Podatkovno rudarjenje in odkrivanje zakonitosti v podatkovnih bazah

Ljupčo Todorovski
Univerza v Ljubljani, Fakulteta za upravo

Maj 2018

Pregled predavanja

Podatkovno rudarjenje (*data mining*)

- Definicije in povezani pojmi
- Proces podatkovnega rudarjenja
- Opis posameznih faz procesa

Seminarske naloge

- Zakup (lizing) avtomobilov
- Prebeg (churn) strank
- Pokrovnost zemljišč
- parlameter.si

Definicije

Friedman (ESL)

Računalniško-podprta, raziskovalna analiza podatkov v (običajno) velikih, kompleksnih podatkovnih množicah.

Witten (Weka)

- Pristop k reševanju problemov z analizo podatkov shranjenih v podatkovnih bazah.
- Odkrivanje implicitnih, predhodno neznanih in potencialno uporabnih informacij iz podatkov.

Aggarwal

Proučevanje zbiranja, čiščenja, obdelave, analize in pridobivanja koristnega znanja iz podatkov.

Podatkovno rudarjenje in strojno učenje

Strojno učenje

Proučevanje algoritmov, ki izboljšujejo svojo uspešnost iz izkušenj.

- Učenje je izboljševanje uspešnosti v nekem okolju skozi pridobivanje znanja, ki je rezultat izkušenj v tem okolju.

Povezava s podatkovnim rudarjenjem

Strojno učenje je bistveni korak podatkovnega rudarjenja, saj algoritmi strojnega učenja omogočajo odkrivanje vzorcev in modelov iz podatkov.

Razlika od podatkovnega rudarjenja

Fokus na algoritmih za učenje in ne na odkrivanju vzorcev in modelov.

Odkrivanje zakonitosti v bazah podatkov

KDD: Knowledge Discovery in Databases

Definicija

Množica tehnologij za odkrivanje ne-trivialnih, implicitnih, predhodno neznanih in potencialno koristnih informacij iz podatkov shranjenih v podatkovnih bazah.

Odkrite informacije so običajno vzorci in modeli, ki omogočajo

1. boljše razumevanje podatkov,
2. napovedovanje bodočega obnašanja sistemov in
3. boljše odločanje.

Povezava s podatkovnim rudarjenjem

- Zajema celotni podatkovni cikel in ne le odkrivanje vzorcev in modelov
- Številni avtorji uporabljata pojma kot sinonima

