

MUNI
ICS

PV177 – DataScience seminář (Úvod do předmětu)

Tomáš Rebok & Martin Macák
Ústav výpočetní techniky MU
Fakulta informatiky MU

Představení vyučujících

RNDr. Tomáš Rebok, Ph.D. (ÚVT MU)

- senior výzkumný pracovník, vedoucí projektů a různých skupin
- dlouhodobá činnost v oblasti náročných výpočtů a zpracování dat
- v posledních letech orientace na oblast datové analytiky



Mgr. Martin Macák (FI MU, ÚVT MU)

- PhD. student
- diplomové i postgraduální téma v oblasti datové analytiky
- člen projektů orientujících se na oblast datové analytiky



Cíle kurzu

Seznámení s metodami/nástroji pro analýzy velkých objemů dat

- tzv. *Big Data*
- oblast obrovská, proto jen vybrané nástroje

Praktické seznámení s dostupnými infrastrukturami pro náročné výpočty a analýzu dat

- superpočítačová a gridová centra v ČR a na MU
- založení účtu + realizace výpočtů na těchto infrastrukturách

Realizace vybraného praktického projektu v oblasti zpracování dat

- případně v oblasti rozvoje nástrojů určených pro zpracování dat
- **vlastní témata vítána!**

DATA & AI LANDSCAPE 2019

INFRASTRUCTURE

- HADOOP ON-PREMISE:** cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, jethro
- HADOOP IN THE CLOUD:** AWS, Microsoft Azure, Google Cloud, SAP Cloud Platform, IBM InfoSphere, IBM BigInsights, IBM Cloud Platform, IBM InfoSphere, IBM Business Partner, Oracle, CAZENA
- STREAMING / IN-MEMORY:** Amazon Kinesis, databricks, SAP, Cloud Platform, ORACLE, confluent, strim, hazelcast, GridGain, GIGASPACE, Wallaroo, FASTDATA, Ix

ANALYTICS & MACHINE INTELLIGENCE

- DATA ANALYST PLATFORMS:** Microsoft, pentaho, alteryx, Digital Reasoning, GUAVUS, AYASDI, ATTIVO, Datameer, incorta, interana, MODE, ENDOR, sisu, switchboard, Starburst
- DATA SCIENCE PLATFORMS:** IBM, databricks, dataiku, DOMINO, rapidminer, TIBCO, ANACONDA, SAS, KNIME, MathWorks

APPLICATIONS - ENTERPRISE

- SALES:** CHORUS, INSIDESALES.COM, people.ai, conversica, clari, avso, tactal, fusesmachines, Clearbit
- MARKETING - B2B:** RADIUS, App Annie, EVERSTRING, MINTIGO, sence, tubular, N GAGGIO, KNOTCH, mrp
- MARKETING - B2C:** zeta, bloomreach, SendGrid, Inzage, ACTION, BLUECORE, CONTENTMODX, TEALUM, importio, Amplero, amperity, QUANTIFIND, Simon, [PERSADO], remesh
- CUSTOMER EXPERIENCE / SERVICE:** qualtrics, MEDALLIA, SurveyMonkey, UserTesting, CLARABRIDGE, zendesk, Customer Freshdesk, INTERCOM, Drift, EVERSON, Gainight, pendio, HEAP, Amplitude, Watson Assistant, Delighted, DigitalGenius, ASAPP, ada, AUTOMAT, ahni, CoDesk, [netomi], [[], frame.ai
- ENTERPRISE PRODUCTIVITY:** slack, ORACLE, GURU lumia, DIFFBOT, clara, talla, Kasisto

APPLICATIONS - INDUSTRY

- ADVERTISING:** AppNexus, critico, xAd, Integral, ORACLE, MOAT, Openx, dataxata, theTradeDesk, algorithm
- EDUCATION:** Leapship, KNEWTON, Clever, Clevera, theTradeDesk, algorithm
- REAL ESTATE:** Redfin, VTS, CREDIFI, GEPHY
- GOVT:** OPENGOV, mark43, LiveStories, Passport, PRIMER
- INTELLIGENCE:** Palantir, Dataminr, Quid, FORGE
- FINANCE - INVESTING:** Kenshuc, Quantopian, ADEFAAR, MANTICORE
- FINANCE - LENDING:** ondeck, affirm, JIANPUAI, Kredtech, AVANT, auro, CLEARBANC, upgrade, 100Credit, WeLab, TrueAccord, MoneyLion, aire, agnifi
- INSURANCE:** Insuramize, Lemonade, CYRENCE, CHippo, Shift Technology, ROOT, CAPE

