Название проекта

Науки о данных

Data Science

Ожидаемый результат проекта

Опишите результаты (изменения), которые Вы ожидаете получить благодаря получению гранта (не более 1500 знаков).

Цель проекта состоит в разработке и внедрении образовательной программы высшего образования уровня магистратуры «Науки о данных». Предполагается углубленное изучение фундаментальной информатики и компьютерных наук, статистики, прикладной математики, теоретической физики и теории информации. Каждый учебный курс будет обеспечен реальными практическими задачами и примерами их решения. Выпускники получат все необходимые компетенции, чтобы быть успешными для работы в области вычислительных сетей передачи данных, хранения данных, статистического анализа и быстрого извлечения аналитической информации из неструктурированных массивов разнотипных данных.

После завершения обучения, выпускники будут:

1. Способны эффективно и компетентно отвечать на возрастающие запросы экономики, касающиеся обработки, хранения, передачи данных.

2. Иметь сбалансированные теоретическую подготовку и практические навыки в прикладной математике, статистической физике, теории информации и компьютерных науках, чтобы анализировать многомерные мультимасштабные неупорядоченные наборы разнотипных данных.

3. Способны описывать и трансформировать информацию для того, чтобы открывать взаимосвязи и стройство сложных массивов данных.

4. Разрабатывать модели и алгоритмы решения задач, математические методы, методы теории информации и параллельной алгоритмизации.

5. Разрабатывать высокопроизводительное программное обеспечение суперкомпьютерных кластеров, распределенных вычислительных сетей и систем хранения данных.

Критерии оценки успешности проекта

Опишите, по каким показателям Вы планируете оценивать, насколько успешным/ востребованным оказался проект (не более 1200 знаков).

1) Доля выпускников магистратуры, успешно защитивших ВКР от общего число поступивших (не менее 70%)

2) Доля трудоустроенных магистров по специальности от общего числа выпускников

(не менее 80%)

3) Доля выпускников магистратуры, поступивших в аспирантуру, от общего числа выпускников (не менее 20%)

4) Доля выпускников магистратуры, имеющих зарегистрированные программные продукты в государственном Реестре программ ЭВМ и баз данных РФ, от общего числа выпускников (не менее 80%)

5) Доля магистров, участвующих в научно-исследовательских проектах (не менее 90%), от общего числа обучающихся

6) Доля выпускников магистратуры, получивших опыт вывода на открытый рынок, программных продуктов анализа данных (не менее 30%)

7) Доля выпускников магистратуры, получивших грамоты, награды, премии, поддержку инновационной деятельности (не менее 50%)

Краткая аннотация проекта

(не более 1500 знаков).

Проект направлен на разработку и внедрение двухлетней образовательной программы высшего образования уровня магистратуры «Науки о данных» в рамках действующего Федерального государственного стандарта высшего образования 09.04.02 «Информационные системы и технологии» на базе Школы естественных наук Дальневосточного федерального университета (ДВФУ). Результатом выполнения проекта станут учебные, учебно-методические материалы, документация.

Учебный план в соответствии с ФГОС 09.04.02 будет содержать: базовую часть, вариативную часть, обязательные дисциплины, дисциплины по выбору, учебные практику, научно-исследовательскую работу и семинары, производственную практику, государственную итоговую аттестацию.

Предполагается, что данная образовательная программа будет обеспечена преподавательским составом Школы естественных наук Дальневосточного федерального университета. Доля научно-педагогических работников, имеющих ученую степень и/или ученое звание, реализующих программу магистратуры, должна быть не менее: 80 процентов для программы академической магистратуры. Преподаватели будут проходить стажировки в ведущих научно-образовательных центрах мира.

Перечень материально-технического обеспечения, необходимого для реализации программы

магистратуры, включает в себя лаборатории, оснащенные лабораторным оборудованием, компьютерные классы, высокопроизводительное вычислительное и серверное оборудование, ленточное хранилище данных, а также системы хранения данных на основе жестких дисков.

Актуальность и востребованность проекта

(не более 1500 знаков).

Актуальность направления характеризует то, что с 2002 года начат выпуск специализированного журнала CODATA Data Science Journal, а с 2003 года выходит журнал The Journal of Data Science. С 2011 года O'Reilly проводит конференции по науке о данных — Strata, корпорация EMC с 2011 года проводит ежегодной саммит по науке о данных. McKinsey в 2011 году спрогнозировал спрос в США на 440—490 тыс. новых специалистов с «глубокими аналитическими навыками по работе с большими данными» к 2018 году и дефицит в 50 % — 60 % в таких специалистах при сохранении существующих образовательных трендов, в связи с этим прогнозом во многом был подогрет интерес к созданию учебных программ.

