

MUNI

# Zpracování a analýza (velkých) dat

CORE042 Data – odpověď na základní otázku života, vesmíru a vůbec...

2. přednáška

**Tomáš Rebok** <[rebok@ics.muni.cz](mailto:rebok@ics.muni.cz)>

Centrum CERIT-SC – Ústav výpočetní techniky MU

# Přednášející



## – Tomáš Rebok

- ÚVT MU, Centrum CERIT-SC & Divize IT infrastruktury
- senior výzkumný pracovník, vedoucí projektů a aplikačně-výzkumných skupin
- dlouhodobá činnost v oblasti výpočetních a datových infrastruktur pro podporu náročných výpočtů a zpracování dat
- primární orientace na oblast datové analytiky

# Obsah přednášky

## Co už znáte ...

- informace, data, metody uložení dat
- životní cyklus dat, otevřená věda a FAIR data

## A co nás dnes čeká?

- seznámení s metodami pro zpracování a analýzu dat
  - typy dat a jejich reprezentace
  - včetně tzv. Big Data přístupů
- seznámení s dostupnými infrastrukturami pro náročné výpočty a analýzu dat
  - superpočítačová a gridová centra v ČR
- příklady datových analýz v rozdílných aplikačně-výzkumných oblastech



## Zpracování (velkých) dat

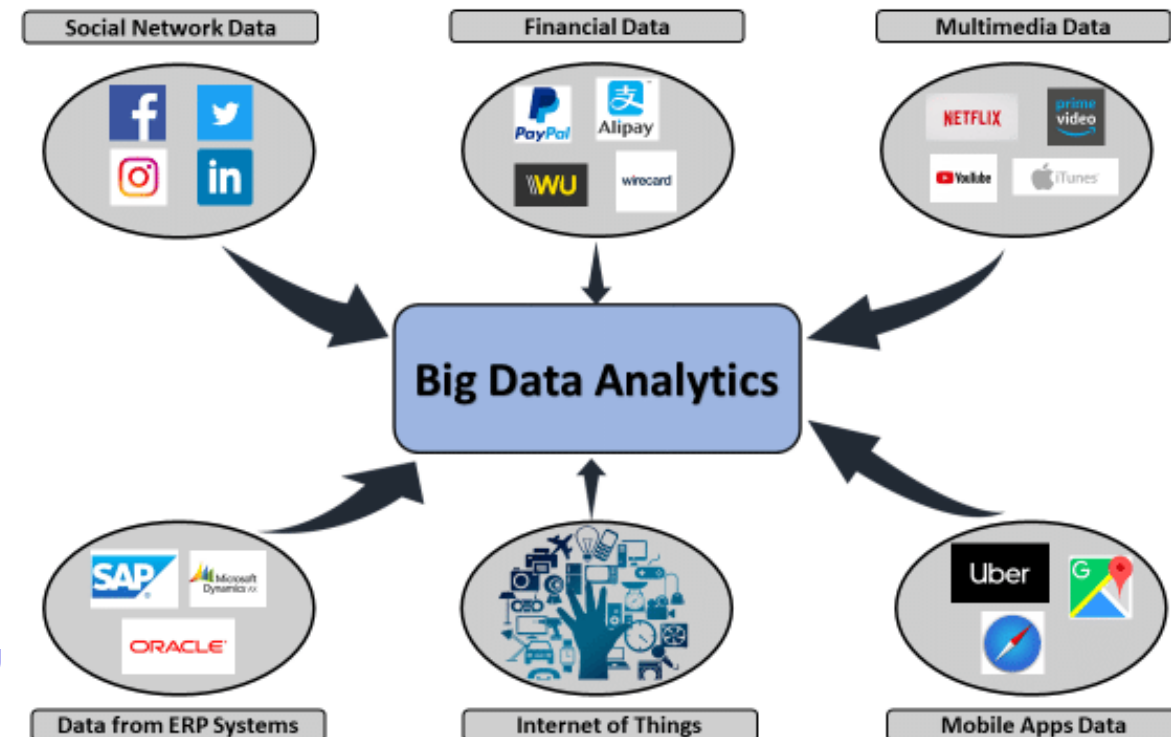
aneb Co jsou to ta velká data?



# Data – fenomén dnešní doby



- potenciálním zdrojem dat je prakticky cokoli (a kdokoli)
- vhodné vytěžování dat může odpovědět na mnoho otázek
  - ovlivňujících jak byznys, tak i pokrok společnosti
- problém není data generovat
- problém již není ani data uložit
- **problém je tato data zpracovat**
  - resp. získat z nich užitečné informace





# Zpracování dat (fáze *Process*)

## – vstup: **co potřebujeme?**

- dobře popsáný problém 😊
- data sesbíraná v rámci realizovaného experimentu  
viz fáze *Plan* a *Collect*
- představu o struktuře dat a požadavcích analýzy dat (předpokládané dotazy)

## – **co musíme vzít v úvahu?**

- citlivá data je nutno anonymizovat/pseudonymizovat  
nutnost odstranit identifikovatelnost zachycených subjektů (i nepřímou!)
- data bychom měli reprezentovat standardizovanými technologiemi  
konverze do požadovaných formátů a příprava pro další analýzu
- anonymizace/pseudonymizace dat i reprezentace jejich formátu musí být pečlivě dokumentovány
- fáze zpracování dat **často úzce souvisí s fází jejich analýzy**

# Základní členění dat



## 1. Strukturovaná data

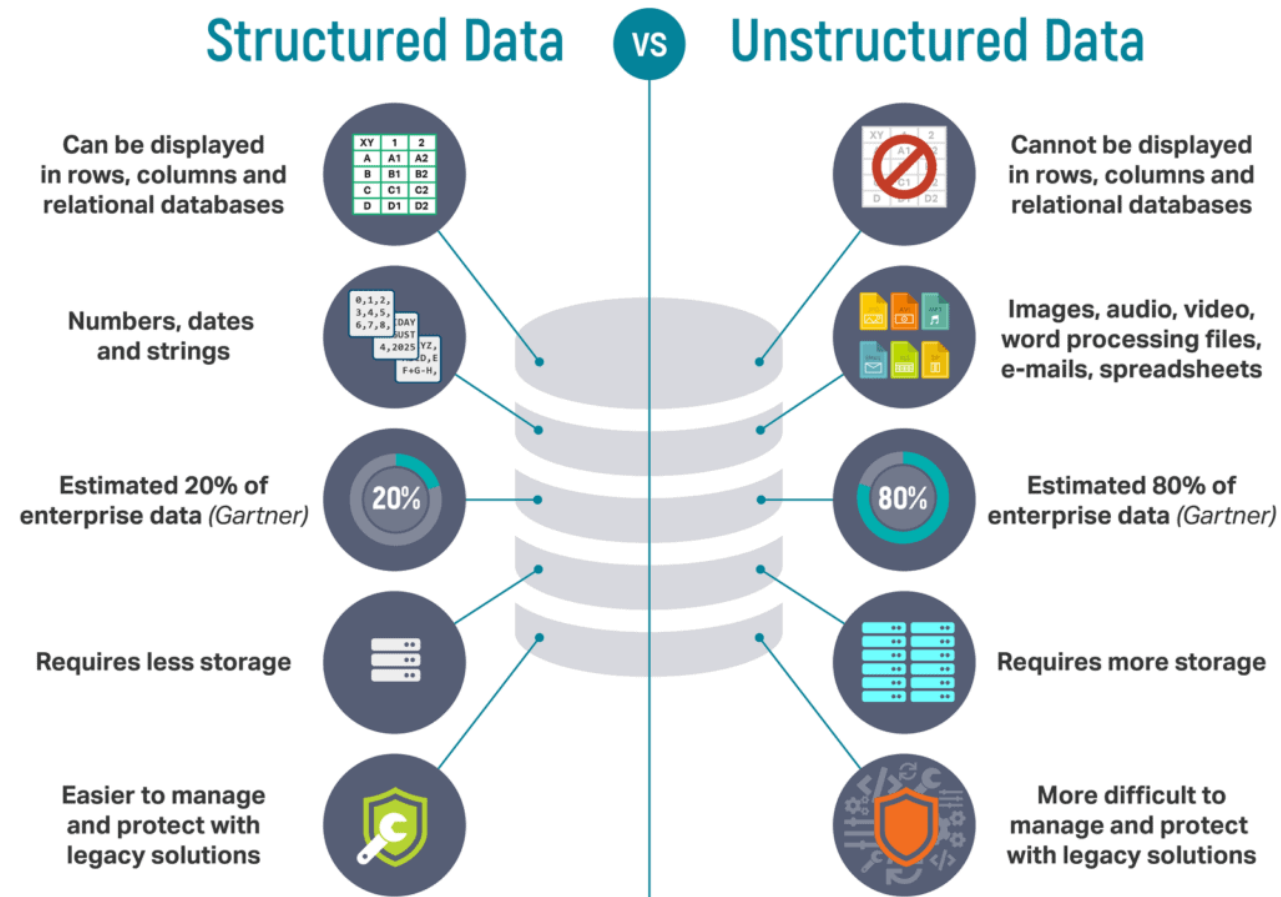
- data s identifikovatelnou strukturou
- typicky reprezentovaná tabulkami

## 2. Nestrukturovaná data

- data bez jasné struktury, bez modelu
- typicky multimediální data

## 3. Semi-strukturovaná data

- kombinace obojího
- nestrukturovaná data s částečnou strukturovanou informací (tzv. metadaty)





# 1. Strukturovaná data

- **typicky tabulková data**, reprezentovaná sloupci a řádky
  - sloupce = vlastnosti konkrétního záznamu, řádky = jednotlivé (různé) záznamy
- data **mají definovanou strukturu**, která je neměnná
  - resp. mění se jen velmi omezeně
  - např. finanční transakce, záznamy prodejů, ...
- reprezentace (a analýza) strukturovaných dat:
  - **jedinou tabulkou** (*MS Excel-style*)
  - **relační databází** (soubor vzájemně provázaných tabulek)
    - tzv. SQL databáze
    - např. *MS Access-style*

	A	B	C	D	E	F	G
1	OrderDate	Region	Rep	Item	Units	Unit Cost	Total
2	9/1/2014	Central	Smith	Desk	2	125	250
3	6/17/2015	Central	Kivell	Desk	5	125	625
4	9/10/2015	Central	Gill	Pencil	7	1.29	9.03
5	11/17/2015	Central	Jardine	Binder	11	4.99	54.89
6	10/31/2015	Central	Andrews	Pencil	14	1.29	18.06
7	2/26/2014	Central	Gill	Pen	27	19.99	539.73
8	10/5/2014	Central	Morgan	Binder	28	8.99	251.72
9	12/21/2015	Central	Andrews	Binder	28	4.99	139.72
10	2/9/2014	Central	Jardine	Pencil	36	4.99	179.64
11	8/7/2015	Central	Kivell	Pen Set	42	23.95	1005.9
12	1/15/2015	Central	Gill	Binder	46	8.99	413.54
13	1/23/2014	Central	Kivell	Binder	50	19.99	999.5
14	3/24/2015	Central	Jardine	Pen Set	50	4.99	249.5
15	5/14/2015	Central	Gill	Pencil	53	1.29	68.37
16	7/21/2015	Central	Morgan	Pen Set	55	12.49	686.95
17	4/10/2015	Central	Andrews	Pencil	66	1.99	131.34
18	12/12/2014	Central	Smith	Pencil	67	1.29	86.43





# 1. Strukturovaná data – relační databáze

- **Relace** = tabulka atributů (sloupce) se záznamy (řádky)
- **Relační databáze** = nástroj pro definici relací, manipulaci s uloženými daty a dotazování (analýzu)

- **SQL** – *Structured Query Language*  
= standardní dotazovací jazyk

```
SELECT column1, column2 FROM table  
WHERE column1='value'
```

