Juhaňák, Libor

Proces analýzy a typy využívaných dat

In: Juhaňák, Libor. *Analytika učení a data mining ve vzdělávání v kontextu systémů pro řízení výuky.* Vydání první Brno: Masarykova univerzita, 2023, pp. 35-48

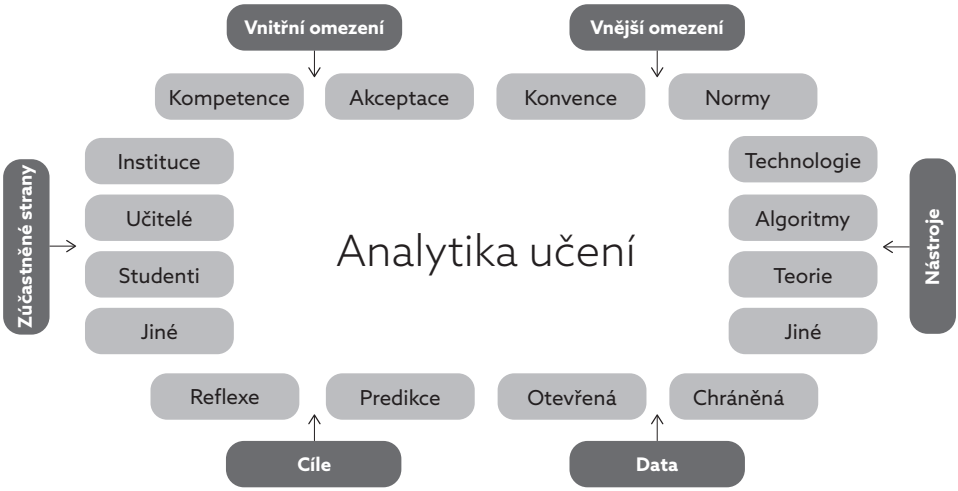
ISBN 978-80-280-0184-1; ISBN 978-80-280-0185-8 (online ; pdf)

Stable URL (handle): https://hdl.handle.net/11222.digilib/digilib.77692
Access Date: 24. 02. 2024
Version: 20230228

4 PROCES ANALÝZY A TYPY VYUŽÍVANÝCH DAT

V této kapitole je věnována podrobnější pozornost tomu, jak lze nahlížet na proces aplikace analytiky učení a data miningu ve vzdělávání a jaké typy dat jsou v těchto oblastech nejčastěji využívány, resp. analyzovány. Před tím je však třeba zdůraznit, že jak data mining ve vzdělávání, tak i analytika učení obvykle vyžadují zapojení několika odlišných oblastí expertízy, což klade důraz nejen na zapojení velkého množství expertů z různých oblastí, nýbrž také na spolupráci několika jinak oddělených součástí (či oddělení) v rámci jedné instituce. Při využití analytiky učení či data miningu ve vzdělávání v rámci určité vzdělávací organizace je tedy třeba zohlednit řadu důležitých faktorů. Před přesunutím pozornosti přímo k procesu aplikace a k analyzovaným datům tak bude nejprve představen obecný rámec analytiky učení (viz Greller a Drachsler, 2012) zahrnující šest různých dimenzí, kterým musí být v rámci vzdělávací instituce věnována pozornost, má-li být dosaženo efektivního zapojení a využívání analytiky učení (či data miningu ve vzdělávání)[19].

19 Citovaní autoři sice o zmiňovaném rámci hovoří pouze v souvislosti s analytikou učení, v naprosté většině aspektů však můžeme tento rámec vztáhnout i na data mining ve vzdělávání.



Obrázek 3: Obecný rámec zahrnující šest základních dimenzí analytiky učení (podle Greller & Drachsler, 2012)

Nejprve můžeme hovořit o cílech využití analytiky učení v rámci dané vzdělávací instituce (srov. obrázek 3 výše). Smyslem analytiky učení je dle Grellera a Drachslera (2012) odhalování skrytých informací a souvislostí ve vzdělávacích datech a jejich následné zpracování pro různé zúčastněné strany. Autoři přitom rozlišují dva základní cíle: reflexi a predikci. Reflexí je míněna především sebe-evaluace. Analytika učení je zde prostředkem pro získání lepšího vhledu do vlastního jednání. Například můžeme pomocí určitého nástroje vizualizovat chování studentů či učitelů v online kurzu, což jim dá možnost reflektovat, zda jejich učení, resp. vyučování probíhá podle jejich představ a cílů. Predikce se pak nachází ještě o krok dál, kdy analytika učení spočívá v modelování chování studentů, což umožňuje např. cíleně upozorňovat učitele na ty studenty, kteří mohou mít na základě odpovídajícího modelu určitý problém se studiem. Učitel tak má podklady k tomu, aby mohl např. s daným studentem pracovat individuálně, nebo mu nabídnout jinou formu podpory při učení, a zabránit tak neúspěchu studenta na konci kurzu.

Za další dimenzi můžeme považovat různé zúčastněné strany. U analytiky učení lze vždy rozlišovat několik zúčastněných stran. Od studentů přes učitele až po vedení instituce či jiné zúčastněné strany (např. investory, donátory, grantové agentury). Sclater (2017) vyjmenovává až 11 různých skupin tzv. stakeholderů, kteří do analytického procesu vstupují. Vedle výše uvedených k nim podle Sclatera patří také:

- IT podpora (*IT Services*). Zajišťují správu vzdělávacího systému, mají na starost vývoj či přizpůsobení (*customization*) technického řešení pro sběr, správu a vizualizaci dat apod.

