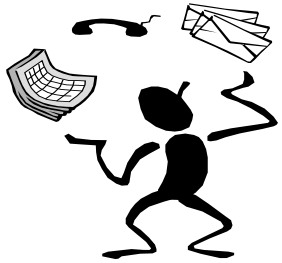
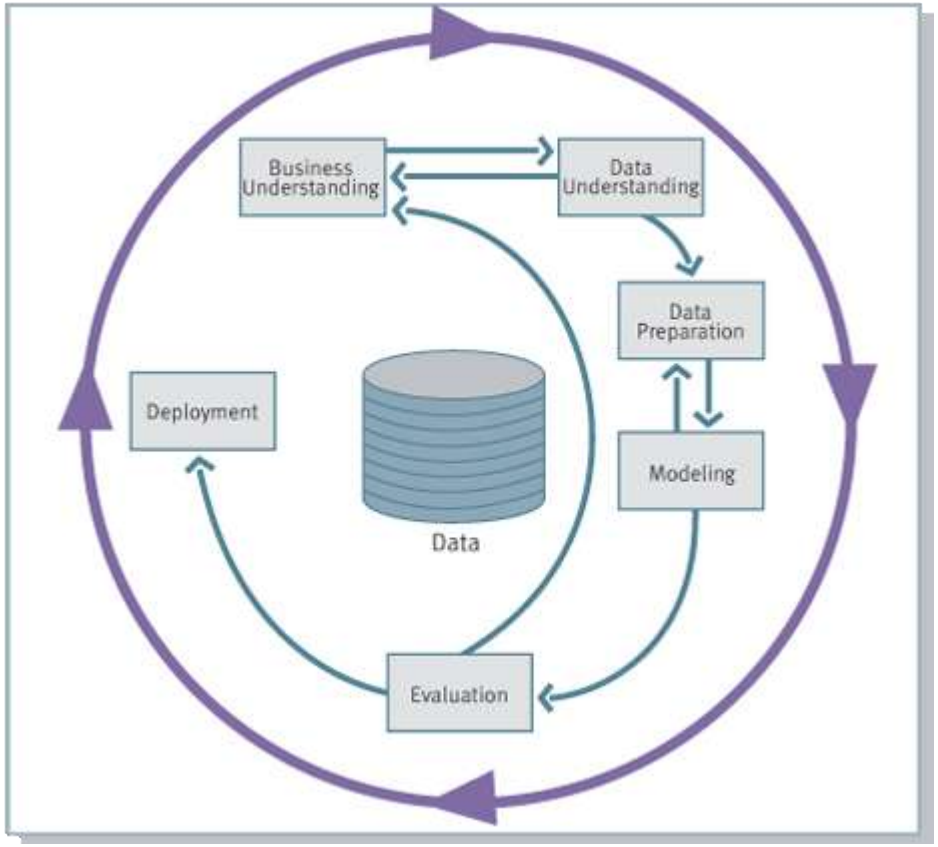
Dobývání znalostí

Lenka Nováková

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

Modelování - Modeling

- použití analytických metod (metody strojového učení)
- používá se více metod
- příklady metod
  - rozhodovací stromy
  - asociační pravidla
  - shluková analýza
  - statistické metody
- často návrat zpět k přípravě dat

