Принципы цифровой аналитики

Школа прикладного анализа данных для исследователей КМНС

Виталий Кашпур

Кашпур Виталий

- Заведующий кафедрой социологии Томского государственного университета
- Один из основателей Центра прикладного анализа больших данных Томского государственного университета
- Член команды Университетского консорциума исследователей больших данных
- В 2020 и 2021 года руководитель образовательной программы «Большие данные и машинное обучение в социальных и когнитивных науках» НТУ «Сириус», Сочи
- Руководитель и модератор групповой проектной работы в рамках стратегических сессий в ТГУ и других организациях
- Эксперт Агентства стратегических инициатив РФ
- Эксперт Федерального Агентства по делам национальностей РФ
- Член Экспертного совета по внутренней политике Администрации Томской области

Как эффективно спроектировать исследование с использованием методов и технологий Big Data?

- принципы цифровой аналитики
- этапы цифровых исследований
- специфика цифровых аналитических команд

Междисциплинарные исследования

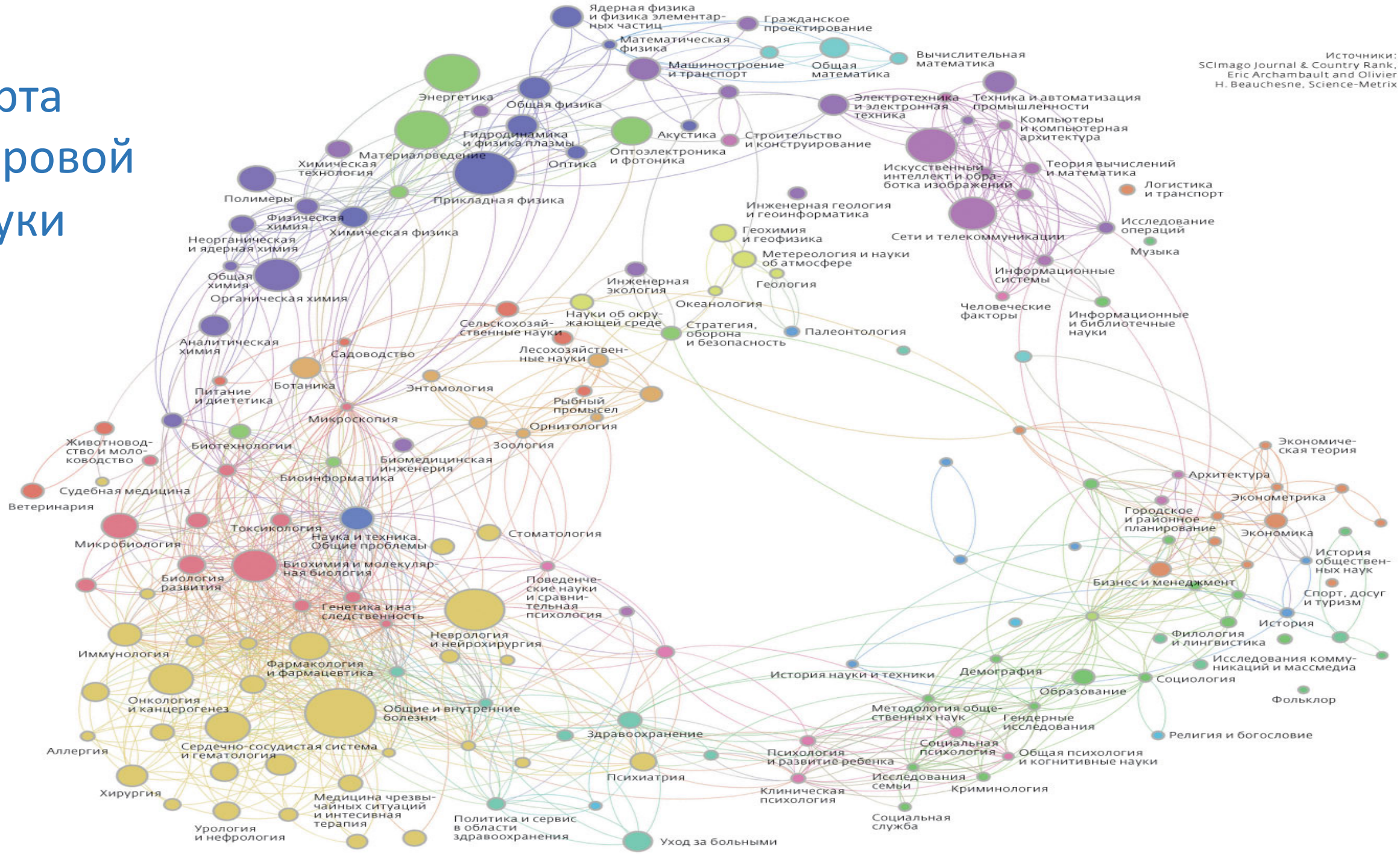
Научное исследование – деятельность по получению научных знаний.