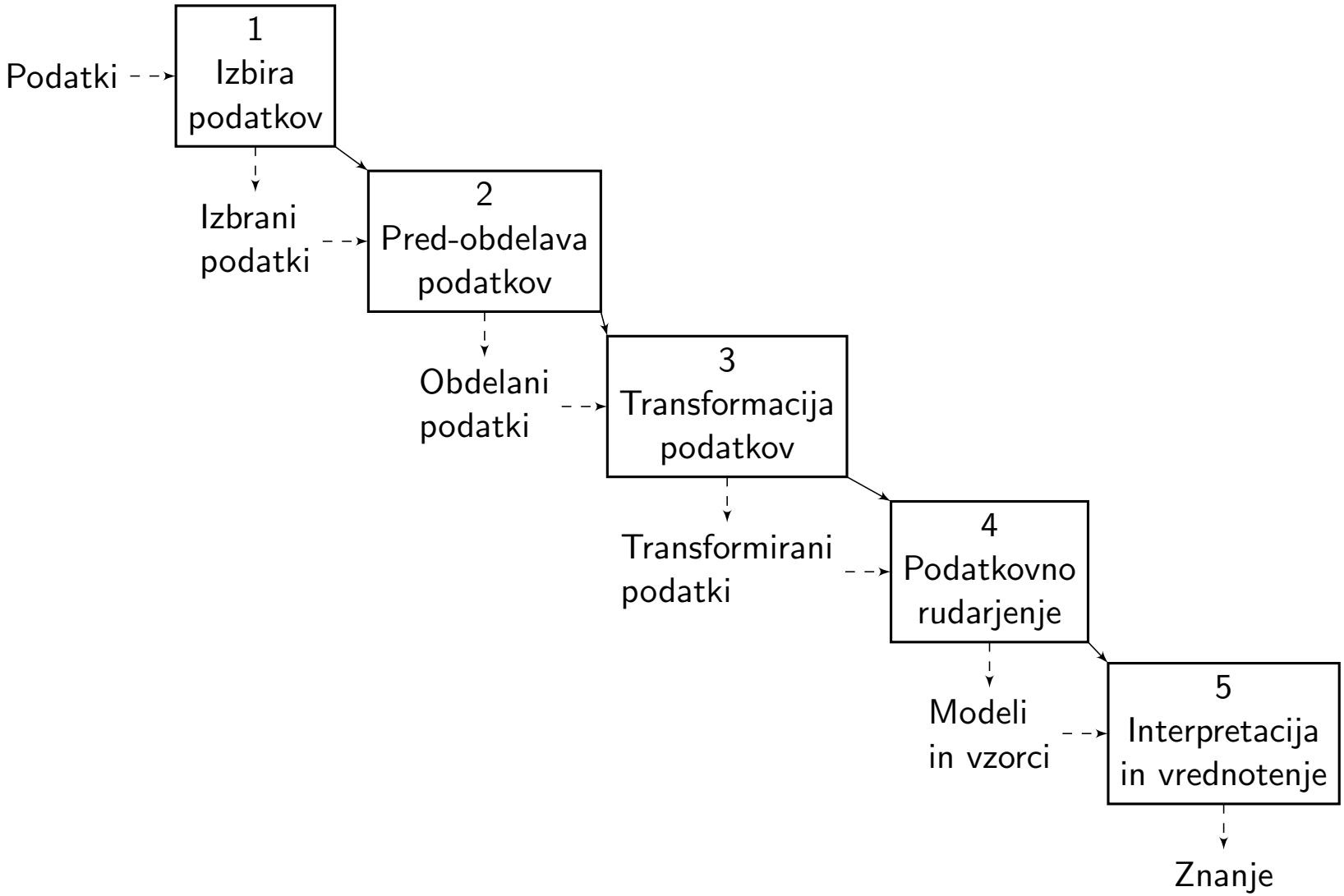
Proces odkrivanja zakonitosti v bazah podatkov

KDD je kompleksen proces izbire, priprave, obdelave in rudarjenja podatkov ter vrednotenja in uporabe odkritih modelov in vzorcev.

Koraki procesa *KDD*

1. *Selection*: izbira podatkov
2. *Preprocessing*: pred-obdelava podatkov
3. *Transformation*: transformacija podatkov
4. *Data mining*: podatkovno rudarjenje, odkrivanje vzorcev in modelov
5. *Interpretation and evaluation*: interpretacija in vrednotenje vzorcev in modelov ter njihova pretvorba v znanje

Koraki procesa *KDD*



1 Izbira podatkov

Ključno vprašanje

Kateri podatki so relevantni za problem, ki ga rešujem?

Tudi zbiranje podatkov

- Posebna strojna oprema: senzorji
- Posebna programska oprema: pajki
- Ročno: uporabniške ankete

Ključna vloga področnih ekspertov.

2 Pred-obdelava podatkov

Ključna naloga

Priprava podatkov v obliko, ki je ustrezna za analizo.

Običajna oblika je tabela

- Vrstice: primeri
- Stolpci: spremenljivke, tudi značilke (*features*) ali atributi

Običajne težave

- Integracija različnih podatkovnih virov
- Ne-strukturirani ali pol-strukturirani podatkovni tipi, predavanje *Obravnava različnih tipov podatkov in vložitve*

3 Transformacija podatkov

Tri ključne naloge

Čiščenje podatkov

- Manjkajoči in napačni podatki
- Predavanje *Obravnava nepopolnih podatkov*

Izbira in transformacija ciljnih in napovednih spremenljivk

Predavanje *Izbira in konstrukcija napovednih spremenljivk*

Vzorčenje in obteževanje primerov

- Izbira primerov za učne in testne podatkovne množice
- Obteževanje primerov pri neenakomernih porazdelitvah
- Predavanje *Neenakomerna porazdelitev vrednosti ciljne spremenljivke*

4 Podatkovno rudarjenje

Ključni nalogi

- Učenje napovednih modelov iz podatkov
- Odkrivanje vzorcev v podatkih

Uporaba metod strojnega učenja

Kreativna uporaba in kombiniranje metod za učenje napovednih modelov in odkrivanje vzorcev iz prvega dela semestra.

