19. Smirnov S.A. Method of controlling access to intellectual switching nodes of telecommunication networks and systems / A.A. Smirnov, Mohamad Abou Taam, S.A. Smirnov // International Journal of Computational Engineering Research (IJCER). – Volume 5, Issue 5. – India. Delhi. – 2015. – P. 1-7.
20. Смирнов С. А. Анализ и исследование методов управления сетевыми ресурсами для обеспечения антивирусной защиты даннях / Мохамад Абу Таам Гани, А. А. Смирнов, С. А. Смирнов // Системи озброєння і військова техніка: наук. журн. – Х.: ХУПС, 2015. – № 3(43). – С. 100-107.
21. Смирнов С. А. Исследование эффективности метода управления доступом к облачным антивирусным телекоммуникационным ресурсам / Мохамад Абу Таам Гани, А. А. Смирнов, С. А. Смирнов // Наука і техніка Повітряних Сил Збройних Сил України: наук. журн. –Х.: ХУПС, 2015. –№ 3(20). – С. 134-141.
22. Смирнов С. А. Комплекс gert-моделей технологии облачной антивирусной защиты телекоммуникационной системы / А. А. Смирнов, А. К. Дидык, А. Н. Дреев, С. А. Смирнов // Безпека інформації: наук. - практ. журн. – К.: НАУ, 2015. – Т. 21, № 3. – С. 251-262.

УДК 004

А.Пилипенко, магістр гр. КН-21М-1,4,
Центральноукраїнський національний технічний університет

ДОСЛІДЖЕННЯ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ СИСТЕМИ BIG DATA НАУКОВИХ ДОСЛІДЖЕНЬ

У роботі розроблено програмне забезпечення, яке призначено для системи big data наукових досліджень. Метою розробки є дослідження та програмна реалізація системи big data наукових досліджень. Об’єктом дослідження є процес big data наукових досліджень. Предметом дослідження є методи big data наукових досліджень. Методи дослідження базуються на методах big data, методах математичної статистики, методах розробки програмного забезпечення. Результат роботи – програмна реалізація системи big data наукових досліджень. В процесі роботи над програмною моделлю виконано аналіз існуючих апаратних та програмних засобів. В повній мірі описані всі компоненти розробленого програмного забезпечення.

комп’ютерні науки, big data, наукові дослідження

Постановка проблеми.

Великі дані (big data) обіцяють революціонізувати виробництво знань у науці та за її межами, забезпечивши нові, високоефективні способи планування, проведення, поширення та оцінки досліджень. Останні кілька десятиліть стали свідками створення нових способів виробництва, зберігання та аналізу даних, кульмінацією яких стала поява галузі даних, яка об’єднує обчислювальні, алгоритмічні, статистичні та математичні методи для екстраполяції знань із великих даних. У той же час рух *відкритих даних,* що виник на основі таких політичних тенденцій, як поштовх до відкритого уряду та відкритої науки, заохочував обмін і взаємозв’язок різнорідних дослідницьких даних через великі цифрові інфраструктури. Наявність величезних обсягів даних у машиночитаних форматах створює стимул для створення ефективних процедур збору, організації, візуалізації та моделювання цих даних. Ці інфраструктури, у свою чергу, служать платформами для розвитку штучного інтелекту з метою підвищення надійності, швидкості та прозорості процесів створення знань. Дослідники з усіх дисциплін бачать, що нова здатність зв’язувати та перехресно посилатися на дані з різних джерел покращує точність і прогностичну силу наукових висновків і допомагає визначити майбутні напрямки дослідження, таким чином, зрештою, забезпечуючи нову відправну точку для емпіричного дослідження. Як свідчить зростання цільового фінансування, навчальних програм і місць публікацій, великі дані широко розглядаються як започаткування нового способу проведення досліджень і кидання виклику існуючому розумінню того, що вважається науковим знанням.

Аналіз останніх досліджень і публікацій. При аналізі останніх досліджень і публікацій [1-10] було виявлено певні прогалини у забезпеченні системи big data наукових досліджень.

Мета й завдання дослідження. Метою роботи є дослідження та програмна реалізація системи big data наукових досліджень.

Для досягнення поставленої мети визначена програма дослідження, що складається з наступних завдань:

– Огляд існуючих систем big data наукових досліджень.
– Дослідження системи big data наукових досліджень.
– Програмна реалізація системи big data наукових досліджень.

Об'єктом дослідження є процес big data наукових досліджень.

Предметом дослідження є методи big data наукових досліджень.

Методи дослідження базуються на методах big data, методах математичної статистики, методах розробки програмного забезпечення.

Виклад основного матеріалу.

У цій роботі досліджуються ці твердження щодо використання великих даних у наукових дослідженнях і з наголосом на філософських питаннях, які виникають у результаті такого використання. З цією метою у роботі обговорюється, як поява великих даних – і пов'язаних з ними технологій, інститутів і норм – інформує про аналіз наступних тем:

– як статистика, формальні та обчислювальні моделі допомагають екстраполювати закономірності з даних і з якими наслідками;
– роль критичного аналізу (людського інтелекту) у машинному навчанні та його зв'язок із зрозумілістю дослідницьких процесів;
– характер даних як компонентів дослідження;
– зв'язок між даними та доказами, а також роль даних як джерела емпіричного розуміння;
– погляд на знання як на теорії;
– розуміння зв'язку між прогнозом і причинністю;
– поділ факту і цінності; і
– ризики та етика науки про дані.

Це сфери, де увага до дослідницьких практик, що обертаються навколо великих даних, може принести користь філософії, і особливо роботі в епістемології та методології науки. Цей запис не охоплює величезну науку в історії та соціальних дослідженнях науки, яка виникла в останні роки на цю тему, хоча посилання на деякі з цієї літератури можна знайти, якщо це концептуально доречно. Доповнюючи історичну та соціальну наукову роботу в дослідженнях даних, філософський аналіз практик обробки даних також може викликати значні проблеми для ажіотажу навколо науки про дані та сприяти критичному розумінню ролі штучного інтелекту, що базується на даних, у дослідженнях.

