Mokslas – Lietuvos ateitis / Science – Future of Lithuania
ISSN 2029-2341 / eISSN 2029-2252
2018 Volume 10, Article ID: mla.2018.932, 1–10
https://doi.org/10.3846/mla.2018.932

Economics and Management
Ekonomika ir vadyba

DIDŽIŲJŲ DUOMENŲ NAUDOJIMAS KLIENTUI PAŽINTI

Simona POLITAITĖ*, Jolanta SABAITYTĖ

Vilniaus Gedimino technikos universitetas, Vilnius, Lietuva

Gauta 2018 m. balandžio 9 d.; priimta 2018 m. gegužės 11 d.

Santrauka. Į klientus orientuotoje rinkoje klientų elgsenos supratimas yra svarbus veiksnys, lemiantis organizacijos sėkmę. Organizacija, siekianti išlikti ir sėkmingai egzistuoti, negali ignoruoti nuolat didėjančių duomenų kiekių – didžiųjų duomenų. Didieji duomenys – sudėtingi duomenų masyvai, kuriuos sunku apdoroti naudojant tradicines duomenų apdorojimo programas. Optimaliai išanalizuoti tokie duomenys suteikia galimybę geriau pažinti klientus, tobulinti sprendimų priėmimo procesą, didinti konkurencinį pranašumą. Organizacijai svarbu suprasti, kaip panaudoti didžiuosius duomenis, kokias apdorojimo priemones ir modelius taikyti. Šiame straipsnyje analizuojamos didžiųjų duomenų koncepcijos ir raida, naudojimo rizikos, gavybos būdai ir taikomi modeliai. Taikomi šie metodai: mokslinių šaltinių sisteminė, loginė analizė, informacijos sugretinimas, sisteminimas.

Reikšminiai žodžiai: didieji duomenys, kliento pažinimas, didžiųjų duomenų analizė, naudojimo rizikos, duomenų tyryba, duomenų valdymas.

Įvadas

Susiformavus žinių visuomenei informacija ir tikslingas jos panaudojimas bei apdorojimas tapo tolimesnio visuomenės vystymosi pagrindu. Informacija išgaunama iš duomenų. Dėl to atsirado poreikis rinkti ir apdoroti didelius duomenų kiekius – didžiuosius duomenis. Sąvoka „didieji duomenys“ – reprezentuoja naujas technologijas, skirtas duomenims, kurie generuojami dideliu greičiu, dideliais kiekiais bei yra įvairios struktūros, apdoroti (Lee, 2017). Optimaliai apdoroti didieji duomenys organizacijoms sukuria palankią informacinę terpę, kuri leidžia pažinti klientą. Dėl to didžiųjų duomenų tyrimams daug dėmesio skiria ne tik akademinė bendruomenė, bet ir verslo atstovai (Marr, 2015; Gartner, 2017; Langkafel, 2016; Beyer ir Laney, 2012; Deloitte, 2015).

Nuolatinį duomenų kiekių augimą lemia technologiniai pokyčiai – didėjanti automatizacija ir augantis internetu sąveikaujančių įrenginių (angl. *Internet of Things*) naudojimas. Kas sekundę visame pasaulyje sukuriama apie 1,7 MB naujų duomenų (Marr, 2015). Šiuose duomenyse užkoduota informacija apie kliento elgseną, interesus ir poreikius. Dėl to vadybininkai patiria nemažą spaudimą, remiantis didžiųjų duomenų analize, identifikuoti klientų požiūrį ir elgesį lemiančius veiksnius.

Optimaliai išanalizuoti duomenys didina įmonės konkurencinį pranašumą. Todėl, neretai didieji duomenys įvardijami kaip vienas strategiškai svarbiausių išteklių XXI amžiuje, svarba prilygstantis auksui ir naftai (Alharthi, Krotov ir Bowman, 2017). Taip pat didieji duomenys tapatinami su skaitmenine revoliucija, kuri, manoma, iš esmės pakeis tai, kaip konkuruoja ir veikia verslo įmonės. Dabar vykstantys pokyčiai apima didžiuosius duomenis įmonės viduje ir išorėje: struktūruotus ir nestruktūruotus, kompiuterinius, internetinius ir mobiliuosius duomenis, kurie leidžia pateikti objekto istorinę apžvalgą ir ateities prognozes.

Didieji duomenys plačiai taikomi įvairiose tradicinėse verslo ir mokslo srityse: prekyboje, medicinoje, fizikoje, bet evoliucionuojant technologijoms atsirado naujų taikymo galimybių (Langkafel, 2016; Dippel, 2017; Jinjiang ir Xueling, 2016). Pastaruoju metu mokslinėje literatūroje aptarinėjamas didžiųjų duomenų naudojimas neuromarketingo srityje (Stanton, Sinnott-Armstrong ir Huettel, 2017; Boksem ir Smidts, 2015). Šioje srityje didieji duomenys naudojami kūno ir smegenų siunčiamiems signalams užfiksuoti, juos išanalizavus galima suprasti, ką sąmoningai ir nesąmoningai mąsto klientas. Taip pat didžiųjų duomenų analizė plačiai taikoma vystant naujus produktus (Hua Tan ir Zhan, 2016; Zhan, Tan, Li ir Tse, 2016).

*Autorius susirašinėti. El. paštas *simona.politaite@stud.vgtu.lt*

Nors duomenų rinkimas tampa vis lengviau prieinamas, organizacijos susiduria su daugybe iššūkių. Duomenų rinkimas nėra pagrindinė kliūtis. Svarbu žinoti, kaip surinktus duomenis analizuoti, kad būtų gaunama naudinga informacija. Tai reikalauja naujų darbuotojų įgūdžių tobulinimo ir IT infrastruktūros plėtojimo, naujos valdymo praktikos įdiegimo arba naujos organizacinės kultūros visoje įmonėje (Manyika et al., 2011). Svarbu pabrėžti, kad ne visos įmonės sugeba prisitaikyti ir optimaliai panaudoti didžiuosius duomenis. Tai patvirtina ir A. McAfee ir E. Brynjolfsson (2012) atliktas tyrimas, kurio metu buvo nagrinėjama 330-ies organizacijų, įsikūrusių Šiaurės Amerikoje, organizacinė ir technologinė veikla. Buvo nustatyta, kad daugelis jų nebuvo pasirengusios panaudoti didžiuosius duomenis organizacinei veiklai gerinti (McAfee ir Brynjolfsson, 2012). Dėl to kyla esminė problema – mokslinėje literatūroje stokojama modelių, skirtų didiesiems duomenims klasifikuoti ir naudoti siekiant pažinti klientą. Šio darbo objektas – didieji duomenys. Darbo tikslas – atsižvelgiant į didžiųjų duomenų naudojimo poreikius, išanalizuoti tokių duomenų klasifikavimo metodus, taikomus klientui pažinti. Siekiant šio tikslo, iškeliami tokie uždaviniai:

– Išanalizuoti didžiųjų duomenų koncepcijas ir raidą.
– Identifikuoti didžiųjų duomenų naudojimo rizikas ir sritis.
– Palyginti skirtingus didžiųjų duomenų analizės metodus.

Darbe taikomi šie tyrimų metodai: mokslinių šaltinių sisteminė, loginė analizė, informacijos gretinimas, sisteminimas ir grafinis atvaizdavimas.

1. Didžiųjų duomenų koncepcijos ir raida

Pažanga informacinių ir komunikacinių technologijų srityje (toliau – IKT) bei didėjantis internetu sąveikaujančių įrenginių naudojimo mastas lėmė nuolatinį atskirų asmenų ir organizacijų generuojamų duomenų kiekių augimą. Dėl to kasdien sukuriama daugybė struktūruotų ir nestruktūruotų duomenų, kuriuos sunku surinkti, valdyti ir analizuoti naudojant esamą IT infrastruktūrą ir priemones. Tokie duomenys vadinami didžiaisiais duomenimis. Surinkti ir išanalizuoti didieji duomenys suteikia galimybę gauti naujų įžvalgų, kurios leidžia įmonei pažinti klientą ir šitaip padidinti konkurencinį pranašumą, todėl dauguma įmonių kaupia didžiuosius duomenis.

