

Яндекс



{IT.IS}

КОНФЕРЕНЦИЯ
КОМПАНИИ "ИНТЕРСВЯЗЬ"
ДЛЯ IT-СПЕЦИАЛИСТОВ

Обработка текста с использованием фреймворка Fast AI

Сотов Алексей, разработчик, специалист по анализу данных

Давайте познакомимся



**Какими задачами по
обработке текста вы
занимались?**

Making neural nets **uncool** again!

* «Сделаем нейронные сети доступными каждому»
– Fast AI



Сделаем нейронные сети
доступными каждому

* «Making neural nets uncool again!»
– Fast AI

Jeremy Howard

- › CEO компании Enlitic
- | **Президент и главный научный сотрудник Kaggle**
- › Основатель успешных стартапов FastMail и др
- | **Создатель и вдохновитель фреймворка Fast AI**



Представление текста

Текст

Представление текста

Текст

Документ 1

Документ 2

• • •

Документ n

Представление текста

Текст

Документ 1

смотрел

Документ 2

гулять

Документ п

невероятно

вышел

красиво

• • •

солнце

однако

фильм

хороший

Мешок слов (Bag of words)

Текст

Документ 1

Документ 2

Документ n

...

смотрел

он

гулять

невероятно

вышел

красиво

солнце

однако

фильм

хороший



Мешок слов (Bag of words)

Может выражать:

- | **Вхождение слова в документ**
- | **Частота слова в документе**
- | **TF-IDF – оценка важности слова
в данном документе**



