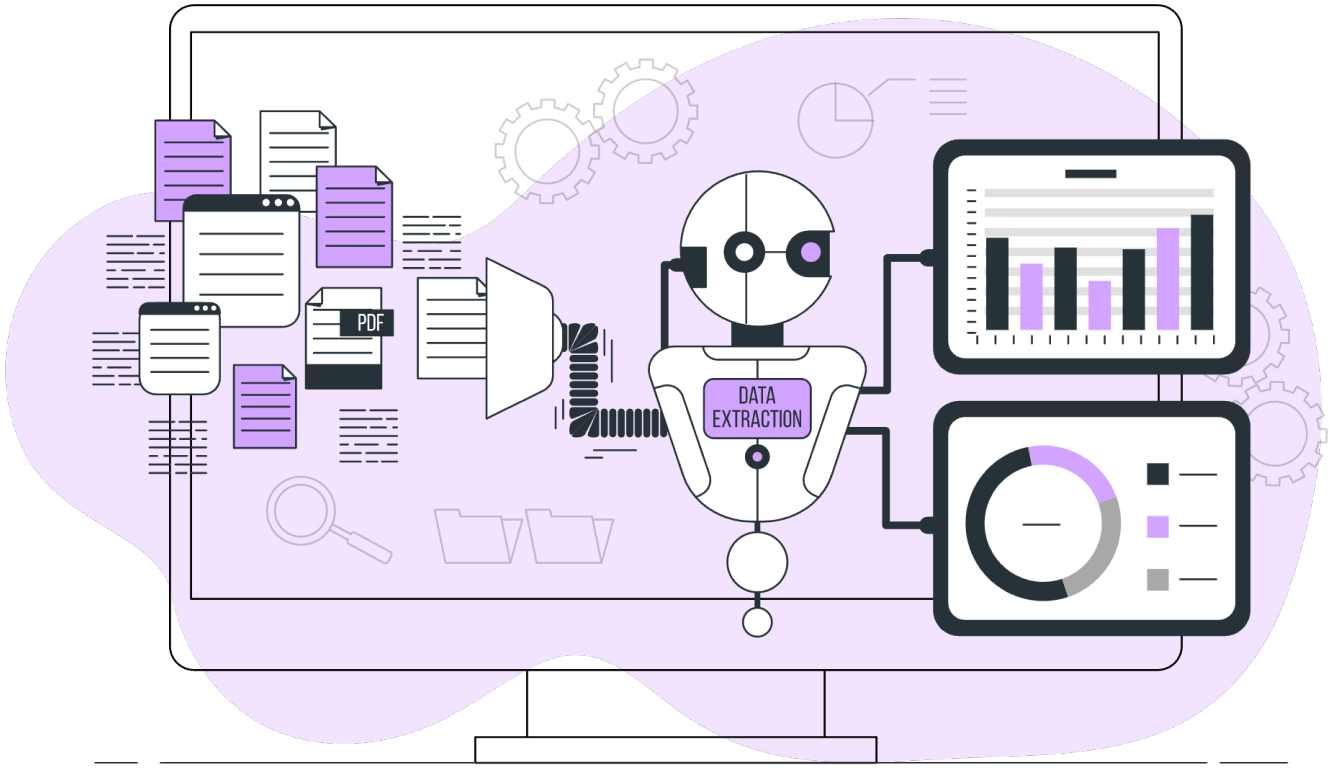
Управление данными и разработкой как основа для применения ИИ