- Tvůrce kurzů (*Learning designer*). Podílí se na designu kurzů a na tom, jakým způsobem se v nich bude promítat analytický nástroj.
- Datový analytik (*Data scientist*). Spolupracuje na návrhu a vývoji analytického systému. Provádí analýzy sbíraných dat, evaluuje fungování analytického systému, podílí se na tvorbě prediktivních modelů, zajišťuje validování modelů apod.
- Výzkumník v oblasti pedagogiky (*Educational researcher*). Do procesu návrhu a vývoje analytického systému, stejně jako následně do jeho evaluace může zasahovat rovněž odborník na pedagogický výzkum. Stejně tak se může podílet i na návrhu formy intervence do edukační reality na základě realizovaných analýz.
- Tutor kurzu (*Tutor*). Zajišťuje podporu studentům přímo v kurzu, monitoruje jejich postup kurzem, případně může navrhovat vylepšení kurzu či analytického nástroje na základě jeho využívání studenty.

To, které zúčastněné strany jsou v konkrétním případě zapojeny, samozřejmě následně ovlivňuje zbývající dimenze.

V závislosti na stanovených cílech a zúčastněných stranách je v rámci další dimenze nutné věnovat pozornost získávaným a analyzovaným datům. Greller a Drachsler (2012) přitom rozlišují data na otevřená a chráněná, v čemž můžeme spatřovat jejich důraz na etické aspekty práce s daty. V souvislosti s daty je však nutné věnovat pozornost i tomu, jaké typy dat jsou vlastně v rámci instituce k dispozici, jaká další data by v případě potřeby bylo možné dále sbírat, jak jsou data zpracovávána a uchovávána apod.

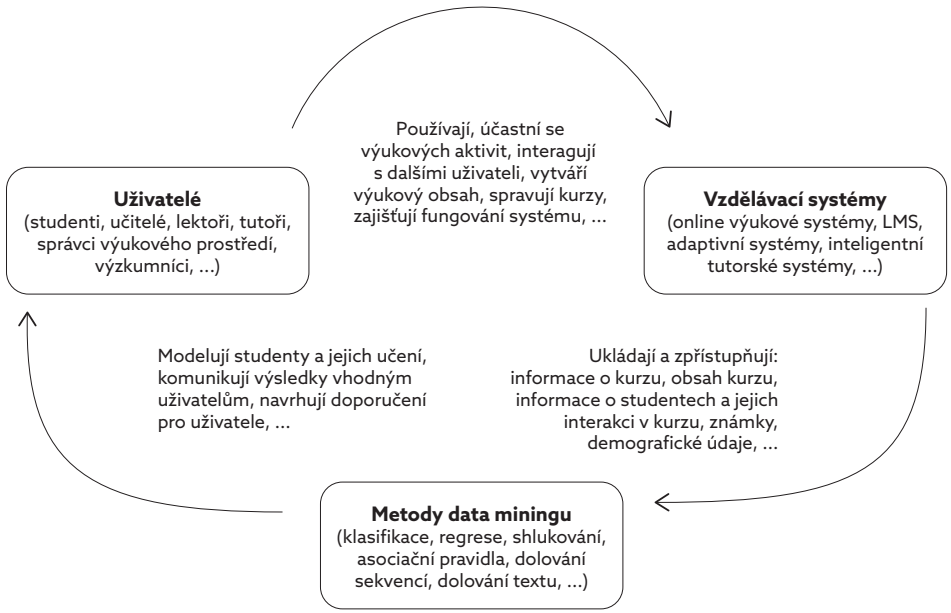
Od dostupných dat se pak odvíjí použité nástroje či prostředky. Mezi ně autoři zahrnují nejen samotné technologie a konkrétní technická řešení, ale také využité algoritmy, metody analýzy, a dokonce i celé teorie či teoretické přístupy. Některé konkrétní nástroje analytiky učení jsou zmíněny a podrobněji popsány v kapitole 6.1, metody a techniky analýzy dat využívané jak v analytice učení, tak i v data miningu ve vzdělávání jsou představeny v kapitole 5.

Poslední dvě dimenze v modelu Grellera a Drachslera (2012) tvoří vnitřní a vnější omezení. Vnitřní omezení vycházejí z možností, limitů a zdrojů instituce, která chce analytiku učení využívat. Zásadní je zde oblast kompetencí, protože kvůli své multidisciplinární povaze vyžaduje analytika učení pro své efektivní fungování nové typy kompetencí (především v souvislosti s moderními technologiemi – typicky programování nebo pokročilá analýza dat). Nedílnou součástí je pak i akceptace analytiky učení napříč zúčastněnými stranami. Vnější omezení pak autoři modelu dělí na konvence, které zahrnují především etiku, a normy, jež spojují s právem a konkrétní legislativou. Vzdělávací instituce se totiž vždy pohybují v určitém společenském a právním kontextu, na který musí brát ohled. Při využití analytiky ve vzdělávacím kontextu je přitom specifické to, že kvůli novosti problematiky nebylo zatím mnoho etických či právních otázek ani položeno, natož zodpovězeno (srov. kapitolu 6.7).

