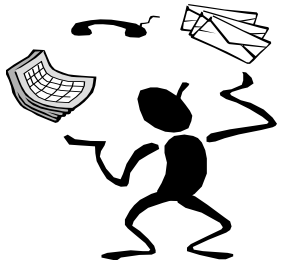
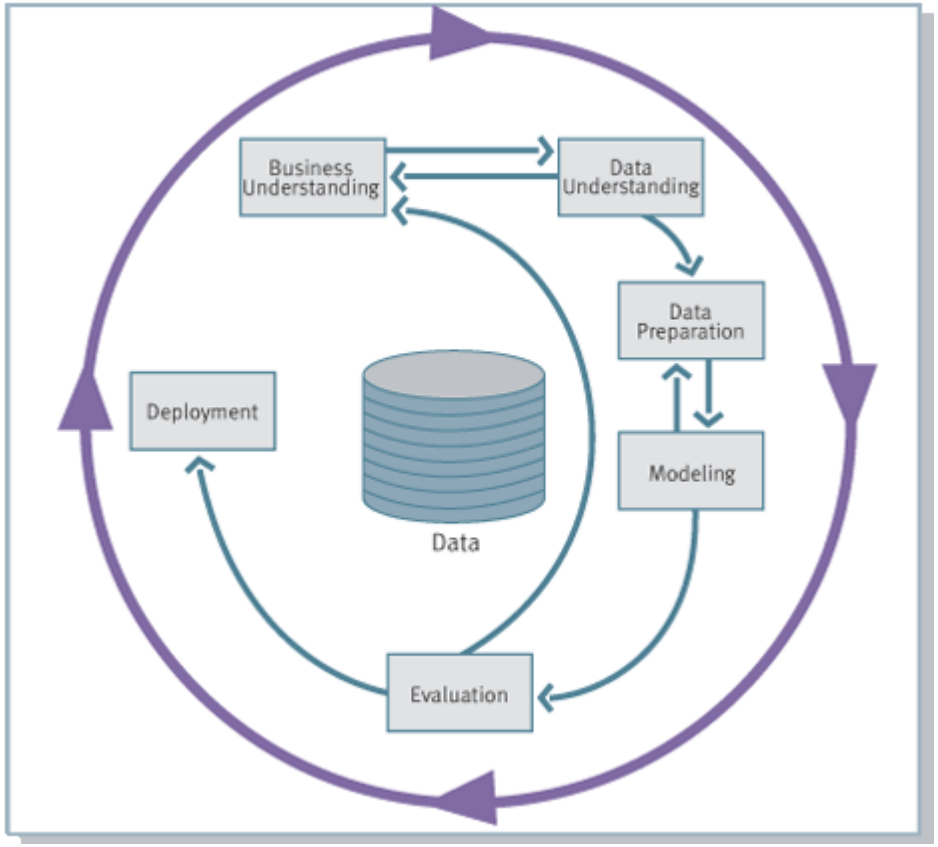
Předzpracování dat

Lenka Vysloužilová

Metodika CRISP-DM (www.crisp-dm.org)





Zadaní – Business Understanding

- pochopení cílů úlohy
- náklady
- hodnotí se přínos
- stanovení předběžného plánu
- forma předání dat
  - anonymizace dat
  - formát dat

