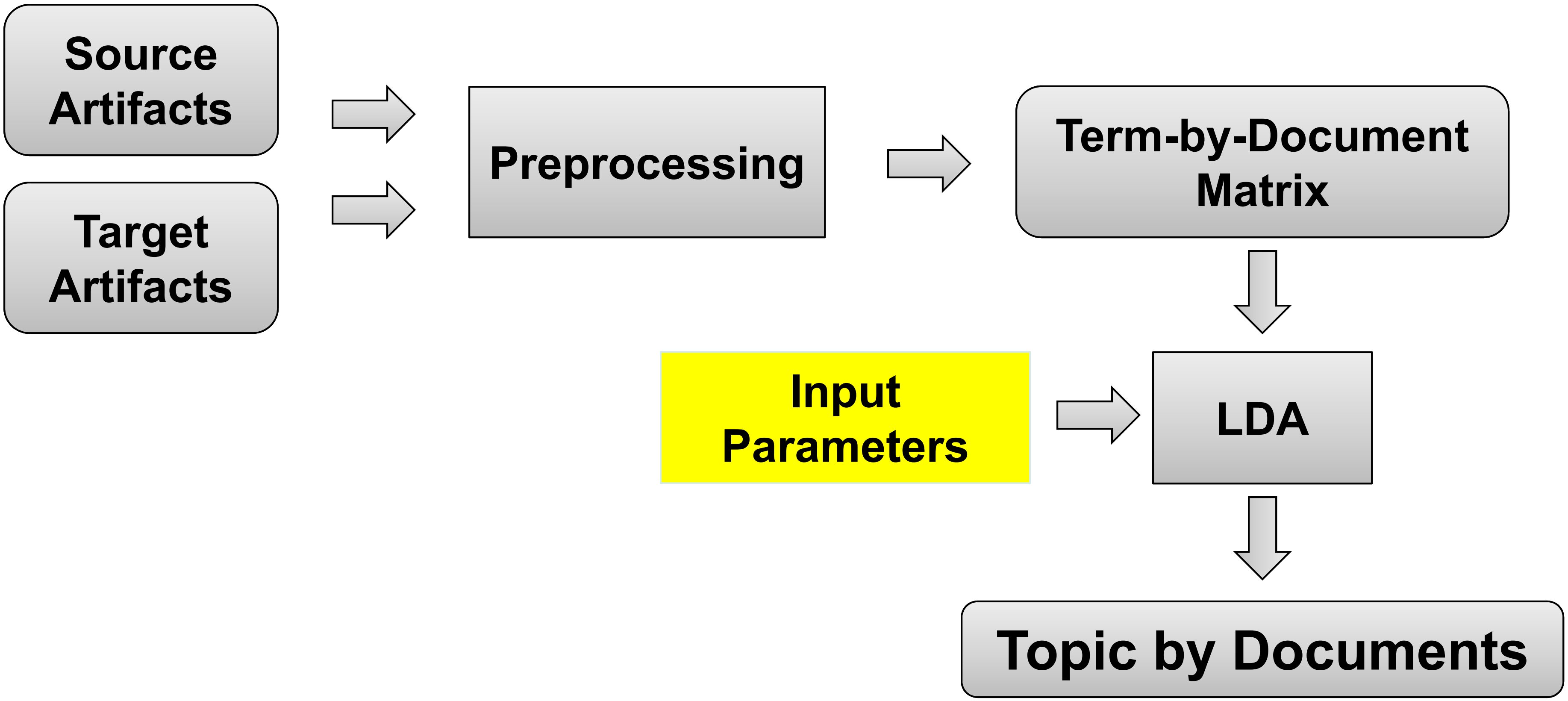


Software Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
Words to Topic
w1 w2 w3 w4 ... wn
t1 0.2 0.1 0.4 0.15 ... 0.1
t2 0.04 0.07 0.01 0.1 ... 0.04
t3 0.03 0 0.17 0.12 ... 0.02
Generate topic “labels” (top 5 words)
Topic by Documents
t1 t2 t3
d1 0.6 0.1 0.3
d2 0.25 0.65 0.1
d3 0.7 0.2 0.1
d4 0.05 0.05 0.9
d5 0.3 0.5 0.2
d6 0.15 0.75 0.1

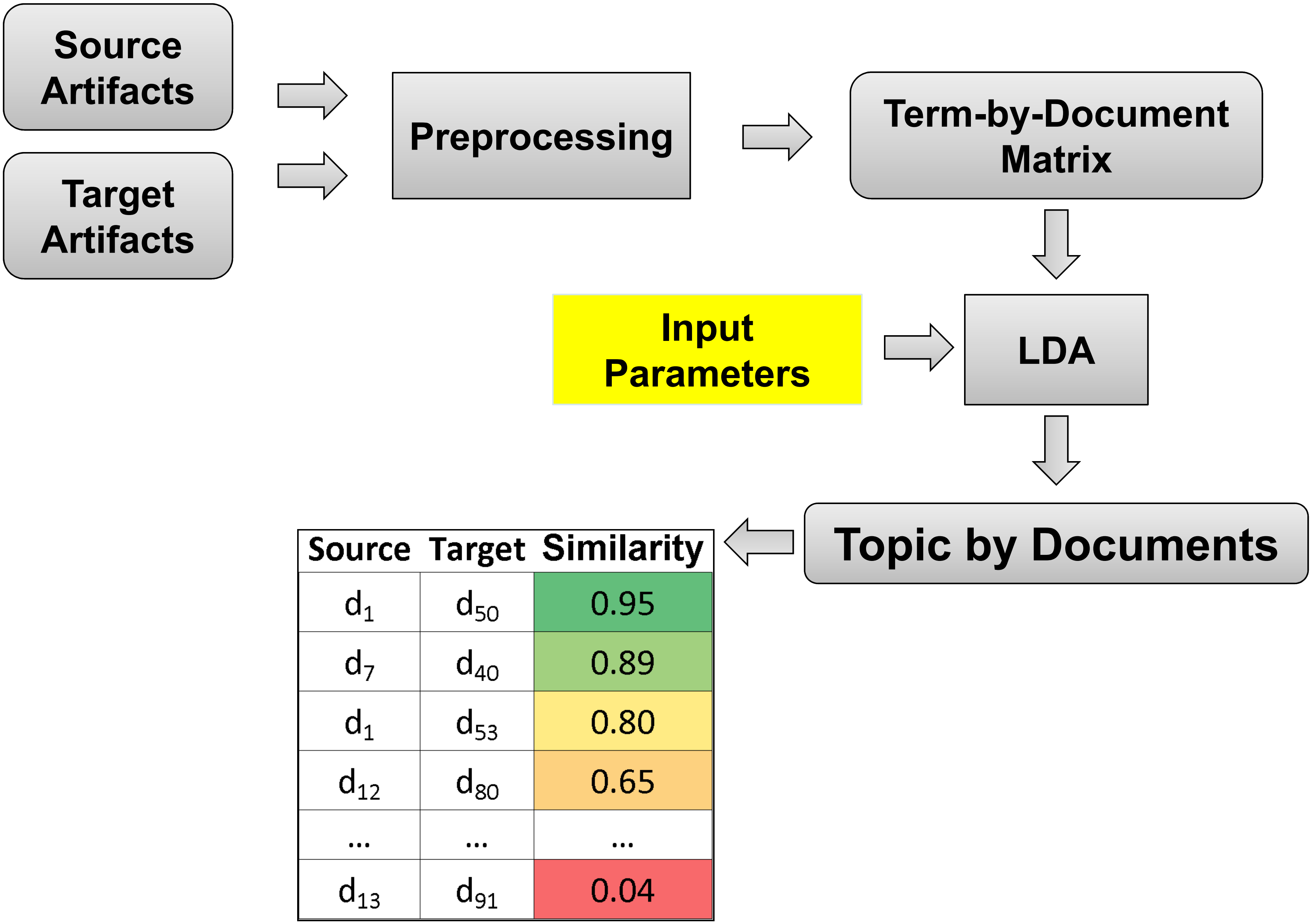

Software Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
Words to Topic
Topic by Documents
Support Software Engineering Tasks
- Traceability Link Recovery
- Feature Location
- Developer recommendation
- Impact Analysis

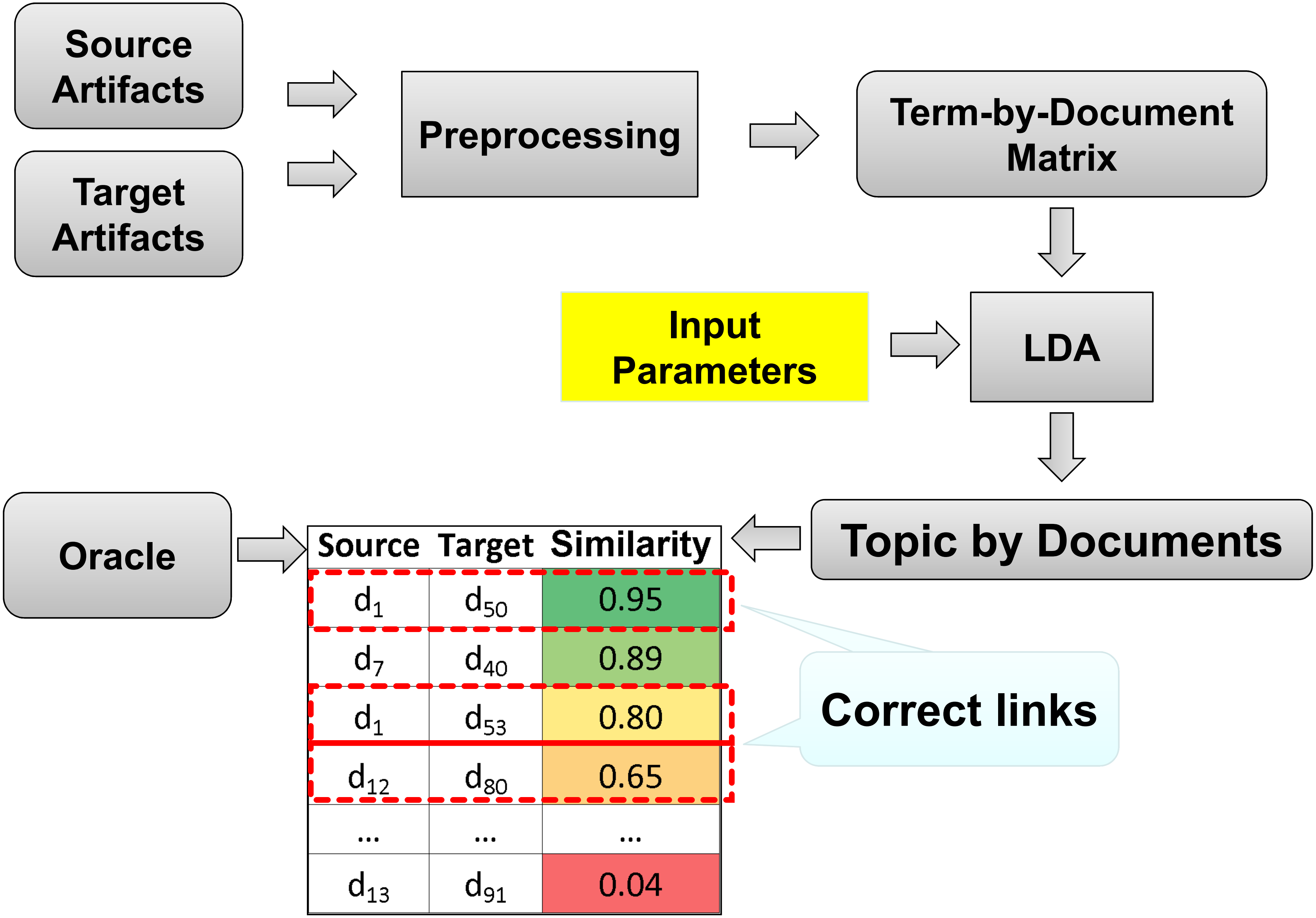

Software Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
Words to Topic
Topic by Documents
Support Software Engineering Tasks
- Traceability Link Recovery
- Feature Location
- Developer recommendation
- Impact Analysis

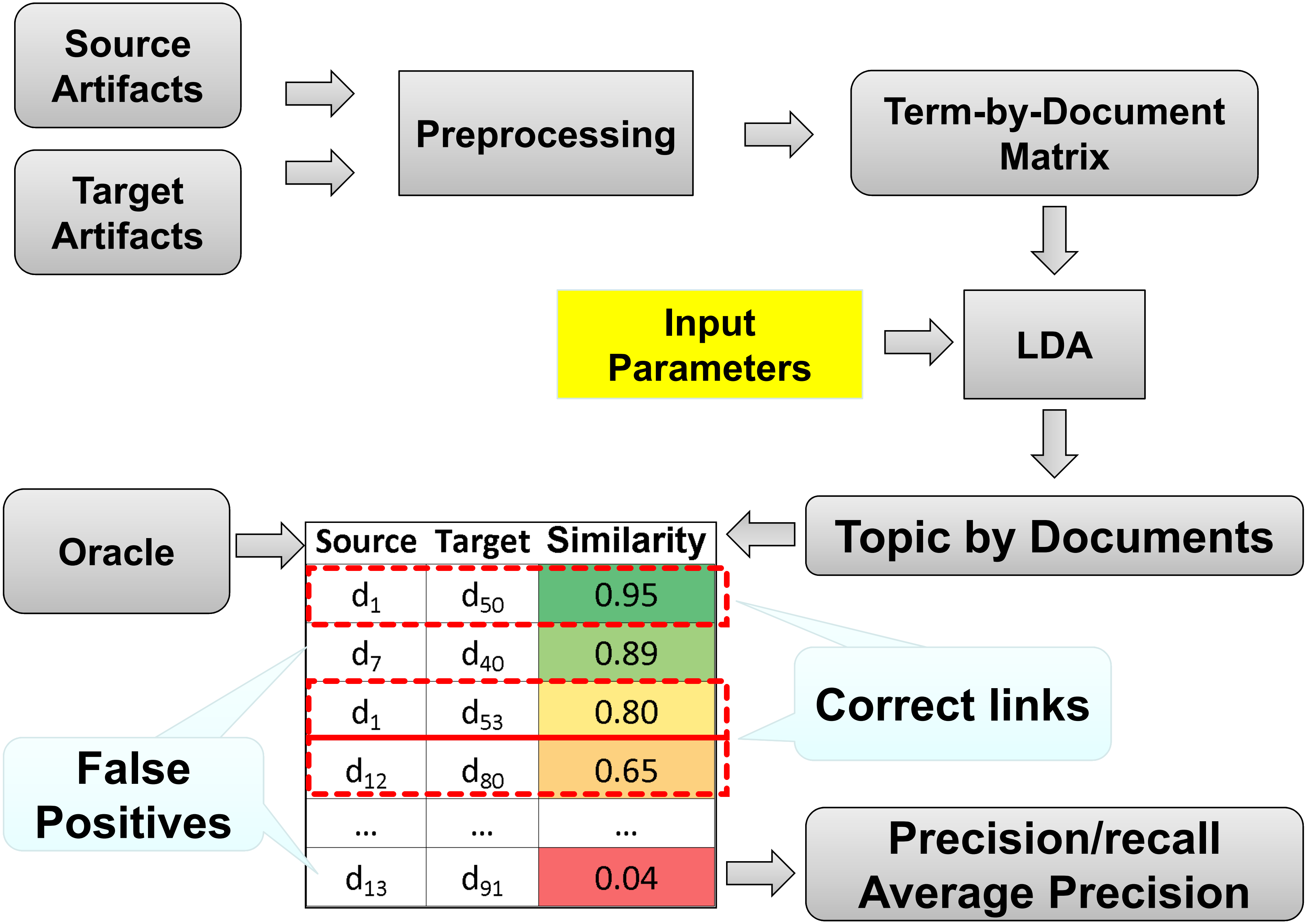

Source Artifacts
Target Artifacts
Use cases, requirements, classes, etc.



Source Artifacts
Target Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
Topic by Documents

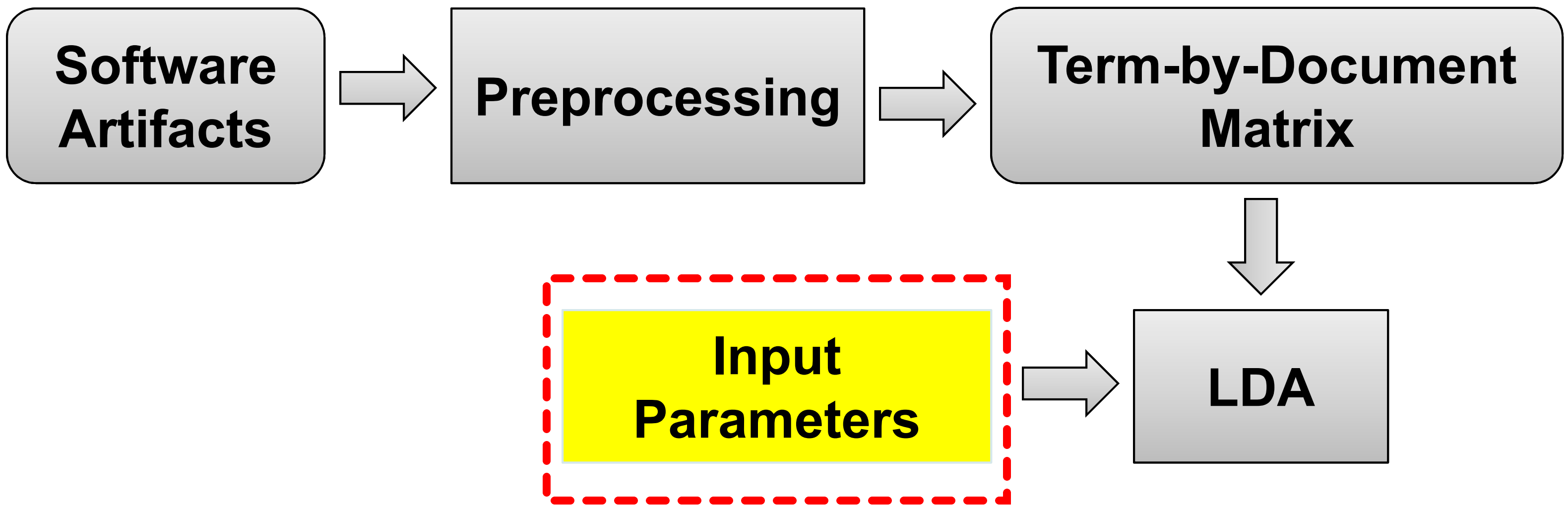

Source Artifacts
Target Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
Topic by Documents
Source Target Similarity
d_1 d_{50} 0.95
d_7 d_{40} 0.89
d_1 d_{53} 0.80
d_{12} d_{80} 0.65
... ...
d_{13} d_{91} 0.04

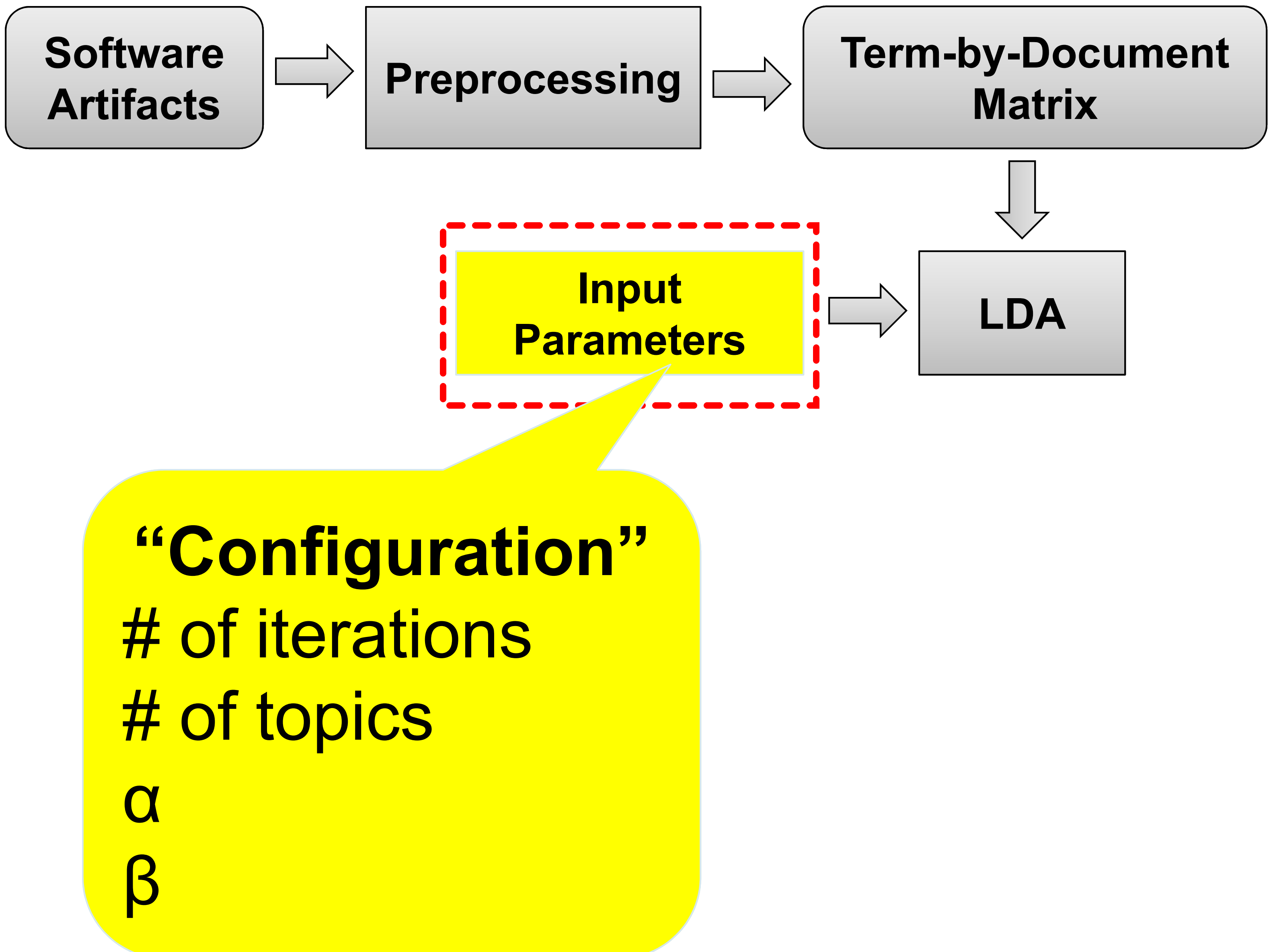

Source Artifacts
Target Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
Oracle
Topic by Documents
Source Target Similarity
d1 d50 0.95
d7 d40 0.89
d1 d53 0.80
d12 d80 0.65
... ...
d13 d91 0.04
Correct links

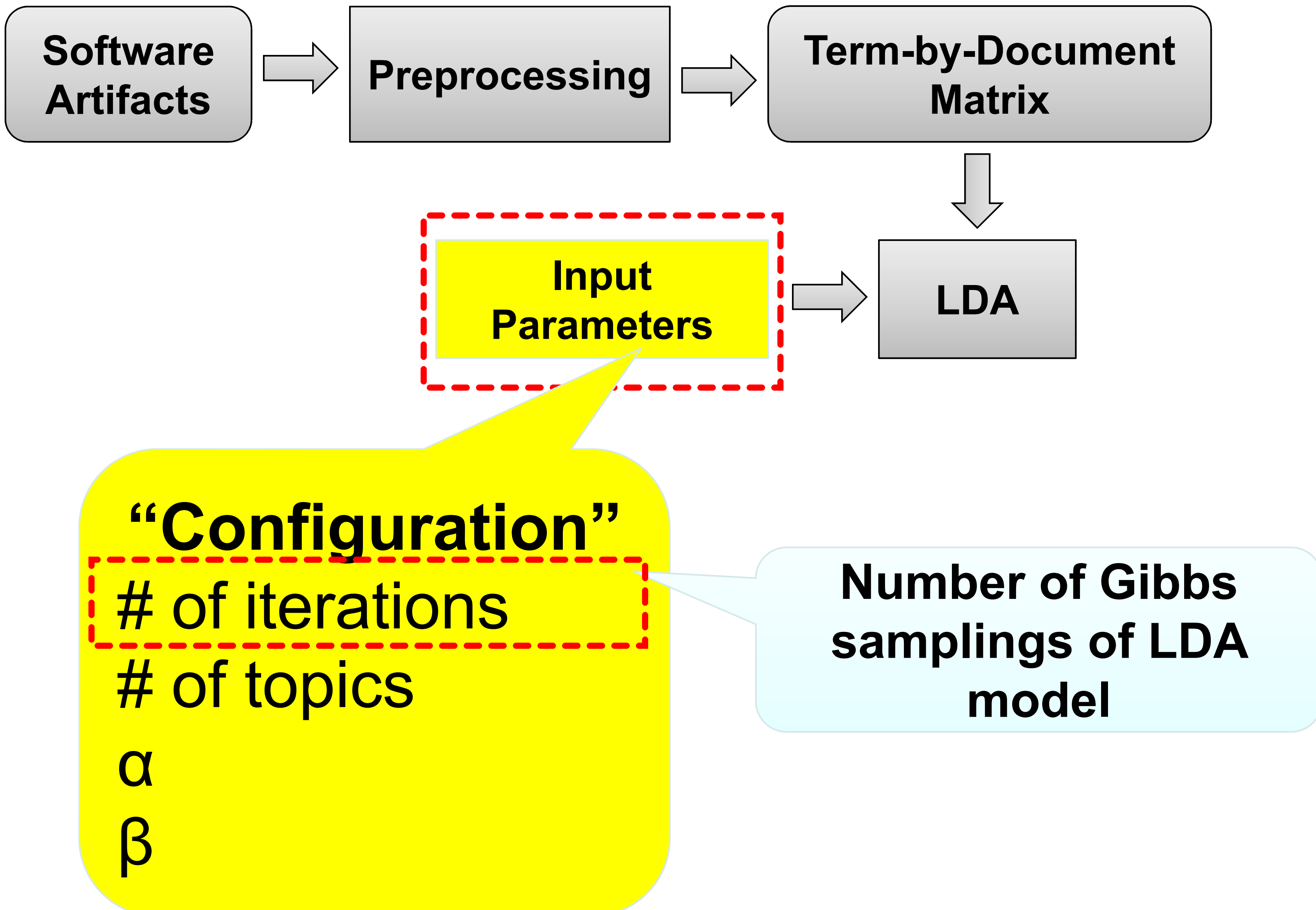


Source Artifacts
Target Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
Topic by Documents
Oracle
Source
Target
Similarity
d1
d50
0.95
d7
d40
0.89
d1
d53
0.80
d12
d80
0.65
...
...
...
d13
d91
0.04
Correct links
False Positives
Precision/recall
Average Precision