Teigiama, kad sąvoka „didieji duomenys“ pirmą kartą pavartota 1998 m. įmonės „Silicon Graphics“ atstovo Johno Mashley pristatymo metu (Diebold, 2012), tačiau dažniau ji pradėta vartoti pastarąjį dešimtmetį. Nors sąvoka šiuo metu plačiai vartojama mokslinėje literatūroje (Alharthi et al., 2017; Sharma, 2015; Gartner, 2017), bendro jos apibrėžimo nėra. Skirtingi autoriai šią sąvoką formuluoja remdamiesi skirtingais aspektais. Pavyzdžiui, A. Alharthi ir kiti (2017) didžiuosius duomenis siūlo traktuoti kaip didžiulį žmogaus veiklos sukurtų, skaitmeninių duomenų kiekį, kurį sudėtinga valdyti įprastomis duomenų analizės priemonėmis. Kiti autoriai didžiųjų duomenų sąvokai apibrėžti taiko 3V modelį, kuris ankstyvuosiuose didžiųjų duomenų vystymosi etapuose apibrėžė didžiųjų duomenų koncepciją. 3V modelį sudaro trys pagrindinės didžiųjų duomenų charakteristikos: kiekis (angl. *Volume*), greitis (angl. *Velocity*) ir įvairovė (angl. *Variety*) (Laney, 2001):

– Kiekis. Ši charakteristika nurodo duomenų kiekį, kurį surenka arba sukuria individas ar organizacija (Lee, 2017). Todėl galima teigti, kad kiekis nurodo duomenų dydį. Didieji duomenys – tai įvairių formatų duomenys, matuojami nebe gigabaitais, o kur kas didesniais vienetais – petabaitais, zetabaitais ar eksabaitais. Kadangi duomenų saugyklų pajėgumai nuolat auga ir tai leidžia kaupti vis didesnius duomenų rinkinius, duomenų kiekio apibrėžimai skiriasi priklausomai nuo veiksnių, tokių kaip laikas ir duomenų tipas (Gandomi ir Haider, 2015). Vadinasi, tai, kas šiandien laikoma didžiaisiais duomenimis, ateityje gali būti nebetraktuojama kaip didieji duomenys.
– Greitis nurodo duomenų generavimo ir apdorojimo spartą (Sharma, 2015). Iš pradžių organizacijos, analizuodamos duomenis, naudojo paketų apdorojimo sistemas (kai atveriamas dokumentas, jam apdoroti iškviečiama atitinkama programa), dėl to duomenų apdorojimo procesas buvo lėtas ir brangus (Lee, 2017). Tačiau laikui bėgant išaugęs duomenų greitis ir skaitmeninių įrenginių, tokių kaip išmanieji telefonai ir jutikliai, naudojimas šį procesą pagreitino. Dėl to įvairioms organizacijoms tapo paprasta atlikti didžiųjų duomenų analizes realiuoju laiku.
– Įvairovė nurodo duomenų tipų skaičių. Duomenys surenkami iš įvairių šaltinių: skaičiuoklių, duomenų bazių, tekstinių dokumentų, skaitmeninių duomenų srautų (Gartner, 2017). Dėl to didieji duomenys įvardijami kaip skirtingų formatų duomenų masyvai. Šie masyvai sudaryti iš struktūruotų, pusiau struktūruotų ir nestruktūruotų duomenų. Struktūruoti duomenys yra skaičiuoklėse ar reliacinėse duomenų bazėse, tai lentelių duomenys; tekstas, nuotraukos, garso ir vaizdo įrašai – nestruktūrizuotų duomenų pavyzdžiai; XML (angl. *Extensible Markup Language*) dokumentų ženklinimo kalba, skirta dokumentų struktūrai aprašyti, yra tipinis pusiau struktūruotų duomenų pavyzdys (Gandomi ir Haider, 2015).

Remiantis 3V modeliu, didieji duomenys apibrėžiami kaip didelės apimties, įvairovės ir didelio greičio informacinis turtas, skirtas sprendimų priėmimo procesui pagerinti bei reikalaujantis ekonomiškai efektyvių ir novatoriškų informacijos apdorojimo formų (Beyer ir Laney, 2012). Didžiųjų duomenų sąvoka yra tiesiogiai susijusi su IKT, todėl joms vystantis, tuo pat metu vystosi ir *didžiųjų duomenų* sąvoka. Laikui bėgant 3V modelis buvo papildytas dviem naujomis charakteristikomis: teisingumu (angl. *Veracity*) ir verte (angl. *Value*), ir tapo naujo – 5V modelio – pagrindu:

– Teisingumas. Ši charakteristika apibūdina reikalavimus, susijusius su surinktų duomenų bei jų analizės rezultatų teisingumu ir tikslumu (Zakir, 2015). Nepa-

tikimi, neteisingi duomenys sukuria papildomų kliūčių priimant sprendimus bei neigiamai veikia visoje organizacijoje vykstančius procesus.
– Vertė. Iš sukauptų duomenų galima gauti naujų įžvalgų apie tiriamą procesą ar reiškinį. Naujos įžvalgos didina organizacijos efektyvumą ir konkurencingumą, todėl šiais laikais duomenys laikomi ekonomiškai vertingu ištekliumi (Kaur et al., 2016).

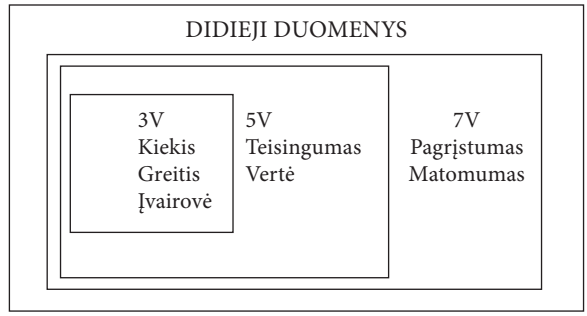
Anot A. De Mauro ir kitų (2014), didieji duomenys reprezentuoja informacinį turtą, kuris charakterizuojamas dideliu kiekiu, greičiu bei įvairove. Norint juos transformuoti į vertę, reikia specifinių technologijų ir analizės metodų. Pabrėžtina, kad svarbų vaidmenį formuodamos 5V modelį atliko IBM ir kitos technologinės kompanijos, kurios investavo į didžiųjų duomenų analizės rinką. Būtent IBM kompanija pasiūlė teisingumo charakteristiką (Lee, 2017).

5V modelį sudarančios charakteristikos pradėtos laikyti gairėmis, į kurias buvo atsižvelgiama sprendžiant problemas, susijusias su didžiųjų duomenų kūrimu, apdorojimu ir valdymu (U. G. Gupta ir A. Gupta, 2016). 2015 m. A. Gandomi ir M. Haider atliktas tyrimas parodė, kad dauguma technologinių kompanijų, apibrėždamos didžiųjų duomenų sąvoką, remiasi 5V arba jo pirmtaku 3V modeliu. Tačiau J. S. Ward ir A. Barker (2013) didžiuosius duomenis apibrėžia kaip sąvoką, apibūdinančią didelių ir (arba) sudėtingų duomenų rinkinių talpinimą ir analizę, taikant daugybę įvairių metodų: NoSQL, MapReduce ir mašininį mokymąsi (angl. *Machine learning*).

