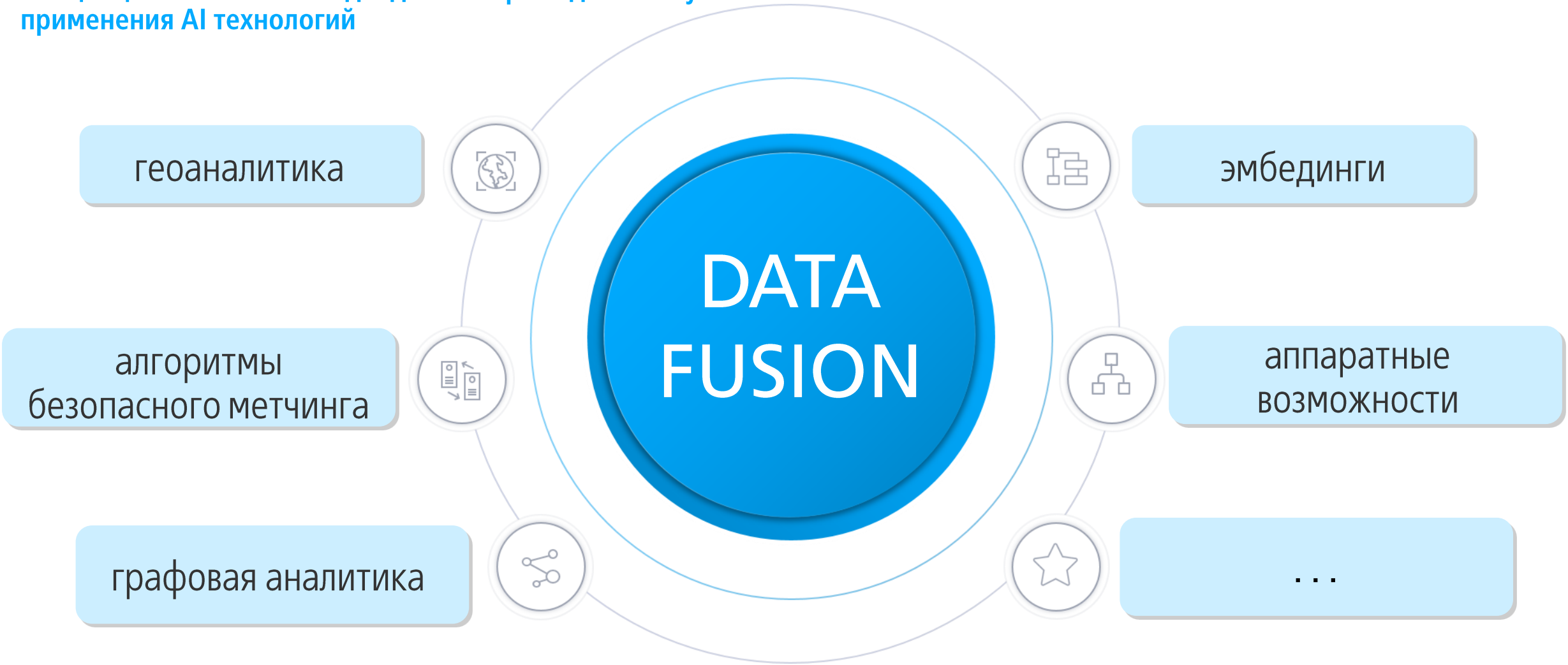
Системный подход к оценке рисков. Переход от классического ML к подвинутому на примере опыта ВТБ

Сергей Голицын

вице-президент, заместитель руководителя департамента анализа данных и моделирования, ВТБ

Пространство Data Fusion

Концепция системного подхода к синергии данных путем применения AI технологий



Системный подход к Data Fusion

Для Data Fusion можно использовать:

Обмен агрегатами данных

Нетривиальные схемы обезличивания данных: гео данные, графы и т.п.

Нейронные сети с целью формирования эмбедингов

Качественный и количественный рост DF задач требует системного подхода:

Сложных процедур обезличивания и подготовки данных

Компетенций в DL от участников обмена данными

Подход к росту DF задач должен быть универсален и операционно эффективен для любых типов данных задач

Можно ли системно подойти к решению Data Fusion задачи, двигаясь в сторону универсального решения?

