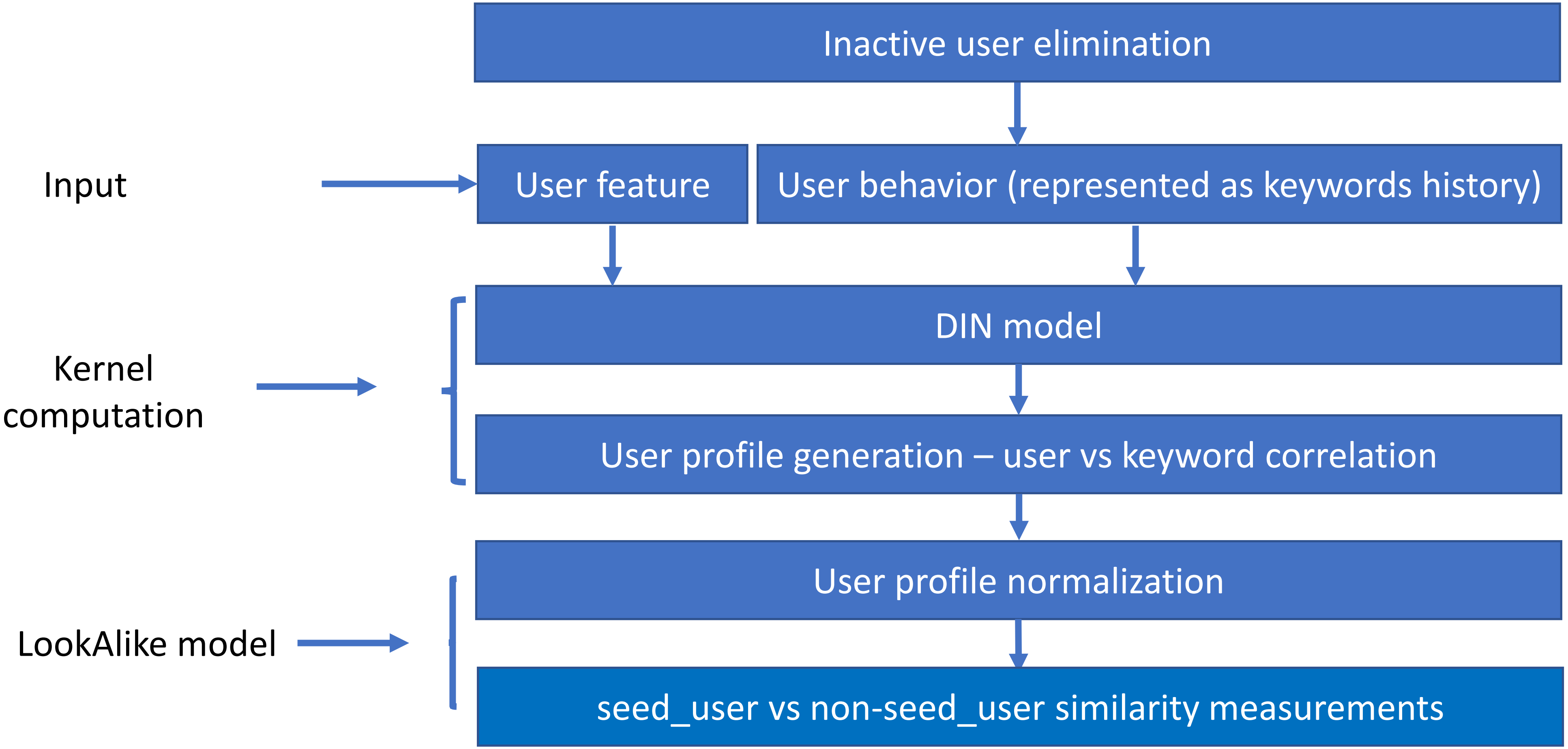
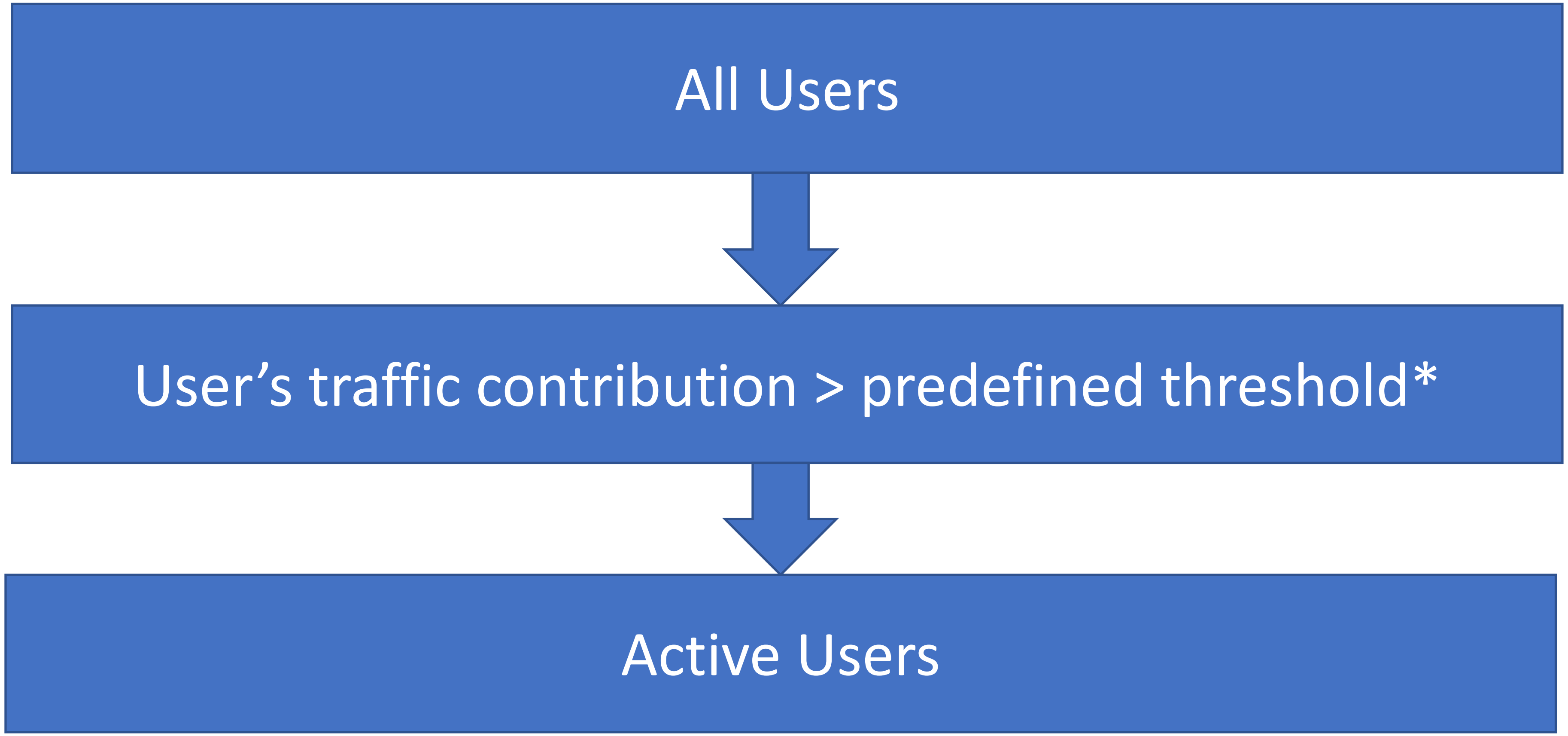
DIN based Look-Alike model

