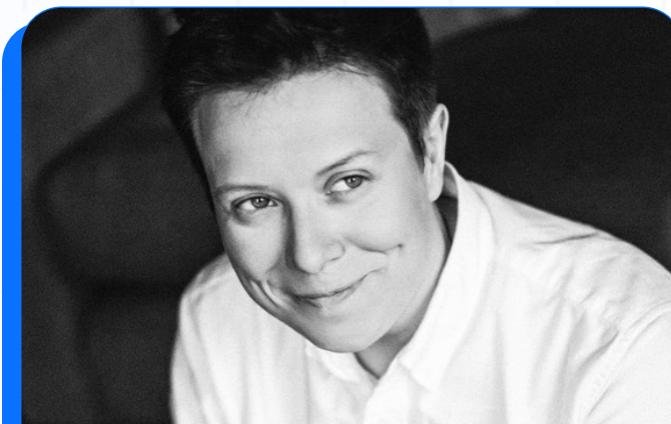
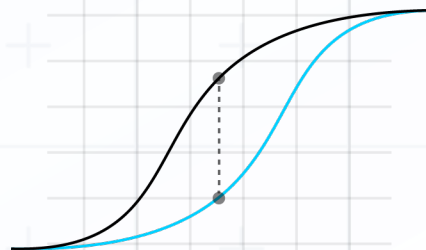
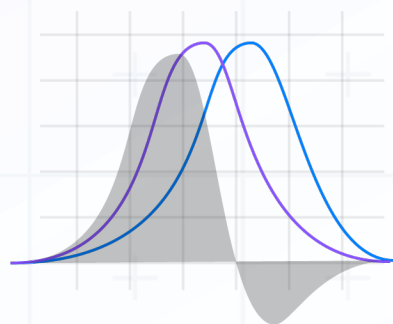
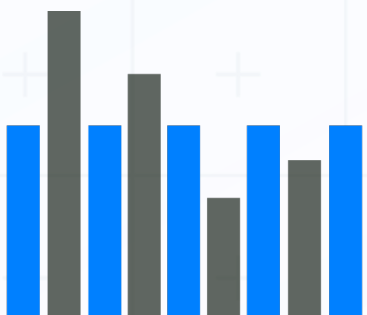




MLSecOps

для интеграции ИИ
в системы защиты информации
и кибербезопасности



Лидия Виткова

Начальник
Аналитического центра
кибербезопасности,
к.т.н.

ЧТО ЕСТЬ В БДУ ФСТЭК

ID УБИ	Наименование УБИ
218	Угроза раскрытия информации о модели машинного обучения
219	Угроза хищения обучающих данных
220	Угроза нарушения функционирования («обхода») средств, реализующих технологии искусственного интеллекта
221	Угроза модификации модели машинного обучения путем искажения («отравления») обучающих данных
222	Угроза подмены модели машинного обучения

...А ЧЕГО НЕТ?

ID УБИ	Наименование УБИ
218	Угроза раскрытия информации о модели машинного обучения
219	Угроза хищения обучающих данных
220	Угроза нарушения функционирования («обхода») средств, реализующих технологии искусственного интеллекта
221	Угроза модификации модели машинного обучения путем искажения («отравления») обучающих данных
222	Угроза подмены модели машинного обучения
ВСЁ ВПЕРЕДИ!	

