Сдвиги в данных, OOD и MLOps

Рындин Максим ИСП РАН

SentiRuEval 2016

- сентимент постов из соцсетей о банках
- между сбором train и test 3-4 месяца
- смена аннотаторов
- разгар кризиса
  =>
- train 85%+
- test 60%-

Что делать

- мониторинг и дообучение
- автоматизация
- дополнительные обратные переходы в жизненном цикле
- дополнительная инфраструктура

ML Pipeline

- постановка задачи
- исследование
- первые эксперименты
- эксперименты
- анализ
- вывод в production

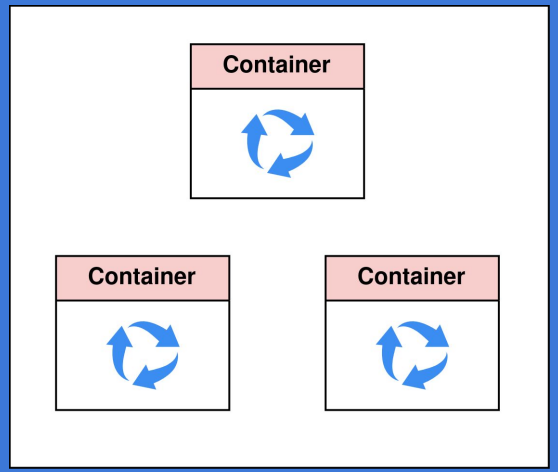
Что мы можем автоматизировать?

- **постановка задачи**
- **исследование**
- первые эксперименты

+ **эксперименты**
+ **анализ**
+ **вывод в production**

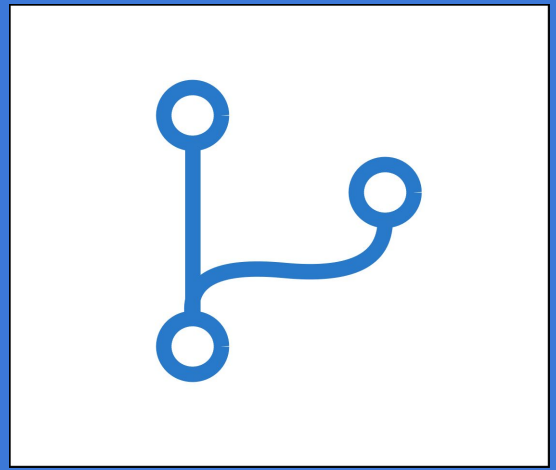
Эксперименты

- **воспроизводимость и изолированность экспериментов**
- версионирование активов
- управление ресурсами
- унификация процессов разработки в команде
- обеспечение доверия



Эксперименты

- воспроизводимость и изолированность экспериментов
- **версионирование активов**
- управление ресурсами
- унификация процессов разработки в команде
- обеспечение доверия



Эксперименты

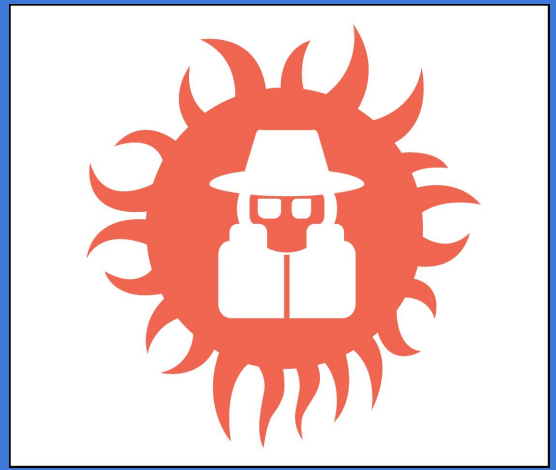
- воспроизводимость и изолированность экспериментов
- версионирование активов
- **управление ресурсами**
- унификация процессов разработки в команде
- обеспечение доверия

Эксперименты

- воспроизводимость и изолированность экспериментов
- версионирование активов
- управление ресурсами
- **унификация процессов разработки в команде**
- обеспечение доверия

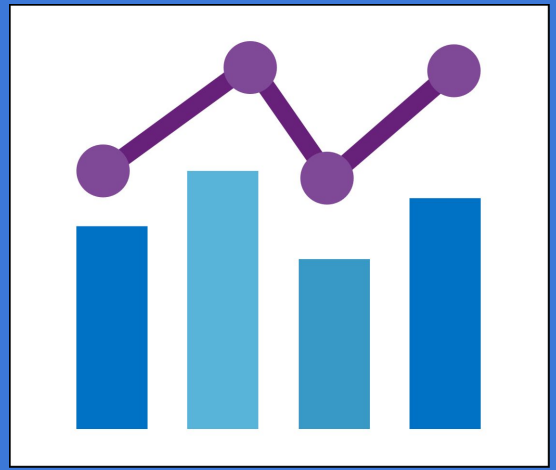
Эксперименты

- воспроизводимость и изолированность экспериментов
- версионирование активов
- управление ресурсами
- унификация процессов разработки в команде
- **обеспечение доверия***



