МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
«Дальневосточный федеральный университет»
(ДВФУ)

ШКОЛА ЕСТЕСТВЕННЫХ НАУК

«СОГЛАСОВАНО»	«УТВЕРЖДАЮ»
Руководитель ОП	Заведующая (ий) кафедрой
д.ф.-м.н., профессор, академик РАН, Гузев М.А.	информатики, математического и компьютерного моделирования
	(название кафедры)
_____	_____ Чеботарев А.Ю.
(подпись) (Ф.И.О. рук. ОП)	(подпись) (Ф.И.О. зав. каф.)
«23» июня 2017 г.	«23» июня 2017 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
Наука о данных и аналитика больших объемов данных

Направление подготовки 09.03.03Прикладная информатика

Форма подготовки очная

курс 4 семестр 7
лекции час.
практические занятия 36 час.
лабораторные работы час.
в том числе с использованием МАО лек._____/пр._____/лаб._____ час.
всего часов аудиторной нагрузки 36 час.
в том числе с использованием МАО _____ час.
самостоятельная работа 36 час.
в том числе на подготовку к экзамену час.
контрольные работы (количество)
курсовая работа / курсовой проект _____ семестр
зачет _семестр
экзамен7семестр
Рабочая программа составлена в соответствии с требованиями образовательного стандарта, самостоятельно установленного ДВФУ, принятого решением Ученого совета Дальневосточного федерального университета, протокол от 28.01.2016 № 01-16, и введенного в действие приказом ректора ДВФУ от 18.02.2016 № 12-13-235.
Рабочая программа обсуждена на заседании кафедры информатики, математического и компьютерного моделирования, протокол № 22 «23» июня 2017 г.
Заведующий кафедрой Чеботарев А.Ю.
Составитель:

Оборотная сторона титульного листа РПУД

I. Рабочая программа пересмотрена на заседании кафедры:

Протокол от «_____» _____ 20___ г. № _____

Заведующий кафедрой _____ _____
(подпись) (И.О. Фамилия)

II. Рабочая программа пересмотрена на заседании кафедры:

Протокол от «_____» _____ 20___ г. № _____

Заведующий кафедрой _____ _____
(подпись) (И.О. Фамилия)

Аннотация
Наука о данных и аналитика больших объемов данных

Целями освоения дисциплины является овладение студентами знаниями о методологиях и технологиях Big Data для обработки, хранения и использования больших данных. Изложены методы обработки неструктурированной информации, серия подходов и инструментарий больших данных. Представлены современное состояние и тенденции развития технологий Big Data.

Минимальные требования к «входным» знаниям, необходимым для успешного усвоении данной дисциплины - удовлетворительное усвоение программы дисциплин: «Объектно-ориентированный анализ и проектирование», «Базы данных», «Сетевые технологии и системное администрирование» - в полном объеме

Компетенции обучающегося, формируемые в результате освоения дисциплины:

ПК-15 Способностью осуществлять ведение базы данных и поддержку информационного обеспечения решения прикладных задач	знает	историю развития методологий проектирования Базовые понятия технологии Big Data, базовые понятия прогнозирования, основные технологии прогнозирования
	умеет	определять массивы больших данных; Анализировать кластеры больших данных, строить различными способами прогнозы развития социально-политических процессов
	владеет	Терминологией курса, современными технологиями создания и обслуживания больших данных, методологией и методикой прогнозирования

Структура и содержание дисциплины

Общая трудоемкость дисциплины составляет 2 зачётных единицы, 72 часа. Из них 36 часов практических работ и 36 часов самостоятельных работ.

Тема 1. Определение больших данных. Технологии хранения больших данных.

Большие данные (big data) в информационных технологиях. Совокупность подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам вычислительной сети, сформировавшихся в конце 2000-х годов, альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence. В данную серию включают средства массово-параллельной обработки неопределённо структурированных данных, прежде всего, решениями категории NoSQL, алгоритмами MapReduce, программными каркасами и библиотеками проекта Hadoop.

качестве определяющих характеристик для больших данных отмечают три: V объём (англ. volume, в смысле величины физического объёма), скорость (англ. velocity в смыслах как скорости прироста, так и необходимости высокоскоростной обработки и получения результатов), многообразие (англ. variety, в смысле возможности одновременной обработки различных типов структурированных и полуструктурированных данных).

