МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

УТВЕРЖДАЮ
Директор ИЯТШ
_____О. Ю. Долматов
«26» 06 2020 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
ПРИЕМ 2019 г.
ФОРМА ОБУЧЕНИЯ очная

ОБРАБОТКА БОЛЬШИХ ОБЪЕМОВ ДАННЫХ

Направление подготовки/ специальность	**01.04.02 Прикладная математика и информатика**		
Образовательная программа (направленность (профиль))	**Математическое моделирование и компьютерные вычисления**		
Специализация			
Уровень образования	высшее образование - **магистратура**		
Курс	**2**	семестр	**2**
Трудоемкость в кредитах (зачетных единицах)	**3**		
Виды учебной деятельности	Временной ресурс		
Контактная (аудиторная) работа, ч	Лекции		**8**
	Практические занятия		**16**
	Лабораторные занятия		**24**
	ВСЕГО		**48**
Самостоятельная работа, ч			**60**
ИТОГО, ч			**108**

Вид промежуточной аттестации	**Диф. зачет**	Обеспечивающее подразделение	**ОЭФ ИЯТШ**

Заведующий кафедрой – руководитель отделения (на правах кафедры)		А.М. Лидер
Руководитель ООП		М.Е. Семенов
Преподаватель		М.Е. Семенов

2020 г.

1. Цели освоения дисциплины

Целями освоения дисциплины является формирование у обучающихся определенного ООП (п. 5.4 Общей характеристики ООП) состава компетенций для подготовки к профессиональной деятельности.

Код компетенции	Наименование компетенции	Индикаторы достижения компетенций		Составляющие результатов обучения	
		Код индикатора	Наименование индикатора достижения	Код	Наименование
УК(У)-4	Способен применять современные коммуникативные технологии, в том числе на иностранном (-ых) языке(-ах), для академического и профессионального взаимодействия	И.УК(У)-4.2	Использует информационно-коммуникационные технологии при поиске необходимой информации в процессе решения стандартных коммуникативных задач на государственном и иностранном (-ых) языках	УК(У)-4.В2	Владеет стратегиями представления результатов анализа и обработки информации
				УК(У)-4.У2	Умеет осуществлять поиск необходимой информации, проводить ее анализ и отбор для решения поставленных задач
				УК(У)-4.З2	Знает правила использования поисковых систем и баз данных для хранения, обработки и передачи информации
ОПК(У)-3	Способен разрабатывать математические модели и проводить их анализ при решении задач в области профессиональной деятельности	И.ОПК(У)-3.1	Использование фундаментальных результатов математики при разработке моделей	ОПК(У)-3.В3	Владеет навыками разработки математических и статистических моделей данных, моделей машинного обучения в области профессиональных деятельности
				ОПК(У)-3.У3	Умеет использовать основные математические модели, умеет строить вычислительные алгоритмы для обработки данных в области профессиональных деятельности
				ОПК(У)-3.З3	Знает методы разработки математических моделей в области профессиональных деятельности
		И.ОПК(У)-3.2	Использование фундаментальных результатов математики для анализа моделей	ОПК(У)-3.В4	Владеет навыками применения общих положений математических дисциплин для анализа моделей при решении задач в профессиональной деятельности
				ОПК(У)-3.У4	Умеет использовать фундаментальные и прикладные знания математических дисциплин для анализа моделей в области профессиональной деятельности
				ОПК(У)-3.З4	Знает методы анализа математических моделей в области профессиональных деятельности
ОПК(У)-4	Способен комбинировать и адаптировать существующие информационно-	И.ОПК(У)-4.1	Применение современных информационно-коммуникационных технологий	ОПК(У)-4.В1	Владеет навыками компьютерной обработки вычислительных задач
				ОПК(У)-4.У1	Умеет строить математические алгоритмы, модели и реализовывать их с помощью

