פרטיות נתונים והתממה

קורס

מגישים:
נוי חיות **********
ינון סגל **********
אמיר חיר **********

תוכן

תקציר

בעבודה זאת, בדקנו מהי התממה (אנונימיזציה) של נתונים ומדוע היא חשובה. בחנו את הטכניקות השונות להתממה אותן אפשר לממש על מנת לשמור על נתונים אנונימיים ולהגן על בטיחות הלקוחות. כמו כן פירטנו על הטכניקות השונות בתהליך ההתממה, סיכונים הכרוכים בתהליך זה ויתרונות מול חסרונות בהתממה וביצענו השוואה ביניהם. בנוסף סיקרנו והסברנו את נושא ההגנה על נתונים פרטיים בענן. לבסוף המלצנו על טכניקת ההתממה העדיפה בעינינו.

סקירת נושא

ההתממה, הידועה גם בשמות "אנונימיזציה" או "עילום נתונים מזהים", היא תהליך מרכזי בתחום הפרטיות ואבטחת המידע. במהלך התממה, יוצרים/ משנים מאגר נתונים על ידי הסרת פרטים מזהים, או החלפתם בערכים שונים מהנתונים המקוריים. התוצאה היא מאגר נתונים חדש, שהוסרה ממנו היכולת לזהות אנשים באופן ישיר.

המטרה העיקרית של התממה היא לאפשר שיתוף מידע רב יותר בצורה שאינה סוכנת לפרטיות האנשים שבעצם הם הקשורים לנתונים. זהו תכלית מרכזית בתחום המחקר, רפואה, ושטחים רבים נוספים, כשהידע המצוין בנתונים יכול לשפר את החקר, התכנון והפעולה, אך במקביל יש לשמור על פרטיות האנשים שנכללים בנתונים. תהליך ההתממה כולל החלפת ערכים, מיפויים עממיים, ושימוש בטכנולוגיות שונות כדי להבטיח שהנתונים שנמצאים בשימוש יהיו בלתי זהירים לזיהוי ישיר.

אולם, יש להדגיש כי התממה אינה תהליך מוחלט וסופי, ולעיתים יכולים פרטים מזהים להישאר בנתונים בצורה חלקית ובכך לסכן באופן בלתי ישיר את פרטיותם. לכן, עיקרי האבטחה והמקצועיות בתהליך התממה נמצאים בשמירה על האינטגריטיות והאנונימיות של הנתונים, ובהבטחת כך שאף אדם מוביל אינו יכול לזהות אנשים מהנתונים באמצעות טכניקות סטטיסטיות או אחרות.

סקירת ספרות

הגנת הפרטיות ואנונימזציה

ישנם סיכונים רבים בחדירה לפרטיות, גישה לא מורשית עלולה להוביל לחשיפה של מידע אישי רגיש וכתוצאה מכך לגניבת זהות, הונאה אפשרית או חשיפת מידע שלא נועד לעיני הציבור.
בנוסף גורמים פנימיים בעלי גישה לנתונים רגישים עלולים לעשות בהם שימוש לרעה או לסכן את סודיותו.

למרות הסיכון, יש צורך הולך וגובר בשיתוף מאגרי נתונים המכילים מידע אישי על פני מספר מסדי נתונים מבוזרים ופרטיים. עם זאת, שיתוף נתונים כזה כפוף למגבלות הנ"ל בסיכון חדירה לפרטיות וחשיפת נתונים, בנוסף, מוסדות אינם צריכים לחשוף את מסדי הנתונים שלהם זה לזה מלבד תוצאות השאילתה. השימוש באנונימזציה המאפשר לספקי נתונים עצמאיים לבנות מסד נתונים אנונימי וירטואלי תוך שמירה על פרטיות אילוצים. (Jurczyk P' X', 2009)

לכן נדרשות פעולות להגנת הפרטיות אשר פועלת על פי שני הנחות יסוד: שמירה על אנשים מפני גישה בלתי מורשית למידע האישי שלהם ושמירה על מצע האמון במערכות דיגיטליות. על ידי אבטחת מידע רגיש מעיניים סקרניות, אנשים מוסמכים לעסוק בתחום הדיגיטלי מבלי לחשוש שהנתונים שלהם ייפלו לידיים הלא נכונות. יתר על כן, עצם מארג האמון שעומד בבסיס האינטראקציות הדיגיטליות נרקם באמצעות יישום אמצעים מחמירים להגנת הפרטיות. כאשר משתמשים יכולים לסמוך על כך שהנתונים שלהם יטופלו באחריות, יש סיכוי גבוה יותר שהם יתקשרו עם פלטפורמות דיגיטליות, ישתפו מידע וישתפו פעולה בביטחון.

הגנת הפרטיות כוללת מערך של אמצעים, אסטרטגיות וטכנולוגיות שנועדו להגן על המידע האישי של אנשים מגישה לא מורשית ולהפחית את סיכוני החשיפה. מאיסוף נתונים ועד אחסון, עיבוד ושידור, אסטרטגיות הגנת הפרטיות שואפות לשמור על סודיות ולהגביל את השימוש לרעה הפוטנציאלי בנתונים אישיים.

אנונימיזציה עומדת כאסטרטגיה מרכזית המוכרת ביכולתה להפוך נתונים נטולי מזהים, ובכך לשמור על סודיותם של אנשים. טכניקה זו היא גישה מתוחכמת שנרתמה כדי להגן על פרטיות הפרט על ידי ביטול או שינוי של מידע אישי מזהה בתוך מערך נתונים. המטרה הבסיסית שלו היא למגר כל מרכיב שיכול להוביל לזיהוי של אנשים ספציפיים, כל זאת תוך שמירה על אמינות במערך הנתונים. תהליך סבוך זה כרוך במניפולציה קפדנית של נתונים, המבטיחה שהנבדקים המקוריים מהם מקור המידע יישארו בלתי ניתנים לאיתור כשהם מונחים לצד הנתונים האנונימיים, גם כשהם משלימים פיסות מידע שונות אחרות.

אנונימיזציה משמשת אמצעי הגנה מרכזי המדגיש את האחריות האתית של ארגונים ושומרי נתונים על ידי כריתה או שינוי נתונים. מתודולוגיה זו שומרת על קדושת פרטיות הפרט בעידן שבו תובנות מונעות נתונים מעודדות חדשנות וקבלת החלטות. עוצמתו נעוצה באיזון העדין שהוא מייצר בין הפיכת הנתונים כמעט בלתי ניתנים לפענוח הנוגעים לזהויות אינדיבידואליות, תוך שמירה על מהות הנתונים למטרות אנליטיות.

