



Big Data и аналитические модели для решения бизнес-задач

28.11.2018

Казань, 2018

Содержание



1. Обработка данных предназначена для решения проблем
2. С чего начинается Big Data?
3. Постановка задачи
4. Технология обработки данных
5. Визуализация данных и принятие решений
6. Масштабирование
7. Предложение
8. Решения, с которыми мы работаем



Обработка данных предназначена для решения проблем

- Правильная **формулировка проблем** является важным этапом работ. Независимо от того, насколько хорошо проводятся последующие этапы, процесс может быть бесполезным, если диагноз проблем был произведен не достаточно корректно.
- Главный вопрос: **в чем заключается проблема?**
- Обработка, анализ данных не является самостоятельным упражнением. **Целью анализа данных является предоставление информации, связанной с компонентами проблемы.**

Обработка данных предназначена для решения проблем

- Когда мы четко сформулируем проблему, то можем обнаружить, что большая часть наших данных не имеет отношения к проблеме!
- В этом смысле вся дальнейшая работа с данными была бы пустой тратой времени и ресурсов.
- Работа с данными должна осуществляться для решения проблем/задач.



С чего начинается Big Data?

Постановка
задачи

Технология
обработки
данных

Масштаби-
рование

Визуализация



Постановка задачи

Важно понимание того, как полученные данные будут использоваться при принятии решений

Диагностика
проблемы



Изучение предметной области, опыта других и т.п., интервьюирование специалистов

Определение
состава
данных



подбор данных,
относящихся к
проблемам

Анализ
данных



Формулировка
гипотез,
ограничений



относительно тех проблем,
по которым подбор данных
признан успешным

Технология обработки данных

Подготовка данных

Нормализация, сортировка и группировка данных, ранжирование, очистка данных, стат. корректировка данных

Построение статистических моделей

SPSS, методы мат. статистики, дисперсионный, регрессионный анализ, ANOVA-модели и т.п.

Проверка качества моделей, оценка вероятности

стат. тесты, проверка на автокорреляцию остатков, гетероскедастичность и т.п.

Адаптация моделей

Преобразование и подбор переменных и т.п.)

