



ГАЗПРОМБАНК

Газпромбанк — входит в тройку крупнейших универсальных банков России и занимает третье место в списке банков Центральной и Восточной Европы по размеру собственного капитала.

Мы не просто банкиры: мы создаем искусственный интеллект, разрабатываем квантовые компьютеры, придумываем новые цифровые продукты, поддерживаем профессиональное комьюнити, не забывая при этом, что в центре инноваций — человек.

Поэтому мы уделяем особую роль развитию. Мы создаем IT-сообщества и с помощью школы спикеров помогаем коллегам делиться успехами. Регулярно проводим митапы и даем возможность повышать квалификацию во внутренних школах разработки.



Что такое MLOps

Ещё больше *Ops



DevOps

DockOps

ArchOps

CloudOps

ChatOps

DevSecOps

UXOps

FinOps

BizOps

ResearchOps

ServerlessOps

DBOps

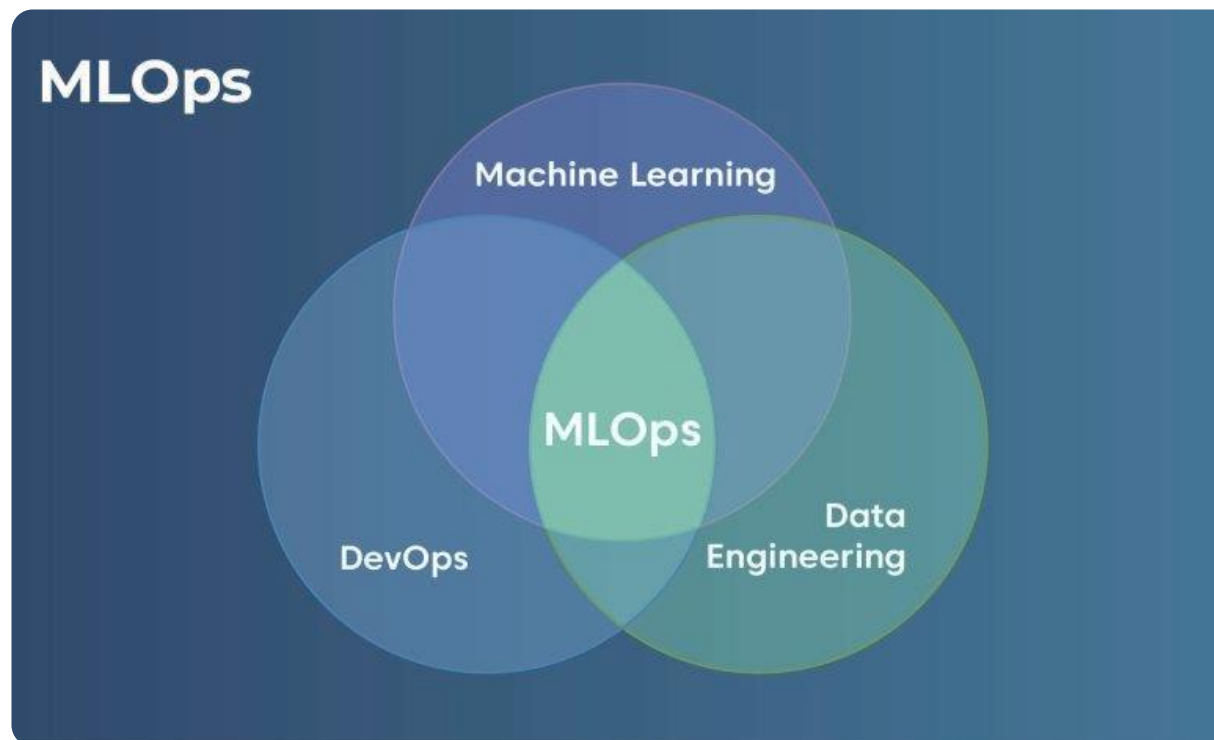
MLOps

DataOps

FrontOps

DesignOps

Что такое MLOps?



Искусственный интеллект



Что такое ИИ?

$$y = f(x)$$

Что такое ИИ?

Искусственный интеллект – наука и технология создания интеллектуальных машин, особенно интеллектуальных алгоритмов и компьютерных программ

- **Машинное обучение** – обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.
- **Глубокое обучение** – совокупность методов машинного обучения, основанных на обучении представлениям, а не специализированным алгоритмам под конкретные задачи.



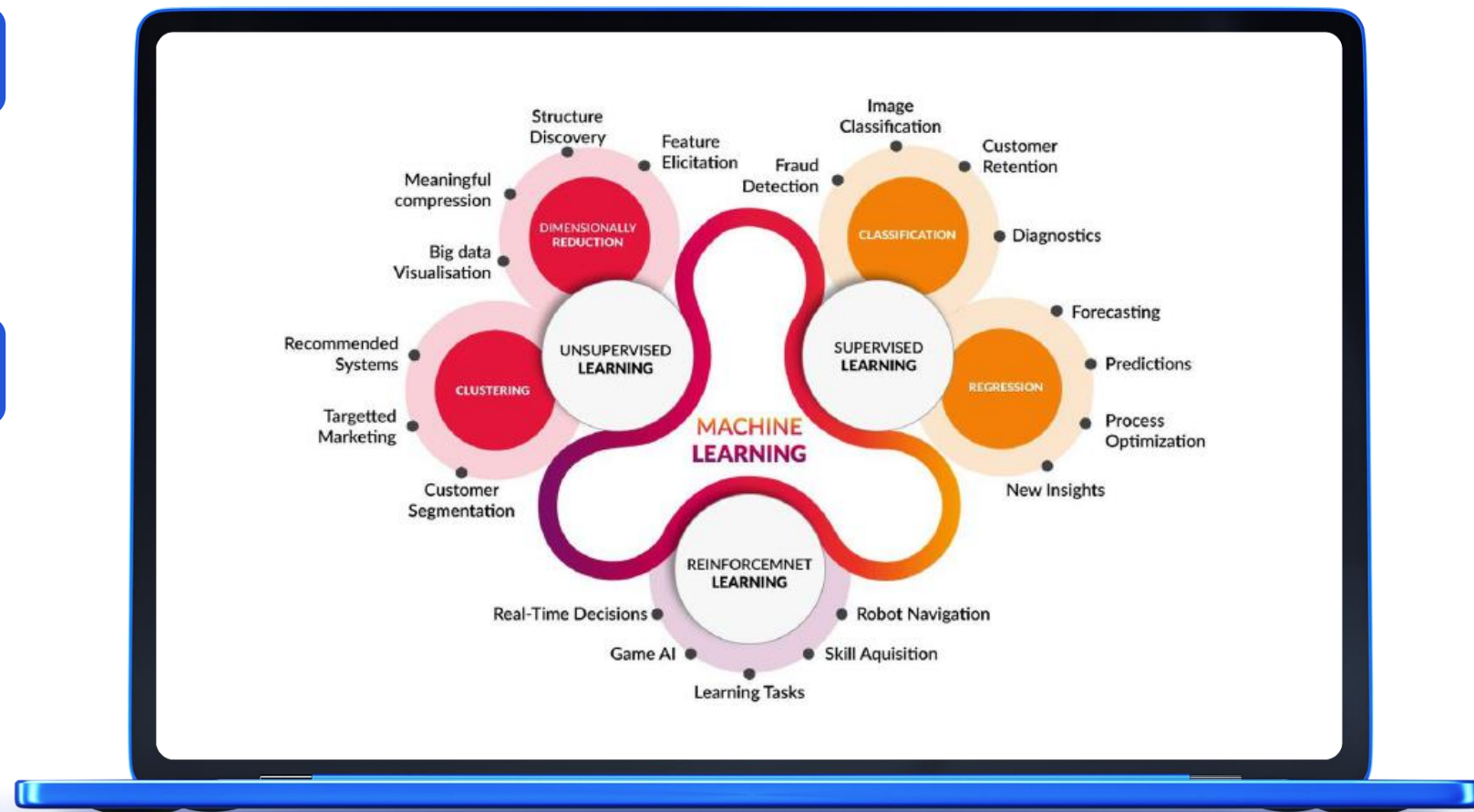
Основные задачи классического ML

С учителем (Supervised)

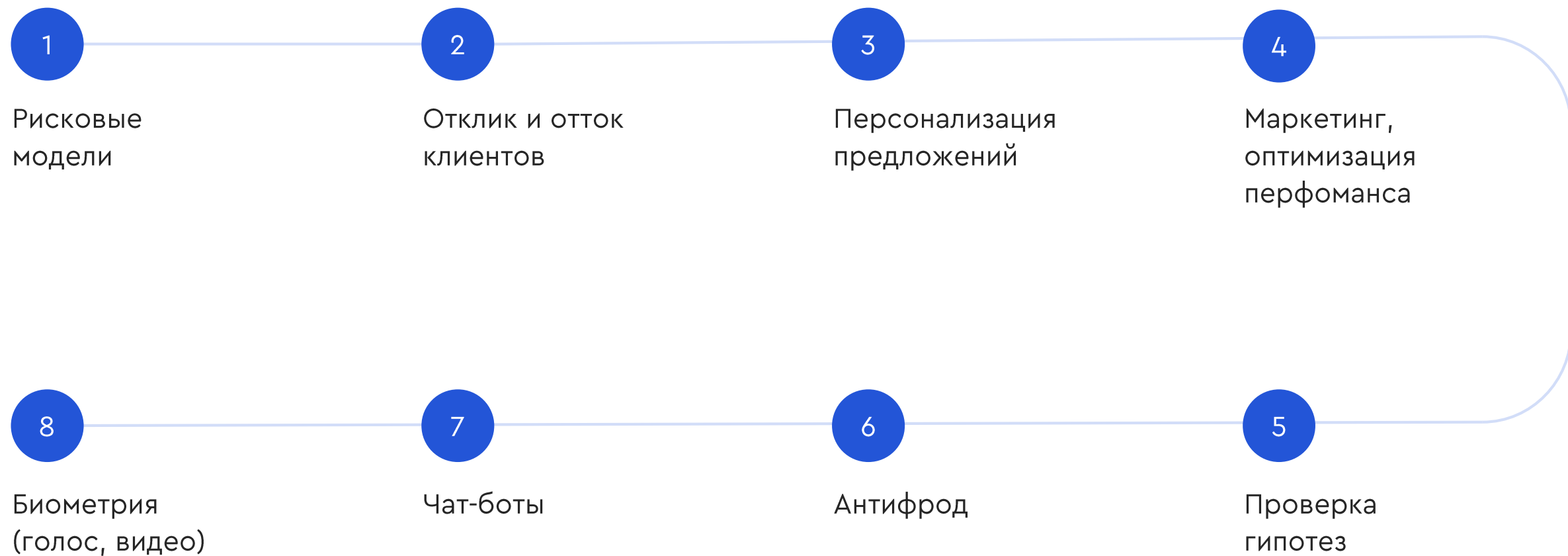
- Регрессия (Regression)
- Классификация (Classification)

Без учителя (Unsupervised)

- Кластеризация (Clustering)
- Снижение размерности (Dimensionality reduction)



