Машинное обучение в бизнесе

Лекция 9
Дмитрий Колодезев
НГУ 2020, весна

Магия

А на самом деле

Родители: «Если все твои друзья прыгнут с крыши, ты тоже прыгнешь?»
Ребёнок: «Нет!»

Алгоритм машинного обучения: «Да!»

ML — не про качество

Почти всегда:

Быстро принимаем
не очень важные решения
среднего качества
дорого и неконтролируемо

Просто, быстро, много

Люди

- Медленные
- Дешево начать
- Не масштабируются
- Адаптируются
- Качество разное
- Издержки

ML-алгоритмы

- Быстрые
- Дорого начать
- Масштабируются
- Не адаптируются
- Качество среднее
- Капитальные вложения

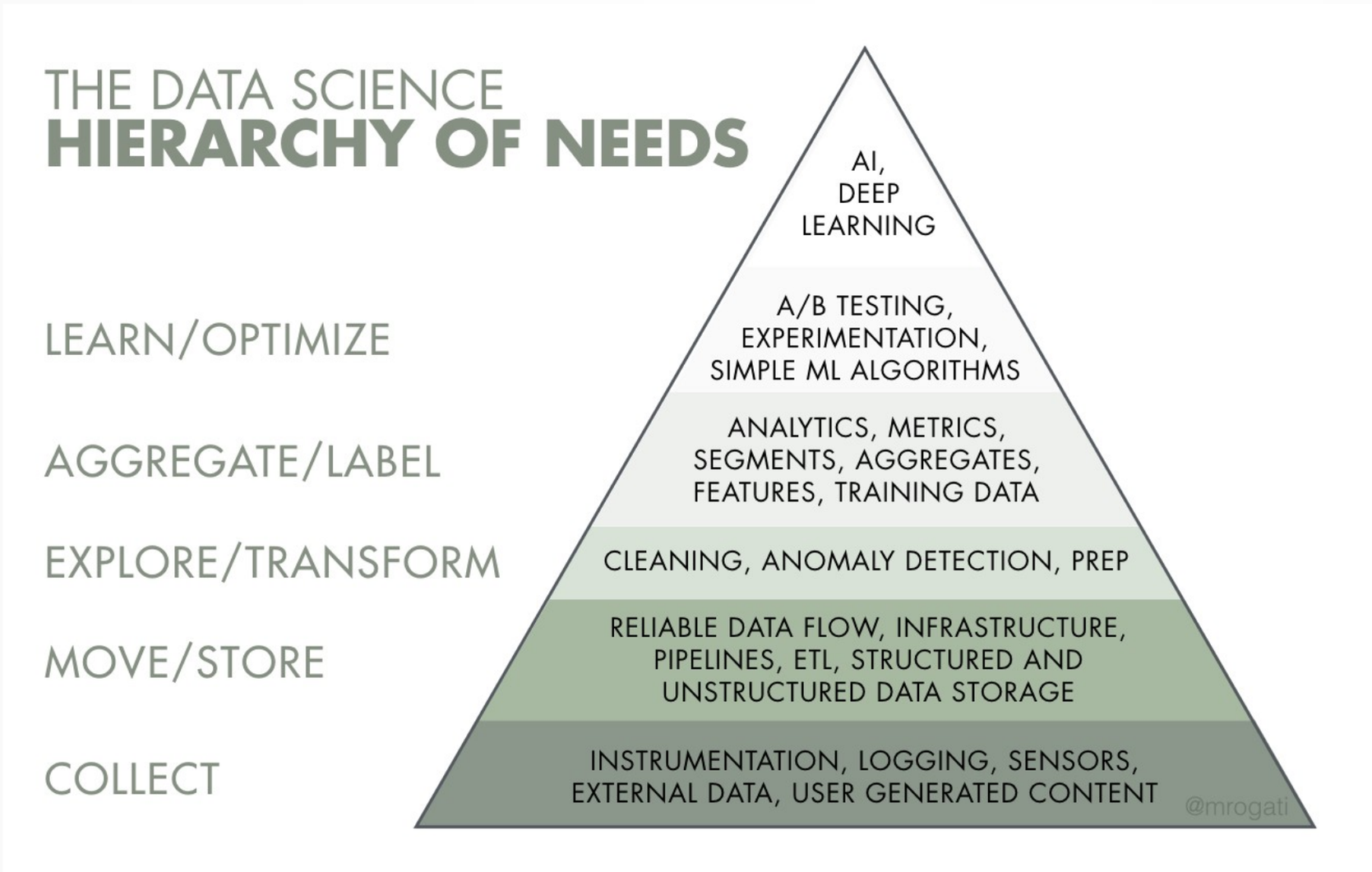
Автоматизируй начальство

- Оценка
- Контроль
- Ресурсное планирование
- Выявление аномалий
- Прогнозирование
- Выявление влияющих факторов
- Отчетность

Автоматизируй клерка

- Скоринг
- Антифрод
- Поиск дубликатов
- Ответы на типовые вопросы
- Анализ рекламных кампаний
- Анализ отчетов

