

Полный цикл проекта по машинному обучению

Елена Кантонистова

ВШЭ, 2022

План рассказа



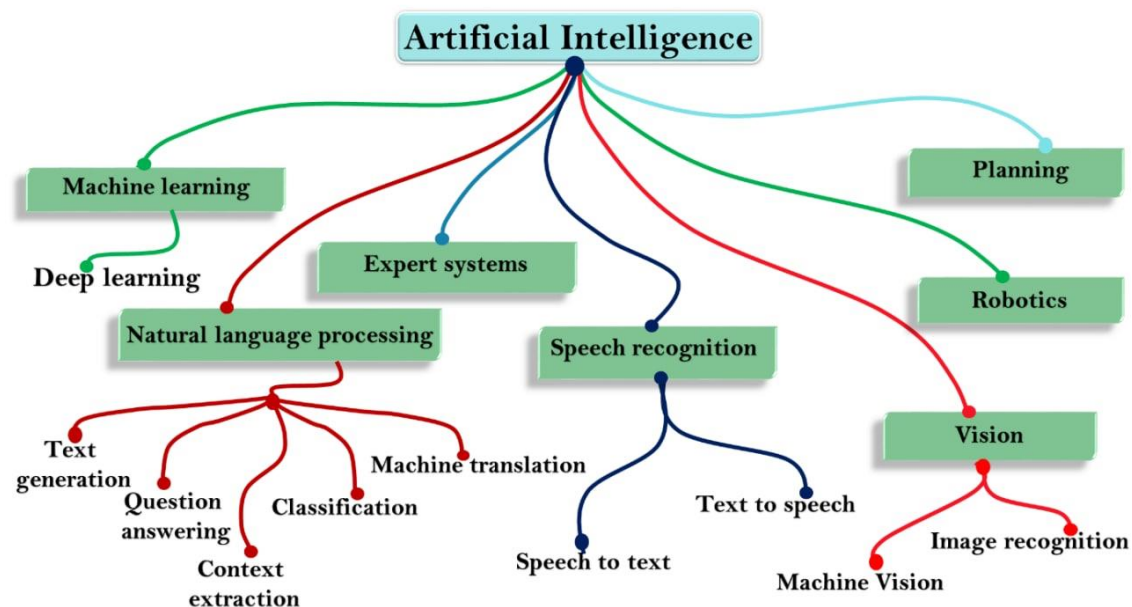
- Что такое машинное обучение. Основные понятия машинного обучения
- Этапы проекта по машинному обучению
- Практика: решение задачи определения оттоковых клиентов
- Оркестрация пайплайнов

Что такое машинное обучение



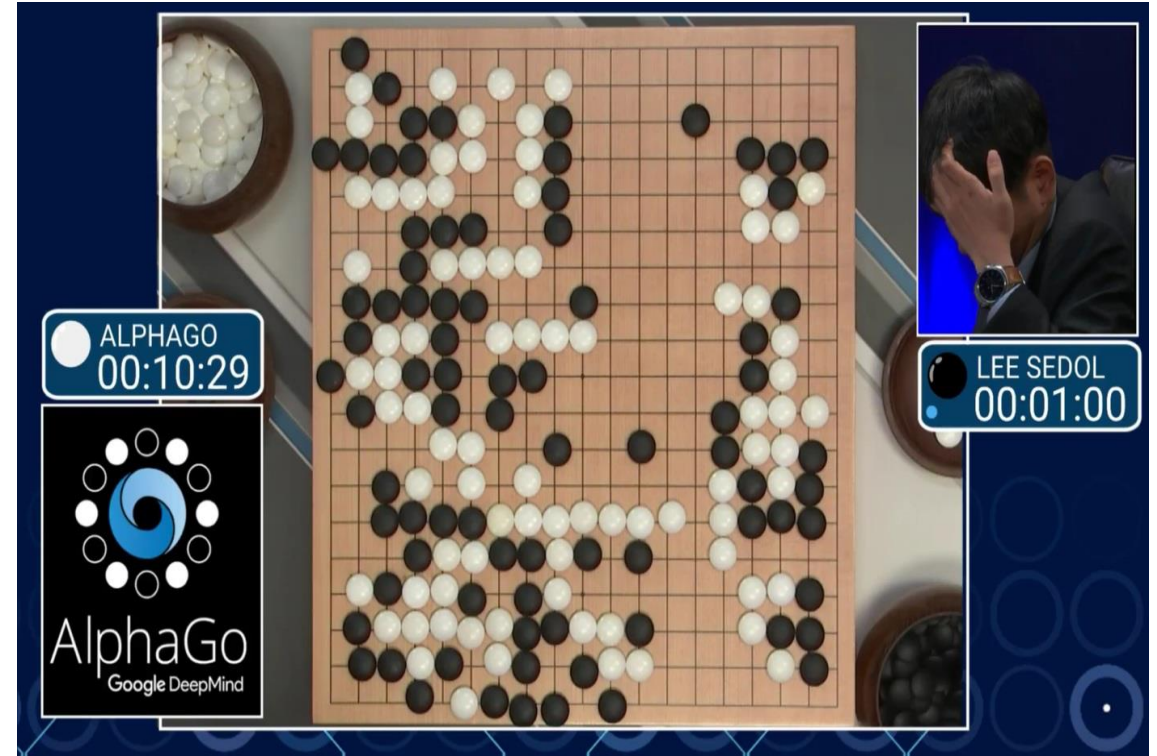
Что такое машинное обучение

Машинное обучение - это часть области искусственного интеллекта, занимающаяся изучением алгоритмов, способных *самостоятельно обучаться, то есть автоматически искать зависимости* в данных. В этом отличие машинного обучения от классической аналитики, когда люди вручную подбирают формулы и правила, по которым определяются зависимости в данных.



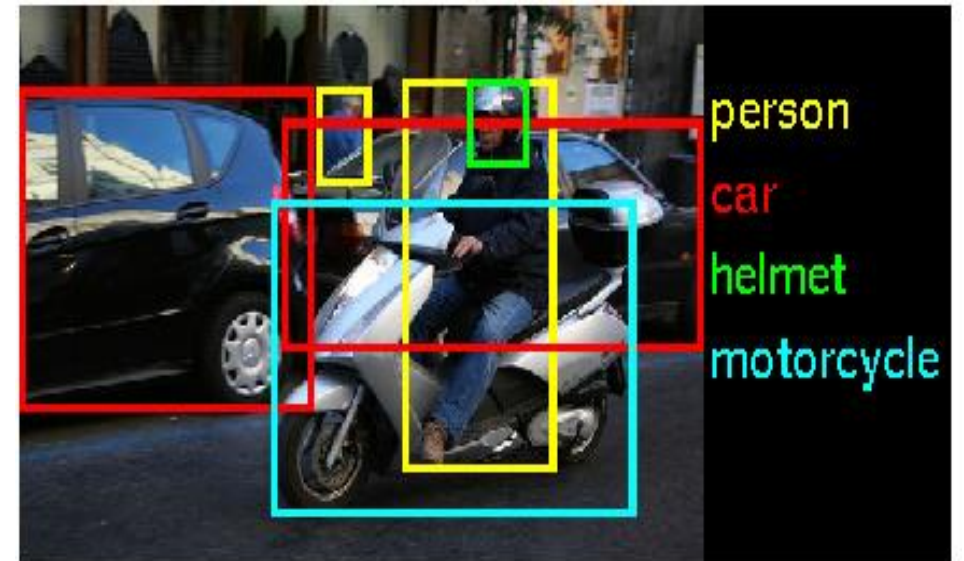
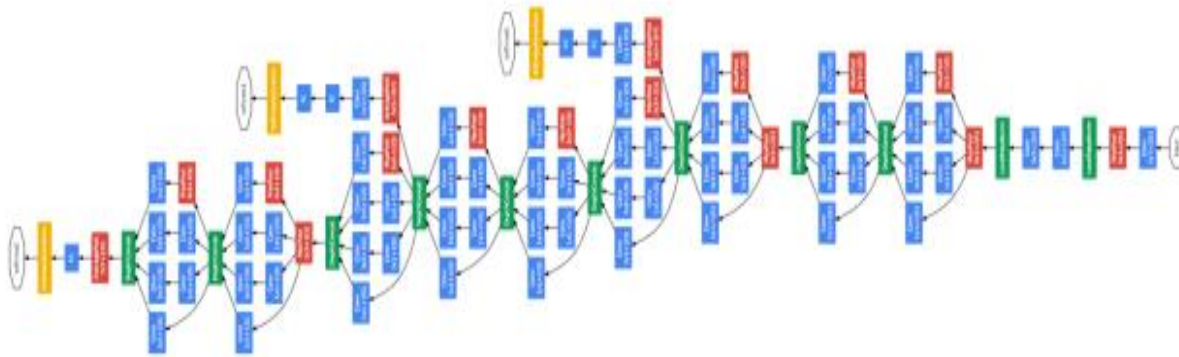
Примеры