Задачи, которые сейчас решает ИИ в банках



Жизненный цикл – процесс обучения



Алгоритм
модели
машинного
обучения

Жизненный цикл – процесс обучения

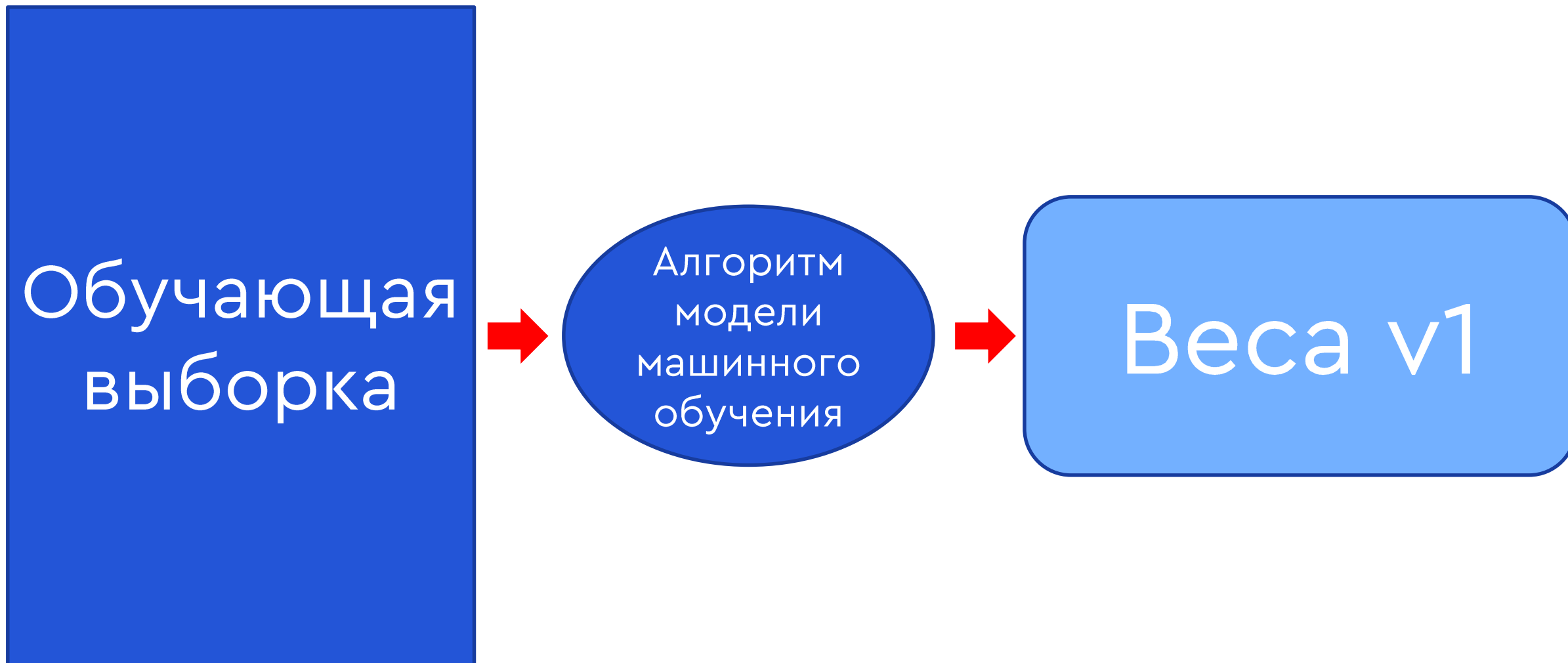


Обучающая
выборка



Алгоритм
модели
машинного
обучения

Жизненный цикл – процесс обучения



Жизненный цикл – процесс обучения



Обучающая
выборка

Алгоритм
модели
машинного
обучения

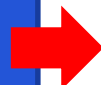
Весы v1

Модель машинного обучения

Жизненный цикл – дообучение



Обновлённая
обучающая
выборка



Алгоритм
модели
машинного
обучения



Весы v2

Модель машинного обучения



Обновлённая
обучающая
выборка

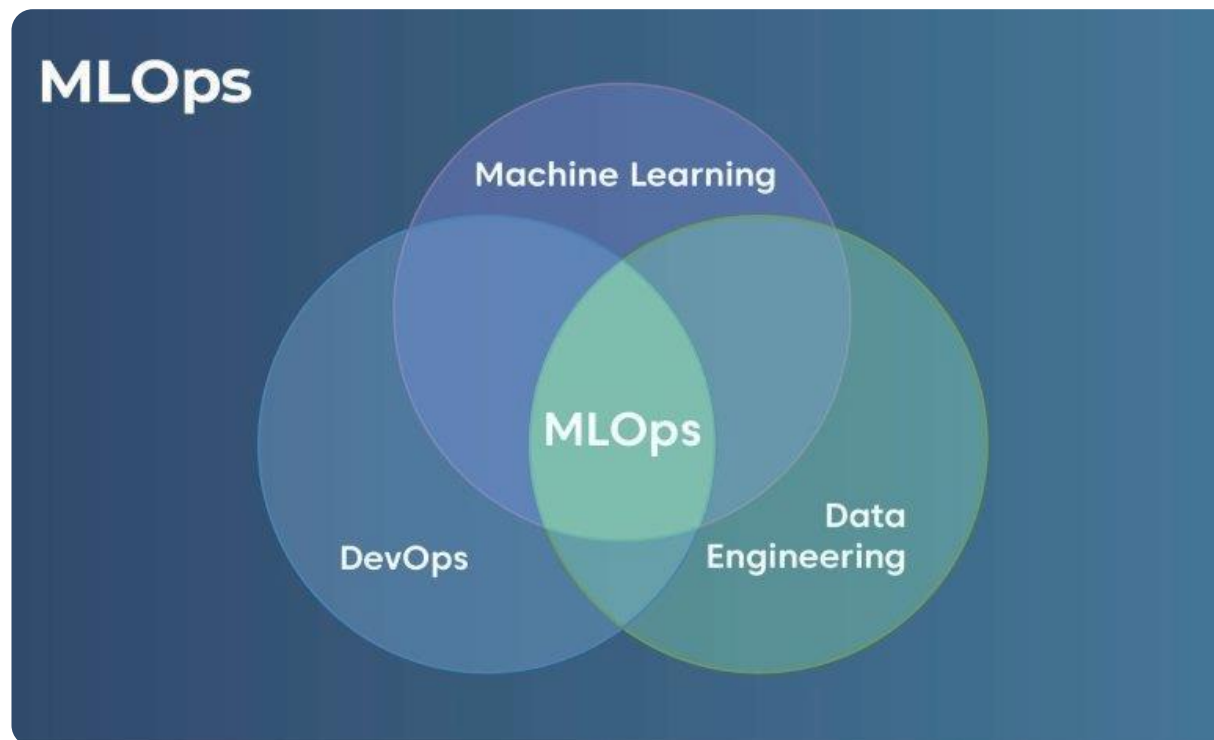
Алгоритм
модели
машинного
обучения

Весы v2

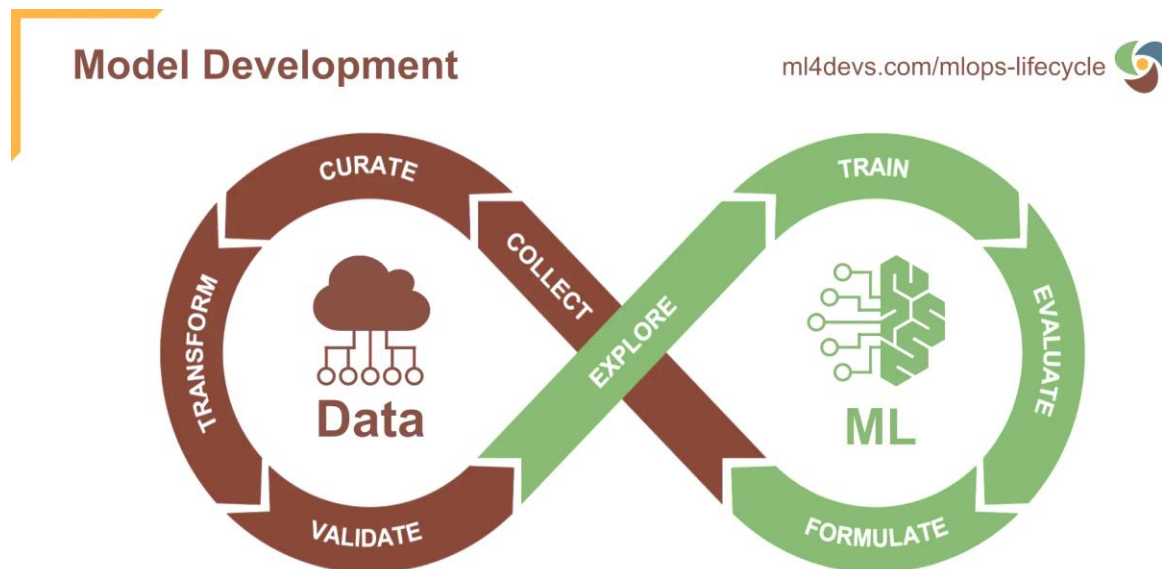
Модель машинного обучения

Эксперимент

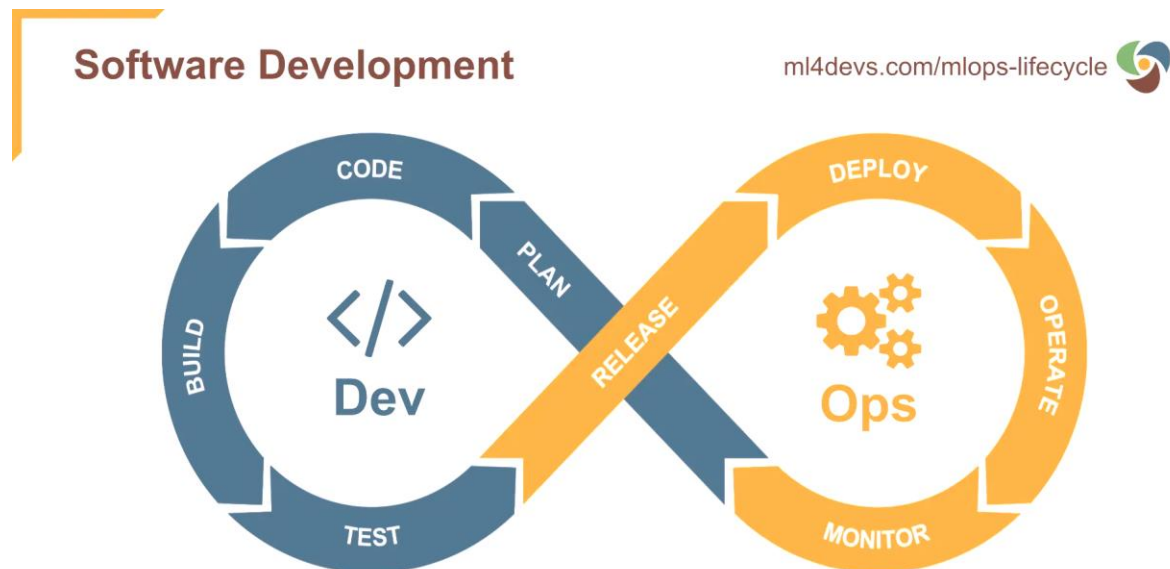
Что такое MLOps?



Что такое MLOps?



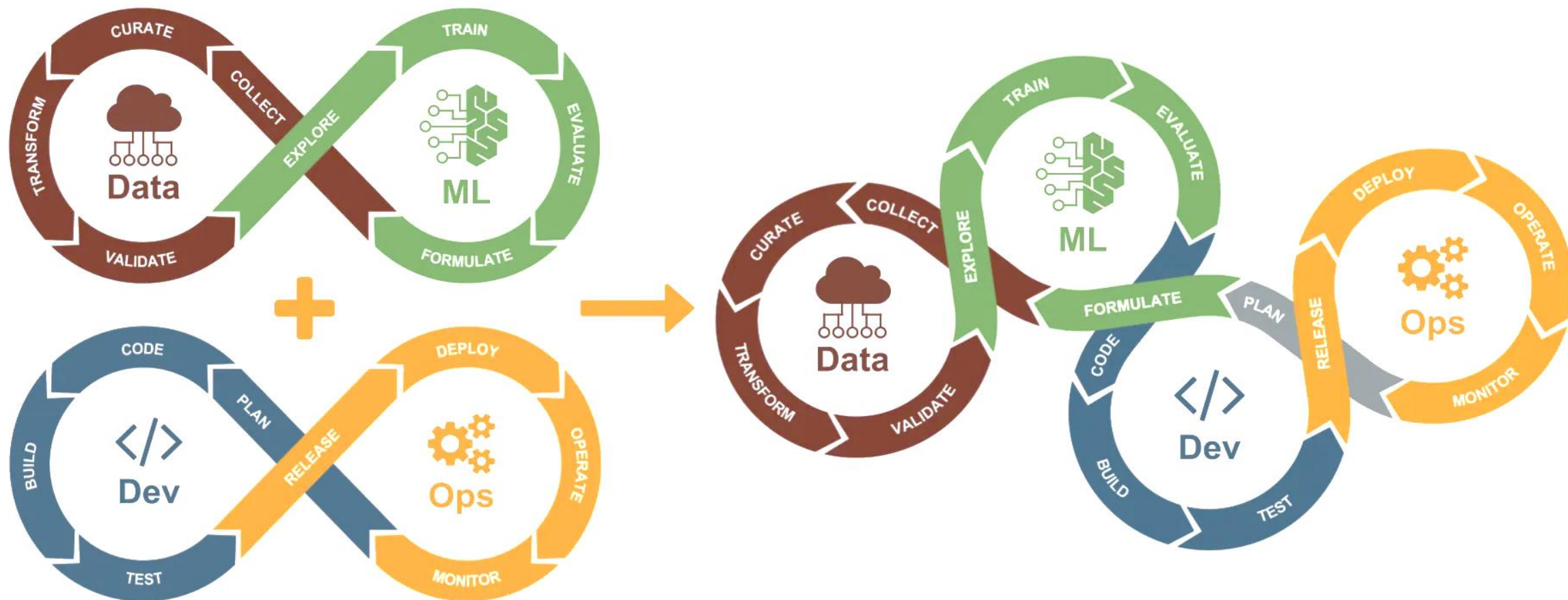
Что такое MLOps?



Что такое MLOps?

MLOps = DataML + DevOps

ml4devs.com/mlops-lifecycle 



Челенжи MLOps



- Подготовка данных (DataEngineering).
- Обучение/дообучение модели.
- Версионирование обучающих выборок и весов (Трекинг эксперимента).
- Организация разработки.
- Безопасность данных.
- Организация CI/CD.

Челенжи MLOps



- Подготовка данных (DataEngineering).
- Обучение/дообучение модели.
- Версионирование обучающих выборок и весов (Трекинг эксперимента).
- Организация разработки.
- Безопасность данных.
- Организация CI/CD.