5 Interpretacija in vrednotenje

Ključna vprašanja

- Kakšen je pomen modelov v kontekstu reševanja problema?
- Kakšen je pomen vzorcev v istem kontekstu?
- Kakšen je prispevek modelov in vzorcev k znanju s področja?

Ključna vloga domenskih ekspertov.

John Snow, London 1854

Kratek opis: ZAKUP

Stopnja zahtevnosti

1: za popolne začetnike (ni bonusa)

Področje

Tveganje finančnega zakupa (lizing) vozil

Koraki procesa *KDD*

3 (Transformacija podatkov) in 4 (Podatkovno rudarjenje)

Opis podatkov: ZAKUP

Binarna klasifikacija

Primeri

- 6.062 predmetov finančnega zakupa v učni množici
- 1.516 predmetov v testni množici

31 spremenljivk

- Predmet zakupa (4): odobren (ciljna), polog, obrestna mera, trajanje
- Vozilo (4): znamka, letnik, cena in ocena vrednosti Eurotax
- Stranka (12): starost, neto plača, ...
- Delodajalec stranke (9): tip, število zaposlenih, ...
- Dobavitelj (2): število pogodb, črna lista

Kriteriji za ocenjevanje: ZAKUP

1. Čim manjša napaka na testnih podatkih
2. Diskusija relevantnosti napovednih spremenljivk

Pozor

Razlika med napakami tipa *FP* in *FN*, več kot 60% odobrenih predmetov in sistematično manjkajoče vrednosti.

Kratek opis: PREBEG

Stopnja zahtevnosti

2: za nabiralce izkušenj (5 bonus točk)

Področje

Prebeg (*churn*) strank k drugemu telekomunikacijskemu operaterju.

Koraki procesa *KDD*

3 (**Transformacija podatkov**) in 4 (Podatkovno rudarjenje)

Opis podatkov: PREBEG

Binarna klasifikacija

Primeri

- 40.000 strank telekomunikacijskega podjetja Orange
- 10.000 strank v testni množici

78 spremenljivk

- Prebeg (ciljna): binarna
- 42 numeričnih in 35 diskretnih
- Napovedne spremenljivke neznanega pomena
- Povprečen delež manjkajočih vrednosti na spremenljivko: 12%

Kriteriji za ocenjevanje: PREBEG

1. Čim manjša napaka na testnih podatkih
2. Čim manj napak tipa FN

Pozor

Razlika med napakami *FP* in *FN*, le dobrih 7% prebeglih strank in ogromno manjkajočih vrednosti.

Kratek opis: POKROV

Stopnja zahtevnosti

3: za izkušene **in pogumne** (10 bonus točk)

