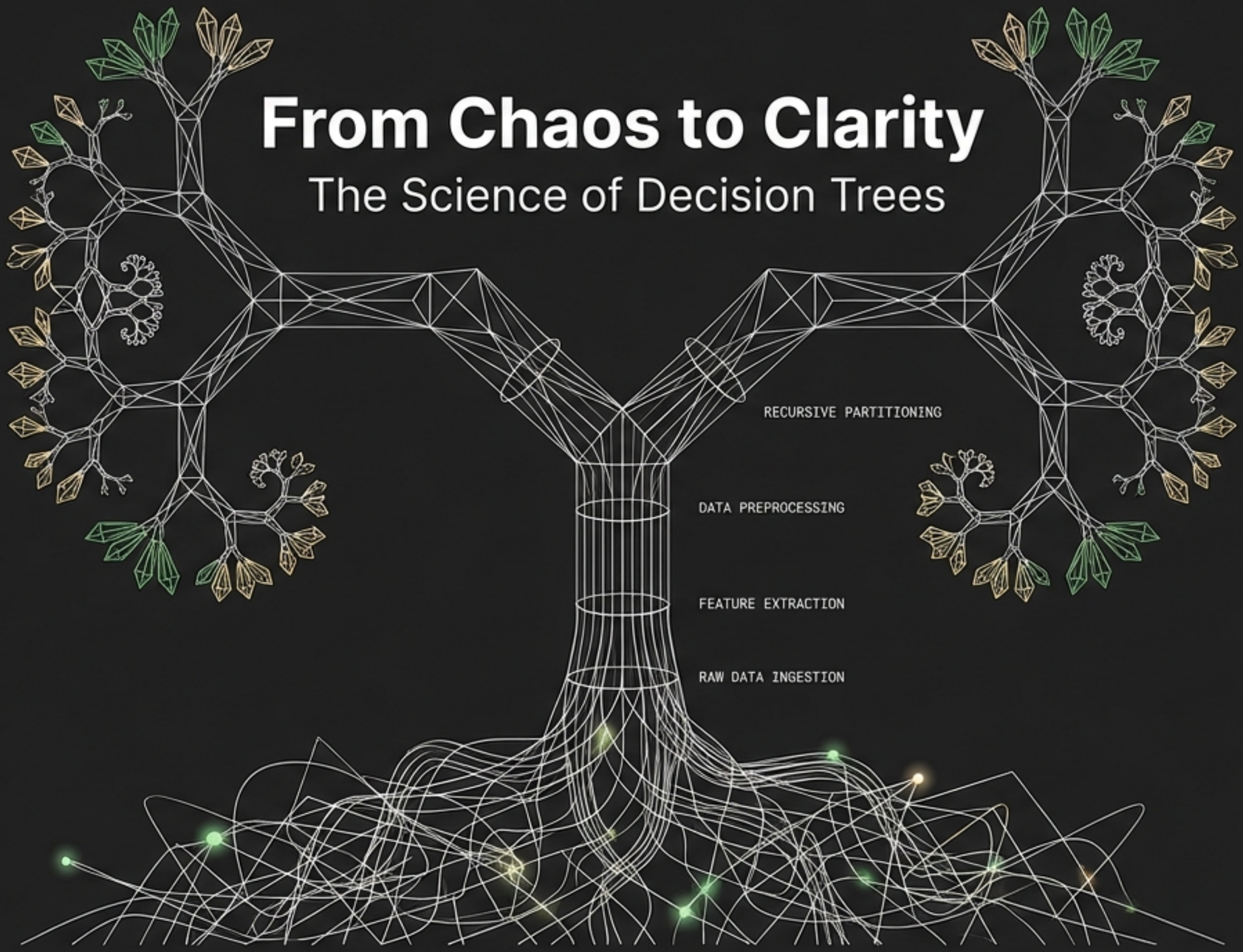
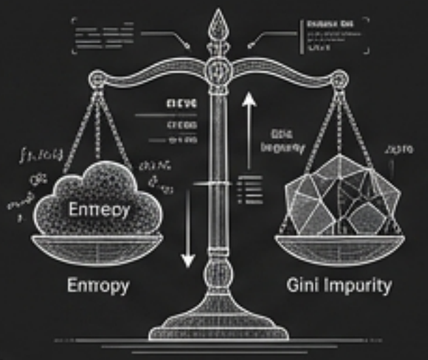
From Chaos to Clarity

