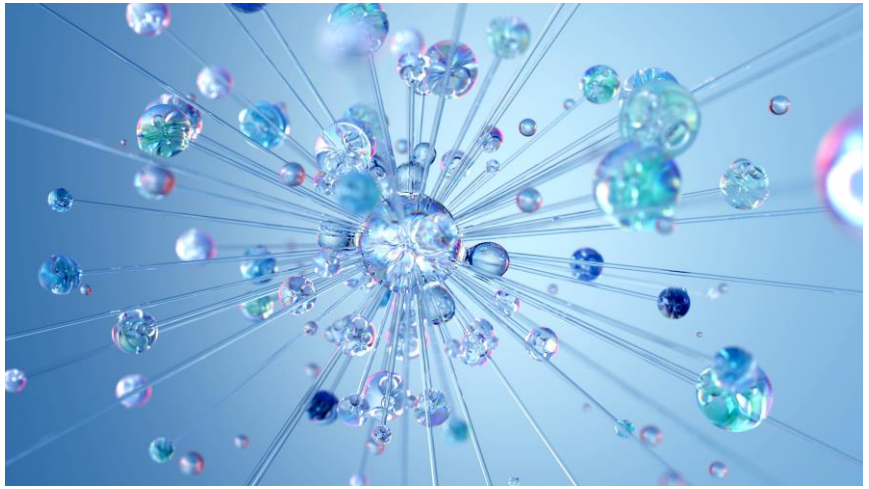
Vilniaus universitetas



https://www.pexels.com

Įvadas į duomenų tyrybą

prof. dr. Olga Kurasova

olga.kurasova@mif.vu.lt

Turinys

- Duomenų tyryba. Kas tai?
- Duomenų rinkimas.
- Duomenų pradinis apdorojimas.
- Duomenų rikiavimas.
- Didieji duomenys.
- Duomenų tyrybos uždaviniai ir metodai.
- Duomenų tyryba ir mašininis mokymasis.
- Duomenų tyryba ir vizualizavimas.

https://www.pexels.com

Duomenų tyryba. Kas tai?

Vilniaus universitetas

Duomenys, informacija, žinios

https://internetofwater.org/valuing-data/what-are-data-informa

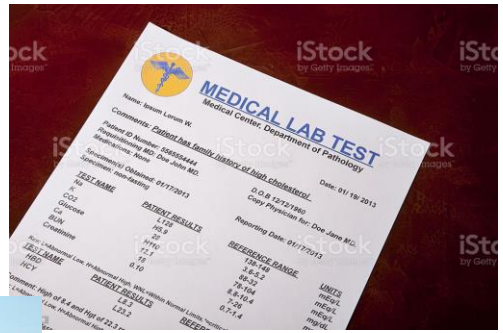
Duomenys, informacija, žinios

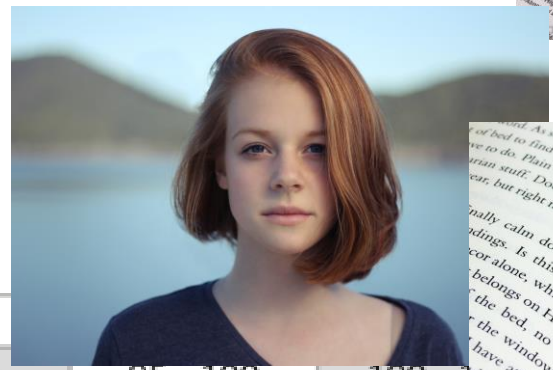
- Visko pradžia yra **duomenys**.
- **Duomenys** – tai objektyviai egzistuojantys faktai, vaizdai, garsai, kurie gali būti naudingi tam tikram uždaviniui spręsti. Dažnai tai **neapdoroti duomenys**.
- **Informacija** – tai duomenys, kurių forma ir turinys yra pateikti būdu, kuris yra tinkamas naudoti sprendimų priėmimo procese. Duomenys virsta informacija, kai **jiems suteikiamas kontekstas** ir jie susiejami su tam tikra problema ar sprendimu.
- **Žinios** – tai gebėjimas spręsti problemas, atnaujinti arba sukurti naujas vertes **remiantis ankstesne patirtimi**, įgūdžiais ar išmokimu. Tai žmogaus proto abstrakcija apie duomenis, jų prasmę, naudą ir sąryšius. Turimos žinios gali virsti informacija, kuri gali būti panaudota **naujoms žinioms įgyti**.

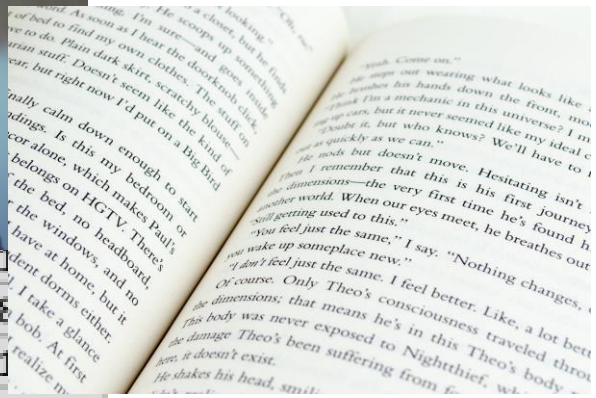
Duomenų pavyzdžiai

- batų numeriai, drabužių dydžiai, prekių kainos, įmonės ekonominiai rodikliai, pacientų kraujo tyrimų rodikliai, nuotraukos, medicininiai vaizdai, įvairūs garsai ir kt.
- duomenys gali būti įvairių formatų:

  skaičiai, tekstas, vaizdai, garsai.





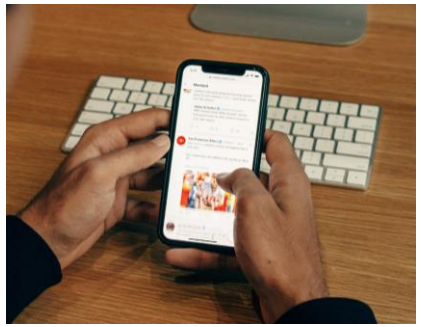


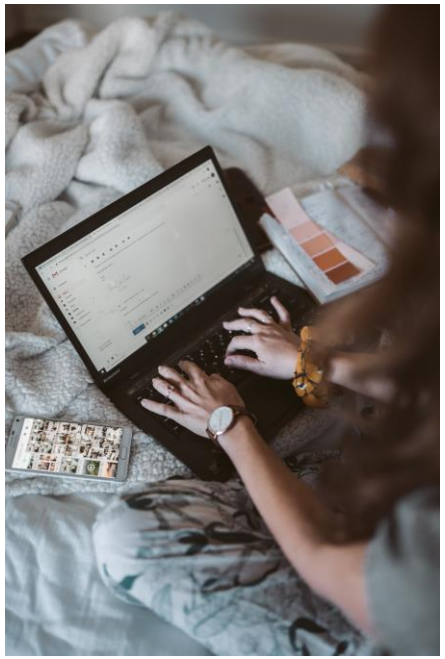
Dydžiai	XS	S	M	L		
A : Krūtinė(cm)	75 - 80	80 - 85	85 - 90	90 - 95	95 - 100	100 - 1
B : Liemuo(cm)	58 - 63	60 - 65	65 - 70	70 - 75	75 - 80	80 - 8
C : Klubai (cm)	82 - 88	87 - 92	92 - 97	95 - 100	98 - 103	100 - 1
D : Torsas (cm)	132 - 140	140 - 148	148 - 156	155 - 164	164 - 172	168 - 176
Ūgis (cm)	150 - 160	155 - 165	160 - 170	165 - 175	170 - 180	175 - 185

https://dancemakers.lt/dydziai/wear-moi-suaugusiuju-ir-vaiku-drabuziu-dyziai/
https://unsplash.com

Informacijos pavyzdžiai

- skelbimas, reklama, bukletas, trumpoji žinutė, elektroninis laiškas ir kt.