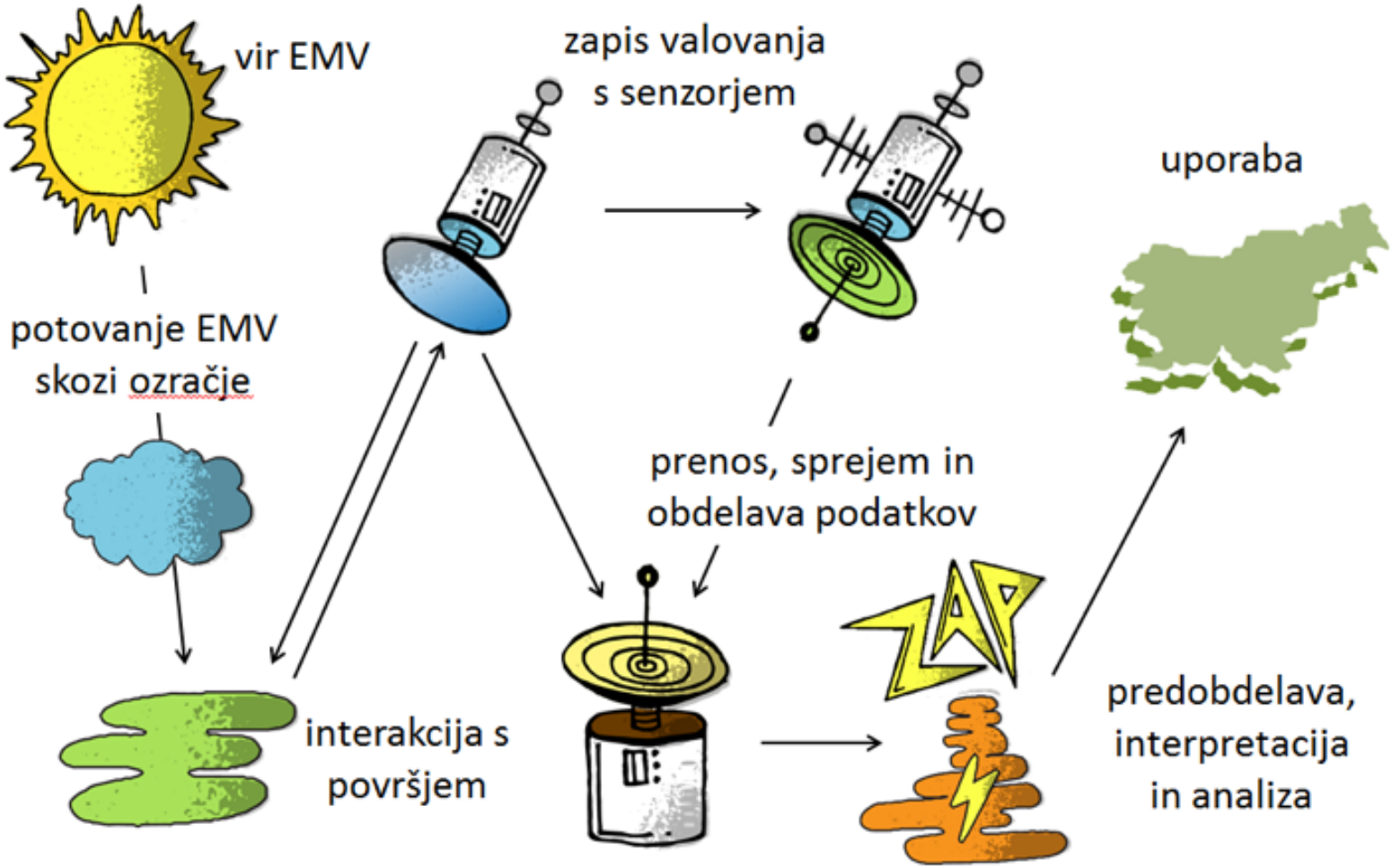
Področje

Ugotavljanje rabe zemljišč (pokrovnosti) iz satelitskih posnetkov.