Да, используя технологию крипто-анклавов

Этапы перехода на продвинутый ML

Выстраивание централизованной системы безопасной обработки данных с применением Auto ML

С ЧЕГО НАЧИНАЛИ

Разрозненные витрины по направлениям и продуктам, обогащение скорами

Классические линейные алгоритмы

Децентрализованная система работы с моделями

Децентрализованное применение моделей

СЕЙЧАС

Широкие объединенные централизованные витрины данных, обмен эмбедингами

Нелинейные алгоритмы – деревья, бустинги

MLops

Централизованная платформа исполнения моделей

В ГОРИЗОНТЕ 1-2 ГОДА

Безопасное объединение сырых данных в едином защищенном контуре

Нейросетевые модели

DLOps + DataOps

Auto ML + платформа полного жизненного цикла модели, включая применение и мониторинг

Графовая аналитика

Ключевая проблема при применении продвинутой графовой аналитики – контроль качества данных в рамках сложной нереляционной структуры используемой информации

Применение графов знаний позволяет

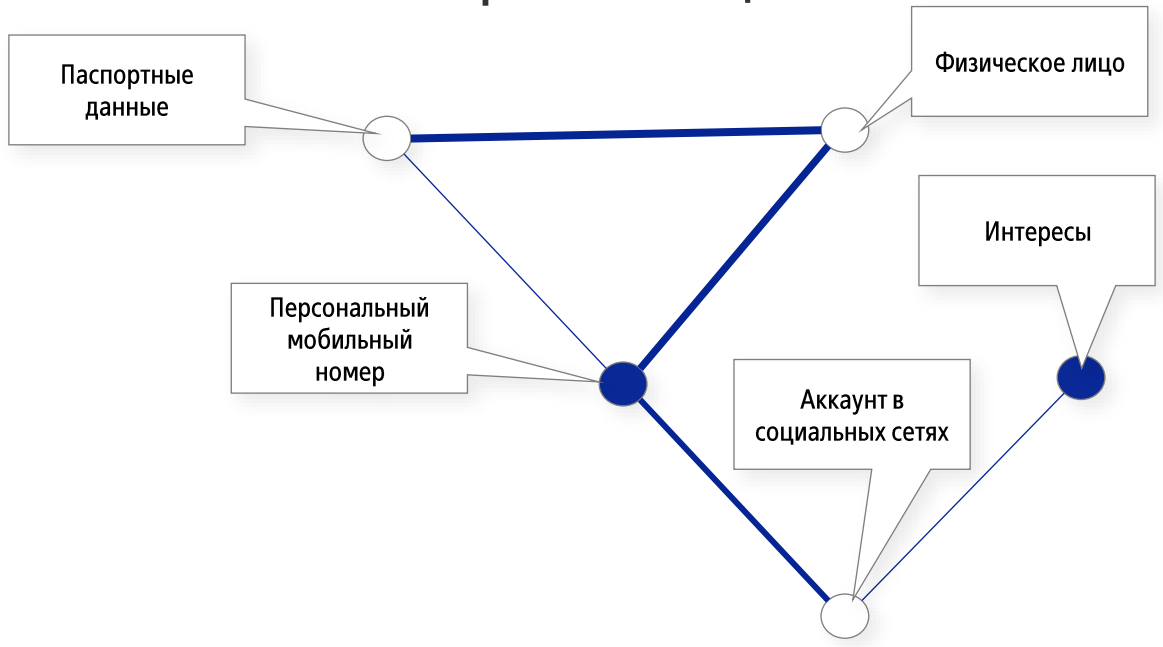
Использовать граф знаний как универсальный инструмент описания данных

Разрабатывать универсальные алгоритмы контроля качества новых данных на основании мета-данных и графа знаний

Верифицировать корректность бизнес-процессов, порождающих данные

Применять разработанные алгоритмы за периметром графовых задач, в том числе, при «слепой» обработки данных внутри крипто-анклавов

Граф знаний о предметной области «физическое лицо»



Эмбединги на графе интерконнект

Признаки рассчитываются по методологии обучения на графах знаний и моделировании языков: в начале на максимально большом объеме данных делается предобучение эмбеддингов (слов, понятий или отношений, в нашем случае – номеров телефонов), после чего признаки периодически обновляются. Также пайплайн позволяет добавлять новые номера в набор признаков с минимальной потерей качества.