Siekiant tikslesnio didžiųjų duomenų sąvokos apibrėžimo, pasiūlytas 7V modelis (Khan et al., 2014; Seddon ir Currie, 2017). Formuojant šį modelį, pridėtos pagrįstumo (angl. *Validity*) ir matomumo (angl. *Visibility*) charakteristikos:
– Pagrįstumas. Ši charakteristika neretai tapatinama su teisingumo charakteristika, tačiau jos skiriasi. Pagrįstumu apibūdinamas duomenų logiškumas, atitiktis faktams (Khan et al., 2014). Yra tikrinama, ar duomenys turi tokią prasmę, kuri jiems priskiriama.
– Matomumas apibūdina duomenis per laiko ir patikimumo prizmę – autorizuotas naudotojas turi prieigą prie duomenų, gali laiku juos rasti, tvarkyti ir patikimai saugoti (Gandomi ir Haider, 2015; Simon, 2014; Walker et al., 2013). Matomumas sukuria kontekstą, kuris suteikiamas naudotojui – turėdamas duomenų rinkinį, naudotojas matys ir kitus susijusius duomenis: laiką, šaltinį, procesus (U.G. Gupta ir A. Gupta, 2016). Be matomumo prarandama galimybė prisijungti prie duomenų ir juos analizuoti.

Galima teigti, kad 3V, 5V ir 7V modeliai didžiųjų duomenų sąvoką apibrėžia siaurąją prasme – kaip objektą, kurio bruožai ilgainiui kinta ir yra aiškinami pagal nustatytas charakteristikas (1 paveikslas).

Be to, šių modelių evoliucija parodė, kad didžiųjų duomenų sąvoka sparčiai vystosi. Tačiau tokie staigūs pokyčiai kognityvine prasme sukėlė daug painiavos. Tai patvirtina ir „Harris Interactive“ įmonės SAP užsakymu 2012 m. atliktas 154 įmonių vadovų tyrimas. Tyrimo metu įmonių



1 paveikslas. Didžiųjų duomenų modeliai (sudaryta autorių, remiantis Laney, 2001; Zakir, 2015; Kaur et al., 2016; Khan et al., 2014; Seddon ir Currie, 2017)
Figure 1. Models of big data (composed by the authors based on Laney, 2001; Zakir, 2015; Kaur et al., 2016; Khan et al., 2014; Seddon & Currie, 2017)

vadovų buvo prašoma apibrėžti didžiųjų duomenų sąvoką. 28 % didžiuosius duomenis įvardijo kaip didelį transakcinių duomenų augimą; 24 % – kaip naują technologiją, skirtą spręsti problemoms, susijusioms su dideliu duomenų kiekiu, greičiu ir įvairove; 19 % teigė, kad didieji duomenys apibrėžia reikalavimus, kaip kaupti ir archyvuoti duomenis; 18 % traktavo didžiuosius duomenis kaip naujus duomenų šaltinius, tokius kaip socialinės medijos, mobilieji įrenginiai (SAP, 2012). Tyrimas parodė, kaip skirtingi įmonių vadovai nevienodai interpretuoja didžiųjų duomenų sąvoką. Dalis vadovų akcentuoja tai, kas yra didieji duomenys, o kita dalis – ką tokie duomenys daro. Kyla problema – vadovai iš tikrųjų nesupranta didžiųjų duomenų koncepcijos, todėl jie nežino, kaip panaudoti tokius duomenis klientui pažinti.

Taip pat svarbu pabrėžti, kad didžiųjų duomenų sąvoka aiškinama ir plačiąja prasme – kaip reiškinys, kurį veikia įvairūs veiksniai. D. Boyd ir K. Crawford (2012) didžiuosius duomenis apibrėžia kaip kultūrinį, technologinį ir mokslinį reiškinį, kuris remiasi:
– technologijomis: renkant, analizuojant, susiejant ir lyginant didelius duomenų rinkinius maksimizuojama skaičiavimo galia ir algoritmais pagrįstas tikslumas;
– analize: norint nustatyti ekonominius, socialinius, techninius ir teisinius reikalavimus, remiamasi dideliais duomenų rinkiniais;
– mitais: egzistuoja įsitikinimas, kad dideliuose duomenų rinkiniuose užkoduota informacija, iš kurios galima gauti objektyvių ir tikslių įžvalgų, kurių negalima gauti iš kitų šaltinių.

Kiti autoriai koncentruoja dėmesį į verslo analitikos sritį, kuri yra didžiųjų duomenų pasekmė. M. Chen ir kt. (2012), apibrėždami didžiuosius duomenis, kartu identifikuoja ir su jais susijusias technologijas: debesų kompiuteriją (angl. *Cloud computing*), internetu sąveikaujančius įrenginius, ir „Hadoop“ didžiųjų duomenų apdorojimo programinę įrangą. Bendrai autoriai tai traktuoja kaip verslo žvalgybą ir analitiką (angl. *Business Intelligence and Analytics*) ir apibūdina kaip technikas, technologijas, siste-

mas, praktikas, metodologijas ir programas, skirtas verslo duomenų analizei, dėl kurios įmonė galėtų laiku priimti sprendimus ir geriau suprastų savo verslo modelį bei rinką (Chen, Chiang ir Storey, 2012). Kadangi verslo analitikos sritis sparčiai plečiasi ir vis daugiau įmonių renka ir analizuoja didžiuosius duomenis, su didžiaisiais duomenimis pradėtos glaudžiai sieti saugumo ir privatumo sąvokos. Pastaruoju metu kyla daug diskusijų asmens duomenų saugumo ir privatumo klausimais. Remiantis TRUSTe internetu sąveikaujančius įrenginių privatumo indeksus (angl. *TRUSTe Internet of Things Privacy Index*) nustatyta, kad tik 20 % interneto vartotojų mano, kad išmaniųjų įrenginių privalumai viršija bet kokius privatumo klausimus (TRUSTe, 2015). Todėl toliau tikslinga analizuoti didžiųjų duomenų naudojimo rizikas.

2. Didžiųjų duomenų naudojimo rizikos

Siekiant geriau suprasti didžiųjų duomenų naudojimo galimybes, svarbu apibrėžti ir galimas šio reiškinio rizikas, kurių pažinimas sudarys prielaidas efektyvesniam taikymui. Nors didžiuosius duomenis kaupti tampa vis lengviau, 60 % tokių duomenų projektų žlunga (Gartner, 2015). Tai lemia kylančios didžiųjų duomenų rizikos.

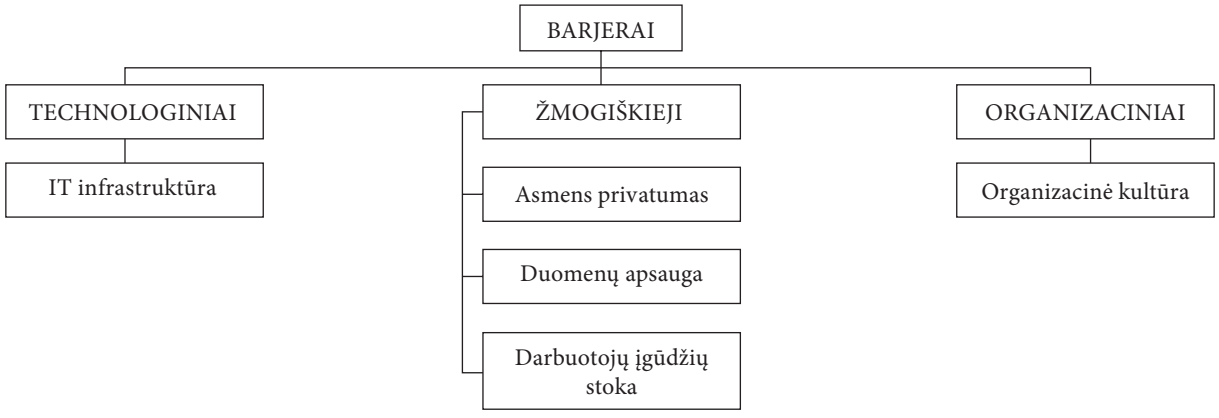
Didieji duomenys tapatinami su pramone 4.0 (S. Wang, J. Wan, Zhang, Li ir Zhang, 2016; Lee, Kao ir Yang, 2014), kuri, manoma, iš esmės pakeis tai, kaip konkuruoja ir operuoja verslo įmonės. Dabar vykstantys pokyčiai apima didžiuosius duomenis įmonės viduje ir išorėje: struktūruotus ir nestruktūruotus, kompiuterinius, internetinius ir mobiliuosius duomenis, kurie leidžia atlikti objekto analizę – pateikti istorinę apžvalgą ir ateities prognozes. Analizės sėkmę lemia potencialių rizikų eliminavimas. Dėl to svarbu identifikuoti ir analizuoti didžiųjų duomenų naudojimo rizikas.

