

# **Active Learning**

# Сложность получения данных

Любой датасет для обучения с учителем нужно размечать **вручную**

Разметка текстов – сложная и дорогая работа  
(может стоить сотни тысяч рублей в зависимости от задачи)

# Сложность получения данных

Любой датасет для обучения с учителем нужно размечать **вручную**

Разметка текстов – сложная и дорогая работа  
(может стоить сотни тысяч рублей в зависимости от задачи)

Сколько времени у вас займет выделение именованных сущностей в тексте?

Нормунд Вилтинис руководил антикоррупционным бюро с 2009 года. За время его пребывания на этом посту, как отмечает "Телеграф", сотрудники КНАВ провели ряд громких задержаний: под стражу были взяты, в частности, бывший глава Latvenergo Карлис Микельсонс и бывший начальник криминального управления таможни Владимир Вашкевич. Одним из последних громких дел КНАВ стала серия обысков в фирмах, связанных с известными латвийскими олигархами.

# Сложность получения данных

Иногда разметка требует высокой квалификации асессора

- Знание двух языков для задачи перевода
- Лингвистическое образование для разметки грамотности текста

Что делать, если размеченных данных требуется много, а бюджет на разметку ограничен?

# Active Learning

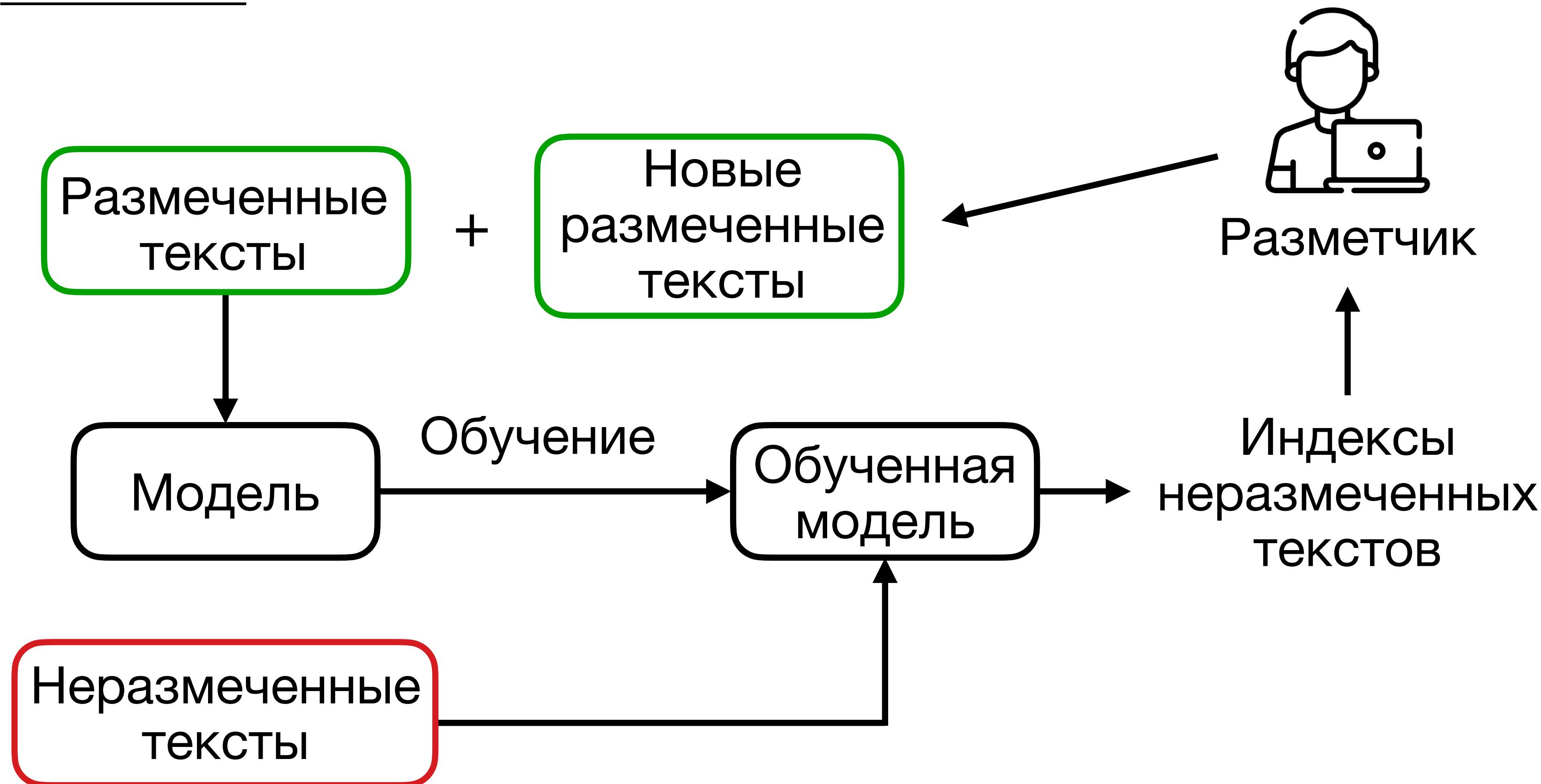
Парадигма обучения моделей в условиях ограниченного бюджета на разметку.

**Идея:** чередуем обучение модели с доразметкой данных, размечая только наиболее полезные тексты

# Active Learning

Парадигма обучения моделей в условиях ограниченного бюджета на разметку.

**Идея:** чередуем обучение модели с доразметкой данных, размечая только наиболее полезные тексты



# Active Learning

Active Learning потерял популярность на время, но снова ее обрел с появлением Трансформера в 2017 году.

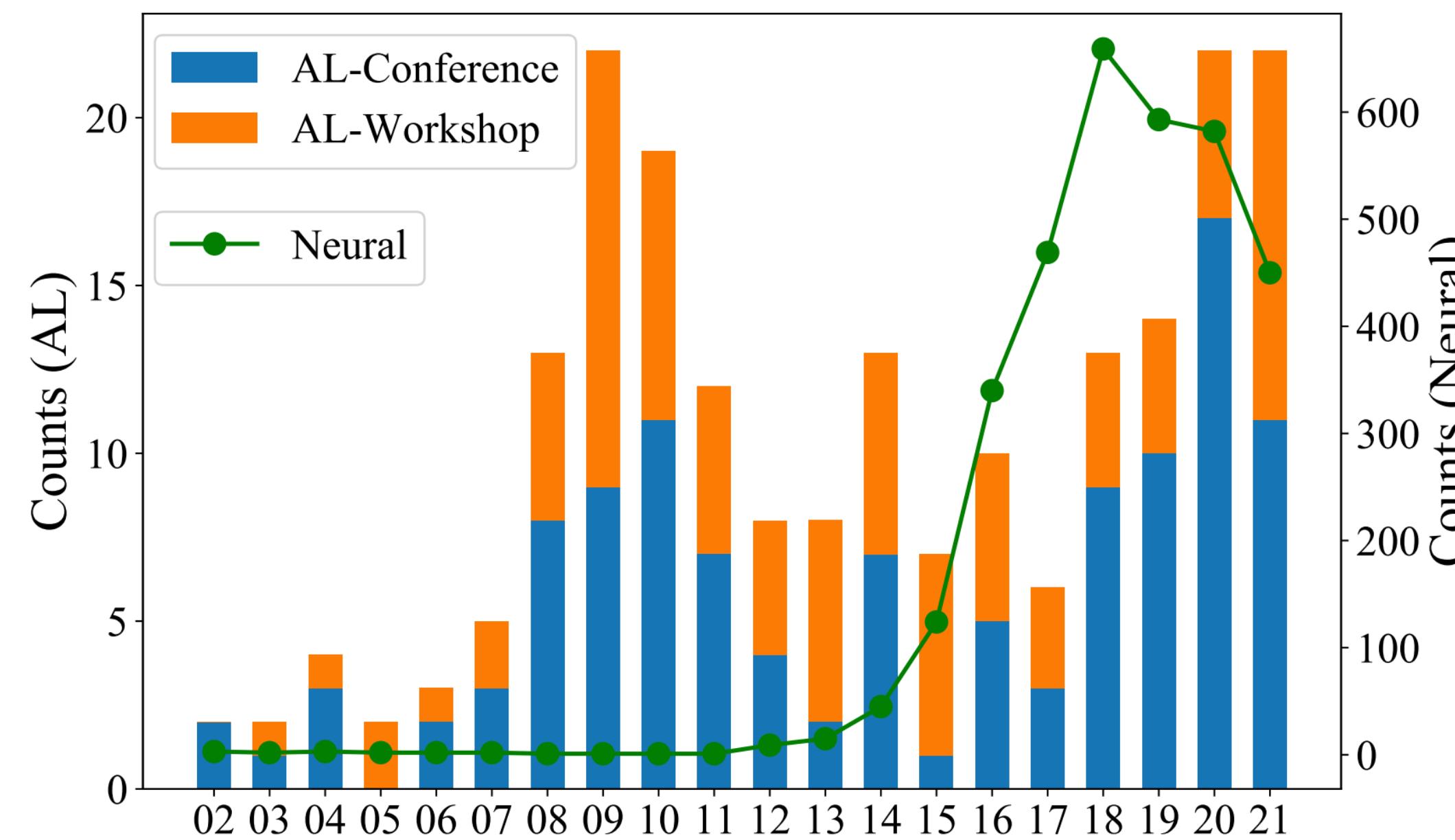


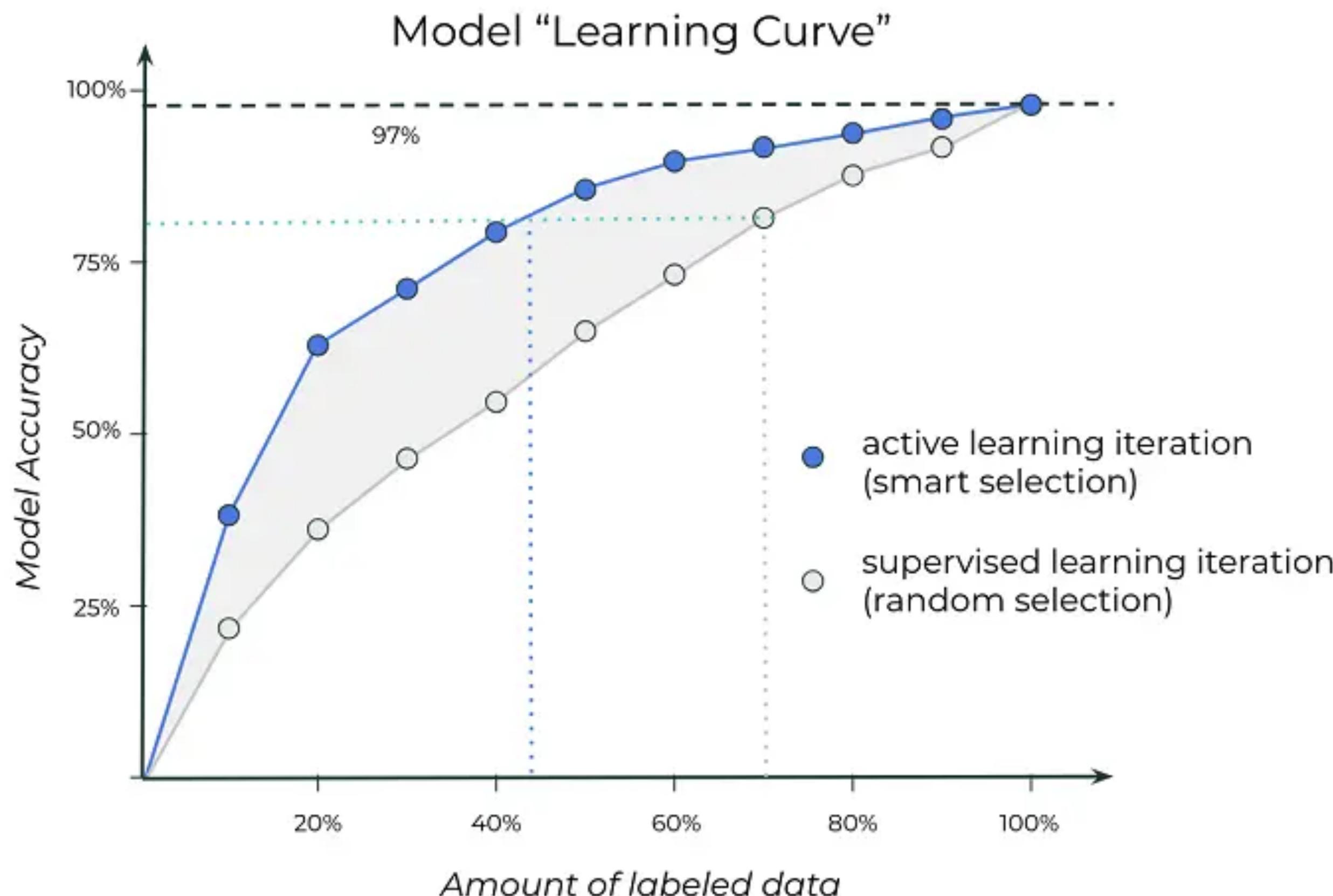
Figure 1: Counts of AL (left) and “neural” (right) papers in the ACL Anthology over the past twenty years.

# Active Learning: основные вопросы

- Как выбирать тексты для разметки?
- Как дообучать модель?
- В какой момент остановиться?
- Как ускорить цикл обучения и разметки?

