M U N I

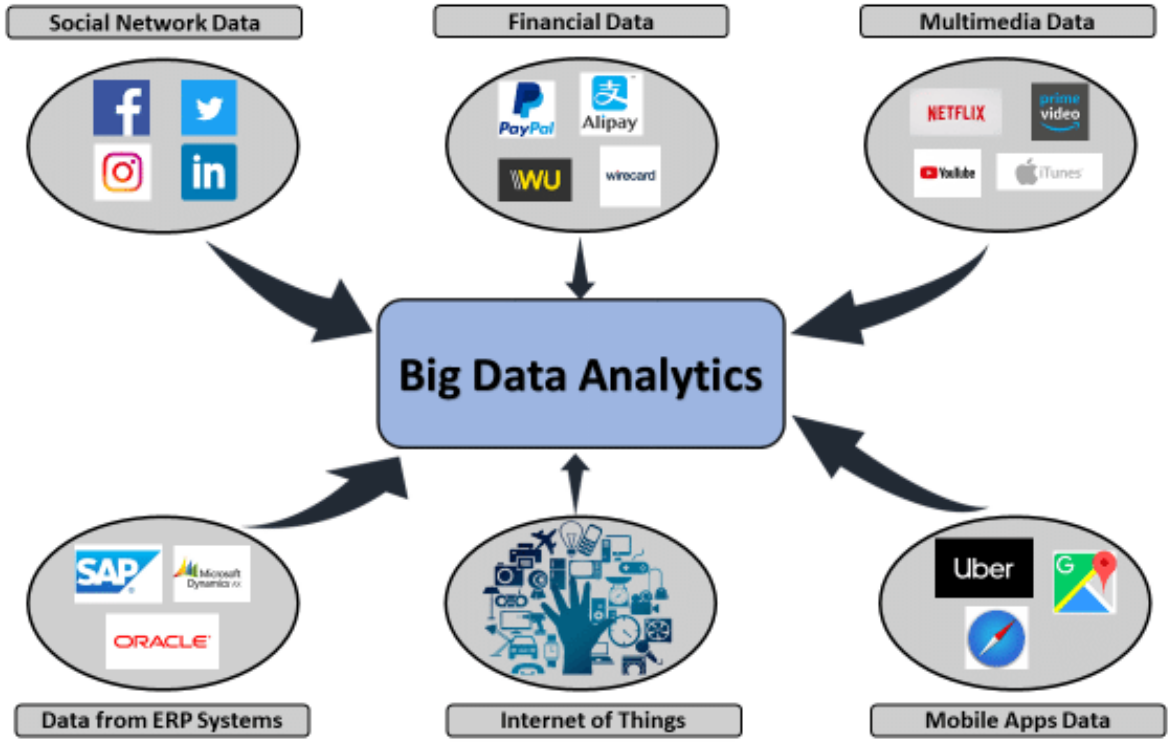
Zpracování a analýza (velkých) dat

Tomáš Rebok <rebok@ics.muni.cz>

Data – fenomén dnešní doby

- potenciálním zdrojem dat je prakticky cokoli (a kdokoli)
- vhodné vytěžování dat může odpovědět na mnoho otázek
  - ovlivňující jak business, 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



Základní členění dat

1. Strukturovaná data

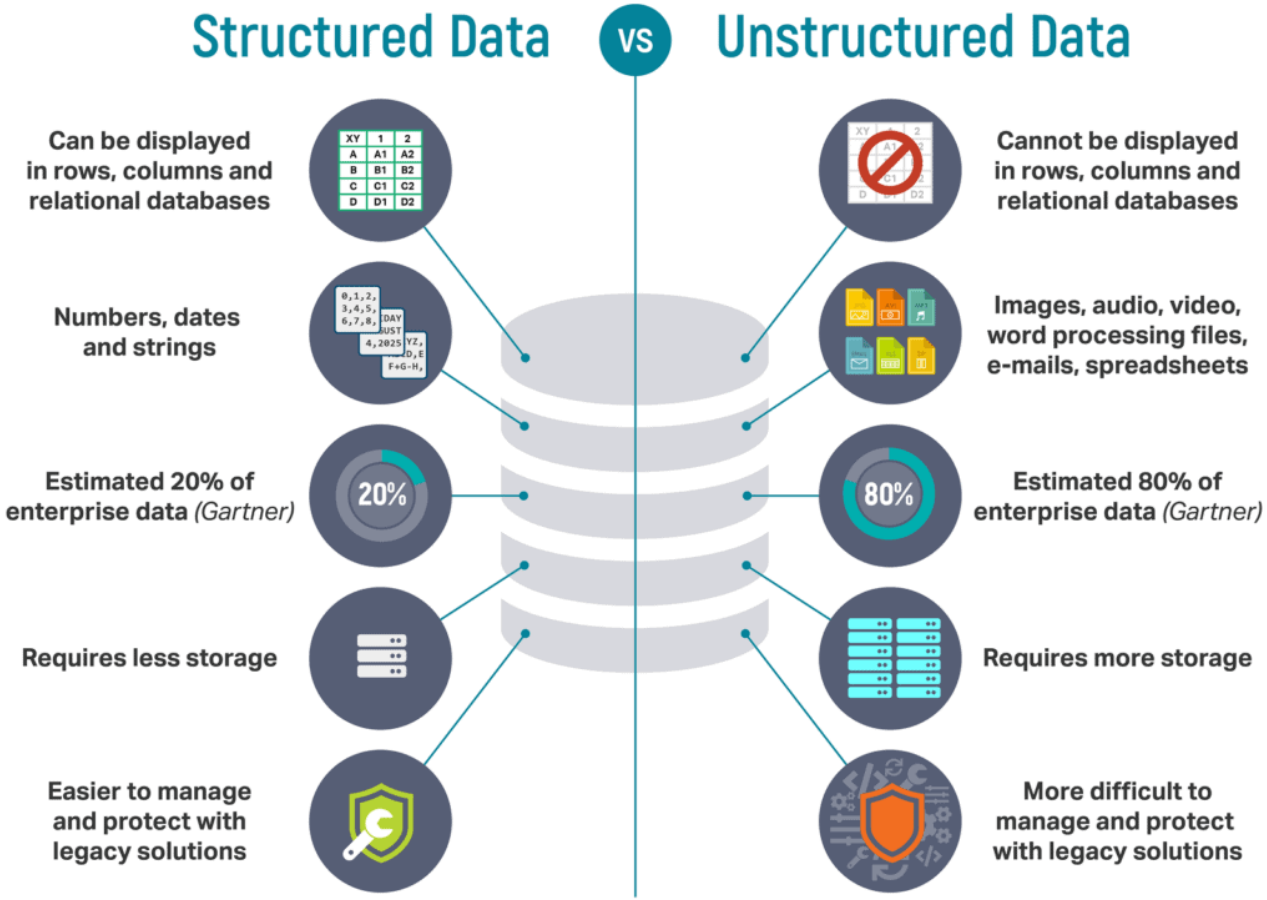
- data s identifikovatelnou strukturou
- typicky reprezentovaná tabulkami

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*

DATASet | OrderDate

	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

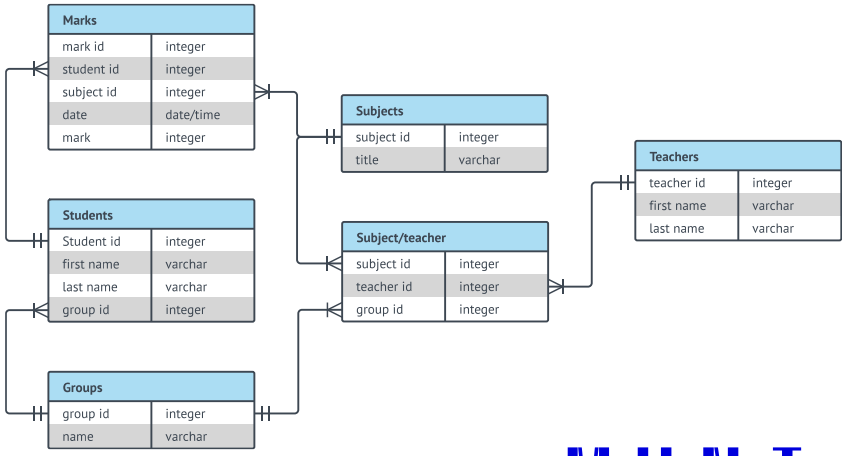
Vybrané modely pro reprezentaci a analýzu STRUKTUROVANÝCH dat

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



MUNI

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**

MUNI

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:jméno = „Jan“`
  - `student:23757:příjmění = „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"
}
```

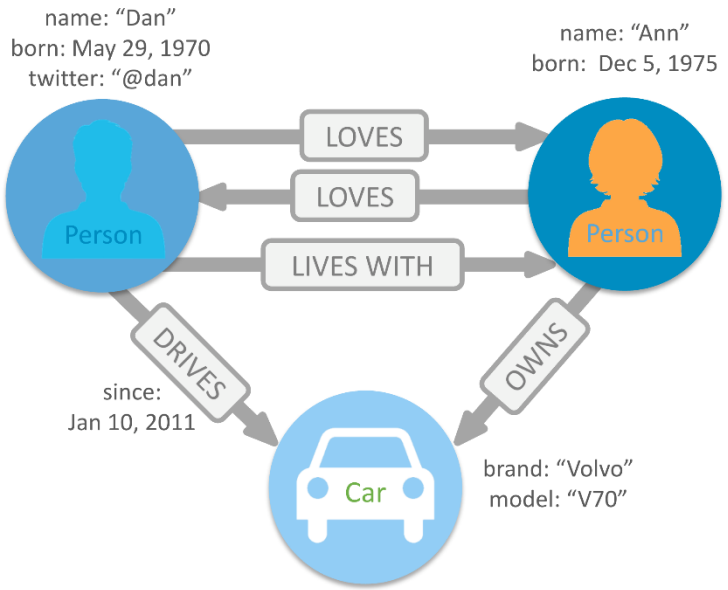
Document 3

