

дисциплина

# Машинное обучение для обработки естественных языков

**Сатмаганбетова Ж.З.**, ст. преподаватель кафедры  
программного обеспечения

---

«Четвёртая технологическая революция строится на вездесущем и мобильном Интернете, искусственном интеллекте и **машинном обучении**» (2016)

*Клаус Мартин Шваб,*  
президент Всемирного экономического форума



Мир наконец поверил в искусственный интеллект.  
Машинное обучение — новый двигатель прогресса.  
Машинное обучение изменит мир. И уже меняет!

# Машинное обучение

Превращение данных в знания

Машинное обучение – это наука о том, как научить машину самостоятельно решать задачи.

Цель машинного обучения – научить машину (точнее, программу) решать задачу, предъявив ей несколько примеров (с правильными и неправильными решениями).

# Данные и Знания

Таблицы

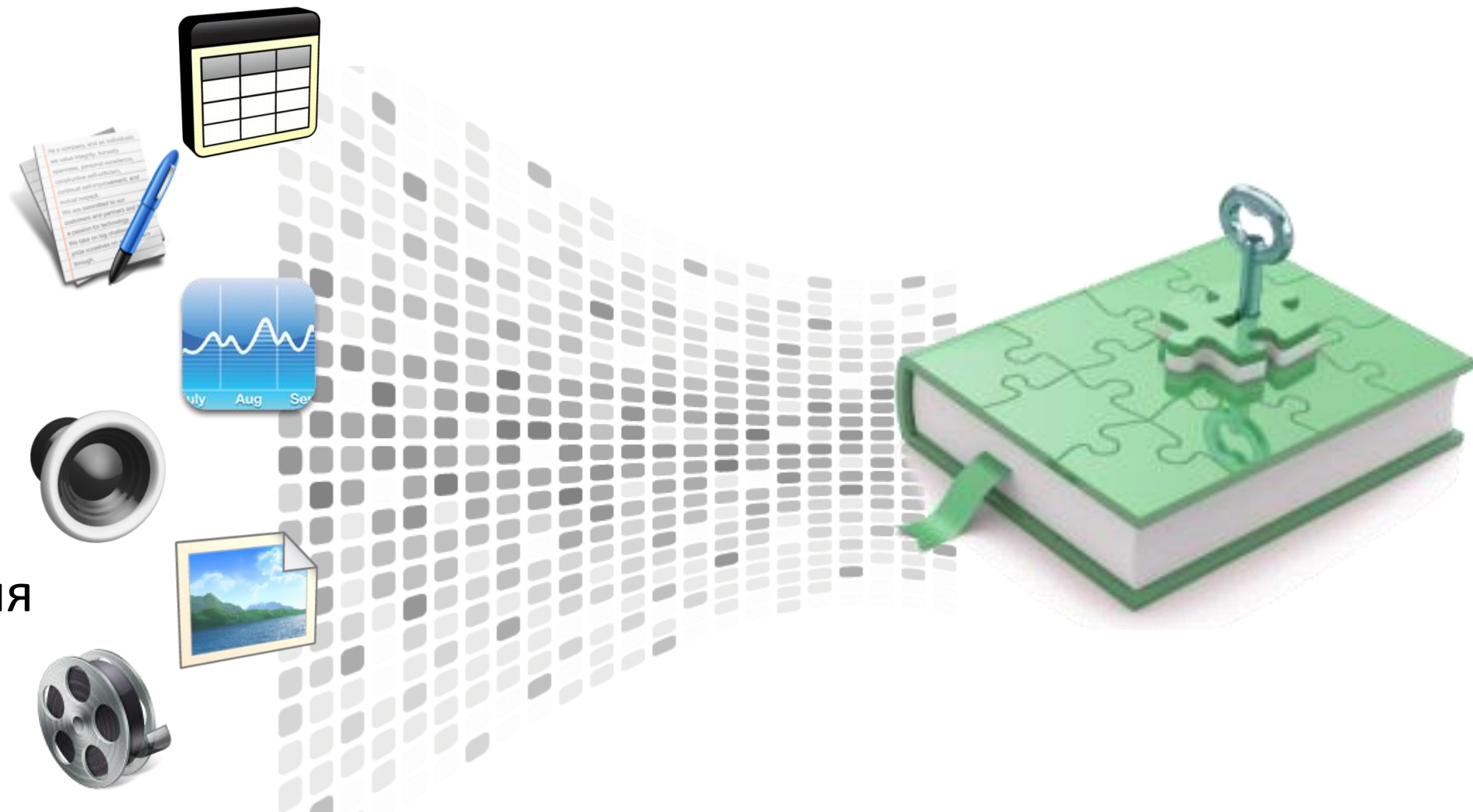
Текст

Сигналы

Звук

Изображения

Видео



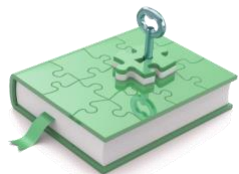
