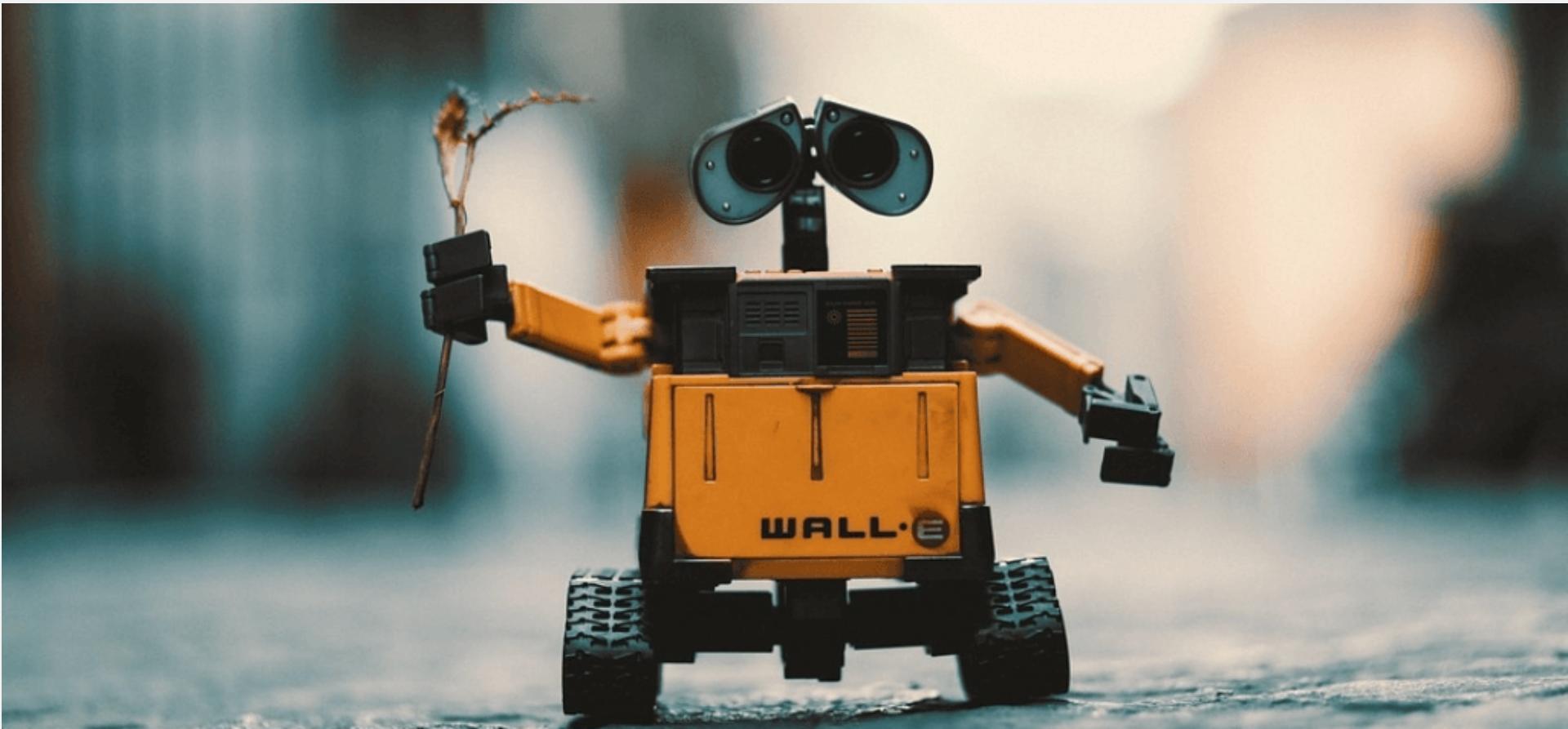




Почта Mail.Ru: ML hands-on

Дмитрий Меркушов

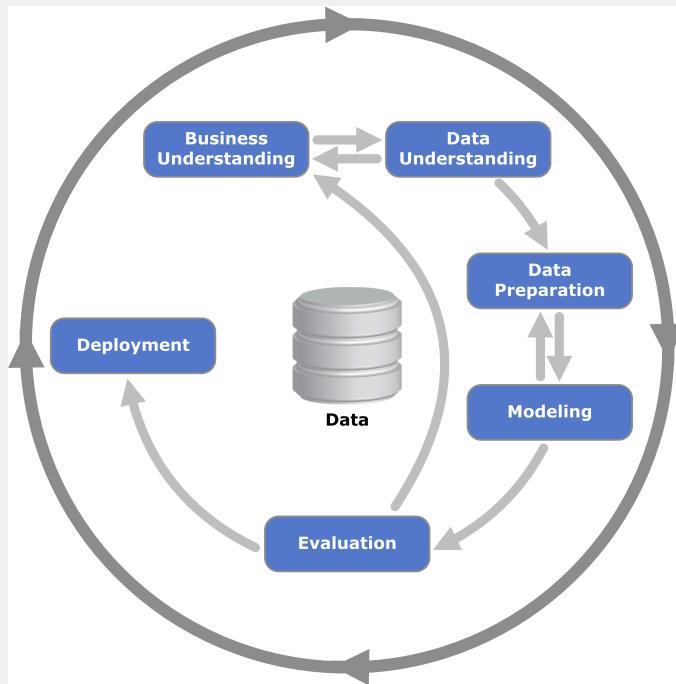
Data Mining: Hands On



Data Mining: CRISP

