Duomenų tyrybos konspektas

1. Įvadas į duomenų tyrybą

Testas prieš pradedant mokytis skyrių „Duomenų tyryba“

1. ***Kokį vaidmenį vaidina žodis „tyryba“ duomenų tyryboje?***

a) Tai reiškia, kad analizuojant duomenis bus taikomi moksliniai metodai.
b) Tik specialistai, turintys atitinkamą išsilavinimą, tai supranta.
c) Toks žodžių junginys gerai skamba.

2. ***Teiginys „Mokytis duomenų tyrybos yra naudinga tik programuotojams“ yra:***

a) Teisingas
b) Klaidingas.

3. ***Ką turėtumėte atlikti, norėdami pademonstruoti, ar krepšinio komandos žaidėjai yra yra aukštesni, negu vidutinio ūgio žmonės?***

a) Surinkti duomenis.
b) Turėti žinių apie statistiką ir tikimybes.
c) Teisingi a) ir b) variantai.

1.1. Duomenys, informacija, žinios

Kas yra duomenys?

Kasdieniame gyvenime mus nuolat supa duomenys. Tekstas, kurį dabar skaitote, yra duomenys. Jūsų draugų telefonų numerių sąrašas išmaniajame telefone yra duomenys, taip pat ir laikrodžio rodomas dabartinis laikas. Duomenimis operuojame skaičiuodami turimus pinigus arba rašydami laiškus draugams.

Sukūrus kompiuterius duomenys tapo daug svarbesni. Pagrindinis kompiuterių vaidmuo – atlikti skaičiavimus, tačiau jiems reikia duomenų, kuriais jie galėtų operuoti. Todėl turime suprasti, kaip kompiuteriai saugo ir apdoroja duomenis.

Atsiradus internetui, kompiuterių, kaip duomenų tvarkymo įrenginių, vaidmuo padidėjo. Pagalvojus, dabar kompiuterius vis dažniau naudojame duomenims apdoroti ir bendrauti, o ne faktiniams skaičiavimams atlikti. Kai rašome elektroninį laišką draugui arba ieškome informacijos internete - iš esmės kuriame, saugome, perduodame ir tvarkome duomenis.

Duomenų, informacijos ir žinių sąryšiai

(https://data.europa.eu/elearning/en/module1/#/id/co-01)

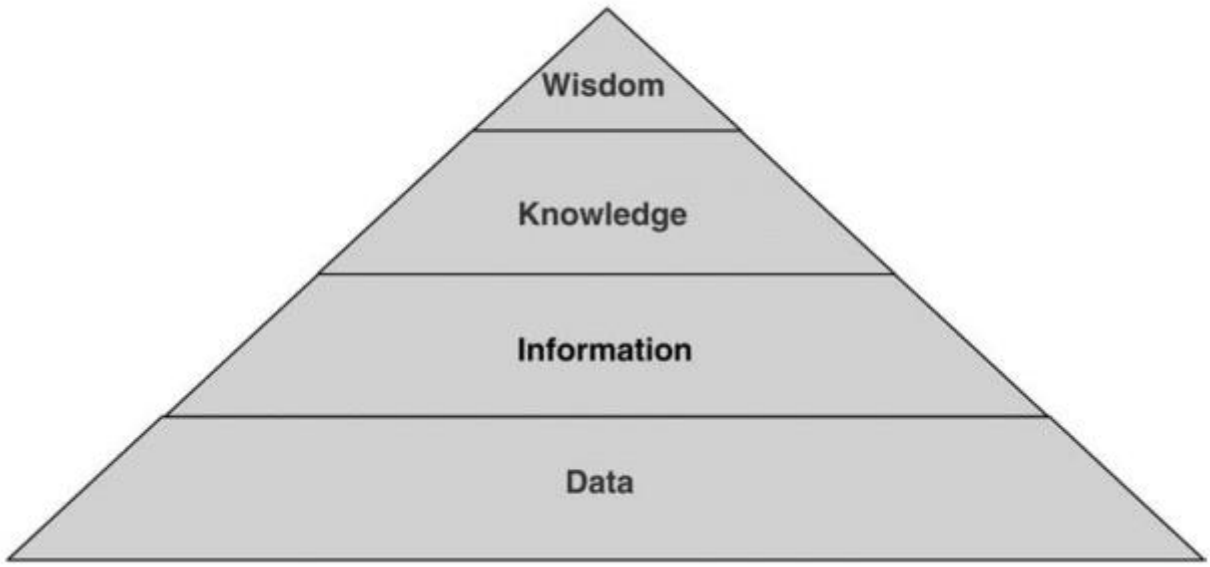
Duomenys yra žaliava, iš kurios galima gauti informacijos ir žinių. Įsivaizduokite duomenis kaip vietoves, vaizdus, aprašymus, atsiliepimus ir kainas, kurie sudaro informacijos, galinčios padėti planuoti, pavyzdžiui, atostogas, pagrindą.

Duomenys tampa **informacija**, kai jiems suteikiamas kontekstas. Remiantis ankstesniu pavyzdžiu, vietos, vaizdai, aprašymai ir kainos – visa tai gali padėti pateikti informaciją, susijusią su turistų lankomomis vietomis. Duomenų rinkimas ir pateikimas padeda formuoti informaciją.

Žinios yra tai, kas gaunama iš informacijos ir pritaikoma jūsų poreikiams. Žinių kaupimas – tai procesas, kurio metu informacija paverčiama pasirinkimais. Remdamiesi ankstesniu pavyzdžiu, žinodami, kad niekas iš jūsų šeimos narių nemėgsta pramogų parkų, galėsite nuspręsti, kurių vietų atostogų metu vengti ir kurios labiau tiktų jūsų šeimai.

DIKW piramidė

Šaltinis: Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of information science*, *33*(2), 163-180.



- **Duomenys**
  - Neturi reikšmės ar vertės, nes jie neturi konteksto ir interpretacijos.
  - Atskiri, objektyvūs faktai arba stebėjimai, kurie yra neorganizuoti ir neapdoroti, ir neperteikia jokios konkrečios prasmės
  - Elementarus ir užfiksuotas daiktų, įvykių, veiklos ir sandorių aprašymas.
  - Dažnai yra didesnių fizinių sistemų elementai (pvz. prietaisų skydeliai), kurie suteikia užuominų apie tai, kokius duomenis reikia pastebėti ir kaip juos skaityti.

Informacija

- Suformatuoti duomenys, kuriuos galima apibrėžti kaip tikrovės atvaizdavimą.
- Duomenys, kurie suteikia pridėtinės vertės dalyko supratimui.

- Duomenys, kurie buvo sutvarkyti taip, kad jie turi prasmę ir vertę gavėjui.
- Duomenys, apdoroti siekiant tam tikro tikslo.

Žinios

- Duomenų ir informacijos derinys, prie kurio pridedama ekspertų nuomonė, įgūdžiai, ir patirtis, kad būtų gautas vertingas turtas, kuriuo galima naudotis priimant sprendimus.
- Tai duomenys ir (arba) informacija, kurie buvo susisteminti ir apdoroti, siekiant perteikti supratimą, patirtį, sukauptą mokymąsi ir kompetenciją, taikomus dabartinei problemai ar veiklai.
- Žinios grindžiamos informacija, kuri išgaunama iš duomenų. Nors duomenys yra daiktų savybė, žinios yra žmonių savybė, kuri juos skatina veikti tam tikru būdu.

Procesai, kurių metu informacija paverčiama į žinias apibūdinami įvairiai:

- daugelio informacijos šaltinių sintezė per tam tikrą laiką;
- įsitikinimų struktūrizavimas;
- studijos ir patirtis;
- organizavimas ir apdorojimas, siekiant perteikti supratimą, patirtį, sukauptą mokymąsi ir patirtį.

Žinios yra:

- kontekstinės informacijos, vertybių, patirties ir taisyklių derinys;
- informacija, ekspertų nuomonė, įgūdžiai ir patirtis;
- informacija kartu su supratimu ir gebėjimais;
- suvokimas, įgūdžiai, mokymas(is), sveikas protas ir patirtis.

Išmintis (įžvalgos)

- Išmintis yra labai neapčiuopiama sąvoka. Ji labiau susijusi su žmogaus intuicija, supratimu, aiškinimu ir veiksmais nei su sistemomis.
- Išmintis laikoma sukauptomis žiniomis, kurios leidžia suprasti, kaip taikyti vienos srities sąvokas naujoms situacijoms ar problemoms.
- Išmintis yra aukščiausias abstrakcijos lygis, pasižymintis įžvalgumu ir gebėjimu matyti už horizonto.
- Tai gebėjimas kritiškai arba praktiškai elgtis bet kurioje situacijoje.

