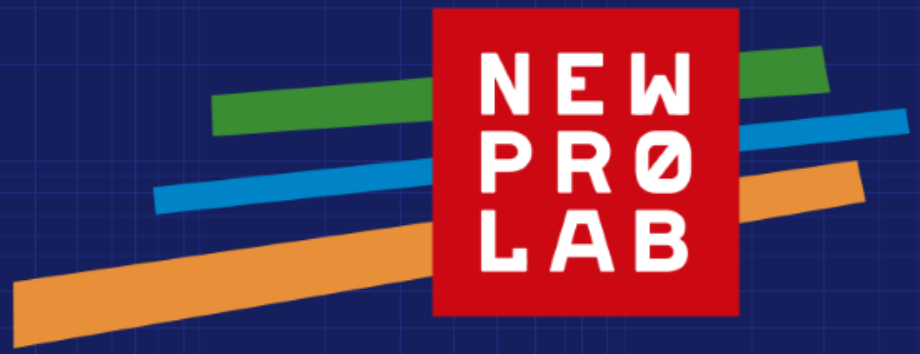


SVD

singular value decomposition

NEWPROLAB.COM

План выступления

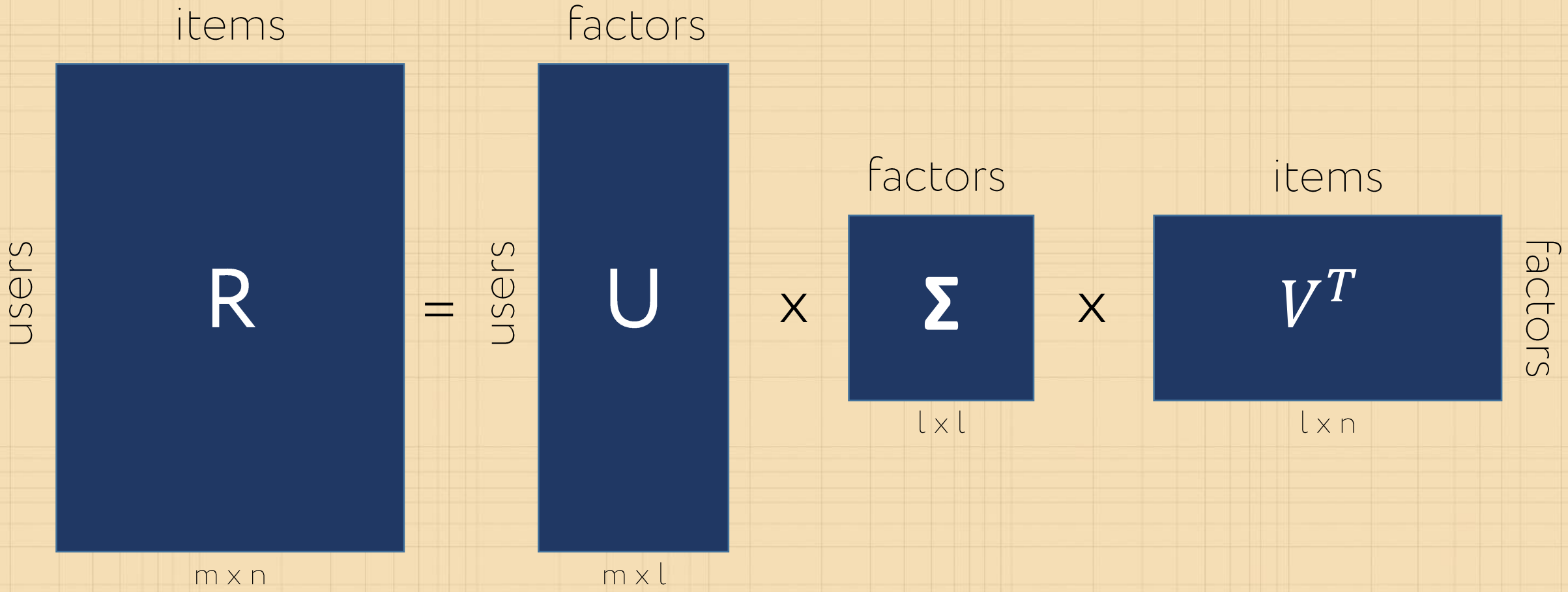
1. Зачем нужны эти 3 буквы?
2. Суть алгоритма
3. Интерпретация на пальцах
4. Небольшой пример в Python
5. ALS – еще 3 буквы


