Configuring Topic Models for Software Engineering Tasks in TraceLab

Bogdan Dit

Annibale Panichella

Evan Moritz

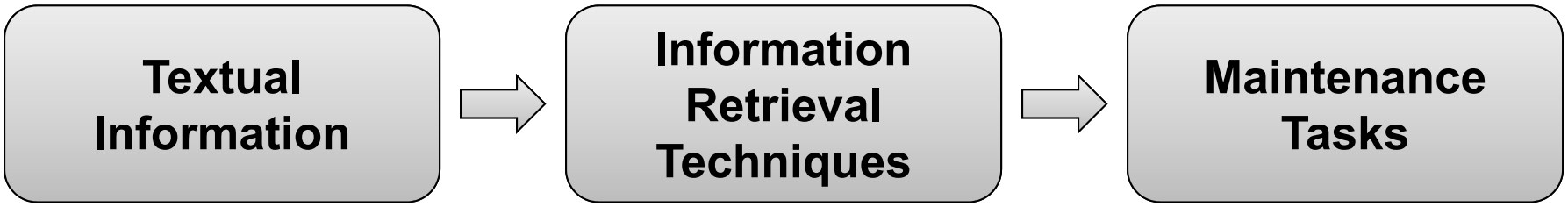
Rocco Oliveto

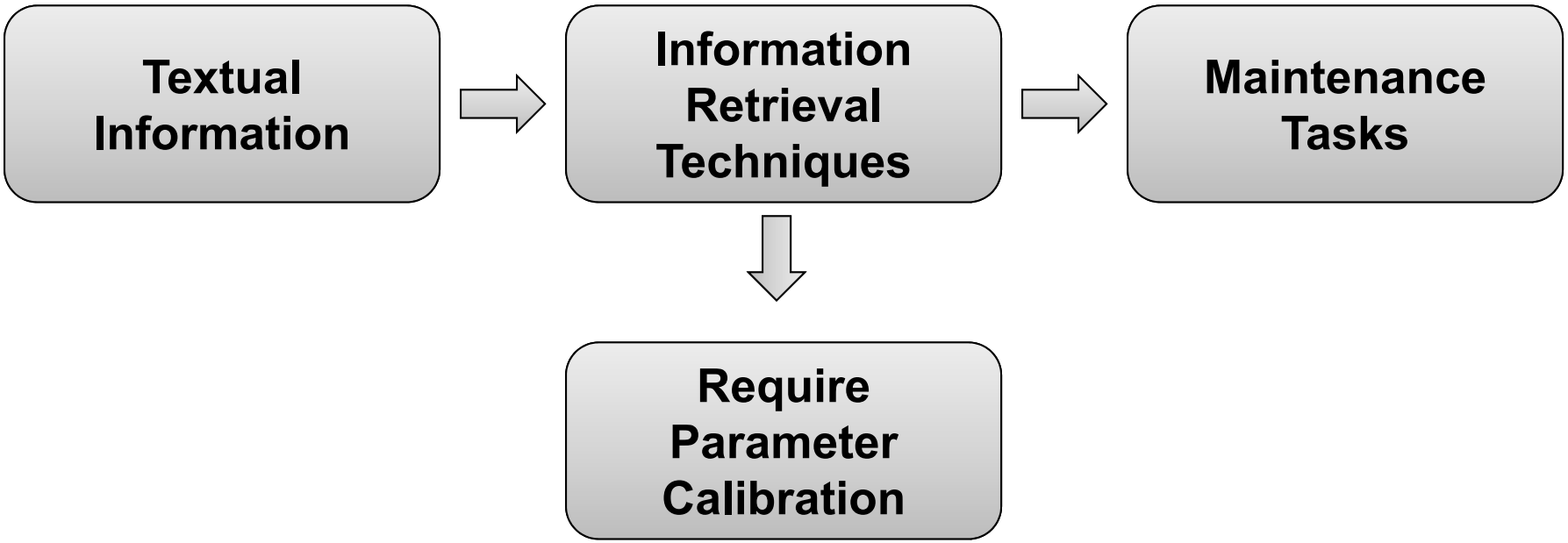
Massimiliano Di Penta

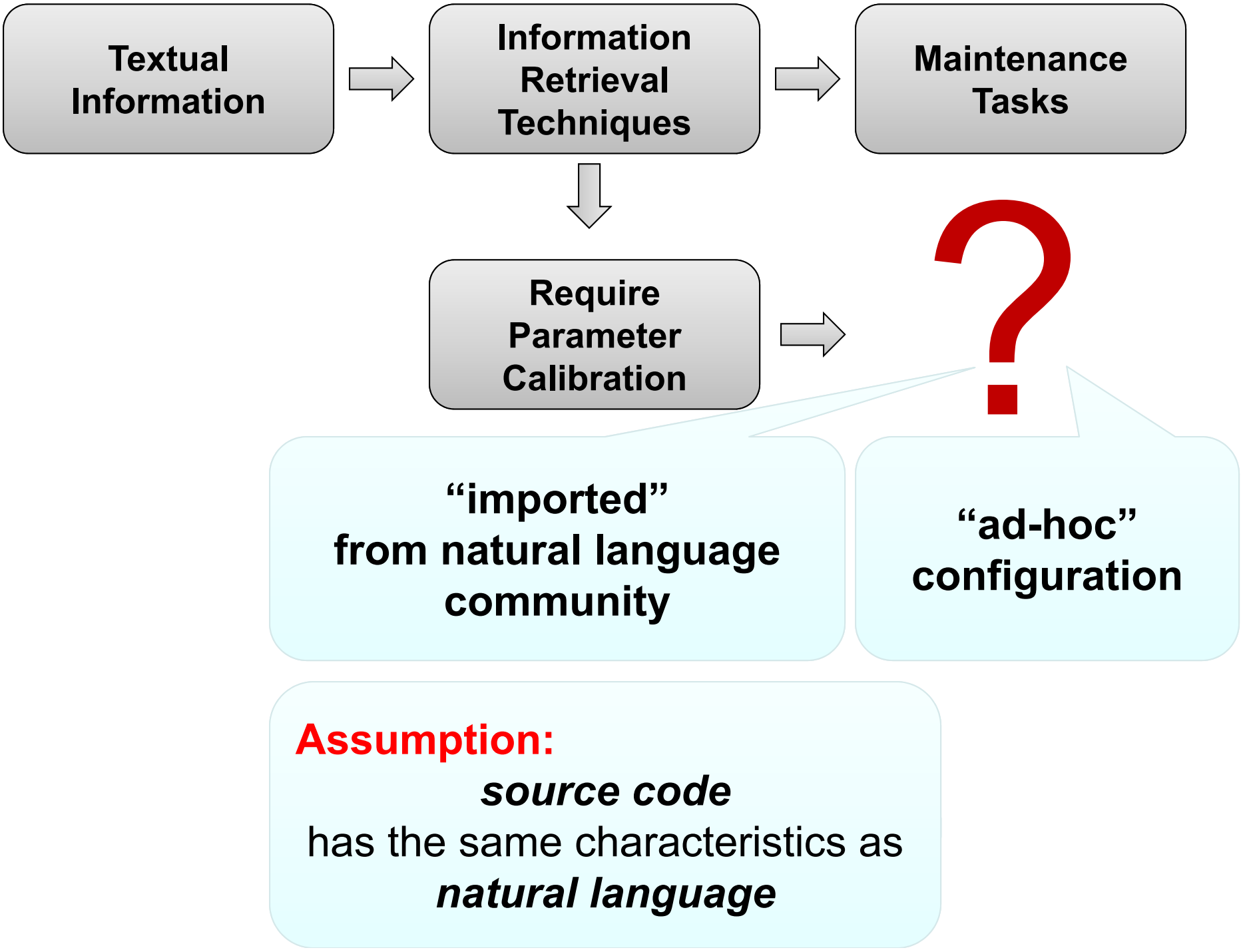
Denys Poshyvanyk

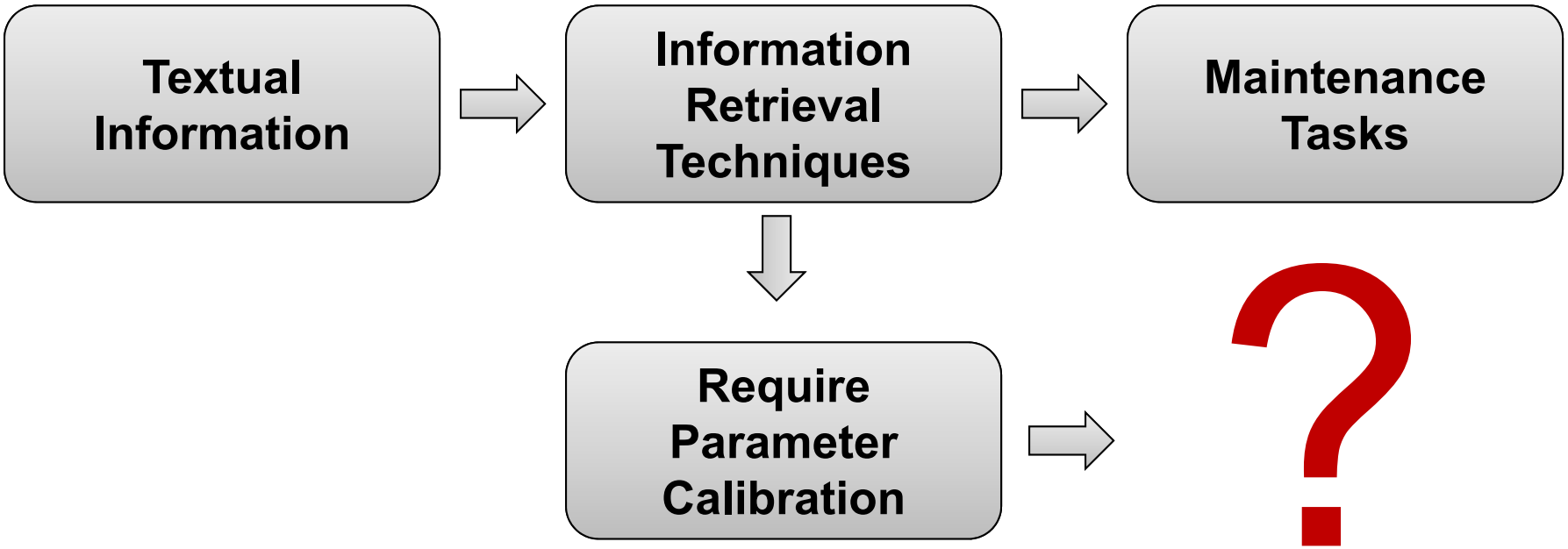
Andrea De Lucia

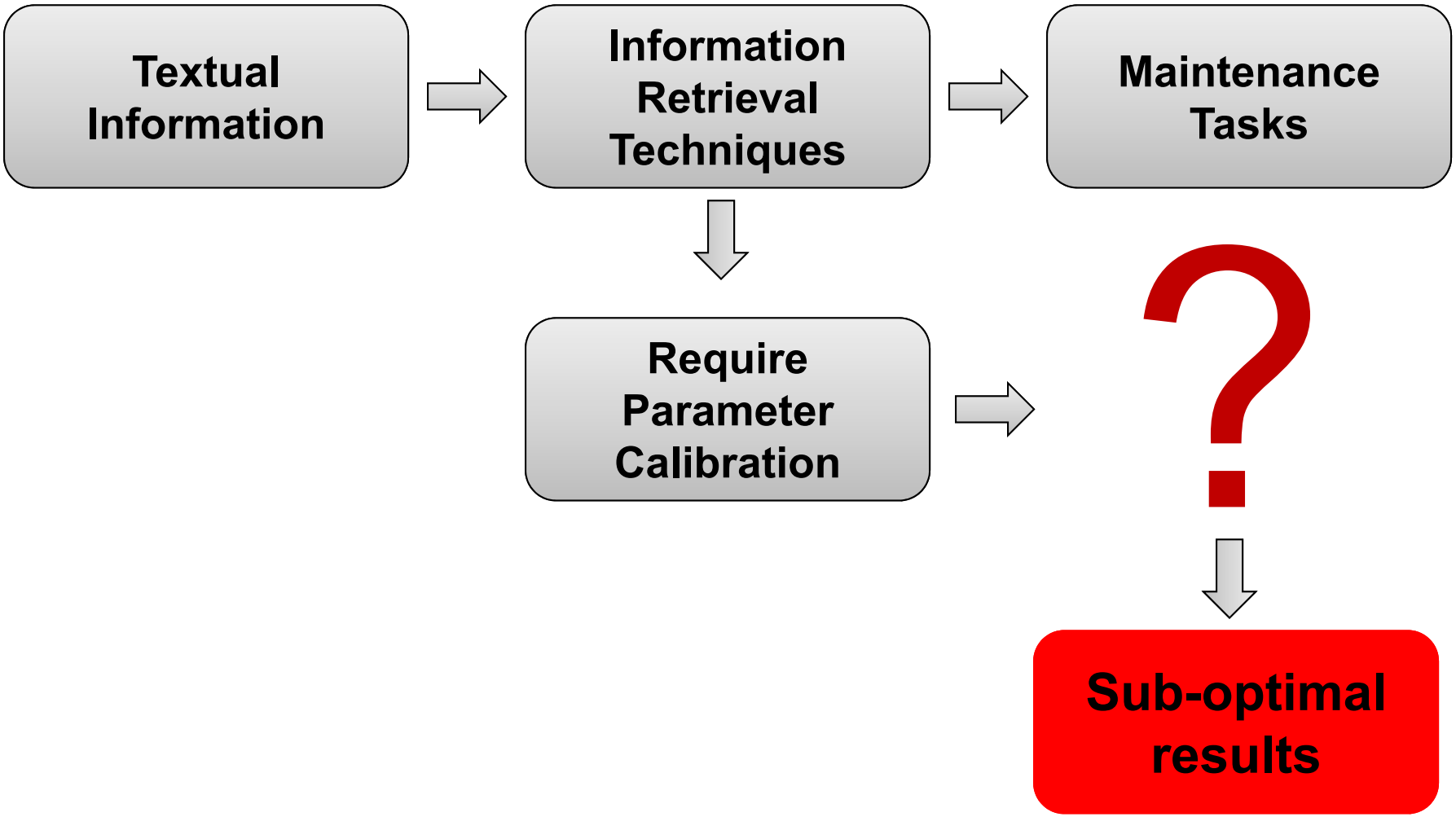
TEFSE'13
San Francisco, CA

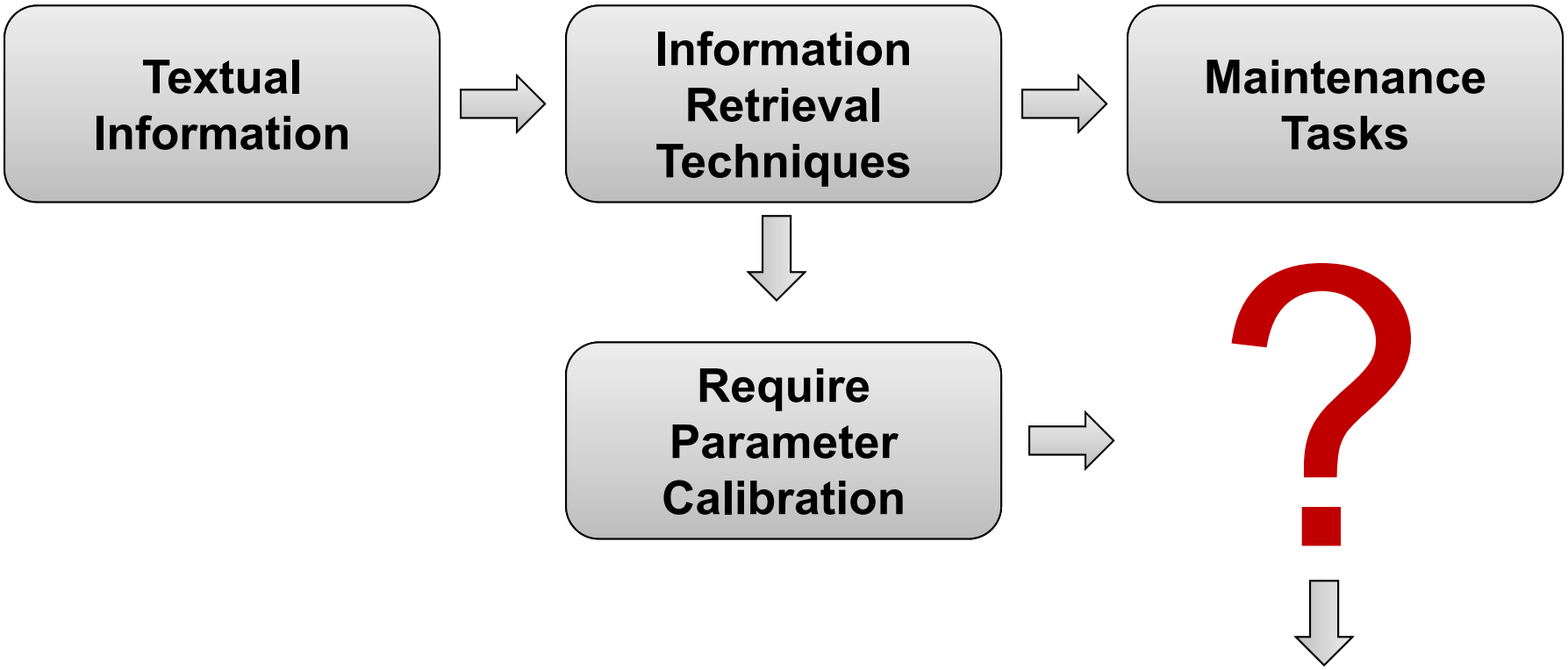

Textual Information
Information Retrieval Techniques
Maintenance Tasks


Textual Information
Information Retrieval Techniques
Maintenance Tasks
Require Parameter Calibration


Textual Information
Information Retrieval Techniques
Maintenance Tasks
Require Parameter Calibration
?
“imported” from natural language community
“ad-hoc” configuration
Assumption:
source code
has the same characteristics as
natural language


Textual Information
Information Retrieval Techniques
Maintenance Tasks
Require Parameter Calibration
?


Textual Information
Information Retrieval Techniques
Maintenance Tasks
Require Parameter Calibration
?
Sub-optimal results

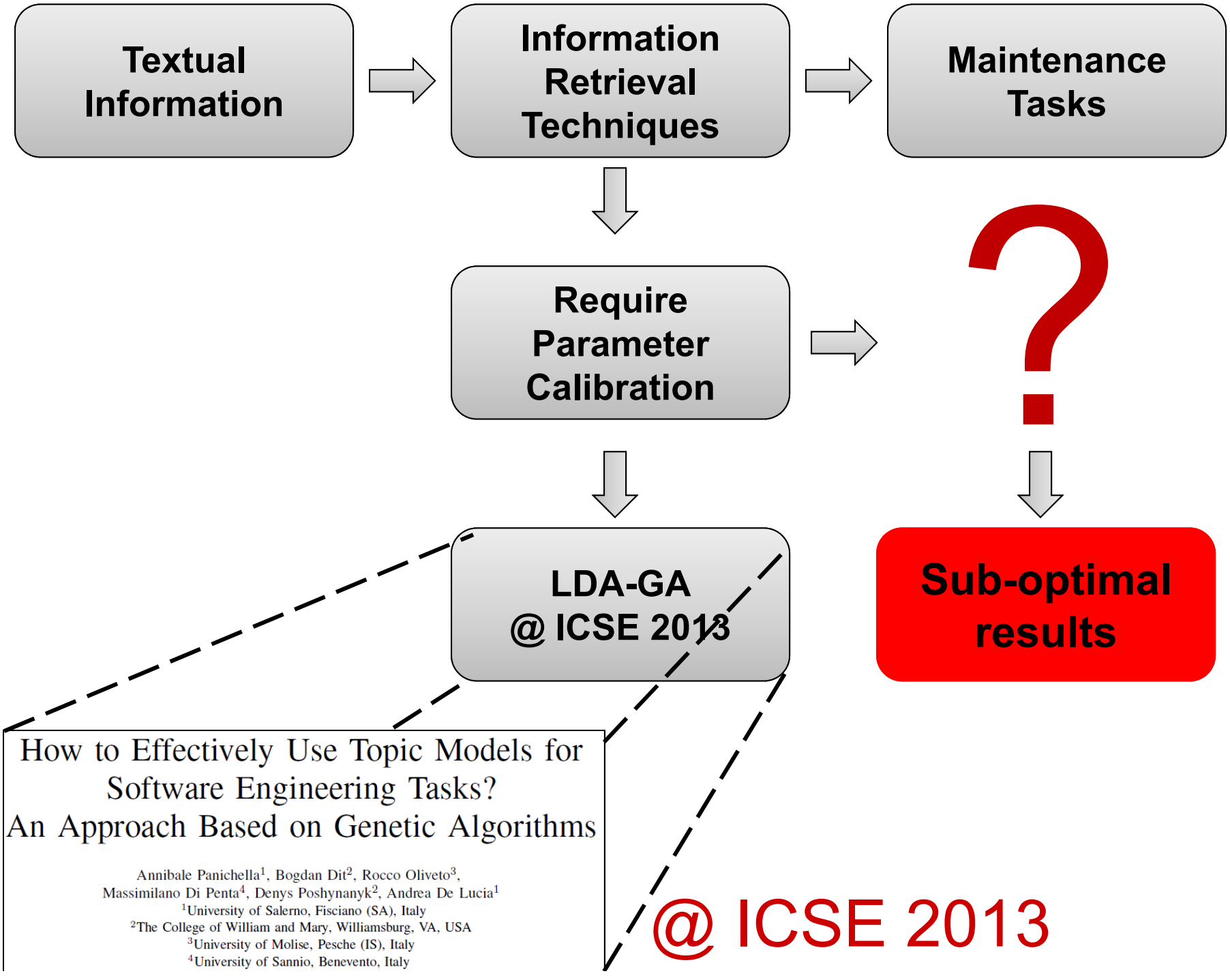


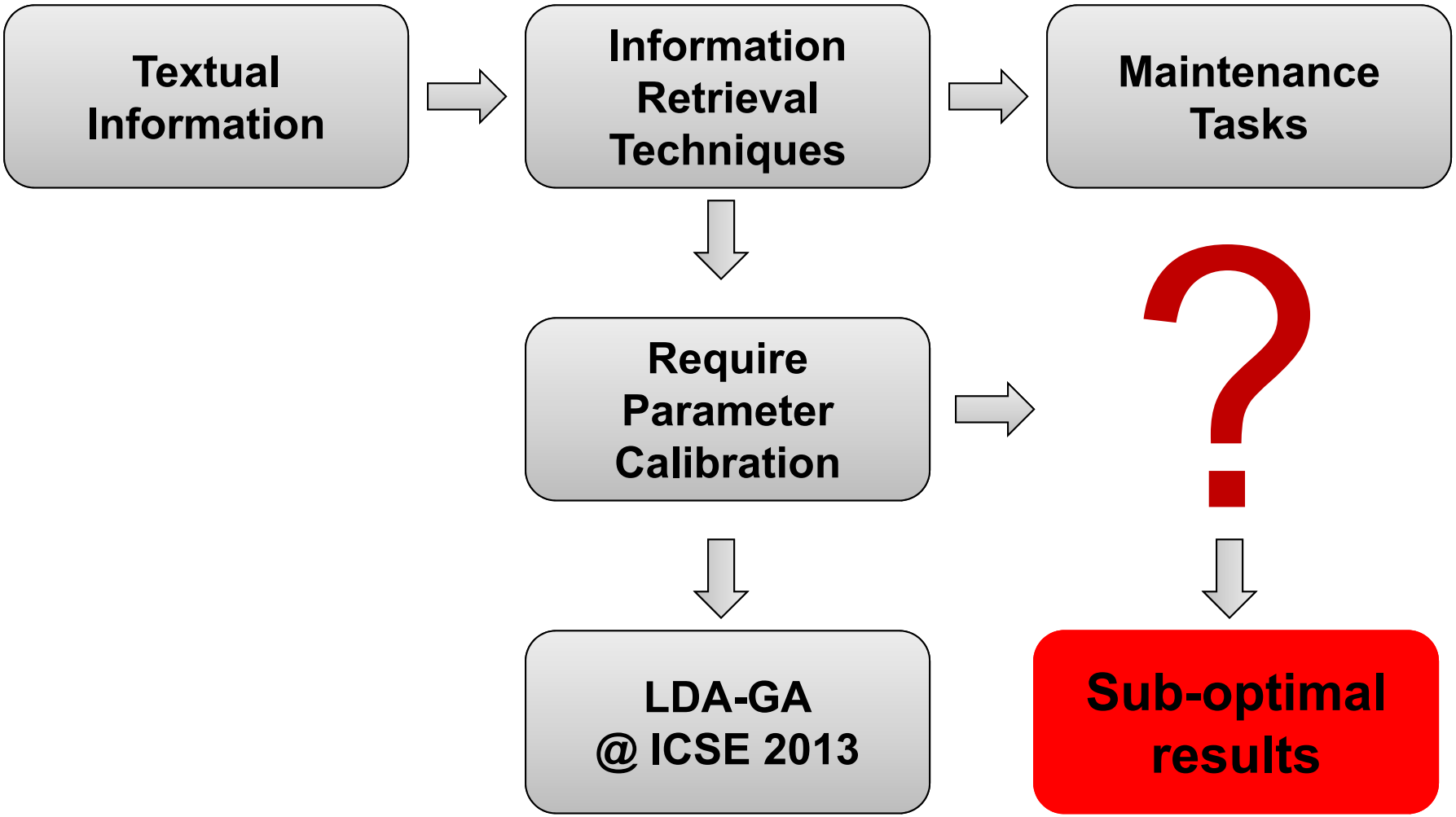
[Hindle et al. @ ICSE'12]:

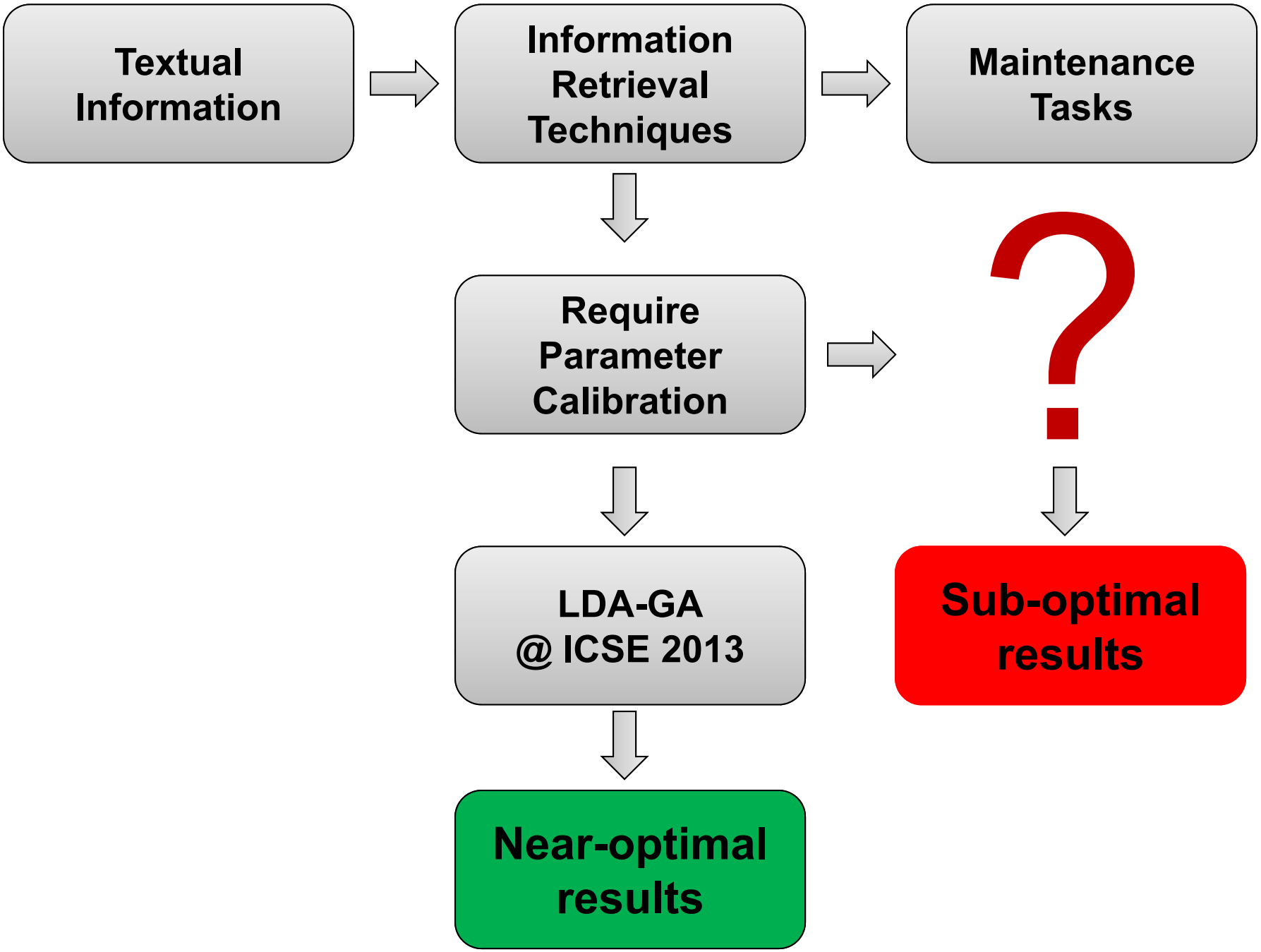
source code

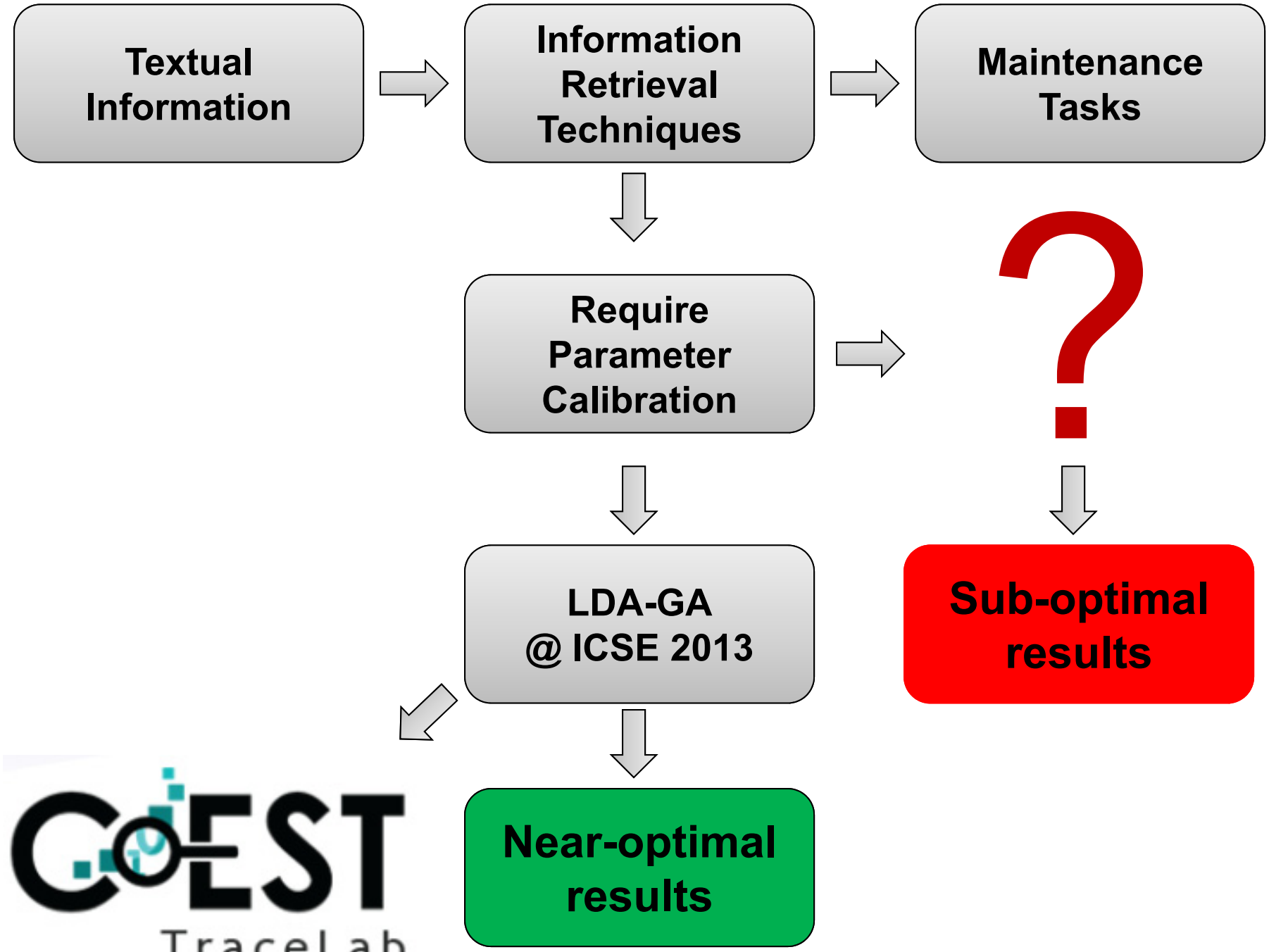
is more ***regular*** and ***predictable*** than

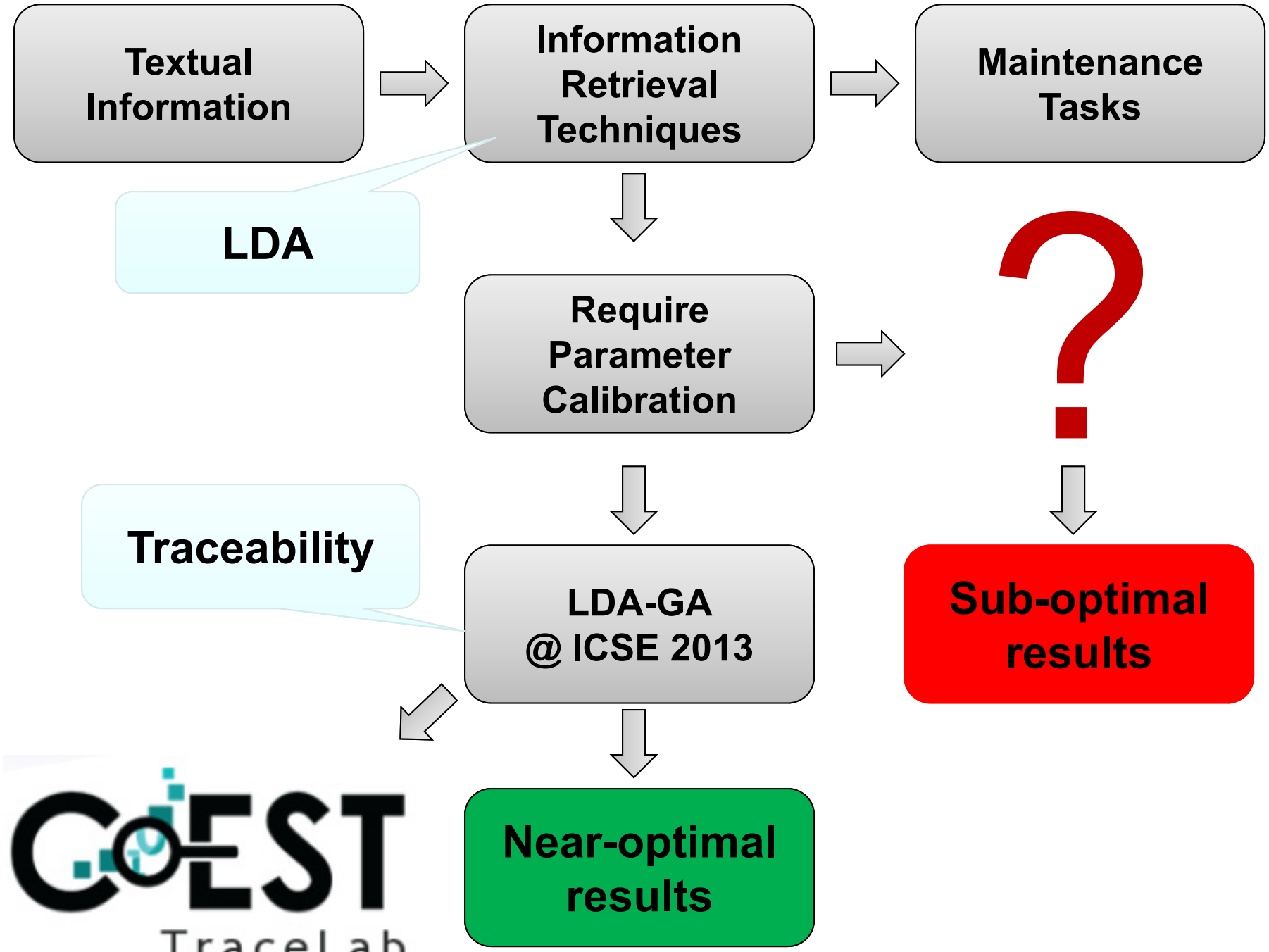
natural language


Textual Information
Information Retrieval Techniques
Maintenance Tasks
Require Parameter Calibration
?
LDA-GA @ ICSE 2013
Sub-optimal results
How to Effectively Use Topic Models for Software Engineering Tasks? An Approach Based on Genetic Algorithms
Annibale Panichella[1], Bogdan Dit[2], Rocco Oliveto[3], Massimilano Di Penta[4], Denys Poshynanyk[2], Andrea De Lucia[1]
[1]University of Salerno, Fisciano (SA), Italy
[2]The College of William and Mary, Williamsburg, VA, USA
[3]University of Molise, Pesche (IS), Italy
[4]University of Sannio, Benevento, Italy
@ ICSE 2013


Textual Information
Information Retrieval Techniques
Maintenance Tasks
Require Parameter Calibration
?
LDA-GA @ ICSE 2013
Sub-optimal results


Textual Information
Information Retrieval Techniques
Maintenance Tasks
Require Parameter Calibration
?
LDA-GA @ ICSE 2013
Sub-optimal results
Near-optimal results


Textual Information
Information Retrieval Techniques
Maintenance Tasks
Require Parameter Calibration
?
LDA-GA @ ICSE 2013
Sub-optimal results
Near-optimal results
CoEST
TraceLab


Textual Information
Information Retrieval Techniques
Maintenance Tasks
LDA
Require Parameter Calibration
?
Traceability
LDA-GA @ ICSE 2013
Sub-optimal results
Near-optimal results
CoEST
TraceLab

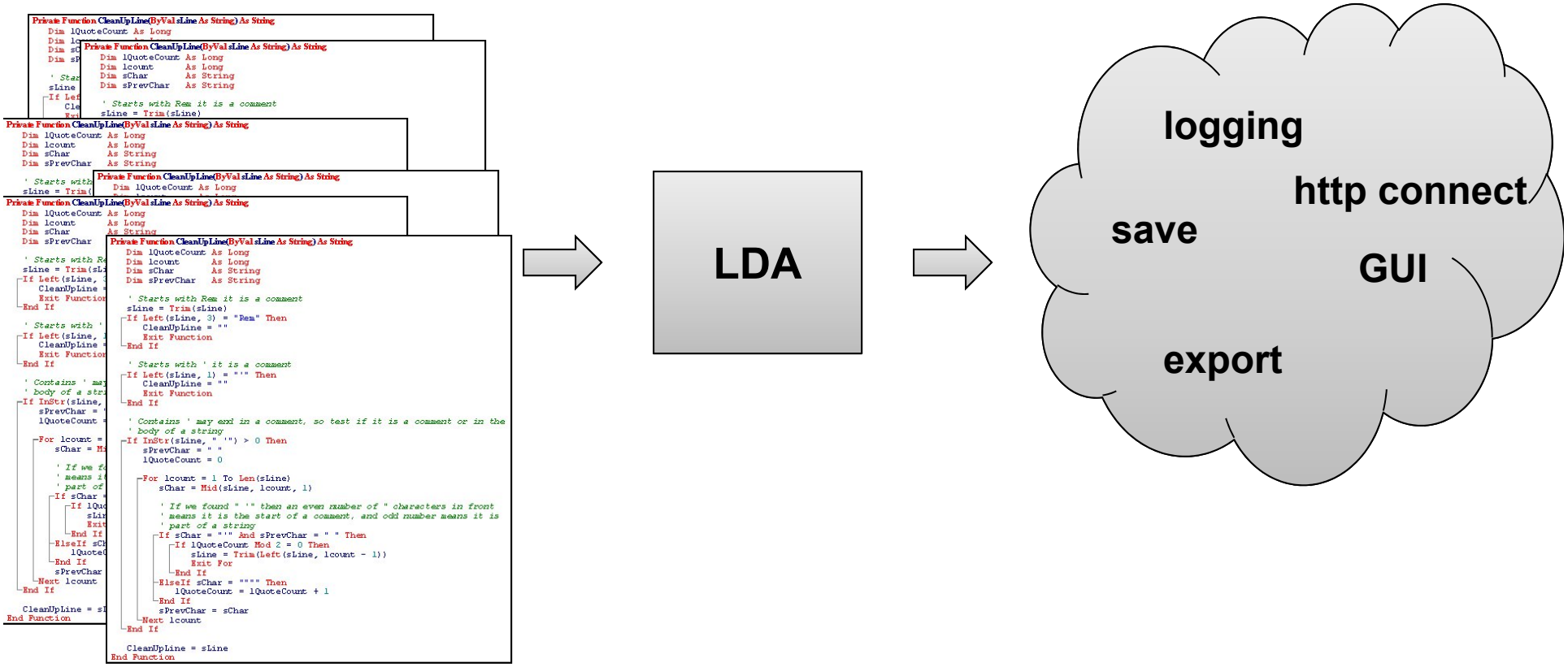
What is LDA?

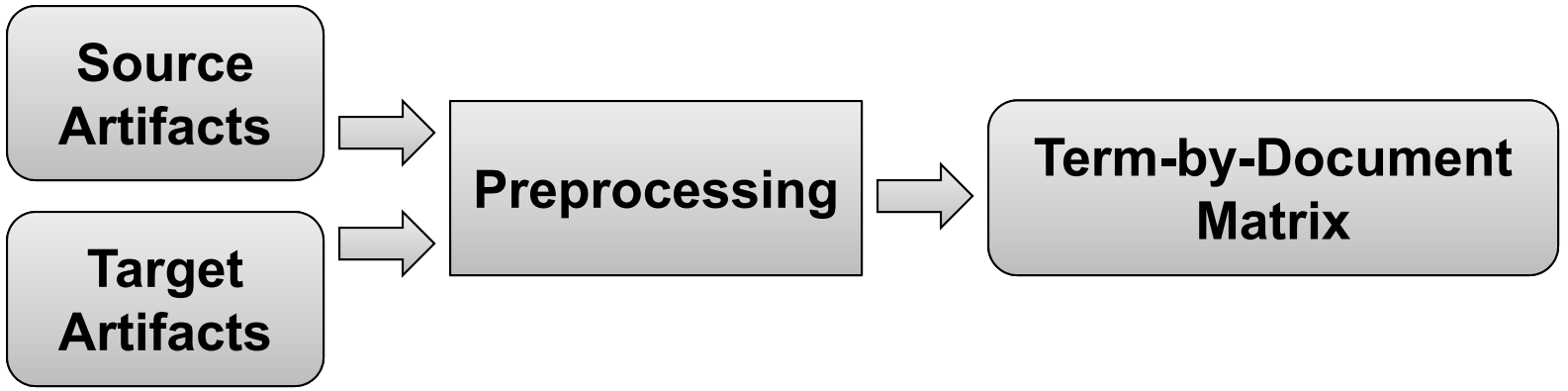
Latent **D**irichlet **A**llocation (LDA)

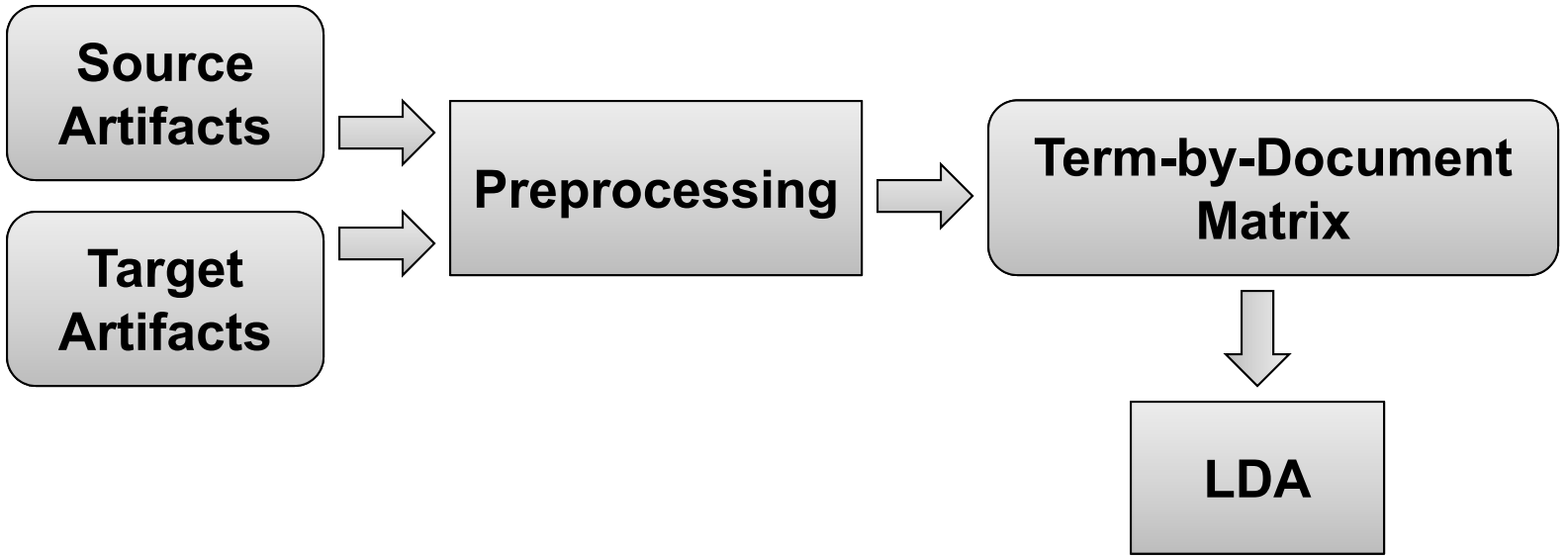
- Topic model that generates the distribution of latent topics from textual documents

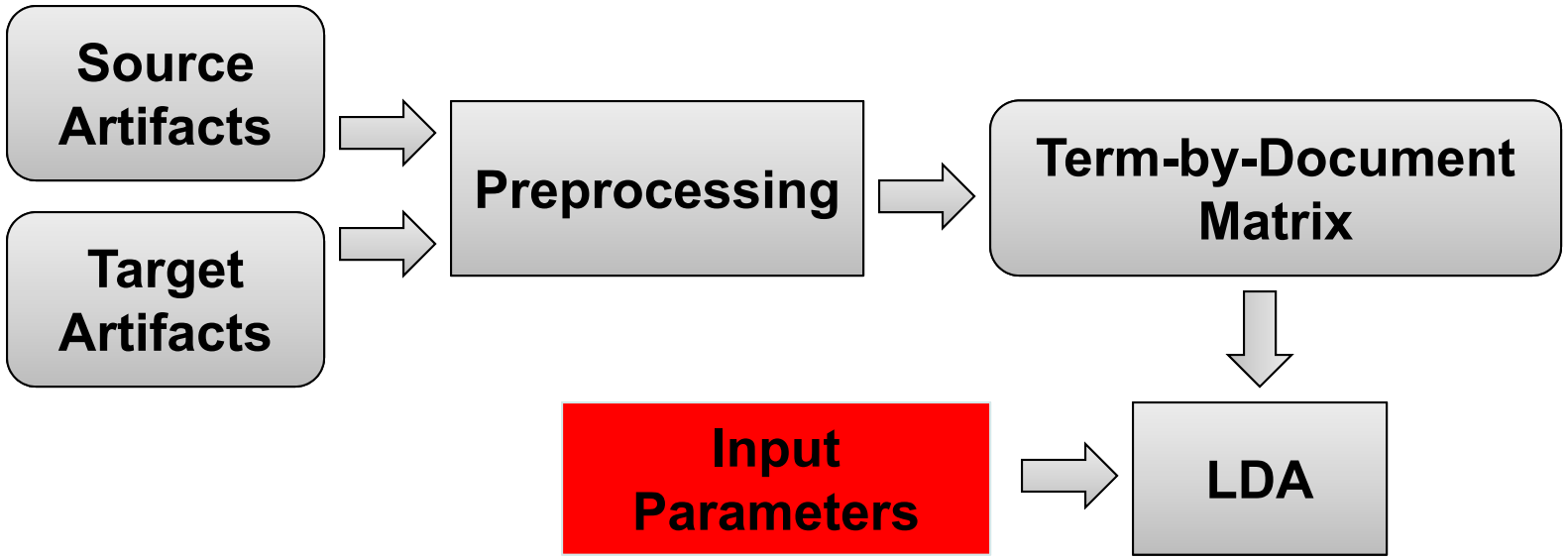
Latent Dirichlet Allocation (LDA)

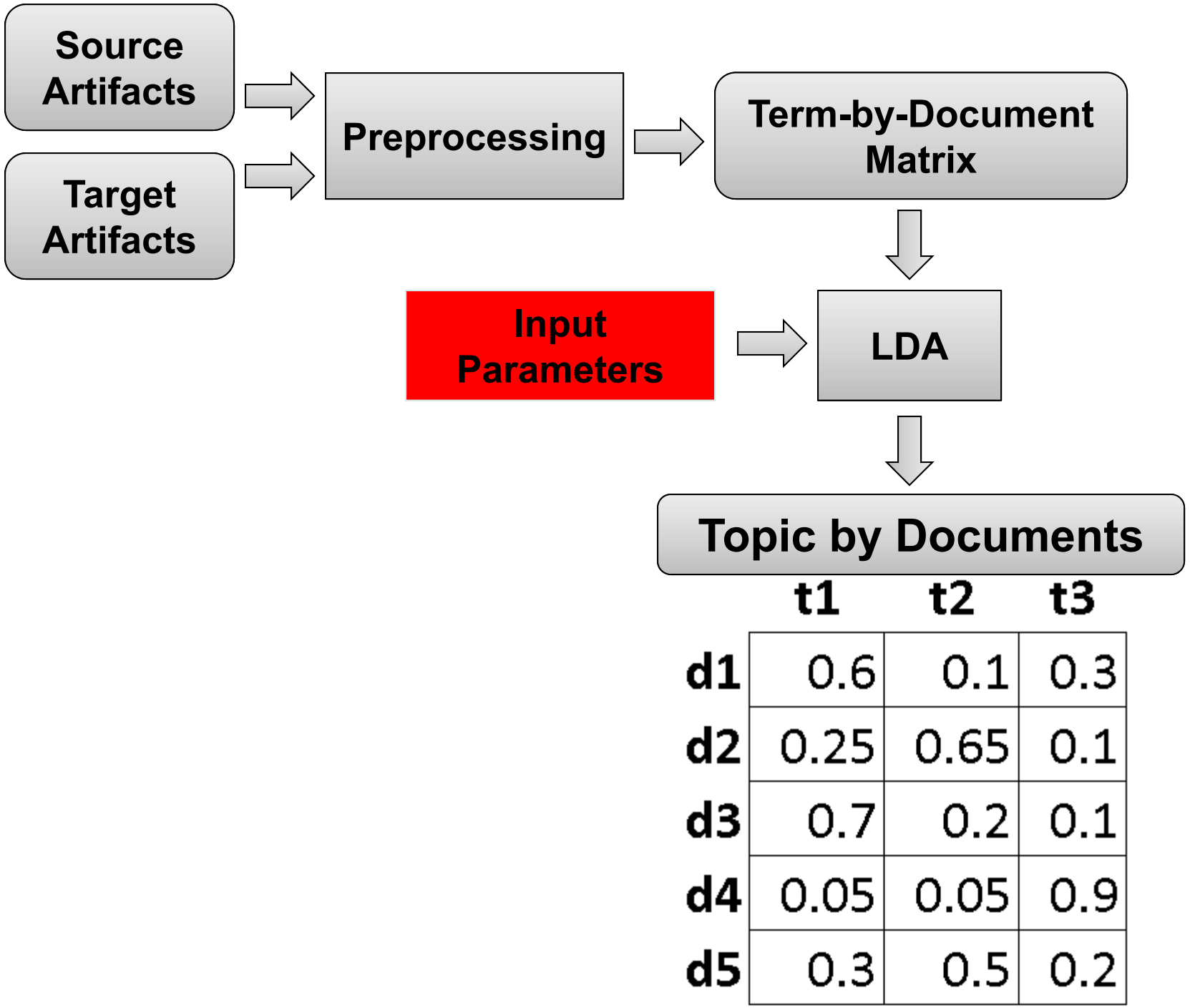
- Topic model that generates the distribution of latent topics from textual documents



