Datamining v praxi

Vratislav Beneš – OptiSolutions s.r.o.

Agenda

1. **Naši zákazníci a jejich problémy**
   - Kdo jsou naši zákazníci
   - Jaká mají data a proč analyzují data
   - Co jim DM přinese
2. **Co předchází analýze**
   - Pochopení problematiky
   - Příprava dat
   - Výběr vhodných
3. **Co s výsledky analýzy aneb jak je předat zákazníkovi**
   - Pochopení výsledků
   - Konverze matematických formulí do řeči manažerů
4. **Nestrukturovaná data**
   - Text mining
   - Propojení se strukturovanými daty
   - Ukázka na hokejových datech
5. **Závěrečné shrnutí**

Agenda

1. **Naši zákazníci a jejich problémy**
   - Kdo jsou naši zákazníci
   - Jaká mají data a proč analyzují data
   - Co jim DM přinese
2. **Co předchází analýze**
   - Pochopení problematiky
   - Příprava dat
   - Výběr vhodných
3. **Co s výsledky analýzy aneb jak je předat zákazníkovi**
   - Pochopení výsledků
   - Konverze matematických formulí do řeči manažerů
4. **Nestrukturovaná data**
   - Text mining
   - Propojení se strukturovanými daty
   - Ukázka na hokejových datech
5. **Závěrečné shrnutí**

Kdo jsou naši zákazníci

- Výrobní firmy
- Prodejní firmy
- Jejich cílem je:
  - Získávat a udržet si perspektivní zákazníky
  - Znát příčiny, které ovlivňují jejich prodeje
  - Identifikovat potenciálně nespolehlivé obchodní partnery
  - Zajistit bezproblémový chod výrobních linek a předcházet jejim výpadkům

Jaká mají data a proč analyzují data

CRM data

- Interní systémy
- Analýza chování zákazníků
- Indetifikace potenciálních zákazníků

Prodejní data

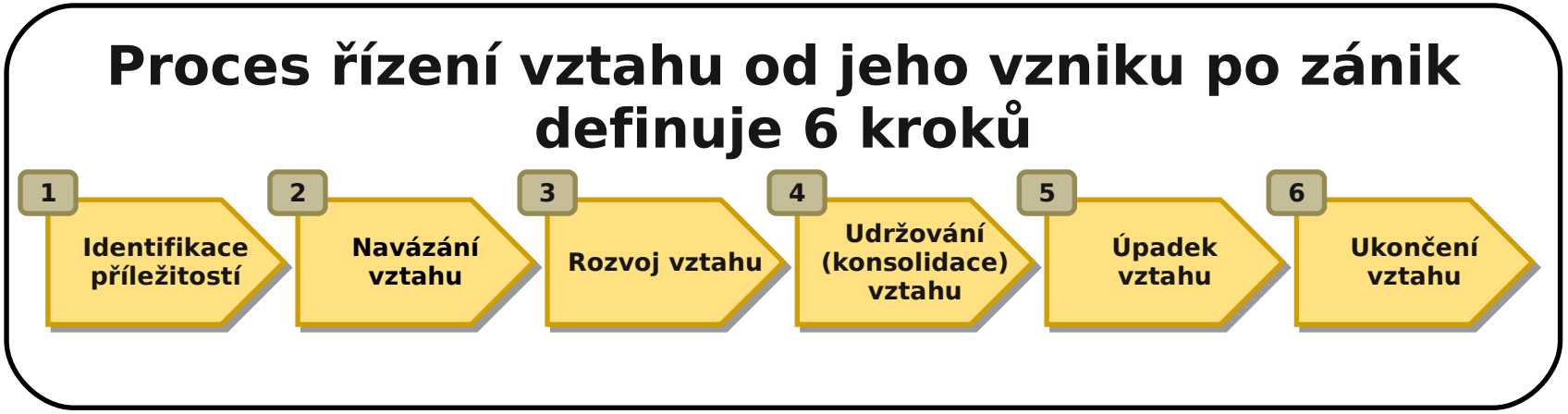
- Agentury na výzkum trhu a interní data
- Informace o dění na trhu vlastních i konkurenčních značek
- Analýza chování celého trhu

ERP systémy (SAP, MS Dynamics apd.)

