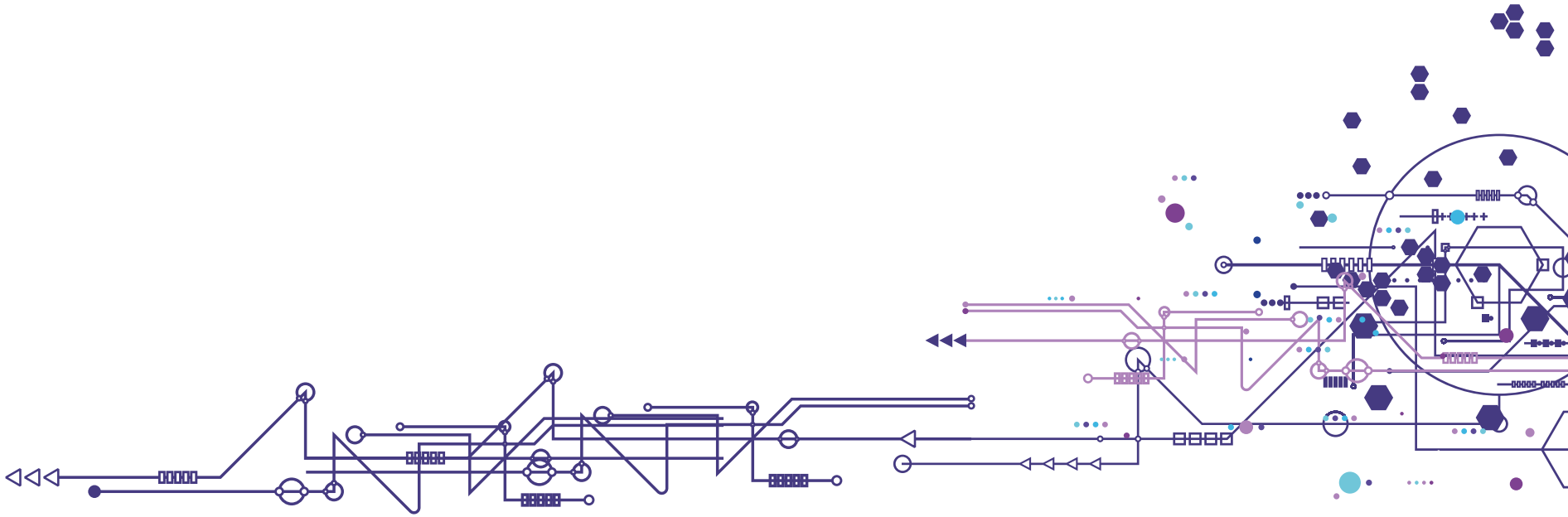
PS Správa citlivých dat

Zdenka Dudová (BBMRI.cz),
Adam Svobodník (CZECRIN)

EOSC Roadshow 2022, Brno
3. 6. 2022

Osnova

- Jak PS Správa citlivých dat funguje
- Vydefinování “citlivých dat”
- Jaké jsou rozdíly oproti správě běžných výzkumných dat
- Co řešíme a kam směřuje naše představa
- Kde se inspirujeme
- Závěrem



