Data Science

(Emil Pelikán)

Data Science (datovou vědu) je možné definovat jako interdisciplinární směr informatického výzkumu, který využívá vědeckých metod, procesů, algoritmů a systémů k extrakci znalostí z dat, která mohou mít strukturovanou nebo i nestrukturovanou podobu. Data Science je koncept sjednocující statistickou metodologii, statistické učení, kvantitativní výzkum, analýzu dat, prediktivní modelování, klasifikační metody, data mining, strojové učení, business analytiku, business inteligenci, apod.

V souvislosti s datovou vědou je často důraz kladen na schopnost analyzovat velká data (Big Data) a potřebu analyzovat data komplexní nejen svým rozsahem co do počtu záznamů a položek, resp. dimenzí v časové doméně (časové řady, videozáznamy, apod.). Komplexita dat může být dále navýšena propojením s dalšími datovými zdroji a databázemi.

Data Science je tedy třeba chápat jako interdisciplinární vědeckou disciplínu, jejímž základem je matematická věda o analýze dat, tj. matematická statistika nebo fuzzy modelování. Datová věda využívá také dalších matematických a informatických disciplín, např. teorii informace, teorii pravděpodobnosti a programování nebo fuzzy logiku. Současně vyžaduje hlubší porozumění oboru, na který se výzkum konkrétně zaměřuje.

Zapojení pracovišť v ČR a téma výzkumu

Většina pracovišť věnujících se oblasti umělé inteligence je zapojena do výzkumu v oblasti Data Science. Representativní příklady jsou:

Pracoviště (abecedně)	**Téma**
AI@VSE	Modelování propojených dat (linked data), jejich shromažďování, publikování a využívání zejména v oblastech veřejné správy, elektronického obchodu, akademické sféry, biomedicíny a encyklopedických znalostí (Wikipedia apod.), analýza, evaluaci a mapování ontologií a datových slovníků, vizualizace ontologií a propojených dat, výzkum v oblasti algoritmů pro data mining a v oblasti srozumitelnosti výsledků data miningu
CIIRC ČVUT	Rozvrhovací algoritmy řízené daty, predikce stavu vozovky, pokročilé metody analýzy biologických signálů pro diagnostiku a terapii se zaměřením na: dlouhodobé mnoha kanálové záznamy; integraci dat z různých modalit; netradiční matematické nástroje (např. Riemannova geometrie); hybridní metody strojového učení; aktivní učení, strukturované reprezentace dat a znalostí v medicíně; standardizace datových struktur

FEL ČVUT	Získávání znalostí z dat (data mining), strojové učení, statistická analýza dat
FIT VUT	Statistické metody strojového učení, modelování a simulace velmi rozsáhlých systémů, zpracování, analýza a extrakce dat z rozsáhlých mediálních dat (audio/řeč, video, databáze obrazů), zpracování velmi rozsáhlých textů přirozeného jazyka (NLP), dolování dat z rozsáhlých textových a mediálních a jiných záznamů
MFF UK	Strojové učení, hluboké učení, reinforcement learning, zpracování přirozeného jazyka, systémy komunikace člověk-stroj
OSU-UVA FM	Teoretický rozvoj metod v oblasti umělé inteligence, fuzzy modelování, numerická a funkcionální analýza, optimalizační úlohy, fuzzy logika, počítačové vidění na základě fuzzy modelování, zpracování velkých dat, jejich redukce, reprezentace, rekonstrukce, získávání znalostí z dat a fuzzy prognostika vývoje systémů, extrakce expertních znalostí, prognostika dynamických systémů
TUL	Automatické zpracování akustických a zejména řečových dat
ÚI AV ČR	Statistické modelování, časově-prostorové analýzy, dynamické modely, latentní komponenty, robustnost, statistické aspekty strojového učení (statistické učení), Bayesovské predikční modely, analýza vysoce-dimenzionálních dat, redukce dimenzionality, komplexní systémy
UPOL AI	Analýza binárních, ordinálních a relačních dat zatížených neurčitostí a nepřesností; faktorové modely; rozsáhlá relační data; dimenzionalita relačních dat; vývoj rychlých algoritmů
ÚTIA AV ČR	Strojové učení, klasifikátory, výběr příznaků a redukce dimenzionality, optimální příznakové podprostory, rozhodování za neurčitosti Bayesovské dynamické rozhodování, decentralizované rozhodování
VŠB TUO	Predikce množství a kvality elektřiny, modelování energetického mixu pro Off-Grid systémy, predikční metody pro elektromobilitu, nabíjecí stanice, analýza velkých dat, škálovatelné metody pro farmakologii, optimalizace dopravních proudů ve městech, chytrá města, autonomní vozidla
ZČU	Sběr, anotace, standardizace, uchovávání, zpracování (machine a deep learning, statistické metody) a interpretace heterogenních biomedicínských (zejména elektrofyziologických) dat, brain-computer interface, open data a open science, infrastruktura pro neuroscience