Представление текста

Текст

Документ 1

смотрел

Документ 2

гулять

Документ п

невероятно

• • •

вышел

красиво

солнце

однако

фильм

хороший

Текст

Документ 1

Документ 2

Документ п

• • •

смотрел

он

гулять

невероятно

вышел

красиво

солнце

однако

фильм

хороший



Текст

Документ 1

Документ 2

Документ п

· · ·

смотрел

он

гулять

невероятно

вышел

красиво

солнце

однако

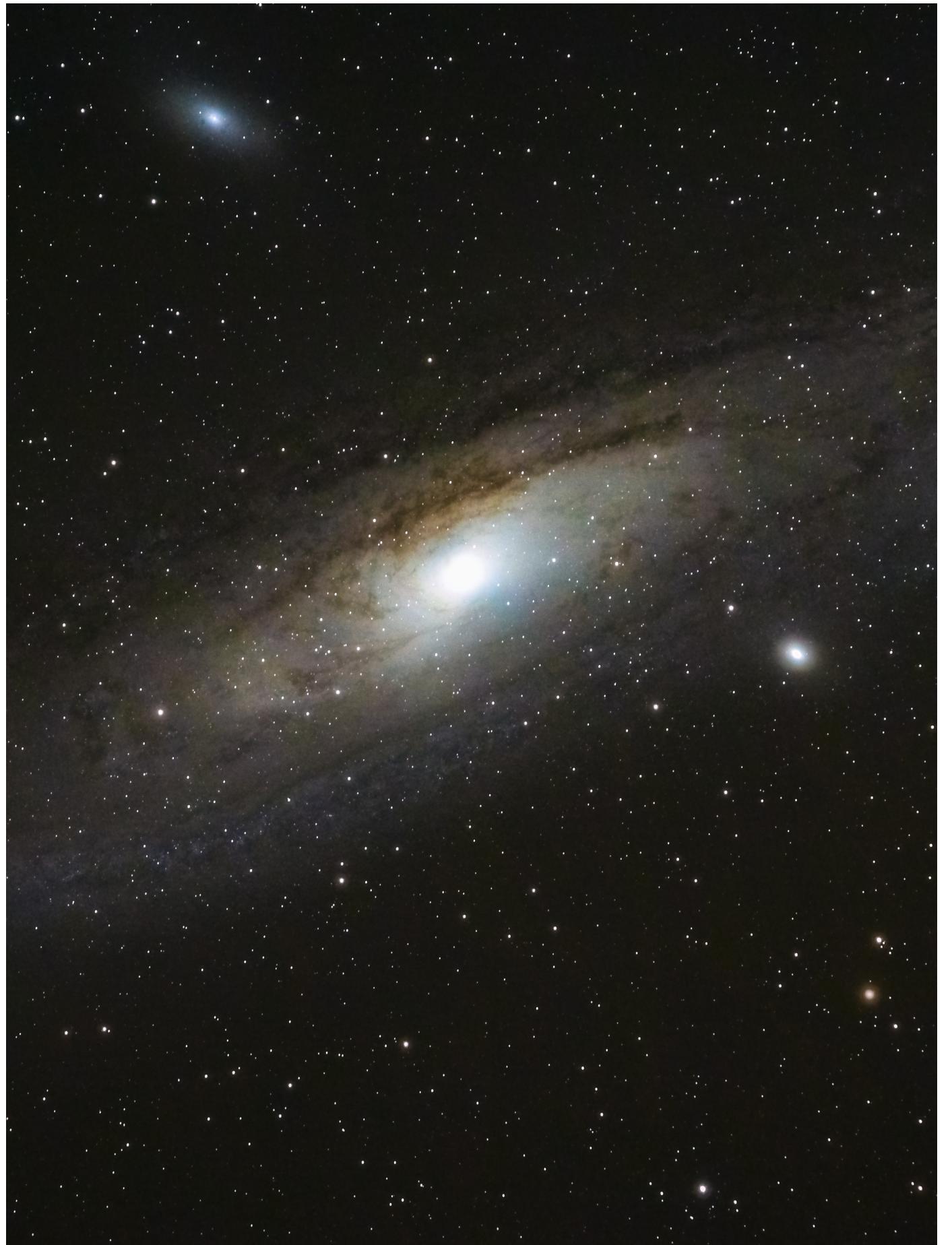
фильм

хороший



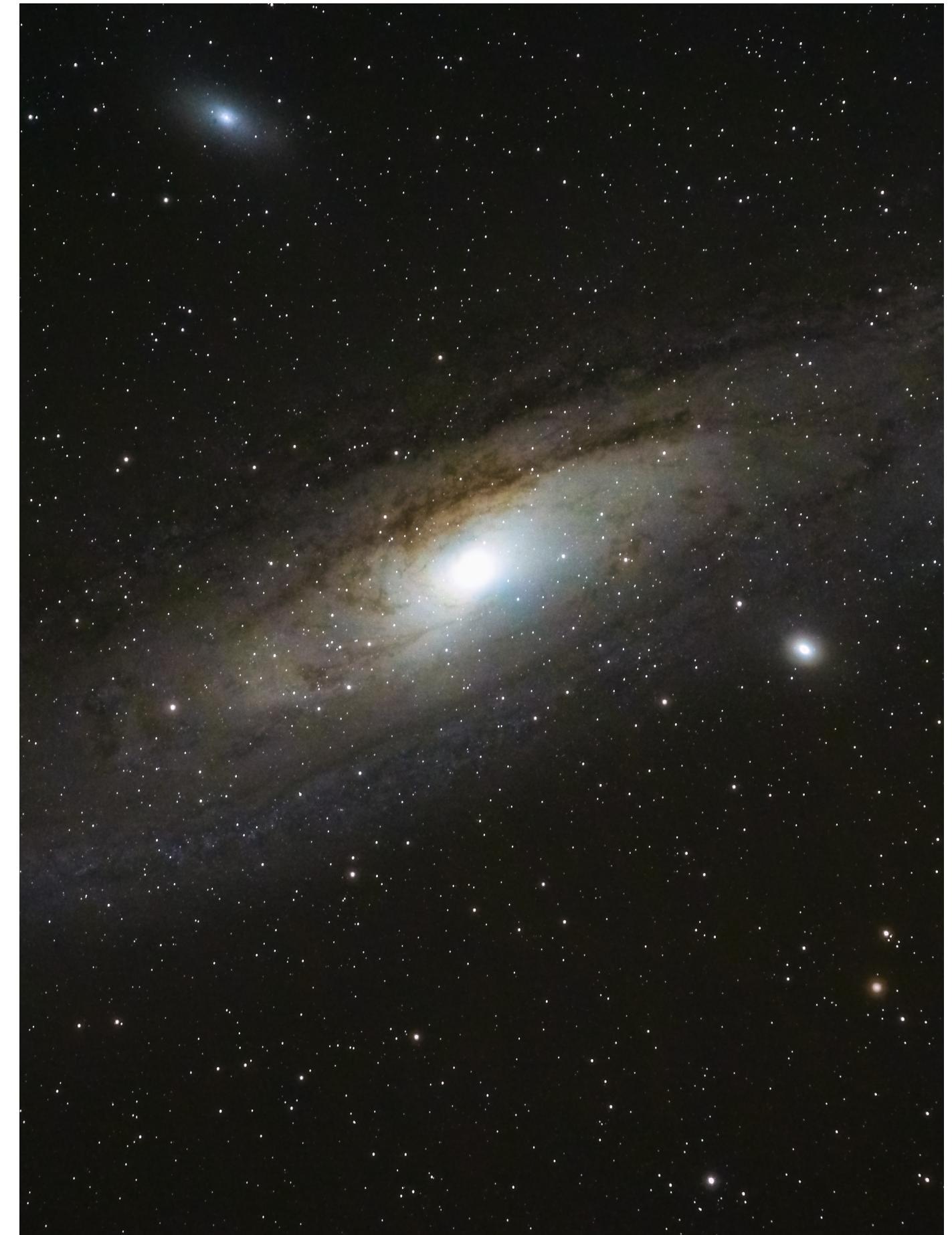
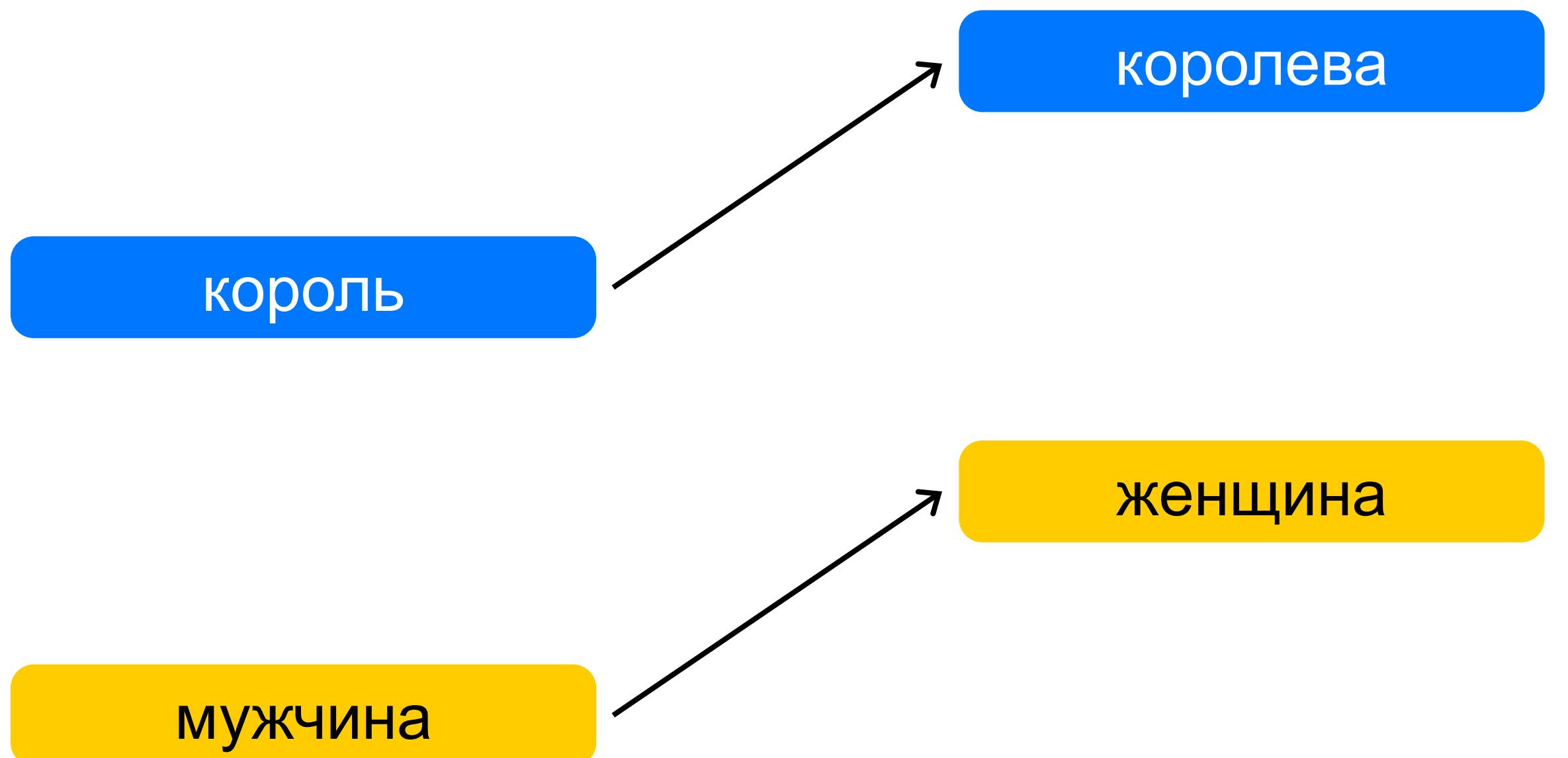
Пример: Word2Vec

