



«Арктика – территория цифровизации»  
г. Мурманск 30 октября 2025 г.

# Извлечение больших открытых данных для организации и поддержки проведения научных исследований

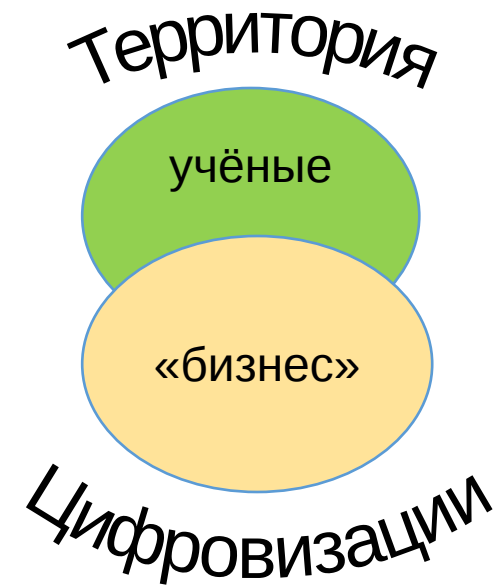
практический опыт  
и результаты регионального сотрудничества

Федеральный исследовательский центр «Кольский научный центр РАН»,  
Институт информатики и математического моделирования им. В.А. Путилова,  
ИИММ КНЦ РАН [www.iimm.ru](http://www.iimm.ru) ,  
зам.директора [Федоров А.М.](mailto:a.fedorov@ksc.ru), уч.секретарь [Датьев И.О.](mailto:i.datyev@ksc.ru),  
[a.fedorov@ksc.ru](mailto:a.fedorov@ksc.ru), [i.datyev@ksc.ru](mailto:i.datyev@ksc.ru)



## О главном

- ❖ Практические наработки
  - система мониторинга
  - технологии работы с открытыми данными социальных сетей
  - аналитика и технологии ИИ
- ❖ Внешнее техническое задание
  - пользовательский интерфейс
  - искусственный интеллект
  - быстрее и дешевле рынка
- ❖ Результаты совместной деятельности
  - доступ к сервису мониторинга
  - данные для отчетов по мониторингу
  - налаживание рабочих контактов
- ❖ Перспективы взаимовыгодного развития
  - обоюдное взаимопонимание целей и задач
  - отработка механизмов взаимодействия
  - неисчерпаемый потенциал совместной работы



## Системный анализ и концептуальное моделирование: структура, динамика,...

Роль аспектов "динамика", "риск" и "неопределенность" в (информационной, мониторинго-центрической) модели бизнес-сообщество-власть (БСВ)