Подготовка данных



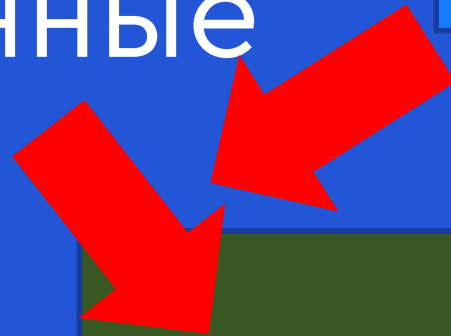
Доступные
данные

Анализ и
планирование

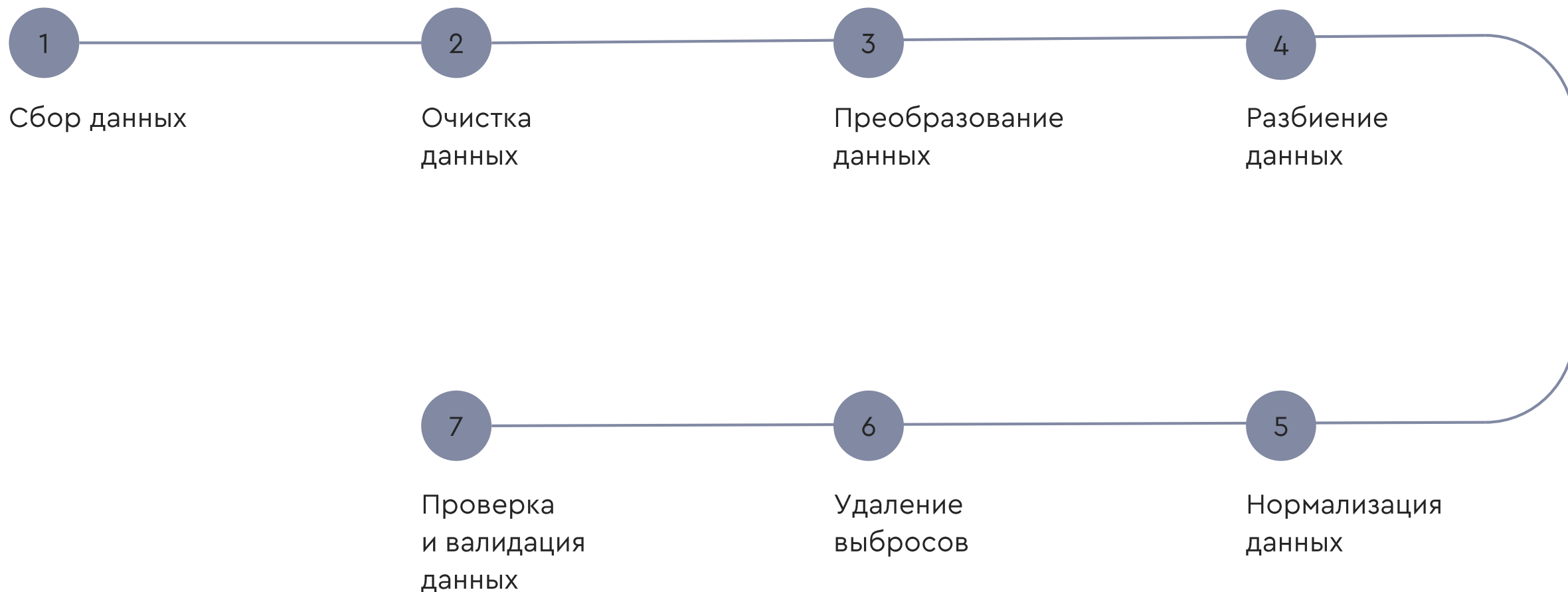
Сбор

Форматирование
и очистка

Агрегация



Подготовка данных с точки зрения ML

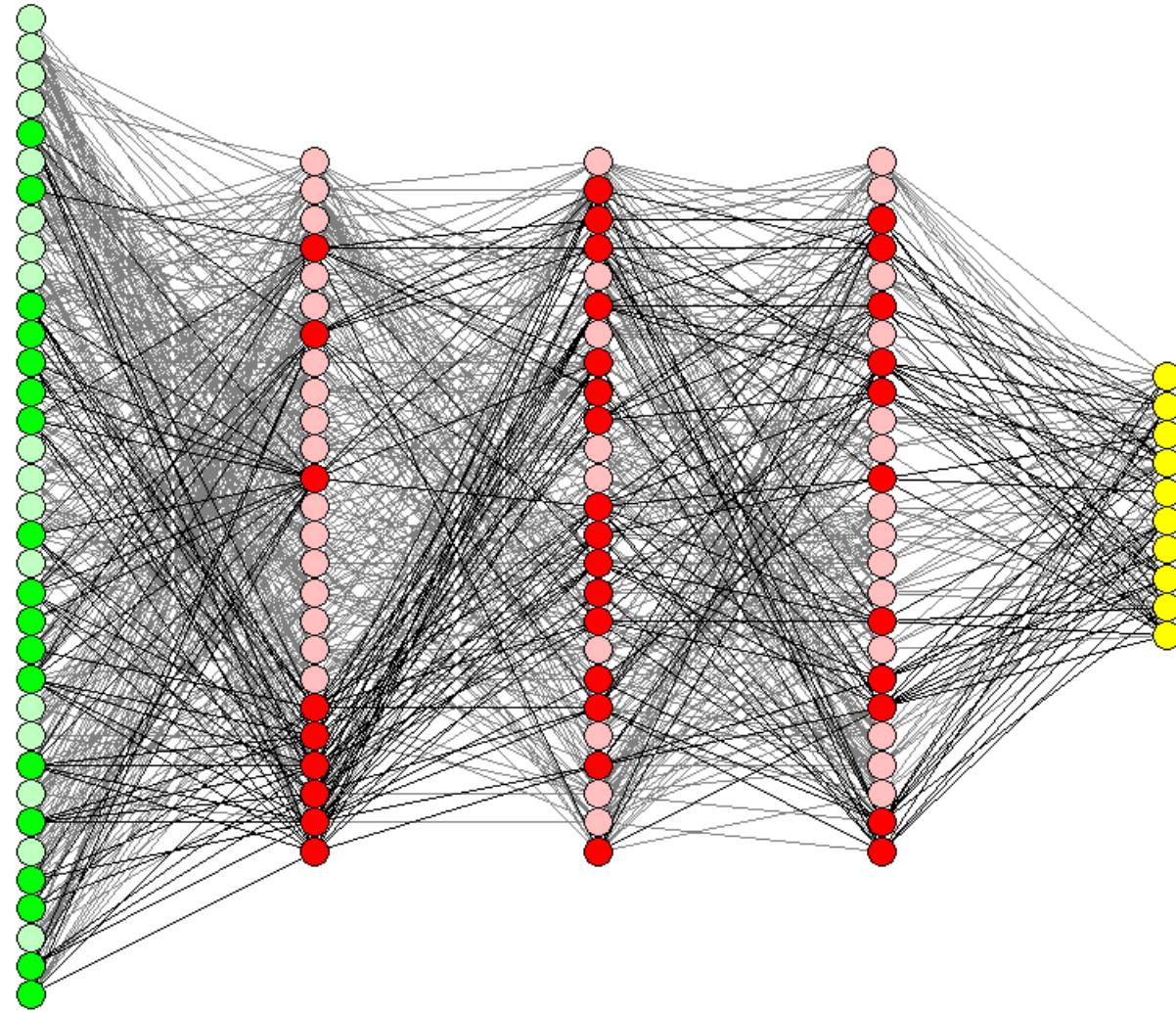


Челенжи MLOps



- Подготовка данных (DataEngineering).
- **Обучение/дообучение модели.**
- Версионирование обучающих выборок и весов (Трекинг эксперимента).
- Организация разработки.
- Безопасность данных.
- Организация CI/CD.

Обучение/дообучение модели



Челенжи MLOps



- Подготовка данных (DataEngineering).
- Обучение/дообучение модели.
- Версионирование обучающих выборок и весов (Трекинг эксперимента).
- Организация разработки.
- Безопасность данных.
- Организация CI/CD.

Версионирование обучающих выборок и весов



#	Lab	Dataset	Size (TB)	Tokens (trillion)	Notes
1	Google	Piper monorepo	86TB	<i>37.9T</i>	DIDACT, code only. From 2016 paper.
2	OpenAI	GPT-4	<i>40TB</i>	<i>20T</i>	1T model ∴ 20T tokens. gdb said 40TB.
3	TTI	RefinedWeb	<i>23.2TB</i>	5.0T	CC-only dataset prepared by UAE.
4	DeepMind	MassiveText (ml)	<i>20TB</i>	5.0T	From Retro paper.
5	Google	PaLM 2	<i>13TB</i>	3.6T	From PaLM 2 CNBC report.
6	Google	Infiniset	<i>12.6TB</i>	2.8T	From LaMDA paper.

Table. 2023 largest dataset estimates to Jun/2023. Rounded. Disclosed in **bold**. Determined in *italics*. For similar models, see my *What's in my AI* paper.

Статистика по размеру моделей в Газпромбанк

1

Размер обучающей
выборки – от 100 мб
до бесконечности

2

Размер весов
от 2 до 6 ГБ

3

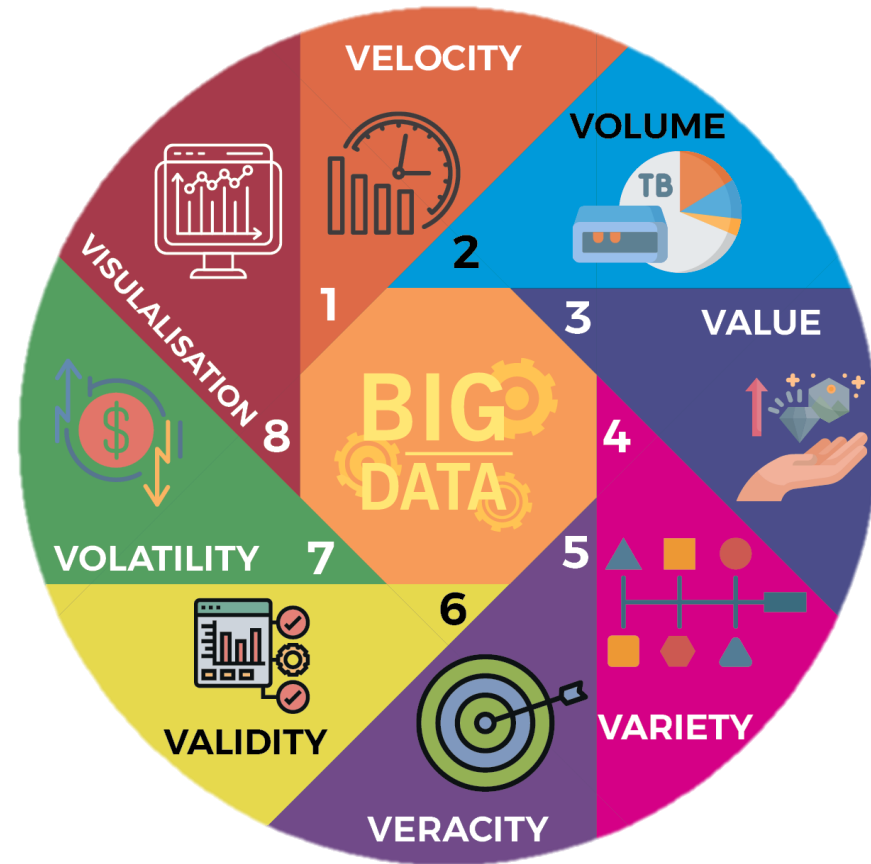
Размер Docker Image
модели от 2.5 до 35 ГБ

Челенжи MLOps



- Подготовка данных (DataEngineering).
- Обучение/дообучение модели.
- Версионирование обучающих выборок и весов (Трекинг эксперимента).
- **Организация разработки.**
- Безопасность данных.
- Организация CI/CD.

Организация разработки



Организация разработки



Челенжи MLOps



- Подготовка данных (DataEngineering).
- Обучение/дообучение модели.
- Версионирование обучающих выборок и весов (Трекинг эксперимента).
- Организация разработки.
- **Безопасность данных.**
- Организация CI/CD.



Безопасность данных



Безопасность данных



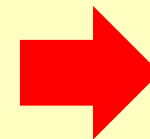
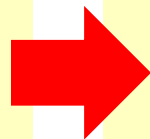
Internet

No Internet

DEV

STAGE

PROD

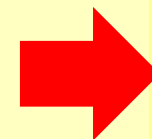
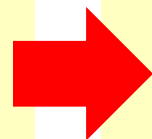




Internet



No Internet

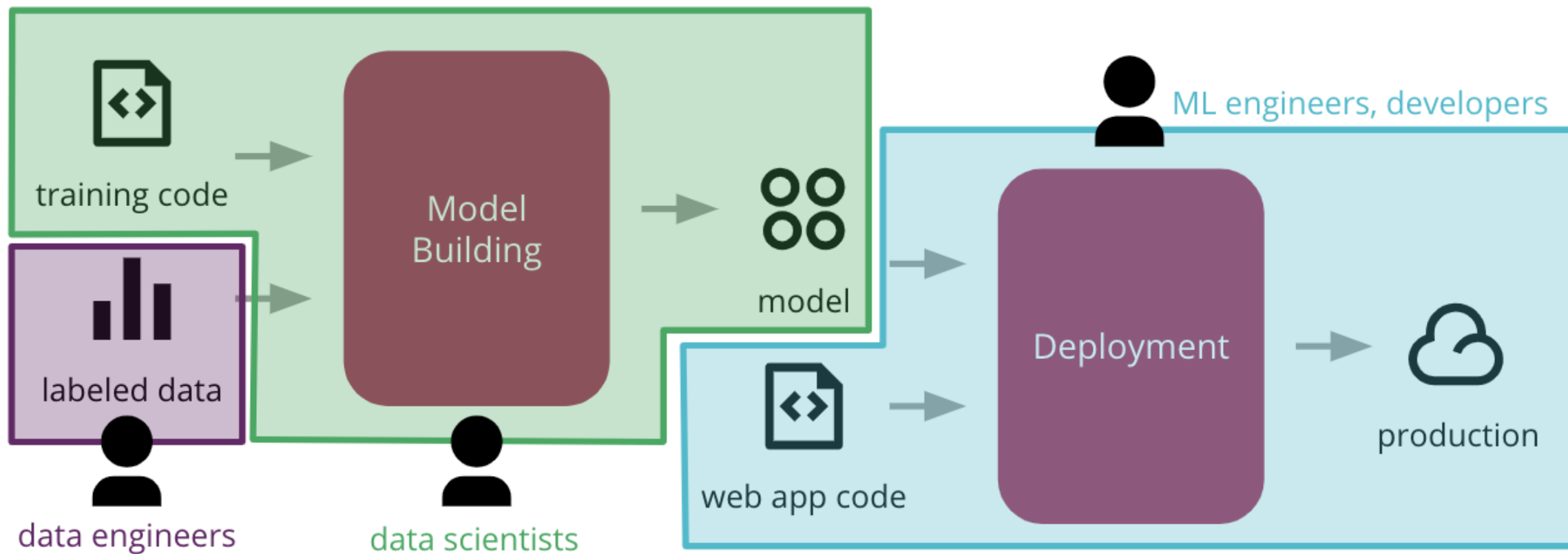


Челенжи MLOps

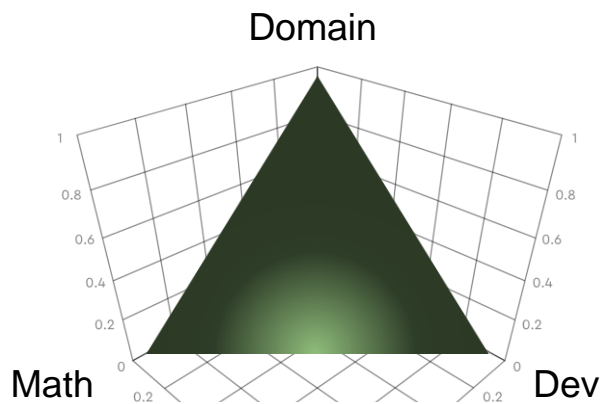


- Подготовка данных (DataEngineering).
- Обучение/дообучение модели.
- Версионирование обучающих выборок и весов (Трекинг эксперимента).
- Организация разработки.
- Безопасность данных.
- **Организация CI/CD.**