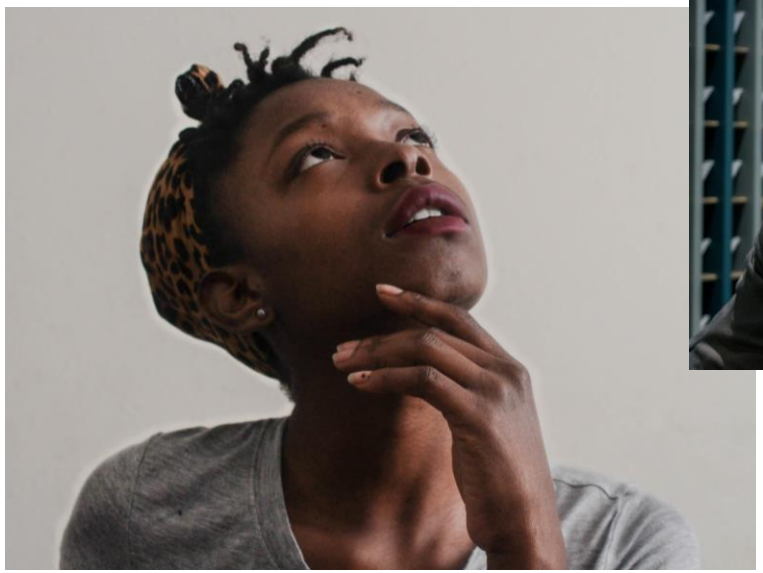


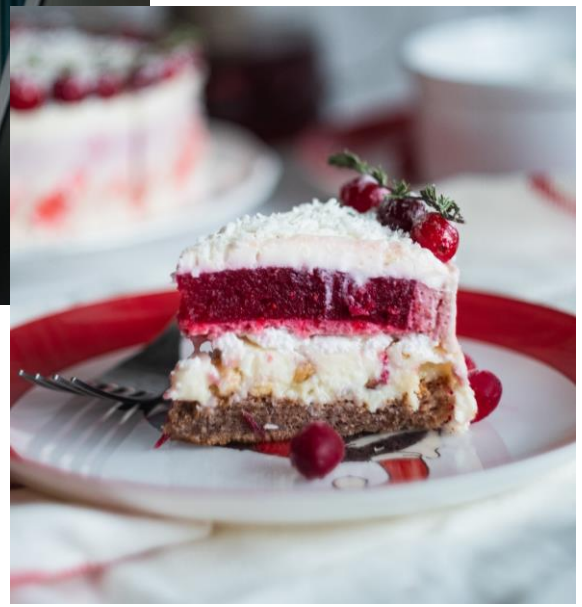


https://unsplash.com, https://www.pexels.com

Žinių pavyzdžiai

- matematinės žinios, žinios apie Lietuvos istorinius faktus, žinojimas kaip pasigaminti Napoleono tortą ir kt.





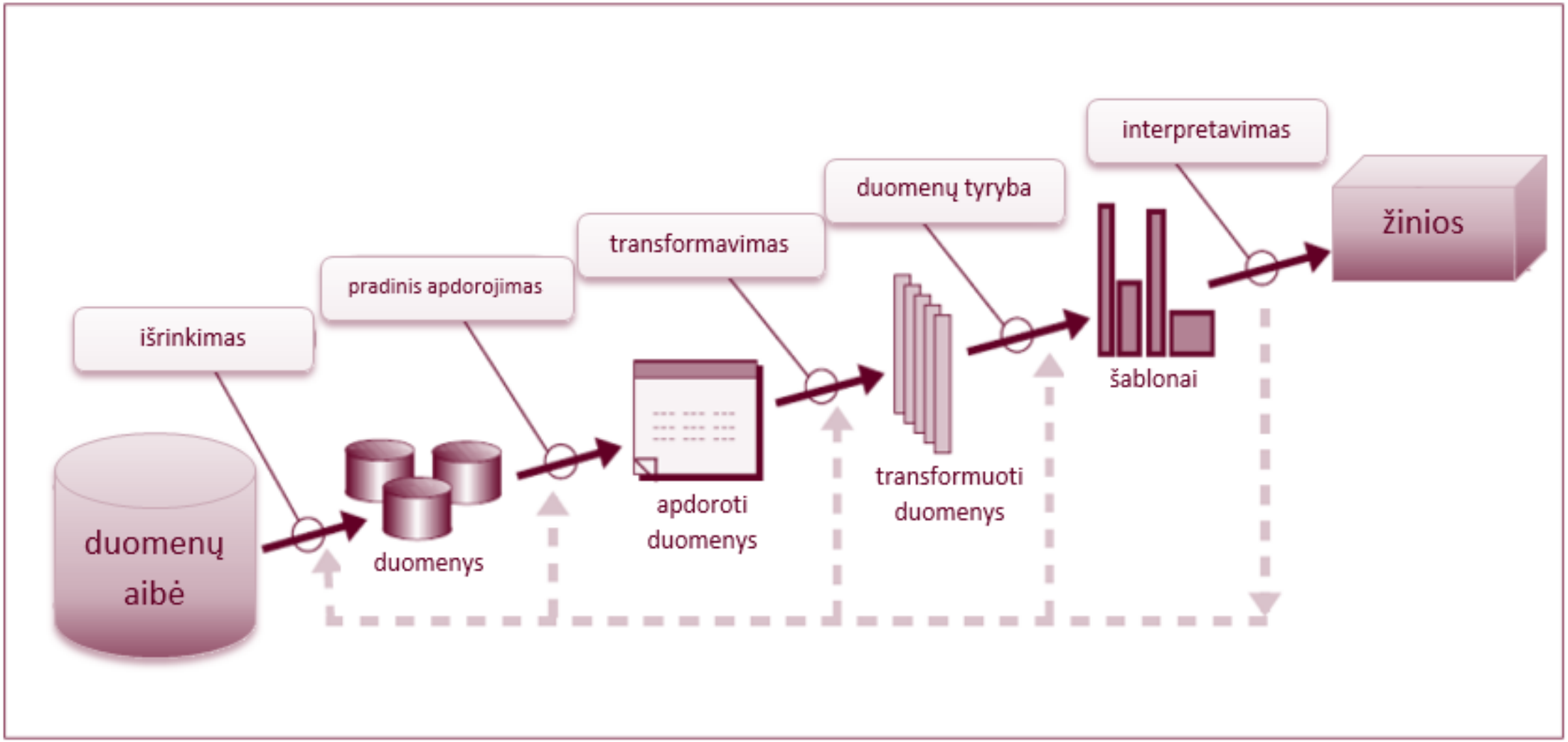
https://unsplash.com

Duomenų tyryba

- **Duomenų tyryba** – tai procesas, kurio metu iš pradinių duomenų, juos apdorojant įvairiais metodais, gaunama naudinga informacija ir žinios.
- Duomenų **tyrybos** (*Data **Mining***) termino atsiradimas. Tiesioginio žodžio „*mining*“ vertimo į lietuvių kalbą nėra, kadangi šis žodis vartojamas metalurgijoje, o šios pramonės šakos Lietuvoje beveik nėra. „*Mining*“ – tai procesas, kurio metu kasinėjant randami naudingi produktai, pvz., iškasenos.
- Labai panašus, bet siauresnis terminas yra „**duomenų analizė**“.

