ЦИПР

ЦИФРОВАЯ
ИНДУСТРИЯ
ПРОМЫШЛЕННОЙ
РОССИИ

ЦИПР 2018

6-8 июня
Иннополис

О ПРОБЛЕМАХ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА БОЛЬШИХ ДАННЫХ

Чехович Юрий Викторович, зав. отделом Интеллектуальных систем ФИЦ ИУ РАН, к.ф.-м.н.

НЕСКОЛЬКО СЛОВ О НАШЕЙ КОМАНДЕ

- Вычислительный центр РАН
- Кафедры МФТИ и ВМК МГУ
- Несколько внедряющих компаний:

  Форексис, Гудфоркаст, Антиплагиат
- Большой опыт решения задач анализа данных

КЛЮЧЕВОЙ ФАКТОР В АНАЛИЗЕ БОЛЬШИХ ДАННЫХ

Что главное

Для успешного создания и развития технологий хранения и анализа больших данных необходимо выделить ключевые факторы будущего успеха:

- Математические методы
- Оборудование для хранения, обработки и передачи
- Подготовка специалистов
- Кооперация

КЛЮЧЕВОЙ ФАКТОР В АНАЛИЗЕ БОЛЬШИХ ДАННЫХ

Что главное

Для успешного создания и развития технологий хранения и анализа больших данных необходимо выделить ключевые факторы будущего успеха:

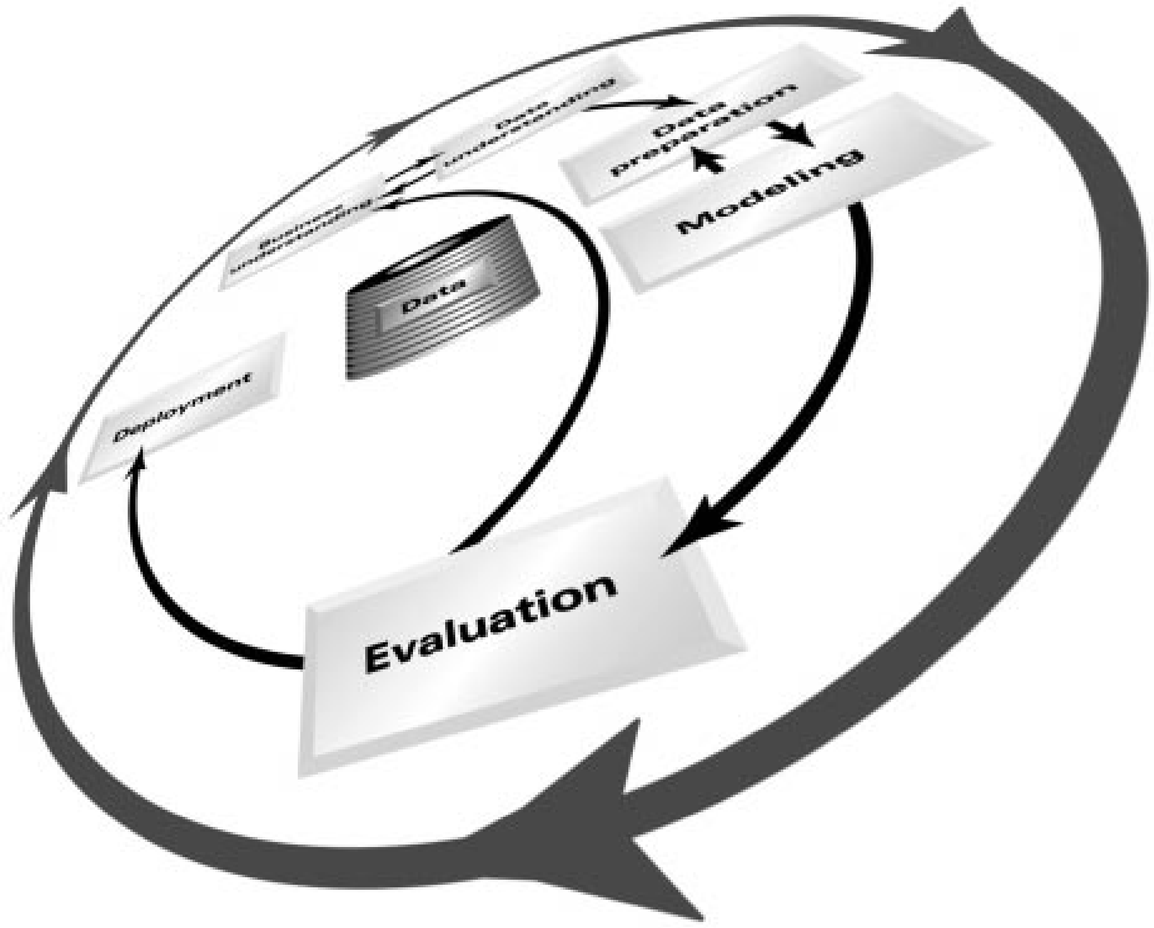
- Математические методы
- Оборудование для хранения, обработки и передачи
- Подготовка специалистов
- Кооперация

Методология!

CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING

CRISP-DM

Методология ведения промышленных проектов интеллектуального анализа данных

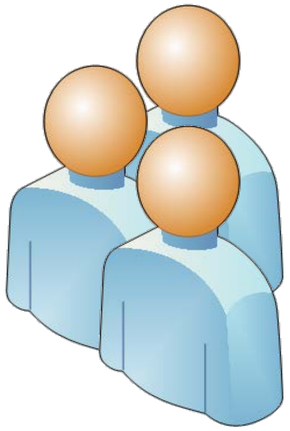


ИСТОРИЯ И ОСОБЕННОСТИ CRISP-DM

Teradata a division of NCR

SPSS

IBM



SIG (Special Interest Group)

1996
Начали задумываться

1997-1999
Проведение конференций, анализ возможных сценариев

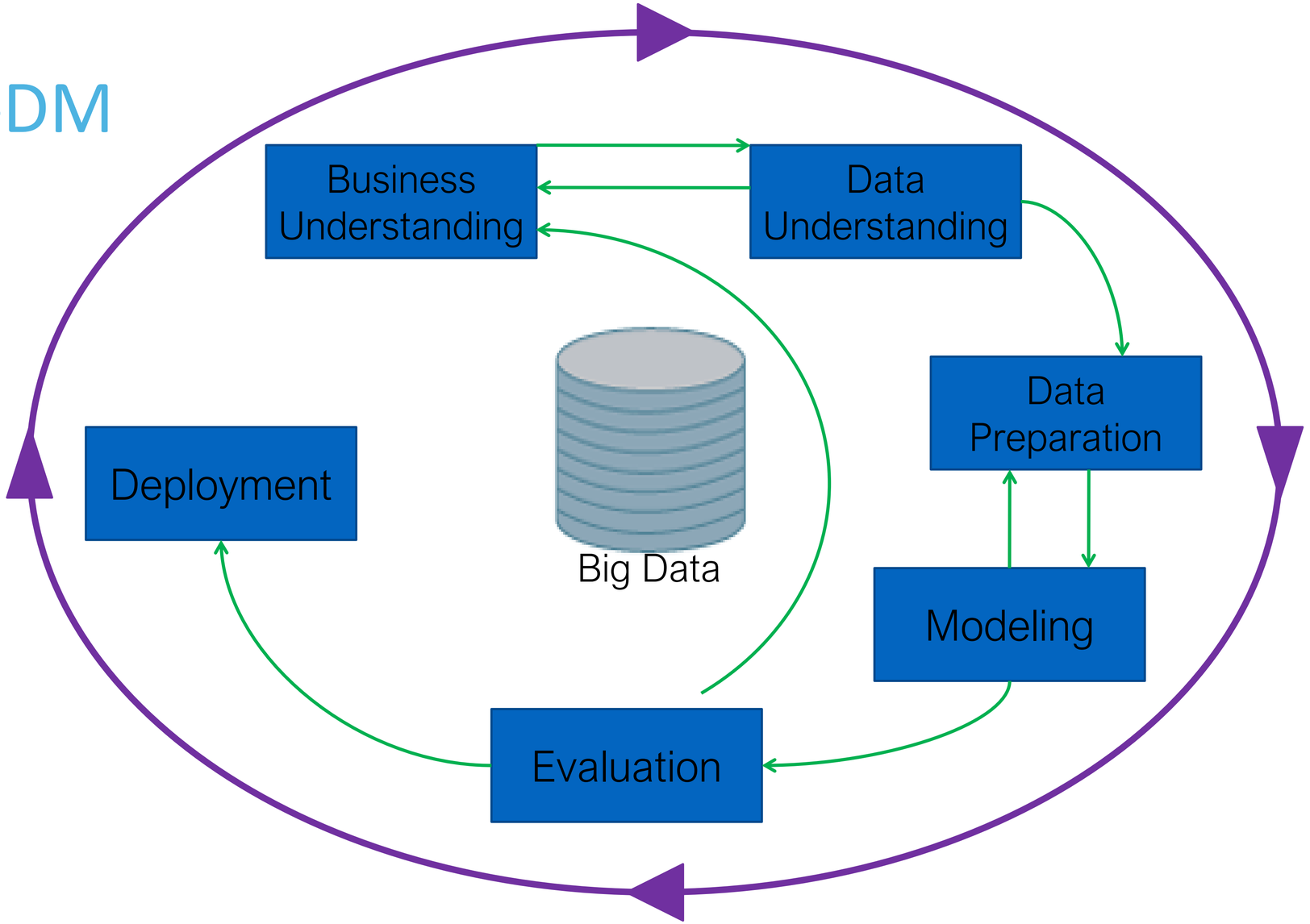
2000
CRISP-DM 1.0

ИСТОРИЯ И ОСОБЕННОСТИ CRISP-DM

- Стандартный процесс для ведения исследовательских проектов в области анализа данных
- Не зависит от прикладной области
- Не зависит от используемого ПО
- Не зависит от применяемых алгоритмов и решаемых задач