Vyhodnocení - Evaluation

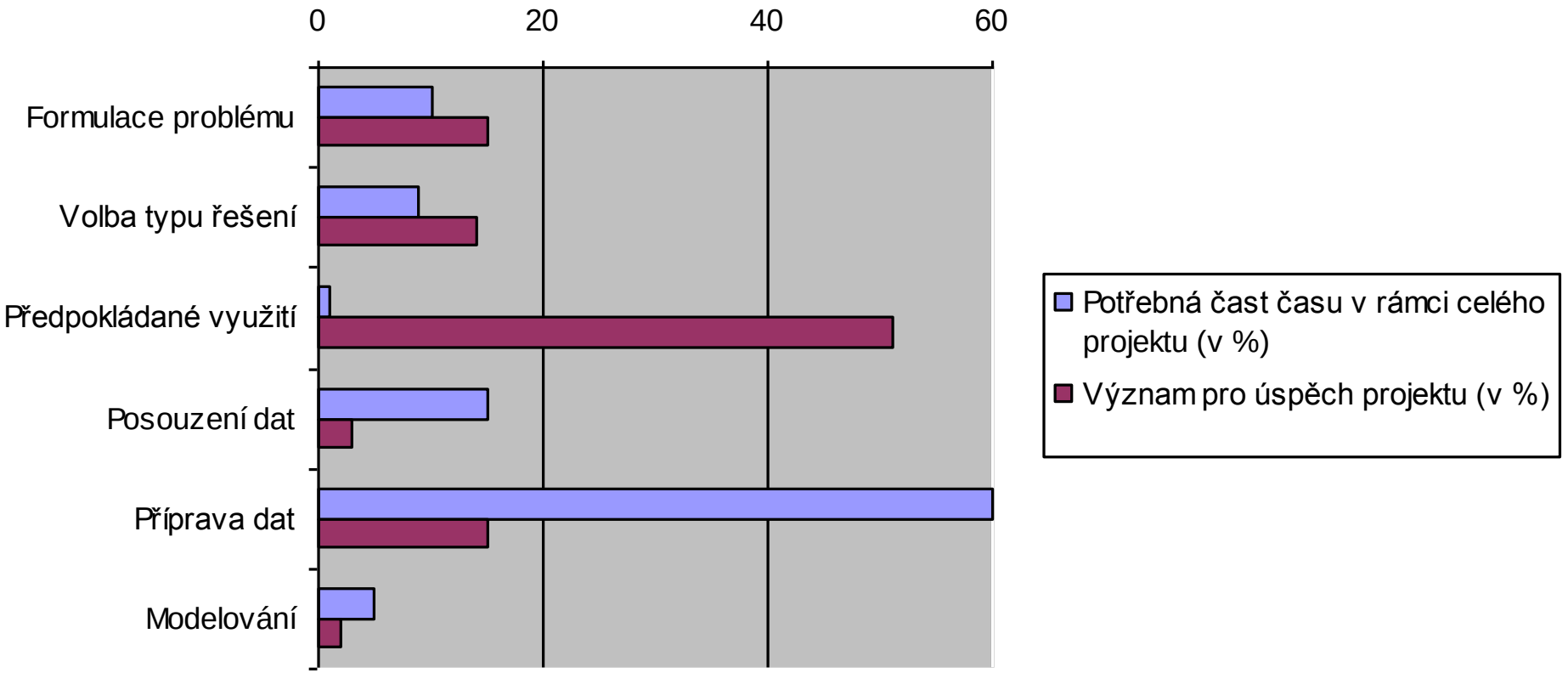
- zhodnocení dosažených výsledků modelování
- zhodnocení výsledků z pohledu zadání
- použití vizualizačních technik

- často návrat zpět na začátek celého procesu a stanovení nových cílů (úprava zadaní)

Použití - Deployment

- úprava získaných znalostí do srozumitelné formy pro zadavatele
- případně pomoc s implementací výsledků do praxe

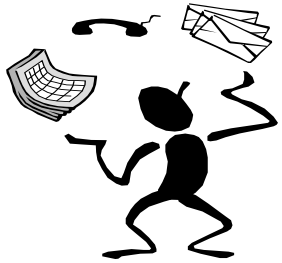
Časové nároky procesu?

