מבוא למדעי הנתונים

- מה הם מדעי הנתונים
- יישומים שונים של מדעי הנתונים
- תהליך העבודה במדעי הנתונים

מדעי הנתונים ועתיד החינוך

הכיתה העתידית - אלגוריתם תומך למידה שעוקב אחרי הלמידה שלכם ומתאים לכם באופן אישי את התכנים שאתם נהנים מהם, אוהבים אותם. האלגוריתם יודע להתאים לכם את השאלות, את הדוגמאות, יודע לנטר אם מישהו מתקשה וצריך עזרה של מורה. כל האלגוריתמים האלה בעוד כמה שנים אנחנו נראה אותם, כלומר יהיו כיתות עתידיות שבהם כל אחד לומד בעצמו עם אלגוריתם שמתאים לו את הלמידה באופן אישי כפי שהוא צריך אותה. היום זה אחד העיסוקים של מדעי הנתונים.

מדעי הנתונים: יוצרים ערך מנתונים

מדעי הנתונים - מקצוע חדש שעוסק בלהפיק ערך מנתונים. אנחנו נמצאים בעידן שבו יש נתונים בכל מקום, נתונים נאספים באינטרנט, ברשתות החברתיות, בחיפושים שאנחנו עושים בגוגל ונתונים נאספים גם מהפלאפונים שלנו, הם מודדים כמה אנחנו הולכים, לאן אנחנו הולכים. נתונים נאספים מחיישנים, רכבת ישראל, למשל, יודעת היכן נמצאת כל רכבת בכל זמן נתון, היא יודעת איזה נוסעים נכנסים לאיזה תחנה ומתי הם יוצאים ממנה. אנחנו חיים בעידן שבו נתונים עוטפים אותנו ונאספים כל הזמן והם הפכו לכמו משאב טבע. מדע מדעי הנתונים עוסק בלהוציא ערך מהנתונים האלה.
לערך מהנתונים יכול להיות ערך מחקרי שבו מנסים להפוך את הנתונים לאיזו שהיא תובנה (insight) או ערך כלכלי שבו חברות מסחריות מנסות להפוך נתונים לכסף.

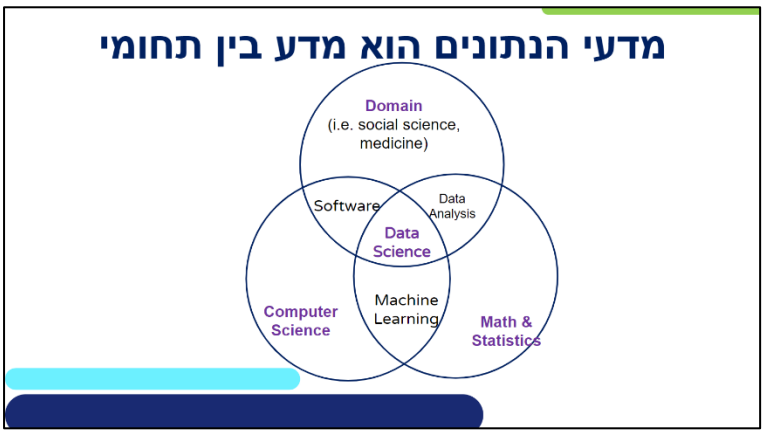
מדעי הנתונים התחילו לפני כמה שנים מהרצון להפוך נתונים לכסף. חברות כמו גוגל גילו שכאשר הפרסומות מותאמות אישית לצופה, המכירות טובות יותר (כמות הלחיצות על הפרסומת הרבה יותר גבוהה). לכן הם הבינו שכדאי להם לחקור אלגוריתמים שיוכלו להתאים פרסומות באופן אישי. באופן דומה, חברות, כמו נטפליקס, עסקו בהתאמה של תכנים, סרטים, סדרות לצופה. חברות אלה מתחרות על היכולת שלהם לספק תוכן יותר מעניין ומותאם וזאת באמצעות אלגוריתמים. דוגמא נוספת זו חברת פייסבוק אשר רוצה לסדר את ה"פיד" שלכם בצורה יותר מעניינת והיא עושה את זה בכך שהיא רואה במה אתם מתעניינים יותר, לאיזה פוסטים אתם נכנסים וכך היא מצליחה להתאים לכם "פיד" והיא מרוויחה יותר כסף.

מדענים, לעומת זאת, מחפשים לקחת את הנתונים האלה ולהפיק מהם תובנות שיש להן ערך אחר (לאו דווקא כלכלי רווחי) לדוגמה- למידה אישית. אם נוכל לספק לתלמידים מערכות למידה אישיות כך שהלמידה תהיה יותר מעניינת עבורם, מותאמת עבורם, בקצב נכון עבורם, ככה נוכל לספק חוויית למידה טובה ויעילה יותר.

