Mendelova univerzita v Brně
Provozně ekonomická fakulta

---

# Detekce duplicit v geoprostorových datech

**Diplomová práce**

Vedoucí práce:
Ing. Pavel Turčínek, Ph.D.

Bc. Adam Prchal

Brno 2024

**Poděkování**

Velké poděkování patří vedoucímu diplomové práce Ing. Pavlovi Turčínkovi, Ph.D. za užitečné rady, vedení a ochotu konzultovat v jakoukoliv hodinu. V neposlední řadě patří poděkování také všem, kteří se jakkoliv podíleli na zlepšení kvality této práce.

**Čestné prohlášení**

Prohlašuji, že jsem práci **Detekce duplicit v geoprostorových datech** vypracoval samostatně a veškeré použité prameny a informace uvádím v seznamu použité literatury. Souhlasím, aby moje práce byla zveřejněna v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách ve znění pozdějších předpisů a v souladu s platnou Směrnicí o zveřejňování závěrečných prací.

Jsem si vědom, že se na moji práci vztahuje zákon č. 121/2000 Sb., autorský zákon, a že Mendelova univerzita v Brně má právo na uzavření licenční smlouvy a užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

Dále se zavazuji, že před sepsáním licenční smlouvy o využití díla jinou osobou (subjektem) si vyžádám písemné stanovisko univerzity, že předmětná licenční smlouva není v rozporu s oprávněnými zájmy univerzity, a zavazuji se uhradit případný příspěvek na úhradu nákladů spojených se vznikem díla, a to až do jejich skutečné výše.

Brno 2024 .....

podpis

## Abstract

.

.



## Abstrakt

.

.

# Obsah

# Todo list



# Seznam obrázků

# 1 Úvod a cíl

## 1.1 Úvod

Společnosti poskytující služby v oblasti Location Intelligence využívají své platformy pro komplexní geoprostorové analýzy, což je zásadní pro efektivní rozhodování v různých typech podnikání. Jako součást těchto platforem, některé společnosti, včetně CleverMaps, nabízí služby zvané jako Data Marketplace, které umožňují uživatelům získávat a integrovat různorodé datové sady. Tyto sady zahrnují např. demografické informace, obchodní statistiky a infrastrukturu měst, což pomáhá klientům lépe cílit své služby a strategická rozhodnutí. (CleverMaps, 2024)

Vzhledem k tomu, že při sběru dat a následnou manipulací s geoprostorovými daty dochází často ke kombinování dat z různých zdrojů (pro co největší úplnost konečných datových sad), vzniká problém s výskytem duplicitních záznamů. Duplicity v datech mohou mít různé podoby a vznikají z několika důvodů, některé z nich jsou:

- **Rozdílné názvy stejných míst** – např. *"Velký městský park"* vs. *"Park v centru města"*.
- **Typografické chyby a nejednotné formáty** – např. ulice *"Masarykova"* vs. *"Masarykova tř."*.
- **Rozdílné souřadnicové systémy** – jeden data set může používat *WGS84*, zatímco jiný *S-JTSK*.

Tyto duplicity snižují kvalitu dat a mohou způsobit chyby v rozhodovacích procesech, např. při plánování dopravy, marketingových analýzách nebo při geokódování obchodních poboček. Proto je detekce a eliminace duplicit v geoprostorových datech klíčová. K identifikaci duplicit mohou být využity různé metody, od jednoduchých textových porovnání až po složitější techniky strojového učení, které analyzují podobnost dat v širším kontextu. (Nauman, 2022; Christen, 2012)

Existují komerční i open-source nástroje pro detekci duplicit, které často využívají cloudové technologie a pokročilé algoritmy. Mezi ně patří např. Data Ladder, Tilores nebo Melissa. Ačkoliv tyto nástroje zlepšují kvalitu dat, často představují vysoké náklady nebo nejsou dostatečně přizpůsobitelné konkrétním datovým sadám. To motivuje společnosti k hledání vlastních řešení. (Christen, 2012)

## 1.2 Cíl

Cílem této práce je prozkoumat a otestovat různé metody detekce duplicit na geoprostorových datech, včetně metod založených na strojovém učení. Na základě analýzy výsledků testů doporučit nejvhodnější metody pro konkrétní typy sad geoprostorových dat, přičemž ověření těchto metod proběhne na datových sadách poskytnutých společností CleverMaps.

Výsledná doporučení by měla společnosti CleverMaps pomoci v rámci zvyšování automatizace a zkvalitnění procesů kontrol kvality dat.

# 2 Literární rešerše

## 2.1 Problematika detekce duplicit v datech

Data jsou v rámci různých ekosystémů ukládána v různých formách. Liší se ve struktuře, pojmenování atributů a hlavně ve způsobu využívání unikátních identifikátorů datových entit. Při agregaci dat z různých zdrojů do centrální databáze vzniká problém. Tím je rozhodnout, které záznamy jsou unikátní a které se ve skutečnosti opakují a mají pouze jinou podobu kvůli odlišné struktuře dat nebo způsobu jejich získání.