**Отношения между словами
в векторном пространстве**



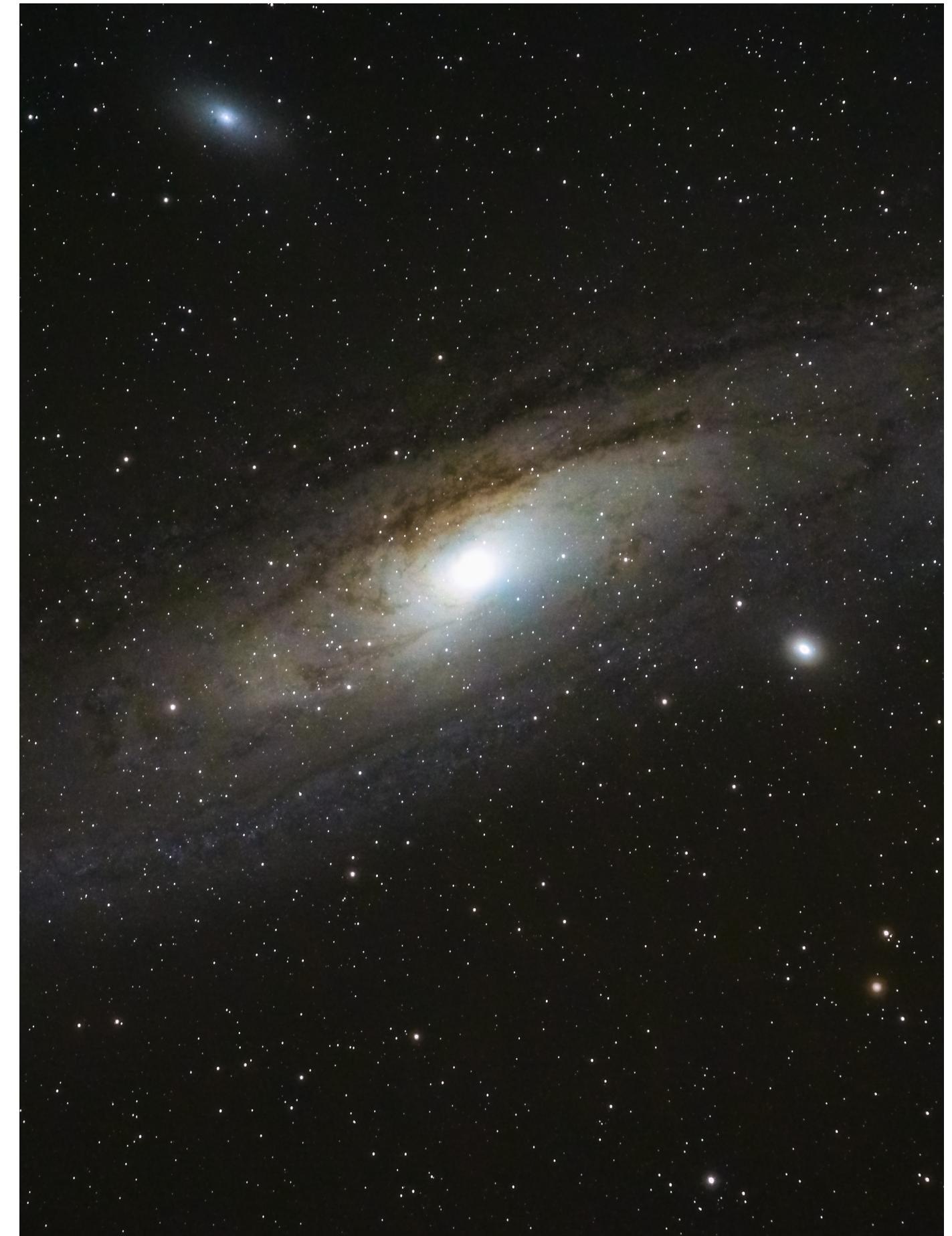
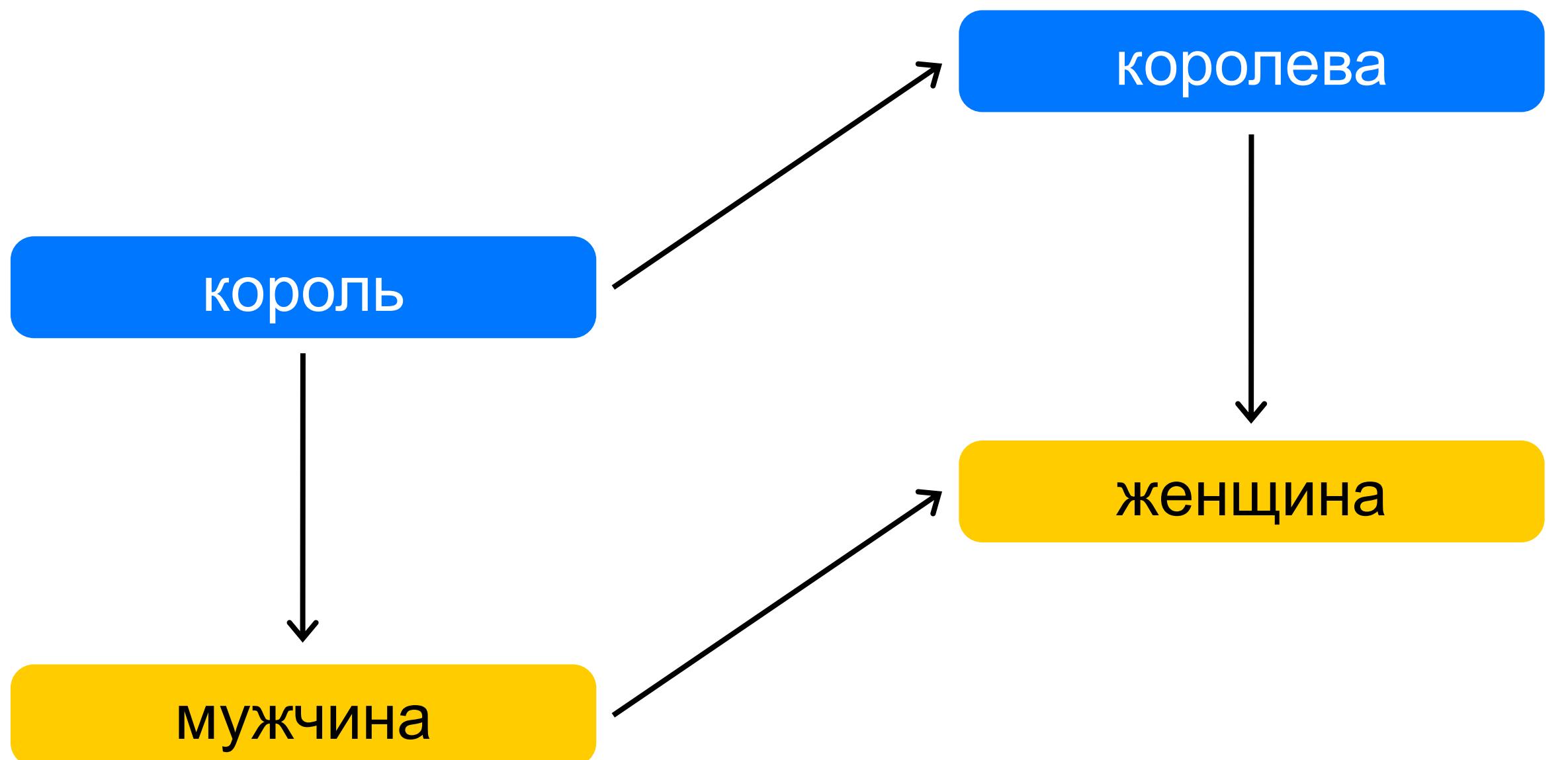
Пример: Word2Vec

Отношения между словами
в векторном пространстве

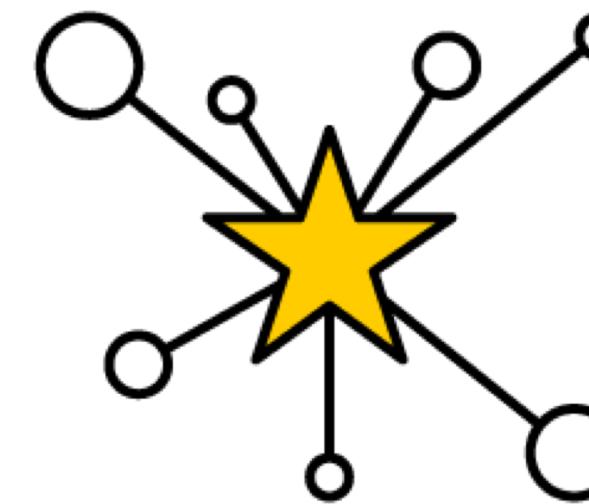
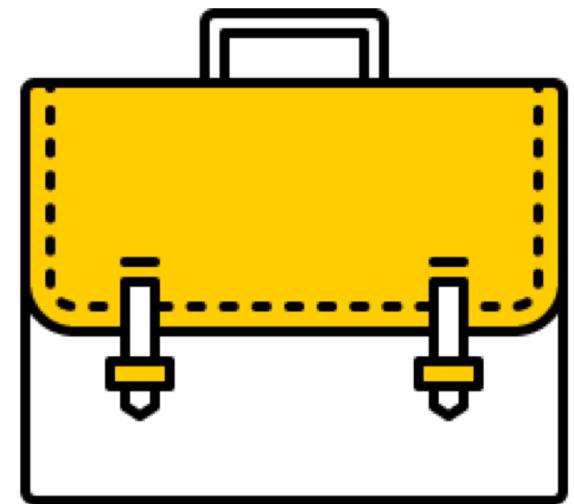


Пример: Word2Vec

Отношения между словами
в векторном пространстве



Представление текста для машинного обучения



Мешок слов (Bag of words)

- › One Hot Encoding
- › Разреженная матрица термы на документы
- › Не учитывает порядок слов
- › Легкость и простота применения – TF/IDF
- › Нельзя использовать для языковой модели – нет информации о последовательности слов
- › Сложность добавления новых слов

Плотные векторные представления (Embedding)

- › Каждое слово представляется в виде вектора чисел
- › Позволяет сохранять информацию о контексте в котором встречается данное слово
- › Примеры – Word2Vec, Fasttext, GloVe
- › Может быть использована для языковой модели
- › Легкость добавления новых слов
- › Требует много текста для обучения

Как учесть
порядок слов
в тексте?

Нам нужна языковая модель!



Что такое «Языковая модель»?

