# מדעי הנתונים וביג-דאטה - מדע חדש בהתהוות

ליאור רוקח ורמי שקד[1]

**המהפכה הדיגיטלית שבה אנו חיים מאפשרת לנו לייצר ולאגור כמויות עצומות של נתונים. בעשור האחרון התפתח המושג של ביג-דאטה (Big Data) או נְתוּנֵי עָתֵק בעברית תקנית. במקור המושג ביג-דאטה מתייחס בעיקרו למאגרי נתונים עצומים כל כך עד ששיטות מסורתיות לעיבוד נתונים אינן מתאימות עוד. אולם לאחרונה רבים נוטים להשתמש במושג ביג-דאטה כשם כולל לשימוש בטכניקות מתקדמות לעיבוד וניתוח נתונים ללא קשר לגודל הנתונים.**

## רקע

לאחרונה עלתה חברת קיימברידג' אנליטיקה לכותרות כאשר התברר כי חברה זו סייעה בתשלום למטה הבחירות של טראמפ לנתח כמות גדולה של נתונים פרטיים שנאספו ברשת החברתית של פייסבוק. במסגרת זו ביצעה החברה הערכת מבנה אישיות ואופי המוני שלא באמצעות שאלון פסיכולוגי אלא באמצעות ניתוח העקבות הדיגיטליים שמשאירים המשתמשים ברשת החברתית. שיטה זו פותחה על ידי חוקרי אקדמיה (Kosinski et al., 2013) ואפשרה למטה של טראמפ לשלוח לכל הבוחרים הפוטנציאליים מידע תעמולתי המותאם לאישיות שלהם. אירוע זה הוא דוגמה מעשית ליכולות שעידן הביג-דאטה מביא.

המהפכה הדיגיטלית שבה אנו חיים מאפשרת לנו לייצר ולאגור כמויות עצומות של נתונים. ההערכה המקובלת היא שכמות הנתונים שנוצרה בעולם בשנתיים האחרונות גדולה משמעותית מכמות הנתונים שנוצרה לפני כן משחר ההיסטוריה ועד שנת 2015. בשנת 2012 העריכה חברת הייעוץ IDC (2012) שבשנת 2010 השאיר אחריו כל אדם עקבות דיגיטליים של כ-3 ג'יגה בַּיִת (באנגלית: byte) של נתונים בכל יום, משמע כשלושה מיליארדי מספרים ואותיות בכל יום. רק לצורך המחשה - אם נדפיס את הנתונים הללו על גבי נייר ונסדר אותם בערימה, נוכל להגיע לשמש ובחזרה כ-10 פעמים. הצפי שעד שנת 2020 תגיע כמות הנתונים ל-146 ג'יגה לאדם ביום, דהיינו גידול של פי 50 בתוך עשור. הנתונים אף הפכו לנכסים משמעותיים של חברות בינלאומיות. הכוח העיקרי של חברות כגון גוגל או פייסבוק גלום בנתונים שהם אוספים אודותינו ובניתוח חכם של הנתונים באופן שעונה על צרכיהן העסקיים. חברות שונות מנפיקות כרטיסי אשראי עם אפשרות להחזר כספי (cash back) רק כדי לקבל גישה להרגלי הצריכה שלנו. אף על פי כן ההערכה המקובלת היא שרק כחצי אחוז מכלל הנתונים שנאספים מנותחים בסופו של דבר.

שלוש מגמות אפשרו את הגידול העצום בכמות הנתונים. תחילה היו אלה מערכות המידע השייכות לארגונים (כגון בנקים וחברות תקשורת) שאפשרו איסוף נתונים מתמשך. בהמשך, עם הופעת האינטרנט והרשתות החברתיות, התעצמה מגמה זו כאשר המשתמשים החלו לתרום נתונים בעצמם. כמות הנתונים שכלל האנושות מייצרת כתוצאה משימוש באינטרנט וברשתות החברתית מגיעה למספרים עצומים. בכל דקה שחולפת נשלחות למעלה מ-16 מיליון הודעות WhatsApp ומבוצעים למעלה מ-3.5 מיליון חיפושים חדשים בגוגל. בכל דקה כחצי מיליון ציוצים חדשים מתפרסמים ברשת ה-Twitter, ואנו מעלים כרבע מיליון תמונות חדשות לפייסבוק. בכל דקה המשתמשים ברחבי העולם

1 פרופ' ליאור רוקח, ראש המחלקה להנדסת מערכות תוכנה ומידע, אוניברסיטת בן גוריון בנגב, מייסד המעבדה לביג-דאטה סא"ל (במיל') רמי שקד, מרצה וחוקר טכנולוגיות למידה, לשעבר מפקד בית הספר למקצועות המחשב של צה"ל

מעלים כ-700 שעות וידאו חדשות בעוד שיתר המשתמשים צופים בכ-5 מיליון סרטונים.