Let’s examine the LDA input parameters in more details

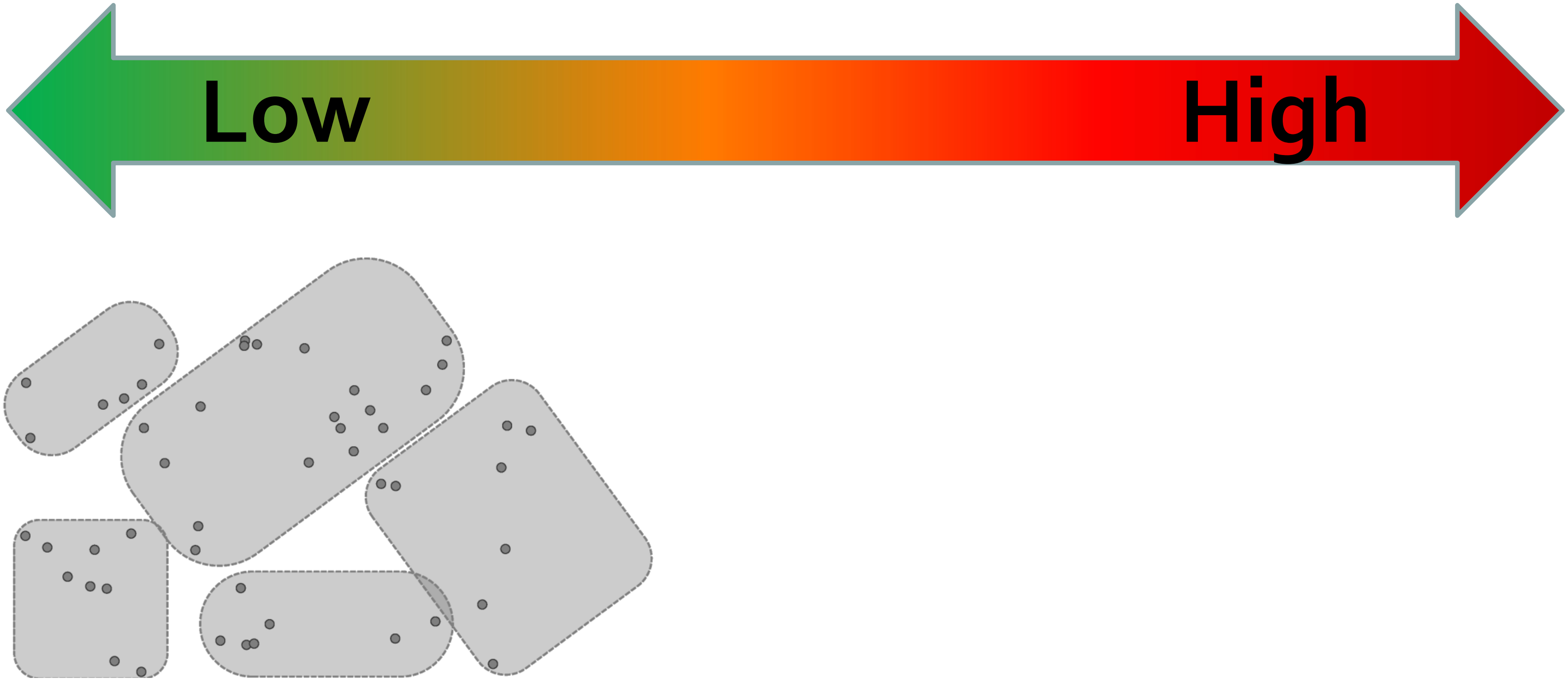

Software Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA


Software Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
"Configuration"
# of iterations
# of topics
α
β


Software Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
"Configuration"
# of iterations
# of topics
α
β
Number of Gibbs samplings of LDA model

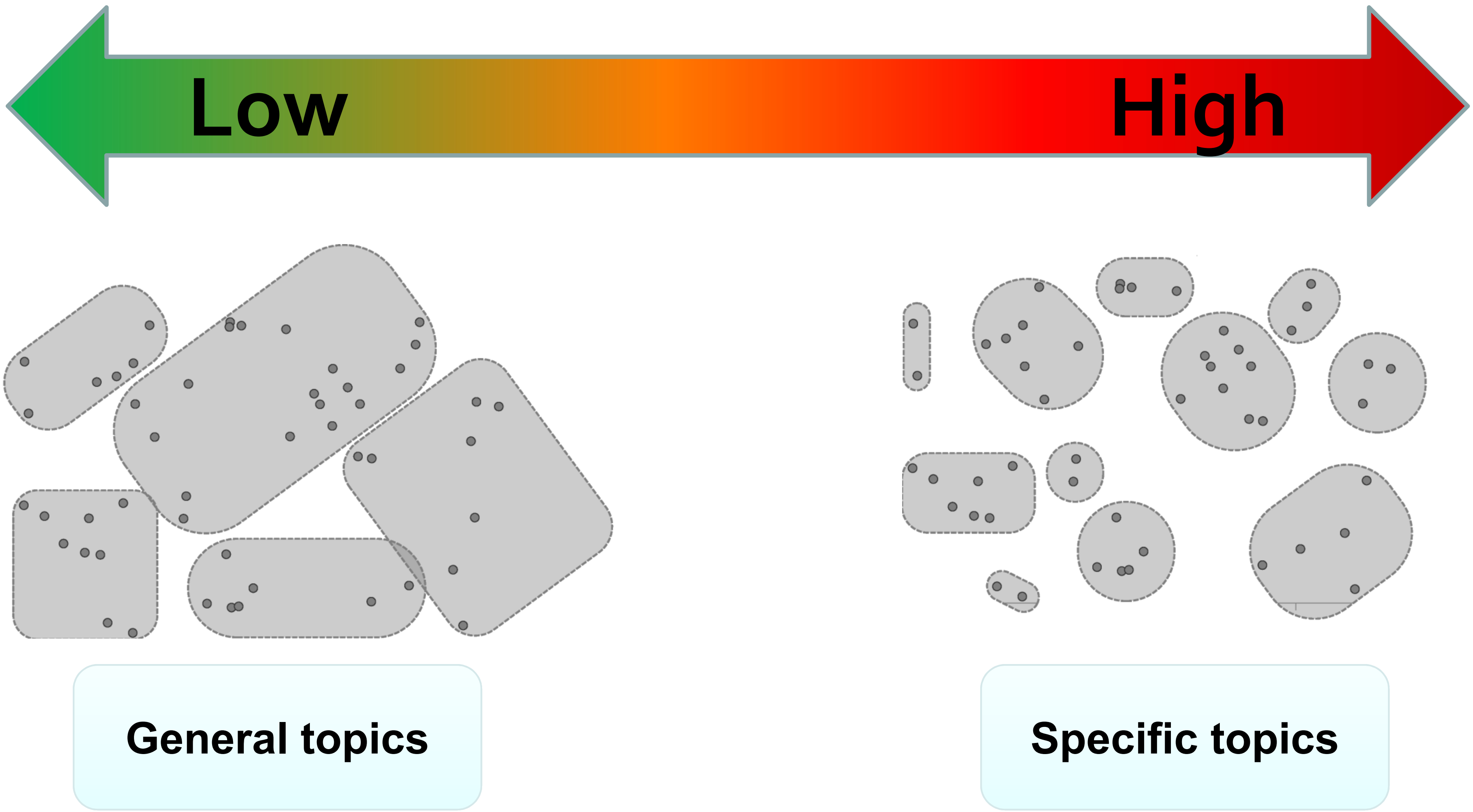
Number of topics…

Number of topics…



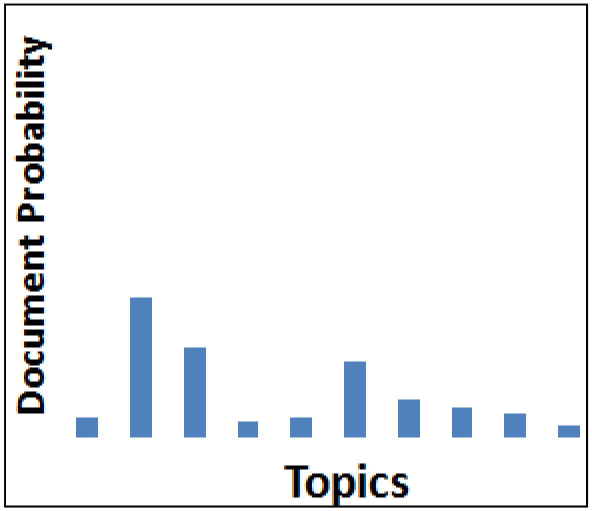
General topics

Number of topics…

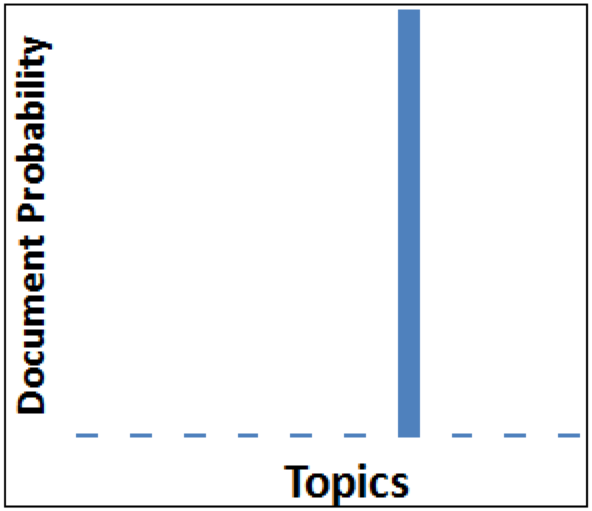


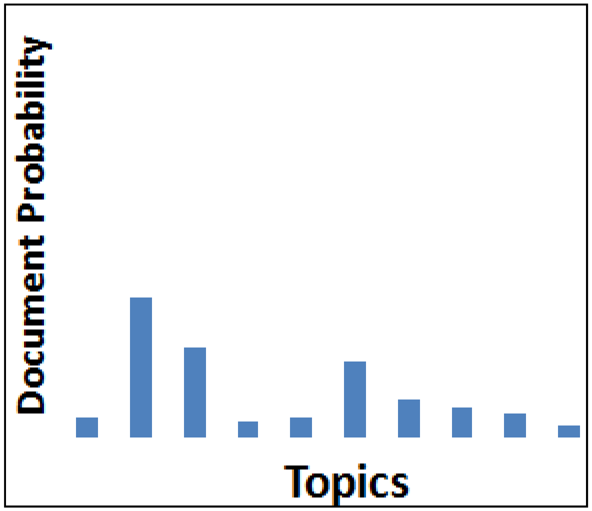
α, influences the “smoothness” of documents to topics distribution

α, influences the "smoothness" of documents to topics distribution

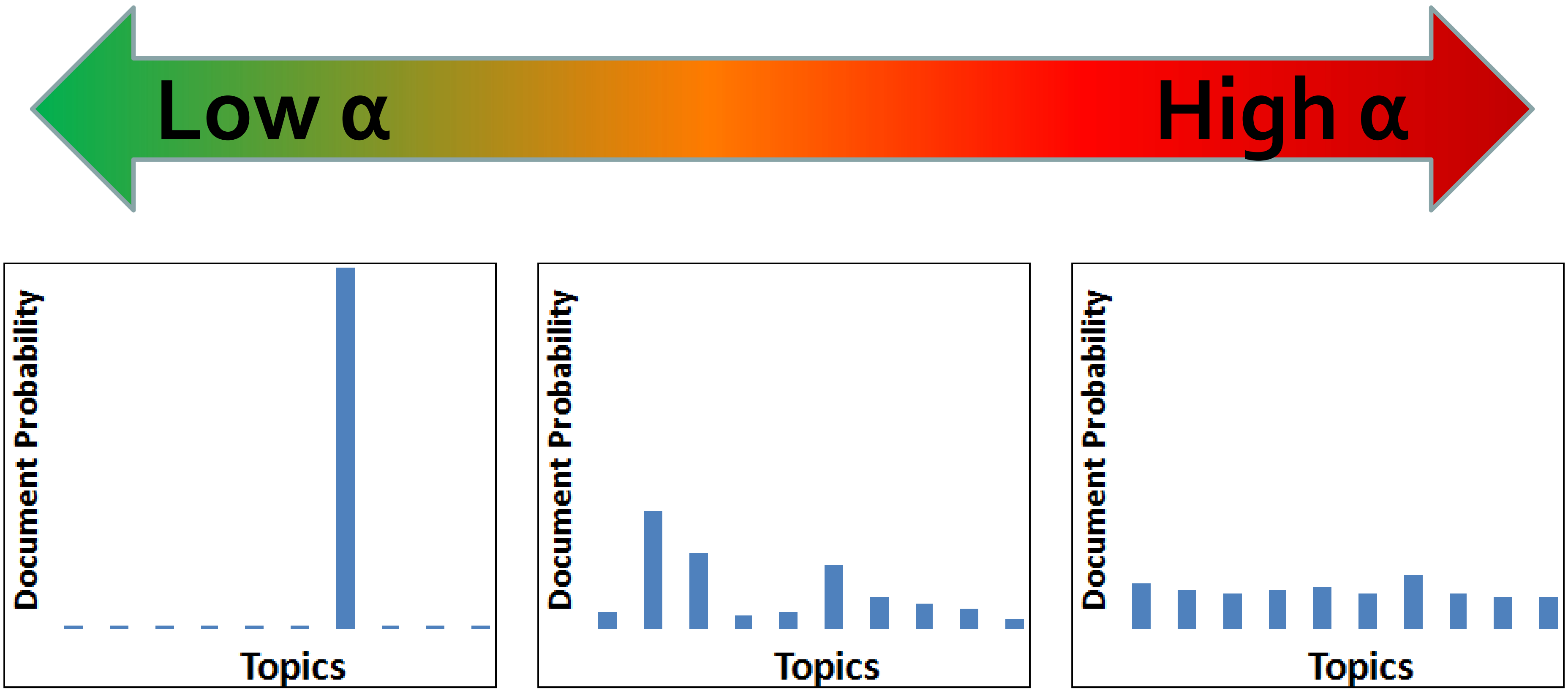


α, influences the "smoothness" of documents to topics distribution



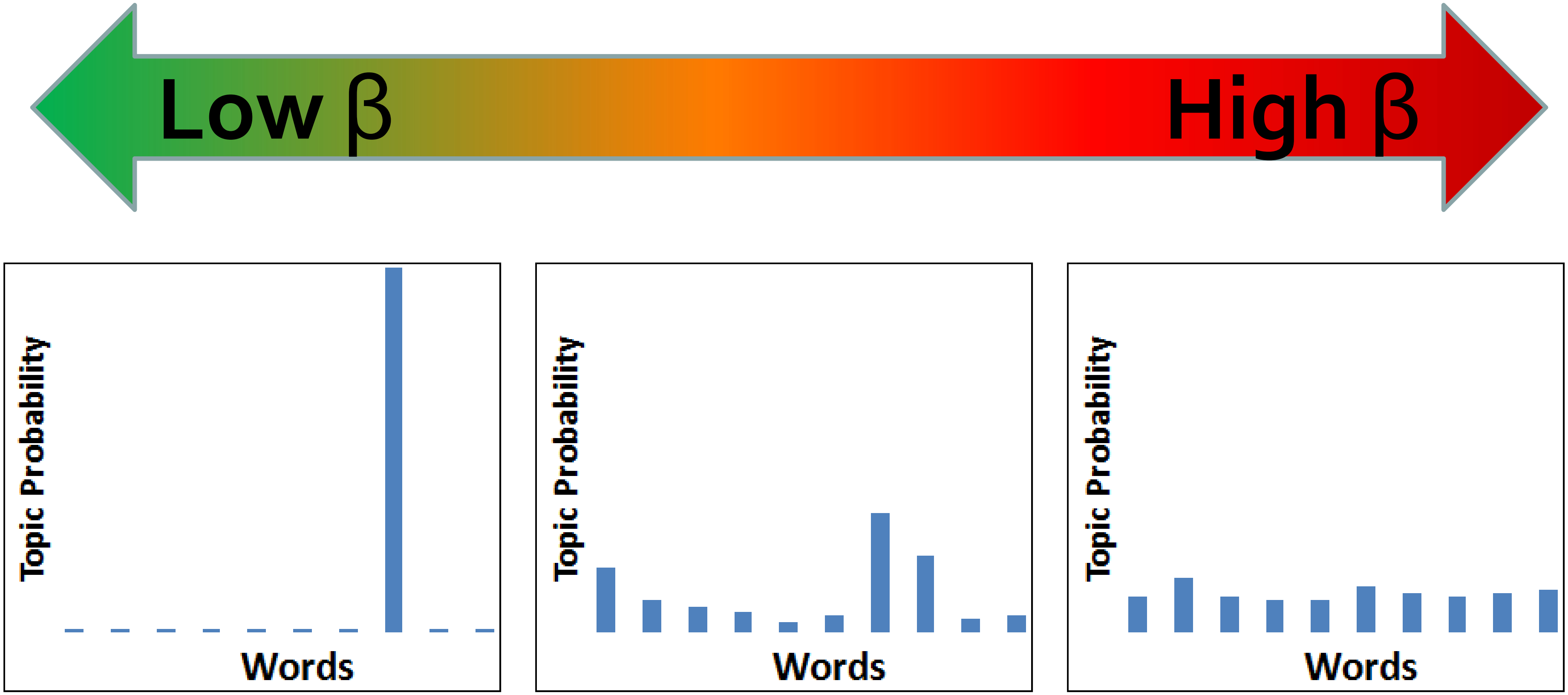


α, influences the “smoothness” of documents to topics distribution



β, influences the "smoothness" of topics to words distribution

β, influences the "smoothness" of topics to words distribution

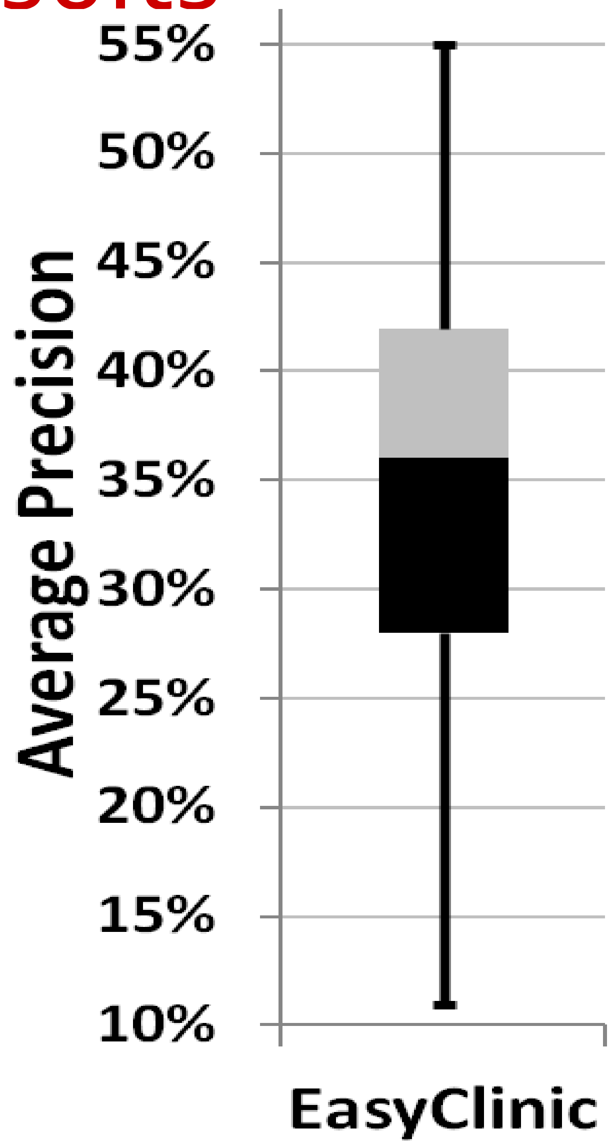


LDA parameters significantly influence the results

- Traceability Link Recovery
- 1,000 different configurations of LDA parameters
  - Evaluate the Average Precision on EasyClinic

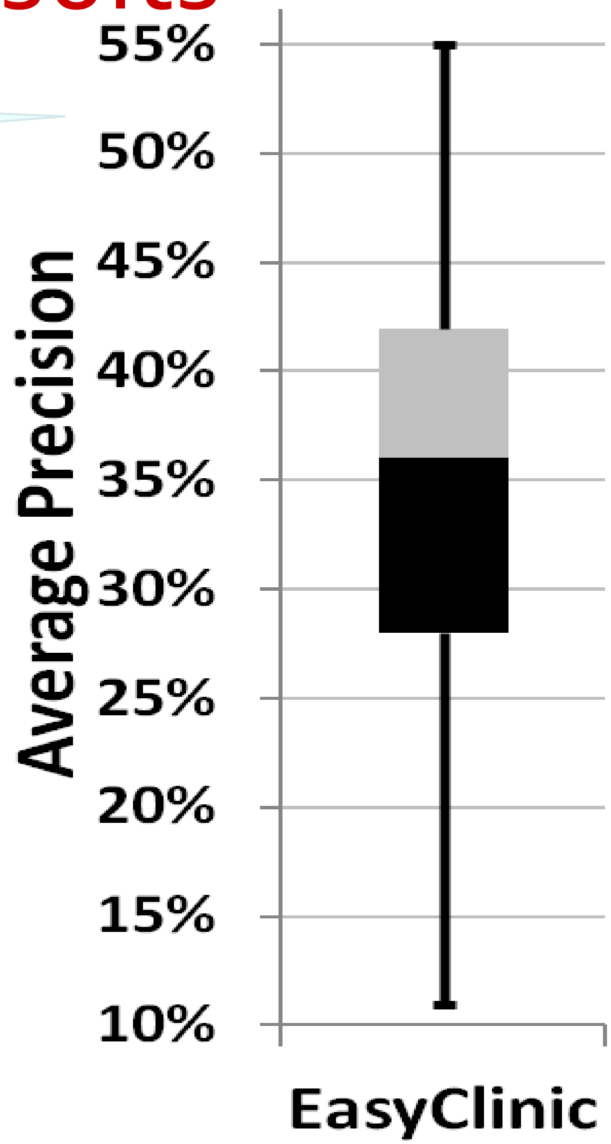
LDA parameters significantly influence the results

- Traceability Link Recovery
- 1,000 different configurations of LDA parameters
  - Evaluate the Average Precision on EasyClinic



LDA parameters significantly influence the results

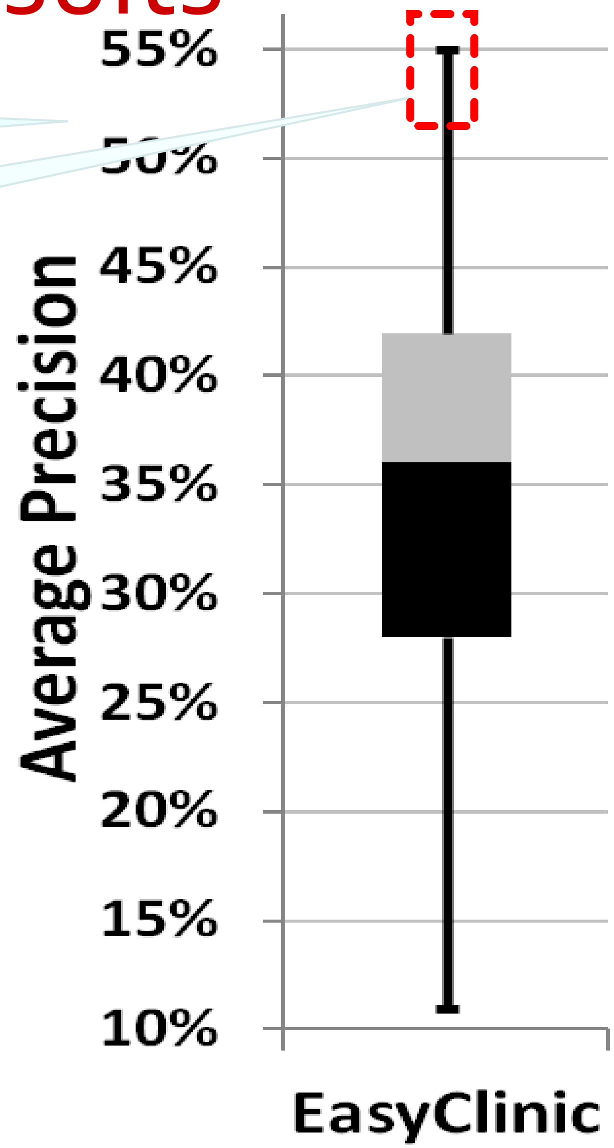
High variability in results



LDA parameters significantly influence the results

High variability in results

Few configurations produce good results



What kind of LDA configurations were used for software?

Software Engineering Tasks?

What kind of LDA configurations were used for software?

“ad-hoc” configurations

Parameters “imported” from natural language community

Assumption:

source code

has the same characteristics as

natural language

What kind of LDA configurations were used for software?

"ad-hoc" configurations

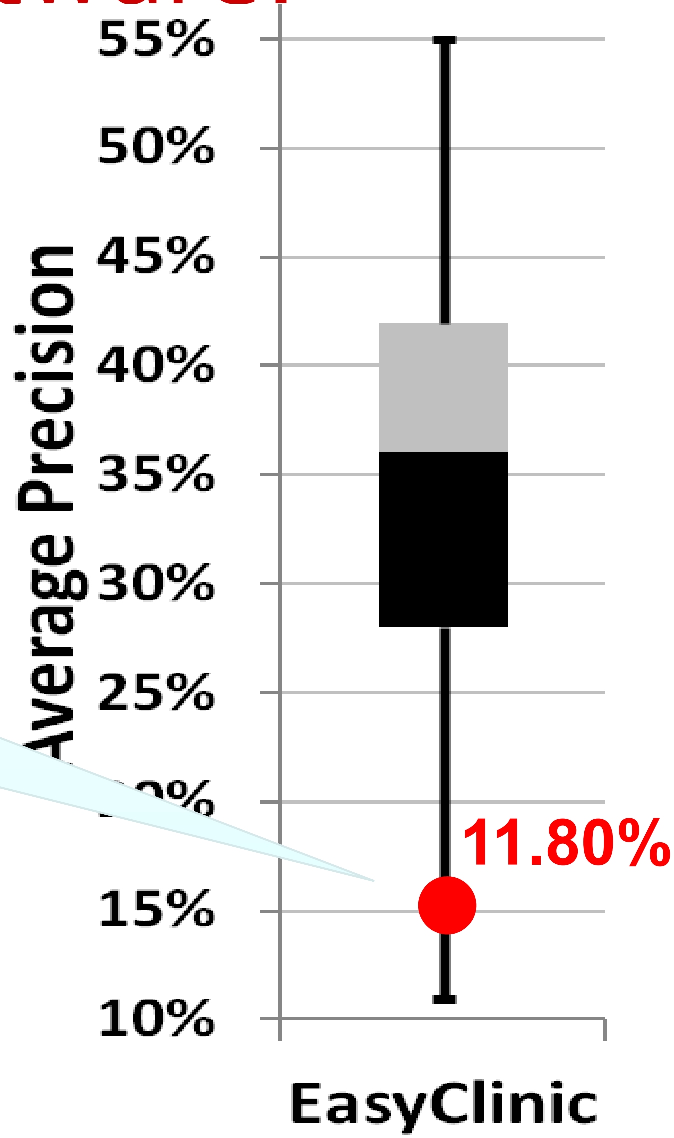
Parameters "imported" from natural language community

Assumption:

source code

has the same characteristics as

natural language



What kind of LDA configurations were used for software?