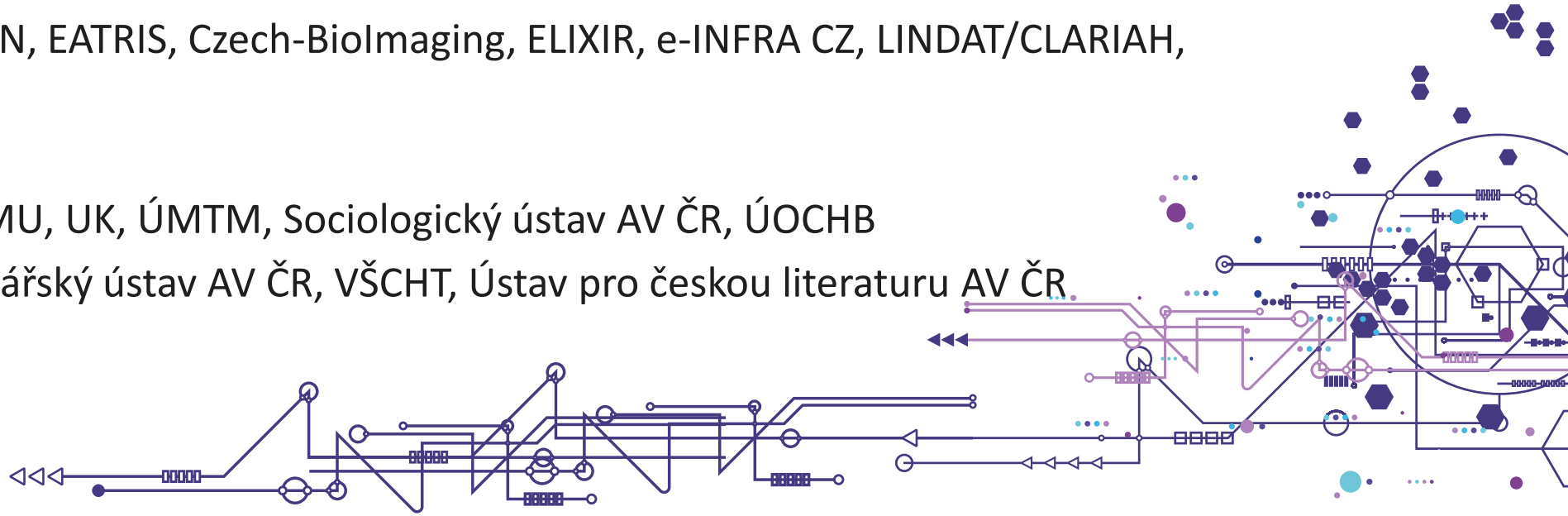
Jak PS Správa citlivých dat funguje

Organizace PS

- Fungování PS je založeno na osobních setkáních
- PS je vedena Zdenkou Dudovou (BBMRI.cz) a Adamem Svobodníkem (CZECRIN)

