Lean ML: рациональный подход к машинному обучению

Андрей Законов

директор по росту и исследованиям,
Команда ВКонтакте

План

I. Когда нужен ML (машинное обучение) и когда не нужен?

План

I. Когда нужен ML (машинное обучение) и когда не нужен?

II. Как быстро и удачно реализовать ML?

План

I. Когда нужен ML (машинное обучение) и когда не нужен?

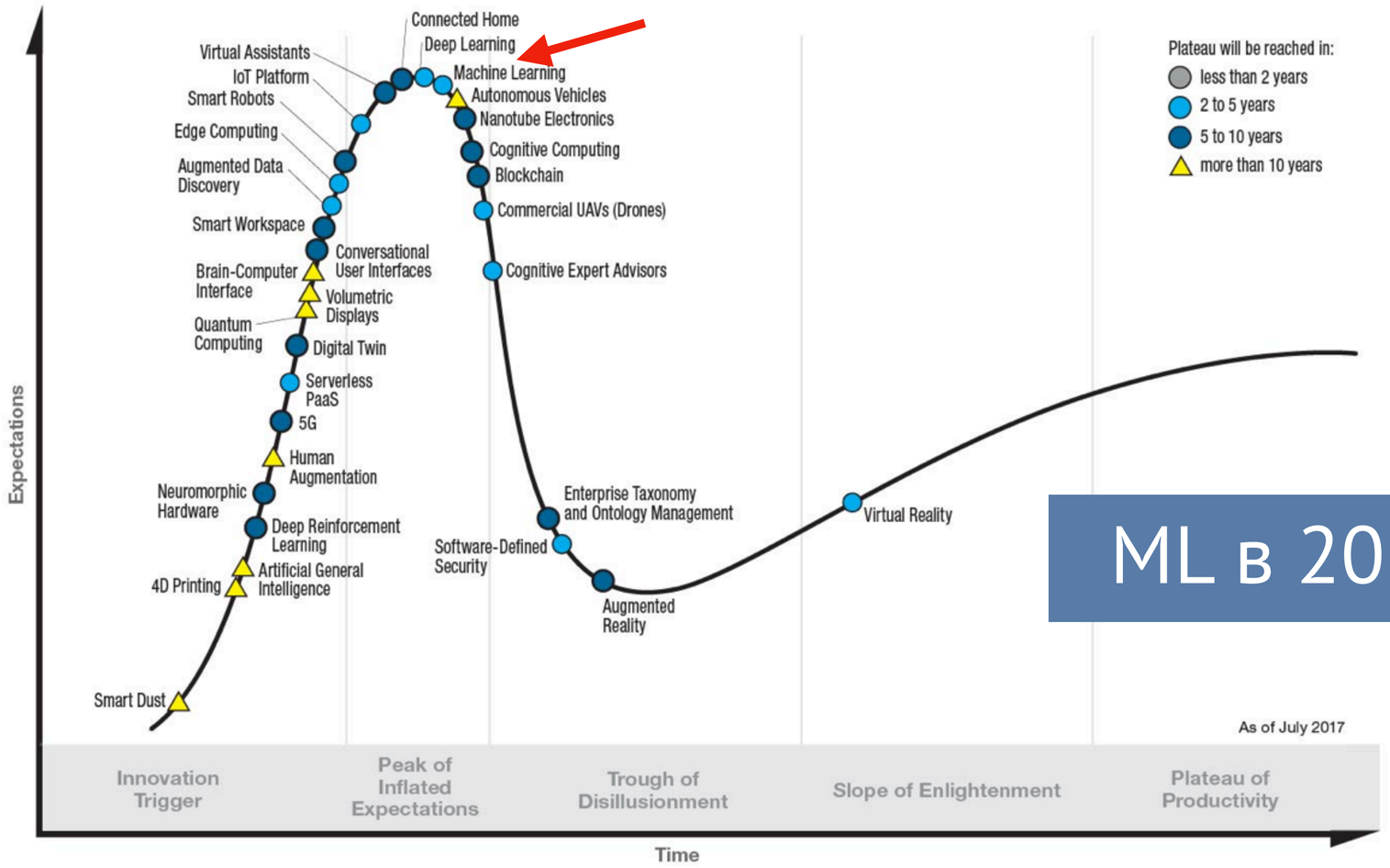
II. Как быстро и удачно реализовать ML?