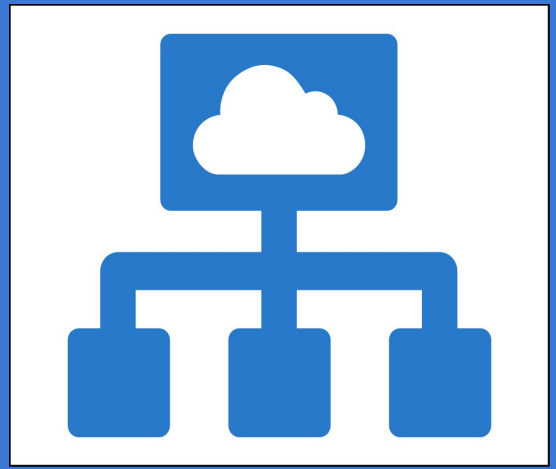
Анализ

- аналитика результатов



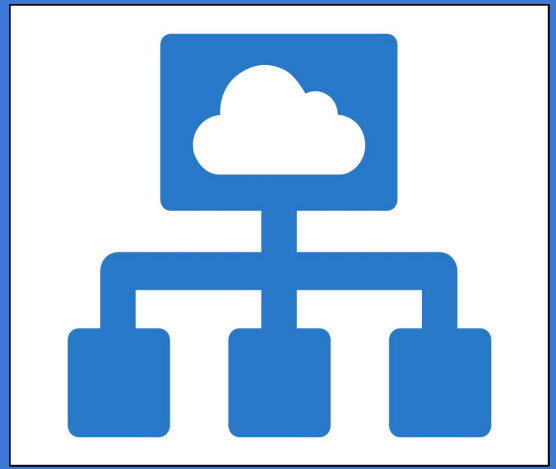
Вывод в production

- **деплой, масштабирование**
- обеспечение доверия
- мониторинг
- мониторинг устаревания



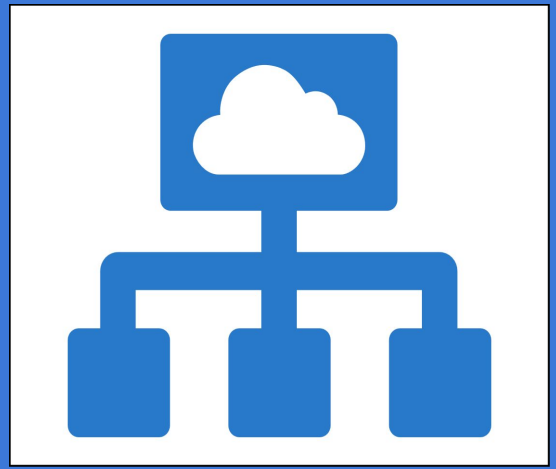
Вывод в production

- деплой, масштабирование
- **обеспечение доверия**
- мониторинг
- мониторинг устаревания



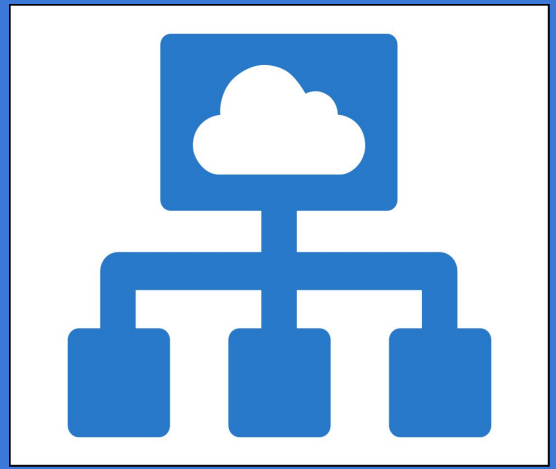
Вывод в production

- деплой, масштабирование
- обеспечение доверия
- **мониторинг**
- мониторинг устаревания



Вывод в production

- деплой, масштабирование
- обеспечение доверия
- мониторинг
- **мониторинг устаревания**



Проблемы при ML-разработке

Эксперименты и анализ:

- воспроизводимость и изолированность экспериментов
- версионирование активов
- аналитика результатов
- управление ресурсами
- унификация процессов разработки в команде
- обеспечение доверия

