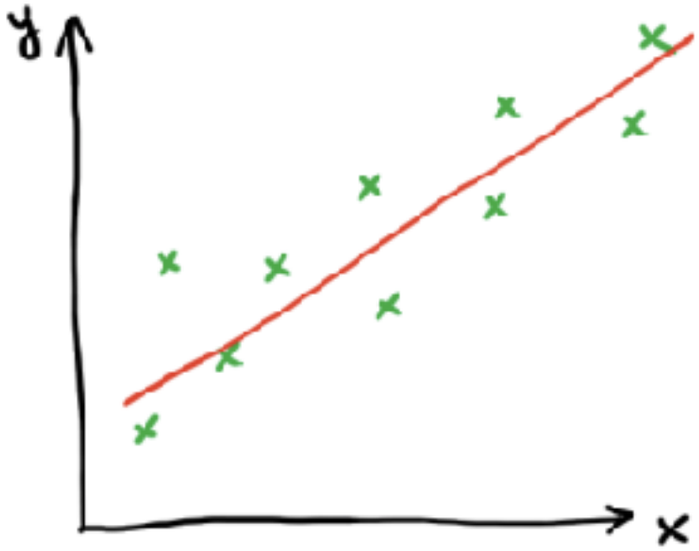
Как решить задачу в ML?

Какие типы задач вообще бывают?

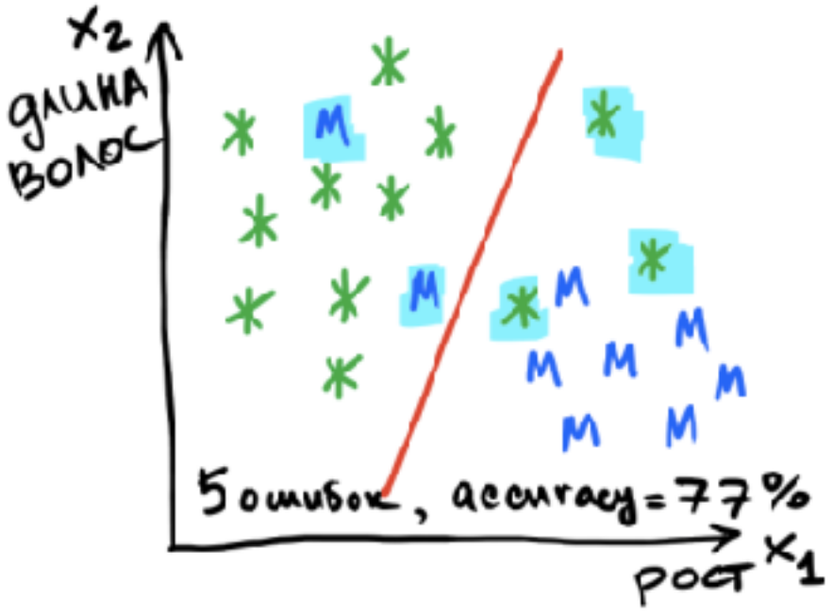
Есть разметка - “обучение с учителем”

x	y
1	2
3	5
-1	-2
5	?



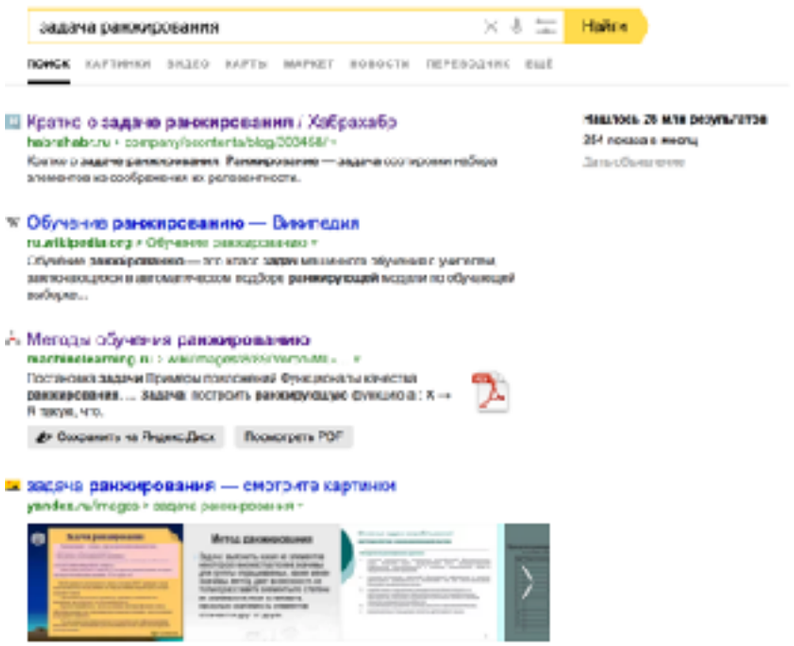
Предсказываем число – регрессия

x_1	x_2	y
180	5	М
170	20	Ж
160	5	М
190	30	?

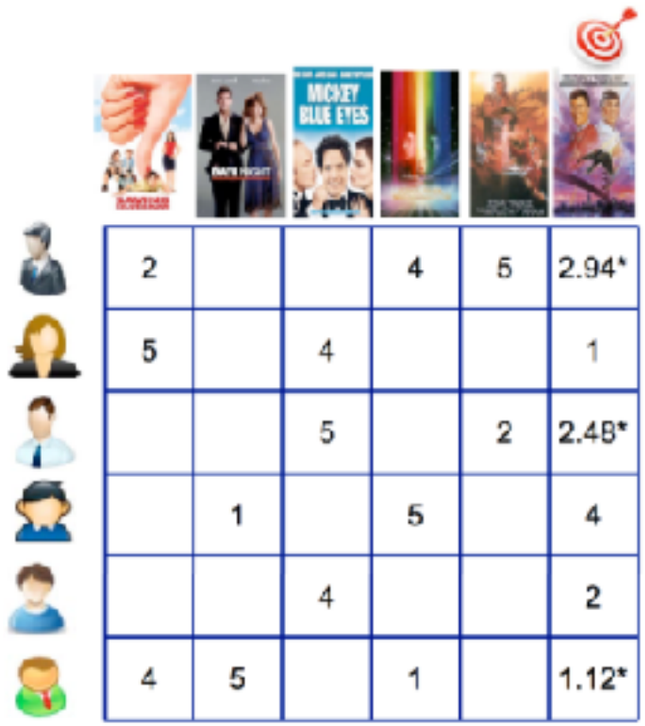


Предсказываем категорию – классификация

Есть разметка - “обучение с учителем”