Представляет модель языка

Хотим присвоить вероятности
последовательностям слов

Вычисляет вероятность появления слов
после данного текста

Применение языковой модели



- › Генерация текста
- › Распознавание речи
скрип колеса vs. скрипка лиса
- › Исправление опечаток
курсовая работа vs. курсовая робота
- › Машинный перевод
- › Классификация текста

Проблемы при обучении языковой модели

Мало данных

- › Невозможно обучать
- › Низкое качество модели

Коробка лежала, коробка лежала, лежала,
лежала на складе, складе складе складе
складе, складе, складе, складе, складе
складе, складе, складе, складе, складе
складе, складе, складе, складе, складе

Проблемы при обучении языковой модели

Мало данных

- › Невозможно обучать
- › Низкое качество модели

Коробка лежала, коробка лежала, лежала,
лежала на складе, складе складе складе
складе, складе, складе, складе, складе
складе, складе, складе, складе, складе
складе, складе, складе, складе, складе

Достаточно данных

- › Сложно обучать
- › Долго обучать

Проблемы при обучении языковой модели

Мало данных

- › Невозможно обучать
- › Низкое качество модели

Коробка лежала, коробка лежала, лежала,
лежала на складе, складе складе складе
складе, складе, складе, складе, складе
складе, складе, складе, складе, складе
складе, складе, складе, складе, складе

Достаточно данных

- › Сложно обучать
- › Долго обучать

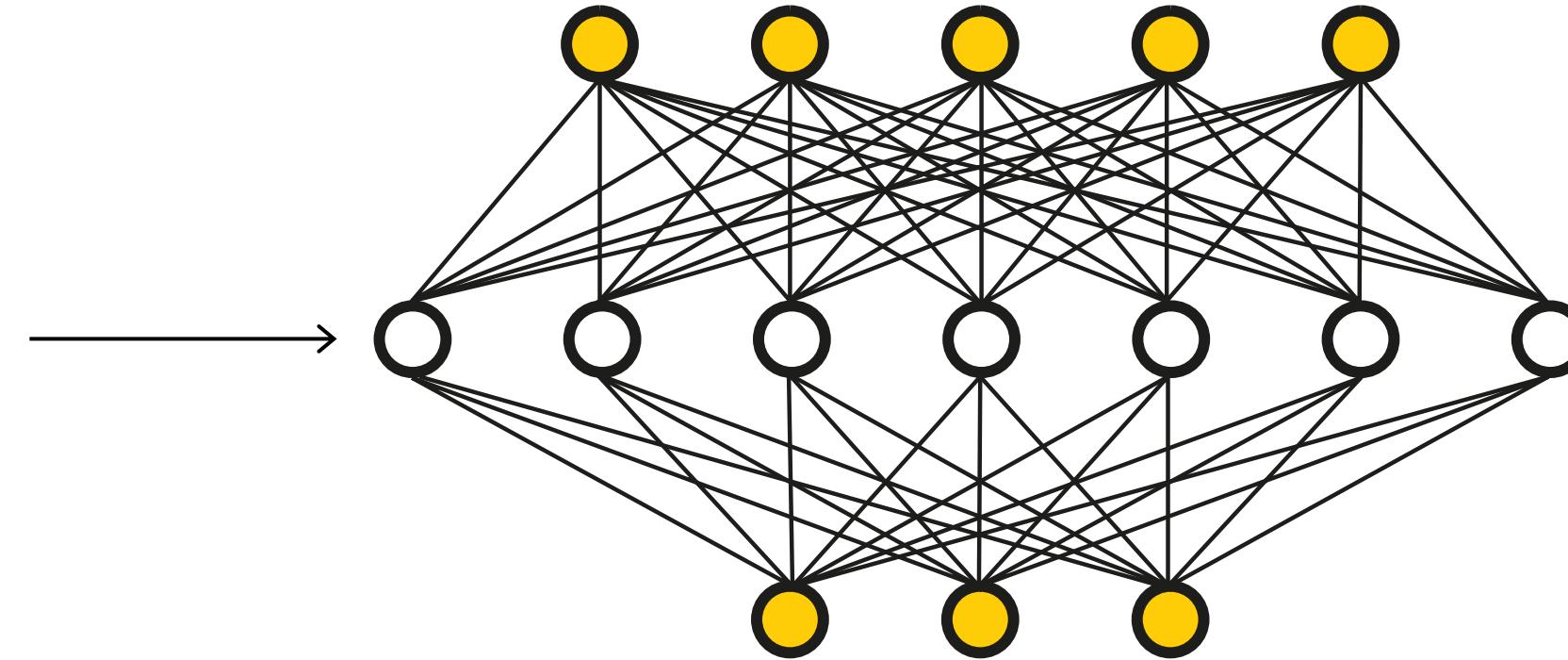
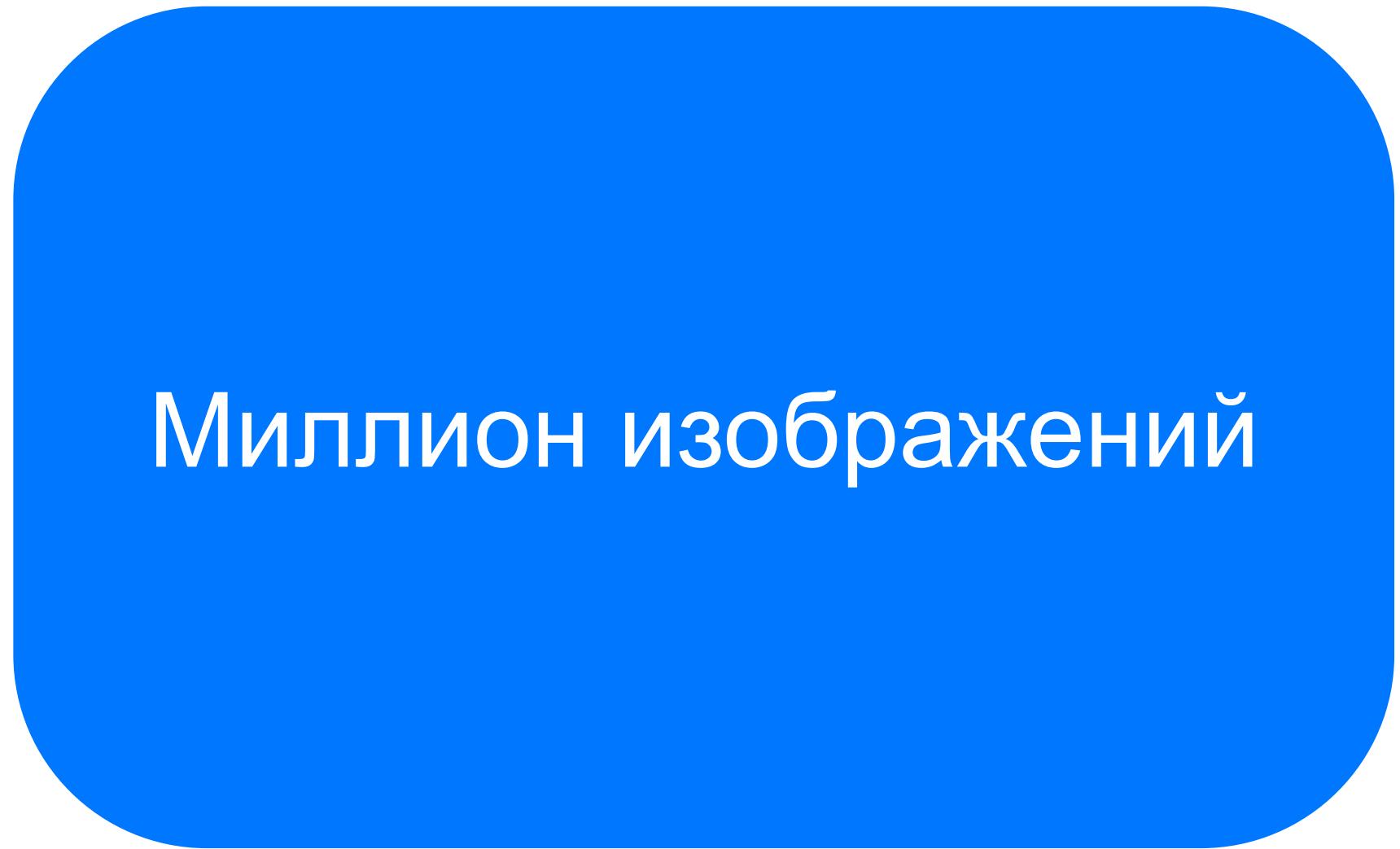