https://www.pexels.com

Duomenų tyryba žinių radimo procese

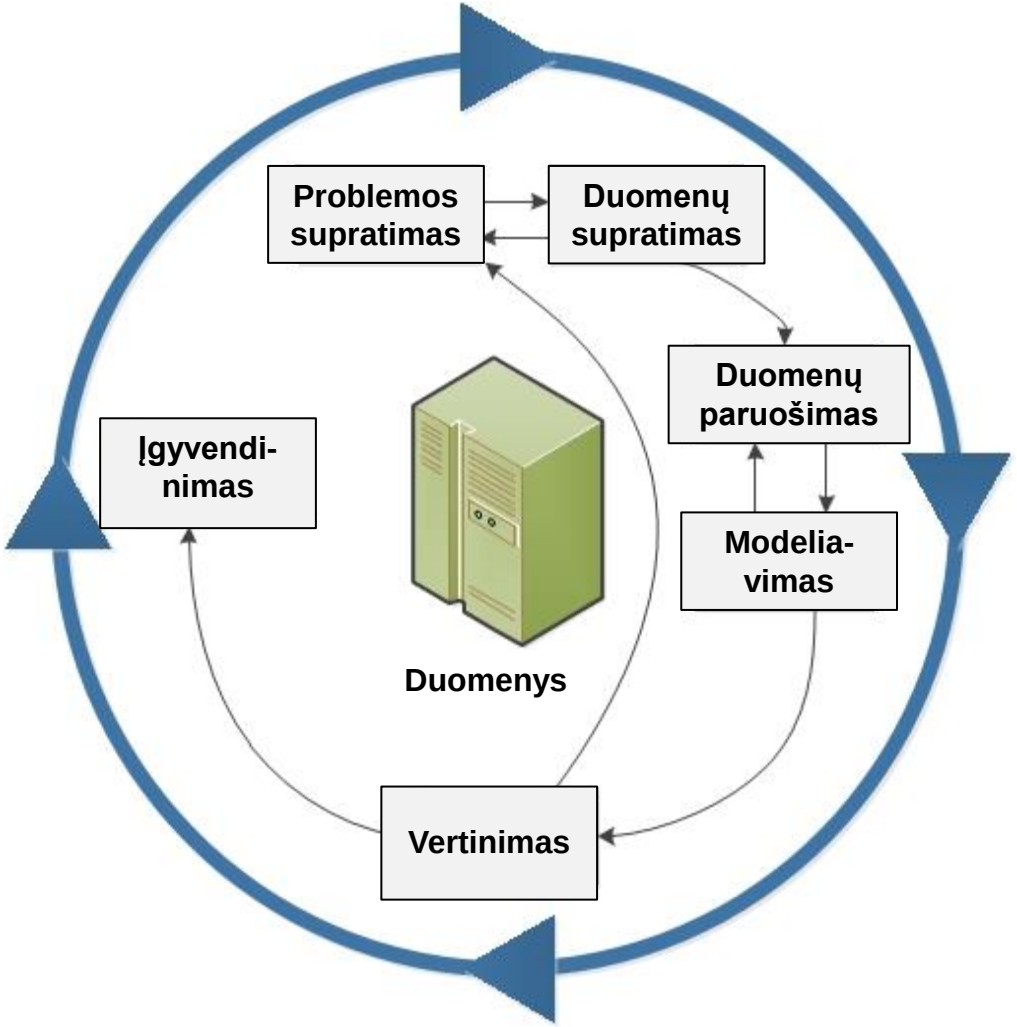


Data Mining in Knowledge Discovery in Databases

Žinių radimo procesą sudarantys žingsniai:

- Iš visos duomenų aibės **išrenkami** analizuojami duomenys;
- Atliekamas **pradinis duomenų apdorojimas** (valymas, filtravimas, transponavimas, požymių atrinkimas, normavimas);
- Atliekamas **duomenų transformavimas**, kurio metu duomenys paruošiami duomenų tyrybos metodui ir programinei įrangai tinkama forma;
- **Duomenys analizuojami** įvairias duomenų tyrybos metodais;
- Interpretuojami ir vertinami gauti rezultatai, ko pasėkoje įgyjamos **naujos žinios**.

Įvairių pramonės šakų duomenų tyrybos standartinis procesas



Adaptuota iš: https://www.ibm.com

Cross-Industry Standard Process for Data mining (CRISP-DM)

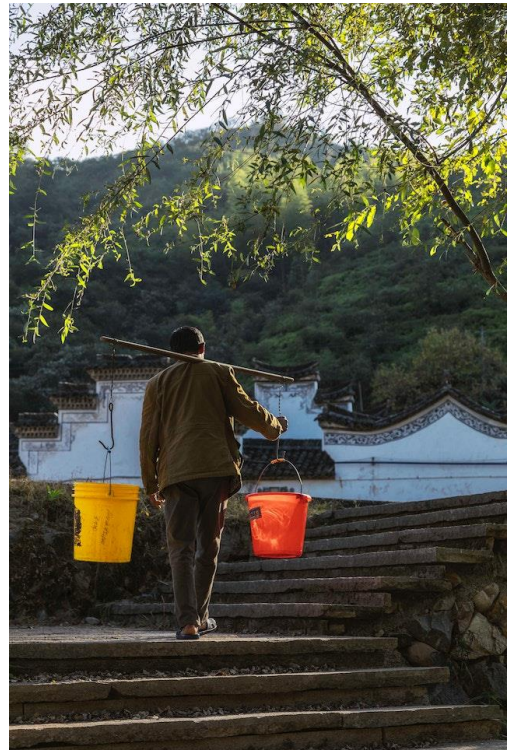
Įvairių pramonės šakų duomenų tyrybos standartinis procesas

- Pradžioje reikia **gerai suprasti numatomą spręsti problemą** – užsakovo tikslus ir reikalavimus (*business understanding*).
- Nustatyti, kokie **duomenys** bus reikalingi šiai problemai spręsti, suprasti duomenis, įvertinti duomenų kokybę (*data understanding*).
- **Paruošti duomenis** modeliavimui: išrinkti reikiamus duomenis, juos išvalyti, integruoti iš kelių šaltinių (jei būtina) (*data preparation*).
- **Modeliavimas** – tai duomenų tyrybos modelių kūrimas ir taikymas nagrinėjamiems duomenims (*modeling*).
- **Modelio rezultatų įvertinimo metu** nustatomas rezultatų tikslumas, patikimumas ir kt. (*evaluation*).
- Paskutinis žingsnis – **modelio įgyvendinimas** (programinės įrangos sukūrimas) (*deployment*).

Duomenų rinkimas

Duomenų rinkimas

- Ar tai **paprasta** ar **sudėtinga** užduotis?
- **Duomenų** yra visur **daug**. Bet ar jie visi **tinkami** specifinei problemai spręsti?
- Galimi **du atvejai**:
  - užsakovas „**atsineša**“ savo turimus duomenis,
  - **reikia gauti** reikalingus duomenis.



https://www.pexels.com

Duomenų rinkimas

- Palankesnė situacija, kai **duomenys pradedami rinkti** suformulavus problemą. Tuomet tikėtina, kad bus surinkti **duomenys, tinkantys** šiai problemai spręsti. Trūkumas – **procesas ilgai užtrunka**.
- Labiau **komplikuotesnis atvejis**, kai nėra laiko rinkti duomenų, reikia „**suktis**“ su esamais.
- Duomenis galima **įsigyti** iš komercinių šaltinių.
- Duomenys gali būti renkami iš įvairių **atvirų šaltinių**:
  - Valstybės duomenų agentūra (https://www.stat.gov.lt/),
  - Lietuvos atvirų duomenų portalas (https://data.gov.lt),
  - Lietuvos hidrometeorologijos tarnyba (http://www.meteo.lt),
  - Atviri duomenys atviram Vilniui (https://open.vilnius.lt/).

Duomenų rinkimas

- Automatizuotas duomenų rinkimas **naudojant API** (*programų programavimo sąsaja*, *application programming interface*).
- Duomenų **rinkiniai parsisiunčiami failuose**. Dažnas formatas **csv** (*comma-separated values*). Šio formato failus galima atsidaryti **bet kokiu teksto redaktoriumi**, **MS Excel** programa. Įprastai programavimo kalbos turi funkcijas šio formato failams nuskaityti.
- Atidarius **csv** failą **MS Excel** programa, failą galima išsaugoti **xlsx** formatu.