SVD
BAHEM

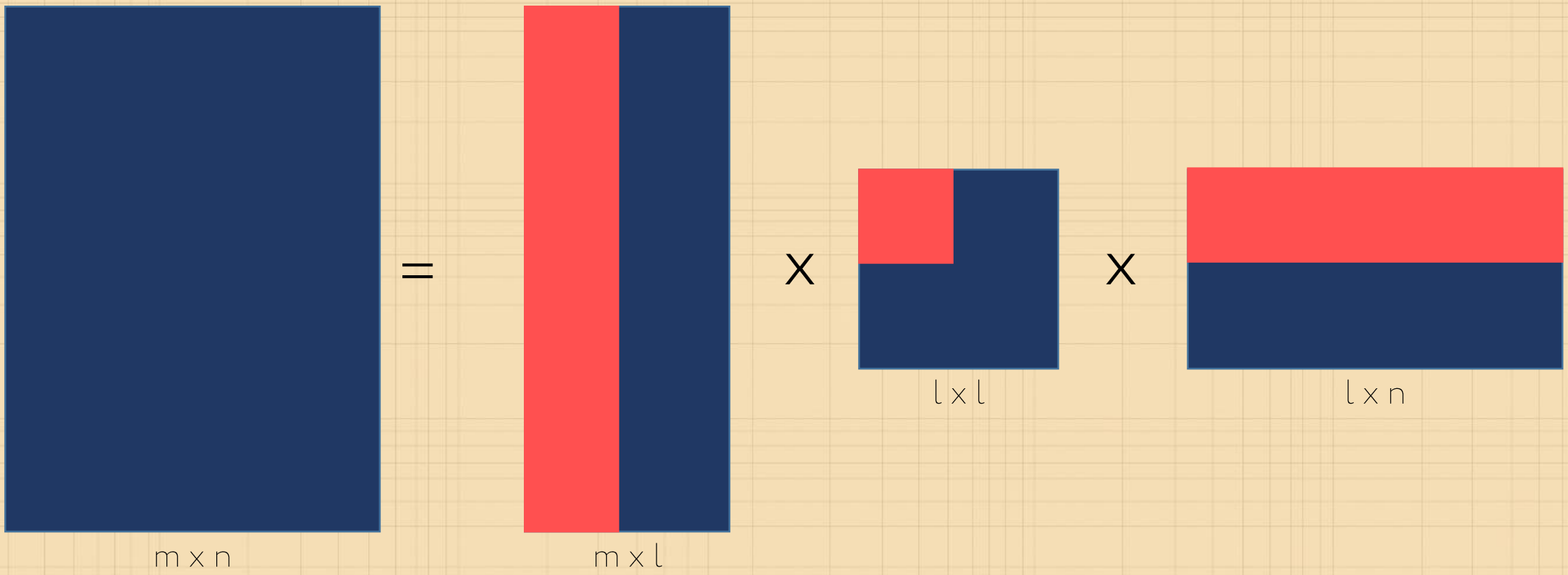
Сценарии использования

1. Слишком большая размерность данных – хочется ее снизить. Находится в одном ряду с PCA, например.
2. Выявить скрытые факторы в данных, которые можно дальше как-то использовать.