III. Как запускать эксперименты и оценивать результаты?

97M пользователей в месяц	**6,5B** сообщений в сутки	**1B** отметок «Нравится» в сутки	**77%** пользователей мобильных платформ
500M просмотров видео в сутки	**86** языковых версий	**140** разработчиков	**8** дизайнеров

Источник: vk.com/about, сентябрь 2018.

Gartner Hype Cycle for Emerging Technologies, 2017



ML в 2018 = хайп

gartner.com/SmarterWithGartner

Source: Gartner (July 2017)

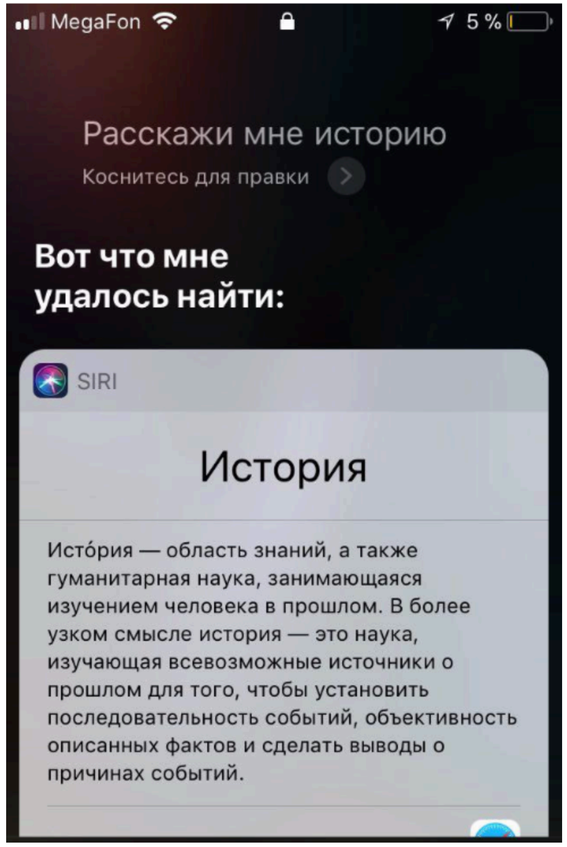
© 2017 Gartner, Inc. and/or its affiliates. All rights reserved.