- Výrobní data
- Logistická data

Co je CRM

<u>CRM (Customer Relationship Management)</u> - představuje aktivní tvorbu, sledování a udržování vztahů se zákazníky. Zahrnuje různé kroky od identifikace příležitostí, přes navázání vztahu, jeho rozvoj až po jeho případné ukončení. Pod CRM rozumíme souhrn nástrojů jako databáze, call centra, automatizace prodeje apod.



Životní cyklus obchodního procesu

<u>Lead</u> – potencionální příležitost, o které víme pouze to, že existuje

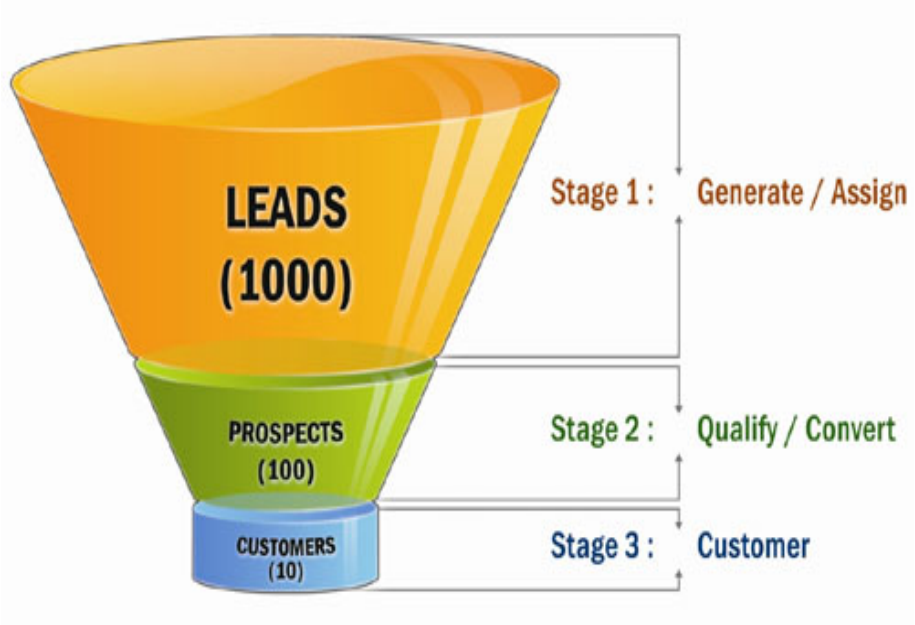
Př: Bude se stavět 10 domů na okraji Horní Dolní

<u>Opportunity</u> – příležitost, kterou již dokážeme ohodnotit a přiřadit ji míru pravděpodobnosti, s kterou nastane

Př: Domy bude stavět 1. stavební a.s. a po předběžné (80%) dohodě s pane Jeřábkem dodáme 10 Universal.

<u>Zákazník – výsledek našeho snažení</u>

Firma 1. stavební se stala našim Zákazníkem a obebrala 8 Universal a 2 Star



Proč dataminingové analýzy nad CRM

- Zjistit zákaznický model
- Na něj aplikovat tzv. ***Business Rules***
  - Stanovení bonity nových zákazníků
  - Vyhodnotit nové příležitosti ihned po jejich vložení do CRM
- Pomoc při nových příležitostech
  - *Jaké pobídky působí v daném regionu?*
  - *Jaké skupiny zákazníků oslovit s výrobkem typu X?*
- Fraud management
  - Identifikace potenciálně nespolehlivých zákazníků
  - Identifikace potenciálně nespolehlivých obch. partnerů
- Získat pohled na funkčnost marketingových kampaní