Российские крупные компании сегодня активно принимают на работу специалистов по данным. Например, представители компании Beeline на собеседовании на позизцию "Ученого по данным" (Data Scientist) задают вопросы по разделам математики, предлагают решить задачи машинного обучения, построить алгоритм, проверяют знания методов машинного обучения и анализа данных.

Средняя зарплата в США Data Scientist — 91 тысяча $ в год. В России от 60-70 т.р. в месяц для сотрудников без опыта работы и 220 т.р. для опытных специалистов. С 2012 года профессия ученого по данным отмечается как одна из самых привлекательных и перспективных в современном мире, поскольку такие специалисты позволят организациям технологических отраслей получить новые конкурентные преимущества благодаря анализу данных.

Описание структуры и содержания образовательного продукта

(не более 3000 знаков).

Образовательная программа магистратуры «Наука о данных» направлена на подготовку высококвалифицированных кадров, способных

1) собирать большие массивы неуправляемых данных и преобразовать их в более удобный формат;

2) решать бизнес-задачи с использованием данных;

3) работать с различными языками программирования, включая C++, Java, SAS, R и Python;

4) работать со статистикой, включая статистические тесты и распределения;

5) использовать аналитические методы, такие как машинное обучение, глубокое обучение, аналитика разнородных данных;

6) сотрудничать с ИТ и бизнесом в равной мере;

7) осуществлять поиск порядка и шаблонов данных, а также выявлять тенденции, которые могут помочь в достижении конечного бизнес-результата;

УЧЕБНЫЙ ПЛАН:

БАЗОВАЯ ЧАСТЬ

Иностранный язык в профессиональной сфере

Философские проблемы науки и техники

Методология научных исследований в информатике, математике и физике (злобина)

Проектирование систем анализа больших данных (нефедев)

ВАРИАТИВНАЯ ЧАСТЬ

ОБЯЗАТЕЛЬНЫЕ ДИСЦИПЛИНЫ

Суперкомпьютерное моделирование и технологии (шевченко)

Языки программирования

Системная инженерия, интеграция и управление большими данными

Дополнительные главы математической статистики, статистические тесты и распределения (кулешов)

ДИСЦИПЛИНЫ ПО ВЫБОРУ

Нейронные сети и машинное обучение (должиков)

Теория информации и кодирования (пустовалов)

Принципы распределенных систем

Высокопроизводительные вычисления и оптимизация (капитан)

Методы параллельной алгоритмизации (злобина)

Методы трансформации и визуализации данных (капитан)

Программно-аппаратные комплексы для численных расчетов (шевченко)

Методы глубокого обучения и искусственный интеллект (должиков)

Обработка и распознавание изображений (пустовалов)

Администрирование суперкомпьютерных систем (шевченко)

Системы поиска и интеллектуального анализа данных (капитан)

Современные методы распределенного хранения и обработки данных (фролов)

ПРАКТИКИ

УЧЕБНАЯ ПРАКТИКА

Практика по получению первичных профессиональных умений и навыков

Научно-исследовательская работа

Научно-исследовательский семинар "Бизнес-информатика"

Научно-исследовательский семинар "Быстрый анализ Больших Данных"

Научно-исследовательский семинар "Системы управления большими данными"

ПРОИЗВОДСТВЕННАЯ ПРАКТИКА

Практика по получению профессиональных умений и опыта организационно-управленческой; инновационной деятельности

Практика по получению профессиональных умений и опыта производственно-технологической; сервисно-эксплуатационной деятельности (в том числе технологическая практика)

Практика по получению профессиональных умений и опыта научно-педагогической деятельности

Практика по получению профессиональных умений и опыта проектной (проектно-конструкторской; проектно-технологической) деятельности

Научно-исследовательская работа

Преддипломная практика

ГОСУДАРСТВЕННАЯ ИТОГОВАЯ АТТЕСТАЦИЯ

Обзор существующих российских и зарубежных образовательных практик

Пожалуйста, опишите, какие образовательные практики и почему будут использованы при создании образовательного продукта. Как Вы планируете их адаптировать для целей образовательного продукта? (не более 3000 знаков).

