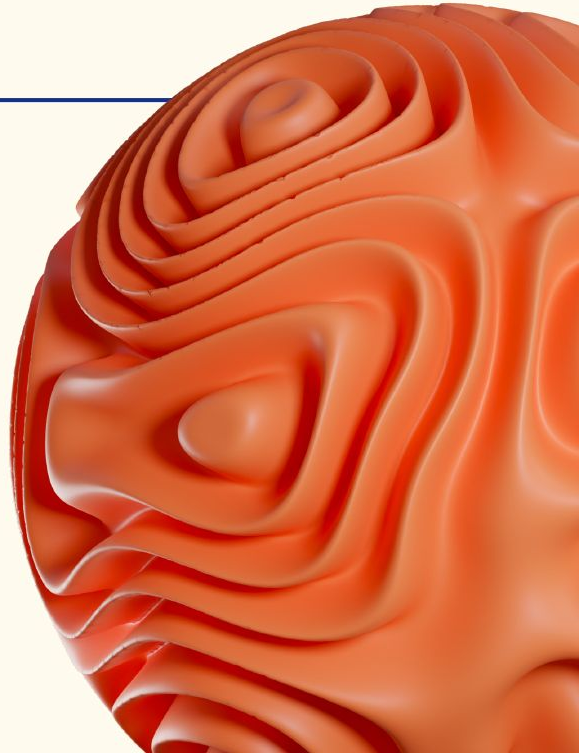
Модели ML в production

МИФИ Национальный исследовательский ядерный университет × SKILLFACTORY

Лекция № 10

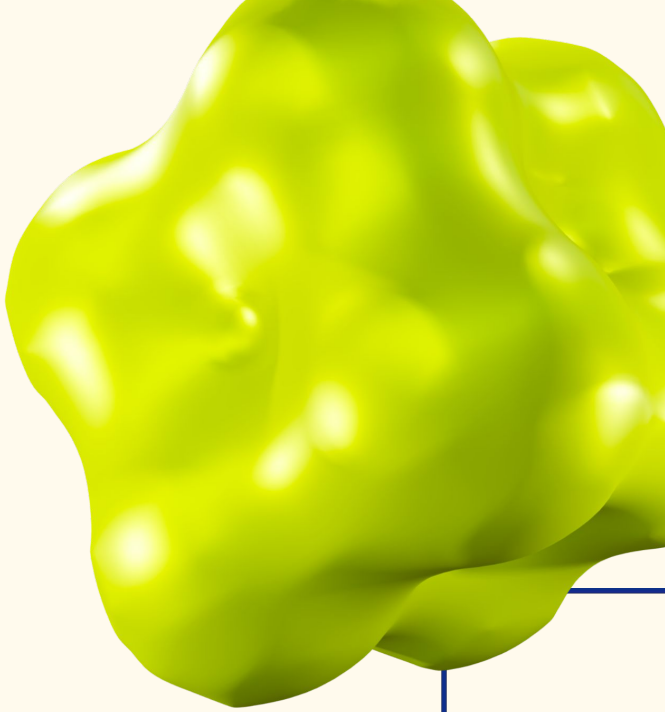
“Kubernetes + бизнес-метрики”



