



Module 03 – Scaling, Encoding, and Distances



175,000 vs 4 – Who Wins?

Computers notice large magnitudes first, not importance. Left unscaled, big numbers dominate distance and model behavior.

- ❑ Goal: Make comparisons fair and not to remove meaning, but to give each feature an equal opportunity to contribute.

Why Size Can Mislead

Different units confuse models

Meters, dollars, and counts on the same scale bias algorithms that use distance or gradients.

Scaling equalizes influence

Transformations put features on comparable numeric footing so learning focuses on patterns, not magnitudes.

Fair comparison, better performance

Proper scaling stabilizes optimization and improves interpretability of distances.

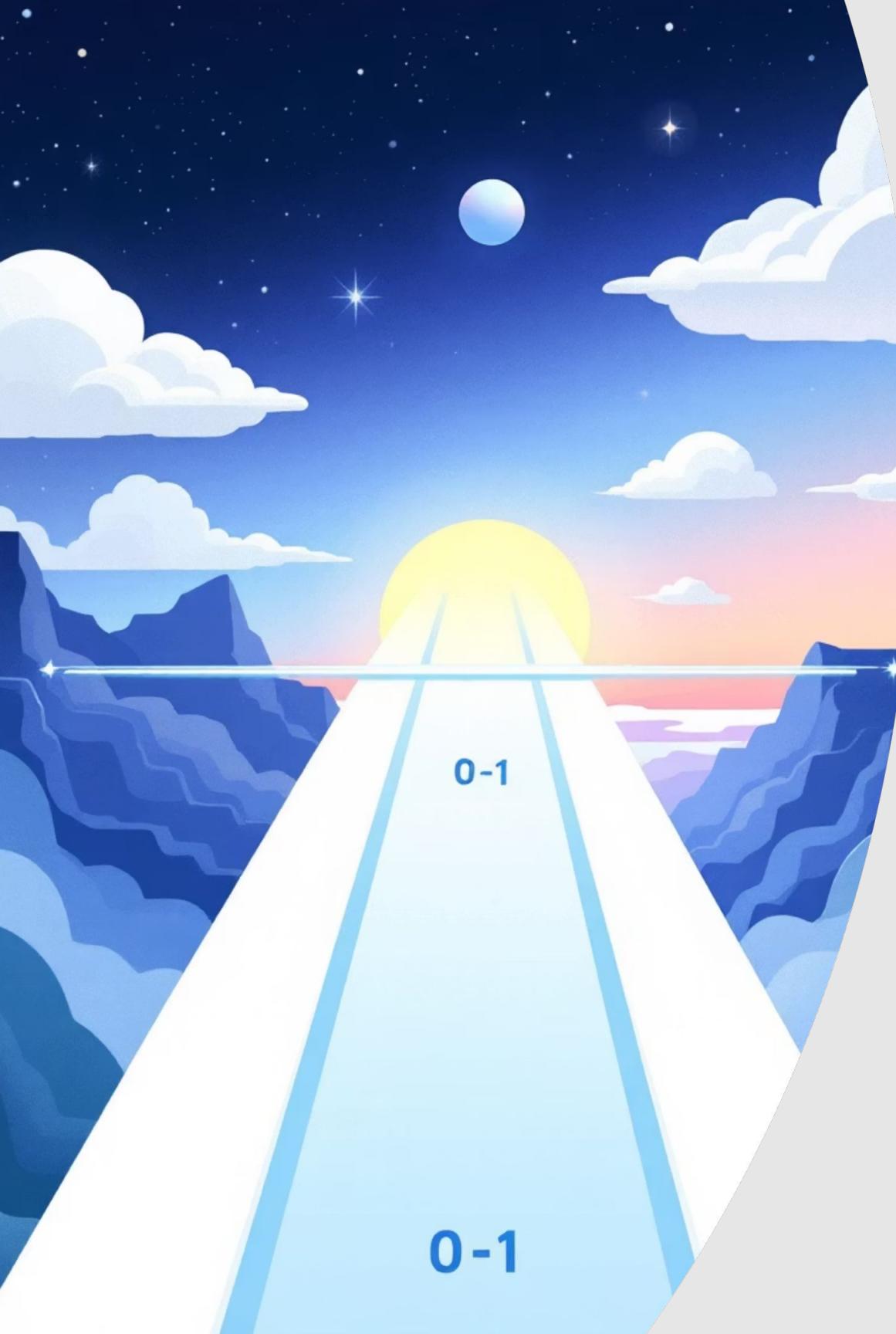


Standardization (z-score)