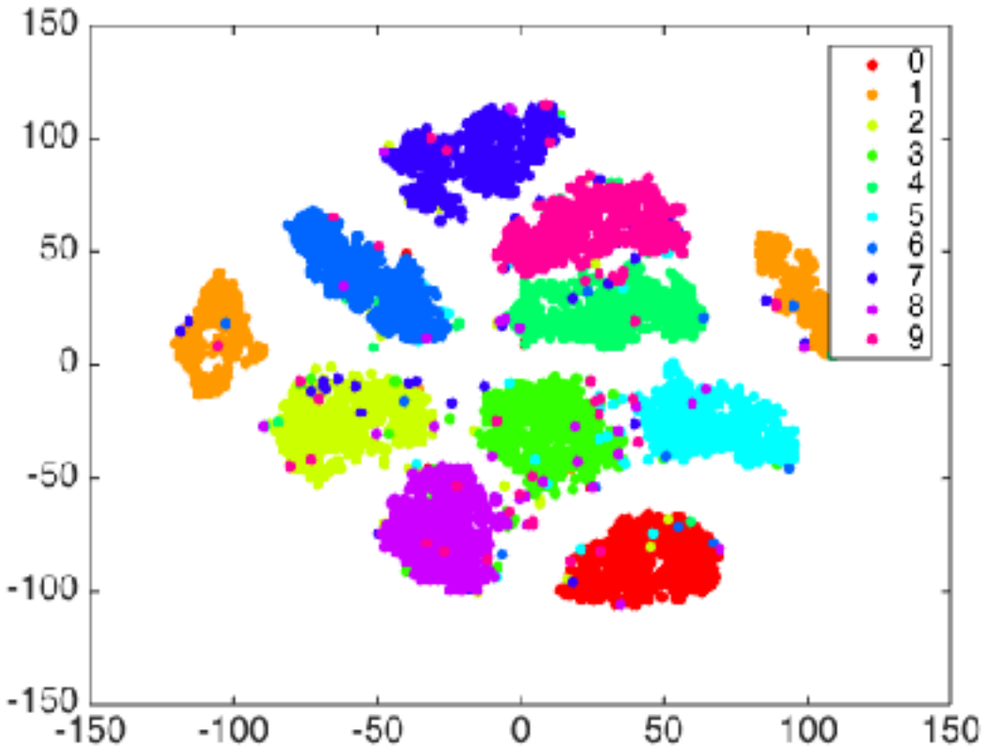


Предсказываем порядок элементов – ранжирование



Решаем эти же задачи для пары пользователь/объект – задача рекомендации

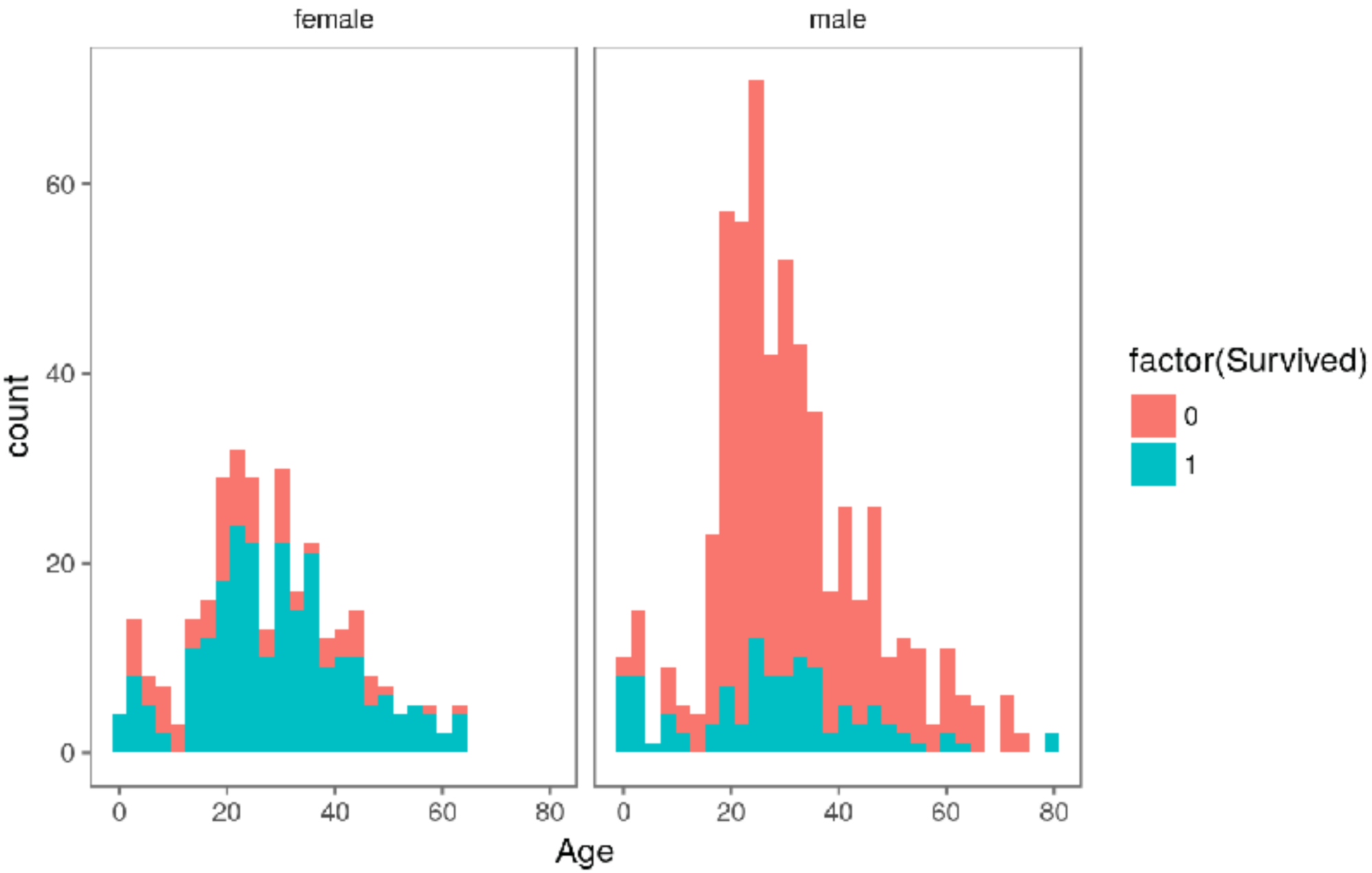
Нет разметки - “обучение без учителя”



Определяем кластер элемента – кластеризация

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S

Преобразовываем сырые данные
в удобный формат (.csv)

