




**České vysoké učení technické v Praze**



**Fakulta elektrotechnická**



**Katedra kybernetiky  
Katedra počítačů**



# Vytěžování dat – přednáška I

## Úvod do vytěžování dat

Filip Železný: [zelezny@fel.cvut.cz](mailto:zelezny@fel.cvut.cz)

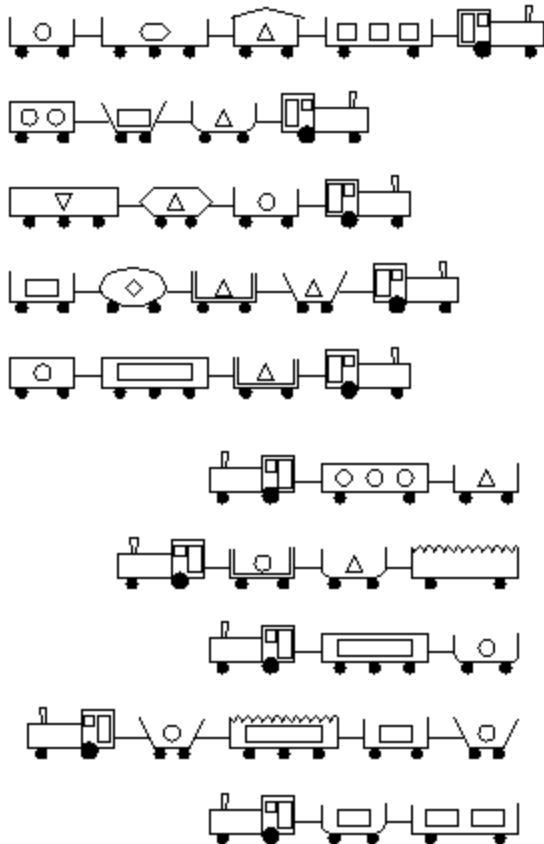
Pavel Kordík: [kordikp@fel.cvut.cz](mailto:kordikp@fel.cvut.cz)

# Vytěžování dat (data mining)

- Fayyad et al: „*Data Mining je netriviální proces identifikace pravdivých, dosud neznámých, potenciálně využitelných a naprosto srozumitelných **vzorů** v datech*“
  - Vzor = obecný princip, souvislost, tvrzení, apod. nalezený v konkrétních datech
  - Vzor reprezentuje **znalost**
  - „Dobývání znalostí z dat“ (Knowledge Discovery in Data, KDD)
- Účel: zlepšení rozhodovacích procesů
- Data mining prakticky: vývoj a využití **počítačových algoritmů** umožňujících vyhledávat vzory

# Vytěžování dat neformálně

## ■ Data

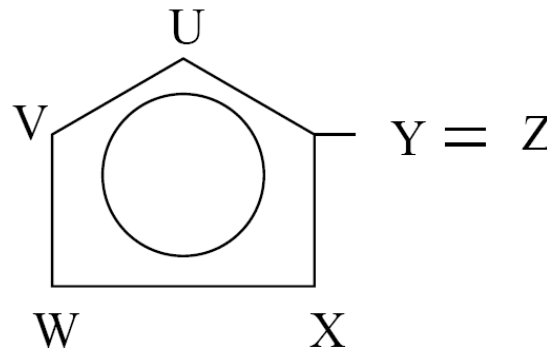


## ■ Nalezené vzory

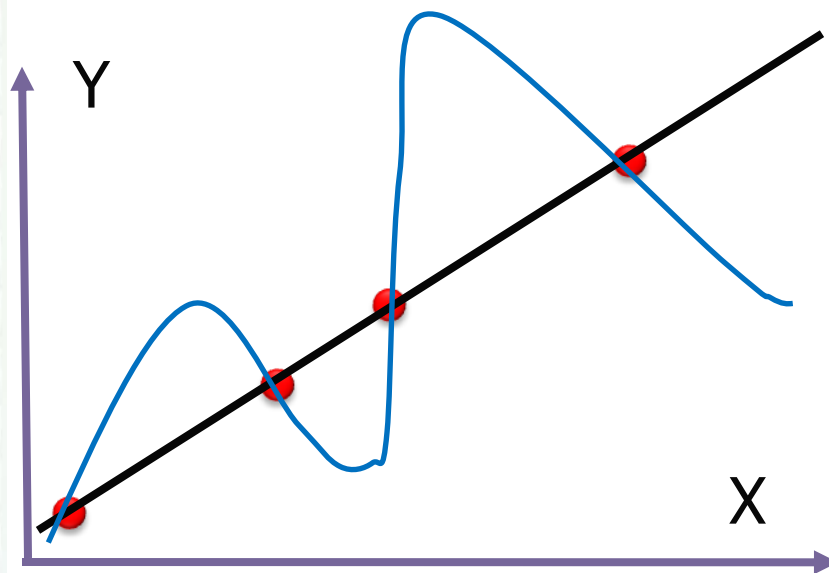
- Má-li vlak 2 vagóny, jede doleva
- Všechny náklady v jednom vagónu mají stejný tvar
- ...

# Příklady (reálnějších) vzorů

- Častá asociace v nákupních koších  
{pivo, dětské pleny}
- Implikace  
IF horečka AND bolest\_svalů THEN chřipka
- Graf



# Který vzor je lepší?



- Data  
 $(x_1, y_1), (x_2, y_2), \dots$
- Vzor 1  
Rovnice přímky
- Vzor 2  
Rovnice polynomu

- Oba vzory platné v zadaných datech.
- Který z nich je „pravdivější“?
- Pravdivost vzorů nelze zaručit.

# Vytěžování dat

- Past data miningu: když se dost snažíme, vždy nějaké vzory najdeme
- V dostupných datech mohou platit jen náhodou
- Nemusí být skutečně charakteristické pro proces, který data generuje
- Google define: Data Mining :  
*"Data mining is the equivalent to sitting a huge number of monkeys down at keyboards, and then reporting on the monkeys who happened to type actual words."*
- Je to tedy k něčemu dobré ?!



