Судя по тому, что вы решили почитать эту книгу, вам приходилось слышать о машинном обучении, пожалуй, сейчас сложнее не услышать это словосочетание, например, термин "Machine Learning" содержится в списке buzzword'ов английской википедии. Модные слова (также гламурная лексика и «умные слова», англ. buzzword) — особый род новых слов и речевых конструкций, часто используемых в коммерции, пропаганде и профессиональной деятельности для оказания впечатления осведомлённости говорящего и для придания чему-либо образа важности, уникальности или новизны. Из-за неумеренного употребления смысл слова размывается, и «модные слова» можно встретить даже в контексте, не имеющем отношения к исходному значению: например, «элитные семинары [слово, в какой-то момент заменившее «элитарные» — доступные самым богатым] по умеренным ценам», «эксклюзивные часы [изготовленные штучно], выпущенные тиражом в 11111 экземпляров». [2]

Машинное обучение попало в этот список не просто так, действительно часто в СМИ появляются новости, что ещё один рубеж покорился программе, иногда можно услышать апокалиптические сценарии о скором восстании машин или наоборот, идеалистические, о скором приходе технологической сингулярности и возможности перенесения сознания в машину или воссоздания сознания умерших людей. Мы же будем в стороне от всего этого ажиотажа и подойдём непосредственно к изучению основ машинного обучения.

Рис. 1: ReCAPTHCA первой версии

Вы наверняка помните то время, когда интернет сайты постоянно заставляли вас разбирать, что написано на данных картинках. Сайты требовали это, чтобы избежать злоупотребления функционалом сайта со стороны ботов, предупреждая таким образом регистрацию множества "пользователей отправку кучи сообщений или злоупотребления иным функционалом сайта или попытки перебора паролей.

Однако несколько лет назад вдруг подобные надписи пропали. Почему? Правоохранительные органы пересажали всех злоумышленников? Или они сами решили уйти в монастырь замаливать грехи. Вряд ли. Произошло то, что всегда происходит в таких случаях: всегда найдётся кто-то, кто захочет злоупотребить вашей системой во вред вам или во вред пользователям, ради кражи их данных, создания повышенной нагрузки на сервера и соответственно замедление или вообще прекращения их работы и других целей. Подобные люди никуда не делись и не денутся.

С появлением сайтов появились и скрипты, которые имитировали действия реальных пользователей, не являясь ими, создателям сайтов это не понравилось, поэтому они попытались отличить ботов от реальных пользователей. Первоначальной идеей было создать задание, с которым бы легко справлялись люди, но плохо справлялись роботы, такой задачей стала задача распознавания символов. Действительно для вас не составляет никакого труда разобрать, что значат все эти закорючки, линии и кружочки из которых состоят слова в этой книге (по крайней мере, я очень на это надеюсь). Но до недавних пор эта задача была для компьютеров совсем не тривиальной.

Так и появилось то, что называется капчей. Капча (от CAPTCHA — англ. Completely Automated Public Turing test to tell Computers and Humans Apart — полностью автоматизированный публичный тест Тьюринга для различения компьютеров и людей) — компьютерный тест, используемый для того, чтобы определить, кем является пользователь системы: человеком или компьютером. Термин появился в 2000 году. Основная идея теста: предложить пользователю такую задачу, которая с лёгкостью решается человеком, но крайне сложна и трудоёмка для компьютера. По состоянию на 2013 год, каждый день пользователями по всей планете вводится примерно 320 миллионов «капчей». [1]

Но разумеется злоумышленники (и не обязательно злоумышлен-

ники, но и многие исследователи из научного интереса) не сдались и решили попытаться обойти защиту. Мы рассмотрим хронологию обхода ReCAPTCHA - пожалуй самую известную реализацию, изображение которой и было представлено выше, на Рис. 1.

- 1 августа 2010 Chad Houck добился точности "взлома"капчи в 31.8%.
- 26 мая 2012 Adam, C-P and Jeffball - 99.1%. Следует уточнить, что ими была взломана аудио версия, при этом за несколько часов до презентации Google выпустил обновление точность упала до 60.95%.
- 27 июня 2012 - Мексиканские студенты Claudia Cruz, Fernando Uceda, and Leobardo Reyes добились точности 82%.
- Август 2017 - Kevin Bock, Daven Patel, George Hughey, Dave Levin - 85.15%.

Исследователи утверждали, что Гугл постоянно менял свою капчу, при этом часто возвращаясь к предыдущей версии[**wiki:captcha_is_hard**]. Однако, всё же сложность капчи постепенно росла в конце концов требование, с которым она создавалась, - быть сложной в решении для программ и лёгкой для людей, стало выполняться всё хуже. Так в августе 2012 – более 90% пользователей находят капчу сложной для ввода[**wiki:captcha_is_hard**]. В Май 2016 была прекращена поддержка данного типа капчи, то есть она перестала обновляться, в 31 марта 2018 выключена окончательно.