למעשה, אנונימיזציה פועלת כאבן יסוד של ניהול נתונים אחראי, תוך התאמה עם תקנות הגנת המידע המתפתחות ושיקולים אתיים. על ידי הסתרת נתונים במעטה האנונימיות הזה, ארגונים לא רק מכבדים את זכויות הפרטיות של אנשים, אלא גם מטפחים סביבה שבה ניתן לרתום תובנות מונעות נתונים לקידום חברתי תוך מזעור הסיכונים הכרוכים בחשיפה לא מורשית של מידע רגיש. (Wallace, 2016)

טכניקות התממה

הסרת תכונות\פיצ'רים (הדחקה) (Suppression)

הטכניקה מסירה את התכונות\פיצ'רים מסוימים מהמערך נתונים\בסיס נתונים (dataset) , הטכניקה צריכה לקרות בכל פעם שיש תכונה לא רלוונטי או הכרחי לניתוח או בכל פעם שהוא אי אפשר לעשות אנונימיזציה\התממה בדרך אחרת . הטכניקה יכול להתרחש גם עבור רשומת dataset שלם המשפיעה על מספר פיצ'רים .היתרון העיקרי של הטכניקה הוא שכאשר מוחקים\מסירים תכונה\פיצ'ר הלא רלוונטי , זה הופך להיות בלתי אפשרי לאחזר את המידע. ניתן להתבונן בדוגמה , בדוגמה מתוארת טבלה של dataset שמכיל שלוש תכונות\פיצ'רים\עמודות : שם תלמיד , שם תלמיד , שם מאמן , ציון . (JF Marques, J Bernardino , 2020)

מאגר לפני טכניקת התממה :

Student	Trainer	Test Score
John	Anna	93
Nicholas	Paul	86
Josh	Paul	54
Taylor	Anna	78

אחרי טכניקת התממה , הפיצ'ר "שם תלמיד" כפי שאנחנו יכולים לראות בטבלה השנייה שנמצאת למטה. (JF Marques, J Bernardino , 2020)

מאגר אחרי התממה :

Trainer	Test Score
Anna	93
Paul	86
Paul	54
Anna	78

מיסוך נתונים (Data Masking)

מה הטכניקה עושה?

הטכניקה משמש להחלפת התווים המורכבת מכיסוי\מיסוך תווים של ערך של נתונים ע"י החלפת התווים האלה בסמל מוגדר מראש (לדוגמה, על ידי X או *). החלפה זו יכולה להיות חלקית , הסתרה חלקית של טקסט או תכונה , שעלולים מספיק כדי להפוך את הנתונים לאנונימיים . (JF Marques, J Bernardino , 2020)

למה להשתמש בטכניקה זאת?

1. דרישות משפטיות : הסביבה הרגולטורית המקיפה את החובות והמחויבות של בעל הנתונים כדי להגן על המידע, שהמידע הופך להיות לקפדני יותר ויותר כמעט בכל תחום שיפוטי משפטי. זוהי הנחה די בטוחה שהתקנים לאבטחה ותחזוקה של נתונים יהפכו למחמירים יותר ויותר בעתיד. (Goyal, C. , 2015).

2.אובדן אמון ואסונות יחסי ציבור : ניתן לומר באופן סביר ברוב המיקומים , שאם מתרחש הדלפת נתונים בארגון מסוים , אזי הסנקציות המשפטיות הפורמלית המופעלות ע"י גופים ממשלתיים אינן הבעיה היחידה שנתמודד איתה. ייתכן זה אפילו לא הדאגה הגדולה ביותר שלנו. חשיפה לא הולמת של נתונים , בין אם מקרית או זדונית , עלולה להיות בעלת השלכות הרסניות. לדוגמה מה יעלה לארגון אם לקוחות פוטנציאלים לא יהיו מוכנים לספק מידע רגיש לחברה מסוימת כי הם קוראים מאמר על הדלפת ובריחת נתונים בעיתון. התמודדות עם יחסי ציבור שלאחר ראיית שם החברות בעיתונות לא תהיה זולה. גם לא צריך הרבה דמיון כדי להבין שההנהלה הבכירה לא תשמח על הצורך לקיים מסיבת עיתונאים כדי להרגיע את הציבור מחדש. עלויות יחסי הציבור של בריחה נתונים בדרך כלל עולות בהרבה מהסנקציות המוטלות ע"י ארגונים ממשלתיים.(Goyal, C. , 2015)

איזה בעיה הטכניקה פותרת את הבעיה ומה המהות של הטכניקה?

טכניקת מיסוך נתונים פותרת את בעיות אבטחת הנתונים בסביבת הבדיקה. המהות למיסוך נתונים הוא החלפת מידע רגיש המועתק ממאגרי מידע ייצור לבסיסי נתונים בדיקה או מאגרי מידע שאינם ייצור בנתונים מציאותיים, אך משופשפים, המבוססים על כללי מיסוך.(Li, M., Liu, Z., Jia, C., & Dong, Z. , 2013).

ניתן להתבונן לדוגמה :

לפני התממה נתון לנו מאגר שמכיל פיצ'רים של מספר מיקוד , מספר ממוצע של הזמנות.

Before anonymisation:

Postal Code	Average No. of Orders/month
100111	2
200222	8
300333	1

After suppressing the "student" attribute:

Postal Code	Average No. of Orders/month
10xxxx	2
20xxxx	8
30xxxx	1

אחרי התממה , ניתן לראות שפיצ'ר של קוד מיקוד עבר תהליך התממה בטכניקה של מיסוך נתונים , שבכל תצפית של מיקוד ניתן לראות שחלק מהקוד המיקוד מוסתרים וזה הרעיון בטכניקה.(JF Marques, J Bernardino , 2020).