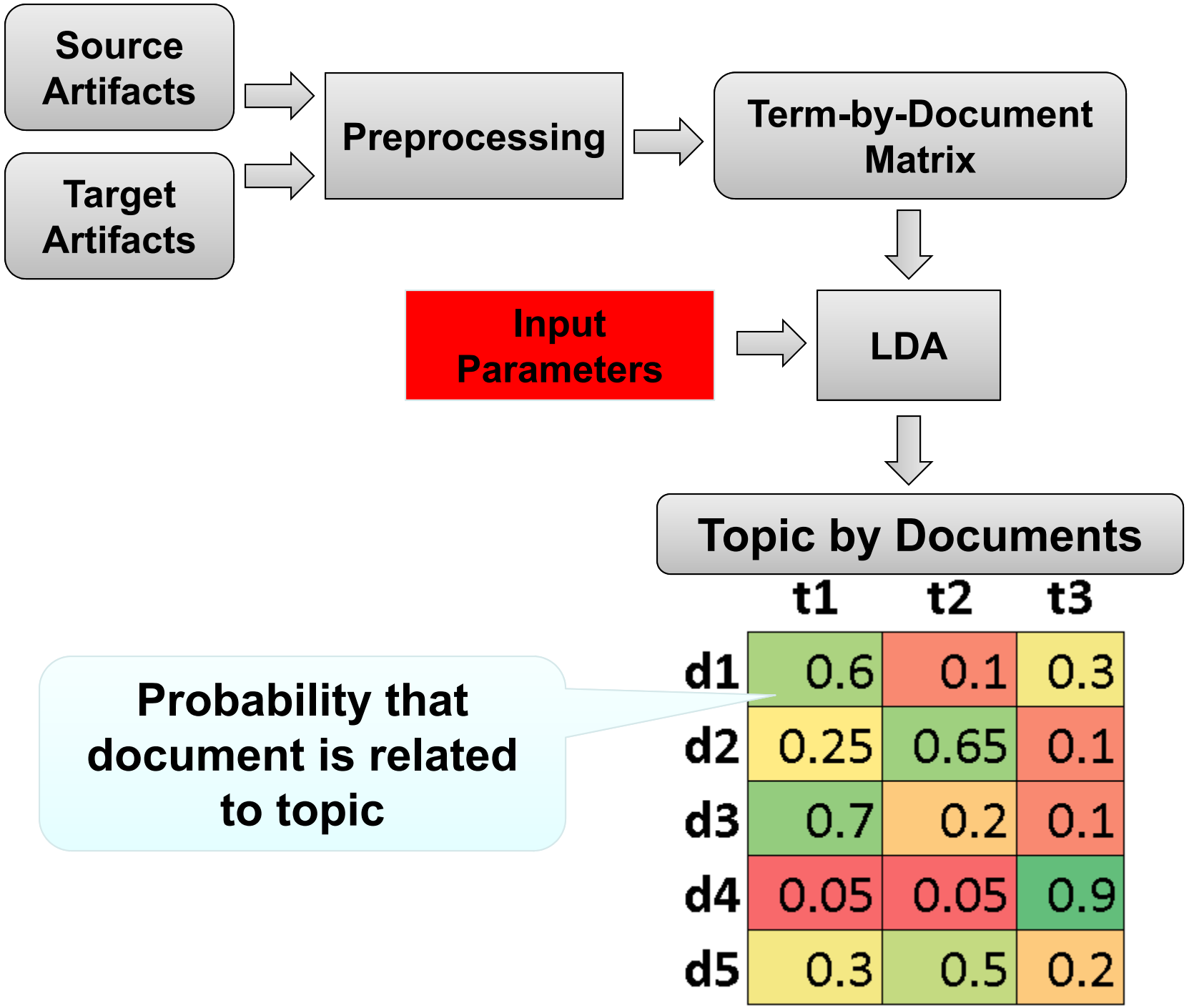

Source Artifacts
Target Artifacts
Preprocessing
Term-by-Document Matrix

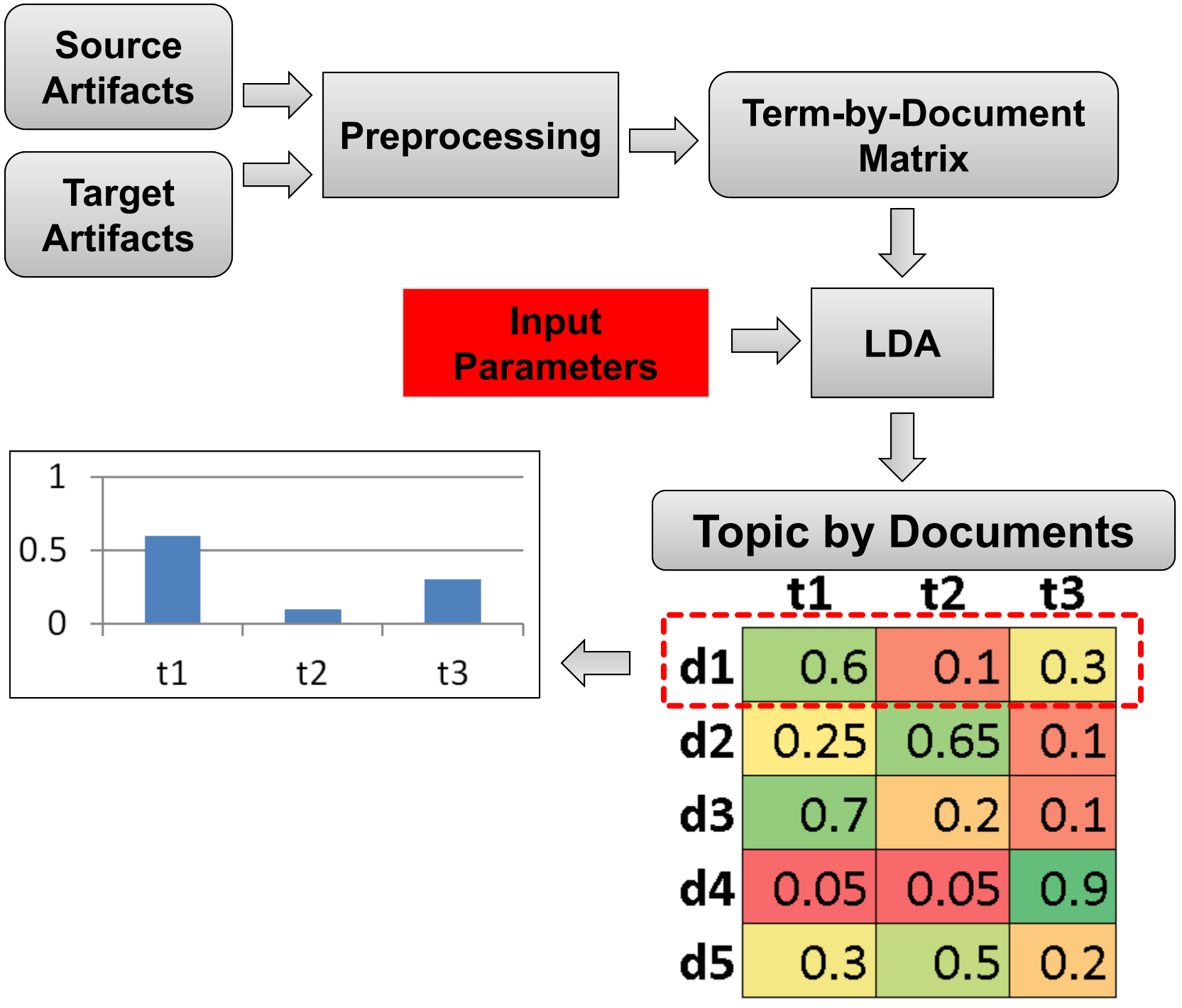

Source Artifacts
Target Artifacts
Preprocessing
Term-by-Document Matrix
LDA

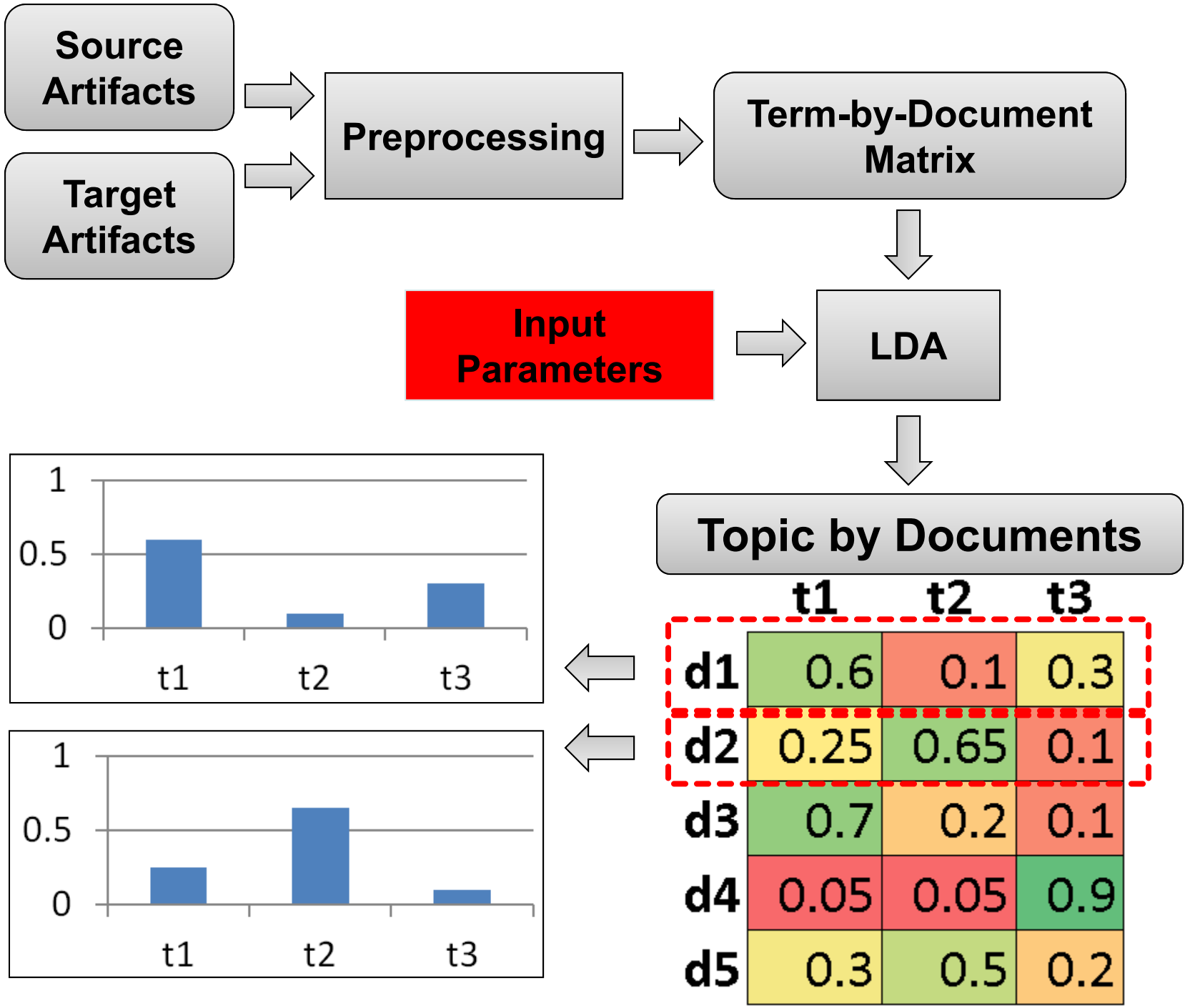

Source Artifacts
Target Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA

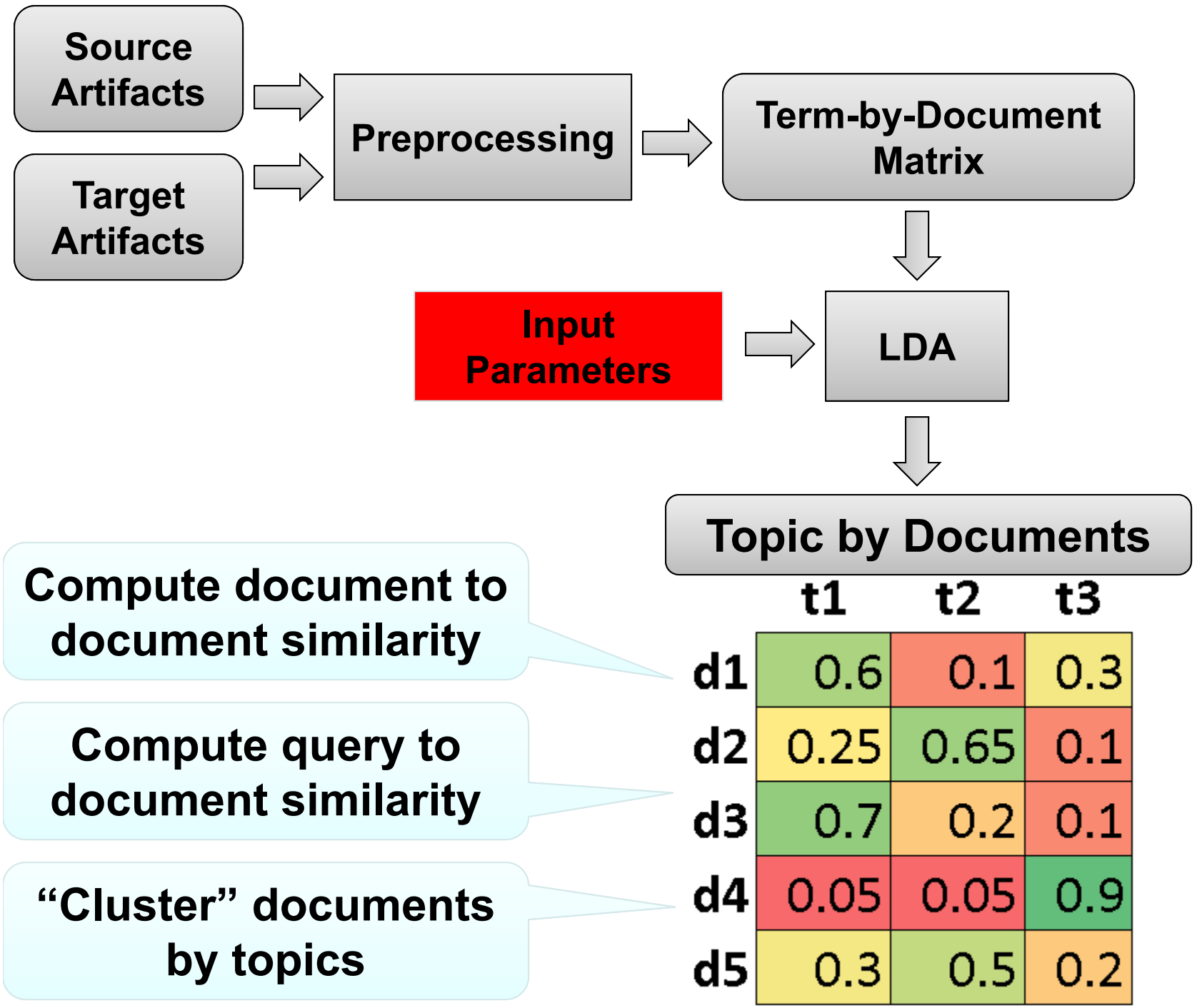


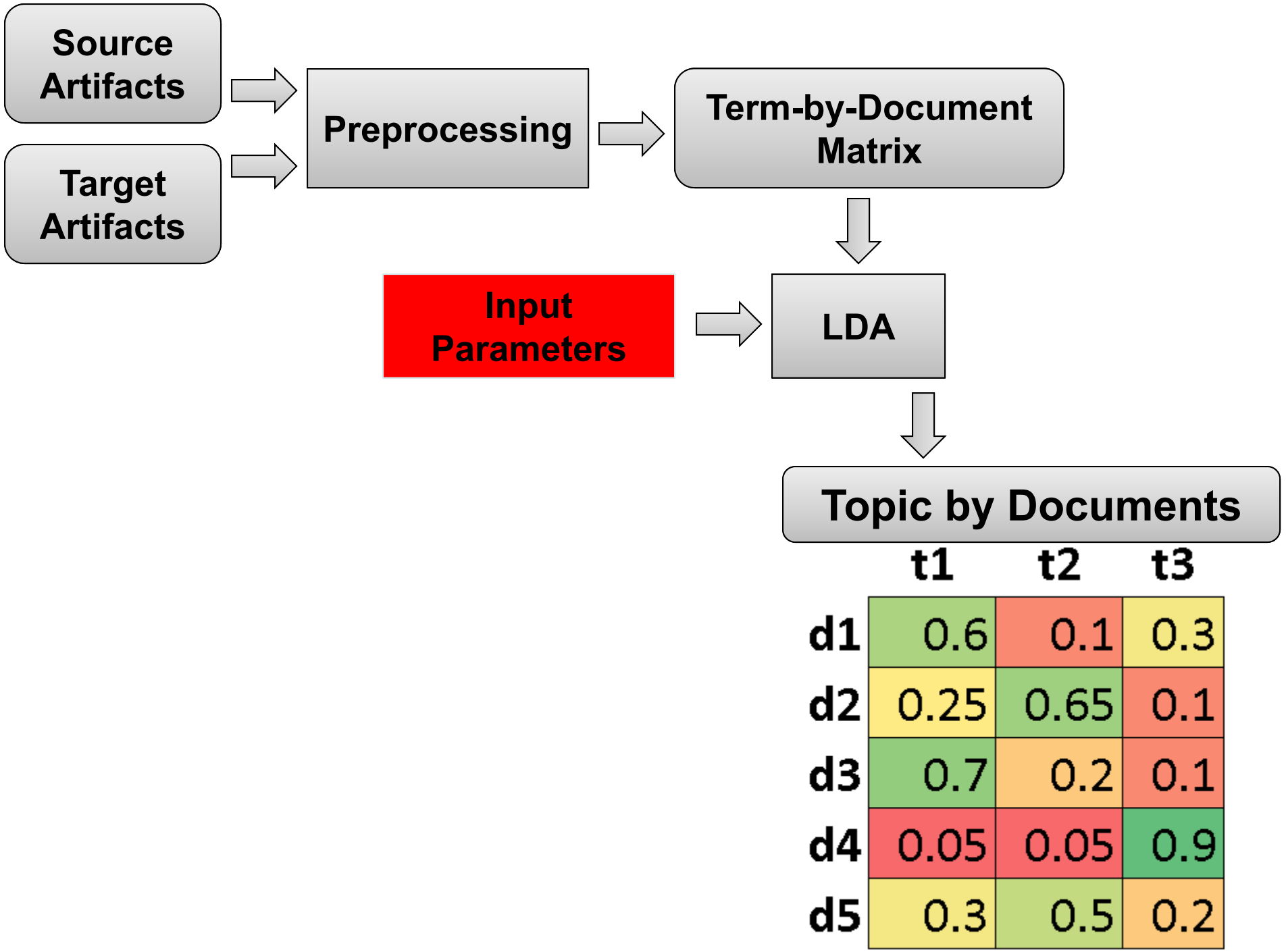
	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2

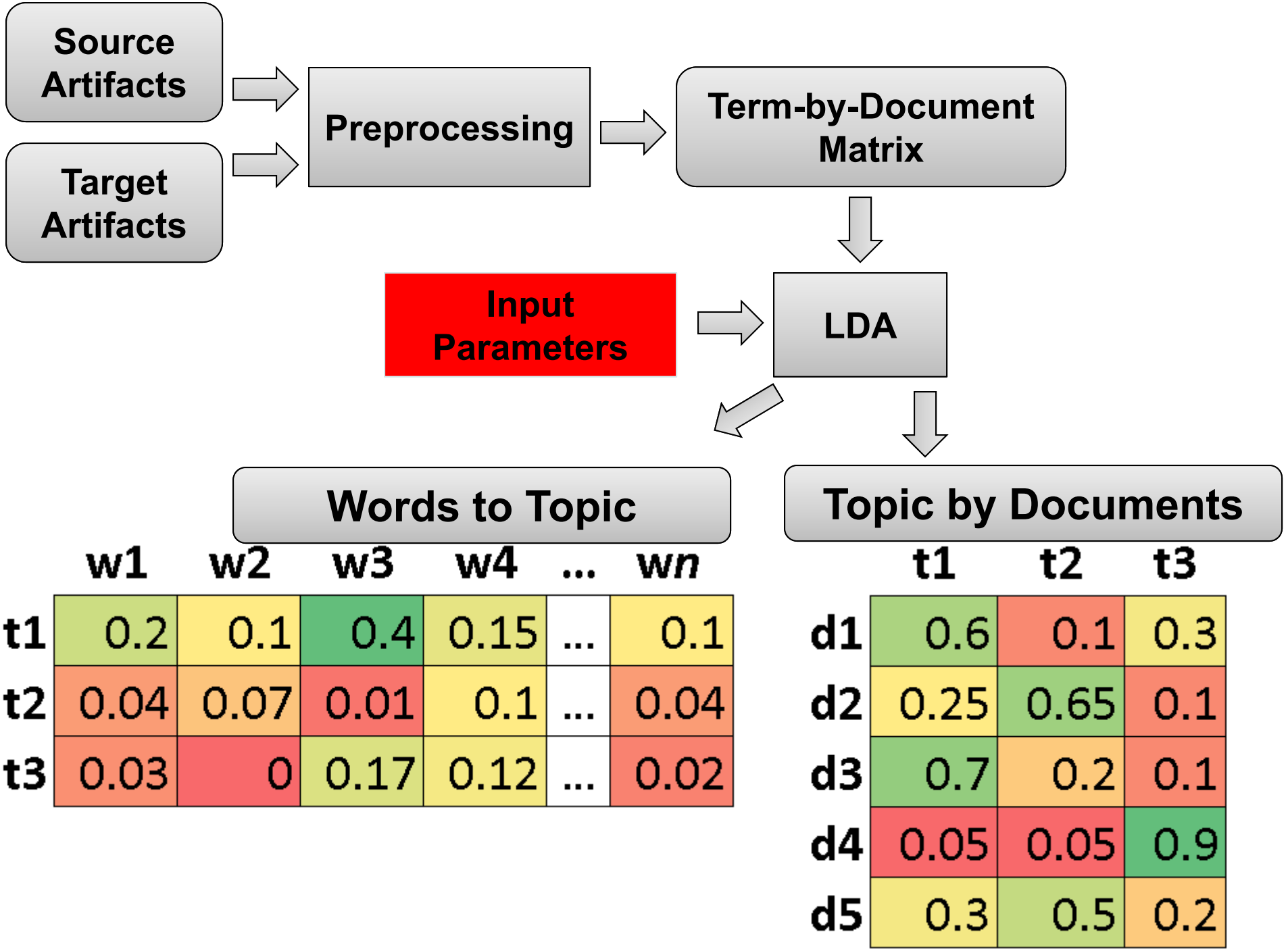

Source Artifacts
Target Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
Topic by Documents
t1 t2 t3
d1 0.6 0.1 0.3
d2 0.25 0.65 0.1
d3 0.7 0.2 0.1
d4 0.05 0.05 0.9
d5 0.3 0.5 0.2
Probability that document is related to topic


Source Artifacts
Target 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
1
0.5
0
t1
t2
t3


Source Artifacts
Target 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
1
0.5
0
t1 t2 t3
1
0.5
0
t1 t2 t3

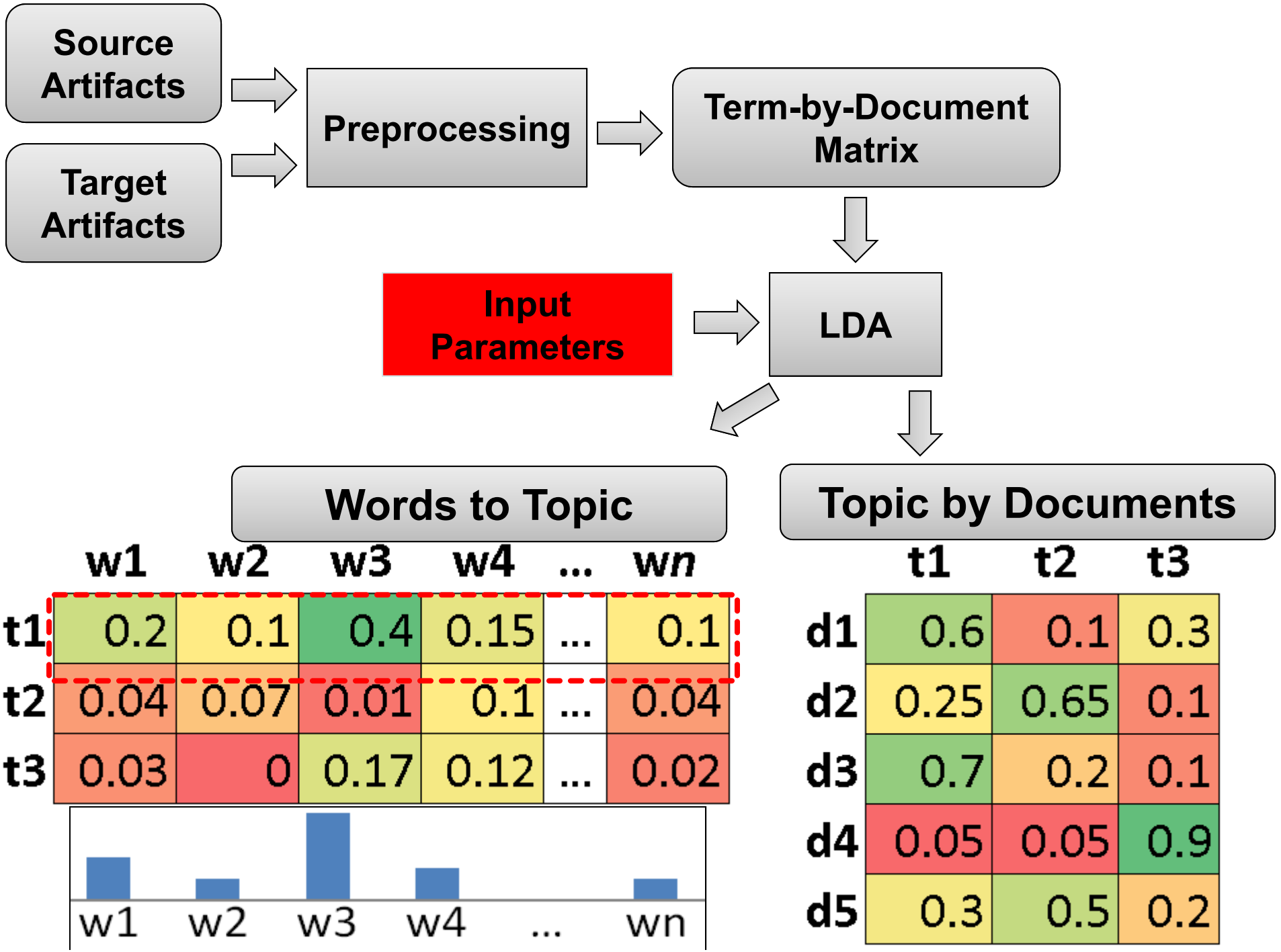

Source Artifacts
Target Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
Topic by Documents
t1 t2 t3
d1 0.6 0.1 0.3
d2 0.25 0.65 0.1
d3 0.7 0.2 0.1
d4 0.05 0.05 0.9
d5 0.3 0.5 0.2
Compute document to document similarity
Compute query to document similarity
“Cluster” documents by topics