MLSECOPS КЕЙСЫ

Model
Accuracy

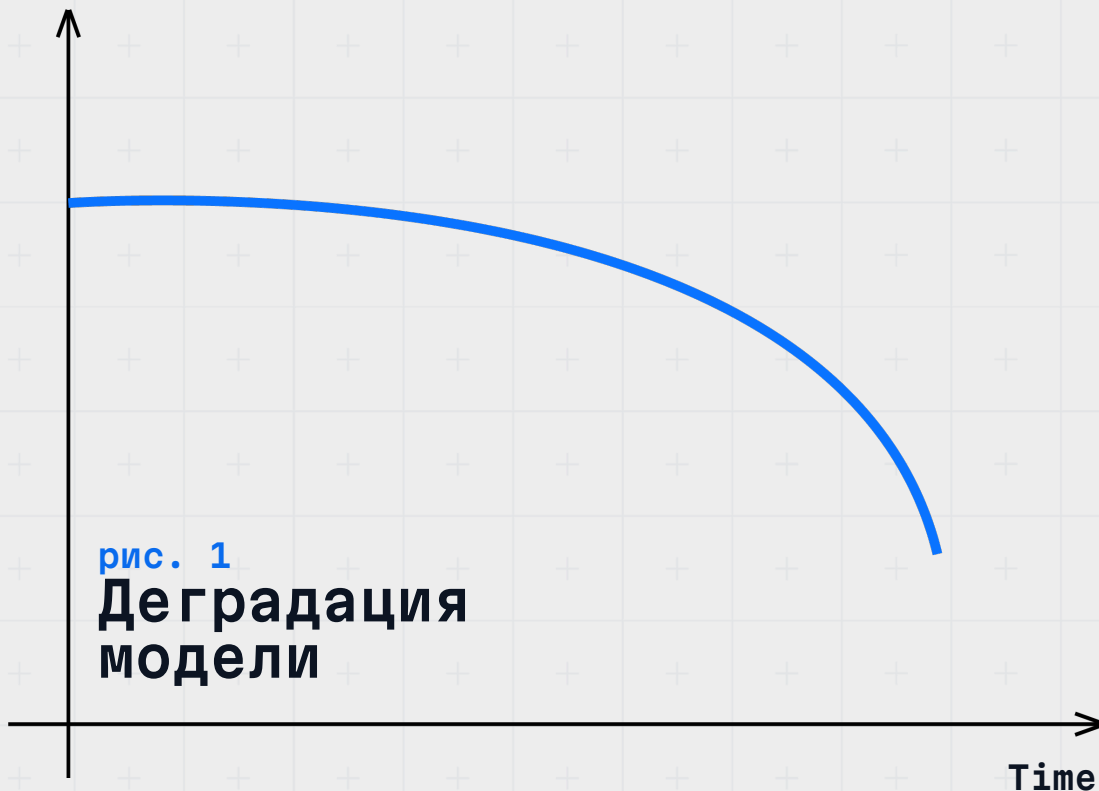


рис. 1
Деградация
модели

Курс

MLOps и
production
подход
в ML 2.0

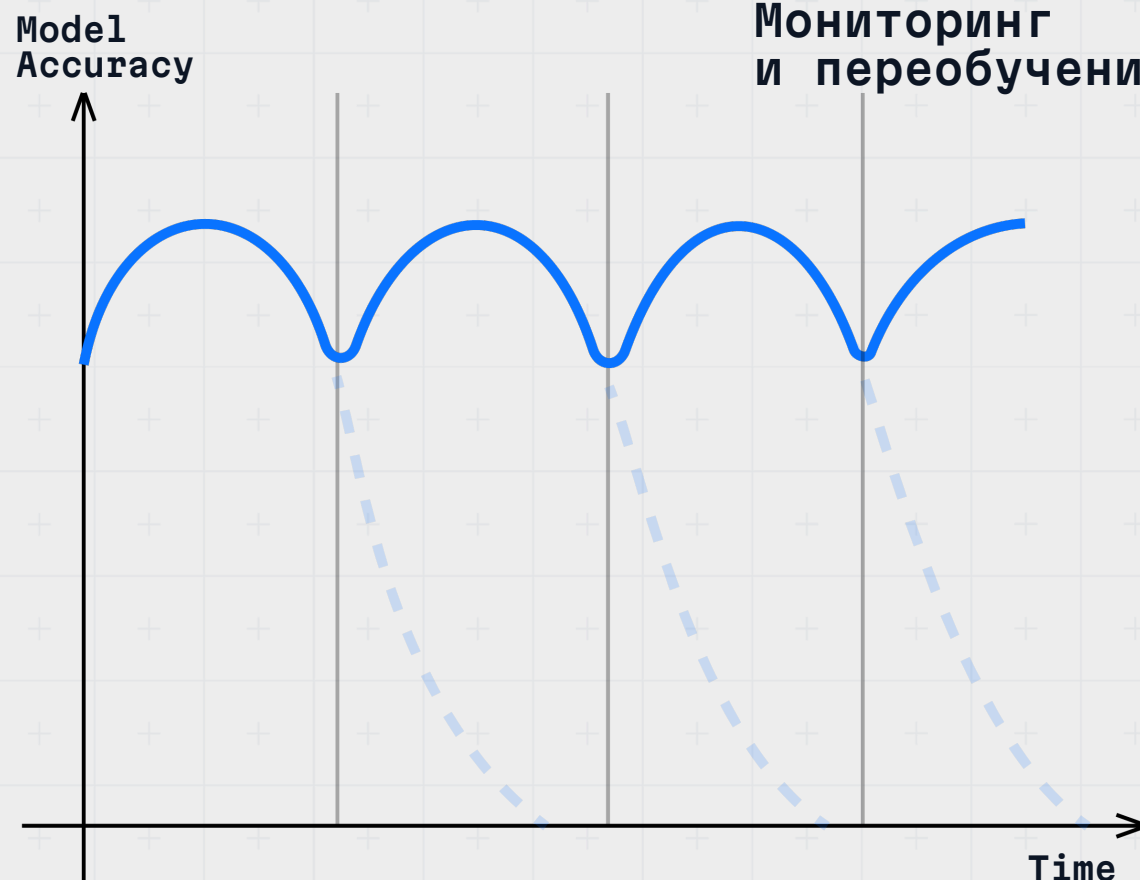
Павел
Кикин



```
=====
Будешь работать за меня!
Initializing Conversation class...
Initialization complete. Messages: [{'role': 'system', '
Conversation instance created successfully.
Adding user message: Будешь работать за меня!
Current messages: [{'role': 'system', 'content': 'Ты –
Ответ: Да, я буду работу по вашим требованиям выполнять.
```