Nekvalitní data jsou komplikací, která může vést k zásadním chybám v datové analýze a rozhodovacích procesech. Například ve zdravotnictví může mít existence duplicitních záznamů o pacientech vážné důsledky od nesprávně předepsaných léků až po chybné zdravotní statistiky.(Bess, 2024) V oblasti e-commerce může duplicita zákaznických účtů znamenat špatně cílené marketingové kampaně nebo mylné vyhodnocení chování uživatelů.(Brown, 2019) A v geoprostorových datech mohou duplicity vést k nesprávné identifikaci lokací, chybným navigačním trasám nebo nesrovnalostem v mapových podkladech.

Pro přesné popsání procesu detekce duplicit je nejprve nutné vyjasnit základní pojmy související s datovými entitami a jejich reprezentací v databázích.

**Entita** je obecný pojem označující reálný objekt nebo koncept, který má své vlastnosti a může být reprezentován v datech. V závislosti na kontextu může entitou být např. osoba, firma, geografický bod zájmu (POI[1]) nebo administrativní oblast.

**Záznam** představuje konkrétní reprezentaci entity v databázi. Jedna entita tak může být v různých databázích reprezentován více různými záznamy, například s mírně odlišnými údaji nebo formátem.

**Atribut** je konkrétní vlastnost nebo charakteristika entity, která ji popisuje a umožňuje její identifikaci v databázích. Každý záznam obsahuje jeden nebo více atributů, přičemž některé atributy mohou být klíčové pro rozpoznání duplicit. Atributy mohou být různých typů, například textové, číselné, kategorické nebo prostorové. V geolokačních datech bývají klíčovými atributy například souřadnice (zeměpisná šířka a délka), adresa, název místa nebo jeho typ (např. restaurace, park, obchodní centrum).

Duplicitní záznamy vznikají tehdy, když databáze obsahuje více reprezentací stejné entity, ať už kvůli chybám v zápisu, rozdílnému formátu dat, nebo rozdílným zdrojům dat. V literatuře se však termíny označující proces identifikace těchto duplicit ne vždy používají jednotně a často se jejich významy překrývají. Zatímco některé zdroje považují následující pojmy za synonyma, jiné mezi nimi rozlišují:

- **Data matching** – proces porovnávání záznamů za účelem nalezení těch, které odpovídají stejné entitě, i když mají různé atributy nebo formáty. (Christen,

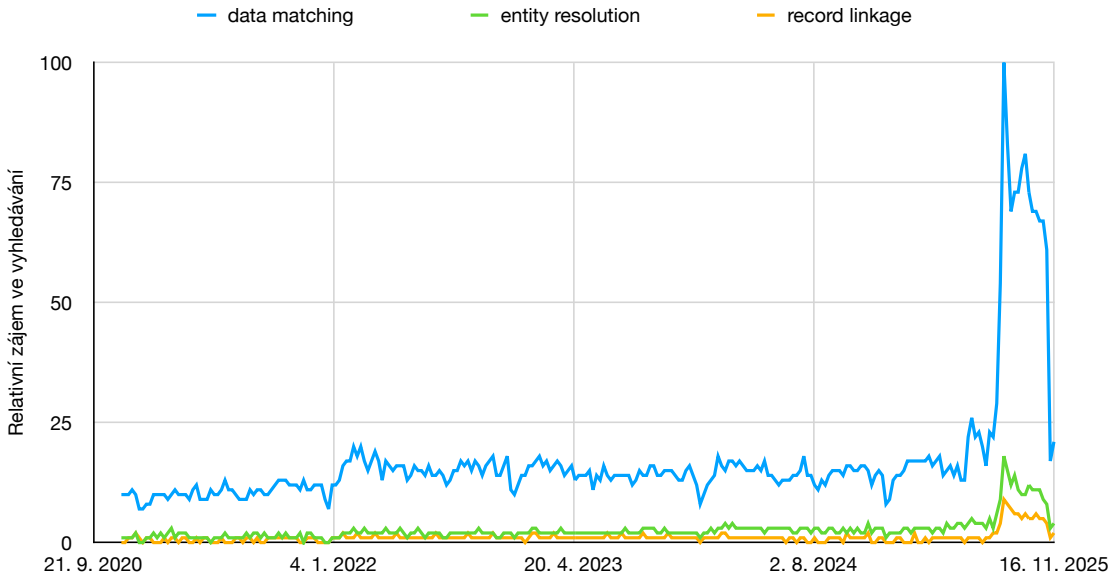
[1]POI – *Point of Interest*, tedy konkrétní místo nebo lokalita, která je nějakým způsobem významná či relevantní pro analýzu.