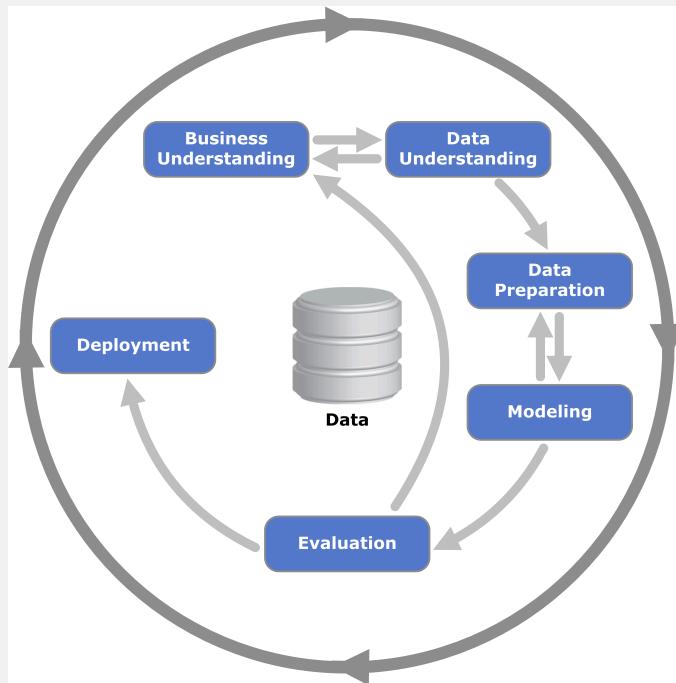


Data Mining: CRISP



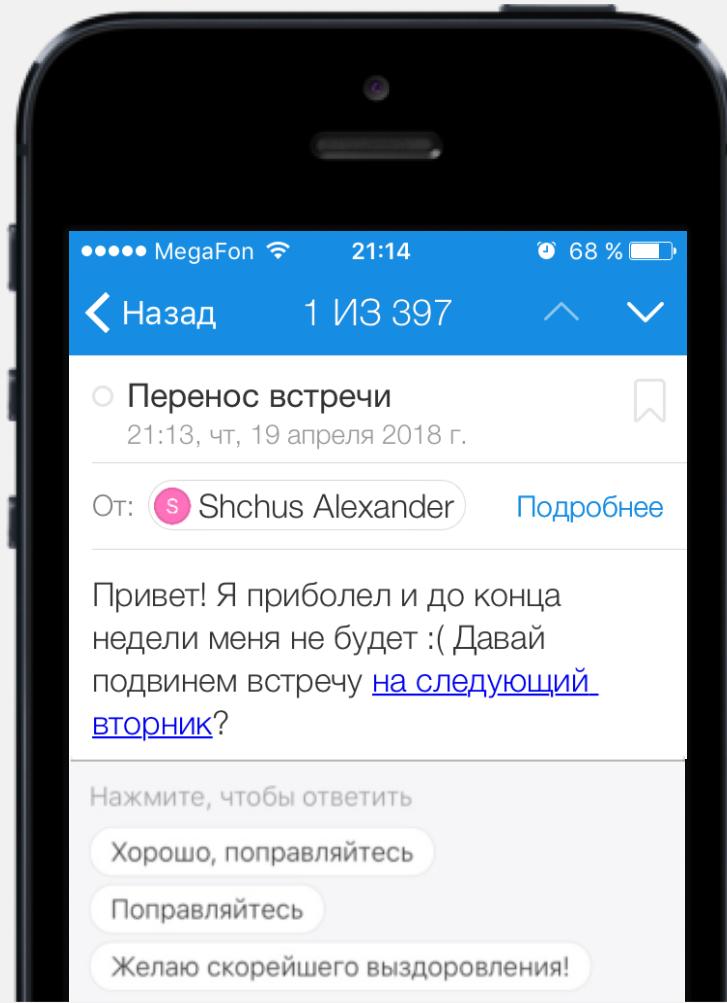
- Бизнес формулировка
- Проверка гипотез
- Сбор данных для обучения
- Модель
- Оценка системы
- Вывод в бой

