ИСП РАН

От идеи до внедрения
Истории успеха ИСП РАН

Арутюн Аветисян
arut@ispras.ru

02.12.2021

ОСНОВА ИСП РАН – ПОСТОЯННО РАЗВИВАЮЩАЯСЯ НАУЧНАЯ ШКОЛА

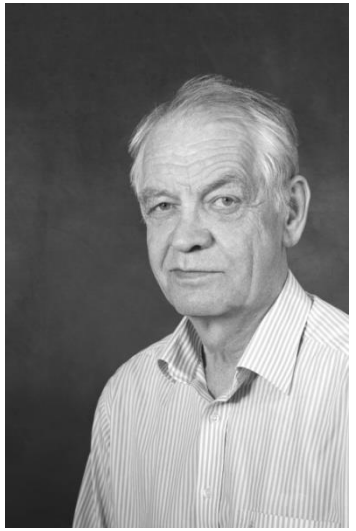
ИСП РАН

С.А. Лебедев

В.А. Мельников

БЭСМ-6 в музее науки Лондона

В.П. Иванников



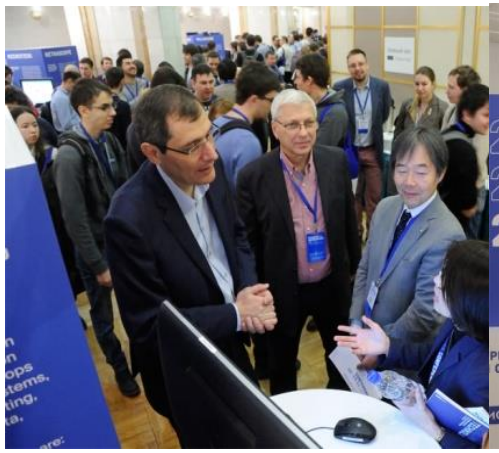
Л.Н. Королёв

«Если мы глубже разберёмся в этом эпохальном советском суперкомпьютере, это позволит пересмотреть заявления времён холодной войны об отставании русской технологии, а также подтвердить или развеять мифы о технологическом совершенстве наших союзников».

Doron Swade, senior curator of computing and information technology

2018: 70 лет IT*
2019: 25 лет ИСП РАН
2023: 75 лет IT*

***в России и странах постсоветского пространства**



2023: ИСП РАН проведёт COMputer, Software and Application Conference (COMPSAC) в Москве. Ключевая конференция IEEE CS, проводится ежегодно в течении 45 лет.

ИСП РАН

Цифровая экономика: приложения

Непрерывный доступ в сеть интернет/интранет

Киберфизические возможности

Большая вычислительная сложность

Умные устройства (дом, офис, завод)

Платформы «интернета вещей», искусственного интеллекта, …

Платформы хранения и обработки «больших» данных

Облачные платформы

Аппаратура

Проблемы современного системного ПО:

1.Эскалация размеров (Astra Linux – более 150 миллионов строк кода).
2.Сложность среды разработки и сборки.
3.Отсутствие изолированных систем.

Необходимые качества системного ПО:

Эффективность

Продуктивность

Безопасность

От фундаментальной* идеи до продукта
Примеры - важнейшие научные результаты РАН 2016-2020

*Наукоемкие инновации основа долгосрочного конкурентоспособного развития ИСП РАН

Svace. История разработки

ИСП РАН

Svace в 2021 году – I

- В 2020 г. добавлены языки Kotlin и Go
- В 2020 г. улучшен анализ чувствительных данных для поиска критических уязвимостей
- Новый графический интерфейс работы с результатами анализа

Svace в 2021 году – II

- С 2009 г. внедряется в компании Samsung в рамках совместной лаборатории
  - С 2015 года внедрен в Samsung и ее дочерние компании как основной статический анализатор
  - 300+ млрд. строк кода проанализировано, 10+ тысяч пользователей
- С 2020 г. ведется внедрение в Huawei и доработка анализатора для программ компании в рамках совместной лаборатории по статическому анализу
- Используется в более чем 30 российских компаниях (Лаборатория Касперского, Код Безопасности, Постгресс…)

ИСП РАН

Основные научные результаты

- Анализ потока данных на основе значений
- Контекстная чувствительность с использованием параметризованных резюме функций
- Чувствительность к путям на основе символьного выполнения с объединением состояний
- ***Алгоритмы Svace вошли в число лучших результатов РАН за 2016 г.***