Production:

- деплой, масштабирование
- обеспечение доверия
- мониторинг
- мониторинг устаревания

MLOps

- практики и инструменты для автоматизации и упрощения процессов жизненного цикла моделей машинного обучения (ML)

• • •

Связь проблемы и инженерии

- умеем автоматизировать обучение и др. => можем автоматизировать мониторинг, дообучение и др.
- даже если есть понимание как поддерживать модели, нужна инфраструктура
- заботимся об устаревании => повышаем уровень разработки

Подзадачи непрерывного обучения

- когда предпринимать действия: ~OOD/AD/etc.
- (опционально) как получить разметку: ~active learning
- как дообучиться, если есть разметка: ~online/multitask learning
- как дообучиться, если нет разметки: ~domain adaptation

Отличие: цель – хорошо работать на всех классах

Зоопарк задач

- AD – Anomaly Detection
- ND – Novelty Detection
- OSR – Open Set Recognition
- OOD – Out Of Distribution (detection)
- OD – Outlier Dection

AD

- Inference time + внешний
- Нет ограничений на ID performance
- feature+concept drift
- supervised
- Собственный confidence “нормальности” примера
- TPR, FPR, AUROC, AUPR at confidence level

ND

Почти как AD, но

- только concept drift
- unsurepvised

OSR

- Только для классификации
- внутри модели => max joint ID+OOD performance
- unsurepvised
- обычно eval на разных *частях одного* набора данных

OOD

OSR без ограничений

- **не только классификация**
- **может быть supervised**
- **обычно eval на разных наборах данных**
- **FPR@TPRx**

OD

- не про “отказ от классификации”, а про подготовку/очистку данных
- unsupervised не по тому, как работает, а по духу

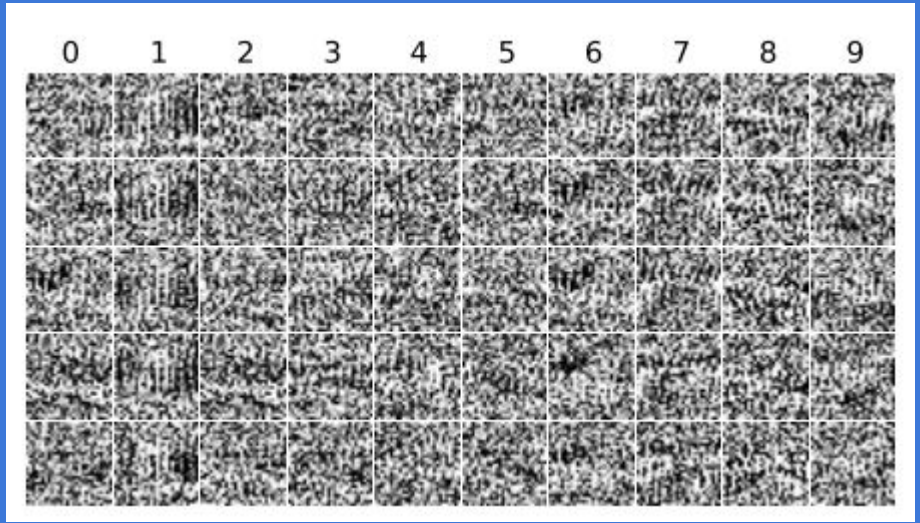
OOD: baselines

- Inference time
  - output based
  - distance based

- Train time
  - явные потери

OOD: почему сложно

- Нейронные сети склонны к overconfidence
  - Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images, Nguyen et al