Vytěžování dat

# PŘÍKLADY APLIKACÍ

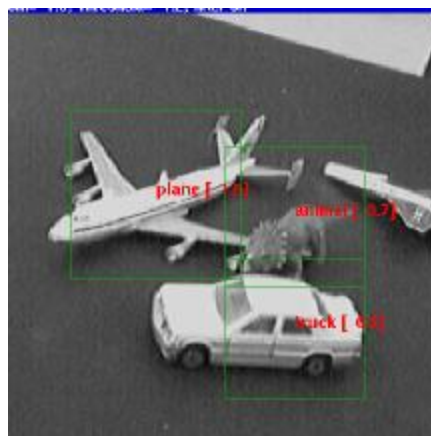
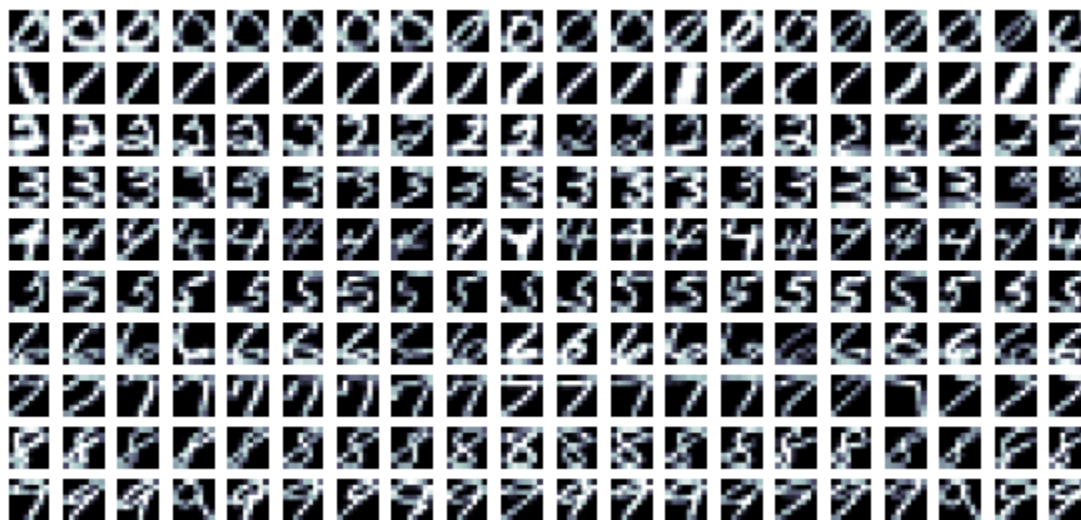


# Rozpoznávání řeči





# Rozpoznávání obrazu



# Vyhledávání relevantních informací

## „Information retrieval“



# Predikce řad a signálů



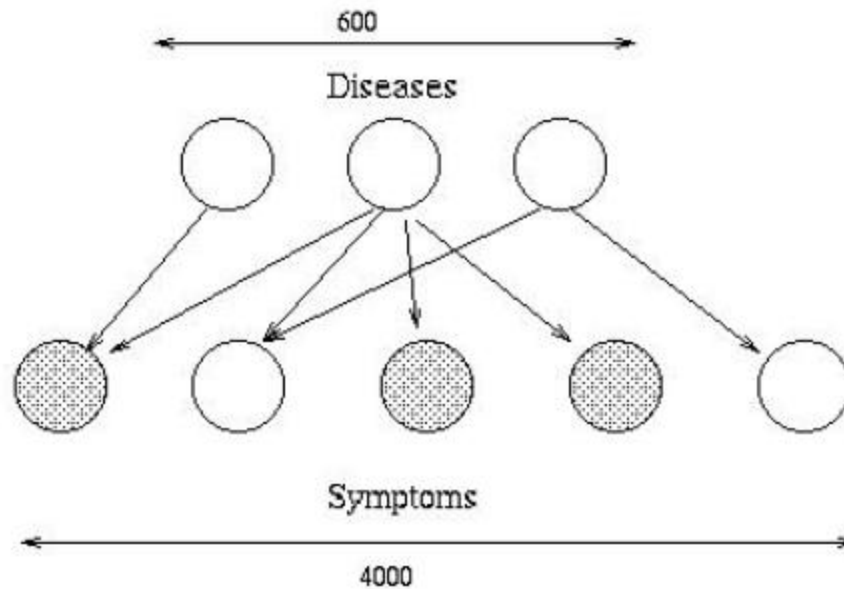
- Finanční analýzy
- Meteorologie
- Biosignály (nejen predikce)
- ...