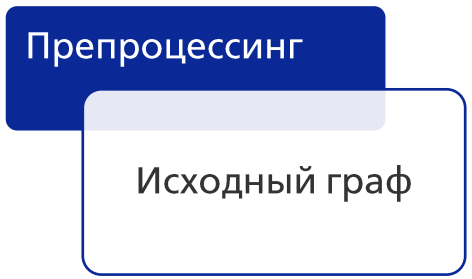
Результаты

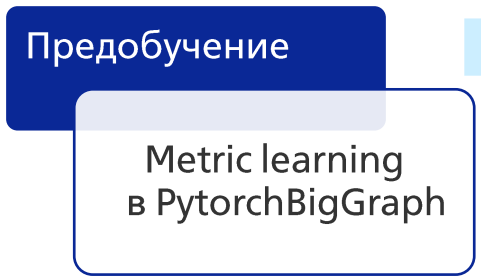
Кредитный риск: Uplift к текущей модели +3%%GINI

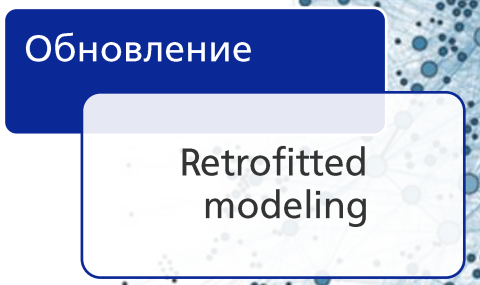
CRM: Uplift к текущей модели +4%%GINI

Для обучения используется metric learning: методология извлечения признаков на основании «топологии» входного датасета путем формирования метрического пространства.
В качестве функции потерь используется margin ranking loss, а пары составляются на основании существующих и несуществующих пар (negative sampling).

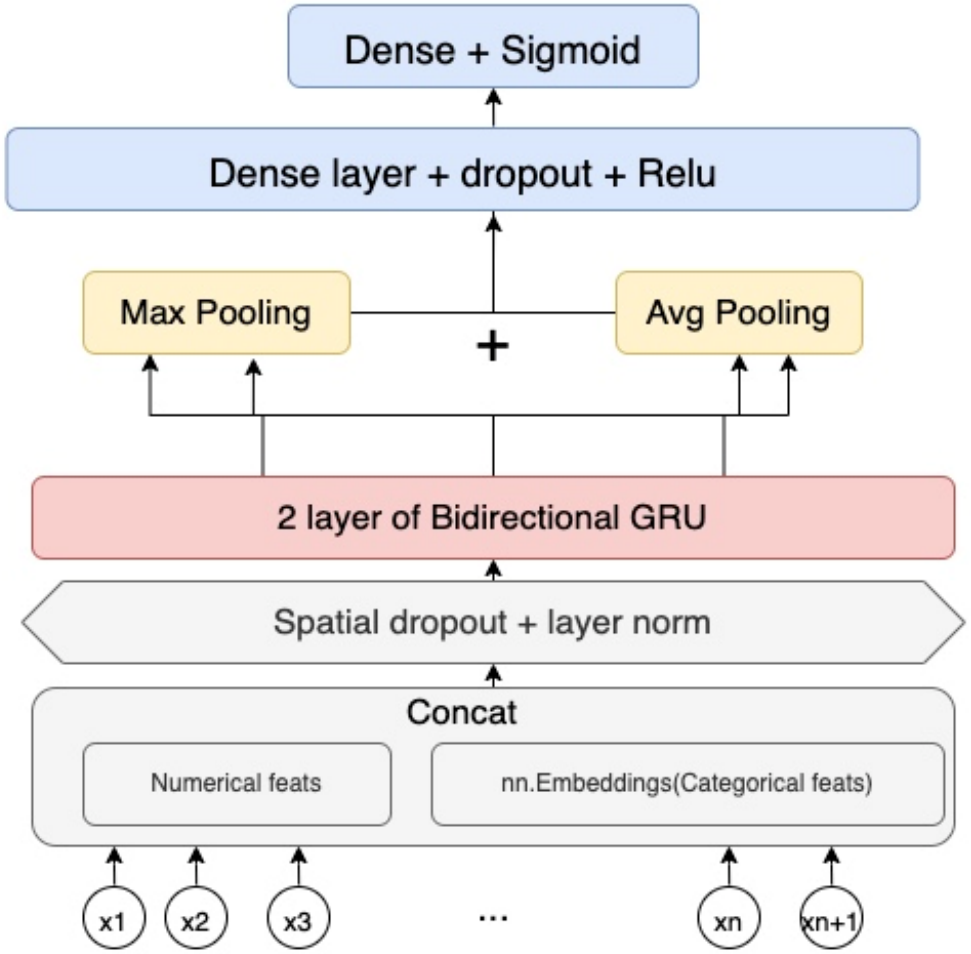
В качестве инструмента обучения используется библиотека PytorchBigGraph. Проверены различные алгоритмы извлечения признаков из графов: TRANSe, DistMult, Complex.
Алгоритм Complex подтвердил академический статус лучшего метода.

