

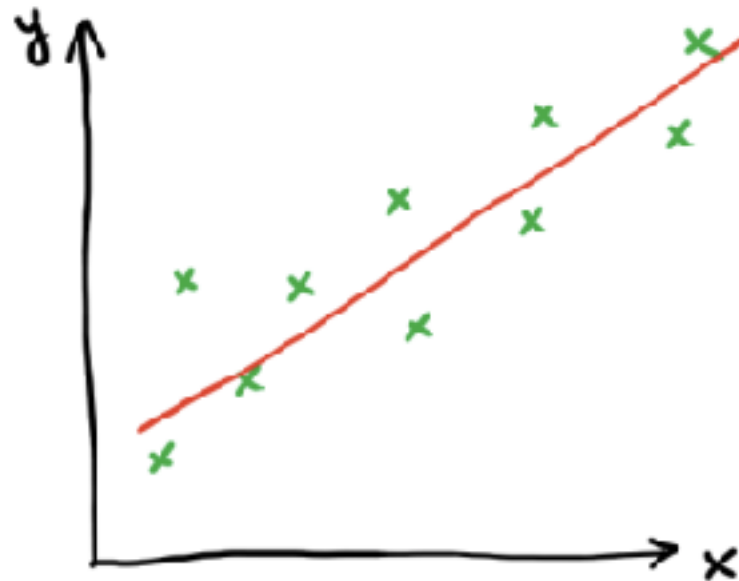
A black and white photograph of the RMS Titanic docked at a pier. The ship's four funnels are visible in the background, and the ship's hull is dark with white upper decks. Several people are visible on the ship's deck and on the pier. The text "Как решить задачу в ML?" is overlaid in white.

Как решить задачу в ML?

Какие типы задач вообще бывают?

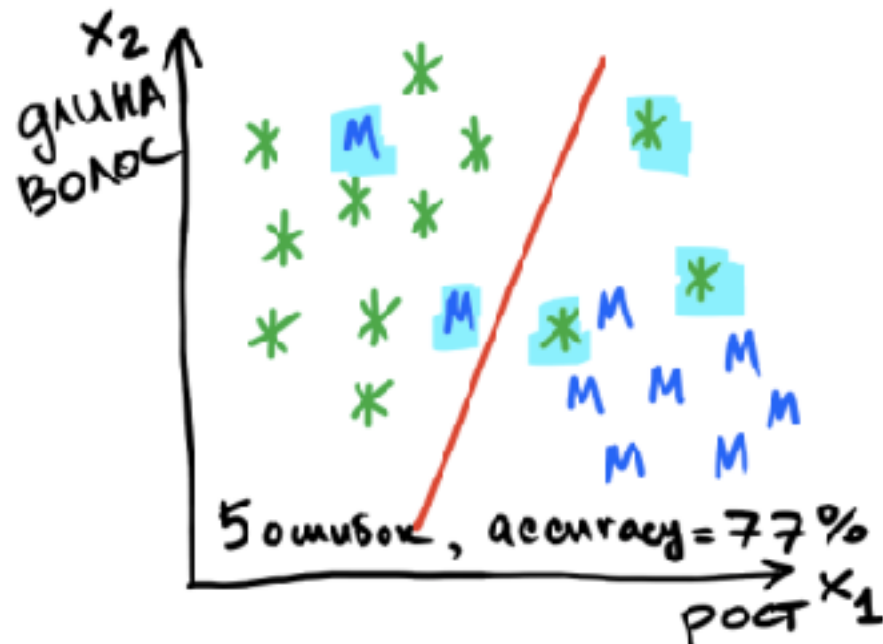
Есть разметка - “обучение с учителем”

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



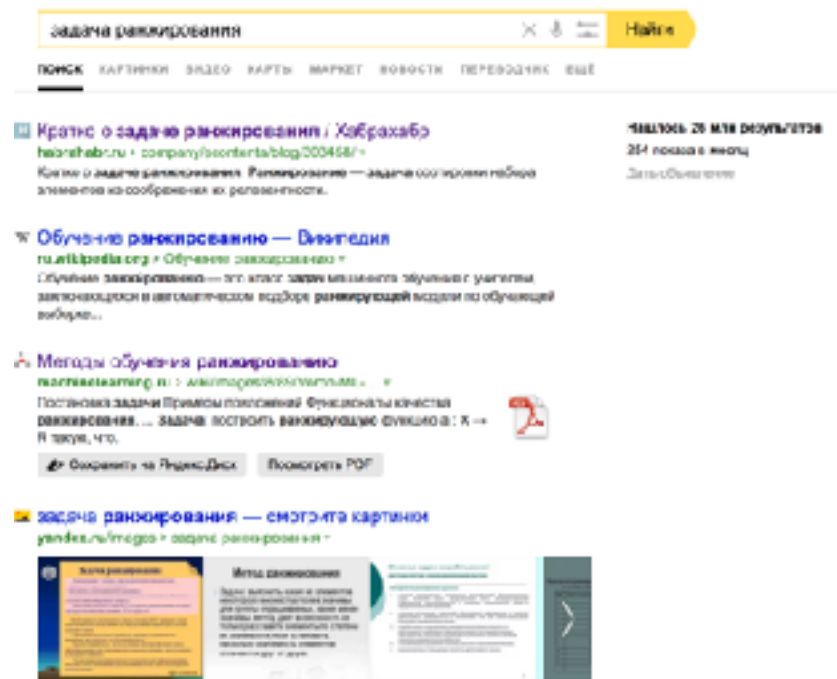
Предсказываем число —
регрессия

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



Предсказываем категорию —
классификация

Есть разметка - “обучение с учителем”



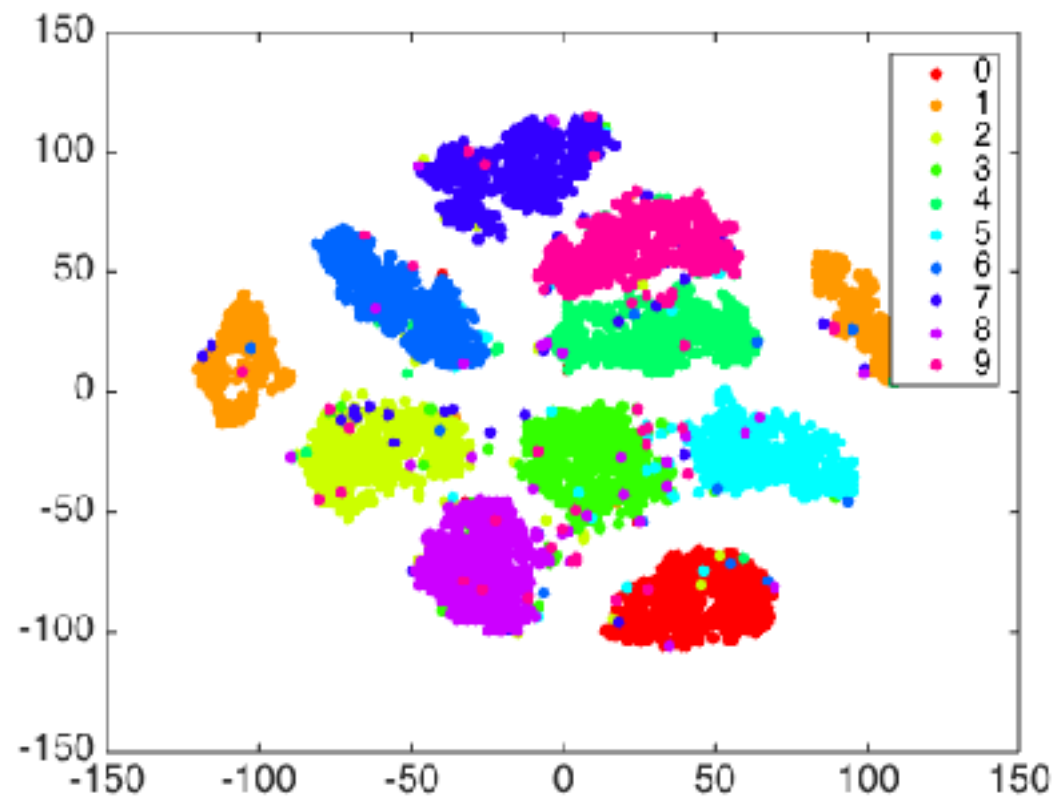
Предсказываем порядок элементов — ранжирование