- korektní návrh relační databáze se řídí několika pravidly
  - tzv. normální formy
  - pomáhají udržet řád, usnadňují analýzu

id	name	age
1	Jim	28
2	Pam	26
3	Michael	42

id	subject	Teacher
1	Languages	John Jones
2	Track	Wally West
3	Swimming	Arthur Curry
4	Computers	Victor Stone

student_id	subject_id	grade
2	1	98
1	2	100
1	4	75
3	3	60
2	4	76
3	2	88



## 2. Nestrukturovaná data

- data, která **nejsou** uspořádána podle předem definovaného datového modelu
  - resp. tento model není znám
- drtivá většina dat (*dle Gartner 80 % všech dat*)
- **typické zdroje** nestrukturovaných dat:
  - dokumenty, faktury, smlouvy, emaily, formuláře, ...
  - obrázky, videa, audiozáznamy, geoprostorová data, ...
  - data ze senzorů a zařízení, data z počítačových systémů (logy) ...
  - binární (= obecné) soubory
- **analýza s využitím specializovaných DB** – tzv. NoSQL databáze
  - jedná se o mnohem větší objemy než v případě strukturovaných dat



# 3. Semi-strukturovaná data

## = částečně strukturovaná data

- např. nestrukturovaná data s doprovodným informacemi (tzv. metadaty)
  - příp. navíc s proměnnou strukturou
- doprovodné informace (metadata) slouží pro prohledávání/analýzu
- **příklad 1: emailové zprávy**
  - tělo emailu (text zprávy) = nestrukturovaná data
  - hlavička emailu (odesílatel, příjemce, datum a čas odeslání, ...) = strukturovaná informace
- **příklad 2: digitální fotografie**
  - zachycený obrázek = nestrukturovaná data
  - datum a čas pořízení, clona, čas závěrky, ID zařízení, ... = strukturovaná informace
  - některé strukturované informace lze doplnit až po zpracování
    - např. informace o zachycených objektech (pes, kočka, osoba, ...) – umělá inteligence
- **drtivá většina nestrukturovaných dat je spíše semi-strukturovaných**

## Analýza (velkých) dat

aneb Kdy se bavíme o tzv. Big Data přístupu?





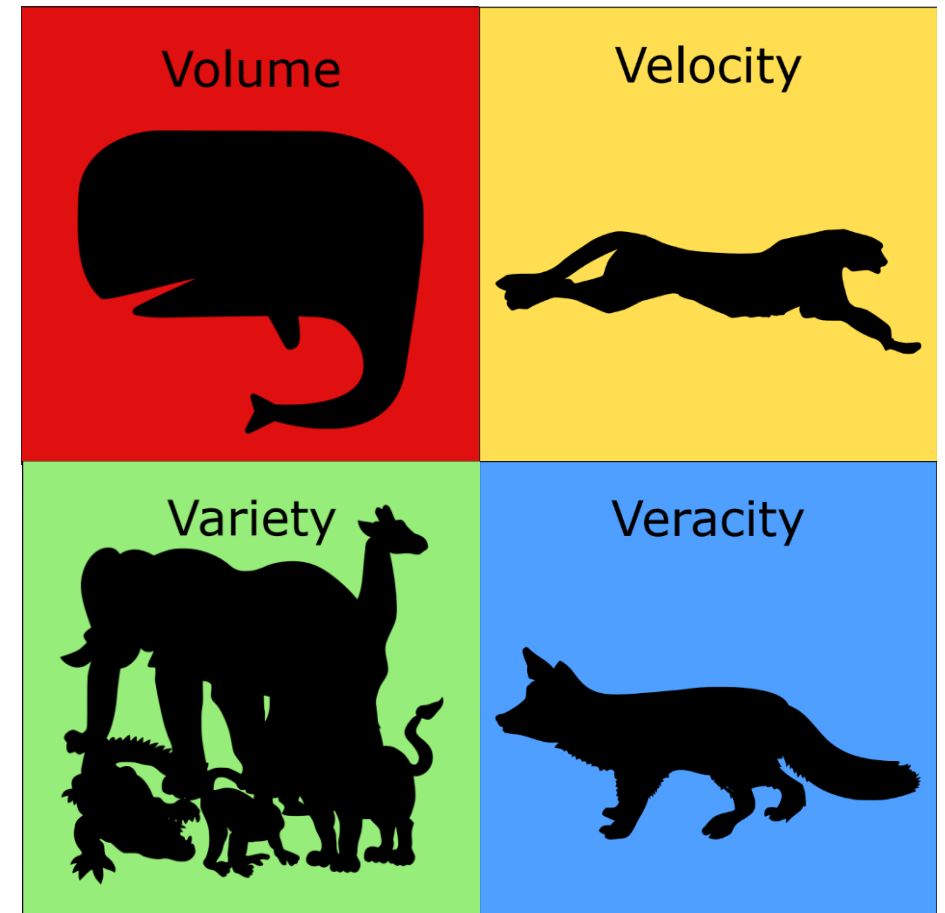
# Big Data

## Co jsou to Big Data?

- data velkých objemů 😊
  - větších, než je možné zpracovávat jednoduchými prostředky
- **ale nejen to:**
  - data, která nelze zachytit jednoduchými strukturami
  - resp. data, jejichž struktura se mění
  - resp. data, která nelze jednoduše zpracovat
- Big Data přístupy byly navrženy v souvislosti s potřebou analyzovat nestructurovaná (resp. semi-structurovaná) data

# Big Data – definice

- data vyhovující (některému z) tzv. **4V**
  - **Volume (objem)** – data velkých objemů
  - **Velocity (rychlost)** – data, která vznikají (přicházejí) rychleji než jak je možno je (standardně) zpracovat
  - **Variety (rozdílnost)** – data různých struktur a typů, různorodého charakteru
  - **Veracity (věrohodnost)** – nutnost čištění nekonzistentních/neúplných dat (např. data ze sociálních sítí)
- občas uváděno jen jako **3V** (bez *Veracity*)
  - ale také jako **7V** (+ *Variability*, *Visualization*, *Value*) nebo též až 42V 😊



THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE



INFRASTRUCTURE: STORAGE (AWS, Microsoft, Google), MPP DBs (Teradata, Vertica), DATA LAKES/LAKEHOUSES (Dremio, Databricks), DATA WAREHOUSES (Snowflake, Oracle, Amazon), STREAMING/IN-MEMORY (Kafka, Flink, Kinesis).

ANALYTICS: BI PLATFORMS (Looker, Tableau, Power BI), VISUALIZATION (Tableau, Power BI, Qlik), DATA SCIENCE NOTEBOOKS (Jupyter, Databricks, Orange3).

MACHINE LEARNING & ARTIFICIAL INTELLIGENCE: DATA SCIENCE PLATFORMS (Databricks, AWS, Azure), ENTERPRISE ML PLATFORMS (Databricks, AWS, Azure), DATA GENERATION & LABELING (Scale AI, Labelbox).

APPLICATIONS - ENTERPRISE: SALES (Salesforce, HubSpot), MARKETING (Marketo, Pardot), CUSTOMER EXPERIENCE (Intercom, Zendesk), HUMAN CAPITAL (Gigamonks, Workday).

DBMS: NoSQL DATABASES (Cassandra, MongoDB), NewSQL DATABASES (CockroachDB, Yugabyte), REAL TIME DATABASES (Couchbase, Redis), GRAPH DBs (Neo4j, Amazon Neptune).

DATA ANALYST PLATFORMS (Alteryx, Tableau), CUSTOMER DATA PLATFORMS (Salesforce, Oracle), PRODUCT ANALYTICS (Mixpanel, Amplitude), LOG ANALYTICS (Datadog, Splunk).

DATA GENERATION & LABELING (Scale AI, Labelbox), MLOPS (Weights & Biases, MLflow), NLP (OpenAI, Google), HORIZONTAL AI / AGI (Anthropic, OpenAI).

APPLICATIONS - HORIZONTAL: CODE & DOCUMENTATION (GitHub, GitLab), TEXT (OpenAI, Google), AUDIO & VOICE (OpenAI, Google), IMAGE (OpenAI, Google), VIDEO EDITING (Runway, Pictory).

ETL / ELT / DATA TRANSFORMATION (Talend, Alteryx, Informatica), REVERSE ETL (Census, Segment), DATA INTEGRATION (Informatica, Talend), DATA GOVERNANCE & CATALOG (Collibra, Alation).

PRODUCT ANALYTICS (Mixpanel, Amplitude), LOG ANALYTICS (Datadog, Splunk), CRYPTO / WEB 3 ANALYTICS (Chainalysis, Elliptic), AI HARDWARE (NVIDIA, Intel).

AI HARDWARE (NVIDIA, Intel), GPU CLOUD (Paperspace, Lambda), CLOSED SOURCE MODELS (OpenAI, Google), EDGE AI (Intel, NVIDIA).

APPLICATIONS - INDUSTRY: FINANCE & INSURANCE (Kenshuc, Upstart), HEALTHCARE (Flatiron, Tempus), LIFE SCIENCES (Moderna, Vertex), TRANSPORTATION (Uber, Tesla).

ORCHESTRATION (Astronomer, Prefect), DATA QUALITY & OBSERVABILITY (Great Expectations, Monte Carlo), FULLY MANAGED (Databricks, Snowflake), MGMT / MONITORING (New Relic, Datadog).

ENTERPRISE SEARCH (Elastic, Algolia), QUERY ENGINE (Dremio, Snowflake), PRIVACY & SECURITY (OneTrust, TrustArc), COMPUTE (AWS, Azure, GCP).

AI MODELS & ARCHITECTURES (OpenAI, Google), SEARCH (Elastic, Algolia), LOGGING & MONITORING (Datadog, New Relic), VISUALIZATION (Tableau, Power BI).

FINANCE & INSURANCE (Kenshuc, Upstart), HEALTHCARE (Flatiron, Tempus), LIFE SCIENCES (Moderna, Vertex), TRANSPORTATION (Uber, Tesla).

OPEN SOURCE INFRASTRUCTURE: FRAMEWORKS (PyTorch, TensorFlow), FORMAT (Parquet, Avro), QUERY / DATA FLOW (Dagster, Airflow), DATA ACCESS (Fivetran, Stitch), DATABASES (PostgreSQL, MySQL), OLAP (ClickHouse, Druid), ORCHESTRATION (Kubernetes, Docker), INFRA-STRUCTURE (AWS, Azure, GCP), DATA OPS (Maven, Gradle), STREAMING & MESSAGING (Kafka, Pulsar), STAT TOOLS & LANGUAGES (Python, R), MLOPS & AI INFRA (MLflow, Kubeflow), AI FRAMEWORKS & LIBRARIES (PyTorch, TensorFlow), AI MODELS & ARCHITECTURES (OpenAI, Google), SEARCH (Elastic, Algolia), LOGGING & MONITORING (Datadog, New Relic), VISUALIZATION (Tableau, Power BI), COLLABORATION (Slack, Microsoft Teams).

DATA SOURCES & APIs: DATA MARKETPLACES & DISCOVERY (Klarity, Arcadia), FINANCIAL & MARKET DATA (Bloomberg, Thomson Reuters), AIR / SPACE / SEA (OpenSky, FlightAware), PEOPLE / ENTITIES (ZoomInfo, Dun & Bradstreet), LOCATION INTELLIGENCE (Foursquare, Mapbox), ESG (Sustainalytics, MSCI).

DATA & AI CONSULTING: QuantumBlack, BCG, Deloitte, IBM, Cambridge Consultants, LeewayHertz, Slalom, Brookline Data Co., Ternary, Tribe AI, Meru, Deepar Insights, Data Root Labs, India Labs, ThirdEye, Azati, Adepto, Upright Analytics, Bytecode IO, Leit Data.