[Hindle et al. @ ICSE'12]:

source code

is more ***regular*** and ***predictable*** than

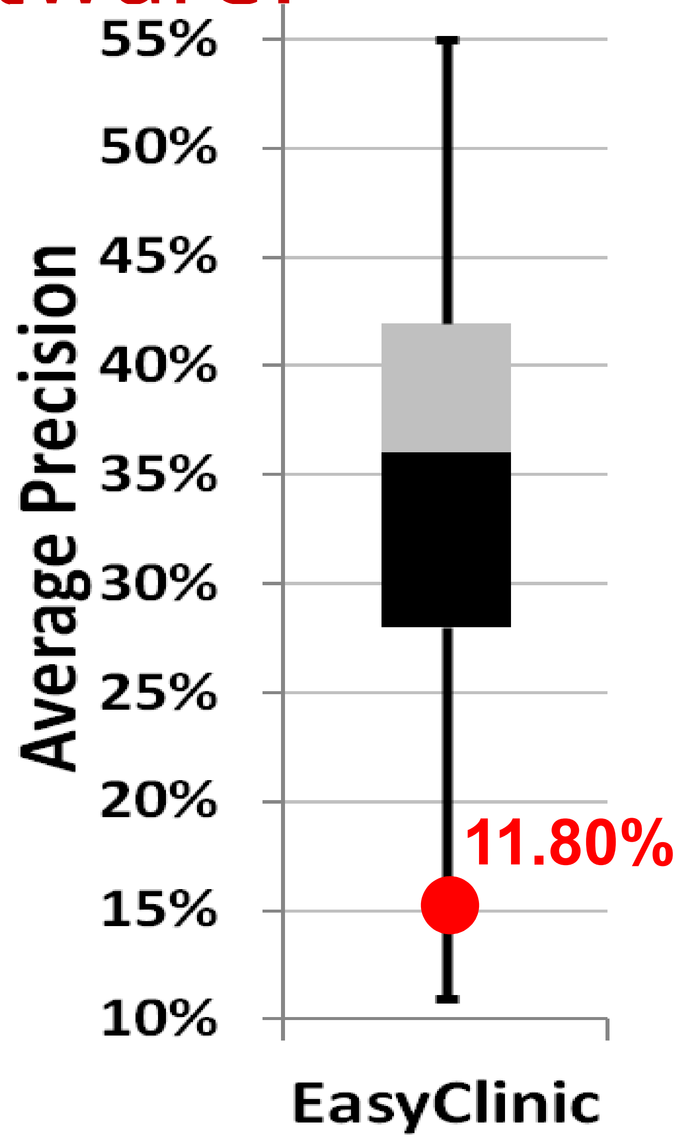
natural language

Assumption:

source code

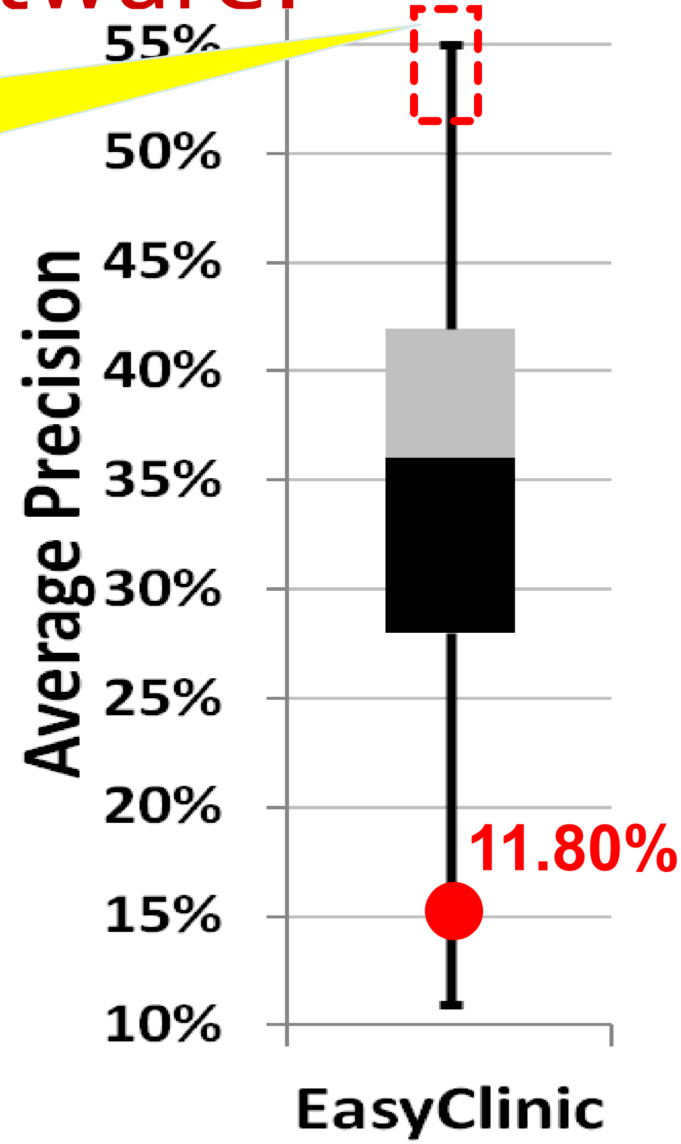
has the same characteristics as

natural language



What kind of LDA configurations were used for software?

We need new techniques to find these configurations



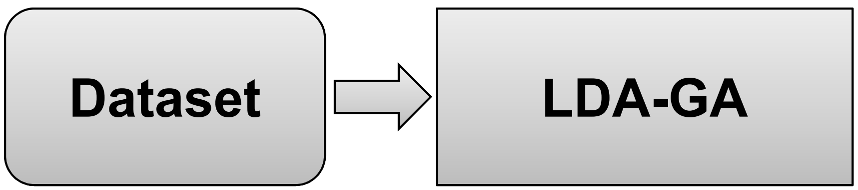
[Hindle et al. @ ICSE'12]:
source code
is more ***regular*** and ***predictable*** than
natural language

~~**Assumption:**~~
~~***source code***~~
~~has the same characteristics as~~
~~***natural language***~~

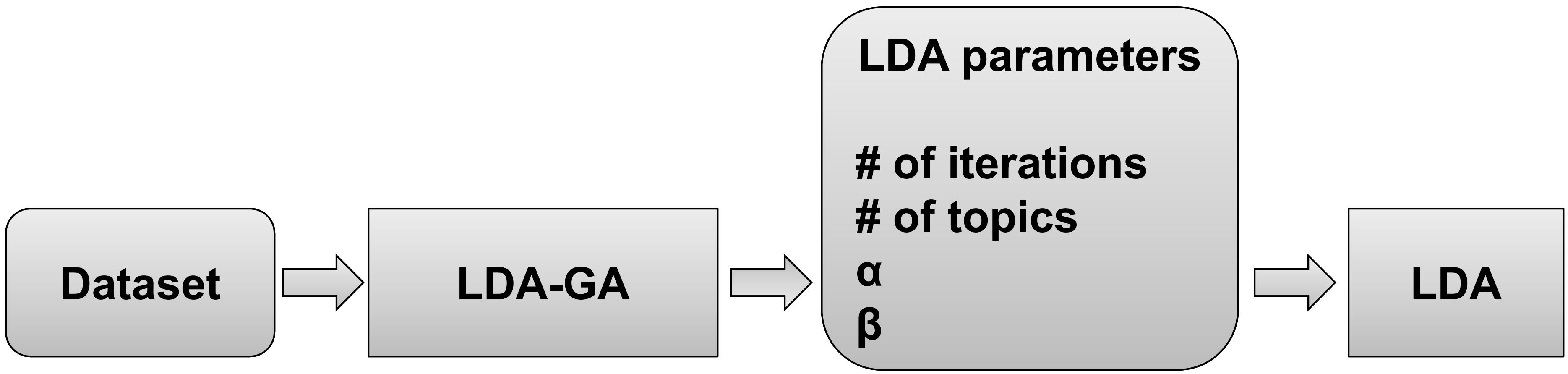
Our contribution…**LDA-GA**

LDA-GA: automatically calibrate the input parameters of LDA using a genetic algorithm

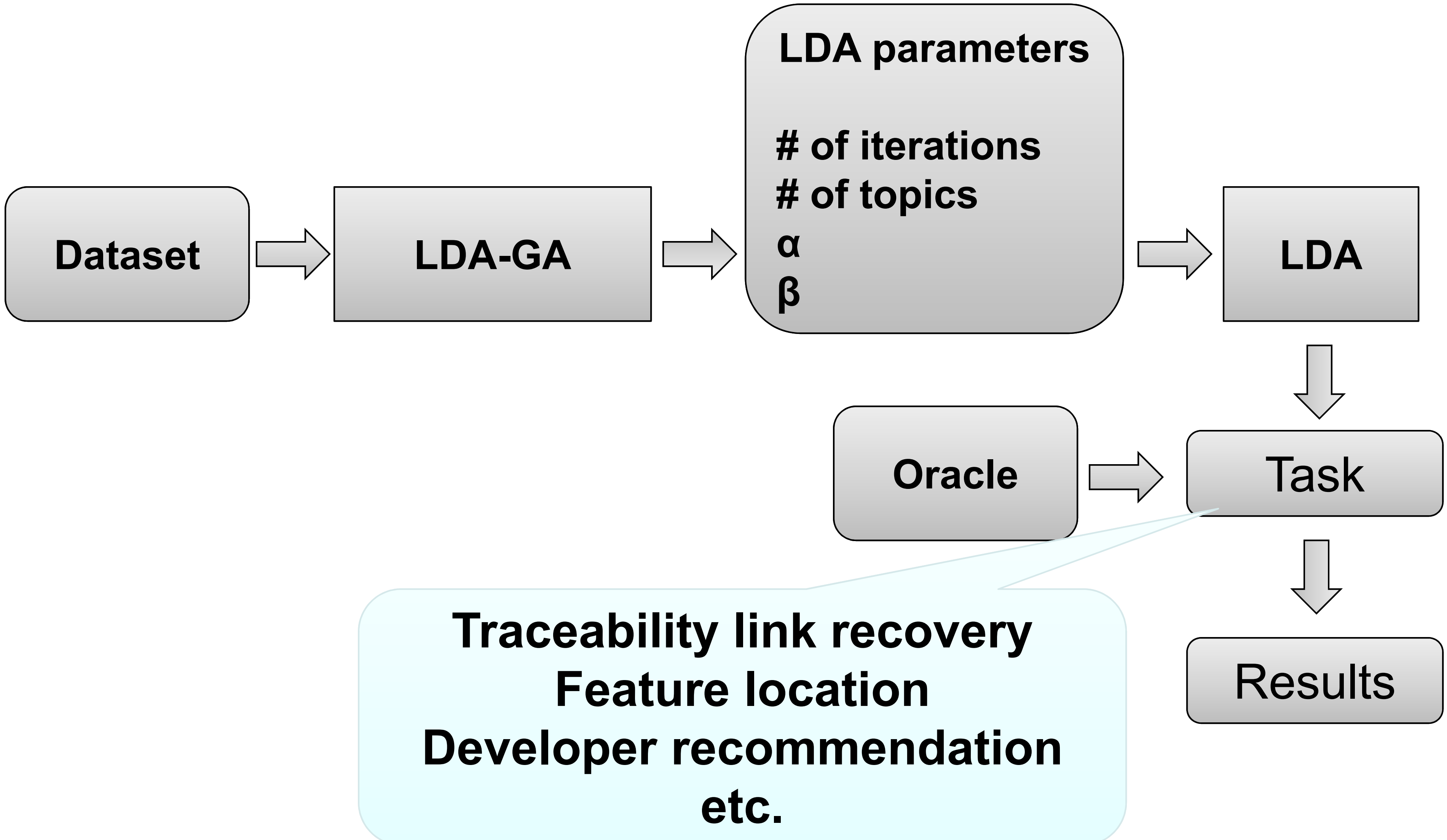
LDA-GA: automatically calibrate the input parameters of LDA using a genetic algorithm



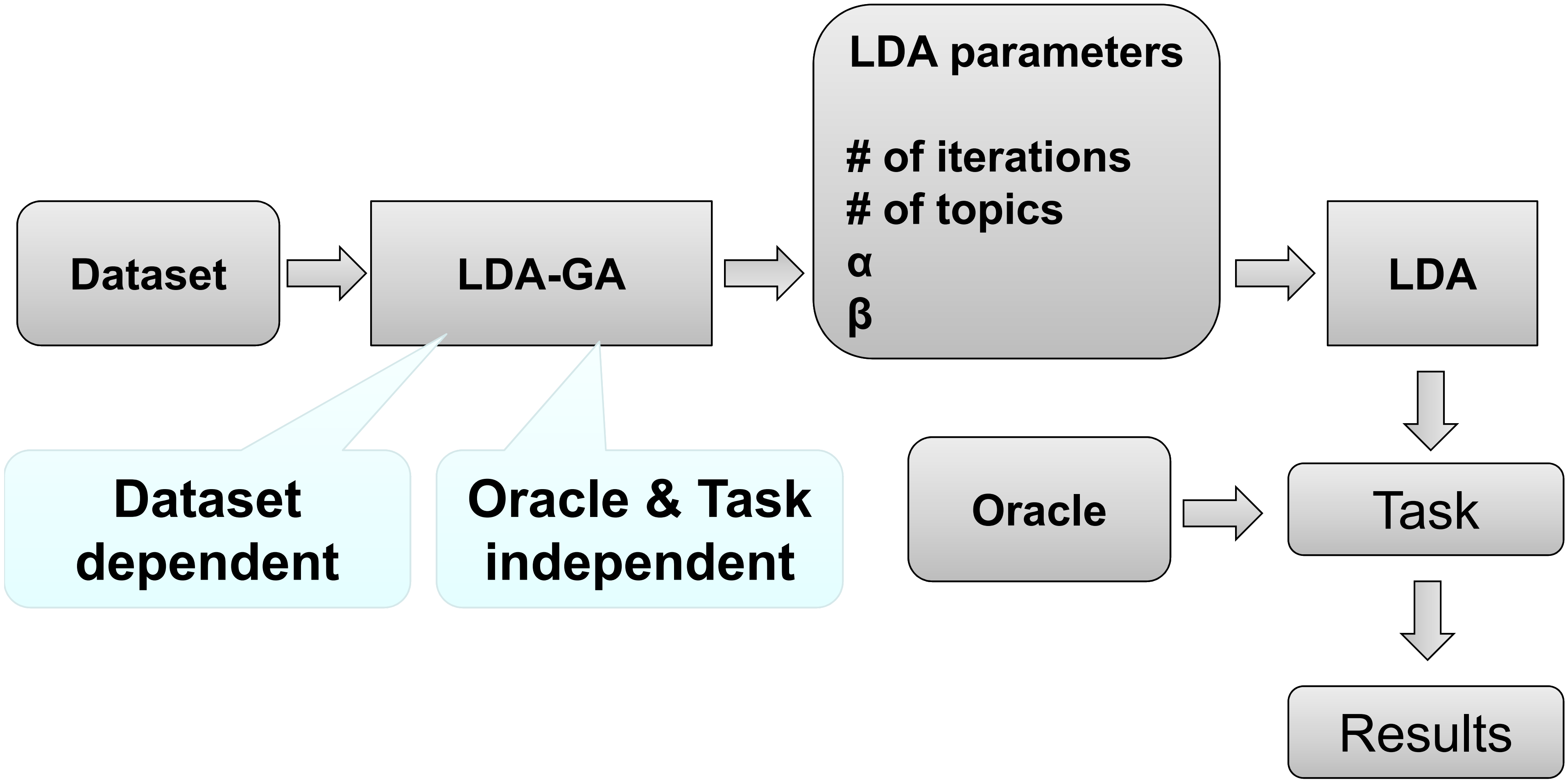
LDA-GA: automatically calibrate the input parameters of LDA using a genetic algorithm



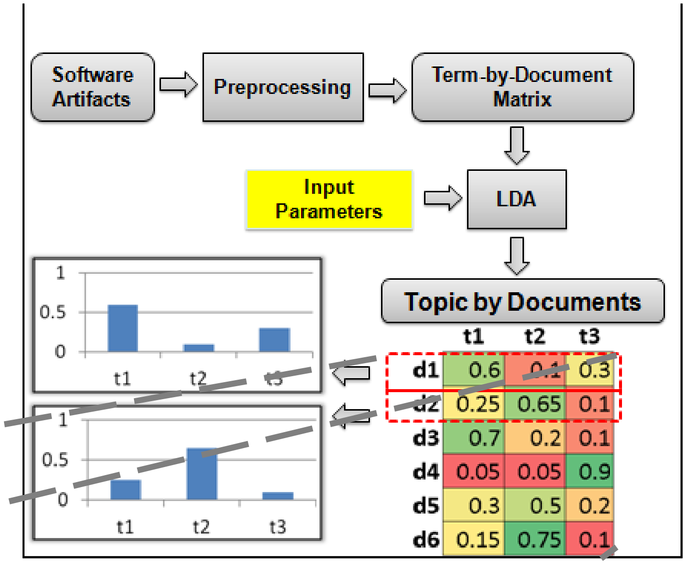
LDA-GA: automatically calibrate the input parameters of LDA using a genetic algorithm



LDA-GA: automatically calibrate the input parameters of LDA using a genetic algorithm



How to **evaluate** how “good” an LDA configuration is?



Topic by Documents

	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1

Topic by Documents

	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1

Topic by Documents

	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1

Dominant Topics

Topic by Documents

	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1

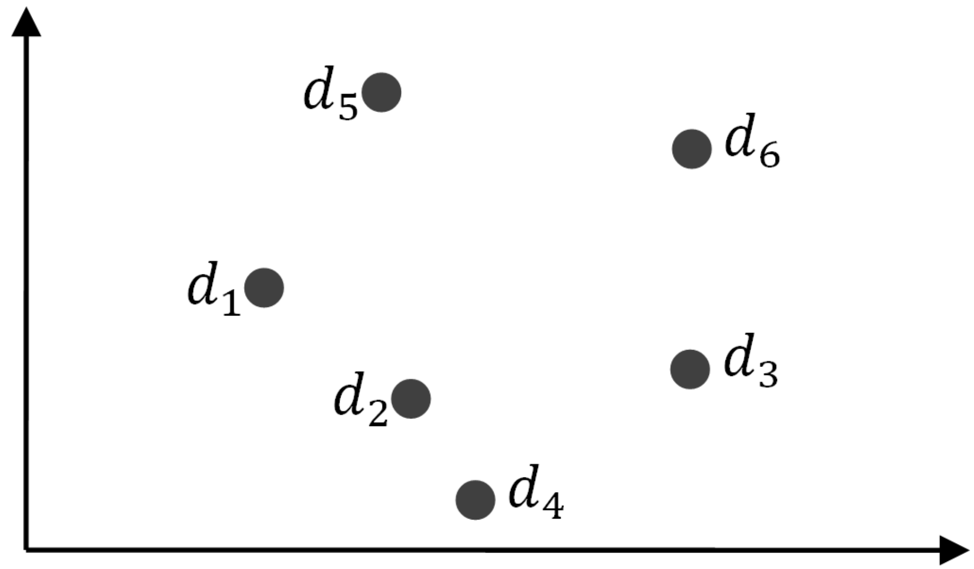
Dominant Topics

Topic by Documents

	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1

Dominant Topics

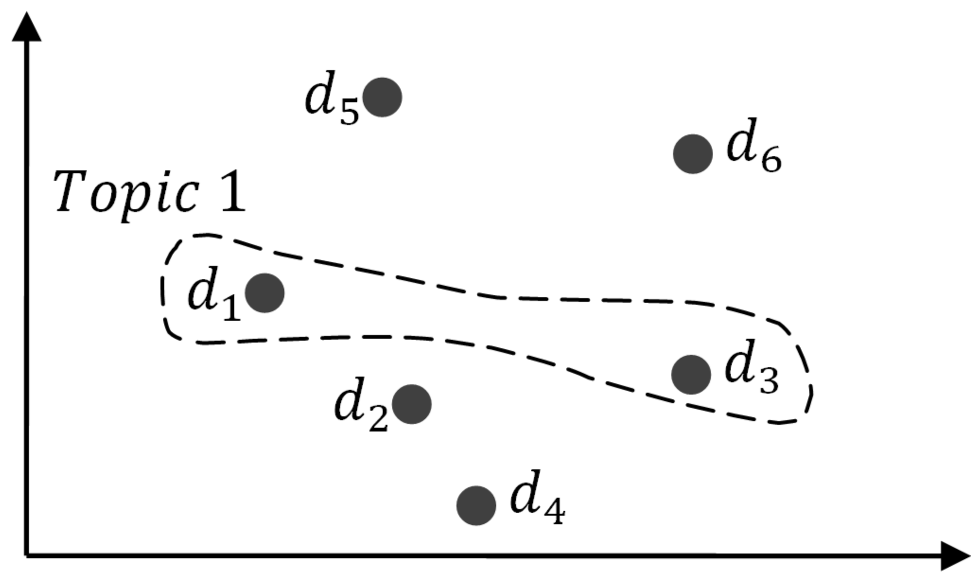
LDA Model



Topic by Documents

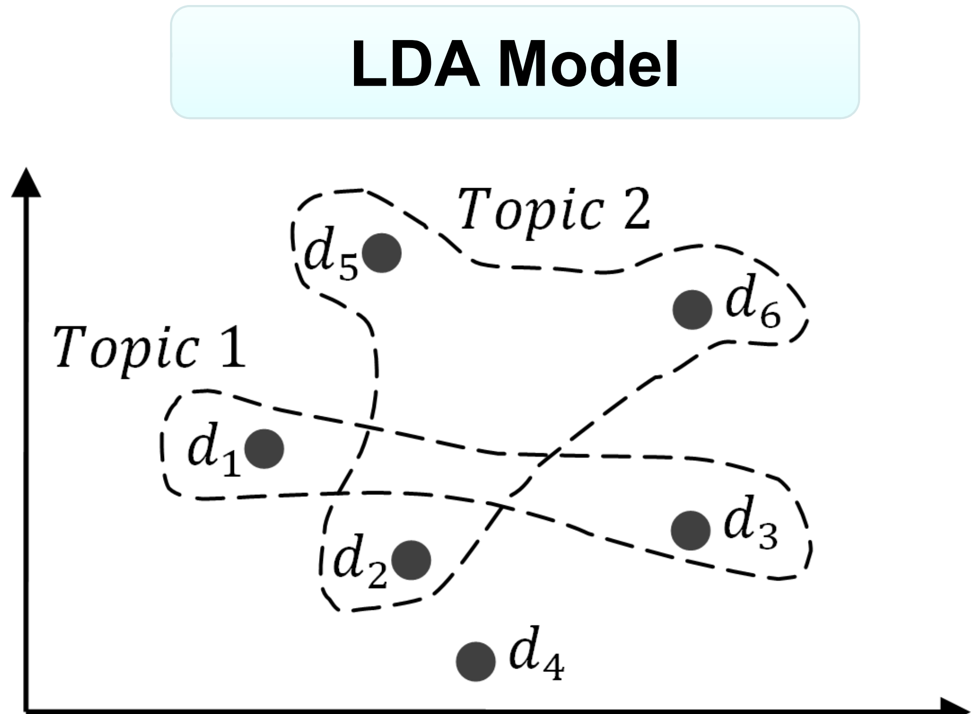
	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1

LDA Model



Topic by Documents

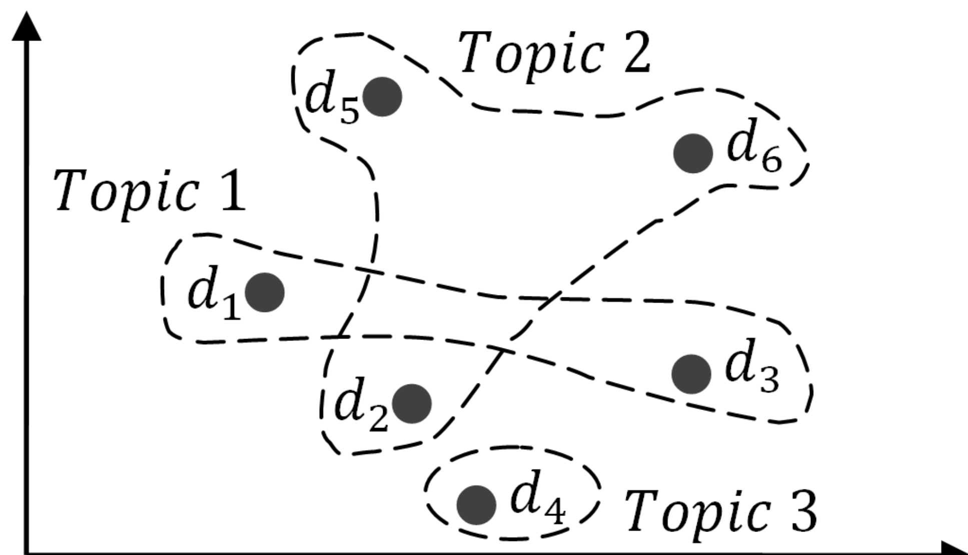
	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1



Topic by Documents

	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1

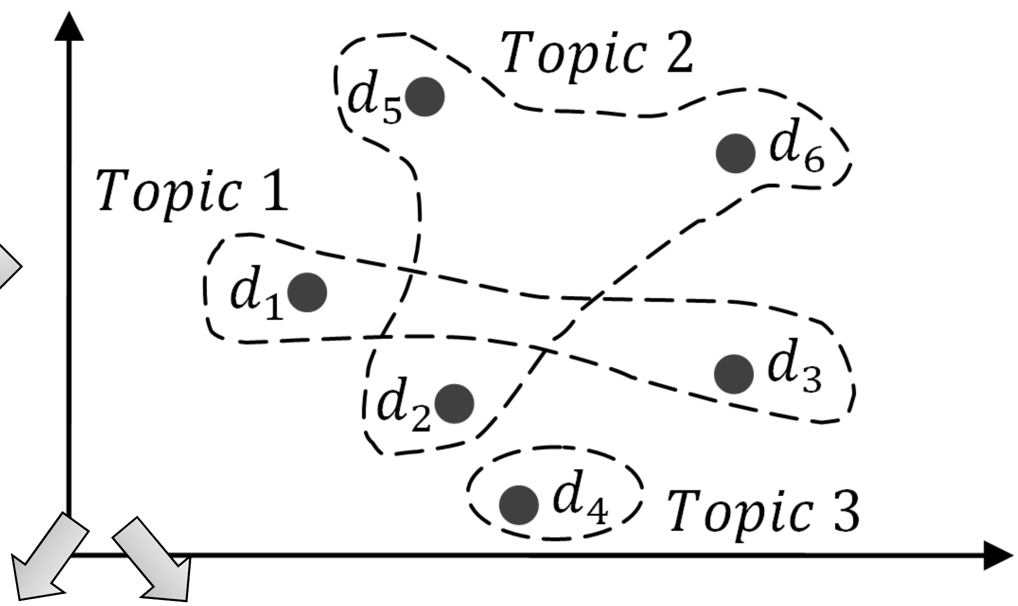
LDA Model



Topic by Documents

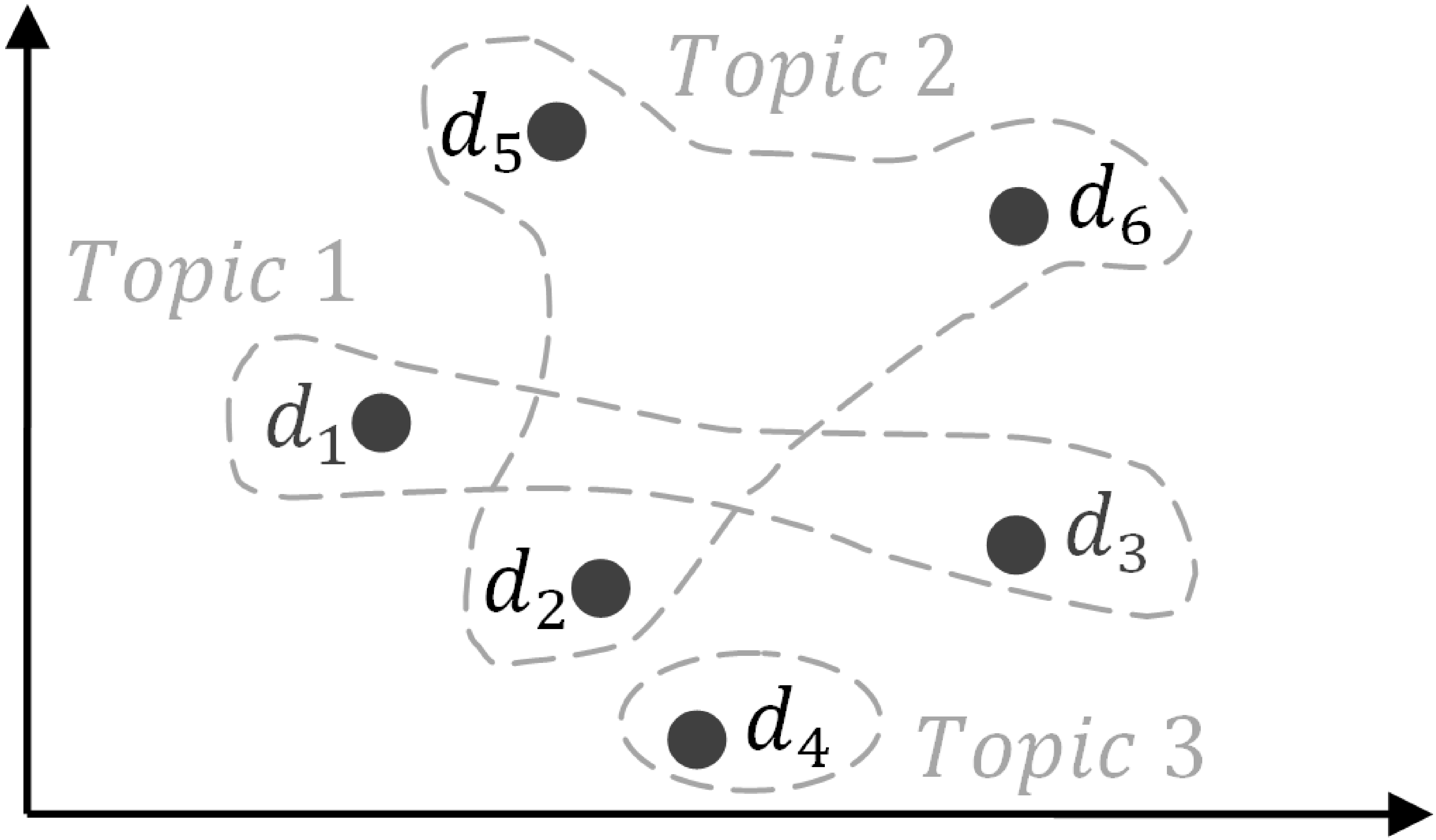
	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2
d6	0.15	0.75	0.1