	2			4	5	2.94*
	5		4			1
			5		2	2.48*
		1		5		4
			4			2
	4	5		1		1.12*

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

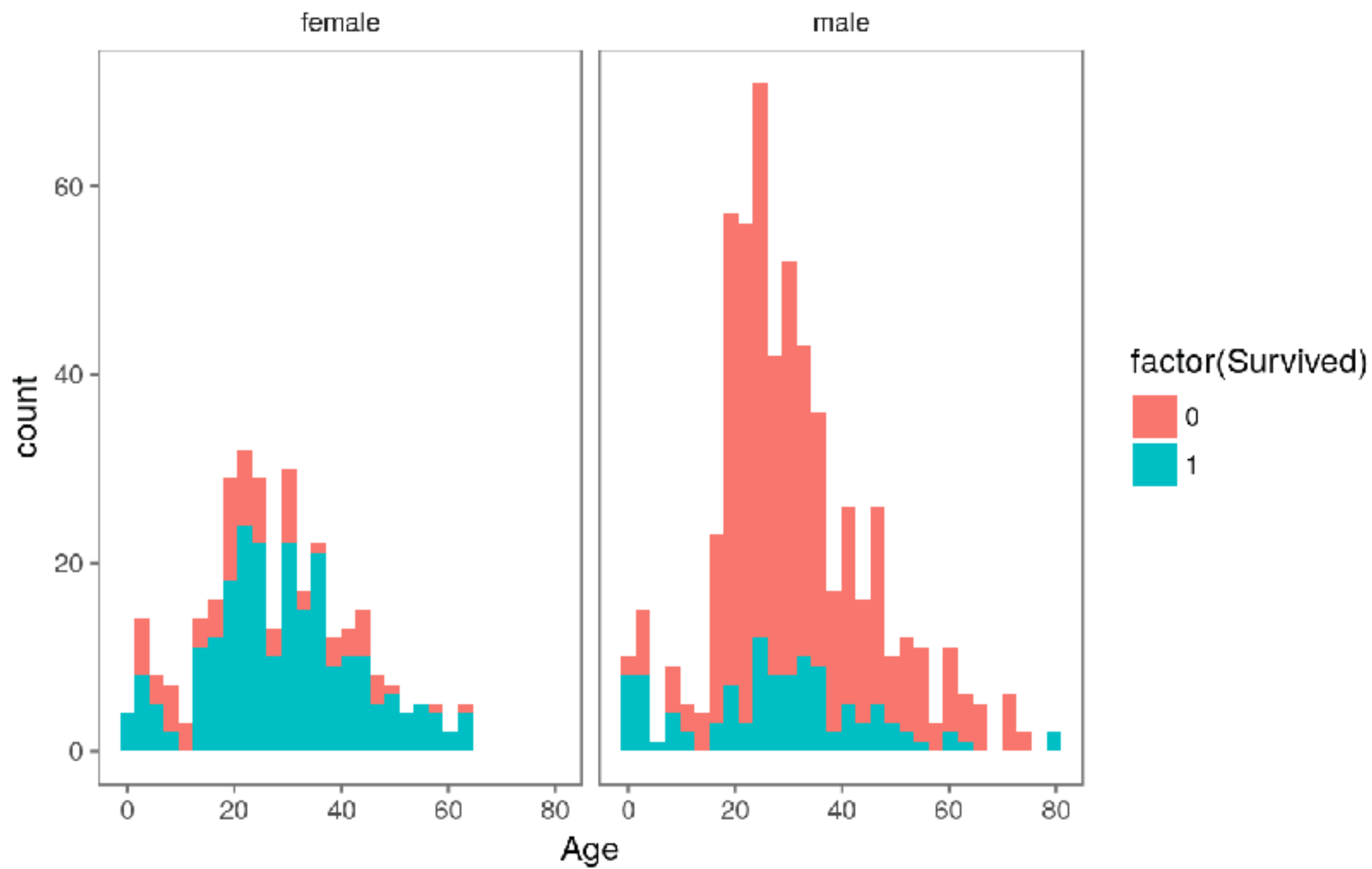
Нет разметки - “обучение без учителя”



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, ...

The screenshot shows the Kaggle website's 'Competitions' page. At the top, there's a navigation bar with links to Competitions, Datasets, Kernels, Discussion, and Jobs. Below this, a large blue banner reads 'Competitions' with 'Learn more' and 'InClass' buttons. The main content area is divided into 'General' and 'InClass' tabs. Under 'General', it shows '1 Entered Competition' and '15 Active Competitions'. The first active competition is 'Titanic: Machine Learning from Disaster', which is a 'Featured' competition. Below it, two other competitions are listed: 'Mercari Price Suggestion Challenge' with a prize of \$100,000 and 1,604 teams, and 'Statoll/C-CORE Iceberg Classifier Challenge' with a prize of \$50,000 and 3,438 teams.

The screenshot shows a Slack channel named 'opendatascience'. The channel has a green header bar with the name 'opendatascience' and a notification bell icon. Below the header, there's a search bar and a list of channels. The channels listed are: # gett_mck_bigdata_hack, # _call_4_collaboration, # _general, # _jobs, # _meetings, # _random_flood, # cancer, # datasets, # deep_learning, # edu_courses, # interesting_links, # kaggle_crackers, # lang_python, # mlcourse_open, # mltrainings_beginners, # mltrainings_live, # nlp, and # proj_kaggle_quora_qs.

The screenshot shows a Slack thread in the '# nlp' channel. The thread is titled 'All Threads' and has 'No new replies'. The thread is started by a user named 'yellowduck' 26 days ago. The message content is: 'Если бы вам дали задачу сделать эмбединги для текстов, как бы вы выстроили исследование?'. Below the message, there's a detailed response in Russian. The response discusses the task of creating text embeddings, mentions the need to get vectors, and talks about the challenges of dealing with noisy information in web pages. It also mentions the use of binary classes and the goal of getting vectors for specific classes and general usage.

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

Где настоящий лось?

