Процесс разработки дата-сервисов

или «CRISP курильщика»

Дмитрий Колодезев
ООО Промсофт, Новосибирск

План выступления

- Определим термины
- Почему это проблема и как это вас касается
- Правильно поставленные задачи
- Процессы CRISP-DM и ASUM
- Проблемы и решения
- CRISP курильщика — до входа в проект
- CRISP курильщика — на проекте
- Рекомендованная литература

"data scientists ... incorporate data science into the operation of a product or service, using data in smart ways to provide value."

Edo Wilder-James https://www.svds.com/how-do-you-build-a-data-product/

Договоримся о терминах

- Дата-продукт — статичные данные
- Дата-сервис — меняющиеся данные
- Небольшой датасервис — 2-3 месяца работы
- Внешний заказчик — соседний департамент
- Вход в проект — работа до денег
- Работа на проекте — работа за деньги

Заказчик (не о нас)

- Ушли, зажав в передних лапах N миллионов.
- Вернулись через полгода с непонятной бесполезной моделью.
- Не будем отдавать задачи на аутсорс

Задача плохо поставлена!

ДНК: Дано, Найти, Критерий

— Здравствуйте. Возьмите меня к себе жить. Я вам буду всё охранять.

— Ещё чего! Мы сами нигде не живём. Ты к нам через год прибегай, когда мы хозяйством обзаведёмся.

(с) Трое из Простоквашино

CRISP-DM

Когда DS назывался DataMining

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf

Вам платят за то, чтобы вы разобрались в бизнесе заказчика, посмотрели его данные, поняли что оттуда можно извлечь и помогли поставить задачу.

Это может сработать.

А может и не сработать.

Проблемы - 1

Предварительный анализ бизнеса:

– Мы хотим нанять профессионалов.

– Мы хотим, чтобы у них был опыт решения нашей уникальной проблемы.

– Мы хотим провести тендер.

Проблемы - 2

Предварительный анализ данных

- Данные нужно найти, собрать или купить.
- Данные во внутренних системах, куда до проекта могут и не пустить
  (могут не пустить и после)
- Данные стоят денег (идеальный рынок)
- Непонятно, какие данные нужны

