Математические и инструментальные методы машинного обучения

3. Визуализация данных

Предпосылки к использованию интеллектуального анализа данных

- Данные имеют неограниченный объем
- Данные являются разнородными (количественными, качественными, текстовыми)
- Результаты должны быть конкретны и понятны
- Инструменты для обработки сырых данных должны быть просты в использовании

Парадокс:

Чем больше данных, тем меньше знаний

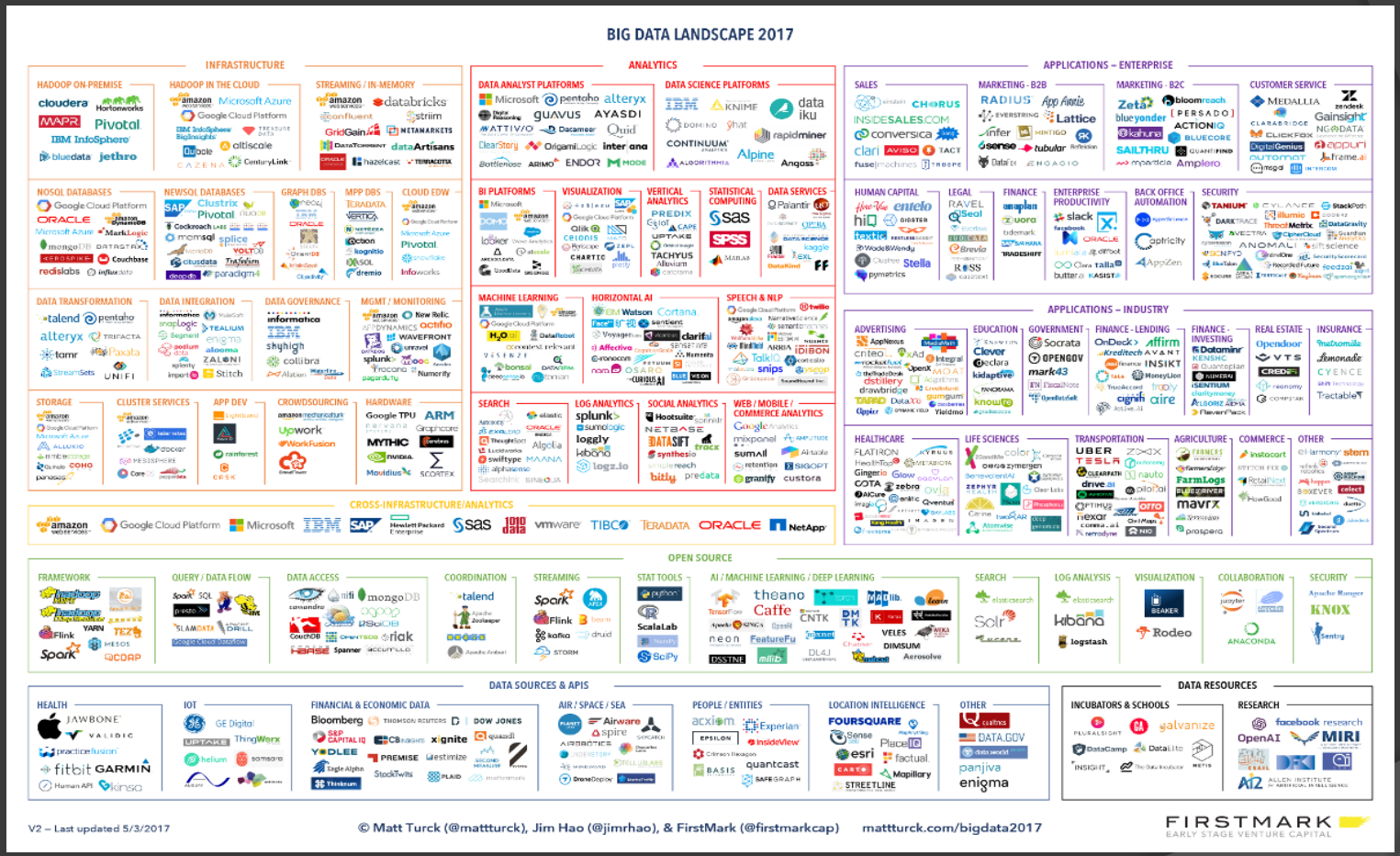
Data Mining - это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Пирамида знаний