Выход есть

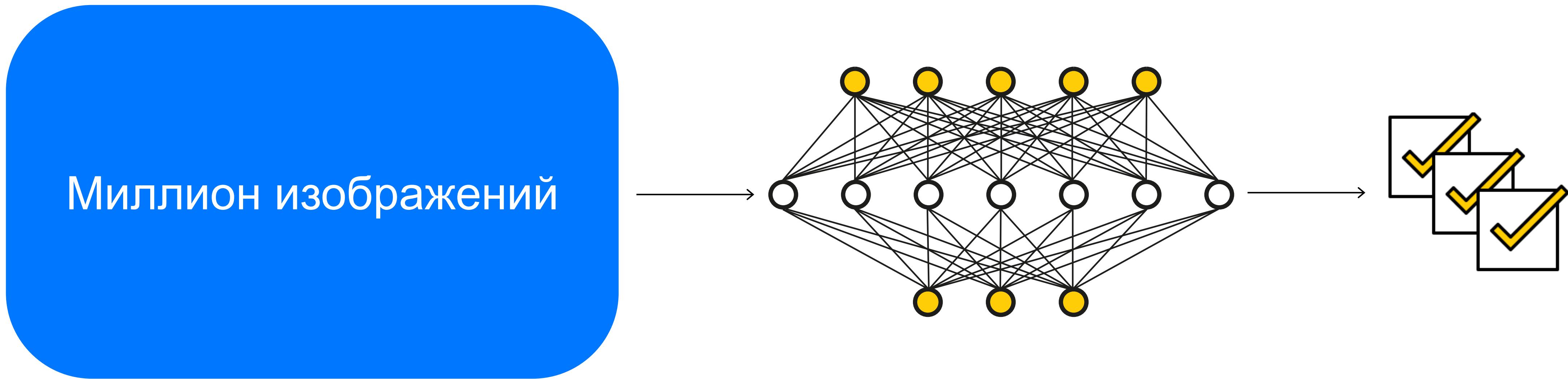
Перенос обучения (Transfer Learning)