Workflow

- Necessary steps


Inactive user elimination
Input
User feature
User behavior (represented as keywords history)
Kernel computation
DIN model
User profile generation – user vs keyword correlation
User profile normalization
LookAlike model
seed_user vs non-seed_user similarity measurements

Inactive user elimination (user prescreen)



All Users

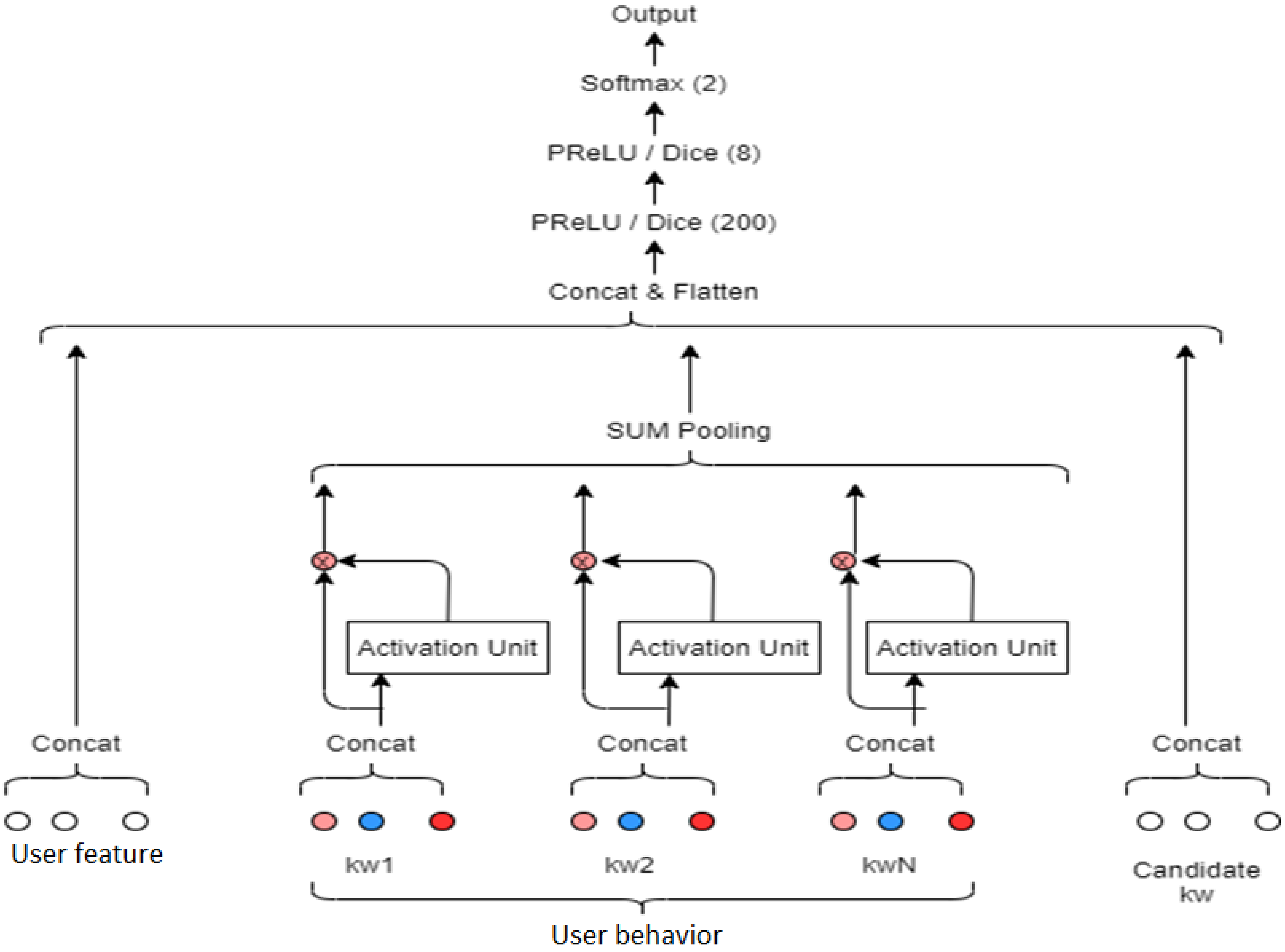
User’s traffic contribution > predefined threshold*