# Данные и Знания



Данные – совокупность зафиксированных фактов



Информация – сведения, уменьшающие неопределённость



Знания – сведения, позволяющие действовать с прогнозируемым результатом

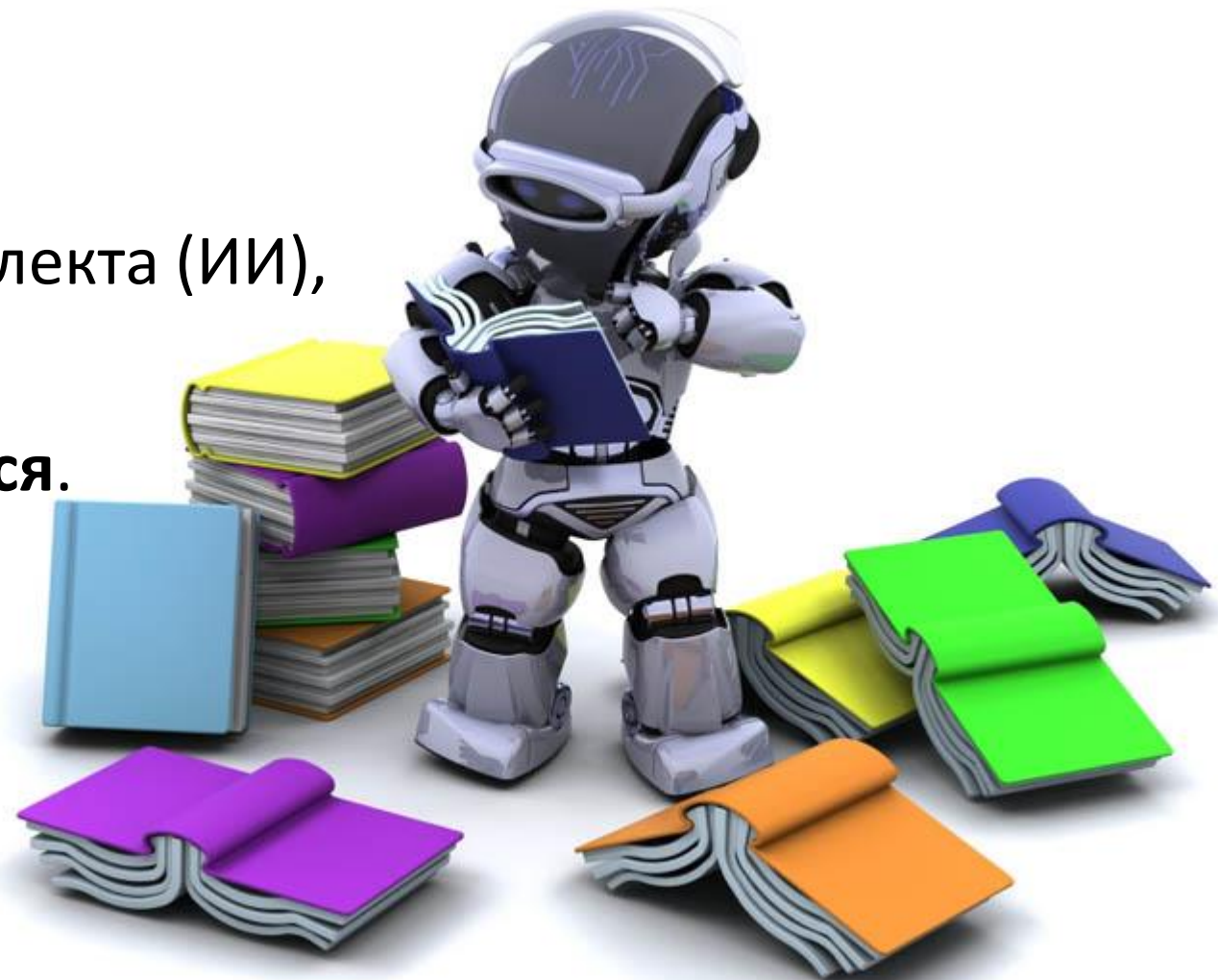
Типичная проблема:

Мы располагаем данными, они хранятся в цифровом виде, но мы не знаем, что в них.

# Что такое машинное обучение?

(англ. Machine learning)

подраздел искусственного интеллекта (ИИ),  
изучающий методы построения  
**алгоритмов, способных обучаться.**



- **Цель и задачи дисциплины**

- Целью дисциплины является изучения основных задач и методов машинного обучения для решения задач автоматической обработки текстов с применением программных систем и инструментов, в которых реализованы данные методы.

- Задачи дисциплины:

- сформировать теоретические знания по основам машинного обучения;
- выработать умения и навыки использования библиотек языка Python для разработки систем машинного обучения для ОЕЯ.

