Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva

Diplomski studij računarstva, 2. semestar, izborni predmet

Ak. god. 2023./2024.

Dubinska analiza podataka

1. predavanje

Nastavnici na predmetu

Nositelj:

izv. prof. dr. sc. **Alan Jović** – ZEMRIS, D-318, alan.jovic@fer.hr

Asistenti:

dr. sc. **Igor Stančin** – ZEMRIS, vanjski suradnik, igor.stancin@fer.hr

Eda Jovičić, mag. ing. – ZEMRIS, vanjska suradnica, eda.jovicic@fer.hr

O predmetu

Sve obavijesti, materijali i ostale informacije o predmetu dostupne su na web stranici:

https://www.fer.unizg.hr/predmet/dap

Izborni predmet profila:

Računarska znanost

Znanost o podacima

Programsko inženjerstvo i informacijski sustavi

ECTS: 5

Broj studenata (max): 60

Nastavne teme – po tjednima

1. tjedan: Administracija predmeta. Uvod u dubinsku analizu podataka. Opis opsega područja.

2. tjedan: Priprema podataka za dubinsku analizu.

3. tjedan: Transformacije podataka i inženjerstvo značajki.

4. tjedan: Odabir značajki.

5. tjedan: Nebalansirani podaci, pomak koncepta u podacima.

6. tjedan: Metode ansambala i postupci tumačenja modela.

7. tjedan: Interpretabilni modeli: sustavi zasnovani na induktivnim pravilima i optimalna stabla odluke.

8a. – 8b. tjedan: ----međuispiti – ne na predmetu

Nastavne teme – po tjednima

9. tjedan: Pronalaženje čestih obrazaca i asocijativna pravila. Sustavi preporučivanja.

10. tjedan: Dubinska analiza vremenskih nizova.

***11. tjedan:** Duboko učenje u dubinskoj analizi podataka.

12. tjedan: --- predavanje se gubi zbog praznika

13. tjedan: Arhitekture dubokog učenja u područjima primjene: vremenski nizovi, slike, tekst, biomedicina

14. tjedan: Prezentacije projekata.

15. tjedan: Završni ispit

- 12. tjedan – predavanje 23. 5. 2024. će se održati u nekom drugom terminu (bit će u obavijesti)

Nastavne obveze na predmetu

- **Predavanja (3 sata tjedno, 7+6 tjedana)**
- **Laboratorijske vježbe (18 sati ukupno)**
- **Projekt (tijekom cijelog semestra) – 60 bodova**
- **Završni ispit (15. tjedan) – 40 bodova**
- **Ispiti na rokovima – 40 bodova**

- **Pragovi za ocjene: 50 (2) – 63 (3) – 75 (4) – 88 (5)**

Predavanja i laboratorijske vježbe

- Predavanja
  - **Četvrtkom od 13 do 16h, D-273**
  - Prisutnost i aktivnost na predavanjima se dodatno boduju: prisutnost do **4** boda, aktivnost do **3** boda
  - Online putem MS Teamsa u ovisnosti o epidemiološkoj situaciji

- Laboratorijske vježbe
  - U okviru **projekta** – pregled i ocjenjivanje obaveznih i dodatnih bilježnica
  - Tjedni pregleda i ocjenjivanja su poznati unaprijed

Projekt

- Nosi **60** bodova, prag za prolaz je **24** boda (40%)
- Traje tijekom čitavog semestra, od 2. do 13. tjedna
- Radi se individualno
- Provedba svih komponenti projekta radi se putem platforme Kaggle
- Komponente projekta:
  - Kompetitivno natjecanje
  - Obavezne bilježnice
  - Dodatne bilježnice
  - Recenzije tuđih obaveznih bilježnica

Detaljnije o projektu

- Putem ovog linka trebate se prijaviti na natjecanje
  - https://www.kaggle.com/t/788af1f078d148e5ac07fa13547e194a
- Sve bitne informacije o projektu nalaze se u ovoj prezentaciji i na Kaggleu
- Sva pitanja vezana uz projekt postavljate u *Discussion* na Kaggleu
  - Prije nego postavite novo pitanje provjerite je li netko drugi već pitao isto
- Sve kodove/bilježnice predajete na Kaggle
  - Na Kaggleu je točno definirano kada bilježnice trebate javno objaviti i kako trebate nazvati javne bilježnice
  - Nema nikakvih ograničenja na privatne bilježnice

Detaljnije o projektu

- Podatci predstavljaju kretanje cijena dionica kroz vrijeme
- Skup za učenje sastoji se od 2 pomoćna stupca (datum i ime dionice) i 6 značajki (opisane na Kaggleu).
- Potrebno je detektirati isplative dane za kupnju dionica. Ako cijena dionice u sljedećih dva mjeseca naraste barem 2%, smatra se da se kupnja isplati.
- Evaluacijska mjera je F1_mjera

Tijek projekta

<table>
<tr><th>Tjedan</th><th>Natjecanje</th><th>Obavezne bilježnice</th><th>Recenzije tuđih bilježnica</th><th>Dodatne bilježnice</th><th>Održavanje labosa</th></tr>
<tr><td>2. tjedan
3. tjedan
4. tjedan</td><td rowspan="9">Aktivno</td><td>Početni pregled podataka i referentni klasifikator (EDA)</td><td>-</td><td>Nisu dozvoljene</td><td>-</td></tr>
<tr><td>5. tjedan</td><td rowspan="3">Inženjerstvo značajki, modeli zasnovani na pravilima i ansambli klasifikatora (FErules)</td><td>Recenzija</td><td rowspan="3">Nisu dozvoljene bilježnice koje odgovaraju na pitanja postavljena u sklopu 2. i 3. obavezne bilježnice</td><td>Lab - EDA</td></tr>
<tr><td>6. tjedan
7. tjedan
MI
MI</td><td>-</td><td>-</td></tr>
<tr><td>9. tjedan</td><td rowspan="3">Analiza vremenskih serija (TS)</td><td>Recenzija</td><td rowspan="2">Nisu dozvoljene bilježnice koje odgovaraju na pitanja postavljena u sklopu 3. obavezne bilježnice</td><td>Lab - FErules</td></tr>
<tr><td>10. tjedan
11. tjedan</td><td>-</td><td>-</td></tr>
<tr><td>12. tjedan</td><td rowspan="3">Nisu dozvoljene</td><td>Lab - dodatne bilježnice</td></tr>
<tr><td>13. tjedan</td><td>-</td><td>-</td><td>Recenzija</td><td>Lab - TS</td></tr>
<tr><td>14. tjedan</td><td>-</td><td>-</td><td>-</td><td>-</td></tr>
<tr><td>ZI</td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

Detaljnije o projektu

- Bodovanje
  - Maksimalan broj bodova na projektu – **60**
  - Maksimalan broj bodova po komponentama projekta

Komponente projekta	Maksimalan broj bodova
Rang lista	30
Tri obavezne bilježnice	36
Recenzija tuđih bilježnica	6
Tuđe recenzije vlastite bilježnice	6
Dodatne bilježnice	12

Detaljnije o projektu

- Natjecanje
  - Pravila i rokovi objavljeni na Kaggleu.
- Bodovanje

Poredak na natjecanju	Broj bodova
1. mjesto	30
2. mjesto	24
3. mjesto	18
4.-7. mjesto	12
7.-15. mjesto	6
>15. mjesta	0

Detaljnije o projektu

- Obavezne bilježnice
  - Tri bilježnice tijekom semestra
  - Teme, zadatci, rokovi i imenovanje su definirani na Kaggleu
  - Recenzije se pišu kao komentar na tuđe bilježnice, a primjer recenzije je vidljiv na Kaggleu

Komponente	Maksimalan broj bodova	Način bodovanja
Bilježnica	12	Dodjeljuje asistent
Recenzirana tuđa bilježnica 1	1	Kvalitetno napravljena recenzija – 1 bod Loše napravljena recenzija – 0 bodova Recenzija nije napravljena – 0 bodova
Recenzirana tuđa bilježnica 2	1	
Tuđe recenzije vlastite bilježnice	2	U slučaju 2 recenzije – ((avg1 + avg2)/2)*2 boda U slučaju 1 recenzije – avg1*2 boda U slučaju 0 recenzija – postotak dodijeljen od asistenta*2 boda

Detaljnije o projektu

- Dodatne bilježnice
  - Temu dodatnih bilježnica sami određujete, a ideja je da objavite nešto čime ćete drugima pomoći da nešto naprave bolje/brže/efikasnije/kvalitetnije
  - Maksimalno dvije po studentu
  - Prije nego objavite provjerite (vidljivo u tijeku projekta, slajd 12) je li u trenutnom tjednu dozvoljena ta tema bilježnice
  - Broj bodova po bilježnici nije ograničen, tj. jedna bilježnica može nositi maksimalnih 12 bodova.
  - Bilježnica mora sadržavati motivaciju/hipotezu, eksperiment/sadržaj, mjerljive rezultate (npr. vrijeme, točnost, F1, zauzeće memorije, linije koda, …) i zaključke

Komponenta	Maksimalan broj bodova	Način bodovanja
Dodatna bilježnica	12	Procjena asistenta. Dozvoljene su maksimalno dvije dodatne bilježnice. Nema ograničenja broja bodova po bilježnici, odnosno jedna bilježnica može nositi svih 12 bodova.