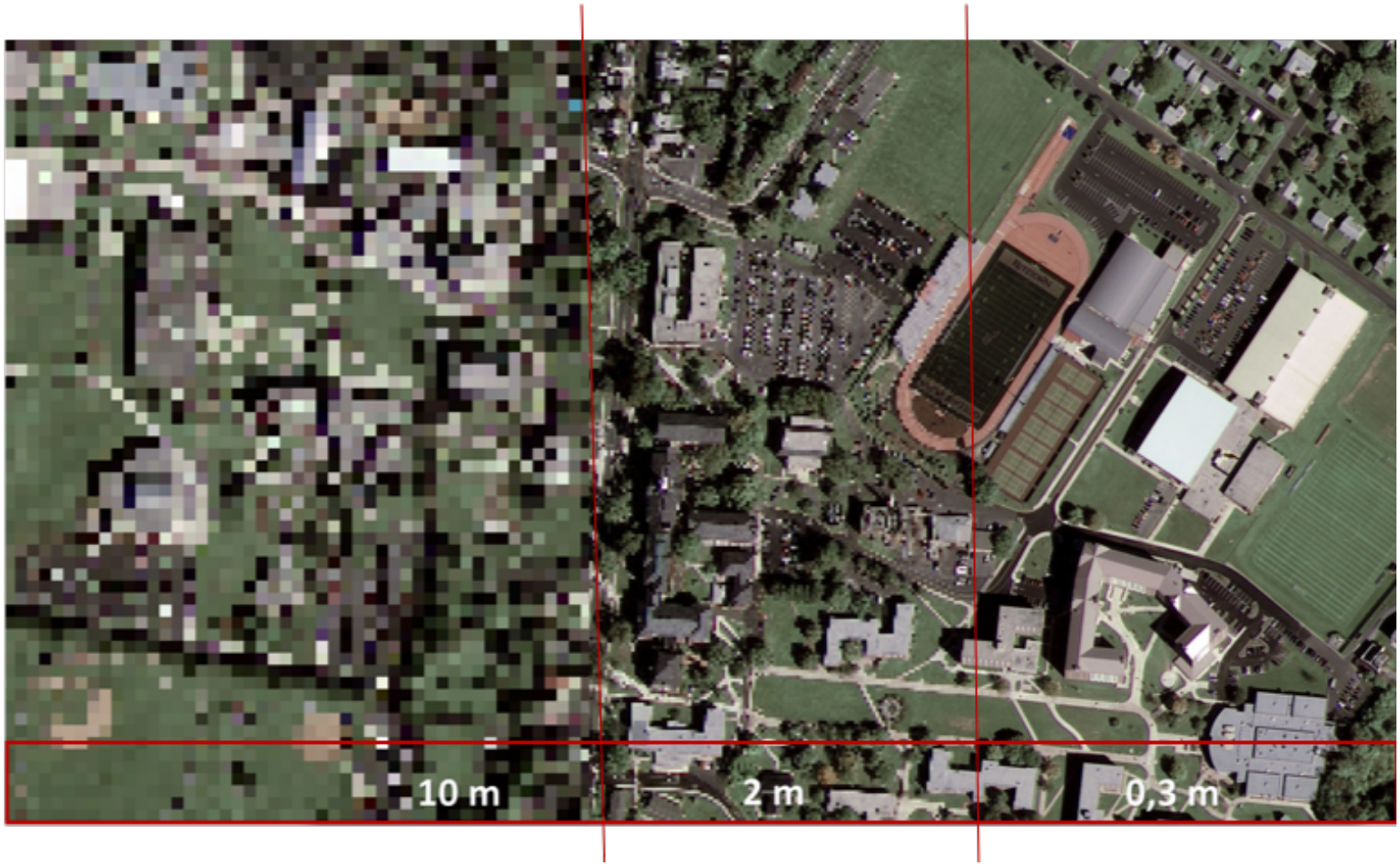
Koraki procesa *KDD*

2 (Pred-obdelava podatkov), 3 (**Transformacija podatkov**) in
4 (Podatkovno rudarjenje)

