

Výskumne orientovaný seminár 2

Mária Bieliková, Róbert Móra, Jakub Ševcech

O nás...



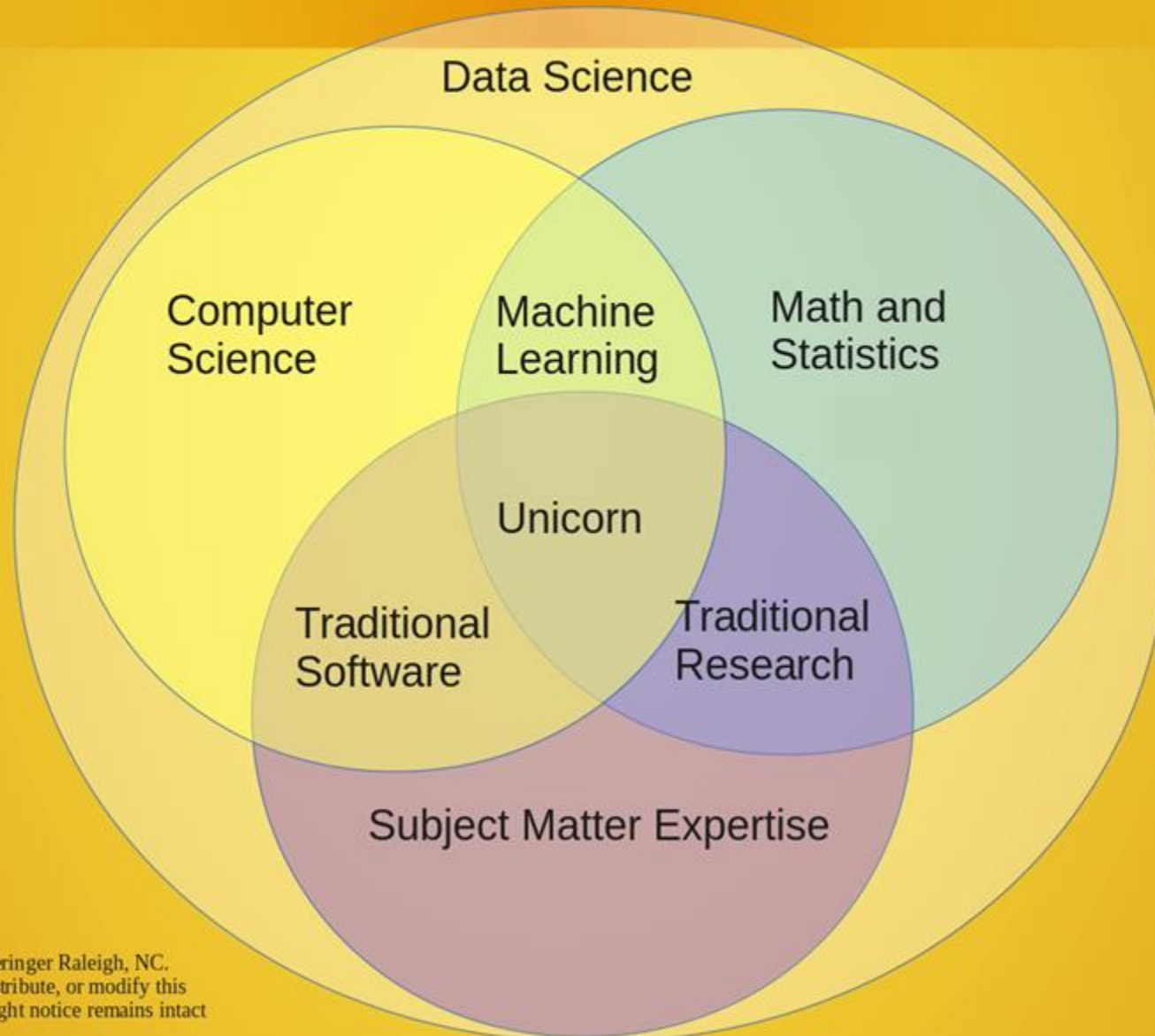
O vás...