Detaljnije o projektu

- Pravila
  - Radi se **samostalno (individualno)**
  - Dozvoljeno je komentirati projekt s kolegama, ali nije dozvoljeno podijeliti kodove/bilježnice s drugima
  - Sve bilježnice bit će provjerene detektorom plagijata. Plagirana bilježnica bit će ocijenjena s nula bodova, studenti koji plagiraju bilježnicu neće moći ostvariti bodove temeljem ranga na natjecanju i dobit će dodatnih 10 negativnih bodova. Također, nastavnici će razmotriti prijavu plagijata studentskoj disciplinskoj komisiji

Završni ispit

- Nosi **40** bodova, prag za prolaz je **16** bodova (40%)
- Uvjet za pristup je **položeni projekt**
- Obuhvaća gradivo čitavog predmeta

- Fokus na razumijevanju gradiva proučavanog na predavanjima
  - Teorijska pitanja s jezgrovitim, opisnim odgovorom
- Popis tema iz kojih se sastavljaju pitanja za završni ispit kao i ogledni primjer ispita bit će dostupni studentima tijekom semestra

Ispiti na rokovima

- Jednaki način polaganja i bodovanje kao i završni ispit
- Uvjet za pristup je **položeni projekt**

Preporučena literatura

- Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques. 4th ed. Morgan Kaufmann, 2016.
- Fuernkranz J, Gamberger D, Lavrač N. Foundations of Rule Learning. Heidelberg : Springer, 2012
- Molnar C, Interpretable Machine Learning, Leanpub, 2020
- James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. Springer, 2014.
- Raschka S, Mirjalili V. Python Machine Learning. 2nd ed. Packt Publishing, Birmingham UK, 2017.
- Kelleher, JD, Duboko učenje, The MIT Essential Knowledge Series, Mate d.o.o., 2021, https://www.mate.hr/product/2310/duboko-ucenje
- Simon J.D. Prince, Understanding Deep Learning, MIT Press, December 2023, https://udlbook.github.io/udlbook/

- **Sve potrebno za položiti predmet obrađuje se na predavanjima i u okviru projekta**

Opis područja

Što je dubinska analiza podataka?

- **Proces pronalaženja skrivenih i korisnih informacija u podacima.**

- *Wikipedia*: “Data mining is a process of extracting and discovering patterns in large data sets”
- *SAS*: “Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes.”
- *Witten, Frank, Hall, Pal*: “Data mining is the extraction of implicit, previously unknown, and potentially useful information from data.”

- Temeljni pojmovi: **pronalaženje** (engl. *extraction, mining*), **obrazac** (engl. *pattern*), **skup podataka** (engl. *data set*)

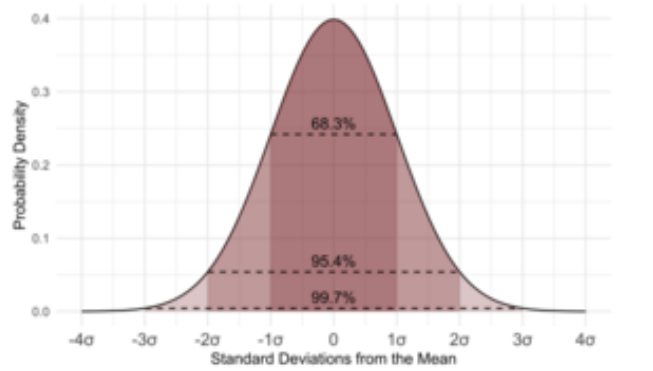
Što je dubinska analiza podataka?

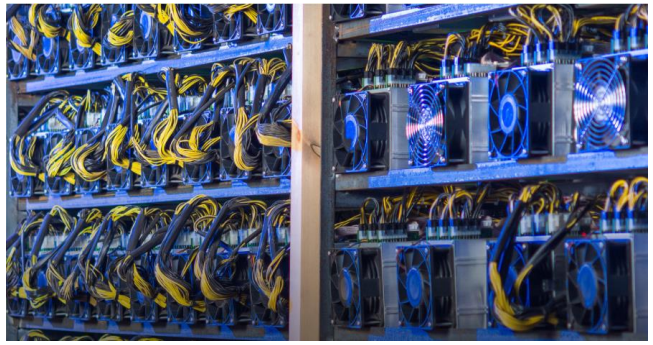
- **Dubinska analiza podataka = primijenjeno strojno učenje**
  - Praktični aspekti primjene metoda strojnog učenja na različite probleme
  - Uzima u obzir značajke skupa podataka i domene od interesa
  - Naglasak nije na algoritmima strojnog učenja nego na njihovoj učinkovitoj primjeni
- Područje omeđeno **strojnim učenjem**, **statistikom** i **bazama podataka**

- Slični pojmovi
  - **Podatkovna analitika**
  - **Znanost o podacima**
  - **Analiza velikih skupova podataka**

Što je dubinska analiza podataka?

- Načini pronalaska korisnih informacija u podacima
  - **Ručni pregled**
  - **Statistika**
  - **Računalna automatizacija otkrivanja znanja**
    - **Metode strojnog učenja** – izgradnja modela, učenje na temelju podataka
    - **Ostale metode** – pronalaženje čestih obrazaca (engl. *frequent pattern mining*), optimizacijski postupci





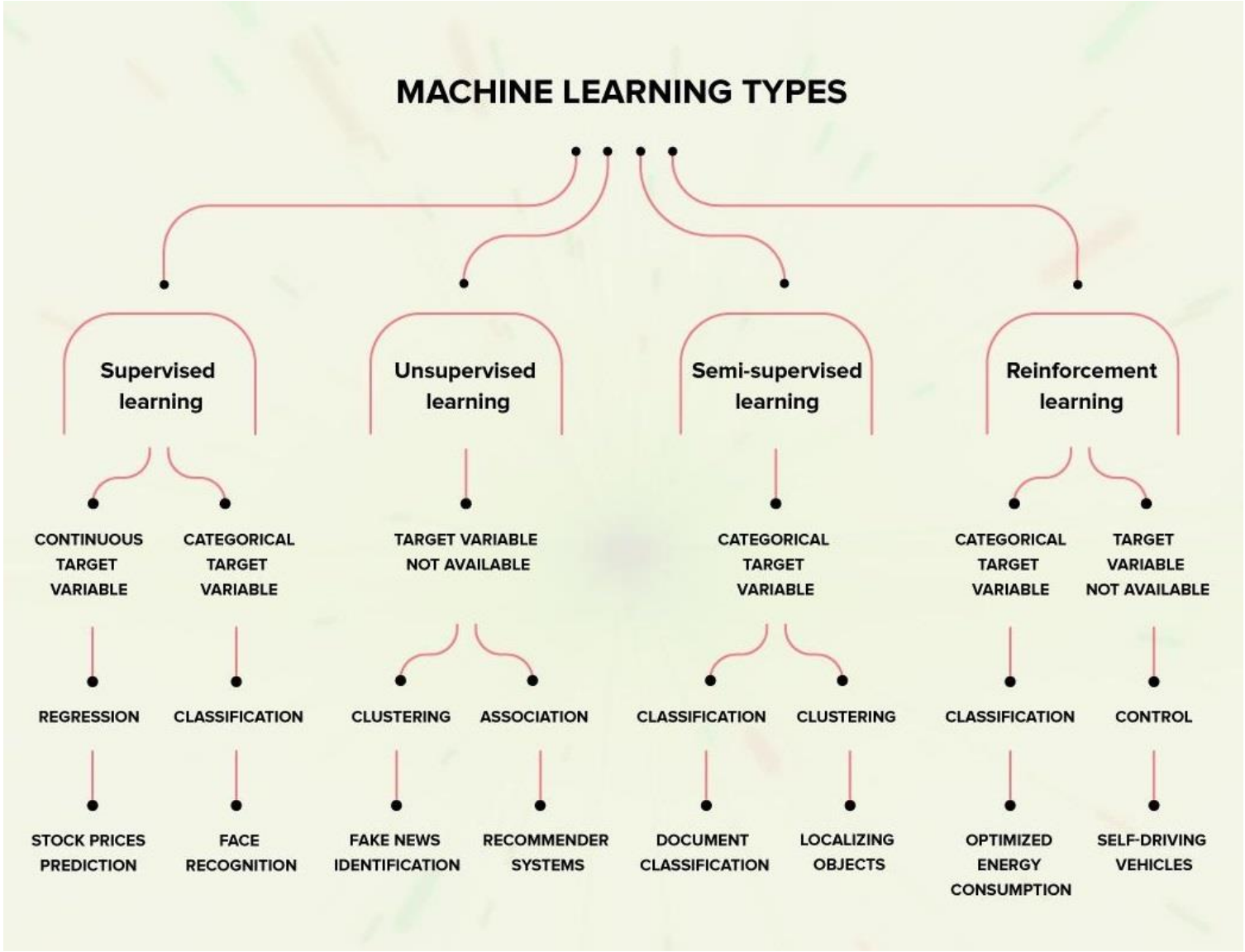
Klasifikacija vrsta strojnog učenja

Četiri vrste strojnog učenja:

Opcije za ciljnu varijablu:

Vrsta rješavanog zadatka:

Primjer primjene:



Prilagođeno od: https://serokell.io/blog/how-to-choose-ml-technique

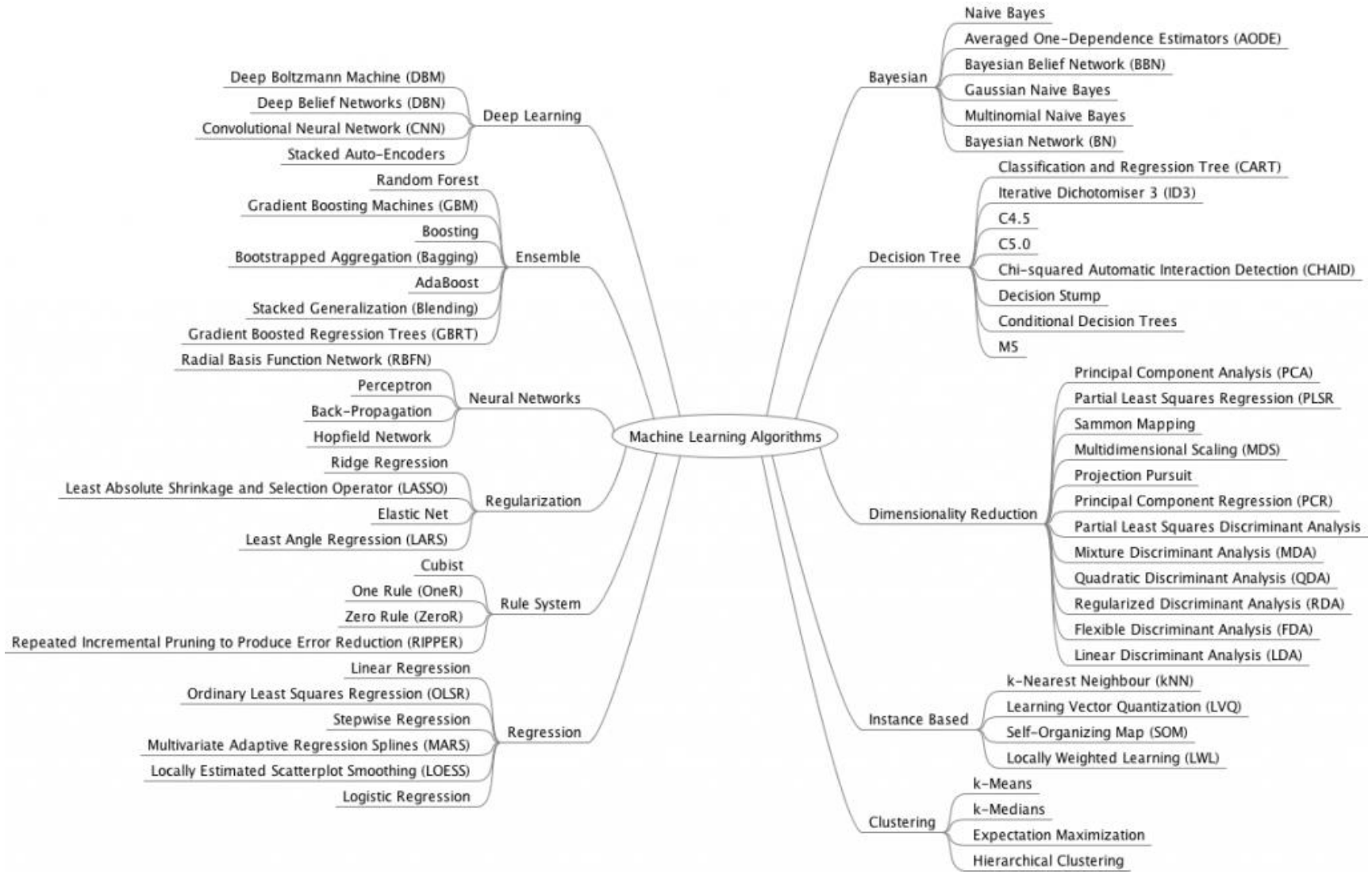
Klasifikacija metoda strojnog učenja

Preuzeto s: https://www.kaggle.com/getting-started/153090

Autor: Jason Brownlee

Vidjeti i članak:

- *Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, Dinani Amorim.* **Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?** *JMLR* 15(90):3133−3181, 2014. https://jmlr.org/papers/volume15/delgado14a/delgado14a.pdf



Šest glavnih zadataka od interesa za dubinsku analizu podataka*

- **Otkrivanje anomalija** (engl. *anomaly detection, change and deviation detection*) – zadatak identifikacije neobičnih, potencijalno zanimljivih podataka
- **Modeliranje ovisnosti** (engl. *dependency modeling, conditional dependency modeling, association rules learning, market basket analysis*) – zadatak traženja odnosa između više varijabli
- **Grupiranje** (engl. *clustering*) – zadatak otkrivanja grupa sličnih podataka
- **Klasifikacija** (engl. *classification*) – zadatak generaliziranja poznate strukture za primjenu na nove podatke
- **Regresija** (engl. *regression*) – zadatak pronalaska funkcije koja modelira ciljne numeričke podatke s najmanjom pogreškom
- **Sažimanje** (engl. *summary*) – zadatak pružanja kompaktne reprezentacije skupa podataka, najčešće u vidu vizualizacije ili izvještaja

* prema: Fayyad U, Piatetsky-Shapiro G, Smyth P (1996). "From Data Mining to Knowledge Discovery in Databases", AI MAGAZINE, pp. 37-54.

Ostali zadaci od interesa za dubinsku analizu podataka

- **Odabir značajki** (engl. *feature selection*) **–** zadatak pronalaska najinformativnijih značajki nekog problema
- **Učenje iz tokova podataka** (engl. *online learning*, *online machine learning*, *data stream mining*) – zadatak pronalaska zanimljivih informacija u podacima koji slijedno pristižu
- **Otkrivanje podgrupa** (engl. *subgroup discovery*) – zadatak pronalaska zanimljivih podskupova podataka u velikim skupovima podataka
- **Učenje strukturiranog izlaza** (engl. *structured output learning*) – pronalaženje modela složenijeg ciljnog podatka, npr. označavanje više ciljnih vrijednosti (engl. *multilabeling*)

Dubinska analiza podataka uključuje pripremu podataka

- **Priprema podataka je iznimno važan korak za uspješnu analizu podataka**
- **Priprema je više od pola posla!**

- Uključuje sve korake od pregleda podataka do predstavljanja pripremljenog skupa postupcima za analizu podataka
- Koraci koji se poduzimaju djelomično ovise o problemu koji se rješava
- Tema 2. – 5. predavanja

Pregled područja od interesa za dubinsku analizu podataka iz perspektive vrsta podataka

- Analiza strukturiranih (tabličnih) podataka (engl. *tabular data analysis*)
- Analiza vremenskih nizova podataka (engl. *time series analysis*)
- Analiza teksta (engl. *text analysis*)
- Analiza slike (engl. *image analysis*)
- Analiza videa (engl. *video analysis*, *content discovery*)
- Analiza sekvenci (engl. *sequence analysis*) – DNA, RNA, peptidi...

