Syllabus on the course "Knowledge Discovery in Data at Scale Technologies"

Approved by Programme Academic Council

Process-verbal 2 from April 10, 2018

Author	Peter Golubtsov
Credits	3
Academic Hours	114
Year of study	1
Mode of study	Full-time

1. **Pre-requisites**

The course is based on the knowledge of foundations of principles and skills:

- Basic computer science principles and skills,
- Scientific programming (e.g., in MatLab or Python)
- Linear algebra,
- Probability and statistics,
- Basics in data analysis.

2. **Course type:** Elective

3. **Abstract**

The main goal of this course is to provide students with an opportunity to acquire conceptual background and mathematical tools applicable to Big Data analytics and real time computation. The course will briefly review specific challenges of Big Data Analytics, such as problems of extracting, unifying, updating, and merging information and specific needs in processing data, which should be highly parallel and distributed. With these specific features in mind we will then study more closely a number of mathematical tools for Big Data analytics, such as regression analysis, linear estimation, calibration problems, real time processing of incoming (potentially infinite) data. We will see how these approaches can be transformed to conform to the Big Data demands.

4. **Learning Objectives**

- Formation of the theoretical knowledge and practical basic skills in the collection, storage, processing and analysis of large data.
- Develop theoretical and practical skills to analyze large data to tackle a wide range of applications.

5. Learning Outcomes

Knowledges Skills	Upon successful completion of this course, students will be able to: • understand main principles of approaching big data problems for large-scale distributed systems; • design an efficient representation of intermediate information for various data processing problems; • Redesign and apply linear regression methods in the context of distributed and emerging data; • Apply optimal linear estimation methods in big data context. • Design and use calibration techniques in the cases where the measurement process is unknown.
Practice	To be able to apply the essential tools and techniques of distributed data processing in practice.

6. Course Plan

1. Introduction to information processing in distributed systems. Simple examples and properties of various forms of information representation.
2. Distributed information management for linear regression problems
3. Linear experiment, optimal estimation problem. Combining linear experiments in “raw” form. Canonical information for linear experiments.
4. Optimal estimation with prior information. Updating prior information. Manipulating information in various forms.
5. Optimal estimation with uncertainty in measurement transformation.
6. Unknown measurement transformation. Calibration problem. Canonical calibration information.

7. Reading List

Most of the material will be given in class. Students will be provided with lecture notes.

Required

1. P.V. Golubtsov, The Concept of Information in Big Data Processing. Automatic Documentation and Mathematical Linguistics, 2018, Vol. 52, No. 1, pp. 38–43. http://dx.doi.org/10.3103/S000510551801003X
2. P.V. Golubtsov, The Linear Estimation Problem and Information in Big-Data Systems. Automatic Documentation and Mathematical Linguistics, 2018, Vol. 52, No. 2, pp. 73–

79. http://dx.doi.org/10.3103/S0005105518020024

3. P.V. Golubtsov, The Transition from A Priori to A Posteriori Information: Bayesian Procedures in Distributed Large-Scale Data Processing Systems. Automatic Documentation and Mathematical Linguistics, 2018, Vol. 52, No. 4, pp. 203–213. http://dx.doi.org/10.3103/S0005105518040064

Optional

1. Albert, A., Regression and the Moore–Penrose Pseudoinverse, New York: Academic Press, 1972.
2. Bekkerman, R., Bilenko, M., and Langford, J., Scaling up Machine Learning: Parallel and Distributed Approaches, New York: Cambridge University Press, 2012.

8. Grading System

The student should demonstrate the knowledge of sections of the discipline and the ability to present the results of homework and tests in accordance with the required competencies.

Evaluation of all forms of monitoring are set on a 10-point scale.

Grade	**10-point scale**
Excellent	10
	9
	8
Good	7
	6
Satisfactorily	5
	4
Bad	3
	2
	1
	0

9. Guidelines for Knowledge Assessment

The final course grade is a sum of the following elements:

1. Attendance record (A);
2. Home assignments (H);
3. Exam (E).

Home assignments:

The course includes six home assignments in the form of research mini projects: three theoretical and three practical (programming projects).

The overall and accumulated course grades Go and Ga (10-point scale each) are calculated as follows:

$$Ga = 0.2\ A + 0.8\ H,$$

$$Go = 0.6\ Ga + 0.4\ E.$$

The overall and accumulated course grades Go and Ga (10-point scale each) include results achieved by students in their attendance record A, home assignments H and exam E; it is rounded up to an integer number of points.

10. Methods of Instruction

Lectures, seminars, practice work (programming projects).