- **При изучении курса студенты должны:**

- ***знать***

- основные уровни обработки текста на ЕЯ и существующие модели анализа текстов;
- - программные системы (приложения) в области АОТ и их архитектурные особенности;
- сущность процедуры разметки и основные стандарты корпусной разметки;
- основные задачи и методы машинного обучения и принципы построения систем машинного обучения для решения задач АОТ;

- ***Уметь***

- навыки использования библиотек языка Python для разработки систем машинного обучения;
- самостоятельно формулировать задачи классификации и кластеризации текстов или предложений;
- пользоваться поисковыми средствами (корпусными менеджерами);

- ***владеть навыками***

- инструментами автоматической обработки естественного языка;
- методикой работы с корпусными базами данных в целях лингвистического анализа;

- ***быть компетентными***

- - использования новых источников материала в лингвистическом исследовании.

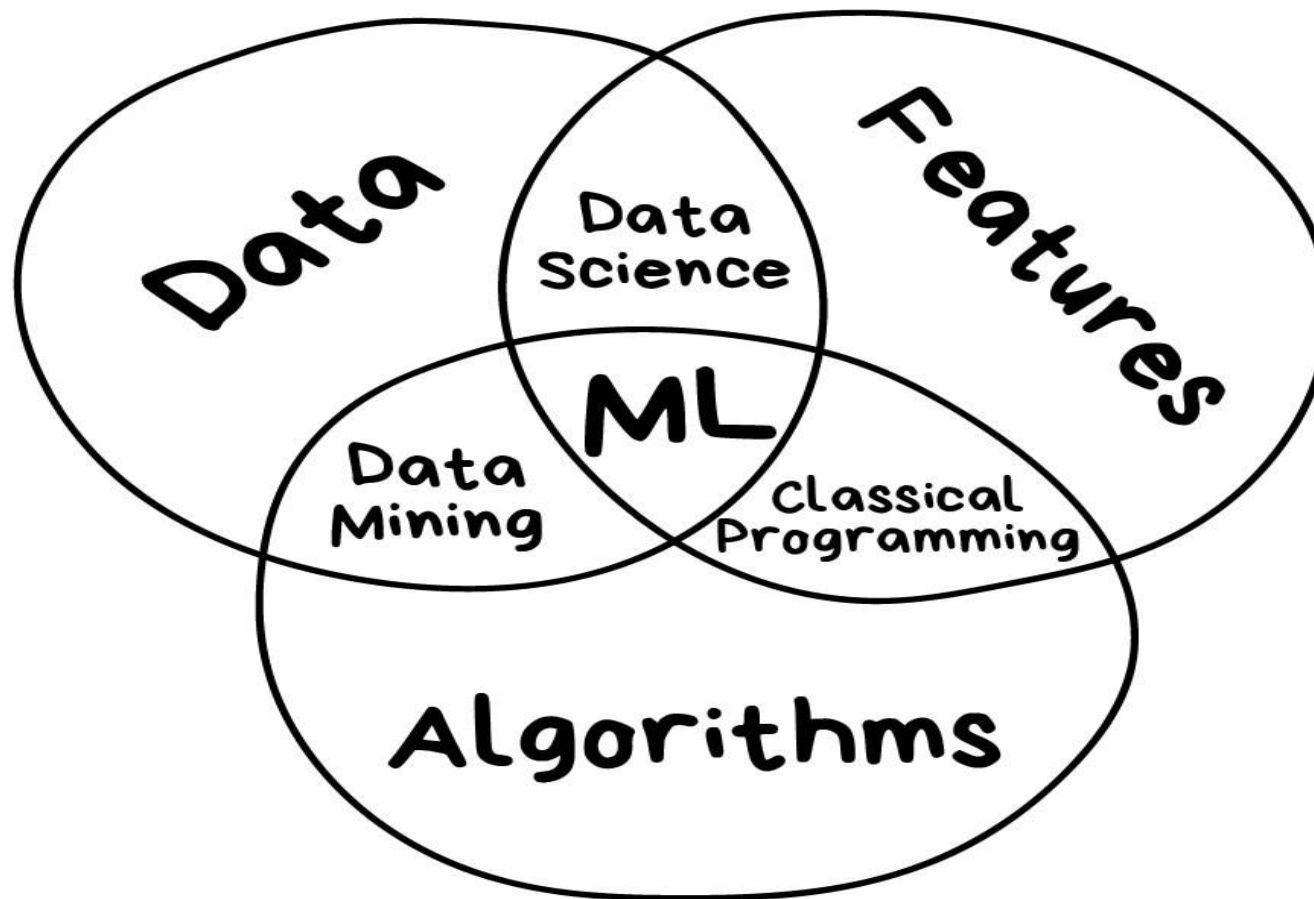
.