Formula: $z = (x - \text{mean}) / \text{standard deviation}$

Centers data at zero and scales by spread.





Min-Max Scaling (0 to 1)

Formula: $x' = (x - \min) / (\max - \min)$

- Maps values into [0,1].
- Great for bounded inputs
- Sensitive to outliers.

Robust Scaling – Median & IQR

Formula (concept): $x' = (x - \text{median}) / \text{IQR}$

- Uses median and IQR to reduce outlier influence.
- A better choice when distributions are skewed or contain extreme values.



Nominal vs Ordinal – Words Need Meaning

Too Nominal (no order)

Examples: City, Color, Species



Ordinal (ordered)

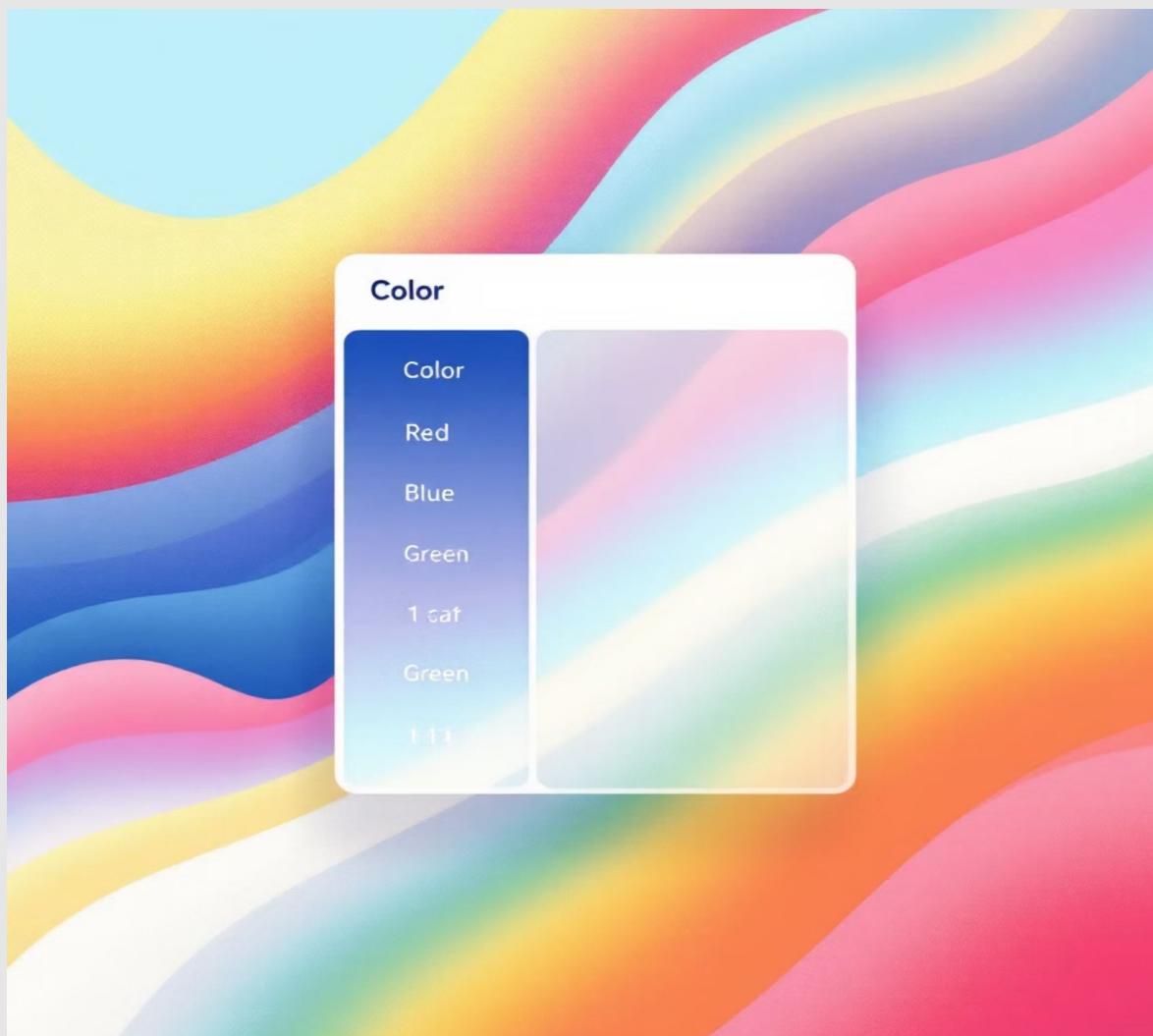
Examples: Small → Medium → Large, rating levels