Alati za dubinsku analizu podataka

- Danas je dostupno više stotina alata za dubinsku analizu podataka
- Podjela na besplatne (možda i otvorenog koda) i komercijalne alate (**IBM SPSS Modeler**, **SAS Enterprise Miner**, **Oracle Data Mining**)
- Podjela na opće (**Weka**, **RStudio**, **Orange**) i specijalizirane alate
  - Specijalizacija na pojedini algoritam (**Random Forest**, **XGBoost**) ili grupu algoritama (**ELKI**, **Keras**), specijalizacija na pojedino primjensko područje (**NLTK**, **MOA**)
- Podjela na:
  - programske jezike (**Python**, **R**, **Java**)
  - programske knjižnice (**scikit-learn, TensorFlow, libsvm**) i
  - radna okruženja (**Matlab**, **RStudio**, **Anaconda**, **RapidMiner**, **Apache Spark**)

Alati za dubinsku analizu podataka

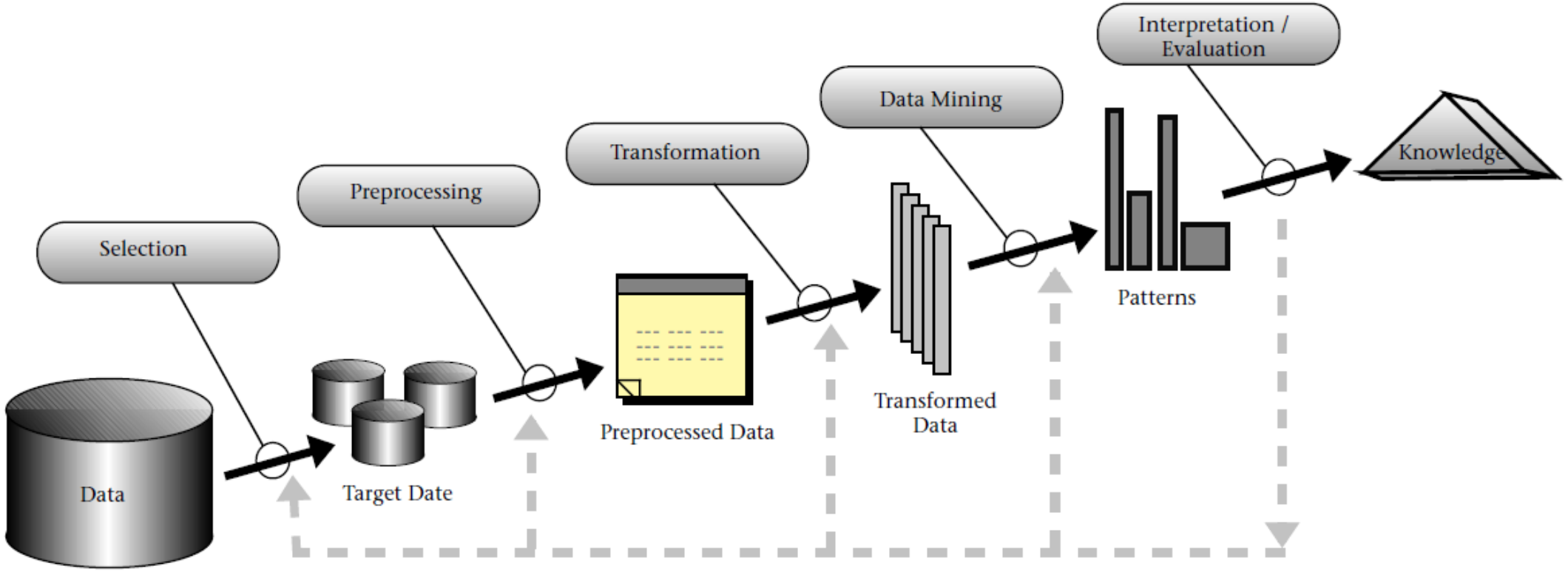
- DAP se provodi u velikom broju alata pisanih u raznim programskim jezicima
- Trenutačno najpopularniji jezici za DAP su **Python, R** i **Java (tim redom)**
- DAP je dosta **komercijalno orijentiran**
- Alati za DAP se stalno mijenjaju, nadograđuju ili nestaju
- **Znanstveni DAP nije isto što i komercijalni DAP**
- Za uspješno razumijevanje DAP-a potrebno je savladati barem nekoliko alata i programskih jezika
- Potrebno je poznavati baze podataka (SQL, NoSQL)

Modeli procesa dubinske analize podataka

Modeli procesa dubinske analize podataka

- Inicijativa za standardizacijom dubinske analize podataka
  - Nedovoljno specificirani proces pronalaženje korisnih informacija
  - I znanost i industrija zahtijevaju organiziraniji pristup!
  - Od sredine 1990-ih do danas
- Glavni modeli procesa
  - **KDD** (engl. *Knowledge Discovery in Databases*) – 1996.
  - **CRISP-DM** (engl. *CRoss-Industry Standard Process for Data Mining*) – 2000.
  - **ASUM-DM** (engl. *Analytics Solutions Unified Method*) – 2015.
  - **CRISP-ML(Q)** (engl. *CRoss-Industry Standard Process model for the development of Machine Learning applications with Quality assurance methodology*) – 2021.

KDD



Fayyad U, Piatetsky-Shapiro G, Smyth P (1996). "From Data Mining to Knowledge Discovery in Databases", AI MAGAZINE, pp. 37-54.

KDD

- Znanstveni pristup otkrivanju znanja u bazama podataka
- ***Data*** (podaci) – skup činjenica, najčešće prikazane kao retci u tablici u bazi
- ***Pattern*** (uzorak) – izraz u nekom jeziku koji opisuje podskup podataka ili model primijenjiv na taj podskup
  - Pronalaženje uzorka je uočavanje ponavljajuće strukture u podskupu podataka
  - Otkriveni uzorak mora biti uočljiv i na novom skupu podataka iz iste razdiobe s nekom sigurnošću
- Koncept zanimljivosti (***Interestingness***) uzorka – mjera vrijednosti uzorka koja uključuje uočljivost, novost, korisnost i jednostavnost
- *Pattern* postaje ***Knowledge*** (znanje) ako prelazi neki definirani prag zanimljivosti – subjektivno!
- ***Data Mining*** – korak KDD-a koji primjenom algoritama analize podataka uz ograničene računalne resursi dovodi do pobrojavanja uzoraka u podacima

CRISP-DM

- Industrijski pristup standardizaciji projekata dubinske analize podataka (1996.+)*
- Proces razložen u šest faza: **razumijevanje poslovnih potreba, razumijevanje podataka, priprema podataka, modeliranje, vrednovanje i puštanje u pogon**
- Faze se načelno provode **slijedno**, ali postoji mogućnost povratka na prethodnu
- Proces je **iterativan** u smislu da iskustvo iz prethodnih iteracija utječe na iduće
- Kompatibilan s UP i agilnim modelima procesa programskog inženjerstva

* Vidjeti: Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R (2000); *The CRISP-DM User Guide,* http://www.statoo.com/CRISP-DM.pdf

CRISP-DM

- **Glavna prednost: neovisan o industriji, alatima i metodama** koji se koriste za DAP
- **Bitan nedostatak:** ne uključuje aktivnosti upravljanja projektom
- Jednog od osnivača, tvrtku ISL je kasnije kupio SPSS, koji je kasnije kupio IBM -> **ASUM DM**
- Postoje specifičniji modeli vezani uz razvoj **prediktivnih modela**

Prikaz povezanosti faza CRISP-DM modela

CRISP-DM

- **Faze** se ostvaruju pomoću jednog ili više **generičkih zadataka**
- **Generički zadatci** (engl. *generic task*)
  - pristupi koji pokrivaju veći broj izazova dubinske analize, npr. čišćenje podataka (engl. *clean data*)
- **Specijalizirani zadatci** (engl. *specialized task*)
  - Specifično izvođenje određenog generičkog zadatak, npr. čišćenje kategoričkih podataka ili grupiranje podataka
- **Instanca procesa** (engl. *process instance*)
  - zapis **stvarno** provedenih akcija, odluka i rezultata, organiziran po specijaliziranim zadacima