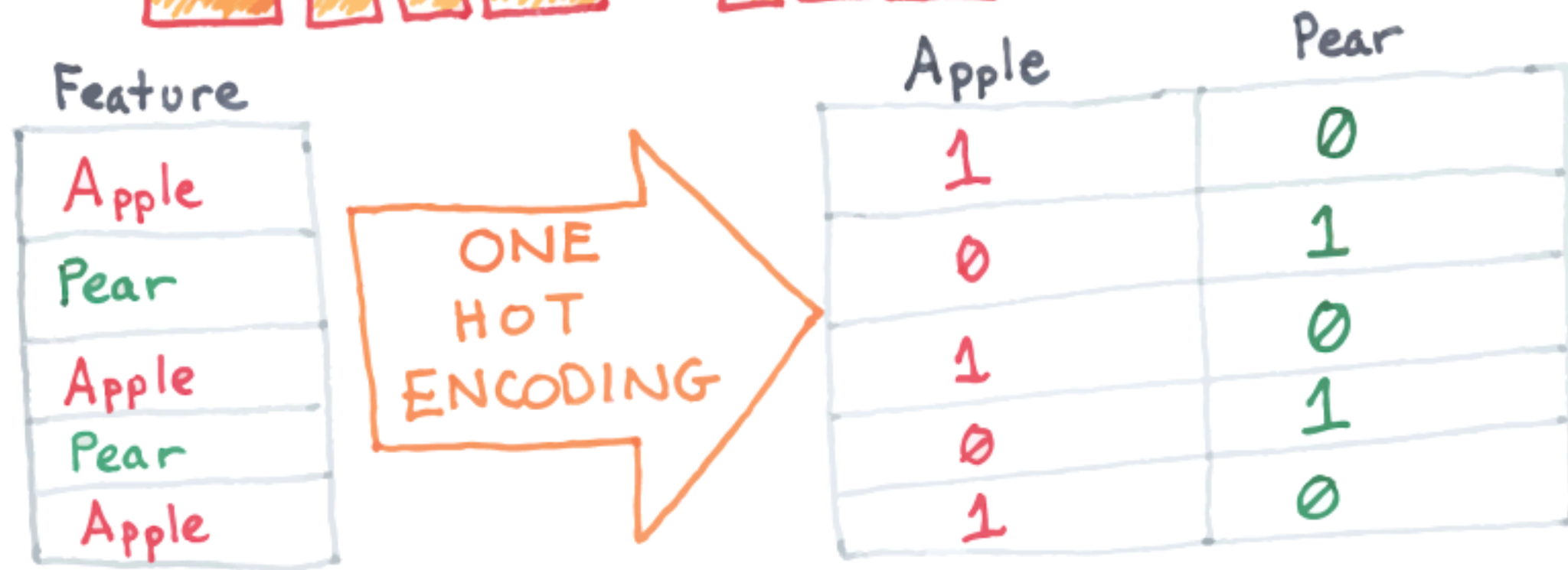


Яндекс.Толока
Яндекс



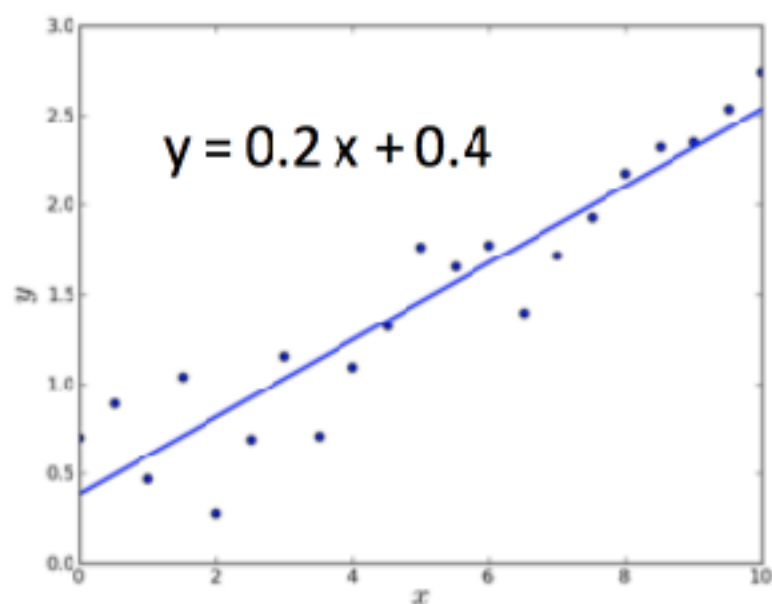
Нужна ручная разметка?
(минимум одна неделя)