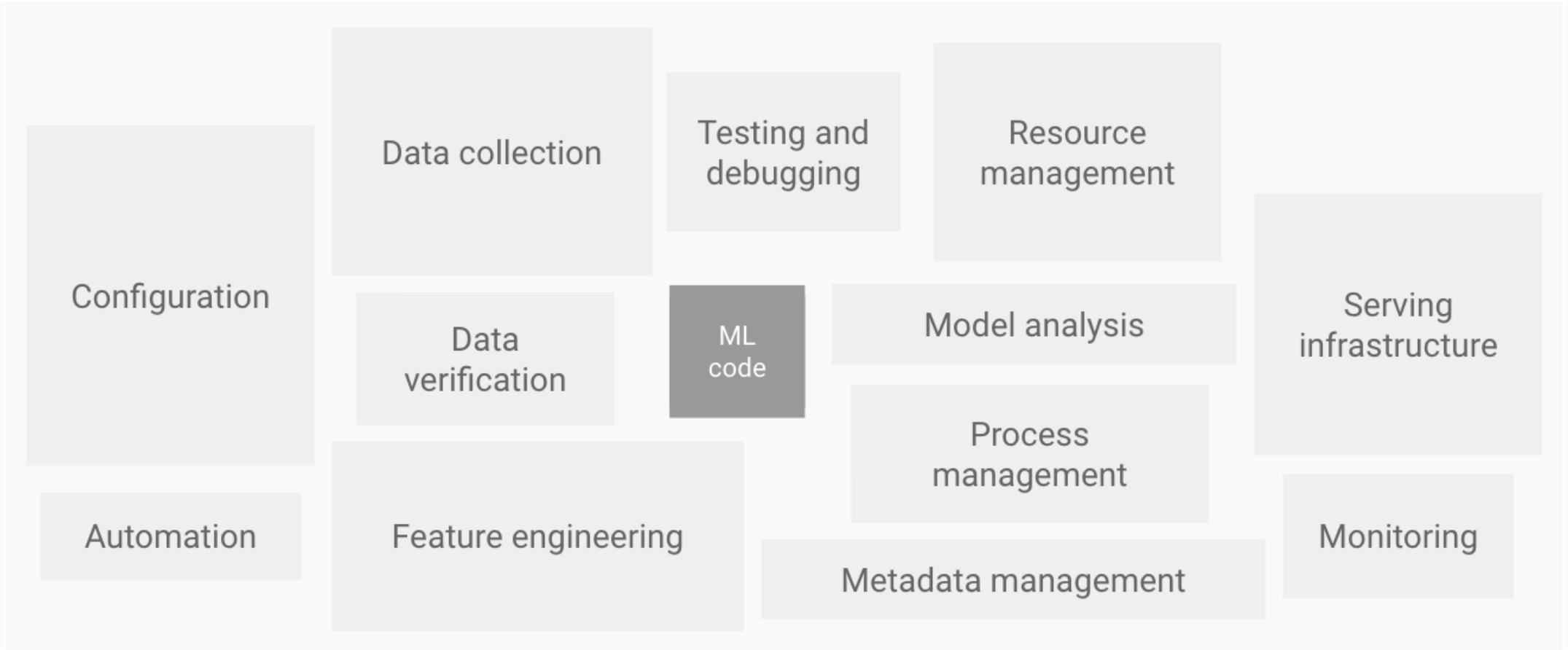
Загорулькин Дмитрий Эдуардович

Заместитель директора Центра стратегической аналитики и больших данных
dzagorulkin@hse.ru



Fall into ML
’24

Составные части ML системы



https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

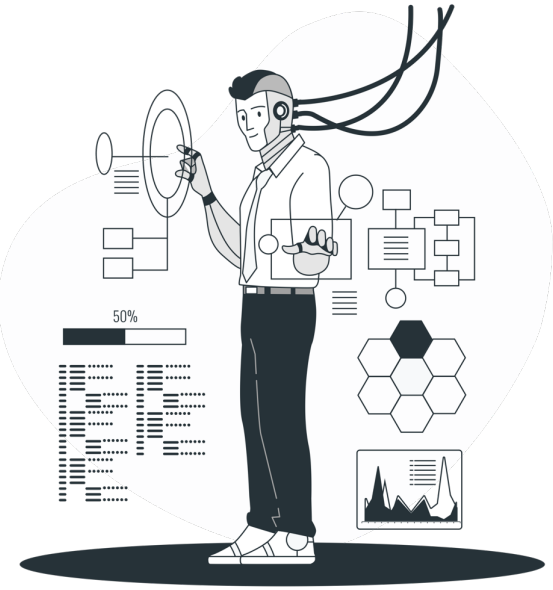
Создание AI проектов

01 Процессы

02 Инфраструктура

03 Данные

04 Моделирование



Понятная постановка задачи

Нам нужно сделать модель, которая определяет уровень технологии.

ПЛОХАЯ ПОСТАНОВКА ЗАДАЧИ

Мы собираем данные об упоминании различных технологий в СМИ. Данные собираются регулярно. Нам необходимо улучшить точность оценки минимум на 20%. Также нужно учитывать, что технологии могут относиться к разным отраслям и иметь разный жизненный цикл. Уровни нужно агрегировать с 10 до 3 классов.

ХОРОШАЯ ПОСТАНОВКА ЗАДАЧИ

Метрики — они для всех разные!

01

Бизнес сфокусирован на получении прибыли!

Бизнес оперирует понятными метриками, такими как процент оттока клиентов, время, проведенное в продукте, количество пользовательских входов в месяц и т.д.

02

Дата-сайентист оперирует модельными метриками (P/R, Accuracy, F1 и другие)

03

Для получения профита от внедрения ML решения важно связать модельные метрики и метрики бизнеса

Важно найти и протестировать это влияние как можно скорее

Высокие модельные метрики не гарантируют улучшение бизнес-метрик!

Управление AI проектом

SCRUM

Подходит под ИТ-проекты

CRISP-DM (Cross-Industry Standard Process for Data Mining)

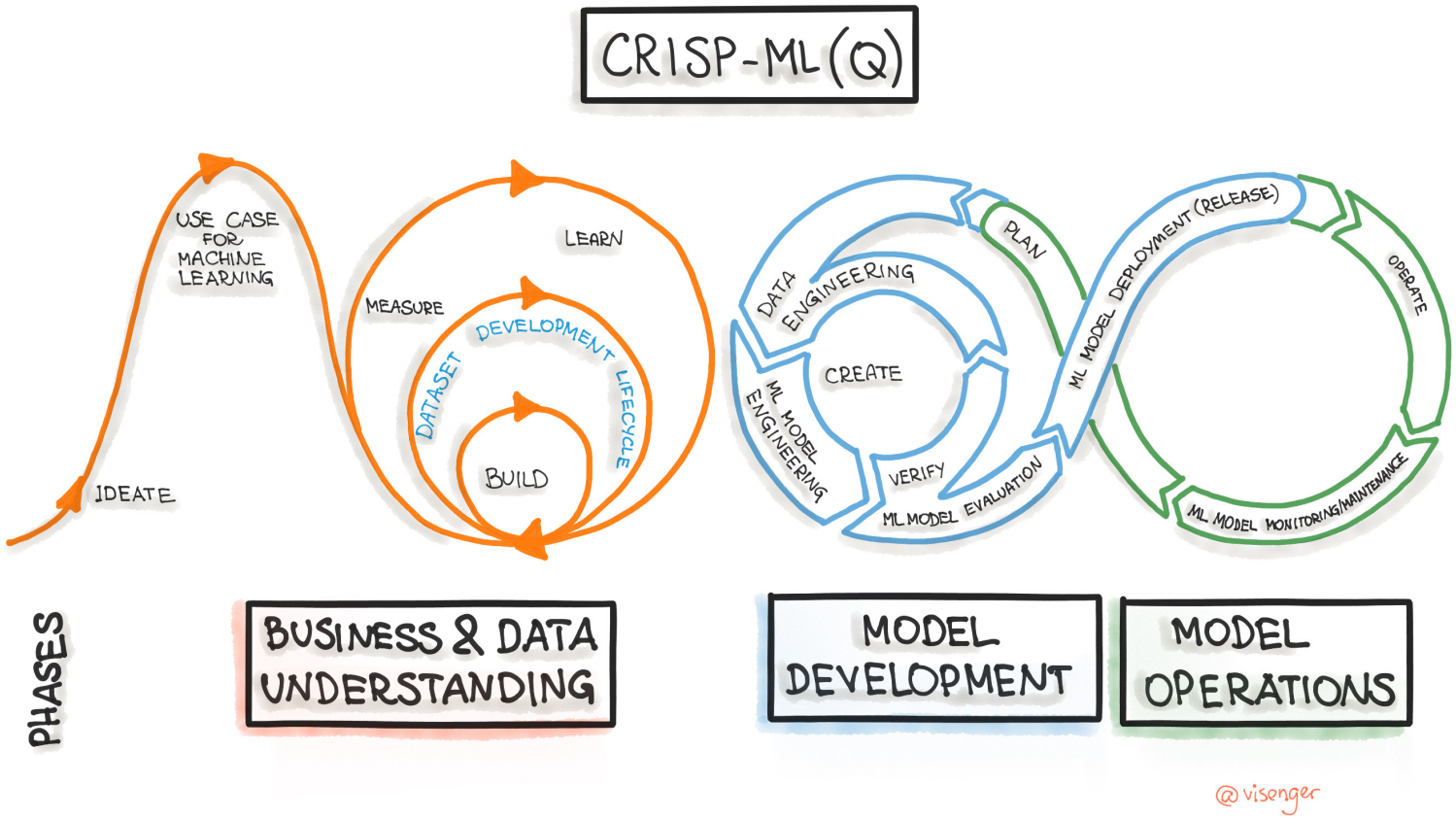
1. Понимание бизнес-целей (Business Understanding) – необходимо привлечение всех заинтересованных сторон
2. Понимание данных (Data Understanding) – проведение разведочного анализа и другие проверки данных
3. Подготовка данных (Data Preparation) – консолидация, агрегация
4. Моделирование (Modeling) – выбор методологии и построения модели
5. Оценка результата (Evaluation)
6. Внедрение модели/процесса (Deployment)