# Typické požadavky na Big Data systémy



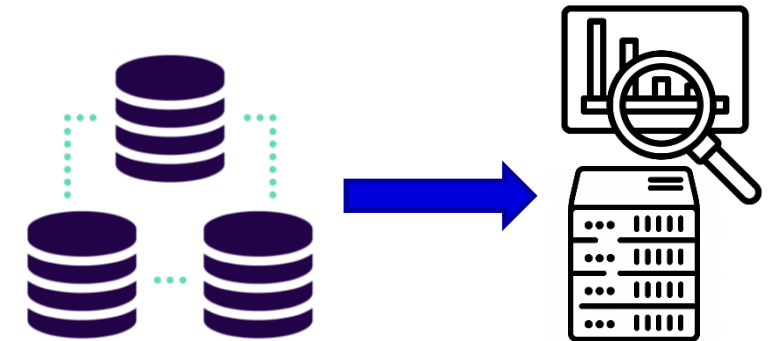
- ukládání velkého množství dat
- zpracování dat v „rozumném“ čase
  - zahrnuje nezbytnost „stěhování“ dat k výpočetním procesům
- škálovatelnost
  - = schopnost systému růst se zvětšujícím se množstvím dat
  - schopnost pojmout dodatečný hardware (rozšíření systému)
  - schopnost využívat tytéž struktury a algoritmy
  - nejčastěji hovoříme o tzv. **distribuovaných systémech**
    - výpočetní infrastruktura sestávající z více fyzických počítačů (výpočetních serverů)



# Škálovatelnost – je vždy nutné stěhování dat?



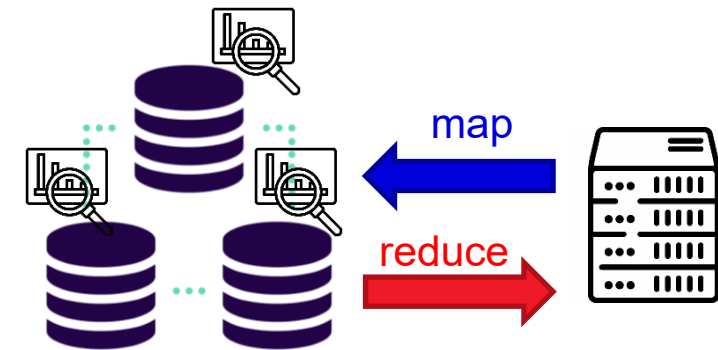
– pro komplexnější zpracování se data typicky přenášejí **od úložného systému k výpočetnímu procesu**



– v případě velkých objemů je toto (časově, datově) velmi náročné

– **alternativa: tzv. Map-Reduce přístup**

- přenos výpočtu k datům (fáze *Map*)
- vyhodnocení dílčích výsledků (fáze *Reduce*)
- vhodné jen pro specifické typy výpočtů  
např. analýzu textových/obrázkových korpusů
- technologie Apache Hadoop





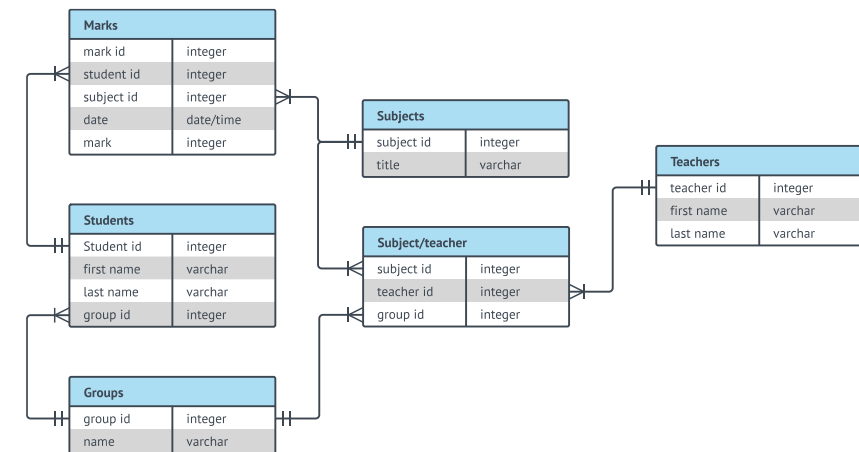
# Vybrané modely pro reprezentaci a analýzu STRUKTUROVANÝCH dat

## Tabulky

- tabulka ve vhodném tabulkovém procesoru  
např. Microsoft Excel, Google Sheets, LibreOffice Writer, ...
- dostupnost základní datové analytiky  
statistické funkce, grafy, atp.

## Relační (SQL) databáze

- schéma tabulek (relací) popsané SQL jazykem  
včetně vzájemných vazeb
- základní analytické funkce dostupné přímo v jazyce SQL  
pokročilé zpracování v návazné aplikaci
- např. PostgreSQL, MySQL, Sqlite, MS Access, ...
- **NewSQL přístup:** škálovatelné SQL databáze  
např. NuoDB, VoltDB, TokuDB, GenieDB



# Vybrané modely pro reprezentaci a analýzu SEMI-STRUKTUROVANÝCH dat (NoSQL)



## Key-value databáze

- ukládají data ve formě „klíč = hodnota“
  - např. „věk = 25“, „rok\_narození = 2011“
- klíč musí být jedinečný
  - hodnoty mohou být jednoduché i složené záznamy
- klíč může nést komplexnější informaci
  - `student:23757:jmeno = „Jan“`
  - `student:23757:prijmeni = „Novák“`
- hlavní výhodou je jednoduchost a rychlost
  - výborně škálují, vhodné pro masové operace



Amazon DynamoDB

Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

# Vybrané modely pro reprezentaci a analýzu SEMI-STRUKTUROVANÝCH dat (NoSQL)



## Dokumentové databáze

- hlavní úložnou jednotkou je **dokument**
  - seskupení „key:value“ hodnot popisujících uloženou entitu
  - klíče v různých dokumentech mohou být odlišné
- podporuje uložení komplexních informací k objektům
  - a jejich prohledávání + analýzu
- velmi rozšířené a hojně používané



elasticsearch

### Document 1

```
{
  "id": "1",
  "name": "John Smith",
  "isActive": true,
  "dob": "1964-30-08"
}
```

### Document 2

```
{
  "id": "2",
  "fullName": "Sarah Jones",
  "isActive": false,
  "dob": "2002-02-18"
}
```

### Document 3

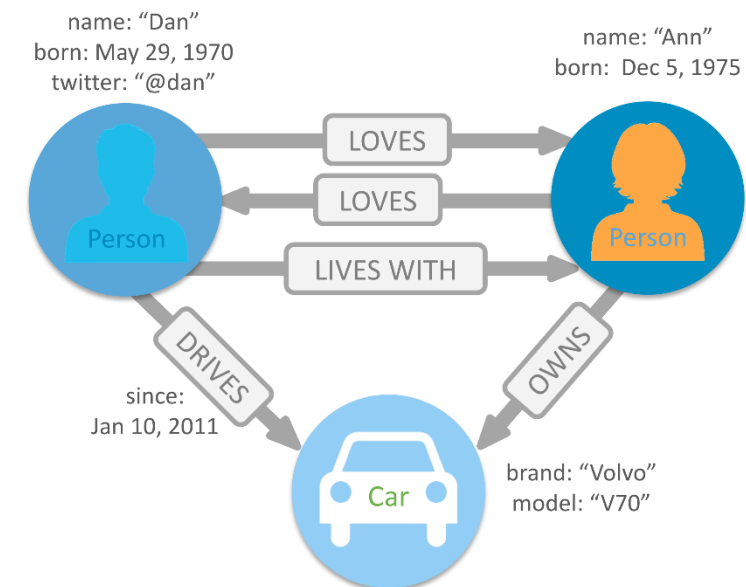
```
{
  "id": "3",
  "fullName": {
    "first": "Adam",
    "last": "Stark"
  },
  "isActive": true,
  "dob": "2015-04-19"
}
```

# Vybrané modely pro reprezentaci a analýzu SEMI-STRUKTUROVANÝCH dat (NoSQL)



## Grafové databáze

- reprezentace uložených dat formou (libovolně komplexního) grafu
  - uzly i hrany podporují uložení dalších metadat nejčastěji formou „key:value“
- extrémně rychlé pro vyhledávání lokálních („vztahových“) informací
  - např. „všichni známí mých přátel“  
rychlost těchto dotazů nezávisí na množství uložených dat viz sociální sítě
- nevhodné pro globální prohledávání
  - např. „průměrný věk všech uložených osob“





# Vybrané modely pro reprezentaci a analýzu SEMI-STRUKTUROVANÝCH dat (NoSQL)

## Existuje řada dalších přístupů

- řádkově-orientované a sloupcově-orientované databáze
- databáze pro uložení časových řad
- databáze pro uložení prostorových dat
- ...

## Vícemodelové databáze

- umožňují využití vícero různých modelů a vícero pohledů (forem dotazů) na tatáž data



## Kde s (velkými) daty pracovat?

aneb Výpočetní infrastruktury v ČR





# Superpočítačová centra

- vysoký hardwarový výkon pro náročné výpočty a zpracování dat
  - seskupení tzv. **výpočetních clusterů**
- specializované výpočetní přístupy
  - kompromis mezi uživatelskou přívětivostí a co nejefektivnějším využitím infrastruktury
    - nejefektivnější využití skrze gridové výpočty
- **akademické vs. komerční výpočty**
  - **pro akademické využití často zdarma**
    - financováno z veřejných zdrojů
  - **pro komerční využití za úplatu**
    - s výjimkou veřejných výzkumných projektů





# Slovníček pojmů – výpočetní cluster

– skupina vzájemně propojených „běžných“ počítačů



(dříve 😊)

# Slovníček pojmů – výpočetní cluster

– skupina vzájemně propojených „běžných“ počítačů



**(dnes)**



# Superpočítačová centra v ČR

– v ČR dostupná ve 3 infrastrukturách (centrech)

– Cesnet/MetaCentrum

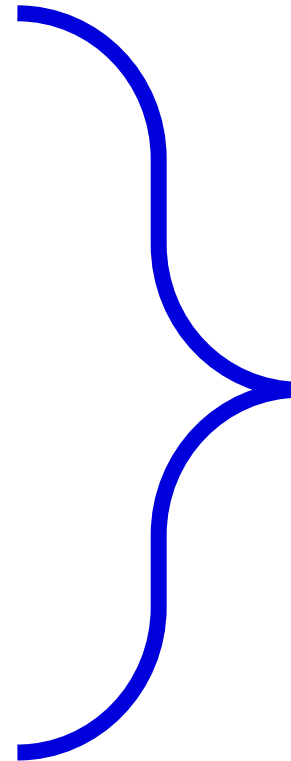
gridový přístup  
cloudový přístup  
specializované výpočty

– MUNI/CERIT-SC

gridový přístup  
cloudový přístup  
specializované výpočty

– VŠB-TUO/IT4Innovations

gridový přístup



**e-INFRA CZ**  
<https://www.e-infra.cz>



# MetaCentrum @ CESNET

- aktivita sdružení CESNET
  - CESNET – sdružení založené (a podporované) českými vysokými školami poskytuje služby vysokým školám + vlastní výzkum
- od roku 1996 **koordinátor Národní Gridové Infrastruktury (NGI)**
  - původně vzniklo na MUNI (Superpočítačové Centrum Brno, SCB, 1994)
- integruje velká/střední HW centra (clustery, výkonné servery a úložiště) několika univerzit/organizací v rámci ČR
  - poskytuje prostředí pro (spolu)práci v oblasti **výpočtů a práce s daty**
- integrováno do **evropské gridové infrastruktury (EGI)**





# MetaCentrum NGI

- přístupné zaměstnancům a studentům VŠ/univerzit, AV ČR, výzkumným ústavům, atp.
- komerční subjekty pouze pro veřejný výzkum
- nabízí:
  - výpočetní zdroje
  - úložné kapacity
  - aplikační programy
- po registraci **k dispozici zcela zdarma**
  - „placení“ formou publikací s poděkováním

<http://metavo.metacentrum.cz>





# NGI – dostupný výpočetní hardware

- výpočetní zdroje: **cca 37600 jader (x86\_64)**
  - uzly s nižším počtem výkonných jader:
    - 2x4-8 jader
  - uzly se středním počtem jader (SMP stroje):
    - 32-80 jader
  - paměť až 10 TB na uzel
- uzly s vysokým počtem jader: SGI UV 2000
  - 504 jader (x86\_64), 10 TB operační paměti
  - 384 jader (x86\_64), 6 TB operační paměti
- další „exotický“ hardware:
  - uzly s GPU kartami (pro AI), Xeon Phi, SSD disky, ...