```
{
  "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

ArangoDB OrientDB®

MUNI

Analýza (velkých) dat

Big Data

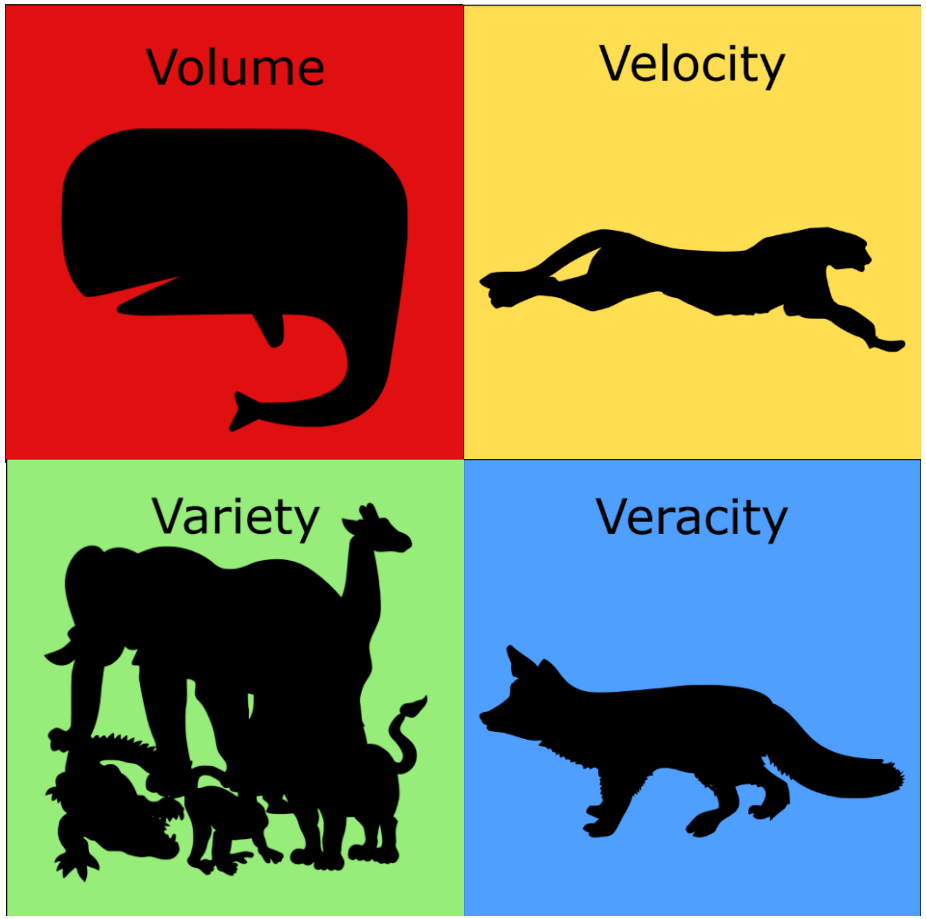
Co jsou to Big Data?

- 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 nestrukturovaná (resp. semi-strukturovaná) data

MUNI

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** ☺



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**

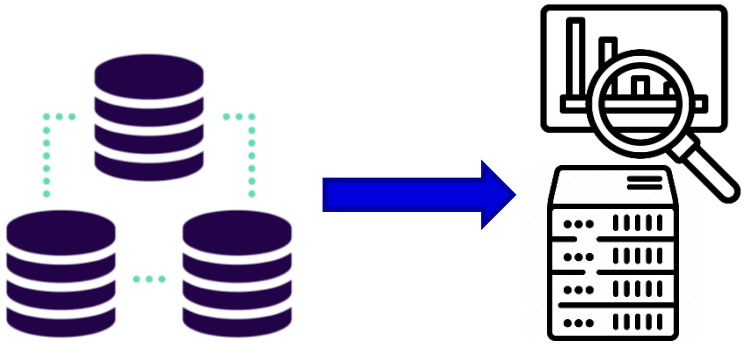
    výpočetní infrastruktura sestávající z více fyzických počítačů (výpočetních serverů)

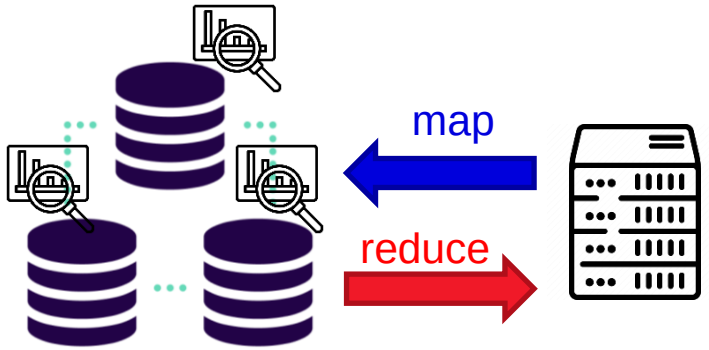
MUNI

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
- 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





MUNI

Co pro analýzu & zpracování dat potřebujeme?

- **Dobře popsaný problém ☺**
- **Vlastní data**
  - ideálně realistická, ladění zpracování lze na menších datech
  - v horším případě syntetická
- **Předpokládané dotazy**
  - mohou mít vliv na vhodný přístup pro zaindexování dat
- **Analýzu a návrh vhodného přístupu**
- **Hardwarovou infrastrukturu**
  - s dostatečnými parametry (výkon, úložiště, …)
- **Implementaci a testování přístupu**

MUNI

MUNI

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 infrastruktur

    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

MUNI

Slovníček pojmů – **výpočetní cluster**

- skupina vzájemně propojených „běžných“ počítačů

(dříve ☺)

MUNI

Slovníček pojmů – **výpočetní cluster**

- skupina vzájemně propojených „běžných“ počítačů

(dnes)

MUNI

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

MUNI

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)

MUNI

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

MUNI

NGI – dostupný výpočetní hardware

- výpočetní zdroje: **cca 34000 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, Xeon Phi, SSD disky, …

MUNI

http://metavo.metacentrum.cz/cs/state/hardware.html

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 uložiště CEPH (analogie k Amazon S3)

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

MUNI

NGI – dostupný software

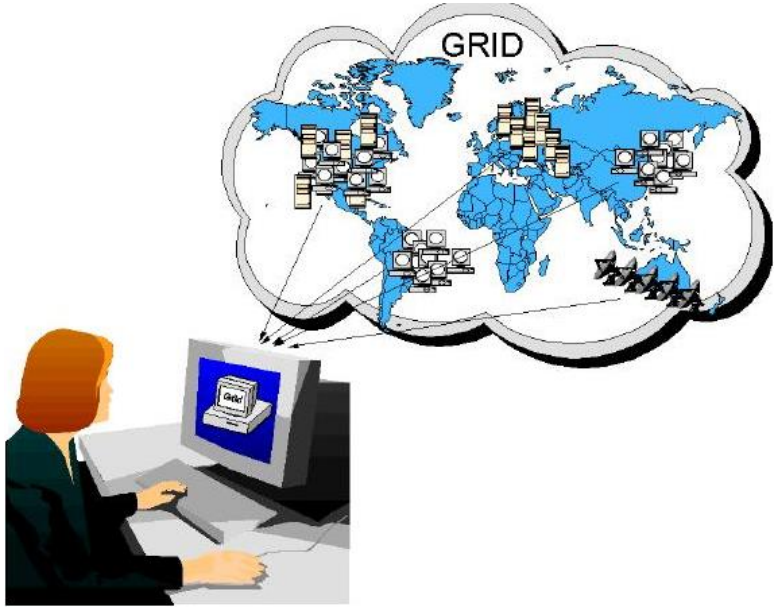
- **~ 300 různých aplikací** (instalováno na požádání)
  - viz http://meta.cesnet.cz/wiki/Kategorie:Aplikace
- průběžně udržované **vývojové prostředí**
  - GNU, Intel, PGI, ladící a optimalizační nástroje (TotalView, Allinea), …
- generický **matematický software**
  - Matlab, Maple, Mathematica, gridMathematica, …
- 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 balíků**
- …

