Анализ данных и машинное обучение, ч. 2

Лекция 1. Процесс анализа данных и применения моделей машинного обучения. Подготовка данных. Анализ аномалий

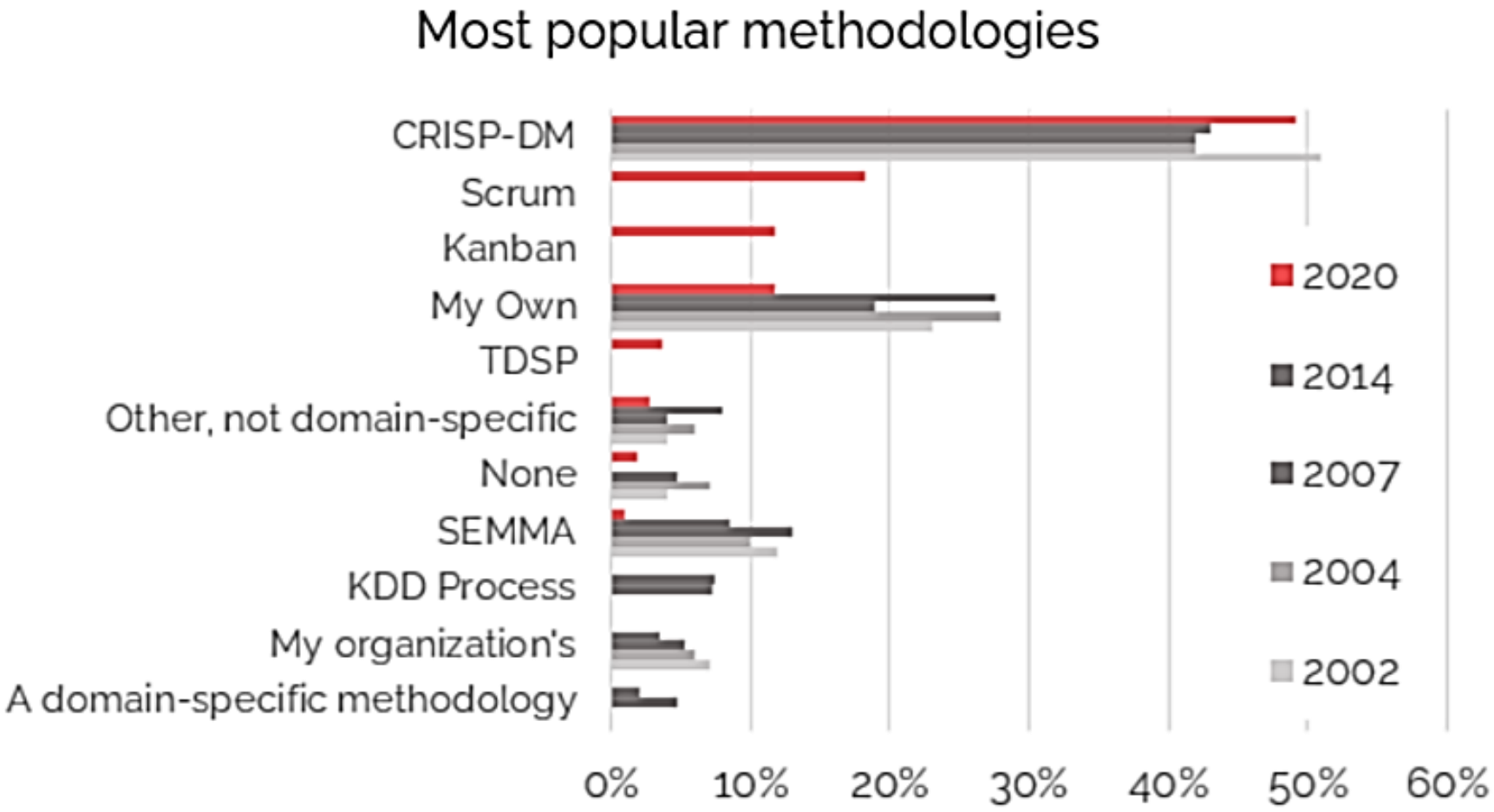
Киреев В.С.,
к.т.н., доцент

Москва, 2025

Жизненный цикл науки о данных

Жизненный цикл науки о данных - это повторяющийся набор шагов, которые вы выполняете для реализации проекта или анализа. Поскольку каждый проект и команда в области науки о данных отличаются друг от друга, каждый конкретный жизненный цикл науки о данных отличается. Однако большинство проектов в области науки о данных, как правило, проходят один и тот же общий жизненный цикл, состоящий из этапов, связанных с наукой о данных.

Современные методологии проектов в DS



Современные методологии проектов в DS

CRISP-DM

Scrum

Data-Driven Scrum

Kanban

SEMMA

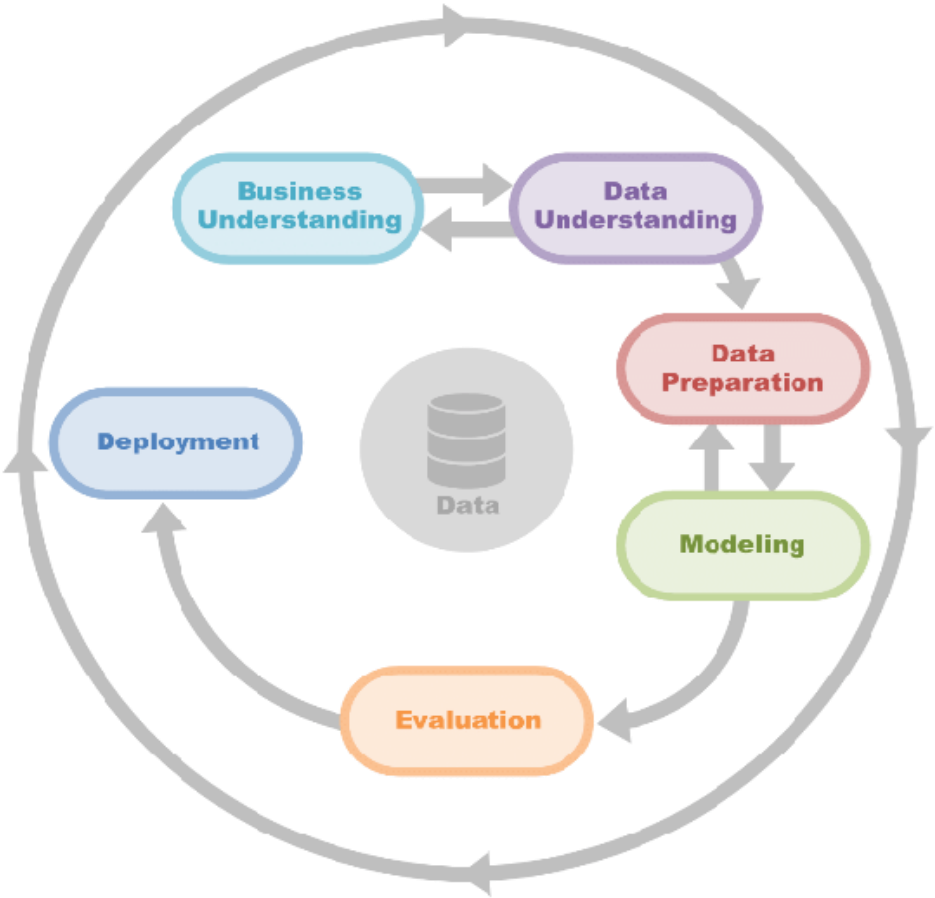
TDSP

KDD

DOMINO

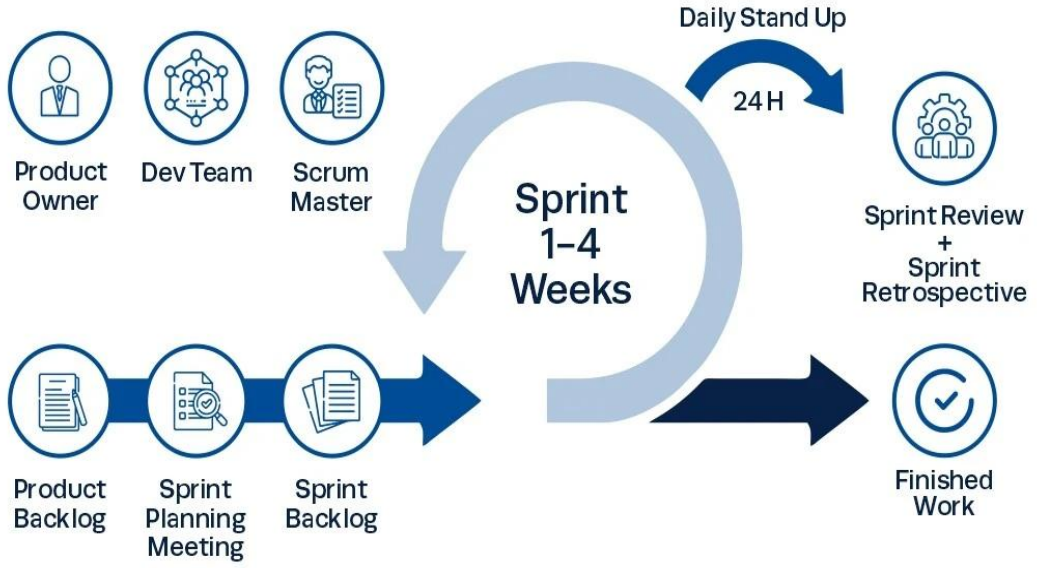
CRISP-DM

Межотраслевой стандартный процесс интеллектуального анализа данных (CRISP-DM) - это модель процесса, состоящая из шести этапов (понимание бизнеса, понимание данных, подготовка данных, моделирование, оценка и внедрение), которая естественным образом описывает жизненный цикл науки о данных.