# NGI – dostupný úložný hardware

- cca **15 PB** pro pracovní data
  - úložiště v Brně, Plzni, ČB, Liberci, Praze
  - uživatelská kvóta 1-3 TB na každém z úložišť
- cca **80+ PB** pro dlouhodobá/archivní data
  - HSM – páskové knihovny
  - objektové uložení CEPH (analogie k Amazon S3)

<http://metavo.metacentrum.cz/cs/state/nodes>



# NGI – dostupný software

- ~ **450 různých aplikací** (instalováno na požádání)
  - viz <http://meta.cesnet.cz/wiki/Kategorie:Aplikace>
- průběžně udržované **vývojové prostředí**
  - GNU, Intel, PGI, ladící a optimalizační nástroje (TotalView, Allinea), ...
- generický **matematický software**
  - Matlab, Maple, Mathematica, gridMathematica, ...
- komerční i volný software pro **aplikační chemii**
  - Gaussian 09, Gaussian-Linda, Gamess, Gromacs, Amber, ...
- **materiálové simulace**
  - ANSYS Fluent CFD, Ansys Mechanical, Ansys HPC...
- **strukturní biologie, bioinformatika**
  - CLC Genomics Workbench, Geneious, Turbomole, Molpro, ...
- **řada volně dostupných balíčků**
- ...



# NGI – jak počítat?

## – dávkové úlohy

- popisný skript úlohy
- oznámení startu a ukončení úlohy

## – interaktivní úlohy

- textový i grafický režim

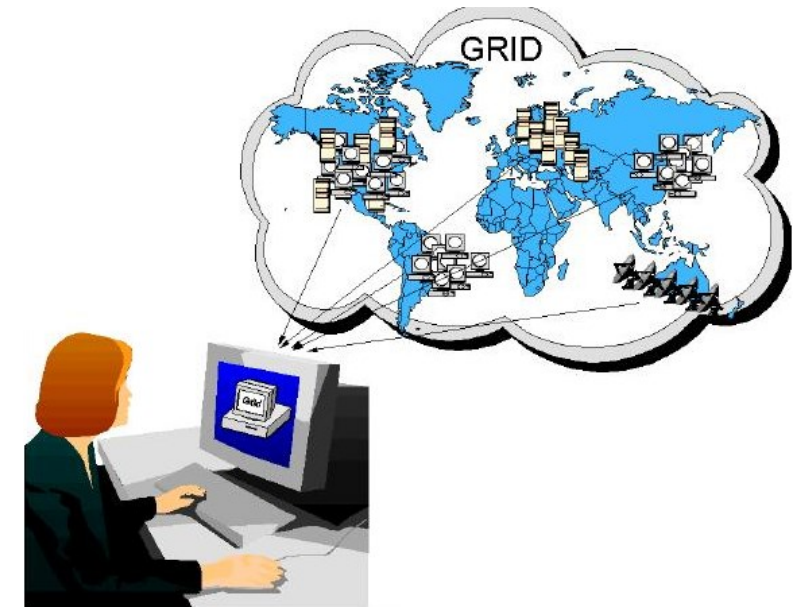
## – cloudové rozhraní

- uživatelé nespouští úlohy, ale virtuální stroje  
pouze pro vědecké výpočty

## – grafické aplikace a virtuální desktopy v prostředí prohlížeče

## – specializovaná prostředí

- Apache Hadoop, Galaxy, ...





# Meta VO – jak se stát uživatelem?

## – **podejte si přihlášku**

- <http://metavo.metacentrum.cz> , sekce „Přihláška“
- EduID.cz => ověření Vaší akademické identity proběhne s využitím Vaší domovské instituce

## – seznamte se s **dokumentací a základy OS Linux**

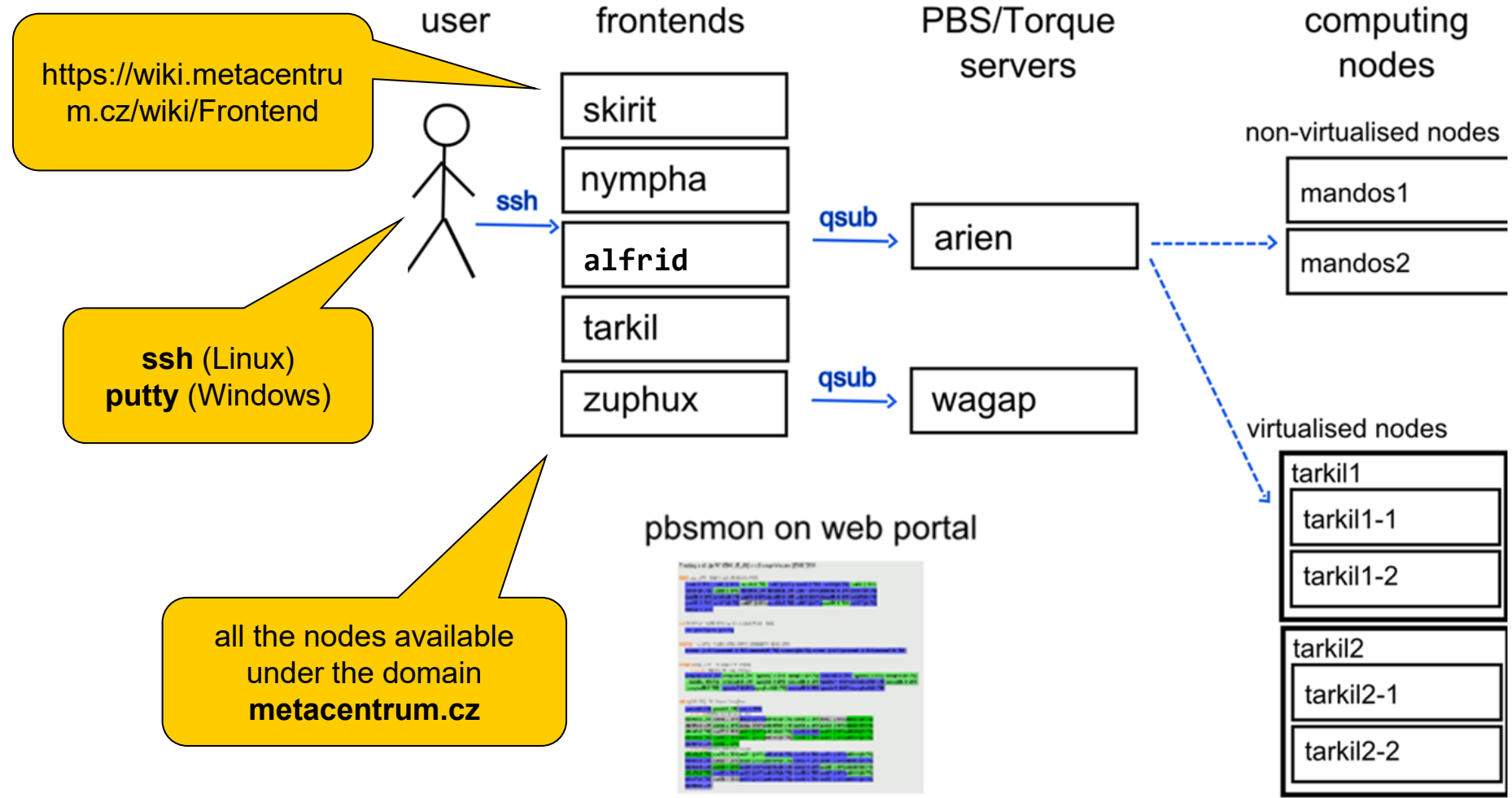
- <http://metavo.metacentrum.cz> , sekce „Dokumentace“
- praktická školení: <https://metavo.metacentrum.cz/cs/seminars/index.html>
- <https://www.abclinuxu.cz/ucebnice/zaklady>

