Демократизация ML

План рассказа

- Часть 1. Пять миров
  в которой мы вспомним, что машинленинг бывает разный.
- Часть 2. Алхимия
  в которой мы вытащим пару древних артефактов,
  оставшихся с времен, когда люди относились к жизни проще.
- Часть 3. Нормально делай - нормально будет
  в которой мы поговорим про машинленинг курильщика.
- Часть 4. Аутсорсь это
  в которой мы напомним, что сколько волка не корми, ишак выносливее.
- Часть 5, заключительная
  ...И немедленно выпил.

Пять миров разработки ПО

JOEL ON SOFTWARE

YOUR HOST

Five Worlds

ROCK STAR DEVELOPER, NEWS

I think there are five worlds here, sometimes intersecting, often not. The five are:

1. Shrinkwrap
2. Internal
3. Embedded
4. Games
5. Throwaway

https://www.joelonsoftware.com/2002/05/06/five-worlds/

Пять миров ML

- Ученые
- Спортсмены
- Стартапы
- Продуктовые компании
- АСУ

Все знают, что такое АСУ?

Сидит в комнатах, заставленных серверами. Совмещают сисадминство с ERP-разработкой. К 35 уходят или становятся важнейшим бизнес-активом
(иногда попадают в рай и становятся начальниками отдела или тех директорами)

Пять миров ML - 2

	Ресурсы	Мультипликатор	Ответственность
Ученые	+	0	0
Спортсмены	+	0	0
Стартапы	++	+++	+
Продуктовые	+++	++	+++
АСУ	+	+	+++

Ответственность — как у инженеров Гугла, ресурсов как у спортсмена.

Большинство красивых рассказов и статей про ученых, спортсменов, продуктовиков.

Конкурентные преимущества доставляют АСУ шники, т. к.:
- Ученые светят всем поровну
- Продуктовики и стартаперы меняют ландшафт для всех
- АСУшники работают только на компанию
- Можно бежать медленнее, чем медведь, но быстрее, чем конкурент.

Пять миров ML -3

Попробовал
Продал
Запустил

Алхимия

- Ali Rahimi, NIPS 2017:

  Machine Learning has become Alchemy
  https://www.youtube.com/watch?v=Qi1Yry33TQE

- Yann LeCun:

  I fundamentally disagree with the message.
  The main message was, in essence, that the current practice in machine learning is akin to "alchemy" (his word).
  It's insulting, yes. But never mind that: It's wrong!
  …..
  Criticizing an entire community (and an incredibly successful one at that) for practicing "alchemy", simply because our current theoretical tools haven't caught up with our practice is dangerous.

