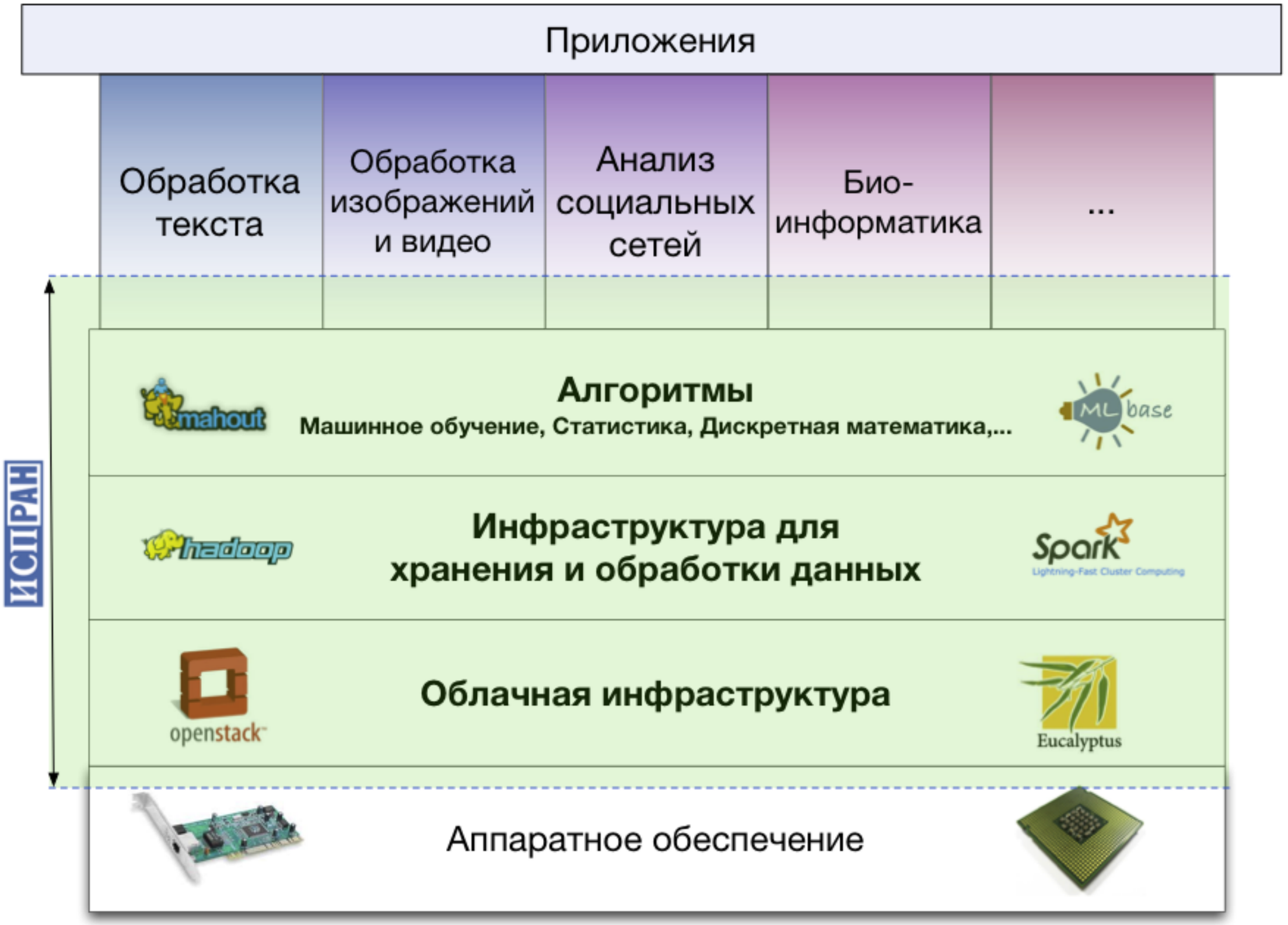
ИСП РАН

Обработка естественного языка и анализ социальных сетей в условиях взрывного роста данных