# Как выбирать тексты для разметки?

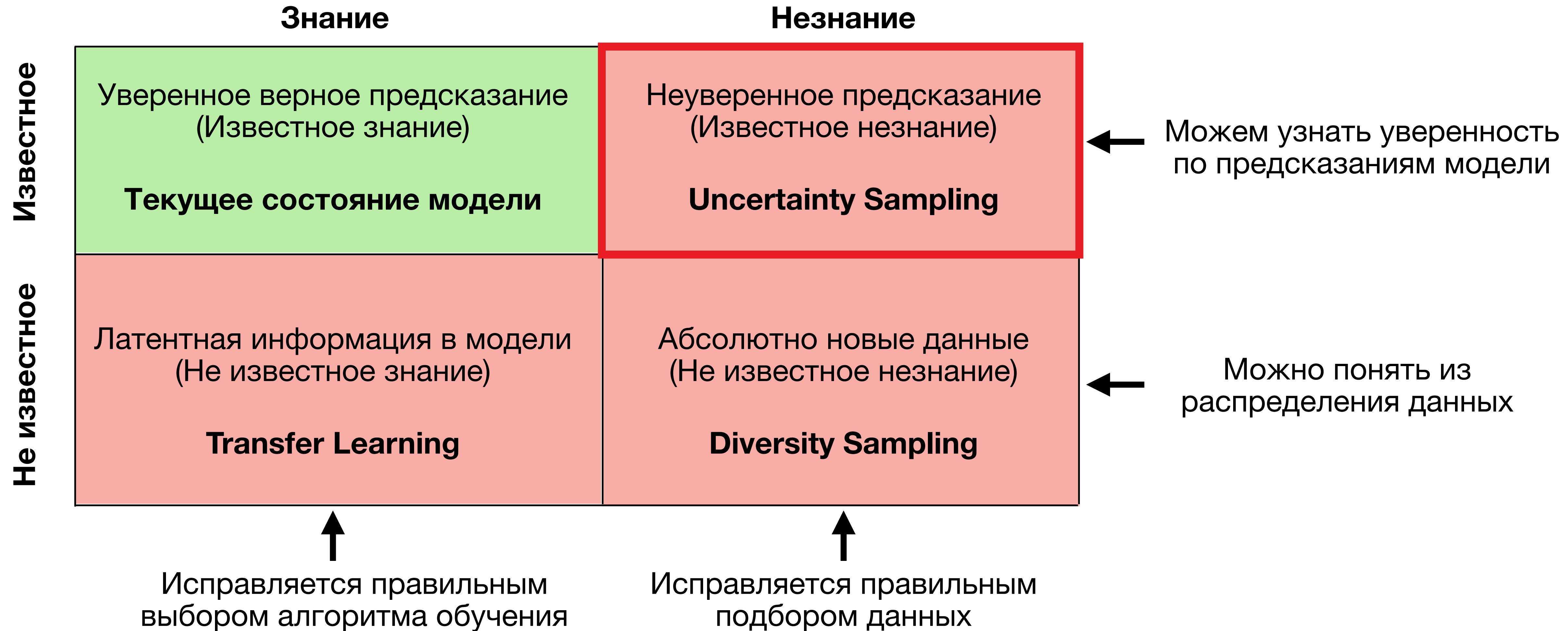
- Выбрать случайные
- Выбирать тексты, на которых важнее всего обучиться



# Квадрат знаний модели

	Знание	Незнание
Известное	Уверенное верное предсказание (Известное знание)	Неуверенное предсказание (Известное незнание)
Не известное	<b>Текущее состояние модели</b>	<b>Uncertainty Sampling</b>
	Латентная информация в модели (Не известное знание)	Абсолютно новые данные (Не известное незнание)
	<b>Transfer Learning</b>	<b>Diversity Sampling</b>

# Квадрат знаний модели

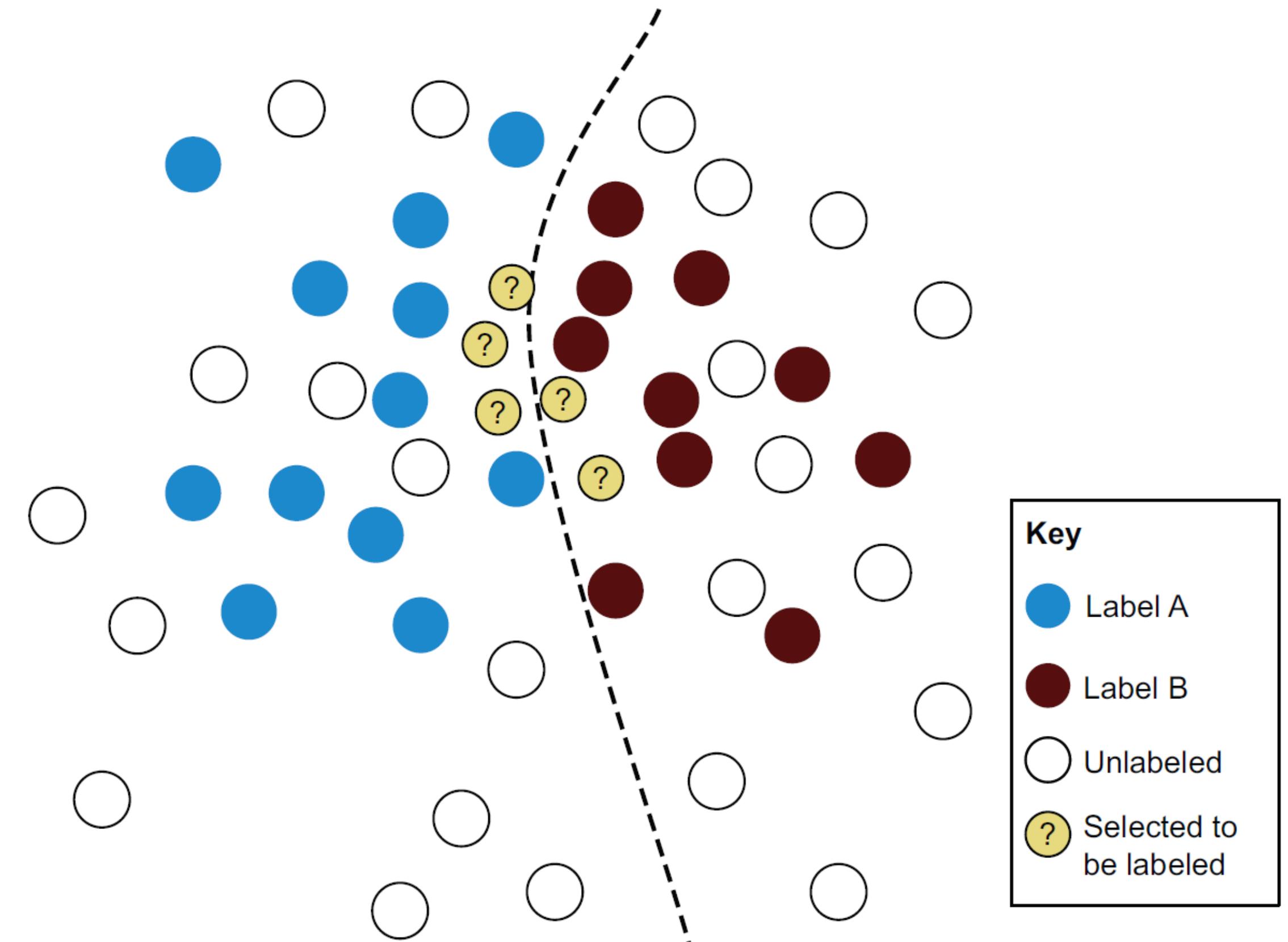


# Uncertainty Sampling

- Важно обучить модель на тех объектах, в которых она не уверена
- Такие объекты располагаются около разделяющей кривой

Алгоритм:

- Считаем уверенность для всех неразмеченных объектов
- Отправляем  $n$  самых неуверенных на разметку
- Обучаем модель заново с этими объектами



# Как считать уверенность?

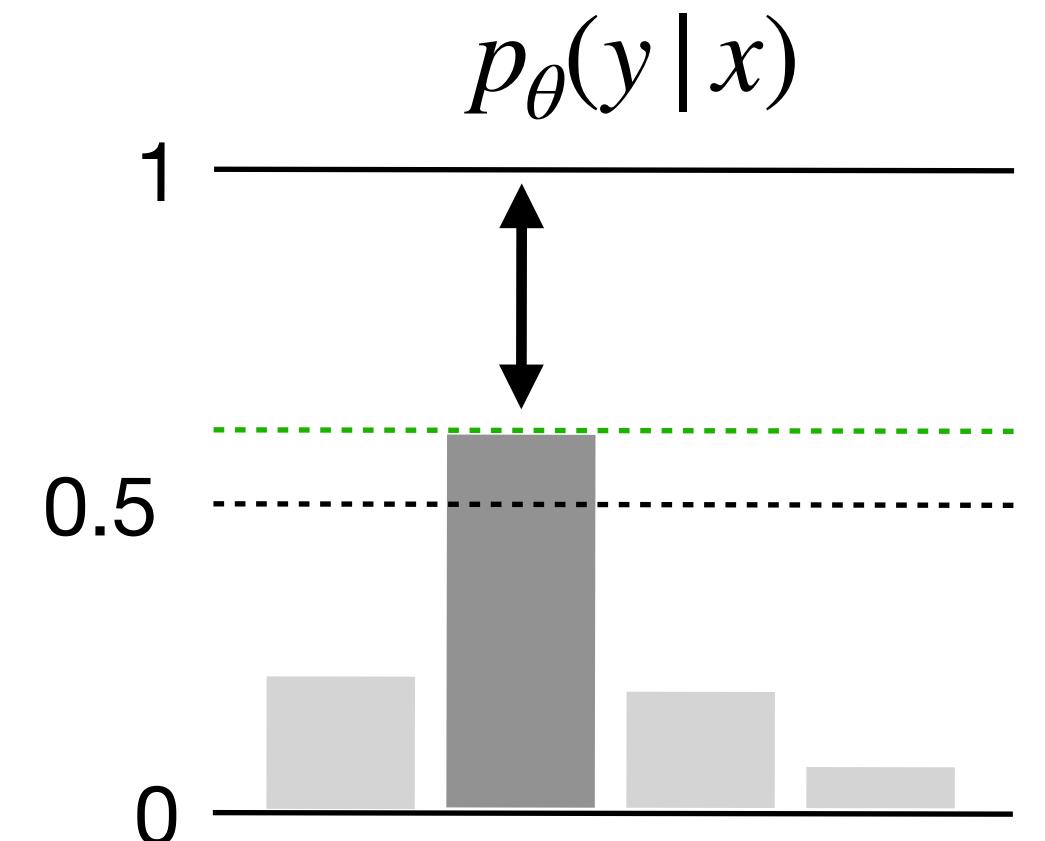
- Least Confident
- Margin of Confidence
- Ratio of Confidence
- Label Entropy
- Query by Committee

# Least Confident

Оцениваем уверенность по вероятности наиболее вероятного класса.

$$\text{score}(x) = 1 - p_{\theta}(y_{(1)} | x),$$

где  $y_{(1)}$  – самый вероятный класс

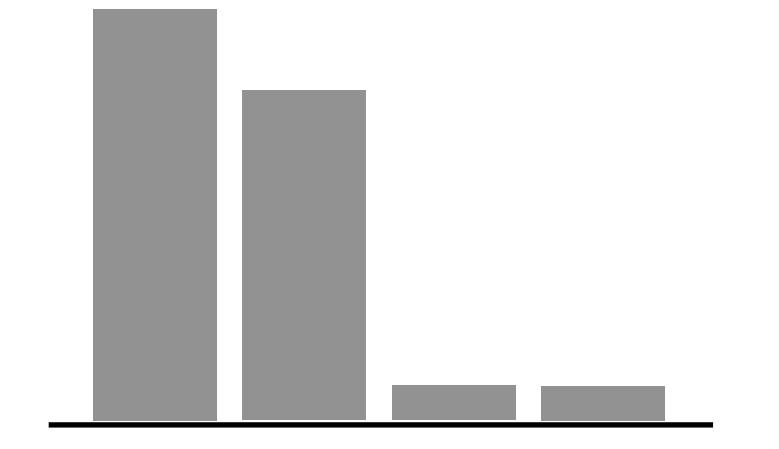
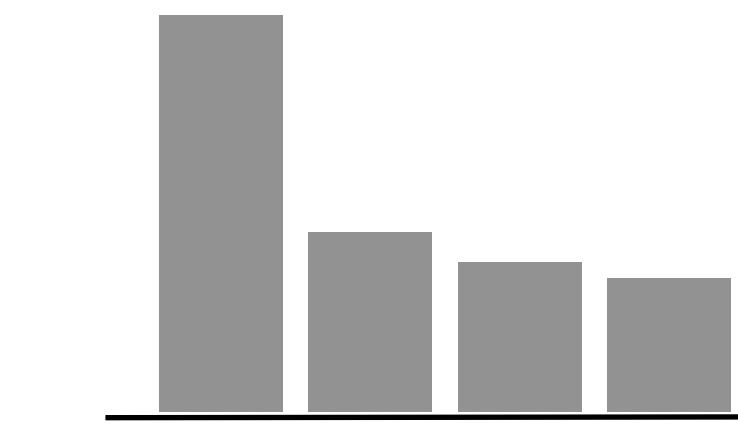


Для разметки выбираются объекты с наибольшим score.

