

DOKTORANTŪROS STUDIJŲ DALYKO SANDAS

Dalyko pavadinimas	Mokslo kryptis, kodas	Fakultetas	Katedra
Didžiųjų duomenų analitika	Informatikos inžinerija (T 007)	MIF	Duomenų mokslo ir skaitmeninių technologijų institutas
Studijų būdas	Kreditų skaičius ECTS	Studijų būdas	Kreditų skaičius
paskaitos	1 (rudens sem.)	konsultacijos	1
individualus	4	seminarai	1

Dalyko anotacija

Dalyko tikslas: Siekiama, kad doktorantas įgytų esminių įgūdžių apie didžiųjų duomenų analitiką bei ugdytų praktinius gebėjimus taikyti didžiųjų duomenų saugojimo ir apdorojimo būdus ir technologijas bei tyrybos metodus, taip pat gebėtų kritiškai vertinti modernias keičiamo dydžio (angl. *scalable*) mašininio mokymosi paradigmas ir paskirstytųjų sistemų architektūras.

Įvadinė dalis. Didžiųjų duomenų apibrėžtys bei jų vystymosi istorija. Pagrindiniai didžiųjų duomenų šaltiniai. Didžiųjų duomenų tipai: struktūriniai ir nestrukūriniai duomenys. Didžiųjų duomenų saugojimo būdai ir technologijos. *Data Lakehouse*, *Data Fabric* ir *Data Mesh* architektūros. Didžiųjų duomenų tyryba: taikymo uždavinių pavyzdžiai.

Didžiųjų duomenų analitika:

Duomenų tyryba žinių gavimo procese, žinių gavimo proceso etapai.

Didelės apimties duomenų pirminio apdorojimo būdai: profiliavimas, valymas, požymių inžinerija (angl. *feature engineering*) ir imčių sudarymas.

Didžiųjų duomenų analitikos uždaviniai.

Didžiųjų duomenų tyryba ir keičiamo dydžio mašininis mokymasis:

- Dažnų sekų paieška ir asociacijų taisyklės.
- Prognozavimo, klasifikavimo ir anomalijų aptikimo metodai, paremti didžiųjų duomenų analize; modelių vertinimas esant koncepcijų poslinkiui (angl. *concept drift*).
- Srautinių duomenų analizė.
- Klasterizavimo metodai didelių apimčių duomenims.
- Dimensijos mažinimo būdai ir reprezentacijų mokymasis (įskaitant įterpinius, angl. *embeddings*).
- Mašininis mokymasis ir gilusis mokymasis didžiųjų duomenų apdorojimui (CNN, RNN, *Transformers*), perkeliama mokymasis (angl. *transfer learning*), paskirstytasis mokymas.
- Didieji kalbos modeliai (LLM) ir generatyvinis dirbtinis intelektas. Paieška papildytas generavimas (angl. *Retrieval-Augmented Generation*, RAG) ir vektorinės duomenų bazės.
- Teksto tyryba ir natūraliosios kalbos apdorojimas (NLP): TF-IDF, transformeriais grįsti modeliai, temų modeliavimas, sentimentų analizė.
- Mašininio mokymosi modelių operacijų valdymas (MLOps).
- Didžiųjų duomenų vizualizavimo iššūkiai.

Didžiųjų duomenų technologijos: *MapReduce* paradigma, *Apache Hadoop* ir *Spark* platformos, *Apache Kafka*, *Spark MLlib* biblioteka mašiniam mokymuisi, Didžiųjų duomenų analitika taikant R paketą, *PySpark* ir *Python*.

Debesų kompiuterijos sprendimai, taikomi didžiųjų duomenų analitikoje. *Microsoft Azure Machine Learning* platforma, *Amazon Web Services (AWS)* sprendimai, *Google Cloud (BigQuery)*,

Dataprocc). Databricks (Lakehouse, Delta Lake), konteinerizacijos ir orkestravimo pagrindai (Docker / Kubernetes)

Didžiųjų duomenų tyrybos praktiniai aspektai. Praktinių taikymų pavyzdžiai sprendžiant uždavinius, kylančius versle, finansuose, medicinoje, kibernetiniame saugume ir kt.

Praktinė užduotis: išspręsti nurodytą duomenų analizės uždavinį, taikant didžiųjų duomenų bei mašininio ar giliojo mokymosi technologijas, įskaitant duomenų surinkimą, pirminį apdorojimą ir vizualizavimą.

Pagrindinė literatūra

Kleppmann, M. (2017). Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems. O'Reilly Media, Inc.

Chambers, B., & Zaharia, M. (2018). Spark: The definitive guide: Big data processing made simple. "O'Reilly Media, Inc."

Favaretto, M., De Clercq, E., Schneble, C. O., & Elger, B. S. (2020). What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PloS one*, 15(2), e0228987.

Reddy, G.T., Reddy, M.P.K., Lakshmana, K., Kaluri, R., Rajput, D.S., Srivastava, G. and Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *IEEE Access*, 8, pp.54776-54788.

Bhatia, P. (2019). Data Mining and Data Warehousing Principles and Practical Techniques. Cambridge University Press. <https://doi.org/10.1017/9781108635592>

Jurafsky, D. and Martin, J.H. 2026. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released January 6, 2026. <https://web.stanford.edu/~jurafsky/slp3>.

Pietsch, W. (2021), Big Data, Cambridge University Press.
<https://doi.org/10.1017/9781108588676>

Mining Massive Data Sets, SOE-YCS0007, Stanford School of Engineering
<https://online.stanford.edu/courses/soe-ycs0007-mining-massive-data-sets>

Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. Mining of massive data sets. Cambridge university press, 2020.
<https://doi.org/10.1017/9781108684163>, <http://www.mmids.org/>

Spark and Python for Big Data with PySpark, Udemy
<https://www.udemy.com/course/spark-and-python-for-big-data-with-pyspark/>

Konsultuojančiųjų dėstytojų vardas, pavardė	Mokslo laipsnis	Svarbiausieji darbai mokslo kryptyje (šakoje) paskelbti per pastaruosius 5 metus
Viktor Medvedev	dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Viktor+Medvedev
Jolita Bernatavičienė	dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Jolita+Bernatavičienė
Olga Kurasova	dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Olga+Kurasova