EXPERIENCE IN ORGAANIZING EDUCATIONAL PROCESS IN BIG DATA ANALYTICS AT BSUIR

M. BATURA, Doctor of Engeneering Sciences
Rector BSUIR, Full Professor, Member of the International Higher Education Academy of Sciences, Honored Worker of Education of the Republic of Belarus

S.K. DZIK, PhD
First BSUIR vice-rector, PhD, associate professor

B. ZIBITSKER, PhD
President and CEO BEZNext, Emeritus professor of BSUIR

D. LIKHACHEVSKY, PhD
BSUIR Faculty of Computer-aided Design Dean, PhD, associate professor

I. TSYRELCHUK, PhD
BSUIR Lifelong and E-learning Studies Faculty Dean, Information and Computer-aided Systems Design Department Head, PhD, associate professor

K. YASHIN, PhD
BSUIR Engineering Psychology and Ergonomics Department Head, PhD, associate professor

The Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: sdick@bsuir.by, bzibitsker@beznext.com, likhachevskyd@bsuir.by, tsyrelchuk@gmail.com, yashin@bsuir.by

Abstract. For the last three years BSUIR has been offering Big Data educational courses. The initial course was conducted by Dr. Boris Zibitsker and Dr. Dominique Heger remotely. One part of the course included theory and the second one was Capstone project. Many students faced the problem of understanding the English materials and so the consequent courses were read in Russian. We always look for the way to improve educational process. Currently we use IBM's Big Data University Ambassodar program and students have a remote access to IBM's Big Data clusters to complete exercises. Fourth year students of Computer-aided Design Faculty took introductory courses and currently there are 58 3rd year students studying Fundamentals of BIG DATA, Introduction to data analysis and R technology, and Machine learning algorithms.

The classes are conducted in Russian and the students complete the exercises at the time suitable for them. This approach of conducting classes in Russian and completing exercises using IBM resources improved the process significantly. We are working on expending Big Data curriculum and offering courses to the students and PhD candidates of different faculties.

Introduction. The demand of modern production, the tasks of the banking sector, the development of nuclear energy and rapidly developing scientific research - all these and many other factors necessitate the training of specialists with higher education in the field of BIG DATA. In order to meet these requirements, as well as pass ahead of the development of the national economics and science, the Belarusian State University of Informatics and Radioelectronics started looking for solutions to the problems of training specialists in Big Data Analytics several years ago.

In this presentation we will share our experience in organizing the educational process in the Big Data Analytics and the plan of expanding the program at different departments and PhD programs.

Training of the specialists for working with large amounts of information - Big Data Analytics

Here are the main steps taken by the University for solving the problem of specialists with higher education training over the past 3 years.

1) There was organized scientific and technical cooperation with BEZNext (USA) specializing in the field of BIG DATA.

2) Specialized scientific and practical conference BIG DATA and Advanced Analytics was organized to be held annually. This conference enables specialists to discuss the development and application of modern information technologies in this area.

3) Commercial courses were set up. The courses provide teaching in English for those who want to learn the basic approaches and technologies of BIG DATA.

4) A supercomputer suitable for processing large amounts of information was purchased.

5) Master students of the Faculty of Computer Systems and Networks conduct research in the field of BIG DATA as a part of Master's theses.

6) Practice-oriented Master course of the Faculty of Computer Systems and Networks opened a new specialty for studying BIG DATA technologies.

7) Business relations with IBM employees (USA), who developed the educational program Ambassador and organized BIG DATA University to study modern technologies for processing large amounts of information, were established.

8) Three teachers were trained to conduct classes on BIG DATA in Russian with students of the Computer-aided Design Faculty.

9) Finally, the classes aimed at studying the main technologies of BIG DATA are organized for the 4th year students of Computer-aided Design Faculty.

The study of BIG DATA technologies was organized in the autumn semester of 2016 for the students of the 4th year (a group of 25 people). Students are trained in the field of Information Systems and Technology (in ensuring industrial safety), the qualification obtained is a system engineer, and the period of studies presupposes 4 years of full-time education.

Together with the covering of two planned IT disciplines 1) System software and 2) Modern programming languages, students were offered the Big Data Analytics materials designed with the practical experience of the IBM BDU program.