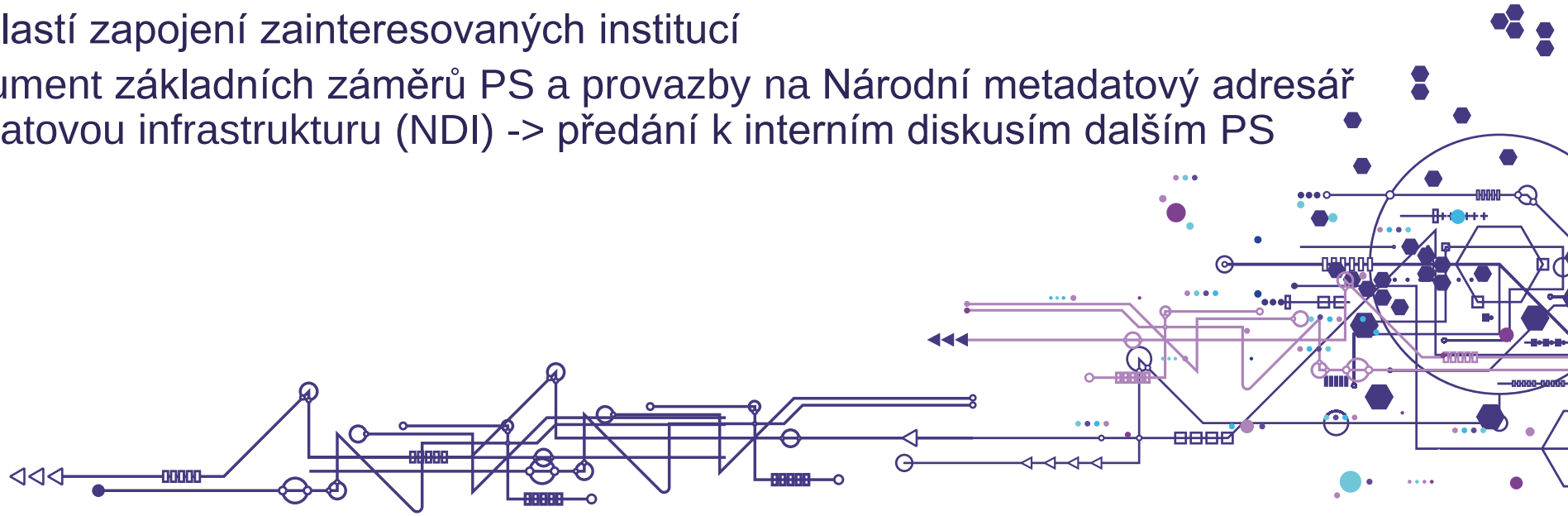
Členové PS

- Velké výzkumné infrastruktury
  - BBMRI, CZECRIN, EATRIS, Czech-BioImaging, ELIXIR, e-INFRA CZ, LINDAT/CLARIAH, RECETOX
- Instituce
  - FN HK, MOÚ, MU, UK, ÚMTM, Sociologický ústav AV ČR, ÚOCHB
  - Národohospodářský ústav AV ČR, VŠCHT, Ústav pro českou literaturu AV ČR



PS Správa citlivých dat v čase

- 5. 4. 2022 online kick-off
- 20. 4. 2022 první pracovní setkání v Brně
  - Sdílení a diskuse nad stávající situací
  - High-level požadavky na infrastrukturu pro správu citlivých dat, idea kompetenčních center, identifikace bariér
- 2. 6. 2022 druhé pracovní setkání v Brně
  - Konkretizace specifických požadavků na infrastrukturu pro správu citlivých dat
  - Identifikace oblastí zapojení zainteresovaných institucí
- Konec 6/2022 dokument základních záměrů PS a provazby na Národní metadatový adresář (NMA) a Národní datovou infrastrukturu (NDI) -> předání k interním diskusím dalším PS
- Design prototypu



Co považujeme za citlivá data

Zaměřeno na GDPR

- **Osobní údaje** = veškeré informace vztahující se k identifikované či identifikovatelné fyzické osobě
  - **Obecné** osobní údaje = jméno, pohlaví, věk a datum narození, osobní stav, ale také IP adresu a fotografický záznam
  - **Organizační** údaje = e-mailová adresa, telefonní číslo či různé identifikační údaje vydané státem
- **Zvláštní kategorie** osobních údajů
  - Údaje o rasovém či etnickém původu, politických názorech, náboženském nebo filozofickém vyznání, členství v odborech, o zdravotním stavu, sexuální orientaci a trestních deliktech či pravomocném odsouzení osob
  - Nově zahrnuje genetické, biometrické údaje a osobní údaje dětí