Проблемы: Не предназначена под итеративный процесс улучшения. Отсутствует QA.

CRISP-ML(Q)

Расширение методологии

CRISP-DM сравнение с CRISP-ML(Q)



CRISP-ML(Q)	CRISP-DM
Business & Data Understanding	Business Understanding
	Data Understanding
Data Preparation	Data Preparation
Modeling	Modeling
Evaluation	Evaluation
Deployment	Deployment
Monitoring & Maintenance	-

https://ml-ops.org/content/crisp-ml | https://arxiv.org/pdf/2003.05155

Данные

Основа построения систем машинного обучения

Существуют разные типы данных. Для одних моделей нужны сильно структурированные данные, для других — нет (картинки, видео, книги и таблицы в БД)

Отсутствие культуры работы с данными зачастую является тормозом внедрения систем машинного обучения в бизнес-процессы

Необходимо развивать культуру работы с данными в части автоматической проверки качества данных и других data governance подходов (ведение дата-каталогов и управление метаданными и др.)

Не все данные подходят для моделирования!

Управление данными

DWH (Data Warehouse)

Сложность внесения изменений • Долго строить • Нужно хорошо понимать данные • В основном для структурированных данных

DataLake

Единая точка получения данных • Нужно следить за качеством и метой • Нет ACID транзакций, эволюции схем

