הכנה לפגישה עם דר' רמי — שאלות ותשובות

*מסמך הכנה אישי. דר' רמי הוא **סטטיסטיקאי** (לכן התשובות הסטטיסטיות מדויקות ונוקבות), אבל **לא מכיר ברידג'** (לכן כל מונח ברידג' מוסבר בפשטות). לכל שאלה: תשובה קצרה, ואם צריך — דוגמה או הערה סטטיסטית.*

מה עשינו עד כה — בקצרה

- **שאלת המחקר:** האם סגנון קבלת-ההחלטות של שחקן ברידג' מנבא את התנהגותו במשא ומתן עסקי? הברידג' הוא *מעבדה*, היעד הוא מו"מ עסקי.
- **הנתונים:** 149,208 שורות מתחרויות ברידג' אירופאיות (2016–2025); אחרי סינון נשארו **563 שחקנים** עם מספיק נתונים (≥50 ידיים).
- **שלב 1 — גילוי פרופילים:** חישבנו 8 תכונות התנהגות לכל שחקן. ניסינו K-Means לחלק לקבוצות — **נכשל** (הנתונים הם רצף, לא אשכולות). במקום זה השתמשנו ב**אחוזון-קיצון + מבחן בינומי** וקיבלנו 4 פרופילים + "כללי".
- **שלב 2 — חילוץ כישורים:** מודל שפה (LLM) קרא את משחקי כל פרופיל והפיק 5–7 כישורים מאפיינים.
- **שלב 3 — בניית סוכנים:** כל פרופיל הפך לסוכן LLM עם "כרטיס אופי" (הכישורים מהברידג', לא תיוג שהמצאנו).
- **שלב 4 — סימולציה כפולה:** כל סוכן שיחק גם *ברידג'* וגם *משא ומתן עסקי*; השווינו את דירוג המנצחים בשני התחומים עם **Spearman ρ**.
- **התוצאה:** מדד ברידג' ראשוני נתן ρ=0.20 (חלש) בגלל חריג אחד; אחרי תיקון המדד → **ρ=0.80** (מעל היעד 0.70). היסוד מובהק (Cohen's d 2.1–4.6).
- **הסתייגות מרכזית:** n=5 פרופילים → כוח סטטיסטי נמוך. ρ הוא **אינדיקציה**, והתרומה האמיתית היא *מתודולוגית*.

🃏 ברידג' ב-90 שניות (כדי שתוכלי לענות לו על הבסיס)

- **המשחק:** 4 שחקנים, 2 זוגות (אתה ושותפך מול 2 יריבים). כל אחד מקבל 13 קלפים.
- **המכרז (bidding):** לפני המשחק יש "מכרז" — כל זוג מבטיח כמה לקיחות (tricks) ייקח. ההבטחה המנצחת נקראת **חוזה (contract)**.
- **רמות 1–7:** חוזה ברמה N = הבטחה לקחת 6+N לקיחות. רמה גבוהה = הבטחה שאפתנית ומסוכנת.

- **מונחי המפתח שלנו:** *סלאם* = חוזה רמה 6–7 (שאפתני, בונוס ענק, סיכון גבוה) · *part-score* = חוזה נמוך 1–3 (בטוח, שמרני) · *NT (NoTrump)* = משחק בלי "צבע מנצח" · *דאבל / הכפלה* = הימור שמכפיל את הנקודות; *הכפלת עונשין* = "אני בטוח שהיריב ייכשל בחוזה שלו".
- **HCP (כוח יד):** נקודות לפי קלפים גבוהים (אס=4, מלך=3, מלכה=2, נסיך=1). מודד כמה היד "חזקה".
- **פאר (par):** החוזה האופטימלי תאורטית לחלוקת קלפים נתונה — "מה היה *נכון* להכריז".