Великі дані (big data)

Діяльність людини та взаємодія з навколишнім середовищем відстежуються та реєструються все ефективніше, де розробляються дедалі складніші обчислювальні інструменти для отримання знань із таких даних. Одним із прикладів є використання різних даних, отриманих від хворих на рак, включаючи геномні послідовності, фізіологічні вимірювання та індивідуальні реакції на лікування, для покращення діагностики та лікування. Іншим прикладом є інтеграція даних про транспортні потоки, навколишні та географічні умови, а також поведінку людини для розробки заходів безпеки для безпілотних транспортних засобів, щоб у разі непередбачуваних подій (наприклад, дитина раптово вибігла на вулицю в дуже холодний день), дані можна швидко проаналізувати, щоб ідентифікувати та сформувати відповідну реакцію (автомобіль повертає достатньо, щоб уникнути дитини, а також мінімізує ризик заносу на льоду та пошкодження інших транспортних засобів). Ще одним прикладом є розуміння харчового статусу та потреб конкретної групи населення, яке можна витягти з об'єднання даних про споживання їжі, отриманих комерційними службами (наприклад, супермаркетами, соціальними мережами та ресторанами), з даними, що надходять від громадської охорони здоров'я та соціальних служб, наприклад результати аналізів крові та госпіталізації, пов'язані з недоїданням. У

кожному з цих випадків доступність даних і відповідних аналітичних інструментів створює нові можливості для дослідження та розробки нових форм дослідження, які широко сприймаються як такі, що мають трансформаційний вплив на науку в цілому.

Корисною відправною точкою для роздумів про значення таких випадків для філософського розуміння дослідження є розгляд того, що насправді означає термін «великі дані» в сучасному науковому дискурсі. Існує кілька способів визначення великих даних (Kitchin 2014, Kitchin & McArdle 2016). Мабуть, найпростішою характеристикою є *великі* набори даних, які створюються в *цифровій* формі та можуть аналізуватися за допомогою *обчислювальних* інструментів. Отже, дві характеристики, які найчастіше асоціюються з великими даними, – це обсяг і швидкість. *Обсяг* означає розмір файлів, які використовуються для архівування та поширення даних. *Швидкість* означає швидкість натискання, з якою дані генеруються та обробляються. Обсяг цифрових даних, створених дослідженнями, зростає шаленою швидкістю та такими способами, які, мабуть, неможливо осягнути людською когнітивною системою, і тому вимагають певної форми автоматизованого аналізу.

Альтернативою є визначення великих даних не через посилання на їхні фізичні атрибути, а скоріше через те, що можна і що не можна з ними робити. З цієї точки зору, великі дані – це різнорідний ансамбль даних, зібраних із різних джерел, як правило (але не завжди) у цифрових форматах, придатних для алгоритмічної обробки, з метою створення нових знань. Наприклад, Бойд і Кроуфорд (2012: 663) ототожнюють великі дані зі «здатністю шукати, агрегувати та перехресно посилатися на великі набори даних», тоді як О'Меллі та Соєр (2012) зосереджуються на здатності опитувати та взаємозв’язувати різні типи даних., щоб мати можливість ознайомитися з ними як з єдиною сукупністю доказів. Приклади трансформаційних «досліджень великих даних», наведені вище, легко вписуються в цю точку зору: це не просто факт, що доступно багато даних, що робить відмінність у цих випадках, а скоріше той факт, що багато даних можна мобілізувати з різноманітних джерел (медичні записи, дослідження навколишнього середовища, вимірювання погоди, поведінка споживачів).

Таке розуміння великих даних сягає корінням у довгу історію дослідників, які стикаються з великими та складними наборами даних, прикладами яких є астрономія, метеорологія, таксономія та демографія (див. колекції, зібрані Daston 2017; Anorova та ін. 2017; Porter & Chaderavian). 2018; також Анорова та ін., Сепкоскі 2013, Стівенс 2019). Подібним чином біомедичні дослідження – і особливо такі підгалузі, як епідеміологія, фармакологія та громадське здоров’я – мають широку традицію роботи з даними великого обсягу, швидкості, різноманітності та мінливості, достовірність, достовірність і цінність яких регулярно обговорюється та оскаржується пацієнтами, урядами., спонсори, фармацевтичні компанії, страхові компанії та державні установи (Bauer 2008). Протягом двадцятого століття ці зусилля стимулювали розвиток методів, установ та інструментів для збору, упорядкування, візуалізації та аналізу даних, таких як: стандартні системи та формати класифікації; вказівки, інструменти та законодавство щодо управління та безпеки конфіденційних даних; та інфраструктури для інтеграції та підтримки зборів даних протягом тривалих періодів часу (Daston 2017).

Кульмінацією цієї роботи стало застосування обчислювальних технологій, інструментів моделювання та статистичних методів до великих даних (Porter 1995; Humphreys 2004; Edwards 2010), що все більше розширює межі аналітики даних завдяки керованому навчанню, підгонці моделі, глибоким нейронним мережам, пошуку та методи оптимізації, складні візуалізації даних та різноманітні інші інструменти, пов’язані зі штучним інтелектом. Багато з цих інструментів базуються на алгоритмах, функціонування та результати яких перевіряються на конкретних зразках даних (цей процес називається «навчанням»). Ці алгоритми запрограмовані на «навчання» з кожної взаємодії з новими даними: іншими словами, вони мають здатність змінювати себе у відповідь на нову інформацію, що вводиться в систему, таким чином стаючи більш налаштованими на явища,

які вони аналізують, і покращуючи свої здатність передбачати майбутню поведінку. Обсяг і ступінь таких змін визначаються припущеннями, які використовуються для побудови алгоритмів, а також здатністю відповідного програмного та апаратного забезпечення ідентифікувати, отримувати доступ і обробляти інформацію, яка має відношення до відповідного навчання. Однак існує певний ступінь непередбачуваності та непрозорості цих систем, які можуть розвиватися до такого рівня, що кидає виклик людському розумінню (докладніше про це нижче).

Також з'явилися нові інституції, комунікаційні платформи та нормативні рамки для збирання, підготовки та підтримки даних для такого використання (Kitchin 2014), наприклад, різні форми інфраструктури цифрових даних, організації, які прагнуть координувати та вдосконалювати глобальний ландшафт даних (наприклад, Research Data Alliance), а також нові заходи щодо захисту даних, як-от Загальний регламент захисту даних, прийнятий у 2017 році Європейським Союзом. Разом ці методи та інституції дають можливість збирати та інтерпретувати дані в набагато ширшому масштабі, а також обіцяють забезпечити вищий рівень деталізації в аналізі даних. [1] Вони розширюють масштаби будь-якого дослідження, даючи дослідникам можливість пов'язувати власні висновки з висновками незліченної кількості інших у всьому світі, як у академічній сфері, так і за її межами. Підвищуючи мобільність даних, вони полегшують їх перепрофілювання для різноманітних цілей, які могли бути непередбачуваними під час початкового створення даних. Трансформуючи роль даних у дослідженнях, вони самі по собі підвищують свій статус цінних результатів дослідження. Ці технологічні та методологічні розробки мають значні наслідки для філософської концептуалізації даних, процесів висновків і наукових знань, а також для того, як дослідження проводяться, організовуються, керуються та оцінюються. Саме до цих філософських проблем я зараз звертаюся.