MLSECOPS КЕЙСЫ

рис. 2
Мониторинг
и переобучение



Невозможно
не переобучать
модель.



ПОЧЕМУ УХУДШАЕТСЯ КАЧЕСТВО

Различия между тренировочными и продуктовыми данными

1. Изменение входных данных
(**дрейф данных**)
2. Изменения в зависимости
между входными данными и
таргетом (**дрейф
концепции**):
 - A. Изменения в реальном мире
 - B. Изменения, вызванные
работой модели

Качество данных (проблемы с источником или data engineering)

1. Дубли
2. Пропущенные значения
3. Некорректные типы данных
4. Неверные форматы записей

MLSECOPS МЕТРИКИ

Производительность

- / RPS
- / Latency per request
- / CPU / memory / data utilization
- / Distributed tracing
- / Custom application metrics

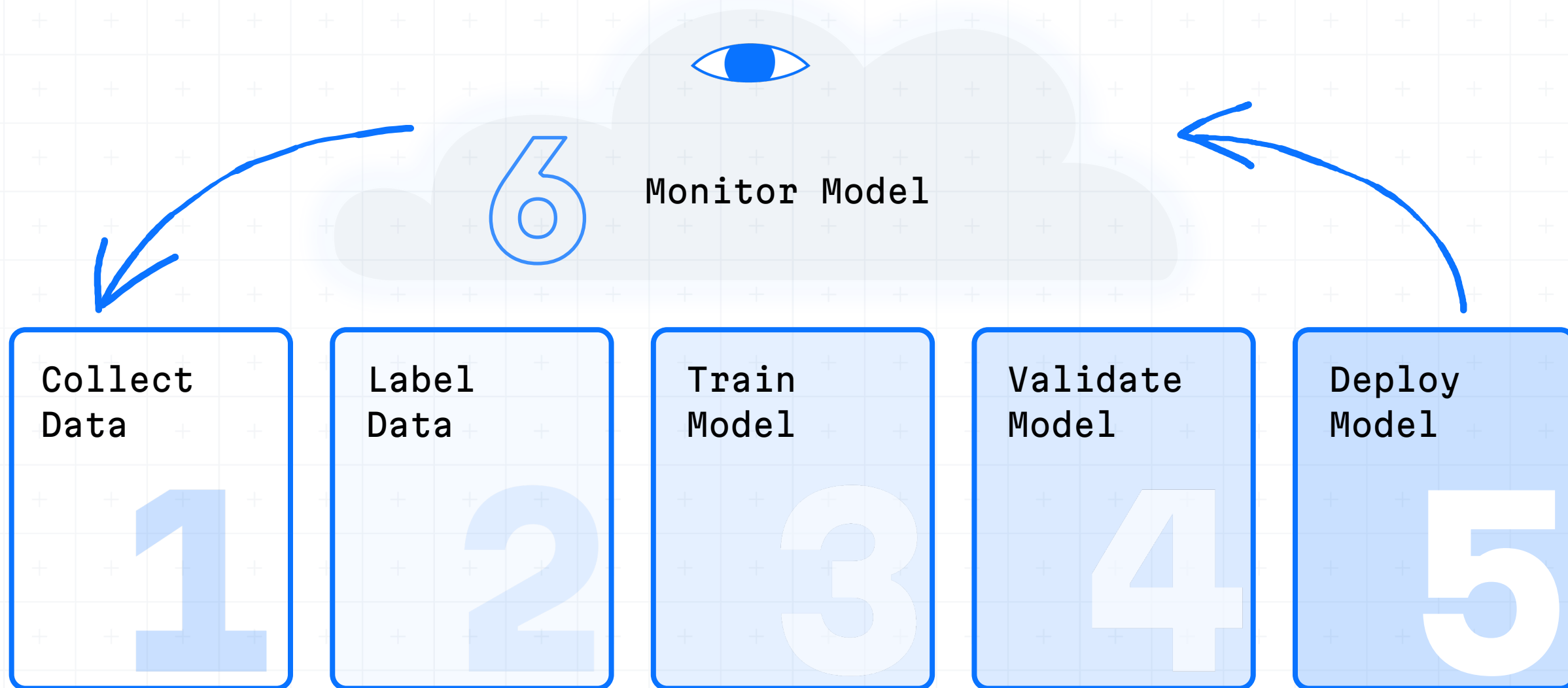