OPEN SOURCE

- FRAMEWORKS:** Spark, Flink, YARN, TEZ, HADOOP, MESOS, Kubernetes, DOCKER, CDAP, HELIX, Redhat
- QUERY / DATA FLOW:** Spark, SQL, presto, SLAMDATA, GraphQL, Flink
- DATA ACCESS & DATABASES:** cassandra, mongoDB, redis, Cockroach LABS, druid, CQL, CHYTHON, Gc4DB, triak, FISHIE, Oceanbase, accumulo
- ORCHESTRATION & MGMT:** talend, Apache Ambari, Apache Airflow, MESOS, etcd, Kong
- STREAMING & MESSAGING:** Spark, nifi, Flink, beam, kafka, STORM, Apache RocketMQ
- STAT TOOLS & LANGUAGES:** R, Python, Scala, Numpy, Studio, SciPy, julia
- AI OPS & INFRA:** miflow, Kubeflow, mlops, DVC, SELDON, Polysyn
- AI / MACHINE LEARNING / DEEP LEARNING:** TensorFlow, Keras, PyTorch, Caffe, Microsoft Cognitive Toolkit, OpenAI, DM, theano, H2O, Apache SINGA, DIMSUM, FeatureFu, Intel, VELES, Chainer, Microsoft, ONNX, LIGU, PyTorch, neon, DSSTNE, milb, DL4, MAHOUT, Aerosolve, TestAI, mir, OpenML
- SEARCH:** elasticsearch, Solr
- LOGGING & MONITORING:** kibana, SENTRY, logstash, Prometheus, fluentbit, fluentd, Grafana, Vector
- VISUALIZATION:** matplotlib, TensorBoard, seaborn, Bokeh
- COLLABORATION:** BeakerX, Jupyter, Anaconda
- SECURITY:** Apache Ranger, KNOX, Sentry, BCCUTULU

CROSS-INFRASTRUCTURE/ANALYTICS

aws, Google Cloud, Microsoft, IBM, SAP, Hewlett Packard Enterprise, SAS, IO10DATA, vmware, TIBCO, TERADATA, ORACLE, NetApp, syncsort, MAPR, cloudera

DATA SOURCES & APIs

- HEALTH:** VALIDIC, practicefusion, fitbit, GARMIN, HUMANA API, kinsa, MIMIC
- IOT:** GE Digital, UPTAKE, thingworx, helium, samsara
- FINANCIAL & ECONOMIC DATA:** Bloomberg, THOMSON REUTERS, DOW JONES, S&P CAPITAL IQ, CB Insights, PLAID, INVESTMET, YODLEE, The Motley Fool, GEstimize, PREMISE, Quantl, Eagle Alpha, StockTwits, xignite, Thinknum, earnest, predata
- AIR / SPACE / SEA:** piconet, SKYCATCH, AIRBÖTICS, spire, INVOICES, kespry, UNDERSTORY, telluslabs, WINDWARD, DroneDeploy, MarineTraffic, LOTUS
- PEOPLE / ENTITIES:** axion, Experian, EPSILON, InsideView, Crism Hexagon, BASIS, Quantcast, SAFEGRAPH
- LOCATION INTELLIGENCE:** FOURSQUARE, mapbox, sense360, playr/bowser, HEXAGON, PlaceIQ, esri, factual, CARTA, Mapillary, Streetline, cuebiq, Radar, OpenStreetMap
- OTHER:** DATA.GOV, IMAGENET, LOBI, LOBI, LOBI, CRUX, Ugraphia

DATA RESOURCES

- DATA SERVICES:** OPERA, DATA SCIENCE, fractal, kaggle, DataKind, INNOPLCUS
- INCUBATORS & SCHOOLS:** PLURALSIGHT, GA, galvanize, DataCamp, DataElite, INSIGHT, The Data Incubator, METIS
- RESEARCH:** facebook research, OpenAI, MIRI, VECTOR INSTITUTE, AIZ, ALLIANCE FOR ARTIFICIAL INTELLIGENCE

Nečekejte, že Vás naučíme všechno. 😊



Zázemí kurzu

Centrum CERIT-SC – výzkumné centrum vybudované na ÚVT MU

- původně Superpočítačové centrum Brno (SCB)

Poskytovatel HW a SW zdrojů (5500+ jader)

- SMP uzly
- HD uzly (2624 jader)
- SGI UV uzel 384 jader, 6 TB paměti
- SGI UV uzel 504 jader, 10 TB paměti)
- Xeon Phi cluster
- úložné kapacity (~ 3,5 PB)

Služby nad rámec „běžného“ HW centra

- zázemí pro kolaborativní výzkum



Zázemí kurzu

Hlavní cíle Centra CERIT-SC:

- flexibilní infrastruktura, vlastní výzkum v infrastrukturních oblastech
- dva hlavní **výzkumné směry**
 - *High-performance computing* – akcelerace výpočtů, GPU computing, ...
 - ***Big Data analytics + AI***
- úzká spolupráce mezi informatiky a partnery centra
 - výpočetní a úložné kapacity jsou pouze nástrojem
 - zaměření na inteligentní a nové použití těchto nástrojů
 - synergický posun informatiky a spolupracujících věd (kolaborativní výzkum)

Zázemí kurzu

Snaha o maximální zapojení studentů

- **bakalářského -> magisterského -> doktorského studia**
 - vedení závěrečných prací v praktických a užitečných oblastech
 - možnost zapojení studentů do řešených projektů
 - možná podpora finančními granty

