Методы автоматизации сбора и обработки статистических данных

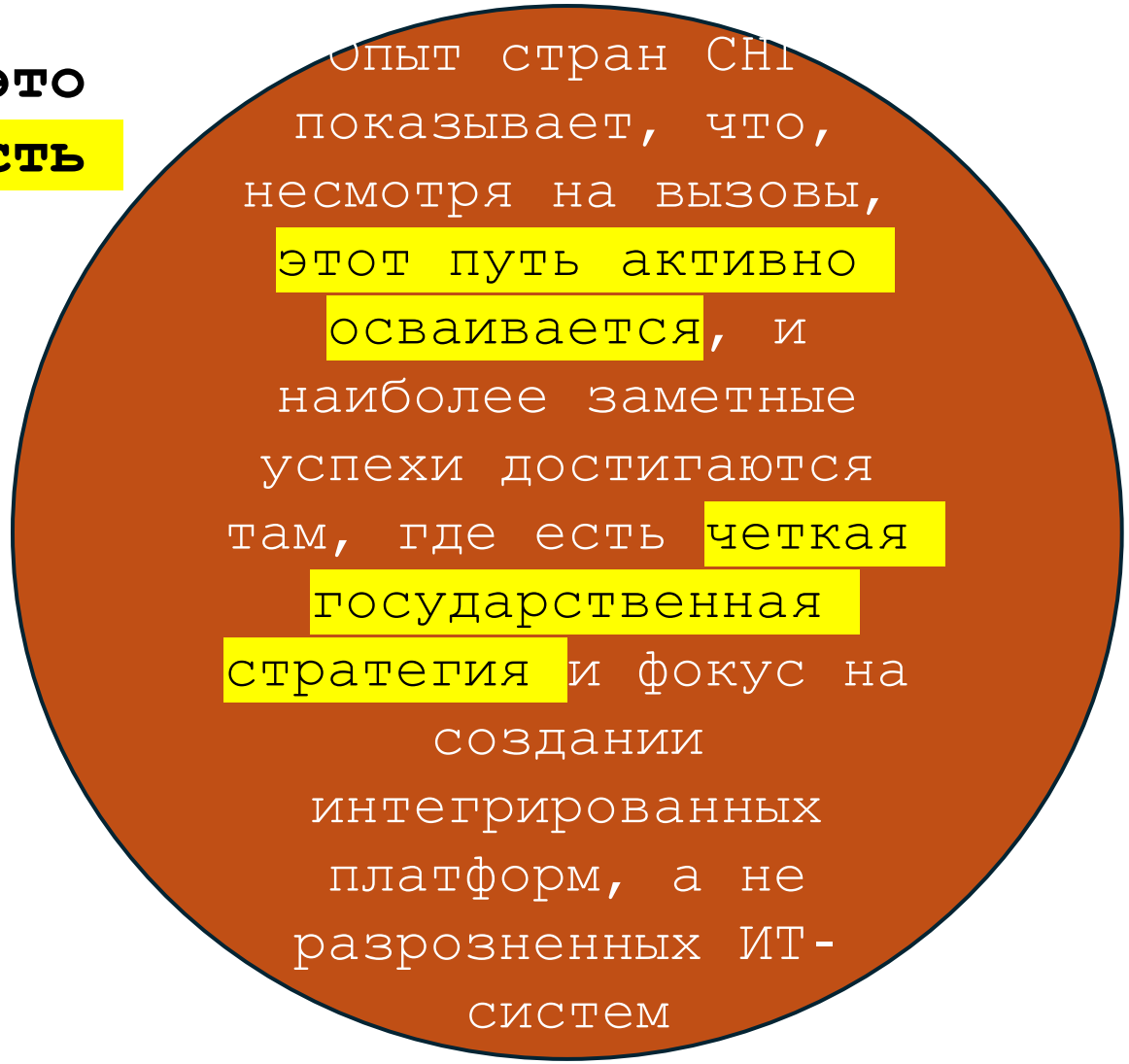
Лекция 2

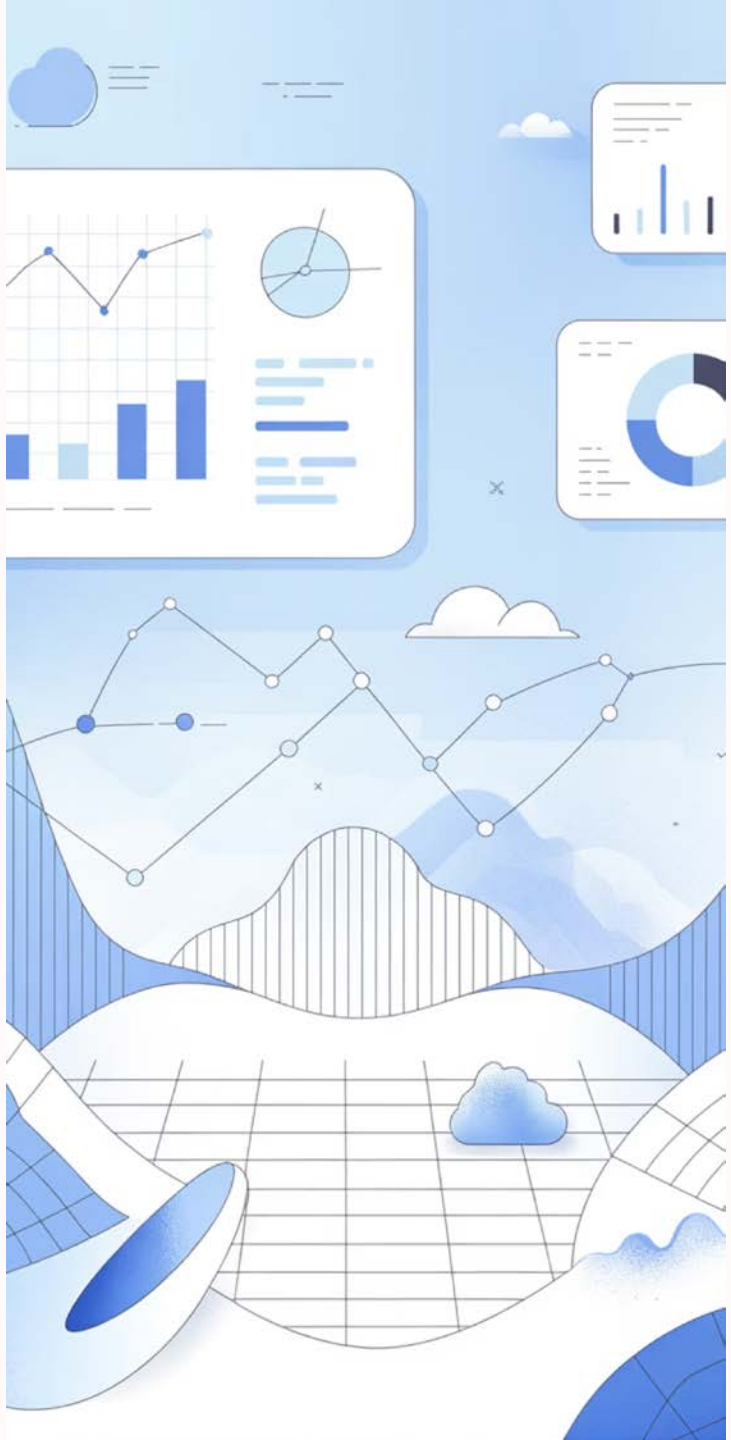
Зарова Елена Викторовна, доктор экономических наук. профессор

План лекции

- Использование цифровых платформ, облачных решений и ИИ для автоматизации сбора и обработки данных
- Практические примеры: автоматизация переписей, обработка административных данных
- Теоретические основы: обработка больших данных, обезличивание и защита информации

Использование цифровых платформ, облачных решений и ИИ для автоматизации сбора данных – это не просто тренд, а необходимость для повышения эффективности государственного управления и конкурентоспособности бизнеса





Применение МО и ИИ на этапах Типовой модели производства статистической информации