# Least Confident: недостатки

- Метод хорошо работает для бинарной классификации
- Для мультиклассовой хуже, так как не учитывает вероятности других классов

это распределение  
более вырожденное,  
но максимальная  
вероятность у него  
ниже



у этого выше  
максимальная  
вероятность, но  
модель не уверена

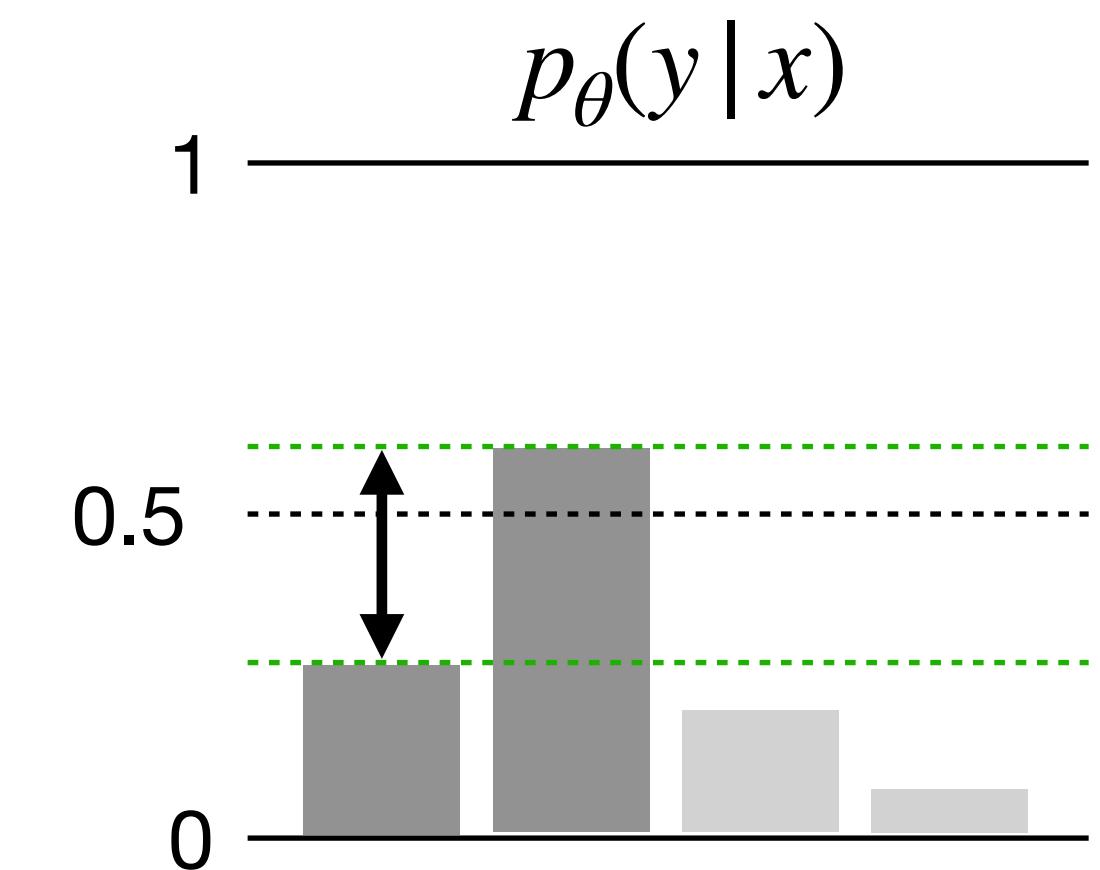
# Margin of Confidence

Считаем разницу между двумя наиболее вероятными классами

$$\text{score}(x) = 1 - (p_\theta(y_{(1)} | x) - p_\theta(y_{(2)} | x)),$$

где  $y_{(1)}$  и  $y_{(2)}$  – первый и второй наиболее вероятные классы

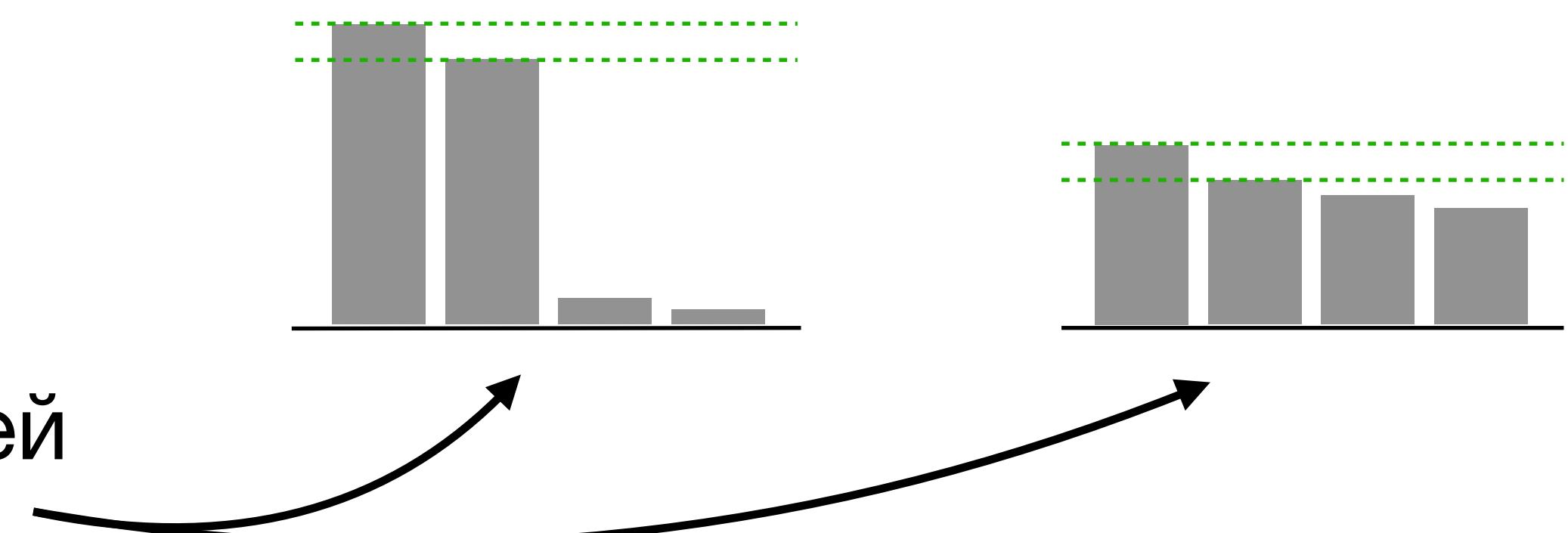
Этот метод используется чаще всего.  
Но он тоже не всегда правильно работает.



# Margin of Confidence: недостатки

- score не зависит от остальных вероятностей
- Смотря на разность, мы не учитываем сами значения вероятностей

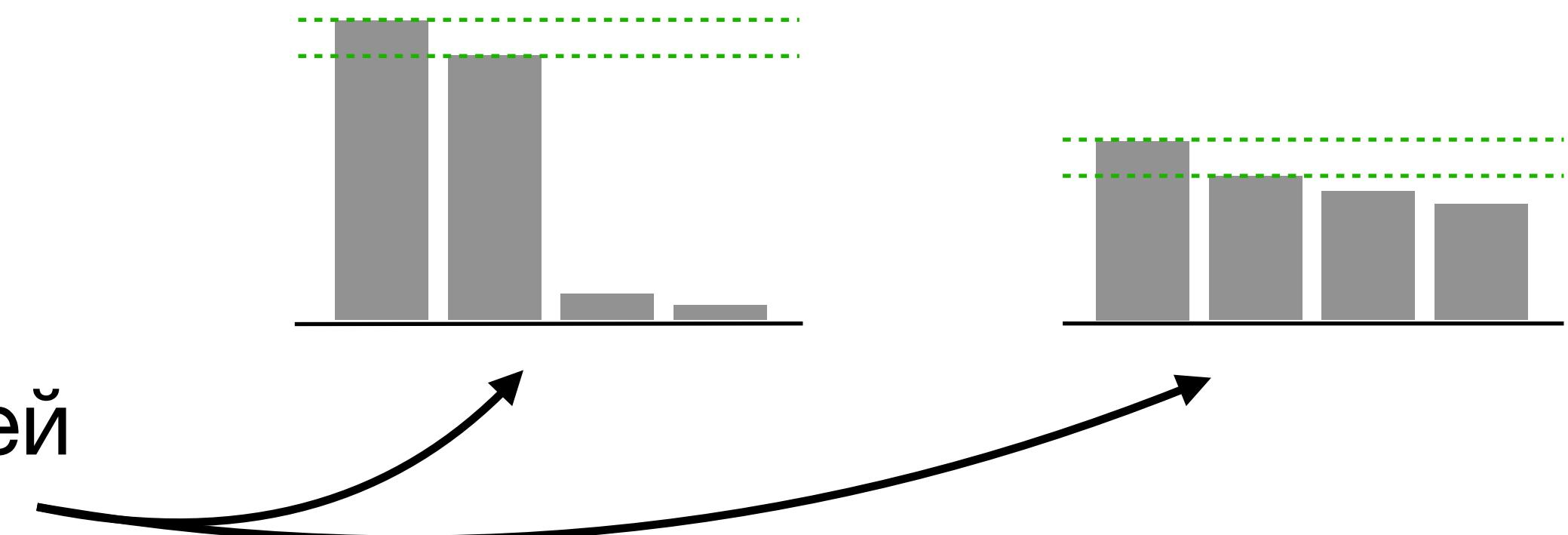
Для этих распределений разность вероятностей двух наиболее вероятных классов совпадает



# Margin of Confidence: недостатки

- score не зависит от остальных вероятностей
- Смотря на разность, мы не учитываем сами значения вероятностей

Для этих распределений разность вероятностей двух наиболее вероятных классов совпадает



- Чтобы исправить этот недостаток, можно настроить параметр температуры в softmax
- Иногда лучше себя показывает Ratio of Confidence

# Ratio of Confidence

$$p_{\theta}(y_i | x) = \text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{k=1}^K e^{z_k}}$$

- Считаем отношение вероятностей вместо разности

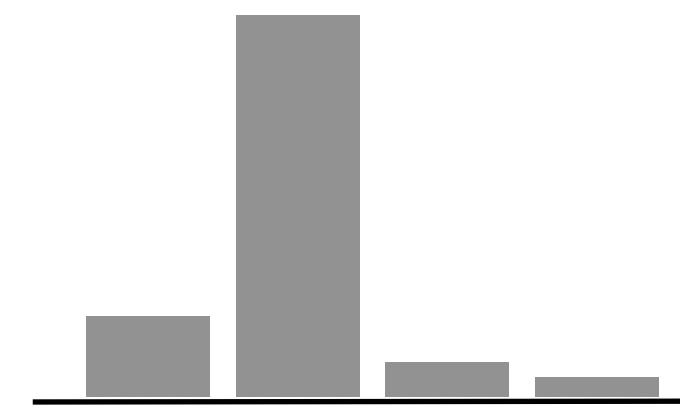
$$\text{score}(x) = \frac{p_{\theta}(y_{(2)} | x)}{p_{\theta}(y_{(1)} | x)} = e^{z_{(2)} - z_{(1)}}$$

