

Соревнования по анализу данных

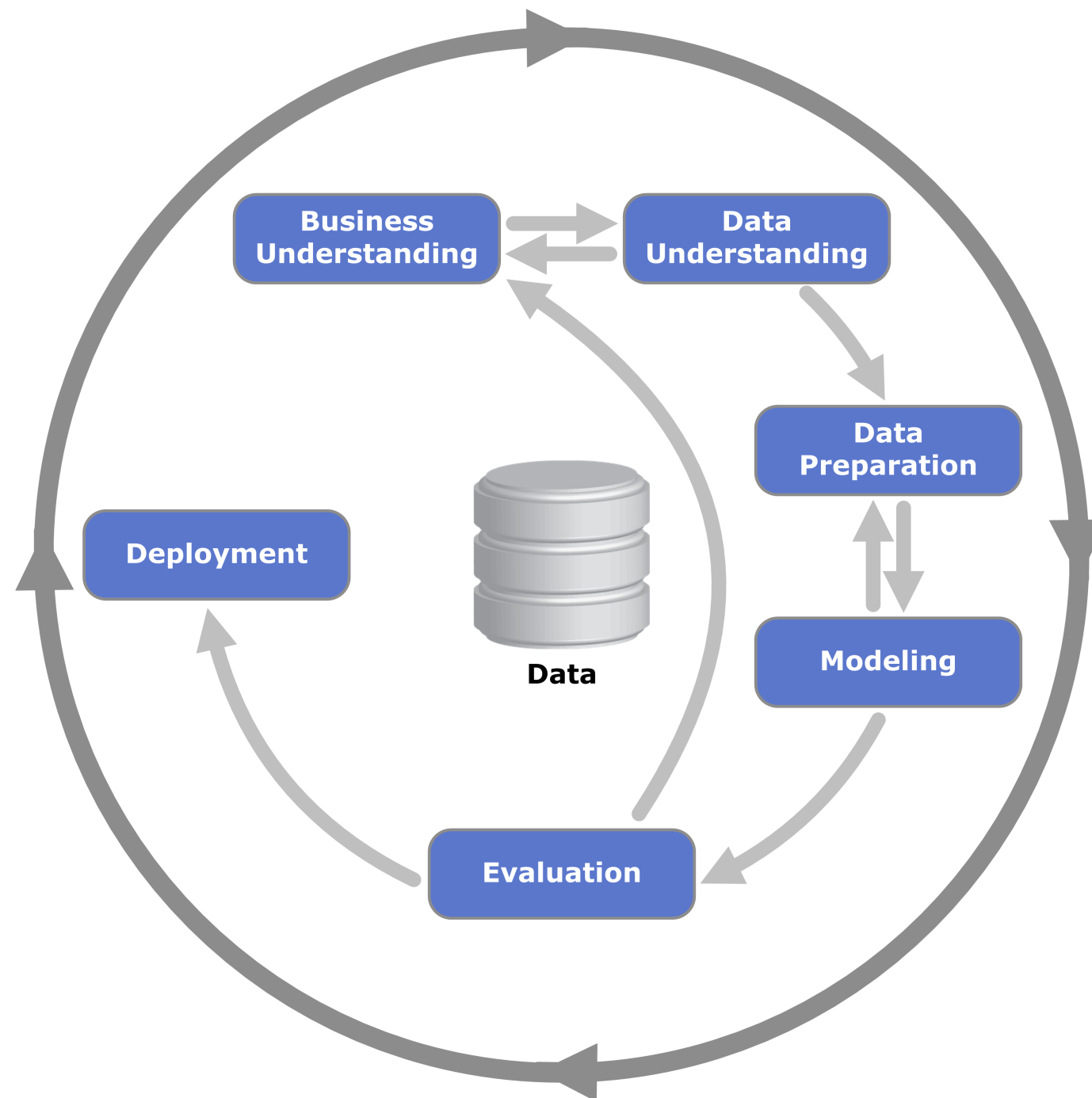
ДПО ВШЭ

Современный анализ данных, глубокое обучение и приложения

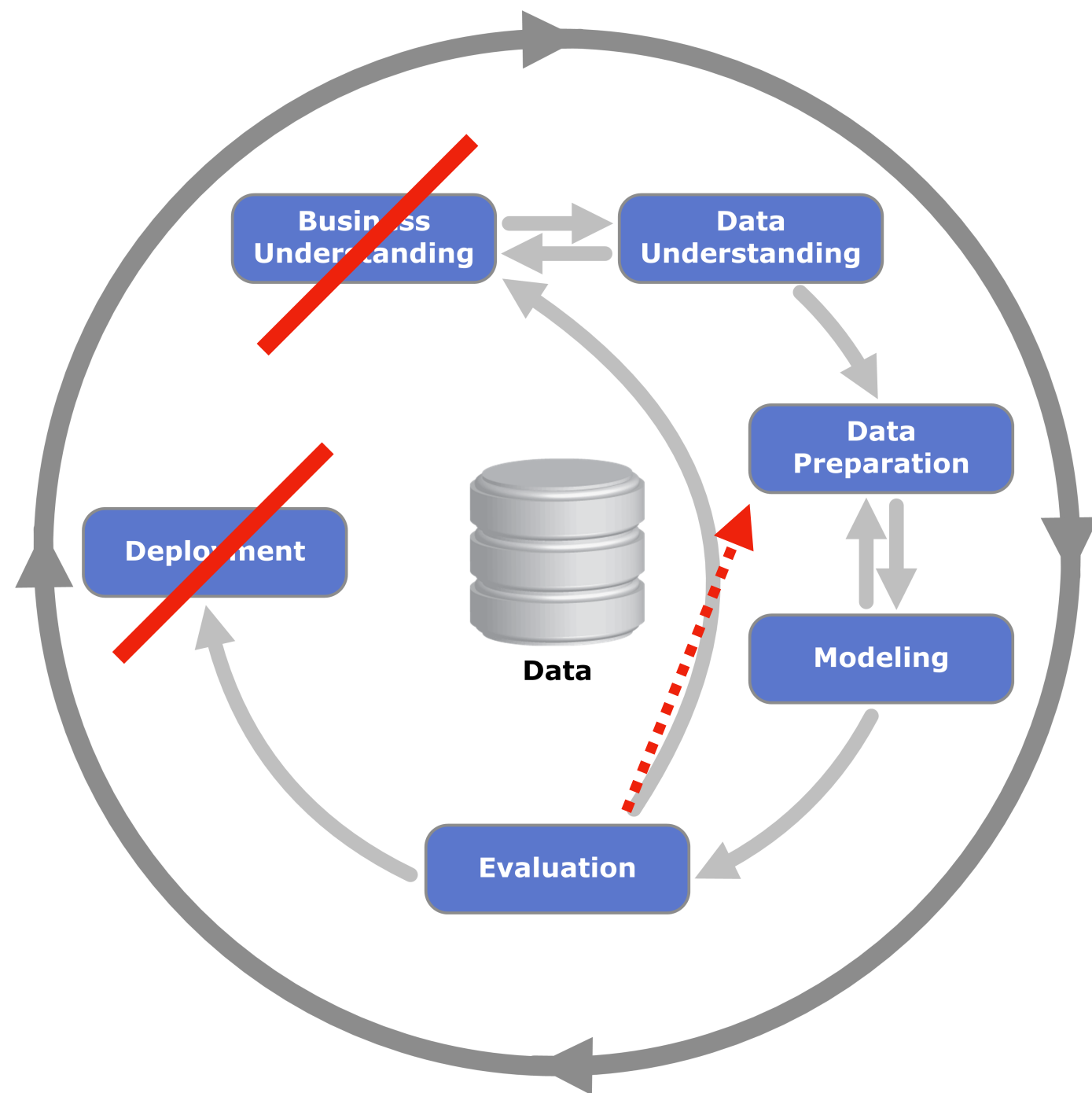
Лето 2019

Эмиль Каюмов

ML процесс в жизни



ML соревнования



ML соревнования

- Не тратится время на постановку и внедрение
- Можно заниматься только частью, связанной напрямую с машинным обучением и анализом данных