Уровни 1 и 2: принципы, задачи, опыт официальной статистики стран СНГ

Введение: Значение МО и ИИ в современной официальной статистике

Рост рынка ИИ

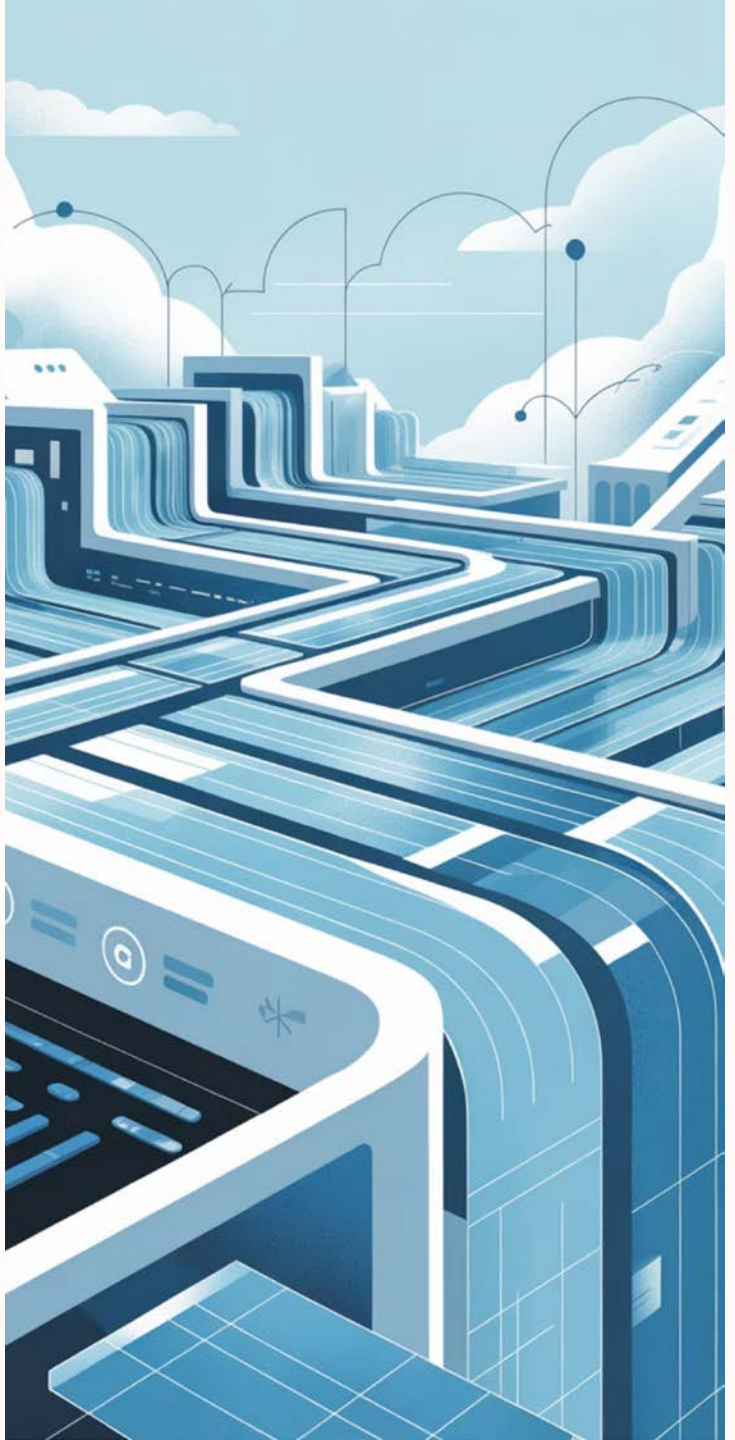
Рынок искусственного интеллекта в России достигнет 300 млрд рублей в 2024 году, становясь мощным драйвером цифровой трансформации экономики

Повышение качества

МО и ИИ позволяют существенно повысить качество, скорость обработки и точность статистических данных

Инновации СНГ

Страны СНГ активно внедряют передовые технологии в национальных статистических службах



Типовая модель производства статистической информации

Уровни 1 и 2

Первый уровень

Сбор и первичная обработка данных — ключевая база для всей статистической системы, требующая особого внимания к качеству

Второй уровень

Интеграция, верификация и подготовка агрегированных данных для последующего анализа

Роль технологий

МО и ИИ автоматизируют рутинные операции, выявляют аномалии и улучшают качество данных на всех этапах

