

גרפי ידע, אונטולוגיות ו-GraphRAG על Amazon Neptune

tlv-dat309 — סיכום סשן, AWS Summit Tel Aviv 2026

Baruch Assif (Senior Partner SA, Data & AI Specialist, AWS) · Daniel Aviram (CTO, GrowersTech) · Erez Drutin (Solutions Architect)

10 בספטמבר 2026 | משך: 33 דקות | הופק מהקלטת הסשן

מסלול: Databases on AWS | רמה: 300

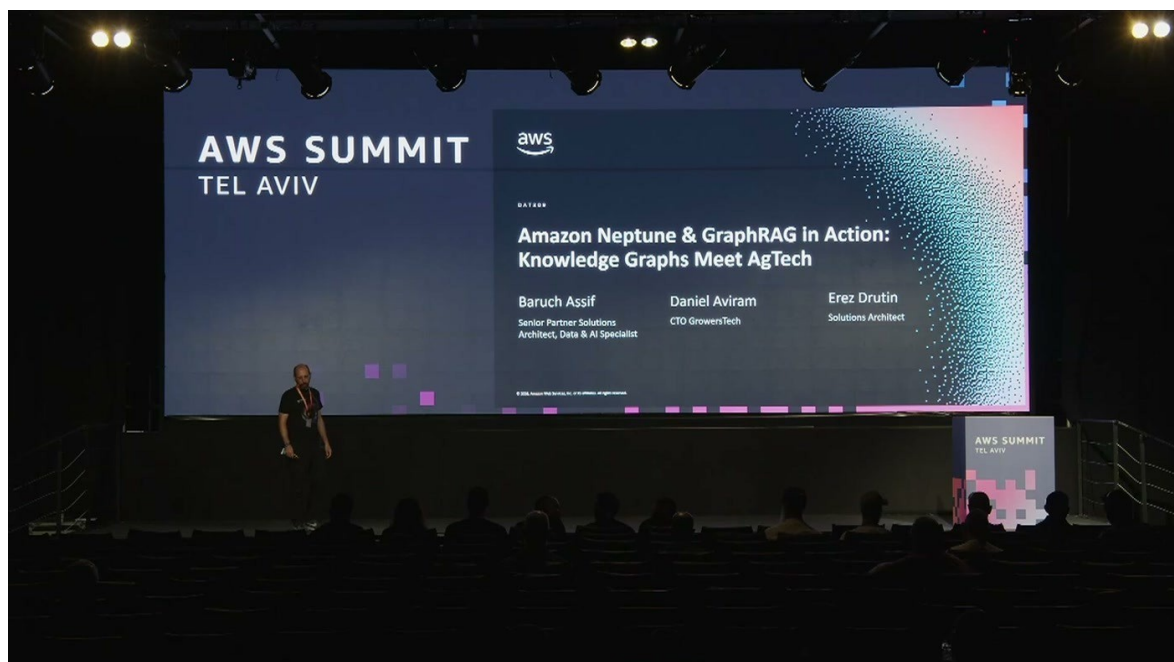
על המסמך הזה

המסמך מסכם את הסשן tlv-dat309 מתוך AWS Summit Tel Aviv 2026 — שלושה דוברים, שלושה חלקים, נושא אחד: למה גרף ידע הוא השכבה שחסרה בין הדאטה הארגוני לבין מודלי שפה וסוכנים. הוא נבנה מתמלול אוטומטי של ההקלטה ומהסליידים שחולצו מהווידאו עצמו, כך שאפשר לחזור לתוכן בלי לצפות שוב בשלושים ושלוש הדקות. חותמות הזמן בגוף המסמך מפנות לנקודה המדויקת בהקלטה.

איך לקרוא את זה

- **פרק 1 — תקציר מנהלים:** חמש דקות קריאה. מה נטען בסשן ועל סמך מה.
- **פרקים 2-5 — התיאוריה (Baruch Assif):** למה SQL נשבר, מה זו אונטולוגיה ומה זו סמנטיקה, ולמה RAG וקטורי לבדו מפספס.
- **פרקים 6-8 — המקרה האמיתי (Daniel Aviram, GrowersTech):** דאטה חקלאי מבולגן, המתכון לגרף ידע, והמספרים שמראים איפה בסיס נתונים רלציוני נעצר.
- **פרקים 9-11 — המימוש (Erez Drutin):** איך בונים GraphRAG בפועל, מהו Hierarchical Lexical Graph, ומהו חיפוש היברידי.
- **פרק 12 — מה לוקחים מכאן:** מסקנות, כלים פתוחים, ומילון מונחים.

הערות מקור המסמך הופק מתמלול אוטומטי של ההקלטה. התמלול סביר בקטעים המקצועיים, אך מונחים לועזיים ושימושים מוצגים מופיעים בו בשיבוש פונטי (למשל "נפטון" = Text-to-Cypher = "text to cipher", Neptune, "אקסיום" = AXIOM). הציטוטים מובאים כלשונם בתמלול ואומתו מול חותמת הזמן המצוינת; מונחים לועזיים נורמלו לאנגלית רק בגוף הטקסט ולא בתוך הציטוטים. כל המספרים במסמך מקורם בסליידים או בדברי הדוברים ומצוין בכל מקרה מהו המקור.



שלושת הדוברים כפי שהופיעו על סליד הפתיחה: ארכיטקט Data & AI מ-AWS, ה-CTO של GrowersTech, וארכיטקט פתרונות

1. תקציר מנהלים

הסשן מציג טענה אחת ומגבה אותה בשלוש שכבות: מודלי שפה לא יודעים את הדומיין שלכם, והדרך לעגן אותם היא לא עוד embedding אלא מודל ידע מפורש — אונטולוגיה שנטענת לגרף, ומעליה חיפוש היברידי שמשלב גרף ווקטורים. הדוגמה שנבחרה היא חקלאות, אבל הדוברים אומרים במפורש שזה בלופרינט שחוזר בתחומים אחרים.