Data Mining: CRISP



- Бизнес формулировка
- Проверка гипотез
- Сбор данных для обучения
- Модель
- Оценка системы
- Вывод в бой
- + Поддержка

Почта: SmartReply



Почта: SmartReply



- Бизнес формулировка
 - Жизнь пользователей удобней – помогаем с простыми ответами
 - Пользователь тратит меньше времени на ответ в mobile
 - Репутационно – решение должно быть AI first

Почта: SmartReply



- Бизнес формулировка
 - Жизнь пользователей удобней – помогаем с простыми ответами
 - Пользователь тратит меньше времени на ответ в mobile
 - Репутационно – решение должно быть AI first
- Проверка гипотез
 - Большинство ответов в mobile – короткие
 - Большинство ответов в mobile – из golden set

Почта: SmartReply



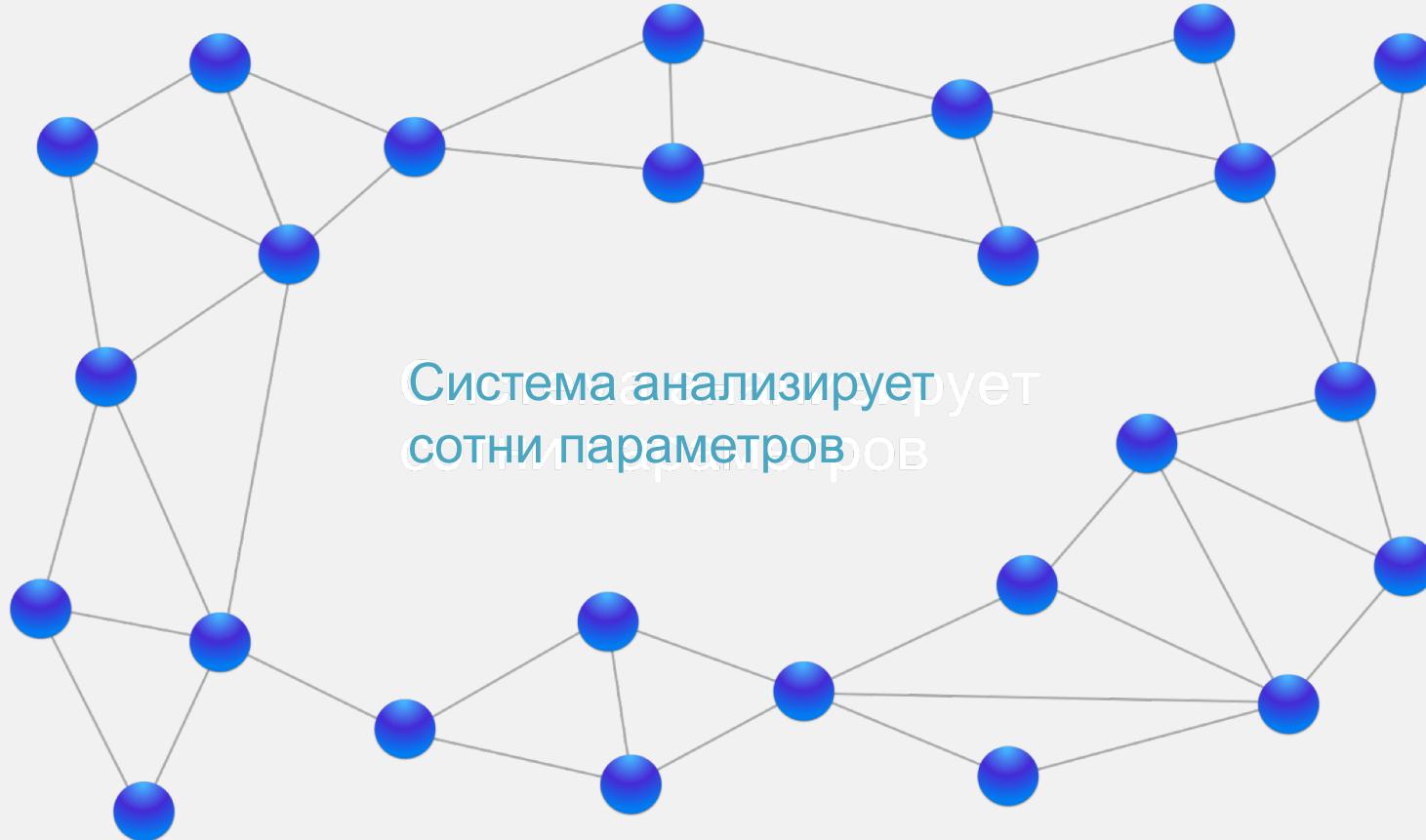
- Бизнес формулировка
 - Жизнь пользователей удобней – помогаем с простыми ответами
 - Пользователь тратит меньше времени на ответ в mobile
 - Репутационно – решение должно быть AI first
- Проверка гипотез
 - Большинство ответов в mobile – короткие
 - Большинство ответов в mobile – из golden set
- Сбор данных
 - Выделяем большой корпус
 - Прочищаем от шлака, цензурируем
- Модель
 - Пробуем разные, от простого к сложному
 - Seq2seq, DSSM