Екстраполяція шаблонів даних: роль статистики та програмного забезпечення

Великі дані часто асоціюють з ідеєю дослідження, *керованого даними*, де навчання відбувається через накопичення даних і застосування методів для вилучення значущих моделей із цих даних. Очікується, що в рамках дослідження, керованого даними, дослідники використовуватимуть дані як відправну точку для індуктивного висновку, не покладаючись на теоретичні упередження – ситуацію, яку прихильники описують як «кінець теорії», на відміну від підходів, керованих теорією, де дослідження складається з перевірка гіпотези (Anderson 2008, Hey et al. 2009). Принаймні в принципі, великі дані становлять найбільший пул даних, який коли-небудь збирався, і, отже, сильну відправну точку для пошуку кореляцій (Mayer-Schönberger & Cukier 2013). Вирішальним для достовірності підходу, керованого даними, є ефективність методів, що використовуються для екстраполяції шаблонів із даних і оцінки того, чи є такі шаблони значущими чи ні, і яке «значення» може включати в себе в першу чергу. Тому деякі філософи та дослідники даних стверджують це

– найважливішою та відмінною характеристикою великих даних є використання статистичних методів і обчислювальних засобів аналізу (Symons & Alvarado 2016: 4).

– наприклад, інструменти машинного навчання, глибокі нейронні мережі та інші «інтелектуальні» практики обробки даних.

Акцент на статистиці як ключовому критерії достовірності та надійності моделей, отриманих із даних, не є новим. Прихильники логічного емпіризму шукали логічно надійні методи для забезпечення та виправдання висновків на основі даних, і їхні зусилля з розробки теорії ймовірності йшли паралельно з укоріненням статистичних міркувань у науках у першій половині двадцятого століття (Romeijn 2017). . На початку 1960-х років Патрік Суппес запропонував фундаментальний зв'язок між статистичними методами та філософією науки завдяки своїй роботі над створенням та інтерпретацією моделей даних. Як філософ, глибоко вкорінений в експериментальній практиці, Суппес цікавився засобами та мотивацією ключових статистичних процедур для аналізу даних, таких як редукція даних і підгонка кривої. Він стверджував, що як тільки дані належним чином *підготовлені* для статистичного моделювання, усі проблеми та вибір, які мотивували обробку даних, стають

неактуальними для їх аналізу та інтерпретації. Це надихнуло його розрізнити моделі теорії, моделі експерименту та моделі даних, зазначивши, що такі різні компоненти дослідження керуються різною логікою і не можуть порівнюватися прямолінійним способом. Наприклад, точне визначення моделей даних для будь-якого даного експерименту вимагає наявності теорії даних у сенсі експериментальної процедури, а також у звичайному розумінні емпіричної теорії явищ, що вивчаються. (Suppes 1962: 253)

Суппес розглядав моделі даних як обов'язково статистичні, тобто як об'єкти призначений для включення всієї інформації про експеримент, яка може бути використана в статистичних перевірках адекватності теорії. (Suppes 1962: 258)

Що повинно входити в моделі даних? Основним обмеженням є потреба в моделях даних, які дозволяють статистичну оцінку відповідності (між прогнозом і фактичними даними); (Mayo 1996: 136) і Бас ван Фраассен, який також прийняв ідею моделей даних як «узагальнення відносних частот, знайдених у даних» (Van Fraassen 2008: 167). Тісно пов'язаний наголос на статистиці як на засобі виявлення помилок у наборах даних щодо конкретних гіпотез, найбільш помітно схвалених статистичним підходом до висновку про помилки, який відстоюють Мейо та Аріс Спанос (Mayo & Spanos 2009a). Цей підхід узгоджується з наголосом на обчислювальних методах для аналізу даних у рамках дослідження великих даних і підтримує ідею про те, що чим кращі інструменти та методи висновків, тим більше шансів витягти надійні знання з даних.

Проте, коли справа доходить до вирішення методологічних проблем, пов'язаних з обчислювальним аналізом великих даних, статистичний досвід має бути доповнений обчислювальною кмітливістю в навчанні та застосуванні алгоритмів, пов'язаних зі штучним інтелектом, включаючи машинне навчання, а також інші математичні процедури для роботи з даних (Bringsjord & Govindarajulu 2018). Розглянемо, наприклад, проблему переобладнання, тобто помилкову ідентифікацію шаблонів у наборі даних, яка може бути значно посилена методами навчання, що використовуються алгоритмами машинного навчання. Немає жодної гарантії, що алгоритм, навчений для успішної екстраполяції шаблонів із даного набору даних, буде таким же успішним, коли його застосувати до інших даних. Загальні підходи до цієї проблеми передбачають перевпорядкування та розділення як даних, так і методів навчання, щоб можна було порівняти застосування одних і тих самих алгоритмів до різних підмножин даних («перехресна перевірка»), об'єднати передбачення, що виникають із по-іншому навчених алгоритмів («ансамблювання») або використовувати гіперпараметри (параметри, значення яких встановлюються до навчання даних), щоб підготувати дані для аналізу.

