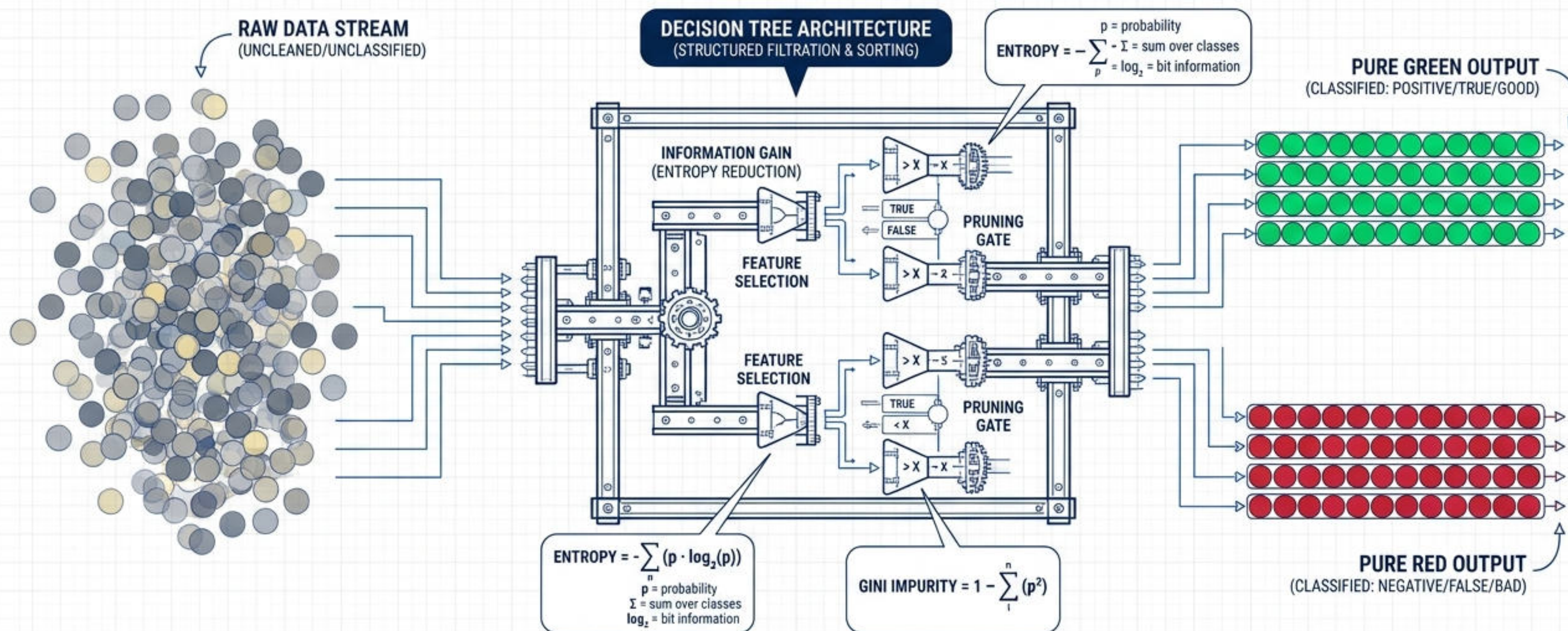


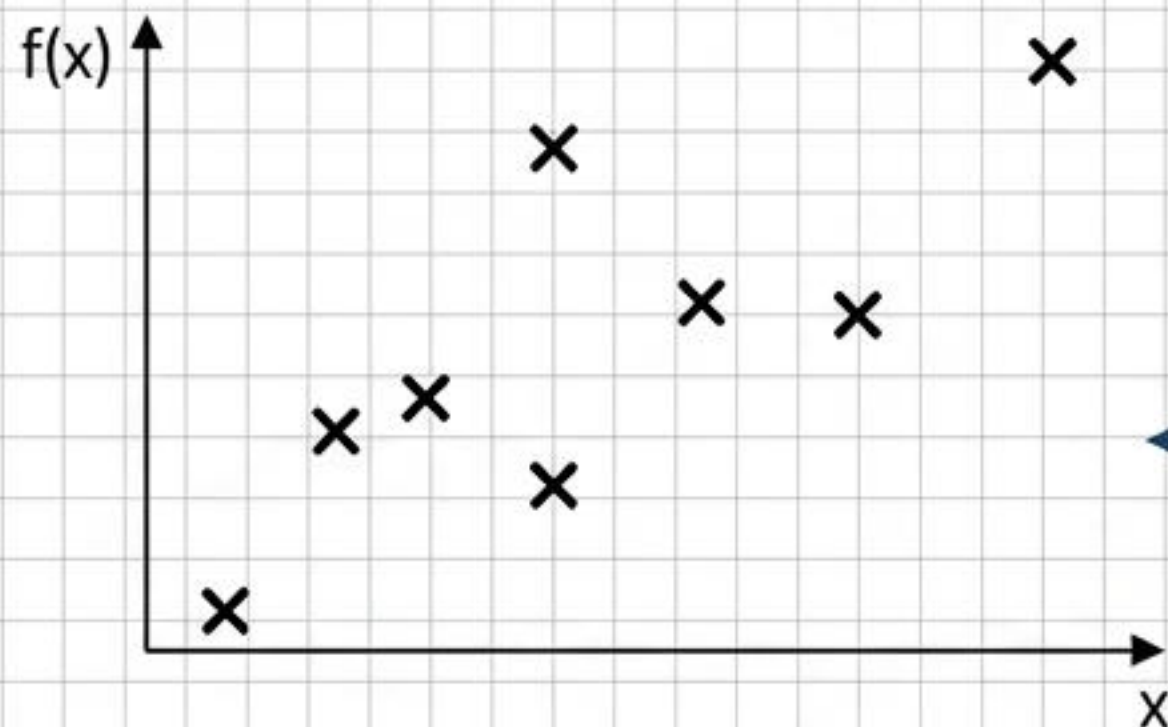
# מסרטוט נתונים להחלטה: הארכיטקטורה של עצי החלטה

מדריך ויזואלי לתורת האינפורמציה ולמידת מכונה (Supervised Learning)



# מטרת העל: לגלות את החוקיות הנסתרת בתוך הכאוס

## הבעיה



יש לנו אוסף של דוגמאות אימון (Set) Training), שנוצרו על ידי פונקציית מטרה שאינה ידועה לנו.