Проблема - 3

Задача поставлена плохо

- Данные не те
- Найти не то
- Критерий не тот

Проблема возникает на стыках модели с реальностью

ASUM

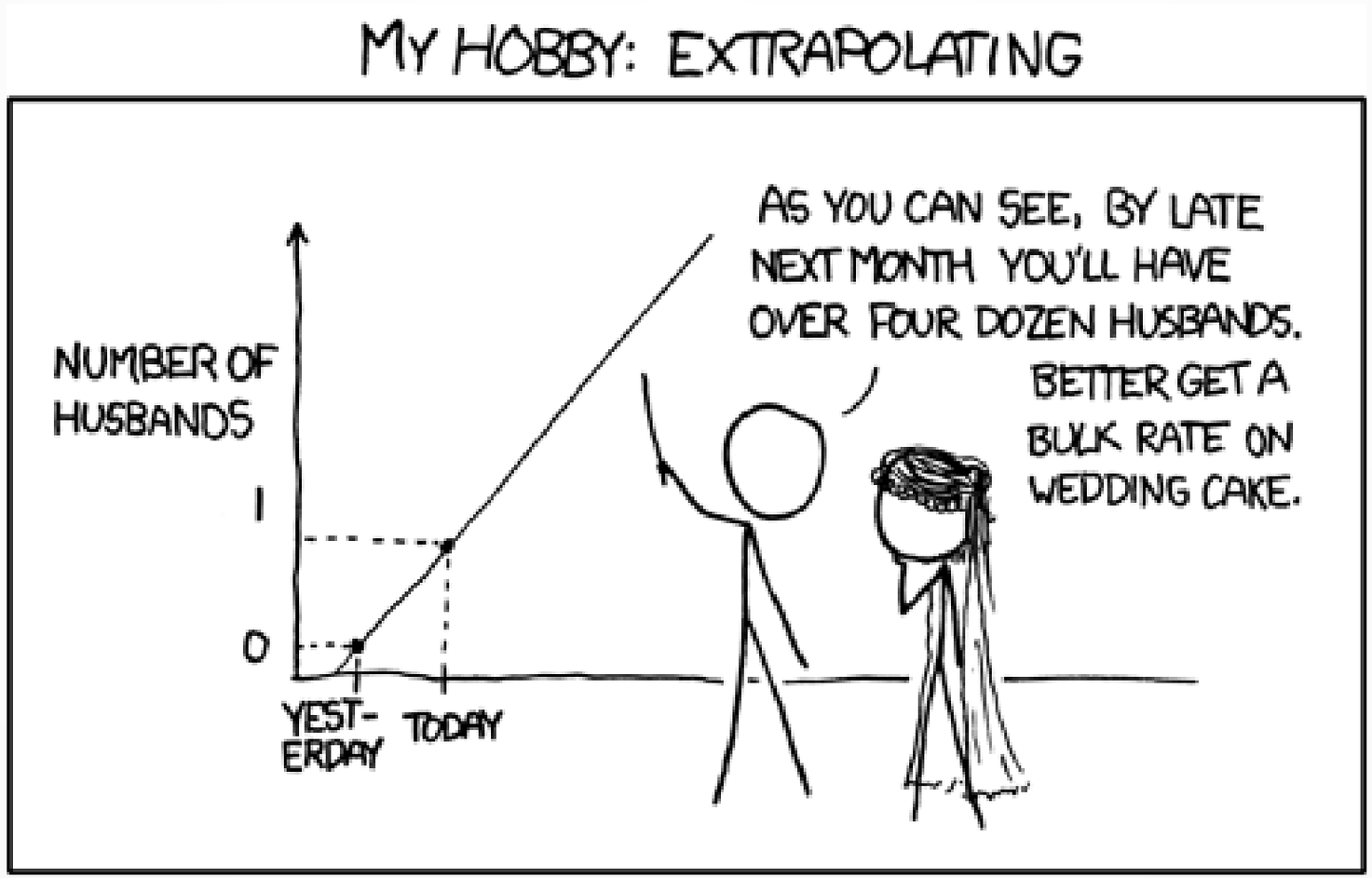
Когда DS и IT слились воедино:

- Analyze
- Design
- Configure & Build
- Deploy
- Operate & Optimize
- Project Management

ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf

Несмотря на слова Agile в тексте, процесс ASUM подозрительно напоминает процессы разработки ПО начала 90-х

Все придет к SCRUM и Co




Реабилитация-Магнитогорск.рф
НАРКОМАНИЯ
ВЫХОД ЕСТЬ
АЛКОГОЛИЗМ
Бесплатные консультации родителям
Телефон доверия
43-12-91
8-922-742-55-33

CRISP курильщика

Прототипируй!

CRISP курильщика

- Быстро делаем прототип.
- Смотрим что получилось.
- Повторяем, пока не кончатся деньги.
- Просим еще.
- Когда деньги кончились, продукт готов.

Прототипируем всё на каждом этапе.

До входа в проект

- Ключевые слова
- Готовые прототипы
- Источники данных
- Суррогатная модель
- Изучаем суррогатную модель
- Допрашиваем заказчика

Ключевые слова

- Бриф, сайт и фейсбук заказчика.
- Запоминаем ключевые слова.
- Гуглим их значение.
- Бегло просматриваем профильные журналы и форумы.
- Контачим со всеми, с кем удастся.