Active Users

* “Prefined threshold” is defined as a range of normal traffic (with low and high bounds) to eliminate:

1. users with consistent low traffic (inactive user, traffic < low bound)
2. users with extremely high traffic for some specific period (robot user, traffic > high bound)

DIN Model


Output
Softmax (2)
PReLU / Dice (8)
PReLU / Dice (200)
Concat & Flatten
SUM Pooling
Activation Unit
Activation Unit
Activation Unit
Concat
Concat
Concat
Concat
Concat
User feature
kw1
kw2
kwN
Candidate
kw
User behavior

DIN Model Output

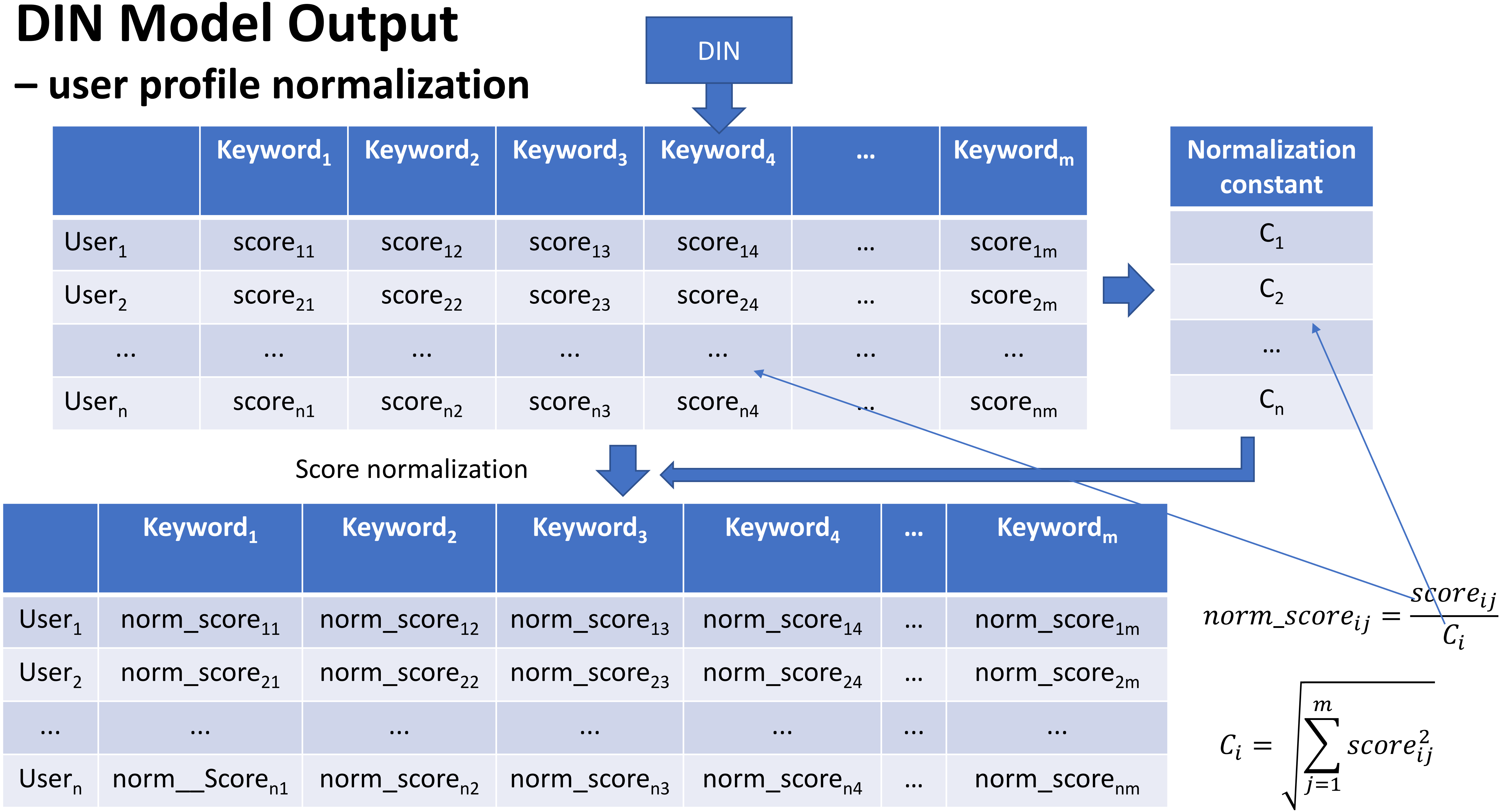
– user vs keyword correlation (user profile generation)

