Рекомендательные системы и Большие Данные

Авторы доклада:

к.т.н., Сергей Папулин

Ольга Недильченко

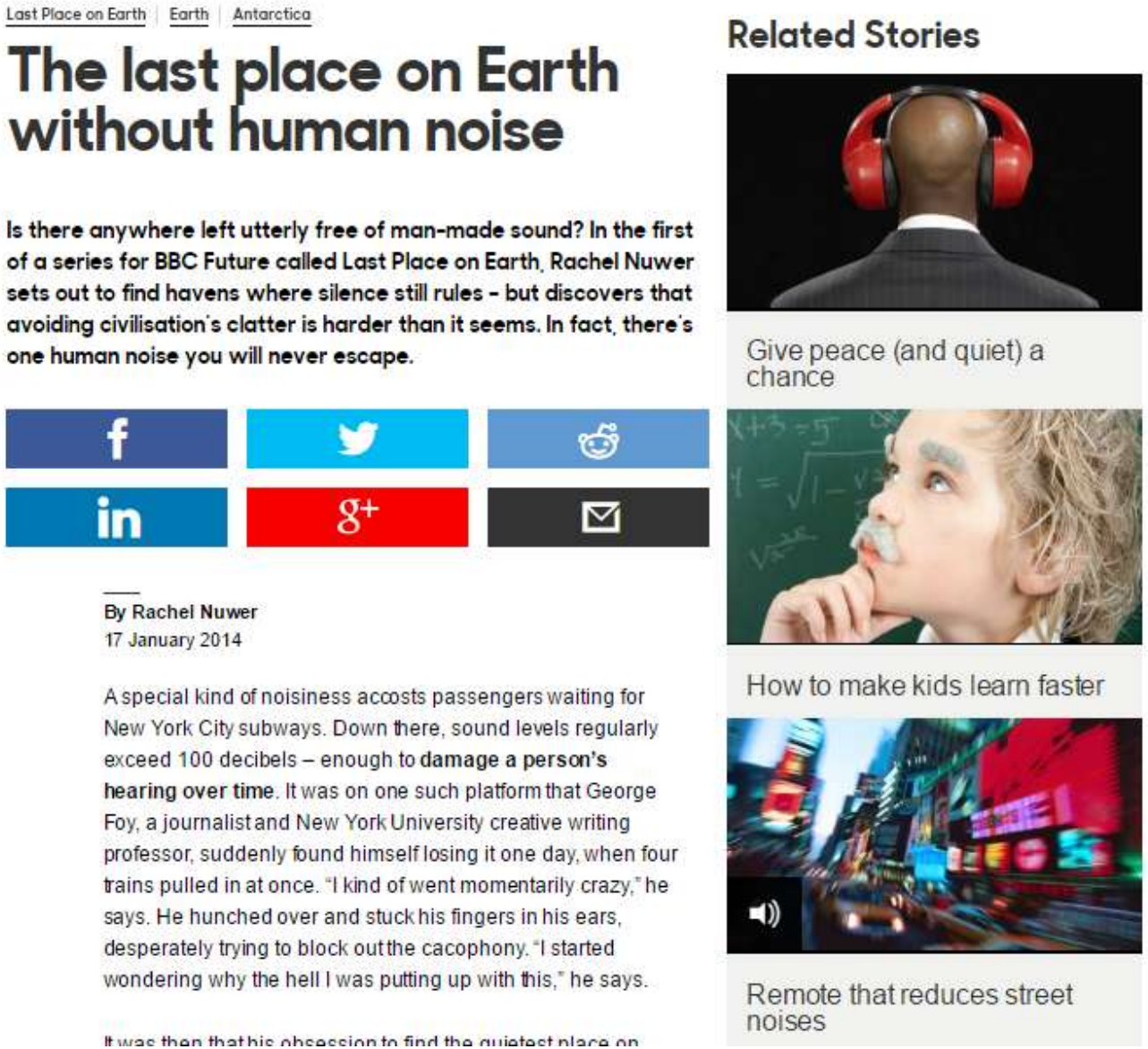
МГТУ им. Н.Э. Баумана, Москва, 2016

План доклада

- Виды рекомендательных систем
- Распределенная контентная фильтрация
- Распределенная коллаборативная фильтрация
- Распределенная факторизация матрицы рейтингов
- Решения на базе Spark

Виды рекомендательных систем

Виды рекомендательных систем



Last Place on Earth | Earth | Antarctica

The last place on Earth without human noise

Is there anywhere left utterly free of man-made sound? In the first of a series for BBC Future called Last Place on Earth, Rachel Nuwer sets out to find havens where silence still rules - but discovers that avoiding civilisation's clatter is harder than it seems. In fact, there's one human noise you will never escape.

By Rachel Nuwer
17 January 2014

A special kind of noisiness accosts passengers waiting for New York City subways. Down there, sound levels regularly exceed 100 decibels – enough to **damage a person's hearing over time**. It was on one such platform that George Foy, a journalist and New York University creative writing professor, suddenly found himself losing it one day, when four trains pulled in at once. "I kind of went momentarily crazy," he says. He hunched over and stuck his fingers in his ears, desperately trying to block out the cacophony. "I started wondering why the hell I was putting up with this," he says.

It was then that his obsession to find the quietest place on

Related Stories

Give peace (and quiet) a chance

How to make kids learn faster

Remote that reduces street noises

Content-based (основанные на контенте)
рекомендации на основе предыдущих оценок пользователя и схожести объектов

Виды рекомендательных систем

Customers Who Bought This Item Also Bought

Recommender Systems: The Textbook
› Charu C. Aggarwal
Hardcover
$63.20 Prime

Statistical Methods for Recommender Systems
Deepak K. Agarwal
★★★★★ 1
Hardcover
$56.99 Prime

Collaborative filtering (основанные на коллаборативной фильтрации) рекомендации на основе оценок других пользователей

Виды рекомендательных систем

Sponsored Links (What's this?)

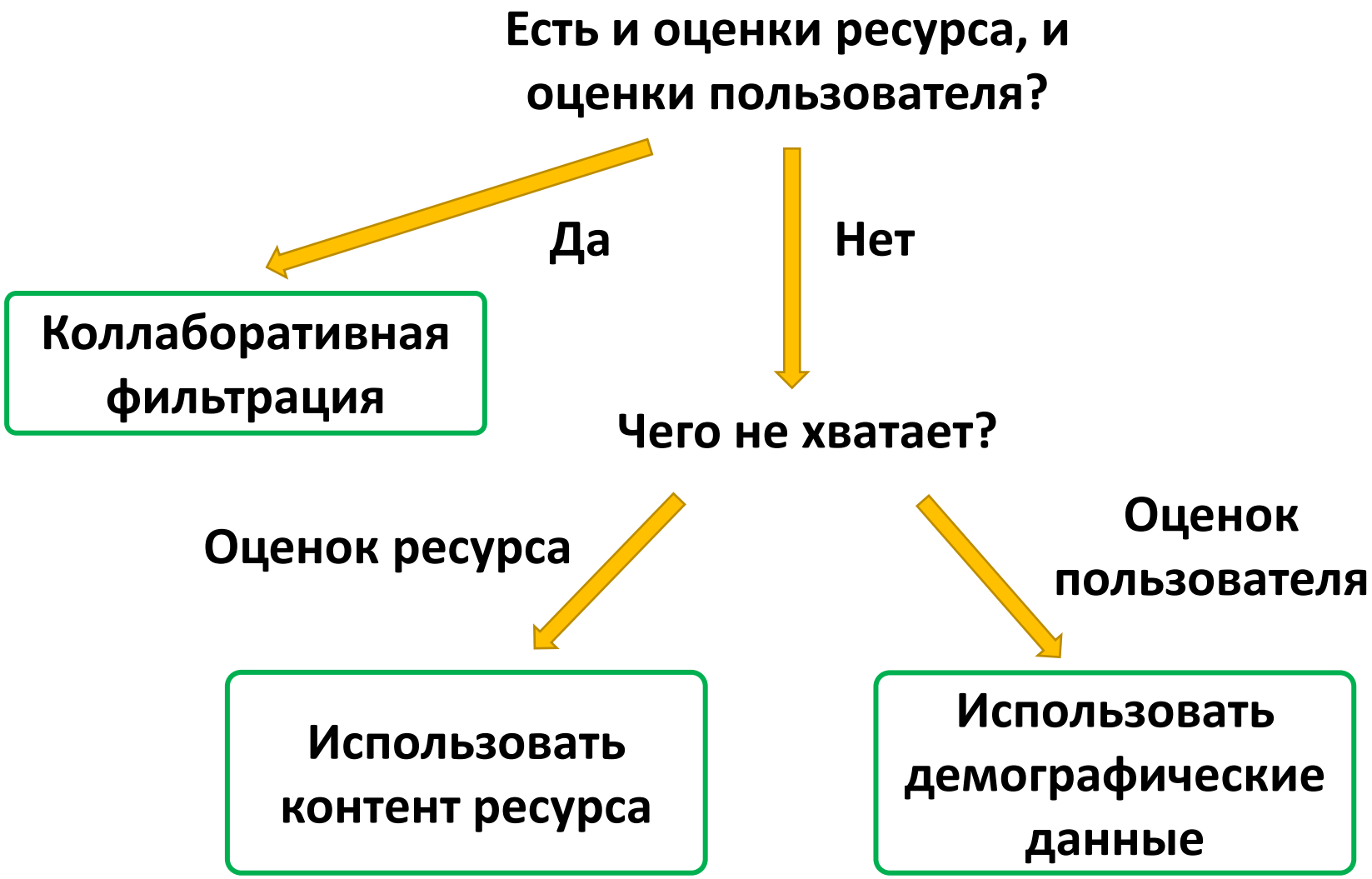
1. Russian Art Auctions
2. Booming Art Market
3. Bernard Buffet Paintings
4. Photographic Art

Автошкола при МГТУ

drivemaster.ru

Social/Demographic based (основанные на социальных данных) используют данные о поле, возрасте, стране пользователя и др.

Виды рекомендательных систем



Hybrid (гибридные) используют смешанные подходы

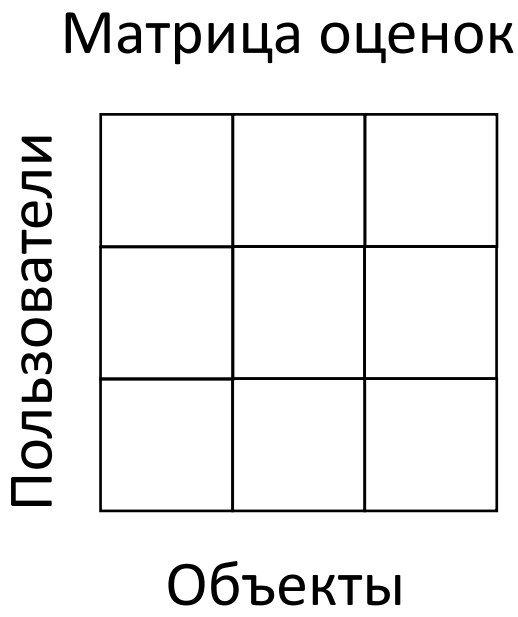
Формализация задачи рекомендательных систем

Дано:

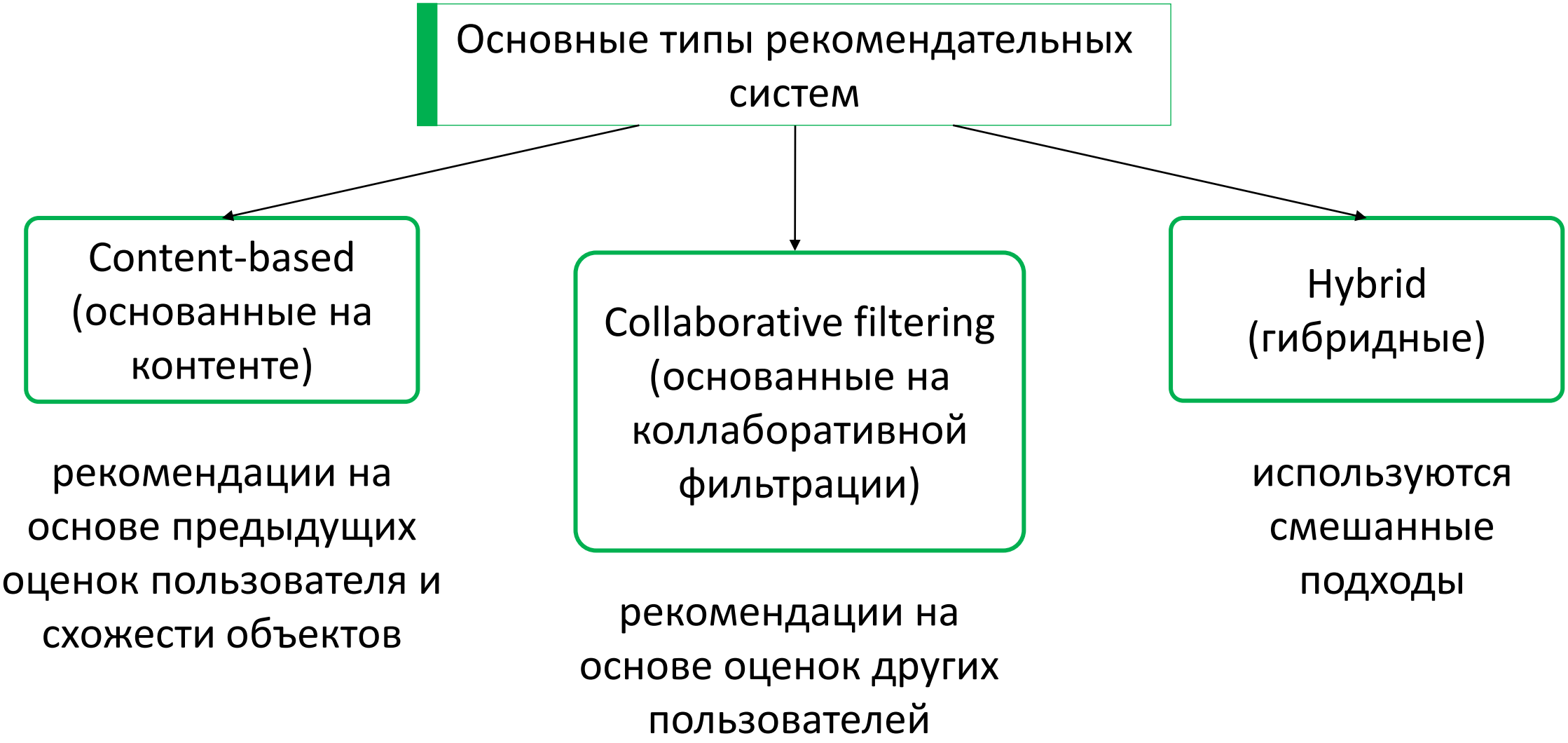
- множество пользователей
- множество объектов
- множество оценок