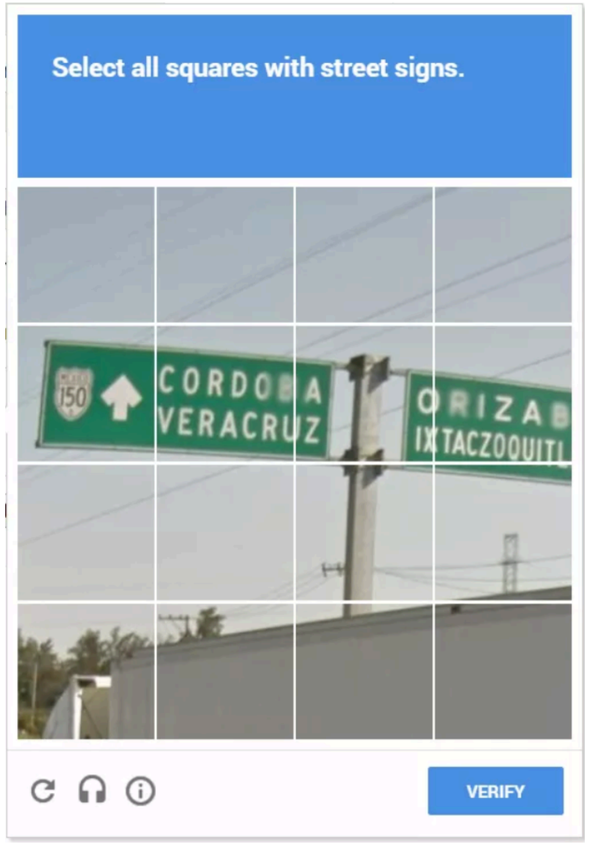
Gartner

пик чрезмерных ожиданий

Тестируем уровень технологий

Тестируем уровень технологий





reCAPTCHA v1

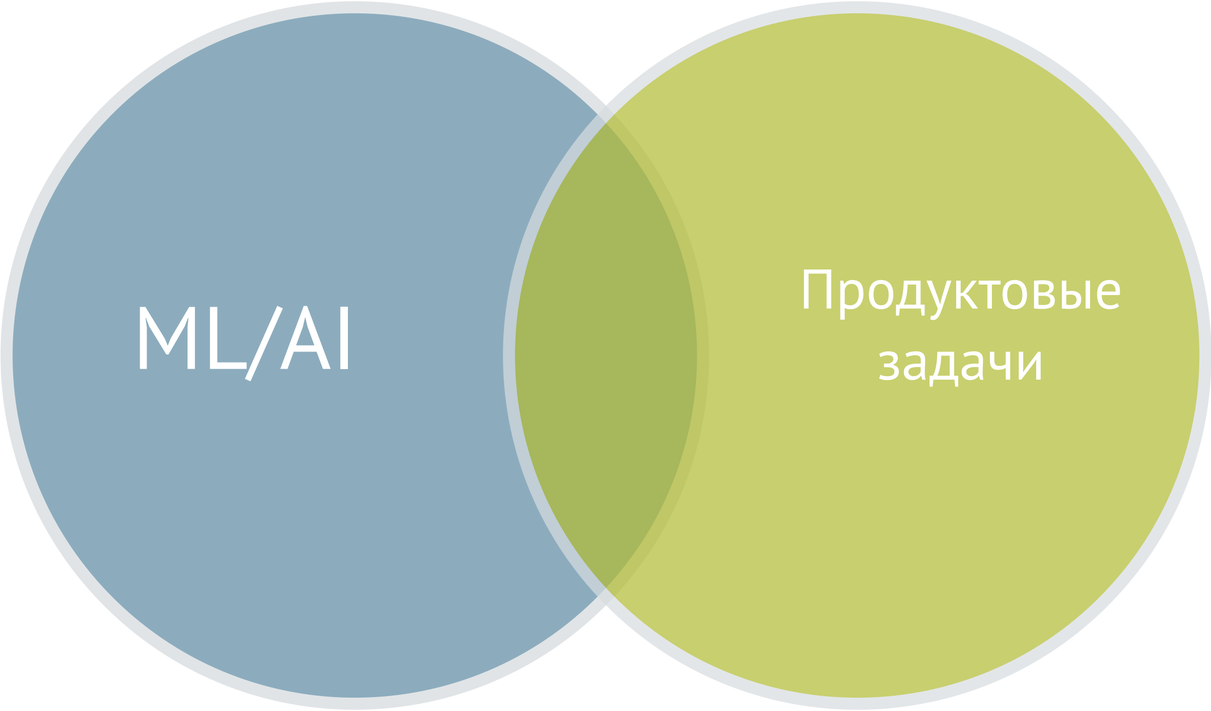
Быстрая проверка гипотез – решаем задачу вручную

AI-first продукты

AI-nice-to-have продукты

AI-first продукты

AI-nice-to-have продукты



Как определить язык пользователя?

Как определить язык пользователя?