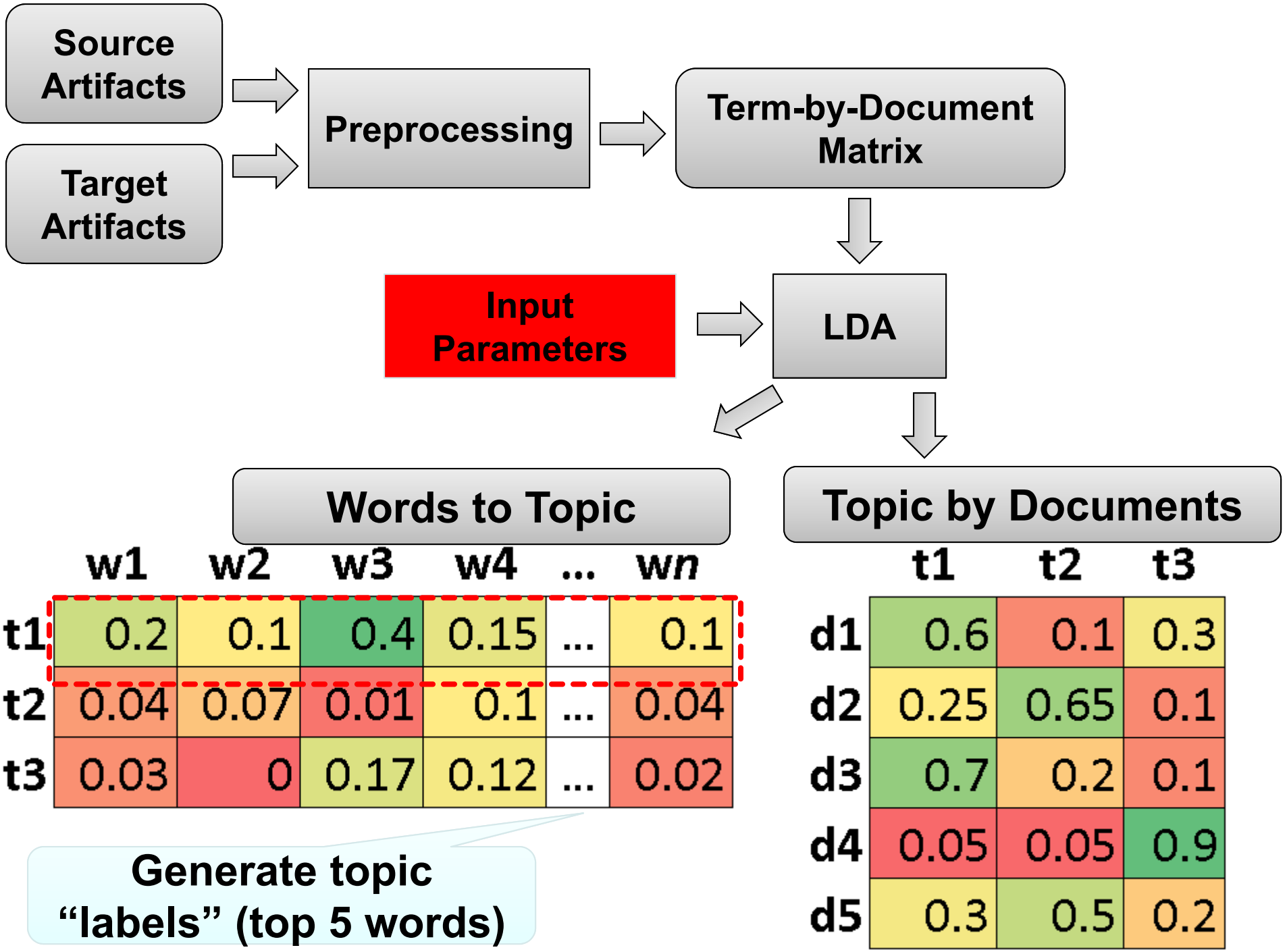

Source Artifacts
Target Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA
Topic by Documents
t1 t2 t3
d1 0.6 0.1 0.3
d2 0.25 0.65 0.1
d3 0.7 0.2 0.1
d4 0.05 0.05 0.9
d5 0.3 0.5 0.2

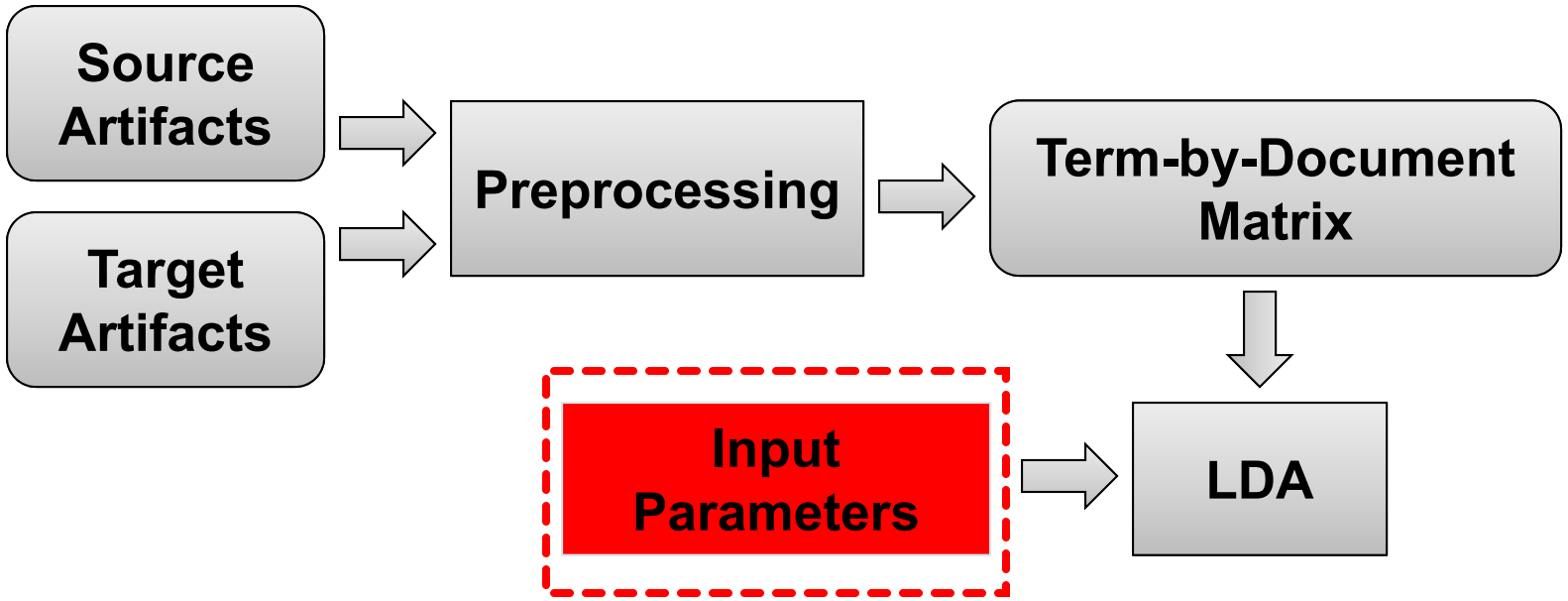


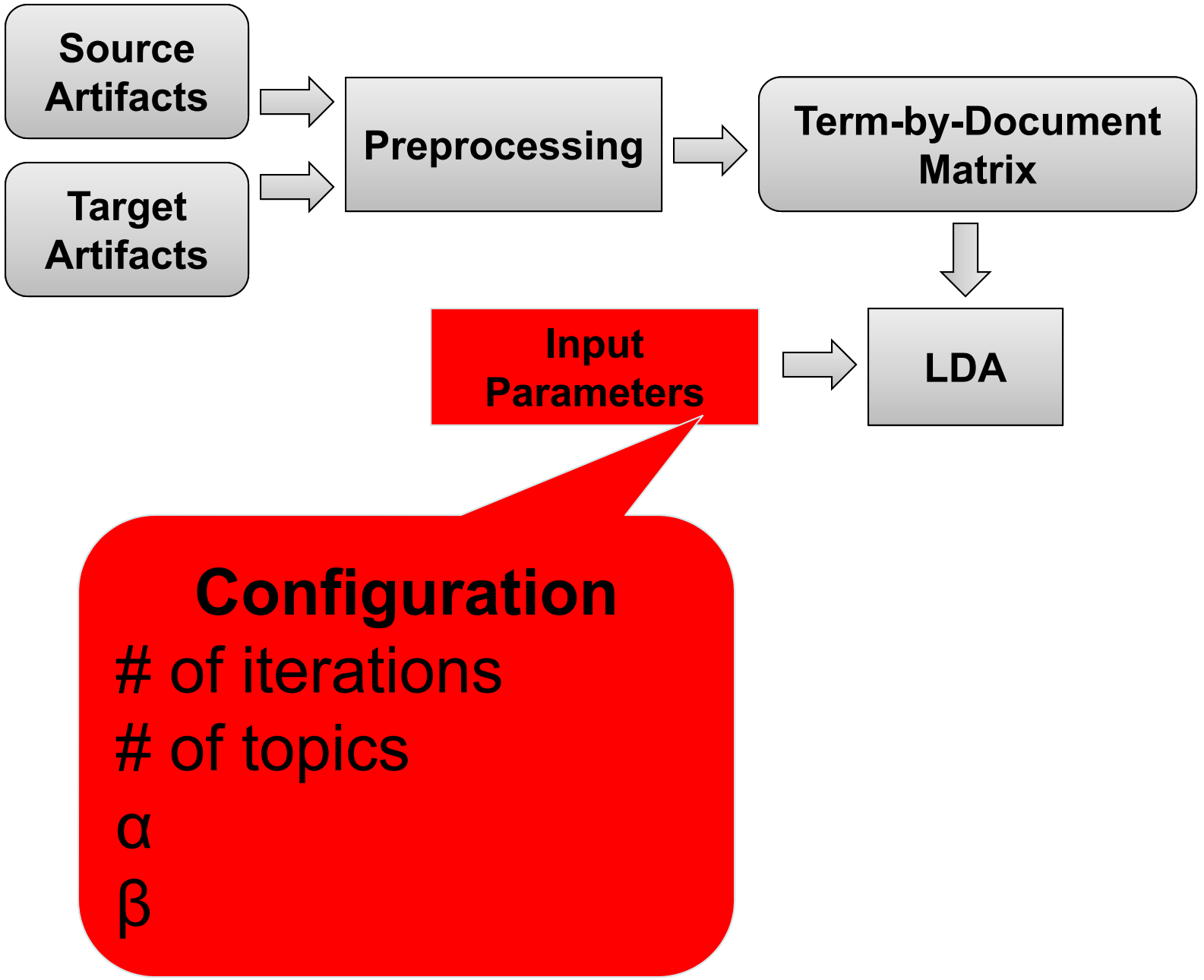
	w1	w2	w3	w4	...	wn
t1	0.2	0.1	0.4	0.15	...	0.1
t2	0.04	0.07	0.01	0.1	...	0.04
t3	0.03	0	0.17	0.12	...	0.02

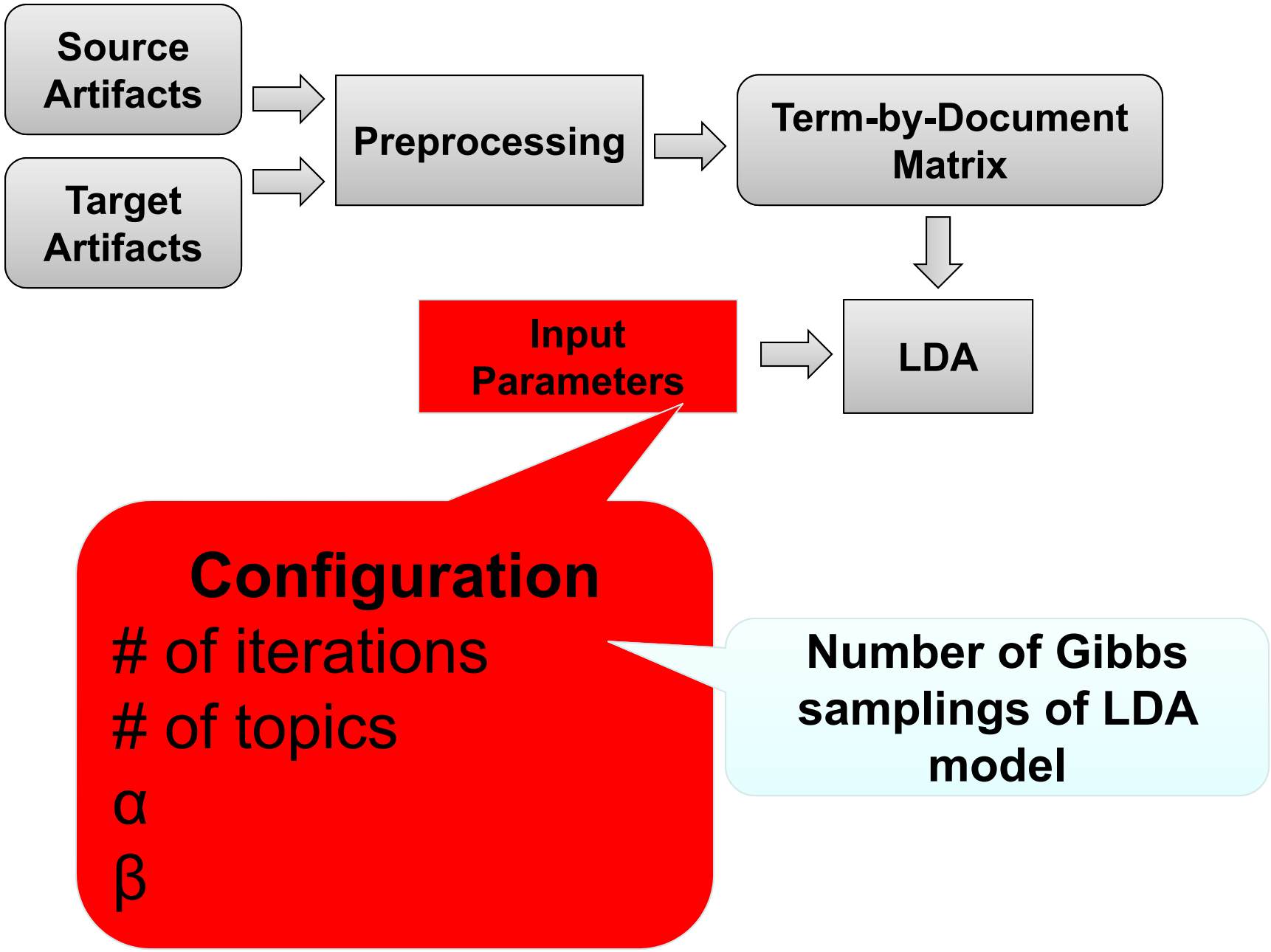
	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2


Source Artifacts
Target 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


Source Artifacts
Target 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
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
Generate topic "labels" (top 5 words)

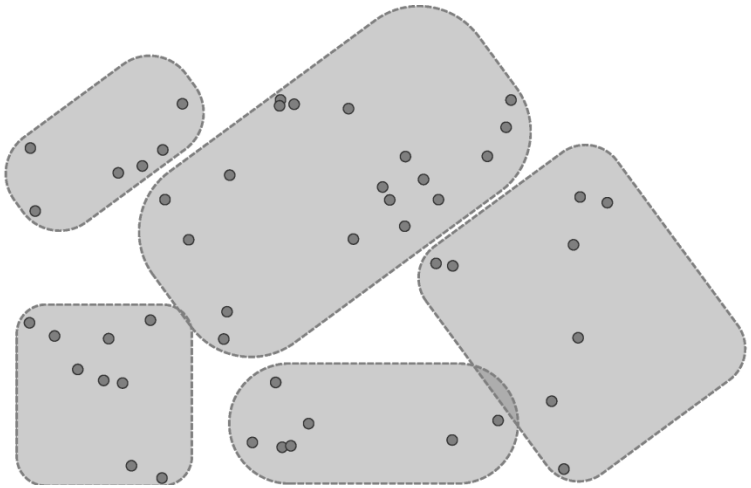

Source Artifacts
Target Artifacts
Preprocessing
Term-by-Document Matrix
Input Parameters
LDA


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


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

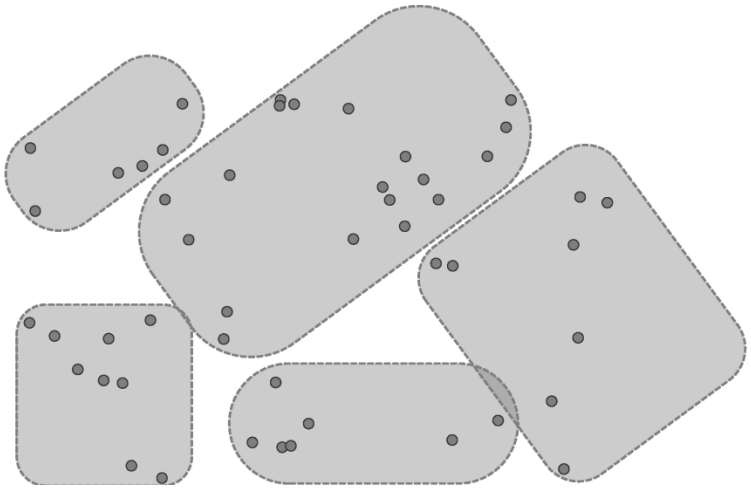
Number of topics…

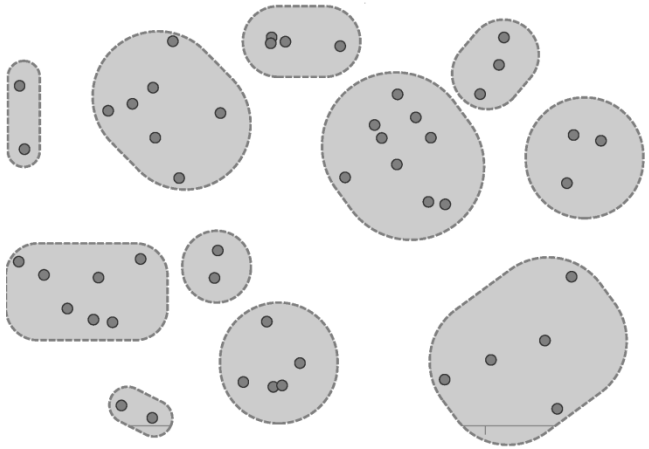
Number of topics…



General topics

Number of topics…



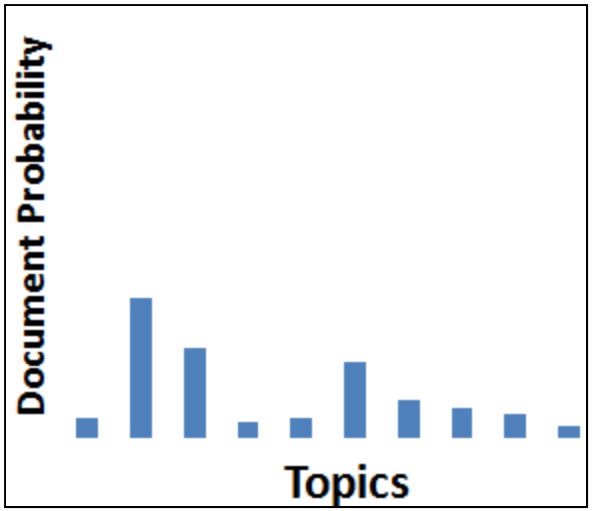


General topics

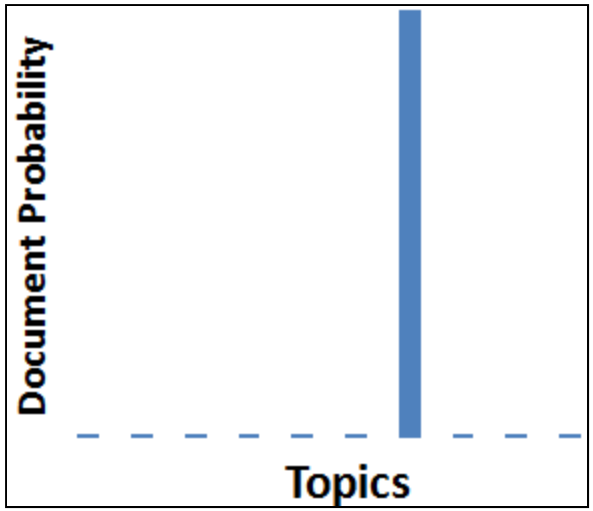
Specific topics

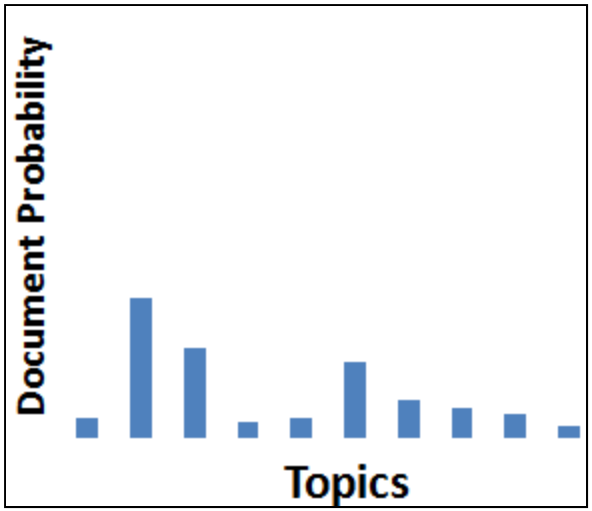
α, influences the "smoothness" of documents to topics distribution

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

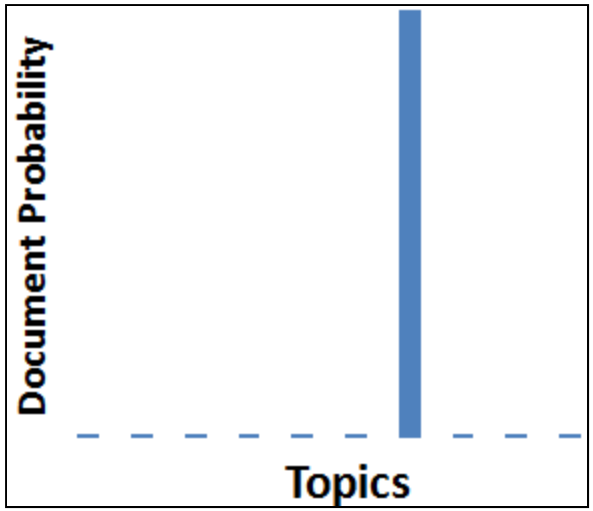


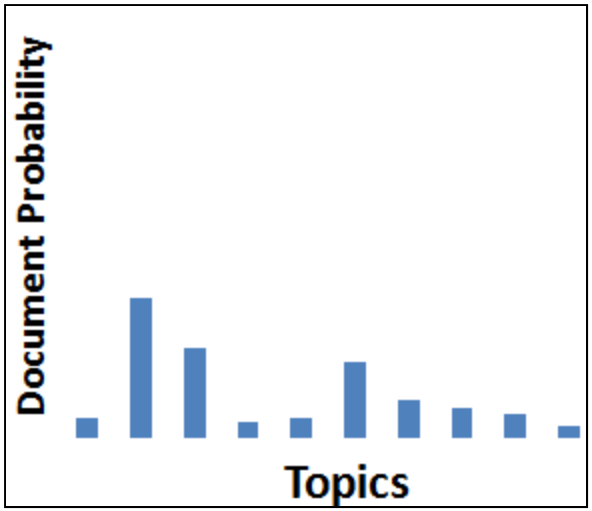
α, influences the "smoothness" of documents to topics distribution