https://www.gdpr.cz/gdpr/osobni-udaje/

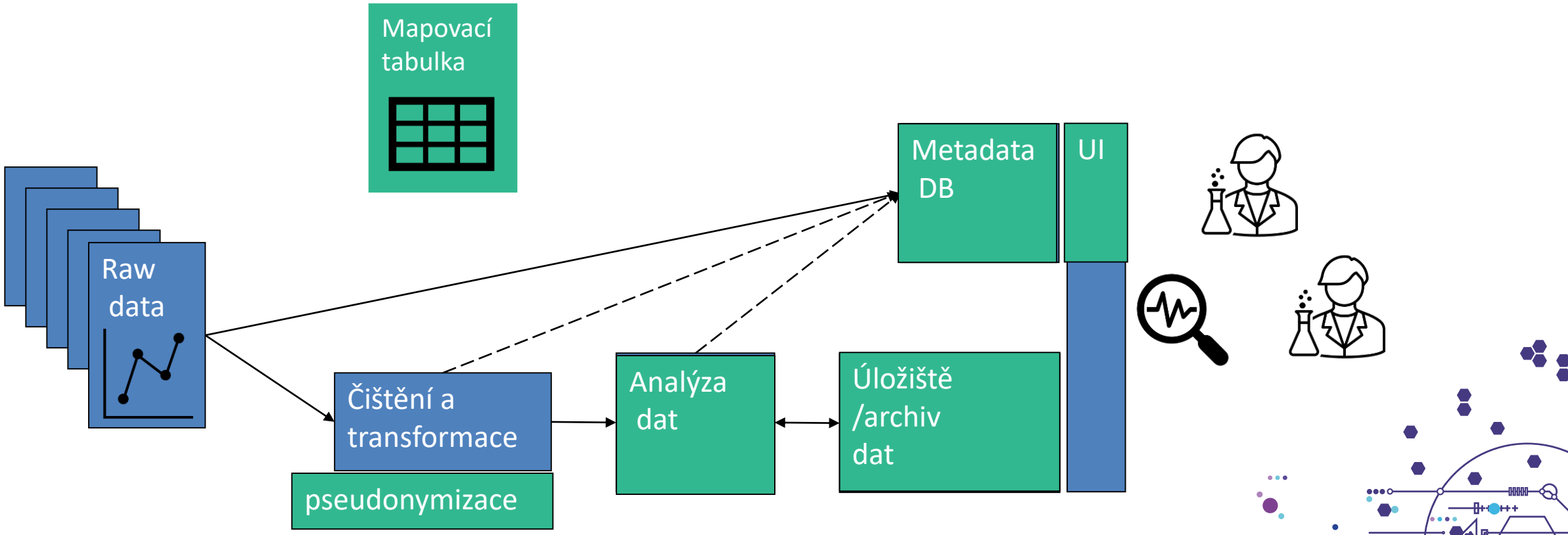
Informovaný souhlas

Zpracování a sdílení dat není možná bez souhlasu “majitele” – donora

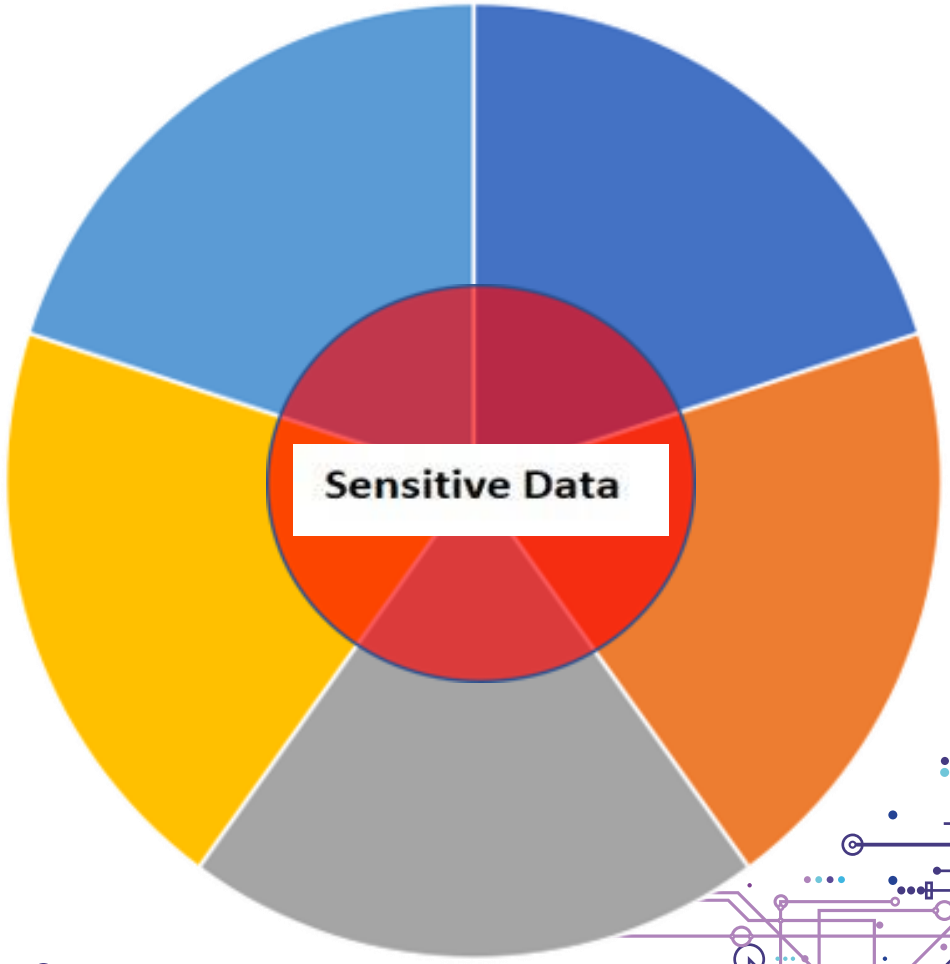
- Osoba, které se data týkají, musí dát souhlas s jejich **využitím pro vědecké účely a sdílení třetí straně**, resp. případný jiný účel, než je primární
- Správce = zákonný důvod zpracování osobních údajů a dat
- Zpracovatel = pověření správcem zpracovávat os. údaje a data
- Sdílení dat je nutné opatřit jasnými pravidly a případně i smlouvou (Data Transfer Agreement)

- Co jsou (jen) metadata, která je možné sdílet bez rizika?