- **הבעיה היא לא המודל אלא הדאטה הארגוני.** Baruch פתח בטענה ש"מה שמאגן באמת מודולי שפה לעולם האמיתי זה הדאטה הארגוני שלנו" [0:00], ושבלי אונטולוגיה שמגדירה את עולם המושגים ובלי סמנטיקה, סוכנים לא יחזירו תשובות רלוונטיות.
- **השאלות שהעסק שואל חוצות חמש-שש ישויות.** "השאלה הזאת מכילה לפחות חמישה או שישה אישויות שונות" [1:37], ו"רוב הזמן אם אנחנו נחבר ונעשה ג'וין של חמש או שש טבלאות, אנחנו כנראה לא נקבל לעולם את התשובה" [1:37].
- **יש לזה מספרים, לא רק אינטואיציה.** סליד של GrowersTech מציג השוואת עומק שאילתה: ב-4 קפיצות SQL רץ כ-26 דקות מול כ-1.4 שניות בגרף — פי 1,135 — ומ-5 קפיצות ומעלה SQL כלל לא מסיים. בעל פה: "בג'וינים החמישי המערכת קורסת" [21:49].
- **RAG וקטורי לבדו מפספס קשרים שאינם דמיון.** "אם אנחנו נשאל שאלה שהיא לא כל כך דומה במרחב הוקטורי, למשל יחסים של סיבה ותוצאה, קשרים של תלויות" [9:19] — התשובה הרלוונטית פשוט לא תגיע. הגרף מוסיף את ההקשר המבני.
- **המערכת בפרודקשן קיימת, לא תרגיל.** GrowersTech מריצה מערכת בשם AXIOM: "אנחנו מנתחים מיליארדי דאטה פוינטס, מנתחים מאות מיליוני טרנזקציות ומחזיקים אלפי אונטולוגיות מדומנים רבים ומגוונים" [23:25].
- **היתרון שנטען הוא הסבירות, לא רק הביצועים.** "היתרון הגדול והחזק ביותר, זה העובדה שיש כמעט אפס אלוסיציות" [20:16] — כי התשובה עוברת דרך מבנה שעבר ולידציה באמצעות האונטולוגיות והטקסונומיות.

2. מתחת לפני השטח: אונטולוגיה, סמנטיקה, דאטה

החלק הראשון של הסשן, בהגשת ארכיטקט ה-AI & Data מ-AWS, נפתח בתמונה אחת שנשארה על המסך לאורך כל הפתיחה: מה שרואים מעל הקרקע הוא ה-AI, ומה שמחזיק אותו מתחת לקרקע הוא שלוש שכבות — אונטולוגיה, סמנטיקה ודאטה.



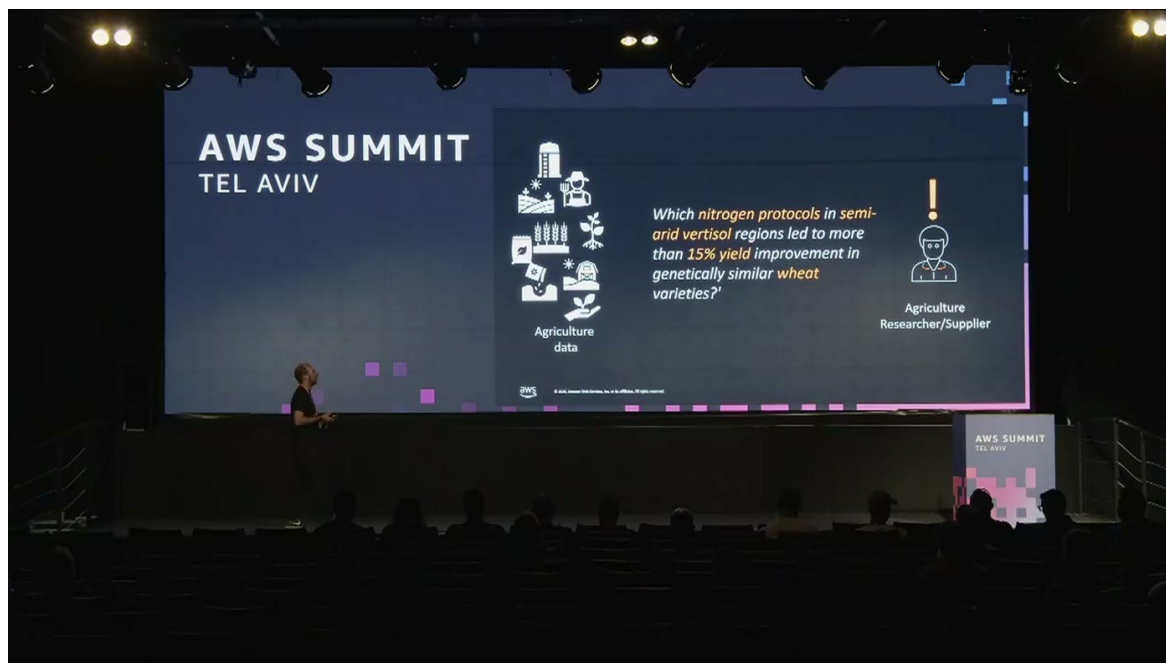
שלוש השכבות שמתחת ל-ONTOLOGY: AI Models / Agents: מעל SEMANTICS מעל DATA, כשהשורשים פרושים על פני מקורות המידע בשדה

הטיעון: "בלי אונטולוג דת של מגדירת עולם המושגים ובלי המשמעות האמיתית של המידע על ידי סימנטיקה ובלי דאטה איכותי ומובנה האג'נטיים לא יצליחו לדעת לנו תשובות שרלוונטיות לעולם האמיתי" [0:00]. התמלול משושב כאן, אך הכוונה ברורה: שלוש השכבות עובדות ביחד, והחוליה החסרה ברוב הארגונים היא דווקא השתיים העליונות.

מדוע חקלאות היא מקרה מבחן טוב? כי היא מייצרת כמות אדירה של מידע מקושר: "אם זה גשמים, סוגי דגנים, דגימות קרקע, תבניות מזג האוויר, גנטיקה של גידולים" [0:00] — וכל אלה קשורים זה בזה.

3. השאלה שמסד נתונים רלציוני לא עונה עליה

כדי להמחיש, הוצגה שאלה אחת שאגרונום או ספק ציוד חקלאי היה רוצה לשאול — והיא הוצגה באנגלית על הסליד.



השאלה כפי שנוסחה על המסך: אילו פרוטוקולי חנקן, באזורי ורטיסול צחיחים-למחצה, הובילו ליותר מ-15% שיפור ביבול בזני חיטה דומים גנטית

הניתוח היה כמותי ולא רטורי: "השאלה הזאת מכילה לפחות חמישה או שישה אישויות שונות" [1:37] — פרוטוקול, סוג קרקע, אזור, זן, גנטיקה, יבול. כדי לענות עליה צריך לחבר אותן, "וכדי לתשל את הדבר הזה ולחבר אותם, זה תהליך מאוד מאוד מסובך" [1:37].

— **Baruch Assif, [1:37]** רוב הזמן אם אנחנו נחבר ונעשה ג'ויין של חמש או שש טבלאות, אנחנו כנראה לא נקבל לעולם את התשובה. ולייצר SQL כזה, זה תהליך גם מאוד מאוד מסובך.

מכאן הוגדרו שני החלקים של הפתרון. הראשון: "ידע של המומחה" — איסוף הידע הארגוני והמשגתו בבסיס הנתונים [1:37]. השני: "מערכת שתדע לתשל את הדאטאבייס הזה" [3:10] — כלומר ממשק שמקבל שאלה בשפה טבעית ויודע להמיר אותה לשאילתה.