ИСП РАН

Svace в цифрах

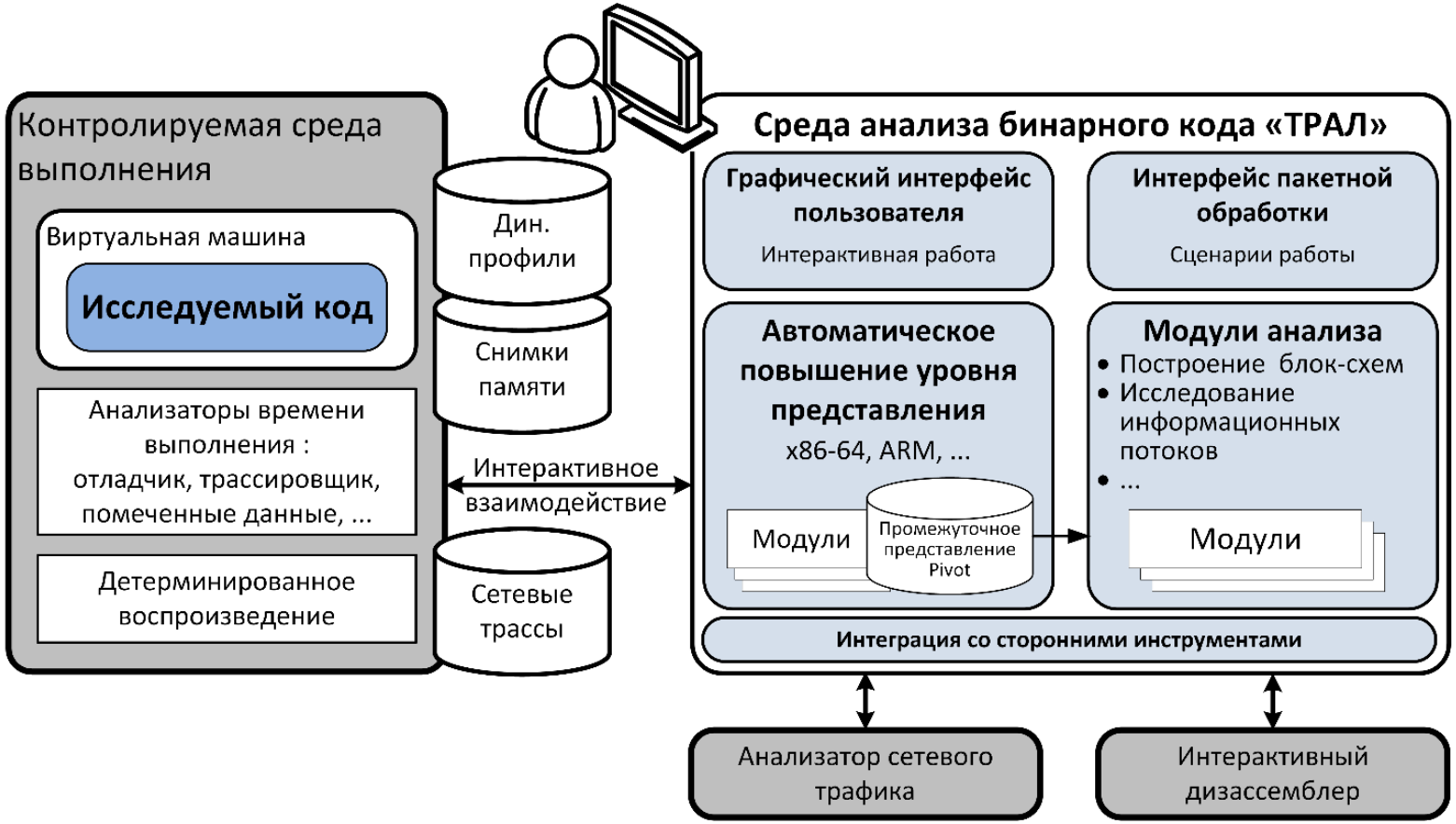
700+ тыс. строк кода на 10 языках

600+ детекторов

50+ статей и регистраций программ

5 кандидатских и 2 докторских диссертации

Среда анализа бинарного кода ТРАЛ



- Среда ТРАЛ включена в важнейшие результаты РАН за 2019 год

ИСП РАН

Трал. Корни

- Начало проектной работы – лето 2007
- Отдел компиляторных технологий ИСП РАН
- Опыт разработки
  - Среда ParJava
  - Статический анализатор Svace
  - Портирование сложных систем на другие ОС и процессорные архитектуры
  - Разработка специализированных программных инструментов