Duomenys

- Tarkime turime **objektus**, $X_1, X_2, \ldots, X_m$, kuriuos apibūdina tam tikri **požymiai** (*features*) $x_1, x_2, \ldots, x_n$.
- **Objektais** gali būti moksleiviai, pacientai, įrenginiai, gaminiai, gamybos procesai, gamtos reiškiniai ir kt. Objektai dar gali būti vadinami **duomenų įrašais** (*data items*, *data samples*).
- Objektus apibūdinančiais **požymiais** gali būti moksleivių pažymiai, pacientų kraujo tyrimų rezultatai, gaminio savybės ir pagaminimo laikas.
- Tokius duomenis galima saugoti **lentelėse**.

Duomenų lentelė

Bendras atvejis:

		Požymiai			
		x_1	x_2	...	x_n
Duomenų įrašai	X_1	x_{11}	x_{12}	...	x_{1n}
	X_2	x_{21}	x_{22}	...	x_{2n}
	...	...	...	...	...
	X_m	x_{m1}	x_{m2}	...	x_{mn}

Duomenų lentelė

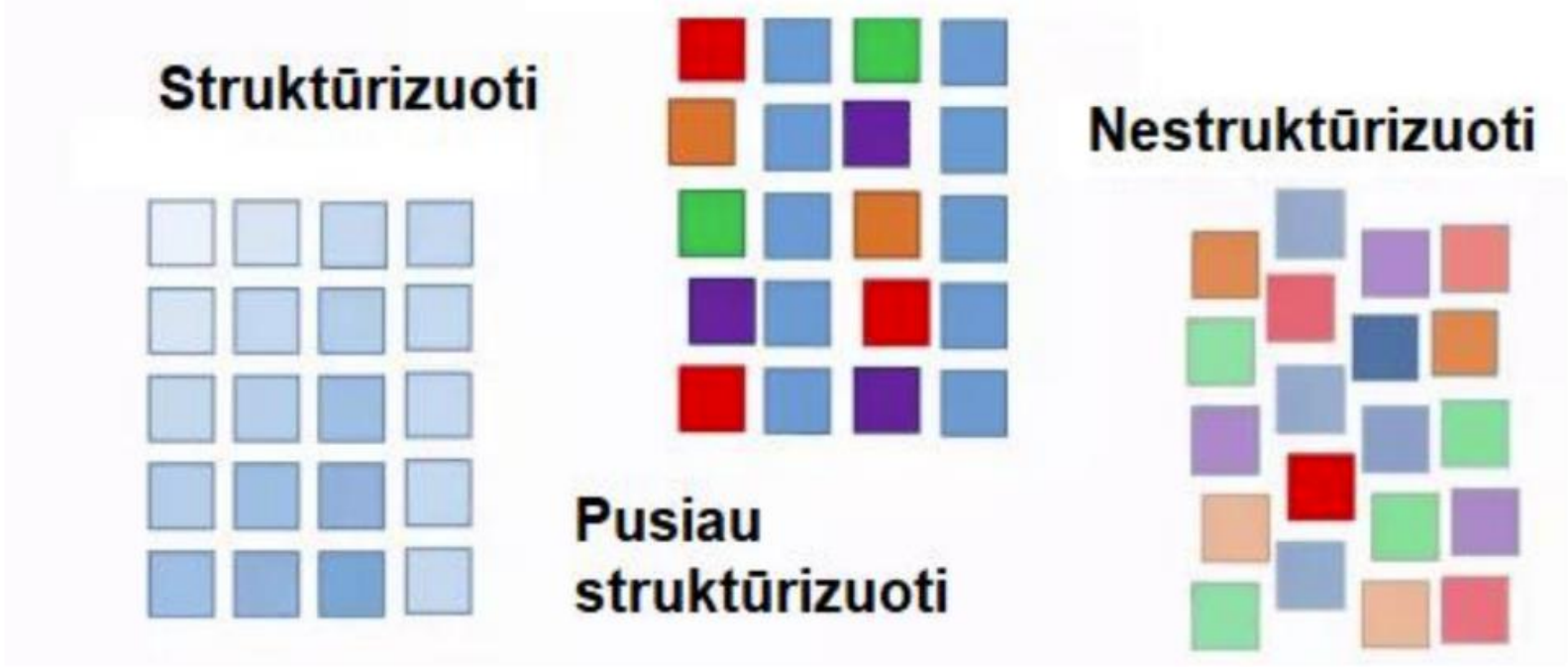
Duomenų pavyzdys – moksleivių dvejų metų pažymių vidurkiai.

		Požymiai			
		Matematika	Lietuvių k.	Fizika	IT
Duomenų įrašai	Moksleivis Nr. 1	8,5	9,2	7,5	9,5
	Moksleivis Nr. 2	7,2	7,7	8,5	7,2
	...	...	...	...	...
	Moksleivis Nr. 20				10

Kito tipo duomenys

- Ne visada duomenys būna **struktūrizuoti** ir patalpinti **lentelėse** (*tabular*).
- Dažnai susiduriama su **nestruktūrizuotais** duomenimis: tekstas, vaizdai, video, garsai ir kt.

Įvairių tipų duomenų vaizdavimas



Šaltinis: https://www.astera.com/type/blog/structured-semi-structured-and-unstructured-data/

Duomenų pradinis apdorojimas

Vilniaus universitetas

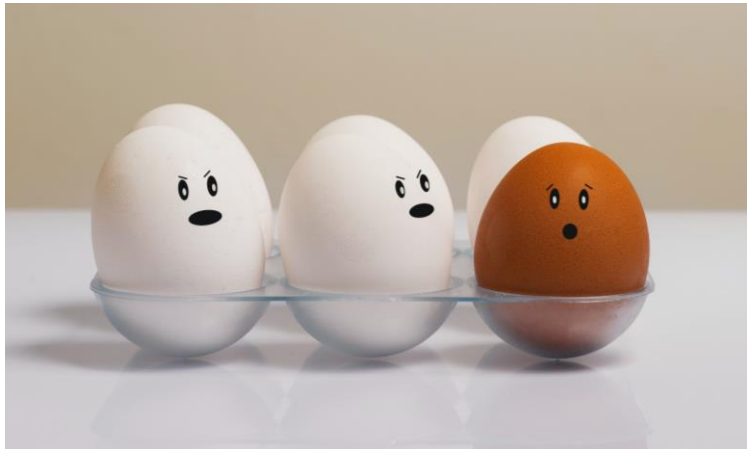
Duomenų valymas

- **Duomenų valymas** (*data cleaning*) – tai neteisingų, pasikartojančių ar kitaip klaidingų duomenų rinkinio duomenų taisymo procesas.
- Viena iš dažnai pasitaikančių situacijų, kai **duomenyse trūksta** tam tikrų reikšmių (*missing values*).
- **Ką daryti**?
- Galimi **du būdai**:
  - **Panaikinti įrašus** su trūkstamomis reikšmėmis.
  - Vietoj trūkstamų reikšmių **įrašyti kitas reikšmes**, pavyzdžiui, to požymio vidurkį.

https://www.pexels.com

Duomenų valymas: išskirčių paieška

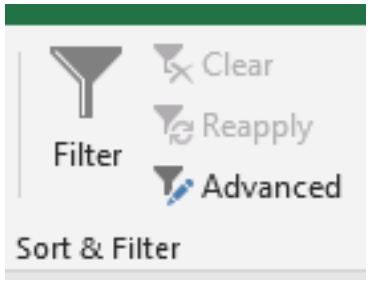
- Kita dažna situacija, kai duomenyse yra reikšmių, kurios **žymiai skiriasi nuo kitų** to paties požymio reikšmių. Šios reikšmės vadinamos **išskirtimis** (*outliers*).
- **Kodėl jos atsiranda**?
- Galimos dvi priežastys:
  - **Klaidingi duomenys** (žmogiškoji klaida, daviklio ar kito įrenginio sutrikimas).
  - **Išskirtys nurodo duomenų specifiką**, pavyzdžiui, retą ligą, gaminio broką ir pan.



https://www.pexels.com

Duomenų filtravimas

- **Duomenų filtravimas** – tai duomenų pradinio apdorojimo procesas, kai iš analizuojamų duomenų aibės **atrenkami** duomenų įrašai **pagal nustatytus kriterijus**.
- **Kriterijai** gali būti gana **paprasti**. Pavyzdžiui, reikia atrinkti tuos duomenų įrašus, kurių požymio (stulpelio) „Ligoninė“ reikšmė yra „VšĮ Vilniaus universiteto ligoninės Santaros klinikos“.
- Tačiau galima suformuoti ir **daug sudėtingesnius kriterijus**, apimančius kelis požymius.