Pradinėje didžiųjų duomenų analizės fazėje susiduriama su rizika, kad gauti rezultatai nebus tikslūs. Tai yra esminė rizika, nagrinėjant didžiųjų duomenų naudojimo rizikas plačiąja prasme. Ši rizika pasireiškia per 5V modelį, kurį sudaro penkios charakteristikos, apibūdinančios didžiuosius duomenis: kiekį, greitį, įvairovę, teisingumą ir vertę; yra nemažai atvejų, kai žinomi analizės metodai negali būti taikomi apdorojant didelius (kiekis), įvairių formatų (įvairovė) duomenų kiekius per priimtiną laiką (mažas greitis), dėl to gaunami netikslūs rezultatai (teisingumas), pagal kuriuos parengiamos klaidingos prognozės (žema vertė) (Krasnow Waterman ir Bruening, 2014). Siekiant tikslesnių rezultatų ir teisingesnių prognozių, didiesiems duomenims tvarkyti turėtų būti naudojamos specialios sistemos ir algoritmai (ur Rehman et al., 2016), kurie garantuotų greitesnį, patikimesnį kompleksiškų duomenų apdorojimą.

Didžiųjų duomenų rizikas nagrinėjant siaurąja prasme, identifikuojami jas sukeliantys technologiniai, žmogiškieji ir organizaciniai barjerai (2 paveikslas).

Technologiniai barjerai apibendrintai apibrėžiami kaip informacinių technologijų (toliau – IT) infrastruktūra. Dauguma dabar naudojamų IT negali atlaikyti augančių didžiųjų duomenų analizės poreikių (Alharthi et al., 2017), kurie sparčiai kinta. Dėl to kyla rizika, kad įmonėje įdiegtos technologijos nebus pritaikytos didžiųjų duomenų analizei. Sumažinti šią riziką padeda techninę įrangą papildančių įrenginių (angl. *Commodity Hardware*) taikymas (Alharthi et al., 2017). Pavyzdžiui, vis augantiems duomenų kiekiams saugoti galima didinti bendrąją duomenų bazės talpą, sujungiant keletą standžiųjų diskų. Įmonei plėtojant esamą IT infrastruktūrą, susiduriama su dar viena rizika – blogu finansinių išteklių paskirstymu. Didžiųjų duomenų analizei skirtos IT infrastruktūros plėtra reikalauja nemažų investicijų į techninę ir programinę įrangą, todėl būtina įvertinti, ar įmonė yra tam finansiškai pasiruošusi. Investicijas naudinga planuoti tarpiniam laikotarpiui, kad būtų galima numatyti naujus didžiųjų duomenų naudojimo atvejus, kurie gali reikalauti naujų duomenų apdorojimo būdų ir priemonių.

Žmogiškieji barjerai siejami su asmens veikla, kuri daro įtaką didžiųjų duomenų naudojimui. Pastaraisiais metais daug dėmesio skiriama duomenų apsaugos ir asmens privatumo klausimams. Pavyzdžiui, vietos nustaty-



2 paveikslas. Didžiųjų duomenų barjerų klasifikacija (sudaryta autorių, remiantis Alharthi et al., 2017)
Figure 2. Classification of big data barriers (composed by the authors based on Alharthi et al., 2017)

mo paslaugų teikėjai gali identifikuoti vartotoją, sekdami informaciją apie jų buvimo vietą, kuri yra galimai susijusi su gyvenamosios ar darbo vietos informacija (Sivarajah, Kamal, Irani ir Weerakkody, 2017). Taip atsiranda rizika pažeisti asmens privatumą. Siekiant eliminuoti šią riziką, teisinis reglamentavimas yra visko pamatas. Leidžiami įstatymai ir kuriami standartai apibrėžia duomenų tvarkymo sąlygas ir sudaro kliūtis atsirasti žmogaus teisių pažeidimams. Didžiųjų duomenų apsaugos aspektu kenkėjiška programinė įranga įvardijama kaip vis auganti grėsmė (Abawajy, Kelarev ir Chowdhury, 2014). Tokios įrangos naudojimas leidžia nutekini duomenis, juos modifikuoti ar panaikinti. Todėl, siekiant apsaugoti sukauptus duomenis, būtina vykdyti saugumo kontrolę, kuri užtikrina kenkėjiškos programinės įrangos prevenciją. Be to, labai svarbu atkreipti dėmesį į darbuotojų, atliekančių didžiųjų duomenų analizę, įgūdžius ir kompetencijas. Daugumą įmonių susiduria su darbuotojų, turinčių didžiųjų duomenų ir jų analizės įgūdžių, stoka (Tole, 2013). Taip kyla pavojus duomenų saugumui. Taip pat duomenų analizė atliekama nekokybiškai, o gaunami rezultatai būna klaidingi. Todėl, ugdant didžiųjų duomenų specialistus, būtinas glaudus bendradarbiavimas tarp švietimo įstaigų ir verslo įmonių, užtikrinant teorinio pasirengimo atitiktį praktiniams poreikiams (Miller, 2014).

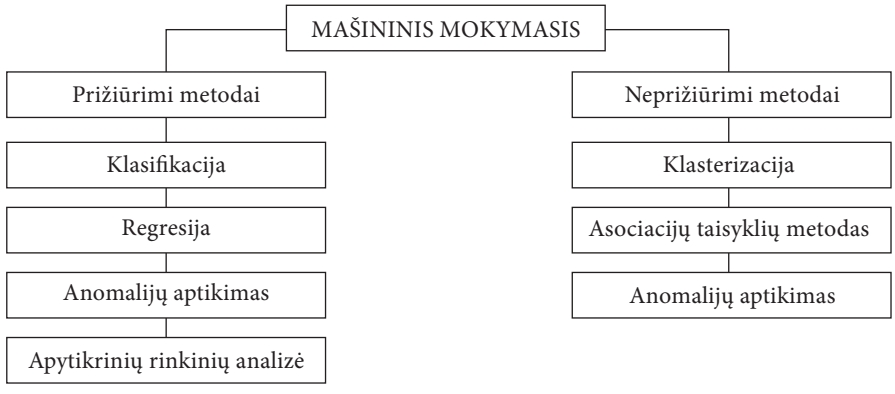
Organizaciniai barjerai – tai organizacinė kultūra, pasireiškianti per įmonės vertybes, normas ir simbolius. Įmonės organizacinę kultūrą iliustruoja vadovybės sudaryta strategija, kuria vadovaujasi visi darbuotojai. Egzistuoja nemažai įmonių, kurios negeba arba nesupranta, kaip panaudoti didžiuosius duomenis ir kokią naudą tai gali suteikti. Atsiranda didelė rizika, kad organizacinė kultūra taps kliūtimi naudoti didžiuosius duomenis įmonėje. Sėkmingiems kultūriniams pokyčiams pasiekti reikalinga su didžiaisiais duomenimis susijusi aiški įmonės vizija ir strategija. Darbuotojai turi žinoti, kokią naudą suteikia didieji duomenys, kokie yra įmonės tikslai naudojant didžiuosius duomenis ir kokių rezultatų tikimasi.

