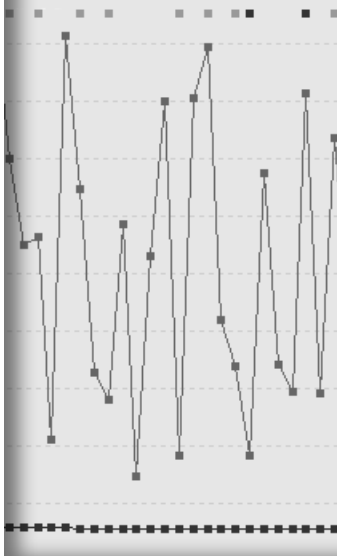
Prezentace pro „majitele dat“

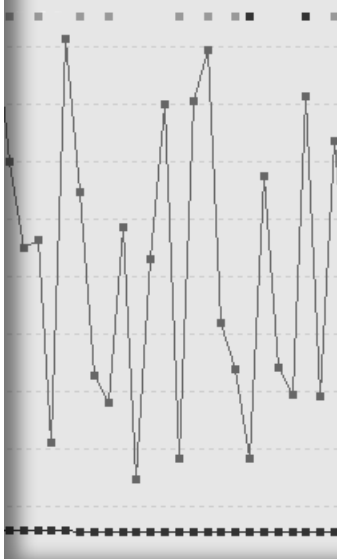
Generování umělých dat

Ludmila Svobodová



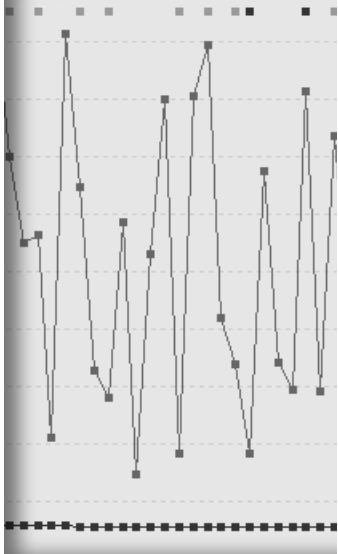
Popis struktury dat

- Vygenerovaná data budou obsahovat hodnoty týkající se nemocenské, osoby a jejího zaměstnání.
- Uvažuji tedy nemocenskou, osobu a zaměstnání, jako tři nejzákladnější skupiny, které se navzájem ovlivňují a tedy má cenu řešit vztahy mezi nimi.
- **Nemocenská**
  - Jako základní charakterizující atributy jsem zvolila důvod, diagnózu a délku trvání nemocenské.
- **Osoba**
  - Jako základní charakterizující atributy jsem zvolila pohlaví a věk.
- **Zaměstnání**
  - Jako základní charakterizující atributy jsem zvolila obor, okres a plat.



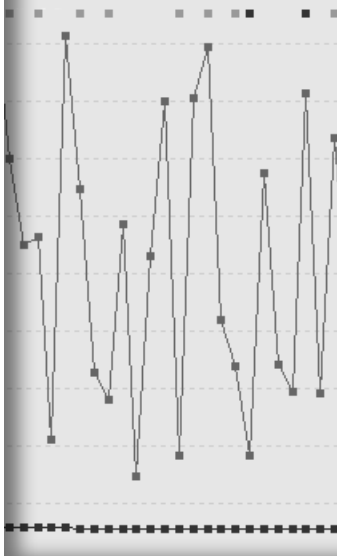
Charakteristiky atributů 1/4

- Data jsou v aplikaci charakterizována pomocí následujících atributů.
- **Věk (ordinální atribut)**
  - Integer number
  - Rozdělení na 11 intervalů délky 5
- **Pohlaví (nominální atribut)**
  - Text
  - 2 kategorie
    - Žena
    - Muž
- **Plat (ordinální atribut)**
  - Integer number
  - Rozdělení na 20 intervalů délky 10 000
  - Plat jsem v žádném zadaném vztahu neuplatnila, ale k doméně zajisté patří, proto jsem ho do generování zanesla.