Duomenų normavimas

Duomenų normavimas – tai procesas, kurio metu suvienodinamos duomenų požymių skalės. Įprastai reikšmės suvedamos į intervalą $[0; 1]$.

	Atlygi-nimas	Stažas
A1	1500	4
A2	1800	10
A3	2300	2
Min	1500	2
Max	2300	10

	Atlygi-nimas	Stažas
A1	0	0,25
A2	0,375	1
A3	1	0
Min	0	0
Max	1	1

P. S. Normavimą atliekant MS Excel, gera proga **prisiminti langelių adresų fiksavimą**.

Duomenų normavimas

- Tegu $X_i = (x_{i1}, x_{i2}, \dots, x_{in}),\ i = 1, \dots, m$. Duomenys **normuojami** kiekvieno požymio reikšmę pakeičiant:

$$x_{ij} \leftarrow \frac{x_{ij} - \min(x_{1j}, x_{2j}, \dots, x_{mj})}{\max(x_{1j}, x_{2j}, \dots, x_{mj}) - \min(x_{1j}, x_{2j}, \dots, x_{mj})}.$$

- Čia
  - $\min(x_{1j}, x_{2j}, \dots, x_{mj})$ yra j-tojo požymio reikšmių **minimali reikšmė**,
  - $\max(x_{1j}, x_{2j}, \dots, x_{mj})$ yra j-tojo požymio reikšmių **maksimali reikšmė**.
- Po normavimo kiekvieno požymio minimalios reikšmės tampa lygios **0**, o maksimalios **1**.

Duomenų rikiavimas

Vilniaus
universitetas

Duomenų rikiavimas

- **Skaitinius duomenis** (skaičius) galima rikiuoti jų didėjimo ar mažėjimo tvarka.
- **Tekstinius duomenis** (tekstą) galima rikiuoti abėcėlės tvarka.
- Paprastoms rikiavimo procedūros atlikti tinka **MS Excel** programa.
- Dažnai duomenų rikiavimas **reikalingas duomenų bazėse** ar programuojant duomenų analizės algoritmus.

Duomenų rikiavimo algoritmai

- Yra daug duomenų rikiavimo algoritmų: **įterpimo**, **MinMax**, **burbuliukio**, **greitojo rikiavimo** ir kt.
- Nurodytame tinklalapyje galima rasti **duomenų rikiavimo algoritmų**, įgyvendintų **Python** kalba. Galima interaktyviai keisti kodą, skaičiavimai atliekami saityno aplinkoje: https://www.tutorialspoint.com/python_data_structure/python_sorting_algorithms.htm
- P.S. Įterpimo algoritme `while` ciklas turi būti `For` viduje, t. y. `while` blokas patrauktas į dešinę.

Didieji duomenys

Didieji duomenys. Kas tai?

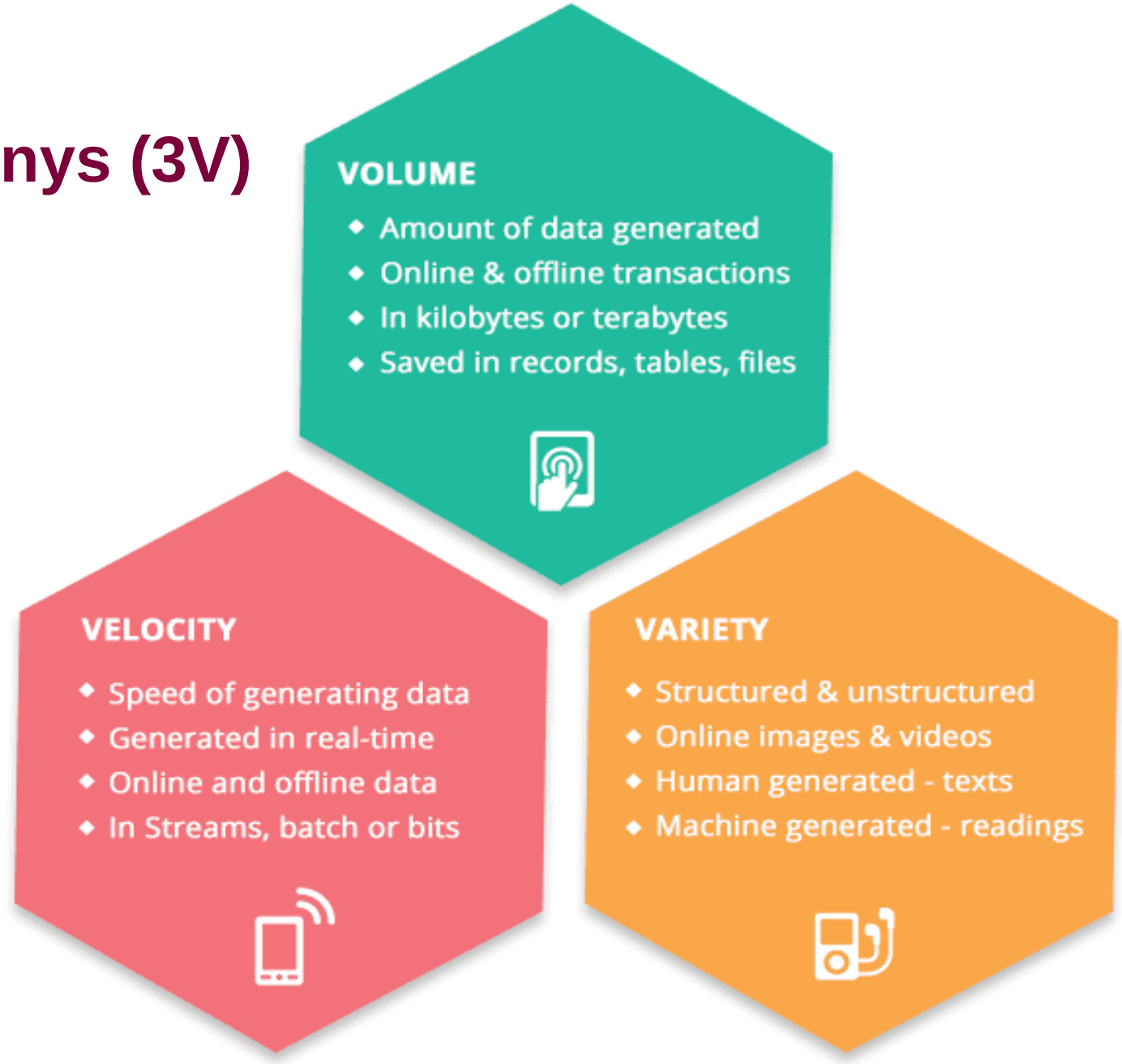
- **Modernios technologijos** leidžia generuoti **milžiniškus duomenų kiekius**.
- Pagrindiniai **didžiųjų duomenų** (*big data*) šaltiniai:
  - Astronomija,
  - Meteorologija,
  - Genų inžinerija,
  - Medicina,
  - Bankinės ir finansinės sistemos,
  - Socialiniai tinklai,
  - Telekomunikacija,
  - Daiktų internetas (*Internet of Things*, IOT)
  - Saitynas (*web*).
  - Kiti

https://www.pexels.com

Didieji duomenys. Ar tai tik didelė duomenų apimtis?

- **Didieji duomenys** (*big data*) – tai tokie duomenys, kurių **apimtis** yra tokia **didelė**, kad jų neįmanoma apdoroti ir analizuoti naudojant įprastus metodus.
- **Didžiuosius duomenis** charakterizuoja pagrindinės savybės (**3V**):
  - didžiulis duomenų kiekis (**Volume**),
  - didelė duomenų įvairovė (**Variety**),
  - nuolat atsirandantys nauji duomenys (**Velocity**).