לפיכך, מדעי הנתונים עוסקים ב:

- הפקת כסף לחברות מסחריות
- תובנות וערך אחר עבור מדענים

מדעי הנתונים הוא מדע בינתחומי



מדעי הנתונים הוא מדע בינתחומי. ישנם שלושה תחומים:

- מדעי המחשב
- מתמטיקה וסטטיסטיקה
- דומיין- שזה תחום הדעת שממנו נלקחים הנתונים

בחיתוך שבין שלושת התחומים נמצאים מדעי המידע data science

מדעי הנתונים מציעים פרדיגמה חדשה

הפרדיגמה המדעית הנפוצה כיום היא:

השערה - ניסוי.... ⟸ תיאוריה.

המדען מעלה השערה ואז הוא צריך לבנות ניסוי כדי לאושש או להפריך את ההשערה הזו וכך בתהליך שחוזר על עצמו מגלים ומפתחים ידע חדש ומייצרים תיאוריה מדעית (לאחר שיש הרבה תובנות וידע).

מדעי הנתונים מציעים ומאפשרים פרדיגמה מדעית חדשה, לפיה מתחילים מהנתונים. כלומר, הנתונים כבר קיימים, לא צריך לתכנן שום ניסוי כדי לאסוף את הנתונים. למשל, אם רוצים לעסוק בנתונים של תחבורה, ווייז אוספת את כל הנתונים שאנחנו רוצים, ובנוסף גם הפלאפונים ורכבת ישראל אוספים מידע על תחבורה. ישנם נתונים על תחומי חיים רבים, כדוגמת ספורט, בריאות.., בתחום החינוך, למשל, MOOC וקמפוסIL אלו הן מערכות ענקיות שמאפשרות לרבים ללמוד קורסים בו זמנית. המערכות האלה אוספות נתונים על הלומדים.

המערכות השונות כבר אוספות נתונים ואנחנו צריכים לשאול את עצמנו מה אנחנו עושים עם הנתונים האלה ואיזה תובנות אנחנו יכולים להפיק מהם. לא צריך להתחיל מההשערה, לחשוב איזה השערה ולאחר מכן לתכנן את הניסוי מכיוון שהניסוי כבר התבצע, הנתונים כבר נאספו. עכשיו נשאר לנו להפיק מהנתונים את התובנות.

באופן זה מתאפשרת פרדיגמה מדעית חדשה שבה לא מתחילים מההשערה אלא מתחילים מהנתונים ובאמצעות חקר נתונים מגיעים לתובנות ובסופו של דבר יכולים לפתח תיאוריה.

היום בתחומים רבים מאוד, כמו ביולוגיה, פיזיקה, כימיה, רפואה, מדעי הנתונים מאפשרים לשנות את שיטת העבודה ולעבוד בצורה אחרת.

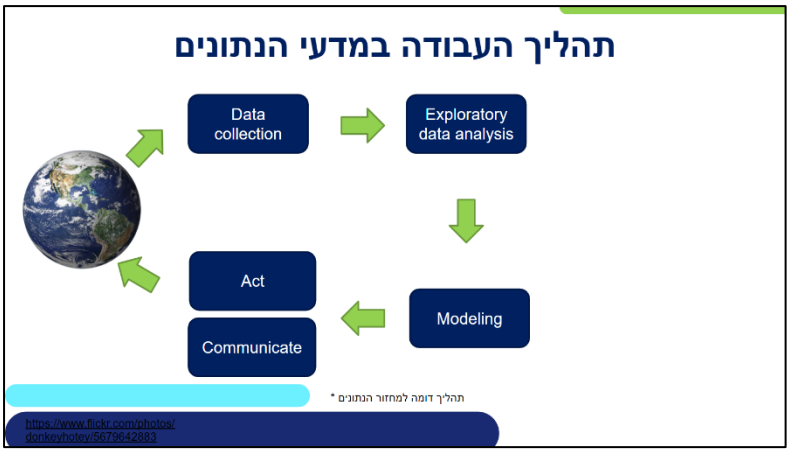
דוגמאות ליישומים

1. **<u>ניטור מגפות</u>** - נתונים נאספים: כמה חולים יש, כמה חולים קשים יש, כמה החלימו... כל קבלת ההחלטות במדינה מתבססת על הנתונים שנאספים. חוקרים ומדענים רבים בעולם מנסים למצוא תובנות מהנתונים האלה, עורכים השוואות בין מדינות שונות ובצורה כזו מפיקים תובנות גם על הטיפולים הנכונים וגם על הדרך הנכונה להתמודד עם הנגיף וגם מנסים לחזות מה יהיה בעתיד.