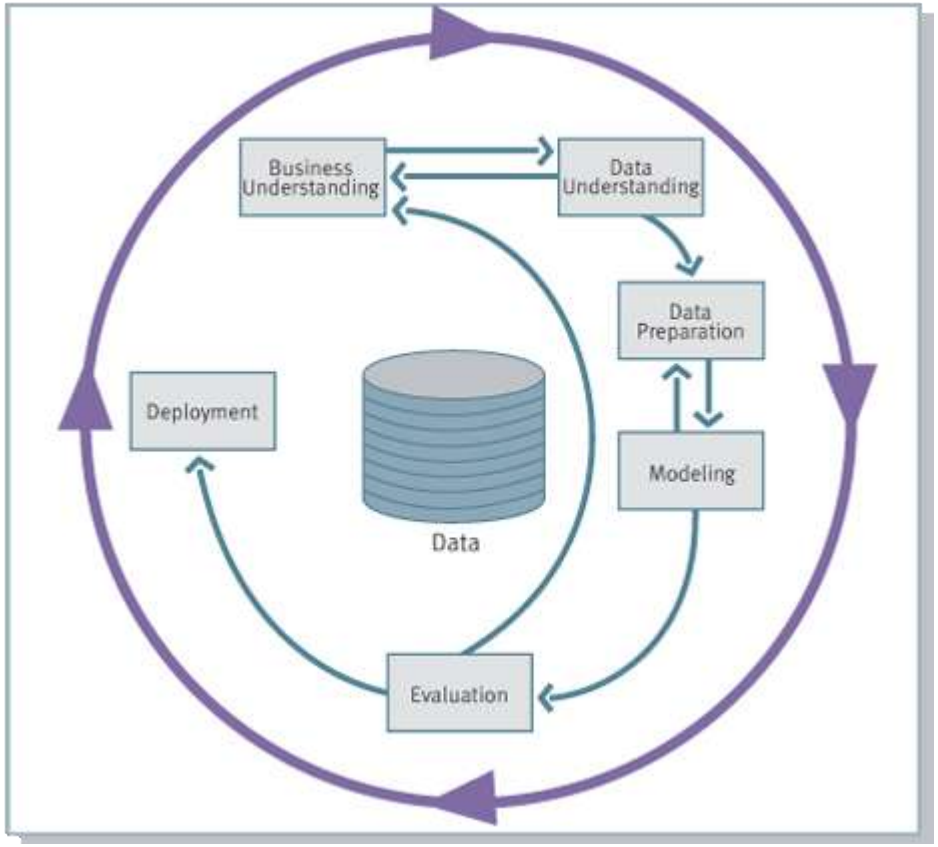


Problémy reálných dat?