Výzkumná vs. Citlivá data



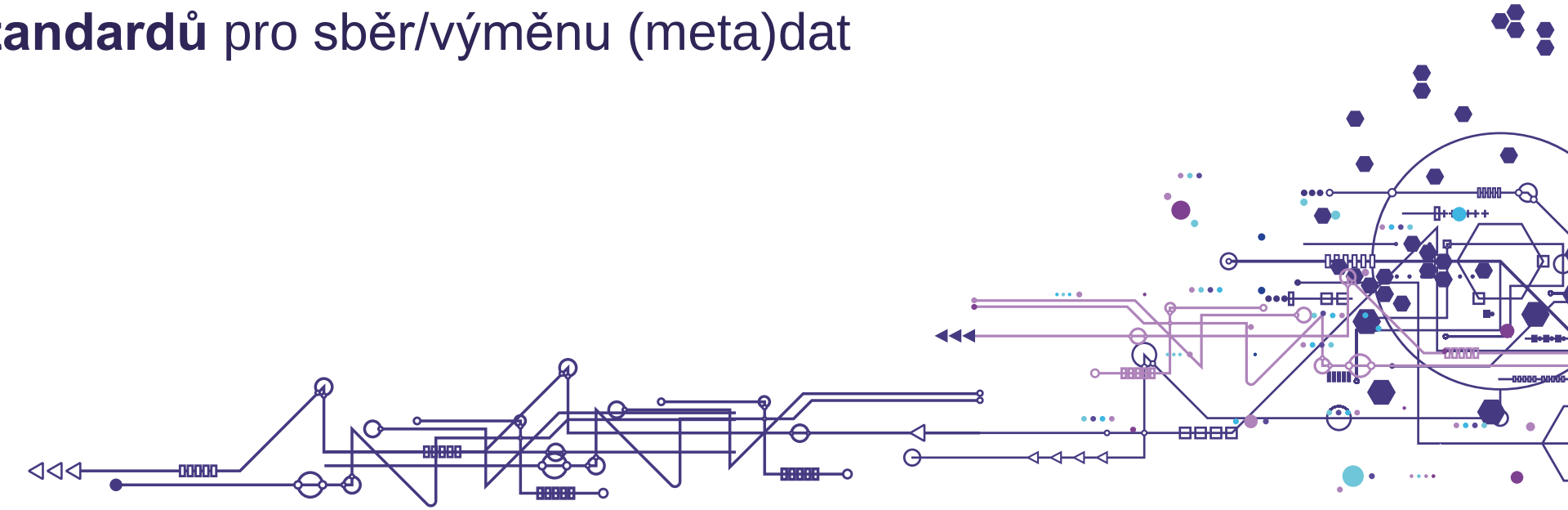
Životní cyklus výzkumných dat



Představa infrastruktury pro správu citlivých dat

Plánování a sběr dat

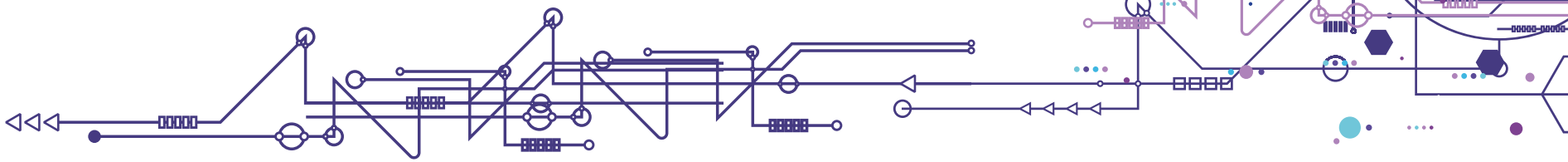
- **Data Management Planning**
  - Komentovaný checklist/templát, jak má vypadat DMP pro správu citlivých dat
- Analýza existujících nástrojů pro **sběr** dat
  - Open-source vs. licencovaný SW
- Doporučení **standardů** pro sběr/výměnu (meta)dat
  - HL7 FHIR