סוגי מיסוך נתונים (Data Masking Types)

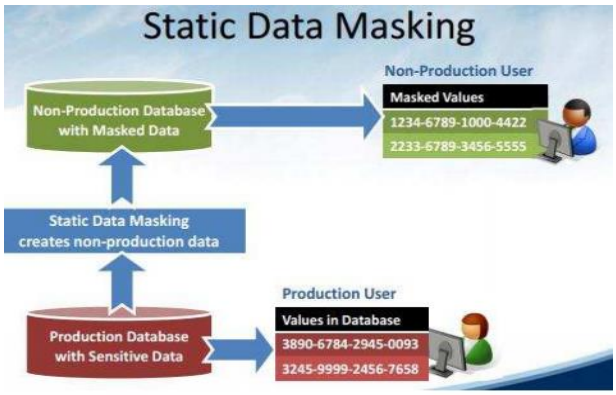
מיסוך נתונים סטטי

רוב הארגונים משתמשים במיסוך נתונים סטטי כאשר הם יוצרים סביבות בדיקה ופיתוח והיא שיטת המיסוך האפשרית היחידה בעת שימוש במפתחים במיקור חוץ במיקום נפרד או בחברה נפרדת. במצבים הללו , יש צורך לשכפל את

מאגר\מסד נתונים. כלים הללו מבטיחים שכל הנתונים הרגישים יהיו מוסכים(masked) לפני שליחתם מהארגון. עם זאת , זה לא פתרון מלא מכיוון שהוא לא מגן על משתמשים מורשים מפני צפייה וחילוץ לא מורשה.(Goyal, C. , 2015).

<u>תהליך של מיסוך נתונים סטטי :</u>

תרשים של תהליך מיסוך נתונים סטטי :



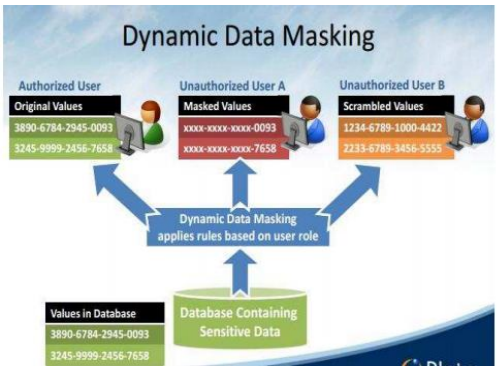
ההסבר לתרשים הוא : 1. מתחיל בשליפת נתונים של Production (ייצור) כקלט , 2.ביצוע טרנספורמציות לביטול זיהוי רשומות ולהסרת מידע רגיש , 3.שמירת על מבנה נתונים ע"י שמירה על שלמות התייחסות בתוך ובין מאגר\מסדי נתונים ,4.מספק נתוני בדיקה מציאותיים באיכות גבוהה לשימוש בסביבות Non-Production.(Goyal, C. , 2015).

מיסוך נתונים דינמי

סוג הטכניקה זאת נועדה לאבטח נתונים בזמן אמת עבור מערכות Production & Non-Production . מיסוך נתונים דינמי עושה את תהליך מיסוך(Masking) את כל הנתונים הרגישים בזמן הגישה אליהם, בזמן אמת והמידע הרגיש לעולם לא יוצא ממסד נתונים. כאשר מסד נתונים או מאגר מסוים מורשה אחר צופים בנתונים בפועל במסד נתונים של Production, נתונים מוסכים(Masked Data) , או נתונים משובשים , כך שהנתונים האמיתיים אינם נחשפים . בדרך זו , בשום שפנים ואופן אף אחד לא נחשף לנתונים פרטיים באמצעות גישה ישירה למסד נתונים. מיסוך נתונים דינמי דורש Proxy , כלומר הוספת רכיב בין שאילתת הנתונים לתגובה. יש לשים לב לחששות הבאים בעת שימוש בפתרונות מיסוך דינמיים של מסד נתונים.(Goyal, C. , 2015).

<u>תהליך של מיסוך נתונים דינמי :</u>

תרשים של תהליך מיסוך נתונים דינמי :



הסבר לתרשים : 1. יצירת שכבת אבטחה נוספת בין מסדי נתונים ויישומים , 2. מיסוך באופן סלקטיבי\אקראי את מידע רגיש ממשתמשים שאינם דורשים ממנו לבצע את עבודתם , 3. הספקת אבטחה דקיקה , מבוססת תפקידים , 4.אפשרות להגדרת תפקידי אבטחה על פני מסדי נתונים ויישומים מרובים.(Goyal, C. , 2015).

החלפת נתונים (Data Swapping/Shuffling)

החלפת נתונים - המכונה לעתים קרובות תמורה וערבוב - מסדרת מחדש את ערכי מאפייני הנתונים כך שהם לא מתאימים למידע המקורי כולל החלפת תכונות (עמודות) הכוללות ערכים מוכרים, כגון תאריך לידה, יכולה להשפיע מאוד על האנונימיזציה.(Team,2023).

היתרון בטכניקה שהיא מקשה על תוקפים לבצע ביטול אנונימיות בגלל חוסר ההתאמה למידע המקורי.(Slavin , 2022).

עם זאת ,טכניקה זו לא תמיד מספקת אנונימיזציה של הנתונים וייתכן שניתן יהיה לארגן אותם מחדש לצורתם המקורית. לכן, יש להשתמש בו בשילוב עם טכניקות אחרות. (JF Marques, J Bernardino , 2020).

הפרעת נתונים (Data perturbation \Noise Addition)

הטכניקה משנה את מערך הנתונים הראשוני באופן שולי ע"י יישום שיטות מספור עגול והוספת רעש אקראי. קבוצת הערכים חייבת להיות פרופורציונלית להפרעה. בסיס קטן יכול לתרום לאנונימיזציה\התממה לקויה , בעוד שבסיס רחב יכול להפחית את השמשיות של מערך הנתונים. לדוגמה , יש להשתמש בבסיס של 5 לעיגול ערכים כמו גיל או מספר בית.(Team , C , 2023)

פסאודונימיזציה (Pseudonymization)

פסאודונימיזציה היא תהליך של הסרת מזהים ממערך נתונים והחלפתם בשם בדוי. המטרה העיקרית של הטכניקה של טכניקת התממה זו היא להבטיח שלא ניתן להתאים נתונים מסוימים לאדם שניתן לזהות , אלא אם כן הם משולבים עם קבוצה נפרדת של מידע. (Dvorin , 2023)

היתרון בטכניקה שהיא שומרת על שלמות הנתונים, הדיוק והדיוק הסטטיסטי ובו בזמן שומרת על סודיות , בנוסף הטכניקה בנתונים שהשתתנו לצרכים למשל : בדיקות , ניתוחים , פיתוח , יצירה ,תוך שמירה על פרטיות הנתונים. (2022 , Slavin).

לדוגמה אם משתמש שולח את השם "ג'ין" במהלך הרישום , מסד הנתונים הראשי יכול פשוט לאחסן אותו כ"אדם 2647". לאחר מכן ניתן לאחסן את האלגוריתם למיפוי אדם 2647 לג'ין במסד נתונים מאובטח אחר. (Dvorin , 2023).