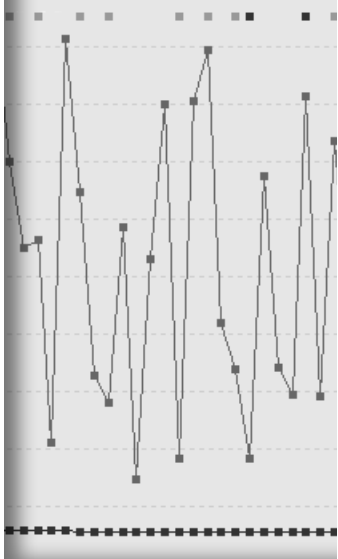
Charakteristiky atributů 2/4

- **Obor (nominální atribut)**
  - Text
  - 8 kategorií
    - Řídící pracovníci
    - Techničtí a odborní pracovníci
    - Úředníci
    - Služby a prodej
    - Kvalifikovaní pracovníci v zemědělství, lesnictví a rybářství
    - Řemeslníci, opraváři
    - Obsluha strojů
    - Pomocní a nekvalifikovaní pracovníci



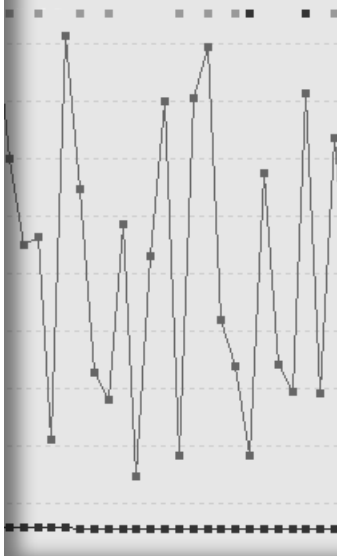
Charakteristiky atributů 3/4

- **Okres (nominální atribut)**
  - Text
  - 77 kategorií
    - Benešov
    - Beroun
    - Kladno
    - … (obsahuje všechny okresy ČR)
- **Délka (ordinální atribut)**
  - Integer number
  - Rozdělení na 17 intervalů délky 10



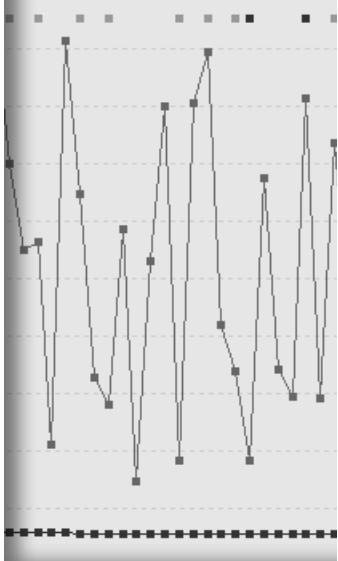
Charakteristiky atributů 4/4

- **Důvod (nominální atribut)**
  - Text
  - 2 kategorie
    - Nemoc
    - Úraz
- **Diagnóza (nominální atribut)**
  - Text
  - 5 kategorií
    - Nemoc dýchací soustavy
    - Těhotenství, porod a šestinedělí
    - Poranění, otravy aj. následky vnějších příčin
    - Novotvary
    - Nemoc svalové a kosterní soustavy



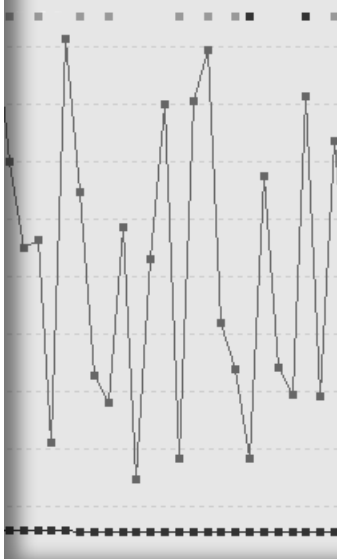
Vztahy v datech

- Vztahy zanesené do generovaných dat jsou dvojího typu.
  - Vztahy, které by se daly považovat za elementární doménové znalosti (např. těhotné jsou pouze ženy).
  - Vztahy, které již nemusí být pro úplného laika na první pohled zřejmé.
    - Jejich zavedení do dat by mělo přispět k tomu, aby se uměle generovaná data začala co nejvíce přibližovat realitě.