- 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

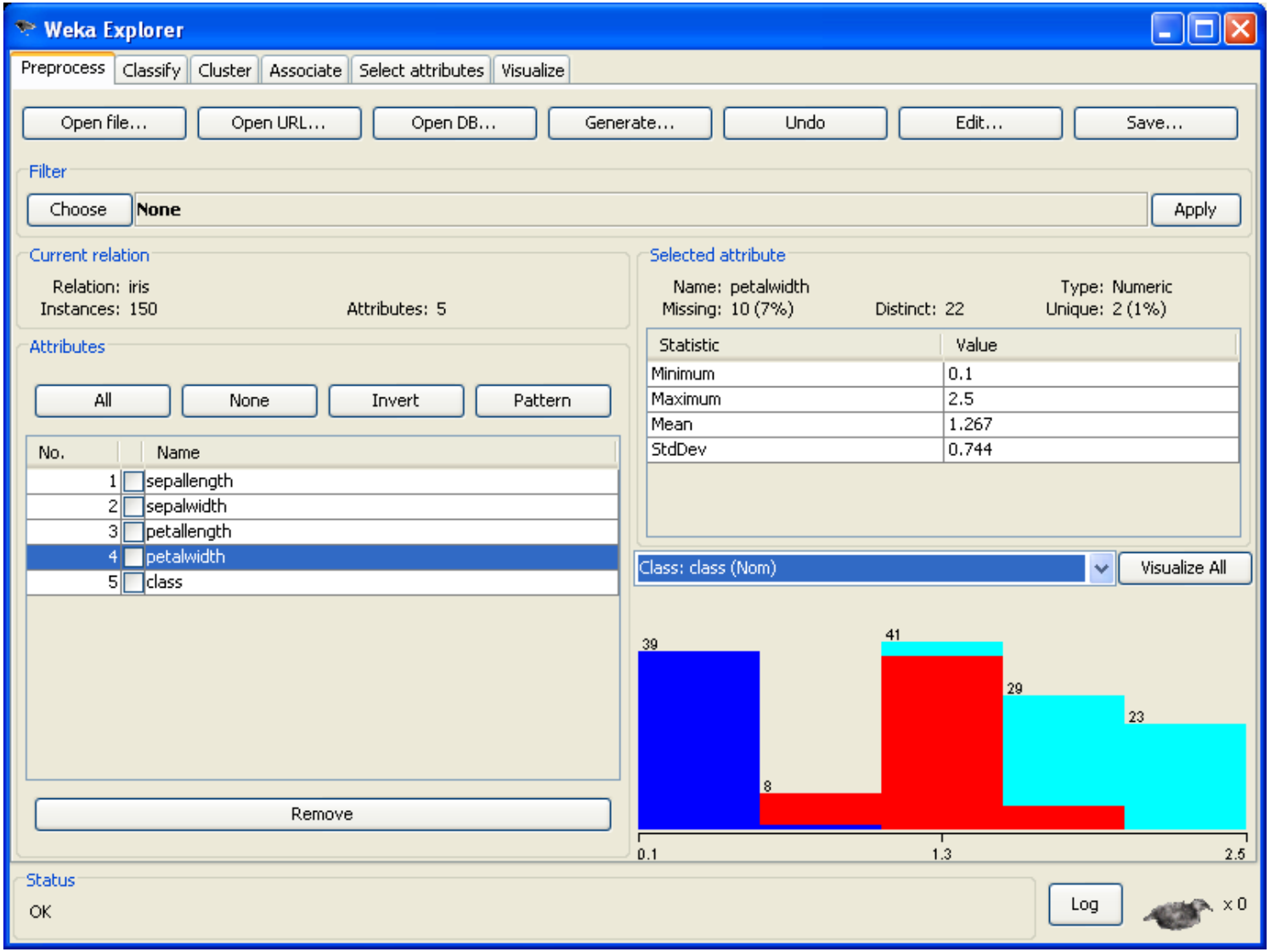
Metodika CRISP-DM (www.crisp-dm.org)





Analýza dat – Data Understanding

Analýza dat


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: petalwidth
Missing: 10 (7%)
Distinct: 22
Type: Numeric
Unique: 2 (1%)
Statistic
Value
Minimum
0.1
Maximum
2.5
Mean
1.267
StdDev
0.744
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Class: class (Nom)
Visualize All
39
8
41
29
23
0.1
1.3
2.5
Remove
Status
OK
Log
x 0

Příprava dat pro modelování

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

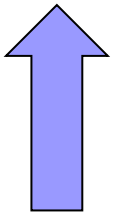
Transformace dat do jedné tabulky

- 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 - Weka

- Nominal
  - 2 hodnoty - muž/žena => binární
  - více hodnot - barva(červená, modrá, zelená) => ordinální
  - ve formátu arff pro Weku se zapisují jako {cervena, modra, zelena}
- Numeric
  - celá čísla
  - reálná čísla – jakou přesnost čísel?
  - dají se řadit
- String
- Date

Převod typů dat

- Datum
  - volba přesnosti – roky, měsíce, dny, hodin, …
  - ChangeDateFormat
  - reprezentace reálným číslem
- Řetězce – String
  - StringToNominal
  - StringToWordVector

