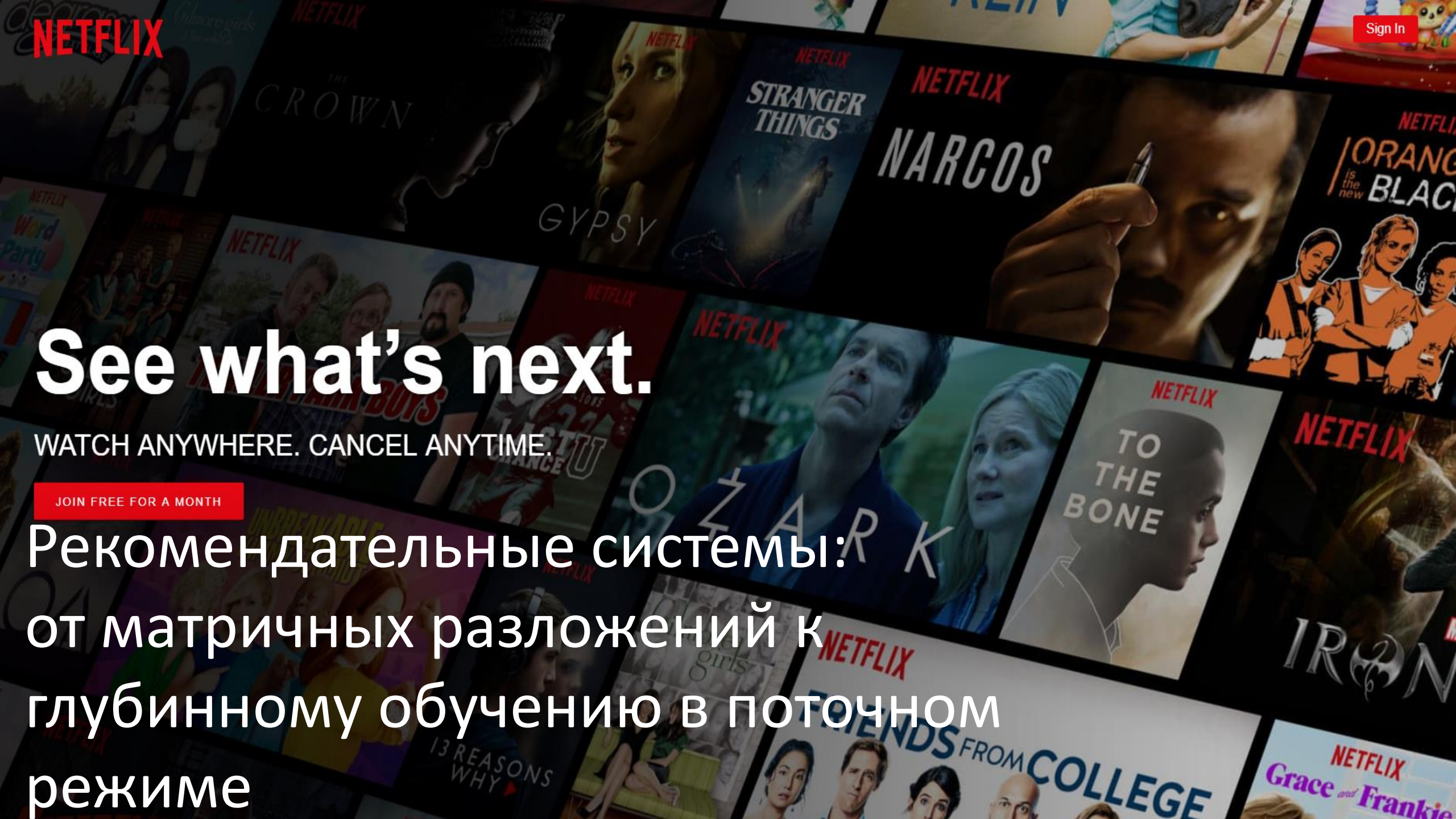
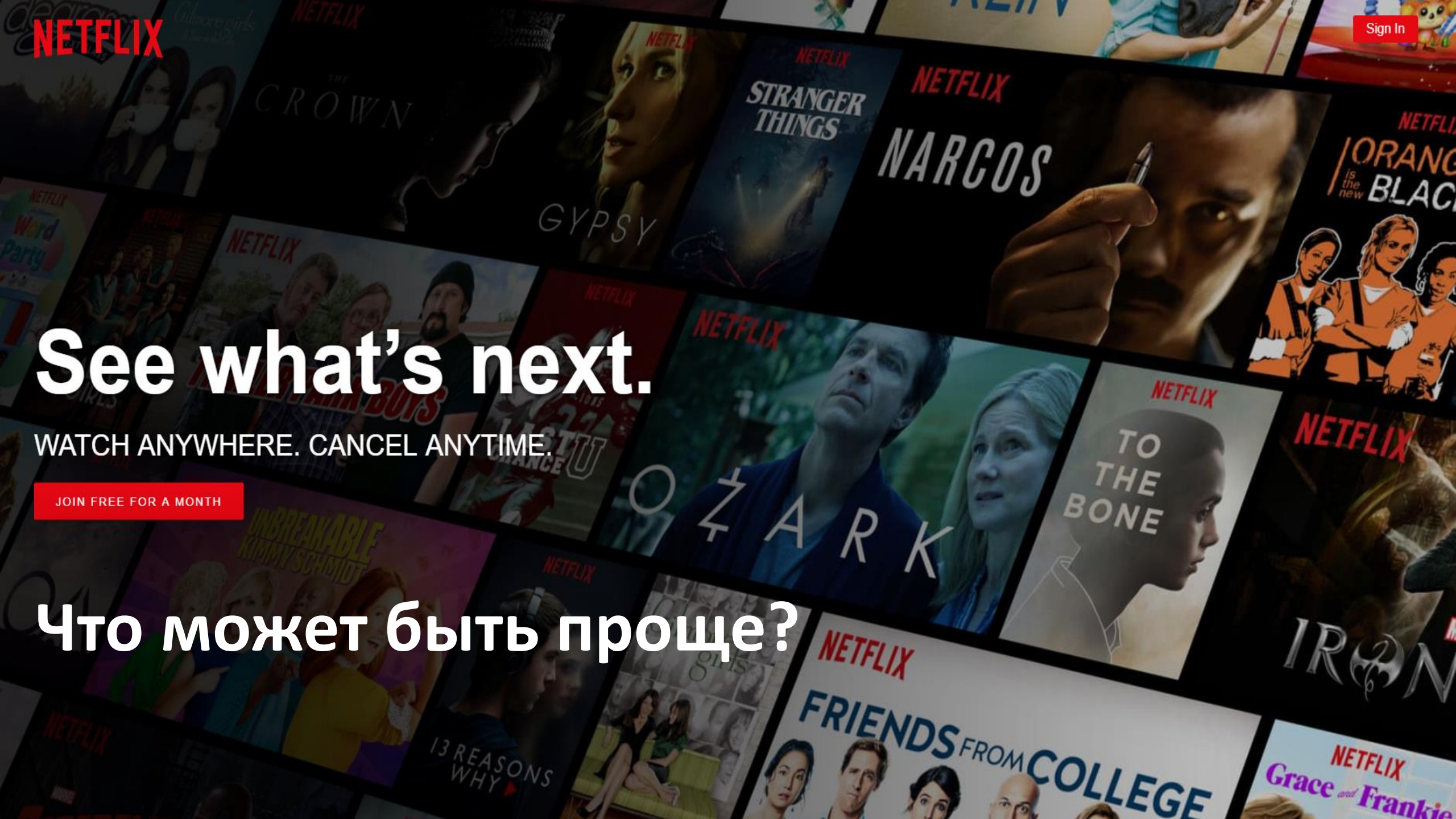


Рекомендательные системы: от матричных разложений к глубинному обучению в поточном режиме


NETFLIX
Sign In
See what's next.
WATCH ANYWHERE. CANCEL ANYTIME.
JOIN FREE FOR A MONTH
Что может быть проще?
THE CROWN
GYPSY
STRANGER THINGS
NARCOS
ORANGE is the new BLACK
OZARK
TO THE BONE
IRON
FRIENDS FROM COLLEGE
Grace and Frankie
13 REASONS WHY
UNBREAKABLE KIMMY SCHMIDT

Михаил Камалов

Рекомендательные системы:
от матричных разложений к глубинному обучению в поточном режиме

Михаил Камалов

EPAM SYSTEMS,

Data Analyst

- Около года работы EPAM`s big data competence center и не менее трех лет участвую в проектах связанных с машинным обучением.
- В основном область моих проектов связанна с обработкой естественного языка и информационным поиском.

Берём данные

Лог юзеров с оценками

Данные о контенте

Данные о юзерах

Строим метрики

Лог юзеров с оценками	Точность
Данные о контенте	Полнота
Данные о юзерах	F-мера

Выбираем модели

Лог юзеров с оценками	Точность
Данные о контенте	Полнота
Данные о юзерах	F-мера

Выбираем модели

Лог юзеров с оценками	Точность	User-based
Данные о контенте	Полнота	
Данные о юзерах	F-мера	

Выбираем модели

Лог юзеров с оценками	Точность	User-based
Данные о контенте	Полнота	Item-based
Данные о юзерах	F-мера	

Выбираем модели

Лог юзеров с оценками	Точность	User-based
Данные о контенте	Полнота	Item-based
Данные о юзерах	F-мера	**Time-based**

Применяем алгоритмы

Лог юзеров с оценками	Точность	User-based	Factorization machine
Данные о контенте	Полнота	Item-based	XGBoost
Данные о юзерах	F-мера	Time-based	

Применяем алгоритмы

Лог юзеров с оценками

Данные о контенте

Данные о юзерах

Точность

Полнота

F-мера

User-based

Item-based

Time-based

Factorization machine

XGBoost

PROFIT?

Как бы не так!

Проблемы с логами

Лог юзеров с ?

Данные о контенте

Данные о юзерах

Точность

Полнота

F-мера

User-based

Item-based

Time-based

Factorization machine

XGBoost

Нет оценок от юзеров

Мультиязычные данные

Ошибки в данных о глубине просмотра

Проблемы с контентом

Лог юзеров с ?

? о контенте

Данные о юзерах

Точность

Полнота

F-мера

User-based

Item-based

Time-based

Factorization machine

XGBoost

Частично без описания и с неизвестной длиной (LIVE)

Много контента с коротким просмотром (< 20 секунд)

Проблемы с юзерами

Лог юзеров с ?

? ? о контенте

? юзерах

Точность

Полнота

F-мера

User-based

Item-based

Time-based

Factorization machine

XGBoost

Нет информации о поле и возрасте

Семья может пользоваться сервисом с одного устройства

Ошибки в числе использованных устройств

Проблемы с метриками

Лог юзеров с ?

? о контенте

? о юзерах

? ? ? ?