Перенос обучения



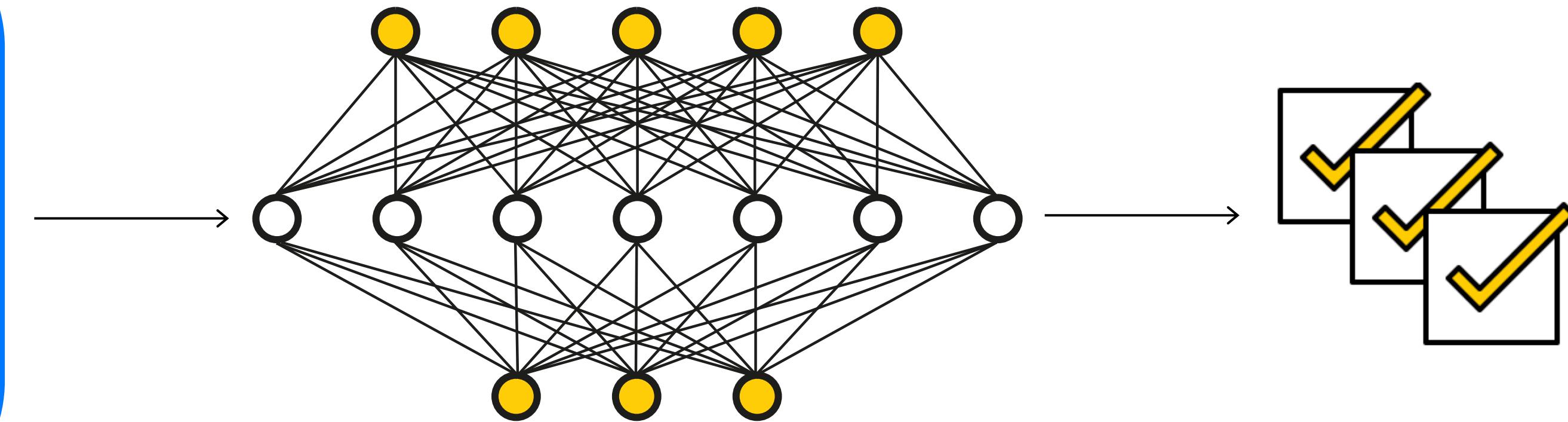
Перенос обучения



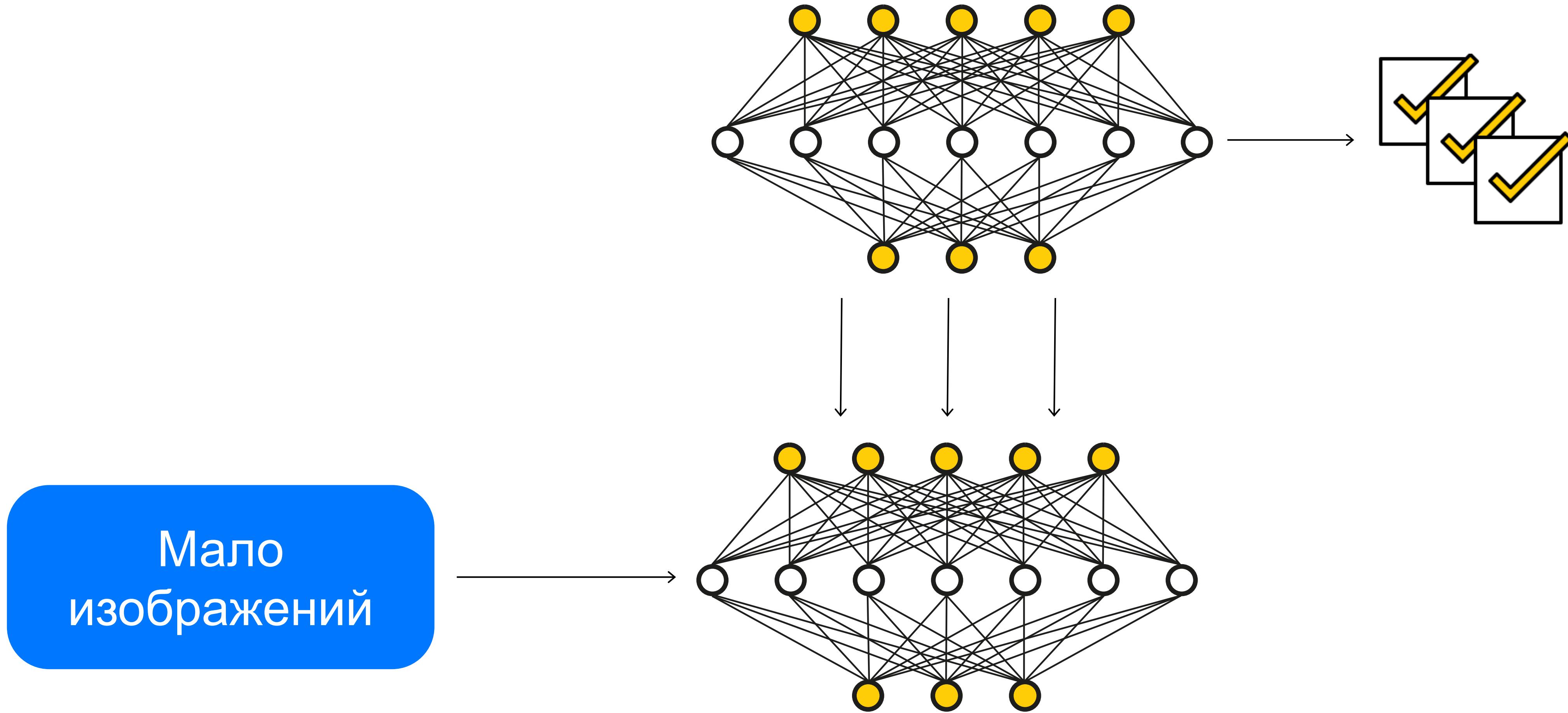
Перенос обучения

Миллион изображений

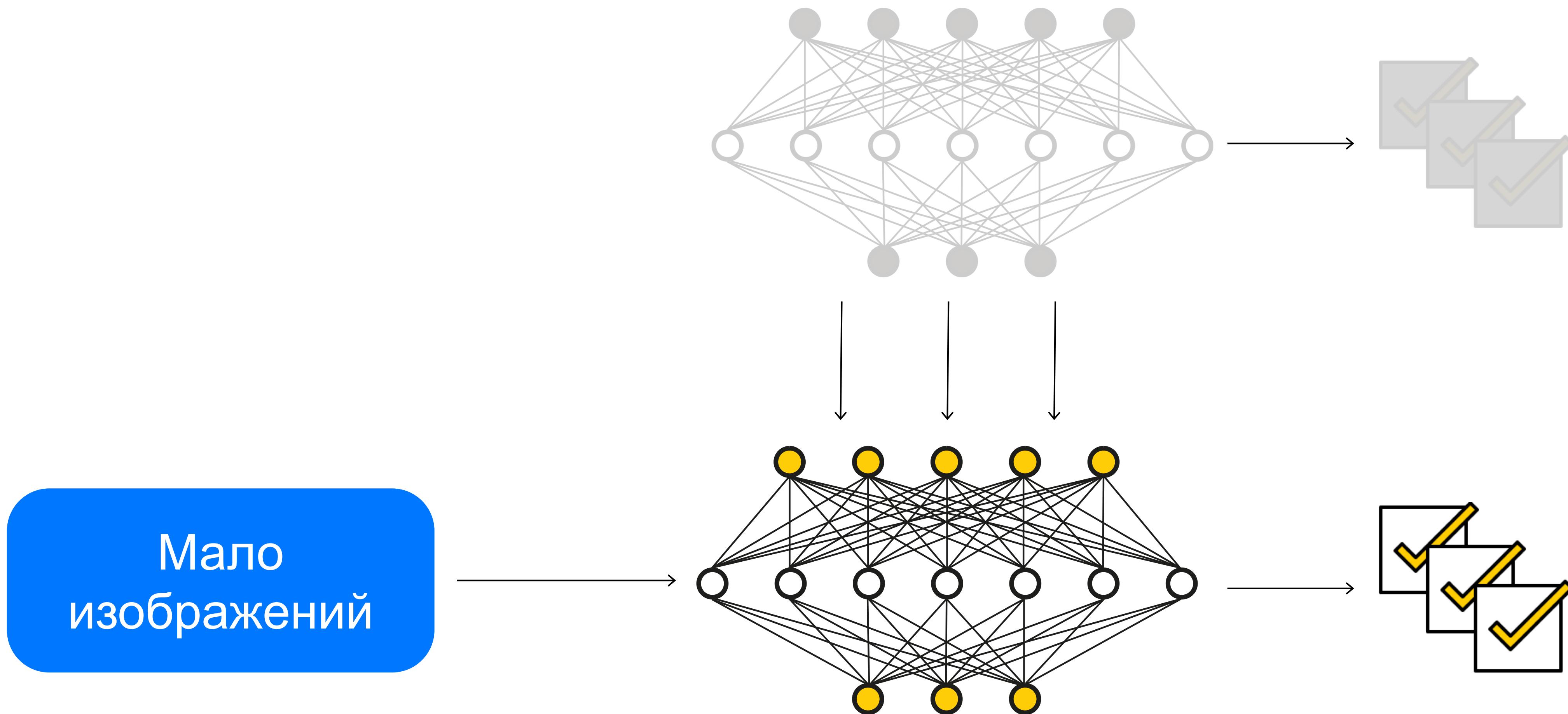
Мало
изображений



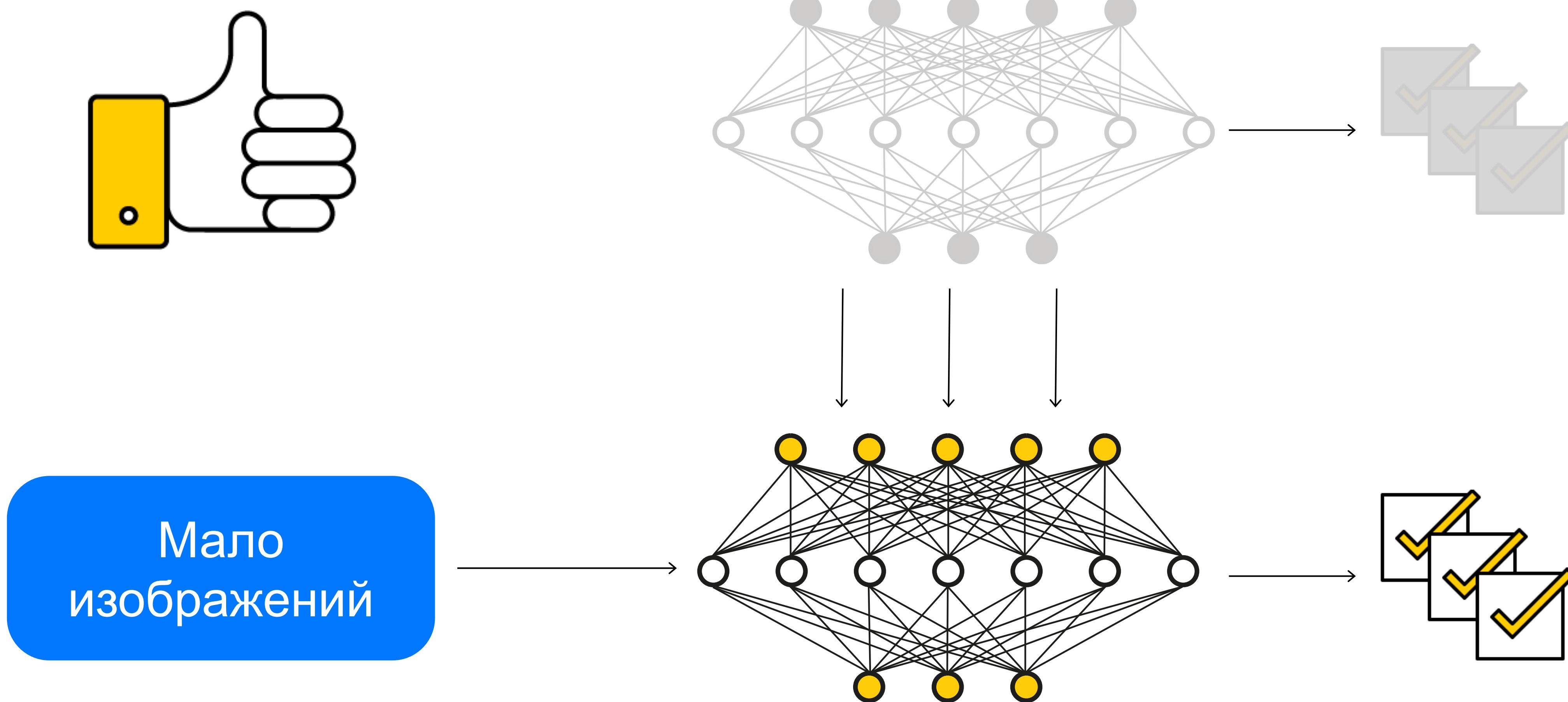
Перенос обучения



Перенос обучения



Перенос обучения



Перенос обучения (Transfer learning)



Сеть заданной архитектуры (Resnet) обучается на миллионах изображений (ImageNet) для решения задачи классификации на 1000 классов



Используя веса обученной сети применяем ее для решения задачи классификации на своем наборе данных



Обучение происходит быстро, полученная модель имеет высокую точность

ULMFiT

Universal Language Model Fine-Tuning
for Text Classification

Использование идеи переноса обучения к задачам обработки текста

- › Берем за основу языковую модель, обученную на большом корпусе текстов
- › «Дотюниваем» языковую модель для работы с собственным текстовым набором



A close-up photograph of a cat's face, focusing on its left eye and ear. The cat has dark fur and is looking slightly away from the camera. Overlaid on the image is a large, white, semi-transparent text area containing the question "Попробуем?".

Попробуем?

Классификация отзывов о фильмах

IMDB Movie Review dataset

- › 25000+25000 положительных и отрицательных отзывов с imdb.com
- › <https://www.kaggle.com/iarunava/imdb-movie-reviews-dataset>

Возьмем русские отзывы с Кинопоиска!