Зачем это нужно участнику

- Опыт в различных областях (CV, RL, NLP, sales forecasting, ...)
- Шашечки, медальки и прочее для резюме
- Призы



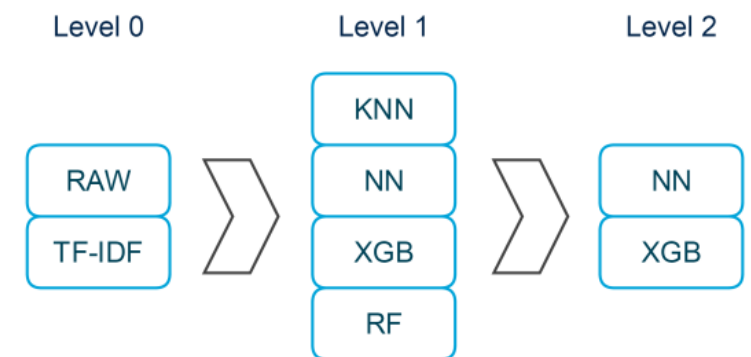
\$30,000
1,675 teams

\$25,000
428 teams

\$1,200,000
1955/3779

Зачем это нужно индустрии

- Новые подходы к решению (факторизационные машины, стэкинг), доказывающие свою мощь
- Новые популярные инструменты и библиотеки
- Внимание сообщества и поиск идей



dmlc
XGBoost

 **RecSys**

Где проходят соревнования

kaggle™



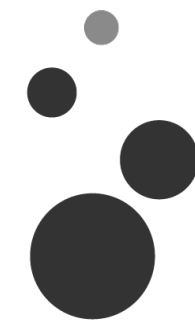
CrowdANALYTIX



topcoder™

crowdAI 

 Boosters.pro



datascience.net

Примеры задач

- Предсказание продаж сети супермаркетов
- Предсказание кликов пользователей
- Выделение фона на изображениях автомобилей
- Обучение модели человека бегу
- Детектирование, подсчёт и измерение длин рыб на видео
- Adversarial attack изображений
- ...


Особенности проведения

- Установлены сроки проведения с дедлайном сдачи
- Имеется небольшое предметное описание задачи
- Набор данных с описанием
- Установлена целевая переменная и метрика оценки
- Лидерборд
- Место для отправки решений
- Правила

Особенности проведения

- Опционально:
 - Форум
 - Публичные скрипты
 - Сдача кода вместо файла с предсказаниями
 - Многоэтапные соревнования

Пример

 Featured Prediction Competition

2018 Data Science Bowl

Find the nuclei in divergent images to advance medical discovery



\$100,000

Prize Money



Booz Allen Hamilton · 1,219 teams · 2 months to go (2 months to go until merger deadline)

Presented by
Booz Allen | Hamilton & kaggle

[Overview](#)

[Data](#)

[Kernels](#)

[Discussion](#)

[Leaderboard](#)

[Rules](#)

[Join Competition](#)

Overview

Description

Evaluation

Prizes

About

Timeline

Spot Nuclei. Speed Cures.

Imagine speeding up research for almost every disease, from lung cancer and heart disease to rare disorders. The 2018 Data Science Bowl offers our most ambitious mission yet: create an algorithm to automate nucleus detection.

We've all seen people suffer from diseases like cancer, heart disease, chronic obstructive pulmonary disease, Alzheimer's, and diabetes. Many have seen their loved ones pass away. Think how many lives would be transformed if cures came faster.