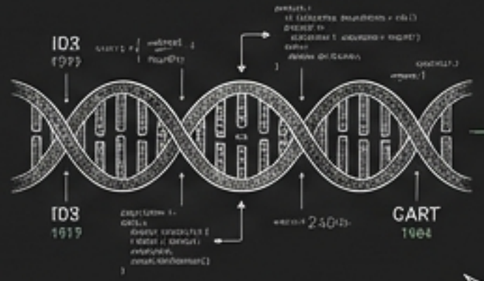
The Science of Decision Trees

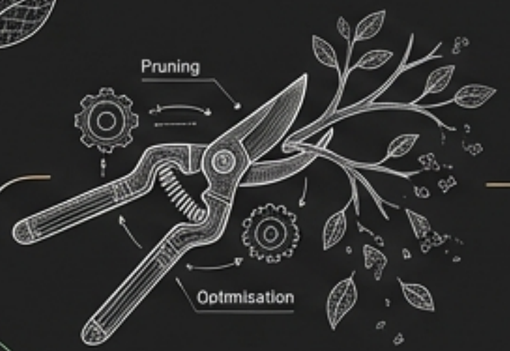


Anatomy, Evolution, and Real-World Application of Recursive Partitioning

The Evolutionary Arc







The Anatomy: Nodes, Leaves, and Branches

Structural elements of decision trees, mapping data flow & splits.

The Physiology: Entropy & Gini Impurity

Metrics for measuring information gain & impurity to optimize splitting.

The Evolution: ID3 to CART

Historical algorithms marking the shift from categorical to regression & classification trees.