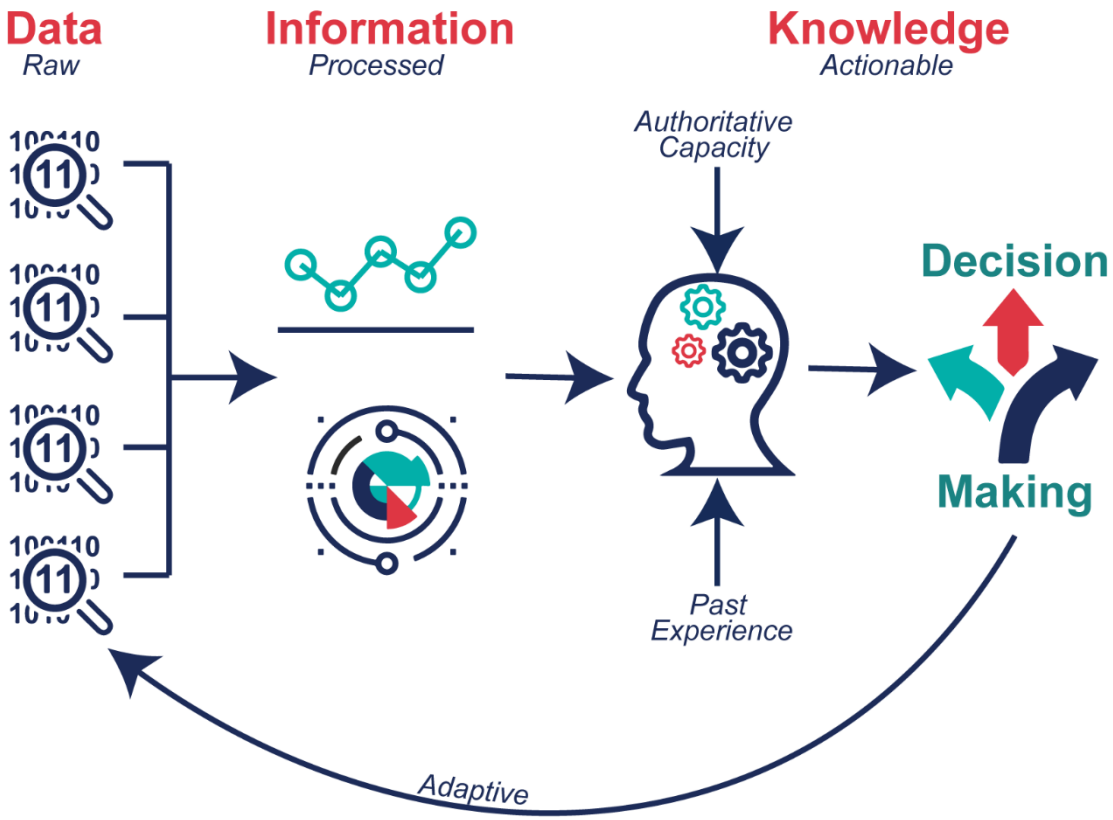
Ši palyginti nedidelė išminties sąvokos aptarimo apimtis rodo, kad platesnėje informacinių sistemų, žinių vadybos ir vadybos literatūroje diskusijoms apie išminties prigimtį ir jos ugdymo būdus skiriama nedaug dėmesio.

Šaltinis: https://internetofwater.org/valuing-data/what-are-data-information-and-knowledge/

Duomenys – tai neapdorotos vertės (skaitinės, tekstinės ir pan.), gautos taikant įvairius duomenų gavimo metodus.

Informacija sukuriama, kai duomenys apdorojami, organizuojami arba struktūrizuojami, siekiant suteikti kontekstą ir prasmę. Informacija iš esmės yra apdoroti duomenys.

Žinios yra tai, ką mes žinome. Kiekvieno žmogaus žinios yra unikalios, tai yra sukaupta ankstesnė patirtis ir įžvalgos, kuriomis remdamiesi aiškiname informaciją ir suteikiame jai prasmę. Kad žinios virstų veiksmais, asmuo turi turėti įgaliojimus ir gebėjimus priimti ir įgyvendinti sprendimą. Žinios (ir įgaliojimai) reikalingos tam, kad būtų gauta informacija, kuri gali turėti poveikį.



Duomenų srautas į informaciją ir žinias nėra vienakryptis. Gautos žinios gali atskleisti surinktų duomenų perteklių ar trūkumą. Dėl to, siekiant geriau patenkinti naudotojų poreikius, gali tekti keisti surinktus duomenis arba tai, kaip šie duomenys paverčiami informacija.

Duomenų, informacijos ir žinių charakteristikos

Duomenys	Informacija	Žinios
Objektyvūs	Turi būti objektyvi	Subjektyvios
Neturi prasmės	Turi prasmę	Turi prasmę, skirtą specifiniam tikslui
Neapdoroti	Apdoroti	Apdoroti ir suprasti
Kiekybiškai įvertinami, gali būti duomenų perteklius	Kiekybiškai įvertinama, gali būti informacijos perteklius	Kiekybiškai neįvertinamos, negali būti pertekliaus

Pavyzdžiai, kuriuose duomenys susiejami su informacija ir žiniomis

- Duomenys: Oro temperatūra yra 25 laipsniai. Informacija: Šiandien yra karšta diena. Žinios: Per karštas oras gali sukelti dehidrataciją ir kitus sveikatos sutrikimus.
- Duomenys: 60% žmonių, gyvenančių miesto centre, naudojasi viešuoju transportu. Informacija: Miesto centro gyventojai labiau linkę naudoti viešąjį transportą nei kitos vietos gyventojai. Žinios: Viešojo transporto sistema miesto centre yra gerai išvystyta, o tai rodo, kad galima sumažinti privačių transporto priemonių naudojimą ir šiltnamio efektą.

- Duomenys: Statistika rodo, kad 70% vaikų turi savo mobilųjį telefoną. Informacija: Daugelis vaikų turi mobilųjį telefoną. Žinios: Vaikų elgesys ir poreikiai pasikeitė, o tai reiškia, kad reikia atnaujinti mokyklų technologijų planus ir svarbu padidinti interneto saugumą, kad apsaugotume vaikus nuo įvairių pavojų.
- Duomenys: Didelis skaičius studentų pasirinko inžinerijos specialybes. Informacija: Inžinerijos studijos populiarėja tarp studentų. Žinios: Dėl didėjančio technologijų ir pramonės sektoriaus vystymosi, inžinerijos specialistams yra didelis poreikis ir jie turi galimybę pasiekti aukštus atlyginimus ir pasirinkti iš įvairių karjeros kelių.

Skyrelio 1.1. medžiagai įtvirtinti siūlomas testas su pasirenkamaisiais atsakymais

Testas (https://www.proprofs.com/quiz-school/quizshow.php?title=data-information-knowledge&q=1**)**

1. _____ gali būti paveikslėliai, skaičiai, tekstas be konteksto
A) Duomenys
B) Informacija
C) Žinios
D) Išmintis (wisdom)
2. Kai duomenys tampa reikšmingais gavėjui, tai vadinama _____
A) Duomenų rinkiniu
B) Informacija
C) Žiniomis
D) Tinkamo atsakymo nėra
3. Kuris iš žemiau pateiktų teiginių yra teisingas?
A) Informacija yra duomenų dalis
B) Duomenys yra informacijos dalis
C) Duomenys ir informacija yra priešingos prigimties.
D) Tinkamo atsakymo nėra.
4. Ar teisingas teiginys: „Kad duomenys būtų naudingi, jie turi virsti informacija“?
A) Taip
B) Ne
5. Kuriuos iš žemiau išvardintų pavyzdžių galima laikyti duomenimis?
A) Datos
B) Kainos
C) Išlaidos
D) Visi išvardinti pavyzdžiai yra duomenys
6. Kurie iš žemiau išvardintų pavyzdžių yra informacija?
A) Ataskaita
B) Receptas
C) Asmens tapatybės kortelė
D) Visi išvardinti pavyzdžiai yra informacija

1.2. Duomenų tyryba

https://microsoft.github.io/Data-Science-For-Beginners/#/1-Introduction/01-defining-data-science/README

Duomenų tyryba apibrėžiama kaip mokslo sritis, kuri naudoja mokslinius metodus žinioms ir įžvalgoms iš struktūrizuotų ir nestruktūrizuotų duomenų išgauti, taip pat žinioms ir įžvalgoms iš duomenų taikyti įvairiose gyvenimo srityse.

Šiame apibrėžime pabrėžiami šie svarbūs duomenų tyrybos aspektai:

Pagrindinis duomenų tyrybos tikslas yra išgauti žinias iš duomenų, kitaip tariant, suprasti duomenis, rasti paslėptus ryšius ir sukurti modelį.

Duomenų tyryba naudoja mokslinius metodus, pavyzdžiui, tikimybių ir statistikos. Tiesą sakant, kai pirmą kartą buvo pradėtas vartoti terminas duomenų tyryba, kai kurie žmonės teigė, kad duomenų tyryba tėra naujas madingas statistikos pavadinimas. Šiandien tapo akivaizdu, kad ši sritis yra daug platesnė.

Gautos žinios turėtų būti pritaikytos tam, kad būtų gauta tam tikrų praktiškai pritaikomų įžvalgų, t. y. praktinių įžvalgų, kurias galima pritaikyti įvairiose situacijose.

Turėtume gebėti dirbti ir su struktūrizuotais, ir su nestruktūrizuotais duomenimis. Prie skirtingų duomenų tipų aptarimo grįšime vėliau.

Taikymo sritis yra svarbi sąvoka, ir duomenų mokslininkams dažnai reikia turėti bent tam tikrų žinių apie probleminę sritį, pavyzdžiui: finansai, medicina, rinkodara ir pan.

Kitas svarbus duomenų tyrybos aspektas yra tai, kad ji tiria, kaip duomenis galima rinkti, saugoti ir valdyti naudojant kompiuterius. Statistika mums suteikia matematinius pagrindus, o duomenų tyryba taiko matematines sąvokas, kad iš tikrųjų iš duomenų būtų galima gauti įžvalgų.