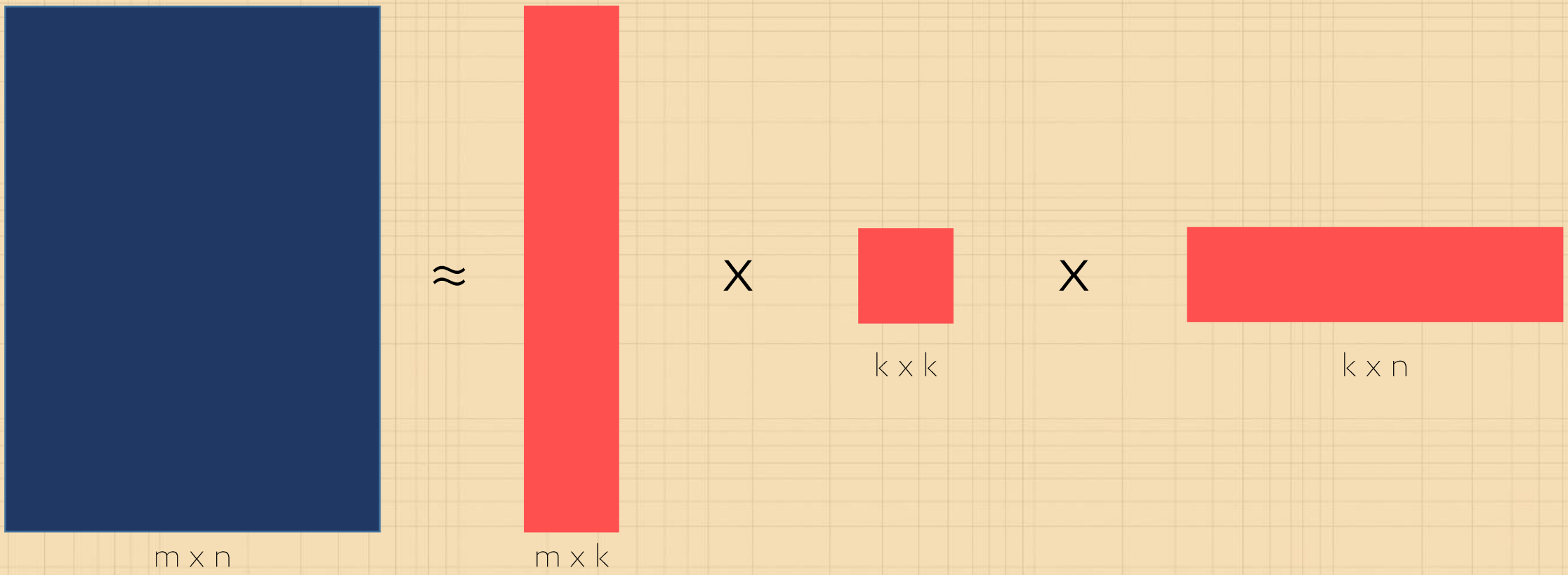
Суть алгоритма



Суть алгоритма



Суть алгоритма

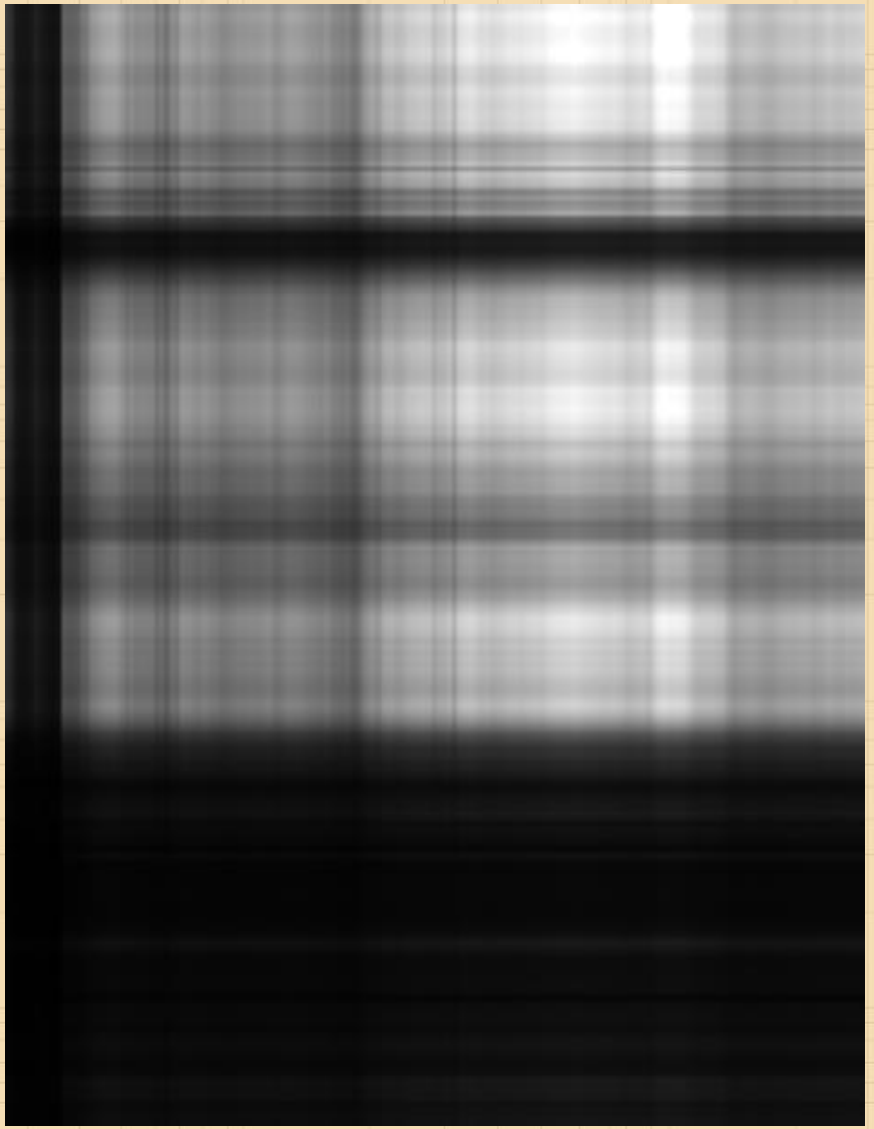


Интерпретация

Картинка ч/б – каждую строку пикселей можно представить как строка числовых значений.

Каждый пиксель – ячейка, в которой указано значение интенсивности.

Интерпретация



Применили SVD и взяли только первый главный фактор.

k = 1

Интерпретация

Взяли 2 главных фактора.

k = 2

Интерпретация

k = 10

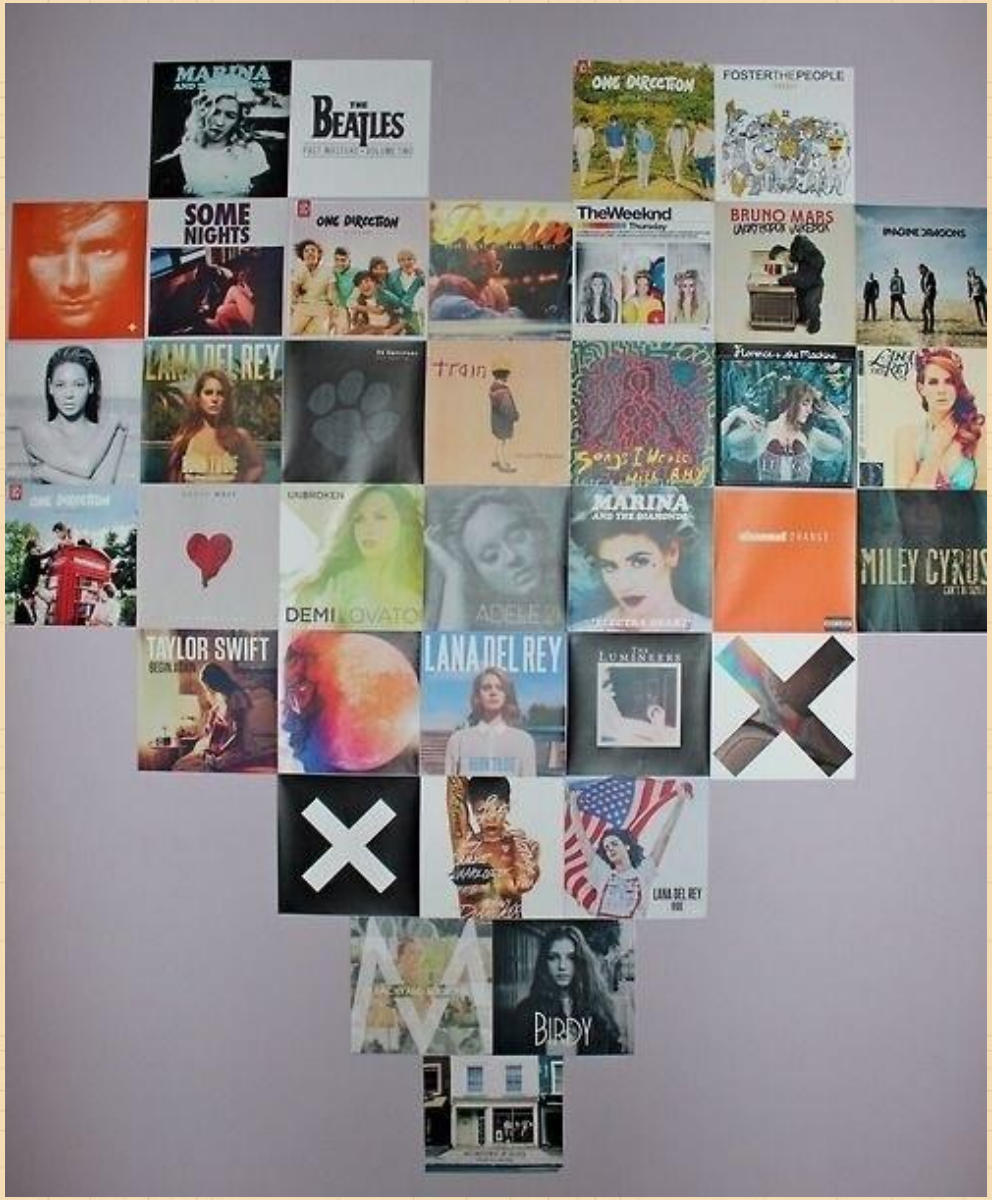
Видите Фейнмана?