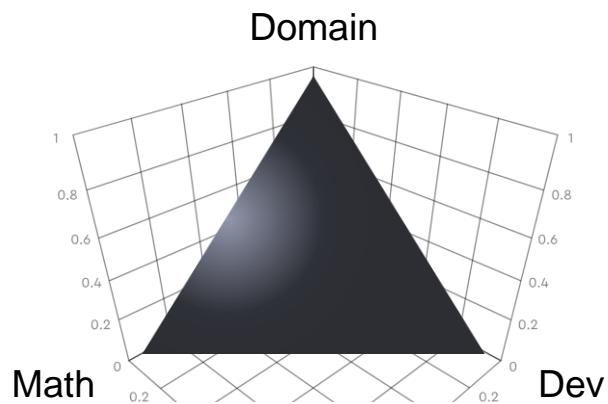
Организация CI/CD для Machine Learning



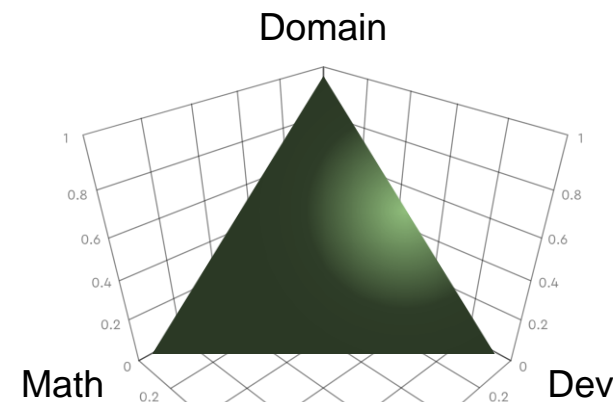
Специалисты в Machine Learning



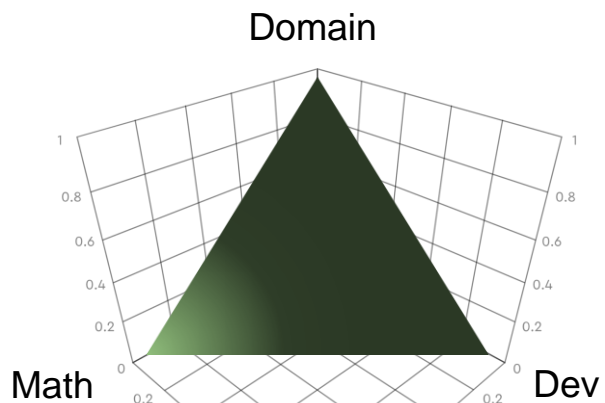
ML Engineer



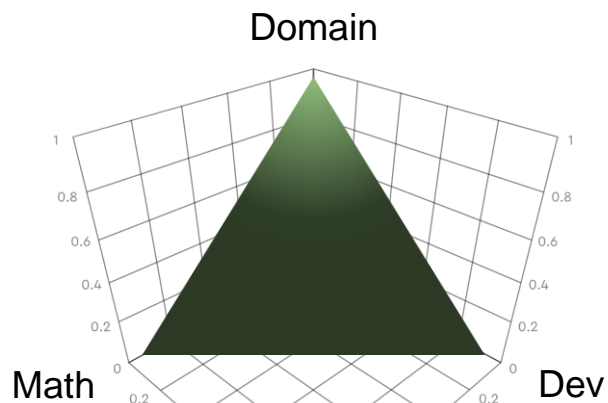
Data Analyst



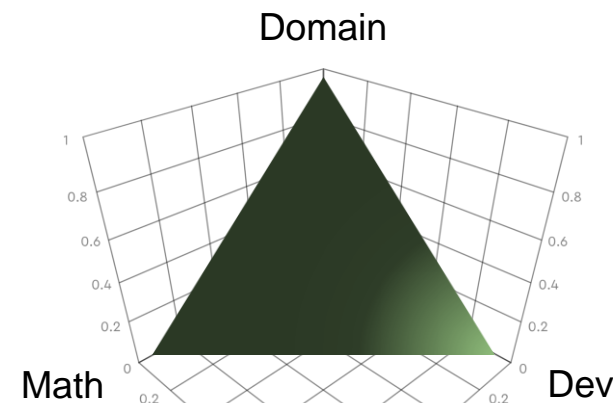
Data Engineer



ML Researcher



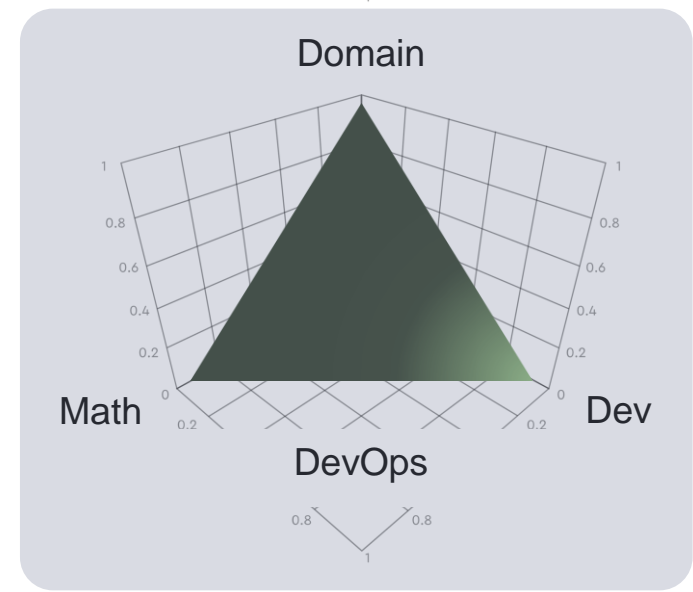
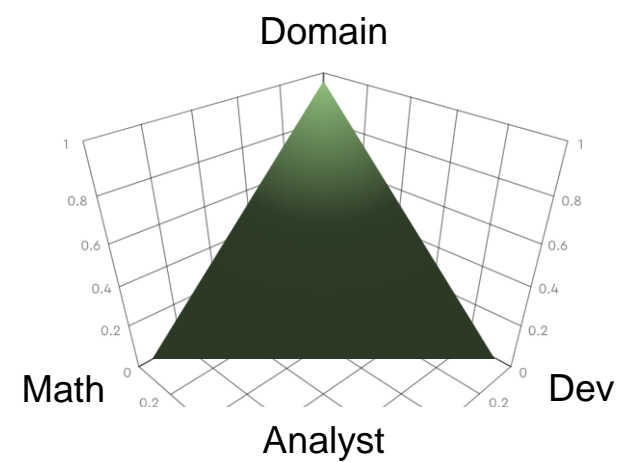
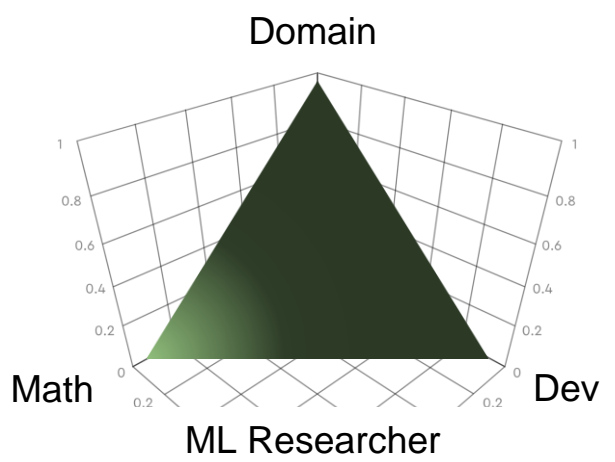
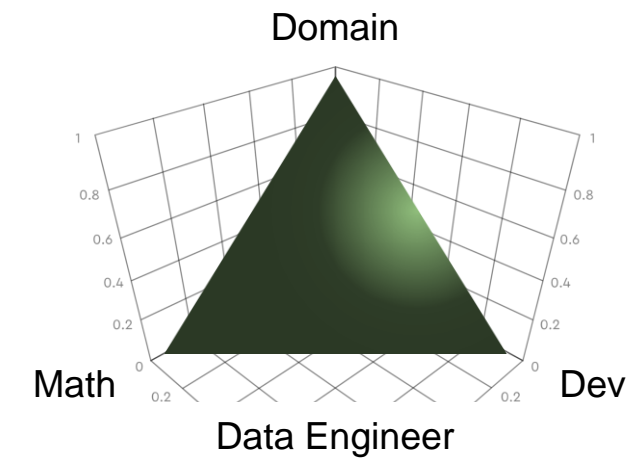
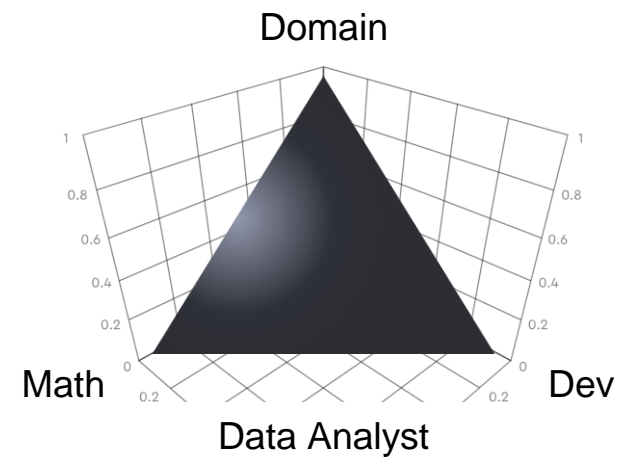
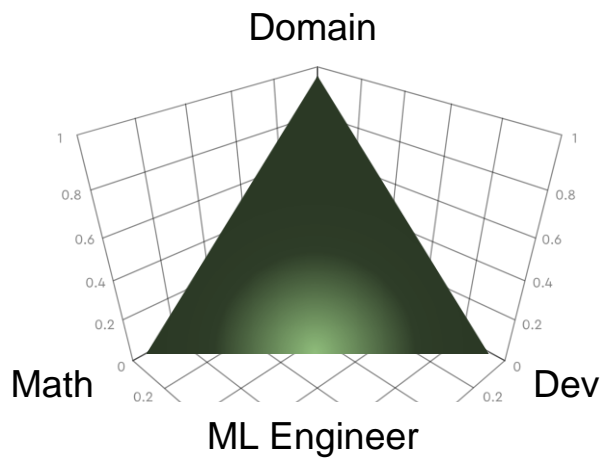
Analyst



DevOps



Специалисты в Machine Learning



DevOps составляющая MLOps



- Источники данных
- Подготовка данных / ETL процессы
- Разработка
- Трекинг эксперимента, сохранение весов.
- Тестирование и доставка моделей
- Мониторинг

DevOps составляющая MLOps



- **Источники данных**
- Подготовка данных / ETL процессы
- Разработка
- Трекинг эксперимента, сохранение весов.
- Тестирование и доставка моделей
- Мониторинг

Типы источников данных



- **SQL**
PostgreSQL/MySQL/MSSQL/Oracle
- **NoSQL**
MongoDB/Kafka/Greenplum
- **Object storage**
S3
- **Block storage**
HDFS
- **WWW**
Yandex metrika