MUNI

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, …



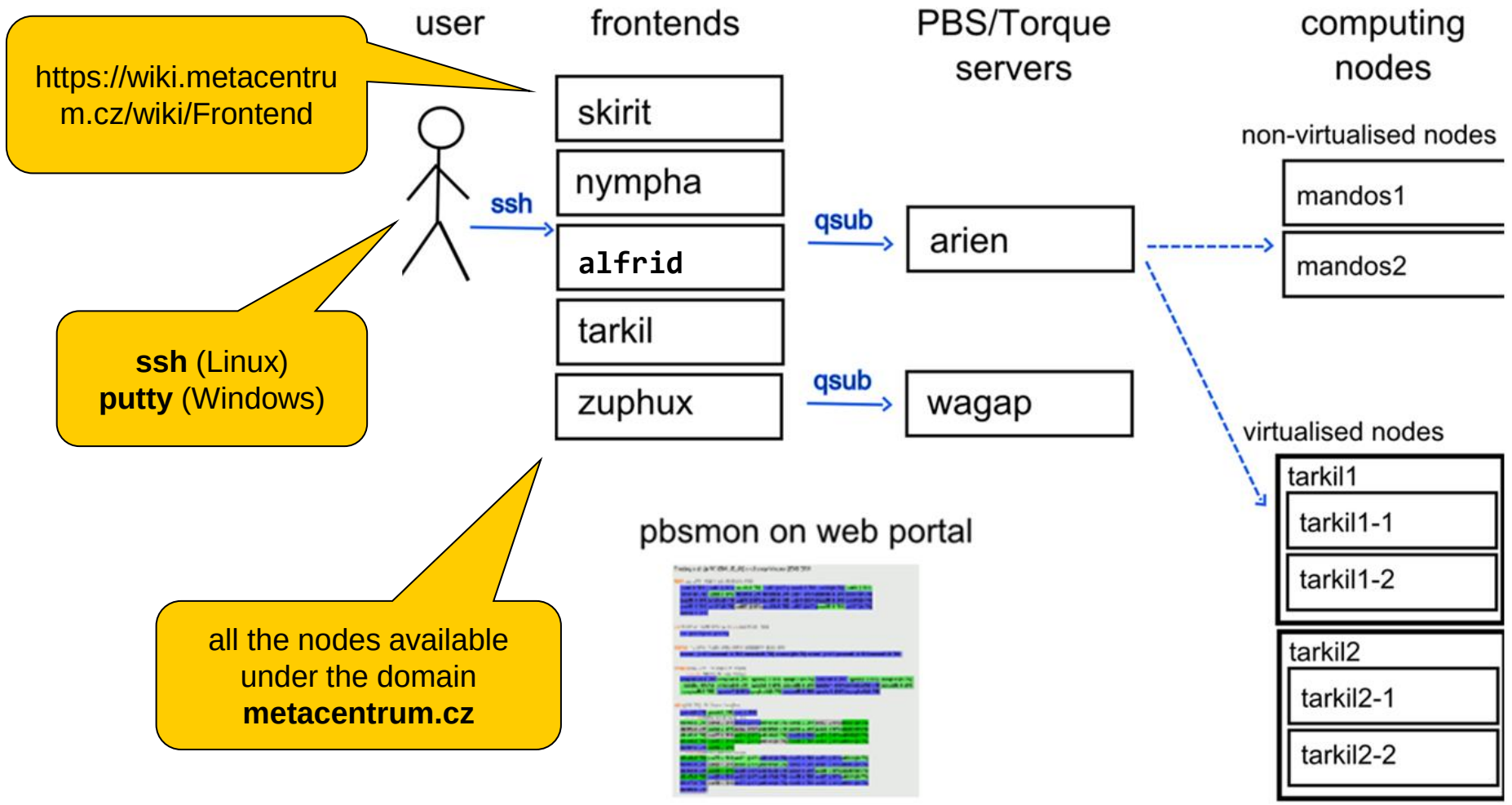
MUNI

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

MUNI

NGI pod pokličkou

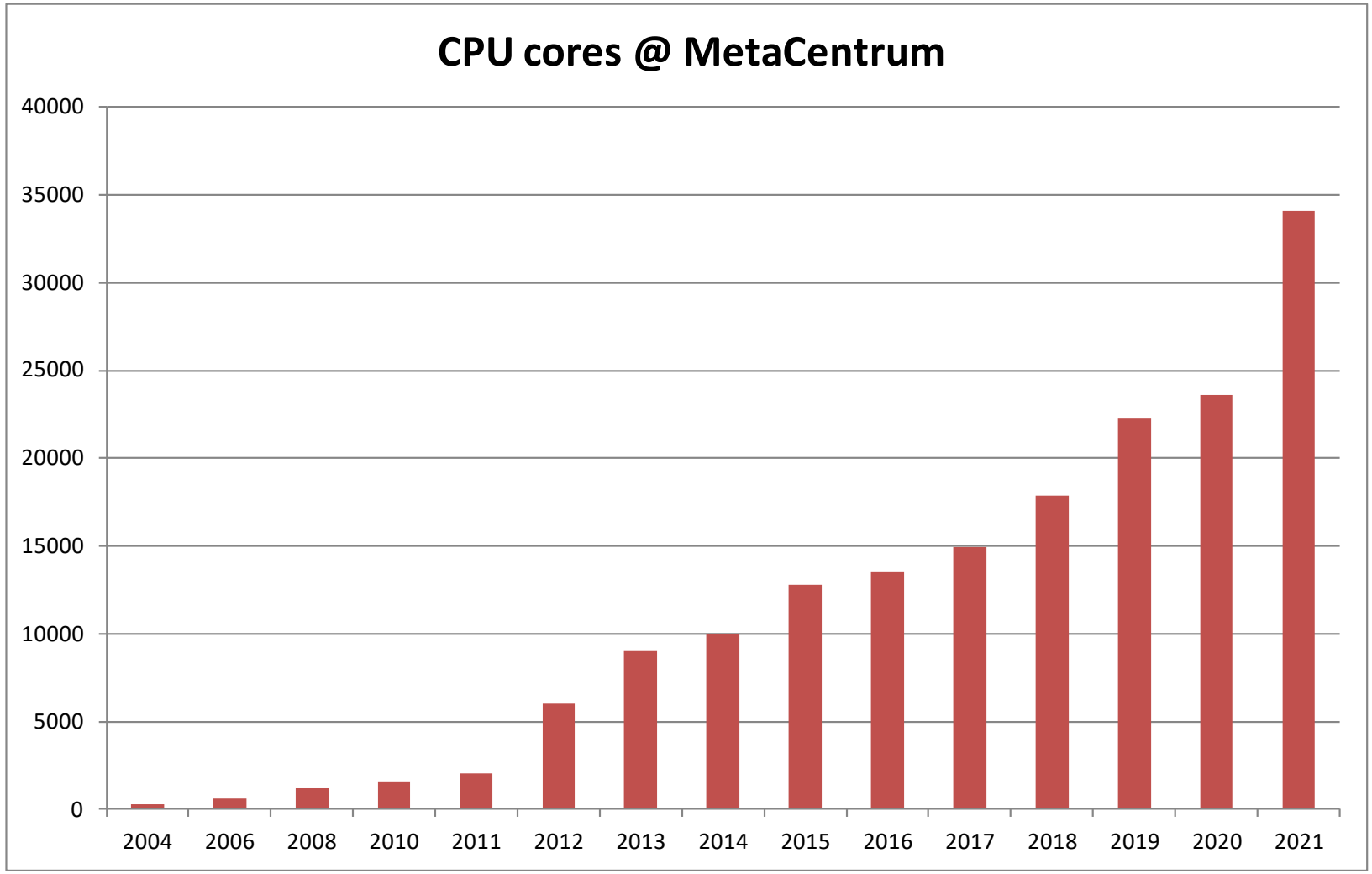


NGI pod pokličkou – **v číslech…**

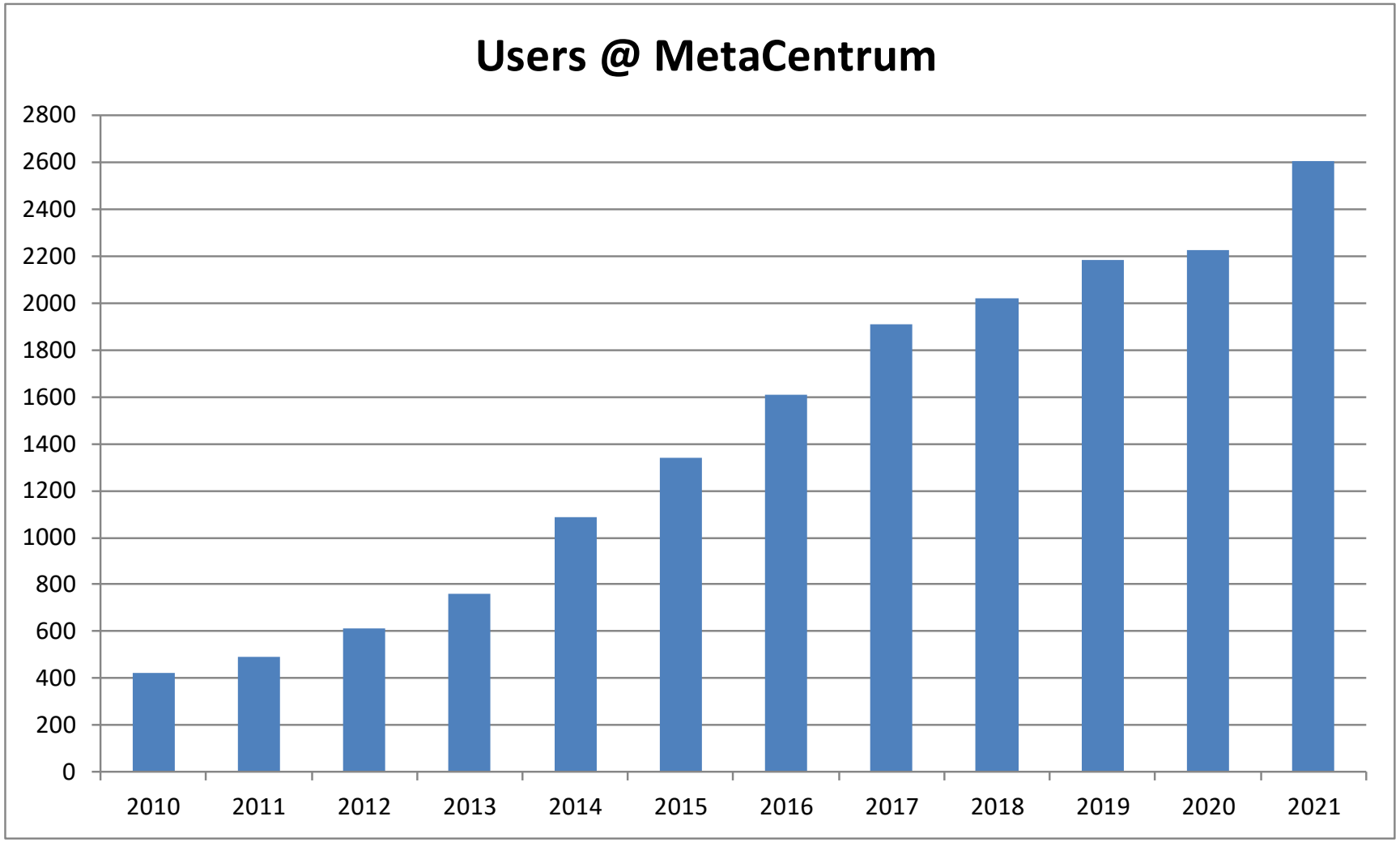
- cca 34084 výpočetních jader, cca 600 uzlů
- za rok 2020:
  - 2606 uživatelů (k 31.12.2021)
  - cca 12 mil. spuštěných úloh
    - cca 33000 úloh denně
    - cca 4600 úloh / uživatel
  - celkem propočítáno cca 22,6 tis. CPUlet

