

Национальный исследовательский университет

Высшая школа экономики

Концепция

Магистерской программы

Науки о данных (Data Science)

НАПРАВЛЕНИЕ 01.04.02

«Прикладная математика и информатика»

Степень (квалификация): магистр

Москва, 2018

1 Общая характеристика

Магистерская программа «Науки о данных» реализуется на факультете компьютерных наук НИУ ВШЭ.

Форма обучения – очная

Набор ведется как на бюджетные места, так и на места с оплатой стоимости обучения.

Основной язык преподавания – русский с возможностью, в случае поступления международных студентов, прочтения большинства курсов на английском и набора всех 120 кредитов на английском языке.

У программы существуют международные партнеры, такие как Университет Клермон Оверни (бывший Blaise Pascal, Clermont-Ferrand (France)), с которыми имеются соответствующие соглашения о студенческом обмене. Российские студенты, обучающиеся на программе, имеют возможность получить второй диплом в вузе-партнере.

2 Актуальность и значимость

Одним из самых важных современных цивилизационных вызовов – анализ растущего объема данных, порождаемых во всех областях современного общества. Благодаря стремительному развитию современных технических средств и вовлечению буквально каждого члена общества в производство данных, эта проблема будет носить всё более и более фундаментальный характер, определяя эффективность всех серьёзных государственных, научных и корпоративных проектов. Современная ИТ-индустрия откликается на этот вызов, поднимая проблематику больших данных (*Big Data*), а академическое сообщество – формируя науку о данных (*Data Science*). На этот вызов откликаются ведущие университеты мира, открывая факультеты и образовательные программы по науке о данных. Эти образовательные программы предусматривают фундаментальную подготовку в области моделей вычислений, математических методов моделирования и прогнозирования, вычислительной архитектуры, современных методов программирования, хранения и извлечения данных.

Наука о данных крайне востребована в различных областях знаний и технологий, и будет иметь всё большее значение в современных исследованиях и бизнесе. Одновременно, возникновение Науки о данных теснейшим образом коррелирует с возникновением и развитием новых математических подходов, алгоритмов и методов, которые активно разрабатываются и эффективно применяются при анализе и обработке данных, в моделировании. Эти современные математические технологии представляют собой критически необходимое звено в технологической (постиндустриальной) революции XXI века.

Несмотря на колоссальную потребность в специалистах в области наук о данных, в нашей стране этот вызов ещё не встретил должного ответа со стороны образовательного сообщества. А ведь наука о данных как область исследований могла бы стать тем объединяющим началом, которое бы связало в одном исследовательско-образовательном проекте представителей различных областей информатики и прикладной математики, в частности, такие области как теорию вычислений, методы математического моделирования, вычислительную архитектуру, программирование, системы принятия решений, технологии хранения и обработки данных, методы формирования и поддержки баз знаний.

3 Цели магистерской программы

Создание магистерской программы «Науки о данных [*Data Science*]» (далее – НОД) отвечает основным вызовам, обращённым к НИУ ВШЭ и Российскому образованию в целом: выход на передовые позиции на международном рынке образовательных программ, подготовка элитных кадров инновационной экономики, обладающих аналитическими,

исследовательскими, методическими и технологическими компетенциями в новых областях науки, техники и экономики.

Овладение выпускниками программы целым спектром новейших, эффективных математических технологий, не включённых в классические математические учебники, по которым традиционно готовятся студенты российских университетов, будет существенно способствовать повышению их конкурентоспособности и востребованности.

В силу принципиальной мультидисциплинарности науки о данных, призванной предоставлять средства анализа информационных, социальных, правовых, психологических, экономических, языковых и других явлений, программа станет одной из системообразующих, интересной выпускникам многих факультетов и сотрудникам многих исследовательских центров НИУ ВШЭ и других академических организаций. Например, с факультетом гуманитарных наук – компьютерная лингвистика, автоматическая обработка текстов, факультетом медиакоммуникаций – визуализация данных и знаний, проектирование интерактивных приложений, факультетом социальных наук и другими общественно-научными факультетами – методы обработки и интерпретации массивов разнородной информации.

