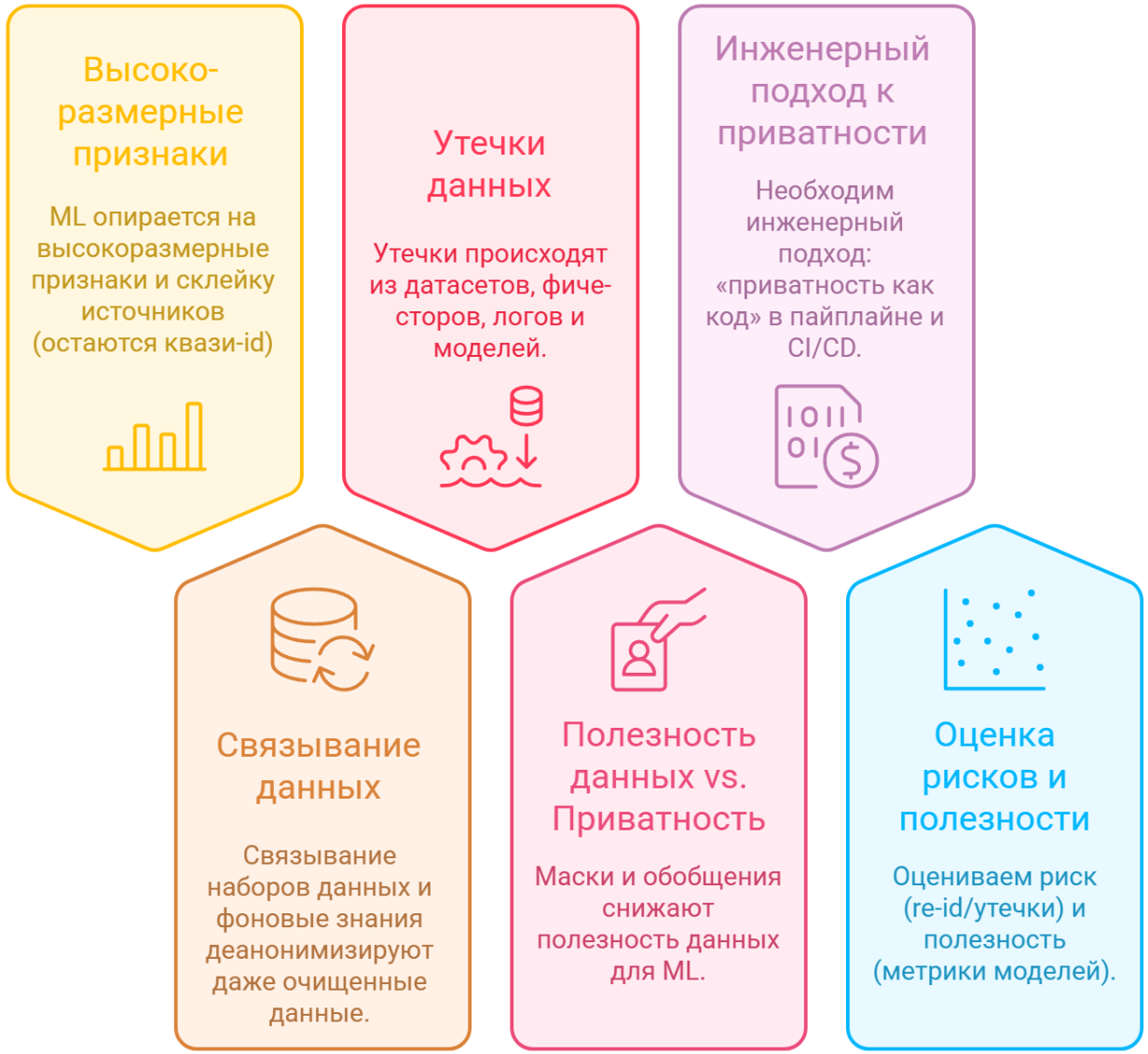
Приватность данных в эпоху ML: обезличивание, анонимизация и синтетические данные на практике