Основная информация

Родной город:	Санкт-Петербург
Языки:	Русский, English

Выбираем продуктовую метрику и формулируем критерий успеха

Сначала интересные

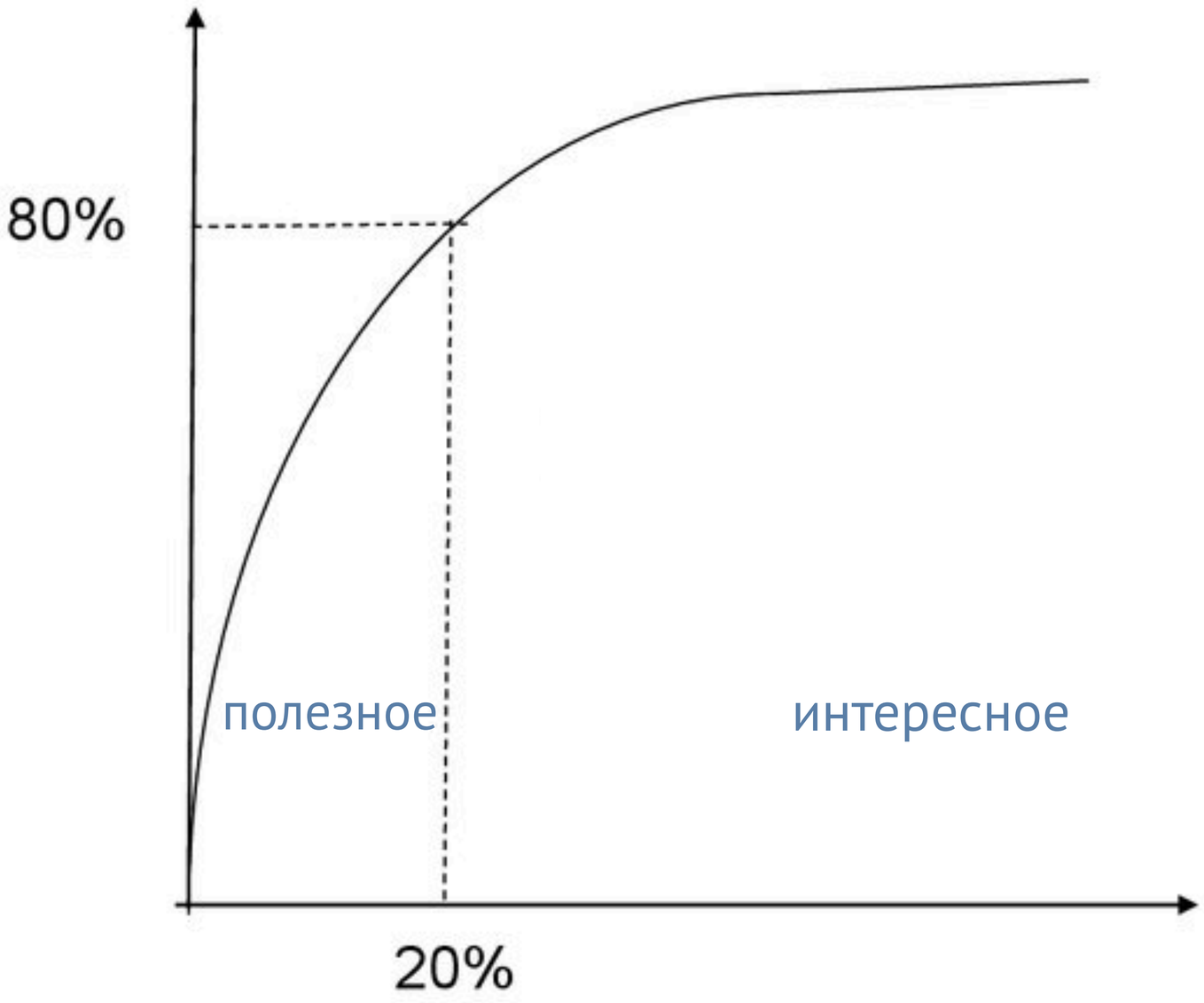
План

I. Когда нужен ML (машинное обучение) и когда не нужен?

II. Как быстро и удачно реализовать ML?

III. Как запускать эксперименты и оценивать результаты?

Ищем лучшие 20%



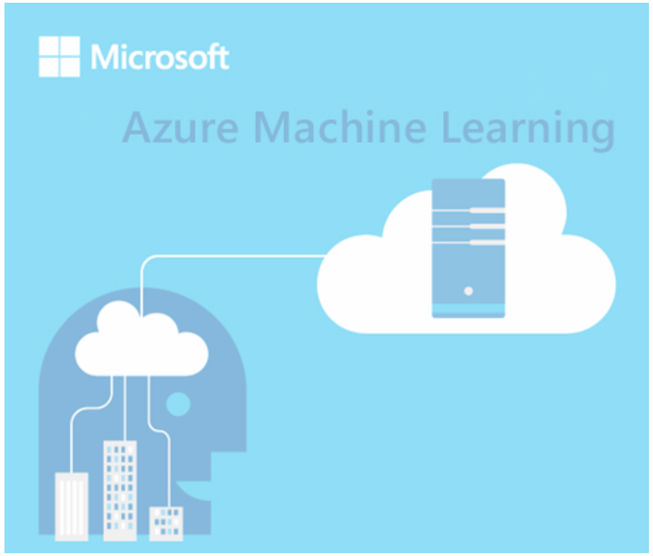
BIG DATA LANDSCAPE 2017

Изучаем готовые решения, помним про OpenSource

V2 – Last updated 5/3/2017

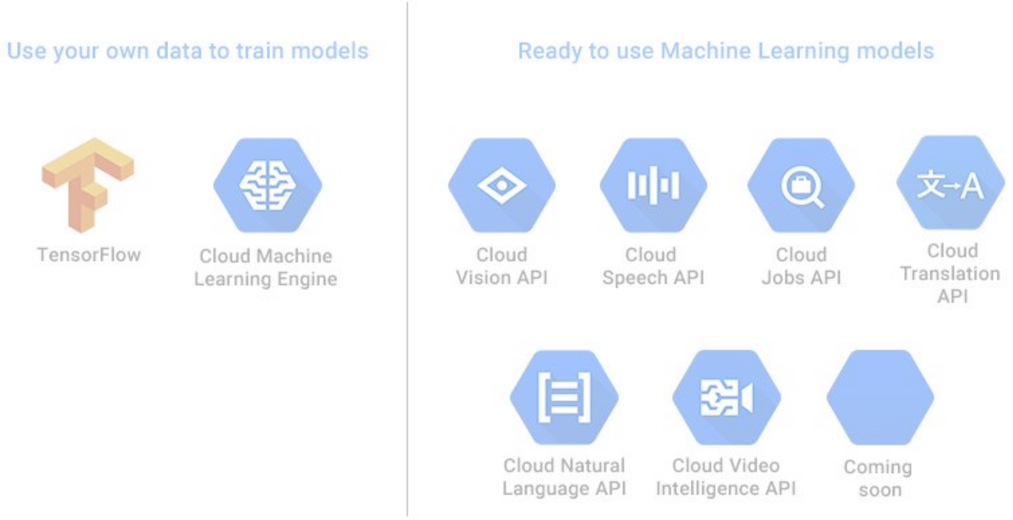
© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap) mattturck.com/bigdata2017

FIRSTMARK
EARLY STAGE VENTURE CAPITAL



ALGORITHMIA

ML as a Service



amazon web services

Amazon Machine Learning

aws.amazon.com/machine-learning

Keep it simple (если решает задачу)

```
In [ ]: import fasttext
        classifier = fasttext.supervised('data.train.txt', 'model')
        texts = ['example very long text 1', 'example very longtext 2']
        labels = classifier.predict(texts)
        print(labels)
```

https://facebook.ai/developers/tools/fasttext

Сразу формулируем ограничения продакшн

- Ограничения на
  - Время работы
  - Ресурсы
- Это повлияет на выбор технологий

Оптимизировать вокруг ML

- Кэшируем
- Предобсчитываем
- Делаем разные уровни сложности/отсечек

Пример: распознавание текста

- Проверяем кэши
- Проверяем наличие текста
- Ограничиваем область изображения
- По внешним факторам предсказываем язык
- И только теперь запускаем распознавание

