# Анализ научных данных: от статистики к data science и обратно

**Преподаватель**

Андрей Владимирович Алексеенко, кандидат биологических наук, доцент (Karolinska Institutet, Stockholm), по образованию биолог со специализацией на генетике растений. Последние 20 лет занимается количественным анализом биомедицинских данных. Автор около 50 научных работ. Создатель методов для анализа множеств и графов, а также веб-сайтов и других публично доступных ресурсов. Преподаватель и организатор ряда курсов по прикладной статистике и системной биологии.

<https://staff.ki.se/people/andrey-alexeyenko> <https://research.scilifelab.se/andrey_alexeyenko/>

**Краткое описание курса**

Современная методология анализа данных может предстать начинающему исследователю оторванной от изученных стандартных курсов математики и статистики. Цель курса – установление связей между теоретическими понятиями и практическими целями и методами исследований в разных областях, от биомедицины до общественных наук. Обзор методов и общение преподавателя со студентами поможет уточнять интересы, задачи курса и терминологию – как междисциплинарно, так и между языковыми доменами. Для желающих получить начальные навыки программирования будет организовано дополнительное занятие по языку R.

Литература по курсу будет даваться в основном на английском языке, в виде отдельных статей или небольших выдержек.

**Длительность:** 1 семестр (8-9 еженедельных занятий по ~1,5 часа; начало – в феврале).

**Формат:** лекции с вопросами и ответами, практические задания.

**Уровень:** базовый, но с (некоторым) опытом изучения либо применения теории вероятностей и/или статистического анализа.

**Пререквизиты**

Знание английского языка на уровне уверенного чтения. Опыт работы с данными или намерение анализировать данные для какой-либо насущной задачи. Умение программировать не требуется, но приветствуется.

**Программа**

**Занятие 1**

Дискуссия: содержание курса, обмен опытом и мнениями. Экзамен, сертификаты.

История анализа данных, концепции и подходы. «Зачем данные?». Big data и data science. Наблюдения. Информация и неопределенность. Изменчивость и ее меры. Нормальное распределение и его причины. Сигнал и шум. Типы данных и форматы их представления. Типы анализа: разведочный, доказательный, причинно-следственный.

**Дополнительное занятие «Начальные навыки программирования, язык R»**

**Занятие 2**

Статистические модели, описывающие зависимости между переменными. Частотный и байесовский подходы, параметрические и непараметрические статистики. Одномерные статистики как меры связи переменных: анализ таблиц 2х2, t-критерий, коэффициенты корреляций и критерии их пригодности. Виды переменных, параметры. Требования к переменным в параметрическом анализе. Сравнение методов одно- и многомерного анализа.

**Занятие 3**

Статистическая значимость и ее меры. P-значение. Факторы, определяющие значимость: сила сигнала, разброс значений, объем выборки. Статистические ошибки I и II рода, главные различия в их использовании. Доверительные интервалы. «Экологическое заблуждение». Статистическая мощность и планирование эксперимента. Гипотезы и их проверка. Метод случайных пермутаций. Поправки на множественность тестирования. Кривые ROC.

**Занятие 4**

Многомерный анализ (МА) с использованием операций на матрицах: факторный, дискриминантный, метод главных компонент и другие. Смысл латентных переменных в МА. Спектральный анализ. Симплексы. Интерпретация и использование латентных переменных. Требования к исходным матрицам. МА в отсутствие критериев значимости. Примеры использования МА.

**Занятие 5**

Линейные модели: дисперсионный анализ и множественная регрессия. F-критерий. Частные корреляции. Взаимодействие и смешение эффектов. Медиаторы, модераторы, конфаундеры, ковариаты и коллайдеры. Проклятие размерности. “The Bell curve”. Успехи (?) глубокого обучения и путь к ним.

**Занятие 6**

Графы и сети (очень большие графы) как подход к визуализации и анализу данных. Основные понятия, типы и структурные элементы графов. Scale-free networks и preferential attachment. Информационное содержание сетей и emerging properties как результат их анализа. Всеведущая сеть? Форматы представления графов и спектральная кластеризация. Алгоритмы анализа, их назначение и вычислительная сложность. Графы как множества. Enrichment: пересечение множеств, статистически значимо отклоняющееся от случайного. Диаграммы Венна и Эйлера. Статистическая значимость результатов.

**Занятие 7**

Состоятельность результатов и их валидация. Мета-анализ, меры состоятельности и кумулятивной значимости. Модели, их назначение, создание, оптимизация и применение. Оценка качества моделей. Информационные критерии. P-hacking. Демонстрация к практическому занятию.

**Занятие 8**

Ковариаты и их роль в анализе. Страты и парадокс Симпсона. Over-fitting и проклятие размерности. Причинно-следственные связи и способы установления причинности.

Дискуссия о прослушанном курсе, обмен мнениями. Экзаменационные работы, сертификаты.