Жарова Мария Александровна
DS WB-tech, Math&Python&DS lecturer
t.me/data_easy

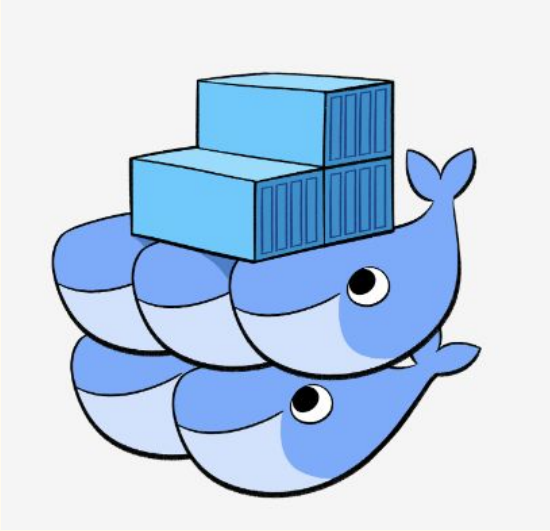
План занятия

1. Обобщение оркестраторов
2. Элементы Kubernetes
3. Бизнес-метрики

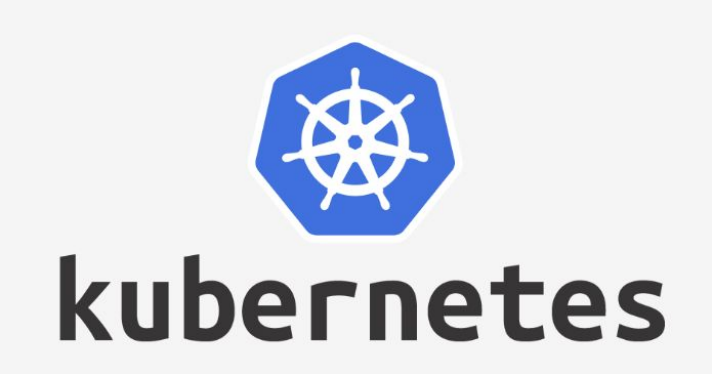


1. Обобщение оркестраторов

Ещё оркестраторы?



Docker Swarm



K8S для распределённых вычислений.

Ещё оркестраторы?

Apache
Airflow

Удобны для ETL, ML workflows, автоматизации.

2. Элементы Kubernetes

Kubernetes, K8S, Кубер...

— это open source инструмент для:

- оркестрации (управления)
- развёртывания
- мониторинга
- масштабирования

приложений в контейнерах.

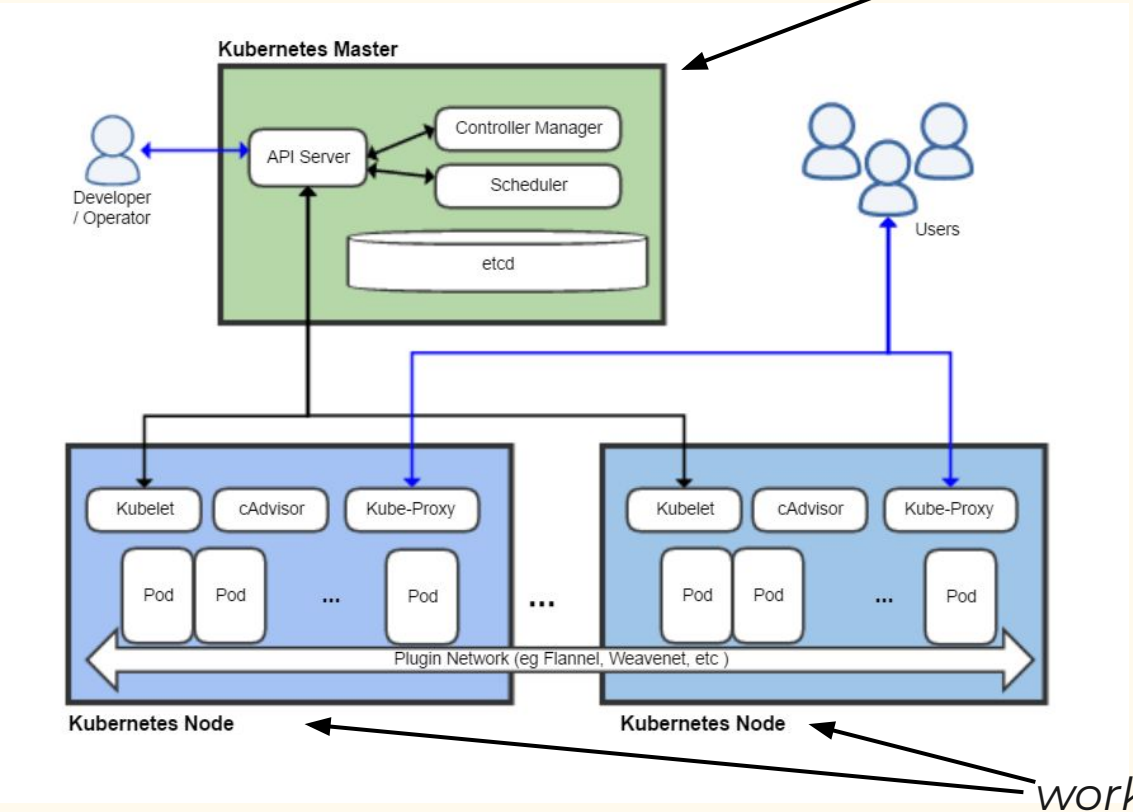
Конкретно, чтобы приложение всегда:

- было доступно
- как можно быстрее работало
- в случае неполадок быстро восстанавливалось

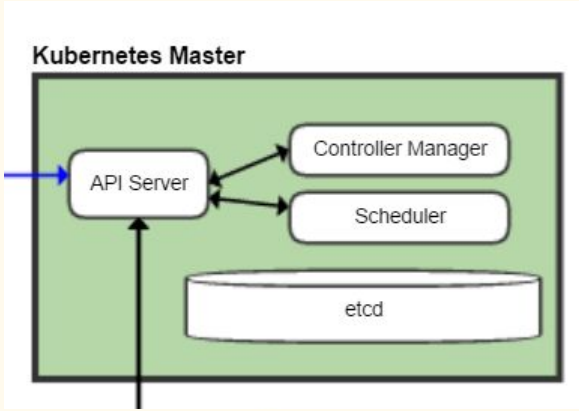
Устройство

control plane
(“мозг” системы)

data plane
(где работают приложения)



Master-нода



API-сервер – способ взаимодействия с кластером (все запросы идут туда).

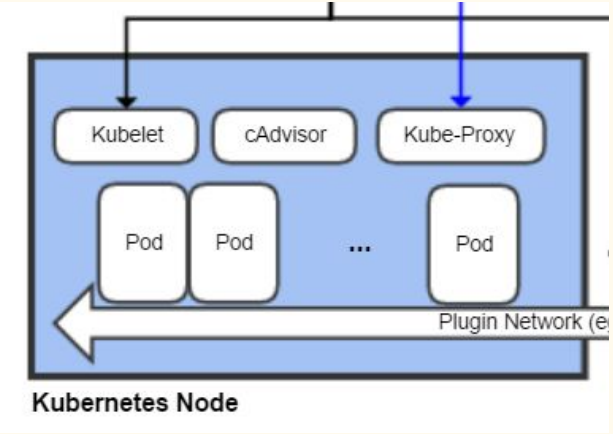
Controller manager – тот, кто следит за сервером (задача: привести current state к desired state), делится на части, отвечающие за узлы, реплики, токены и т.д.

Scheduler – планирует, как расположить контейнеры на нодах кластера (смотрит на загруженность и доступность ресурсов).

Etcd – хранилище данных (хэш-таблица), данные о текущем состоянии кластера (конф. файлы, статусы нод и контейнеров).

Для отказоустойчивости лучше создавать не одну master-ноду!

Worker-нода



Kubelet – получает инструкции от master + передаёт их на *container runtime*.

Kube-proxy – отвечает за коммуникацию и балансировку во внутренней сети.

Container runtime – исполняемая среда контейнера: образы, их запуск, остановка и управление ресурсами.

Pod – там, где запускаются контейнеры; в одном поде может быть запущено несколько контейнеров. У каждого пода собственный IP-адрес.

По сути, K8S управляет подами – в нужный момент перераспределяет между ними нагрузку, уничтожает или создаёт новые.

Как общаться с K8S?

При помощи *конфигурационных YAML-файлов* (тот самый декларативный стиль).