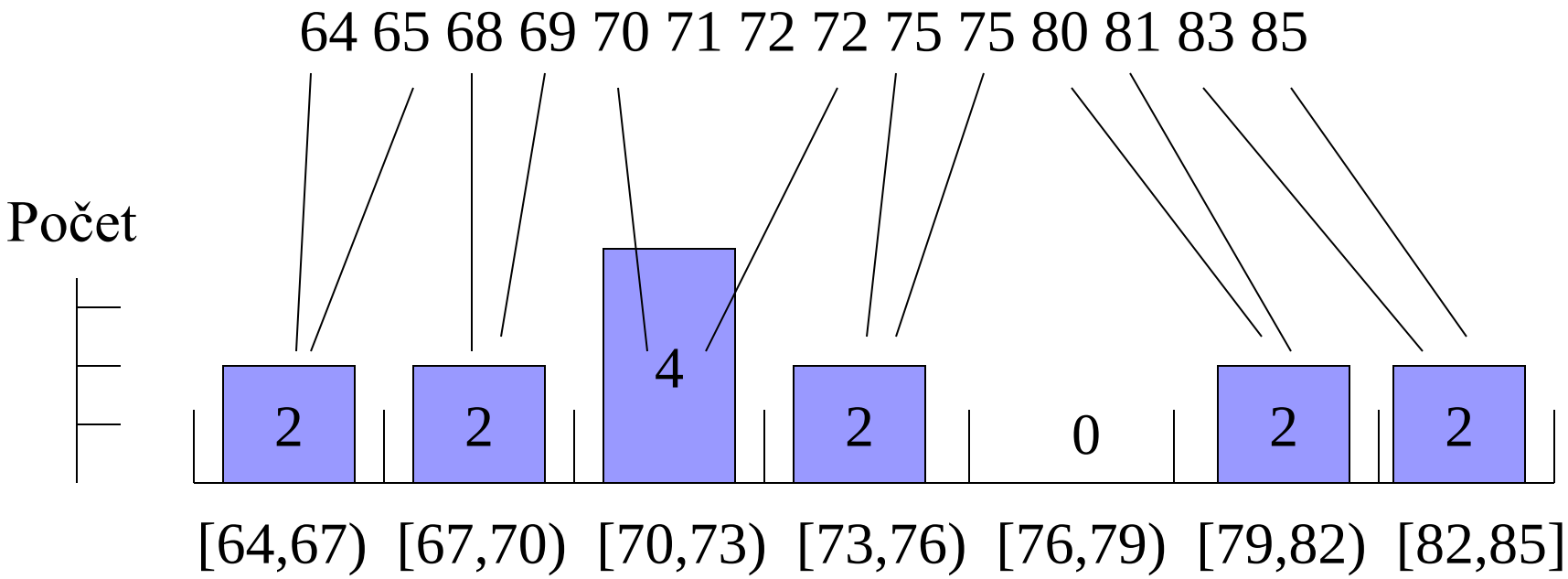
Převod typů dat

- Nominální hodnoty
  - nominální -> binární MakeIndicator, NominalToBinary
  - výběr nejfrekventovanějších hodnot
  - spojení do větších přirozených celků – města -> kraje MergeTwoValues
  - SwapValues
- Numerické hodnoty
  - diskretizace dat – volba intervalů
  - normalizace dat – volba intervalu

