

# Машинное обучение в бизнесе

**Дмитрий Колодезев, ООО Промсофт  
CompTech @2022**

# Здравствуйте

- Дмитрий Колодезев
- Работаю в Промсофте
- Анализирую данные
- Консультирую и учу
- Собираю датазавтраки по вторникам
- Участвую в работе OpenDataScienceLab

# План лекции

- Что такое «Машинное обучение»
- Промышленный ML vs Kaggle
- ML-разработка vs «Обычная» разработка
- Перед употреблением взбалтывать
- Baseline от края до края
- Допиливаем и доразмечаем
- Бухгалтерия, этика и здравый смысл

# Искусственный интеллект

... Способность приобретать, обрабатывать, создавать и применять знания, определенные в форме модели, для выполнения одной или нескольких поставленных задач; ...

ПНСТ 553-2021

# Машинное обучение



Machine learning is the science of getting computers to act without being explicitly programmed \*



Software 2.0 is code written by the optimization based on an evaluation criterion (such as “classify this training data correctly”) \*



Машинное обучение -  
автоматизированное построение удобной  
проекции признакового пространства

# Предположения ML

Машиное обучение — это автоматизированный подход к **выявлению сложных шаблонов** в **имеющихся данных** и использование этих шаблонов для **предсказаний** на **новых данных**. [cs329s](#)

- Мы можем выявить шаблоны
- Шаблоны сложные
- Данные имеются
- Можем предсказывать
- Будут новые данные

# Просто о сложном

**Родители: «Если все твои друзья прыгнут с крыши, ты тоже прыгнешь?»**

**Ребёнок: «Нет!»**

**Алгоритм машинного обучения: «Да!»**

# ML — про ЭКОНОМИЮ

Почти всегда:

Быстро принимаем  
не очень важные решения  
среднего качества  
дорого и неконтролируемо

# Просто, быстро, много

## Люди

- Медленные
- Дешево начать
- Не масштабируются
- Адаптируются
- Качество разное
- Издержки

## ML-алгоритмы

- Быстрые
- Дорого начать
- Масштабируются
- Не адаптируются
- Качество среднее
- Капитальные вложения

# А оно нам надо?

## THE DATA SCIENCE **HIERARCHY OF NEEDS**

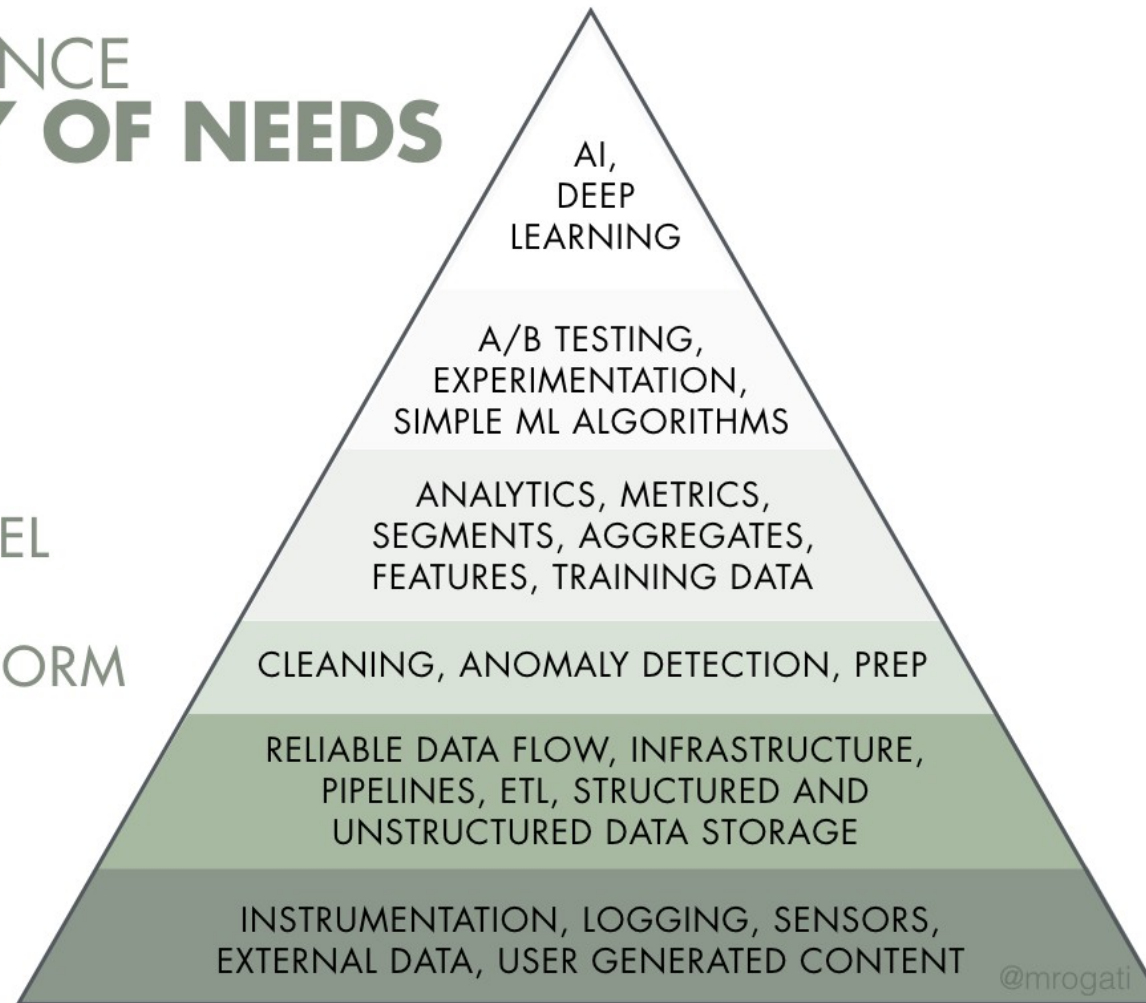
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT



# Примерно как на Kaggle

- На Kaggle: получить наибольшее качество на отложенной выборке
- В бизнесе: решить проблемы
  - заинтересованных лиц
  - имеющимися силами
  - в срок
  - в рамках имеющегося бюджета
  - в условиях конфликта интересов

# Как в Kaggle, правда?

- Можем ли мы улучшить текущий результат?
- Оправдано ли это экономически?
- Является ли это узким местом?
- Будут ли доступны данные вовремя?
- Сможем ли мы использовать предсказания?
- Сможем ли мы внедрить?
- Управление изменениями без C\*\* не работает

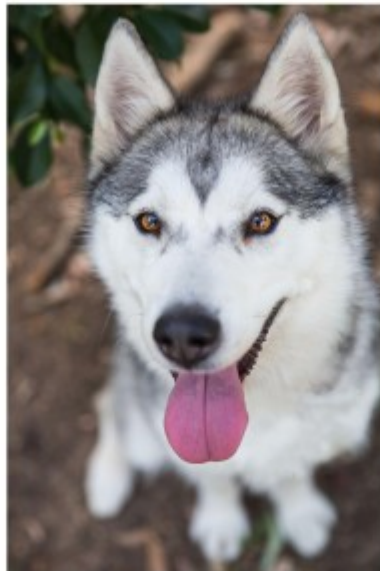
# Правда же?

## Stakeholder objectives

**ML team**  
highest accuracy



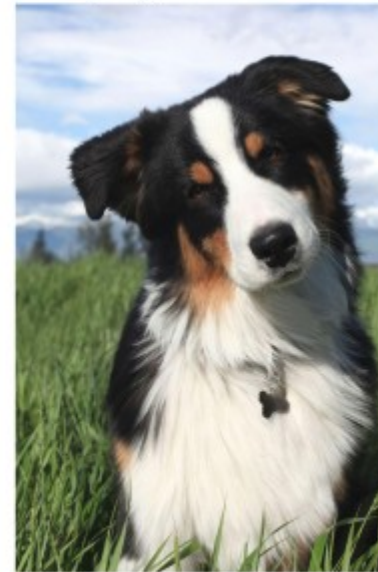
**Sales**  
sells more ads



**Product**  
fastest inference



**Manager**  
maximizes profit  
= laying off ML teams



# Иногда нас просто любят

- Большому бизнесу может быть удобно иметь в отчете строчку про AI-трансформацию.
- Растущему стартапу нужно очаровать инвесторов своей прорывной прорывностью.
- Друг моего знакомого рассказывал своей подруге, что так тоже бывает.

# Обычная разработка ПО

- DS в бизнесе = ML инженер
- ML инженер = ML разработчик
- ML разработчик = python разработчик
- Тут нет ничего особенного
- Ну почти нет
- Кроме трех маленьких деталей:
  - Данные
  - Управление проектами
  - Мониторинг