### - динамика

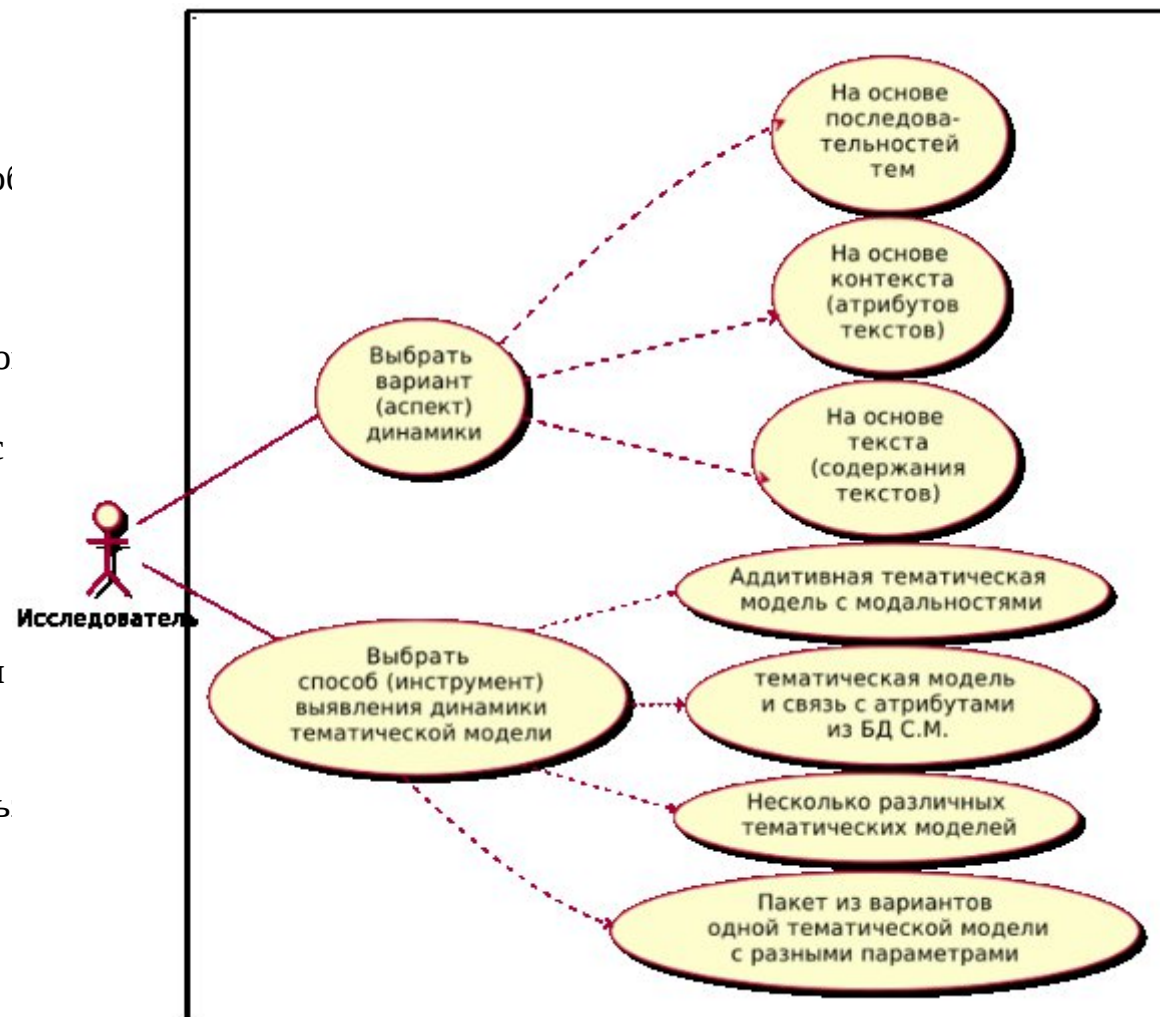
- информационная повестка постоянно меняется во времени
- отношение элементов БСВ друг к другу тесно связаны с изменяющейся информационной повесткой
- необходим постоянный мониторинг информационной повестки, для того, чтобы опираться на правильную картину мира

### - риск

- критическая скорость и пороги изменения информационной повестки
- риск - "упустить" правильную/адекватную/объективную информационную по
- риск - на основе недостоверной информационной повестки начать неправильно/некорректно/неэффективно/контрпродуктивно взаимодействовать с компонентами БСВ

### - неопределённость

- все ли источники информации использованы?
- задержка в информационной реакции
- информационная повестка получена из мониторинга, но ее не достаточно для формирования "картины мира"
- информационная повестка НЕ получена из мониторинга
- в информационной повестке еще нет "отголосков" того, что принято и исполн другими компонентами БСВ
- информационная повестка кем-то/чем-то намеренно искажается
- внутреннее представление о компонентах БСВ входит в противоречие с представлением, полученным на основе мониторинга