# Analýza transakčních dat

EAN 1	Název 1	EAN 2	Název 2	EAN 3	Název 3	Výskyty
8590338901657	'MR.MATTES WC NÚHRADA CITRON 1X40G'	8590338901671	'MR.MATTES WC NÚHRADA LES 1X40G'	8590338901664	'MR.MATTES WC NÚHRADA MOŮE 1X40G'	7366
8585000703424	'VOUX MÍDLO MLÚKO A MED, 100G'	8585000703394	'VOUX MÍDLO S VITAMŇEM E, 100G'	8585000703332	'VOUX MÍDLO ZELENÍ -AJ A CITRON 100G'	3748
8585000703424	'VOUX MÍDLO MLÚKO A MED, 100G'	8585000703363	'VOUX MÍDLO S ALOE VERA, 100G'	8585000703332	'VOUX MÍDLO ZELENÍ -AJ A CITRON 100G'	3441
8585000703394	'VOUX MÍDLO S VITAMŇEM E, 100G'	8585000703363	'VOUX MÍDLO S ALOE VERA, 100G'	8585000703332	'VOUX MÍDLO ZELENÍ -AJ A CITRON 100G'	3419
8594003602498	'KOUPELOVÚ SUL RELAXA-NŇ 70G'	8594003602481	'KOUPELOVÚ SUL ZKLIDŮJŮČŮ 70G'	8594003602504	'KOUPELOVÚ SUL ZVLÚ-ŮJŮČŮ 70G'	3111
8585000703424	'VOUX MÍDLO MLÚKO A MED, 100G'	8585000703394	'VOUX MÍDLO S VITAMŇEM E, 100G'	8585000703363	'VOUX MÍDLO S ALOE VERA, 100G'	3099
8693495004886	'PALMOLIVE MIDLO MILK AND HONEY100G'	8693495004848	'PALMOLIVE MIDLO BŮLÚ-ZVLÚ-ŮJŮČŮ 100G'	8693495004862	'PALMOLIVE MIDLO ZELENÍ -ZVLÚ-ŮJŮČŮ 100G'	2715
8594001698981	'HAMÚ KOJENECKÚ VÍÚIVA S MERUÚKAMI 190G'	8594001698974	'HAMÚ KOJENECKÚ VÍÚIVA S BROSKVEMI 190G'	8594001698837	'HAMÚ KOJENECKÚ VÍÚIVA S JOHODAMI 190G'	2651
8594002670122	'TWIGGY ČVESTKOVÚ V JOGURTU 35G'	8594002670139	'TWIGGY BRUSINKOVÚ V JOGURTU 31G'	8594002670405	'TWIGGY JAHODOVÚ V JOGURTU'	2609
8595000910562	'LARRIN-PLUS FIALOVÍ 40G N.V.'	8595000910500	'LARRIN-PLUS ZELENÍ 40G N.V.'	8595000910555	'LARRIN WC PRIM MODRÍ N.V. 40G'	2554
8594001698981	'HAMÚ KOJENECKÚ VÍÚIVA S MERUÚKAMI 190G'	8594001698974	'HAMÚ KOJENECKÚ VÍÚIVA S BROSKVEMI 190G'	8595139718879	'HAMÚ KOJENECKÚ VÍÚIVA S HRUČKAMI 190G'	2525
8594001698974	'HAMÚ KOJENECKÚ VÍÚIVA S BROSKVEMI 190G'	8594001698837	'HAMÚ KOJENECKÚ VÍÚIVA S JOHODAMI 190G'	8595139718879	'HAMÚ KOJENECKÚ VÍÚIVA S HRUČKAMI 190G'	2486
8595121403653	'DROXI MINI SG S VIT.E,B5 SPORT&RELAX 35 ML'	8595121403639	'DROXI MINI SG S ES.OLEJÍ AROMA ANTI STRESS 35 ML'	8595121403646	'DROXI MINI KRÚM GEL HONEY&MILK S HYDRÚTKAMI 35ML'	2485
8594003602511	'KOUPELOVÚ SUL -ISTŮČŮ 70G'	8594003602481	'KOUPELOVÚ SUL ZKLIDŮJŮČŮ 70G'	8594003602504	'KOUPELOVÚ SUL ZVLÚ-ŮJŮČŮ 70G'	2442
8717163607268	'DOVE MÍDLO EXFOLIATING 100G'	4000388170704	'DOVE MÍDLO S OLEJEM 100G'	8000700000005	'DOVE MÍDLO 100G'	2432
5900452071854	'SUNÚREK MASOZEL. PŮŮKRM ZELENINA S VEPÚOVÍM MASEM'	5900452071465	'SUNÚREK MASOZEL. PŮŮKRM ZELENINA S JEHNÚŮM MASEM'	5900452072134	'SUNÚREK MASOZEL. PŮŮKRM ZELENINA S TELECŮM MASEM 1'	2389
8594001698981	'HAMÚ KOJENECKÚ VÍÚIVA S MERUÚKAMI 190G'	8594001698837	'HAMÚ KOJENECKÚ VÍÚIVA S JOHODAMI 190G'	8595139718879	'HAMÚ KOJENECKÚ VÍÚIVA S HRUČKAMI 190G'	2362
8594003602511	'KOUPELOVÚ SUL -ISTŮČŮ 70G'	8594003602498	'KOUPELOVÚ SUL RELAXA-NŇ 70G'	8594003602481	'KOUPELOVÚ SUL ZKLIDŮJŮČŮ 70G'	2352
8717163978238	'LUX MÍDLO S VITAÚKY Z RTÚI 90G'	8717163978320	'LUX MÍDLO S VITAÚKY Z BYLIN 90G'	8717163978399	'LUX MÍDLO S MANDLOVÍM OLEJEM 90G'	2307
8594003602511	'KOUPELOVÚ SUL -ISTŮČŮ 70G'	8594003602498	'KOUPELOVÚ SUL RELAXA-NŇ 70G'	8594003602504	'KOUPELOVÚ SUL ZVLÚ-ŮJŮČŮ 70G'	2290
8717163978283	'LUX MÍDLO S VITAÚKY Z OVOCE 90G'	8717163978320	'LUX MÍDLO S VITAÚKY Z BYLIN 90G'	8717163978399	'LUX MÍDLO S MANDLOVÍM OLEJEM 90G'	2238

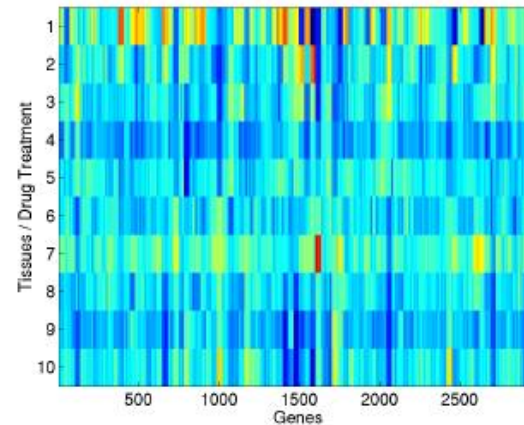
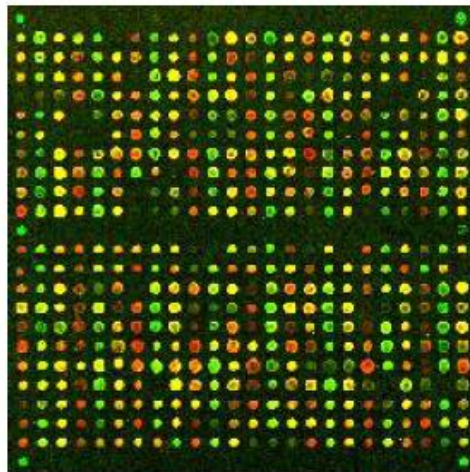
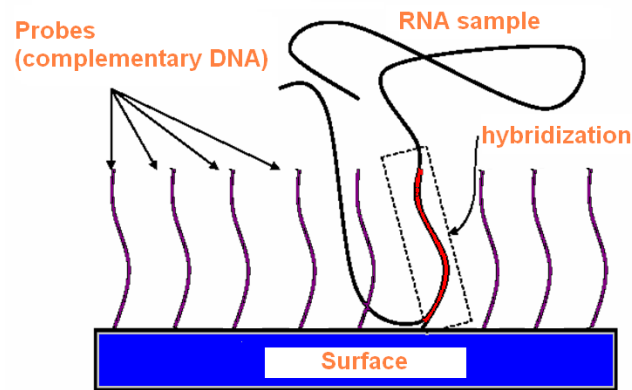
- Časté asociace mezi výrobky v nák. koších
- Segmentace zákazníků