к. ф.-м. н. Турдаков Денис Юрьевич

Общий взгляд



ИСП РАН

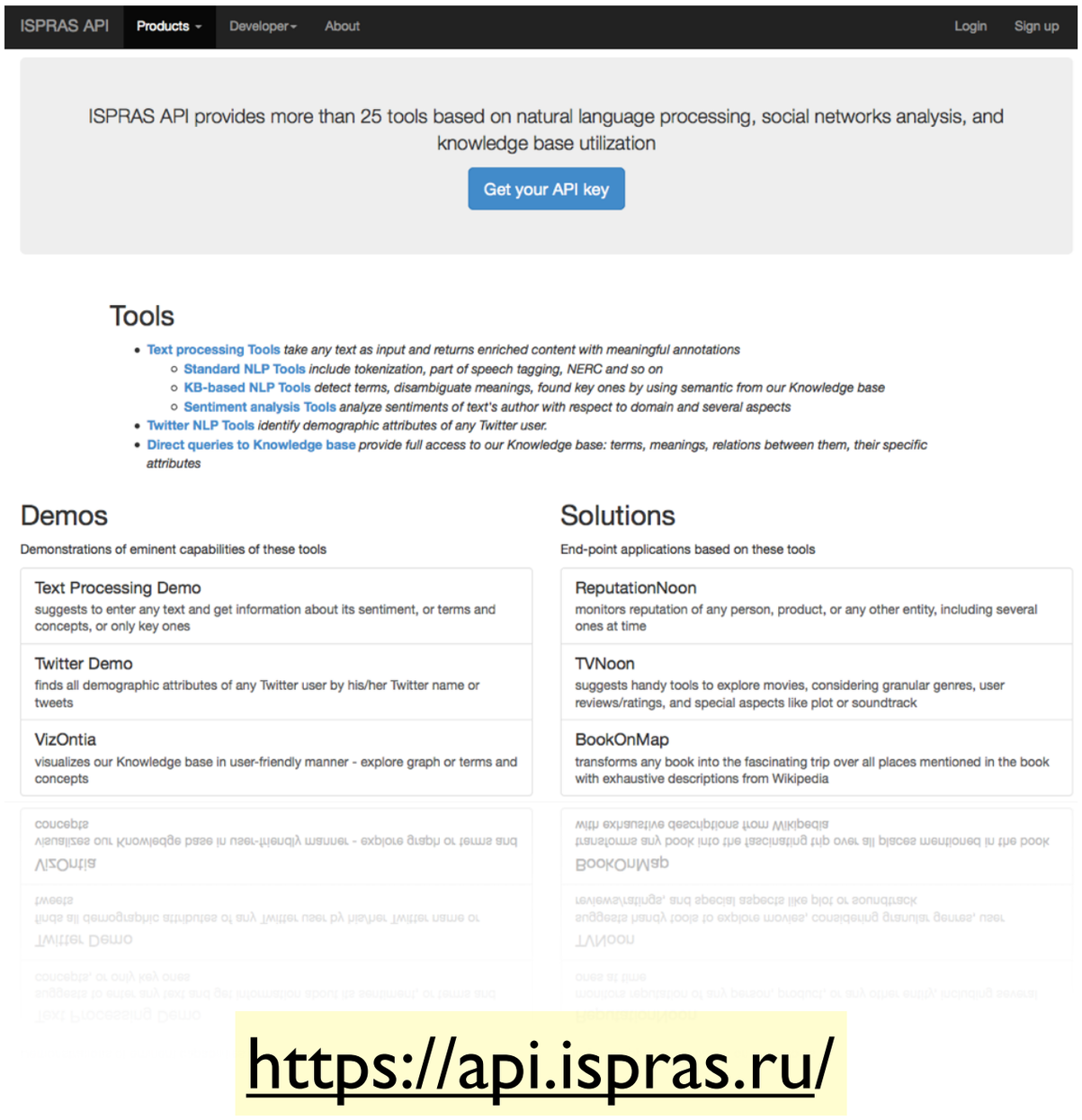
Технологии ИСП РАН

ИСП РАН

- XML СУБД с открытым кодом
- Международное сообщество пользователей и разработчиков
- Более 100 загрузок ежедневно

Сервисы ИСП РАН

ИСП РАН предоставляет более 25 сервисов для обработки естественного языка, анализа социальных сетей и работы с базами знаний



https://api.ispras.ru/

ИСП РАН

Обработка естественного языка

ИСП РАН

Приложения обработки текстов

- Семантический поиск
- Автоматическое реферирование
- Анализ эмоциональной окраски сообщений
- Определение демографических атрибутов пользователей Интернета
- ...