Интерпретация

k = 50

Круто?
Исходный размер был: 475x620.

Интерпретация



Мы пытаемся на самом деле узнать вкус пользователя. И пытаемся это сделать, глядя на данные.

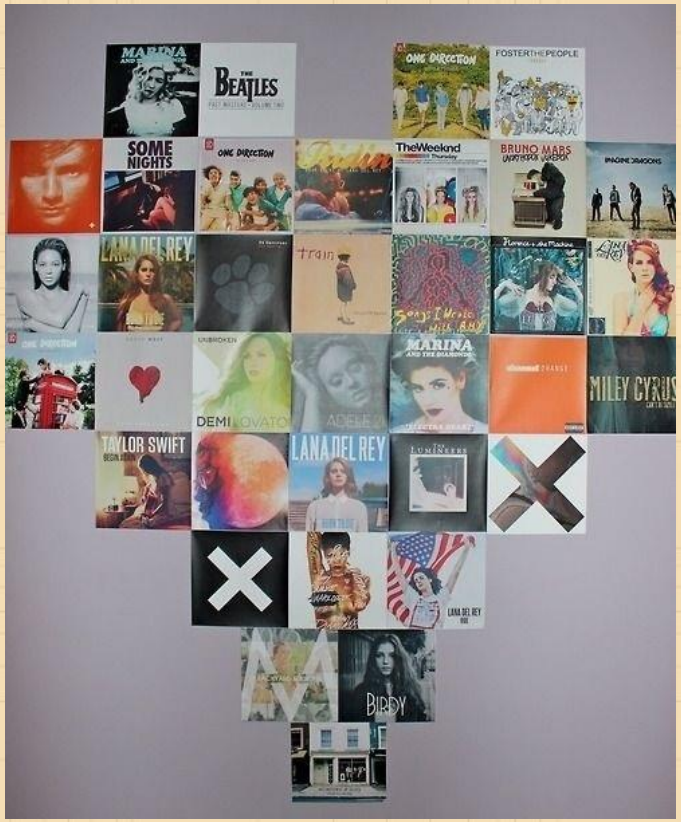
Интерпретация

Есть другой вариант – спросить…

Интерпретация

Есть другой вариант – спросить…
И она скажет: «Я люблю современную поп-музыку».

Сокращение размерности



Я люблю современную поп-музыку

Пример

	Yesterday, Beatles	Summertime Sadness, Lana Del Rey	November Rain, Guns 'n Roses	Diamonds, Rihanna	Highway to Hell, AC/DC	What a Wonderful World, Louis Armstrong	Hit the Road Jack!, Ray Charles
Артем	5	2	5	2	5	4	4
Вася	4	3	3	3	3	5	5
Маша	3	5	2	5	2	4	3
Саша	5	2	4	2	5	4	4
Клара	5	5	3	4	3	5	5

Пример

U

	f1	f2	f3
Артем	0.3262	0.5236	-0.455
Вася	-0.719	-0.052	-0.44
Маша	0.5477	-0.644	-0.392
Саша	0.1567	0.4523	-0.44
Клара	-0.23	-0.322	-0.503

$\sum$

	f1	f2	f3
f1	22.73	0	0
f2	0	5.3211	0
f3	0	0	1.7328

V^T

	Yesterday, Beatles	Summertime Sadness, Lana Del Rey	November Rain, Guns 'n Roses	Diamonds, Rihanna	Highway to Hell, AC/DC	What a Wonderful World, Louis Armstrong	Hit the Road Jack!, Ray Charles
f1	0.0205983	0.23140511	0.29370711	0.36389253	0.38414103	-0.35664957	-0.67274252
f2	0.2122569	-0.57029559	0.37908873	-0.50977121	0.46409138	-0.10192274	0.01912943
f3	-0.43649738	-0.33353455	-0.33632132	-0.31142383	-0.35567176	-0.43374969	-0.41651737

Пример

f1 – не джаз
f2 – рок и не поп
f3 – не очень информативен

U

	f1	f2	f3
Артем	0.3262	0.5236	-0.455
Вася	-0.719	-0.052	-0.44
Маша	0.5477	-0.644	-0.392
Саша	0.1567	0.4523	-0.44
Клара	-0.23	-0.322	-0.503