LDA Model

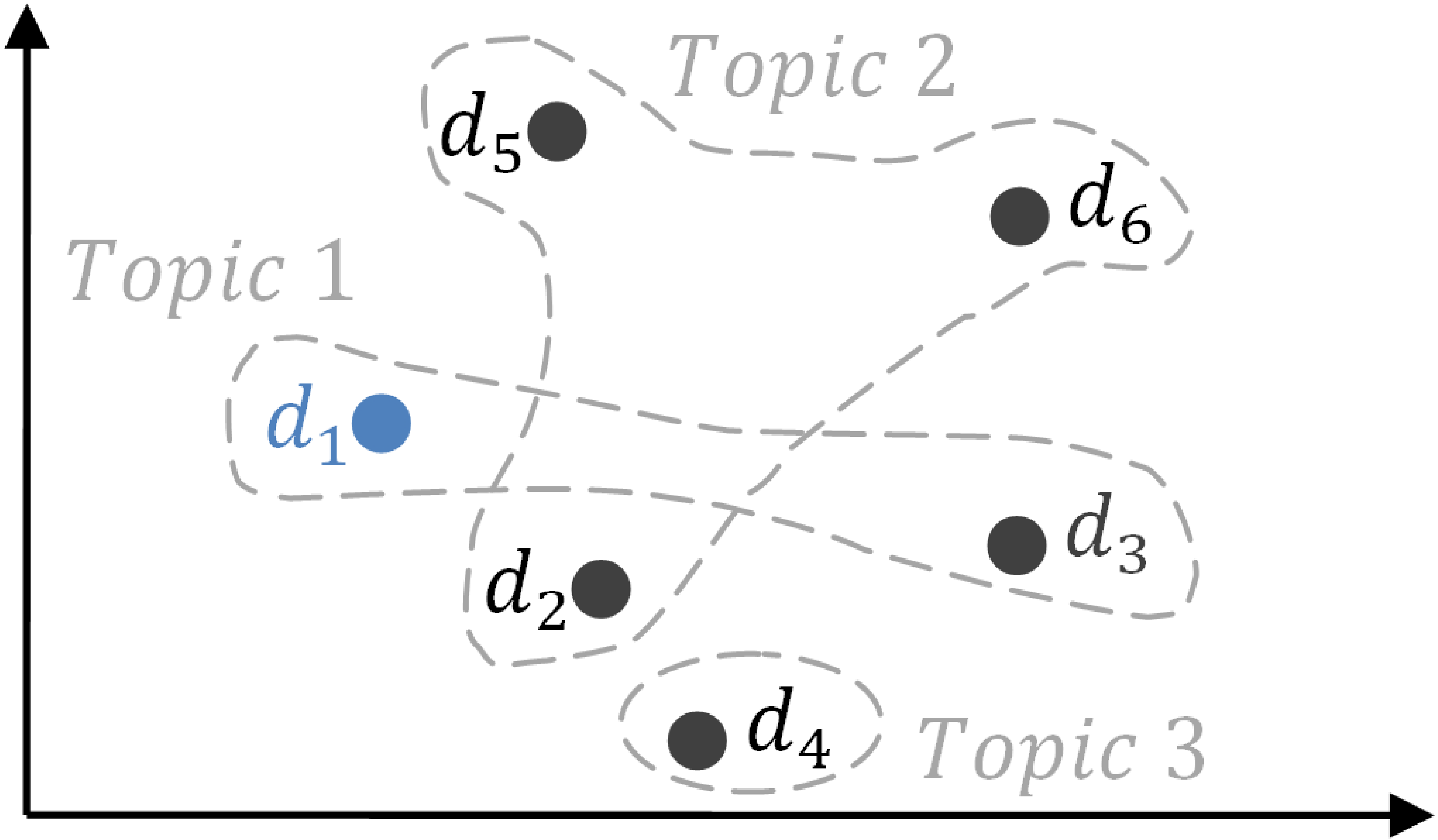


Cohesion (similarity): how related the documents in the same clusters are

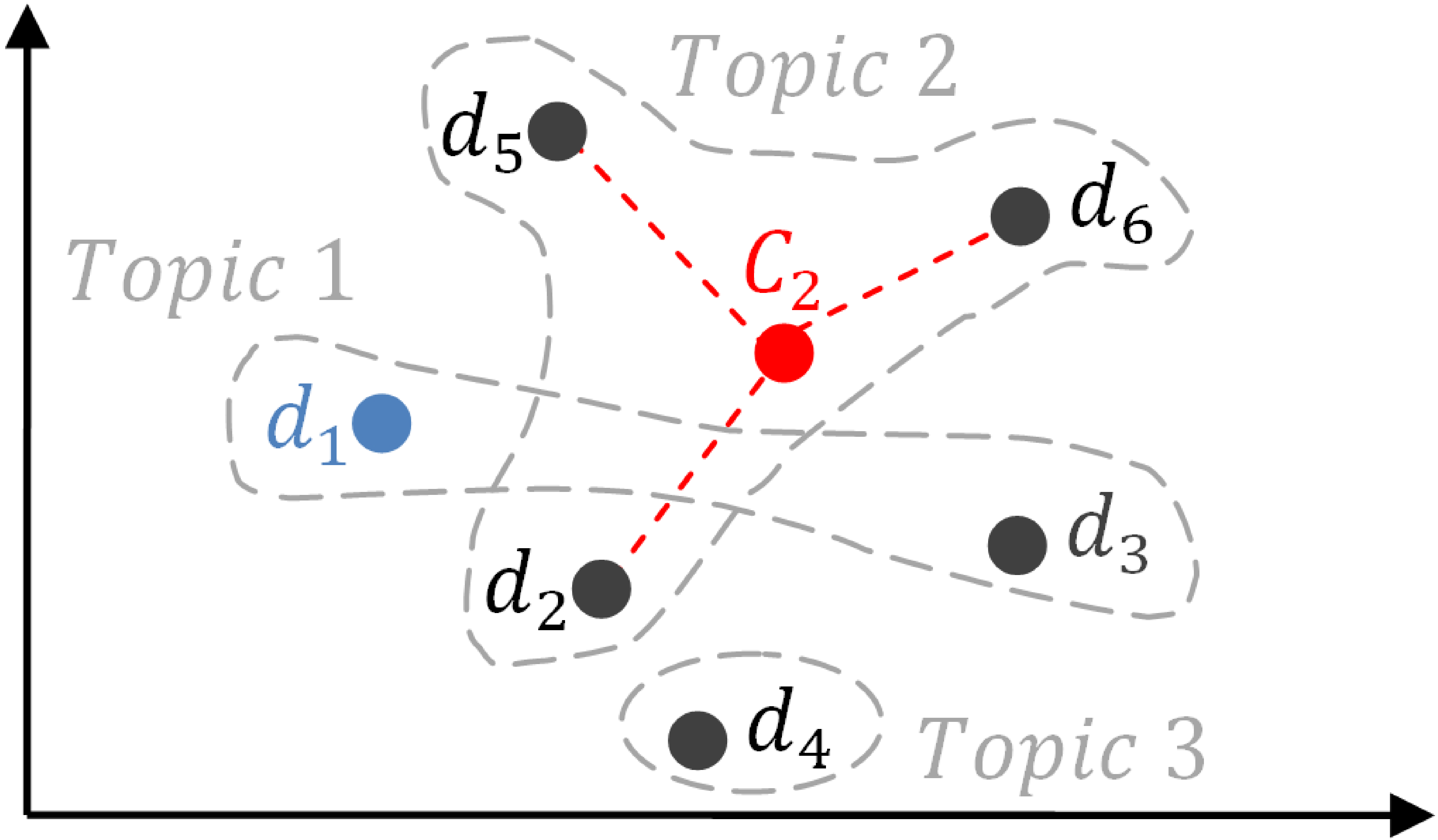
Separation (dissimilarity): how distinct a cluster is from other clusters

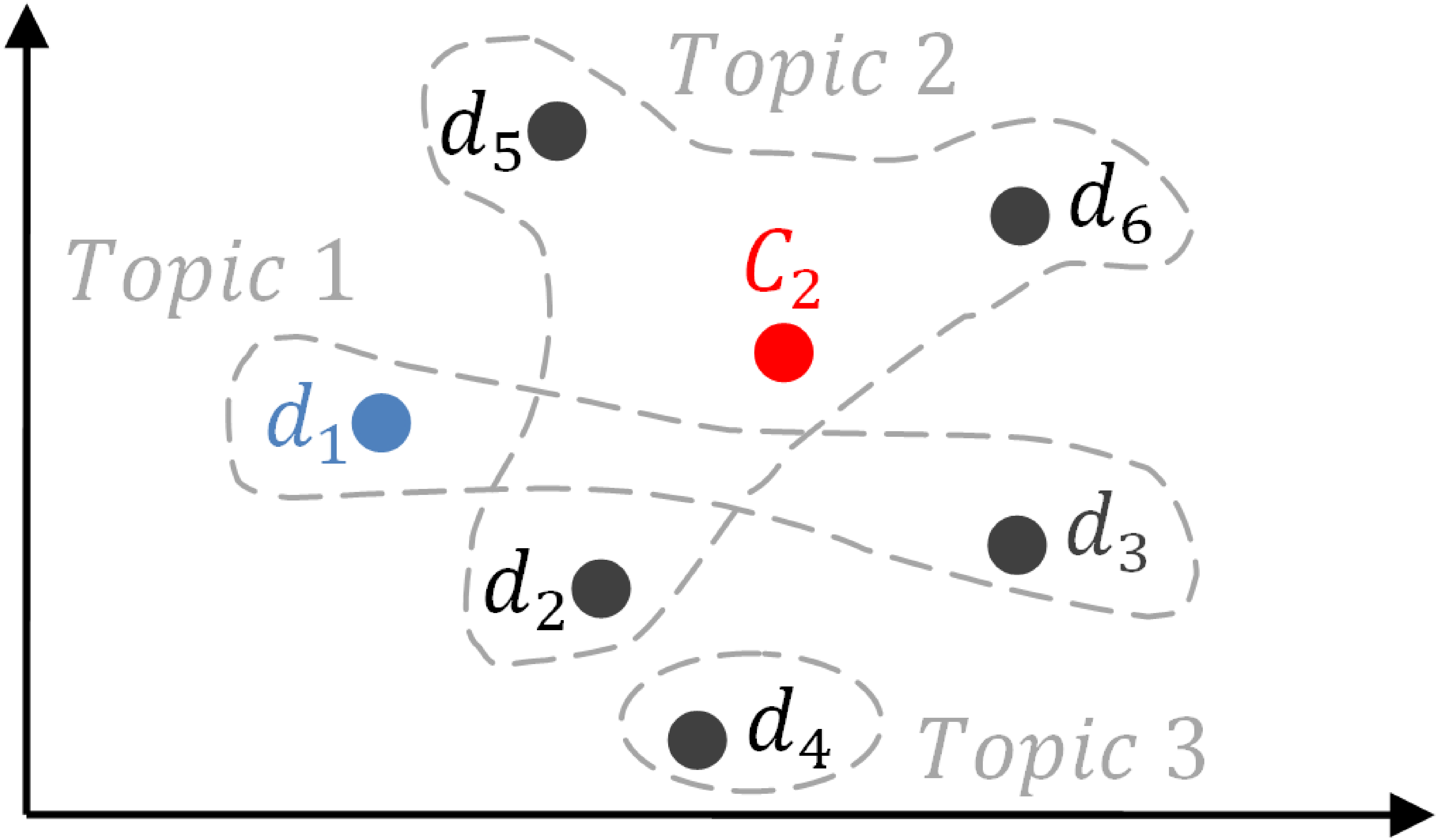


Topic 2
d_5
d_6
Topic 1
d_1
d_3
d_2
d_4
Topic 3

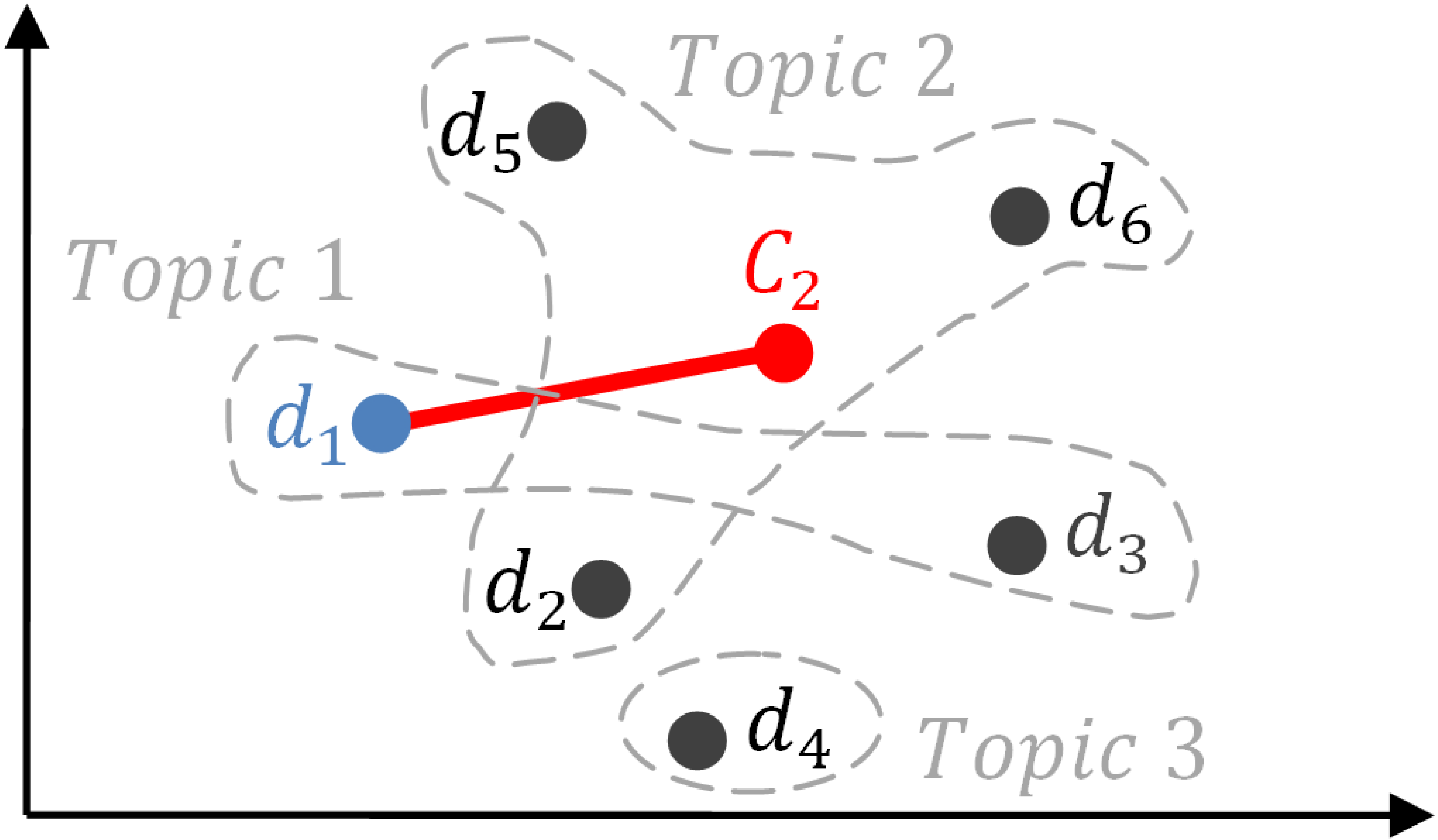


Topic 1
Topic 2
Topic 3
d_1
d_2
d_3
d_4
d_5
d_6

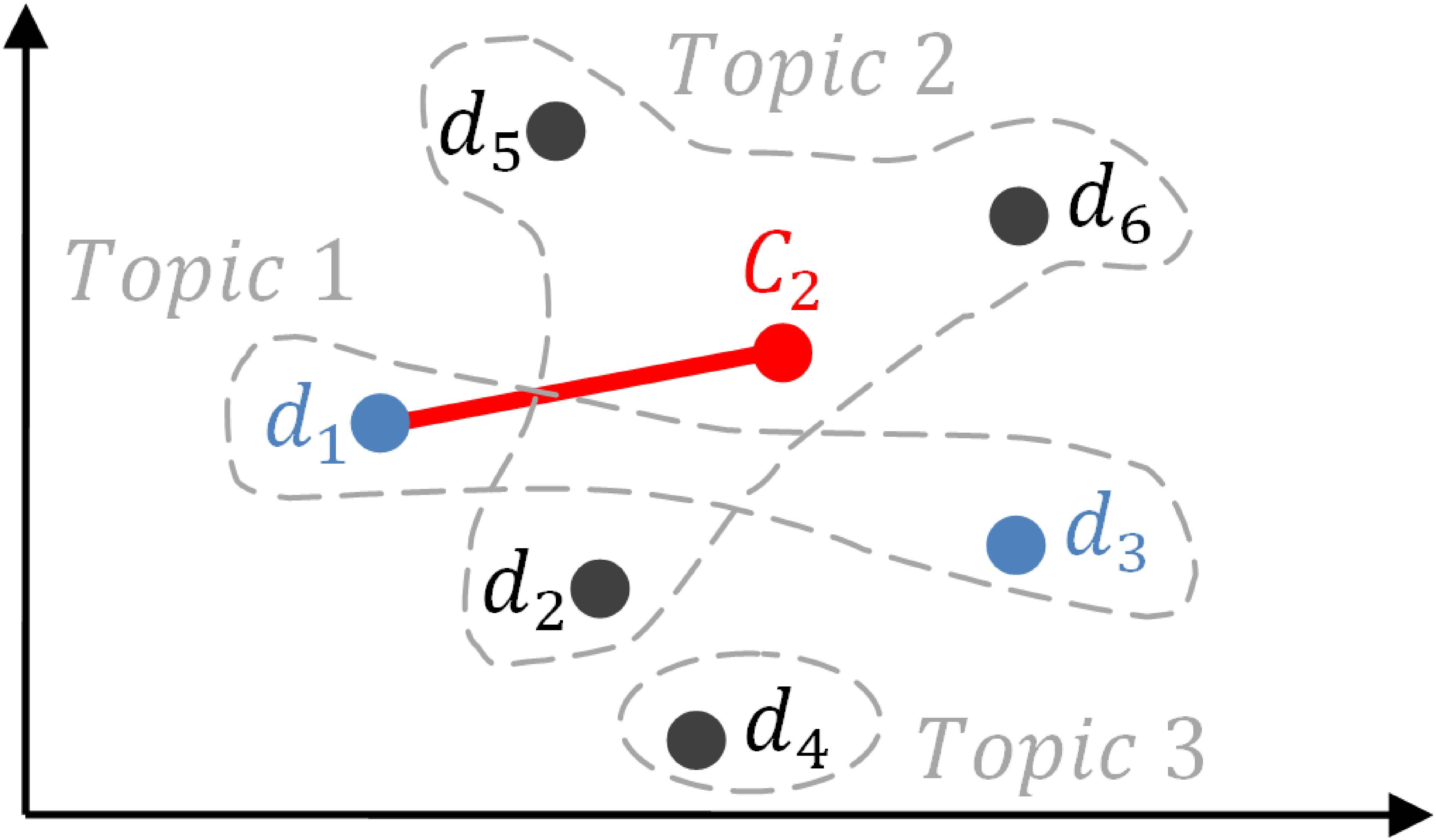

Topic 2
d_5
d_6
c_2
Topic 1
d_1
d_3
d_2
d_4
Topic 3



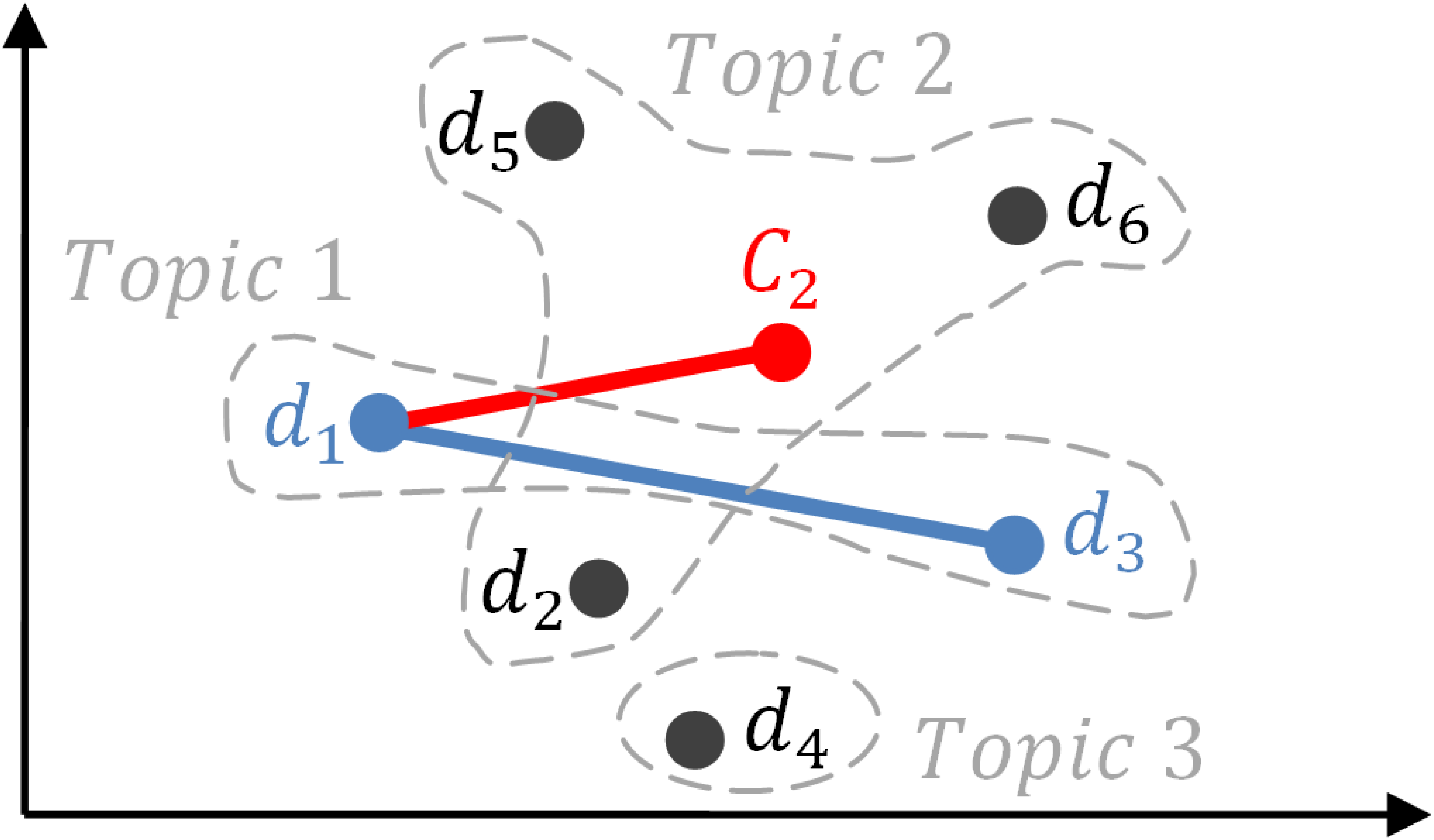
Topic 2
d_5
d_6
C_2
Topic 1
d_1
d_3
d_2
d_4
Topic 3



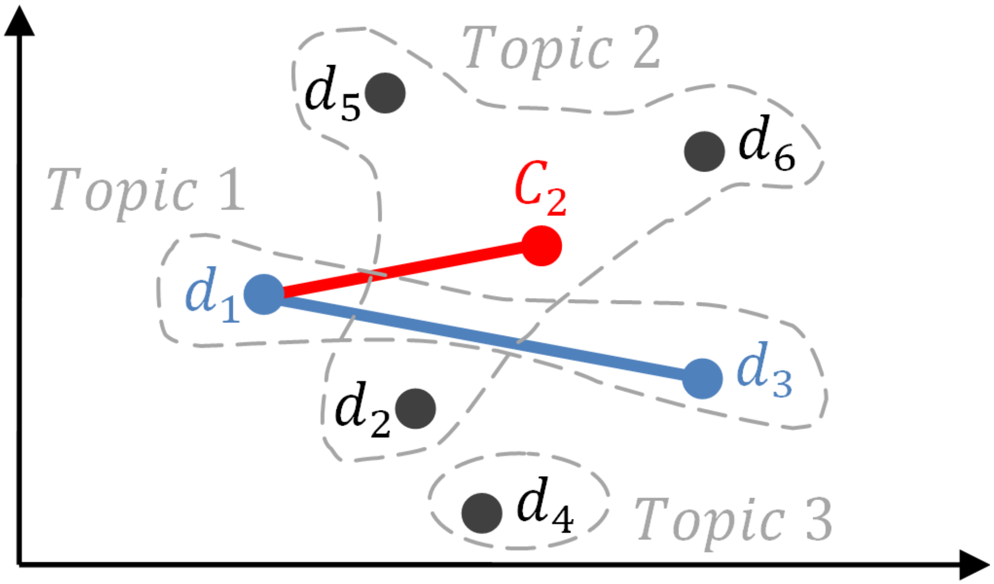
Topic 2
d_5
d_6
Topic 1
C_2
d_1
d_3
d_2
d_4
Topic 3