2012)

- **Entity resolution** – proces slučování duplicitních záznamů do jednoho sjednoceného záznamu reprezentujícího danou entitu. (Quantexa, 2024)
- **Record linkage** – propojení odpovídajících záznamů napříč různými databázemi, aniž by došlo ke sloučení do jedné reprezentace. (Stepanenko, 2024)

To, že existují různé termíny pro podobné procesy, ukazuje, že problematika detekce duplicit vznikala nezávisle v různých oborech, a proto je důležitější se soustředit na samotné postupy, než na přesné označení. (Christen, 2012)

V této práci bude hlavní pozornost věnována *data matching*, tedy samotnému procesu hledání duplicitních záznamů v datech, spíše než jejich následnému slučování (*entity resolution*). Pro zajímavost, podle Google Trends je termín *data matching* v posledních pěti letech globálně vyhledávanější než ostatní zmiňované termíny. Viz obr. 1.




Obrázek 1: Zájem o termíny ”data matching”, ”entity resolution” a ”record linkage”. *Zdroj*: (Google Trends, 2025)

## 2.2 Příčiny vzniku duplicit

Některé důvody vzniku duplicit v datech již byly zmíněny a následuje podrobnější pohled na vybrané z nich. Duplicity lze rozdělit na ty vznikající v rámci jednoho datového zdroje a na ty vznikající spojením více datových zdrojů.

### Duplicity v samostatném zdroji

Přestože by se na první pohled zdálo, že v rámci jednoho datového zdroje šance na vznik duplicity minimální, je zde i tak několik případů, které vedou ke vzniku duplicit. Tady jsou některé z nich:

- Lidský faktor – při ručním zadávání dat je vysoká pravděpodobnost typografických chyb, nekonzistentních formátů nebo přímo neplatných informací. V prostředích, kde data zadává více osob (například projekty typu OpenStreetMap), se riziko ještě zvyšuje, protože jednotliví přispěvatelé nemusí zaznamenat již existující záznam vložený dříve.
- Nedostatečné standardizace formátů – různé formuláře, dotazníky, nebo části/-moduly jednoho systému mohou využívat různé formáty pro datum, telefonní čísla, rodná čísla a další typy dat. Pokud není správně ošetřen převod těchto dat do jednotného formátu před vložením záznamu, může opět docházet ke vzniku duplicit.
- Chybějící integritní omezení databáze – v případě, že daná databáze nemá nastavená integritní omezení (např. na unikátní hodnoty v rámci rodného čísla, IČO nebo třeba unikátnosti dvou a více atributů v rámci jednoho záznamu), bude každý vložený záznam brán jako nový, doposud neexistující, záznam, přestože se může jednat o identickou duplicitu.

### Duplicity ve sloučeném zdroji

Při pokusu o spojení více zdrojů do jednoho vzniká několik výzev. V rámci business inteligence se tento proces spojování různých datových zdrojů nazývá ETL[2]a v rámci něj může docházet k duplicitám v několika případech:

- Integrace různých systémů – v případě systémů, které mají data o entitách kvůli různým důvodům (např. inventární systém školy ukládá záznamy o vybavení, online bazar má záznamy o prodávaném vybavení), může vzniknout situace, kdy záznamy z více systémů nelze spojit na základě společného unikátního klíče, protože má každý systém svůj unikátní pro danou entitu.
- Průběžné spojování dat – některé data sety jsou pravidelně aktualizované, aby co nejpřesněji odrážely realitu. Pokud však entita z reálného světa v rámci aktuality změnila své atributy zásadně (např. pekárna se přestěhovala na jinou

[2]Extract, transform, load

adresu a upravila své jméno), může vzniknout situace, kdy některý ze zdrojů bude mít stále staré údaje, a jiný nové, a tím pádem nepůjde tento rozdíl spojit s původním záznamem a bude se mylně jevit, že jde o novou entitu.

- Rozdílné definice entit a identifikační metody – při integraci dat z různých zdrojů se může stát, že jednotlivé systémy definují a identifikují stejné entity odlišně. Jeden systém může například využívat kombinaci jména, adresy a rodného čísla k jednoznačné identifikaci osoby, zatímco jiný systém pracuje s interním identifikačním číslem nebo používá jiné atributy (např. e-mailovou adresu). Tato nejednotnost vede k tomu, že algoritmy pro slučování dat nemusí správně rozpoznat, že se jedná o stejnou entitu, a následně dochází k vytvoření duplicitních záznamů v cílovém datovém skladu.

Toto zajisté nejsou všechny možné příčiny vzniku duplicit v datech, měly by však nastínit, jak se jednotlivé příčiny mohou lišit, a proč se jedná o problém, který nelze vyřešit univerzálním způsobem.