MUNI

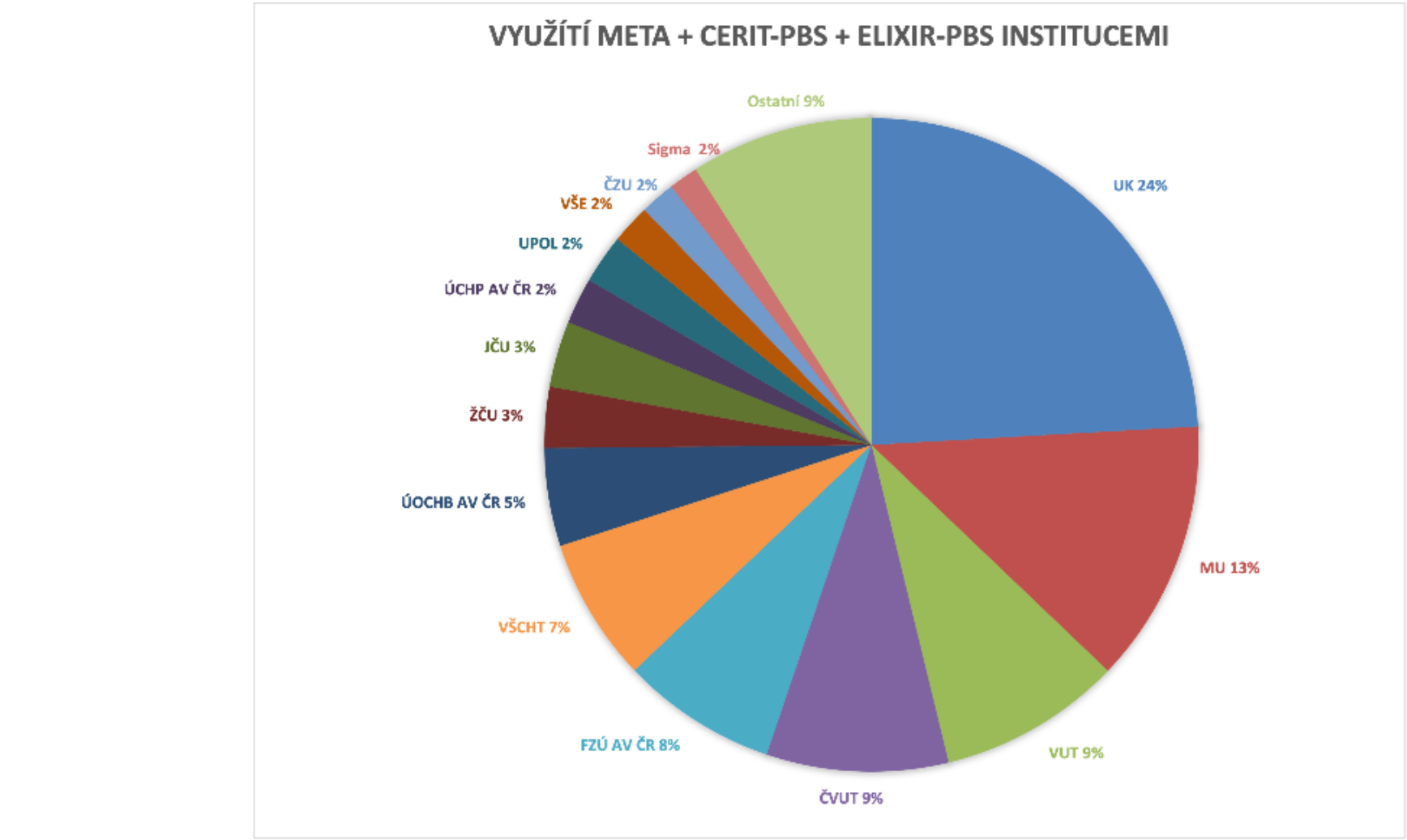
NGI pod pokličkou – **a grafech…**



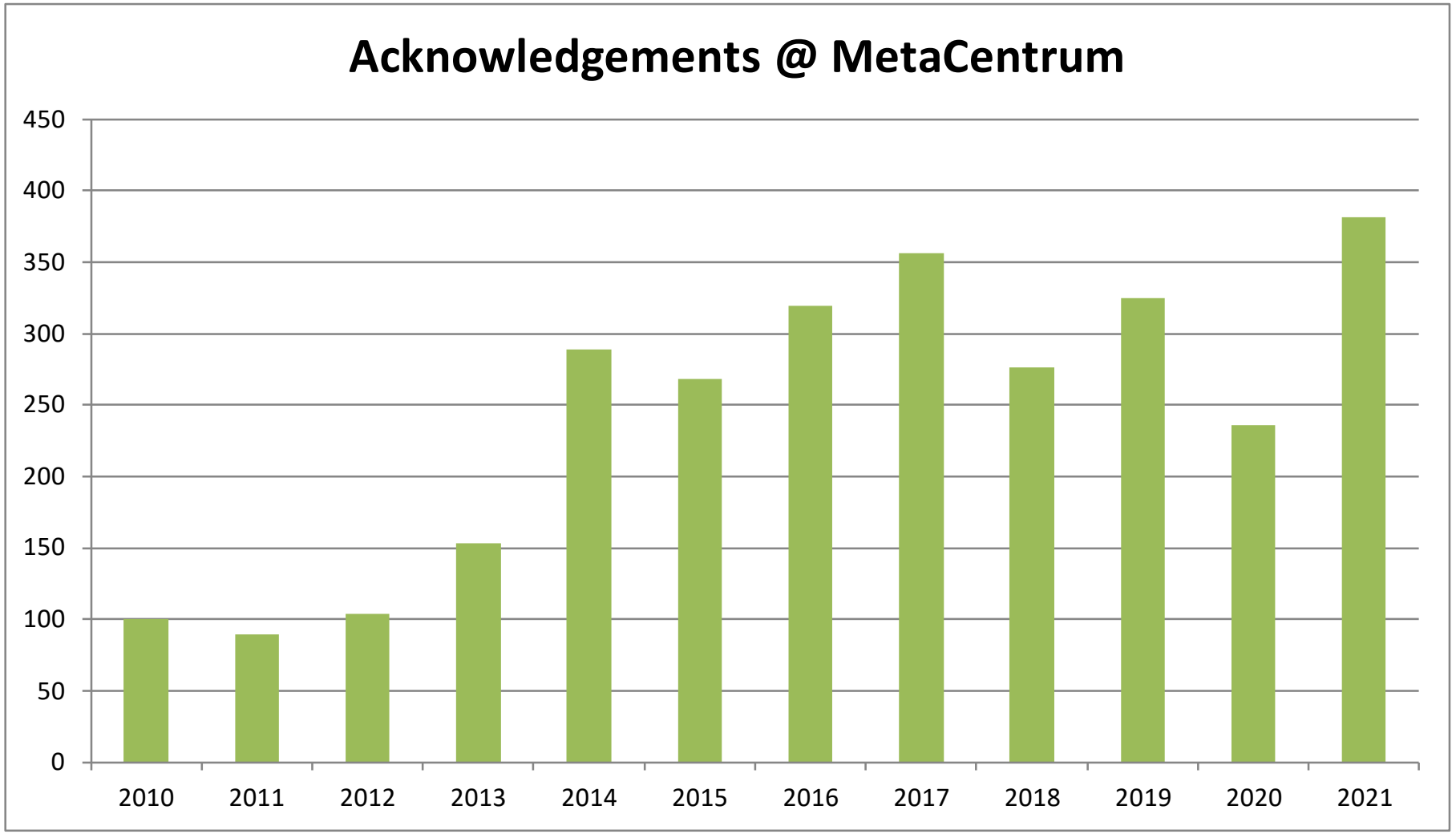
NGI pod pokličkou – **a grafech…**



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

MUNI

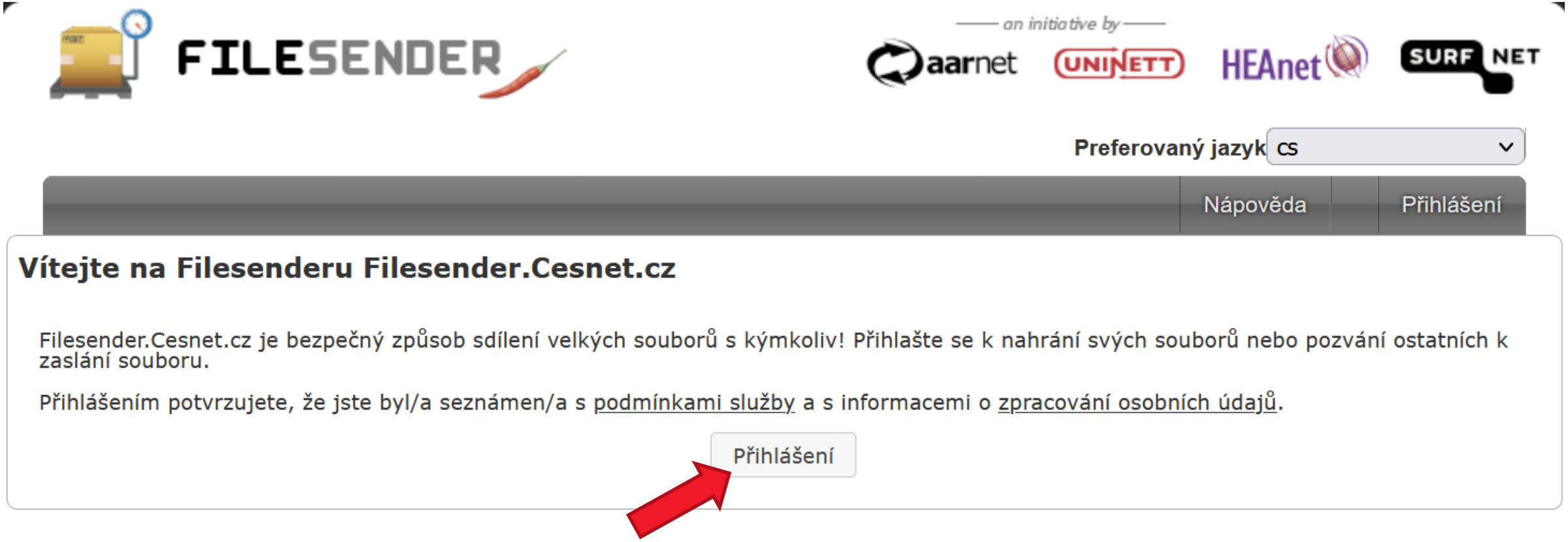
IT4Innovations

- **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.

