Методология ведения DS-проектов

О чём поговорим?

Сегодня на лекции

1

CrispDM

Методология ведения DS-проектов

2

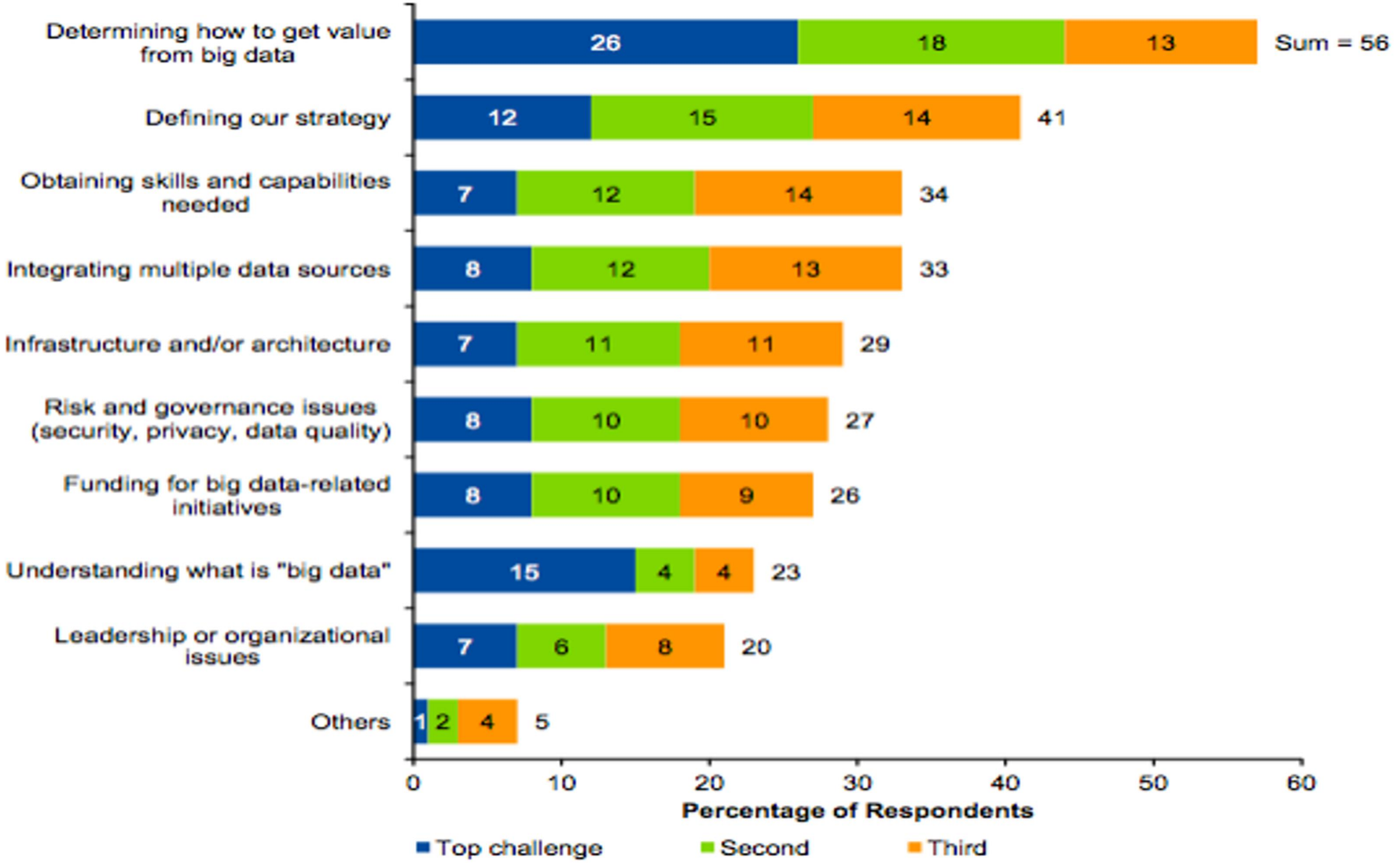
Смотрим ноутбук

С небес на землю

Crisp DM

Главные вызовы в области больших данных

Основной вызов — как получить value из больших данных?



Gartner, 2013

Кто помнит основные шаги Crisp DM? =)

Crisp DM


Бизнес-анализ
Анализ данных
Подготовка данных
Моделирование
Оценка решения
Внедрение
Данные

Business Understanding

1

Сбор справочной информации:

- составление бизнес-фона
- определение бизнес-целей
- критерии успеха проекта с точки зрения бизнеса

Business Understanding

2

Оценка ситуации:

- инвентаризация ресурсов
- требования, предположения и ограничения
- риски и непредвиденные обстоятельства
- анализ затрат, выгод

Business Understanding

3 Определение целей DS:

- цели DS
- критерии успеха DS

4 Создание плана проекта

Пример

Фаза	Время	Ресурсы	Риски
Business Understanding	1 неделя	Аналитики	Изменение условий
Data Understanding	3 недели	Аналитики	Проблемы с данными, технологиями
Data Preparation	4 недели	DS, DE	Проблемы с данными, технологиями
Modeling	2 недели	DS	Не удастся построить модель
Evaluation	1 неделя	Аналитики	Изменение условий, отсутствие результатов
Deployment	1 неделя	DS, Разработчик	Изменение условий, отсутствие результатов

Готовность к Data Understanding

С точки зрения бизнеса:

- что бизнес надеется получить от этого проекта?
- как будет определяться успех проекта?
- есть бюджет и ресурсы?
- есть ли доступ ко всем данным, необходимым для этого проекта?
- обсуждали ли вы и ваша команда риски и непредвиденные обстоятельства, которые могут возникнуть?
- оправдывают ли результаты вашего анализа затрат, выгод этот проект?