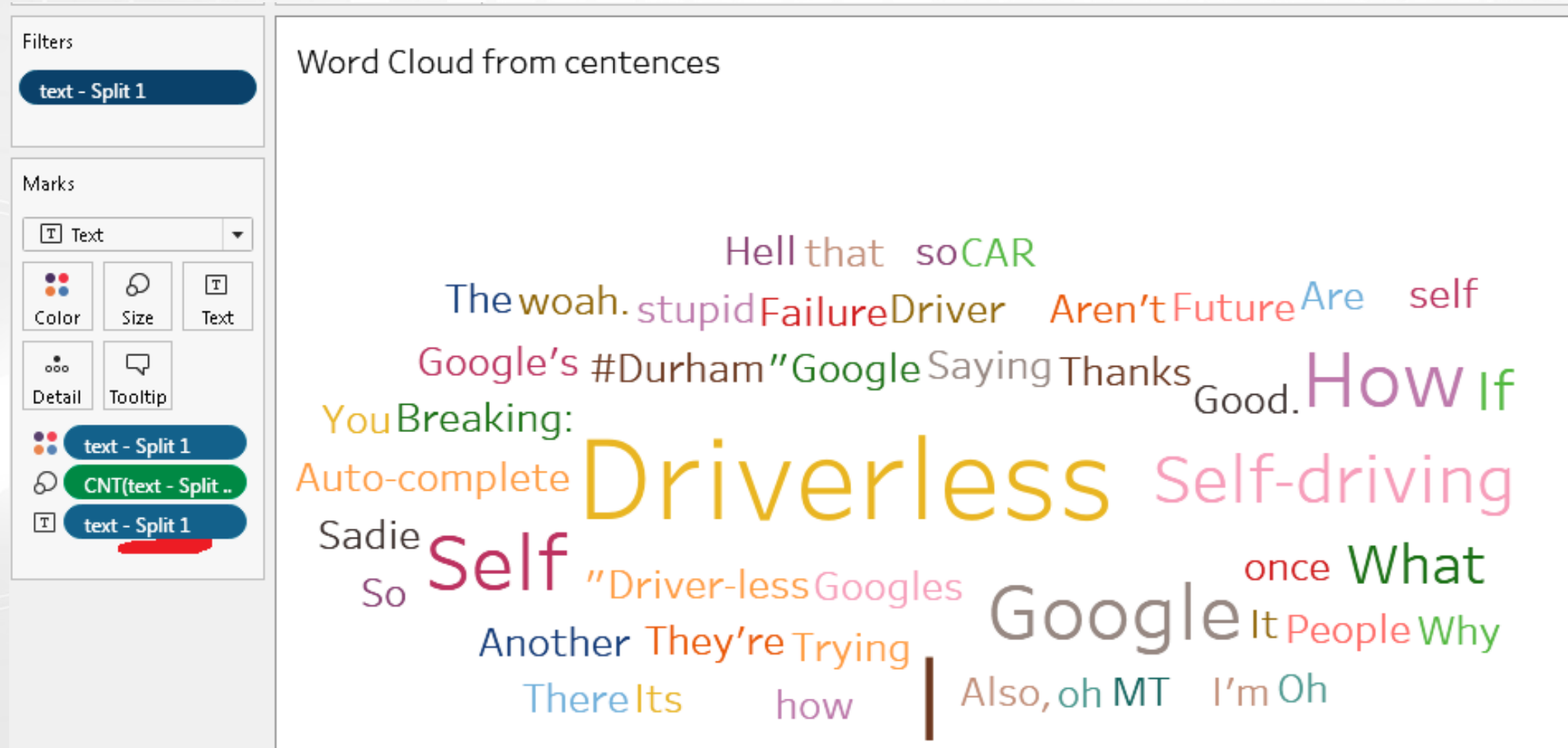
Интерпретация результатов

относительно сформулированных проблем

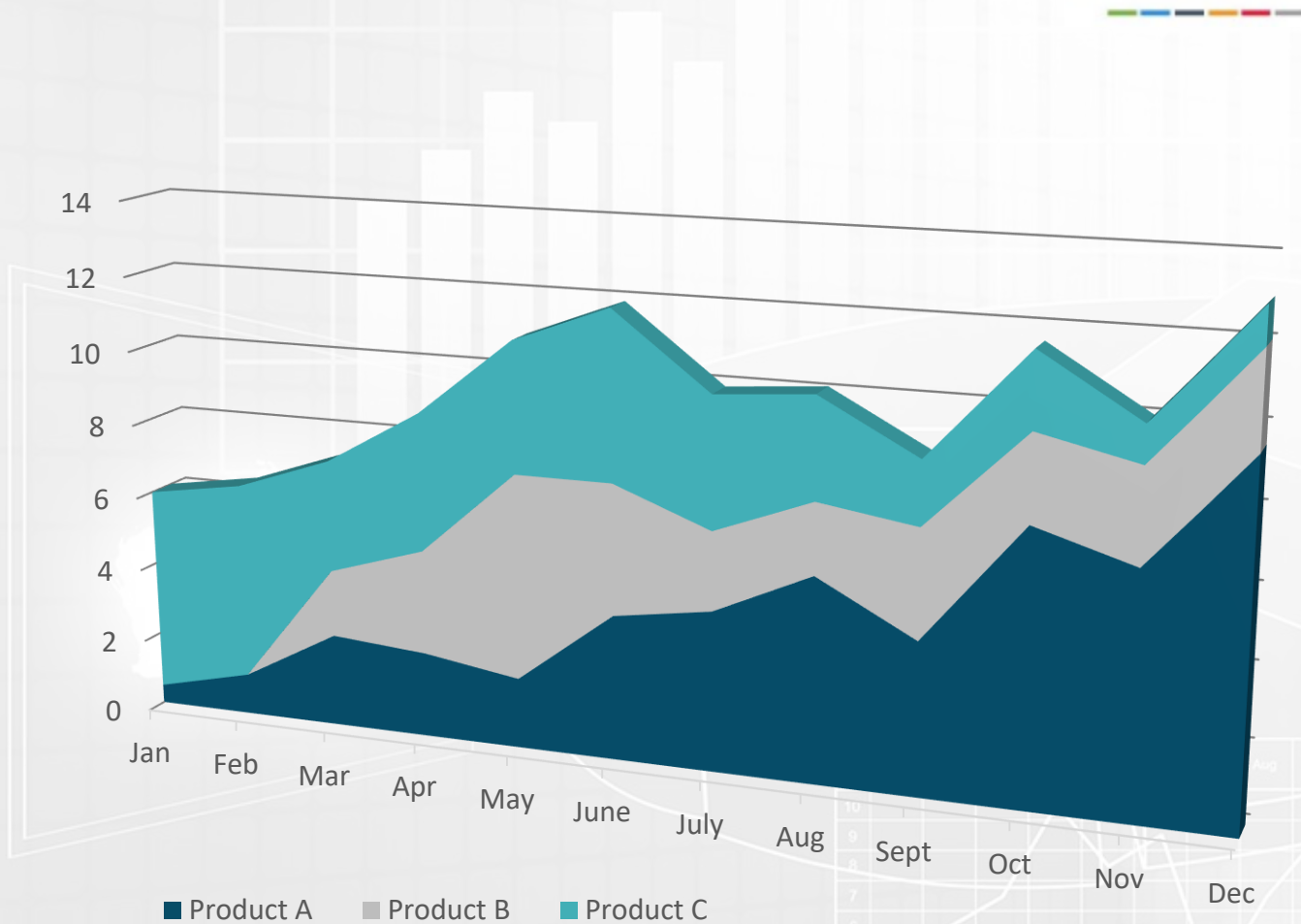
Выводы качественной статистически значимой модели можно использовать для оперативного управления