4.1 Proces aplikace analytiky učení a data miningu ve vzdělávání

Na proces aplikace data miningu ve vzdělávání se můžeme dívat vícero způsoby. V dosavadní odborné literatuře zaměřující se na tuto oblast lze však rozlišit dva základní způsoby konceptualizace procesu aplikace data miningu v kontextu vzdělávání. V prvním případě vychází konceptualizace z chápání data miningu ve vzdělávání jakožto formativní evaluační techniky (viz Romero & Ventura, 2010), jejímž cílem je zlepšení učení studentů v rámci určitého vzdělávacího systému či programu. Metody a techniky data miningu zde pak slouží především k tomu, aby díky nim bylo možné získat důležité informace o tom, jakým způsobem studenti systém používají. Tyto informace pak mohou být využity tvůrci daného systému či vzdělavateli v daném systému jako jeden z podkladů pro rozhodování o případných úpravách systému či změnách ve způsobech jeho využívání. Proces aplikace data miningu ve vzdělávání je pak v tomto smyslu chápán jako iterativní cyklus, který je znázorněn na následujícím obrázku (viz obrázek 4).

Jak je z obrázku patrné, můžeme uvažovat o různých typech uživatelů (studenti, učitelé, administrátoři kurzů) a dalších subjektech zainteresovaných v procesu vzdělávání. Tyto zúčastněné strany jsou zároveň uživateli určitého vzdělávacího systému (příp. i více systémů současně). Tímto systémem může být např. online



Obrázek 4: Konceptualizace procesu aplikace data miningu ve vzdělávání jakožto formativní evaluační techniky (podle Romero, Ventura, Pechenizkiy, & Baker, 2010)

systém pro řízení výuky (LMS), systém typu ITS aj.[20] Každá ze zúčastněných stran však využívá daný vzdělávací systém jinými způsoby. V případě LMS tak např. studenti participují v online kurzech, pracují se studijními materiály, plní výukové aktivity apod., učitelé vytváří studijní obsah kurzu a vedou kurzy, administrátoři systému zajišťují jeho provoz apod.

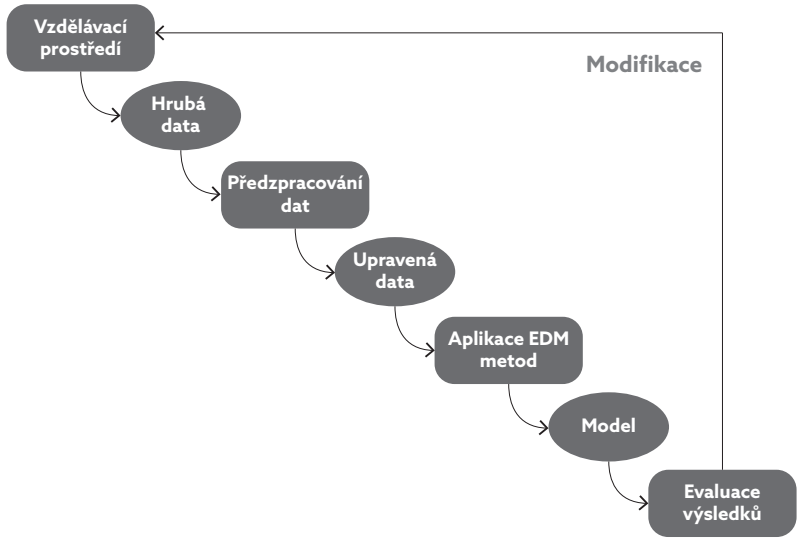
Různé typy vzdělávacích systémů přitom sbírají různé typy dat, která jsou analyzovatelná s využitím data miningových metod a technik. Může jít např. o data týkající se obsahu kurzů či o tom, jak s obsahem studenti interagují, data o vzájemné komunikaci studentů, o získaných známkách apod. Tato data pak mohu být ze vzdělávacích systémů extrahována a mohou na ně být aplikovány různé data miningové techniky jako shlukování, klasifikace, dolování z textu, dolování dat s využitím asociačních pravidel a další (podrobněji viz kapitolu 5). Výsledky získané aplikací data miningových metod jsou následně prezentovány jeho (různým) uživatelům. Tak může být studentům například zobrazeno personalizované doporučení určitých studijních materiálů, učitelé naopak mohou být informováni o studentech, kteří mají v kurzu nějaké problémy nebo se v kurzu chovají nějakým nestandardním způsobem apod. Tím se zároveň uzavírá jeden běh cyklu a může začít běh nový. Na základě zobrazených výsledků se totiž uživatelé ve vzdělávacích systémech opět nějakým způsobem chovají, což je znovu zaznamenáváno a může být následně analyzováno.

Druhý způsob konceptualizace data miningu ve vzdělávání vychází do značné míry z obecného procesu dobývání znalostí z databází, jak je obvykle prezentován v odborné literatuře zaměřené na data mining obecně (viz např. Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Jak je patrné ze znázornění procesu na obrázku níže (viz obrázek 5), je zde pozornost primárně zaměřena na samotný proces práce s daty. V určitém smyslu lze tento způsob konceptualizace procesu data miningu chápat jako úžeji zaměřený, než je způsob popisovaný výše.

