Úvod do Miningu

Martin Macák

Fakulta informatiky, Masarykova univerzita, Brno

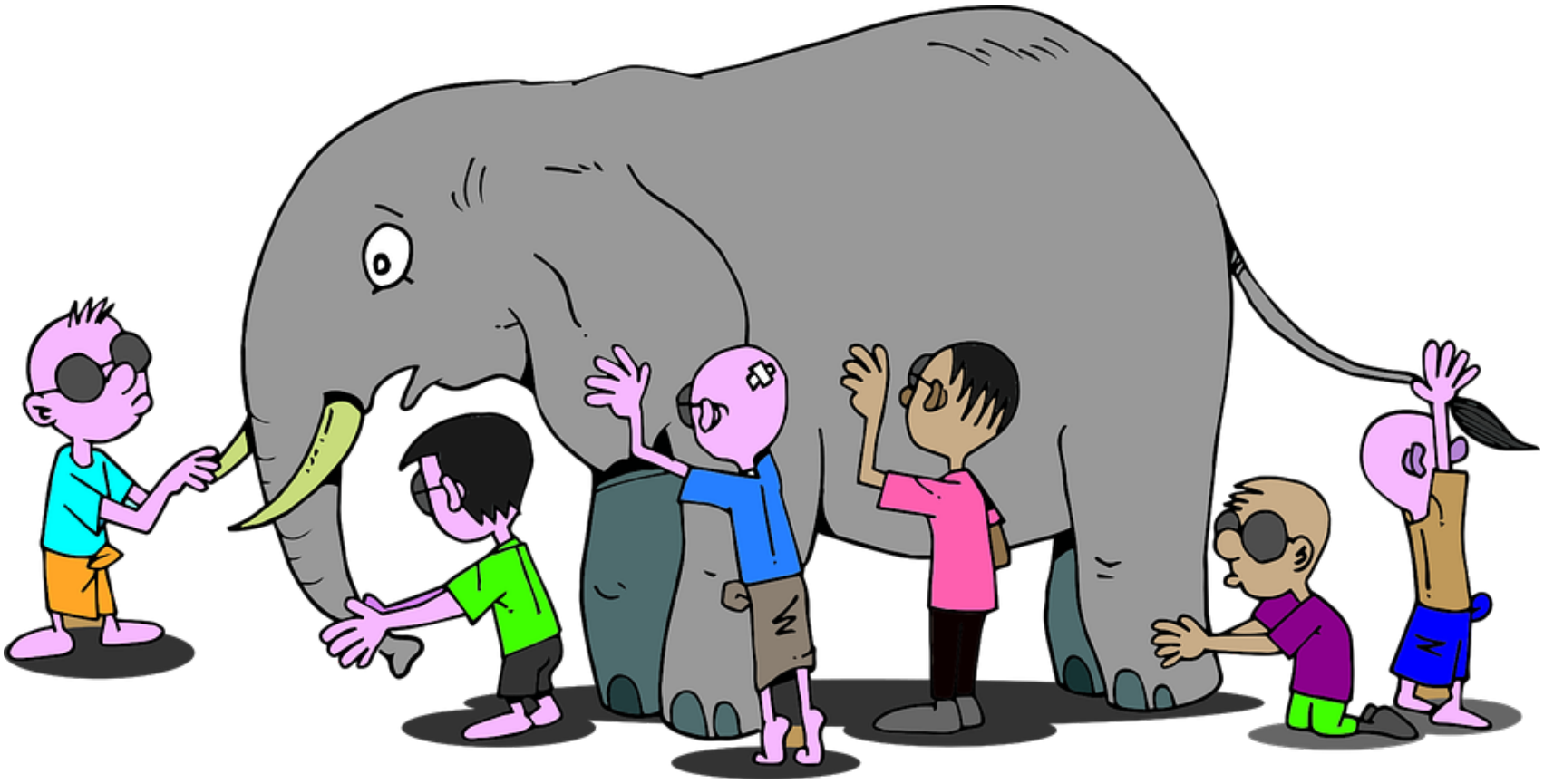
3. 10. 2019

Mining

Osnova prednášky

1. Data Mining
2. Process Mining

Blind Men and an Elephant



- nie je dobré pozerať sa na dáta iba z jednej perspektívy

Mining

1. **Data Mining**
2. Process Mining

Príklad data miningu

Názov	Žáner	Dátum vydania	Počet likov na FB	AVG vek hráčov
World of Warcraft	MMORPG	23.11.2004	6 417 874	22
Witcher 3	RPG	19.5.2015	1 421 134	20
Pokémon Go	AR Mobile	6.7.2016	2 001 838	35
PUBG	Battle royale	20.12.2017	1 883 340	11
League of Legends	MOBA	27.10.2009	14 981 158	16

- čo sú dôležité vlastnosti pre to, aby bola hra úspešná?
- ako závisí priemerný vek hráčov od dátumu vydania hry?
- vieme určiť žáner hry na základe ostatných atribútov?
- prečo má LoL toľko likov na FB?
- môžeme odhadnúť priemerný vek hráčov pre novú hru?
- ...

Data mining

1. pracujeme so štruktúrovanými dátami
2. tabuľka záznamov
3. každý záznam má nejaké premenné
   - kategorické
     - ordinálne (ľahký, priemerný, ťažký)
     - nominálne (pop, rock, metal)
   - numerické (1, 2, 3, . . .) alebo (1.32442, 2.4514, . . .)

Data mining techniky

1. supervised learning
   - snažíme sa popísať závislú premennú (*response variable*) na základe nezávislých premenných (*predictor variables*)
   - označíme, ktorú premennú chceme ako *response variable*
2. unsupervised learning
   - nemáme *response variable*

Supervised learning

- regresné techniky
  - numerická *response variable*
  - hľadáme funkciu, ktorá dokáže popísať *response variable* na základe *predictor variables*
- klasifikačné techniky
  - kategorická *response variable*
  - chceme klasifikovať záznamy na základe *predictor variables*

Regresné techniky

