



FEDERAL STATE
STATISTICS SERVICE

ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ ИИ И МАШИННОГО ОБУЧЕНИЯ ПРИ РАБОТЕ С НОВЫМИ ИСТОЧНИКАМИ ДАННЫХ

Климаев Станислав Николаевич

02.10.2024

Начальник Аналитического управления

КОНЦЕПЦИЯ ИСПОЛЬЗОВАНИЯ БОЛЬШИХ ДАННЫХ В СТАТИСТИКЕ

ROSSTAT



ПРЕИМУЩЕСТВА

- Меньше времени на подготовку данных
- Гораздо более подробная статистика
- Повышенное качество данных
- Снижение нагрузки на отчетность



ИСТОЧНИКИ ДАННЫХ

- Сотовые операторы
- Экосистемы (Яндекс, Mail.ru, СБЕР)
- Открытые библиотеки (OpenStreetMap, АИС «Реформа ЖКХ» и др.)
- ККТ
- Данные спутниковых снимков
- Данные РЖД и авиакомпаний

ПРИОРИТЕТНЫЕ ОБЛАСТИ



Статистика потребительских цен

- Статистическая практика
- Обследования
- Большие данные
- Административные данные



Статистика туризма

- Численность туристов
- Количество поездок
- Продолжительность пребывания
- Количество посетителей



Статистика розничной торговли и услуг

- Данные ККТ (ФНС России)
- Данные системы маркировки



Статистика населения

- Постоянное население
- Временное население
- Потоки маятниковых трудовых мигрантов



Статистика сельского хозяйства



ЦЕЛЬ

- Автоматизация процесса сбора цен
- Оперативность формирования показателей



ПРОБЛЕМЫ

- Полная автоматизация мониторинга цен
- Отсутствие единой маркировки товаров и услуг



РЕАЛИЗОВАНО

- Разработка новой методологии ИПЦ
- Разработка системы приема и обработки «Больших данных» в рамках ИПЦ
- Веб-скрейпинг
- Нейросеть для идентификации товаров
- Экспериментальный расчет по новой методологии



ПРЕИМУЩЕСТВА

- Точность наблюдений
- Репрезентативность показателей
- Актуализация системы весов для расчета ИПЦ
- Увеличение количества товаров (услуг) представителей и ценовых котировок



ЦЕЛЬ

Отслеживание изменения стоимости небольшой группы схожих товаров вместо отслеживания отдельного конкретного товара



ИСТОЧНИКИ ДАННЫХ

Данные чеков
контрольно-кассовой
техники

~ **230 Гб**
ежедневно



ДЕЙСТВИЯ:

Предобработка данных

с целью ускорения и упрощения дальнейших операций

Формирование подвыборок

по ключевым словам

Обработка подвыборки

используется модель CountVectorizer, по результатам анализа которой получают вектора из названий позиций в чеках. Далее вектора кластеризуются.

Модель CountVectorizer

Тест

‘Картофель свежий’

‘Картофель мытый’

‘Морковь свежая’