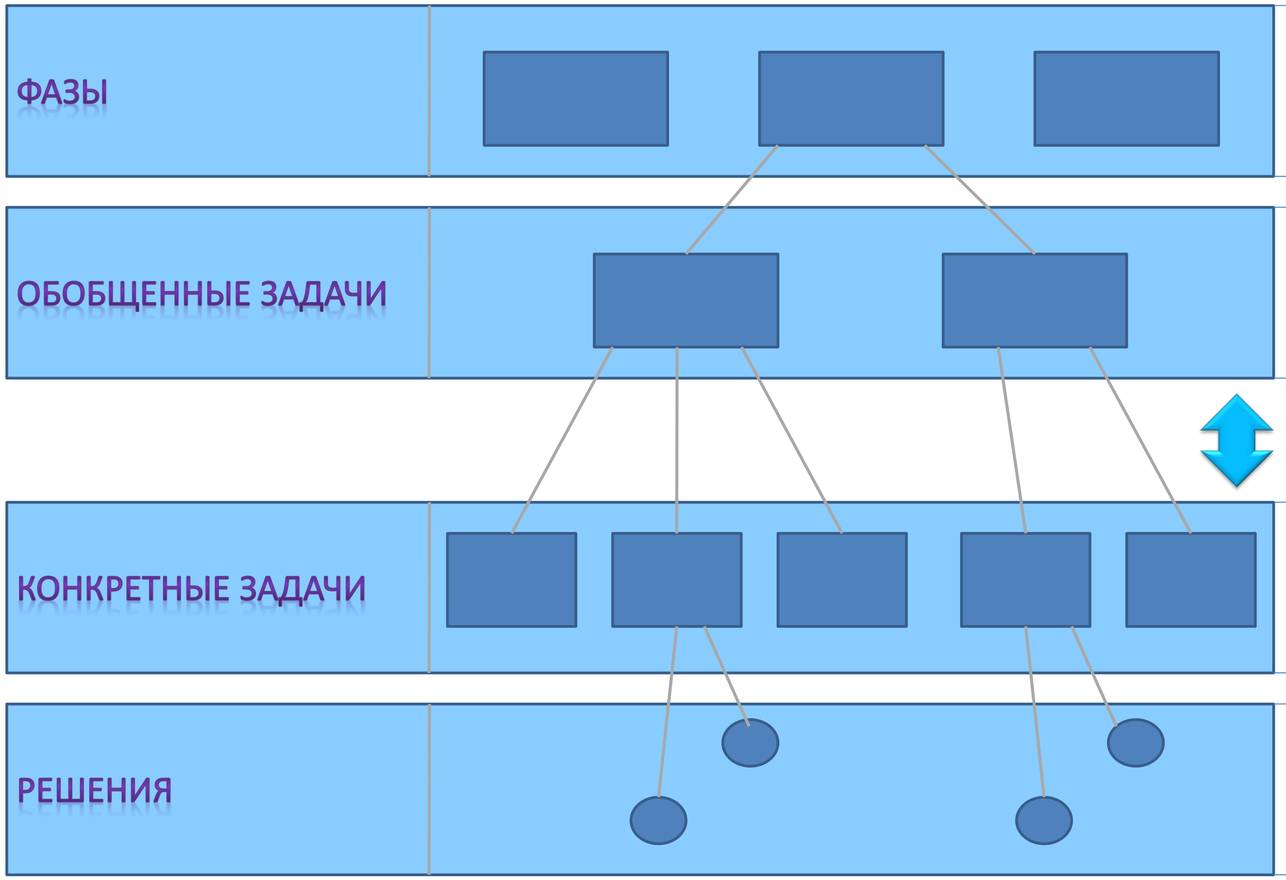
ОСНОВНЫЕ ПРИНЦИПЫ CRISP-DM

Фазы CRISP-DM



ОСНОВНЫЕ ПРИНЦИПЫ CRISP-DM

Иерархическая декомпозиция



Состав фаз определен стандартом
Пример фазы: **Подготовка данных**

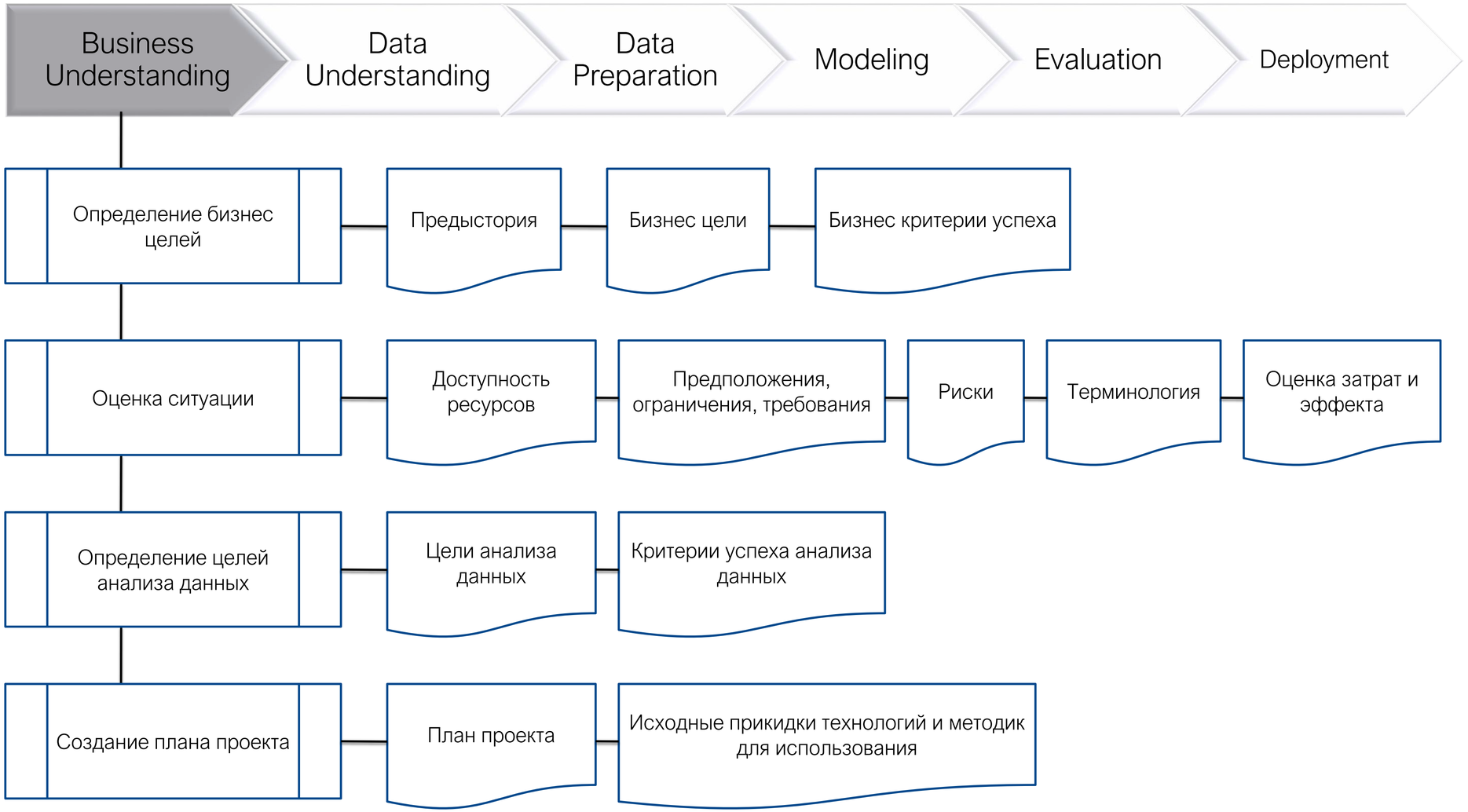
Состав обобщенных задач определен стандартом
Пример: **Очистка данных**