  https://www.facebook.com/yann.lecun/posts/10154938130592143

Что это за Али

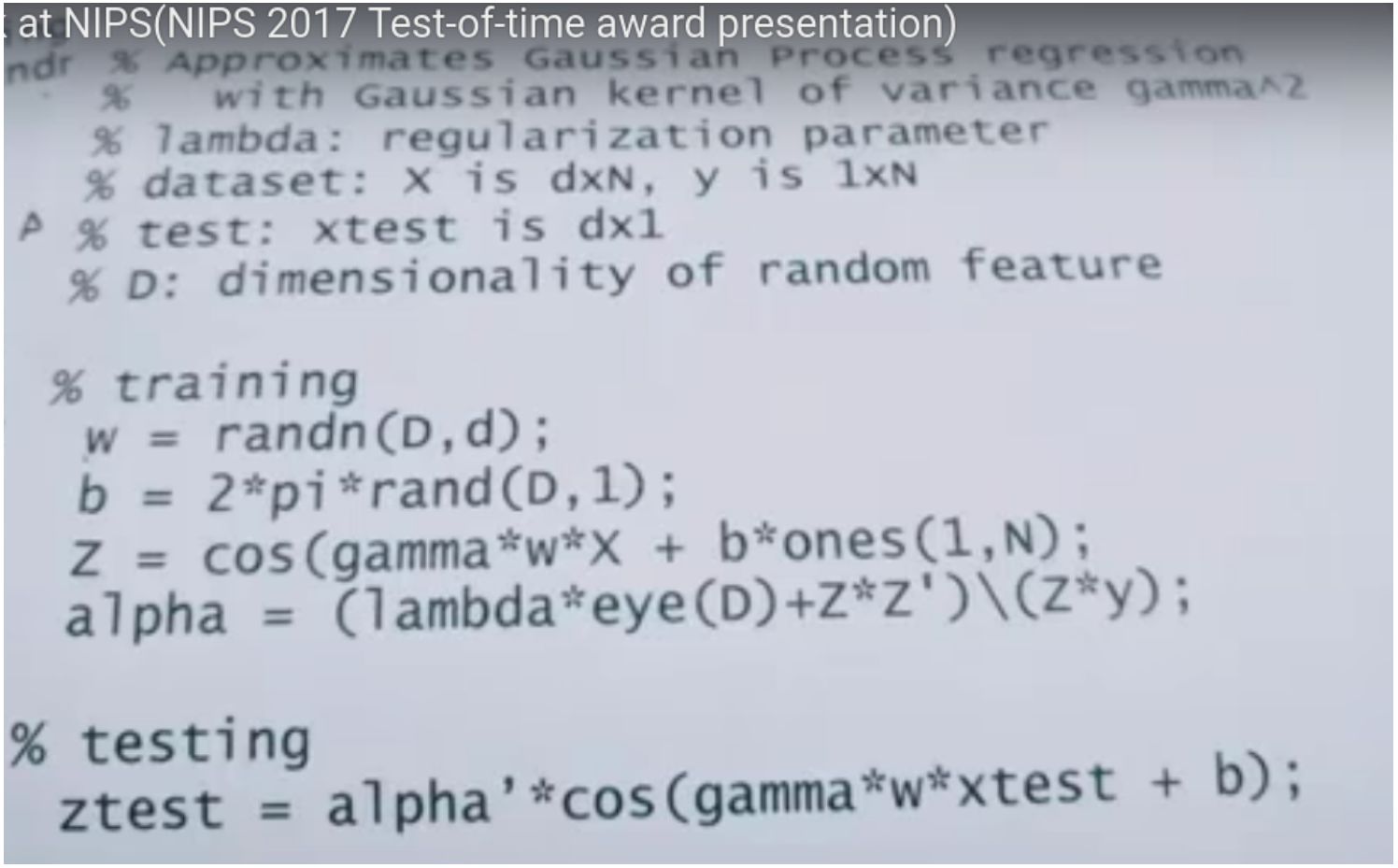
- Test of Time" award за работы 2007-2008 о случайных фичах.

Random Features for Large-Scale Kernel Machines

И в более общем виде

Weighted Sums of Random Kitchen Sinks

RBF SVM vs Ведро с болтами



Ведро с болтами -2

- Ali Rahimi + Benjamin Recht, 2007
- Из хулиганских побуждений
- Хотели показать, что на достаточно большом количестве случайных фич качество будет почти как у SVM.
- Качество оказалось лучше.
- «обучение» значительно быстрее.

О чем спор-то?

- Али говорит, что современные глубокие сети — то же самое ведро с болтами. Любая достаточно большая куча гладких функций подойдет.
- Ян несогласный ;-)

Не такой глубокий пример

- The Robust Beauty of Improper Linear Models in Decision Making
- Estimating Coefficients in Linear Models: It Don't Make No Nevermind
- Устойчивая красота неприличных моделей

Неглубокий пример-2

- Стандартизируем (делим на sd, например)
- «По причине отсутствия данных для определения значений β1...βm мы поступим волюнтаристски: назначим им значения +1 (влияет положительно), -1 (влияет отрицательно) и 0 (не влияет).»
- Оно работает. Без обучения. Хуже чем RF. Немного хуже LR. Но без обучения совсем.