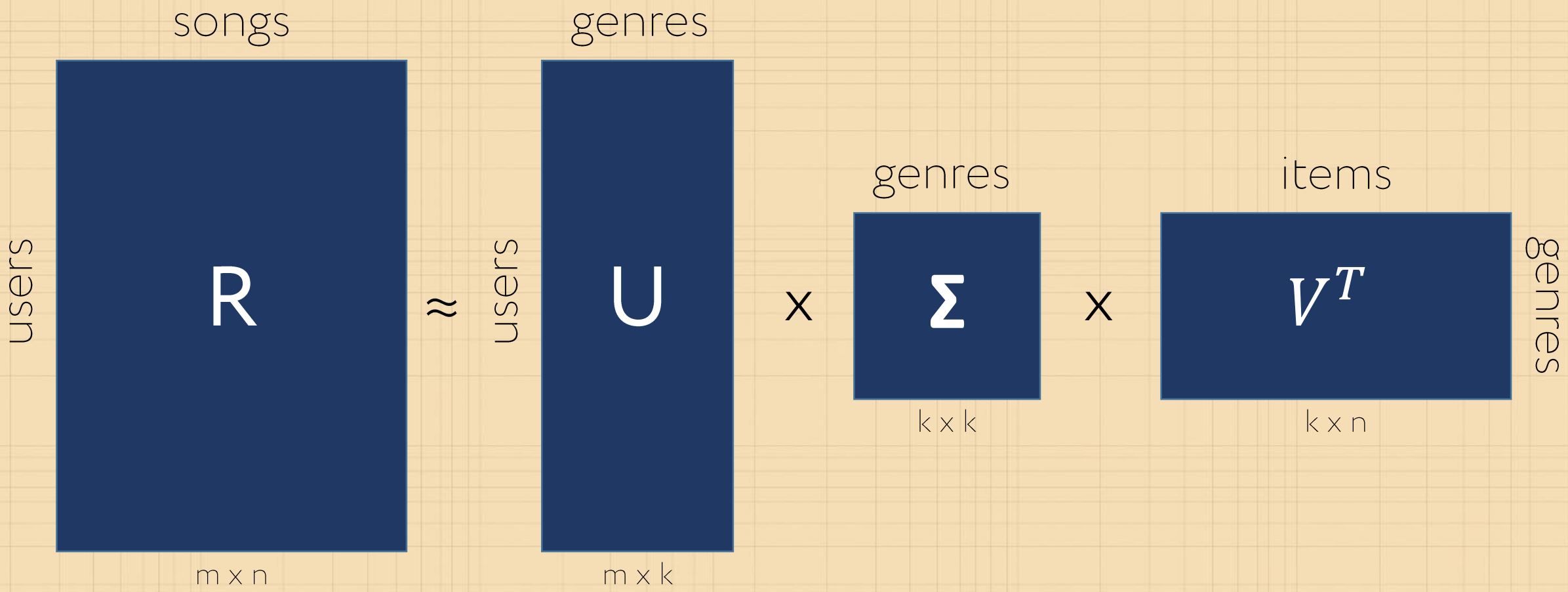
$\sum$

	f1	f2	f3
f1	22.73	0	0
f2	0	5.3211	0
f3	0	0	1.7328

V^T

	Yesterday, Beatles	Summertime Sadness, Lana Del Rey	November Rain, Guns 'n Roses	Diamonds, Rihanna	Highway to Hell, AC/DC	What a Wonderful World, Louis Armstrong	Hit the Road Jack!, Ray Charles
f1	0.0205983	0.23140511	0.29370711	0.36389253	0.38414103	-0.35664957	-0.67274252
f2	0.2122569	-0.57029559	0.37908873	-0.50977121	0.46409138	-0.10192274	0.01912943
f3	-0.43649738	-0.33353455	-0.33632132	-0.31142383	-0.35567176	-0.43374969	-0.41651737

Интерпретация



Пример

k	RMSE
4	0.095
3	0.173
2	0.340
1	0.961

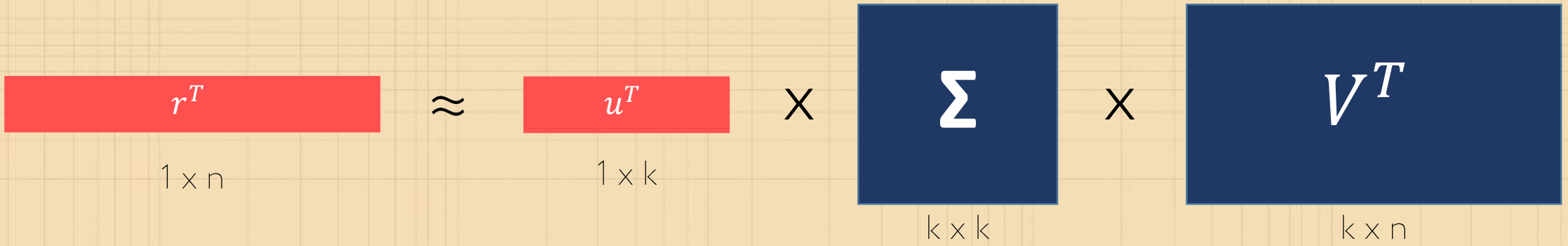
Восстановленная матрица

	Yesterday, Beatles	Summertime Sadness, Lana Del Rey	November Rain, Guns 'n Roses	Diamonds, Rihanna	Highway to Hell, AC/DC	What a Wonderful World, Louis Armstrong	Hit the Road Jack!, Ray Charles
Артем	5	2	5	2	5	4	4
Вася	4	3	3	3	3	5	5
Маша	3	5	2	5	2	4	3
Саша	5	2	4	2	5	4	4
Клара	5	5	3	4	3	5	5