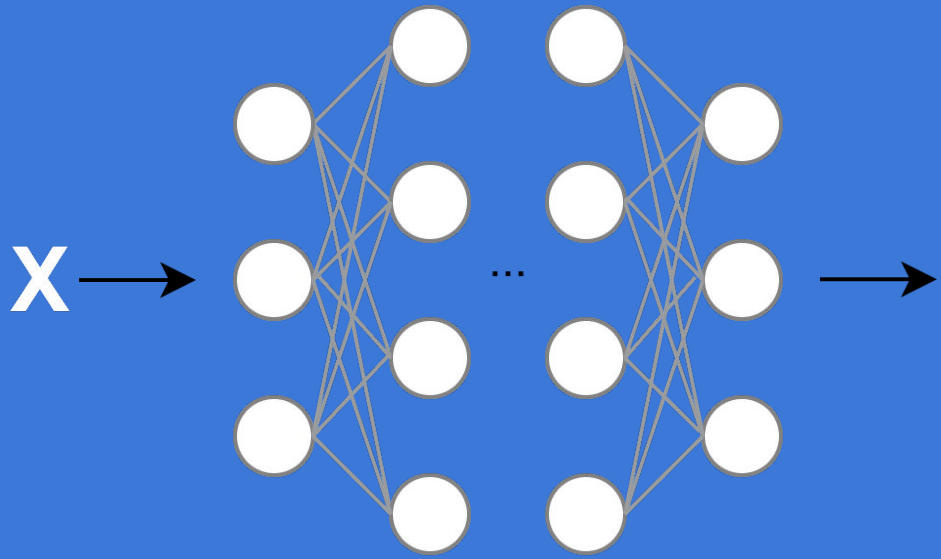


OOD: почему сложно

- Нейронные сети склонны к overconfidence
- OOD примеры сложно придумать
- OOD примеры сложно собирать
- Часто задача не на уровне примеров, а на уровне объектов

