RecSysChallenge 2016

RecSysChallenge 2016

International competition organized for the 10° RecSys Conference, held in Boston (at MIT) in September 2016

Focus on job recommendation problem:

Given a XING website user, predict those job postings the user will positively interact with in the near future (one week)

Team Pumpkin-Pie

M. Edemanti

R. Pagano

E. Sacchi

T. Carpi

E. Kamberoski

M. Quadrana

Dataset

- 1 500 000 users
- 1 300 000 items
- 8 850 000 interactions
- 10 350 000 rows of impressions

- Make prediction for 150 000 users

Total size of the dataset alone: about 10 Gb!

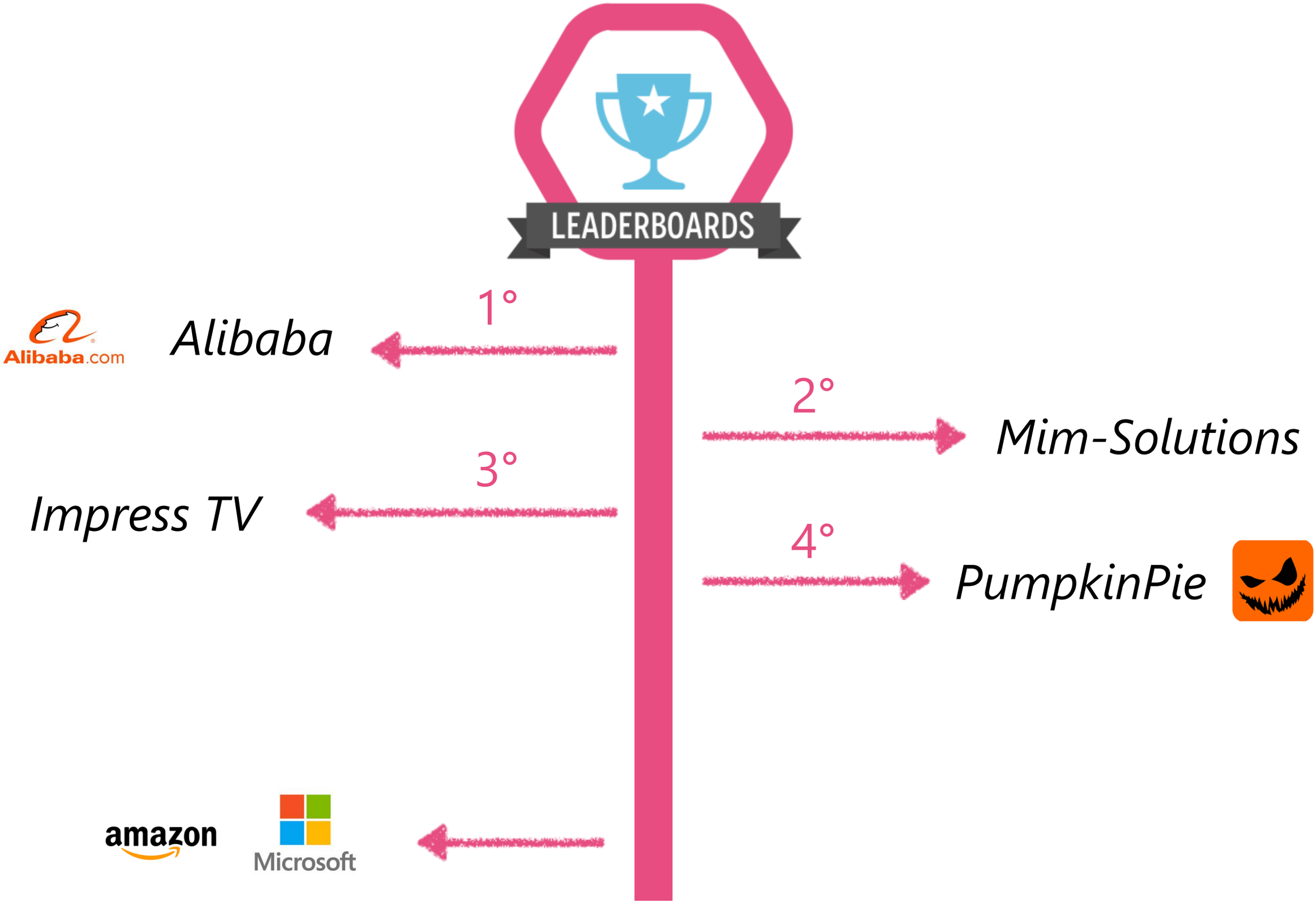
Computational Resources

From PoliCloud:

- 44 cores
- 94 Gb ram
- 760 Gb storage
- 6 VM

From TU Delft:

- 64 cores
- 128 Gb ram
- 100 Gb storage
- 1 VM

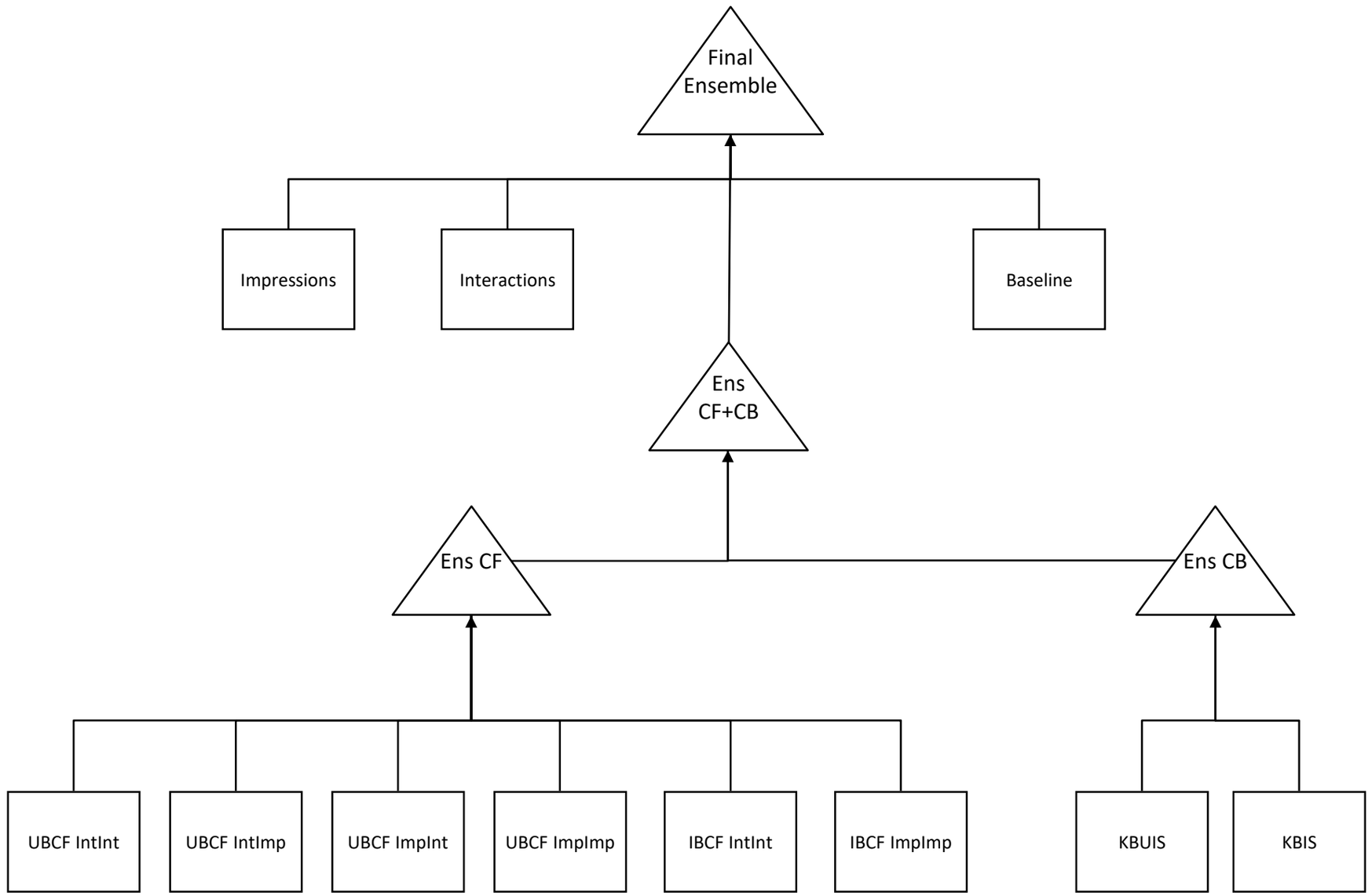

LEADERBOARDS
1°
Alibaba
2°
Mim-Solutions
3°
Impress TV
4°
PumpkinPie
amazon
Microsoft