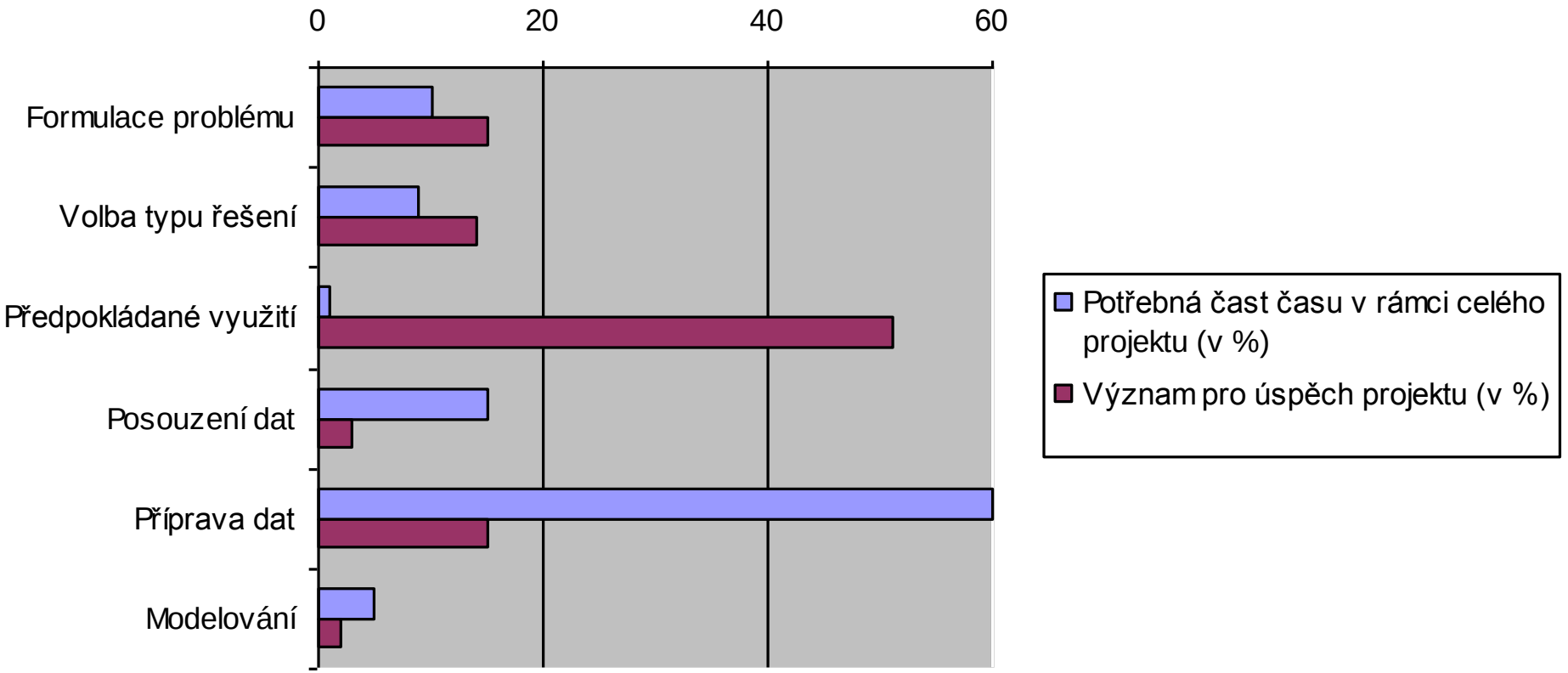
Analýza dat – Data Understanding

- získání základní představy o datech
- kvalita dat (chybějící údaje)
- descriptivní charakteristiky dat
  - četnosti hodnot (histogramy)
  - minima, maxima, průměry
- použití vizualizačních technik

Příprava dat – Data Preparation

- příprava dat pro modelování
  - selekce atributů – výběr relevantních atributů
  - čištění dat
  - získávání odvozených atributů
  - převod typů dat
  - transformace dat do jedné velké tabulky
  - formátování pro jednotlivé modelovací techniky
- nejpracnější část celého procesu
- často se provádí opakovaně

Časové nároky procesu?

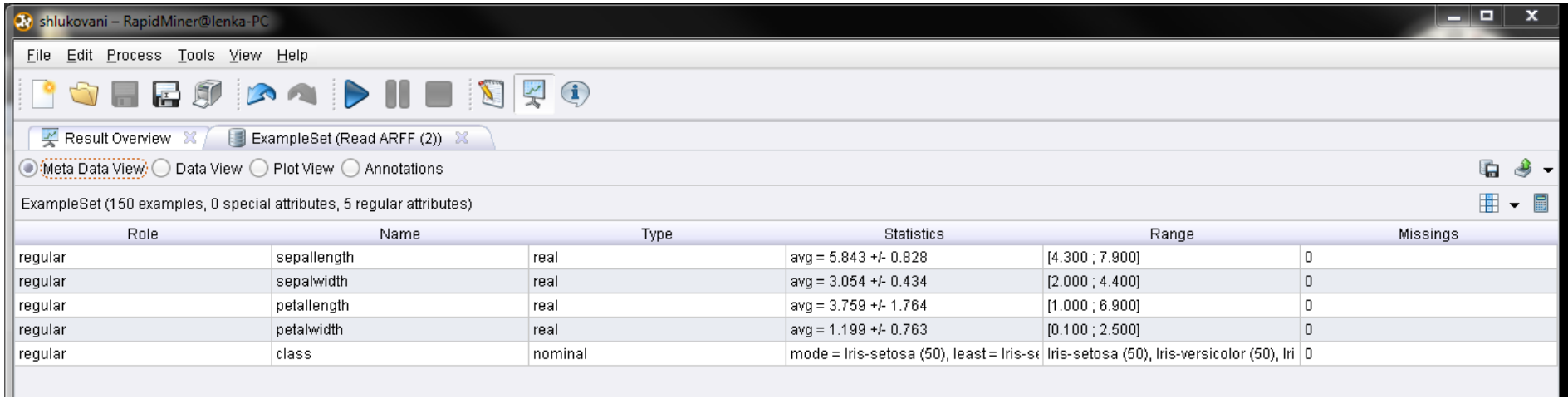


Problémy reálných dat?