## הפתרון המבוקש

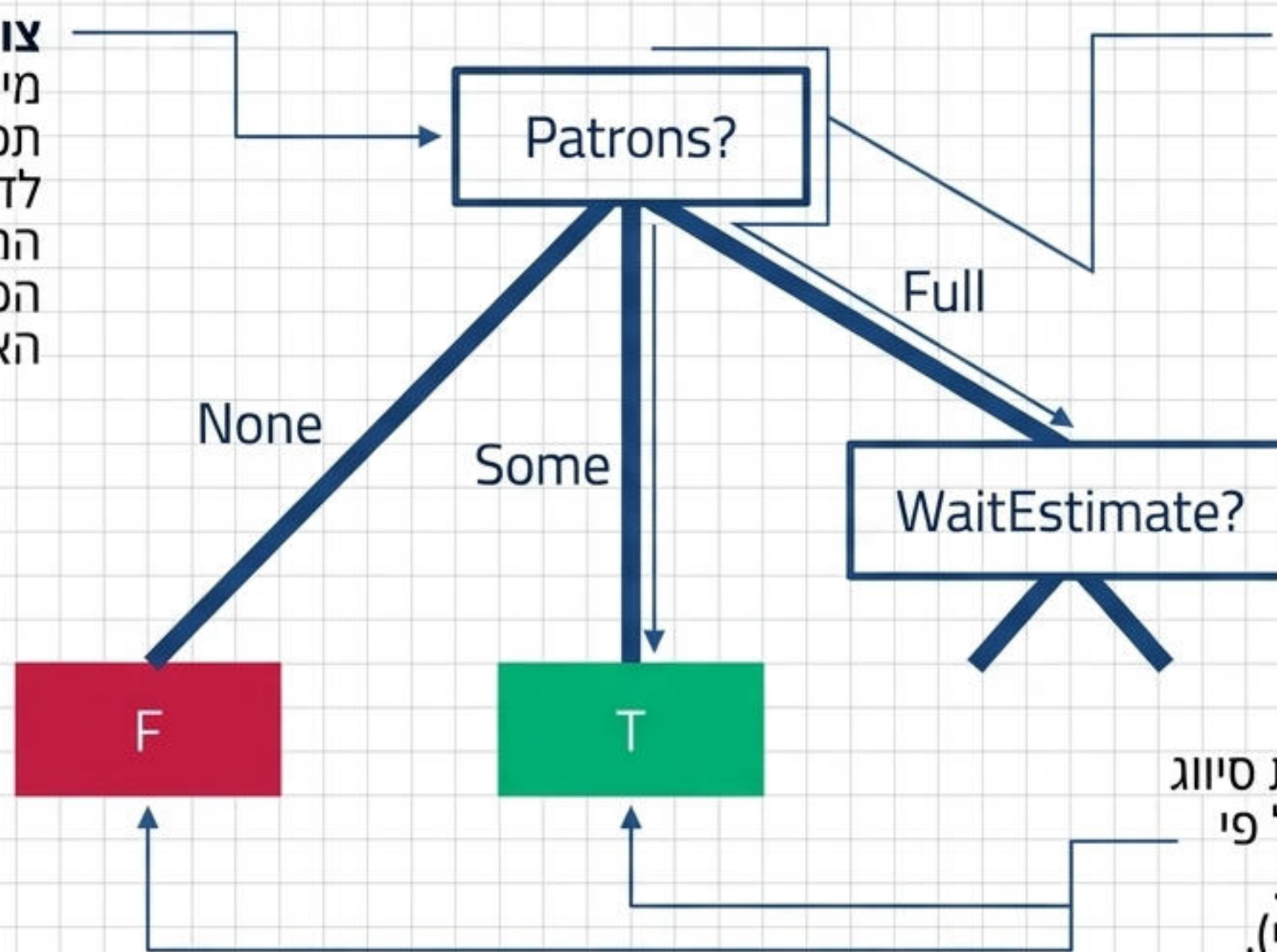


הבעיה: למצוא השערה (h) שתהיה קרובה ככל האפשר לפונקציית המטרה (f).

**תערו של אוקאם (Ockham's Razor):** נעדיף תמיד את ההשערה הפשוטה ביותר שתואמת לנתונים. עקרון זה יוביל את כל הדרך בה נבנה מודלים.

# הכלי שלנו: האנטומיה של עץ החלטה

**צומת פנימי (Internal Node):**  
מייצג שאלת בדיקה על  
תכונה מסוימת (Attribute).  
לדוגמה: "מה מספר  
החיות?" "מוחל" הלקוחות?".  
הפיצולים (ענפים) הם התשובות  
האפשריות.



**מסלול (Path):** מייצג כלל  
(Rule) ספציפי: צירוף של  
תנאים שמוביל להחלטה.

**עלה (Leaf Node):**  
התחנה הסופית. מכיל תווית סיווג  
(Class Label) שנקבעת על פי  
רוב הדוגמאות שהגיעו אליו.  
(ירוק = חיובי, אדום = שלילי).