Linear Ensemble

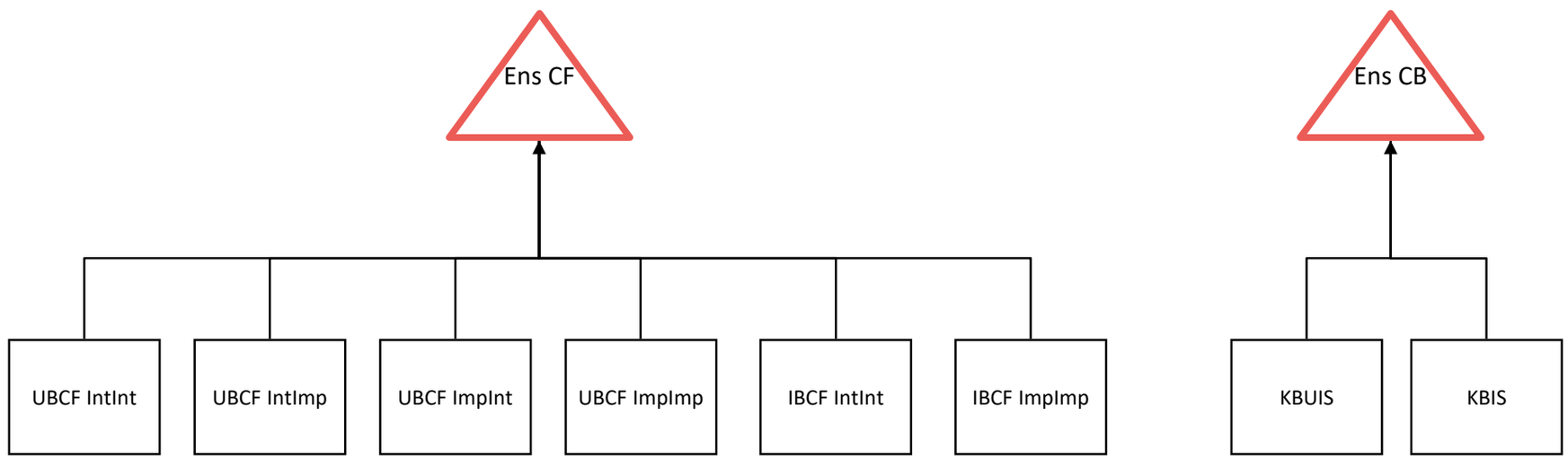

RECSYS / BOSTON / 2016
The 10th ACM Conference on
Recommender Systems
September 15 - 19, 2016
DIAMOND SUPPORTERS
Google
Spotify
Alibaba Group
PLATINUM SUPPORTERS
NETFLIX
PANDORA
Quora
OpenTable
GOLD SUPPORTERS