Задача:

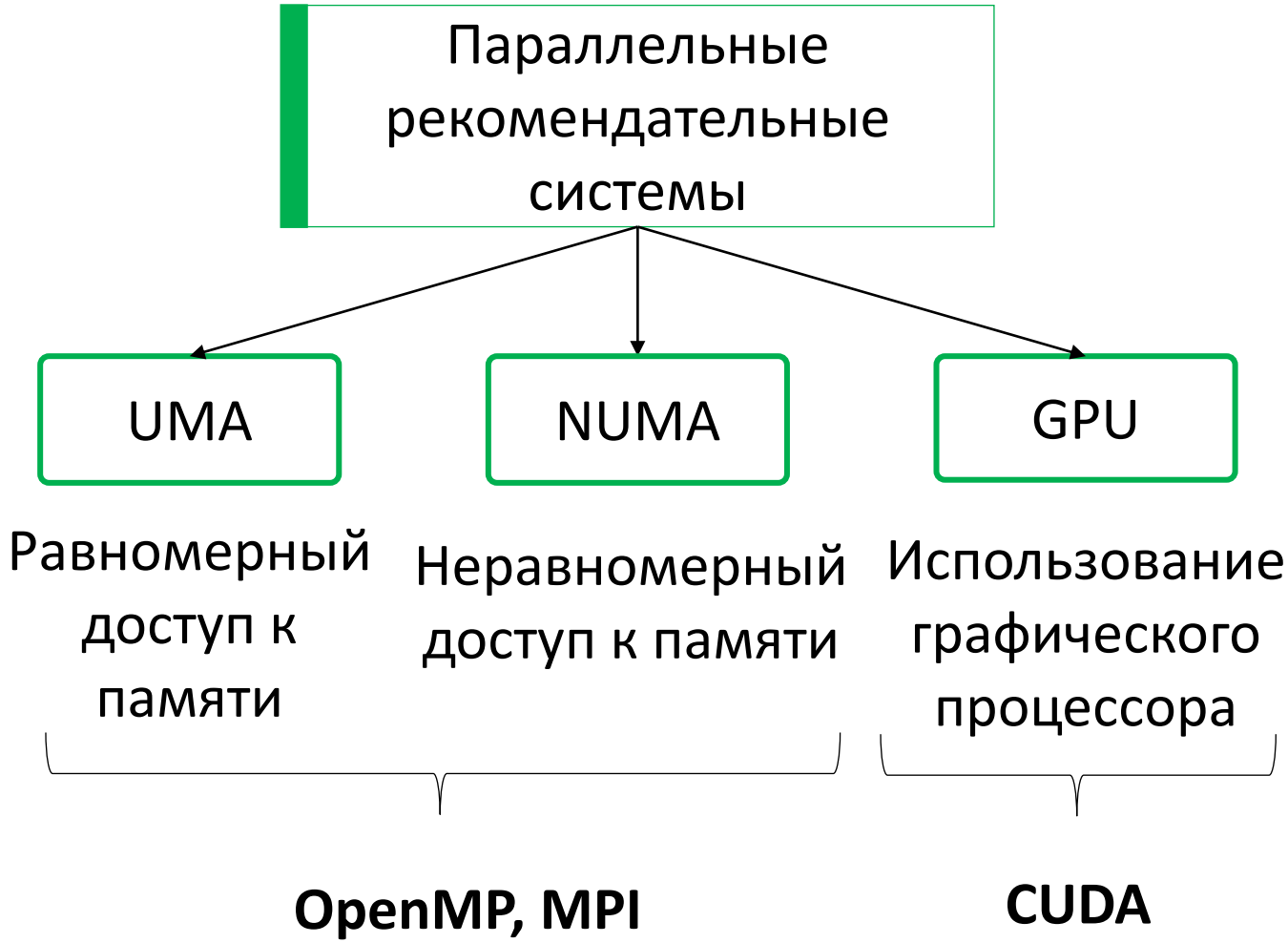
для пользователя и объекта
предсказать оценку

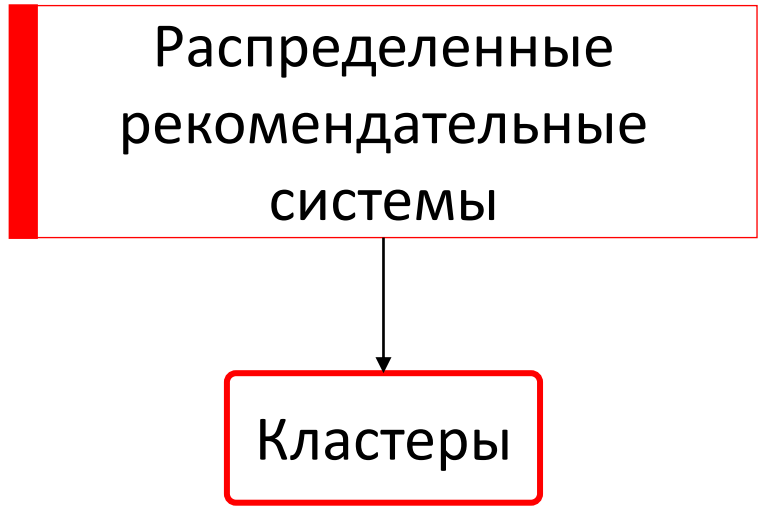


Виды рекомендательных систем



Параллелизм рекомендательных систем

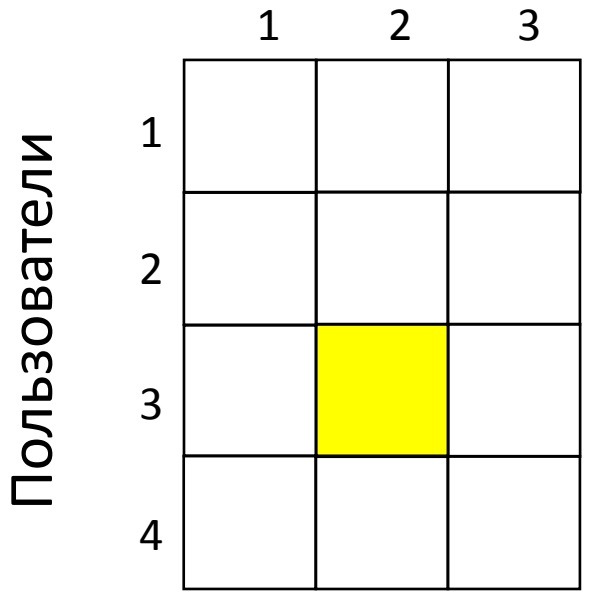




Распределенная контентная фильтрация

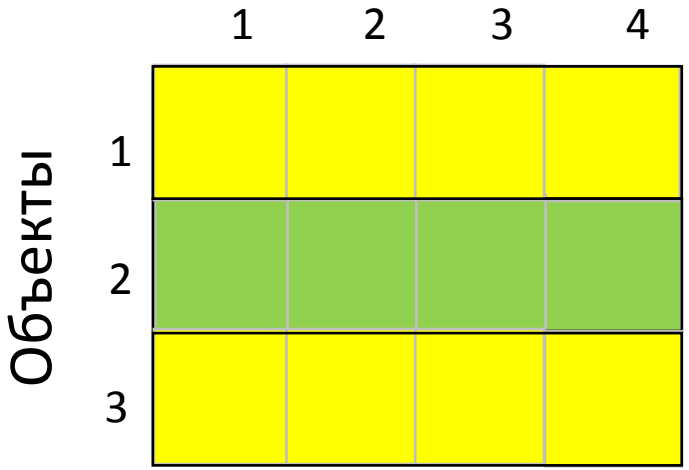
Контентная фильтрация

Объекты



Матрица оценок

Признаки объектов



Матрица признаков объектов

Общая формула

Контентная фильтрация

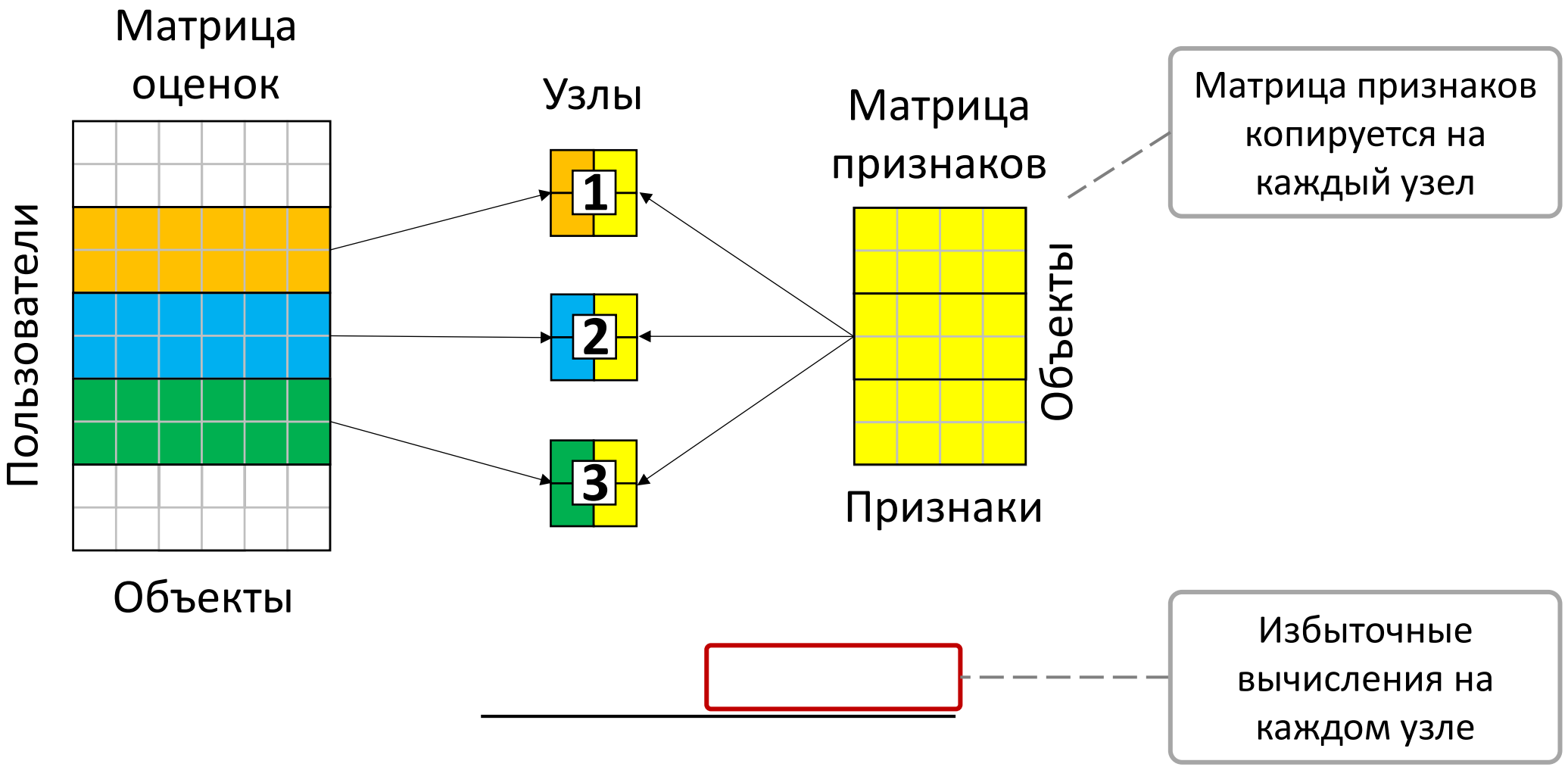
Преимущества

- простота вычислений
- гибкость (можно выбирать, какие признаки объектов использовать)
- прозрачность (рекомендации просто объяснить)