Základní vztahy v datech

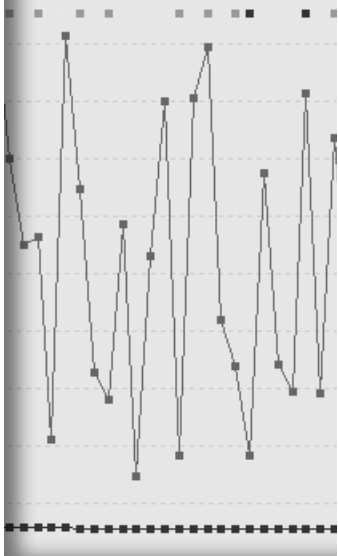
- Déle na nemocenské bývají ženy.
- Nejdelší nemocenské jsou ty s diagnózou novotvarů a tuberkulózy
- Nejkratší dobu trvají nemocenské z důvodu nemoci dýchacích cest.
- Nemocenská z důvodu porodu se vyskytuje pouze u žen a to především do věku 45 let. Nejvíce těchto diagnóz se potom vyskytuje u žen od 25 do 35 let.
- Absolutní většina nemocenských je z důvodu nemoci. Zbývající případy připadají na pracovní a ostatní úrazy.
- Čím je člověk starší, tím je delší trvání jeho nemocenské.



Pokročilé vztahy v datech 1/2

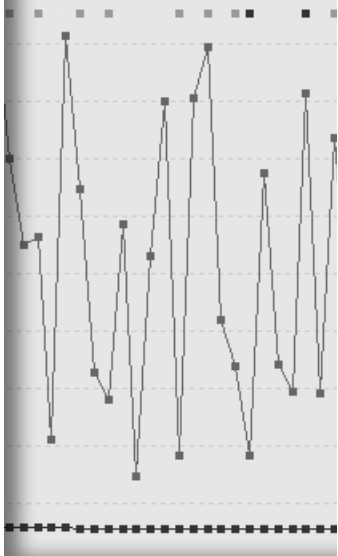
- V rámci těchto vztahů bylo cílem co nejvíce upravit vztahy mezi oborem zaměstnání a diagnózou.
- V reálných datech jsou tyto vztahy výrazné a nepopiratelné, tedy jsem chtěla, aby do výsledných vygenerovaných dat bylo zaneseno aspoň to v jakém oboru, jaká diagnóza převládá či je naopak nejméně častá.
  - **Zanesení těchto vztahů do generovaných dat se ukázalo jako největší problém evoluce. O tomto problému více na slidu Výsledná evoluce.**
- Tyto vztahy jsou znázorněné v tabulce viz níže.

Diagnóza	Nejvíce případů	Nejméně případů
Nemoc dýchací soustavy	Řemeslníci, úředníci	Řídící pracovníci, pracovníci v zemědělství, lesnictví a rybářství
Těhotenství, porod	Úřednice, služby a prodej	Obsluha strojů
Poranění aj. následky vnějších příčin	Řemeslníci, opraváři	Techničtí a odborní pracovníci
Novotvary	Úředníci	Obsluha strojů
Nemoc svalové soustavy	Nekvalifikovaní pracovníci, řemeslníci	Techničtí a řídící pracovníci



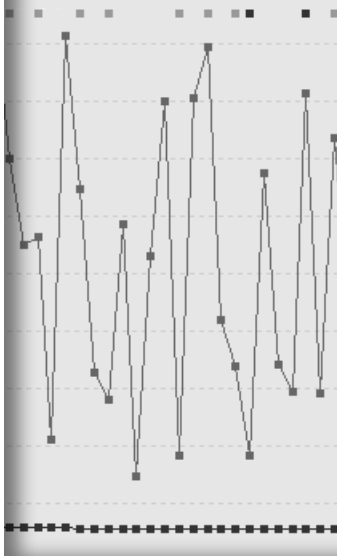
Pokročilé vztahy v datech 2/2

- Vztahy týkající se délky a počtu nemocenských v jednotlivých okresech.
- Opět jsem se zaměřila na pokrytí extrémů, tedy že dlouhé nemocenské jsou v Šumperku a krátké v Mladé Boleslavi a Praze.
- Nejvíce nemocenských je v okresech Prachatice a Šumperk, nejméně se jich vyskytuje v Praze a Jeseníku.
  - **Tento vztah také nebyl do dat nakonec zanesen viz slide Výsledná evoluce.**