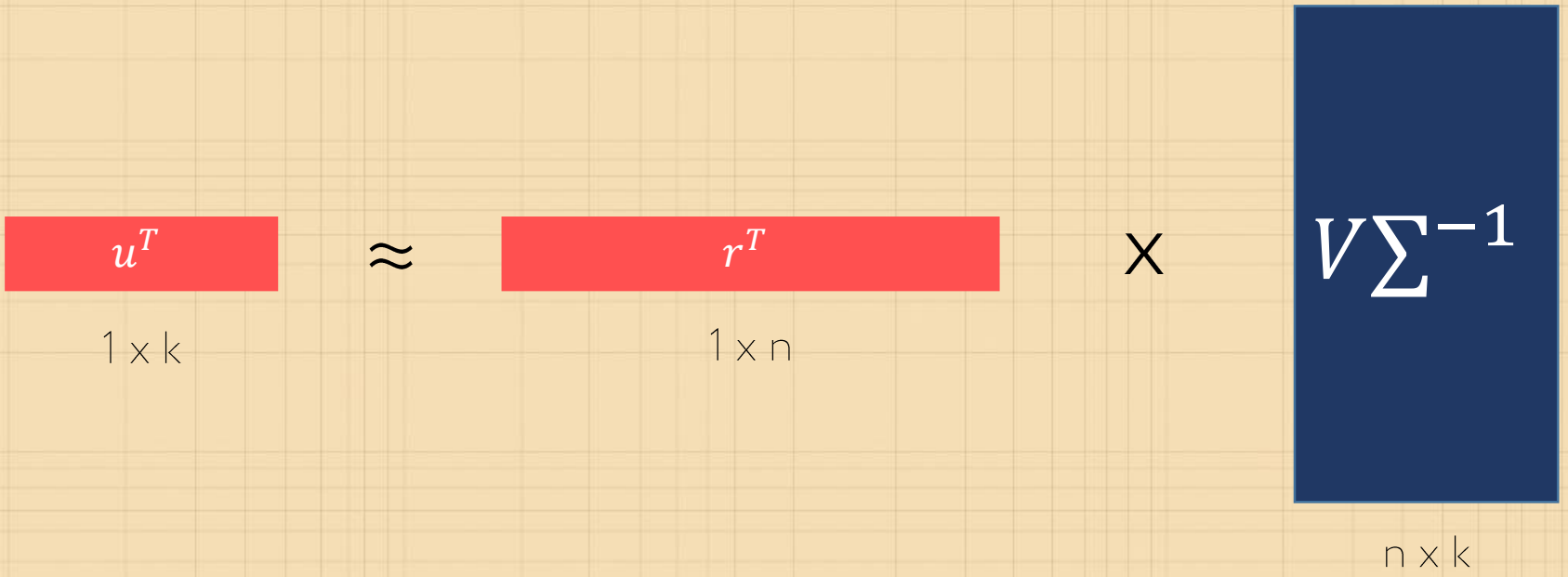
	Yesterday, Beatles	Summertime Sadness, Lana Del Rey	November Rain, Guns 'n Roses	Diamonds, Rihanna	Highway to Hell, AC/DC	What a Wonderful World, Louis Armstrong	Hit the Road Jack!, Ray Charles
Артем	5	2	5	2	5	4	4
Вася	4	3	3	3	3	4	4
Маша	3	5	2	5	2	4	4
Саша	5	2	4	2	5	4	4
Клара	5	5	3	4	3	5	5

$k = 2$

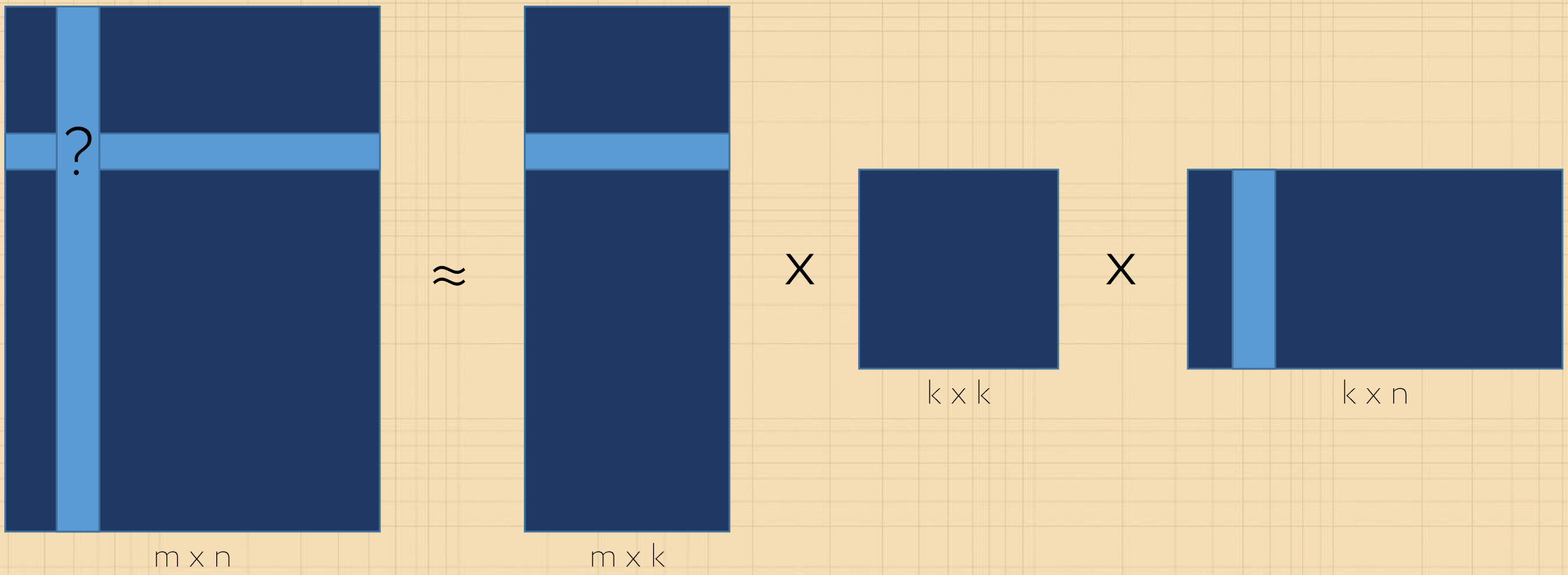
Как обновляется профиль?



Как обновляется профиль?



Как делать рекомендации?



SVD++

базовые предикторы
user и item

$$\hat{r}_{u,i} = \mu + b_u + b_i + u_u^T v_i$$

глобальное
среднее

профили user и item в
пространстве факторов

Недостатки

1. Медленный
2. Не работает с пропущенными значениями

Достоинства

1. Часто делает более осмысленные рекомендации. В CF юзеры, которые смотрели разные части одного сиквела, будут непохожи по косинусной мере. При помощи SVD – будут.