Vybrané výsledky

AI@VSE

- EntityClassifier - nástroj pro vyhledávání entit v textu s využitím Linked Hypernym Dataset (http://entityclassifier.eu/)
- Public Contracts Ontology, v původní podobě nebo s úpravami se použila pro reálné publikování dat o veřejných zakázkách v Itálii a Španělsku

CIIRC ČVUT

- Kucewicz, M.T.; Doležal, J.; Křemen, V.; Berry, B.M.; Miller, L.R.; Magee, A.L.; Fabián, V.; Worrell, G.A. Pupil size reflects successful encoding and recall of memory in humans
  Scientific Reports. 2018, 8 ISSN 2045-2322.
  https://www.nature.com/articles/s41598-018-23197-6
- V. Křemen, et al. Behavioral state classification in epileptic brain using intracranial electrophysiology. Journal of Neural Engineering. 2017, 14(2), ISSN 1741-2560.
- V. Gerla, et al. Hybrid Hierarchical Clustering Algorithm Used for Large Datasets: A Pilot Study on Long-Term Sleep Data. In: Precision Medicine Powered by pHealth and Connected Health. International Conference on Biomedical and Health Informatics 2017. Springer Nature Singapore Pte Ltd.. 2017, pp. 3-7. 1. vol. 66.
- E. Saifutdinova, V. Gerla, and L. Lhotska. Riemannian Geometry in Sleep Stage Classification. In: M. Bursa et al. (Eds.): ITBAM 2017, LNCS 10443, pp. 92–99, 2017.

FEL ČVUT:

- Ryšavý P., Železný F.: Estimating sequence similarity from read sets for clustering next-generation sequencing data. Data Mining and Knowledge Discovery, 33(1):1-23, 2019
- Šourek G., Aschenbrenner V., Železný F., Schockaert S., Kuzelka. O: Lifted Relational Neural Networks: Efficient Learning of Latent Relational Structures. Journal of Artificial Intelligence Research 62:69-100, 2018
- Hubáček O., Šourek G., Železný F.: Learning to predict soccer results from relational data with gradient boosted trees. Machine Learning 2018

FIT VUT

- ZEINALI Hossein, SAMETI Hossein, BURGET Lukáš a ČERNOCKÝ Jan. Text-dependent speaker verification based on i-vectors, Neural Networks and Hidden Markov Models. Computer Speech and Language. Amsterdam: Elsevier Science, 2017, roč. 2017, č. 46, s. 53-71. ISSN 0885-2308.
- SOCHOR Jakub, JURÁNEK Roman, ŠPAŇHEL Jakub, MARŠÍK Lukáš, ŠIROKÝ Adam, HEROUT Adam a ZEMČÍK Pavel. Comprehensive Data Set for Automatic Single Camera Visual Speed Measurement. IEEE Transactions on Intelligent Transportation Systems. 2018, roč. 2018, č. 99, s. 1-11. ISSN 1524-9050.
- DYTRYCH Jaroslav a SMRŽ Pavel. Advanced User Interfaces for Semantic Annotation of Complex Relations in Text. Lecture Notes in Computer Science. 2018, roč. 2017, č. 10839, s. 205-221. ISBN 978-3-319-93581-2. ISSN 0302-9743.

MFF UK

- Václavík, R. - Novák, A. - Šůcha, P. - Hanzálek, Z.: Accelerating the Branch-and-Price Algorithm Using Machine Learning, European Journal of Operational Research, Volume 271, Issue 3, December 2018, Pages 1055-1069, doi: 10.1016/j.ejor.2018.05.046, Elsevier.

- Václavík, R. - Šůcha, P. - Hanzálek, Z.: Roster evaluation based on classifiers for the nurse rostering problem, Journal of Heuristics, October 2016, Volume 22, Issue 5, Pages 667–697, doi: 10.1007/s10732-016-9314-9, Springer.

OSU-UVA FM

- Novák, V., Perfilieva, I., Dvořák, A.: Insight into Fuzzy Modeling. Wiley & Sons, Hoboken 2016.