Основными целями НОД являются следующие.

1. Получение лицами, имеющими базовое образование в области прикладной математики и смежных областях, специализации магистерского уровня в сфере наук о данных.
2. Дополнение аналитических компетенций, полученных при изучении экономики, социологии, компьютерной лингвистики и т.п., математическими и технологическими компетенциями по работе с многомерными данными сложной структуры, включая большие данные, представленные в различных форматах.
3. Получение профессионалами в области системной аналитики или аналитики предметной области дополнительных технологических навыков в области специализации (при попадании профессиональных интересов в один из профилей).

3.1 Целевая аудитория

1. Выпускники бакалавриата НИУ ВШЭ.
2. Выпускники бакалавриата математических факультетов ведущих университетов.
3. Выпускники бакалавриата ведущих университетов по направлениям и специальностям, связанным с необходимостью анализа больших данных (экономика, социология, управление сложными системами и др.).
4. Специалисты и магистры, желающие получить дополнительную специализацию и системно подготовиться к серьёзной аналитической или исследовательской деятельности в области анализа сложных систем и больших данных, а также разработке интеллектуальных систем.
5. Выпускники иностранных вузов с достаточной базовой математической подготовкой.

3.2 Обзор текущего состояния образования в области *Data Science* и основные конкуренты

С 2011 года вопрос адекватной подготовки специалистов в области наук о данных становится «модным», но, в отличие от многих других «модных трендов», он интересует исследователей и профессионалов практически всех научных и технологических ниш. В апреле 2013 года тема «*Data Science Curriculum*» активно обсуждалась на *European Data Forum*.

Дополнительные программы уже запущены во некоторых известных университетах (Университетах Нью-Йорка (<http://cds.nyu.edu/academics/ms-in-data-science/curriculum>), Вашингтона (<http://www.pce.uw.edu/certificates/data-science.html>), Карнеги-Меллон (<http://www.heinz.cmu.edu/school-of-information-systems-and-management/information-systems-management-mism/business-intelligence-data-analytics/index.aspx>), Сиракуз (<http://ischool.syr.edu/future/cas/datascience.aspx>) и др.). В прошлом году появились первые полноценные online-курсы (например, <https://www.coursera.org/course/datasci>), а с 2014 года запускается online-программа в Беркли (<http://datascience.berkeley.edu>).

Отметим, что во многих случаях профессионалы в отдельных областях просто не используют словосочетание «*Data Science*» по историческим причинам.

4 Обоснование потребности в магистрах

На рынке труда востребованы специалисты, способные работать в области анализа многомерных данных сложной структуры, включая большие и текстовые данные. Уже не раз отмечалось, что организациями накоплены огромные массивы данных, многие из которых плохо структурированы. Их обработка и анализ становятся всё актуальней по мере того, как ускоряются бизнес-процессы, ожесточается конкуренция и возрастает цена своевременного и правильно принятого решения. В последние годы также становятся всё более доступны для анализа личные и персональные данные, размещённые в сети Интернет, особенно в виде «социальных сетей».

Классическая схема подготовки аналитиков не соответствует этим вызовам, поскольку системно не охватывает дополнительные задачи обработки и анализа данных, включая неструктурированные данные больших объёмов. При этом очевиден дефицит специалистов, готовых системно подходить к решению задач, связанных именно с методологией обработки данных разных видов и типов, упорядочением доступа к хранилищам данных, перестройкой структуры хранилищ, эффективностью процессов обработки, анализом больших данных (требующих снижения размерности, спец. схем проведения статистических экспериментов, приближённых методов, эффективных алгоритмов) и т.п. Дефицит обостряется с развитием смежных технологий: 3D-печати, дополненной реальности, облачных вычислений, «умной» среды и т.д.

В качестве примеров можно привести компетенции, указываемые в вакансиях на ведущих online-площадках: Работа с большими объемами данных, анализ данных, BI, Big Data, Distributed Cache, Data-Warehouse, ETL, Business Intelligence, Hadoop, MapReduce, Опыт анализа социальных сетей, Опыт работы с Big Data, и т.д.