Didieji duomenys (3V)



https://www.wishworks.com

Didieji duomenys. Jų tipai.

- **Struktūrizuoti** duomenys
  - Lentelės, reliacinės duomenų bazė ...
- **Nestruktūrizuoti** duomenys
  - Tekstas, vaizdai, video, audio ...
- **Dalinai struktūrizuoti** duomenys
  - Dokumentai lentelėse ...
- **Metaduomenys**
  - Struktūrizuoti duomenis apie duomenis
- **Srautiniai** duomenys
  - Laike atsirandantys duomenys iš daviklių ir kt. ...
- **Geografiniai** duomenys
  - Duomenys, apimantys informaciją apie poziciją erdvėje

https://eihdigital.com/wordpress-design-and-development-service/

Duomenų tyrybos uždaviniai ir metodai

Vilniaus
universitetas

Duomenų tyryba: paprasčiausi uždaviniai

- **Statistinių charakteristikų skaičiavimas**: vidurkis, dispersija, mediana, maksimali reikšmė, minimali reikšmė ir kt. – **aprašomoji statistika**.
- **Matematinė statistika** yra **duomenų tyrybos** dalis.

Duomenų tyryba: klasifikavimas

- **Duomenų klasifikavimas** – vienas iš dažniausių sprendžiamų duomenų tyrybos uždavinių.
- **Klasifikavimo tikslas** – priskirti duomenis tam tikrai klasei.
- Įprastai daliai duomenų klasės yra žinomos. Pritaikius klasifikavimo metodą, **klasės yra nustatomos** duomenims, kurių klasės nebuvo žinomos.

Klasifikavimas medicinoje

- Klasifikavimo uždaviniai dažnai sprendžiami **medicinoje**, siekiant nustatyti **preliminarią diagnozę**.
- Tarkime, turime pacientų širdies veiklos rodiklius ir kraujo tyrimų duomenis ir žinome, kad dalis pacientų serga tam tikra liga, kiti pacientai yra sveiki. Vadinasi, turime dviejų klasių duomenis: **sergantys**, **sveiki**.
- Taip pat turime pacientus, kurių širdies veiklos rodikliai yra išmatuoti ir kraujo tyrimo rezultatai yra žinomi, tačiau **jie nėra priskirti** nei vienai klasei (t. y., nėra diagnozės).
- **Klasifikavimo tikslas** – priskirti šiuos pacientus vienai iš dviejų klasių (**sveiki** ar **sergantys**).

Duomenų pavyzdys

	SKS	*DKS*	*ŠD*	*MTL_Ch*	**Klasė**
Pacientas Nr. 1	120	85	75	3,5	sveikas
Pacientas Nr. 2	125	75	62	5,0	sveikas
Pacientas Nr. 3	110	70	70	2,8	sveikas
...	...	...	...	...	...
Pacientas Nr. 101	145	85	90	5,5	serga
Pacientas Nr. 102	152	80	75	2,7	serga
...	...	...	...	...	...
Pacientas Nr. 201	110	90	65	4,0	**???**
Pacientas Nr. 202	135	75	62	2,6	**???**

SKS – sistolinis kraujo spaudimas,

DSK – diastolinis kraujo spaudimas,

ŠD – širdies dažnis,

MTL_Ch – Cholesterolis, esantis mažo tankio lipoproteinų sudėtyje.

Duomenų tyryba: prognozavimas

- Dar vienas populiarus duomenų analizės uždavinys yra **prognozavimas**, kurio metu žinant duomenų dalies požymių reikšmes, nustatomos reikšmės požymiui, kurio reikšmė nežinoma.
- **Prognozavimo tikslas** – iš „istorinių“ duomenų nustatyti reikšmes „ateities“ duomenims.
- **Prognozavimo uždaviniai** sprendžiami įvairiose srityse. Pavyzdžiui, meteorologijoje remiantis ankstesnių metų to paties laikotarpio duomenimis bei pastarojo laikotarpio pokyčių prognozuojama oro temperatūra ateinančiai savaitei.
- Taip pat **prognozavimas** atliekamas vertybinių popierių biržoje, įvairiose finansinėse rinkose ir kitur.

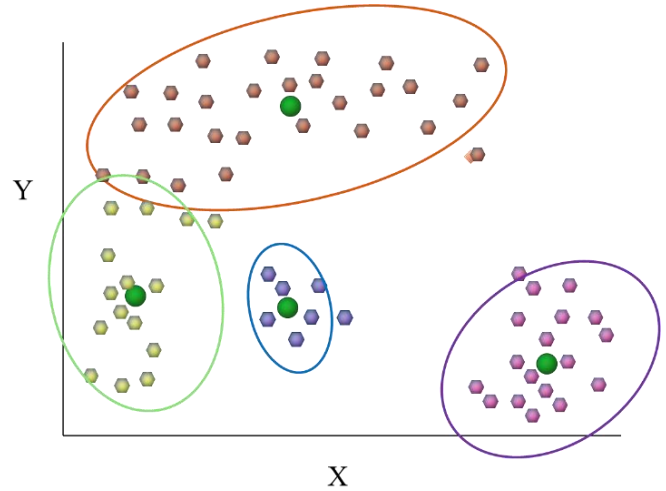
Duomenų prognozavimas: pavyzdys

- **Žaliuose langeliuose** nurodytas vienos įmonės metinis pelnas (tūkstančiais). Pagal juos bus norima prognozuoti pelną, nurodytą **raudonuose langeliuose**, t. y., pagal penkių metų duomenis bus prognozuojamas pelnas sekantiems metams;
- **Mėlyni langeliai** nurodytas pastarųjų penkių metų pelnas;
- Tikslas – prognozuoti pelną sekantiems metams (**oranžinis langelis**).

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
1	832	738	891	566	411	774	260	379	510	973	946	
2	832	738	891	566	411	774	260	379	510	973	946	
3	832	738	891	566	411	774	260	379	510	973	946	
4	832	738	891	566	411	774	260	379	510	973	946	
5	832	738	891	566	411	774	260	379	510	973	946	
6	832	738	891	566	411	774	260	379	510	973	946	
7	832	738	891	566	411	774	260	379	510	973	946	???

Duomenų tyryba: klasterizavimas

- **Klasterizavimo** tikslas – suskirstyti objektus taip, kad panašūs objektai patektų į tą patį klasterį, o skirtingi – į skirtingus.
- **Klasteris** – tai panašių objektų grupė.
- Čia svarbu nustatyti objektų **panašumo matą**. Vienas paprasčiausių panašumo matų – gerai žinomas Euklido atstumas.