Pipeline решений

1. Прочитать описание задачи и области (могут выигрывать люди той же области, а могут не прочитавшие про задачу)
2. Изучить метрику
3. Проверить форум и лидерборд на наличие проблем в задаче

Pipeline решений (2)

4. Скачать и изучить данные (EDA)
5. Сделать бейзлайн-решение, определиться и проверить валидацию
6. Придумывать, пробовать, улучшать результат до окончания конкурса, возвращаться к форуму, искать особенности в данных
7. Подготовка финального сабмита (тонкая настройка параметров, стэкинг)

Основные проблемы

- Переобучение под лидерборд

#	Δ pub
1	▲ 3
2	▲ 848
3	▲ 6
4	▲ 1
5	▲ 2024

- Невоспроизводимое решение (от «забыл, что за сабмит» до «не получается тот же результат»)

Валидация

- Стандартные техники валидации (K-fold, holdout)
- Time Series Split в задачах со временем
- Важно, чтобы валидационное множество было похоже на тестовую выборку!
 - Похожее распределение признаков
 - Похожее распределение целевой переменной
 - Влияние выбросов
- Ваша локальная валидация должна коррелировать с публичным лидербордом

Кейс: распределение целевой переменной

- Задача классификации на 2 класса, метрика LogLoss

$$LogLoss = -\frac{1}{N} \sum_{i=1}^N (y_i \log p_i + (1 - y_i) \log(1 - p_i))$$

- Можно отправить $p=\text{const}$ и восстановить баланс классов
- Аналогично Accuracy, F-мера за 4 сабмита

Leakage

- Особенности в проведении и ошибки организаторов могут вызывать «особую» информацию, помогающую в решении, но не в реальном применении
 - Информация в ID или порядке строк
 - Мета-данные к файлам
 - «Заглядывание» в будущее

Кейс: Kaggle Expedia

- Задача прогнозирования типа отеля, которые бронирует пользователь
- В данных нет координат отелей, но есть **города пользователей и расстояния до отелей**
- С некоторой точностью можно было восстановить реальное местоположение отеля