ИСП РАН

Текущее состояние области

- Коммерческие сервисы

CALAIS Powered by Thomson Reuters

AlchemyAPI™

- Библиотеки и системы с открытым кодом

GATE general architecture for text engineering

- Решаются задачи уровня морфологии и синтаксиса
- Для решений уровня семантики нужны знания о мире в понятном для компьютера виде
- Построение баз знаний вручную чрезвычайно трудоемко и дорого

ИСП РАН

Источники знаний

- Источники знаний о многих предметных областях
- Вызов: Специализированные области не покрыты

ИСП РАН

Технология Texterra

Texterra - масштабируемое решение для быстрой обработки текстов, основанное на использовании знаний, извлекаемых из Веб-ресурсов и коллекций текстовых документов

Преимущества

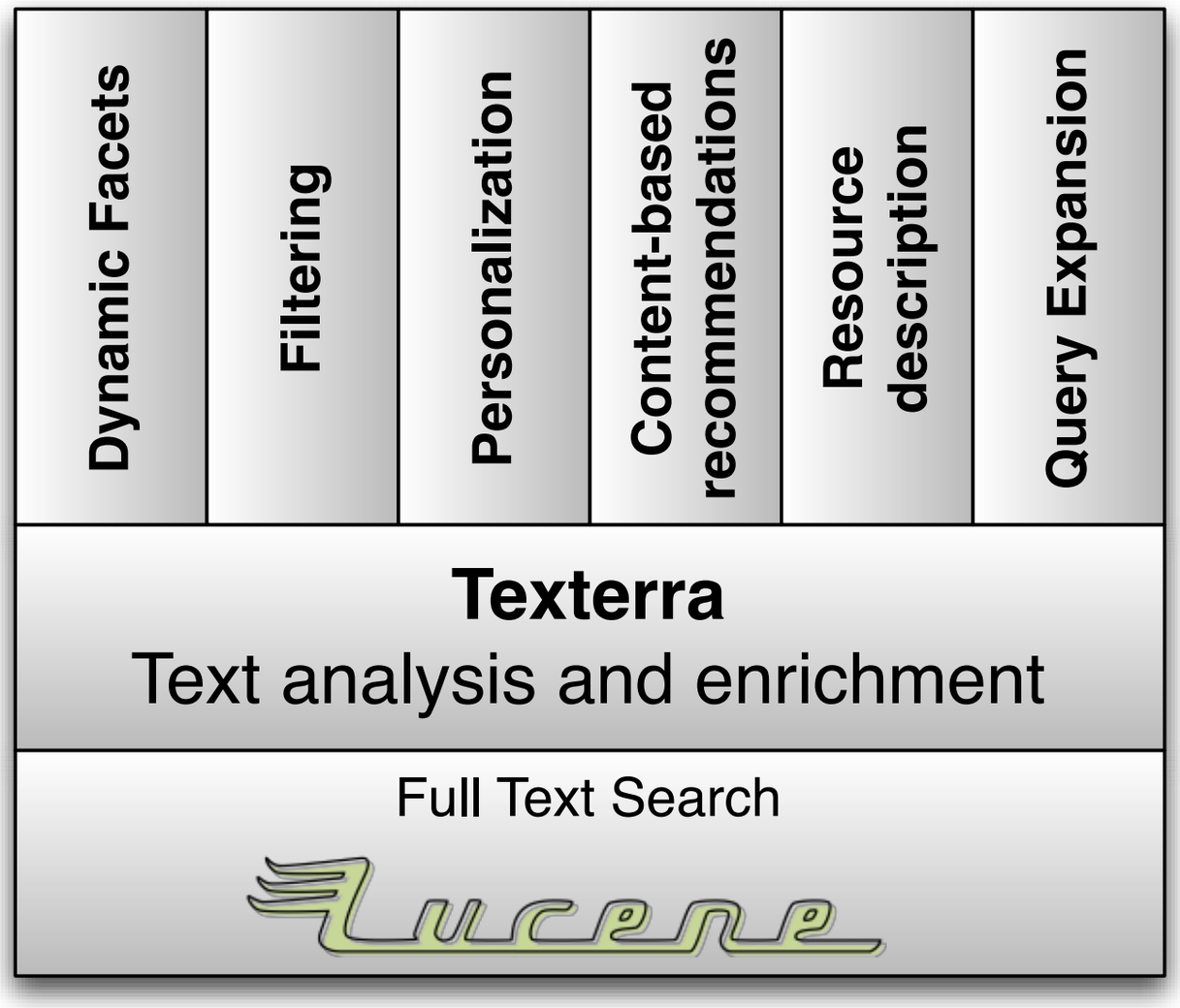
ИСП РАН

Используется подход аналогичный наиболее известным инструментам, но дополнительно предоставляется

- Инструменты для автоматического
  - построения базы знаний, используя открытые Веб-ресурсы
  - обогащения базы знаний предмено-ориентированной информацией на основе анализа текстовых документов
- Использование знаний о мире позволяет получить более точные результаты при решении прикладных задач
- Многоязычность: поддержка русского, английского, корейского + возможность добавления новых языков
- Высокая скорость работы

Технология *Noon

Набор инструментов для быстрой разработки предметно-ориентированных семантических поисковых и навигационных систем



ИСП РАН

*Noon: поиск

- Texterra преобразует поиск по ключевым словам в семантический поиск по значениям
- Эта операция улучшает точность ранжирования результатов поиска