נתונים סינתטיים (Synthetic data)

נתונים סינתטיים הם טכניקת אנונימיזציה\התממה של נתונים הכוללת מידע מיוצר באופן אלגוריתמי שאין לו קשר לאירועים אמיתיים. מודלים מתמטיים נבנים על סמך התבניות של מערך הנתונים המקורי. שיטות סטטיסטיות כמו רגרסיה לינארית , סטיית תקן וחציון בין היתר משמשות כדי להמציא אב טיפוסים סינתטיים.

חלק ממנהלי מסדי נתונים רואים בשיטה זו דרך מפוארת יותר לבצע אנונימזציה\התממה של נתונים במקום לבצע שינויים ישירות בערכות הנתונים המקוריות.(Slavin,2022)

הכללה (Generalization)

הטכניקה כוללת אי הכללה מכוונת של חלקים מסוימים של הנתונים כדי להפוך אותם לפחות מזוהים. בטכניקה זו , הנתונים ישונו לקבוצה של טווחים או אזורים גדולים בגבולות המתאימים.(Team , C , 2023).

לדוגמה אם קיים מערך נתונים המציין את גיל של כל אדם , ניתן להכליל אותו באמצעות קטגוריות כגון 21 עד 25 , ו-26 עד 30. הטכניקה יכולה גם להכליל כתובת ע"י הסרת מספר הבית תוך שמירה על הרחוב שם , עיר או מיקוד .(2023 , Dvorin)

ישנן שתי טכניקות שיכולות להיחשב כהכללה: K-אנונימיות (K-Anonymity) ו-L-גיוון (L-Diversity) . (JF , 2020 Marques, J Bernardino).

K-Anonymity

טכניקה זו מורכבת מקיבוץ הרשומות של K יחידים לקטגוריות מה שהופך אותם ליפול תחת אותם שילובים. (JF , 2020 Marques, J Bernardino).

לדוגמה, אם המאפיינים המזהים הם גיל ומחלה ו-K=3, למערך הנתונים האנונימי בשיטה זו יהיו לפחות 3 רשומות עבור כל שילוב של התכונות המזהות. התחשבות בשני הפרטים בדוגמה המוצגת בתוצאה של k אנונימיות עם K=3 תהיה זו המומחשת בטבלה המתוארת כאן. (JF Marques, J Bernardino , 2020).

Paul	
Age	Disease
21	Heart Disease

Mark	
Age	Disease
38	Cancer

Figure 4: K-Anonymity: original dataset.

Age range	Disease
20-30	Heart Disease
20-30	Heart Disease
20-30	Heart Disease
30-40	Cancer
30-40	Cancer
30-40	Cancer

Figure 5: K-Anonymity: original dataset.

L-Diversity

L-Diversity היא התפתחות של K-Anonymity שבה לפחות L ערכים נפרדים חייבים להתקיים עבור כל קבוצה שווה ותכונה מזהה.כלומר, מובטח שבכל קבוצה שווה לכל תכונה יש לפחות L ערכים שונים. המטרה של טכניקה זו היא להגביל את התרחשותם של מחלקות שקילות עם שונות נמוכה של התכונה. לפיכך, כפולש\פורץ שיש לו גישה לנתונים עבור אדם ספציפי נשאר תמיד עם מידה של אי ודאות. עם זאת, טכניקה זו רגישה להתקפות של הסקה הסתברותית. (JF Marques, J Bernardino , 2020).

הגנה על נתונים פרטיים בענן

הבנה כללית- אחסון בענן

בנוף המתפתח של ההתקדמות הטכנולוגית, עליית הטכנולוגיה בתחום ה IOT (האינטרנט של הדברים), הופעתן של ערים חכמות, טרנספורמציות דיגיטליות בארגונים והזינוק בכלכלה הדיגיטלית העולמית. בתוך מגמות טרנספורמציות אלו, אחסון כמויות עצומות של נתונים הפך לדאגה להרבה ארגונים. בשביל לענות על הצורך בנפחי נתונים מסיביים, מערכות אחסון בענן הופיעו כמרכיב הכרחי בעידן החדש. ממשלות, ארגונים ומשתמשים בודדים כאחד מעבירים באופן פעיל את הנתונים שלהם לענן. בפשטות, אחסון בענן הוא בעצם אחסון נתונים בשרתים מרוחקים מהארגון אליהם המשתמש ניגש דרך האינטרנט, בניגוד לאחסון מקומי. השינוי הזה הביא מספר גדול של יתרונות, כולל הגמישות של הגדלה או ירידה

בכמות האחסון בהתבסס על צרכי הארגון, הוזלה של עלויות לארגונים באמצעות צמצום של הוצאות על חומרה ותחזוקה של השרתים, והנוחות של גישה לנתונים מכל מקום עם חיבור לאינטרנט. עם זאת, לצד הנוחות הללו, העברת נתונים זו מביאה איתה גם סיכונים פוטנציאליים, כגון גישה לא מורשית, הפרות נתונים, חשיפת מידע רגיש והפרות פרטיות .