Алхимическая мораль

- Это магия. Проще относитесь к жизни
- Ищите артефакты, которые отзываются на вашу магию, и используйте их
- Пробуйте разное, побольше
- Черная магия сеток
- Зеленая магия деревьев
- Фиолетовая магия гауссовских процессов и вероятностных графов.

Поле боя

Бизнес «на земле»

Оптовая торговля.
Производство.
Не-IT услуги для бизнеса.
Издательство.
Агентство недвижимости.
Сеть кафе, автосервисов, автомоек.
Небольшая транспортная компания.

Все будет классно

- 1000 товаров
- 20000 клиентов
- B2B
- Менеджеры подбирают пары, звонят, продают.
- В 99% случаев их отправляют в сад
- Очень успешный бизнес

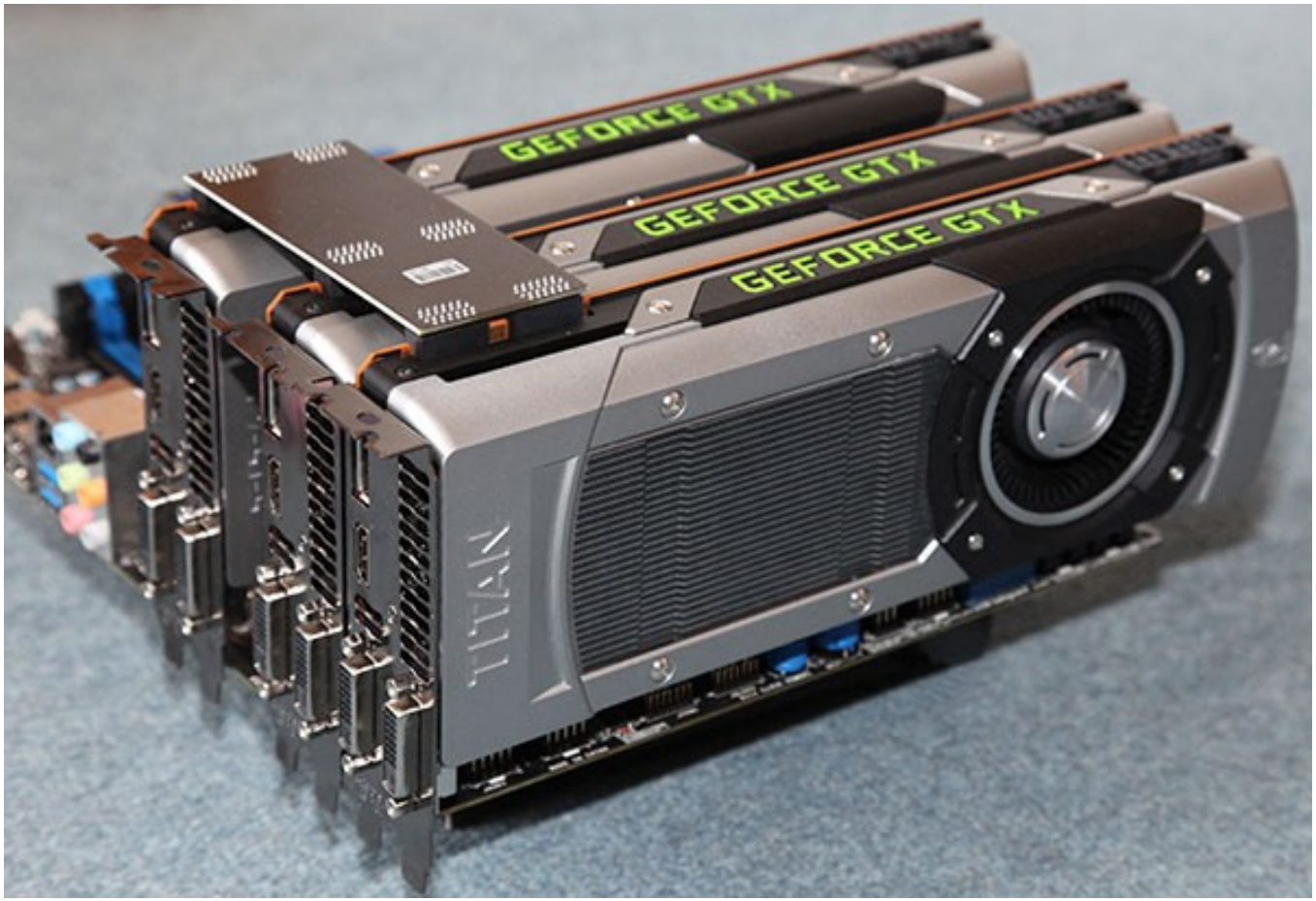
Все будет классно - 2

- Поможем продавать больше.
- Давайте будем подбирать пары клиент-товар, по которым велика вероятность продажи.
- Давайте отсортируем эти пары по вероятности и дадим менеджерам
- Обычно по клиенту рекомендуют товар.
- Или по товару — клиента.

Идея!

- F (товар, купец) вероятность сделки
- Считаем 20 000 000 вариантов, исключаем уже сделанные предложения, динамически рассчитываем, дообучаемся после каждого клика
- Фильтруем по менеджерам, регионам, отраслям….
- Сколько у вас RAM? 8 Gb, и на нем еще CRM + 2 вспомогательных БД + сервер MQ + десяток микросервисов.

ML здорового человека



ML курильщика

