

neu.ro

Gartner

COOL
VENDOR
2019

In AI Core Tech

**“Зелёные” DGX, виртуальные GPU,
оркестрация, интеграция и
взаимозаменяемость, – или как мы строим
MLOps платформу будущего**

Мария Давыдова, Head of Product @ Neu.ro, COO @ AIIA

30 октября 2021 года

neu.ro

Обо мне

- Разработчик > 15 лет
- Инженер MLOps > 2 лет
- Участник конференций и митапов > 15 лет
- Community manager > 10 лет
- Блоггер > 10 лет

*Разработчикам ПО и инженерам МашО
необходимы качественные инструменты для
того, чтобы делать качественные продукты*

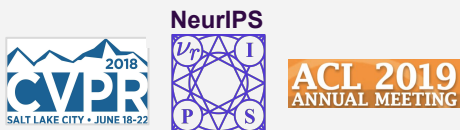


Мария Давыдова, aka NeuroMasha

0 Neu.ro

Мы - признанная команда специалистов по ML и разработке ПО, нацеленная решать проблемы MLOps и поддерживать ответственный подход к ИИ

Компания



Competition Winners



170+

Papers in ML/DL

2019

Beta Launch

Партнерства



Опыт

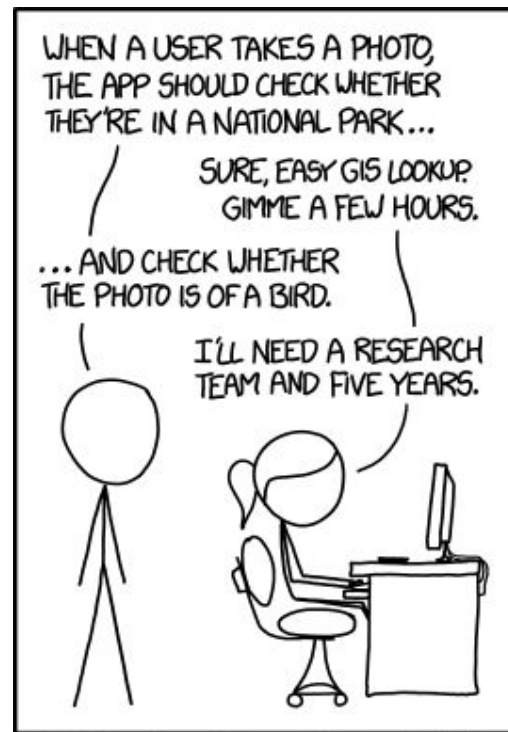


О чём речь?

Машинное обучение != Разработка ПО

Некоторые задачи MLOps

- **Масштабирование API**
 - Классический DevOps
- **Автоматизация выкладки моделей**
 - DevOps + тестирование моделей
- **Отслеживание экспериментов**
 - Множество переменных и выходов данных
- **Мониторинг качества прогнозов моделей**
 - Как отслеживать? Как реагировать?
- **Дрейф данных (data drift)**
 - Данные пользователей со временем меняются
- **Управление обновлением данных**
 - Отслеживание изменения бинарных файлов



Компоненты системы MLOps

Hardware

- Высокопроизводительные GPU
- Хранилища
- CPU, память, сеть

Software

- Решение ряда разнородных задач, от сбора и версионирования данных до деплоя и мониторинга API
- Средства автоматизации процессов

Wetware

- Data architects, data engineers, data scientists, software developers, devops engineers, mlops engineers, ...

Почему ML != SD?

Критерий	Традиционная разработка ПО	Модели машинного обучения
Статус проекта	Версия кода	Версия кода + Версия данных
Время сборки	Несколько минут	Несколько часов/дней
Стоимость сборки	Несколько центов/долларов	Несколько сотен/тысяч долларов
Поведение	Детерминированное	Недетерминированное
Производительность со временем	Стабильная	Обычно падает
Мониторинг	Системные показатели	Системные показатели + качество модели
Ответственность	Разработчики	Распределена между командами

Требования к идеальной системе MLOps

01	Воспроизводимость Reproducibility	<ul style="list-style-type: none">• Повторяемость экспериментов через версионирование ассетов и метаданных
02	Масштабирование Scalability	<ul style="list-style-type: none">• Получение необходимых ресурсов как в облаке, так и в гибридной инфраструктуре
03	Непрерывность Continuity	<ul style="list-style-type: none">• Автоматизация• CI/CD/CT
04	Коллаборация Collaboration	<ul style="list-style-type: none">• Совместная работа разных команд в одном окружении
05	Ответственность Accountability	<ul style="list-style-type: none">• Управление правами доступа• Логи аудита
06	Этичность Ethics	<ul style="list-style-type: none">• Мониторинг и анализ работы моделей• Объяснение результатов (explainability)
07	Перспектива Future-proof	<ul style="list-style-type: none">• Способность к развитию системы• Избежание vendor lock-in

И как этого добиться?