## – **počítejte**

- netřeba oficiálních žádostí o výpočetní čas



# NGI pod pokličkou

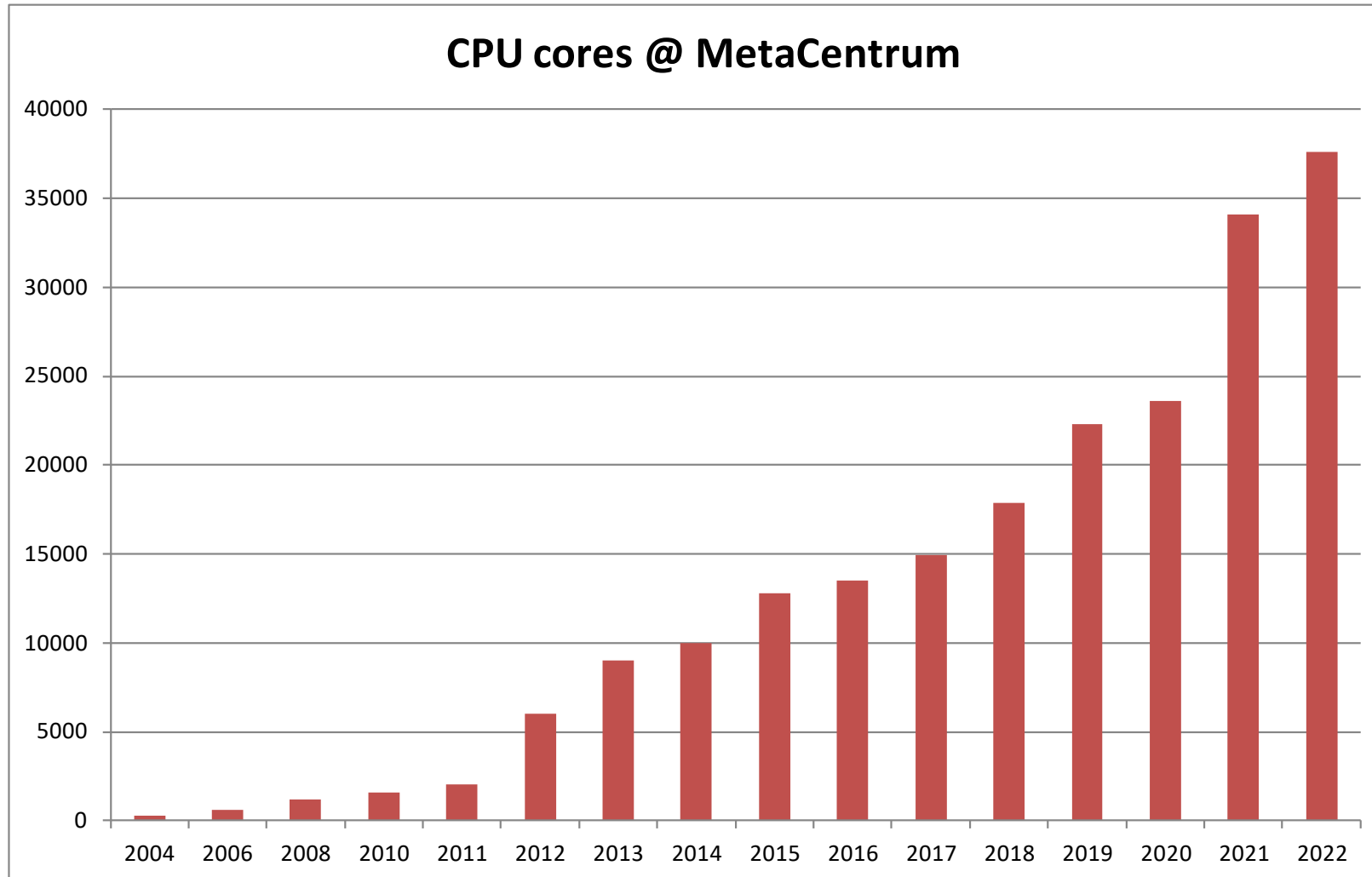




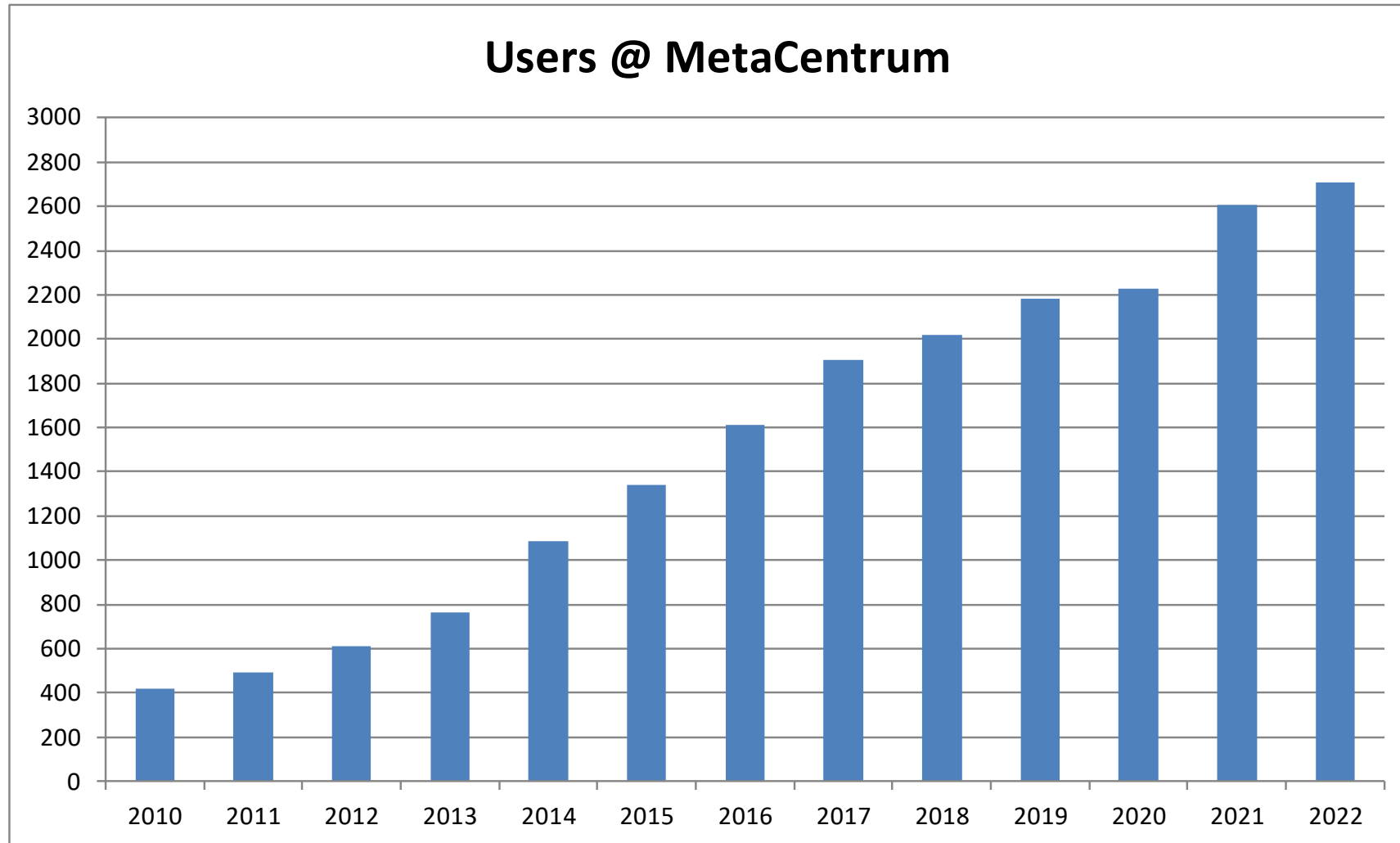
# NGI pod pokličkou – v číslech...

- cca 37600 výpočetních jader, cca 700 uzlů
  - a 455 GPU karet
- za rok 2022:
  - 2710 uživatelů (k 31.12.2022)
  - cca 11 mil. spuštěných úloh
    - cca 30500 úloh denně
    - cca 4100 úloh / uživatel
  - celkem propočítáno cca 27,5 tis. CPUlet
    - a 314 GPUlet

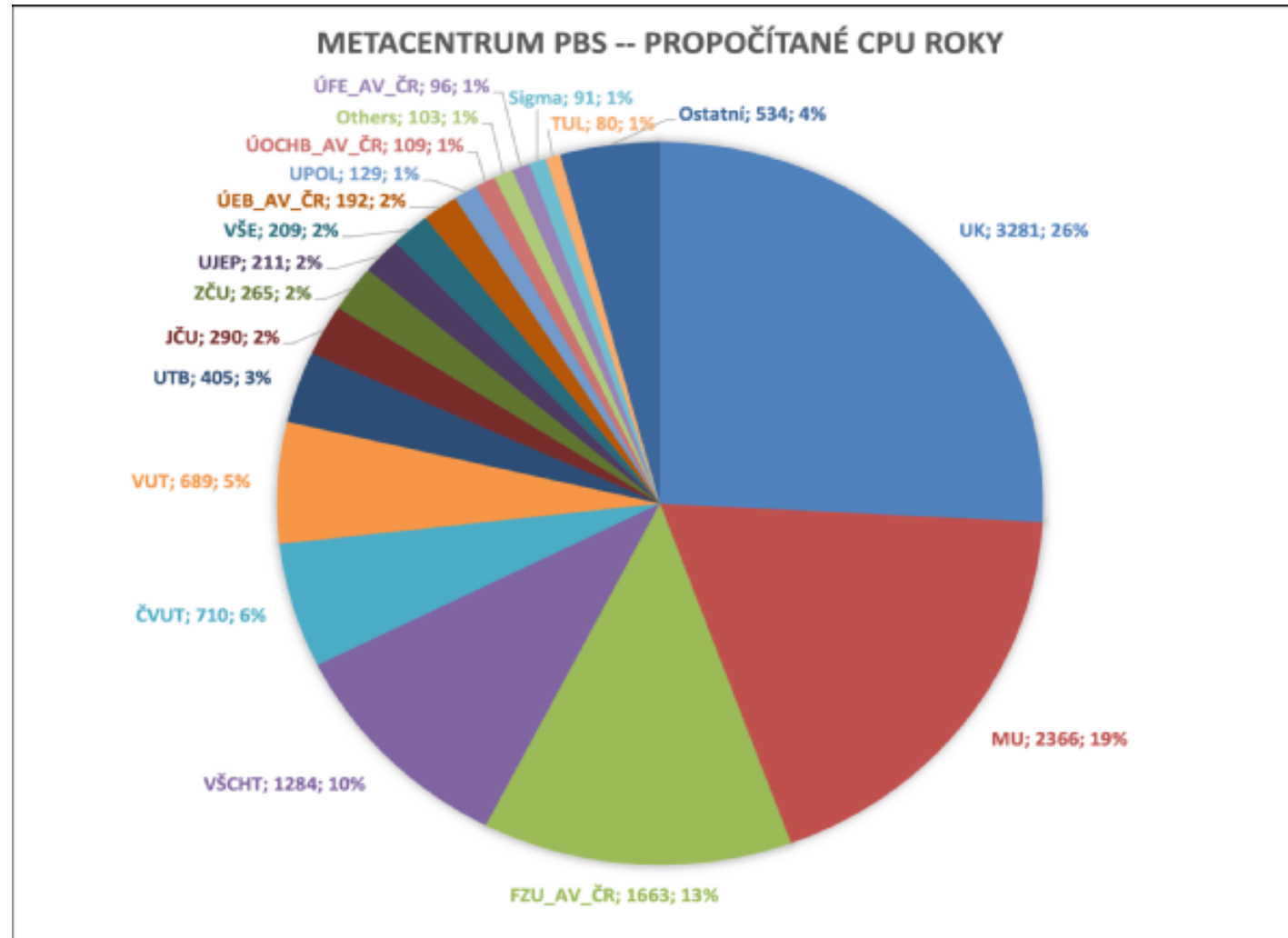
# NGI pod pokličkou – a grafech...



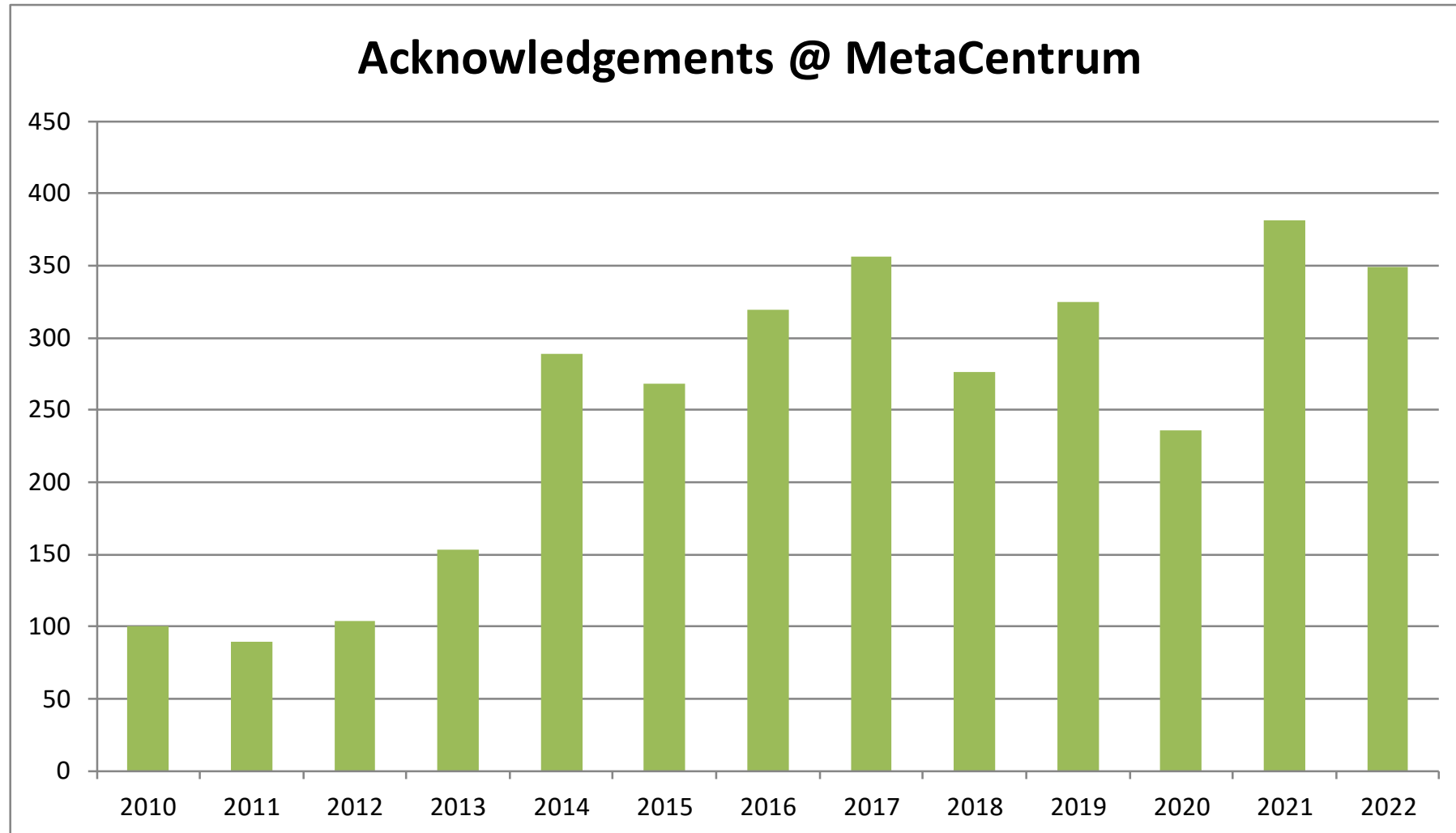
# NGI pod pokličkou – a grafech...



