Věda o datech

Věda o datech

- data scholarship – komplexní množina vztahů mezi daty a vědou
- poprvé řešeno pod názvem „data-intensive research“, politické iniciativy po roce 2000: eScience, eSocial Science, eHumanities → eResearch, eInfrastructure, cyberinfrastructure
- věda – činnost zahrnující to, jak se učíme, uvažujeme o intelektuálních problémech a interpretujeme evidenci
- problémy se sdílením dat: špatně použití, dezinterpretace, odpovědnost, nedostatek expertízy, chybějící nástroje a zdroje, nízká důvěra, ztráta kontroly, znečištění společného fondu dat, udržitelnost
- napětí kolem dat: vlastnictví, kontrola, přístup k datům, obtížnost přenosu dat mezi kontexty a v čase, rozdíly v e formách a žánrech vědecké komunikace, změny technologií, praktik a politik

Pojem data

- první použití pojmu 1646 v teologii, později v matematice. Použití:
- (1) množina principů přijatých jako základ argumentu
- (2) fakta, konkrétně ta vědecká
- v 18. st debatován singulár x plurál
- fakta ve formě vědecké evidence získané z experimentů, pozorování a dalších způsobů bádání
- data nejsou pravdivá, ani nejsou realitou
- fakta, zdroje evidence či principy argumentů používané k prosazování pravdy o realitě, údajná evidence (Buckland)

Pojem data

- data nejsou čisté či přírodní objekty
- existují v kontextu, získávají význam v kontextu, z perspektivy jejich pozorovatele
- míra reprezentovatelnosti kontextu a významu ovlivňuje přenositelnost dat
- **Definice**: Data jsou formalizované a znovu interpretovatelné reprezentace informací vhodné pro komunikaci, interpretaci a zpracování (OAIS)
- datument – otevřená data strukturovaná pro za účelem čitelnosti stroji, která jsou volně dostupná
- The Data Documentation Initiative (DDI) – množina metadatových standardů pro management životního cyklu dat

Typy dat

- data můžeme seskupovat podle:
- ➢ **stupeň zpracování** (EOS DIS – datový informační systém NASA pro systém pozorování země)
- ✓ úroveň 0 – syrová data z přístrojové techniky v plném rozlišení
- ✓ úroveň 1A – metadata k datům v plném rozlišení
- ✓ úroveň 1B – data dělena na sensorové jednotky instrumentů
- ✓ úroveň 2 – přidána další metadata
- ✓ úroveň 3 – sladění datových produktů s časoprostorovými souřadnicemi
- ✓ úroveň 4 – agregace dat do modelů

Typy dat

Data Level	Description
Level 0	Reconstructed, unprocessed instrument and payload data at full resolution, with any and all communications artifacts (e.g., synchronization frames, communications headers, duplicate data) removed. (In most cases, the EOS Data and Operations System (EDOS) provides these data to the data centers as production data sets for processing by the Science Data Processing Segment (SDPS) or by a SIPS to produce higher-level products.)
Level 1A	Reconstructed, unprocessed instrument data at full resolution, time-referenced, and annotated with ancillary information, including radiometric and geometric calibration coefficients and georeferencing parameters (e.g., platform ephemeris) computed and appended but not applied to Level 0 data.
Level 1B	Level 1A data that have been processed to sensor units (not all instruments have Level 1B source data).
Level 2	Derived geophysical variables at the same resolution and location as Level 1 source data.
Level 3	Variables mapped on uniform space-time grid scales, usually with some completeness and consistency.
Level 4	Model output or results from analyses of lower-level data (e.g., variables derived from multiple measurements).

Figure 2.1

NASA EOS DIS Processing Levels. *Credit:* Figure redrawn by Jillian C. Wallis.