המגמה השלישית שמשפיעה רבות על גידול בכמות הנתונים היא האינטרנט של הדברים (IoT = Internet of Things). מדובר בנתונים שנוצרים אוטומטית על ידי המכונה וללא התערבות ישירה של המשתמשים האנושיים. כבר היום כל טלפון נייד מייצר כמות עצומה של נתונים באמצעות החיישנים שנמצאים בתוכו. במחקר שביצענו לאחרונה הסכימו חמישים נבדקים להתקין תוכנה על גבי הטלפון הנייד שלהם כדי שזו תדגום חלק קטן מהנתונים שמיוצרים על ידי החיישנים. לאחר כשנה של דגימה התקבל בסיס נתונים הכולל 10 מיליארדי רשומות (Mirsky et al. 2017). הרכבים האוטונומיים הצפויים לכבוש את כבישי ארצנו בעשור הבא כוללים מספר גדול עוד יותר של חיישנים העוקבים בצורה רציפה אחר הסביבה. נתונים אלו נאספים על ידי חיישני רכב המשותפים למספר כלי רכב, וזאת כדי למנוע היווצרות של גודש תנועה או כדי להתריע בפני סכנות בכביש.

ההתפתחות האדירה הזו המלווה אותנו בשנים האחרונות בכל תחומי החברה, תעשייה, כלכלה, שירות, בידור, חינוך ורפואה יוצרת לארגונים ממשלתיים, ציבוריים ופרטיים אתגרים לא פשוטים, כאשר הם באים לנתח, להבין ולעשות שימוש בנתונים הנאגרים במערכות אלה כדי לשפר את יעילותם ורווחיותם. מורכבות הנתונים וכמויות הנתונים הנאספים מדי יום בארגונים אלו גורמים לקושי, ומאידך תורמים באיכותם להצלחתו של הארגון.

## ביג-דאטה

על רקע זה התפתח בעשור האחרון המושג של ביג-דאטה (Big Data) או נְתוּנֵי עָתֵק בעברית תקנית. במקור המושג ביג דאטה מתייחס בעיקרו למאגרי נתונים עצומים כל כך עד ששיטות מסורתיות לעיבוד נתונים אינן מתאימות עוד. אולם לאחרונה רבים נוטים להשתמש במושג ביג-דאטה כשם כולל לשימוש בטכניקות מתקדמות לעיבוד וניתוח נתונים ללא קשר לגודל הנתונים. אפשר לאפיין כל פתרון בסביבת Big Data על בסיס מספר ממדים שונים. מדעני הנתונים נוהגים לאפיין זאת כך:

א. נפח (Volume) - מתייחס לכמות הנתונים שנאגרת בבסיס הנתונים.

ב. מהירות (Velocity)- מתייחס לקצב שבו מתווספים נתונים חדשים למאגר.

ג. גיוון (Variety)- מתייחס למגוון הנתונים הנשמרים במאגר. כיום ניתן לאסוף באותו המאגר מגוון רחב של נתונים הכולל: נתונים מספריים, טקסטואליים, תמונה, שמע, וידאו, חישה וכו'

פרט לשלושת הממדים העיקריים, יש שמוסיפים שבעה ממדי V נוספים עד להשלמת 10 ממדים, בפרט: Variability (חוסר עקביות של בסיס הנתונים), Veracity (אי-אמינות), Volatility (אי-זמינות), Validity (נכונות), Vulnerability (פגיעות), Value (ערך), Visualization (ויזואליזציה).

השלבים העיקרים בפיתוח מערכות מבוססי נתוני עתק הם אלה (Maimon & Rokach, 2010):

א. איסוף הנתונים הגולמיים - בשלב זה הנתונים הגולמיים נאספים ומאוחסנים בתוך בסיסי הנתונים. לרוב שלב זה נעשה ממילא כחלק מהתפעול השוטף של מערכת המידע ולפני שמחליטים על השימושים העתידיים שייעשו בנתונים. במרבית הארגונים נוקטים בגישה של "אסוף כפי יכולתך" ושומרים את כל הנתונים.

ב. הגדרת מטרות המערכת - בשלב זה מחליטים מה המטרה העיקרית של מערכת הביג-דאטה ואילו מטרות מצפים להשיג.

ג. עיבוד מקדים - בשלב זה מכינים את הנתונים לצורך ניתוחם. העיבוד המקדים כולל מספר רב של שלבי משנה כגון:

1. טיוב הנתונים - בשלב זה בודקים את איכות הנתונים ומבצעים תיקון לנתונים משובשים כגון: ערכים מחוץ לטווח האפשרי (למשל, גיל שלילי של לקוח) או צירופים לא הגיוניים (לקוח בן 12 שיש לו שלושה ילדים).
2. זיהוי והגדרה של מאפיינים בעלי משמעות שניתן לחלצם מתוך הנתונים הגולמיים ואשר עשויים לסייע בניתוח מושכל של הנתונים. בשלב זה מקובל להיעזר בידע של מומחי התוכן במערכת.
3. חילוץ מאפיינים - בשלב זה מיישמים את המאפיינים שהוגדרו בשלב הקודם וממירים את הנתונים הגולמיים למבנה החדש.
4. בחירת הנתונים שישמשו לניתוח - בשלב זה אנו בוחרים את הרשומות שתשתתפנה בבניית המדגם ואת המאפיינים שישמשו לייצוגם, וזאת כדי להקל על השלב

הבא ולהבטיח את טיב המודל שיתקבל.

ה. ניתוח הנתונים - בשלב זה נעזרים בשיטות ואלגוריתמים לאפיון הנתונים באופן שיאפשר חיזוי או הערכה של נתונים חדשים בשלבים הבאים. לרוב התוצר של שלב זה הוא מודל או מודלים המשרתים את המטרות שהוגדרו.