Анализ естественного языка – облако тегов



Визуализация результатов и принятия решений



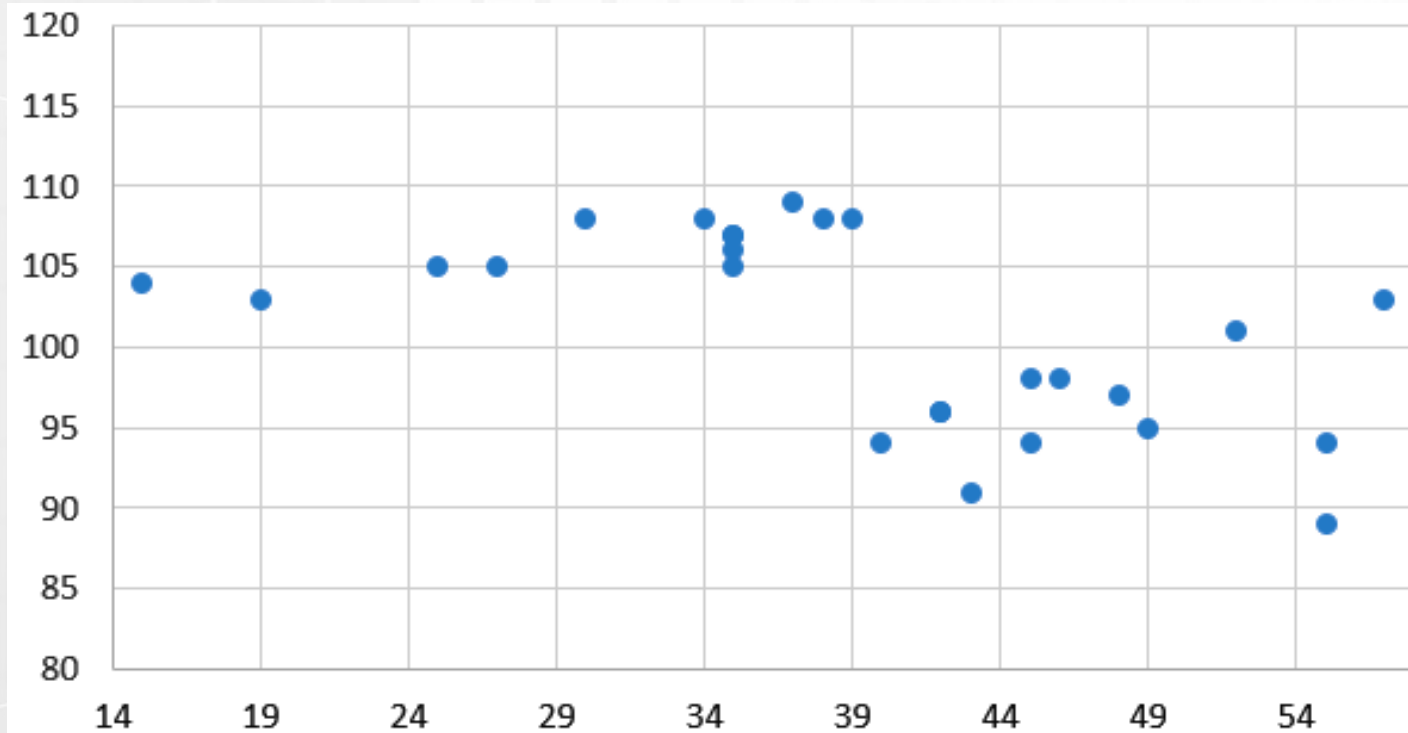
- **Визуализация**
результатов
моделирования
для принятия
решений