Силаев Юрий Владимирович
Старший преподаватель ДПИ ФКН НИУ ВШЭ

Почему «удалить ФИО» не спасает в эпоху ML



Термины, которые нельзя смешивать

Термин	Кратко «что это»	Статус по ПДн	Обратимость	Для ML
Псевдонимизация	Замена идентификаторов на коды; ключ соответствия хранится отдельно	**Остаётся ПДн**	Обратимо при доступе к ключу	Удобно для связки записей и аудита, но сохраняет риск linkage или pe-id
Обезличивание (деперсонализация)	Обработка, при которой без доп. инф. нельзя определить субъекта	**Не ПДн**, если действительно исключена идентификация	Должно быть необратимо для оператора и третьих лиц	Требует количественных проверок риска; иначе мнимая защита
Полная анонимизация	Необратимая потеря связи с субъектом при разумных ресурсах	**Не ПДн**	Необратимо	Часто ломает полезность (особенно по редким классам или взаимодействиям)
Методы (маскирование, генерализация, FPE/токенизация)	Техники преобразования данных	**Не статус**, а инструменты	Зависит от метода	Влияют на распределения и качество моделей; методы НЕ результат

Что фиксирует приказ РКН № 140

1	**Цель** Унифицировать требования к обезличиванию и контроль ответственности оператора
2	**Процесс** Подготовка, выбор методов, проверка результата, документирование
3	**Категории методов** Замена/кодирование идентификаторов; модификация значений; удаление/супрессия; перестановка/перемешивание; агрегирование
4	**Организационные меры** Локальные акты, регистры операций, раздельное хранение ключей/массивов
5	**Контроль доступа** Разграничение прав, протоколирование обращений к ключу соответствия
6	**Документирование** Акты/протоколы обезличивания, описание применённых методов и условий
7	**Области применения** Внутренняя разработка, тестирование, обмен в рамках закона

Почему приказ РКН № 140 недостаточен для ML

Количественные критерии риска

Отсутствуют количественные критерии риска re-id и linkage для конкретных наборов данных.

CI/CD-гейты

Отсутствуют CI/CD-гейты и регрессионные тесты приватности в Dev/MLOps.

Дифф. приватность

Отсутствует дифференциальная приватность: ε/δ, бюджет, учёт потерь.

Синтетические данные

Для синтетических данных нет критериев уместности и проверок утечек или похожести.

Совместная оценка «риск-полезность»

Отсутствует совместная оценка «риск-полезность» на целевых ML-метриках.