Typy dat

- důsledky pro datové kurátorství:
- nejnižší úrovně – potřeba algoritmu zpracování a dokumentace transformace dat na vyšší úroveň
- software pro datovou kolonu – čištění, kalibrace, redukce, kontrola verzí
- nesyrovější přístrojová data příliš objemná k skladování
- kurátorské snahy zaměřeny na nejvíce zpracované produkty

Typy dat

- **původu a hodnoty** (Národní vědecká rada US) – jaká data si zaslouží uchování a na jak dlouho?
- ✓ data z pozorování – rozpoznání, povšimnutí nebo zaznamenání faktů či výskytů fenoménů, obvykle pomocí přístrojů. Nejdůležitější data k uchování – ta nejméně replikovatelná
- ✓ výpočetní data – produkty tvorby počítačových modelů, simulací a průběhu práce . Dokumentace hardwaru, softwaru, stupních či výstupních dat. Někdy uchovány jen výstupy, někdy jen algoritmus.
- ✓ Experimentální – výsledky procedur v kontrolovaných podmínkách. Data může být snadnější replikovat než uchovat

Typy dat

- **původu a hodnoty dat pro komunitu – sbírky**

(Národní vědecká rada US)

- záznamy – 14. stol., svědectví či důkaz faktu, implikuje svědectví, evidenci, důkaz
- Fyzická x digitální sbírky, digitální x digitalizovaná, surogáty x plný obsah, statické obrázky x prohledatelné reprezentace, prohledatelné série x vylepšený obsah
- ✓ Sbírky výzkumných dat – miminální zpracování, kurátorství, nedodržují vždy standardy komunity – formát, struktura
- ✓ Sbírky zdrojových a komunitních dat – standard komunity, přímé financování pro bezprostřední potřeby. Př. PlasmoDB (genomika parazita malárie), Ocean Drilling Program
- ✓ Sbírky referenčních dat – velké, různorodé a distribuované komunity, robustní standardy, zajištěná udržitelnost, řídící struktura. Př. Protein data Bank, astronomická SIMBAD

Malá x velká věda

- **Velká věda** – neviditelná univerzita (komunitní vztahy, výměna informací), internacionální, kolaborativní
- **Malá věda** – heterogenní metody, data, lokální kontrola a analýza

- **Velká data** – data velkého rozsahu (relativní), logistické problémy s manipulací a managementem
- Data jsou velká: množstvím, varietou, rychlostí, kombinací, všudypřítomností
- **Velká x malá data** - Co s nimi lze dělat? Co mohou odhalit? Jaký rozsah analýzy potřebují?

Důsledky trendu otevřenosti

- otevřené modely – software, správa, standardy, publikace, data, služby, spoluprací produkované znalostí
- změny mezi podílníky ve všech sférách
- podporuje informační tok (př. vědecký obsah), modularitu systémů a služeb, interoperabilitu
- ekonomické a sociální výdaje – free software, open access
- propojení komerční sféry (výzkumná data s komerční hodnotou) a akademického výzkumu (komerční data pro akademické účely)

Důsledky trendu otevřenosti

- změna politiky výzkumu – vláda, financující agentury a časopisy podporují sdílení vědeckých dat
- proti komodifikaci informačních zdrojů → vznik nových trhů:

✓ zdravotní dokumentace
✓ nákupní chování spotřebitelů
✓ sociální média
✓ vyhledávání informací
✓ vědecké publikování
✓ genomika