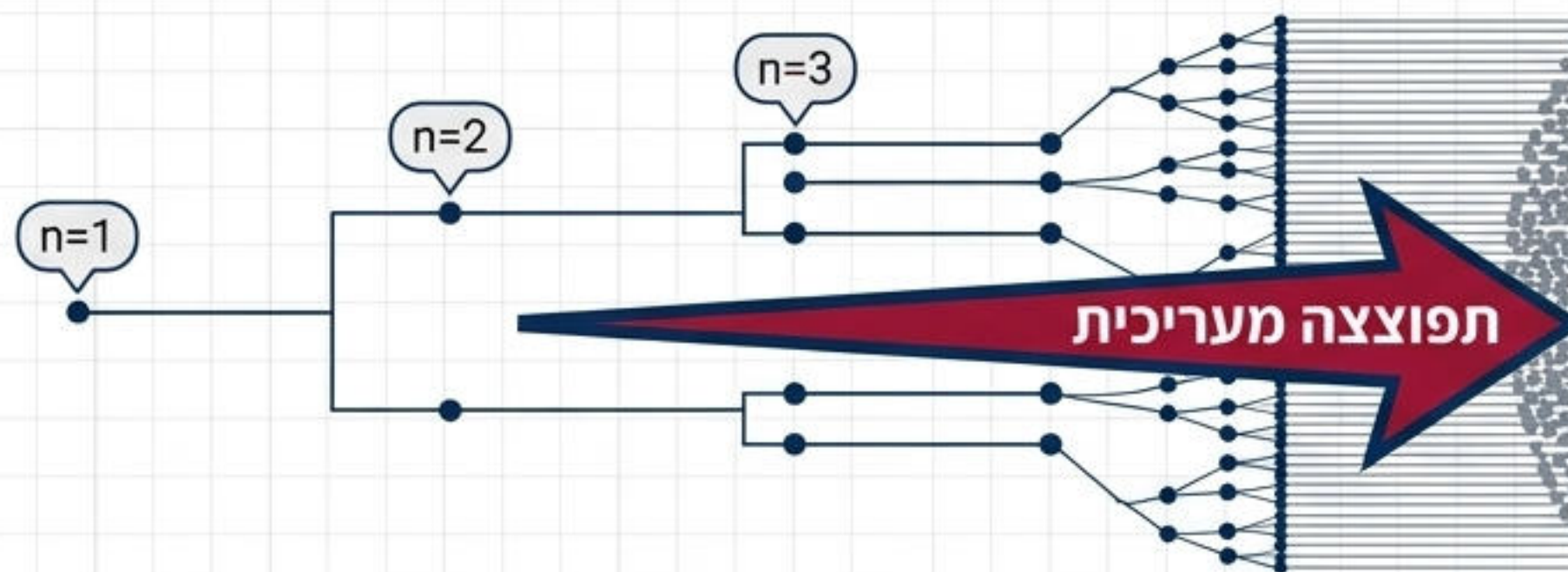
# מחסום החישוב: התפוצצות מרחב ההשערות

כמה עצי החלטה שונים ניתן לבנות עבור  $n$  תכונות בוליאניות?

מספר העצים = מספר הפונקציות הבוליאניות =  $2^{2^n}$

מספר הבנות

מספר הפונקציות



**עבור 6 תכונות בלבד**

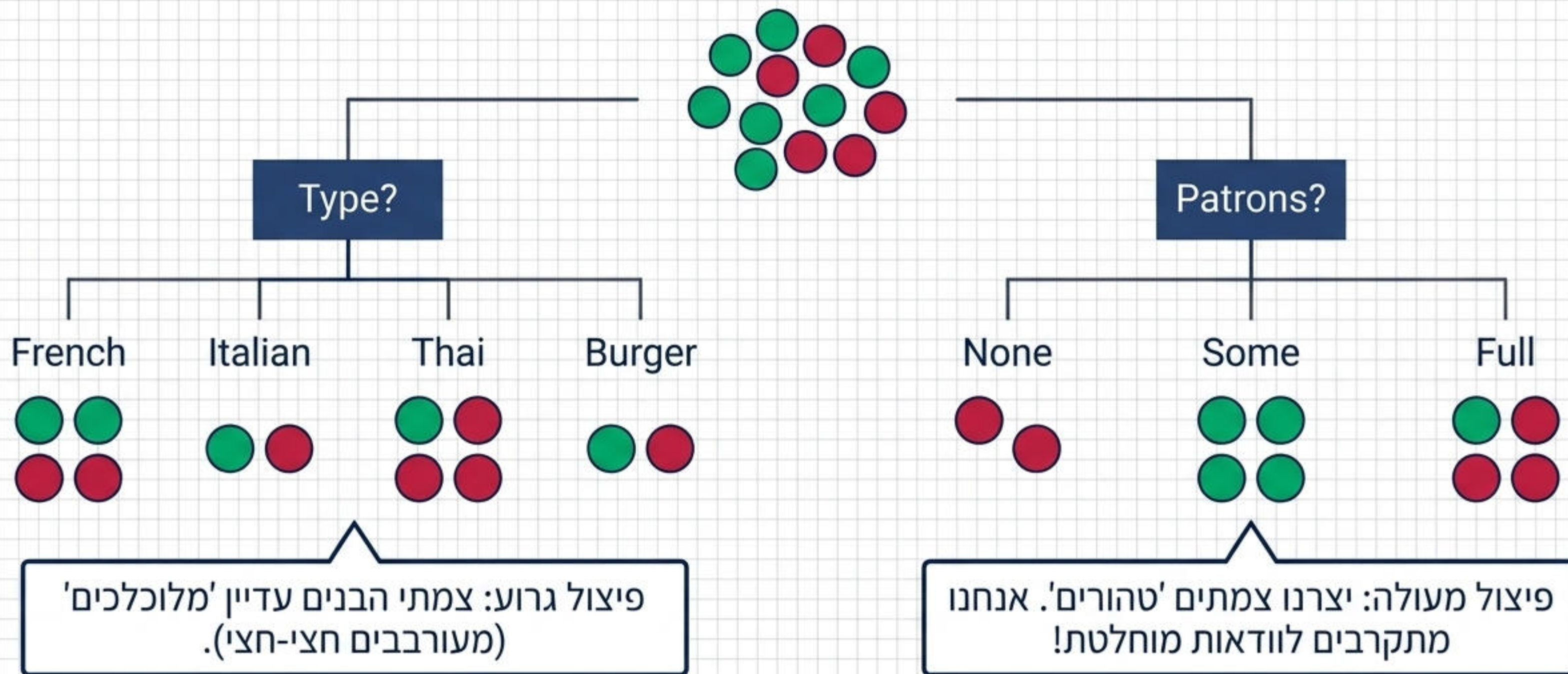
(לדוגמה: מזג אוויר, זמן, רעב...) ישנם:

**18,446,744,073,709,551,616**

**עצים אפשריים!**

מציאת העץ הקטן ביותר האפשרי היא בעיה בלתי פתירה מבחינה חישובית (NP-complete).  
אנו זקוקים להיוריסטיקה (Heuristic) חכמה שתבנה עץ 'מספיק טוב' צעד אחר צעד.