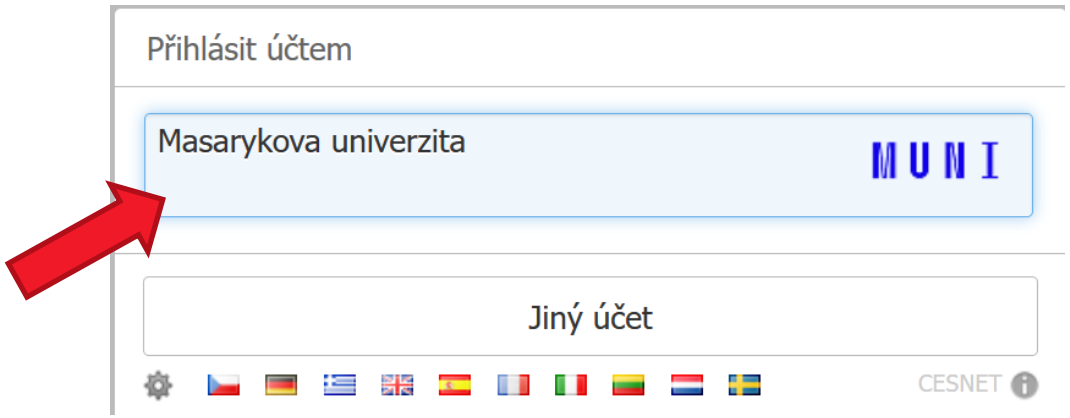
- **FileSender** – webová služba pro zasílání velkých souborů
- aktuální limit je 2 TB (~ 2000 GB)
  - doba expirace až 1 měsíc
- http://filesender.cesnet.cz
- odesílatel nebo příjemce musí být autorizovaným akademickým pracovníkem
  - autorizovaný uživatel **může odesílat datové soubory** libovolnému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

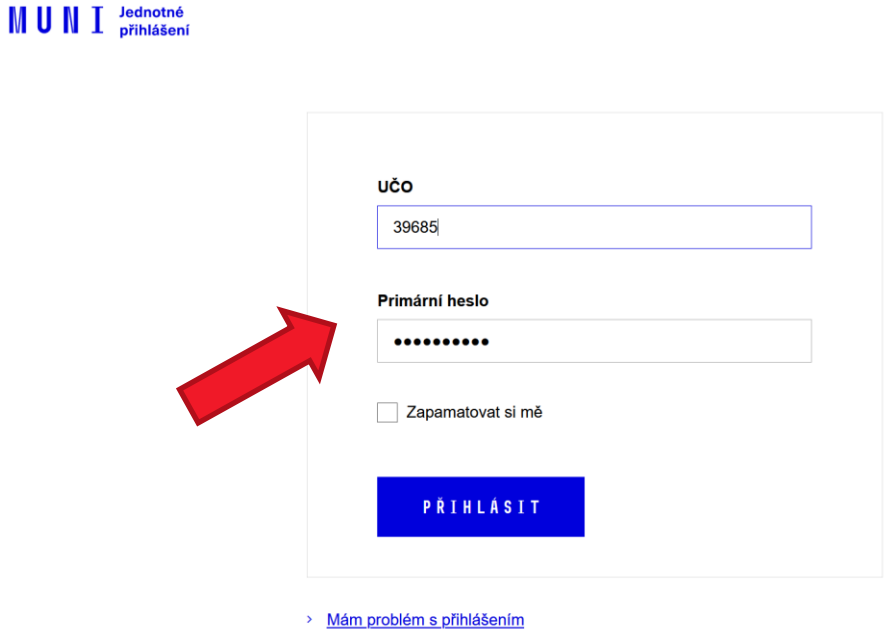
FileSender – ukázka využití



MUNI

FileSender – ukázka využití

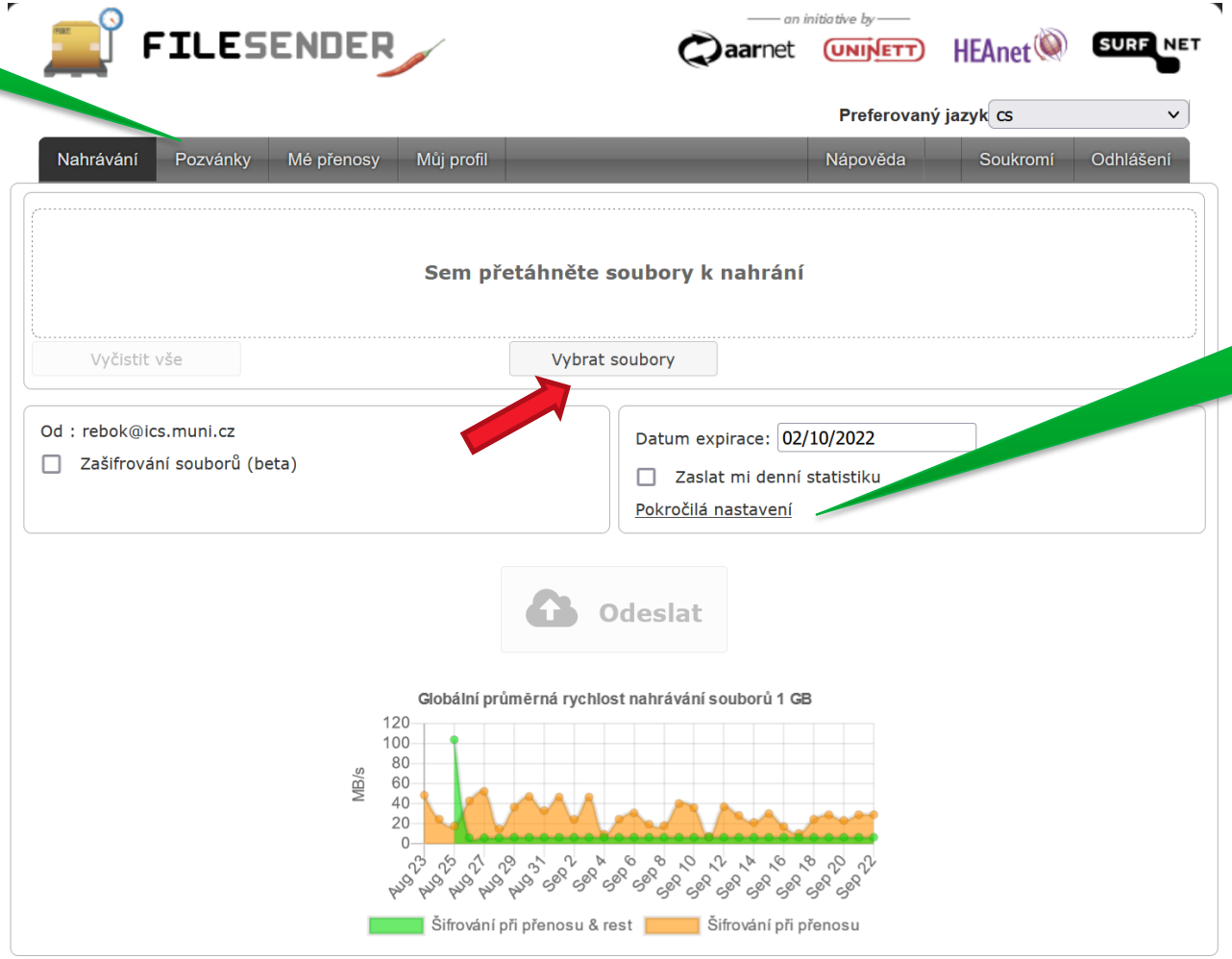




MUNI

FileSender – ukázka využití

pozvánky



pokročilé
notifikace,
získání odkazu, atp.

MUNI

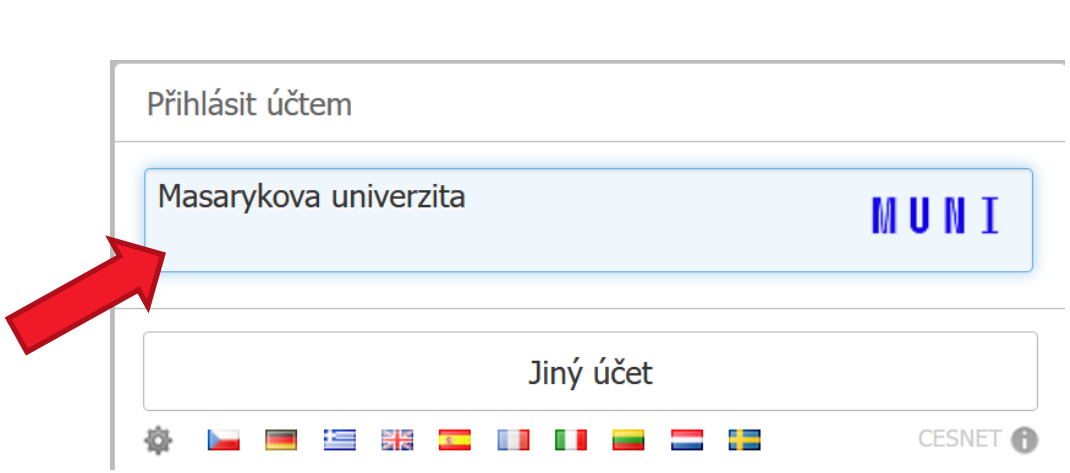
Datové služby e-INFRA CZ pro koncové uživatele II.