(c) Леонид Каганов https://lleo.me/dnevnik/2010/04/27.html

Ну и ладно

- В соответствии с бизнес-правилами выкинем маловероятные сочетания клиент-товар
- Выкинем те, что уже предлагались в ближайшие полгода
- Посчитаем вероятность, оставим те, у которых $P > 0.9$
- Рассчитываем прогноз раз в час, обучаемся ночью раз в сутки, когда сервер не загружен

Магия внедрения

- Ищем лояльного нам продавца
- Даем ему список пар (в html)
- Расспрашиваем, где очевидная чушь
- Исправляем
- Когда другие продавцы начнут просить — идем сдаваться начальству и впиливаем это в CRM.

Глаз да глаз

- Обучаемся на 90% данных, на 10% считаем тест, записываем в базу
- По итогам дня считаем logloss
  $-1/N\ (y * \log(p) + (1-y) * \log(1-p))$
- Расспрашиваем продавцов
- Подкручиваем фичи
- Посматриваем на графики теста и итогов дня
- Ассерт+алерт на резкое падение качества

Гига-метрика

- По какому проценту успешных сделок у нас был прогноз > 0.9
- Бесполезна для обучения, тестирования,
- Очень, очень полезна при внедрении
- Высокая с самого начала, т. к. самосбывающееся пророчество.

Мораль 3 части

- Делайте прототип тайком
- Отслеживайте метрики
- Ищите понятные менеджменту показатели
- Вербуйте сторонников
- Грабьте корованы

У кого процессор толще

- Microsoft
- Google
- Amazon
- IBM
- …

Больше данных, больше серверов, больше инженеров, больше денег.