Níže znázorněné fáze procesu aplikace data miningu ve vzdělávání lze podrobněji popsat následovně (srov. Bousbia & Belamri, 2014; Romero & Ventura, 2013):

1. Vzdělávací prostředí. V první fázi začínáme s určitým vzdělávacím prostředím, v jehož rámci mohou být sbírány určité typy dat. Opět může jít o poměrně různorodá vzdělávací prostředí od online vzdělávacích systémů přes různé administrativní systémy až po prostředí školní třídy. V závislosti na typu prostředí jsou pak dostupné různé typy dat. Tato data jsou však v jednotlivých vzdělávacích systémech obvykle ukládána v takové podobě, že na ně nelze přímo aplikovat metody data miningu. Jde tedy o tzv. hrubá data (*crude data* či *raw data*), která musí být nejprve určitým způsobem zpracována, aby bylo možné jejich další využití.

20 V širším smyslu pak lze za systém považovat i tradiční výuku ve školní třídě a jakékoli další uspořádání výuky, ve kterém nejsou pro komunikaci a interakci využívány (pouze) elektronické prostředky.



Obrázek 5: Konceptualizace procesu aplikace data miningu ve vzdělávání inspirovaná obecným procesem dobývání znalostí z databází (podle Bousbia & Belamri, 2014)

2. Předzpracování dat. Další fází je tzv. předzpracování dat. Tím jsou myšleny úpravy dat, které umožní transformovat původní hrubá data do takové podoby, aby na ně mohly být aplikovány vybrané data miningové techniky. Nutno podotknout, že v kontextu vzdělávání je tato fáze často velmi náročná a zároveň zásadně důležitá. Data ve vzdělávacích systémech se obvykle vyskytují v různých úrovních granularity (viz níže), jsou ukládána v různých formátech, mají víceúrovňovou strukturu apod. Zároveň je ve vzdělávacích systémech k dispozici velké množství dat různých typů, ve kterých je třeba se nejprve zorientovat a vybrat pouze ta data, která jsou podstatná pro řešení daného problému či zodpovězení položené otázky. Výsledkem této fáze jsou upravená či transformovaná data, jež jsou nachystaná pro aplikaci data miningových metod.
3. Aplikace data miningových metod. V rámci aplikace konkrétních data miningových metod na připravená data může jít o využití obecných data miningových metod, které jsou použitelné i v jiných oblastech (např. shlukování, klasifikace apod.). Často je však zapotřebí zohlednit právě kontext vzdělávání, jenž může mít určitá specifika a může vyžadovat využití specifických data miningových metod. Může jít např. o metody zohledňující hierarchickou povahu dat v kontextu vzdělávání či o metody modelování longitudinálních dat apod. Zároveň je ale třeba říci, že data mining ve vzdělávání je i dnes poměrně mladou a rozvíjející se výzkumnou oblastí, tudíž je porozumění specifickým vlastnostem vzdělávacího kontextu

a vývoj odpovídajících přizpůsobených či zcela nových data miningových technik stále jedním z primárních cílů výzkumníků v této oblasti. Výsledkem aplikace data miningových technik je obvykle nějaký model. Může jít např. o model, podle kterého jsou studentům doporučovány další studijní materiály k rozšíření znalostí, či model predikující úspěšnost studentů v kurzu apod.

4. Evaluace výsledků. Závěrečnou fází data miningového procesu je pak evaluace či interpretace vytvořených modelů. Tato fáze je zcela zásadní pro následné uplatnění modelů a získaných poznatků ve vzdělávací praxi (tj. pro účely rozhodování, pro účely modifikace systému atd.). Uživatelé výstupů modelování, ať už to jsou přímo studenti a učitelé, či jiné zainteresované subjekty, totiž musí být schopni výstupům porozumět, aby mohli zvolit odpovídající reakci[21]. Aplikací vytvořených modelů do vzdělávacího prostředí zároveň dochází k úpravě či modifikaci tohoto prostředí, čímž se uzavírá celý cyklus, a může tak začít jeho nový běh.

Jak je z výše uvedeného patrné, v pozadí obou konceptualizací procesu data miningu ve vzdělávání je rozpoznatelný obecný iterativní proces výzkumné práce založený na formulování hypotéz, jejich testování, resp. ověřování a následném vyhodnocení a formulování hypotéz nových (Romero & Ventura, 2013). Zároveň, jak bude patrné z následujících pasáží, obdobným způsobem jako data mining ve vzdělávání je konceptualizován i proces analytiky učení.

Již bylo naznačeno, že analytika učení, podobně jako akční výzkum, bývá konceptualizována v podobě opakujícího se cyklu sestávajícího z několika odlišných fází. Doposud však není obecně přijímáno jedno konkrétní pojetí tohoto analytického cyklu (*learning analytics cycle*). Jednotlivé fáze rozlišované v rámci procesu analytiky učení se tak u různých autorů liší. Například Campbell a Oblingerová (2007) představili cyklus akademické analytiky v pěti krocích zahrnujících získávání dat, informování, predikování, jednání a následné zlepšování. Chatti et al. (2012) později zúžili počet etap jen na tři, zatímco Siemens (2013) uvádí cyklus až o sedmi fázích, v němž podrobněji rozlišuje různé fáze zpracování dat předcházející samotné analýze. Lal (2014) se pak pokouší o určité sjednocení předchozích přístupů a uvádí fází šest. Srovnání zmíněných pojetí procesu analytiky učení je možné vidět v tabulce 3.