Междисциплинарное исследование – исследование, совместно реализуемое представителями разных научных дисциплин.

Основания междисциплинарности:
-проблема;
-объект;
-метод.

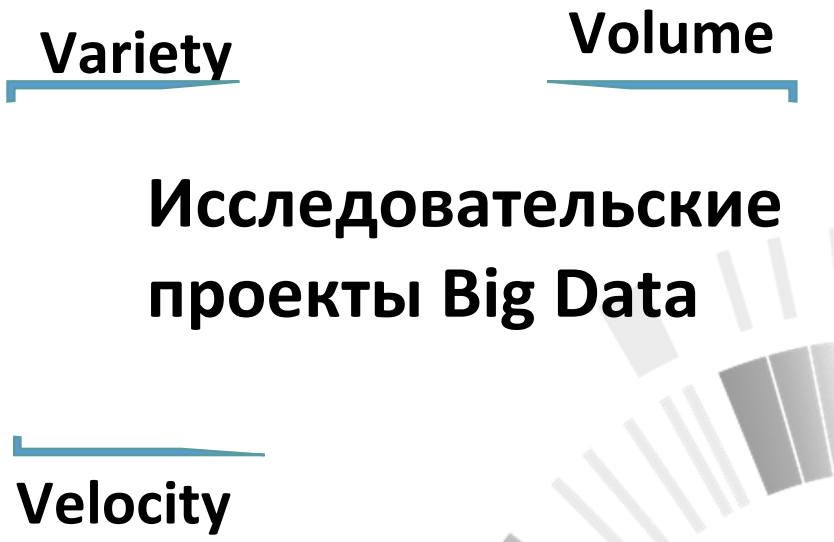
Big Data — междисциплинарное поле исследований