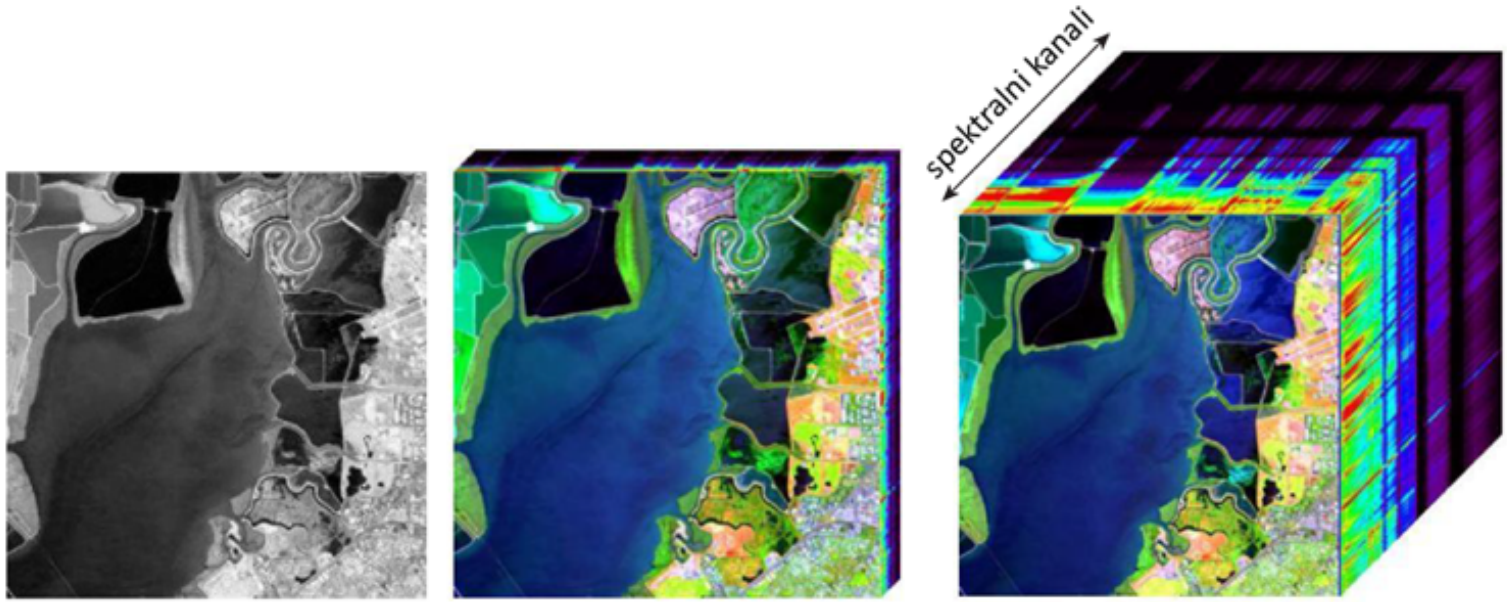
POKROV: Daljinsko zaznavanje



POKROV: Satelitski posnetki in prostorska ločljivost



POKROV: Satelitski posnetki in spektralna ločljivost



Opis podatkov (**okvirni**): POKROV

Binarna in navadna klasifikacija

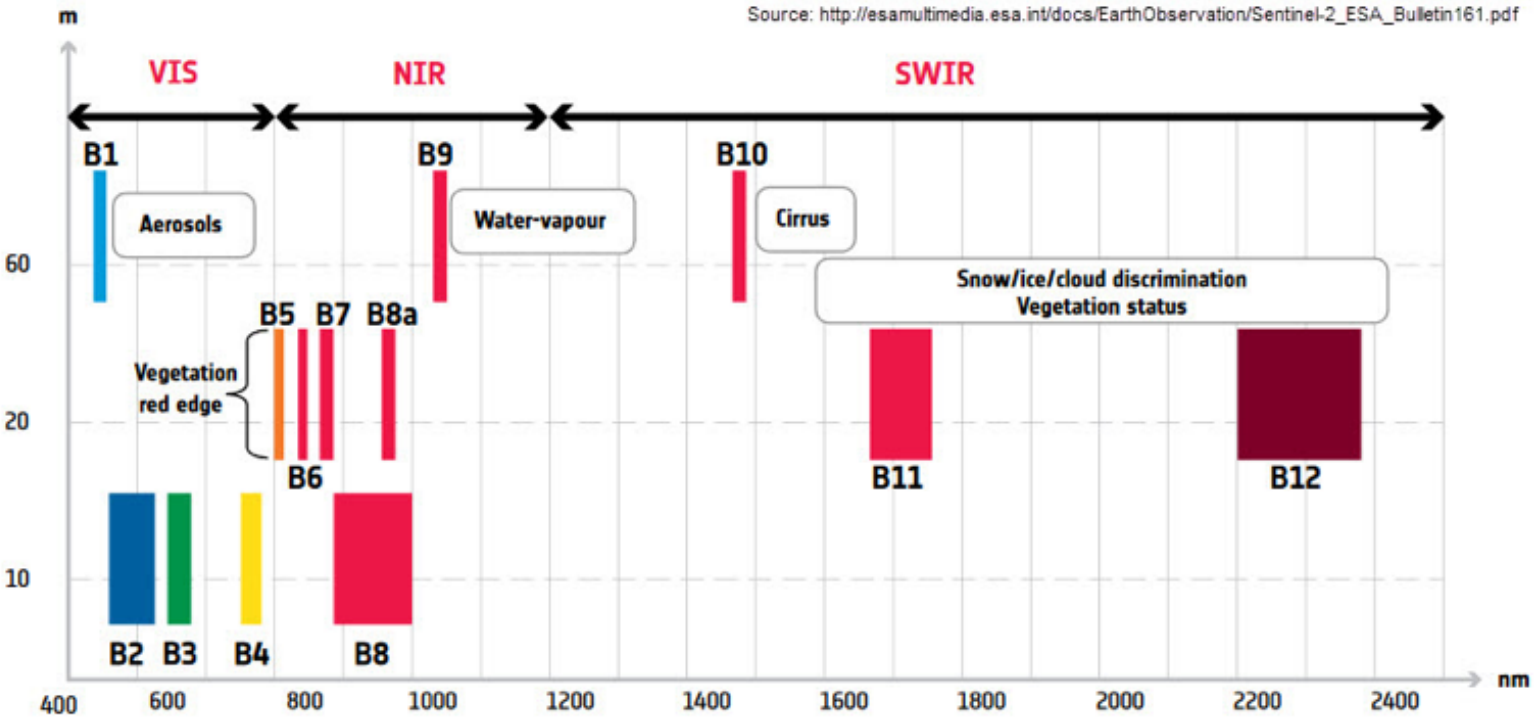