ו. בחינת טיב המודל - בשלב זה מוודאים כי המודל שהתקבל אכן תקף. אחת השיטות המקובלות היא לבחון את תקפות המודל על גבי נתונים שלא שימשו בשלבים הקודמים וזאת כדי להימנע מתופעות לא רצויות כגון התאמת יתר (Overfitting). התאמת יתר מתרחשת כאשר המודל מורכב יתר על המידה. תופעה זו גורמת למודל ללמוד רעשים סטטיסטיים בנתונים כאילו הם מייצגים תופעות אמיתיות ובכך לאבד את היכולת להכליל.

ז. יישום המודל - בשלב זה משתמשים במודל באופן שוטף ורציף.

התהליך בכללותו הנו תהליך איטרטיבי ובהתאם להיזון החוזר המתקבל מיישום המודל, אפשר לחזור לכל אחד מהשלבים הקודמים ולשפר את המודל.

## מדעי הנתונים

השלב העיקרי ובעל היכולת הטובה ביותר להשיג ערך מוסף הוא שלב ניתוח הנתונים. פיתוח טכניקות ניתוח נתונים רבות ושונות במהלך השנים הוביל לתחום חדש המכונה כיום מדע נתונים (Data Science). תחום זה עוסק בניתוח נתונים לשם הפקת מידע וידע, קבלת החלטות ומיכון של מערכות מתוך מקורות פנימיים וחיצוניים לארגון במטרה לתמוך ולשפר את ההחלטות הארגון. העובד שאחראי על ניתוח הנתונים בארגון, מכונה מדען הנתונים (Data Scientist), והוא בדרך כלל משלב יכולות מקצועיות משלושה תחומים עיקריים: פיתוח תוכנה, מתמטיקה והבנה עסקית.

מדע הנתונים הפך עם השנים לגורם מכריע בסביבה התחרותית ומשמש את כל הרבדים בארגון, החל בהחלטות תפעוליות וכלה בשיפור התכנון האסטרטגי. בארגונים מודרניים מבינים היום שהנתונים הרבים הנאגרים במערכות המידע של הארגון (למשל, מידע על לקוחות, על תהליכים ועל עסקאות) הם אחד מנכסיו העיקריים של הארגון, ושניתוח מושכל שלהם מייצר יתרון גדול לבעליו. ככל שכמויות הנתונים הנאספים וכוח המחשוב גדלים (שני דברים הקורים כבר שנים ברציפות ובקצב מואץ), כך יכולת המערכות לפתור בעיות ולענות על הצרכים משתפרת בצורה משמעותית. בענפים מסוימים ניתן להכריז על מהפכות למחצה בפיתוח מערכות ובהפקת ידע ותובנות חדשות עקב הכנסת כלים הלומדים בעזרת שימוש בנתונים. למשל, שיפור המעשה החינוכי, יכולת החיפוש במאגרי נתונים עצומים, יכולת ניהוג כלי רכב אוטונומיים, ביולוגיה חישובית, עיבוד שפה טבעית, רובוטיקה, ראיית ושמיעת מכונה, תחבורה וערים חכמות, רפואה אישית, סייבר ועוד.

ניתוח הנתונים הנאספים על אדם מסוים עשוי לעיתים לגלות דברים אודותיו לפני שהוא בעצמו מודע להם. הדוגמה הידועה ביותר בתחום היא המקרה שהתרחש בארצות הברית. רשת הקמעונות Target החליטה לנתח את נתוני הרכישות של לקוחותיה (Duhiggfeb, 2012). הם הציבו מטרה לאתר משפחות צעירות עוד בשלבים הראשונים של ההיריון כדי להציע להן מוצרים מתאימים. לשם כך הם ניתחו את הרגלי הצריכה של הנשים ברשת, זמן רב לפני שהן החלו לרכוש מוצרי תינוקות. ניתוח זה גילה כי הרגלי הצריכה משתנים לעיתים עוד לפני שהנשים עצמן גילו שהן בהיריון. על בסיס יכולת החיזוי הזו החלה חברת Target לשלוח קופונים מתאימים לנשים שהמודל מעריך שהן כרגע בהיריון. הדבר אף גרם לתקרית לא נעימה שבה לקוח כועס נכנס לאחד מסניפי הרשת ודרש לדבר עם מנהל: "הבת שלי קיבלה את זה בדואר!", אמר. "היא עדיין בתיכון, ואתם שולחים לה קופונים לבגדי תינוקות ועריסות? האם אתם מנסים לעודד אותה להיכנס להריון?". מנהל הסניף התנצל והבטיח להסיר את הבת מרשימת התפוצה לקופונים. מספר ימים לאחר מכן התקשר האב שוב לסניף ובקול נבוך אמר כי "מסתבר שהיו כמה פעילויות בבית שלי שלא הייתי מודע להן לגמרי. בתי אמורה ללדת באוגוסט. אני חייב לך התנצלות."

מדעי הנתונים (Data Science) עושים שימוש בשיטות שמגיעות מדיסציפלינות אקדמיות שונות ובעיקר מדיסציפלינות אלה:

א. מתמטיקה בכלל, סטטיסטיקה וחקר ביצועים בפרט.