Карта мировой науки



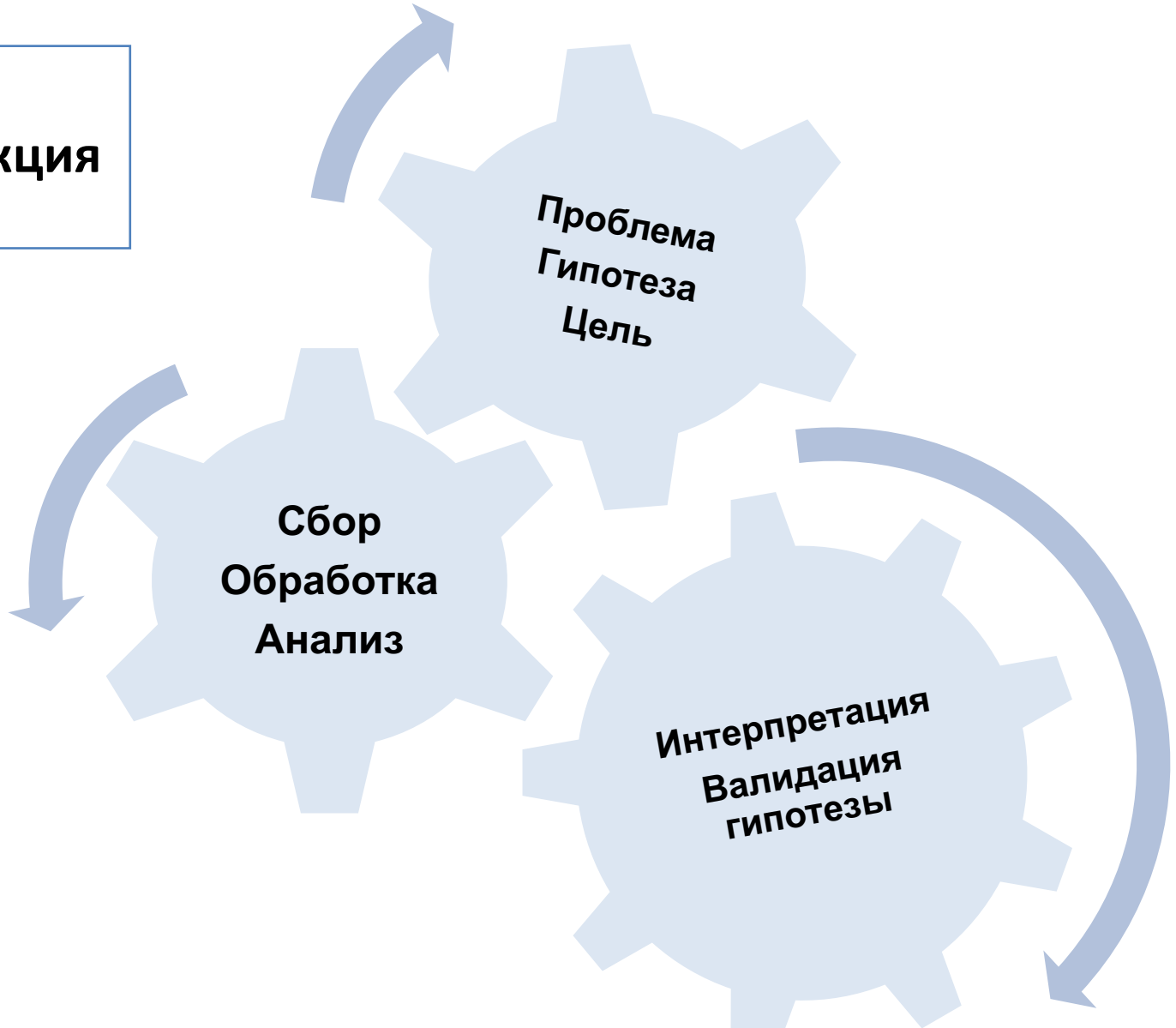
Принципы цифровых исследований

- ✓1. Big Data: объем, скорость обработки, разнообразие данных
- ✓2. Принцип использования всех возможных, в том числе и неструктурированных данных
- ✓3. Принцип приоритета внешних источников данных
- ✓4. Охват всех единиц наблюдения
- ✓5. Результат для пользователя
- ✓6. Этичность



Структура исследовательской деятельности

Цифровое исследование – абдукция = дедукция + индукция



Инициация исследовательского проекта

1. Описание ситуации и постановка проблемы/исследовательского вопроса
2. Формулирование гипотезы
3. Формирование команды
4. Идентификация стейкхолдеров и заказчика

Проблема - отсутствие/противоречивость знаний об объекте

Ключевые вопросы при проектировании:

1. Зачем проект? Какую проблему решаем и для кого? В чем его ценность? Кто его заказчик и стейкхолдеры?

2. На основе каких материалов будем делать? Каких методов, данных и инструментов - почему именно они?

3. Почем будем делать проект – цена и ресурсы?

Планирование исследовательского проекта

Откуда взять конкурентоспособную тематику исследований?

Ставка ЦПАБД ТГУ:

а) Мы претендуем на свой вклад в решение глобальных проблем человечества!