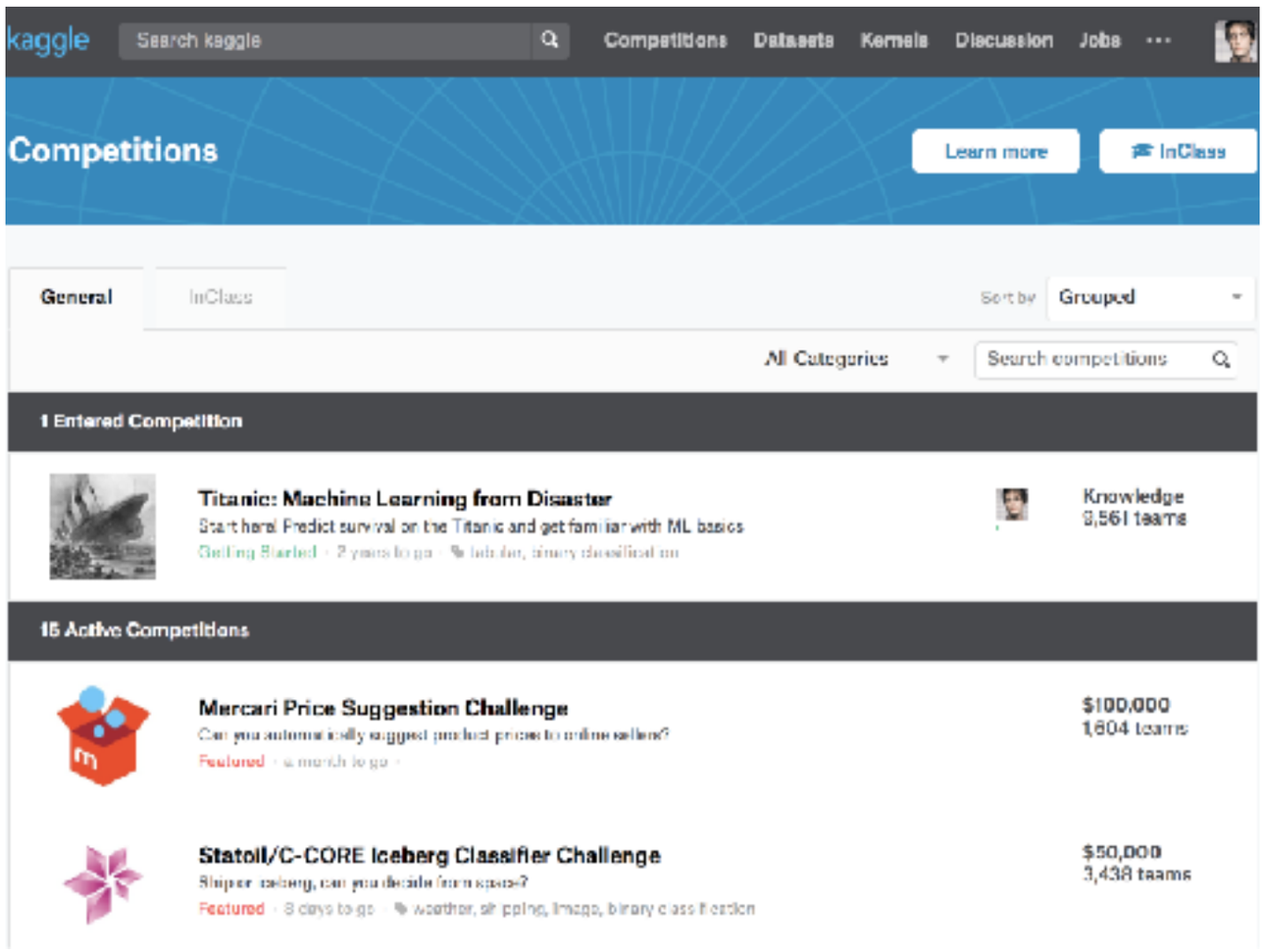


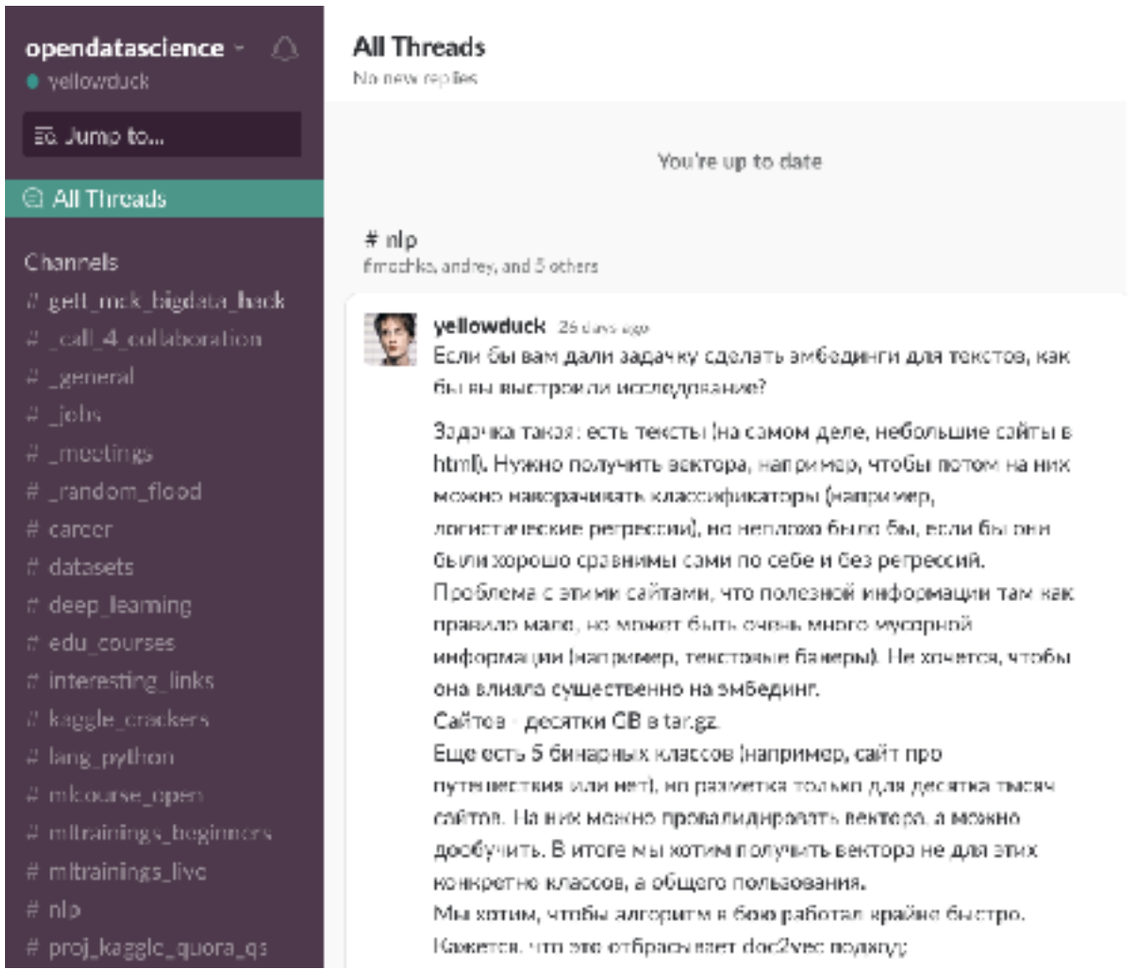
Визуализируем и изучаем данные

Согласовываем метрику

Есть бизнес-метрики: деньги, отток, …

Есть метрики качества алгоритма: accuracy, MSE, …





Размышляем о способах решения:
смотрим прошедшие соревнования на kaggle.com,
спрашиваем в slack ods.ai

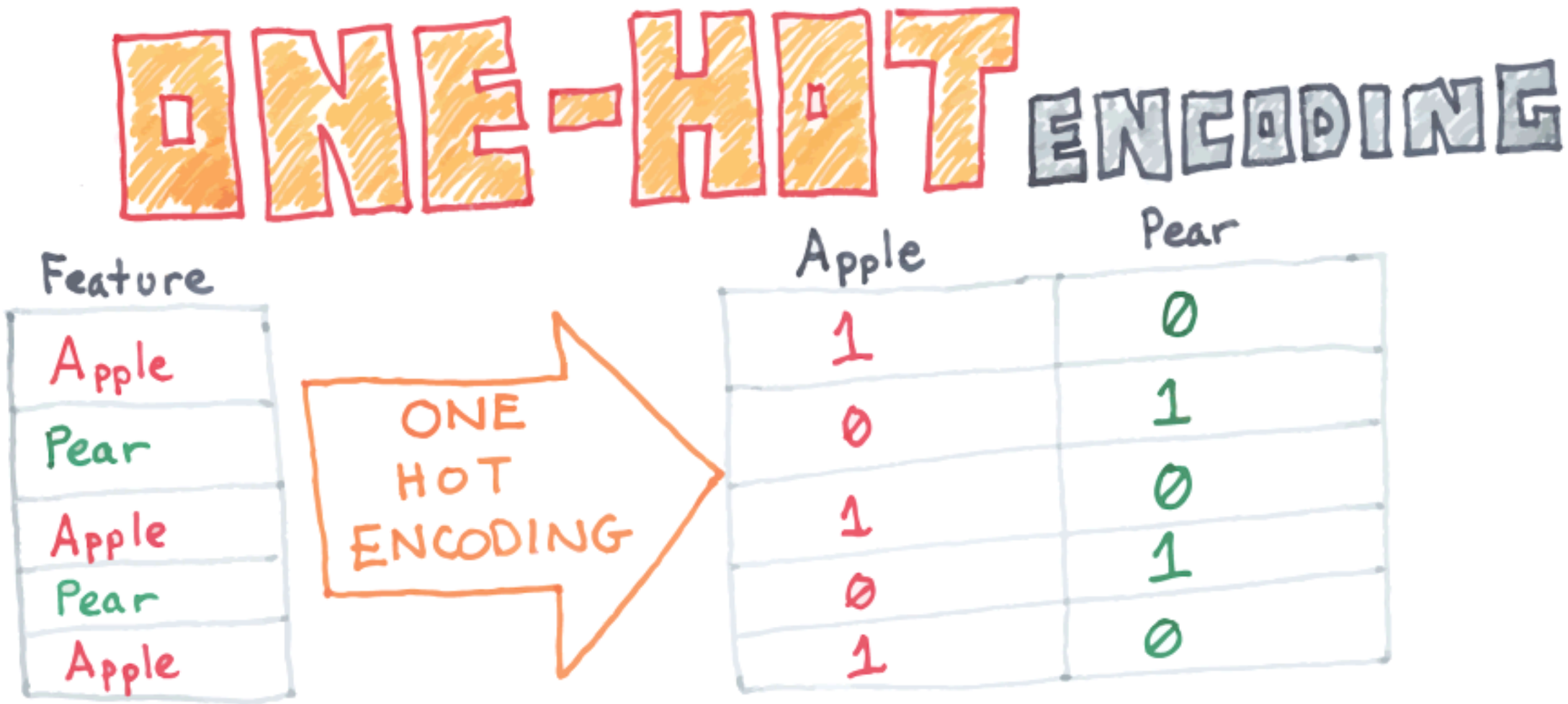
Где настоящий лось?

Яндекс.Толока
Яндекс

amazon mechanical turk

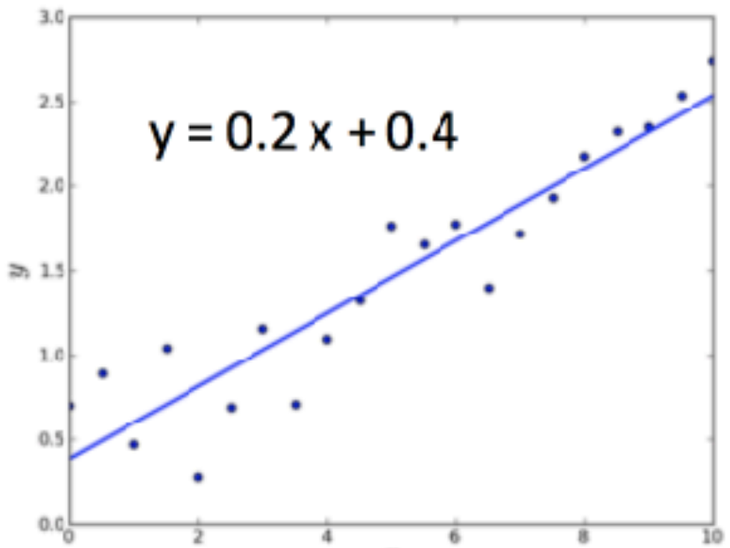
Нужна ручная разметка?