	Keyword$_1$	Keyword$_2$	Keyword$_3$	Keyword$_4$	…	Keyword$_m$
User$_1$	score$_{11}$	score$_{12}$	score$_{13}$	score$_{14}$	…	score$_{1m}$
User$_2$	score$_{21}$	score$_{22}$	score$_{23}$	score$_{24}$	…	score$_{2m}$
…	…	…	…	…	…	…
User$_n$	score$_{n1}$	score$_{n2}$	score$_{n3}$	score$_{n4}$	…	score$_{nm}$

DIN

DIN Model Output

– user profile normalization



	Keyword$_1$	Keyword$_2$	Keyword$_3$	Keyword$_4$	…	Keyword$_m$
User$_1$	score$_{11}$	score$_{12}$	score$_{13}$	score$_{14}$	…	score$_{1m}$
User$_2$	score$_{21}$	score$_{22}$	score$_{23}$	score$_{24}$	…	score$_{2m}$
…	…	…	…	…	…	…
User$_n$	score$_{n1}$	score$_{n2}$	score$_{n3}$	score$_{n4}$	…	score$_{nm}$

Normalization constant
C$_1$
C$_2$
…
C$_n$

Score normalization

	Keyword$_1$	Keyword$_2$	Keyword$_3$	Keyword$_4$	…	Keyword$_m$
User$_1$	norm_score$_{11}$	norm_score$_{12}$	norm_score$_{13}$	norm_score$_{14}$	…	norm_score$_{1m}$
User$_2$	norm_score$_{21}$	norm_score$_{22}$	norm_score$_{23}$	norm_score$_{24}$	…	norm_score$_{2m}$
…	…	…	…	…	…	…
User$_n$	norm__Score$_{n1}$	norm_score$_{n2}$	norm_score$_{n3}$	norm_score$_{n4}$	…	norm_score$_{nm}$

$$norm\_score_{ij} = \frac{score_{ij}}{C_i}$$

$$C_i = \sqrt{\sum_{j=1}^{m} score_{ij}^2}$$

DIN Model Output

– user similarity measurement

$\boldsymbol{User_i's\ normalized\ profile}$:

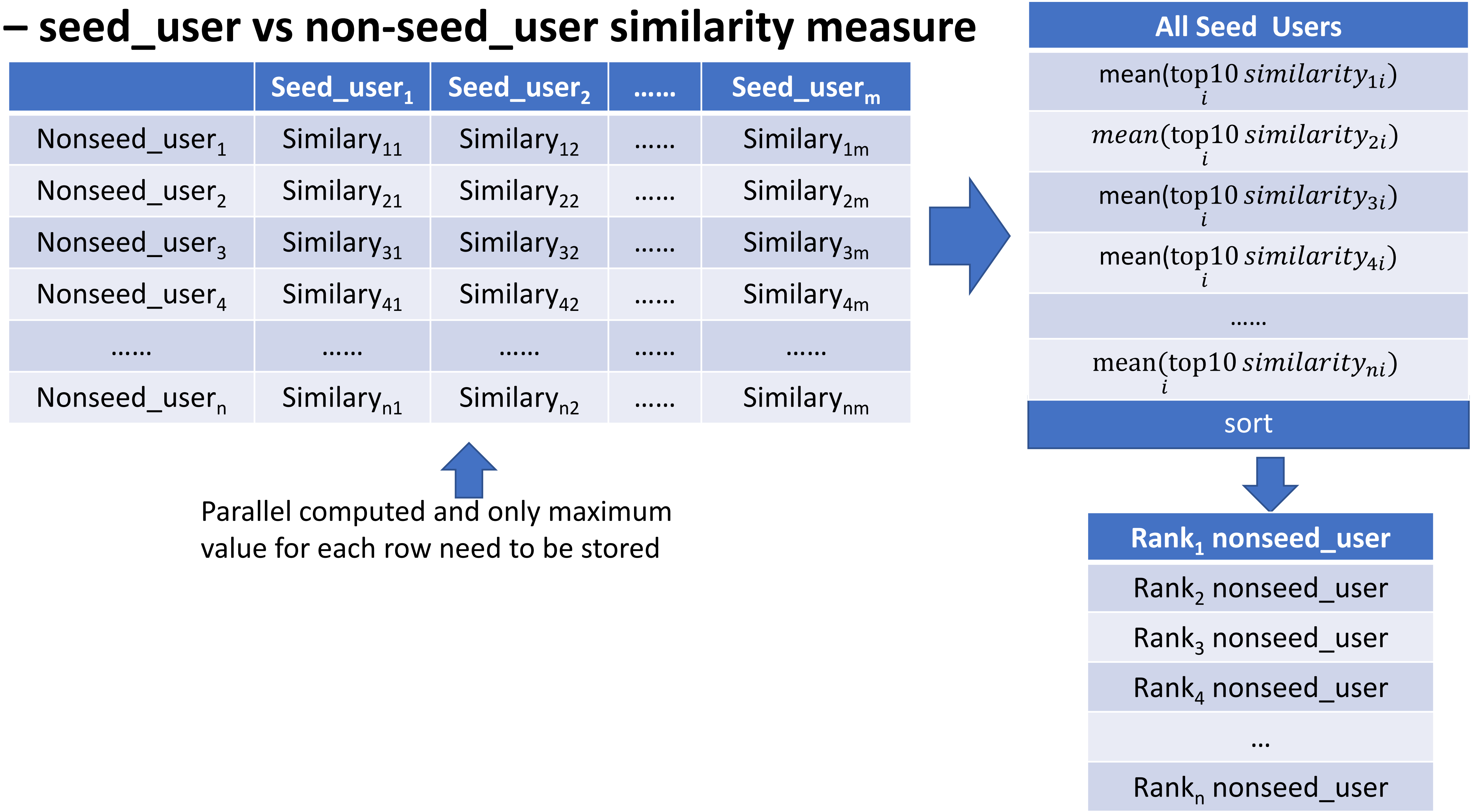
$$S_i = \{norm\_score_{i1}, \quad norm\_score_{i2}, \quad \dots \quad norm\_score_{im}\}$$