21 Doplňme, že interpretace data miningových modelů je do značné míry obecným problém. Některé data miningové techniky sice vedou k dobrým výsledkům, co se týče úspěšnosti predikce, ale na druhou stranu produkují modely, které jsou jen obtížně interpretovatelné (někdy se používá označení *black-box models*; typickým příkladem takové techniky jsou neuronové sítě). V některých případech proto může být preferován takový model, který má sice horší predikční schopnosti, ale je velmi snadno interpretovatelný, a tudíž srozumitelný (tzv. *white-box model*; příkladem může být modelování s využitím techniky rozhodovacích stromů). Velmi užitečnou publikací je v tomto kontextu volně dostupná kniha Molnara (2019).

Tabulka 3: Srovnání různých konceptualizací procesu analytiky učení

<table>
<tr><th>Campbell & Oblinger, 2007</th><th>Chatti et al., 2012</th><th>Siemens, 2013</th><th>Lal, 2014</th></tr>
<tr><td rowspan="4">získávání</td><td rowspan="4">sběr dat a předzpracování (pre-processing)</td><td>sběr a akvizice</td><td rowspan="2">získávání dat</td></tr>
<tr><td>ukládání</td></tr>
<tr><td>čištění</td><td rowspan="2">strukturování a agregování dat</td></tr>
<tr><td>integrace</td></tr>
<tr><td>informování</td><td rowspan="4">analýzy a akce</td><td rowspan="2">analýza</td><td rowspan="2">analyzování</td></tr>
<tr><td rowspan="2">predikování</td></tr>
<tr><td>zobrazení a vizualizace</td><td>zobrazování a vizualizace</td></tr>
<tr><td>jednání</td><td rowspan="2">akce</td><td>akce</td></tr>
<tr><td>zlepšování</td><td>následné zpracování (post-processing)</td><td>zlepšování</td></tr>
</table>

Vzhledem k tomu, že zatím žádná z konceptualizací není všeobecně přijímána, můžeme cyklus analytiky učení při určitém zobecnění popsat ve čtyřech základních krocích:

1. Sběr a zpracování dat samozřejmě záleží na tom, co je v daném případě naším cílem a jaké otázky si pokládáme. Campbell a Oblingerová (2007) uvádějí poměrně rozsáhlý výčet typů dat, která mohou být ve vzdělávacích institucích dostupná. Od demografických dat přes výsledky studia až po chování žáků či studentů v konkrétních kurzech. Pozornost je však nutné věnovat i tomu, která data zatím sbírána nejsou, ale mohla by být pro naše účely potenciálně užitečná. Získaná data je následně nutné předzpracovat a případně převést do potřebných formátů. Stěžejní roli zde hraje čištění dat. Pro komplexnější pohled na zkoumaný fenomén může být využita integrace dat z několika různých zdrojů (Siemens, 2013).
2. Analýza a vizualizace je druhou fází analytického procesu. V této fázi jsou používány různé metody a techniky ke zkoumání získaných dat za účelem objevení potenciálně užitečných informací a jejich následné prezentace (často v podobě vizualizace) zúčastněným stranám. Co se týče využívaných metod, Lal (2014) je v kontextu analytiky učení rozděluje do tří hlavních oblastí: základní statistické metody, data miningové metody, kde jde především o klasifikaci, klastrování a asociační pravidla, a metody sociální analytiky učení, kam lze zařadit např. analýzu sociálních sítí, obsahovou analýzu či analýzu diskurzu.
3. Akce zde znamená zásah do edukační reality, jenž je proveden určitou zainteresovanou stranou na základě výsledků analýz, které jí byly prezen-

továny. Akce přitom může mít nejrůznější charakter. Může jít o jednoduché oznámení, upozornění či varování, ale také o komplexnější formu intervence do současné podoby edukační reality. Tak může být student například automaticky upozorněn na to, že v porovnání s ostatními je výrazně pozadu. Nebo může být naopak upozorněn učitel, který následně zvolí vhodný způsob intervence. Akcí může být také optimalizace stávajícího výukového systému či jiné systémové úpravy a vylepšení. V případě adaptivních systémů lze za akci považovat přizpůsobení či personalizaci, jež jsou vykonány na základě výsledků provedených analýz. U doporučovacích systémů lze za akci považovat automatické doporučení vhodného rozšiřujícího studijního materiálu (Chatti et al., 2012; Lal, 2014; Siemens, 2013).

4. Reflexe a revize označují poslední etapu cyklu, kdy probíhá zhodnocení či evaluace provedené akce a zároveň dochází k plánování dalšího cyklu. Může tak jít např. o sběr jiných dat nebo doplnění dat stávajících. Stejně tak může jít o provedení doplňujících analýz, volbu jiných analytických metod či vykonání jiných akcí na základě provedených analýz (srov. Campbell & Oblinger, 2007; Chatti et al., 2012).