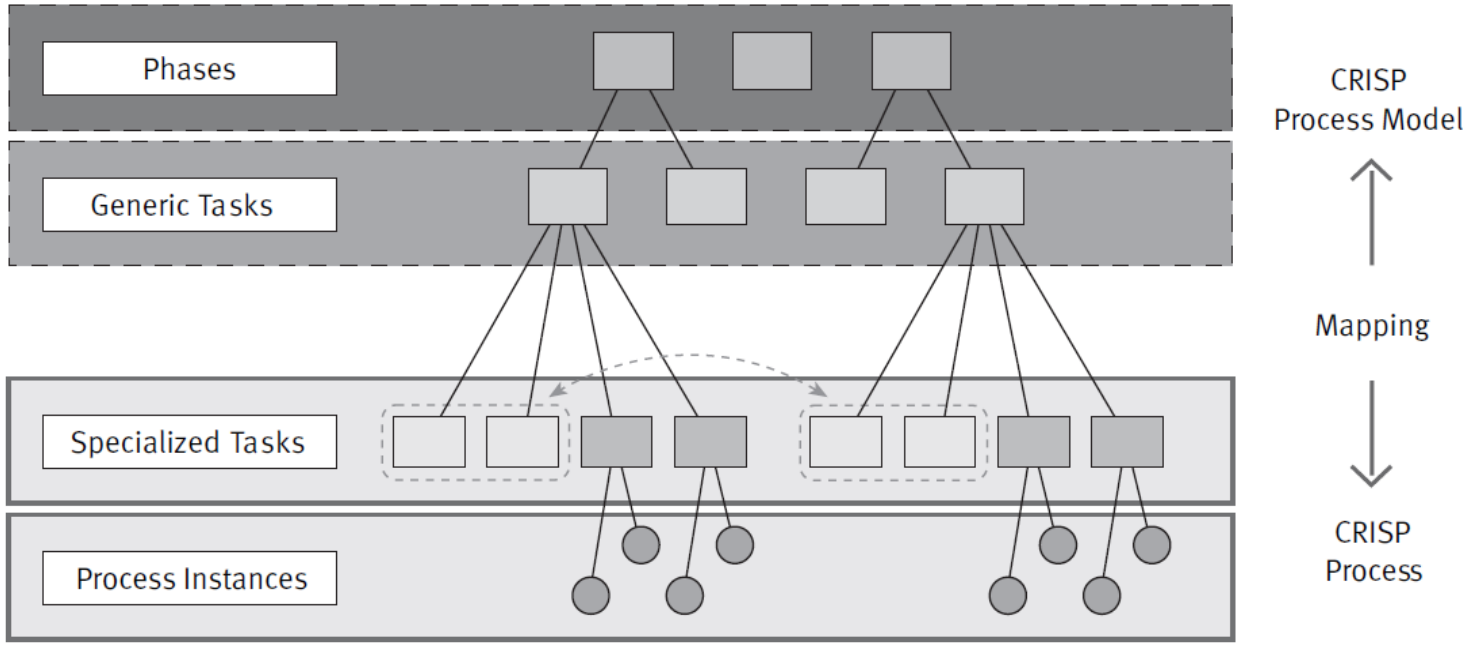
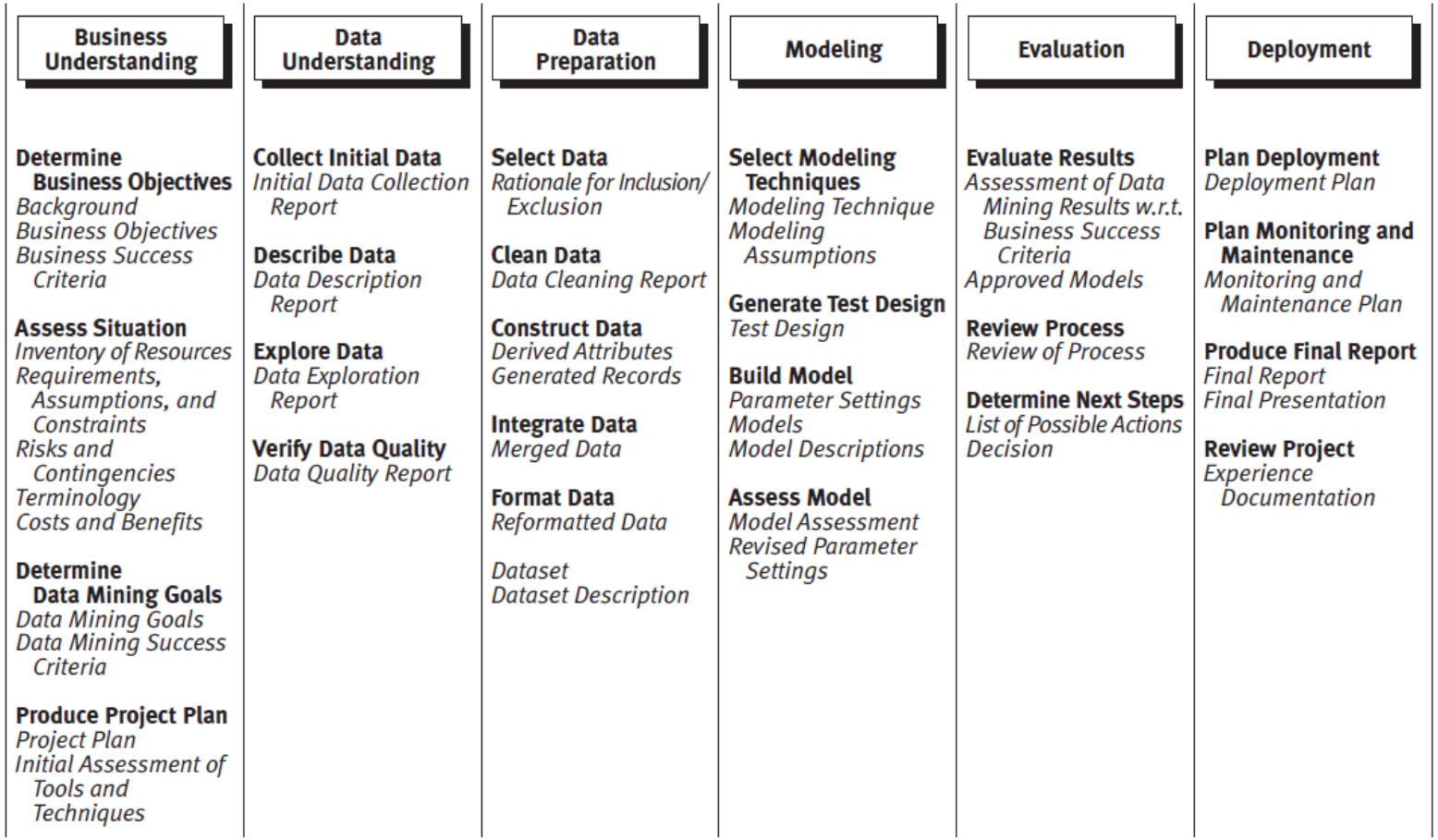


Figure 1: Four level breakdown of the CRISP-DM methodology

CRISP-DM



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria*	**Collect Initial Data** *Initial Data Collection Report*	**Select Data** *Rationale for Inclusion/ Exclusion*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models*	**Plan Deployment** *Deployment Plan*
Assess Situation *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits*	**Describe Data** *Data Description Report*	**Clean Data** *Data Cleaning Report*	**Generate Test Design** *Test Design*	**Review Process** *Review of Process*	**Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan*
Determine Data Mining Goals *Data Mining Goals* *Data Mining Success Criteria*	**Explore Data** *Data Exploration Report*	**Construct Data** *Derived Attributes* *Generated Records*	**Build Model** *Parameter Settings* *Models* *Model Descriptions*	**Determine Next Steps** *List of Possible Actions* *Decision*	**Produce Final Report** *Final Report* *Final Presentation*
Produce Project Plan *Project Plan* *Initial Assessment of Tools and Techniques*	**Verify Data Quality** *Data Quality Report*	**Integrate Data** *Merged Data*	**Assess Model** *Model Assessment* *Revised Parameter Settings*		**Review Project** *Experience Documentation*
		Format Data *Reformatted Data*			
		Dataset *Dataset Description*			

Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

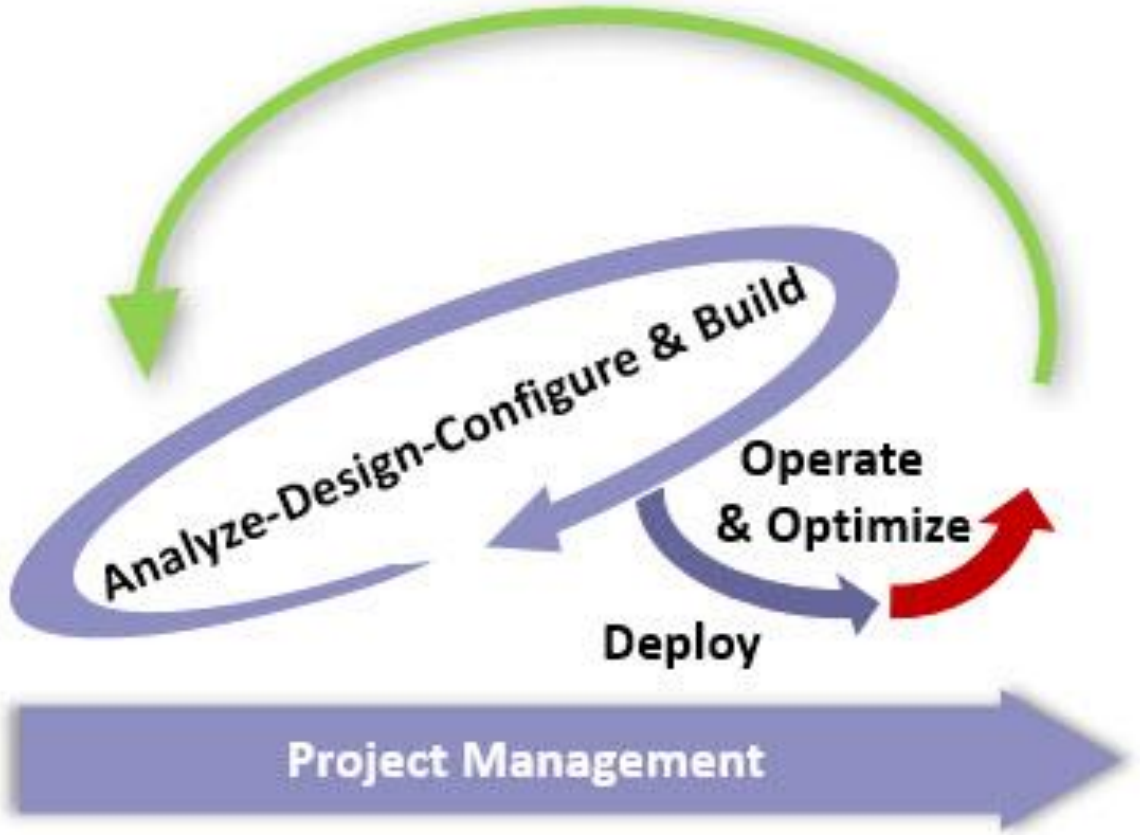
- Svaki od generičkih zadataka i njihovih izlaznih rezultata se opisuje detaljno u modelu CRISP-DM