- Нейронная сеть, играющая в Го
- **Март 2016** – победа над мировым чемпионом
- Нейронная сеть обучалась, играя сама с собой для увеличения объёмов входных данных (принцип обучения с подкреплением, reinforcement learning)



Примеры

- **ImageNet** — задача распознавания объектов на изображении
- Решается с помощью нейронных сетей с точностью, превышающей точность работы человека



Примеры

- Аннотирование изображений



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



"two young girls are playing with legos toy."



"boy is doing backflip on wakeboard."



"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



"young girl in pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."

Примеры

- Чтение по губам

Google Deepmind в **2017** году создали модель, обученную на телевизионном датасете, которая смогла превзойти профессионального lips reader'а с канала BBC.



Трансформеры в задачах анализа текстов



✓ В **2017** году в Google

Выпустили статью “Attention is all you need”, описывающую механизм внимания – механизм, используемый в нейронных сетях для извлечения информации из текста.

✓ Трансформер – архитектура нейронной сети, основанная на механизме внимания. Она даёт state-of-the-art (SOTA) результаты во многих задачах машинного обучения, связанных с обработкой естественного языка:

- Определение тональности текста
- Перевод с одного языка на другой
- Определение связности предложений в тексте и др.

Пример: ответы на вопросы по тексту

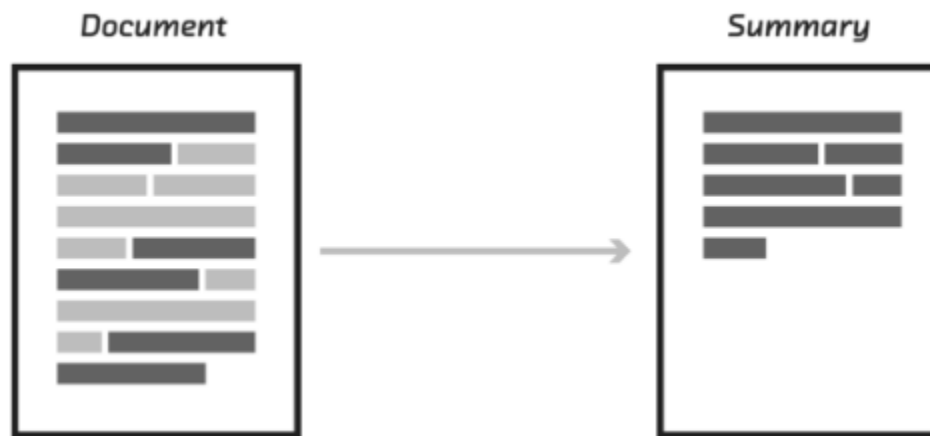
- Context: *Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.*
- Question: *The Basilica of the Sacred heart at Notre Dame is beside to which structure?*

Пример: ответы на вопросы по тексту

- Context: Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to **the Main Building** is **the Basilica of the Sacred Heart**. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.
- Question: *The Basilica of the Sacred heart at Notre Dame is beside to which structure?*
- Answer: *start_position: 49, end_position: 51*

Пример: суммаризация текстов

Суммаризация – получение смысловой выжимки из текста. С 2019 года сильно улучшилось качество суммаризации, благодаря внедрению Deep Learning подхода, основанного на механизме внимания.



[сервис для сокращения текста](#)

Пример: рекомендации

Рекомендации Netflix:

Profile Type	Score Image A	Score Image B
Comedy	5.7	6.3
Romance	7.2	6.5



Image A



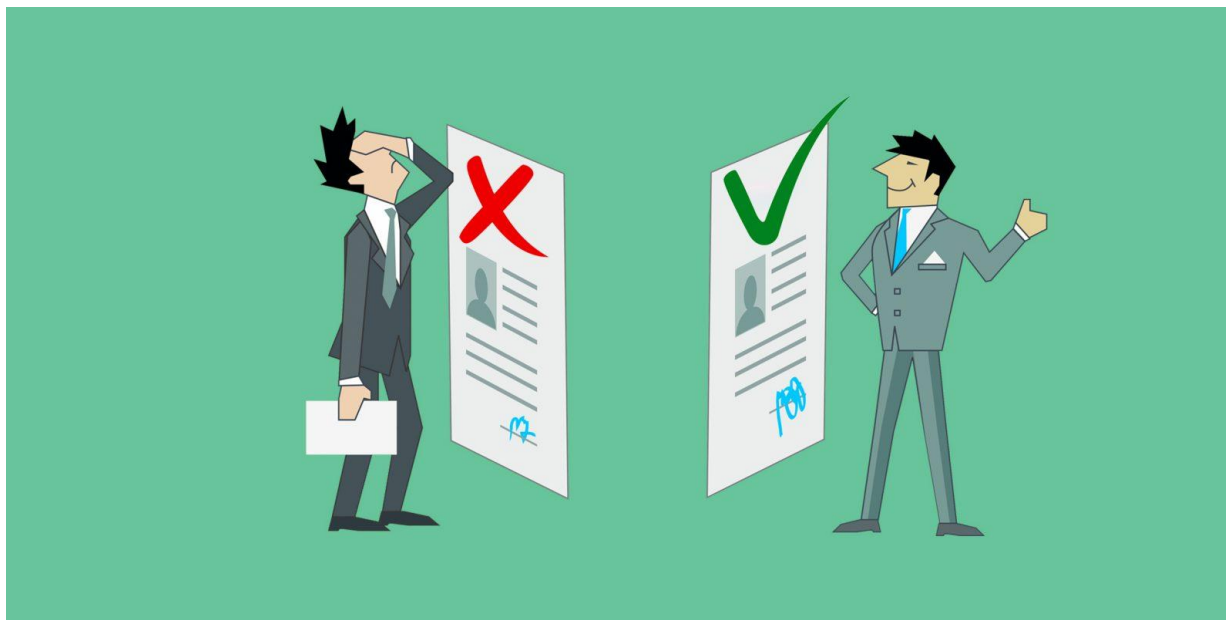
Image B

Основные понятия машинного обучения



Пример: задача скоринга

- Пусть по характеристикам клиента (пол, возраст, средний доход, рейтинг кредитной истории и так далее) мы хотим предсказать, **вернёт клиент кредит или не вернёт.**



Пример: задача скоринга

- **Целевая переменная (target)**, то есть величина, которую хотим предсказать - это число (например, 1 - если человек вернет кредит, и 0 иначе).
- Характеристики клиента, а именно, его пол, возраст, доход и так далее, называются **признаками (features)**.
- Сами же клиенты - сущности, с которыми мы работаем в этой задаче - называются **объектами (objects)**.

Обучение алгоритма

- На **этапе обучения** происходит анализ большого количества данных, для которых у нас имеются правильные ответы (например, клиенты, про которых мы знаем - вернули они кредит или нет; пациенты и их анализы, где про каждого пациента мы знаем, болен он или здоров и так далее).