Supervised Learning – Regresné techniky

- Demo: lineárna regresia -> $y = a_0 + a_1 x$
- polynomiálna regresia
- ...

Supervised learning – Lineárna regresia

Demo time :)

Klasifikačné techniky

Supervised Learning – Klasifikačné techniky

- Demo: Decision trees
- K-Nearest Neighbour
- ...

Supervised Learning – Decision trees

- typicky máme viacero *predictor variables*
- rozdelíme záznamy do stromu podľa nejakej *predictor variable* tak, aby sme znížili entrópiu
- entrópia – miera neurčitosti

$$E = -\sum_{i=1}^{k} p_i \ log_2(p_i)$$

- k – počet rozličných hodnôt
- p – pravdepodobnosť, že prvok má hodnotu *i*

Supervised Learning – Decision trees príklad

- mám 6 ľudí, nejaké info o nich (či má auto, či má psa, . . .)
- 3 sú vo vzťahu, 3 nie sú vo vzťahu
- chceme klasifikovať, či je človek vo vzťahu na základe informácií, čo máme
- entrópia koreňového uzlu nášho stromu:

$$E = -\sum_{i=1}^{k} p_i \ log_2(p_i) = \ -(\frac{3}{6} \ log_2(\frac{3}{6}) + \frac{3}{6} \ log_2(\frac{3}{6}))$$

Supervised Learning – Decision trees

- pri stromoch s viac uzlami ako 1 určujeme entrópiu stromu ako vážený priemer entrópií listov
- *information gain*
  - rozdiel medzi entrópiou stromu pred a po rozdelení
  - chceme ho mať čo najväčší
  - pokiaľ by bol záporný, neoplatí sa rozdelovať strom

Supervised Learning – Decision trees algoritmus

1. začneme s koreňom, ktorý zodpovedá všetkým záznamom
2. pre každý uzol skúsime, či jeho rozdelenie podľa všetkých jeho *predictor variables* poskytuje v nejakých prípadoch *information gain*
3. vyberieme možnosť s najvyšším *information gain* a rozdelíme uzol
4. ak sme nejaký uzol rozdelili, GOTO 2
5. máme náš strom

Supervised Learning – Decision trees konfigurácie

- minimálna veľkosť uzlu
- minimálny *information gain*
- výška stromu
- ...