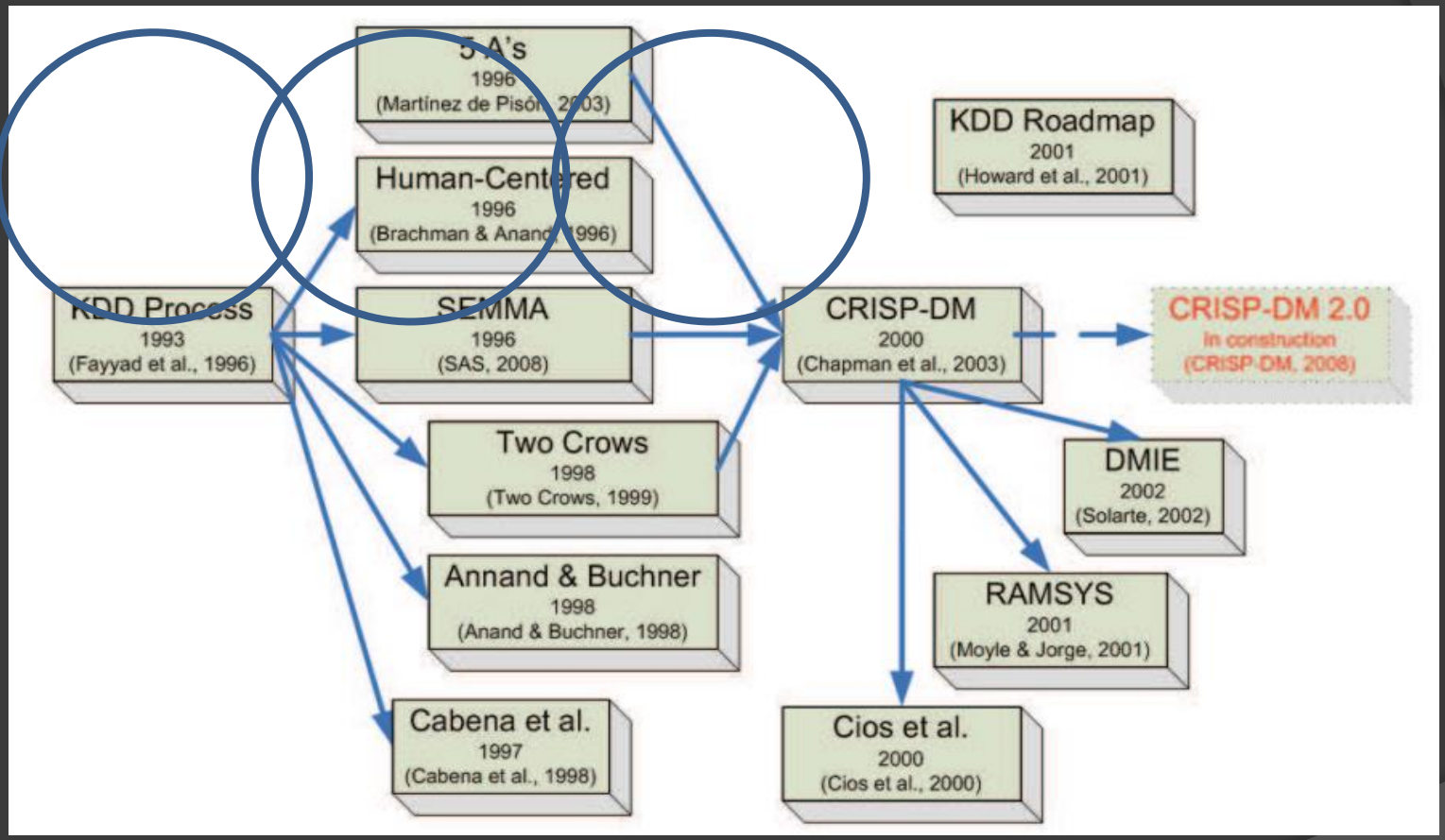
Применение интеллектуального анализа данных

- Реклама и продвижение товара
  - Какова эффективность рекламы?
- Перекрестные продажи
  - Какие продукты покупатель готов дополнительно приобрести?
- Обнаружение мошенничества
  - Правильные ли сведения были поданы?
- Удержание клиента
  - Какие клиенты готовы разорвать договор?
- Управление рисками
  - Выдавать ли кредит данному заёмщику?
- Сегментирование потребителей
  - Выдавать ли кредит данному заёмщику?