Типы источников данных



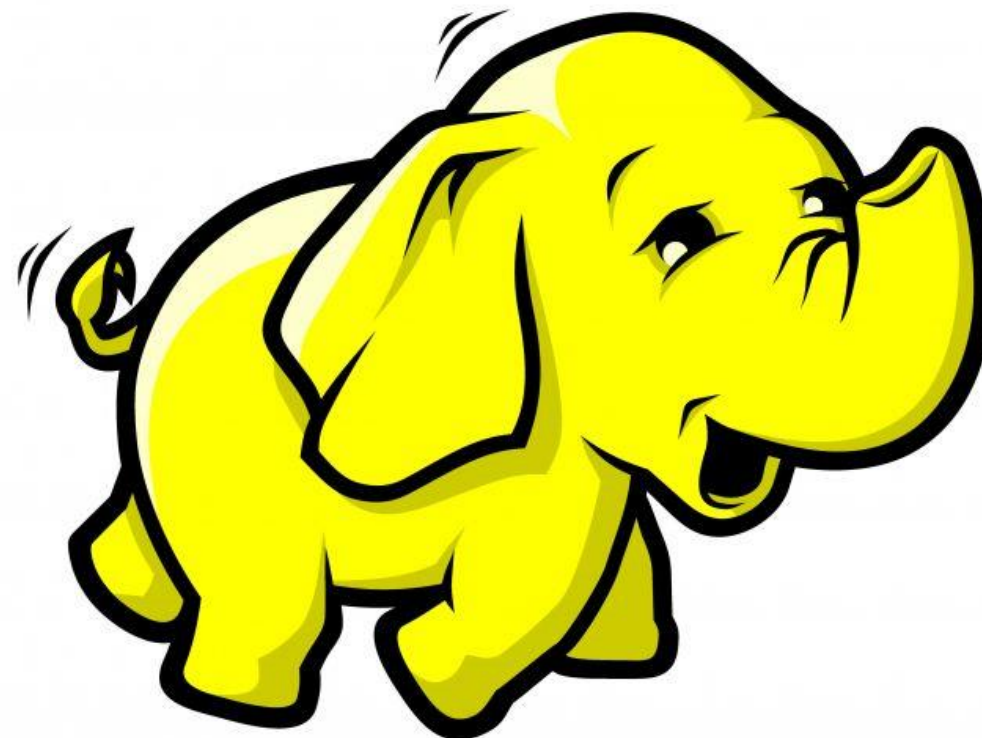
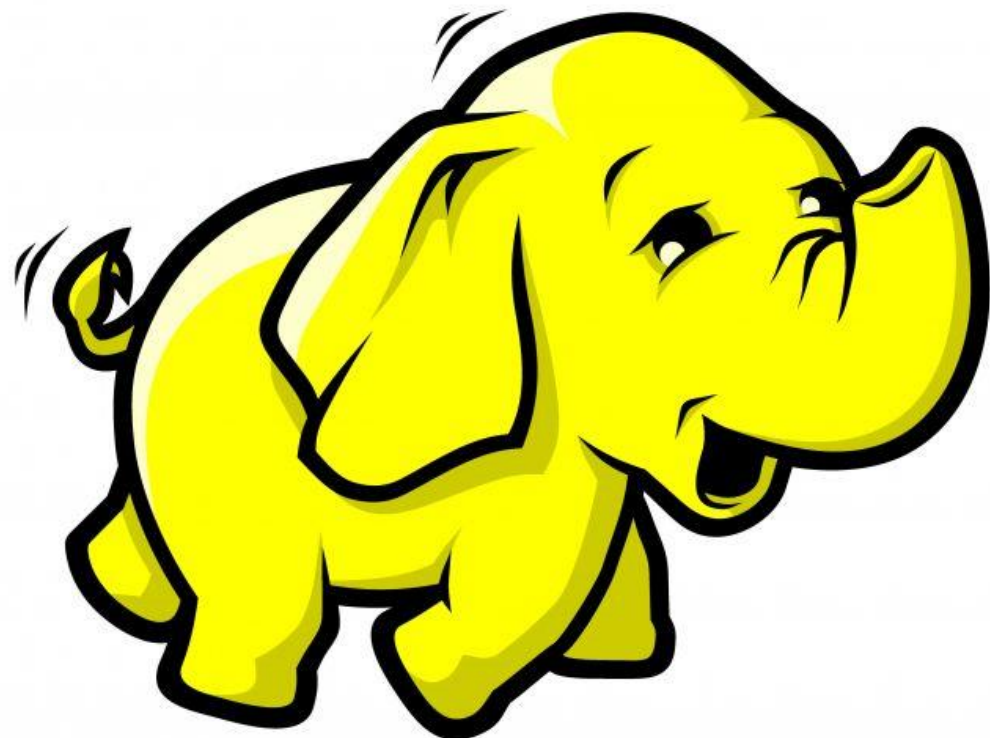
- **SQL**
PostgreSQL/MySQL/MSSQL/Oracle
- **NoSQL**
MongoDB/Kafka/Greenplum
- **Object storage**
S3
- **Block storage**
HDFS
- **WWW**
Yandex metrica

**Уменьшение
накладных
расходов**

Надоор в Газпромбанк



2 x 1 ПБайт

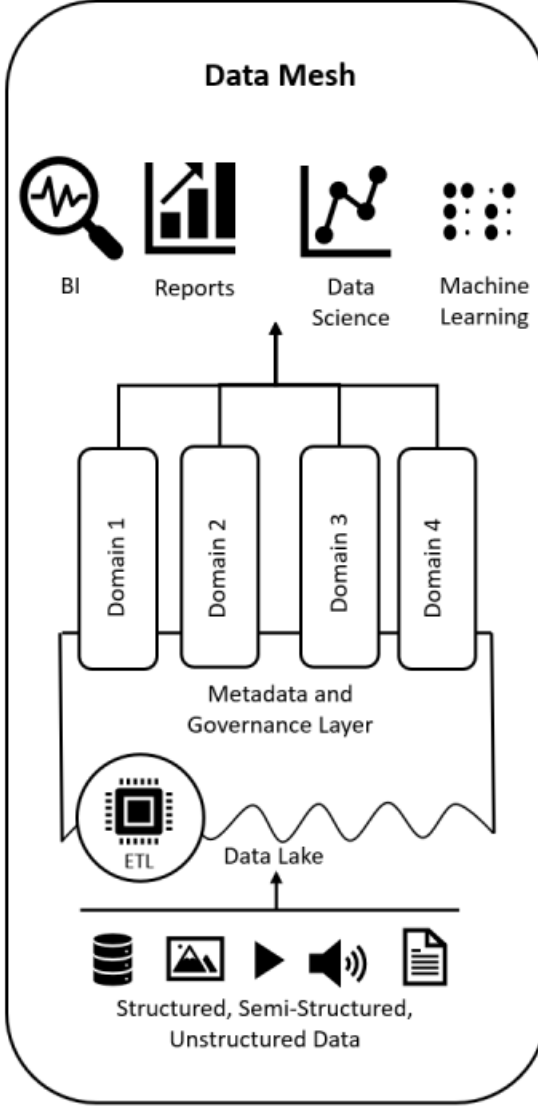
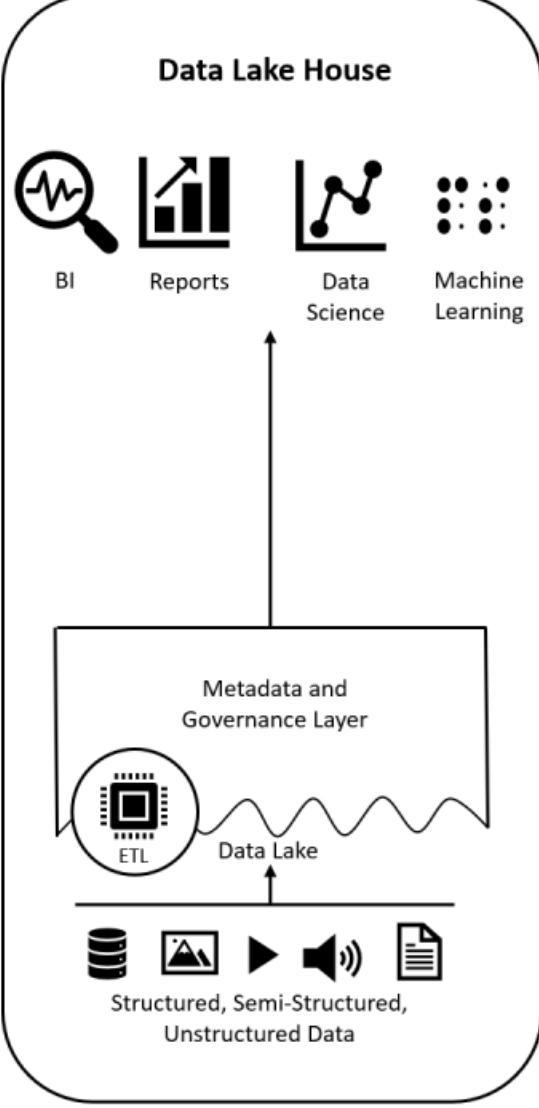
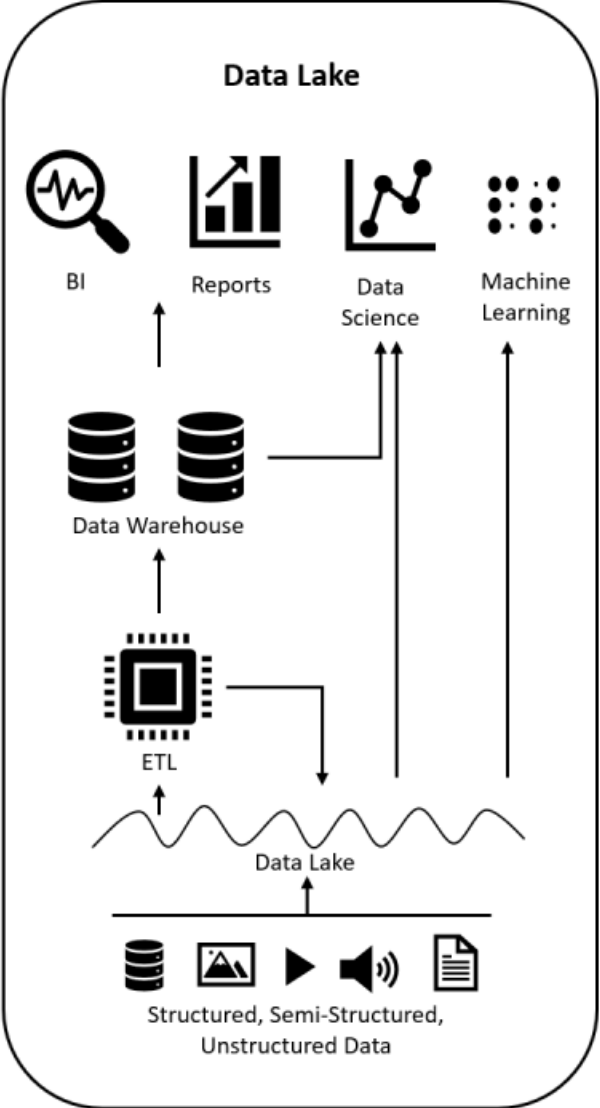
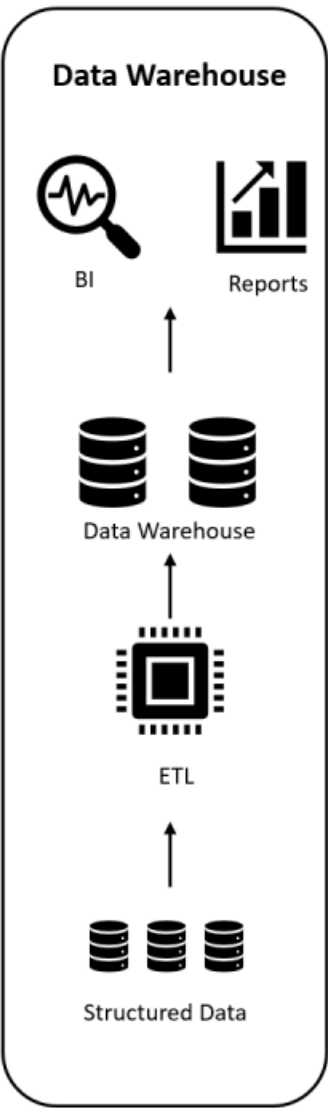