# המפתח להיוריסטיקה: חתירה ל'טוהר' (Purity) מלא



השאלה הטובה ביותר היא זו שמפצלת את הנתונים לתתי-קבוצות בהן כל הדוגמאות שייכות (כמעט) לאותו סיווג. איך נמדוד זאת מתמטית?

# המנוע המתמטי: אנטרופיה (H) כמדד לכאוס

$$H(Y) = - \sum p_i \log_2 p_i$$

מייצג את מספר שאלות ה'כולא' (ביטים) הנדרשות בממוצע כדי לזהות תוצאה אקראית.

סכום על פני כל הסיווגים האפשריים (למשל: כן/לא).

ההסתברות (השכיחות) של סיווג i בצומת הנוכחי.

מטבע הוגן



$p(\text{head})=0.5, p(\text{tail})=0.5.$   
 $H=1.$   
אי ודאות מקסימלית.

מטבע מוטה



$p(\text{head})=0.51.$   
 $H=0.9997.$

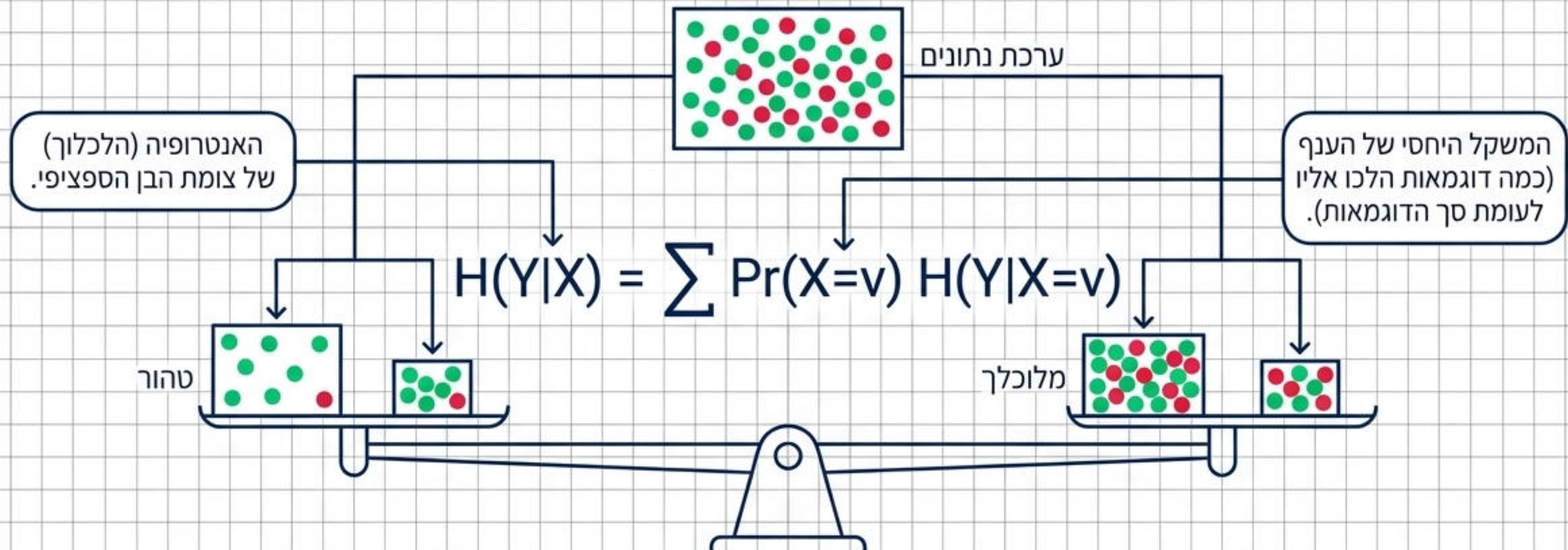
מטבע קבוע



$p=1. H=0.$   
ודאות מוחלטת (צומת טהור!).