דרכים להגן על הנתונים בענן

1. הצפנת נתונים: אחד ההיבטים הבסיסיים של אבטחת נתונים באחסון ענן הוא ההצפנה. ההצפנה מבטיחה שהנתונים יועברו לפורמט מקודד במהלך השידור (העברה של הקבצים אל הענן או הורדה ממנו) ובזמן 'מנוחה' של הקבצים על שרתי הענן. תהליך זה כולל אלגוריתמים מורכבים שהופכים את הנתונים לבלתי קריאים ללא מפתח הפענוח המתאים. על ידי שימוש בהצפנה, ספקי ענן מסכלים הפרה פוטנציאלית או ניסיונות גישה לא מורשית, ומבטיחים שגם אם נתונים יירטו, הם יישארו בלתי ניתנים לפענוח וחסרי תועלת לגורמים לא מורשים. (Calvin Chong Kun Lee & Gouher Ahmed, 2021)
2. בקרות גישה: יישום בקרות גישה חזקות הוא חיוני. לדוג' אימות רב-שלבי (MFA) (אימות אלקטרוני אשר משתמש בשתי דרכים או יותר על מנת לאפשר גישה לחשבון/מערכת, ותפקידו להגן מפני גישה של גורמים בלתי רצויים גם במידה ואחד או יותר מדרכי האימות כשלו) מוסיף שכבת אבטחה נוספת על ידי דרישה מהמשתמשים לספק צורות אימות מרובות לפני גישה לנתונים שלהם. בנוסף ישנן גם בקרות גישה מבוססות תפקידים (לטובת מידור הנתונים) שמאפשרות למנהלי מערכת להגדיר הרשאות מדויקות עבור משתמשים שונים, ומגבילה את הגישה רק למה שנדרש.
3. עמידה ברגולציה: ספקי שירותי ענן מחוייבים לציית לתקנות ולעמוד בתקנים ספציפיים, תוך הבטחת הגנת נתונים והפרטיות. כגון תקנות הגנת המידע הכללית -GDPR) אוסף של הוראות מחייבות שהוסדרו על ידי הפרלמנט האירופי, מועצת האיחוד האירופי והנציבות האירופית על מנת להגן על נושאי המידע, הנקראים) "Data Subjects"ויקיפדיה ((וחוק פרטיות הצרכן של קליפורניה ,(CCPA) מחייבות הנחיות קפדניות לאופן הטיפול וההגנה על נתוני המשתמש. תקנות אלו מטילות עונשים כבדים על ארגונים שאינם עומדים בתקני פרטיות הנתונים, מה שמדגיש את חומרת השמירה על הסודיות וגורם מצד אחד לשקיפות ואחריות מצד ספקי הענן ומד שני מספק למשתמשים את הביטחון שהנתונים שלהם מטופלים בזהירות. (PAN YANG, NAIXUE XIONG, & JINGLI REN, 2020)
4. חיסיון נתונים: כל ספקי שירותי הענן (המוכרים) מחויבים בהתחייבות משפטית לשמור על סודיות הפרטים והנתונים המועלים לענן. נאסר עליהם להשתמש בנתוני המשתמשים למטרות רווח אישי, למטרות שיווקיות או כל פעילות בלתי מורשית אחרת. מחויבות זו מודגשת על ידי מדיניות פרטיות מחמירה ותקנות הגנה על נתונים המחייבות ספקי ענן לכל שימוש לרעה בנתוני משתמשים.
5. השכלת משתמשים: העצמת המשתמשים עם ידע על שיטות עבודה מומלצות לאבטחה היא חיונית. חינוך משתמשים לגבי היגיינת סיסמאות, זיהוי ניסיונות דיוג והימנעות משיתוף מידע רגיש משפר את עמדת האבטחה הכוללת.
6. ביקורות וניטור: ביקורות אבטחה, הערכת פגיעויות וניטור בזמן אמת חיוניים כדי לזהות איומים פוטנציאליים ולטפל בהם באופן מידי. הניטור עוזר לזהות פגיעויות לפני שהן מנוצלות על ידי גורמים זדוניים.

בעולם שבו פרצות נתונים נראות כמעט בלתי נמנעות, חשוב מאוד הן לספקי שירותי הענן והן למשתמשים לשתף פעולה בהגנה על נתוני משתמשים פרטיים. הטמעת גישה רב-שכבתית לאבטחה, שילוב של הצפנה, בקרות גישה, עמידה בדרישות רגולציה וניטור מתמשך, הם המפתח להפחתת סיכונים ולהבטחת סודיות ושלמות המידע הרגיש. מכיוון שאחסון בענן ממשיך לעצב את הנוף הדיגיטלי, זוהי אחריות משותפת לתעדף את אבטחת הנתונים בראש ולהגן על הדבר החשוב ביותר: הנתונים הפרטיים שלנו.

השוואת טכניקות להתממה

שם הטכניקה	מה הטכניקה עושה?	יתרונות	חסרונות
מיסוך נתונים – Data Masking	הטכניקה משמש להחלפת התווים המורכבת מכיסוי\מיסוך תווים של ערך של נתונים ע"י החלפת התווים האלה בסמל מוגדר מראש. (JF , 2020 Marques, J Bernardino)	1.מיסוך נתונים יכול להגן על נתונים רגישים מפני גישה או חשיפה בלתי מורשית. 2.מיסוך נתונים יכול לשמר את השלמות והעקביות של הנתונים המקוריים. 3.ניתן להתאים אישית ולהתאים את מיסוך הנתונים לצרכים ספציפיים. (Dilmegani , 2022)	1.הטכניקה לא תמיד מספקת הגנה מלאה על נתונים רגישים. 2.מיסוך נתונים עשוי שלא תמיד להיות תואם הדדי עם מסדי נתונים\מערכות אחרות. (Dilmegani , 2022)
נתונים סינתטיים – Synthetic Data	טכניקה שבו מידע מיוצר באופן אלגוריתמי שאין לו קשר לאירועים אמיתיים. (Slavin,2022)	1. ניתן לשתף נתונים סינתטיים ולהשתמש בהם למטרות מחקר, ועוד, מבלי להפר חוקי פרטיות אתיים כלשהם; למשל, להכשרת מודלים של למידה עמוקה. 2. הטכניקה משפרת את האיכות וההכללה של מודלים ששייכים ל(Data Driven Models). 3. ניתן להפיק נתונים סינתטיים לפי דרישה ובקנה מידה. 4. נתונים סינתטיים ניתנים להתאמה אישית והתאמה לצרכים ספציפיים. (Dilmegani , 2022)	1. איכות הנתונים הסינתטיים תלויה בדיוק ובחוסן של האלגוריתמים ומאגר הנתונים הבסיסיים. 2. נתונים סינתטיים עשויים שלא תמיד לשמר את המאפיינים הייחודיים של הנתונים המקוריים. 3. הנתונים לא תמיד יכולים להיות אמינים. 4.נתונים סינתטיים לא תמיד להיות תואמים לפעולה הדדית עם מאגר נתונים. (Dilmegani , 2022)
הדחקה – Suppression	הדחקה הוא תהליך של הסרת פיצ'ר מסוים לחלוטין ממאגר הנתונים. (JF , 2020 Marques, J Bernardino)	היתרון העיקרי הוא שכאשר מוחקים\מסירים תכונה\פיצ'ר הלא רלוונטי , זה הופך להיות בלתי אפשרי לאחזר את המידע. (JF Marques, J , 2020 Bernardino)	הטכניקה לבדה לא מבטיחה התממה, משום שיש סיכוי להוביל לסיכון מוגבר לזיהוי מחדש בשילוב עם מקורות נתונים חיצוניים. (El Emam et al,2012)
הפרעת נתונים - Noise Addition	שינוי את מאגר הנתונים הראשוני באופן שולי ע"י יישום שיטות מספור עגול והוספת רעש אקראי. (Team , C , 2023)	הטכניקה משנה מעט את התכונות של מאגר הנתונים הראשוני לפחות מדויקות. (JF Marques, J , 2020 Bernardino)	במידה ויש בחירה לא נכונה לבסיס אם הבסיס הוא קטן, לכן בסיס קטן יכול לתרום להתממה לקויה\גרועה, בעוד שבסיס רחב יכול להפחית את השמשיות של מערך הנתונים . (Team , C , 2023)