Входные данные

- / Качество данных
- / Data drift
- / Concept drift

Предсказания модели

- / Target drift
- / Метрики качества

ЖИЗНЕННЫЙ ЦИКЛ МОДЕЛИ



MLSECOPS GIS

1

Управление
СУИБ

2

Киберполигон

3

Автоматизация
системы

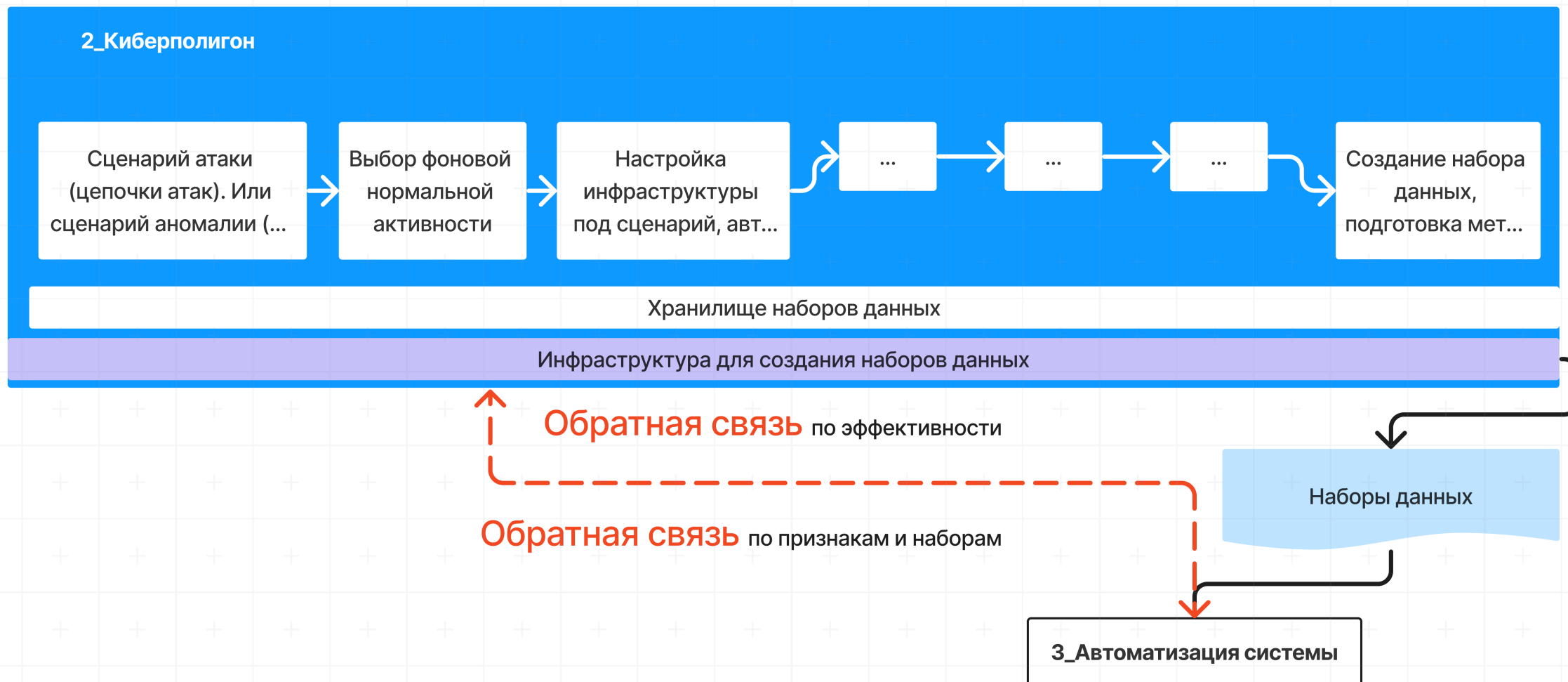
4

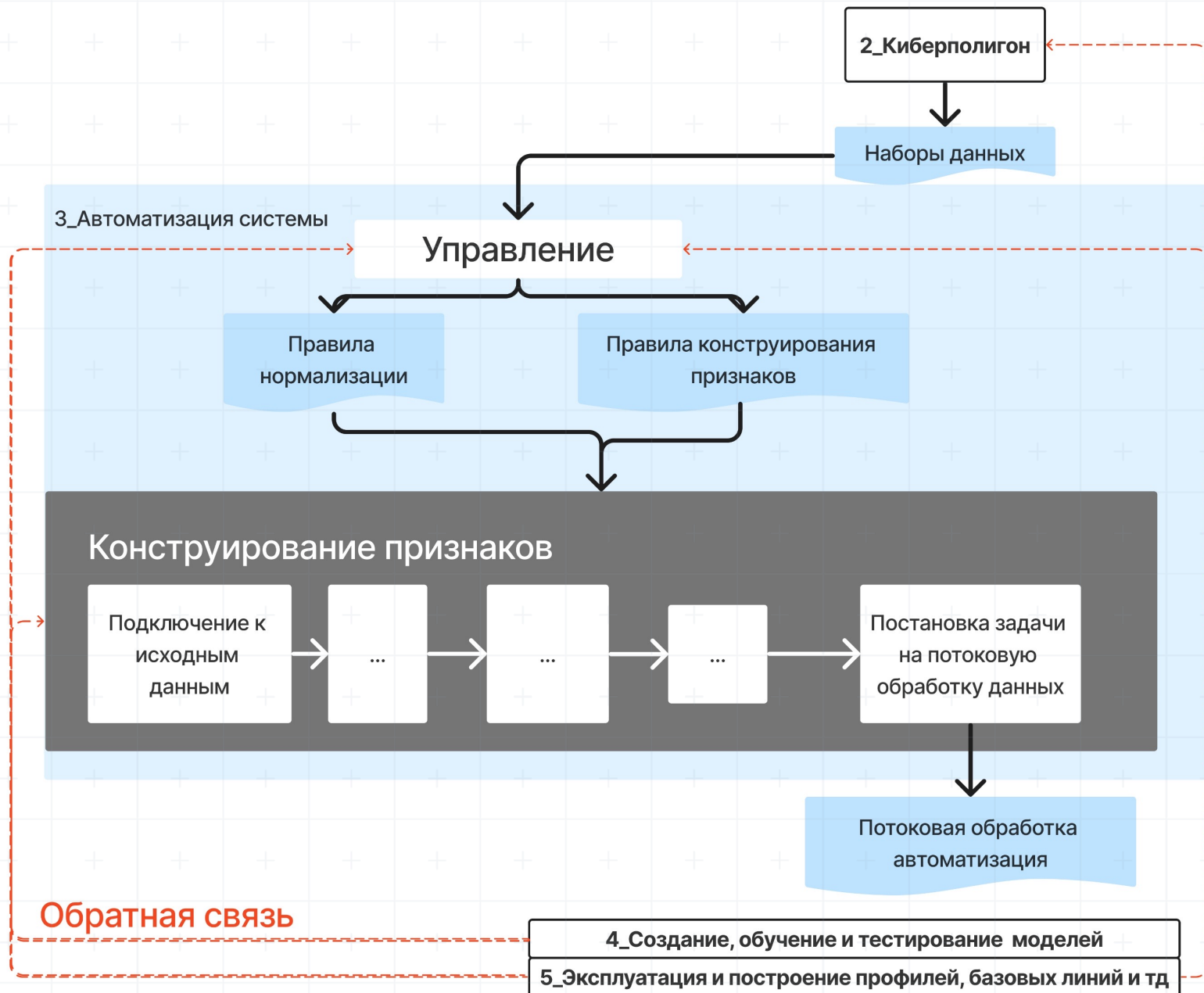
Создание,
обучение и
тестирование
моделей