Конкретизация

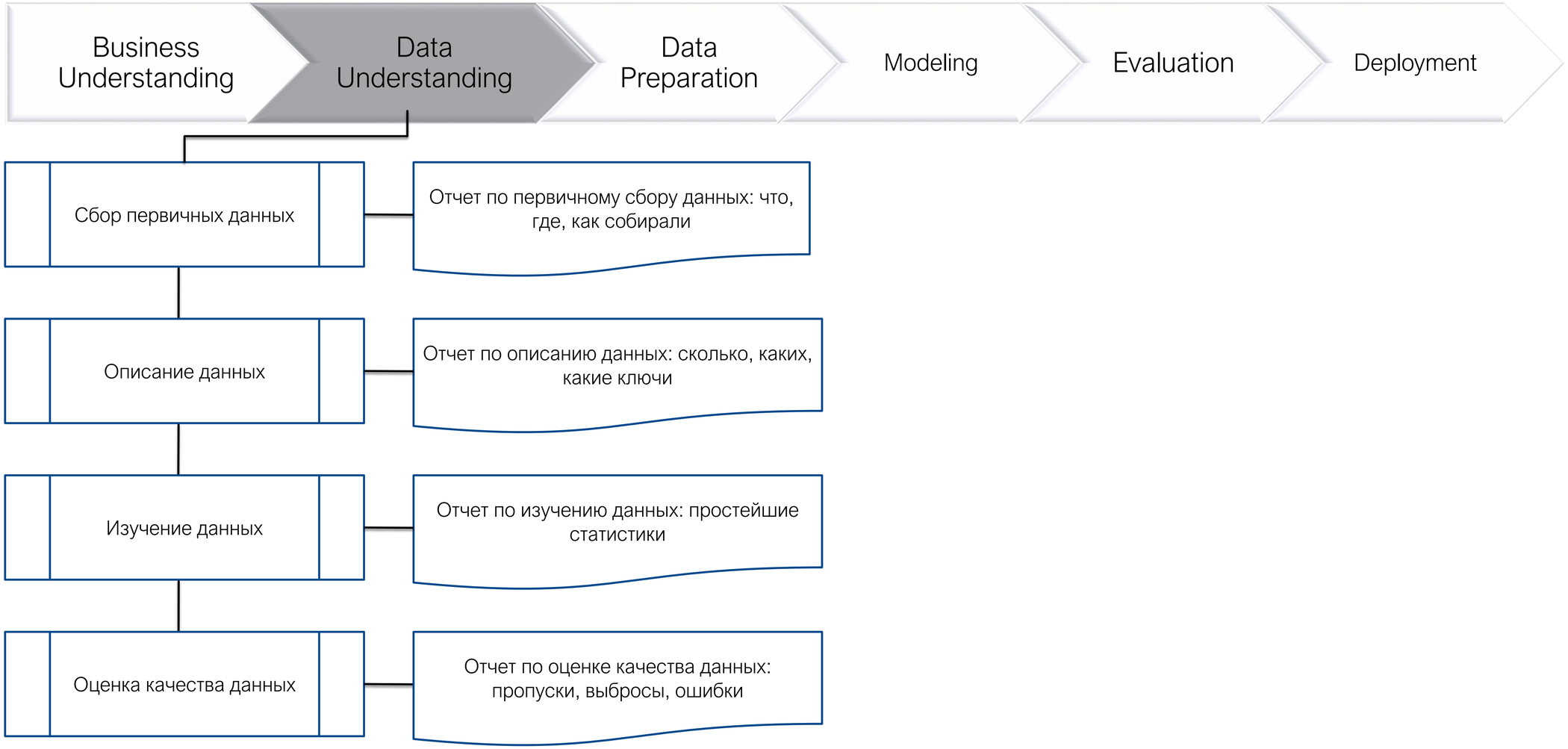
Состав конкретных задач определяются при планировании
Пример: **Борьба с пропусками**

Конкретные решения определяют исследователи