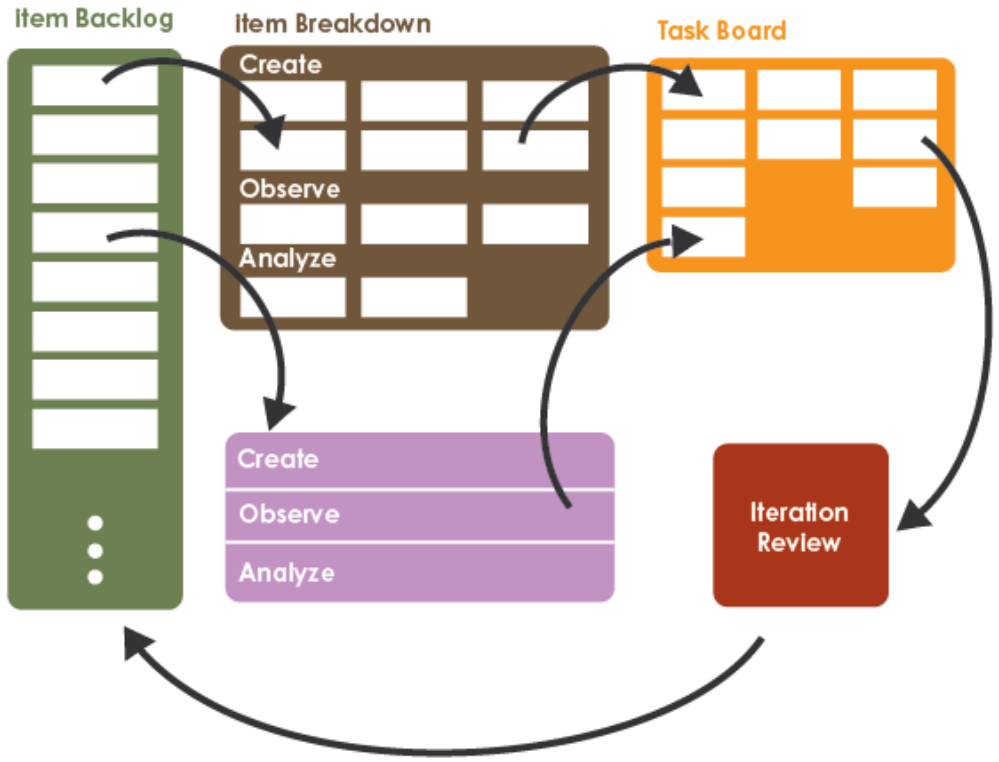
Scrum

Scrum - это наиболее часто используемая платформа для проектов разработки программного обеспечения и фактически гибкая платформа управления проектами. Она делит проект на серию мини-проектов, каждый из которых является последовательным и фиксированной продолжительности, называемых спринтами. Scrum также определяет встречи и роли, которые помогают руководить командой при выполнении проекта.



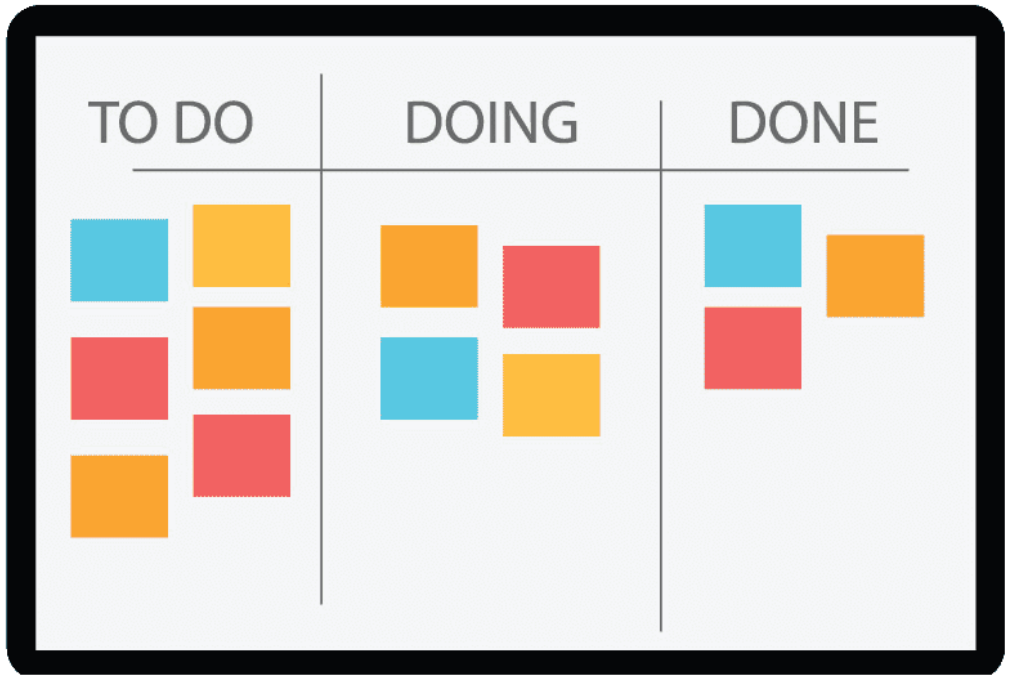
Data Driven Scrum

Data Driven Scrum (DDS) - это гибкая платформа, специально разработанная для команд, работающих в области науки о данных. Короче говоря, DDS направлена на улучшение взаимодействия и коммуникации в команде, работающей в области науки о данных. Джефф Зальц и Алекс Сазерленд создали Data Driven Scrum, чтобы обратить внимание на тот факт, что другие хорошо известные гибкие подходы (такие как Scrum и Kanban) часто не соответствуют уникальным потребностям проектов, связанных с наукой о данных.



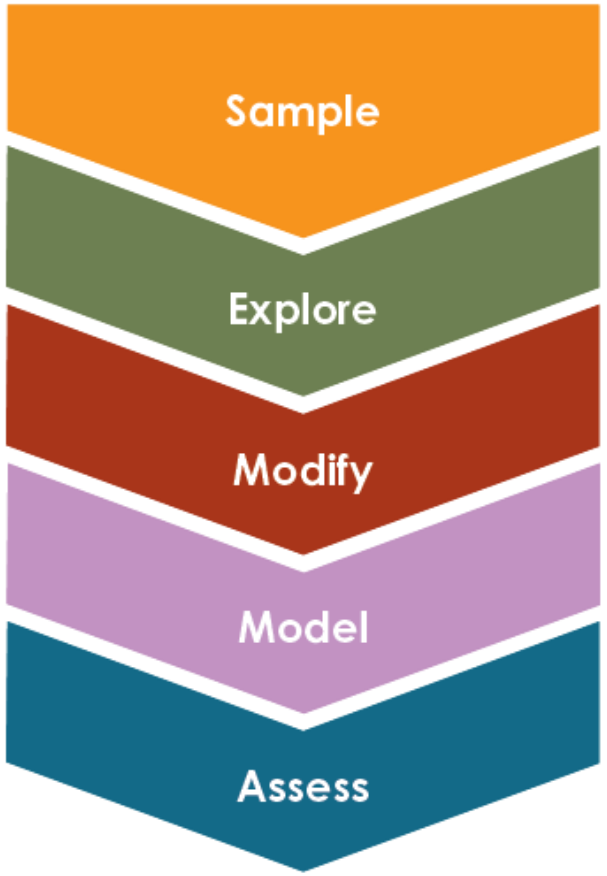
Kanban

Два ключевых принципа Kanban заключаются в том, чтобы (1) визуализировать процесс и (2) минимизировать количество незавершенных работ. Короче говоря, ограничивая количество задач, которые выполняются одновременно, Kanban обеспечивает гибкость.