Вирішення цих проблем, у свою чергу, вимагає знайомство з математичними операціями, про які йде мова, їх реалізацією в коді та апаратною архітектурою, що лежить в основі таких реалізацій. (Лоурі 2017: 3)

Наприклад, машинне навчання націлене на створення програм, які розробляють власні аналітичні або описові підходи до сукупності даних, а не використовують готові рішення, такі як дедукція на основі правил або регресії більш традиційної статистики. (Лоурі 2017: 4)

Іншими словами, статистика та математика мають бути доповнені досвідом у програмуванні та комп'ютерній інженерії. Сукупність навичок, витлумачених таким чином, призводить до специфічного епістемологічного підходу до дослідження, який загалом характеризується наголосом на засобах дослідження як найважливішому рушії дослідницьких цілей і результатів. Цей підхід, який Сабіна Леонеллі охарактеризувала як *орієнтований на дані*, передбачає «більше зосередження на процесах, за допомогою яких здійснюється дослідження, ніж на його кінцевих результатах» (Leonelli 2016: 170). З цієї точки зору, процедури, техніки, методи, програмне забезпечення та апаратне забезпечення є основними двигунами дослідження та головним впливом на його результати. Зосереджуючись більш конкретно на обчислювальних системах, Джон Саймонс і Джек Хорнер стверджували, що більша частина дослідження великих даних складається з *наукових досліджень, які інтенсивно займаються програмним забезпеченням,* а не досліджень,

керованих даними: тобто наука, яка залежить від програмного забезпечення для свого проектування, розробки, розгортання та використання, і, таким чином, охоплює процедури, типи міркувань і помилки, які є унікальними для програмного забезпечення, наприклад, проблеми, породжені спробами відобразити величини реального світу на дискретних автоматах або наближення числових операцій (Symons & Horner 2014: 473). Наука, яка інтенсивно займається програмним забезпеченням, мабуть, підтримується *алгоритмічною раціональністю,* зосередженою на здійсненності, практичності та ефективності алгоритмів, які зазвичай оцінюються на основі конкретних ситуацій дослідження (Lowrie 2017).

Людський і штучний інтелект

Алгоритми надзвичайно різноманітні за своїми математичними структурами та концептуальними зобов'язаннями, тому необхідно провести більше філософської роботи над специфікою обчислювальних інструментів і програмного забезпечення, що використовується в науці про дані та пов'язаних із ними додатках. Нові роботи з філософії інформатики пропонують чудовий спосіб вперед (Turner & Angius 2019). Тим не менш, зрозуміло, що те, чи буде певний алгоритм успішно застосовуватися до наявних даних, залежить від факторів, які неможливо контролювати за допомогою статистичних чи навіть обчислювальних методів: наприклад, розмір, структура та формат даних, природа класифікатори, які використовуються для поділу даних, складність меж прийняття рішень і самі цілі дослідження.

У сильній критиці, заснованій на філософії математики, Крістіан Калуде та Джузеппе Лонго стверджували, що існує фундаментальна проблема з припущенням, що більше даних обов'язково дасть більше інформації:

дуже великі бази даних повинні містити довільні кореляції. Ці кореляції виникають лише через розмір, а не через природу даних. (Calude & Longo 2017: 595)

Вони прийшли до висновку, що аналіз великих даних за визначенням не здатний відрізнити хибні кореляції від значущих і тому становить загрозу для наукових досліджень. Пов'язане занепокоєння, яке інколи називають «прокляттям розмірності» дослідниками обробки даних, стосується того, наскільки аналіз певного набору даних можна збільшити за складністю та за кількістю змінних, що розглядаються. Добре відомо, що чим більше вимірів враховується при класифікації вибірок, наприклад, тим більший набір даних, на якому такі виміри можна точно узагальнити. Це демонструє постійну, тісну залежність між обсягом і якістю даних, з одного боку, і типом і широтою дослідницьких питань, для яких дані повинні служити доказами, з іншого боку.

Визначення відповідності між методами логічного висновку та даними вимагає високого рівня знань і контекстуального судження (ситуація, відома в машинному навчанні як «теорема про відсутність безкоштовного обіду»). Дійсно, надмірна залежність від програмного забезпечення для висновків і моделювання даних може призвести до дуже проблематичних результатів. Саймонс і Хорнер відзначають, що використання складного програмного забезпечення в аналізі великих даних робить межі похибки невизначеними, оскільки немає чіткого способу їх статистичного тестування (Саймонс і Хорнер 2014: 473). Складність шляху програм із високою умовністю накладає обмеження на стандартні методи виправлення помилок. Як наслідок, не існує ефективного методу для характеристики розподілу помилок у програмному забезпеченні, окрім тестування всіх шляхів у коді, що є нереалістичним і важкорозв'язаним у переважній більшості випадків через складність коду.

Замість того, щоб діяти як заміна, ефективне та відповідальне використання інструментів штучного інтелекту в аналізі великих даних вимагає стратегічного вправляння людського інтелекту, але для цього системи штучного інтелекту, які застосовуються до великих даних, повинні бути доступними для перевірки та модифікації. Так це чи ні, і хто найкраще кваліфікований для здійснення такого контролю, залишається предметом суперечок. Томас Ніклс стверджував, що все більш складні та розподілені алгоритми, які використовуються для аналізу даних, йдуть слідами давніх наукових спроб вийти за межі людського пізнання. Отримані в результаті епістемічні системи можуть більше не бути

зрозумілими для людей: «інопланетний інтелект», у межах якого «людські здібності більше не є основним критерієм епістемічного успіху» (Ніклз готовий до публікації). Таке необмежене пізнання обіцяє можливість потужного логічного висновку на основі раніше немислимих обсягів даних. Однак труднощі в контекстуалізації та ретельному аналізі таких міркувань ставлять під сумнів надійність результатів. Не тільки алгоритми машинного навчання стають дедалі недоступнішими для оцінювання: крім складності програмного коду, аналіз обчислювальних даних вимагає цілої екосистеми класифікацій, моделей, мереж і інструментів логічного висновку, які зазвичай мають різну історію та цілі, і які пов'язані з один одного – і ефекти, коли вони використовуються разом – далекі від розуміння і цілком можуть бути непростеженими.

Це ставить питання про те, чи знання, створені такими аналітичними системами даних, взагалі зрозумілі людям, і якщо так, то які форми зрозумілості вони дають. Безумовно, отримання знань із великих даних може не передбачати підвищення людського розуміння, особливо якщо розуміння розуміти як епістемічну навичку (de Regt 2017). Це може не бути проблемою для тих, хто чекає на появу нового виду розумних машин, які можуть оволодіти новими когнітивними інструментами так, як не можуть люди. Але, як зазначали Ніклс, Ніколас Решер (1984), Вернер Каллебаут (2012) та інші, навіть у такому випадку «ми б не досягли науки без перспективи» (Ніклз буде випущено). Хоча людські історії та припущення, вплетені в ці системи, може бути важко роз'єднати, вони все одно впливають на їхні результати; і незалежно від того, чи є ці процеси дослідження відкритими для критичного розгляду, їх телос, наслідки та значення для життя на планеті, мабуть, мають бути такими. Як стверджував Ден МакКвіллан (2018), зростаюча автоматизація аналітики великих даних може сприяти прийняттю неоплатонічної *машинної метафізики*, у рамках якої математичні структури, «розкриті» штучним інтелектом, переважатимуть будь-яке звернення до людського досвіду. Лучано Флоріді повторює цю інтуїцію у своєму аналізі того, що він називає *інфосферою* :