ASUM-DM

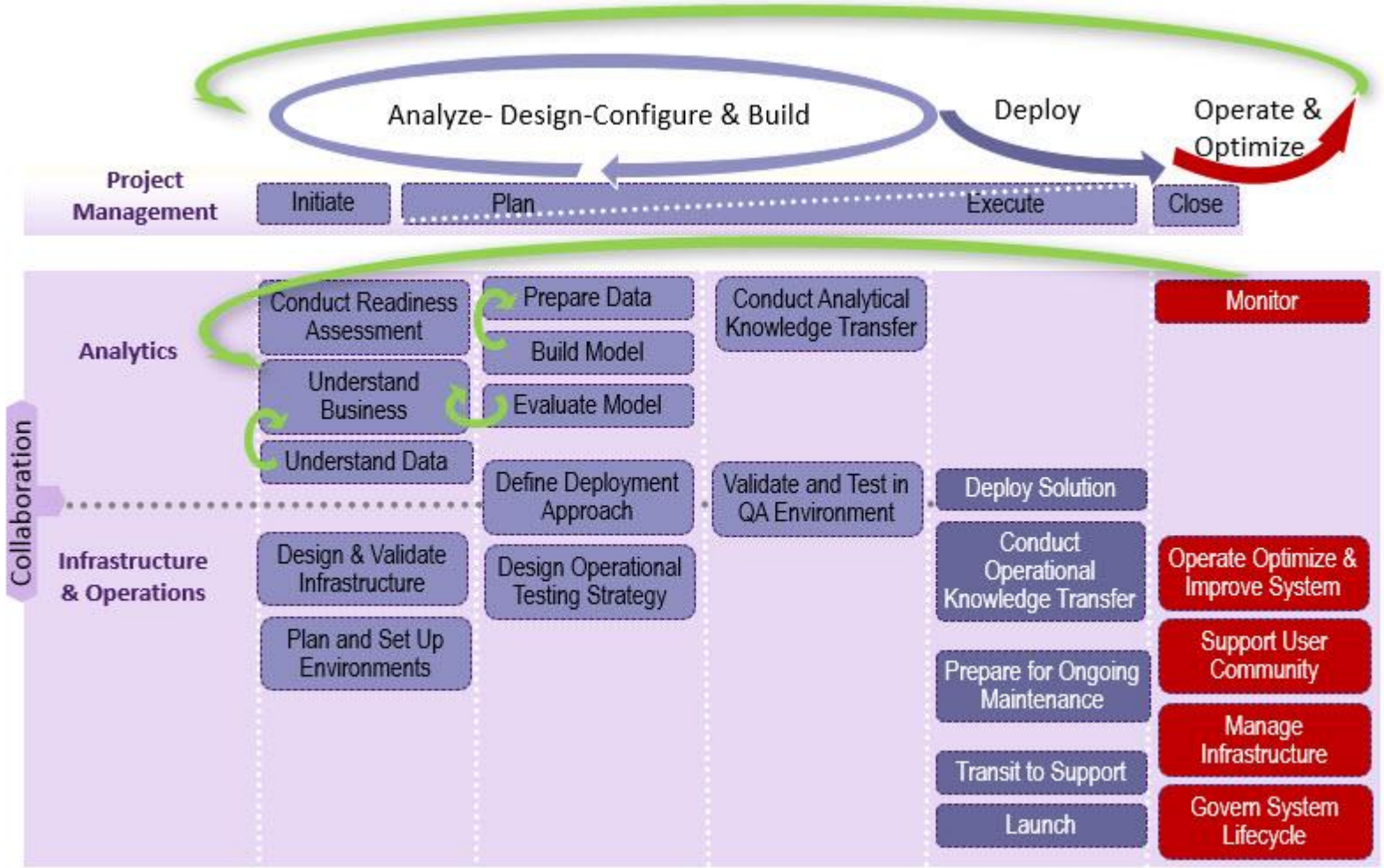
- *Analytics Solutions Unified Method*
- ASUM-DM je IBM-ovo proširenje metodologije CRISP-DM (2015.)
- ASUM-DM ima bolje razrađen operativni dio **puštanja u pogon**, a uključuje i poslovni aspekt **upravljanja projektom** u obzir
- Kod puštanja u pogon, nove značajke su dodane u model: **suradnja** (engl. *collaboration*), **kontrola verzija koda** (engl. *version control*), **sigurnost** (engl. *security*) i **usklađenost s regulativama** (engl. *compliance*)
- http://gforge.icesi.edu.co/ASUM-DM_External/index.htm

ASUM-DM

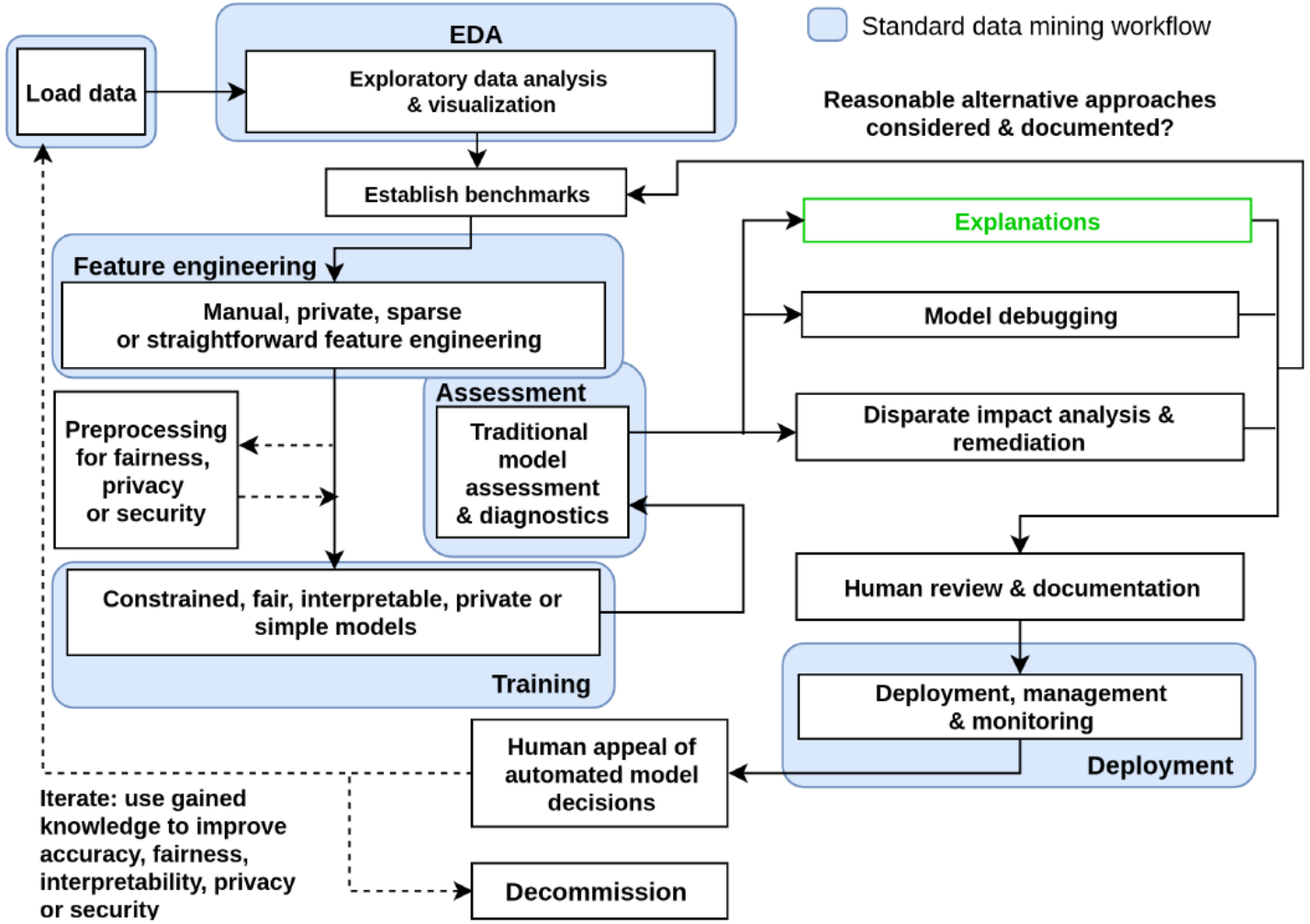
- **Pet faza**: analiza, oblikovanje, konfiguracija i izgradnja (modela), puštanje u pogon, djelovanje i optimizacija
- **Prve tri faze su združene** zbog iterativne prirode DAP projekata
- Upravljanje projektima je opcionalno