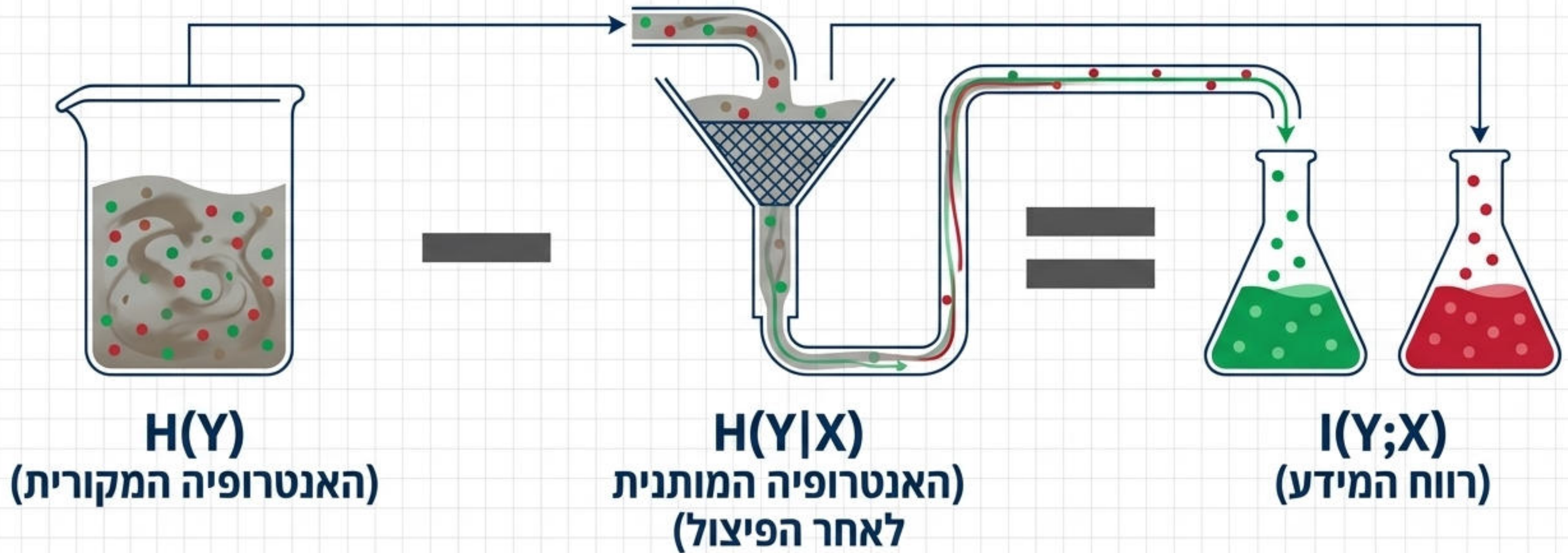
# אנטרופיה מותנית: מדידת הכאוס לאחר הפיצול

האנטרופיה המותנית,  $H(Y|X)$ , מודדת כמה 'כאוס' נשאר במערכת לאחר שפיצלנו אותה בעזרת תכונה  $X$ .



זהו ממוצע משוקלל של האנטרופיות של כל צמתי הבנים. אנו מחפשים תכונה  $X$  שתניב את האנטרופיה המותנית הנמוכה ביותר.

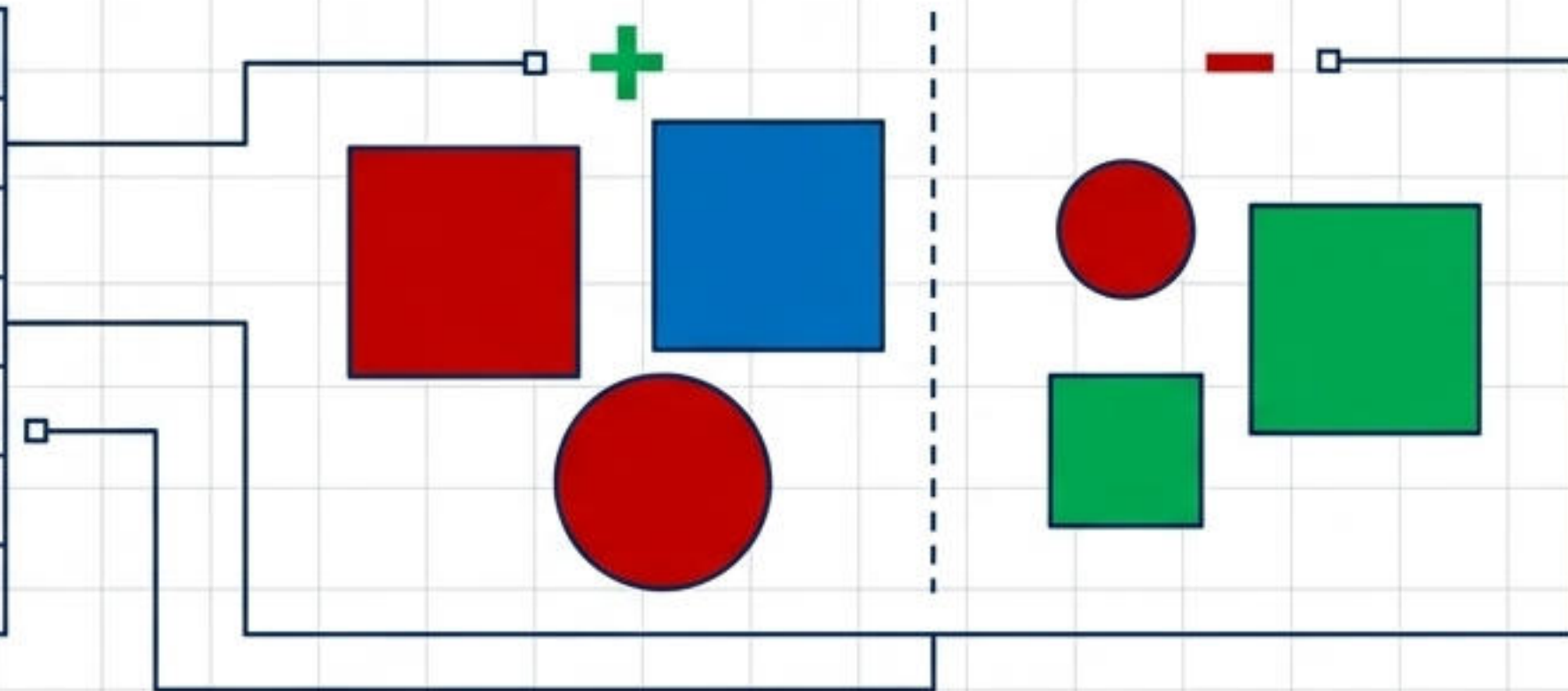
# הזיקוק הסופי: רווח מידע (Information Gain)



האלגוריתם תמיד יבחר בשאלה (תכונה)  $X$  שממקסמת את רווח המידע  $I(Y;X)$ .  
רווח מידע גבוה משמעותו ירידה דרסטית ברמת אי-הוודאות.