User-based

Item-based

Time-based

Factorization machine

XGBoost

Проблемы с алгоритмами

Лог юзеров

о контенте

о юзерах

User-based

Item-based

Time-based

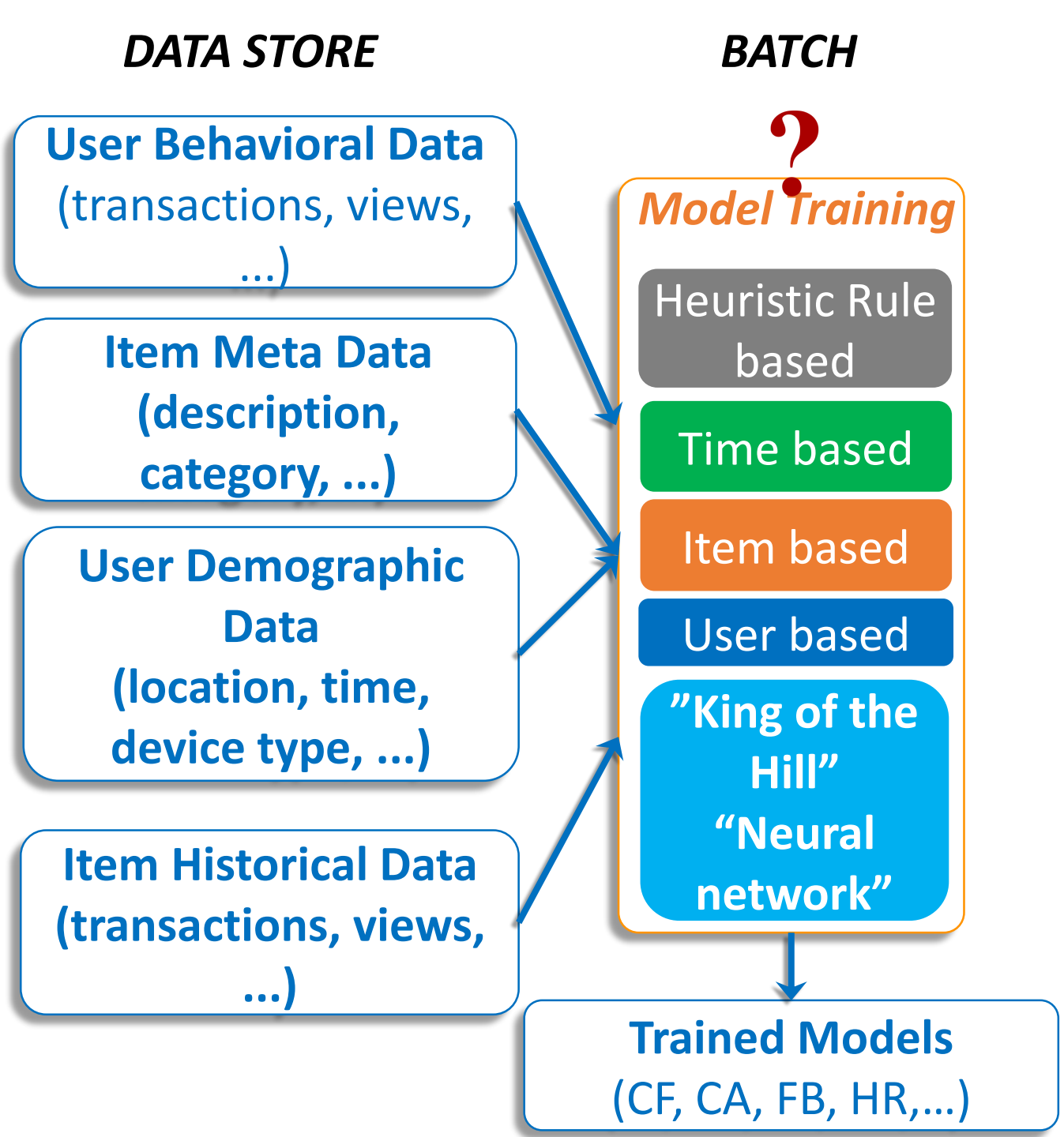
DATA STORE

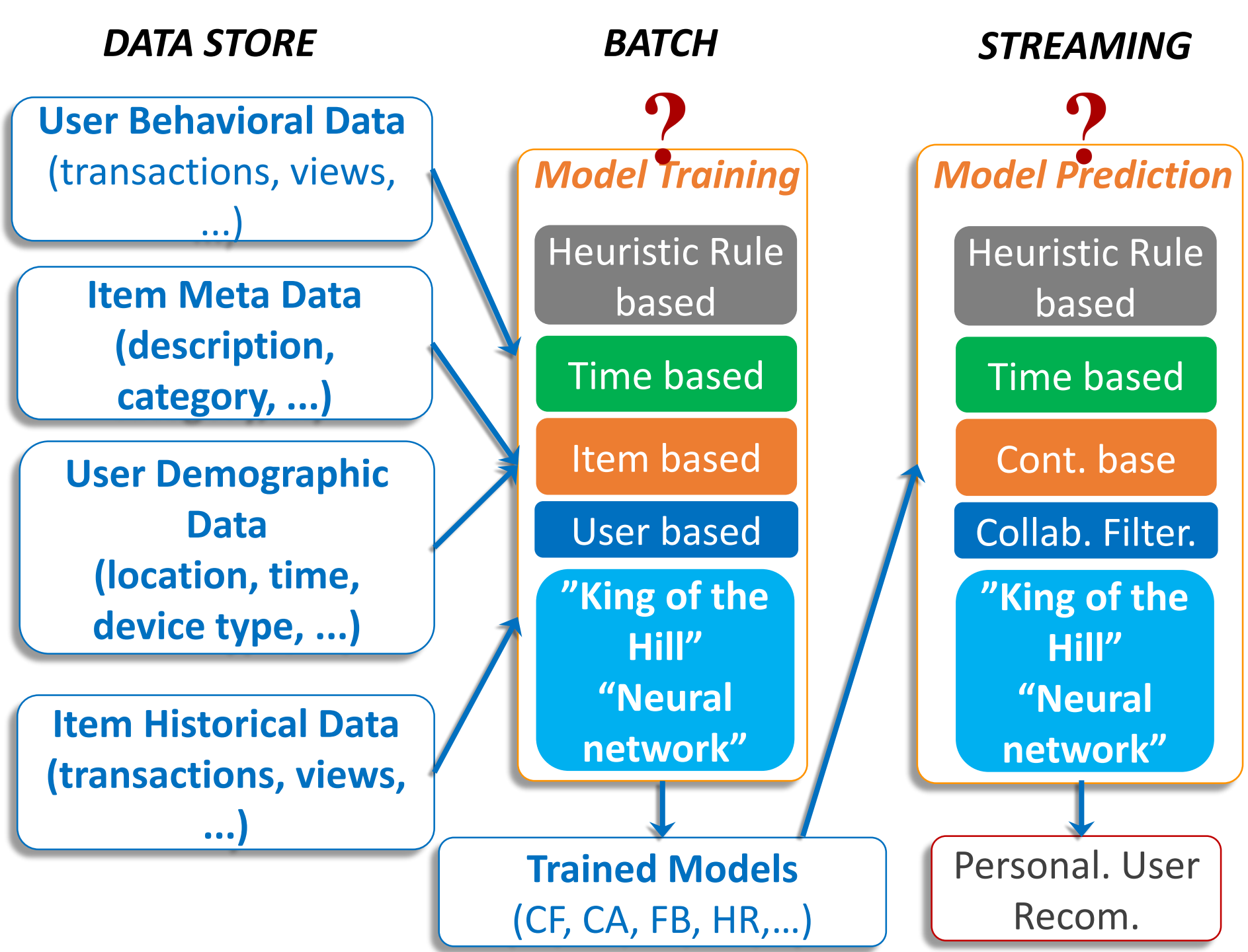
User Behavioral Data
(transactions, views, ...)

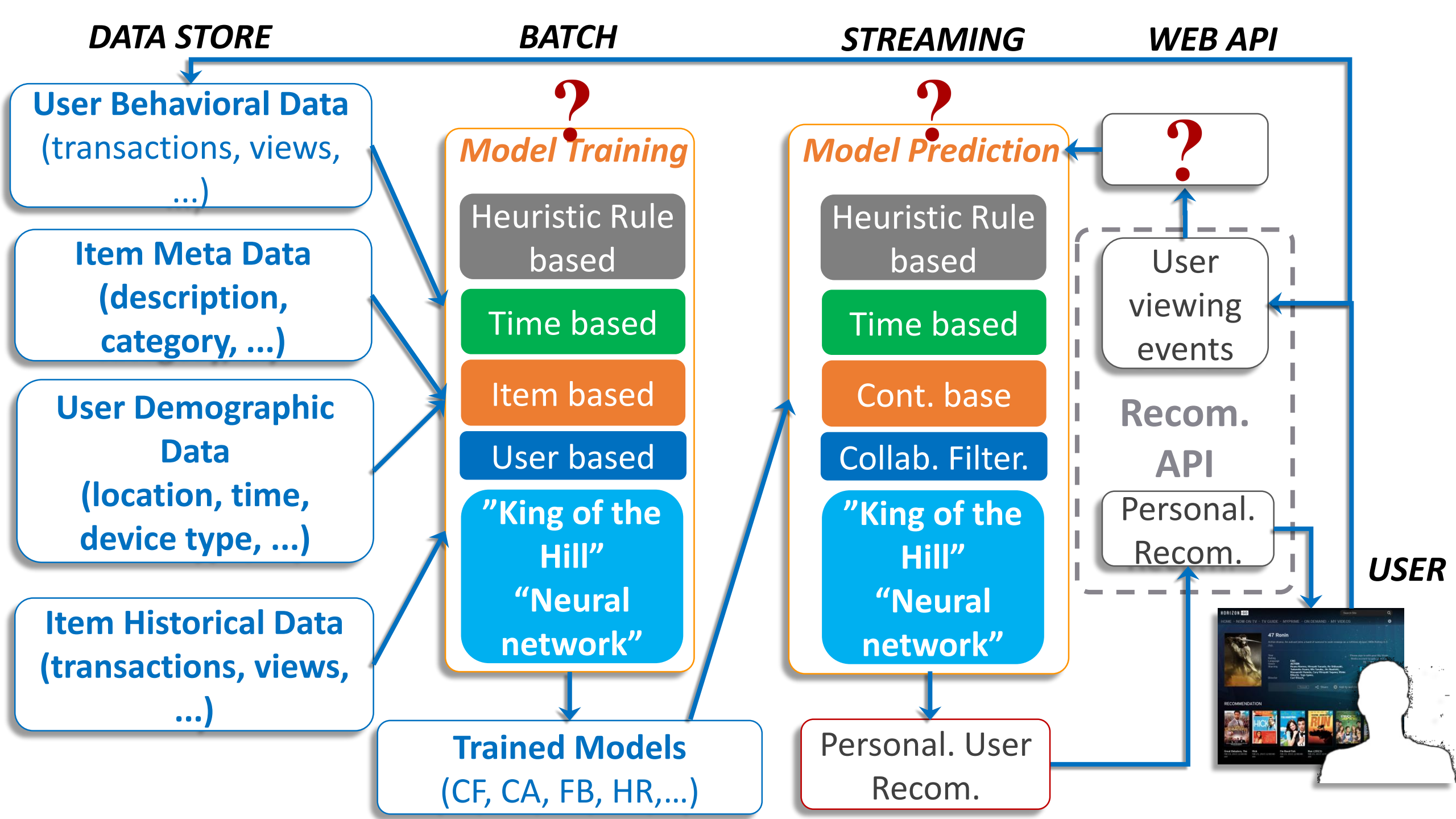
Item Meta Data (description, category, ...)

User Demographic Data (location, time, device type, ...)

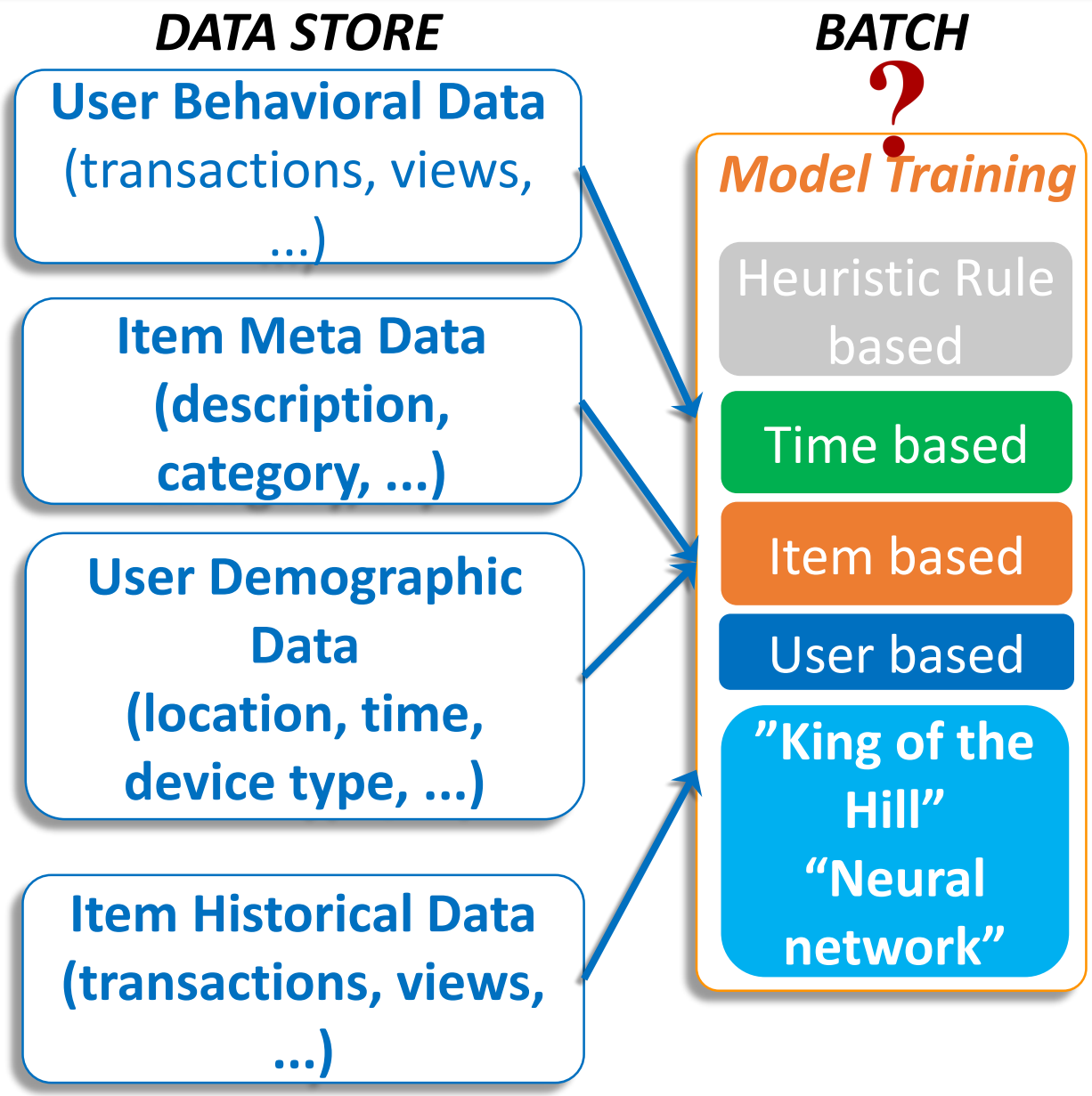
Item Historical Data (transactions, views, ...)

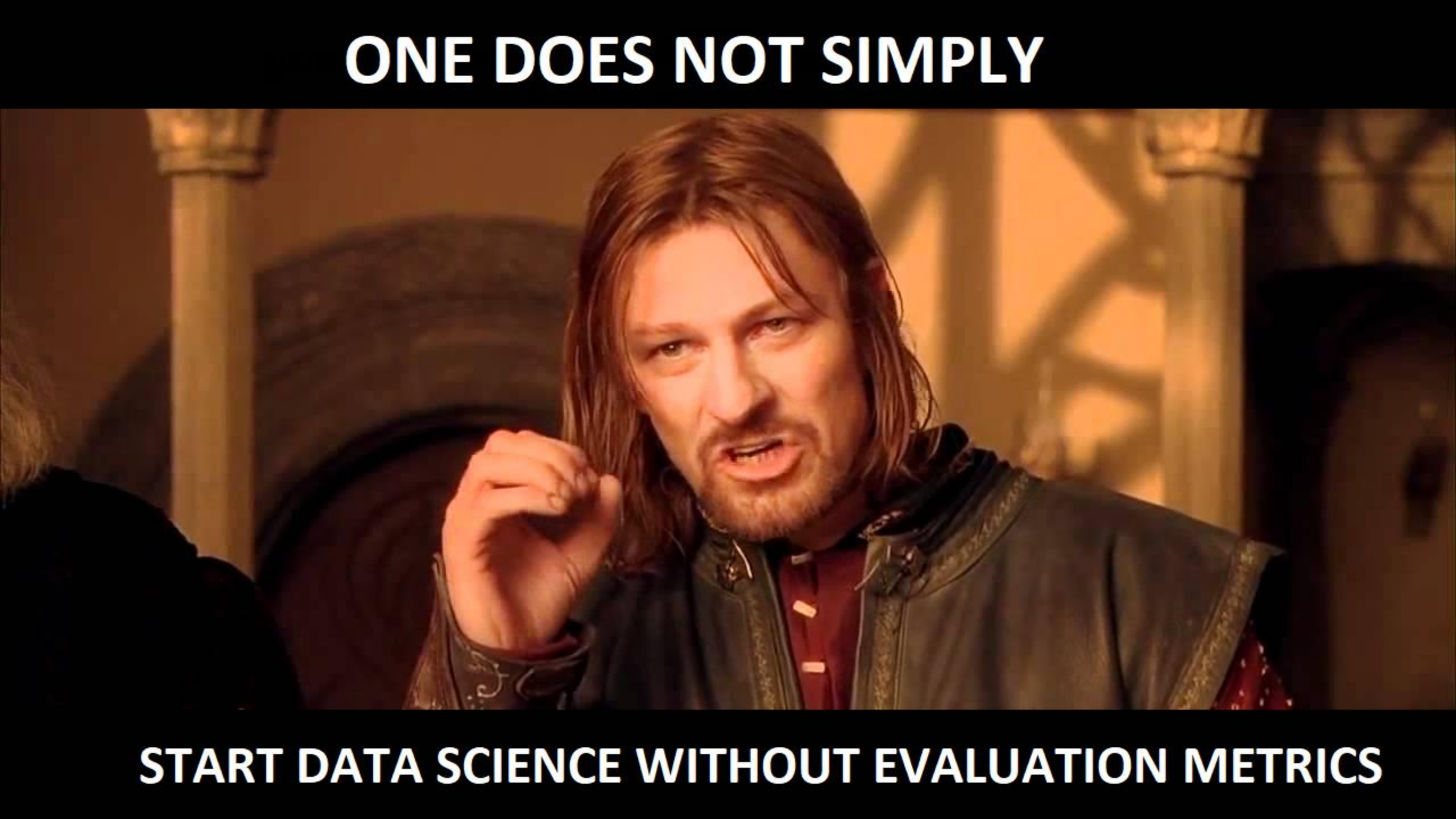

DATA STORE
BATCH
User Behavioral Data (transactions, views, ...)
Item Meta Data (description, category, ...)
User Demographic Data (location, time, device type, ...)
Item Historical Data (transactions, views, ...)
?
Model Training
Heuristic Rule based
Time based
Item based
User based
"King of the Hill" "Neural network"
Trained Models (CF, CA, FB, HR,...)


