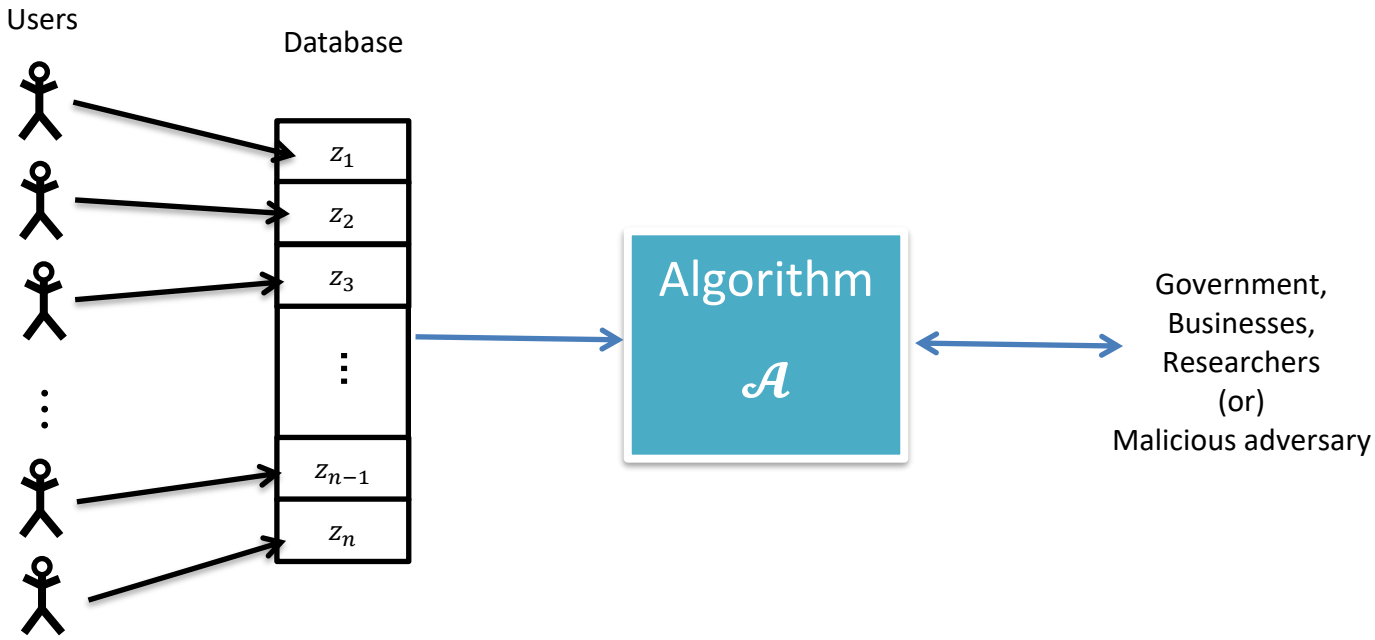
27.10.2022 | ***פרטיות מידע***

הרצאה 1: מבוא

Textbook: Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy

מרצה: אורי שטמר

<u>סיפור המסגרת בקורס שלנו:</u>



איך נוכל להבטיח פרטיות לאנשים שתרמו את המידע שלהם לדטהבייס?

מסתבר שזאת שאלה מאוד עדינה. היו הרבה נסיונות לפתור אותה בעבר, חלק גדול מהם הובילו לכשלונות של פרטיות...

רעיון כושל ראשון: אולי נמחוק מהדטהבייס מזהים ברורים כמו שמות + ת.ז. ?

הבעייה עם הרעיון הזה היא שגם שדות אחרים פחות ברורים עשויים להתברר בדיעבד כשדות מזהים.

Example: NYC Taxi and Limo Commission 2014

- זאת רשות שאחראית על רגולציה של מוניות ב NYC (מוניות צריכות אישור מהם)
- יש להם נתונים על כל הנסיעות של כל המוניות ב NYC
- בשנת 2014 הם פרסמו גרסה "אנונימית" של ה data שלהם (בסה"כ 19GB).
- כל שורה מייצגת נסיעה בודדת וכוללת: נק' התחלה, נק' סיום, זמן התחלה, זמן סיום, כמה אנשים, מרחק נסיעה. בנוסף, כל שורה הכילה hash (MD5) של מספר המונית.