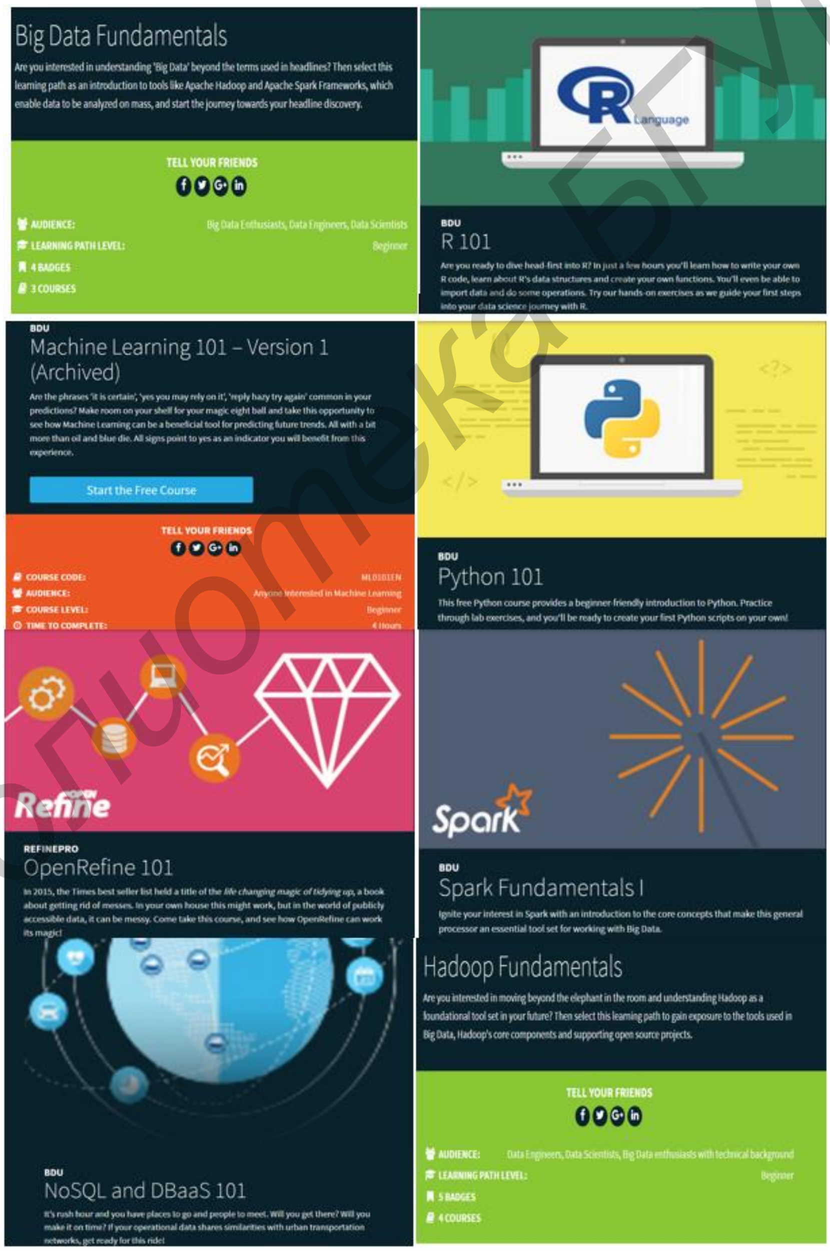
The Ambassador program and BDU are organized by IBM specialists. The founders of the noted educational programs are IBM specialists working in the USA, Canada, India and other countries.

Universities of the USA, China, and Russia have been using educational resources of IBM Big Data University (BDU); BSUIR joined the BDU programs at the end of 2016.

In July-August 2016 three teachers of the Engineering Psychology and Ergonomics Department (BSUIR), which trains system engineers in two specialties: 1) Engineering and psychological support of information technology and 2) Information systems and technologies (in ensuring industrial safety), began the preparation of BIG DATA training materials for working with the students.

The educational program Ambassador BDU includes eight following sections (modules).