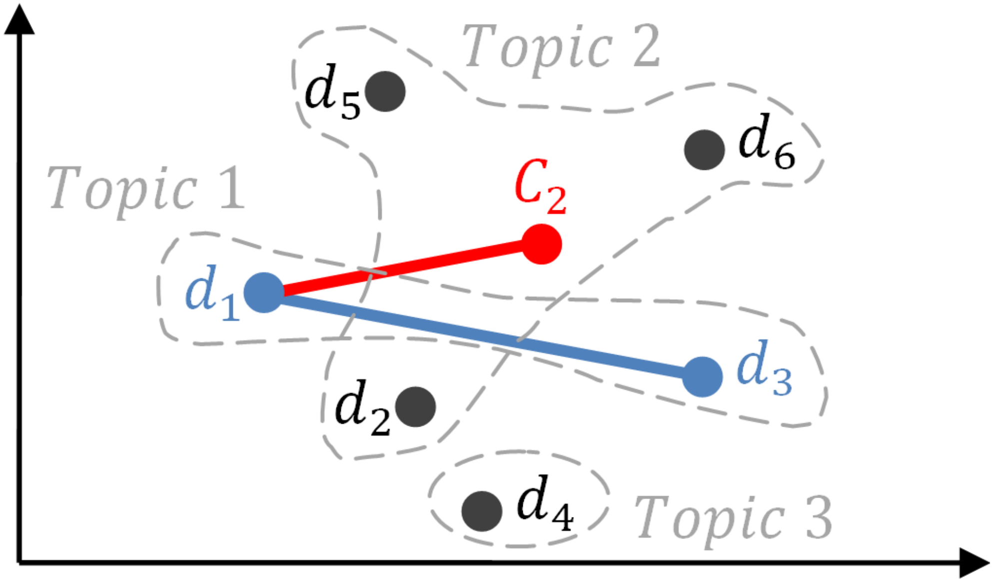
Topic 1
Topic 2
Topic 3
d_1
d_2
d_3
d_4
d_5
d_6
C_2



Topic 2
d_5
d_6
Topic 1
C_2
d_1
d_3
d_2
d_4
Topic 3



$$silhouette(doc) = \frac{red - blue}{\max(red, blue)}$$



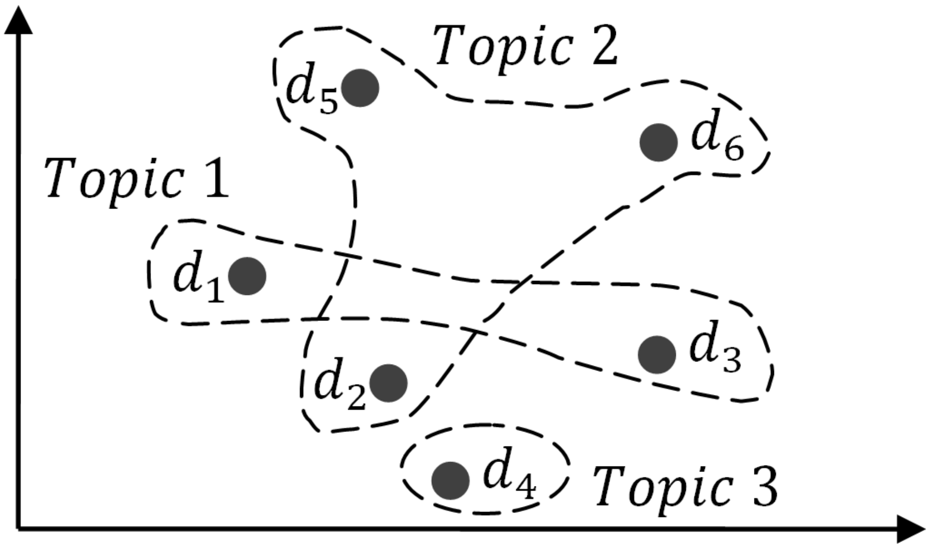
$$\boxed{silhouette(doc) = \frac{red - blue}{\max(red, blue)}}$$

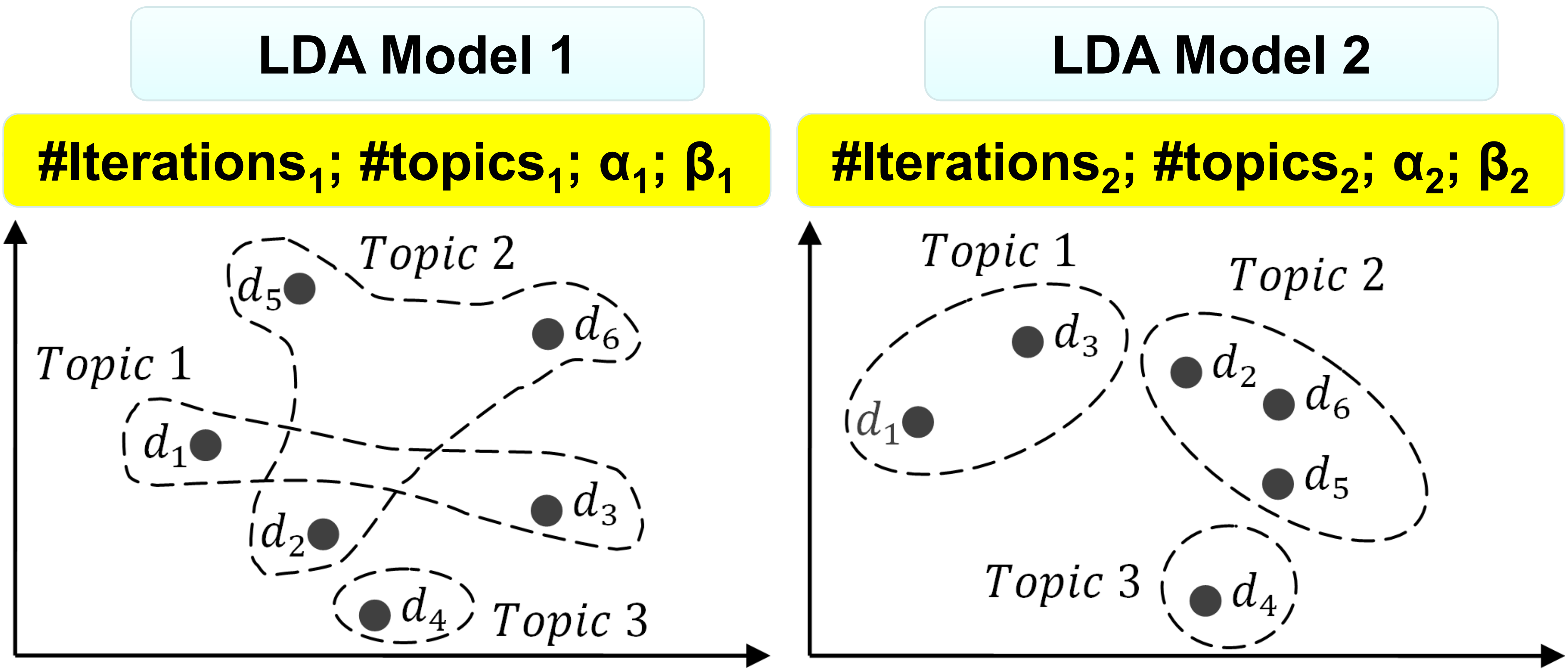
$$\boxed{silhouette(LDA\ model) = \frac{silhouette(doc_1) + \cdots + silhouette(doc_n)}{number\ of\ documents}}$$

Higher silhouette coefficients are better

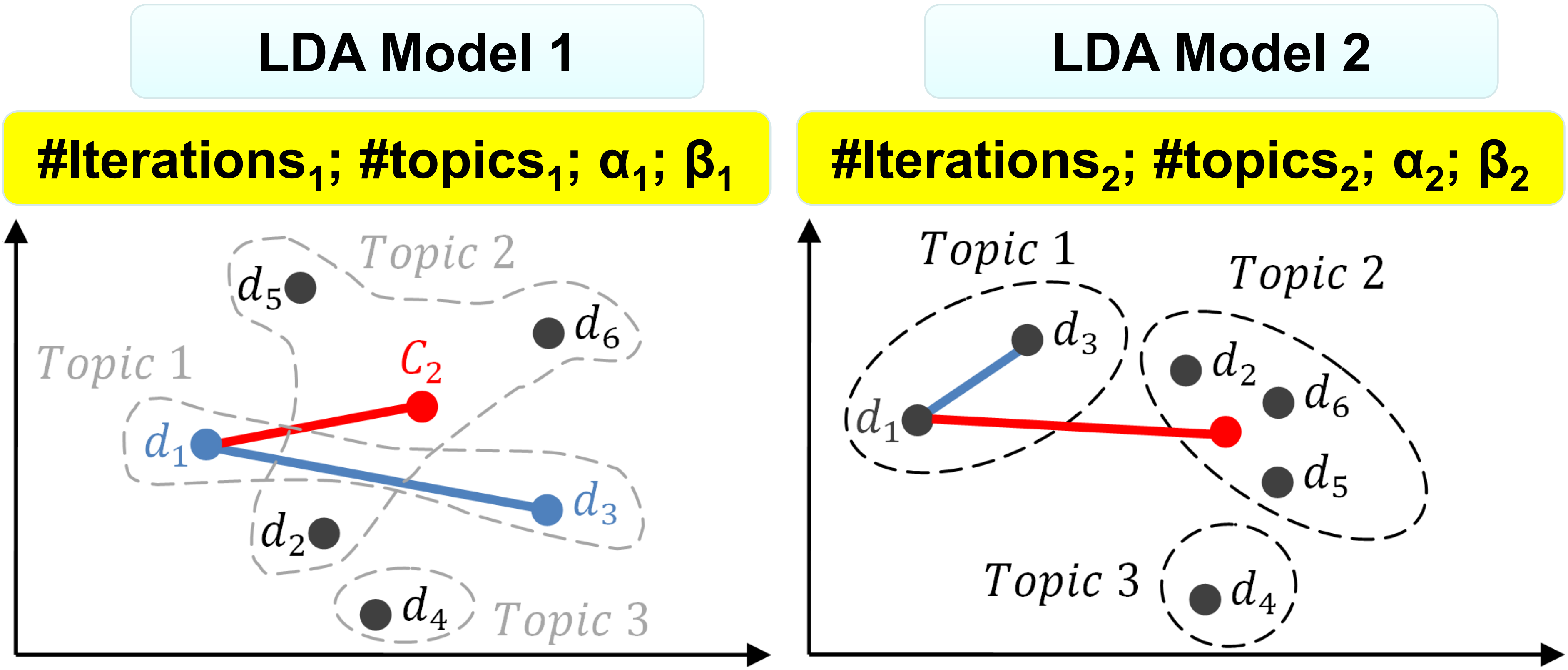
LDA Model 1

#Iterations$_1$; #topics$_1$; α_1; β_1





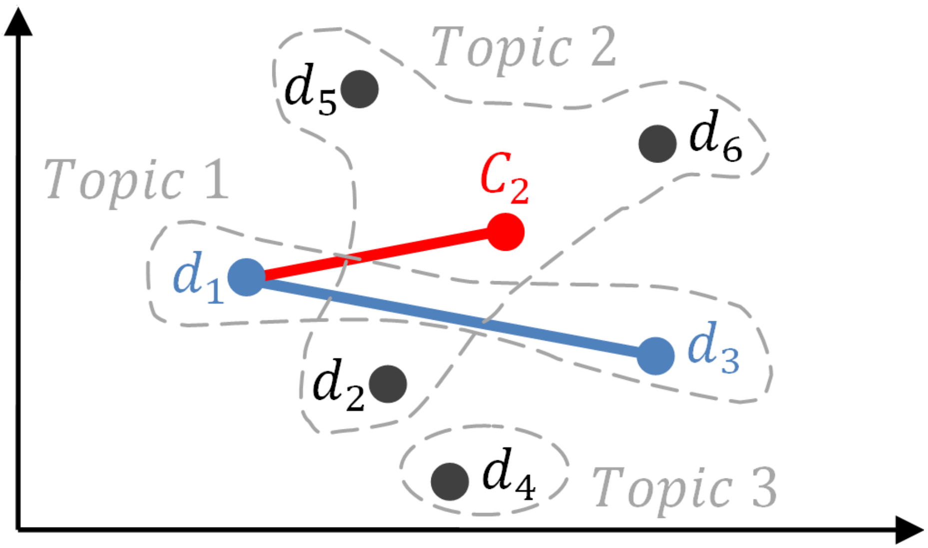
LDA Model 1
#Iterations1; #topics1; α1; β1
Topic 2
Topic 1
Topic 3
d5
d6
d1
d3
d2
d4
LDA Model 2
#Iterations2; #topics2; α2; β2
Topic 1
Topic 2
Topic 3
d3
d1
d2
d6
d5
d4



LDA Model 1
#Iterations$_1$; #topics$_1$; α$_1$; β$_1$
Topic 2
d_5
d_6
Topic 1
C_2
d_1
d_3
d_2
d_4
Topic 3
LDA Model 2
#Iterations$_2$; #topics$_2$; α$_2$; β$_2$
Topic 1
Topic 2
d_3
d_2
d_6
d_1
d_5
Topic 3
d_4

LDA Model 1

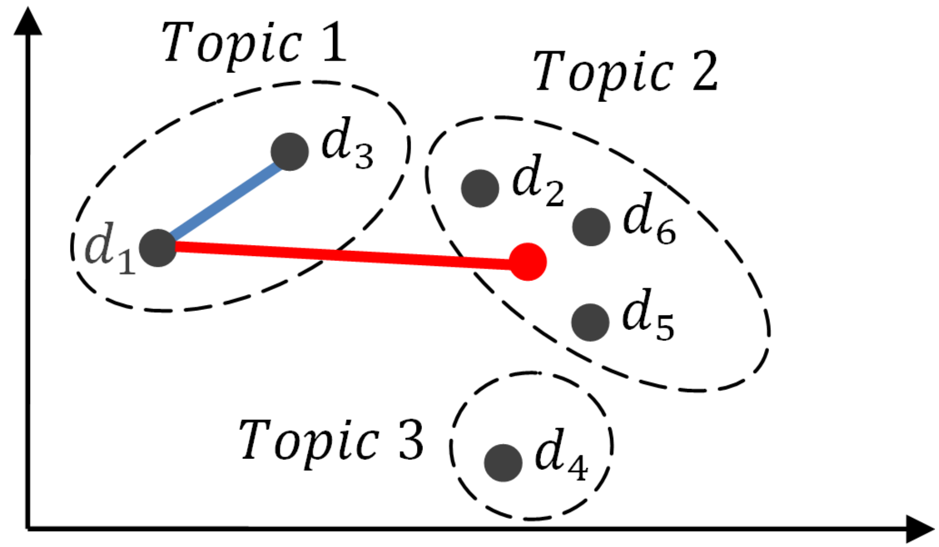
#Iterations$_1$; #topics$_1$; α_1; β_1



silhouette = 0.3

LDA Model 2

#Iterations$_2$; #topics$_2$; α_2; β_2



silhouette = 0.9

Clusters more cohesive
Clusters well separated

How to **evaluate** how "good" an LDA configuration is?

How to **identify** the "good" LDA parameter configurations?

How to identify the "good" LDA parameter configurations?

for numIter in [500, …]
    for numTopics in [5, …]
        for α in [0.01, …]
            for β in [0.01, …]
                LDA[numIter , numTopics , α, β]

Exhaustive approach:
- Discretize search space & iterate?

How to identify the "good" LDA parameter configurations?

~~**for** numIter in [500, …]~~
~~**for** numTopics in [5, …]~~
~~**for** α in [0.01, …]~~
~~**for** β in [0.01, …]~~
~~LDA[numIter , numTopics , α, β]~~

Exhaustive approach:
- Discretize search space & iterate?

Too many possibilities

Use a Genetic Algorithm

What is a **G**enetic **A**lgorithm (GA)?

- Stochastic search technique based on the process of natural evolution to identify *near-optimal solutions* to search problems

Choose a random population of LDA parameters

Choose a random population of LDA parameters

First generation: Population of random chromosomes

	iteration	topics	α	β
LDA Cfg. 1	510	74	0.34	2.5
LDA Cfg. 2	725	128	1.28	0.4
LDA Cfg. 3	814	97	0.43	0.9
...	...	...	...	...
LDA Cfg. n	618	250	1.14	0.74

Choose a random population of LDA parameters

	iteration	topics	α	β
LDA Cfg. 1	510	74	0.34	2.5
LDA Cfg. 2	725	128	1.28	0.4
LDA Cfg. 3	814	97	0.43	0.9
…	…	…	…	…
LDA Cfg. n	618	250	1.14	0.74

First generation: Population of random chromosomes

Individual (chromosome): Represents one possible LDA parameter configuration

Choose a random population of LDA parameters

	iteration	topics	α	β
LDA Cfg. 1	510	74	0.34	2.5
LDA Cfg. 2	725	128	1.28	0.4
LDA Cfg. 3	814	97	0.43	0.9
…	…	…	…	…
LDA Cfg. n	618	250	1.14	0.74

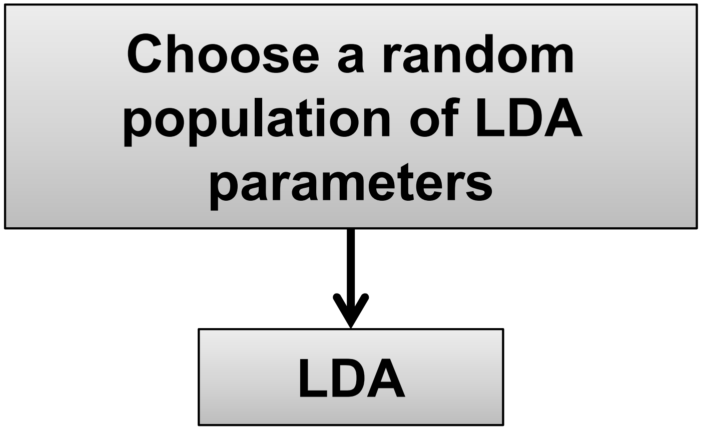
First generation:
Population of random chromosomes

Individual (chromosome):
Represents one possible LDA parameter configuration

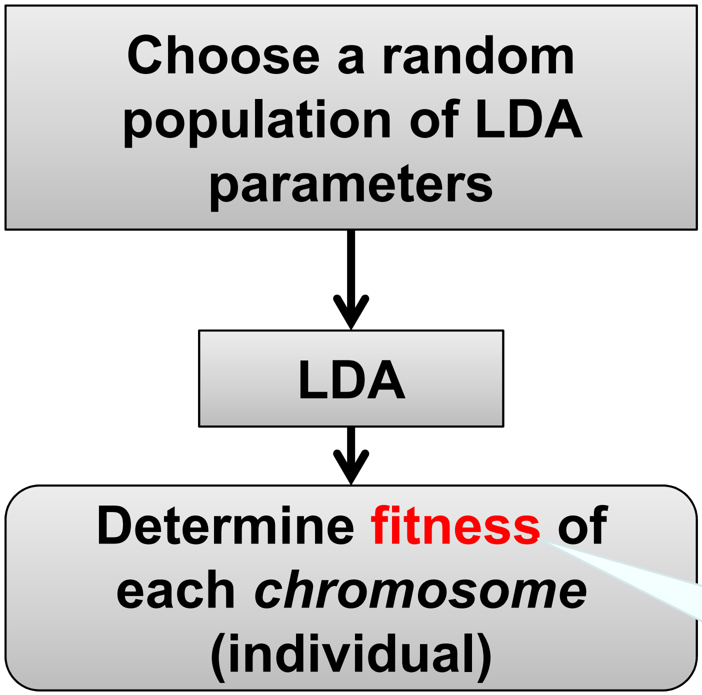
Gene:
LDA parameter

Choose a random population of LDA parameters

	iteration	topics	α	β
LDA Cfg. 1	510	74	0.34	2.5
LDA Cfg. 2	725	128	1.28	0.4
LDA Cfg. 3	814	97	0.43	0.9
...	...	...	...	...
LDA Cfg. n	618	250	1.14	0.74

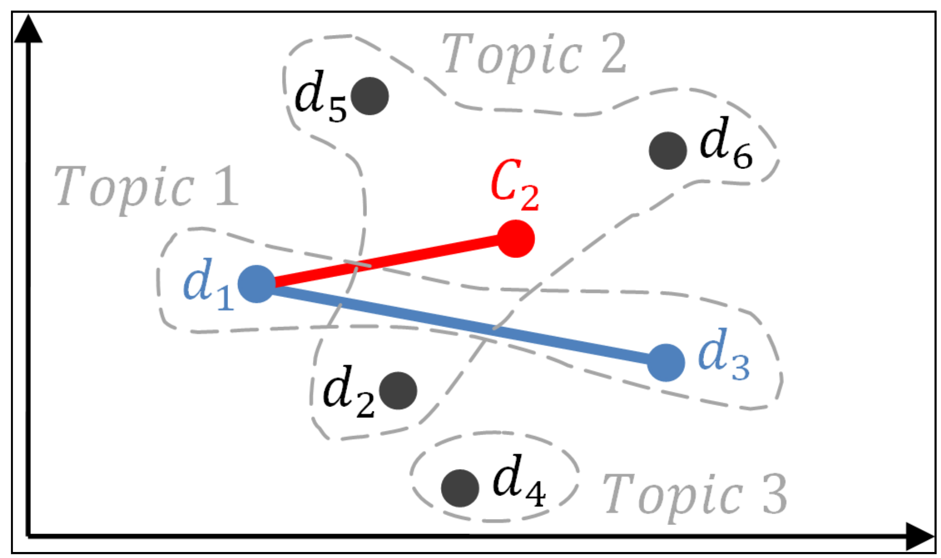


	iteration	topics	α	β
LDA Cfg. 1	510	74	0.34	2.5
LDA Cfg. 2	725	128	1.28	0.4
LDA Cfg. 3	814	97	0.43	0.9
…	…	…	…	…
LDA Cfg. n	618	250	1.14	0.74



	iteration	topics	α	β
LDA Cfg. 1	510	74	0.34	2.5
LDA Cfg. 2	725	128	1.28	0.4
LDA Cfg. 3	814	97	0.43	0.9
...	...	...	...	...
LDA Cfg. n	618	250	1.14	0.74

Fitness = silhouette coefficient



Choose a random population of LDA parameters

↓

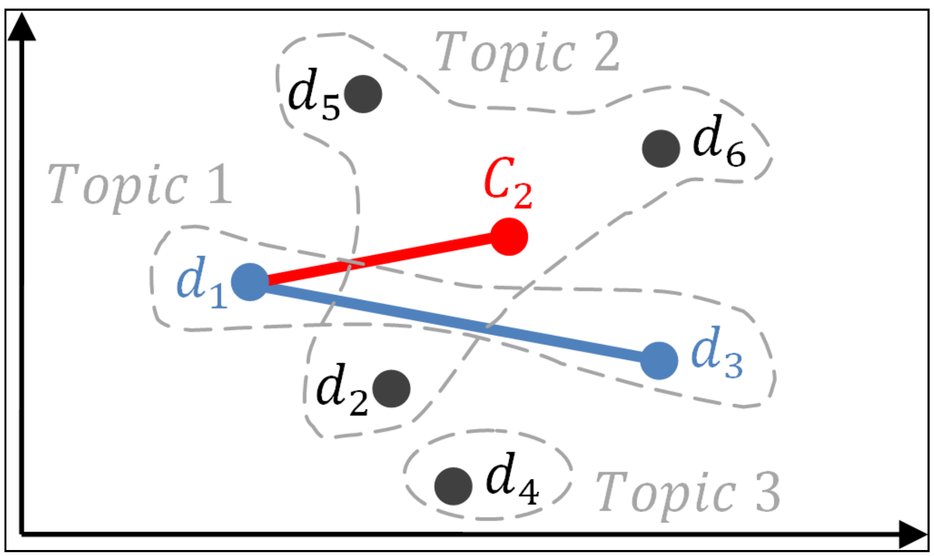
LDA

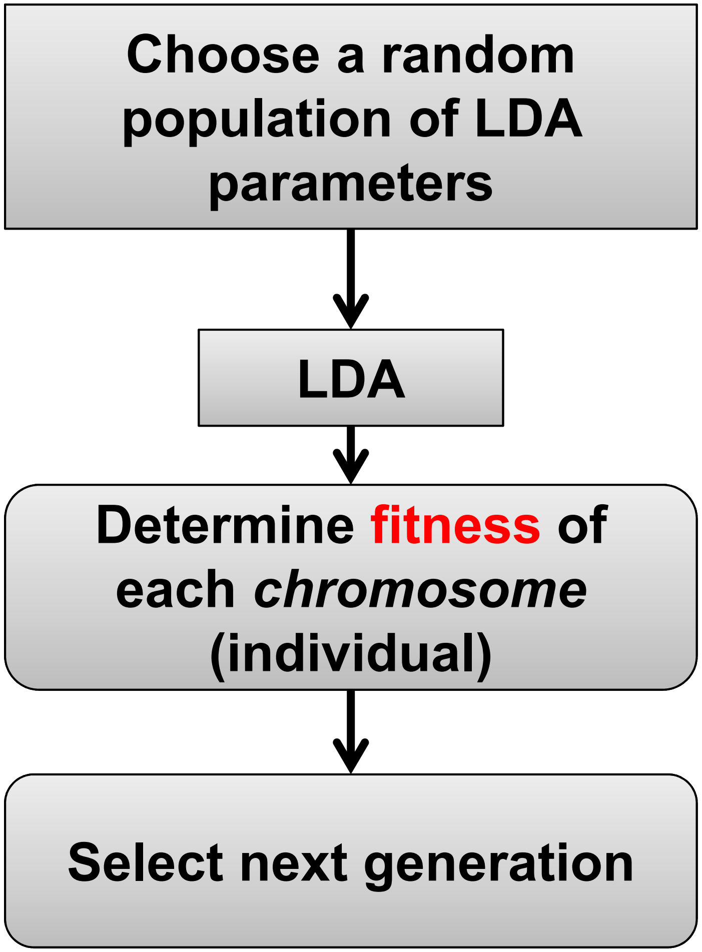
↓

Determine fitness of each *chromosome* (individual)