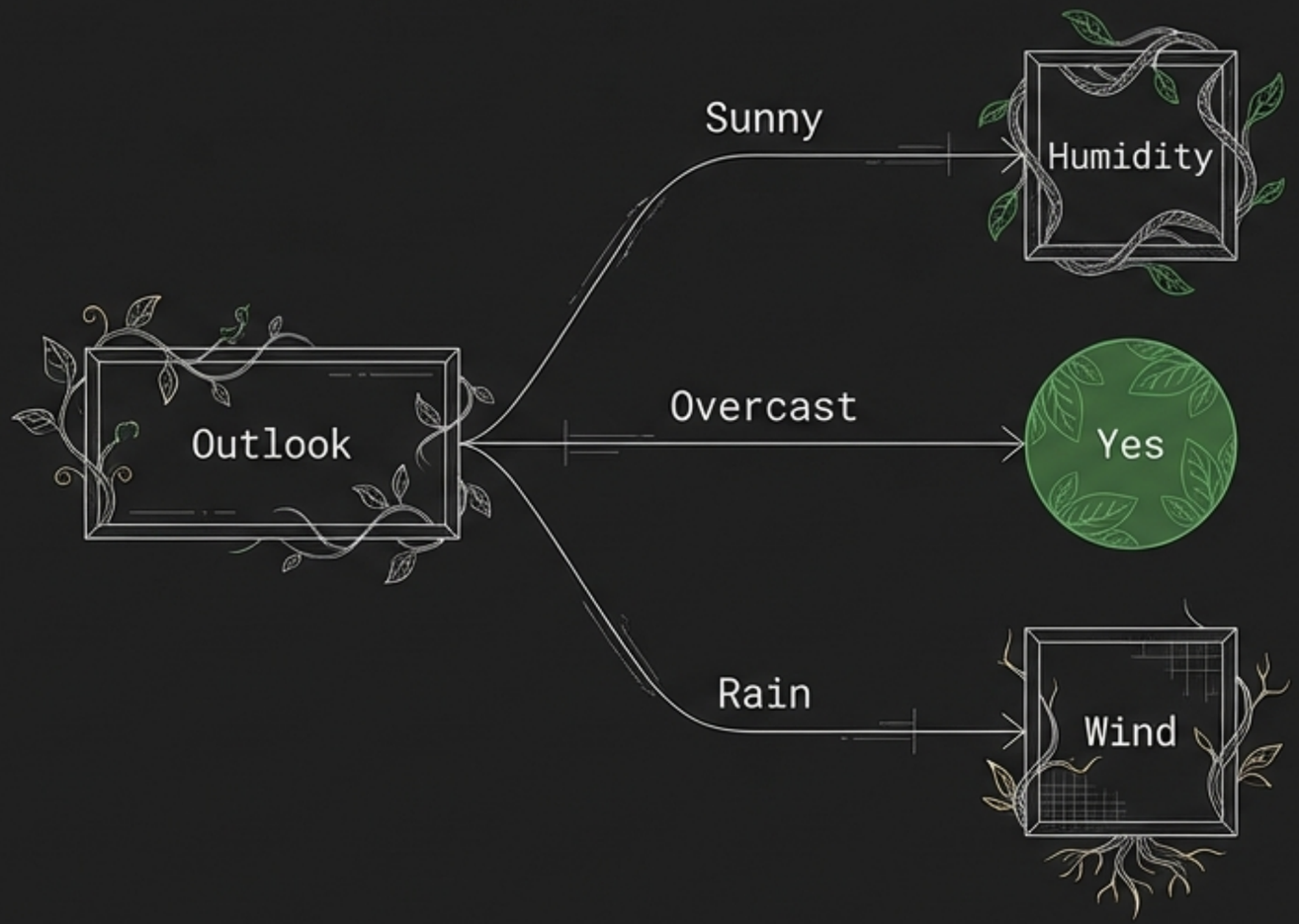
The Maintenance: Pruning & Optimization

Techniques to reduce complexity & prevent overfitting for better generalization.

The Habitat: Banking & Medicine Use Cases

Real-world applications including credit risk assessment, disease diagnosis, & treatment planning.