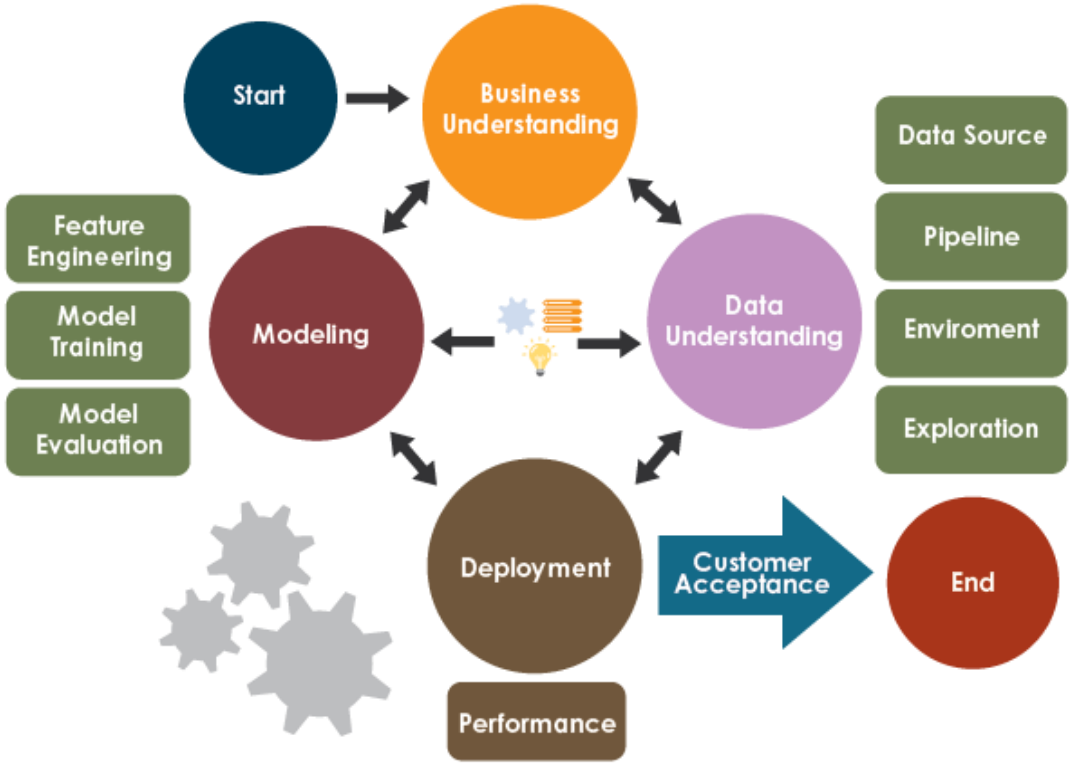
SEMMA

Программа SEMMA, разработанная компанией SAS, определяет 5 этапов проекта (отбор образцов, исследование, модификация, моделирование и оценка). Несмотря на то, что она предназначена для того, чтобы помочь пользователям в использовании инструментов SAS Enterprise Miner для решения задач интеллектуального



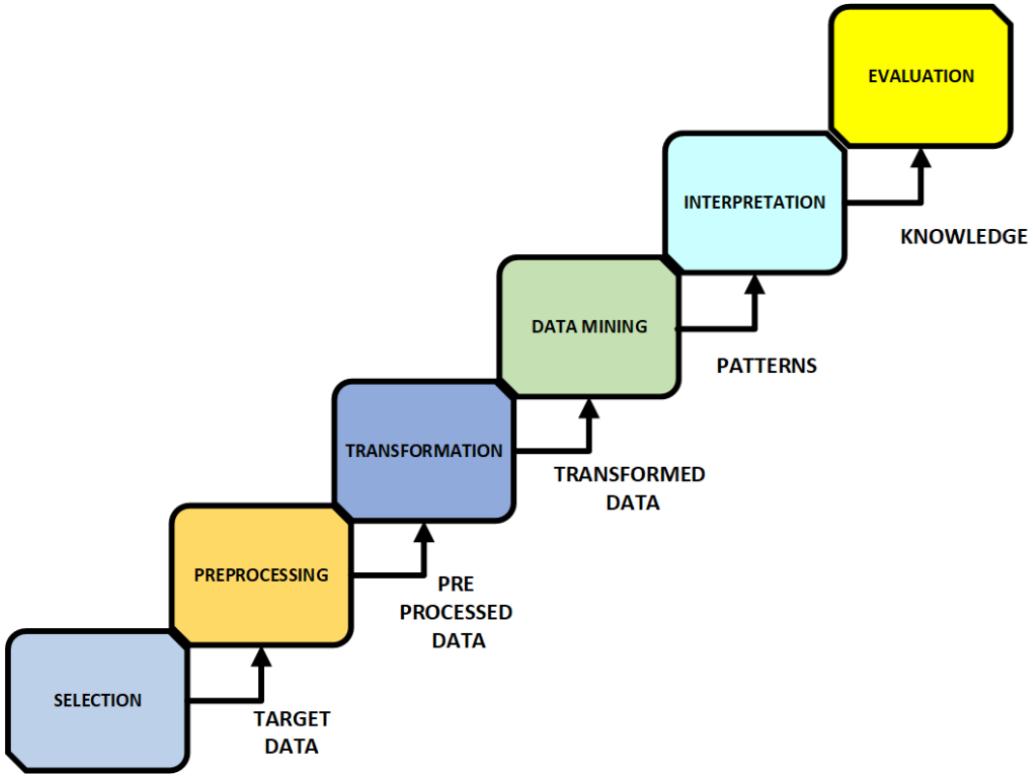
TDSP

TDSP (Team Data Science Process) - запущенный Microsoft в 2016 году, TDSP определяет 5 этапов жизненного цикла data science (понимание бизнеса, сбор и обработка данных, моделирование, развертывание и принятие заказчиком).



KDD

Обнаружение знаний в базе данных (Knowledge Discovery in Database, KDD) - это общий процесс обнаружения знаний в данных посредством интеллектуального анализа данных или извлечения шаблонов и информации из больших наборов данных с использованием машинного обучения, статистики и систем баз данных.



Жизненный цикл DOMINO

Domino Data Lab , поставщик из Кремниевой долины, предоставляющий платформу для науки о данных, разработал структуру жизненного цикла своего проекта по науке о данных в техническом документе 2017 года. Жизненный цикл науки о данных Domino основан на трех руководящих принципах:

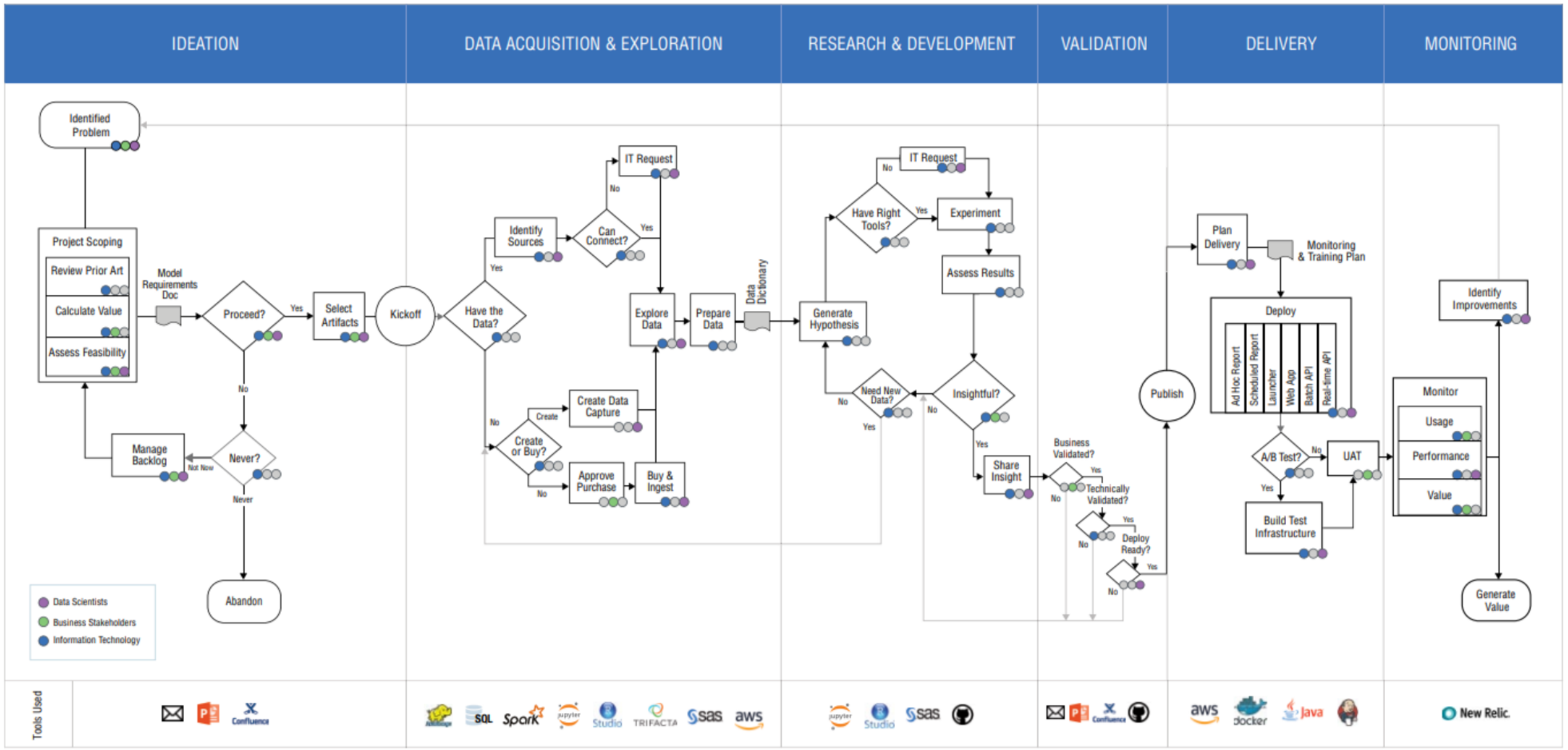
«Ожидайте и принимайте итерации», но «не допускайте, чтобы итерации существенно задерживали проекты или отвлекали их от поставленной цели»;