Пример: **Если пропущен числовой атр. – среднее**

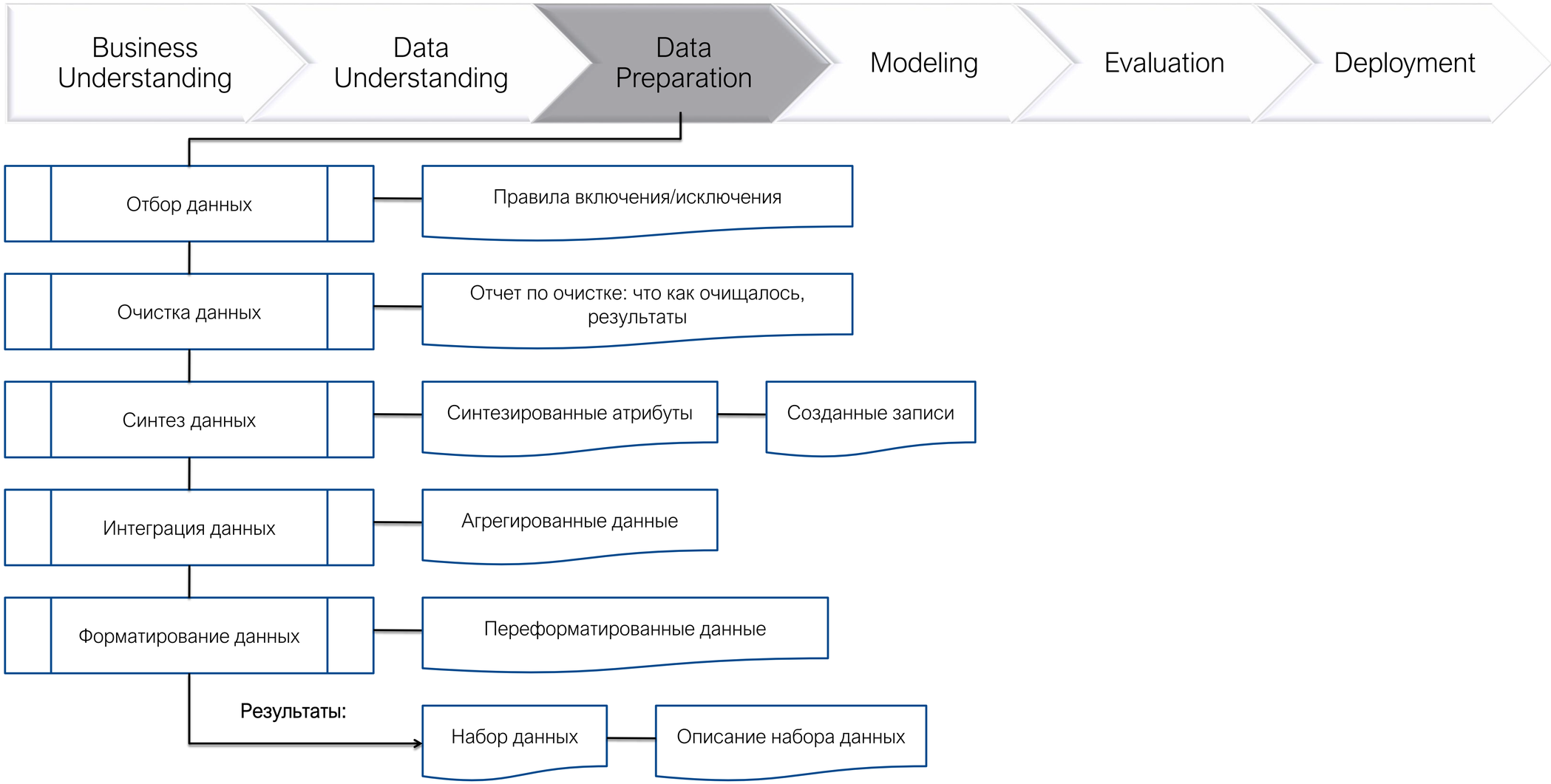
СОДЕРЖАТЕЛЬНАЯ ПОСТАНОВКА ЗАДАЧИ (BUSINESS UNDERSTANDING)