החלפת נתונים – Data Swapping/Shuffling	סידור מחדש את ערכי מאפייני הנתונים כך שהם לא מתאימים למידע המקורי כולל החלפת תכונות (עמודות) הכוללות ערכים\תצפיות מוכרים. (Slavin , 2022)	הטכניקה מקשה על תוקפים לבצע ביטול אנונימיות בגלל חוסר ההתאמה למידע המקורי.(Slavin , 2022).	טכניקה לא תמיד מספקת התממה של הנתונים וייתכן שניתן יהיה לארגן אותם מחדש לצורתם המקורית. לכן, יש להשתמש בו בשילוב עם טכניקות אחרות. (JF Marques, J , 2020 Bernardino).
הכללה - Generalization	אי הכללה מכוונת של חלקים מסוימים של הנתונים. (Team , C , 2023)	אם הנתונים(נתונים מזוהים שמאפיין בן אדם כלשהו) ישונו לקבוצה של טווחים או אזורים גדולים בגבולות המתאימים אז הנתונים הופכים ללא מזוהים ופחות ספציפיים, כך מונעים את זיהוי אדם ספציפי. (JF , 2020 Marques, J Bernardino).	1.ייתכן שהטכניקה לא תגרום לאנונימיזציה יעילה. 2. טכניקה זו רגישה להתקפות של הסקת מסקנות הסתברותיות. 3.סכנת זיהוי מחדש מבחינת K-Anonymity : ההסתברות לזהות אדם שווה או קטנה מ-K/1. לכן, ככל שה-K גבוה יותר, ההסתברות לזיהוי נמוכה יותר , וככל שה-K קטן יותר ההסתברות לזיהוי גדולה. (JF Marques, J , 2020 Bernardino).
פסבדונימיזציה - Pseudonmization	תהליך של הסרת מזהים ממערך נתונים והחלפתם בשם בדוי. (Dvorin , 2023)	1.שמירת על שלמות הנתונים, הדיוק והדיוק הסטטיסטי ובו בזמן שומרת על סודיות , בנוסף הטכניקה בנתונים שהשתנו לצרכים למשל : בדיקות , ניתוחים , פיתוח , יצירה ,תוך שמירה על פרטיות הנתונים. (Slavin , 2022). 2.האלגוריתם ממפה\מפענח את האדם במסד נתונים\מאגר לפי שם בדוי ושם בדוי מסוים הוא שייך לאדם מסוים כל שם בדוי שונה אחד משני . (Dvorin , 2023)	1. הסיכון לזיהוי מחדש של נתונים אנונימיים תמיד קיים. תוקפים נחושים, המשלבים נתונים בדויים עם מידע זמין אחר (כגון מפתחות הצפנה ועוד), יכולים לזהות נתונים מקוריים. 2.ייתכן טכניקה תגרום לאובדן איכות הנתונים, מה שמקשה על ארגונים לבצע ניתוח מדויק. 3.עבור ארגונים מסוימים, יישום פסבדונימיזציה דורש מומחיות ומשאבים נוספים. העלות והמורכבות של יצירת מידע בדוי עולים ככל שגודל מערכי הנתונים גדל. (Richman , 2023)

יתרונות וחסרונות של התממת נתונים

יתרונות

1. **הגנה על פרטיות**:

אנונימיזציה משמשת בעיקר כדי להגן על פרטיותם של אנשים (לדוג'- לקוחות החברה) על ידי הסרה או שינוי של מידע מזהה. ע"י שימוש באחת הטכניקות ל- אנונימיזציה ניתן למנוע גישה לא מורשית לנתונים אישיים ולהפחית את הסיכון לגניבת זהות, מעקב והפרות פרטיות אחרות. (Karatas, 2023)

2. **ציות לרגולציה**:

קיימות תקנות רבות להגנת מידע, כגון GDPR (General Data Protection Regulation) ו-HIPAA (סדרת תקנות המיועדות לשימוש בכל הקשור ל-נתונים דיגיטליים רפואיים), אשר דורשות מארגונים להגן על פרטיות הנתונים האישיים של הלקוחות שלהם. אנונימיזציה של נתונים מאפשרת לארגונים לעמוד בדרישות הרגולטוריות. (Richman, 2023)

3. **מחקר וניתוח**:

ע"י ביצוע התממה לנתונים, חוקרים יכולים להשתמש בנתונים האנונימיים כדי לבצע ניתוחים שונים, כגון לימוד מגמות ודפוסים. כיום, בעולם שמונע מנתונים, מחקר וניתוח ממלאים תפקיד מרכזי בהבנת תופעות מורכבות, קבלת החלטות מושכלות והנעת התקדמות בתחומים שונים. אנונימיזציה, תהליך של הסתרת מידע אישי מזהה (PII- מידע אישי רגיש או מידע המאפשר זיהוי אישי) ממאגרי נתונים, התגלה ככלי קריטי המאפשר לחוקרים למנף נתונים רגישים תוך כיבוד הפרטיות. בתחומים כמו בריאות, מדעי החברה, כלכלה וכיו"ב שבהם הגישה לנתונים מהעולם האמיתי חיונית לביצוע צעדים משמעותיים ולקבלת החלטות מושכלות. (Richman, 2023)

4. **כריית נתונים**:

כריית נתונים (תהליך של חילוץ דפוסים, תובנות וידע ממערכי נתונים גדולים) הפך לכלי הכרחי עבור עסקים מודרניים השואפים להשיג יתרון תחרותי. אנונימיזציה משחקת תפקיד מכריע במתן אפשרות לארגונים לרתום את הכוח של כריית נתונים תוך שמירה על פרטיות הלקוחות ועמידה בתקנות הגנת מידע.

לדוג' שיפור חוויית לקוח - אנונימיזציה מאפשרת לחברות לנתח דפוסים בהתנהגות המשתמשים, להבין את העדפותיהם ולהתאים מוצרים, שירותים ומסעות שיווק בהתאם מבלי לחשוף פרטים אישיים של לקוחות בודדים.