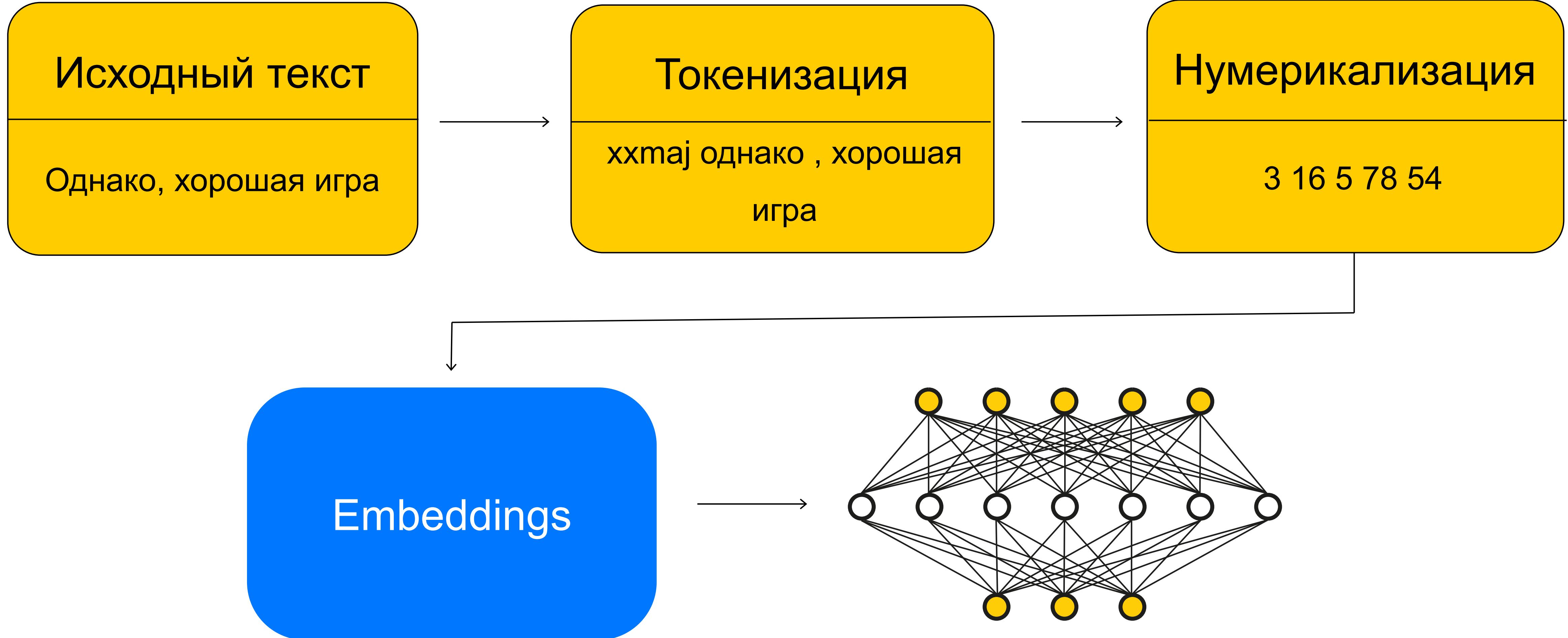
- › 15000 положительных и отрицательных отзывов с kinopoisk.ru



itis.sotov.xyz



Подготовка данных



Архитектура AWD-LSTM

Рекуррентная нейронная сеть

Размер словаря – 60 004 слова

Размер векторов Embedding-ов – 400

Содержит 3 скрытых слоя,
каждый по 1150 LSTM-блоков

Использует лучшие практики
регуляризации при обучении



Обучение нейронной сети

Шаг обучения –
Важный гиперпараметр

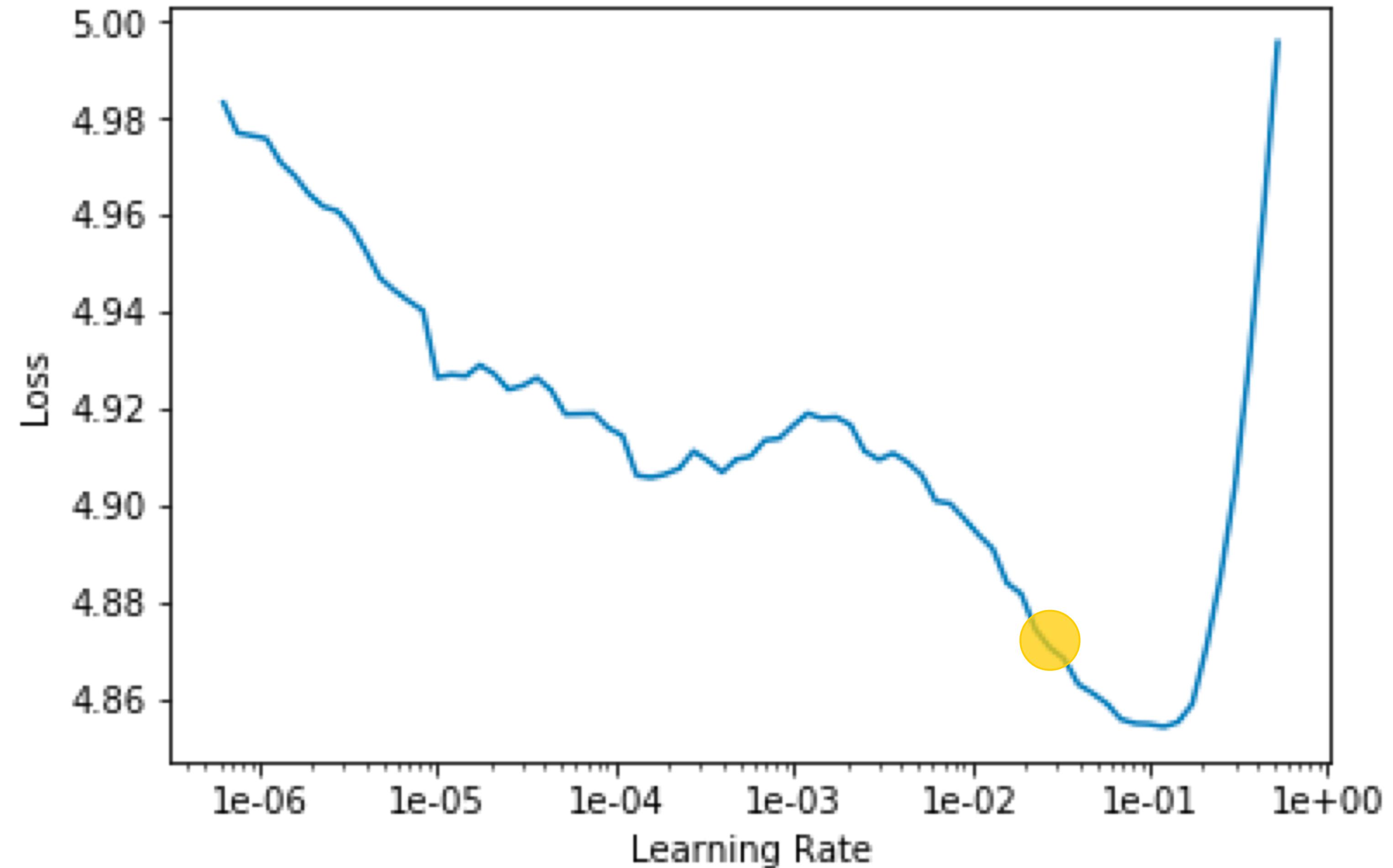
Он определяет скорость обучения нейронной сети

Выбираем оптимальный шаг обучения

Leslie Smith's One Cycle Policy

Обучаем сеть,
изменяя
learning rate
во время обучения

[https://sgugger.github.io/
the-1cycle-policy.html](https://sgugger.github.io/the-1cycle-policy.html)

