# Математика для Data Science

Владимир Анатольевич Судаков

доктор технических наук,

Профессор кафедры 806 МАИ

sudakov@ws-dss.com

Telegram: @vladimir_255

# Содержание курса

1. Цели и задачи Data Science
2. Математические основы
3. Манипулирование данными
4. Оценки и ранги
5. Статистический анализ
6. Визуализация данных
7. Математические модели
8. Линейная алгебра
9. Линейная и логистическая регрессии
10. Методы измерения расстояний и сетей
11. Машинное обучение с учителем и без учителя
12. Методы машинного обучения с подкреплением
13. Вероятностные графовые модели

# Литература


- https://github.com/sudakov/math-for-ds
- Steven S. Skiena. The Data Science. Design Manual
- Джоэл Грас. Data Science. Наука о данных с нуля
- Венцель Е.С. Теория вероятностей
- Пугачев В.С. Теория вероятностей и математическая статистика
- Triola, Mario F. Elementary statistics
- Олег Ларичев. Теория и методы принятия решений, а также Хроника событий в Волшебных странах.
- Стюарт Рассел, Питер Норвиг. Искусственный интеллект: современный подход (AIMA-2)
- Аллен Б. Дауни. Изучение сложных систем с помощью Python

# Правила

- Решения публиковать на github. Присылать мне ссылку на почту или в telegram, каждый раз когда хотите мне их показать.
- Решения присылать за 48 часов до занятий
- Лучше прислать недоделанное решение, чем вообще ничего
- Можно и нужно спрашивать. Лучше если вопрос будет коротким
- У нас будут контрольные, но когда – заранее неизвестно
- Пропустить можно, не более 2-х занятий в семестр (и лекции и практика). Опоздание больше 15 минут – пропуск. Отсутствие ответа на вопрос при онлайн-встрече – пропуск.
- За активность/ответы на занятиях будут дополнительные плюсы.
- Вовремя решенные ДЗ, контрольные без ошибок, и не более 2-х пропусков – автомат на зачете

# Data Science

раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме

Объединяет методы по обработке данных в условиях больших объёмов и высокого уровня параллелизма, статистические методы, методы интеллектуального анализа данных и приложения искусственного интеллекта для работы с данными, а также методы проектирования и разработки баз данных

# Отличия Data Scientist от программистов

- Данные против ориентированности на метод. Ученые ориентируются на данные, а программисты - на алгоритмы.
- Забота о результатах. Реальных ученых заботят ответы.
- Достоверность. Реальные ученые привычны к идее, что в данных есть ошибки, а программисты - обычно нет: "мусор на входе, мусор на выходе", как способ сказать, что это не моя проблема.
- Точность. Ничто никогда не является совершенно истинным или ложным, в то время как в информатике либо математике все или истинно, или ложно. 8/13 = 0,61538461538

# Вопрос для обсуждения

Многие методы Data Science и Machine Learning появились достаточно давно - 50-70 года прошлого века, но активно использоваться в бизнесе начали только сейчас

С чем это связано?

Что такого случилось?

Что есть сейчас и чего не было тогда?