Пример в Python

датасет movielens100k


ALS
BAHEM

ALS – alternating least squares

Хьюстон, у нас оптимизационная проблема:

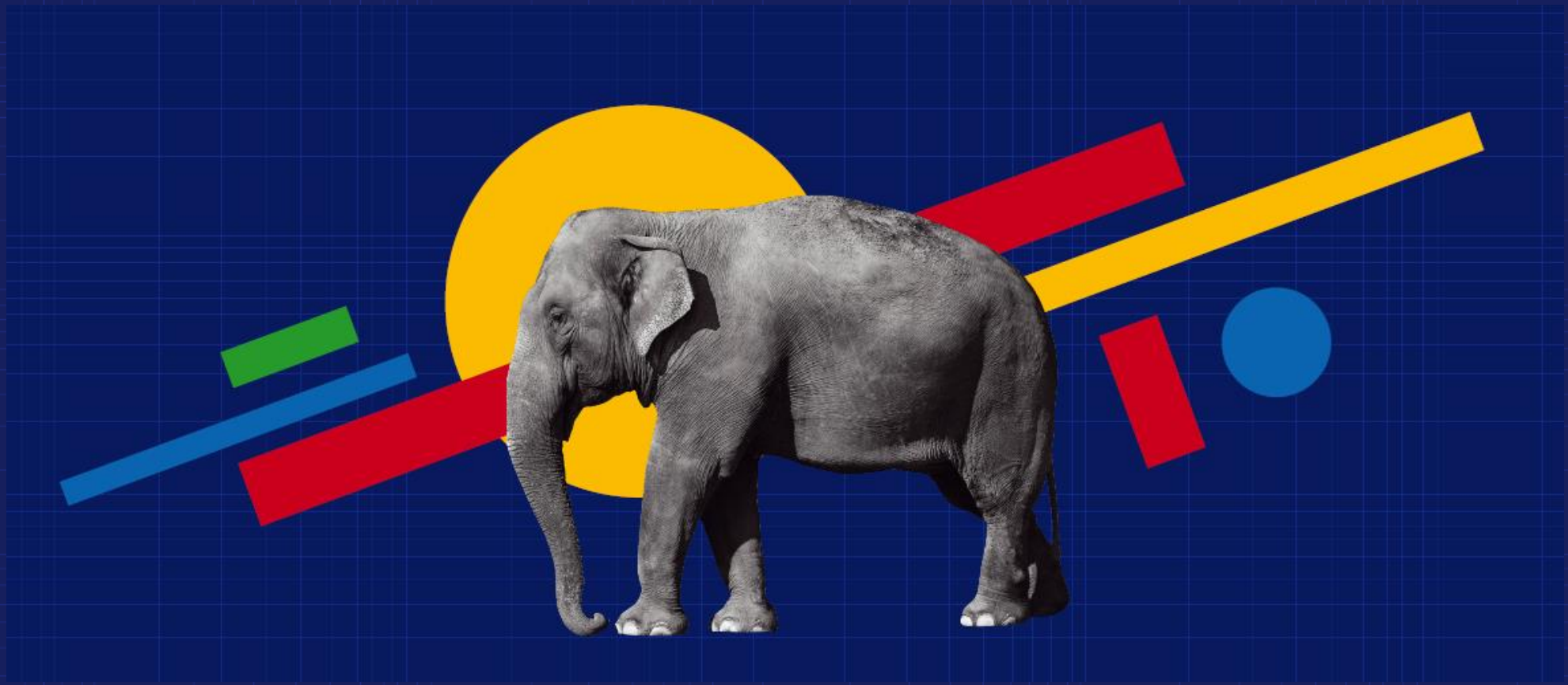
$$f = \left(r_{u,i} - \hat{r}_{u,i}\right)^2 + \lambda \sum \theta^2 \rightarrow min$$

ALS – alternating least squares

Хьюстон, у нас оптимизационная проблема:

$$f = \left(r_{u,i} - \hat{r}_{u,i}\right)^2 + \lambda \sum \theta^2 \rightarrow min$$

$$f = \left(r_{u,i} - u_u^T v_i\right)^2 + \lambda \left(\sum \|u_u^2\| + \sum \|v_i^2\| \right) \rightarrow min$$