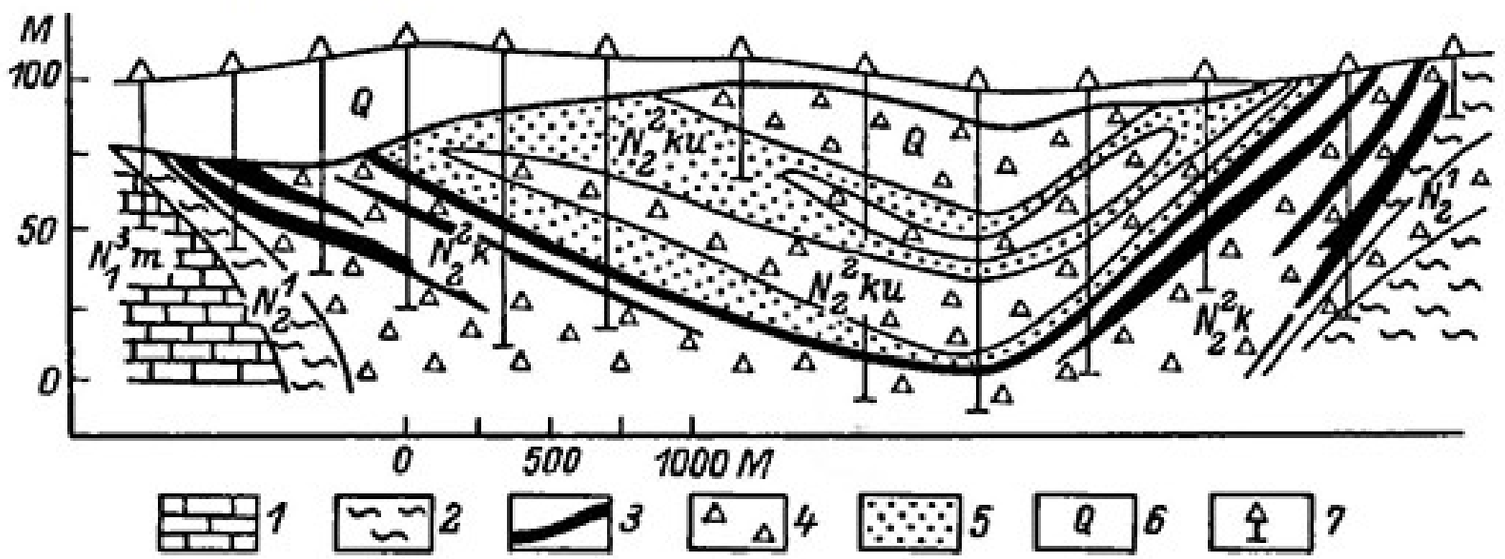
Сделано другими

- Kaggle
- Crunchbase
- Producthunt
- ODS Slack
- Upwork
- На выходе:
  набор предположений о проблемах отрасли

Геология данных

Время и деньги

1. Данные, которые вам показали
2. Данные, которые можно купить
3. Данные, которые можно собрать
4. Хорошие данные, которые где-то есть
5. Открытые данные



Суррогатная модель

- Пробуем предсказать хоть что-то
- Нет разметки?
  Попробуйте предсказать косвенный признак.
  (рекомендательная система → цена)
  (здоровье → возраст)
- Неинтерпретируемая, без валидации
- Какой-нибудь бустинг
- Чтобы понять, как устроен мир

Допрос модели

- ELI5
  https://eli5.readthedocs.io/en/latest/index.html
- LIME
  https://github.com/marcotcr/lime
- SHAP
  https://github.com/slundberg/shap

  Объяснение типичных точек даст набор инсайтов средней руки для начала разговора

Допрос заказчика

- Есть учебники по допросу (см «Книги»)
- Главное — список гипотез
- Идеально — мимоходом
- Несколько человек
- Слушать и смотреть по сторонам
  Идите в гембу.
- Что не укладывается в вашу картину мира?

На проекте

- Визуализация
- Пайплайн
- Метрика и бейзлайн
- Обучение модели
- Валидация данных
- Деплой и версионирование
- Оценка и мониторинг
- Дообучение на новых данных

Визуализация

- Визуализируйте входные данные
- Pandas profiling
  https://github.com/pandas-profiling/pandas-profiling
- Гистограммы
- Попарные графики
- В любой непонятной ситуации визуализируйте
- Даже если и так все понятно