Тема 2. Процесс анализа больших данных. Технологии анализа больших данных. Научные проблемы в области больших данных. Методы и техники анализа, применимые к большим данным:

- методы класса Data Mining: обучение ассоциативным правилам (англ. association rule learning), классификация
- (методы категоризации новых данных на основе принципов, ранее применённых к уже наличествующим данным), кластерный анализ, регрессионный анализ;
- краудсорсинг - категоризация и обогащение данных силами широкого, неопределённого круга лиц, привлечённых на основании публичной оферты, без вступления в трудовые отношения;
- смешение и интеграция данных (англ. data fusion and integration) - набор техник, позволяющих интегрировать разнородные данные из разнообразных источников для возможности глубинного анализа, в качестве примеров таких техник, составляющих этот класс методов приводятся цифровая обработка сигналов и обработка естественного языка (включая тональный анализ);
- машинное обучение, включая обучение с учителем и без учителя, а также Ensemble learning (англ.) - использование моделей, построенных на базе статистического анализа или машинного обучения для получения комплексных прогнозов на основе базовых моделей (англ. constituent models, ср. со статистическим ансамблем в статистической механике);
- искусственные нейронные сети, сетевой анализ, оптимизация, в том числе генетические алгоритмы;
- распознавание образов;
- прогнозная аналитика;
- имитационное моделирование;

- пространственный анализ (англ. Spatial analysis) - класс методов, использующих топологическую, геометрическую и географическую информацию в данных;
- статистический анализ, в качестве примеров методов приводятся A/B-тестирование и анализ временных рядов;
- визуализация аналитических данных - представление информации в виде рисунков, диаграмм, с использованием интерактивных возможностей и анимации как для получения результатов, так и для использования в качестве исходных данных для дальнейшего анализа.

Тема 3. Прогнозирование и предвидение в социально-политических и медиа процессах. Методы прогнозирования.

Понятие прогноза и предвидения. Отличие прогнозирования от предвидения. Закон распределения случайной величины. Статистические оценки параметров. Доверительные области. Теория моментов. Корреляционный анализ. Использование модели множественной линейной регрессии для прогнозирования экономических показателей. Доверительные интервалы для зависимой переменной. Сглаживание временных рядов. Динамические модели с распределенными лагами. Стационарные временные ряды. Тестирование стационарности. Коинтеграция. Анализ временных рядов. Адаптивные и мультипликативные методы прогнозирования. Экспоненциальное сглаживание. Авторегрессионные модели. Модели скользящего среднего. Интегрированные процессы. Идентификация авторегрессионной модели скользящего среднего. Прогнозирование с моделями временных рядов. Доверительные интервалы прогноза. Дисперсионный анализ влияния качественных факторов. Ранговые методы. Факторный анализ. Метод главных факторов. Многомерное шкалирование. Классическая модель многомерного шкалирования. Неметрические методы. Кластерный анализ. Дискриминантный анализ. Многомерный статистический анализ.

Тема 4. Программы статистической обработки информации. Представление возможностей пакета SPSS Statistics для целей анализа социально-политических процессов.

SPSS Statistics (аббревиатура англ. "Statistical Package for the Social Sciences", "статистический пакет для социальных наук") - компьютерная программа для статистической обработки данных, один из лидеров рынка в области коммерческих статистических продуктов, предназначенных для проведения прикладных исследований в социальных науках. Применение программы для решения прикладных задач прогнозирования: ввод и хранение данных; возможность использования переменных разных типов; частотность признаков, таблицы, графики, таблицы сопряжённости, диаграммы; первичная описательная статистика; маркетинговые и медиа исследования; анализ данных маркетинговых и медиа исследований.

Методические указания по организации самостоятельной работы студентов

Планируются следующие виды самостоятельной работы (внеаудиторной):
- подготовка к лабораторным работам,
- оформление отчётов по лабораторным работам,
- работа с конспектом лекций и изучение рекомендованной литературы при подготовке к экзамену.

Примерный перечень вопросов и заданий к экзамену

1. Понятие Большие данные. Роль цифровой информации в 21 веке.
2. Виды массивов данных.

3. Базовые принципы обработки больших данных.
4. Технологии обработки больших данных: NoSQL, MapReduce, Hadoop, R.
5. Технологии Business Intelligence и реляционные системы управления базами данных.
6. Прогнозирование и предвидение: общее и особенное.
7. Виды прогнозов
8. Общие методы анализа социально-политических и медиа процессов.
9. Специальные методы анализа социально-политических и медиа процессов.
10. Предварительный анализ данных.
11. Проверка гипотез о законе распределения случайной величины.
12. Статистические оценки параметров. Доверительные области.
13. Теория моментов.
14. Корреляционный анализ.
15. Использование модели множественной линейной регрессии для прогнозирования экономических показателей. Доверительные интервалы для зависимой переменной.
16. Сглаживание временных рядов. Динамические модели с распределенными лагами.
17. Стационарные временные ряды. Тестирование стационарности.
18. Коинтеграция. Анализ временных рядов.
19. Адаптивные и мультипликативные методы прогнозирования. Экспоненциальное сглаживание.
20. Авторегрессионные модели. Модели скользящего среднего.
21. Интегрированные процессы. Идентификация авторегрессионной модели скользящего среднего.
22. Прогнозирование с моделями временных рядов. Доверительные интервалы прогноза.
23. Предсказание и прогнозирование социально-экономических прогнозов.
24. Дисперсионный анализ влияния качественных факторов. Ранговые методы.
25. Факторный анализ. Метод главных факторов.
26. Многомерное шкалирование. Классическая модель многомерного шкалирования.