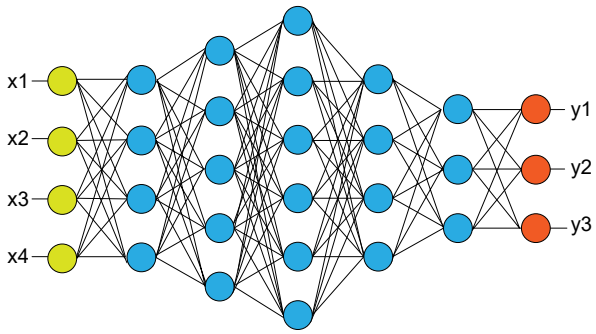
ב. מדעי המחשב בכלל ובינה מלאכותית ולמידה חישובית בפרט.

## הקשר בין ביג דאטה לבינה מלאכותית

במילים פשוטות, בינה מלאכותית (Artificial Intelligence) היא דיסציפלינה שנועדה לחקות ולדמות מגוון יכולות אנושיות באמצעות מחשב ובכך להפכו למחשב "חכם". למידה חישובית (Machine Learning) היא תת-תחום בבינה מלאכותית שנועדה לאפשר למערכות ויישומי מחשב ללמוד ולהשתפר בהשגת מטרותיהן מתוך נתונים שנאספו ואשר מייצגים את התנסויות העבר. למידה חישובית אחראית במידה רבה לפריצות הדרך שאנו חווים בשנים האחרונות בתחום הטכנולוגי.

אולי השימוש הפופולרי הראשון בלמידה חישובית הוא מנוע החיפוש של גוגל. היכולת שלנו להזין שאילתת חיפוש או סתם שאלה ולקבל תשובות רלוונטיות, נובעת מהעובדה שמנוע החיפוש של גוגל למד מה התשובה הרלוונטית מתוך טריליוני חיפושים שבוצעו על ידי משתמשים אחרים. כך גוגל מסוגלת אף להשלים את מילות החיפוש עוד לפני שהספקנו להקליד אותן. כל חיפוש שאנו מבצעים בגוגל מאפשר למערכת שלהם להמשיך ללמוד ולהשתפר. מאחורי השירות הפשוט לכאורה עומדות חוות שרתים הכוללות מיליוני מחשבים עם כוח עיבוד חזק במיוחד.

בשנים האחרונות חלו התפתחויות משמעותיות בתחום הלמידה החישובית עם הפיכת השיטה של למידה עמוקה (Deep learning) למעשית. למידה עמוקה אפשרה להשיג פריצות דרך המאפשרות למחשב לחקות יכולות אנושיות. למשל, לפי האתגר ImageNet, החל משנת 2015 יכולת המחשב לזהות עצמים בתמונה עולה על זו של אדם ממוצע (He et al., 2015). אחת היכולות השכליות הראשונות והחשובות ביותר שרוכשים תינוקות בשנתם הראשונה היא זיהוי פנים של אנשים שונים. אך גם בזה המחשב עולה לאין שיעור על בני אדם. שערו בנפשכם את יכולתכם להבחין בין שני אנשים ממוצא סיני או שני אחים תאומים. המחשב עולה על האדם בפעולה זו, בין היתר כיוון שהוא הוזן בכמות עצומה של דוגמאות, כמות רבה מזו שאנו בני התמותה נחשפים אליה במהלך חיינו. בתחום הראייה הממוחשבת המחשב מצליח גם במשימות מורכבות יותר כגון פענוח אוטומטי של תצלומי ממוגרפיה באיכות דומה, ולעיתים באיכות שאף עולה על זו של רופא רדיולוג מומחה. גם במקרה זה התקבלה היכולת העל-אנושית מתוך למידה חישובית של מאגרים הכוללים מיליוני תצלומי ממוגרפיה מתויגים עם אבחנות קודמות. ההצלחות אינן רק נחלתם של יישומי הראייה הממוחשבת. בשנים האחרונות חלו התפתחויות משמעותיות ביכולת המחשב להבין שפת דיבור כמו גם היכולת לנתח טקסטים כתובים.




איור 1: המחשה של רשת עצבית

למידה עמוקה מבוססת בעיקרה על מודל של רשת עצבית מלאכותית (איור 1) (ANN-Artificial Neural Network) שקיבלה את ההשראה שלה מהתהליכים המתקיימים במוחם של בעלי החיים מתקדמים (Bengio et al., 2015). למותר לציין כי המודל הראשון של רשת עצבית מלאכותית פורסם עוד בשנות ה-40 של המאה הקודמת, אך השימוש במודל הפך למעשי רק בשנות האלפיים בגלל מספר התפתחויות שחברו להן יחדיו:

א. התפתחויות בחומרת המחשב - השיפור בביצועים של המעבדים ובפרט של המעבדים הגרפיים (שבמקור שימשו בעיקר בתחום משחקי המחשב) כמו גם הגידול בנפח זיכרונות המחשב מאפשרים לאמן רשת עצבית גדולה בזמן סביר.

ב. התפתחויות באלגוריתמים לאימון רשת עצבית אשר מאפשרות לאמן רשתות עמוקות של קשרים בצורה נכונה יותר.