Na seminároch sa budeme
venovať dátovej vede

Neexistuje presná definícia, čo je to dátová veda. Dátový vedec je...

- Sexy označenie pre štatistika
 - <http://www.statisticsviews.com/details/feature/5133141/Nate-Silver-What-I-need-from-statisticians.html>
- Programátor v oblasti strojového učenia / dolovania v dátach, ktorý má znalosť o doméne
 - <http://shakthydoss.com/technical/what-is-the-difference-between-artificial-intelligence-machine-learning-statistics-and-data-mining/#comment-261>
- Pokus o rebranding; niekto, kto ovláda štatistiku lepšie ako ktorýkoľvek softvérový inžinier a je lepší v softvérovom inžinierstve ako ktorýkoľvek štatistik
 - <https://www.quora.com/What-is-data-science/answer/Rahul-Agarwal-10>

Data Science Venn Diagram v2.0



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 Endpoints for Insights
arm

STREAMING / IN-MEMORY

Amazon Kinesis Databricks
SAP Cloud Platform Oracle Confluent
strim hazelcast Goinc4n
GIGASPACE Wallaroo FASTOMA Ix

ANALYTICS & MACHINE INTELLIGENCE

DATA ANALYST PLATFORMS

Microsoft pentaho alteryx

Digital guavus AYASDI

ATTIVO Datameer incorta.

interana MODE ENDOR

DATA SCIENCE PLATFORMS

IBM databricks dataiku

DOMINO rapidminer TIBCO

ANACONDA sas

KNIME MathWorks

APPLICATIONS – ENTERPRISE

BI PLATFORMS

- looker
- Alteryx
- Tableau
- aws
- Qlik
- ARC42 DATA
- ThoughtSpot
- ATSCALE
- Microsoft
- Olik
- Google Analytics
- Information Builders
- ibirst
- MicroStrategy
- Klarna IQ

VISUALIZATION

- tableau
- Microsoft Power BI
- SAP
- Google Cloud
- Perceptics
- CELONIS
- zepl
- VISION4DATA
- plioty
- CHARTIO
- FLUXUS TIO

MACHINE LEARNING

- Alteryx
- Microsoft Azure Machine Learning
- Amazon SageMaker
- Google Cloud
- H2O
- dataRobot
- gamalon
- VIZEN ELEMEN
- deeppense.ai

DATA TRANSFORMATION

- talend
- pentaho
- alteryx
- TRIPACTA
- stream
- Paxata
- StormSets
- UNIFI

DATA INTEGRATION

- SAP Data Services
- Informatica
- Microsoft
- TERADATA
- inprolog
- enigma
- segment
- ATTENITY
- zaloni
- import.io
- Infovision
- Flowize
- SHOWLOW
- MATELION

DATA GOVERNANCE

- Informatica
- IBM
- McAfee
- colibra
- Alation
- drimio
- Altera
- OKERA
- data.world

AWS / MONITORING

- mgmt
- New Relic
- octlo
- rubrik
- APDDYNAMICS
- dynatrace
- WAVEFORM
- Signifia
- esprimo
- spunk
- Moogsoft
- padguru
- unicon
- Namify
- dataScience
- zoores
- servave
- UNIFI

The collage is organized into three distinct sections, each with a title in a red banner at the top:

- COMPUTER VISION:** This section includes logos for Microsoft Azure, Amazon Rekognition, darfai, EVERAI, deepomatic, neurio, twentybin, URBINO, AODE, 视觉云, YITU, trac, and 爱康.
- HORIZONTAL AI:** This section features logos for AWS Watson, Cortana, PACE, 智视, sentient, Voyagex, 智云, Affective, PROGRESS, 智云, Numeta, FUTU, 智云, nara.io, 智云, OSARO, BLUE VISION, and 爱康.
- SPEECH & NLP:** This section displays logos for Google Cloud, twitter, Amazon Alexa, Comma.ai, narrative, 智云, Malivo, Eiger, SoundHound Inc., PRIMA, Verbit, 智云, URBINO, cogito, snips, and 智云.