Výsledná evoluce 1/3

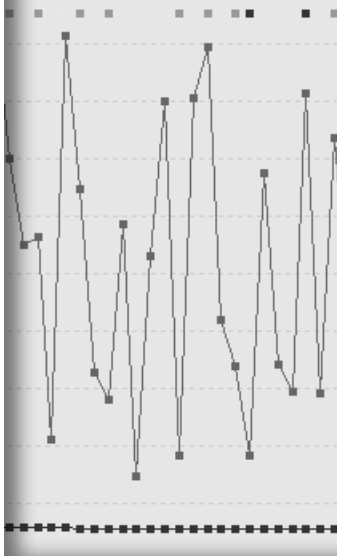
- Po zadání všech vztahů evoluce dohromady, se některé vztahy navzájem brzdí.
  - Největší problém nastal u vztahů mezi oborem zaměstnání a diagnózou.
- Pokud udržuji frekvence diagnóz podle reálných dat, tak převládá diagnóza nemoci dýchacích cest. Rozdělení četnosti diagnózy tak, aby u některých oborů byla méně a u jiných více, se ukázalo jako veliká brzda evoluce.
  - Prvně jsem se tedy snažila vyřešit tento problém.
- Když ale snížím rozdíly mezi jednotlivými typy diagnóz, tak to evoluci trochu pomůže.
  - Ale problém to stále neřeší a evoluce ani cca po 3 000 iteracích není na cestě k výsledku.
- Prioritou je zanést do dat, co nejvíce to bude možné, vztahy mezi oborem a diagnózou.
  - Zrušila jsem tedy nerovnoměrné rozložení frekvencí u diagnózy.



Výsledná evoluce 2/3

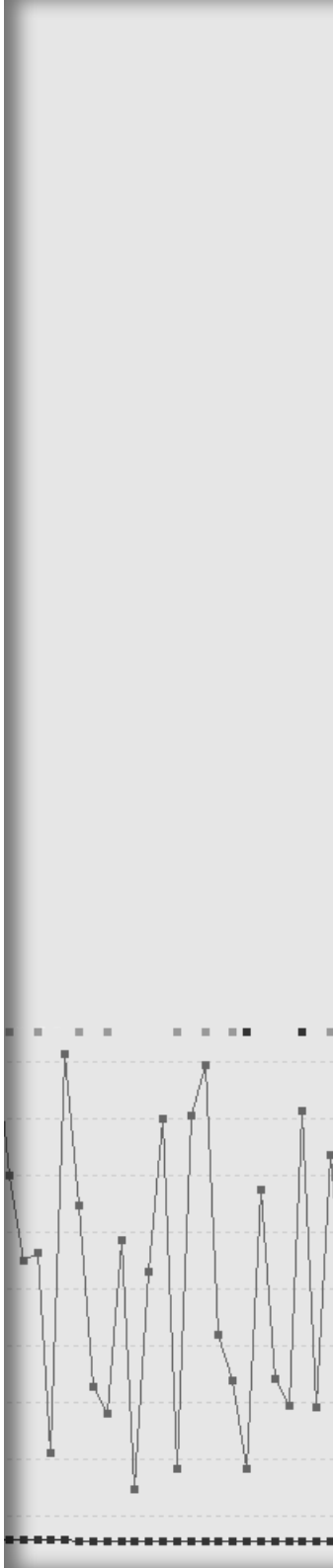
- Problém to ale pořád zcela neřeší.
  - Proto jsem se rozhodla celou problematiku ještě více zjednodušit.
- Nové zjednodušené zadání vztahů viz tabulka níže.
- Až po zadání takto zjednodušených vztahů začala být výsledná evoluce úspěšnější.

Diagnóza	Nejvíce případů	Nejméně případů
Nemoc dýchací soustavy	Řemeslníci, opraváři	Pracovníci v zemědělství, lesnictví a rybářství
Těhotenství, porod	Služby a prodej	Obsluha strojů
Poranění aj. následky vnějších příčin	Řemeslníci, opraváři	Řídící pracovníci
Novotvary	Úředníci	Obsluha strojů
Nemoc svalové soustavy	Nekvalifikovaní pracovníci	Techničtí pracovníci



Výsledná evoluce 3/3

- Vztah nejvíce nemocenských je v okresech Prachatice a Šumperk, nejméně se jich vyskytuje v Praze a Jeseníku.
  - Nebyl zanesen do výsledné evoluce, protože počet okresů na generovaná data je poměrně velký.
  - Pravidla, která by hlídala frekvenci daného vztahu by tedy byla dost omezující a v této fázi by znamenala značné zpomalení evoluce.