דוג' נוספת- הבנת מגמות בשוק- באמצעות אנונימיזציה למאגרי נתונים המכילים מידע על דמוגרפיה של לקוחות, הרגלי רכישה ומיקומים גיאוגרפיים, ארגונים יכולים לקבל תובנות לגבי מגמות שוק רחבות יותר. דבר המאפשר לעסקים לצפות שינויים בביקוש, לזהות פלחי שוק מתעוררים ולהתאים את האסטרטגיות שלהם להעדפות הצרכנים המשתנות. כמו כן, גם נתונים אנונימיים מפלטפורמות מדיה חברתיות או פורומים מקוונים יכולים לספק ערך רב ולעזור לעסקים לאמוד את דעת הקהל לגבי מוצרים ושירותים.

5. **הפחתת סיכונים**:

פרצות נתונים ומתקפות סייבר הפכו לדאגות משמעותיות עבור ארגונים בתעשיות שונות. אנונימיזציה מופיעה כאסטרטגיה קריטית למזער את ההשפעה של הפרות פוטנציאליות תוך שמירה על השימושיות של מערכי נתונים יקרי ערך, מהסיבות הבאות: (Richman, 2023)

a. הקלה על ההשלכות של הפרת נתונים:

אנונימיזציה מפחיתה את האטרקטיביות של נתונים גנובים לפושעי סייבר על ידי הסרת מידע אישי מזהה (PII). במקרה של הפרה, גם אם גורמים לא מורשים מקבלים גישה למערך הנתונים האנונימי, היעדר קישורים ישירים לאנשים פרטיים מאתגר לשייך את הנתונים לאנשים ספציפיים. דבר אשר מגביל את הנזק הפוטנציאלי שיכול להיגרם כתוצאה מפרצת נתונים, הגנה על אנשים מפני גניבת זהות, הונאה והפרות פרטיות אחרות.

b. שמירה על שלמות מוניטין:

פרצות נתונים לא רק גורמות להפסדים כספיים אלא גם עלולה לפגוע במוניטין של הארגון ולשחק באמון הלקוחות. על ידי שימוש בטכניקות אנונימיזציה, חברות מפגינות את מחויבותן לאבטחת מידע ופרטיות, תוך טיפוח תחושת אמון בקרב לקוחות ומחזיקי עניין.

חסרונות ואתגרים

1. **הסיכון עדיין קיים**:
אנונימיזציה אינה חסינה מתקלות ותמיד קיים סיכון שתוקפים נחושים יוכלו לזהות פרטים אישיים על אנשים ע"י פריצה להצפנה או מיסוך הנתונים או על ידי מידע זמין אחר, כגון פוסטים במדיה חברתית או מידע אחר שזמין לציבור. (Richman, 2023)
2. **אובדן תועלת וניתוח לא מדויק**:
אחד האתגרים המרכזיים של אנונימיזציה הוא למצוא את האיזון העדין בין שמירה על איכות הנתונים לבין הבטחת הפרטיות. טכניקות אנונימיזציה אגרסיביות עשויות להגן על הפרטיות ביעילות אך להפוך את הנתונים כמעט חסרי תועלת לצורך ניתוח מידע ועבדה פנימית בארגון, לפעמים אנונימיזציה עלולה להוביל לתוצאות לא מדויקות או מוטות בברית נתונים. זה נכון במיוחד כאשר מדובר על מידע דמוגרפי או התנהגותי המשפיע על מהימנות או תוצאות הממצאים. יצירת האיזון כרוכה בהתייחסות לגורמים כמו פירוט הנתונים, הוספת רעש לנתונים ומקרי השימוש הספציפיים עבור הנתונים האנונימיים. (Karatas, 2023) (Richman, 2023)
3. **מורכבות**:
אנונימיזציה אינה פתרון חד-פעמי שמתאים לכולם, מדובר תחום מתפתח עם טכניקות מתפתחות כל הזמן. ככל שמתפתחות שיטות זיהוי חדשות, חייבים להתאים את גישות האנונימיזציה כדי להקדים את האיומים הפוטנציאליים. אנונימיזציה נכונה של נתונים דורשת הבנה עמוקה של מבני נתונים, קשרים וסיכונים פוטנציאליים לזיהוי מחדש. הטמעת טכניקות אנונימיזציה יעילות עשויה להיות מורכבת וגוזלת זמן ולהישאר מעודכן כל הזמן. (Karatas, 2023)
4. **מומחיות וכישורים**:
יישום אסטרטגיות אנונימיזציה אפקטיביות דורש כוח עבודה מיומן המצויד במומחיות בפרטיות נתונים, הצפנה, סטטיסטיקה וניהול נתונים. ארגונים זקוקים לאנשי מקצוע שיכולים לנווט במורכבות של אנונימיזציה של נתונים, להעריך נקודות תורפה אפשריות ולבחור טכניקות מתאימות בהתבסס על ההקשר הספציפי.
5. **עלויות**:
הטמעת תהליכי אנונימיזציה איכותיים והבטחת עמידה מתמשכת בתקנות המשתנות עשויה להיות ייקרה מאוד לארגונים. זה כולל השקעות בטכנולוגיה, כוח אדם והדרכה. ארגונים זקוקים לצוות שבקיא בתקנות פרטיות, בשיטות עבודה מומלצות לטיפול בנתונים ובטכניקות אנונימיזציה. גיוס והכשרת אנשי מקצוע מיומנים בתחום זה יכולים לעלות ביוקר. בנוסף, יש לקחת בחשבון את העלות שצוותים קיימים יזדקקו להכשרה מתמשכת ולמידה חדשה כדי לעמוד בקצב המתפתח של פרטיות נתונים. (Richman, 2023)

סיכום ,דיון, המלצות ומסקנות להמשך

עבודתנו מתעמקת בתחום הקריטי של טכניקות הגנת הפרטיות והאנונמיזציה, במיוחד בהקשר של שיתוף נתונים אישיים על פני מסדי נתונים מבוזרים ופרטיים. המוקד הוא הסיכונים הקרובים הקשורים לגישה לא מורשית, שעלולים להוביל לחשיפה של מידע אישי רגיש. פגיעות זו פותחת דלת לבעיות רבות כולל גניבת זהות, פעילויות הונאה וחשיפת נתונים סודיים בשוגג. מתוך הכרה בצורך הגובר בשיתוף נתונים בעולם המקושר של ימינו, העבודה מדגיש בו-זמנית את החשיבות של התייחסות לדאגות הפרטיות הללו.