ИСП РАН

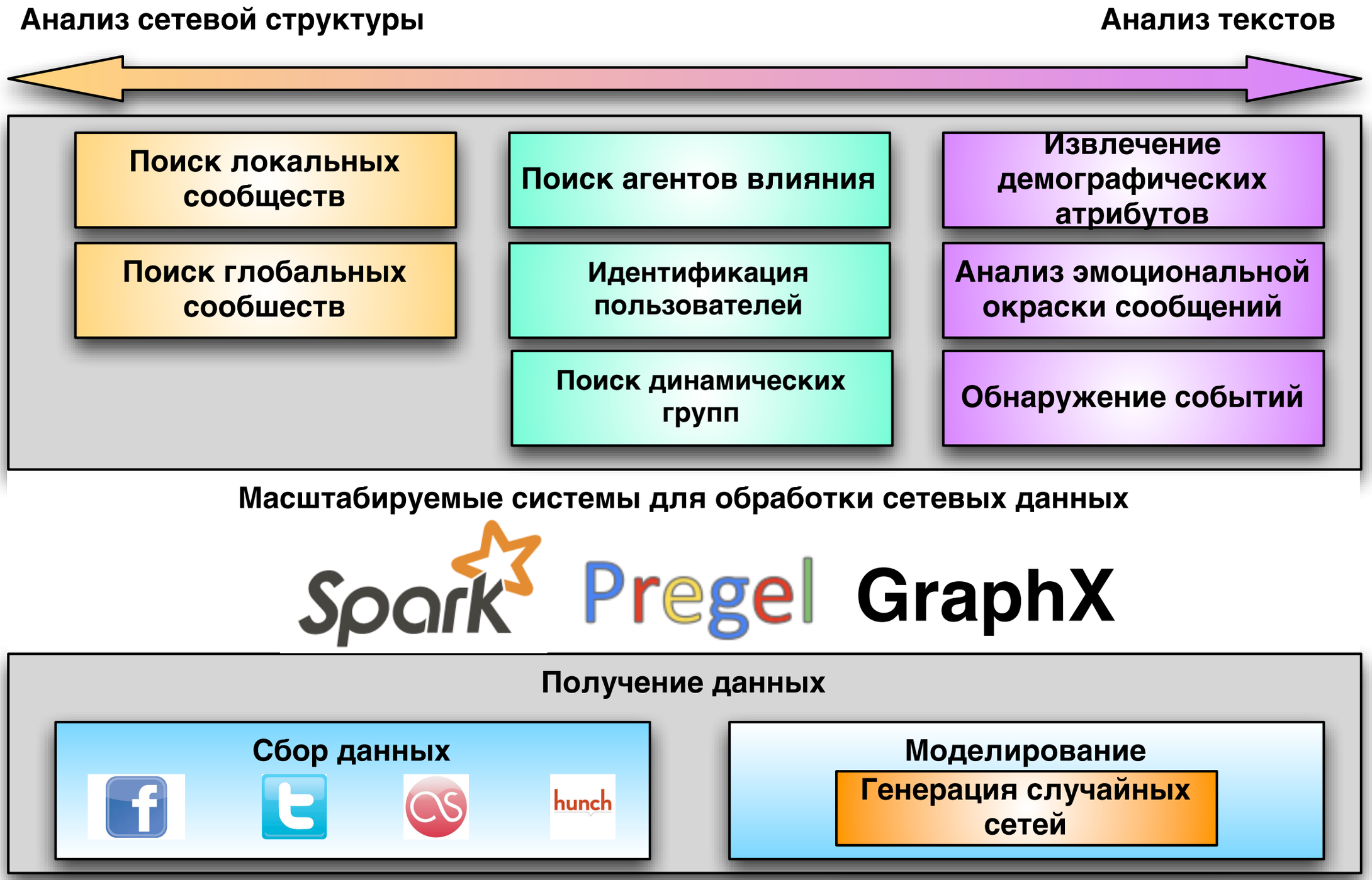
*Noon: навигация

- Что делать если ключевые слова для запроса неизвестны?
- *Noon предоставляет инструменты для навигации по коллекциям документов
  - Аналог навигации по карте в Веб-сервисах картографии (maps.yandex.ru)
  - Позволяет исследовать новую область и понять ключевые слова для последующего поиска

ИСП РАН

Анализ социальных сетей

Решения ИСП РАН



ISPRAS API | Products | Developer | About

ИСП РАН

Twitter Processing Demo

Twitter Link | Custom Info

BillGates

Autodetect language

Submit

Result

User is single male republican over 23 with higher education who believes in god.

Description

User Language: English

Extracted Attributes

- **Age:** over 23
- **Gender:** male
- **Relationship:** single
- **Politics:** republican
- **Education:** higher education
- **Religious:** yes

User Info

@BillGates

Bill Gates

Sharing things I'm learning through my foundation work and other interests...

Example Tweets

- 8 out of 10 people say they want to give to a charity. #GivingTuesday makes it easy: http://t.co/a0LGvYyEPw http://t.co/VOFZYc2uzb
- Progress towards the MDGs has been impressive. Are we ready to take the next step? @JeffDSachs weighs in: http://t.co/LZzin2by6G
- When you see the term “global health,” think “saving lives.” That's what it means. | @nytimes: http://t.co/1YPmfasp9J
- The idea that aid is wasted is just wrong. I recently spoke with the Sun newspaper about the impact of the UK's aid: http://t.co/KH8tTWLBL0
- Here's my wrap-up of day 2 in NYC. Got to see leaders from #Chad, #Pakistan, and much more: http://t.co/HXXcLiP9S6