Siekdama išvengti didžiųjų duomenų naudojimo rizikas sukeliančių barjerų, organizacija pirmiausia turi skatinti tokių duomenų analize grįstų sprendimų priėmimą. Taip pat turi diegti atitinkamas didžiųjų duomenų apdorojimo technologijas ir mokyti darbuotojus suprasti bei analizuoti didžiuosius duomenis. Pabrėžtina, kad laikas, skirtas didžiųjų duomenų gavybai ir analizei, gali būti neproporcingai didelis gautai jos naudai. Dėl to būtina pasirinkti organizacijos renkamiems duomenims apdoroti tinkamus didžiųjų duomenų analizės metodus ir modelius bei nuspręsti, ar įmonei iš viso reikia analizuoti didžiuosius duomenis.

3. Didžiųjų duomenų analizės metodai klientams pažinti

Didieji duomenys – tai nuolat augantys duomenų rinkiniai, kurie yra įvairios struktūros (Akoka, Comyn-Wattiau ir Laoufi 2017; Oussous, Benjelloun, Lahcen ir Belfkih, 2017). Dėl to, ankstesnės duomenų apdorojimo technologijos – paketinis duomenų apdorojimas – nėra tinkamos didžiųjų duomenų gavybai ir analizei. Didžiųjų duomenų gavybai apibūdinti vartojama sąvoka „duomenų tyryba“. Duomenų tyryba – tai procesas, kurio metu, naudojant įvairias duomenų analizės priemones, išgaunami duomenų modeliai (Gaber, 2010). Duomenų tyrybos tikslas – rasti naujus, dar nežinomus duomenų modelius, kurie būtų naudojami priimant plėtros sprendimus, besiremiančius klientų priimtais sprendimais praeityje. Žmogui tai padaryti nėra lengva, todėl turi būti taikomi tam tikri mašininio mokymo metodai, pagal kuriuos nestruktūrizuoti duomenys būtų organizuojami ir klasifikuojami. Yra dvi mašininio mokymo metodų grupės – prižiūrimi (angl. *Supervised*) ir neprižiūrimi (angl. *Unsupervised*) metodai (Pandey, Kumar ir Srivastava, 2016) (3 paveikslas).

Prižiūrimi mašininio mokymosi metodai susiję su klasifikacija ir reikalauja žmogaus įsikišimo. Pirmame prognozavimo etape, kompiuteriui (mašinai) pateikiamas klasifikuotų duomenų paketas, iš kurio sukuriamas algoritmas, naudojamas kitais prognozavimo etapais; neprižiūrimi metodai apima duomenų skirstymą į kategorijas, pagrįstas duomenų įvesties parametrų panašumu (Koturwar,



3 paveikslas. Mašininio mokymosi metodai (sudaryta autorių, remiantis Makhdoomi, 2017)
Figure 3. Machine learning methods (composed by the authors based on Makhdoomi, 2017)

Girase ir Mukhopadhyay, 2015). M. Makhdoomi (2017) nuodugniau apibūdina toliau pateikiamus mašininio mokymosi metodus.

Klasifikacija – tai viena populiariausių duomenų gavybos technikų, naudojanti iš anksto klasifikuotų pavyzdžių rinkinį modeliui, kuris vėliau gali būti naudojamas naujiems neklasifikuotų duomenų įrašams klasifikuoti, sukurti. Klasifikavimo procesas apima du etapus: mokymąsi ir klasifikaciją. Mokymosi pakopoje duomenys analizuojami pagal klasifikavimo algoritmą. Klasifikavimo pakopoje naudojami bandymų duomenys, skirti klasifikavimo taisyklių tikslumui įvertinti. Buvo sukurti keli klasifikavimo metodai, populiariausias jų, taikomas sprendžiant realias pasaulio problemas – klasifikavimas naudojant sprendimų medžius, Bajeso klasifikaciją, palaikymo vektoriaus mašinas, asociacijomis grįstą klasifikaciją. Dar viena labai populiari klasifikavimo technika – neuroniniai tinklai. Jie sėkmingai naudojami klasifikacijai, nes padeda efektyviai suklasifikuoti sudėtingus duomenis. Naive Bayes klasifikatorius yra tiesioginis tikimybinis klasifikatorius, kuris taiko Bajeso teoremą ir imasi stiprių nepriklausomų sąryšių tarp funkcijų. K artimiausių kaimynų metodas – kitas populiarus klasifikavimo metodas. K artimiausių kaimynų metodo idėja yra naujo objekto palyginimas su mokymo aibės objektais, kurie yra panašūs į jį (Fan, Han ir Liu, 2016). Norint naują objektą priskirti kuriai nors klasei, skaičiuojami atstumai nuo to objekto iki visų mokymo aibės objektų.

Regresija gali būti pritaikyta skaitmeniniams objektams prognozuojant. Regresija gali būti naudojama modeliuojant santykius tarp vieno ar kelių nepriklausomų bei priklausomų kintamųjų. Yra įvairių tipų regresijos metodai: tiesinė regresija, daugiamatė tiesinė regresija, netiesinė regresija ir daugiamatė netiesinė regresija.

Asociacijų taisyklių metodas. Asociacijos nustatymo arba suradimo problema ta, kad reikia rasti aibę požymių ar atributų, kurie atskirti dideliu objektų rinkiniu duotoje duomenų bazėje. Tokio tipo analizė padeda priimti sprendimus ir analizuoti klientų elgesį. Ji turi keletą naudojimo būdų, pavyzdžiui, gautais rezultatais galima vadovautis pateikiant skirtingus produktus parduotuvėje tam, kad padidėtų pardavimo apimtis, būtų gaunama informacija apie pomėgius žmonių, lankančių tinklalapius, ar atrandant naujus ryšius tarp biologinių duomenų. Kai kurie iš visuotinai žinomų asociacijos taisyklių sudarymo būdų yra šie: daugiapakopė asociacijos taisyklė, daugiamatė asociacijos taisyklė ir kiekybinės asociacijos taisyklė.

Klasterizavimo metodas taikomas duomenų objekto panašumui rasti ir tokiems objektams grupuoti pagal rastus panašumus. Grupuojama taip, kad duomenų taškai, priklausantys tam pačiam klasteriui, yra labiau tarpusavyje panašūs nei duomenys, priklausantys skirtingam klasteriui. Yra sukurtos įvairios grupavimo metodikos: apibrėžto klasterizavimo (angl. *Crisp clustering*) ir neapibrėžto klasterizavimo (angl. *Fuzzy clustering*). Paprastai klasterizavimo algoritmai klasifikuojami į paskirstymo klasterizavimo algoritmus ir hierarchinę klasterizaciją. „K-means“ – vienas iš populiariausių klasterių algoritmų, kuris taiko suskaidymo metodą klasifikuojant duomenis į iš anksto nustatytą klasterių skaičių.