4.2 Typy dat využívaných v oblasti analytiky učení a data miningu ve vzdělávání

Jak bylo již naznačeno výše, při aplikaci data miningu ve vzdělávání či analytiky učení[22] se lze setkat s různými typy dat. To, o jaká data se jedná, zcela zásadně ovlivňuje, které data miningové či analytické postupy a techniky lze použít. Podobně také typ dostupných dat ovlivňuje to, jaké pedagogické otázky lze s využitím těchto metod vůbec zodpovědět. K vytvoření typologie využívaných dat je však možné využít poměrně různorodá kritéria. Níže proto není podána jediná typologie dat, nýbrž je představeno několik odlišných způsobů, jakým jsou obvykle rozlišovány různé typy dat v kontextu data miningu ve vzdělávání a analytiky učení.

Do jisté míry výchozím kritériem pro rozlišení různých typů dat může být obsah, který data reprezentují. Zde lze využít typologii Sclatera (2017), již sice představuje primárně ve spojení s analytikou učení, ovšem je dobře využitelná i v souvislosti s data miningem ve vzdělávání. Sclater rozlišuje čtyři základní kategorie dat. První kategorií jsou základní demografické údaje, které škola o svých

22 Na tomto místě je vhodné podotknout, že co se týče typů využívaných, resp. analyzovaných dat, je možné obsah této části považovat za platný jak pro data mining ve vzdělávání, tak i pro analytiku učení. V obecném pohledu lze totiž říci, že se obě tyto výzkumné oblasti zaměřují na stejné typy dat. A to i přes to, že při bližším pohledu mohou být jednotlivé typy dat v daných oblastech akcentovány v odlišné míře.

studentech uchovává. Sem mohou spadat údaje jako datum narození, pohlaví, bydliště či místo narození, rodinný stav apod. Druhou kategorií jsou tzv. akademická data, která zahrnují údaje týkající studia jednotlivých studentů. Jde tak např. o volbu předmětů, známky, resp. hodnocení z absolvovaných předmětů, odevzdané úkoly či závěrečné práce apod. Spadají sem údaje o „cestě" studenta studiem (tj. např. volba předmětů) a o jeho pokrocích (tj. absolvování předmětů) i data reprezentující obsah generovaný studenty v průběhu plnění studia či dílčích kurzů (tj. seminární práce, eseje apod.). Třetí kategorií jsou data o učební aktivitě studentů. Zde má Sclater na mysli primárně data z různých typů online výukových systémů (např. LMS), jež jsou obvykle generována automaticky v podobě tzv. logů. Poslední typ dat pak lze na základě Sclatera nazvat daty o vzdělávacím kontextu. Sem spadají všechny údaje, které mohou poskytovat potřebný kontext pro výše uvedené typy dat. Může jít např. o detailní informace o zamýšleném kurikulu či studijním plánu, jež by poskytovaly nezbytný kontext pro analýzy zaměřující se na volbu jednotlivých předmětů a studijních cest studentů. Jiným příkladem mohou být informace doplňující kontext k jednotlivým kurzům, např. trvání kurzu, způsob ukončení kurzu, studijní materiály a aktivity v kurzu včetně plánovaného termínu jejich splnění apod. Taková „kontextová" data mohou být relevantní zvláště v případech, kdy je věnována pozornost predikci úspěšnosti studentů v jednotlivých kurzech.

Za další ze základních kritérií pro rozlišení různých typů dat lze považovat vzdělávací prostředí či systém, ze kterého data pocházejí (srov. např. Bousbia & Belamri, 2014). Na jedné straně zde máme různé typy počítačem podporovaného vzdělávání, kde může jít o data pocházející ze systémů pro řízení výuky (LMS), ale i jiných typů virtuálních vzdělávacích prostředích (VLE). Zmíněny byly také například inteligentní tutorské systémy (ITS) či adaptivní hypermediální systémy (AHS), které také produkují data, jež mohou být analyzována. Dále lze zmínit například vzdělávací systémy na podporu spolupráce a počítačem podporovaného kolaborativního učení (CSCL) či tzv. osobní vzdělávací prostředí (PLE) a e-portfolia, jejichž data mohou být rovněž vytěžována. Data dále mohou pocházet i z výukových her či tzv. vážných her (*serious games*), případně se může jednat o data z různých online testovacích systémů, hlasovacích systémů (*student response systems*) a jiných elektronických výukových platforem. Na druhé straně ale může jít i o různé typy dat pocházejících z tradičního (myšleno „offline") vzdělávacího prostředí. V těchto případech mohou být data dostupná v různých vzdělávacích informačních systémech (EIS) či studentských informačních systémech (SIS). Také může být nutné tato data nejprve převést do elektronické podoby (jsou-li např. zaznamenávána pouze v papírové podobě) či nejprve navrhnout, jak by vůbec mohla být sbírána (v případě, že jde o data, která se doposud nesbírají).

Jiným rozlišovacím kritériem může být způsob sběru dat (Bousbia & Belamri, 2014). Velmi často bývají data sbírána automatizovaně v digitální podobě, jak je tomu u zmiňovaných online vzdělávacích systémů. Zde samozřejmě záleží na

daném systému, která konkrétní data automaticky zaznamenává. Obvykle však jde o nějakou podobu tzv. logů, které zaznamenávají určité typy událostí, jež v systému nastávají. K dispozici jsou i data v digitální podobě, která nejsou generována automaticky. Zde může jít např. o již zmiňované seminární práce odevzdávané studentem v digitální podobě, příspěvky studentů v online diskuzních fórech, ale také např. videozáznamy výuky apod. Data však mohou být sbírána i manuálně, případně mohou být kombinovány různé typy sběru dat. Typickým příkladem je situace, kdy výzkumník provádí pozorování dění ve školní třídě, zatímco studenti pracují s určitým vzdělávacím softwarem, přičemž tato manuálně zaznamenávaná data z pozorování jsou následně spojena s automaticky sbíranými daty ze vzdělávacího softwaru.