# NGI pod pokličkou – a grafech...



# NGI pod pokličkou – a grafech...







# Centrum CERIT-SC

- **Centrum CERIT-SC** – výzkumné centrum vybudované na ÚVT MU
  - původně Superpočítačové centrum Brno (SCB)
- poskytovatel HW a SW zdrojů
  - **součást MetaCentrum NGI**
- služby nad rámec „běžného“ HW centra
  - **mezioborový (interdisciplinární) výzkum**  
spolupráce IT výzkumníků a partnerů z jiných oborů



# Centrum CERIT-SC



- hlavní cíle Centra CERIT-SC @ MUNI:
  - flexibilní infrastruktura, vlastní výzkum v infrastrukturních oblastech
  - tři hlavní **výzkumné směry**:
    - High-performance computing* – akcelerace výpočtů, GPU computing, ...
    - Artificial Intelligence* – aplikace metod umělé inteligence a strojového učení
    - Big Data analytics***
- 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

# IT4Innovations



- **IT4Innovations** – superpočítačové centrum při VŠB TUO v Ostravě
  - aktuálně dostupné superpočítače:  
Karolina, Barbora, NVIDIA DGX-2
- služby dostupné **akademickým pracovníkům i komerčním subjektům**
- jak HW centrum, tak **výzkumné služby**
  - vlastní výzkumné laboratoře
  - výzkumné spolupráce s uživateli centra
- o výpočetní čas **nutno oficiálně žádat**
  - tzv. *grantové soutěže* (každých 6 měsíců)  
následně dedikovaný výpočetní čas  
vhodná finanční participace



# Datové služby e-INFRA CZ pro koncové uživatele I.



- **FileSender** – webová služba pro zasílání velkých souborů
- aktuální limit je 2 TB (~ 2000 GB)
  - doba expirace až 1 měsíc
- <http://filesender.cesnet.cz>
- odesílatel nebo příjemce musí být autorizovaným akademickým pracovníkem
  - autorizovaný uživatel **může odesílat datové soubory** libovolnému uživateli
    - emailové notifikace o životním cyklu dat
  - autorizovaný uživatel **může odeslat pozvánku pro příjem datových souborů** od libovolného uživatele



# FileSender – ukázka využití



Preferovaný jazyk CS

Nápověda

Přihlášení

## Vítejte na Filesenderu Filesender.Cesnet.cz

Filesender.Cesnet.cz je bezpečný způsob sdílení velkých souborů s kýmkoliv! Přihlašte se k nahrání svých souborů nebo pozvání ostatních k zaslání souboru.

Přihlášením potvrzujete, že jste byl/a seznámen/a s podmínkami služby a s informacemi o zpracování osobních údajů.

Přihlášení

# FileSender – ukázka využití




Přihlásit účtem

Masarykova univerzita **MUNI**

Jiný účet

🇨🇪 🇩🇪 🇫🇮 🇬🇧 🇪🇸 🇫🇷 🇮🇹 🇭🇺 🇸🇪

CESNET ⓘ



**MUNI** Jednotné přihlášení

English 🇺🇦 Українська

UČO


39685

Primární heslo

••••••••

Zapamatovat si mě

**PŘIHLÁSIT**



> [Mám problém s přihlášením](#)

# FileSender – ukázka využití



pozvánky

The screenshot shows the FileSender web interface. At the top, there's a navigation bar with 'Nahrávání', 'Pozvánky', 'Mé přenosy', 'Můj profil', 'Nápověda', 'Soukromí', and 'Odhlášení'. The 'Pozvánky' tab is active. Below the navigation bar, there's a large dashed box with the text 'Sem přetáhněte soubory k nahrání'. Below this, there are two buttons: 'Vyčistit vše' and 'Vybrat soubory'. A red arrow points to the 'Vybrat soubory' button. Below the buttons, there's a form with the following fields: 'Od : rebok@ics.muni.cz', a checkbox for 'Zašifrování souborů (beta)', 'Datum expirace: 02/10/2022', a checkbox for 'Zaslat mi denní statistiku', and a link for 'Pokročilá nastavení'. A green arrow points to the 'Pokročilá nastavení' link. Below the form, there's a large 'Odeslat' button with a cloud icon. At the bottom, there's a line graph titled 'Globální průměrná rychlost nahrávání souborů 1 GB'. The graph shows upload speed in MB/s over time from August 23 to September 22. The legend indicates two series: 'Šifrování při přenosu & rest' (green) and 'Šifrování při přenosu' (orange).

pokročilé notifikace, získání odkazu, atp.

# Datové služby e-INFRA CZ pro koncové uživatele II.



- **OwnCloud** – cloudové uložení a-la Google Drive nebo Dropbox
  - aktuální kvóta je 100 GB / uživatel
- <https://owncloud.cesnet.cz/>
- **synchronizace a dostupnost dat mezi zařízeními**
  - klienti dostupní pro OS Windows, Linux, OS X
  - také pro chytré telefony a tablety
  - umožňuje sdílení dat mezi uživateli
  - poskytuje zálohování
  - atp.





# OwnCloud – ukázka využití



The screenshot shows the login page for OwnCloud @ CESNET. The page features a blue background with a 3D bar chart graphic. The main heading is "ownCloud @ CESNET" with the tagline "Sync, Share & Backup all of your academic data." Below this is a "PŘIHLÁSIT SE" button and a link for login. The top navigation bar includes "Uživatelská dokumentace", "FAQ", "Kontakt", and "Přihlásit se". A red arrow points to the "Přihlásit se" link.

# OwnCloud – ukázka využití




Přihlásit účtem

Masarykova univerzita **MUNI**

Jiný účet

Settings icons: Czech Republic, Germany, Finland, United Kingdom, Spain, France, Italy, Hungary, Sweden. CESNET



MUNI Jednotné přihlášení

English Українська

UČO


39685

Primární heslo

.....

Zapamatovat si mě

**PŘIHLÁSIT**



> [Mám problém s přihlášením](#)

# OwnCloud – ukázka využití



Soubory ownCloud@CESNET DataCare RNDr. Tomáš Rebok Ph.D.

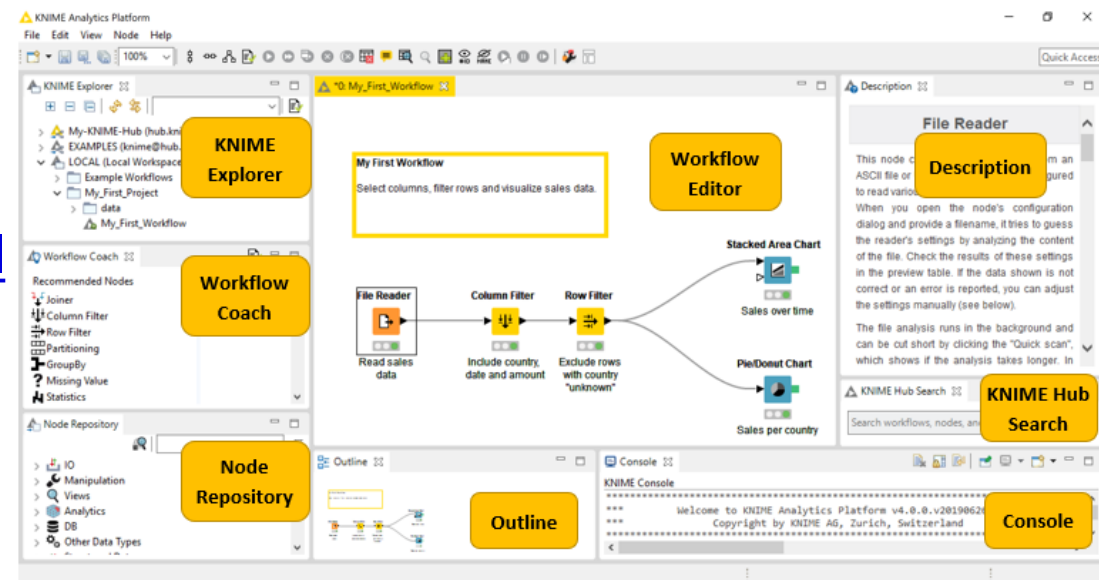
Všechny soubory > Shared > +

<input type="checkbox"/>	Název ▲		Velikost	Upraveno
<input checked="" type="checkbox"/>	MetaCentrum	Mgr. Miroslav Ruda	4.7 MB	před 5 měsíci
<input checked="" type="checkbox"/>	prezentace-tabor	Jan Růžička	281 KB	před 3 měsíci
	2 adresáře		5 MB	

# TIP: Nástroj vizuální datové analýzy



- **KNIME** – open-source nástroj vizuální datové analýzy (a zpracování)
  - vizuálně přehledná datová analytika, mnoho integrovaných funkcí a možností datová analýza formou workflow
  - rozšiřitelné moduly včetně vlastních funkcí (Python)
- <https://www.knime.com/>
- **desktopová aplikace**
  - dostupná zdarma (server za poplatek)
  - pro běžné operační systémy
- **dostupnost i v rámci e-INFRA CZ**
  - <http://docs.cerit.io/docs/rancher-applications.html> na požádání vypomůžeme
  - v budoucnu přes vyvíjený **CloudApp Store**



MUNI

# Ilustrativní příklady životního cyklu dat v rámci výzkumných spoluprací ÚVT MU

## Analýza dat stavu krajiny

aneb Výzkumná spolupráce ÚVT s partnerem CzechGlobe



# Ústav výzkumu globální změny Akademie věd ČR

- alias **CzechGlobe**
- veřejná výzkumná instituce, **evropské centrum excelence**
- dlouhodobý výzkum probíhající **globální změny, jejich projevů v atmosféře a dopadů na biosféru a lidskou společnost**
  - atmosféra – ekosystém – socio-ekonomický systém
- hlavní zdroje dat:
  - atmosférické stanice – monitoring skleníkových plynů
  - ekosystémové stanice (v ČR i zahraničí) – toky uhlíku v základních typech ekosystémů
  - růstové komory
  - letecká laboratoř
  - laboratoře
  - atp.

# Ústav výzkumu globální změny Akademie věd ČR





# Plánování sběru a shromažďování dat

## Planování sběrů dat

- pravidelný sběr dat
  - zahrnuje mj. plánování lokalit ekosystémových a atmosférických stanic
  - nejstarší záznamy z roku 1996
- nepravidelný sběr dat
  - plánované „kampaně“ – např. nálety vybraných ekosystémů leteckou laboratoří



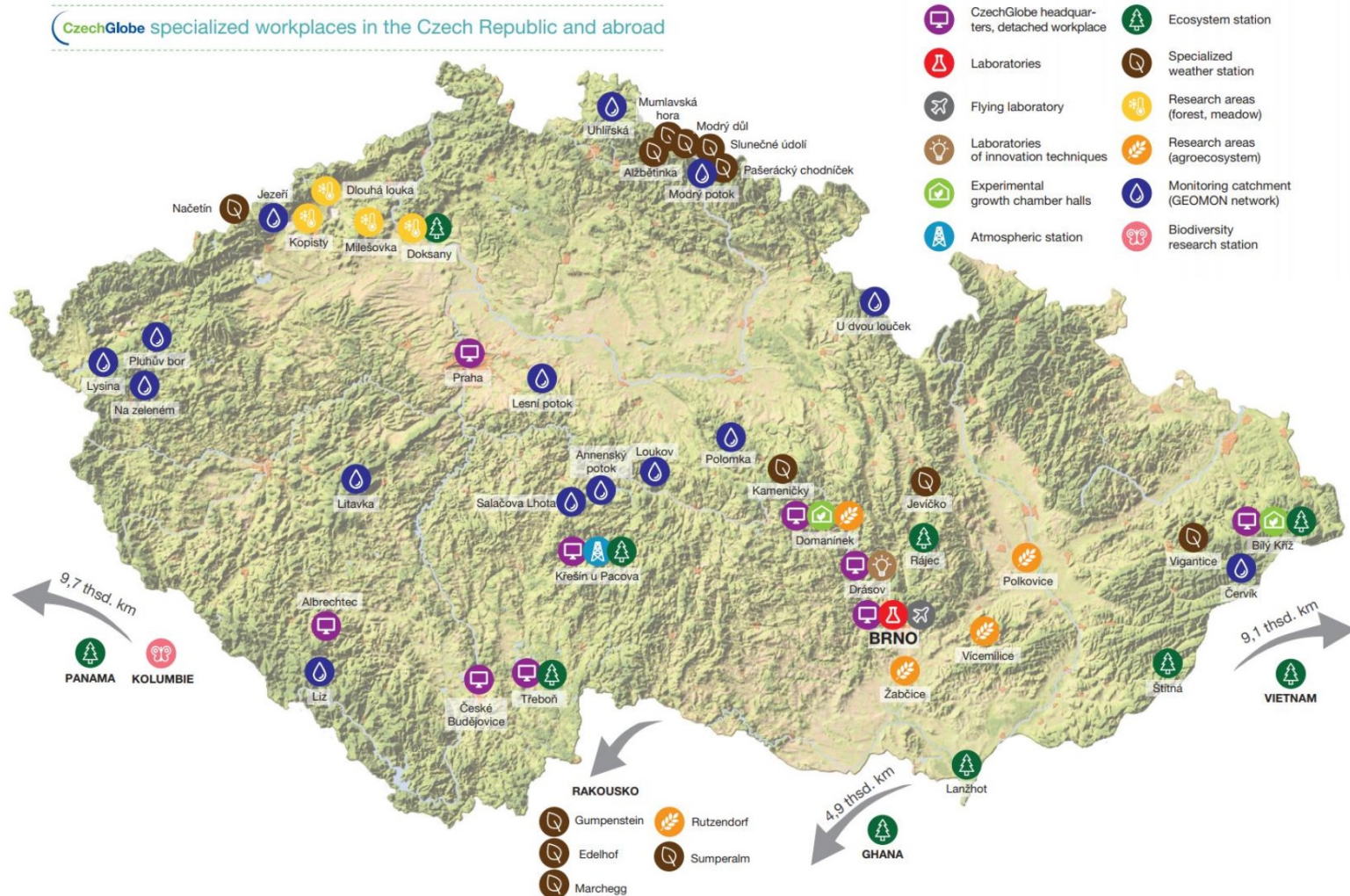
## Shromažďování dat

- online sběr z měřících ekosystémových stanic
  - každých cca 10 minut desítky parametrů, zasíláno do datových center
- datové nosiče – ad-hoc sběr