Великі можливості, які пропонують інформаційно-комунікаційні технології, супроводжуються величезною інтелектуальною відповідальністю зрозуміти їх і правильно скористатися ними. (2014: vii)

Ці міркування відповідають давній критиці Пола Хемфріса комп'ютерного моделювання як *епістемічно непрозорого* (Humphreys 2004, 2009) – і зокрема його визначення того, що він називає *суттєвою* епістемічною непрозорістю:

Процес по суті є епістемічно непрозорим для *X* тоді і тільки тоді, *коли* для *X неможливо* знати всі епістемічно релевантні елементи процесу. (Хамфріс 2009: 618)

Різні аспекти загальної проблеми епістемічної непрозорості наголошуються в широкому філософському дослідженні про роль моделювання, обчислення та моделювання в науці: наслідки відсутності експериментального доступу до конкретних частин світу, що моделюється, наприклад (Morgan 2005). ; Паркер 2009; Раддер 2009); труднощі у перевірці надійності обчислювальних методів, що використовуються в рамках моделювання (Winsberg 2010; Morrison 2015); зв'язок між непрозорістю та виправданням (Durán & Formanek 2018); форми чорного ящика, пов'язані з механістичними міркуваннями, реалізованими в обчислювальному аналізі (Craver and Darden 2013; Bechtel 2016); і дебати щодо внутрішніх обмежень обчислювальних підходів і відповідного досвіду (Коллінз 1990; Дрейфус 1992). Роман Фрігг і Джуліан Райс стверджували, що такі проблеми не є фундаментальними проблемами для природи дослідження та моделювання, а фактично існують у континуумі з традиційними методологічними проблемами, добре відомими в науці (Frigg & Reiss 2009). Незалежно від того, погоджується хтось із цією позицією чи ні (Humphreys 2009; Beisbart 2012), аналіз великих даних явно розширює можливості обчислювальних і статистичних методів, таким чином підкреслюючи межі того, що навіть технологічно вдосконалені люди здатні знати та розуміти.

Природа (великих) даних

Таким чином, дослідження аналізу великих даних проливає світло на елементи дослідницького процесу, які неможливо повністю контролювати, раціоналізувати чи навіть розглянути за допомогою офіційних інструментів.

Одним із таких елементів є робота, необхідна для представлення емпіричних даних у машиночитаному форматі, сумісному з наявним програмним забезпеченням та аналітичними інструментами. Дані потрібно відібрати, очистити та підготувати для статистичного та обчислювального аналізу. Процеси, пов'язані з відокремленням даних від шуму, кластеризацією даних, щоб їх можна було простежити, та інтеграцією даних різних форматів виявилися дуже складними та теоретично структурованими, як продемонстрували, наприклад, Джеймс Макаллістер (1997, 2007, 2011) та Ульяна Фіст. (2011) робота над моделями даних, порівняння Марселем Бумансом і Леонеллі принципів кластеризації в різних галузях (готується до публікації), а також аналізи особливостей наборів даних Джеймсом Ґріземером (готується до друку) і Мері Морган (готується до друку). Суппес був настільки стурбований тим, що він назвав «дивовижною складністю» виробництва та обробки даних, що він хвилювався, що філософи не оцінять способи, якими статистика може і допомагає вченим абстрагувати дані від такої складності. Він описав велику групу дослідницьких компонентів і заходів, які використовуються для підготовки даних для моделювання, як «прагматичні аспекти», що охоплюють «кожне інтуїтивне розгляд експериментального плану, що не передбачає формальної статистики» (Suppes 1962: 258), і позиціонує їх як найнижчий рівень його ієрархія моделей – на протилежному кінці її вершини, якою є моделі теорії. Незважаючи на нещодавні спроби реабілітації методології індуктивно-статистичного моделювання та логічного висновку (Mayo & Spanos 2009b), цей підхід поділяється багатьма філософами, які вважають процеси виробництва та обробки даних настільки хаотичними, що не піддаються систематичному аналізу. Це пояснює, чому дані отримали так мало уваги у філософії науки порівняно з моделями та теорією.

Однак питання про те, як дані визначаються та ідентифікуються, є вирішальним для розуміння ролі великих даних у наукових дослідженнях. Давайте тепер розглянемо дві філософські точки зору – *репрезентативну* та *реляційну* – обидва сумісні з появою великих даних, але при цьому акцентуємо увагу на різних аспектах цього явища, що має значні наслідки для розуміння ролі даних у висновках. і, як ми побачимо в наступному розділі, як доказ. Репрезентативний *погляд* тлумачить дані як надійні уявлення про реальність, створені через взаємодію між людьми та світом. Взаємодії, які генерують дані, можуть відбуватися в будь-якому соціальному середовищі незалежно від цілей дослідження. Приклади варіюються від біолога, який вимірює окружність клітини в лабораторії та записує результат у файл Excel, до вчителя, який підраховує кількість учнів у своєму класі та записує це в класний журнал. Даними в цих взаємодіях вважаються об'єкти, створені в процесі опису та/або вимірювання світу. Ці об'єкти можуть бути цифровими (файл Excel) або фізичними (реєстр класів) і формувати відбиток певної взаємодії з природним світом. Цей відбиток – «слід» або «мітка», за словами Яна Хекінга (1992) і Ханса-Йорга Рейнбергера (2011), відповідно, є важливою точкою відліку для аналітичного дослідження та для отримання нових ідей. Ось чому дані формують законну основу для емпіричного знання: виробництво даних еквівалентно «захопленню» особливостей світу, які можна використовувати для систематичного вивчення. Відповідно до репрезентативного підходу, дані – це об'єкти з фіксованим і незмінним змістом, значення яких через те, що вони представляють реальність, необхідно досліджувати та розкривати крок за кроком за допомогою адекватних методів висновку. Дані, що документують форму клітини, можна моделювати, щоб перевірити відповідність форми еластичності, проникності та стійкості клітин, створюючи доказову базу для розуміння передачі сигналів між клітинами та розвитку. Отримані дані підрахунку учнів у класі можна об'єднати з аналогічними даними, зібраними в інших школах, створюючи доказову базу для оцінки щільності учнів у цьому районі та частоти відвідування ними школи.