הארכיטקטורה כפי שהוצגה

ארבעת המרכיבים, בסדר שבו הוצגו [3:10]: הידע הארגוני נשמר כאונטולוגיות; "ה-LLM מספק את הממשק לשפה הטבעית" וממיר את השאלה לשאילתה מעל האונטולוגיה; Amazon Neptune הוא בסיס הנתונים — גם למידע הארגוני וגם לאונטולוגיות, ולפי הדובר מסוגל להחזיק גם גרף וגם וקטורים; ו-Amazon Bedrock מספק את המודל.

שימו לב שימו לב להפרדה: האונטולוגיה היא מה שהארגון מביא, ה-LLM הוא רק המתרגם מהשפה הטבעית אל השאילתה. הטענה של הסנן היא שברוב המימושים היחס הזה הפוך — משקיעים במודל ומדלגים על מודל הידע.

4. גרף, גרף ידע, אונטולוגיה וסמנטיקה — ההגדרות

החלק הזה של ההרצאה היה מכוון לקהל שלא מגיע מעולם הגרפים, וסידר ארבעה מונחים שנוטים להתערבב.

- **גרף:** מודל מידע משני סוגי אובייקטים — nodes שמחזיקים את המידע עצמו, ו-edges שמקשרים ביניהם [4:40].
 - **גרף ידע (Knowledge Graph):** מוסיף משמעות למבנה — "הנודים מייצגים אישויות בעולם האמיתי והאגים מייצגים את החיבורים בין האישויות האלה בעולם האמיתי" [4:40]. הוצג כצורה מודרנית של ה-Semantic Web, שחוזר לרלוונטיות דווקא בגלל מודלי השפה.
 - **אונטולוגיה:** "האונטולוגיה היא מודל אוג שמאגד את הסמנטיקה של התחום" [6:17] — כלומר מודל לוגי שמגדיר אובייקטים, constraints, properties, relationships.
 - **סמנטיקה:** "סמנטיקה בעצם היא הגדרות שמאפשרות את התוכנה לפרש את הכוונה" [7:48] — מה המספר או הערך אומרים בפועל.
- דוגמת האונטולוגיה שהוצגה הייתה של Crop — יבול: אובייקט מסוג יבול, property של ה-growth state שלו, relationship מסוג "נשתל ב-" אל מקום, ו-constraints מעל האובייקט, למשל שהיבול לעולם לא יהיה עשב שוטה [6:17]. המסקנה: "הכוח של האונטולוגיה זה בעצם לספק את הידע על גבי הדאטה שיש לנו" [7:48] — ובזכותה אפשר ללמד גם את ה-LLM וגם את העובדים בארגון את המטא-דאטה של הארגון עצמו.

הדוגמה שהכי נדבקת בזיכרון: 32 מעלות

— **Baruch Assif, [7:48]** אז 32 מעלות מצד אחד יכול להיות יום קיץ חם אבל מצד שני יכול להיות נקודת הכפייה של מים. ללא סמנטיקה אנחנו לא נבין בעצם מה ה-32 הזה אומר.

("נקודת הכפייה" בתמלול = נקודת הקפיאה; הכוונה להבדל בין 32 מעלות צלזיוס ל-32 מעלות פרנהייט.) מיד אחרי זה הוצג הייצוג בגרף: node של חיטה, קשת is a אל crop, קשת אל האזור שבו נשתלה, וקשת אל סוג הקרקע באותו אזור. "אפילו מהמבנה הפשוט הזה הם יכולים לשאול שאלות מאוד מורכבות" [9:19] — באיזה סוג קרקע נשתלה החיטה, או אילו זני חיטה יכולים לגדול באותו soil type.

5. למה RAG וקטורי לבדו לא מספיק

זה הטיעון המרכזי של החלק הראשון, והוא הוצג כניגוד ישיר ל-RAG הסטנדרטי שכולם מכירים. בחיפוש סמנטי לוקחים מסמכים, מייצרים embeddings, מחפשים במרחב הווקטורי ומוצאים את הקרוב ביותר — "חיפוש סמנטי בעצם נוצר לנו נתונים שהם מאוד מאוד דומים" [9:19]. הבעיה מתחילה כשהקשר הרלוונטי אינו קשר של דמיון.



ההבחנה שהסליד מעמיד: וקטור מחזיר תוכן דומה לשאלה, גרף מחזיר תוכן רלוונטי מבנית — גם כשהוא כלל לא דומה לה

— **Baruch Assif, [9:19]** אם אנחנו נשאל שאלה שהיא לא כל כך דומה במרחב הווקטורי, למשל יחסים של סיבה ותוצאה, קשרים של תלויות, קשרים של ארגונים, אז הם לא יופיעו במרחב הווקטורי שלנו, ולכן אנחנו נפספס את התשובה הרלוונטית.

הפתרון שהוצע הוא שילוב של שני סוגי הדאטה — וקטורים לחיפוש דמיון, גרף לחיפוש הקשרי-מבני: "על ידי הייבריד סרט, חיפוש היבריד, של חיבור של שתי הדאטאבסים ביחד, נוכל למצוא את התשובה היותר מדויקת. ככה נוכל לשפר את הריקול, מבי להקריב את הפרסיזן של הדאטה" [10:52]. כלומר: שיפור recall בלי להקריב precision.

סוגי קשרים שאינם סמנטיים

- parent-child — היררכיות
- cause and effect — סיבה ותוצאה
- legal taxonomy-ו social organization — מבנים ארגוניים ומערכות רגולטוריות
- מידע דומייני שהמודל לא אומן עליו

על שני האחרונים הדובר היה מפורש: "אם נחפש חיפוש סמנטי על מערכות רגולטוריות, אנחנו כנראה לא נמצא נתונים רלוונטיים, כי המודל עצמו לא מכיר את המערכת הרגולטורית הרלוונטית" [10:52], ובהמשך — "המודלים לא מאומנים על מידע חקלאי, ולכן אנחנו צריכים לעזור להם עם מידע נוסף" [10:52]. בנוסף, מבנה הגרף מאפשר שכבה אנליטית נוספת: תשאול הטופולוגיה, מקור הנתונים, ותלויות — "אם למשל נוד אחד נופל או נוד אחד משתנה, מה האפקט על שאר הנודים האחרים" [12:25].

6. הדוגמה: GreenFields Farm

כדי להראות את ההבדל בפועל, הוצג תרחיש מומצא אך מפורט. השאלה: מה היבול הצפוי של חיטה בחווה בעונה הזו.