- חוקר בשם Vijay Pandurangan לקח את כל מספרי המוניות ב NYC (זה מידע פומבי) וחישב את ערך ה hash שלהם. ככה הוא יכל לשחזר את מספרי המוניות בדטהבייס הזה! בעצם זה גם מדביק שמות לדטהבייס הזה, כי החיבור בין מספר מונית לשם של בעל המונית זה מידע פומבי ב NYC!
- זה אומר שהחוקר הזה גילה בדיוק כמה הרוויחו כל נהגי המוניות ב NYC!!!
- שאלה: האם הבעיה הייתה נפתרת אם במקום להשתמש ב MD5 הם היו מדביקים random ID לכל מונית? (זה היה מונע את המתקפה של Pandurangan)
- הבעיה לא נפתרת:
(1) כשאני עולה על מונית אני לומד את השם של הנהג, ואז יכול למצוא את הנסיעה בדטהבייס וככה לקשר בין השם של הנהג ל random ID שלו וככה אוכל ללמוד על כל הנסיעות שלו.
(2) אם אני רואה חבר שלי עולה על מונית, אני יכול למצוא את הנסיעה שלו ולבדוק שהוא באמת נסע לאן שהוא אמר שהוא נוסע...

Example: Group Insurance Commission (GIC) 2002

- GIC זה גוף ממשלתי במסצ'וסטס שאחראי על רכישת ביטוח בריאות לעובדים ממשלתיים
- הם פרסמו גרסה "אנונימית" של המידע הרפואי של העובדים הממשלתיים (ושל המשפחות שלהם). זה לא כלל שמות או ת.ז., אבל זה כן כלל: **מיקוד, תאריך לידה, מין,** מוצא, תאריכי ביקור, דיאגנוזות, ניתוחים, תרופות, תשלומים.
- בשנת 2002, חוקרת בשם Latanya Sweeney הצליבה את הדטהבייס הזה עם דטהבייס פומבי (voter registration list) שמכיל את השדות: **מיקוד, תאריך לידה, מין,** שם, כתובת, תאריך רישום, מפלגה, תאריך הצבעה אחרונה.
- היא גילתה שהשילוב הזה של מיקוד+תאריך לידה+מין מזהה בצורה ייחודית כמעט את כל תושבי ארה"ב ולכן התהליך הזה איפשר לה להדביק שמות וכתובות למידע הרפואי של ה GIC.

Example: Netflix Prize 2006-2009

- תחרות לשיפור מערכת ההמלצה של נטפליקס
- ב 2007 וב 2008 הם העניקו פרסים של 50,000$ עבור התקדמויות
- בשנת 2009 הם העניקו פרס של מיליון דולר לחוקרים שהצליחו לשפר את מערכת ההמלצה של נטפליקס ביותר מ 10%
- הדטה שנטפליקס נתנו לחוקרים היה אמור להיות "אנונימי" והכיל שורות מהצורה:
<user pseudo-ID, movie, date of rating, rate>
- בשנת 2008, שני חוקרים (Narayanan and Shmatikov) הצליבו את הדטה הזה עם הדטה של imdb שמכיל שורות מהצורה:
<real name, movie, date of rating, rate>
- זה הוביל לתביעה ייצוגית נגד נטפליקס וגרם להם לבטל את התחרות הבאה שהם תכננו

רעיון כושל שני: במקום לפרסם דטהבייס בלי שדות מזהים, אולי רק נענה על שאילתות?

דוגמה בעייתית: נניח שהדטהבייס מכיל מידע רפואי של אנשים, ונניח שאנחנו מאפשרים לשאול את השאילתא הבאה:

כמה אנשים בדטהבייס נולדו בתאריך 1/1/2000, גרים ברחוב חיים לבנון בתל אביב, והיה להם קורונה?

נניח ששאלנו את השאילתא הזאת וקיבלנו את התשובה 1. מה זה אומר לנו לגבי חבר שלנו שנולד בתאריך הזה וגר בחיים לבנון?

רעיון כושל שלישי: אולי נרשה רק שאילתות שהתשובה עליהן גדולה?

אנחנו עדיין באותה בעיה. נוכל לשאול את שתי השאילתות הבאות:

(1) לכמה אנשים שגרים בחיים לבנון ונולדו לפני 1/1/200 היה קורונה? נניח נקבל תשובה 317
(2) לכמה אנשים שגרים בחיים לבנון ונולדו לפני 2/1/200 היה קורונה? נניח נקבל תשובה 318

זה נקרא differencing attack. זה אמנם נראה טריוויאלי, אבל כדאי לזכור את זה.

דוגמה למתקפת שחזור נתונים

הטבלה הבאה מתארת סטטיסטיקות שה UC Census אולי היו רוצים לפרסם (זאת דוגמה דמיונית). בטבלה הזאת מחקו את כל הסטטיסטיקות שמבוססות על פחות מ 3 אנשים.

Table 1. Fictional statistical data for a fictional block.