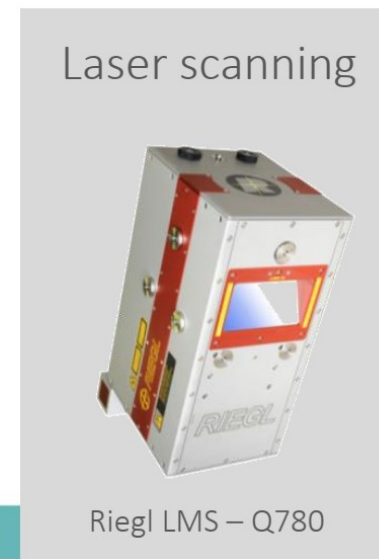
# Sběr dat – pozemní měřicí stanice



CzechGlobe specialized workplaces in the Czech Republic and abroad



# Sběr dat – letecká laboratoř pro dálkový průzkum



# Sběr dat – typy dat dálkového průzkumu Země



*Imaging spectroscopy*



*Laser scanning*



*Thermal scanning*



# Zpracování dat



- data **velkých i malých objemů**

- data ze senzorů měřících věží vs. satelitní/letecká data

- příklady úpravy a čištění dat

- detekce chyb v datech měřících stanic

- proces odhalování chybějících či nesmyslných hodnot (častá chybovost senzorů)

- ne vždy snadno odhalitelné chyby

- nefunkční senzor vs. chybující senzor vs. zakrytý senzor

- hodnocení dat indikátorem kvality

- prostor pro uplatnění metod strojového učení a umělé inteligence

- např. M. Moravčík: *Použití neuronových sítí pro doplňování chybějících dat meteorologických měření*. DP

- 2017, vedoucí Rebok, <https://is.muni.cz/th/d09hs/>

- zarovnávání leteckých snímků

- eliminace pohybů letadla vůči Zemi

# Zpracování dat

Doplňování chybějících hodnot s využitím neuronových sítí (M. Moravčík)

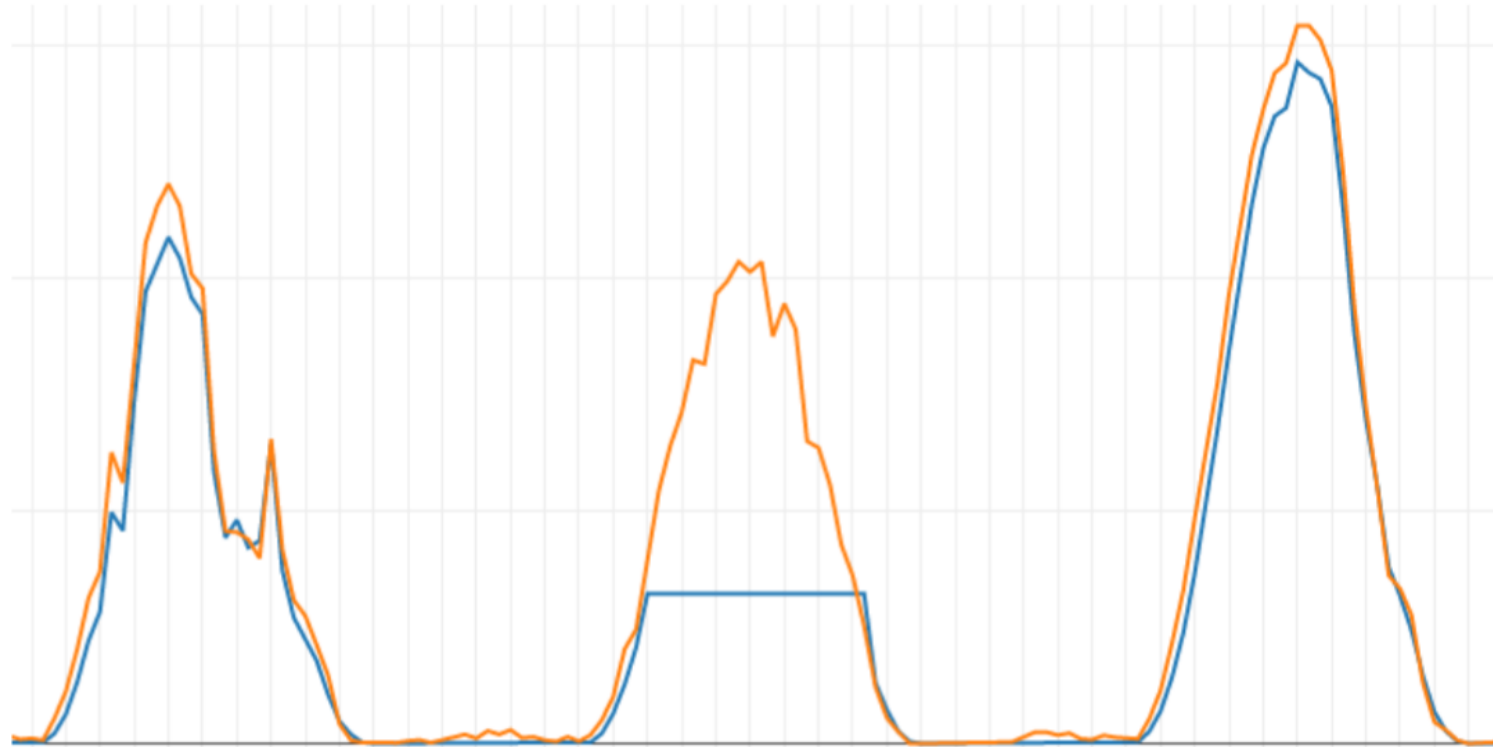


Figure 5.1: Correction of UVB measurement (update interval 30 min). Blue line shows input data, yellow line is gap-filling result.

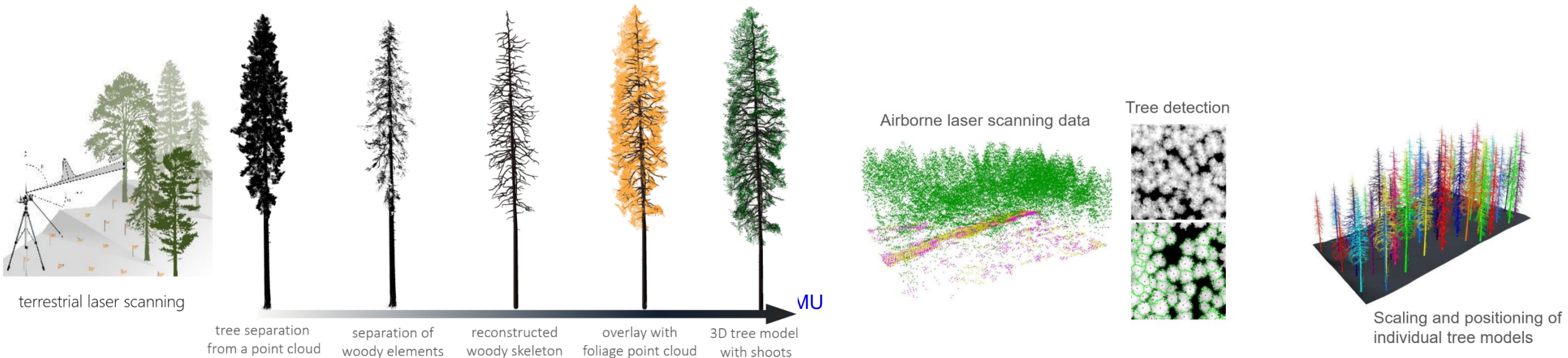
# Analýza dat – příklady

realizované ve spolupráci CzechGlobe a ÚVT MU



## Rekonstrukce 3D modelů stromů a lesů

- vstupem mrak bodů z laserového skenu (LiDAR)
  - pozemní (individuální stromy) a letecký (les)
- výstupem 3D struktura (model) stromu / lesa
  - výstupy jsou vstupem pro návazné výzkumné aplikace





# Analýza dat – příklady

realizované ve spolupráci CzechGlobe a ÚVT MU

## Vytváření bezoblačných mozaik z družicových dat

- v definovaném časovém rozsahu a prostoru
  - omezení na sledované vegetační období
- vstupem jsou data z družice Sentinel-2
- více metod:
  - per-pixel
  - per-dlaždice
- výstup je vstupem pro návaznou analýzu





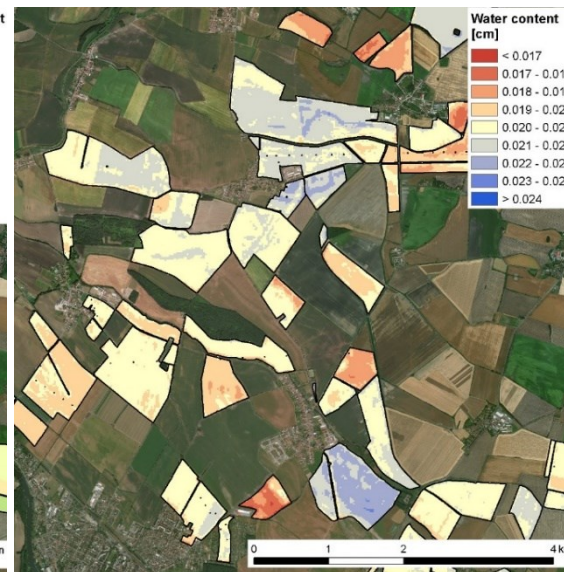


# Analýza dat – příklady

realizované ve spolupráci CzechGlobe a ÚVT MU

## Odhadování vegetačních parametrů zemědělských plodin

- např. obsah chlorofylu, vody, index listové plochy
- vstupem jsou bezoblačné mozaiky družicových snímků nebo snímků z letadla
- per-pixel analýza:
  - porovnávání vůči spektrální databázi