# Lékařská diagnostika



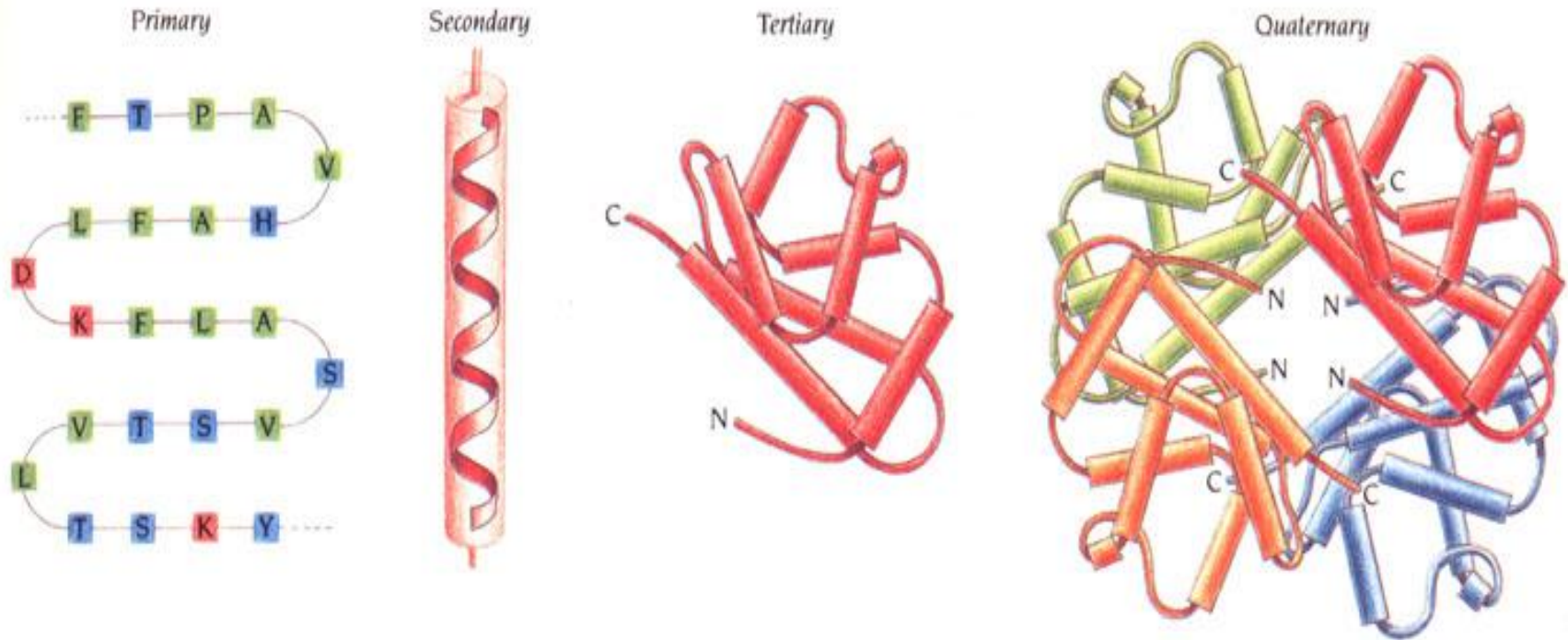


# Bioinformatika



## ■ Objevování funkcí genů

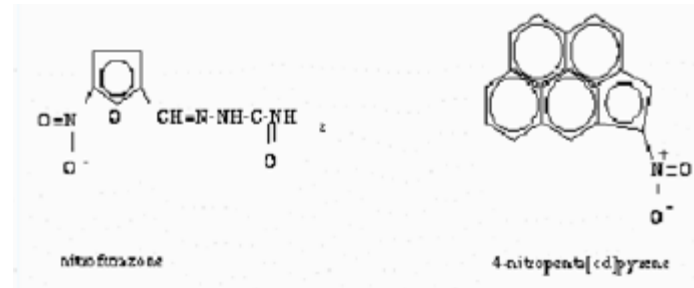
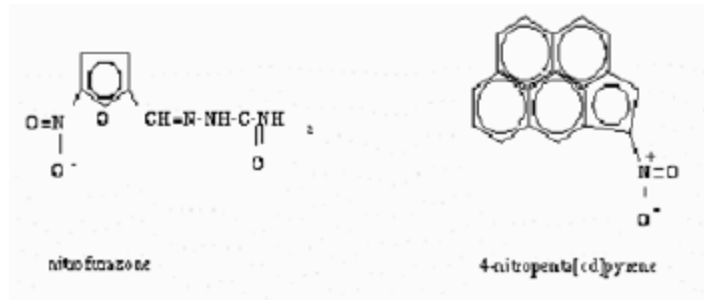
# Bioinformatika



- Predikce sekundární struktury bílkovin



# Bioinformatika



## ■ Predikce karcinogenity sloučenin

# Systemy pro doporučení

The screenshot shows the Amazon.com homepage in a Windows Internet Explorer browser. The page features a navigation bar with links to 'Your Account', 'Cart', 'Your Lists', and 'Help'. A search bar is prominently displayed. Below the navigation bar, there are several promotional banners and category links. The 'Browse' section on the left lists categories like Books, Movies & Music, Clothing & Accessories, Computer & Office, Consumer Electronics, Food & Household, Health & Beauty, Home & Garden, and Kids & Toys. The main content area includes a 'Rewards You'll Love' banner for the Amazon Visa Card, a 'Check Out the Newest Arrivals in Toys & Games' section, and a 'Books Bestsellers' section. A 'New Sony Alpha DSLR' advertisement is also visible on the right side.

The screenshot shows the Netflix Prize website. The header features the 'Netflix Prize' title and navigation links like 'Home', 'Rules', 'Leaderboard', 'Register', 'Update', 'Submit', and 'Download'. Below the header, there is a 'Welcome!' message and a section titled 'Movies For You' which lists recommended movies. A large banner on the right side of the page reads 'You really liked it...' and mentions a prize of \$50,000. The page also includes a 'Welcome!' message and a 'You really liked it...' section.

# Analýza sociálních sítí

- Elektronická komunikace
- LinkedIn, FaceBook, ...
- Citace ve výzkumu
- Organizovaný zločin
- Znalostní sítě (Wiki)
- ...