$\boldsymbol{Cross\ user\ similarity}$:

$$Similarity(S_i, \quad S_j) = S_i \quad \cdot \quad S_j = \sum_{k=1}^{m} norm\_score_{ik} \times norm\_score_{jk}$$

DIN based Look-Alike model

– seed_user vs non-seed_user similarity measure



	Seed_user$_1$	Seed_user$_2$	……	Seed_user$_m$
Nonseed_user$_1$	Similary$_{11}$	Similary$_{12}$	……	Similary$_{1m}$
Nonseed_user$_2$	Similary$_{21}$	Similary$_{22}$	……	Similary$_{2m}$
Nonseed_user$_3$	Similary$_{31}$	Similary$_{32}$	……	Similary$_{3m}$
Nonseed_user$_4$	Similary$_{41}$	Similary$_{42}$	……	Similary$_{4m}$
……	……	……	……	……
Nonseed_user$_n$	Similary$_{n1}$	Similary$_{n2}$	……	Similary$_{nm}$

Parallel computed and only maximum value for each row need to be stored

All Seed Users
$\text{mean}(\underset{i}{\text{top10}}\, similarity_{1i})$
$mean(\underset{i}{\text{top10}}\, similarity_{2i})$
$\text{mean}(\underset{i}{\text{top10}}\, similarity_{3i})$
$\text{mean}(\underset{i}{\text{top10}}\, similarity_{4i})$
……
$\text{mean}(\underset{i}{\text{top10}}\, similarity_{ni})$
sort

Rank$_1$ nonseed_user
Rank$_2$ nonseed_user
Rank$_3$ nonseed_user
Rank$_4$ nonseed_user
…
Rank$_n$ nonseed_user