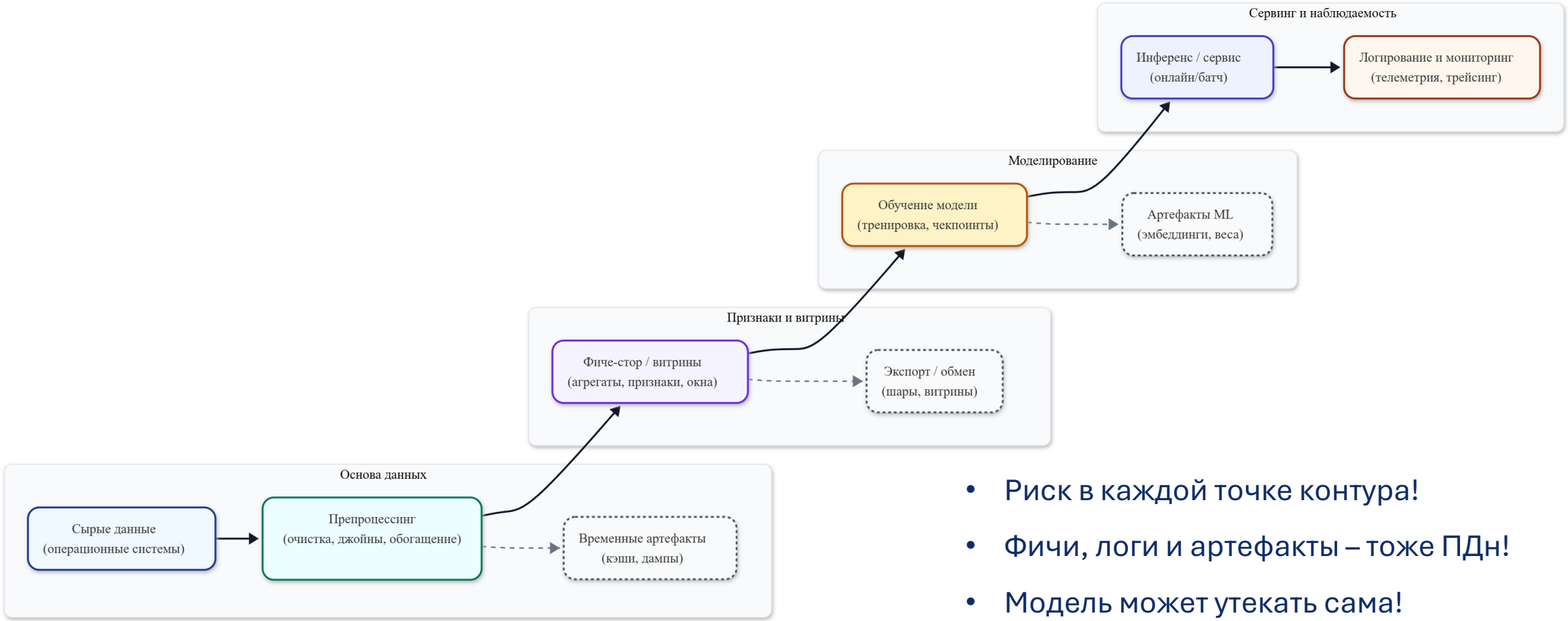
Модельные утечки

Модельные утечки (membership-inference, inversion) не покрыты.

Фиче-сторы

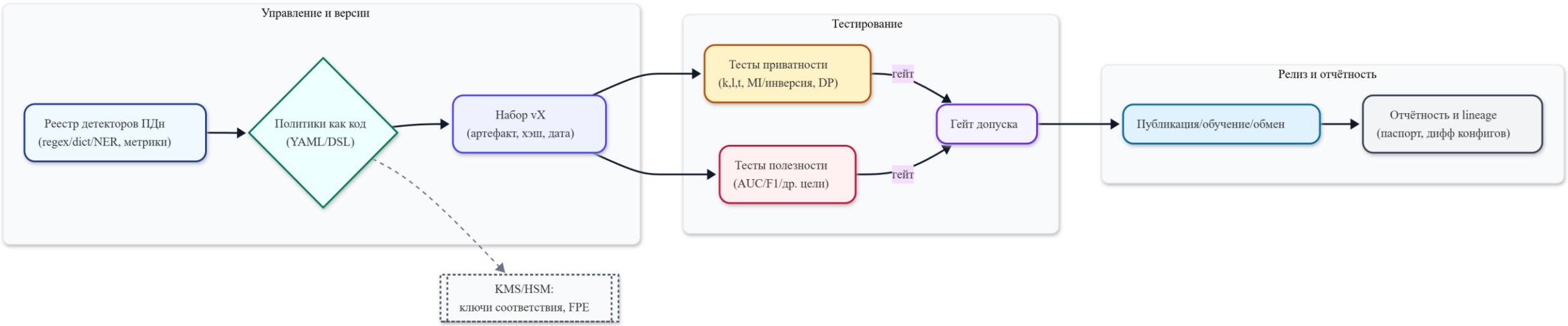
Фиче-сторы, логи, артефакты обучения полностью слепая зона требований.

Где ломается приватность в ML-пайплайне



- Риск в каждой точке контура!
- Фичи, логи и артефакты – тоже ПДн!
- Модель может утекать сама!

Приватность как код



- Декларативные политики и версионирование артефактов
- Автотесты приватности и полезности в CI/CD
- Гейты с порогами, блокировкой релиза, исключениями по процедуре
- Паспорт набора и lineage для аудита
- KMS, HSM для ключей и соответствий

Детекторы ПДн

Тип детектора	Что ловит	Слабые места	Метрики, порог	Точки запуска
Регулярные выражения, шаблоны	e-mail, телефоны, паспорта, ИНН, СНИЛС, карты и т.п.	Ложные срабатывания на коды, артикулы	Precision/Recall ≥ целевых; FPR ≤ порога	Вход ETL, экспорт, логи
Структурные валидаторы (контрольные суммы)	ИНН, СНИЛС, карты, ИБАН и т.п.	Локальные форматы, исключения	Rate валидных и невалидных; coverage по полям	Препроцессинг, фиче-стор
Словари, справочники	ФИО, адреса, организации, домены	Омографы, редкие имена	Recall по эталонам; OOV-доля	Вход, выход, текстовые поля
NER-модели	PERSON/ORG/LOC/DATE/ID-like в тексте	Доменный дрейф, жаргон	F1 на тест-корпусе; drift-алерты	Документы, заметки, логи
OCR и парсинг вложений	ПДн в PDF, сканах, изображениях	Качество OCR, верстка	% успешно распознанных; error-rate	Приёмка файлов, архивы
Семантические, эмбеддинг-детекторы	«Смыслы» (диагнозы, места работы)	Шум, близкие синонимы	Top-k accuracy; threshold	Текст, чат, поддержка
Квазиидентификаторы, редкость	Уникальные комбинации, linkage-риск	Выбор квази-полей	k-anon./uniqueness; risk-score	Перед публикацией набора