The Algorithm that Mimics Human Logic

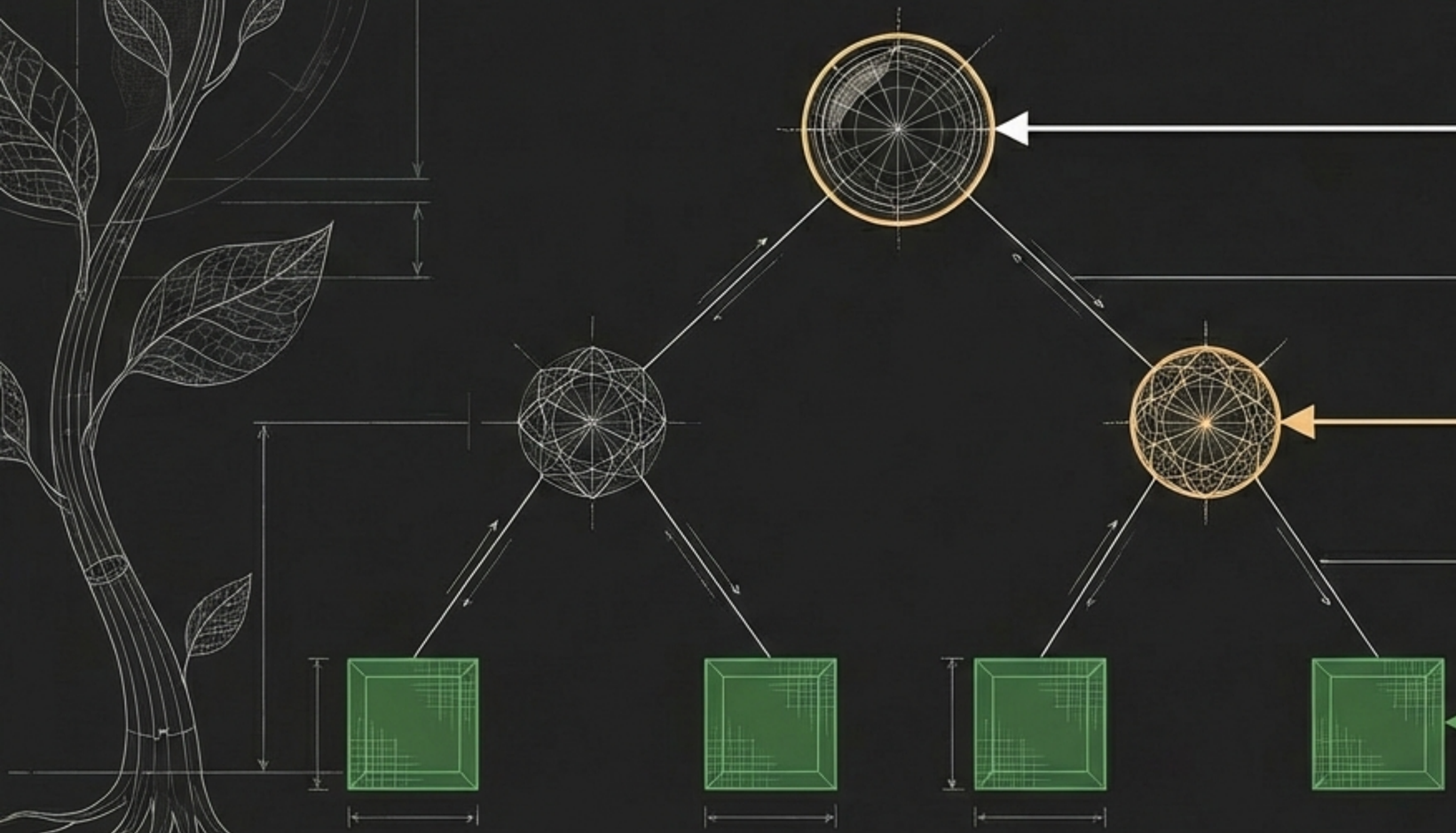


DEFINITION: A supervised learning algorithm for classification and regression.

MECHANISM: Recursive splitting of data into non-overlapping subspaces.

GOAL: Homogeneity.

Anatomy of a Decision Tree



ROOT NODE
(Entire Population)

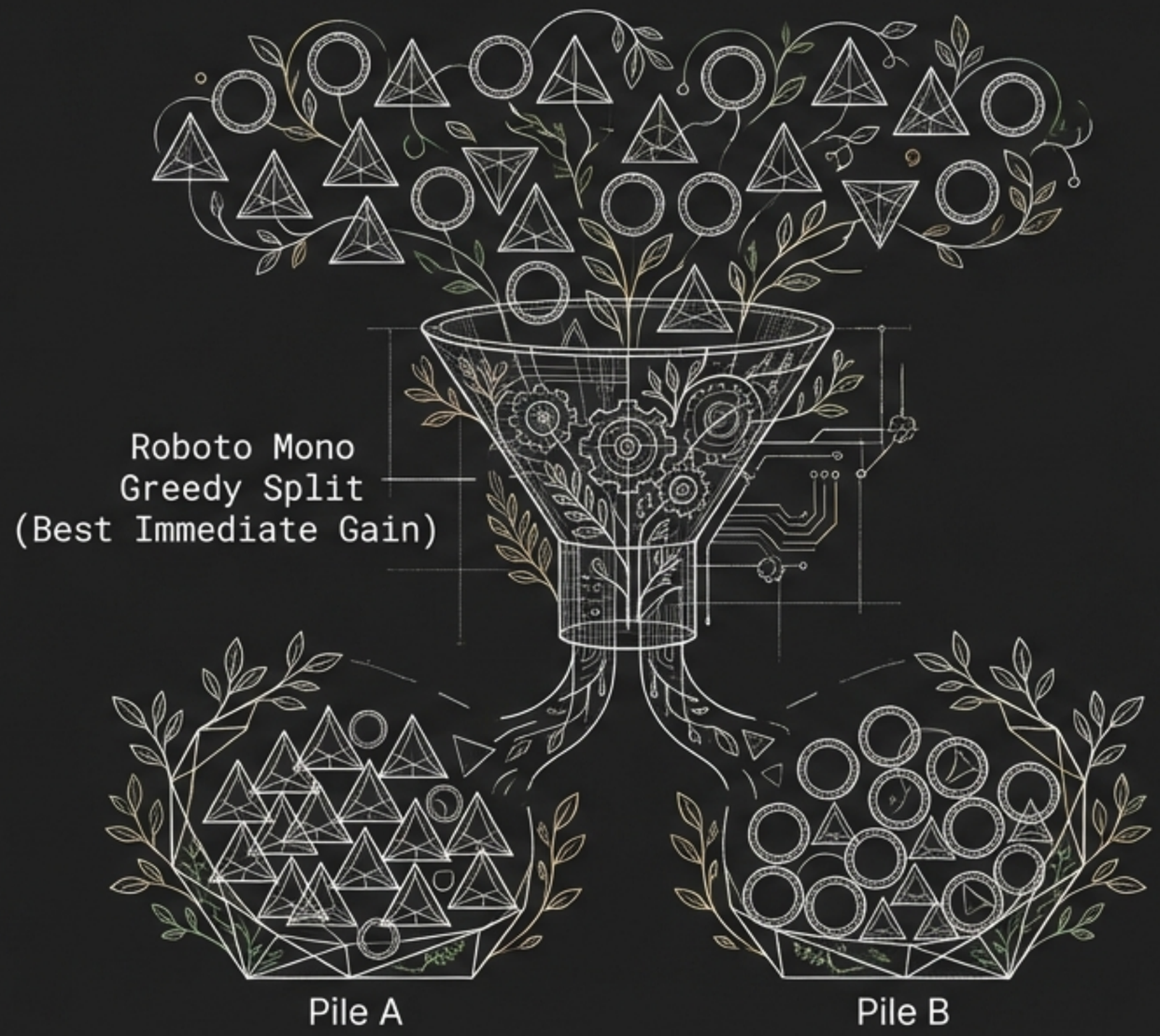
BRANCH (Decision Rule)