надо есть слона по частям

ALS – alternating least squares

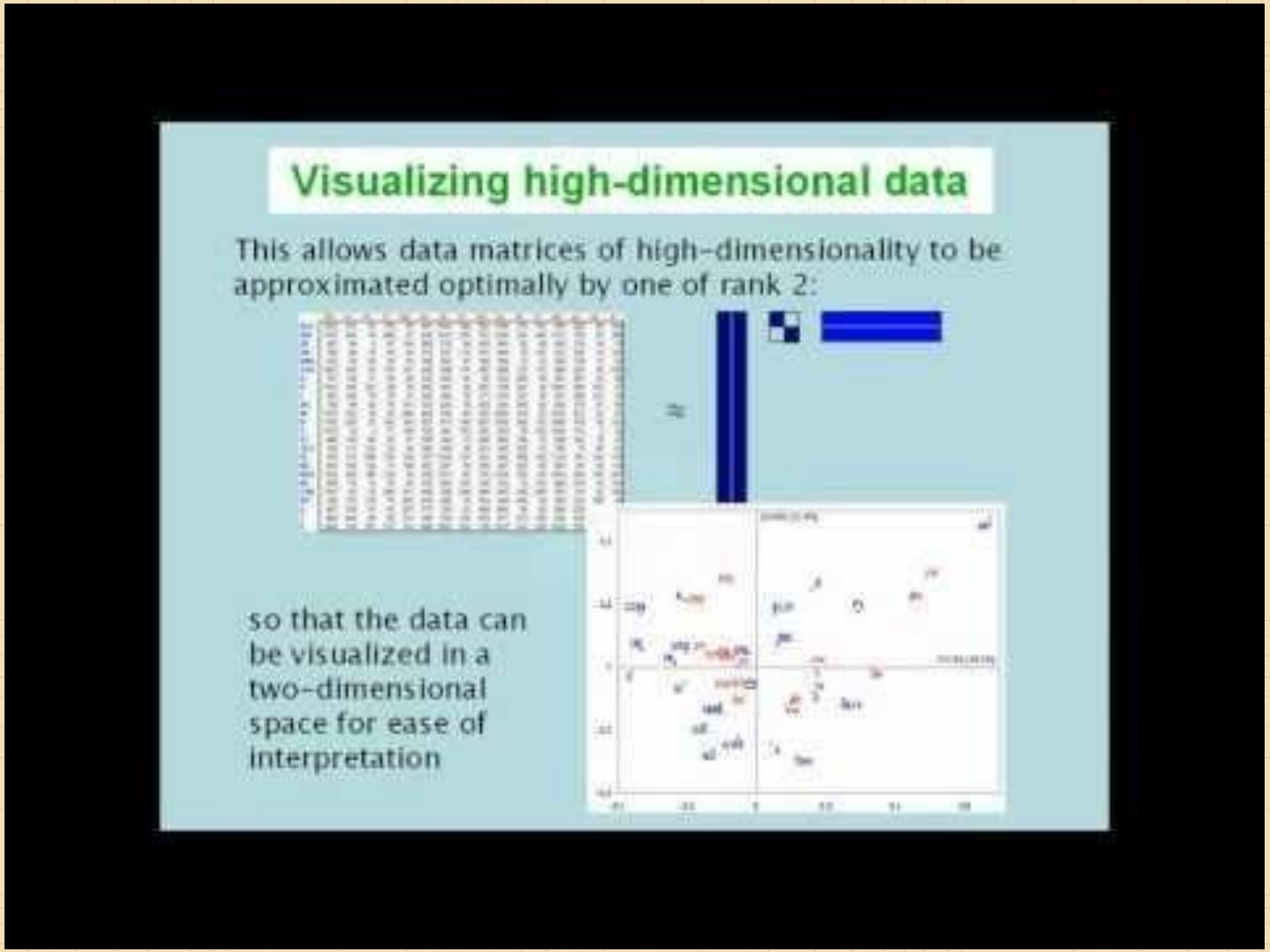
На каждой итерации давайте менять что-то одно:

$$f = \left(r_{u,i} - u_u^T v_i\right)^2 + \lambda\left(\sum \|u_u^2\| + \sum \|v_i^2\|\right) \to min$$

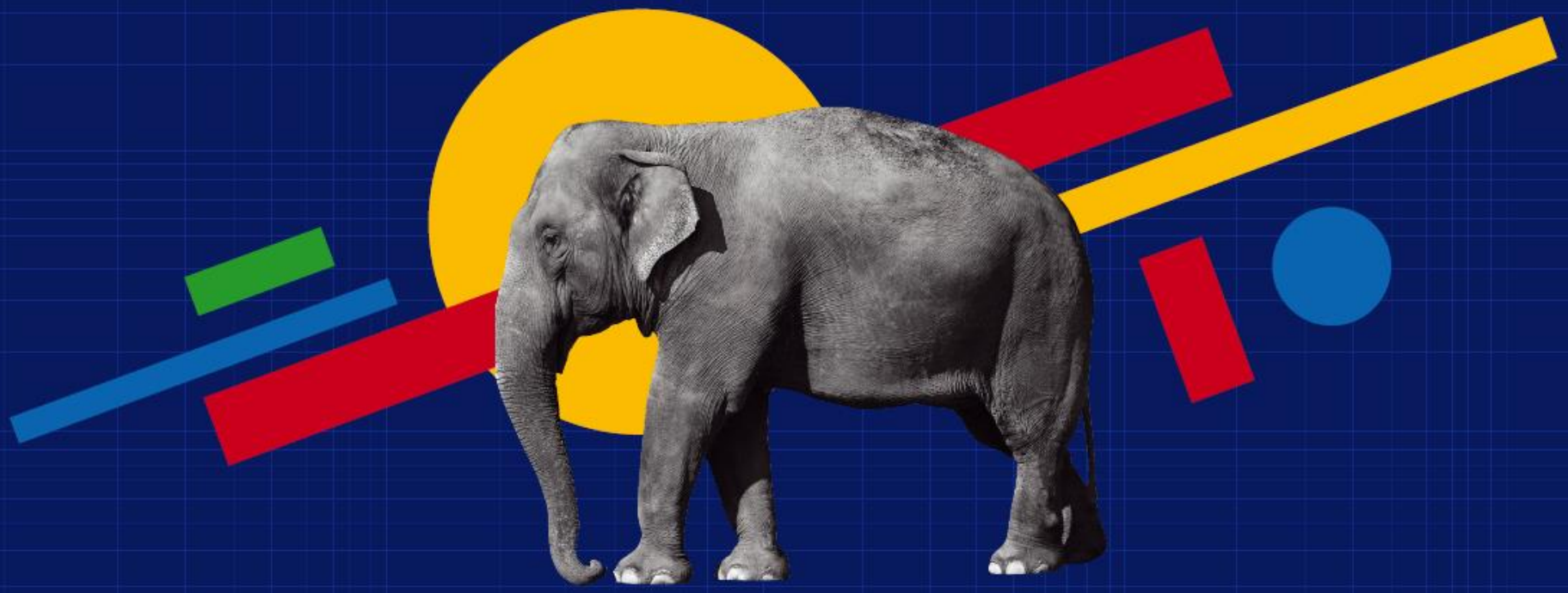
Преимущества ALS

1. Можно параллелить.
2. Есть реализация в Apache Spark!
3. Рекомендуют использовать для предсказания неявных рейтингов.

На десерт



https://www.youtube.com/watch?v=JEYLfIVvR9I



Big Data is Love

NEWPROLAB.COM