«Возможность комплексного сотрудничества» путем создания компонентов, которые можно повторно использовать в других проектах.;

«Ожидайте потребности в аудите» и «сохраняйте все соответствующие артефакты, связанные с разработкой и развертыванием модели».

Основной жизненный цикл делит проект на шесть итеративных этапов, которые отражают этапы CRISP-DM.

Жизненный цикл DOMINO



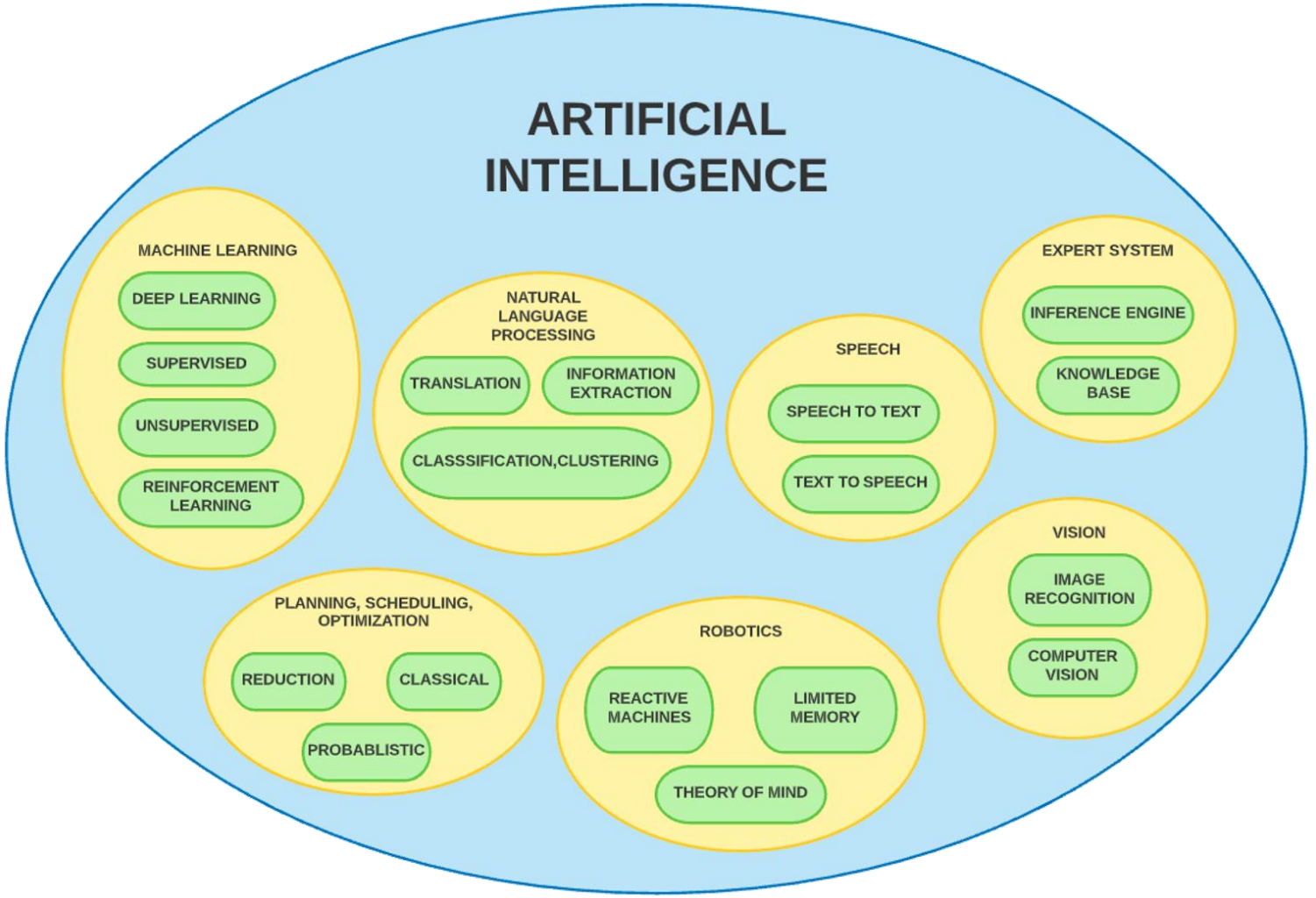
Искусственный интеллект

Иску́сственный интелле́кт (*Artificial intelligence (AI)*)— раздел информатики, изучающий возможность обеспечения разумных рассуждений и действий с помощью вычислительных систем и иных искусственных устройств.

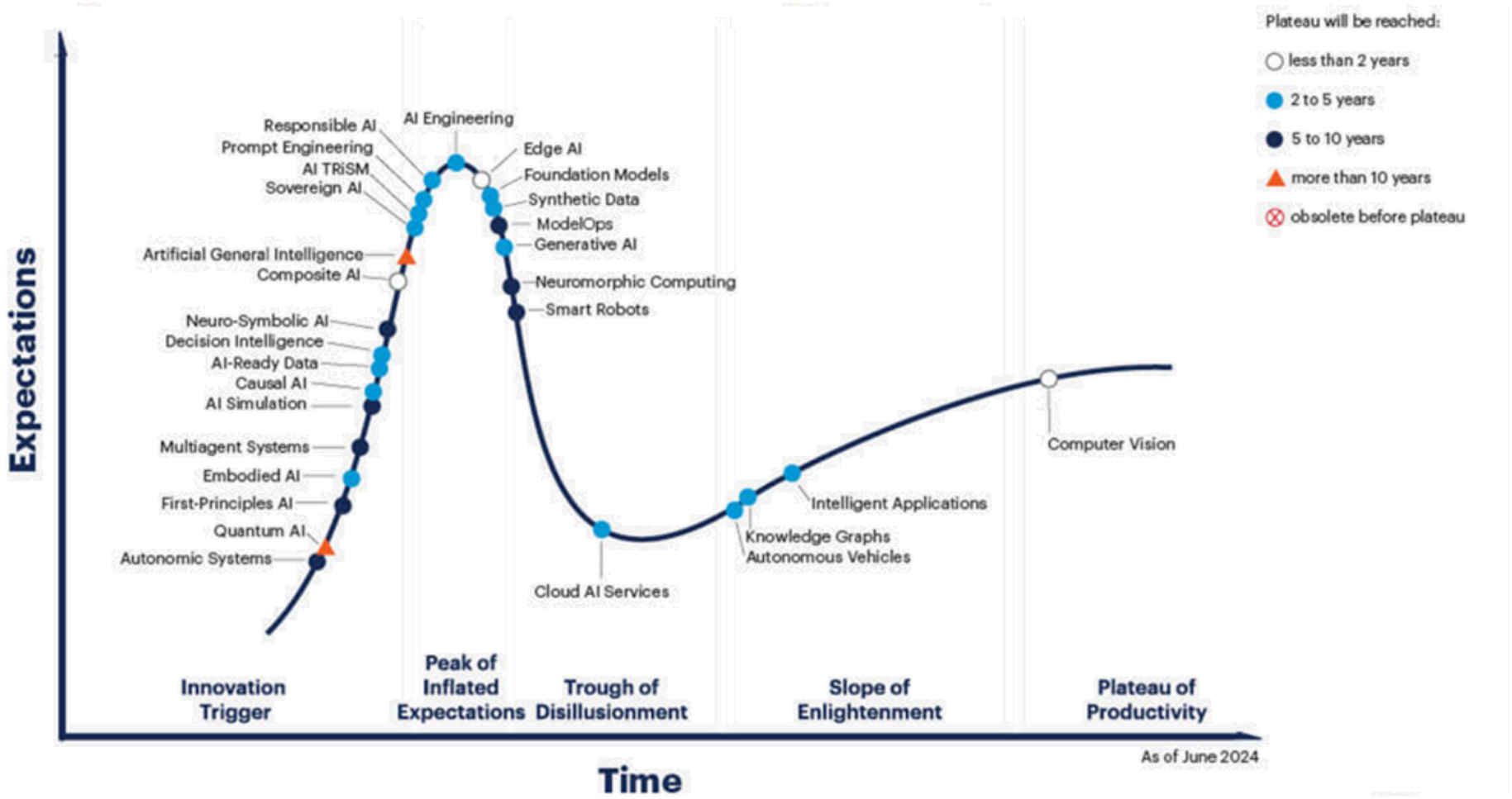
Искусственный интеллект общего назначения (AGI) - это форма ИИ, которая обладает способностью понимать, усваивать и применять знания в широком спектре задач и областей. Он может быть применен к гораздо более широкому набору вариантов использования и включает в себя когнитивную гибкость, адаптивность и общие навыки решения проблем.