Industry	Applications
ADVERTISING	10
EDUCATION	10
REAL ESTATE	10
GOVT	10
INTELLIGENCE	10
FINANCE-INVESTING	10
FINANCE-LENDING	10
INSURANCE	10

Category	Providers and Services
STORAGE	aws, Google Cloud, Microsoft Azure, Oracle Cloud, IBM Cloud, Amazon S3, Google Drive, Microsoft OneDrive, Oracle Cloud, IBM Cloud, Amazon S3, Google Drive, Microsoft OneDrive, Oracle Cloud, IBM Cloud, Amazon S3, Google Drive, Microsoft OneDrive
CLUSTER SVCS	IBM, Microsoft, Amazon, Google, Oracle, IBM, Microsoft, Amazon, Google, Oracle, IBM, Microsoft, Amazon, Google, Oracle, IBM, Microsoft, Amazon, Google, Oracle
DATA GENERATION & LABELLING	Amazon, Microsoft, Google, Oracle, IBM, Microsoft, Amazon, Google, Oracle, IBM, Microsoft, Amazon, Google, Oracle, IBM, Microsoft, Amazon, Google, Oracle
AI OPS	Google, Microsoft, Amazon, Oracle, IBM, Microsoft, Amazon, Google, Oracle, IBM, Microsoft, Amazon, Google, Oracle, IBM, Microsoft, Amazon, Google, Oracle
GPU DBs & CLOUD	Microsoft, Amazon, Google, Oracle, IBM, Microsoft, Amazon, Google, Oracle, IBM, Microsoft, Amazon, Google, Oracle, IBM, Microsoft, Amazon, Google, Oracle
HARDWARE	Google, Microsoft, Amazon, Oracle, IBM, Microsoft, Amazon, Google, Oracle, IBM, Microsoft, Amazon, Google, Oracle, IBM, Microsoft, Amazon, Google, Oracle

The diagram is organized into four main categories, each with a list of associated companies:

- SEARCH**
 - Elasticsearch
 - Algolia
 - Lucidworks
 - Swifttype
 - Alphasense
 - Omni-us
- LOG ANALYTICS**
 - Splunk
 - Sumologic
 - Solarwinds
 - Timber
 - Kibana
 - Logzio
- SOCIAL ANALYTICS**
 - Hootsuite
 - Sprinklr
 - Netbase
 - Synthesio
 - SimpleReach
 - Bitly
- WEB / MOBILE / COMMERCE ANALYTICS**
 - Google Analytics
 - Mixpanel
 - Amplitude
 - Airtable
 - Resc
 - Sigopt
 - Granify

HEALTHCARE
 Flatiron Clover Zebra 3D Med Tempus
 Flatiron Clover Zebra 3D Med Tempus
 Flatiron Clover Zebra 3D Med Tempus

LIFE SCIENCES
 Flatiron Clover Zebra 3D Med Tempus
 Flatiron Clover Zebra 3D Med Tempus
 Flatiron Clover Zebra 3D Med Tempus

TRANSPORTATION
 Flatiron Clover Zebra 3D Med Tempus
 Flatiron Clover Zebra 3D Med Tempus
 Flatiron Clover Zebra 3D Med Tempus

AGRICULTURE
 Flatiron Clover Zebra 3D Med Tempus
 Flatiron Clover Zebra 3D Med Tempus
 Flatiron Clover Zebra 3D Med Tempus

COMMERCE
 Flatiron Clover Zebra 3D Med Tempus
 Flatiron Clover Zebra 3D Med Tempus
 Flatiron Clover Zebra 3D Med Tempus

INDUSTRIAL
 Flatiron Clover Zebra 3D Med Tempus
 Flatiron Clover Zebra 3D Med Tempus
 Flatiron Clover Zebra 3D Med Tempus