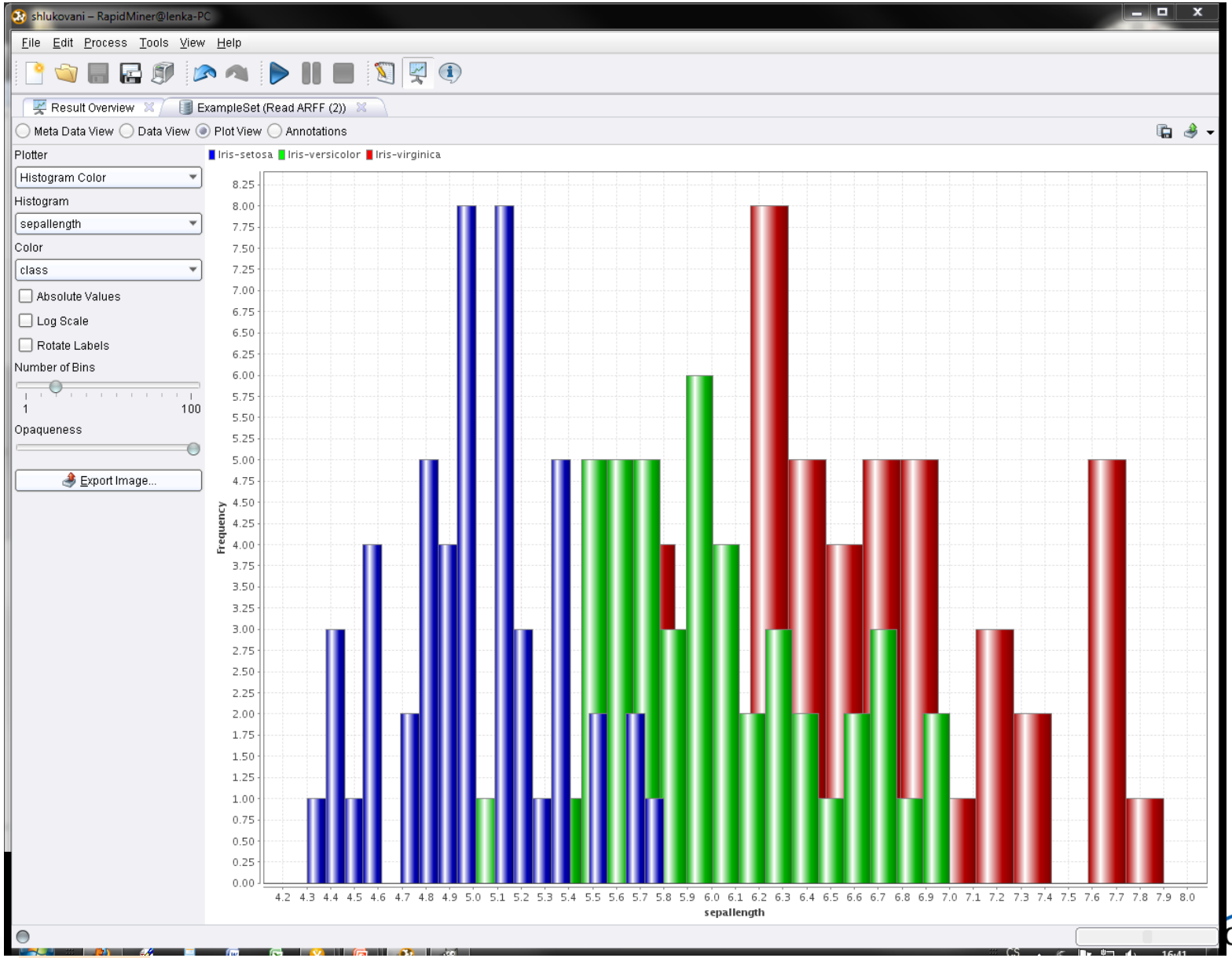
- Data obsahují **špatné údaje** způsobené chybami měřicích přístrojů i lidské obsluhy (outliers)
- **Nevyplněné údaje**
- Data jsou popsána pomocí **příliš mnoha atributů** - není zřejmé, které z nich jsou pro řešení zvolené úlohy relevantní. Úspěch modelování závisí na volbě vhodné množiny atributů (PAC učení)
- Data mají formu **složitého relačního schématu**, nikoliv jediné tabulky předpokládané atributovými metodami strojového učení