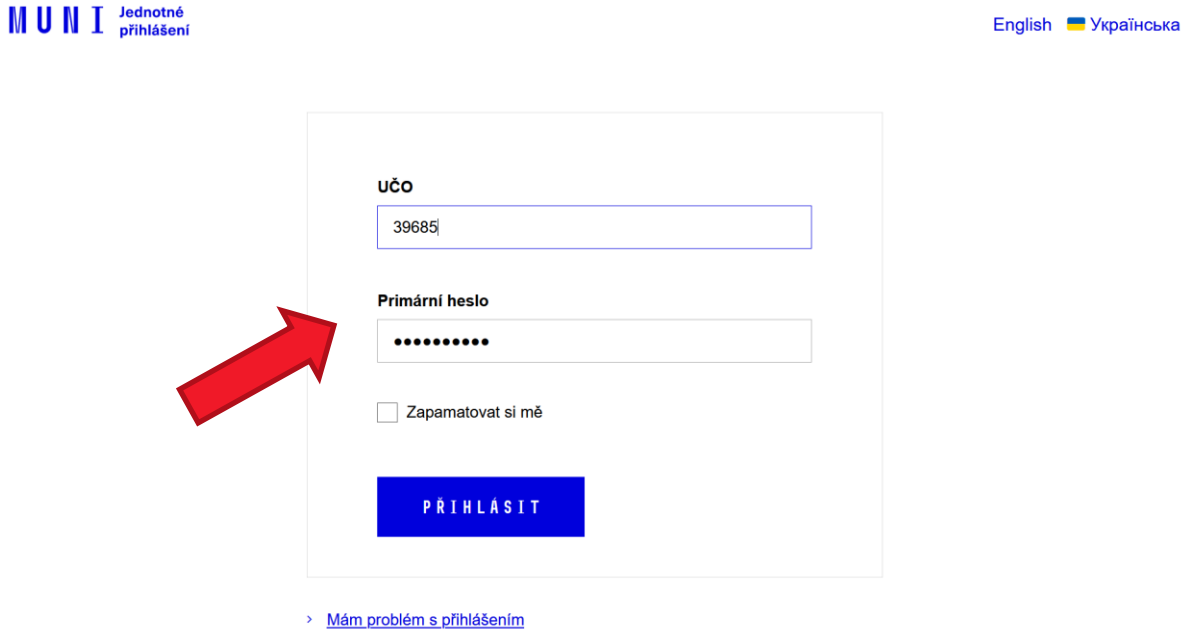
- **OwnCloud** – cloudové uložiště a-la Google Drive nebo Dropbox
  - aktuální kvéta je 100 GB / uživatel
- https://owncloud.cesnet.cz/
- **synchronizace a dostupnost dat mezi zařízeními**
  - klienti dostupní pro OS Windows, Linux, OS X
  - také pro chytré telefony a tablety
  - umožňuje sdílení dat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

OwnCloud – ukázka využití

MUNI

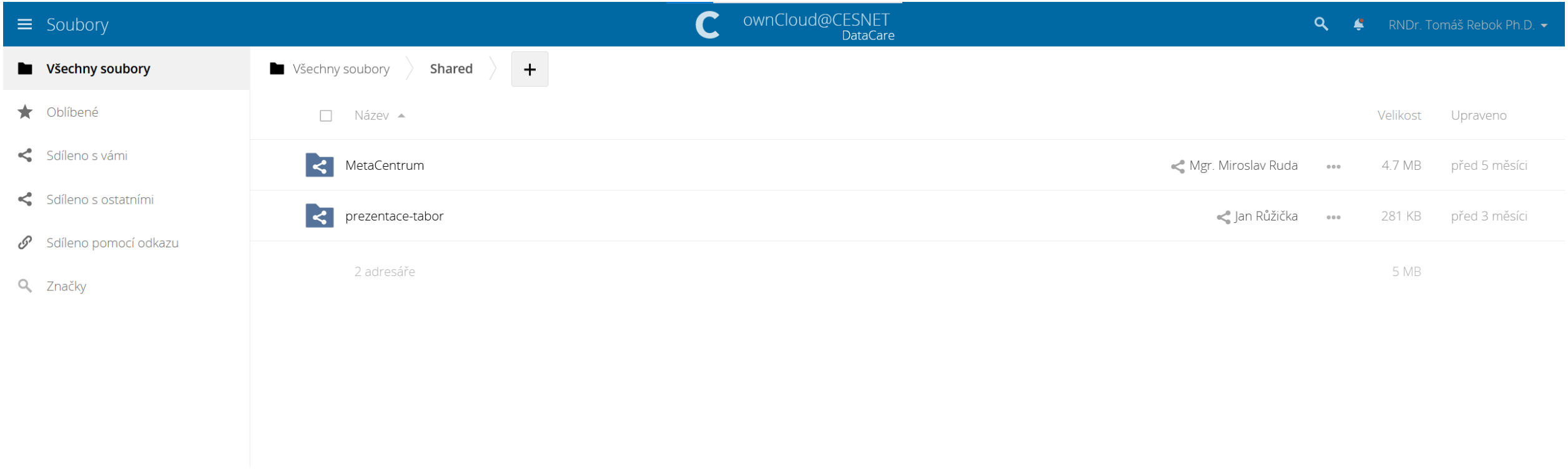
OwnCloud – ukázka využití





MUNI

OwnCloud – ukázka využití

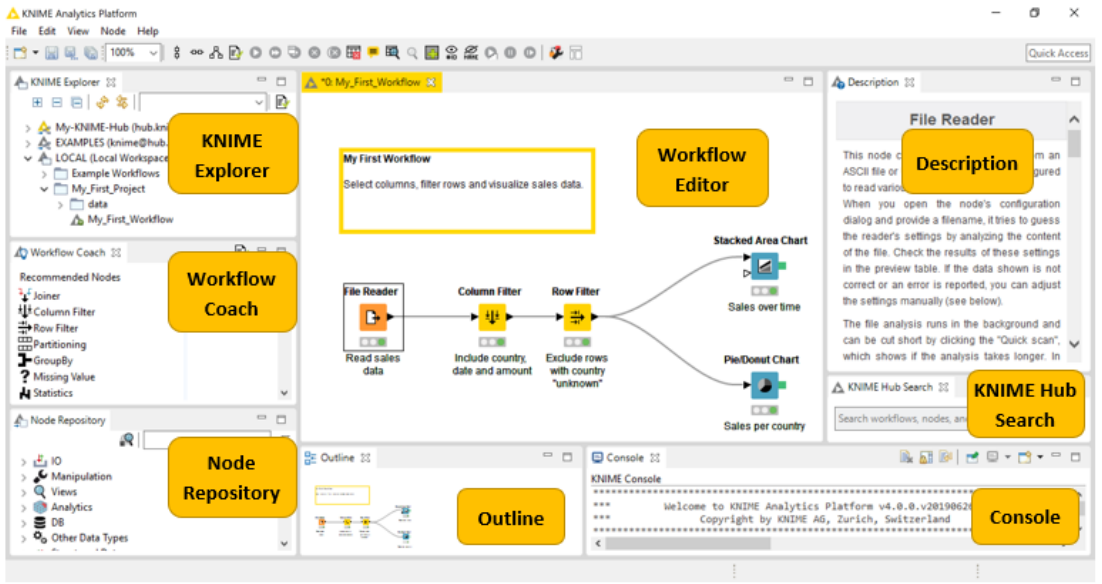


MUNI

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**

Open for Innovation
KNIME



Vybrané datově-analytický výzkum realizovaný v rámci spolupráce CERIT-SC

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 ekosystémůr
  - růstové komory
  - letecká laboratoř
  - laboratoře
  - atp.

Ústav výzkumu globální změny Akademie věd ČR

MUNI

Plánování sběru a shromažďování dat

Planování sběrů dat

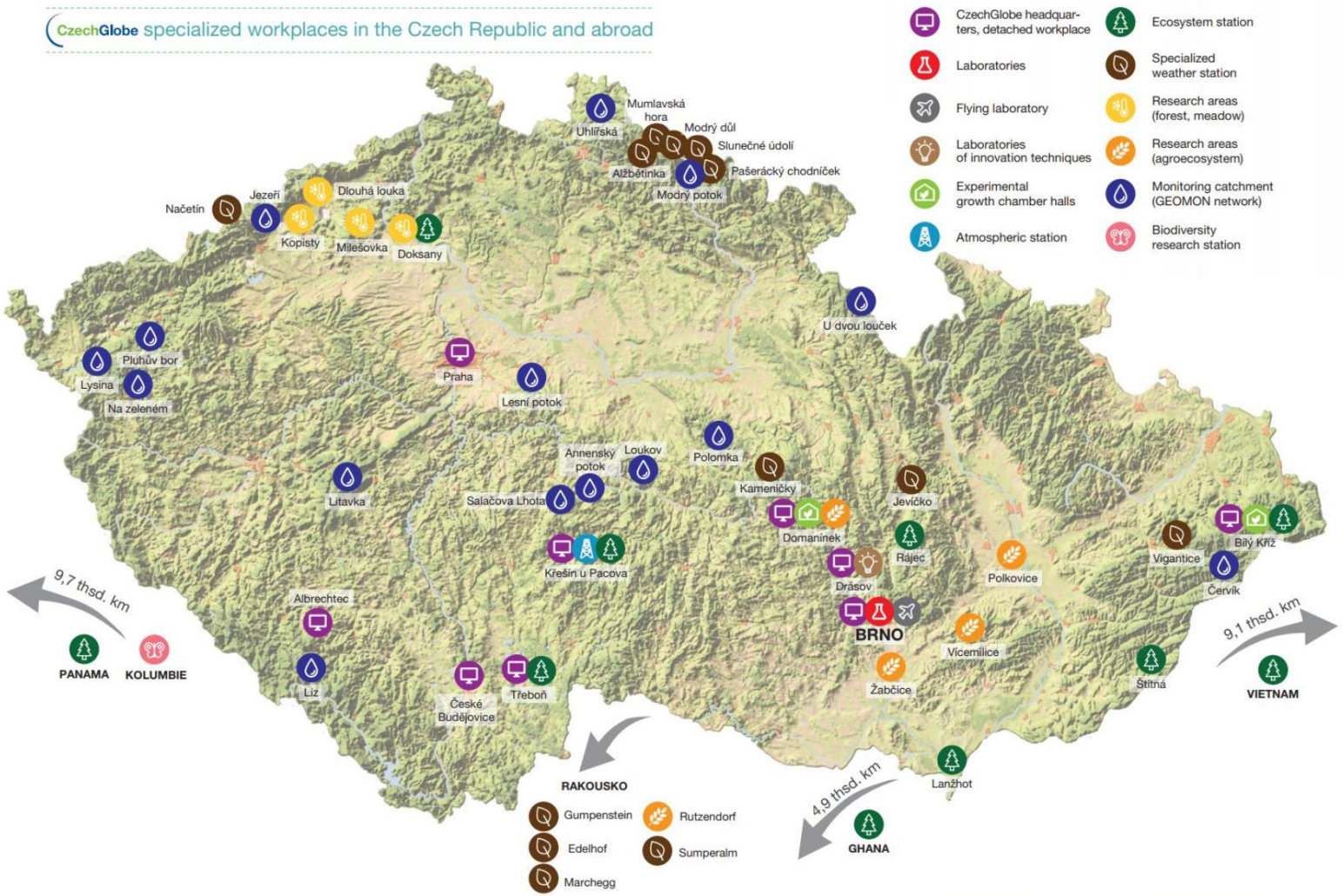
- pravidelný sběr dat
  - zahrnuje mj. plánování lokalit ekosystémových a atmosférických stanic
  - nejstarší záznamy z roku 1996
- nepravidelný sběr dat
  - plánované „kampaně“ – např. nálety vybraných ekosystémů leteckou laboratoří