Encoding: One-Hot vs Ordinal

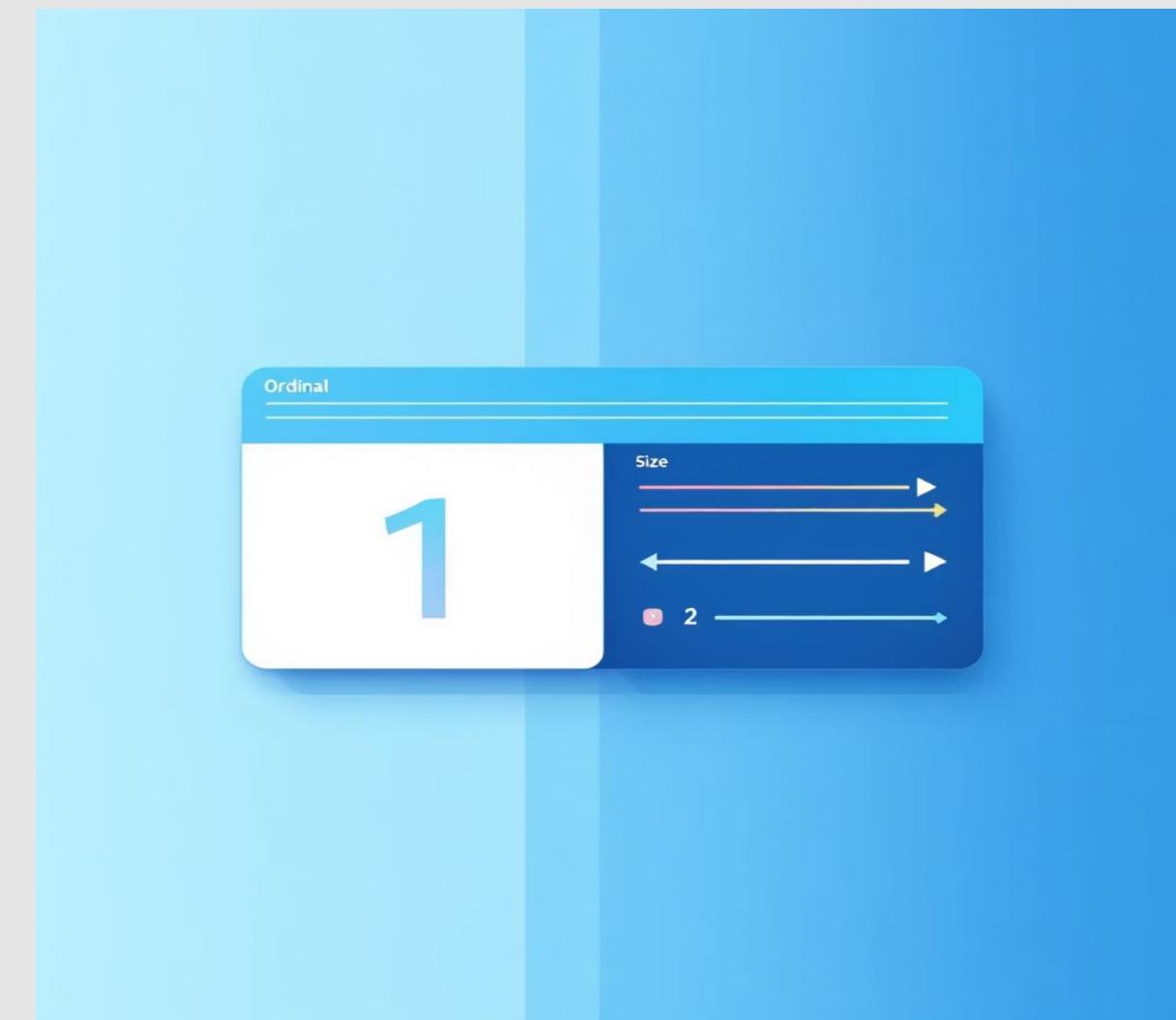
One-Hot Encoding

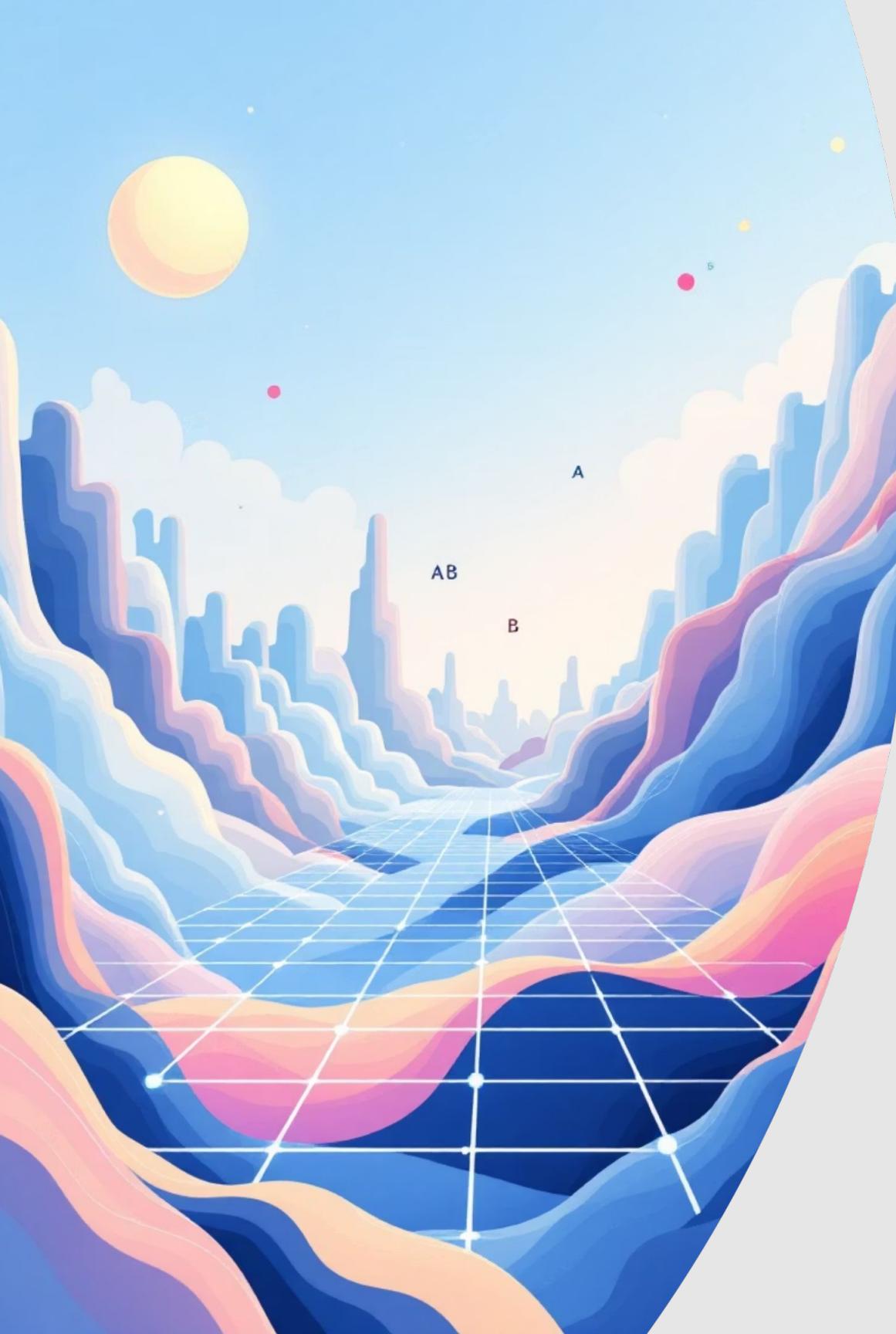
Color → Red, Blue, Green becomes separate 0/1 columns.
Best for nominal features without order.