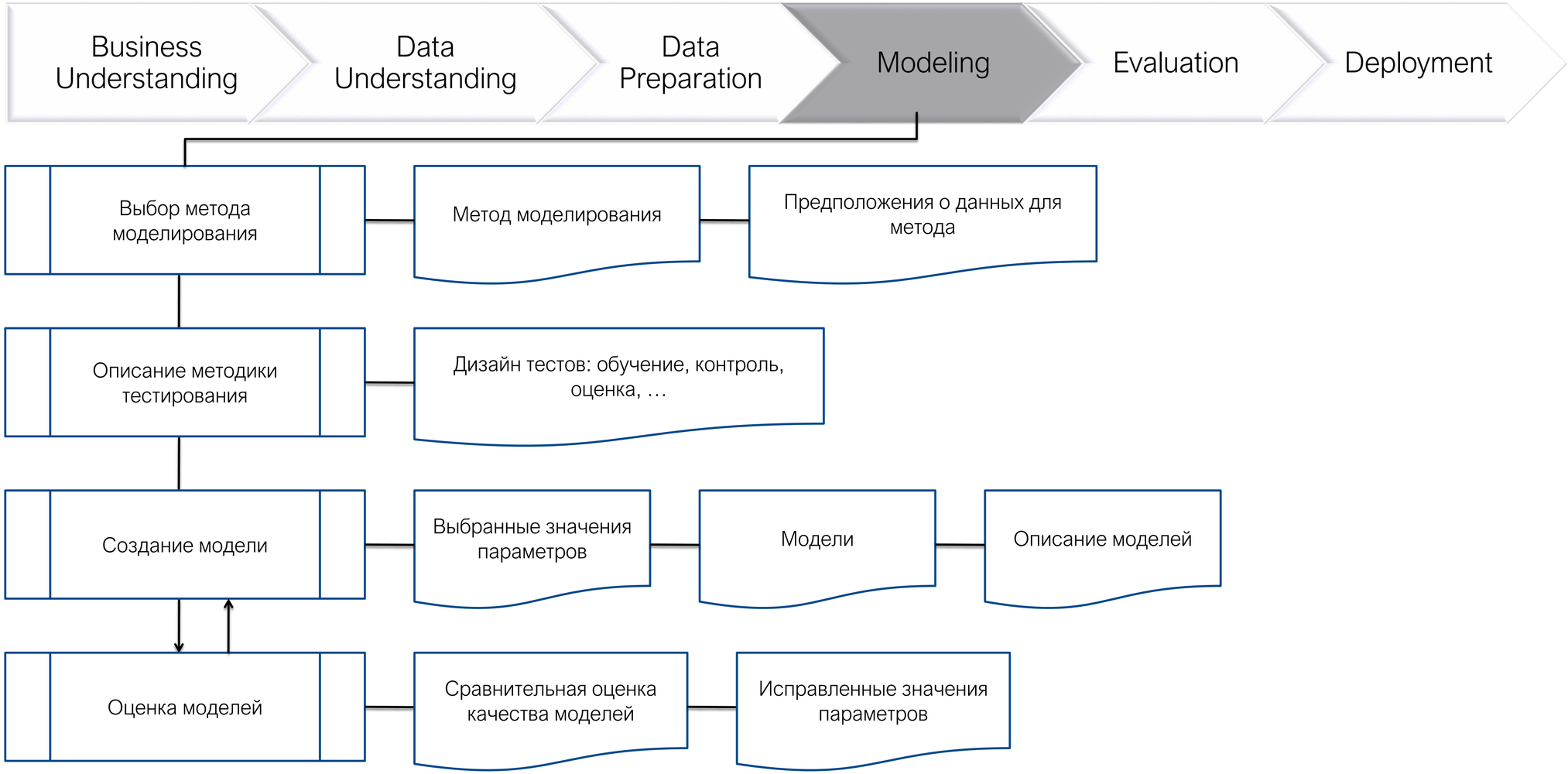
АНАЛИЗ ИСТОЧНИКОВ ДАННЫХ (DATA UNDERSTANDING)