חלק א' — תכונות, נתונים, ויחידת הניתוח

ש: אילו פיצ'רים (תכונות) יש לך? (השאלה ששאל בפעם הקודמת)

ת: 10 תכונות התנהגות לכל שחקן, ב-2 משפחות; 8 מהן נכנסו למודל. כל תכונה היא **שיעור** (יחס בין 0 ל-1).

מתוך "מה הוכרז" (חוזה)	מתוך "איך הוכרז" (מכרז)
`slam_rate` — שיעור חוזי סלאם	`opening_rate` — שיעור פתיחת מכרז
`nt_rate` — שיעור חוזי NT	`preempt_rate` — פתיחה גבוהה (אגרסיבי)
`partscore_rate` — שיעור עצירה נמוך/בטוח	`intervention_rate` — התערבות במכרז היריב
`game_rate` — שיעור חוזי "משחק"	`penalty_double_rate` — הכפלות עונשין
`double_rate` — שיעור חוזים שהוכפלו	`avg_bids_per_board` — מס' הכרזות לחלוקה

ש: מהי יחידת הניתוח? מהו n?

ת: יחידת הניתוח היא **שחקן בודד**. **563** = n **שחקנים** שעברו סף כשירות (≥50 ידיים שהכריזו עליהן וגם ≥50 ידיים עם נתוני מכרז). כל שחקן הוא וקטור של 8 מספרים.

ש: איך בחרת אילו תכונות להכניס? למה 8 ולא 10?

ת: הסרנו 2 תכונות עם **שונות נמוכה** — מקדם השונות (CV = סטיית תקן ÷ ממוצע) קטן מ-0.10. תכונה שכמעט זהה אצל כולם לא מבדילה בין שחקנים, אז היא רק מוסיפה רעש.

> **דוגמה:** `avg_level` — כל שחקני העילית מכריזים בממוצע סביב רמה 3.3. אם לכולם אותו ערך, התכונה חסרת תועלת לסיווג, כמו לשאול "למי יש שתי עיניים?".

ש: האם התכונות מתואמות ביניהן? יש מולטי-קולינאריות?

ת: בדקנו מטריצת מתאמים והסרנו כפילויות. בנוסף, לפני האשכול הרצנו **PCA** (3 רכיבים) שמייצר צירים אורתוגונליים — כך שגם אם נשאר מתאם שיורי, האשכול עצמו עבד על מרחב לא-מתואם.

> **הערה סטטיסטית:** זו גם הסיבה שאנחנו מדווחים על PCA(3) כתצורה הטובה ביותר מתוך 5 שנבדקו — היא מנטרלת קורלציה ומפחיתה ממדיות לפני K-Means.

חלק ב' — למה K-Means נכשל ומהו "רצף"

ש: למה K-Means נכשל? (השאלה ששאל בפעם הקודמת)

ת: K-Means מחפש **קבוצות נפרדות** ("איים"). בנתונים שלנו אין איים — יש ענן רציף אחד. מדד ה**סילואט** (silhouette) הגיע לכל היותר ל-**0.24**, בעוד שמבנה אשכולות אמיתי דורש ≥ 0.5. ניסינו k=2 עד 6 — כולם נכשלו.

> **הערה סטטיסטית:** סילואט מודד לכל נקודה כמה היא קרובה לאשכול שלה לעומת האשכול השכן: s = (b−a)/max(a,b). ערך סביב 0 = הנקודה על הגבול בין אשכולות = אין הפרדה אמיתית.

ש: מה זה "רצף" (continuum)? (השאלה ששאל בפעם הקודמת)

ת: רצף = הנתונים משתנים בהדרגה חלקה, בלי קפיצות או פערים שמפרידים קבוצות. אין "סוגים", יש *ספקטרום*.

> **דוגמה פשוטה: גובה** של אנשים הוא רצף — אין שתי קבוצות נפרדות "נמוכים" ו"גבוהים" עם פער ביניהן; יש מעבר חלק מ-150 ל-200 ס"מ. לעומת זאת **סוג דם** (A/B/AB/O) הוא קטגוריות נפרדות = אשכולות אמיתיים. שחקני הברידג' שלנו הם כמו גובה, לא כמו סוג דם.

