Data Science: co a k čemu to je

Jan Hučín

2022

Osnova

PROFINIT

1. Co je Data science
2. Kdo je Data scientist
3. Typické a netypické úlohy Data science
4. Příklady z praxe
5. Jak se to dělá: metodika CRISP-DM

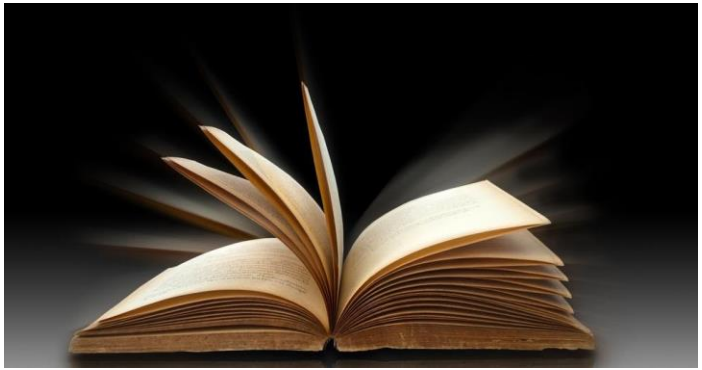
1. Co je Data science

Wikipedie o DS

PROFINIT

Data science (DS) neboli datová věda je **interdisciplinární obor**, který využívá vědecké metody, procesy, algoritmy a systémy pro **získávání znalostí a poznatků z dat** v různých podobách, jak strukturovaných, tak nestrukturovaných, podobně jako data mining.

https://cs.wikipedia.org/wiki/Data_science



DS v bodech

PROFINIT

1. interdisciplinární obor
2. získávání znalostí a poznatků
3. data
4. + podpora rozhodování

Kdo to je?

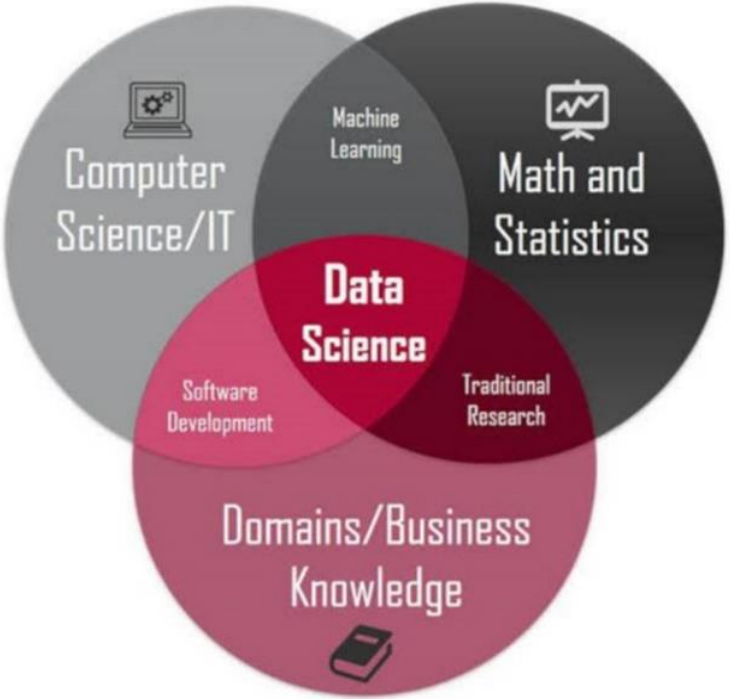
Kdo to je?

A co mají společné?