ג. אך התרומה המשמעותית ביותר של הבינה המלאכותית היא ללא ספק ה-Big Data - דהיינו, זמינות של כמות גדולה של נתונים לאימון. התפתחות ה-Big Data היא הכרחית מכיוון שבחלק מאתגרי הבינה המלאכותית נדרשת כמות עצומה של נתונים כדי לאמן את המכונה. זאת בניגוד לבני אדם שדי להם במספר מצומצם של דוגמאות כדי ללמוד מושג חדש (Lehrach et al., 2017). למשל, כדי ללמד פעוטות להבחין בין כיסא לשולחן, די בדוגמאות ספורות. אולם את המכונה יש להזין בכמות גדולה של דוגמאות כדי שתוכל להבחין

בין כיסא בעל ארבע רגליים לבין שולחן בעל ארבע רגליים או כדי שתוכל להבין שגם שולחן מרובע וגם שולחן עגול הם שולחנות. בעידן ה-Big Data איסוף הכמות הדרושה של דוגמאות ואחסונן בבסיס נתונים הפך למשימה ישימה, וכך ניתן להעמיד בפני המכונה כמות גדולה של נתונים שלא הייתה ברשותנו בעבר.

קיימים מודלים שונים של רשתות עצביות. המשותף לכולם הוא קיומם של צומתי עיבוד המייצגים את הנוירונים הביולוגיים שקשורים זה לזה. רשת עצבית מלאכותית מאופיינת על ידי מספר הנוירונים, מבנה הרשת, מספר השכבות, אופן החיבור בין הנוירונים ברשת וכדומה. תהליך הלמידה נועד לקבוע את עוצמת (משקל) הקישור של כל קשר ברשת העצבית. הלמידה מתבצעת על-ידי "תגמול" "ועניש" של קשרים שונים ועל ידי חשיפת רשת הנוירונים לדוגמאות רבות. "תגמול" ו"ענישה" של הקשרים מתבצע על ידי שינוי המשקל של אותו הקשר, כך שכל קשר ש"מתוגמל" - יגדל משקלו, וכל קשר ש"נענש" - ירד משקלו. לרוב רשתות עצביות מורכבות ועמוקות יותר יכולות ללמוד משימות מורכבות יותר.

### ביג-דאטה בחינוך

פורמט בית הספר אשר עוצב בימי המהפכה התעשייתית לפני כ-150 שנה מתבסס על עקרונות ההומניזם, השוויון והטכנולוגיה הדלה דאז. עיקרון השוויון והצורך בכוח עבודה משכיל ומיומן הוביל לפתרון המהיר והקל שלפיו כולם לומדים את אותו הדבר, כך שהמורים מכוונים את הוראתם לתלמידים הממוצעים באותה דרך ובאותו זמן (זלקוביץ וגולדשטיין, 2011). עיקרון השוויון נשמר ברובו גם היום, אך כיום הטכנולוגיה המתקדמת מאפשרת לנו להתייחס גם לייחודיות של כל התלמידים וכל המורים, אם נחפוץ בכך.

המחויבות של מערכת החינוך להכין את הבוגרים למציאות המשתנה נושאת עמה הבטחה לחולל שינויים מהותיים. גישות חינוכיות חדשות, אשר שוללות את גישת ההוראה הזהה לכל התלמידים, חרטו על דגלן את הרעיון שלפיו במרכז הלמידה יש להציב את ייחודיות הלומדים, ובשל כך בית הספר חייב להתאים עצמו לצרכים, ליכולות ולרקע הקודם של כל תלמיד ותלמיד (Gardner, 1983). מימוש השינוי הנדרש והתאמתו לתלמידים הוא מורכב, כיוון שהתלמידים השונים נבדלים זה מזה במגוון רב של פרמטרים: מגדר, תרבות, הרגלים, השפעת החברה והבית, גיל, מצב סוציו-אקונומי, תכונות אישיות ועוד.

כחלק מהצורך לתת מענה לשינוי זה התפתח תחום חדש יחסית המכונה כריית מידע בחינוך או EDM- Educational Data Mining. תחום זה מוגדר ע"י חוקרי חינוך רבים כאחד התחומים פורצי הדרך בתחום החינוך המתקדם (אבני ואברום, 2015). בבסיס הרעיון עומדת ההנחה שניתן ללמוד מתוך כריית נתונים איכותית ומתמשכת ברשת החינוכית ובמערכות הלמידה הדיגיטאליות (LMS) על התנהגות הלומדים וסגנונות הלמידה השונים, על אינטראקציות בין הלומדים ועל האינטראקציה בין הלומדים לבין המערכת. כל זאת מתוך הביג-דאטה הנאגר במערכות הללו.
"חֲנֹךְ לַנַּעַר עַל-פִּי דַרְכּוֹ גַּם כִּי-יַזְקִין לֹא-יָסוּר מִמֶּנָּה" (משלי כ"ב פסוק ו').

### למידה הסתגלותית (אדפטיבית) ולמידה מותאמת אישית (פרסונאלית)

בכיתה אופיינית ניתן למצוא תלמידים בעלי צרכים וסגנונות לימוד שונים, והאחריות הכמעט בלתי אפשרית להתאמת החומר הלימודי לכולם מונחת על כתפיהם של המורים. למידה הסתגלותית (הידועה גם בוריאציה מעט שונה כלמידה פרסונלית או מותאמת) מבוססת על התאמה אישית של התוכן, תפיסת ההוראה, אמצעי הלימוד ושיטות הלימוד לכל הלומדים, והיא מבצעת התאמה זו לאורך כל שלבי הלימוד על בסיס מידע שנאסף לאורך התהליך. תחום זה מיושם ע"י שיטות של כריית מידע, למידת מכונה וסטטיסטיקה. מטרת התחום היא לשפר את השיטות לחקר הנתונים הללו על מנת לגלות תובנות חדשות על האופן שבו תלמידים לומדים, וכך להתאים את תהליך הלמידה באופן מתמשך (בן-צדוק 2011; שמש, 2017).