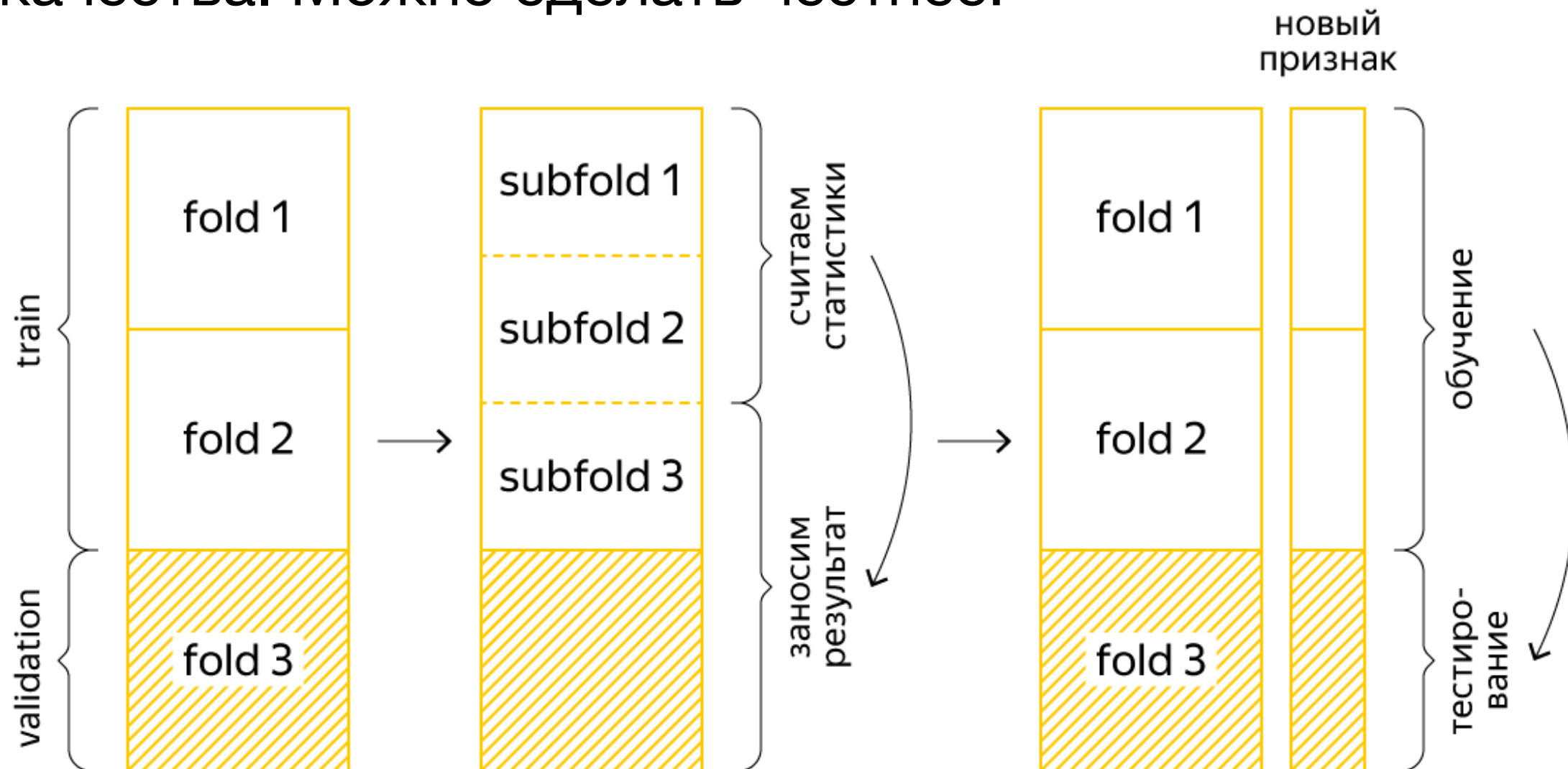
Категориальные признаки

- Обычные подходы:
 - label encoding
 - one-hot encoding
 - mean target encoding

$$\textit{Smoothed Likelihood} = \frac{\textit{mean(target)} * \textit{nrows} + \textit{global mean} * \textit{alpha}}{\textit{nrows} + \textit{alpha}}$$

Валидация с mean target encoding

- Использование K-fold и для валидации, и для подсчёта счётчиков может вносить искажение в оценку качества. Можно сделать честнее:



Тюнинг гиперпараметров

- Погоня за долями процентов качества требует тонкой настройки алгоритмов
 - Grid search — долго
 - HyperOpt, BaeysianOpt — удобнее
 - Вручную — быстрее всего, но требует опыта

Кейс: ручной тюнинг градиентного бустинга

1. Зафиксировать learning rate и параметры случайности (количество итераций лучше не фиксировать)
2. Найти баланс между недообучением и переобучением через сложность (depth) и регуляризацию (min split gain)
3. Настроить параметры случайности
4. Понизить learning rate и увеличить количество итераций для финального алгоритма