OTHER
 Flatiron Clover Zebra 3D Med Tempus
 Flatiron Clover Zebra 3D Med Tempus
 Flatiron Clover Zebra 3D Med Tempus

CROSS-INFRASTRUCTURE/ANALYTICS

aws Google Cloud Microsoft IBM SAP Hewlett Packard Enterprise SAS 1010DATA vmware TIBCO TERADATA ORACLE NetApp syncsort MAPR cloudera

OPEN SOURCE

The diagram illustrates a comprehensive ecosystem of data science and machine learning tools, organized into 12 functional categories:

- FRAMEWORKS:** Includes TensorFlow, Spark, PyTorch, Flink, Hadoop, HBase, Hive, and Mahout.
- QUERY / DATA FLOW:** Includes Spark SQL, Presto, Tez, SLAMDATA, and GraphQL.
- DATA ACCESS & DATABASES:** Includes MongoDB, Redis, Cockroach Labs, Druid, and others.
- ORCHESTRATION & MGMT:** Includes Talend, Apache Airflow, and others.
- STREAMING & MESSAGING:** Includes Spark Streaming, Flink, Kafka, and others.
- STAT TOOLS & LANGUAGES:** Includes R, Python, Scala, and others.
- AI OPS & INFRA:** Includes MLOps tools like MLflow, DVC, and others.
- AI / MACHINE LEARNING / DEEP LEARNING:** Includes TensorFlow, Keras, PyTorch, and others.
- SEARCH:** Includes Elasticsearch, Solr, and others.
- LOGGING & MONITORING:** Includes ELK Stack (Elasticsearch, Logstash, Kibana) and others.
- VISUALIZATION:** Includes Tableau, Matplotlib, and others.
- COLLABORATION & SECURITY:** Includes tools for team collaboration and security like Apache Ranger and Sentry.

DATA SOURCES & APIs

HEALTH Apple Validic practice fusion fitbit GARMIN HUMAN API kinso MIMIC

IOT GE Digital UPTAKE thingworx helium samsara

FINANCIAL & ECONOMIC DATA Bloomberg THOMSON REUTERS DOW JONES SEP CAPITAL IQ CBINSIGHTS PLAID SIFTED EVEREST ESTIMIZE PREMISE Quandt Engage Alpha StockTwits xignite Thinknum earnest predata

AIR / SPACE / SEA Airbotics spire kesyri UNDERSTORY telluslabs WINDWARD DroneDeploy MarineTraffic L3HARRIS Pylonet SKYWATCHER

PEOPLE / ENTITIES axiometer experian EPSILON InsideView Crism Hexagon BASIS Quantcast SAFEGRAPH

DATA RESOURCES

DATA INTELLIGENCE

- RSquare
- Mapbox
- 560
- active data
- HEXAGON
- esri
- factual
- Mapillary
- OpenStreetMap
- A Radar
- OpenStreetMap

OTHER

- DATA.GOV
- IMAGENET
- DATA
- CRUX
- graffiti.io

DATA SERVICES

- OPERA
- IMAGENET
- DATA SCIENCE
- fractal
- kaggle
- DataKind
- EXEL
- innopactus

INCUBATORS & SCHOOLS

- PLURALSIGHT
- GA
- galvanize
- DataCamp
- DataElite
- INSIGHT
- The Data Incubator
- METIS

RESEARCH

- OpenAI
- facebook research
- MIRI
- VECTOR INSTITUTE
- AI2
- ALLEN INSTITUTE
- ARTIFICIAL INTELLIGENCE

July 16, 2019 - FINAL 2019 VERSION

© Matt Turck (@mattturck), Lisa Xu (@lisaxu92), & FirstMark (@firstmarkcap) mattturck.com/data2019