Jim Gray teigia, kad duomenų tyrybą galima laikyti atskira mokslo paradigma, kuri yra:

- **Empirinė**, kurioje daugiausia remiamasi stebėjimais ir eksperimentų rezultatais.
- **Teorinė**, kai naujos sąvokos atsiranda iš esamų mokslo žinių.
- **Kompiuterinė**, kai naujus principus atrandame remdamiesi tam tikrais kompiuteriniais eksperimentais.
- **Duomenimis grindžiama**, pagrįsta ryšių ir dėsningumų atradimu duomenyse.

https://www.javatpoint.com/data-mining

Informacijos gavybos procesas, kurio metu iš didžiulių duomenų rinkinių nustatomi dėsningumai, tendencijos ir naudingi duomenys, leidžiantys verslui priimti duomenimis pagrįstus sprendimus, vadinamas **duomenų tyryba**.

Kitaip tariant, galima sakyti, kad **duomenų tyryba** – tai paslėptų informacijos modelių tyrimo procesas, kurio metu įvairiais aspektais informacija klasifikuojama į naudingus duomenis, kurie renkami ir kaupiami tam tikrose srityse, pavyzdžiui, duomenų saugyklose, veiksminga analizė, duomenų gavybos algoritmas, padedantis priimti sprendimus ir kitus duomenų reikalavimus, kad galiausiai būtų galima sumažinti išlaidas ir gauti pajamų.

Duomenų tyryba – tai automatinė didelių informacijos sankaupų paieška siekiant rasti tendencijas ir dėsningumus, kurie pranoksta paprastas analizės procedūras. Duomenų tyryba naudoja sudėtingus matematinius algoritmus duomenų segmentams ir įvertina būsimų įvykių tikimybę. Duomenų tyryba taip pat vadinama duomenų žinių atradimu (angl. Knowledge Discovery of Data, KDD).

Duomenų tyryba – tai procesas, kurį organizacijos naudoja siekdamos iš didžiulių duomenų bazių išgauti konkrečius duomenis verslo problemoms spręsti. Jis pirmiausia neapdorotus duomenis paverčia naudinga informacija.

Duomenų tyryba yra panaši į duomenų mokslą, kurį vykdo asmuo, konkrečioje situacijoje, konkrečiam duomenų rinkiniui, turėdamas tam tikrą tikslą. Šis procesas apima įvairių rūšių paslaugas, pavyzdžiui, teksto tyrybą, žiniatinklio tyrybą, garso ir vaizdo įrašų tyrybą, vaizdinių duomenų tyrybą ir socialinės žiniasklaidos tyrybą. Tyryba atliekama naudojant paprastą arba labai specifinę programinę įrangą.

Didžiausias iššūkis – išanalizuoti duomenis ir iš jų išgauti svarbią informaciją, kurią galima panaudoti problemai spręsti arba įmonės plėtrai. Yra daug galingų priemonių ir metodų, kuriais galima išgauti duomenis ir rasti iš jų geresnių įžvalgų.

Duomenų tyrybos privalumai

- Duomenų tyrybos metodas leidžia organizacijoms gauti žiniomis pagrįstus duomenis.
- Duomenų tyryba leidžia organizacijoms atlikti pelningus veiklos ir gamybos pakeitimus.
- Palyginti su kitomis statistinių duomenų taikymo sritimis, duomenų tyryba yra ekonomiškai efektyvi.
- Duomenų tyryba padeda organizacijos sprendimų priėmimo procese.
- Ji palengvina automatinį paslėptų dėsningumų atradimą, taip pat tendencijų ir elgsenos prognozavimą.
- Ją galima įdiegti tiek naujoje sistemoje, tiek esamose platformose.
- Tai greitas procesas, kuris naujiems naudotojams leidžia per trumpą laiką išanalizuoti didžiulius duomenų kiekius.

Duomenų tyrybos trūkumai

- Yra tikimybė, kad organizacijos gali parduoti naudingus klientų duomenis kitoms organizacijoms už pinigus. Kaip teigiama pranešime, "American Express" pardavė savo klientų kredito kortelių pirkinius kitoms organizacijoms.
- Daugeliu duomenų tyrybos analitinės programinės įrangos yra sudėtinga naudotis, o darbui su ja reikia išankstinio mokymo.
- Skirtingos duomenų tyrybos priemonės veikia skirtingai dėl jų konstrukcijoje naudojamų skirtingų algoritmų. Todėl tinkamų duomenų tyrybos priemonių pasirinkimas yra labai sudėtinga užduotis.
- Duomenų tyrybos metodai nėra tikslūs, todėl tam tikromis sąlygomis gali sukelti sunkių padarinių.

Duomenų tyrybos sąryšiai su kitais mokslais (sritimis)

https://microsoft.github.io/Data-Science-For-Beginners/#/1-Introduction/01-defining-data-science/README

Duomenų bazės

Svarbiausias aspektas yra **kaip saugoti** duomenis, t. y. kaip juos struktūrizuoti taip, kad būtų galima greičiau apdoroti. Yra įvairių tipų duomenų bazių, kuriose saugomi struktūrizuoti ir nestruktūrizuoti duomenys, kuriuos ir aptarsime savo kurse.

Didieji duomenys

Dažnai mums reikia saugoti ir apdoroti labai didelius duomenų kiekius, kurių struktūra yra palyginti paprasta. Yra specialių metodų ir įrankių, skirtų tiems duomenims saugoti paskirstytu būdu kompiuterių klasteryje ir juos efektyviai apdoroti.

Mašininis mokymasis

Vienas iš būdų suprasti duomenis – **sukurti modelį**, kuris galėtų numatyti norimą rezultatą. Modelių kūrimas iš duomenų vadinamas **mašininiu mokymusi**. Mašininį mokymąsi nagrinėsime skyriuje „Dirbtinis intelektas“.

Dirbtinis intelektas

Mašininio mokymosi sritis, vadinama dirbtiniu intelektu (DI), taip pat remiasi duomenimis ir apima didelio sudėtingumo modelių, imituojančių žmogaus mąstymo procesus, kūrimą. Dirbtinio

intelekto metodai dažnai leidžia nestruktūrizuotus duomenis (pvz., natūralią kalbą) paversti struktūrizuotomis įžvalgomis.

Vizualizavimas

Didžiuliai duomenų kiekiai žmogui yra nesuprantami, tačiau, sukūrę naudingas vizualizacijas, naudodami šiuos duomenis, galime juos geriau suprasti ir padaryti tam tikras išvadas. Taigi, svarbu žinoti daugybę informacijos vizualizavimo būdų. Duomenų vizualizavimui bus skirtas atskiras skyrius.

Duomenų tyrybos taikymai
https://www.javatpoint.com/data-mining

Duomenų tyrybą visų pirma naudoja organizacijos, turinčios intensyvius vartotojų poreikius – mažmeninės prekybos, komunikacijos, finansų, rinkodaros bendrovės, nustatyti kainą, vartotojų pageidavimus, produkto pozicionavimą ir poveikį pardavimams, klientų pasitenkinimą ir įmonės pelną. Duomenų tyryba leidžia mažmeninės prekybos įmonei naudoti pardavimo vietoje esančius įrašus apie klientų pirkinius ir kurti produktus bei akcijas, kurios padeda organizacijai pritraukti klientą.

Sveikatos priežiūros srityje

Duomenų tyryba sveikatos priežiūroje turi puikių galimybių pagerinti sveikatos sistemą. Joje naudojami duomenys ir analizė, siekiant geresnių įžvalgų ir nustatyti geriausią praktiką, kuri pagerintų sveikatos priežiūros paslaugas ir sumažintų išlaidas. Analitikai naudoja tokius duomenų tyrybos metodus kaip mašininis mokymasis, daugiamatė duomenų bazė, duomenų vizualizavimas, minkštieji skaičiavimai ir statistika. Duomenų tyryba gali būti naudojama kiekvienos kategorijos pacientams prognozuoti. Šios procedūros užtikrina, kad pacientai gautų intensyvią priežiūrą tinkamoje vietoje ir tinkamu laiku. Duomenų tyryba taip pat leidžia sveikatos priežiūros draudikams atpažinti sukčiavimą ir piktnaudžiavimą.

Atliekant rinkos krepšelio analizę

Rinkos krepšelio analizė yra hipoteze pagrįstas modeliavimo metodas. Jei perkate tam tikrą produktų grupę, tai labiau tikėtina, kad pirksite ir kitą produktų grupę. Šis metodas gali leisti mažmenininkui suprasti pirkėjo pirkimo elgseną. Šie duomenys gali padėti mažmenininkui suprasti pirkėjo reikalavimus ir atitinkamai pakeisti prekių parduotuvėje išdėstymą. Naudojant kitą analitinį metodą galima palyginti rezultatus tarp skirtingų parduotuvių, tarp skirtingų demografinių grupių pirkėjų.

Švietime

Duomenų tyryba švietime yra naujai besiformuojanti sritis, susijusi su metodų, kuriais tiriamos žinios iš duomenų, gautų švietimo aplinkoje, kūrimu. Pripažįstama, kad EDM tikslai yra patvirtinti būsimą mokinių mokymosi elgseną, tirti švietimo pagalbos poveikį ir skatinti mokymąsi mokytis. Organizacija gali naudoti duomenų tyrybą siekdama priimti tikslius sprendimus, taip pat numatyti mokinio rezultatus. Turėdama rezultatus, įstaiga gali susitelkti į tai, ko ir kaip mokyti.