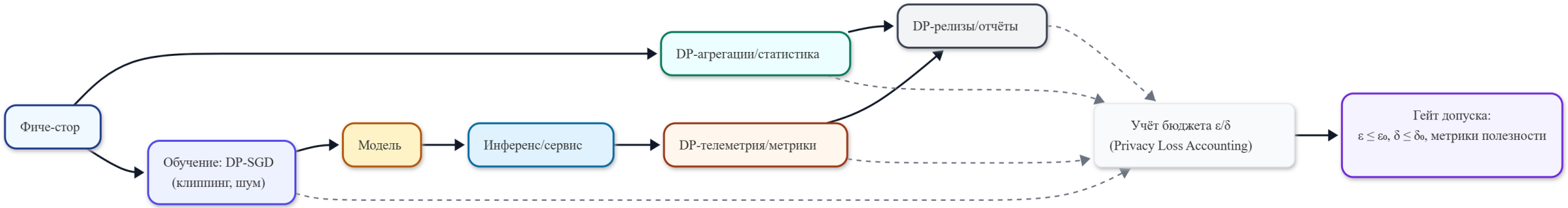
Политики обезличивания

Метод	Где применять	Плюсы	Основные риски	Влияние на ML
Токенизация (детермин., рандом.)	Связка записей, аудит, реидентификация по запросу	Сохраняет связность, управляемая обратимость	Утечка ключей, маппинга; детерминизм даёт linkage между доменами	Сохраняет join'ы; как фича – бесполезна и опасна
FPE (симм. шифрование с сохранением формата)	Устойчивые идентификаторы (account_id, phone), межсистемные джойны	Формат и равенство сохраняются, удобно для интеграций	Детерминизм выдаёт равенство/частоты; риск компрометации ключей	Джойны ок; как признак – поддерживает уникальность, а значит – риск
Хеширование и соль/«перец»/HMAC	Безвозвратная защита; join при общем секрете (HMAC)	Односторонность; можно делать согласованные join'ы	Без соли – атаки словарями; детерминизм даёт linkage	Как фича бесполезно; годится только для связки
Генерализация, биннинг, округление	Квазиидентификаторы: возраст, гео, время, редкие категории	Снижает уникальность, упрощает политику доступа	Ломает взаимодействия и «хвосты»; риск смещения	Потери по точности и recall, страдают минорные классы
Супрессия, удаление, редакция	Чрезмерно чувствительные поля/редкие значения	Устраняет источник риска полностью	Паттерны пропусков становятся идентификаторами; потеря сигналов	Рост пропусков вызовет импутацию, деградацию качества
Шум, пертурация, микроагрегация	Числовые поля, отчёты/обмен, статистика	Простая маскировка, можно калибровать	Некоррелированный шум убивает структуру; «слишком малый» не спасает	Баланс «шум-метрики» критичен; возможна деградация
Перестановка, свапинг (атрибут, строки)	Разрыв прямых связей, снижение linkage	Сохраняет маргинали, прост в реализации	Ломает временные и реляционные зависимости; обратим паттернами	Плохо для временных и графовых моделей; ок для агрегатов