		Age		
Statistic	**Group**	**Count**	**Median**	**Mean**
1A	Total Population	7	30	38
2A	Female	4	30	33.5
2B	Male	3	30	44
2C	Black or African American	4	51	48.5
2D	White	3	24	24
3A	Single Adults	(D)	(D)	(D)
3B	Married Adults	4	51	54
4A	Black or African American Female	3	36	36.7
4B	Black or African American Male	(D)	(D)	(D)
4C	White Male	(D)	(D)	(D)
4D	White Female	(D)	(D)	(D)
5A	Persons Under 5 Years	(D)	(D)	(D)
5B	Persons Under 18 Years	(D)	(D)	(D)
5C	Persons 64 Years or Over	(D)	(D)	(D)

Note: Married persons must be 15 or over

- אנחנו רואים שבבלוק הזה ישנם 3 גברים. ננסה לשחזר את הגילאים שלהם.
- נסמן את הגילאים שלהם כ-

$$A \le B \le C$$

- אנחנו יודעים שהחציון הוא 30 ולכן

$$\boldsymbol{A \le 30 \quad \& \quad B = 30 \quad \& \quad C \ge 30}$$

- אנחנו יודעים שהממוצע הוא 44 ולכן

$$44 = \frac{A+B+C}{3} = \frac{A+30+C}{3}$$

כלומר

$$\boldsymbol{A + C = 102}$$

- אז למדנו כל מיני אילוצים על הגילאים האלה (באדום). הנה טבלה עם כל האפשרויות שמקיימות את האילוצים שמצאנו:

Table 2. Possible ages for a median of 30 and a mean of 44.

A	B	C	A	B	C	A	B	C
1	30	101	11	30	91	21	30	81
2	30	100	12	30	90	22	30	80
3	30	99	13	30	89	23	30	79
4	30	98	14	30	88	24	30	78
5	30	97	15	30	87	25	30	77
6	30	96	16	30	86	26	30	76
7	30	95	17	30	85	27	30	75
8	30	94	18	30	84	28	30	74
9	30	93	19	30	83	29	30	73
10	30	92	20	30	82	30	30	72

- בעצם מה שעשינו כאן זה לצמצם את מרחב כל הגילאים האפשריים מבערך $100^3 = 1{,}000{,}000$ אפשרויות ל- 30 אפשרויות. ועשינו את זה בעזרת סטטיסטיקה אחת בלבד...

<u>תרגיל:</u> נניח שהיינו מפרסמים שישנם שני Black or African American Males וששממוצע הגילאים שלהם הוא 28. האם זה היה מספיק לנו כדי לשחזר במדוייק את הגילאים A,B,C, או שעדיין ישנם 2 או יותר אפשרויות?

בעצם מה שעשינו כאן זה יותר כללי. זה מתאר תבנית למתקפת שחזור נתונים:

- ישנו דטהבייס "אמיתי" X שאיננו ידוע לנו
- ישנם סטטיסטיקות $f_1, f_2, \dots, f_k$ שרוצים לחשב או להעריך על X (למשל ספירות או שאילתות)
- אנחנו מקבלים הערכות (אולי רועשות) לסטטיסטיקות האלה:

$$a_1 \approx f_1(X) \quad , \quad a_2 \approx f_2(X) \quad , \quad \dots \quad , \quad a_k \approx f_k(X)$$

- ההערכות האלה מגדירות לנו <u>אילוצים</u> שמצמצמים את מרחב האפשרויות עבור X

<u>בעיית שחזור נתונים:</u> בהינתן אילוצים $\{f_i(X) \approx a_i\}$, מצא דטהבייס $\tilde{X}$ שהוא עקבי עם האילוצים.

האינטואיציה כאן היא שאם נקבל מספיק אילוצים (מספיק ספציפיים/בדיוק מספיק גבוהה) אז זה יהיה אפשרי לשחזר את X או לקבל משהו קרוב אליו.

<u>הגדרה:</u> מערכת תקרא **<u>לא פרטית באופן בוטה</u>** אם יריב יכול לשחזר 99% מהדטהבייס.

<u>דוגמה להפרה בוטה של פרטיות:</u>

- נחשוב על מקרה שבו הדטהבייס מכיל ביט בודד עבור כל אחד מ- n אנשים. נסמן את הדטהבייס שלנו על ידי $X \in \{0,1\}^n$.
- נניח שאנחנו רוצים לענות על שאילתות מהצורה הבאה. שאילתא מוגדרת על ידי וקטור $q \in \{0,1\}^n$ והתשובה לשאיתלא q היא $A(q,X) = \sum_{i=1}^{n} q_i \cdot X_i$

הצעה למכניזם שעונה על שאילתות ומוסיף רעש במטרה לשמר "פרטיות":
בהינתן שאילתא q, חשב והחזר $\lfloor \text{רעש} \rceil + A(q,X)$.

האם זה רעיון טוב? זה תלוי כמה רעש מוסיפים ועל כמה שאילתות עונים...