Gamybos inžinerijoje

Žinios yra geriausias gamybos įmonės turimas turtas. Duomenų tyrybos įrankiai gali būti naudingi norint rasti dėsningumus sudėtingame gamybos procese. Duomenų tyryba gali būti naudojama sistemos lygmens projektavime, siekiant gauti sąsajas tarp gaminio architektūros, gaminių portfelio ir klientų duomenų poreikių. Be kitų užduočių, ji taip pat gali būti naudojama prognozuojant gaminio kūrimo laikotarpį, sąnaudas ir lūkesčius.

CRM (ryšių su klientais valdymo) srityje

Ryšių su klientais valdymas (CRM) yra susijęs su klientų pritraukimu ir išlaikymu, klientų lojalumo didinimu ir į klientus orientuotų strategijų įgyvendinimu. Norėdama užmegzti tinkamus santykius su klientais, verslo organizacija turi rinkti duomenis ir juos analizuoti. Naudojant duomenų tyrybos technologijas, surinktus duomenis galima naudoti analizei.

Nustatant sukčiavimą

Dėl sukčiavimo prarandama milijardai dolerių. Tradiciniai sukčiavimo aptikimo metodai reikalauja daug laiko ir yra sudėtingi. Duomenų tyryba suteikia prasmingus modelius ir paverčia duomenis informacija. Ideali sukčiavimo aptikimo sistema turėtų apsaugoti visų naudotojų duomenis. Prižiūrimus metodus sudaro pavyzdinių įrašų rinkimas, o šie įrašai klasifikuojami kaip sukčiavimo arba nesukčiavimo. Naudojant šiuos duomenis sudaromas modelis, o pagal šį metodą nustatoma, ar dokumentas yra apgaulingas, ar ne.

Nustatant melą

Nusikaltėlio sulaikymas nėra didelė problema, tačiau išaiškinti tiesą yra labai sudėtinga užduotis. Teisėsaugos institucijos gali naudoti duomenų tyrybos metodus nusikaltimams tirti, įtariamų teroristų ryšiams stebėti ir pan. Šis metodas taip pat apima teksto tyrybą ir juo ieškoma prasmingų dėsningumų duomenyse, kurie paprastai yra nestruktūruotas tekstas. Lyginama ankstesnių tyrimų metu surinkta informacija ir sudaromas melo aptikimo modelis.

Bankininkystėje

Bankų sistemos skaitmeninimas turėtų generuoti milžinišką duomenų kiekį, kai atliekama kiekviena nauja operacija. Duomenų tyrybos metodai gali padėti bankininkams sprendžiant su verslu susijusias bankininkystės ir finansų problemas, nustatant tendencijas, atsitiktinumus ir sąsajas verslo informacijoje ir rinkos kaštus, kurie vadovams ar vadybininkams nėra iš karto akivaizdūs, nes duomenų apimtis yra per didelė arba juos ekspertai per greitai sukuria ekrane. Vadovas gali rasti šiuos duomenis, kad galėtų geriau orientuotis į pelningą klientą, jį įsigyti, išlaikyti, segmentuoti ir išlaikyti.

Duomenų tyrybos įgyvendinimo iššūkiai

https://www.javatpoint.com/data-mining

Nors duomenų tyryba yra labai galinga, ją vykdant susiduriama su daugybe iššūkių. Įvairūs iššūkiai gali būti susiję su našumu, duomenimis, metodais, technikomis ir kt.

Neišsamūs ir triukšmingi duomenys

Duomenų tyryba – tai naudingų duomenų išgavimo iš didelio kiekio duomenų procesas. Duomenys realiame pasaulyje yra nevienalyčiai, neišsamūs ir triukšmingi. Didžiuliais kiekiais pateikti duomenys paprastai būna netikslūs arba nepatikimi. Šios problemos gali kilti dėl duomenų matavimo prietaiso arba dėl žmogiškųjų klaidų. Tarkime, mažmeninės prekybos tinklas renka klientų, kurie išleidžia daugiau nei 500 JAV dolerių, telefono numerius, o apskaitos darbuotojai šią informaciją įtraukia į savo sistemą. Įvesdamas telefono numerį žmogus gali padaryti klaidą, todėl duomenys bus neteisingi. Net kai kurie klientai gali nenorėti atskleisti savo telefono numerių, todėl duomenys būna neišsamūs. Duomenys gali būti pakeisti dėl žmogaus ar sistemos klaidos. Visos šios pasekmės (triukšmingi ir neišsamūs duomenys) apsunkina duomenų tyrybą.

Duomenų pasiskirstymas

Realaus pasaulio duomenys paprastai saugomi įvairiose platformose paskirstytosios kompiuterijos aplinkoje. Jie gali būti duomenų bazėje, atskirose sistemose arba net internete. Praktiškai visus

duomenis perkelti į centralizuotą duomenų saugyklą yra gana sudėtinga užduotis, daugiausia dėl organizacinių ir techninių priežasčių. Pavyzdžiui, įvairūs regioniniai biurai gali turėti savo serverius duomenims saugoti. Neįmanoma visų biurų duomenis saugoti centriniame serveryje. Todėl duomenų tyrybai reikia sukurti priemones ir algoritmus, kurie leistų išgauti paskirstytus duomenis.

Sudėtingi duomenys

Tai gali būti daugialypės terpės duomenys, įskaitant garso ir vaizdo įrašus, vaizdus, sudėtingus duomenis, erdvinius duomenis, laiko eilutes ir pan. Šių įvairių tipų duomenų valdymas ir naudingos informacijos išgavimas yra sudėtinga užduotis. Dažniausiai, norint gauti konkrečią informaciją, reikėtų tobulinti naujas technologijas, naujas priemones ir metodikas.

Veikimas

Duomenų tyrybos sistemos našumas visų pirma priklauso nuo naudojamų algoritmų ir metodų efektyvumo. Jei sukurtas algoritmas ir metodai neatitinka reikalavimų, tai neigiamai paveiks duomenų tyrybos proceso efektyvumą.

Duomenų privatumas ir saugumas

Duomenų tyryba paprastai sukelia rimtų problemų, susijusių su duomenų saugumu, valdymu ir privatumu. Pavyzdžiui, jei mažmenininkas analizuoja informaciją apie įsigytas prekes, jis atskleidžia duomenis apie pirkėjų pirkimo įpročius ir pageidavimus be jų sutikimo.

Duomenų vizualizavimas

Duomenų tyrybos srityje duomenų vizualizavimas yra labai svarbus procesas, nes tai yra pagrindinis metodas, kuriuo išvestis vartotojui parodoma patraukliai. Duomenys turi tiksliai perteikti tai, ką jais norima išreikšti. Tačiau daug kartų tiksliai ir paprastai pateikti informaciją galutiniam naudotojui yra sudėtinga. Kadangi įvesties duomenys ir išvesties informacija yra sudėtingi, norint, kad tai pavyktų, reikia įgyvendinti labai veiksmingus ir sėkmingus duomenų vizualizavimo procesus.

1.3. Duomenų tipai

Skiriami 3 pagrindiniai duomenų tipai: **struktūrizuoti, pusiau struktūrizuoti** ir **nestruktūrizuoti**. Toliau pateikiami įvairių duomenų tipų apibrėžimai iš įvairių šaltinių.

Struktūrizuoti duomenys

- Duomenys, kurių lengviausia ieškoti ir tvarkyti, nes jie paprastai pateikiami lentelėse, o jų elementus galima suskirstyti į fiksuotus iš anksto apibrėžtus laukus, vadinami struktūrizuotais duomenimis. Pagalvokite, kokius duomenis galite saugoti "Excel" skaičiuoklėje, ir turėsite struktūrizuotų duomenų pavyzdį. Struktūrizuoti duomenys gali atitikti duomenų bazės dizainerio sukurtą duomenų modelį - prisiminkite pardavimo įrašus pagal regionus, produktus ar klientus. Struktūrizuotuose duomenyse esybės gali būti sugrupuotos ir sudaryti ryšius („klientai“, kurie taip pat yra „patenkinti paslauga“). Dėl to struktūrizuotus duomenis lengva saugoti, analizuoti ir ieškoti, todėl dar visai neseniai jie buvo vieninteliai įmonėms lengvai pritaikomi duomenys. Šiandien, daugumos vertinimu, struktūrizuoti duomenys sudaro mažiau nei 20 proc. visų duomenų. Dažnai struktūrizuoti duomenys tvarkomi naudojant struktūrizuotų užklausų kalbą (SQL) - programavimo programinės įrangos kalbą, kurią praėjusio amžiaus septintajame dešimtmetyje sukūrė IBM, skirtą reliacinėms duomenų bazėms.

  Struktūrizuotus duomenis gali kurti mašinos ir žmonės. Struktūrizuotų duomenų pavyzdžiai - finansiniai duomenys, pavyzdžiui, apskaitos operacijos, adresų duomenys, demografinė informacija, klientų žvaigždučių įvertinimai, mašinų žurnalai, išmaniųjų telefonų ir išmaniųjų įrenginių vietos nustatymo duomenys ir kt.