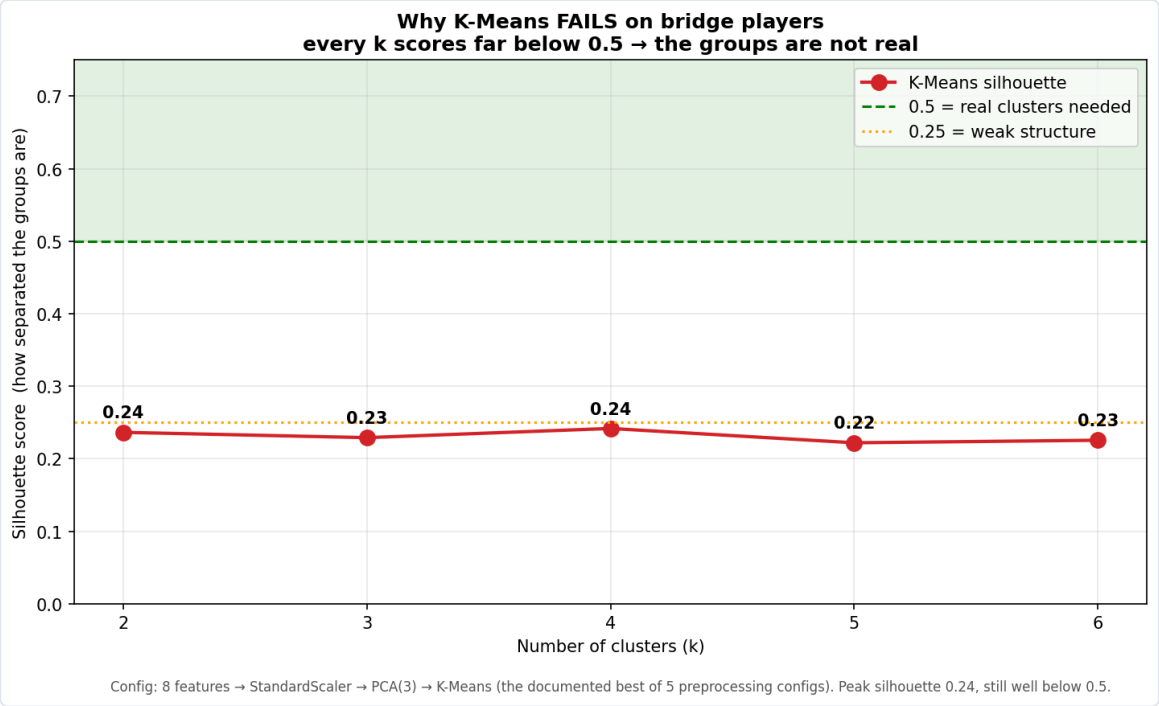
ש: איך את יודעת שזה באמת רצף, ולא שפשוט בחרת k לא נכון?

ת: 3 ראיות מתכנסות: (1) **כל** k מ-2 עד 6 נתן סילואט ≥ 0.24; (2) ניסינו גם **HDBSCAN** (שלא מניח מספר אשכולות מראש) **וגם GMM** — שניהם נכשלו; (3) ויזואלית, מפת ה-PCA מראה ענן יחיד בלי פערים. שלוש שיטות שונות שמסכימות = לא ארטיפקט של פרמטר בודד.

ש: אם K-Means נכשל — איך בכל זאת הגדרת פרופילים?

ת: בשיטת **אחוזון-קיצון (extreme-percentile)**: במקום לחתוך את הענן לקבוצות מזויפות, לקחנו את ה-**10% העליונים** בכל ציר התנהגות, ואישרנו כל קבוצה במבחן מובהקות בינומי.

> **דוגמה:** "צייד סלאמים" = ה-10% עם שיעור הסלאם הגבוה ביותר. "לוחם" = ה-10% עם הכי הרבה הכפלות עונשין. זה לא ממיין את כולם בכוח — זה מנקד את הזנבות הקיצוניים של הרצף.



