



# Лекция 1. Введение



# Специфика ML проектов

- ML это ново модно молодёжно, но
- Сложно прогнозировать качество
- Сложно оценить сроки
- Сложно внедрять

В общем много рисков



# Как подойти к решению этих проблем?



# Этапы ML проекта

- Постановка задачи
- Сбор данных
- Обучение модели
- Внедрение в продакшн
- Постаналитика

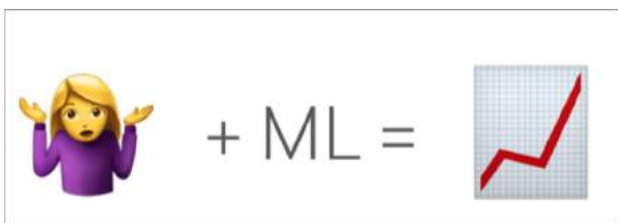


# Постановка задачі



# Постановка задачи

## Ожидание



## Реальность



## Работает



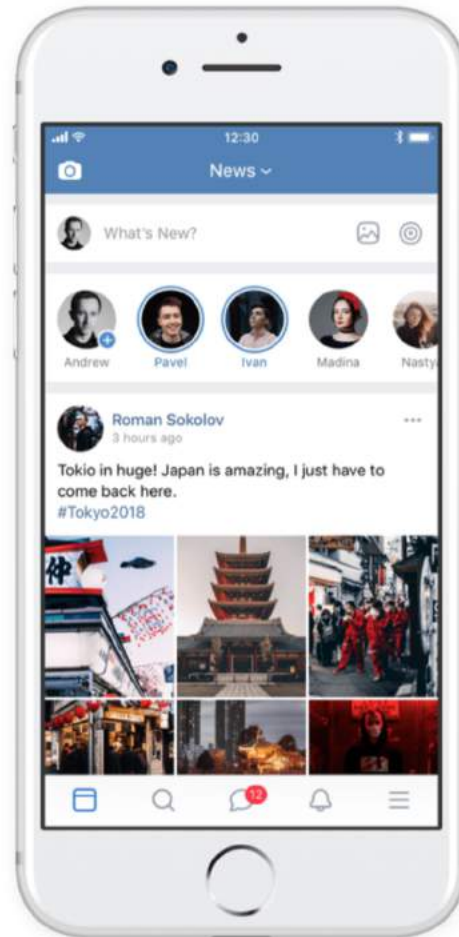
# Постановка задачи

- Какую задачу пользователя решаем?

# Постановка задачи

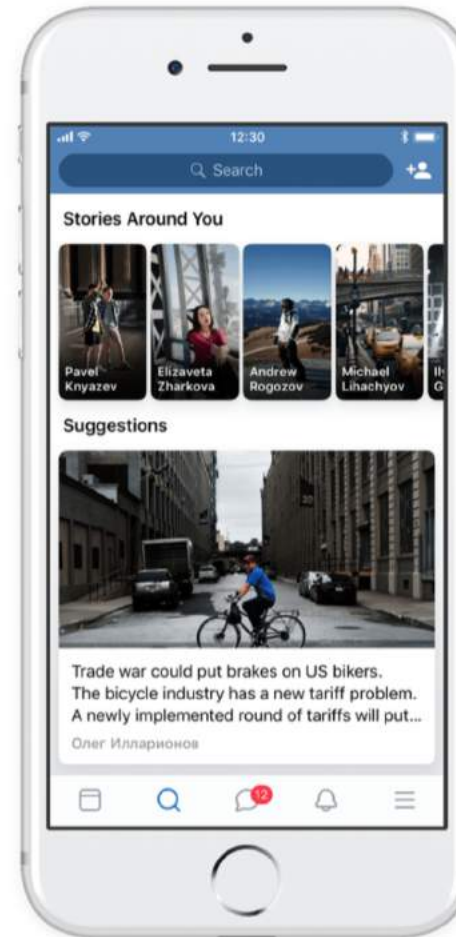
## Выбор среди друзей

- $X$  человекомесяцев
- $Y$  серверов
- $Z$  потраченных нервов



## Выбор среди всех

- $5X$  человекомесяцев
- $3Y$  серверов
- ☹️ потраченных нервов





# Постановка задачи

- Какую задачу пользователя решаем?
- Нужен ли ML?

# Постановка задачи

Холодильники — купи


market.yandex.ru Холодильники — купить на Яндекс.Маркете

← Я ↻ 🔒

🔍 📁 📧 ⚙️ 🔋 ⬇️

## Холодильники

Сортировать: [по популярности](#) [по цене](#) [по рейтингу](#) [по отзывам](#) [по размеру скидки](#) [по новизне](#) ☐ Сначала предложения в моём регионе



**Холодильник SUPRA TRF-030**

**4.5** 4 отзыва

42x40.2x48.5 см  
однокамерный  
без морозильника  
общий объем 30 л


68 человек купили этот товар

**5 400 Р**

17 предложений от 5 301 Р

*Малюсенький, дешевый, легкий эээ... относительно тихий, похож на...*

**Арсений**



**Холодильник Бирюса 50**

**4.5** 4 отзыва

47.2x45x49.2 см  
однокамерный  
класс A+  
без морозильника  
общий объем 46 л


176 человек купили этот товар

**5 400 Р**

44 предложения от 5 370 Р

*Самый дешевый на момент покупки. Спустя год - работает. ...*

**Magikan**



**Холодильник CENTEK CT-1700-47**

44.4x48.5x49.5 см  
однокамерный  
класс A+  
морозильник сверху  
общий объем 47 л