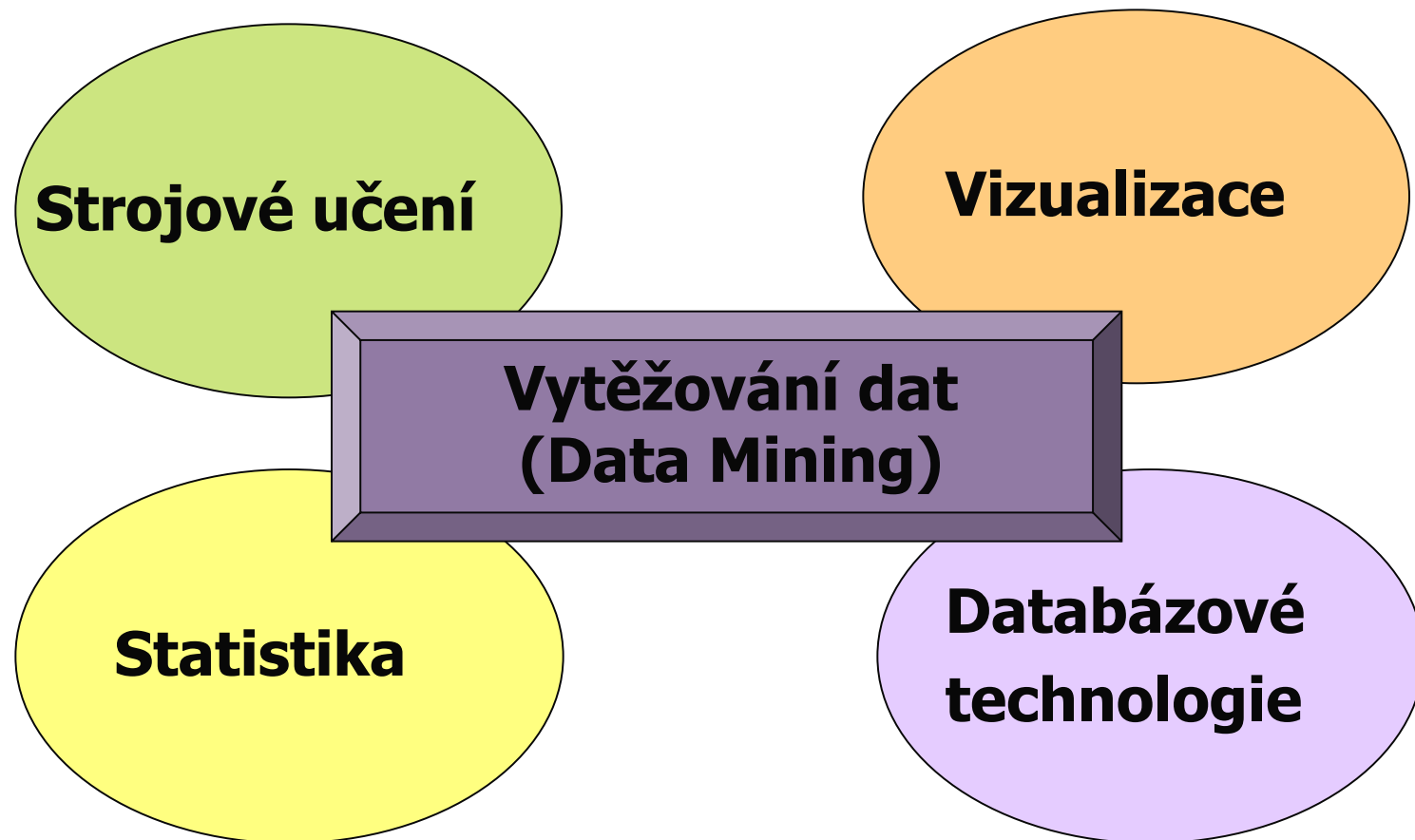
# K něčemu to tedy dobré je!

- Data mining prokázal užitečnost (\$\$\$) v mnoha reálných aplikacích.
- Obor disponuje metodami pro důkladné ověřování a hodnocení výsledků vytěžování.
- Jejich důsledné využití odlišuje „seriózní“ DM od šarlatánství.
- I proto tento předmět.

Vytěžování dat

# **SOUVISLOSTI S DALŠÍMI OBORY**

# Souvislosti



# Souvislosti: Strojové učení

- Obor **umělé intelligence**
- Systémy (algoritmy) zlepšující svoji činnost na základě zkušenosti
  - Zkušenost = data
  - Zlepšování činnosti: obvykle pomocí hledání vzorů v datech
- Podobný cíl s VD, algoritmy SU lze využít
  - Ale rozdíly (srozumitelnost vzorů, datové transformace, ...)

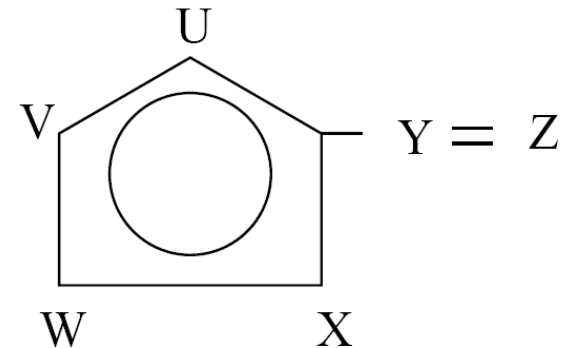
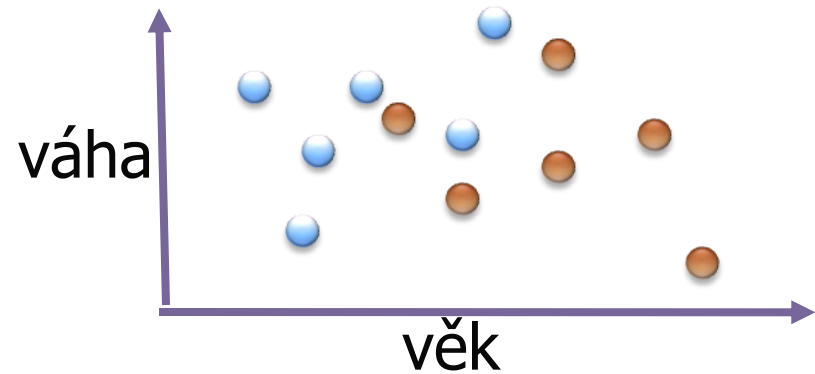


# Souvislosti: Statistika

- Statistická analýza
  - Odhad statistik (průměr, variance, ...)
  - Ověřování **hypotéz**
    - Staří volí stranu X, mladí stranu Y
    - Potvrují to nasbíraná data?
- Explorační analýza
  - **Člověk** vymýšlí a ověřuje množství hypotéz
- Data mining
  - **Počítač** vymýšlí a ověřuje množství hypotéz

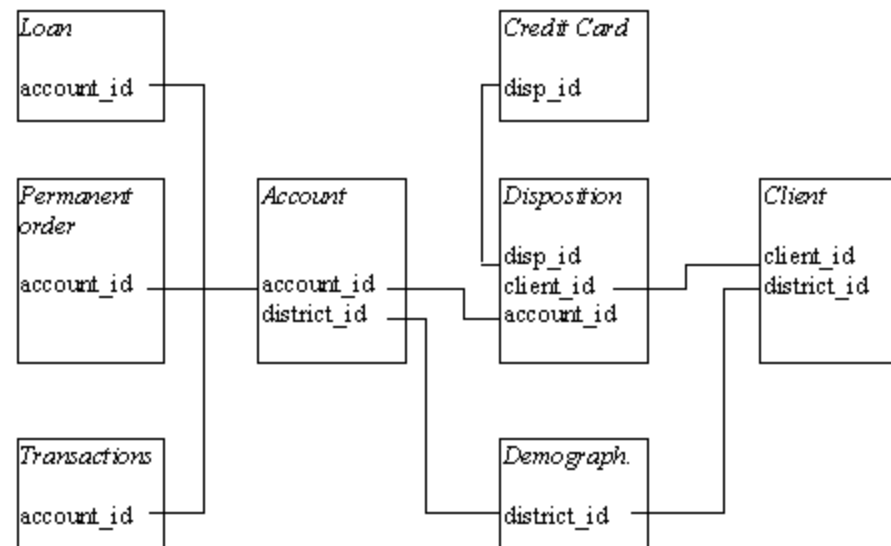
# Souvislosti: Vizualizace

- **Vizualizace dat**  
může odhalit  
hypotézy
- Často je nutné  
**vizualizovat**  
**vzory**