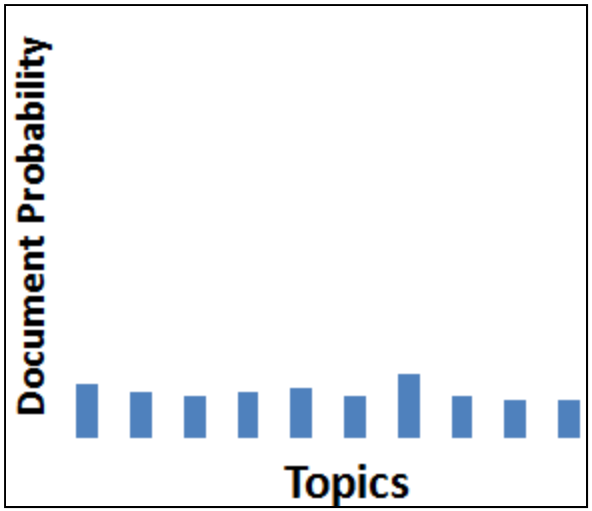




α, influences the "smoothness" of documents to topics distribution

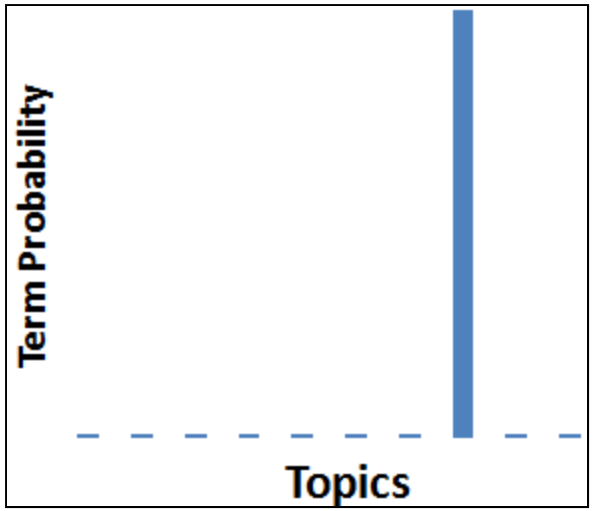


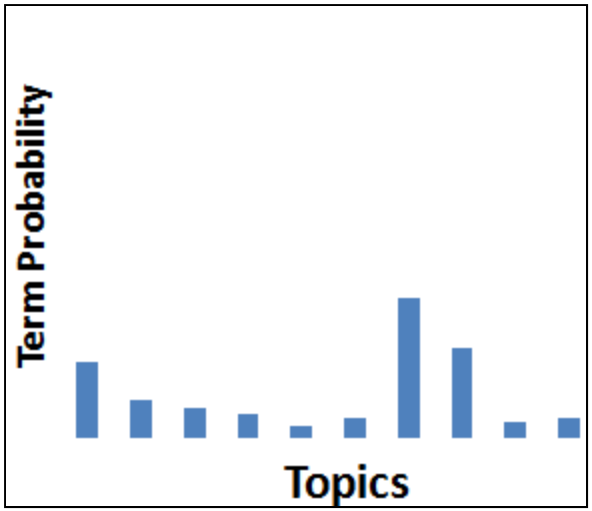


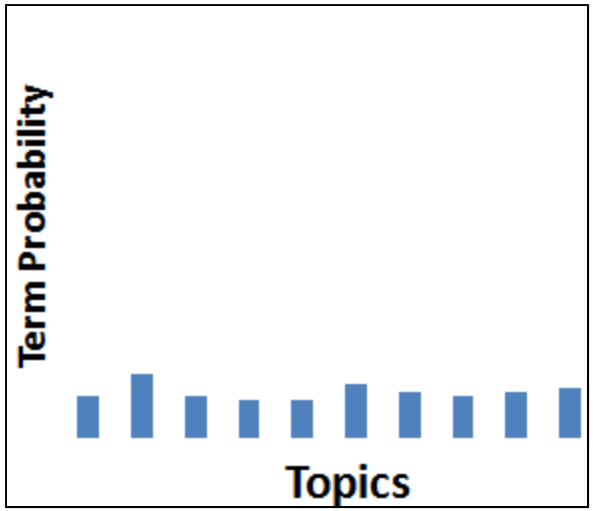


β, influences the "smoothness" of topics to words distribution

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





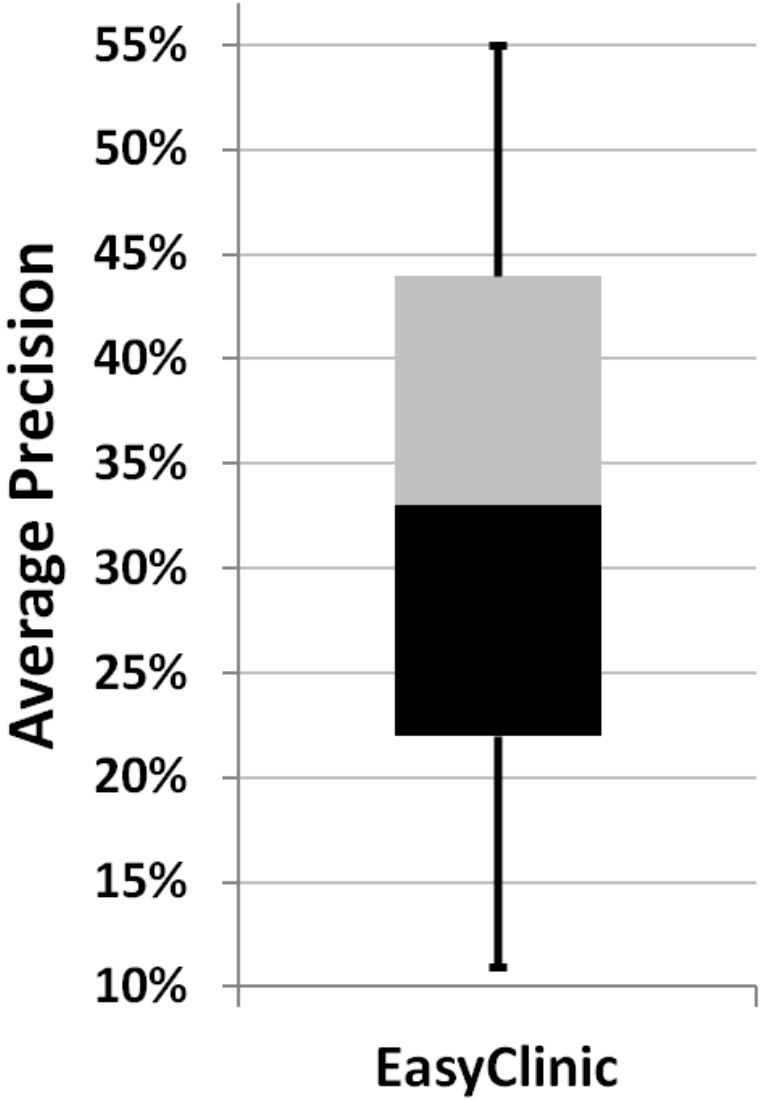


LDA parameters significantly influence the results

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

LDA parameters significantly influence the results

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



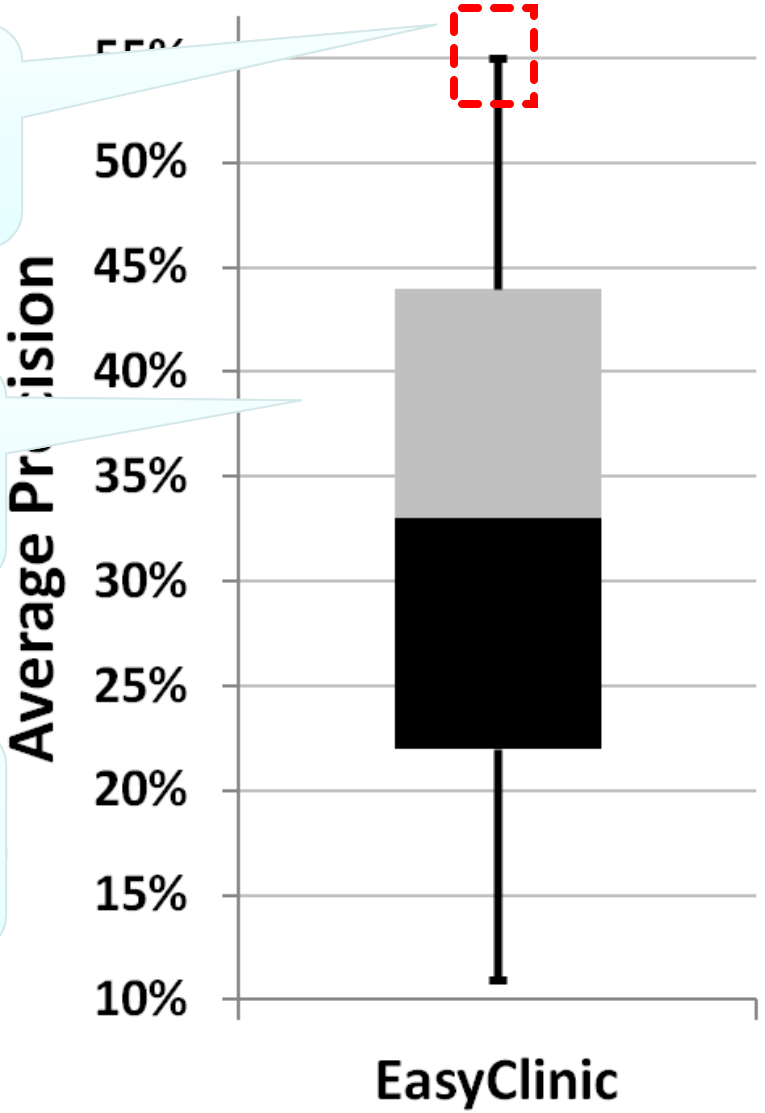
LDA parameters significantly influence the results

- **Few configurations produce good results**

LDA parameters

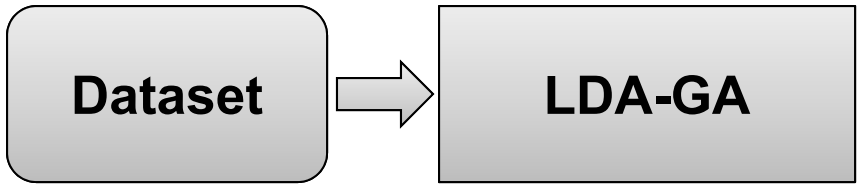
High variability in results

LDA-GA

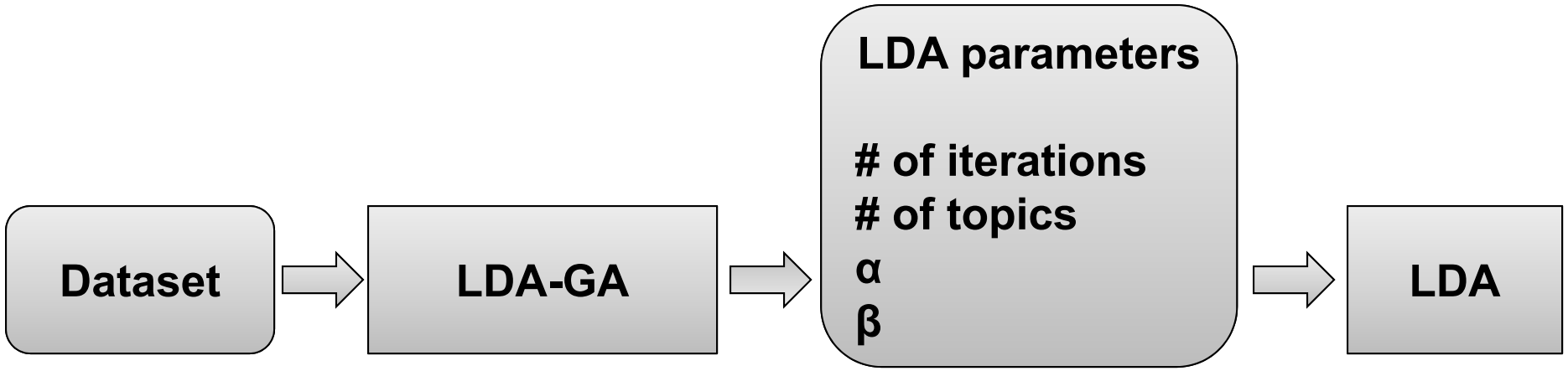


LDA-GA: automatically calibrate the input parameters of LDA using genetic algorithms

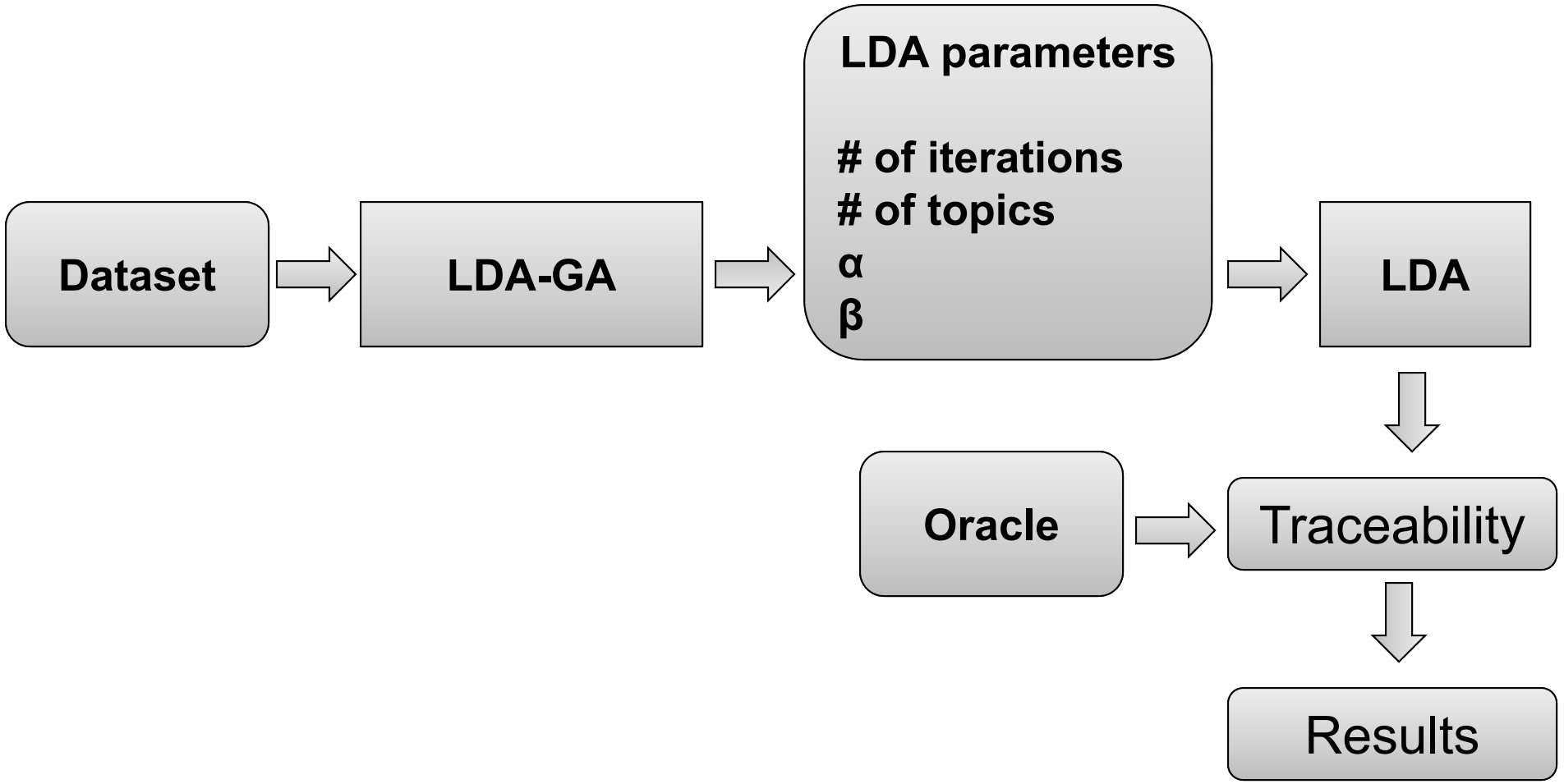
LDA-GA: automatically calibrate the input parameters of LDA using genetic algorithms



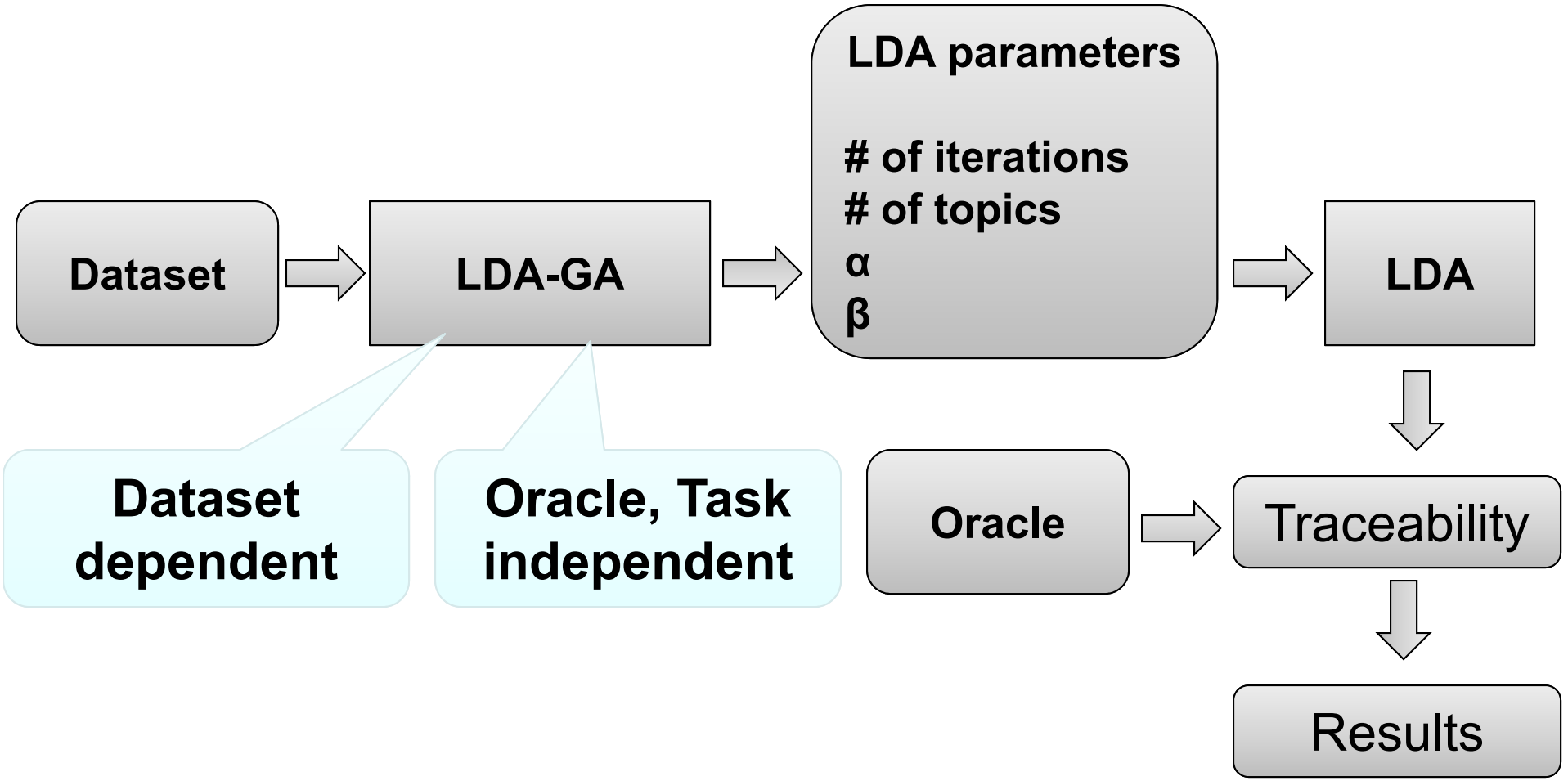
LDA-GA: automatically calibrate the input parameters of LDA using genetic algorithms



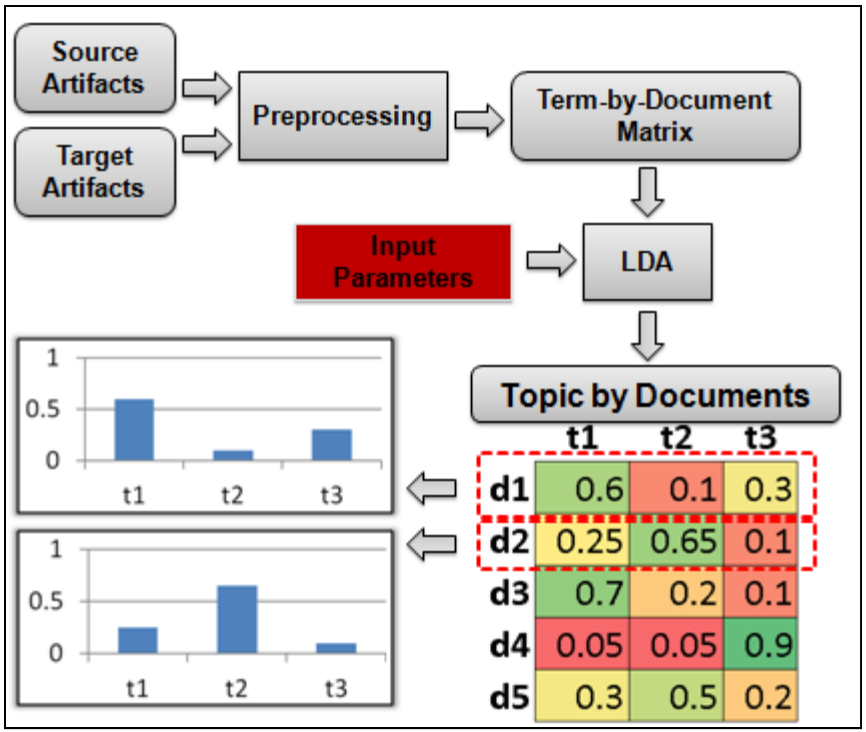
LDA-GA: automatically calibrate the input parameters of LDA using genetic algorithms



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

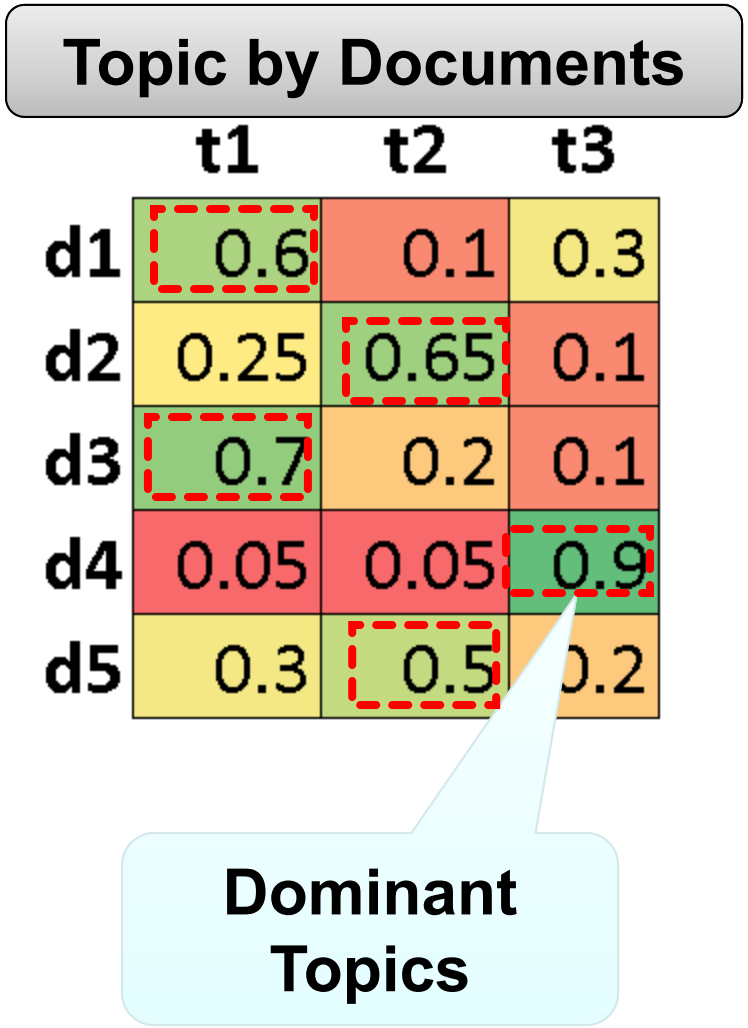


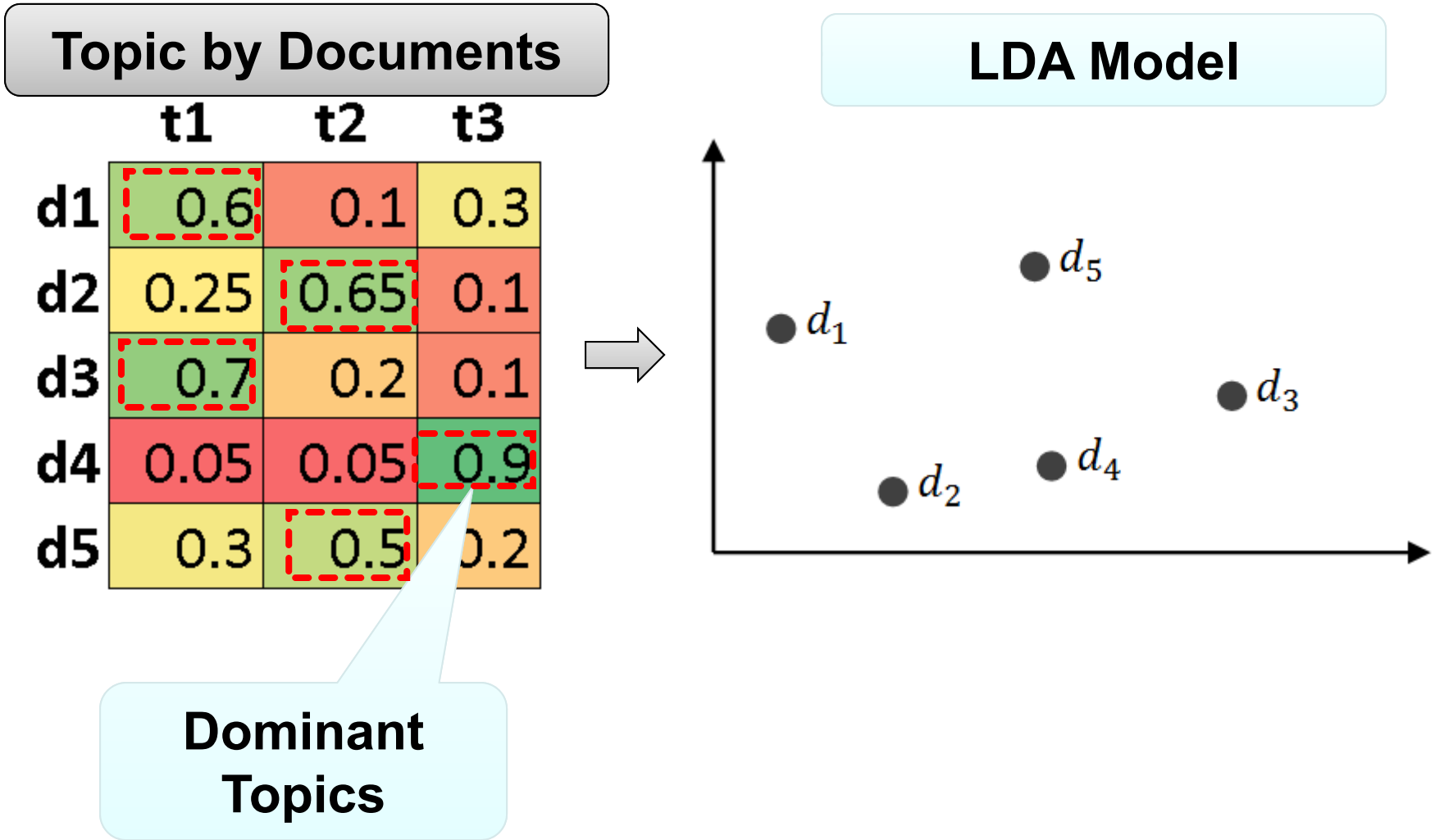
Evaluating the LDA parameter configuration



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


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
Dominant Topics

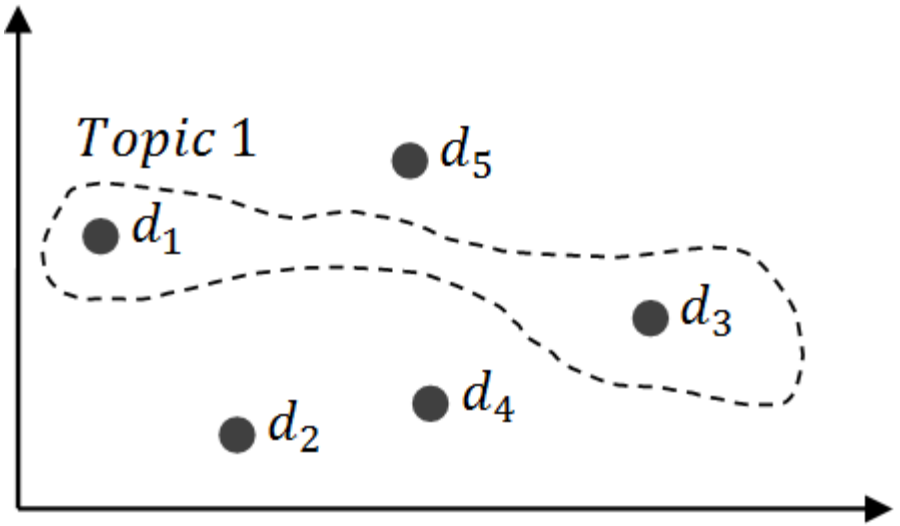

Topic by Documents
t1
t2
t3
d1
0.6
0.1
0.3
d2
0.25
0.65
0.1
d3
0.7
0.2
0.1
d4
0.05
0.05
0.9
d5
0.3
0.5
.2
Dominant Topics
LDA Model
d_1
d_2
d_3
d_4
d_5