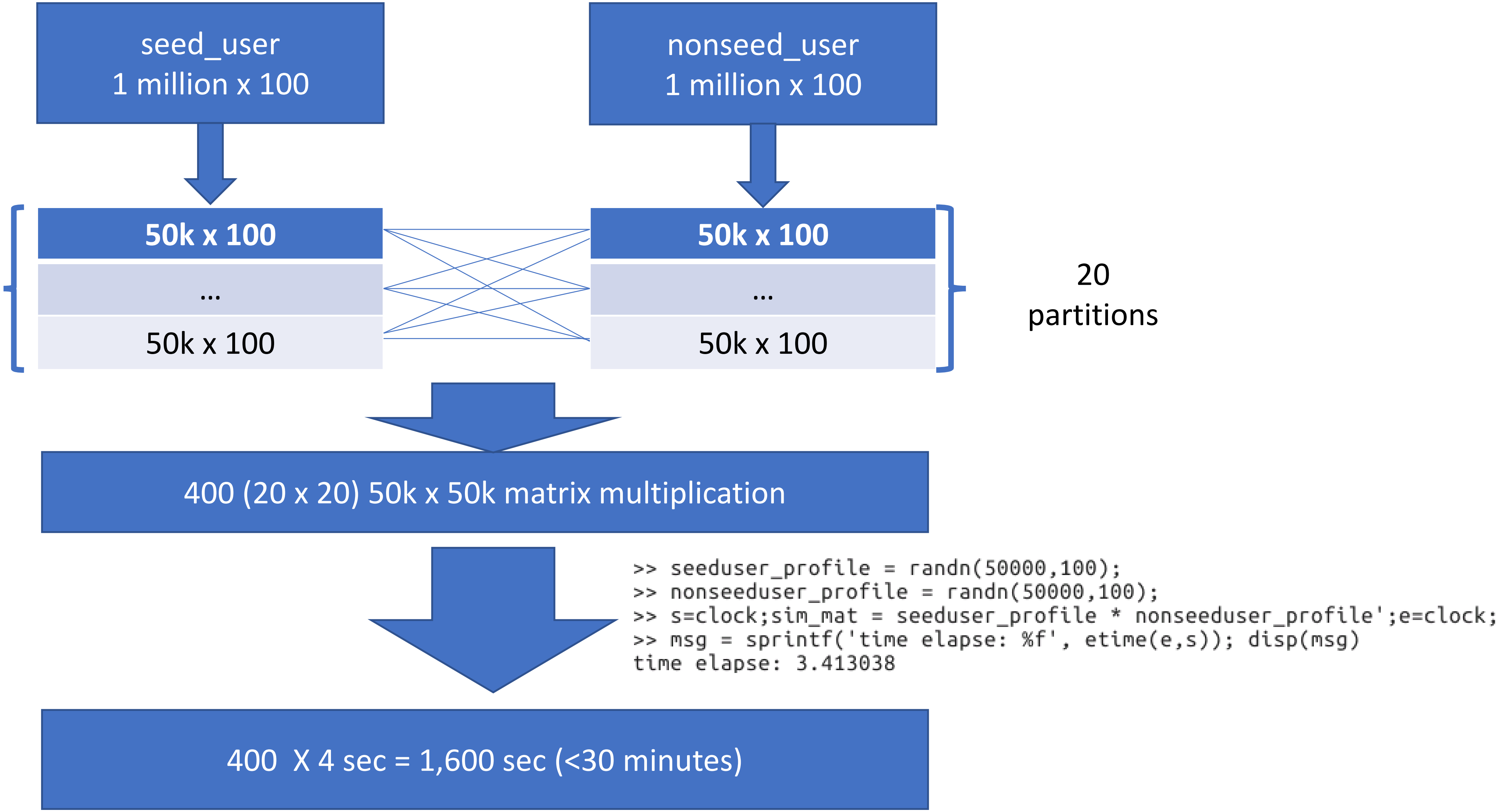
Similarity computation estimation

$$M_{seed} = \begin{bmatrix} norm\_score_{1,1} & \cdots & norm\_score_{1,m} \\ \cdots & \cdots & \cdots \\ norm\_score_{n_{seed},1} & \cdots & norm\_score_{n_{seed},m} \end{bmatrix}$$

$$M_{nonseed} = \begin{bmatrix} norm\_score_{1,1} & \cdots & norm\_score_{1,m} \\ \cdots & \cdots & \cdots \\ norm\_score_{n_{nonseed},1} & \cdots & norm\_score_{n_{nonseed},m} \end{bmatrix}$$

$$M_{similarity} = M_{seed} \quad \times \quad M_{nonseed}^{T}$$

Similarity computation estimation



seed_user
1 million x 100

nonseed_user
1 million x 100

20
partitions

50k x 100

…

50k x 100

50k x 100

…

50k x 100

20
partitions

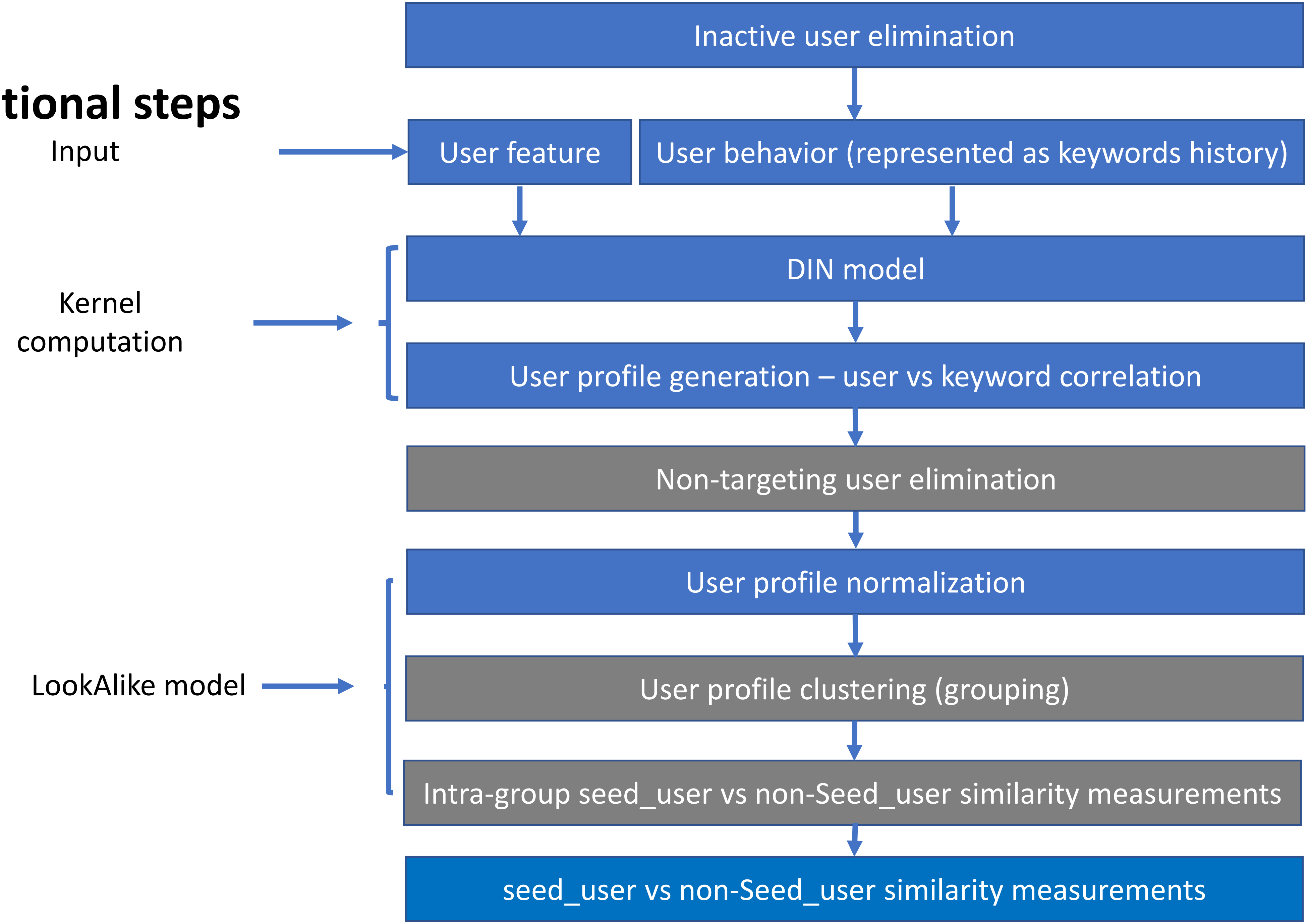
400 (20 x 20) 50k x 50k matrix multiplication

```
>> seeduser_profile = randn(50000,100);
>> nonseeduser_profile = randn(50000,100);
>> s=clock;sim_mat = seeduser_profile * nonseeduser_profile';e=clock;
>> msg = sprintf('time elapse: %f', etime(e,s)); disp(msg)
time elapse: 3.413038
```