Код компетенции	Наименование компетенции	Индикаторы достижения компетенций		Составляющие результатов обучения	
		Код индикатора	Наименование индикатора достижения	Код	Наименование
	коммуникационные технологии для решения задач в области профессиональной деятельности с учетом требований информационной безопасности				языков программирования
				ОПК(У)-4.З1	Знает стратегии тестирования и отладки программного обеспечения
				ОПК(У)-4.В2	Владеет навыками использования прикладного программного обеспечения для решения задач в профессиональной деятельности
				ОПК(У)-4.У2	Умеет применять математический язык, методы при построении моделей объектов профессиональной деятельности с использованием инструментальных средств компьютерного моделирования
				ОПК(У)-4.З2	Знает профессиональную терминологию, содержание ключевых понятий и определений, используемых в теории и практике применения информационных технологий в науке и образовании
				ОПК(У)-4.В3	Владеет навыками работы с программными продуктами и информационными ресурсами
				ОПК(У)-4.У3	Умеет самостоятельно расширять и углублять знания в области информационно-коммуникационных технологий
				ОПК(У)-4.З3	Знает средства интеграции приложений и операционных систем
ПК(У)-1	Способен проводить научные исследования и получать новые научные и прикладные результаты самостоятельно и в составе научного коллектива	И.ПК(У)-1.2	Формирует и создает перечень возможных методов решения, обеспечивающих проведение научных исследований	ПК(У)-1.В2	Владеет наукоемкими технологиями и пакетами прикладных программ для решения прикладных задач
				ПК(У)-1.У2	Умеет самостоятельно выбирать эффективные методы решения поставленных задачи разрабатывать новые методы для получения новых научных и прикладных результатов
				ПК(У)-1.З2	Знает классические методы, применяемые в прикладной математике и информатике; необходимые и достаточные условия их реализации

2. Место дисциплины (модуля) в структуре ООП

Дисциплина относится к вариативной части Блока 1 учебного плана образовательной программы.

3. Планируемые результаты обучения по дисциплине

После успешного освоения дисциплины будут сформированы результаты обучения:

Планируемые результаты обучения по дисциплине		Компетенция
Код	Наименование	
РД-1	Выполнять исследования процессов создания, накопления и обработки информации, включая анализ и создание моделей данных и знаний, языков их описания и манипулирования.	ОПК(У)-3
РД-2	Владеть методами исследования и обработки данных и их применению в самостоятельной научно-исследовательской и профессиональной деятельности.	ОПК(У)-4
РД-3	Владение методами и инструментами визуализации и ординации многомерных объектов	ПК(У)-1, УК(У)-4

Оценочные мероприятия текущего контроля и промежуточной аттестации представлены в календарном рейтинг-плане дисциплины.

4. Структура и содержание дисциплины

Основные виды учебной деятельности

Разделы дисциплины	Формируемый результат обучения по дисциплине	Виды учебной деятельности	Объем времени, ч.
Раздел (модуль) 1. Стратегии работы с большими объемами данных	РД-1	Лекции	**2**
		Практические занятия	**4**
		Лабораторные занятия	**8**
		Самостоятельная работа	**20**
Раздел (модуль) 2. Обработка больших объемов данных	РД-2, РД-3	Лекции	**4**
		Практические занятия	**8**
		Лабораторные занятия	**8**
		Самостоятельная работа	**20**
Раздел (модуль) 3. Визуализация и ординация многомерных объектов	РД-2, РД-3	Лекции	**2**
		Практические занятия	**4**
		Лабораторные занятия	**8**
		Самостоятельная работа	**20**

Содержание разделов дисциплины:

Раздел 1. ***Стратегии работы с большими объемами данных***

Форматы данных и типы моделей. Преобразование форматов. Структурированный язык запросов. Широкий и длинный формат данных. Предпроцессинг. Манипуляция. Чтение/запись, кэширование. Функции чтения/записи данных из/в текстовый файл. Справочная система R/Python. Платформы Hadoop, Spark, библиотеки sparklyr, modeldb, фреймворк h2o. Репрезентативная выборка ограниченного размера. Разбиение данных на части. Выполнение вычислений на стороне базы данных. Кэширование вычислений.

Темы лекций:

1. Введение в большие данные. Стратегии работы с большими объемами данных

Темы практических занятий:

1. Языки, программные среды, фреймворки для обработки данных.
2. Платформы для выполнения распределенных вычислений.

Темы лабораторных занятий:

1. Функции чтения/записи данных из/в текстовый файл.

2. Справочная система R/Python.
3. Стратегии работы с большими массивами данных.
4. Кэширование вычислений.

