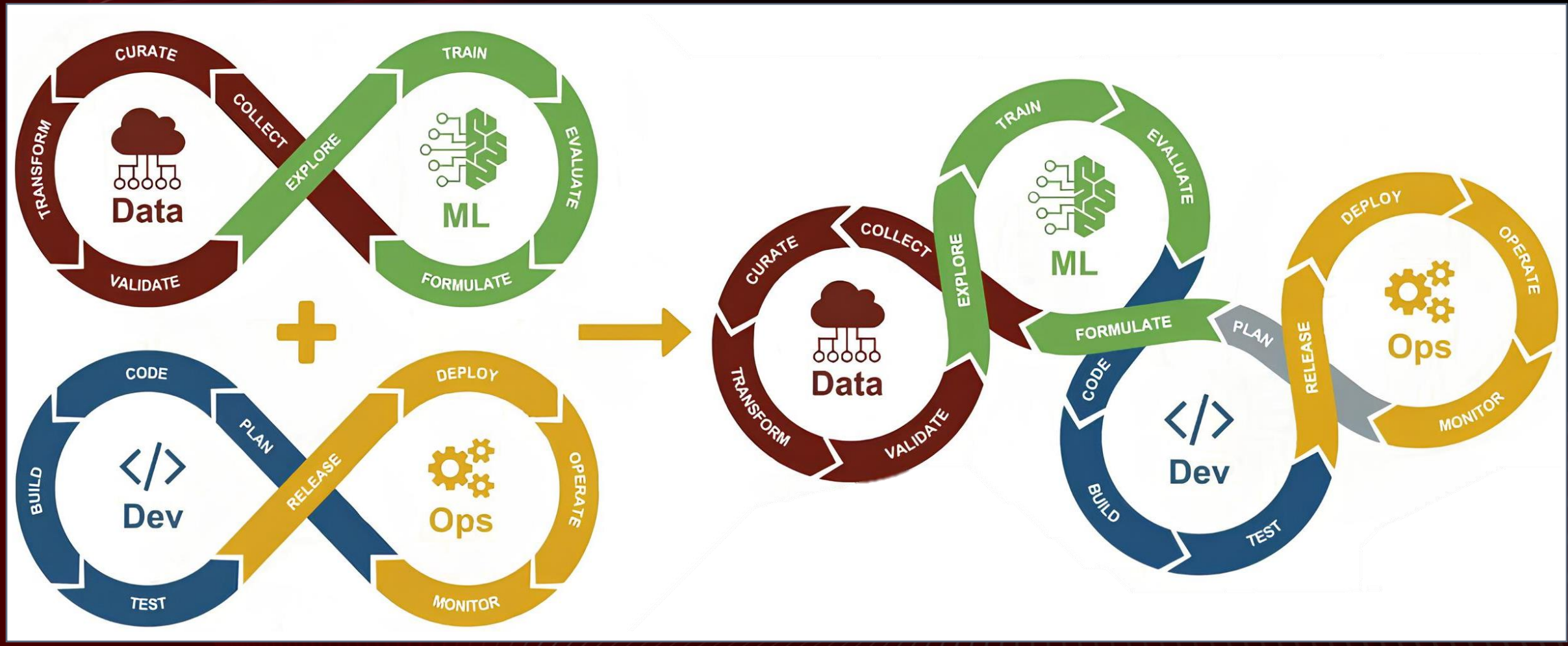
MLSecOps-подход

Газизова Светлана
Positive Technologies

«Конвейер»?


CURATE
COLLECT
TRANSFORM
EXPLORE
Data
VALIDATE
TRAIN
EVALUATE
ML
FORMULATE
CODE
PLAN
BUILD
RELEASE
Dev
TEST
DEPLOY
OPERATE
Ops
MONITOR

Кейсы атак на ИИ

1 Ray Framework vulnerability

Злоумышленники воспользовались CVE-2023-48022 – уязвимостью в фреймворке RAY, который используется для оркестрирования моделями машинного обучения. Воспользовавшись уязвимостью злоумышленники смогли исполнить код на хостах, где расположена модель.

https://www.oligo.security/blog/shadowray-attack-ai-workloads-actively-exploited-in-the-wild