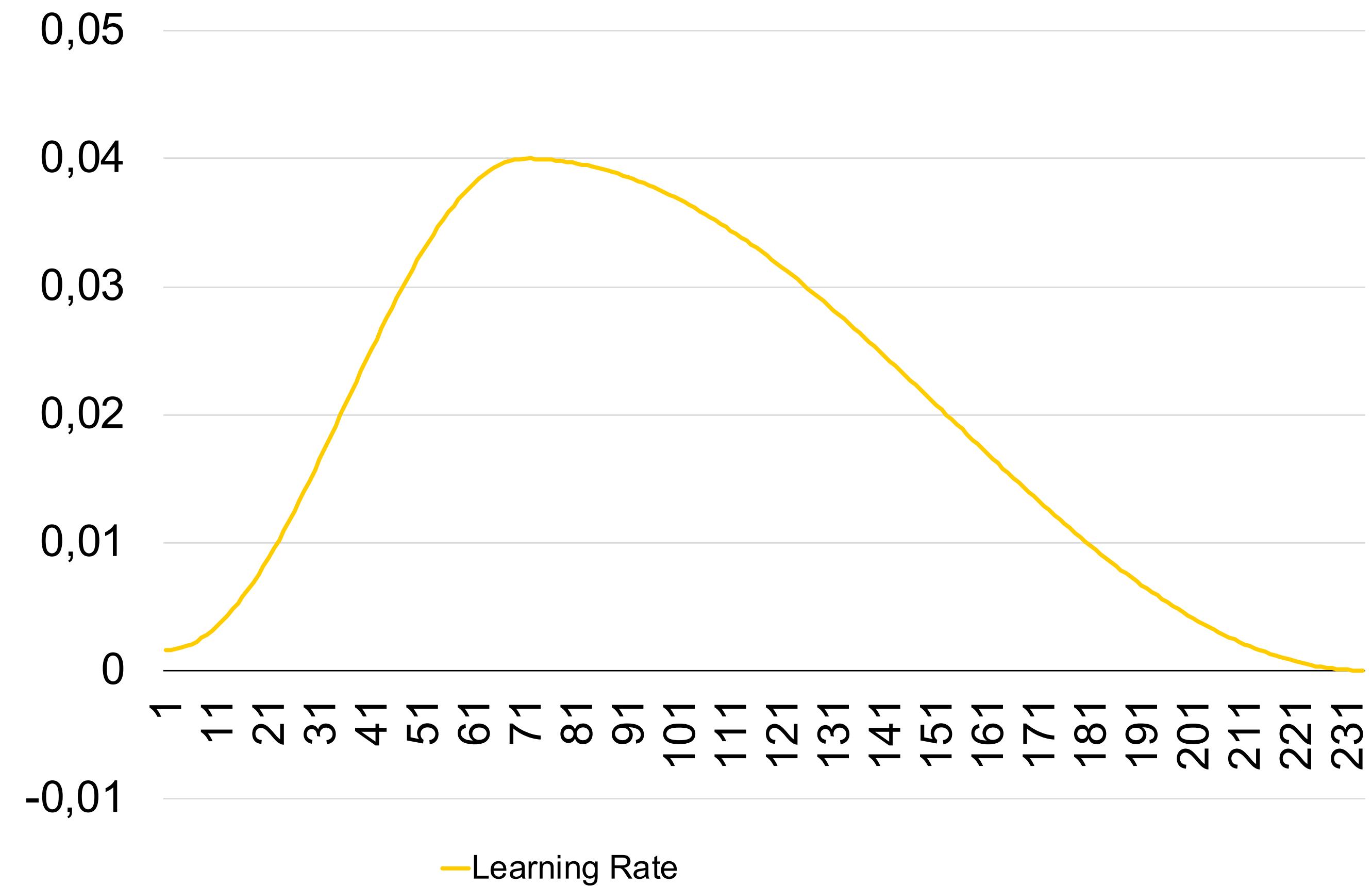


Правильно обучаем сеть

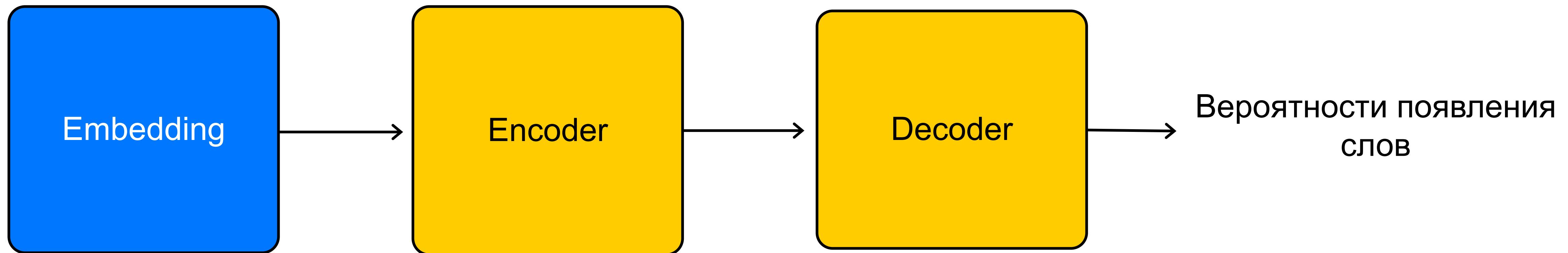
Leslie Smith's One Cycle Policy

Обучаем сеть,
изменяя
learning rate
во время обучения

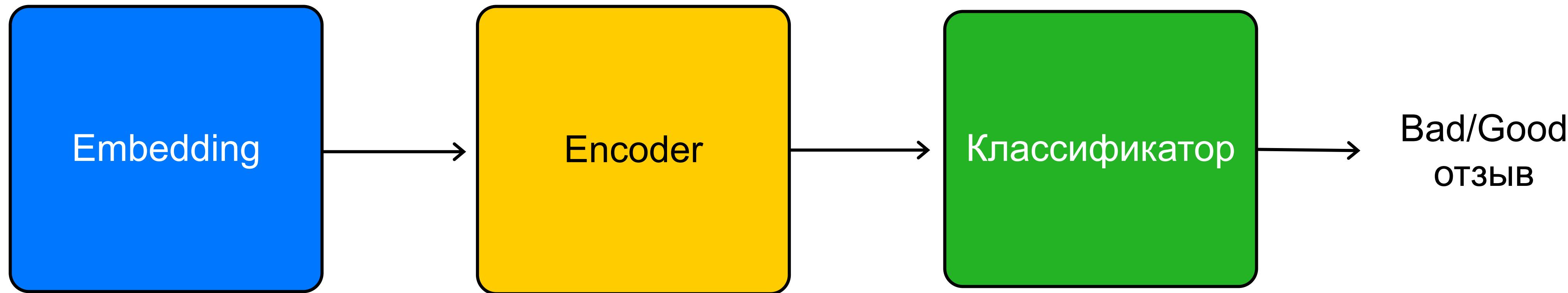
[https://sgugger.github.io/
the-1cycle-policy.html](https://sgugger.github.io/the-1cycle-policy.html)



Архитектура сети (Языковая модель)



Архитектура сети (Классификатор)



Современные архитектуры

Transformer – не рекуррентная сеть

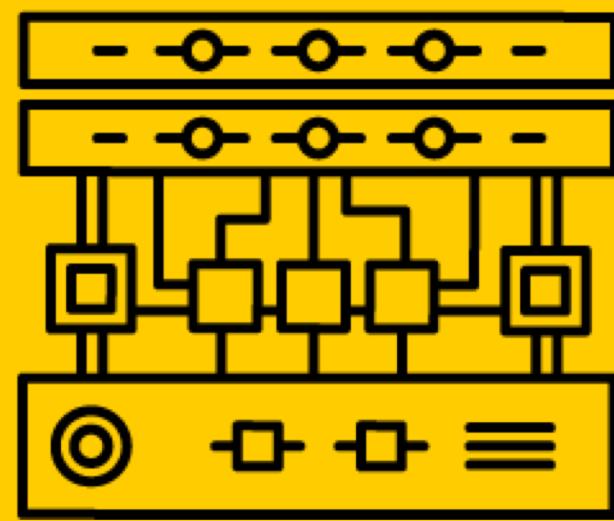
| **Все что нужно – это «внимание»!**

Примеры реализаций:

- › BERT от Google
- › GPT-2 от Open AI



Есть над чем задуматься



Этичность использования AI



{IT.IS}

КОНФЕРЕНЦИЯ
КОМПАНИИ "ИНТЕРСВЯЗЬ"
ДЛЯ IT-СПЕЦИАЛИСТОВ

Спасибо

Алексей Сотов

разработчик,
специалист по анализу данных

 ptenec@yandex-team.ru

 @asotov

Ссылки на материалы

| **Код Jupyter-ноутбука для запуска:**

› <https://github.com/xechehot/itis-fast-ai-text>

| **Google Collab:**

› <https://colab.research.google.com>

| **Документация Fast AI**

› <https://docs.fast.ai>