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

Dominant Topics

LDA Model

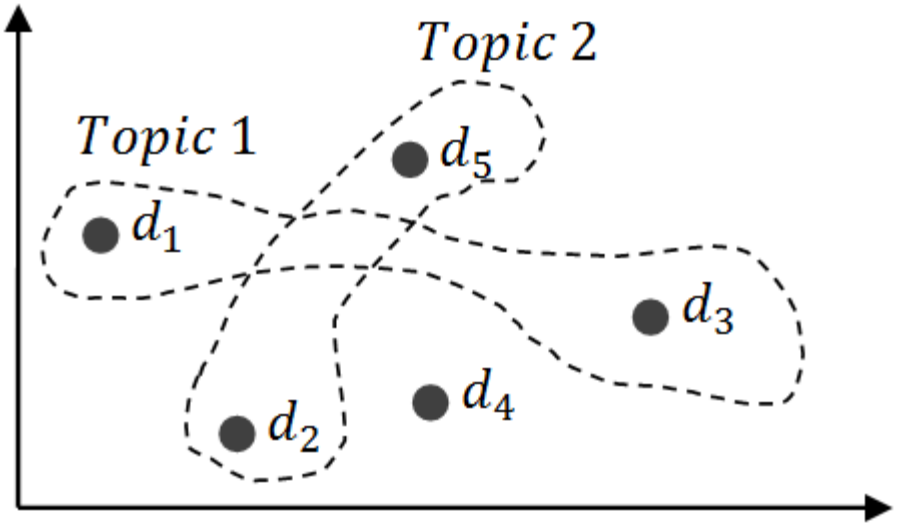


Topic by Documents

	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

Dominant Topics

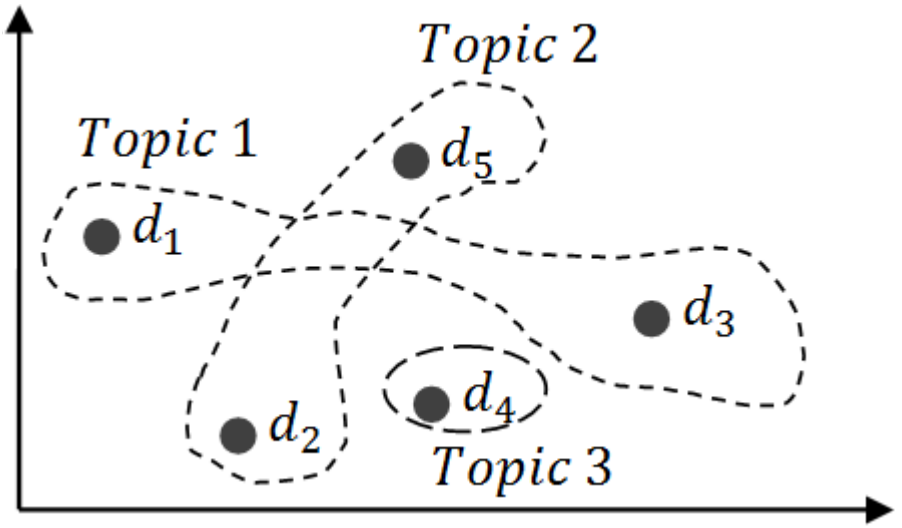
LDA Model



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

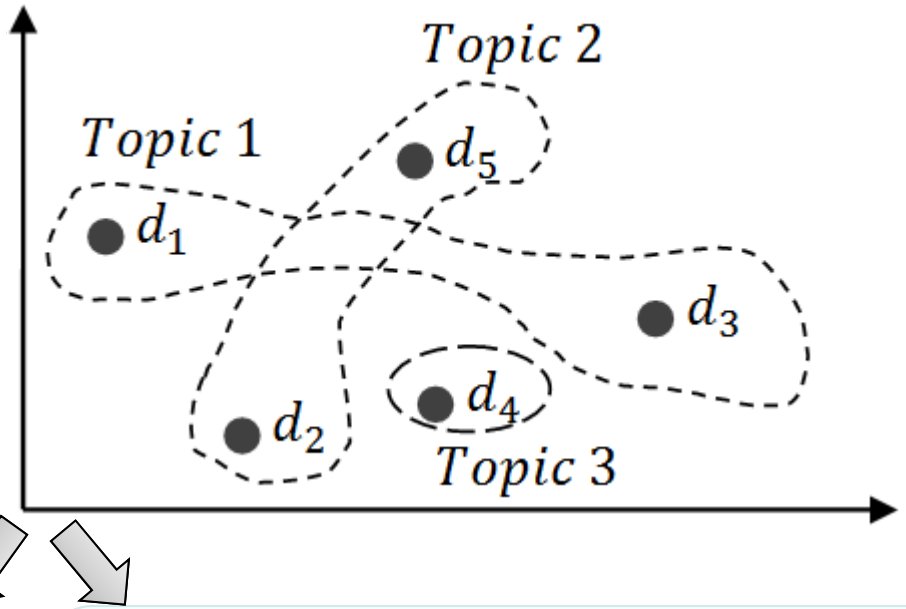
LDA Model



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

LDA Model



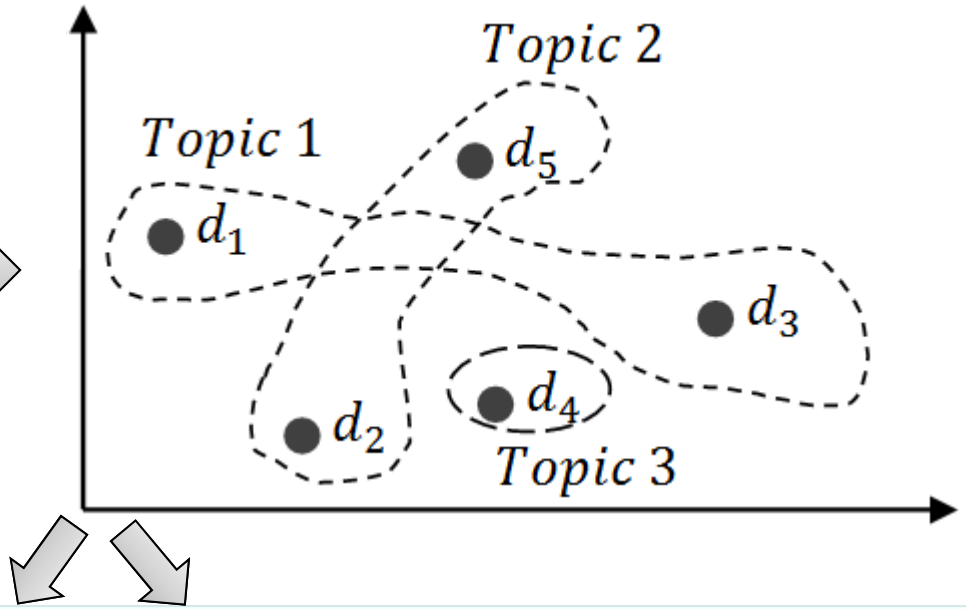
Cohesion (similarity): how related the documents in the same clusters are

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

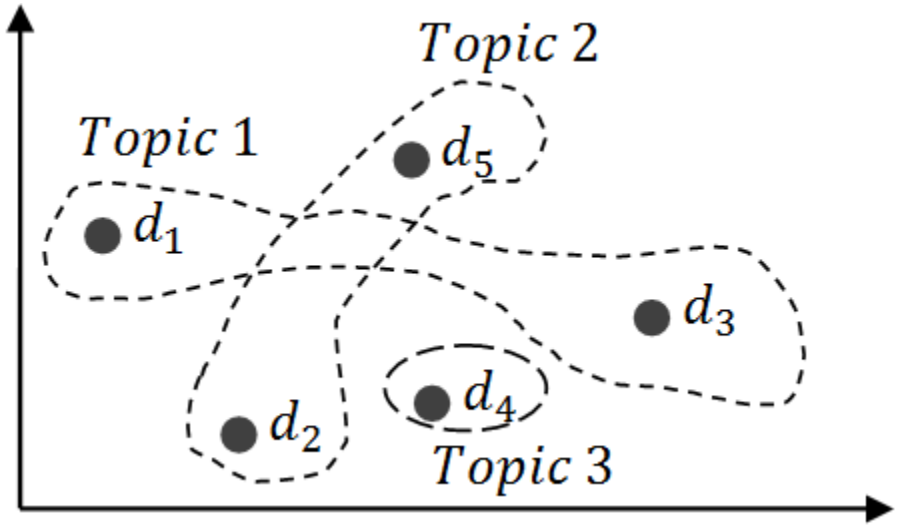
Topic by Documents

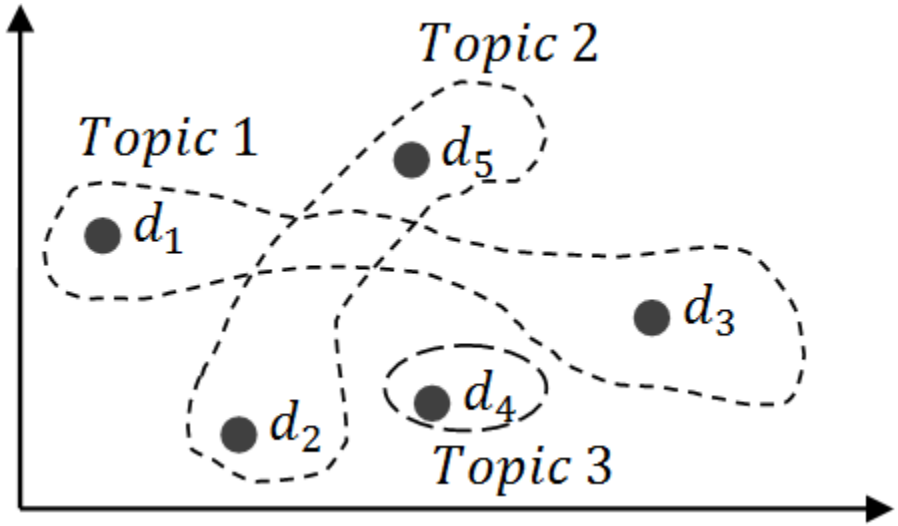
	t1	t2	t3
d1	0.6	0.1	0.3
d2	0.25	0.65	0.1
d3	0.7	0.2	0.1
d4	0.05	0.05	0.9
d5	0.3	0.5	0.2

LDA Model

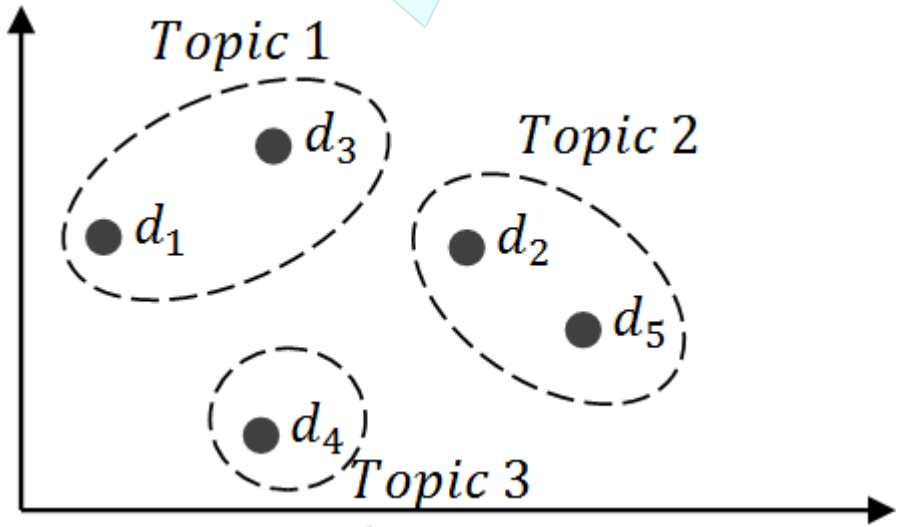


Silhouette coefficient


Topic 2
Topic 1
d_5
d_1
d_3
d_4
d_2
Topic 3



Cohesive
Separated



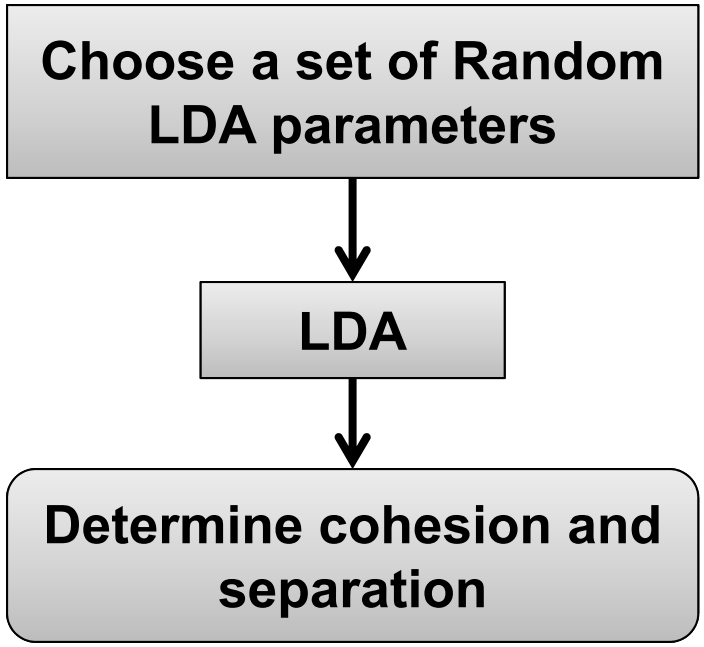
Documents are rearranged more closely in the n-dimensional space

We propose a genetic algorithm to find those solutions

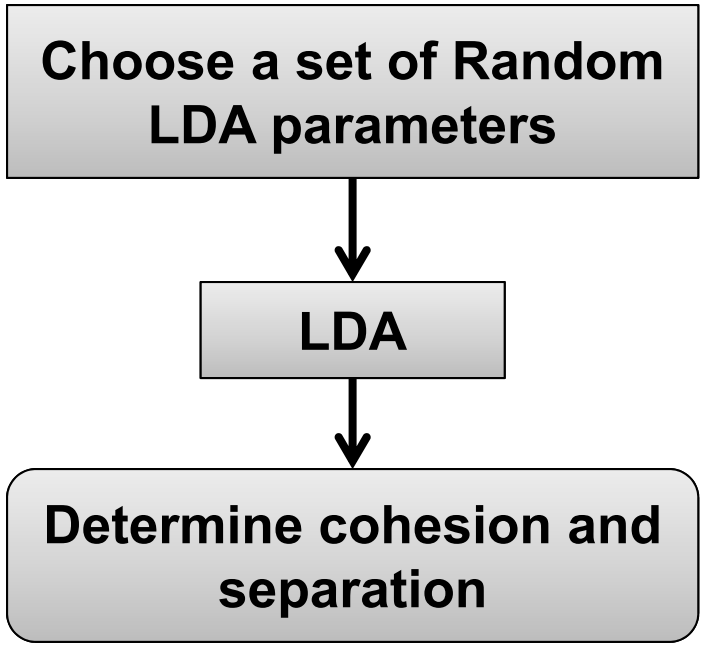
Choose a set of Random LDA parameters

Choose a set of Random 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. n	618	250	1.14	0.74

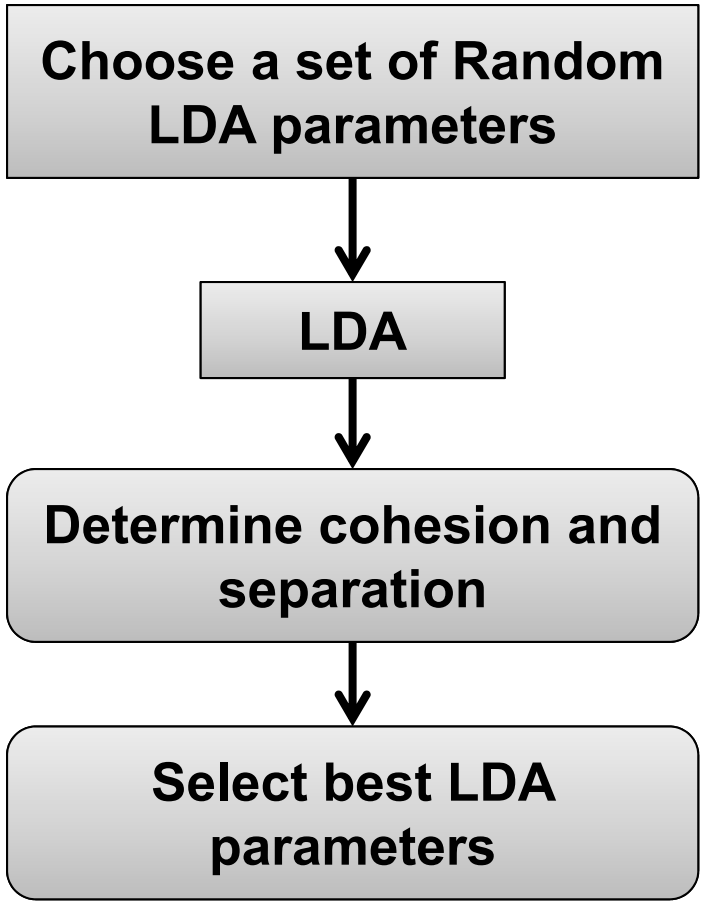


	iteration	topics	α	β
LDA Cfg. 1	510	74	0.34	2.5
LDA Cfg. 2	725	128	1.28	0.4
...	...	...	...	...
LDA Cfg. n	618	250	1.14	0.74

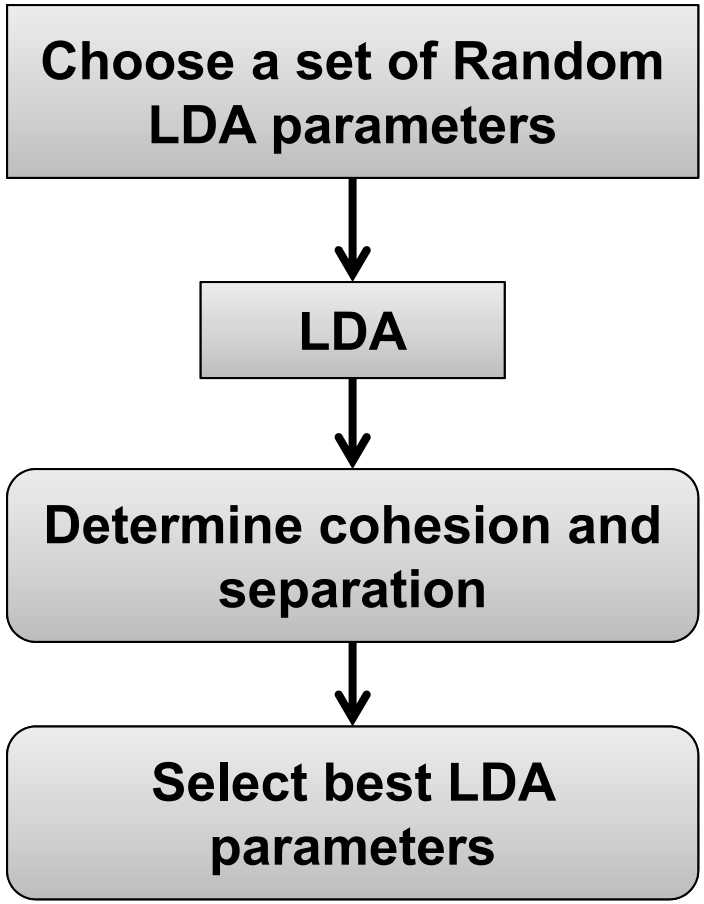


	iteration	topics	α	β	Fitness
LDA Cfg. 1	510	74	0.34	2.5	0.2
LDA Cfg. 2	725	128	1.28	0.4	0.3
...	...	...	...	...	
LDA Cfg. n	618	250	1.14	0.74	0.1

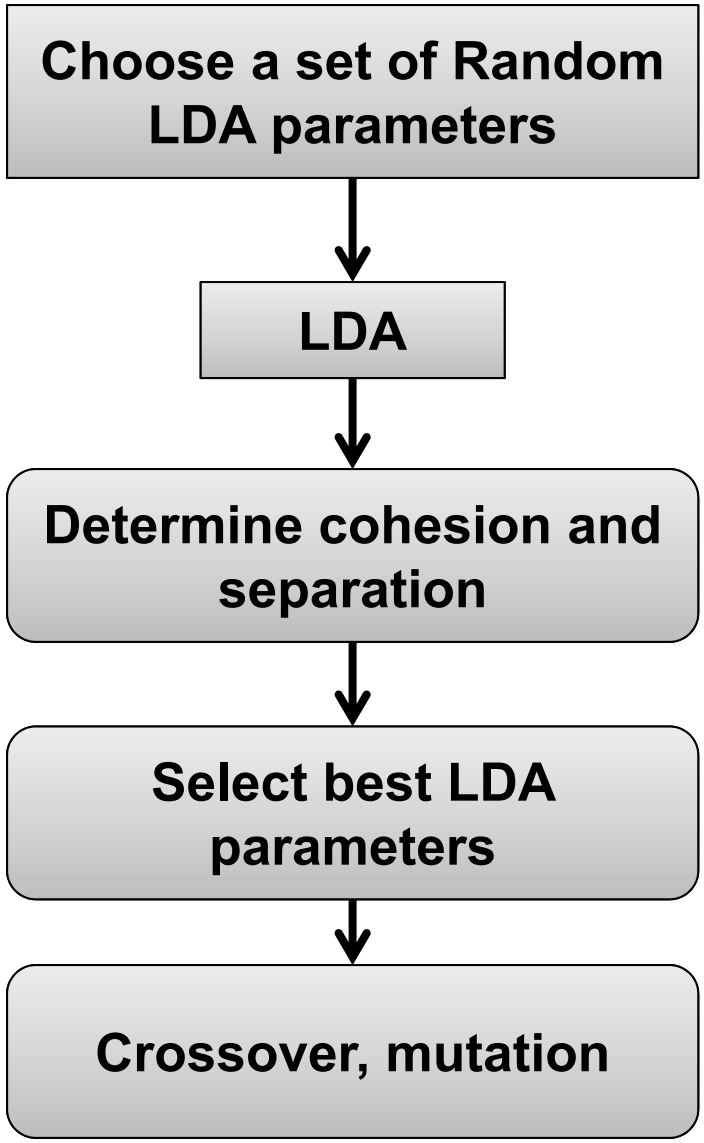
Cohesion and Separation (Silhouette)

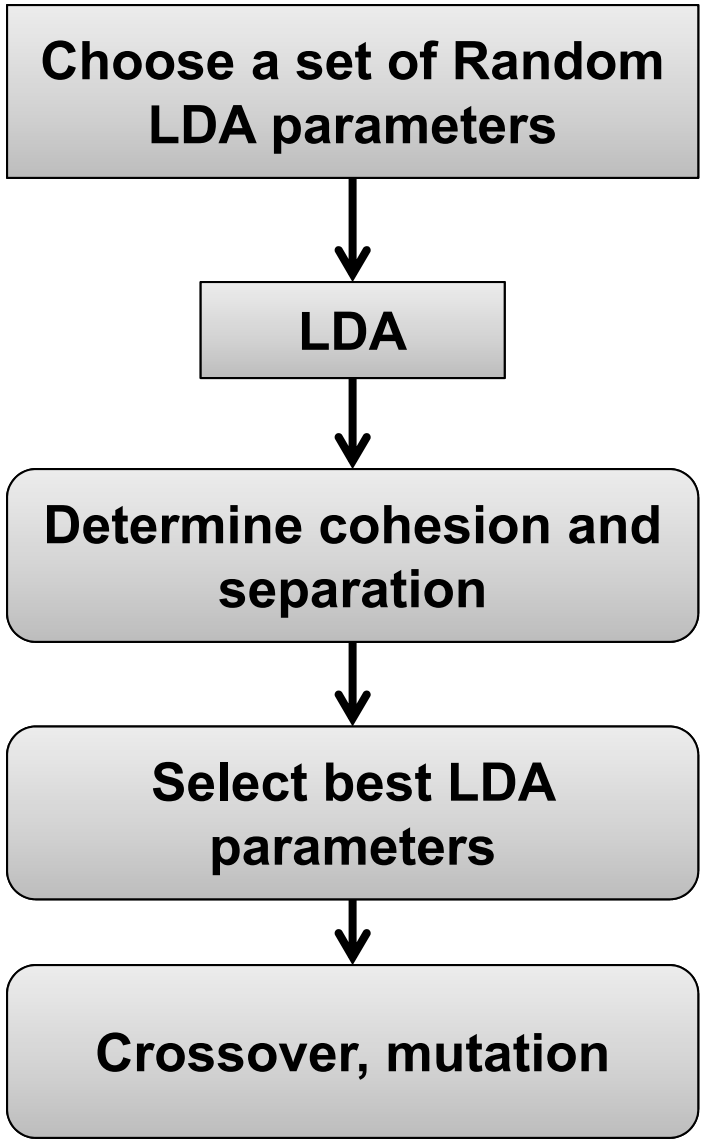


	iteration	topics	α	β	Fitness
LDA Cfg. 1	510	74	0.34	2.5	0.2
LDA Cfg. 2	725	128	1.28	0.4	0.3
…	…	…	…	…	
LDA Cfg. n	618	250	1.14	0.74	0.1



	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	0.3
…	…	…	…	…	
LDA Cfg. n	618	250	1.14	0.74	0.1

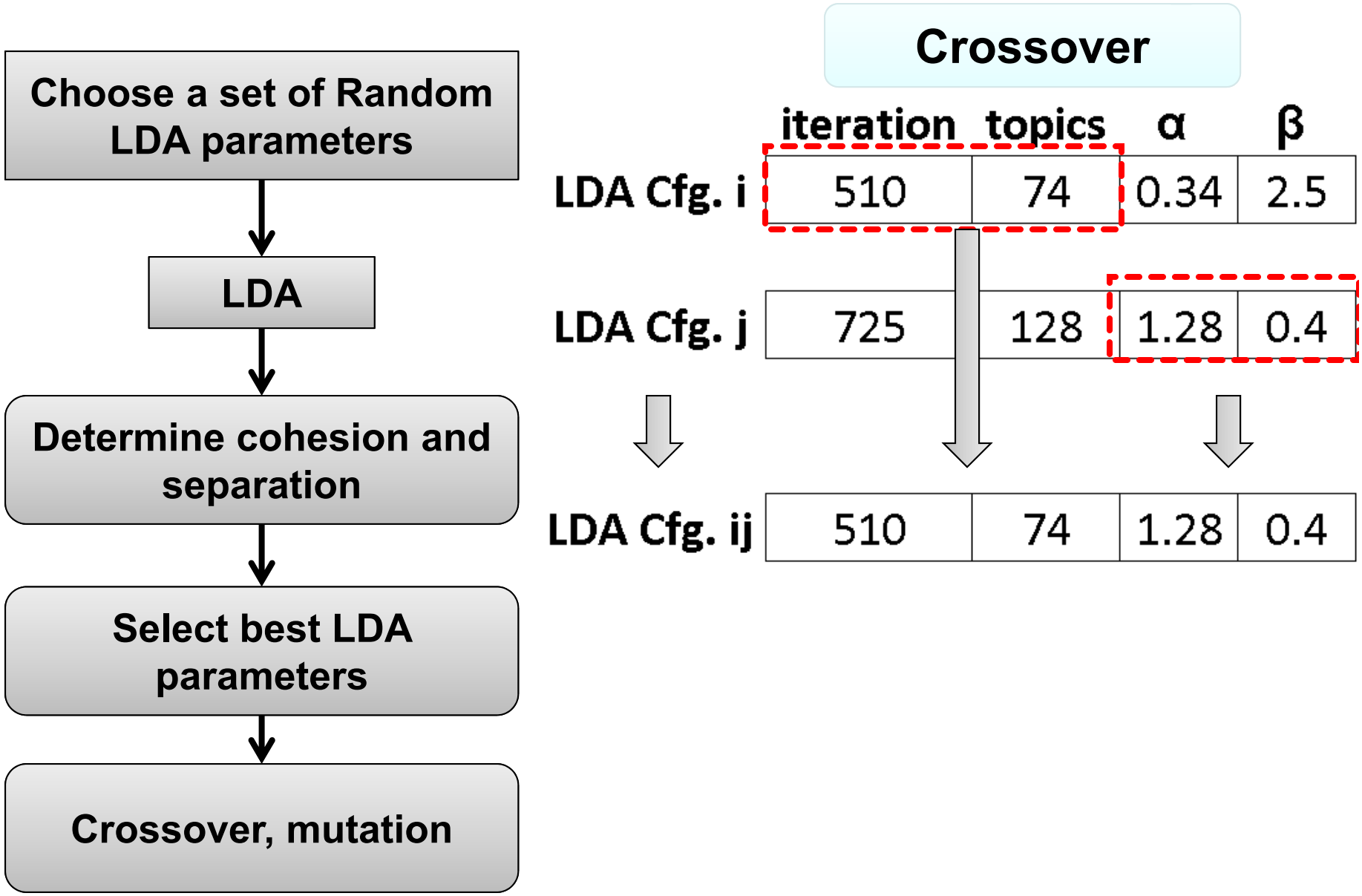

Choose a set of Random LDA parameters
LDA
Determine cohesion and separation
Select best LDA parameters
Crossover, mutation