Anomalijų aptikimo metodas taikomas norint nustatyti duomenų perdavimo taškus. Išėjimo taškai yra tie taškai, kurie gerokai skiriasi nuo kitų duomenų. Pavyzdžiui, kreditinių kortelių sukčiavimo atvejų aptikimas, tinklo įsi-

1 lentelė Mašininio mokymosi metodų palyginimas
Table 1. Comparison of machine learning methods

Eil. Nr.	1	2	3	4
Metodas	Klasifikacija	Regresija	Klasterizacija	Asociacijų taisyklių
Technika	Sprendimų medžio sudarymas	Daugiamatė tiesinė regresija	„K-means“ algoritmas	Daugiapakopė asociacijos taisyklė
Tikslas	Sukurti modelį, kuris prognozuoja priklausomo kintamojo vertę, remdamasis keliais nepriklausomais kintamaisiais	Sudaryti statistinį modelį, apibūdinantį įtaką dviejų ar daugiau kiekybinių veiksnių X priklausomam kintamajam	Sujungti objektus į grupes (klasterius), kur kiekvienas objektas priklauso tam klasteriui, kurio vidurkis jam arčiausias	Rasti priklausomybes tarp atributų duomenų bazėje
Privalumai	– Gauti modeliai lengvai suprantami – Lengva sukurti modelį – Nereikalauja pradinio duomenų apdorojimo – Tinka didelėms duomenų apimtims	– Nustatoma vieno ar daugiau prognozuojamų kintamųjų įtaka kriterijų vertei – Identifikuojamos neatitiktys ar anomalijos	– Paprasta taikyti – Greitai apdorojami duomenys	– Paprasta taikyti – Nustatomi ryšiai tarp duomenų didelėje duomenų bazėje
Trūkumai	– Ne toks tikslus metodas kaip kiti – Gali būti sukuriami pernelyg sudėtingi modeliai	– Nustatomi ryšiai, bet ne priežastys – Ne visada lengvai suprantami modeliai	– Iš karto turi būti žinomas klasterių skaičius – Netinka darbui su netiksliais ir iškraipytais duomenimis	– Reikalauja daugkartinio duomenų bazės nuskaitymo – Metodo taikymas kelia grėsmę asmens privatumui

laužimo aptikimas ir gedimų nustatymas – tai tik keletas išskirtinio aptikimo metodų. Šie metodai nustato įprastą bendrą duomenų taškų elgesį ir nustato skirtingo elgesio duomenų taškus. Yra keletas išskirtinių tipų aptikimo metodų, kurie gali būti suskirstyti į vienarūšius metodus, daugialypius metodus, parametrinius metodus, neparametrinius metodus, grafinius, statistinius, nuotolinius.

Apytikrių rinkinių formavimas – klasifikacijos duomenų gavimo metodas, daugiausia susijęs su neaiškios ir neišsamios informacijos analize. Dauguma realių duomenų rinkinių yra sudėtingi, paprastai neapibrėžti ir neišsamūs, todėl juos galima analizuoti pagal šią klasifikavimo schemą. Matematiniai skaičiavimai naudojami paslėptiems duomenų modeliams ištirti apytikrių rinkinių atveju. Apytikriai rinkiniai turi keletą naudojimo būdų, pavyzdžiui, duomenims mažinti, funkcijų atrankai ir išgavimui bei sprendimų taisyklėms kurti.

Apžvelgus mašininio mokymosi metodus, toliau pateikiami keli dažniausiai mokslinėje literatūroje nagrinėjami (Koturwar et al., 2015; Panigrahi, 2012; Sunita ir Lobo, 2012; Makhdoomi, 2017): klasifikacijos, regresijos, klasterizacijos ir asociacijų taisyklių metodų lyginamoji analizė. 1 lentelėje pateikiami keturių mašininio mokymosi technikų: sprendimų medžio sudarymo, daugiamatės tiesinės regresijos, „K-means“ algoritmo ir daugiapakopės asociacijos taisyklės – apibūdinimai. Pateikiami jų tikslai, pagrindiniai privalumai ir trūkumai (1 lentelė).

Atlikus metodų lyginamąją analizę, buvo identifikuoti prižiūrimų ir neprižiūrimų mašininio mokymosi technikų tikslai, privalumai ir trūkumai. Kiekviena jų turi savų privalumų ir trūkumų, todėl vienos, tinkamos visoms situacijoms technikos išskirti negalima. Organizacija, siekianti pažinti klientus, taikydama didžiųjų duomenų analizę, turi pasirinkti metodą arba jų derinį, kuris tiks jos renkamiems duomenims analizuoti.

Apibendrinus galima teigti, kad duomenų tyryba – tai naujų duomenų modelių paieška. Siekiant supaprastinti ir pagreitinti šią paiešką, taikomi mašininio mokymosi metodai, kurie skirstomi į dvi grupes: prižiūrimus ir neprižiūrimus metodus.

4. Didžiųjų duomenų taikymo sritys

Didžiųjų duomenų naudojimo būdai sparčiai tobulėja, įmonėms stengiantis išgauti informacijos iš kiekvieno surinktų duomenų masyvo. Didžiųjų duomenų naudojimo proveržis daro įtaką daugeliui sričių ir suteikia daug galimybių. Siekiant geriau suvokti didžiųjų duomenų taikymo principą, tikslinga atlikti šios priemonės taikymo įvairiose srityse analizę, todėl toliau išskiriamos pagrindinės taikymo sritys:

Internetinė reklama. Vartotojas, apsilankęs internetinėje svetainėje, gali matyti tam tikrus skelbimus. Paspaudus ant skelbimo, jo susidomėjimas tam tikra preke fiksuojamas įmonės duomenų bazėje. Remiantis surinktais duomenimis, vartotojui pateikiama daugiau reklamų apie jo susidomėjimą sukėlusią prekių kategoriją (Sharma, 2015). Pavyzdžiui, socialiniame tinkle „Facebook“ reklaminiai skelbimai pagrindiniame naudotojo puslapyje rodomi, remiantis duomenimis, surinktais iš jo veiklos: mygtukų „patinka“ paspaudimų, paieškos laukelyje vestų reikšminių žodžių. Tokių duomenų analizei gali būti taikomas „K-means“ metodas.

Rekomendacijos internetinėje mažmeninėje prekyboje ir socialiniuose tinkluose. Rekomendacijos yra svarbus internetinių parduotuvių, muzikos ir filmų parduotuvių, tokių kaip „Amazon“, „Netflix“, „Ebay“ ir kt., bei socialinių tinklų: „Facebook“, „Youtube“ – bruožas (Sharma, 2015). Čia analizuojamos vartotojų charakteristikos, ieškoma panašumų ir skirtumų tarp įvairių vartotojų. Duomenys, surinkti iš ankstesnės vartotojo patirties, ankstesnių pasirinkimų ar pasirinkimų, kuriuos padarė kiti panašaus profilio ir interesų vartotojai, naudojami jo naujiems pasirinkimams prognozuoti ir pateikti kaip rekomendacijas. Tam fiksuojamos vartotojo filmų anonsų ar prekių peržiūros, kurių analizei gali būti taikomi klasterizacijos metodai.

Telekomunikacijų sistemos. Telekomunikacijų operatoriai gali naudoti didžiųjų duomenų analizę, kad sugeneruotų klientų pageidavimus ir įvertintų tinklo efektyvumą realiuoju laiku. Tai suteikia galimybę priimti faktais pagrįstus sprendimus realiuoju laiku. Yra išskiriami keli aspektai (Deloitte, 2015):

- pagerinta klientų patirtis. Įmonė įgauna geresnį supratimą apie klientus ir gali patobulinti jų patirtį teikdama aukštos kokybės paslaugas, palaikydama greitą grįžtamąjį ryšį ir teikdama individualius pasiūlymus;
- duomenų valdymas. Naudojant organizacijoje turimą informaciją ir kombinuojant ją su įžvalgomis iš rinkos, galima pagerinti priimamus sprendimus ir sumažinti išlaidas;
- duomenų skatinamas augimas. Didieji duomenys suteikia galimybę atsirasti naujiems pajamų srautams, kuriuos sukuria teikiami nauji pasiūlymai klientams.

Čia fiksuojami vartotojo atlikti mokėjimai, paieškose vartoti reikšminiai žodžiai, skambučių išklotinės. Tokio tipo duomenys analizuojami taikant klasifikacijos, klasterizacijos metodus.