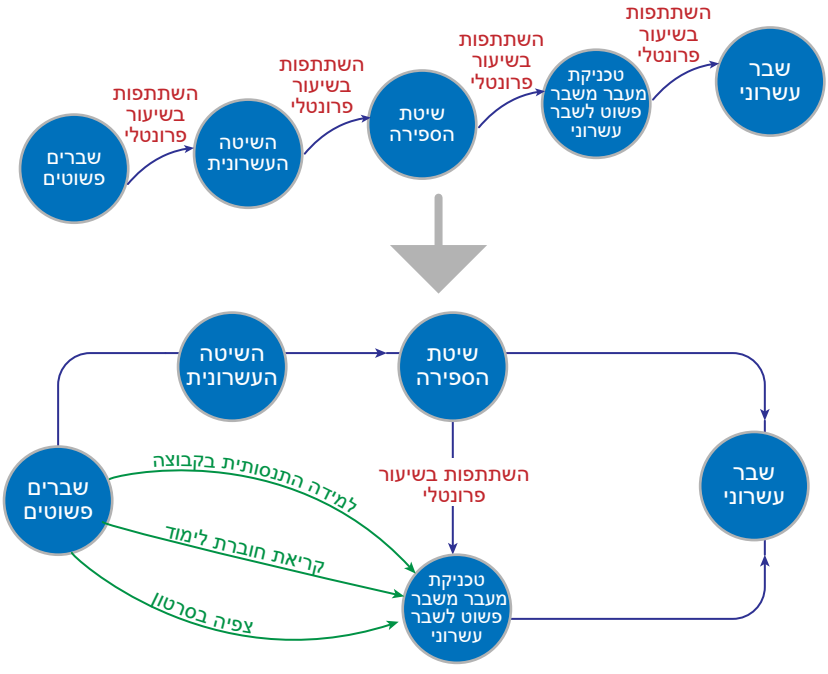
השיטה שבה המורים בבתי הספר נוהגים לשפר את הלמידה, מתבססת בעיקרה על ציונים, מדדים ומחוונים שלאורם הם תופסים את המתרחש בכיתה. גישה זו היא גישה "תגובתית", מאוחרת מדיי ותוצאתית למצב הנתון. לעומת זאת EDM דוגל במעבר מפסיביות לאקטיביות במעשה החינוכי. היכולת לאפשר לנו להיות פרואקטיביים יותר, לצפות בעוד מועד את מה שעומד

לקרות ולהתאים את המעשה החינוכי למציאות המתהווה.
מערכות ה-EDM מאפשרות לממש בקלות יחסית את העקרונות המרכזיים העומדים בבסיס הלמידה האדפטיבית והפרסונאלית, והמפורטים כאן:

1. שקיפות הנתונים לכלל המעורבים במעשה החינוכי (תלמידים, מורים, הורים, מנהלים).
2. אבחון מדויק ושוטף של מיקומם של התלמידים במסלול הלמידה וזיהוי תבניתי של אפיוני הלמידה המתאימים להם ביותר. האבחון יכול להתייחס לקצב הלימוד, העניין שהלומדים מגלים, חולשות, קשב, השפעת מאפייני הקבוצה וגודלה וכו'.
3. התאמת תוכנית למידה אישית לתלמידים על בסיס האבחון כדי שהמערכת תאפשר למורים להתאים את תוכנית הלמידה באופן אישי לכל התלמידים באופן שוטף ומתמשך.
4. משוב "איכותי" מבוסס-נתונים תדיר ומיָדי לתלמידים ולמורים המעיד על התקדמותם של התלמידים בתוכנית הלמידה האישית שעוצבה עבורם, וכך ניתן בכל פעם מחדש להתאים להם את מסלול הלמידה.
5. סְקָיְילָבִילִיוּת (scalability או סילוּמיוּת) ושכפול הפתרונות המותאמים למערכות חינוך נוספות.
6. אג'ילִיות (Agility או גמישות) - תוכנית הלמידה גמישה ומשתנה ומסוגלת להתאים את עצמה באופן תדיר להתפתחות ולשינוי שהתלמידים עוברים.

ניתן לעשות שימוש בתהליכי EDM גם למטרות חינוכיות נוספות. להלן מספר דוגמאות:

- תהליכי הוראה-למידה-הערכה (הֶלָ"ה) מדויקים יותר בדגש על מתן משוב למורות ולמורים בזמן אמת. המשוב מבוסס על איסוף נתונים ועיבודם בצורה מהירה (לדוגמה, זיהוי התנהגויות חריגות של תלמידים). בצורה זו ההתמקדות היא בתהליך הלמידה ולא רק בהישג הסופי. התמקדות זו נוהגת ביתר "צדק" כלפי הלומדים ומבטאת התייחסות מקיפה יותר למאמציהם המתמשכים וליחסם למקצוע.
- חיזוי נשירת סטודנטים באמצעות ניתוח הנתונים הרבים שמגיעים ממערכות ניהול למידה (LMS). המערכות הללו אוספות מידע על התנהגות הסטודנטים במערכת, לדוגמה:




איור 2: דוגמה קונספטואלית למעבר מלמידה "ליניארית" מבוססת הוראה לרשת למידה אדפטיבית

היקף ההשתתפות בפורומים, הזמן שלוקח לסטודנטים להגיש עבודה וכו'. שימוש זה נפוץ מאוד בסביבות MOOC's- massive open online course שבהן אחוז הנשירה גבוה במיוחד.