С 2013 учебного года Университет Данди, Оклендский университет, Университет Южной Калифорнии запустили магистерские программы по науке о данных, а бизнес-школа Имперского колледжа Лондона — программу подготовки «магистров наук по науке о данных и менеджменту» (англ. MSc Data Science & Management). В том же году Вашингтонский университет, Университет Калифорнии в Беркли и Нью-Йоркский университет получили грант в размере $37,8 млн на развитие науки о данных, в рамках которого в течение пяти лет должны будут, в том числе, выстроить учебные программы и создать возможности для академической карьеры в данной области

В курсе введения в науку о данных Вашингтонского университета, опубликованном в системе Coursera, выделены следующие разделы:

- модели данных: отношения, «ключ — значение», деревья, графы, изображения, тексты;
- реляционная алгебра и параллельное выполнение запросов;
- NoSQL-системы и хранилища «ключ-значение»;
- компромиссы между SQL-, NoSQL- и NewSQL-системами;
- проектирование алгоритмов для Hadoop (и для MapReduce в общем случае);
- базовый статистический анализ: семплирование, регрессии;
- введение в data mining: кластеризация, ассоциативные правила, деревья решений;
- приложения: социальные сети, биоинформатика, анализ текста.

Блок науки о данных программы магистерской программы по «науке о данных и менеджменту» Императорского колледжа включает подготовительный курс «продвинутой статистики» (англ. advanced statistics), непосредственно в курс по науке о данных входят следующие дисциплины:

- машинное обучение;
- системы управления базами данных;
- инженерия программного обеспечения;
- анализ данных (англ. intelligent data) и вероятностный вывод (англ. probabilistic inference), в описании дисциплины даются ссылки на байесовский вывод и алгоритмические методы моделирования, классификации и дискриминантного анализа данных на его основе;
- вероятностные модели и продвинутая статистика.

После курсов по науке о данных и основам менеджмента в программе предусмотрен прикладной курс, разбитый на два потока, в финансово-технологический поток включены управление рисками, управление активами и производные финансовые инструменты, а в консалтинговый — обработка больших массивов данных (англ. large datasets), сетевой анализ, эконометрический анализ, приложения в сфере услуг и консалтинге, энергетике, здравоохранении, политике.

Программа Университета Данди делает упор на «большие данные», прежде всего, в противовес «табличной обработке», и фокусируется на интеллектуальном анализе данных, моделировании баз данных и хранилищ, статистике, в рамках программы изучаются языки SQL, MDX, R, Erlang, Java, инструменты Hadoop и NoSQL.

В российском сегменте высшего образования магистерская программа "Науки о данных (Data Science)" реализуется Национальным исследовательским университет «Высшая школа экономики». Образовательная программа предусматривает подготовку в области современных методов извлечения знаний из данных, математических методов моделирования и прогнозирования, современных программных систем и методов программирования для анализа данных. Бюджетных мест 55, 15 платных мест, 6 платных мест для иностранцев.

http://aboutdata.ru/2017/02/12/bigdato/

Методический задел

Перечислите и опишите опубликованные и/или разработанные Вами или членами проектной команды учебно-методические материалы (учебные пособия, методические пособия и т.п.) и иные методические разработок по теме проекта, которые будут использоваться. (не более 3000 знаков).

Разработаны методические материалы по следующей тематике:

1. Модели и методы интеллектуального анализа данных.

2. Нечеткие системы.

3. Методы и инструментальные средства управления проектами.

4. Высокопроизводительные сети.

5. Системный подход к проектированию информационных систем.

6. Основы высокопроизводительных вычислений

Учебные пособия и монографии руководителя проекта:

1) Афремов Л.Л., Белоконь В.И., Нефедев К.В., Кириенко Ю.В., Магнитные свойства нанодисперсных магнетиков, Издательство Дальневосточного федерального университета, Владивосток, 2010, 112 с.

2) Нефедев К.В. Введение в пакет Wien2k. Учебное пособие. Владивосток, Изд-во ДВГУ, 2009, 149 с., 150 экз.

3) Нефедев К.В. Введение в численные методы. Базовые конструкции языка С++. Учебное пособие. Владивосток, Изд-во ДВГУ, 2008, 136 с., 150 экз.

4) Нефедев К.В. Теория информации и основы кодирования, Учебное пособие. Владивосток, Изд-во ДВГУ, 2008, 136 с., 150 экз.