- Получаем разность логитов вместо разности вероятностей

# Label Entropy

Энтропия – мера неопределённости распределения вероятностей

$$\text{score}(x) = - \sum_{k=1}^K p_\theta(y_k | x) \log p_\theta(y_k | x)$$



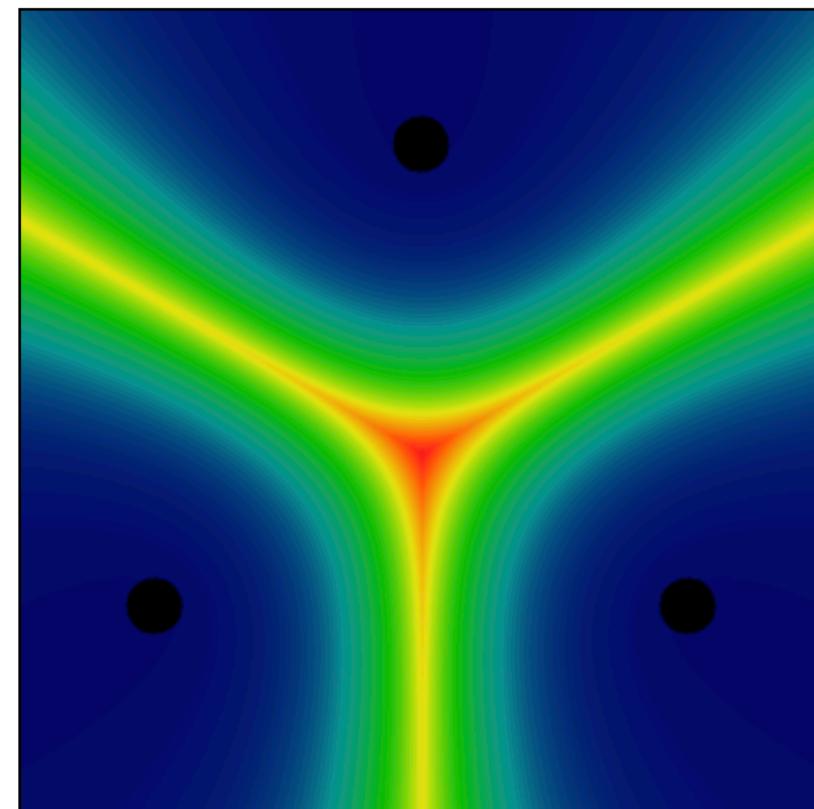
низкая энтропия



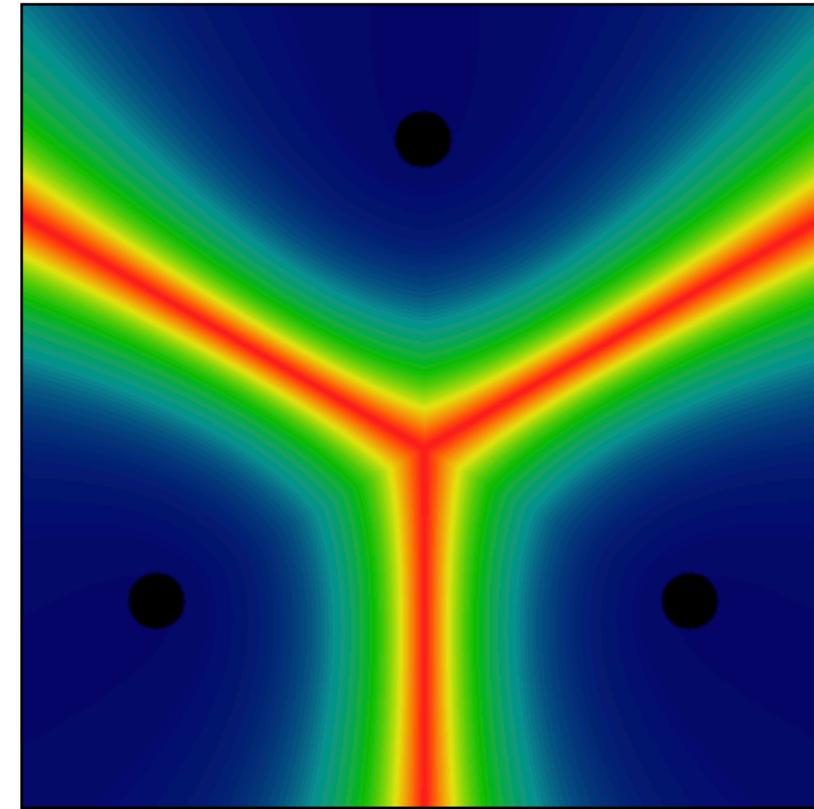
высокая энтропия

# Визуальное сравнение методов

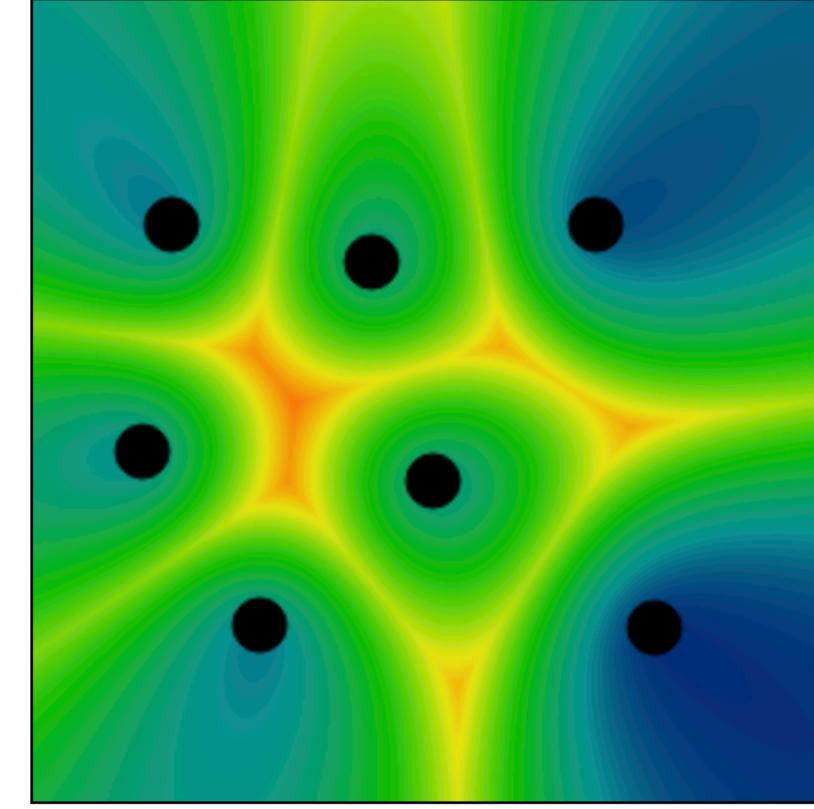
Least Confident



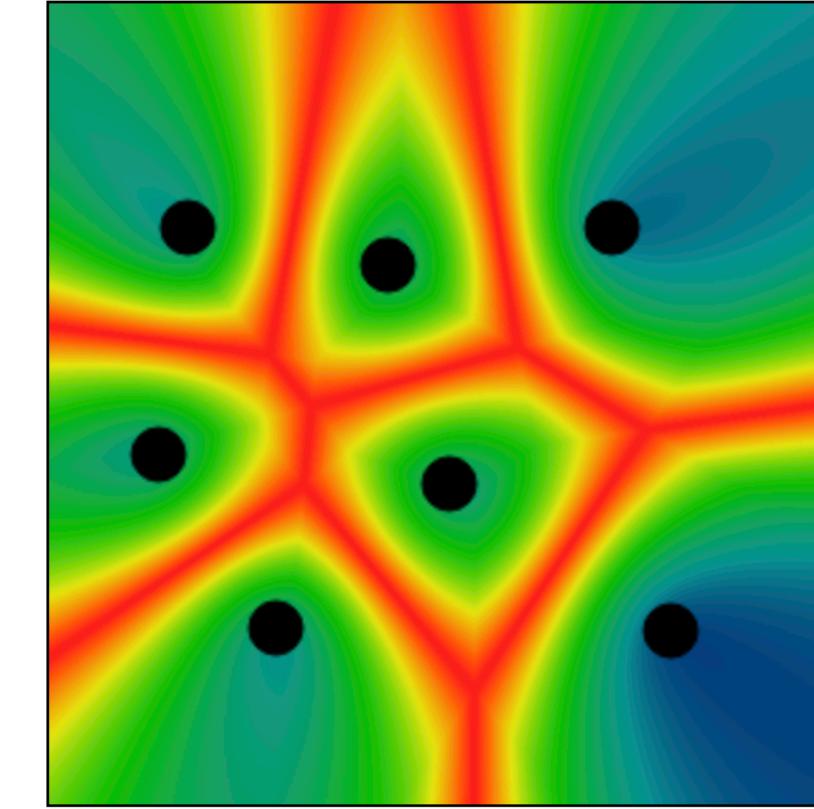
Margin of Confidence



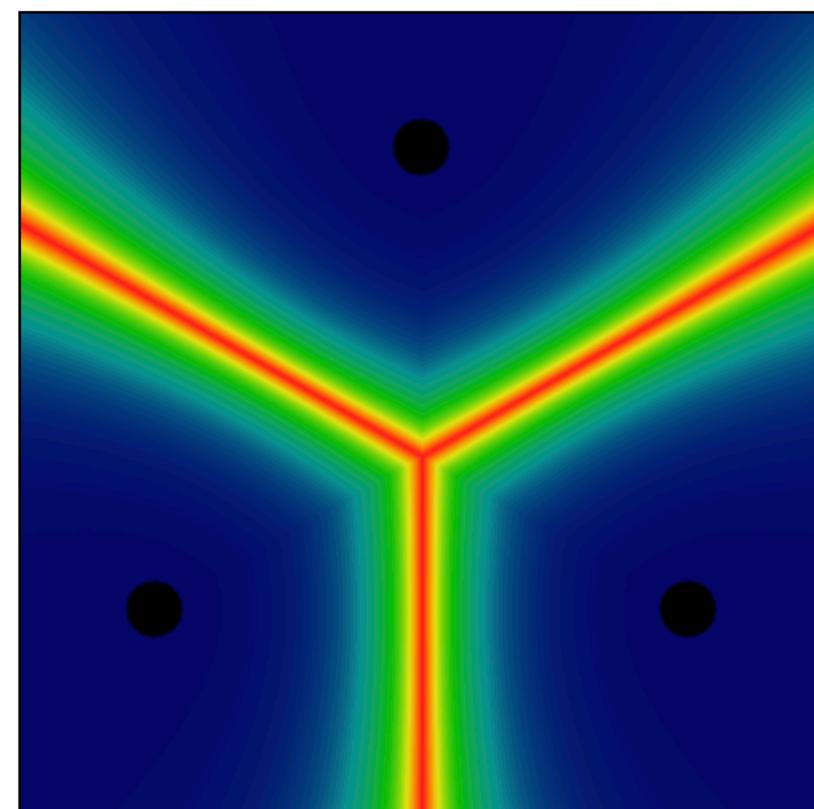
Least Confident



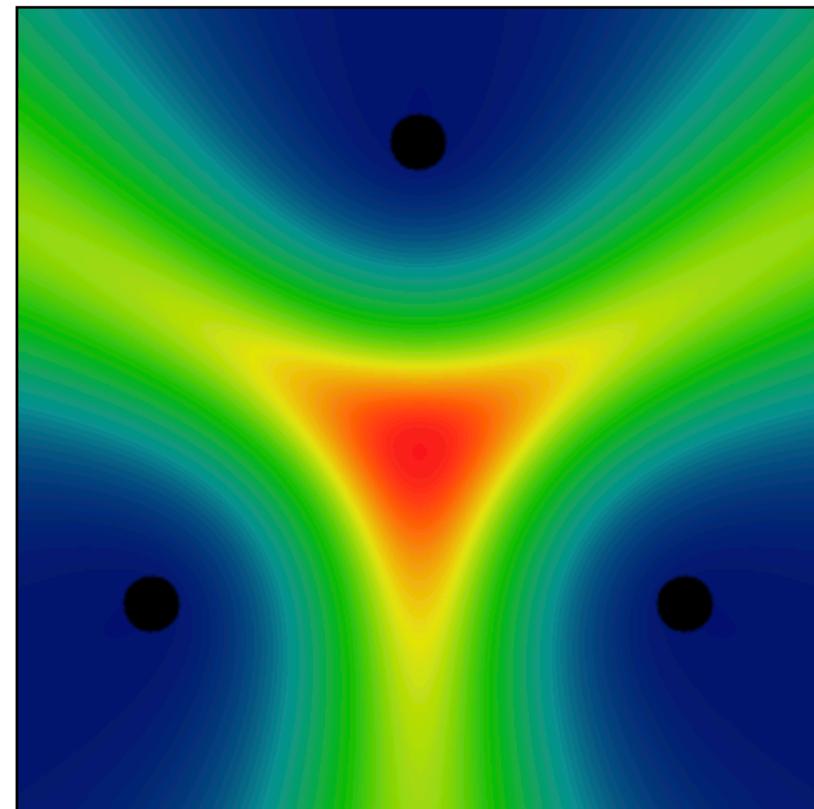
Margin of Confidence



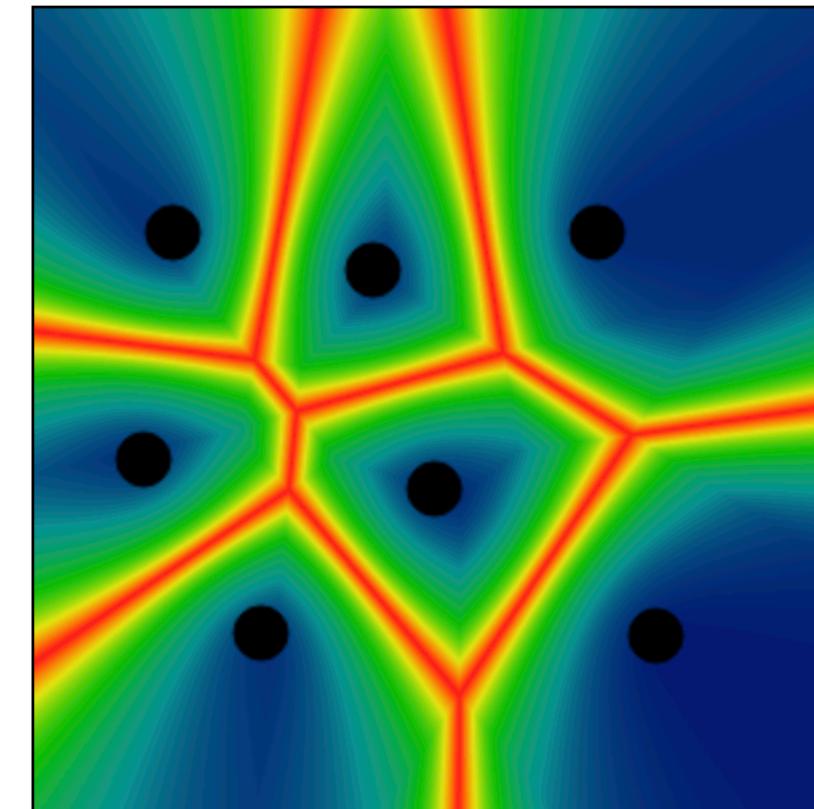
Ratio of Confidence



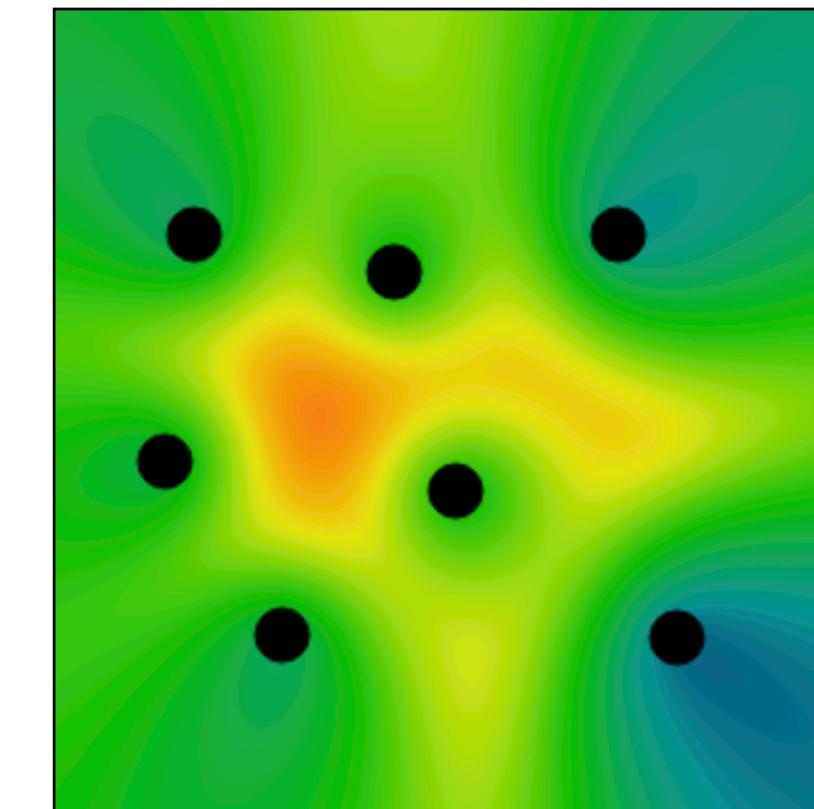
Label Entropy



Ratio of Confidence



Label Entropy



# Query by Committee

- Обучаем ансамбль из  $m$  моделей и смотрим на согласованность их предсказаний
- Чаще всего используется энтропия

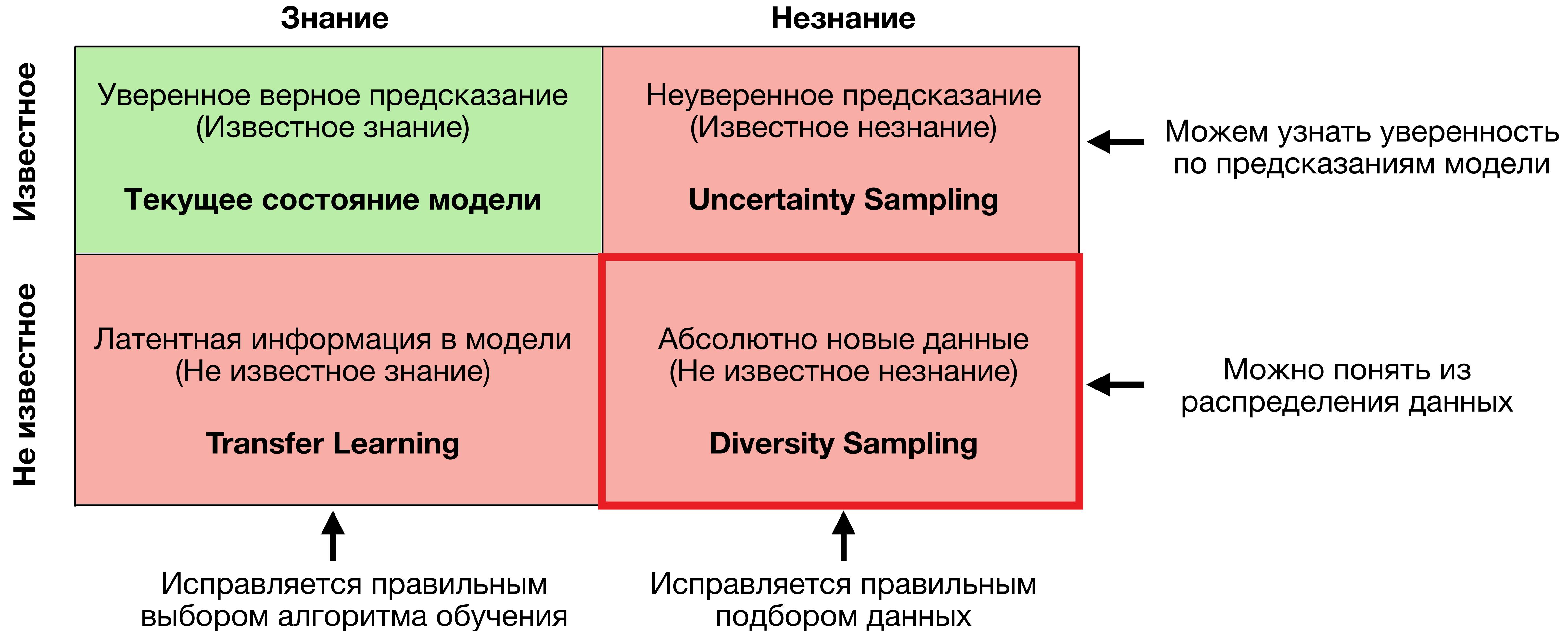
$$\text{score}(x) = - \sum_{k=1}^K \frac{V(y_k)}{m} \log \frac{V(y_k)}{m},$$

$V(y_k)$  – число моделей, предсказавших класс  $y_k$ .