Усложняем шаг за шагом

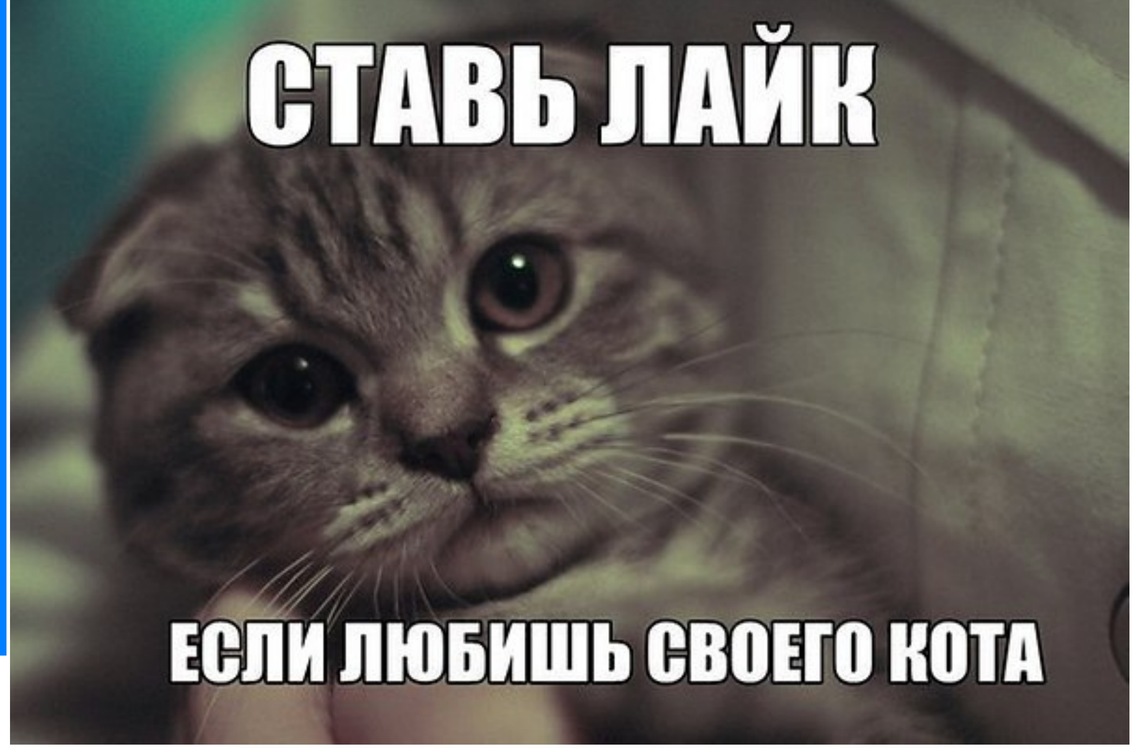
1. Запускаемся без ML
2. Берем простую метрику и простое решение
3. Проверяем новые гипотезы простыми экспериментами
4. Усложняем метрики
5. Усложняем модели
6. Читаем статьи
7. AI Research

План

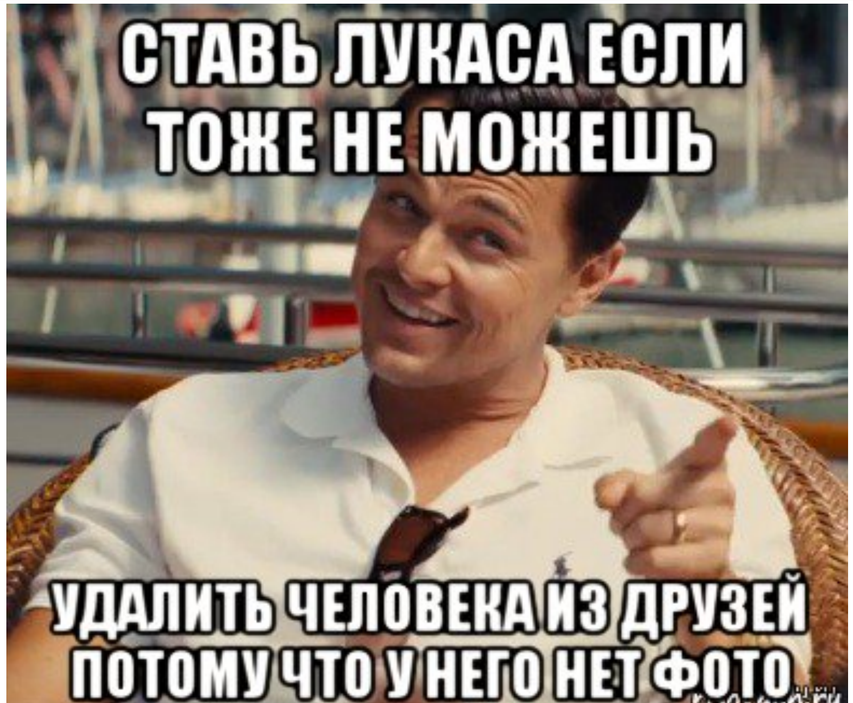
I. Когда нужен ML (машинное обучение) и когда не нужен?