# Машинное обучение



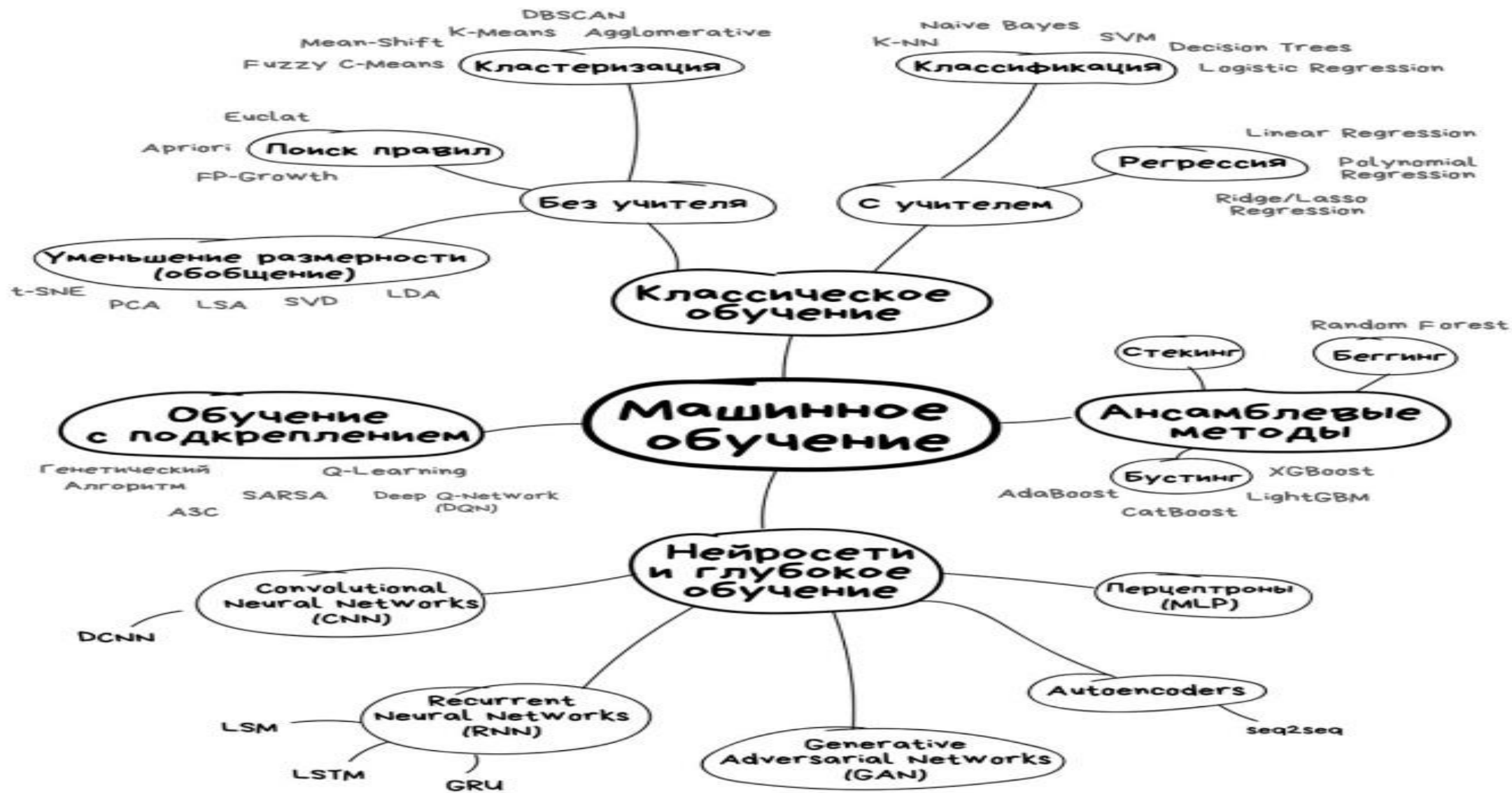


# Три составляющие обучения

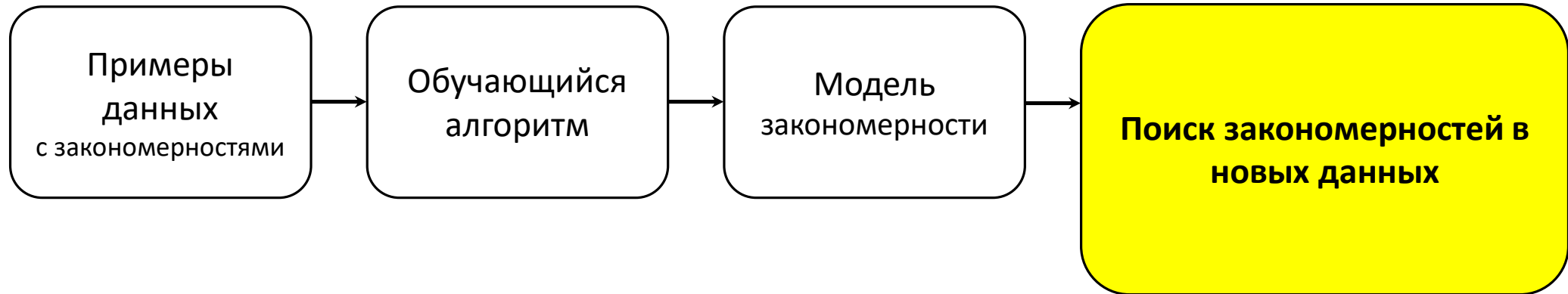


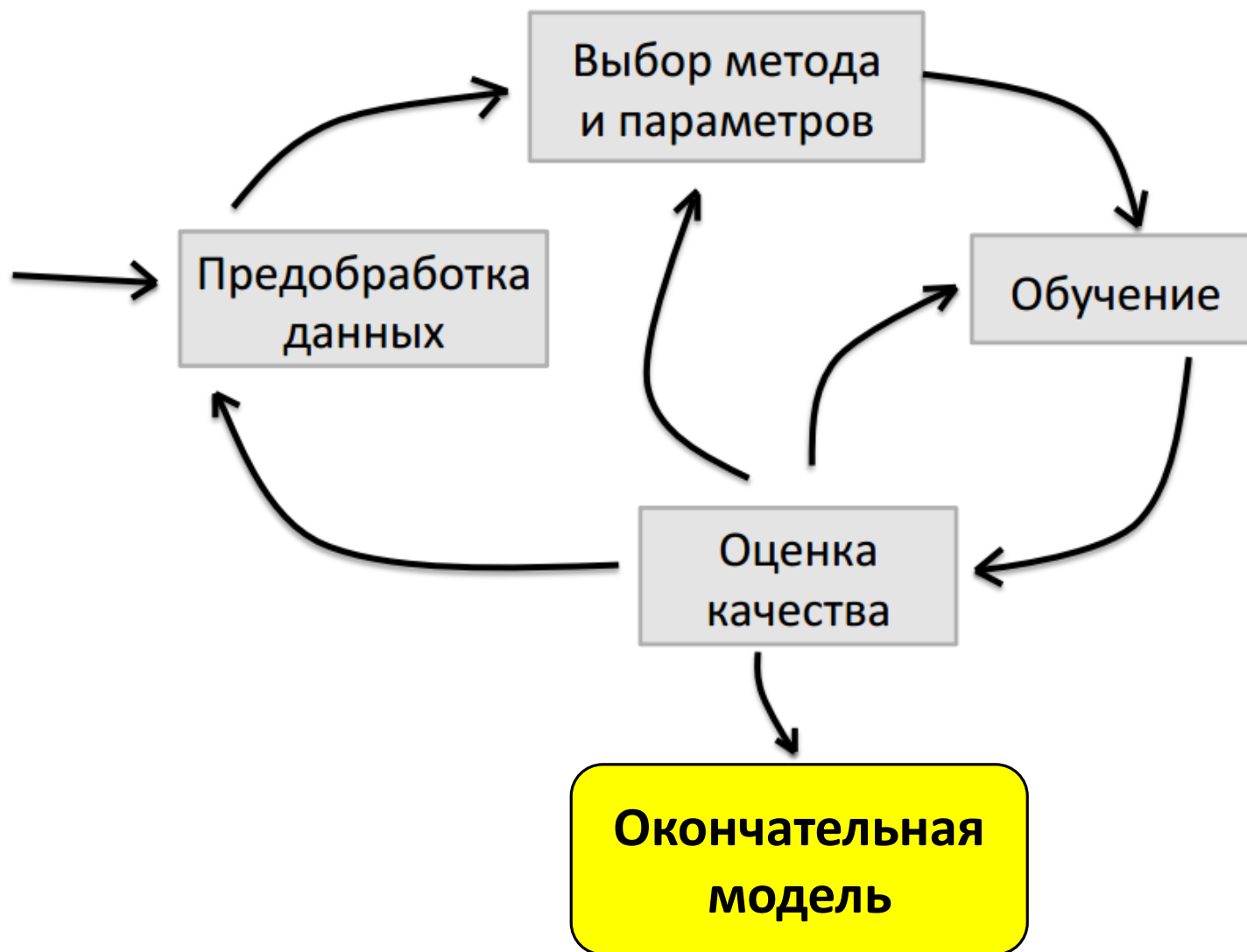
# Обучение vs Интеллект



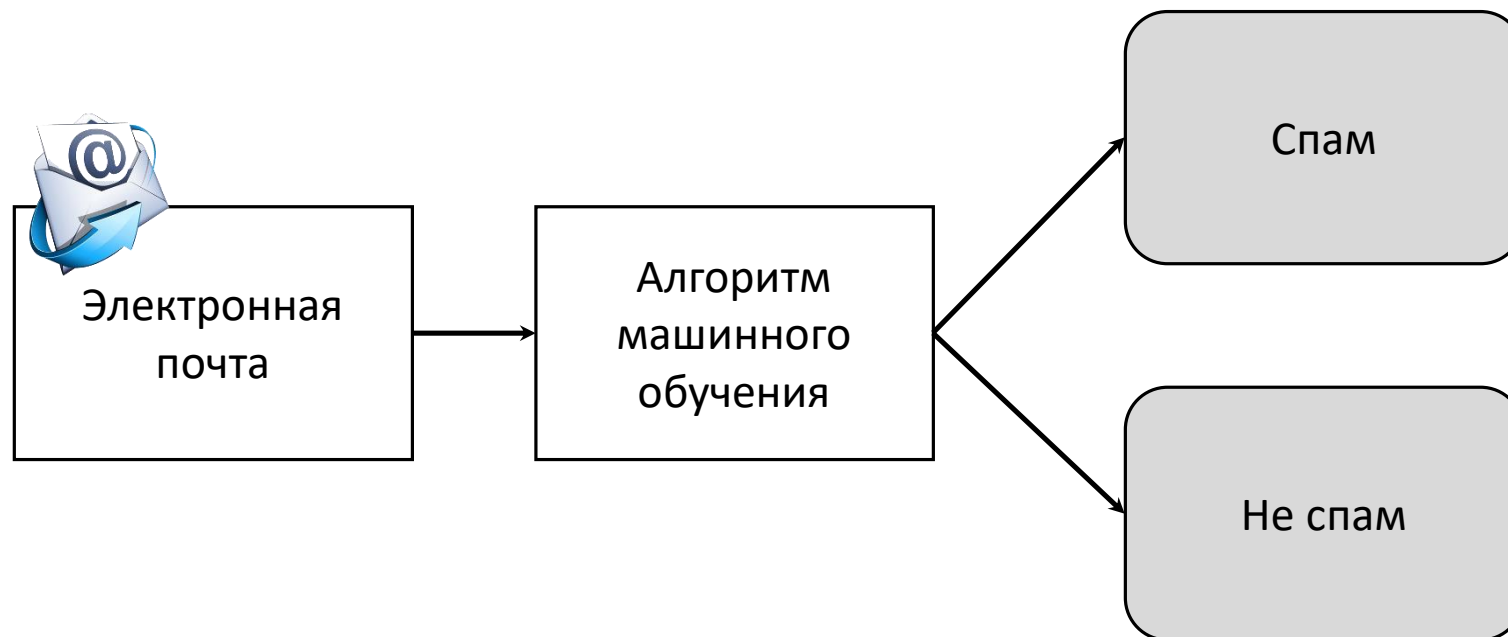