## 2.3 Důsledky duplicitních dat

Duplicitní data mohou mít na organizaci zásadní a mnohostranné dopady. Především se výrazně snižuje kvalita dat, informace se stávají nepřesnými, neúplnými a méně aktuálními. To ovlivňuje nejen tvorbu reportů, ale i analytické procesy, které jsou základem strategických rozhodnutí. Rozhodnutí vycházející z chybně interpretovaných dat mohou vést k nesprávným závěrům, což se může projevit například ve špatném plánování, neefektivním řízení zdrojů nebo dokonce v riziku finančních ztrát.

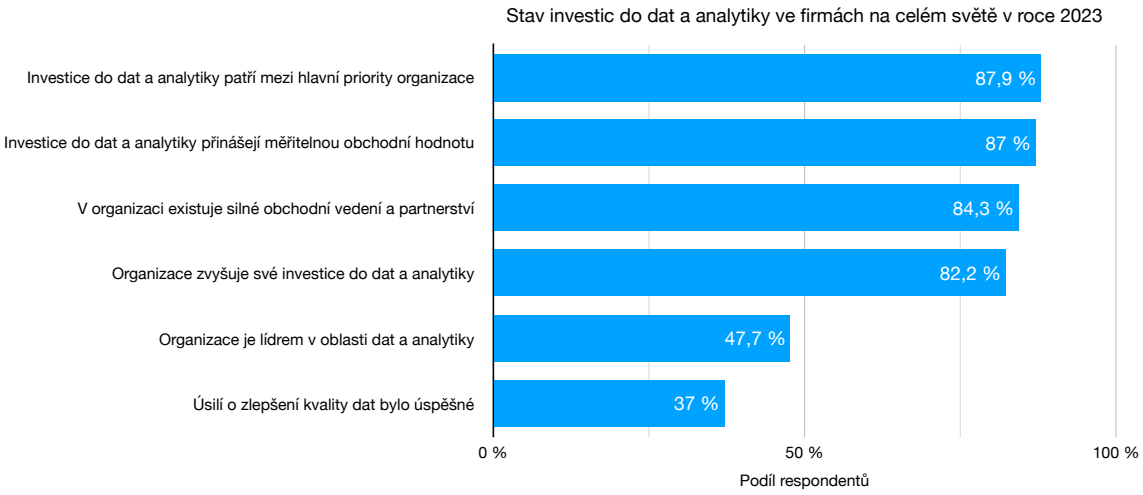
Dalším významným důsledkem je snížení efektivity pracovníků. V situaci, kdy je třeba trávit spoustu času identifikací a odstraňováním duplicit, se zaměstnanci odklánějí od jiných úkolů, které by mohly přinést přidanou hodnotu. Kromě rutinní práce, jako je kontrola a aktualizace záznamů, je také nutné dohledávat různé verze stejných dat, což výrazně komplikuje správu informací. Takový stav nejenže zvyšuje provozní náklady, ale také vede k vyšší chybovosti a celkovému zhoršení kvality práce.

Nelze opomenout ani problémy spojené s dodržováním regulatorních požadavků. Regulace často vyžadují přesné, konzistentní a bezpečné uchovávání dat. Duplicitní záznamy komplikují tvorbu přesných reportů, což může mít za následek sankce, pokuty nebo dokonce ztrátu důvěry ze strany regulačních orgánů, spolupracujících firem a zákazníků. Správa citlivých údajů, jako jsou osobní nebo finanční informace, se také stává náročnější, což představuje další riziko v případě případných narušení bezpečnosti nebo při nutnosti vyhovět zákonným požadavkům na přístup a správu těchto dat.

Duplicitní data také negativně ovlivňují i zákaznický servis a správu IT infrastruktury. Když jsou informace o zákaznících roztříštěny mezi několika duplicitními záznamy, je pro pracovníky obtížné získat ucelený přehled o jejich historii, což

vede ke zhoršení kvality poskytovaných služeb. Současně se tím zvyšuje náročnost správy datových logů, zálohovacích a archivních procesů.

Celkově lze říct, že neřešené duplicitní záznamy mají rozsáhlý negativní vliv na efektivitu, bezpečnost a spolehlivost informací v organizaci. Firmy se snaží aktivně problém s kvalitou dat řešit, a to ukazuje i průzkum "Big Data and AI Executive Survey 2024" od NewVantage Partners (2024) za rok 2023. Průzkumu se účasnilo více než 100 firem z globálního Fortune 1000, a ukazuje, že pro většinu organizací (87,9 %) má investice do dat a analytiky vysokou prioritu. Ze stejného výzkumu však i vychází informace, že snaha o zlepšení kvality dat byla úspěšná pouze u 37 % dotázaných firem.




Obrázek 2: Výsledky výzkumu o investicích do dat a analytiky.
*Zdroj*: (NewVantage Partners, 2024)

## 2.4 Obecné metody detekce duplicitních záznamů

K řešení detekce duplicit je třeba využít kombinaci několika metod, které se vzájemně doplňují a umožňují zvýšit celkovou přesnost a robustnost systému. Následující podkapitola rozebírá jednotlivé skupiny přístupů, jejich fungování, výhody i nevýhody, a možnosti jejich vzájemného propojení.