Prodejní data

Prodejní data

- Informace o promo aktivitách: letáky, sekundární umístění, cenové akce
- Prodeje jednotlivých značek a jejich balení
- Data podle:
  - Formátů obchodů (supermarket, hypermarket apd)
  - Řetězeců (Tesco, Ahold apd)

Ostatní data

- Ostatní promoakce (promo týmy, ochutnávky ...)
- Demografická data (skladba obyvatel, velikost města)
- Geografická data (blízkost konkurenta, umístěná prodejny)
- Nezaměstnanost regionu (ČSÚ)
- Průměrná teplota a vlhkost za týden (ČHÚ)
- Kupní síla regionu (Incoma/GfK, vypočtená podle ekonomických ukazatelů)
- apd

Proč dataminingové analýzy nad prodejními daty

Otázka:

- Existuje spojitost mezi růstem prodejů výrobku X, letákovou podporou a druhotným umístěním na prodejnách různých typů?

Odpověď: (informace jsou smyšlené)

- Jestliže je v aktuálním týdnu pouze leták, pak prodeje vzrostou o více jak 123%.
- Jestliže je i sekundární umístění, pak prodeje vzrostou o více jak 456%.
- Na obchodech typu Supermarket ve městech do 15.000 obyvatel zároveň platí: je–li sekundární umístění podpořeno in-store aktivitou, pak prodeje vzrostou o další 789%.

Efektivní využití těchto znalostí

- Možnost efektivního rozhodování se o promo akcích
- Lepší porozumění promo akcí u jednotlivých značek, regionů a typu obchodů
- Odhad předpokládaných prodejů
- Simulace vlivu promo aktivit
- Simulace dopadu akce konkurenta
- Simulace dopadu změny ceny
- Testování rozdílných scénářů
- One to one marketing

Agenda

1. **Naši zákazníci a jejich problémy**
   - Kdo jsou naši zákazníci
   - Jaká mají data a proč analyzují data
   - Co jim DM přinese
2. **Co předchází analýze**
   - Pochopení problematiky
   - Příprava dat
   - Výběr vhodných metod
3. **Co s výsledky analýzy aneb jak je předat zákazníkovi**
   - Pochopení výsledků
   - Konverze matematických formulí do řeči manažerů
4. **Nestrukturovaná data**
   - Text mining
   - Propojení se strukturovanými daty
   - Ukázka na hokejových datech
5. **Závěrečné shrnutí**

Co předchází analýze

- Pochopení problematiky
- Perfektně porozumět zadání
- Co nejvíce se seznámit s analyzovaným světem
  - ***Hledat nové, zajímavé znalosti, ne obecné a již notoricky známé***
- Znát motivaci zákazníka pro danou analýzu
- CRM
  - Co je *příležitost* a *stav příležitosti*?
  - Je vztah mezi produktem a službou?
  - Co je důležité pro obchodníka a co pro management firmy
- Prodejní data
  - Význam jednotlivých ukazatelů (*Sales Value, Sales Volume, Sales Units...*)
  - Soustruhy se neprodávají v supermarketu
  - Zákazník hypermarketu reaguje jinak než zákazník e-shopu

Příprava dat

- Data jsou všude
- Vhodná data nenajdete na ulici
- Některá data musíte dohledat
- Každý zdroj znamená jiný formát
- Musíte je dát do jedné matice
- Vhodná znalost DB systémů a základy SQL

Příklad: Analýza hokejových zápasů s cílem vybrat vhodné atributy pro model predikující výsledky budoicích zápasů. Zdroj: *Internet*