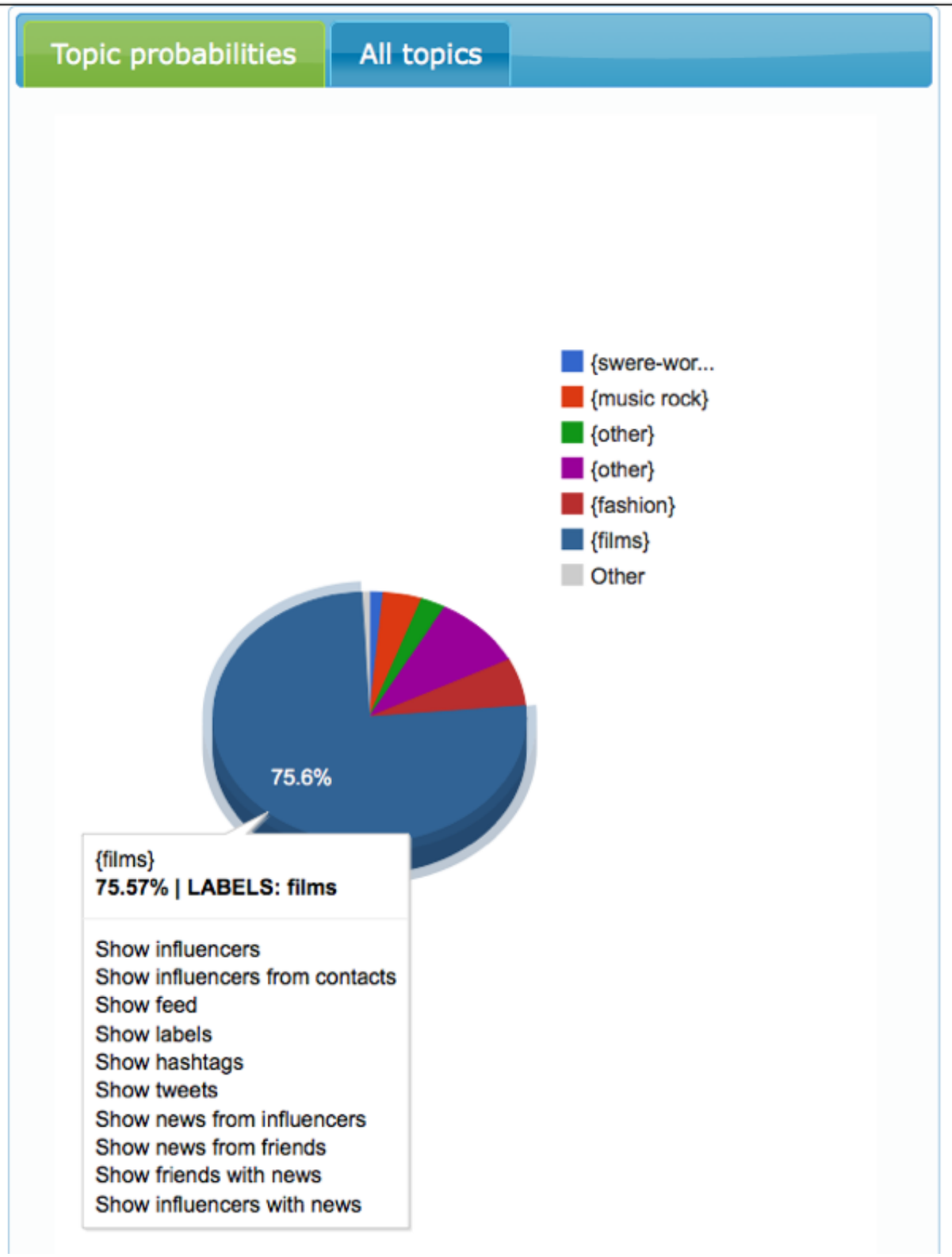
Извлечение демографических атрибутов

- Пол
- Возраст
- Семейное положение
- Политические взгляды
- Образование
- Религиозные взгляды