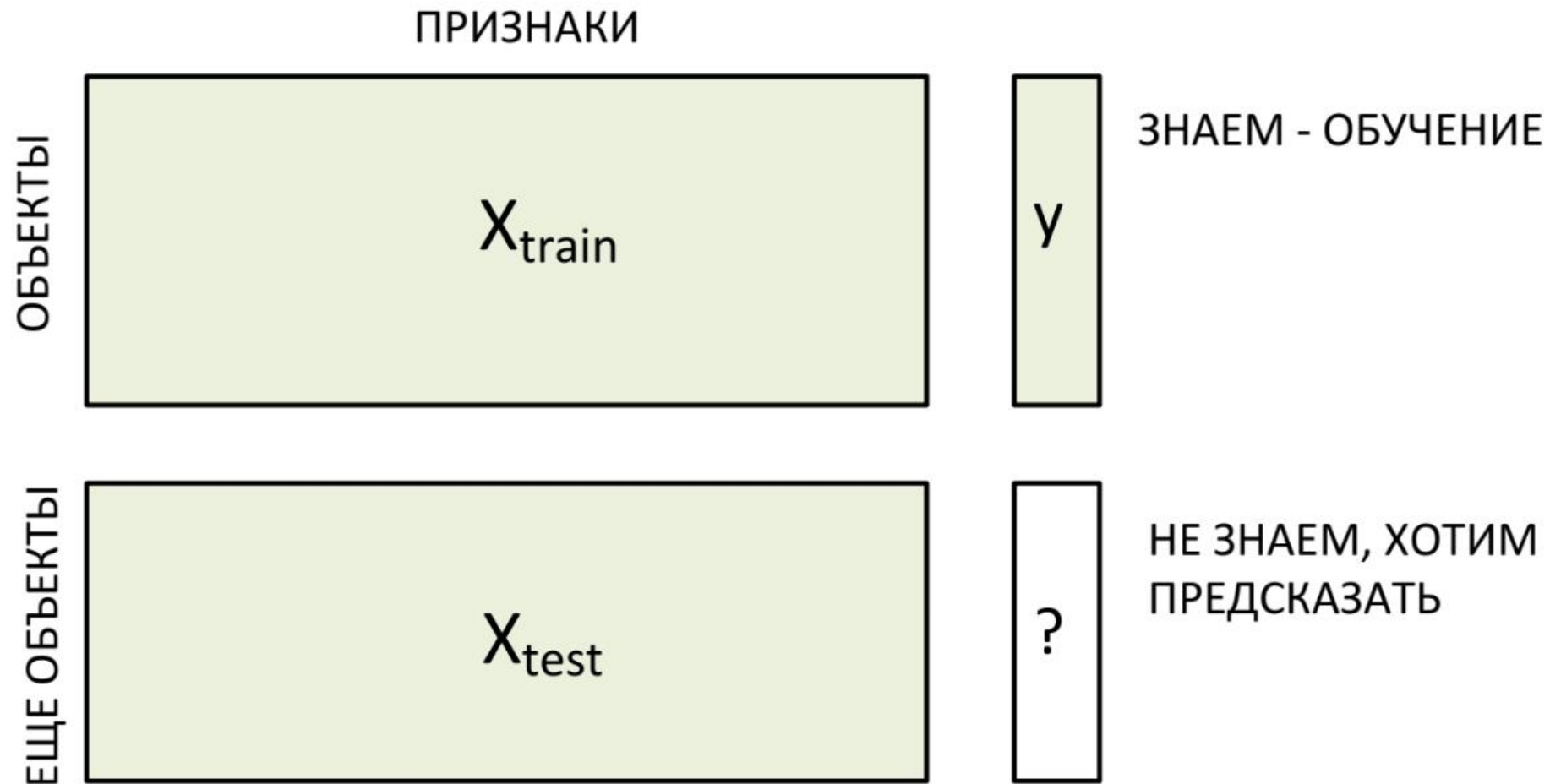
- Модель машинного обучения изучает эти данные и старается научиться делать предсказания таким образом, чтобы для каждого объекта предсказывать как можно более точный ответ. Все данные с известными ответами называются **обучающей выборкой**.

Применение алгоритма



- На **этапе применения** готовая (уже обученная) модель применяется для того, чтобы получить ответ на новых данных. Например, у нас есть подробная информация о клиентах, и мы применяем модель, чтобы она предсказала, кто из них вернет кредит, а кто нет.

Этапы машинного обучения



Типы задач в ML: Классификация

В задачах **классификации** целевая переменная - это класс объекта. То есть в задачах классификации ответ может быть одним из конечного числа классов.

Например, в задаче бинарной (или двухклассовой) классификации мы можем предсказывать:

- пол клиента (мужчина или женщина)
- уйдет клиент из компании или нет
- вернет человек кредит или нет
- болен пациент или здоров и т. д.



Примеры задач классификации

- Задачи медицинской диагностики (пациент здоров или болен)
- Задачи кредитного скоринга (выдаст банк кредит данному клиенту или нет)
- Задача предсказания оттока клиентов (уйдет клиент в следующем месяце или нет)
- Предсказание поведения пользователя (кликнет пользователь по данному баннеру или нет)
- Классификация изображений (на изображении кошка или собака)

Примеры задач классификации

Мультиклассовая классификация

- Определение типа объекта на изображении



Pedestrian



Car



Motorcycle



Truck

- Определение наиболее подходящей профессии для данного кандидата

Типы задач в ML: Регрессия

В задачах **регрессии** целевая переменная может принимать бесконечно много значений. Например, прибыль фирмы может быть любым числом (как очень большим, так и очень маленьким) - даже отрицательным или нецелым.



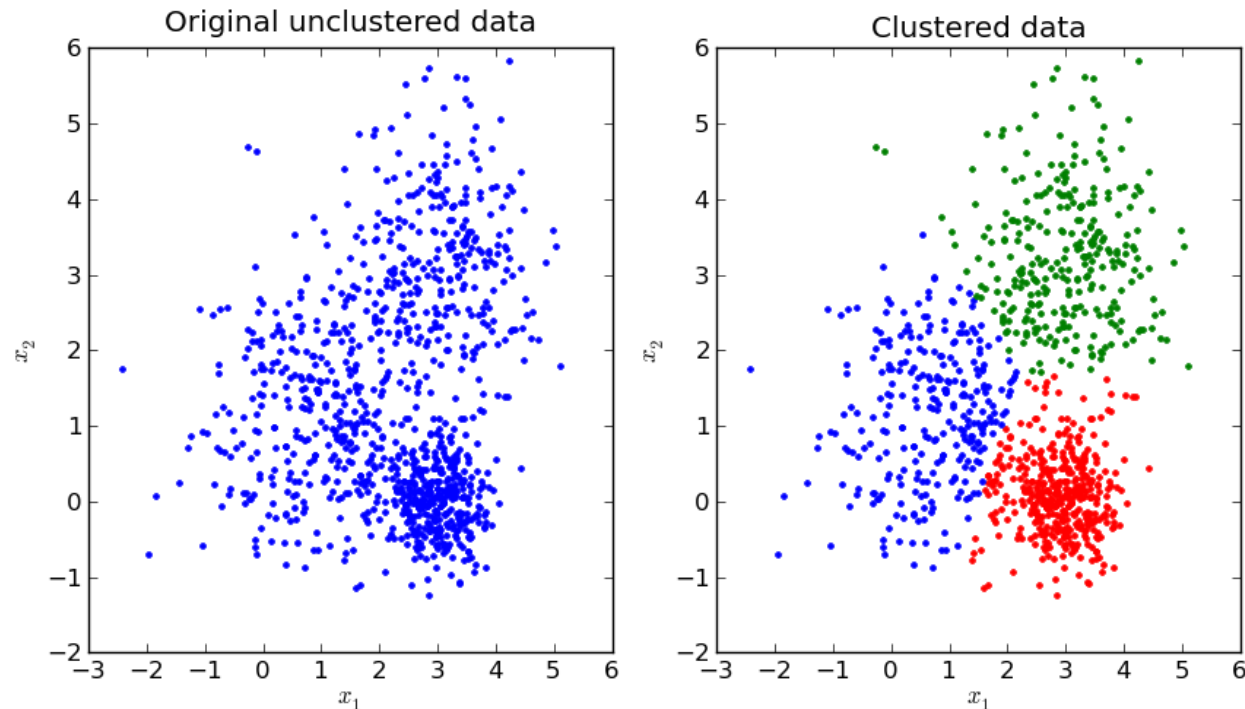
Примеры задач регрессии



- Предсказание стоимости недвижимости (стоимость квартиры в Москве)
- Предсказание прибыли ресторана
- Предсказание поведения временного ряда в будущем (стоимость акций)
- Предсказание зарплаты выпускника вуза по его оценкам

Типы задач в ML: кластеризация

Кластеризация – задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаков описаний объектов.