**5 725 Р**

10 предложений от 5 725 Р

### Магазины на карте



Орск

### Категории

- Бытовая техника
- Крупная техника для кухни
- Холодильники, морозильники, винные шкафы
  - Холодильники
  - Все результаты поиска

### Цена, Р

от 5 301 до 959 990

☐ Цена с учётом доставки

☐ Покупка в кредит

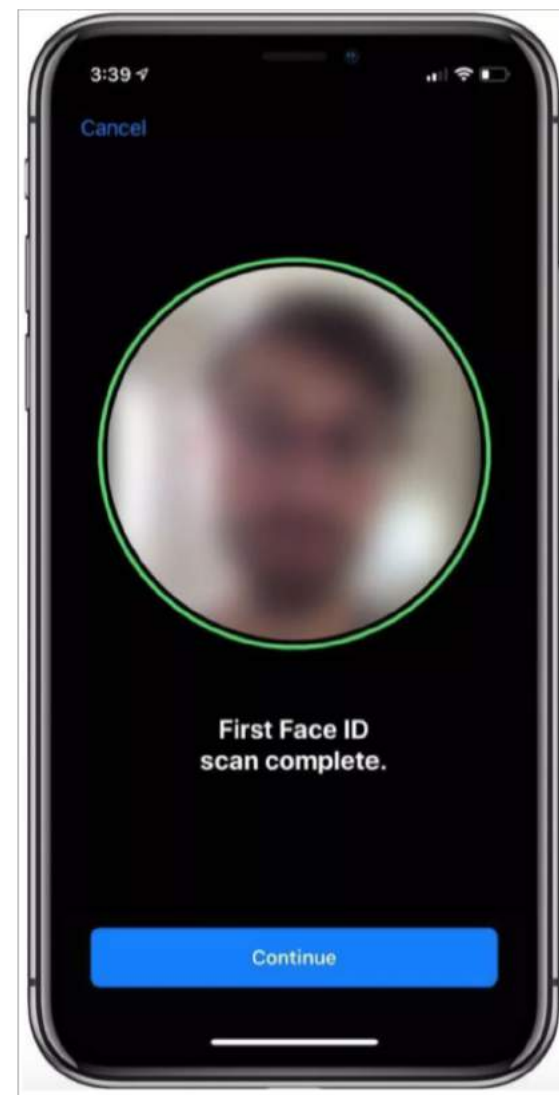
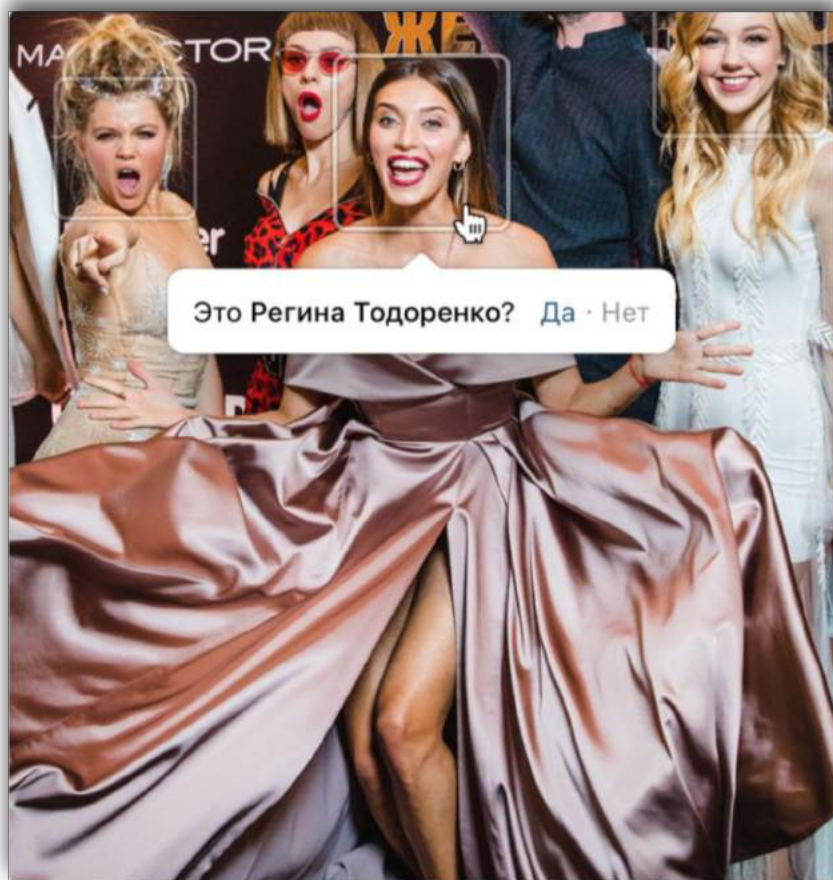
☒ В продаже

### Производитель

# Постановка задачи

- Какую задачу пользователя решаем?
- Нужен ли ML?
- Как выглядят входные данные и как часто меняются?
- Какие требования по скорости работы?

# Постановка задачи



# Постановка задачи

- Какую задачу пользователя решаем?
- Нужен ли ML?
- Как выглядят входные данные и как часто меняются?
- Какие требования по скорости работы?
- Что будет если ошибаемся?