DataLakehouse = DataLake + DWH

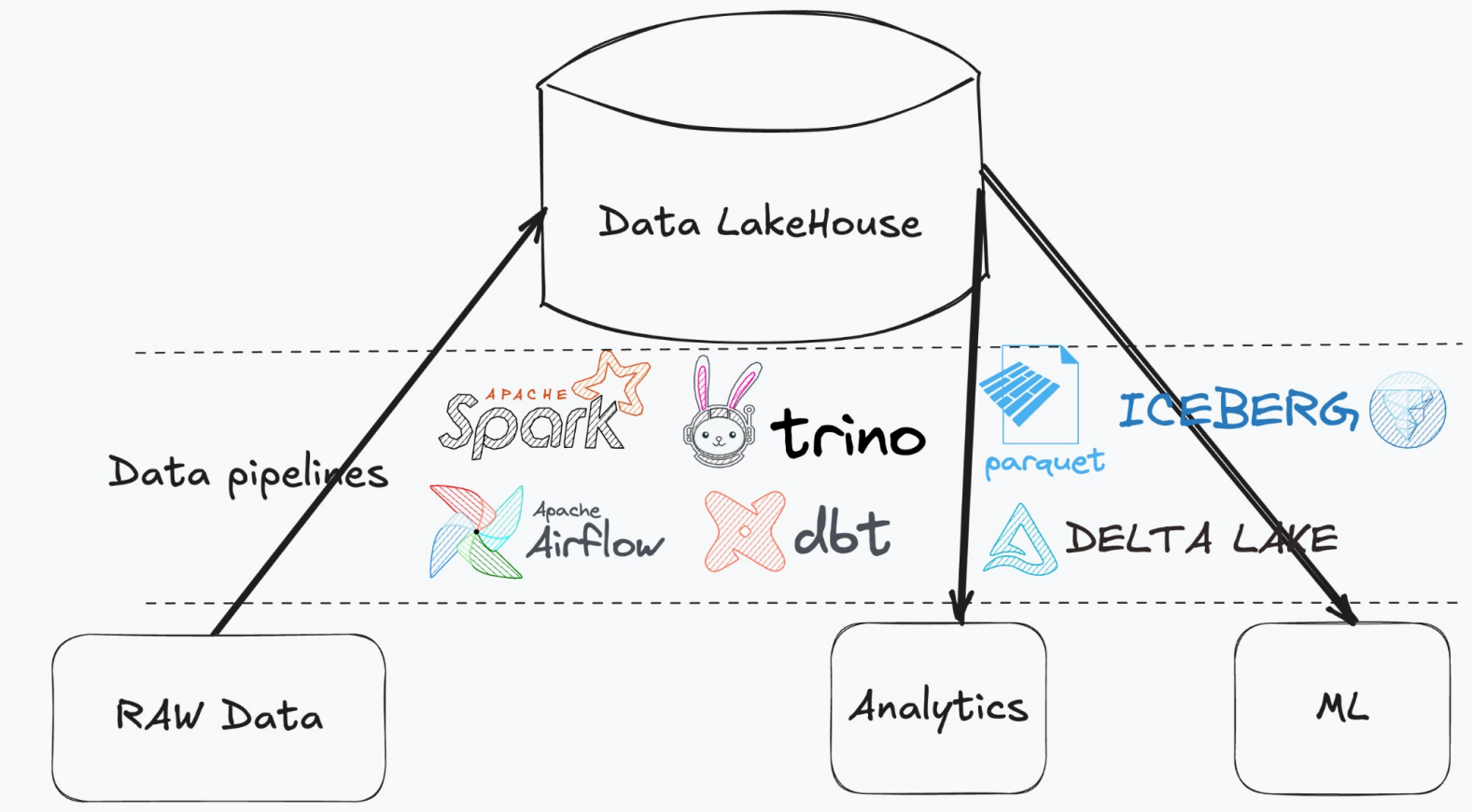
DataMesh

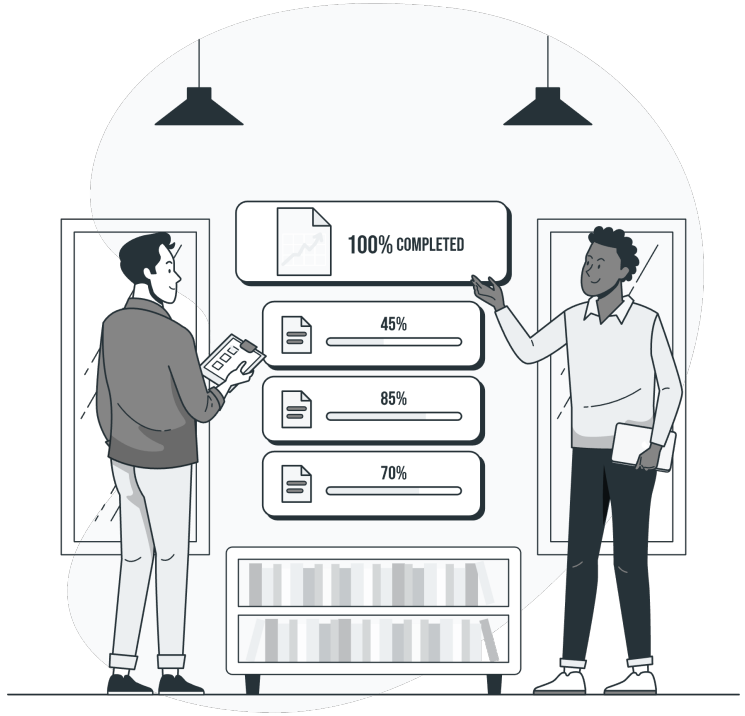
Организационный подход к управлению данными • Полностью децентрализованный • За отдельные данные отвечает владелец, который передает их при необходимости потребителям

DataFabric

Включает в себя элементы DataLakehouse и DataMesh и вносит дополнительно элементы data governance

Процесс работы с данными





Инфраструктура для ML (DL)

Критерий	Своя	Облачная
Обслуживание инфраструктуры	Нужны квалифицированные инженеры (высокие операционные издержки)	–
Инженеры DevOps/MLOps	+	+
Готовность крупного бизнеса передавать чувствительные данные (безопасность)	Нужны специалисты по безопасности	Нет
Расширяемость (Managed решения)	- / +	Ограничена предоставляемыми сервисами
Vendor lock	Нет	Да
Гибкая масштабируемость	Ограничена физическими серверами и используемыми инструментами	Да
Отказоустойчивость	Нужно несколько ЦОДов	+
Цена (особенно при использовании GPU ускорителей)	Получается на порядок дешевле на длинной дистанции	Pay as you go

Вывод: использование только облачной инфраструктуры для DL выходит дорого, идеально использовать гибрид, если это возможно для бизнеса. Разработка внутри, инференс для клиентов снаружи.

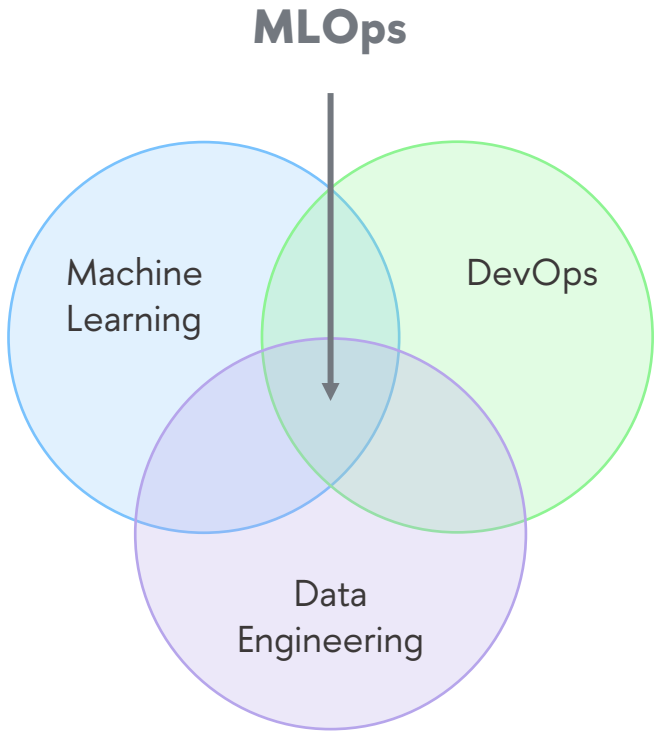
Всегда проводите оценку, какие ресурсы потребуются под вашу систему и будет ли готов бизнес на такие расходы!