DS – interdisciplinární obor

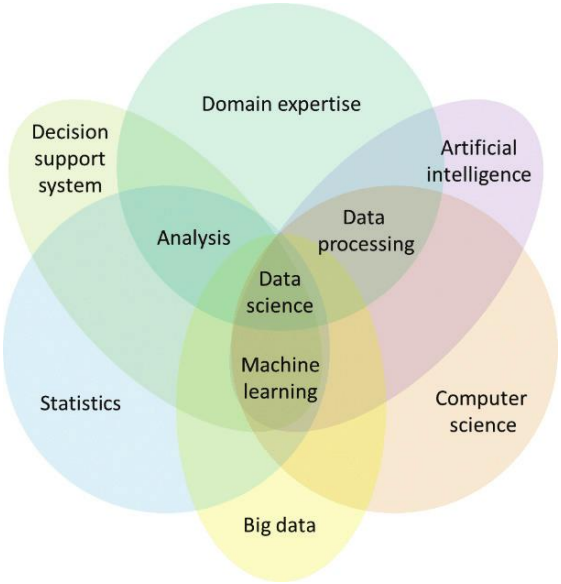
PROFINIT

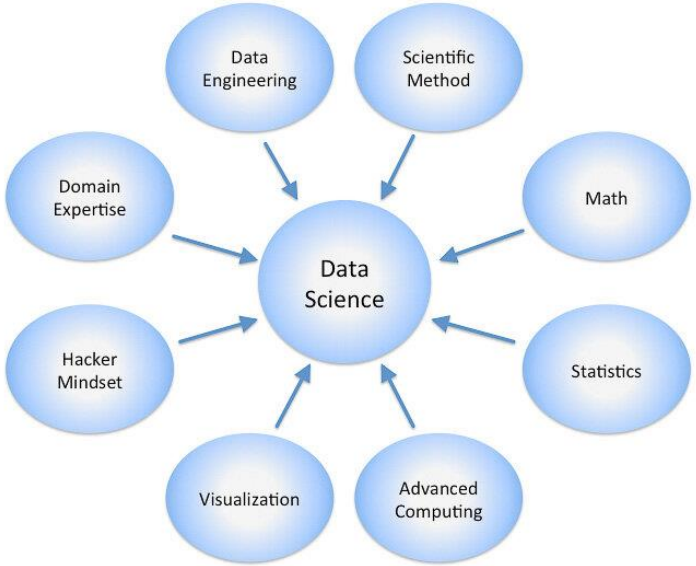
https://www.kdnuggets.com/2020/08/top-10-lists-data-science.html



DS – interdisciplinární obor

PROFINIT

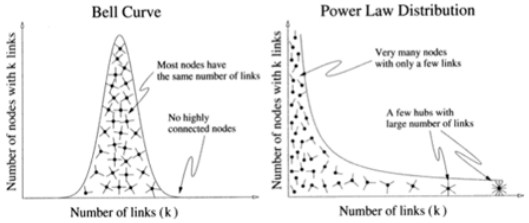




Palmer S. Data Science for the C-Suite. New York: Digital Living Press; 2015

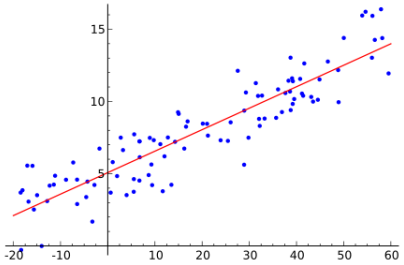
https://commons.wikimedia.org/wiki/File:DataScienceDisciplines.png

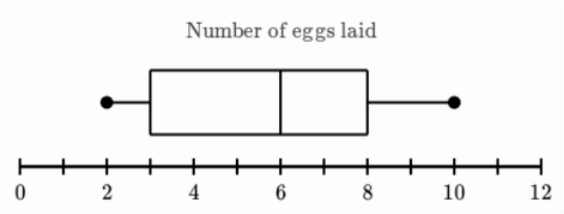
Matematika a statistika



PROFINIT

- Základní otázka
  - Je za tím něco víc, nebo je to jen náhoda?
- Práce s nahodilostí
  - pravděpodobnost, odhady, simulace, predikce
  - Kolik informace je v datech? (signál vs. šum)
- Příklady statistických úloh
  - Kolika let se s pravděpodobností 0,9 dožiji?
  - Jsou muži chytřejší než ženy?
  - Vyplatí se mi tankovat prémiové palivo?
  - Kdo by vyhrál volby, kdyby se konaly dnes?
  - Měl bych skočit z mostu po pozitivním HIV testu?