Почта: SmartReply



- Бизнес формулировка
 - Жизнь пользователей удобней – помогаем с простыми ответами
 - Пользователь тратит меньше времени на ответ в mobile
 - Репутационно – решение должно быть AI first
- Проверка гипотез
 - Большинство ответов в mobile – короткие
 - Большинство ответов в mobile – из golden set
- Сбор данных
 - Выделяем большой корпус
 - Прочищаем от шлака, цензурируем
- Модель
 - Пробуем разные, от простого к сложному
 - Seq2seq, DSSM
- Оценка
 - Технические – метрики на выборках, семплах prec@3
 - Продуктовые – % пользования фичей, % показанных ответов

Антиспам: Marshal



Антиспам: Marshal



Антиспам: Marshal



- Бизнес формулировка
 - Оперативный детект взломов

Антиспам: Marshal



- Бизнес формулировка
 - Оперативный детект взломов
- Проверка гипотез
 - У пользователя есть характерный профиль поведения
 - Взломы сопровождаются сменой пользовательского поведения

Антиспам: Marshal



- Бизнес формулировка
 - Оперативный детект взломов
- Проверка гипотез
 - У пользователя есть характерный профиль поведения
 - Взломы сопровождаются сменой пользовательского поведения
- Сбор данных
 - Очень много данных сессий пользователей
 - Выделить явно хорошие, явно плохие
 - Профит – выделить граничные плохие, возможно ассессорами

Антиспам: Marshal



- Бизнес формулировка
 - Оперативный детект взломов
- Проверка гипотез
 - У пользователя есть характерный профиль поведения
 - Взломы сопровождаются сменой пользовательского поведения
- Сбор данных
 - Очень много данных сессий пользователей
 - Выделить явно хорошие, явно плохие
 - Профит – выделить граничные плохие, возможно ассесорами
- Модель
 - Пробуем разные, от простого к сложному
 - Pattern Mining: Apriori, FP-Growth

Антиспам: Marshal



- Бизнес формулировка
 - Оперативный детект взломов
- Проверка гипотез
 - У пользователя есть характерный профиль поведения
 - Взломы сопровождаются сменой пользовательского поведения
- Сбор данных
 - Очень много данных сессий пользователей
 - Выделить явно хорошие, явно плохие
 - Профит – выделить граничные плохие, возможно ассесорами
- Модель
 - Пробуем разные, от простого к сложному
 - Pattern Mining: Apriori, FP-Growth
- Оценка
 - Технические – метрики на выборках
 - Технические – оценка precision/recall на семплах
 - Продуктовые – число регулярных детекторов, число фолзов системы

Почта: Категоризация



Почта: Категоризация



Andrey

Защищено | https://e.mail.ru/messages/inbox/

Mail.Ru Почта Мой Мир Одноклассники Игры Знакомства Новости Поиск Все проекты ▾

• a.shamne@inbox.ru ▾ выход

@mail.ru БЕТА Письма Контакты Файлы Темы Ещё

Написать письмо Удалить Спам Переместить Ещё

Входящие

- Социальные сети
- Рассылки
- Важное
- Гитара
- Друзья
- Рыбалка
- Семья
- Фото
- Отправленные
- Черновики
- Архив
- Спам
- Корзина