По сути это и есть *desired state*.

```yaml
apiversion: extensions/v1beta1
kind: Deployment
metadata:
  name: nginx-deployment
spec:
  revisionHistoryLimit: 5
  minReadySeconds: 10
  selector:
      matchLabels:
        app: nginx
        deployer: distelli
  strategy:
    type: RollingUpdate
    rollingUpdate:
      maxUnavailable: 1
      maxSurge: 1
    replicas: 3
    template:
      metadata:
        labels:
            app: nginx
            deployer: distelli
      spec:
        containers:
          - name: nginx
            image: nginx: 1.7.9
```

```yaml
apiVersion: apps/v1
kind: Deployment
metadata:
 name: nginx-deployment
 labels:
  app: nginx
spec:
 replicas: 3
 selector:
  matchLabels:
   app: nginx
 template:
  metadata:
   labels:
    app: nginx
  spec:
   containers:
   - name: nginx
     image: nginx:1.14.2
     ports:
     - containerPort: 80
```

3. Бизнес-метрики

Метрики: что важно для бизнеса

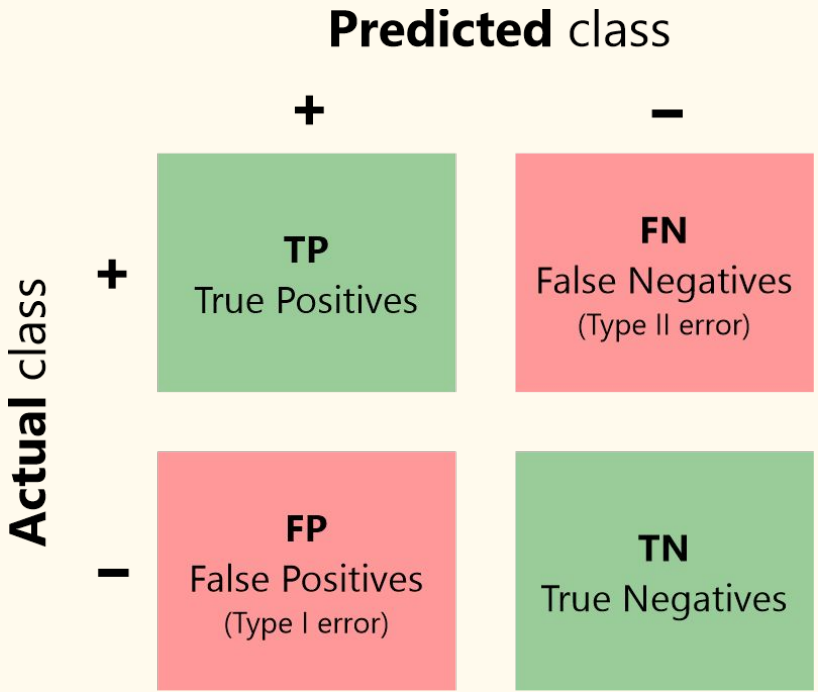
Правда жизни:

- бизнес-заказчиков не интересуют технические метрики (по типу roc-auc, precision, recall, mse и т.д...), т.к. они не совсем их понимают
- бизнес-заказчикам важна полученная прибыль🤑

Необходимо “переводить” наши технические метрики и результаты в *бизнес-ценность*, объясняя это *бизнес-языком*

Это умение является одним из главных, отличающих джуна от мидла от сеньёра от тимлида:)

Вспоминаем confusion matrix



False Positive - ошибка первого рода

False Negative - ошибка второго рода

Точность $$precision = \frac{TP}{TP + FP}$$

(оптимизирует ошибку 1 рода)

Полнота $$recall = \frac{TP}{TP + FN}$$

(оптимизирует ошибку 2 рода)

Вспоминаем confusion matrix

F-мера:

$$F_\beta = (1 + \beta^2) \frac{precision \times recall}{\beta^2 precision + recall}$$

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

Про вес метрик:

- Если $0 < \beta < 1$, то *precision* важнее.
- Если $\beta > 1$, то *recall* важнее.
- Если $\beta = 1$, это среднее гармоническое.