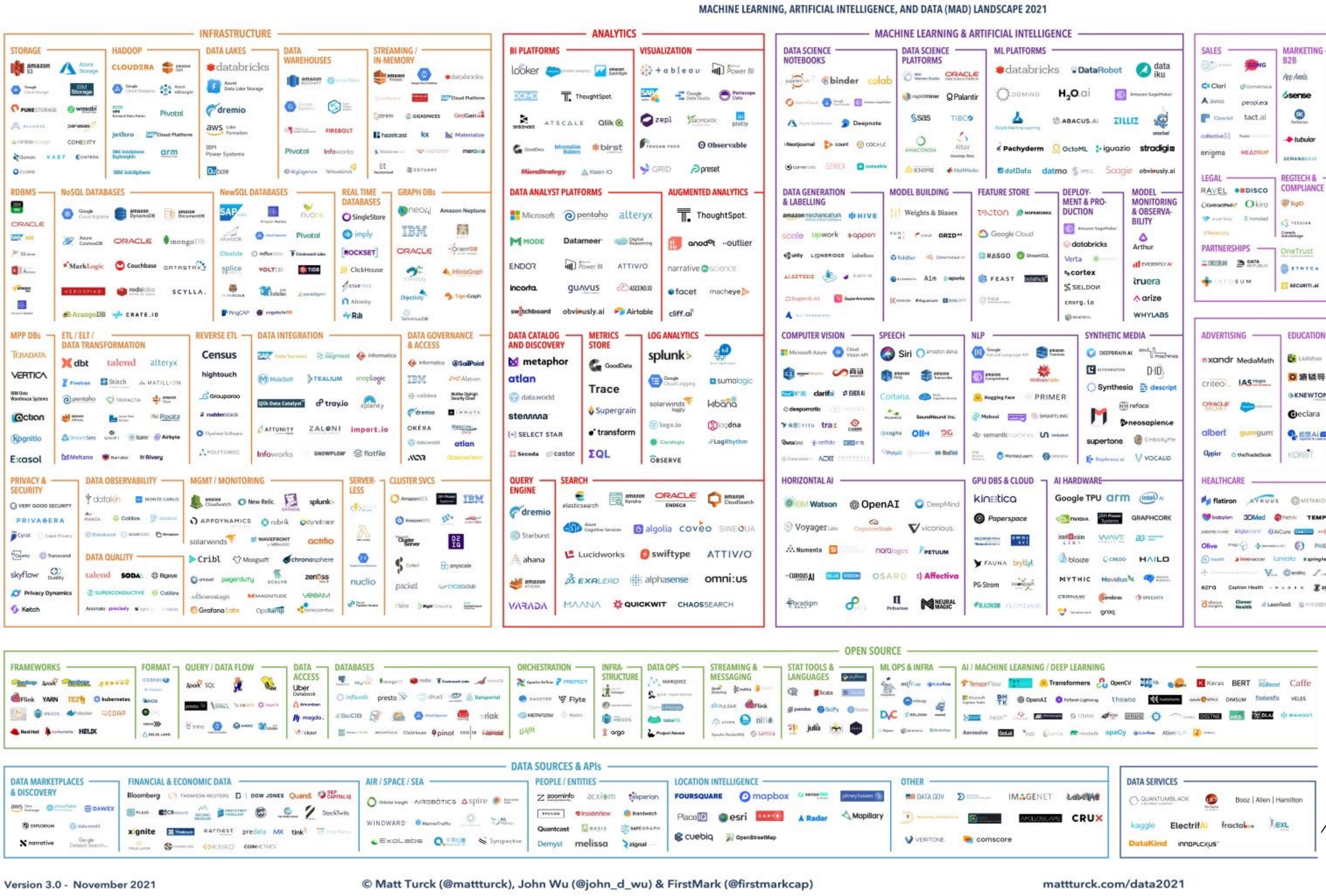
MLOps

Основные плюшки:

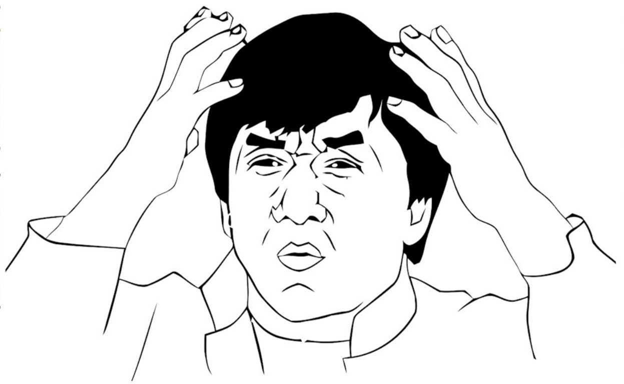
- Автоматизация рутинных процессов
- Масштабируемость
- CI/CD для моделей (canary deploy)
- Воспроизводимость
- Улучшения взаимодействия между командами (DS,OPS,DEV)
- Мониторинг решений

Большое количество инструментов разного уровня зрелости.
Может быть сложно интегрировать их между собой и в бизнес-процессы компании.