Shromažďování dat

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

MUNI

Sběr dat – pozemní měřící stanice



Sběr dat – letecká laboratoř pro dálkový průzkum

Visible and near infra-red imaging spectroscopy

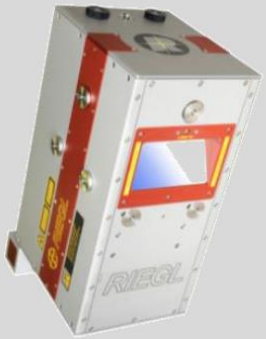
CASI-1500

SASI-600

Thermal imaging spectroscopy

TASI-600

Laser scanning



Riegl LMS – Q780

MUNI

Sběr dat – typy dat dálkového průzkumu Země

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/
  - zarovnávaní leteckých snímků
    - eliminace pohybů letadla vůči Zemi

MUNI

Zpracování dat

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

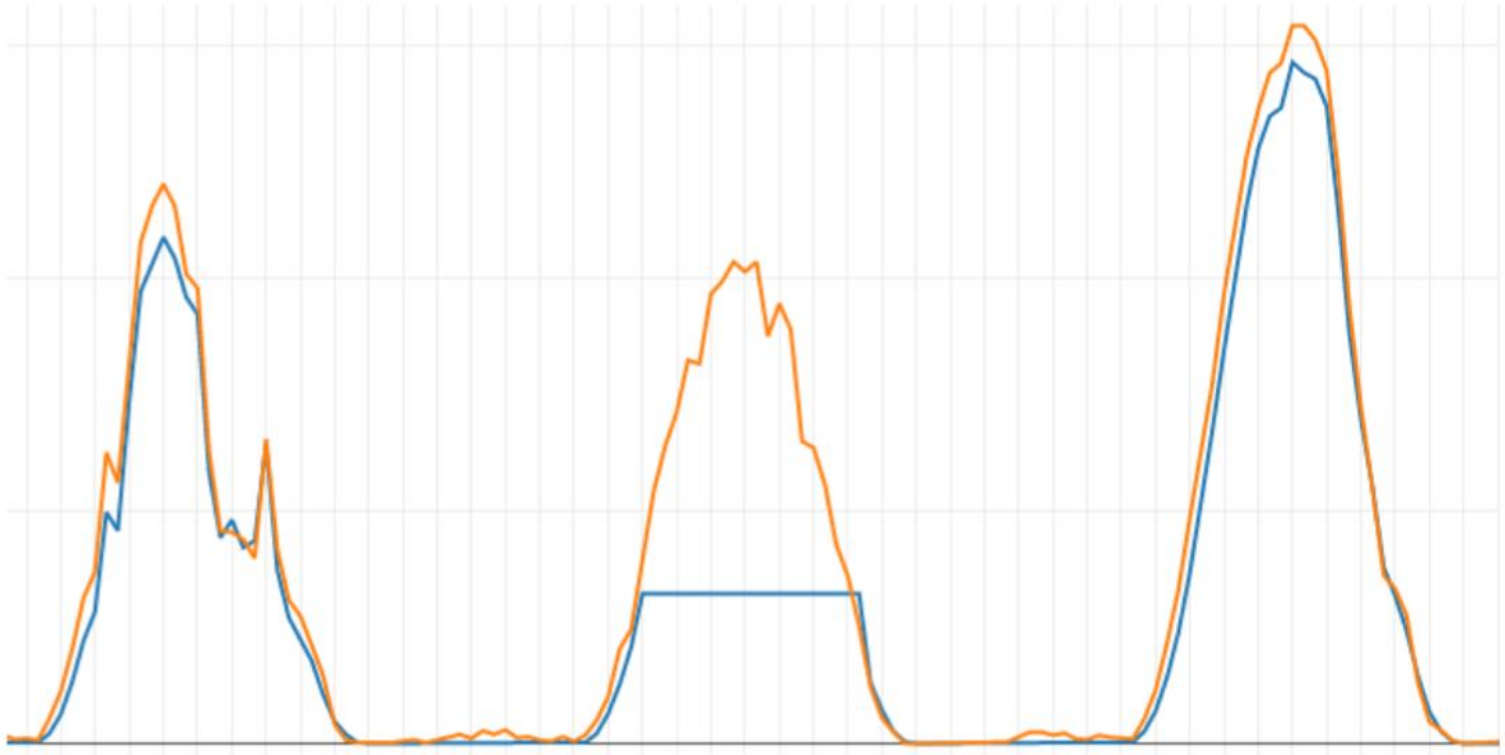


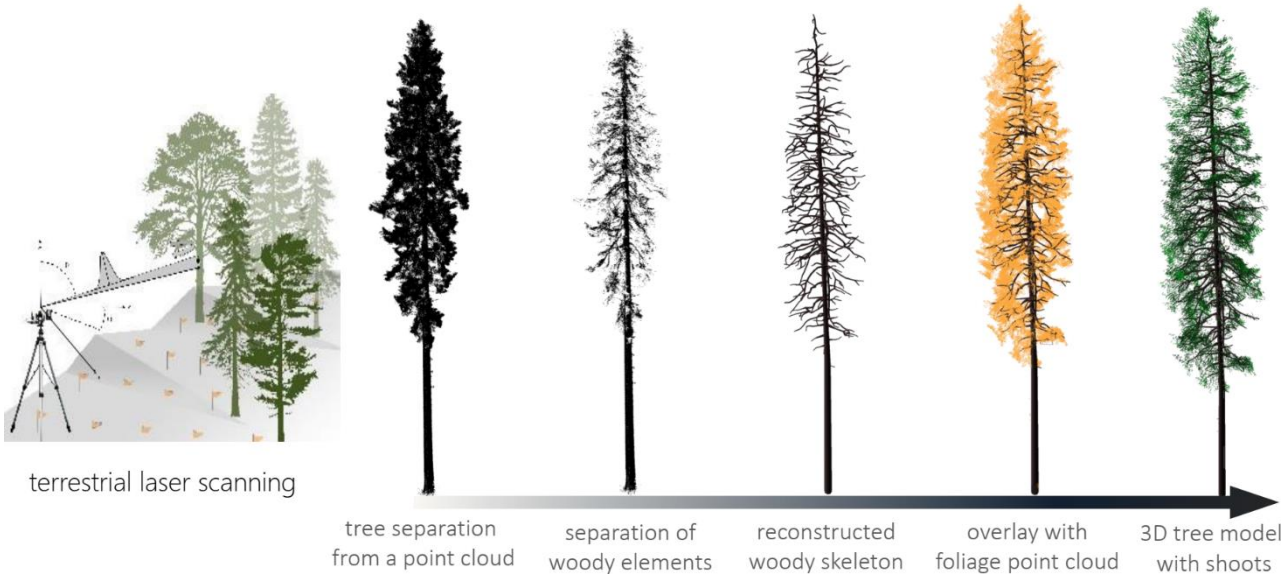
Figure 5.1: Correction of UVB measurement (update interval 30 min).
Blue line shows input data, yellow line is gap-filling result.

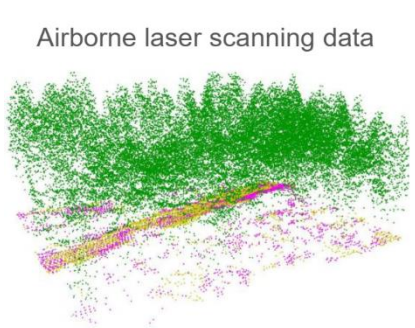
Analýza dat – příklady

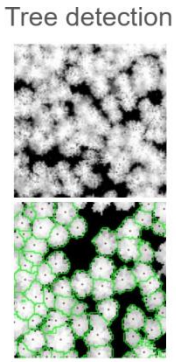
realizované ve spolupráci CzechGlobe a ÚVT MU