חמש טענות מחמישה מקורות שונים — מחקר, אנליסט, דומיין, דוח וחדשות — שרק החיבור ביניהן נותן את התשובה הנכונה

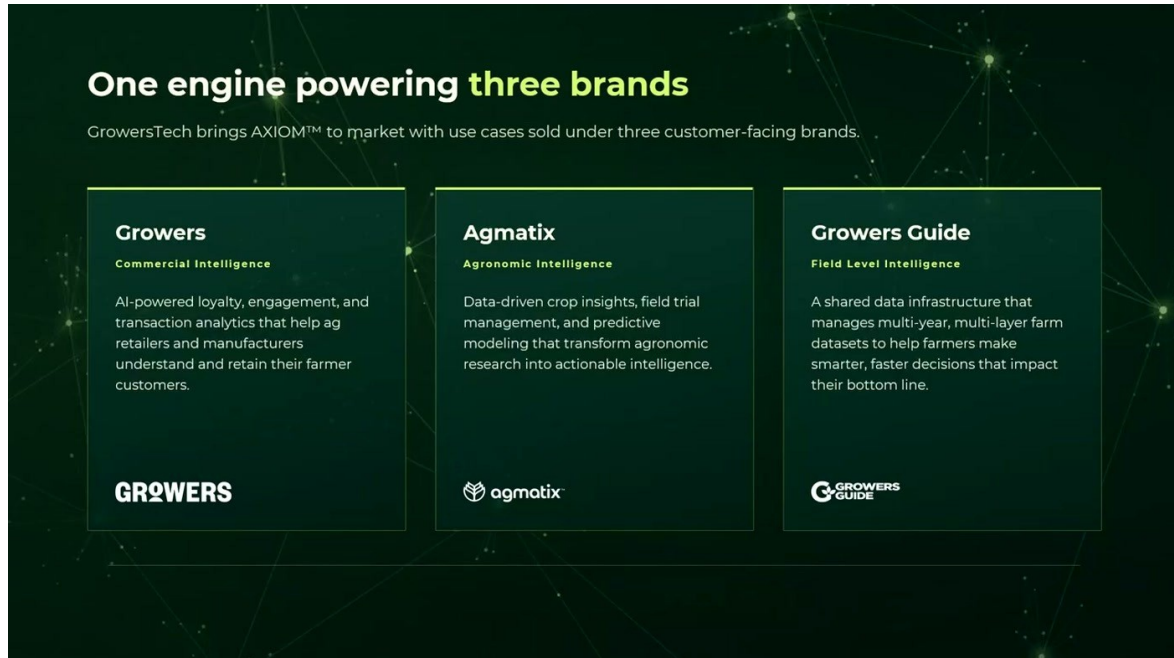
הסליד מפרק את התרחיש לטענות לפי מקור: חוות GreenFields שתלה זן חיטה עתיר-יבול בשם HY200 (מחקר); זן HY200 מראה עלייה של 25% ביבול בניסויים בתנאים צחיחים-למחצה (אנליסט); GreenFields משתמשת בתוכנית דיסון מדויק של NutriChem (דומיין); NutriChem רוכשת אשלג מ-AquaMin Mining (דוח); ו-AquaMin Mining הושעתה בשל הפרות זיהום מים (חדשות).

וכאן ההבדל: "האג'נט יבוא ויעשה חיפוש סמנטי עם רק רגיל ובעצם ימצא לנו רק את הנתונים שקשורים לחיפוש הסמנטי, בעצם רק מה שקשור לגרינפילד. ולכן האג'נט יחשוב שבעצם ההילד שלנו יהיה מאוד מאוד גבוה והכל טוב" [14:03]. הוספת שכבת הגרף מגלה את שרשרת האספקה — הדשן תלוי בספק שהושעה — ולכן "אנחנו נקבל תשובה לא מספקת" [14:03], מסקנה שמשנה את ההחלטה העסקית.

שימו לב שימו לב שהמידע הקריטי — ההשעיה של ספק המים — הוא הפריט הכי פחות "דומה" לשאלה על יבול חיטה. בדיוק זה מה שחיפוש וקטורי לא היה מחזיר, ובדיוק זה מה שמהלך של שלוש קפיצות בגרף מחזיר מיד.

7. GrowersTech: למה דאטה חקלאי מבולגן במיוחד

החלק השני הועבר על ידי ה-CTO של GrowersTech: "אני דניאל אווירם, ה-CTO של חברת גרורסטק" [14:03]. הוא פתח במיפוי שרשרת הערך החקלאית: "יש לנו שלושה שחקנים מרכזיים בקצה האחד יש את החקלאים בקצה השני יש את היצרנים" [15:38] — ובאמצע המשווקים (retailers), שאמורים לחבר ביניהם. החברה פונה לכל אחד מהם דרך מותג אחר.



מנוע אחד ושלושה מותגים — Growers לאינטליגנציה מסחרית, Agmatix לאינטליגנציה אגרנומית, Growers Guide לרמת השדה

- **חקלאים → Growers Guide:** מידע שמגיע מהמכונות — "כל טרקטור מלא בחיישנים" — מחיישני קרקע ומדגימות קרקע [15:38].
- **יצרנים → Agmatix:** עריכת ניסויים למוצרים חדשים שחייבים לעמוד ברגולציה, כולל human safety בתכשירי הדברה, וניתוח הניסויים להפקת תובנות אגרנומיות [15:38].
- **משווקים → Growers:** "אלה הם הלקוח היותר מסחרי, שם קוראים כל העסקאות" — חיבור ל-CRM ול-ERP שלהם, משיכת הטרנזקציות והפקת המלצות [17:11].

מעל אלה מתווספים מקורות חיצוניים: נתוני מזג אוויר וקרקע, remote sensing ולוויינות, ומאגרי open source. "מקורות המידע הם שונים ומגוונים" [17:11], ולכך מצטרפת בעיה היסטורית: "כשלוקחים בחשבון שגם עולם החקלאות זה בערך התעשייה הכי ותיקה בהיסטוריה, היא משתנה בין אזור גיאוגרפיקה זה לאחר" [17:11].



Agriculture data is uniquely messy

Six data families. Zero shared language. Every insight requires crossing between them.

Research Trials Controlled, scientific, structured, but locked in institutional silos.	Commercial Transactions Sales, inventory, pricing, product codes with no agronomic meaning.
Satellite & Remote Sensing Raster imagery, vegetation indices, a different paradigm entirely.	Soil & Weather Environmental, geospatial, time-series with no standard schema.
Farmer Practices Informal, regional, passed down, rarely digitized, never standardized.	Regulatory & Labels Compliance-driven, jurisdiction-specific with changes every season.