Představa infrastruktury pro správu citlivých dat

Zpracování a datová analýza

- Jak **identifikovaná data** zpracovávat v rámci infrastruktury EOSC a v jakých případech je vhodné kterou úpravu použít
  - Anonymizace vs. Pseudonymizace vs. Primární identifikátory (r. č.)
- Informovaný souhlas a jeho **parametrizaci**!
  - Automatizované kontroly obsahu – např. případ “incidental findings” a jak na ně reagovat
- **Čištění a validace** dat před datovou analýzou
  - Nástroje a jejich promazávání, tzv. pre-processing
- **Statistical Analysis Plan**
- Bezpečné **analytické** prostředí s vhodnými nástroji



Představa infrastruktury pro správu citlivých dat

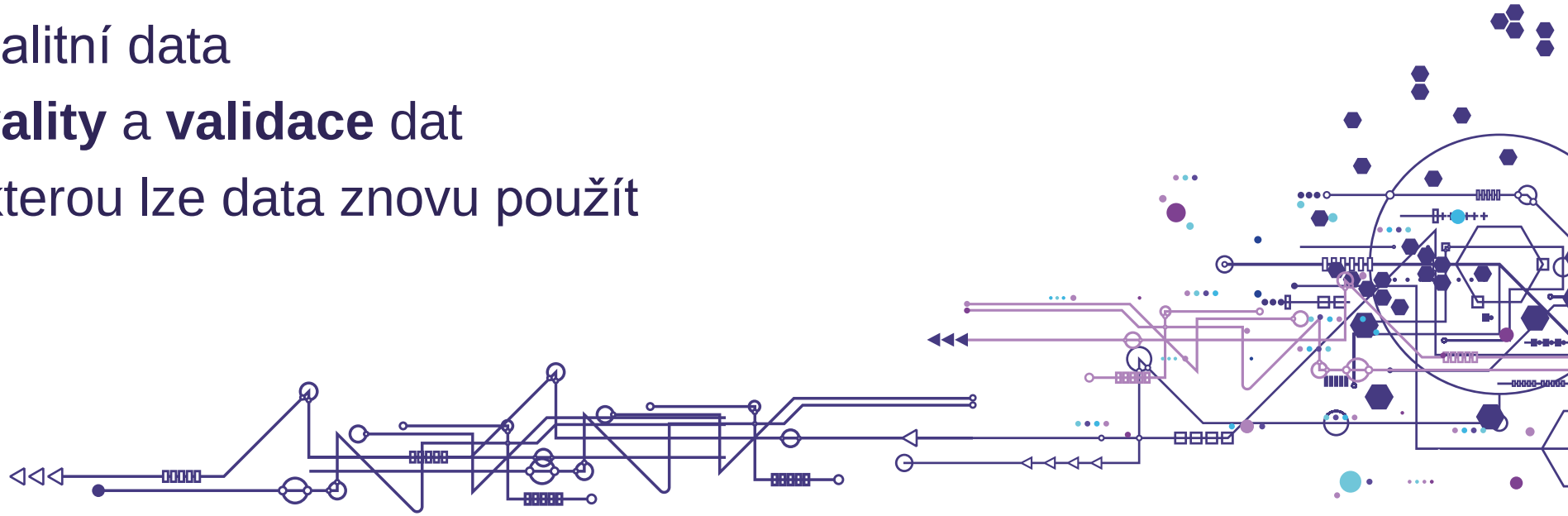
Uchování

- **Dlouhodobé** uchování
  - Nároky na zabezpečení a průběžné konverze formátů
- Kdy **mazat/anonymizovat** uložená data
  - Notifikace z infrastruktury na základě DMP
- Nastavení mechanismu **být zapomenut**
- Zachování metadat i po zániku dat
  - Jasné vymezení, **co jsou metadata a co už data**
  - Perzistentní identifikátory dat i metadat
- Nakládání s **“živými” daty**
  - Sdílená zodpovědnost při nastavování **přístupů** (granularita)
- **Šifrování** a patřičné zabezpečení i **backupů zlatých kopií**

Představa infrastruktury pro správu citlivých dat

Sdílení a opětovné (po)užití

- Vzorový **Data Transfer Agreement**
- Vhodné evidovat **komu a jaká data** byla vydána
  - Měnit identifikátory – jednotlivé řádky i datové sady
  - Každému žadateli poskytnout v souhrnu určité množství dat
- **Vytěžování metadat** ze všech kroků zpracování -> katalogizace, ukládání
- Jak poznám kvalitní data
  - Kontrola **kvality** a **validace** dat
- **Licence**, pod kterou lze data znovu použít



Inspirace I

UK Data Service – 5 safes framework

Researchers are trained and authorised to use data safely.