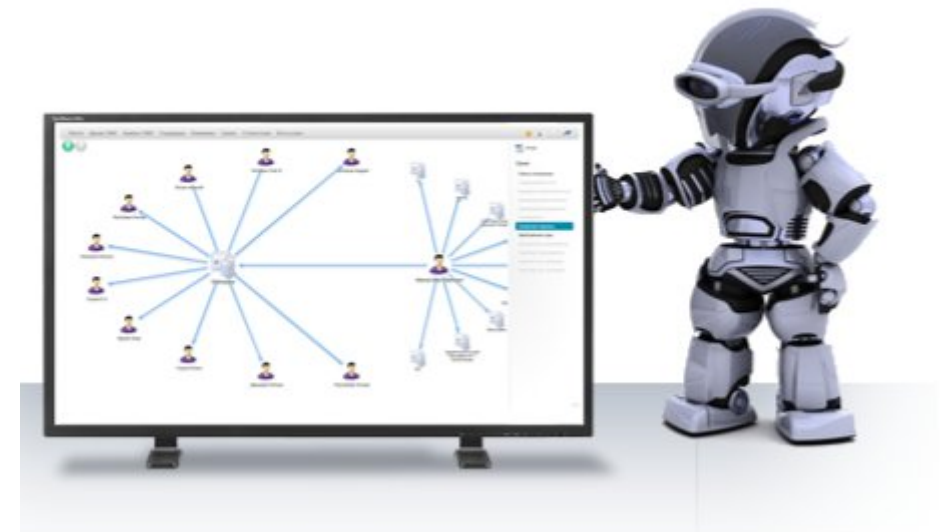


- Первой технологией, позиционированной **именно для получения данных из Интернет**, является *web scraping* — технология извлечения данных со страниц веб-ресурсов, которая чаще всего представляет собой автоматизированный процесс выполнения программным кодом *GET*-запросов к целевому веб-ресурсу.
- Наиболее известные сканирующие виртуальные роботы (боты): *Xenon*, *BingBot*, *Googlebot*, *Yandex*, *ChatGPT* и др.
- Веб-сканирование предлагается и в виде услуги: программное обеспечение как услуга или данные как услуга. Эти услуги позволяют автоматически собирать любые общедоступные данные в Интернете. Сканирование используется для агрегирования данных — процесса, позволяющего извлекать, преобразовывать, анализировать и визуализировать данные из нескольких источников.
- **К особенностям, осложняющим сбор данных, относятся динамически загружаемый и дублирующийся контент, а также защита от ботов.**

### Методы фокусировки при обработке веб-документов

- *Фокусированный сканер* — это сканер, который собирает веб-страницы, удовлетворяющие определённому свойству, например, сканировать страницы только определённого домена, или нечёткими, например, сканировать страницы о футболе или сканировать страницы с большими значениями рейтинга.
- Темы страницы - важное свойство, которое привело к появлению термина «*тематический сканер*». Сбор данных также может производиться по заданной эмоциональной окраске текста – тональности. В *семантическом сканере* для фокусировки используется информация о семантике, чаще всего — онтологии ПрО для представления тематических карт и связывания веб-страниц с соответствующими онтологическими концепциями, что позволяет производить категоризацию веб-документов.
- **Целью применения методов фокусировки является повышение объёма обладающих определёнными характеристиками собранных данных и сокращение времени сбора с учётом необходимости обхода блокировок со стороны администраторов веб-ресурсов. В фокусированные сканеры для повышения эффективности сбора данных всё чаще используются алгоритмы из области искусственного интеллекта.**

## База: информационная роботизация





## Информационный роботизированный мониторинг: фокусировка, адаптивность, ..

Доступность больших открытых данных для проведения мониторинга

- Интернет ; Специализированные приложения ; Веб-приложения

Роботизация информационного мониторинга

- SELENIUM - инструмент роботизации
- - Альтернатива на платформе Node.js - Puppeteer

Разметка css как средство идентификации объектов мониторинга

- DOM, HTML, XPATH и пр. как средство фокусирования мониторинга

Практический мониторинг чатов ватсап:

- Выявление разметки интересующих элементов
- Конфигурирование роботов для работы с разметкой
  - вариант: адаптивная самоконфигурация в процессе работы
- Организация сетевой (агентной) инфраструктуры для мониторинга
- Мониторинг
  - оперативная реакция на события:
    - оповещения оператору, ...

Способы конфигурации роботов для мониторинга

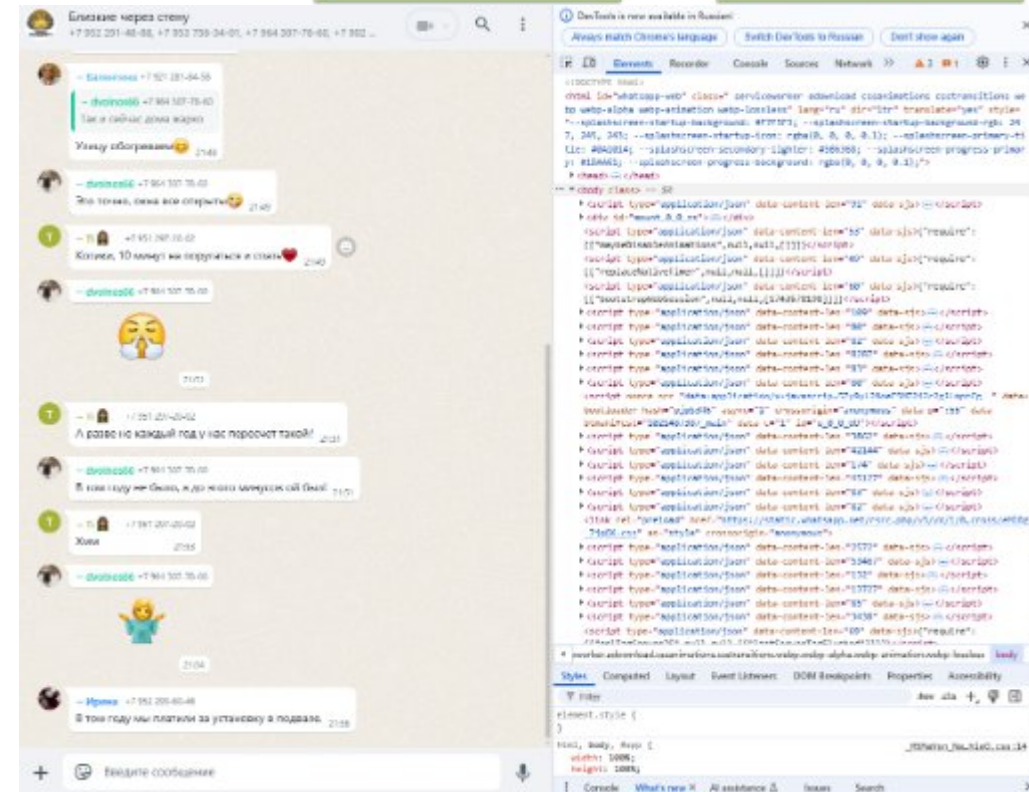
- - программист, на этапе конфигурирования
- - способ «письмо себе»
- - ии, промпт-инжиниринг