400 X 4 sec = 1,600 sec (<30 minutes)

Workflow

- Necessary + optional steps



Inactive user elimination

Input

User feature

User behavior (represented as keywords history)

Kernel computation

DIN model

User profile generation – user vs keyword correlation

Non-targeting user elimination

LookAlike model

User profile normalization

User profile clustering (grouping)

Intra-group seed_user vs non-Seed_user similarity measurements

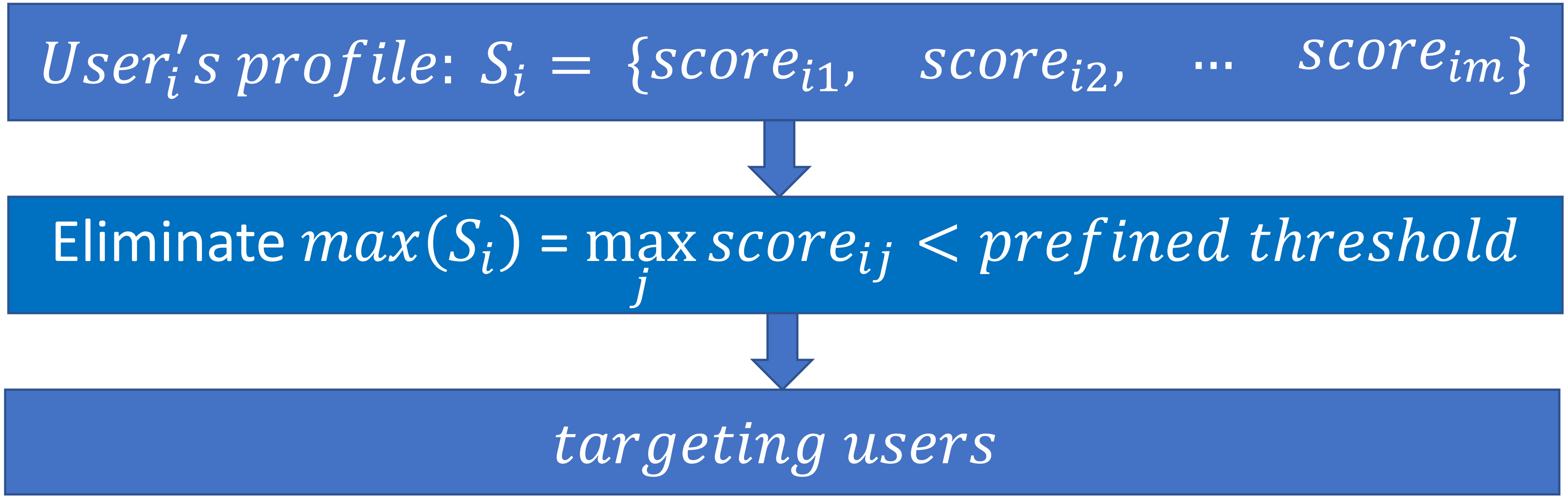
seed_user vs non-Seed_user similarity measurements

DIN Model Output

– non-targeting user elimination

	Keyword$_1$	Keyword$_2$	Keyword$_3$	Keyword$_4$	…	Keyword$_m$
User$_1$	score$_{11}$	score$_{12}$	score$_{13}$	score$_{14}$	…	score$_{1m}$
User$_2$	score$_{21}$	score$_{22}$	score$_{23}$	score$_{24}$	…	score$_{2m}$
…	…	…	…	…	…	…
User$_n$	score$_{n1}$	score$_{n2}$	score$_{n3}$	score$_{n4}$	…	score$_{nm}$

DIN



$User_i's\ profile\colon\ S_i = \{score_{i1},\ \ score_{i2},\ \ \dots\ \ score_{im}\}$