Воплощенный ИИ (Embodied AI)- это интеграция машинного обучения, компьютерного зрения, технологий обучения роботов и языковых технологий, кульминацией которой является “воплощение” искусственного интеллекта: роботы, способные воспринимать, действовать и сотрудничать.

Искусственный интеллект

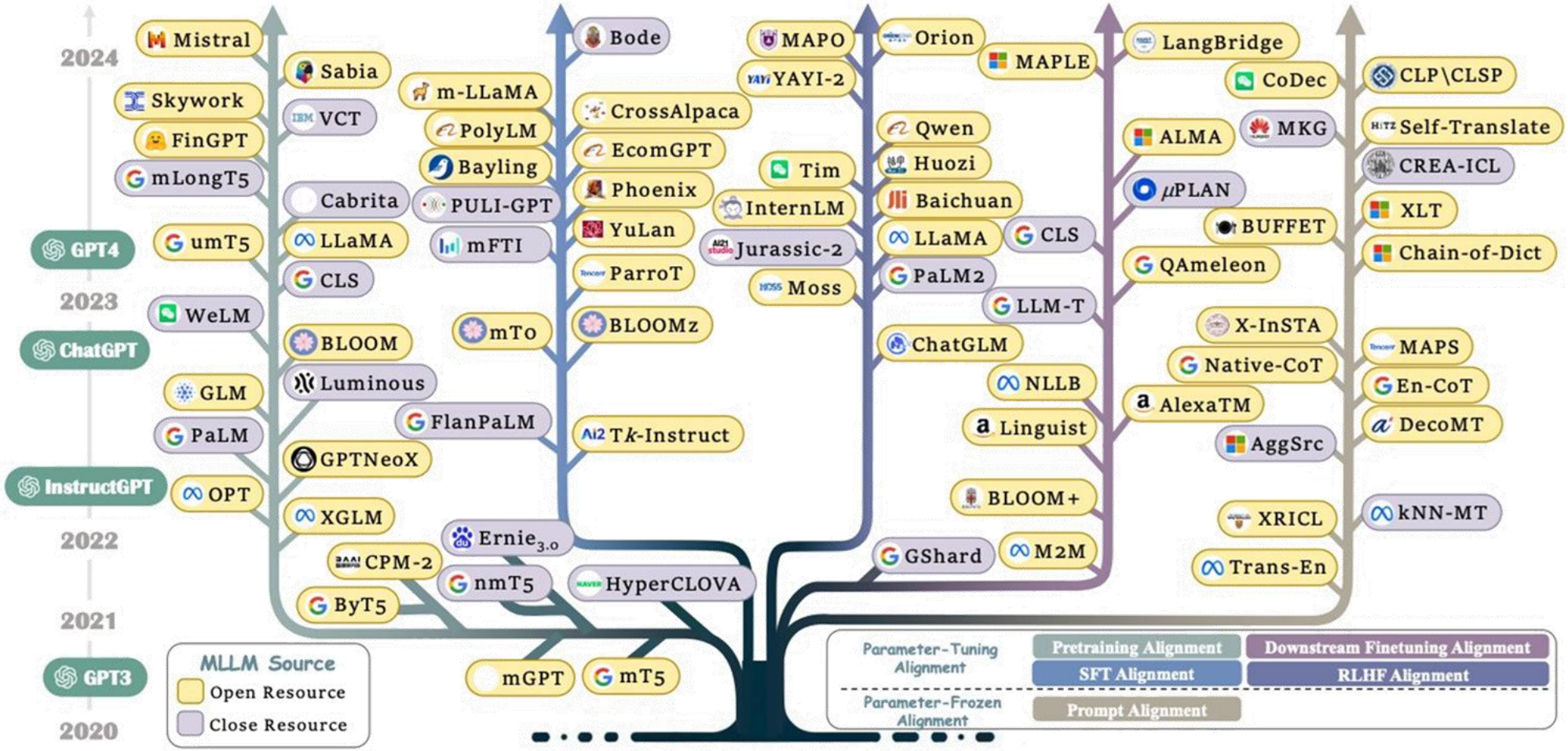


ИИ. Цикл хайпа по Gartner

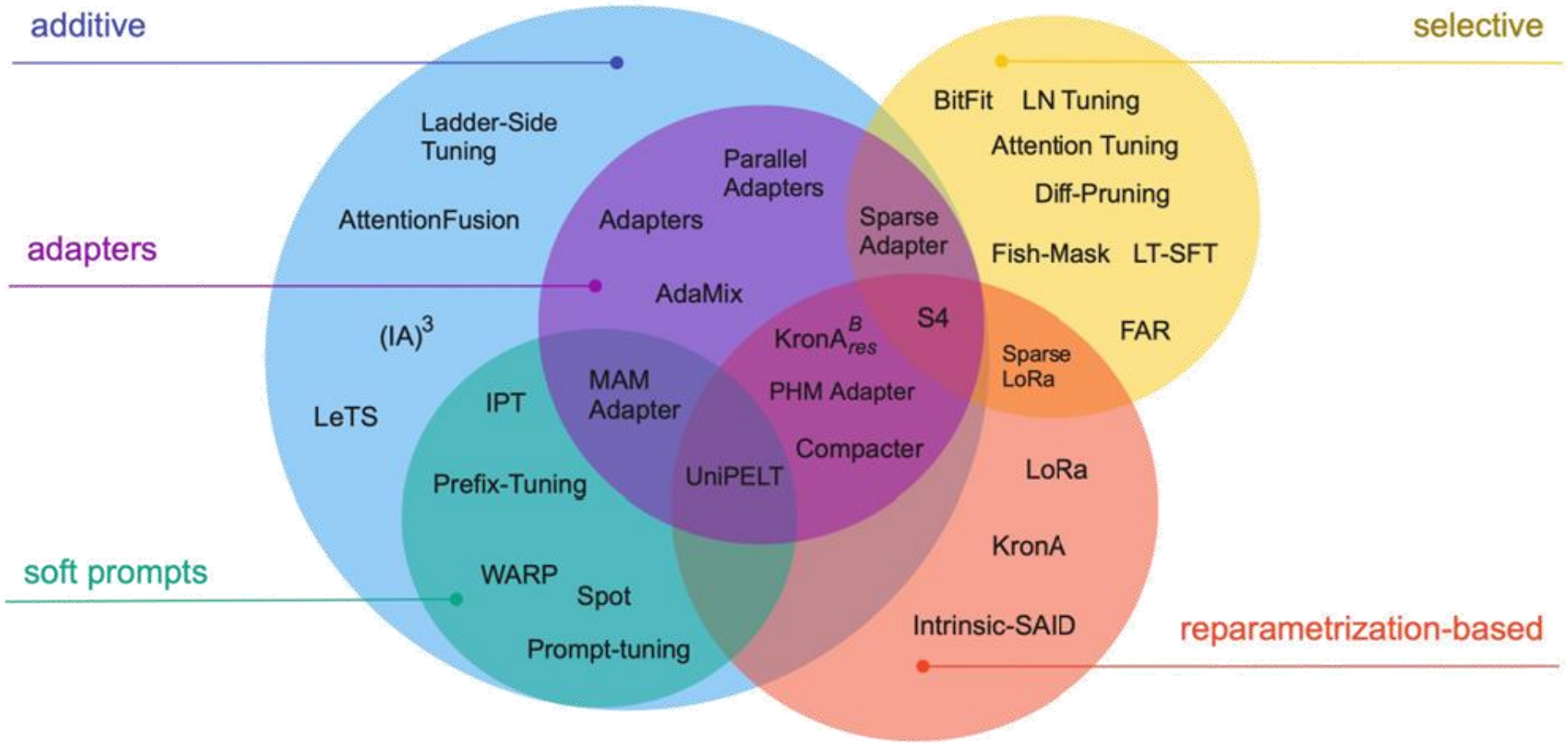


Машинное обучение

Современные LLM



Современные LLM. Методы тюнинга



Машинное обучение

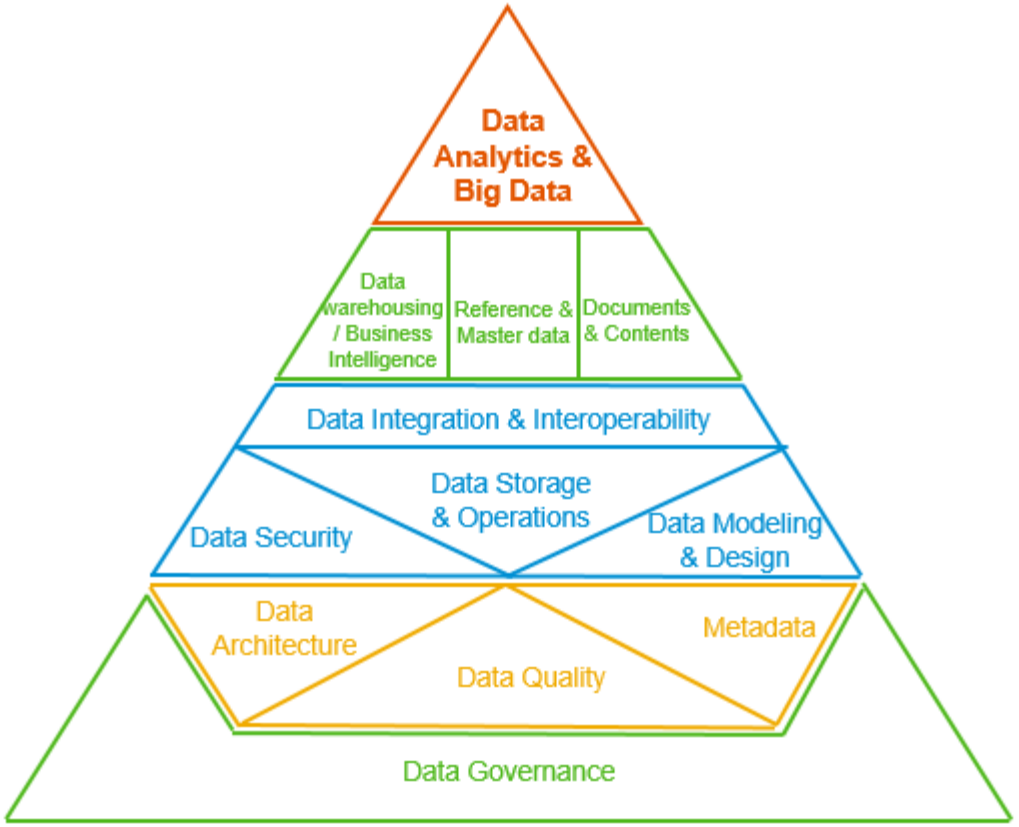
Data Governance

Управление данными - это принципиальный подход к управлению данными на протяжении всего их жизненного цикла, от приобретения до использования и утилизации. Управление данными необходимо каждой организации. По мере того, как предприятия во всех отраслях переходят к цифровой трансформации, данные быстро становятся самым ценным активом, которым они обладают.

Например, руководителям высшего звена нужны точные и своевременные данные для принятия стратегических бизнес-решений. Специалистам по маркетингу и продажам нужны достоверные данные, чтобы понимать, чего хотят клиенты. Специалистам по закупкам и управлению цепочками поставок необходимы точные данные для поддержания товарно-материальных запасов и минимизации производственных затрат. Специалисты по соблюдению требований должны доказать, что обработка данных осуществляется в соответствии как с внутренними, так и с внешними требованиями.

DATA-DMBOK

Международный свод знаний по управлению данными DAMA (DAMA-DMBOK) представляет собой комплексную платформу для управления данными. Он был создан DAMA International, некоммерческой организацией, занимающейся продвижением концепций и практик управления данными.