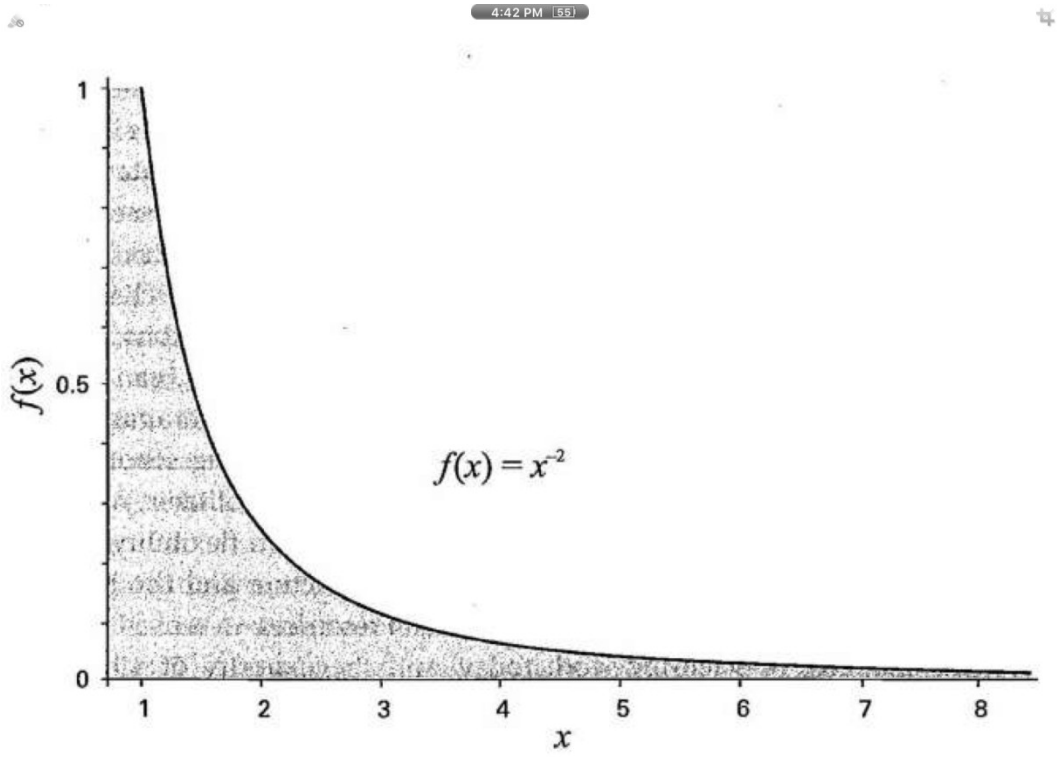
Dlouhý chvost

- Chris Anderson: statistická distribuce: mocninný zákon
- dostupnost a používání dat ve výzkumném a ekonomickém sektoru
- malé množství vědeckých týmů pracuje s velkým množství dat

  (astronomie, fyzika, genomika,

  makroekonomie, digital humanities)
- velké množství týmů má málo dat

Dlouhý chvost



- velká data – nejen málo vědců, ale i nízká varieta

Důsledky pro vědu

- **velká věda** – generování velkého množství dat vyžaduje:
- společný instrument (teleskop, DNA sekvencer)
- společný formát (metadata, databázová struktura)
- velká, distribuovaná, mezinárodní spolupráce týmů
- komunitní repozitáře velkých dat
- **malá věda** – větší varieta obsahu, struktury, reprezentací, výzkumné metody a sbírky flexibilnější k aktuálním problémům
- nedostatek standardů pro sdílenou infrastrukturu, chybějící kritické množství pro vývoj sdílených zdrojů dat
- jedinci a malé skupiny - většina vědecké práce, minimální financování
- explorativní, lokální, různorodý výzkum, málo sdílených komunitních zdrojů

No data

- na data bohaté obory: spojené datové zdroje, společná metoda, nástroje, infrastruktura
- na data chudé obory: ceněno vlastnictví dat, ovlivňuje volbu metod a teorii
- důvody neexistence či nedostupnosti dat:

➢ **data nejsou dostupná**: entity nemusejí nebo podle zákona nesmí uvolnit data (obchodní, vzdělávací, muzejní kurátorská dokumentace, patentový proces, data o lidském subjektu apod.)

➢ **data nejsou zveřejněná**: období embarga, ochrana vzácných zdrojů, vzorků, kulturního dědictví

- př. trivializace sběru a skladování dat v chemii.

No data

- Mechanismus uvolnění: příspěvek do komunitního archivu, podpůrné materiály v časopisu, zaslání na webovou stránku, uvolnění po požadavku
- **data nejsou použitelná**: chybí motivace věnovat datům úsilí – informace potřebné k interpretaci dat: vztah ke zkoumanému problému, výzkumné doméně, expertíze
- seznamy kódů, modely, popis metod sběru, čištění, analýzy, dostupnost softwaru či statistických nástrojů, informace o původu a transformacích dat, zastaralý hardware

- Teorie, evidence a postupy jsou vzájemně hluboce propojeny

Rozdíly kategorií

- Rozdíl mezi typy dat – lidský vynález: rozhodnutí o kritériích a jménu každé kategorie – století vyjednávání
- ➢ Syrová x zpracovaná data: syrová data – relativní pojem, závisí na tom, kde začíná zkoumání
- Přístroje – designovány a projektovány k detekci určitého fenoménu za určitých podmínek → určuje co může být detekováno. Detekce nejsyrovějších dat – nekonečný regres epistemologických voleb
- ➢ Primární x sekundární data: primárním i reprezentace originálu, který již neexistuje, editované kompilace vylepšující čitelnost originálu
- Co je primární závisí na kontextu a startovní pozici. Sekundární zdroj pro jednoho může být primárním pro druhého
- Entita se může stát daty pouze když ji někdo použije jako evidenci fenoménu