# Query by Committee: недостатки

- Обучение  $t$  моделей вычислительно не эффективно
- Модели могут иметь высокую корреляцию
  - Можно обучать разные модели на разных подмножествах данных (bagging)
  - Можно обучать модели из разных семейств: нейронные сети, TF-IDF, ...
- Объекты с наибольшим скором часто оказываются выбросами
- Важно аккуратно подбирать метод сэмплирования объектов для разметки

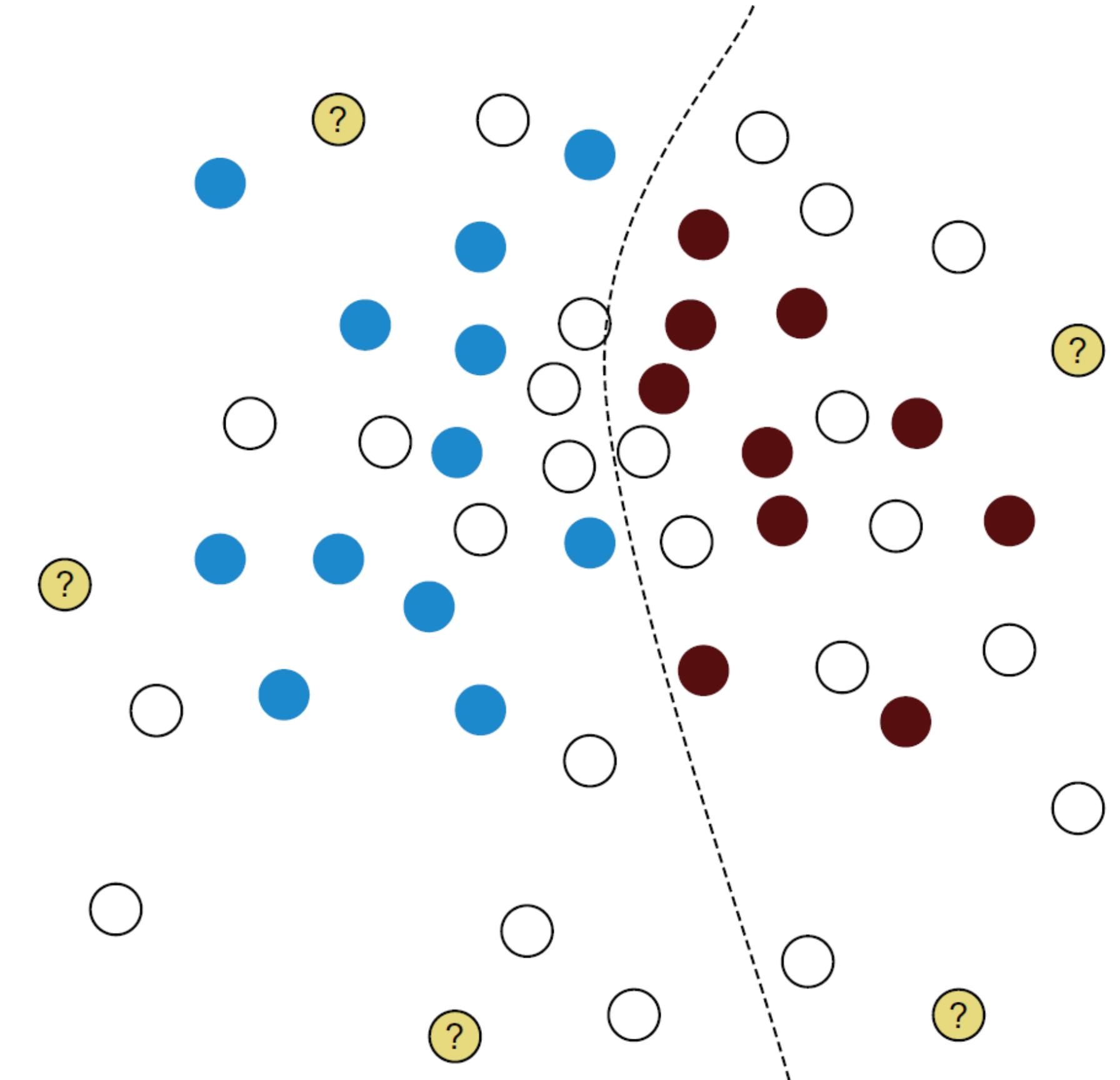
# Квадрат знаний модели



# Diversity Sampling

- Тексты в реальных данных очень часто смещены:
  - По языку
  - По источнику
  - По классу (отрицательных отзывов больше положительных)
- Причин для смещения бывает много
  - Более развитые страны создают больше данных
  - Популярные источники владеют ресурсами на создание контента

Мы хотим, чтобы модель умела работать со всеми данными



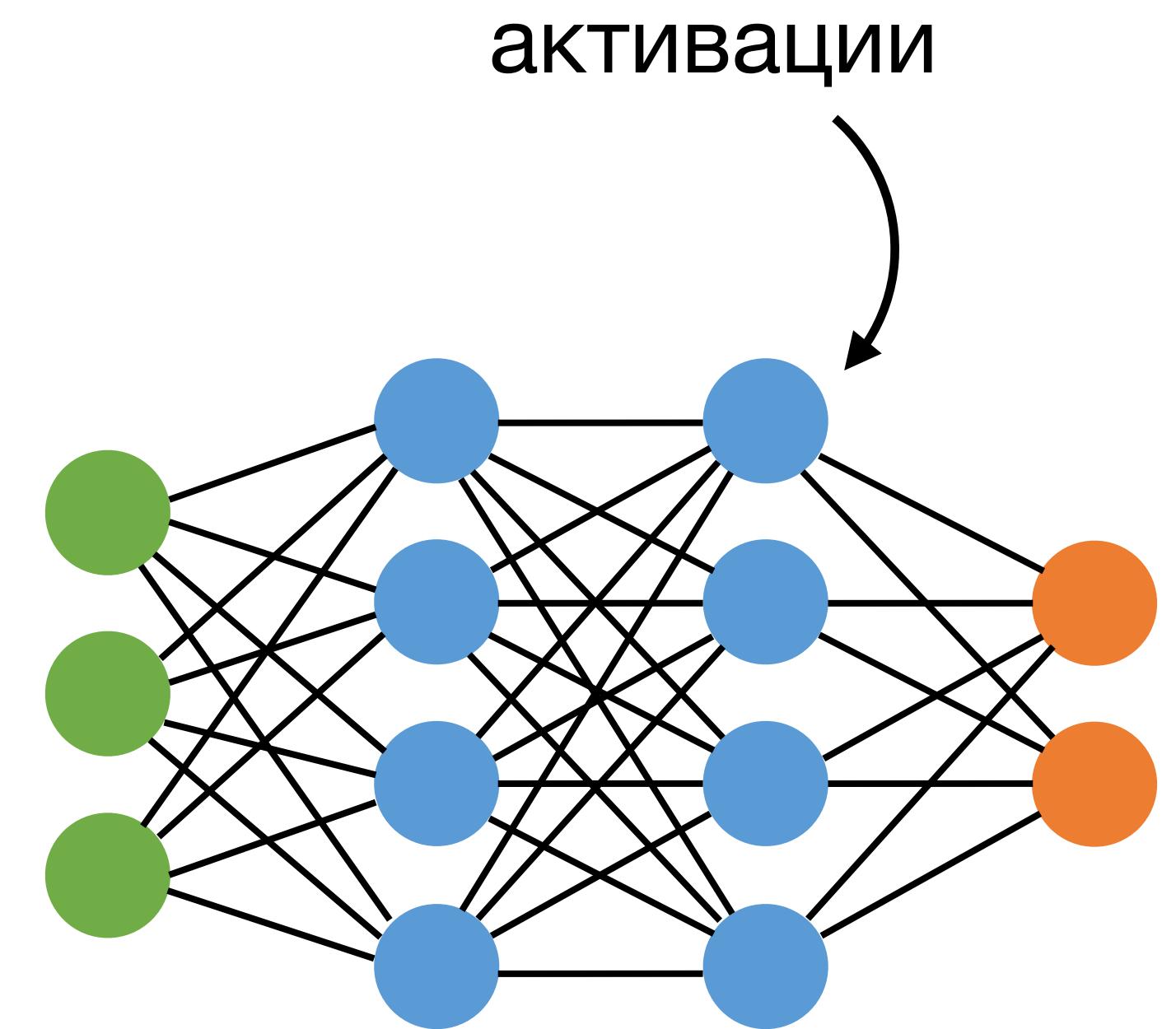
# Diversity Sampling

Наша задача – увеличить разнообразие данных.  
Для этого существует несколько способов:

- Model-based sampling
- Cluster-based sampling
- Representative sampling

# Model-based sampling

- Так как модели – это нейронные сети, можно посмотреть на активации их слоев
- Если входной текст активирует нейрон, то этот текст вероятно знаком модели
- Если ни один из нейронов не активируется, то модель ничего не знает о тексте



# Что значит активируется?

- Если значение активации равно 1.34, это много или мало?

# Что значит активируется?

- Если значение активации равно 1.34, это много или мало?
- Для оценки магнитуды активаций можно использовать валидационную выборку (неразмеченную)
- Пусть определенный нейрон для разных валидационных текстов получает значения
  - [2.52, 1.95, 1.18, 0.22, -0.12]
- Скажем, что 2.52 соответствует 100% активации, а -0.12 – 0% активации
- Тогда новый текст со значением 0.05 будет лежать ровно между 20% и 0%. Он активируется на 10%

# Агрегация нейронов

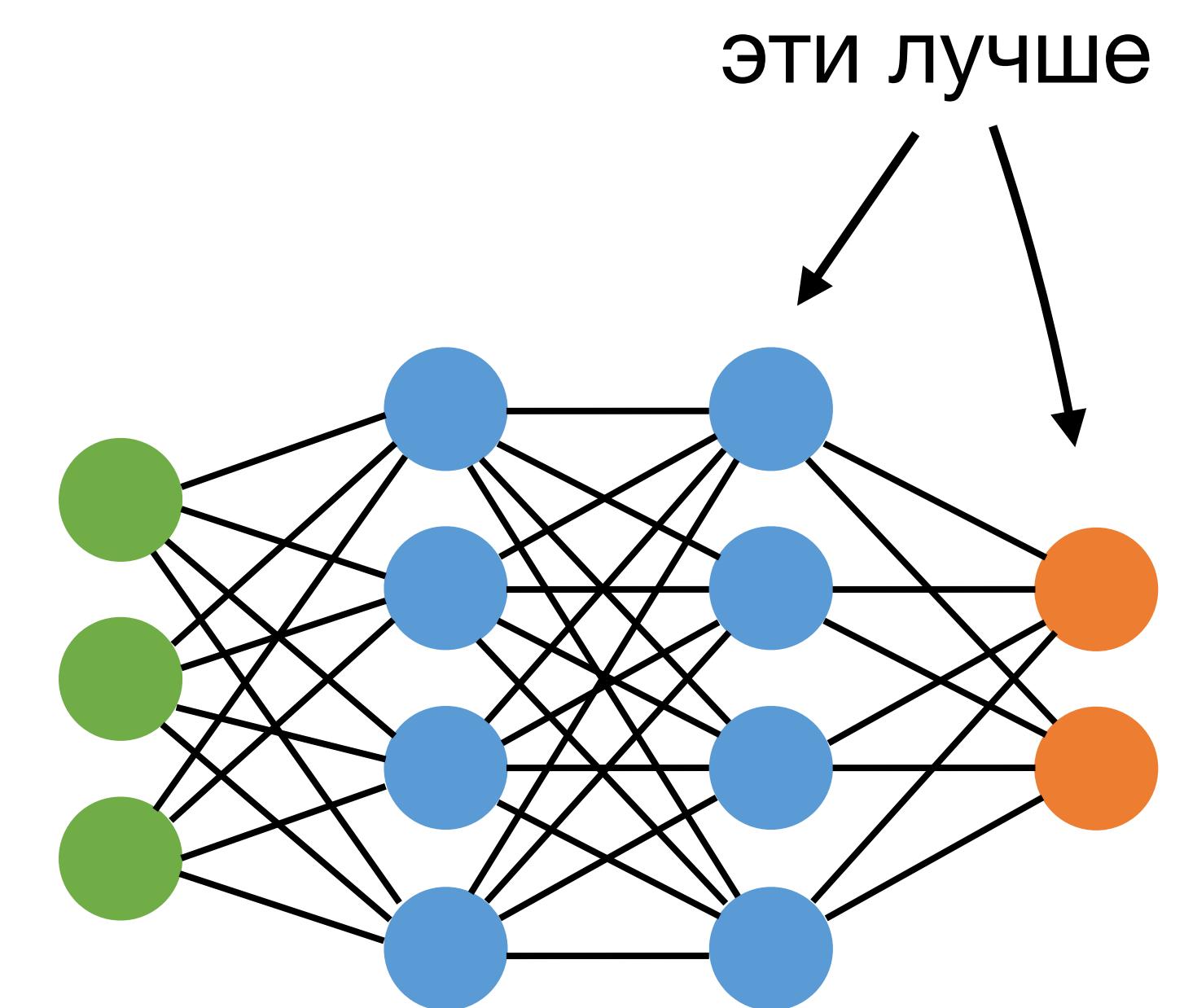
Самый надежный способ агрегации рангов – усреднение

Пусть у нас есть  $d$  выходных нейронов. Посчитаем для каждого уровень активации и усредним значения.

$$\text{score}(z) = -\frac{1}{d} \sum_{i=1}^d \text{rank}(z_i)$$

# Активации каких слоев брать?

- Чем глубже слой, тем больше семантической информации он извлекает
- Обычно более глубокие слои подходят лучше
- Можно брать несколько слоев
- Надежнее брать выходы слоев до применения нелинейности!