Analýza dat – Data Understanding

Analýza dat



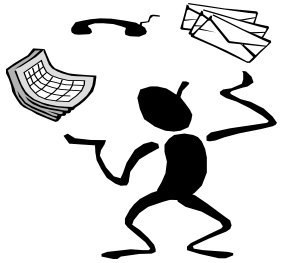
Role	Name	Type	Statistics	Range	Missings
regular	sepallength	real	avg = 5.843 +/- 0.828	[4.300 ; 7.900]	0
regular	sepalwidth	real	avg = 3.054 +/- 0.434	[2.000 ; 4.400]	0
regular	petallength	real	avg = 3.759 +/- 1.764	[1.000 ; 6.900]	0
regular	petalwidth	real	avg = 1.199 +/- 0.763	[0.100 ; 2.500]	0
regular	class	nominal	mode = Iris-setosa (50), least = Iris-se	Iris-setosa (50), Iris-versicolor (50), Iri	0

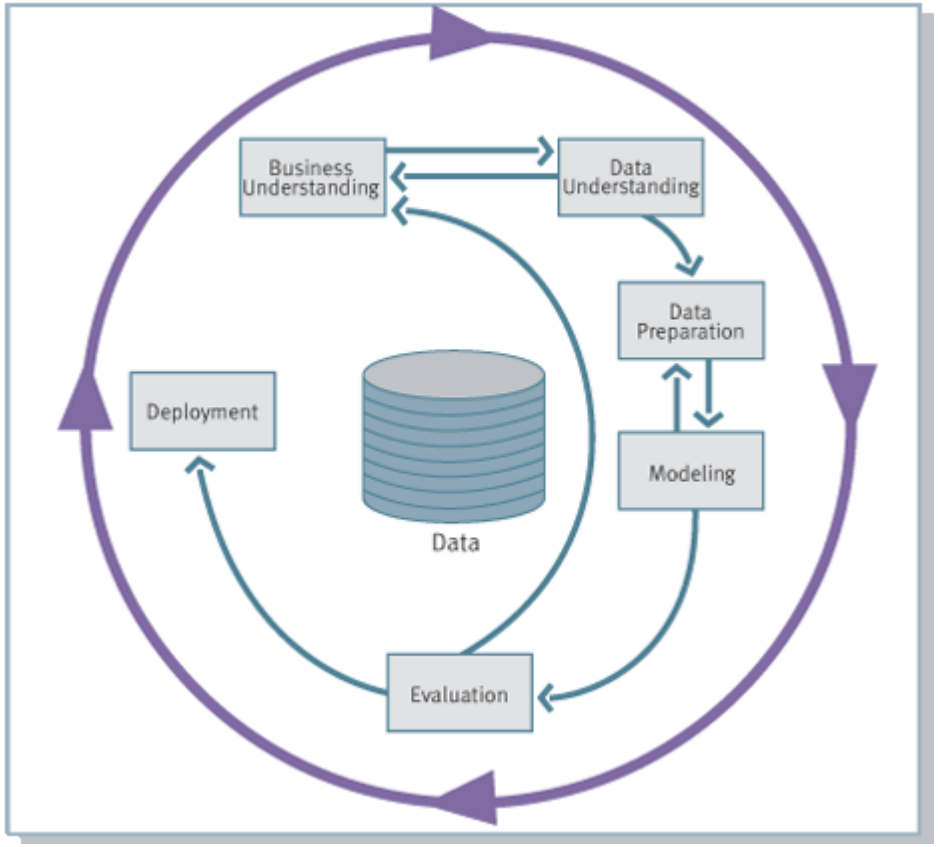


laboratory Gerstner

Příprava dat pro modelování

Metodika CRISP-DM (www.crisp-dm.org)





Příprava dat – Data Preparation

- transformace dat do jedné velké tabulky
- čištění dat
  - oprava chyb a odlehlých hodnot
  - převod typů dat
  - náhrada chybějících hodnot
- získávání odvozených atributů
- vzorkování dat
- selekce atributů – výběr relevantních atributů
  - dekompozice x selekce
- formátování pro jednotlivé modelovací techniky

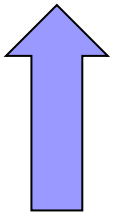
Transformace dat do jedné tabulky

- 1:1
  - prakticky pouze doplnění tabulky o nové atributy
- 1:N
  - vytvoření agregovaných hodnot
  - součet, min, max, průměr, regresní křivka
  - majoritní hodnota, počet různých hodnot, výskyt konkrétní hodnoty
  - do této skupiny patří časové řady
- M:N
  - nutná volba úlohy, zda chceme 1:N nebo 1:M
- Propozicionalizace

Datová tabulka

Filtrování instancí

Sepallength	Sepalwidth	Petallength	Petalwidth	Class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.7	1.5	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
6.3	3.3	6.0	2.5	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica



Filtrování a úprava atributů

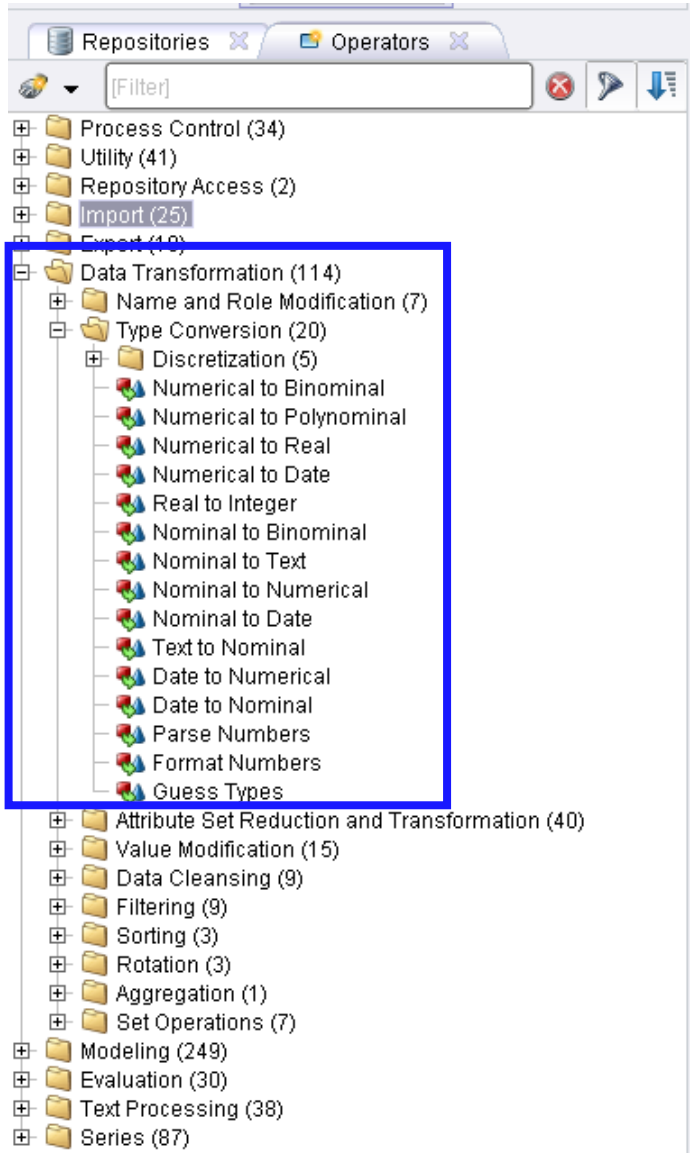
ÚPRAVA ATRIBUTŮ

Typy atributů - RapidMiner

- Binominal/polynominal
  - 2 hodnoty - muž/žena => binominal
  - více hodnot - barva(červená, modrá, zelená) => polynominal
- Numeric
  - celá čísla => integer
  - reálná čísla => real
  - jakou přesnost čísel?
  - dají se řadit
- Text
- Date

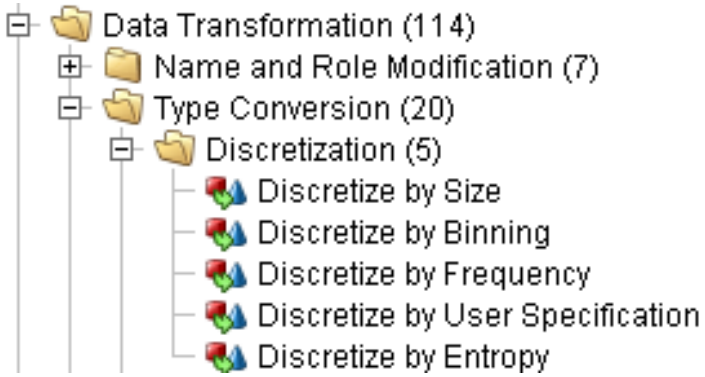
Převod typů dat

- Datum
  - reprezentace reálným číslem
  - pomocí tří atributů
  - Generate attribute umí další matematické operace
- Text
  - Text to Nominal