ОЧИСТИТЬ ОЧИСТИТЬ

Удалить Спам Переместить Ещё

□ РАССЫЛКИ Biglion Москва, The New York Times

□ Tatyana Shamne Fwd: Документы Начало г

□ Штрафы ГИБДД М Штраф ГИБДД погашен

□ PlayStation Благодарим за покупку

□ iHerb Представляем доставку

□ Бандеролька Перемены всегда к лучшему

□ Kickstarter Projects We Love: The White

□ PlayStation Благодарим за покупку

□ PlayStation Благодарим за покупку

□ AliExpress Order Andrey Shamne, the s

□ Tatyana Shamne Fwd: Welcome Начало пер

□ OZON.ru Ваш заказ 22971728-0027 в OZON.ru оплачен

Перейти к отслеживанию 4 июн

□ OZON.RU Андрей, электронный чек по вашему заказу 22971728-0027 Инфор

4 июн

□ secure.payment@z Информация о платеже

Уважаемый покупатель Шамне Андрей И 4 июн

ВЫ НЕДАВНО ИСКАЛИ

От: info@twitter.com
От: Sony@email.sonyentertainmentnetwork.com
От: playstation@eu.playstationmail.net
От: follow-suggestion-noreply@quora.com
От: a.sergeev@corp.mail.ru

КАТЕГОРИИ

- Заказы
- Финансы
- Регистрации
- Путешествия
- Билеты
- Штрафы

Расширенный поиск

Почта: Категоризация



- Бизнес формулировка
 - Разобрать ящик пользователей от шума
 - Сделать работу с письмами удобней

Почта: Категоризация



- Бизнес формулировка
 - Разобрать ящик пользователей от шума
 - Сделать работу с письмами удобней
- Проверка гипотез
 - Пользователи взаимодействуют с ограниченным типом писем

Почта: Категоризация



- Бизнес формулировка
 - Разобрать ящик пользователей от шума
 - Сделать работу с письмами удобней
- Проверка гипотез
 - Пользователи взаимодействуют с ограниченным типом писем
- Сбор данных
 - Набираем выборки по каждому классу с помощью эвристик и регулярных выражений
 - Добавляем ошибки модели после ее первичной выкатки

Почта: Категоризация



- Бизнес формулировка
 - Разобрать ящик пользователей от шума
 - Сделать работу с письмами удобней
- Проверка гипотез
 - Пользователи взаимодействуют с ограниченным типом писем
- Сбор данных
 - Набираем выборки по каждому классу с помощью эвристик и регулярных выражений
 - Добавляем ошибки модели после ее первичной выкатки
- Модель
 - Разное, от простого к сложному
 - Bag of words, fastText классификаторы

Почта: Категоризация



- Бизнес формулировка
 - Разобрать ящик пользователей от шума
 - Сделать работу с письмами удобней
- Проверка гипотез
 - Пользователи взаимодействуют с ограниченным типом писем
- Сбор данных
 - Набираем выборки по каждому классу с помощью эвристик и регулярных выражений
 - Добавляем ошибки модели после ее первичной выкатки
- Модель
 - Разное, от простого к сложному
 - Bag of words, fastText классификаторы
- Оценка
 - Алгоритмические – метрики на выборках
 - Технические – оценка precision/recall на семплах
 - Продуктовые – число регулярных детекторов, число фолзов системы

Почта: Категоризация + NER



Заказ № 12533680 от 28.09.2019 поступила оплата

Заказы

Заказ

Вентилятор вытяжной ELECTROLUX Basic EAEB 120

- В процессе 28 сен
- Заказ в пути
- Доставлен

[Перейти в магазин](#)

	ОНЛАЙН ТРЕЙД.РУ	Интернет-магазин ОНЛАЙН ТРЕЙД.РУ Актуально на: 29.09.2019 0:04 Здравс...	29 сен
	ОНЛАЙН ТРЕЙД.РУ	Интернет-магазин ОНЛАЙН ТРЕЙД.РУ Актуально на: 28.09.2019 23:57 Здрав...	28 сен
	ОНЛАЙН ТРЕЙД.РУ	Интернет-магазин ОНЛАЙН ТРЕЙД.РУ Актуально на: 28.09.2019 20:25 Здрав...	28 сен
	ОНЛАЙН ТРЕЙД.РУ	Уважаемый клиент! По заказу 12533680-1 сформирована ссылка для доплат...	28 сен
	ОНЛАЙН ТРЕЙД.РУ	Интернет-магазин ОНЛАЙН ТРЕЙД.РУ Актуально на: 28.09.2019 20:05 Здрав...	28 сен

<https://new.mail.ru>
пробуйте