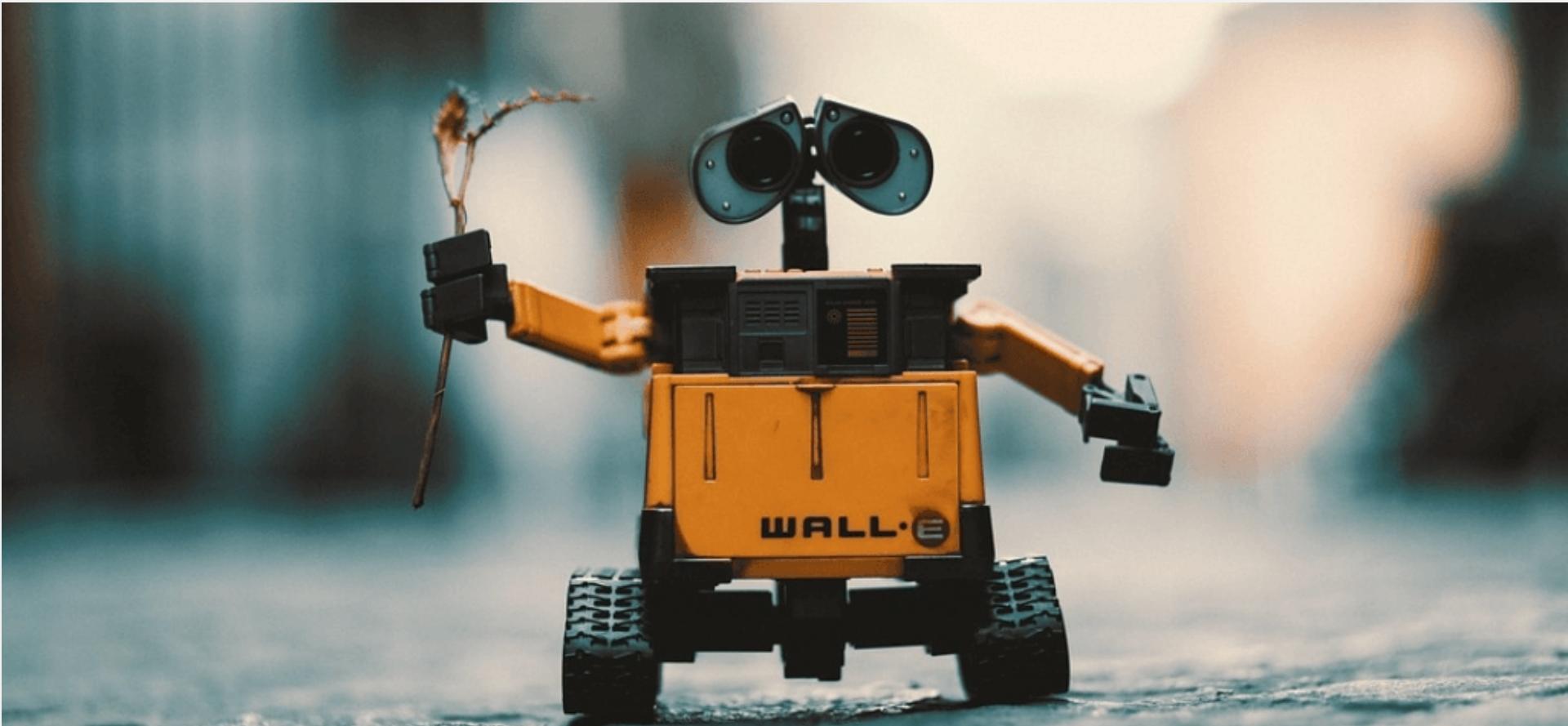




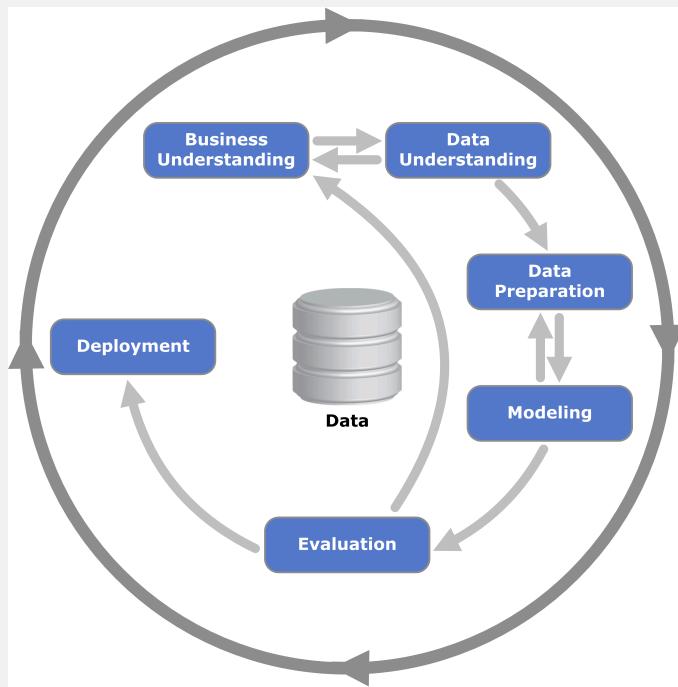
Почта Mail.Ru: ML hands-on

Дмитрий Меркушов

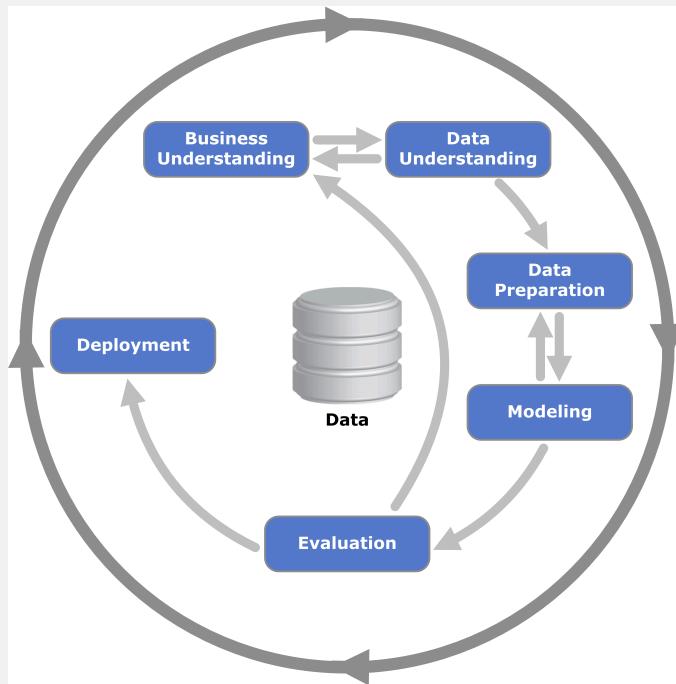
Data Mining: Hands On



Data Mining: CRISP