DECISION NODE
(Splitting Attribute)

BRANCH (Decision Rule)

LEAF NODE
(Final Outcome/Class)

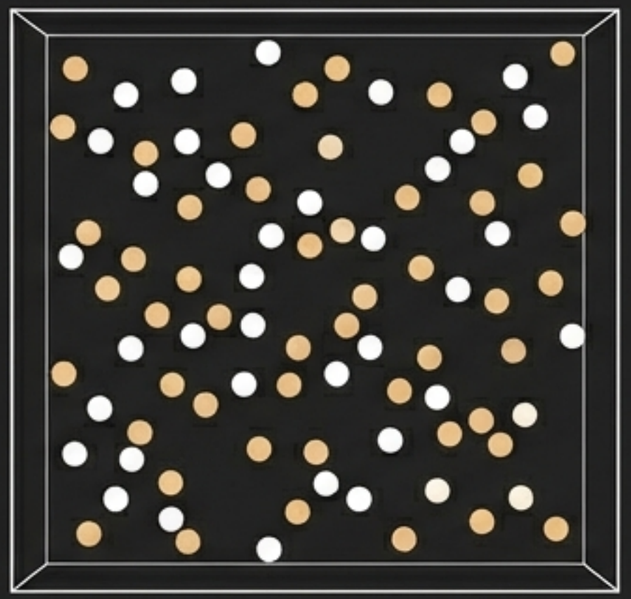
The Brain: Recursive Partitioning



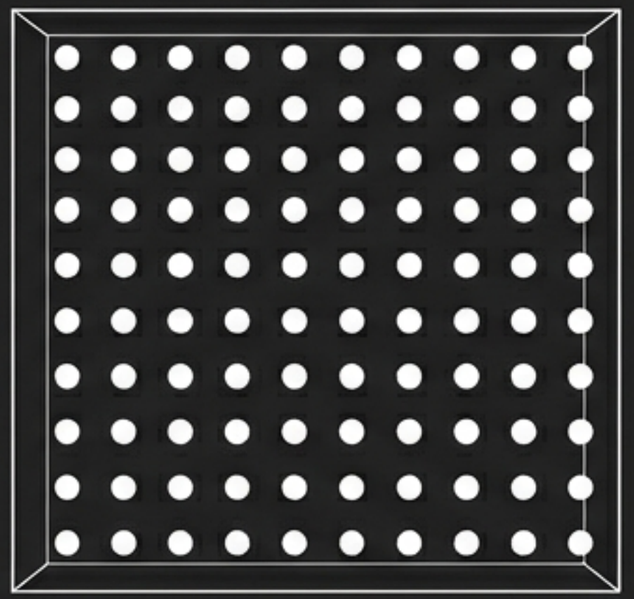
- **STRATEGY:** Top-Down, Greedy Approach.
- **ACTION:** Test all features. Select the threshold that maximizes purity.
- **CONSTRAINT:** Never backtracks to change previous splits.