Типы задач машинного обучения

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это **обучение с учителем**. Сюда относят классификацию, регрессию и ранжирование.
- Если нам неизвестны значения целевой переменной или целевая переменная вообще отсутствует, то есть алгоритм обучается только по признакам объектов, то это **обучение без учителя**. Примерами обучения с учителем являются кластеризация, понижение размерности и др.

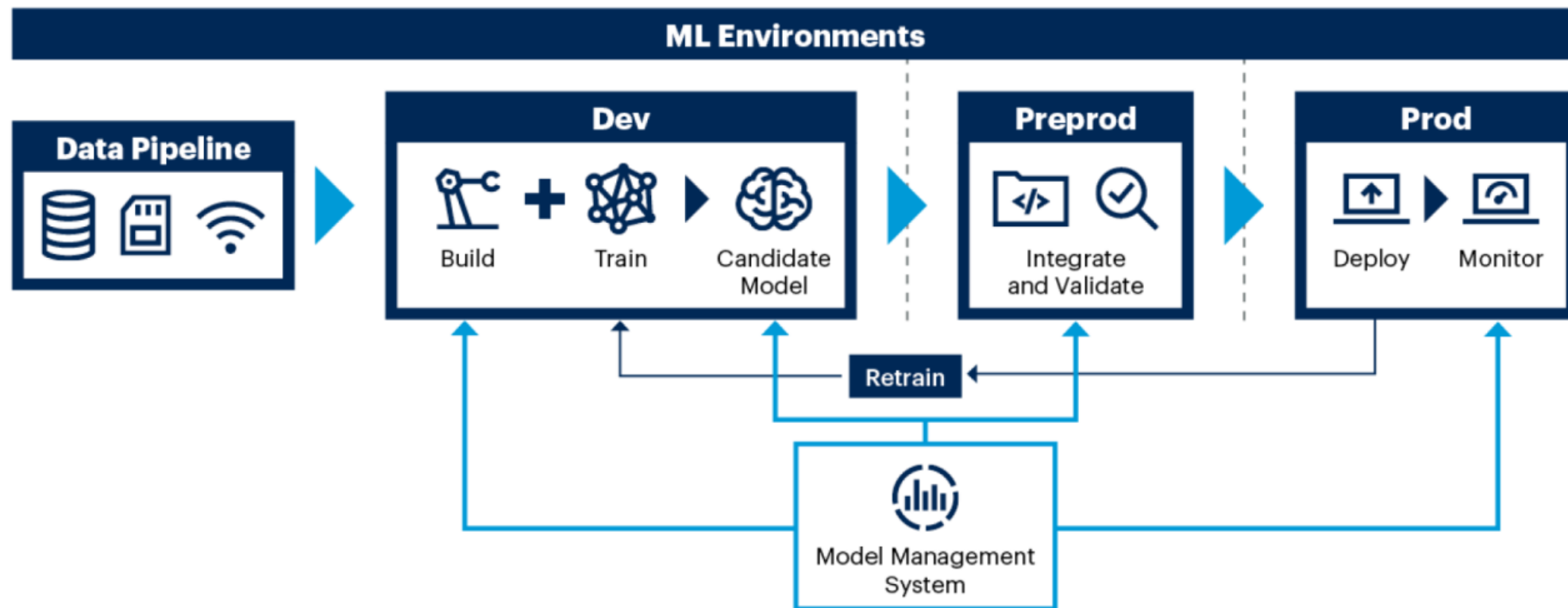


Что такое машинное обучение



Схема проекта по машинному обучению

Typical ML Pipeline



Source: Gartner

718951_C

Схема проекта по машинному обучению



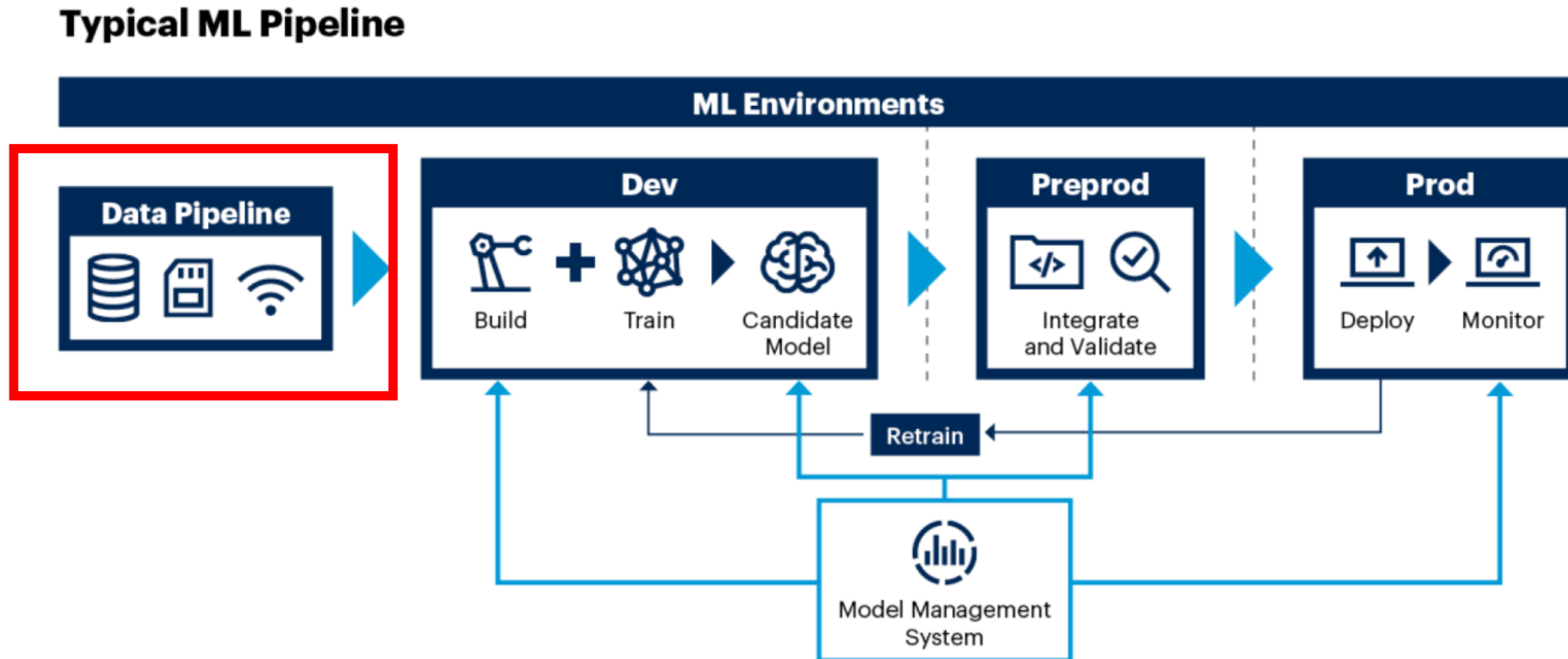
1. Постановка задачи
2. Работа с данными
3. Обучение и валидация модели
4. Тестирование модели на новых пользователях
5. Внедрение модели и мониторинг
6. Оркестрация процессов

1. Постановка задачи



- Что нужно сделать?
- Какие есть данные?
- Где хранятся данные?
- Какие метрики качества решения?
- Когда и в каком виде предоставить решение?
- Какие технологии необходимо использовать в проекте?