K-Means: כל k נותן סילואט ~0.24, הרחק מתחת ל-0.5 הנדרש. אין אשכולות אמיתיים.

ש: המבחן הבינומי — בשביל מה הוא, ולמה $p<0.05$?

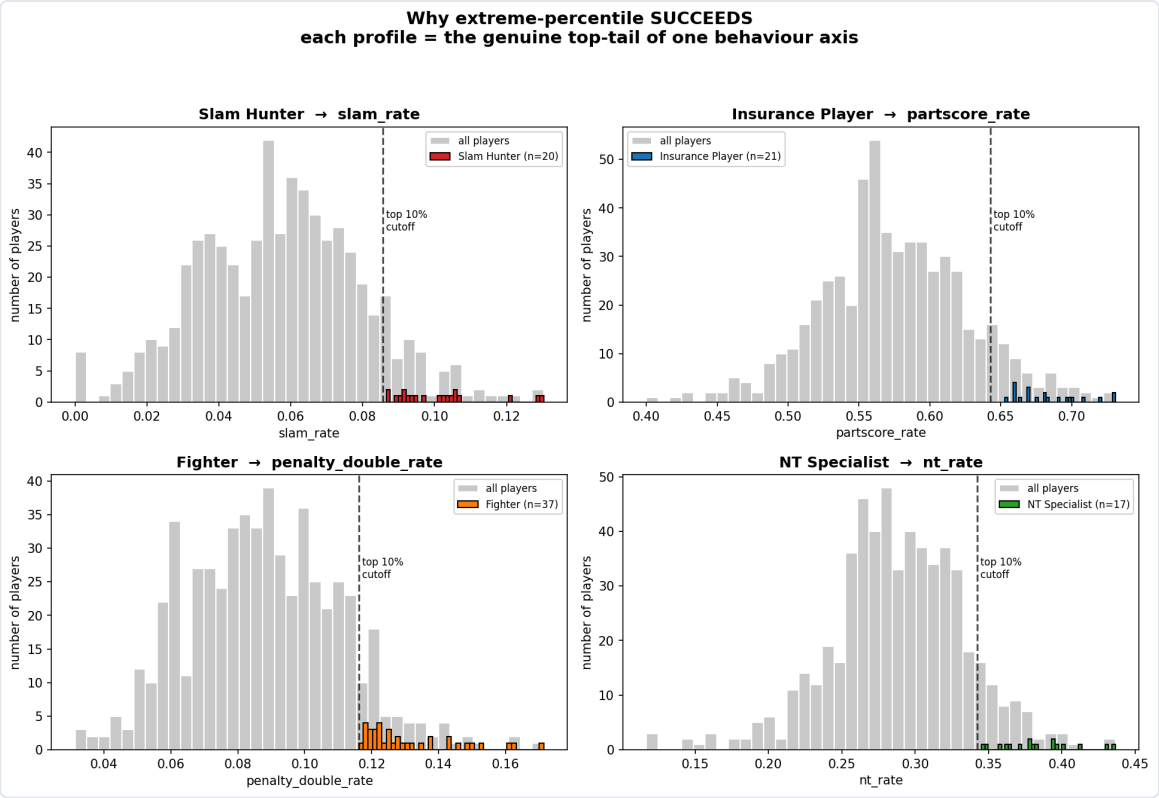
ת: הוא מאשר שהציון הקיצוני של שחקן הוא **אמיתי ולא מקריות של מדגם קטן**. השאלה: "בהינתן השיעור הבסיסי באוכלוסייה, מה ההסתברות שהשחקן הזה הגיע לשיעור כה גבוה במקרה?". אם $p<0.05$ — לא סביר שזה מקרה, אז הוא באמת שייך לפרופיל.