(https://www.forbes.com/sites/bernardmarr/2019/10/18/whats-the-difference-between-structured-semi-structured-and-unstructured-data/?sh=6727d9952b4d).

- Struktūrizuoti duomenys, paprastai priskiriami kiekybiniams duomenims, yra labai gerai organizuoti ir lengvai iššifruojami mašininio mokymosi algoritmais. 1974 m. IBM sukurta struktūrizuotų užklausų kalba (SQL) yra programavimo kalba, naudojama struktūrizuotiems duomenims tvarkyti. Naudodami reliacinę (SQL) duomenų bazę, verslo naudotojai gali greitai įvesti, ieškoti ir tvarkyti struktūrizuotus duomenis (https://www.ibm.com/cloud/blog/structured-vs-unstructured-data).
- Struktūrizuoti duomenys yra kiekybiniai, labai gerai organizuoti ir lengvai analizuojami naudojant duomenų analizės programinę įrangą. Jie formatuojami į sistemas, kurios turi taisyklingą dizainą, telpa į nustatytas eilutes, stulpelius ir lenteles. Struktūrizuotų užklausų kalba (SQL) yra standartinė kalba, naudojama bendrauti su duomenų baze ir ypač naudinga tvarkant struktūrizuotus duomenis. Naudojama duomenų paieškai, pridėjimui, atnaujinimui ir pašalinimui, be kita ko, SQL palengvina struktūrizuotų duomenų tvarkymą. Pagalvokite apie viešbučio duomenų bazę, kurioje galima ieškoti svečių pagal vardą, telefono numerį, kambario numerį ir pan. Arba brūkšninius kodus, naudojamus produktams tvarkyti ir klasifikuoti gamybos, platinimo ir pardavimo vietose.
  Struktūrizuoti duomenys paprastai saugomi reliacinėse duomenų bazėse (RDBMS). Duomenų bazėse esančią informaciją gali įvesti žmonės arba mašinos, o paiešką galima lengvai atlikti pagal rankiniu būdu įvestas užklausas arba algoritmus. Struktūrizuotiems duomenims saugoti ir tvarkyti taip pat naudojamos programos, pavyzdžiui, "Excel", kurias galima lengvai sujungti su kitomis analitinėmis priemonėmis tolesnei analizei atlikti. Struktūrizuoti duomenys puikiai tinka pagrindiniam organizavimui ir kiekybiniams skaičiavimams, tačiau turi atitikti griežtus, iš anksto nustatytus parametrus. Struktūrizuotų duomenų pavyzdžiai – tai duomenys, kurių lengva ieškoti pagal nustatytą struktūrą ir kuriuos galima susieti su kitomis duomenų bazėmis. Galite atlikti paiešką pagal klientų adresą, kad sužinotumėte, kurie produktai yra populiariausi tam tikroje vietovėje, arba išsiaiškinti, kuriuos produktus daug kartų užsako keli klientai (https://monkeylearn.com/blog/structured-data-vs-unstructured-data/).

Pusiau struktūrizuoti duomenys

- Pusiau struktūrizuotiems duomenims priskiriami duomenys turi tam tikrų apibrėžiančių ar nuoseklių savybių, tačiau neatitinka tokios griežtos struktūros, kokios tikimasi iš reliacinės duomenų bazės. Todėl yra tam tikrų organizacinių savybių, pavyzdžiui, semantinių žymų arba metaduomenų, kad būtų lengviau juos organizuoti, tačiau duomenys vis tiek yra kintantys.

  Geras pavyzdys yra el. pašto pranešimai. Nors tikrasis turinys yra nestruktūrizuotas, jame yra struktūrizuotų duomenų, pavyzdžiui, siuntėjo ir gavėjo vardas ir el. pašto adresas, išsiuntimo laikas ir pan. Kitas pavyzdys - skaitmeninė nuotrauka. Pats vaizdas yra nestruktūrizuotas, bet jei nuotrauka buvo padaryta, pavyzdžiui, išmaniuoju telefonu, joje būtų nurodyta data ir laikas, geografinė žyma ir įrenginio ID. Išsaugojus nuotrauką, jai taip pat galėtų būti suteiktos žymos, kurios suteiktų struktūrą, pavyzdžiui, „šuo“ arba „augintinis“.

  Daugelis duomenų, kuriuos žmonės paprastai priskiria nestruktūrizuotiems duomenims, iš tiesų yra pusiau struktūrizuoti, nes juose yra tam tikrų klasifikavimo požymių (https://www.forbes.com/sites/bernardmarr/2019/10/18/whats-the-difference-between-structured-semi-structured-and-unstructured-data/?sh=6727d9952b4d).

- Pusiau struktūrizuoti duomenys (pvz., JSON, CSV, XML) yra "tiltas" tarp struktūrizuotų ir nestruktūrizuotų duomenų. Jie neturi iš anksto nustatyto duomenų modelio ir yra sudėtingesni už struktūrizuotus duomenis, tačiau juos lengviau saugoti nei nestruktūrizuotus duomenis.

Pusiau struktūrizuoti duomenys naudoja metaduomenis (pvz., žymas ir semantines žymes) konkrečioms duomenų savybėms nustatyti ir duomenims suskirstyti į įrašus ir iš anksto nustatytus laukus. Metaduomenys galiausiai leidžia pusiau struktūrizuotus duomenis geriau kataloguoti, ieškoti ir analizuoti nei nestruktūrizuotus duomenis.

Metaduomenų naudojimo pavyzdys: internetiniame straipsnyje rodoma antraštė, ištrauka, rodomas paveikslėlis, paveikslėlio alt-tekstas, slug ir t. t., kurie padeda atskirti vieną žiniatinklio turinio dalį nuo panašių dalių.

Pusiau struktūrizuotų duomenų ir struktūrizuotų duomenų pavyzdys: Klientų duomenų failas su lentelėmis ir duomenų bazė su CRM lentelėmis.

Pusiau struktūrizuotų duomenų ir nestruktūrizuotų duomenų pavyzdys: Kliento Instagram komentarų sąrašas (https://www.ibm.com/cloud/blog/structured-vs-unstructured-data).

- Pusiau struktūrizuoti duomenys, kuriuos sudaro daugiausia nestruktūrizuotas tekstas, tačiau juos galima laisvai suskirstyti pagal metažymes. Pavyzdys galėtų būti el. paštas, kuriame galima ieškoti pagal kategorijas Gautieji, Išsiųstieji, Juodraščiai ir t. t. arba socialinė žiniasklaida, kurią galima suskirstyti į kategorijas Draugai, Žinutės, Viešos žinutės, Privačios žinutės ir t. t. Pusiau struktūrizuotus duomenis galima lengvai suskirstyti į iš anksto nustatytas kategorijas, tačiau informacija šiose kategorijose pati savaime yra nestruktūrizuota. Analizuojant el. laiškus, ketinimų klasifikavimas gali būti naudingas automatiškai skaitant verslo el. laiškus pagal kliento ketinimus, kad būtų galima pasakyti, ar jis į užklausą atsako tikrai susidomėjęs, ar ne (https://monkeylearn.com/blog/structured-data-vs-unstructured-data/).

Nestruktūrizuoti duomenys

- Nestruktūrizuoti duomenys – tai duomenys, kurių negalima sutalpinti į eilutės ir stulpelio duomenų bazę ir kurie neturi susijusio duomenų modelio. Pagalvokite apie el. pašto žinutės tekstą. Dėl struktūros nebuvimo nestruktūrizuotus duomenis sunkiau ieškoti, valdyti ir analizuoti, todėl įmonės plačiai atmetė nestruktūrizuotus duomenis, kol pastaruoju metu paplitus dirbtiniam intelektui ir mašininio mokymosi algoritmams juos apdoroti tapo lengviau.

  Nestruktūrizuotų duomenų pavyzdžiai: nuotraukos, vaizdo ir garso failai, tekstiniai failai, socialinės žiniasklaidos turinys, palydoviniai vaizdai, prezentacijos, PDF, atviri apklausos atsakymai, interneto svetainės ir skambučių centrų stenogramos / įrašai. (https://www.forbes.com/sites/bernardmarr/2019/10/18/whats-the-difference-between-structured-semi-structured-and-unstructured-data/?sh=6727d9952b4d).

- Nestruktūrizuotų duomenų, paprastai priskiriamų prie kokybinių duomenų, negalima apdoroti ir analizuoti įprastinėmis duomenų analizės priemonėmis ir metodais. Kadangi nestruktūrizuoti duomenys neturi iš anksto nustatyto duomenų modelio, juos geriausia tvarkyti nereliacinėse (NoSQL) duomenų bazėse. Kitas būdas tvarkyti nestruktūrizuotus duomenis – naudoti duomenų ežerus, kad jie būtų išsaugoti neapdoroti.

  Nestruktūrizuotų duomenų svarba sparčiai didėja. Naujausios prognozės rodo, kad nestruktūrizuoti duomenys sudaro daugiau kaip 80 % visų įmonės duomenų, o 95 % įmonių teikia pirmenybę nestruktūrizuotų duomenų valdymui (https://www.ibm.com/cloud/blog/structured-vs-unstructured-data).