> ניבוי הצלחה בקורס - מערכות חיווי אזהרה מוקדמת (EWIS - early warning indicator system) מאפשרות לנבא מגמות כמו מוכנות לאקדמיה של סטודנטים פוטנציאליים, הגדלת פוטנציאל גיוס תלמידים למגמה מסוימת, לספק התראה מוקדמת אודות סיכונים ובעיות הנגזרים מפוטנציאל הנשירה, והחשוב מכך - מערכות אלו מספקות המלצות למניעת הנשירה של סטודנטים ותלמידים בסיכון.

עפ"י דו"ח Horizon מ-2017 נראה שב 3-5 השנים הקרובות ייעשה שימוש הולך וגובר בניתוח הנתונים ובשימוש בתובנות המערכת אשר ישפיעו כמעט מידית על הדרך שבה המורים מארגנים את הרצף, המהלך והמרחבים של הלמידה כמו גם את אופן השימוש בטכנולוגיה זו באופן רציף. EDM הנו המונח אשר מעצב ויעצב מחדש את תהליכי ההוראה ואת התהליכים הפדגוגיים, וניתן לשער שרעיונות אלו ימומשו יותר ויותר במערכות החינוך.

## לא הכול ורוד

עידן ה-Big Data טומן בחובו גם סכנות. הקלות הבלתי נסבלת שבה נאספים נתונים אודותינו, עלולה לפגוע משמעותית בפרטיות שלנו. נתון בודד כשלעצמו עשוי להיות בלתי מזיק. אך השילוב של נתונים שנאספים לאורך זמן עשויים לגלות טפחים רבים, גם כאלה שלא היינו רוצים לגלות. חלק מהנתונים אנו מנדבים בעצמנו למשל, כאשר אנו מפרסמים Post ברשת החברתית. אך חלק אחר מהנתונים מתפרסמים על ידי אחרים, למשל, כאשר החברים שלנו ברשת החברתית מזכירים אותנו ב-Post שלהם או מעלים תמונה שבה אנו מופיעים. זו הסיבה שבעידן הנוכחי קשה הרבה יותר להסתיר, למשל, נטייה מינית, כי די בסקירת פרופילים של חבריו של פלוני ברשת חברתית, די בה כדי להסיק פרטים גם לגביו ('כלשון הפתגם 'אמור לי מי הם חבריך - ואומר לך מי אתה").
לפיכך ברור שלעיתים איסוף הנתונים וניתוחם באתרי האינטרנט השונים מיטיבים עם בעלי האתר על חשבון האינטרס שלנו. למעשה, בכל פעם שאנו רוכשים מוצר באינטרנט או בוחרים את הטיסה הבאה, סביר להניח שמאחורי הקלעים קיים מנגנון מבוסס-למידה חישובית המעריך את המחיר שאנו נהיה מוכנים לשלם עבור המוצר או השירות, ובהתאם לכך קובע את מחירו.

לאור כל זאת החלו ממשלות ברחבי העולם לפתח רגולציה שתגביל את איסוף הנתונים הקשורים בבני אדם ולהשתמש בהם. למשל, חוק הגנת הפרטיות בישראל מסדיר את פעילותם של מאגרי מידע הכוללים מידע פרטי ורגיש. במסגרת חוק זה כל בעל מאגר מידע מסוג זה חייב להירשם אצל רשם מאגרי המידע, למנות מנהל שיהיה אחראי עליו, לפרט את המטרות שלשמן הוקם המאגר ולפעול אך ורק לפיהן. בנוסף כל אדם זכאי לעיין במידע המתייחס אליו המוחזק במאגר מידע (למעט מאגרי מידע של מערכת הביטחון), ובעל המאגר חייב לאפשר לו לעיין במידע זה. האיחוד האירופי הוא ללא ספק מוביל הרגולציה המתקדמת בתחום. בזמן כתיבת שורות אלה כללי ה-GDPR (General Data Protection Regulation) שנחקקו בשנת 2016, הופכים לבני אכיפה וכוללים מספר זכויות ובכללן: הזכות להישכח, שלפיה לכל אדם עומדת הזכות למחיקת כל הנתונים שנאספו אודותיו או הזכות להתנגד לעיבוד אוטומטי של נתוניו, דהיינו, פלוני רשאי לסרב שתתקבלנה החלטות בעניינו כתוצאה מניתוח אוטומטי של הנתונים שנאספו לגביו.

ביג דאטה ובינה מלאכותית גם מעצבים מחדש את שוק העבודה. מקצועות הצווארון הכחול (כגון נהגי משאית ופועלי ייצור) צפויים להיעלם מהעולם. אך גם מקצועות הצווארון הלבן עשויים להיות מושפעים מהתפתחויות בתחום. למשל, לא נזדקק עוד לרופא רדיולוג כדי לפענח תוצרים של דימות רפואי. המחשב יכול למלא את המשימה לאחר שהוא אומן באמצעות כמות גדולה של נתונים ממקרי עבר שפוענחו ותויגו על ידי רדיולוגים.