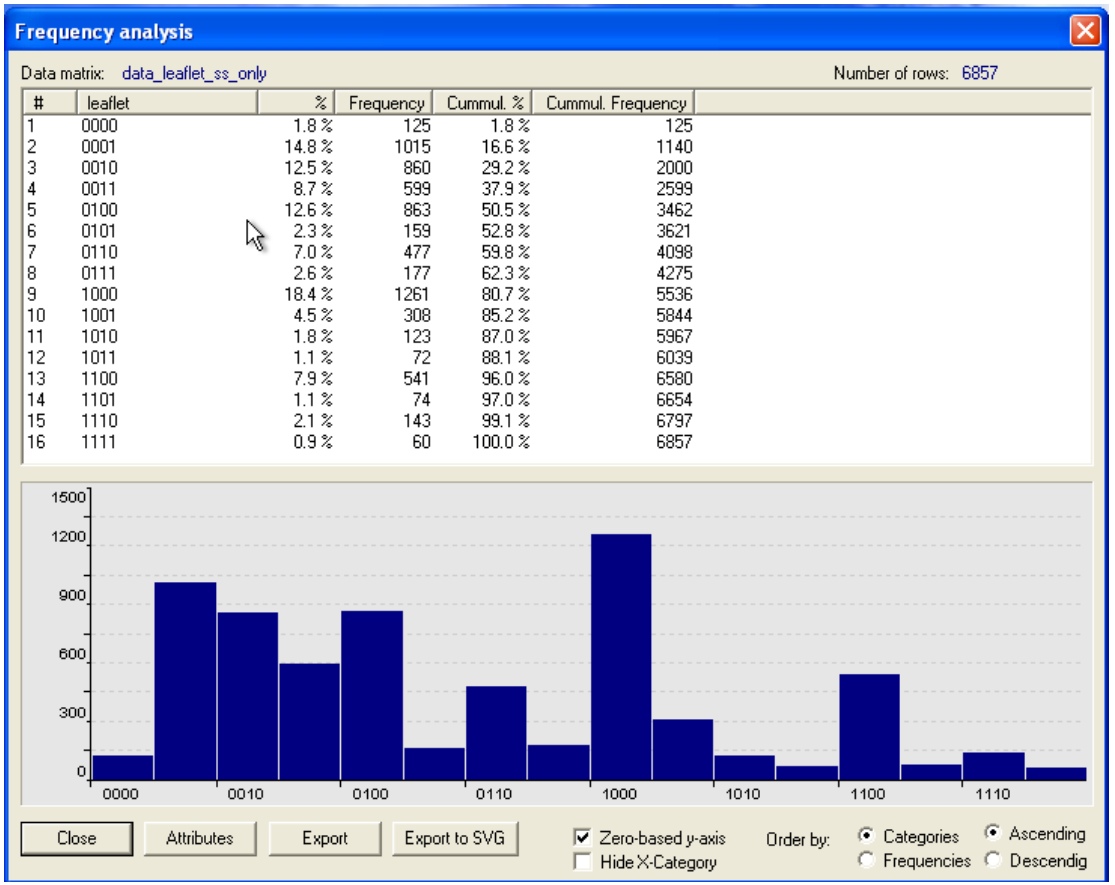
kolo	domaci	hoste	vysledek	1.tretina	2.tretina	3.tretina	pp	strely na branu	vylouceni
2	HC ČSOB Pojišťovna Pardubice	Bílí Tygři Liberec	5:2	2:0	2:2	1:0		40:23	5:8
3	HC Energie Karlovy Vary	HC ČSOB Pojišťovna Pardubice	3:7	0:3	1:3	2:1		29:34	7:5
4	HC ČSOB Pojišťovna Pardubice	HC Sparta Praha	2:6	1:4	1:2	0:0		32:35	4:10
5	PSG Zlín	HC ČSOB Pojišťovna Pardubice	2:1	0:0	0:1	2:0		28:30	2:3
6	HC ČSOB Pojišťovna Pardubice	HC MOUNTFIELD	2:1	0:1	0:0	1:0	1:0	27:29	4:4
7	HC Kometa Brno	HC ČSOB Pojišťovna Pardubice	1:4	0:1	0:1	1:2		33:26	9:9
8	HC ČSOB Pojišťovna Pardubice	BK Mladá Boleslav	2:3	2:2	0:0	0:1		33:23	4:5
9	HC Slavia Praha	HC ČSOB Pojišťovna Pardubice	1:2	0:1	1:1	0:0		26:26	6:3

Příprava atributů

- Příprava metrik
- Vhodný model kódování

Příklad: Převodem výskytu letáku ve 4 po sobě jdoucích týdnech do jednoho ukazatele.