Раздел 2. ***Обработка больших объемов данных***

Числовые и текстовые данные. Отсутствующие данные. Ошибки в данных. Выбросы в данных. Дублирующие наблюдения (строки). Мультиколлинеарность. Цифровизация данных. Методы фильтрации, обертки, смешанные методы. Предсказание значений. График boxplot. Предиктивные модели. Дискретная, непрерывная, категориальная переменная. Шкалы. Машинное обучение с учителем, без учителя. Переобучение. Метод опорных векторов. Метод k-ближайших соседей. Выборочные характеристики положения и их использование для заполнения пропусков. Графический способ определения выбросов. Принципы разделения выборки на обучающую и тестовую. Методы преобразования непрерывных данных в категориальные. Классификация и регрессия. Линейные, нелинейные модели.

Темы лекций:

2. Базовые алгоритмы интеллектуального анализа данных (Data mining). Подготовка исходных данных к обработке.
3. Выбор признаков (Feature Selection), экземпляров (Instance Selection), дискретизация для классификации (Discretization).

Темы практических занятий:

3. Графический разведочный анализ данных.
4. Статистические методы заполнения пропусков и определения выбросов.
5. Определение важности признаков.
6. Классификация и регрессия. Линейные, нелинейные модели.

Темы лабораторных занятий:

5. Обработка пропусков в данных.
6. Графический способ определения выбросов.
7. Преобразование и дискретизация данных.
8. Формирование обучающей и тестовой выборки.

Раздел 3. Визуализация и ординация многомерных объектов

Снижение размерности. Метод главных компонент. Метод главных координат. Неметрическое многомерное шкалирование, анализ соответствий. Функции многомерного факторного анализа. Кластерный анализ. Дендограммы. Оптимальное проецирование. Ординационные диаграммы.

Темы лекций:

4. Алгоритмы сжатия информационного пространства

Темы практических занятий:

7. Методы визуализации многомерных объектов
8. Методы ординации многомерных объектов.

Темы лабораторных занятий:

9. Системы визуализации ggplot2, GGVIS. Ординационные диаграммы.
10. Метод главных компонент, координат.
11. Кластерный анализ. Дендограммы.
12. Неметрическое многомерное шкалирование.

5. Организация самостоятельной работы студентов

Самостоятельная работа студентов при изучении дисциплины (модуля) предусмотрена в следующих видах и формах:

- Работа с лекционным материалом, поиск и обзор литературы и электронных источников информации по индивидуально заданной проблеме курса;
- Изучение тем, вынесенных на самостоятельную проработку;
- Поиск, анализ, структурирование и презентация информации;
- Перевод текстов с иностранных языков;
- Выполнение домашних заданий;
- Программные расчеты;
- Подготовка к практическим занятиям;
- Подготовка к лабораторным занятиям;
- Подготовка к оценивающим мероприятиям.

6. Учебно-методическое и информационное обеспечение дисциплины

6.1. Учебно-методическое обеспечение

1. Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R. Москва: ДМК Пресс, 2014. — 588 с. // Лань: электронно-библиотечная система. — URL: https://e.lanbook.com/book/58703
2. Буховец, А. Г. Алгоритмы вычислительной статистики в системе R: учебное пособие / А. Г. Буховец, П. В. Москалев. — 2-е изд., перераб. и доп. — Санкт-Петербург : Лань, 2015. — 160 с. // Лань: электронно-библиотечная система. — URL: https://e.lanbook.com/book/68459
3. Введение в статистическое обучение с примерами на языке R / Г. Джеймс, Д. Уиттон, Т. Хасти, Р. Тибширани; перевод с английского С. Э. Мастицкого. — Москва : ДМК Пресс, 2017. — 456 с. // Лань: электронно-библиотечная система. — URL: https://e.lanbook.com/book/93580
4. Москвитин, А. А.. Данные, информация, знания: методология, теория, технологии: монография [Электронный ресурс] / Москвитин А. А.. — Санкт-Петербург: Лань, 2019. — 236 с. https://e.lanbook.com/book/113937
5. Шитиков В.К., Мастицкий С.Э. (2017) Классификация, регрессия и другие алгоритмы Data Mining с использованием R. 351 с. – Электронная книга, адрес доступа: https://github.com/ranalytics/data-mining