Готов ли ты к ML?



https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

Необходимые условия

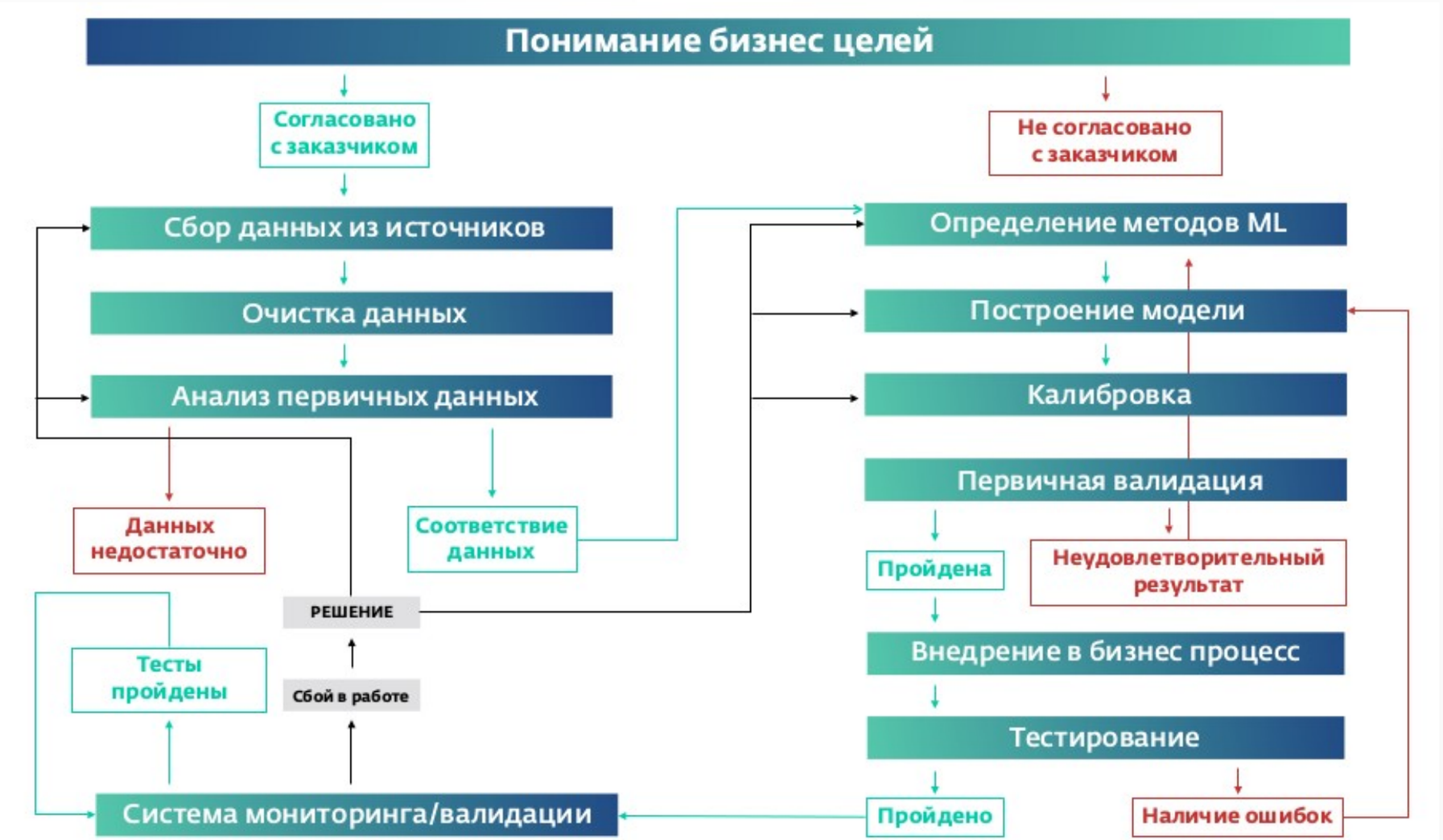
- Владелец процесса
- Неэффективность
  - Медленно. Дорого. Трудоемко. Сложно
- Данные, много
  - Мы про анализ данных
- Возможность внедрения
- Прямое влияние на бизнес-показатель
  - 1% значимо — надо делать
  - 10% значимо — надо считать

Античность: CRISP-DM

- Понимание бизнеса
- Понимание данных
- Подготовка данных
- Моделирование
- Оценка
- Развертывание
- 1996 г. Ровесник DVD

ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf

Современный подход

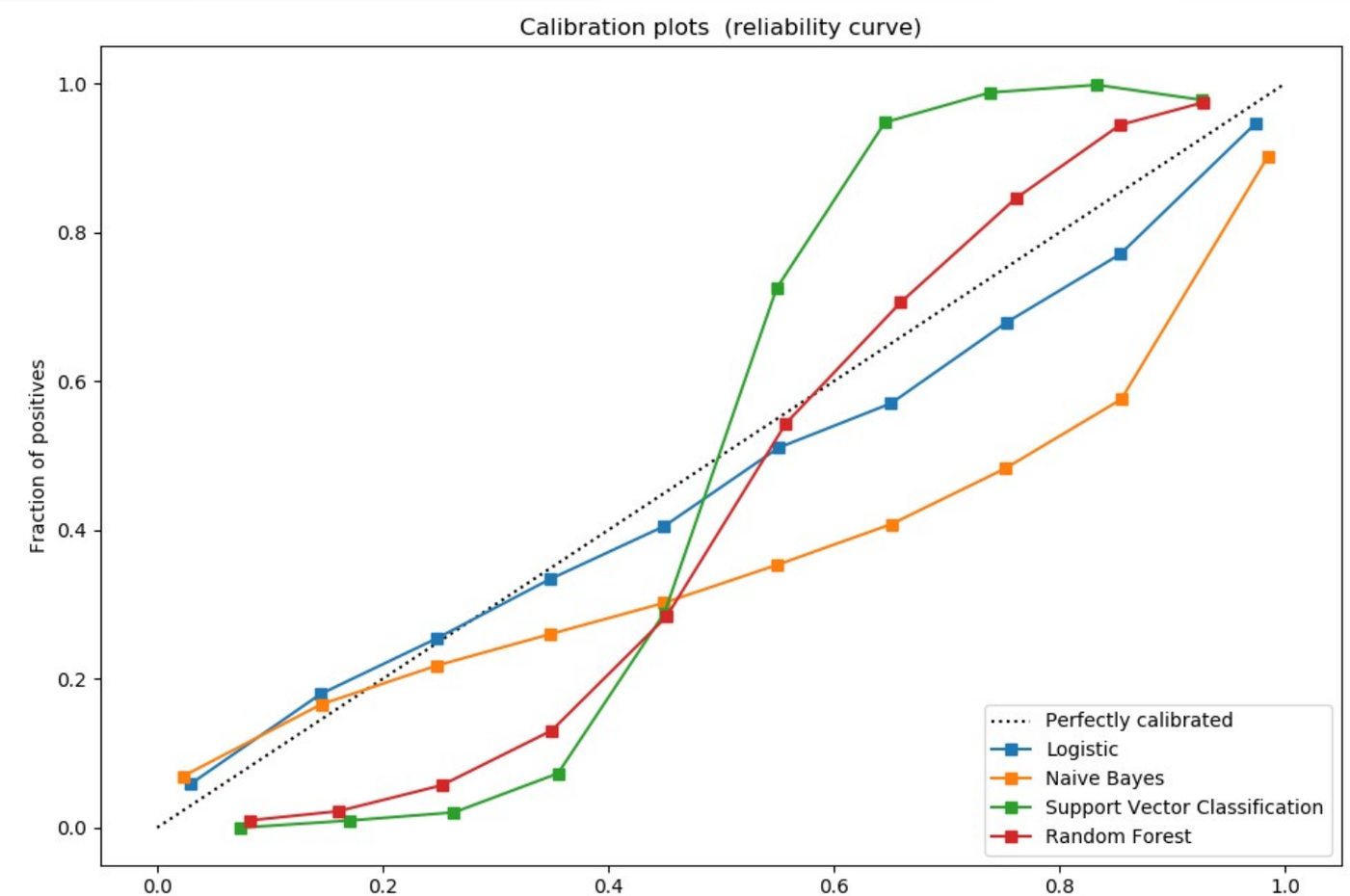


Источник: https://www.coursera.org/learn/machinnoe-obuchenie-v-finansah/

Придут черные лебеди

- Мониторинг входных данных
  - Распределение признаков
  - Детектор новинок
- Мониторинг работы модели
  - Распределение предсказаний
  - Оценка уверенности модели
- Дообучение модели
  - Когда?
  - Как проверить, что хорошо?

Калибровка



- predict_proba - не вероятность.
- Не вероятность!

https://scikit-learn.org/stable/modules/calibration.html

Функция потерь в рублях

- FP и FN имеют разную цену
- Пропустить фрод vs
  Отказать хорошему клиенту
- Пропустить брак vs
  Выкинуть хороший товар
- Пропустить аварию турбины vs
  Зря остановить ее на месяц

Функция потерь в рублях

- Функция потерь может быть сложной
- Недифференцируемой
- Зависящей от внешних условий

- Оптимизировать ROC-AUC
- Подбирать порог по своей функции потерь

Еще проблемы

- Интерпретируемость
- Предсказуемое качество
- Извлечение данных
- Фиксация прошлого
- Отказ от предсказаний (эскалация)

Почитать и посмотреть

- Машинное обучение в финансах (coursera)
- Проблема калибровки уверенности
- Прощадь под кривой ошибок
- Интерпретация черных ящиков

Вопросы

Слайды тут

dkolodezev

promsoft

dkolodezev

d_key

dmitry_kolodezev

https://kolodezev.ru/download/slides_nsu_mlb2020.pdf