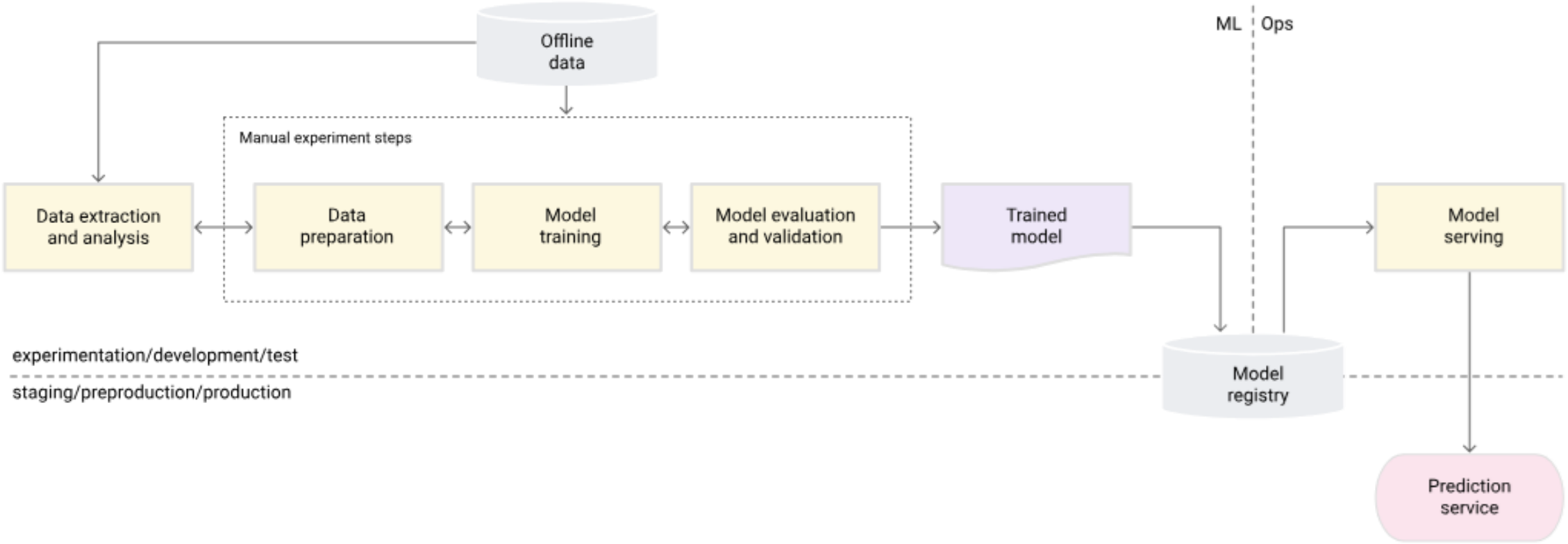


MLOPS

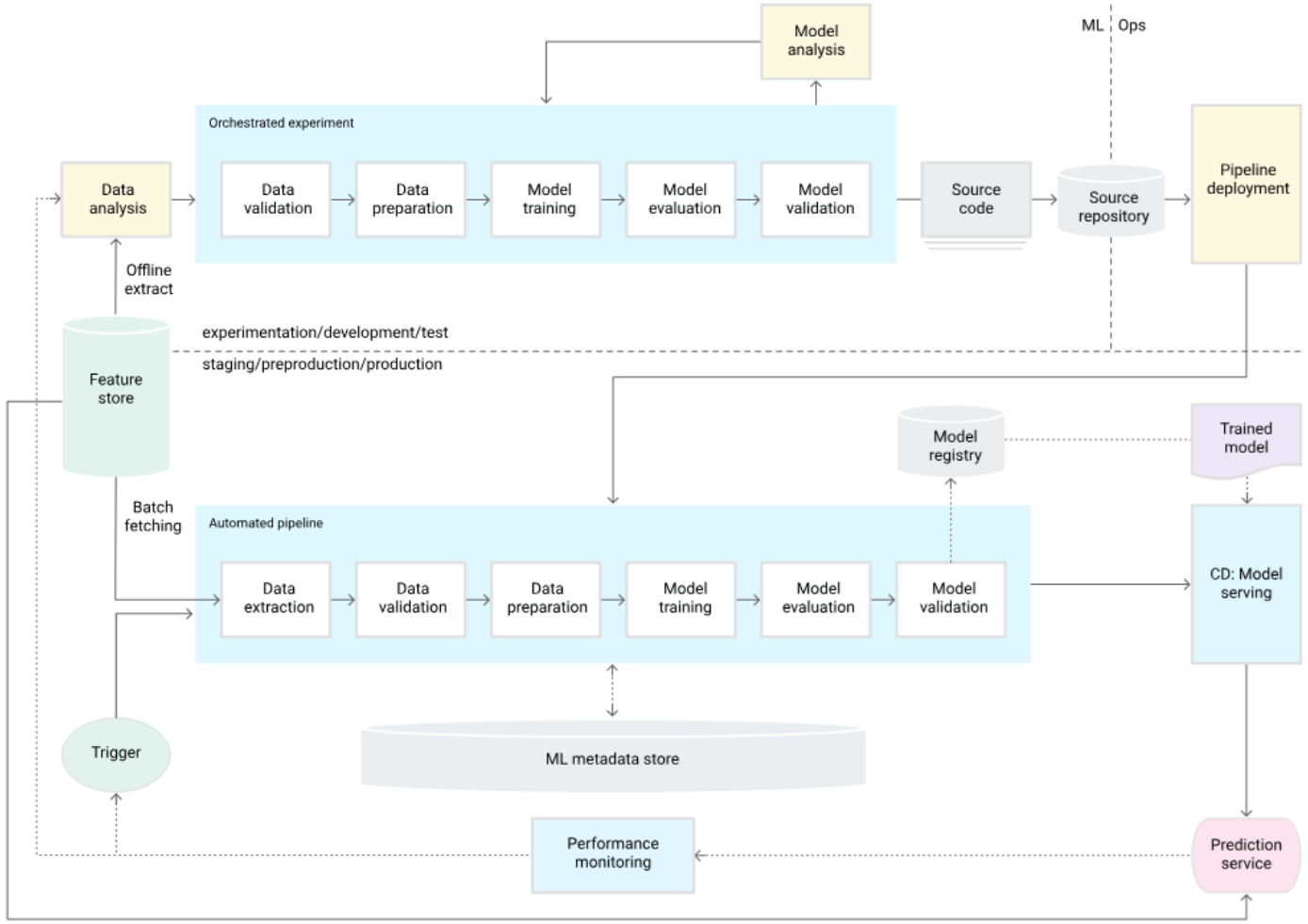
MLOps - это культура и практика ML-инжиниринга, целью которых является объединение разработки ML-систем (Dev) и эксплуатации ML-систем (Ops). Практика MLOps означает, что вы выступаете за автоматизацию и мониторинг на всех этапах создания ML-систем, включая интеграцию, тестирование, выпуск, развертывание и управление инфраструктурой.