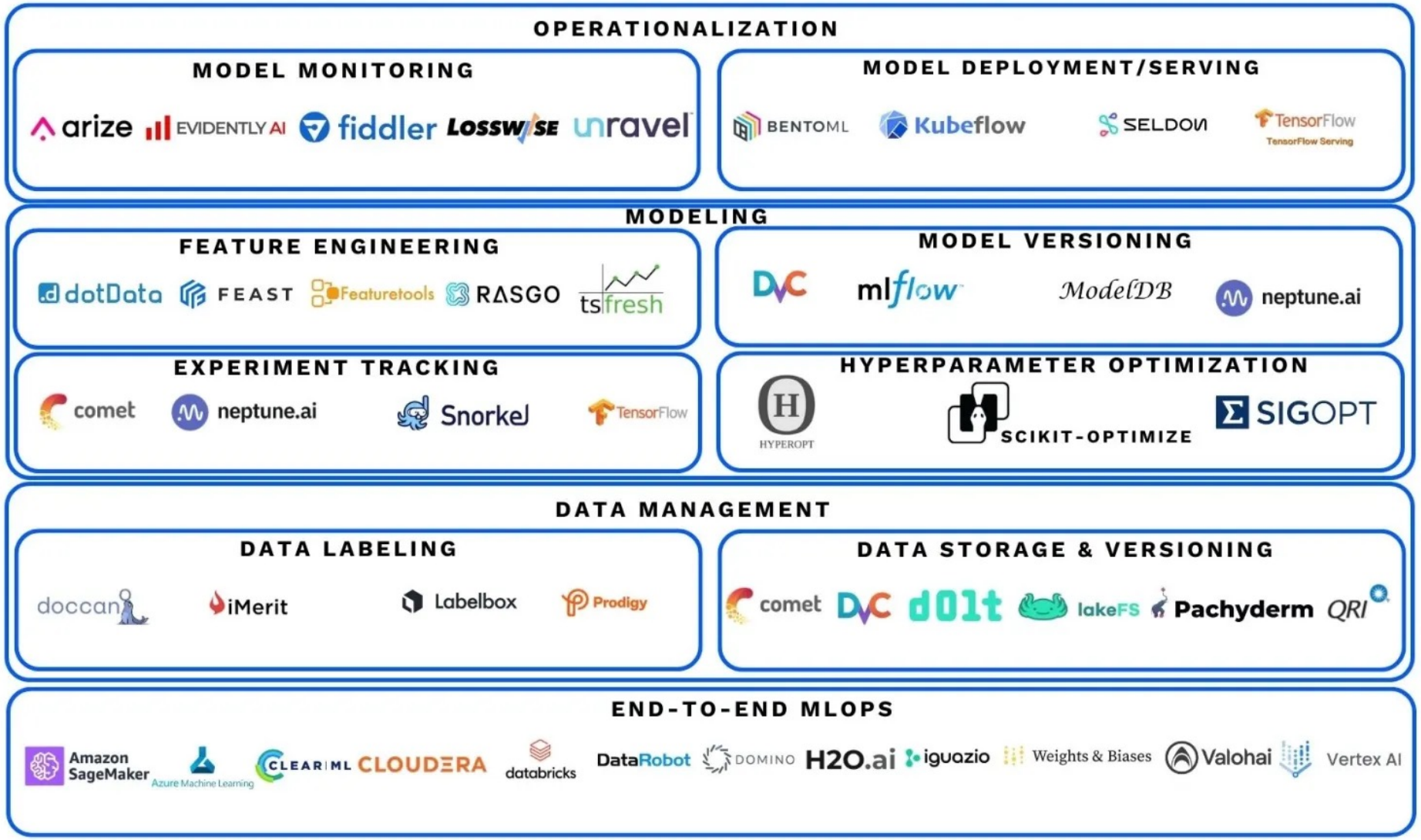




MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021
INFRASTRUCTURE
STORAGE
HADOOP
DATA LAKES
DATA WAREHOUSES
STREAMING / IN-MEMORY
RDBMS
NoSQL DATABASES
NewSQL DATABASES
REAL TIME DATABASES
GRAPH DBs
MPP DBs
ETL / ELT / DATA TRANSFORMATION
REVERSE ETL
DATA INTEGRATION
DATA GOVERNANCE & ACCESS
PRIVACY & SECURITY
DATA OBSERVABILITY
DATA QUALITY
MGMT / MONITORING
SERVER-LESS
CLUSTER SVCS
ANALYTICS
BI PLATFORMS
VISUALIZATION
DATA ANALYST PLATFORMS
AUGMENTED ANALYTICS
DATA CATALOG AND DISCOVERY
METRICS STORE
LOG ANALYTICS
QUERY ENGINE
SEARCH
MACHINE LEARNING & ARTIFICIAL INTELLIGENCE
DATA SCIENCE NOTEBOOKS
DATA SCIENCE PLATFORMS
ML PLATFORMS
DATA GENERATION & LABELLING
MODEL BUILDING
FEATURE STORE
DEPLOYMENT & PRODUCTION
MODEL MONITORING & OBSERVABILITY
COMPUTER VISION
SPEECH
NLP
SYNTHETIC MEDIA
HORIZONTAL AI
GPU DBS & CLOUD
AI HARDWARE
APPLICATIONS – ENTERPRISE
SALES
MARKETING - B2B
MARKETING - B2C
CUSTOMER EXPERIENCE / SERVICE
HUMAN CAPITAL
LEGAL
REGTECH & COMPLIANCE
FINANCE
AUTOMATION & RPA
SECURITY
PARTNERSHIPS
APPLICATIONS – INDUSTRY
ADVERTISING
EDUCATION
REAL ESTATE
GOV'T & INTELLIGENCE
COMMERCE
FINANCE - LENDING
INSURANCE
FINANCE - INVESTING
HEALTHCARE
LIFE SCIENCES
TRANSPORTATION
AGRICULTURE
INDUSTRIAL
OTHER
OPEN SOURCE
FRAMEWORKS
FORMAT
QUERY / DATA FLOW
DATA ACCESS
DATABASES
ORCHESTRATION
INFRA-STRUCTURE
DATA OPS
STREAMING & MESSAGING
STAT TOOLS & LANGUAGES
ML OPS & INFRA
AI / MACHINE LEARNING / DEEP LEARNING
DATA SOURCES & APIs
DATA MARKETPLACES & DISCOVERY
FINANCIAL & ECONOMIC DATA
AIR / SPACE / SEA
PEOPLE / ENTITIES
LOCATION INTELLIGENCE
OTHER
DATA SERVICES
Version 3.0 - November 2021
© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)
mattturck.com/data2021



Распространенные MLOps-инструменты



Основные инструменты

Версионирование и отслеживание экспериментов: MLflow, DVC (Data Version Control), WANDB

CI/CD пайплайны: Jenkins, GitLab CI, CircleCI, Kubeflow Pipelines

Контейнеризация и оркестрация: Docker, Kubernetes

Мониторинг: Prometheus, Grafana, AWS CloudWatch, Azure Monitor

DataOps: Apache Airflow, Prefect, Kafka, DBT

ML Платформы: Weights & Biases, Neptune.ai, ClearML.