DevOps составляющая MLOps

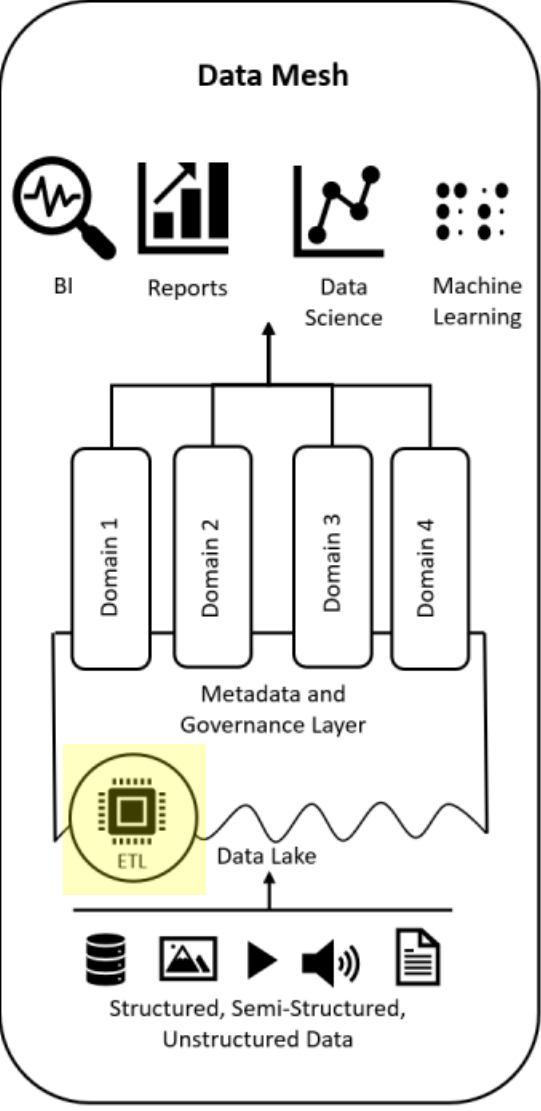
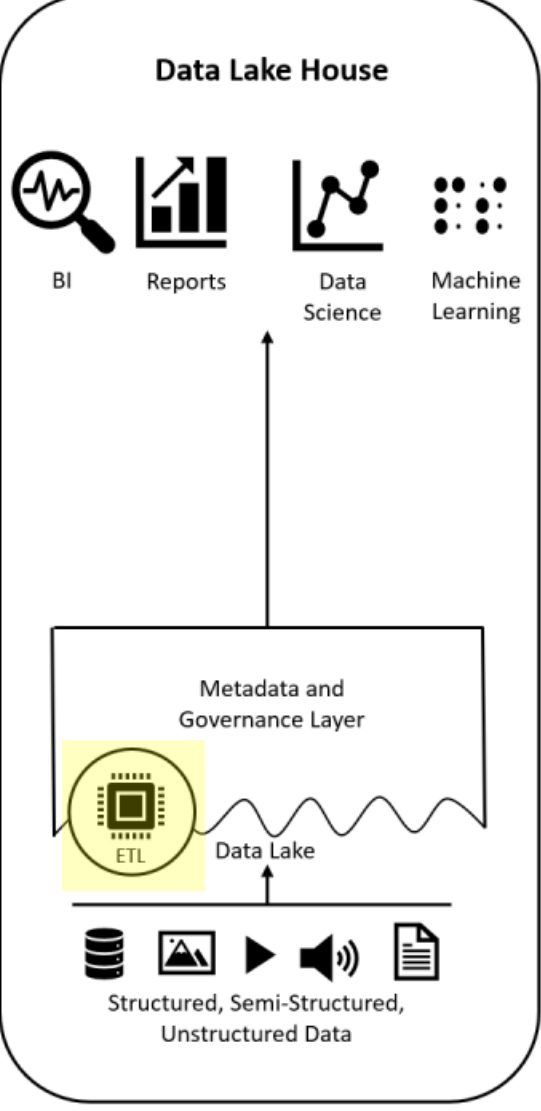
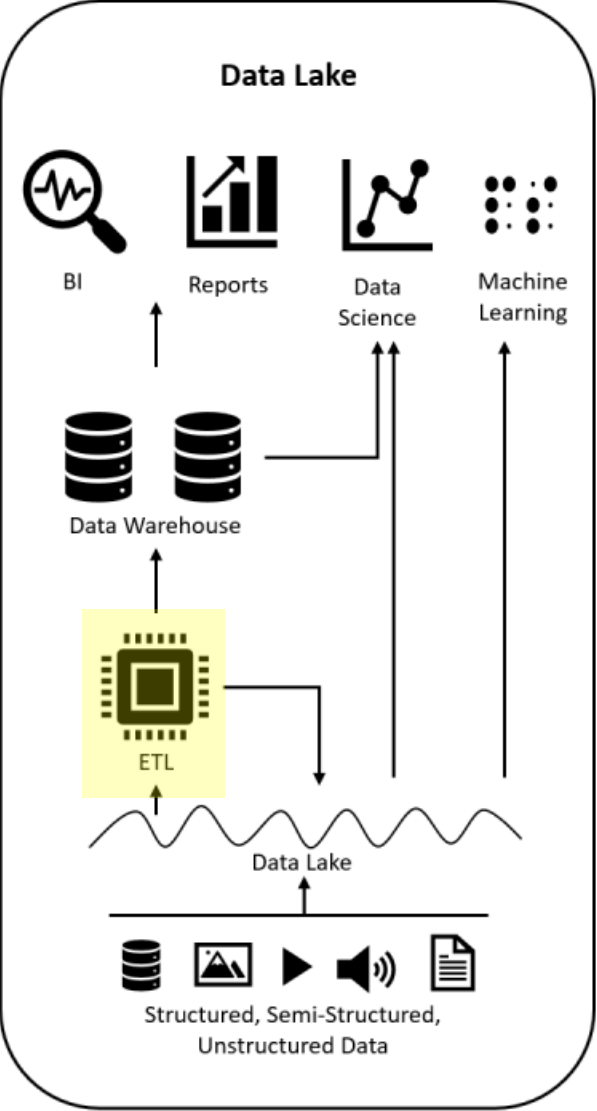
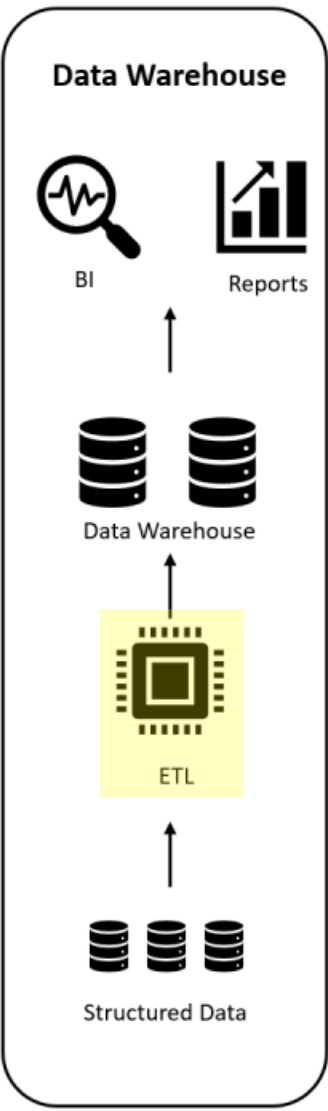


- Источники данных
- Подготовка данных / ETL процессы
- Разработка
- Трекинг эксперимента, сохранение весов.
- Тестирование и доставка моделей
- Мониторинг

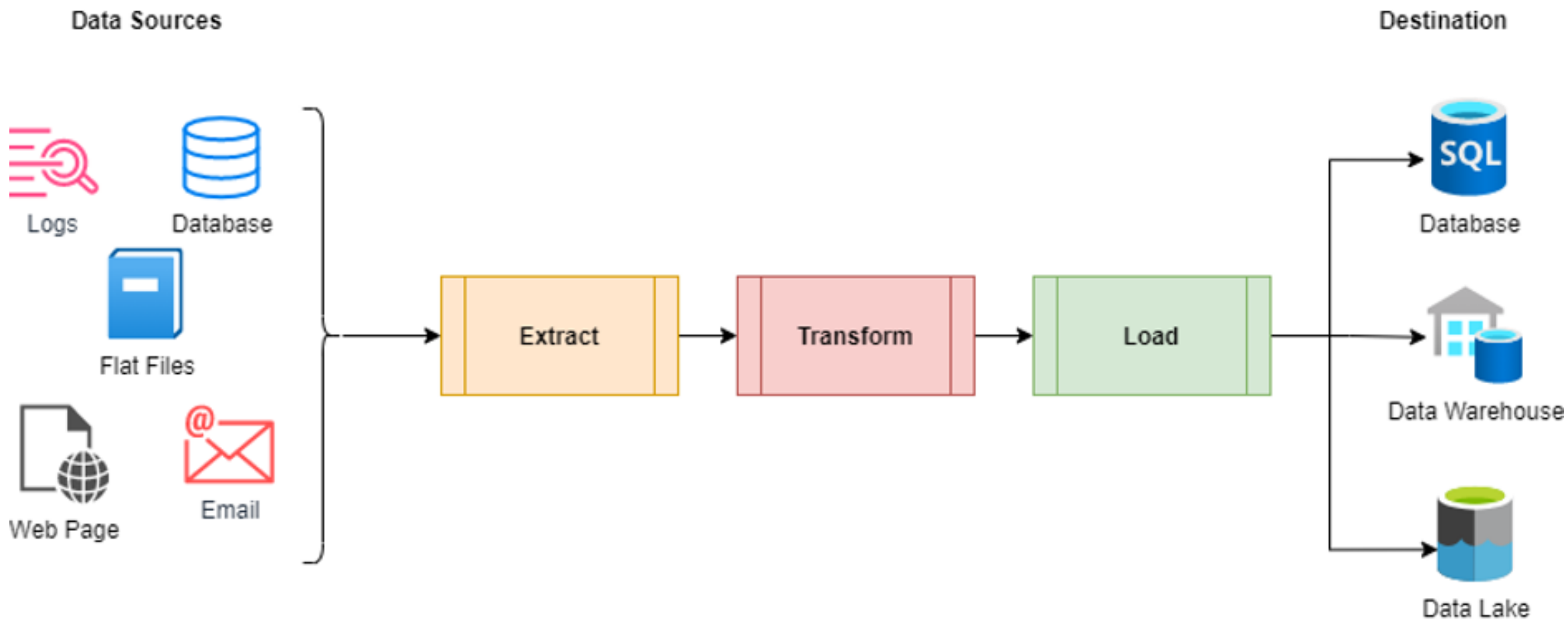
Подходы к организации источников данных



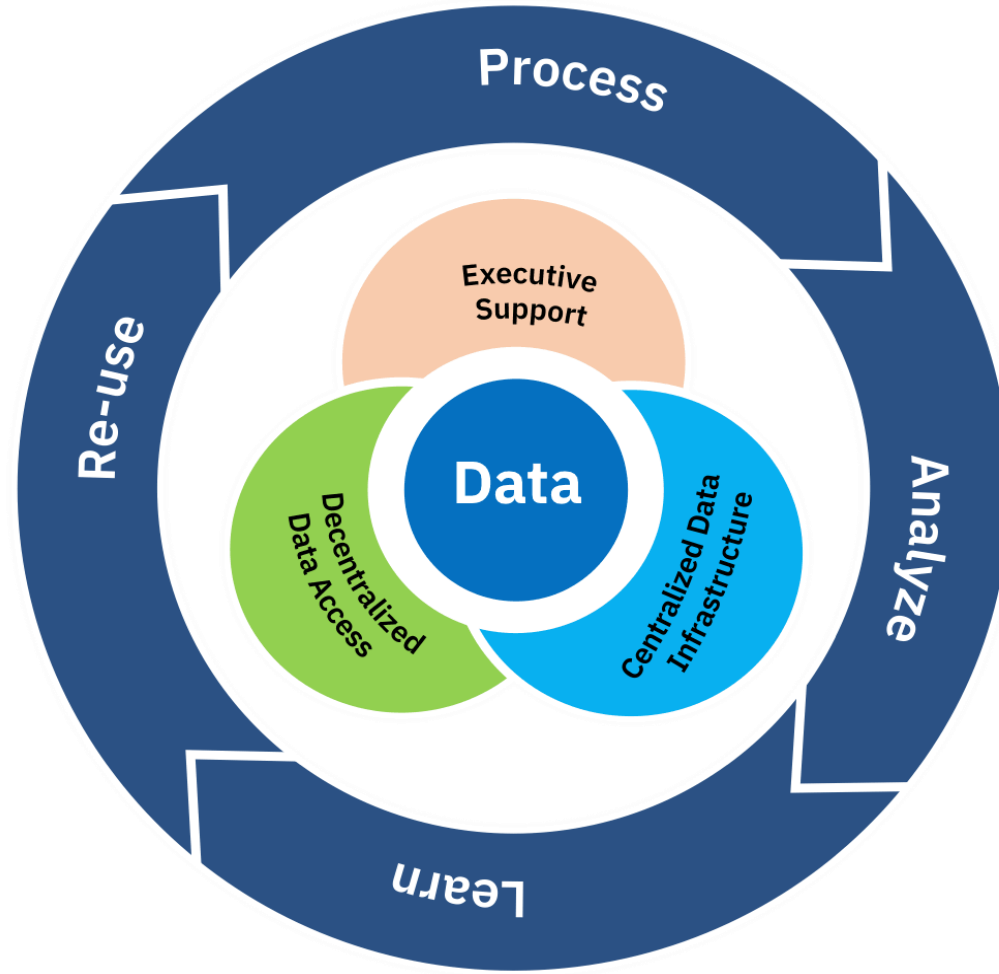
Подходы к организации источников данных



ETL процессы



DataOps





SAS Data Integration Studio 4.901 - GCI Linux - gciadmin

File Edit View Check Outs Actions Debug Tools Window Help

New [Icons] Workspace Server

Folders Inventory Transformations

- Model Manager
- SAS BI Report Services
- SAS Business Rules Manager
- SAS Customer Experience Analytics
- SAS Customer Experience Targeting
- CXT ECV STP
 - CXT ETL
 - Data Source
 - Jobs
 - ECV Jobs
 - ETL Jobs
 - Deployed Jobs
 - cxt_1000_pre_extract
 - cxt_1001_extract_attrb
 - cxt_1001_extract_basket
 - cxt_1001_extract_basket_action
 - cxt_1001_extract_basket_item
 - cxt_1001_extract_custom_attributes
 - cxt_1001_extract_cxt_errors
 - cxt_1001_extract_documents
 - cxt_1001_extract_form_path
 - cxt_1001_extract_goal
 - cxt_1001_extract_int_prch
 - cxt_1001_extract_media
 - cxt_1001_extract_media_hr
 - cxt_1001_extract_new_sessions
 - cxt_1001_extract_page_errors
 - cxt_1001_extract_product_view
 - cxt_1001_extract_promotions

ct_1000_pre_extract

Up Run Stop [Icons]

Diagram Code Log Output

Details Mappings Status Warnings and Errors Statistics Control Flow

Source table: All Tables

#	Column	Column Description	Table	Table Description	Type	Length
1	Process_nm		ETL_SIGNAL (ETL...	CXTCNFG	Character	20 (Nk
2	process_...	PROCESS_DTTM	ETL_SIGNAL (ETL...	CXTCNFG	Numeric	8 DA
3	status		ETL_SIGNAL (ETL...	CXTCNFG	Character	20 (Nk
4	NORM_ID		NORMALIZATION...	CXTCNFG	Numeric	8 (Nk

Basic Properties

Name	Value
Name	cxt_pre_extract
Description	
Node Type	cxt_pre_extract
Transform Version	86
User Written	No
Diagnostic Mode	No
Checkpoint	No
Run always when restarting	No
Control Order	1

Apache Nifi



The screenshot displays the Apache NiFi web interface for a process group named "Dati GTFS (PTR-01-A)". The interface includes a top navigation bar with the NiFi logo, a toolbar with various icons, and a status bar showing system metrics like "359" tasks and "2" alerts. A search bar is located in the top right corner.

The main workspace shows a data flow graph with several components:

- Create tables with no PK:** A process that creates tables in a database.
- Fetch file:** A process that retrieves data from a file system.
- RouteToProcessors:** A routing process that directs data to different downstream processors based on attributes.
- WRITE_AGENCY, WRITE_STOPS, WRITE_STOP_TIMES:** Database write processes that store data into specific tables.
- LogAttribute:** A process that logs specific attributes of the data.
- Copy to GTFS tables with PK:** A process that copies data into tables that have primary keys.
- Drop tables without PK:** A process that drops tables that do not have primary keys.

Each component in the flow is represented by a card showing its name, version, and current status (e.g., "In: 0 (0 bytes)", "Out: 0 (0 bytes)", "Tasks/Time: 0 / 00:00:00.000"). The flow is connected by lines representing data paths, with some paths labeled with names like "Name success" or "Name routes".