Data is treated to protect any confidentiality concerns.

A SecureLab environment prevents unauthorised use.

Research projects are approved by data owners for the public good.

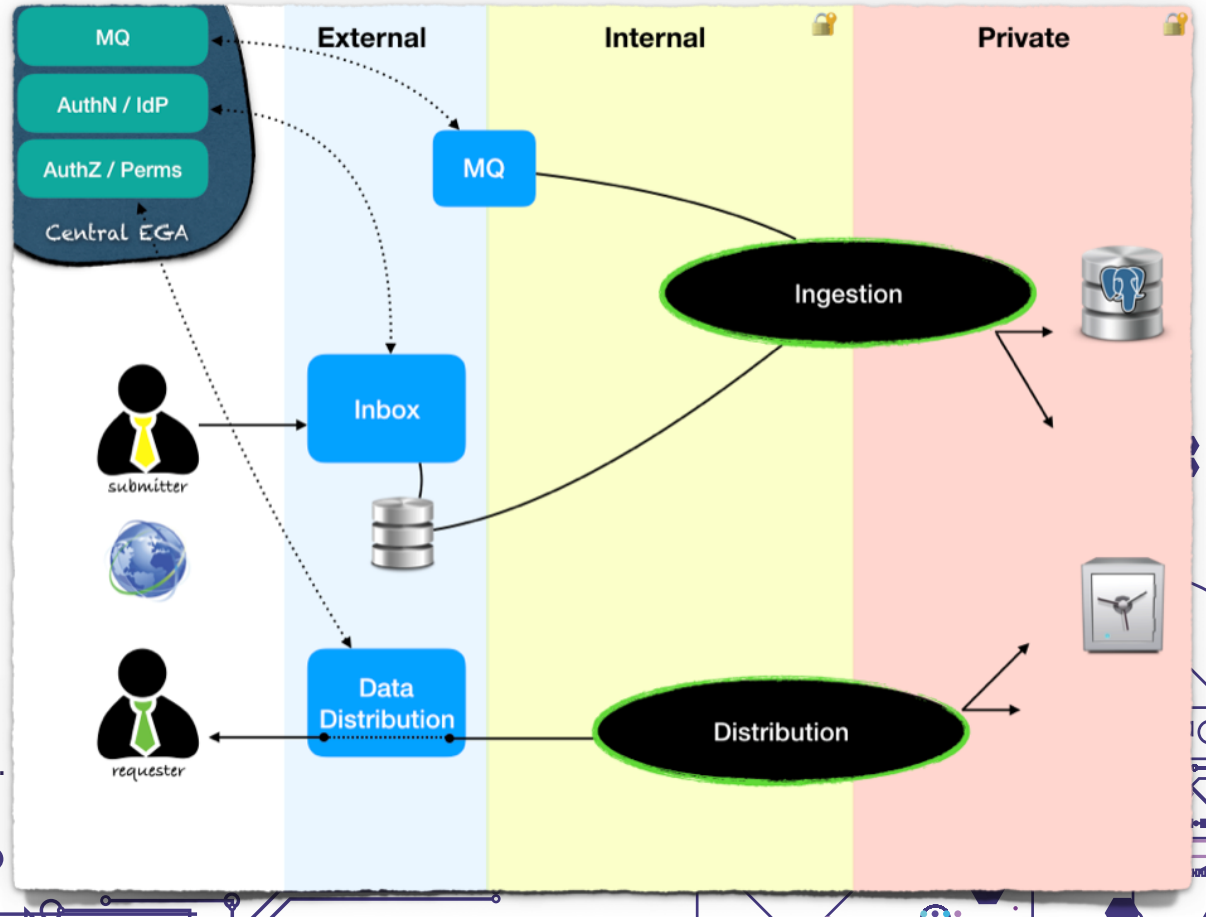
Screened and approved outputs that are non-disclosive.