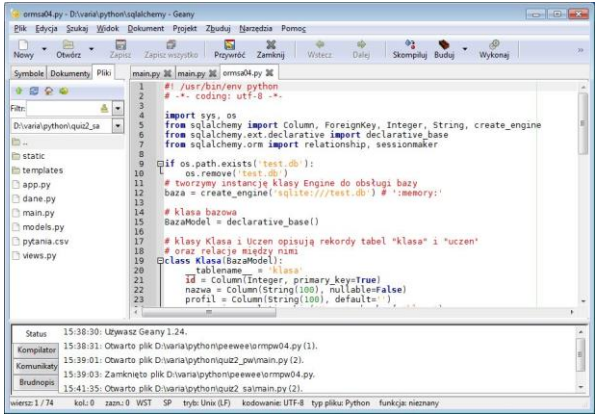




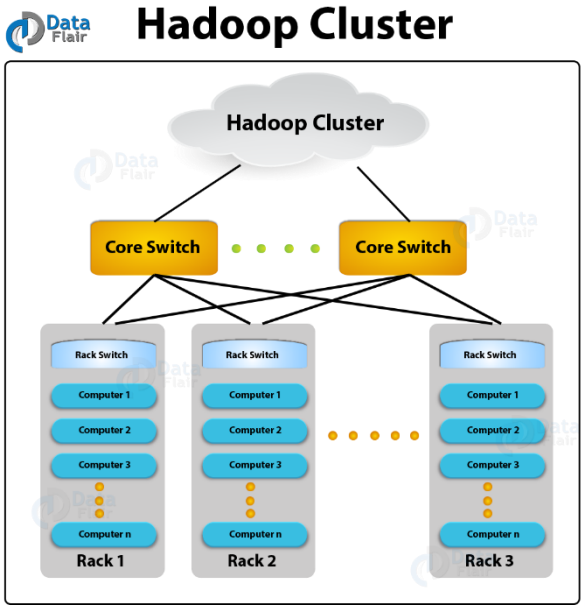
Informatika

PROFINIT

- Základní potřeby:
  - potřeba dat a přístupu k nim
  - potřeba výpočetního výkonu
  - potřeba nasazení v reálném čase

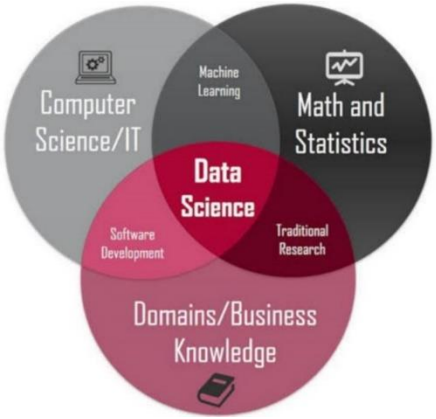


- SQL i NoSQL databáze
- Big Data (→ Big Data Science)
- programování
- analytické nástroje (Rstudio, SPSS)



Machine Learning

- Řešení úloh s využitím:
  - matematických modelů
  - dat, někdy složitých a velkých
  - efektivních algoritmů
  - v praktickém kontextu
- Účel světí prostředky:
  - řešení hrubou silou místo jemné matematiky
  - důležitý je výsledek