> **דוגמה:** אם השיעור הבסיסי של סלאם הוא 5.5%, ושחקן עשה סלאם ב-10% מתוך 200 ידיים — המבחן הבינומי שואל אם 10% מתוך 200 חורג מובהקית מ-5.5%. כן → הוא "צייד סלאמים" אמיתי.

ש: בדקת הרבה שחקנים והרבה צירים — מה עם בעיית ההשוואות המרובות?

ת: שאלה מצוינת וצפויה ממנו. כן, זה רלוונטי: 563 שחקנים × כמה צירים = הרבה מבחנים. הדרכים שלנו: (1) סף שמרני (10% עליונים *וגם* $p<0.05$, לא רק אחד מהם); (2) גדלי האפקט הם **עצומים** (Cohen's d 2.1–4.6) — הם שורדים כל תיקון סביר; (3) נשמח ליישם תיקון **FDR (Benjamini-Hochberg)** אם הוא ימליץ — זה לא ישנה את הפרופילים בגלל גודל האפקט.

> **המסר לסטטיסטיקאי:** "אני מודעת לבעיית ה-multiple comparisons; האפקטים כל כך גדולים שהם עמידים, ואני פתוחה ליישם FDR."



כל פרופיל = הזנב הימני (10% העליונים) של ציר התנהגות אחד. אפור=כלל האוכלוסייה, צבע=הפרופיל.

חלק ג' — ולידציה של הפרופילים

ש: מה זה Cohen's d ולמה הערכים 2.1–4.6 חשובים?

ת: Cohen's d מודד **גודל אפקט** — בכמה סטיות-תקן נבדל הפרופיל מהממוצע. d=0.8 כבר נחשב "גדול". הערכים שלנו **2.1–4.6** הם עצומים — הפרדה כמעט מלאה בין הפרופיל לשאר.

> **דוגמה:** d=2.1 פירושו שכמעט כל "צייד סלאם" נמצא מעל כמעט כל שחקן רגיל בציר הסלאם. החפיפה זעירה.

ש: גדלי הקבוצות קטנים (20, 21, 37...) — זו בעיה?

ת: לוולידציה של שלב 1 — לא, כי גודל האפקט עצום וה-p מובהק גם ב-n קטן. הקבוצות: צייד סלאם 20, ביטוח 21, לוחם 37, מומחה NT 17, והשאר (הרוב) "כללי", מתוך 563.

> **הערה:** ה-n הקטן *כן* הופך לבעיה בשלב 4, כששאלת היחידה היא הפרופיל (n=5). שם נדבר על כוח סטטיסטי.

חלק ד' — המשא ומתן, התהליך, והיישור (שלב 4)

ש: איך עובד המשא ומתן? מי מנהל אותו?

ת: כל פרופיל (סוכן LLM) משחק **קונה** ב-4 תרחישים עסקיים, מול **מוכר קבוע שאינו LLM** — המוכר הוא נוסחה: מוותר בהדרגה לכיוון "רצפה" ולא מתחת לה. כך ההבדלים שמדדנו נובעים מהקונה (הפרופיל), לא מרעש של מוכר משתנה.

> **דוגמה לתרחיש:** רכישת חברת SaaS — מחיר הוגן 9M, המוכר פותח ב-13M, רצפה 8M. הקונה מנסה להוריד מחיר ככל האפשר עד 4 סבבים.

ש: איך מוגדר "ניצחון" בכל תחום?

ת: בברידג': כמה קרוב השחקן הכריז ל"חוזה הנכון" (par) לפי כוח היד. **במו"מ:** איזה אחוז מהפער מחיר תפס הקונה — (פתיחה − רצפה) / (פתיחה − מחיר סופי). בשני המקרים: מספר בין 0 ל-1, ממוצע על כל הידיים/התרחישים.

ש: מה זה ρ Spearman ולמה לא Pearson?

ת: ρ Spearman הוא מתאם על **דירוגים** (מי ראשון, שני...), לא על הערכים עצמם. בחרנו בו כי השאלה שלנו היא על **סדר**: "האם אותו פרופיל שמנצח בברידג' מנצח גם במו"מ?". הוא גם עמיד לחריגים ולא מניח קשר ליניארי.