II. Как быстро и удачно реализовать ML?

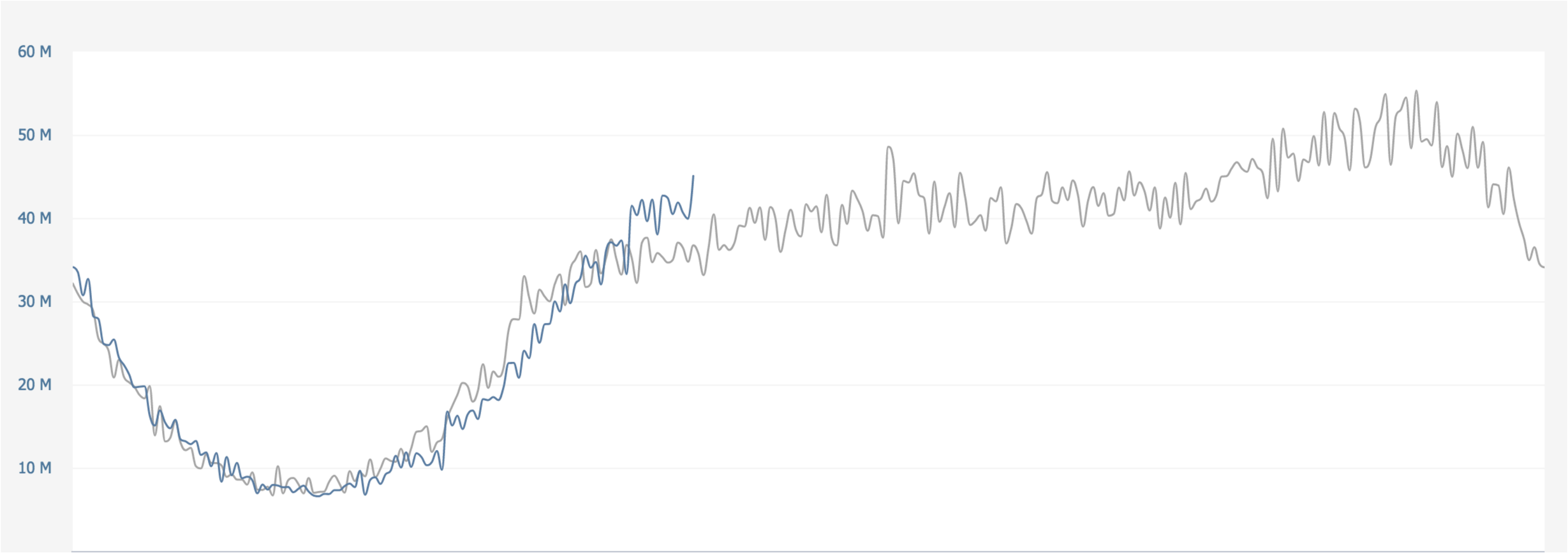
III. Как запускать эксперименты и оценивать результаты?