Формирование векторов

Хеш-функция f

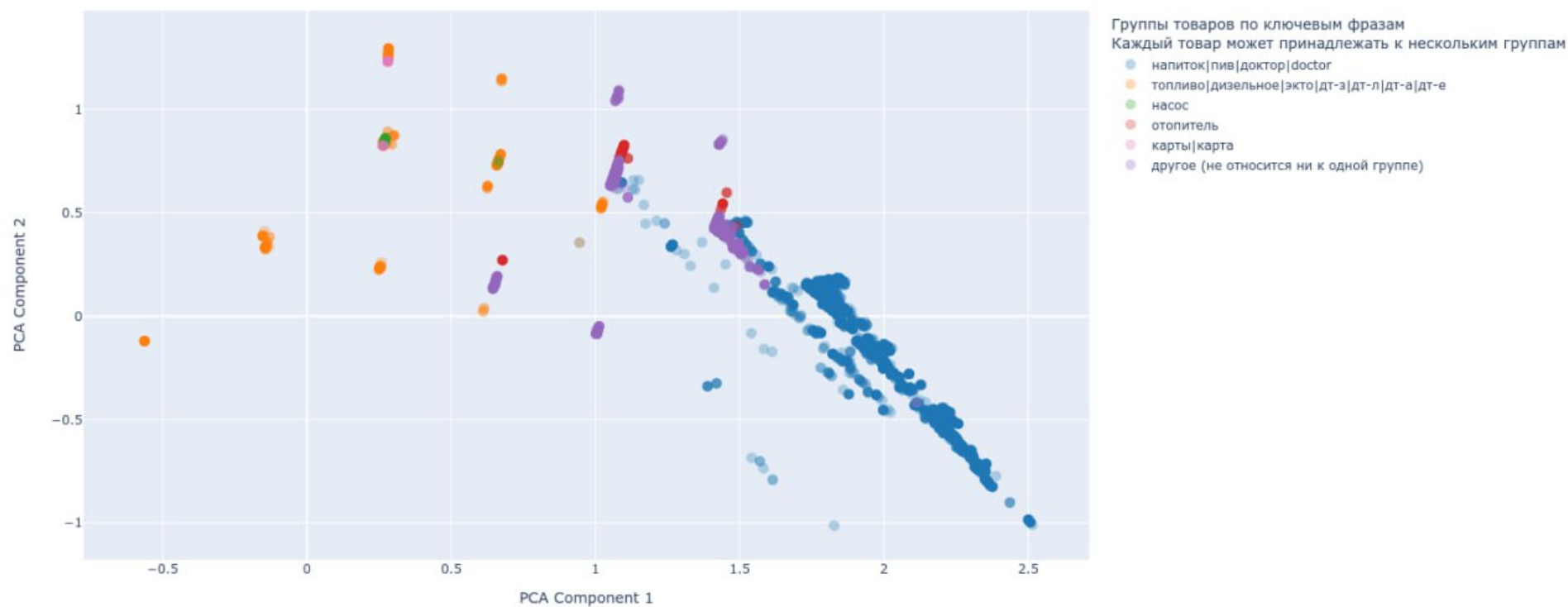
f(‘картофель’):1
f(‘морковь’):2
f(‘свежий’):3
f(‘мытый’):4

1	2	3	4
1	0	1	0
1	0	0	1
0	1	1	0

С целью оценки кластеризации, а также простоты восприятия и изображения информации – **понижается размерность до 2**.

Построен **интерактивный график** с возможностью просмотра предложения для каждой точки.

Анализ товаров с ключевым словом 'дизель' в названии после понижения размерности методом главных компонент (PCA).
Для векторизации названий товаров используется модель CountVectorizer





ЦЕЛЬ

- Переход на использование регистра населения и других административных источников данных
- Реестровые модели, цифровой профиль гражданина



ПРОБЛЕМЫ

- Низкая автоматизация использования первичных данных переписи в межпереписной период
- Отсутствие единого хранилища административных данных
- Использование геотреков повсеместно
- Двойной учет абонентов связи (несколько сим-карт)



ПРЕИМУЩЕСТВА

- Формирование методологической базы
- Увеличение частоты расчета отраслевых показателей
- Создание новых отраслевых показателей



РЕАЛИЗОВАНО

- Верификация данных Всероссийской переписи населения 2020 г. с данными сотовых операторов
- Ведение подсистемы «Демография»



ЦЕЛЬ

Использование модели геопространственных данных для формирования модели оптимального использования ресурсов



ИСТОЧНИКИ ДАННЫХ

- Переписи и микропереписи Росстата
- Обследования Росстата
- OpenStreetMap
- АИС «Реформа ЖКХ»

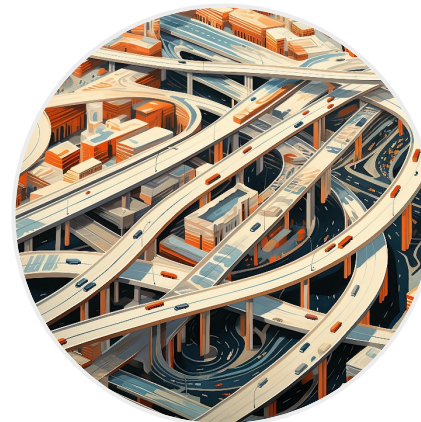
ВОЗМОЖНОСТИ:



Геопространственный анализ



Городское моделирование



Обработка больших массивов геоданных

Территориальные кластеры

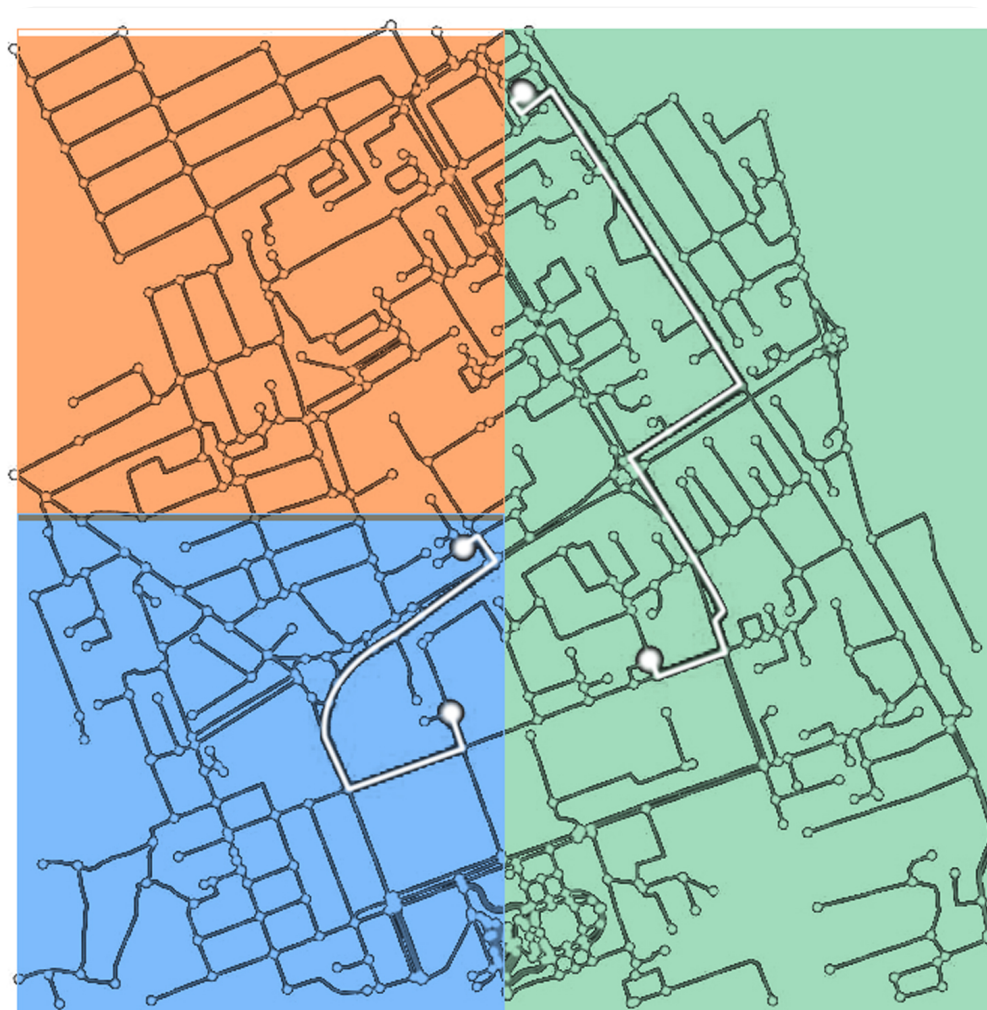
Адреса разбиваются на **кластеры**, включающие в себя всех респондентов, которых может обойти человек за один рабочий день.

Вершины и пути

Модель в основе имеет представление адресов в виде набора **вершин** и пути (дорог) между ними в виде **ребер**.

Оптимальный путь

На графе по алгоритмам (алгоритм Дейкстры и алгоритм Беллмана) можно найти **оптимальный** (самый короткий покрывающий наибольшее количество адресов) **путь** внутри каждого кластера и рассчитать **время**, которое уделяется на этот путь. **В случае, если время больше рабочего дня, то путь оптимально декомпозируется на более короткие пути.**





ЦЕЛЬ

- Верификация «Больших данных» в официальной статистике
- Типовая модель работы с данными сотовых операторов
- Организация правовой, финансово-экономической и технологической модели взаимодействия



ПРОБЛЕМЫ

- Неполнота охвата объектов наблюдения
- Отсутствие единой методологии
- Несопоставимость реестров
- Отсутствие единства понятийного аппарата