Технологии больших данных

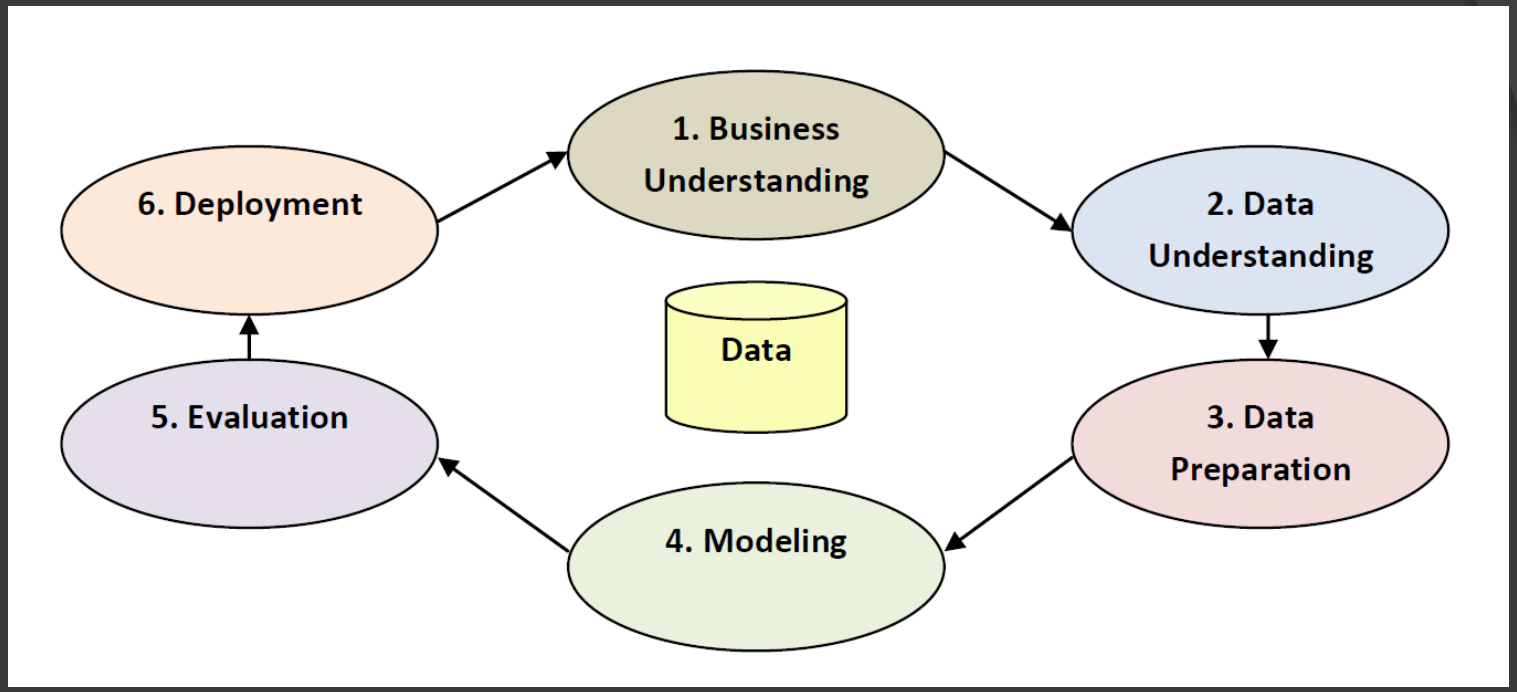


Развитие методологий анализа данных

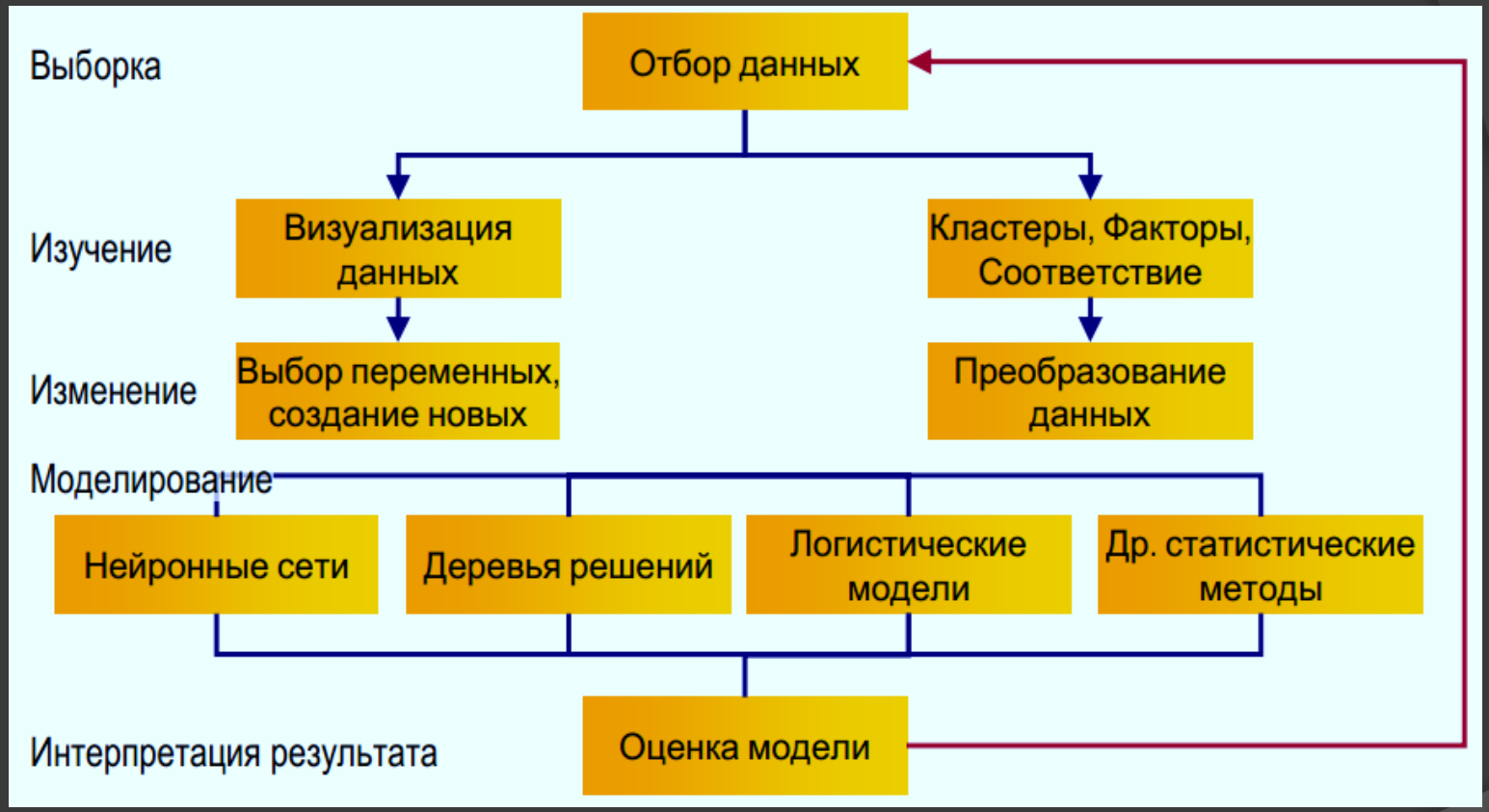


Этапы процесса анализа данных по методологии KDD

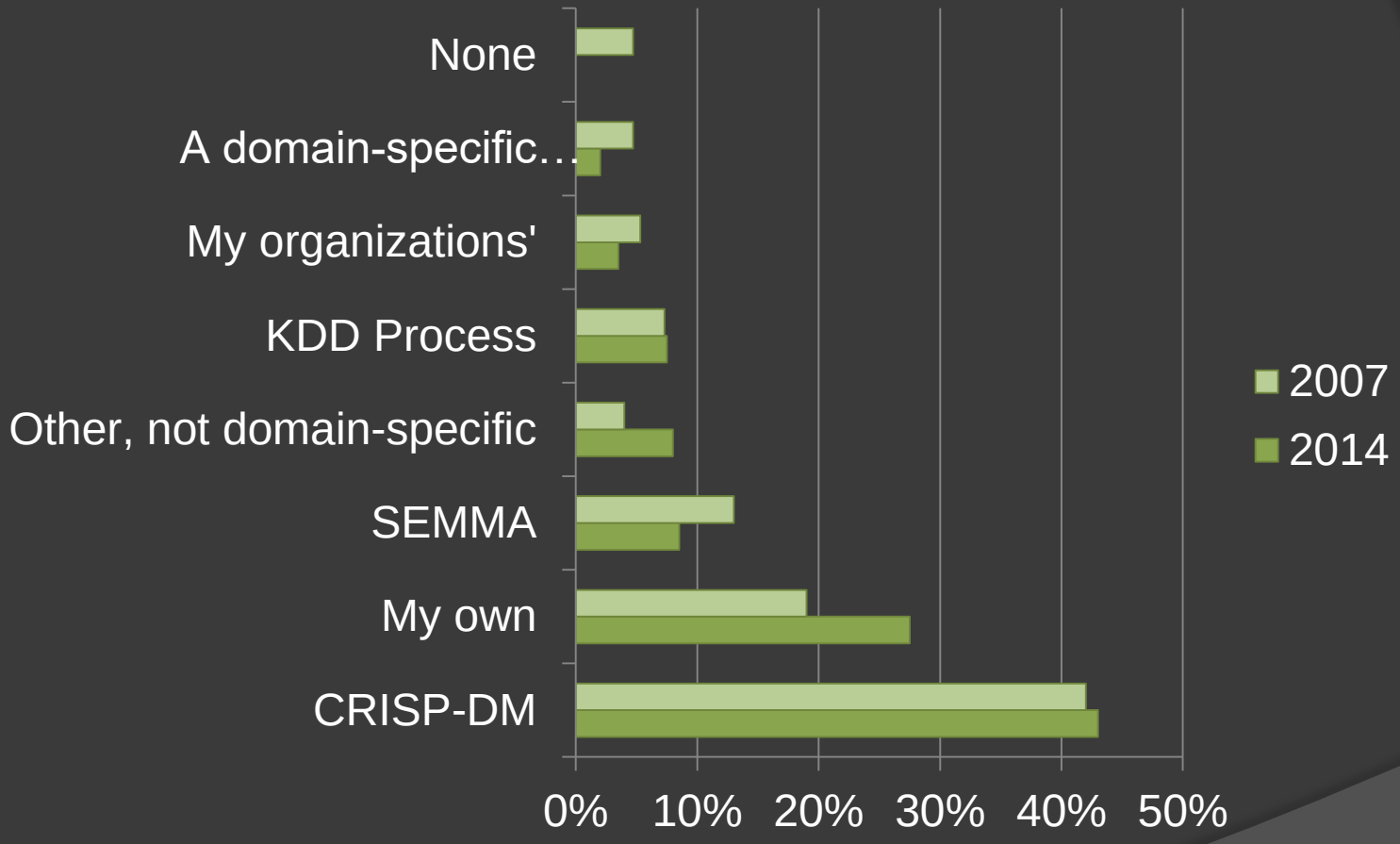
Этапы процесса анализа данных по стандарту CRISP-DM



Этапы процесса анализа данных по методологии SEMMA



Использование различных методологий в анализе данных



http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Типы задач анализа данных

Data-Intensive Domains

Основной целью анализа в областях с интенсивным использованием данных (DID) является извлечение знаний из комплексных данных, организованных в сетевые инфраструктуры, таких как хранилища данных, сети, облака.

Анализ Больших Данных на наличие аномалий

- Анализ социальных сетей
- Анализ большого потока информации из научных экспериментов
- LSST(Large Synoptic Survey Telescope, 2020) – планируется 200,000 изображений в год (1.28 ПБ)
- LHC(Large Hadron Collider) – генерирует сам и в результате вычислений около 10-15 ПБ в год

Data-Intensive Domains