Ověření vygenerovaných dat

- Vygenerovaná data jsem nahrála do LMWorkspace a ověřila, že obsahují požadované vztahy.
- Pro vybrané 3 vztahy jsem vytvořila v LMWorkspace klasickou úlohu a ověřila, že v datech byl evolucí vygenerovaný vztah nalezen.
  - Obor úředníci mají mít větší četnost novotvarů.

    0.608 **Obor**(*úředníci*) >÷< **Diagnoza**(*novotvary*)

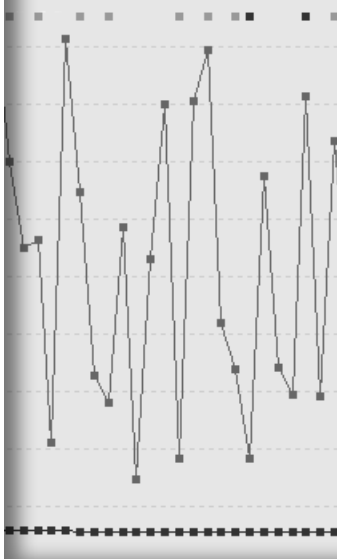
  - Obor techničtí a odborní pracovníci mají mít menší četnost nemocí svalové soustavy.

    -0.560 **Obor**(*techničtí a odborní pracovníci*) >÷< **Diagnoza**(*nemoc svalové a kosterní soustavy*)

  - Diagnóza těhotenství má nejvíce případů u osob ve věku <25;34>.

    0.604 **Diagnoza**(*těhotenství, porod a šestinedělí*) >÷< **Vek <25;35>**(*<25;34>*)

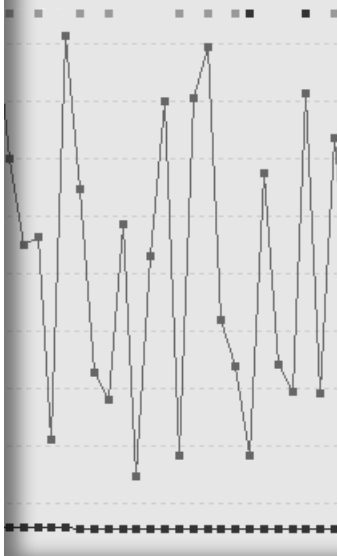
- Vztahy jsou tedy do dat opravdu zanesené.



Zhodnocení kvality dat

- Data v základu splňují to, co by měla a reflektují nejzákladnější doménové znalosti.
  - V datech není nic, co by vyloženě šlo proti logice domény.
- Velkou slabinou je, že se některé vztahy musely pro finální evoluci zjednodušit.
- Řešením je vztahy opět nabalovat zpátky.
  - Pokud se tak bude provádět postupně, bude snadnější podchytit momenty, kdy se vztahy v evoluci začnou již neslučitelně prát a omezovat.
- Slabinou dat je také fakt, že se jedná o širokou doménu.

  Šlo by vztahy rozšiřovat a přidávat další data, na kterých by mohly být nemocenské závislé.
- Optimalizace celé evoluce.
  - Již nyní je evoluce pomalá a dochází k výsledkům po desítkách hodin.
  - Zadání vztahů jinak např. pomocí jiného typu úloh by mohlo vést k optimalizaci a zrychlení celé evoluce.



Závěr

- Data nejsou dokonalá, ale v základu drží logiku domény.
- Nejvíce problémové jsou vztahy mezi oborem a diagnózou.
  - Těchto vztahů je v evoluci hodně a vzájemně se přetahují.
  - Řešeno zjednodušením, ale teď by bylo lepší postupně vztahy zkoušet a přitvrzovat. Takto by se dalo jádro konfliktu zúžit.
- Moje evoluce je zatím velice pomalá a fáze ladění je velice zdlouhavá a časově náročná a i malá změna v evoluci nebo vztazích může vést k úplně jinému výsledku.