F-мера достигает **максимума** при precision и recall, равных **единице.**

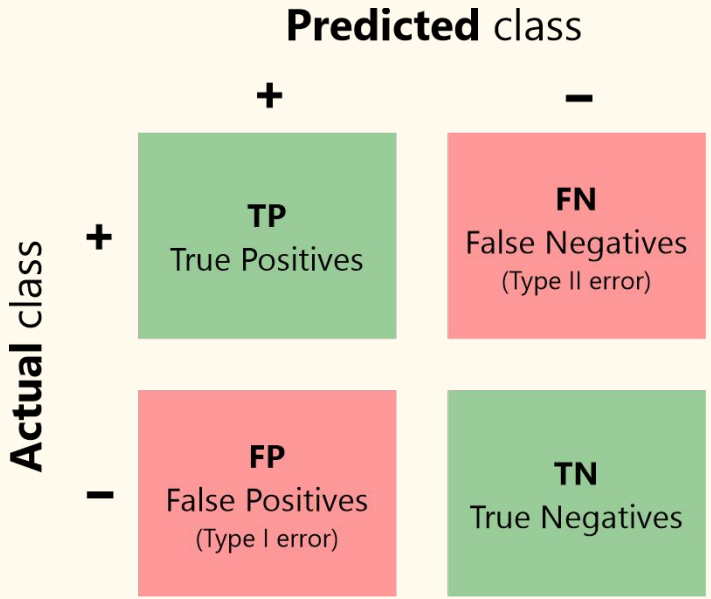
F-мера **близка к нулю**, если один из аргументов **близок к нулю.**

Пример кейса

Мы владеем компанией, предоставляющей услуги сотовой связи. Нам нужно, чтобы клиенты от нас не уходили к конкурентам.

Давайте попробуем предсказать, когда клиент захочет уйти, чтобы вовремя предложить им скидку (решаем **задачу оттока**).

Итак, цель — выявлять клиентов в группе риска, чтобы компания вовремя предлагала им скидку.



Пример кейса

Модель оттока:

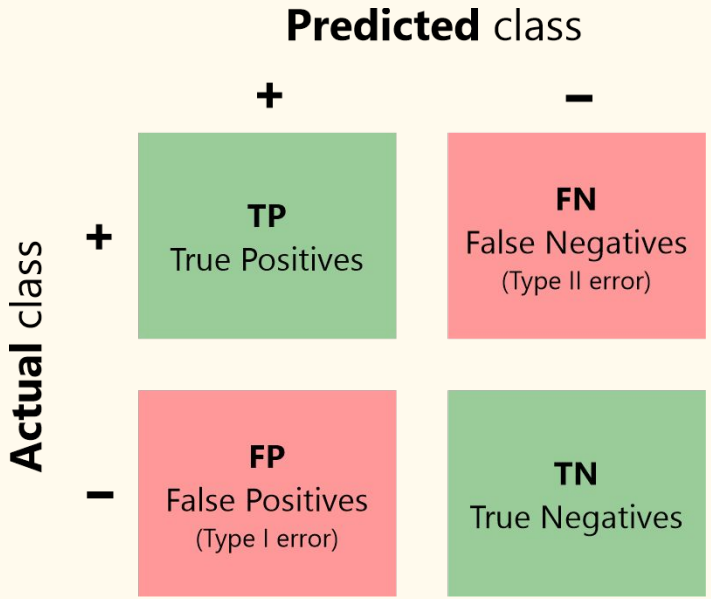
- предсказывает вероятность ухода
- таргет:
  - 1 - клиент правда хочет уйти
  - 0 - клиент не собирается уходить

Ошибка первого рода:

- модель: клиент собирается уходить, даём скидку
- клиент: я не собирался уходить:)

Ошибка второго рода:

- модель: клиент не собирается уходить, не даём скидку
- клиент: я собрался уходить:(



Какая ошибка хуже?