Для всего есть API

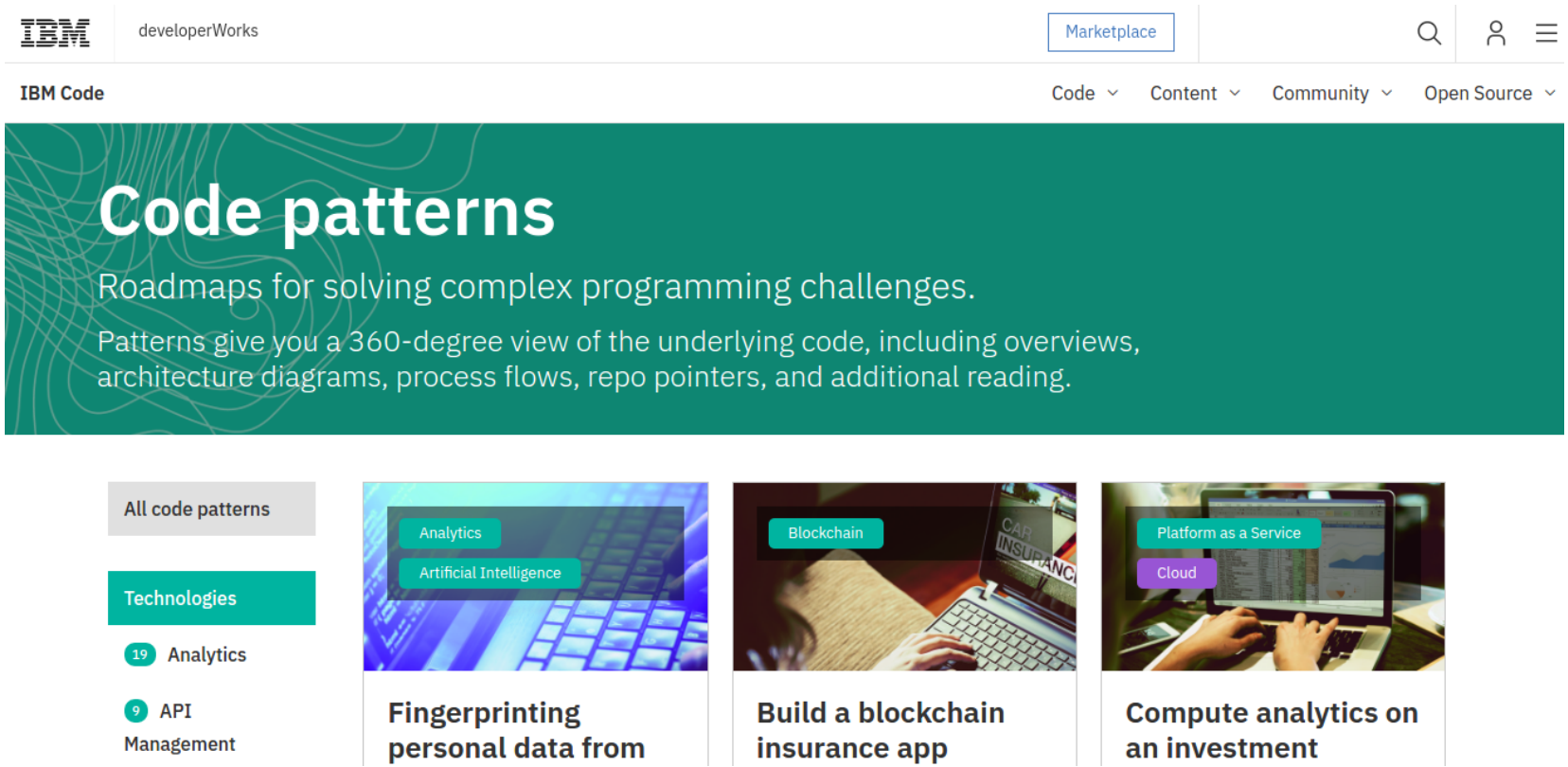
- Распознавание и классификация картинок
- Распознавание голоса
- Прогноз временных рядов
- Поиск аномалий
- Классификаторы и регрессии
- Анализ текстовых документов
- Для чего еще нет — скоро будет

Работа с возражениями

- Да, стоит денег, но на пробу можно взять немножко, и если заработает — денег дадут.
- Не стартап, не важна IP
- Не продуктовка, не важен контроль над инфраструктурой
- Важно внедрить как можно быстрее, пока это еще дает конкурентные преимущества

И немедленно выпил!

- Все уже ~~украдено~~ разжевано до нас https://developer.ibm.com/code/patterns/



И немедленно выпил!

- Брать и немедленно делать
- Скоро каждый 1С-разработчик сможет это сделать двумя строками кода.
  Успей до него, делай прямо сейчас.
- Трех человекодней достаточно (почти) для любого ML-прототипа.

- В любой непонятной ситуации
  читайте Rules of ML от Гугла.

Вопросы?