The same product, three systems

RESEARCH	RETAILER	ERP
"Urea 46-0-0"	"Granular nitrogen"	"FRT-4600"

Every insight requires crossing between these families. **That's the hard part.**

שש משפחות דאטה ללא שפה משותפת — וחמשת התווים שמדגימים את הבעיה: אותו מוצר, שלושה שמות בשלוש מערכות

אותו דשן, שלושה שמות

הדוגמה הקונקרטית שהוא נתן היא הדשן הבסיסי ביותר בחקלאות: "בעולם האגרונומי יקרה יוריה 46.00, אצל הריטיילר הוא יקרה גרנולה נייטרוגן כפרודקט" [17:11], "ובמערכת ה-ERP הוא יקרה [18:44] "FRT 4.600. על הסליד: "Urea 46-0-0" ≠ "Granular nitrogen" ≠ "FRT-4600". "ואנחנו בעצם צריכים להבין שמדובר באותו הדבר בדיוק כשאנחנו רוצים לנתח ולתת תובנות" [18:44] — וזו בדיוק הנקודה שבה אונטולוגיה וגרף ידע נכנסים.

8. המתכון: אונטולוגיה → גרף ידע → AI נוירו-סימבולי

הדובר הדגיש שהתהליך אינו ייחודי לחקלאות: "ולמעשה מה שתוארתי כרגע זה בלו פרינט מאוד נפוץ לא רק בעולם החקלאות, זה משהו שהוא מאוד מאוד שימושי בעולמות כמו Health Care, Financial Services ו-"Supply Chain [18:44]. שלושת השלבים:

- **שלב 1 — למדל את הדומיין:** בניית האונטולוגיות. "אנחנו בעצם מייצרים איזושהי אמת אחידה שמגדירה את הדומיין" [18:44].
- **שלב 2 — לקודד אותו כגרף ידע:** החלק האופרטיבי — הטמעת המידע ממקורות שונים לתוך המבנה [20:16].
- **שלב 3 — רק אז להוסיף את שכבת ה-AI:** שעובד מעל גרף נקרא נוירו סימבוליק [20:16] "AI — הצד הגרפי הוא הסימבולי, ה-LLM או ה-Machine Learning הם הצד הנוירוני שרץ מעליו.

— **Daniel Aviram, [20:16]** לחיבור הזה יש המון המון יתרונות. היתרון הגדול והחזק ביותר, זה העובדה שיש כמעט אפס אלוסינציות. כי אנחנו עוברים דרך מערכת שעברה ולידציה אינרנטית, דרך האונטולוגיות והטקסונומיות.

השאלה העסקית שממחישה את הצורך

אבל, הוא הוסיף, מניעת הזיות אינה הסיבה היחידה: "יש מקרים שבהם אי אפשר לספק תובנות בדרך המסורתית באמצעות דאטאבייסים [20:16] "SQL. השאלה שהוצגה כדוגמה: אילו משווקים צריכים להגדיל את מלאי קוטל הפטריות שלהם, באזורים שבהם הזנים שייזרעו בעונה הבאה רגישים למחלות.

When business meets agronomy

A crop protection manufacturer asks:

"Which retailers should increase shelf stock of our fungicide in regions where next season's planted varieties are highly susceptible to disease pressure?"

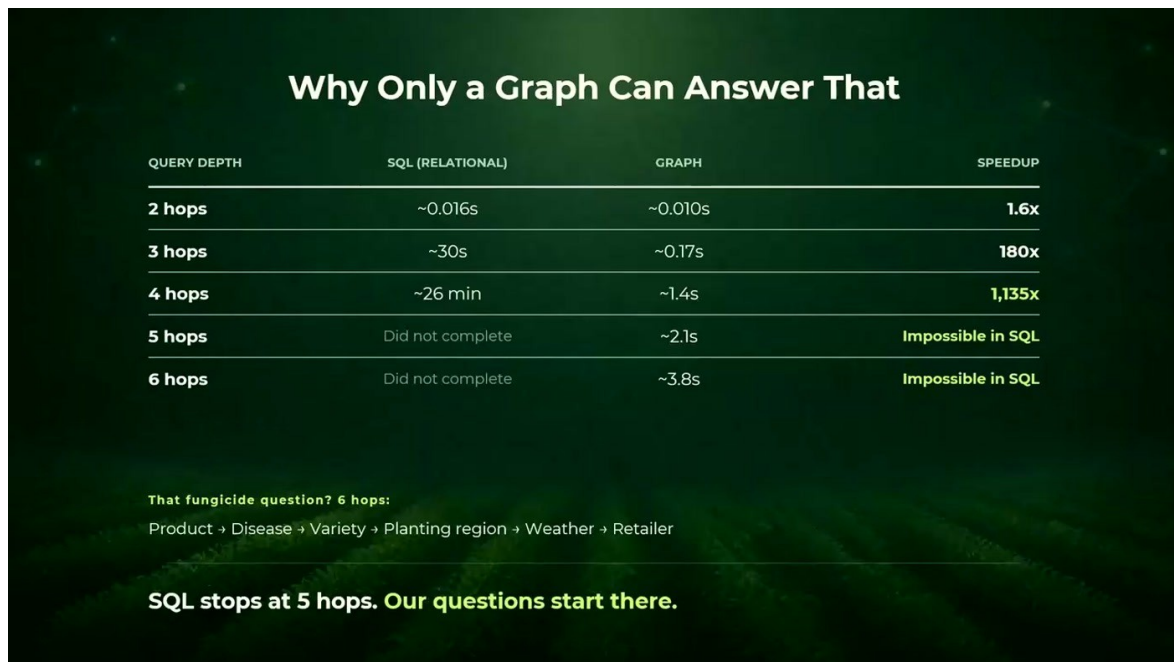
1 Product catalog Which diseases does this fungicide treat?	2 Crop varieties Which varieties are susceptible to those diseases?	3 Planting data Where are those varieties planted next season?
4 Weather forecasts Which regions favor disease development?	5 Retail network Which retailers serve those regions?	6 Transaction history What is their stock and sales velocity?

Six data families. One question.
Only a knowledge graph that connects commercial and agronomic data can answer it.

פירוק השאלה לשש משפחות דאטה: קטלוג מוצרים, זני גידול, נתוני זריעה, תחזיות מזג אוויר, רשת המשווקים והיסטוריית מלאי

הפירוק בעל פה תאם: "קטלוג המוצרים, איזה מחלה החומר מרפא, נתוני זנים, איזה זנים בכלל רגישים למחלה, נתוני שטילה, מקומות, תחזית מזג האוויר, ריסת רשת הכימונאים, והיסטוריית המחירות והמלאי שלהם" [21:49].

זהו הסליד היחיד בסשן שמציג מדידת ביצועים, והוא שווה עיון בפני עצמו.