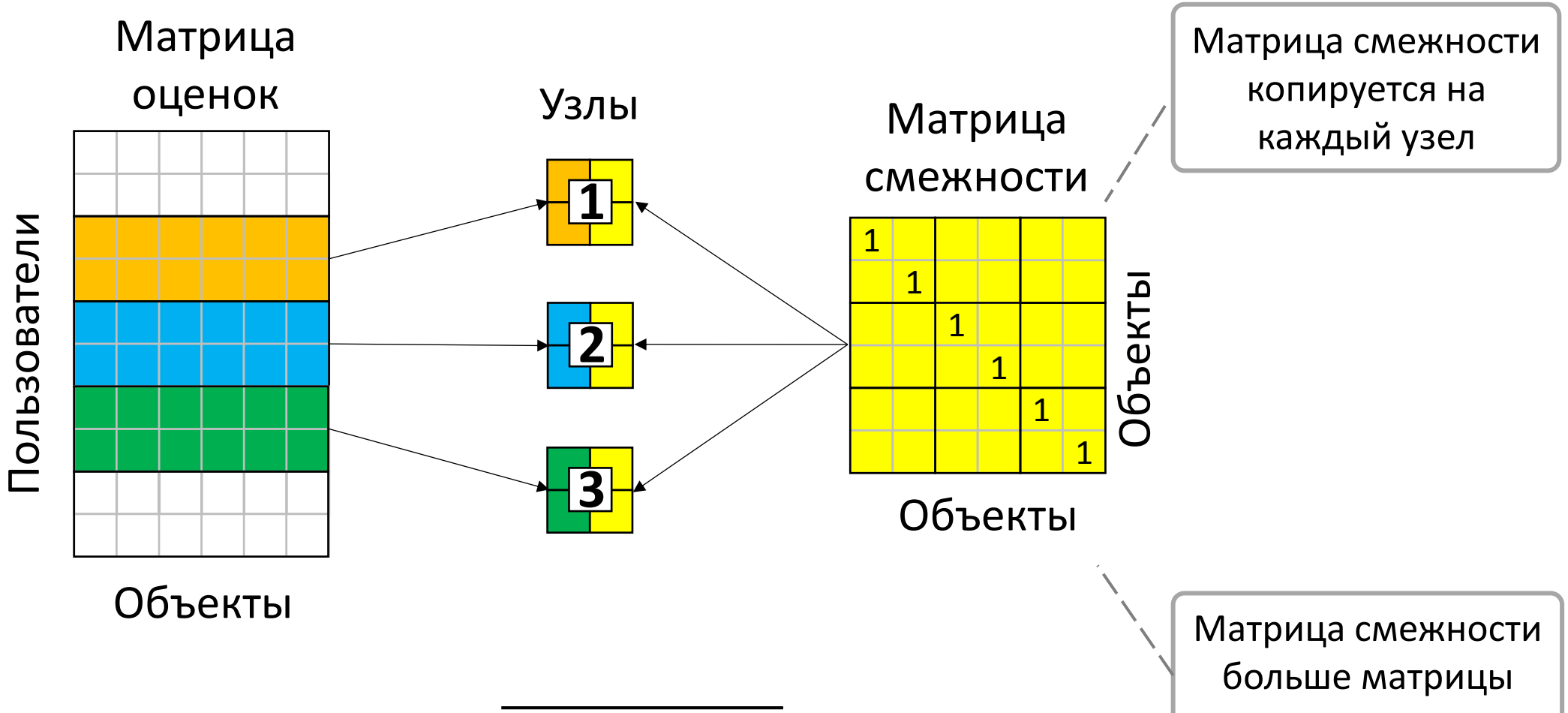
Недостатки

- не учитываются сложные случаи
- нельзя рекомендовать объекты вне текущих предпочтений пользователя
- проблема "холодного старта" (новые объекты никому не рекомендуются)

Распределенная контентная фильтрация. Вариант 1

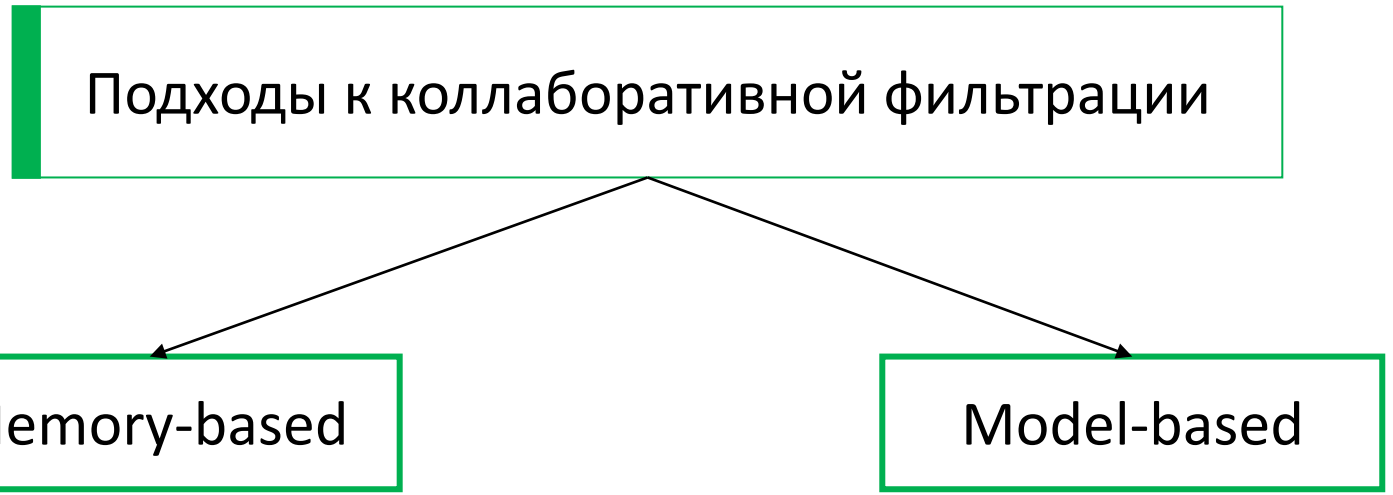


Распределенная контентная фильтрация. Вариант 2



Распределенная коллаборативная фильтрация

Коллаборативная фильтрация



Методы, основанные на нахождении ближайших соседей (kNN = k Nearest Neighbours)

Методы, основанные на факторизация матриц, байесовых сетях, методах кластеризации

Коллаборативная фильтрация (Memory Based)

Матрица оценок

	1	2	…	k	…	m
1	r_{12}	r_{12}	…	r_{1k}	…	r_{1m}
2	r_{21}	r_{22}	…	r_{2k}	…	r_{2m}
…	…	…	…	…	…	…
k	r_{k1}	r_{k2}	…	r_{kk}	…	r_{km}
…	…	…	…		…	…
n	r_{n1}	r_{n2}	…	r_{nk}	…	r_{nm}

Основанная на пользователях (user-based)

Основанная на объектах (item-based)

Коллаборативная фильтрация (Memory Based)

Преимущества

- позволяет выявлять сильные зависимости между пользователями/объектами
- прозрачность (рекомендации просто объяснить)

Недостатки

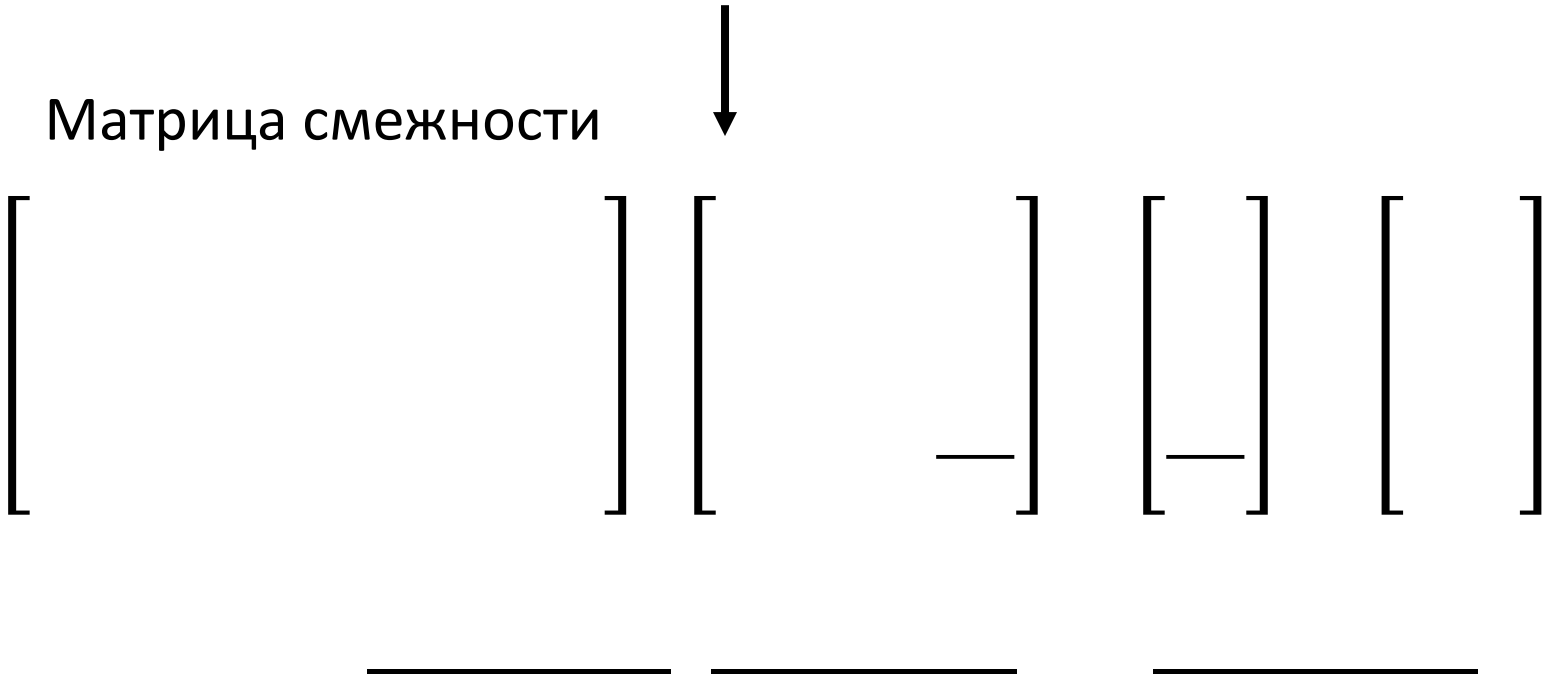
- сложность вычислений (на деле берут лишь k ближайших соседей)
- нет поддержки пользователей с уникальными предпочтениями
- проблема "холодного старта" (новые объекты никому не рекомендуются)
- рекомендации часто тривиальны

Распределенная коллаборативная фильтрация (Memory Based)

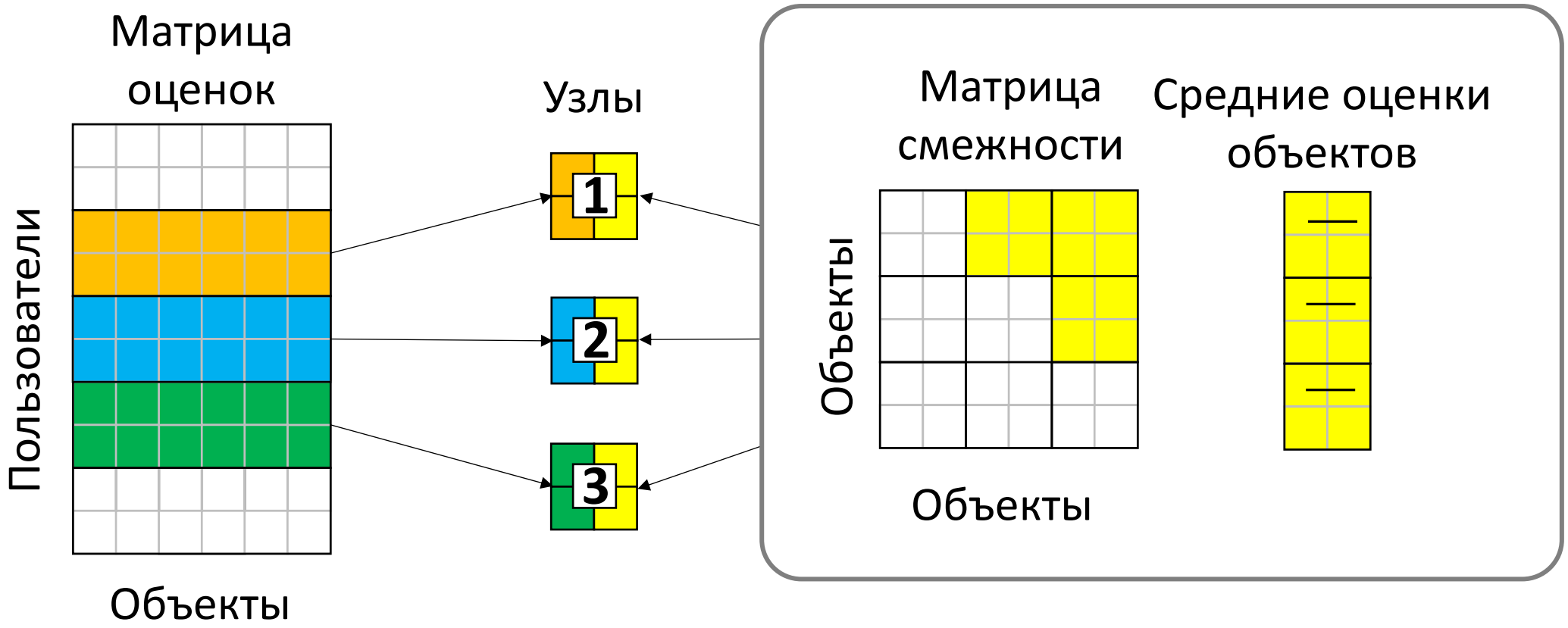
Item-based

( )

Матрица смежности



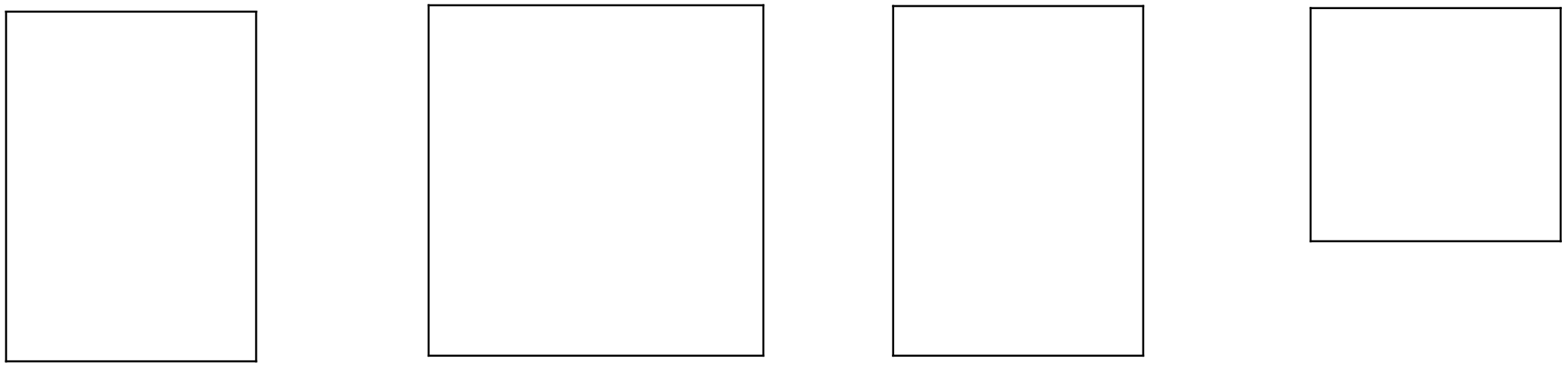
Распределенная коллаборативная фильтрация (Memory Based)