Primeri

- 10.000 pikslov velikosti 10×10 metrov (v naravi)
- Dobra polovica učnih primerov, ostalo testni

48 spremenljivk

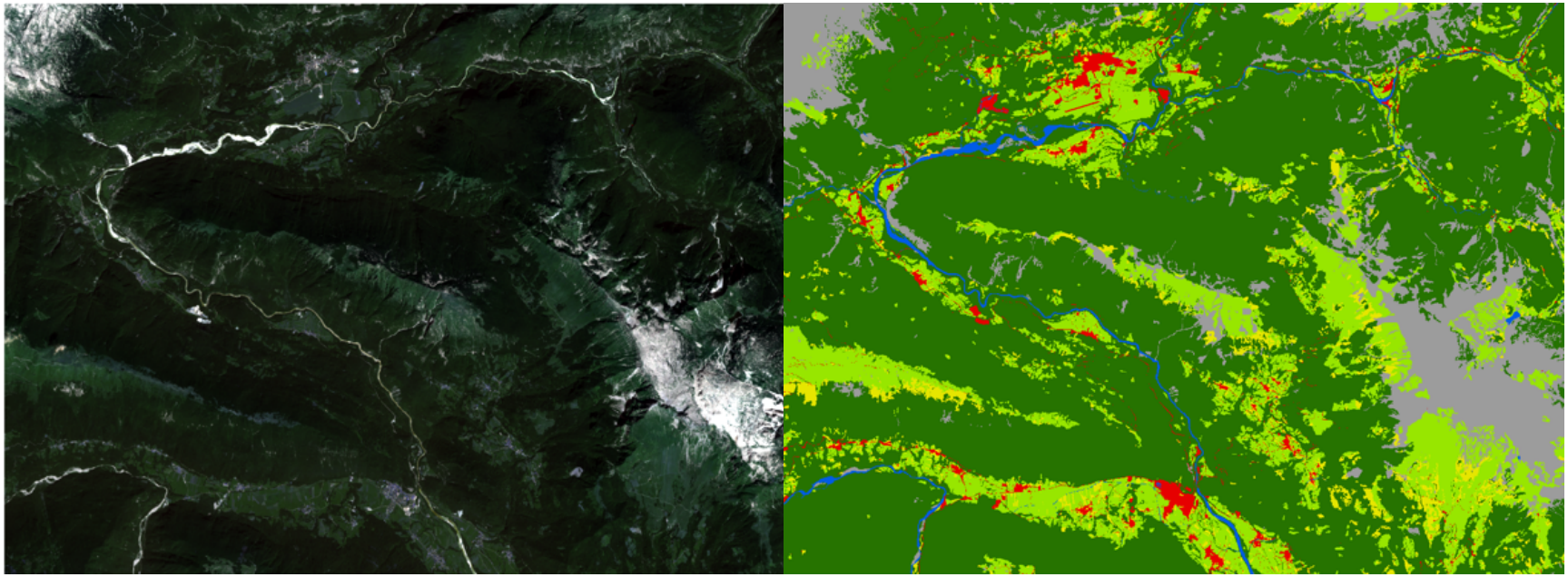
- Pokrovnost: prevladujoča raba zemljišča v pikslu
- 13 spektralnih kanalov (svetlobnih frekvenc)
- 34 indikatorjev izračunanih iz kanalov