Специалисты по обработке данных и внедрению и обучению ML-модели с прогнозируемой производительностью на основе автономного набора данных, предоставляющего соответствующие обучающие данные для их варианта использования. Однако реальная проблема заключается не в создании ML-модели, а в создании интегрированной ML-системы и ее непрерывной эксплуатации в производственной среде.

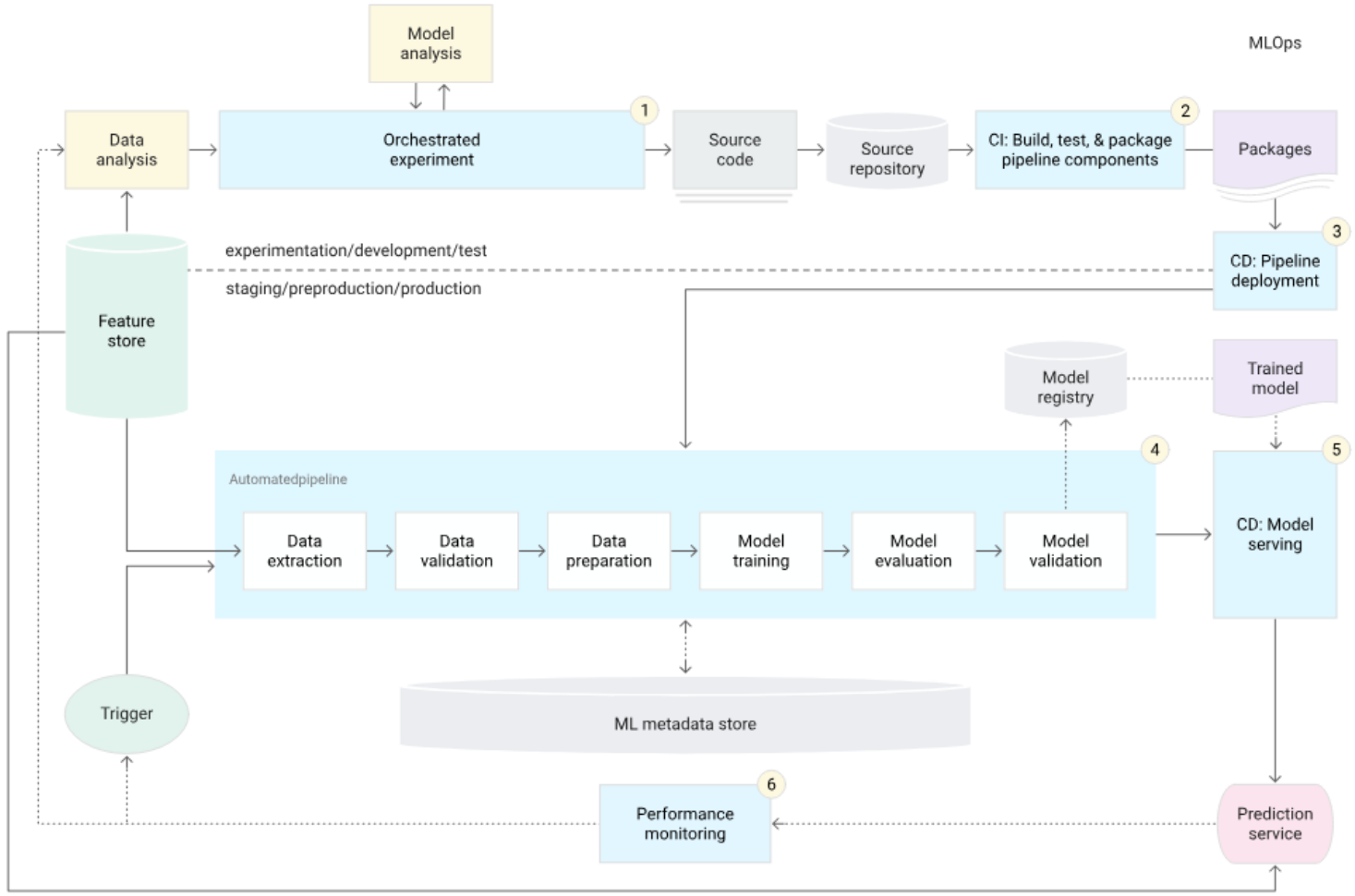
MLOPS (нулевой уровень зрелости)



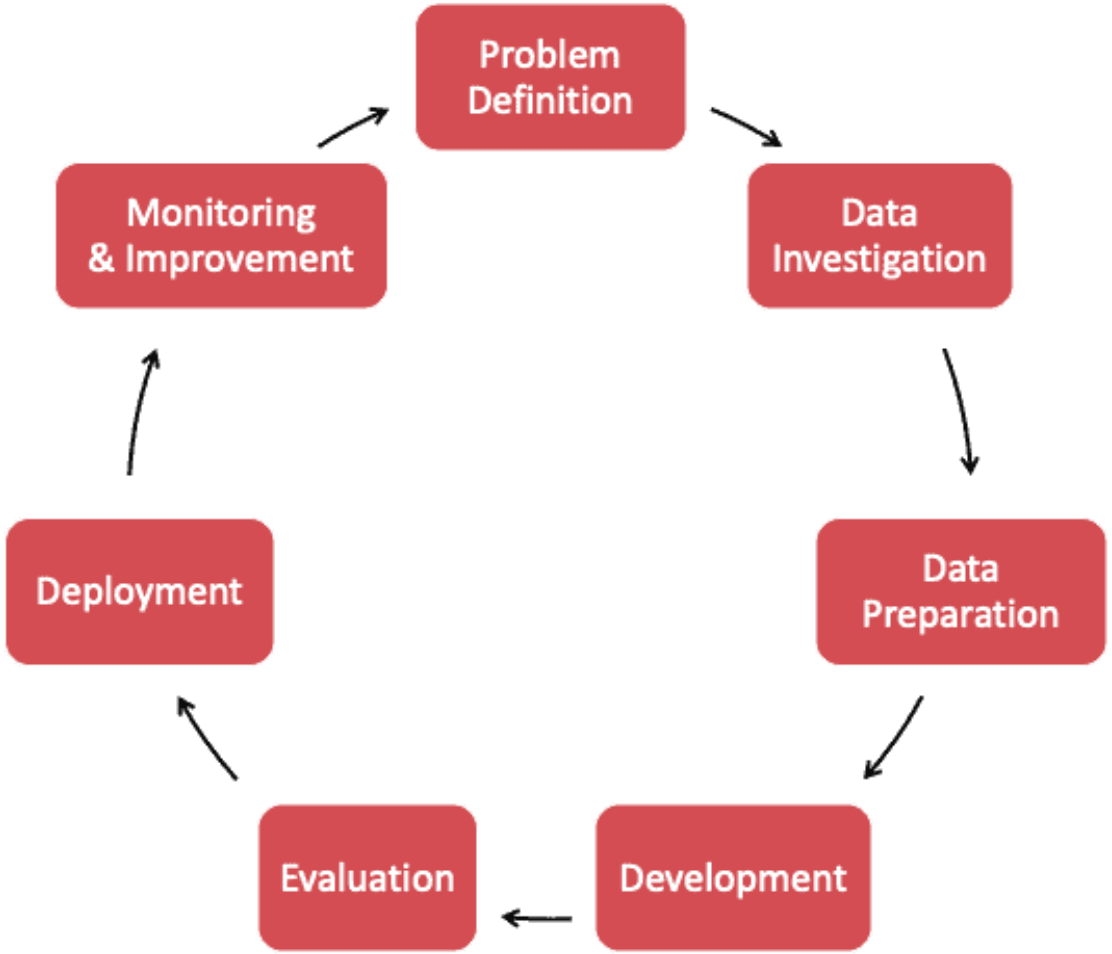
MLOPS (первый уровень зрелости)