Це відображає інтуїцію про те, що дані, особливо коли вони надходять у формі числових вимірювань або зображень, таких як фотографії, якимось чином віддзеркалюють явища, для документування яких вони створені, створюючи таким чином моментальний знімок цих явищ, який можна вивчати в контрольованих умовах. досліджень. Це також відображає ідею даних як «необроблених» продуктів дослідження, які максимально наближені до безпосереднього знання реальності. Це має сенс істинного значення, яке іноді приписують даним як неспростовним джерелам доказів – ідея Поппера про те, що якщо знайдено дані, що підтверджують дане твердження, то це твердження підтверджується як істинне принаймні до тих пір, поки не знайдено інших даних, спростувати це. У цій точці зору дані являють собою об'єктивну основу для отримання знань, і саме ця об'єктивність – здатність отримувати знання з людського досвіду, виходячи за його межі – робить знання емпіричними. Ця позиція добре узгоджується з ідеєю, що великі дані є цінними для науки, оскільки вони сприяють індуктивному накопиченню знань (у широкому розумінні): збір даних, зібраних за допомогою надійних методів, створює гору фактів, готових до аналізу, і чим більше фактів створені та пов'язані один з одним, тим більше знань можна отримати.

Філософи давно визнали, що дані не говорять самі за себе, а різні типи даних вимагають різних інструментів для аналізу та підготовки для інтерпретації (Bogen 2009 [2013]). Згідно з репрезентативною точкою зору, існують правильні та неправильні способи інтерпретації даних, які особи, відповідальні за аналіз даних, повинні розкрити. Але що таке «правильна» інтерпретація у сфері великих даних, де дані послідовно розглядаються як мобільні об'єкти, які, принаймні в принципі, можна повторно використовувати незліченною кількістю способів і для досягнення різних цілей? Можливо, більше, ніж будь-коли в історії науки, нинішня мобілізація та повторне використання великих даних підкреслює ступінь, до якого інтерпретація даних – а разом з цим і будь-які дані, які представляють – можуть відрізнятися залежно від концептуальних, матеріальних і соціальні умови розслідування. Аналіз того, як великі дані переміщуються між контекстами, показує, що очікування та здібності тих, хто бере участь, визначають не лише спосіб інтерпретації даних, але й те, що в першу чергу вважається «даними» (Леонеллі та Темпіні, готові до публікації). Представницький погляд на дані як на об'єкти з фіксованим і контекстуально незалежним значенням суперечить цим спостереженням.

Альтернативний підхід полягає в тому, щоб прийняти ці висновки та повністю відмовитися від ідеї даних як фіксованого відображення реальності. У рамках *реляційного погляду* дані – це об'єкти, які розглядаються як потенційні чи фактичні докази наукових тверджень у спосіб, який, принаймні в принципі, можна ретельно перевірити та врахувати (Leonelli 2016). Значення, яке надається даним, залежить від їхнього походження, фізичних особливостей і того, що ці характеристики представляють, а також мотивів та інструментів, які використовуються для їх візуалізації та захисту конкретних інтерпретацій. Таким чином, надійність даних залежить від достовірності та чіткості процесів, які використовуються для їх отримання та аналізу. Подання даних; спосіб їх визначення, відбору та включення (або виключення) до баз даних; та інформація, яка надається користувачам для їх реконтекстуалізації, є фундаментальною для отримання знань і значно впливають на їх зміст. Наприклад, зміни у форматі даних – що, очевидно, пов'язано з процедурами оцифрування, стиснення даних або архівування – можуть мати значний вплив на те, де, коли та хто використовує дані як джерело знань.

Цей фреймворк визнає, що будь-який об'єкт можна використовувати як даний або припинити використовувати як такий, залежно від обставин – міркування, знайоме аналітикам великих даних, які звикли вибирати та змішувати дані, що надходять із великої різноманітності джерел. Реляційний погляд також пояснює, як, залежно від дослідницької перспективи, що його інтерпретує, один і той самий набір даних може використовуватися для представлення різних аспектів світу («явища», як їх знаменито охарактеризували Джеймс Боген і Джеймс Вудворд, 1988). Розглядаючи повний цикл наукового дослідження з точки зору виробництва та аналізу даних, саме на етапі *моделювання* даних даним надається певна

репрезентативна цінність (Leonelli 2019b).

Реляційний погляд на дані спонукає звернути увагу на історію даних, висвітлюючи їх постійну еволюцію, а іноді й радикальні зміни, а також вплив цієї функції на здатність даних підтверджувати чи спростовувати гіпотези. Це пояснює критичну важливість документування процесів управління даними та перетворення, особливо з великими даними, які передаються далеко й широко цифровими каналами та групуються та інтерпретуються різними способами та форматами. Це також пояснює зростаюче визнання досвіду тих, хто створює, курує та аналізує дані, як необхідного для ефективної інтерпретації великих даних у межах науки та за її межами; і нерозривний зв'язок між соціальними та етичними проблемами щодо потенційного впливу обміну даними та науковими проблемами щодо якості, дійсності та безпеки даних (boyd & Crawford 2012; Tempini & Leonelli, 2018).

Залежно від погляду на дані, очікування щодо того, що великі дані можуть зробити для науки, різко відрізнятимуться. Репрезентативний погляд враховує ідею великих даних як забезпечення найповнішої, надійної та генеративної бази знань, яку будь-коли бачили в історії науки, завдяки її величезному розміру та неоднорідності. Реляційний погляд не бере на себе таких зобов'язань, зосереджуючись замість цього на тому, які висновки робляться з таких даних у будь-який момент, як і чому.

Великі дані та докази

Єдине, в чому погоджуються репрезентативні та реляційні погляди, – це ключова епістемічна роль даних як емпіричних доказів претензій щодо знань або втручань. Хоча існує велика кількість філософської літератури про природу доказів (наприклад, Achinstein 2001; Reiss 2015; Kelly 2016), проте зв'язку між даними та доказами приділено менше уваги. Можливо, це пов'язано з неявним прийняттям багатьма філософами репрезентативного погляду на дані. У рамках репрезентативного погляду ідентифікація того, що вважається даними, передує дослідженню того, для чого ці дані можуть бути доказами: іншими словами, дані є «даними», як вказує етимологія слова, а методи висновку відповідають за визначення того, чи і як дані, доступні слідчим, можуть бути використані як докази та для чого. Таким чином, фокус філософської уваги зосереджений на формальних методах виділення помилок і оманливих інтерпретацій, а також на ймовірнісному та/або пояснювальному відношенні між тим, що без проблем вважається набором доказів, і даною гіпотезою. Тому велика частина обширних філософських робіт про докази взагалі уникає терміна «дані». Основна робота Пітера Ахінштейна є яскравим прикладом: у ній обговорюються спостережувані факти та експериментальні результати, а також те, чи будуть у вчених підстави вірити таким фактам і за яких умов, але в ній не згадується дані та відповідна практика обробки (Achinstein 2001).