ONE-HOT ENCODING



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

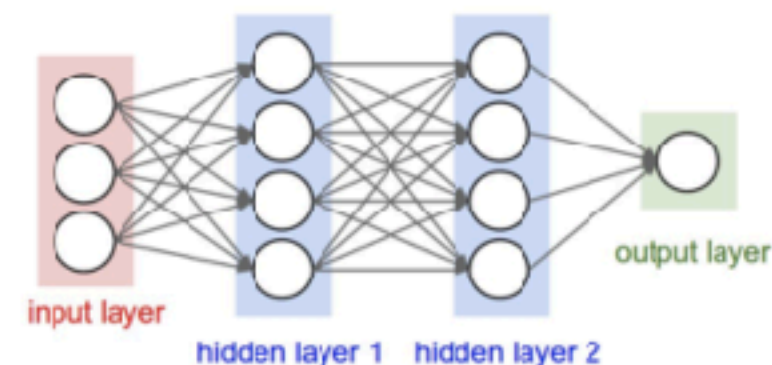
Линейные модели



Решающие деревья



Нейронные сети



PYTORCH
Deep Learning with PyTorch

 **Yandex
CatBoost**

 **scikit
learn**

 **TensorFlow™**

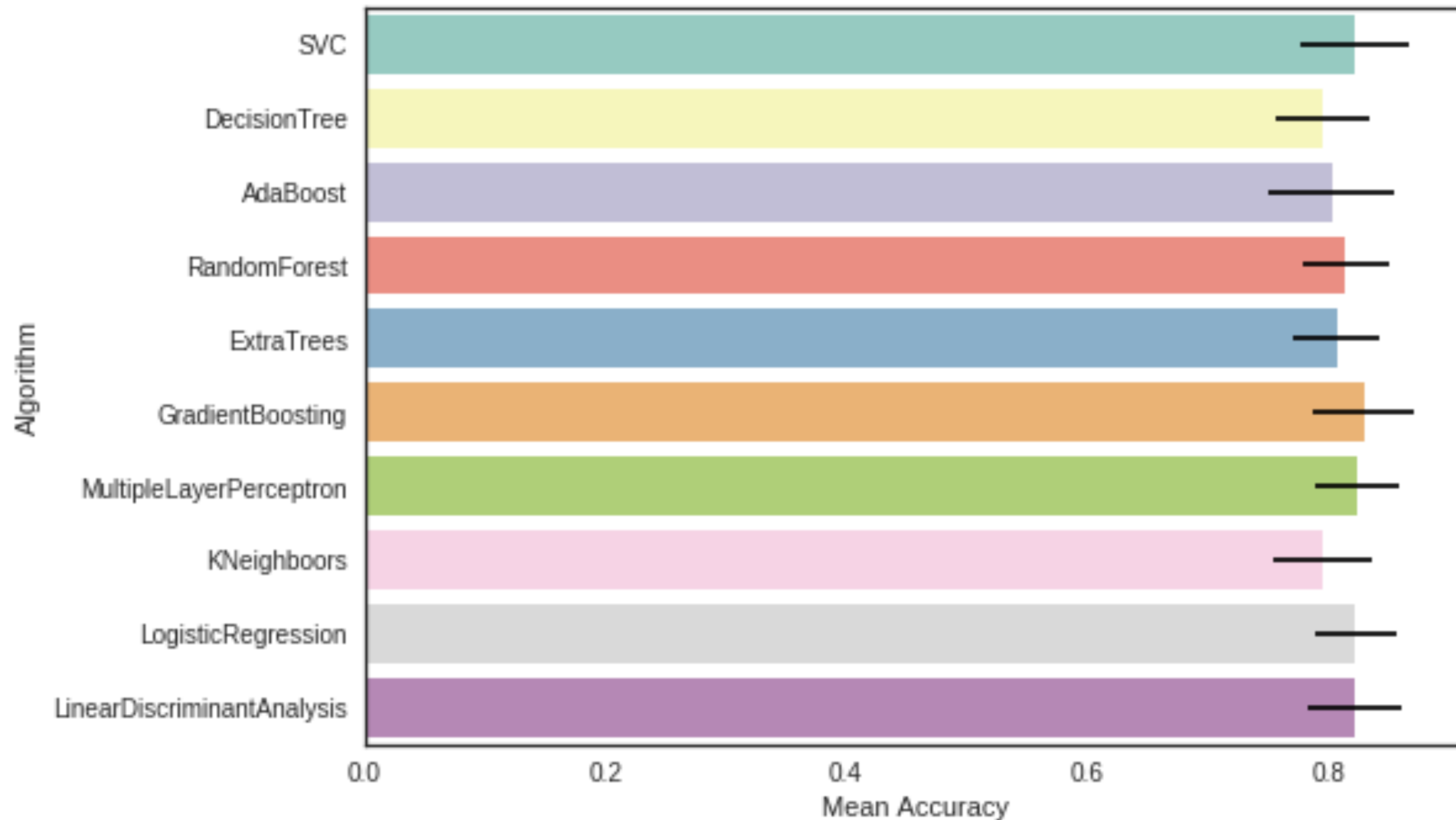
Обучаем модель

**dmlc
XGBoost**

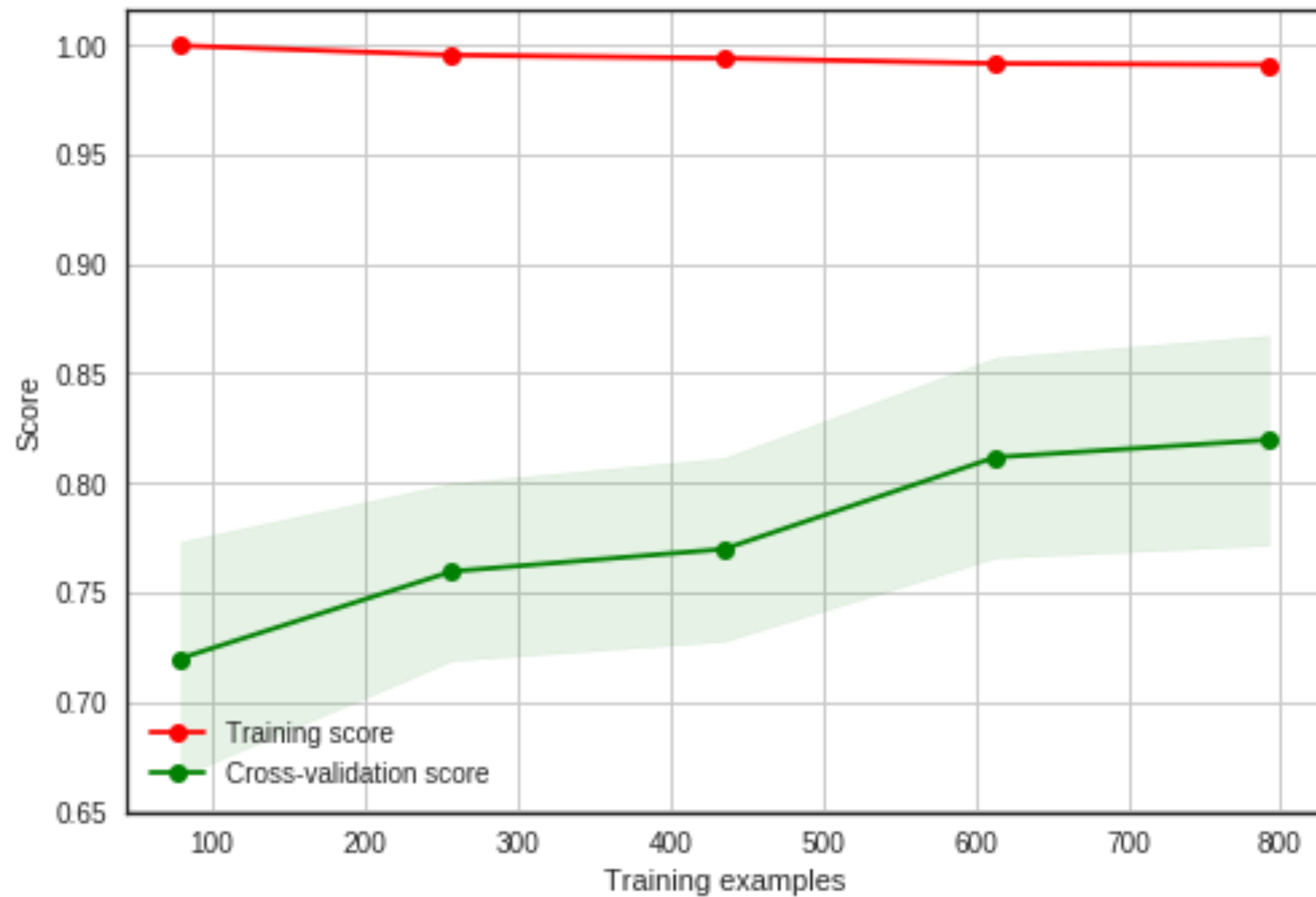
Accuracy = 88%

Получаем предварительный результат.
Часто работает принцип “20/80”.
(Где-то здесь прошли первые 2 недели + неделя на разметку)

Cross validation scores



Получаем предварительный результат.
Часто работает принцип “20/80”.
(Где-то здесь прошли первые 2 недели + неделя на разметку)

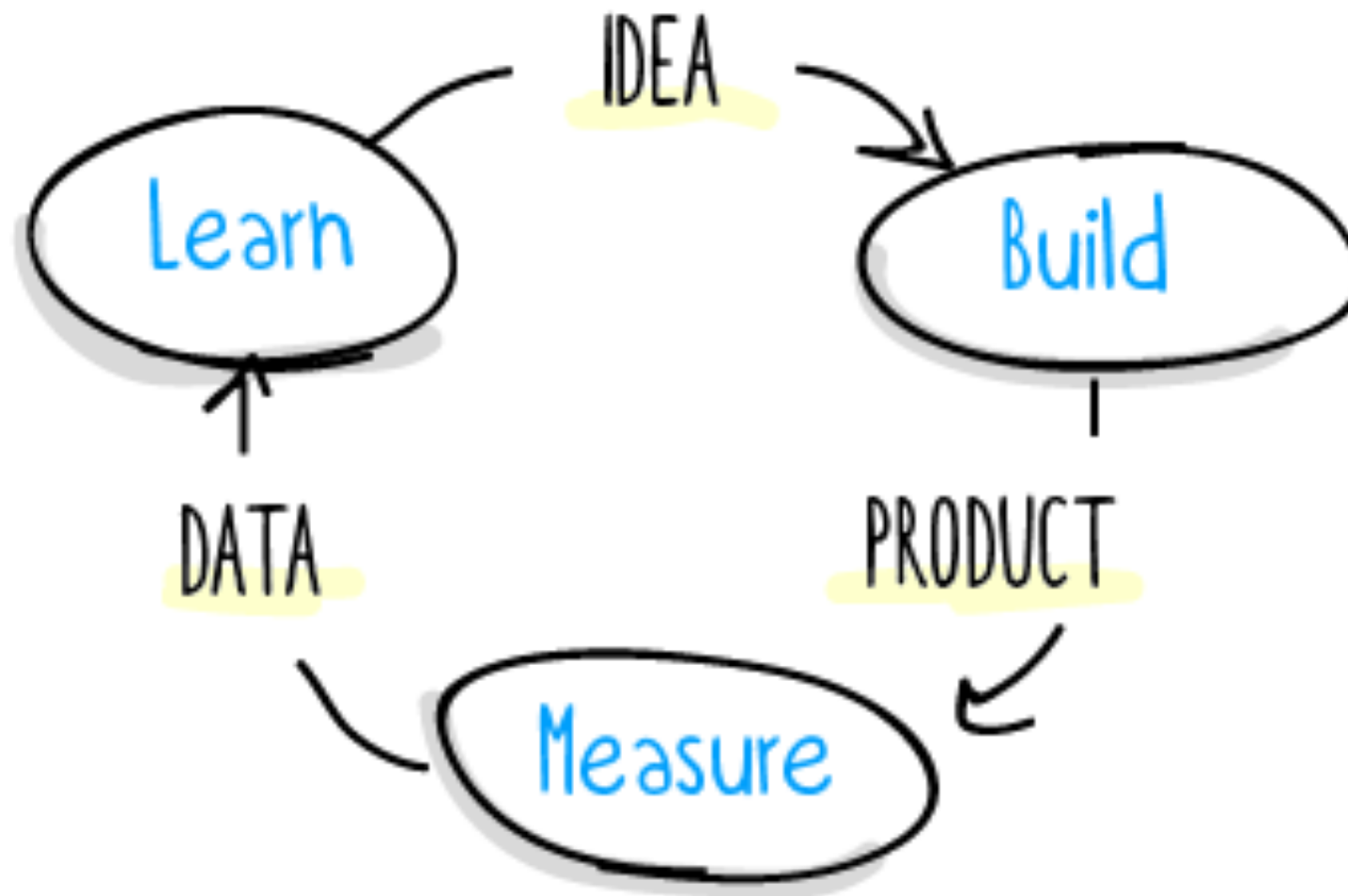


Может мало данных?