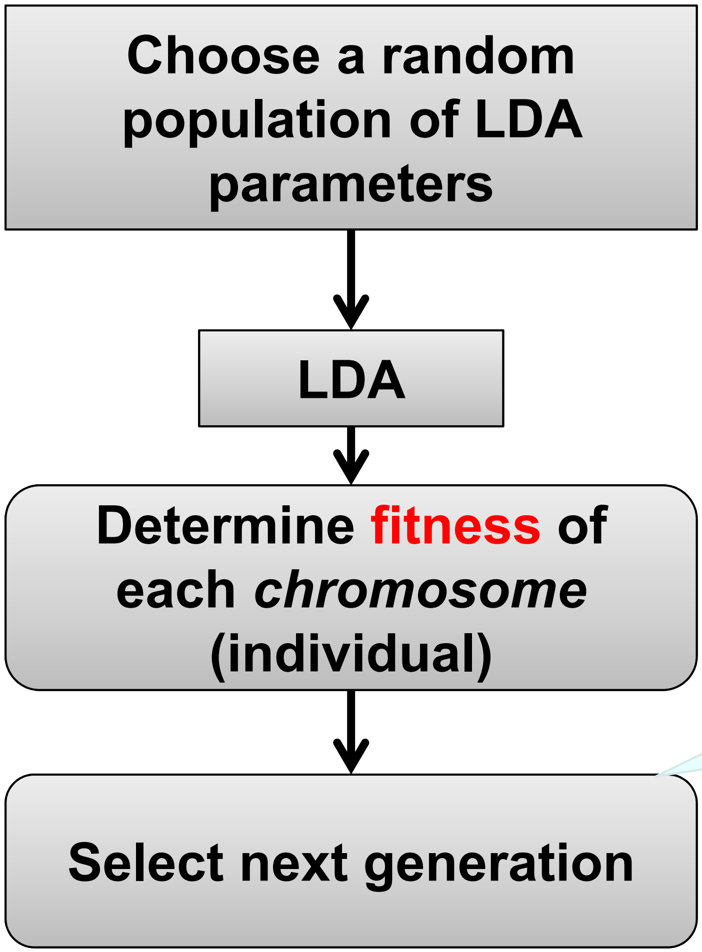
	iteration	topics	α	β	Fitness
LDA Cfg. 1	510	74	0.34	2.5	0.2
LDA Cfg. 2	725	128	1.28	0.4	0.4
LDA Cfg. 3	814	97	0.43	0.9	0.35
…	…	…	…	…	…
LDA Cfg. n	618	250	1.14	0.74	0.1

Fitness = silhouette coefficient



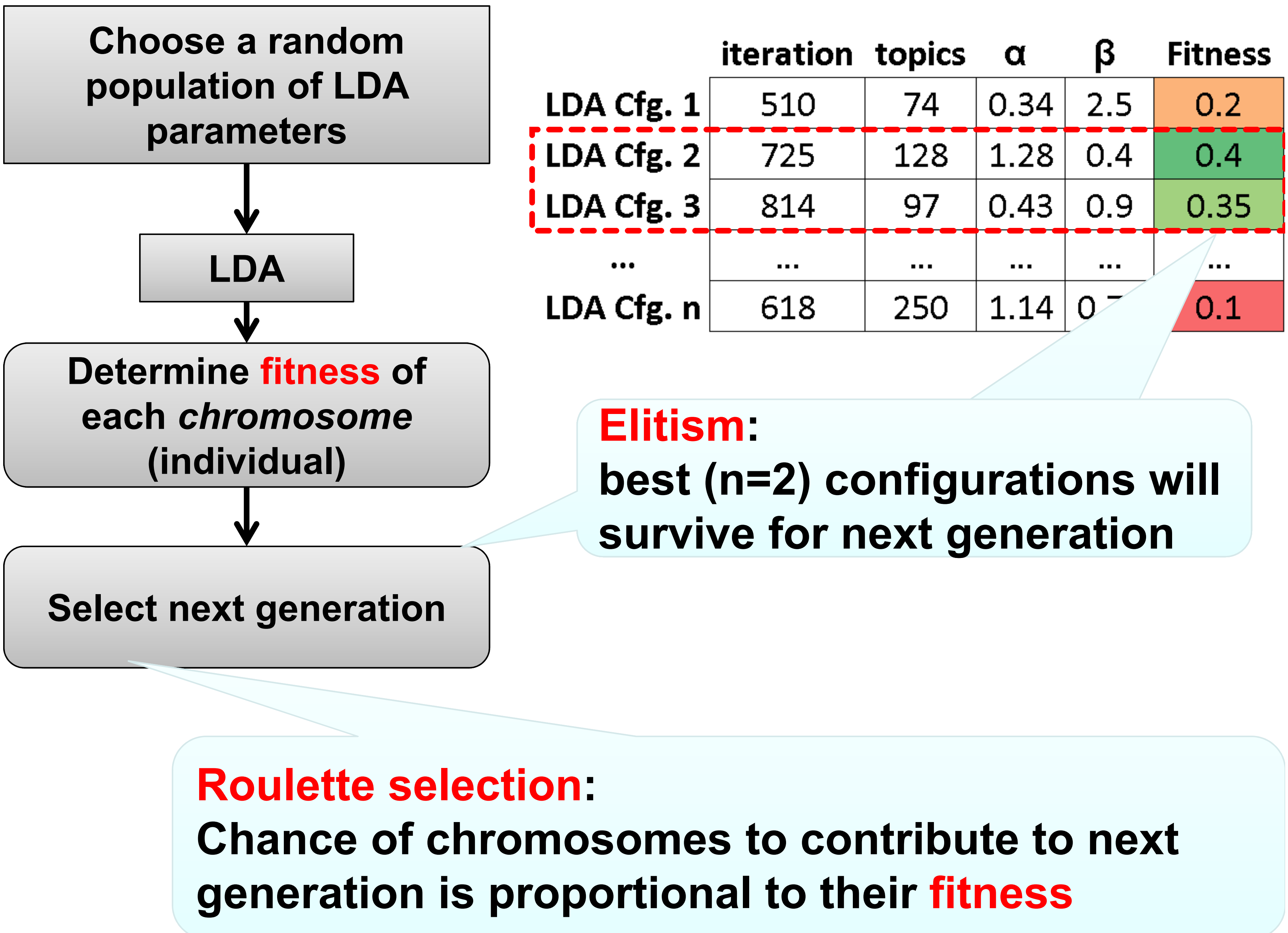


	iteration	topics	α	β	Fitness
LDA Cfg. 1	510	74	0.34	2.5	0.2
LDA Cfg. 2	725	128	1.28	0.4	0.4
LDA Cfg. 3	814	97	0.43	0.9	0.35
...	...	...	...	...	...
LDA Cfg. n	618	250	1.14	0.74	0.1

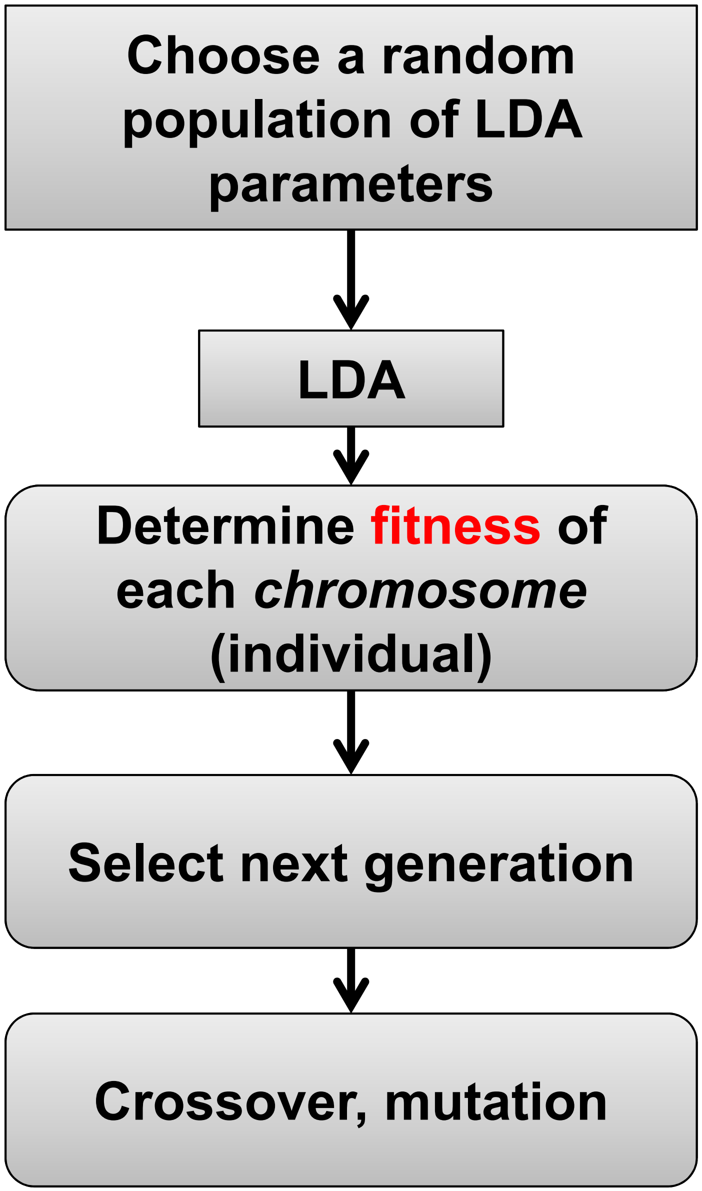


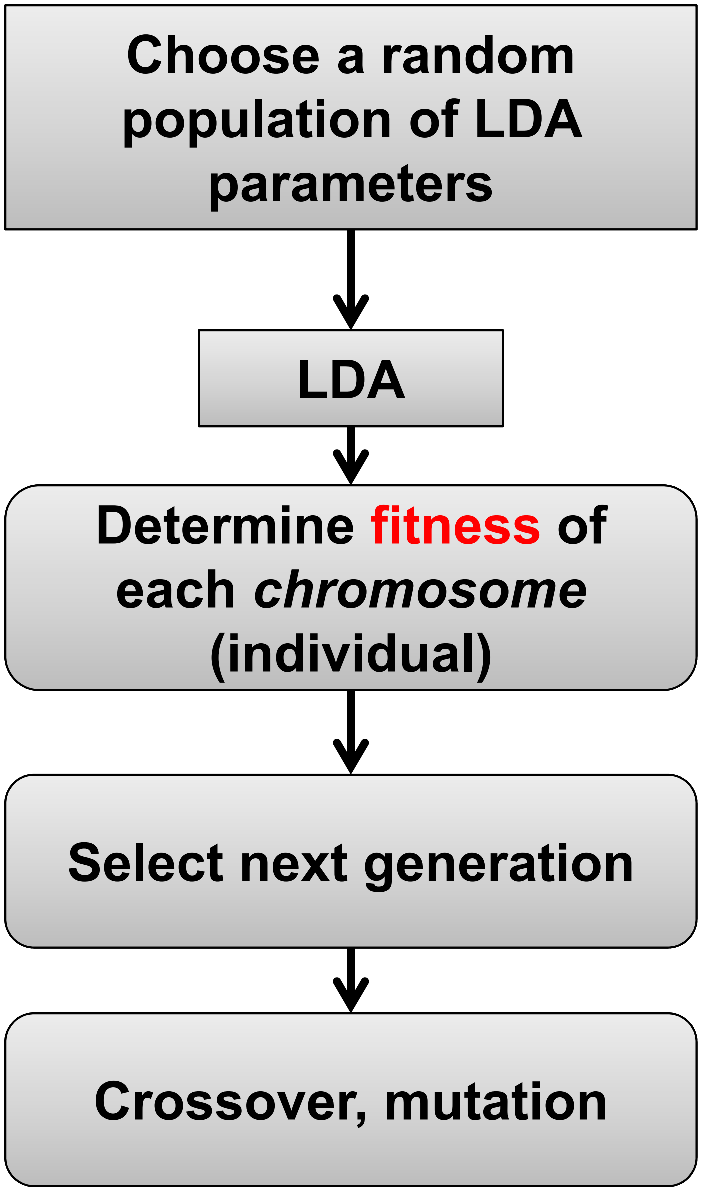
	iteration	topics	α	β	Fitness
LDA Cfg. 1	510	74	0.34	2.5	0.2
LDA Cfg. 2	725	128	1.28	0.4	0.4
LDA Cfg. 3	814	97	0.43	0.9	0.35
…	…	…	…	…	…
LDA Cfg. n	618	250	1.14	0.[illegible]	0.1

Elitism:
best (n=2) configurations will survive for next generation



Choose a random population of LDA parameters
LDA
Determine fitness of each chromosome (individual)
Select next generation
iteration topics α β Fitness
LDA Cfg. 1 510 74 0.34 2.5 0.2
LDA Cfg. 2 725 128 1.28 0.4 0.4
LDA Cfg. 3 814 97 0.43 0.9 0.35
... ...
LDA Cfg. n 618 250 1.14 0. 0.1
Elitism:
best (n=2) configurations will survive for next generation
Roulette selection:
Chance of chromosomes to contribute to next generation is proportional to their fitness

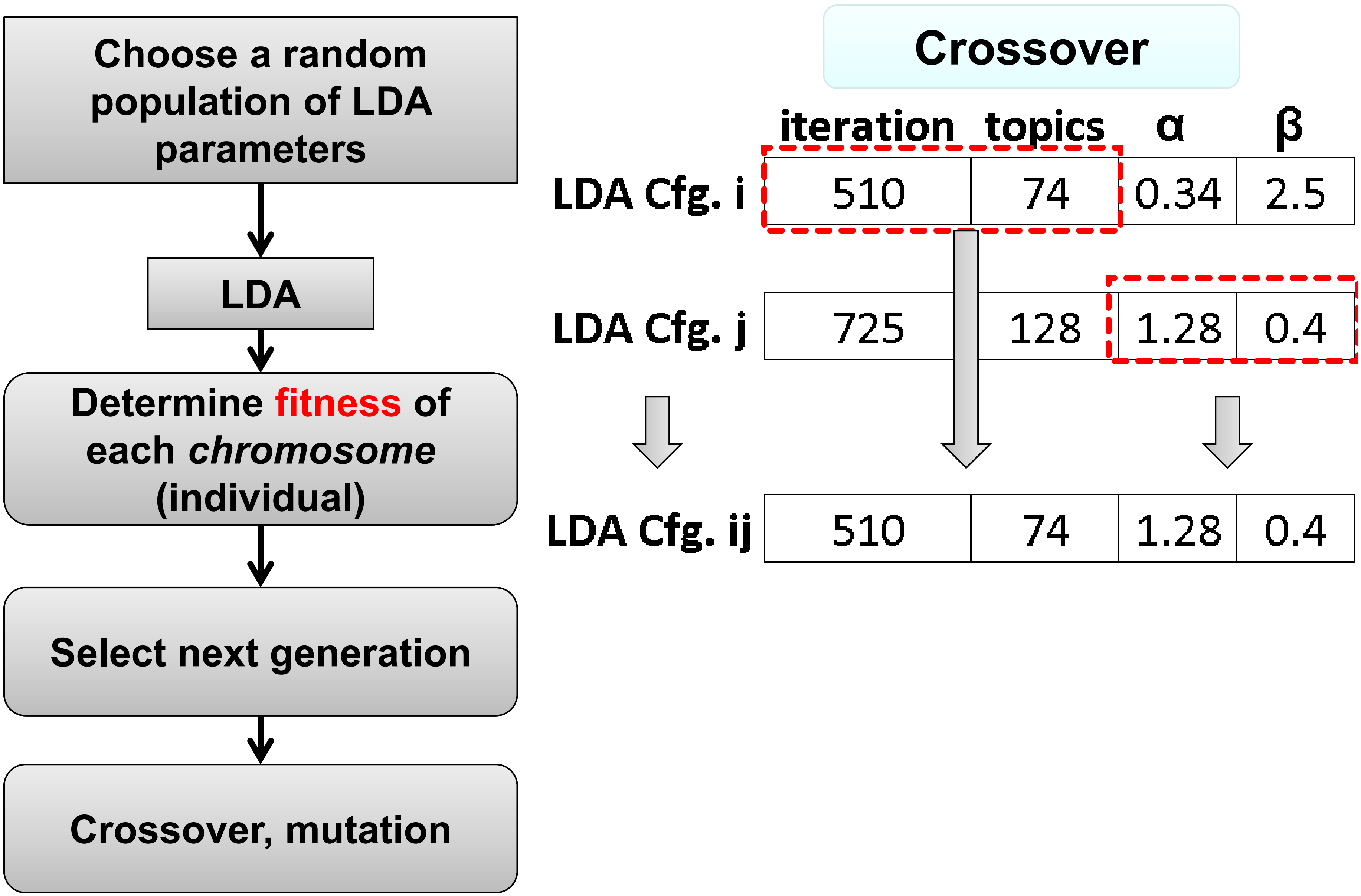

Choose a random population of LDA parameters
LDA
Determine fitness of each chromosome (individual)
Select next generation
Crossover, mutation

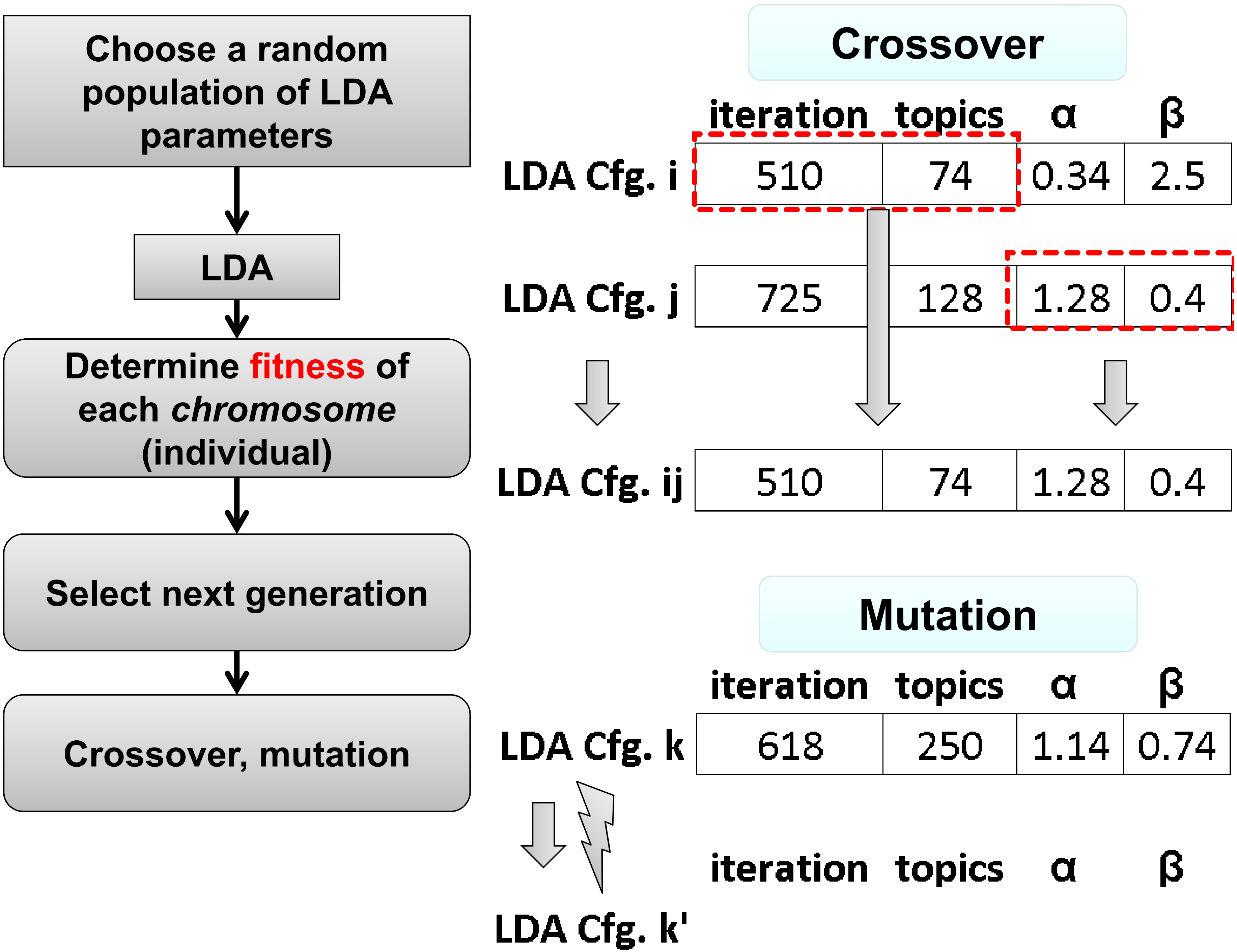


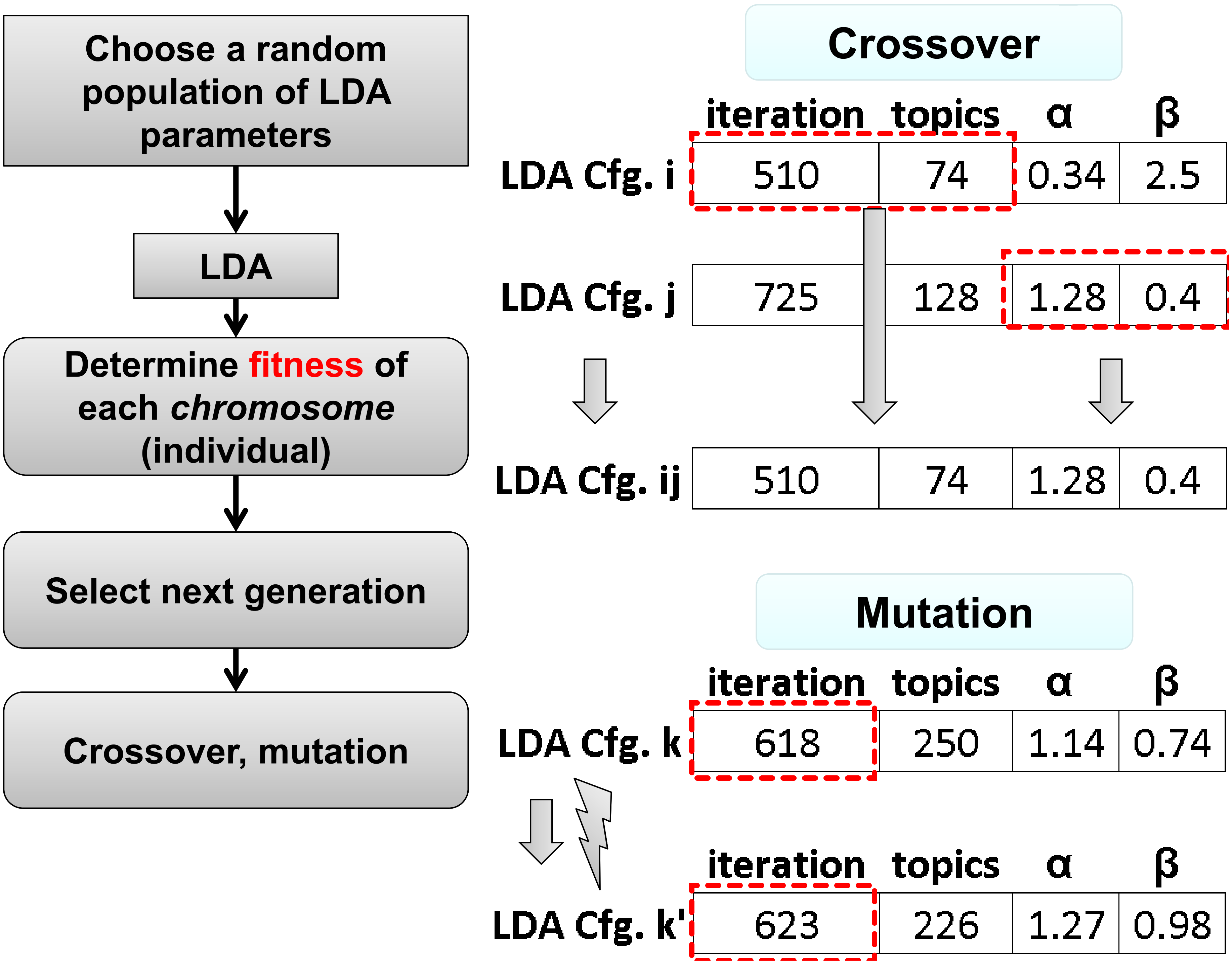
Crossover

	iteration	topics	α	β
LDA Cfg. i	510	74	0.34	2.5
LDA Cfg. j	725	128	1.28	0.4

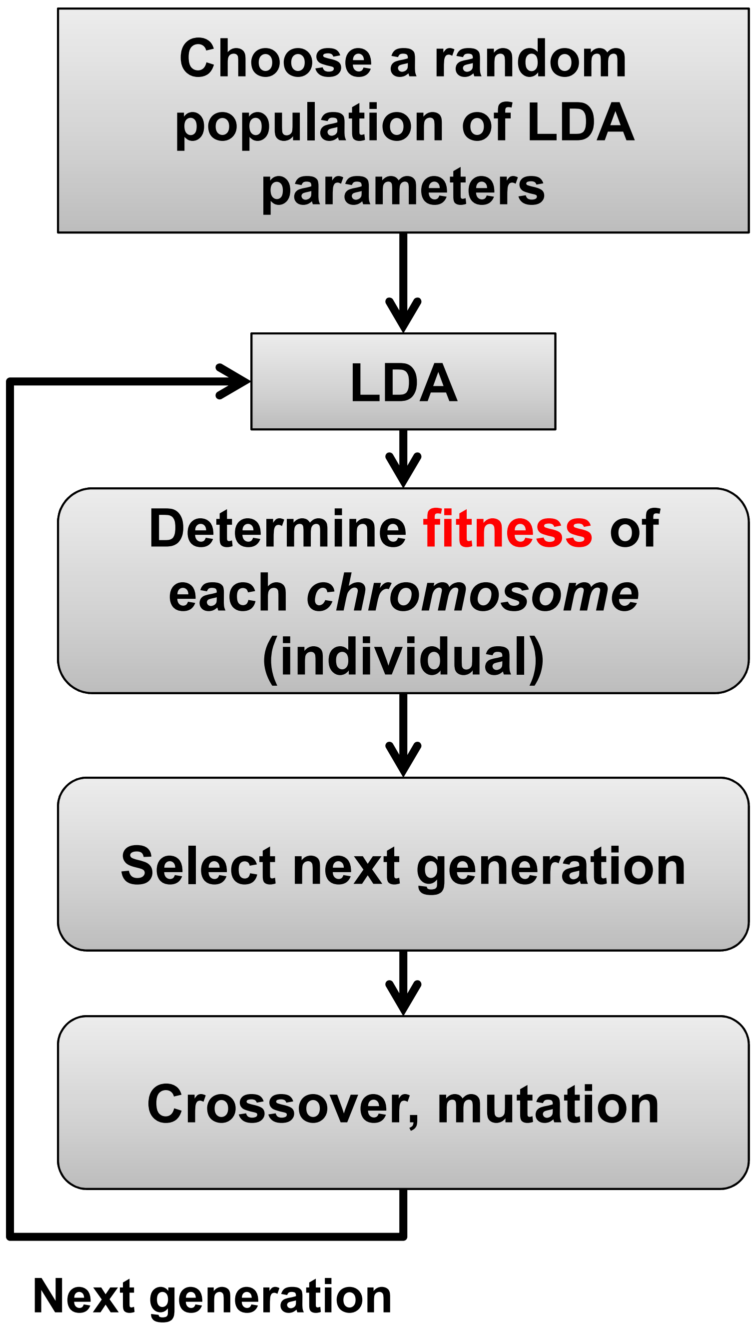
LDA Cfg. ij


Choose a random population of LDA parameters
LDA
Determine fitness of each chromosome (individual)
Select next generation
Crossover, mutation
Crossover
iteration
topics
α
β
LDA Cfg. i
510
74
0.34
2.5
LDA Cfg. j
725
128
1.28
0.4
LDA Cfg. ij
510
74
1.28
0.4