Ordinal Encoding

Size → Small=1, Medium=2, Large=3. Preserve order when numeric progression makes sense.





Rows Become Vectors – Measure the Distance

After scaling and encoding, each row is a numeric vector. Distances between vectors quantify similarity.



Euclidean

$\sqrt{\sum(x_i - y_i)^2}$ — geometric straight-line distance. Sensitive to scale.

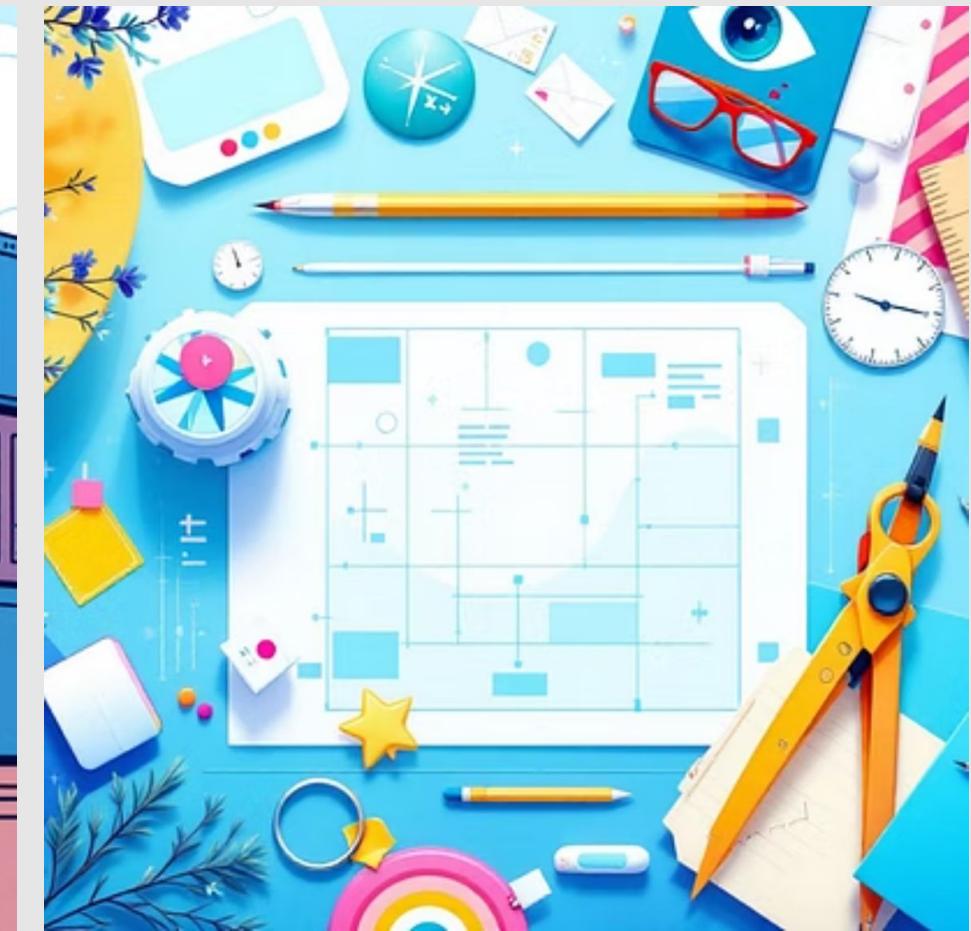
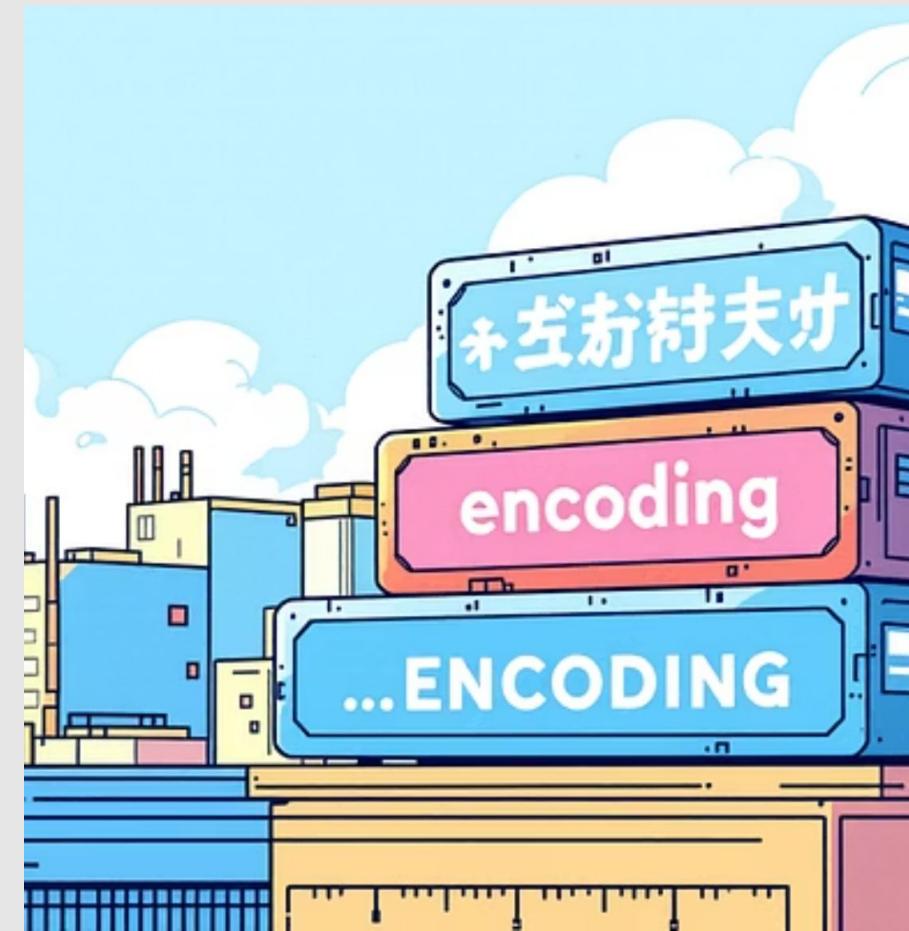


Manhattan

$\sum|x_i - y_i|$ — grid-style path distance. Often more robust to single-coordinate differences.

- ❑ Always scale before measuring distances. Otherwise magnitude dominates similarity.

This Module's Agenda



Scaling = Fairness

Normalize magnitudes so features contribute meaningfully.



Encoding = Meaning

Translate words into numbers that reflect type and order.



Distances = Comparison

Choose distance wisely and scale first.