Ингредиент 1. Надежная инфраструктура

- **Облака**
 - AWS
 - Azure
 - GCP
 - Другие провайдеры с VMware либо managed K8s
- **On-premise**
- **DGX**
 - 8 x A100 - это масштабируемость от **56 x 1/7 A100 (MIG)** для инференса до **8 x A100 с 320 GB RAM** для действительно больших моделей, в одном корпусе
 - DGX можно также объединять в суперкомпьютеры
 - У Neu.ro есть кластер DGX в Исландии на **зелёной энергии**

Ингредиент 2. K8s + Tools

- **Базовая оркестрация ресурсов и контейнеров**
 - Все выполняемые на платформе задания (jobs) контейнеризованы
- **Ingress out-of-the-box**
 - У любого контейнера предсказуемый URL
- **Prometheus + Grafana для репортов**
- **Docker Registry для хранения контейнеров**
- **Хранилища**
 - NFS для быстрого сетевого хранилища
 - PVC для монтируемых дисков
 - Object storage для длительного хранения ассетов
- **Базы данных**

Ингредиент 3. Программный слой Neu.ro

- **Продвинутая оркестрация ресурсов**
 - Round robin
 - Preemptible nodes
 - Различные ресурсные квоты
 - Отчёты
- **Оркестрация прав**
 - Все элементы системы (задания-контейнеры, хранилище, образы Docker, роли, проекты) находятся под управлением системы прав доступа
 - Role-based access control & SSO
 - Секреты для доступа к third-party services
- **Оркестрация процессов**
 - Собственный движок пайплайнов (Neu.ro Flow) с YAML
 - Многоуровневые ациклические графы исполнения
 - Модульность

Ингредиент 4. Интерфейсы

The screenshot displays the neu.ro dashboard interface. At the top left is the 'neu.ro' logo. The top right corner contains links for 'Docs', 'Log Out', and a flag icon. A left sidebar lists navigation items: Dashboard, Jobs, Projects, Secrets, Information, and Cluster management. Below the sidebar, a 'Cluster:' dropdown menu is set to 'neuro-compute'. The main content area is titled 'Dashboard' and includes a 'Credits: unlimited' indicator. Under the 'Actions' heading, there are four panels: 'Terminal' with a 'RUN A JOB' button; 'Storage' with 'RUN A JOB' and 'OPEN RUNNING' buttons; 'IDEs' with a 'RUN A JOB' button; and 'Docker registry' with 'RUN A JOB' and 'OPEN RUNNING' buttons. On the right, a 'Running jobs: 2' section lists two jobs: 'registry-browser-6acec' (started 14 hours ago) and 'filebrowser-d86c9' (started 19 hours ago), each with a 'KILL' button. A 'KILL ALL' button is also present. At the bottom, a note provides instructions on installing the CLI and contacting support at team@neu.ro.

neu.ro

Docs Log Out

Dashboard

Credits: unlimited

Actions

Terminal
RUN A JOB

Storage
RUN A JOB OPEN RUNNING

IDEs
RUN A JOB

Docker registry
RUN A JOB OPEN RUNNING

Running jobs: 2 KILL ALL

registry-browser-6acec
about 14 hours ago Job Details • HTTP URL KILL

filebrowser-d86c9
about 19 hours ago Job Details • HTTP URL KILL

To install CLI and start using Neuro with it please visit our [Getting started](#) guide. Use buttons above to start jobs without CLI.
If you have any issues, please contact us at team@neu.ro

Ингредиент 4. Интерфейсы

run

Run a job with predefined resources...

Usage

```
neuro job run [OPTIONS] IMAGE [CMD]...
```

Run a job with predefined resources configuration.

`IMAGE` container image name.

`CMD` list will be passed as commands to model container.