### Pravidla

Jedná se o principiálně jednoduchý přístup, který spočívá v tom, že se vytváří jednoznačná pravidla nebo sada pravidel, které podle daných kritérií určují duplicitní skupiny záznamů. Taková pravidla bývají navržena experty v dané oblasti, kteří vychází ze znalostí z oboru a charakteristik konkrétních dat.

Pravidla mají zpravidla jasně definované atributy, nebo skupinu atributů, které porovnávají a to buď striktně (hodnoty atributů musí být identické), kdy se jedná

o tzv. exact matching, nebo volněji s použitím např. matematické logiky či prahových hodnot. Lze však kombinovat striktní a volnější porovnání v rámci složeného pravidla. To může slovně být popsáno třeba takto: *„Dva záznamy jsou duplicity, pokud se shodují v atributu jméno a příjmení, mají shodné datum narození a současně rozdíl v jejich příjmech za poslední rok nepřesahuje 5 %.“*. (Tejada, 2002)

K volnějšímu porovnání lze uvést i fonetické metody, které se používají primárně u textových údajů, které zní velice podobně, ale zapisují se různě. Jedná se například o Soundex, Metaphone a Double Metaphone, které z textového vstupu generují fonetické kódy, které lze následně porovnávat. (Howard, 2020)

Výhodou takových pravidel je snadná implementace a jasná interpretace výsledků. Nicméně, vytvořená pravidla jsou málo flexibilní, jelikož jsou psaná vždy pro specificky typ dat, nebo pro konkrétní odvětví. V jednoduchých případech může takové řešení stačit, nicméně ve složitějších situacích může komplexnost takových pravidel růst a mohou být složitá na případná ladění a upravování. Pravidla však mohou sloužit jako dobrý první krok při zpracování dat, a na ně lze dále navázat další pokročilejší metody, které budou popsány v dalších sekcích.

### Metody založené na podobnosti

Metody založené na podobnosti identifikují duplicitní záznamy pomocí měření vzájemné podobnosti jednotlivých záznamů. Podobnost lze definovat jako míru blízkosti nebo podobnosti dvou záznamů na základě jejich atributů. Tyto metody jsou založeny na matematických metrikách, které kvantifikují míru podobnosti mezi dvojicemi atributů či celých záznamů.

Podobnost mezi dvěma záznamy si lze představit jako číselnou hodnotu vyjadřující jejich vzájemnou shodu. Hodnota podobnosti bývá často definována v intervalu od 0 (žádná podobnost) do 1 (přesná shoda). K vyhodnocení podobnosti se používají různé metriky, které se zvolý na základě typu a charakteru porovnávaných atributů (text, čísla, časový údaj, atd.).

Mezi jednu z nejznámějších metrik patří **Levenshteinova vzdálenost**, která měří, kolik znaků je nutno přidat, odebrat, nebo změnit, aby se jeden řetězec stal identický s druhým. Výsledkem je číselná hodnota, která udává, kolik těchto operací se musí minimálně provést. Jde o metriku, která umožňuje snadno rozeznat například překlepy v krátkých řetězcích, ale v případě dlouhých řetězců mohou být vypočítané vzdálenosti vysoké. (Christen, 2012)

Další známou metrikou je **Jaro-Winklerova podobnost**, která měří podobnost na základě počtu shodných znaků, jejich relativní vzdálenosti a počtu transpozic (špatného pořadí znaků). Krom toho také zohledňuje to, jestli porovnávané řetězce mají stejný začátek (prefix) a pokud ano, tak výrazně zvýší výslednou hodnotu podobnosti, která je v rozmezí 0 (zcela odlišné) a 1 (zcela shodné). Tato metrika je tedy lépe aplikovatelná na delší řetězce než zmíněná Levenshteinova vzdálenost, ale kvůli výpočetní náročnosti se často uvádí, že není ideální na porovnání dlouhých řetězců. (Christen, 2012)

**Kosinová podobnost s TF-IDF** reprezentací je často používanou kombinací metod vhodnou především pro porovnávání delších textových dokumentů. TF-IDF[3] transformuje texty do číselných vektorů, přičemž každému slovu v dokumentu přiřazuje váhu podle toho, jak často se vyskytuje v rámci konkrétního dokumentu (term frequency - TF) a jak vzácně nebo běžně se vyskytuje napříč celou kolekcí dokumentů (inverse document frequency - IDF). Slova, která se objevují často v mnoha dokumentech (např. běžné spojky či předložky), dostávají nižší váhu, zatímco specifická a méně častá slova obdrží váhu vyšší, čímž je zajištěno lepší zachycení významové relevance jednotlivých dokumentů. Kosinová podobnost následně měří podobnost dvou takto vzniklých vektorů dokumentů jako kosinus úhlu mezi nimi, přičemž výsledek se pohybuje mezi hodnotami 0 a 1, kde hodnota blízká 1 označuje velmi podobné dokumenty a hodnota blízká 0 označuje dokumenty s minimální nebo žádnou podobností. Velkou výhodou této metody je schopnost efektivně porovnávat obsah dokumentů bez zkreslení způsobeného různými délkami textů, a proto je tato technika široce využívána v oblasti vyhledávání informací, doporučovacích systémech, klasifikaci dokumentů a identifikaci duplicit, kde je potřeba přesněji zachytit tematickou nebo obsahovou podobnost textových záznamů. (Jurafsky, 2025)