# Машинное обучение



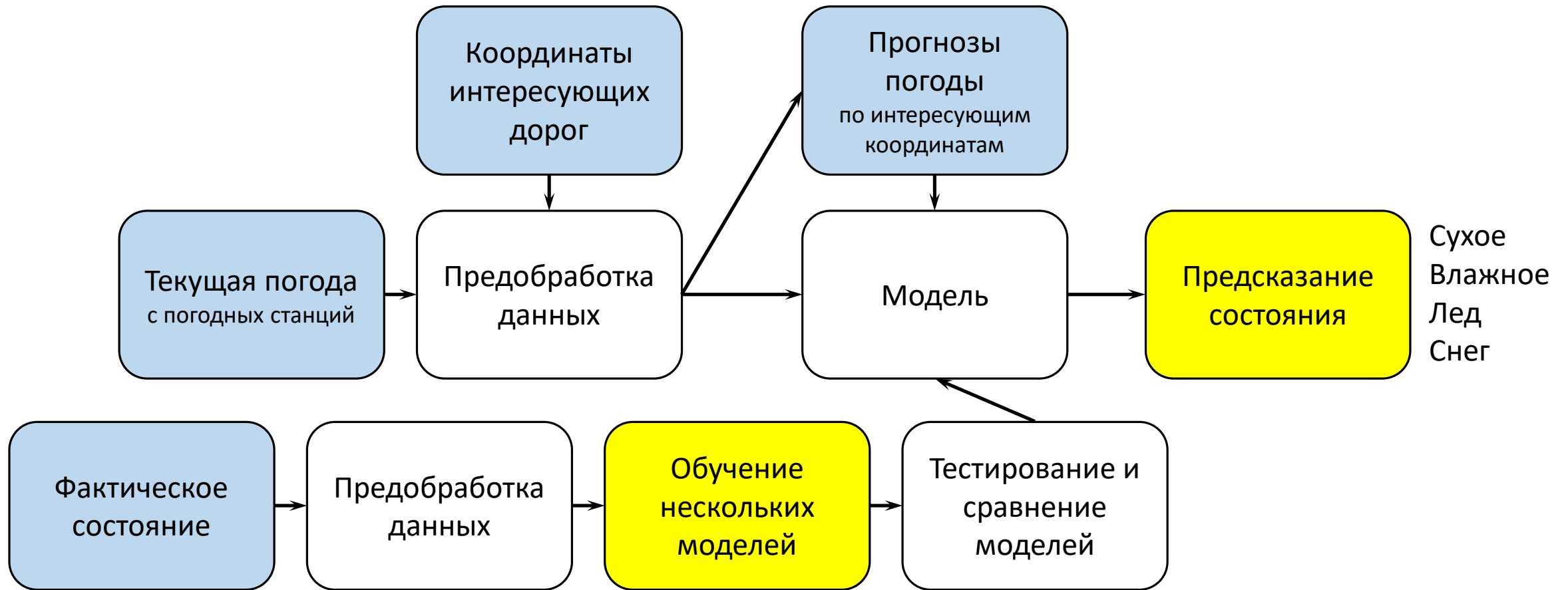


# Пример из жизни

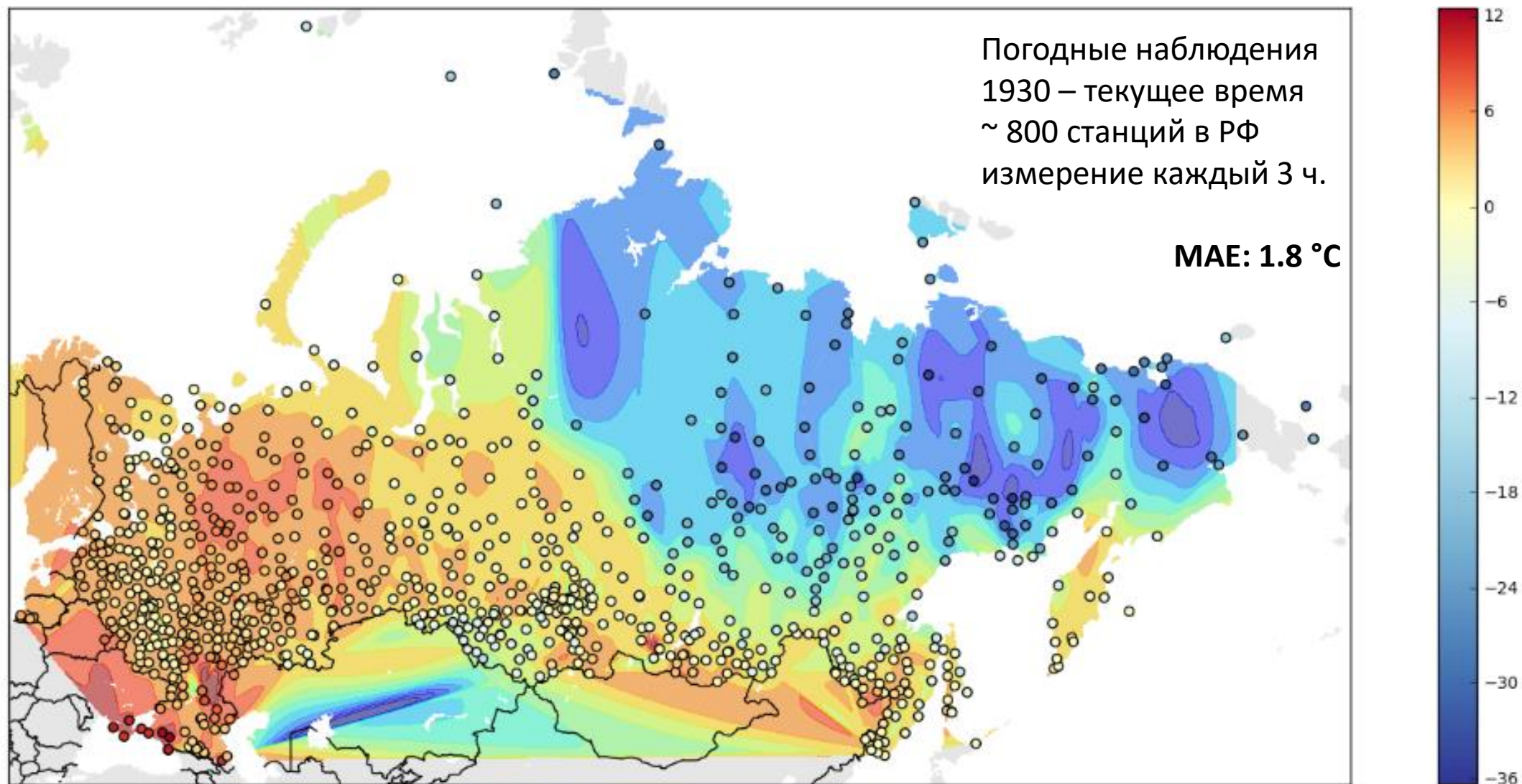


# Пример из жизни

Прогнозирование состояния дорожного полотна на 6 и 24 часа