# Постановка задачі



Цена ошибки



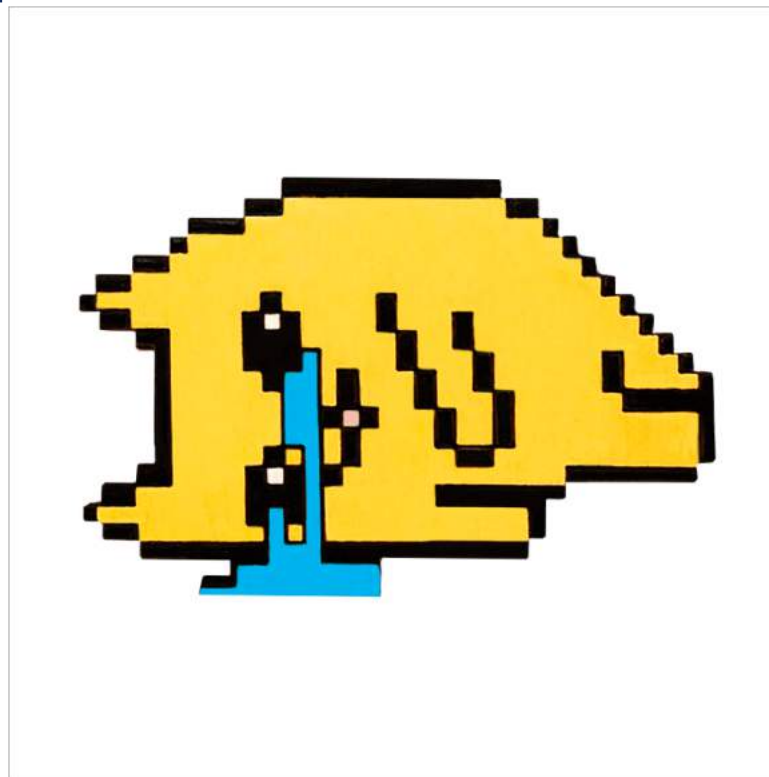
# Постановка задачи

А ещё нужно думать о смежных задачах

Выбрать правильный KPI

И представлять сроки

И риски

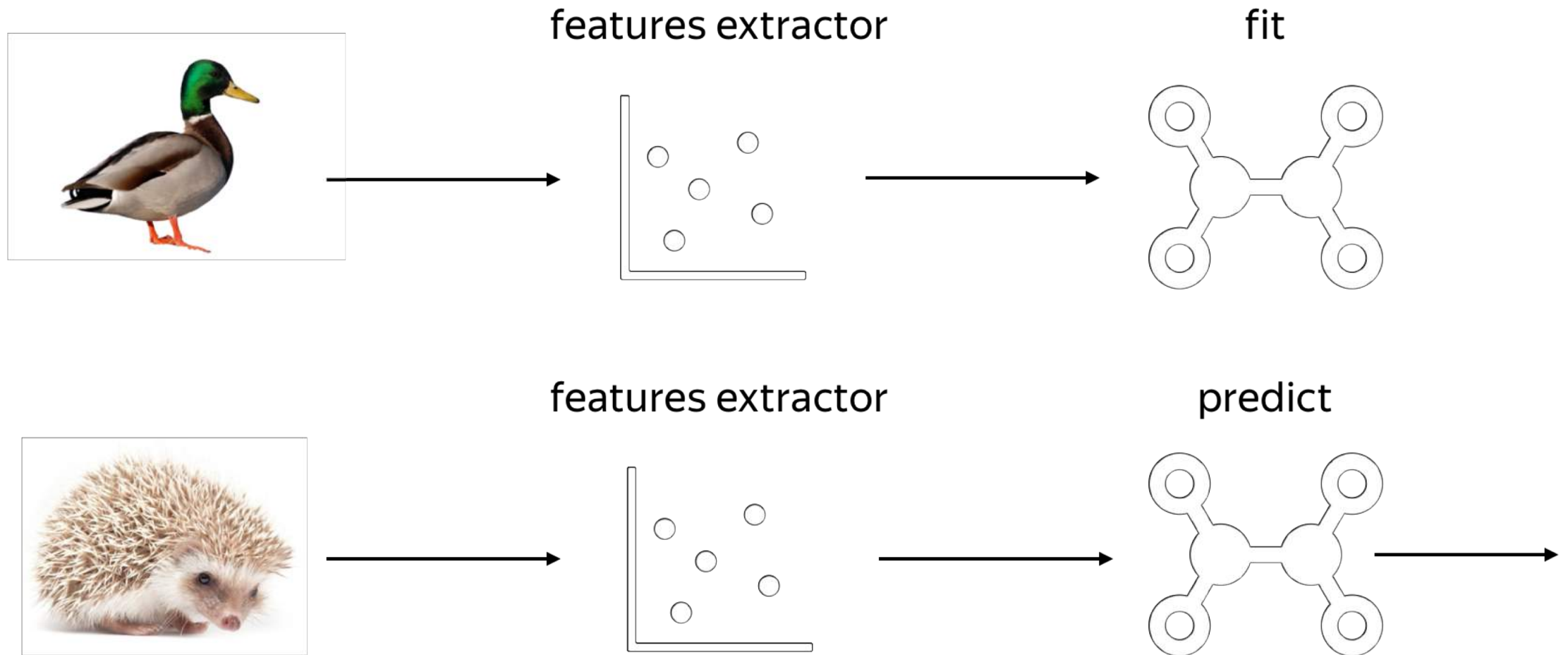


# Сбор данных

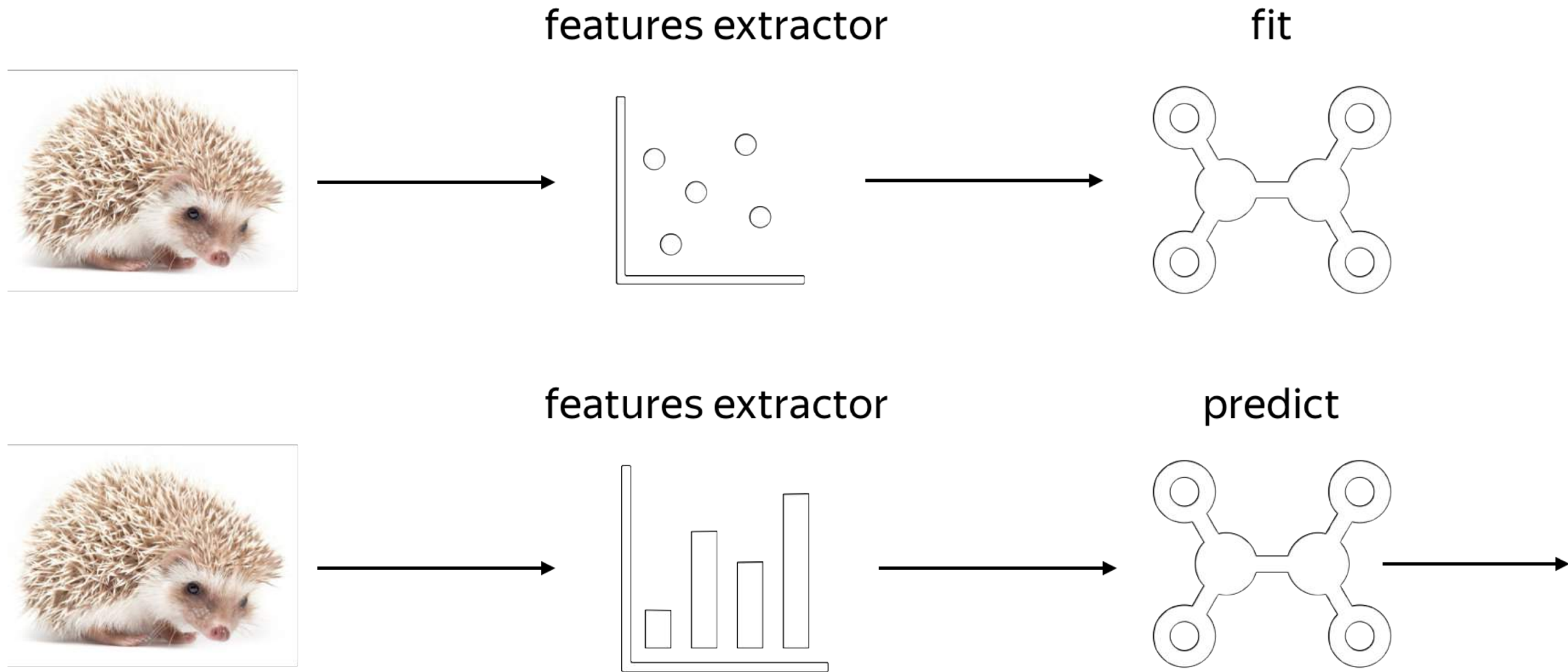




# Сбор данных



# Сбор данных



# Сбор данных

Копипаста это плохо

Копипаста это плохо

Копипаста это плохо

Копипаста это плохо