Навпаки, у реляційному поданні об'єкт можна ідентифікувати як дані лише тоді, коли він розглядається як такий, що має цінність як доказ. Докази стають категорією ідентифікації даних, а не категорією використання даних, як у репрезентативному погляді (Canali 2019). Таким чином, доказ є конститутивним для самого поняття даних і не може бути відокремлений від нього. Це передбачає прийняття того, що умови, за яких даний об'єкт може служити доказом – і, отже, розглядатися як дані – можуть змінюватися; і якщо ця доказова роль повністю припиниться, об'єкт знову перетвориться на звичайний елемент, що не є датою. Наприклад, фотографія рослини, зроблена туристом у віддаленому регіоні, може стати доказом для дослідження морфології рослин з цієї конкретної місцевості; проте більшість фотографій рослин ніколи не розглядаються як докази для дослідження особливостей і функціонування світу, а з тих, які є, багато з них можуть згодом бути відкинуті як нецікаві або більше не стосуються поставлених питань.

Ця точка зору враховує мобільність і перепрофілювання, які характеризують використання великих даних, а також можливість того, що об'єкти, які спочатку не були згенеровані для того, щоб служити доказами, можуть бути згодом прийняті як такі. Розглянемо «мінімальний науковий принцип доказів» Мейо та Спаноса, який вони визначають таким чином:

Дані $\mathbf{x}_0$ надають погані докази для *H,* якщо вони є результатом методу чи процедури,

які мають незначну або зовсім не здатну знаходити недоліки в *H*, навіть якщо *H* є хибним. (Mayo & Spanos 2009b)

Цей принцип сумісний із реляційним поглядом на дані, оскільки він включає випадки, коли методи, що використовуються для генерування та обробки даних, можливо, не були спрямовані на перевірку гіпотези H: все, що він вимагає, це те, щоб такі методи могли бути релевантними для тестування H, у той момент, коли дані використовуються як докази для H (я повернуся до ролі гіпотез у обробці доказів у наступному розділі).

Розробка структурної схеми

Для побудови структурної схеми ми розглянемо автоматизовані системи наукових досліджень, яка зображена на рисунку 1.

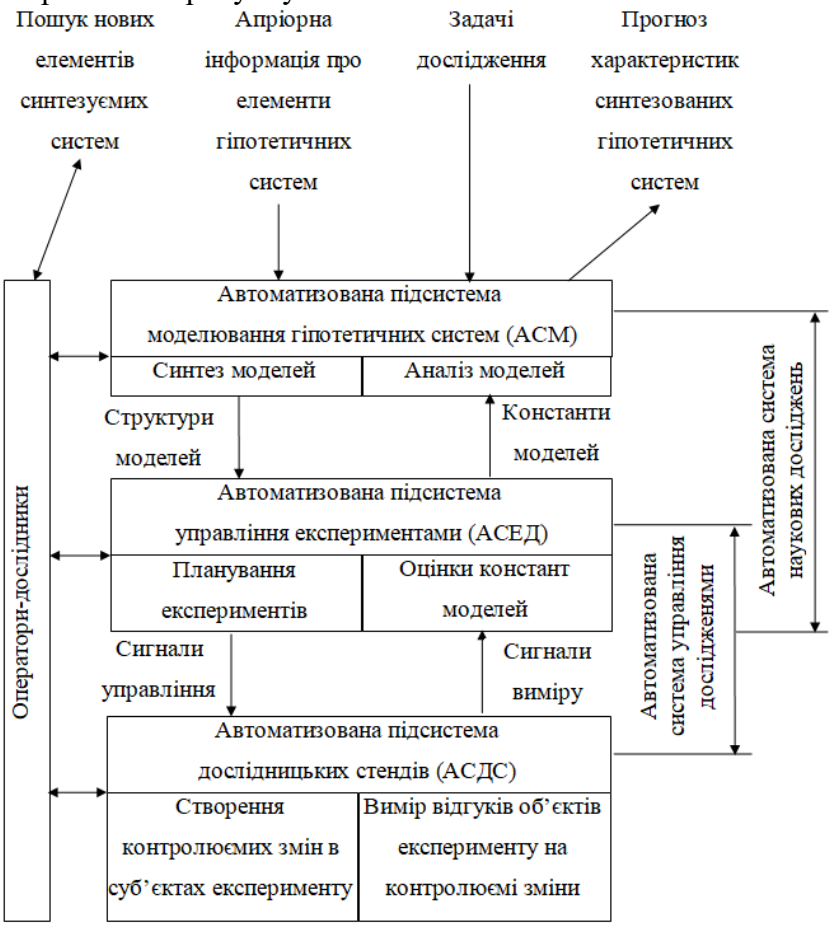

Рисунок 1 – Структурна схема системи

Висновки. У роботі наведені теоретичне узагальнення й рішення наукового завдання дослідження методів big data наукових досліджень.

Рішення даного завдання полягало у вирішенні наступних задач:

– Був проведений огляд існуючих систем big data наукових досліджень.
– Досліджена система big data наукових досліджень.
– На основі отриманих результатів досліджень створена програмна реалізація системи big data наукових досліджень.

Розроблені під час виконання випускної кваліфікаційної роботи за другим (магістерським) рівнем вищої освіти алгоритми дозволяють успішно вирішувати завдання big data наукових досліджень.

Проведено аналіз предметної галузі в ході якого були виявлені об'єкти, взаємодія яких носить істотний характер для функціональної діяльності предметної галузі, і їхні основні характеристики; побудована алгоритм і вибраний середовище розробки.

Розроблене програмне забезпечення має простий, дружній та зручний інтерфейс користувача, що забезпечує легкість у освоєнні роботи програмного продукту, зручність у використанні, і не потребує особливих спеціальних знань.