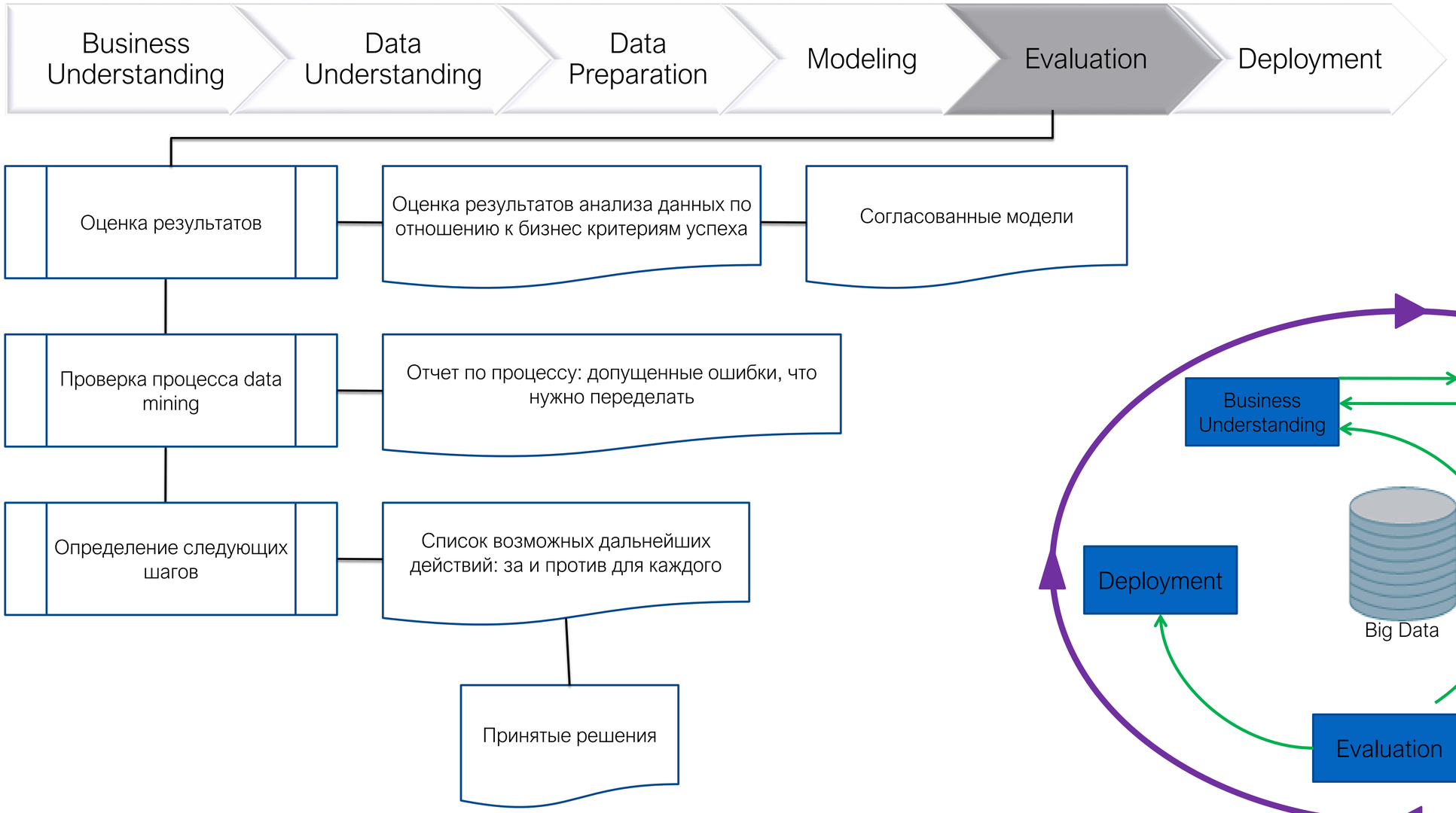
ПРЕДОБРАБОТКА ДАННЫХ (DATA PREPARATION)