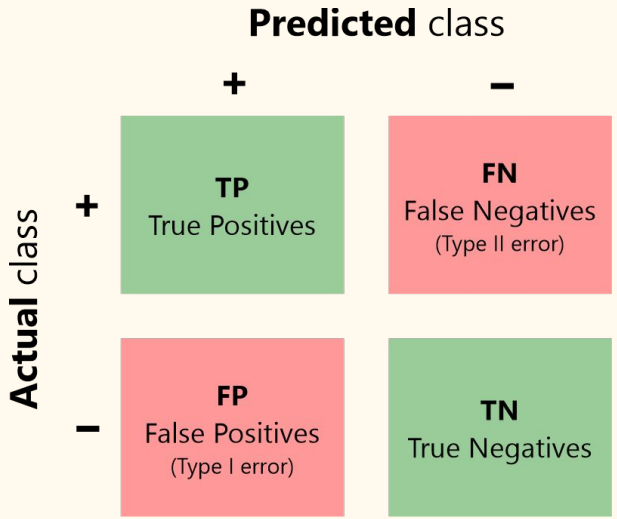
Потерять клиента == потерять выручку от него (LTV).

$$TP\,(LTV - D_{cost}) + TN\,(LTV) + FP\,(LTV - D_{cost}) + FN\,(-LTV) \rightarrow max$$

LTV - Life Time Value - суммарная выручка от клиента

ИЛИ

“пожизненная ценность” — предсказание чистого дохода, связанного со всеми будущими отношениями с клиентом.



Ещё примеры бизнес-метрик

Return on Investment (ROI) – показатель эффективности для оценки инвестиций или сравнения разных проектов. Показывает сумму прибыли на конкретный проект относительно суммы инвестиций.

$$ROI = \frac{Current\ Value\ of\ Investment - Cost\ of\ Investment}{Cost\ of\ Investment}$$

$$\textbf{AOV} = \frac{\textbf{Сумма всех покупок}}{\textbf{Количество покупок}}$$

Average Revenue Per User = общий доход / количество клиентов за месяц

$$\text{CPO} = \frac{\text{затраты на рекламу}}{\text{число продаж}}$$

+ GMV (Gross Merchandise Value) — валовая сумма всех покупок
+ CR (Conversion Rate)

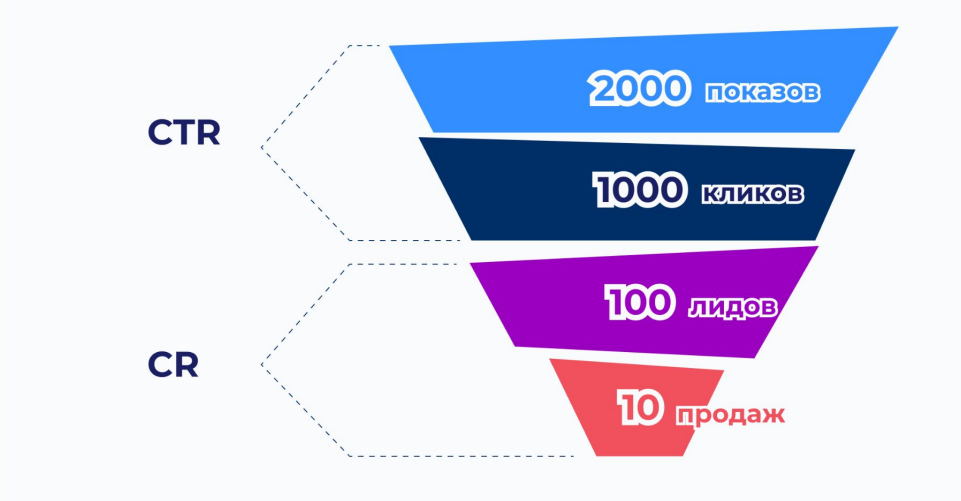
Конверсия и воронка

CR — может вычисляться на различных этапах пути клиента — *воронки*.

То, что относится к кликам, обычно обозначают CTR.

$$CR = \frac{\text{Количество целевых действий}}{\text{Общее число посетителей}} \times 100\%$$

$$CTR = \frac{\text{Клики}}{\text{Показы}} \times 100$$



Спасибо
за внимание!

МИФИ Национальный исследовательский ядерный университет × SKILLFACTORY