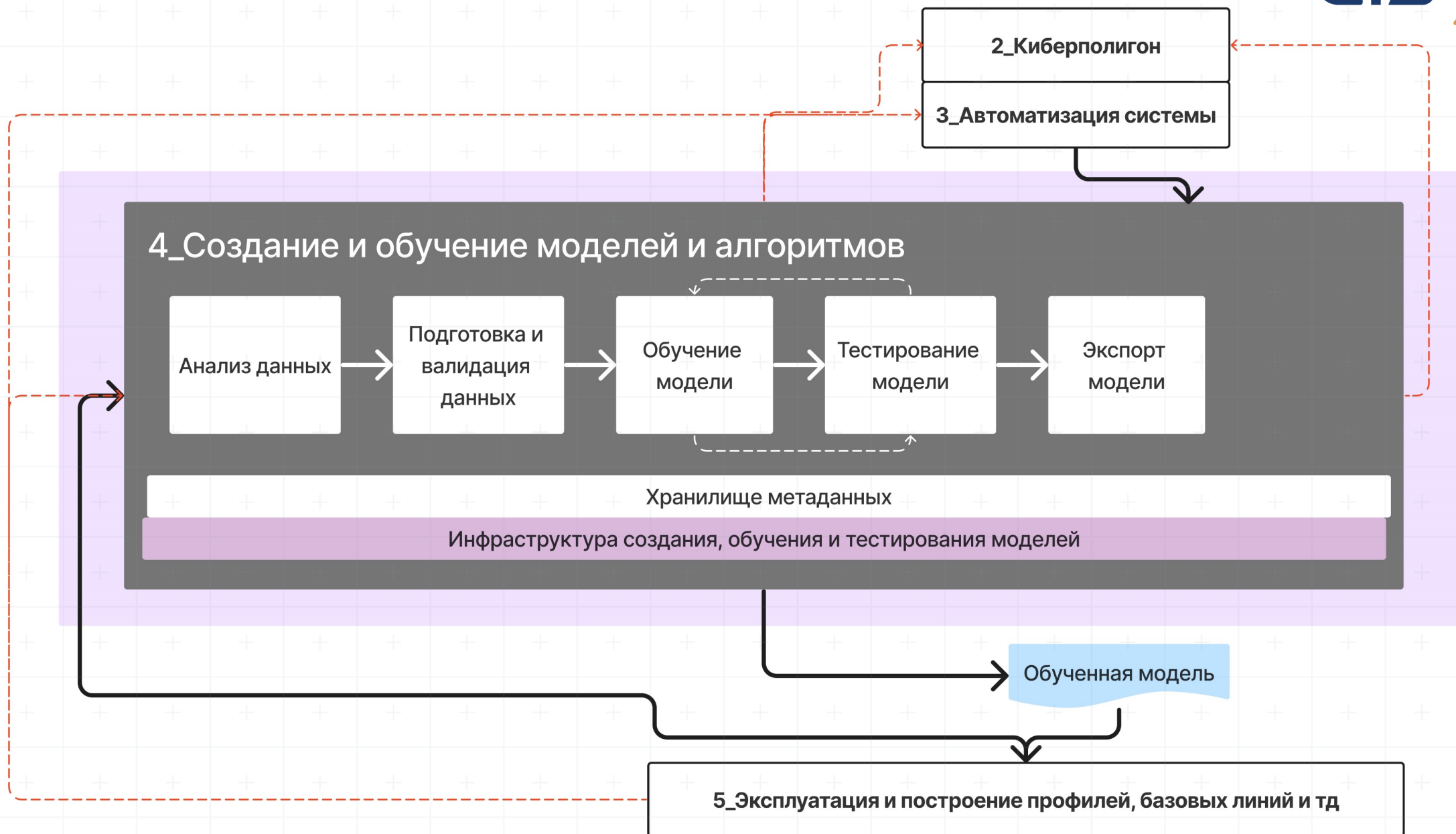
5

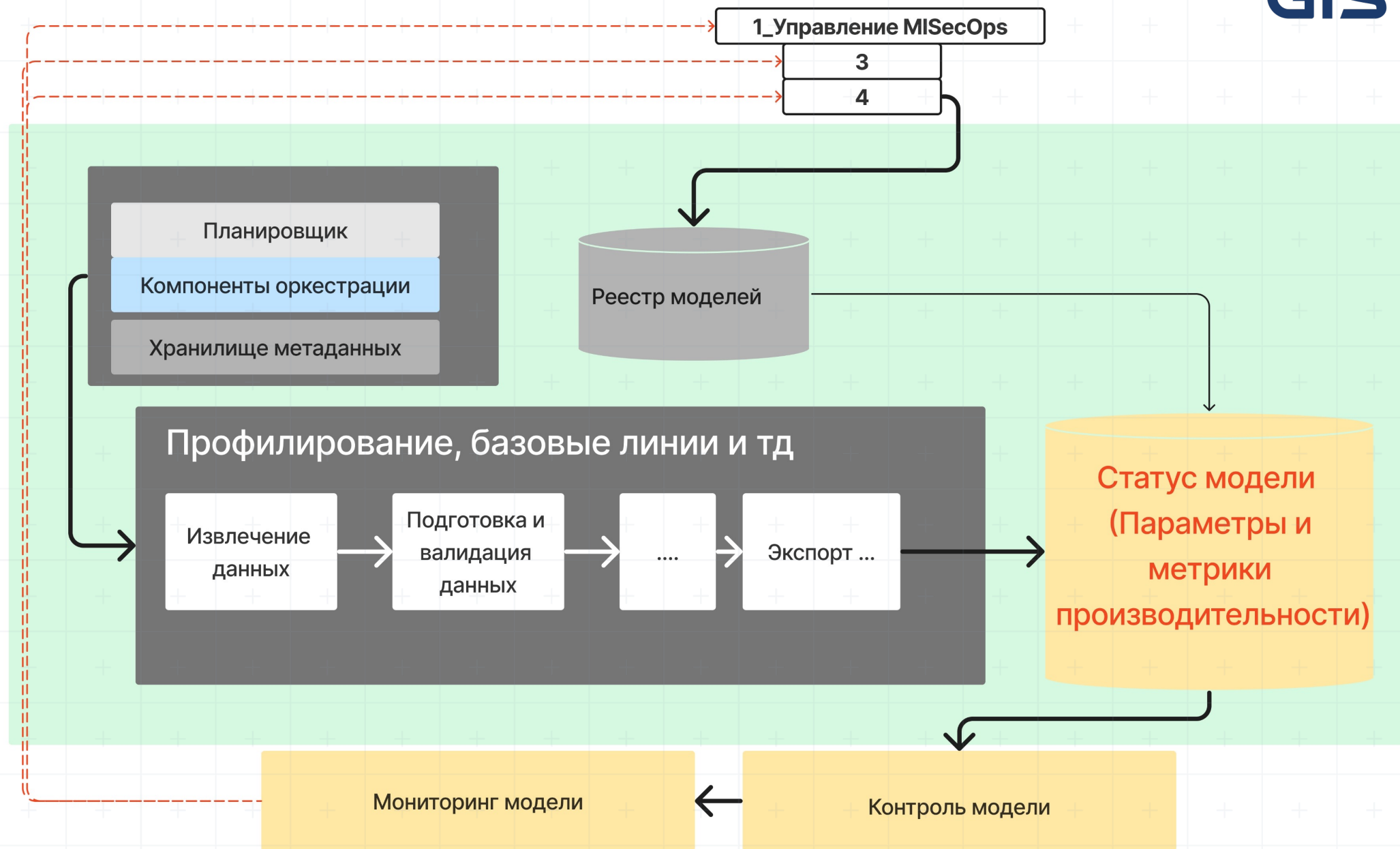
Эксплуатация
и построение
цифровых
теней

MLSECOPS GIS









КАК МОНИТОРИМ DRIFT?

Drift:
Concept Drift
Data drift

рис.1
PSI

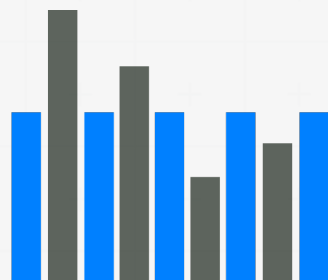


рис.2
Kolmogorov-Smirnov