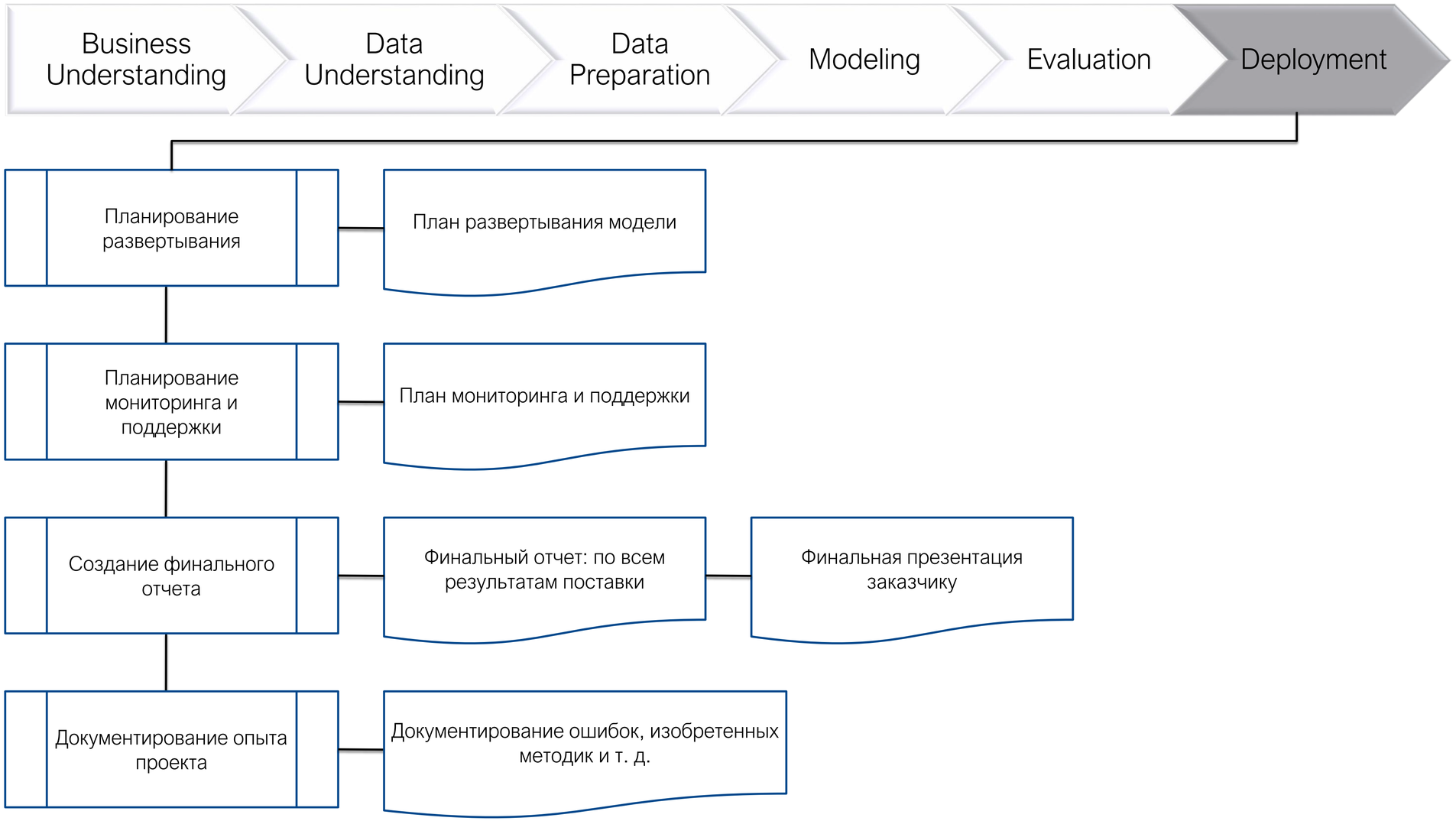
ПОИСК РЕШЕНИЯ (MODELING)



ОЦЕНКА РЕШЕНИЯ (EVALUATION)



РАЗВЕРТЫВАНИЕ (DEPLOYMENT)



ЗАМЕЧАНИЯ И ВЫВОДЫ

- CRISP-DM не панацея
- CRISP-DM не помешает таланту
- CRISP-DM позволяет накапливать и передавать опыт
- CRISP-DM создавался как стандарт анализа данных. Большие данные не проблема?
- Не проблема!

ВОПРОСЫ

Спасибо за внимание!