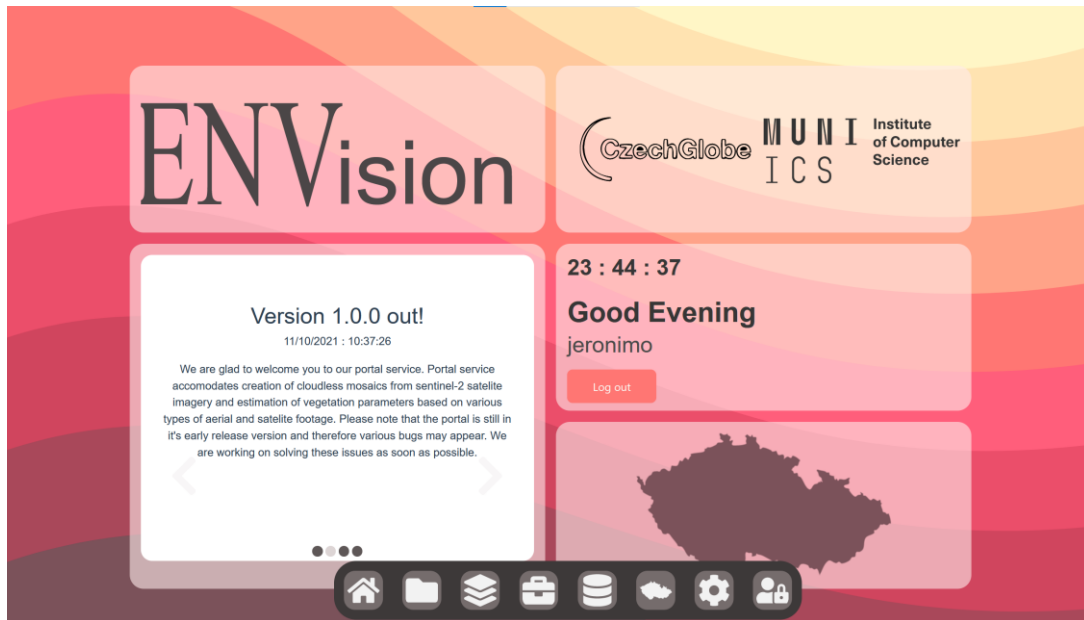
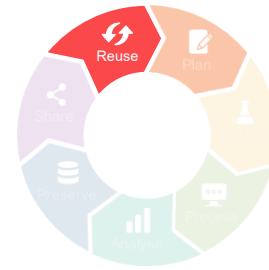
# Sdílení a prezentace dat

realizované ve spolupráci CzechGlobe a ÚVT MU



## Platforma ENVision (<https://envision.cerit-sc.cz>)

- vytvořený portál pro sdílení a analýzu ekosystémových dat ČR
- existují i nadnárodní portály: Google Earth Engine, Sentinel-Hub, atp.



## Analýza dat kriminálních činů

aneb Aplikačně-výzkumná spolupráce ÚVT s Policií ČR



# Policie České republiky

- netřeba blíže představovat 😊
- obrovské objemy různorodých dat
- výrazná variabilita hledaných informací
- výrazná specifika proti standardním přístupům k analýze dat



## Plánování

- ad-hoc
- vlastní proces sběru dat precizně plánovaná činnost

# Policie České republiky

## Sběr dat

- musí podléhat předchozímu schválení (soudní příkazy)
- velký důraz na transparentnost a precizní popis průběhu sběru
  - prokazatelnost korektního zajištění dat



## Zpracování dat

- opět důraz na transparentnost a průkaznost postupů
- minimální filtrace dat



# Policie České republiky

## Analýza dat

- hledané informace (často) předem neznámé
  - vyžaduje iterativní (a ideálně i interaktivní) prohledávání
  - „hledání jehly v kupce sena“
- vyžaduje budování tzv. „situačního povědomí“
  - *tradiční přístup*: využití izolovaných aplikací
    - iterativní analýza dat s využitím izolovaných specializovaných aplikací
    - budování situačního povědomí „v hlavě“ datového analytika (s využitím podpurných aplikací)
  - *moderní přístup*: využití pokročilých distribuovaných systémů
    - všechna data „na jedné hromadě“
    - analýzy dat napříč různorodými datovými sadami (např. hledání organizovaných skupin)
    - podpora budování situačního povědomí přímo v systému



# Policie České republiky

## Konzervace, udržování dat

- dlouhodobé uchovávání nemá význam, spíše se neuplatňuje



## Sdílení dat

- velmi precizně kontrolovaný přístup k datům, vč. jejich přenosů
- mnohdy nesdíleno ani mezi kolegy



## Znovuvyužití dat

- většinou se neuplatňuje
  - data zajištěná pro účely případu A nelze využít v případě B
  - nanejvýš pro „studijní“ či rozvojové potřeby

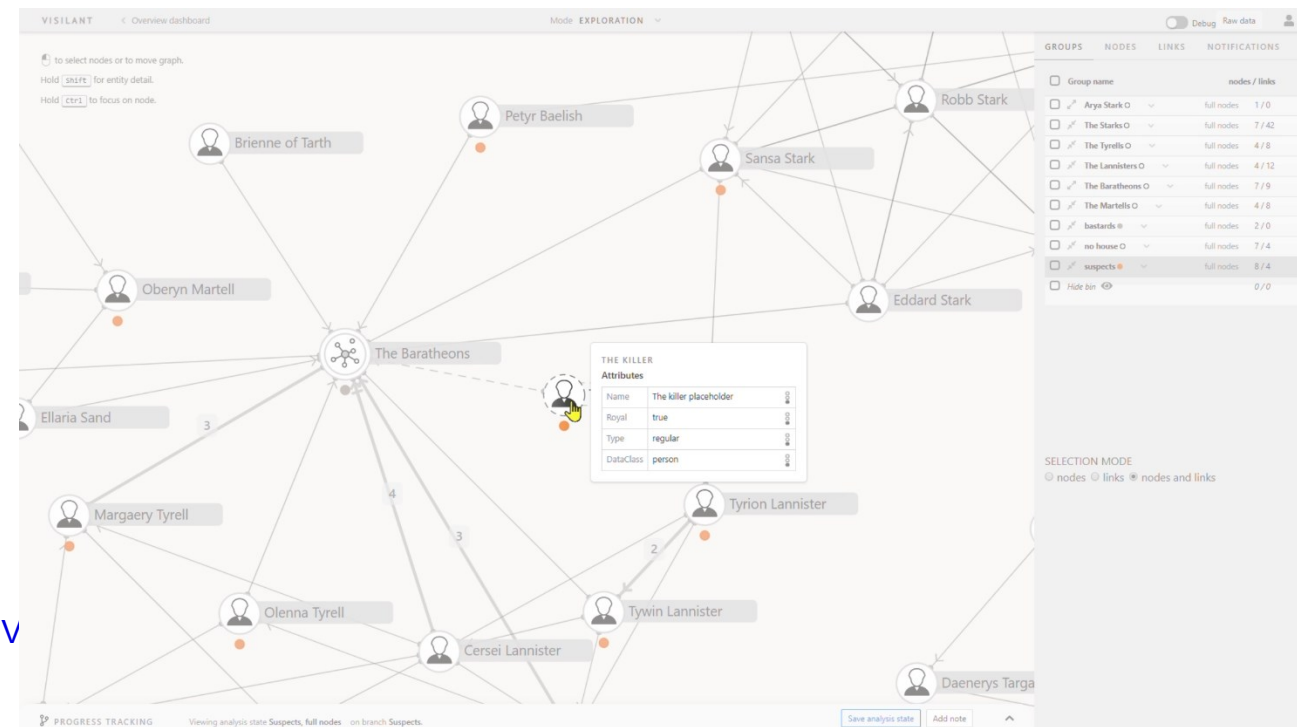


# Platforma ANALYZA

Realizovaná ÚVT MU prostřednictvím projektu pro potřeby Policie ČR

## Platforma ANALYZA

- projekt realizovaný ÚVT MU s podporou Ministerstva vnitra ČR (2017–2020)
- *Cíl projektu:* vyvinout distribuovaný systém podporující komplexní analýzy heterogenních dat velkého rozsahu
  - podpora budování situačního povědomí v jednotném systému
  - analýzy a vizualizace komplexních vztahů
- demonstrace možností nového přístupu
  - od 1.1.2023 řešíme **nový projekt AFoLab**  
*Automatizovaná forenzní laboratoř digitálních dat pro odhalování komplexní trestné činnosti*





MUNI

# Shrnutí

2. přednáška CORE042

# Shrnutí

## Zpracování a analýza dat

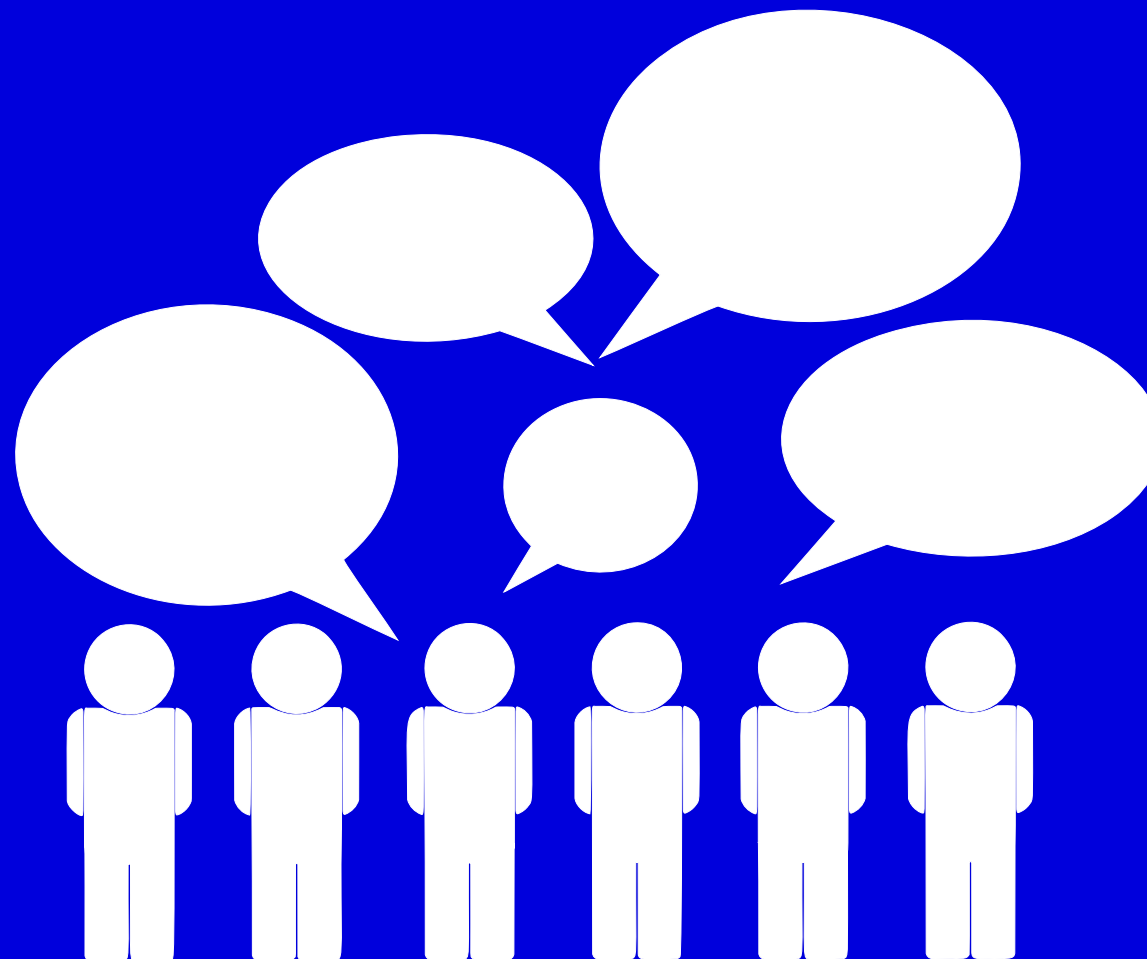
- jak se na data dívat?
  - strukturovaná vs. nestrukturovaná vs. semi-strukturovaná
  - výběr vhodného modelu pro zpracování a analýzu
    - důležitá je i znalost předpokládaných dotazů
      - tabulkové procesory, SQL databáze, NewSQL databáze, NoSQL databáze
- nebojte se být Big(Data) 😊

## Výpočetní a úložné infrastruktury v ČR

- dostupné prostřednictvím **e-INFRA CZ**
  - CESNET, CERIT-SC, IT4I
  - výpočetní a úložné kapacity pro náročné zpracování
    - akademikům dostupné zdarma
- doplňkové služby pro podporu datového zpracování a analýzy

# MUNI

## Diskuze



Zdroj: [Communicate\\_communication\\_conference\\_2028004](#) od [OpenClipart-Vectors](#) z [Pixabay](#)