DATA STORE
BATCH
STREAMING
User Behavioral Data (transactions, views, ...)
Item Meta Data (description, category, ...)
User Demographic Data (location, time, device type, ...)
Item Historical Data (transactions, views, ...)
Model Training
Heuristic Rule based
Time based
Item based
User based
"King of the Hill" "Neural network"
Trained Models (CF, CA, FB, HR,...)
Model Prediction
Heuristic Rule based
Time based
Cont. base
Collab. Filter.
"King of the Hill" "Neural network"
Personal. User Recom.

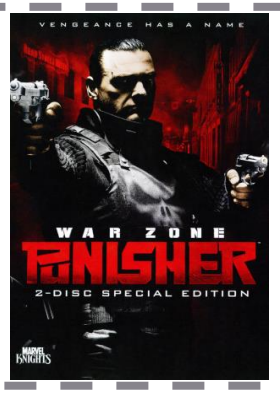

DATA STORE
BATCH
STREAMING
WEB API
User Behavioral Data (transactions, views, ...)
Item Meta Data (description, category, ...)
User Demographic Data (location, time, device type, ...)
Item Historical Data (transactions, views, ...)
Model Training
Heuristic Rule based
Time based
Item based
User based
"King of the Hill"
"Neural network"
Trained Models (CF, CA, FB, HR,...)
Model Prediction
Heuristic Rule based
Time based
Cont. base
Collab. Filter.
"King of the Hill"
"Neural network"
Personal. User Recom.
User viewing events
Recom. API
Personal. Recom.
USER

Архитектура



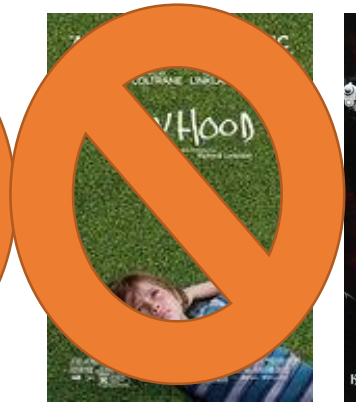

ONE DOES NOT SIMPLY
START DATA SCIENCE WITHOUT EVALUATION METRICS

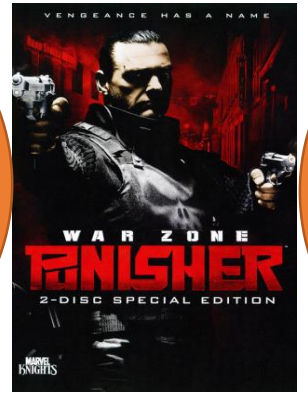
Mean average precision

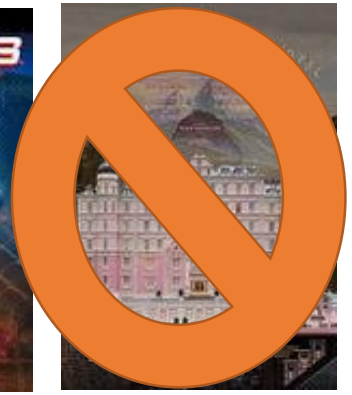


Релевантные фильмы для пользователя Alex

Выдача рекомендаций







Mean average precision

Выдача рекомендаций

Precision:

0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38

Average Precision = (0.5 + 0.4 + 0.43) / 3 = 0.44

Mean Average Precision = 0.44 / 1

Решение проблем

Лог юзеров

о контенте

о юзерах

User-based

Item-based

Time-based

Решение проблем

Лог юзеров

о контенте

о юзерах

Mean Average Precision

User-based

Item-based

Time-based

Catalog Coverage

Рекомендации для
1 пользователя

Рекомендации для
2 пользователя

Все возможные рекомендации

Решение проблем

Лог юзеров

о контенте

о юзерах

Mean Average Precision

User-based

Item-based

Time-based

Решение проблем

Лог юзеров

о контенте

о юзерах

Mean Average Precision

Catalog coverage

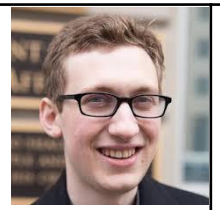
User-based

Item-based

Time-based

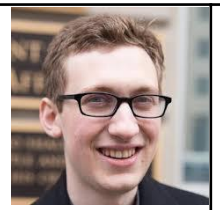
Модификации User based (Коллаборативная фильтрация)

User-item matrix

	Alex 	Mike 	Kate 
			
			
			

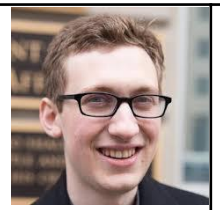
Модификации User based (Коллаборативная фильтрация)

User-item matrix

	Alex 	Mike 	Kate 
	0		
	1.0		
	0.0		

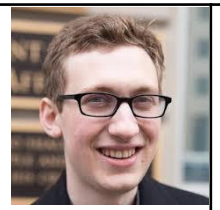
Модификации User based (Коллаборативная фильтрация)

User-item matrix

	Alex 	Mike 	Kate 
	0	1.0	
	1.0	0.0	
	0.0	1.0	

Модификации User based (Коллаборативная фильтрация)

User-item matrix

	Alex 	Mike 	Kate 
	0	1.0	0
	1.0	0.0	0.0
	0.0	1.0	1.0

Модификации User based (Коллаборативная фильтрация)

Item matrix

	Dim 1
	**0.231**
	**0.122**
	**0.443**

Модификации User based (Коллаборативная фильтрация)

Item matrix

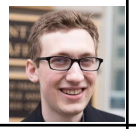
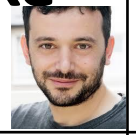
	Dim 1	Dim 2
	0.231	0.125
	0.122	0.008
	0.443	0.512

Модификации User based (Коллаборативная фильтрация)

Item matrix

	Dim 1	Dim 2
	0.23	0.12
	0.12	0.43
	0.44	0.51

User matrix

	Alex 	Mike 	Kate 
Dim 1	0.34	0.18	0.86

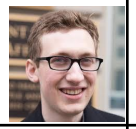
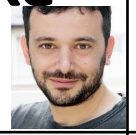
Модификации User based (Коллаборативная фильтрация)

Item matrix

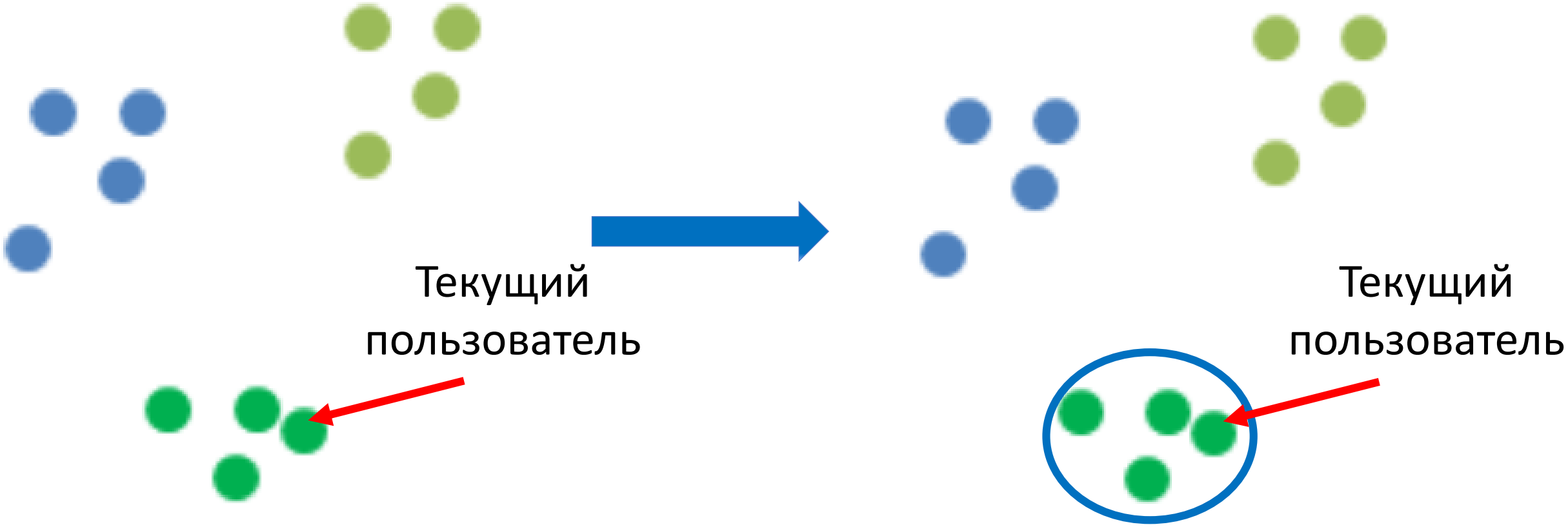
	Dim 1	Dim 2
	0.23	0.12
	0.12	0.43
	0.44	0.51

×

User matrix

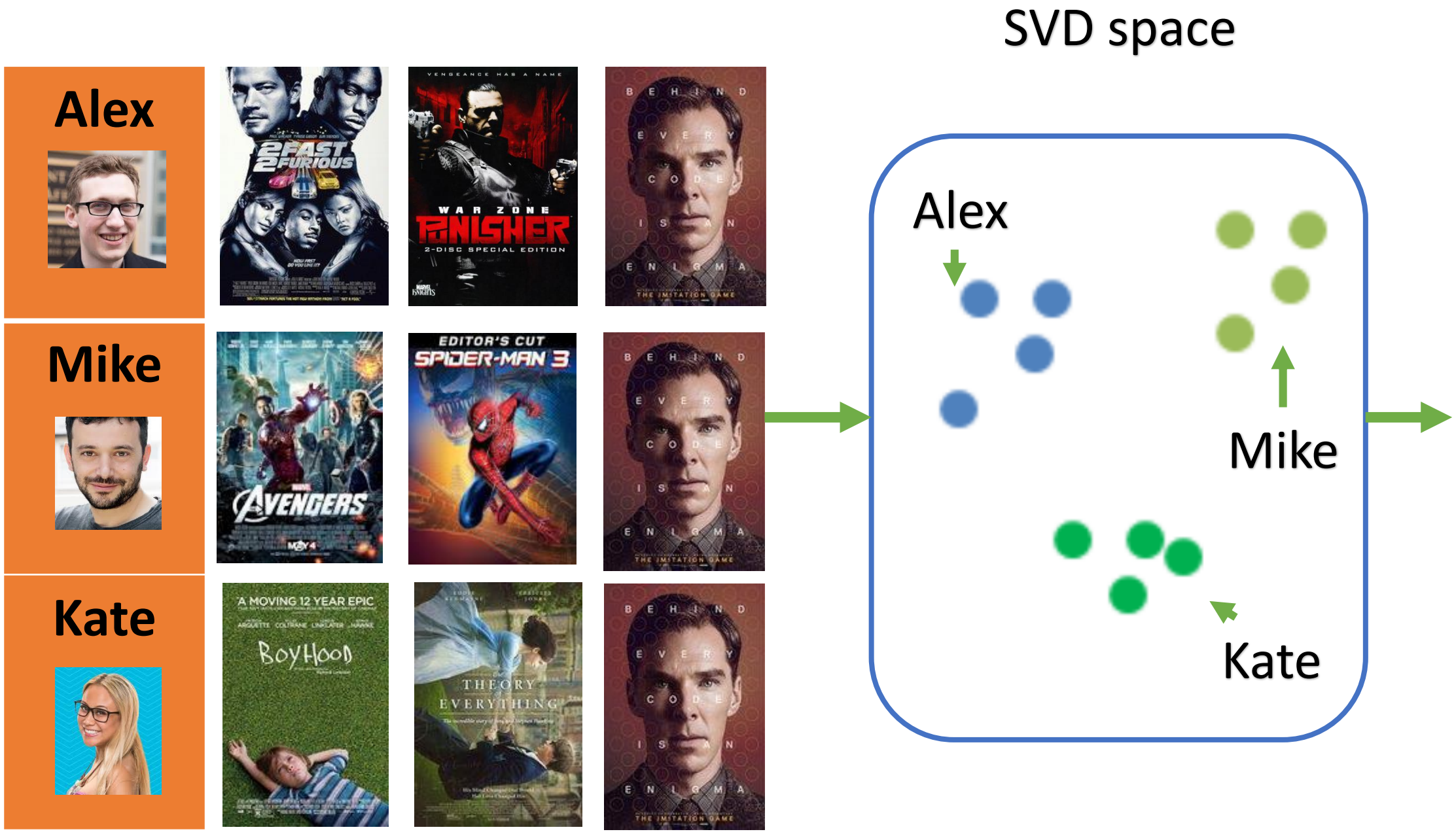
	Alex 	Mike 	Kate 
Dim 1	0.34	0.18	0.86
Dim 2	0.14	0.73	0.23