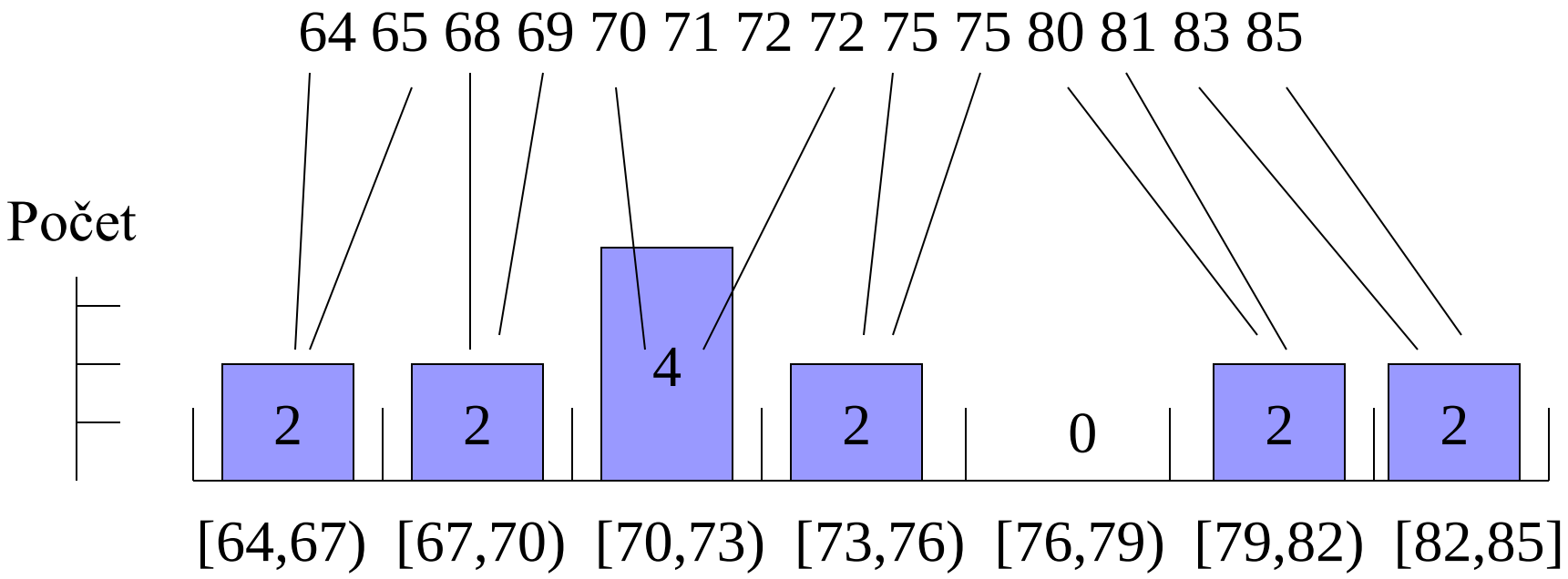


Diskretizace dat

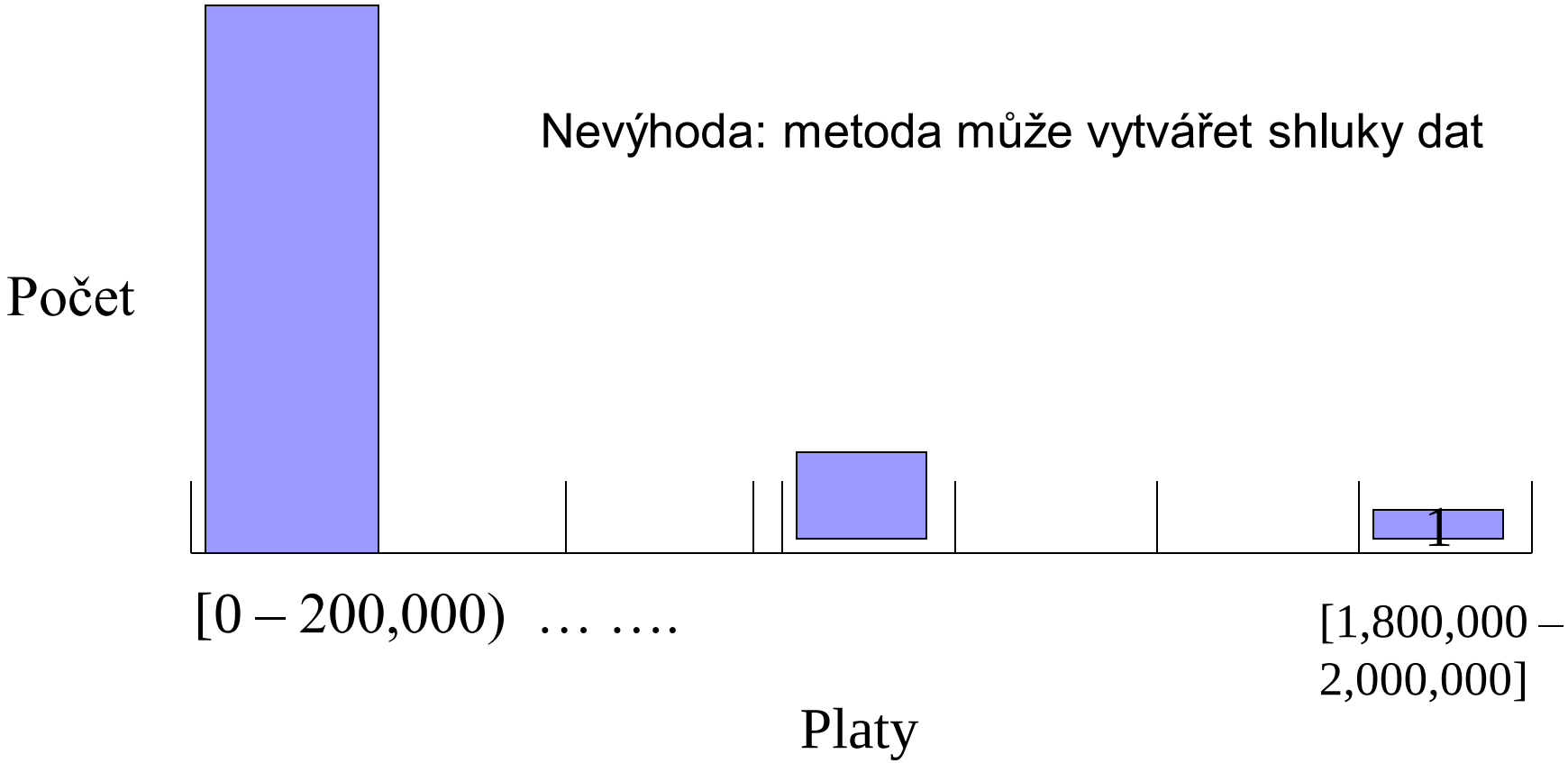
- Neinformované metody
  - ekvidistantní intervaly
  - ekvifrekvenční intervaly
- Informované metody
  - využití znalosti o příslušnosti objekt -> třída
  - strategie rozdělování nebo spojování intervalů
- Rapid Miner