Будет интереснее, чем вы думаете

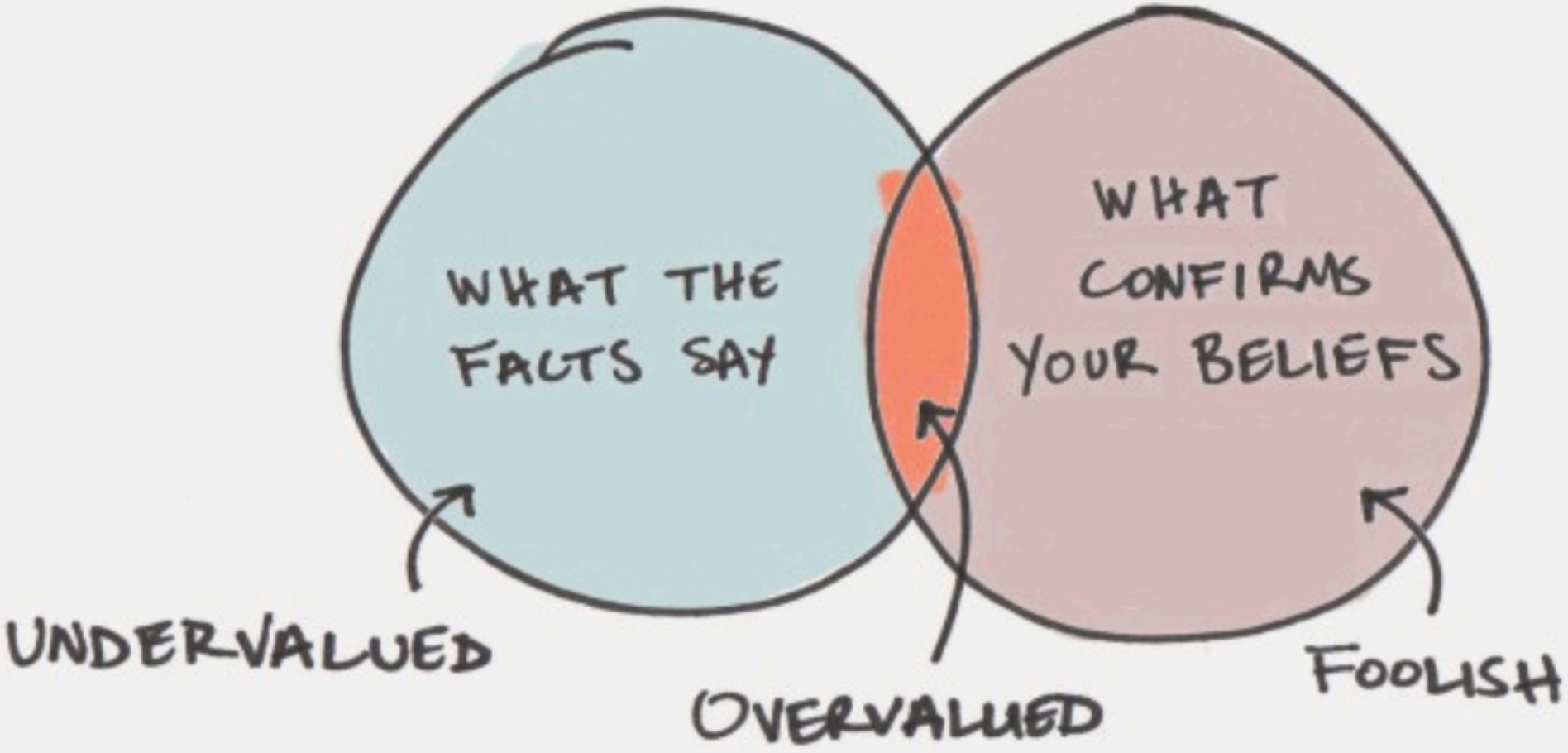


Что будем считать успехом?
Определяем до начала эксперимента


60 M
50 M
40 M
30 M
20 M
10 M

Считаем стат. значимость и делаем A-A-B-B

THE CONFIRMATION BIAS



Shop Back to School deals

Account & Lists Orders Try Prime 0 Cart

Sort b

- ✓ Featured
- Price: Low to High
- Price: High to Low
- Avg. Customer Review
- Newest Arrivals

Wall Bullet

Lowest Priced

Skateboarding Wall Bullet snowboard wall mount display hanger rack

short term != long term

привыкание, wow-эффект, насыщение

Data-Driven

Data-Informed

Summary

1. Разумно оцениваем пользу ML в задаче
2. Ищем 20% усилий, которые дадут 80% результата
3. Критично оцениваем результаты

ВКонтакте с вами

vk.com/andrewz