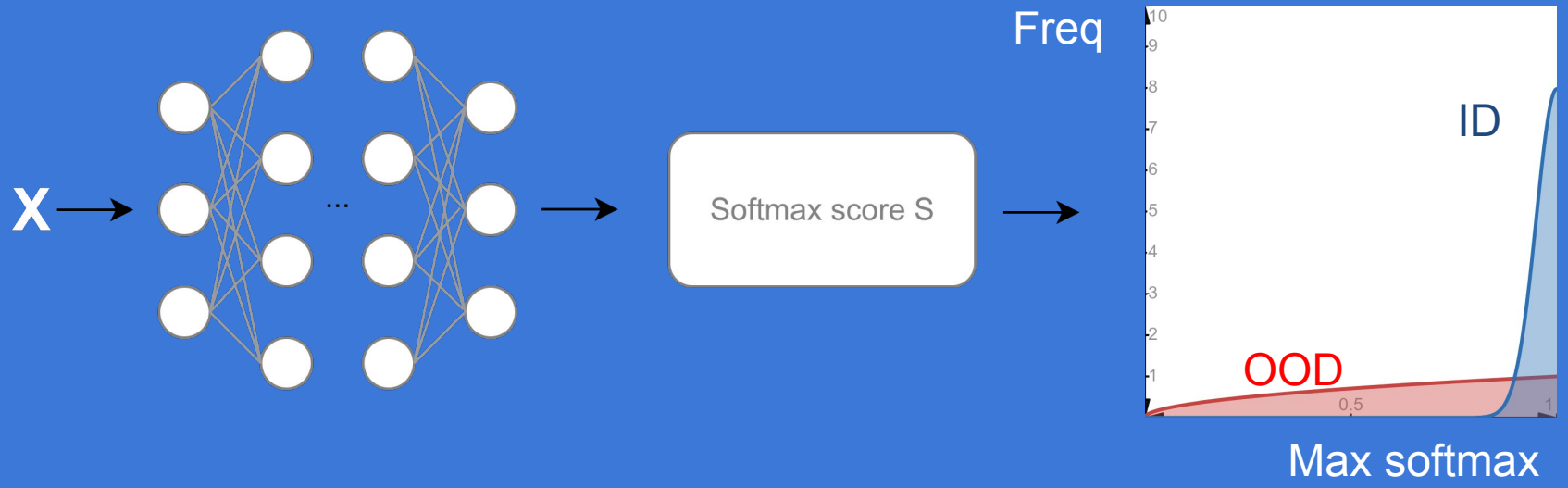
Inference time OOD

- **Конструирование scoring function S**

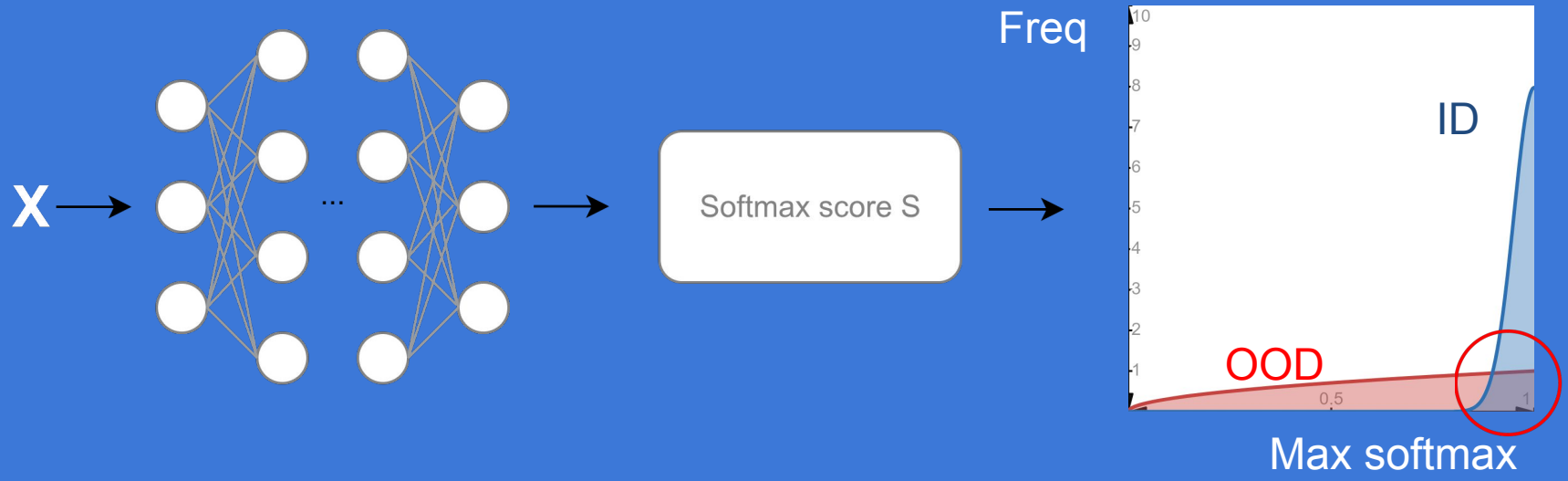


$$G_\lambda = \begin{cases} in & S(x, f) \geq \lambda \\ out & S(x, f) < \lambda \end{cases}$$

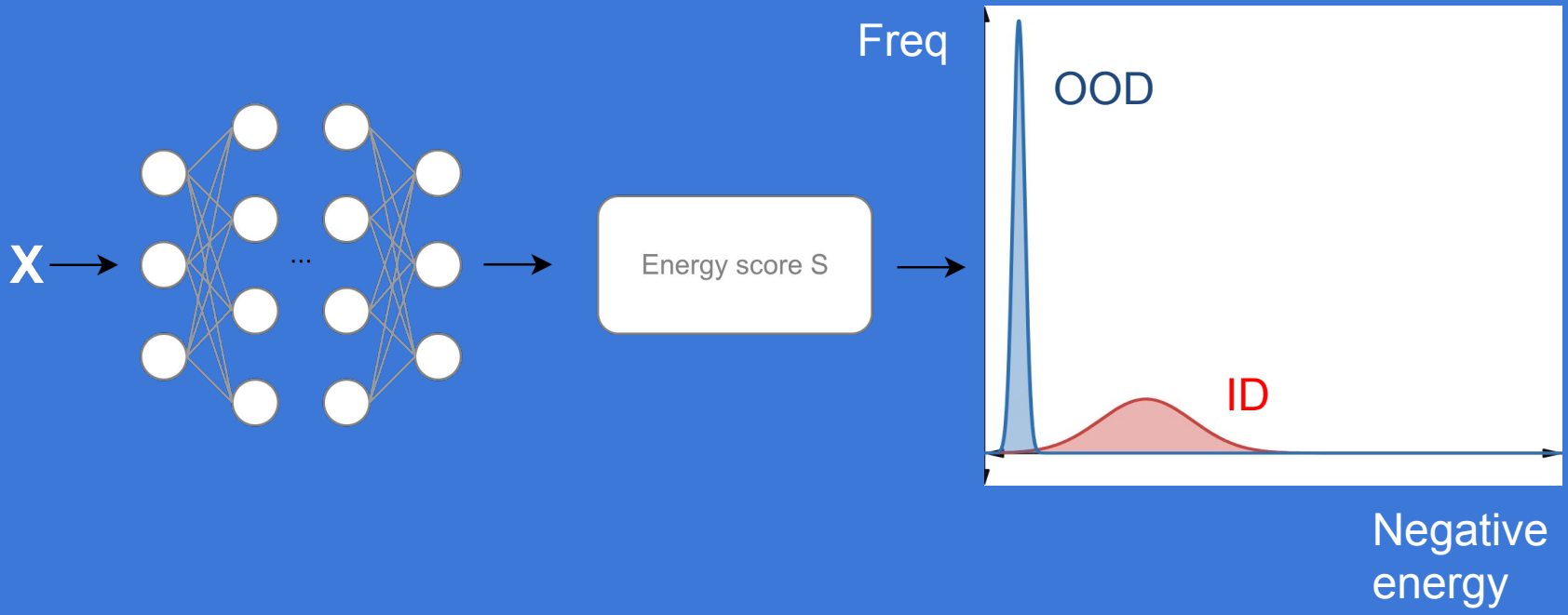
Softmax baseline



Softmax baseline



Energy baseline



Energy

- **Функция от x и y**
  - **тем меньше чем более “совместимы” вход и выход**
  - **какая-то скалярная, может быть невероятностая**
  - **распределение Гиббса – вероятность, что x ассоциирован с y**