**Jaccardova podobnost** představuje další často používanou metriku, která se zaměřuje na měření podobnosti množin atributů nebo znaků dvou záznamů. Tato metoda je založena na porovnání velikosti průniku (tedy počtu společných prvků) a velikosti sjednocení (tedy celkového počtu unikátních prvků v obou množinách). Výsledná hodnota Jaccardovy podobnosti se pohybuje mezi 0 (žádná shoda mezi množinami) a 1 (úplná shoda množin). Výpočet Jaccardovy podobnosti je jednoduchý, rychlý a dobře interpretovatelný, což ji dělá vhodnou pro různé aplikace, zejména tam, kde je potřeba porovnávat množiny diskrétních atributů, například množiny klíčových slov, tokenů, kategorií nebo uživatelů. Jednou z nevýhod této metriky však je to, že nezohledňuje frekvenci výskytu jednotlivých prvků, ale pouze jejich přítomnost nebo absenci. Jaccardova podobnost proto není ideální tam, kde je důležitá četnost prvků, například u analýzy textů, kde může být užitečnější kombinace s TF-IDF nebo jinými frekvenčními přístupy. Přesto se často využívá v systémech pro detekci duplicit, shlukování dokumentů a doporučovacích systémech. (Jurafsky, 2025)

Metody založené na podobnosti tedy představují efektivní přístupy k identifikaci duplicitních záznamů prostřednictvím různých matematických metrik, jejichž výběr závisí na typu a charakteru analyzovaných dat. Zatímco Levenshteinova vzdálenost a Jaro-Winklerova podobnost jsou ideální pro porovnávání krátkých nebo středně dlouhých textů, například při detekci překlepů či podobnosti názvů, Kosinová podobnost s TF-IDF reprezentací se osvědčuje především při práci s delšími dokumenty, kde je důležité zachytit obsahovou relevanci. Jaccardova podobnost zase vyniká svou jednoduchostí a rychlostí při porovnávání množin diskrétních atributů, jako jsou např. klíčová slova nebo tagy. V praxi se často osvědčuje vhodně kombi-

[3] term frequency-inverse document frequency

novat různé metriky nebo jejich varianty podle konkrétní situace, aby bylo dosaženo co nejpřesnějšího a nejspolehlivějšího výsledku při identifikaci duplicitních entit.

### Pravděpodobnostní metody

Pravděpodobnostní metody se zaměřují na modelování pravděpodobnosti, že dva záznamy odpovídají stejné entitě. Tyto metody využívají statistické postupy, které umožňují hodnotit míru shody mezi záznamy nikoliv pouze na základě přímého porovnání atributů, ale také na základě pravděpodobnostních vztahů, jež odrážejí chování dat v reálném prostředí. Jedním z nejznámějších a nejpoužívanějších přístupů v této oblasti je **Fellegi-Sunterův model**, který tvoří základ moderních metod identifikace duplicitních záznamů.

Fellegi-Sunterův model vychází z principu, že každá dvojice záznamů může spadat do jedné ze dvou hypotéz: buď představuje záznamy popisující stejnou entitu, nebo jde o záznamy různých entit. Každý atribut se při porovnání hodnotí podle toho, jak pravděpodobné je, že by se jeho shoda či neshoda vyskytla v případě skutečné duplicity, případně v případě dvou nesouvisejících záznamů. Na základě těchto informací se z jednotlivých atributů stanoví celkové **pravděpodobnostní skóre**, které vyjadřuje, do jaké míry pozorované shody podporují jednu či druhou hypotézu. (Winkler, 2014)

Model následně rozlišuje tři **rozhodovací oblasti**. Pokud je skóre dostatečně vysoké, dvojice záznamů je označena jako duplicita. Pokud je skóre naopak nízké, dvojice je považována za odlišné záznamy. Zbývající dvojice spadají do takzvané ***nejisté oblasti***, která vyžaduje ruční přezkoumání nebo dodatečné metody k upřesnění výsledku. Tento postup umožňuje minimalizovat oba typy chyb – **falešně pozitivní** i **falešně negativní** – a zároveň poskytuje strukturovaný způsob, jak pracovat s nejednoznačnými případy. (Winkler, 2014)