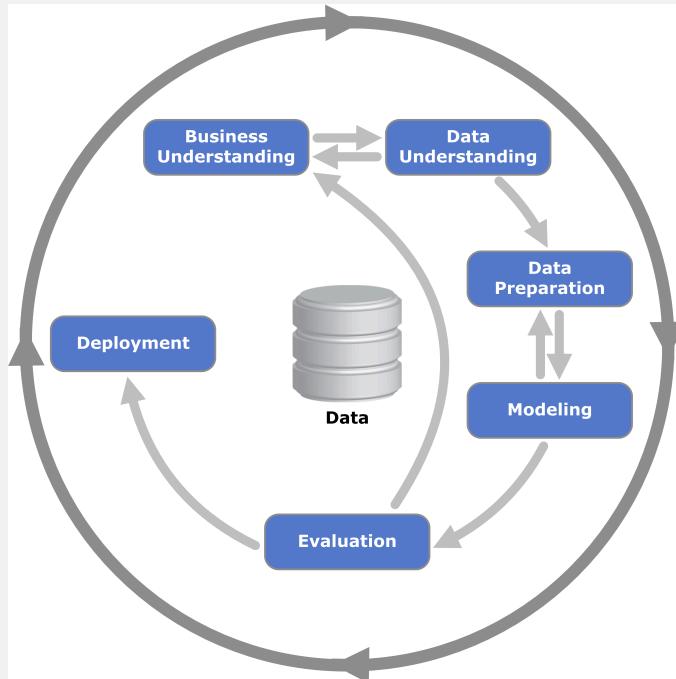


Data Mining: CRISP



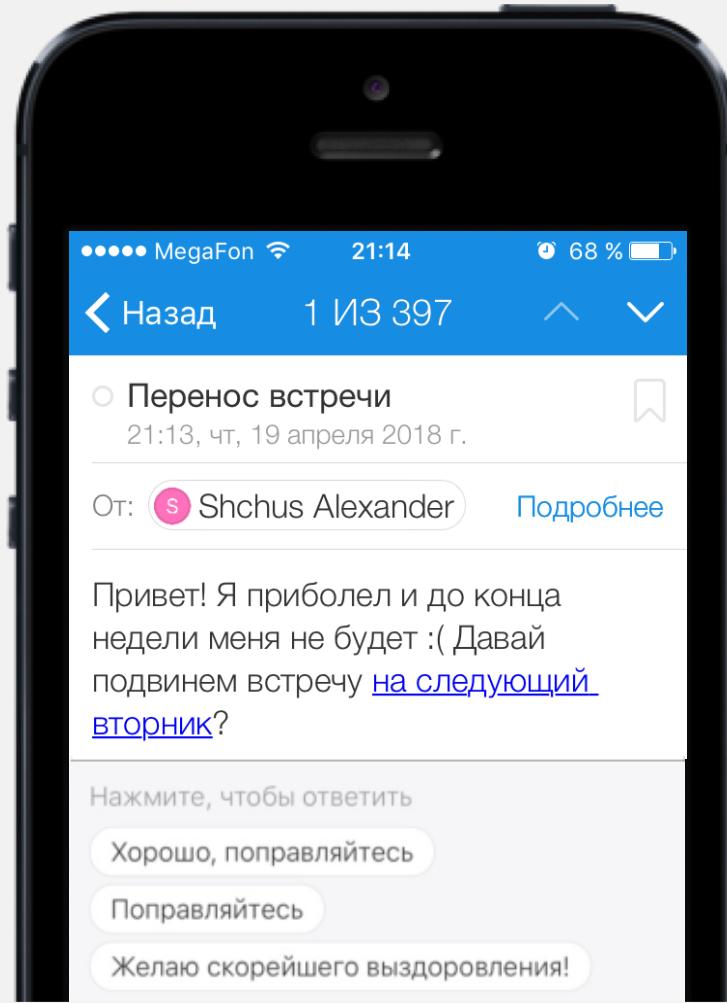
- Бизнес формулировка
- Проверка гипотез
- Сбор данных для обучения
- Модель
- Оценка системы
- Вывод в бой