2. Этап работы с данными



Source: Gartner

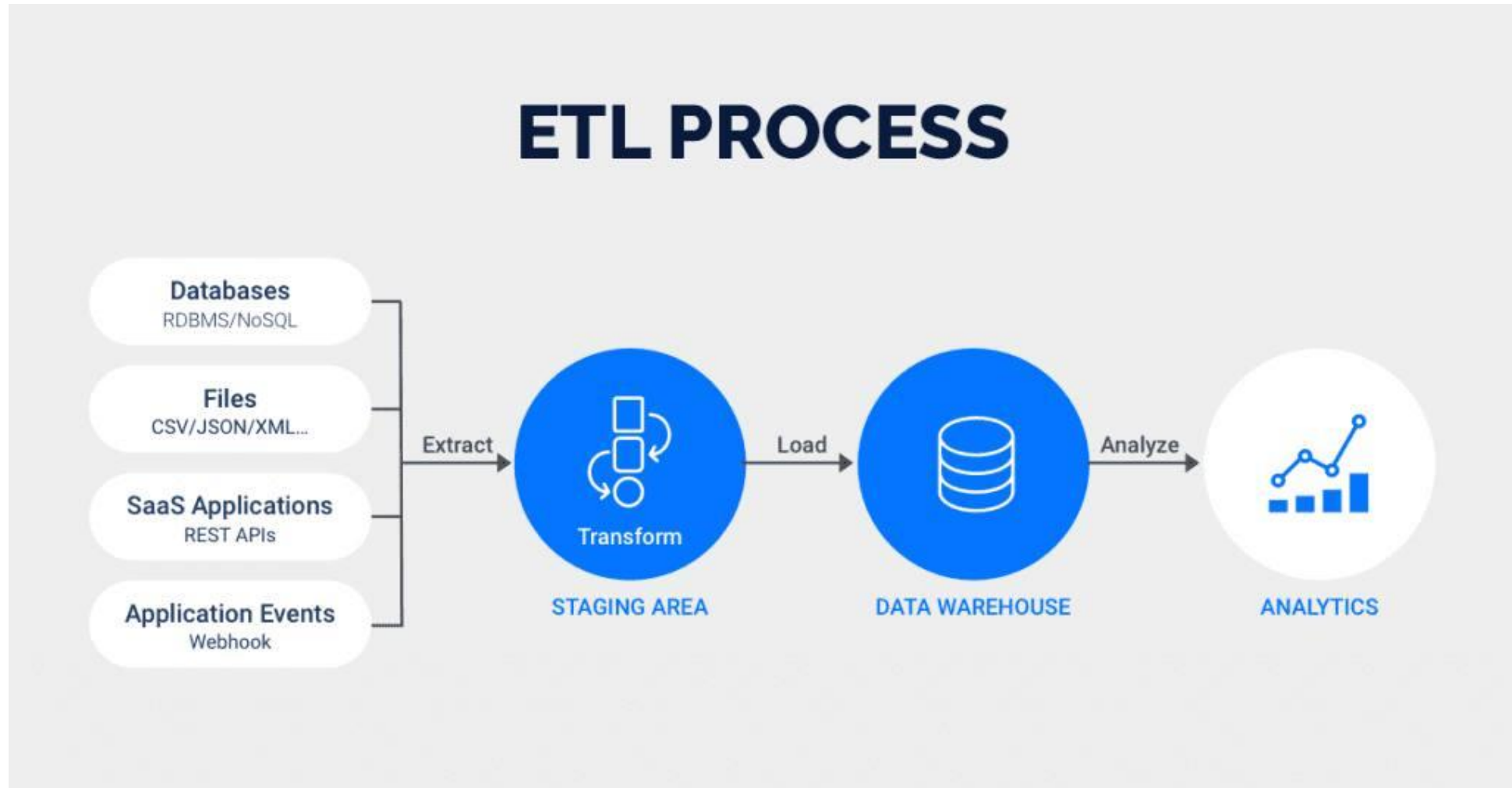
718951_C

2. Этап работы с данными

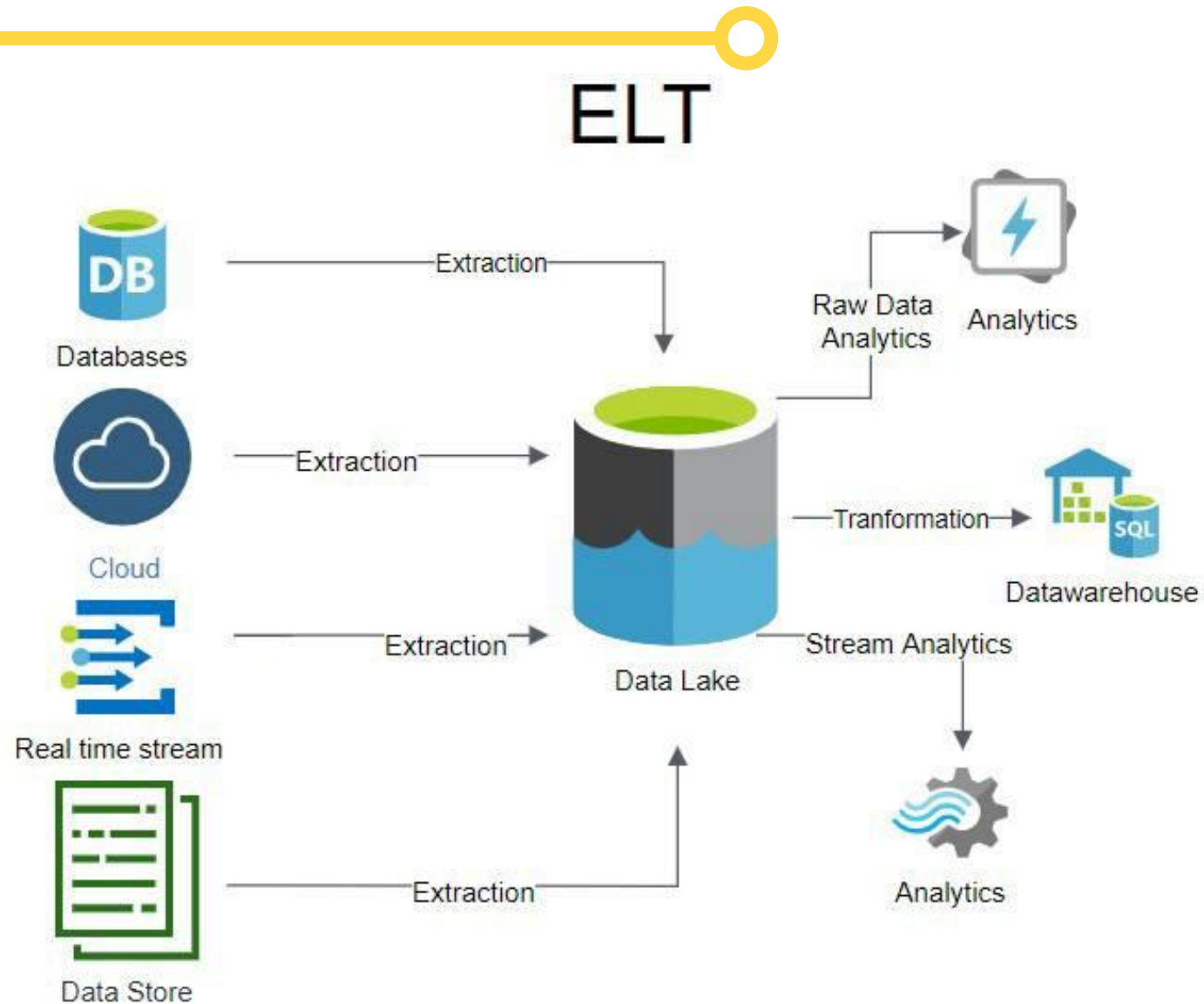


1. *Сбор данных:* в каких источниках хранятся данные? Есть ли к ним доступы?
2. *Обработка данных:*
 - Проверка качества данных
 - Очистка данных
 - Feature engineering
 - Агрегация данных
3. *Загрузка данных в хранилище*