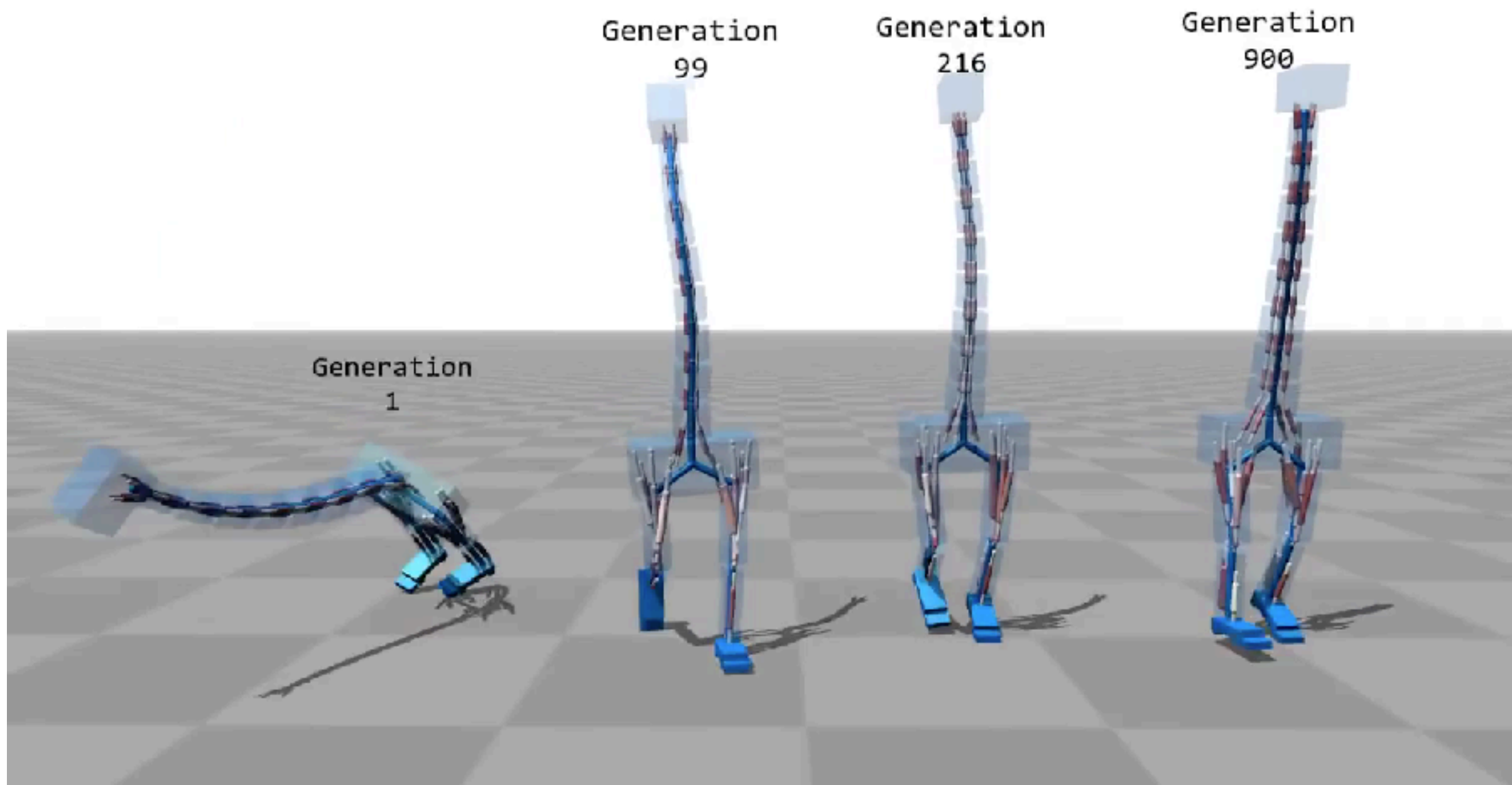


1997 Red River flood. Grand Forks

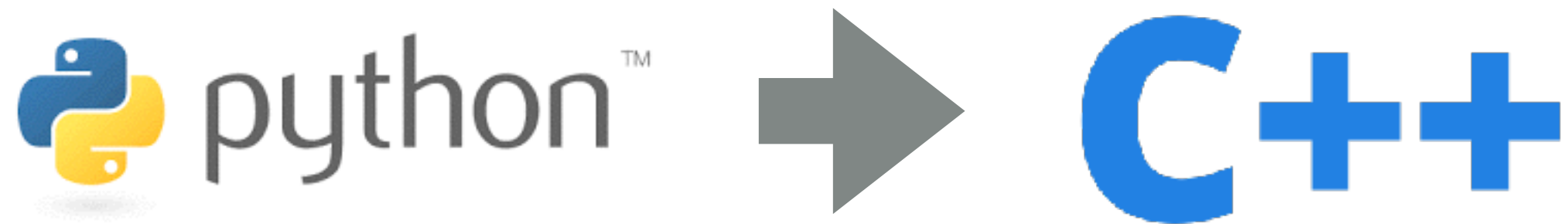
Иногда интересуется не только предсказание,
но и степень уверенности в нем



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



Тюним модель
(бесконечно долго)



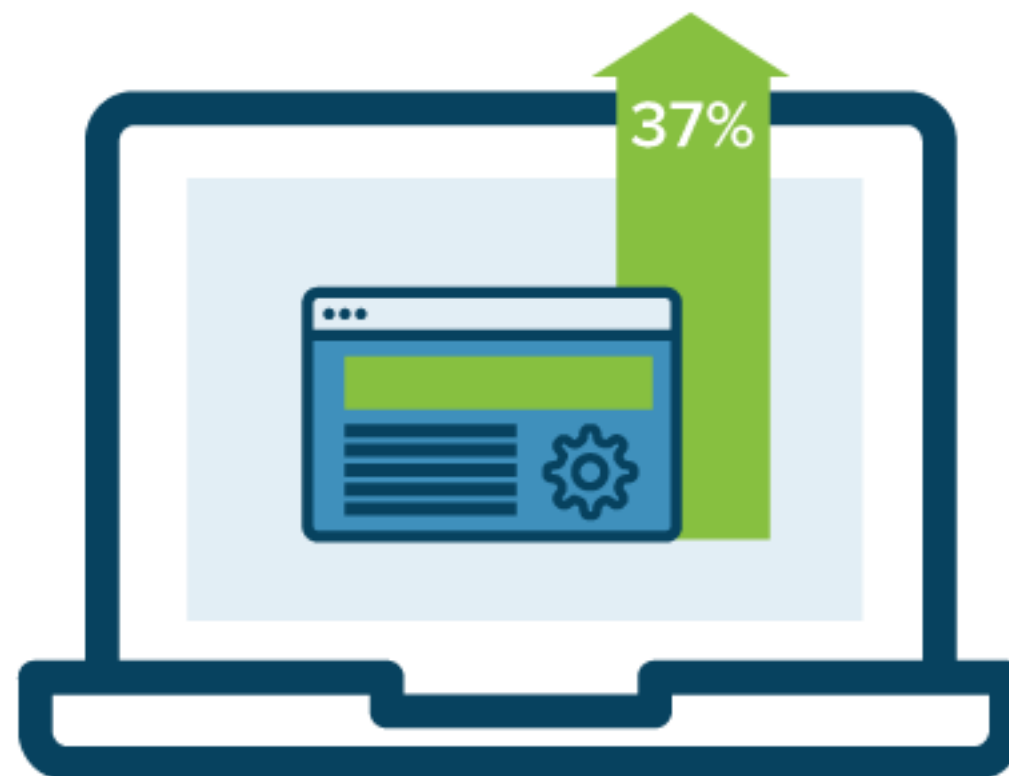
Интегрируем модель в продакшен
(обычно дольше, чем вы предполагали вначале)