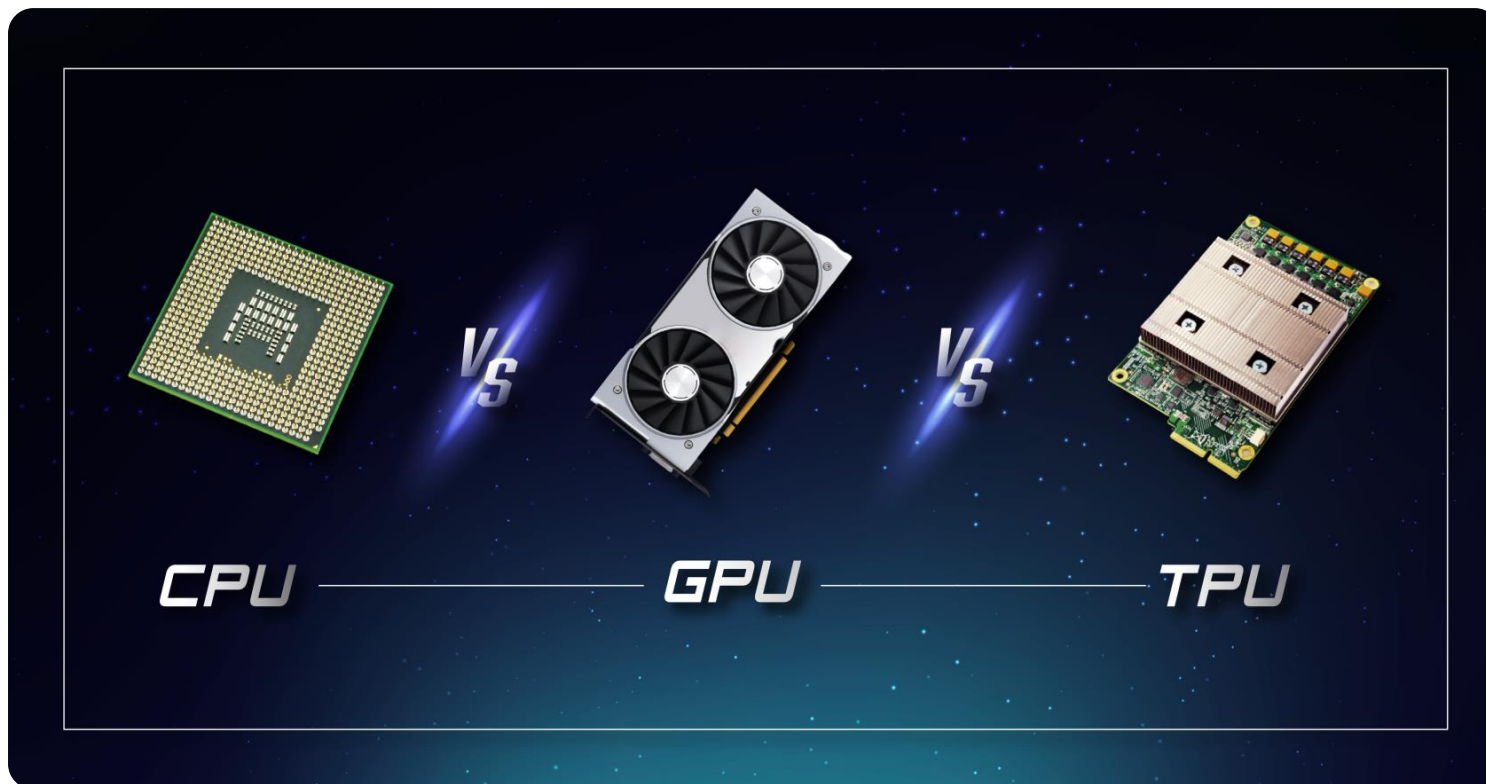
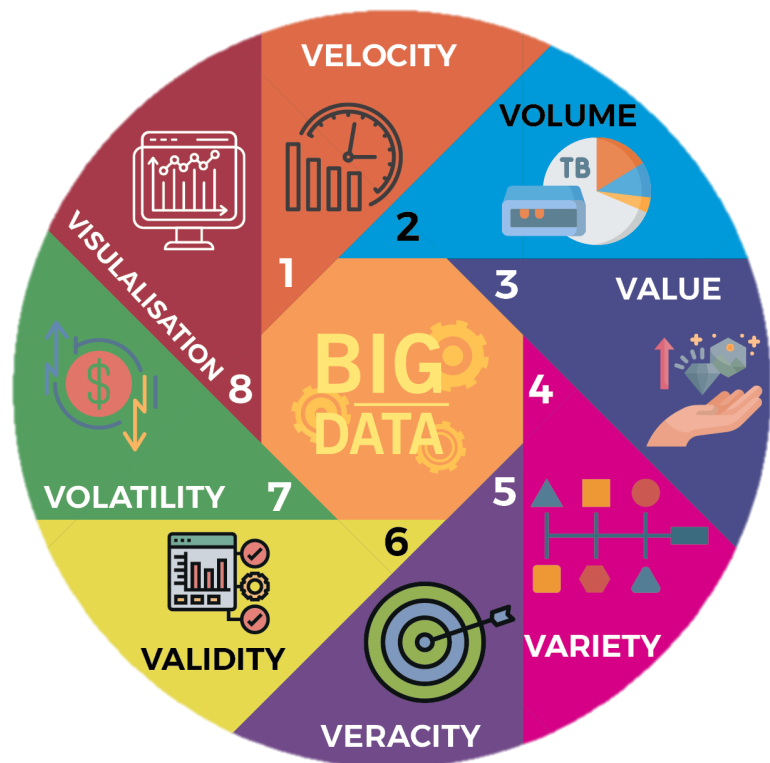
On the left side, there is a "Navigate" panel with a search and zoom function, and an "Operate" panel with various control icons (stop, play, refresh, etc.) and a "DELETE" button. The bottom of the interface shows the breadcrumb path: "NIFI Flow » SMASH » Ferrara2 » Dati GTFS (PTR-01-A)".

MLOps с точки зрения DevOps



- Источники данных
- Подготовка данных / ETL процессы
- **Разработка**
- Трекинг эксперимента, сохранение весов.
- Тестирование и доставка моделей
- Мониторинг

Разработка в MLOps



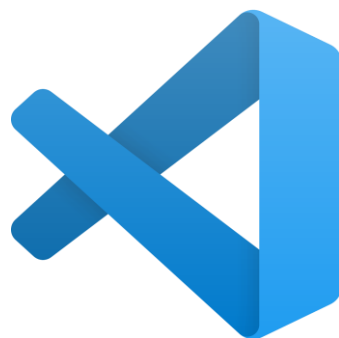
Jupyter notebooks



The image displays a collage of Jupyter notebook interfaces. The top-left window shows a notebook titled "In Depth: Linear Regression" with text explaining regression models. Below it, a code cell shows a simple linear regression model: `m = np.polyfit(x, y, 2); plt.scatter(x, y)`. The middle window shows a "Launcher" interface with buttons for Python 3, C++11, C++14, C++17, Julia 1.10, phylogenetics (Python 3.7), and R. The top-right window shows a "Seattle Weather: 2012-2015" plot with a scatter plot of maximum daily temperature and a bar chart of the number of records. The bottom row shows three smaller notebook windows: "Julia" with a scatter plot of Sepal.Length vs Petal.Length, "python notebook" with the Lorenz system equations $\dot{x} = \sigma(y - x)$, $\dot{y} = \rho x - y - xz$, and $\dot{z} = -\beta z + xy$, and "R" with a scatter plot of Sepal.Length vs Petal.Length.

Jupyter notebooks с точки зрения MLOps

- Плохо для прома (PyCharm, Visual Studio Code)
- Хорошо для исследований
- Разные контейнеры для разработки и применения
- Отдельные пайплайны для разных этапов



MLOps с точки зрения DevOps



- Источники данных
- Подготовка данных / ETL процессы
- Разработка
- Трекинг эксперимента, сохранение весов.
- Тестирование и доставка моделей
- Мониторинг

Трекинг эксперимента



git



git

DVC



```
$ dvc add cats-dogs

$ dvc remote add storage s3://bucket/dvc-cache

$ dvc push
5000 files pushed
```

```
$ dvc exp run
"data/data.xml.dvc" didn't change, skipping
Stage "prepare" didn't change, skipping
Stage "featurize" didn't change, skipping
Running stage "train":
> python src/train.py data/features model.pkl
|
```

```
$ dvc exp show
```

Experiment	roc_auc.train	roc_auc.test	train.min_split
workspace	0.96464	0.93989	0.04
main	0.98667	0.94602	0.01
├─ 0417ff0 [fuggy-zack]	0.98667	0.94602	0.01
├─ cb5c648 [fishy-mene]	0.97652	0.94309	0.02
├─ c0dca21 [atrip-teds]	0.96957	0.9407	0.03
├─ 3d2c9c7 [brisk-aged]	0.96464	0.93989	0.04
└─ 6e910c4 [chirk-main]	0.96193	0.93979	0.05



git

DVC

nexus | S

MLOps с точки зрения DevOps



- Источники данных
- Подготовка данных / ETL процессы
- Разработка
- Трекинг эксперимента, сохранение весов.
- **Тестирование и доставка моделей**
- Мониторинг

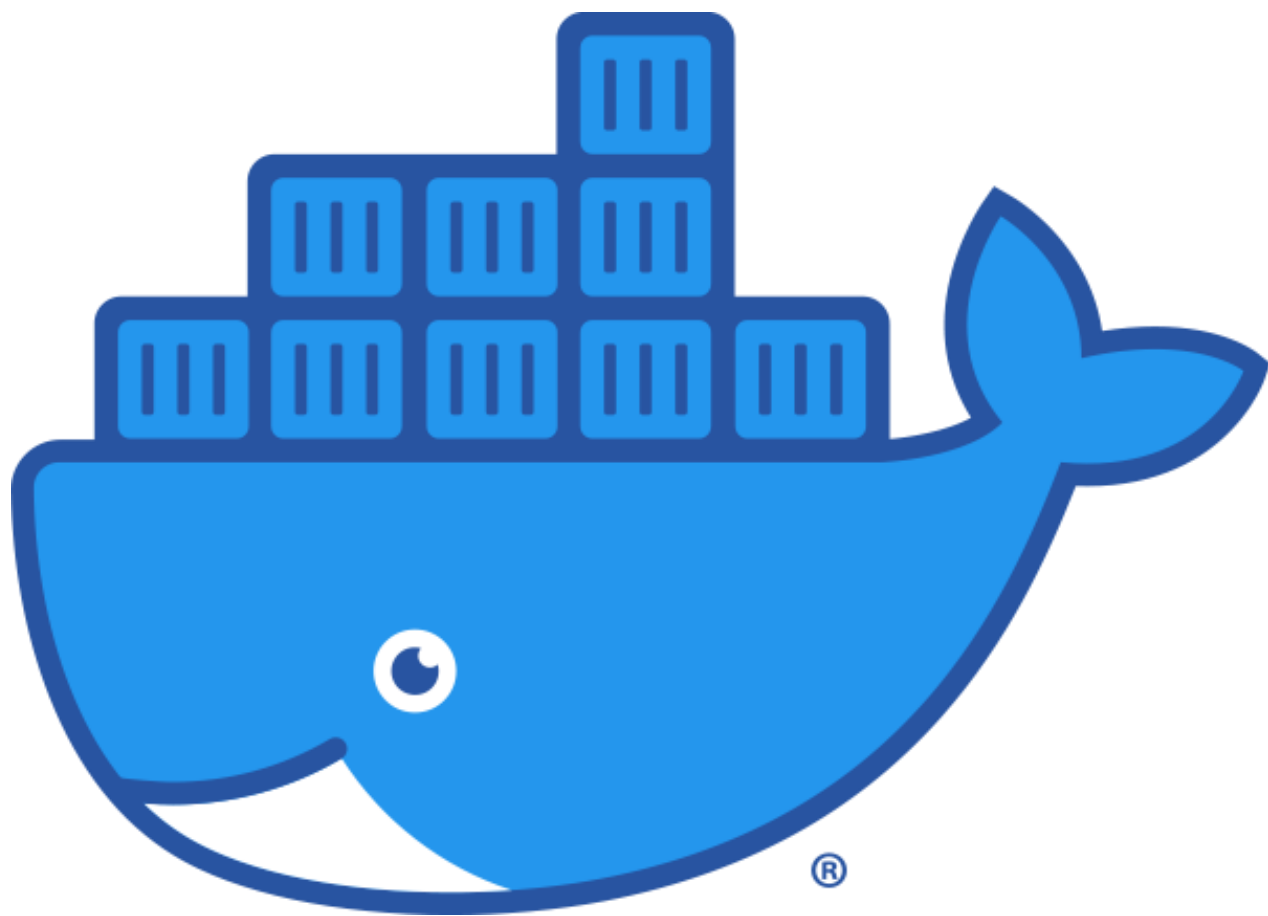
Тестирование моделей

Сделать

Тестирование ML модели должно быть сопряжено с тестированием самих данных, на которых построена и будет применяться модель.

- Обеспечить наличие датасета на всех средах.
Гарантировать воспроизводимость.