# מאלגורה לפרקטיקה: מעבדת הניסויים

| Example | Color | Shape  | Size  | Class |
|---------|-------|--------|-------|-------|
| 1       | Red   | Square | Big   | +     |
| 2       | Blue  | Square | Big   | +     |
| 3       | Red   | Circle | Big   | +     |
| 4       | Red   | Circle | Small | -     |
| 5       | Green | Square | Small | -     |
| 6       | Green | Square | Big   | -     |



## Step 1: Calculate Global Entropy

• יש לנו 3 תוויות '+' ו-3 תוויות '-'.

$$H(Class) = -(3/6 \log_2(3/6)) - (3/6 \log_2(3/6)) = 1$$

תרומת 3 תוויות '+'  
(חצי מהנתונים)

תרומת 3 תוויות '-'  
(חצי מהנתונים)

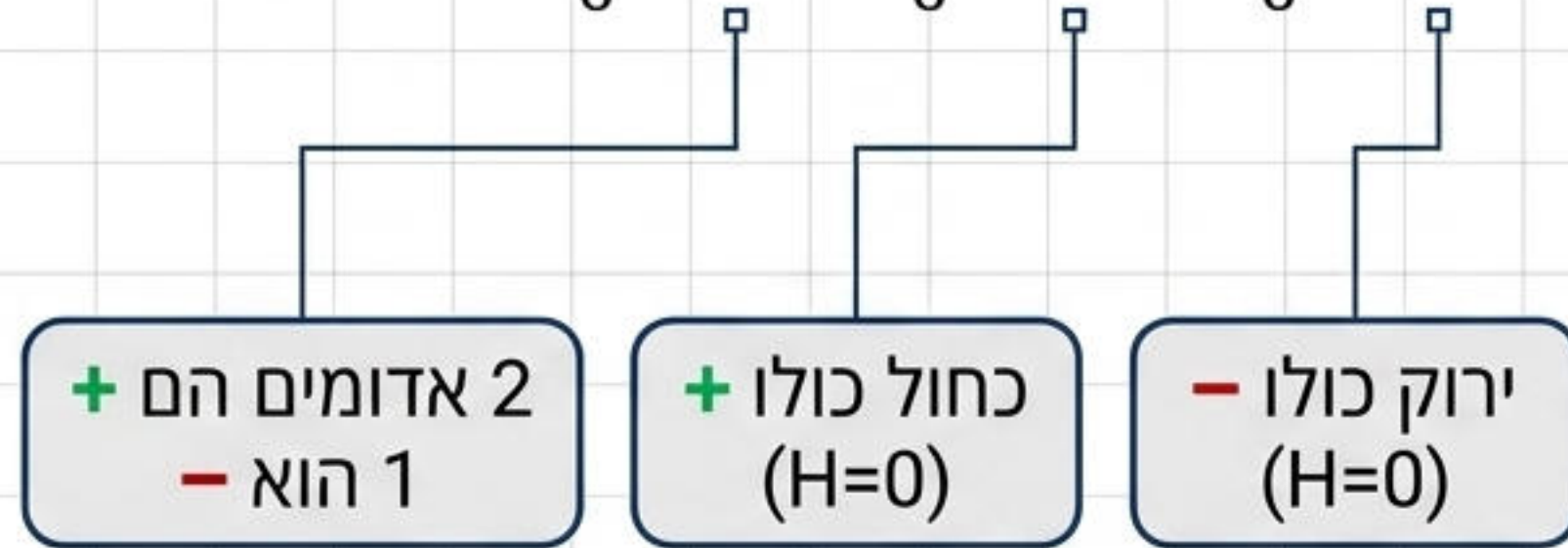
כאוס מקסימלי  
(אנטרופיה 1)

המערכת כרגע בכאוס מוחלט. משימתנו: למצוא את השאלה לשורש העץ שתעשה את הסדר הטוב ביותר.

# התחרות על השורש: צבע מול צורה

## Testing COLOR

$$H(Class | Color) = \frac{3}{6}H(\text{Red}) + \frac{1}{6}H(\text{Blue}) + \frac{2}{6}H(\text{Green})$$