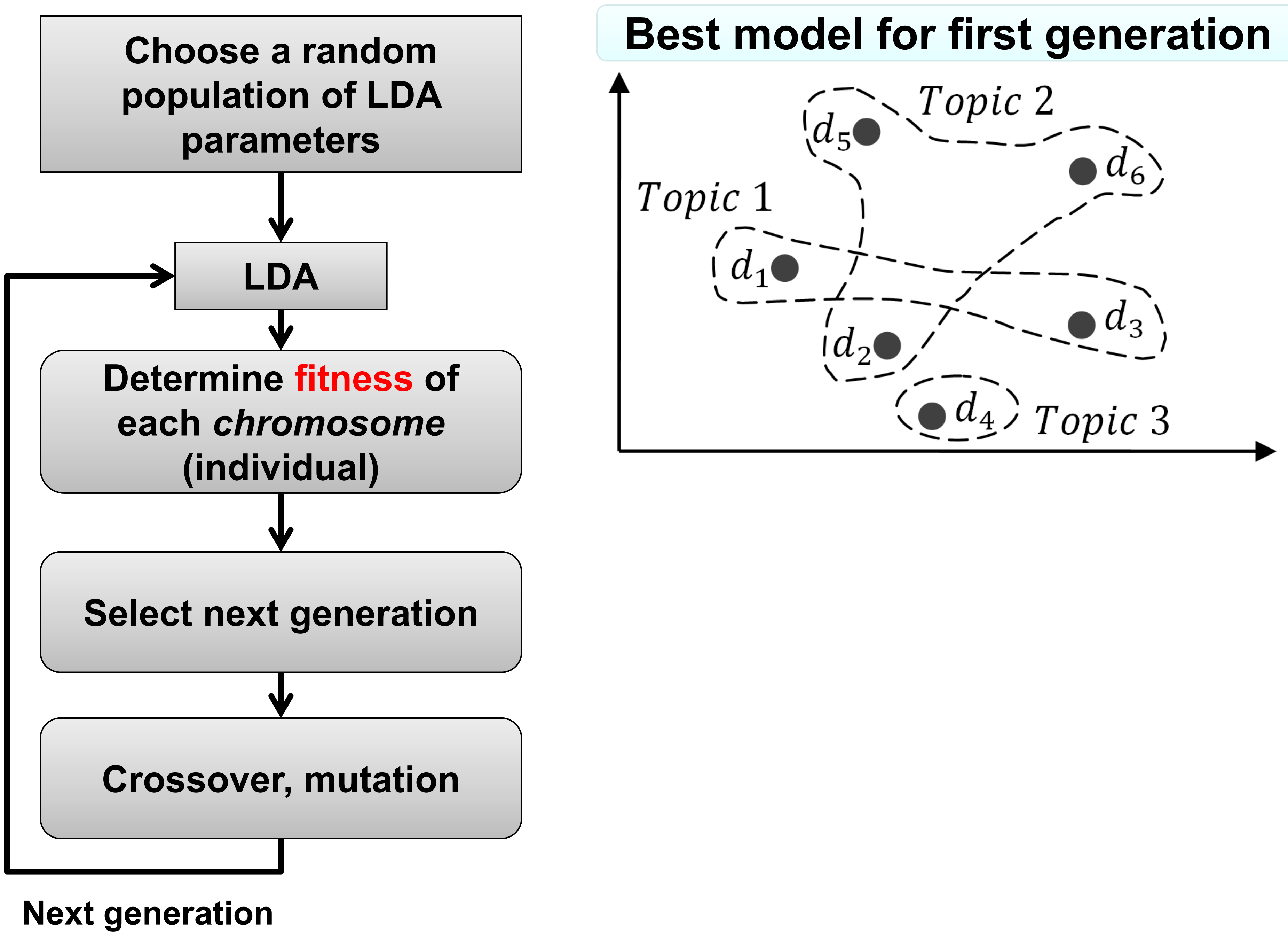

Choose a random population of LDA parameters
LDA
Determine fitness of each chromosome (individual)
Select next generation
Crossover, mutation
Crossover
iteration topics α β
LDA Cfg. i 510 74 0.34 2.5
LDA Cfg. j 725 128 1.28 0.4
LDA Cfg. ij 510 74 1.28 0.4
Mutation
iteration topics α β
LDA Cfg. k 618 250 1.14 0.74
iteration topics α β
LDA Cfg. k'

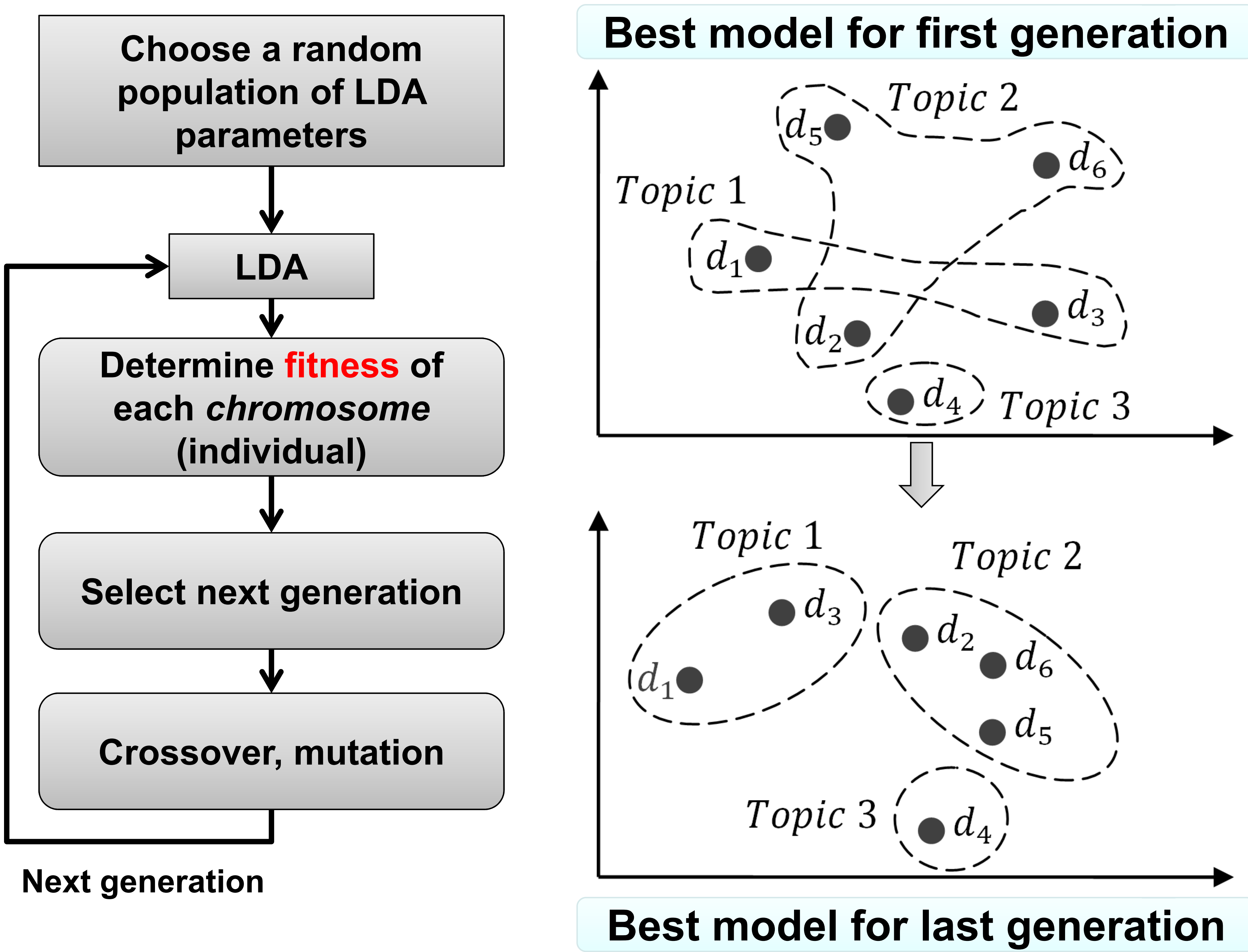


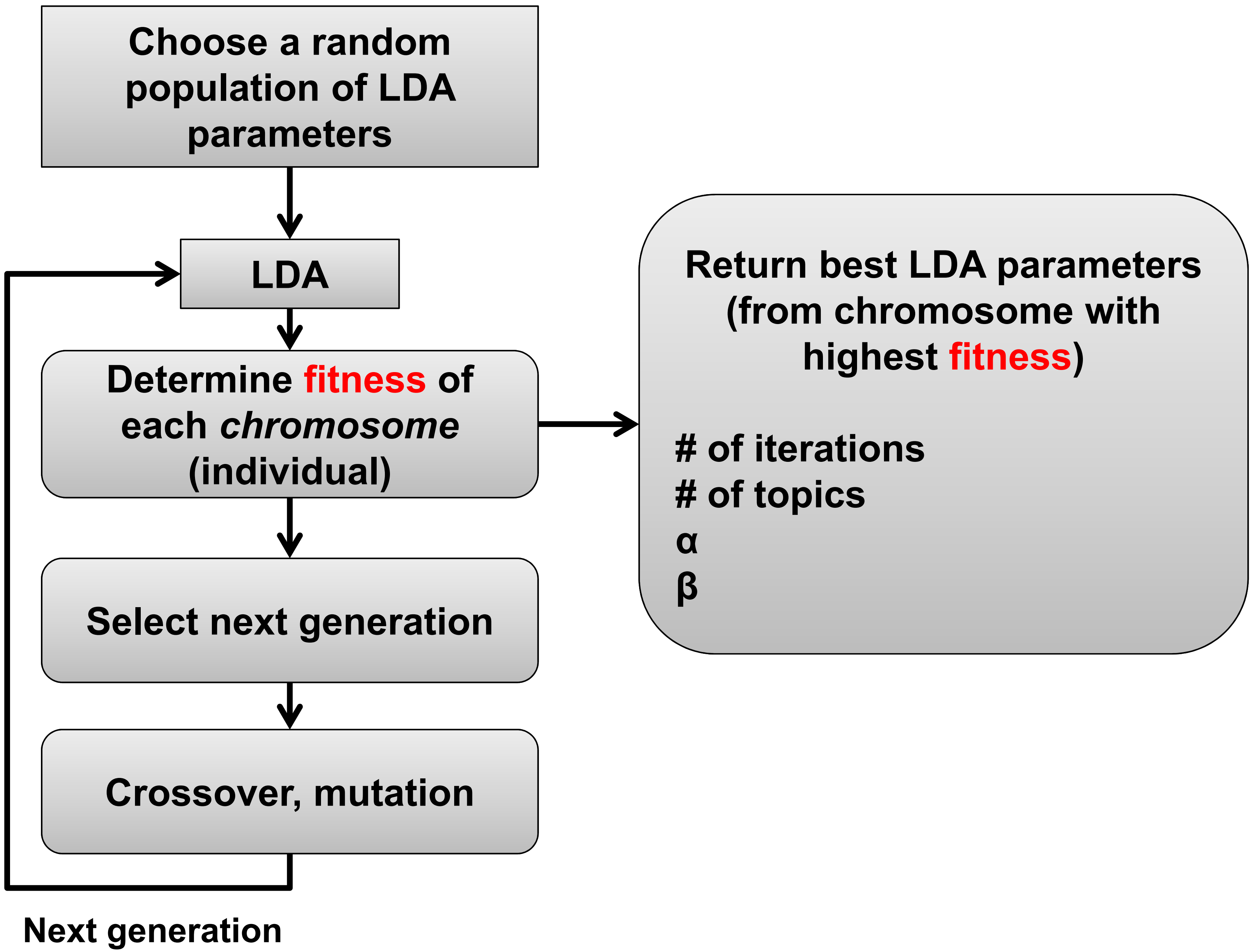
Choose a random population of LDA parameters
LDA
Determine fitness of each chromosome (individual)
Select next generation
Crossover, mutation
Crossover
iteration
topics
α
β
LDA Cfg. i
510
74
0.34
2.5
LDA Cfg. j
725
128
1.28
0.4
LDA Cfg. ij
510
74
1.28
0.4
Mutation
iteration
topics
α
β
LDA Cfg. k
618
250
1.14
0.74
iteration
topics
α
β
LDA Cfg. k'
623
226
1.27
0.98



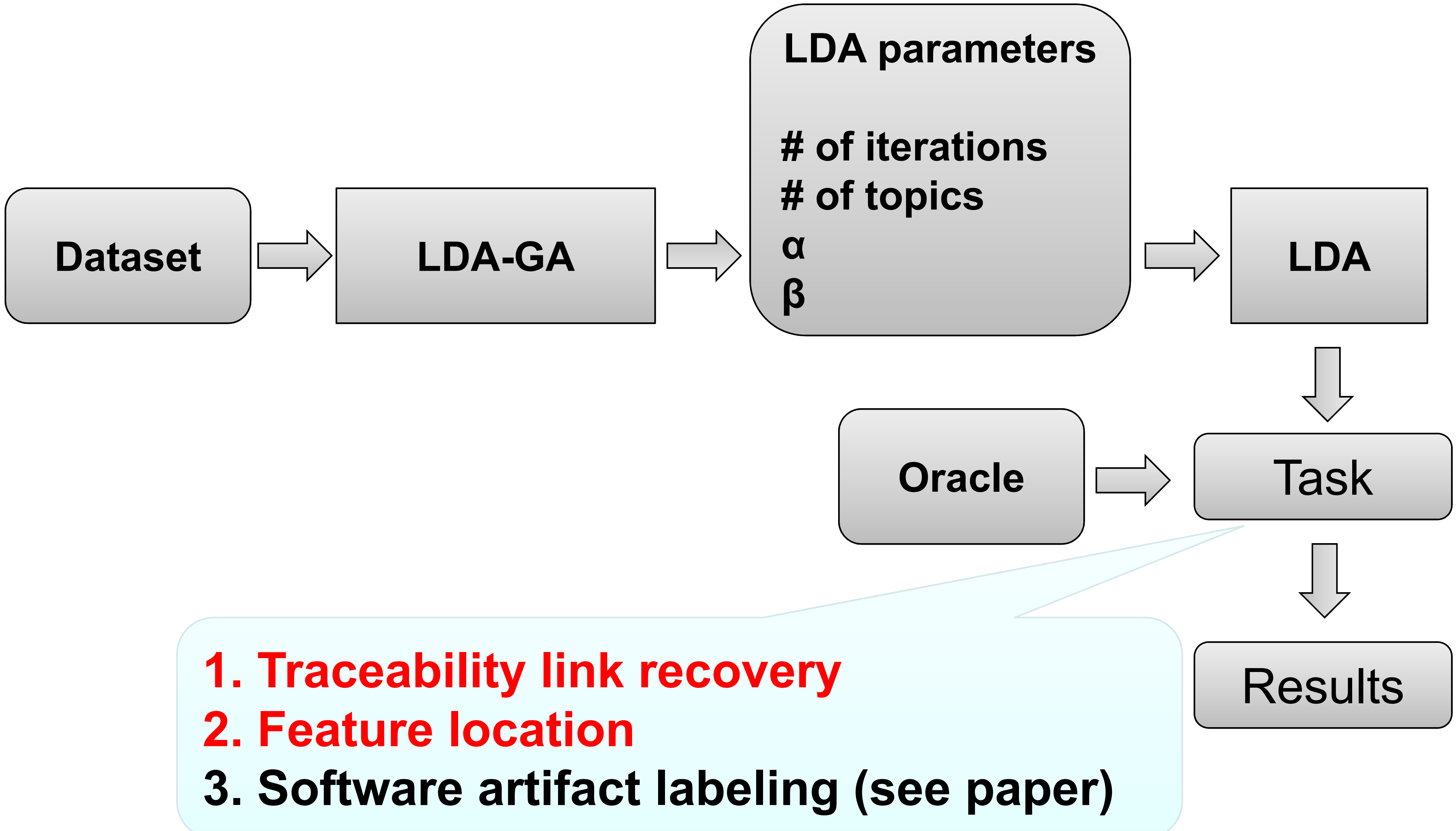
Choose a random population of LDA parameters
LDA
Determine fitness of each *chromosome* (individual)
Select next generation
Crossover, mutation
Next generation


Choose a random population of LDA parameters
LDA
Determine fitness of each *chromosome* (individual)
Select next generation
Crossover, mutation
Next generation
Best model for first generation
$Topic\ 1$
$Topic\ 2$
$Topic\ 3$
d_1
d_2
d_3
d_4
d_5
d_6


Choose a random population of LDA parameters
LDA
Determine fitness of each chromosome (individual)
Select next generation
Crossover, mutation
Next generation
Best model for first generation
Topic 1
Topic 2
Topic 3
d_1
d_2
d_3
d_4
d_5
d_6
Best model for last generation


Choose a random population of LDA parameters
LDA
Determine fitness of each chromosome (individual)
Select next generation
Crossover, mutation
Next generation
Return best LDA parameters (from chromosome with highest fitness)
# of iterations
# of topics
α
β

Evaluation...


Dataset
LDA-GA
LDA parameters
# of iterations
# of topics
α
β
LDA
Oracle
Task
Results
1. Traceability link recovery
2. Feature location
3. Software artifact labeling (see paper)

Evaluation: Traceability Link Recovery

- Recover links between ***use cases*** and ***code classes***

System	Size	# use cases	# code classes	# correct links
EasyClinic	20KLOC	30	47	93
eTour	45KLOC	58	174	366

Combinatorial:

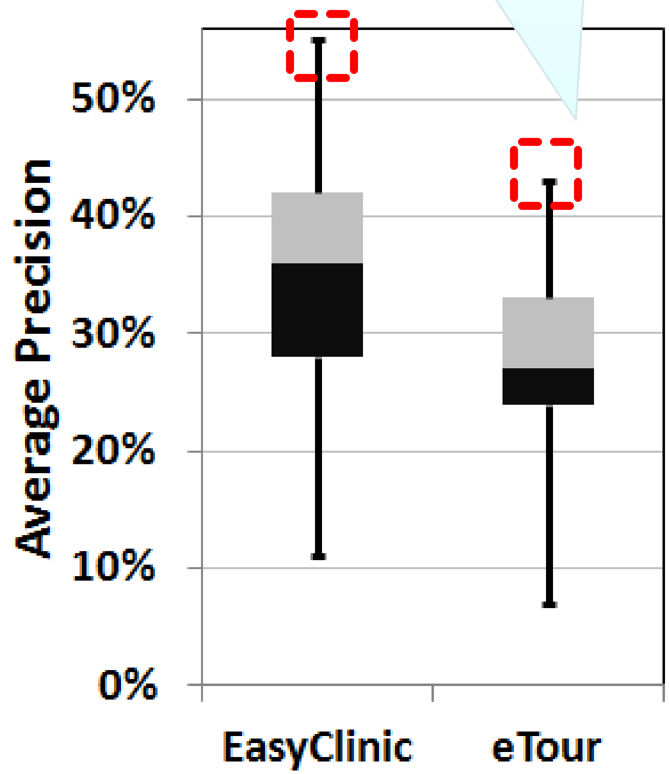
```
for numIter in [500, …]
    for numTopics in [5, …]
        for α in [0.01, …]
            for β in [0.01, …]
                LDA[numIter , numTopics , α, β]
```

Choose LDA parameters with best average precision using an *oracle*

Combinatorial:

for numIter in [500, …]
  for numTopics in [5, …]
    for α in [0.01, …]
      for β in [0.01, …]
        LDA[numIter , numTopics , α, β]

Choose LDA parameters with best average precision using an *oracle*



Combinatorial:

for numIter in [500, …]
  for numTopics in [5, …]
    for α in [0.01, …]
      for β in [0.01, …]
        LDA[numIter , numTopics , α, β]

Choose LDA parameters with best average precision using an *oracle*

LDA-GA:

run LDA-GA 30 times (to account for randomness)

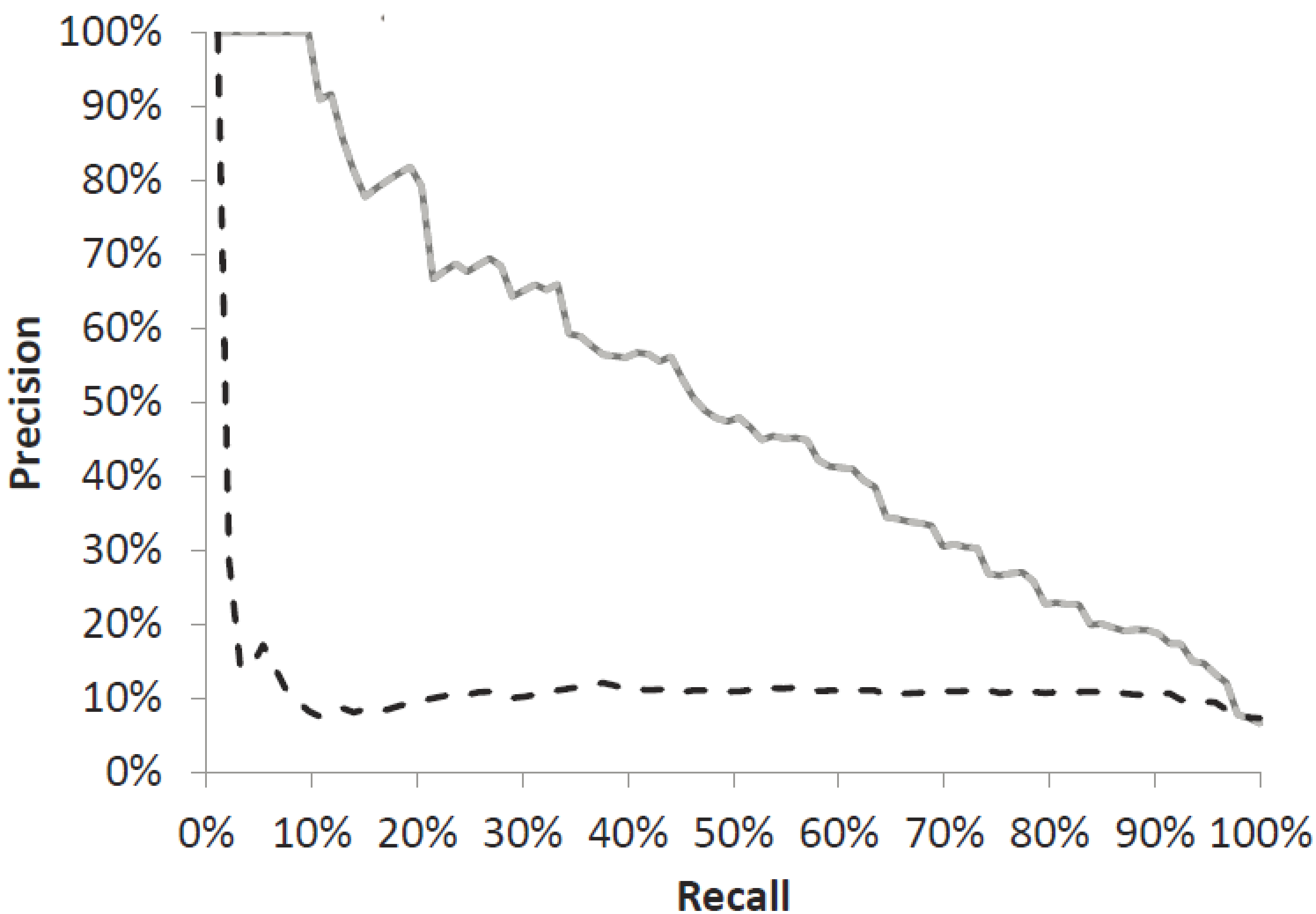
Choose LDA parameters corresponding to median fitness over 30 runs

Baseline:

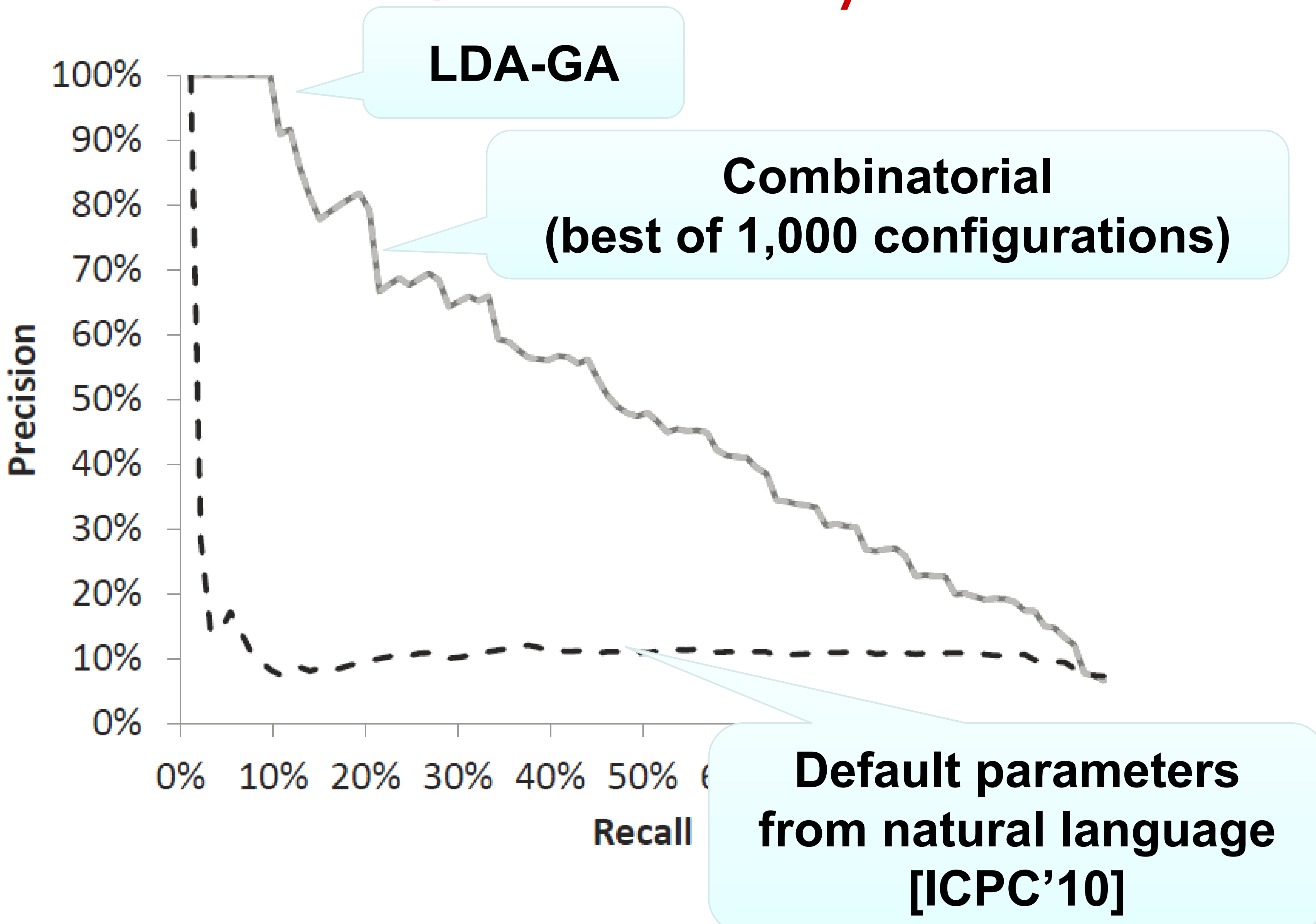
[ICPC'10]