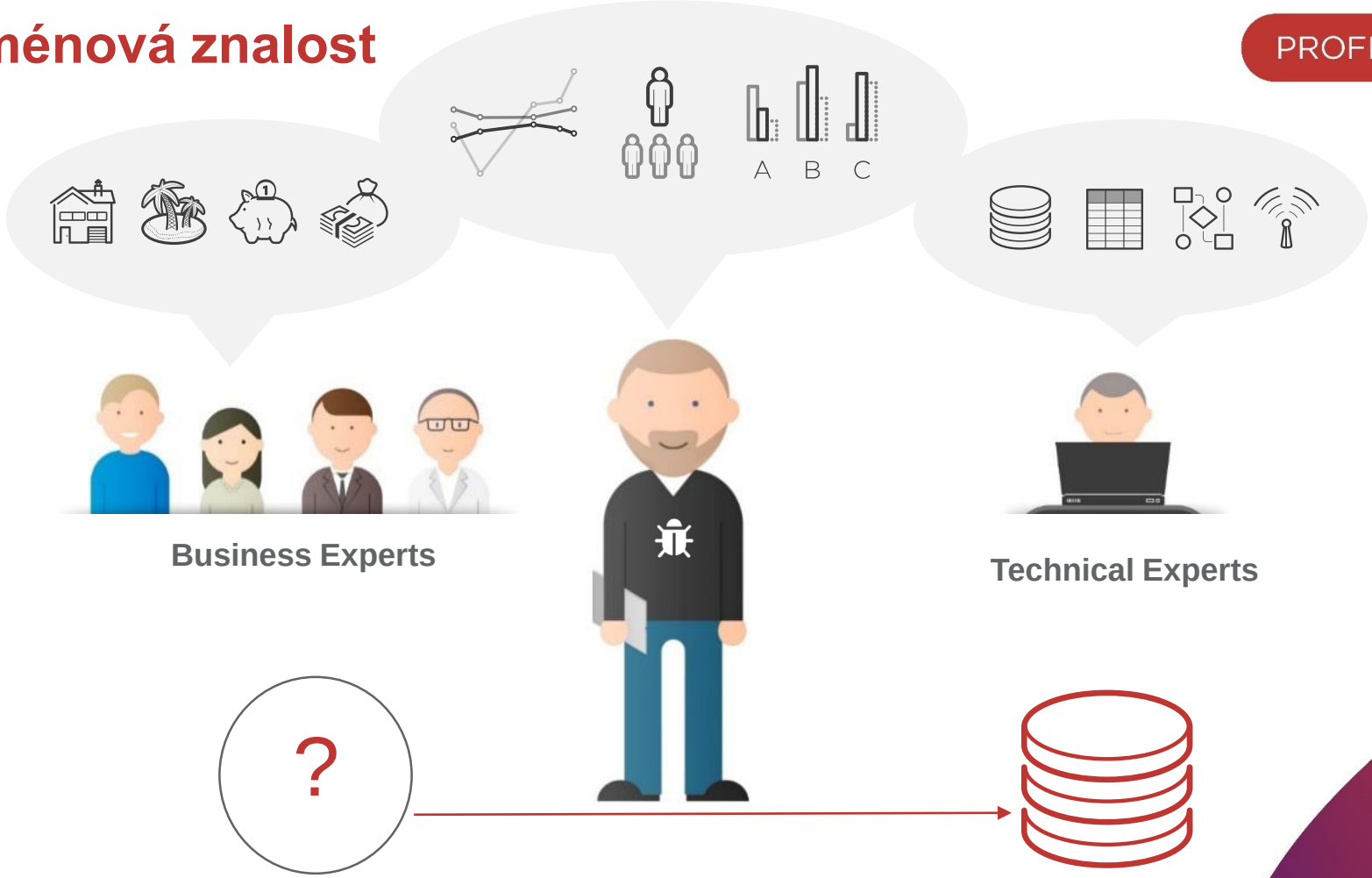


PROFINIT

2. Kdo je Data scientist

Doménová znalost

PROFINIT



Nutné schopnosti Data scientisty

PROFINIT

- orientace v doméně
- komunikace s byznysovými experty
- komunikace s technickými experty
- prezentační dovednosti
- překlad zadání byznysu → požadavky na data, na analytiku → požadavky na implementaci
- překlad analytických výstupů → byznys

3. Typické a netypické úlohy DS

Úlohy z pohledu byznysu

PROFINIT

- **bankovnictví**
  - Komu mám půjčit a s jakým úrokem? Kdo si ode mne nejspíš půjčí?
  - Je tato transakce podvodná?
- **telco**
  - Kdo je s kým partner? Ke kterému rodiči patří tento student?
  - Který klient chce doopravdy utéct jinam?
  - Jaké jsou typické způsoby využívání mobilu?
- **retail, energie**
  - Kdy začíná a končí špička? Kolik zařízení musím mít kdy připraveno?
  - Jak poznám nestandardní situaci? Kdy vyhlásit pohotovost či poplach?
- **jiné**
  - Kolik letadel nestihne přistát před dobou nočního klidu?
  - V kolik hodin dorazím do cíle?

Úlohy z pohledu data science

PROFINIT

- **klasifikace**
  - Je to případ A, nebo B?
- **predikce**
  - Jaká je pravděpodobnost, že nastane událost?
- **odhady**
  - Kolik bude X? Jaký podíl pozitivních se objeví?
- **segmentace**
  - Jaké jsou typické skupiny, typické vzorce chování?
- **detekce**
  - Je to normální případ, nebo anomalita?
- **podobnost**
  - Které případy se sobě podobají? Kdo má s kým vztah?

Netypické úlohy

PROFINIT

- Je tento text odněkud opsaný?
  - analýza výskytu podobných úseků
  - clustering dokumentů podle obsahu

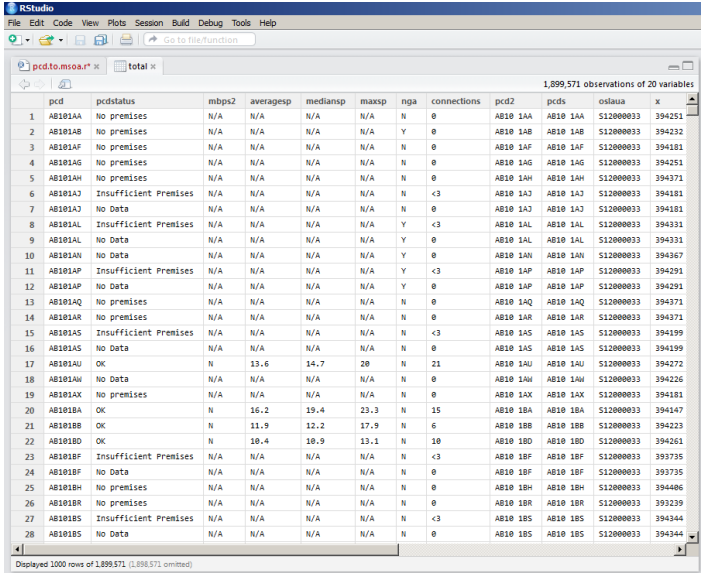
- Co prodává obchod s tímto názvem terminálu?
  - klasifikace do kategorií MCC podle názvu

- Které zboží mám dát v obchodě blízko sebe?
  - časté sady položek, asociační pravidla
- Co dalšího mám klientovi nabídnout?
  - doporučovací systémy