(минимум одна неделя)



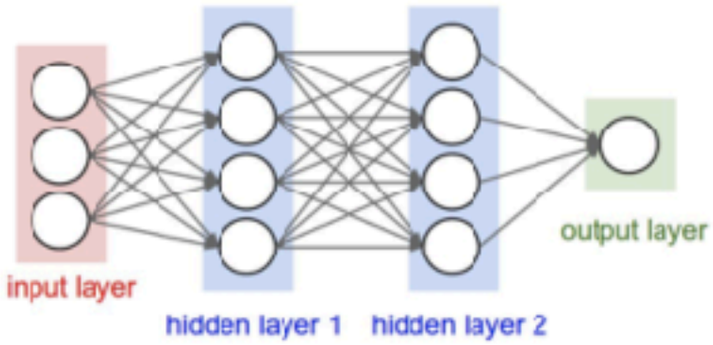
Предобрабатываем данные:
выявляем выбросы, заполняем пропуски,
преобразовываем категориальные признаки

Линейные модели



Решающие деревья

Нейронные сети



PYTORCH
Deep Learning with PyTorch

Yandex CatBoost

scikit learn

TensorFlow

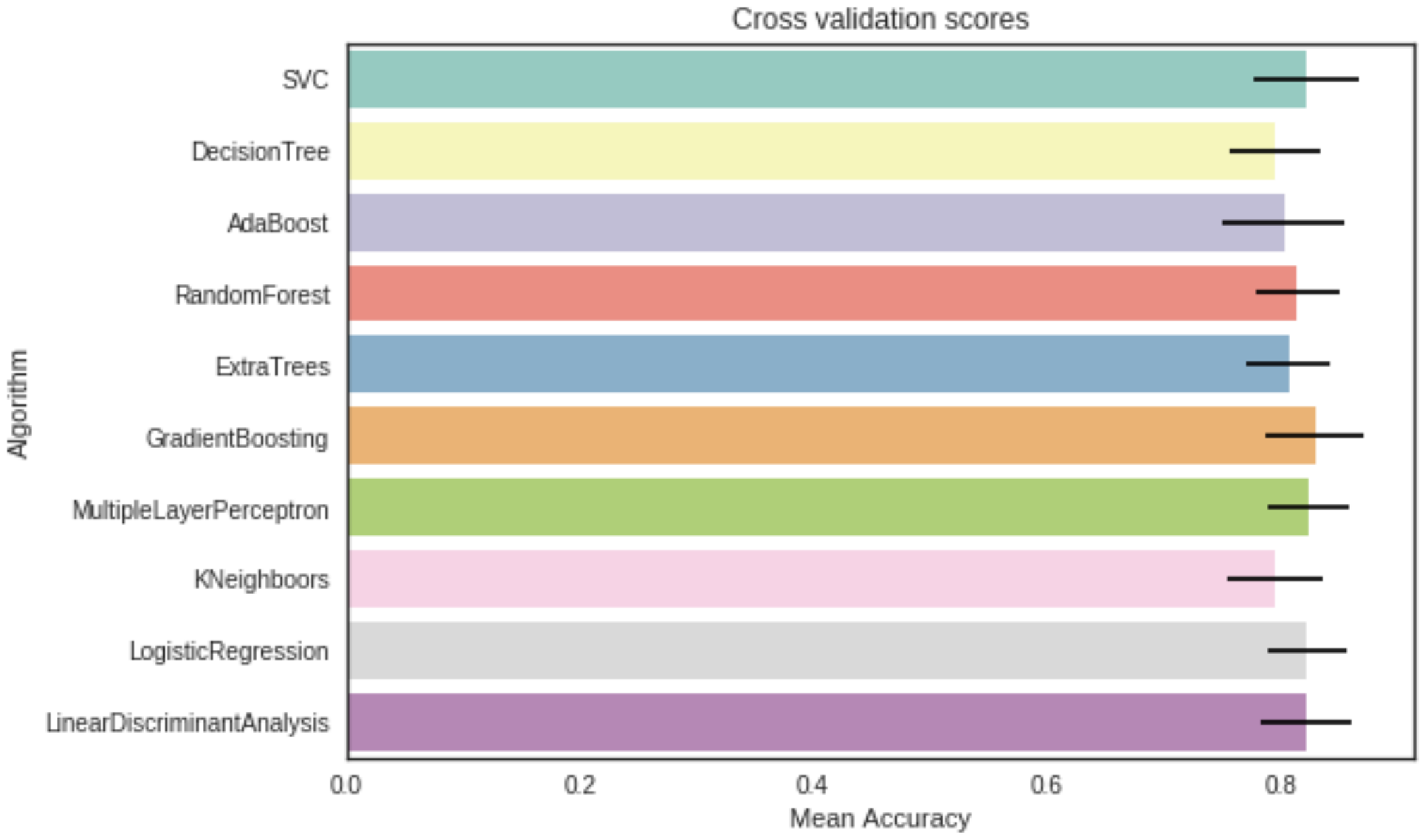
Обучаем модель

dmlc XGBoost

Accuracy = 88%

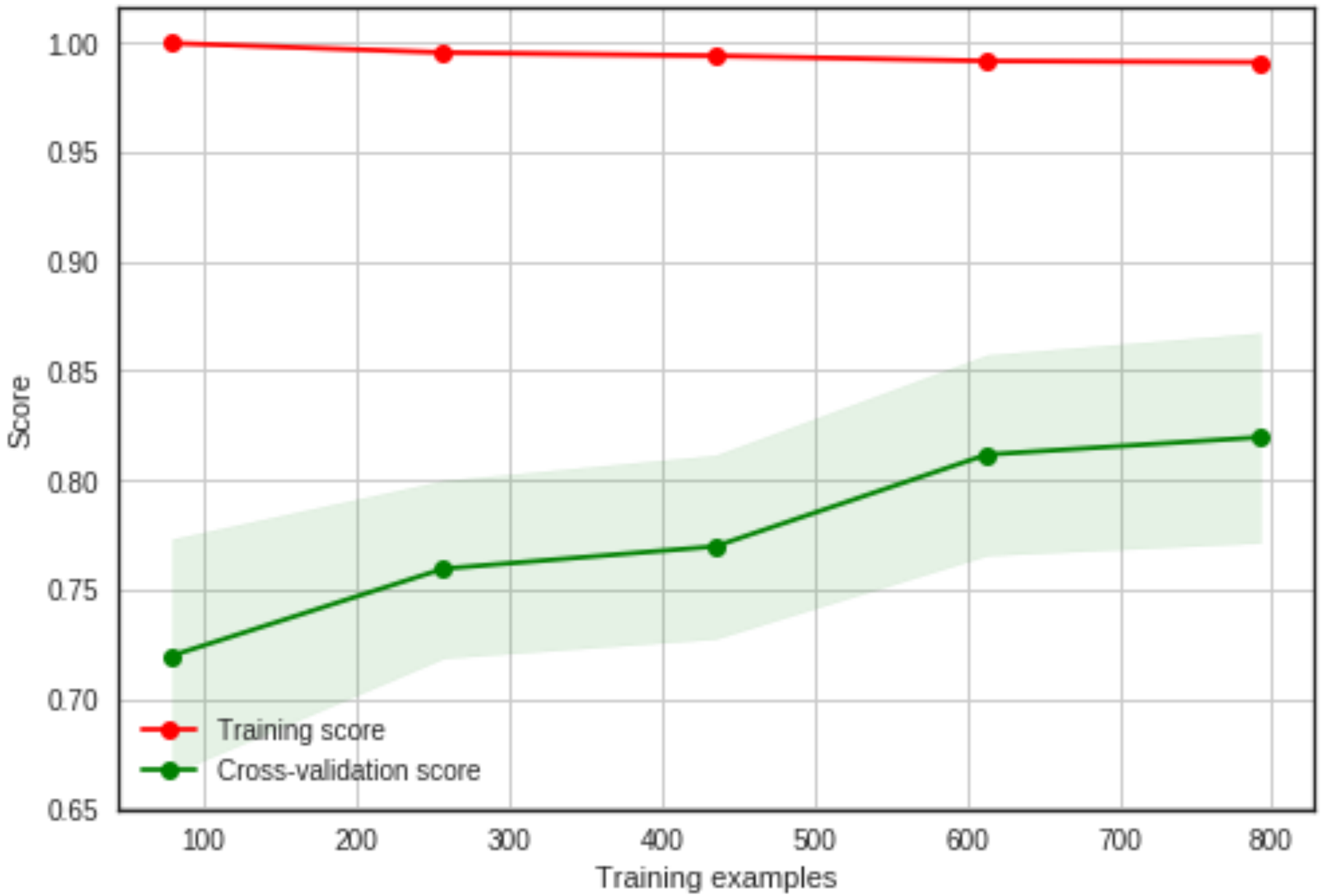
Получаем предварительный результат.
Часто работает принцип “20/80”.