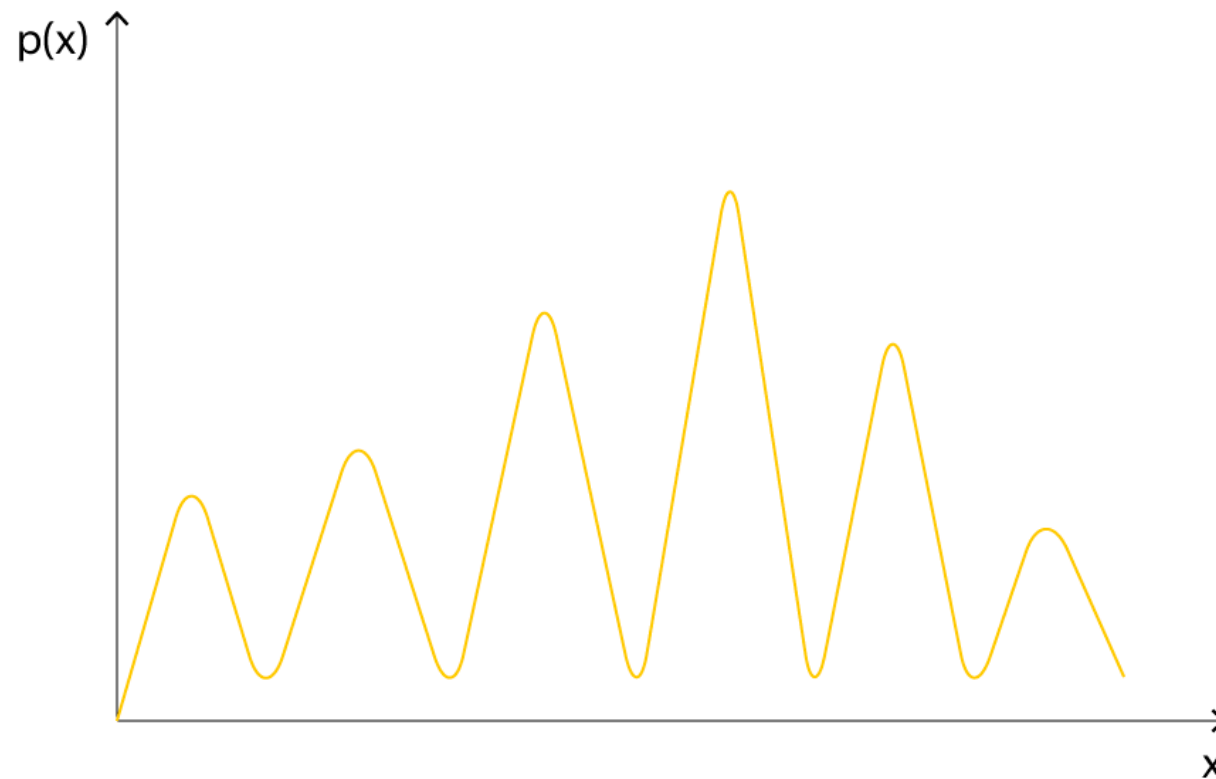
(Не единственно верный, но рабочий способ)

Постпроцессинг

- В некоторых случаях необходимо проводить обработку предсказаний
- Простой пример: бинаризация вероятностей для задачи с F-мерой
- Можно подобрать по кросс валидации порог бинаризации для максимизации результата

Кейс: анонимизированные признаки

- В некоторых задачах компании анонимизируют и шифруют признаки.
- Пример: Kaggle BNP Paribas. Распределение многих признаков:



- Можно было избавиться от искусственного шума

Кейс: клиппинг вероятностей для LogLoss

$$\text{logloss} = -\frac{1}{N} \sum_{i=1}^N (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