זמני שאלתה לפי עומק קפיצות — SQL רלציוני מול גרף, עם גורם ההאצה בעמודה הימנית

האצה	גרף	SQL רלציוני	עומק
1.6x	~0.010 שני	~0.016 שני	2 קפיצות
180x	~0.17 שני	~30 שני	3 קפיצות
1,135x	~1.4 שני	~26 דקות	4 קפיצות
בלתי אפשרי ב-SQL	~2.1 שני	לא הסתיים	5 קפיצות
בלתי אפשרי ב-SQL	~3.8 שני	לא הסתיים	6 קפיצות

שאלת קוטל הפטריות היא, לפי הסליד, שאלה בעומק 6 קפיצות: מוצר ← מחלה ← זן ← אזור זריעה ← מזג אוויר ← משווק. שורת המחץ על הסליד: "SQL stops at 5 hops. Our questions start there." בעל פה זה נשמע כך: "הנתונים אומרים שאחרי ארבע ג'וינים מורכבים כאלה, לוקח בערך חצי שעה, בג'וינים החמישי המערכת קורסת" [21:49].

הערות מקור המספרים מגיעים מסליד של GrowersTech ולא מבנצ'מרק ציבורי. הסליד אינו מציין גדלי דאטה, סכימה, אינדוקסים או חומרה, ולכן הם מדגימים את סדר הגודל של הבעיה — התפוצצות קומבינטורית בעומק join — ולא מהווים השוואת מוצרים.

AXIOM — המערכת בפרודקשן

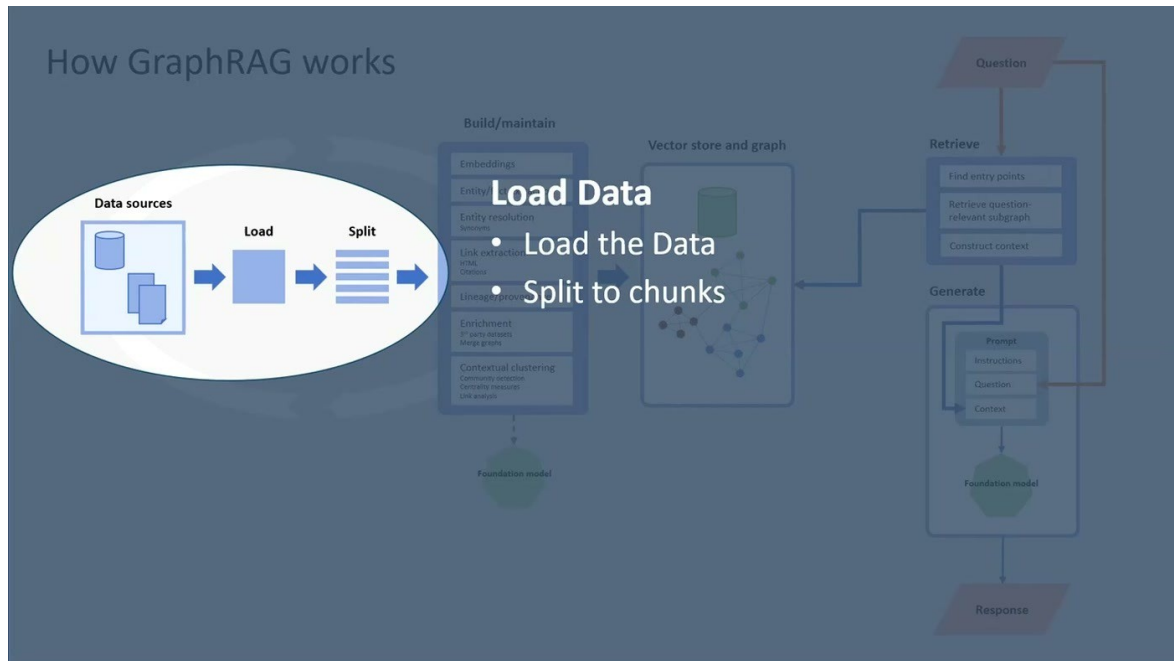
כל המידע משלושת המותגים מתנקז למערכת שנקראת AXIOM (בתמלול "אקסיום"), שמבצעת תהליך שהוגדר כסטנדרטי בתעשייה: איסוף ממקורות שונים, נרמול, ניקוי, חיבור באמצעות האונטולוגיות, הטמעה בגרף, ואז reasoning מעל הגרף [21:49]. הסקאלה: "אנחנו מנתחים מיליארדי דאטה פוינטס, מנתחים מאות מיליוני טרנזקציות ומחזיקים אלפי אונטולוגיות מדומנים רבים ומגוונים" [23:25].

הזרימה מצד הלקוח: שאלה עסקית מגיעה לסוכן, "האג'נט יכול להחליט אם השאלה שנשאלת היא שאלה שאנחנו יכולים לעשות מה שנקרא direct retrieval, מה שנקרא text to cipher ולגשת ישר גרף, או לעבור דרך גרף רג בצורה קצת יותר מורכבת" [23:25]. כלומר שני מסלולים: Text-to-Cypher ישיר מול הגרף, או GraphRAG מלא. בשני המקרים — "המידע הגיע מהגרף, זה אומר שהוא עבר את כל התהליכים והוא מקושר ו-explained, והוא "validated" [23:25].

— Daniel Aviram, [23:25] לסיכום there is no AI strategy without data strategy.

9. איך בונים GraphRAG בפועל

החלק השלישי, בהגשת ארכיטקט הפתרונות, הוגדר במפורש כחלק המעשי: "מטרה שלי זה שתבינו שגם אתם יכולים לעשות את זה" [24:56]. נקודת הפתיחה מרגיעה — "תהליך של בנייה של GraphRag כמערכת זה תהליך מאוד דומה בהתחלה שלו ל-Rag Classy" [24:56].



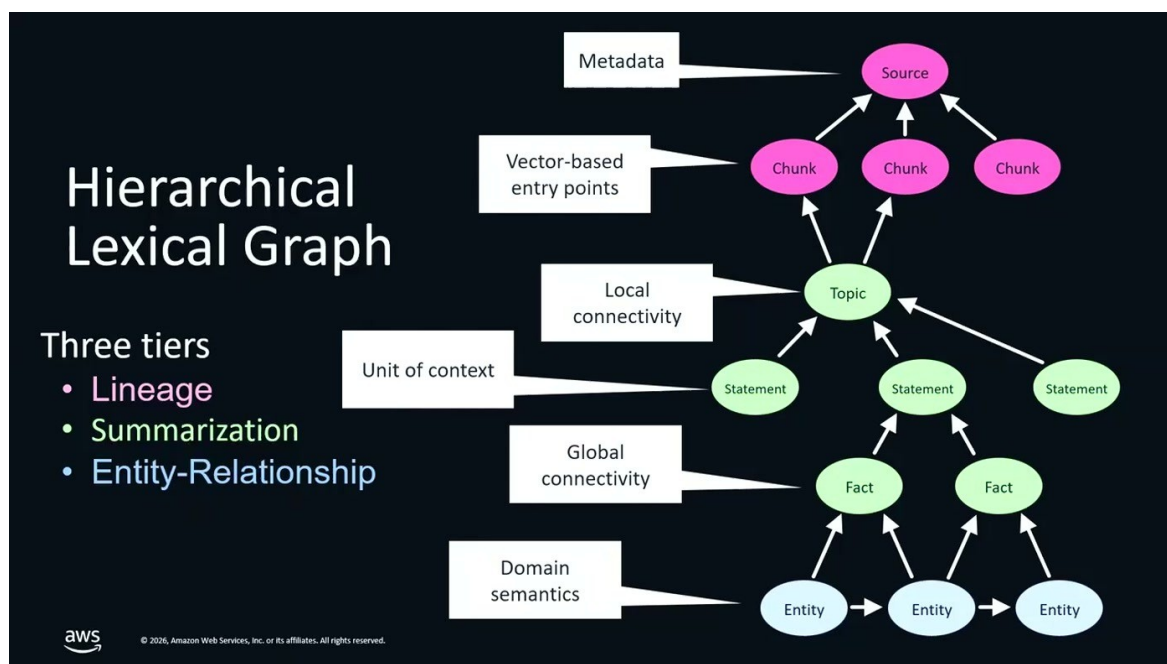
השלב הראשון בזרימה המלאה, זהה ל-RAG רגיל: טעינת מקורות הדאטה ופיצולם ל-chunks