The background is a dark blue-grey color with faint, semi-transparent overlays of various data visualizations. These include a bar chart at the top left, a world map on the right, a line graph with a grid at the bottom right, and a map of Australia on the left. The main text is centered and reads:

*“Цель визуализации —
понимание,
а не картинка”*

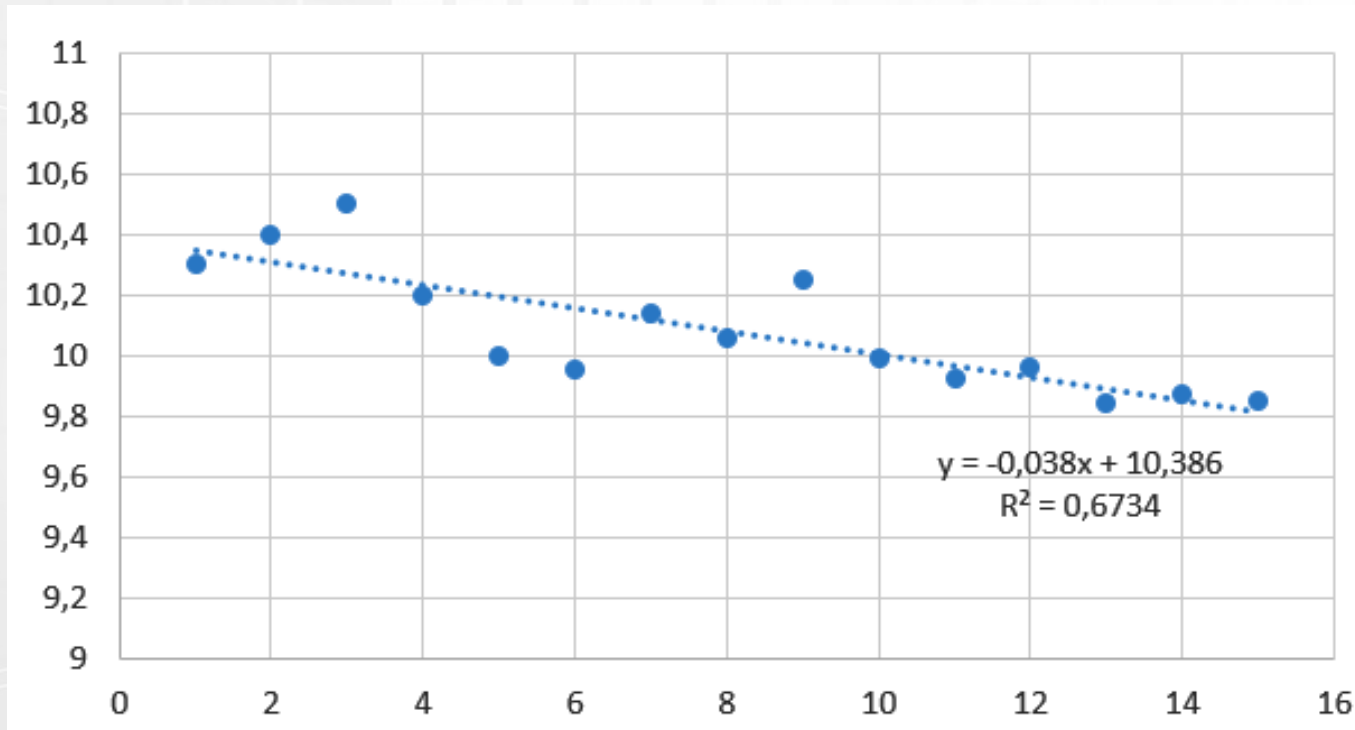
*Бен Шнейдерман,
известный американский научный деятель*