# Souvislosti: Databáze

- Většina aplikací data miningu předpokládá vstupní data ve formě relační databáze



- (PKDD 1999 challenge, <http://lisp.vse.cz/pkdd99/>)

# Souvislosti: Databáze

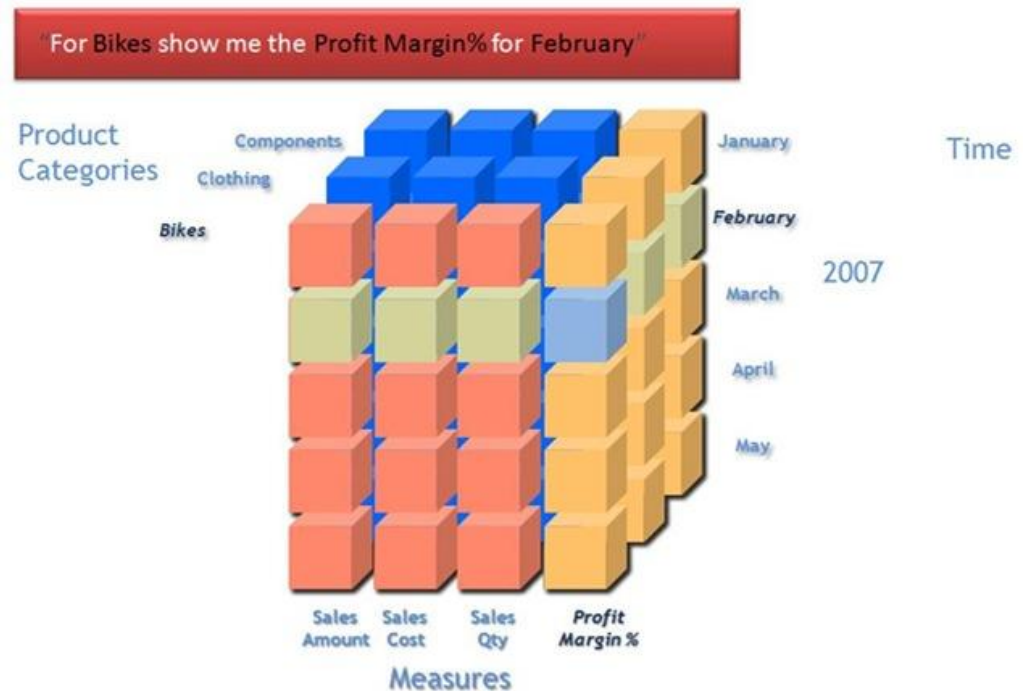
- Vzor může mít podobu **relačního databázového dotazu** (SQL)

```
SELECT Client.ID FROM Client, Disposition, Account  
WHERE Client.District = Praha and Client.Client_ID =  
Disposition.Client_ID AND Account.Account_ID =  
Disposition.Account_ID and Account.Balance > 1000000
```

- Napr. tento vzor charakterizuje „bezproblemové klienty“
- Efektivita DM závisí na schopnosti DB systému rychle vyhodnotit tyto dotazy

# Souvislosti: Databáze

- OLAP: Online Analytical Processing
- Predpocítané odpovědi na časté dotazy do relační databáze
- Pro „manazerský data mining“
- [obr.: Wiki]



Vytěžování dat

# VYTĚŽOVÁNÍ JAKO PROCES

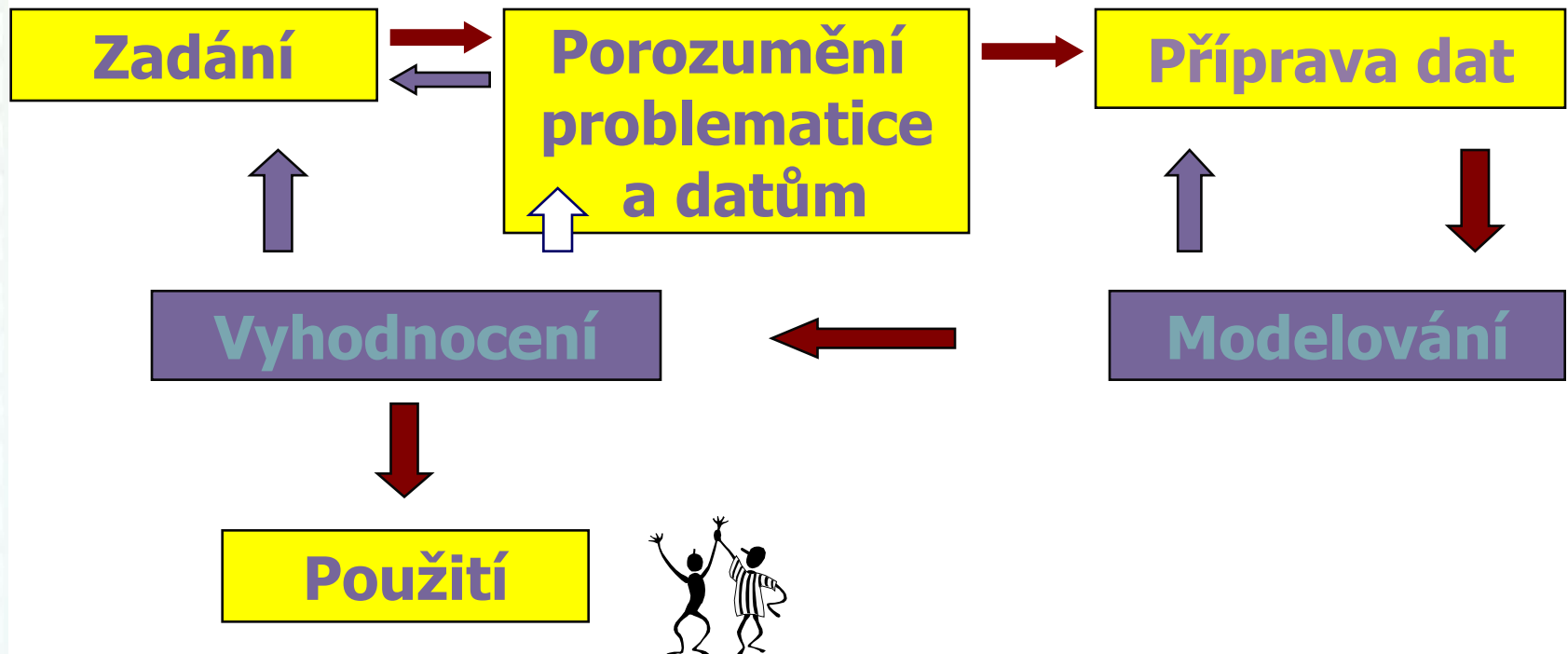
# VD: praktické problémy

- Proč je náročné pracovat s reálnými daty?
- Data nejsou sbírána jako zdroj trénovacích příkladů, ale především kvůli podnikové dokumentaci a archivaci. Z tohoto hlediska bývá sběr i uložení optimalizováno.
- Většina času v data miningových projektech je strávena mimo aktivity hledání vzorů (modelování)
- Téměř vždy opakování cyklu **pokus-omyl**