ASUM-DM – uključivanje suradnje na projektu

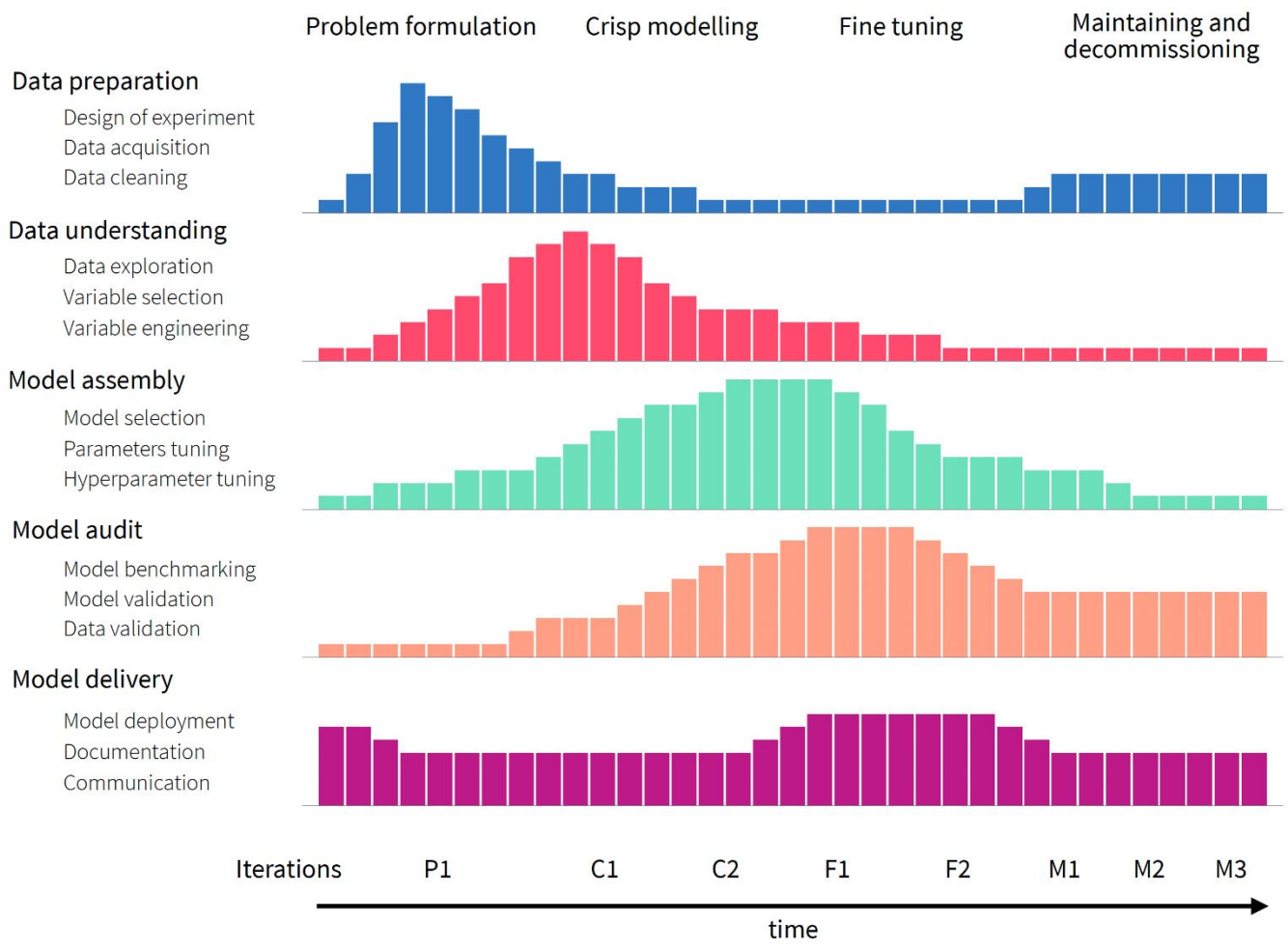


Modeli procesa s fokusom na strojno učenje



- Naglasak na **objašnjivosti** modela strojnog učenja
- **Hall P.**, On Explainable Machine Learning Misconceptions & A More Human-Centered Machine Learning, 2019
- https://github.com/jphall663/xai_misconceptions/blob/master/xai_misconceptions.pdf

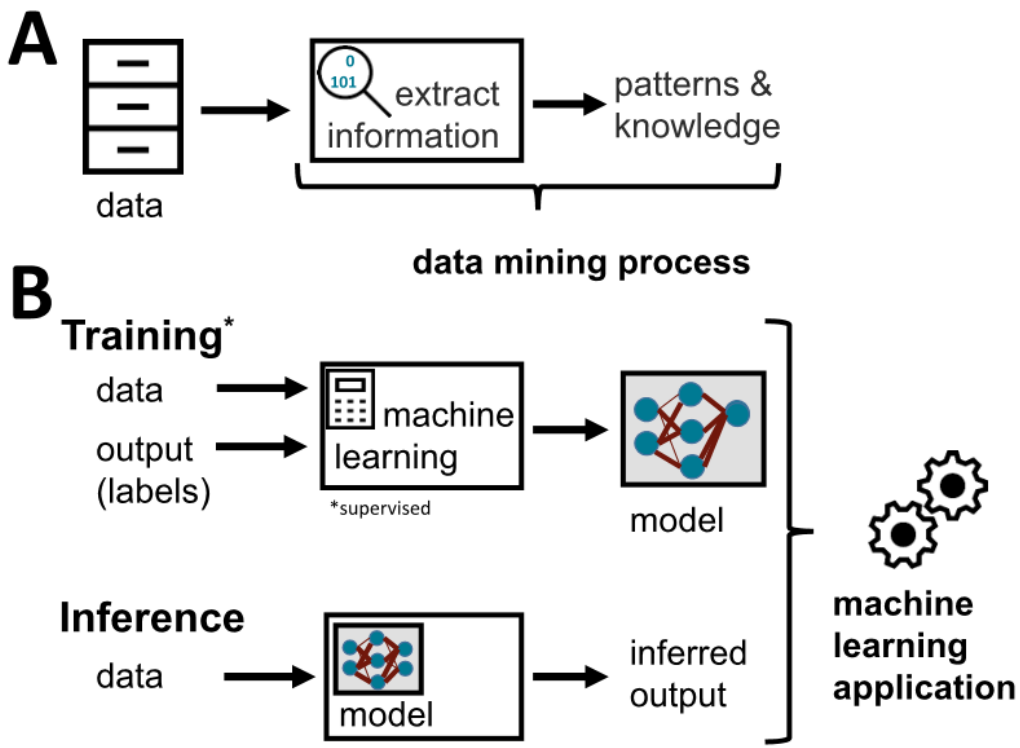
Model Development Process



- Razrada **životnog ciklusa** (engl. *lifecycle*) prediktivnog modela
- Analogan modelu unificiranog procesa (engl. *Unified Process*, UP) kod razvoja programske potpore
- **Biecek P.**, Model Development Process, 2019.
- https://ema.drwhy.ai/modelDevelopmentProcess.html
- Biecek P., Burzykowski T. Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models. With examples in R and Python, CRC Press, 2020.