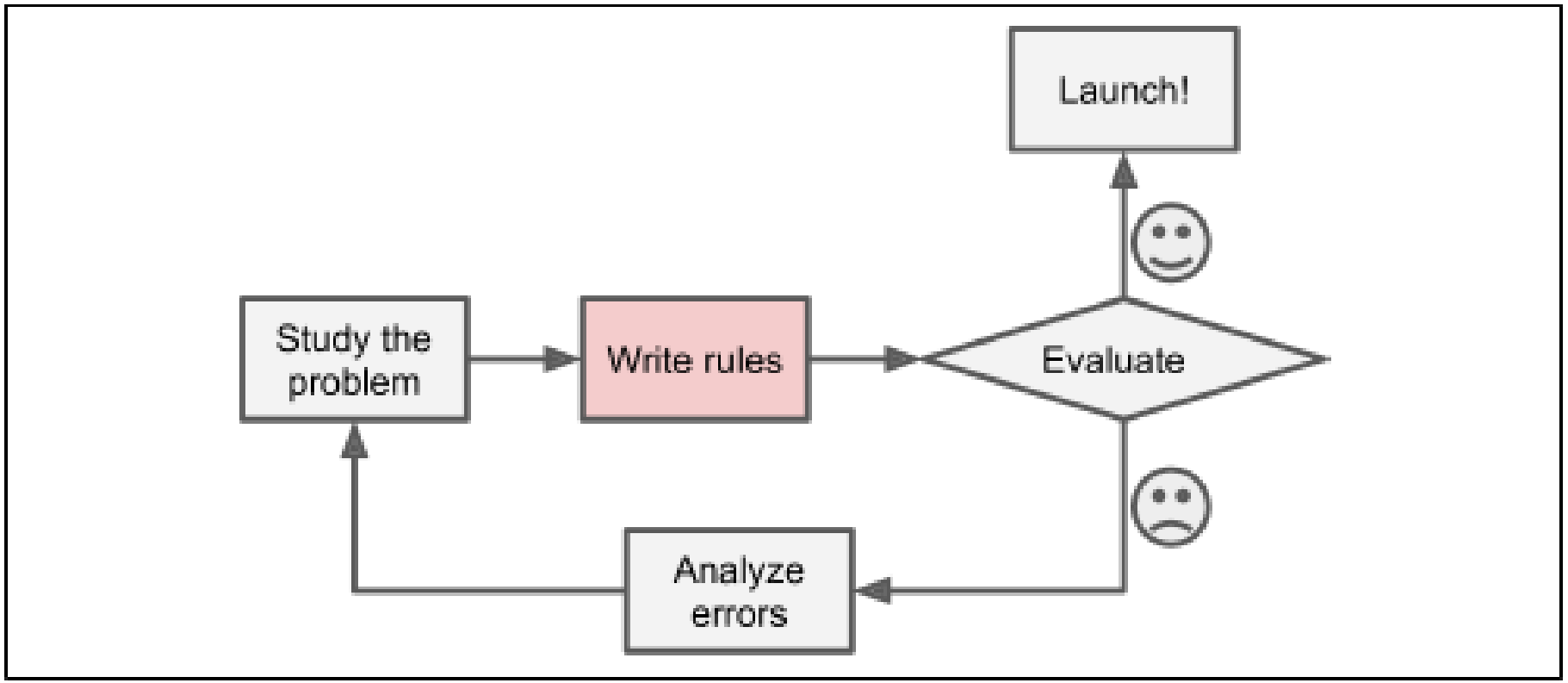
Duomenų tyryba ir mašininis mokymasis

Vilniaus universitetas

Mašininis mokymasis. Kas tai?

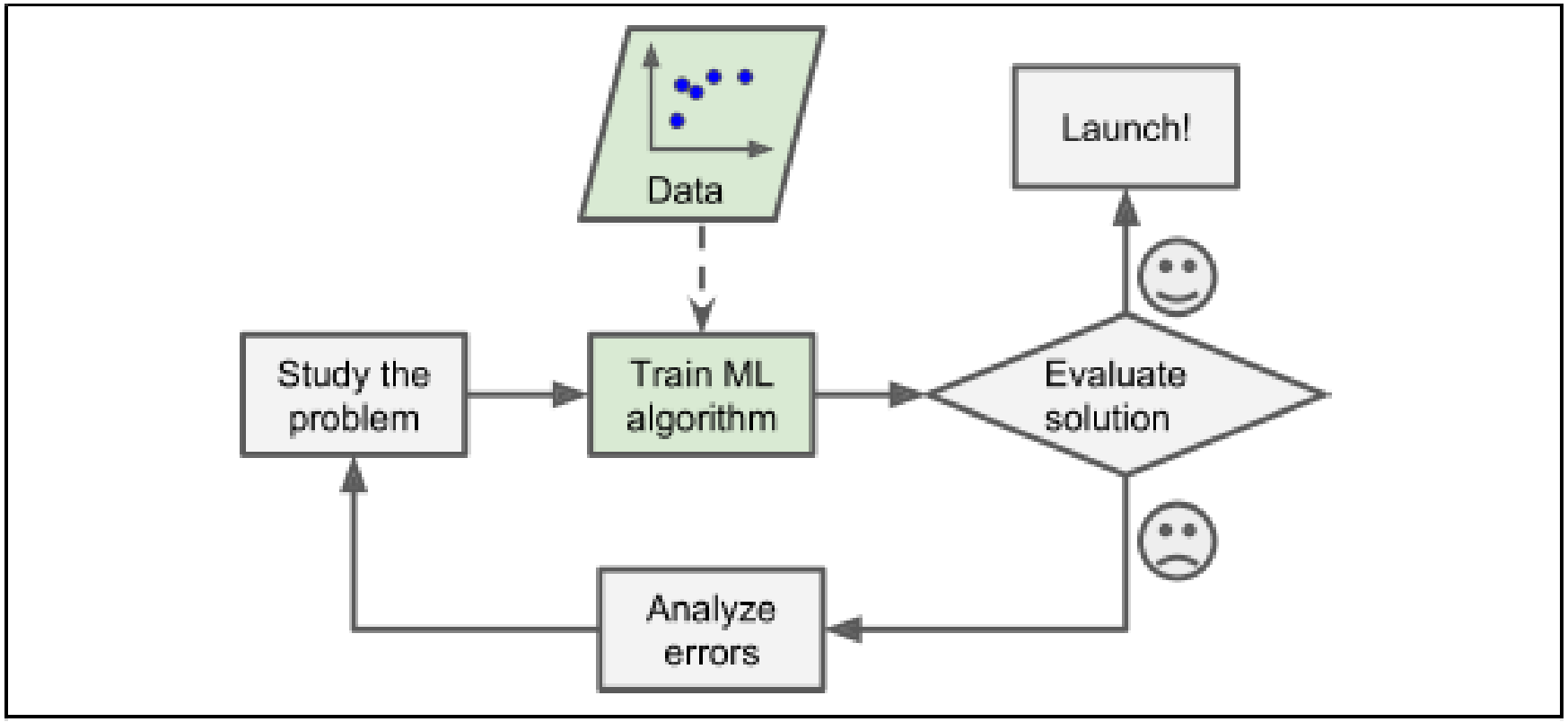
- **Mašininis mokymasis** (*machine learning*) – tai informatikos sritis, suteikianti kompiuteriams **galimybę mokytis be aiškių instrukcijų**. Tai algoritmai, kurių veikimas gerėja, kai jie gauna **daugiau duomenų**.
- **Gilusis mokymasis** (*deep learning*) – tai mašininio mokymosi poaibis, kuris efektyvus mokant naudojant didžiulius duomenų kiekius.
- **Dirbtinis intelektas** (*artificial intelligence*) – tai sistema, kuri gali mąstyti, jausti, veikti ir prisitaikyti, atlikti funkcijas, kurios įprastai siejamos su žmogaus gebėjimais.

Tradicinis (taisyklėmis grįstas) būdas



Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc.

Mašininiu mokymusi grįstas būdas



Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc.

Mašininis mokymasis ir duomenų tyryba

- **Panašumai**:
  - Abu konceptai apima analitinius procesus.
  - Abu mokosi iš duomenų.
  - Abiem reikia daug duomenų, kad pasiekti gerą rezultatą.
- **Skirtumai** (vienas požiūris):
  - **Duomenų tyrybos** metodais nagrinėjami turimi duomenys. **Mašininio mokymosi** algoritmais iš esamų duomenų bandoma nuspėti, kas bus ateityje.
  - **Duomenų tyrybos** procese yra daugiau „rankinio“ darbo ir sprendimų priėmėjo įsikišimo. **Mašininio mokymosi** algoritmai yra daugiau automatizuoti, čia eliminuojamas žmogaus įsikišimas.

Mašininis mokymasis ir duomenų tyryba

Kitu požiūriu **duomenų tyryba** gali būti dvejopa:

- arba tik nagrinėjanti duomenis (**statistika**),
- arba, pasitelkus **mašininio mokymosi** algoritmus, nuspėjanti (numatanti), kas bus ateityje.