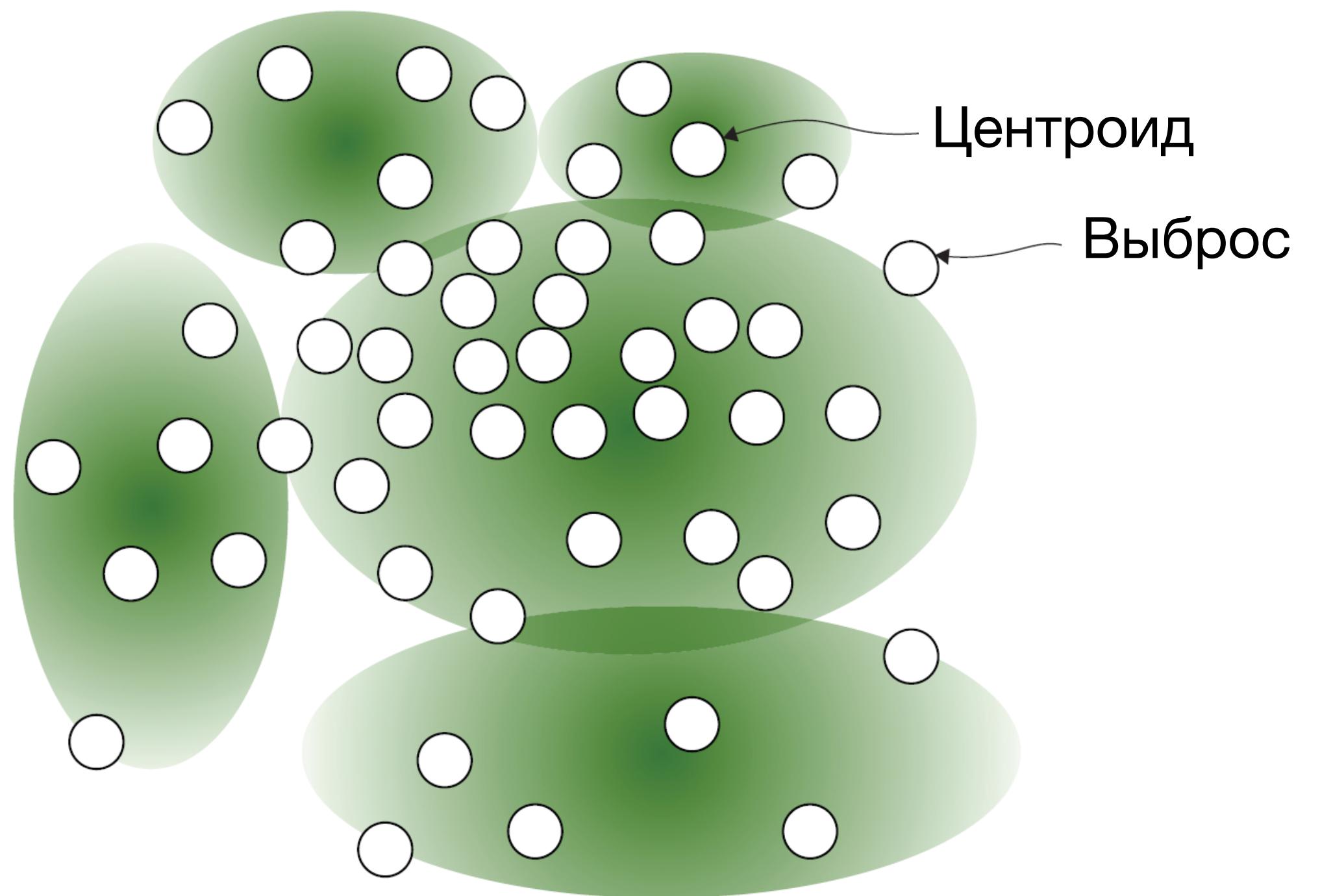


# Недостатки Model-based sampling

- Найденные тексты могут быть очень похожи
- Метод плохо работает вначале, когда данных мало

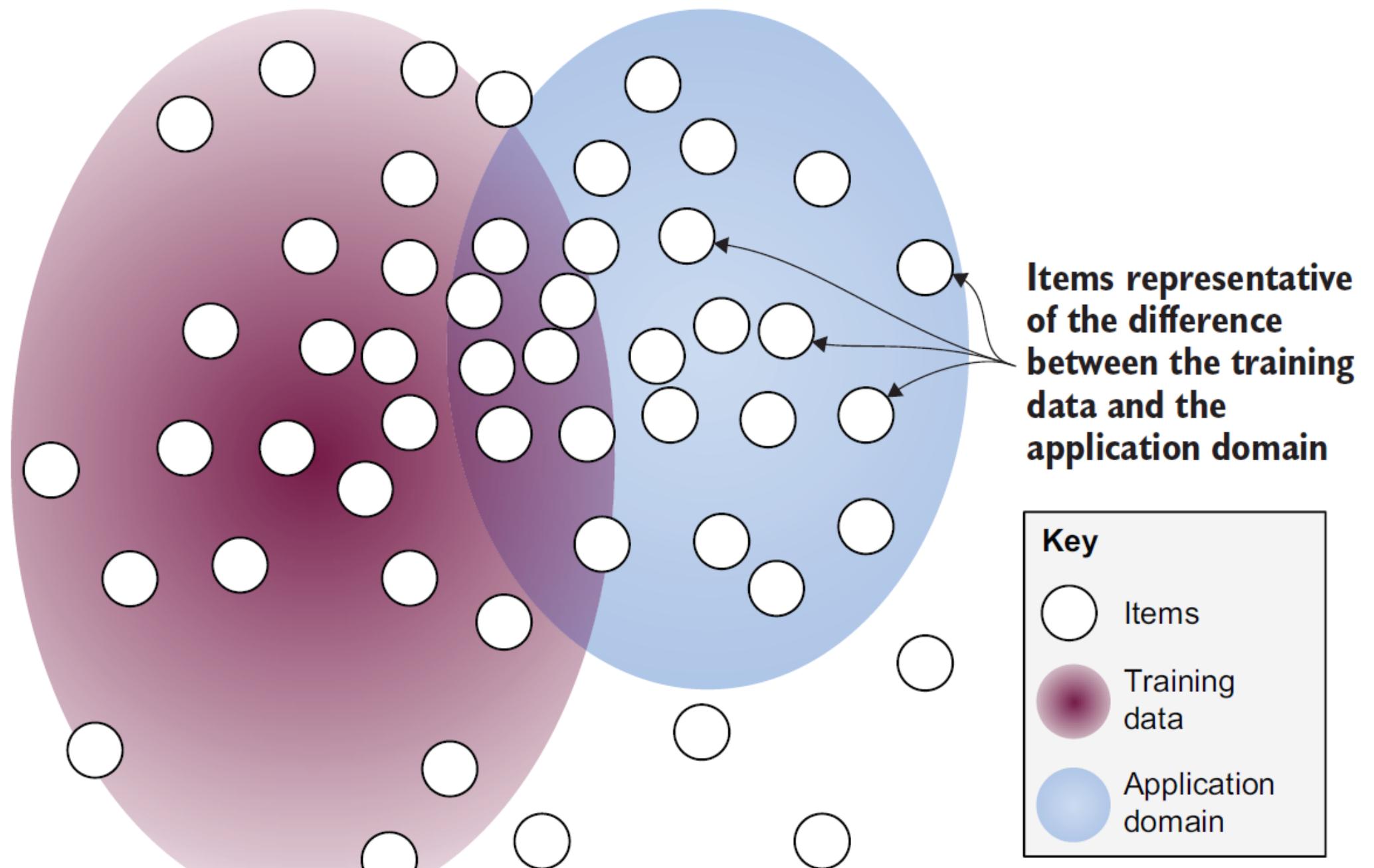
# Cluster-based sampling

- Разбиваем выборку на много кластеров
- Берем по несколько текстов из каждого кластера
- При выборе текстов можно брать **случайный, центроид или выброс**
- Обычно берут всего понемногу
- В качестве расстояния лучше брать косинусное



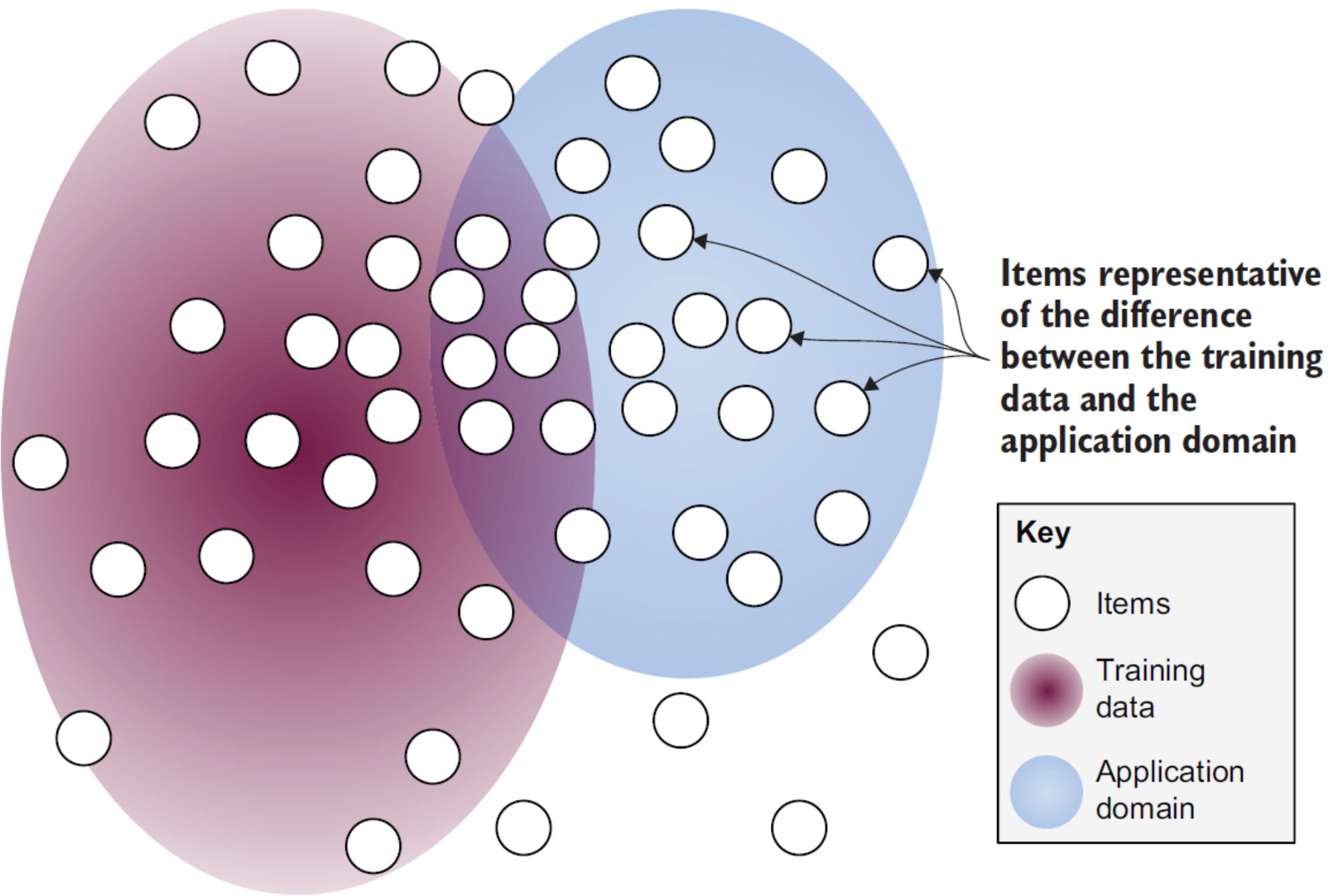
# Representative sampling

- В реальных данных распределение неразмеченных данных часто отличается от целевого домена
- Если мы знаем целевое распределение, то будем брать тексты, которые лежат ближе к нему
- Для оценки близости можно кластеризовать обе выборки и сравнивать расстояние текста до ближайших кластеров



# Недостатки Representative sampling

- Representative sampling часто приводит к переобучению и однообразности текстов
- Лучше всего использовать его в комбинации с Uncertainty Sampling