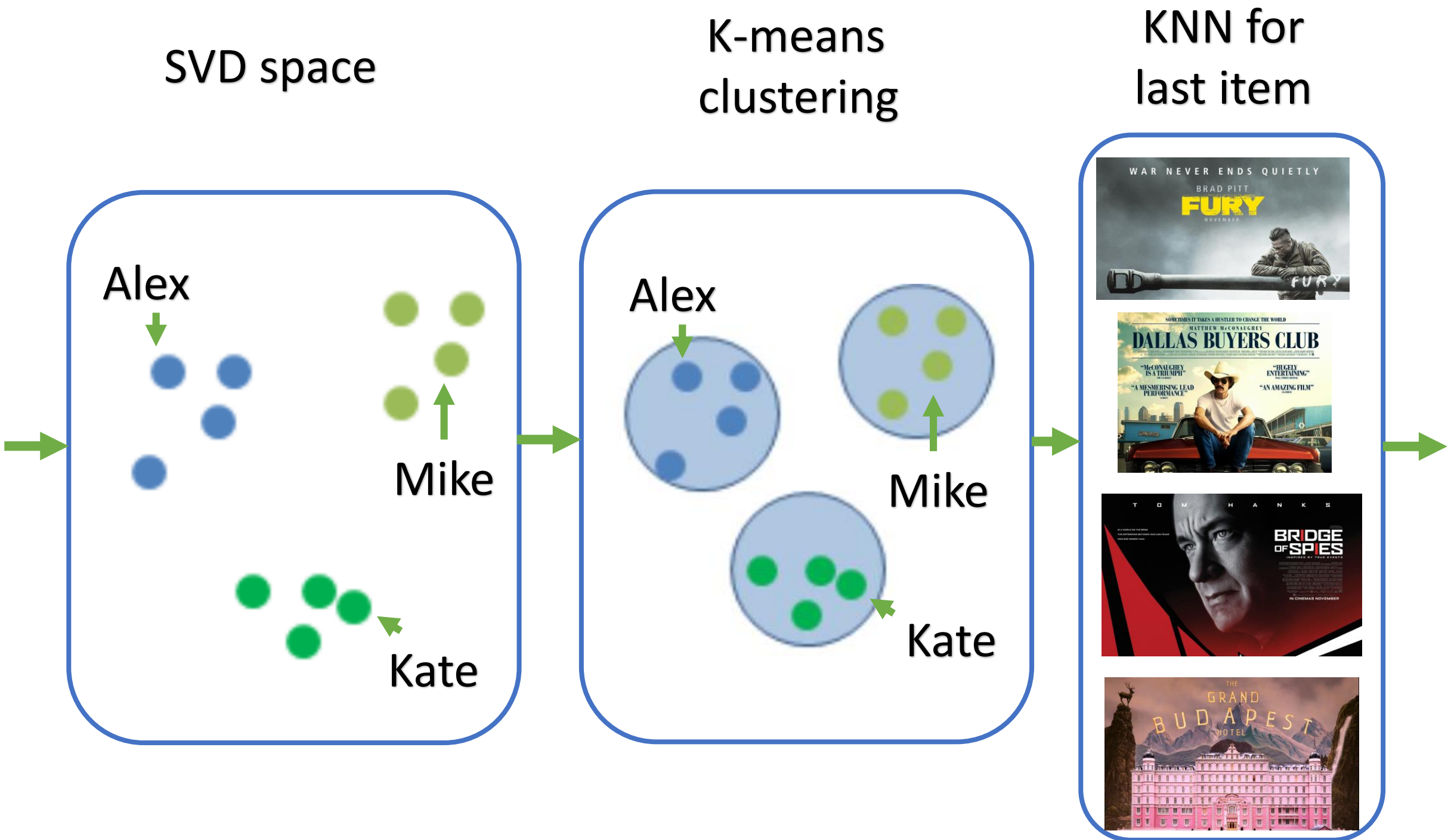
Модификации User based



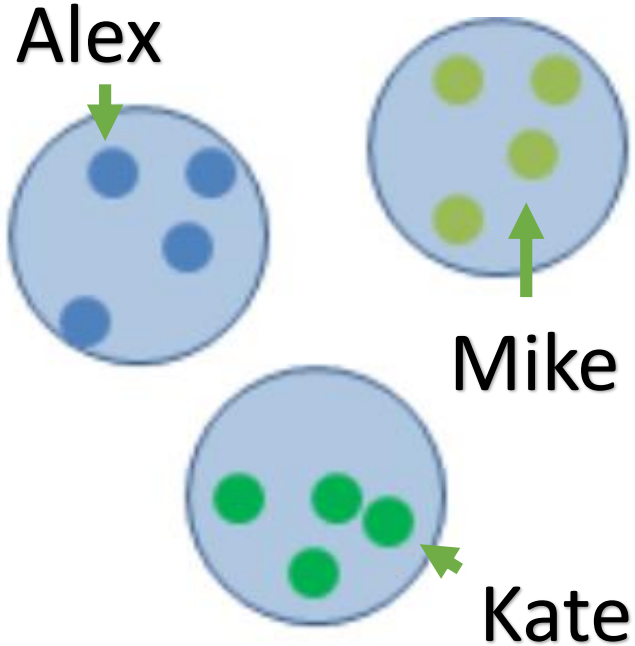
SVD
представление

Генерация
рекомендаций для NN

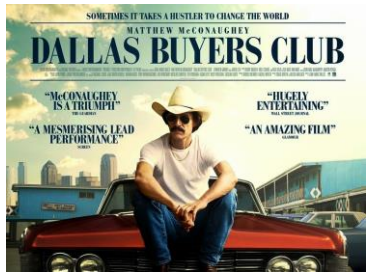

SVD space
Alex
Mike
Kate
2 FAST 2 FURIOUS
WAR ZONE
PUNISHER
2-DISC SPECIAL EDITION
BEHIND EVERY CODE IS AN ENIGMA
THE IMITATION GAME
AVENGERS
EDITOR'S CUT
SPIDER-MAN 3
A MOVING 12 YEAR EPIC
BOYHOOD
THEORY OF EVERYTHING
Alex
Mike
Kate


SVD space
Alex
Mike
Kate
K-means clustering
Alex
Mike
Kate
KNN for last item
WAR NEVER ENDS QUIETLY
BRAD PITT
FURY
NOVEMBER
DALLAS BUYERS CLUB
MATTHEW McCONAUGHEY
"McCONAUGHEY IS A TRIUMPH"
"A MESMERISING LEAD PERFORMANCE"
"HUGELY ENTERTAINING"
"AN AMAZING FILM"
T O M H A N K S
BRIDGE OF SPIES
IN CINEMAS NOVEMBER
THE GRAND BUDAPEST HOTEL

K-means clustering



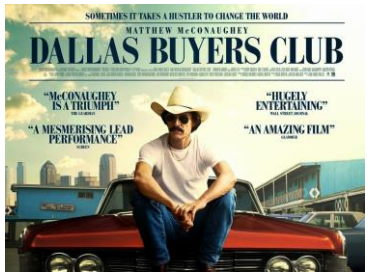
KNN for last item



Personalization

Alex

Kate



Результаты

Recommender	MAP@20 %	Catalog coverage %
random	0.0	5.9
Baseline-top-20	2.4	0.0

Результаты

Recommender	MAP@20 %	Catalog coverage %
random	0.0	5.9
Baseline-top-20	2.4	0.0
SVD	8.4	3.1

Результаты

Recommender	MAP@20 %	Catalog coverage %
random	0.0	5.9
Baseline-top-20	2.4	0.0
SVD	8.4	3.1
SVD+NN	**8.9**	**3.6**

Результаты

Recommender	MAP@20 %	Catalog coverage %
random	0.0	5.9
Baseline-top-20	2.4	0.0
SVD	8.4	3.1
SVD+NN	**8.9**	**3.6**
SVD+KNN+KMEANS	3.9	1.3

Решение проблем

Лог юзеров

о контенте

о юзерах

Mean Average Precision

Catalog coverage

User-based

Item-based

Time-based

Решение проблем

Лог юзеров ?

? о контенте

? о юзерах

Mean Average Precision

Catalog coverage

User-based

Item-based

Time-based

SVD + NN

?

Модификации Item based

Признаки контента	Представление

Модификации Item based

Признаки контента	Представление
• meta (актёры, режисер, тематика) • Тексовое описание • combine (meta + long description)	**Term frequency – Inverted document frequency (TF-IDF)**

Модификации Item based

Признаки контента	Представление
• meta (актёры, режисер, тематика) • Тексовое описание • combine (meta + long description)	**Term frequency – Inverted document frequency (TF-IDF)**
• meta (актёры, режисер, тематика)	**One hot encoding**

Модификации Item based

Признаки контента	Представление
• meta (актёры, режисер, тематика) • Тексовое описание • combine (meta + long description)	**Term frequency – Inverted document frequency (TF-IDF)**
• meta (актёры, режисер, тематика)	**One hot encoding**
• История пользовательского контента	**Word2vec** –words embedding.

Модификации Item based

1. $\mathbf{Tf - Idf} = \mathbf{tf} * \mathbf{log}\left(\frac{|D|}{df}\right)$, tf - частота слова в документе,

df – число документов содержащих слово. D – корпус документов.

Alex 

Модификации Item based

1. $\mathbf{Tf - Idf} = \mathbf{tf} * \mathbf{log}\left(\frac{|D|}{df}\right)$, tf - частота слова в документе,

df – число документов содержащих слово. D – корпус документов.

Alex 

Lee Tamahori ACTION Samuel L.
Jackson ACTIE Samuel L. Jackson

Модификации Item based

1. $\mathbf{Tf - Idf} = \mathbf{tf} * \mathbf{log}\left(\frac{|D|}{df}\right)$, tf - частота слова в документе,

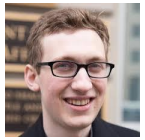
df – число документов содержащих слово. D – корпус документов.

Alex 

Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

Rob Cohen ACTION Lee Tamahori Drama

Модификации Item based

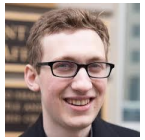
Alex 

Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

Rob Cohen ACTION Lee Tamahori Drama

Lee Tamahori ACTION Rob Cohen Samuel Jackson Drama

Модификации Item based

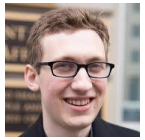
Alex 

Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

Rob Cohen ACTION Lee Tamahori Drama

	Lee	Tamahori	ACTION	Rob	Cohen	Samuel	Jackson	Drama
	0.29							
	0.13							

Модификации Item based

Alex 

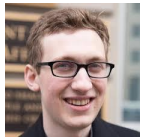
Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

Rob Cohen ACTION Lee Tamahori Drama

||

	Lee	Tamahori	ACTION	Rob	Cohen	Samuel	Jackson	Drama
	0.29	0.69						
	0.13	0.12						

Модификации Item based

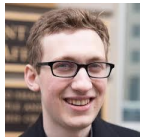
Alex 

Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

Rob Cohen ACTION Lee Tamahori Drama

	Lee	Tamahori	ACTION	Rob	Cohen	Samuel	Jackson	Drama
	0.29	0.69	1.38					
	0.13	0.13	1.38					

Модификации Item based

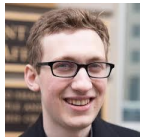
Alex 

Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

Rob Cohen ACTION Lee Tamahori Drama

	Lee	Tamahori	ACTION	Rob	Cohen	Samuel	Jackson	Drama
	0.29	0.69	1.38	0				
	0.13	0.13	1.38	0.64				

Модификации Item based

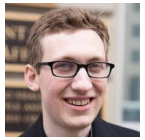
Alex 

Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

Rob Cohen ACTION Lee Tamahori Drama

	Lee	Tamahori	ACTION	Rob	Cohen	Samuel	Jackson	Drama
	0.29	0.69	1.38	0	0			
	0.13	0.13	1.38	0.64	0.64			

Модификации Item based

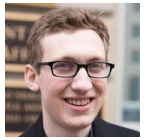
Alex 

Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

Rob Cohen ACTION Lee Tamahori Drama

	Lee	Tamahori	ACTION	Rob	Cohen	Samuel	Jackson	Drama
	0.29	0.69	1.38	0	0	1.38		
	0.13	0.13	1.38	0.64	0.64	0		

Модификации Item based

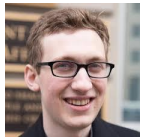
Alex 

Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

Rob Cohen ACTION Lee Tamahori Drama

	Lee	Tamahori	ACTION	Rob	Cohen	Samuel	Jackson	Drama
	0.29	0.69	1.38	0	0	1.38	1.38	
	0.13	0.13	1.38	0.64	0.64	0	0	

Модификации Item based

Alex 

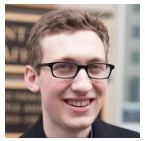
Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

Rob Cohen ACTION Lee Tamahori Drama

	Lee	Tamahori	ACTION	Rob	Cohen	Samuel	Jackson	Drama
	0.29	0.69	1.38	0	0	1.38	1.38	0
	0.13	0.13	1.38	0.64	0.64	0	0	0.23

Модификации Item based

1. $One - hot - encoding$ преобразование категориальных признаков

Alex 

Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

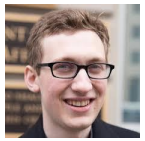
Rob Cohen ACTION Lee Tamahori Drama

||

Lee Tamahori	**ACTION**	**Rob Cohen**	**Samuel Jackson**	**DRAMA**

Модификации Item based

1. $One - hot - encoding$ преобразование категориальных признаков

Alex 

Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

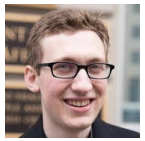
Rob Cohen ACTION Lee Tamahori Drama

||

	Lee Tamahori	ACTION	Rob Cohen	Samuel Jackson	DRAMA
	1				
	1				

Модификации Item based

1. $One - hot - encoding$ преобразование категориальных признаков

Alex 

Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

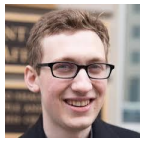
Rob Cohen ACTION Lee Tamahori Drama

||

	Lee Tamahori	ACTION	Rob Cohen	Samuel Jackson	DRAMA
	1	2			
	1	2			

Модификации Item based

1. $One - hot - encoding$ преобразование категориальных признаков

Alex 

Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

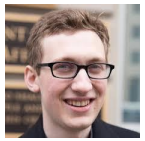
Rob Cohen ACTION Lee Tamahori Drama

||

	Lee Tamahori	ACTION	Rob Cohen	Samuel Jackson	DRAMA
	1	2	0		
	1	2	1		

Модификации Item based

1. $One - hot - encoding$ преобразование категориальных признаков

Alex 

Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

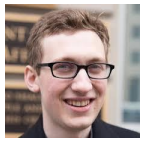
Rob Cohen ACTION Lee Tamahori Drama

||

	Lee Tamahori	ACTION	Rob Cohen	Samuel Jackson	DRAMA
	1	2	0	2	
	1	2	1	0	

Модификации Item based

1. $One - hot - encoding$ преобразование категориальных признаков

Alex 

Lee Tamahori ACTION Samuel L. Jackson ACTIE Samuel L. Jackson

Rob Cohen ACTION Lee Tamahori Drama

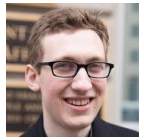
||

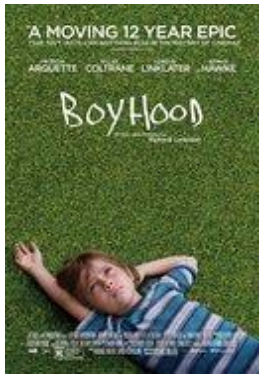
	Lee Tamahori	ACTION	Rob Cohen	Samuel Jackson	DRAMA
	1	2	0	2	1
	1	2	1	0	0