# Пример из жизни





## Формирование документации

### Проблемы

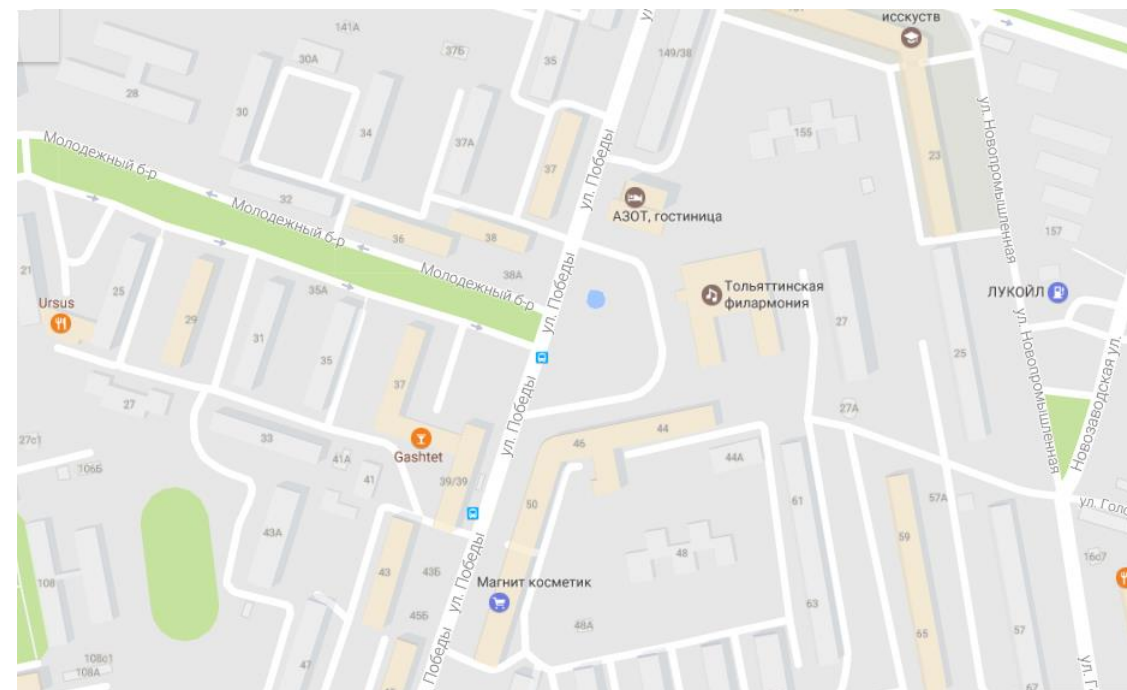
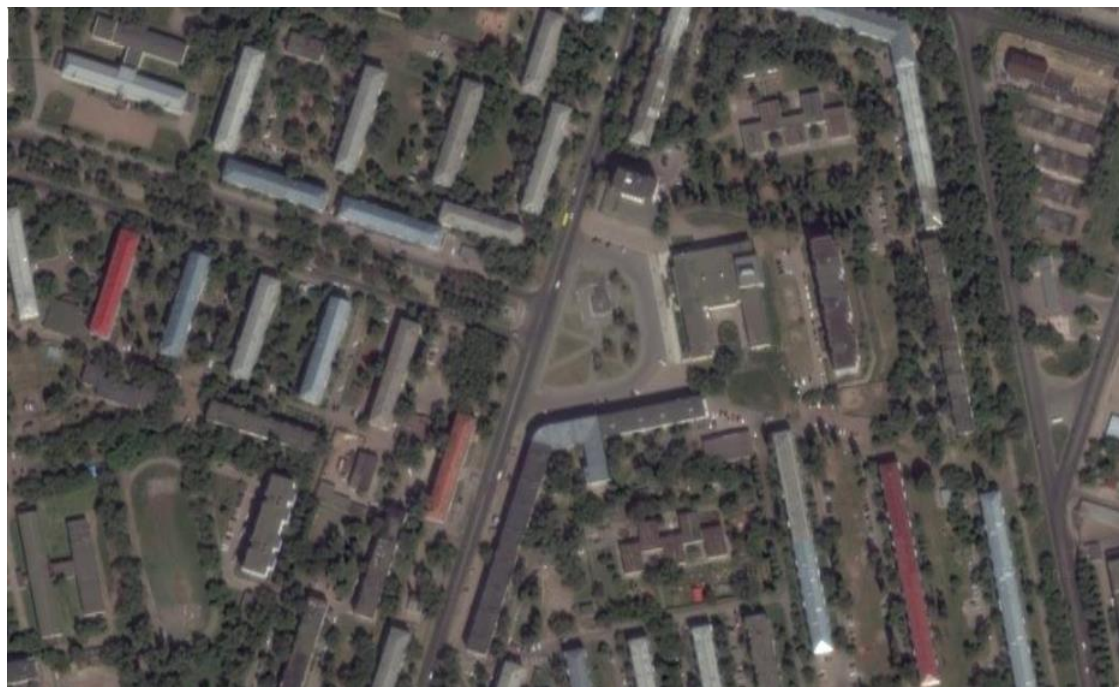
- Генерация строк требует обучающей базы >1млн. записей
- Помнит и хранит контекст до 100 слов

### Достижения

- Разбиение текста на значимые единицы
- Выходной контроль

# Пример из жизни

## Сегментация космических снимков



# Пример из жизни

## Сегментация космических снимков

Космические снимки

- ортогональные 4х-канальные (RGB+NIR)
- с разрешением 0,5-1м на пиксель

Обрабатываемые типы объектов:

- дороги
- постройки
- деревья (лес)
- травяное(растительное) покрытие
- земля (пашня)
- вода
- железные дороги