2. Этап работы с данными

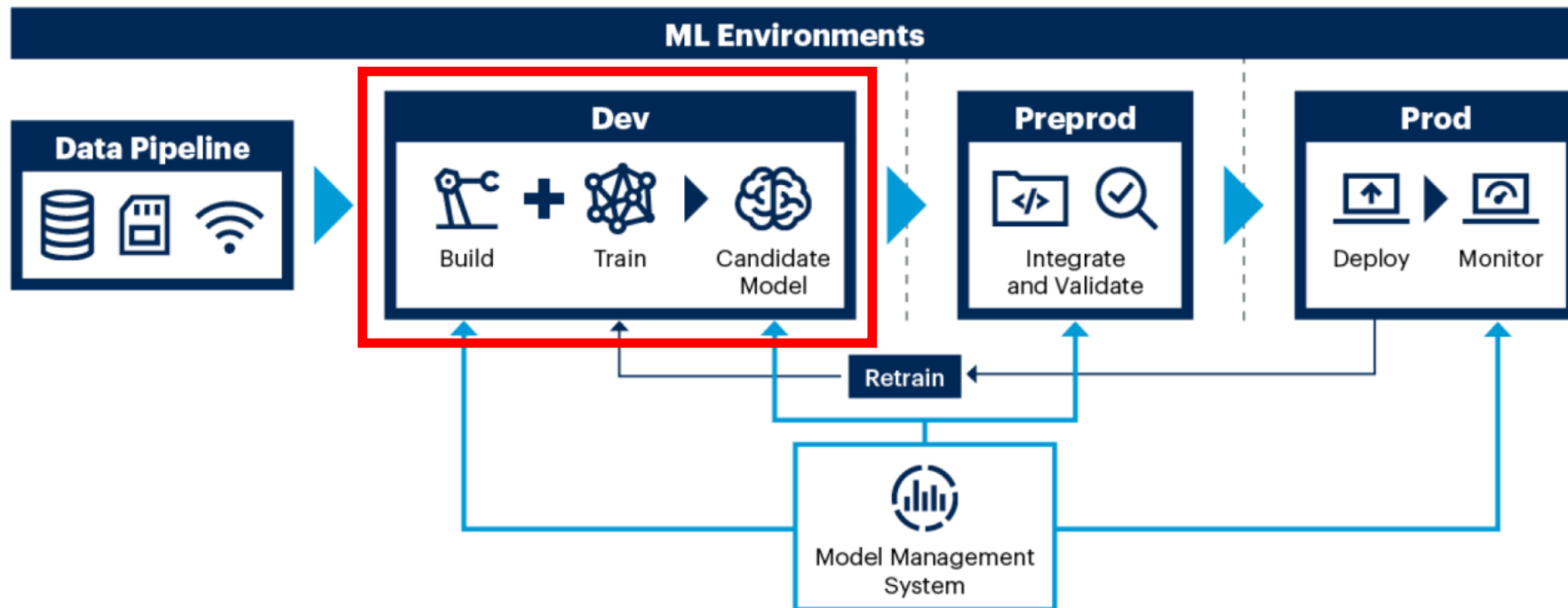


2. Этап работы с данными



3. Обучение и валидация модели

Typical ML Pipeline



Source: Gartner

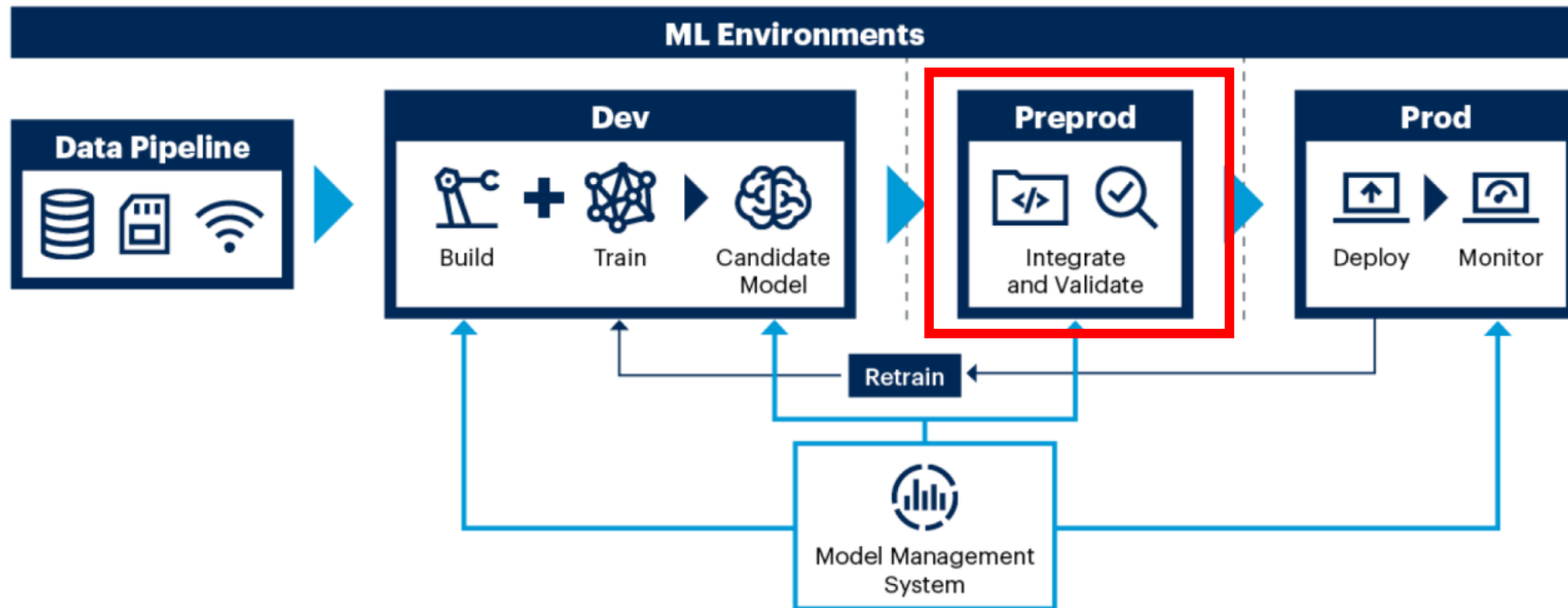
718951_C

3. Обучение и валидация модели

1. *Выбор модели* (линейные модели, деревья, бустинги, нейронные сети)
2. *Обучение модели*
3. *Валидация модели* (оценка качества модели на тестовых данных)
4. *Подбор гиперпараметров модели*
5. *Выбор наилучшей модели*