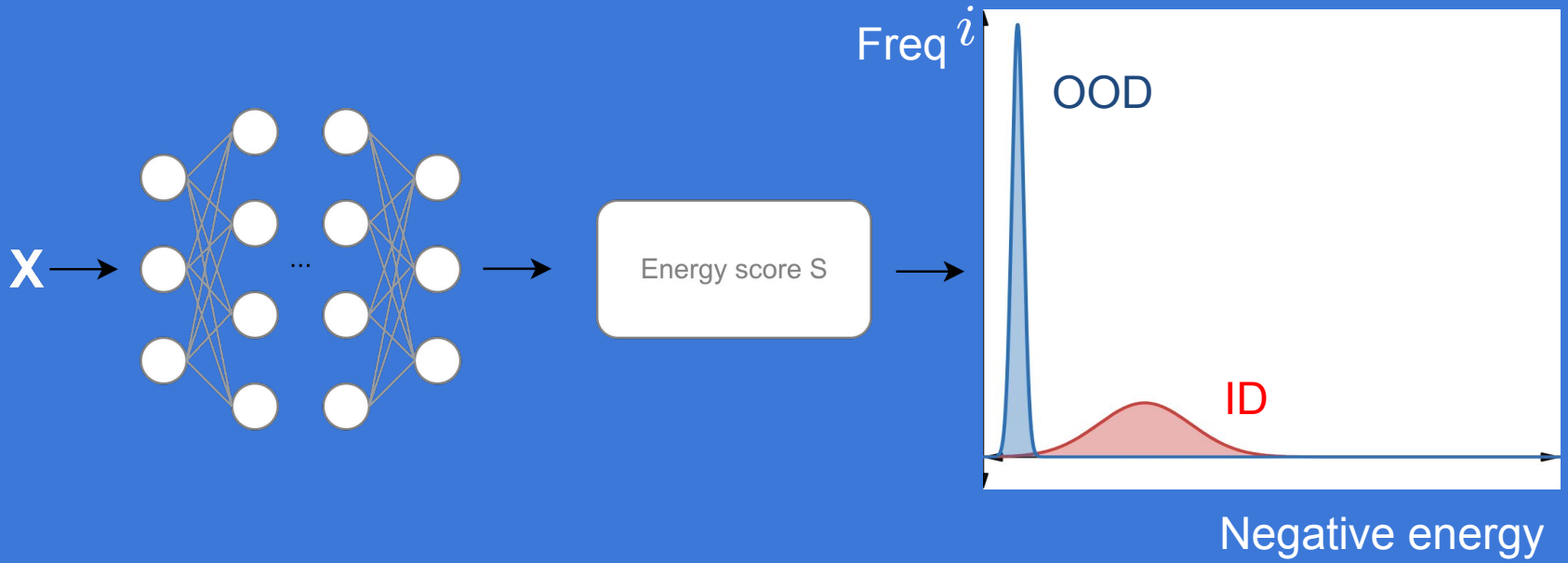
$$p(y|x) = \frac{e^{-E(x,y)/T}}{\int_{y'} e^{-E(x,y')/T}} = \frac{e^{-E(x,y)/T}}{e^{-E(x)/T}}$$

Energy

$$p(y|x) = \frac{e^{f_y(x)/T}}{\sum_i e^{f_i(x)/T}} \qquad p(y|x) = \frac{e^{-E(x,y)/T}}{\int_{y'} e^{-E(x,y')/T}}$$