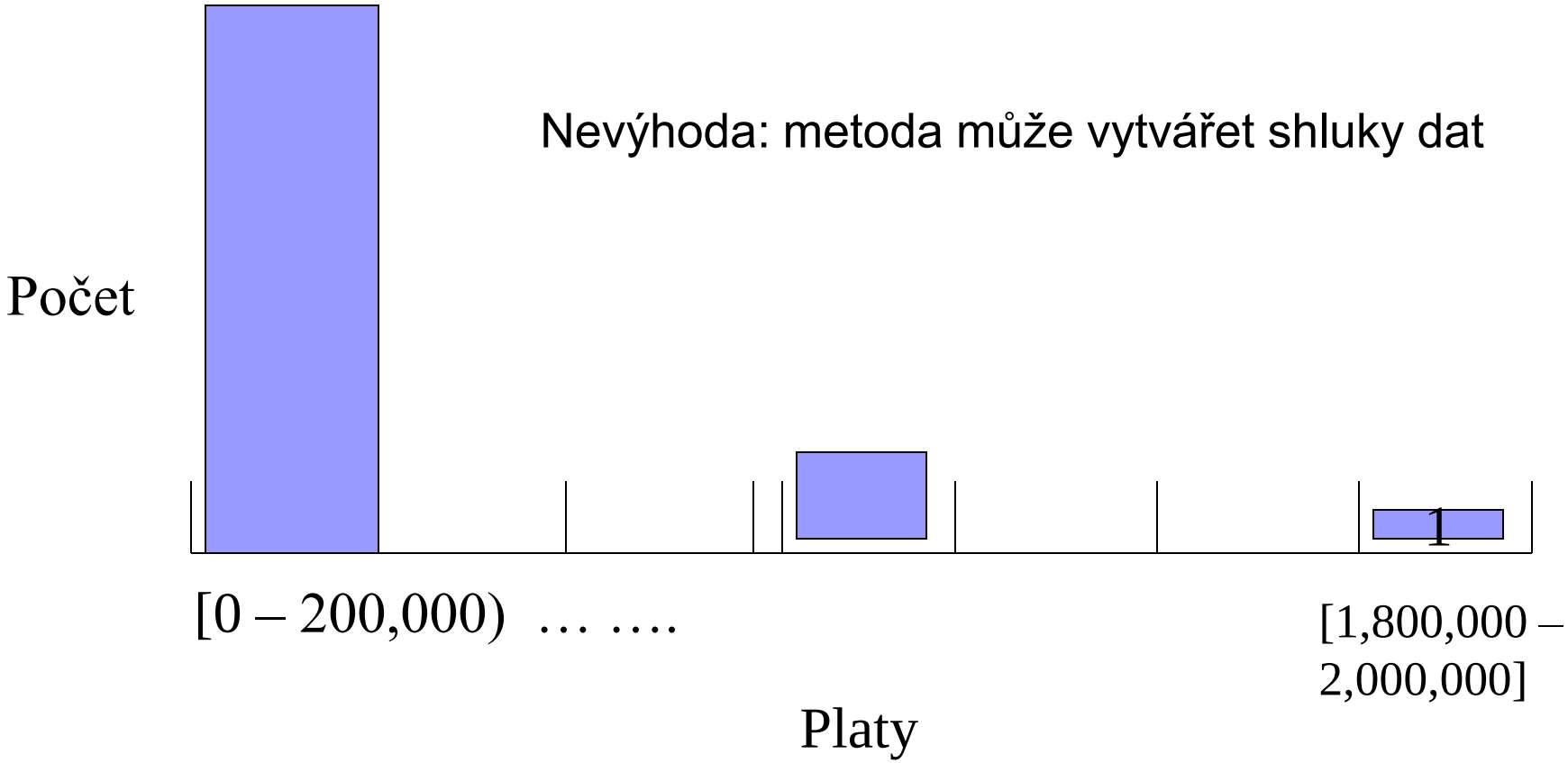
Diskretizace dat

- Neinformované metody
  - ekvidistantní intervaly
  - ekvifrekvenční intervaly
- Informované metody
  - využití znalosti o příslušnosti objekt -> třída
  - strategie rozdělování nebo spojování intervalů
- Weka
  - Discretize unsupervised i supervised