Vieta pagrįstos paslaugos. Galimybė naudotis didelės spartos internetu ir GPS sistemomis, esančiomis išmaniuosiuose įrenginiuose, leidžia vartotojui pasiekti daugybę programų rinkti ir naudoti vietovės duomenis bei juos apdoroti, pavyzdžiui, eismo įvertinimo sistemas, kuriose vertinamas eismo scenarijus. Tokios sistemos miestuose yra plačiai naudojamos. „Google“ sukūrė „Google “ žemėlapius, kurie ne tik rodo vietą, bet ir įvertina atstumą ir laiką, reikalingą keliauti iš vienos vietos į kitą. Šios paslaugos apima realiojo laiko kelių ir oro eismo įvertinimo sistemas (Sharma, 2015).

Sveikatos priežiūros paslaugos. Remiantis ankstesniais ligų ir infekcijų duomenimis, sveikatos priežiūros specialistai gauna galimybę optimizuoti pacientų gydymą, nuspėti tam tikros ligos proveržį ar tam tikrų vaistų efektyvumą (Langkafel, 2016).

Žemės ir atmosferos jutiklių sistemos. Tokios sistemos saugo petabaitus duomenų, gautus iš jutiklių. Šios sistemos apskaičiuoja žemės drebėjimo epicentrus ir įvertina atmosferos pokyčius (Sharma, 2015). Didieji duomenys sukūrė naujus prognozavimo būdus. Yra keletas veiksnių, kurie lėmė išaugusį mokslininkų susidomėjimą didžiaisiais duomenimis (Sharma, 2015):

- galimybė nustatyti santykius tarp skirtingų duomenų šaltinių;
- galimybė prognozuoti vartotojo elgseną;
- produktų ar paslaugų pardavimo prognozavimas;
- sukčiavimo ar finansinės rizikos prognozavimas;
- socialinio tinklo duomenų analizė vartotojų jausmui ir pasitenkinimui nustatyti;
- galimybė analizuoti didelės apimties kompiuterinius duomenis iš jutiklių tinklaraščių, paspaudimų skaičiaus.

Taigi apibendrinus galima teigti, kad didžiųjų duomenų naudojimo galimybės yra labai plačios. Nors didžiųjų duomenų valdymo strategijos ir praktinio pritaikymo būdai vis dar vystosi, didžiųjų duomenų analizė tapo būtinybe daugeliui įmonių, įvairiose pramonės šakose. Didieji duomenys plačiausiai naudojami internetinėje reklamoje, prekyboje, telekomunikacijose, sveikatos priežiūroje ir žemės bei atmosferos jutiklių sistemose. Efektyviausiai pritaikomi klasifikacijos ir klasterizacijos duomenų analizės metodai, kurie tinka nestruktūrizuotiems duomenims analizuoti.

Išvados

- Atlikta mokslinės ir kitos literatūros analizė leido identifikuoti didžiųjų duomenų sampratą. Nustatyta, kad literatūroje nėra vieno didžiųjų duomenų apibrėžimo. Apibendrintai galima teigti, kad didieji duomenys – tai dideli duomenų masyvai, kuriems apdoroti netinka įprastinės duomenų apdorojimo priemonės. Nustatyta, kad tai kompleksinė sąvoka, kuri vis dar vystosi. Pabrėžtina, kad autoriai, apibūdindami didžiųjų duomenų sąvoką, ją apibrėžia siaurąja ir plačiąja prasme ir akcentuoja skirtingas tokių duomenų charakteristikas. Tai išplečia didžiųjų duomenų sąvokos koncepciją ir sukelia daug painiavos kognityvine prasme.
- Organizacijos, analizuojančios didžiuosius duomenis, susiduria su įvairiomis rizikomis. Esminė rizika – gauti rezultatai nėra tikslūs. Siekiant identifikuoti kylančias rizikas, išskiriami didžiųjų duomenų naudojimo barjerai: technologiniai, žmogiškieji ir organizaciniai. Atsižvelgusios į šiuos barjerus, organizacijos gebės užtikrinti tinkamų technologijų naudojimą, suteikti reikiamus įgūdžius darbuotojams ir garantuoti didžiaisiais duomenimis grindžiamos organizacinės kultūros vystymą. Organizacijos, kurios sugebės eliminuoti didžiųjų duomenų naudojimo rizikas, išliks konkurencingos augančioje, duomenimis grindžiamoje ekonomikoje.
- Duomenų gavyba ir analizė – sudėtingas procesas. Duomenų analizės metodai turi būti parenkami atsižvelgiant į tai, kaip ir kokioje srityje organizacija renka didžiuosius duomenis. Jie analizuojami ne tik norint gauti naujų įžvalgų apie klientus, bet ir tobulinant verslo modelį. Didžiųjų duomenų analizei taikomi prižiūrimi ir neprižiūrimi mašininio mokymosi metodai. Dažniausiai taikomas didžiųjų duomenų analizės metodas – klasifikacija. Efektyviai pritaikius analizės metodus, kurie tinkami organizacijos renkamiems duomenims, atsiranda galimybės tobulinti sprendimų priėmimo procesą, dėl to didėja konkurencinis organizacijos pranašumas.
- Didieji duomenys ir jų analizė – nauja sparčiai besivystanti sritis, kuri kasmet sukuria vis didesnę vertę įmonėms, gebančioms veiksmingai naudoti tokius duomenis. Jų analizė taikoma įvairiose srityse: internetinėje reklamoje, medicinoje, telekomunikacijose. Pabrėžtina, kad patys duomenys nekuria vertės organizacijai. Vertė gaunama duomenų analizės proceso metu, gautas įžvalgas naudojant sprendimams priimti.

Literatūra

Abawajy, J. H., Kelarev, A., & Chowdhury, M. (2014). Large iterative multitier ensemble classifiers for security of big data. *IEEE Transactions on Emerging Topics in Computing, 2*(3), 352-363. https://doi.org/10.1109/TETC.2014.2316510

Akoka, J., Comyn-Wattiau, I., & Laoufi, N. (2017). Research on Big Data – a systematic mapping study. *Computer Standards & Interfaces, 54*(2), 105-115. https://doi.org/10.1016/j.csi.2017.01.004

Alharthi, A., Krotov, V., & Bowman, M. (2017). Addressing barriers to big data. *Business Horizons, 60*(5), 285-292. https://doi.org/10.1016/j.bushor.2017.01.002

Beyer, A. M., & Laney, D. (2012). The Importance of Big Data: a definition. Gartner, Stamford, CT. Retrieved from https://www.gartner.com/doc/2057415/importance-big-data-definition

Boyd, D., & Crawford, K. (2012). Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon, information. *Communication & Society, 15*(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878

Boksem, M. A. S., & Smidts, A. (2015). Brain responses to movie trailers predict individual preferences for movies and their population-wide commercial success. *Journal of Marketing Research, 52*(4), 482-492. https://doi.org/10.1509/jmr.13.0572

Chen, M., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From Big Data to Big Imptact. *MIS Quarterly: Management Information Systems, 36*(4), 1165-1188.

De Mauro, A., Greco, M., & Grimaldi, M. (2014). What is big data? A consensual definition and a review of key research topics. *4th International Conference on Integrated Information AIP Proceedings.* Madrid.

Deloitte. (2015). *Opportunities in Telecom Sector: arising from Big Data*. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/in/Documents/technology-media-telecommunications/in-tmt-opportunities-in-telecom-sector-noexp.pdf

Diebold, F. X. (2012). *A personal perspective on the origin(s) and develop-ment of "big data": The phenomenon, the term, and the discipline, second version*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2202843

Dippel, A. (2017). Das Big Data Game, NTM Zeitschrift für Geschichte der Wissenschaften. *Technik und Medizin, 25*(4), 485-517.