1) BIG DATA Fundamentals
2) Introduction to Data Analysis Using R
3) Machine Learning Algorithms
4) Python
5) Introduction to Open Refine
6) Spark Fundamentals
7) NoSQL
8) Hadoop Fundamentals



Big Data Fundamentals

Are you interested in understanding 'Big Data' beyond the terms used in headlines? Then select this learning path as an introduction to tools like Apache Hadoop and Apache Spark Frameworks, which enable data to be analyzed on mass, and start the journey towards your headline discovery.

TELL YOUR FRIENDS

AUDIENCE: Big Data Enthusiasts, Data Engineers, Data Scientists
LEARNING PATH LEVEL: Beginner
4 BADGES
3 COURSES

BDU
R 101

Are you ready to dive head-first into R? In just a few hours you'll learn how to write your own R code, learn about R's data structures and create your own functions. You'll even be able to import data and do some operations. Try our hands-on exercises as we guide your first steps into your data science journey with R.

BDU
Machine Learning 101 – Version 1 (Archived)

Are the phrases 'it is certain', 'yes you may rely on it', 'reply hazy try again' common in your predictions? Make room on your shelf for your magic eight ball and take this opportunity to see how Machine Learning can be a beneficial tool for predicting future trends. All with a bit more than oil and blue die. All signs point to yes as an indicator you will benefit from this experience.

Start the Free Course

TELL YOUR FRIENDS

COURSE CODE: ML0101EN
AUDIENCE: Anyone interested in Machine Learning
COURSE LEVEL: Beginner
TIME TO COMPLETE: 4 Hours

BDU
Python 101

This free Python course provides a beginner-friendly introduction to Python. Practice through lab exercises, and you'll be ready to create your first Python scripts on your own!

REFINEPRO
OpenRefine 101

In 2015, the Times best seller list held a title of the *life changing magic of tidying up*, a book about getting rid of messes. In your own house this might work, but in the world of publicly accessible data, it can be messy. Come take this course, and see how OpenRefine can work its magic!

BDU
Spark Fundamentals I

Ignite your interest in Spark with an introduction to the core concepts that make this general processor an essential tool set for working with Big Data.

BDU
NoSQL and DBaaS 101

It's rush hour and you have places to go and people to meet. Will you get there? Will you make it on time? If your operational data shares similarities with urban transportation networks, get ready for this ride!

Hadoop Fundamentals

Are you interested in moving beyond the elephant in the room and understanding Hadoop as a foundational tool set in your future? Then select this learning path to gain exposure to the tools used in Big Data, Hadoop's core components and supporting open source projects.

TELL YOUR FRIENDS

AUDIENCE: Data Engineers, Data Scientists, Big Data enthusiasts with technical background
LEARNING PATH LEVEL: Beginner
5 BADGES
4 COURSES

The following three sections (modules) were selected as the primary steps for mastering the

BDU materials by IBM in Russian with the students of the 4th year:

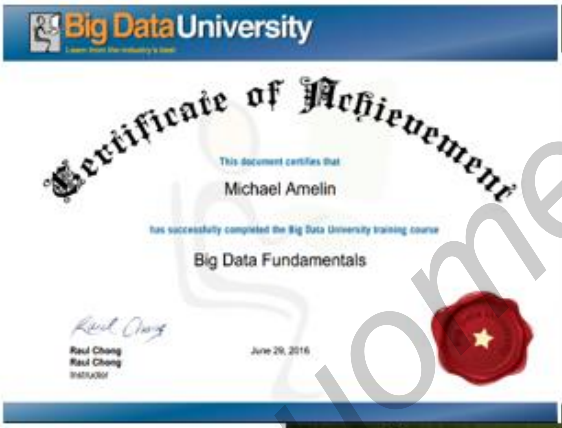
1 The Basics of BIG DATA
2 Introduction to the analysis of data using R technology
3 Algorithms of machine learning

As already mentioned above, in June 2016 three teachers were assigned to train students. All the three teachers completed their master's and postgraduate studies and are fluent in English which is necessary for covering BDU educational materials provided by IBM. The beginning of the classes with students was scheduled for November 1, 2016.

Teachers studied the BDU materials in July-August 2016 and in September-October they developed lectures and practical classes for students in Russian. In preparing the training materials the teachers used not only BDU resources (provided by IBM), but also studied educational resources on relevant topics at other universities around the world.