Список літератури

1. Smirnov O., Kovalenko O., Kovalenko A., Kavun S. «Quantitative Risk Assessment Method Development in the Context of the SDLC-model». 2021 IEEE 8th International Conference on Problems of Infocommunications, Science and Technology (PIC S&T), 2021, pp. 203-208, doi: 10.1109/PICST54195.2021.9772143 (Scopus).
2. Smirnov O., Neskorodieva T., Fedorov E., Rymar P. «Neural Network Modeling Method of Transformations Data of Audit Production with Returnable Waste». CEUR Workshop Proceedings Volume 3101, 2021, Pages 192-207. (Scopus).
3. Smirnov O., Kuznetsov A., Kiian A., Kuznetsova K. «Data hiding scheme based on spread sequence addressing». CEUR Workshop Proceedings Volume 2805, 2020, Pages 44-58. (Scopus).
4. Smirnov, O., Kuznetsov, A., Potii, O., Poluyanenko, N., Stelnyk, I., Mialkovsky, D. «Combining and filtering functions in the framework of nonlinear-feedback shift register». International Journal of Computing; 2020, Volume 19, Issue 2 – Research Institute for Intelligent Computer Systems – 2020. – P. 247-256. (Scopus).
5. Smirnov O., Kuznetsov A., Kiian A., Kuznetsova T. «Non-binary constant weight coding technique». CEUR Workshop Proceedings. Volume 2740, 2020, Pages 102-114. (Scopus).
6. Smirnov O.A., Alimseitova Zh., Adranova A., Akhmetov B., Lakhno V., Zhilkishbayeva G. «Models and algorithms for ensuring functional stability and cybersecurity of virtual cloud resources». Journal of theoretical and applied information technology Vol.98. No 21, 2020, P. 3334-3346. (Scopus).
7. Smirnov O., Kuznetsov A., Arischenko A., Chepurko I., Onikiychuk A., Kuznetsova T. «Pseudorandom sequences for spread spectrum image steganography». CEUR Workshop Proceedings Volume 2654, 2020, Pages 122-131. (Scopus).
8. Smirnov O., Kuznetsov A., Kovalchuk D., Kuznetsova T. «New technique for data hiding in cover images using adaptively generated pseudorandom sequences». CEUR Workshop Proceedings Volume 2654, 2020, Pages 1-14. (Scopus).
9. Smirnov O., Lutsenko M., Kuznetsov A., Kiian A., Kuznetsova T., «Biometric cryptosystems: overview, state-of-the-art and perspective directions». Lecture Notes in Networks and Systems, vol 152. Springer, Cham. 2021, pp 66-84. (Scopus).
10. Smirnov O., Kuznetsov A., Pushkar'ov A., Serhiienko R., Babenko V., Kuznetsova T., «Representation of Cascade Codes in the Frequency Domain». In: Radivilova T., Ageyev D., Kryvinska N. (eds) Data-Centric Business and Applications. Lecture Notes on Data Engineering and Communications Technologies, vol 48. Springer, Cham. 2021. pp 557-587. (Scopus).
11. Smirnov, O., Markovets, O. Vovk, N., Turchyn, Y., «Model of informational support for social network administrators' content creation». CEUR Workshop Proceedings Volume 2616, 2020, Pages 125-136. (Scopus).
12. Smirnov, O., Drieieva, H., Drieiev, O., Polishchuk, Y., Brzhanov, R., Aleksander, M. «Method of fractal traffic generation by a model of generator on the graph». CEUR Workshop Proceedings Volume 2616, 2020, Pages 366-379. (Scopus).
13. Smirnov, O., Shekhanin, K., Kuznetsov, A., Krasnobayev, V. «Detecting Hidden Information in FAT». International Journal of Computer Network and Information Security (IJCNIS). Vol. 12, No. 3, 2020. PP.33-43. (Scopus).
14. Smirnov, O., Drieieva, H., Drieiev, O., Simakhin, V., Bondar, S., Odarchenko, R. «Managing multifractal properties of the binary sequence generated with the Markov chains», CEUR Workshop Proceedings Volume 2608, 2020, Pages 633-645. (Scopus).
15. Smirnov, O., Kuznetsov, A., Gorbacheva, L., Babenko, V., «Hiding data in images using a pseudo-random sequence», CEUR Workshop Proceedings Volume 2608, 2020, Pages 646-660., (Scopus).
16. Smirnov, O., Kuznetsov, A., Kolovanova, I., Kuznetsova, T., «Noise immunity of the algebraic geometric codes». International Journal of Computing; 2019, Volume 18, Issue 4 – Research Institute for Intelligent Computer Systems – 2019. – P. 393-407. (Scopus).
17. Smirnov, O., Kuznetsov, A., Kiian, A., Kuznetsova, K., Ivko, T., Prokopovych-Tkachenko, D., «Soft Decoding Based on Ordered Subsets of Verification Equations of Turbo-Productive Codes», CEUR Workshop Proceedings Volume 2353, CEUR Workshop Proceedings 2019, Pages 873-884. (Scopus).
18. Smirnov, O., Kuznetsov, A., Prokopovych-Tkachenko, D. «Hiding Data in Images Using a Pseudo-Random Sequence». ISCI'2020: Information Security in Critical Infrastructures. Collective monograph. Edited by Ivan D. Gorbenko, Victor A. Krasnobayev and Alexandr A. Kuznetsov. ASC Academic Publishing, USA, 2020. pp. 46-59. – ISBN: 978-1-7362833-0-1 (Hardback), ISBN: 978-1-7362833-1-8 (Ebook).
19. Smirnov, O., Kuznetsov, A., Shekhanin, K., Chepurko, I. Detecting Hidden Information in FAT. Монографія: In.: ISCI'2019: Information Security in Critical Infrastructures. Collective monograph. Edited by Ivan D. Gorbenko and Alexandr A. Kuznetsov, ASC Academic Publishing, USA, 2019, pp. 412-429. – ISBN: 978-0-9989826-8-7 (Hardback), ISBN: 978-0-9989826-9-4 (Ebook).
20. Smirnov, O., Kuznetsov, A., Kuznetsova., K. Synthesis of Discrete Signals with Improved Correlation Properties. Монографія: In.: ISCI'2019: Information Security in Critical Infrastructures. Collective monograph. Edited by Ivan D. Gorbenko and Alexandr A. Kuznetsov, ASC Academic Publishing, USA, 2019, pp. 281-299. – ISBN: 978-0-9989826-8-7 (Hardback), ISBN: 978-0-9989826-9-4 (Ebook).