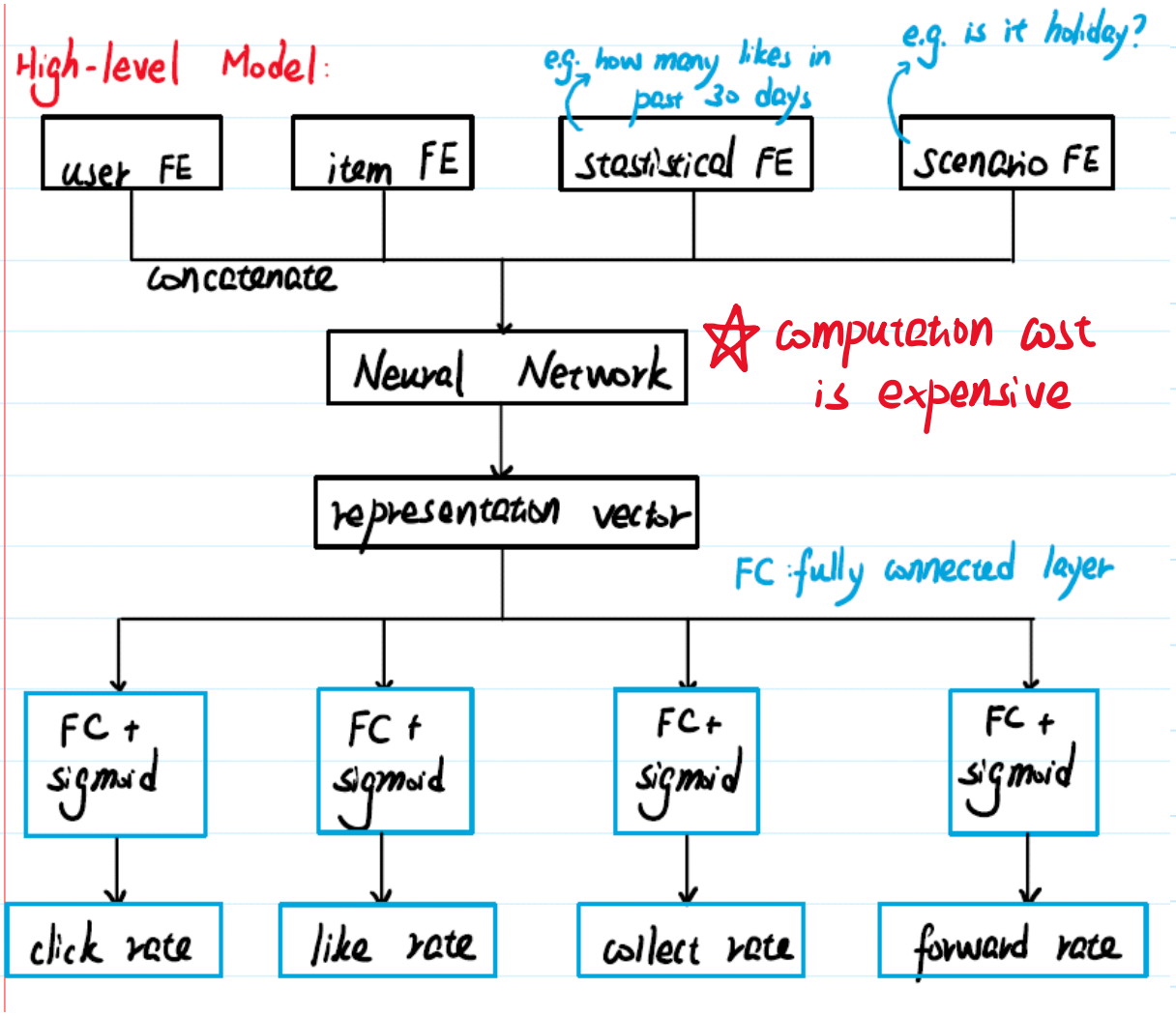
Pre-rank Model

Monday, April 1, 2024 12:19 PM

Pre-rank	Rank
score ~1000 items	score ~100 items
single inference cost is small	cost is large
OK to have low accuracy	need high accuracy