Netypické úlohy

PROFINIT

- Jaký význam mají data v tabulce?
  - nedokumentovaná DB tabulka
  - sémantická analýza neznámých dat
  - shoda sloupce s různými vzory
  - heuristiky
  - nástroj Sémantický analyzér



	pcd	pcdstatus	mbps2	averagesp	medianspp	maxsp	nga	connections	pcd2	pcds	oslaua	x
1	AB101AA	No premises	N/A	N/A	N/A	N/A	N	0	AB10 1AA	AB10 1AA	S12000033	394251
2	AB101AB	No premises	N/A	N/A	N/A	N/A	Y	0	AB10 1AB	AB10 1AB	S12000033	394232
3	AB101AF	No premises	N/A	N/A	N/A	N/A	N	0	AB10 1AF	AB10 1AF	S12000033	394181
4	AB101AG	No premises	N/A	N/A	N/A	N/A	N	0	AB10 1AG	AB10 1AG	S12000033	394251
5	AB101AH	No premises	N/A	N/A	N/A	N/A	N	0	AB10 1AH	AB10 1AH	S12000033	394371
6	AB101AJ	Insufficient Premises	N/A	N/A	N/A	N/A	N	<3	AB10 1AJ	AB10 1AJ	S12000033	394181
7	AB101AJ	No Data	N/A	N/A	N/A	N/A	N	0	AB10 1AJ	AB10 1AJ	S12000033	394181
8	AB101AL	Insufficient Premises	N/A	N/A	N/A	N/A	Y	<3	AB10 1AL	AB10 1AL	S12000033	394331
9	AB101AL	No Data	N/A	N/A	N/A	N/A	Y	0	AB10 1AL	AB10 1AL	S12000033	394331
10	AB101AN	No Data	N/A	N/A	N/A	N/A	Y	0	AB10 1AN	AB10 1AN	S12000033	394367
11	AB101AP	Insufficient Premises	N/A	N/A	N/A	N/A	Y	<3	AB10 1AP	AB10 1AP	S12000033	394291
12	AB101AP	No Data	N/A	N/A	N/A	N/A	Y	0	AB10 1AP	AB10 1AP	S12000033	394291
13	AB101AQ	No premises	N/A	N/A	N/A	N/A	N	0	AB10 1AQ	AB10 1AQ	S12000033	394371
14	AB101AR	No premises	N/A	N/A	N/A	N/A	N	0	AB10 1AR	AB10 1AR	S12000033	394371
15	AB101AS	Insufficient Premises	N/A	N/A	N/A	N/A	N	<3	AB10 1AS	AB10 1AS	S12000033	394199
16	AB101AS	No Data	N/A	N/A	N/A	N/A	N	0	AB10 1AS	AB10 1AS	S12000033	394199
17	AB101AU	OK	N	13.6	14.7	20	N	21	AB10 1AU	AB10 1AU	S12000033	394272
18	AB101AW	No Data	N/A	N/A	N/A	N/A	N	0	AB10 1AW	AB10 1AW	S12000033	394226
19	AB101AX	No premises	N/A	N/A	N/A	N/A	N	0	AB10 1AX	AB10 1AX	S12000033	394181
20	AB101BA	OK	N	16.2	19.4	23.3	N	15	AB10 1BA	AB10 1BA	S12000033	394147
21	AB101BB	OK	N	11.9	12.2	17.9	N	6	AB10 1BB	AB10 1BB	S12000033	394223
22	AB101BD	OK	N	10.4	10.9	13.1	N	10	AB10 1BD	AB10 1BD	S12000033	394261
23	AB101BF	Insufficient Premises	N/A	N/A	N/A	N/A	N	<3	AB10 1BF	AB10 1BF	S12000033	393735
24	AB101BF	No Data	N/A	N/A	N/A	N/A	N	0	AB10 1BF	AB10 1BF	S12000033	393735
25	AB101BH	No premises	N/A	N/A	N/A	N/A	N	0	AB10 1BH	AB10 1BH	S12000033	394406
26	AB101BR	No premises	N/A	N/A	N/A	N/A	N	0	AB10 1BR	AB10 1BR	S12000033	393239
27	AB101BS	Insufficient Premises	N/A	N/A	N/A	N/A	N	<3	AB10 1BS	AB10 1BS	S12000033	394344
28	AB101BS	No Data	N/A	N/A	N/A	N/A	N	0	AB10 1BS	AB10 1BS	S12000033	394344

Netypické úlohy

PROFINIT

- Co znamená tento zápis adresy?
  - nestrukturované, neúplné zápisy, nekonsistence, překlepy
  - dohledávání adresy v registru RÚIAN (po částech)
  - nástroj Geolocator