Diskretizace: Ekvidistantní intervaly

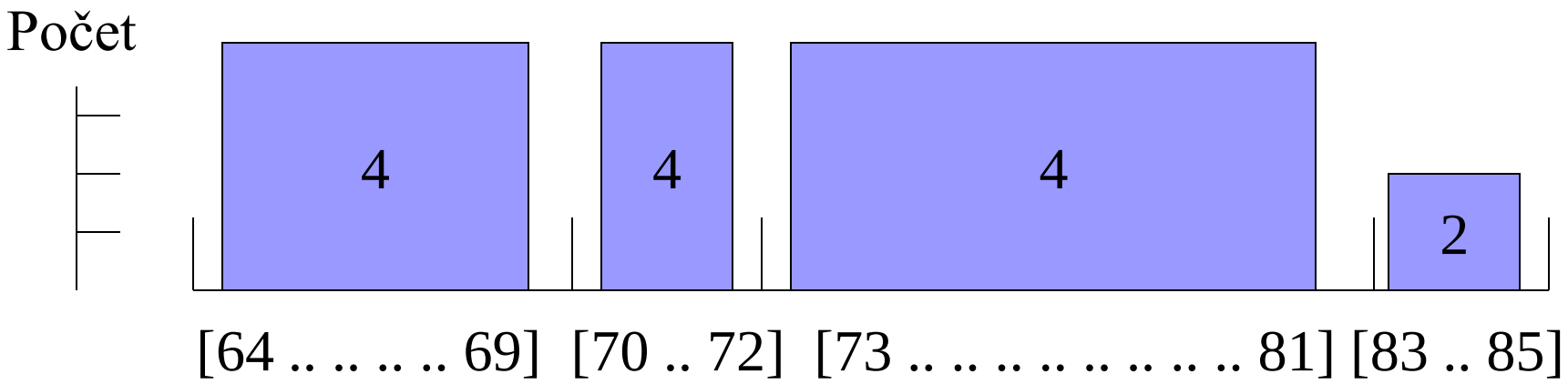


Diskretizace: Ekvidistantní intervaly



Diskretizace: Ekvifrekvenční intervaly

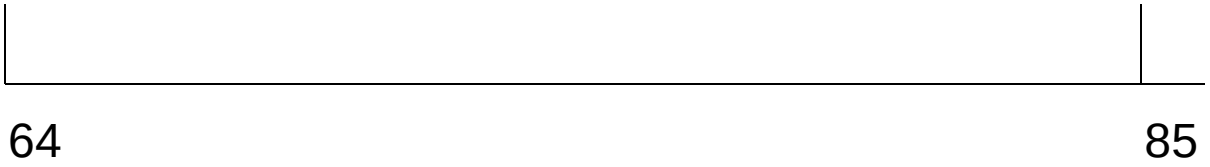
64 65 68 69 70 71 72 72 75 75 80 81 83 85