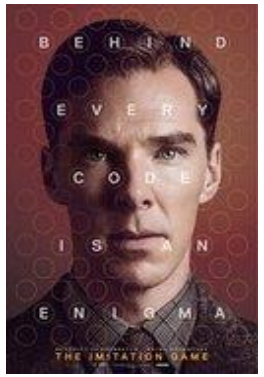
Модификации Item based

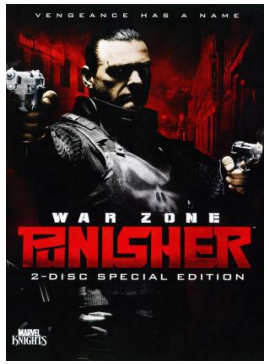
1. $Word\ 2\ Vec$ – представление слова вектором составленным по ближайшим словам (word imbedding)

Alex



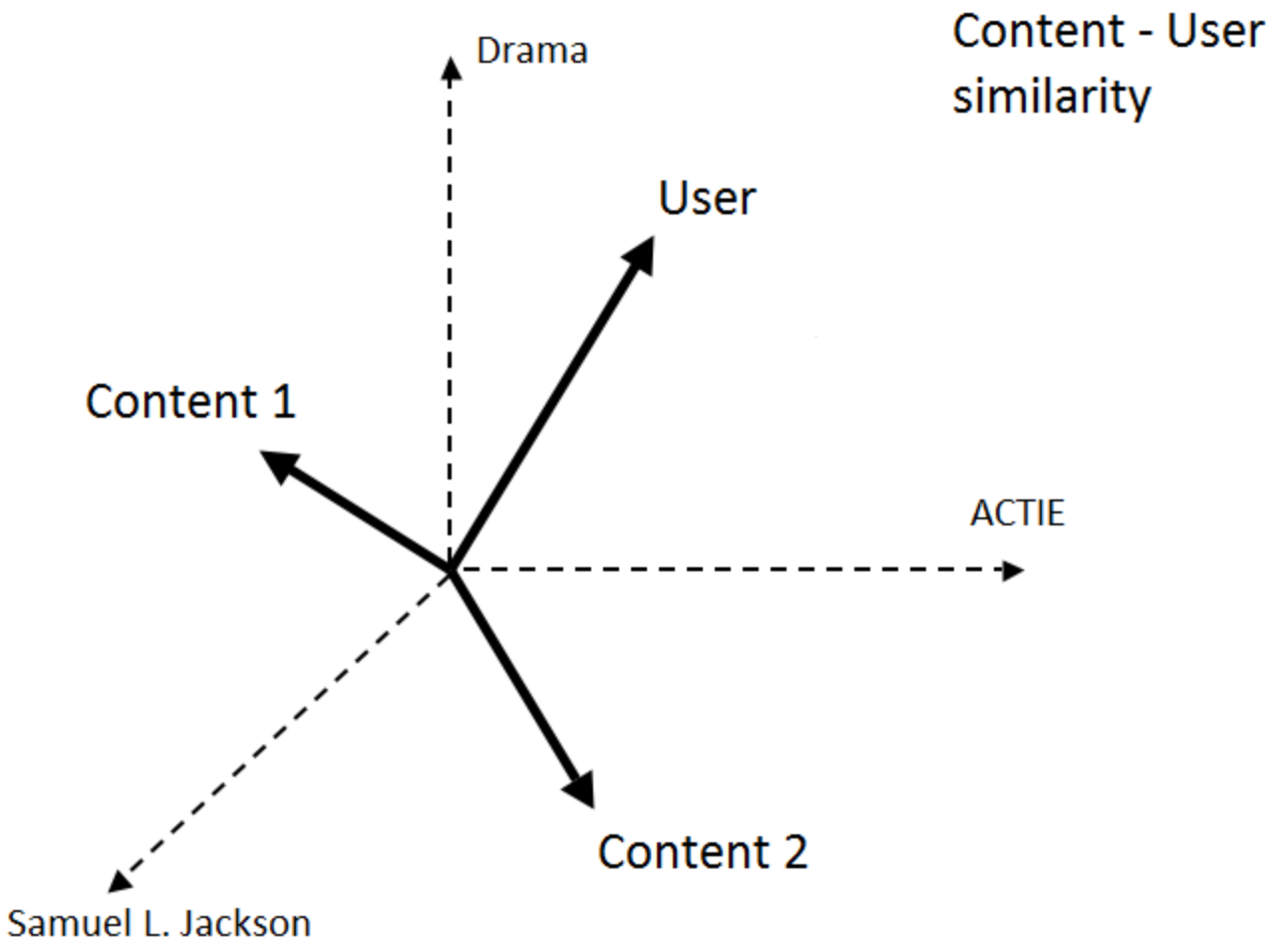






0.678	0.124	0.351	0.276

Результаты


Content - User similarity
Drama
User
Content 1
ACTIE
Content 2
Samuel L. Jackson

Результаты

Model	MAP@20 %	Catalog coverage %
Baseline-top-20	2.4	0.0

Результаты

Model	MAP@20 %	Catalog coverage %
Baseline-top-20	2.4	0.0
Meta – One Hot Enc.	**3.5**	**3.2**

Результаты

Model	MAP@20 %	Catalog coverage %
Baseline-top-20	2.4	0.0
Meta – One Hot Enc.	**3.5**	**3.2**
Combine – Tf-Idf	3.2	4.3

Результаты

Model	MAP@20 %	Catalog coverage %
Baseline-top-20	2.4	0.0
Meta – One Hot Enc.	**3.5**	**3.2**
Combine – Tf-Idf	3.2	4.3
Long desc. – Tf-Idf	3.1	4.1

Результаты

Model	MAP@20 %	Catalog coverage %
Baseline-top-20	2.4	0.0
Meta – One Hot Enc.	**3.5**	**3.2**
Combine – Tf-Idf	3.2	4.3
Long desc. – Tf-Idf	3.1	4.1
Meta – Tf-Idf	3.2	5.2

Результаты

Model	MAP@20 %	Catalog coverage %
Baseline-top-20	2.4	0.0
Meta – One Hot Enc.	**3.5**	**3.2**
Combine – Tf-Idf	3.2	4.3
Long desc. – Tf-Idf	3.1	4.1
Meta – Tf-Idf	3.2	5.2
Top1-w2v	0.8	6.3

Решение проблем

Лог юзеров ?

? о контенте

? о юзерах

Mean Average Precision

Catalog coverage

User-based

Item-based

Time-based

SVD + NN

?

Решение проблем

Лог юзеров

о контенте

о юзерах

Mean Average Precision

Catalog coverage

User-based

Item-based

Time-based

SVD + NN

KNN

Модификации state of the art алгоритмов

1) ВРЕМЕННЫЕ ИНТЕРВАЛЫ

- День был поделен:
  - **5 am – 11 am**
  - **11 am – 4 pm**
  - **4 pm – 9 pm**
  - **4 pm – 5 am**

- Рекомендации составлялись на основе частоты контента в определенный временной интервал.

Модификации state of the art алгоритмов

2) ДНЕВНЫЕ ИНТЕРВАЛЫ

- Разбиение по дням:
  - **All days**
  - **Working days (Monday – Friday)**
  - **Weekend days (Saturday, Sunday)**
- Рекомендации составлялись на основе частоты контента в определенный день.

Модификации state of the art алгоритмов

Recommender	MAP@20 %	Catalog coverage %

Модификации state of the art алгоритмов

Recommender	MAP@20 %	Catalog coverage %
Time dependency	0.2	2.1

Модификации state of the art алгоритмов

Recommender	MAP@20 %	Catalog coverage %
Time dependency	0.2	2.1
Day dependency	**2.4**	**4.6**

Решение проблем

Лог юзеров

о контенте

о юзерах

Mean Average Precision

Catalog coverage

User-based

Item-based

Time-based

SVD + NN

KNN

Решение проблем

Лог юзеров

о контенте

о юзерах

Mean Average Precision

Catalog coverage

User-based

Item-based

Time-based

SVD + NN

KNN

Day dependency

Модификации Heuristic

Модификации Heuristic

Работа в центре

Модификации Heuristic

Recommender	MAP@20 %	Catalog coverage %
Командировки	1.3	3.1
Работать в центре	1.4	2.6