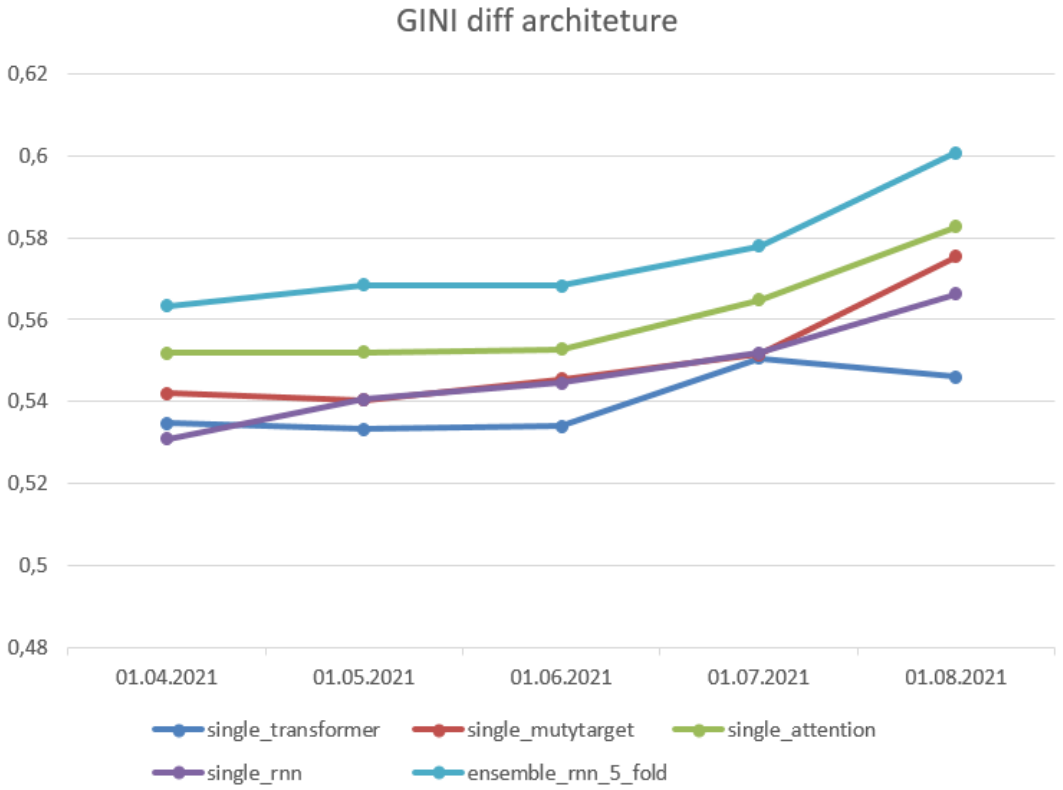






Нейросетевая модель на транзакциях для скоринга ФЛ





Совместно со Сколтехом разработана библиотека NAS по поиску оптимальных нейросетевых архитектур