Кластеризация



- Зависимость расходов на отдых от возраста
- Отсутствие линейной взаимосвязи

Простая регрессия



ЦЕЛЬ:

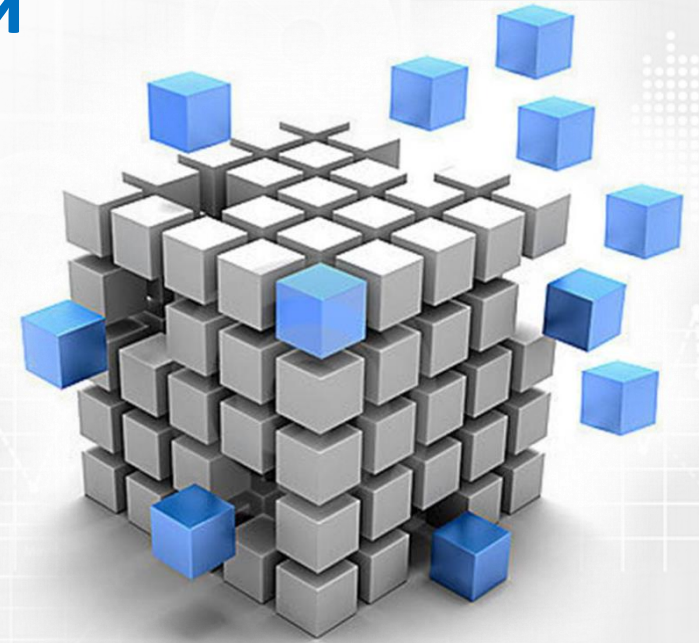
Прогноз времени
забега 100 м.

победителя среди
мужчин на

Олимпийских играх, на
основании данных с
1948 г.

Масштабирование

- Успешное моделирование позволит определить **какие данные следует собирать и обрабатывать в дальнейшем** на регулярной основе **для решения проблем** и поддержки принятия решений.
- Начнется процесс накопления больших данных, Ваши данные станут расти в нужном Вам направлении.
- Вы сможете заставить эти данные работать на Вас (делать прогнозы и принимать решения).



Дальнейшее развитие

**Аутсорсинг, услуга
«Построение
аналитической
модели» -**

если опыт
использования
результатов
моделирования будет
требоваться на
**периодической
основе -**



**ТЗ на систему работы с
Big Data, прогнозные
системы и проект её
использования –**

если опыт
использования
результатов
моделирования будет
требоваться на
оперативной основе

POLYMATICA – аналитическая платформа



Платформа **AggreGate** для сбора данных с физических устройств



Big data в организациях: резюме

1. Каждый программный продукт по работе с данными надо **адаптировать** под конкретную бизнес-задачу*
2. Продвинутая аналитика сокращает время на подготовку отчётов и принятие решений
3. Ни одна система **не даёт значимые результаты без участия человека**
4. **80% времени** занимает **процесс подготовки и очистки данных****
5. Вопрос доверия к данным носит субъективный характер

* IBM Institute business value

** Dasu T, Johnson T (2003). [Exploratory Data Mining and Data Cleaning. Wiley-IEEE](#)

Мы можем пройти этот путь вместе





Свяжитесь с нами

Компания АБАК
Казань,
ул. Аделя Кутуя, д. 159

+7 (843) 299-75-00
abak@abak.ru