# Искусственный интеллект

**ГОСТ Р 59277— 2020:**

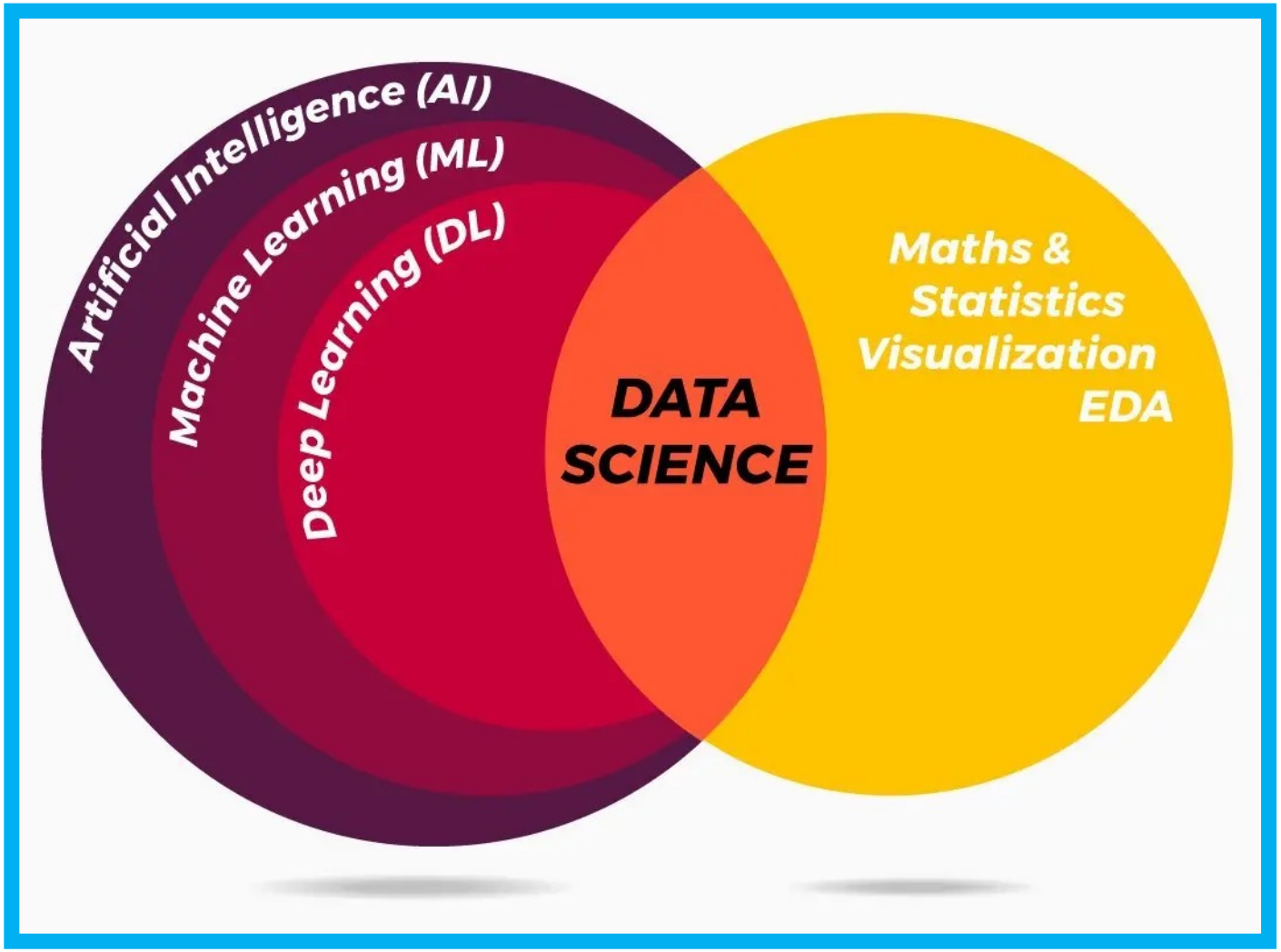
Искусственный интеллект (artificial intelligence):

комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение, поиск решений без за ранее заданного алгоритма и достижение инсайта) и получать при выполнении конкретных практически значимых задач обработки данных результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека

# Машинное обучение

- Машинное обучение (англ. machine learning, ML) — это исследование компьютерных алгоритмов, которые автоматически улучшаются благодаря опыту и использованию данных
- Алгоритмы машинного обучения создают модель на основе выборочных данных, известных как «обучающие данные», чтобы делать прогнозы или предлагать решения, не будучи явно запрограммированными на это