# Данные

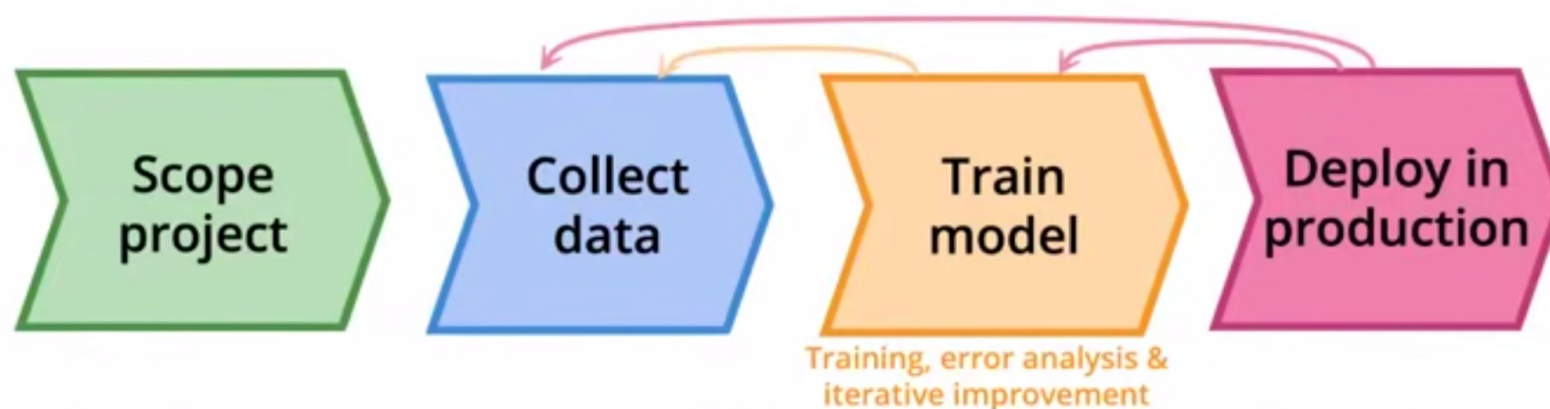
- ML — модель сцеплена с данными
- Данные **обычно** не версионировуются
  - Версионировуются схемы данных, но не данные
- Данные **обычно** не обозримы
  - Какая точка данных портит вашу модель?
  - Данных **обычно** больше, чем кода
  - Diff **обычно** не работает
- Предположения о данных нужно проверять
  - Нужно, но трудно

# Вот, например, красивое



# Данные важны

## Train model: Speech Recognition



Error analysis shows your algorithm does poorly in speech with car noise in the background. What do you do?

### Model-centric view

How can I tune the model architecture to improve performance?

### Data-centric view

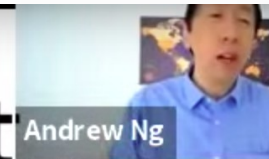
How can I modify my data (new examples, data augmentation, labeling, etc.) to improve performance?

Andrew Ng

# Данные важнее моделей

## Improving the code vs. the data

|               | Steel defect detection | Solar panel        |
|---------------|------------------------|--------------------|
| Baseline      | 76.2%                  | 75.68%             |
| Model-centric | +0%<br>(76.2%)         | +0.04%<br>(75.72%) |
| Data-centric  | +16.9%<br>(93.1%)      | +3.06%<br>(78.74%) |



Andrew Ng

# Управление проектом

- Фаза Discovery, time-boxed
  - Копаем отсюда и до обеда
- Фаза Delivery, спринты
  - Нормальный проект по разработке ПО
  - Как будто бы у вас каждую неделю меняют схему баз данных
  - И раз в полгода меняют движок баз данных
  - В остальном все то же самое

# МОНИТОРИНГ

- Вы уверены, что оно работает так же хорошо, как вчера?
- Вы уверены, что оно работает хорошо?
- Вы уверены, что оно вообще работает?
- Подруга моей знакомой рассказывала своему другу, что это не всегда очевидно.
- Посмотрите, например, [Evidently AI](#)

# Как делать

- Продуктовая гипотеза
- Доступные данные
- Анализ осуществимости
- Критерии оценки
- Базовая модель
- Доработка модели
- Развертывание
- Мониторинг
- Дообучение

# Когда не делать

- Если это неэтично
- Если простое правило решает проблему
- Если данные недоступны
- Если цена ошибки высока
- Если каждое решение должно быть объяснимо
- Если дешевле нанять человека ;-(

# Вроде как правила

- 50% эффекта ML-модели можно достичь без ML, правилами, регулярами и т.д.
- Если эксперт не видит в данных закономерностей, ML тоже не увидит
- Большинство проблем — на границах системы
- Основные затраты — данные
- Если метрика не в рублях, она не важна.

<https://www.youtube.com/watch?v=tIB4CL9Esl0>

# Почитать

- Rules of ml

[http://martin.zinkevich.org/rules\\_of\\_ml/rules\\_of\\_ml.pdf](http://martin.zinkevich.org/rules_of_ml/rules_of_ml.pdf)

- Machine Learning Yearning

[https://github.com/yennlh/ml-yearning/blob/master/Ng\\_MLY01\\_13.pdf](https://github.com/yennlh/ml-yearning/blob/master/Ng_MLY01_13.pdf)

- Evidently AI Blog

<https://evidentlyai.com/blog>

- CS 329S: Machine Learning Systems Design

<https://stanford-cs329s.github.io/2021/syllabus.html>

# Вопросы

Слайды тут



dkolodezev



promsoft



dkolodezev



d\_key



dmitry\_kolodezev

<https://kolodezev.ru/download/slides-mlb-2022.pdf>