Sbíraná data obvykle obsahují osobní či citlivé údaje, proto se lze na sběr dat dívat i z hlediska toho, nakolik dochází k získávání a zpracovávání dat osobní či citlivé povahy. Sclater (2017) v tomto kontextu zmiňuje tzv. poskytované údaje (*provided data*), tzn. data, která individuální subjekt (obvykle student) vědomě poskytuje. Příkladem takto vědomě poskytnutých dat mohou být údaje získávané s využitím online dotazníků či formulářů, jež student dobrovolně vyplní a odešle. Jiným příkladem jsou data pramenící z pozorování v širokém slova smyslu (*observed data*). Zde má Sclater na mysli jakákoli automaticky zaznamenávaná data, která se nějakým způsobem týkají činnosti sledovaných subjektů. Spadají sem tedy i již zmiňované logy a další typy automaticky zaznamenávaných dat v rámci online výukových systémů. Rovněž zde můžeme hovořit např. o různých videozáznamech výuky či o záznamech z nejrůznějších senzorů použitých v rámci výukových experimentů apod. Za další typ dat pak lze považovat derivovaná data (*derived data*), která nejsou sbírána přímo, nýbrž jsou generována na základě jiných sbíraných dat. Zde jsou typickým příkladem různé metriky vypočítávané ze sbíraných dat (např. počet zobrazených studijních materiálů za jednu návštěvu online kurzu, průměrný počet slov na jeden online diskuzní příspěvek apod.). Posledním typem dat jsou tzv. vyvozovaná data (*inferred data*), která jsou do jisté míry podobná derivovaným datům. Opět jde o data, jež nejsou sbírána přímo, nýbrž jsou nějakým způsobem odvozena od sbíraných dat. Oproti derivovaným datům však nejde o jednoduché metriky, ale o data pramenící z aplikace analytických a data miningových metod na sbíraná data. Příkladem tak mohou být data, jež jsou výstupem nějakého prediktivního systému a která uvádějí pravděpodobnost, s jakou daný student úspěšně ukončí určitý kurz. Jiným příkladem vyvozených dat může být aplikace shlukovacího algoritmu a zařazení studenta do určité skupiny či kategorie na základě jeho studijních strategií při plnění online kurzu.

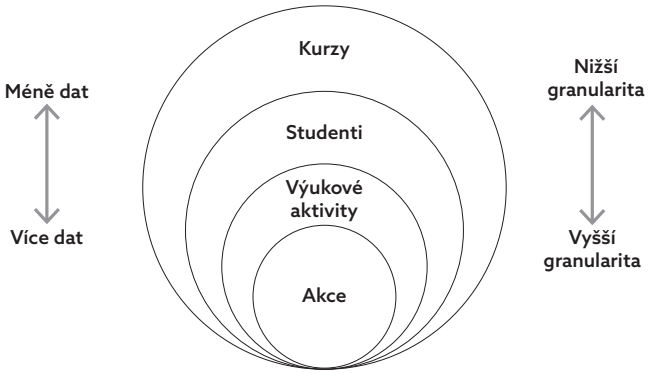
Dále lze hovořit i o dostupnosti dat (Bousbia & Belamri, 2014; Sclater, 2017). Častým případem v kontextu data miningu ve vzdělávání a analytiky učení je situace, kdy jsou v rámci instituce využita již existující a dostupná data, která však zatím pro tyto účely nebyla využívána nebo nebyla doposud dostatečně vytěžena.

Typicky tak jde o data z různých vzdělávacích a výukových online systémů, která byla v průběhu let automaticky zaznamenávána v databázích daných systémů. Jiným příkladem využití již dostupných dat mohou být data, jež jsou volně k dispozici. Zde může jít např. o data dostupná v některém z široké škály obecných datových repositářů či o konkrétní datové sady zveřejněné obvykle pro nějaké specifické účely či v souvislosti s konkrétním realizovaným výzkumem[23]. Nutno však zmínit, že existují i pokusy o aplikaci data miningových metod na volně dostupná data z velkých mezinárodních výzkumů (*international large scale assessments – ILSA*) jako jsou PISA, TIMSS, PIRLS či ICILS, která jsou s využitím „klasických" statistických metod v rámci pedagogického výzkumu využívána poměrně běžně (viz např. Juhaňák et al., 2018, 2019). Čím dál tím častěji se v kontextu data miningu ve vzdělávání a analytiky učení objevují také typy dat, která jsou generována v průběhu výzkumných experimentů za využití specifických měřicích přístrojů[24]. Taková data jsou samozřejmě primárně nedostupná a mohu být k dispozici teprve po realizaci daného experimentu.