ИСП РАН

Трал

50+ статей и регистраций программ

7 кандидатских и 1 докторская диссертация

Инструменты жизненного цикла разработки безопасного ПО

- Статический анализ – одна из технологий, нужная для инструментальной поддержки жизненного цикла ПО
- Методы и алгоритмы, вошедшие в Svace, были дополнены методами динамического анализа (Crusher, ИСП Фаззер, Sydr, Трал), статического анализа бинарного кода (Binside), методами диверсификации и обфускации, безопасным компилятором на основе GCC
- Разрабатываются ГОСТы по применению статического и динамического анализа в практическом жизненном цикле безопасного ПО

Лаборатория «Системного программирования» в Великом Новгороде

- Открыта в 2010 году
  - Численность на момент открытия – 4 человека
  - 2021 год – 12 человек
- Программный эмулятор как средство контролируемого выполнения
  - Базовое средство – QEMU
  - Детерминированное воспроизведение
    - В 2015 году патчи приняты сообществом QEMU, лаборатория поддерживает механизм воспроизведения в официальном дистрибутиве эмулятора
  - Интроспекция виртуальной машины
  - Обратная отладка
  - Полносистемная отладка без поддержки со стороны гостевой ОС
  - Платформа динамического анализа

Лаборатория «Безопасное программное обеспечение» в Орле

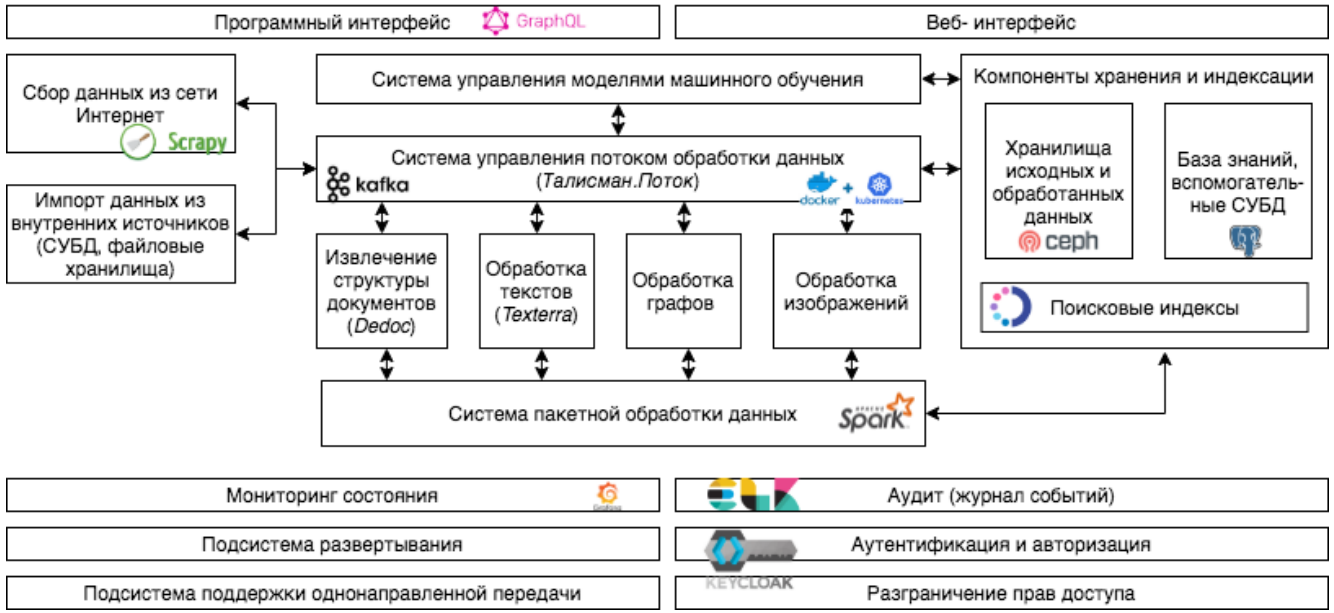
- Открыта в 2019 году
  - Численность на момент открытия – 4 человека
  - 2021 год – 9 человек
- Технологии разработки безопасного ПО
  - Практические задачи аудита безопасности
  - Внедрение инструментов анализа кода в учебный процесс
- Разработка тестового ПО для оценки программных инструментов
- Стеганография
- Анализ сетевого трафика