2 VirusTotal Poisoning

Атака на систему анализа вредоносного ПО, где злоумышленник использует образец вредоносного ПО, пропускает его через метаморфический код и получает файлы, которые не запускаются, но распознаются антивирусом. Эти файлы попадают в датасет антивируса, отравляя его. В результате оригинальное вредоносное ПО становится труднее распознавать.
https://atlas.mitre.org/studies/AML.CS0002

3 Bypassing Cylance's AI Malware Detection. AI Cylance

Проведя анализ работы антивируса Cylance, изучив открытые источники, патенты и включив подробное ведение журнала злоумышленники выяснили, что ансамбль моделей (называемый "первой моделью") определяет вредоносное ПО, но его результаты могут отменяться второй моделью. Благодаря этому, злоумышленники смогли обойти основную модель, используя вторую модель.
https://atlas.mitre.org/studies/AML.CS0003

4 Compromised PyTorch Dependency Chain

Была скомпрометирована цепочка зависимостей. Вредоносная зависимость «torchtriton» на PyPI имеет то же имя, что и библиотека в репозитории PyTorch-nightly. Поскольку PyPI имеет приоритет в экосистеме Python, на машину попадает вредоносный пакет вместо PyTorch. Вредоносный «torchtriton» сканирует систему для получения базовой информации и крадет конфиденциальные данные.
https://atlas.mitre.org/studies/AML.CS0015

Атаки на ИИ

Все классические угрозы приложений, а также…

Угрозы данных:

Манипуляция данными (Data Poisoning): Утечка, некачественные данные, нарушение целостности данных.

Угрозы модели:

Атаки на алгоритмы (Model Manipulation): Злоумышленники могут использовать уязвимости в алгоритмах для манипуляции результатами.

Атаки на обучение (Training Attacks): Вредоносные вмешательства в процесс обучения могут повлиять на параметры модели.

Функции отторжения (Feature Poisining): Добавление вредоносных характеристик может привести к компрометации модели.

Недостаточный контроль доступа: Без правильных мер доступа злоумышленники могут вмешаться в процесс обучения.

Эксплуатация приложения с моделью:

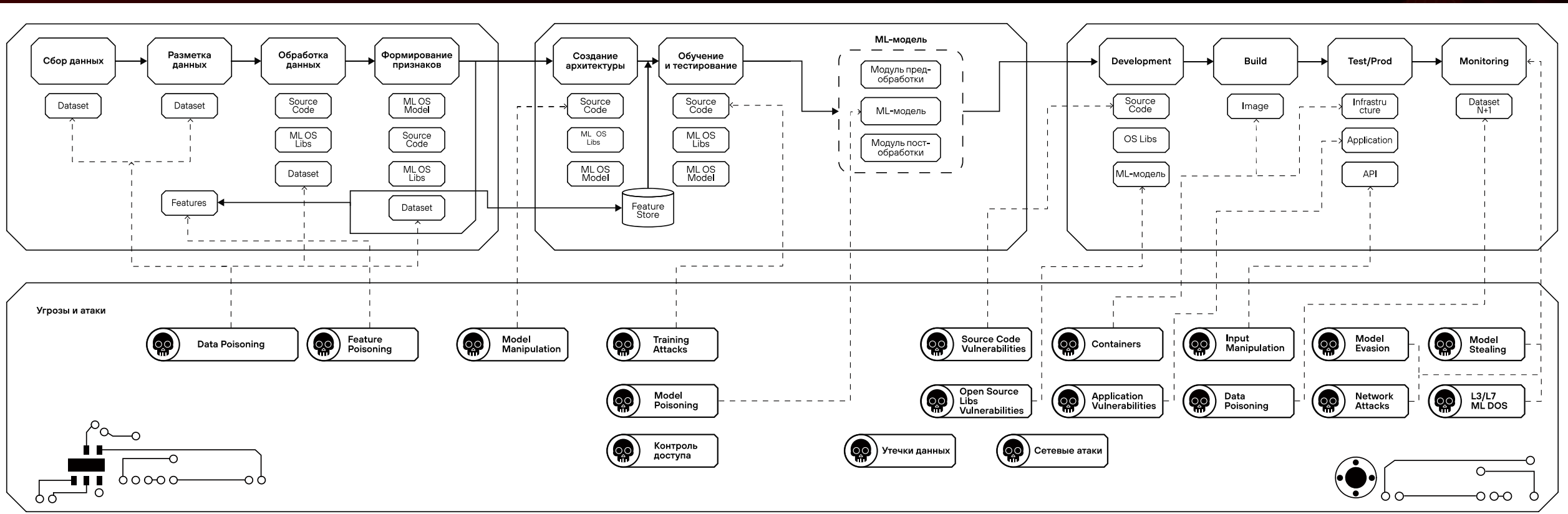
Атаки на модель в рабочем состоянии (Model Evasion): Злоумышленники могут создавать входные данные, которые обманывают модель.