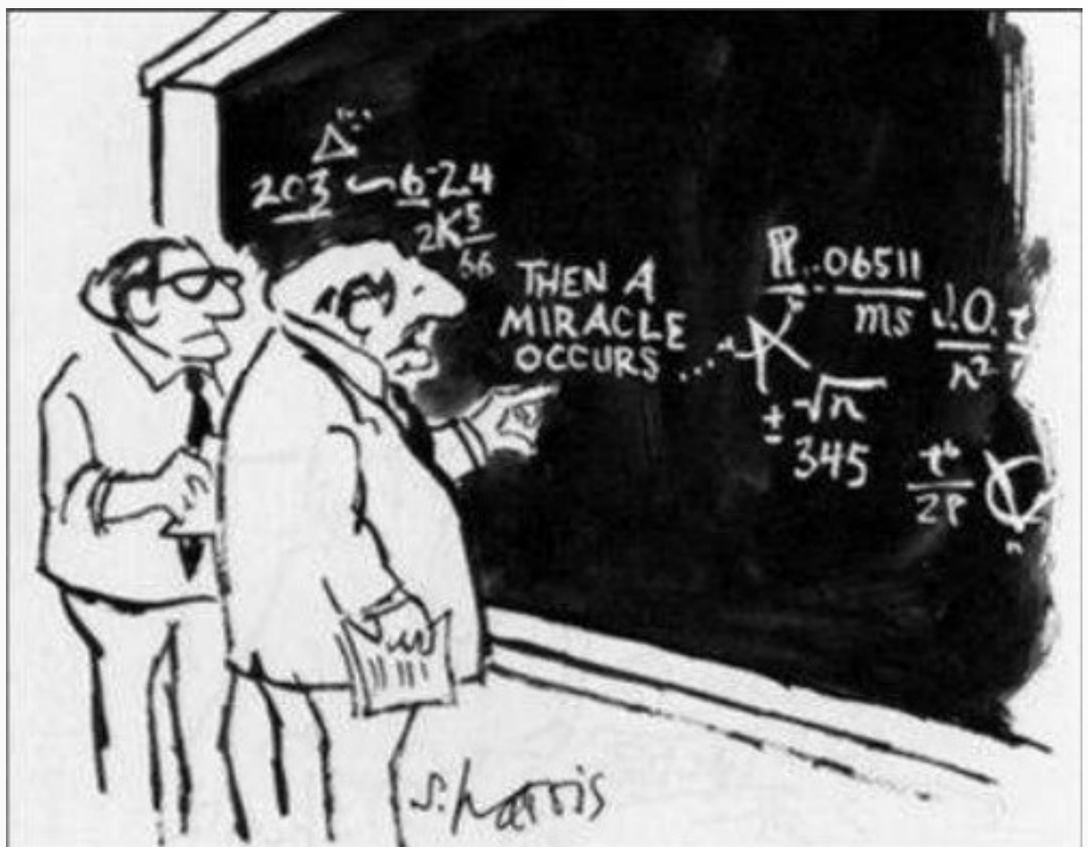
Pipeline

- Из конца в конец как можно быстрее
- В середине что подвернется — константа, таблица значений, простой if, random, простая модель.
- Проблемы обычно на входе и выходе
- Пайплайн обучения
- Пайплайн инференса

Метрика и бейзлайн

- Метрика до начала работы над моделью.
- Идеально — понятная заказчику
- Не F1, а что-то в терминах денег, времени
- Основа успеха проекта
- И немедленно зафиксировал бейзлайн!
- Бейзлайн до работы над моделью,
  чтобы понимать — где вы сейчас.

Обучение модели



"I think you should be more explicit here in step two."

- Простые фичи
- Простые модели
- Грубые ошибки
- Свежие данные
- Тюнинг потом

Валидация данных

- Пропадают и добавляются колонки в базе
- Меняется формат данных, из текста в строку
- Проверять предположения о диапазоне значений, пропущеных данных, дублирующихся идентификаторах и проч.
  https://hypothesis.readthedocs.io/en/latest/numpy.html
- .....
- Валидация — часть пайплайна обучения
- Разумный отвал в продакшене

Деплой и версионирование

- Докер ваш друг
- Версионирование API ваш второй друг
- Как вы будете выкатывать новую модель?
- Как вы будете откатывать модель назад?
- Как вы будете поддерживать несколько версий модели?

Оценка и мониторинг

- Метрики после обучения:
  - Качество на отложенной выборке
  - Скорость работы
  - ....
- Метрики работающей модели (какие?)
- Мониторинг ошибок (sentry?)
- Показатели в динамике, привязанные к версиям модели

Дообучение

- Мир меняется
- Распределение входных данных меняется
- Как модель будет дообучаться?
  - По запросу? Как поймем, что пора?
  - Ежемесячно руками?
  - Ежедневно автоматом?
  - Онлайн?
- Как контролировать качество?

Книги

- Rules of ml
  http://martin.zinkevich.org/rules_of_ml/rules_of_ml.pdf
- Machine Learning Yearning
  http://www.mlyearning.org/
- Спроси маму
  https://www.alpinabook.ru/catalog/StartupsInnovativeEntrepreneurship/125907/

Вопросы?

dk.promsoft@mail.ru

telegram.me/Promsoft

Skype: dk_promsoft

ODS @d_key