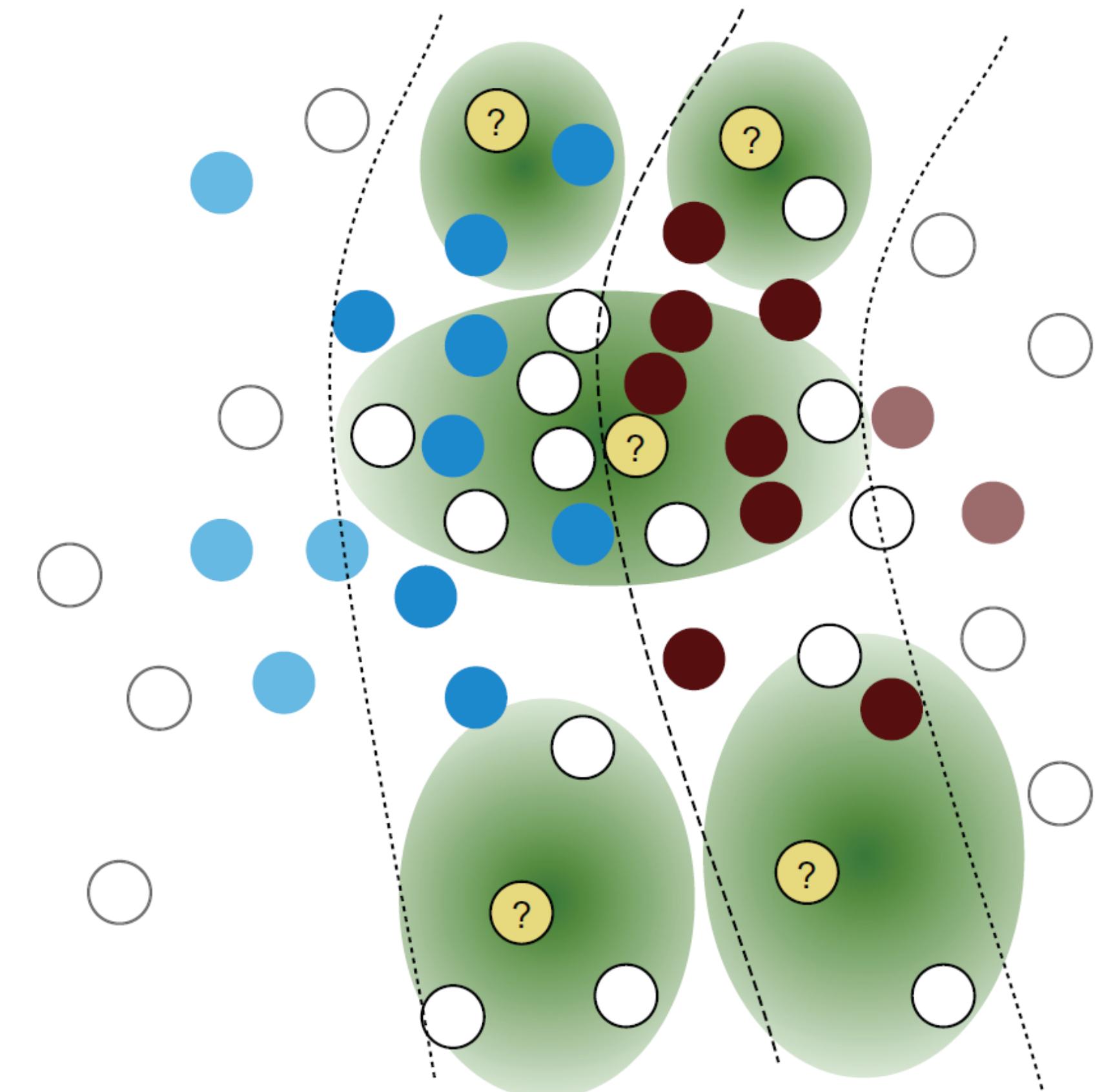


# Комбинация методов

- Uncertainty Sampling часто приводит к похожести текстов и плохой обобщаемости модели
- Diversity Sampling находит недостаточно сложных примеров
- Очень логично объединить методы для устранения недостатков
- На ранних циклах важнее Diversity Sampling, а на более поздних – Uncertainty Sampling

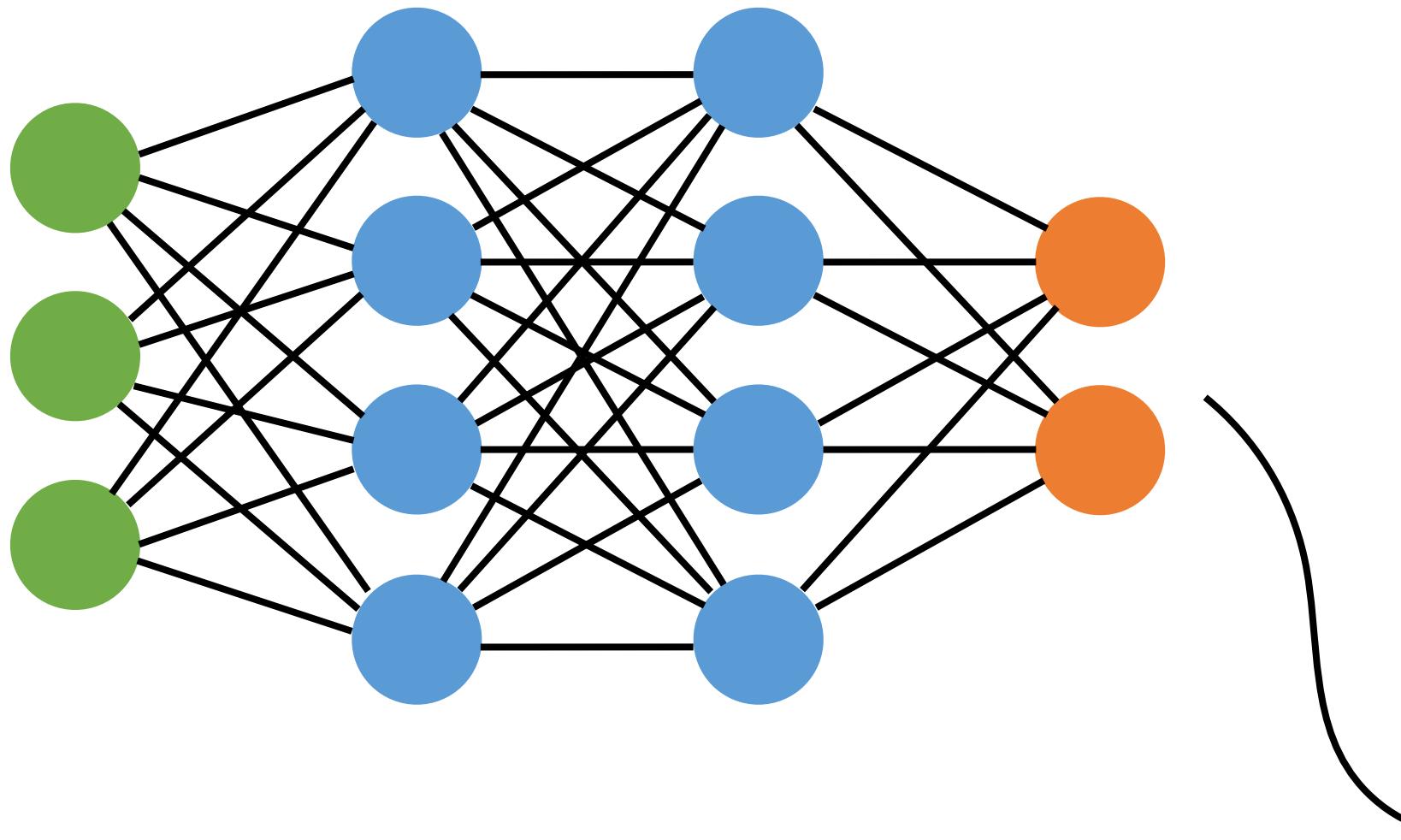
# Uncertainty и Кластеризация

- Самый простой способ объединения
- С помощью Uncertainty Sampling находим тексты рядом с разделяющей кривой
- Кластеризуем все эти тексты
- Выбираем тексты из каждого кластера



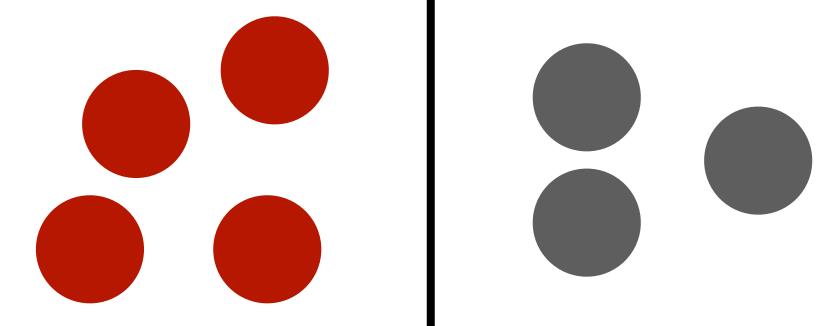
# Active transfer learning

- Предсказываем ответы для валидационной выборки и записываем правильность



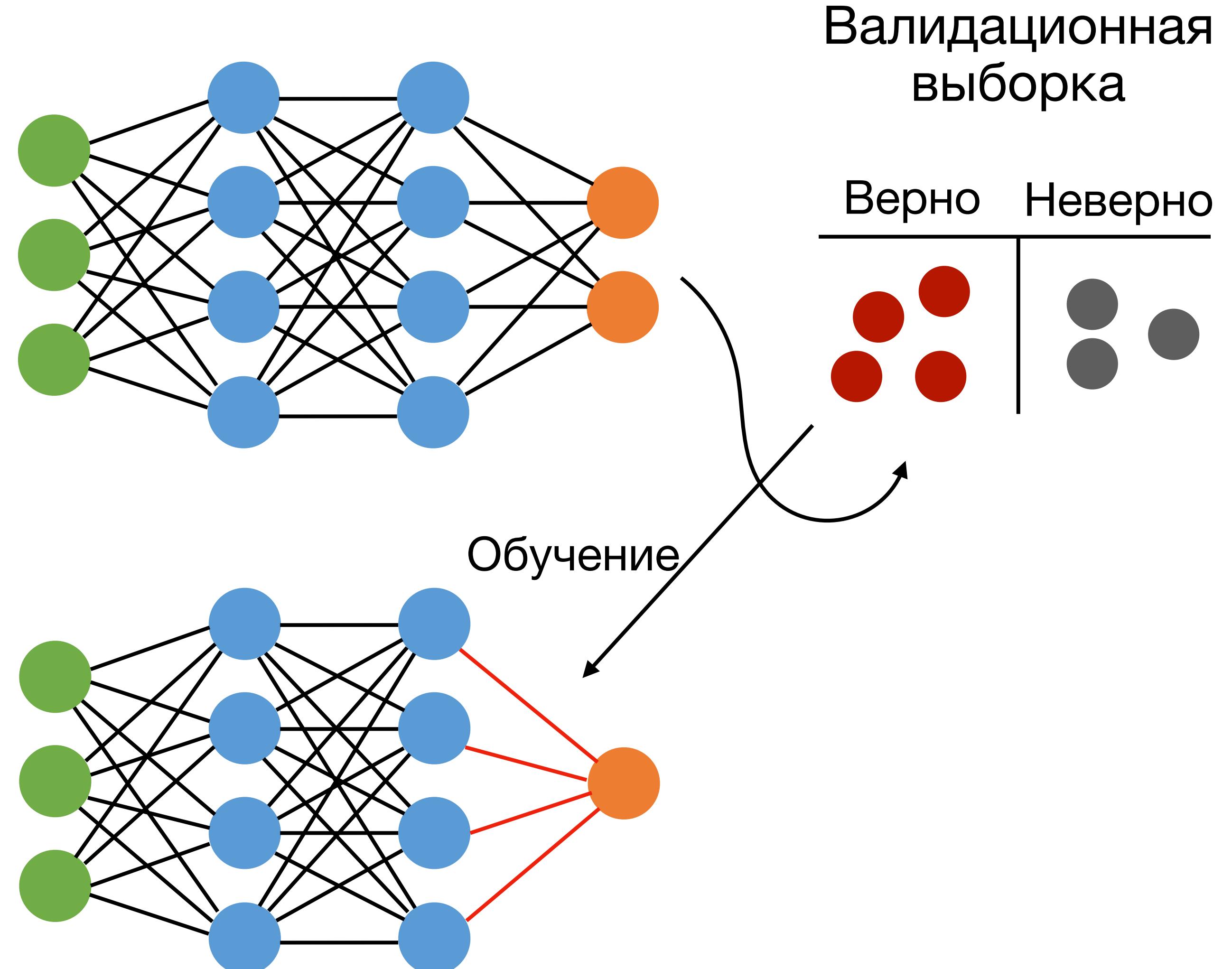
Валидационная  
выборка

Верно Неверно



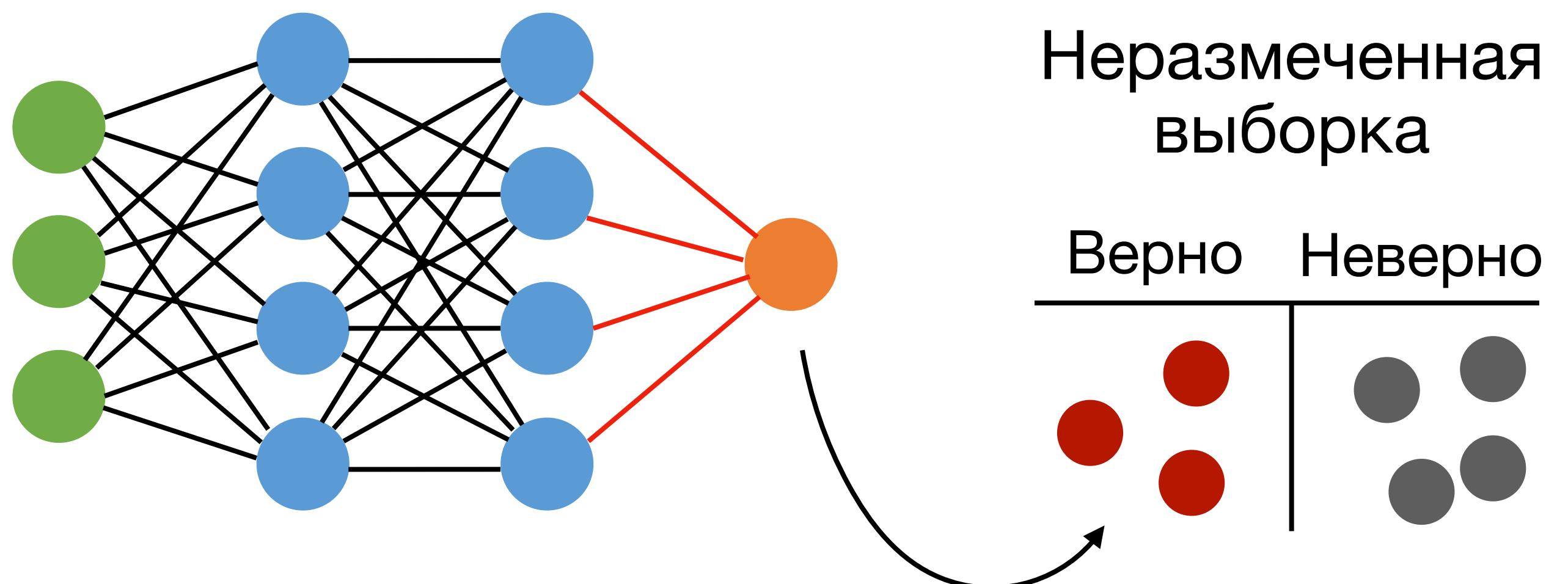
# Active transfer learning

- Предсказываем ответы для валидационной выборки и записываем правильность
- Обучаем голову предсказывать правильность по этому датасету