27. Неметрические методы. Кластерный анализ. Дискриминантный анализ.

28. Многомерный статистический анализ.

29. Статистический анализ в пакете SPSS Statistics.

30. Основные возможности пакета SPSS Statistics.

Основная литература:

1. Прогнозирование и планирование в условиях рынка : учеб. пособие / Т.Н. Бабич, И.А. Козьева, Ю.В. Вертакова, Э.Н. Кузьбожев. / М. : ИНФРА-М, 2017. / 336 с. / (Высшее образование: Бакалавриат) - Режим доступа: http://znanium.com/catalog/product/851194
2. Социально-экономическое прогнозирование: Учебное пособие / Герасимов А.Н., Громов Е.И., Скрипниченко Ю.С. - М.:СтГАУ - 'Агрус', 2017. - 144 с.: Режим доступа: http://znanium.com/catalog/product/975933
3. Методы хранения и обработки данных: Учебник / Дадян Э.Г. - М.:НИЦ ИНФРА-М, 2018: - Режим доступа: http://znanium.com/catalog/product/989190

Дополнительная литература:

1. Современные базы данных. Основы. Часть 1: Учебное пособие / Дадян Э.Г. - М.:НИЦ ИНФРА-М, 2017. - 88 с.: 60x90 1/16 ISBN 978-5-16-106526-6 (online) - Режим доступа: http://znanium.com/catalog/product/959289
2. Базы данных : учеб. пособие / О.Л. Голицына, Н.В. Максимов, И.И. Попов. ? 4-е изд., перераб. и доп. ? М. : ФОРУМ : ИНФРА-М, 2019. ? 400 с. ? (Высшее образование: бакалавриат). - Режим доступа: http://znanium.com/catalog/product/1019244
3. Проектирование современных баз данных. Практикум: Учебно-методическое пособие / Дадян Э.Г. - М.:НИЦ ИНФРА-М, 2017. - 84 с.: 60x90 1/16 ISBN 978-5-16-106528-0 (online) - Режим доступа: http://znanium.com/catalog/product/959294
4. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля) The Gallup Organization World Wide Web Server - www.gallup.com/index.html Большие данные - https://habrahabr.ru/hub/bigdata/
5. Центр управления финансами - Методы прогнозирования - http://center-yf.ru/data/Marketologu/Metody-prognozirovaniya.php

Перечень информационных технологий, используемых при осуществлении образовательного процесса

- Операционная система Microsoft Windows Professional 7 Russian
- Пакет офисного программного обеспечения Microsoft Office 2010 Professional Plus Russian
- Браузер Mozilla Firefox

- Браузер Google Chrome
- Adobe Reader XI

Описание материально-технической базы

Мультимедийная аудитория. Мультимедийная аудитория состоит из интегрированных инженерных систем с единой системой управления, оснащенная современными средствами воспроизведения и визуализации любой видео и аудио информации, получения и передачи электронных документов. Типовая комплектация мультимедийной аудитории состоит из: мультимедийного проектора, автоматизированного проекционного экрана, акустической системы, а также интерактивной трибуны преподавателя, включающей тач-скрин монитор с диагональю не менее 22 дюймов, персональный компьютер (с техническими характеристиками не ниже Intel Core i3-2100, DDR3 4096Mb, 500Gb), конференц-микрофон, беспроводной микрофон, блок управления оборудованием, интерфейсы подключения: USB,audio, HDMI. Интерактивная трибуна преподавателя является ключевым элементом управления, объединяющим все устройства единую систему, и служит полноценным рабочим местом преподавателя. Преподаватель имеет возможность легко управлять всей системой, не отходя от трибуны, что позволяет проводить лекции, практические занятия, презентации, вебинары, конференции и другие виды аудиторной нагрузки обучающихся в удобной и доступной для них форме с применением современных интерактивных средств обучения, в том числе с использованием в процессе обучения всех корпоративных ресурсов. Мультимедийная аудитория также оснащена широкополосным доступом в сеть интернет. Компьютерное оборудованием имеет соответствующее лицензионное программное обеспечение.

Компьютерный класс, представляющий собой рабочее место преподавателя и не менее 15 рабочих мест студентов, включающих компьютерный стол, стул, персональный компьютер, лицензионное программное обеспечение. Каждый компьютер имеет широкополосный доступ в сеть Интернет. Все компьютеры подключены к корпоративной компьютерной сети КФУ и находятся в едином домене.