б) Научные темы «на хайпе», в которых у команды есть экспертиза и компетенции

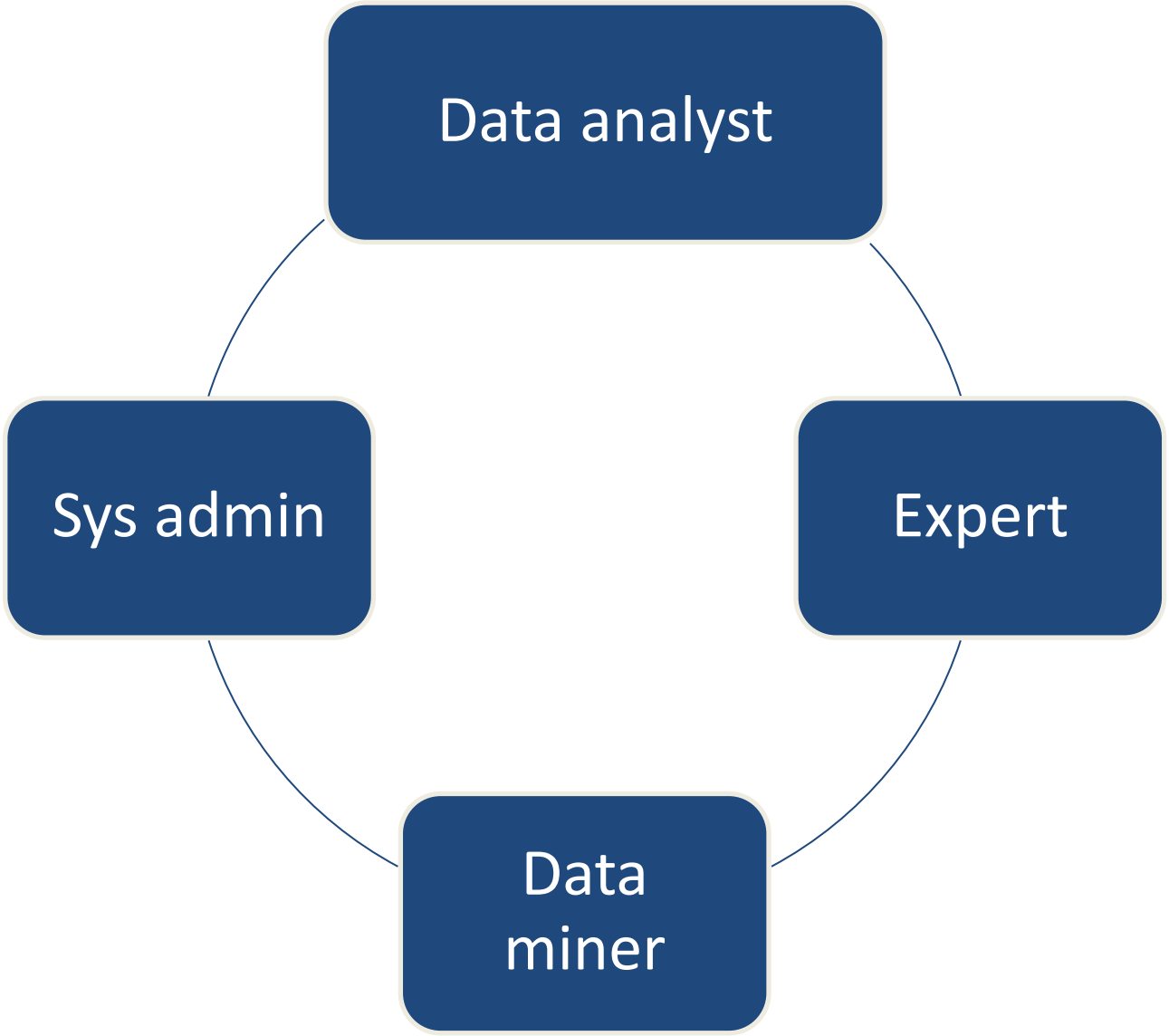
Реализация исследовательского проекта

Важнейшие элементы:

1. Работа по проекту
2. Сохранение и развитие команды
3. Обеспечение ресурсами
4. Эффективные коммуникации
5. Минимизация рисков

В современных проектах граница между проектированием и реализацией проекта условна!