https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/

Inspirace II

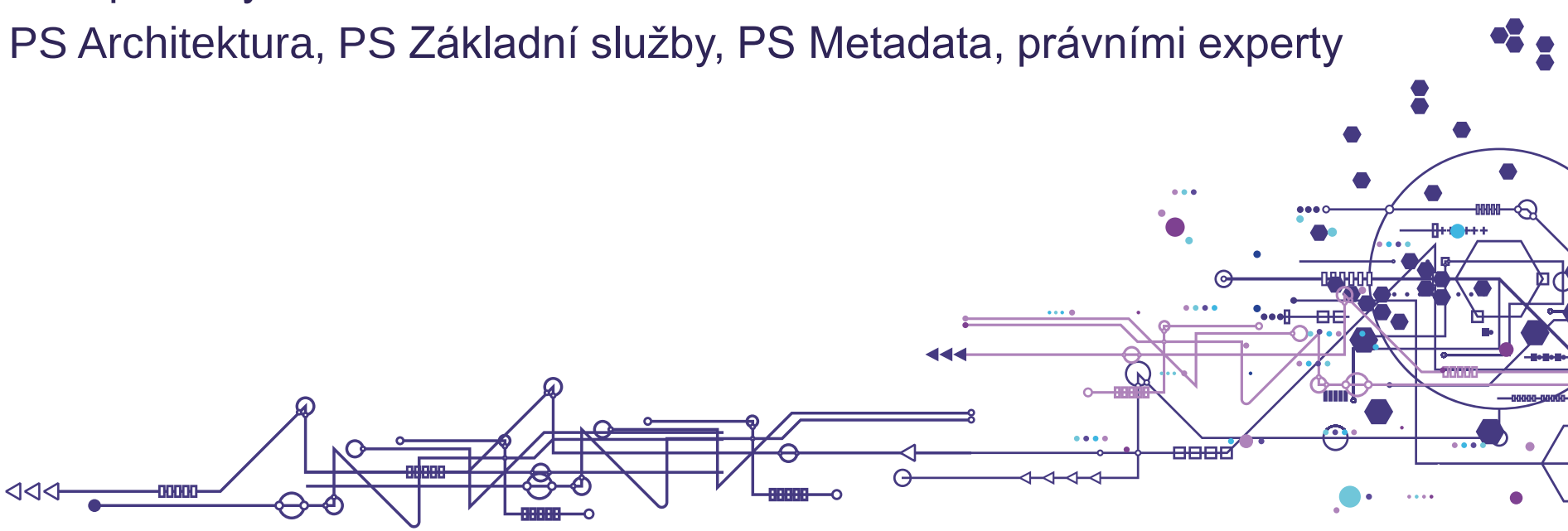
Central & Local European Genome-phenome Archive

- Centrální EGA
  - Databáze uživatel
    - IDs, hesla, ssh public klíče, crypt4gh-compatibilní public klíče
- Lokální EGA
  - Inbox – login a vložení dat
  - Dlouhodobá databáze a úložiště dat
  - Primární zpracování dat (Ingestion pipeline)
    - DB: Databáze PostgreSQL pro ukládání vnitřních částí pipeline.
    - MQ: Zprostředkovatel komunikace (RabbitMQ) mezi příslušnými účty, frontami a vazbami. Používá federovanou frontu k získávání zpráv od CentralEGA, posílá odpovědi zpět.
    - INGEST: Rozdělí hlavičku Crypt4GH, dešifruje soubor, provede kontrolní součty jeho obsahu, přesune data do backendu úložiště.
    - BACKUP: Vytvoří více záloh datové sady.
  - Distribuce dat



Závěrem

- PS funguje od 5/4/2022
- Indentifikujeme **uživatelské potřeby** pro správu citlivých dat
- Hledáme **specifika** pro citlivá data napříč obory
- Pracujeme na **doporučení**, jak by měla infrastruktura pro správu citlivých dat nabízena v rámci EOSC v ČR vypadat
  - Podzim 2022 vznik patřičných dokumentů
- Spolupracujeme s PS Architektura, PS Základní služby, PS Metadata, právními experty



Napište nám

eosc-info@e-infra.cz

Děkuji za pozornost