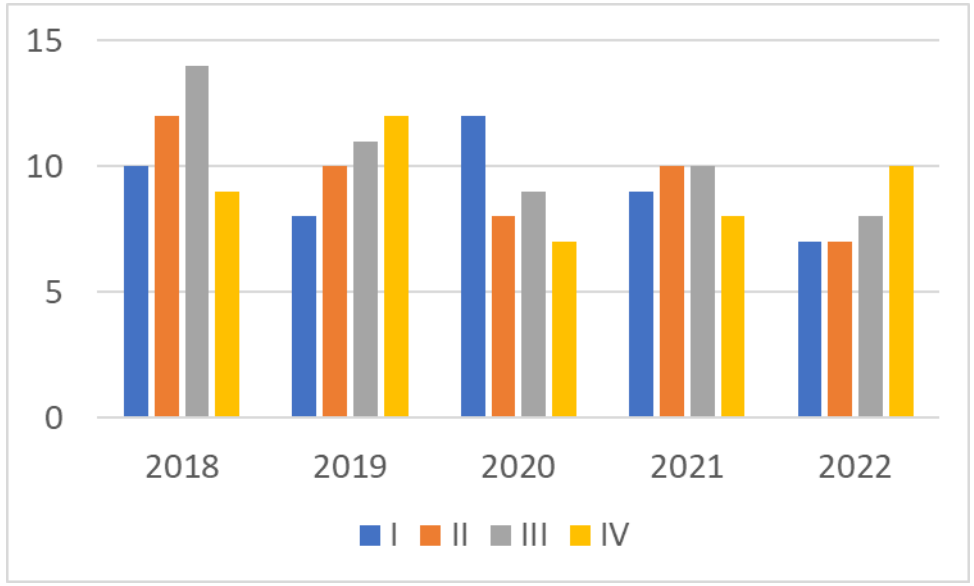
https://www.sas.com/en_nz/insights/analytics/data-mining.html

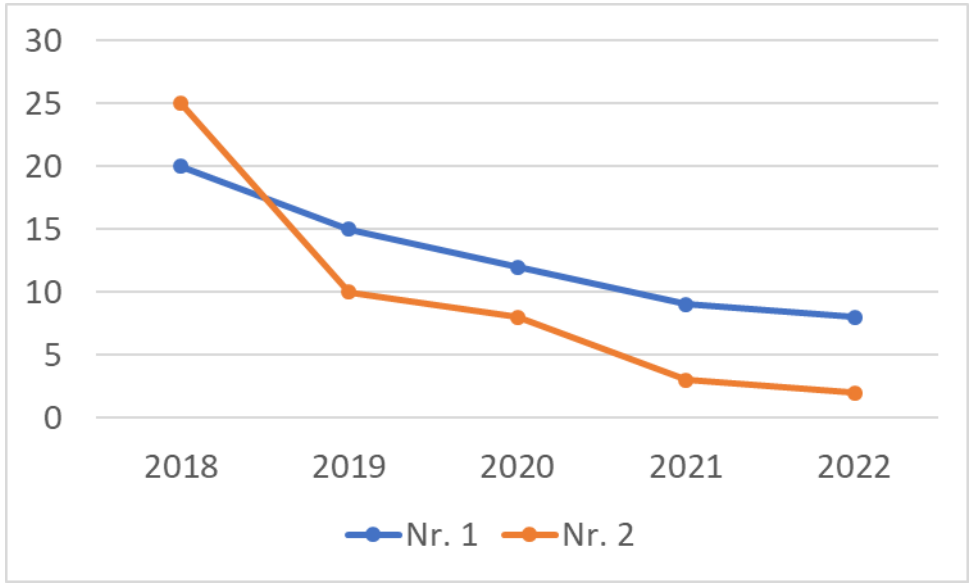
Duomenų tyryba ir vizualizavimas

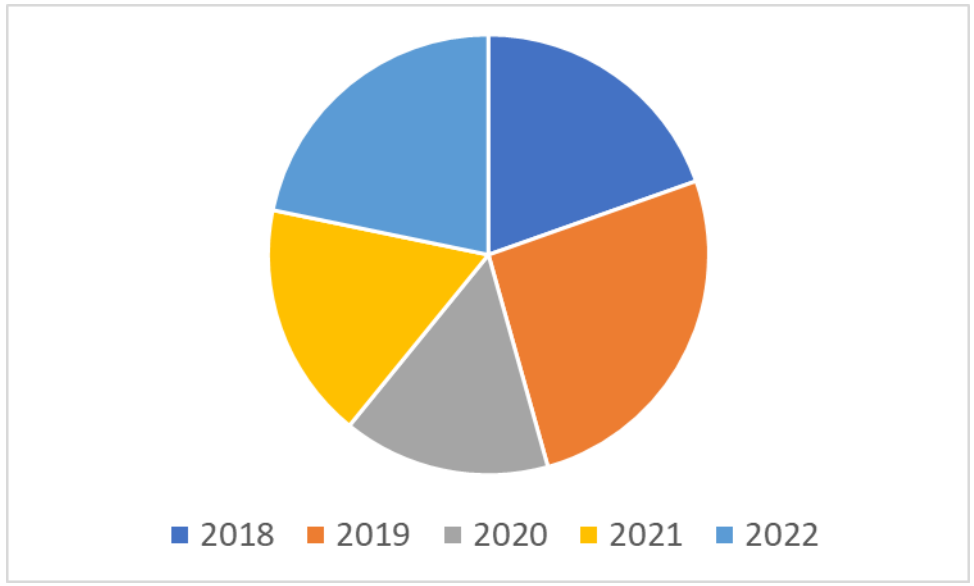
Vilniaus
universitetas

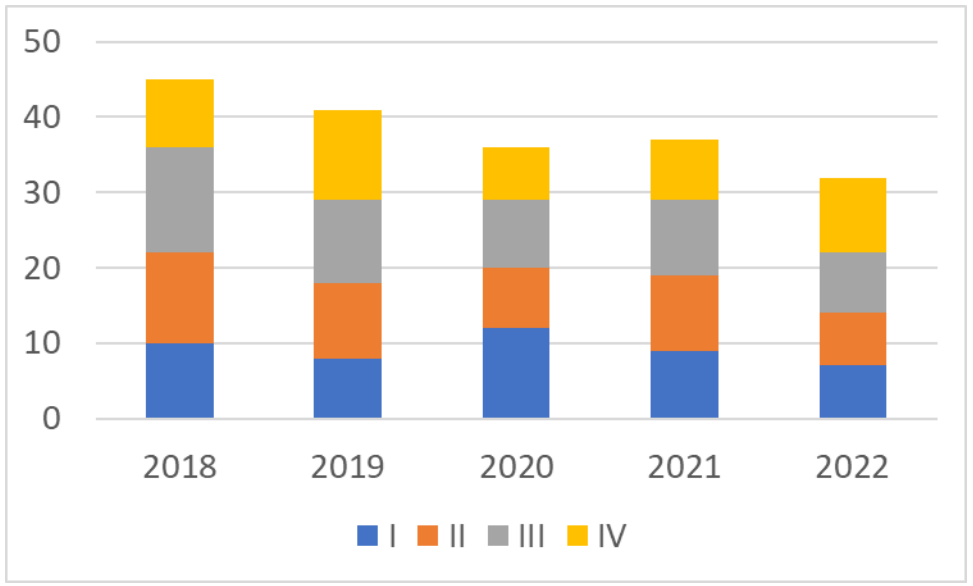
Duomenų vizualizavimas

- **Duomenų vizualizavimas** gali padėti tam tikra prasme struktūrizuoti turimus duomenis ir neleisti pasiklysti duomenų labirintuose.
- **Duomenų vizualizavimas** gali būti naudingas vaizduojant duomenų tyrybos rezultatus.
- Standartiniai grafikai **MS Excel** programoje gali puikiai atlikti **duomenų vizualizavimo** ir **duomenų tyrybos rezultatų vizualizavimo** užduotis.


15
10
5
0
2018
2019
2020
2021
2022
I
II
III
IV


30
25
20
15
10
5
0
2018
2019
2020
2021
2022
Nr. 1
Nr. 2


2018
2019
2020
2021
2022


50
40
30
20
10
0
2018
2019
2020
2021
2022
I
II
III
IV