Исследовательская команда: позиции и компетенции



Эксперт – специалист, имеющий компетентное предметное знание об изучаемом объекте

Дата-аналитик – специалист по методам анализа цифровых данных

Дата-майнер - специалист по добыче и обработке больших данных

Сисадмин – специалист по хранению и управлению большими данными

Уровни исследовательской работы

Важно

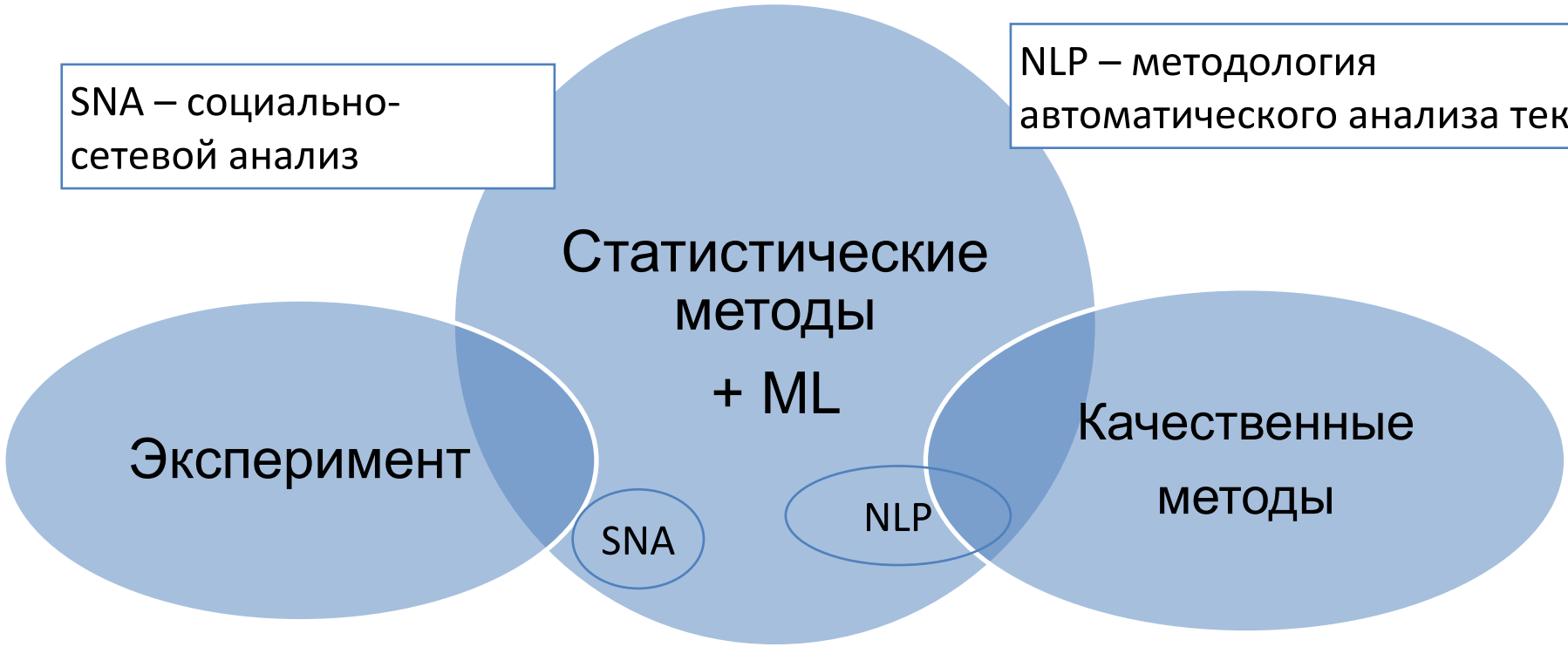
Формируйте общий язык исследовательской команды:

- *проблемные и методические семинары;*
- *совместная постановка исследовательских задач и интерпретация результатов;*
- *неформальное общение*