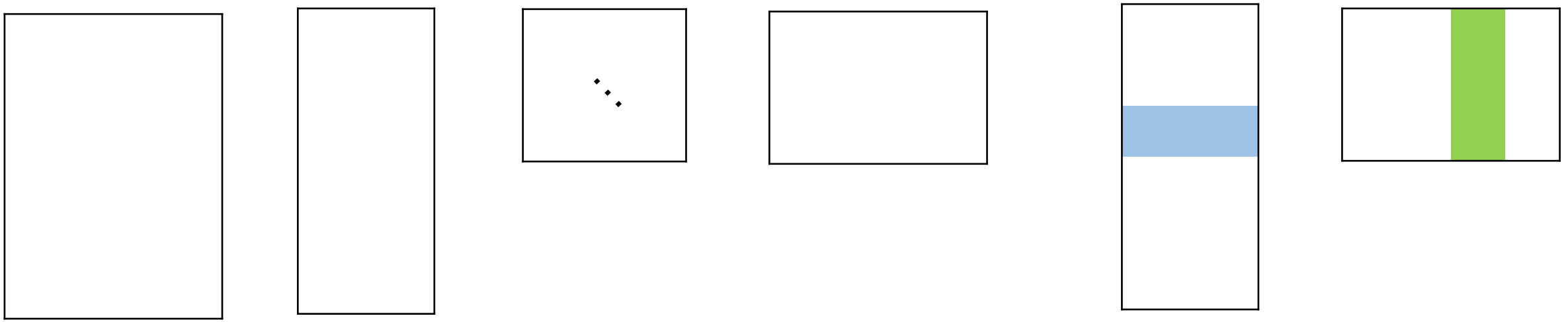
( ) ( )

Коллаборативная фильтрация (Model Based)

( )



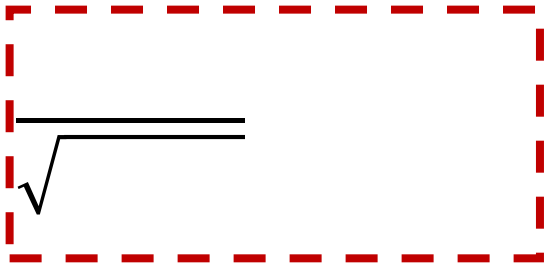
Коллаборативная фильтрация (Model Based)



Модели SVD и SVD++

:

:



– средняя оценка по всем объектам

– отклонения средней оценки пользователя u и объекта

– отклонение средней оценки товара

– объекты, просмотренные пользователем u

– дополнительный набор факторов, характеризующий пользователя на основе того, что он просматривал

( )

( )

Распределенная коллаборативная фильтрация (Model Based)

Преимущества

- позволяет выявлять общие зависимости, учитывать все имеющиеся данные
- высокая точность

Недостатки

- высокие временные затраты на обучение модели
- не прозрачный (рекомендации сложно объяснить)

Распределенная факторизация матрицы рейтингов

Факторизация матрицы. Задача оптимизации

Задача оптимизации

( )

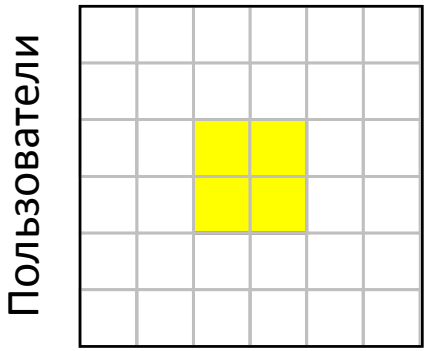
Функция потерь

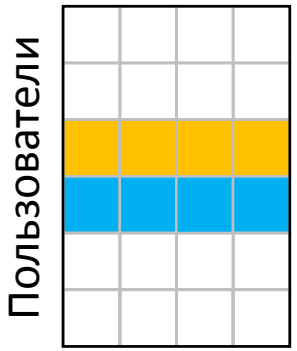
( ) ( )

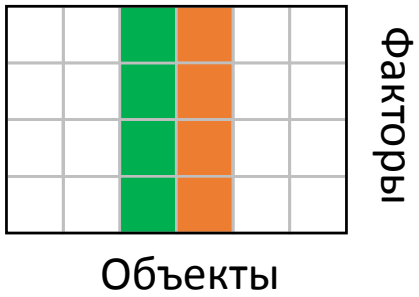
( )

( ) ( )

Матрица рейтингов





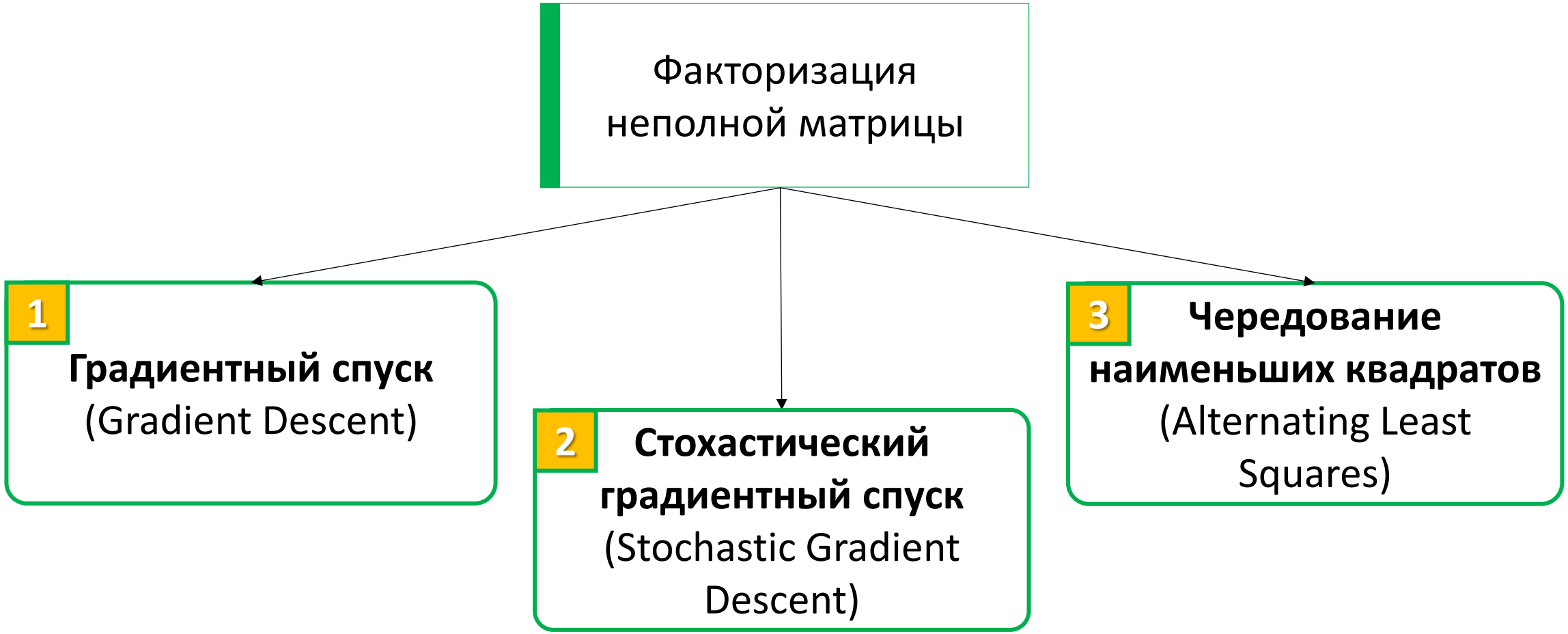


Пример функции потерь (без регуляризации)

( ) ( )

( )

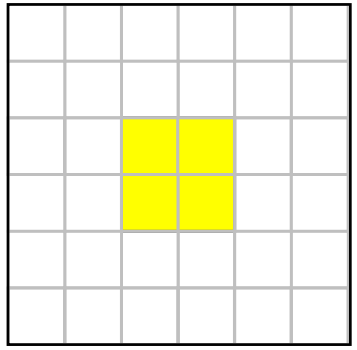
Факторизация матрицы. Распределенные алгоритмы

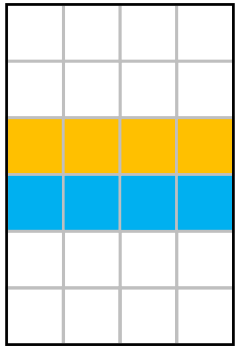


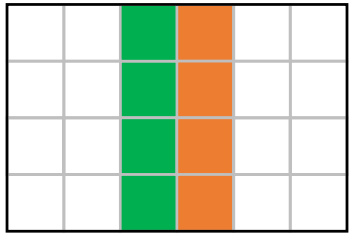
Градиентный спуск

Задача оптимизации

( )



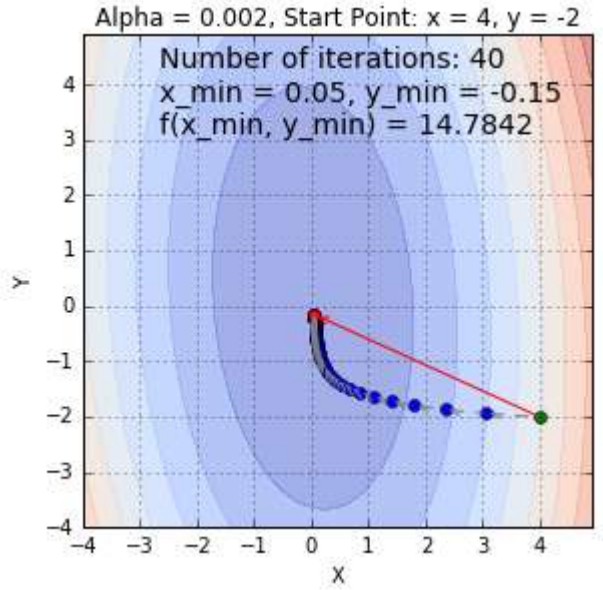




Градиентный спуск для минимизации функции потерь

– итерация

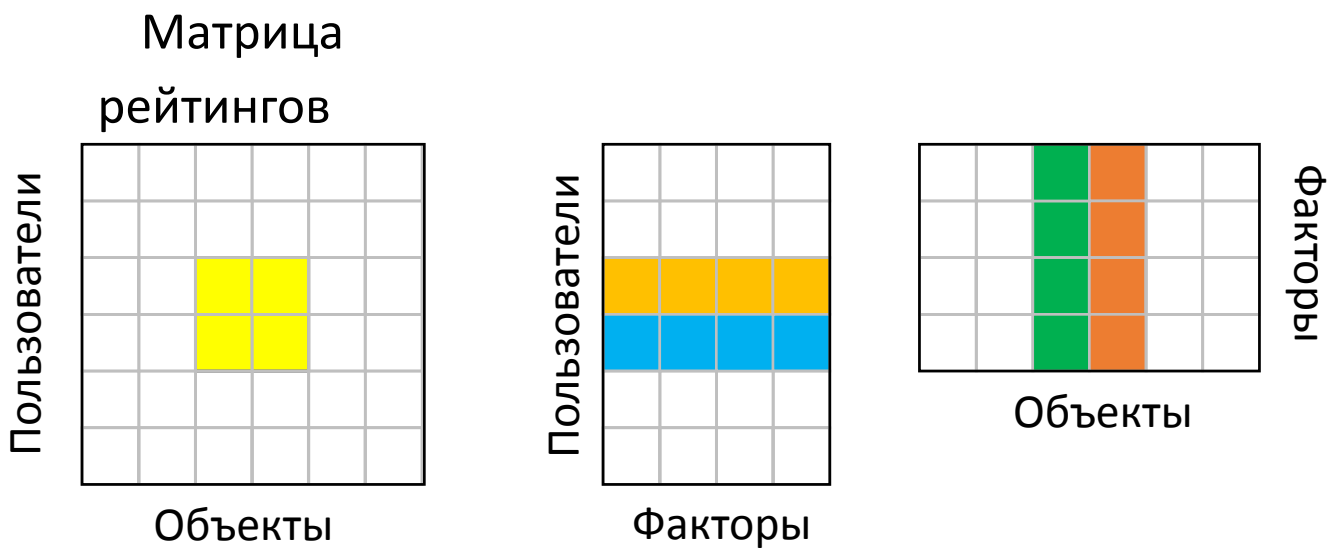
– шаг обучения



Градиентный спуск

Градиентный спуск

Частные производные



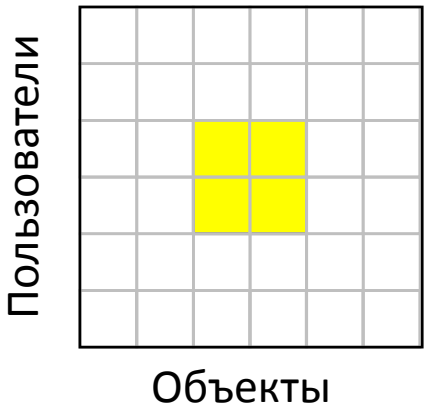
—— ( ) —— ( ) —— ( ) —— ( )