Выпускники НОД будут востребованы в следующих сферах деятельности.

1. Научно-исследовательская деятельность, как в сфере профессиональной специализации, так и в других сферах в составе коллективов, имеющих дело с большими данными.
2. Аналитическая деятельность (в её классическом понимании) в организациях всех форм собственности, включая промышленные предприятия, консалтинговые фирмы, ассоциации и фонды, органы государственного управления.
3. Экспертная деятельность, связанная с методологией, методами, задачами и технологиями управления и анализа больших данных в сфере профессиональной специализации.
4. Технологическая поддержка аналитической, консалтинговой и оптимизационной деятельности коллективов, имеющих дело с большими данными.
5. Руководство коллективами в аналитических, исследовательских и управленческих подразделениях организаций всех форм собственности.

5 Набор компетенций, формируемых программой

НОД формирует следующие базовые компетенции:

1. Способен применять на практике системные знания о структуре и особенностях некоторой предметной области в сфере специализации.
2. Способен строить математические модели и вычислительные алгоритмы обработки (получения, хранения, переработки) данных.
3. Способен строить математические модели и вычислительные алгоритмы анализа данных как в общей постановке, так и в сфере специализации.
4. Способен использовать на практике современные технологии обработки больших объёмов сложно или недостаточно структурированных данных.

В целом выпускник НОД может решать задачи поиска, сбора, хранения, подготовки, анализа данных и интерпретации результатов в области специализации.

6 Особенности структуры и содержания программы

6.1 Оригинальный образовательный стандарт

Магистерская программа «Науки о данных [*Data Science*]» опирается на оригинальный образовательный стандарт (ОрОС) НИУ ВШЭ по направлению 01.04.02 «Прикладная математика и информатика». Стандарт обновлен в соответствии с требованиями ФГОС 3+ в 2018 году.

Стандарт включает теоретическое обучение – цикл дисциплин направления и цикл дисциплин программы, в который входит цикл дисциплин специализации. В блок дисциплин направления входят дисциплины «Современные методы анализа данных» и «Современные методы принятия решений».

Также стандартом предусмотрено обязательное освоение студентами следующих элементов:

1. Научно-исследовательский семинар в соответствии со специализацией.
2. Междисциплинарная курсовая работа.
3. Научно-исследовательская и проектно-технологическая практики.
4. Проект и проектный семинар.
5. Итоговая государственная аттестация.
 - а. Подготовка и защита выпускной квалификационной работы.

Поскольку целевой аудиторией НОД являются достаточно разнородные бакалаврские программы различных университетов и институтов, в 1 модуле 1 курса предусмотрены адаптационные дисциплины: «Дискретная математика для разработки алгоритмов и программ» и «Теория вероятностей и математическая статистика».

6.2 Специализации

Учебный план НОД предусматривает возможность выбора специализации. Специализации представляют актуальные сферы профессиональной специализации выпускников.

В настоящий момент на программе реализуются 3 специализации.

1. Интеллектуальные системы и структурный анализ (на английском языке).
2. Анализ Интернет-данных (базовый и продвинутый треки).
3. Технологии моделирования сложных систем.

6.3 Учебный план

Базовый учебный план программы НОД имеет следующую структуру
Адаптационные курсы:

1. Дискретная математика для разработки алгоритмов и программ
2. Теория вероятностей и математическая статистика
Блок дисциплин направления:
 1. Современные методы анализа данных.
 2. Современные методы принятия решений
 Обязательные элементы учебного плана:
 1. Научно-исследовательский семинар (по специализациям)
 2. Курсовая работа
 3. Проектный семинар
 4. Дисциплина по выбору из общеуниверситетского пула «МАГОЛЕГО»
 5. Проектно-технологическая практика
 6. Научно-исследовательская практика
 7. Проект
 8. Подготовка выпускной квалификационной работы
 9. Защита выпускной квалификационной работы

Базовая дисциплины специализаций:

Базовая часть специализации ИССА:

Упорядоченные множества в анализе данных

Анализ сетевых структур

Введение в методы машинного обучения и майнинга данных

Методы машинного обучения и майнинга

Базовая часть специализации ТМСС:

Упорядоченные множества в анализе данных

Математические основы современных телекоммуникаций

Предсказательное моделирование

Современные матричные вычисления

Вариативная дисциплины специализаций ИССА и ТМСС:

I семестр

Методы автоматической верификации программ

Медицинская информатика

Робастные методы в статистике

Неопределенность и нечеткость при анализе данных и принятии решений

Прикладная линейная алгебра

II семестр

Компьютерная лингвистика и анализ текста

Теория информации и комбинаторная теория поиска

Игры и решения в задачах анализа данных и моделирования

Анализ данных в медицине

Прикладная логистика и эволюционные

III семестр

Глубинное обучение

Инжиниринг данных и сервисов для автоматизации бизнес процессов

Введение в технологию блокчейн

Анализ и прогнозирование временных рядов: методы и приложения

Базовая часть специализации АИД

Стандартный трек

Машинное обучение

- Алгоритмы и структуры данных
- Методы и системы обработки больших
- Продвинутый трек
 - Дополнительные главы статистической
 - Методы и системы обработки больших
 - Обучение с подкреплением
 - Глубинное обучение: дополнительные главы
- Вариативная часть АИД (I семестр)
 - Стандартный трек
 - Восстановление функциональных зависимостей
 - Основы стохастики. Стохастические динамики сложных многокомпонентных систем
 - Язык Python
 - Продвинутый трек
 - Функциональный анализ
 - Выпуклый анализ и оптимизация
- Вариативная часть АИД (II семестр)
 - Стандартный трек
 - Вероятностно-статистические методы анализа данных
 - Параллельные и распределённые вычисления
 - Методы оптимизации в машинном
 - Продвинутый трек
 - Нейробайесовские методы в машинном обучении
 - Информационный поиск
- Вариативная часть АИД (III семестр)
 - Стандартный трек
 - Анализ изображений и видео
 - Автоматическая обработка текста
 - Обучение представлений и глубинное обучение
 - Продвинутый трек
 - Машинный перевод
 - Практическая аналитика

6.4 Концепция научно-исследовательского семинара

Научно-исследовательский семинар (НИС) магистерской программы состоит из 2 компонентов.

1. Общенаучная подготовка.
2. Семинары специализации.

Цель первого компонента – подготовить студента к участию в научных совещаниях и дискуссиях как одному из основных элементов функционирования науки. Программа семинара включает три основных вида деятельности, ориентированных на то, чтобы научить студента умению изложить свои результаты, понять и изложить чужие результаты, а также умению принять участие в научной дискуссии. Это:

- 1) подготовка и проведение доклада о собственной курсовой работе или ВКР;
- 2) изучение материалов и подготовка доклада об опубликованной в международной научной прессе работе;

- 3) активное участие в заседаниях студенческого семинара, научных семинаров лабораторий НИУ ВШЭ, семинаров различных НИИ РАН (например, ИППИ РАН, ИСА РАН, ВЦ РАН) и др., посещение конференций в сфере специализации.

Оцениваются все три вида деятельности. Для повышения эффективности подготовки студентам даётся дополнительная информация по подготовке презентаций, правилам ведения дискуссии и т.п.

Семинары специализации зависят от направленности конкретной специализации и могут проводиться на различных площадках, включая промышленные. Их главная цель – поддержать основную научно-исследовательскую деятельность студента. Формат проведения обсуждается собранием научных руководителей.

7 Отличительные особенности программы

Программа «Науки о данных» имеет следующие особенности.