Методы цифровых исследований

SNA – социально-сетевой анализ

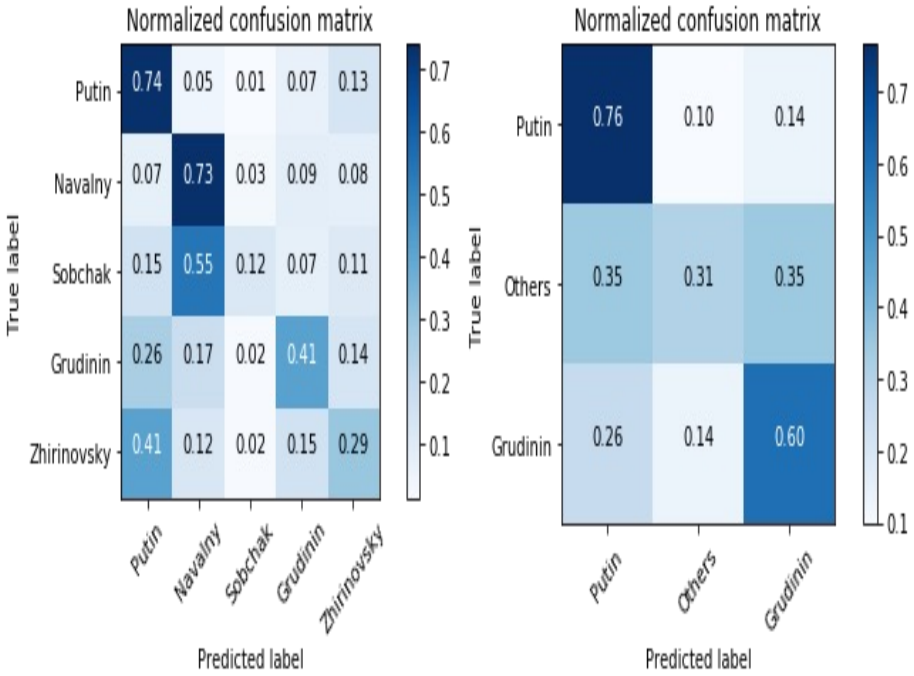
NLP – методология автоматического анализа текста



Пример: Модель прогноза результатов выборов на основе анализа пользователей ВК

Гипотеза - пользователь будет подписываться в ВК на источники информации исходя из своих политических предпочтений.

С использованием трехслойной искусственной нейронной сети (ANN), проведен анализ и классификация 23 000 000 пользователей Вконтакте в соответствии с их политическими предпочтениями и сделан прогноз результатов президентских выборов 2018 года.



Результаты

Кандидат	Прогноз 1	Прогноз 2*	Исход
В. В. Путин	43%	61.6%	76.7%
П. Н. Грудинин	18.8%	21.3%	11.8%
В. В. Жириновский	27.8%	17.1%	5.7%
К. А. Собчак	1.2%		1.7%
А. А. Навальный	9.2%	—	—
Остальные кандидаты	—	—	1%

*При обучении не учитывались сторонники А.А. Навального

Интерпретация данных

Интерпретация – ответы на вопросы:

1. Что мы видим из полученных данных?

Описание объекта (ситуации/системы/процесса/траектории) на основе полученных данных:

- норма/выброс показателя;
- характер динамики.

2. Почему мы получили этот результат?

Определение факторов влияния, причинно-следственных связей;

3. Что с этим результатом делать?

Прогноз

Проверка гипотезы и валидация решений (выбор наиболее эффективной альтернативы)

Принципы организации исследовательских центров

- **Научная эффективность**: исследовательский протокол, релевантность, актуальность и альтернативность исследований
- **Ориентация на рынок/заказчика**: не только исследовательский результат, но и продукт
- **Разделение труда** и наличие разных позиций в команде
- **Включенность в сети**: промышленные партнеры, образовательные программы, другие университеты

Спасибо за внимание

Виталий Кашпур, vitkashpur@mail.ru