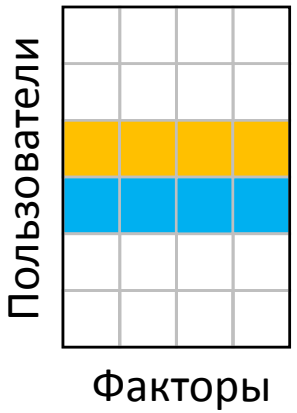
{ |( ) } { |( ) }

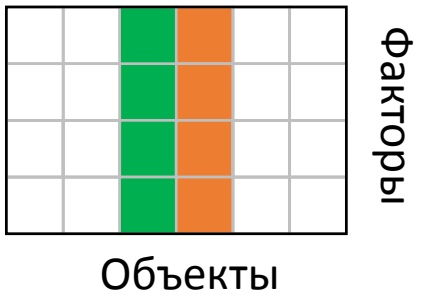
Пример частной производной для функции потерь (без регуляризации)

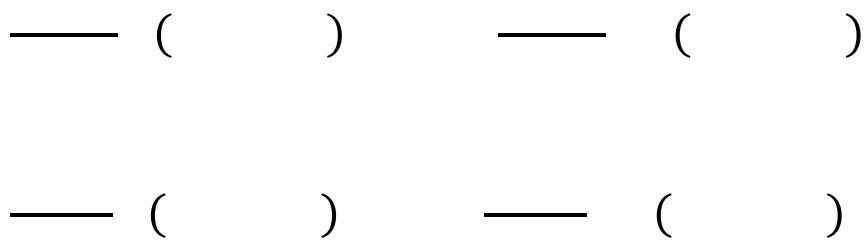
—— ( ) ( )