Решение проблем

Лог юзеров

о контенте

о юзерах

Mean Average Precision

Catalog coverage

User-based

Item-based

Time-based

SVD + NN

KNN

Day dependency

Решение проблем

Лог юзеров

о контенте

о юзерах

Mean Average Precision

Catalog coverage

User-based

Item-based

Time-based

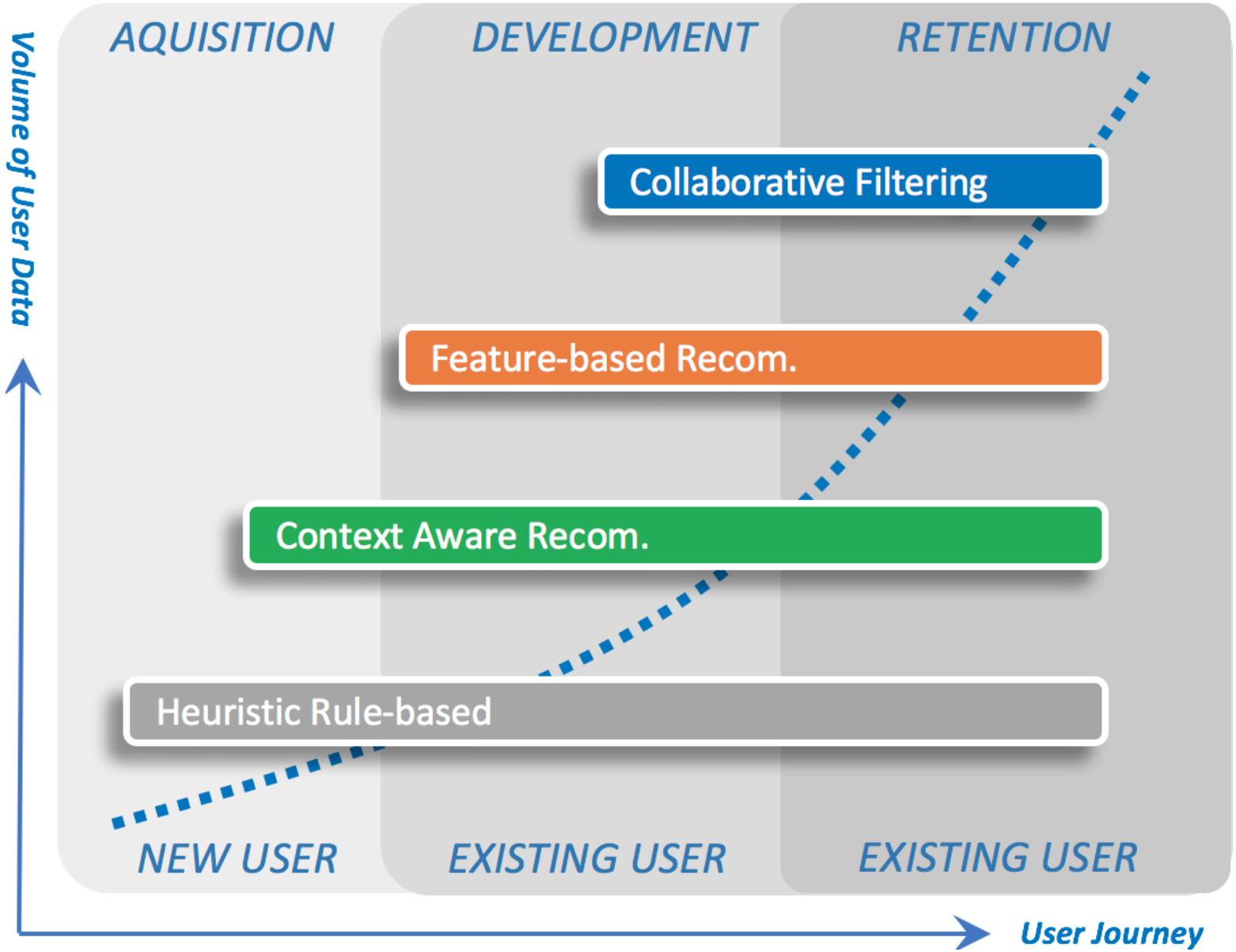
SVD + NN

KNN

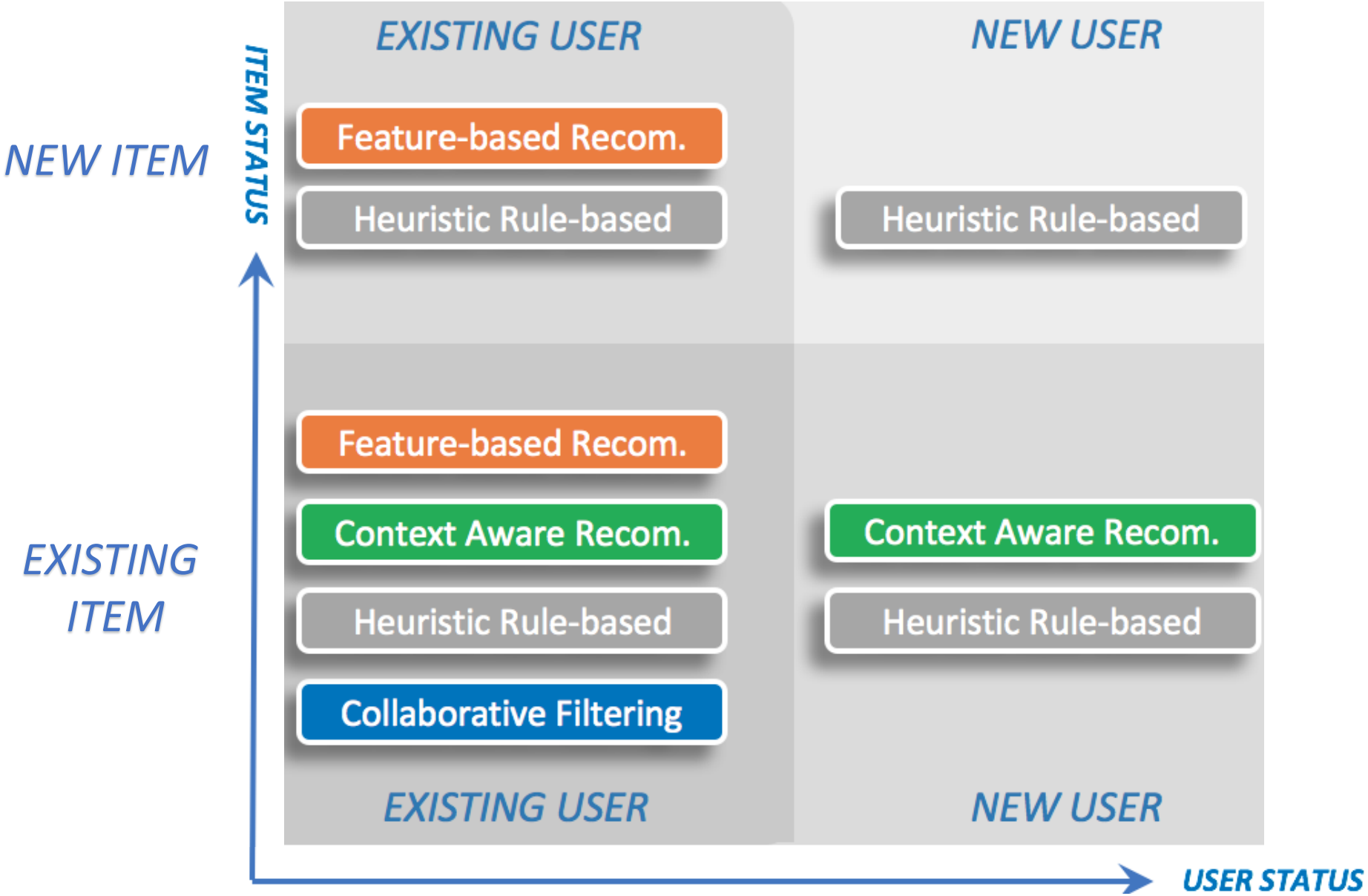
Day dependency

Heuristic

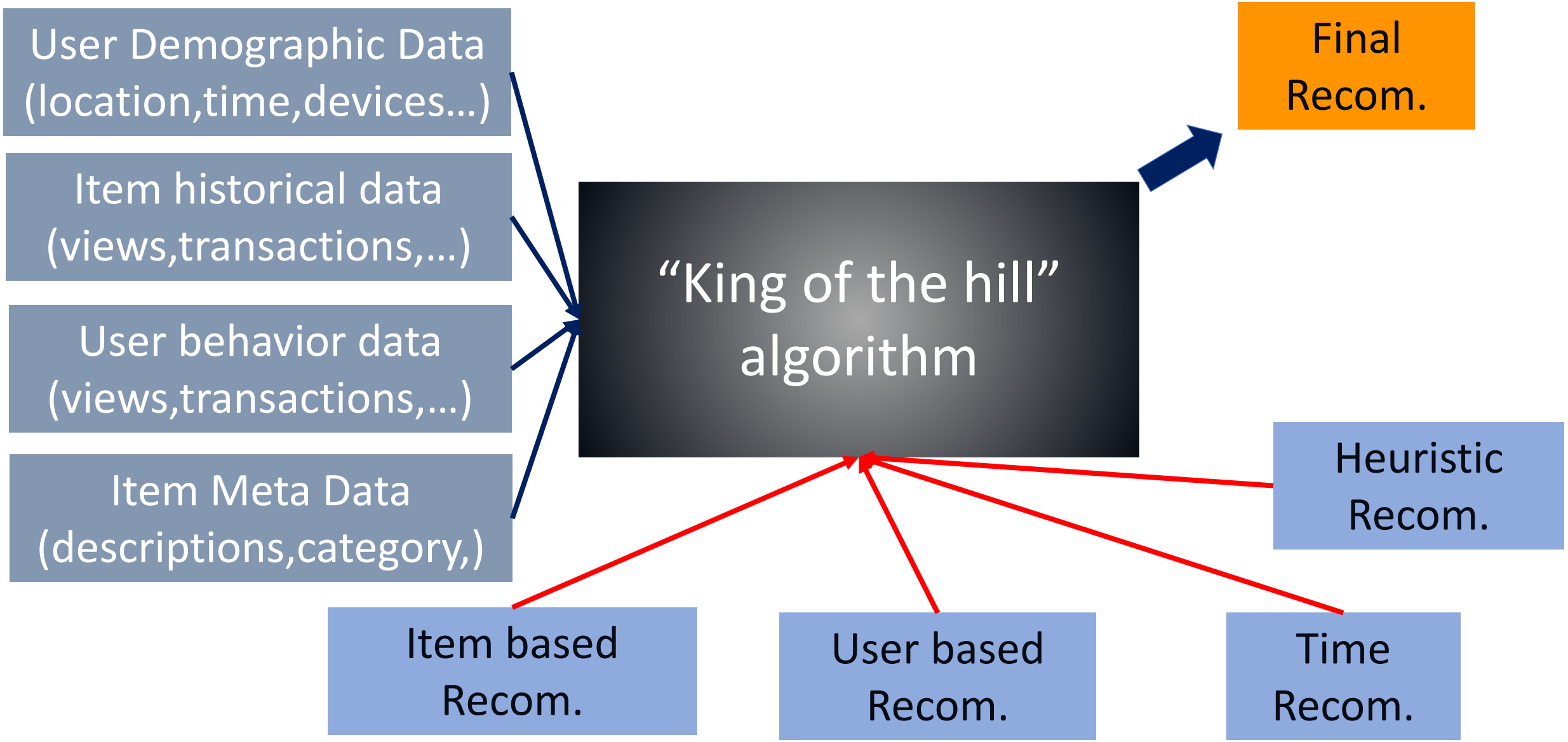
Результаты


AQUISITION
DEVELOPMENT
RETENTION
Volume of User Data
Collaborative Filtering
Feature-based Recom.
Context Aware Recom.
Heuristic Rule-based
NEW USER
EXISTING USER
EXISTING USER
User Journey

Результаты


EXISTING USER
NEW USER
ITEM STATUS
NEW ITEM
Feature-based Recom.
Heuristic Rule-based
Heuristic Rule-based
EXISTING ITEM
Feature-based Recom.
Context Aware Recom.
Heuristic Rule-based
Collaborative Filtering
Context Aware Recom.
Heuristic Rule-based
EXISTING USER
NEW USER
USER STATUS

Модификации Ensembles



Результаты

- SVD отлично коррелируется с различными модификациями SVD
- Переранжированный SVD-UNN при помощи word2Vec, увеличивает производительность

Результаты

- SVD отлично коррелируется с различными модификациями SVD
- Пере
  прои

Algorithm	MAP@20 %	% of max

Результаты

Algorithm	MAP@20 %	% of max
best of two (initial minimum)	14.2	72.32

Результаты

Algorithm	MAP@20 %	% of max
best of two (initial minimum)	14.2	72.32
“king of the hill”, take 1 alg.	16.8	80.26

Результаты

Algorithm	MAP@20 %	% of max
best of two (initial minimum)	14.2	72.32
“king of the hill”, take 1 alg.	16.8	80.26
“king of the hill”, mixing of ranks (linear score)	16.5	82.16

Результаты

Algorithm	MAP@20 %	% of max
best of two (initial minimum)	14.2	72.32
“king of the hill”, take 1 alg.	16.8	80.26
“king of the hill”, mixing of ranks (linear score)	16.5	82.16
“king of the hill”, mixing of ranks (nonlinear score)	16.3	82.41

Результаты

Algorithm	MAP@20 %	% of max
best of two (initial minimum)	14.2	72.32
“king of the hill”, take 1 alg.	16.8	80.26
“king of the hill”, mixing of ranks (linear score)	16.5	82.16
“king of the hill”, mixing of ranks (nonlinear score)	16.3	82.41
“king of the hill”, mixing of normalized scores	17.3	84.55

Результаты

Algorithm	MAP@20 %	% of max
best of two (initial minimum)	14.2	72.32
“king of the hill”, take 1 alg.	16.8	80.26
“king of the hill”, mixing of ranks (linear score)	16.5	82.16
“king of the hill”, mixing of ranks (nonlinear score)	16.3	82.41
“king of the hill”, mixing of normalized scores	17.3	84.55
max of two (possible maximum)	20.2	100

Решение проблем

Лог юзеров

о контенте

о юзерах

Mean Average Precision

Catalog coverage

User-based

Item-based

Time-based

SVD + NN

KNN

Day dependency

Heuristic

Решение проблем

Лог юзеров

о контенте

о юзерах

Mean Average Precision

Catalog coverage

User-based

Item-based

Time-based

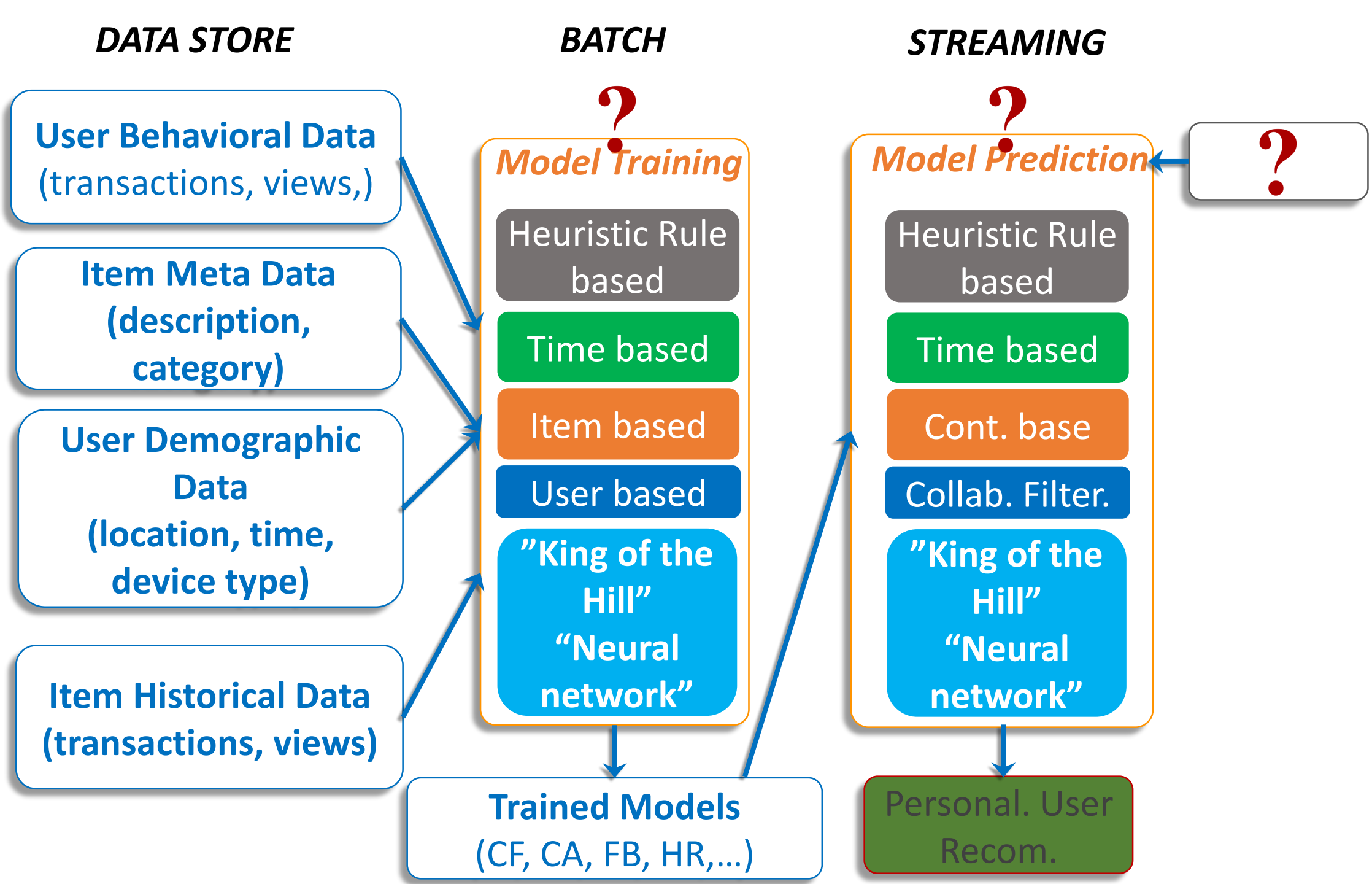
SVD + NN

KNN

Day dependency

Heuristic

King of the hill


DATA STORE
BATCH
STREAMING
User Behavioral Data
(transactions, views,)
Item Meta Data
(description,
category)
User Demographic
Data
(location, time,
device type)
Item Historical Data
(transactions, views)
Model Training
Heuristic Rule
based
Time based
Item based
User based
"King of the
Hill"
"Neural
network"
Trained Models
(CF, CA, FB, HR,...)
Model Prediction
Heuristic Rule
based
Time based
Cont. base
Collab. Filter.
"King of the
Hill"
"Neural
network"
Personal. User
Recom.