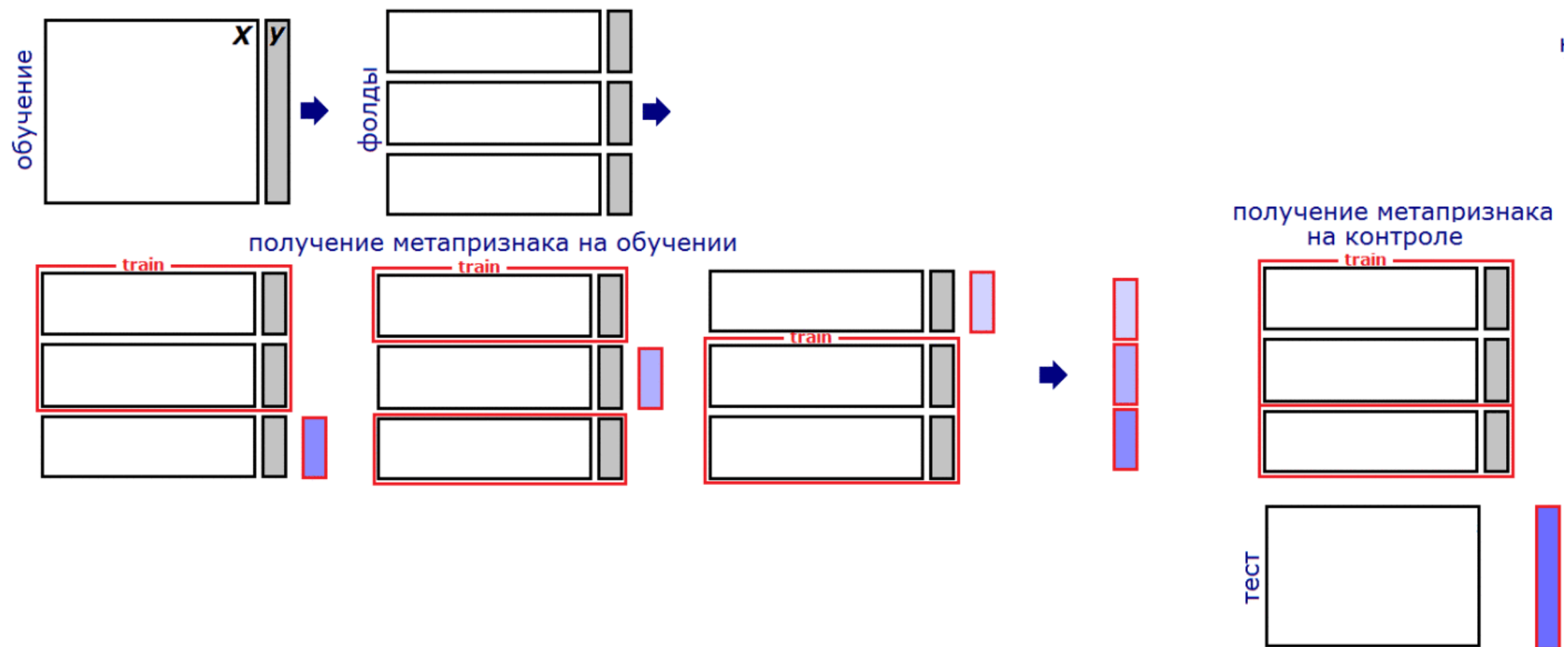
- Логарифм чувствителен к малым числам
- Даже если алгоритм работает хорошо, может встретиться объект с неправильной разметкой
- Если на 1 из 100 объектов предсказать $p=1e-15$ (у объекта положительного класса), то ухудшим ошибку на 0.345 (для сравнения в Kaggle Quora Question Pairs у победителя 0.116)
- Если ограничим с каждой стороны вероятности на $1e-5$, то ошибка на таких объектов уменьшится в 3 раза, а на правильных объектах потеряем лишь по $1e-7$ качества.