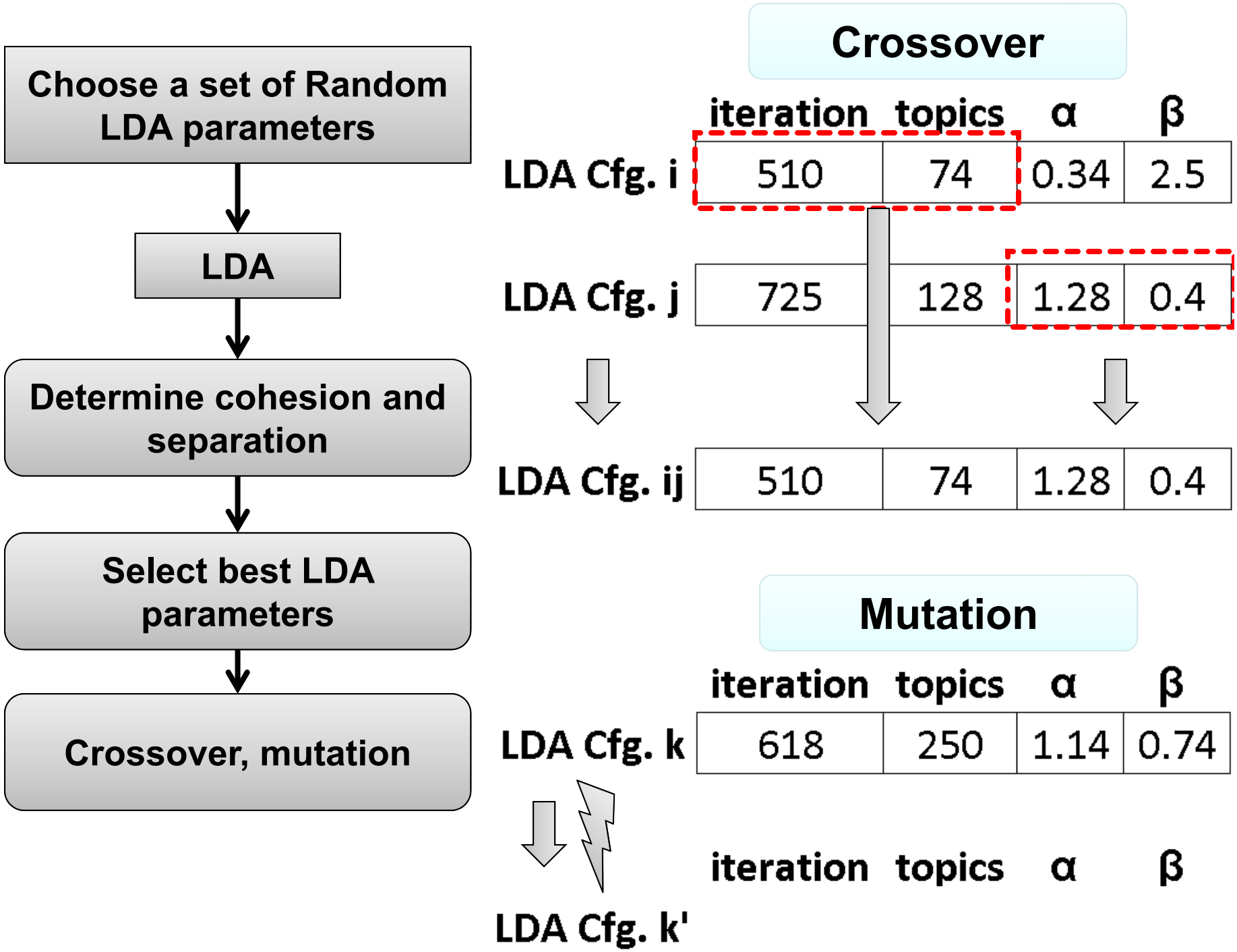


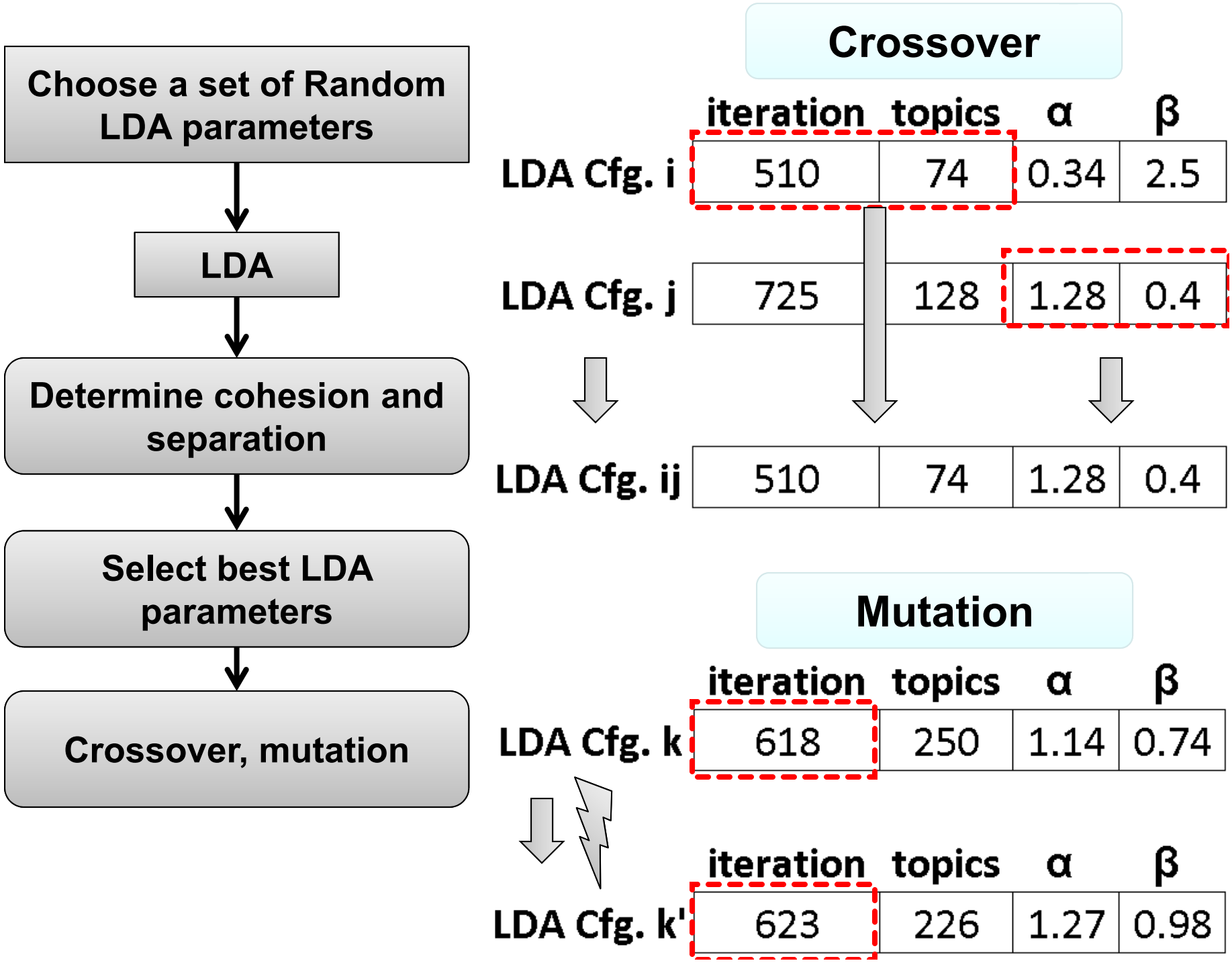
Crossover

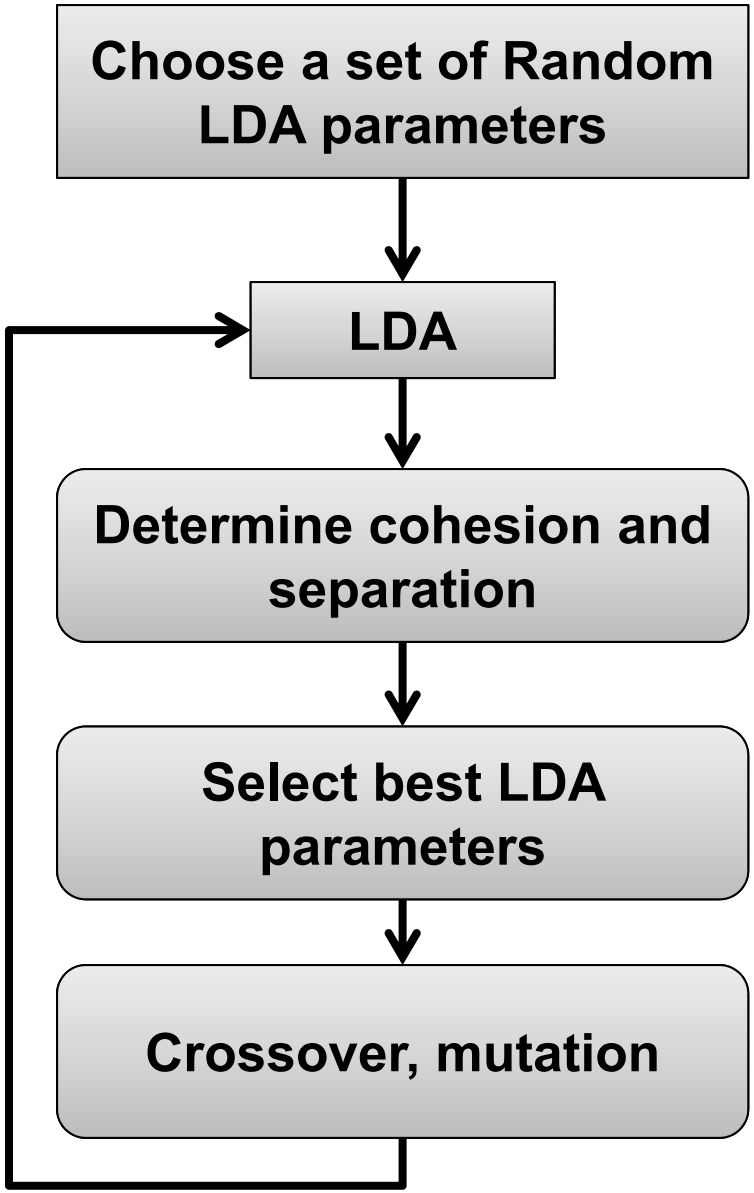
	iteration	topics	α	β
LDA Cfg. i	510	74	0.34	2.5
LDA Cfg. j	725	128	1.28	0.4

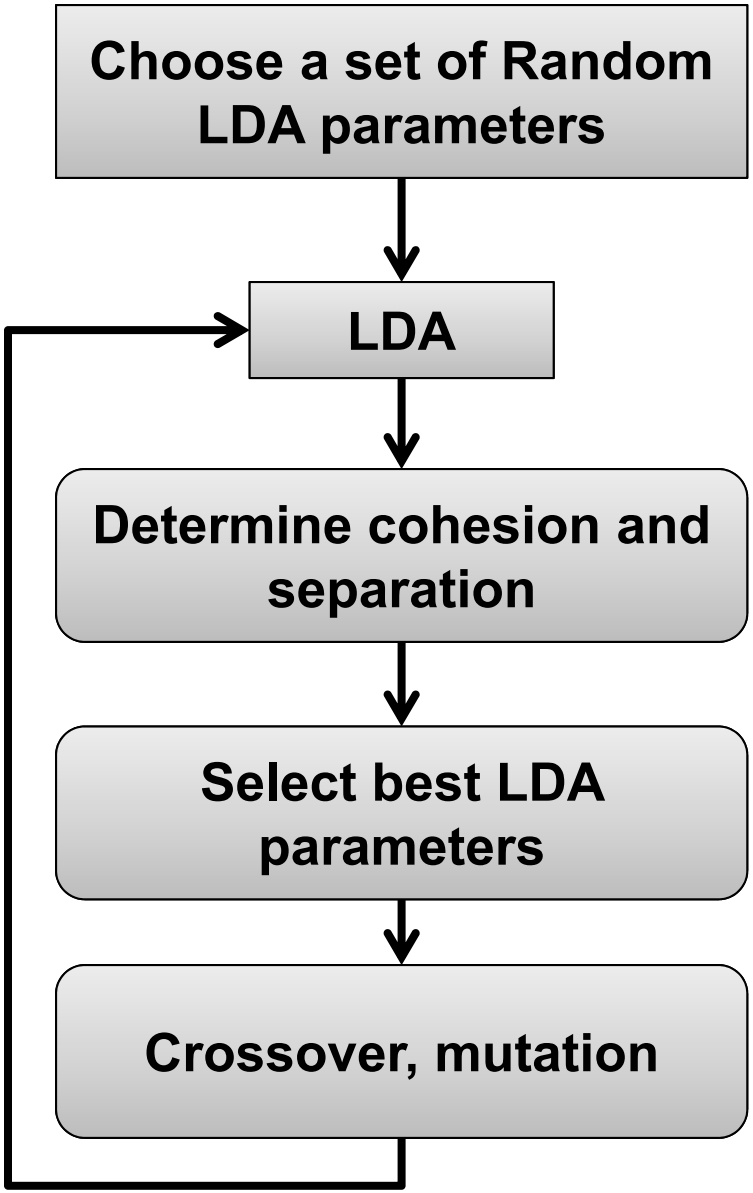
LDA Cfg. ij

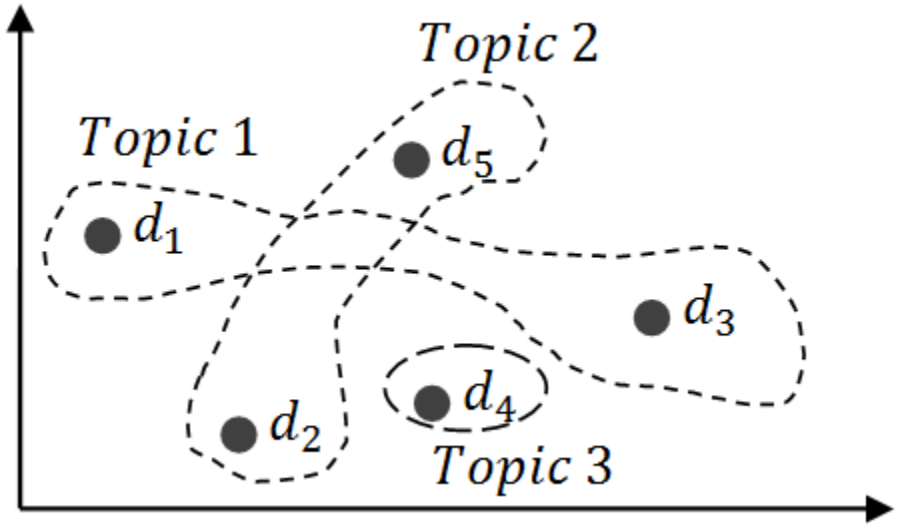

Choose a set of Random LDA parameters
LDA
Determine cohesion and separation
Select best LDA parameters
Crossover, mutation
Crossover
iteration
topics
α
β
LDA Cfg. i
510
74
0.34
2.5
LDA Cfg. j
725
128
1.28
0.4
LDA Cfg. ij
510
74
1.28
0.4

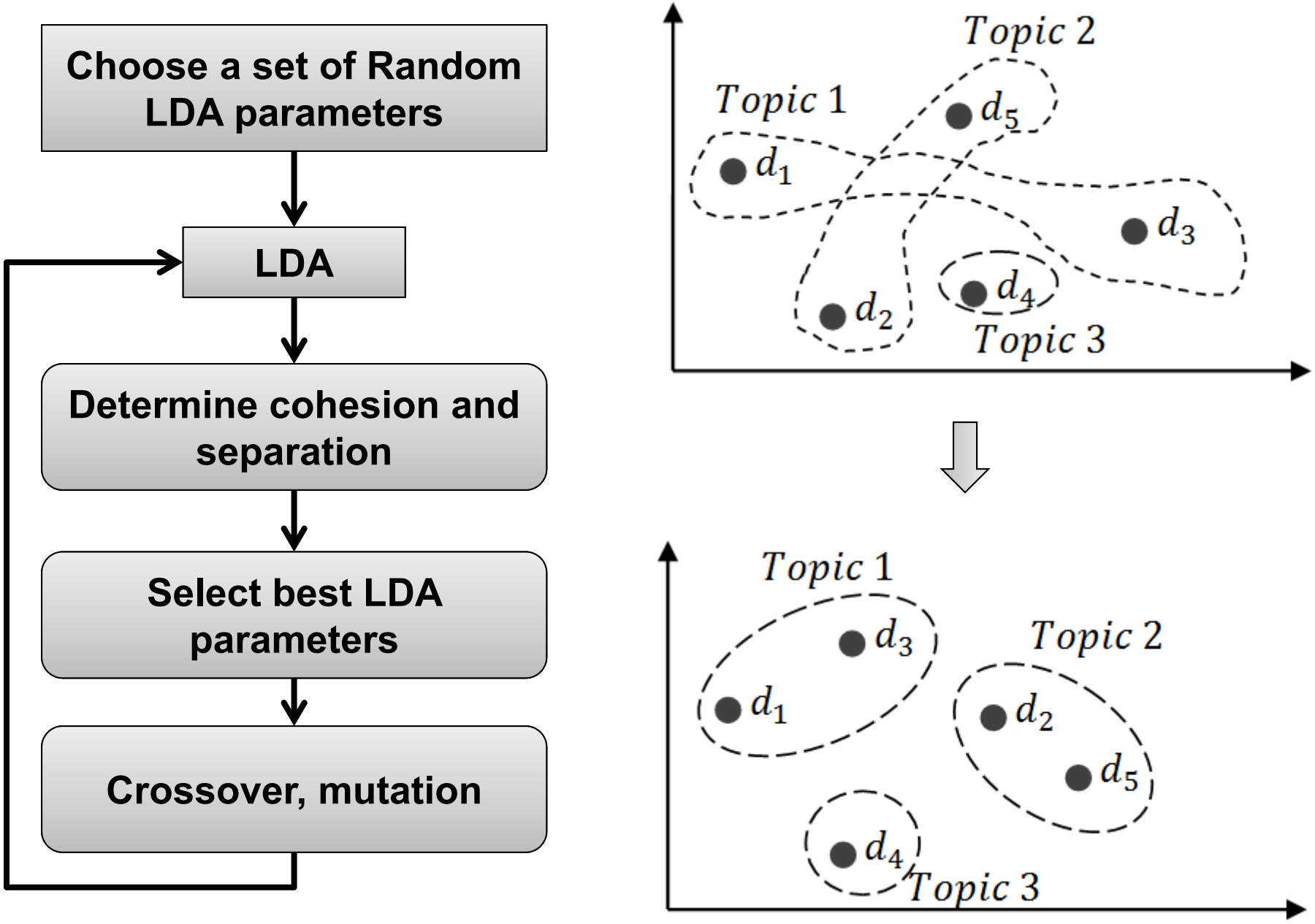

Choose a set of Random LDA parameters
LDA
Determine cohesion and separation
Select best LDA parameters
Crossover, mutation
Crossover
iteration
topics
α
β
LDA Cfg. i
510
74
0.34
2.5
LDA Cfg. j
725
128
1.28
0.4
LDA Cfg. ij
510
74
1.28
0.4
Mutation
iteration
topics
α
β
LDA Cfg. k
618
250
1.14
0.74
iteration
topics
α
β
LDA Cfg. k'

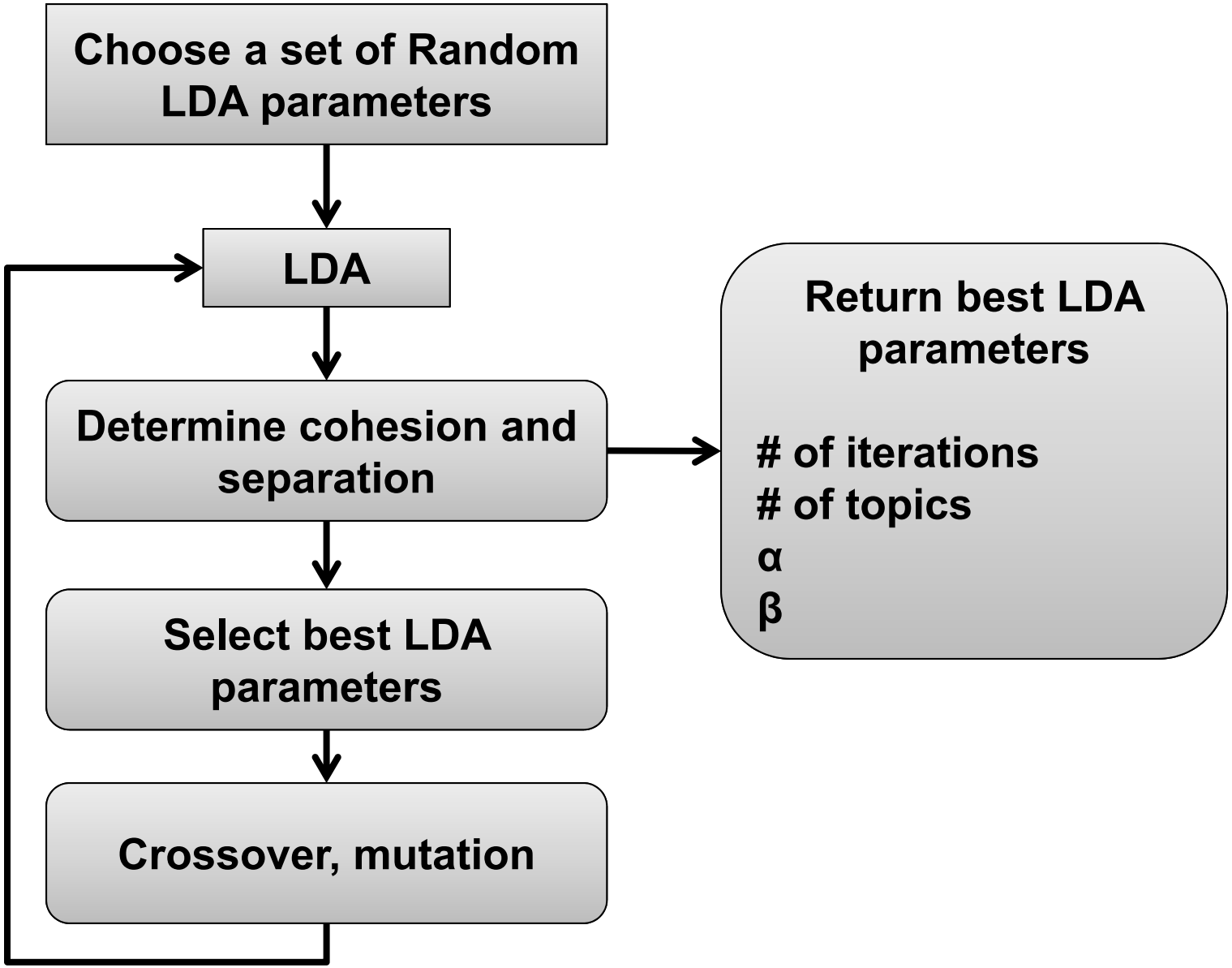

Choose a set of Random LDA parameters
LDA
Determine cohesion and separation
Select best LDA parameters
Crossover, mutation
Crossover
iteration
topics
α
β
LDA Cfg. i
510
74
0.34
2.5
LDA Cfg. j
725
128
1.28
0.4
LDA Cfg. ij
510
74
1.28
0.4
Mutation
iteration
topics
α
β
LDA Cfg. k
618
250
1.14
0.74
iteration
topics
α
β
LDA Cfg. k'
623
226
1.27
0.98


Choose a set of Random LDA parameters
LDA
Determine cohesion and separation
Select best LDA parameters
Crossover, mutation

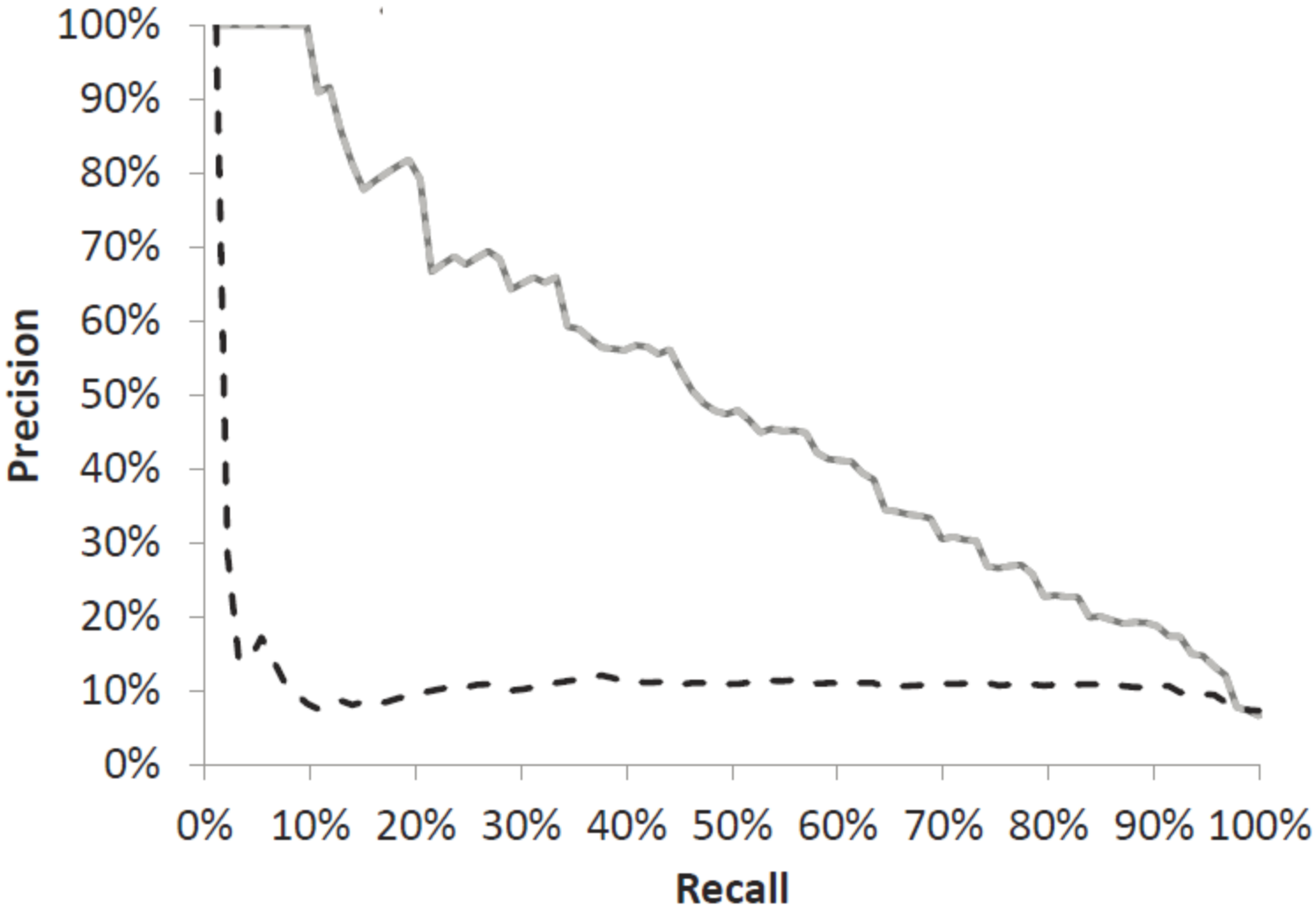

Choose a set of Random LDA parameters
LDA
Determine cohesion and separation
Select best LDA parameters
Crossover, mutation