РЕАЛИЗОВАНО

Инвентаризация около 300 понятий, относящихся к отрасли



ПРЕИМУЩЕСТВА

- Обогащение методологии по расчету внутреннего туризма
- Совершенствование системы сбора, обработки и анализа статистических данных в рамках реализации Стратегии развития туризма в Российской Федерации на период до 2035 года
- Повышение частоты и детализации расчета отраслевых показателей
- Формирование новых показателей отрасли



НЕДОСТАТКИ

- Наличие необходимых **технических мощностей** не во всех регионах РФ
- **Недостоверность информации** о владельце мобильного устройства (общая информация об абоненте, наличие 2 и более SIM-карт разных операторов, частая смена места проживания и т.д.)
- **Неравномерное распределение** доли покрытия операторов сотовой связи в регионах РФ
- **Отсутствие единого ТЗ и понятийного аппарата** у всех операторов сотовой связи для распространения алгоритма расчета показателей в сфере туризма в регионы РФ
- Необходимость разработки **проекта методики, сбора, алгоритма верификации и агрегации данных** для расчета показателей в сфере туризма (разные справочники и структура данных у каждого оператора)



ПРЕИМУЩЕСТВА

- **Обогащение данных** федерального статистического наблюдения в сфере туризма (возможно частичная или полная замена);
- Выявление **основных точек притяжения** туристов/экскурсантов (российские и иностранные граждане);
- Отслеживание **частоты перемещений** абонентов в разрезе МО (регионов РФ) (въездные, внутренние и выездные поездки) и **длительность их пребывания** в определенных локациях;
- Создание **детального профиля абонента**, с возможностью моделирования возраста, пола, дохода, места жительства и работы, и многих других атрибутов;
- Возможность **настроить критерии отбора абонентов** в разрезе типов посещения для принятия различных управленческих решений

Альтернативный способ расчета по оценке туристического потока осуществляется на основе банковских данных.

Планируется заключение соглашения **«Мегафон»** с Правительством **Ставропольского края**



Федеральная служба государственной статистики (Росстат), Комитет по туризму города Москвы, СберАналитика подписали **соглашение о совместной работе** над методологией аналитики туристического потока. Сотрудничество сторон направлено на улучшение качества и оперативности формирования статистики, а это, в свою очередь, позволит повысить уровень услуг для путешественников.

ПОЛНОЕ ПОКРЫТИЕ ДАННЫМИ ПО РФ НА ОСНОВЕ ДАННЫХ КРЕДИТНЫХ ОРГАНИЗАЦИЙ

ROSSTAT

 **108 млн+**
физических лиц

 **6 млн+**
юридических лиц

 **10 млн+**
терминалов*
*более 6 млн с эквайрингом Сбера

 **7,7 млн+**
торговых точек

 **1,1 тыс.+**
городов
Российской Федерации

 **Все**
регионы
Российской Федерации

70+ источников данных
обогащение собственных данных
из внешних источников

60+ месяцев глубина аналитики
ретроспектива данных с ежедневной детализацией
ключевых показателей

Высокая точность и технологии

- Big data, machine learning, AI

Различные варианты представления исследований

- Инфо-панели
- Статические отчеты
- Дата-сети

*При анализе используются только агрегированные обезличенные данные. Персональная информация и любая информация, составляющая коммерческую тайну, не раскрываются.



Терминалы
эквайринга



Банковские карты
клиентов



Финансовые
продукты банка



Сведения
из мобильных
приложений



Зарплатные
проекты клиентов



Данные
расчетно-
кассовых счетов



Количественные
и качественные
исследования

01

Сбор данных

Источники данных:

- Терминалы
- Транзакции по картам
- Геоданные Сбербанка онлайн

02

Определение групп туристов

Дополнительные критерии:

- Критерий активности
- Критерий домашнего региона
- Критерий туристической поездки

03

Профиль туриста

Дополнительные данные:

- Социально-демографический профиль
- Агрегированные и обезличенные данные о финансах

04

Траты туристов

Дополнительные данные:

- До 50 категорий обезличенных и агрегированных трат туристов

05

Дооценка

Дополнительные коэффициенты:

- Коэффициент дооценки

ВАЛИДАЦИЯ РОССТАТОМ МОДЕЛИ ОЦЕНКИ ТУРИСТИЧЕСКОГО ПОТОКА

Многоступенчатый процесс, охватывающий этап сбора информации/данных, изучение модели, подготовку валидационной выборки, всесторонний анализ модели, завершающийся подготовкой отчета, фиксирующего выявленные слабые зоны модели и рекомендации по их возможному решению.