ИИ и МО в стратегиях развития государственной статистики стран СНГ

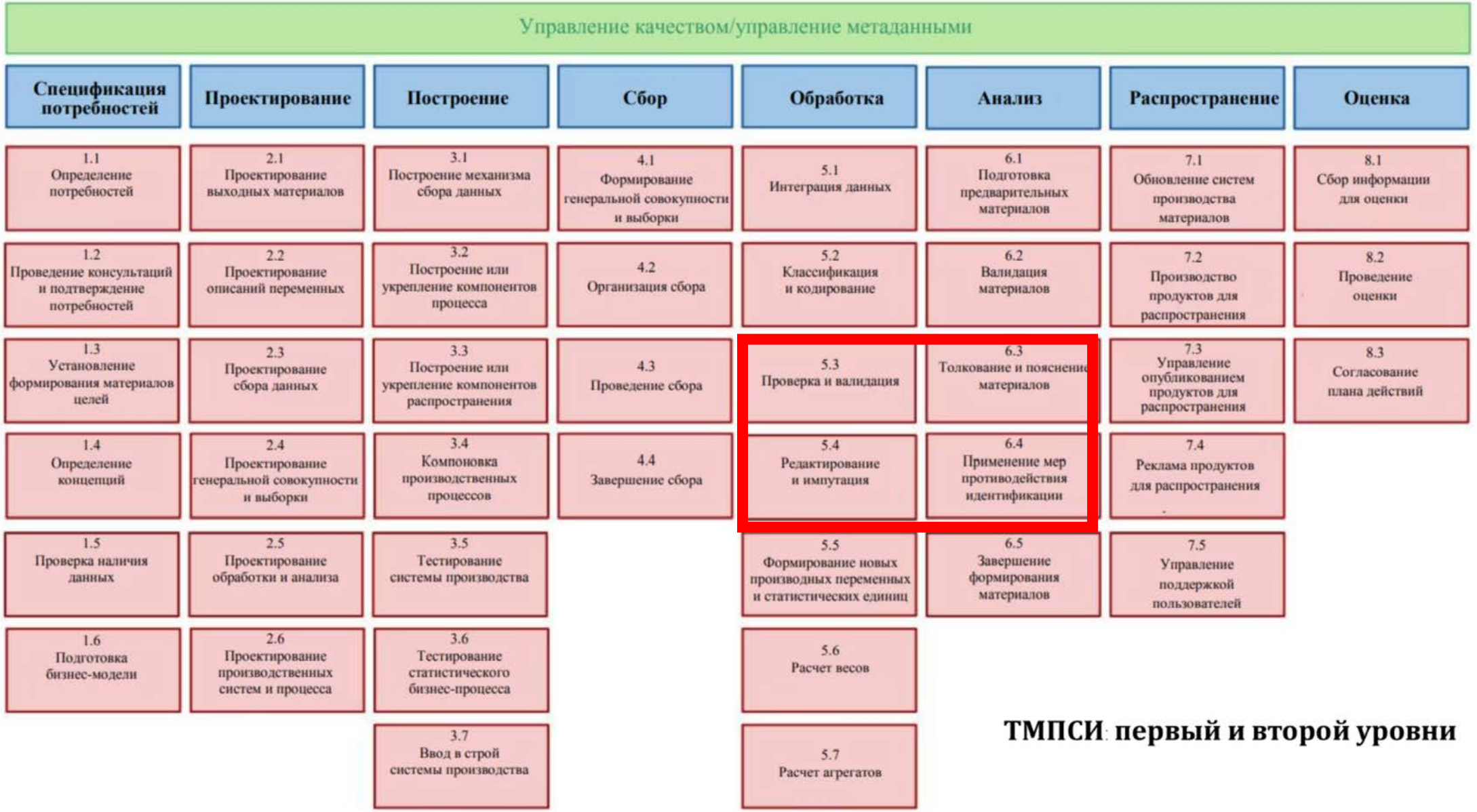
Страна	Наличие и ссылка на стратегию	Год утверждения	Упоминание ИИ/МО в стратегии
Азербайджан	**Да.** Государственная программа по развитию статистики в Азербайджанской Республике на 2018-2025 годы	2018	**Косвенное указание.** Упоминается развитие IT-инфраструктуры и использование "альтернативных источников данных", но прямого указания на ИИ/МО нет.
Армения	**Да.** Стратегическая программа развития статистики Армении на 2022-2026 гг.	2022	**Прямое указание.** Содержит задачу "исследовать и внедрять передовые технологии, [illegible]е как искусственный интеллект и машинное обучение, для [illegible]анных и

Сообщение для DeepSeek

Требует проверки

Беларусь	**Да.** Концепция развития государственной статистики до 2027 года	2020	**Косвенное указание.** Упоминается "анализ больших данных" и "совершенствование ИТ-инфраструктуры", но прямого указания на ИИ/МО нет.
Казахстан	**Да.** Концепция развития государственной статистики до 2026 года	2021	**Прямое указание.** Планируется "внедрение методов больших данных, машинного обучения и искусственного интеллекта для повышения оперативности, точности и эффективности статистических процессов".
Кыргызстан	**Да.** Стратегия развития государственной статистики Кыргызской Республики на 2019-2030 годы	2019	**Прямое указание.** Запланировано "использование методов интеллектуального анализа данных, машинного обучения и искусственного интеллекта для обработки больших ма ов информации".