Atribut letáků je kódovaný 4 bitovým číslem podle výskytu v čase. První bit znamená výskyt v aktuálním týdnu, druhý v předchzím, třetí v předminulé a čtvrtý v předpřed minulém týdnu.

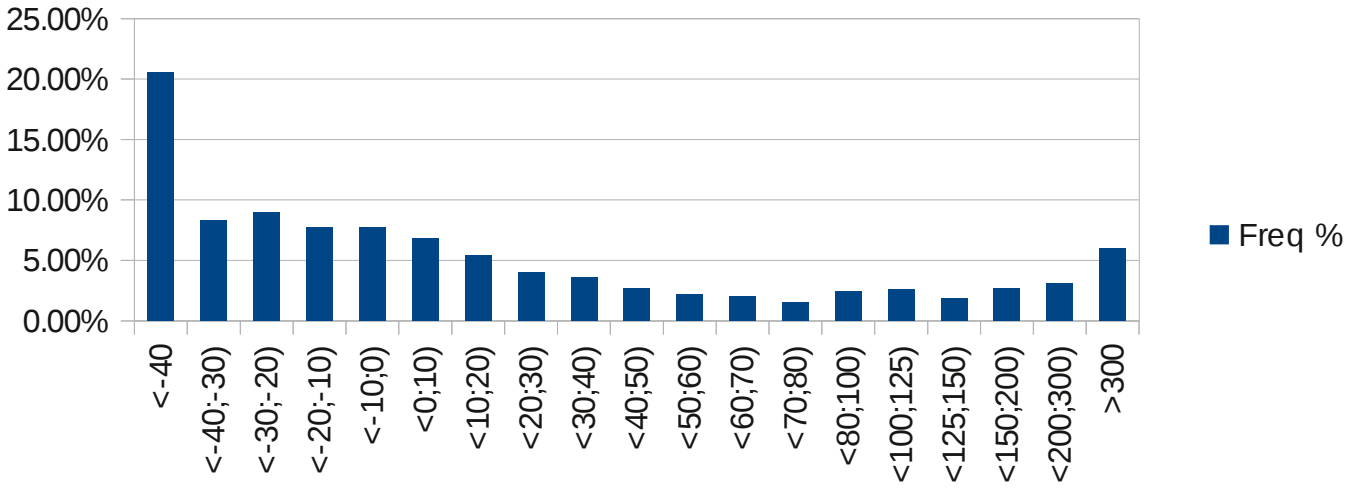


#	leaflet	%	Frequency	Cummul. %	Cummul. Frequency
1	0000	1.8 %	125	1.8 %	125
2	0001	14.8 %	1015	16.6 %	1140
3	0010	12.5 %	860	29.2 %	2000
4	0011	8.7 %	599	37.9 %	2599
5	0100	12.6 %	863	50.5 %	3462
6	0101	2.3 %	159	52.8 %	3621
7	0110	7.0 %	477	59.8 %	4098
8	0111	2.6 %	177	62.3 %	4275
9	1000	18.4 %	1261	80.7 %	5536
10	1001	4.5 %	308	85.2 %	5844
11	1010	1.8 %	123	87.0 %	5967
12	1011	1.1 %	72	88.1 %	6039
13	1100	7.9 %	541	96.0 %	6580
14	1101	1.1 %	74	97.0 %	6654
15	1110	2.1 %	143	99.1 %	6797
16	1111	0.9 %	60	100.0 %	6857