Метрики приватности и утечек

Метрика, тест	Что измеряет	Когда применять	Типичные ловушки	Что идёт в проверку
k-анонимность	Минимальный размер «неотличимых» групп по квазиидам	Табличные наборы с ограниченным числом квази-полей	Высокая размерность, редкие категории «разбивают» группы	$k \geq k_0$ по ключевым квазидам
l-разнообразие	Разнообразие чувствительных значений внутри k-группы	Когда важна защита от атрибутной атаки	«Разнообразие» фиктивное при дисбалансе классов	$l \geq l_0$ на всех k-группах
t-близость	Насколько распределение в k-группе близко к общему	Для тонкой настройки при смещениях	Нестойкость к плохому выбору метрики близости	$t \leq t_0$ (дистанция)
Uniqueness, редкость	Долю уникальных записей по квазиидам	Для первичной оценки linkage-риска	«Безвредные» поля в квазиидах занижают риск	$U \leq U_0$ (процент уникальных)
Nearest-Neighbor Radius (DCR)	«Близость» записи к ближайшей в исходных данных	При публикациях и обмене, синтетике	Неверная нормализация, одна метрика на все типы	$r \geq r_0$ (минимальный радиус)
Membership Inference (MI)	Выявляемость факта участия записи в обучении	Для моделей и витрин после обучения	Непредставительный атакующий датасет	AUC(attacker) ≤ α; Adv ≤ β
Model Inversion (sanity)	Восстановление чувствит. атрибутов из ответов модели	Для сервисов и онлайн-инференса	Оверсемплинг редких классов «подсказывает» инверсию	Acc(inv) ≤ γ на чёрном ящике

Где включать дифференциальную приватность и как контролировать бюджет



- Калибруем шум по чувствительности и цели ε, δ
- Распределяем бюджет между тренингом, аналитикой, телеметрией
- Ведём PLA: композиция, остаток бюджета, алерты
- Валидируем качество до и после шума на эталонных задачах
- Порог ε, δ являются частью гейта; любые исключения только по процедуре

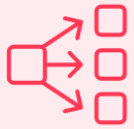
Когда синтетические данные «самое то»

Анонимизация

Классическая анонимизация ломает полезность, а обмен всё равно нужен.

Балансировка данных

Балансировка редких классов, снятие «острых» признаков.



Проверки приватности

«Синтетика ≠ приватность»: нужны проверки приватности, похожести и полезности.

Стандарты

Опираемся на 3 стандарта ПНСТ 1.11.164 - термины, архитектура и методы синтеза, методы оценки качества.

Разработка и тестирование

Dev, Test, демо без ПДн: репро, A/B, обучение новых команд.

Частичный синтез

Частичный синтез: только рискованные столбцы, срезы.

Встраивание в конвейер

Встраиваем в конвейер: генерация, тесты, гейт, паспорт.

Оценка синтетики: похожесть, полезность, приватность

Класс метрик	Примеры	Что показывает	Частые ошибки
Похожесть: маргинали	KS/χ^2 по числовым/категориальным, энтропии	Сохранение одномерных распределений	Игнор зависимостей; сравнение на «грязных» шкалах
Похожесть: зависимости	Корреляции (P/S), Cramér's V, MI	Парные связи и сила ассоциаций	Оценка только парных связей
Похожесть: многомерная геометрия	PCA/UMAP-дистанции, MMD/energy distance	Кластеры и глобальную структуру	Смешение типов без нормализации; чувствительность к масштабу
Полезность	TSTR/TRTS, ΔAUC/ΔF1, cross-fit	Насколько синтетика «заменяет» реальные для задач	Утечки через фичи, сплиты; оверфит к генератору
Стабильность, калибровка	ECE/Brier, lift/PR-кривые, subgroup gap	Сохранение калибровки и fairness по подгруппам	Игнор дисбаланса и редких классов
Приватность: меморизация	NN-radius/DCR, exact/near-dup rate, canaries	Близость к исходным записям, утечки шаблонов	Неверные метрики расстояния; дубликаты в исходнике
Приватность: атаки на модели	MI AUC/adv., attrib.-inference acc.	Выявляемость участия, восстановление чувствительных атрибутов	Непредставительный «атакующий» датасет; не тот пайплайн