Examples

```
1 # Starts a container pytorch:latest on a machine with smaller GPU resources
2 # (see exact values in 'neuro config show') and with two volumes mounted:
3 #   storage://<home-directory> --> /var/storage/home (in read-write mode),
4 #   storage://neuomation/public --> /var/storage/neuomation (in read-only mode).
5 $ neuro run --preset=cpu-small --volume=HOME pytorch:latest
6
7 # Starts a container using the custom image my-ubuntu:latest stored in neuomation
8 # registry, run /script.sh and pass arg1 and arg2 as its arguments:
9 $ neuro run -s cpu-small image:my-ubuntu:latest --entrypoint=/script.sh arg1 arg2
```

Options

Name	Description
<code>-s</code> ,	
<code>--preset</code>	Predefined resource configuration (to see available values, run
<code>PRESET</code>	<code>neuro config show</code>)

```
async-for monitor(id: str) → AsyncIterator[bytes] [source]
```

Get job logs as a sequence of data chunks, e.g.:

```
async for chunk in client.jobs.monitor(job_id):
    print(chunk.encode('utf8', errors='replace'))
```

Parameters: `id (str)` – job id to retrieve logs.

Returns: `AsyncIterator` over `bytes` log chunks.

```
async-with port_forward(id: str, local_port: int, job_port: int, *,
no_key_check: bool = False) → None [source]
```

Forward local port to job, e.g.:

```
async with client.jobs.port_forward(job_id, 8080, 80):
    # port forwarding is available inside with-block
```

Parameters: • `id (str)` – job id.

• `local_port (int)` – local TCP port to forward.

• `job_port (int)` – remote TCP port in a job to forward.

```
coroutine resize(id: str, *, w: int, h: int) → None [source]
```

Resize existing TTY job.

Parameters: • `id (str)` – job id.

• `w (int)` – New screen width.

• `h (int)` – New screen height.

```
coroutine run(container: Container, *, name: Optional[str] = None, tags: Sequence[str] = (), description: Optional[str] = None, is_preemptible: bool = False,
schedule_timeout: Optional[float] = None, life_span: Optional[float] = None) → JobDescription [source]
```

Start a new job.

Parameters: • `container (Container)` – container description to start.

• `name (str)` – optional container name.

• `name` – optional job tags.

• `description (str)` – optional container description.