$$I(Class; Color) = 1 - 0.46 = 0.54 \text{ bits}$$

## Testing SHAPE

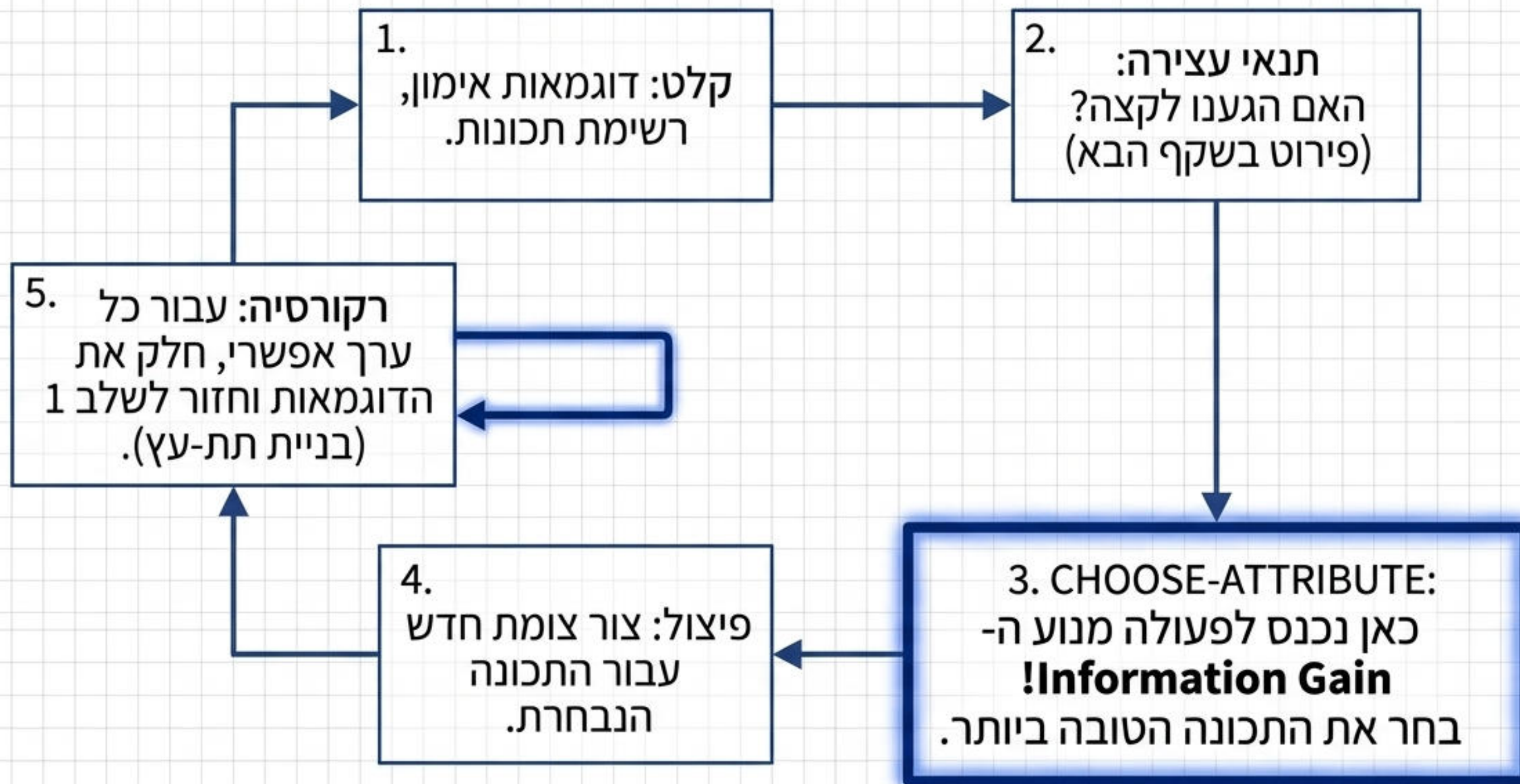
$$H(Class | Shape) = \frac{4}{6}H(\text{Square}) + \frac{2}{6}H(\text{Circle})$$



$$I(Class; Shape) = 1 - 1 = 0 \text{ bits}$$

הצבע מנצח! נהפיוית כצומת השורש,  
הצבע מנצח! נבחר בו כצומת השורש, כיוון שהוא מספק את רווח המידע המקסימלי.