Распределенный градиентный спуск

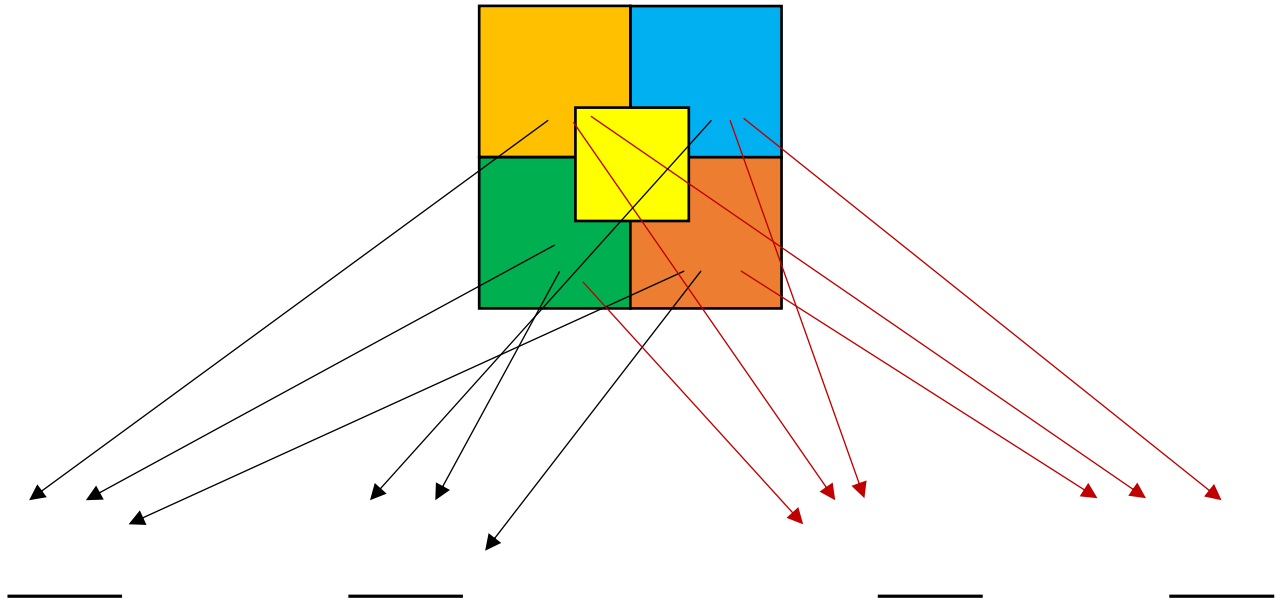




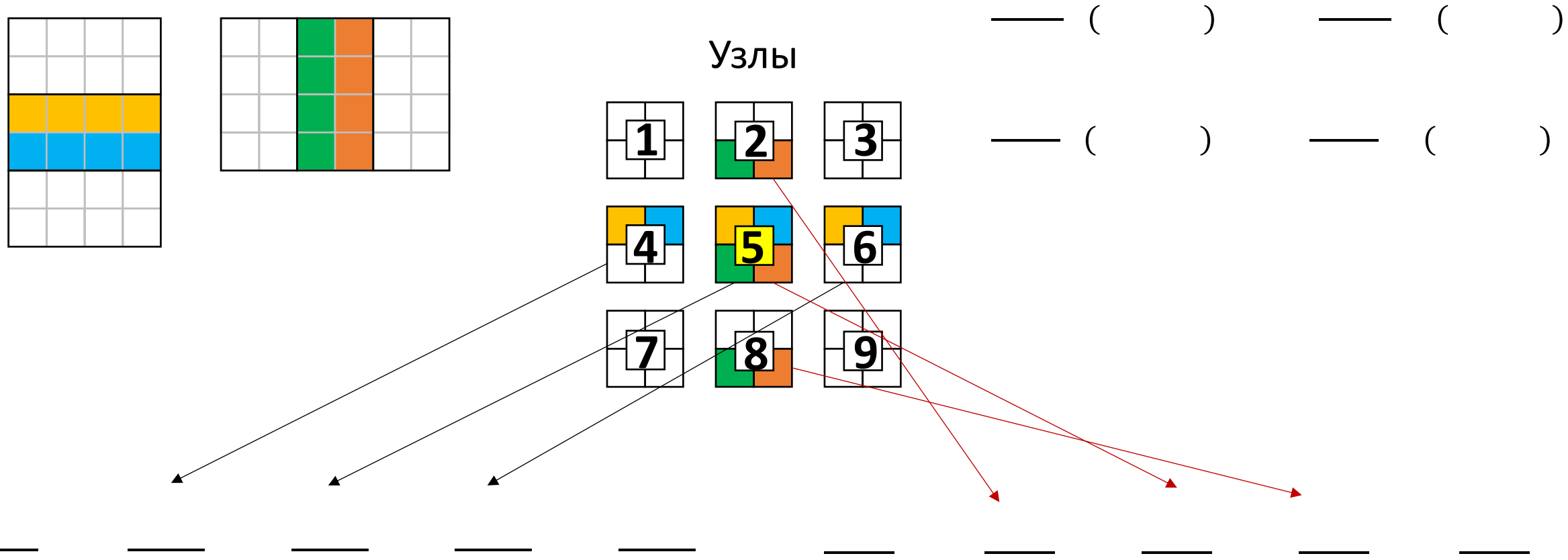




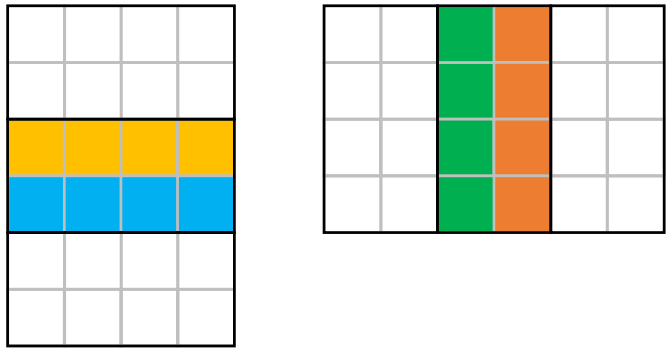
Узел 5



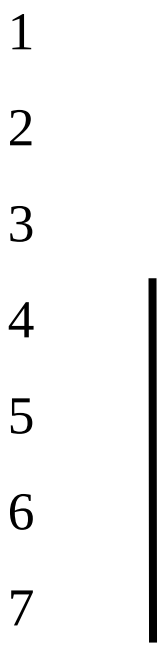
Распределенный градиентный спуск



Распределенный градиентный спуск

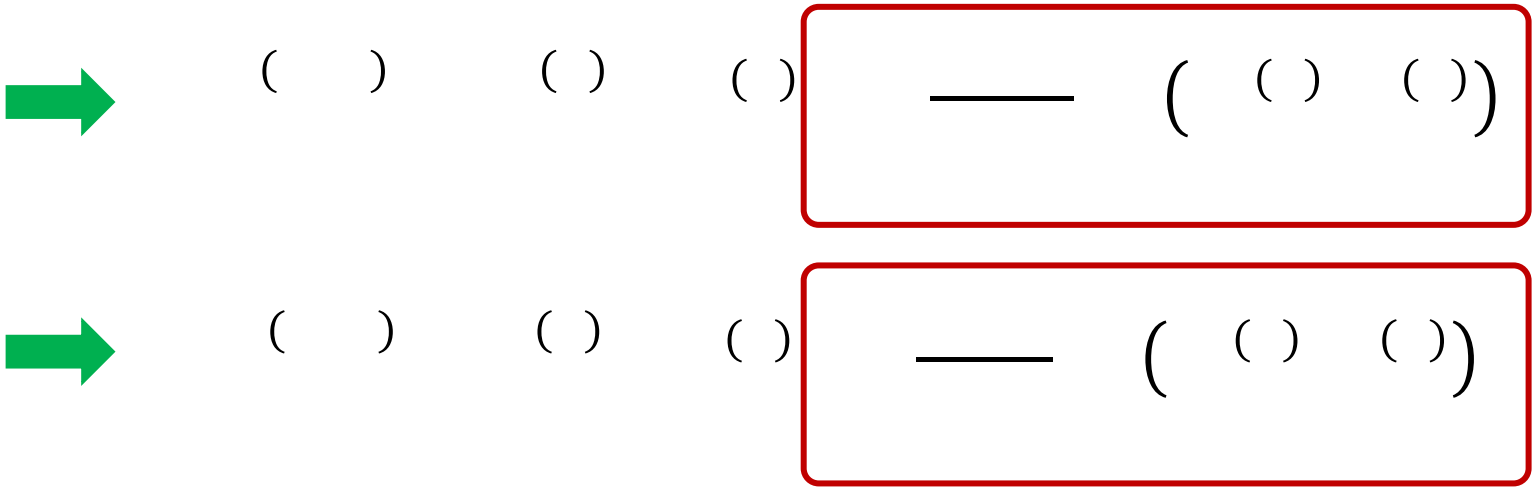


Распределенный градиентный спуск

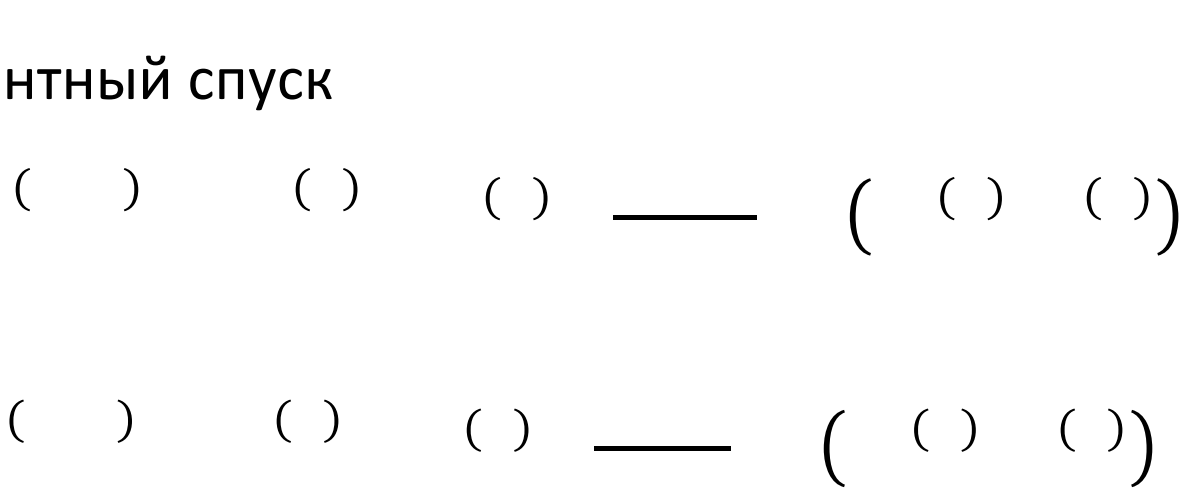


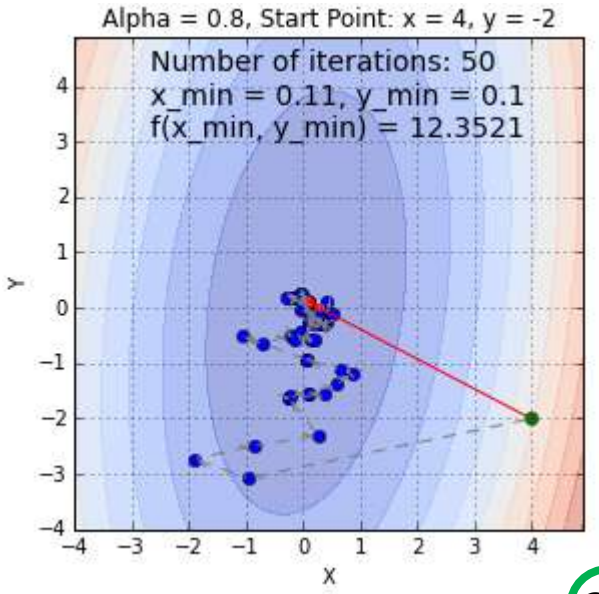
Стохастический градиентный спуск

Градиентный спуск

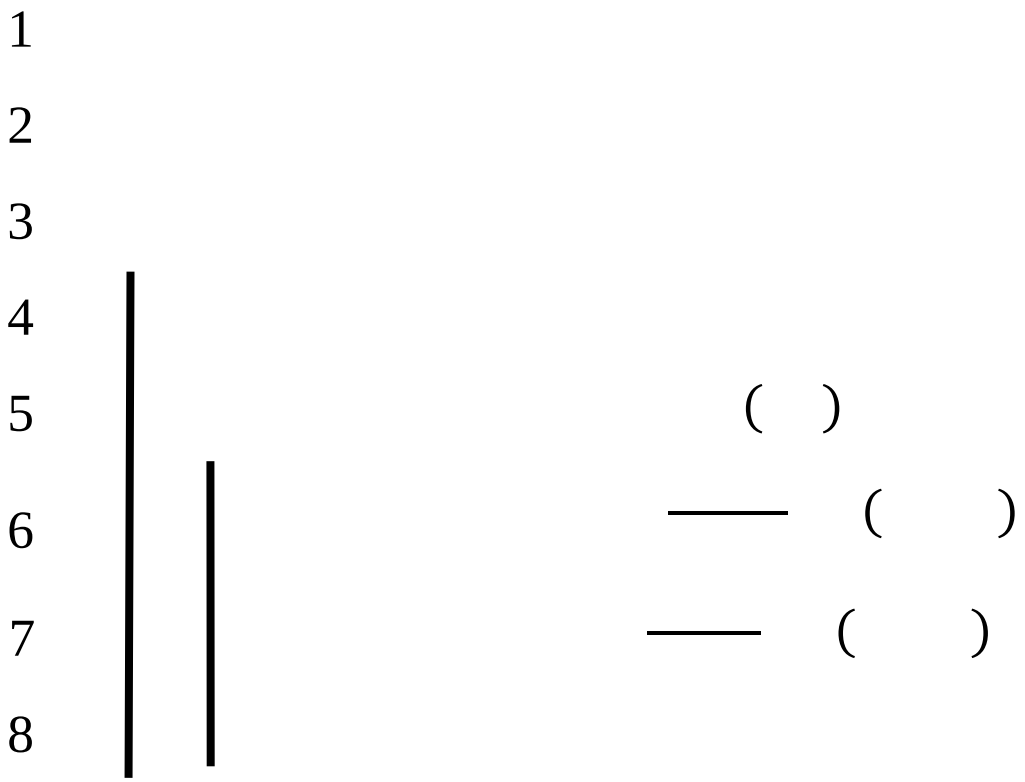


Стохастический градиентный спуск

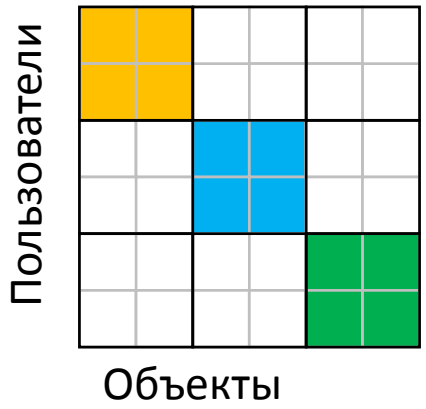


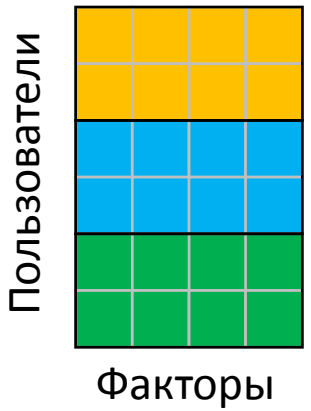


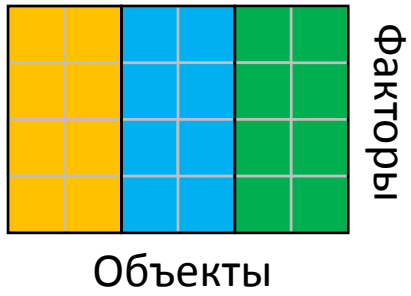
Стохастический градиентный спуск

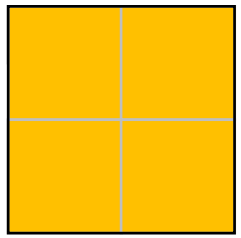


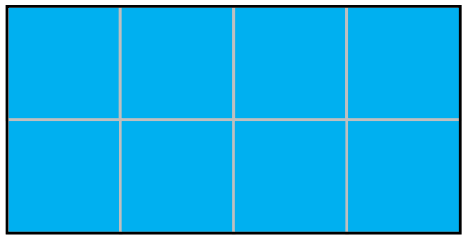
Распределенный стохастический градиентный спуск

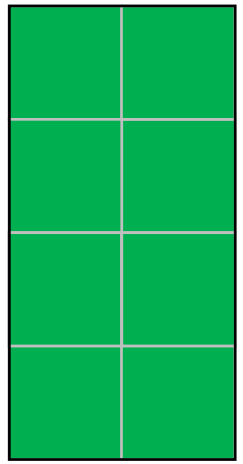




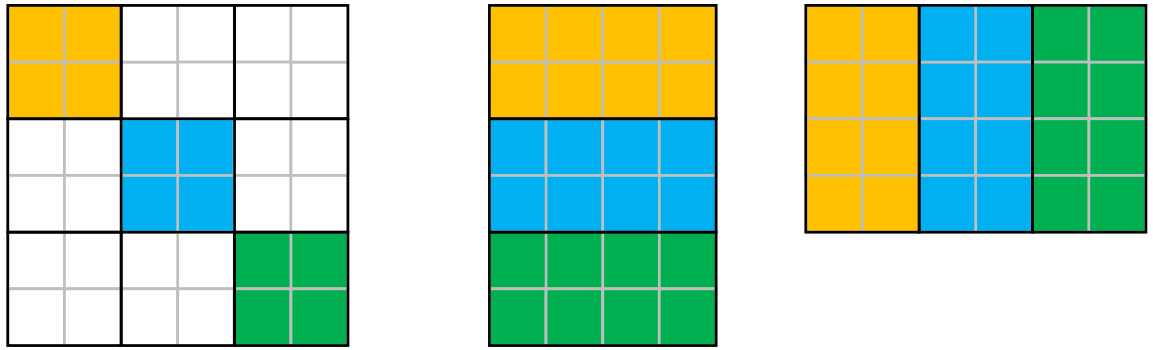








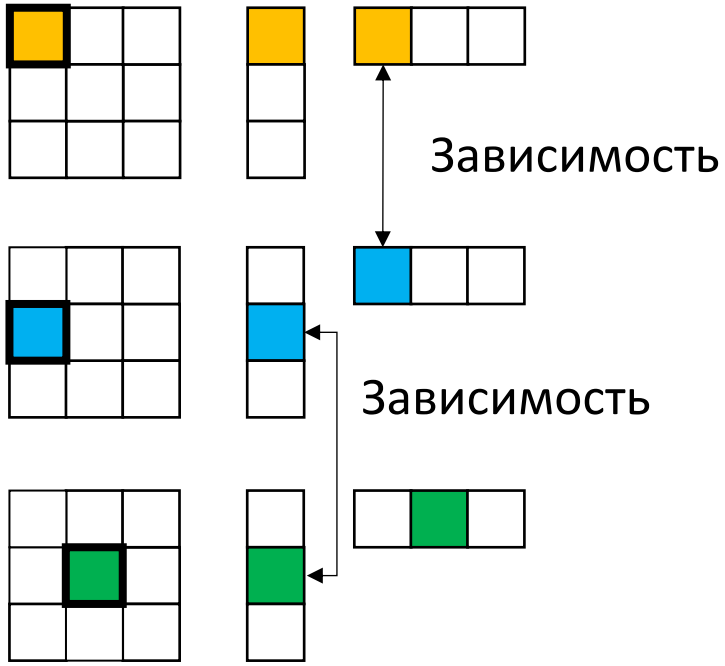
Распределенный стохастический градиентный спуск



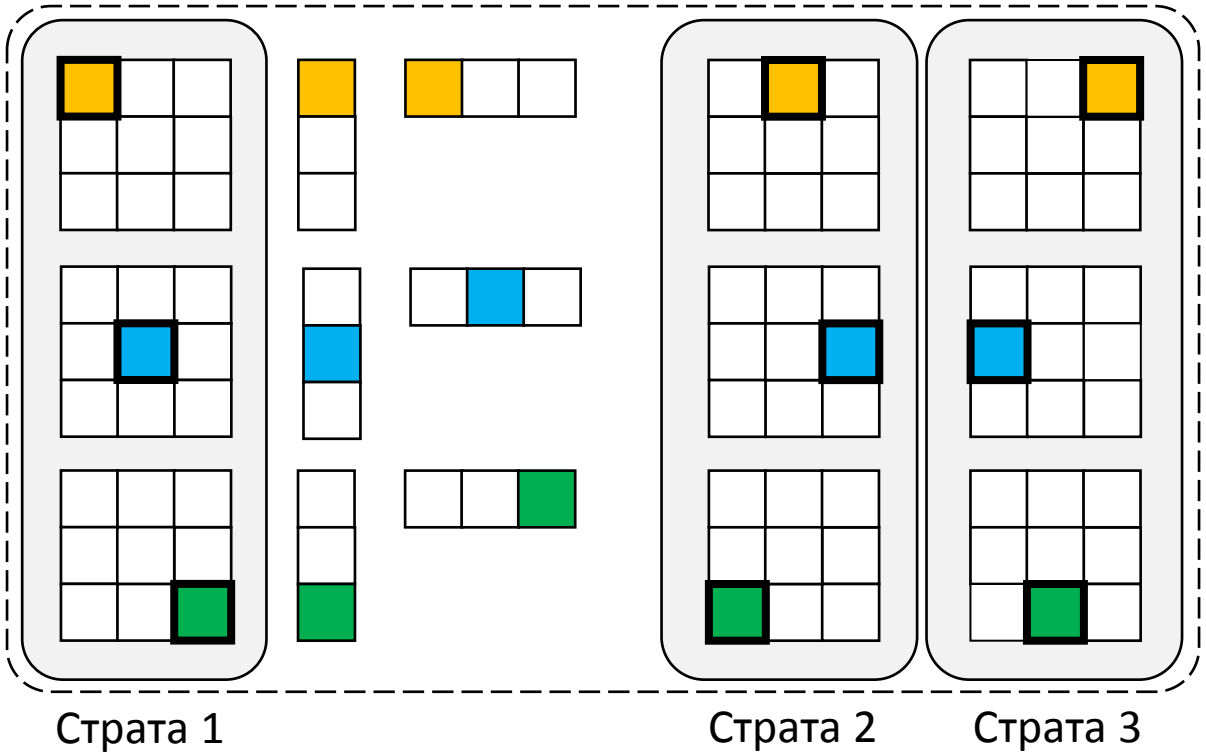
( ) ( ) ( ) —— (( ) ( ))

( ) ( ) ( ) —— (( ) ( ))