# CRISP cyklus

- Cross Industry Standard Process for DM
  - 1997-9 navrženo jako prům. standard konsorciem Evrop. firem ([www.crisp-dm.org](http://www.crisp-dm.org))



Vytěžování dat

# **ZÁKLADNÍ POJMY A ÚLOHY**

# Termíny: Model, Vzor, Hypotéza

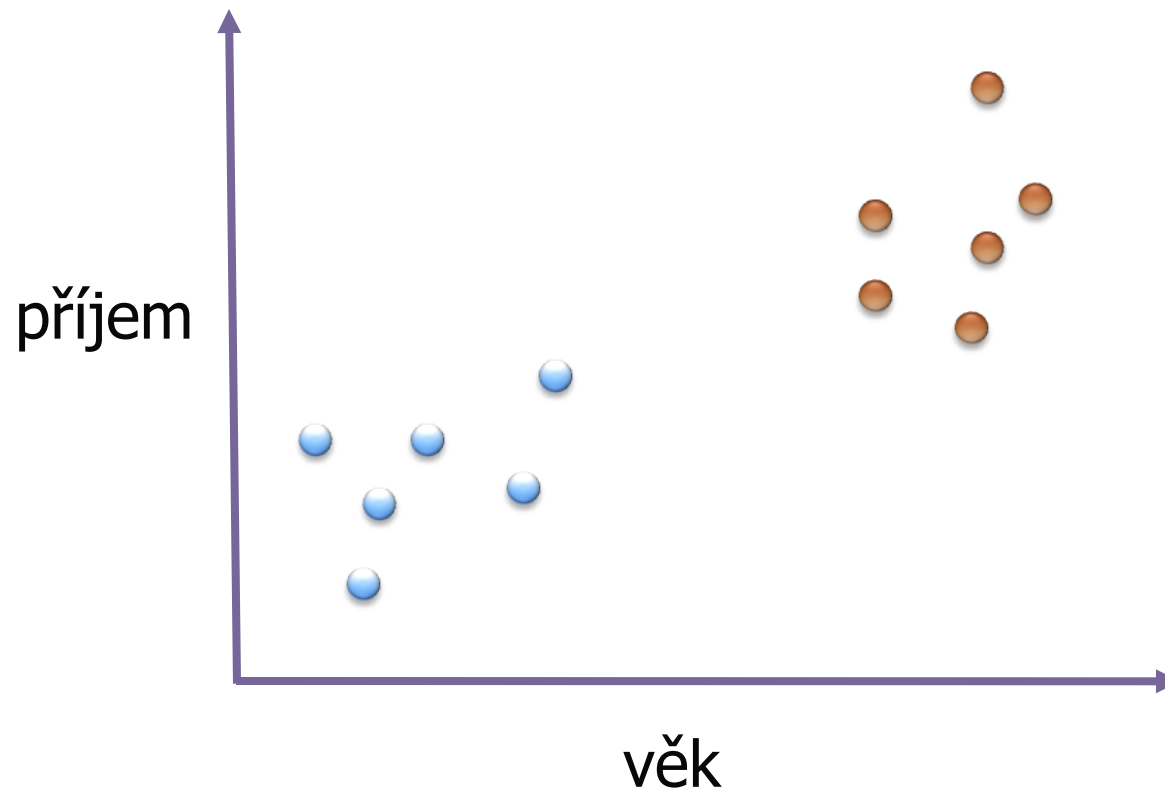
- Vzor (Pattern)
  - Formální **popis obecného principu** v datech
- Model
  - Vzor (příp. množina vzorů) postačující k řešení definované úlohy modelování dat
  - Extrémní případ: vzor(y) umožňující **generovat** pozorovaná data („generativní model“)
- Hypotéza
  - model/vzor, který lze **ověřit/ohodnotit** na datech
- Pozor: „fuzzy“ terminologie. Užití termínů závisí zejména na **kontextu**.

# Obvykle kategorie uloh

- Deskripce (popisné úlohy)
  - Model má charakterizovat vstupní data jako celek
- Hledání nuggetů
  - Hledáme zajímavé vzory, které se mohou týkat třeba jen malých podmnožin vstupních dat
- Klasifikace, Regrese, Predikce
  - Jedna veličina  $V$  je vybrána jako cílová
  - Modelujeme vliv ostatních veličin na  $V$

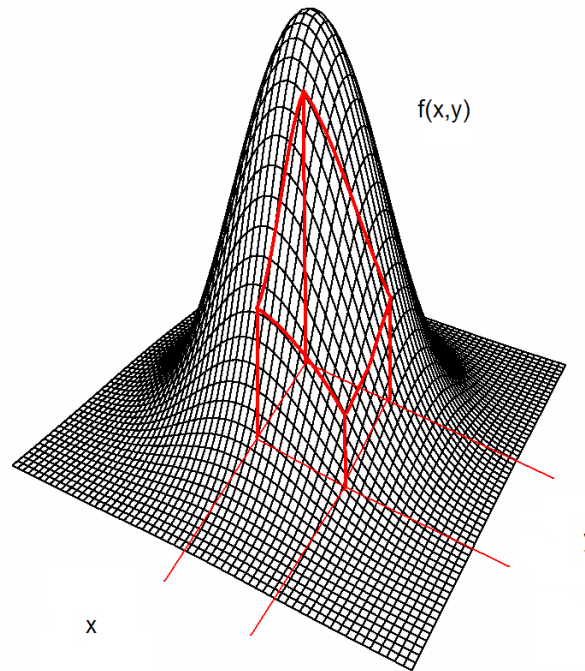
# Deskripce: příklady

- Shluková analýza (aka: segmentace)