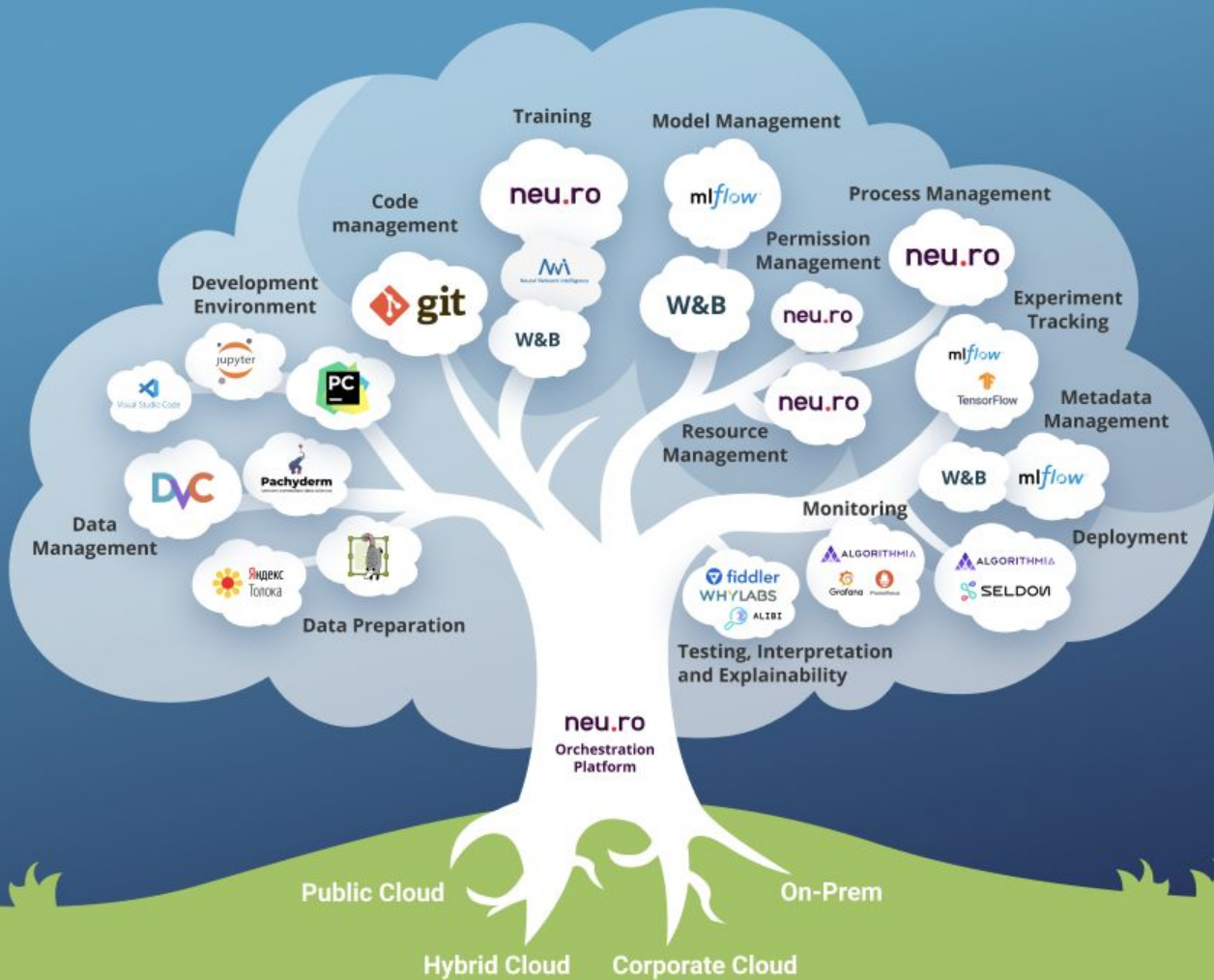
• `is_preemptible (bool)` – a flag that specifies is the job is *preemptible* or not, see [Preemption](#) for details.

• `schedule_timeout (float)` – minimal timeout to wait before reporting that job cannot be scheduled because the lack of computation cluster resources (memory, CPU/GPU etc).

• `life_span (float)` – job run-time limit in seconds. Pass *None* to disable.

Ингредиент 5. Инструменты

- **Библиотеки и фреймворки**
 - TensorFlow, Pytorch, etc
- **Пакеты (CLI)**
 - DVC, etc
- **Docker-native**
 - Jupyter, Label Studio, etc
- **K8s-native**
 - Pachyderm, Seldon, etc
- **SaaS**
 - Weights & Biases, etc
- **Desktop**
 - PyCharm, VS Code, etc



Ингредиент 6. Интеграции

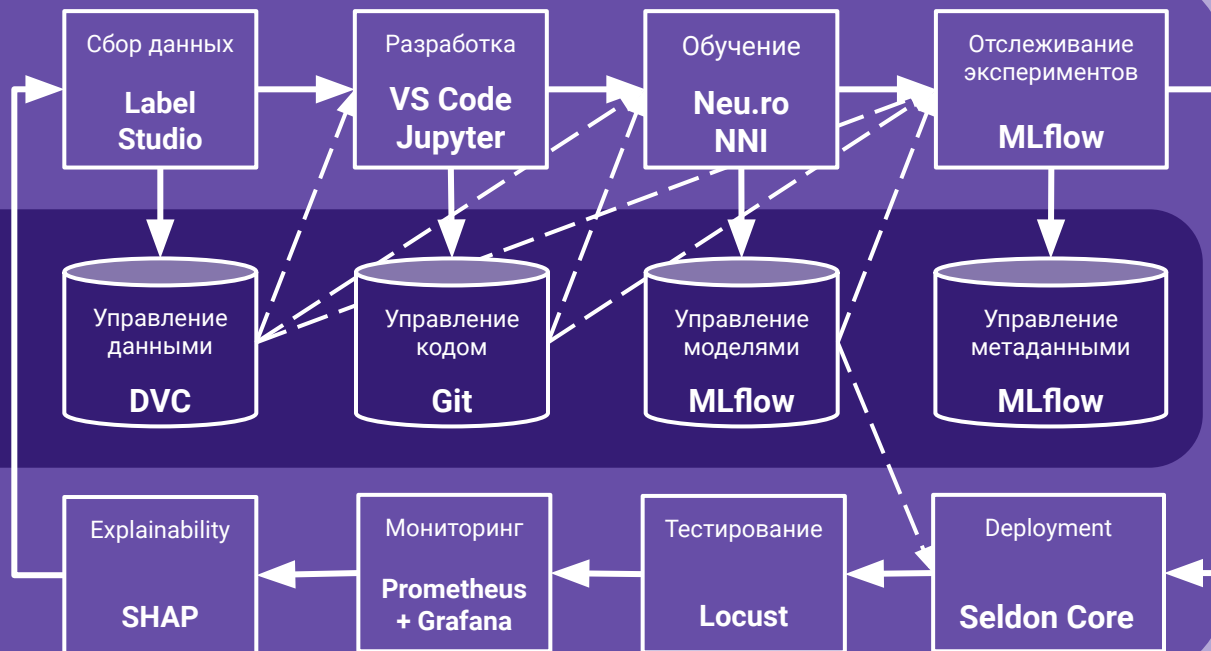
- **В веб-интерфейс**
 - Jupyter Notebooks, JupyterLab, FileBrowser, etc
- **В Neu.ro Flow actions**
 - MLflow, TensorBoard, etc
- **В документацию и рецепты**
 - См. <https://docs.neu.ro> и <https://github.com/neuro-inc>
- **Специализированные интеграции K8s-native инструментов**
 - mlflow-to-seldon - автоматизированный деплоймент моделей из хранилища в MLflow “на горячую” в Seldon