4. Тестирование модели

Typical ML Pipeline

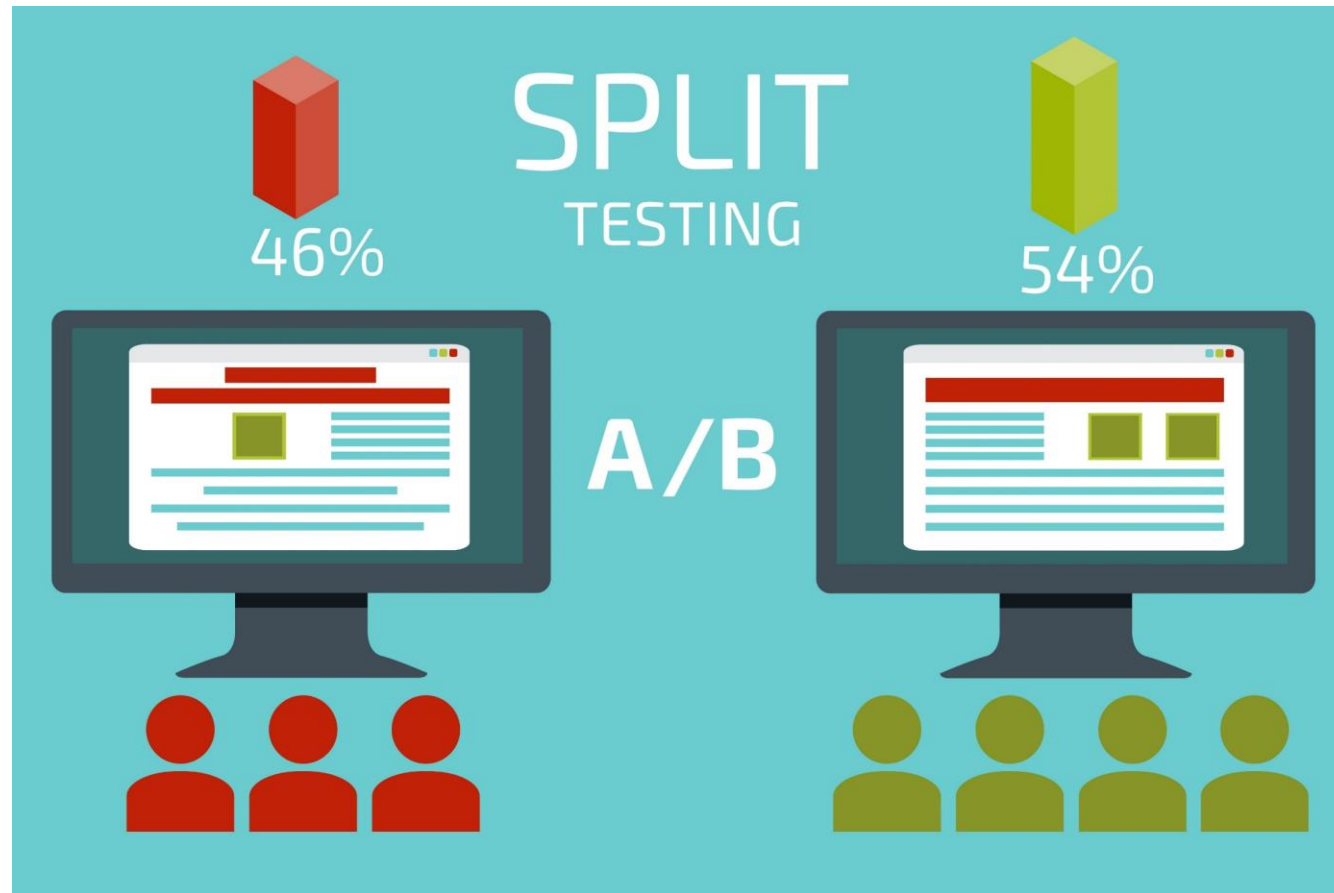


Source: Gartner

718951_C

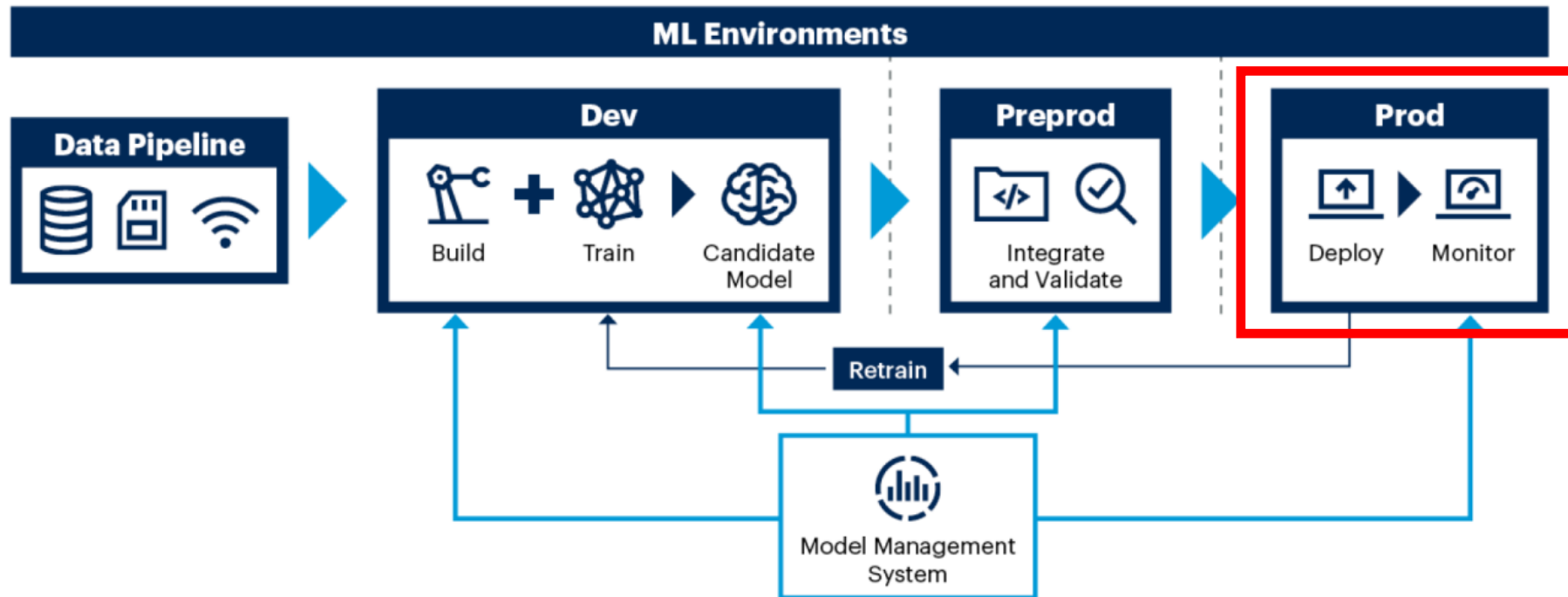
4. Тестирование модели

А/В-тестирование модели на новых пользователях



5. Внедрение модели и мониторинг

Typical ML Pipeline



Source: Gartner

718951_C

5. Внедрение модели и мониторинг

Внедрение модели:

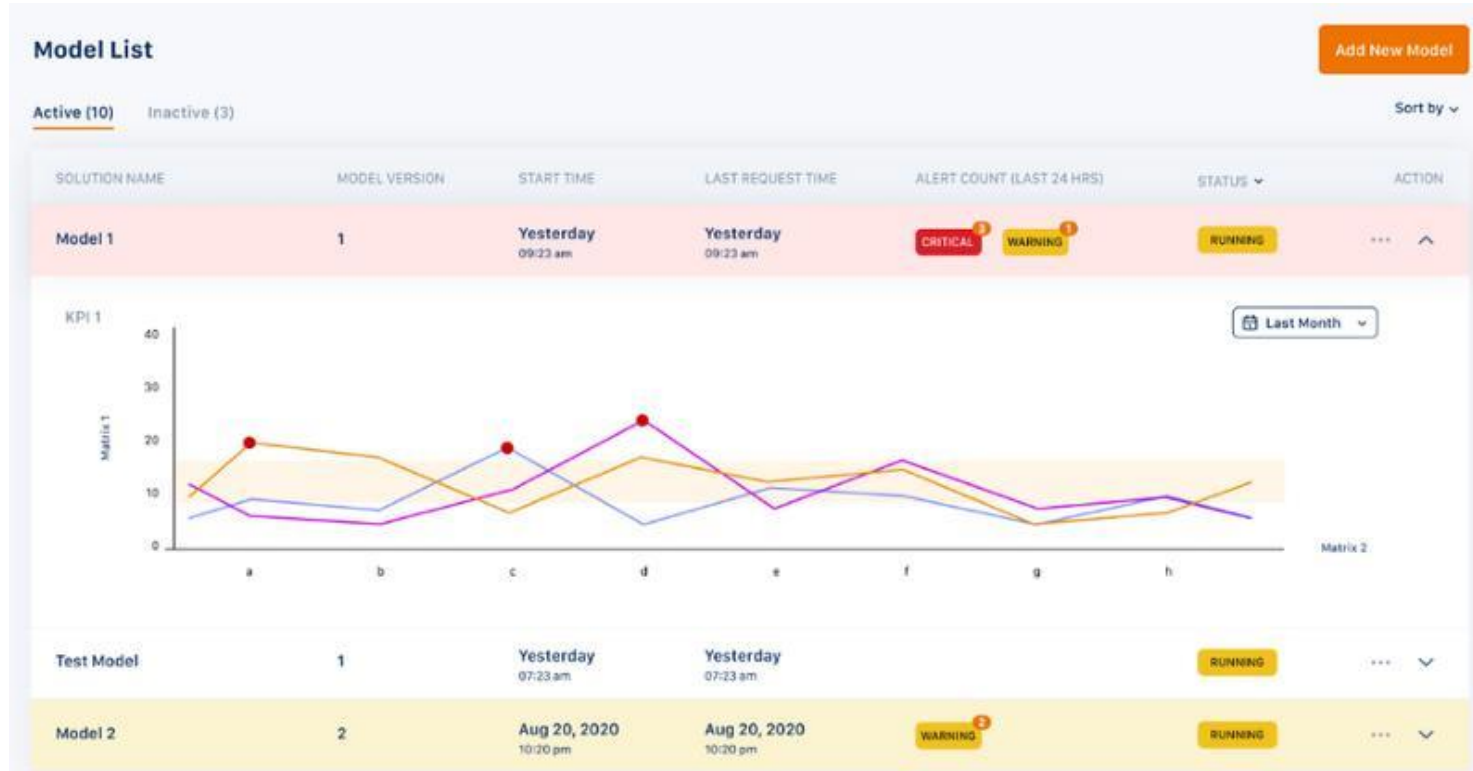
в зависимости от бизнес-целей это может быть