Topic 2
Topic 1
d_5
d_1
d_3
d_4
d_2
Topic 3

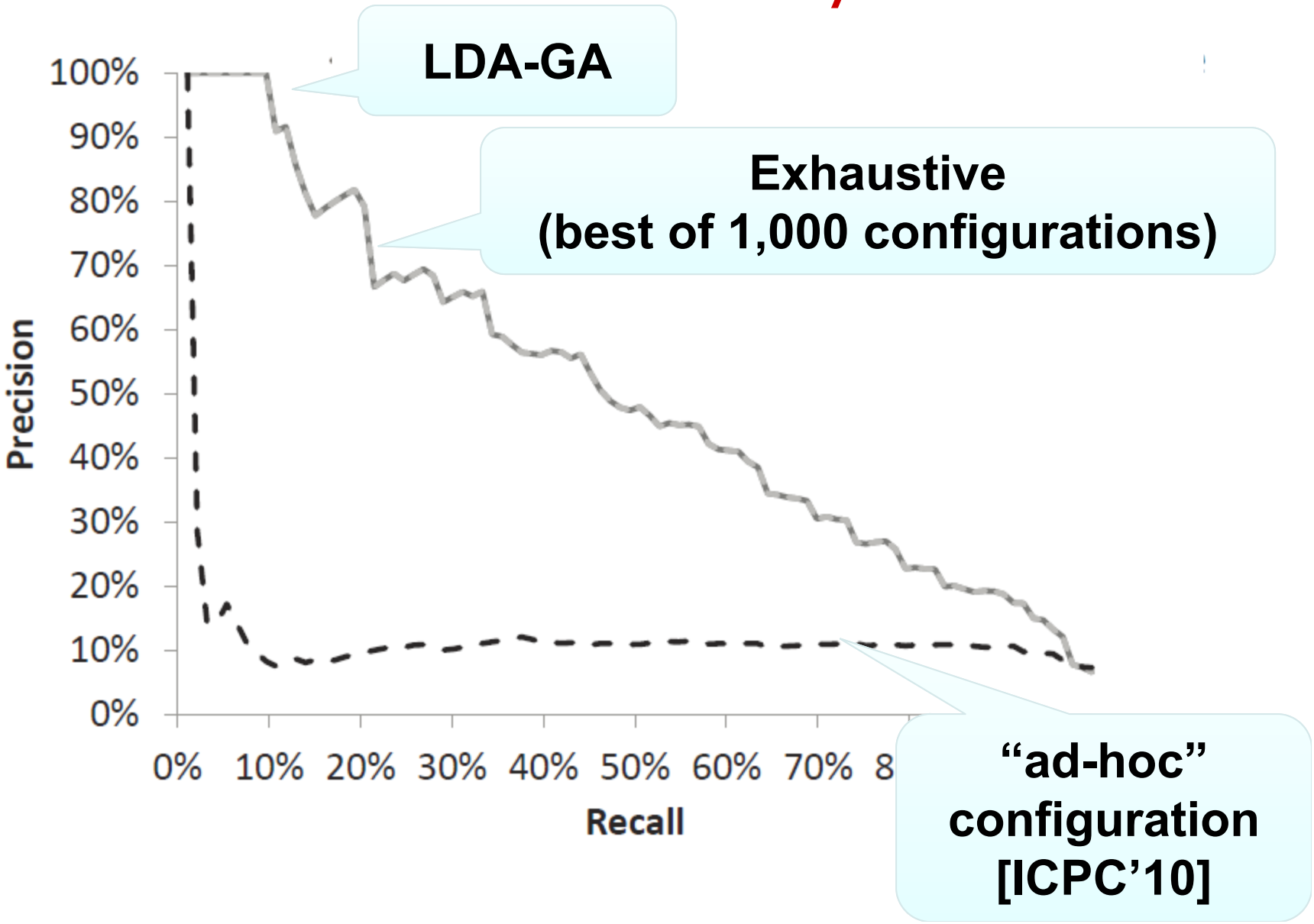

Choose a set of Random LDA parameters
LDA
Determine cohesion and separation
Select best LDA parameters
Crossover, mutation
Topic 1
Topic 2
Topic 3
d_1
d_2
d_3
d_4
d_5
Topic 1
Topic 2
Topic 3
d_1
d_2
d_3
d_4
d_5


Choose a set of Random LDA parameters
LDA
Determine cohesion and separation
Select best LDA parameters
Crossover, mutation
Return best LDA parameters
# of iterations
# of topics
α
β

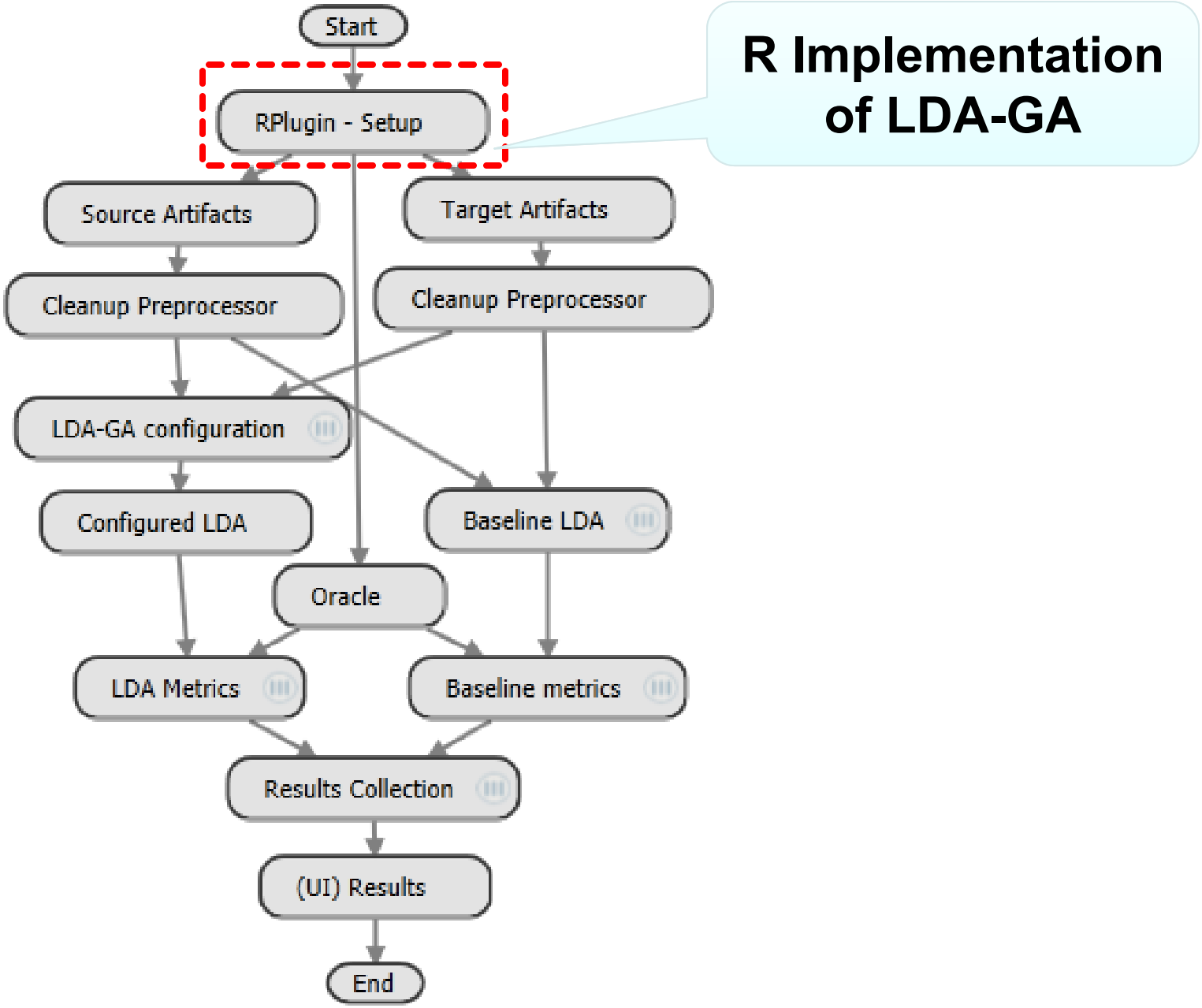
Precision/Recall EasyClinic



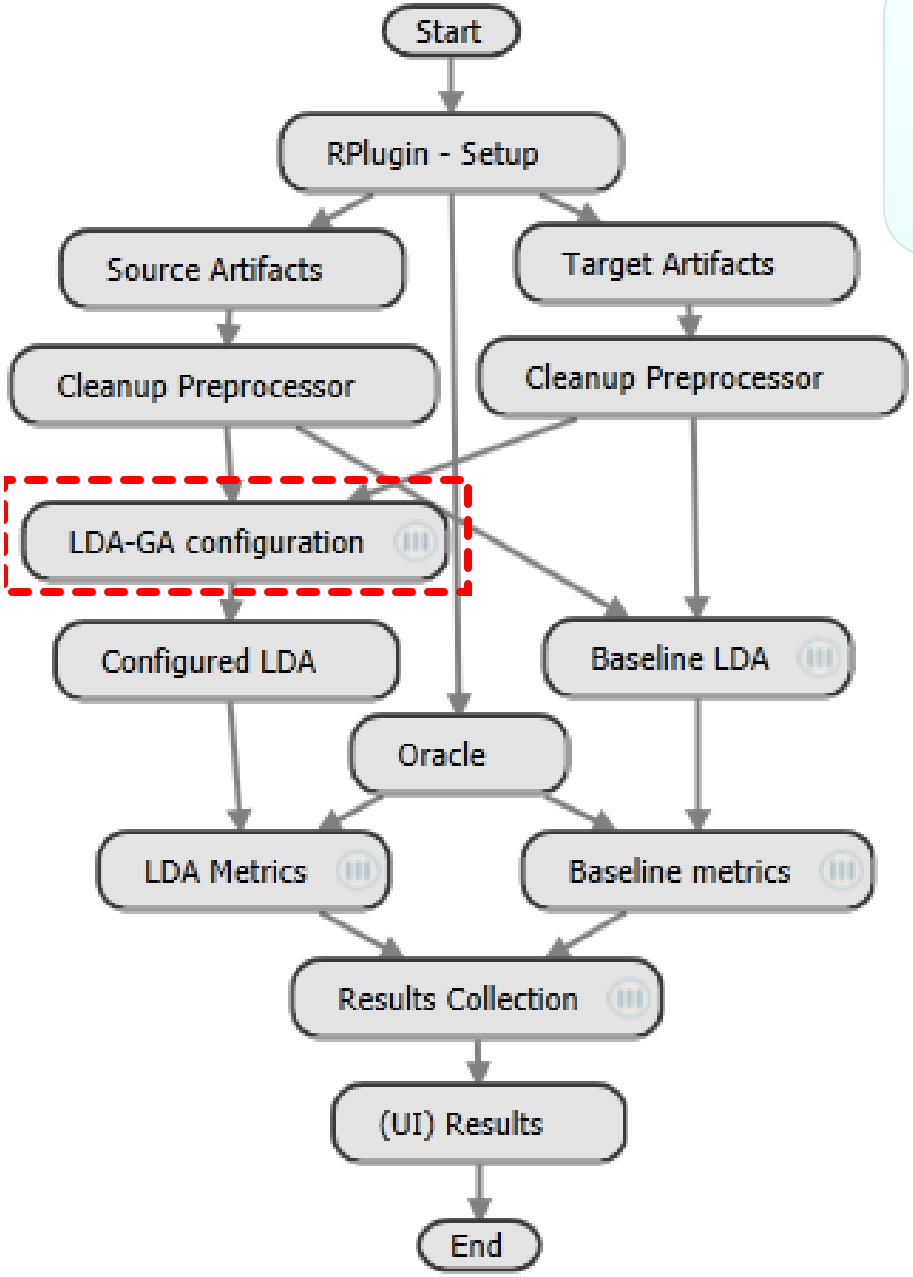
Precision/Recall EasyClinic



LDA-GA in TraceLab

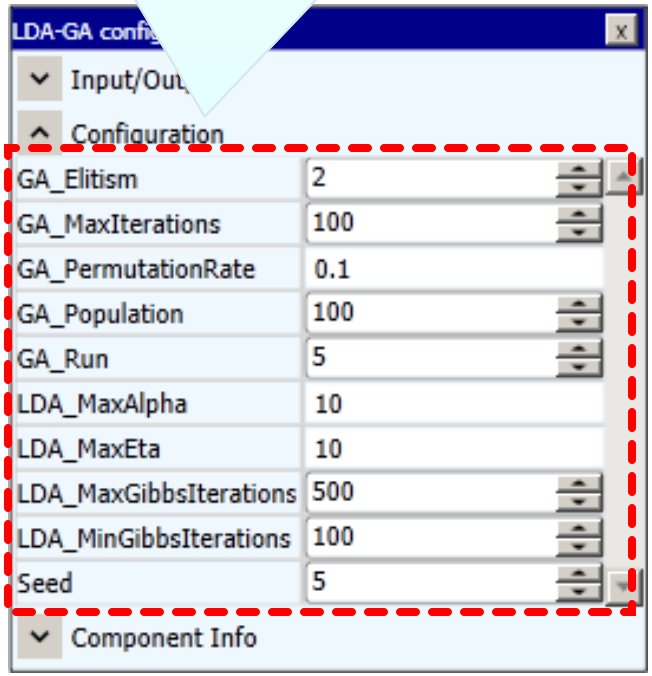


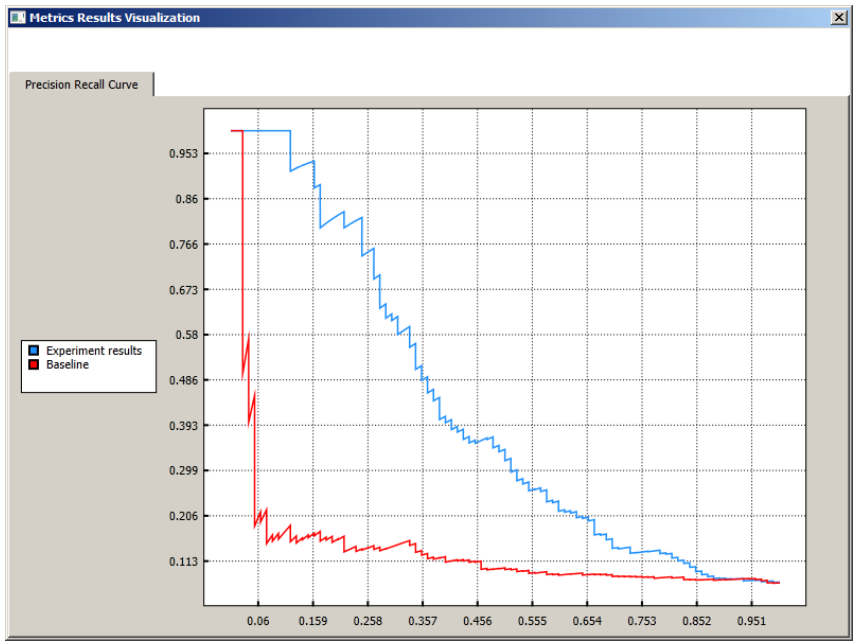
LDA-GA in TraceLab

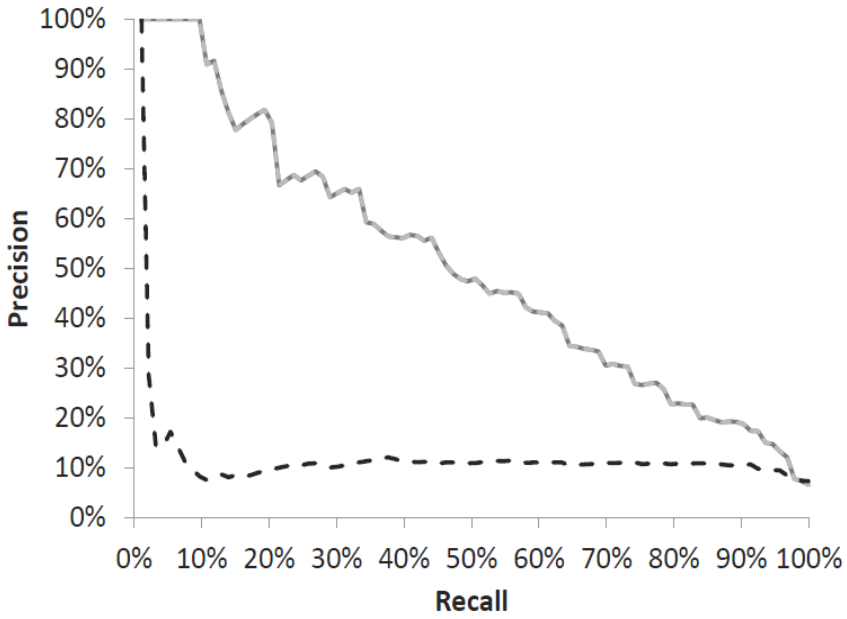


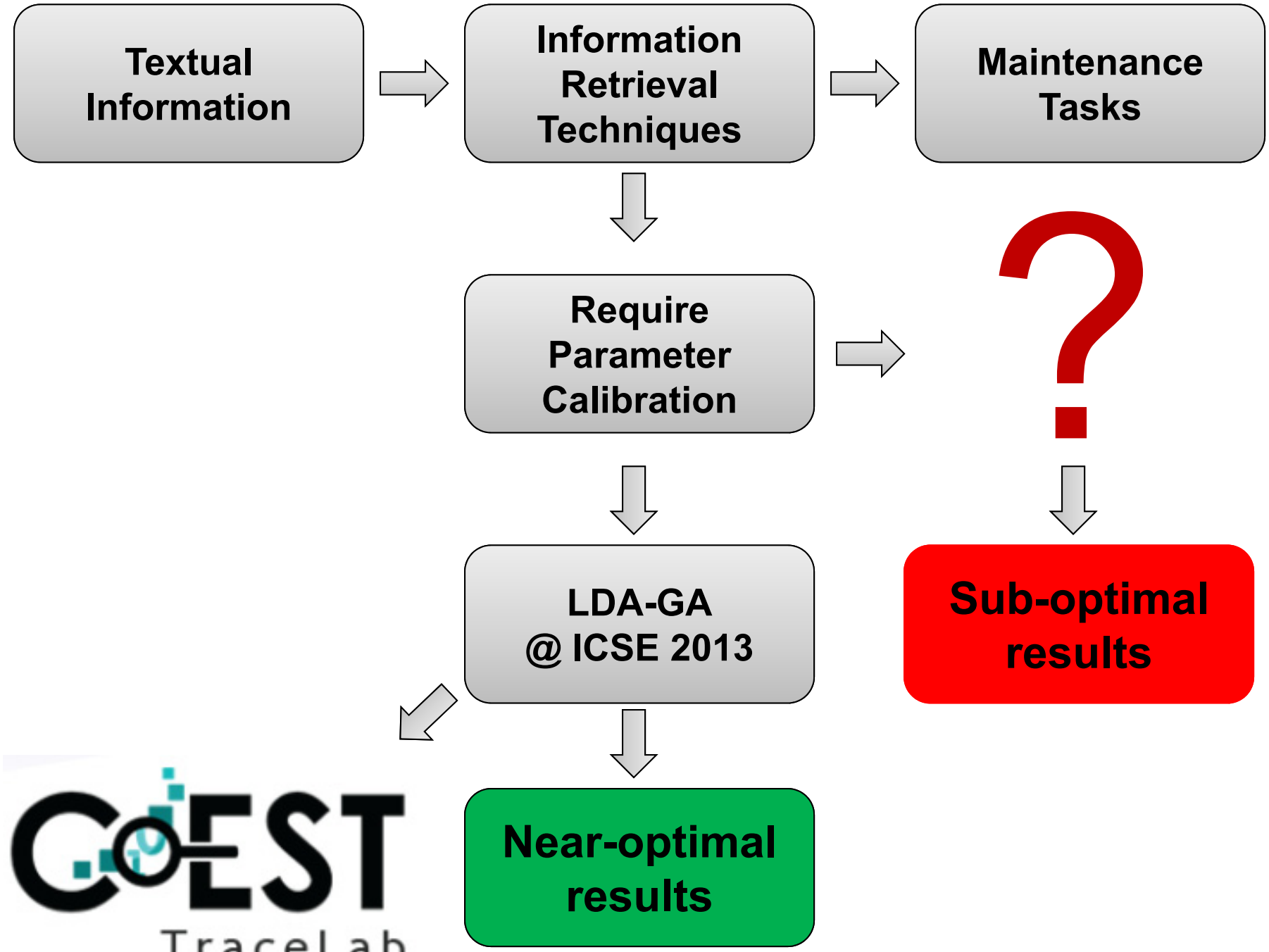
Genetic algorithm settings

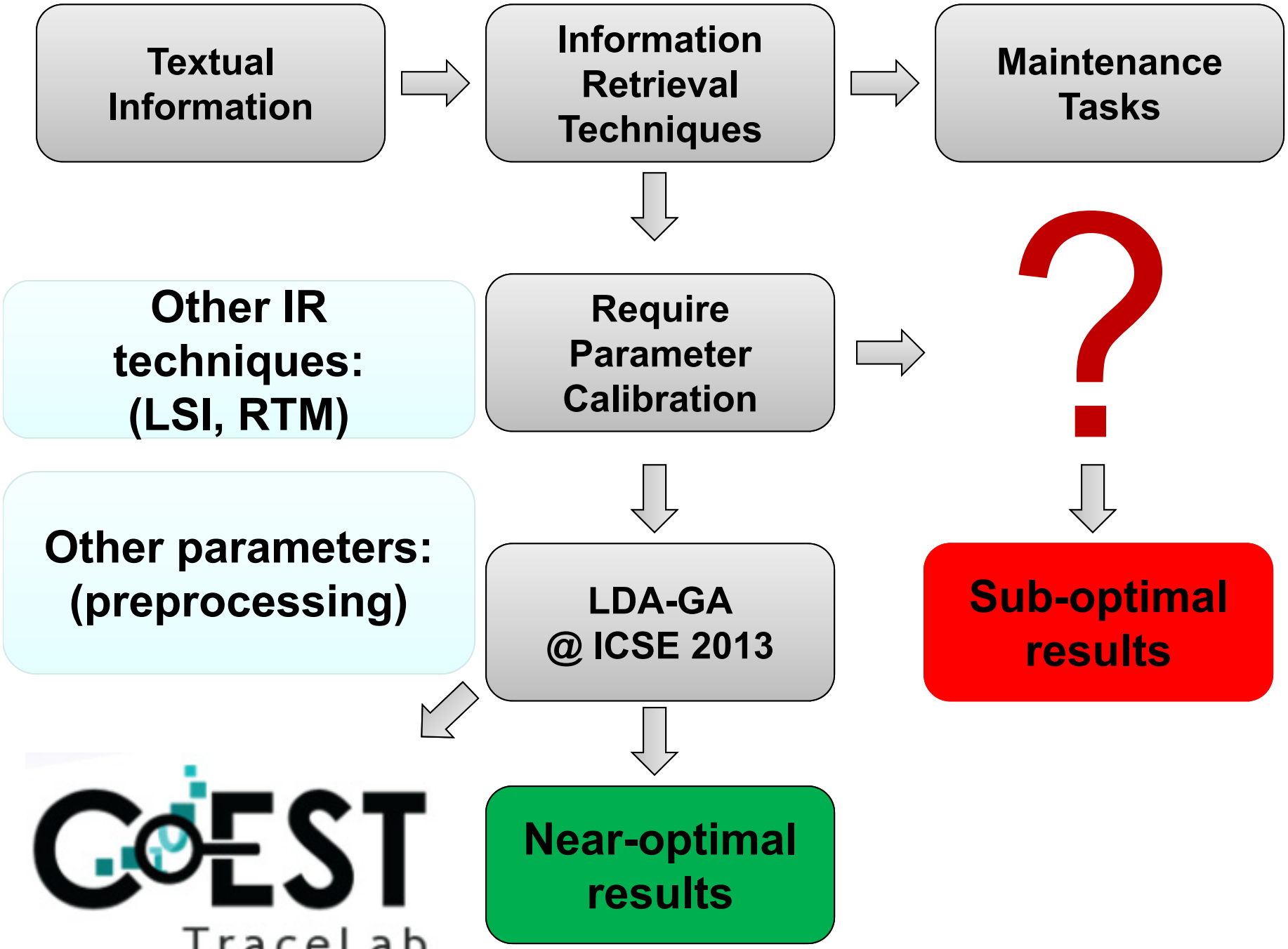
LDA-GA settings




Metrics Results Visualization
Precision Recall Curve
Experiment results
Baseline
0.953
0.86
0.766
0.673
0.58
0.486
0.393
0.299
0.206
0.113
0.06
0.159
0.258
0.357
0.456
0.555
0.654
0.753
0.852
0.951


Precision
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Recall


Textual Information
Information Retrieval Techniques
Maintenance Tasks
Require Parameter Calibration
?
LDA-GA @ ICSE 2013
Sub-optimal results
Near-optimal results
CoEST
TraceLab


Textual Information
Information Retrieval Techniques
Maintenance Tasks
Other IR techniques: (LSI, RTM)
Require Parameter Calibration
?
Other parameters: (preprocessing)
LDA-GA @ ICSE 2013
Sub-optimal results
Near-optimal results
CoEST
TraceLab

Thank you! Questions?

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

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