Выделяют **2 основных типа** валидации модели:

Первичная валидация

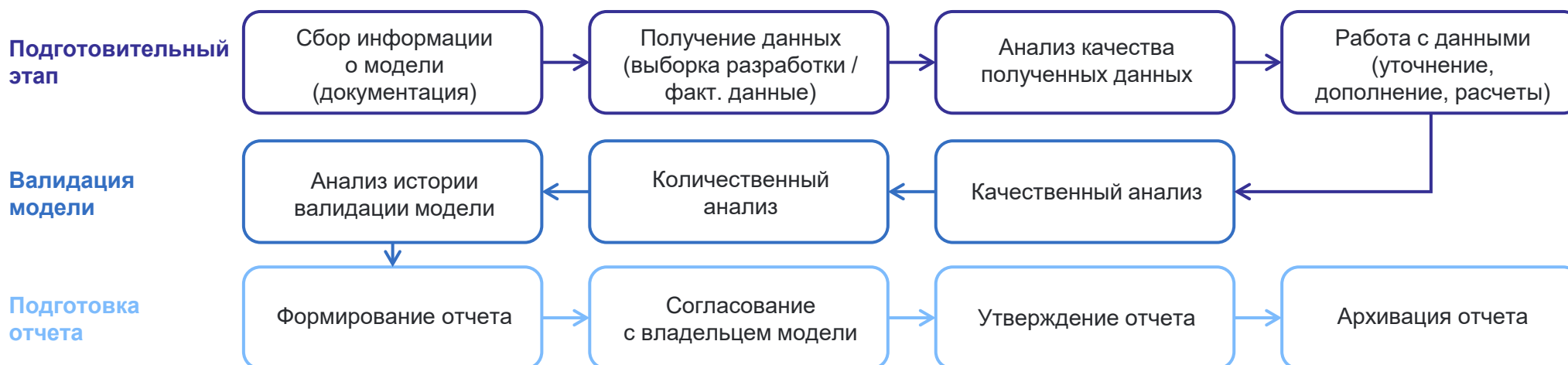
Осуществляется на этапе согласования новой модели, перед передачей готовой модели на утверждение в соответствии с Регламентом и ее внедрением.

Периодическая валидация

Анализ эффективности применения модели за определенный период времени после ее внедрения в промышленные системы Банка. Периодическая валидация проводится не реже 1 раза в год в соответствии с утвержденным планом по валидации.

Валидация модели проводится на актуальной валидационной выборке, отражающей структуру портфеля за соответствующий период.

Перечень шагов валидации модели



ЧТО МОЖНО УЗНАТЬ О ТУРИСТИЧЕСКОМ ПОТОКЕ ПО АНАЛИТИКЕ БОЛЬШИХ ДАННЫХ

ROSSTAT



Объем туристического
и экскурсионного потока



Географию туристов



Продолжительность поездки



Лояльные категории туристов



Социально-демографический
обезличенный профиль туриста



Финансовый обезличенный
профиль, поведение
и потребительские предпочтения



Оборот туристической индустрии



Точки притяжения: где больше
всего тратят туристы



ЦЕЛЬ

- Использование предварительно обученной языковой нейронной сети для получения векторных представлений каждого предложения



ПРОБЛЕМЫ

- Отсутствие единства в НСИ
- Дублирующие позиции в справочниках

В методе целевые и новые предложения проходят через обученную языковую нейронную сеть, результатом которой являются два вектора большой размерности. Далее между этими векторами рассчитывается косинусное расстояние.



Чем ближе к 1 — тем ближе по смыслу предложения



Вино и коньяк

Вина и коньяки	0,952
Коньяк и вино	0,974
Спиртные напитки	0,627
Продукт винный	0,725
Виноград	0,707



Унификация базы справочников



Единая НСИ



Уменьшение размеров
справочников



**Снижение
трудозатрат
на их обработку**



Очищенные данные с сохранёнными
векторными представлениями можно
использовать для поиска схожих
позиций при добавлении новой позиции
в справочник



**Предотвращение
дублирования элементов
справочника**

ЦЕЛИ ПРОЕКТИРОВАНИЯ И РАЗВИТИЯ GIGACHAT ПО СТАТИСТИКЕ ДЛЯ РОССТАТА

ROSSTAT

Значительное усиление роли Росстата в предоставлении статистической информации, улучшение взаимодействия с гражданами и обеспечение более эффективного решения поставленных задач