The multi-stack ensemble

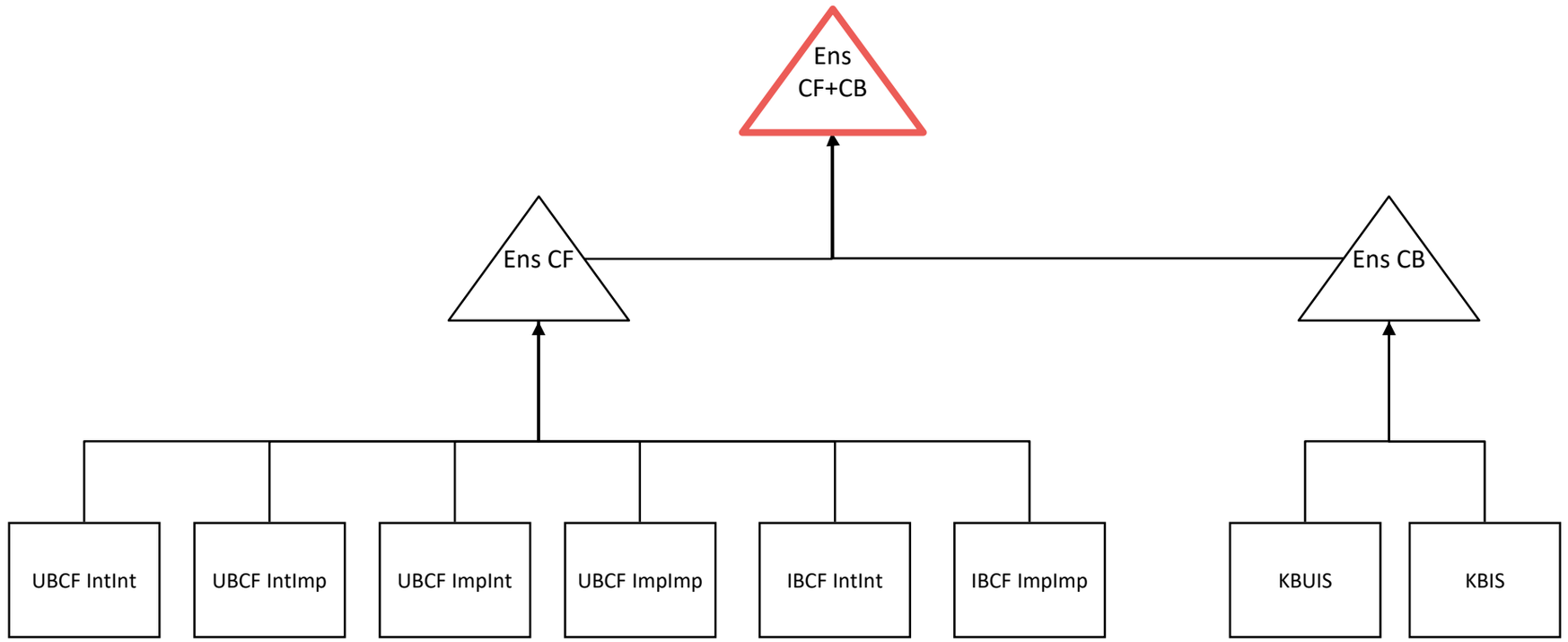
Multi-Stack Ensemble



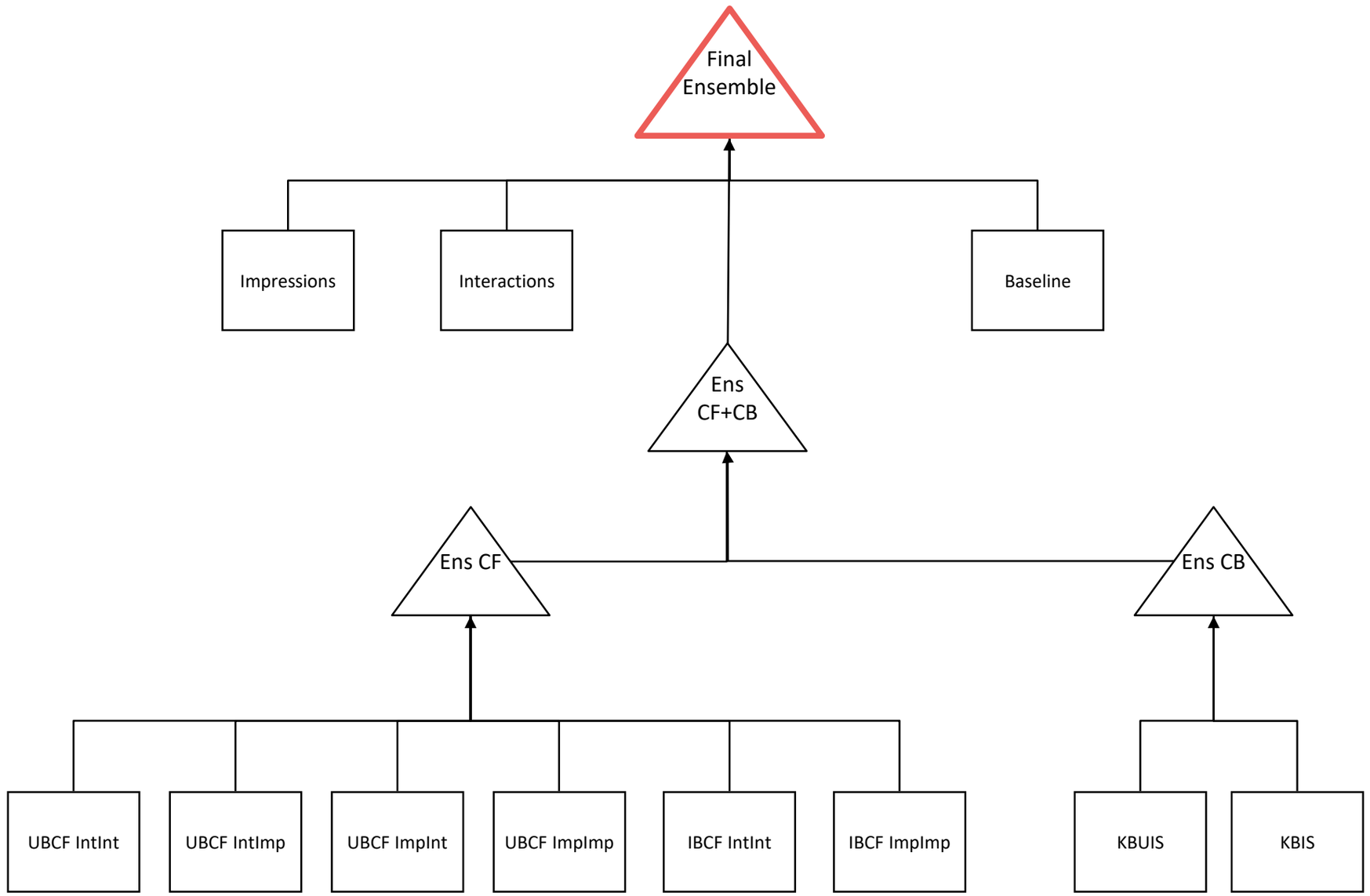
Multi-Stack Ensemble



Multi-Stack Ensemble



Multi-Stack Ensemble



2-Step Algorithm

Voting-Based Technique

+

Reduce Function

Linear Ensemble

$$s_{aui} = w_a - \text{rank}_{aui} \cdot d_a$$

w_a — *weight* for algorithm *a*

rank_{aui} — *rank* of item *i* in algorithm *a*

d_a — *decay* for algorithm *a*

Linear Ensemble

Algorithm A			*Algorithm B*		
Weight		*2*	*Weight*		*2*
Decay		*0.001*	*Decay*		*0.0015*
		red 3.9930			
1	*red*	*1.9990*	*1*	*green*	*1.9985*
2	*pink*	*1.9980*	*2*	*orange*	*1.9970*
3	*yellow*	*1.9970*	*3*	*purple*	*1.9955*
4	*blue*	*1.9960*	*4*	*red*	*1.9940*

Evaluation-Score Ensemble

local test score

$$w_a = \frac{l_a}{n_a}$$

# of elements

$$s_{aui} = w_a \cdot e(\text{rank}_{aui})$$

$$e(\text{rank}_{aui}) = \begin{cases} 37.83, & \text{rank}_{aui} \in [1, 2] \\ 27.83, & \text{rank}_{aui} \in [3, 4] \\ 22.83, & \text{rank}_{aui} \in [5, 6] \\ 21.17, & \text{rank}_{aui} \in [7, 20] \\ 20.67, & \text{rank}_{aui} \in [21, N] \end{cases}$$

Evaluation-Score Ensemble

Algorithm A		*Algorithm B*	
l_a	*200k*	*l_b*	*200k*
n_a	*1 Mln*	*n_b*	*800k*
Weight	*0.2*	*Weight*	*0.25*

red	*7.5660*	yellow	*9.4575*
pink	*7.5660*	green	*9.4575*
purple	*5.5660*	blue	*6.9575*
orange	*5.5660*	grey	*6.9575*