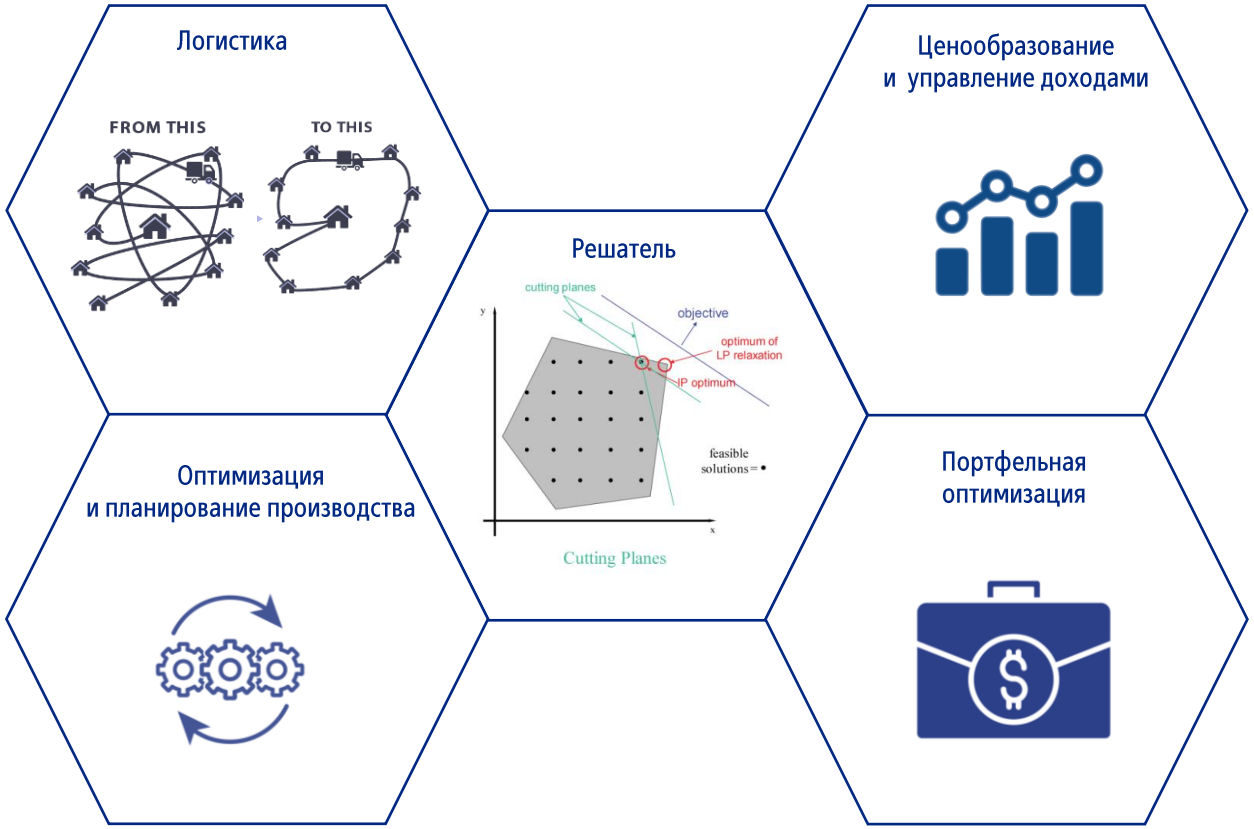
- https://arxiv.org/pdf/1911.02496.pdf - статья E.T.-RNN: Applying Deep Learning to Credit Loan Applications
- https://habr.com/ru/company/alfa/blog/551130/ - Нейросетевой подход к моделированию карточных транзакций

Оптимизатор. Применимость в индустрии

Финансовые рынки — важный элемент экономики страны. Алгоритмы оптимизации играют ключевую роль в задачах диверсификации.

Именно алгоритмы оптимизации позволяют не только существенно повышать эффективность стратегий, но и создавать индивидуальные стратегии с учетом капитала и риск-аппетита инвесторов.

Дальнейшее развитие оптимизационных алгоритмов и агентного моделирования могут позволить более эффективно распределять финансовые и материальные потоки в рамках экономики в целом.



Крипто-анклав

Партнеры передают данные в крипто-защищенную область, получая гарантии:

Отсутствия доступа у других участников анклава к их данным.