> **דוגמה:** אם הדירוג בברידג' הוא A>B>C והדירוג במו"מ גם ρ=1 — A>B>C, גם אם הערכים המוחלטים שונים לגמרי.

ש: n=5 פרופילים — מה עם כוח סטטיסטי? ה-p לא מובהק!

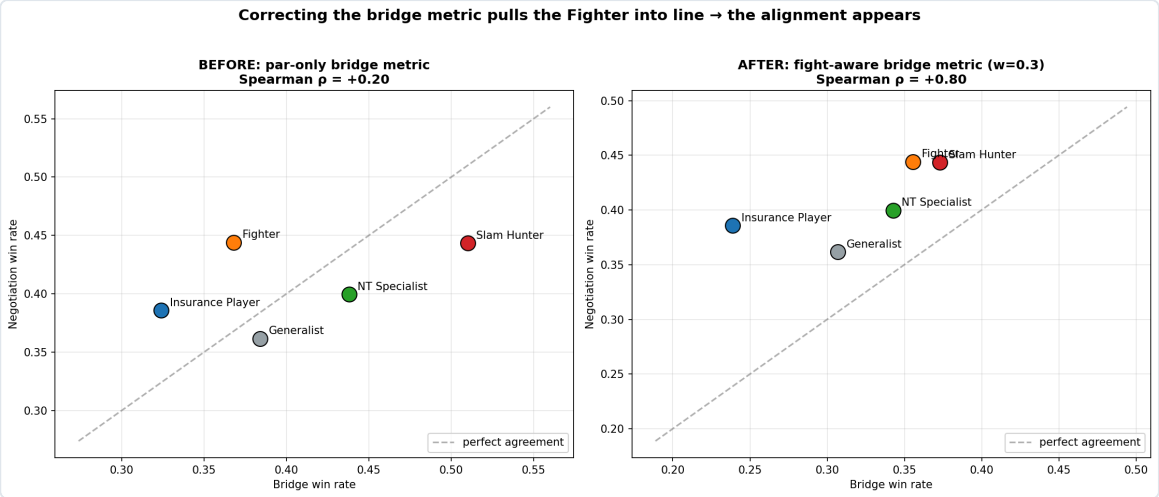
ת: נכון, וזו ההסתייגות המרכזית שלנו, שאנחנו **מציינים מפורשות**. עם n=5, גם ρ=0.80 נותן p≈0.10. לכן אנחנו אומרים ש-ρ הוא **אינדיקציה ולא הוכחה**, והתרומה האמיתית היא **מתודולוגית** (ראי השאלה הבאה). זה proof-of-concept, והדוקטורט ירחיב את ה-n.

> **המסר:** "אני לא טוענת למובהקות; אני טוענת לעיקרון מתודולוגי שהודגם, עם תוכנית להגדיל את ה-n."

ש: שינית את המדד אחרי שראית את התוצאה — זה לא p-hacking?

ת: השאלה הכי חשובה שלו, ויש לנו תשובה חזקה. **לא**, מ-3 סיבות: (1) ה"לוחם" **מוגדר** ע"י הכפלות כבר משלב 1 — המדד המתוקן פשוט מודד את הכישור המגדיר שלו עצמו; (2) **מומחה הברידג' חזה את הבעיה מראש**, לפני שראינו את (3) ;ρ בחרנו משקל **מתון (0.3)**, לא את זה שממקסם את ρ (0.4 היה נותן 0.90), והצגנו **ניתוח רגישות** מלא.

> **המסר לסטטיסטיקאי:** "התיקון הוא a-priori ומוגדר-תאורטית, לא חיפוש אחורה. דיווחתי משקל מתון + sensitivity analysis כדי להראות שזה לא garden-of-forking-paths."



לפני (par בלבד, ρ=0.20) מול אחרי (מדד מודע-לחימה, ρ=0.80). ה-Fighter (כתום) זז אל האלכסון.