Модификации Personalize User

Признаки пользователя
(e.g. device type)

Признаки контекста
(e.g. meta data)

Распределение по времени
(Obsolescence of items)

Current item

Next item

Модификации Personalize User

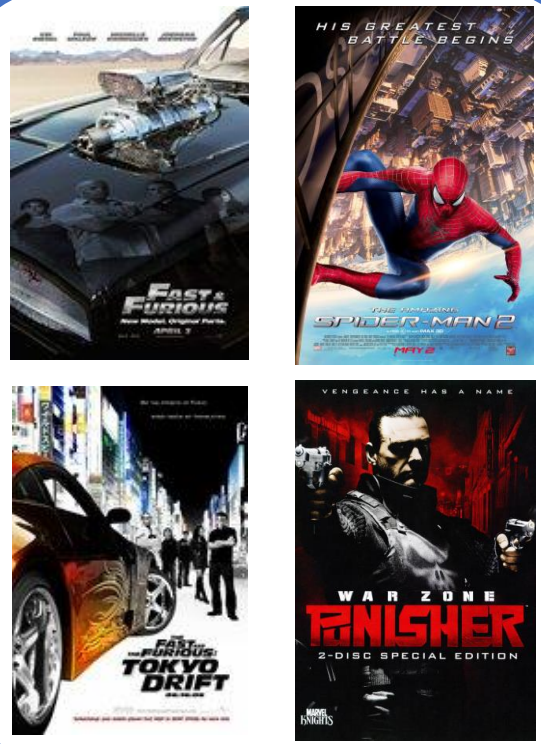
Признаки контекста
(e.g. meta data)

Распределение по времени
(Obsolescence of items)

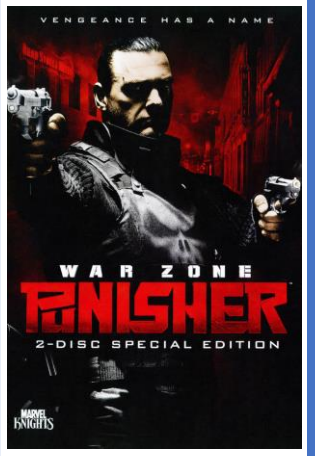
Признаки пользователя
(e.g. device type)

Hidden Layer

Masked layer

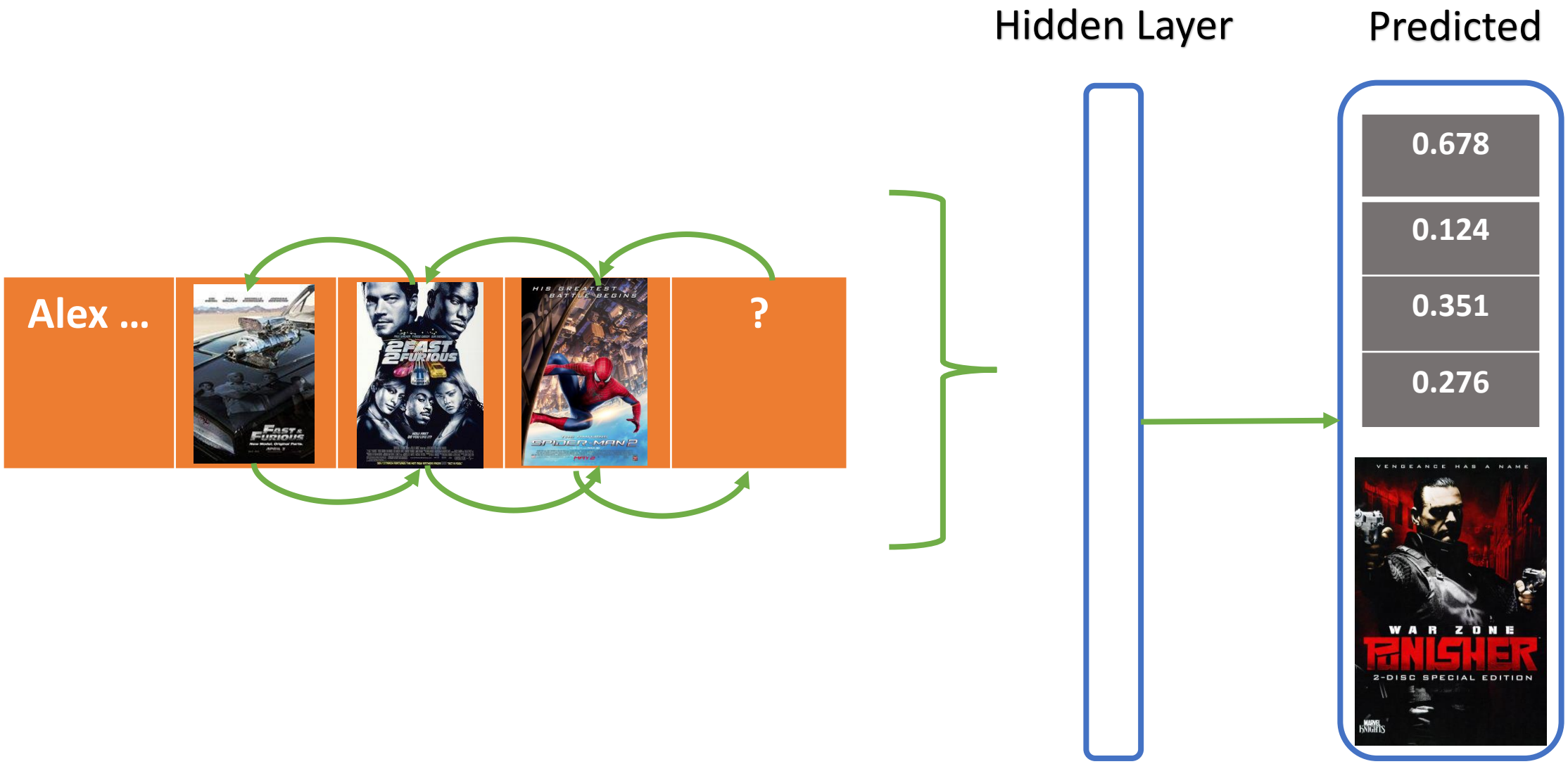


Predicted



Лучшие рекомендации
(коллаборативная фильтрация)

Модификации Personalize User



Модификации Personalize User

Model	MAP@20 %	Catalog coverage %
Baseline-top-20	2.4	0.2
random	0.0	5.9

Модификации Personalize User

Model	MAP@20 %	Catalog coverage %
Baseline-top-20	2.4	0.2
random	0.0	5.9
Full_con_probability	**10.6**	**7.2**

Модификации Personalize User

Model	MAP@20 %	Catalog coverage %
Baseline-top-20	2.4	0.2
random	0.0	5.9
Full_con_probability	**10.6**	**7.2**
Full_con_word2vec	0.0	5.2

Модификации Personalize User

Model	MAP@20 %	Catalog coverage %
Baseline-top-20	2.4	0.2
random	0.0	5.9
Full_con_probability	**10.6**	**7.2**
Full_con_word2vec	0.0	5.2
Recurent_word2vec	0.1	0.6