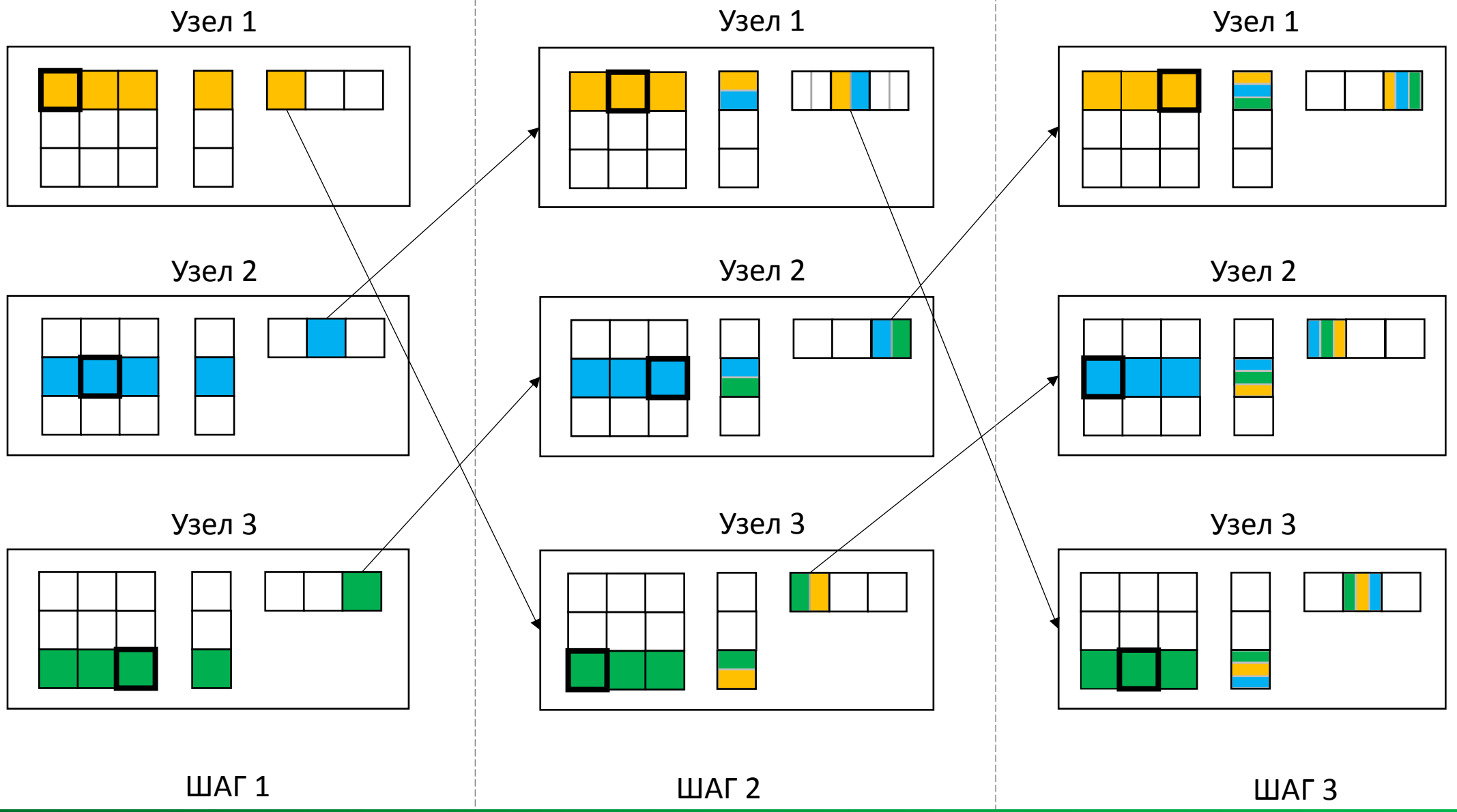
Проблема



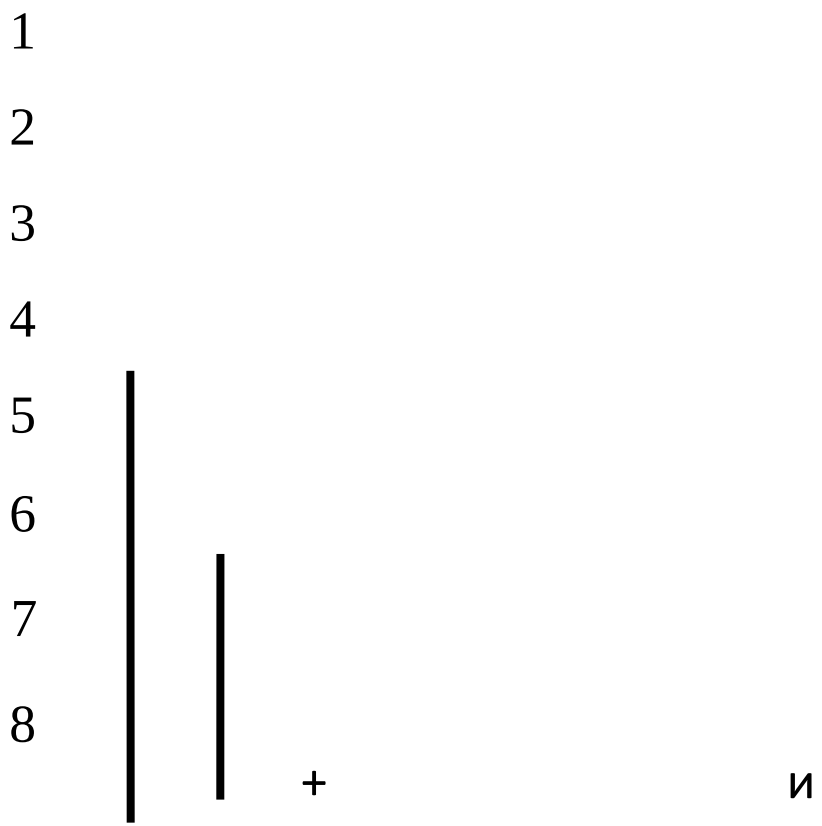
Решение



Распределенный стохастический градиентный спуск



Распределенный стохастический градиентный спуск



Метод наименьших квадратов с чередованием (ALS)

Alternating Least Squares – ALS

Задача оптимизации

( )

Чередующиеся наименьшие квадраты (ALS)

( )

( )

( )

( )

Метод наименьших квадратов с чередованием (ALS)

Вычисление при фиксированном

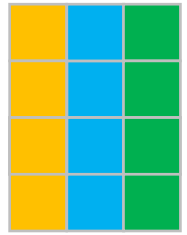
( ) ( ))

	2	2	3		
		5		5	
4		3	5		
	3	3			5
2				4	
5	2		3		2

4	3	5

С регуляризацией

( ) ( ))



Метод наименьших квадратов с чередованием (ALS)

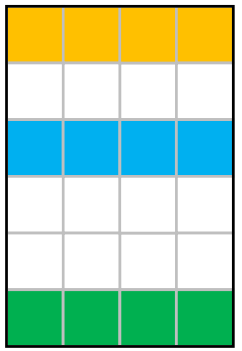
Вычисление при фиксированном

$$\left(\quad (\quad) \quad (\quad) \right) \qquad (\quad)$$

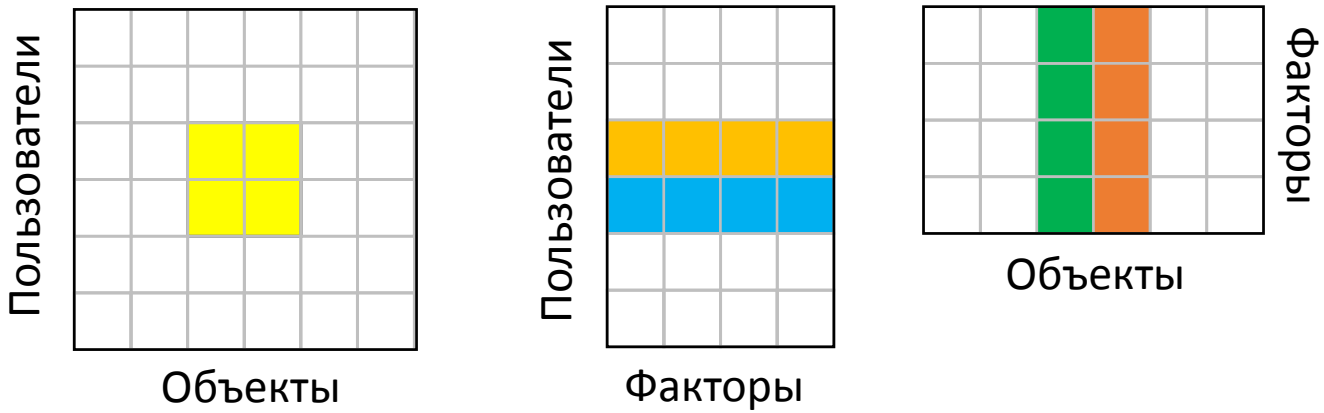
	2	2	3		
		5		5	
4		3	5		
	3	3			5
2				4	
5	2		3		2

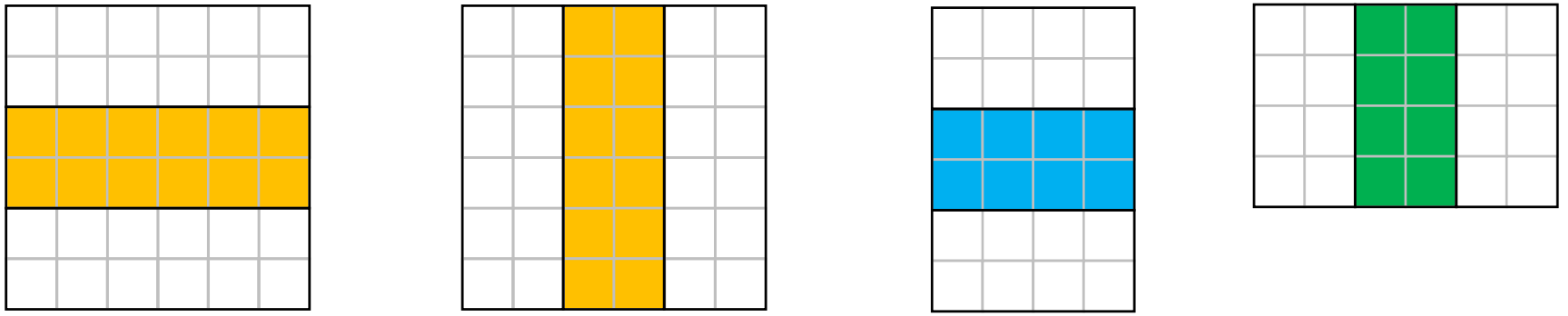
С регуляризацией

$$\left(\quad (\quad) \quad (\quad) \right) \qquad (\quad)$$



Распределенный ALS

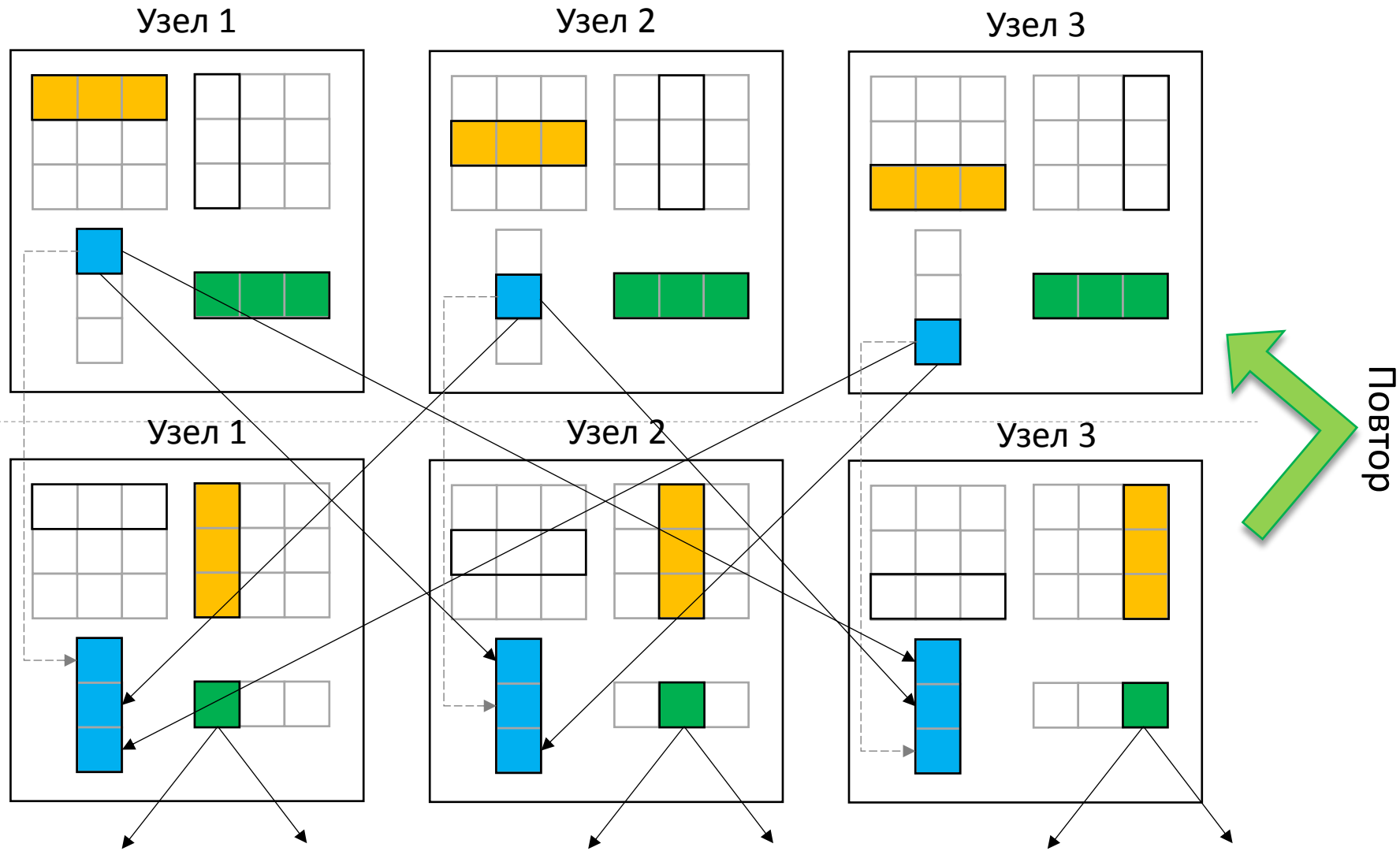




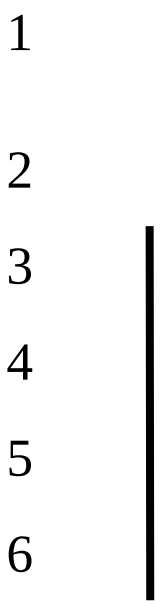
Распределенный ALS

1. Вычисление при фиксированном

2. Вычисление при фиксированном



Распределенный ALS



Сравнение методов

В эксперименте [R.Gemulla и др, 2013] использовался R-кластер:

- 16 узлов, каждый
  - Intel Xeon E5530 (2.4GHz, 8 ядер)
  - 48ГБ ОЗУ

Для сравнения методов:

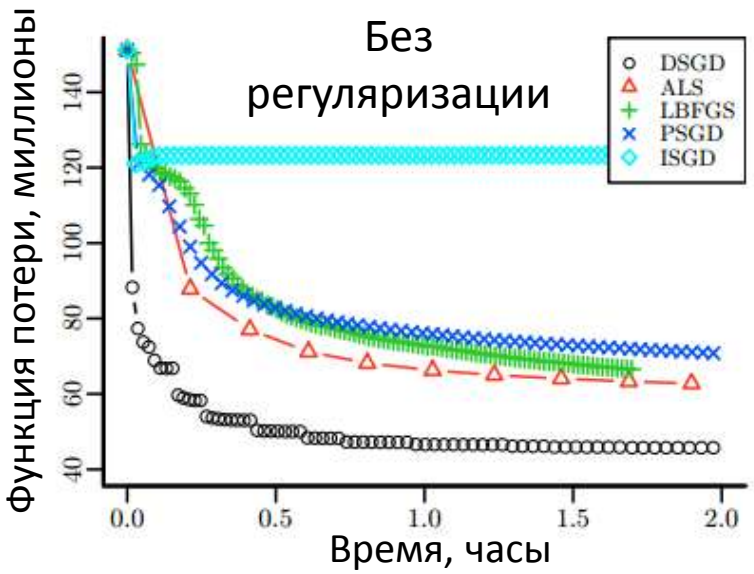
- 8 узлов, каждый с 8 параллельными задачами