2. **<u>חקר הסביבה</u>** - הסביבה משתנה באופן תדיר. באמצעות מעקב אחרי נתונים, למשל תצלומי לוויין ומזג אוויר, ניתן לחזות תופעות עתידיות. חיזוי מזג אוויר באמצעות אלגוריתמים מתוחכמים התפתח מאוד.

3. **<u>רפואה מותאמת אישית</u>** - הפלאפונים שלנו מנטרים הרבה מהפעילות שלנו: כמה אנחנו הולכים, כמה אנחנו זזים, יש שעון חכם מיוחד שמודד צעדים, שעון שמודד את קצב הלב... אם נהיה חולים, ניתן יהיה לקחת את כל הנתונים שנאספו עלינו בעבר ולראות איזה טיפולים עזרו לחולים שדומים לנו ובכך ניתן להתאים את המינון התרופתי או את התרופה עצמה בצורה אישית לכל חולה ולהצליח לטפל בצורה יותר טובה ומותאמת.

<u>תהליך העבודה במדעי הנתונים</u>



תהליך העבודה במדעי הנתונים מתחיל בעולם האמיתי ומסתיים בעולם האמיתי.
האנושות מספקת את הנתונים ובסוף אנחנו רוצים לעשות משהו עם הנתונים האלה ולהחזיר את זה לעולם האמיתי. תמיד יש לנתונים קשר ומשמעות לעולם האמיתי ולתוצאות הניתוח שלנו תמיד יש קשר לעולם האמיתי ומשמעות בעולם האמיתי.

איסוף נתונים - data collection - בשלב הראשון אוספים את הנתונים. לפעמים השלב הזה פשוט ולפעמים מורכב. נתונים יכולים להגיע במגוון גדול של צורות: תמונה, קול, מספר, קצב למידה של תלמידים או דופק (ביג דאטה).
שלב איסוף הנתונים הוא תהליך לא פשוט וצריך להשקיע בו תשומת לב. נתונים יכולים להיות חסרים, מלוכלכים ולכן <u>בשלב הזה נשקיע גם בלטייב את בסיס הנתונים שאנחנו אוספים ולוודא שאנחנו אוספים נתונים שהם באמת מדוייקים</u> ויוכלו לאפשר לנו להריץ את האלגוריתמים בשלב מאוחר יותר ולהסיק תובנות שהן תובנות נכונות.

שלב exploratory data analysis - זהו שלב שבו אנחנו מנתחים את הנתונים בעצמנו, עדיין אנחנו לא מפעילים אלגוריתמים. אנחנו נשתמש בכלים סטטיסטיים, נפעיל שיטות של ויזואליזציה, המחשה ויזואלית של הנתונים, נשרטט גרפים, נחפש קשרים בין נתונים שונים ובשלב הזה ננסה להשיג תובנות מהנתונים ולהבין אותם בצורה טובה.
כחוקרים אנושיים- אנחנו מנסים לקבל איזושהי אינטואיציה על מה שקורה במאגר הנתונים שלנו.

שלב ה- modeling - בשלב הזה מנסים לבנות מודלים. אלו יכולים להיות מודלים סטטיסטיים פשוטים, מתוחכמים או מודלים של למידת מכונה (נושא שיפורט בהמשך). המודלים האלה מנסים לחזות את העתיד, להבין את התופעה. למשל, תלמיד פתר את שאלה 3 מהר ואת שאלה 4 לאט, איזה שאלה כדאי עכשיו לתת לו? שאלה 5 או לחזור לשאלה דומה לשאלה 3. אנחנו מקבלים את ההחלטה באמצעות המודל.

שלב ה-act - בשלב הזה אנחנו פועלים. ניתן לתלמיד את השאלה הבאה שהכי מתאימה לו או לחולה את התרופה שהכי מתאימה לו. זאת, לפי החיזוי של המודלים.

שלב ה- communicat - בשלב הזה אנו נרצה לקשר את התוצאות שלנו עם העולם. למשל במחקרים בנושא הקורונה- רוצים להעביר את המודלים והתחזיות שלנו למקבלי החלטות, למשל משרד הבריאות.

לסיכום, המודל הזה דומה למחזור הנתונים. כמו במחזור נתונים התהליך מתחיל באיסוף הנתונים מהעולם האמיתי ממשיך לניתוח, בניית מודלים, החזרת התוצאות לעולם האמיתי ומשם מחליטים על פעולה כלשהי או מפיקים דוח לקבלת החלטות.

תרגול