The Physics of Decision Making: Entropy

High Entropy (1.0)
Pale Amber



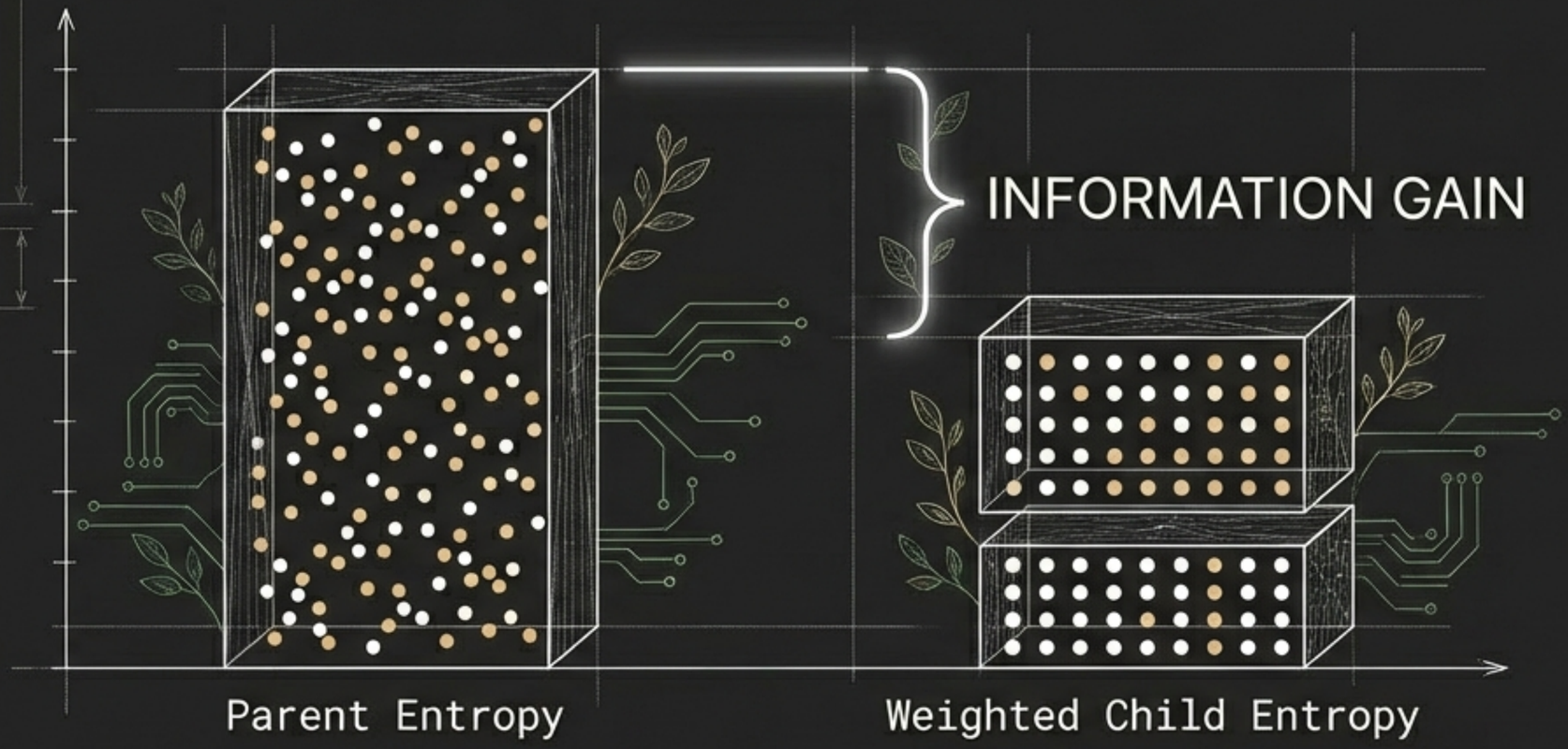
Low Entropy (0.0)
Crisp White



$$H(S) = - \sum p(x) \log_2 p(x)$$

The algorithm seeks to reduce Entropy (Uncertainty) at every step.