Копипаста это плохо

Копипаста это плохо

Копипаста это плохо

Копипаста это плохо

Копипаста это плохо

Копипаста это плохо

Копипаста это плохо

Копипаста это плохо

Копипаста это плохо

Копипаста это плохо

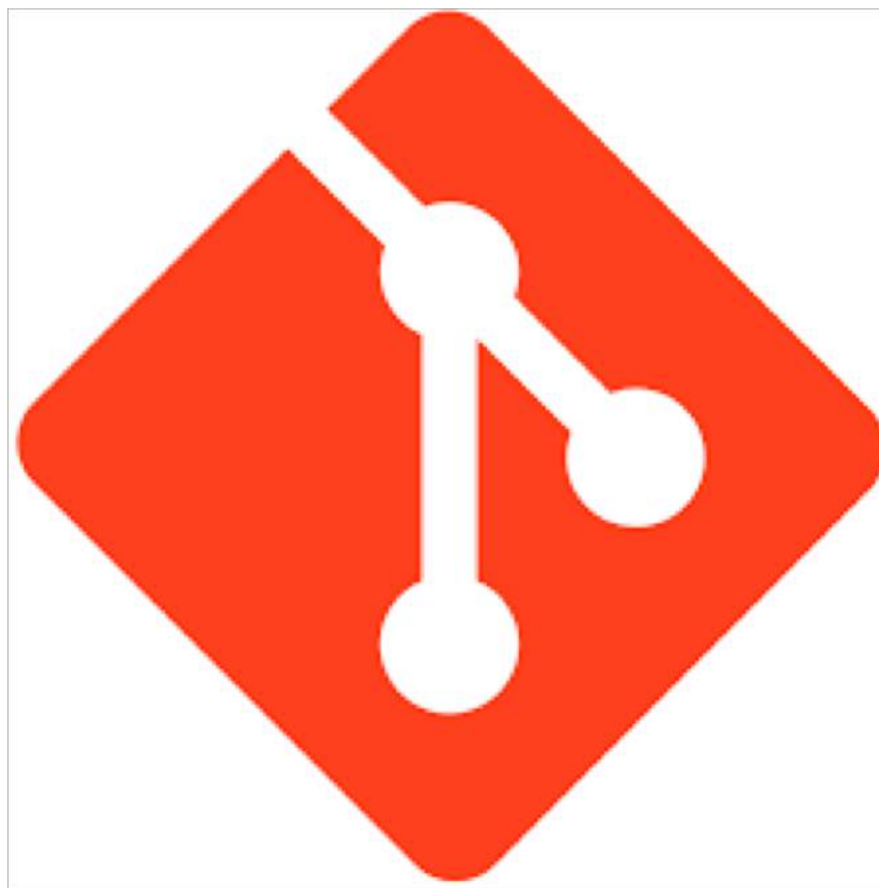
Копипаста это плохо

Копипаста это плохо

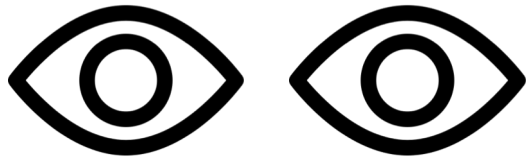
Копипаста это плохо

Копипаста это плохо

# Сбор данных

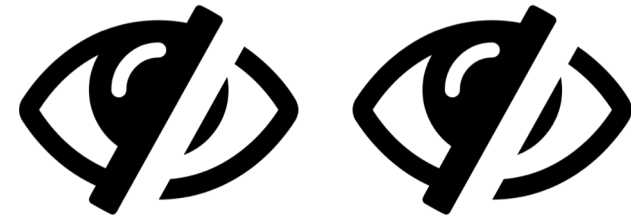