A



CONTROL

B



VARIATION

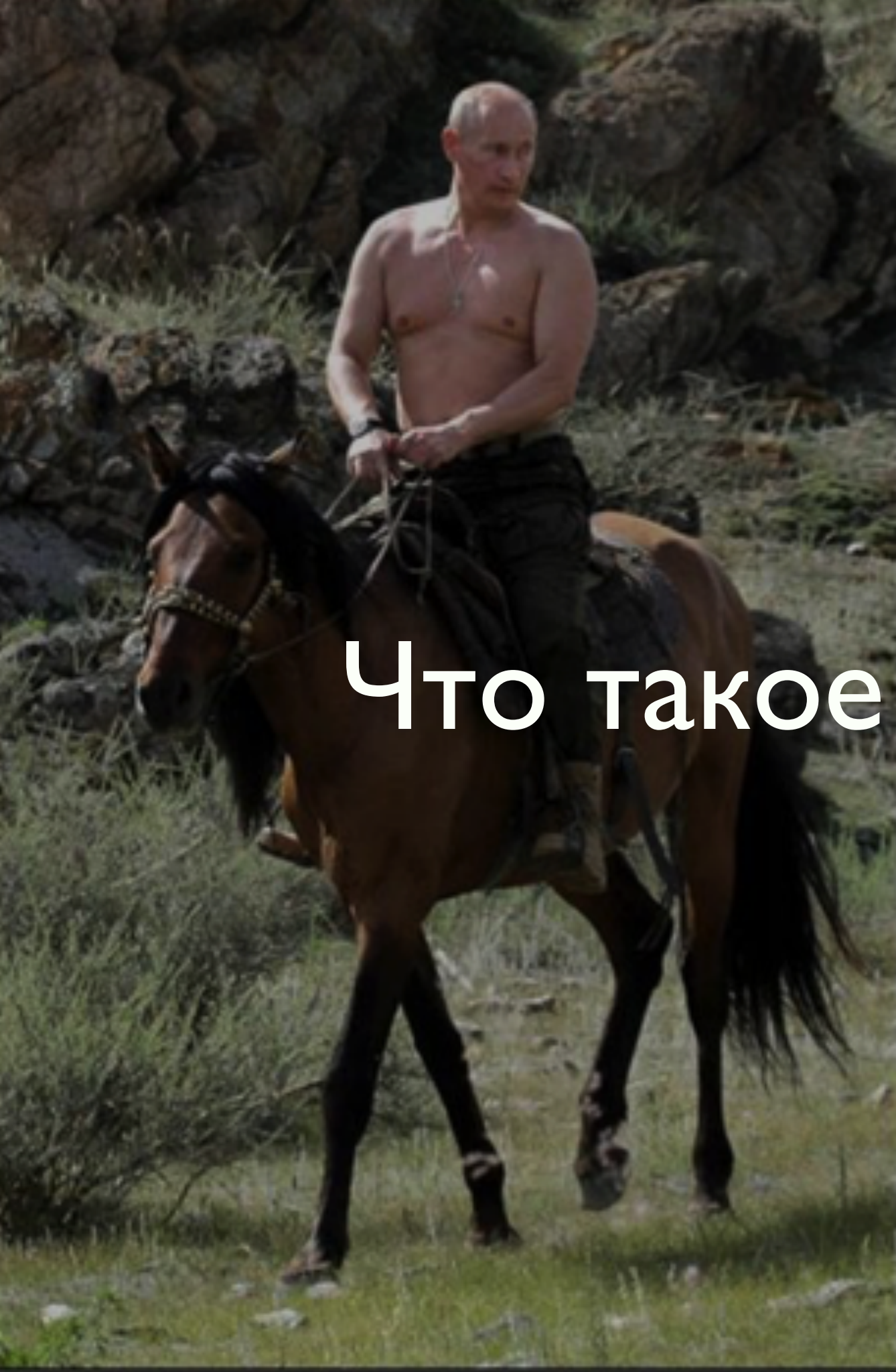
Тестируем на живом потоке
(минимум 2 недели)



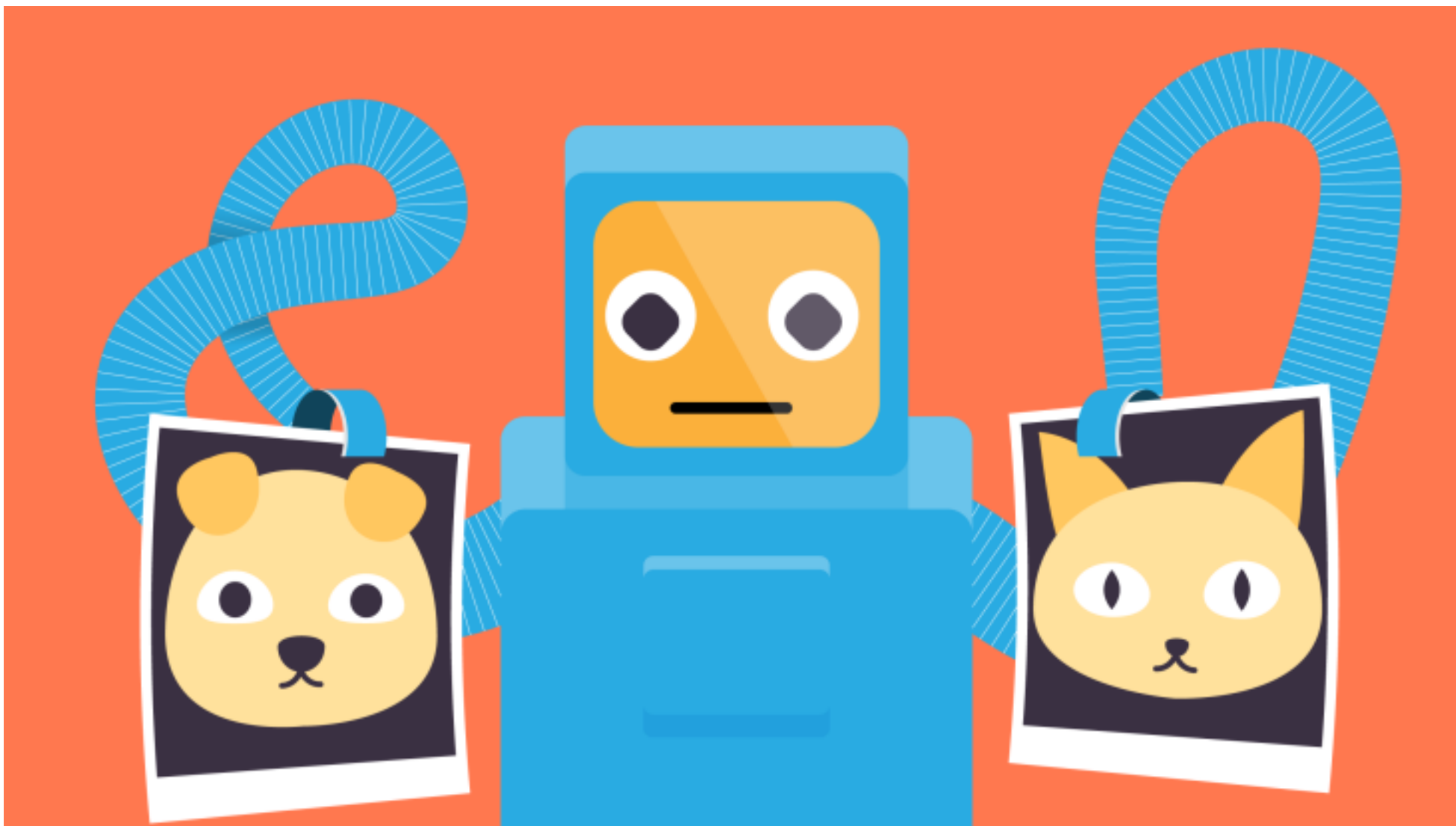
Когда всё в продакшене – работа еще не закончена!
Модели устаревают. Очень важно вести документацию и коммитить свой код