(Где-то здесь прошли первые 2 недели + неделя на разметку)



Получаем предварительный результат.
Часто работает принцип “20/80”.

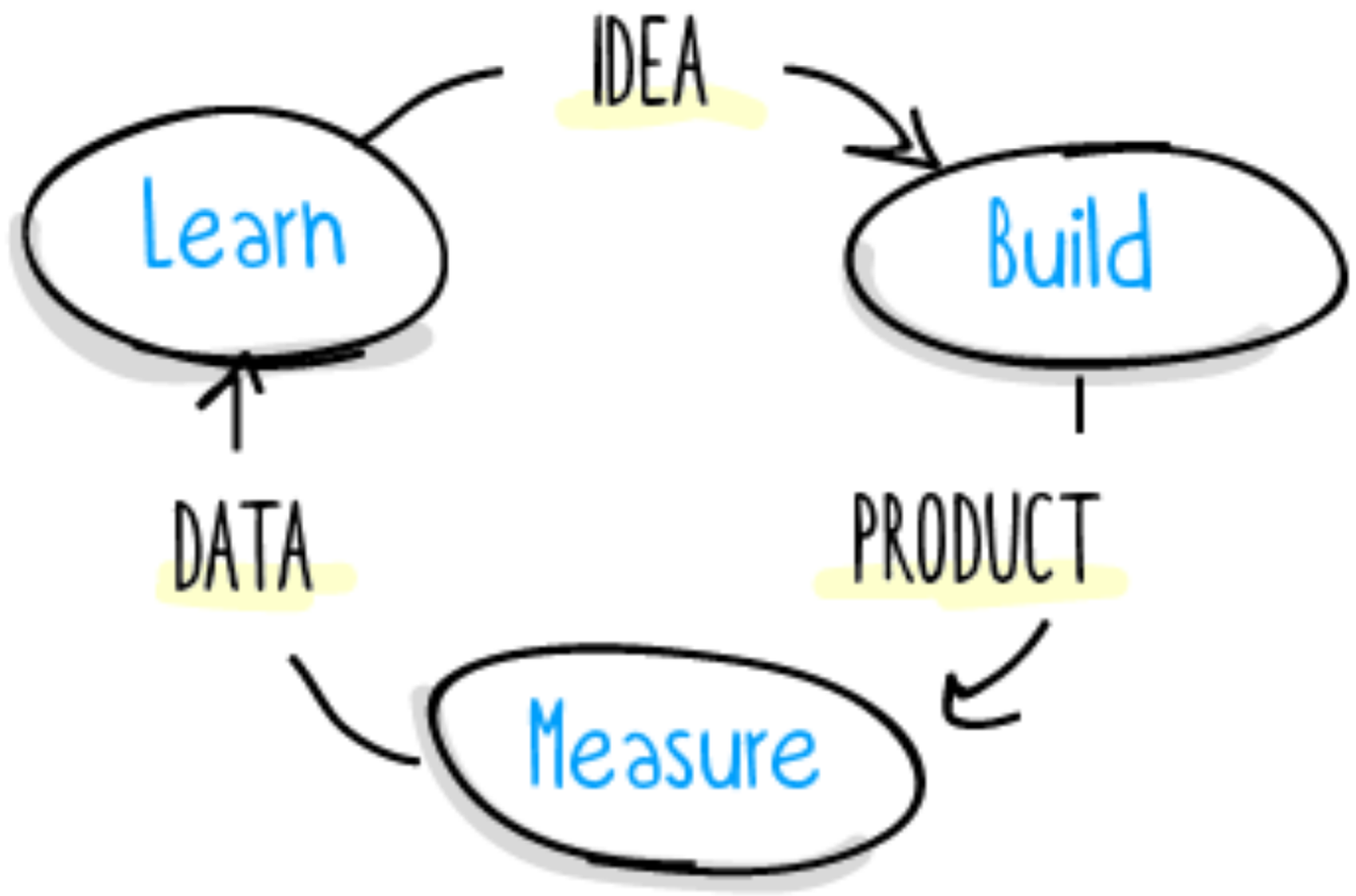
(Где-то здесь прошли первые 2 недели + неделя на разметку)



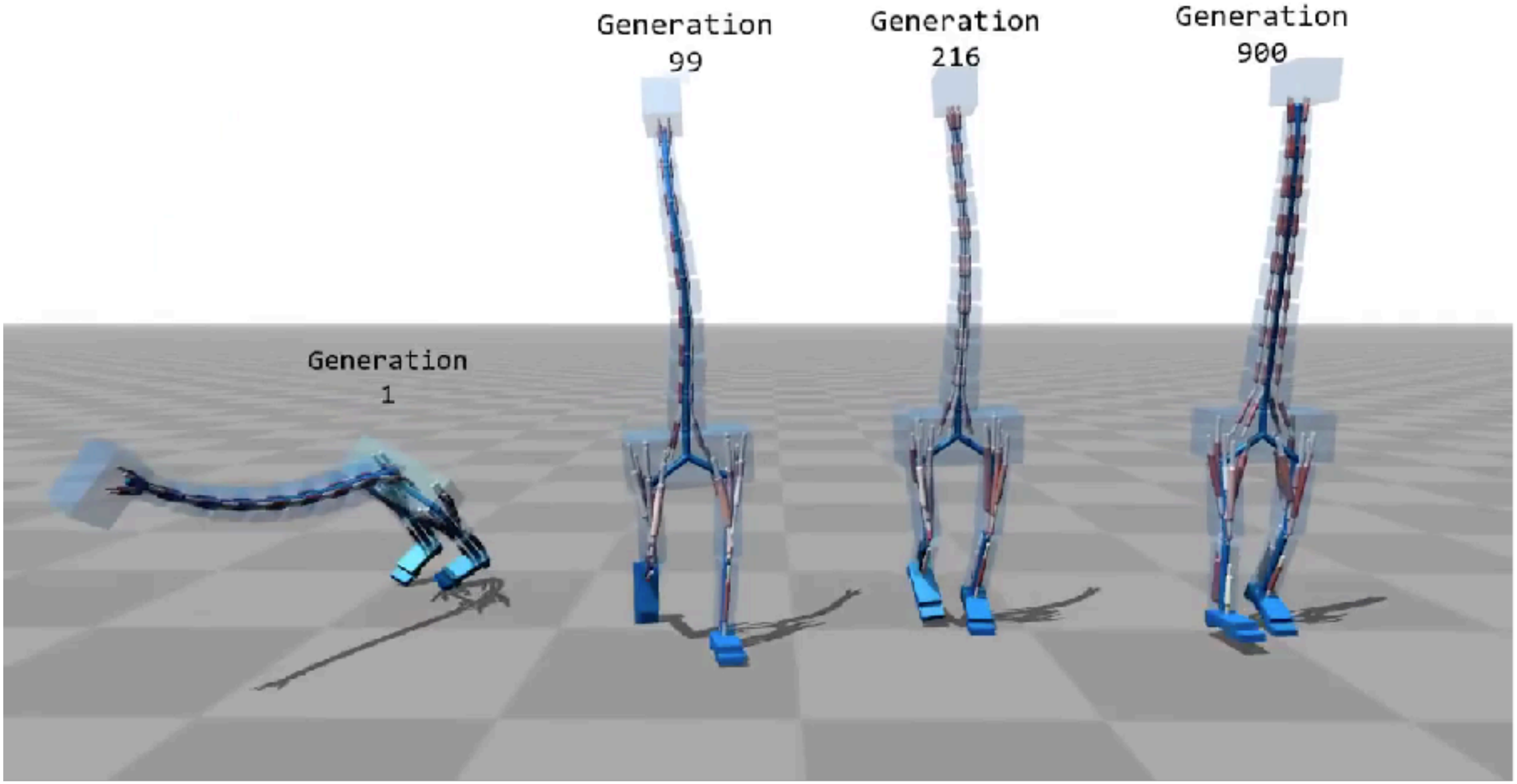
Может мало данных?