- сервис, применяющий модель ([пример](#))
- телеграм-бот с моделью
- использование специальных serving-инструментов (например, [Seldon](#))

5. Внедрение модели и мониторинг

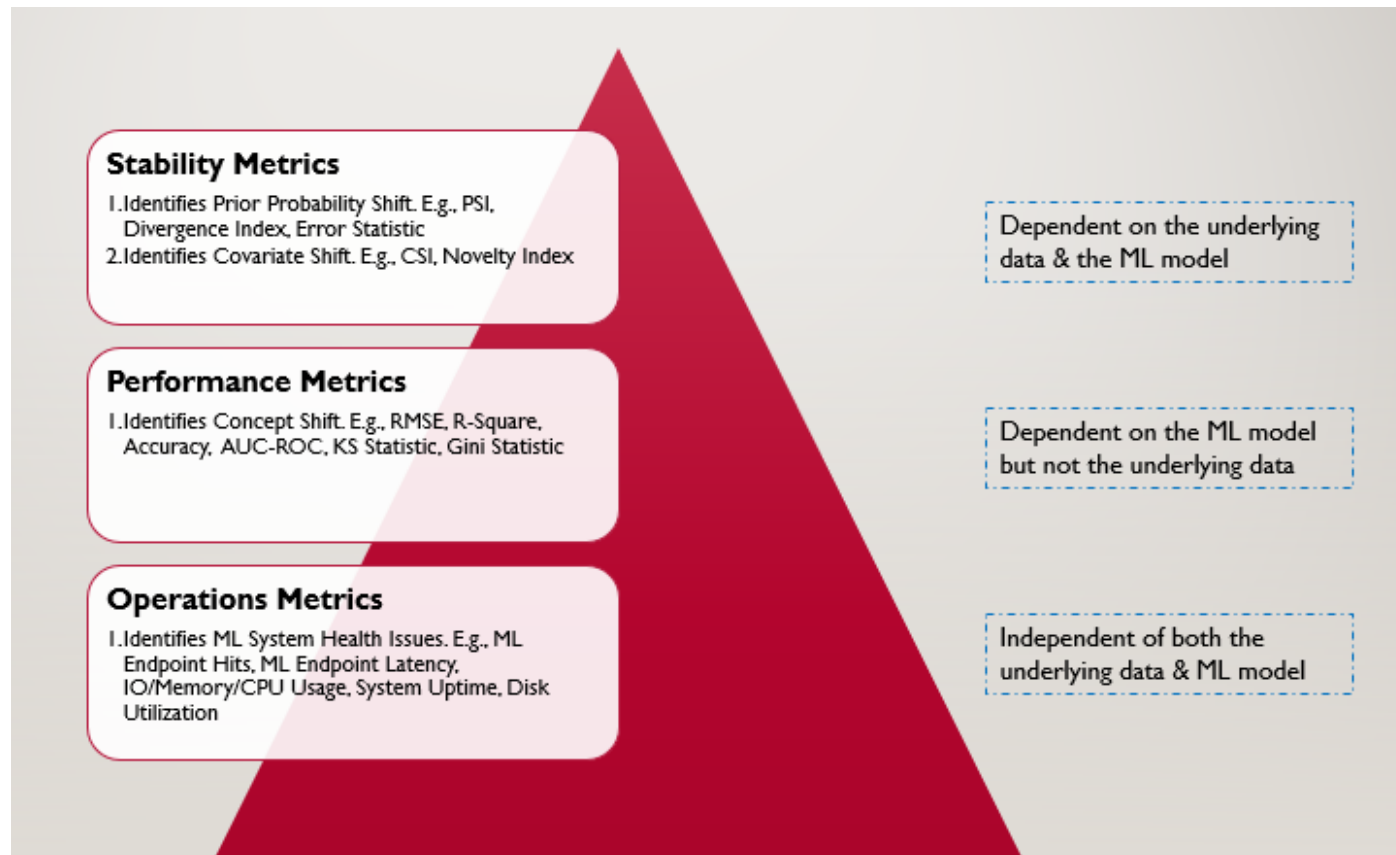
Мониторинг модели

- Цель состоит в том, чтобы отслеживать модели по различным метрикам

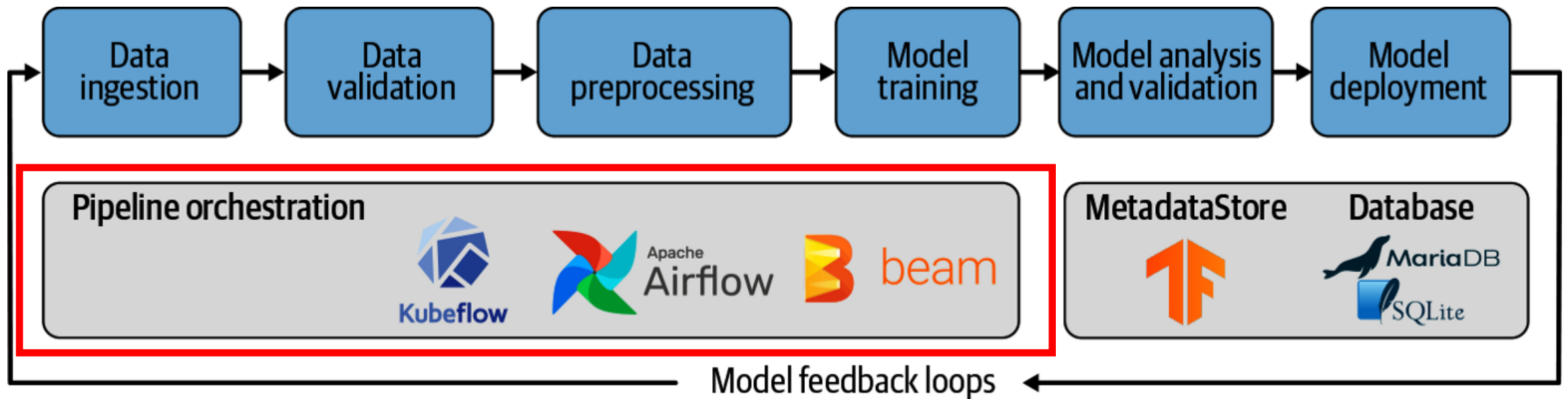


5. Внедрение модели и мониторинг


Мониторинг модели – какие показатели измеряем:



6. Оркестрация процессов



Практика: решение задачи определения ОТТОКОВЫХ КЛИЕНТОВ



Оркестрация пайплайна



Оркестрация

Будем использовать инструмент Apache Airflow.

С его помощью можно:

- Запланировать регулярные запуски пайплайна
- Оценивать успешность выполнения шагов пайплайна и их время



Планирование времени запуска

- Регулярный запуск пайплайна осуществляется при помощи Cron.
- Формат времени в Cron:



[перевод времени онлайн в Cron](#)