<u>משפט 1:</u> אם הרעש האקראי שמוסיפים חסום על ידי $\frac{n}{401}$ אזי המע' הנ"ל מפרה פרטיות בצורה בוטה אם היריב מסוגל לשאול את כל 2^n השאילתות האפשריות.

<u>הוכחה:</u> היריב תוקף ב- 2 שלבים:
(1) לכל שאילתא אפשרית $q \in \{0,1\}^n$ שאל את השאילתא q וקבל תשובה (רועשת) a_q
(2) מצא והחזר דטהבייס $\hat{X} = (\hat{x}_1, \dots, \hat{x}_n)$ כך שלכל $q \in \{0,1\}^n$ מתקיים:

$$\left|a_q - A(q,\hat{X})\right| \leq \frac{n}{401}$$

כאן a_q היא התשובה הרועשת שקיבלנו מהמערכת כששאלנו את השאילתא q, ו- $A(q,\hat{X})$ הוא ערך השאילתא q על $\hat{X}$.

למה קיים כזה דטהבייס $\hat{X}$? כי בפרט הדטהבייס האמיתי X הוא כזה, לפי ההבטחה שהרעש חסום ע"י $\frac{n}{401}$.

<u>ניתוח:</u>
בשלב (1) היריב שאל (וקיבל תשובה רועשת עבור) כל שאילתא אפשרית. בפרט, היריב שאל את 2 השאילתות הבאות:

- $q^0 = (q_1^0, q_2^0, \dots, q_n^0) \in \{0,1\}^n$ כך ש- $q_i^0 = 1$ אם ורק אם $x_i = 0$
  (כאן x_i היא הכניסה ה-i-ית בטדהבייס האמיתי X)
- $q^1 \in \{0,1\}^n$ כך ש- $q_i^1 = 1$ אם ורק אם $x_i = 1$

לדוגמה, אם $X = (1,1,0,1,0)$ אזי $q^0 = (0,0,1,0,1)$ ו- $q^1 = (1,1,0,1,0)$

שאלת איפוס: מה הערך של $A(q^0,X)$ ומה הערך של $A(q^1,X)$?

כעת, לפי בחירת $\hat{X}$ מתקיים:

$$\left|a_{q^0} - A(q^0,\hat{X})\right| \leq \frac{n}{401}$$

בנוסף, לפי ההנחה שהרעש חסום:

$$\left|a_{q^0} - A(q^0, X)\right| \leq \frac{n}{401}$$

לכן, לפי אי-שוויון המשולש מתקיים:

$$\left| \underbrace{A(q^0, \hat{X})}_{\text{מזה?}} - \underbrace{A(q^0, X)}_{=0} \right| \leq \frac{2n}{401}$$

$\hat{X}$ מספר ה−1−ים ב במקומות שבהם אמורים להיות אפסים

כי $q_i^0=1$ אם ורק אם $x_i=0$

כלומר, מכל השורות שאמורות להיות 0, לכל היותר $\frac{2n}{401}$ מהן "התהפכו" ונהיו 1 בטעות. באופן דומה עבור q^1.
סה"כ $X, \hat{X}$ לא מסכימים על לכל היותר $\frac{4n}{401}$, כלומר פחות מ *1%* טעות.

מ.ש.ל.

החיסרון העיקרי של המתקפה שראינו עכשיו הוא שהיינו צריכים לשאול 2^n שאילתות, שזה לא סביר (למשל, אם יש לנו דטהבייס בגודל 1000 אז ברור שלא נוכל לשאול 2^{1000} שאילתות...). עדיין, יש למתקפה הזאת משמעות קונספטואלית חשובה:

"בשביל שיהיה לנו סיכוי למנוע מתקפת שחזור, אנחנו חייבים להגביל באיזושהי צורה את מספר הסטטיסטיקות שנפרסם מהדטהבייס"

או במילים אחרות

"יש גבול לכמות האינפורמציה שאנחנו יכולים לשחרר בצורה פרטית מהדטהביס"