Use default LDA parameters from natural language

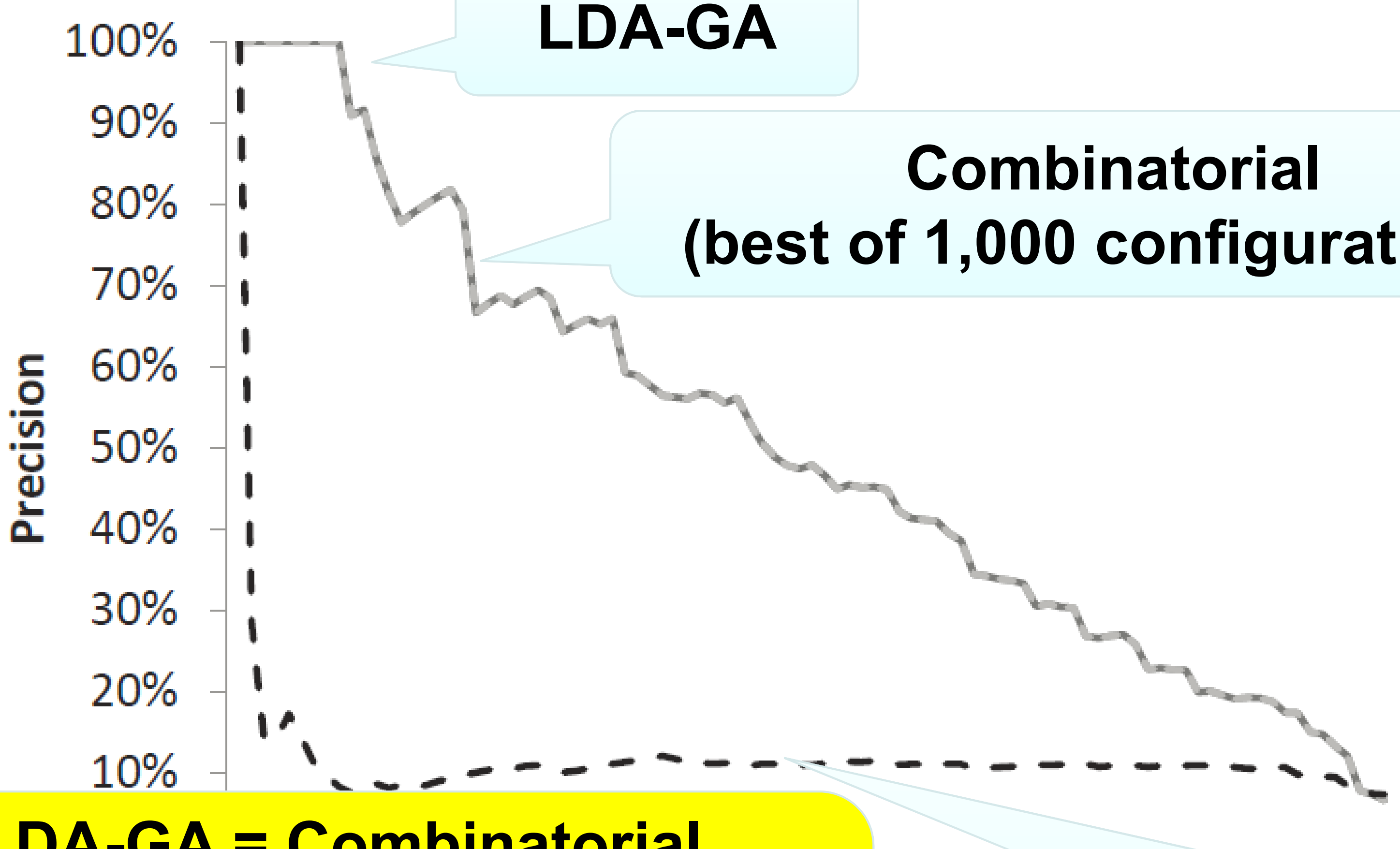
Precision/Recall EasyClinic



Precision/Recall EasyClinic



Precision/Recall EasyClinic



LDA-GA

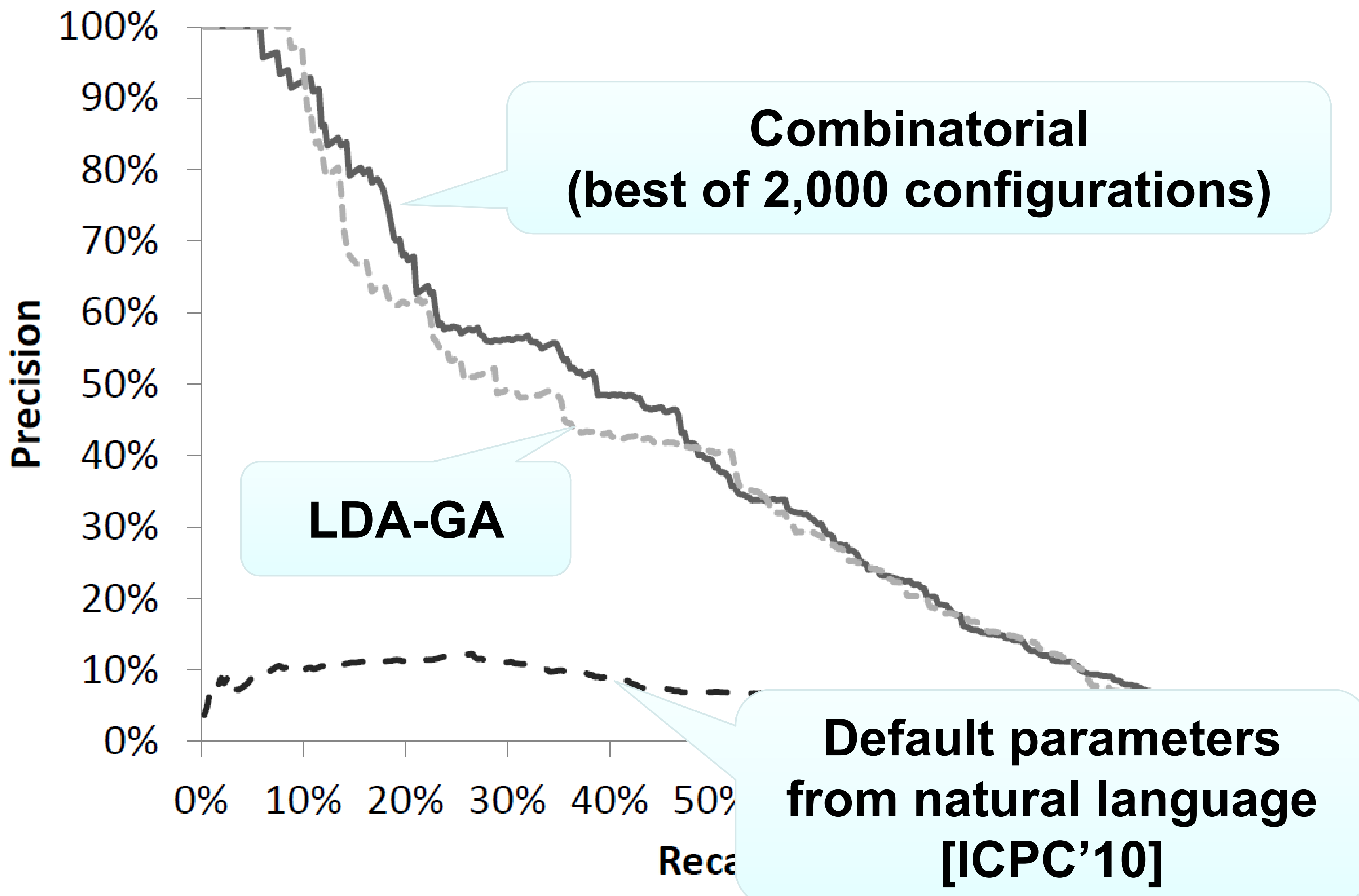
Combinatorial
(best of 1,000 configurations)

LDA-GA = Combinatorial

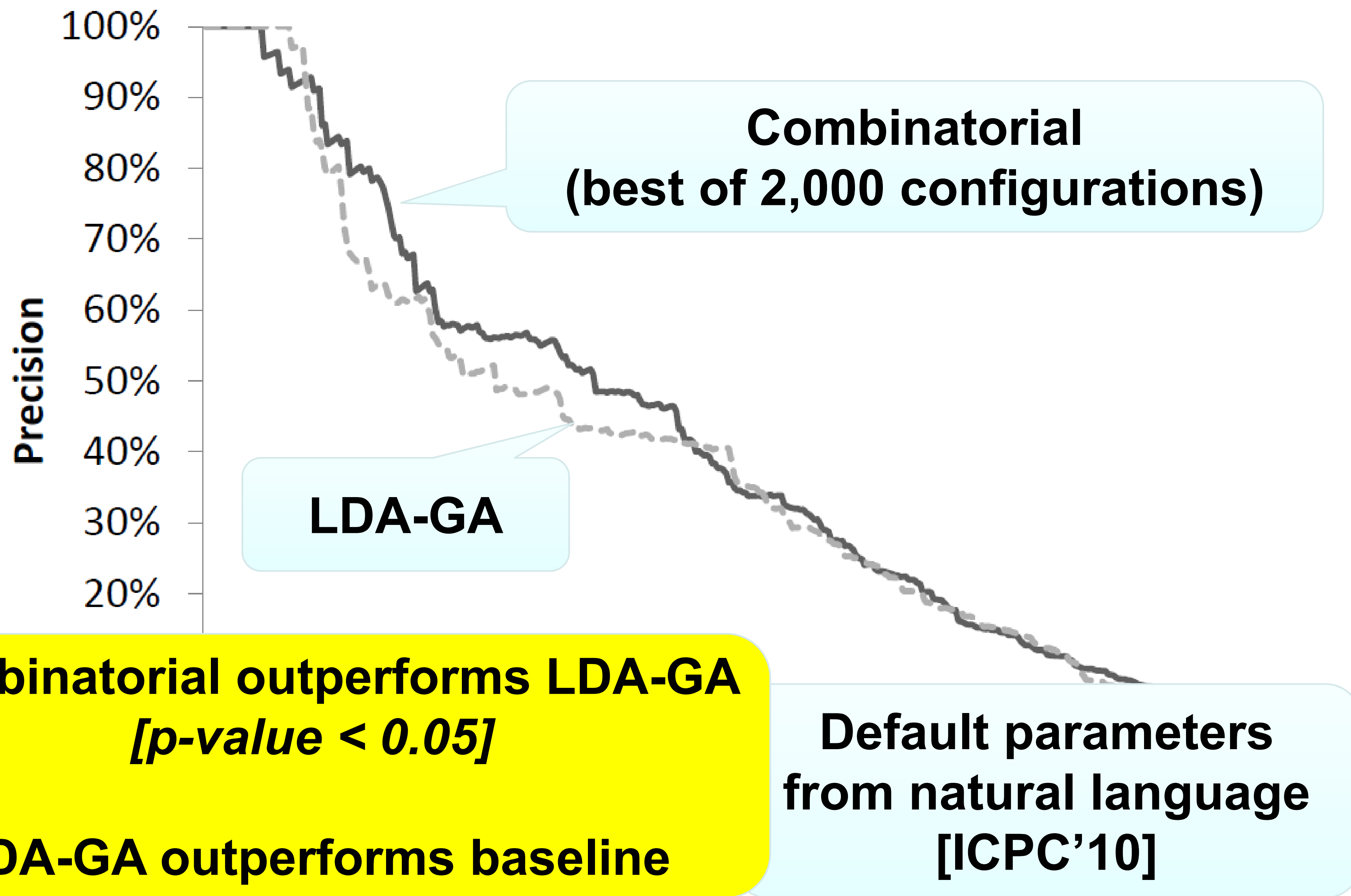
LDA-GA outperforms baseline
[p-value < 0.05]

Default parameters
from natural language
[ICPC'10]

Precision/Recall eTour



Precision/Recall eTour



Combinatorial outperforms LDA-GA
[p-value < 0.05]

LDA-GA outperforms baseline

Evaluation: Feature location

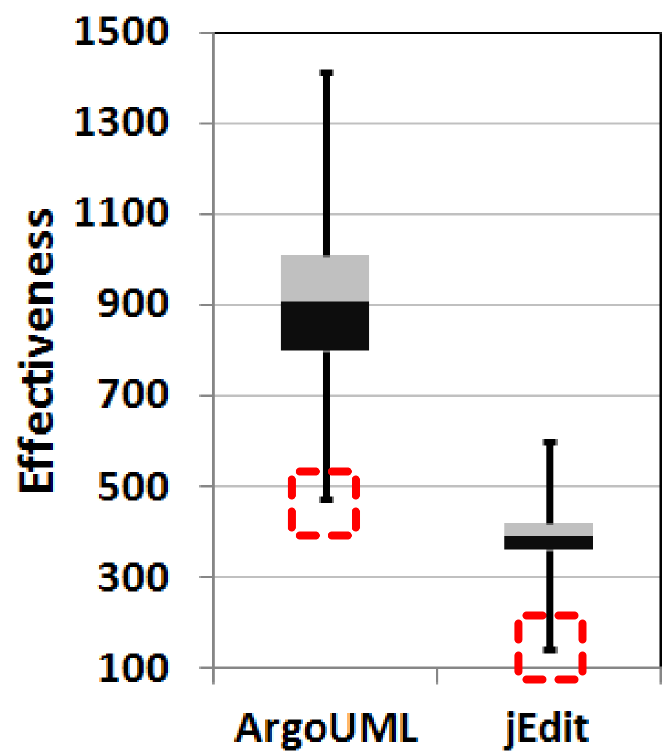
- Identify methods related to a maintenance task (e.g., bug, feature)

System	Size	# features	# methods
jEdit	104KLOC	150	6,413
ArgoUML	149KLOC	91	11,000

Combinatorial:

```
for numIter in [500, …]
    for numTopics in [5, …]
        for α in [0.01, …]
            for β in [0.01, …]
                LDA[numIter , numTopics , α, β]
```

Choose LDA parameters with best average effectiveness using an *oracle*



Combinatorial:

```
for numIter in [500, …]
    for numTopics in [5, …]
        for α in [0.01, …]
            for β in [0.01, …]
                LDA[numIter , numTopics , α, β]
```

Choose LDA parameters with best average effectiveness using an *oracle*

LDA-GA:

run LDA-GA 30 times (to account for randomness)

Choose LDA parameters corresponding to median fitness over 30 runs

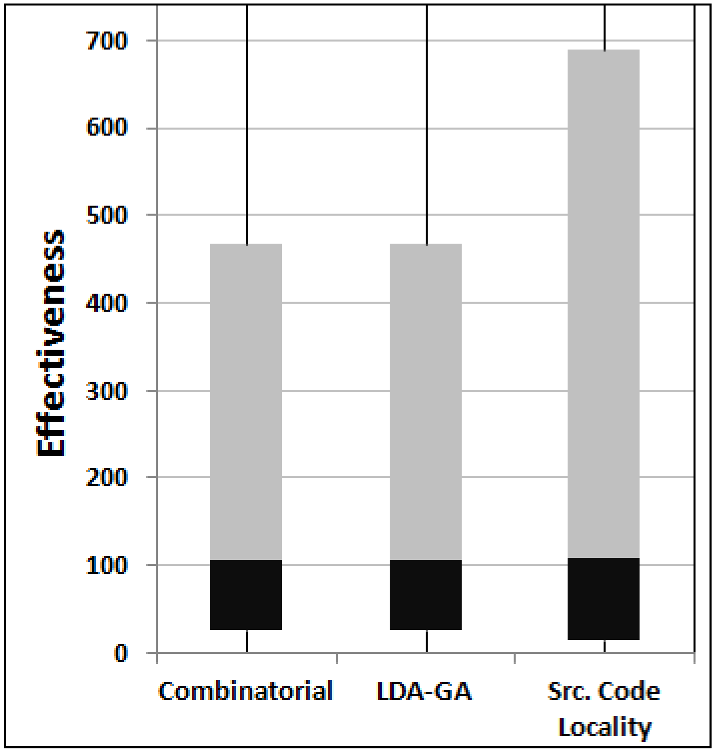
Baseline:

[SCAM'10]

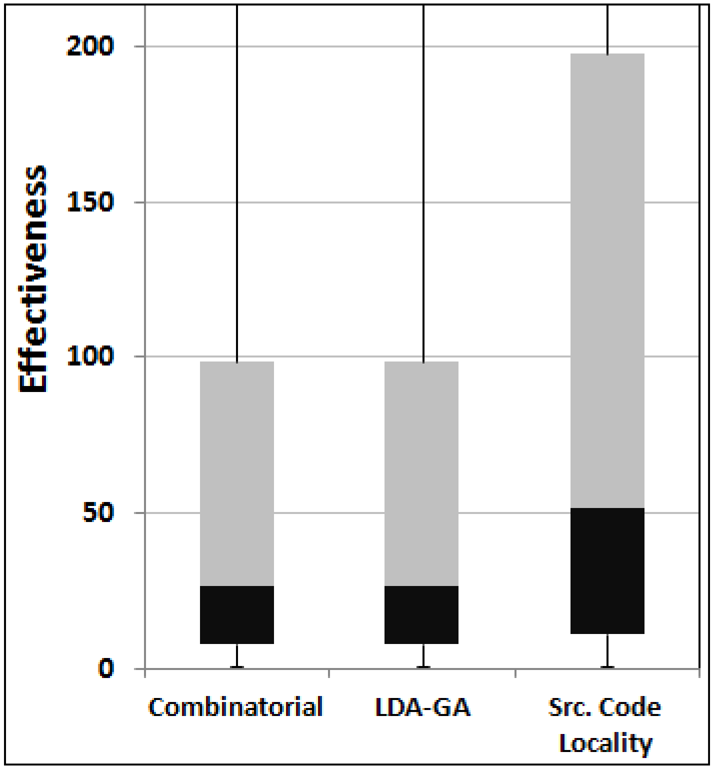
Use LDA parameters from *source locality heuristic*

Effectiveness measure

ArgoUML

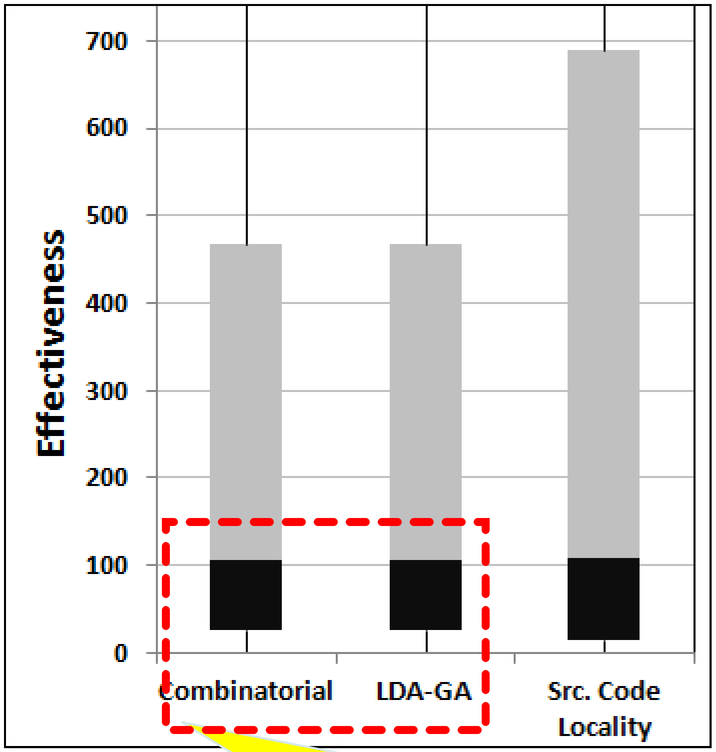


jEdit

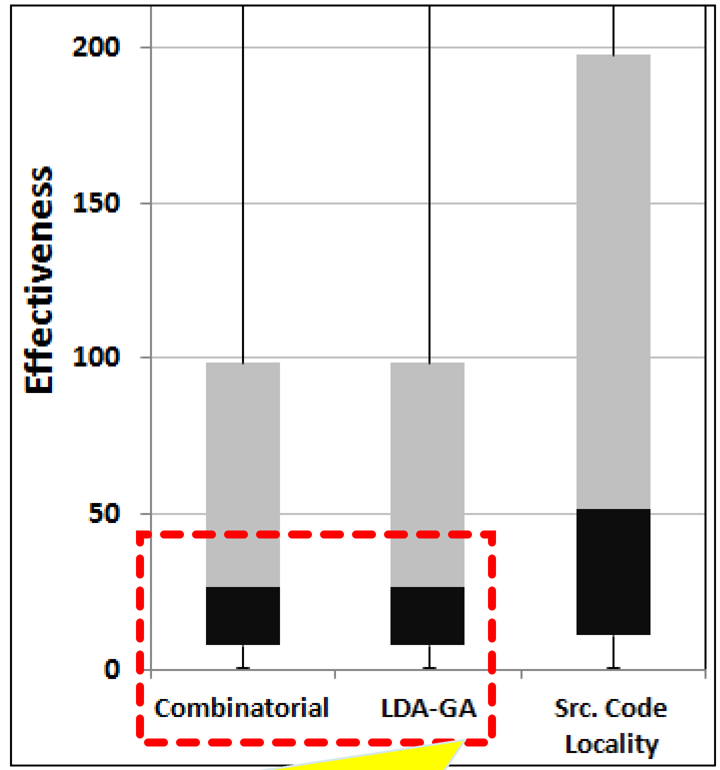


Effectiveness measure

ArgoUML



jEdit

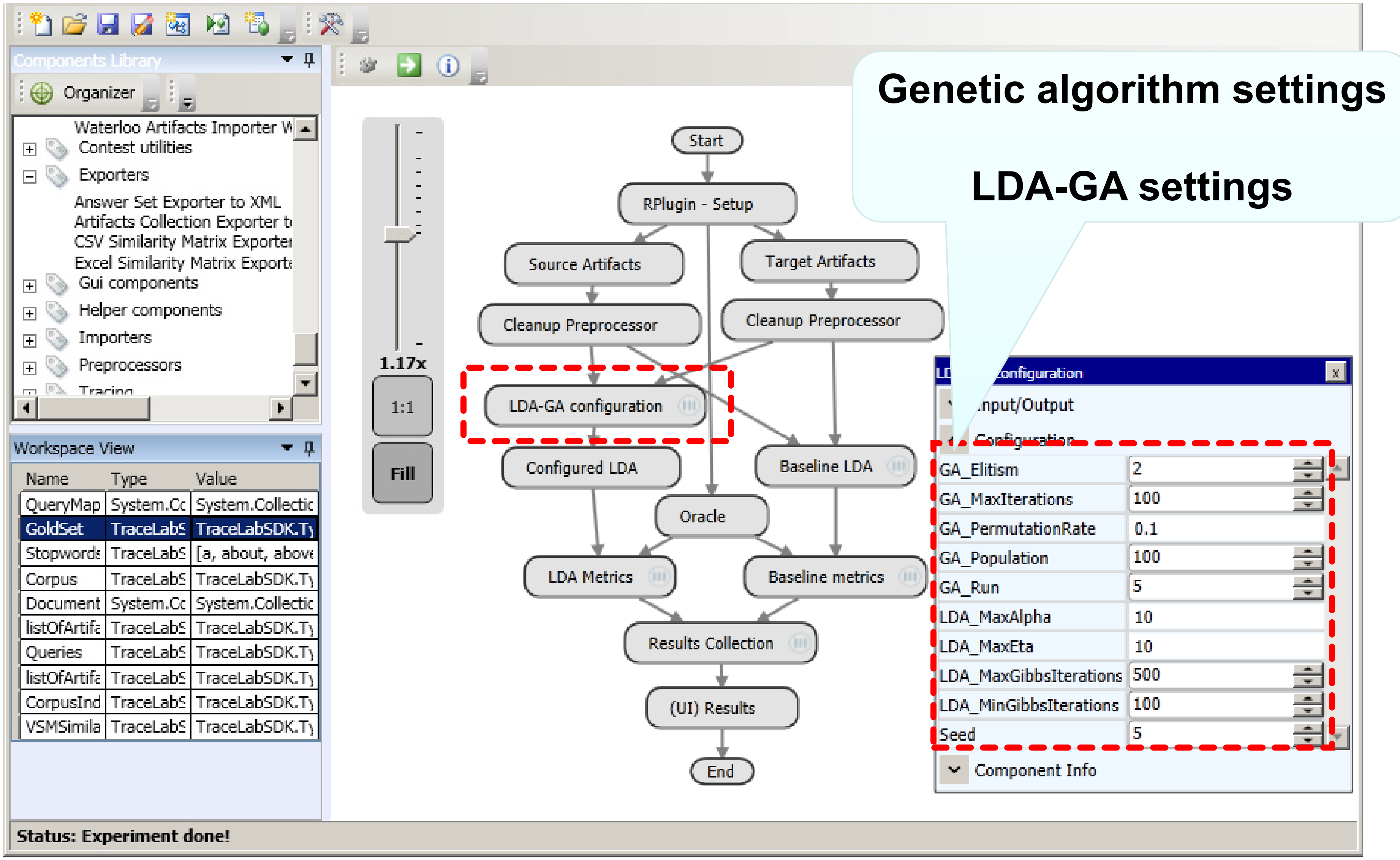


LDA-GA = Combinatorial
LDA-GA outperforms baseline *[p-value < 0.05]*

Conclusions

- Showed the impact of setting the LDA parameters on the results
- We proposed LDA-GA, a genetic based approach to automatically configure and find the near-optimal solution for LDA parameters
  - Dataset dependent
  - Oracle & task independent
- The approach was evaluated on three maintenance tasks

LDA-GA in TraceLab



Thank you! Questions?

http://www.distat.unimol.it/reports/LDA-GA/

http://www.cs.wm.edu/semeru/data/tefse13/

References

- [Hindle et al., ICSE'12] A. Hindle, E. T. Barr, Z. Su, M. Gabel, and P. T. Devanbu, "On the naturalness of software," in Proc. of the 34th IEEE/ACM International Conference on Software Engineering (ICSE'12), Zurich, Switzerland, June 2-9, 2012, pp. 837–847.
- [ICPC'10] R. Oliveto, M. Gethers, D. Poshyvanyk, and A. De Lucia, "On the equivalence of information retrieval methods for automated traceability link recovery," in Proc of the 18th IEEE International Conference on Program Comprehension (ICPC'10), Braga, Portugal, 2010, pp. 68–71.
- [SCAM'10] S. Grant and J. R. Cordy, "Estimating the optimal number of latent concepts in source code analysis," in Proc. of the 10th International Working Conference on Source Code Analysis and Manipulation (SCAM'10), 2010, pp. 65–74.

Threats to Validity

- We used datasets that have been used in other studies
- We ran GA 30 times to account for randomness
- Non-parametric statistical test
- Generalizability of results to other SE tasks

GA Settings

- Implementation: GA library in R
- Population size: 100
- Elitism of 2 individuals
- Roulette wheel selection operator
- Crossover probability: 0.6
- Mutation probability: 0.01
- Stop criteria:
  - No improvement in 10 generations
  - When reaching 100 generations

Software Artifact Labeling

TABLE IV
AVERAGE OVERLAP BETWEEN AUTOMATIC AND MANUAL LABELING.

exVantage					
	LDA		De Lucia et al. [13]		
	LDA-GA	Combinatorial	n = M	n = M/2	n = 2
Max	**100%**	**100%**	100%	100%	100%
3rd Quartile	**95%**	**95%**	71%	70%	69%
Median	67%	**70%**	59%	60%	54%
2nd Quartile	60%	**67%**	34%	50%	41%
Min	**50%**	**50%**	0%	0%	40%
Mean	74%	77%	52%	56%	60%
St. Deviation	19%	17%	31%	34%	23%

JHotDraw					
	LDA		De Lucia et al. [13]		
	LDA-GA	Combinatorial	n = M	n = M/2	n = 2
Max	**100%**	**100%**	100%	100%	100%
3 Quartile	81%	**82%**	73%	70%	66%
Median	71%	**75%**	65%	61%	56%
2 Quartile	47%	**50%**	46%	45%	41%
Min	14%	14%	0%	38%	**29%**
Mean	65%	66%	59%	60%	59%
St. Deviation	28%	26%	28%	20%	24%