Blending

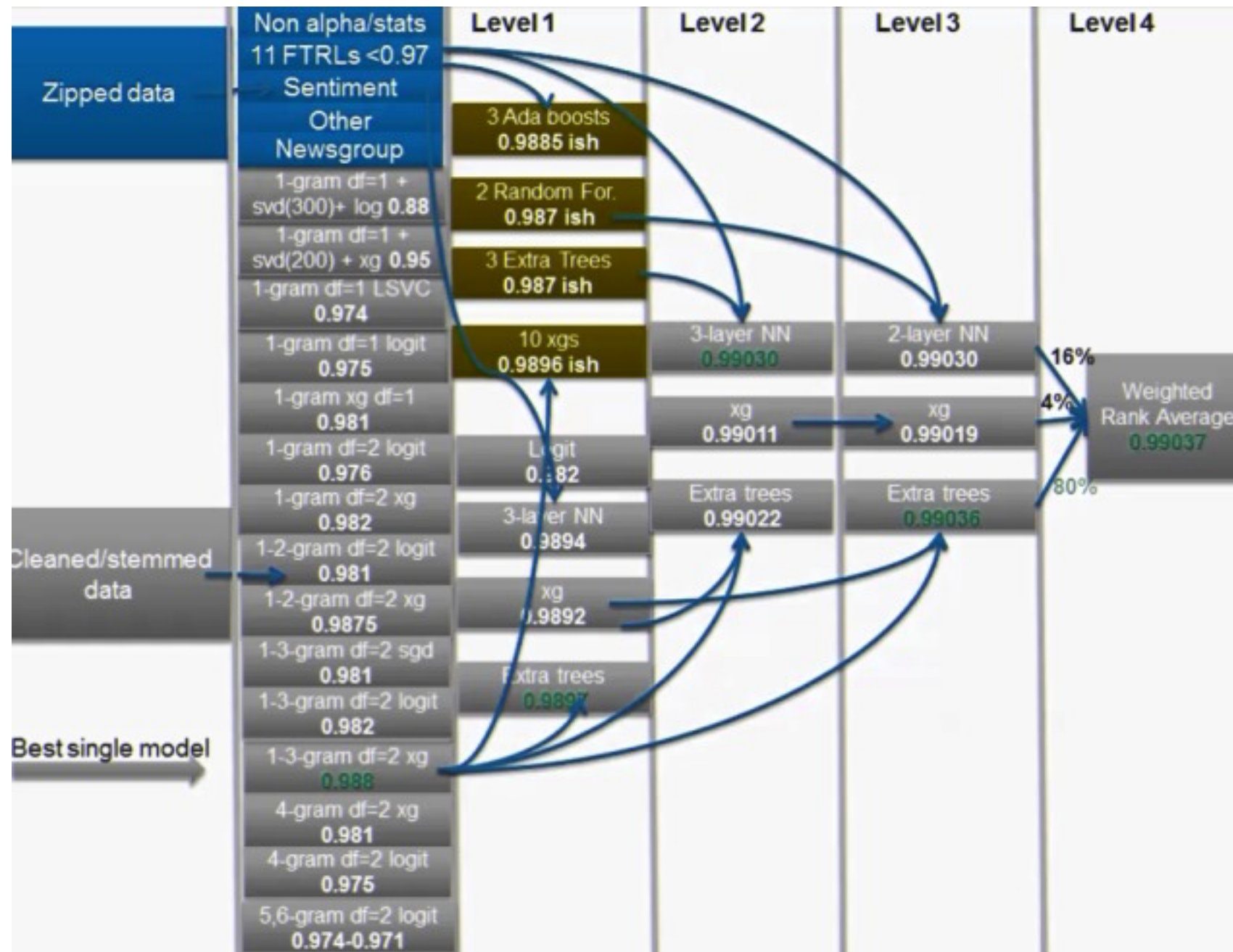
- Разные алгоритмы ведут себя по-разному, хочется компенсировать ошибки одного другим
- Простой способ: усреднить предсказания всех алгоритмов
- Сложнее: подобрать веса для смешивания алгоритмов
- Основной эффект достигается от мало скоррелированных алгоритмов
- Слабый относительно остальных алгоритм может как испортить, так и заметно улучшить результат

Stacking

- Идея: подавать предсказания алгоритмов как новые признаки для других алгоритмов



Кейс: многоуровневый stacking



Kaggle Dato winner

Silent mode

- В некоторых случаях можно скрывать свой результат на лидерборде, чтобы лидеры чувствовали себя спокойно
- Если в задаче с AUC отправить все вероятности как 1-р, то получим результат $1 - \text{AUC} < 0.5$, который никто не увидит, если будет более высокий результат с $\text{AUC} > 0.5$
- Для некоторых метрик можно попробовать проверять результат в 2 сабмита по половине выборки

Материалы

- Coursera: How to Win a Data Science Competition: Learn from Top Kagglers
- Сообщество Open Data Science (каналы #mltrainings_beginners, #mltrainings_live, #kaggle_crackers)
- Youtube: канал ML тренировок (приходите на открытые встречи в офисе Яндекса)