# Сбор данных



**Изучаем данные глазами**

**Но не заглядываем в будущее**



**И не подменяем понятия**

**СТАВЬ ЛАЙК**

**ЕСЛИ ЛЮБИШЬ СВОЕГО КОТА**

# Обучение модели



# Обучение модели

- **MVP и инфраструктура для экспериментов первостепенно**



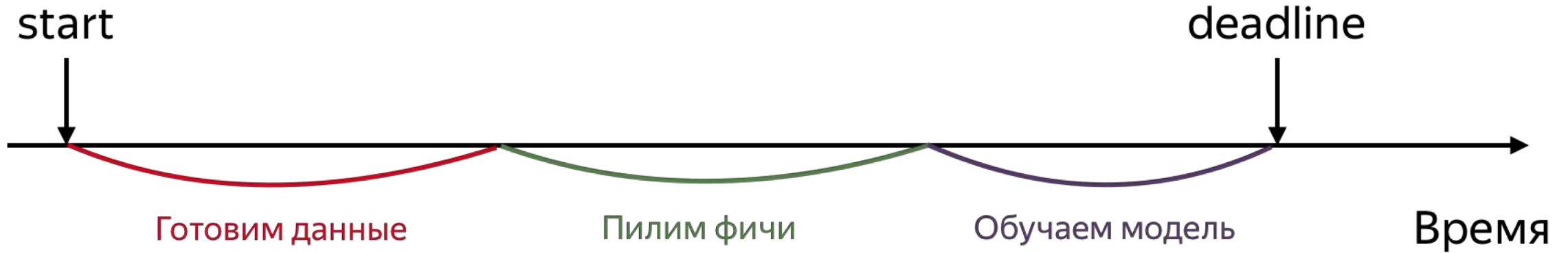
# Обучение модели



# Обучение модели

- **MVP и инфраструктура для экспериментов первостепенно**
- **Много быстрых итераций лучше малого числа долгих**