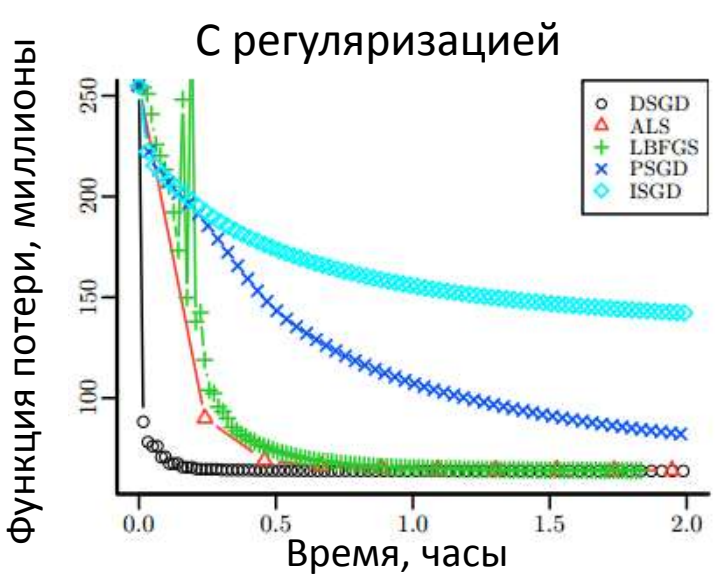
Набор данных Netflix:

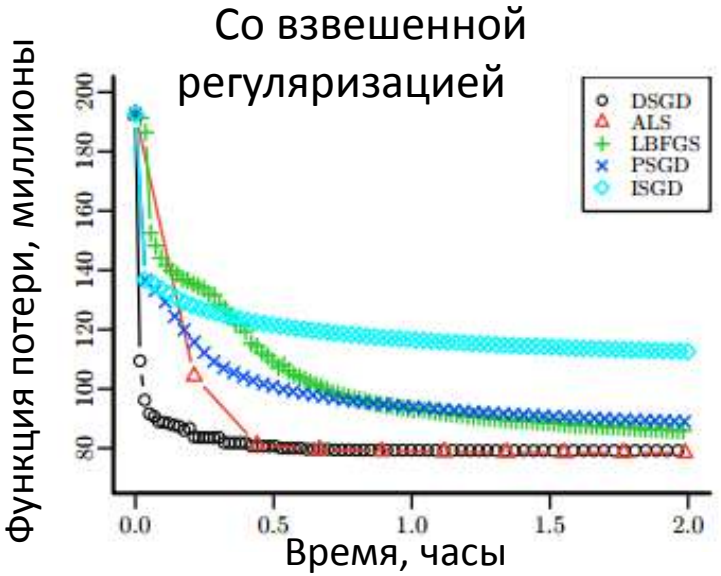
- 100 миллионов рейтингов
- 480 тысяч пользователей
- 18 тысяч фильмов

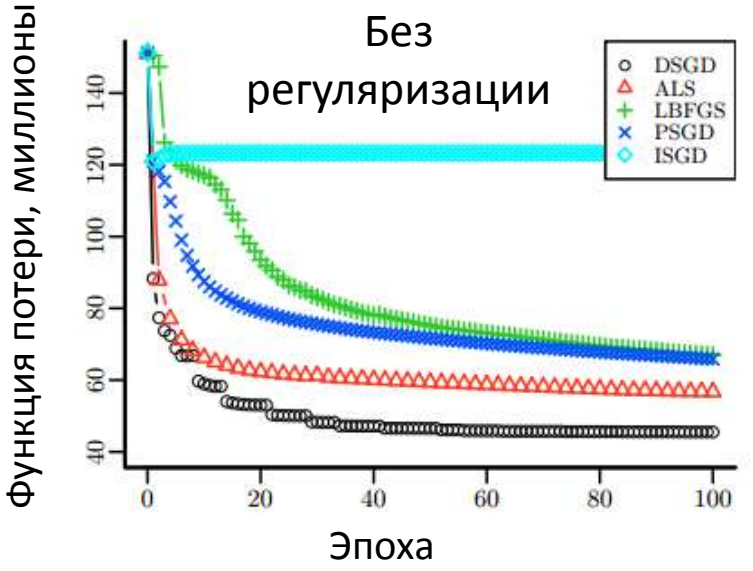
Сравнение методов

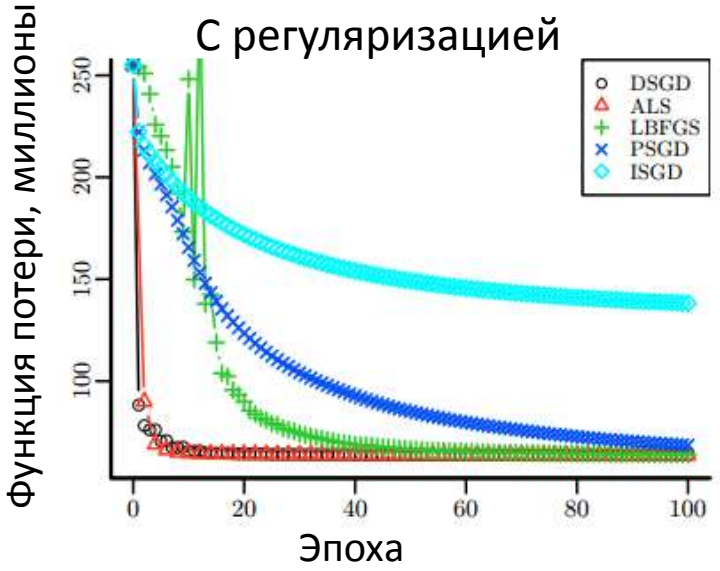
[R.Gemulla и др, 2013]

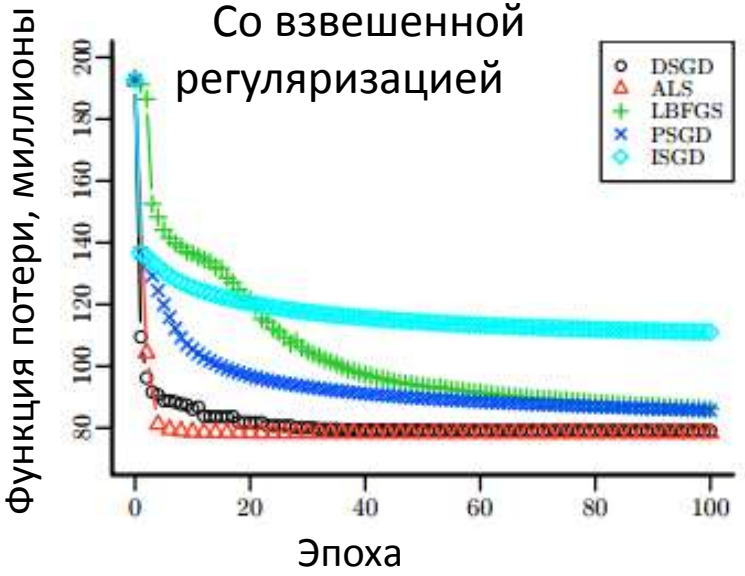










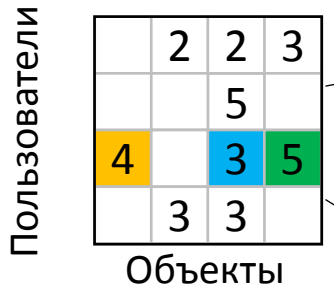


Рекомендательные системы на базе Spark

Основные действия

MLlib

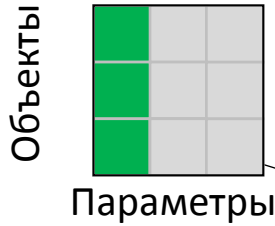
Матрица рейтингов



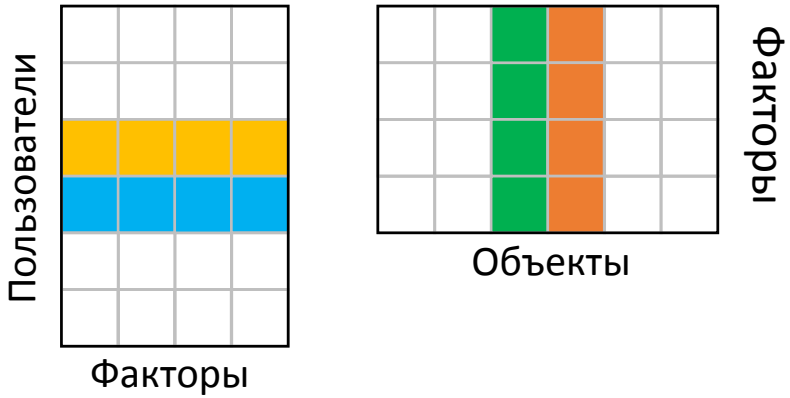
Выбор модели (параметров)

{ }

Матрица объектов



Обучение лучшей модели

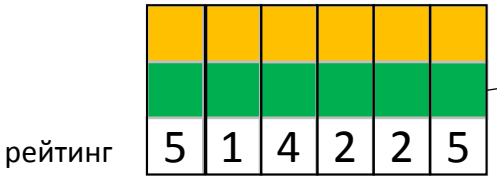


Список для рекомендации

Рекомендации (прогнозирование)

)

Список новых рейтингов



Добавление новых рейтингов

Повторное обучение

ALS в MLlib

Библиотека Mllib включает модуль Recommendation, который содержит класс для реализации распределенного ALS

Метод train класса ALS

`train`(*ratings*, *rank*, *iterations=5*, *lambda_=0.01*, *blocks=-1*, *nonnegative=False*, *seed=None*)

`trainImplicit`(...)

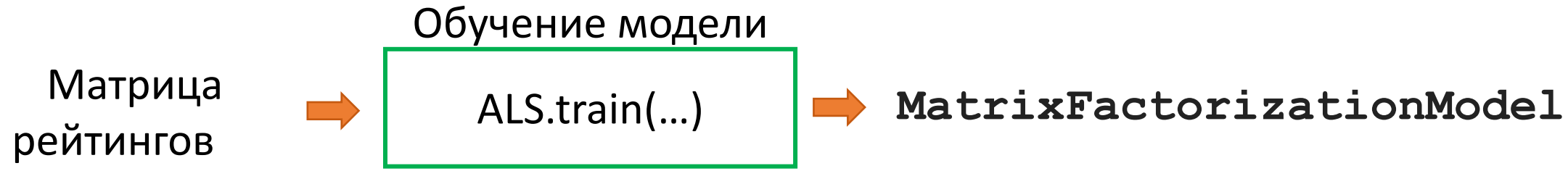
Основные параметры обучения ALS

- *numBlocks* – число блоков
- *rank* - количество факторов
- *iterations* - количестов итераций
- *lambda* - параметр регуляризации

MLlib

Рекомендации на базе Spark и MLlib

Модель факторизованной матрицы



Рекомендации:

- **`predict`**(*user*, *product*)
- **`predictAll`**(*user_product*)
- **`recommendProducts`**(*user*, *num*)
- **`recommendProductsForUsers`**(*num*)
- **`recommendUsers`**(*product*, *num*)
- **`recommendUsersForProducts`**(*num*)

Другое:

- **`load`**(*sc*, *path*)
- **`productFeatures`**()
- **`userFeatures`**()

MLlib

Другие решения

ALS

Item-Based и User-Based коллабортаивная фильтрация

MLlib

Frequent Itemsets

SVD++

Источники

Efthalia Karydi and Konstantinos G. Margaritis «Parallel and Distributed Collaborative Filtering: A Survey», University of Macedonia, Department of Applied Informatics Parallel and Distributed Processing Laboratory

Qun Liu, Xiaobing Li «A New Parallel Item-Based Collaborative Filtering Algorithm Based on Hadoop» School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China. Journal of Software, Volume 10, Number 4, April 2015

Simon Dooms, Pieter Audenaert, Jan Fostier, Toon De Pessemier, Luc Martens «In-Memory, Distributed Content-Based Recommender System». Agency for Innovation by Science and Technology (IWT Vlaanderen)

Boduo Li, Sandeep Tata, Yannis Sismanis «Sparkler: supporting large-scale matrix factorization». EDBT '13 Proceedings of the 16th International Conference on Extending Database Technology. PP 625-636, 2013

Yehuda Koren «Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model». AT&T Labs, August 24–27, 2008, Las Vegas, Nevada, USA

Xavier Amatriain «Recommender Systems». Machine Learning Summer School, Published on Jul 21, 2014

Xavier Amatriain «Big & Personal: data and models behind Netflix recommendations». BigMine’13 August 2013. Chicago, Illinois, USA

Источники

R. Gemulla, P. J. Haas, E. Nijkamp, Y. Sismanis «Large-Scale Matrix Factorization with Distributed Stochastic Gradient Descent», IBM Research Report RJ10481, March 2011 Revised February, 2013

Yunhong Zhou, Dennis Wilkins, Robert Schreiber, Rong Pan «Large-Scale Parallel Collaborative Filtering for the Netflix Prize» AAIM '08 Proceedings of the 4th international conference on Algorithmic Aspects in Information and Management. PP 337–348, Springer-Verlag Berlin, Heidelberg, 2008

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, Ciyou Zhu «A limited memory algorithm for bound constrained optimization». SIAM Journal on Scientific Computing, Volume 16 Issue 5, Sept. PP 1190-1208, 1995

Boduo Li, Sandeep Tata, Yannis Sismanis «Sparkler: supporting large-scale matrix factorization». EDBT '13 Proceedings of the 16th International Conference on Extending Database Technology. PP 625-636, 2013

Trevor Hastie, Rahul Mazumder, Jason D. Lee, Reza Zadeh «Matrix completion and low-rank SVD via fast alternating least squares». The Journal of Machine Learning Research, Volume 16, Issue 1, PP 3367-3402, 2015

Hsiang-Fu Yu, Cho-Jui Hsieh, Si Si, Inderjit Dhillon «Scalable Coordinate Descent Approaches to Parallel Matrix Factorization for Recommender Systems». ICDM '12 Proceedings of the 2012 IEEE 12th International Conference on Data Mining, PP 765-774, 2012

Спасибо за внимание!