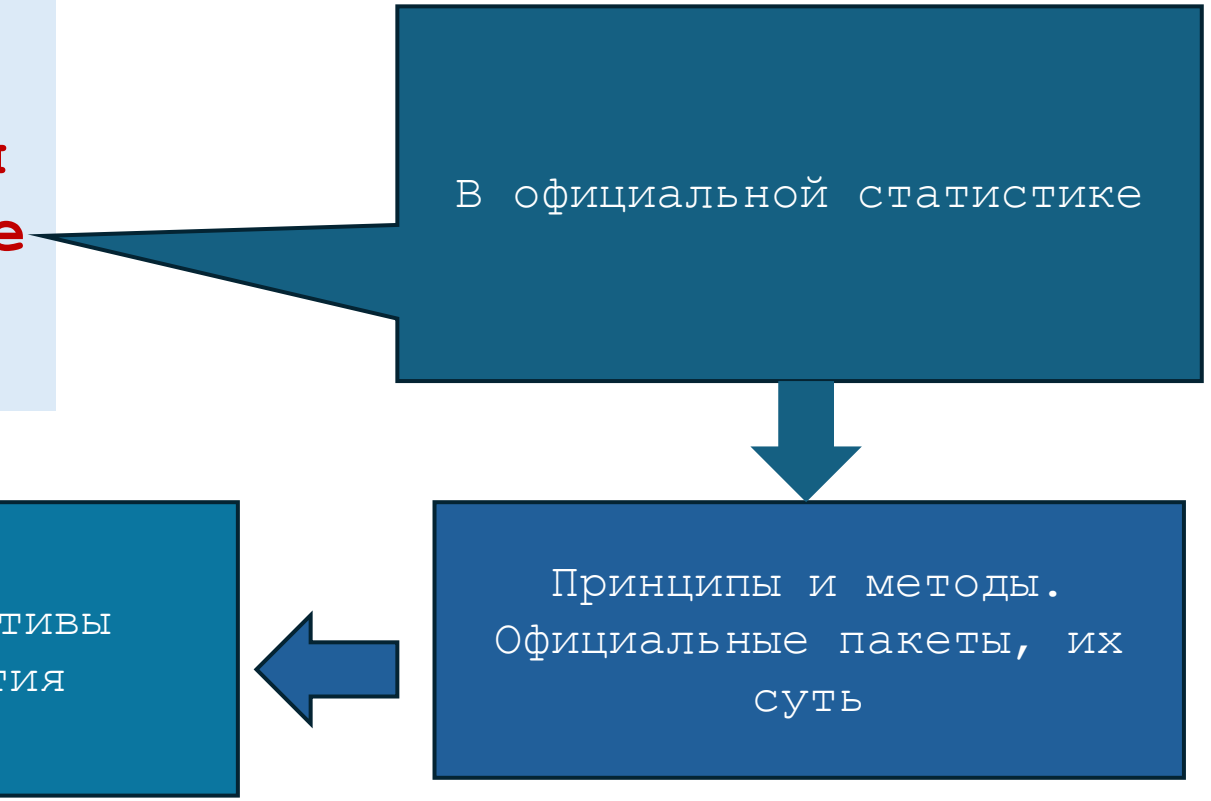
Молдова	**Да.** Стратегия развития национальной статистической системы на 2021-2025 годы	2021	"внедрении новых технологий, таких как искусственный интеллект и машинное обучение, для анализа больших данных и повышения эффективности производства статистики".
Россия	**Да.** Концепция развития государственной статистики на период до 2024 года и дальнейшую перспективу до 2030 года	2021	**Прямое указание.** Включено "использование методов машинного обучения и искусственного интеллекта для анализа данных, в том числе больших данных, и автоматизации процессов".
Узбекистан	**Да.** Стратегия развития государственной статистики на 2020-2025 годы	2020	**Прямое указание.** Включено "внедрение передовых методов, включая машинное обучение и искусственный интеллект, для анализа больших данных и повышения качества статистической информации".