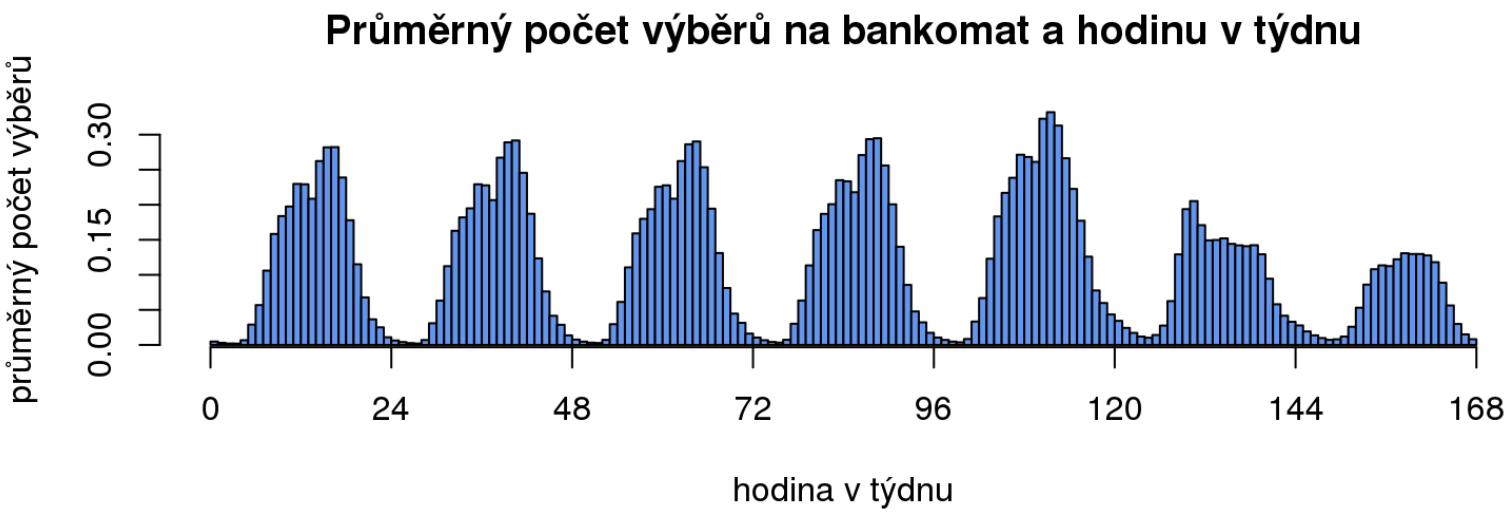
Praha, Vinohrady, Římská 1240/39, 120 00 Praha 2
79692 Prostějov,Fanderlíkova 25
U rybníka 301, Lukov u Zlína, 28507
Šípková 9/,Jaktař,Opava,746 01,Opava,Moravskoslezský,CZ
43151, Klášterec, Krátká 641
392/9 a 393/11
Střední 135/16
Babice 48
Liberec, Rujancova 11

4. Příklady z praxe

Typologie klientů podle výběru z bankomatu

PROFINIT

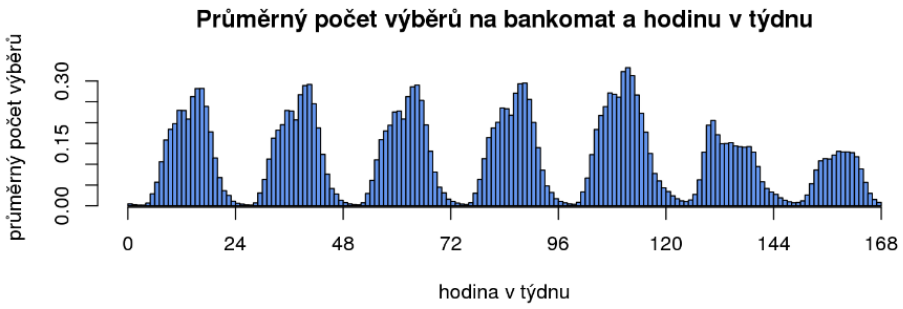
- data z malé české banky
- id klienta, název bankomatu, čas výběru, částka