Model Stealing: Использование API модельных сервисов для воспроизведения и кражи модели.

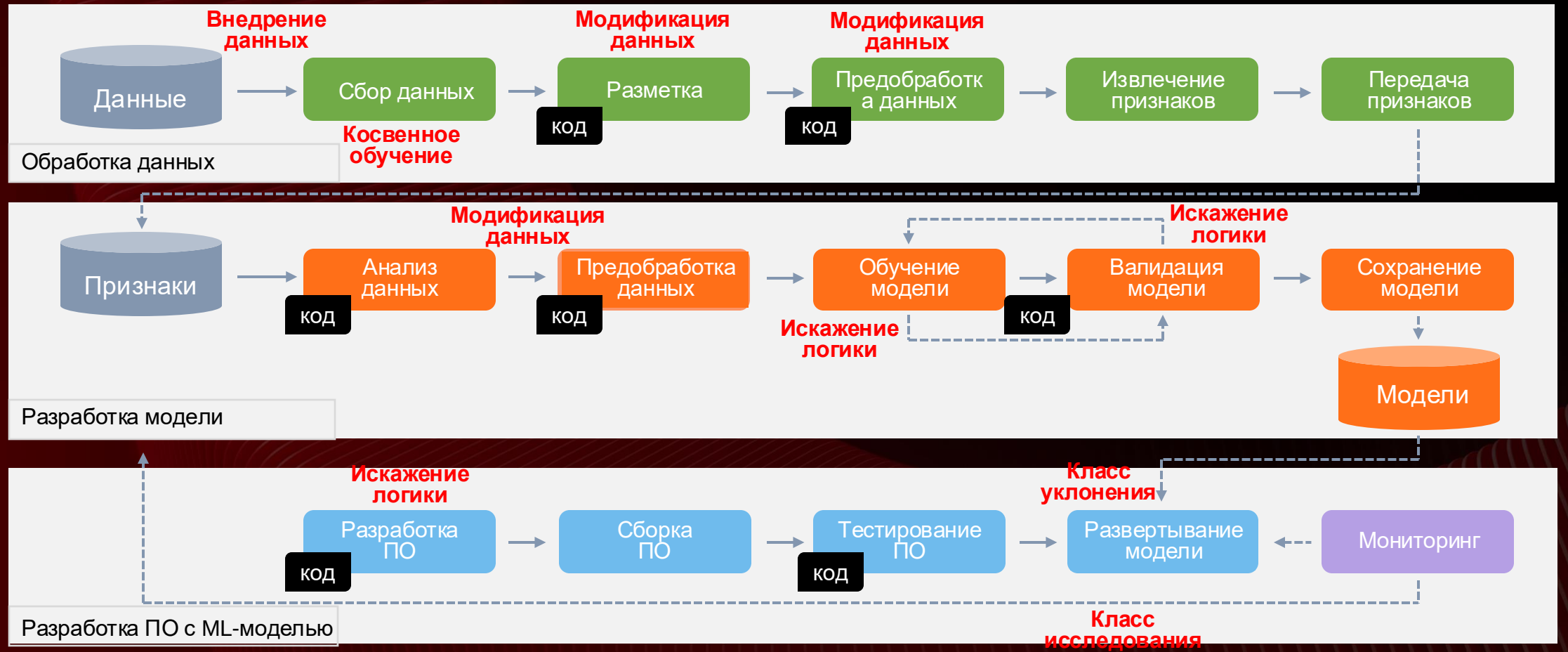
Неправильные обновления: Некорректно обновленные модели могут вести себя ненадежно.

DDOS атаки (Distributed Denial of Service): Направленные к модели сервисы могут быть перегружены.

А если детализируем?



Атаки внутри конвейера



Защита внутри конвейера