כאמור, גם בתחום החינוך ביג-דאטה ולמידה חישובית צפויים לעשות מהפך. אולם מערכות מידע חכמות ככל שתהיינה אינן תחליף למורים! יש להיזהר מהתפיסה שלפיה הנתונים הם חזות הכול ושמחשב צמוד לתלמידים ונתונים בידי המורים הם הפתרון לכל בעיות החינוך. לעולם לא יחליפו הנתונים את הקשר האישי בין מורים לתלמידים ואת המסרים העוברים באמצעות קשר זה. אם כך אין זה פלא שעל פי הדו"ח האחרון של ארגון ה-OECD,

מקצועות החינוך הם בעלי הסיכוי הנמוך ביותר מכלל המקצועות המועדים לאוטומציה, בכלל זה מקצועות הרפואה או הניהול (Nedelkoska and Quintini, 2018). ברם תפקידם של המורים עתיד להשתנות. מערכות המידע מבוססות הביג-דאטה יעצימו את המורים ויאפשרו להם להגיע למלוא הפוטנציאל כאנשי מקצוע, המכירים את תלמידיהם היטב, מלווים ומנחים אותם בדרך להישגים הטובים ביותר.

ההשערה הרווחת היא שבעשור הבא יוכל המחשב לפתח מודעות עצמית ויכולות קוגניטיביות אוטונומיות אשר יאפשרו לו לעצב עצמאית מודלים לקבלת החלטות. מודלים כאלו עשויים להוביל גם למשברים שכיום עדיין מנוהלים ידי בני אדם. למשל, בשנת 2010 התרחש אירוע המכונה Flash Crash שבו מדד הבורסה הראשי בניו-יורק נפל פתאומית וללא כל הסבר, ולאחר זמן קצר התאושש כאילו מעולם לא התרחש. חקירה שבאה בעקבות האירוע מגלה שהגורם העיקרי לקריסה היו מחשבי אַלְגו-טריידינג (או בעברית מסחר אלגוריתמי) אשר מחליטים בעצמם לבצע פעולות של קנייה ומכירה בקצב גבוה על ידי ניתוח אוטומטי של נתוני המסחר. מאז התרחשו עוד מספר אירועי קריסה זמניים שנבעו מניתוח אוטומטי של נתונים. אירוע מעניין נוסף התרחש בשנת 2016 כאשר חברת מייקרוסופט שילבה צ'טבוט (רובוט שיחה) שנועד לצייץ אוטונומית ברשת ה-Twitter. האינטראקציה שהייתה לרובוט עם משתמשים אנושיים הפכה אותו בתוך 24 שעות למיזנתרופ וגזען עד שהמהנדסים של מיקרוסופט נאלצו לנתקו לאלתר מהרשת.

סינגולריות טכנולוגית היא נקודת הזמן שבה תשיג הבינה המלאכותית יכולת אינטלקטואלית הגבוהה מזו של בני אדם. יש הרואים בכך את האמצאה האחרונה של האנושות. מאותו הרגע תוכל המכונה להפיק בעצמה את ההמצאות הבאות. כך אולי תוכל האנושות למצוא מזור לבעיות שבני התמותה לא הצליחו עדיין לפתור, כגון מציאת תרופות למחלות חשוכות מרפא. כיום מספר המדענים שעוסקים בבעיית מחקר מסוימת הנו מוגבל. אך כאשר תשיג מכונה את היכולת האינטלקטואליות של מדען, הרי שניתן יהיה להמשיך לשפרה כי אין לה מחסום ביולוגי. נקודת הזמן הזו מעלה הרבה שאלות פילוסופיות ודילמות מוסריות כגון: האם ניתן לפתח מצפון במכונות? כך או כך אנו חיים ללא ספק בעידן מרתק.

## רשימת ספרות


אבני, ע., ו אברום, ר. (2015). ביג-דאטה, חינוך ואתיקה - מנתונים לתובנות. ביג-דאטה, חינוך ואתיקה, 1-33.
https://ianethics.com/wp-content/uploads/2015/02/EduBigData-AI1-2015.pdf

בן-צדוק, ג. (2011). כריית נתונים למטרת חקר התנהגויות תלמידים בסביבות מתוקשבות. מכון מופ"ת. http://portal.macam.ac.il/ArticlePage.aspx?id=4419

זליקוביץ", צ., וגולדשטיין, א. (2011). למידה מותאמת אישית. בתוך מלמד, ע. וגולדשטיין, א. (עור'). הוראה ולמידה בעידן הדיגיטלי, 77-96.

שמש, א. (2017). סקירת ספרות בנושא פרסונאליזציה בחינוך - (/http://piechallenge.org.il)

Duhiggfeb, C. (20012). How Companies Learn Your Secrets, The New-York Times Magazine, 16.

Gardner, H. (1983). Frames of Mind: The Theory of Multiple Intelligences.

He, K., Zhang, X., Ren, S. and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026-1034).

IDC Digital Universe Study: Big Data, Bigger Digital Shadows and Biggest Growth in the Far East, IDC (December 2012).

Kosinski, M. Stillwell, D. & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. PNAS USA. 110 (15): 5802–5805.

Maimon, O. & Rokach, L. (2010) Data Mining and Knowledge Discovery Handbook, 2nd ed. ISBN 978-0-387-09822-7

Mirsky, Y., Shabtai, A., Shapira, B., Elovici, Y., & Rokach, L. (2017) Anomaly detection for smartphone data streams. Pervasive and Mobile Computing 35: 83-107.

Nedelkoska, L. & Quintini, G. (2018). Automation, skills use and training, OECD Social, Employment and Migration Working Papers, No. 202, OECD Publishing, Paris.