- Nestruktūrizuoti duomenys – tai informacija, kuri neturi nustatytos struktūros ir netelpa į apibrėžtą sistemą. Nestruktūrizuotų duomenų pavyzdžiai yra garso ir vaizdo įrašai, vaizdai ir įvairūs tekstai: ataskaitos, el. laiškai, įrašai socialiniuose tinkluose ir kt. Rasti įžvalgų nestruktūrizuotuose duomenyse nėra lengva, tačiau tinkamai išanalizuoti tekstiniai duomenys gali

būti labai vertingi siekiant išgauti kokybinius rezultatus, pavyzdžiui, klientų nuomones, arba organizuoti verslo duomenis, pavyzdžiui, klientų aptarnavimo bilietus, pagal atskiras kategorijas, kad juos būtų galima nukreipti tinkamam darbuotojui (https://monkeylearn.com/blog/structured-data-vs-unstructured-data/).

Struktūrizuotų, pusiau struktūrizuotų ir nestruktūrizuotų duomenų pavyzdžiai, savybės ir jų palyginimas

1. **Darbo pokalbis** (https://www.forbes.com/sites/bernardmarr/2019/10/18/whats-the-difference-between-structured-semi-structured-and-unstructured-data/?sh=6727d9952b4d)

Kad būtų lengviau suprasti duomenų klasifikavimo skirtumus, pasitelkime šią analogiją. Sakykime, kad vykstant pokalbiui dėl darbo yra trys skirtingos pokalbių klasifikacijos: struktūrizuotas, pusiau struktūrizuotas ir nestruktūrizuotas.

Struktūrizuoto pokalbio metu pokalbio vedėjas vadovaujasi griežtu scenarijumi, kurį nustatė žmogiškųjų išteklių skyrius ir kurio laikomasi su kiekvienu kandidatu. Kita pokalbio forma - nestruktūrizuotas pokalbis. Per nestruktūrizuotą pokalbį interviu vedėjas pats sprendžia, kokius klausimus ir kokia tvarka jie bus užduodami (ar apskritai bus užduodami) kiekvienam kandidatui. Iš dalies struktūrizuotas interviu turi ir struktūrizuoto, ir nestruktūrizuoto interviu elementų. Jame naudojamas nuoseklumas ir kiekybiniai elementai, kuriuos leidžiama naudoti struktūrizuotame pokalbyje, tačiau suteikiama laisvė pritaikyti pagal aplinkybes, kurios labiau atitinka nestruktūrizuotą pokalbį.

Taigi, kalbant apie duomenis, struktūrizuotus duomenis lengva organizuoti ir jie atitinka griežtą formatą; nestruktūruoti duomenys yra sudėtinga ir dažnai kokybinė informacija, kurios neįmanoma redukuoti į reliacinę duomenų bazę ar organizuoti joje, o pusiau struktūrizuoti duomenys turi abiejų elementų.

2. **Įvairių tipų duomenų pavyzdžiai** (https://microsoft.github.io/Data-Science-For-Beginners/#/1-Introduction/01-defining-data-science/README)

Struktūrizuoti	Pusiau struktūrizuoti	Nestruktūrizuoti
Žmonių sąrašas su jų telefono numeriais.	Vikipedijos puslapiai su nuorodomis	Encyclopedia Britannica tekstas
Pastarųjų 20 metų visų pastato patalpų temperatūra kiekvieną minutę.	Mokslinių straipsnių rinkinys JSON formatu su autoriais, publikacijos duomenimis ir santrauka	Failų bendrinimas su įmonių dokumentais
Visų į pastatą įeinančių žmonių amžiaus ir lyties duomenys	Interneto puslapiai	Neapdorotas vaizdo įrašas iš stebėjimo kameros

3. **Struktūrizuotų duomenų pavyzdžiai** (https://www.coursera.org/articles/structured-vs-unstructured-data)
    - **Skrydžio užsakymas**. Skrydžio ir užsakymo duomenys, pvz., datos, kainos ir paskirties vietos, tvarkingai saugomi "Excel" skaičiuoklės formatu. Rezervuojant skrydį ši informacija saugoma duomenų bazėje.
    - **Ryšių su klientais valdymas** (CRM, Customer relationship management). CRM programinė įranga, pavyzdžiui, "Salesforce", struktūrizuotus duomenis apdoroja analitinėmis priemonėmis, kad sukurtų naujus duomenų rinkinius, skirtus įmonėms analizuoti klientų elgseną ir pageidavimus.
4. **Nestruktūrizuotų duomenų pavyzdžiai** (https://www.coursera.org/articles/structured-vs-unstructured-data)
    - **Pokalbių robotai**. Pokalbių robotai suprogramuoti atlikti teksto analizę, kad atsakytų į klientų klausimus ir suteiktų reikiamą informaciją.

- **Rinkos prognozės**. Duomenys gali būti panaudoti siekiant numatyti akcijų rinkos pokyčius, kad analitikai galėtų pakoreguoti savo skaičiavimus ir investicinius sprendimus.

Struktūrizuotų ir nestruktūrizuotų duomenų privalumai ir trūkumai
https://www.ibm.com/cloud/blog/structured-vs-unstructured-data

Duomenų tipas	Privalumai	Trūkumai
Struktūrizuoti	• **Lengvai pritaikomi mašininio mokymosi algoritmai**: Specifinė ir organizuota struktūrizuotų duomenų architektūra palengvina manipuliavimą ML duomenimis ir užklausų teikimą. • **Lengvai pritaikomi versle**. Struktūrizuoti duomenys nereikalauja gilaus supratimo apie skirtingus duomenų tipus ir jų veikimą. Naudotojai, turėdami pagrindinį supratimą apie su duomenimis susijusią temą, gali lengvai pasiekti ir interpretuoti duomenis. • **Galimybė naudoti daugiau įrankių duomenų apdorojimui**. Kadangi struktūrizuoti duomenys pradėti analizuoti anksčiau už nestruktūrizuotus duomenis, yra daugiau įrankių, skirtų naudoti ir analizuoti struktūrizuotus duomenis.	• **Ribotas naudojimas.** Duomenys, kurių struktūra iš anksto nustatyta, gali būti naudojami tik pagal numatytą paskirtį, o tai riboja jų lankstumą ir panaudojimo galimybes. • **Ribotos saugojimo galimybės**: Duomenys paprastai saugomi duomenų saugojimo sistemose, turinčiose griežtas schemas (pvz., duomenų saugyklose). Todėl, pasikeitus duomenų reikalavimams, būtina atnaujinti visus struktūrizuotus duomenis, o tai reikalauja didžiulių laiko ir išteklių sąnaudų.
Nestruktūrizuoti	• **Pirminis formatas.** Nestruktūrizuotus duomenis, saugomus pirminiu formatu, kol jų prireikia, galima laikyti neapibrėžtus. Dėl jo pritaikomumo duomenų bazėje padaugėja failų formatų, o tai praplečia duomenų fondą ir leidžia duomenų mokslininkams rengti ir analizuoti tik tuos duomenis, kurių jiems reikia. • **Greita kaupimo sparta.** Kadangi duomenų nereikia iš anksto apibrėžti, juos galima kaupti greitai ir lengvai. • **Duomenų ežero saugykla***. Leidžia naudoti didžiulę saugyklą ir mokėti už naudojimąsi, o tai mažina išlaidas ir leidžia lanksčiai keisti duomenų dydį.	• **Reikalinga patirtis**. Dėl neapibrėžto/nesuformuoto duomenų pobūdžio, norint parengti ir analizuoti nestruktūrizuotus duomenis, reikia duomenų mokslo žinių. Tai naudinga duomenų analitikams, tačiau sudaro sunkumų naudotojams, kurie gali ne iki galo suprasti specializuotas duomenų temas arba tai, kaip panaudoti savo duomenis. • **Specializuotos priemonės**. Duomenų tvarkytojams reikia specializuotų įrankių nestruktūrizuotiems duomenims tvarkyti, o tai apriboja produktų pasirinkimą.

*Duomenų ežero saugykla (Data lake storage): https://www.valdas.blog/2019/03/31/duomenu-ezeras/#palyginkime-duomen%C5%B3-saugykl%C4%85-su-duomen%C5%B3-e%C5%BEeru

Struktūrizuotų ir nestruktūrizuotų duomenų pagrindiniai skirtumai
https://www.ibm.com/cloud/blog/structured-vs-unstructured-data

Struktūrizuoti (kiekybiniai) duomenys leidžia pamatyti klientus „iš paukščio skrydžio“, o nestruktūrizuoti (kokybiniai) duomenys padeda geriau suprasti klientų elgseną ir ketinimus. Panagrinėkime kai kuriuos pagrindinius skirtumus ir jų reikšmę:

Šaltiniai: Struktūrizuoti duomenys gaunami iš GPS jutiklių, internetinių formų, tinklo žurnalų, žiniatinklio serverių žurnalų, OLTP sistemų ir t. t., o nestruktūrizuotų duomenų šaltiniai yra el. pašto pranešimai, tekstų apdorojimo dokumentai, PDF failai ir kt.

Formos: Struktūrizuotus duomenis sudaro skaičiai ir reikšmės, o nestruktūrizuotus duomenis – jutikliai, tekstiniai failai, garso ir vaizdo failai ir kt.

Modeliai: Struktūrizuoti duomenys turi iš anksto nustatytą duomenų modelį ir prieš patalpinant į duomenų saugyklą yra suformatuojami pagal nustatytą duomenų struktūrą (pvz., schema įrašant), o nestruktūrizuoti duomenys saugomi savo prigimtiniu formatu ir neapdorojami tol, kol nėra naudojami (pvz., schema skaitant).