לאחר ניתוח יסודי, הטכניקה האופטימלית להשגת הגנת פרטיות חזקה ולאפשר שיתוף נתונים מאובטח היא מיסוך נתונים . גישה זו מבטיחה גישה מבוקרת למידע רגיש, ומאפשרת למשתמשים שונים רמות נראות שונות על סמך רמות ההרשאה. על ידי אספקת הנתונים הדרושים למשתמשים ספציפיים בלבד, מיסוך נתונים מייצר איזון בין שמירה על כלי השירות למידע ושמירה על פרטיות הפרט.

עבודה זו האירה את הסכנות של גישה לא מורשית לנתונים, ושפכה אור על ההשלכות האפשריות - החל מגניבת זהות ועד לחשיפת מידע בשוגג. ככל שהצורך בשיתוף נתונים אישיים על פני מסדי נתונים פרטיים ומבוזרים הולך וגובר, כך עולה הדחיפות בתכנון אסטרטגיות יעילות להפחתת סיכונים אלו.

בין שלל טכניקות אנונימיזציה (מיסוך נתונים , הדחקה ועוד...) הזמינות שסקרנו, לדעתנו מיסוך נתונים-Data Masking מתגלה כבחירה האופטימלית ומועדפת לשימור פרטיות חזק תוך הקלה על שיתוף נתונים מאובטח. יתרונות הטכניקה שהיא יכולה להגן על נתונים רגישים מפני גישה או חשיפה בלתי מורשית, מיסוך נתונים יכול לשמר את השלמות והעקביות של הנתונים המקוריים, מיסוך נתונים יוצר איזון חיוני. הטכניקה זו מבטיחה שנתונים של המשתמשים חשופים ברמות שונות של נראות לתוך הנתונים, בהתאם להתאמה אישית או לצרכים ספציפיים. גישה זו לא רק שומרת על הפרטיות, אלא גם שומרת על שירות הנתונים, מה שהופך אותה לפתרון משכנע עבור תרחישי שיתוף נתונים מודעים לפרטיות.

עם זאת למיסוך נתונים יש חסרונות אך לא מדובר בחסרונות קריטיים שנמצאות בטכניקות אחרות , החסרונות הם : 1. הטכניקה לא תמיד מספקת הגנה מלאה על נתונים רגישים ,אך יכול להיות שהטכניקה יכולה לספק הגנה על נתונים בצורה כמעט מלאה או קרוב למלאה.2. מיסוך נתונים עשוי שלא תמיד להיות תואם הדדי עם מסדי נתונים.

לסיכום לפי מסקנתנו שמיסוך נתונים היא הטכניקה להגנת פרטיות חזקה ואפשרות שיתוף נתונים מאובטח המועדפת ביותר, שנכון החסרונות לא קריטיים מול חסרונות של טכניקות אחרות שסקרנו שכן הן יש להן חסרונות קריטיות(למשל : סיכוי לזיהוי מחדש , התממה לקויה ועוד) .

בעולם שבו נתונים מעודדים חדשנות וקידמה, השמירה על פרטיות הפרט נותרה בעלת חשיבות עליונה. על ידי קידום מתמשך של אנונימזציה בשלל הטכניקות ויישומן המעשי, חוקרים ואנשי מקצוע יכולים לפלס דרך הרמונית, כזו שבה תובנות מונעות נתונים מתקיימות במקביל להגנת הפרטיות, מה שיוצר עתיד המכבד את הפרטיות האישית מבלי לפגוע בצווי הניתוח וההתקדמות.

ביבליוגרפיה

- Calvin Chong Kun Lee & ,Gouher Ahmed .(2021) .Improving Internet Privacy, Data Protection and Security Concerns .*International Journal of Technology, Innovation and Management*.
- Jurczyk, P' X .(2009) .'Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers.
- Jurczyk, P' X .(2009) .'*Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers* אוחזר מתוך .https://link.springer.com/chapter/10.1007/978-3-642-03007-9_13#citeas
- Karatas, G .(2023 8) .'*Data anonymization: Pros, Cons & Techniques in 2023* אוחזר מתוך . research.aimultiple: https://research.aimultiple.com/data-anonymization/
- PAN YANG, NAIXUE XIONG, & JINGLI REN. (2020). Data Security and Privacy Protection. *National Natural Science Foundation*.
- Richman, A .(2023 4) .'*The Advantages and Disadvantages of Pseudonymized Data* אוחזר מתוך . k2view: https://www.k2view.com/blog/pseudonymized-data/#Protecting-User-Privacy
- Wallace, S' E .(2016) .'*What Does Anonymization Mean? DataSHIELD and the Need for Consensus on Anonymization Terminology* אוחזר מתוך . https://www.liebertpub.com/doi/full/10.1089/bio.2015.0119
- Marques, J. F., & Bernardino, J. (2020). Analysis of Data Anonymization Techniques. *KEOD*, 235-241. https://www.scitepress.org/Papers/2020/101423/101423.pdf
- Li, M., Liu, Z., Jia, C., & Dong, Z. (2013, September). Data masking generic Model. In *2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies* (pp. 724-727). IEEE. https://ieeexplore.ieee.org/abstract/document/6631710
- Goyal, C. (2015). Data masking: need, techniques & solutions. *Int. Res. J. Manag. Sci. Technol.(IRJMST)*, *6*(5), 221-229. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Data+masking%3A+need%2C+techniques+%26+solutions+&btnG=
- El Emam, K., Arbuckle, L., Koru, G., Eze, B., Gaudette, L., Neri, E., ... & Gluck, J. (2012). De-identification methods for open health data: the case of the Heritage Health Prize claims dataset. *Journal of medical Internet research*, *14*(1), e33. https://www.jmir.org/2012/1/e33

- Dvorin, T. (2023, May 24). *Data Anonymization Techniques: Pros and Cons*. Duality Technologies. https://dualitytech.com/blog/data-anonymization-techniques-pros-and-cons/
- Team, C. (2023, May 31). *Data Anonymization*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/business-intelligence/data-anonymization/
- Slavin, B. (n.d.). *Data Anonymization: Overview, Techniques, Plus Pros And Cons - DuoCircle*. DuoCircle. https://www.duocircle.com/email-security/data-anonymization-overview-techniques-plus-pros-and-cons
- Dilmegani, C. (2022, December 22). *Synthetic Data vs Data Masking: Benefits & Challenges in 2023*. AIMultiple. https://research.aimultiple.com/synthetic-data-vs-data-masking/
- Richman, A. (2023, May 2). *The Advantages and Disadvantages of Pseudonymized Data*. https://www.k2view.com/blog/pseudonymized-data/#Challenges-and-Limitations