1997 Red River flood. Grand Forks

Иногда интересует не только предсказание,
но и степень уверенности в нем

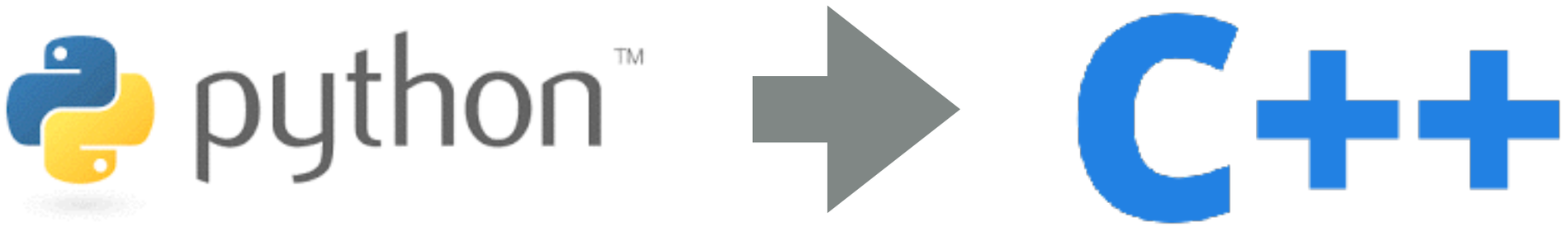


Корректируем метрики, запрашиваем новые данные, уточняем задачу, делаем предположения об итоговом качестве.



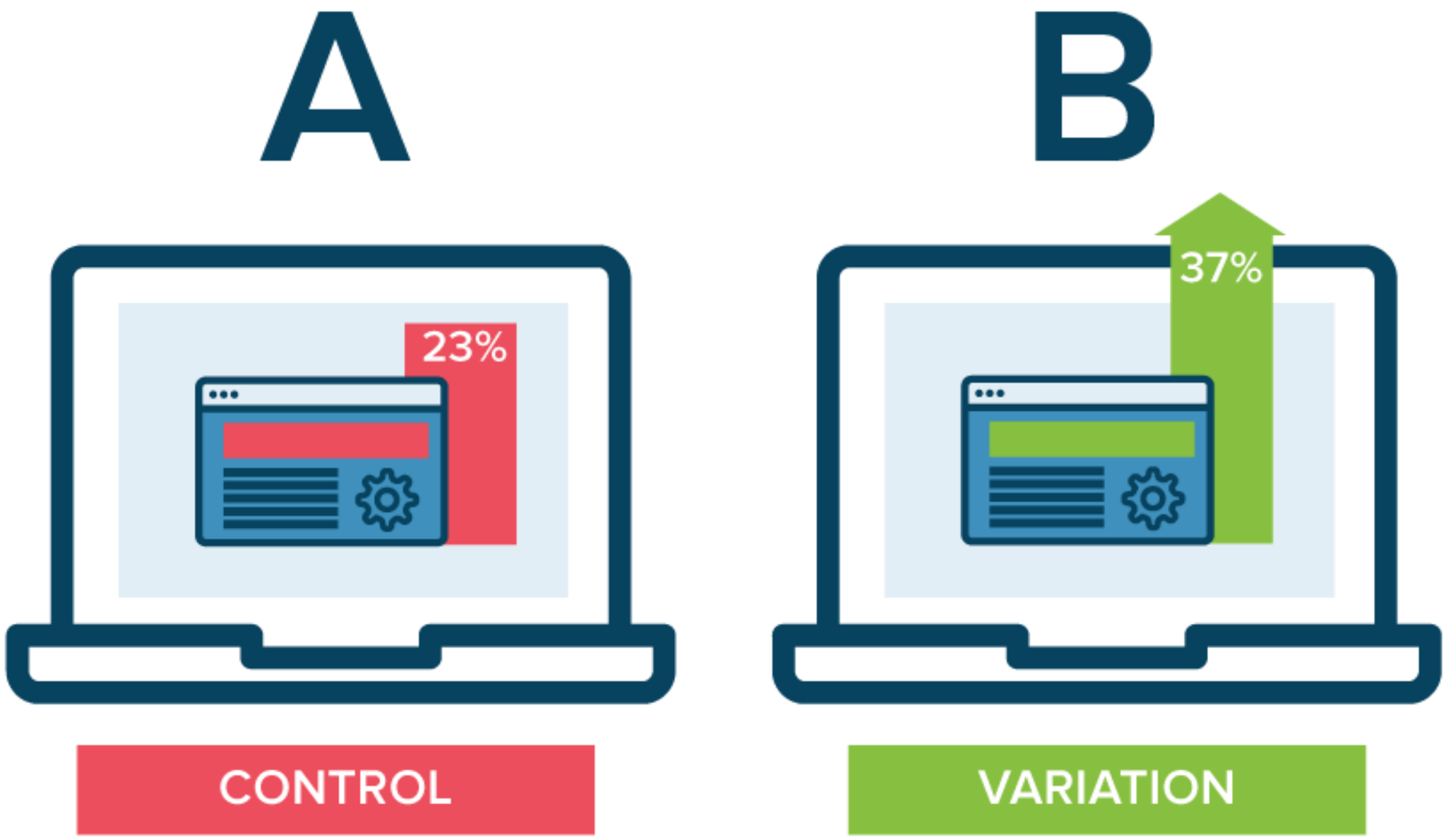
Тюним модель

(бесконечно долго)



Интегрируем модель в продакшен

(обычно дольше, чем вы предполагали вначале)



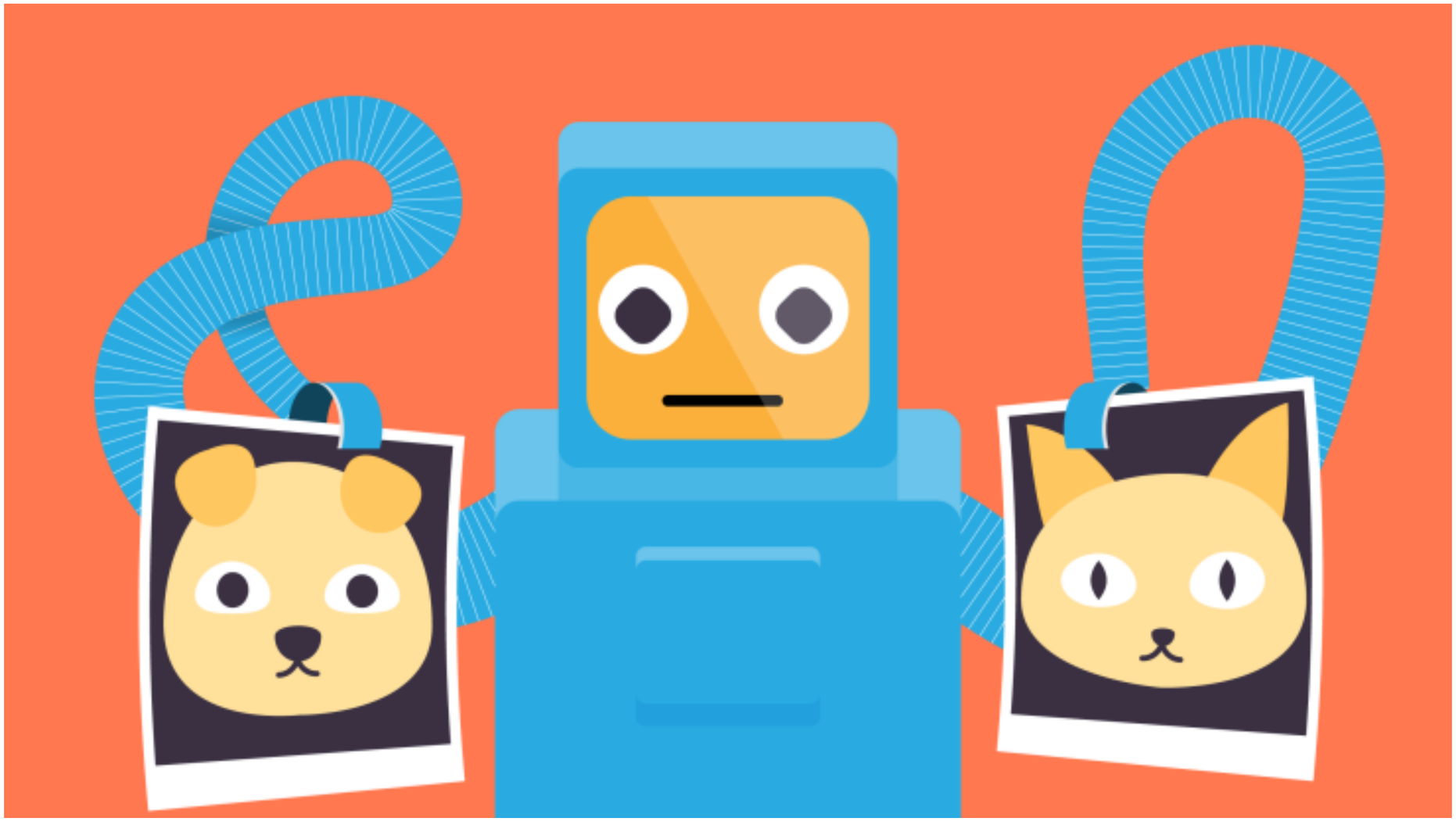
Тестируем на живом потоке

(минимум 2 недели)