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



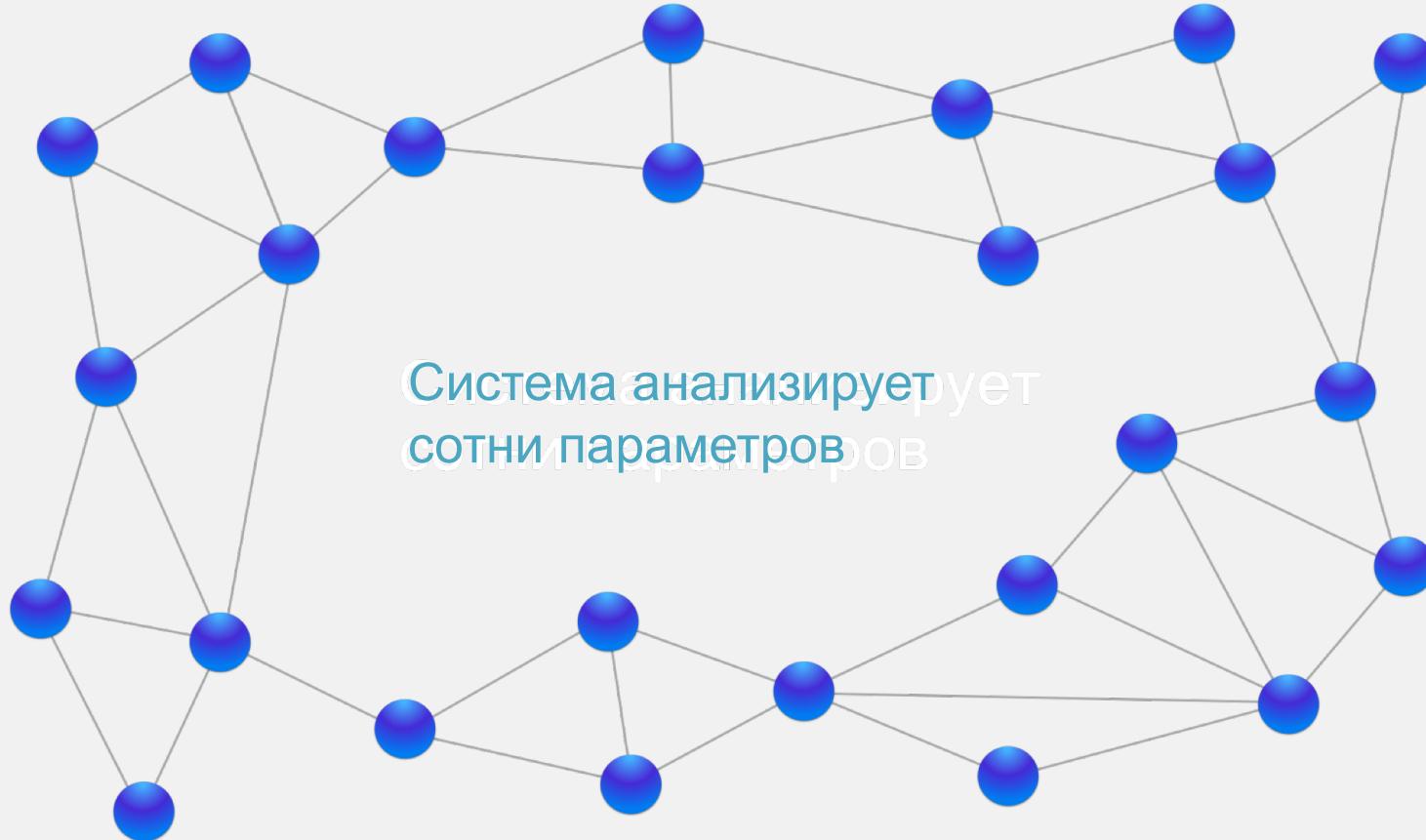
- Бизнес формулировка
 - Жизнь пользователей удобней – помогаем с простыми ответами
 - Пользователь тратит меньше времени на ответ в mobile
 - Репутационно – решение должно быть AI first
- Проверка гипотез
 - Большинство ответов в mobile – короткие
 - Большинство ответов в mobile – из golden set
- Сбор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



- 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



<https://new.mail.ru>
пробуйте