5) Нефедев К.В., Шевченко Ю.А., Капитан В.Ю., Суперкомпьютерное моделирование магнитных наноархитектур, Суперкомпьютерные

технологии в науке, образовании и промышленности / Под редакцией академика В.А. Садовничего, академика Г.И. Савина, чл.-

корр. РАН Вл.В. Воеводина. — М.: Издательство Московского университета, 2012, стр.79-86

6) Капитан В.Ю., Нефедев К.В., Высокопроизводительные расчеты магнитных свойств и моделирование неравновесных явлений в

нанопленках, 2014, Modeling, Simulation and Optimization of Complex Processes, Под редакцией: Hans Georg Bock, Xuan Phu Hoang,

Rolf Rannacher, Johannes P.Schlöder, Springer International Publishing, pp. 95-107

7) Нефедев К.В., Методические указания по выполнению лабораторных работ по курсу « Основы высокопроизводительных вычислений», изд. ДВГУ, 2011, рекомендовано к изданию на заседании кафедры компьютерных систем ШЕН ДВФУ протокол № 2 от 26.10.11., 61 стр.

8) Нефедев К.В., Учебное пособие «Основы высокопроизводительных вычислений», рекомендовано к изданию на заседании кафедры компьютерных систем ШЕН ДВФУ протокол № 2 от 26.10.11., 70 стр.

9)

Методология и методическая новизна продукта

(не более 1500 знаков).

Основная практическая цель профессиональной деятельности в науке о данных — обнаружение закономерностей в данных, извлечение знаний из данных в обобщённой форме, поэтому педагогические методы, которые планируется ипользовать для обучения магистров, будут направлены на эти виды профессиональной деятельности. Для получения компетенций в этой области деятельности необходимо использовать интегрированные взаимодополняющие методики, которые позволят получить общепредметный опыт, практический опыт в информационных технологиях (hacking skills) и знания математической статистики.

Особенность дисциплины состоит в приоритете практической применимости результатов, то есть, успешности предсказаний. Наука о данных основывается на методах классической статистики, в ней подразумевается исследование сверхбольших разнородных массивов цифровой информации и неразрывная связь с информационными технологиями, обеспечивающими их обработку. Профиль специалиста по науке о данных в меньшей степени требует концентрации на содержании предметных областей, но требует более глубоких знаний в математической статистике, машинном обучении, программировании, и в целом более высокого образовательного уровня (магистры, кандидаты наук, Ph.D в сравнении с бакалаврами и специалистами).

Для освоения студентами дисциплин образовательный программы магистратуры «Наука о данных» предполагается использование обучающих методов и методик авторской разработки.

Используемые технологии

Пожалуйста, опишите, какие технологии уже имеются в вузе; какое оборудование и программное обеспечение используется, а какое необходимо закупить для реализации проекта (не более 2000 знаков).

Для реализации образовательных программ ИТ направления и научных суперкомпьютерных вычислений в Дальневосточном федеральном университете имеется полностью оснащенный центр обработки данных созданный в соответствии с современными стандартами.

Центр обработки данных (ЦОД) ДВФУ размещается на первом этаже корпуса А в помещении А218 общей площадью 517 кв. м. Помещение ЦОД разбито на отсеки:

-серверная 53 кв. м;

-вент. камера 26 кв. м;

-электрощитовая 33 кв. м;

-помещение АУГПТ 12 кв. м.

Также имеются два технических помещения:

-техническое помещение 91 кв. м;

-техническое помещение 287 кв. м.

В центре установлен суперкомпьютерный вычислительный кластер, включающий в себя блэйд-центр (шасси для установки 10 лезвий Sun Blade 6000 в количестве 6 шт, лезвия Sun Blade X6250 – 60 шт, четырехядерные процессоры Xeon Model E5345 Quad-core 2.33 – 120 шт) для кластеризуемых задач и специального сервера (сервер x64 Sun Fire X4600 M2, восемь двухядерных процессоров AMD Opteron Model 8220 2.8GHz-dual-core) для некластеризуемых задач, требующих повышенной производительности процессора и больших объемов оперативной памяти. Общее количество вычислительных ядер блэйд-центра 480 ядер. Обмен данными между лезвиями осуществляется с помощью высокоскоростного соединения InfiniBand на 10 Гб/с.

Кафедра компьютерных систем Школы естественных наук ДВФУ, преподаватели которой будут осуществлять образовательную деятельность по данной программе магистратуры, также имеет компьютерные классы. Например, компьютерный укомплектованный 15 ПК, NVIDIA Quadra K2000 для образовательного процесса в качестве основы высокопроизводительной вычислительной лаборатории.