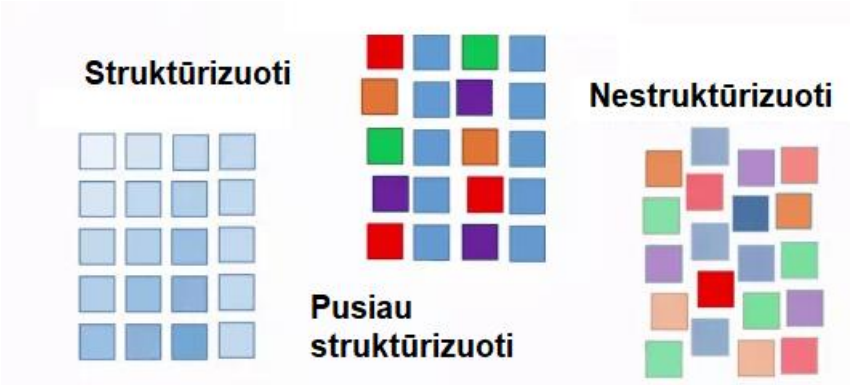
Saugojimas: Struktūrizuoti duomenys saugomi lentelių formatais (pvz., skaičiuoklių lentelėse arba SQL duomenų bazėse), kuriems saugoti reikia mažiau vietos. Juos galima saugoti duomenų saugyklose, todėl juos galima labai lengvai glaudinti. Kita vertus, nestruktūruoti duomenys saugomi kaip medijos failai arba NoSQL duomenų bazės, kurioms reikia daugiau vietos. Juos galima saugoti duomenų ežeruose, todėl juos sunku glaudinti.

Naudojimo būdai: Struktūrizuotus duomenis naudoja mašininis mokymasis (ML) ir jų algoritmai, o nestruktūrizuotus duomenis - natūralios kalbos apdorojimas (NLP) ir teksto gavyba.

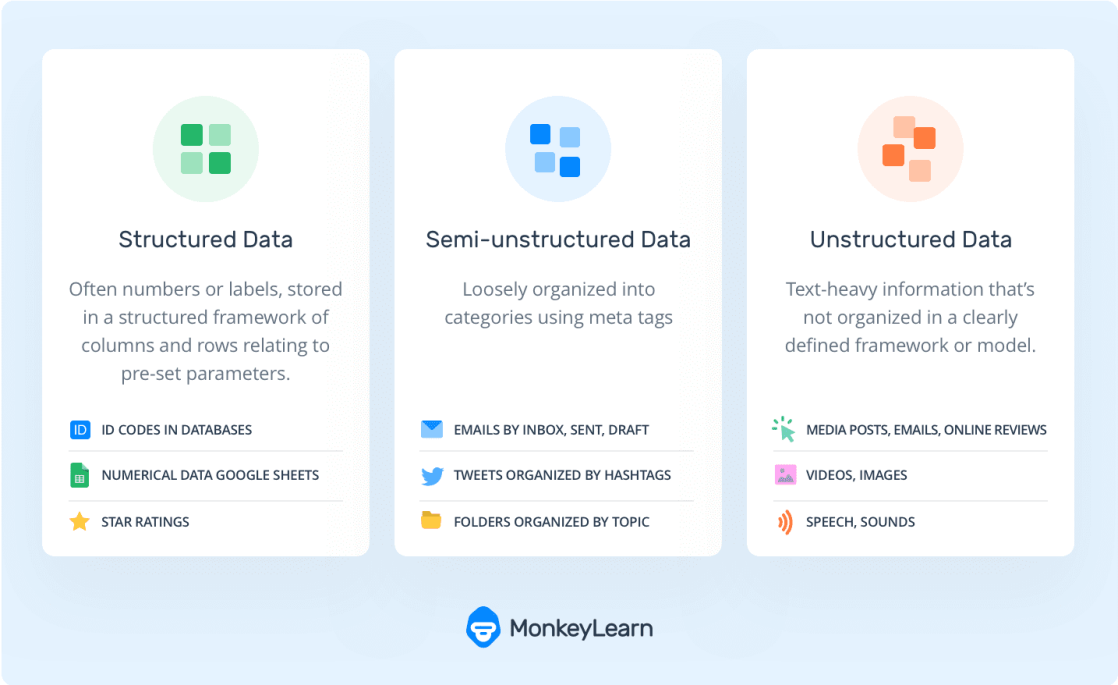
Iš šaltinio: (https://monkeylearn.com/blog/structured-data-vs-unstructured-data/)

Struktūrizuoti duomenys	**Nestruktūrizuoti duomenys**
Struktūrizuoti duomenys yra kiekybiniai ir dažnai pateikiami kaip skaičiai, datos, reikšmės ir eilutės.	Nestruktūrizuotieji duomenys yra kokybiniai duomenys, apimantys tekstą, vaizdo įrašus, garso įrašus, vaizdus ir kt.
Struktūrizuoti duomenys saugomi eilutėmis ir stulpeliais.	Nestruktūrizuoti duomenys saugomi kaip garso, teksto ir vaizdo failai arba NoSQL duomenų bazės.
Apytikriai 20 proc. verslo duomenų.	Apytikriai 80-90 % verslo duomenų (ir jų vis daugėja!).
Randami uždaro tipo apklausose, pavyzdžiui, NPS ir CSAT rezultatuose, ir žiniatinklio analizėje.	Klientų aptarnavimo užklausose, atsiliepimuose internete, el. pašto žinutėse, atviruose klausimuose ir kt.
Saugomi duomenų saugyklose.	Saugomi taikomosiose programose, NoSQL (nereliacinėse) duomenų bazėse, duomenų ežeruose ir duomenų saugyklose.
Atskleidžia modelius ir tendencijas, kurie parodo, kas vyksta.	Atskleidžia modelius ir tendencijas, kurie paaiškina, kodėl kažkas vyksta.
Reikalauja mažiau vietos saugykloje.	Reikia daugiau vietos saugykloje.
Lengva analizuoti naudojant tokias priemones kaip skaičiuoklė.	Sunku analizuoti be dirbtinio intelekto įrankių.

Įvairių tipų duomenų vaizdavimas



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



Šaltinis: https://monkeylearn.com/blog/structured-data-vs-unstructured-data/

Duomenų šaltiniai

https://microsoft.github.io/Data-Science-For-Beginners/#/1-Introduction/01-defining-data-science/README

1. **Struktūrizuotų duomenų šaltiniai**
   - Daiktų internetas (IoT), įskaitant įvairių jutiklių, pavyzdžiui, temperatūros ar slėgio ir kt., duomenis, suteikia daug naudingų duomenų. Pavyzdžiui, jei namuose ar mokykloje įrengti daiktų interneto jutikliai, galime automatiškai valdyti šildymą ir apšvietimą, kad sumažintume išlaidas.
   - Apklausos, kurias prašome naudotojų užpildyti po pirkimo arba apsilankius interneto svetainėje.
   - Elgesio analizė gali, pavyzdžiui, padėti suprasti, kaip giliai naudotojas įeina į svetainę ir kokia yra tipinė svetainės palikimo priežastis.
2. **Nestruktūrizuotų duomenų šaltiniai**
   - Tekstai gali būti turtingas įžvalgų šaltinis, pavyzdžiui, bendras nuotaikų įvertinimas arba raktinių žodžių ir semantinės reikšmės išskyrimas.

- Vaizdai arba vaizdo įrašai. Stebėjimo kameros vaizdo įrašas gali būti naudojamas siekiant įvertinti eismo intensyvumą kelyje ir informuoti žmones apie galimas spūstis.
- Interneto serverio žurnalai gali būti naudojami siekiant suprasti, kurie mūsų svetainės puslapiai lankomi dažniausiai ir kiek laiko.

3. **Pusiau struktūrizuotų duomenų šaltiniai**
   - Socialinių tinklų grafikai gali būti puikus duomenų apie naudotojų asmenybes ir galimą informacijos sklaidos veiksmingumą šaltinis.
   - Kai turime krūvą nuotraukų iš vakarėlio, galime pabandyti išgauti grupės dinamikos duomenis sudarydami žmonių, fotografuojančių vienas kitą, grafiką.
   - Žinodami įvairius galimus duomenų šaltinius, galite pabandyti pagalvoti apie įvairius scenarijus, kuriuose galima taikyti duomenų tyrybos metodus, kad geriau pažintumėte situaciją ir patobulintumėte verslo procesus.

Veiksmai su duomenimis
https://microsoft.github.io/Data-Science-For-Beginners/#/1-Introduction/01-defining-data-science/README

1. **Duomenų rinkimas**

Pirmasis žingsnis – surinkti duomenis. Nors daugeliu atvejų tai gali būti nesudėtingas procesas, pavyzdžiui, duomenys, gaunami į duomenų bazę iš žiniatinklio programos, kartais turime naudoti specialius metodus. Pavyzdžiui, iš daiktų interneto jutiklių gaunamų duomenų gali būti per daug, todėl prieš tolesnį apdorojimą pravartu naudoti buferizavimo galinius taškus, pavyzdžiui, IoT Hub, kad būtų galima surinkti visus duomenis.