CRISP-ML(Q)

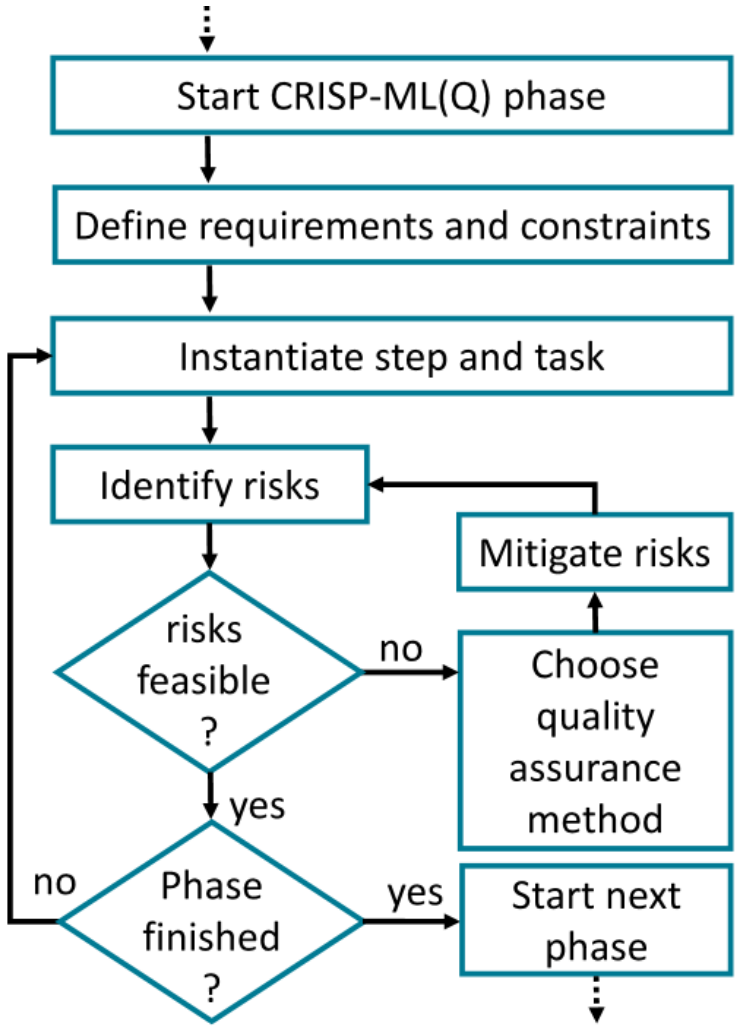
- Suvremeni model procesa namijenjen **razvoju modela strojnog učenja**, njegovom **puštanju u pogon** i **održavanju**
- Šest koraka (faza):
  - Razumijevanje poslovne strane i podataka
  - Priprema podataka
  - Modeliranje
  - Vrednovanje modela
  - Puštanje u pogon
  - **Nadzor i održavanje**



Studer, S.; Bui, T.B.; Drescher, C.; Hanuschkin, A.; Winkler, L.; Peters, S.; Müller, K.-R. Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. Mach. Learn. Knowl. Extr. 2021, 3, 392–413. https://doi.org/10.3390/make3020020

CRISP-ML(Q)

- Svaki korak ima **kontrolu kvalitete**
- **Nadzor i održavanje modela**
  - Novost u odnosu na CRISP-DM
  - Prate se performanse modela, model se treba kontinuirano osvježavati ako dođe do promjene
  - Najčešći uzroci smanjenjih performansi:
    - Distribucija podataka se promijenila (tzv. promjena koncepta)
    - Degradacija sklopovlja o kojoj ovisi model
    - Nadogradnja programske potpore



Studer, S.; Bui, T.B.; Drescher, C.; Hanuschkin, A.; Winkler, L.; Peters, S.; Müller, K.-R. Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. Mach. Learn. Knowl. Extr. 2021, 3, 392–413. https://doi.org/10.3390/make3020020

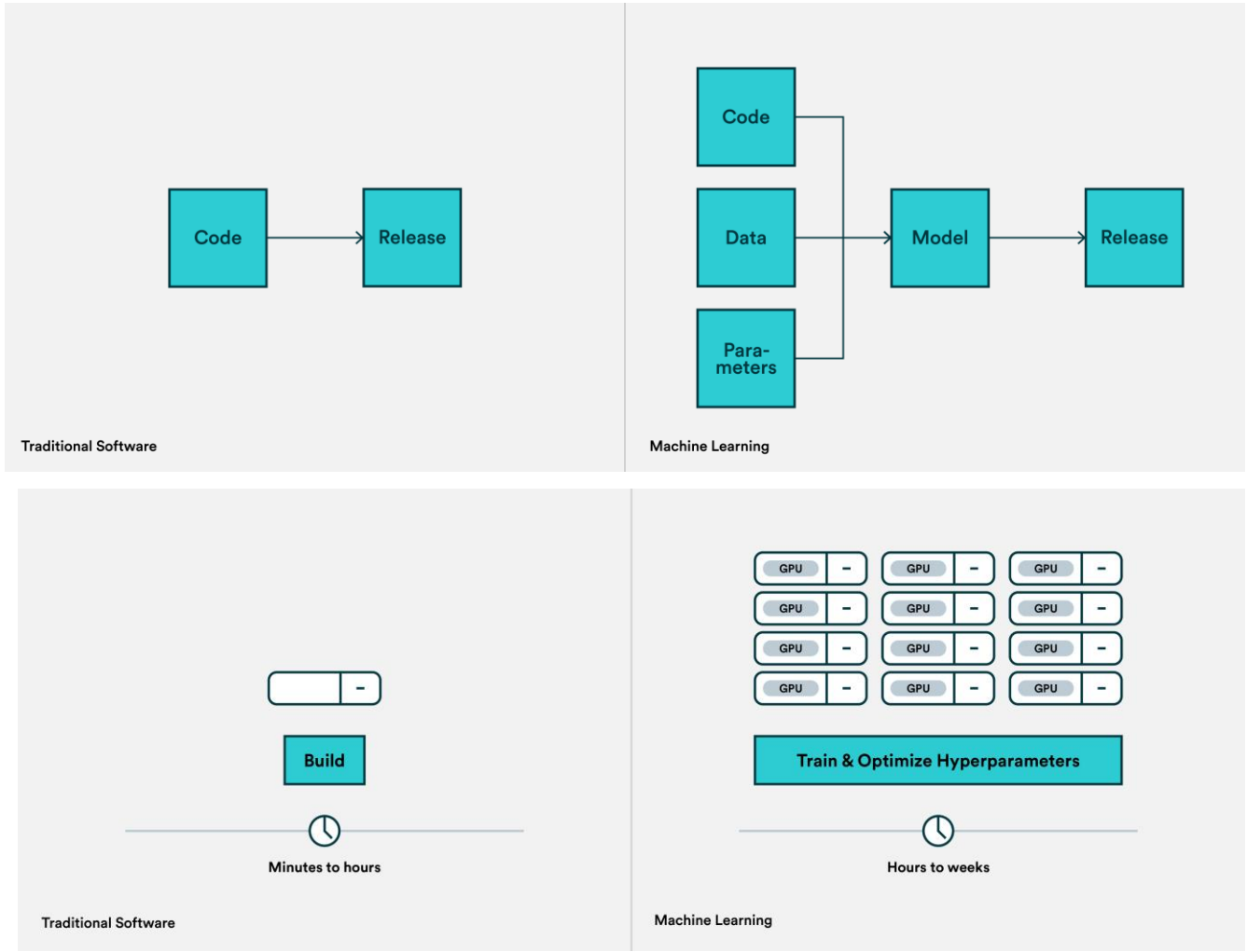
Platforme za automatizaciju izgradnje modela strojnog učenja

- Fokus na korisnike bez **naprednog** znanja o strojnom učenju i programiranju

- Google AutoML (dio Google Cloud Platforme, GCP) – više usluga, ovisno o podacima
  - https://cloud.google.com/automl
- H2O AutoML (dio H2O.ai) – fokus na optimizaciji modela neuronskih mreža
  - https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html
- DataRobot AI Cloud – komercijalna platforma za prediktivni i generativni AI
  - https://www.datarobot.com/

DevOps i MLOps

- DevOps: automatizacija procesa razvoja, ispitivanja i puštanja u pogon programske potpore
- MLOps: isto to, samo s modelima strojnog učenja
  - Izazovno zbog prirode problema
  - Uklapa se u postojeće modele procesa prog. inženjerstva (agilna metodologija)
  - Alati za kontrolu verzija koda su isti (npr. Git – Github, Gitlab)
  - Razni ostali alati ne moraju biti isti, npr. vidjeti: https://neptune.ai/blog/best-software-for-collaborating-on-machine-learning-projects



Preuzeto s: https://valohai.com/blog/difference-between-devops-and-mlops/

Što (uglavnom) ne razmatramo na predmetu

- Sustave za rad s velikim skupovima podataka (Apache Hadoop, Apache Spark, RDD, HDFS, NoSQL baze) i algoritme specifične za rad s takvim skupovima
- Paralelizaciju postupka analize podataka i alati za rad s tokovima podataka (Docker, Kubernetes, Swarm, Apache Kafka...)
- Dubinsku analizu podataka s weba (*web mining*)
- Grupiranje, podržano učenje
- …

Zaključak

- Dubinska analiza podataka koristi se za pronalaženje skrivenih informacija u podacima s ciljem dobivanja znanja
- Danas postoji razvijeno mnoštvo tehnologija i alata za otkrivanje informacija
- Razvijene su standardne metodologije za znanstveno i industrijsko otkrivanje znanja koje se i dalje razvijaju
- Veliki naglasak kod dubinske analize podataka je na poslovnoj primjeni
- Predmet je praktično orijentiran, prema stvarnim problemima i pristupima analizi podataka, a projekt na predmetu potiče kompetitivnost i inovativnost