Diskretizace: Ekvidistantní intervaly

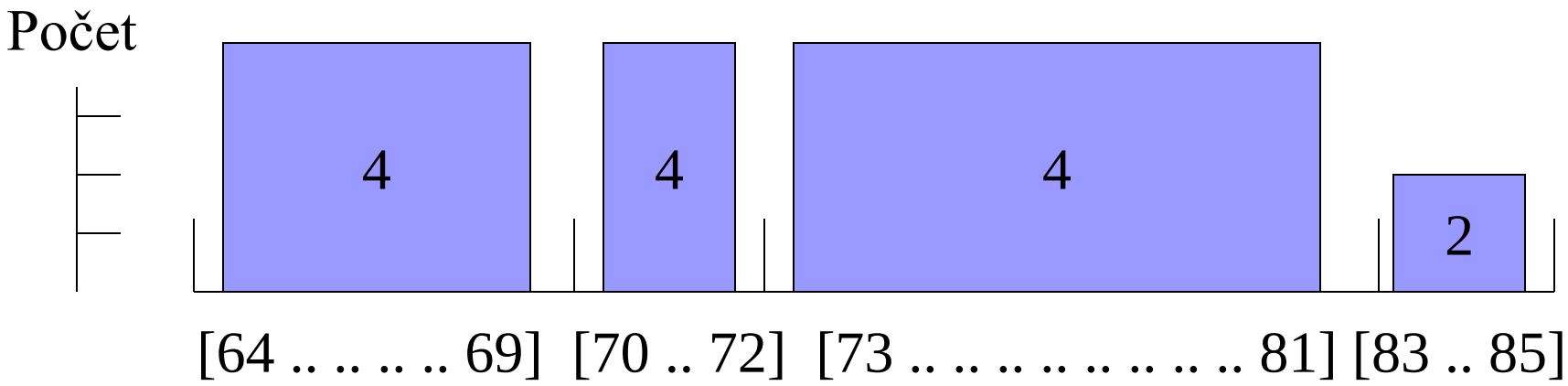


Diskretizace: Ekvidistantní intervaly



Diskretizace: Ekvifrekvenční intervaly

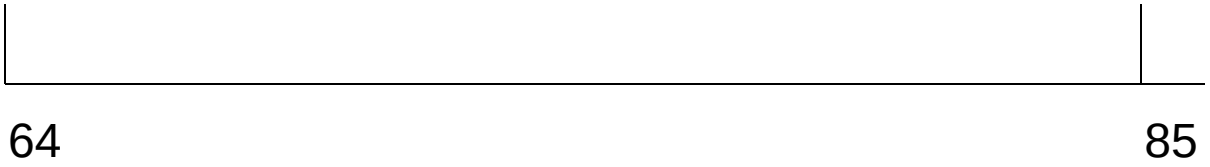
64 65 68 69 70 71 72 72 75 75 80 81 83 85



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

- Weka
  - Standardize (avg=0, stdev=1)
  - Center (avg=0)
  - Normalize (<0,1>, <A,B> A-B=scale, B=translation)

Odvozené atributy

- výpočet nového atributu ze stávajících
- BMI = váha(kg) / výška(m)^2
- rodné číslo => věk a pohlaví
- agregační hodnoty
- Weka
  - AddExpression
  - MathExpression
  - NumericTransform

Redukce počtu atributů

- Analýza hlavních komponent (PCA)
  - nové atributy nelze interpretovat
  - využití pro vizualizaci dat – použijeme n nejlepších komponent
  - PrincipalComponents
- 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 – množiny atributů
  - metoda obálky – použití metod strojového učení
  - sekce Select Attributes

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
  - RemoveWithValues
- náhrada hodnotou „nevím“
  - AddValues (nominal),
- náhrada
  - nejčetnější hodnotou
  - průměrem, mediánem ReplaceMissingValues
  - nalezení nejbližšího souseda
  - využití algoritmu pro modelování

Outliers

- Výrazně odlišné hodnoty atributu pro danou instanci
  - Outlier pro jeden atribut nemusí být outlier i pro kombinaci atributů a naopak!
  - NumericCleaner
  - InterquartileRange

Vzorkování dat

- obrovský počet instancí - pro algoritmy pracující v dávkovém režimu nutnost
  - redukce počtu dat Resample, ReservoirSample
  - tvorba modelů na základě podmnožin a jejich následná kombinace volí se přímo u volby modelu (Classify) například
- rozdělení dat na trénovací a testovací část
  - volí se přímo u volby modelu (Classify)
- 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