Очистка (восстановление) данных от разметки

Проектирование прикладных сетевых мультиагентных систем:

- агенты - докеры;
- взаимосвязанный сеть/комплекс докеров: бекэнд, фронтэнд, селениум, мониторинг;
- размещение автономных комплексов на разных серверах
- взаимодействие комплексов: обмен данными, согласованное управление ...

Сетевая централизация обеспечивает лучшие показатели фокусировки



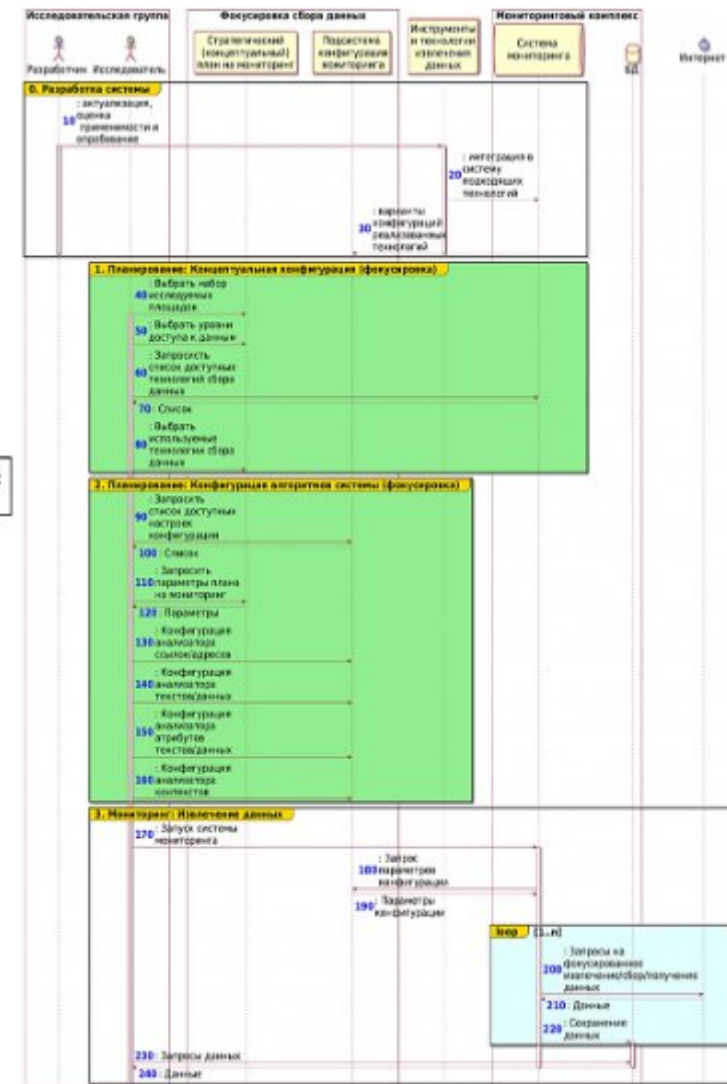
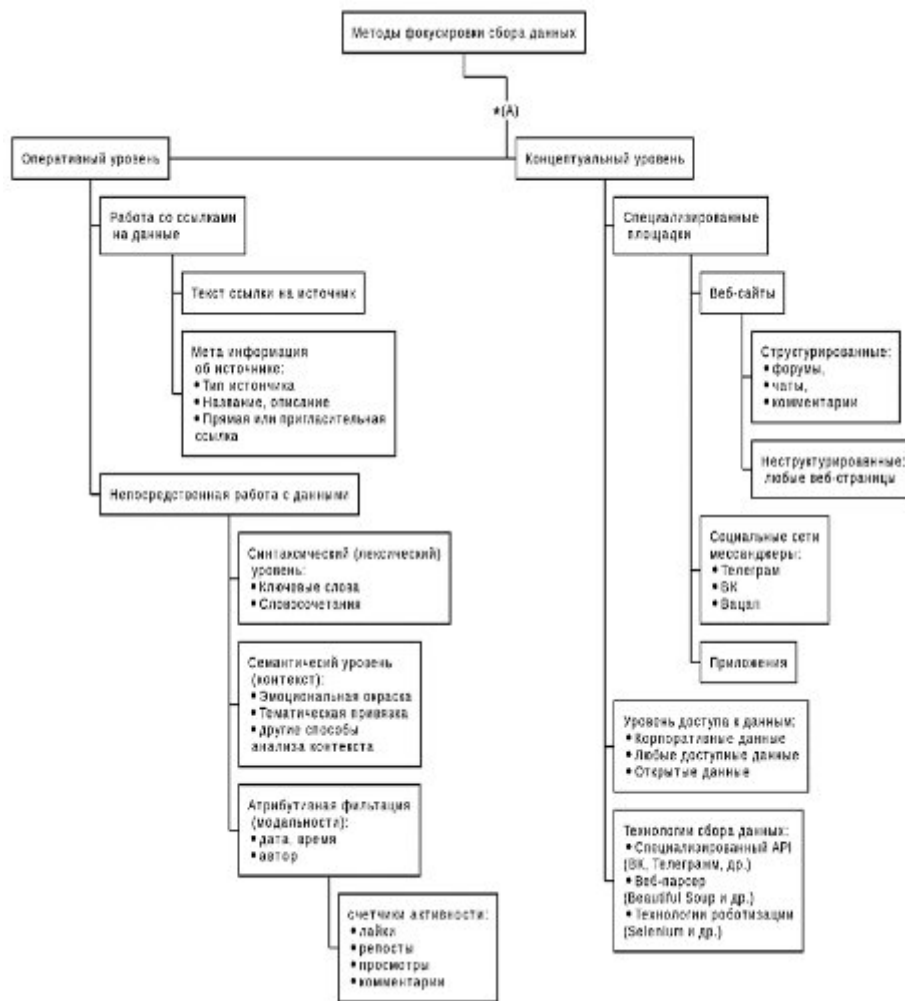