ИСП РАН

Учебная деятельность

- Архитектура ЭВМ и язык ассемблера
  - 1ый курс ВМК МГУ, часть сквозного курса программирования
- Информационная безопасность и анализ кода (Анализ кода и надежность программ)
  - Магистратура ВМК МГУ
- Информационная безопасность и компьютерные сети
  - Магистратуры ВМК МГУ, ФУПМ МФТИ, ФКН ВШЭ
- Проведение курсов повышения квалификации, в том числе – участие в совместных курсах с ФСТЭК России

Платформа Talisman



- Накопление знаний о предметной области
  - Построение баз знаний
- Создание мониторинговых систем
  - Мониторинг репутации персон и организаций
  - Выявление неявных информационных кампаний
- Конкурентная разведка
  - Поиск информации об объектах в открытых источниках
  - Выявление неявных взаимосвязей между объектами
  - Интеграция открытой и закрытой информации

ИСП РАН

Анализ социальных сетей

Вызовы

1. Big Data: Большой объем связанной между собой информации
2. Специфичный язык
   - Сленг,
   - Внешний контекст
   - Высокая скорость изменения
3. Недостоверность информации
   - Боты
   - Дезинформация

- Поиск сообществ (кластеризация графов): 1 млрд. вершин, 100 млрд. ребер
- Моделирование сложных сетей (complex networks)
- Сопоставление профилей в разных сетях (F_1 - 89% при сопоставлении контактов вершины)
- Определение значений демографических атрибутов (*пол, возраст, регион проживания, семейное положение, образование*; достоверность ~80%)
- Инструменты распределенного подсчета векторных представлений вершин (graph embedding)
- Выявление ботов (F_1– 65%-82%)
- Обнаружение информационных кампаний
- **Результаты в диссертациях**
  - **А.В. Коршунов (2015)**
  - **М.Д. Дробышевский (2019)**
  - ***А.Г. Гомзин (2021)***

ИСП РАН

Анализ текстов

Вызовы

1. Извлечение информации из неструктурированных текстов и полу-структурированных документов
2. Автоматическое построение баз знаний
3. Анализ эмоциональной окраски сообщений

- **Texterra:** Система анализа текстов
  - Более 20 инструментов анализа текстов
  - http://texterra.ispras.ru
- **DEREK:** открытый нейросетевой фреймворк для извлечения сущностей и отношений
  - https://github.com/ispras-texterra/derek
- **TIE:** набор инструментов извлечения информации для пополнения предметно-ориентированных баз знаний (часть платформы Talisman)
- Прикладные направления:
  - Анализ текстов социальных медиа
  - Автоматический анализ законодательства
  - Поиск заимствований
- **Результаты в диссертациях**
  - **Д.Ю. Турдаков (2010)**
  - **Н.А. Астраханцев (2015)**
  - **Анонс: Ц.Г. Гукасян (2021)**

ИСП РАН

Сбор информации из открытых источников

Вызовы

1. Обеспечение заданной **полноты** и точности в условиях ограниченных ресурсов
2. Отсутствие фиксированных программных интерфейсов
3. Обеспечение актуальности данных
4. Защита от сбора данных
5. Обработка динамических страниц

- Создана облачная инфраструктура сбора данных из открытых источников
  - социальных сетей (VK, Facebook, Instagram, Twitter)
  - блогов (LiveJournal)
  - сайтов СМИ
  - новостных агрегаторов (Яндекс.Новости)
  - Форумы Dark web
- Разработано расширения для браузеров, позволяющих визуально разметит Веб-сайт для обхода и извлечения нужной информации
- Разработаны методы автоматического построения сборщиков для некоторых категорий ресурсов (новостные сайты)