Пример стека: Open source

Управление разрешениями
Neu.ro

Управление процессами
Neu.ro

Управление артефактами
Neu.ro



Управление ресурсами
Neu.ro

Инфраструктура

Onprem
Hybrid
AWS
GCP
Azure
VMware
Other

Трудности

- **Контроль ресурсов для K8s-native инструментов**
- **SSO и единое управление правами на всём стеке**
- **Мифические security & governance**

Об AIIA (AI Infrastructure Alliance)

Формирование стандартов и канонического стека в ML разработке



- Основан в начале 2020
- Активное развитие с конца 2020
- Более 60 компаний

Determined AI

Determined AI lets data science teams train models faster using state-of-the-art distributed training. It also lets teams build better models with advanced hyperparameter tuning. It makes scaling training across distributed GPUs on-prem or in cloud straightforward and easy to learn from and reproduce their work back in experiment tracking.

ALGORITHMIA

For machine learning, business threat and fraud ML, model, into production faster, more securely, and effectively within their existing operational processes, Algorithmia's machine learning operations (MLOps) software helps teams at every stage of the ML lifecycle, unlike traditional, expensive, and resource-intensive MLOps management solutions that lock users into a single technology stack. Algorithmia automates ML deployment, streamlines collaboration between operations and development, streamlines DevOps and CI/CD processes, and provides advanced security and governance. Over 11000 engineers and data scientists have used Algorithmia's platform to date, including the United Nations, government intelligence agencies, and Fortune 500 companies. For more information, visit www.algorithmia.com.

Superb AI

Superb AI is a training data platform built by a team of experts and engineers. With over 70 datasets and over 1000 requests, the platform is designed to deliver the most accurate and comprehensive training data platform for all users on the frontier of artificial intelligence.

YData

YData is a data science platform that accelerates the access to valuable data and makes it easier to explore. YData is an AI Solutions with high-quality and synthetic data.

Tecton

Tecton provides an enterprise feature store that makes it easy to build, deploy, and share features for machine learning. Tecton automatically transforms raw data into feature tables, stores the feature, and serves from a unified training and online production. It offers data scientists to build and deploy features within hours instead of weeks. Tecton was founded by the creators of Uber ML Hub.

Pachyderm

Pachyderm delivers a cloud data science and data ingestion platform for K8s, that unifies the Git for data. It's a robust pipeline system using Kubernet and Docker to quickly build data transformation and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

INCIGHT

INCIGHT makes ML researchers and engineers more effective by giving them tools to manage their experiments, data, and pipeline development. The company's open source CloudML MLops platform automates and simplifies developing, managing and deploying machine learning systems by giving data scientists simple plug & play tools for experiment management, automated pipeline creation, operation (Features store, Data pipeline) and operations (Features store, Data pipeline) and operations (Features store, Data pipeline) and operations (Features store, Data pipeline).