POKROV: Napovedne spremenljivke (spektralni kanali)



↑ Spatial resolution versus wavelength: Sentinel-2's span of 13 spectral bands, from the visible and the near-infrared to the shortwave infrared at different spatial resolutions ranging from 10 to 60 m on the ground, takes land monitoring to an unprecedented level

POKROV: Testno območje je predalpski svet



Kriteriji za ocenjevanje: POKROV

1. Čim manjša napaka na testnih podatkih
2. Diskusija relevantnosti napovednih spremenljivk

Pozor

Izbor relevantnih spremenljivk, transformacija podatkov ter prostorska in časovna komponenta.

Kratek opis: PARLA

Stopnja zahtevnosti

4: za **pogumne** eksperte (15 bonus točk)

Področje

Prebeg poslancev in druge politične igre v državnem zboru RS.

Koraki procesa *KDD*

1 (Izbira podatkov), 2 (Pred-obdelava podatkov),
3 (Transformacija podatkov) in 4 (Podatkovno rudarjenje)

Opis podatkovnega vira: PARLA

parlameter.si

Osem glavnih entitet (povezave med njimi v oklepajih)

1. Poslanke in poslanci (povezave s 4)
2. Poslanske skupine (povezave z 1)
3. Delovna telesa (povezave z 1)
4. Seje državnega zbora in delovnih teles (povezave s 3)
5. Govori poslank in poslancev (povezave z 1 in 4)
6. Glasovanja (povezave z 1 in 4)
7. Zakoni in amandmaji (povezave s 6)
8. Poslanska vprašanja (povezave z 1)

Cilji podatkovne analize: PARLA

Raziskovalna/eksplorativna analiza

- Trendi političnega diskurza: katere besede in besedne zveze prevladujejo v govorih poslancev in kako se te razširjajo med poslanci
- Sponzorstvo zakonov: kdo med poslanci, poslanskimi skupinami in vladajoče večine oziroma opozicije podpira posamezne zakone oziroma zakone z določenega področja
- Usklajenost poslanskih skupin: koliko poenoteno glasujejo poslanske skupine o zakonih, ali obstajajo podobni poslanci iz različnih skupin

Napovedno modeliranje

- Napovedovanje prehoda poslancev med poslanskimi skupinami oz. med vladajočo večino in opozicijo
- Napovedovanje volilnih rezultatov (potem, ko bodo ti na voljo)

Kriteriji za ocenjevanje: PARLA

1. Kreativnost in inovativnost analize
2. Čim manjša napaka napovednih modelov

Pozor

Ni pripravljenih podatkov, le dostop do vira,
različni tipi podatkov, tudi besedila, redki prebegi.

Kratek opis: IZBIRA

Stopnja zahtevnosti

Po analogiji s štirimi nalogami opisanimi zgoraj.

Področje

Pripravite kratek opis področja oz. problema za podatkovno rudarjenje.

Koraki procesa *KDD*

Vsaj dva: 3 (Transformacija podatkov) in 4 (Podatkovno rudarjenje)