# Взаимосвязи между науками



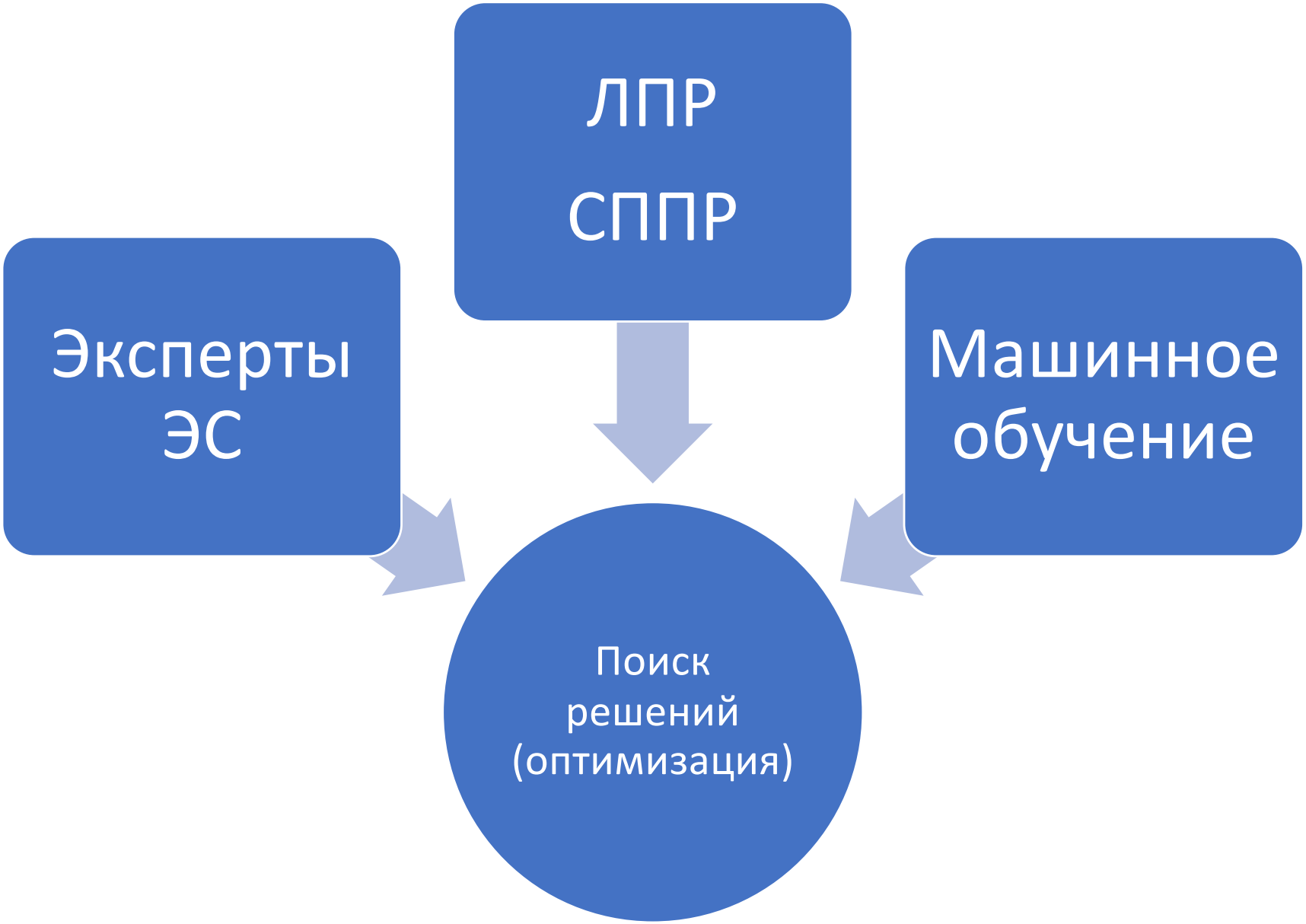
# Давайте подумаем про цели

- Зачем мы учим машины? Например, получить прогноз продаж
- Какие у разработчика моделей?
- А какие цели у бизнеса?