Сервинг моделей: TensorFlow Serving, TorchServe, Flask, FastAPI, KServe, Nvidia Triton, Nuclio (Serveless)

Различие в инфраструктуре при обучении модели и для инференса

	Обучение	Инференс
Цель	Улучшение качества модели	Меньшая задержка на ответ
Вычислительные ресурсы	Необходимы большие ресурсы, в т.ч. GPU	Ресурсов требуется меньше, GPU тоже меньше объема
Загрузка	Батч загрузка, много длинных задач	Запрос/Ответ, может быть батч. Оптимизировано под быстрый ответ
Необходимость масштабирования	Большие кластеры для обучения крупных моделей	Автомасштабирование при увеличении нагрузки
Задержка	-	Адекватное для человека время на ответ
Дисковое пространство	Большие данные, озера данных	Минимальные объемы (модель, конфиг, логи, дополнительный код)
Отказоустойчивость	-	Высокая избыточность, важность безотказной работы
Мониторинг	Mlflow и подобные для трекинга экспериментов	Постоянно следить за качеством модели, оценивать data drift и performance degradation

Деплоймент моделей (инференс)

По типу выполнения:

Онлайн инференс:

- Специализированные облачные решения (AWS SageMaker, GCP Cloud Run, Selectel Inference Platform и др.)
- Или развернутые там же opensource инструменты

ОБЛАКО

- Существует множество решений, подходящих под разные технологии, но нет единого стандартного
- Необходимо подбирать инструмент под конкретную технологию
- С развитием LLM начали появляться специализированные решения для инференса с возможностью запуска (квантизованных) и обычных моделей ollama, vllm и т.д.

СВОЯ ИНФРАСТРУКТУРА

По типу:

- REST/GRPC
- Model as a Service (MAS)
- Model-on-Demand
- ...

Оффлайн инференс:

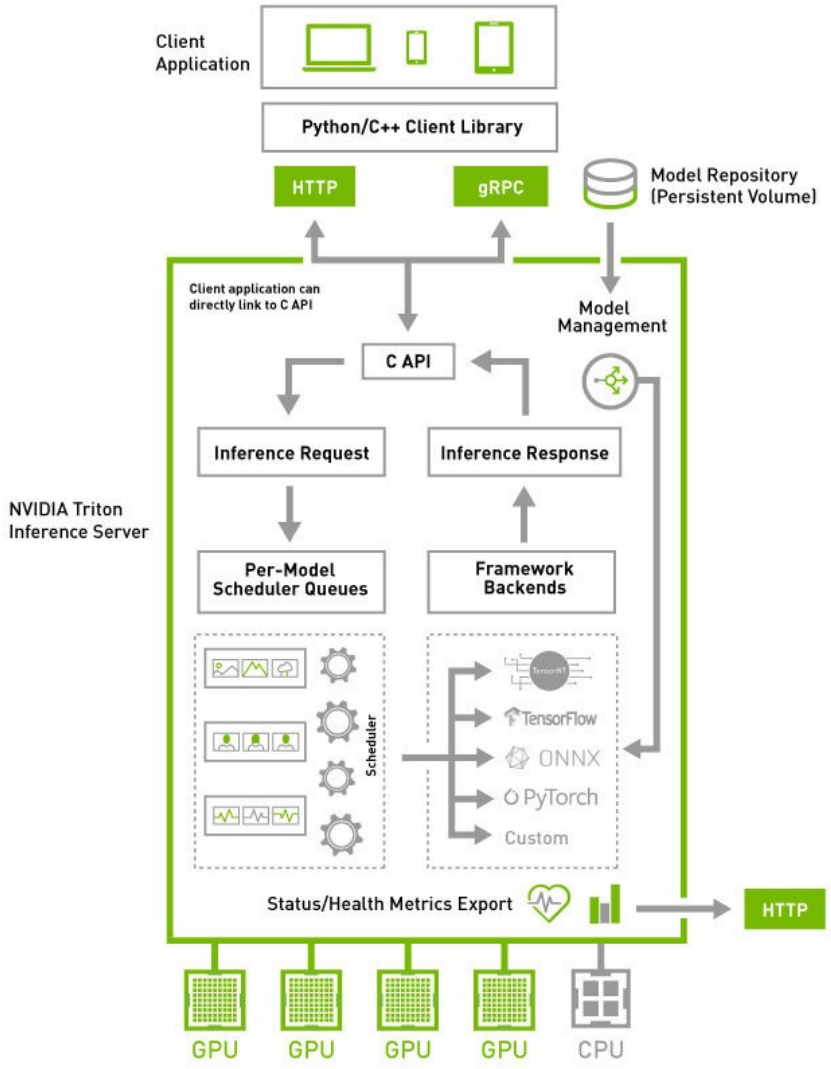
- Батч обработка
- REST/GRPC

Деплоймент Model as a Service (REST/GRPC)


```
from fastapi import FastAPI
from pydantic import BaseModel

class ModelRequest(BaseModel):
    name: str
    price: float

app = FastAPI()

@app.post("/predict/")
async def next_best_offer(request: ModelRequest):
    return modelService.getModel().predict(request)
```

+	Ds-way подойдет для быстрого прототипа решения
–	Нет поддержки cloud native и интеграции с k8s, как следствие, решение не масштабируется
–	Непонятно, как будет работать под нагрузкой?
–	Быстрее будет собирать запросы в батчи или отправлять в модель только по одному?
–	Если потребуется добавить новый функционал (например, авторизацию и аутентификацию), нужно внести изменения в код. Изменения могут быть довольно частыми
–	Нет мониторинга качества и непонятно, когда модель станет деградировать
–	Может меняться как код, так и сама модель, придется все переписывать заново