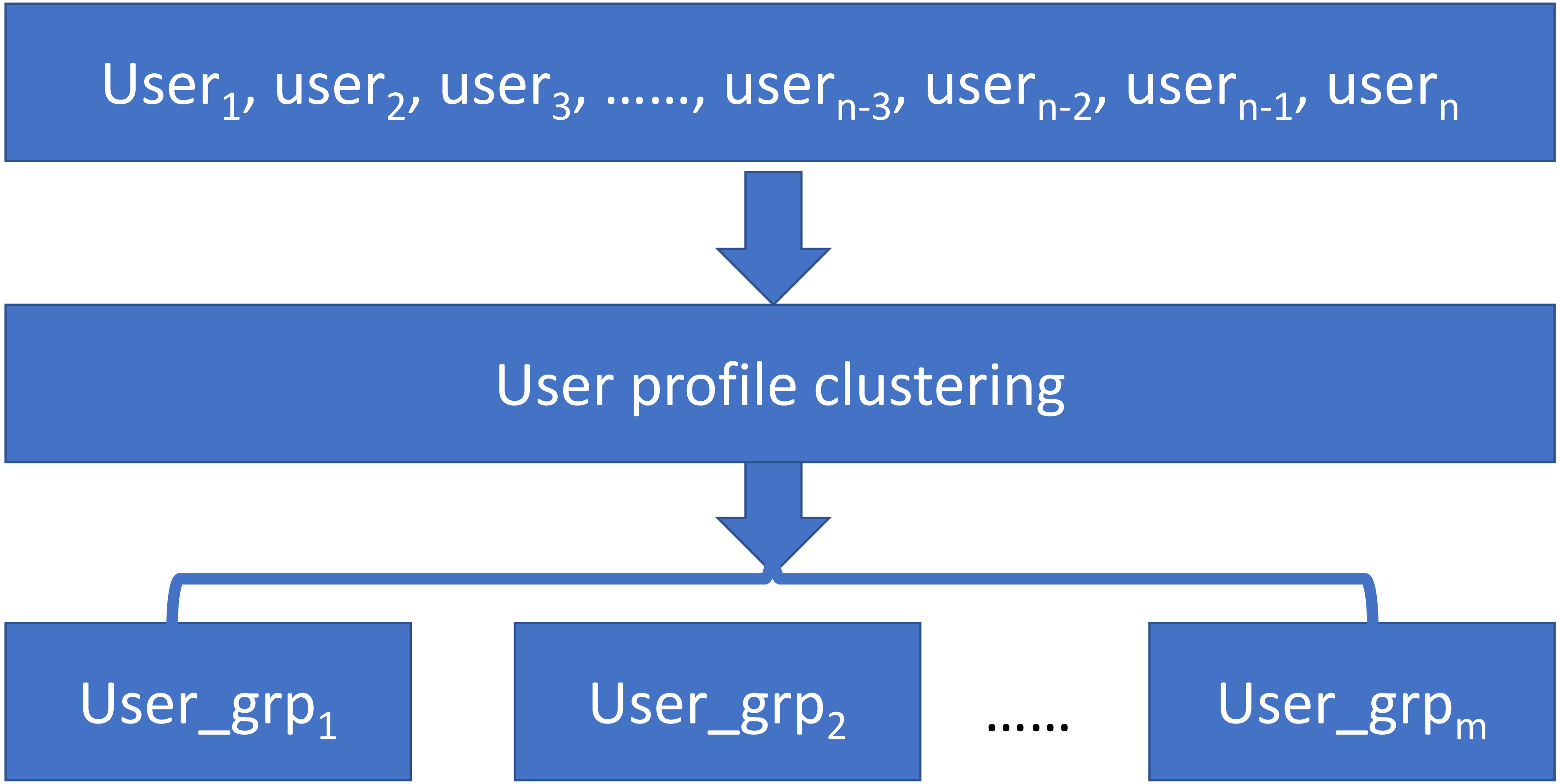
Eliminate $max(S_i) = \max_j score_{ij} < prefined\ threshold$

$targeting\ users$

* The purpose is to eliminate users that have no interest of any keywords (ineffective traffic)

DIN Model Output

– user clustering



User$_1$, user$_2$, user$_3$, ……, user$_{n-3}$, user$_{n-2}$, user$_{n-1}$, user$_n$

User profile clustering

User_grp$_1$

User_grp$_2$

……

User_grp$_m$

DIN based Look-Alike model

– group-wise seed_user vs non-seed_user similarity measure (active user only)

	Seed user in grp$_1$	**Seed user in grp$_2$**	**……**	**Seed user in grp$_m$**
Nonseed user in grp$_1$	similarity matrix$_{11}$	0	……	0
Nonseed user in grp$_2$	0	similarity matrix$_{22}$	……	0
……	……	……	……	0
Nonseed user in grp$_m$	0	0	……	similarity matrix$_{mm}$

DIN based Look-Alike model

– within group seed_user vs non-seed_user similarity measure

Similarity matrix$_{ij}$

	Seed_user$_{grpi,1}$	Seed_user$_{grpi,2}$	……	Seed_user$_{grpi,m}$
Nonseed_user$_{grpi,1}$	Similary$_{11}$	Similary$_{12}$	……	Similary$_{1m}$
Nonseed_userg$_{rpi,2}$	Similary$_{21}$	Similary$_{22}$	……	Similary$_{2m}$
Nonseed_user$_{grpi,3}$	Similary$_{31}$	Similary$_{32}$	……	Similary$_{3m}$
Nonseed_user$_{grpi,4}$	Similary$_{41}$	Similary$_{42}$	……	Similary$_{4m}$
……	……	……	……	……
Nonseed_user$_{grpi,n}$	Similary$_{n1}$	Similary$_{n2}$	……	Similary$_{nm}$

Parallel computed and top 10 values for each row need to be stored

All Seed Users in grpi
$\mathrm{mean}(\underset{i}{\mathrm{top}10}\, similarity_{1i})$
$mean(\underset{i}{\mathrm{top}10}\, similarity_{2i})$
$\mathrm{mean}(\underset{i}{\mathrm{top}10}\, similarity_{3i})$
$\mathrm{mean}(\underset{i}{\mathrm{top}10}\, similarity_{4i})$
……
$\mathrm{mean}(\underset{i}{\mathrm{top}10}\, similarity_{ni})$
sort

Rank$_1$ nonseed_user
Rank$_2$ nonseed_user
Rank$_3$ nonseed_user
Rank$_4$ nonseed_user
…
Rank$_n$ nonseed_user