## Фокусировка извлечения данных: обобщение, концептуализация, новые задачи

Информационная поддержка принятия решений в сфере управления регионом на основе анализа больших открытых данных онлайн источников (в т.ч. социальных сетей):

Разработка методов и технологий фокусированного сбора и интеллектуализированного извлечения информации из открытых онлайн источников (в т.ч. социальных сетей) в рамках исследовательского поиска для поддержки принятия решений в сфере государственного и корпоративного управления.

- Анализ и систематизация современных технологий и актуальных программно-технических средств фокусированного сбора и интеллектуализированного извлечения информации из открытых онлайн источников.
- Разработка компонентов комплексной технологии моделирования, хранения и автоматизированного роботизированного доступа к открытым онлайн источникам.
- Создание комплекса концептуальных моделей, профильных хранилищ данных и прототипов программных средств для автоматизации исследовательского поиска при решении отдельных задач государственного и корпоративного управления.



## Технологии исследовательского поиска: семантические и геоданные,...

### Определение типа пользователей

Есть лог системы [[проект-sman]]

- последовательность действий пользователя в интерфейсе

На основе лога динамически определять тип пользователя:

- тематическое моделирование, формирование псевдо-текстов на основе логов работы аккаунтов

-полученная тематическая модель используется для определения потенциальных типов пользователей других аккаунтов.

– Для таргетирования рекламы.

– Типов пользователей может быть разное количество. Можно по-разному "нарезать" лог.