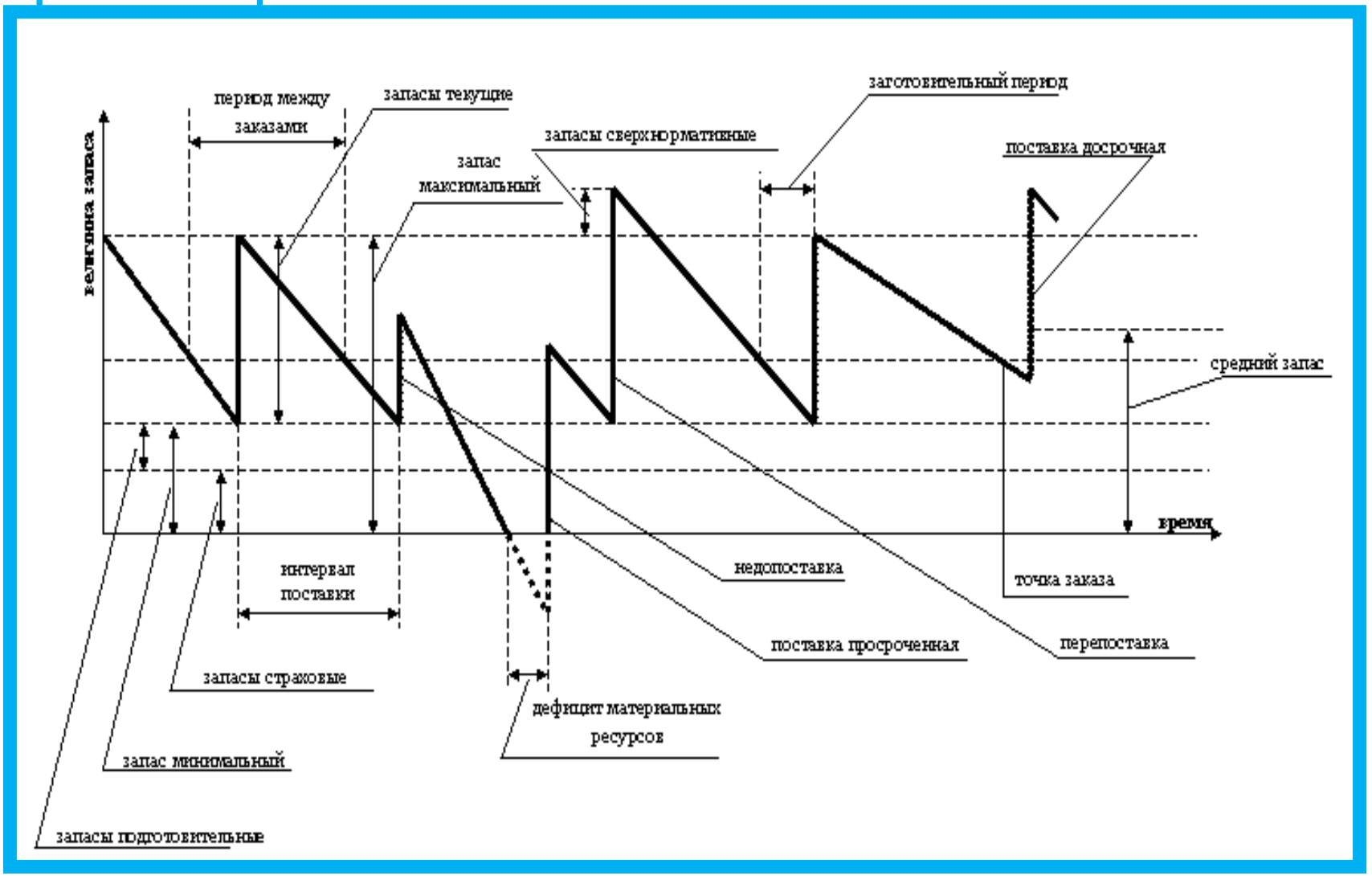
# Цели

- Прибыль (часто важность не велика в краткосрочной перспективе)
- Доля рынка (важность велика)
- Сделать людей счастливыми
- Прославится
- Обеспечить долгую, стабильную жизнь компании
- Не знаю....