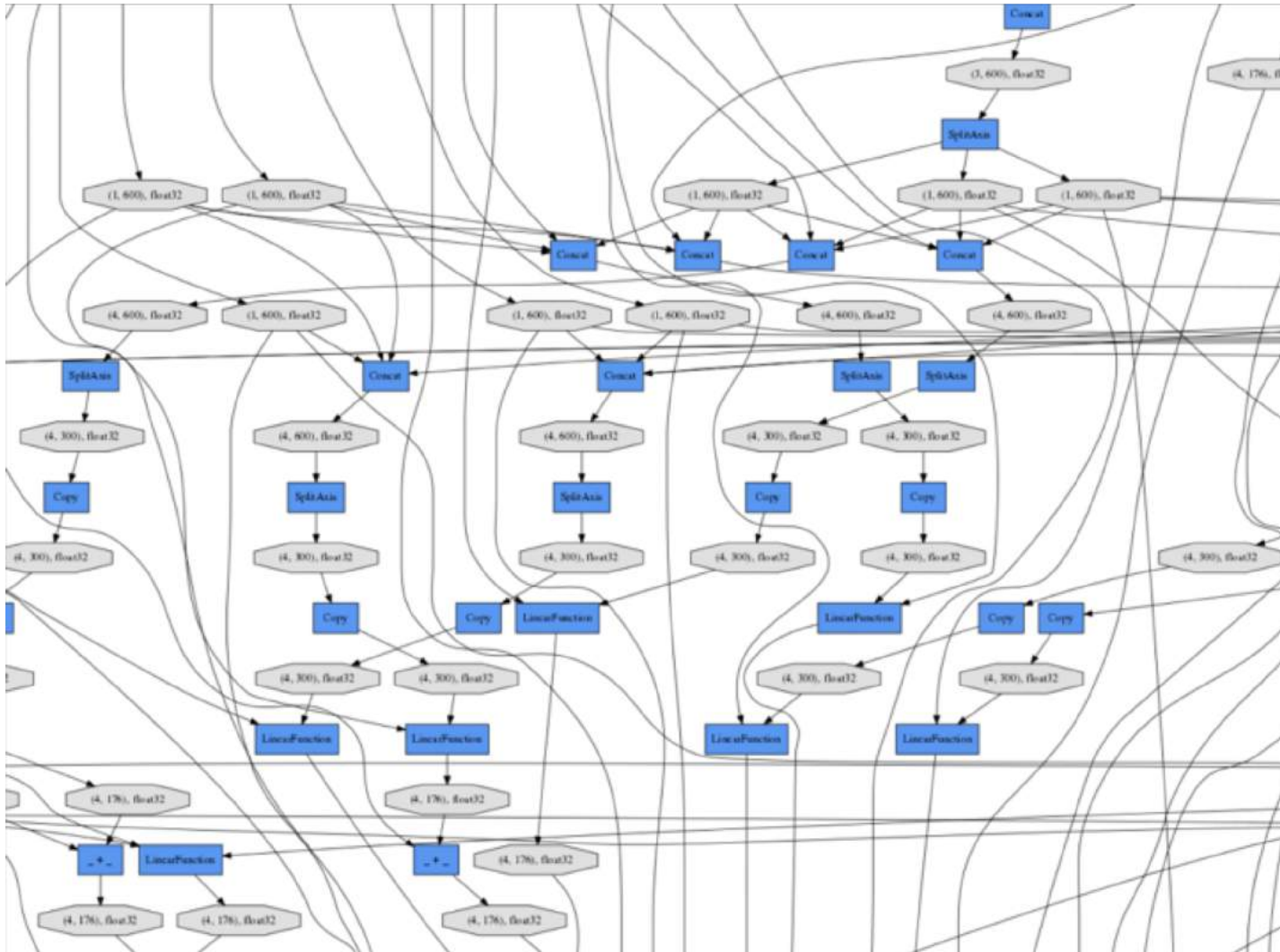
# Обучение модели



# Обучение модели

- **MVP и инфраструктура для экспериментов первостепенно**
- **Много быстрых итераций лучше малого числа долгих**
- **В ML воспроизводимость супер важна**

# Обучение модели

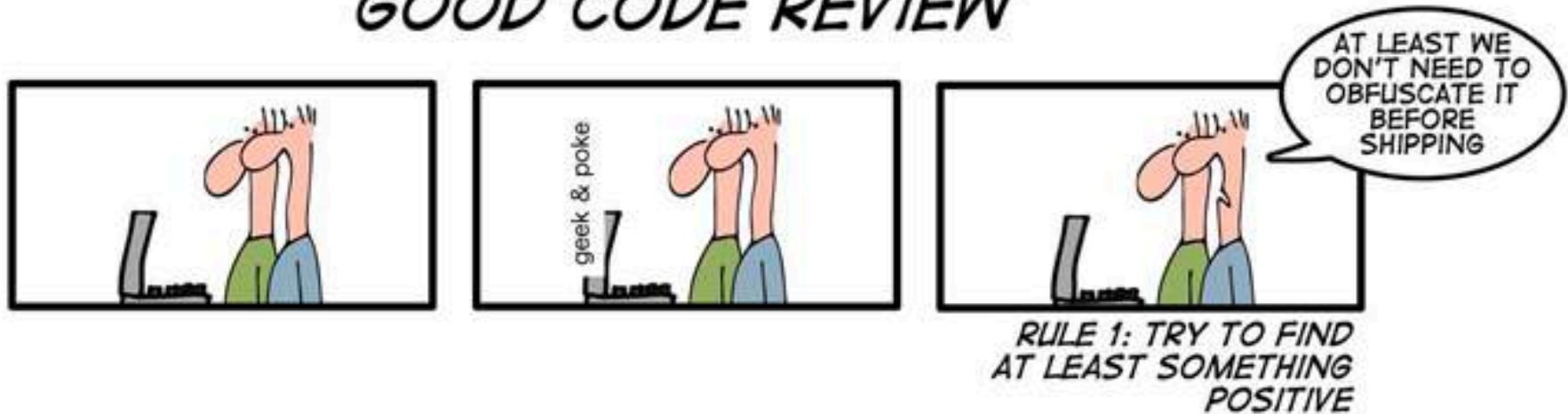


# Обучение модели

- **MVP и инфраструктура для экспериментов первостепенно**
- **Много быстрых итераций лучше малого числа долгих**
- **В ML воспроизводимость супер важна**
- **Тесты и ревью после удачных итераций**

# Обучение модели

## *HOW TO MAKE A GOOD CODE REVIEW*



# Внедрение модели





# Внедрение модели

- **Требования к коду значительно возрастают**
- **Нужна отдельная инфраструктура для выкатки моделей**
- **Могут проявляться неожиданные с точки зрения DS проблемы**
- **Нужно всегда иметь план Б**