TUL

- Z. Koldovský and P. Tichavský, "Gradient Algorithms for Complex Non-Gaussian Independent Component/Vector Extraction, Question of Convergence, " IEEE Transactions on Signal Processing, vol. 67, no. 4, pp. 1050-1064, Feb. 2019.
- Z. Koldovský and F. Nesta, "Performance Analysis of Source Image Estimators in Blind Source Separation," IEEE Transactions on Signal Processing, Vol. 65, No. 16, pp. 4166-4176, ISSN:1053-587X, Aug 2017.
- P. Cerva, J. Silovsky, J. Zdansky, J. Nouza, L. Seps, "Speaker-adaptive speech recognition using speaker diarization for improved transcription of large spoken archives", Speech Communication, vol. 55, no. 10, pp. 1033-1046, 2013

ÚI AV ČR

- Runge, J., Petoukhov, V., Donges, J.F., Hlinka, Jaroslav, Jajcay, Nikola, Vejmelka, Martin, Hartman, David, Marwan, N., Paluš, Milan, Kurths, J. Identifying Causal Gateways and Mediators in Complex Spatio-Temporal Systems. Nature Communications. 2015, 6(7 October), Article 8502. ISSN 2041-1723, doi: 10.1038/ncomms9502.
- Brabec, Marek, Procházka, Pavel, Maturkanič, Dušan. Semiparametric Statistical Analysis of the Blade Tip Timing Data for Detection of Turbine Rotor Speed Instabilities. Quality and Reliability Engineering International. 2018, 34(7), 1308-1314. ISSN 0748-8017, doi: 10.1002/qre.2327
- Valenta, Zdeněk, Kalina, Jan. Exploiting Stein's Paradox in Analysing Sparse Data from Genome-Wide Association Studies. Biocybernetics and Biomedical Engineering. 2015, 35(1), 64-67. ISSN 0208-5216, doi: 10.1016/j.bbe.2014.10.004

UPOL

- R Belohlávek, V Vychodil: Discovery of optimal factors in binary data via a novel method of matrix decomposition. J. Comput. Syst. Sci. 76(1): 3-20 (2010)
- R Belohlavek, M Trnecka: From-below approximations in Boolean matrix factorization: Geometry and new algorithm. J. Comput. Syst. Sci. 81(8): 1678-1697 (2015)
- R Belohlavek, M Trnecka: Handling Noise in Boolean Matrix Factorization. IJCAI 2017: 1433-1439

ÚTIA AV ČR

- Quinn A., Kárný Miroslav, Guy Tatiana Valentine: Optimal design of priors constrained by external predictors, International Journal of Approximate Reasoning vol. 84, 1 (2017), p. 150-158 [2017]
- Kárný Miroslav, Herzallah R.: Scalable Harmonization of Complex Networks With Local Adaptive Controllers, IEEE Transactions on Systems Man Cybernetics-Systems vol. 47, 3 (2017), p. 394-404 [2017]
- Kárný Miroslav : Approximate Bayesian recursive estimation, Information Sciences vol. 285, 1 (2014), p. 100-111 [2014]

VŠB TUO

- T Vantuch, S Misak, T Jezovicz, T Burianek, V Snasel (Q1). The Power Quality Forecasting Model for Off-Grid System Supported by Multi-objective Optimization. IEEE Transactions on Industrial Electronics, 2017, IF 7.168
- V OJHA, V Snasel, A Abraham (Q1), Multiobjective Programming for Type-2 Hierarchical Fuzzy Inference Trees, IEEE Transactions on Fuzzy Systems, 2017, IF 7.671
- VK Ojha, A Abraham, V Snasel (Q1), Metaheuristic design of feedforward neural networks: A review of two decades of research, Engineering Applications of Artificial Intelligence 60, 97-116, 2017, IF 2.894
- V Snášel, J Nowaková, F Xhafa, L Barolli (Q1), Geometrical and topological approaches to Big Data, Future Generation Computer Systems 67, 286-296, 2017, IF 3.997

ZČU

- C. Cerisara, P. Kral, L. Lenc, On the Effects of using Word2Vec Representations in Neural Networks for Dialogue Act Recognition, in Computer Speech & Language, IF: 1.90, January 2018, vol. 47, pp. 175-193, Elsevier, ISSN: 0885-2308, doi: 10.1016/j.csl.2017.07.009.
- Mouček, R., Vařeka, L., Prokop, T., Štěbeták, J., Brůha, P. Event-related potential data from a guess the number brain-computer interface experiment on school children. Scientific Data, 2017, Volume 4, Article Number: UNSP 160121, March 2017, DOI: 10.1038/sdata.2016.121
- Vařeka, L. and Mautner P. Stacked Autoencoders for the P300 Component Detection. Front. Neurosci. 11:302., 2017, doi: 10.3389/fnins.2017.00302