V praxi existuje množství **variant Fellegi-Sunterova modelu**, které se snaží řešit limity původní metody a lépe reflektovat povahu moderních dat. Patří sem například přístupy využívající **EM algoritmus** pro automatické odhadování parametrů modelu, rozšíření umožňující pracovat s **chybějícími hodnotami**, metody zohledňující závislosti mezi atributy či modely určené pro specifické typy dat, jako jsou číselné, textové nebo časové údaje. Některé novější varianty dále využívají **bayesovské přístupy**, které umožňují zahrnout předběžné znalosti o datech a vytvářet flexibilnější pravděpodobnostní modely. (Christen, 2012)

Výhodou pravděpodobnostních metod je jejich schopnost pracovat s neurčitostí, zohledňovat různou **informativnost jednotlivých atributů** a dosahovat vysoké přesnosti při správně nastavených parametrech. Díky tomu jsou vhodné pro rozsáhlé databáze a různorodé datové zdroje, kde se mohou vyskytovat překlepy, různé formáty či jiné formy nekonzistence. Nevýhodou těchto metod je vyšší výpočetní náročnost a potřeba kvalitního odhadu parametrů, který nemusí být triviální u dat s omezeným množstvím příkladů nebo u dat s výraznými závislostmi mezi atributy. V některých případech může také vzniknout potřeba ručního zásahu u záznamů

spadajících do nejisté oblasti, což snižuje celkovou míru automatizace. Přesto však pravděpodobnostní přístupy představují dobrý základ pro moderní systémy detekce duplicit, často používaný v kombinaci s dalšími metodami.

### Metody strojového učení

- Definice a typy (supervizované, semi-supervizované, nesupervizované)
- Typické algoritmy: rozhodovací stromy, Random Forest, Gradient Boosting
- Klíčové aspekty přípravy dat a výběr atributů
- Výhody, nevýhody

### Metody založené na hlubokém učení

- Princip neuronových sítí pro entity matching (např. embeddings, transformer modely)
- Scénáře nasazení a případové studie z literatury
- Omezení a nevýhody (black-box přístup, náročnost na data a výpočetní zdroje)

### Grafové metody

- Grafové přístupy k modelování entit
- Komunitní detekce, grafové neuronové sítě (GNN)
- Specifika aplikace, výhody, nevýhody

## 2.5 Specifické metody pro rozpoznání duplicit v geolokačních datech

- Specifika geolokačních dat
- Prostorová přesnost a chyby měření
- Problematika výběru vhodných vzdálenostních metrik a projekcí

### Metody založené na vzdálenosti

- Popis a využití Haversinovy a Vincentyho vzdálenosti
- Aplikace vzdálenostních metod, výběr prahů

### Shlukovací (clusteringové) metody

- DBSCAN, HDBSCAN a OPTICS
- Typické scénáře použití v praxi
- Výhody a omezení metod

### Metody založené na prostorové mřížce (grid-based)

- Principy Geohashe a prostorových mřížek
- Scénáře využití a důsledky výběru velikosti mřížky
- Omezení metody a problémy s přesností

### Pravděpodobnostní a bayesovské metody

- Modelování prostorové nepřesnosti (GPS chyby)
- Příklady bayesovských přístupů ke geolokaci
- Výhody a limitace těchto metod

### Metody strojového učení a hlubokého učení v geolokaci

- Stručně zmínit možnost aplikace ML přístupů, pokud to téma umožňuje
- Zvýraznit případové studie (pokud existují relevantní)

### Geokódování a porovnávání adres

- Standardizace adresních údajů
- Geokódovací nástroje a postupy

# 3 Literatura

Albayrak, Osman Semih; Aytekin, Tevfik; Kalayci, Tolga Ahmet *Duplicate product record detection engine for e-commerce platforms*. Expert Systems with Applications. 2022, roč. 193. ISSN 0957-4174. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0957417421017073..

Bess, Oleg *The problem with duplicate and mismatched patient records* [online]. 2024 [cit. 2025-02-09]. Dostupné z: https://www.physicianspractice.com/view/the-problem-with-duplicate-and-mismatched-patient-records..

Brown, Forrest *8 Problems That Result from Data Duplication • Profisee* [online]. 2019 [cit. 2025-02-09]. Dostupné z: https://profisee.com/blog/8-business-process-problems-that-result-from-data-duplication/..

Christen, Peter *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Berlin Heidelberg, 2012. 270 s. ISBN 978-3-642-43001-5 978-3-642-31163-5..

CleverMaps *Location Insights for Business* [online]. 2024 [cit. 2024-04-28]. Dostupné z: https://www.clevermaps.io..

Draisbach, Uwe; Naumann, Felix *On Choosing Thresholds for Duplicate Detection* [online]. 2013 [cit. 2024-04-28]. Dostupné z: https://hpi.de/oldsite/fileadmin/user_upload/fachgebiete/naumann/publications/PDFs/2013_draisbach_on.pdf..