týden	leták	leták transf
1	0	0000
2	1	1000
3	0	0100
4	1	1010
5	0	1101
6	1	1010

Příprava atributů – LISp miner - intervaly

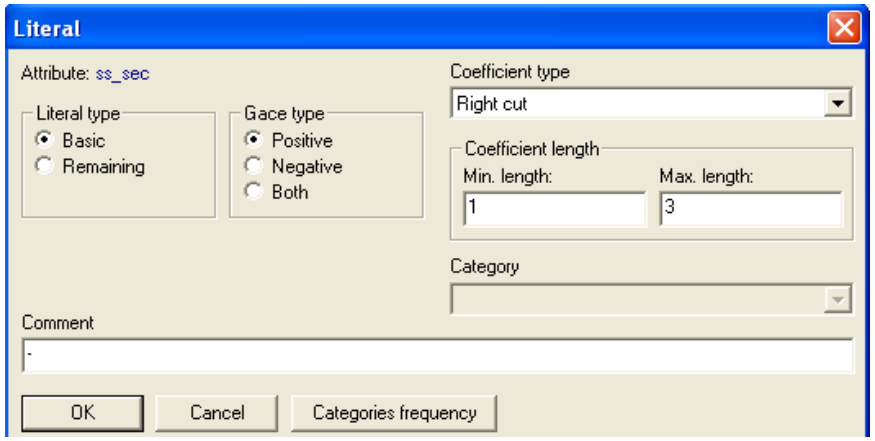
- Vývoj prodejů v %
- Je rozdíl mezi 31.2% a 31.4%?
  - ~~0.2 pb~~
  - Zanedbatelný
- Vytvořit intervaly o délce 10