Jednotka dat

- jednotka dat – tradiční papírové jednotky – knihy a články, dnes lze rozdělit na mnohem menší jednotky.
- vhodná jednotka závisí na zamýšleném využití – jednotka vhodná k šíření, citování, použití a kurátorství.
- data reprezentována v jednotkách různé velikosti: pixely, fotony, znaky či jejich segmenty, písmena, slova, buňka v tabulce, množina dat, datový archiv
- vědecký obsah je stále více atomizovaný a zpracovávaný ve formě dat

Vznik dat

- pozorování je třeba nějakým způsobem reprezentovat jako data, činnosti dokumentovány jako entity na papíře či v počítači
- vloženy rovnou do analytického nástroje nebo transkribovány pro pozdější vyhodnocení
- odpovědi mohou být kombinovány s poznámkami
- kontextuální informace – podmínky v čase pozorování
- pozorování – selektivní, záznam detailů podle zájmů studie, ignorace ostatních detailů
- výzkumný kontext ovlivňuje, co bude pokládáno za data a jak budou zaznamenána a reprezentována. Vztah mezi znalostí a kontextem.

Původ dat

- původ – problémový pojem, vzdálenost od vzniku dat, tj. prvního zacházení s něčím jako s daty
- roli mohou hrát: čas, kontext, metoda, teorie, jazyk, expertíza
- čím blíže vědecká práce původu dat, tím méně interpretace závisí na formální reprezentaci
- osobní a důvěrný vztah ke svým datům mají etnogtrafové: např. ručně psané poznámky nejsou reprezentovány způsobem, který by mohli interpretovat ostatní
- opačný extrém: vědci pracující s množinami dat posbíranými jinými vědci, např. demografové
- datoví analytici mají málokdy důvěrnou znalost jak, kdy a proč byla data shromážděna

Prameny x zdroje

- prameny x zdroje – rozdíl mezi novými a existujícími daty
- prameny – data vznikající během výzkumného projektu
- výzkumník má větší kontrolu nad prameny než nad zdroji
- zdroje – závisí na zpřístupnění dalšími stranami a právy k jejich zveřejnění, publikování a znovuvyužití
- záznamy nevytvořené pro výzkumné účely mohou být také zdroji
- dnes vědci tvoří nová data zachycením informací, které nemohly být zachyceny dříve: zakódování ručně psaných vzorů, extrakce textů z historických materiálů pomocí algoritmů pro optické rozeznání znaků apod.

Metadata

- metadata jsou strukturované informace popisující, vysvětlující, lokalizující či jinak zjednodušující vyhledání, využití nebo správu informačních zdrojů (NISO - National Information Standards Organization)
- typy metadat:
- NISO: deskriptivní, strukturální, administrativní
- archiváři: deskriptivní, administrativní, prezervační, technická (dokumentace hardwaru a softwaru, autentizační data a další systém specifikující informace), metadata používání (sledování uživatele, znovupoužití obsahu, informace o verzích)

Metadata

- hodnocení vědeckých metadat: dle abstrakce a rozšířitelnosti schématu, flexibility, modularity, úplnosti, dostatečnosti, jednoduchosti, výměny, vyhledávání, publikování a archivování dat
- zda je konkrétní jednotka informace považována za data či metadata záleží na tom, co je popisováno a za jakým účelem.
- Data jedné osoby mohou být metadata druhé osoby: př. bibliografie, anonymizované informace o obsahu telefonátů (telefonní čísla komunikujících, výrobní číslo telefonu, délka a trvání hovoru) – metadata pro datový model komunikační sítě x osobní data

LITERATURA

- BORGMAN, Christine L. – Big data, Little data, No data: Scholarship in the Networked World. Cambridge: MIT, 2015. ISBN 978-0-262-02856-1.
- PRICE, Derek John de Solla. *Little Science, Big Science*. 2. ed. New York: Columbia University Press, 1965.