Fan, J., Han, F., & Liu, H. (2016). Challenges of Big Data analysis. *National ScienceReview, 1*(2), 293-314. https://doi.org/10.1093/nsr/nwt032

Gaber, M. M. (2010). *Scientific data mining and knowledge discovery – principles and foundations* (397 p.). New York: Springer. ISBN: 97 836 42027871.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gartner. (2015). *Gartner says business intelligence and analytics leaders must focus on mindsets and culture to kick start advanced analytics.* Retrieved from https://www.gartner.com/newsroom/id/3130017

Gartner. (2017). *Big Data*. Retrieved from http://www.gartner.com/it-glossary/big-data/

Gupta, U. G., & Gupta, A. (2016). Vision: a missing key dimension in the 5V Big Data framework. *Journal of International Business Research and Marketing, 1*(3), 46-52.

Hua Tan, K., & Zhan, Y. (2016). Improving new product development using big data: a case study of an electronics company. *R&D Management, 47*(4), 570-582. https://doi.org/10.1111/radm.12242

Jinjiang, Y., & Xueling, Z. (2016). The research on China's current online retail statistics based on Big Data's Perspective. *Innovation, entrepreneurship and strategy in the era of internet* (pp. 388-393).

Kaur, K., Kaur, I., Kaur, N., Tanisha, Gurmeen, & Deepi. (2016). Big data management: characteristics, challenges and solutions. *International Journal of Computer Science and Technology, 7*(4), 54-57.

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, W. K. M., Alam, M., Shiraz, M., & Gani, A. (2014). Big data: survey, technologies, opportunities, and challenges. *The Scientific World Journal*. Retrieved from https://www.hindawi.com/journals/tswj/2014/712826/

Koturwar, P., Girase, S., & Mukhopadhyay, D. (2015). A survey of classification techniques in the area of big data. *International Journal of Advance Foundation and Research in Computer, 1*(11), 1-7.

Krasnow Waterman, K., & Bruening P. J. (2014). Big Data analytics: risks and responsibilities. *International Data Privacy Law, 4*(2), 89-95. https://doi.org/10.1093/idpl/ipu002

Laney, D. (2001). *3-D data management: controlling data volume, velocityand variety*. Application delivery strategies by META Group Inc. Retrieved from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Langkafel, P. (2016). *Big Data in medical science and healthcare management: diagnosis, therapy, side effects* (248 p.). Germany: Berlin. ISBN: 978-3-11-044528-2.

Lee, I. (2017). Big Data: dimensions, evolution, impacts, and challenges. *Business Horizons, 60*(3), 293-303. https://doi.org/10.1016/j.bushor.2017.01.004

Lee, J., Kao, H. A., & Yang, S. (2014). Service innovation and smart analytics for Industry 4.0 and big data environment. *Procedia CIRP, 16*, 3-8. https://doi.org/10.1016/j.procir.2014.02.001

Makhdoomi, M. (2017). Data mining approach for Big Data Analysis: a theoretical discourse. *International Journal of Advanced Research in Computer Science, 8*(7), 104-109. https://doi.org/10.26483/ijarcs.v8i7.4032

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: the next frontier for innovation, competition, and productivity*. Retrieved from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

Marr, B. (2015). *Big Data: 20 mind-boggling facts everyone must read*. Retrieved from https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#5947d77c17b1

McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review, 90*(10), 60-68.

Miller, S. (2014). Collaborative approaches needed to close the big data skills gap. *Journal of Organization Design, 3*(1), 26-30. https://doi.org/10.7146/jod.9823

Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2017). Big Data technologies: a survey. *Journal of King Saud University – Computer and Information Sciences.* Retrieved from http://www.sciencedirect.com/science/article/pii/S1319157817300034

Pandey, P., Kumar, M., & Srivastava, P. (2016). *Classification techniques for Big Data: a survey*. Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7724938

Panigrahi, P. K. (2012). A comparative study of supervised machine learning techniques for spam e-mail filtering. *Fourth International Conference on Computational Intelligence and Communication Networks* (pp. 506-512). https://doi.org/10.1109/CICN.2012.14

SAP. (2012). *Small and midsize companies look to make big gains with "big data"*. Retrieved from http://global.sap.com/corporate-en/news.epx?PressID=19188

Seddon, J. J. J. M., & Currie, W. L. (2017). A model for unpacking big data analytics in high-frequency trading. *Journal of Business Research, 70*, 300-307. https://doi.org/10.1016/j.jbusres.2016.08.003

Sharma, S. (2015). *Rise of Big Data and related issues*. IEEE. Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7443346

Simon, P. (2014). *The visual organization: data visualization, big data, and the quest for better decisions* (28 p.). USA: SAS Institute.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research, 70*, 263-286. https://doi.org/10.1016/j.jbusres.2016.08.001

Stanton, S. J., Sinnott-Armstrong, W., & Huettel, S. A. (2017). Neuromarketing: ethical implications of its use and potential misuse. *Journal of Business Ethics, 144*(4), 799-811. https://doi.org/10.1007/s10551-016-3059-0

Sunita, B. A., & Lobo, L. M. R. J. (2012). A comparative study for selecting the best unsupervised learning algorithm in e-learning system. *International Journal of Computer Applications, 41*(3), 27-34. https://doi.org/10.5120/5523-7562

Tole, A. A. (2013). Big Data challenge. *Database Systems Journal, 4*(3), 31-40.

TRUSTe. (2015). *2015 US IoT Privacy Index*. Retrieved from https://www.truste.com/resources/privacy-research/us-internet-of-things-index-2015/

ur Rehman, M. H., Liew, C. S., Abbas, A., Jayaraman, P. P., Wah, T. Y., & Khan, S. U. (2016). Big Data reduction methods: a survey. *Data Science and Engineering, 1*(4), 265-284. https://doi.org/10.1007/s41019-016-0022-0

Walker, R., ap Cenydd, L., Pop, S., Miles, H. C., Hughes, C., Teahan, W. J., & Roberts, J. C. (2013). *Storyboarding for visual analytics, Information Visualization, 14*(1), 27-50. https://doi.org/10.1177/1473871613487089

Wang, S., Wan, J., Zhang, D., Li, D., & Zhang, C. (2016). Towards smart factory for industry 4.0: a self-organized multi-agent system with big data based feedback and coordination. *Computer Networks, 101*(4), 158-168. https://doi.org/10.1016/j.comnet.2015.12.017

Ward, J. S., & Barker, A. (2013). *Undefined by data: a survey of big data definitions*. Retrieved from http://arxiv.org/abs/1309.5821

Zakir, J. (2015). Big Data analytics. *International Association for Computer Information Systems, 16*(2), 81-90.

Zhan, Y., Tan, K. H., Li, Y., & Tse, Y. K. (2016). Unlocking the power of big data in new product development. *Annals of Operations Research* (pp. 1-19). https://doi.org/10.1007/s10479-016-2379-x

MODEL OF THE BIG DATA USE FOR CUSTOMER COGNITION

S. Politaitė, J. Sabaitytė

Abstract

In a customer-oriented market, understanding customer behavior is an important determinant of the success of an organization. An organization that strives to survive and succeed can not ignore increasing amounts of data – big data. Big data is complex data arrays that are difficult to process using traditional data processing applications. Optimal analysis of such data enables organizations for better understanding of its customers, improve the decision-making process and increase its competitive advantage. It is important for the organization to understand how to use big data, which processing tools and models to apply. This article analyzes the concepts and evolution of big data, the risks of exploitation, mining methods and applied models. Applied methods: systematic, logical analysis of information sources, comparison of information, systemization.

Keywords: Big data, customer cognition, big data analytics, risks of exploitation, big data mining, big data management.