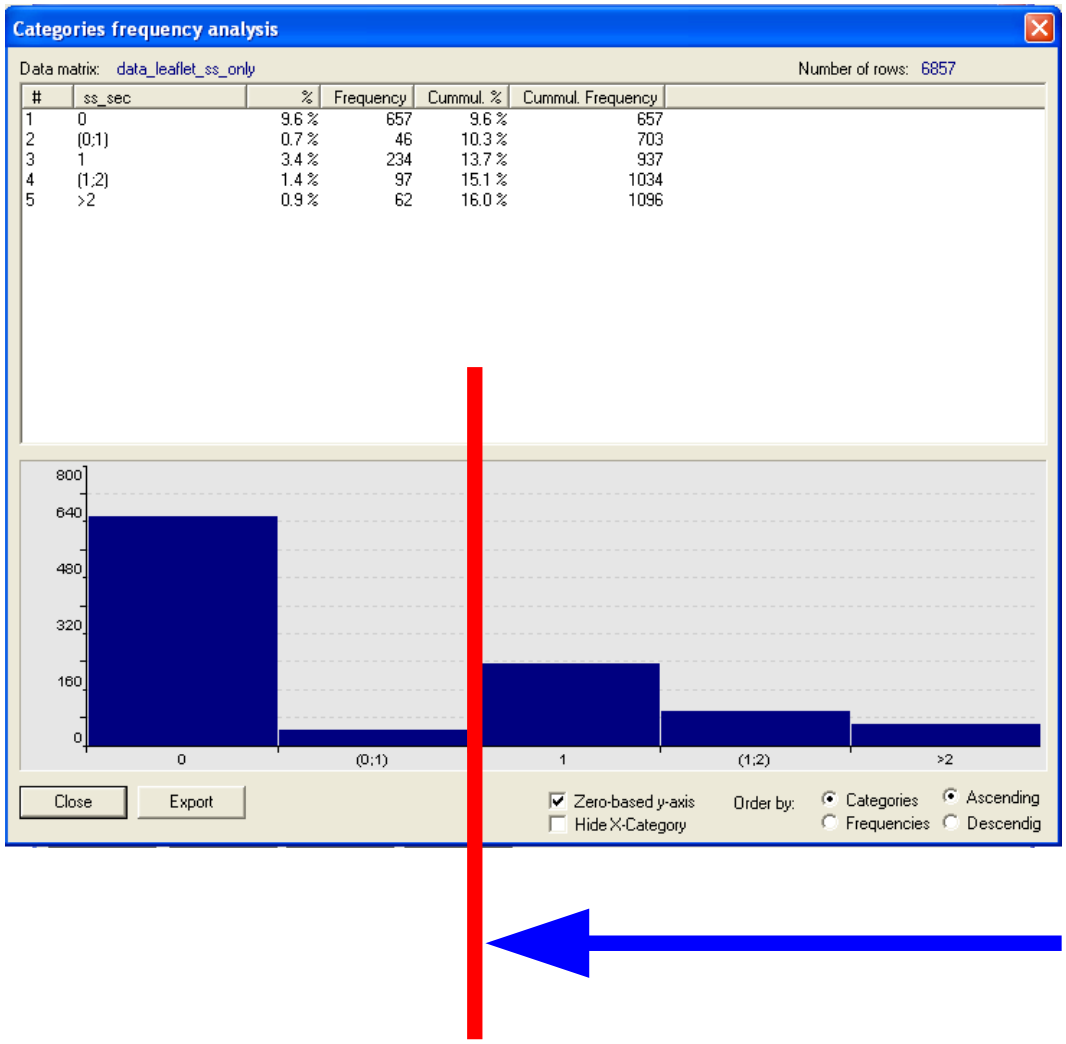


Příprava dat – LISp miner - řezy

- Pravé, levé řezy
- Délka řezu

Příklad: Zajímají nás pouze ty případy, kdy byl výskyt minimálně jedné celé palety.





Agenda

1. **Naši zákazníci a jejich problémy**
   - Kdo jsou naši zákazníci
   - Jaká mají data a proč analyzují data
   - Co jim DM přinese
2. **Co předchází analýze**
   - Pochopení problematiky
   - Příprava dat
   - Výběr vhodných
3. **Co s výsledky analýzy aneb jak je předat zákazníkovi**
   - Pochopení výsledků
   - Konverze matematických formulí do řeči manažerů
4. **Nestrukturovaná data**
   - Text mining
   - Propojení se strukturovanými daty
   - Ukázka na hokejových datech
5. **Závěrečné shrnutí**

Co s výsledky analýzy aneb jak je předat zákazníkovi

- **Výsledky většinou poslouchají manažeři**
- V drtivé většině případů to nejsou matematici
- **Mluvit jejich řečí**
- Neznají technické pojmy z DM
- Pečlivě volit, které technické argumenty použijete
- Všechny **důkladně a srozumitelně vysvětlit**

Příklad:

$$\cancel{typ\,obchodu(Supermarket) \wedge velikost\,mesta(0{,}15000) \wedge secss(0;\infty) \wedge instore(TRUE) \rightarrow sales(789,\infty)}$$

Na obchodech typu Supermarket ve městech do 15.000 obyvatel zároveň platí: je–li sekundární umístění podpořeno in-store aktivitou, pak prodeje vzrostou o další 789%.

Agenda

1. **Naši zákazníci a jejich problémy**
   - Kdo jsou naši zákazníci
   - Jaká mají data a proč analyzují data
   - Co jim DM přinese
2. **Co předchází analýze**
   - Pochopení problematiky
   - Příprava dat
   - Výběr vhodných
3. **Co s výsledky analýzy aneb jak je předat zákazníkovi**
   - Pochopení výsledků
   - Konverze matematických formulí do řeči manažerů
4. **Nestrukturovaná data**
   - Text mining
   - Propojení se strukturovanými daty
   - Ukázka na hokejových datech
5. **Závěrečné shrnutí**

Nestrukturovaná data

- Velké množství informací není strukturováno
- e-maily, dokumenty, články na webu
- IBM Watson
  - Jeopardy!
  - zdravotní pojišťovna WellPoint
- Obrovská výzva

Příklad:

Vážený pane Vomáčka,

nabízíme vám vámi poptávaný výrobek A za 20.000Kč.