ש: למה נשאר היפוך בין Insurance ל-Generalist?

ת: זה ההיפוך היחיד שמחזיק את ρ על 0.80 ולא 1.0, וזה **ממצא ולא רעש**: שחקן ה"ביטוח" (השמרן) חלש יותר בברידג' אבל חזק יותר במו"מ. **שמרנות פוגעת בברידג'** (מכריז נמוך מדי → דיוק נמוך) אבל **עוזרת במו"מ** (סבלנות, סגירת עסקה יציבה). אותה תכונה עוברת בכיוון הפוך בין התחומים.

ש: איך התמודדת עם זה שה-LLM לא דטרמיניסטי?

ת: גם ב-temperature=0 ה-LLM נותן תשובות שונות (ריצה מקבילית על GPU). הפתרון: כל יד רצה **3 פעמים** והוכרעה ב**הצבעת רוב**; כל פלט גולמי נשמר ל-JSONL. הרפרודוקטיביות היא ברמת ה**מדגם** (התפלגות על הרבה ידיים), לא ברמת הקריאה הבודדת.

ש: איך מנעת "תאום מלאכותי" — שה-LLM פשוט מציית לפרומפט?

ת: הזרקנו לסוכנים רק את הכישורים **שנגזרו מהברידג'** (כמו "הכפלת עונשין אגרסיבית"), אף פעם לא מילות אופי שהמצאנו (כמו "תהיה אגרסיבי"). יש גם **בקרת פרומפט הפוך** מתוכננת: אם פרופיל הפוך נותן יישור גבוה באותה מידה — סימן שזה ציות ולא העברה אמיתית.

חלק ה' — התמונה הגדולה והדוקטורט

ש: למה ברידג' בכלל? איך זה קשור לעסקים?

ת: נתוני משא ומתן אמיתיים **סודיים ולא נמדדים** — אי אפשר לאמן עליהם מודל. ברידג' הוא מעבדה נקייה לאותן החלטות (מתי לקחת סיכון, מתי לעצור, מתי ללחוץ), עם 149K תוצאות מדידות. אם הפרופיל מהברידג' מנבא מו"מ — קיבלנו דרך ללמוד אסטרטגיה עסקית בלי נתונים פרטיים.

ש: מה התרומה אם המספר לא מובהק?

ת: התרומה היא **מתודולוגית**: *מדד ההצלחה בכל תחום חייב לתפוס את הכישור הרלוונטי לאותו תחום.* מדדנו את הציר הלא-נכון בברידג' וזה הסתיר קשר אמיתי; ברגע שהמדד תפס את הלחימה — הקשר הופיע. זה עיקרון שמתקף את עצמו ולא תלוי ב-n.

ש: מה הצעדים הבאים לדוקטורט?

ת: (1) להגדיל את ה-n — יותר פרופילים ויותר תרחישים, כדי להגיע למובהקות; (2) מדד ברידג' עשיר יותר או מנוע double-dummy אמיתי במקום קירוב הפאר; (3) תרחישי מו"מ אמיתיים יותר; (4) בקרת הפרומפט-ההפוך בפועל.

💡 **טיפ אחרון לפגישה:** אם הוא לוחץ על נקודה סטטיסטית — אל תתגוננו. תגידו "נקודה טובה, אני מודעת לזה, וזו בדיוק הסיבה ש-X". כל ההסתייגויות (n קטן, multiple comparisons, p-hacking) כבר **מתועדות מראש** במסמכים שלך — זה מראה בגרות מחקרית.

מסמך הכנה אישי לפגישה. מבוסס על: `features_and_hypothesis.md`, `RESEARCH_INSIGHTS.md`, `results/stage4/alignment_corrected_report.md`, ופרק התוצאות ב- `reports/main.tex`. נתונים: 149,208 שורות EuroBridge; 563 שחקנים; 8 תכונות; 5 פרופילים.