2. **Duomenų saugojimas**

Saugoti duomenis gali būti sudėtinga, ypač jei kalbame apie didelius duomenis. Nusprendžiant, kaip saugoti duomenis, tikslinga numatyti, kokiu būdu ateityje norėsite pateikti užklausą apie duomenis. Duomenis galima saugoti keliais būdais:

- Reliacinėje duomenų bazėje saugomas lentelių rinkinys, o užklausoms atlikti naudojama speciali SQL kalba. Paprastai lentelės yra suskirstytos į skirtingas grupes, vadinamas schemomis. Daugeliu atvejų duomenis reikia konvertuoti iš pradinės formos, kad jie atitiktų schemą.
- NoSQL duomenų bazės, pavyzdžiui, CosmosDB, duomenims schemų neįtvirtina ir leidžia saugoti sudėtingesnius duomenis, pavyzdžiui, hierarchinius JSON dokumentus ar grafikus. Tačiau NoSQL duomenų bazės neturi plačių SQL užklausų galimybių ir negali užtikrinti referencinio vientisumo, t. y. duomenų struktūros lentelėse ir ryšių tarp lentelių taisyklių.
- Duomenų ežero saugykla naudojama didelėms neapdorotų, nestruktūrizuotų duomenų kolekcijoms saugoti. Duomenų ežerai dažnai naudojami dideliems duomenims, kai visi duomenys netelpa viename kompiuteryje, todėl juos reikia saugoti ir apdoroti serverių klasteryje. Parquet yra duomenų formatas, kuris dažnai naudojamas kartu su dideliais duomenimis.

3. **Duomenų apdorojimas**

Tai žingsnis, apimantis duomenų konvertavimą iš pradinės formos į formą, kurią galima naudoti vizualizavimui ir (arba) modelių mokymui. Dirbant su nestruktūrizuotais duomenimis, pavyzdžiui, tekstu ar vaizdais, gali tekti naudoti tam tikrus dirbtinio intelekto metodus, kad iš duomenų būtų galima išgauti požymius ir taip juos paversti struktūrizuota forma.

4. **Vizualizavimas / žmogaus įžvalgos**

Dažnai, norėdami suprasti duomenis, turime juos vizualizuoti. Turėdami daug įvairių vizualizavimo metodų savo įrankių rinkinyje galime rasti tinkamą atvaizdavimą, kad padarytume įžvalgą. Dažnai duomenų mokslininkui reikia „žaisti su duomenimis“, daug kartų juos vizualizuojant ir ieškant tam tikrų sąsajų. Taip pat galime naudoti statistinius metodus, kad patikrintume hipotezes arba įrodytume ryšį tarp skirtingų duomenų.

5. Prognozavimo modelio mokymas

Kadangi galutinis duomenų mokslo tikslas – gebėti priimti sprendimus remiantis duomenimis, galime norėti pasinaudoti mašininio mokymosi metodais, kad sukurtume prognozavimo modelį. Paskui jį galėsime naudoti prognozėms atlikti naudodami naujus panašios struktūros duomenų rinkinius.

Žinoma, priklausomai nuo faktinių duomenų, kai kurių žingsnių gali nebūti (pavyzdžiui, kai duomenis jau turime duomenų bazėje arba kai mums nereikia modelio mokymo) arba kai kurie žingsniai gali būti kartojami kelis kartus (pavyzdžiui, duomenų apdorojimas).

Skaitmenizacija ir skaitmeninė transformacija
(https://microsoft.github.io/Data-Science-For-Beginners/#/1-Introduction/01-defining-data-science/README

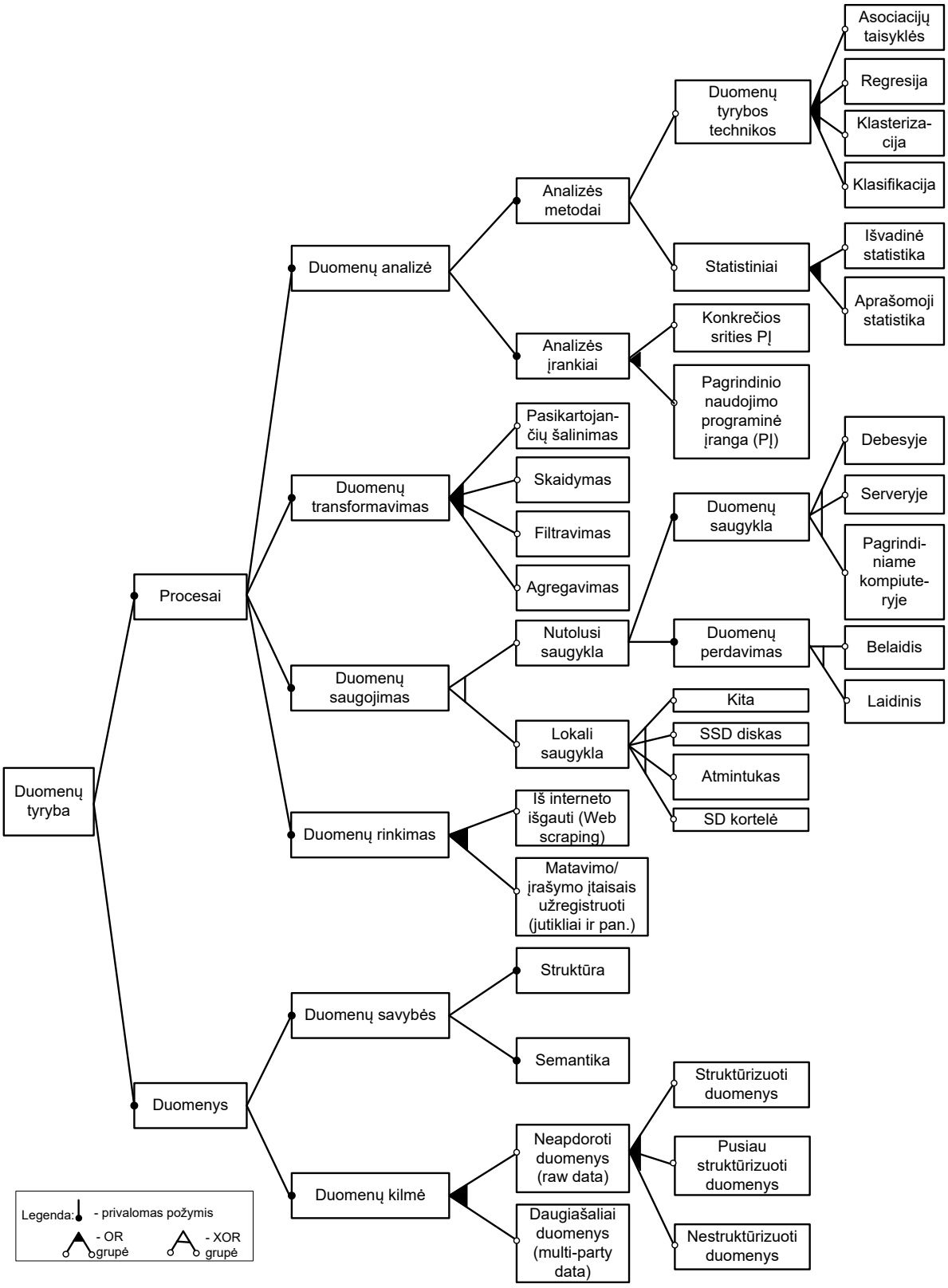
Pastarąjį dešimtmetį daugelis įmonių pradėjo suprasti duomenų svarbą priimant verslo sprendimus. Norint taikyti duomenų tyrybos principus valdant verslą, pirmiausia reikia surinkti tam tikrus duomenis, t. y. verslo procesus perkelti į skaitmeninę formą. Tai vadinama skaitmeninimu. Duomenų tyrybos metodų taikymas šiems duomenims, kuriais remiantis priimami sprendimai, gali lemti reikšmingą produktyvumo padidėjimą (ar net verslo posūkį), vadinamą skaitmenine transformacija.

Panagrinėkime pavyzdį. Tarkime, turime duomenų tyrybos kursą (kaip šis) ir norime jį patobulinti pasitelkdami duomenų tyrybą. Kaip tai galime padaryti?

Galime pradėti nuo klausimo „Ką galima skaitmenizuoti?“. Paprasčiausias būdas būtų matuoti laiką, per kurį kiekvienas mokinys įveikia kiekvieną pamoką, ir įvertinti įgytas žinias pateikiant testą su keliais atsakymų variantais kiekvienos pamokos pabaigoje. Vidutiniškai įvertinę visų mokinių užimamą laiką, galime išsiaiškinti, kurios pamokos mokiniams kelia daugiausia sunkumų, ir jas supaprastinti.

Pradėję analizuoti testų su keliais atsakymų variantais rezultatus, galime pabandyti nustatyti, kurias sąvokas mokiniai sunkiai supranta, ir panaudoti šią informaciją turiniui tobulinti. Kad tai padarytume, turime sudaryti testus taip, kad kiekvienas klausimas atitiktų tam tikrą sąvoką ar žinių dalį.

1.4. Duomenų tyrybos modelis



Šaltinis: Stuikys, V., Burbaite, R., Ziberkas, G., & Kubiliunas, R. (2022). Development and Evaluation of an Approach for Integrating Data Science Concepts into High School STEM Curriculum. *INTERNATIONAL JOURNAL OF ENGINEERING EDUCATION*, *38*(3), 756-773.