Supervised learning – Decision trees

Demo time :)

Unsupervised learning

- hľadáme vzťahy, patterny, ...
  - Demo: association rule learning
  - Sequential pattern mining
  - ...
- zoskupujeme
  - Demo: k-means clustering
  - k-medoids clustering
  - ...

Association rule learning

Unsupervised learning – Association rule learning

- forma: X => Y

Zákazník	Chipsy	Cola	Mrkva	Brokolica
1	X	X	-	-
2	X	X	-	X
3	X	X	-	-
4	-	-	X	X
5	X	X	-	-
6	X	-	X	X
7	X	X	-	-
8	-	-	X	X

- Chipsy => Cola
- Cola => Chipsy
- Cola, Chipsy => Brokolica
- Mrkva => Brokolica
- ...

Unsupervised learning – Association rule learning

- vlastnosti opisujúce kvalitu pravidla
  - support

$$support(X => Y) = \frac{N_{X \cup Y}}{N}$$

  - confidence

$$confidence(X => Y) = \frac{N_{X \cup Y}}{N_X}$$

  - lift (určuje koreláciu)

$$lift(X => Y) = \frac{N_{X \cup Y}\ N}{N_X\ N_Y}$$

- využívajú sa na filtrovanie alebo usporiadanie pravidiel

Unsupervised learning – Association rule learning

- typicky chceme:
  - veľký *support*
  - *confidence* blízko 1
  - *lift* > 1

Unsupervised learning – Association rule learning

Demo time :)

k-means clustering

Unsupervised learning – k-means clustering

- chceme poskupinkovať dáta
- máme **k** centroidov
- iteratívne pridružujeme dáta k najbližšiemu centroidu, následne presuniem centroid do stredu pridružených dát a opakujem proces
- opakujeme, kým sa mení skupinkovanie
- výsledný zoznam pridružených dát k danému centroidu tvorí cluster

Unsupervised learning – k-means clustering

Demo time :)

Mining

1. Data Mining
2. **Process Mining**

Process mining

- pracuje s dátami
- je process-centric (typicky pracujeme s logami udalostí)
- každá udalosť musí obsahovať *caseId*, *activity* a *timestamp*
- je možné ho kombinovať s data mining technikami

Process mining – príklad logu

Suárez	Faul	10:30
Messi	Gól	10:57
Suárez	Žltá karta	15:21
Coutinho	Faul	21:49
Messi	Gól	22:41

- *caseId* | *activity* | *timestamp*

Suárez	Camp Nou	Faul	10:30
Messi	Camp Nou	Gól	10:57
Suárez	Santiago Bernabéu	Žltá karta	10:21
Messi	Mestalla	Faul	21:49
Coutinho	Mestalla	Gól	4:41
Suárez	Santiago Bernabéu	Faul	21:24
Messi	Santiago Bernabéu	Faul	23:11
Dembélé	Camp Nou	Gól	12:39