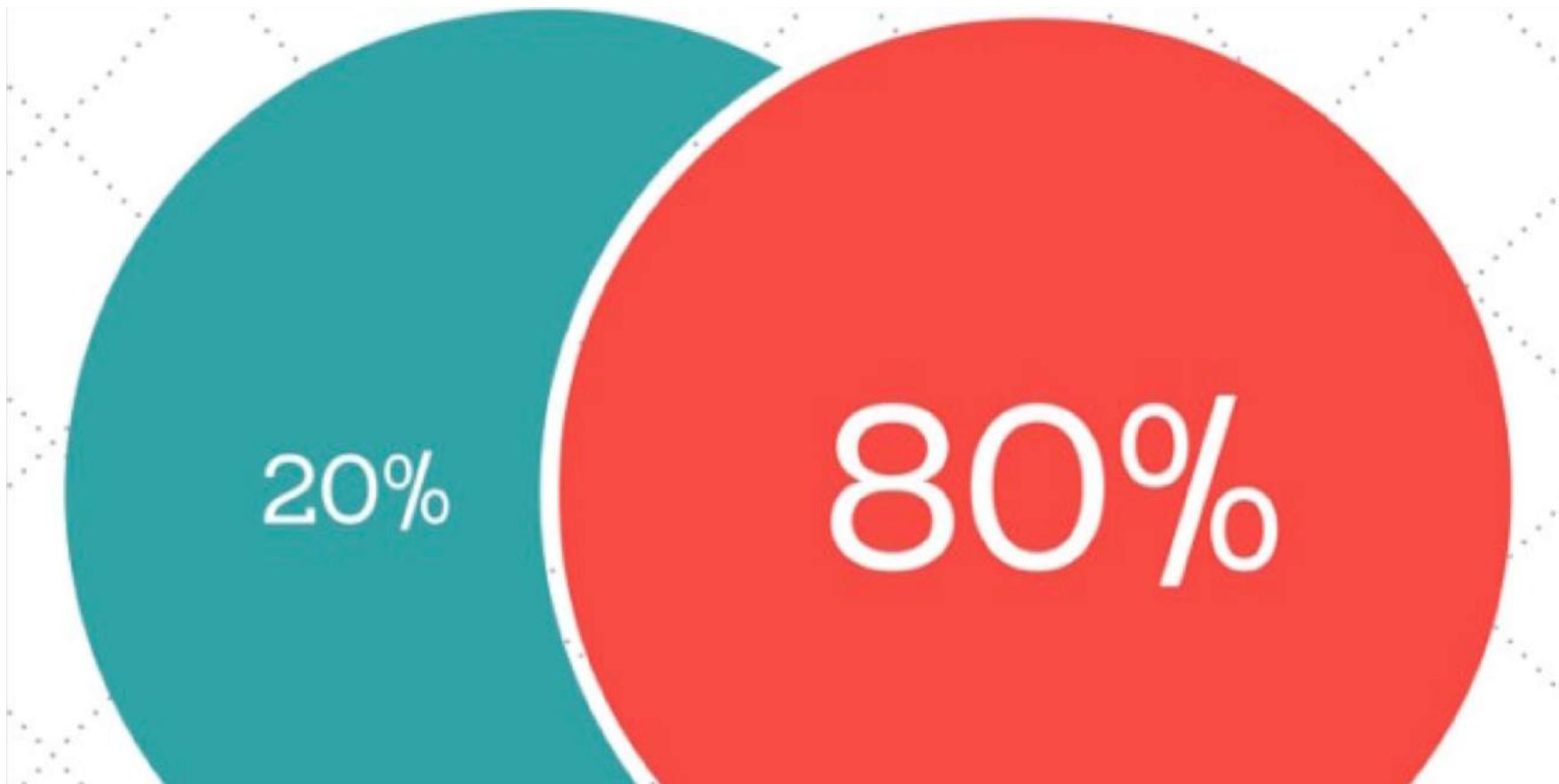
# Постаналитика



# Постаналитика

- **Метрики это хорошо, но пользователи важнее**

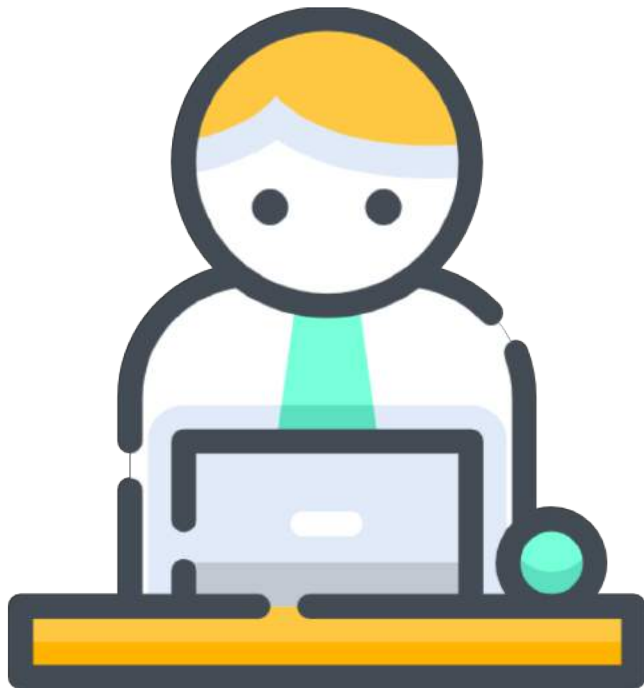
# Постаналитика



# Постаналитика

- **Метрики это хорошо, но пользователи важнее**
- **Используйте статистику, ~~врите~~ честно**