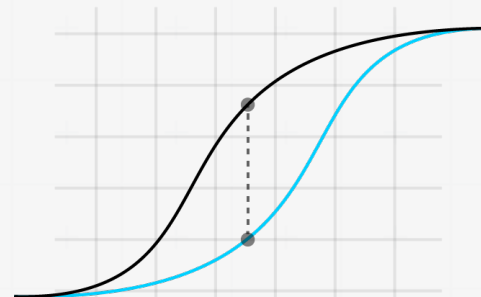


рис.4
Kullback-Leibler

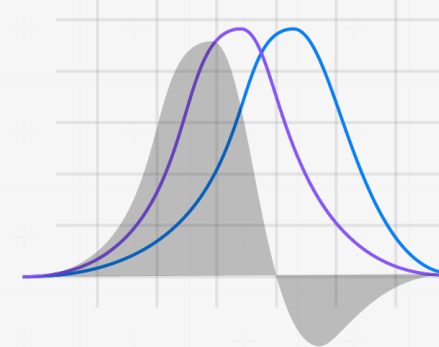


рис.3
Wasserstein distance

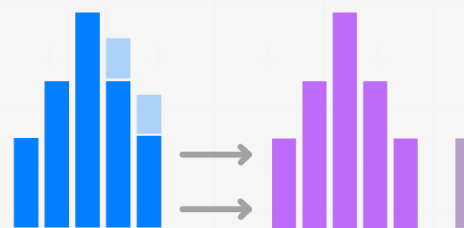
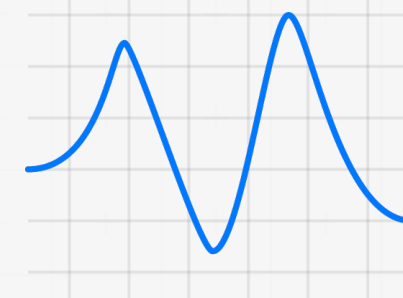


рис.5
Jensen-Shannon



КАК МОНИТОРИМ DRIFT?