Typologie klientů podle výběru z bankomatu

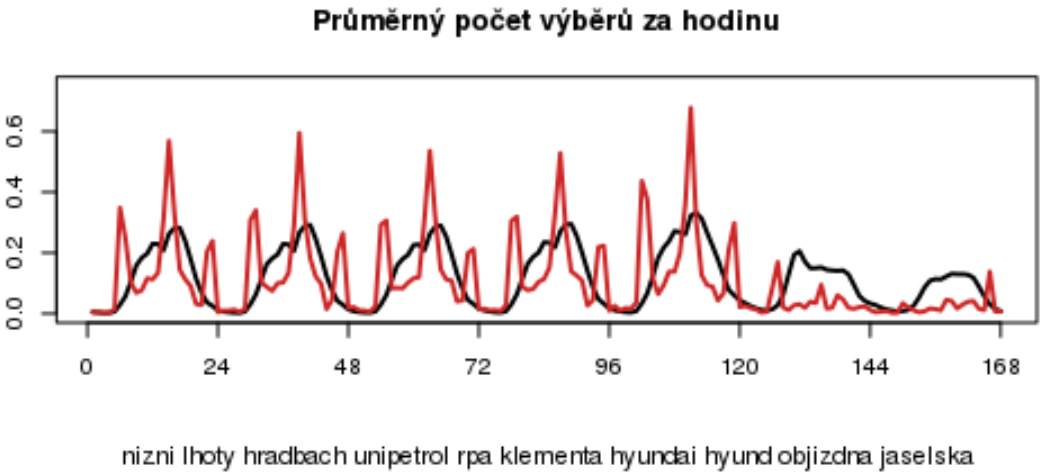
PROFINIT

- bankomaty se liší průběhem týdenní křivky výběrů
- bankomat → vektor průměrných počtů výběrů přes týden
- redukce dimenze UMAP do 2D
- clustering DBScan
- mnoho zajímavých clusterů



Typologie klientů podle výběru z bankomatu

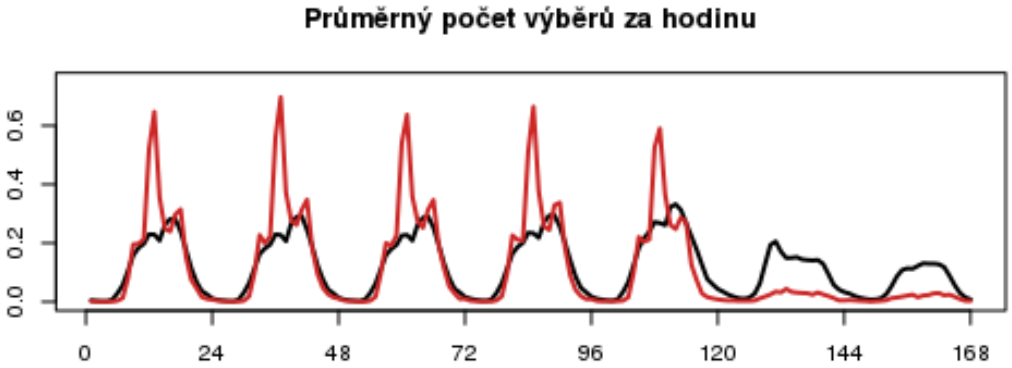
PROFINIT



CZ:Plzen::CSOB 2052 PLZEN
CZ:KOLIN::KB ATM NA HRADBACH
CZ:NIZNI LHOTY::CS, NIZNI LHOTY, HYUND
CZ:PRELOUC::CS, JASELSKA 593
CZ:NIZNI LHOTY::CS, HYUNDAI 700/1
CZ:MLADA BOLESLA::CS, VACLAVA KLEMENTA
CZ:LITVINOV::CS, UNIPETROL RPA
CZ:OTROKOVICE::KB ATM OBJIZDNA 1628
CZ:MLADA BOLESLA::CS,VACL.KLEMENTA,SKODA
CZ:Mlada Bolesla::CSOB 1210 ML BOLESLAV

Typologie klientů podle výběru z bankomatu

PROFINIT

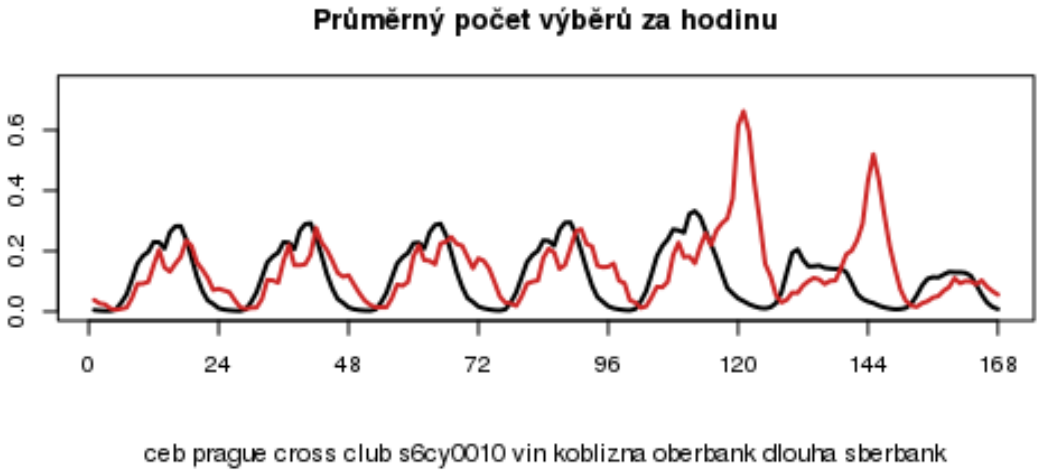


xankraci holandska kolejni her sberbanks1as1306 mps churchilla hrebenech parku kavc