לוקחים מקורות — "בין אם זה כפצת Text, PDF, HTML, you name it" — שוברים אותם ל-chunks וטוענים לבסיס הנתונים [24:56]. אסטרטגיית ה-chunking יכולה להיות לפי כמות תווים או לפי תווים פלוס סוף משפט, "ויש עולם שלם של תוכן שאפשר לצלול אליו מבחינת הסטרטגיות של [24:56] chunking".

מכאן הדרכים נפרדות. ב-RAG קלאסי יחידת המידע הבסיסית היא ה-embedding; ב-GraphRAG מוסיפים מעליה את האונטולוגיות ואת גרף הידע. שלב ה-Build/maintain כולל טכניקות כמו fact extraction, entity resolution ו-link extraction — "אנחנו מתפססים על הפאונדיישן מודל שלנו, כדי לחלץ ולהרכיב באמצעותם את המידע שעמד להיכנס לדאטאבייס" [26:26]. כלומר: מודל היסוד עצמו הוא הכלי שמחלץ את הישיות והקשרים מהטקסט הגולמי.

10. Hierarchical Lexical Graph — המבנה

הדפוס הפופולרי שהוצג לבניית הגרף מטקסט נקרא Hierarchical Lexical Graph, והוא נבנה בשלושה שכבות (tiers).



שלוש השכבות והתפקיד של כל אחת: Lineage מן המקור אל ה-chunks לטופיקים וטענות, ו-Entity-Relationship בתחתית

- **Lineage**: מסמך המקור וה-chunks שלו. לכל מסמך יש מטא-דאטה שאפשר לבחור — כולל גרסאות: "אפשר לשמור גרסה א', גרסה ב', גרסה latest" ולתשאל לפי timestamp של נקודה בזמן [29:28].
- **Summarization**: טופיקים, statements ו-facts. כל topic הוא אוסף של statements שמהם אפשר להבין משהו רחב יותר [26:26].
- **Entity-Relationship**: בתחתית — domain semantics, הקשרים בין הישויות עצמן. "כמו שאמרנו החלק הגרפי הקלאסי" [26:26].

local connectivity מול global connectivity

ההסבר ניתן דרך דוגמה מחוץ לעולם החקלאות: "אז הדוגמה של הטופיק שלנו היום תהיה הקריירה של לברון ג'יימס" [27:57]. כל עובדה על השחקן — שהוא נכנס לעונה ה-24 שלו, שהוא עומד לשחק בפילדלפיה, ששיחק בלוס אנג'לס לייקרס — היא statement, וכולן מתקבצות תחת טופיק אחד: הקריירה שלו.

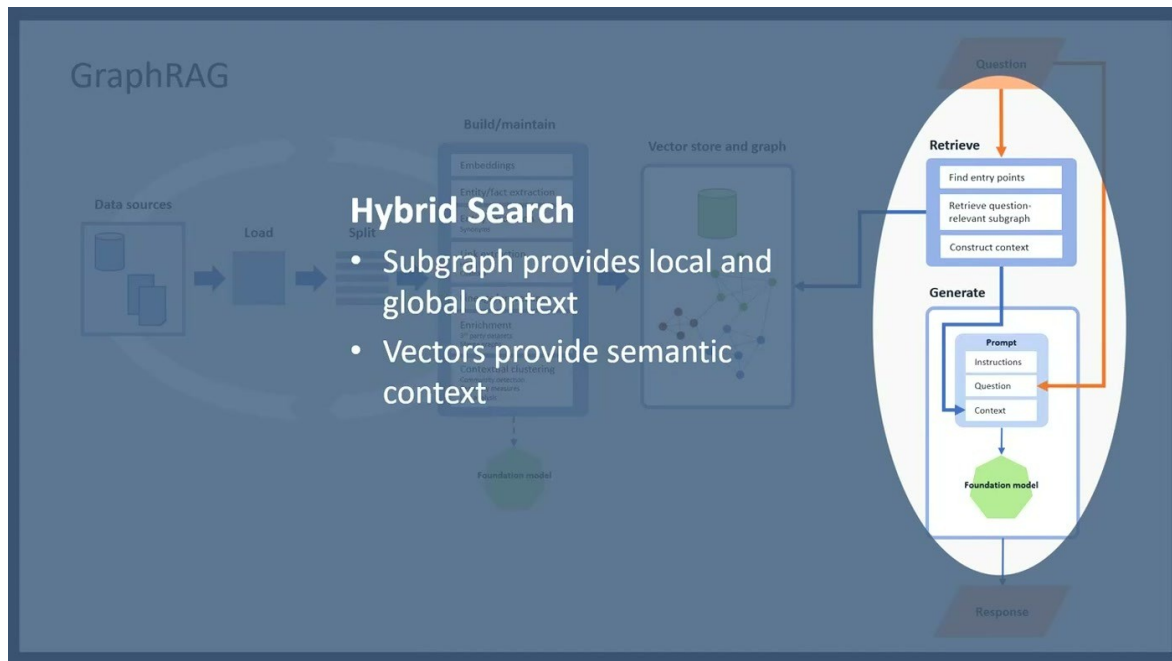
אם כל ה-statements הגיעו מאותה כתבה, זו local connectivity — "שאני עדיין עובד מול אותו מסמך מקור שאיתו התחלתי את התהליך" [27:57]. אבל למסקנה רחבה יותר צריך כמה מסמכים:

— **Erez Drutin, [27:57]** אני אקח כתבה מספין, אני אקח כתבה מספורט חמש וכן הלאה, וכשאני אגיע לסטייטמנטים דומים מכל הכתבות השונות האלו אני אוכל בעצם לקבוע איזושהי עובדה. לדבר הזה קוראים גלובל קונקטיביטי.