# Постаналитика



<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
$\geq 0.1$	

# Постаналитика

- **Метрики это хорошо, но пользователи важнее**
- **Используйте статистику , врите честно**
- **Даже эксперимент с отличным результатом может всё испортить**







# Этапы ML проекта

- Постановка задачи
- Сбор данных
- Обучение модели
- Внедрение в продакшн
- Постаналитика



**DS команда**

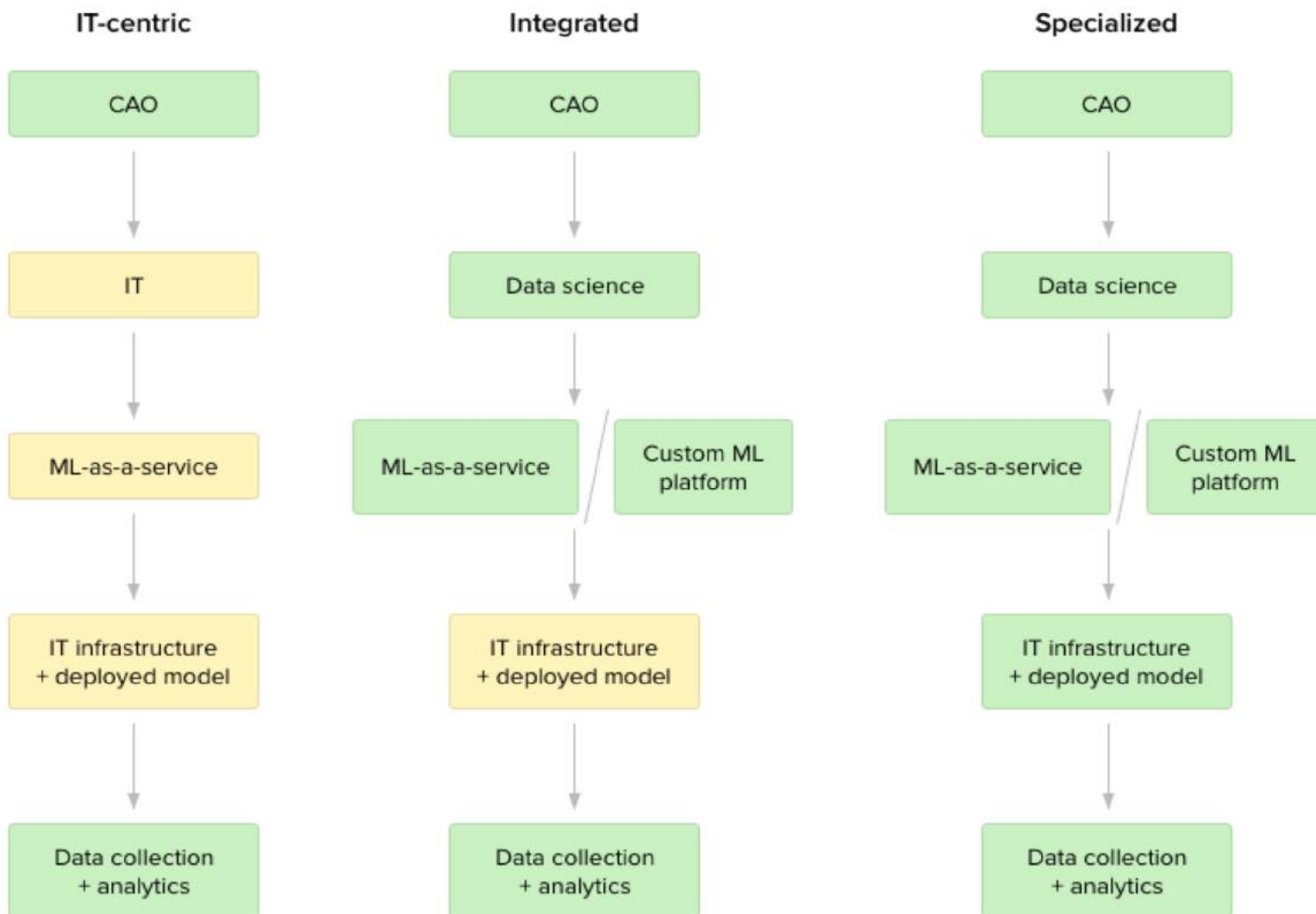


	Data Analyst	Machine Learning Engineer	Data Engineer	Data Scientist
Programming Tools	Very important	Very important	Very important	Very important
Data Visualization and Communication	Very important	Somewhat important	Somewhat important	Very important
Data Intuition	Somewhat important	Very important	Somewhat important	Very important
Statistics	Somewhat important	Very important	Somewhat important	Very important
Data Wrangling	Not that important	Not that important	Very important	Very important
Machine Learning	Not that important	Very important	Not that important	Very important
Software Engineering	Not that important	Somewhat important	Very important	Somewhat important
Multivariable Calculus and Linear Algebra	Not that important	Very important	Not that important	Somewhat important
<div> <div>Not that important</div> <div>Somewhat important</div> <div>Very important</div> </div>				

## Роли в DS команде

## Data Science Team Structures

IT Data Science



# Организация процессов

# Ещё несколько ролей из практики

**Разработчик**

**Менеджер**

**Исследователь**

**Каглер**

**Бизнесмен**

**Статистик**

**Математик**

**Астронафт**



**А кем хотите  
стать вы?**