TerminusDB

Terminus DB brings a distributed, revision control database to data science teams across the enterprise. The TerminusDB is an open source, full featured in-memory graph database. The platform is designed to help data scientists to explore their data. It allows for the versioning of data science models, features, graphs, queries, workflows, APIs, and more. It allows teams to manage data to databases and to collaboratively work on shared resources.

fiddler

Founded in October 2013, Fiddler mission is to enable business to build, deploy, and maintain trustworthy AI solutions. Fiddler's product portfolio includes a platform for machine learning operations and technical tools to monitor, evaluate and manage machine learning pipelines and customer. Fiddler works with pioneering Fortune 500 companies as well as emerging tech companies. For more information please visit www.fiddler.ai or follow us on Twitter @fiddlerai or follow us on LinkedIn.

superwise.ai

Superwise enables data science and operational teams to monitor and analyze the health of their AI-based systems. By automating platform monitoring in real-time, superwise helps data science teams to quickly identify and resolve any issues from their machine learning pipelines that are automatically retrained and available. The platform includes built-in dashboards, workflow, and compliance tools. The company's main client groups include AI Scientists, operational team, and compliance teams. The company's main client groups include AI Scientists, operational team, and compliance teams. The company's main client groups include AI Scientists, operational team, and compliance teams.

SELDON

Seldon accelerates the adoption of machine learning technologies to solve the most challenging problems with a diverse and growing set of applications. Seldon simplifies the process of testing, monitoring and deploying machine learning models. Seldon provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

CLEAR ML

Allegro AI makes ML researchers and engineers more effective by giving them tools to manage their experiments, data, and pipeline development. The company's open source CloudML MLops platform automates and simplifies developing, managing and deploying machine learning systems by giving data scientists simple plug & play tools for experiment management, automated pipeline creation, operation (Features store, Data pipeline) and operations (Features store, Data pipeline) and operations (Features store, Data pipeline).

neu.ro

Neuro.ai provides an end-to-end platform for AI solutions. The company's open source CloudML MLops platform automates and simplifies developing, managing and deploying machine learning systems by giving data scientists simple plug & play tools for experiment management, automated pipeline creation, operation (Features store, Data pipeline) and operations (Features store, Data pipeline) and operations (Features store, Data pipeline).

DAGHub

DAGHub is a data science platform for building experiments, monitoring ML models, and automating, simplifying the build process and making it easy to get up and running. DAGHub helps data scientists to explore their data. It allows for the versioning of data science models, features, graphs, queries, workflows, APIs, and more. It allows teams to manage data to databases and to collaboratively work on shared resources.

Valohai

Valohai is an end-to-end ML Ops platform for machine learning and deep learning. The platform provides everything you need to take your machine learning from data science to production. Valohai ML Ops platform is used by the platform is automatically retrained and available. The Valohai ML Ops platform is used by the platform is automatically retrained and available. The Valohai ML Ops platform is used by the platform is automatically retrained and available.

InfuseAI

InfuseAI is a cloud-native platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

gretel

Gretel automates privacy engineering, enabling companies to meet the strictest privacy laws like GDPR. The platform provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

Honeyjob

Honeyjob is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

acceldata

Accelerdata is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

activeroop

Activeroop is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

Toloka

Toloka is a crowdsourcing platform, designed by engineers to solve the most challenging problems with a diverse and growing set of applications. Toloka simplifies the process of testing, monitoring and deploying machine learning models. Toloka provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

mindsdb

MindsDB is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

bodywork

Bodywork is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

deepchecks

Deepchecks is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

LABELLER

Labeller is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

Solutions Integrators

Solutions Integrators are a group of companies that provide a range of services to help businesses integrate their AI solutions. They provide a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

DataArt

DataArt is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

PLATFORM9

Platform9 is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

MLops community

MLops community is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

DATASCIENCESALON

Data Science Salon is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

INOQ

INOQ is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

TRIOR

TRIOR is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

JUICE

JUICE is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

DataTalks Club

DataTalks Club is a platform for building and deploying machine learning models. It provides a robust dashboard and greater collaboration between data scientists and analysts. MLHub helps explain the production of back machine learning models and guide the confidence of their predictions, such as healthcare, biosci and defense agencies.

v **Спасибо!**