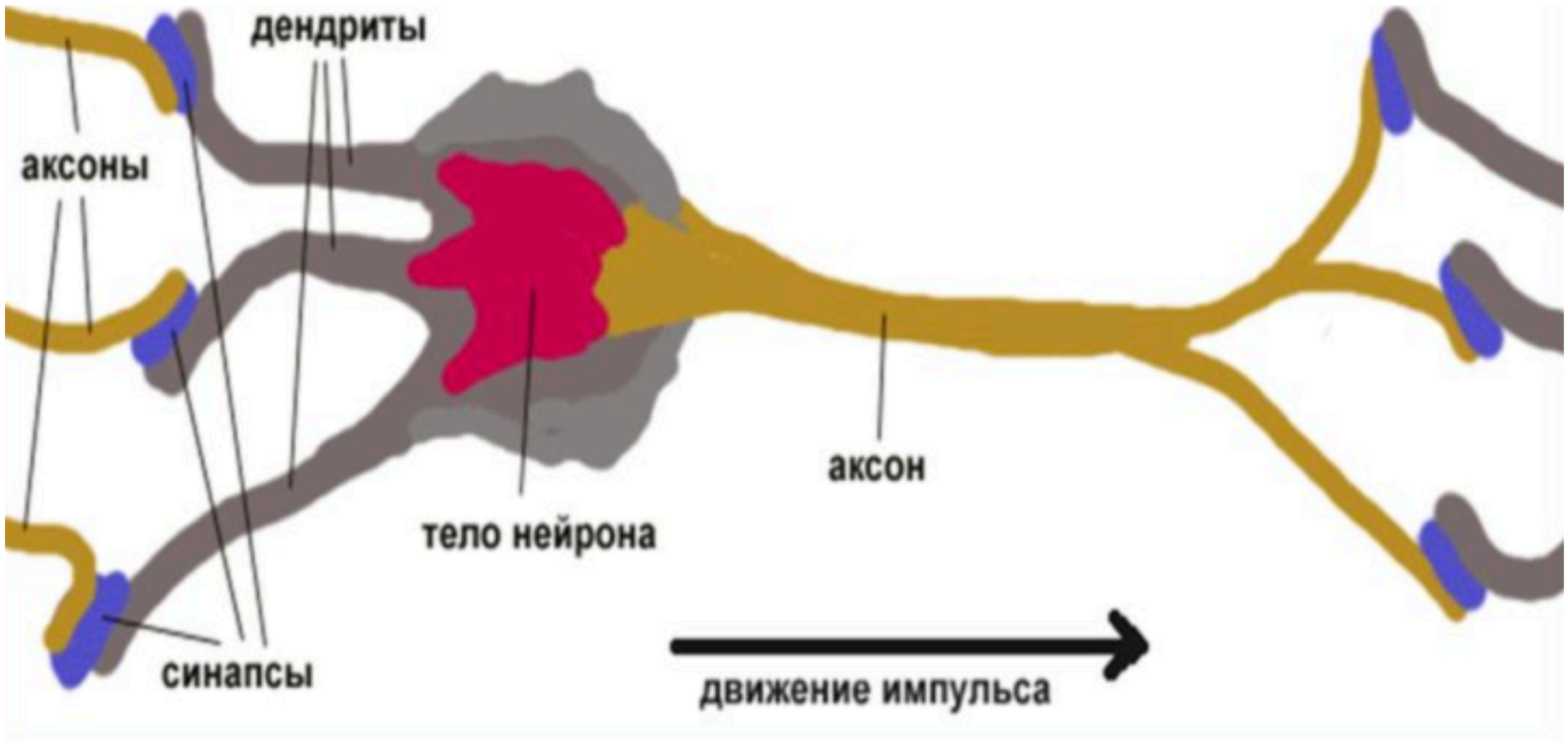
Когда всё в продакшене – работа еще не закончена!
Модели устаревают. Очень важно вести
документацию и коммитить свой код

Workshop

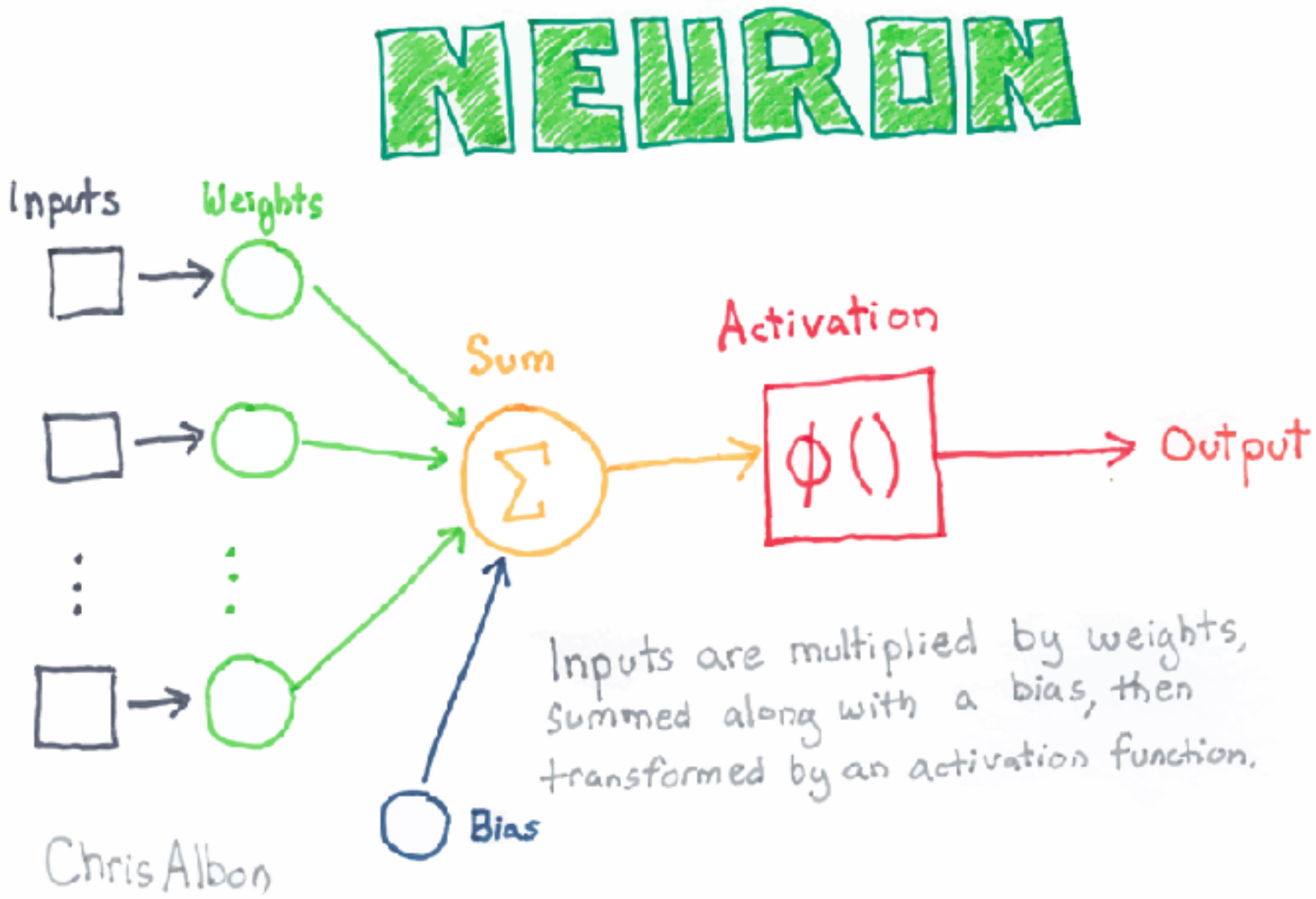
Что такое нейросети?