Решение проблем

Лог юзеров

о контенте

о юзерах

Mean Average Precision

Catalog coverage

User-based

Item-based

Time-based

SVD + NN

KNN

Day dependency

Heuristic

King of the hill

Решение проблем

Лог юзеров

о контенте

о юзерах

Mean Average Precision

Catalog coverage

User-based

Item-based

Time-based

SVD + NN

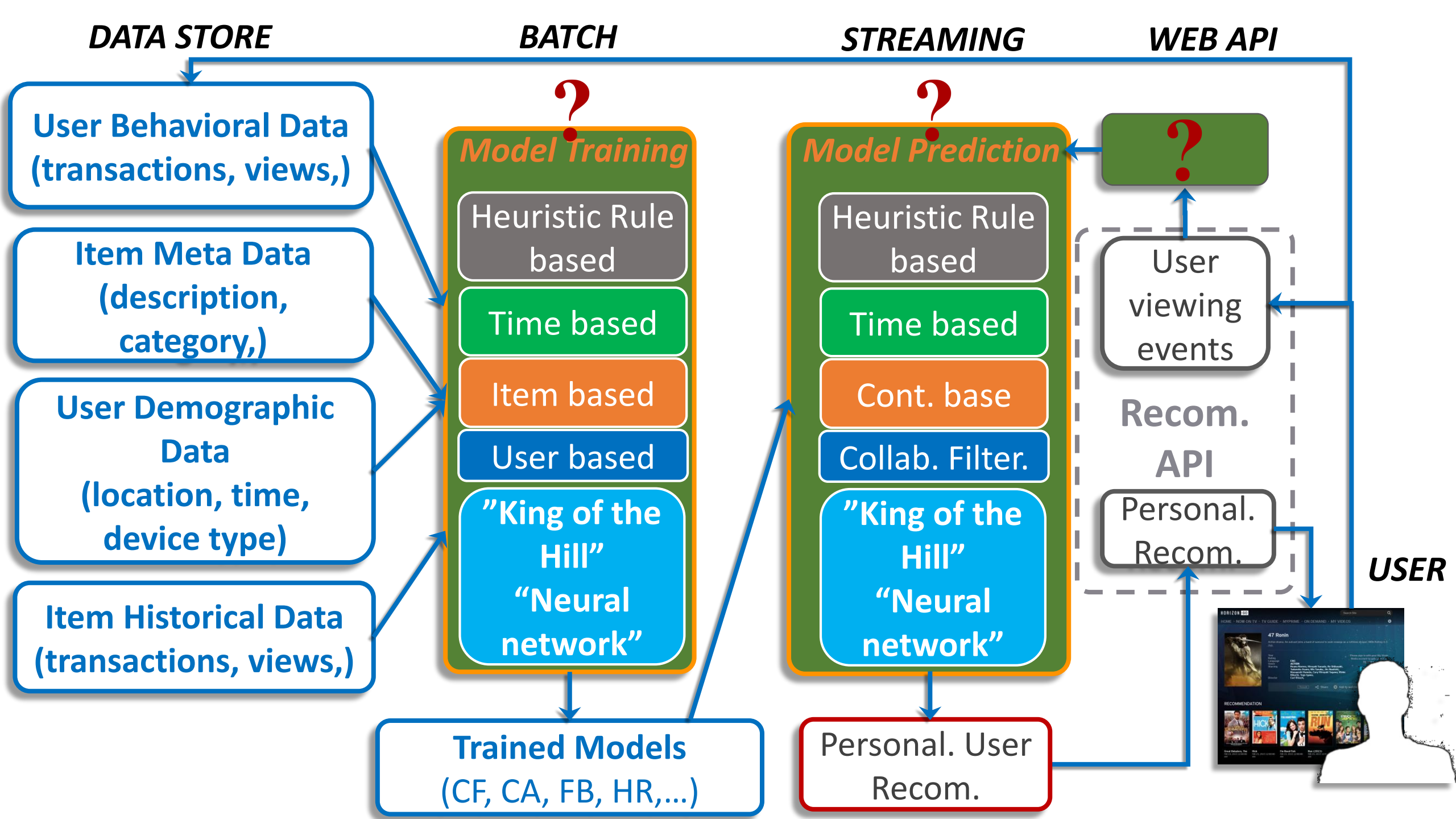
KNN

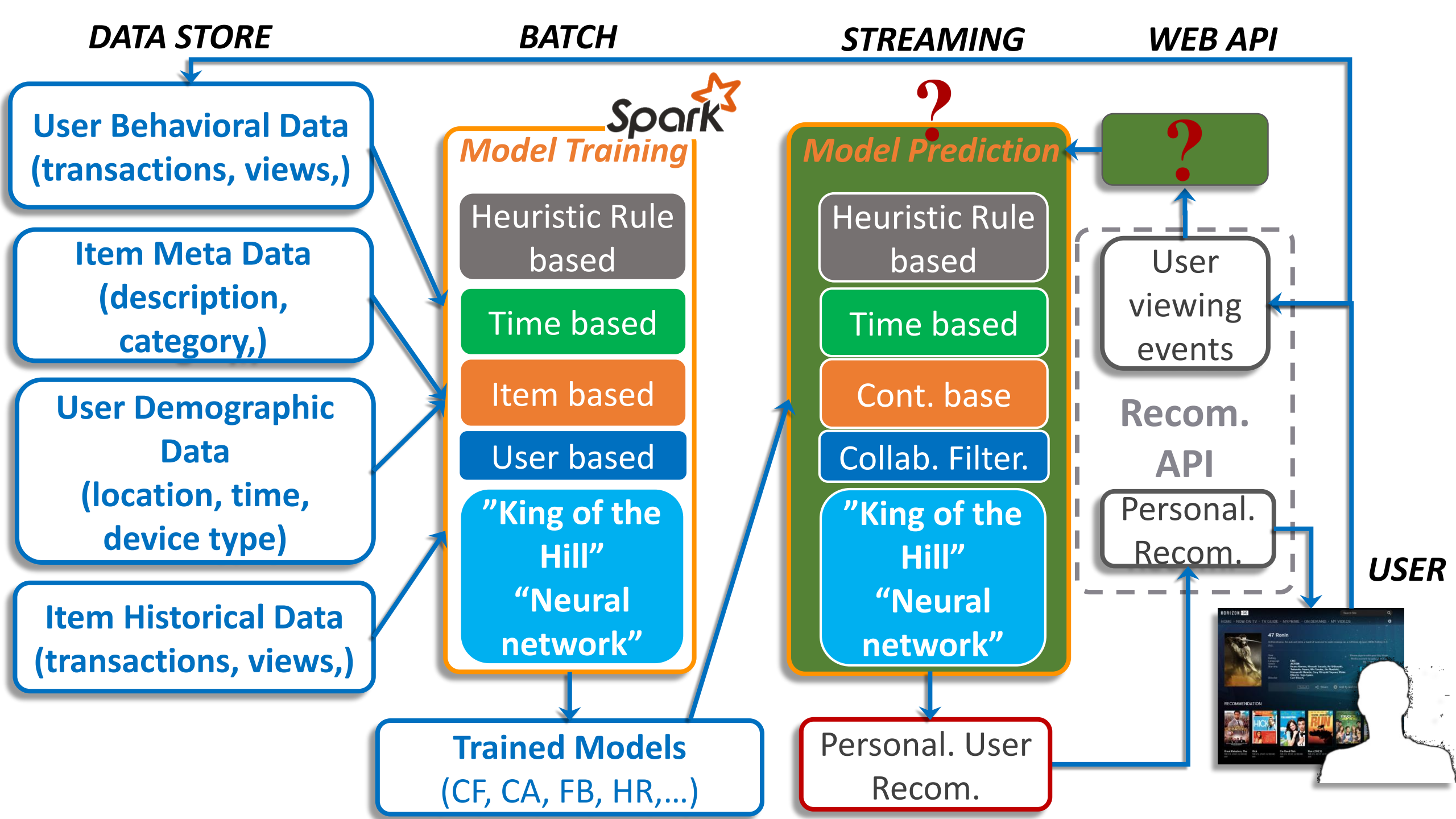
Day dependency

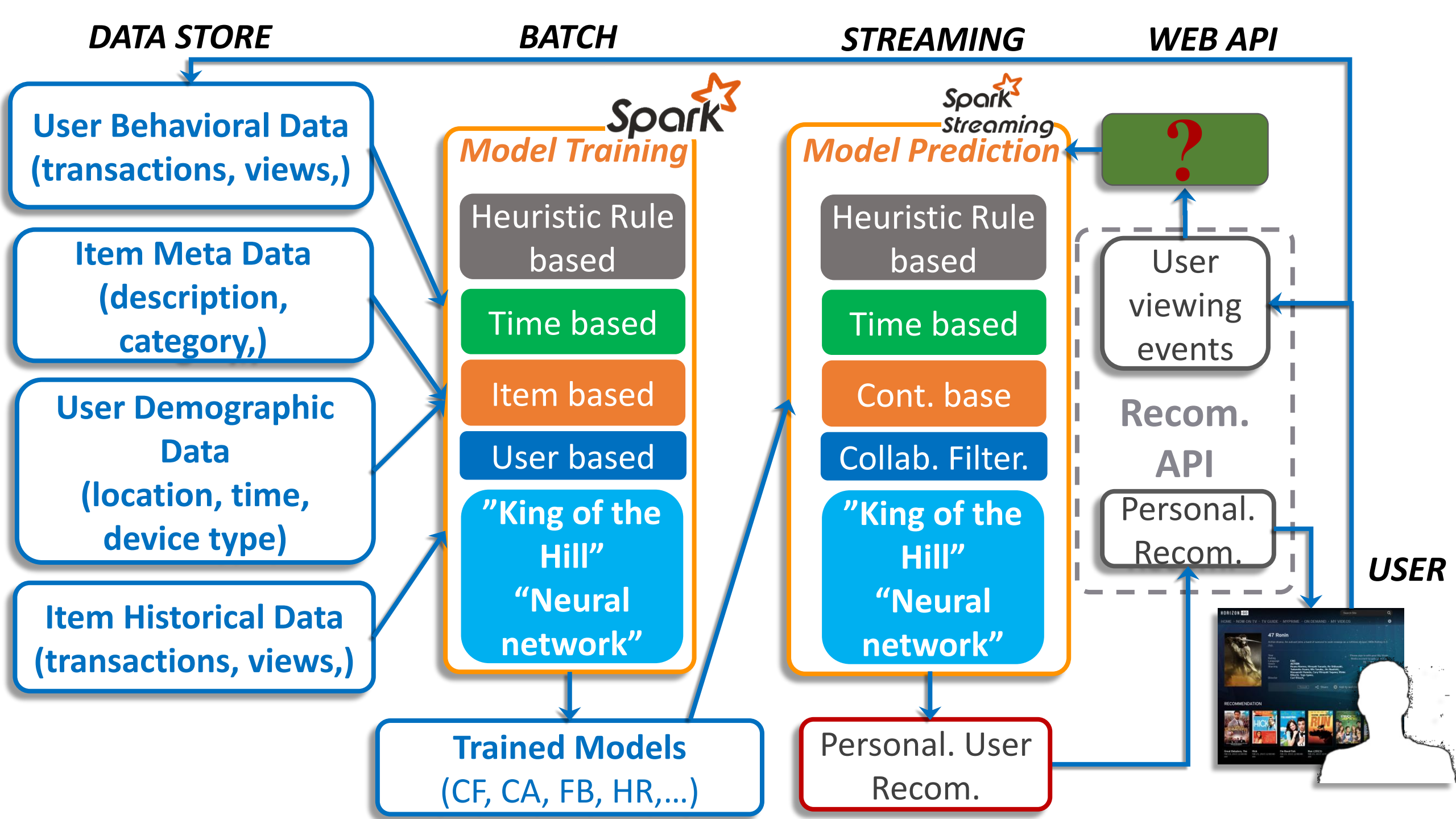
Heuristic

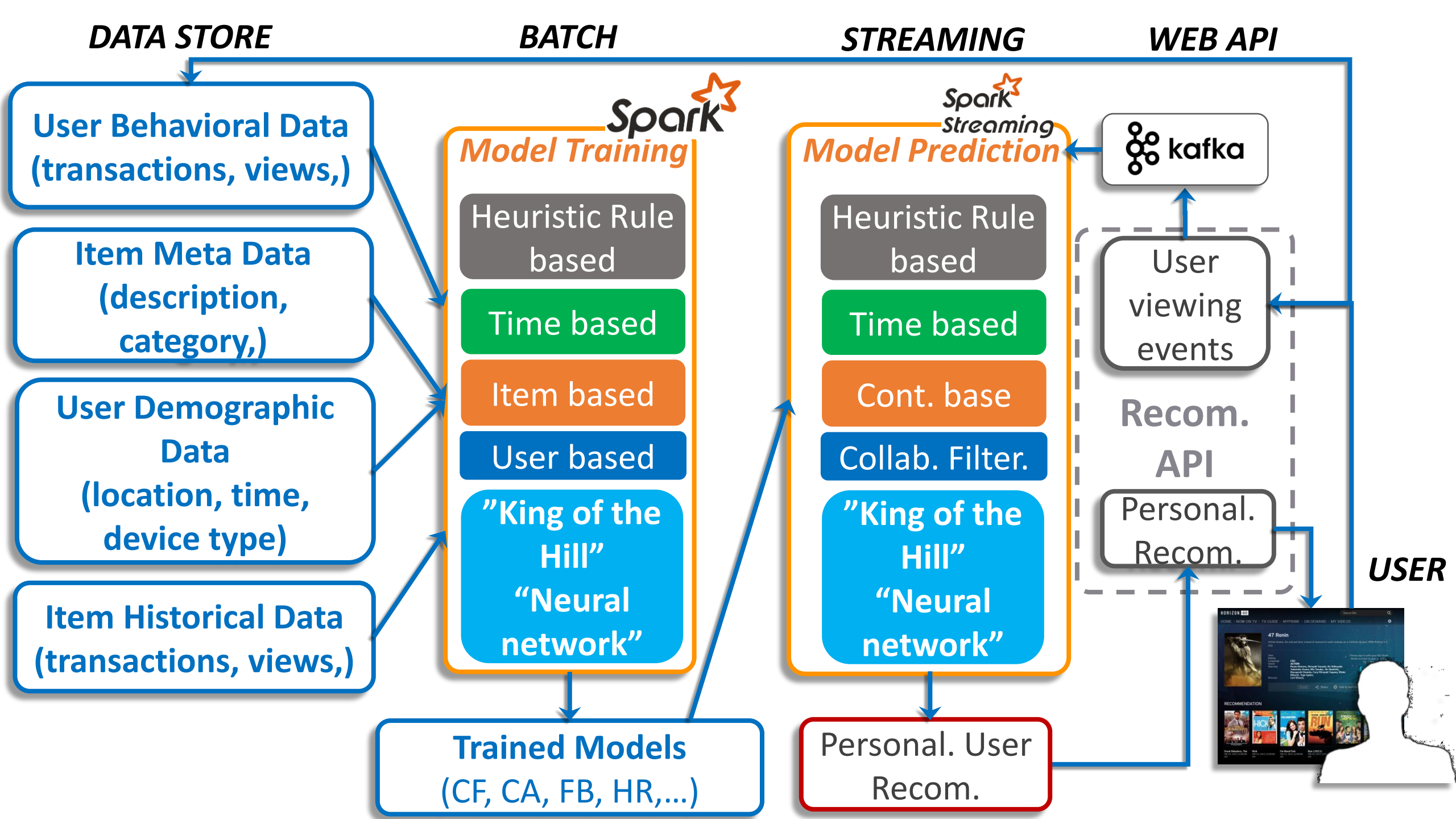
King of the hill

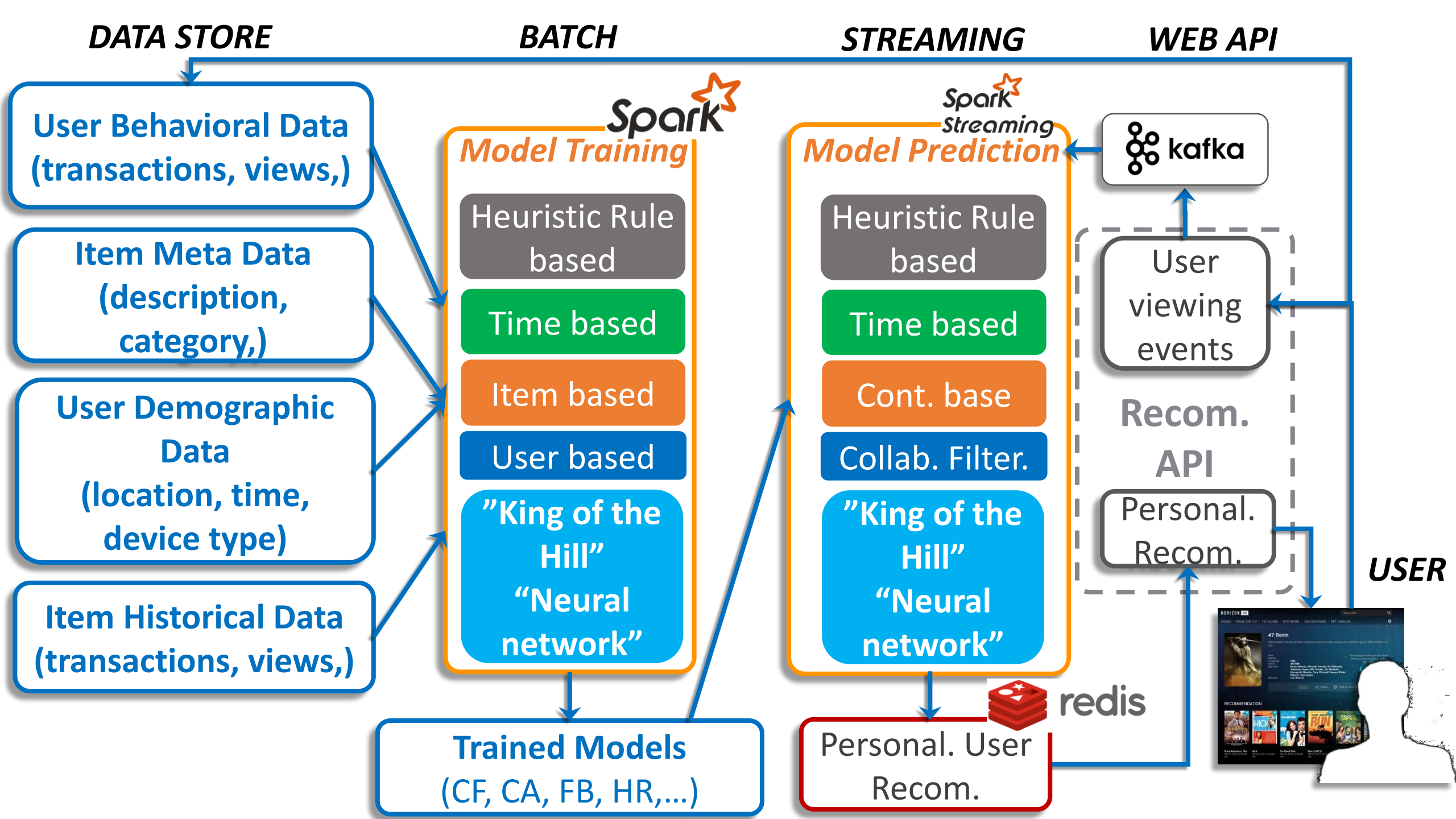
Neural network


DATA STORE
BATCH
STREAMING
WEB API
User Behavioral Data (transactions, views,)
Item Meta Data (description, category,)
User Demographic Data (location, time, device type)
Item Historical Data (transactions, views,)
Model Training
Heuristic Rule based
Time based
Item based
User based
"King of the Hill" "Neural network"
Model Prediction
Heuristic Rule based
Time based
Cont. base
Collab. Filter.
"King of the Hill" "Neural network"
User viewing events
Recom. API
Personal. Recom.
USER
Trained Models (CF, CA, FB, HR,...)
Personal. User Recom.


DATA STORE
BATCH
STREAMING
WEB API
User Behavioral Data (transactions, views,)
Item Meta Data (description, category,)
User Demographic Data (location, time, device type)
Item Historical Data (transactions, views,)
Spark
Model Training
Heuristic Rule based
Time based
Item based
User based
"King of the Hill" "Neural network"
Trained Models (CF, CA, FB, HR,...)
?
Model Prediction
Heuristic Rule based
Time based
Cont. base
Collab. Filter.
"King of the Hill" "Neural network"
Personal. User Recom.
?
User viewing events
Recom. API
Personal. Recom.
USER


DATA STORE
BATCH
STREAMING
WEB API
User Behavioral Data (transactions, views,)
Item Meta Data (description, category,)
User Demographic Data (location, time, device type)
Item Historical Data (transactions, views,)
Spark
Model Training
Heuristic Rule based
Time based
Item based
User based
"King of the Hill" "Neural network"
Trained Models (CF, CA, FB, HR,...)
Spark Streaming
Model Prediction
Heuristic Rule based
Time based
Cont. base
Collab. Filter.
"King of the Hill" "Neural network"
Personal. User Recom.
?
User viewing events
Recom. API
Personal. Recom.
USER


DATA STORE
BATCH
STREAMING
WEB API
User Behavioral Data (transactions, views,)
Item Meta Data (description, category,)
User Demographic Data (location, time, device type)
Item Historical Data (transactions, views,)
Spark
Model Training
Heuristic Rule based
Time based
Item based
User based
"King of the Hill" "Neural network"
Trained Models (CF, CA, FB, HR,...)
Spark Streaming
Model Prediction
Heuristic Rule based
Time based
Cont. base
Collab. Filter.
"King of the Hill" "Neural network"
Personal. User Recom.
kafka
User viewing events
Recom. API
Personal. Recom.
USER


DATA STORE
BATCH
STREAMING
WEB API
User Behavioral Data (transactions, views,)
Item Meta Data (description, category,)
User Demographic Data (location, time, device type)
Item Historical Data (transactions, views,)
Spark
Model Training
Heuristic Rule based
Time based
Item based
User based
"King of the Hill" "Neural network"
Trained Models (CF, CA, FB, HR,...)
Spark Streaming
Model Prediction
Heuristic Rule based
Time based
Cont. base
Collab. Filter.
"King of the Hill" "Neural network"
redis
Personal. User Recom.
kafka
User viewing events
Recom. API
Personal. Recom.
USER

Проблемы с потоками

- Тест работы с потоком:

~ 800 т. пользователей

~ 12 млн строк логов

- Item based: K – near neighbors (~ 2 Gb) vs LSH Forest (~ 170 m)
- Предсказание нейронными сетями в поточном режиме
  - Сохранение графа и весов (Tensorflow)
  - Tensorflow (~ 200 Mb) vs Keras (~ 700 Mb)

Модель	Батч/Время
Tensorflow	100/~ 100 sec
Keras	100/~ 180 sec
KNN	100/~ 210 sec (NaN)
LSH Forest	100/~ 50 sec
SVD	100/~ 140 sec

Выводы

- Определитесь с архитектурой системы
- Подберите метрики, исходя из ваших данных и задачи
- Не жалейте времени на ручную проверку эвристик
- Экспериментируйте и модифицируйте state of the art алгоритмы
- Обратите внимание на комбинирование алгоритмов

Выводы

Спасибо за внимание!