Для работы с большими данными в ЦОД установлены

1) система хранения данных HP 3PAR 10800 (6 node) 300 Тб
2) ленточная библиотека HP 103 eml 2,4 Пб.

Место в существующем учебном процессе

Пожалуйста, поясните, как проект соотносится с образовательными стандартами вуза (не более 2000 знаков).

Дальневосточный федеральный университет имеет лицензию на осуществление образовательной деятельности по направлению подготовки уровня магистратуры, соответствующему Федеральному Государственному Образовательному Стандарту 09.04.02 «Информационные системы и технологии». В ДВФУ разработан собственный образовательный стандарт магистратуры, который дополняет ФГОС 09.04.02.

Объем программы магистратуры должен составлять 120 зачетных единиц вне зависимости от формы обучения, применяемых образовательных технологий, за один учебный год 60 з.е.

Объектами профессиональной деятельности выпускников, освоивших программу магистратуры в соответствии с п. 4.2 ФГОС, являются информационные процессы, технологии, системы и сети, их инструментальное (программное, техническое, организационное) обеспечение, способы и методы проектирования, отладки, производства и эксплуатации информационных технологий и систем в областях: машиностроение, приборостроение, наука, техника, образование, медицина, административное управление, юриспруденция, бизнес, предпринимательство, коммерция, менеджмент, банковские системы, безопасность информационных систем, управление технологическими процессами, механика, энергетика, ядерная энергетика, связь, телекоммуникации, управление инфокоммуникациями, легкая промышленность, медицинские и биотехнологии, системы массовой информации и др., а также предприятия различного профиля и все виды деятельности в условиях экономики информационного общества.

На формирование каких компетенций направлен проект

Перечислите, пожалуйста, компетенции, которые будут формироваться у слушателей, и поясните, каким образом они будут влиять на востребованность выпускников данной магистерской программы на рынке труда: региональном, национальном, международном (не более 3000 знаков).

Data Scientist (ученый по данным) — это и аналитик данных и специалист по интеллектуальной обработке данных. Кроме того, предполагается узкая специализация по вычислительным высокопроизводительным системам, вычислительным сетям и средствам передачи данных, системам хранения данных.

Data Scientist должен знать и уметь применять:

- методы визуализации и трансформации данных;
- методы машинного обучения, искусственного интеллекта;
- методы глубокого обучения;
- теорию и методы распознавания образов;
- средства подготовки данных;
- методы текстовой аналитики;
- технологии быстрой обработки больших данных и облачных вычислений на предприятиях реальных секторов экономики, компенсировать расходы на хранение и доступ к информации.

Помимо прочего, нужно знать и понимать:

- Статистику и машинное обучение.
- Языки программирования SAS, R или Python.
- Базы данных MySQL и Postgres.
- Технологии визуализации данных и отчетности.

Выпускники программы магистратуры "Науки о данных" должны быть способны:

1) модернизировать или разрабатывать новые научные методы и способы деятельности

2) создавать новые теории, изобретать новые способы и инструменты профессиональной деятельности

3) самостоятельно осваивать новые методы исследований, изменять научный и производственный профиль своей деятельности

4) совершенствовать и развивать свой интеллектуальный и культурный уровень, строить траекторию профессионального развития и карьеры

5) анализировать, верифицировать, оценивать полноту информации в ходе профессиональной деятельности, при необходимости восполнять и синтезировать недостающую информацию

6) организовать многостороннюю коммуникацию и управлять ею

7) вести профессиональную и научно-исследовательскую деятельность в международной среде

8) организовать научно-исследовательскую деятельность

9) осуществлять целенаправленный многокритериальный поиск информации о новейших научных и технологических достижениях в сети Интернет и в других источниках.

10) использовать в профессиональной деятельности знания в области естественных наук, математики и информатики, понимание основных фактов, концепций, принципов теорий, связанных с прикладной математикой и информатикой.

11) разрабатывать математические модели и высокопроизводительные алгоритмы анализа данных

12) реализовывать алгоритмы и разрабатывать программное обеспечение

13) применить современные языки программирования и языки манипулирования данными, операционные системы, электронные библиотеки и пакеты программ, сетевые

технологии

14) разрабатывать теорию, модели, алгоритмы и инструментарий для работы с высокопроизводительными вычислительными системами

15) разрабатывать теорию, модели, алгоритмы и инструментарий для работы с системами хранения данных

16) разрабатывать теорию, модели, алгоритмы и инструментарий для работы с вычислительными сетями