– **В зависимости от типа пользователя будут формироваться интерфейсы.**

### Использование карты (или ГИС)

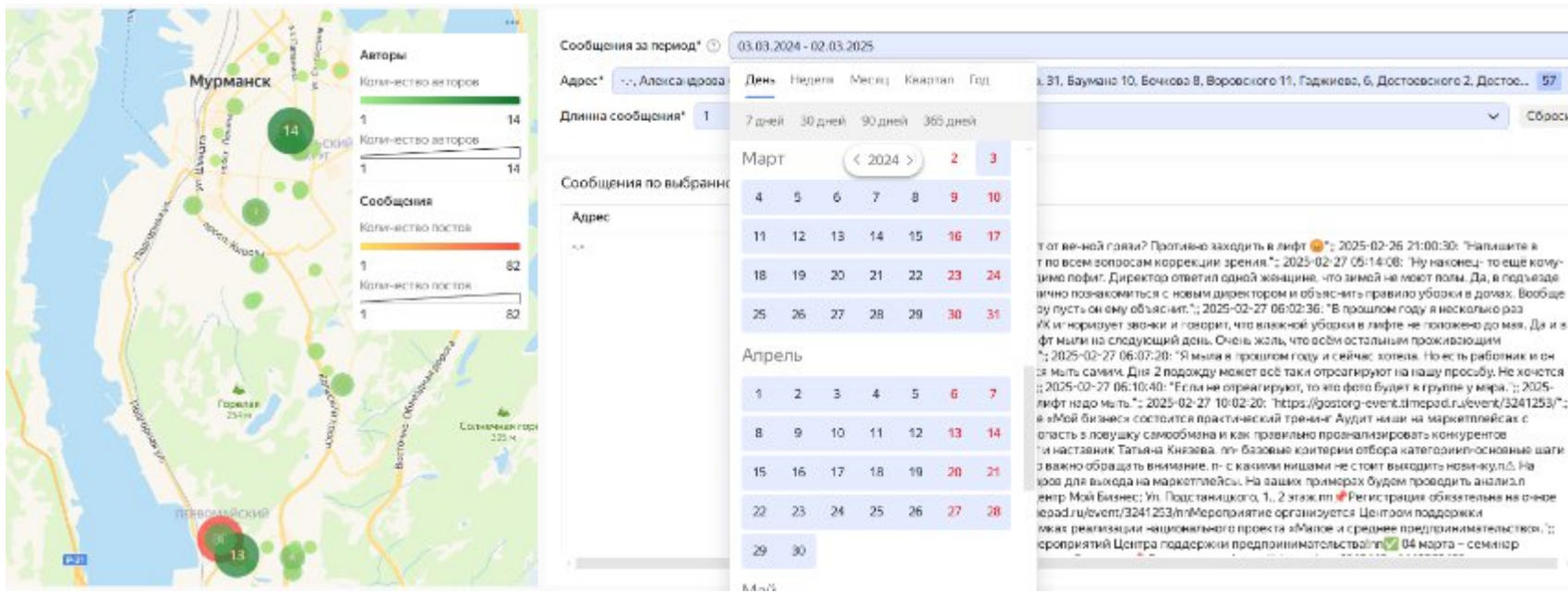
Для работы с геоданными полнофункциональная ГИС не нужна.

– Яндекс-DataLens - инструмент с возможностью построения дашбордов , одним из типов которых является "Карта"

– Для отображения точки на карте нужна только "Геоточка"

На карту можно нанести произвольную семантическую информацию из практически любых источников - SQL базы, файлы с таблицами и пр.

– Узкое место - процедура геокодирования - т.е. перевод строки с адресом (названием места) в координаты



## Технологии искусственного интеллекта *или искусственный интеллект*

### Обывательское представление:

- умные компьютеры
- гениальные результаты
- людская безработица
- восстание машин

### Планы применять супер-технологий в народном хозяйстве:

- GPT
- чат-боты
- языковые модели
- генераторы текстов/картинок/идей/...

### Особенности:

- вероятностный характер, недетерминированность;
- зависимость от обучающей выборки;
- системные пред-промпты;
- поддержка специалистов;

### Практические аспекты:

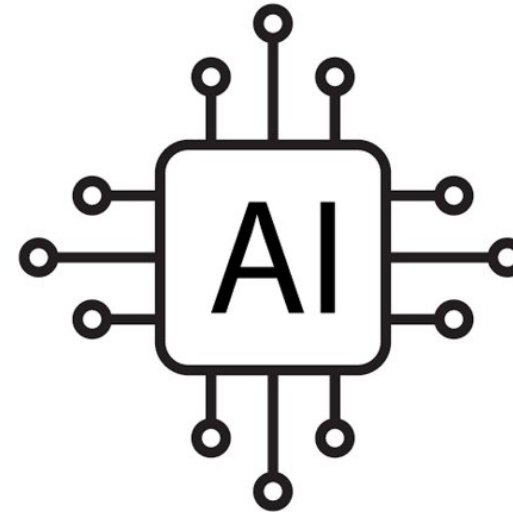
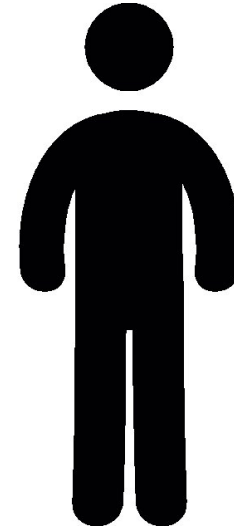
- «Алиса сказала...»
- «ГигаЧат подтвердил ..»
- «ЯндексGPT разложил всё по полочкам»