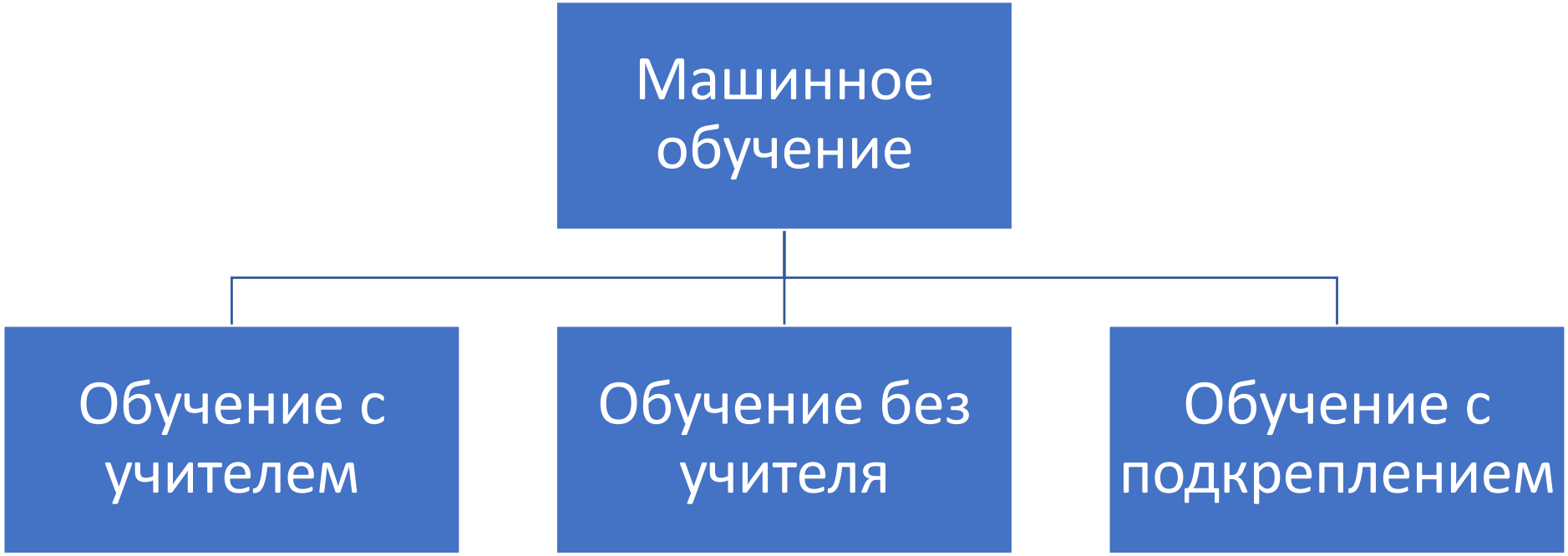
# Подходы ИИ


ЛПР
СППР
Эксперты
ЭС
Машинное
обучение
Поиск
решений
(оптимизация)

# Пример из бизнеса FMCG



# Машинное обучение



# Задачи машинного обучения

- Регрессия
- Классификация
- Ранжирование
- Кластеризация
- Понижение размерности

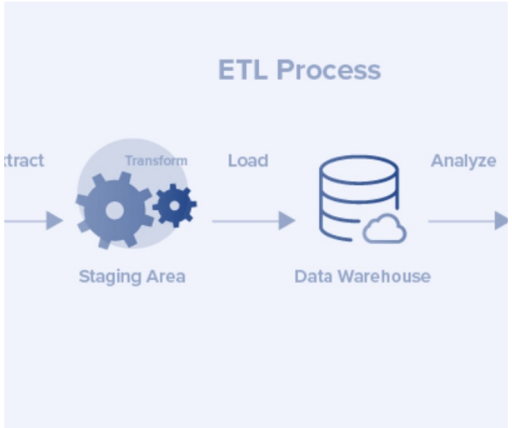
# Где брать данные

- kaggle.com
- T-100 Domestic Segment Data
- opensky-network.org
- …..

# Итоги

Нужно справиться с тремя сложностями:

Критерии и цели не определены



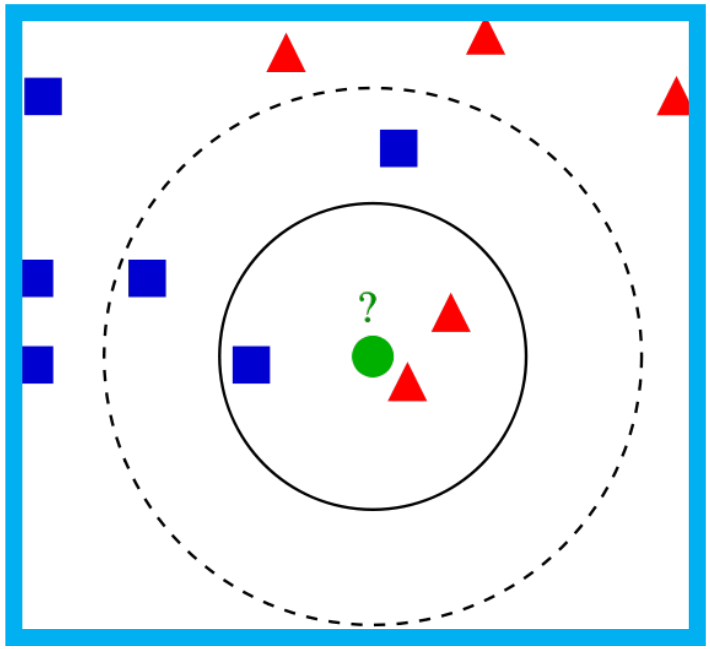

Нет данных

Пространство поиска решений колоссально – даже суперкомпьютер не справится

# Метод ближайших соседей

Правило классификации: объект принадлежит тому же классу что и его k-ближайший соседей.

Близость  определяется в пространстве признаков.



# Задача № 1

- Давайте познакомимся:
  - Что Вы пьете по утрам? Чай или Кофе? Научите модель прогнозировать утренний напиток методом k ближайших соседей.
- Разбиваемся на команды 3-4 человека:
  - Распределение ролей
  - Парное программирование
  - Подготовка исходных данных
  - Тестирование
  - Анализ – какое k лучше?
  - Показ решения
  - Code review чужой бригадой.
- Обсуждение
  - Какое решение лучше и почему?

# Попрактикуемся

- Запишем метрику близости
- Запишем алгоритм