One of the teachers completed the training, passed exams and received certificates on the following 6 courses in July-August 2016:

1 BIG DATA Fundamentals (BDU от IBM)
2 Introduction to R-DataCamp Course (BDU от IBM)
3 Introduction to Machine Learning (DataCamp)
4 Intermediate R (DataCamp)
5 Intro to Statistics with R: Correlation and Linear Regression
6 Hadoop Fundamentals I (BDU от IBM)



Big Data University
Learn from the industry's best

Certificate of Achievement

This document certifies that

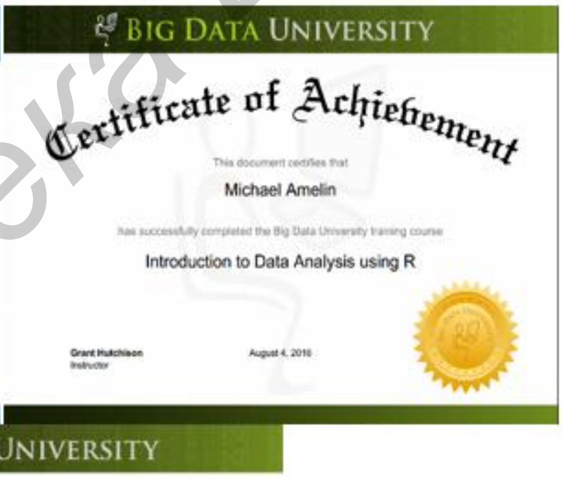
Michael Amelin

has successfully completed the Big Data University training course

Big Data Fundamentals

Raul Chong
Raul Chong
Raul Chong
Instructor

June 29, 2016



BIG DATA UNIVERSITY

Certificate of Achievement

This document certifies that

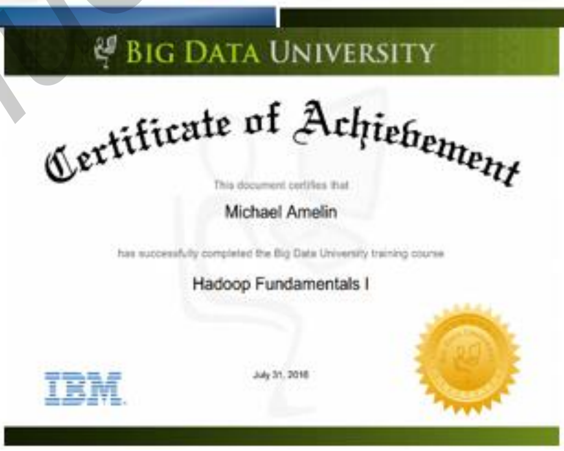
Michael Amelin

has successfully completed the Big Data University training course

Introduction to Data Analysis using R

Grant Hutchison
Instructor

August 4, 2016



BIG DATA UNIVERSITY

Certificate of Achievement

This document certifies that

Michael Amelin

has successfully completed the Big Data University training course

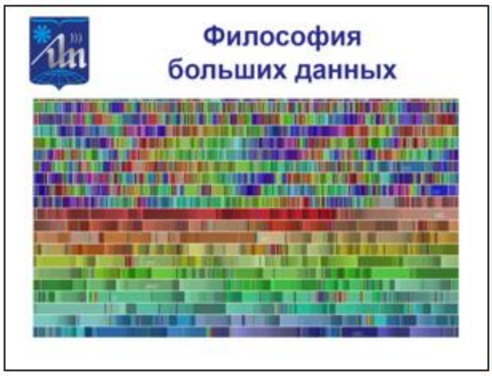
Hadoop Fundamentals I

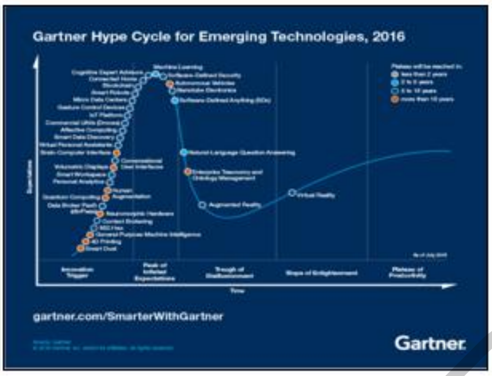
IBM

July 31, 2016

Further on, we will consider the questions the teachers gave to the students in the process of studying the individual sections (modules) of BIG DATA.