Measuring Success: Information Gain



`Information Gain = Entropy(Parent) - Entropy(Children)`

Used by: ID3 Algorithm

The Alternative Metric: Gini Impurity

$$Gini = 1 - \sum (Pi)^2$$

- DEFINITION: Probability of a randomly chosen element being classified incorrectly.
- TARGET: 0.0 (Perfect Purity).
- ADVANTAGE: Computationally faster than Entropy (no logarithms).
- USED BY: CART (Classification and Regression Trees).

Evolution of the Algorithm

ID3 (The Original)	C4.5 (The Successor)	CART (The Standard)
Metric: Information Gain	Metric: Gain Ratio	Metric: Gini Index
Data: Categorical Only	Data: Continuous & Categorical	Data: Classification & Regression
Missing Values: No	Missing Values: Yes	Structure: Binary Splits Only

Optimization: Pruning & Scaling

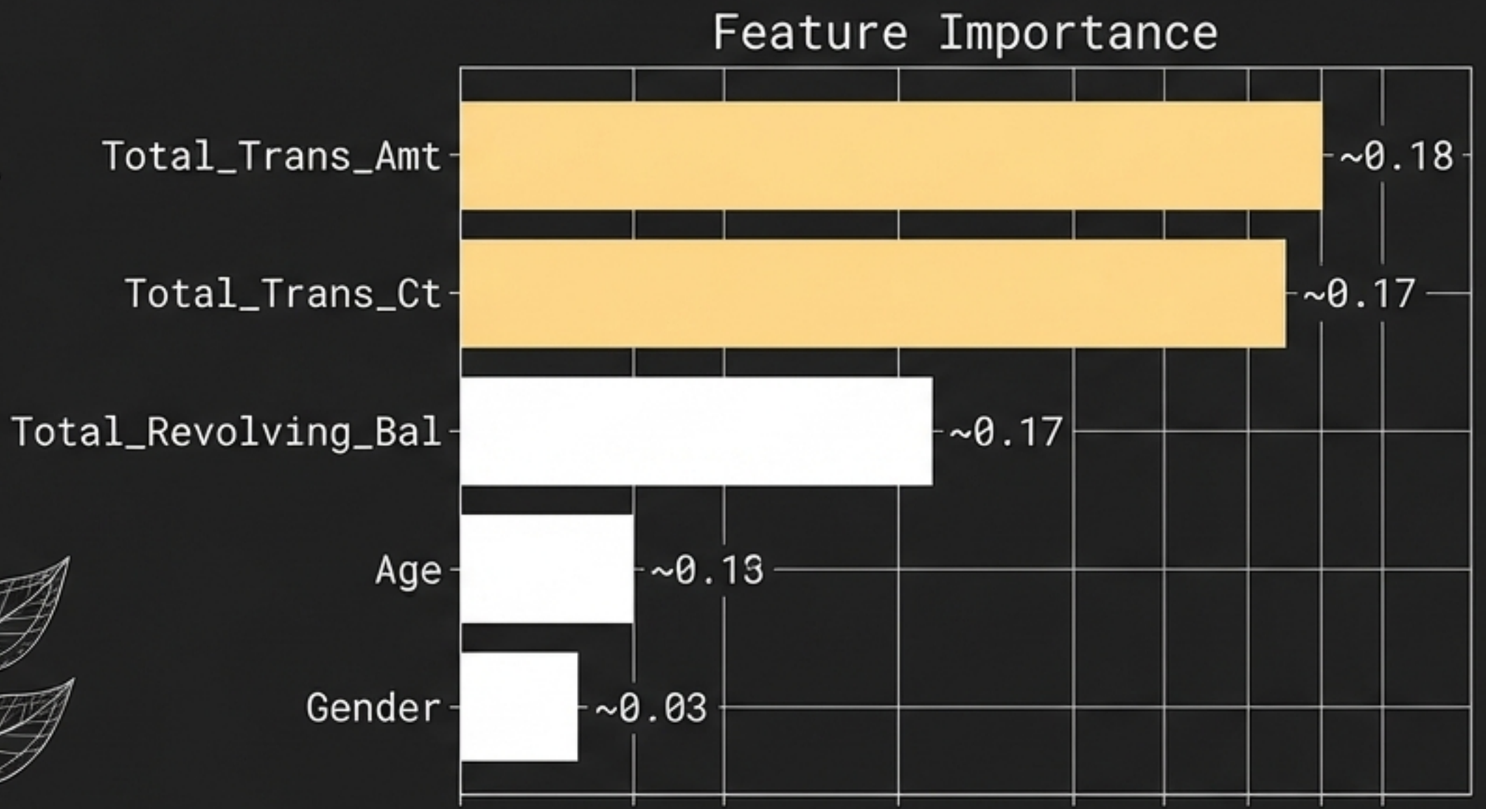
Overfitting (High Variance).