ТМПСИ: первый и второй уровни

Выявление и устранение выбросов является важной производства статистической информации

Задача – рассмотреть применение методов МО и технологий ИИ при работе с выбросами



«Выбросы»

Выбросом (*outlier*) **считается значение в данных, которое находится далеко за пределами других наблюдений**

Экстремальные значения – это устранимые или неустранимые ошибки, возможные фиктивные значения

Статистический выброс – это наблюдение, которое существенно отклоняется от основной массы данных и расположено далеко от кривой плотности распределения вероятностей, к которой относятся основной объем данных. Формально, выброс – это точка данных с низкой вероятностью в данном распределении

Типы ошибок

Причины выбросов разнообразны:

- **Ошибки измерения**
- **Ошибки ввода данных**
- **Ошибки обработки данных**

ОШИБКИ РЕГИСТАЦИИ

- **Ошибки выборки**
- **Ошибки эксперимента**
- **Естественные выбросы**. ***Эти отклонения не являются ошибками, хотя и «выбиваются» на фоне остальных данных***

ОШИБКИ РЕПРЕЗЕНТАТИВНОСТИ

ROSA: выявление выбросов (одномерный и многомерный подходы, применение методов МО)

ROSA (R FOR OFFICIAL STATISTICS AND DATA ANALYSIS):

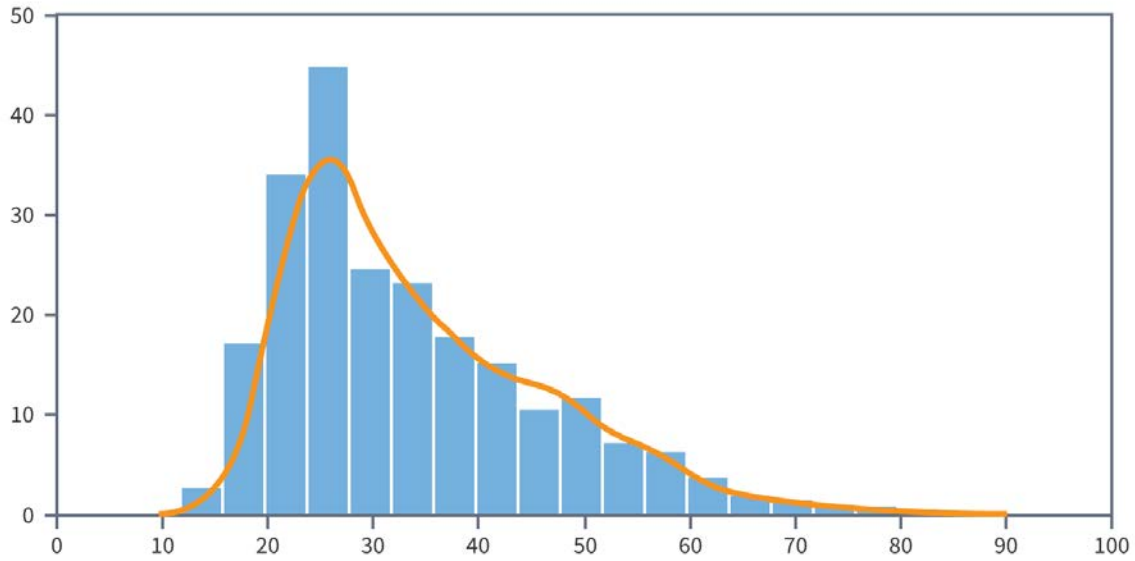
Создатель и целевая аудитория:

- **Разработчик:** Евростат (Eurostat), статистическая служба Европейского Союза.
- **Для кого:** Специально создан для национальных статистических офисов и официальных статистиков.

Основные функции пакета univOutl для одномерного выявления выбросов:

- LocScaleB(): Выявляет выбросы на основе робастного расположения и масштаба (аналог метода "медиана ± 3 MAD").
- QCD(): Использует робастный квартильный коэффициент дисперсии (QCD).
- adjbox(): Строит "скорректированные" диаграммы размаха (boxplot), которые лучше учитывают асимметрию данных.
- HDoutliers(): Алгоритм для обнаружения выбросов на основе теории больших отклонений (Heavy-Depth), эффективен для многомодальных распределений.

Асимметричное распределение



IQR – метод установления выбросов