Section "BIG DATA Basics". For the theoretical study of the section "Fundamentals of BIG DATA" by the students 18 hours of lectures (9 lectures) were allocated.





The section included the covering of the following 8 questions:

1. Definition and description of big data; its role in the economics and in the activity of enterprises.

2. Stages of big data technologies development.

3. Holistic approach to the big data development.

4. Examples of the big data formation: data and indicators of sensors in production; information from social media sources; results and arrays of scientific research results; information coming up with the expansion of data warehouses.

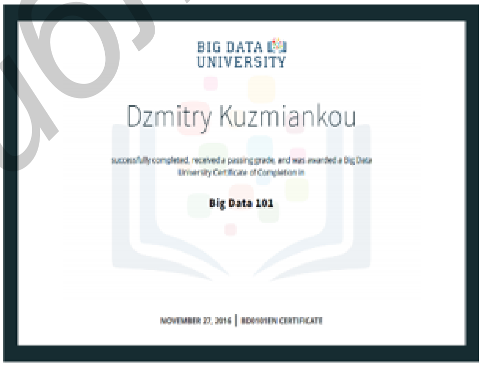
5**.** Explanation of the reasons for forming an integrated big data Platform for a general merger of what, otherwise, would be separate information stores with their own separate and autonomous analytics; in other words, the reason why the integrated Big Data Platform requires the connection of all types of information, individual before, into a single unit and a powerful information space (the creation of an information lake).

6. Identification of the importance of data management that will subsequently ensure big data control.

7. Description and determination of the components necessary for the formation of the Big Data Platform.

8. Comparison and confrontation of the following concepts: inactive data processing (data-at-rest processing); data stream processing (data-in-motion processing); data warehouses processing; contextual search.

Upon mastering the section "Fundamentals of BIG DATA" students passed the exam successfully. After the local BSUIR exam students were able to log in at BDU by IBM and cover the discipline "BIG DATA 101" and to get BDU certificates (examples of certificates are presented below).

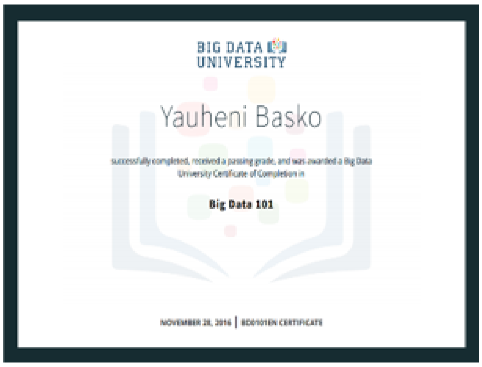

BIG DATA UNIVERSITY

Dzmitry Kuzmiankou

successfully completed, received a passing grade, and was awarded a Big Data University Certificate of Completion in

Big Data 101

NOVEMBER 27, 2016 | BD0101EN CERTIFICATE


BIG DATA UNIVERSITY

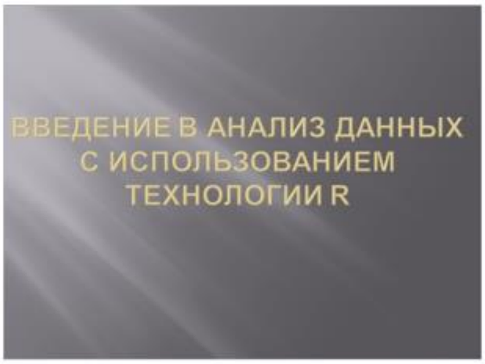
Yauheni Basko

successfully completed, received a passing grade, and was awarded a Big Data University Certificate of Completion in

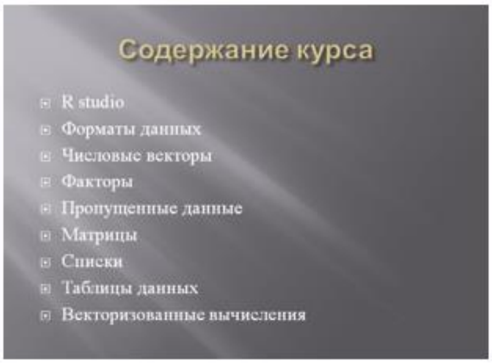
Big Data 101

NOVEMBER 28, 2016 | BD0101EN CERTIFICATE

Section "Introduction to data analysis using R technology". For students' studying this section 24 hours of lectures and 36 hours of practical training were allocated.



ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИИ R



Содержание курса

- R studio
- Форматы данных
- Числовые векторы
- Факторы
- Пропущенные данные
- Матрицы
- Списки
- Таблицы данных
- Векторизованные вычисления

The content of the discipline is as follows:

Topic 1 Introduction to R Technology
§ 1.1 Numerical vectors
§1.2 Factors
§1.3 Arithmetic operations

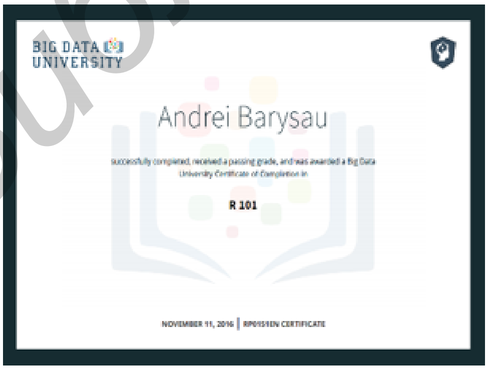
Topic 2 Simple manipulations: numbers and vectors
§2.1 Vector arithmetic
§2.2 Logical vectors
§2.2 Symbol vectors

Topic 3 Objects, their modes and attributes
§3.1 Attributes, Modes
§3.2 Getting and installing

Topic 4 Ordered and unordered factors
§4.1 Obtaining factors from a categorical variable
§4.2 Obtaining factors from a quantitative variable

Topic 5 Arrays and Matrices
§5.1 Product
§5.2 Transportation
§5.3 Matrix Tools
§5.4 Linking Arrays

Upon mastering the section "Introduction to data analysis using R technology" students successfully passed the exam. After the local BSUIR exam students were able to log in at BDU by IBM and cover the discipline "BIG DATA 101" and to get BDU certificates (examples of certificates are presented below).



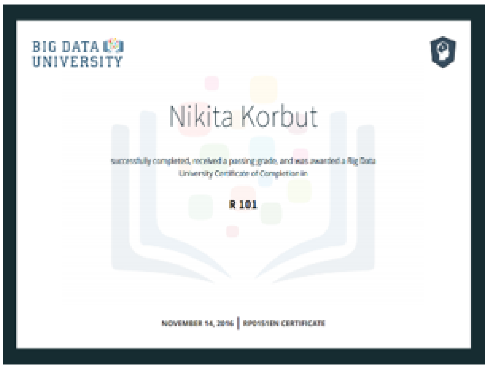
BIG DATA UNIVERSITY

Andrei Barysau

successfully completed, received a passing grade, and was awarded a Big Data University Certificate of Completion in

R 101

NOVEMBER 11, 2016 | RP0151EN CERTIFICATE



BIG DATA UNIVERSITY

Nikita Korbut

successfully completed, received a passing grade, and was awarded a Big Data University Certificate of Completion in

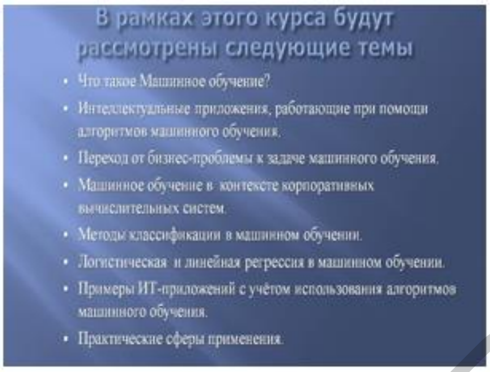
R 101

NOVEMBER 14, 2016 | RP0151EN CERTIFICATE

Section "Algorithms of machine learning". For the students studying this section 24 hours of lectures and 36 hours of laboratory work were allocated.