Workshop

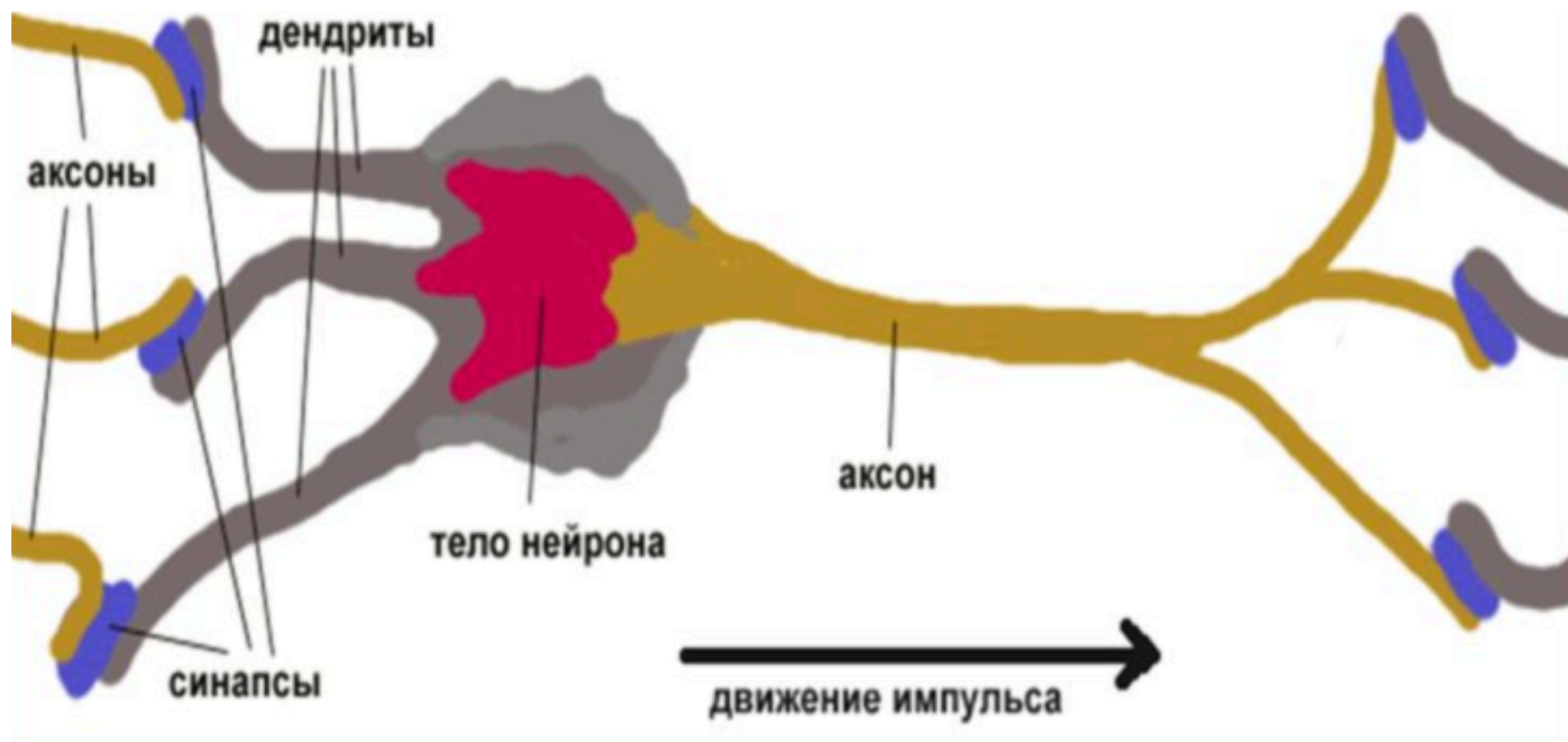
A detailed view of a workshop interior. The left wall is covered with numerous green and blue plastic storage bins arranged in rows. A red metal step ladder is positioned in the center-left. The background is filled with wooden shelves holding various tools, equipment, and materials. A workbench with a drill press is visible in the center. The floor is covered with wood shavings and debris. A large white tub is in the foreground on the left, and a grey plastic bin filled with wood scraps is in the foreground on the right. The lighting is warm and focused on the central area.



Что такое нейросети?

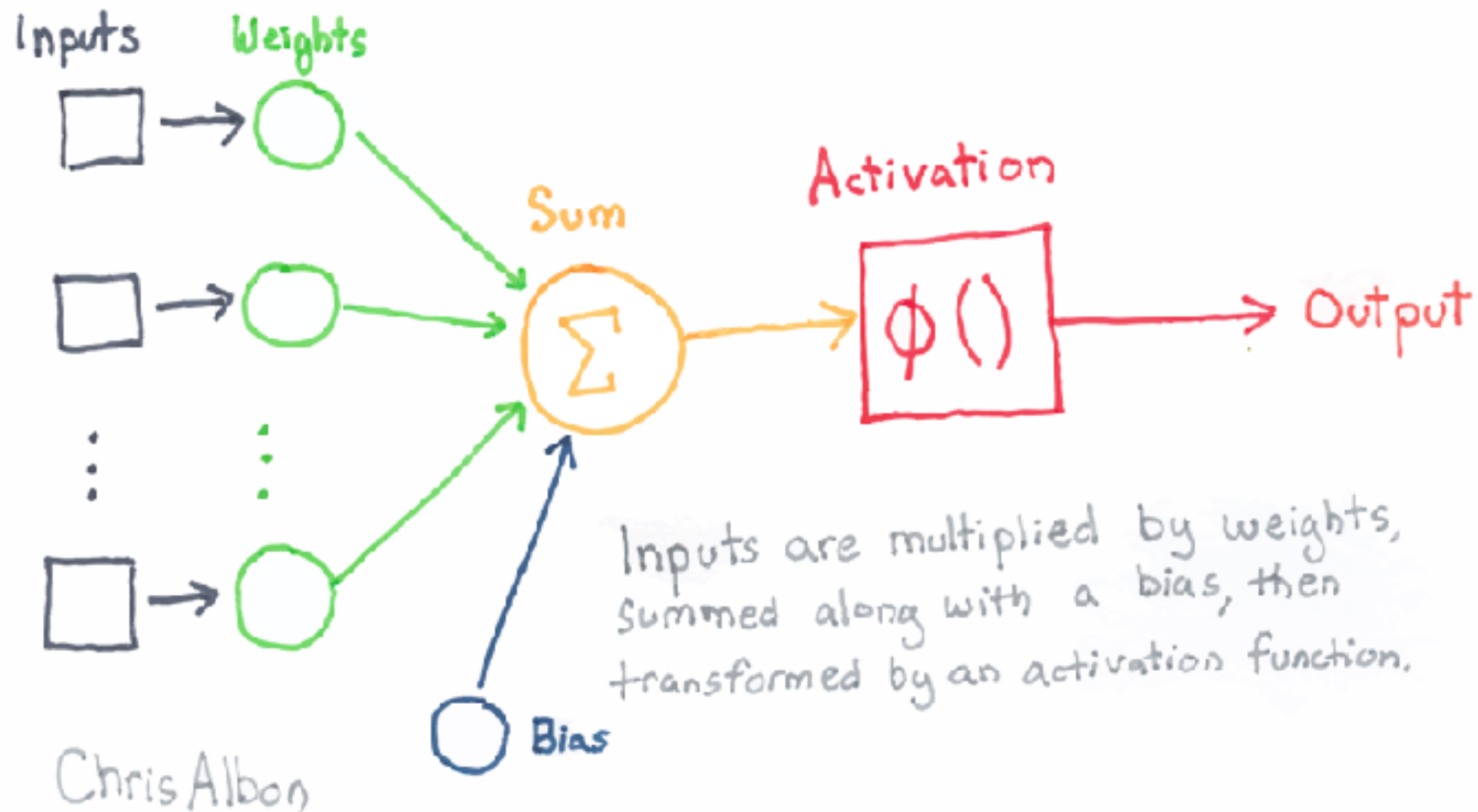


Нейросети хорошо работают **ТОЛЬКО** на
структурированных данных:
изображения, звук, текст, временные ряды