Algorithms

Collaborative-Filtering

IDF value as a *rate* for the job

$$r_{ui} = \begin{cases} 0 \text{ if user } u \text{ did not interact with item } i \\ \log\left(\frac{\text{total \# of interactions}}{\text{\# of interactions with job } i}\right) \quad \text{otherwise} \end{cases}$$

Collaborative-Filtering

User-based e Item-based

$$sim\_user(u,w) = \frac{\sum_i r_{ui} r_{wi}}{\sqrt{\sum_i r_{ui}^2}\sqrt{\sum_i r_{wi}^2} + \beta}$$

$$sim\_item(i,j) = \frac{\sum_{u \in \mathcal{U}} r_{ui} r_{uj}}{\sqrt{\sum_{u \in \mathcal{U}} r_{ui}^2}\sqrt{\sum_{u \in \mathcal{U}} r_{uj}^2} + \beta}$$

Content-Based

Concept-based joint *User-Item* similarity

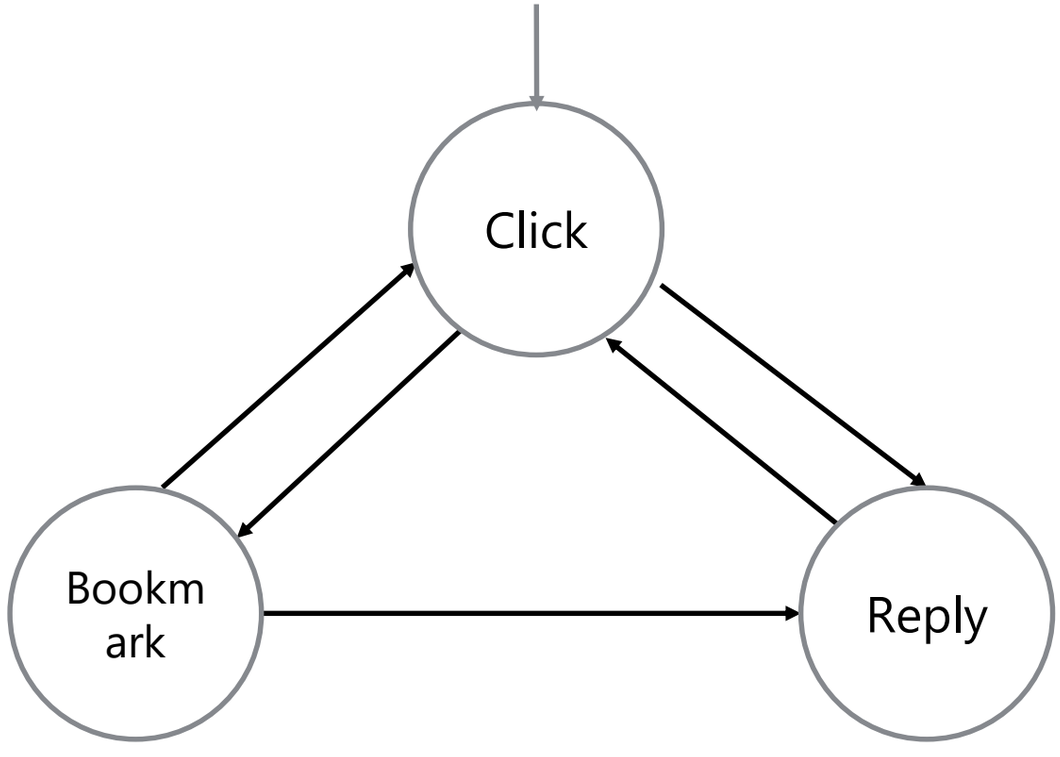
$$\mathrm{UF}(c,u) = \mathrm{TF}_U(c,u) \cdot \mathrm{IDF}_U(c)$$