סוגים שונים של כישלונות של פרטיות

- היריב מצליח לשחזר 99% מהדטהבייס
- היריב מצליח לשחזר שורה אחת מהדטהבייס
- היריב אולי לא לשחזר שורה שלמה, אבל הוא משחזר אותה חלקית. למשל, כל כניסה בדטהבייס זה המידע הרפואי של אדם אחד, והיריב משחזר חלק מהמידע הרפואי שלי.
- היריב לא מצליח ממש לשלוף שורה או חלק ממנה, אבל הוא כן לומד מידע על שורות בדטהבייס. אפשר לפרמל את זה בכל מני דרכים. אחת האפשרויות היא בעזרת הניסוי:
  - נניח שיש התפלגות מסוימת P שממנה נדגמות שורות בדטהבייס. למשל אולי P זאת ההתפלגות האחידה מעל אנשים במדינת ישראל.
  - נדגום דטהבייס $X = (x_1, x_2, \dots, x_n)$ המכיל n דגימות מתוך P
  - נריץ את M על X ונקבל פלט y
  - נגריל ביט $b \in \{0,1\}$
  - אם $b = 0$ אז נגריל עוד נקודה x_0 מההתפלגות P ואחרת נגריל x_0 בהתפלגות אחידה מתוך X
  - היריב מקבל את y, x_0 ופולט ניחוש $\hat{b}$. היריב מנצח אם $\hat{b} = b$.
  - שימו לב: לנצח בהסתברות ½ זו לא חוכמה... אבל נניח שהיריב מצליח לנצח בהסתברות 99%. האם זו פגיעה בפרטיות? מה אם הוא מצליח לנצח בהסתברות ¾ ?

עכשיו אנחנו רוצים להתחיל לדבר על תוצאות חיוביות ועל אלגוריתמים שכן מבטיחים פרטיות. אנחנו רוצים להיות מסוגלים לפרסם סטטיסטיקות לגבי הנתונים שלנו ולהיות מסוגלים להבטיח שזה לא יאפשר מתקפות כפי שראינו, או סוגים אחרים של מתקפות. איך אפשר להבטיח כזה דבר? מה אנחנו צריכים להוכיח בדיוק?

צעד אחורה: בשביל להיות מסוגלים להוכיח שמתקיימת "פרטיות", אנחנו קודם חייבים להגדיר בדיוק למה אנחנו מתכוונים כשאנחנו אומרים "פרטיות". (אחרת, אם לא נגדיר מתי פרטיות נשמרת, איך נוכל להוכיח שזה מתקיים?)

איך נגדיר מתי פרטיות נשמרת?

שתי גישות עיקריות:

גישה 1:	נזהה משפחה מסוימת של מתקפות שאנחנו רוצים להתגונן מפניהן, ונאמר שמכניזם "משמר פרטיות" אם הוא חסין בפני המתקפות האלה
גישה 2:	נזהה תכונה כללית ונוכיח שאלגוריתמים שמקיימים את התכונה הזאת הם באיזשהו מובן חסינים בפני כל מתקפה אפשרית, גם מתקפות עתידיות שאינן ידועות לנו כרגע. נאמר שמכניזם "משמר פרטיות" אם הוא מקיים את התכונה הזאת.

דוגמה לגישה 1: k-אנונימיות

- הקלט הוא טבלה בה כל שורה מייצגת את המידע של אדם אחד (השורה יכולה להכיל הרבה שדות).
- אנחנו מניחים שאנחנו יודעים מהם השדות שעשויים להופיע בדטהבייסים אחרים שלתוקף אולי יש גישה אליהם. אלו הם שדות שהתוקף עלול להשתמש בהם כדי לבצע linkage attack (כמו למשל המתקפה עם ה GIC משיעור שעבר). לשדות האלה קוראים "שדות מזהים".
- הפלט הוא טבלה מאותו גודל, בה חלק מהערכים "הוכללו", כלומר הוחלפו "בקבוצות" של ערכים אפשריים במקום ערכים ספציפיים. לדוגמה:
  - 2 ← [1,3]
  - גבר ← {גבר, אישה}
  - אברהם ← אב*
- טבלה נקראת "k-אנונימית" אם לכל שורה בטבלה יש לפחות k שורות בטבלה עם בדיוק אותם ערכים בשדות המזהים.

Identifying attributes			
Age	***Zip***	***Nationality***	***Condition***
21	*13092*	*American*	*AIDS*
23	*13083*	*German*	*Heart*
28	*13029*	*Italian*	*Viral*
29	*13079*	*Israeli*	*Viral*
30	*13032*	*American*	*Cancer*
33	*13002*	*Israeli*	*Cancer*
34	*13021*	*Russian*	*Cancer*
38	*13094*	*Lebanese*	*Cancer*
41	*13089*	*German*	*Cancer*
47	*13044*	*American*	*Heart*
48	*13041*	*Egyptian*	*Viral*
49	*13058*	*American*	*Viral*

⟹

Identifying attributes			
Age	***Zip***	***Nationality***	***Condition***
<30	*130***	*	*AIDS*
<30	*130***	*	*Heart*
<30	*130***	*	*Viral*
<30	*130***	*	*Viral*
*3**	*130***	*	*Cancer*
*3**	*130***	*	*Cancer*
*3**	*130***	*	*Cancer*
*3**	*130***	*	*Cancer*
≥40	*130***	*	*Cancer*
≥40	*130***	*	*Heart*
≥40	*130***	*	*Viral*
≥40	*130***	*	*Viral*

מה טוב ב *K*-אנונימיות?
נראה קשה לבצע מתקפות *linkage* אם כל מה שיש לנו זה מידע לגבי "השדות המזהים"