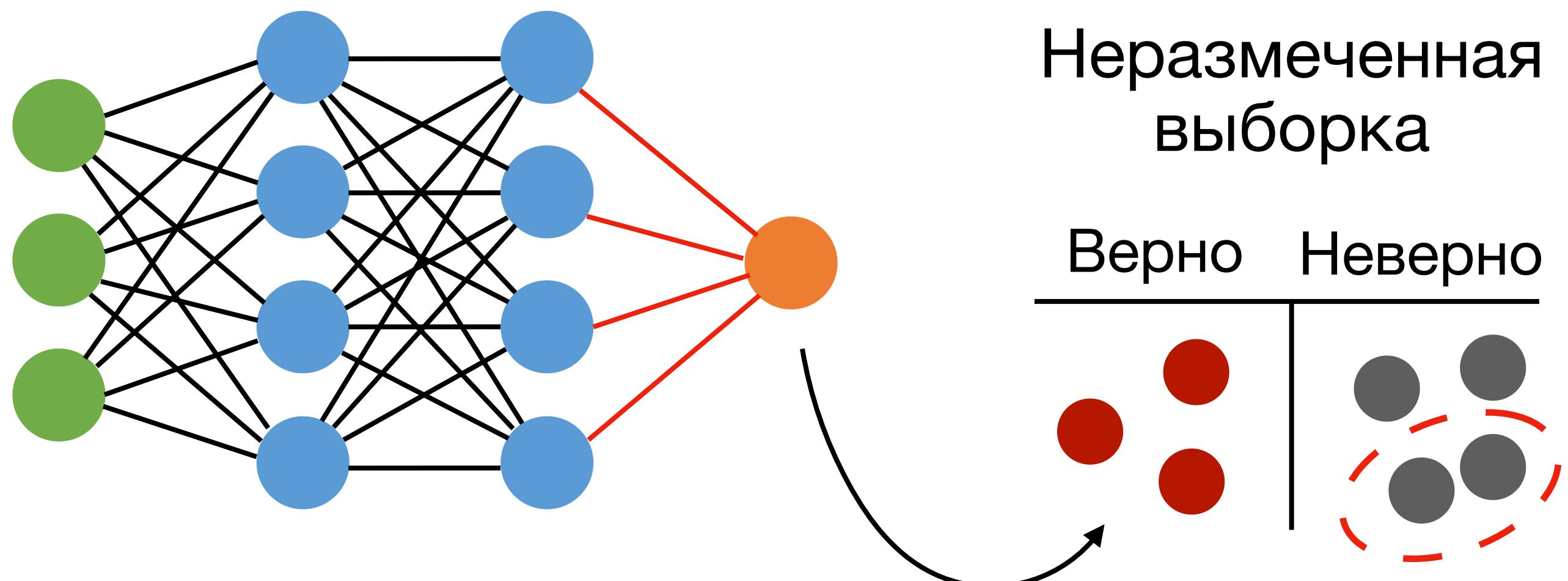
# Active transfer learning

- Предсказываем ответы для валидационной выборки и записываем правильность
- Обучаем голову предсказывать правильность по этому датасету
- Предсказываем правильность для неразмеченной выборки



# Active transfer learning

- Предсказываем ответы для валидационной выборки и записываем правильность
- Обучаем голову предсказывать правильность по этому датасету
- Предсказываем правильность для неразмеченной выборки
- Берем тексты, которые вероятнее всего будут неверно предсказаны



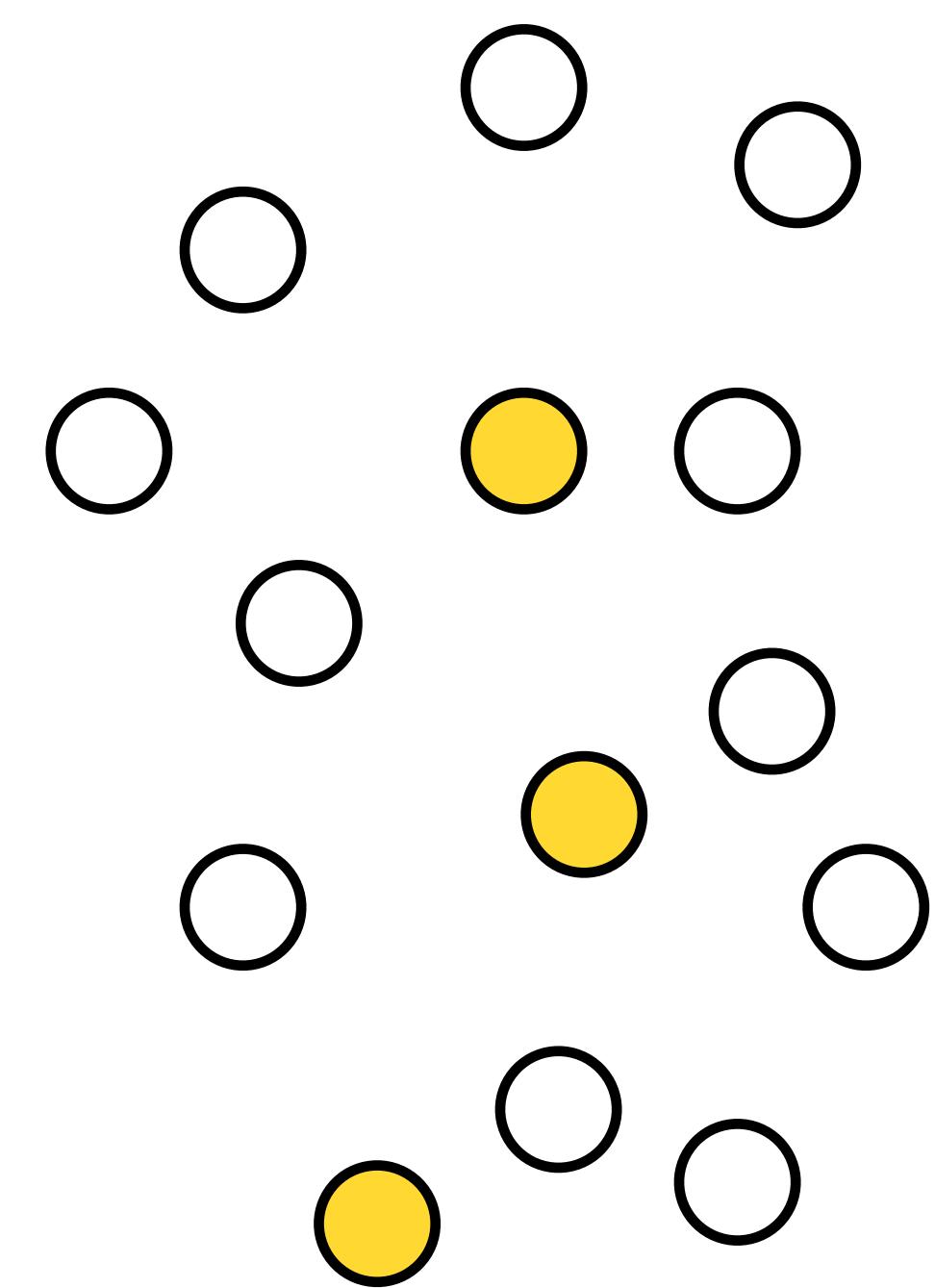
# Active transfer learning

-  Не нужно настраивать никаких параметров, модель все делает за нас
-  Не нужно много данных, так как учим только голову
-  Обучение очень быстрое
-  Тексты оказываются недостаточно разнообразными
-  Модель может переобучиться, если валидационная выборка слишком маленькая

# **Детали Active Learning**

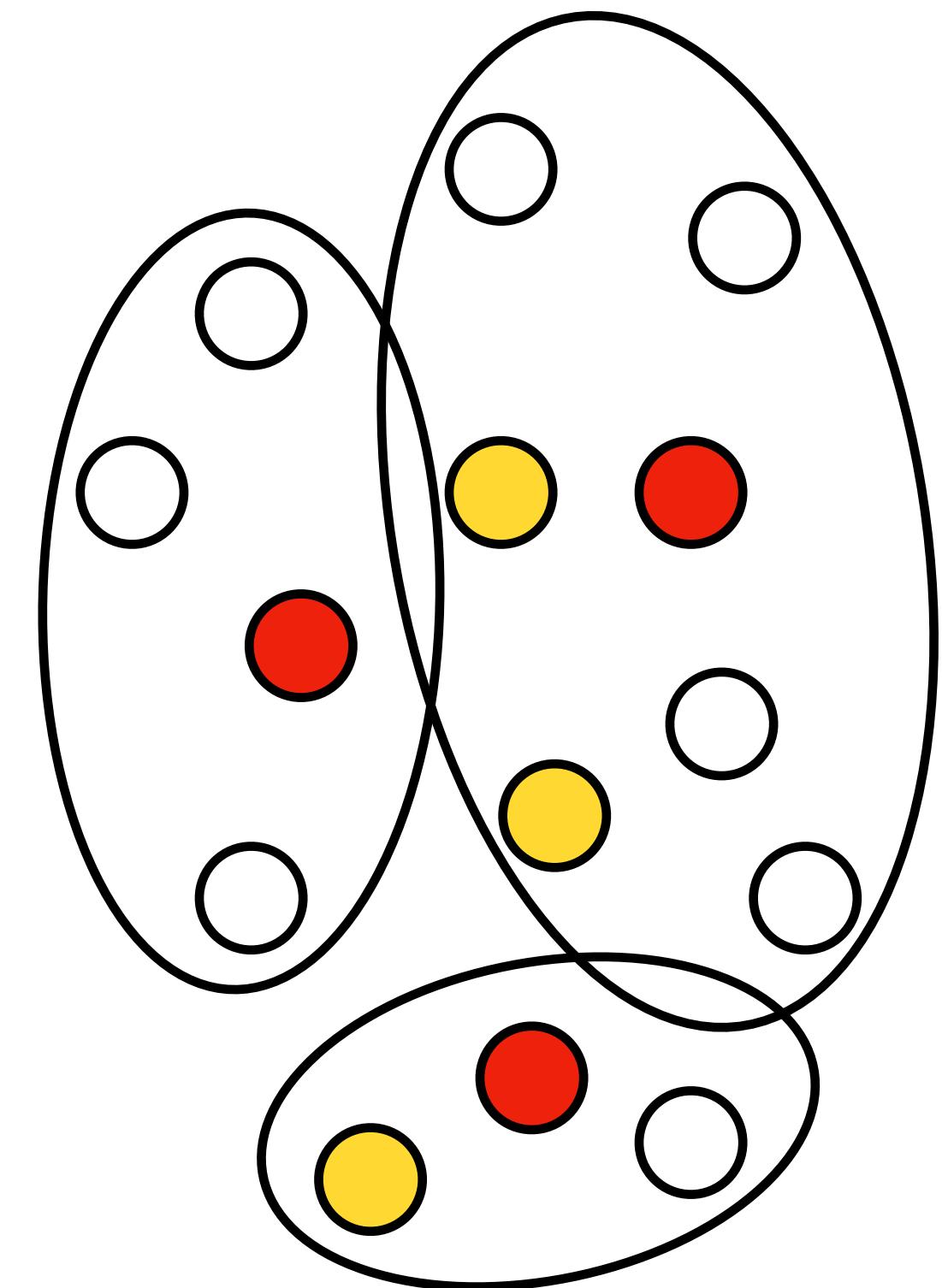
# Какие тексты размечать на старте?

- Наиболее распространенный случай – случайное семплирование
- Так мы не контролируем ничего



# Какие тексты размечать на старте?

- Наиболее распространенный случай – случайное семплирование
  - Так мы не контролируем ничего
- 
- Лучше семплировать центры кластеров
  - Получаем хорошие разнообразные примеры



# Когда останавливать active learning?

- Проще всего останавливаться при израсходовании бюджета

# Когда останавливать active learning?

- Проще всего останавливаться при израсходовании бюджета
- Можно останавливаться при достижении определенного качества
  - Для этого нужно иметь большой тестовый датасет ← может быть дорого
  - Или использовать кросс-валидацию ← размеченные данные смещены

# Когда останавливать active learning?

- Проще всего останавливаться при израсходовании бюджета
- Можно останавливаться при достижении определенного качества
  - Для этого нужно иметь большой тестовый датасет ← может быть дорого
  - Или использовать кросс-валидацию ← размеченные данные смещены
- Можно останавливаться, когда модель становится уверенной
  - ← желательно отнормировать значения уверенности

# Когда останавливать active learning?

- Проще всего останавливаться при израсходовании бюджета
- Можно останавливаться при достижении определенного качества
  - Для этого нужно иметь большой тестовый датасет ← может быть дорого
  - Или использовать кросс-валидацию ← размеченные данные смещены
- Можно останавливаться, когда модель становится уверенной
  - ← желательно отнормировать значения уверенности
- Можно замерять разницу моделей между итерациями

# Оценка стоимости разметки текста

- Можно считать стоимость каждого текста одинаковой
  - Тексты имеют разную длину и сложность

# Оценка стоимости разметки текста

- Можно считать стоимость каждого текста одинаковой
  - Тексты имеют разную длину и сложность
- Можно оценивать стоимость по числу токенов
  - Плохо работает, так как сложность не всегда зависит от длины

# Оценка стоимости разметки текста

- Можно считать стоимость каждого текста одинаковой
  - Тексты имеют разную длину и сложность
- Можно оценивать стоимость по числу токенов
  - Плохо работает, так как сложность не всегда зависит от длины
- Лучше обучить простую модель для предсказания сложности на основе разных текстовых статистик и компетенции разметчиков

# Оценка стоимости разметки текста

- Можно считать стоимость каждого текста одинаковой
  - Тексты имеют разную длину и сложность
- Можно оценивать стоимость по числу токенов
  - Плохо работает, так как сложность не всегда зависит от длины
- Лучше обучить простую модель для предсказания сложности на основе разных текстовых статистик и компетенции разметчиков
- Предсказывать стоимость разметки всех текстов может быть сложно, так как AL выбирает наиболее сложные тексты

# Как использовать стоимость разметки?

- Не размечать тексты дороже определенного порога

# Как использовать стоимость разметки?

- Не размечать тексты дороже определенного порога
- Можно оценить вклад от разметки текста
  - растет при убывании уверенности модели
- Поделить вклад на стоимость (return-on-investment)

$$ROI(x) = \frac{\text{uncertainty}(x)}{\text{price}(x)}$$

- Брать тексты с наибольшим ROI
- Тут полезно отнормировать значение неуверенности

# Ускорение цикла

- Дообучать модель на каждой итерации, а не учить с нуля
- Обучать модель поменьше, потому что важна не точность, а неуверенность
- Обучать модель, пока разметчики заняты разметкой
  - Датасет будет устаревшим, но это не так страшно