Доставка моделей - контейнеры для MLOps



MLOps с точки зрения DevOps

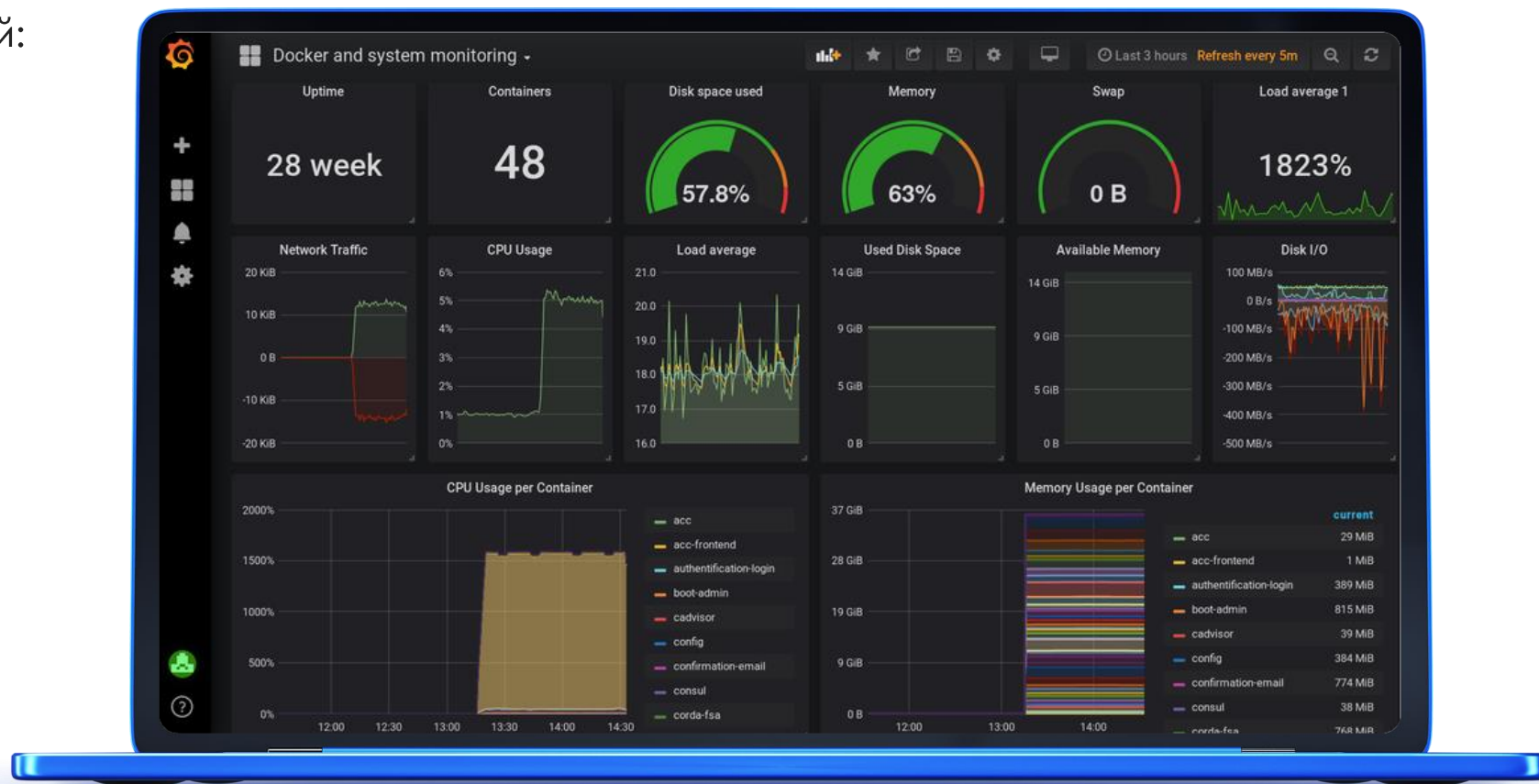


- Источники данных
- Подготовка данных / ETL процессы
- Разработка
- Трекинг эксперимента, сохранение весов.
- Тестирование и доставка моделей
- **Мониторинг**

Мониторинг ML моделей

Уровни мониторинга ml моделей:

- ИТ мониторинг (Prometheus)
- ML мониторинг (Gini, ROC AUC, NDCG, Lift)
- Бизнес мониторинг (А/В тесты, эффекты)



Мониторинг ML моделей

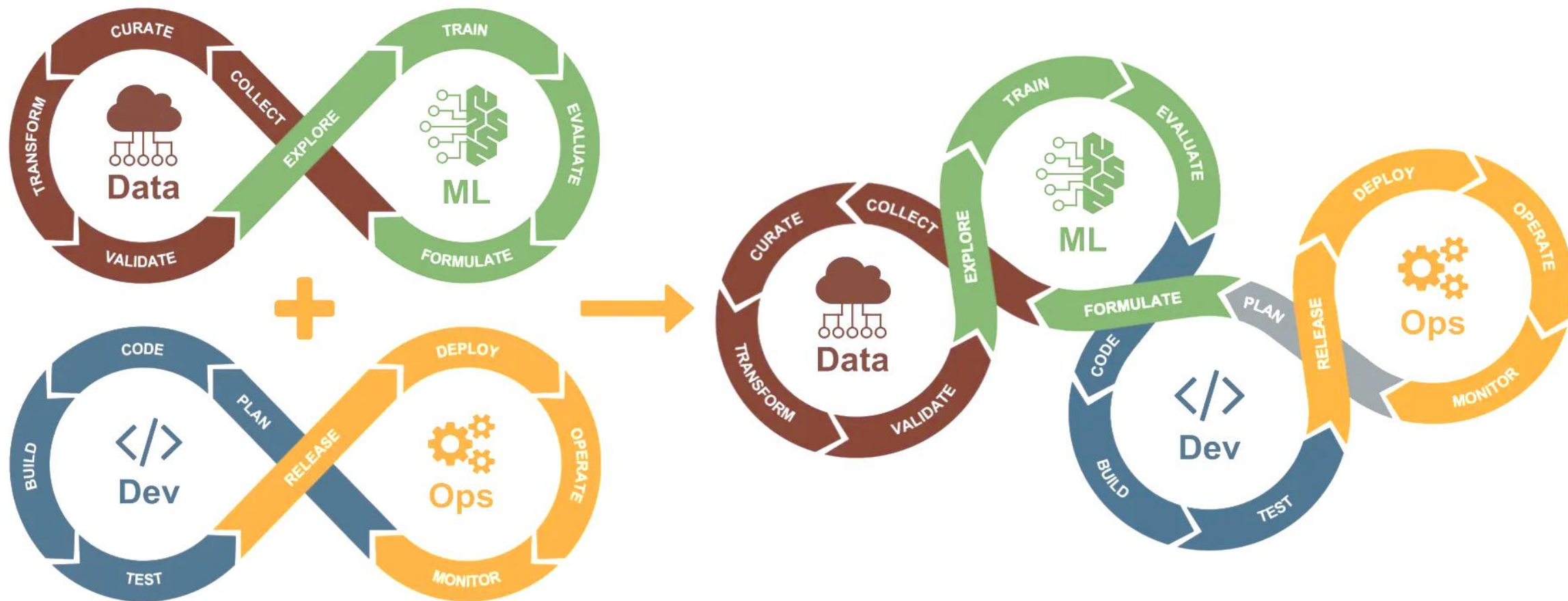
Batch

Online

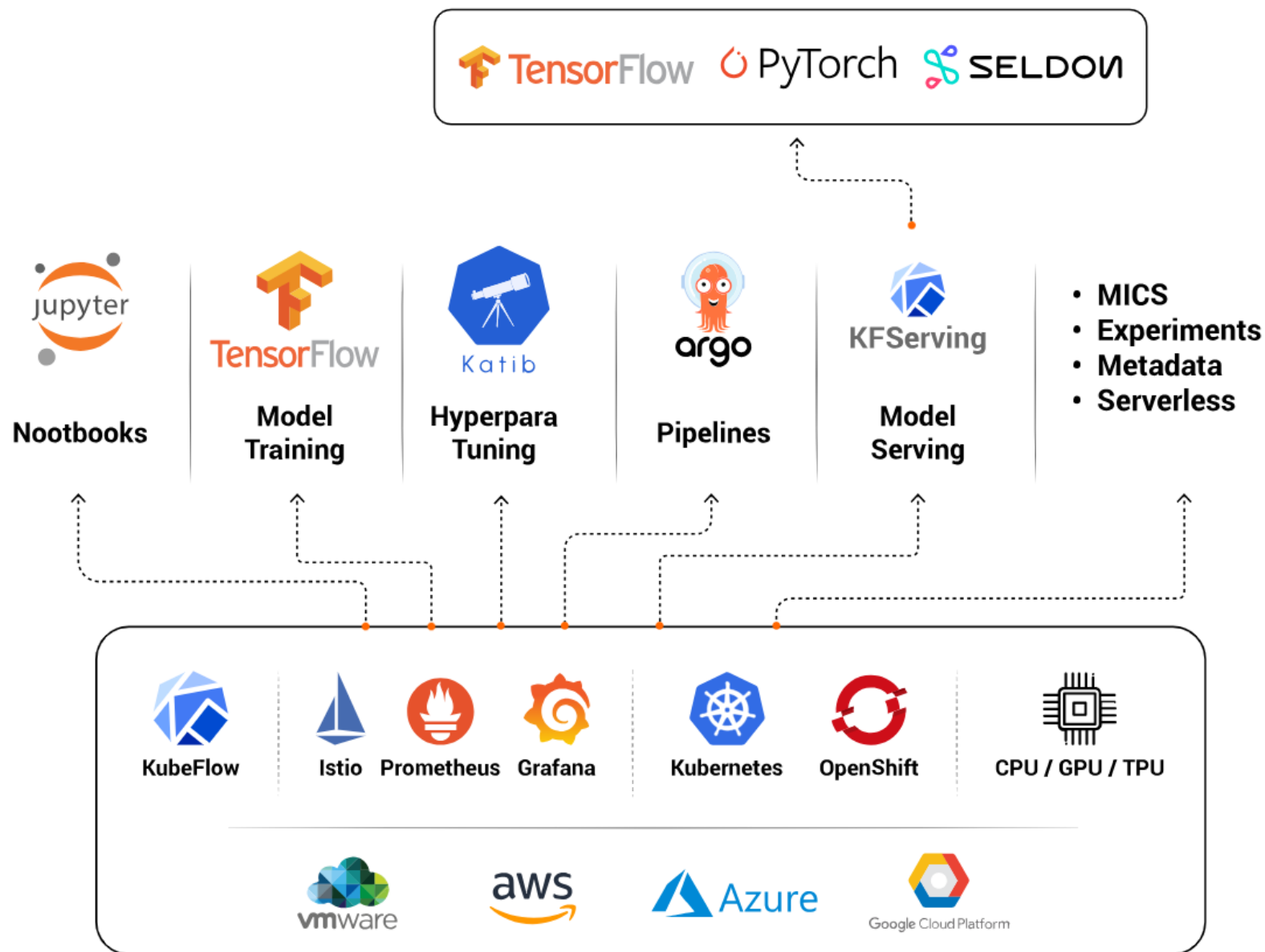
Это MLOps

MLOps = DataML + DevOps

ml4devs.com/mlops-lifecycle 



Единая система для MLOps?





Kubeflow

Home

Notebooks

Tensorboards

Models

Volumes

Experiments (AutoML)

Experiments (KFP)

Pipelines

Runs

Recurring Runs

Artifacts

Privacy • Usage Reporting

build version dev_local

kubeflow-user (Owner)



Dashboard

Activity

Quick shortcuts

- Upload a pipeline**
Pipelines
- View all pipeline runs**
Pipelines
- Create a new Notebook server**
Notebook Servers
- View Katib Experiments**
Katib

Recent Notebooks

- kale.log**
Accessed 10/12/2021, 2:06:43 PM
- lost+found**
Accessed 10/12/2021, 2:06:00 PM

Recent Pipelines

- open-vaccine-model**
Created 5/6/2021, 12:32:25 PM
- [Tutorial] DSL - Control structures**
Created 5/6/2021, 1:42:51 AM
- [Tutorial] Data passing in python components**
Created 5/6/2021, 1:42:49 AM
- [Demo] TFX - Taxi tip prediction model trainer**
Created 5/6/2021, 1:42:48 AM
- [Demo] XGBoost - Iterative model training**

Documentation

Getting Started with Kubeflow

Get your machine-learning workflow up and running on Kubeflow



MiniKF

A fast and easy way to deploy Kubeflow locally



Microk8s for Kubeflow

Quickly get Kubeflow running locally on native hypervisors



Minikube for Kubeflow

Quickly get Kubeflow running locally



Kubeflow on GCP

Running Kubeflow on Kubernetes Engine and Google Cloud Platform



Kubeflow on AWS

Running Kubeflow on Elastic Container Service and Amazon Web Services



Requirements for Kubeflow

Get more detailed information about using Kubeflow and its components



Новая платформа данных



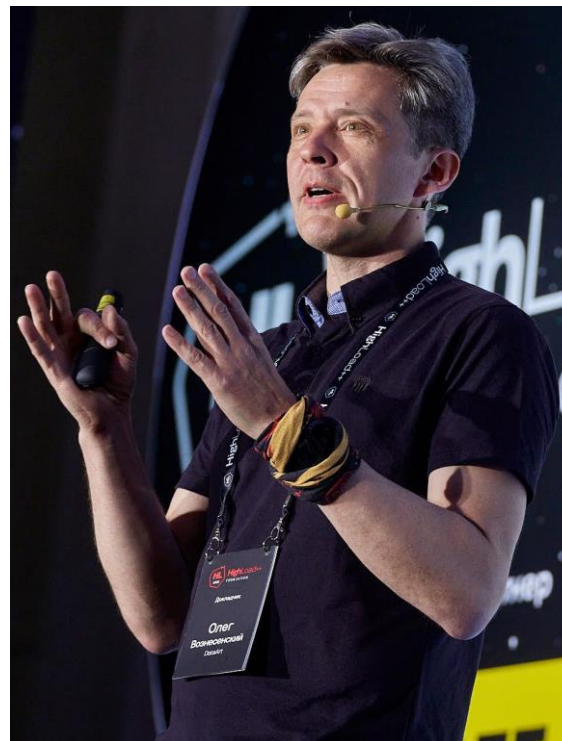
Platform engineering

Денис Занков



podsyb

Олег Вознесенский



seasadm