Zázemí kurzu

CERIT-SC – vybrané příklady spoluprací (datová analytika)

- spolupráce s Policií ČR
 - vývoj nástrojů pro datovou analytiku kriminálních činů
 - uživatelé jsou policejní analytici
 - příležitostná spolupráce na analýze dat reálných kauz
- spolupráce s výzkumnými partnery (uvnitř i vně MUNI)
 - mnoho spoluprací na pomezí IT a spolupracujících věd
 - Ústav výzkumu globální změny, bioinformatika a analýzy genomu, Ústav fyziky materiálů AV ČR, ...
- spolupráce s komerčními subjekty a státními organizacemi
 - aktuálně se rozvíjející spolupráce s RedHat a.s.
 - dlouhodobá spolupráce se společností MycroftMind a.s.

Zázemí kurzu

Zázemí kurzu nabízí možnost skloubení výuky s pokročilou praxí

- můžete podpořit svou přípravu do budoucí (komerční) praxe či akademické kariéry
 - případně rozvíjet se i v non-IT oblasti, která je pro Vás zajímavá
- spolupracemi již prošlo mnoho studentů
 - dlouhodobé zkušenosti (cca 8 let)
 - velmi pozitivní zpětná vazba
 - získaná praxe pro hledání zaměstnání, nabídka pracovních pozic u partnerů, dlouhodobější spolupráce s ÚVT/CESNETem, ...
- **Vaše dosavadní znalosti a zkušenosti nejsou podmínkou, důležitá je vlastní motivace**



Průběh kurzu

aneb „O zajímavý obsah se podělíme“

Teoretické přednášky

- úvodních cca 6 týdnů
- doplníme o zvanou/é přednášku/y dalších kolegů
 - specialistů na oblast analýzy/zpracování dat
 - máte nějaké podněty, co byste chtěli slyšet?
- slidy budou dostupné v IS MU (po proběhlém kurzu)

Zadání praktického projektu

- práce ve skupinách (3 studenti)
- témata nejen zajímavá, ale i užitečná (Vaše výsledky budou prospěšné)

Průběžné konzultace při zpracování projektů

Prezentace výsledků, závěrečné shrnutí

Podmínky úspěšného ukončení

- účast na přednáškách (na většině)
- realizace praktického projektu a prezentace výsledků
- **vlastní krví stvrzená celoživotní oddanost datové analytice a budoucí spolupráce 😊 😊 😊**

Historické praktické projekty – jaro 2019

Ústav výzkumu globální změny AV ČR (CzechGlobe)

- příprava analytických pohledů pro analýzu dat meteorologických měření
 - v nástroji Kibana (+ Elasticsearch)
- návaznost běžící diplomovou prací

Policie ČR

- příprava nástroje pro nahrávání dat do skladu rozsáhlých heterogenních dat
 - podpora běžícího projektu
- 4 studenti zapojeni do běžícího projektu

Masarykova univerzita – CESNET

- analýza dat jednotného přihlášení
 - odhalování abnormálního chování uživatele
- předběžná domluva na zapojení studentů zapojených do tohoto projektu

Historické praktické projekty – podzim 2019

Sběr a analýza dat využití IT infrastruktury MU

- MU disponuje rozsáhlou IT infrastrukturou, jejíž využití je nezbytné sledovat pro účely rozhodování o dalších investicích do ní
 - disková úložiště, výpočetní kapacity (servery), síťová infrastruktura, služby, ...
- cíle projektu: příprava infrastruktury a produkčního systému pro sběr a vyhodnocení dat + příprava typových datových analýz

Analýza/sběr/vizualizace dat z energetického sektoru

- ve spolupráci se společností MycroftMind a.s.
- cíl projektu: analýza dat topologie smartmeterů energetické soustavy
 - a odhalování abnormalit v nich
- pokračující spolupráce s 1 studentem

Průzkum a dokumentace (pokročilých) vlastností analytického nástroje Kibana (ElasticSearch) a rozvoj integrující platformy CopAS

- s primárním zaměřením na analýzu síťových toků
- pokračující spolupráce se 2 studenty

Aktuální praktické projekty – jaro 2020

Prozatimní témata (budou dopřesněna)

- škálovatelnost grafových databází – až na samou hranici jejich schopností
 - výsledky budou využity pro reprezentaci a analýzy proteinů
- analýza obrazových dat ve spolupráci s Moravskou zemskou knihovnou
 - identifikace obrázků v digitalizovaných dokumentech a jejich klasifikace
 - vyhledávání podobných obrázků
 - identifikace osob na obrázcích
 - vyhledávání knih podle obálky
- zpracování leteckých dat Ústavu výzkumu globální změny
- ...

a možná i překvapení ...

- praktické využití **největšího holografického displeje v ČR**
 - včetně ovládání Kinectem
 - možné využití vícero týmy



M U N I

I C S

M U N I

I C S