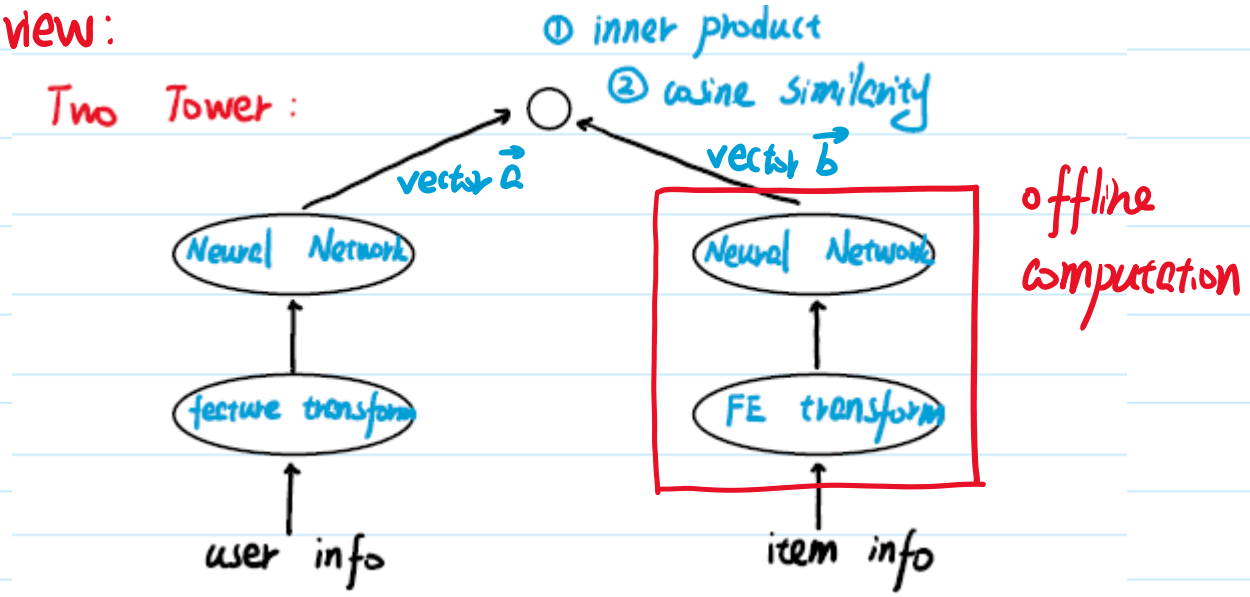
Ranking Model Review:



Pre-rank Model

Monday, April 1, 2024 12:19 PM

Review:



Rank Model	Retrieval Model (Two Tower)
① early concatenate	① late "fusion"
② online inference cost is large	② item vector is offline computed. online inference is only for user tower

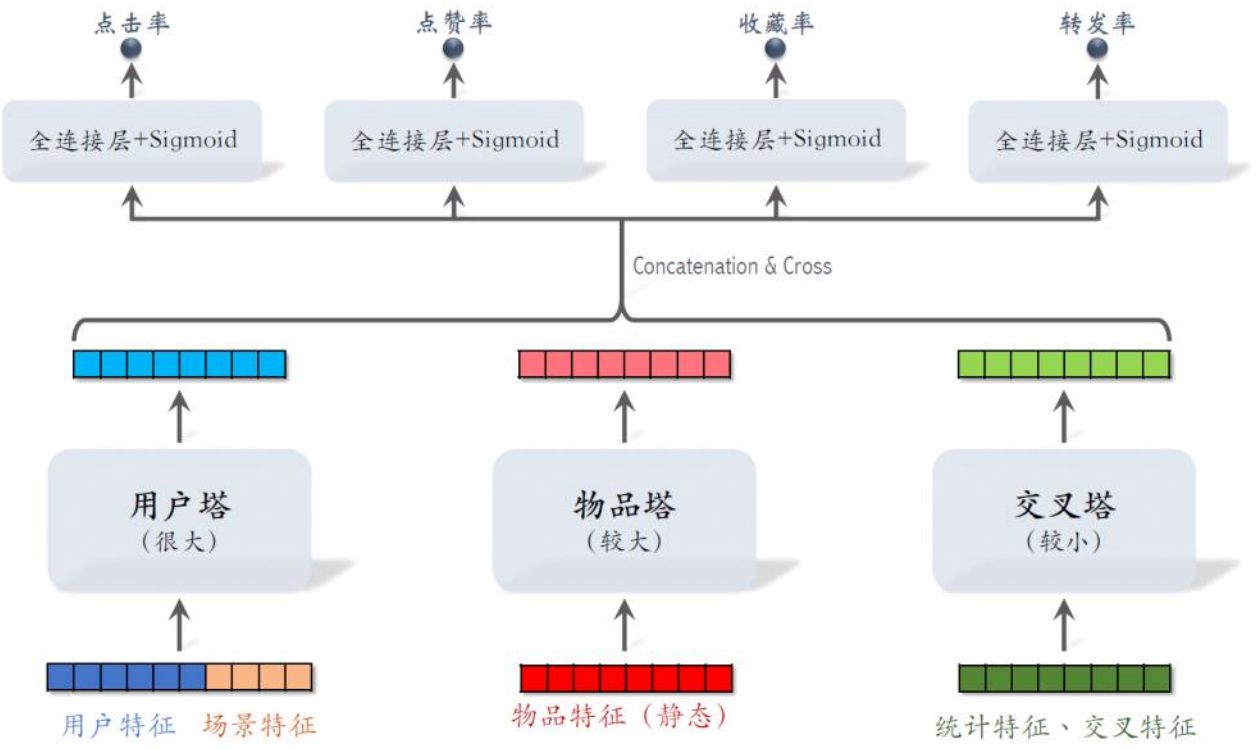
Balance between rank model and retrieval model

→ we will have "pre-rank" model.

Pre-rank Model

Monday, April 1, 2024 12:19 PM

Pre-rank Model (Three Tower Model)



(Picture from Shusen Wang on Bilibili)

User Tower:

Each time we only serve one user. so cost is ok.

Item Tower:

parameter server.

N items → N inference (use cache)

item info is static

Cross Tower:

all rely online inference (network needs to be small)

FC + sigmoid:

N times of inference (where most cost comes from)