Důležitým kritériem pro rozlišování různých typů dat ve vzdělávání je rovněž úroveň či podrobnost popisu, tzv. granularita dat (Romero & Ventura, 2013). Jak naznačuje obrázek níže (viz obrázek 6), data využívaná v rámci data miningu ve vzdělávání se mohou týkat různé úrovně obecnosti, resp. mohou být různě podrobná. Můžeme mít k dispozici data týkající se jednotlivých kurzů (např. počet studentů v kurzu, zaměření kurzu, požadované ukončení kurzu aj.), ale také data na úrovni studentů v rámci kurzu (např. získané známky v kurzu). Zároveň lze věnovat pozornost ještě i nižším úrovním a sledovat např. jen konkrétní výukovou aktivitu v rámci kurzu či konkrétní akce studentů v rámci dané aktivity. Anebo se lze naopak zaměřit na obecnější úroveň a pracovat s daty na úrovni studijních programů či celých vzdělávacích institucí. Data v různých online vzdělávacích systémech (ale i v rámci jednoho systému) se přitom vyskytují v rozdílných úrovních granularity, proto je obvykle nejprve nutné převést data na požadovanou úroveň granularity na základě položené otázky či záměrů výzkumu. Případně je možné provádět analýzy zohledňující více úrovní a hierarchický charakter dat (např. úro-

23 Příkladem datového repozitáře může být např. repozitář *DataShop @CMU* (pslcdatashop.web.cmu.edu). Jako příklady specificky zaměřených dat lze uvést např. *LAK Dataset* (solaresearch.org/initiatives/dataset), který obsahuje metadata a strukturované plné texty klíčových výzkumných publikací v oboru analytiky učení a data miningu ve vzdělávání. Dále lze zmínit data publikovaná českým týmem, jenž se věnuje data miningu ve vzdělávání na Open university – viz *Open University Learning Analytics dataset* (www.nature.com/articles/sdata2017171). Tato data pochází z online vzdělávacího systému využívaného na Open university. V neposlední řadě lze zmínit i data zveřejněná na stránkách výzkumné skupiny *Adaptive Learning*, která působí na Fakultě informatiky Masarykovy univerzity a která se věnuje především problematice adaptivních vzdělávacích systémů, z nichž pak pochází publikovaná data (viz www.fi.muni.cz/adaptivelearning/?a=data).

24 Příkladem mohou být relativně často využívaná zařízení pro snímání očních pohybů (tzv. eye tracking). Pro širší přehled specifických měřicích přístrojů, které jsou či mohou být v tomto kontextu využívány, se lze obrátit na webové stránky laboratoře HUME Lab (humelab.cz/services/equipment).



Obrázek 6: Různé úrovně granularity a množství dat (podle Romero & Ventura, 2013)

veň studentů a úroveň kurzů, ve kterých jsou zapsáni). Zmiňovaný obrázek zároveň naznačuje vztah mezi granularitou a množstvím dat. Pohybujeme-li se na vysoké úrovní detailu (tj. vysoká granularita), pracujeme obvykle s velkými objemy dat, kdežto na obecnějších úrovních (nižší granularita) je množství dat nižší.[25]

V neposlední řadě lze rozlišovat různé typy dat na základě struktury dat (Bousbia & Belamri, 2014), resp. dle toho, v jaké podobě jsou údaje ukládány a uchovávány či do jaké podoby musí být údaje před aplikací data miningových metod transformovány. Za do značné míry výchozí strukturu je možné považovat tabulku, kdy jednotlivé řádky tvoří vybrané případy (studenti v kurzu, různé kurzy apod.) a sloupce tvoří sledované proměnné (známky získané v kurzu, množství studentů v kurzu apod.). S takovou strukturou dat se lze setkat i při využití tradičních metod pedagogického výzkumu (např. při dotazníkovém šetření). Obvykle se přitom v této souvislosti hovoří o tzv. atribučních datech. V rámci analytiky učení a data miningu ve vzdělávání se však výzkumníci relativně často setkávají i s jinak strukturovanými daty. Např. při práci s daty kvalitativního (tj. textového) charakteru může být potřeba nejprve vytvořit odpovídající textový korpus (např. korpus odevzdaných úkolů či odpovědí v diskuzním fóru), než je možné data analyzovat. Data mohou zachycovat určité vztahy či sociální interakci (např. data z online diskuzních fór), a tudíž mohou být strukturována v podobě tzv. matice sousednosti či v podobě seznamu uzlů a seznamu hran. V tomto kontextu se hovoří o tzv. relačních datech. Jiným příkladem mohou být již zmiňované logy, u nichž lze

25 Pro lepší představu: Budeme-li se např. pohybovat na úrovni jednoho kurzu a studentů v něm, pak se u běžného univerzitního kurzu můžeme pohybovat v řádu desítek až stovek studentů. Naopak při zaměření pozornosti na úroveň konkrétních aktivit studentů v online kurzu se u běžného univerzitního kurzu dostáváme k desetitisícům až statisícům záznamů (v závislosti na počtu studentů a na množství studijních materiálů a výukových aktivit v online kurzu).

hovořit o datech zachycujících nějaký proces, resp. řadu událostí v čase. Tato data mají vlastní specifickou strukturu, kdy jednotlivé řádky v tabulce tvoří události, přičemž každá událost musí obsahovat minimálně informaci o případu, kterého se týká, o aktivitě, která je provedena, a o času, kdy událost nastala. Mimo uvedené příklady lze pak samozřejmě v realizovaných výzkumech narazit i na další typy dat se specifickou strukturou.