Готовность к Data Understanding

С точки зрения DS:

- как конкретно анализ данных может помочь достичь целей бизнеса?
- есть ли представление о том, какие методы DM могут дать наилучшие результаты?
- как узнать, что результаты являются достаточными для нужд бизнеса?
- как будут поставлены результаты моделирования: отчёт, сервис?
- включает ли план проекта все этапы CRISP-DM?
- обозначены ли риски и зависимости в плане?

Data Understanding

1. Соберите имеющиеся данные: собственные, внешние и т. п.
2. Опишите данные: количество, значения, связи и т. п.
3. Исследуйте данные
4. Изучите качество данных: отсутствующие данные, ошибки в данных, плохие метаданные и т. п.

Готовность к Data Preparation

- Все ли источники данных чётко определены и доступны? Известно ли о каких-либо проблемах или ограничениях?
- Определены ли ключевые атрибуты? Эти атрибуты помогли вам сформулировать гипотезы?
- Определён ли размер всех источников данных? Можно ли использовать подмножество данных, где это уместно?
- Рассчитаны базовые статистики для каждого интересующего атрибута?
- Каковы проблемы качества данных для этого проекта?
- Чётко ли определены этапы подготовки данных?

Data Preparation

- Выберите данные: разделите на train, test, выберите признаки
- Очистите данные: заполните пропуски, исправьте ошибки и т. п.
- Расширьте данные: новые признаки и т. п.
- Сохраните данные: подготовьте data frame для обучения модели и сохраните его

Готовность к Modeling

- На основании вашего первоначального исследования смогли ли вы выбрать подходящие подмножества данных для моделирования?
- Эффективно ли вы очистили данные?
- Правильно ли соединены разные наборы данных?
- Задокументировали ли вы все сделанные шаги по подготовке данных?

Modeling

- Выберите метод моделирования
- Постройте модель
- Оцените модель
- Опишите результат

Готовность к Evaluation

- Дала ли модель понятные результаты? Есть ли очевидные несоответствия, которые требуют дальнейшего исследования?
- Исследовано более одного типа модели и результаты сравнены?
- Можно ли поставить модель заказчику?

Evaluation

1 Оцените результаты:

- чётко ли представлены результаты?
- есть ли какие-нибудь новые инсайты?
- может ли модель и результаты быть применимыми к бизнесу?
- какие дополнительные вопросы появились после моделирования?

2 Просмотрите процесс:

- что пошло не так, и как это можно исправить?
- есть ли альтернативные решения, действия, которые могли бы быть выполнены?

3 Определите следующие шаги

Deployment

1. Планирование внедрения:

- для каждой модели создайте план внедрения
- определите все проблемы внедрения и составьте план на случай непредвиденных обстоятельств

2. Планирование мониторинга и технического обслуживания:

- определите модели и результаты, которые требуют поддержки
- Как понять, что модель перестала быть актуальной?
- Что делать в этом случае?

3. Проведение итогового обзора проекта

CRISP-DM

Business Understanding/ **Бизнес-анализ**	Data Understanding/ **Анализ данных**	Data Preparation/ **Подготовка данных**	Modeling/ **Моделирование**	Evaluation/ **Оценка решения**	Deployment/ **Внедрение**
Determine Business Objectives/ **Определение бизнес-целей**	Collect Initial Data/ **Сбор данных**	Select Data/ **Выборка данных**	Select Modeling Techniques/ **Выбор алгоритмов**	Evaluate Results/ **Оценка результатов**	Plan Deployment/ **Внедрение**
Assess Situation/ **Оценка текущей ситуации**	Describe Data/ **Описание данных**	Clean Data/ **Очистка данных**	Generate Test Design/ **Подготовка плана тестирования**	Review Process/ **Оценка процесса**	Plan Monitoring and Maintenance/ **Планирование мониторинга и поддержки**
Determine Data Mining Goals/ **Определение целей аналитики**	Explore Data/ **Изучение данных**	Construct Data/ **Генерация данных**	Build Model/ **Обучение моделей**	Determine Next Steps/ **Определение следующих шагов**	Produce Final Report/ **Подготовка отчета**
Produkt Project Plan/ **Подготовка плана проекта**	Verify Data Quality/ **Проверка качества данных**	Integrate Data/ **Интеграция данных**	Assess Model/ **Оценка качества моделей**		Review Project/ **Ревью проекта**
		Format Data/ **Форматирование данных**			

Кейс.
Анализ данных по методологии CrispDM

https://colab.research.google.com/drive/1tpM0qvQGidnBoo6aVZU0_ObotzntgEpH?usp=sharing

Итоги

Итоги

Методология CrispDM включает в себя 6 этапов работы над задачей:

- ✓ бизнес-анализ,
- ✓ анализ данных,
- ✓ подготовку данных,
- ✓ моделирование,
- ✓ оценку решения,
- ✓ внедрение