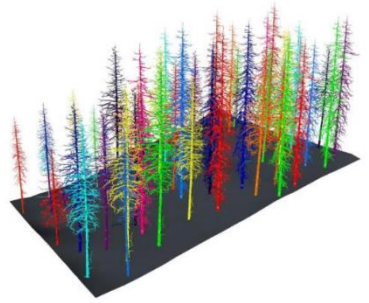
Rekonstrukce 3D modelů stromů a lesů

- 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

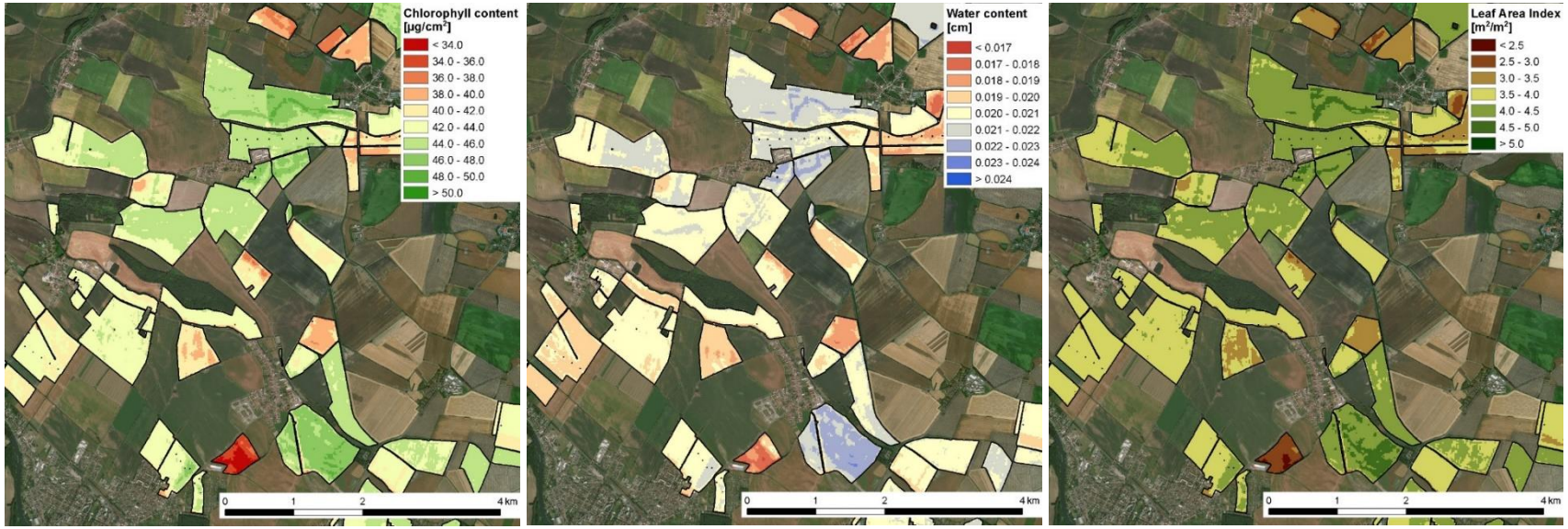
- v definovaném časovém rozsahu a prostoru
  - omezení na sledované vegetační období
- vstupem jsou data z družice Sentinel-2
- více metod:
  - per-pixel
  - per-dlaždice
- výstup je vstupem pro návaznou analýzu

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

MUNI

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

MUNI

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 podpůrných 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 případu B
  - nanejvýš pro „studijní“ či rozvojové potřeby

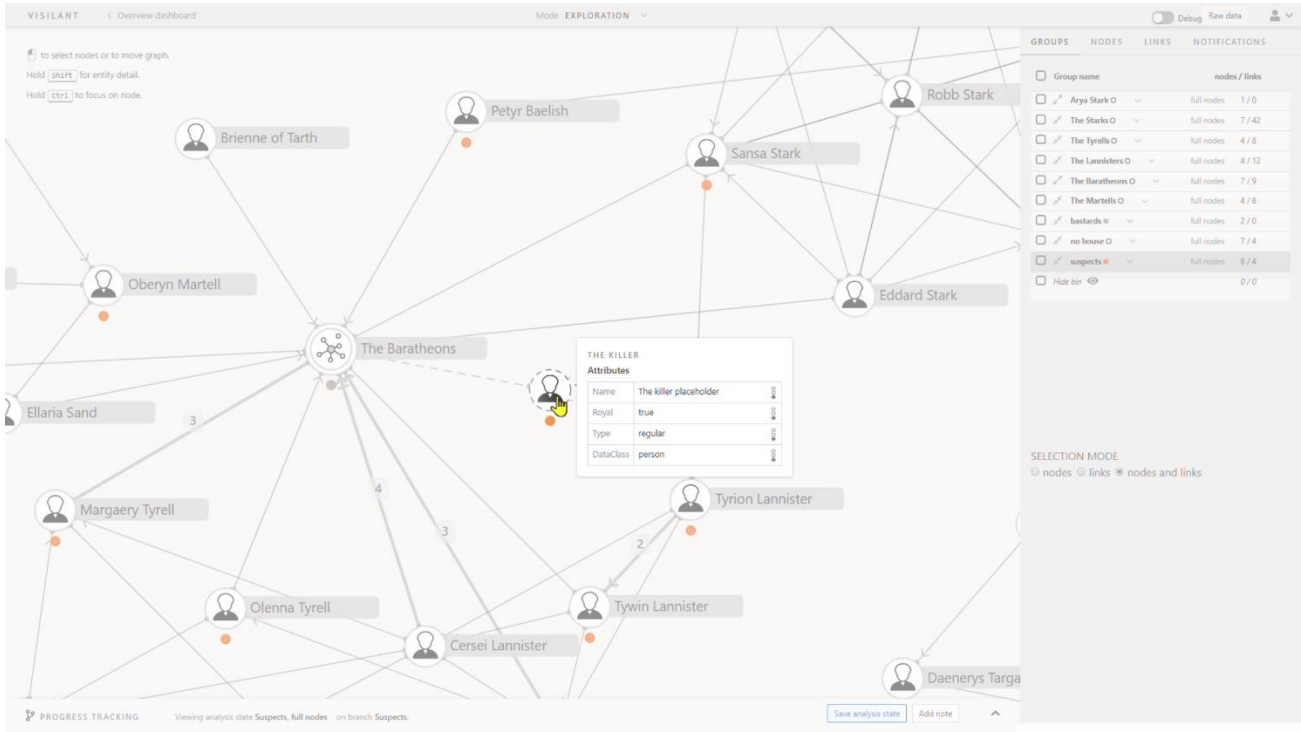
MUNI

Platforma ANALYZA

Realizovaná projektem ÚVT MU pro účely 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, podán návazný projekt



Shrnutí

teoretické části …

Shrnutí

Zpracování a analýza dat

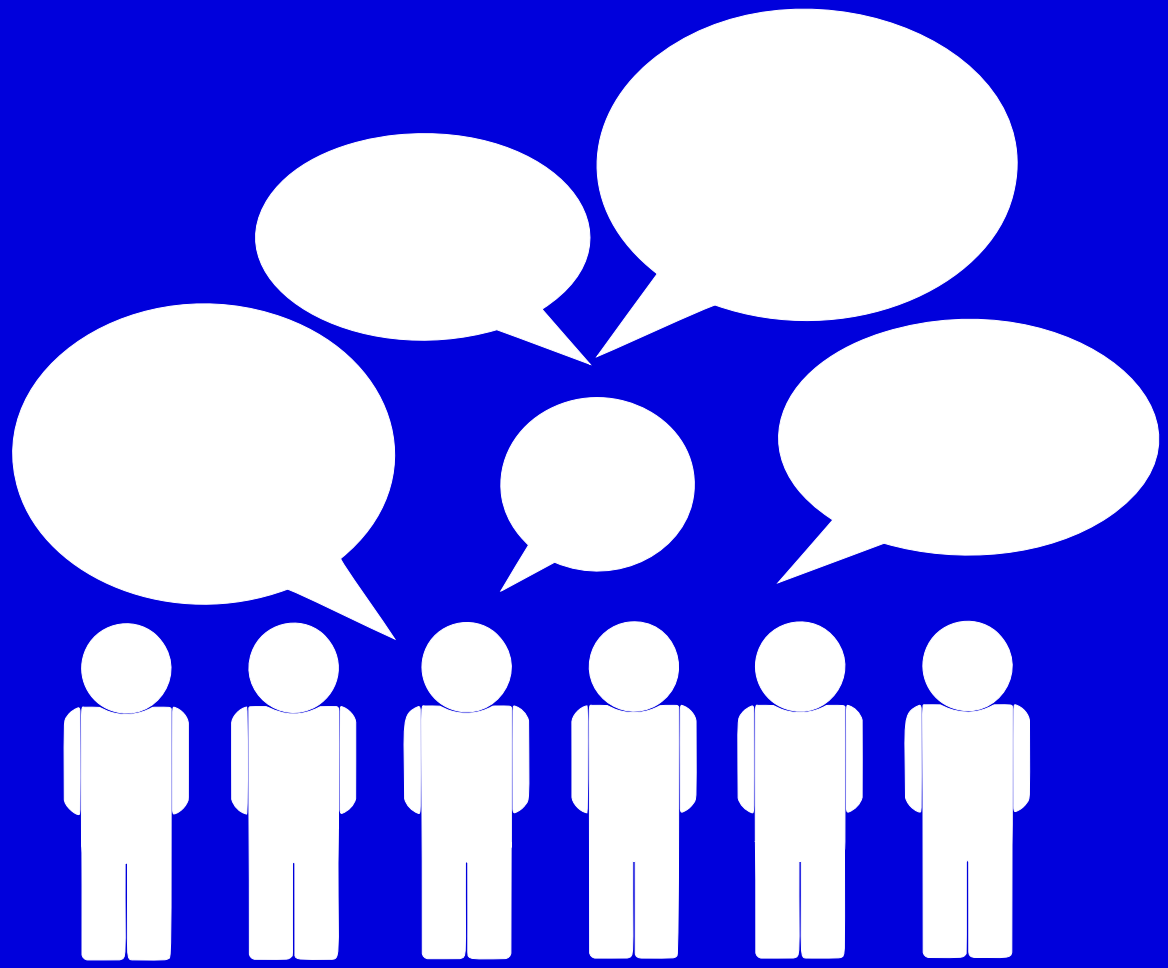
- jak se na data dívat?
  - strukturovaná vs. nestrukturovaná vs. semi-strukturovaná
  - výběr vhodného modelu pro zpracování a analýzu
    - důležitá je i znalost předpokládaných dotazů
      - tabulkové procesory, SQL databáze, NewSQL databáze, NoSQL databáze
- nebojte se být Big(Data) ☺

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