Energy

$$E(x, f) = -Tlog \sum e^{f_i(x)/T}$$

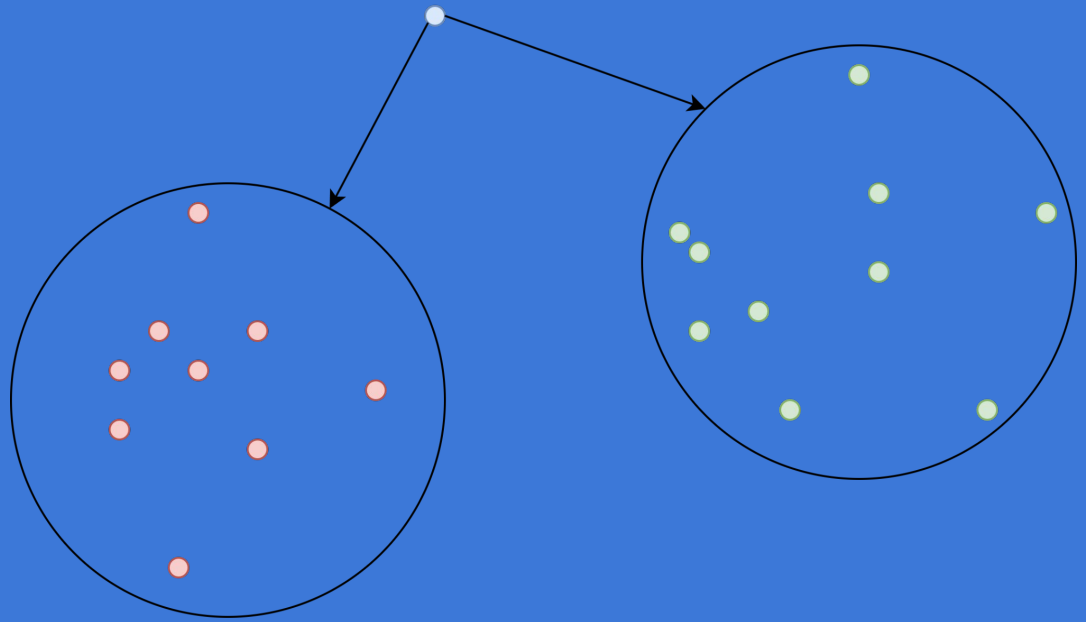


Distance baseline

- **Работаем в пространстве признаков**
  - **обычно последний слой до софтмакса**

- **Принимаем решение по расстоянию до ближайшей центроиды (ID близко, OOD равномерно далеко)**

Distance baseline



Mahalanobis distance

$$M(f, x) = \max_i -(x - \mu_i)^T \Sigma^{-1} (x - \mu_i)$$

- **Параметрическое**
- **Предполагает гауссовость фичей, если не похожи – сабоптимально**
- **Фичи из сети => оптимальны для целевой задачи, но неоптимальны для OOD**

Снимаем ограничения

- kNN для снятия предположений о распределении
- учим фичи получше (contrastive learning/etc.)
  - Cider – все фичи на сфере
  - максимизируем угол между классами (обеспечиваем кластеризуемость)
  - малая дисперсия в кластере

OOD: baselines

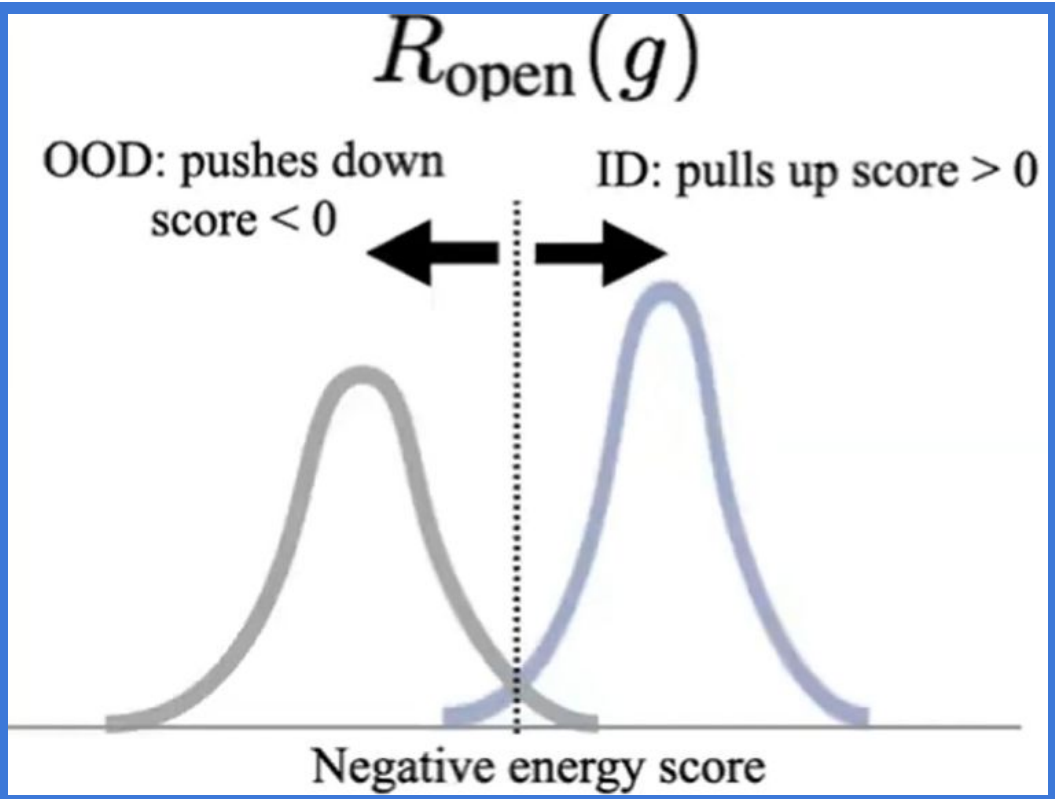
- Inference time
  - output based
  - distance based