Алгоритмы машинного обучения

Амелин Михаил Александрович,
магистр экономических наук,
ассистент кафедры инженерной психологии и эргономики



В рамках этого курса будут рассмотрены следующие темы

- Что такое Машинное обучение?
- Интеллектуальные приложения, работающие при помощи алгоритмов машинного обучения.
- Переход от бизнес-проблемы к задаче машинного обучения.
- Машинное обучение в контексте корпоративных вычислительных систем.
- Методы классификации в машинном обучении.
- Логистическая и линейная регрессия в машинном обучении.
- Примеры ИТ-приложений с учётом использования алгоритмов машинного обучения.
- Практические сферы применения.

The content of the discipline is as follows:

Topic 1 Introduction to data analysis and machine learning. Logical classification methods

§1.1 Examples of machine learning application

§1.2 Problems of retraining

§ 1.3 Python for data analysis

§1.4 Working with vectors and matrices in NumPy

§1.5 Decision trees

Topic 2 Metric and linear classification methods

§2.1 Nearest neighbors algorithms

§2.2 Parzen window method

§2.3. Emissions detection

§2.4 The stochastic gradient method

§2.5 The SAG algorithm

Topic 3 Support vector method. Logistic regression

§3.1 The essence of support vectors method

§3.2 Application of support vectors method

§3.3 The essence of logistic regression

§3.4 Application of logistic regression

Topic 4 Linear regression. Dimension reduction and principal components method

§4.1 Singular decomposition

§4.2 Crestal regression

§4.3 The LASSO Method

§4.4 Approach to the characteristics selection

Topic 5 Compositions of algorithms. Neural networks

§5.1 Bagging and Random Forest

§5.2 Gradient boosting

§5.3 Back propagation

Topic 6 Clustering and visualization

§6.1 Lowering the dimension

§6.2 Solving semi-supervised learning tasks

Topic 7 Machine Learning in Applied Problems

§7.1 Working with numerical characteristics

§7.2 Categorical and textual features

§7.3 Data preprocessing

Prospects. In future, in spring semester (April and May 2017), two new groups of students will be trained in BIG DATA basics under the IBM Big Data University (BDU) program. The first group will consist of 32 3rd year students specializing in Engineering and psychological support of information technology (system engineers); the second group is represented by 26 students of the 3rd year specializing in Information systems and technologies (in industrial safety) (system engineers). Both groups study 4 years full-time.

Together with the development of the planned IT discipline "Virtual Reality Technologies" the students of the first group will be offered the materials of the same three sections (modules) of BIG DATA developed last semester. These are: 1) The fundamentals of BIG DATA; 2) Introduction to data analysis using R technology; 3) Algorithms of machine learning. 24 hours of lectures and 32 hours of practical training are allocated to master all these three sections (modules). This is supposed to be a good experience for teaching in consolidating and improving the material, although with other students.

The students of the second group will be offered the same three sections (module) for parallel learning, but with two other IT disciplines already: 1) System software and 2) Modern programming languages. 24 hours of lectures and 18 hours of practical classes will be allocated for covering the section (module) "Fundamentals of BIG DATA". The two other modules will require the total 12 hours of lectures and 12 hours of laboratory work.

The number of sections (modules) is expected to be increased from 3 to 5 in the 2017-2018 academic year. The teachers will prepare two more new modules: 1) Python; 2) Introduction to Open Refine.

The following key technologies are required to prepare a high-level specialist for BIG DATA: Kafka, Spark, Storm, R, Cassandra, NewSQL, Columnar Database, In Memory, NoHadoop, NoSQL, OLTP, OLAP, ERP, Map Reduce, Scala, etc. Teachers pay special attention to the preparation of practical classes following the recommendations of the IBM BDU educational program and the educational programs for BIG DATA developed by other universities.

When we started Big Data program, it was a new field. Today, 3 years later, there is nearly no University teaching Computer Science, Business Administration, Finance or Economics with the students unaware of applying Big Data analytics to solving practical tasks.

Most of the requests for new projects financing by new startups include plans for Big Data analytics application. BSUIR invested a lot of efforts and energy into the introduction of Big Data Analytics education. Its success will depend on implementing this technology into the teaching courses at BSUIR and applying this technology by all PhD students in their theses.

We are planning to establish partnership with several vendors in open source projects using Machine Learning Algorithms. This will provide the students with the access to the state-of-the-art technology and prepare them to the challenging opportunities in the future.