Поддерживается 8 языков

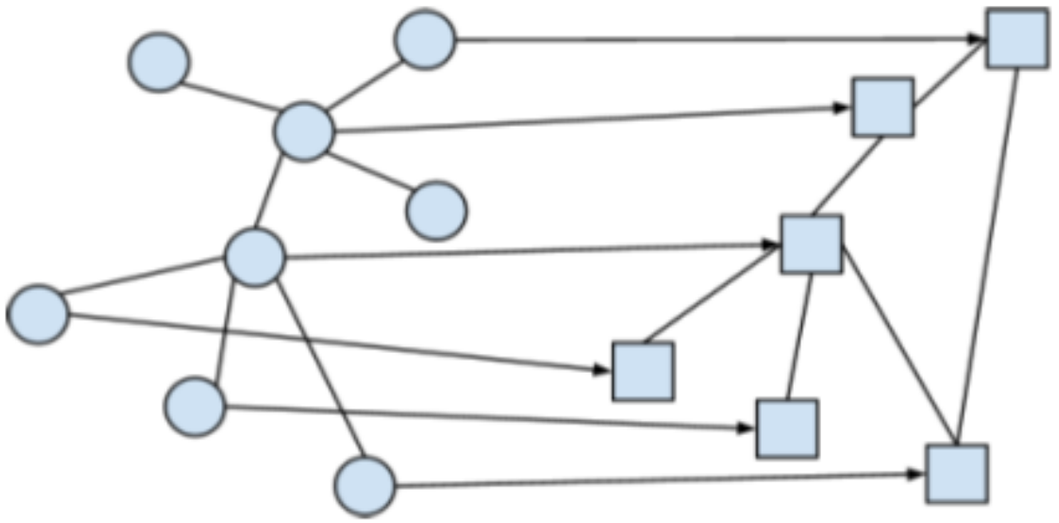
Поиск агентов влияния

- Граф влияния в Twitter с привязкой по темам
- На основе распределенных алгоритмов тематического моделирования
  - PLSA
  - robust PLSA



ИСП РАН

Идентификация пользователей

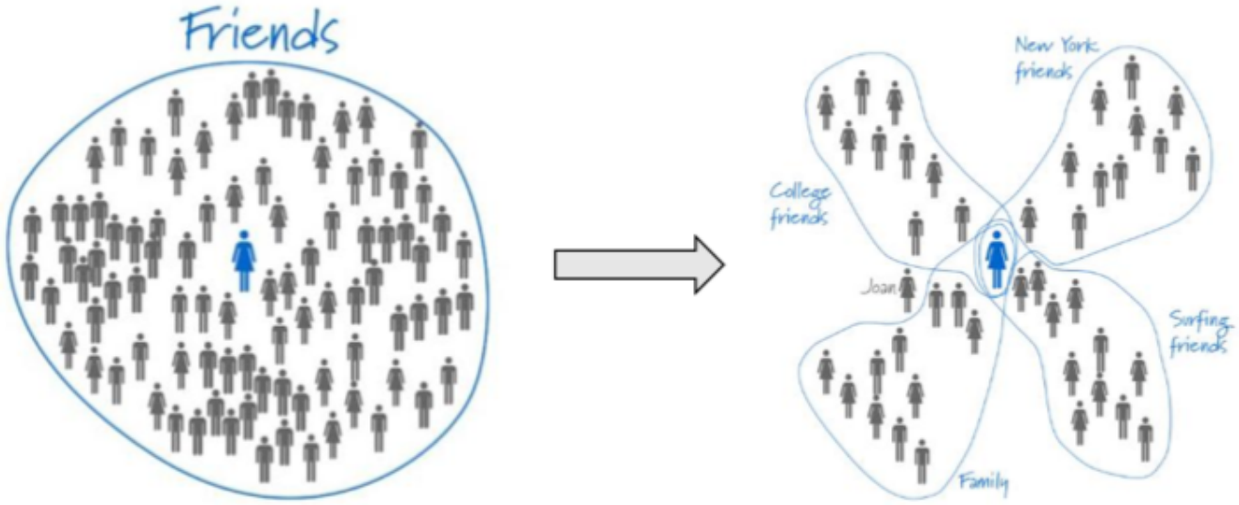


- Сопоставление аккаунтов пользователей в разных социальных сетях, используя
  - текстовые данные
  - структуру сети
- Позволяет производить деанонимизацию

ИСП РАН

Поиск сообществ

Локальные сообщества



- Автоматическое структурирование контактов
- Улучшение рекомендательных систем

Глобальные сообщества

- Оптимизация траффика
- Улучшение рекомендательных систем
- Поиска спама

Обработка x100 млн. пользователей!

ИСП РАН

Обработка “больших данных”

- Необходимы технологии для обработки больших объемов данных
- Использование систем с открытым кодом позволяет решать задачи уже сейчас

ИСП РАН

Участие ИСП РАН

- Во время решения практических задач возникают новые требования к открытым системам
- ИСП РАН активно участвует в развитии международных проектов
- Разработанные дополнения и модификации признаны сообществом

ИСП РАН