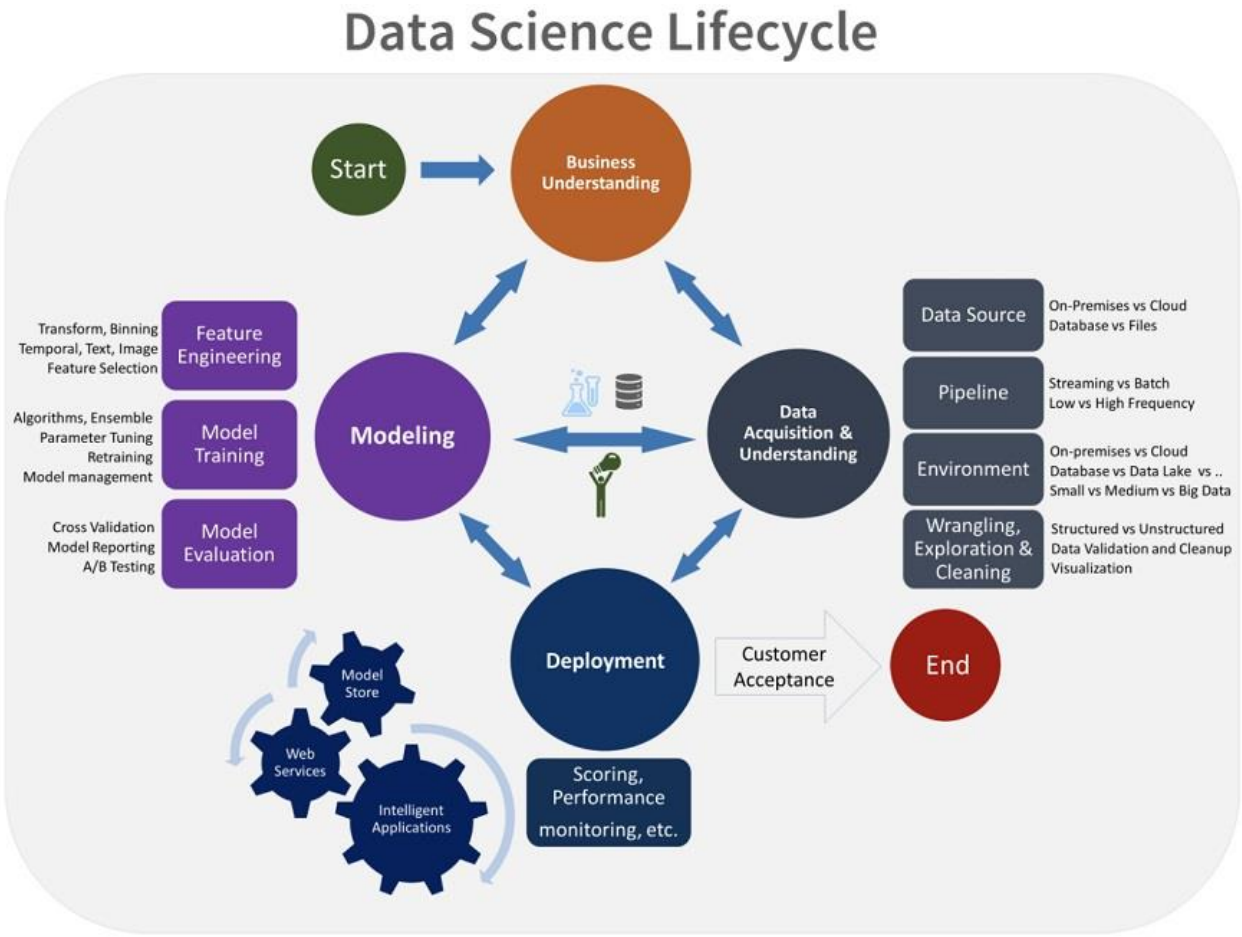
Применения **строго** регламентированных алгоритмов для анализа их данных.

Возможности **получения** «на выходе» из анклава **исключительно результатов** применения моделей на данных участников анклава **без доступа к исходной информации.**

AutoML подходы

Ключевой тренд: задача DS **создавать** не отдельные модели, а **фабрики** по разработке моделей под типовые задачи.

AutoML платформа обеспечивает не только разработку модели, но и **автоматизирует полный жизненный цикл модели**, включая внедрение и мониторинг модели как сервис.