S pozdravem,
Karel Jedlička

Zákazník má bonitu 1 a obchod se realizuje.

Jelínku,

nabízíme vám vámi poptávaný výrobek A za 19.000Kč.

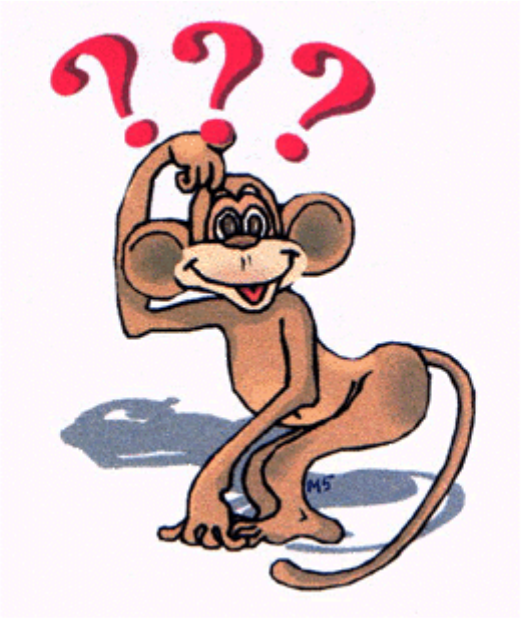
Pepa Vohnátko

Zákazník má bonitu 1 a obchod se **nerealizuje**.

Nestrukturovaná data

kolo	domaci	hoste	vysledek	1.tretina	2.tretina	3.tretina	pp	strely na branu	vylouceni
2	HC ČSOB Pojišťovna Pardubice	Bílí Tygři Liberec	5:2	2:0	2:2	1:0		40:23	5:8
3	HC Energie Karlovy Vary	HC ČSOB Pojišťovna Pardubice	3:7	0:3	1:3	2:1		29:34	7:5
4	HC ČSOB Pojišťovna Pardubice	HC Sparta Praha	2:6	1:4	1:2	0:0		32:35	4:10
5	PSG Zlín	HC ČSOB Pojišťovna Pardubice	2:1	0:0	0:1	2:0		28:30	2:3
6	HC ČSOB Pojišťovna Pardubice	HC MOUNTFIELD	2:1	0:1	0:0	1:0	1:0	27:29	4:4
7	HC Kometa Brno	HC ČSOB Pojišťovna Pardubice	1:4	0:1	0:1	1:2		33:26	9:9
8	HC ČSOB Pojišťovna Pardubice	BK Mladá Boleslav	2:3	2:2	0:0	0:1		33:23	4:5
9	HC Slavia Praha	HC ČSOB Pojišťovna Pardubice	1:2	0:1	1:1	0:0		26:26	6:3

K	L	M	N
kolo	komentare obecne	komentar Pardubice	komentar protivnika
2.	**Hokejisté Pardubic vstoupili do** nové extraligové **sezony výhrou** nad Libercem 5:2. Mezi střelce se zapsali i dvě zámořské posily Corey Elkins a Jonathan Sim.	Chtěl bych poděkovat hráčům za předvedený výkon a jsem rád, že **jsme** v domácím prostředí **zahájili sezonu vítězstvím**. Pro vývoj utkání **bylo důležité, že jsme v první třetině ubránili skoro minutovou přesilovku** Liberce **pět na tři a** pak v **následné přesilovce jsme dokázali otevřít skóre**. Díky tomu **jsme vyhráli první třetinu** a tím se zklidnili. Myslím, že jsme dnes podali dobrý výkon.	Hodnotí se mi těžko. Bylo to první utkání a musím říci, že **domácí se** na něj **připravili dobře** a v podstatě naprosto zaslouženě zvítězili. **My jsme** dneska s naším výkonem **neměli šanci**.

- Jak identifikovat klíčová slova?
- Jak extrahovat jejich význam?
- Jak získaný význam strukturovat?
- Jak data provázat se strukturovanými?

Otázky