מה יכול להשתבש?
זה לא מבטיח חסינות בפני מתקפות אחרות. למשל, נניח שאנחנו רואים את הדטהבייס הימני ("האנונימי") ונניח שאנחנו יודעים שחבר שלנו בגיל 30 ביקר בבית החולים. ← גילינו שיש לו סרטן

בעייה נוספת – קומפוזיציה: נניח שאנחנו רואים 2 טבלאות, כל אחת מהן *K*-אנונימית:

Identifying attributes			
Age	***Zip***	***Nationality***	***Condition***
<35	*130***	*	*AIDS*
<35	*130***	*	*Tuberculosis*
<35	*130***	*	*Flu*
<35	*130***	*	*Tuberculosis*
<35	*130***	*	*Cancer*
<35	*130***	*	*Cancer*
≥35	*130***	*	*Cancer*
≥35	*130***	*	*Cancer*
≥35	*130***	*	*Cancer*
≥35	*130***	*	*Tuberculosis*
≥35	*130***	*	*Viral*
≥35	*130***	*	*Viral*

Identifying attributes			
Age	***Zip***	***Nationality***	***Condition***
<30	*130***	*	*AIDS*
<30	*130***	*	*Heart*
<30	*130***	*	*Viral*
<30	*130***	*	*Viral*
*3**	*130***	*	*Cancer*
*3**	*130***	*	*Cancer*
*3**	*130***	*	*Cancer*
*3**	*130***	*	*Cancer*
≥40	*130***	*	*Cancer*
≥40	*130***	*	*Heart*
≥40	*130***	*	*Viral*
≥40	*130***	*	*Viral*

נניח שאנחנו יודעים שחבר שלנו בן 28 ביקר בשני בתי החולים האלה. ← גילינו שיש לו איידס

בעיה נוספת: ההגדרה של *K*-אנונימיות מפרטת תנאים על הפלט, אבל לא מגבילה את תהליך החישוב (האלגוריתם) שמייצרת את הפלט הזה. זה עשויי להוביל לבעיות נוספות אם נדע פרטים נוספים על האלגוריתם עצמו.

- למשל, אם אנחנו יודעים שהאלגוריתם מבצע "הכללה מינימלית אפשרית", אז מהדטהבייס האנונימי הראשון שראינו אנחנו לומדים שהיה מישהו בדטהבייס בגיל 30 והיה שם מישהו בגיל 39.

<u>מסקנה:</u> אולי בהמשך נרצה להגביל גם את תהליך החישוב עצמו ולא רק את הפלט שלו...

נסיון כושל לפרמל את גישה 2

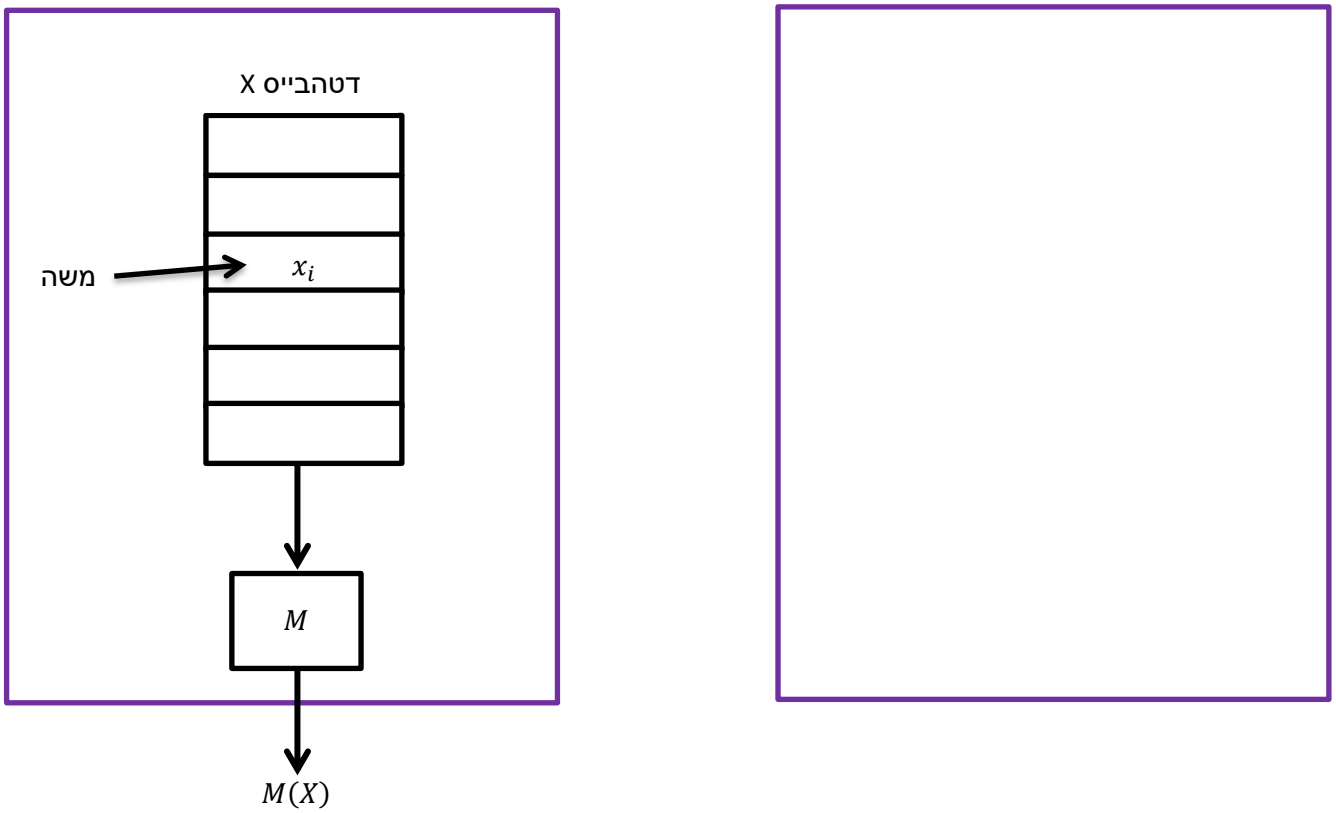
בשנת 1977 מתמטיקאי בשם תור דלניוס הציע את ההגדרה הבאה:

<u>הצעה *(Dalenius 77)*:</u> "הפרטיות שלי נשמרת אם אי אפשר ללמוד עלי כלום"

זה אמנם לא ממש פורמלי, אבל זה נשמע טוב נכון?

אבל ההגדרה הזאת חזקה מדי ולא נוכל לעשות איתה שום דבר. במילים אחרות, אם יש איזושהי תועלת שאפשר להסיק מהדטהבייס, אז אי אפשר לעמוד בהגדרה הזאת...

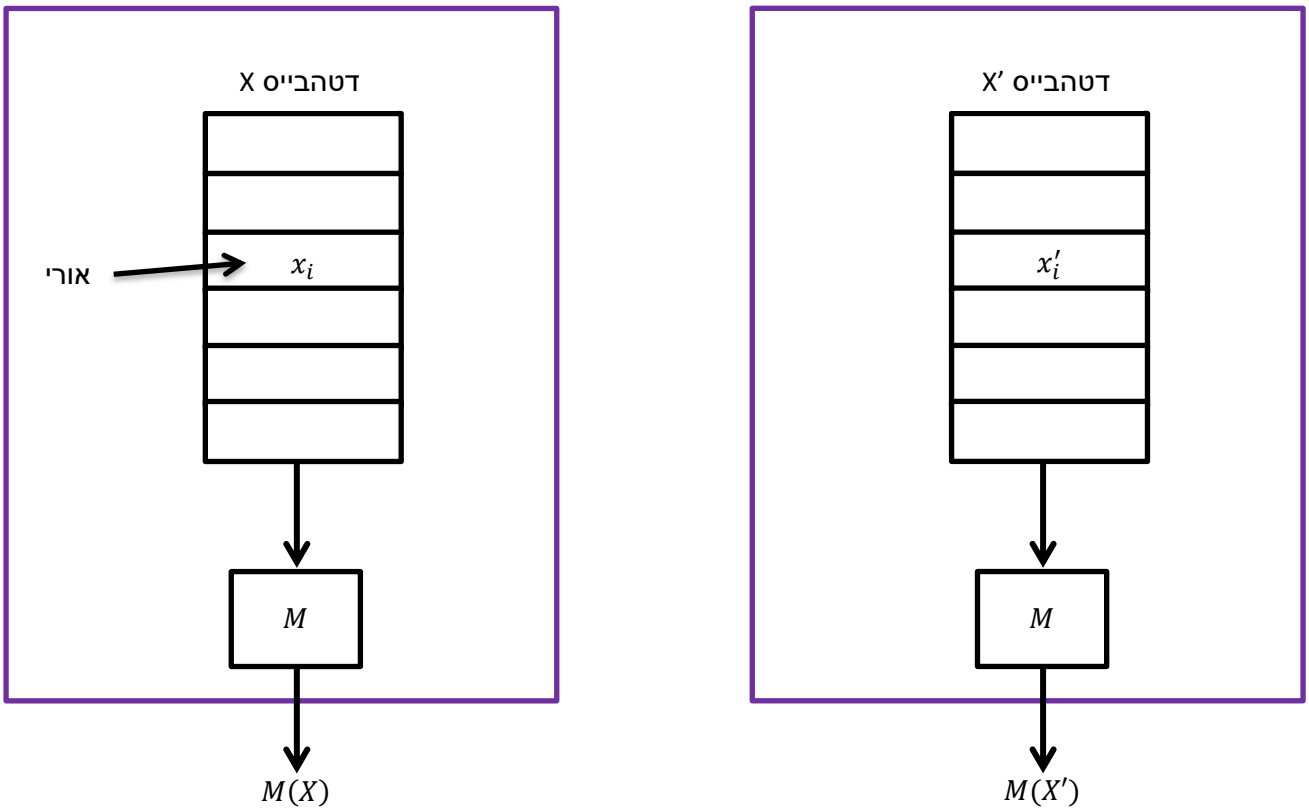
איפה הבעיה בהגדרה הזאת? נקודת הייחוס שלה לא טובה. בציור, דלניוס השווה את 2 המצבים הבאים:



כלומר, ההגדרה של דלניוס אמרה שאי אפשר ללמוד על משה שום דבר שלא ידענו עליו גם קודם, כלומר שום דבר שלא היינו יכולים ללמוד עליו גם מהתמונה הימנית. אנחנו רוצים למצוא נקודת ייחס אחרת להשוות אליה.

פרטיות דפרנציאלית

נסתכל על דטהבייס *X′* זהה ל- *X* פרט לכך שהמידע שלי הוחלף במשהו אחר. (כל שאר השורות זהות)



אם היריב לא יכול להבדיל בין *M(X)* לבין *M(X′)* אזי נאמר שפרטיות נשמרת.

<u>הגדרה:</u> מכניזם $M: D^n \to R$ מקיים (ε, δ)-פרטיות דיפרנציאלית אם לכל $X, X' \in D^n$ המקיימים $x_i \neq x_i'$ בדיוק עבור כניסה i אחת (דטהבייסים כאלה יקראו <u>שכנים</u>), ולכל תת קבוצה $S \subseteq R$ מתקיים:

$$\Pr[M(X) \in S] \leq e^{\varepsilon} \cdot \Pr[M(X') \in S] + \delta$$

כאשר ההסתברות היא על פני הבחירות האקראיות של המכניזם M.

אינטואיציה: אפילו אם היריב מכיר את כל הכניסות בדטהבייס חוץ מאת הכניסה שלי, ואז הוא רואה את הפלט של המכניזם, אז הוא לא לומד הרבה על הכניסה שלי מכיוון שלא משנה מה היה שם הפלט של המכניזם היה נראה בערך אותו דבר.

<u>הערות:</u>

- פרטיות דפרנציאלית זו תכונה של <u>האלגוריתם</u> שמנתח את המידע.
- ההגדרה סימטרית ביחס ל X,X' בגלל הכמת לכל.
- נחשוב על ε בתור קבוע קטן, למשל $\varepsilon = 0.1$. זכרו כי עבור ε קטן מתקיים $e^{\varepsilon} \approx (1 + \varepsilon)$.
- על הפרמטר δ אפשר לחשוב בתור "הסתברות לכישלון של פרטיות" ונרצה שהוא יהיה מאוד קטן. למשל 2^{-n} או $n^{-\log n}$...
- חייבים לדרוש $\delta \ll 1/n$ אחרת ההגדרה לא מספקת.
- ככל ש- ε, δ קטנים יותר מקבלים יותר פרטיות.
- אם $\varepsilon = \delta = 0$ אז חזרנו להגדרה של דלניוס ולא נוכל לעשות איתה כלום... (למה?)
- עבור המקרה בו $\delta = 0$ במקום לסמן (ε, δ)-פרטיות דפרנציאלית, לפעמים נסמן בקיצור ε-פרטיות דפרנציאלית.
- עבור המקרה בו $\delta = 0$ ניתן לפשט קצת את ההגדרה (אם R בת מנייה) ולקבל את ההגדרה השקולה הבאה: מכניזם $M: D^n \to R$ מקיים ε-פרטיות דפרנציאלית אם לכל $X, X' \in D^n$ שכנים ולכל **<u>איבר</u>** $s \in R$ מתקיים

$$\Pr[M(X) = s] \leq e^{\varepsilon} \cdot \Pr[M(X') = s]$$

כלומר כאן s הוא <u>איבר</u> ב- R ולא תת קבוצה...

- המכניזם M הוא אקראי ואקראיות <u>הכרחית</u>!

<u>תרגיל:</u> הראו כי לכל אלגוריתם דטרמניסטי A מתקיים: או שהאלגוריתם לא משמר פרטיות דפרנציאלית (לכל ε, δ סבירים) או שהאלגוריתם לא תלוי בקלט שלו (כלומר לא תלוי בדטהבייס).