Pruned (Generalization).

PRUNING: Cost-Complexity Pruning removes branches that do not justify their complexity cost.

SCALING: Not Required. Trees process data based on order, not distance.

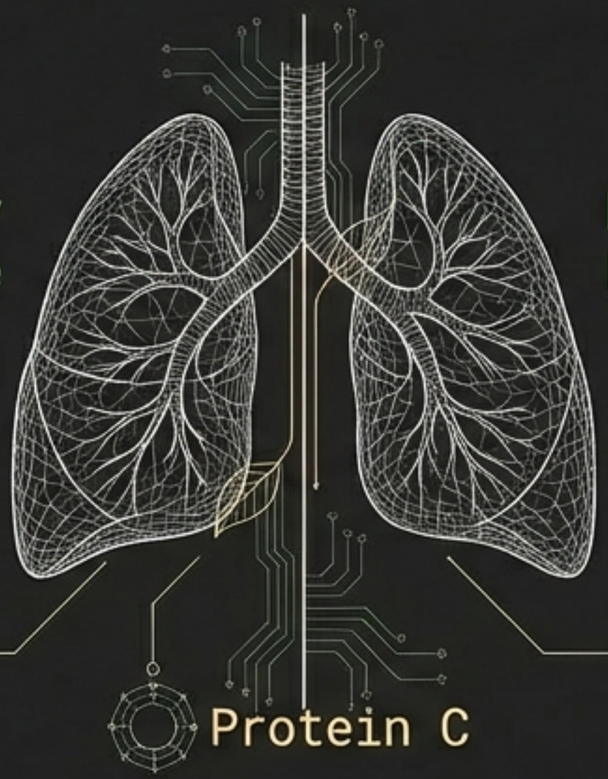
Case Study: Banking Customer Churn



Result: 91% Accuracy in predicting attrition.

Case Study: Precision Medicine in ARDS

Hyper-inflammatory Phenotype



Hypo-inflammatory Phenotype

IL-8

Protein C

Bicarbonate

Impact: Decision Trees classify phenotypes with >95% accuracy using only 3 variables, enabling targeted therapy.

Advantages & Limitations

PROS

- White Box: Highly interpretable.
- Versatile: Mixed data types.
- No Scaling needed.

CONS

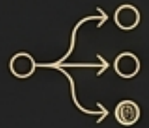
- Instability: Sensitive to small data changes.
- High Variance: Prone to overfitting.
- Greedy: Local optimization only.

Strength in Numbers: The Random Forest

Single Tree
(High Variance)

Random Forest
(Low Variance, High Accuracy)

BAGGING: Bootstrap Aggregating builds multiple trees on random samples.



RANDOM FOREST: Decorrelates trees by splitting on random subsets of features.

RESULT: Reduced variance and higher accuracy.

Summary: The Foundation of Interpretable AI

Structure:

Hierarchical splits to maximize purity.

Metrics:

Entropy (ID3) & Gini (CART) drive the logic.

Utility:

Critical for Explainable AI in Finance & Medicine.

Legacy: →

The building block for Random Forests & XGBoost.