Inference Servers



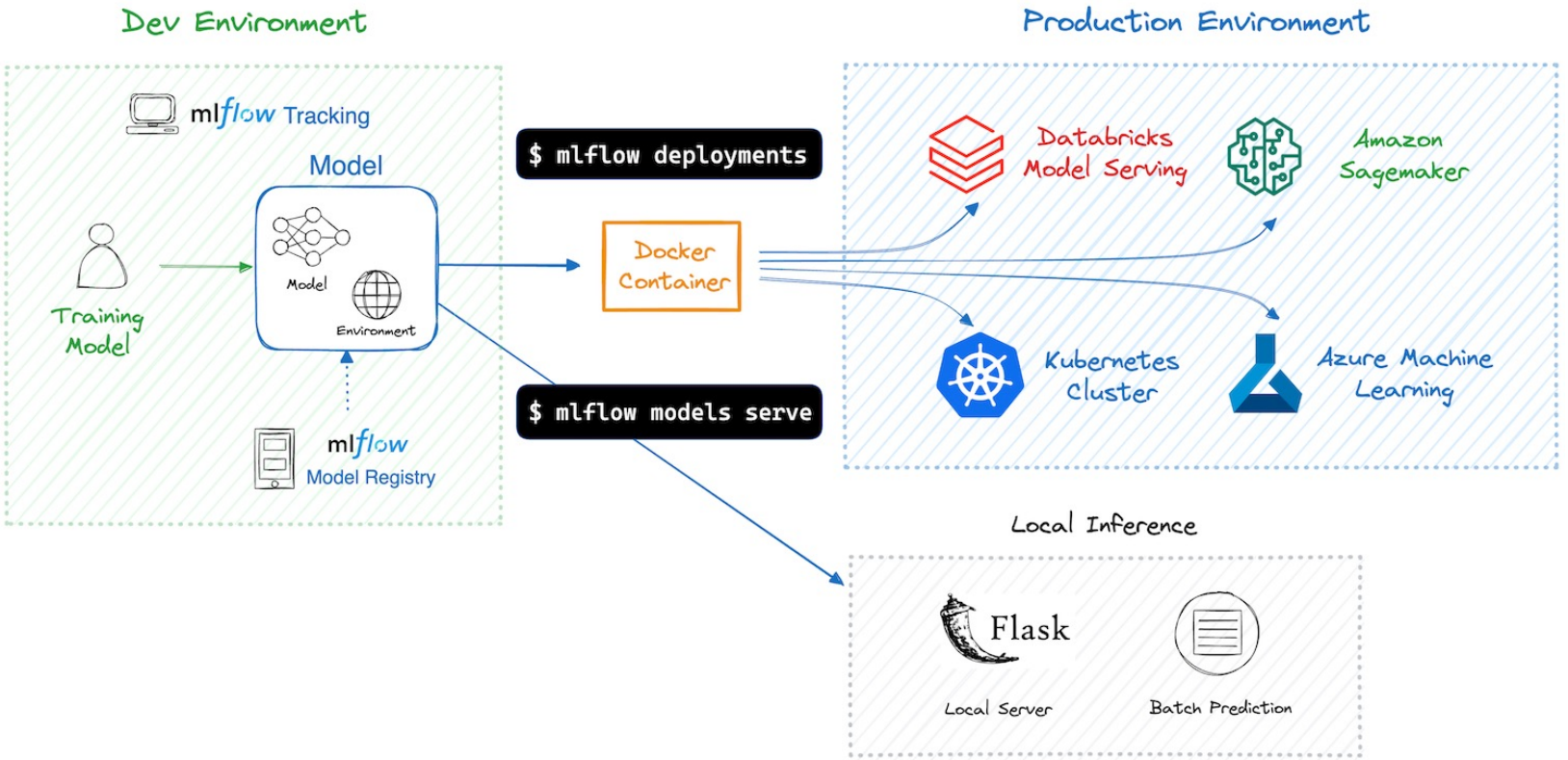
Triton Inference Server

+	Одновременное выполнение моделей
+	GPU/CPU
+	Адаптивный (динамический) батчинг
+	Поддержка практически всех известных бэкендов (TensorRT, TensorFlow, PyTorch, ONNX, OpenVINO, Python и другие)
+	Горячая замена моделей
+	Много дополнительных функций
–	Весьма сложен

Другие:

- TFServing - tensorflow
- TorchServe - pytorch

MLFlow Serving



Локальный инференс

–	Для тестирования
–	Не подходит под прод нагрузку
+	Быстро и просто развернуть

Прод инференс

+	Лишен недостатков локального
+	Асинхронный, отдельный пул обработчиков, можно одновременно сервить несколько моделей

Model Serving. Kserve

- Поддерживает большинство известных райнтаймов TFServing, TorchServe, Triton Inference Server

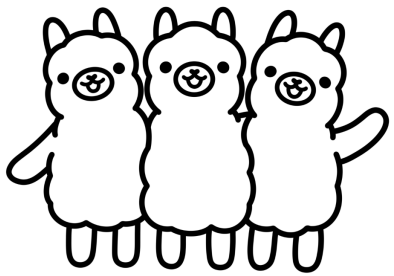
- Cloud native
  - Масштабирование на CPU/GPU
  - Управление версиями
  - Батчинг
  - Логирование (запрос/ответ)
  - Управление трафиком
  - Метрики

и др.

Специализированные решения

LLM - vLLM, Ollama, llama.cpp

- llama3.2 1b, 3b
- llama3.1 8b,70b,450b
- gemma2 2b,9b,27b
- qwen2.5 0.5...72b
- mistral-nemo 12b
- mistral
- mixtral
- llava 7b,13b,34b



https://ollama.com/library

ЗАГОРУЛЬКИН ДМИТРИЙ ЭДУАРДОВИЧ

Заместитель директора Центра стратегической аналитики и больших данных ИСИЭЗ НИУ ВШЭ

Заместитель руководителя системы интеллектуального анализа больших данных iFORA

dzagorulkin@hse.ru

Сайт iFORA

iFORA в Telegram

iFORA-экспрессы

Хотите у нас работать или пройти стажировку?
Сканируйте QR-код