### Разработчики:

- ИИ – Имитация Интеллекта;
- точечное применение технологии

### Ученые:

- «вау»-эффект
- предварительный системный анализ



## Бизнес и наука : общий знаменатель

### Ключевые показатели эффективности:

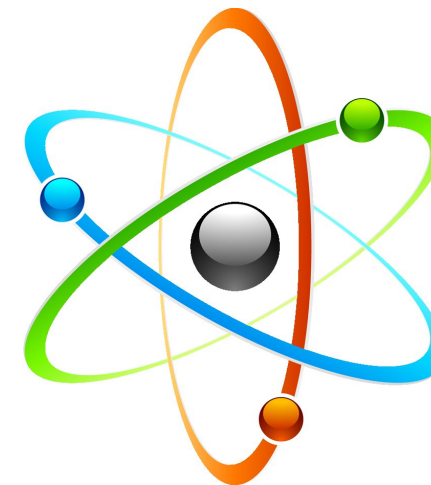
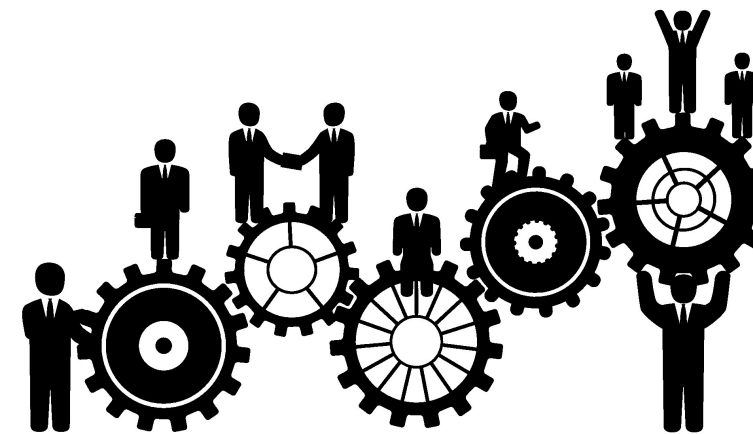
- **Бизнес:**
  - снижение затрат
  - повышение доходов
  - общественные связи, ...
- **Наука:**
  - публикации
  - привлеченные средства
  - популяризация науки, образование, ...

### Перспектива сотрудничества:

- взаимодействие на альтернативной основе;
- региональные интересы;
- образовательная синергия, ...

### Объективная реальность:

- **Бизнес**
  - ИТ-разработка как бизнес
  - движение представителей ИТ-индустрии в регионы
  - персональные предложения для ЛПР
- **Наука:**
  - фундаментальная абстрактность;
  - технологическая некомпетентность;
  - кадровая неполнота;
  - неготовность к рискам