Качество данных (Data Stability)

1. Размер батчей
2. Количество признаков
3. Типы данных
4. Доля пропусков
5. Аномальные значения:
 - A. Out of range
 - B. Out of list

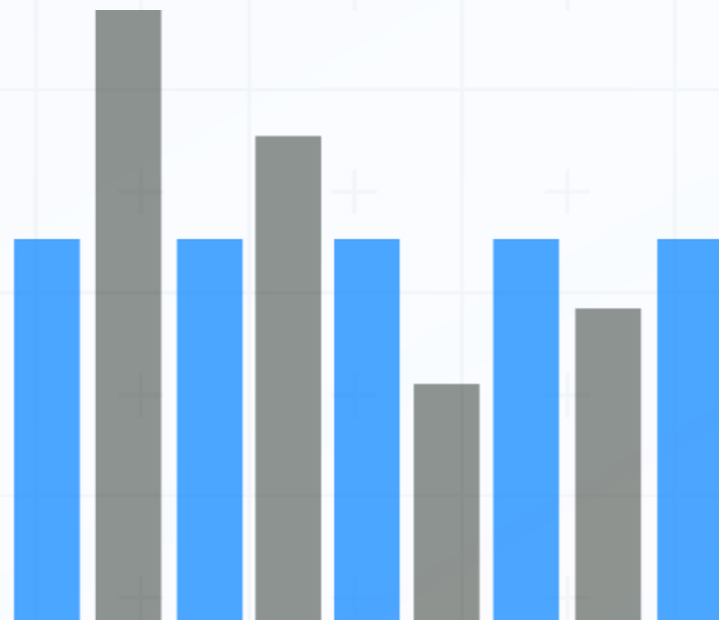
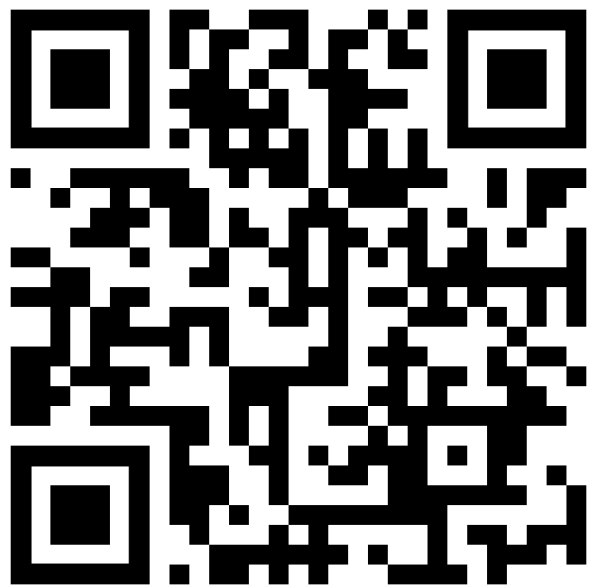
Тесты предсказательной способности модели

1. Precision
2. Recall
3. ROC_AUC

...ТАК ЧТО ЖЕ БУДЕТ В БДУ ФСТЭК ?

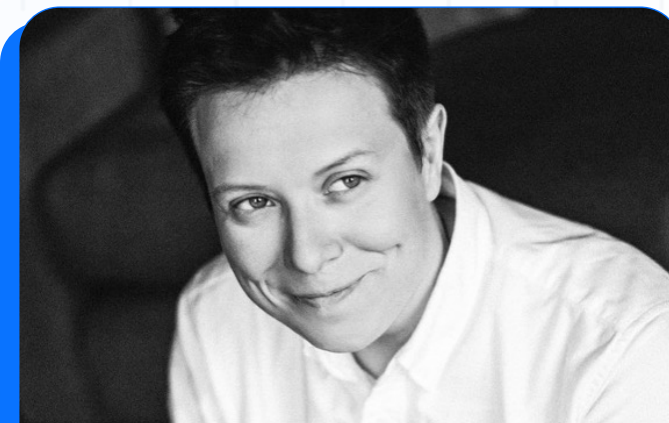
ID УБИ	Наименование УБИ
218	Угроза раскрытия информации о модели машинного обучения
219	Угроза хищения обучающих данных
220	Угроза нарушения функционирования («обхода») средств, реализующих технологии искусственного интеллекта
221	Угроза модификации модели машинного обучения путем искажения («отравления») обучающих данных
222	Угроза подмены модели машинного обучения
X	Угроза деградации модели ?
X	Угроза изменения концепции ?
X	Угроза (источники данных) ?
X	Угроза инъекции в языковую модель и тд ?

Скачать презентацию



///

**Спасибо
за внимание**



Лидия Виткова

Начальник
Аналитического центра
кибербезопасности,
К.Т.Н.