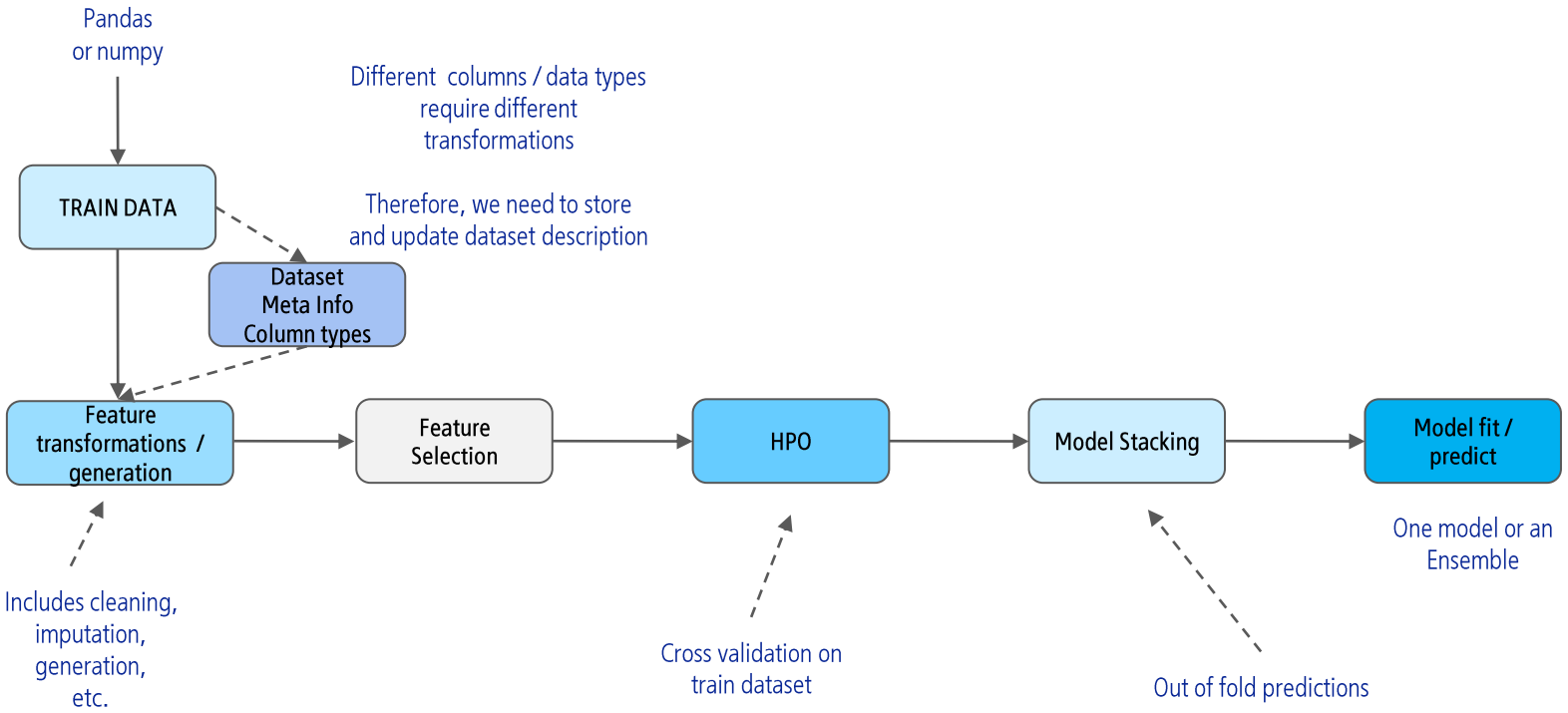
Библиотека Auto ML

Библиотека представляет собой конструктор, состоящий из разных блоков.

Новые блоки легко добавлять с помощью общих классов адаптеров и файла конфигурации.

Под каждую задачу можно быстро собрать свой уникальный пайплайн.

Блоки можно переиспользовать по отдельности, вне библиотеки.



Особенности и ограничения библиотеки:

- оптимальный подбор гиперпараметров;
- предобработка данных и фичегенарация;
- поиск выбросов, выявления функциональных зависимостей;
- работает исключительно с табличными данными

Принципы Data Fusion

Data Fusion – точка сближения науки и бизнеса

ВНУТРЕННИЙ R&D

R&D С НАУЧНЫМ СООБЩЕСТВОМ

**ПАРТНЕРСТВА –
КРОСС-ОТРАСЛЕВАЯ СИНЕРГИЯ**

DATA
FUSION

Спасибо за внимание