## Разработки на основе научных результатов

1. Кулешов С.В., Зайцева А.А., Левашкин С.П. Технологии и принципы сбора и обработки неструктурированных распределенных данных с учетом современных особенностей предоставления медиа-контента // Информатизация и связь. 2020. № 5. С.22-28. DOI 10.34219/2078-8320-2020-11-5-22-28. EDN FMQNTT.
2. Кулешов С.В., Зайцева А.А. Феноменологическое описание процессов сбора и обработки интернет-документов // Изв. вузов. Приборостроение. 2023. Т.66, № 12. С.1002-1010. DOI:10.17586/0021-3454-2023-66-12-1002-1010.
3. Федоров А.М., Датьев И.О., Вишняков И.Г. Проектирование информационной системы комплексного тематического анализа больших данных социальных медиа // Онтология проектирования. 2024. Т.14, №1(51). С. 55-70. DOI:10.18287/2223-9537-2024-14-1-55-70.
4. Датьев И.О., Федоров А.М., Ревякин А.А. Фокусированный сбор и обработка открытых данных социальных медиа. Онтология проектирования. 2024. Т.14, №4(54). С.569-581. DOI:10.18287/2223-9537-2024-14-4-569-581.
5. Олейник А. Г., Федоров А. М., Датьев И. О., Зуенко А. А., Шестаков А. В., Вишняков И. Г. Об использовании RAG-технологии для исследовательского поиска в справочных и нормативных текстах // Труды Кольского научного центра РАН. Серия: Технические науки. 2024 Т. 15, № 3. С. 5–26. doi:10.37614/2949.1215.2024.15.3.001.
6. Датьев И.О., Федоров А.М. Аддитивная регуляризация при тематическом моделировании текстов сообществ онлайн-социальных сетей // Онтология проектирования. 2022. Т. 12, №2(44). С.186-199. DOI:10.18287/2223-9537-2022-12-2-186-199.
7. Федоров А.М., Датьев И.О. Конфигуратор системы мониторинга чатов ватсап // Роспатент: Свидетельство о государственной регистрации программы для ЭВМ №2024666744 от 16 июля 2024 г. (заявка №2024666117 от 11 июля 2024 г.); Регистрационный номер: 624112800028-8; [https://new.fips.ru/registers-doc-view/fips\\_servlet?DB=EVM&DocNumber=2024666744&TypeFile=pdf](https://new.fips.ru/registers-doc-view/fips_servlet?DB=EVM&DocNumber=2024666744&TypeFile=pdf)
8. Федоров А.М., Датьев И.О., Вишняков И.Г. Программный комплекс для построения, анализа результатов и визуализации динамических тематических моделей на основе данных социальных медиа // Роспатент: Свидетельство о государственной регистрации программы для ЭВМ №2023687646 от 18 декабря 2023 г. (заявка № 2023682819 от 30 октября 2023 г.); ЕГИСУ: 624011802376-4
9. Федоров А.М., Датьев И.О. Анализатор списка источников для мониторинга социальных сетей // Роспатент: Свидетельство о государственной регистрации программы для ЭВМ №2024667647 от 26 июля 2024 г. (заявка №2024666529 от 11 июля 2024 г.); Регистрационный номер: 624112800033-2; [https://new.fips.ru/registers-doc-view/fips\\_servlet?DB=EVM&DocNumber=2024667647&TypeFile=pdf](https://new.fips.ru/registers-doc-view/fips_servlet?DB=EVM&DocNumber=2024667647&TypeFile=pdf)

РОССИЙСКАЯ ФЕДЕРАЦИЯ **RU** **2024667647**



ФЕДЕРАЛЬНАЯ СЛУЖБА  
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) ГОСУДАРСТВЕННАЯ РЕГИСТРАЦИЯ ПРОГРАММЫ ДЛЯ ЭВМ

Номер регистрации (свидетельства): <b>2024667647</b>	Авторы: Федоров Андрей Михайлович (RU), Датьев Игорь Олегович (RU)
Дата регистрации: <b>26.07.2024</b>	Правообладатель: ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР "КОЛЬСКИЙ НАУЧНЫЙ ЦЕНТР РОССИЙСКОЙ АКАДЕМИИ НАУК" (RU)
Номер и дата поступления заявки: 2024666529 11.07.2024	
Дата публикации и номер бюллетеня: <b>26.07.2024</b> Бюл. № 8	
Контактные реквизиты: fedorov@iimm.ru	

Название программы для ЭВМ:  
Анализатор списка источников для мониторинга социальных сетей

РОССИЙСКАЯ ФЕДЕРАЦИЯ **RU** **2024666744**



ФЕДЕРАЛЬНАЯ СЛУЖБА  
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) ГОСУДАРСТВЕННАЯ РЕГИСТРАЦИЯ ПРОГРАММЫ ДЛЯ ЭВМ

Номер регистрации (свидетельства): <b>2024666744</b>	Авторы: Федоров Андрей Михайлович (RU), Датьев Игорь Олегович (RU)
Дата регистрации: <b>16.07.2024</b>	Правообладатель: ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР "КОЛЬСКИЙ НАУЧНЫЙ ЦЕНТР РОССИЙСКОЙ АКАДЕМИИ НАУК" (RU)
Номер и дата поступления заявки: 2024666117 11.07.2024	
Дата публикации и номер бюллетеня: <b>16.07.2024</b> Бюл. № 7	
Контактные реквизиты: fedorov@iimm.ru; +79212844848	

Название программы для ЭВМ:  
Конфигуратор системы мониторинга чатов ватсап

# спасибо за внимание

Теория и практика научно-исследовательских разработок для поддержки управления региональным развитием

Федеральный исследовательский центр «Кольский научный центр РАН»,

Институт информатики и математического моделирования им. В.А. Путилова,  
ИИММ КНЦ РАН [www.iimm.ru](http://www.iimm.ru) ,

зам.директора [Федоров А.М.](mailto:a.fedorov@ksc.ru), уч.секретарь [Датьев И.О.](mailto:i.datyev@ksc.ru),

[a.fedorov@ksc.ru](mailto:a.fedorov@ksc.ru), [i.datyev@ksc.ru](mailto:i.datyev@ksc.ru)