Vybrané aplikace

CIIRC ČVUT

- Inteligentní monitorování samostatně žijících osob s využitím dat ze senzorů v daném prostředí, detekce nežádoucích stavů (funkční vzorek, spolupráce s firmou MediWare)
- Sledování očních pohybů (spolupráce CIIRC s firmou Medicton Group); https://i4tracking.cz/oblasti-vyzkumu/
- Doležal, J. & Fabian, V. 41. Application of eye tracking in neuroscience. Clin. Neurophysiol. 126, e44 (2015)

- Dobiáš, M.; Doležal, J.; Klesalová, A.; Chytrý, V.; Erlebach, J.; Fabián, V.; Urban, M. Metodika sledování očních pohybů pro testování kompetencí v personalistice. [Applied Certified Methodology] 2017

FEL ČVUT

- CyberCalc: systém využívající techniky umělé inteligence pro kombinatorický návrh velkých soustav mechanických klíčů a zámků. V roce 2016 získal cenu Wernera von Siemense za nejlepší výsledek vývoje v ČR. Autoři: Filip Železný, Radomír Černoch, Vladislav Král, Vyacheslav Kungurtsev

FIT VUT

- Systém pro sémantické obohacování plných textů o vazbu na jmenné autority, software, 2017. Autoři: Otrusina Lubomír, Smrž Pavel

MFF UK

- Road condition prediction – Contract with Porsche Engineering Services, 2017-2018

OSU-UVA FM

- Automatický systém pro čtení registračních značek projíždějících automobilů z digitálních obrazů zaznamenaných v reálném provozu. Realizace na základě smlouvy s fy CGI IT Czech Republic, s.r.o. Systém je v provozu od r. 2017.

TUL

- Program pro diktování do počítače vyvinutý ve spolupráci s firmou Newton Technologies a komerčně nasazený v 5 evropských zemích
- Komplexní technologická platforma pro automatický přepis, monitoring a analýzu televizních a rozhlasových stanic v současné době pracující s 20 evropskými jazyky a komerčně nasazená firmou Newton Media
- Systém pro automatické vyhledávání ve zvukovém archivu Českého rozhlasu (pracuje se souborem cca 250.000 záznamů, které byly automaticky přepsány a zaindexovány)

ÚI AV ČR

- "Malware Classification of Executable Files by Convolutional Networks" (patent application filed with the United States Patent and Trademark Office under No. 62/583,366, joint patent application of AVAST and UI AV ČR)

UPOL

- *Regulátor na bázi fuzzy logiky pro automatické dávkování relaxantů.*

Regulátor byl vyvinut s podporou IGA Ministerstva zdravotnictví ČR ve spolupráci s Fakultní nemocnicí Olomouc (FNOL); Bělohlávek (UP) + prof. MUDr. Milan Adamus, CSc., MBA (FNOL). Po roce 2005 byl rutinně používán při dlouhotrvajících operacích mozku ve FNOL.

VŠB TUO

- Platforma HyperLoom sloužící ke spouštění velkého množství malých úloh na velkých HPC systémech
- Predikce a simulace pro velká data, metoda certifikovaná ministerstvem dopravy. Výsledky lze využít například v oblasti energetiky. http://modata.vsb.cz

ZČU

- „Hardware and software infrastructure for research in electrophysiology“ – soubor nástrojů pro sběr, uchovávání, popis a analýzu elektrofyziologických dat“

Významné projekty v posledních 5 letech

AI@VSE

- GA CR 18-23964S „Fokusovaná kategorizační síla webových ontologií“, 2018-2020
- EU H2020 OpenBudgets.eu, „Financial Transparency Platform for the Public Sector“, 2015-2017
- EU FP7 LinkedTV „Television Linked To The Web“, 2011-2015

CIIRC ČVUT

- 2015 – 2019 AZV No. 15-31398A Charakteristiky elektromechanické dyssynchronie predikující efekt srdeční resynchronizační terapie
- 2015 – 2018 AZV No. 15-25710A Identifikace individuální dynamiky glykemických exkurzí u pacientů s diabetem pro zlepšení rozhodovacích postupů ovlivňujících dávkování inzulínu
- 2017 – 2020 Osobní zdravotní a asistenční systémy (MPO Trio)
- 2017 – 2019 GAČR Temporal context in analysis of long-term non-stationary multidimensional signal
- 2016 – 2018 GAČR Processing of complex sounds in the central auditory system under normal and pathological conditions