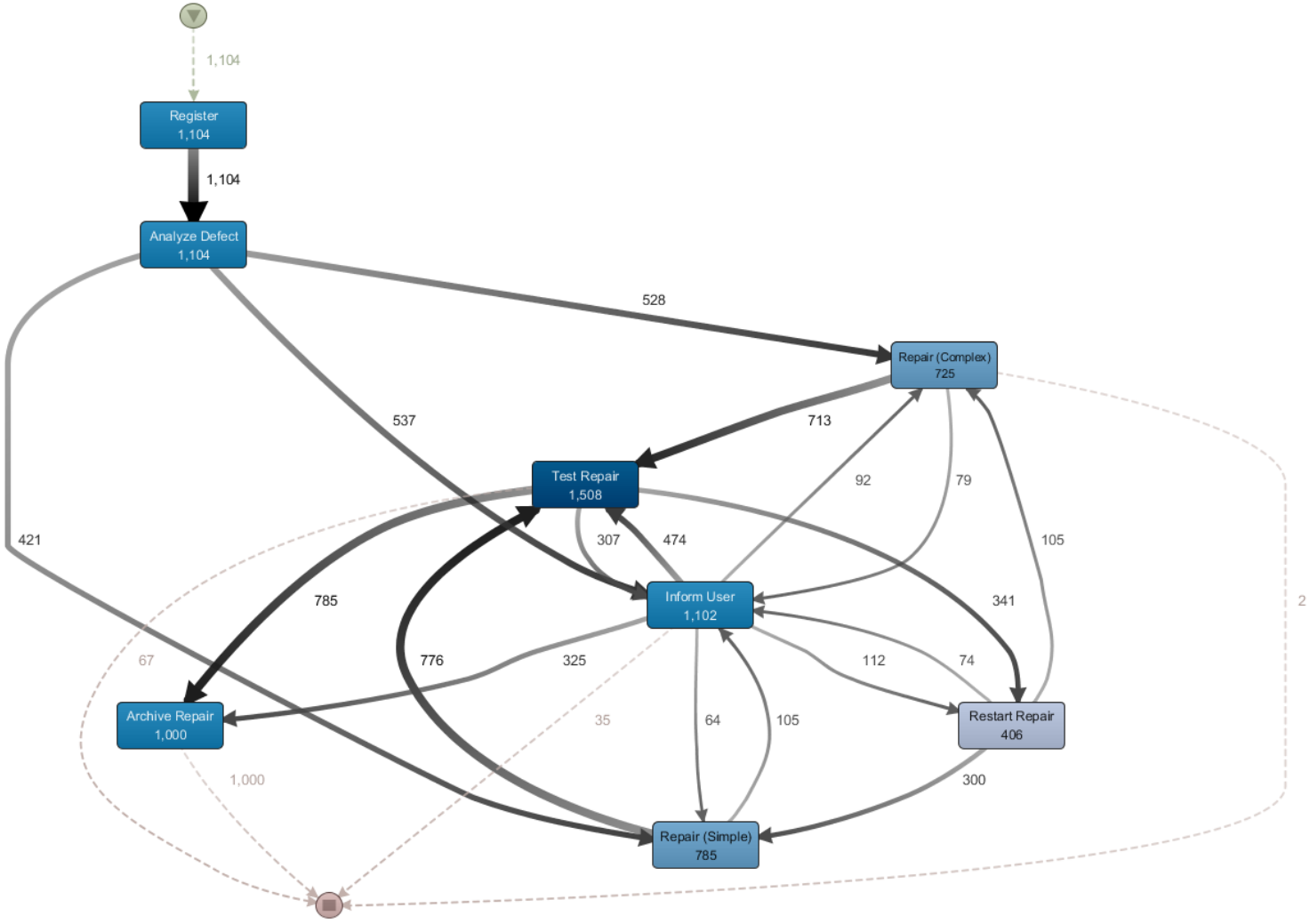
- závisí od kontextu, čo je *caseId*

Process mining techniky

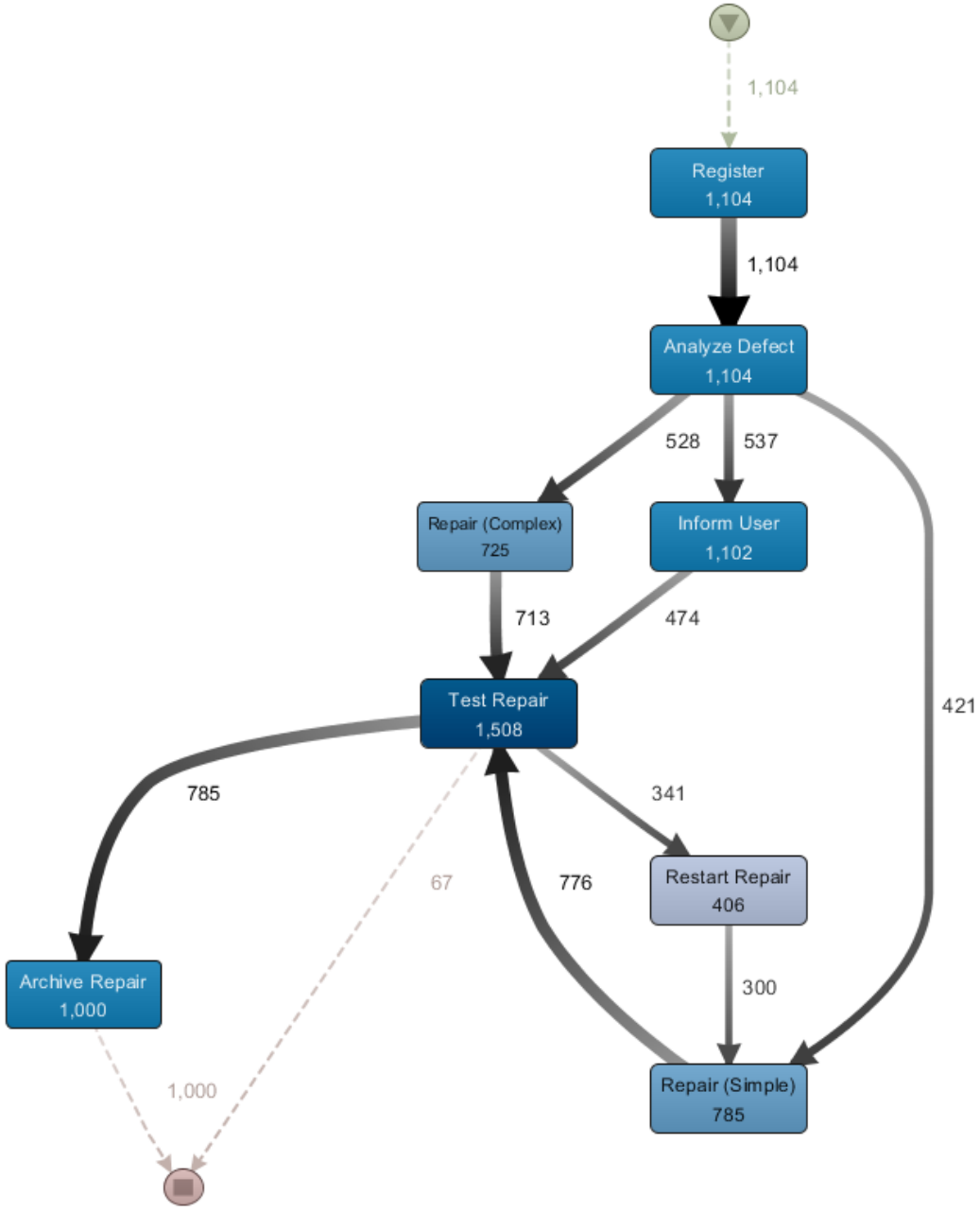
- Process Discovery
- Conformance Checking
- Enhancement

Process discovery

- z logov vytvoríme model, ktorý zodpovedá realite



Process discovery



Conformance checking

- prehráme konkrétny *case* na našom modely

Enhancement

- pridávame data perspektívu
- využijeme napr. data mining techniky

Rekapitulácia

1. Data Mining
   - data-centric
   - supervised – regresné a klasifikačné techniky
   - unsupervised – association rule learning, clustering
2. Process Mining
   - process-centric
   - process discovery
   - conformance checking
   - enhancement

macak@mail.muni.cz