CZ:PRAHA 4::CS. NA HREBENECH II 11
CZ:PRAHA 8::CS. POBREZNI 665/21
CZ:Ceske Budejov::CSOB 2118 C BUDEJOVIC
CZ:PRAHA 4::CS. NA PANKRACI 1683/1
CZ:Praha::RB Hvezdova
CZ:PRAHA 2 - MPS::UNICREDIT - PRAHA 2
CZ:Praha 1::CSOB 1820 PRAHA 1
CZ:Praha 4::MMB NA PANKRACI 1724/1
CZ:Brno-Kralovo::CSOB 0882 BRNO
CZ:Praha 4::RB CT Kavci hory
CZ:Praha::RB Prosek Point
CZ:Brno::SBERBANK 7013 HER.
CZ:Brno::CSOB 4029 BRNO
CZ:Praha 8::CSOB 1989 PRAHA 8
CZ:PRAHA 6::CS. NAR.TECH.KNIHOVNA

Typologie klientů podle výběru z bankomatu

PROFINIT



CZ:Brno::CZE00421
CZ:Ostrava::CZE00236
CZ:PRAHA 1::CZE00487
CZ:PRAHA 1::CS. DLOUHA 743/9
CZ:Praha 2::SBERBANK 7020 VIN.
CZ:Praha 1::CSOB 1216 PRAHA 1
CZ:Praha 2::CZE00183
CZ:PRAHA 1::KB ATM DLOUHA 34
CZ:Ceske Budejov::OBERBANK 7507 CEB
CZ:BRNO::KB ATM KOBLIZNA 3
CZ:Praha 1::SBERBANK 7007 P1-VN
CZ:PRAHA 1::ATM S6CY0010
CZ:Prague::ATM Cross club

Typologie klientů jinak

PROFINIT

- podle plateb na čerpacích stanicích (má auto?)
- podle plateb ve spec. obchodech (cestuje? je to kutil? má doma zvíře?)
- podle plateb za služby (má dům? má chalupu? má děti?)
- podle plateb za sázky, gambling apod.
- atd.

Další příklady

PROFINIT

- fraud detection v IB
  - jak poznat neobvyklé chování?
- risk z online žádosti
  - prozrazuje chování klienta rizikovost?
- text2bank
  - krátký text v IB → určení požadované akce
- clever pricing
  - vysoká úroková sazba (cena) = nepůjčí si nikdo
  - nízká úroková sazba = nic nevydělám
  - někde mezi tím je optimum

5. Jak se to dělá: metodika CRISP-DM

CRISP-DM

Cross Industry Standard Process for Data Mining

PROFINIT

- Staré (1996), ale stále dobré
  - „nákupní seznam“ pro plánování
- Iterativní povaha
  - Data science = science, tj. upřesňování
  - předchůdce agilních metodik

CRISP-DM

PROFINIT

Fáze CRISP-DM

PROFINIT

1. Business Understanding
   - Co zákazník potřebuje? Jaká k tomu má data? Jak se pozná úspěch?
2. Data Understanding
   - posbírání dat, exploratorní analýza, kvalita dat, první testy hypotéz
3. Data Preparation
   - konstrukce datové sady pro modelování; sestavení, transformace a výběr příznaků, redukce dimenzionality atd.
4. Modeling
   - aplikace modelovacích technik, výběr modelu, kalibrace parametrů, měření výkonnosti modelu
5. Evaluation
   - vyhodnocení úspěšnosti modelu vzhledem k věcným kritériím
6. Deployment
   - produkční nasazení modelu v datovém workflow zákazníka; vyřešení administrace, údržby, zaškolení, rekalibrace…

Bonus: aktuální problémy DS

Aktuální problémy DS

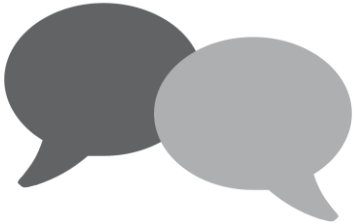
PROFINIT

- Využívání osobních dat – pro a proti
  - personifikovaná reklama, produkt na míru
  - ochrana před podvody, zneužitím
  - big brother
- Přesnost a aktuálnost modelů
  - overfitting
  - globální změny znehodnotí modely
- Slepé uličky
  - mnoho úloh nemá předchůdce
  - stovky PoC → desítky jednorázových projektů → jednotky v produkci
  - vyvinout si interně, nebo outsourcovat?

Reference

PROFINIT

- Sinan Ozdemir: Principles of Data Science
- Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, Abhijit Dasgupta: Practical Data Science Cookbook
- Frank Kane: Hands-On Data Science and Python Machine Learning



Diskuse