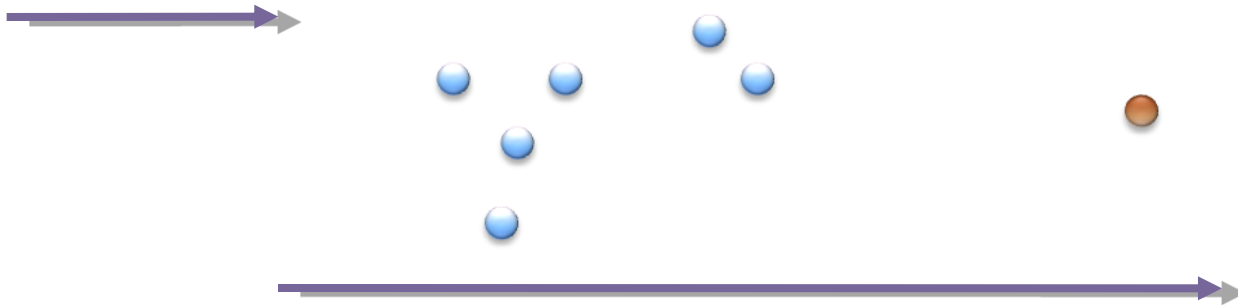
# Deskripce: příklady

- Modelování pravděpodob. rozdělením



# Nuggety: příklady

- Asociace  
    {pivo, párky, horčice}
- Asociační pravidla  
    {párky, horčice} -> {pivo}
- Odlehlé instance (outlier detection)



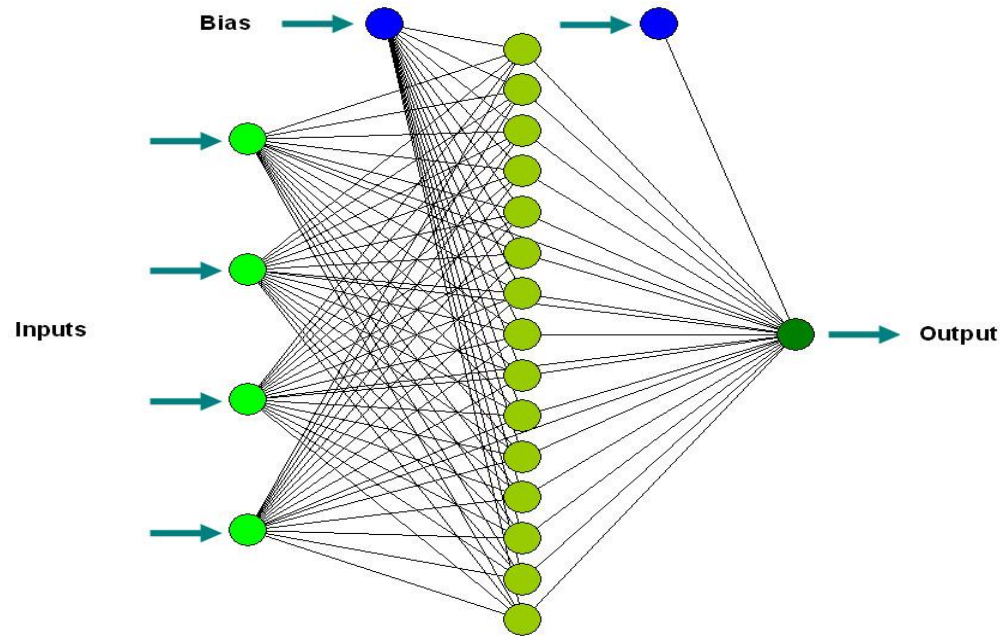


# Klasifikace: příklady

- **Cilový** atribut má **konečnou** množinu hodnot
- Příklad: data: tabulka pacientů
  - Cilový atribut **infarkt**, hodnoty z {ano, ne}
  - Klasifikátor:  
IF váha > 100 and věk > 50 THEN **infarkt-ano**
  - Tento klasifikátor je **symbolický**
  - -> čitelný člověkem

# Klasifikace: příklady

## ■ Příklad **nesymbolického** klasifikátoru



## ■ Neuronová síť s dopředným řetezením

# Regrese a Predikce

## ■ Regrese

- Jako klasifikace, ale cílová veličina **reálné** číslo
- Příklad regresoru

$$\text{brzdná\_dráha} = c * \text{hmotnost} * \text{rychlost}$$

## ■ „Predikce“

- Klasifikační i regresní modely mohou být využity pro **předpověď** hodnoty cílové veličiny
- Z ostatních veličin a/nebo z hodnot cílové veličiny v minulosti

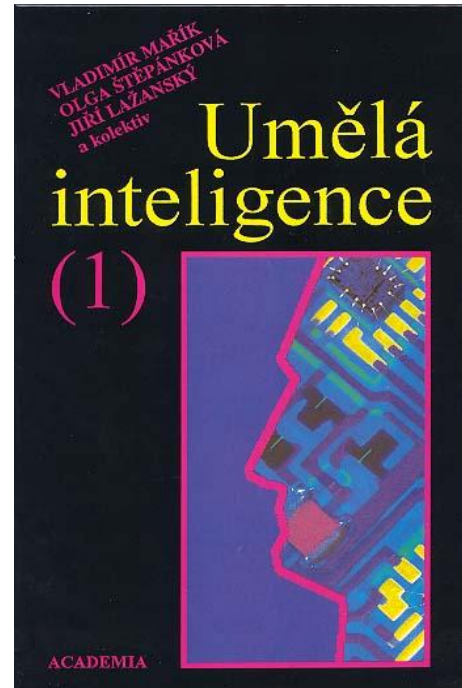
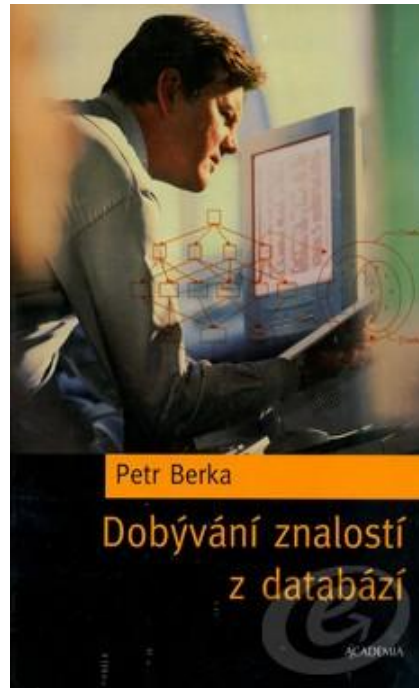
## ■ „Časové rady“

Y336VD Vytěžování dat

# To nejdůležitější teprve přijde!

- Jak modely/vzory **hledat/konstruovat**?
- Jak je **hodnotit**, tj.
  - Jak **odhadovat** jejich kvalitu?
  - Jak **definovat** jejich kvalitu?

# Literatura



<http://cw.felk.cvut.cz/doku.php/courses/y336vd/start>