FIRSTMARK 
EARLY STAGE VENTURE CAPITAL

<https://mattturck.com/data2019/>

Štatistika vs. strojové učenie vs. dolovanie v dátach

- <http://shakthydoss.com/technical/what-is-the-difference-between-artificial-intelligence-machine-learning-statistics-and-data-mining/>
- Ak opisujete dáta (deskriptívna štatistika) alebo odvodzate závery o populácii na základe jej vzorky (inferenčná štatistika), tak ste zrejme *štatistik*
- Ak využívate inferenčnú štatistiku na tvorbu algoritmov, ktoré sú schopné samé sa učiť, tak zrejme robíte *strojové učenie*
- Ak využívate strojové učenie na riešenie konkrétneho problému a deskriptívnu štatistiku na opísanie dát a výsledkov, tak zrejme *dolujete v dátach*

Čo robí dátový vedec?

1. Definuje (formuluje) otázky a hypotézy
2. Definuje ideálnu dátovú sadu
 - a. Zisťuje, k akým dátam má prístup
 - b. Získava dáta
 - c. Čistí ich
 - d. Skúma ich vlastnosti (exploratívna analýza)
3. Identifikuje a realizuje vhodný typ analýzy
 - a. Štatistické modelovanie/predikcia
4. Interpretuje a komunikuje výsledky (dátový produkt)
5. Automatizuje kroky 2-4 pre predikcie na nových dátach

Úvodný kvíz na zamyslenie

1. „Keď chceš mať viac lajkov, musíš v statuse používať emotikony,“ tvrdí kamarát. Ako by ste jeho tvrdenie ako dátový vedec overili? Ako by ste pri tom argumentovali? Na čo si pritom treba dať pozor?
2. Vlastníte byt, ktorý chcete predať. Potrebujete teda určiť jeho cenu.
 - a. Ako by ste ako dátový vedec postupovali?
 - b. Aké dáta by ste na to potrebovali? Kde by ste ich získali?
3. Banka zbiera údaje o klientoch (rádovo tisíce klientov) a chcela by zistiť, ako na základe nich rozhodnúť, komu poskytnúť úver, a komu nie.
 - a. Aké dáta by ste od bankára vyžadovali? V akej forme?
 - b. Ako by ste ako dátový vedec postupovali pri návrhu spôsobu rozhodovania?
 - c. Ako by ste bankárovi vysvetlili/prezentovali vami navrhnutý model?
 - d. Ako by ste presviedčali banku, že je vami navrhnutý spôsob spoľahlivý?

Organizácia seminárov a podmienky absolvovania

VOS2 vs. IAU

Organizácia a témy seminárov

- Nástroje na analýzu a spracovanie dát
 - Python a R
- **Blok I: Základy analýzy dát a strojového učenia**
 - Exploratívna analýza a predspracovanie dát
 - Transformácie a tvorba odvodených črt
 - Tvorba a optimalizácia parametrov modelov strojového učenia
 - Vyhodnocovanie a výber modelov
 - Komunikovanie výsledkov
 - *Prípadová štúdia 1*
- **Blok II: Metódy racionality**
 - *Esej* o identifikovaných metódach racionality v zadanom štúdijnom texte (vypracovaná v 2-3 členných tímoch)
 - <http://yudkowsky.net/rational>, <http://www.hpmor.com>
- **Blok III: Pokročilé témy**
 - *Prípadová štúdia 2* na zadanú tému vypracovaná v 2-3 členných tímoch

Podmienky absolvovania

Klasifikovaný zápočet

Základy R	10b	(zadanie 2. týždeň, odovzdanie do semináru v 3. týždni)
Základy Pythonu	10b	(zadanie 2. týždeň, odovzdanie do semináru v 3. týždni)
Prípadová štúdia 1	30b	(zadanie v 3. týždni, odovzdanie v 7. týždni)
Analýza metód racionality	10b	(zadanie v 7. týždni, odovzdanie v 10. týždni)
Prípadová štúdia 2	40b	(zadanie v 7. týždni, záväzné nahlásenie témy v 9. týždni, odovzdanie v 12. týždni)
Spolu	100b	