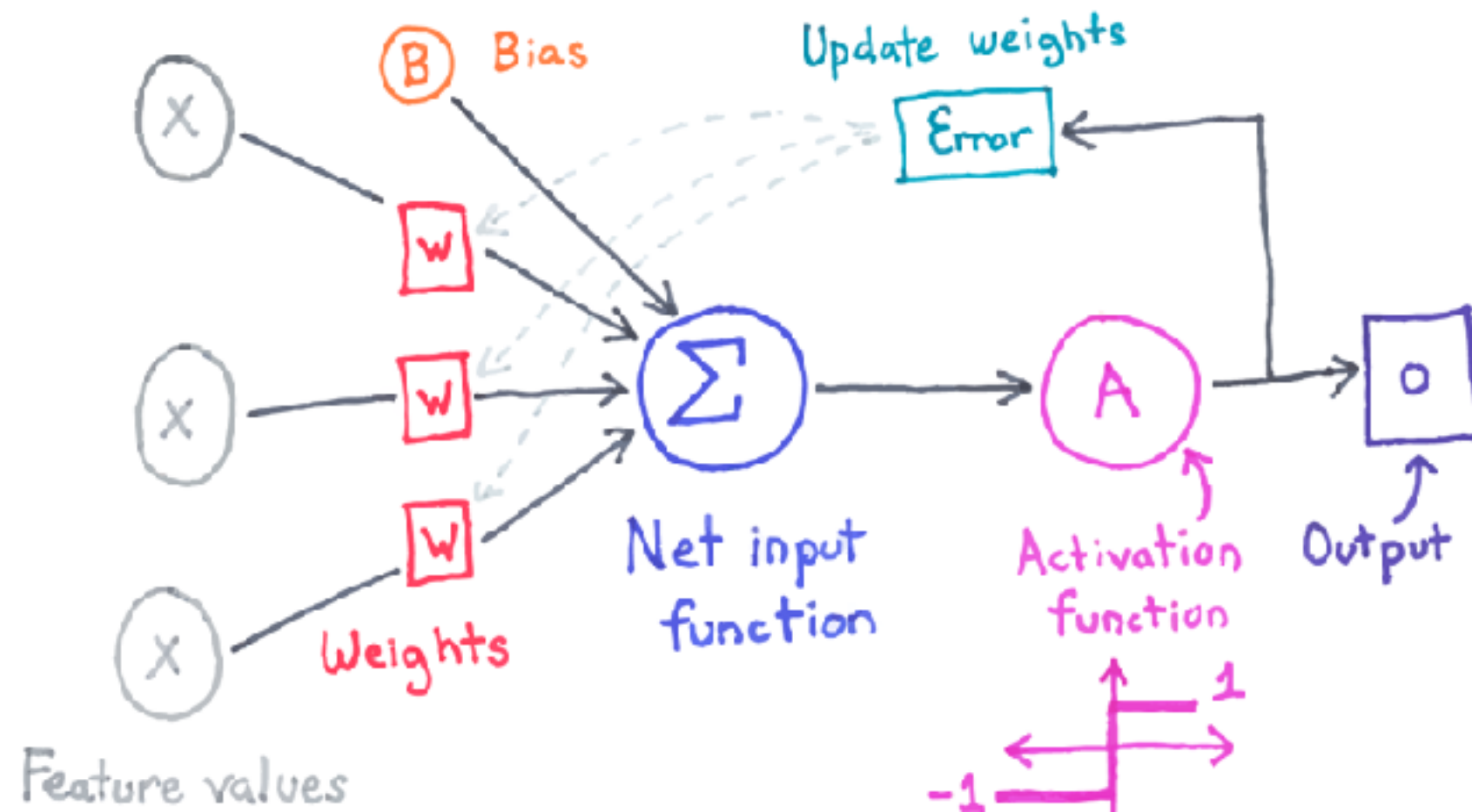
Биологический нейрон

NEURON

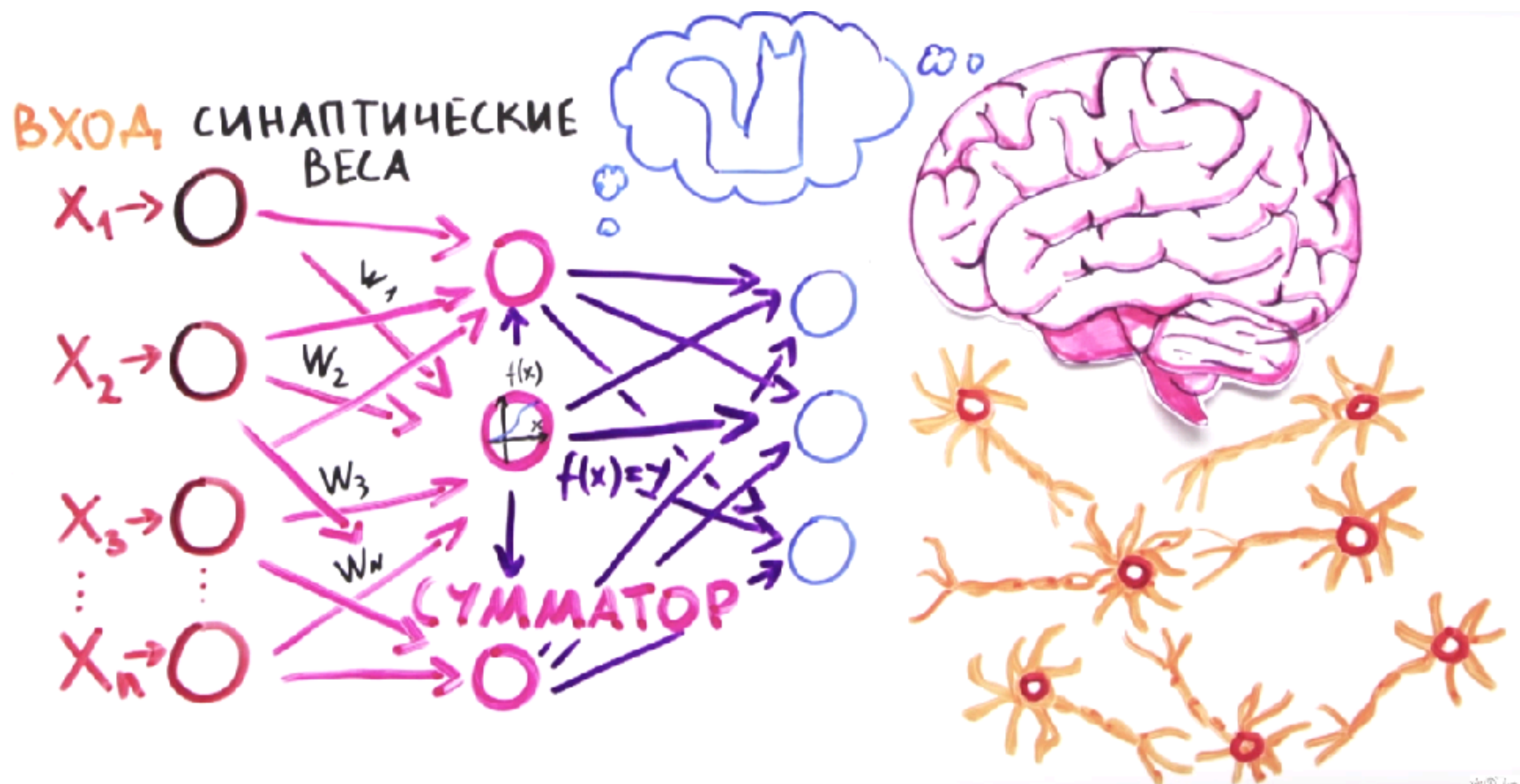


Математическая модель нейрона

PERCEPTRON



Нейросеть с одним нейроном



Нейронная сеть

Слайд для вопросов