- Train time
  - явные потери

OOD loss

$$\text{argmin} \left[\underbrace{R_{\text{closed}}(f)}_{\text{Classification error on ID}} + \alpha \cdot \underbrace{R_{\text{open}}(g)}_{\text{Error of OOD detector}} \right]$$

OOD loss

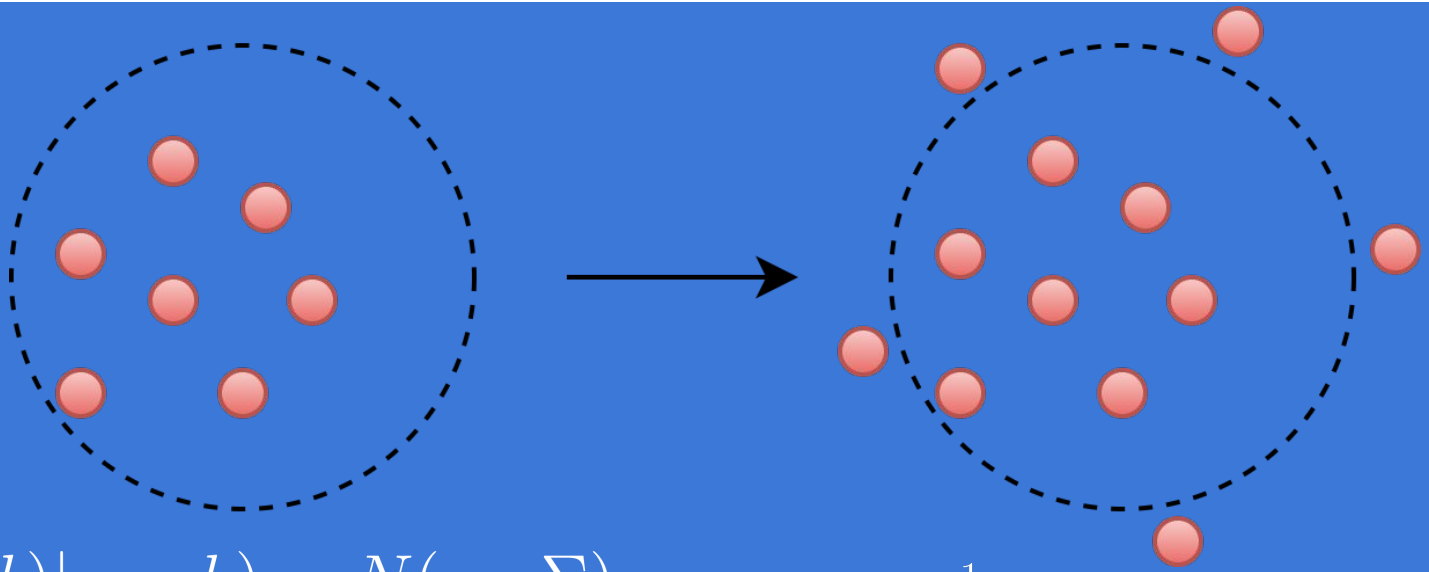

$R_{\text{open}}(g)$
OOD: pushes down
score < 0
ID: pulls up score > 0
Negative energy score

OOD: baselines

- Inference time
  - output based
  - distance based

- Train time
  - явные потери
  - синтез OOD данных

OOD: синтетика



$$p_\theta(h(x,b)|y=k) = N(\mu_k, \Sigma)$$

$$\nu_k = \{v_k | \frac{1}{(2\pi)^{m/2}|\hat{\Sigma}|^{1/2}} e^{-\frac{1}{2}(v_k-\hat{\mu_k})^T\hat{\Sigma}^{-1}(v_k-\hat{\mu_k})}\}$$

OOD: сертифицируемые потери

- будем бороться с overconfidence
  - кросэнтропия с равномерным распределением
- для надежности хотим равномерность не в точке, а в окрестности
- плохо сходится, но доказуемо