FIT VUT

- Intelligent Management Platform for Advanced Real-Time media processes, EU-7FP-ICT - Sedmý rámcový program Evropského společenství pro atomovou energii (Euratom) v oblasti jaderného výzkumu a vzdělávání, 7E13044, 316564, 2012-2015
- MegaModelling at Runtime - scalable model-based framework for continuous development and runtime validation of complex systems., ECSEL JU - Horizon 2020, 737494, 2017-2020

MFF UK

- GACR P103-16-23509S. FOREST - Flexible Scheduling and Optimization Algorithms for Distributed Real-time Embedded Systems, 2016-2018

OSU-UVA FM

- Výzkumný záměr MSM 6198898701 „Logické a algebraické metody pro zpracování informací zatížených neurčitostí a jejich použití ve fuzzy modelování“ (2005-2010)
- Partner projektu „CZ.1.05/1.1.00/02.0070 Centrum excelence IT4Innovations“ Operační programy EU (2011-2015)
- Partner výzkumného centra 1M6798555601 „Data - Algoritmy – Rozhodování“ (2005-2009)
- Výzkumný záměr MSM 179000002 „Modelování složitých systémů ve fuzzy a v nejistém prostředí“ (2000-2004)
- GRANT N62909-12-1-7039 of the Department of the NAVY, USA, "F-transform - A New Promising technique for image and Signal Processing: How to make its Applications more reliable" (2012-2013)

TUL

- TAČR (TA04010199)- MULTILINMEDIA - Multilingual platform for automatic monitoring and analysis of media, (2015-2017)
- TAČR (TH03010018) – DeepSpot - Multilingual technology for spotting and instant alerting, 2018-2021
- N62909-18-1-2040, Office of Naval Research Global, NICOP – Adaptive Algorithms for Independent Component/Vector Extraction (2018)
- GAČR 17-00902S, The Czech Science Foundation, Advanded Joint Blind Source Separation Methods (2017-2019)
- DA-15-114599, California Community Foundation, Noise reduction of far field speech recordings using two or more microphones (2014-2017)

ÚI AV ČR

- GBP202/12/G061 Centrum excelence - Institut teoretické informatiky (CE-ITI)
- GA18-18080S Objevování znalostí v datech o aktivitě člověka založené na fúzi
- GAP202/11/1368 Učení funkcionálních vztahů z vysoce dimenzionálních dat
- GA13-17187S Konstrukce pokročilých srozumitelných klasifikátorů
- LG12020 Využití pokročilé statistické analýzy a nestatistických separačních metod pro detekování fyzikálních procesů v datech snímaných urychlovači elementárních částic

UPOL

- GAČR 2015–2017 GA15–17899S, „Rozklady matic s booleovskými a ordinálními daty: teorie a algoritmy“
- GAČR 2014–2016 GA14–11585S, „Relační podobnostní databáze“

ÚTIA AV ČR

- Hierarchical models for detection and description of anomalies, Doc. Ing. Václav Šmídl, Ph.D, Duration: 2018 - 2020 Grantor: GACR
- Rationality and Deliberation, Ing. Miroslav Kárný, DrSc, Duration: 2016 - 2018 Grantor: GACR

VŠB TUO

- Systémy založené na rozšířené realitě. Projekt PACMAN H2020 – jsme spoluřešitelem
- H2020 projekt LEXIS – jsme hlavním řešitelem
- IT4Innovations národní superpočítačové centrum je od roku 2016 vedeno společností Intel jako jedno z tzv. Intel Parallel Computing Center
- H2020 projekt ANTAREX, partneři
- H2020 projekt ExCAPE, partneři

ZČU

- INTERREG V-A Počítačový asistenční systém řízený mozkovými vlnami pro osoby s omezenou hybností

Vize rozvoje a příspěvku k ekonomice

Datová věda má potenciál využití v různých oborech vědeckého výzkumu a společenského života v kontextu porozumění a řešení komplexních problémů, zejména v interdisciplinárních oblastech. Metodologie datové vědy pronikají do vývoje metod umělé inteligence a stávají se inteligentními nástroji pro řízení procesů, rozhodování, optimalizaci technologických postupů, extrakci užitečné informace, pro zlepšení a vývoj nových diagnostických a léčebných metod v medicíně, monitorování, modelování, predikci spotřeby energie a strategických průmyslových surovin, zajištění kybernetické bezpečnosti, odhalování falešných zpráv (‘fake news’) i pro analýzu a stanovení vládních strategií.