$$\mathrm{TF}_U(c,u) = \frac{\sum_i b_{uic}}{\sum_i b_{ui}}$$

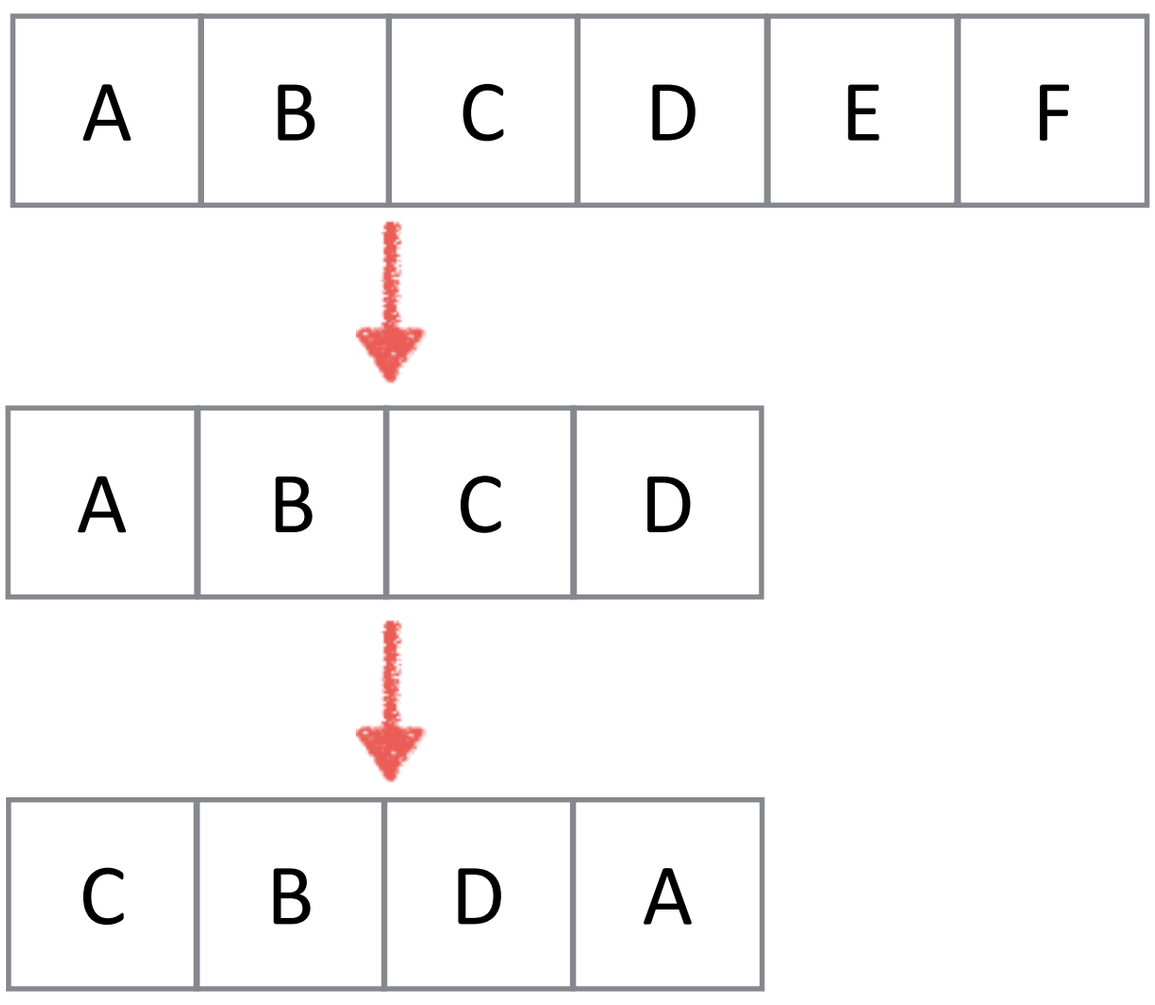
$$\mathrm{IDF}_U(c) = \log\left(\frac{\#\text{ users} \in \mathcal{B}}{\#\text{ users} \in \mathcal{B} \text{ having concept } c}\right)$$

Interactions & Impressions

Already clicked jobs are likely to be clicked again



Interactions & Impressions



Dataset

Filtering-step

Reordering-step