Diskretizace: v závislosti na třídě

požadujeme minimálně 3 instance na interval

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No



Normalizace dat

- Převod numerických hodnot do intervalu <0,1>
- Numerické atributy

$$a_i = \frac{v_i - \min v_i}{\max v_i - \min v_i} \quad \text{nebo} \quad a_i = \frac{v_i - Avg(v_i)}{StDev(v_i)}$$

v_i: aktuální hodnota atributu I

- RapidMiner - Normalize
  - Z-transformation (μ,σ²)
  - range transformation (min,max)

Odvozené atributy

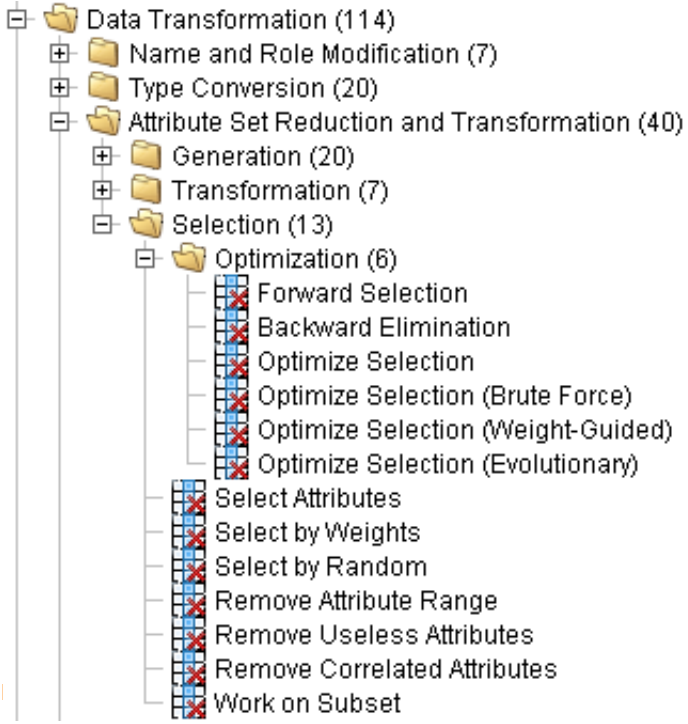
- výpočet nového atributu ze stávajících
- BMI = váha(kg) / výška(m)^2
- rodné číslo => věk a pohlaví
- agregační hodnoty
- RapidMiner
  - Generate Attribute

Redukce počtu atributů

- Analýza hlavních komponent (PCA)
  - nové atributy nelze interpretovat
  - využití pro vizualizaci dat – použijeme n nejlepších komponent
  - Principal Component Analysis
- Singular Value Decomposition (SVD)
  - Singular Value Decomposition

Selekce atributů

- hledáme takové atributy, které nejlépe přispějí ke klasifikaci
- metoda filtru
  - spočteme charakteristiku vyjadřující vhodnost atributu
  - chi-kvadrát, entropie, informační míra závislosti
  - vychází z kontingenční tabulky
  - nevýhoda: posuzujeme každý atribut samostatně
- metoda obálky
  - použití metod strojového učení



ÚPRAVA INSTANCÍ

Náhrada chybějících hodnot

- nedělat nic ?
  - některým algoritmům chybějící hodnoty nevadí, např. rozhodovací stromy
- ignorovat celou instanci
  - ideální pro data s minimem chybějících hodnot
- náhrada hodnotou „nevím“
  - none ReplaceMissingValues
- náhrada
  - průměrem, danou hodnotou, min, max ReplaceMissingValues
  - využití algoritmu pro modelování ImputeMissingValues

Outliers

- Výrazně odlišné hodnoty atributu pro danou instanci
  - Outlier pro jeden atribut nemusí být outlier i pro kombinaci atributů a naopak!
- bude v jedné z příštích přednášek

Vzorkování dat

- obrovský počet instancí - pro algoritmy pracující v dávkovém režimu nutnost
  - redukce počtu dat Sample
  - výběr podskupiny dat podle nějakého kritéria Filter Examples
  - tvorba modelů na základě podmnožin a jejich následná kombinace volí se přímo u volby modelu například
- rozdělení dat na trénovací a testovací část
  - Více v přednášce o validaci modelů
- nevyvážená data např třída A 95%, třída B 5%
    - každý objekt patří do majoritní třídy
  - různé ceny chybného rozhodnutí
  - výběr dat pro různé třídy s různou pravděpodobností

Summary

Dobrá příprava dat je klíčem k vytvoření platného a spolehlivého modelu