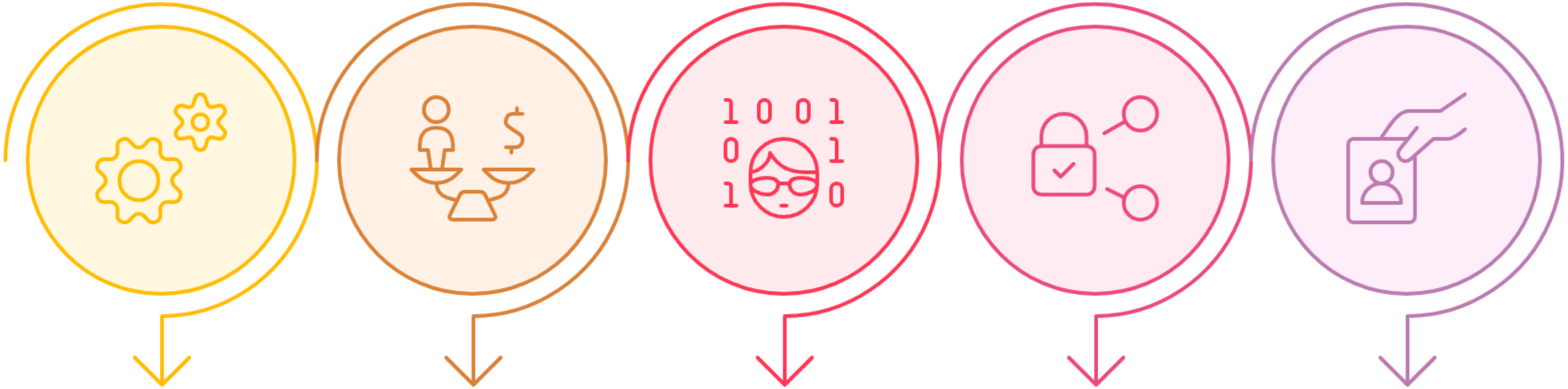
Полная отчётность: паспорт набора и гейты допуска

Раздел паспорта	Обязательные поля	Артефакты, доказательства	Ответственный
Идентификация набора	ID, версия vX, дата, владелец, цель использования	Хэш артефакта, URI хранения	Data Owner
Состав ПДн (до, после)	Перечень полей, классификация, квази-ID	Отчёты детекторов (версии, метрики)	Privacy Engineer
Применённые политики	Policy YAML/DSL, параметры, KMS-ссылки	Коммит или хэш политики, журнал применений	Privacy Engineer или Platform
Метрики приватности	k/l/t, % уникальных (U), NN-radius/DCR, MI AUC/Adv, DP ε/δ	Репорты тестов, протоколы PLA	Privacy Engineer и ML Lead
Метрики полезности	Бейзлайн vs текущие (ΔAUC/ΔF1/калибровка)	Бенчмарки, протокол валидации	ML Lead
Решение гейта	Пороги, PASS/FAIL, исключения (обоснование и срок)	Запись согласования (тикет и подписи)	DPO и Data Owner
Lineage и изменения	Источники, трансформации, diff vX-vY	Схема lineage, changelog	Platform или Data Eng

Модель зрелости приватности в ML

Уровень	Что уже есть	Пробелы, риски	Следующий шаг
L0 – AD-hoc	Ручные маски и удаление ФИО, локальные акты	Пропуски ПДн, нет метрик и отчётности, «времянки» утекают	Каталог детекторов, базовые политики, раздельное хранение ключей
L1 – Политики как код	Детекторы и политики (YAML/DSL), KMS/HSM, базовая документация	Нет гейтов и метрик полезности, слабый контроль артефактов	Автотесты приватности и полезности, пороги гейта, паспорт набора
L2 – CI/CD-гейты	Гейты допуска, метрики риска и качества, паспорт и lineage	Модельные тесты (MI, инверсия) точечны, DP отсутствует	Стандартные тесты утечек на моделях, DP на чувствительных узлах, синтетика с проверками
L3 – Приватность как платформа	DP с учётом бюджета (PLA), синтетика с гейтами, мониторинг дрейфа приватности, регулярные аудиты	Масштабирование практик и обучение команд	Privacy SLO/KPI, tabletop-учения, внешняя сертификация

Итоги



Инженерная практика

Приватность это инженерная практика: детекторы, политики, тесты, гейты, отчётность

Риск и полезность

Риск и полезность считаем вместе, решения только через гейт допуска

Дифференциальная приватность

Дифференциальная приватность – формальный слой там, где нужен гарантийный бюджет ε/δ

Синтетика и приватность

Синтетика ≠ приватность: три проверки – похожесть, полезность, приватность

Паспорт и lineage

Паспорт набора и lineage – обязательные артефакты каждого релиза

Вопросы?

Силаев Юрий Владимирович

https://t.me/SilaevYV

ysilaev@hse.ru

+7-929-533-3406