Elahi, Ehsan *Why Data Duplicates Exist and How to Get Rid of Them? - Data Ladder Data Duplicates: Why Do They Exist and How to Eliminate Them?* [online]. 2021 [cit. 2025-02-09]. Dostupné z: https://dataladder.com/why-data-duplicates-exist-how-to-get-rid-of-them/..

Esri *GIS Dictionary* [online]. 2024 [cit. 2024-04-28]. Dostupné z: https://support.esri.com/en-us/gis-dictionary/search..

Evensen, Adrian *Entity Resolution — An Introduction* [online]. 2024 [cit. 2024-10-13]. Dostupné z: https://medium.com/@adev94/entity-resolution-an-introduction-fb2394d9a04e..

Google Trends *Google Trends* [online]. 2025 [cit. 2025-11-21]. Dostupné z: https://trends.google.com/trends/explore?date=today\%205-y\&q=data\%20matching,entity\%20resolution,record\%20linkage\&hl=cs..

HEAVY.AI *What is Geodata? Definition and FAQs* [online]. 2024 [cit. 2024-08-18]. Dostupné z: https://www.heavy.ai/technical-glossary/geodata..

Howard, James P., II *Phonetic Spelling Algorithm Implementations for R.* .

2020, roč. 95. ISSN . Dostupné z: https://doi.org/10.18637/jss.v095.i08..

Huyen, Chip *Designing machine learning systems: an iterative process for production-ready applications*. Beijing Boston Farnham Sebastopol Tokyo, 2022. 367 s. ISBN 978-1-0981-0796-3..

Jurafsky, Daniel; Martin, James H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. , 2025. s. ISBN ..

Keras *Keras documentation: Getting started with Keras* [online]. 2024 [cit. 2024-04-29]. Dostupné z: https://keras.io/getting_started/..

McClain, Bonny P. *Python for geospatial data analysis: theory, tools, and practice for location intelligence*. Beijing Boston Farnham Sebastopol Tokyo, 2023. 262 s. ISBN 978-1-0981-0479-5..

McGregor, Susan E. *Practical Python data wrangling and data quality*. Sebastopol, CA, 2022. 395 s. ISBN 978-1-4920-9150-9..

Nauman, Felix; Herschel, Melanie *An Introduction to Duplicate Detection*. , 2022. 84 s. ISBN 978-3-031-01835-0..

NewVantage Partners *Global state of data/analytics investment 2023* [online]. 2024 [cit. 2025-02-12]. Dostupné z: https://www.statista.com/statistics/1453262/global-state-of-data-analytics-investment/..

Pašek, Jan; Sido, Jakub; Konopík, Miloslav; Pražák, Ondřej *MQDD – Pre-training of Multimodal Question Duplicity Detection for Software Engineering Domain* [online]. 2022 [cit. 2024-04-29]. Dostupné z: https://www.researchgate.net/publication/359518098_MQDD_--_Pre-training_of_Multimodal_Question_Duplicity_Detection_for_Software_Engineering_Domain\#fullTextFileContent..

Quantexa *What is Data Matching & How Does it Work?* [online]. 2024 [cit. 2024-11-13]. Dostupné z: https://www.quantexa.com/resources/data-matching/..

scikit-learn developers *scikit-learn: machine learning in Python — scikit-learn 1.4.2 documentation* [online]. 2024 [cit. 2024-04-30]. Dostupné z: https://scikit-learn.org/stable/index.html..

Shearer, Michael *Hands-on entity resolution: a practical guide to data matching with python*. Beijing Boston Farnham Sebastopol Tokyo, 2024. 1 s. ISBN 978-1-0981-4848-5 978-1-0981-4845-4..

Stepanenko, Roman *What is Entity Resolution* [online]. 2024 [cit. 2024-11-14]. Dostupné z: https://recordlinker.com/what-is-entity-resolution/..

Tejada, Simon *Learning Object Identification Rules for Information Extraction*. Austin, TX, 2002. . . Dostupné z: https://usc-isi-i2.github.io/papers/tejada02-thesis.pdf..

TensorFlow *Machine learning education | TensorFlow* [online]. 2024 [cit. 2024-04-29]. Dostupné z: https://www.tensorflow.org/resources/learn-ml..

Tilores *Fuzzy Matching Algorithms for Data Deduplication and Linking* [online]. 2023 [cit. 2024-04-28]. Dostupné z: https://tilores.io/fuzzy-matching-algorithms..

Winkler, William *Matching And Record Linkage*. Wiley Interdisciplinary Reviews: Computational Statistics. 2014, roč. 6. ISSN 9780471598527. Dostupné z: ..

Witten, Ian H.; Frank, Eibe; Hall, Mark A.; Pal, Christopher J. *Data mining: practical machine learning tools and techniques*. Amsterdam Boston Heidelberg London New York Oxford Paris San Diego San Francisco Singapore Sydney Tokyo, 2017. 621 s. ISBN 978-0-12-804291-5..