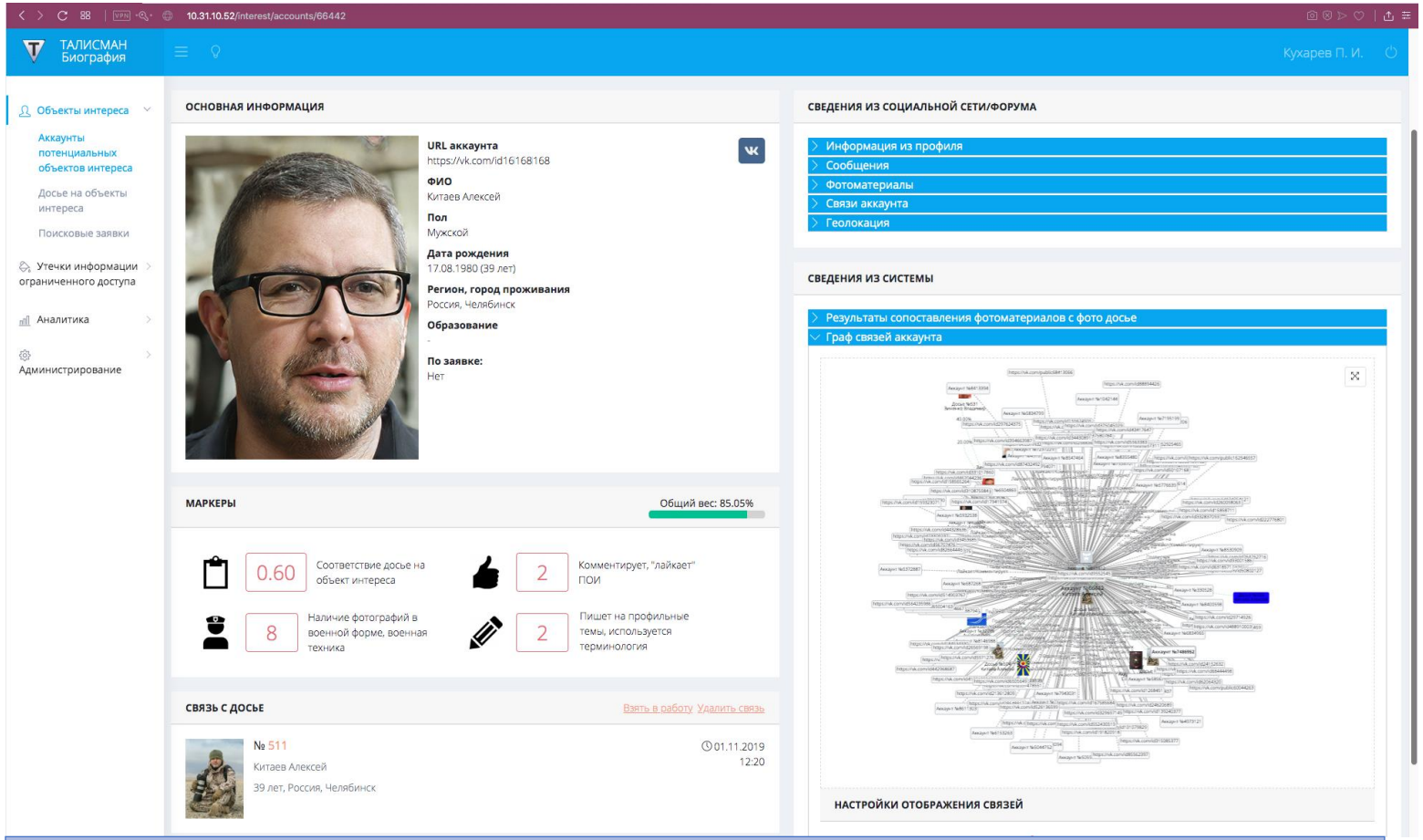
Машинное обучение

Вызовы

1. Минимизация усилий на получение размеченных данных для обучения моделей
2. Проблема устаревания моделей машинного обучения (concept drift, feature drift) в работающей системе
3. Объяснение моделей (Explainable machine learning)
4. Безопасность использования ИИ

1. Инструменты для автоматизация получения размеченных данных
   - Активное обучение
   - Краудсорсинг
2. Создано решение проблемы устаревания моделей
   - Проактивное обучение
   - Адаптация к новой предметной области
   - Автоматическое дообучение и контроль версий моделей
3. Разрабатываются методы объяснения моделей
4. Изучаются методы создания доверенного ИИ

Решение: Talisman.Биография



- Регистрация в едином реестре российских программ (№ 5547)
- Сертификация НДВ2

Область применения

- Задачи отдела кадров
- Задачи отдела по связям с общественностью
- Связывание данных сотрудников или соискателей с их аккаунтами в социальных сетях
- Верификация анкетных данных
- Обнаружение утечки корпоративной информации через аккаунты сотрудников

ИСП РАН

Результаты

Включены в лучшие результаты РАН	• 2017 Texterra - масштабируемая платформа для извлечения семантики из текста • 2018 Talisman - платформа анализа социальных медиа
Включены в Единый реестр российских программ	• Talisman.Биография • Talisman.Поток • Облачная среда «Асперитас»
Внедрение фундаментальных результатов через платформенные решения	• Внедрение решений на основе платформы Talisman • Платформа для цифровой медицины

МОДЕЛЬ КОНКУРЕНТОСПОСОБНОГО РАЗВИТИЯ ЦЕНТРОВ КОМПЕТЕНЦИЙ

ИСП РАН

ИСП РАН

Спасибо!