שימו לב ההבחנה הזו היא לב העניין: fact בשכבת ה-summarization אינו טענה של מסמך בודד אלא טענה שאוששה על פני כמה מסמכי מקור — ובכל זאת שומרת קשר חזרה לכל אחד מהם. זה מה שמייצר תשובה גם מגובה וגם ניתנת לייחוס.

11. חיפוש היברידי והתמונה המלאה

כשכל השכבות במקומן, שלב ה-retrieval יכול לשלוב בין השניים.



שלב ה-Retrieve: תת-גרף מספק הקשר לוקלי וגלובלי, והווקטורים מספקים את ההקשר הסמנטי — ושניהם נכנסים יחד ל-prompt

"חיפוש היברידי בעצם מאפשר את החיפוש גם לפי המרכיב הגרפי שלנו, שכמו שאמרנו החיפוש הלוקלי באמצעות הטופיקים והחיפוש הגלובלי עם אפקטים, וחיפוש ספקטורי" [29:28]. הזרימה השלמה: מקורות דאטה ← טעינה ← chunking ← בניית הגרף על שכבותיו ← vector store וגרף ← retrieval היברידי ← prompt ← תשובה.

נקודה ארכיטקטונית שהודגשה במפורש: אין חובה לאחד את שני האחסונים.

— Erez Drutin, [30:59] כן חשוב לי לציין שגם החלק של הגרף והחלק של הווקטורים לא חייבים לשאלת באותו דאטאבייס. יכול להיות שיהיה לכם דאטאבייס שבו אתם תשמרו את הרכיבים הווקטורים שלכם, וחלק שבו יש לכם את הגרף שתשמרו נניח בנפטון.

המסר הסופי היה מכוון נגד ההתלהבות מהחדש: "אנתולוגיות וסחאלג' גרף זה משהו חדש שמתחיל לצבור תאוצה. הוא לא באמת חדש, התחום הזה קיים כבר הרבה מאוד שנים, ולאחרונה זה מתחיל לצבור תאוצה בגלל כל ההתפתחות של האג'נטים" [30:59]. ושלוש זוויות שימוש הוצעו: שכבה אפליקטיבית-תפעולית (כמו אצל GrowersTech), "company brain" ארגוני, ודוקומנטציה עבור הסוכנים עצמם.

12. מה לוקחים מכאן

הסשן נסגר בשלושה מצביעים קונקרטיים, כולם בקוד פתוח או שירות מנוהל.

Call for action

GraphRAG Toolkit



Amazon Neptune



✦ Context Ontology Accelerator



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

שלושת ה-QR שהוצגו בסיום: ה-toolkit, שירות הגרף המנוהל, והמאיץ לבניית אונטולוגיה

- **GraphRAG Toolkit**: "פריימורק אופן סורסי שאתם יכולים להשתמש בו כבר עכשיו כדי להתחיל לפתח את הגרף רג שלכם" [32:30].
- **Context Ontology Accelerator**: "גם הוא אופן סורסי, פריימורק חדש שעוצנו לפני כמה חודשים, שעוזר לכם לייצר טאונטולוגיה לטובת הגרף שלכם" [32:30].
- **Amazon Neptune**: "השירות שאנחנו מציעים פה ב-AWS לטובת שמירה של מידה גרפי" [32:30].

ארבע מסקנות מעשיות

- **התחילו מהאונטולוגיה, לא מהמודל**. סדר השלבים שהוצג הוא מפורש ולא מקרי — מידול הדומיין, קידוד כגרף, ורק אז שכבת ה-AI. ארגונים שמתחילים מהסוף מקבלים סוכן שלא מכיר את עולם המושגים שלהם.
- **השתמשו בעומק השאילתה כמבחן החלטה**. אם השאלות העסקיות שלכם עוברות בעקביות יותר מארבע-חמש ישויות, זה האינדיקטור המעשי לכך שגרף אינו אופציה אלא הכרח.
- **היברידי, לא או-או**. אף צד בסשן לא הציע להחליף את ה-vector store. הטענה היא recall גבוה יותר בלי פגיעה ב-precision, ושני האחסונים יכולים לחיות בנפרד.
- **ה-lineage הוא התוצר, לא רק התשובה**. מבנה שלוש השכבות שומר קשר מכל fact אל מסמכי המקור. בסביבה מפוקחת זה מה שהופך תשובה של LLM למשהו שאפשר להציג בביקורת.

מה לא כוסה בסשן

הוגן לציין מה נשאר פתוח: לא הוצג דמו חי, לא הוצג קוד, ולא נדונו עלויות, תחזוקת האונטולוגיה לאורך זמן, או תהליך ה-entity resolution ברמת פרטים. השאלה הקשה מכולן בפרויקטים כאלה — כמה מאמץ אנושי נדרש לבנות ולתחזק אלפי אונטולוגיות — הוזכרה כעובדה ("אלפי אונטולוגיות") אך לא פורקה.

מילון מונחים

מונח	פירוש כפי שהוצג בסשן
Node / Edge	היחידות הבסיסיות של גרף — node מחזיק את הישות והמידע, edge מחזיק את הקשר בין ישויות
Knowledge Graph	גרף שבו ה-nodes מייצגים ישויות בעולם האמיתי וה-edges את הקשרים ביניהן; הוצג כצורה מודרנית של Semantic Web
אונטולוגיה	מודל לוגי שמאגד את הסמנטיקה של התחום — אובייקטים, properties, relationships ו-constraints
סמנטיקה	ההגדרות שמאפשרות לתוכנה לפרש את הכוונה שמאחורי נתון (דוגמת 32 המעלות)
Neuro-Symbolic AI	שילוב של רשת נוירונים או LLM (הצד הנוירוני) הפועלת מעל גרף ידע (הצד הסימבולי)
GraphRAG	RAG שבו ה-retrieval מתבצע גם מעל גרף ידע ולא רק מעל vector store
Hierarchical Lexical Graph	דפוס בניית גרף מטקסט בשלוש שכבות: Lineage, Summarization, Entity-Relationship
Local / Global connectivity	קישוריות בתוך מסמך מקור בודד (טופיקים) לעומת קישוריות על פני מספר מסמכים (facts)
Text-to-Cypher	המרה ישירה של שאלה בשפה טבעית לשאילתת גרף, כמסלול חלופי ל-GraphRAG מלא
Hybrid Search	שילוב חיפוש וקטורי (דמיון סמנטי) עם שליפת תת-גרף (רלוונטיות מבנית)
Amazon Neptune	שירות בסיס הנתונים המנוהל של AWS לאחסון מידע גרפי, ולפי הדובר גם וקטורים
AXIOM	מערכת ה-ingestion והנרמול של GrowersTech, שמטמיעה את המידע לגרף ומבצעת מעליו reasoning