# המכונה בפעולה: אלגוריתם למידת עץ החלטה (DTL)

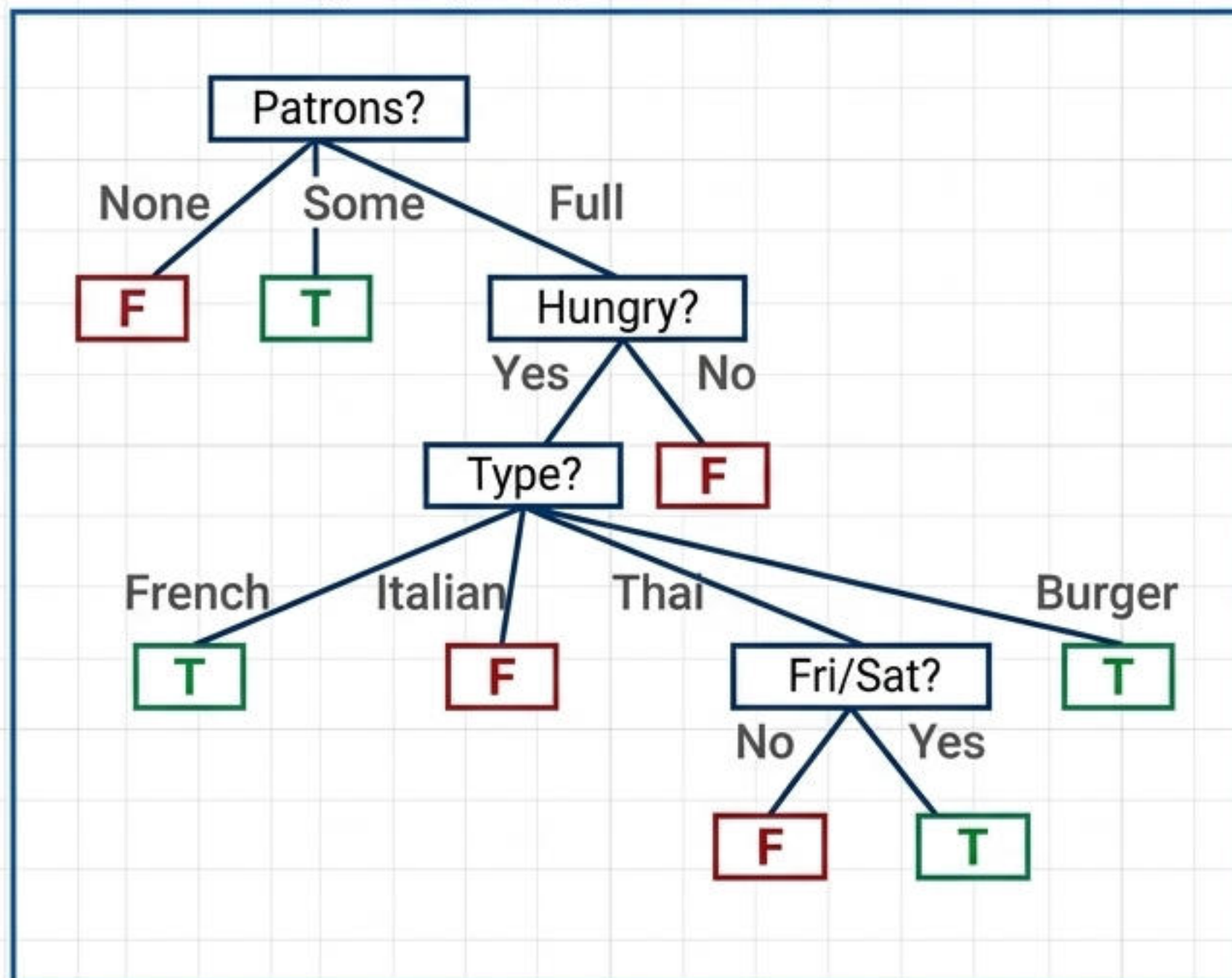


# מקרי קצה ותנאי עצירה: מתי העץ מפסיק לצמוח?

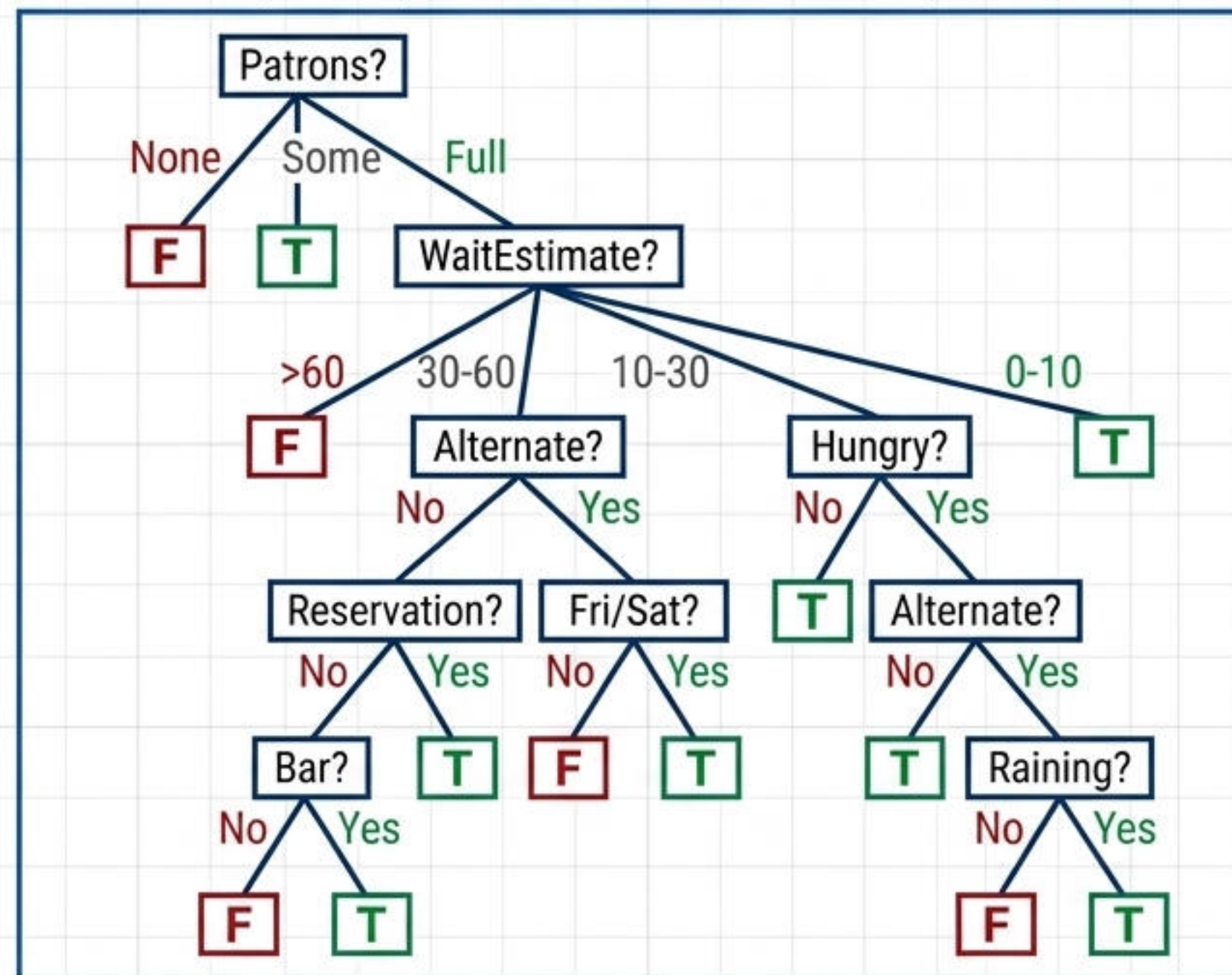
| המצב (If)                                                      | הפעולה (Then)                                                | היגיון (Why)                                                                  |
|----------------------------------------------------------------|--------------------------------------------------------------|-------------------------------------------------------------------------------|
| כל הדוגמאות בצומת שייכות לאותו סיווג (ענף טהור).               | עצור. החזר עלה עם הסיווג הזה.                                | הגענו ל-0 אנטרופיה. אין צורך לשאול יותר שאלות.                                |
| רשימת התכונות ריקה (נגמרו השאלות), אך הדוגמאות עדיין מעורבבות. | עצור. החזר עלה עם הסיווג הסיווג הנפוץ ביותר (Majority vote). | הנתונים מכילים רעש או מספיק מידע כדי להפריד. מתפשרים על הרוב.                 |
| קבוצת הדוגמאות ריקה (ענף שלא קרו בו מקרים באימון).             | עצור. החזר עלה עם סיווג ברירת המחדל (הנפוץ ביותר בצומת האב). | אין נתונים ללמוד מהם; אנו משתמשים בסטטיסטיקה הכללית ביותר הזמינה מהשלב הקודם. |

# עקרון התער של אוקאם: העץ האמיתי מול העץ הנלמד

העץ הנלמד (Simpler)



העץ 'האמיתי' המורכב (True)



ההיוריסטיקה שלנו (Information Gain) מייצרת עצים קומפקטיים מאוד. כמות קטנה של נתוני אימון אינה מצדיקה מודל מסובך מדי. עץ פשוט יותר הוא כללי יותר, נוטה פחות ללמוד "רעש", ומהווה את היישום המושלם של עקרון התער של אוקאם בלמידת מכונה.