Дополнительная литература

1. Olvera-López J., Carrasco-Ochoa J. Martínez-Trinidad J. F. and Kittler J. (2010). A review of instance selection methods. Artif. Intell. Rev. 34. 133-143. https://mafiadoc.com/a-review-of-instance-selection-methods-soft-computing-and-_5b054f698ead0ed4758b4586.html
2. X. Wu et al. (2008) Top 10 algorithms in data mining. Knowl. Inf. Syst. 14. 1–37. http://www.cs.umd.edu/~samir/498/10Algorithms-08.pdf
3. Лесковец, Юре. Анализ больших наборов данных: пер. с англ. / Ю. Лесковец, А. Раджараман, Дж. Ульман. — Москва: ДМК Пресс, 2016. — 498 с.
4. Фрэнкс, Билл. Укрощение больших данных. Как извлекать знания из массивов информации с помощью глубокой аналитики: пер. с англ. / Б. Фрэнкс. — Москва: Манн, Иванов и Фербер, 2014. — 340 с.
5. Хименко, Виталий Иванович. Случайные данные: структура и анализ: учебник / В. И. Хименко. — Москва: Техносфера, 2019. — 424 с.

6. Орельен Жерон, Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow. Концепции, инструменты и техники для создания интеллектуальных систем. М.: Вильямс. - 2018. - 688 с.
7. Хэдли Уикем, Гарретт Гроулмунд, «Язык R в задачах науки о данных: импорт, подготовка, обработка, визуализация и моделирование данных». М.: Вильямс. - 2018. - 592 с.
8. Черняк Л. Серьезно о технологиях для Больших Данных // Открытые системы. СУБД, - 2014. - № 1. http://www.osp.ru/os/2014/01/13039646/

6.2. Информационное и программное обеспечение

Internet-ресурсы (в т.ч. в среде LMS MOODLE и др. образовательные и библиотечные ресурсы):

- Персональная страница Семенова М.Е. *http://portal.tpu.ru/SHARED/s/SME/work*
- Курс «Big Data» *https://www.coursera.org/specializations/big-data*
- Анализ данных в Spark-кластере с помощью пакета dplyr https://r-analytics.blogspot.com/2020/03/spark-dplyr.html
- Локальный Spark-кластер: устанавливаем, подключаемся, пробуем https://r-analytics.blogspot.com/2020/02/spark-r-connect.html
- Spark и sparklyr для работы с большими данными в R https://r-analytics.blogspot.com/2020/02/spark-intro.html
- Конференция по большим данным и искусственному интеллекту *https://bigdatadays.ru/ru/*
- Профессиональные базы данных и информационно-справочные системы доступны по ссылке: https://www.lib.tpu.ru/html/irs-and-pdb

Лицензионное программное обеспечение (в соответствии с **Перечнем лицензионного программного обеспечения ТПУ)**:

1. 7-Zip;
2. Adobe Acrobat Reader DC;
3. Adobe Flash Player;
4. AkelPad;
5. Cisco Webex Meetings;
6. Document Foundation LibreOffice;
7. Google Chrome;
8. Microsoft Office 2007 Standard Russian Academic;
9. Mozilla Firefox ESR;
10. PTC Mathcad Prime 6 Academic Floating;
11. Tracker Software PDF-XChange Viewer;
12. WinDjView
13. Zoom Zoom

7. Особые требования к материально-техническому обеспечению дисциплины

В учебном процессе используется следующее лабораторное оборудование для практических и лабораторных занятий:

№	Наименование специальных помещений	Наименование оборудования
1.	Аудитория для проведения учебных занятий всех типов, курсового проектирования, консультаций, текущего контроля и промежуточной аттестации	Доска аудиторная настенная - 1 шт.;Шкаф для одежды - 1 шт.;Шкаф для документов - 1 шт.;Комплект учебной мебели на 10 посадочных мест;

		Office 2007 Standard Russian Academic; Mozilla Firefox ESR; Tracker Software PDF-XChange Viewer; WinDjView; Zoom Zoom

Рабочая программа составлена на основе Общей характеристики образовательной программы по направлению 01.04.02 Прикладная математика и информатика, профиль «Математическое моделирование и компьютерные вычисления» (приема 2019 г., очная форма обучения).

Разработчик(и):

Должность	Подпись	ФИО
доцент		Семенов М.Е.

Программа одобрена на заседании отделения экспериментальной физики ИЯТШ (протокол № 6 от 20.06.2019).

Заведующий кафедрой – руководитель отделения (на правах кафедры) экспериментальной физики ИЯТШ:
д. т. н. _____/Лидер А. М./
подпись