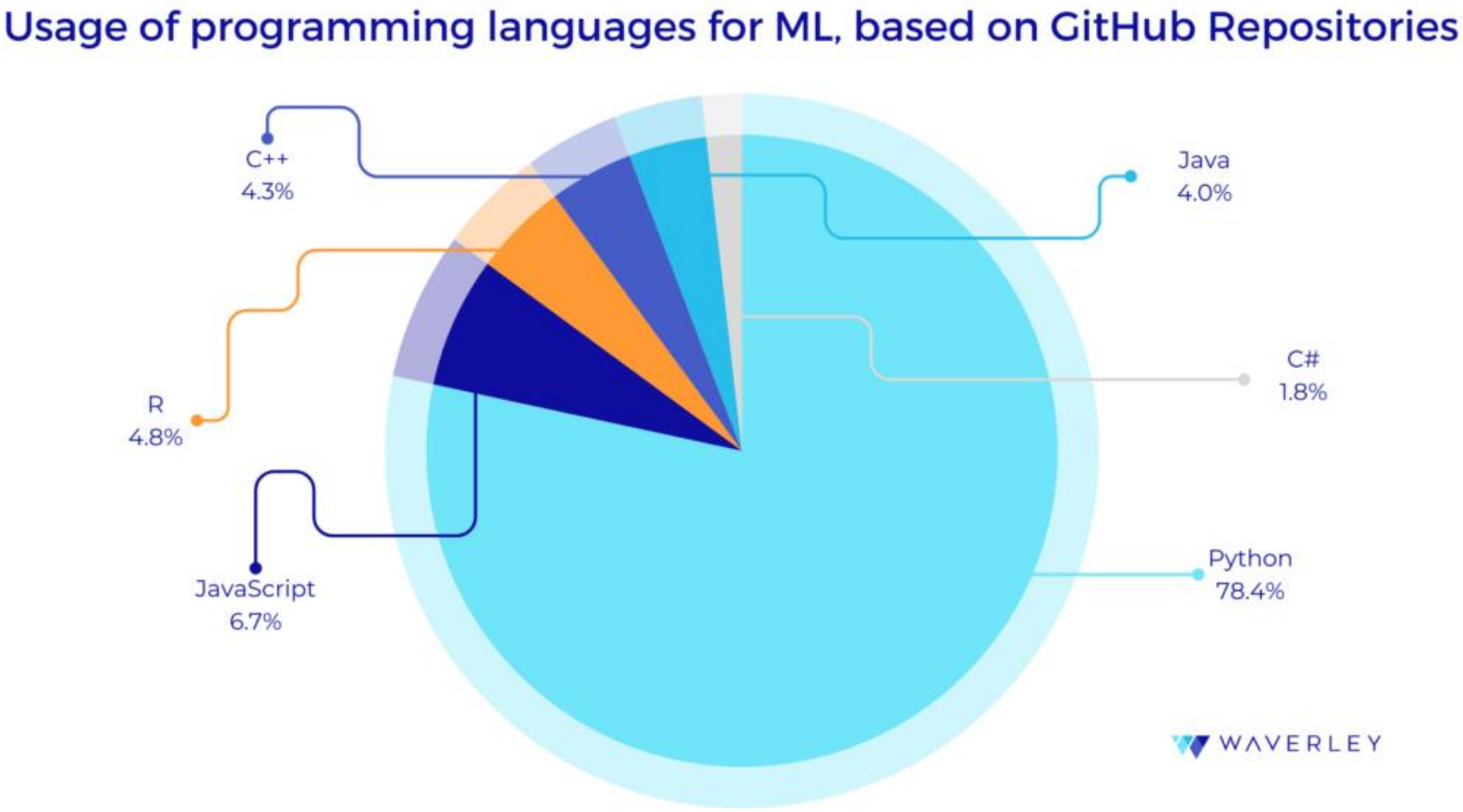
MLOPS (второй уровень зрелости)



Общий процесс DS

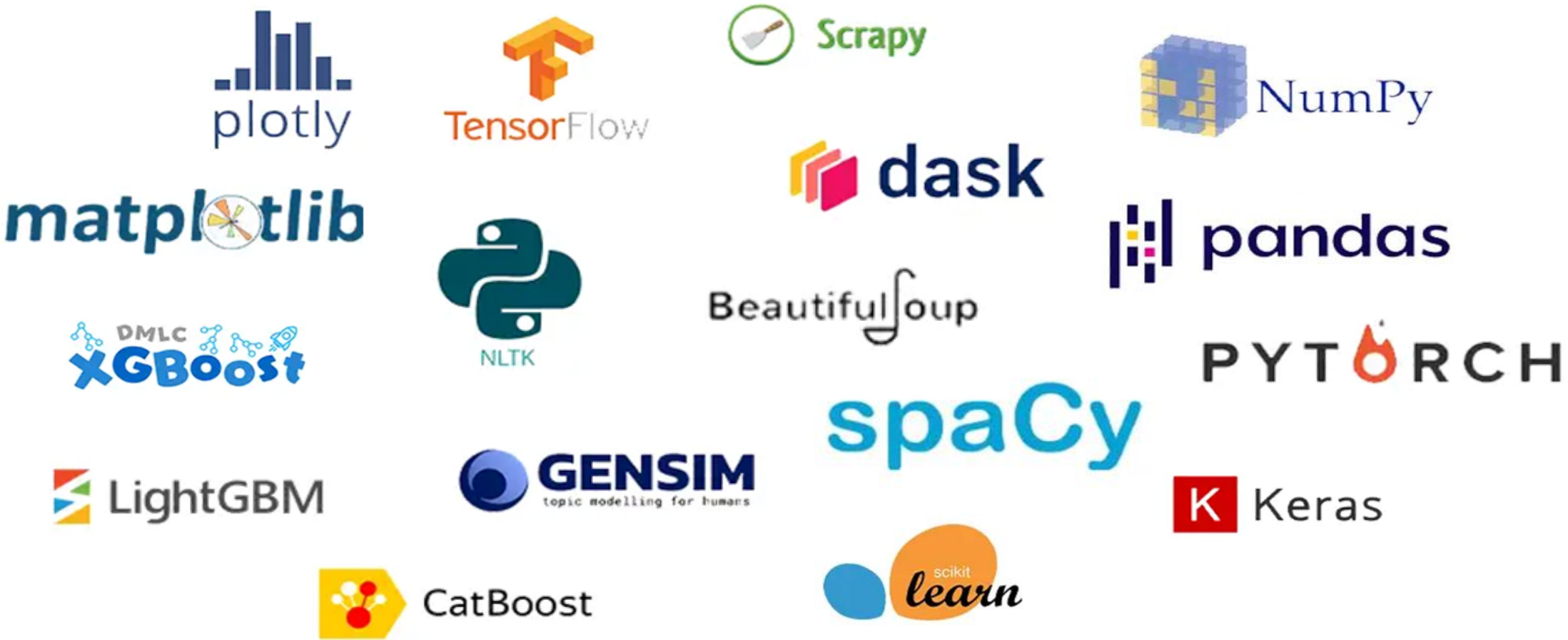


Использование языков программирования в DS/ML/AI



https://waverleysoftware.com/blog/python-for-ai-and-ml/

Библиотеки Python



Спасибо за внимание!

Спасибо за внимание!