Цели:

1. Повышение доверия к данным, увеличение охвата и прозрачности:

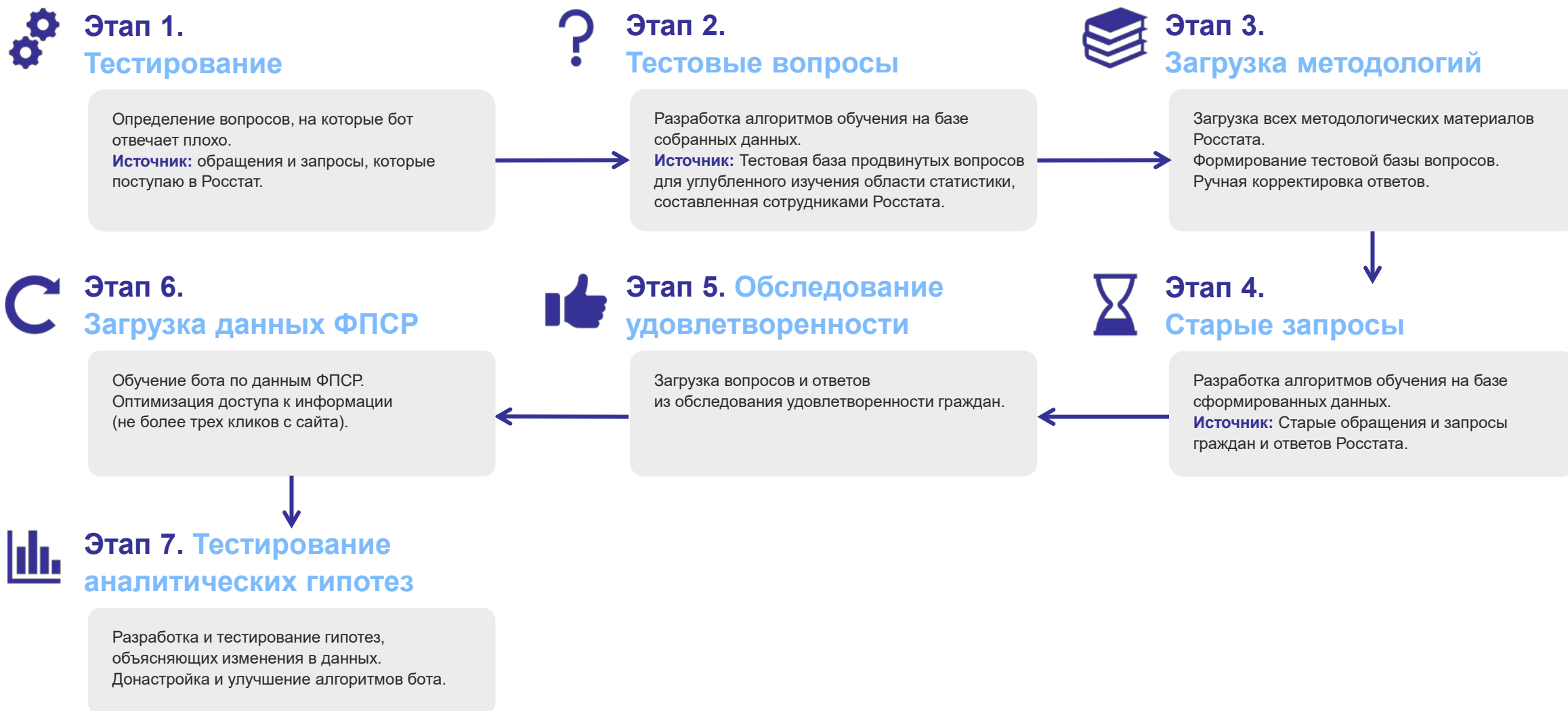
Обеспечить пользователей **достоверной и актуальной** статистической информацией.

2. Создание канал распространения официальной статистики:

Создать **удобный и доступный** канал для предоставления данных через GiGaChat по статистике.

3. Улучшение пользовательского опыта:

Оптимизировать взаимодействие с пользователями через их вопросы и запросы для **повышения качества** работы, с целью удовлетворения потребностей.



THANK YOU FOR YOUR ATTENTION!