1. Концентрация на математических основах, методологии, методах и алгоритмах анализа данных и обработки больших данных.
2. Углублённое изучение базовых моделей для системного взгляда на развитие наук о данных.
3. Акцент на данных сложной структуры, неструктурированных данных и данных на естественных языках, что требует привлечения методов искусственного интеллекта (например, онтологий, социальных сетей, историй болезней, баз патентов, блогов).
4. Сочетание траекторий для «аналитиков» (создателей математических моделей и методов) и «разработчиков» (создателей интеллектуальных систем).
5. Широкий спектр прикладных областей (от поиска в Интернет до медицинской информатики).
6. Опытные преподаватели, одновременно являющиеся известными исследователями и практиками в областях анализа и обработки данных в различных областях (например, Б.Г. Миркин, С.О. Кузнецов, Л.Е. Жуков).
7. Международные академические партнёры, включая университеты, с которыми заключены договоры об обмене (например, University BLAISE PASCAL, France, University of Texas, USA, KU Leuven, Belgium, TU Dresden, Germany).
8. Российские и зарубежные партнёры в промышленности, не просто применяющие известные методы и технологии на практике, а ведущие серьёзные научно-исследовательские и опытно-конструкторские разработки в области хранения и обработки больших объемов данных, внося вклад в развитие наук о данных на мировом уровне (такие как, Yandex, Abbyy, Infobright).

8 Ресурсное обеспечение

8.1 Обеспеченность профессорско-преподавательским составом

Руководитель магистерской программы – заведующий отделением Прикладной математики и информатики НИУ ВШЭ, д.ф.-м.н. С.О. Кузнецов, специалист в области математических моделей майнинга данных и машинного обучения, автор более 50 научных работ, индексируемых в системах Scopus и Web of Knowledge.

Ядром профессорско-преподавательский состава (ППС) МПНД является ППС департамента анализа данных и искусственного интеллекта, базовой кафедры Яндекс, кафедры технологий моделирования сложных систем (кафедра, на которой преподают сотрудники ИППИ РАН), Математические методы системного анализа (базовая кафедра ИСА РАН) и др. К проведению занятий в рамках МПНД планируется привлечь преподавателей и других подразделений НИУ ВШЭ.

К участию в преподавании на магистерской программе привлекаются лучшие специалисты в области анализа данных, майнинга данных и интеллектуальных систем, построению и анализу алгоритмов, среди них. отметим проф. Б.Г. Миркина, проф. Л.Е. Жукова, проф. К.В. Воронцова, проф. Г.А. Кабатянского, проф. И.И. Цитовича, доц. С.А. Обьедкова, доц. М.А. Бабенко, проф.О.С.Пьяных, проф.О.Ю.Реброву, проф.В.А.Громова и др.

Из зарубежных партнёров следует отметить Университет Блеза Паскаля (Prof. L.Nourine, University of Blaise Pascal (UBP), Clermont-Ferrand, France), Университет Техаса в Браунсвиле (Prof. F.Khan, Brownsville, USA), с которыми заключён договор об обмене студентами, Dr. J.Poelmans, Католический университет Лёвена (KU Leuven, Belgium), Prof. B.Ganter, S.Rudolph, Политехнический университет Дрездена (TU-Dresden, Germany), Prof. A.Napoli, Лотарингская лаборатория по информатике LORIA, Nancy, France), Prof. J.-F. Boulicaut, университет Лион-2 (Lyon-2, France), Медицинская школа Гарварда (Prof. O. Ryanykh), и другие исследователи, имеющие широкую международную известность, с которыми поддерживается тесное взаимодействие в области исследовательских и образовательных проектов. Предполагается, что сотрудники этих и других известных международных центров будут регулярно читать курсы и вести мастер-классы для студентов магистерской программы. Учеба на магистерской программе предусматривает участие в студенческих обменах с партнерскими университетами с возможностью получения второго диплома за рубежом.

8.2 Обеспеченность оборудованием

НИУ ВШЭ обладает классическими средствами для проведения занятий с использованием компьютеров: компьютерными классами с современными рабочими станциями и доступ к сети Интернет, как высокоскоростной проводной, так и беспроводной.

Однако для НОД необходим совершенно другой уровень обеспеченности доступом к вычислительным ресурсам и хранилищам данных. Планируется использовать мощности центра обработки данных компании Яндекс в режиме удалённого доступа. Также участникам МПНД будут доступны новейшие программные средства, от продуктов MathWorks (MatLab) до платформы InfoBright.