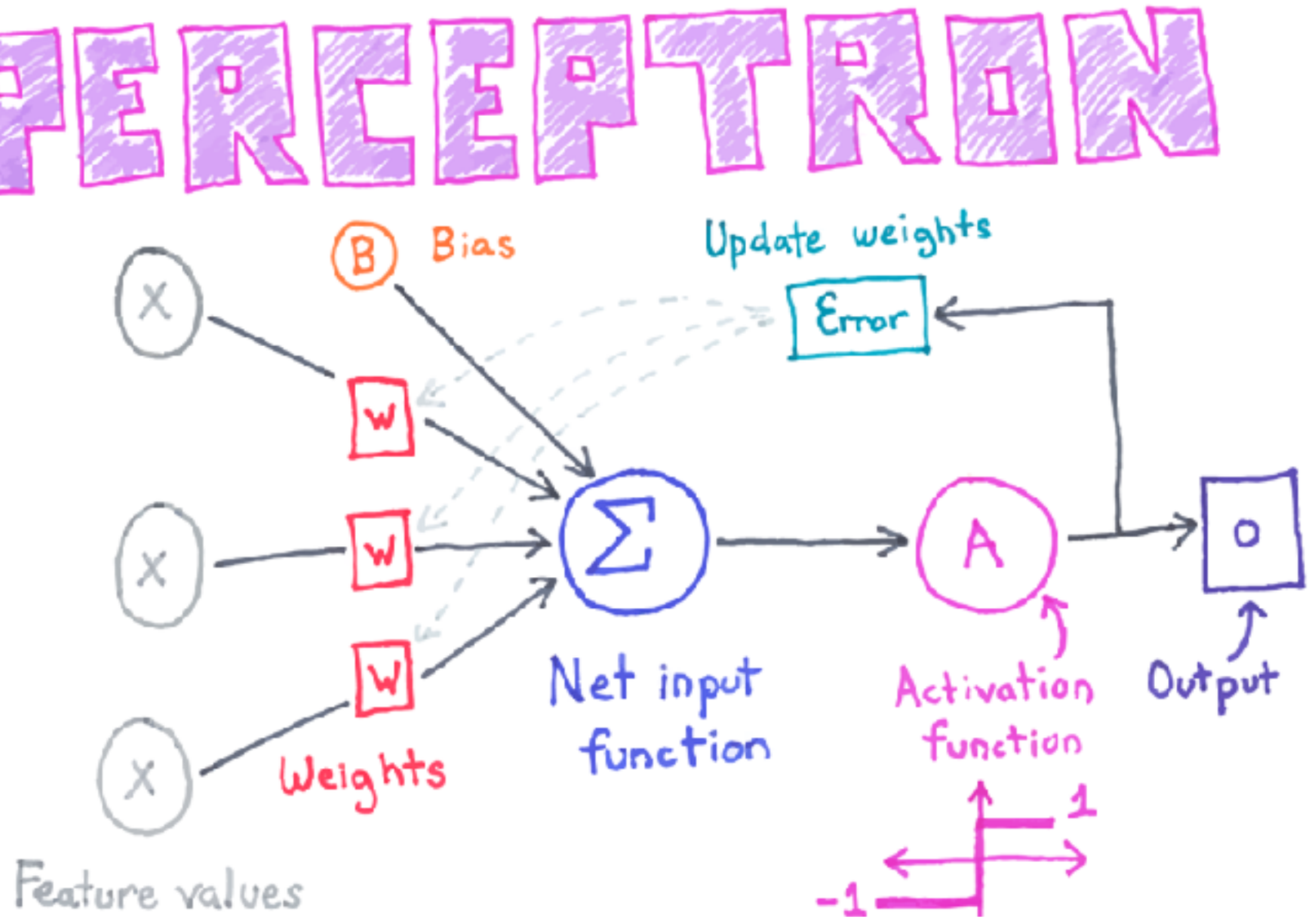
Нейросети хорошо работают **только** на структурированных данных:
изображения, звук, текст, временные ряды



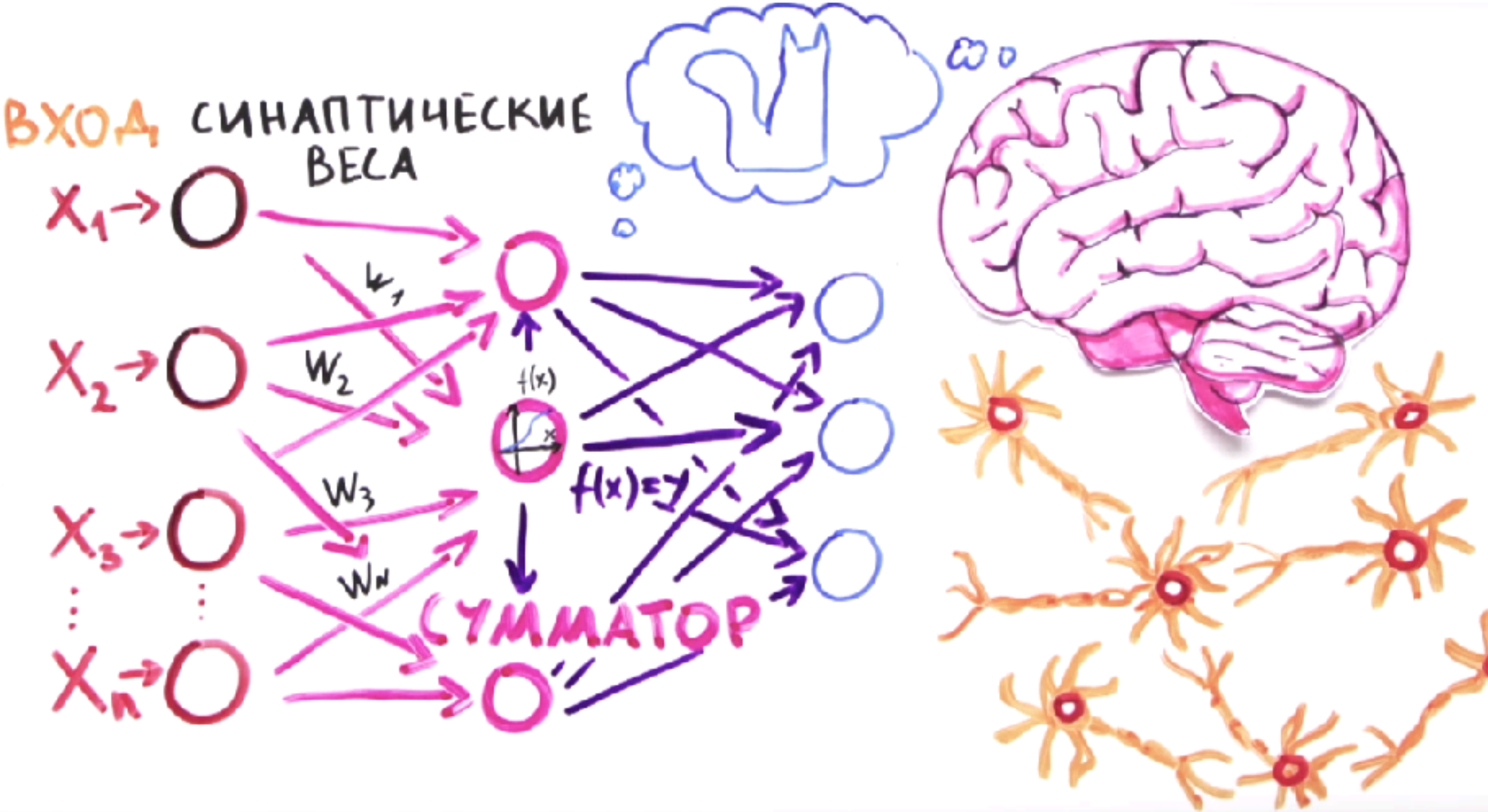
Биологический нейрон



Математическая модель нейрона



Нейросеть с одним нейроном



Нейронная сеть

Слайд для вопросов