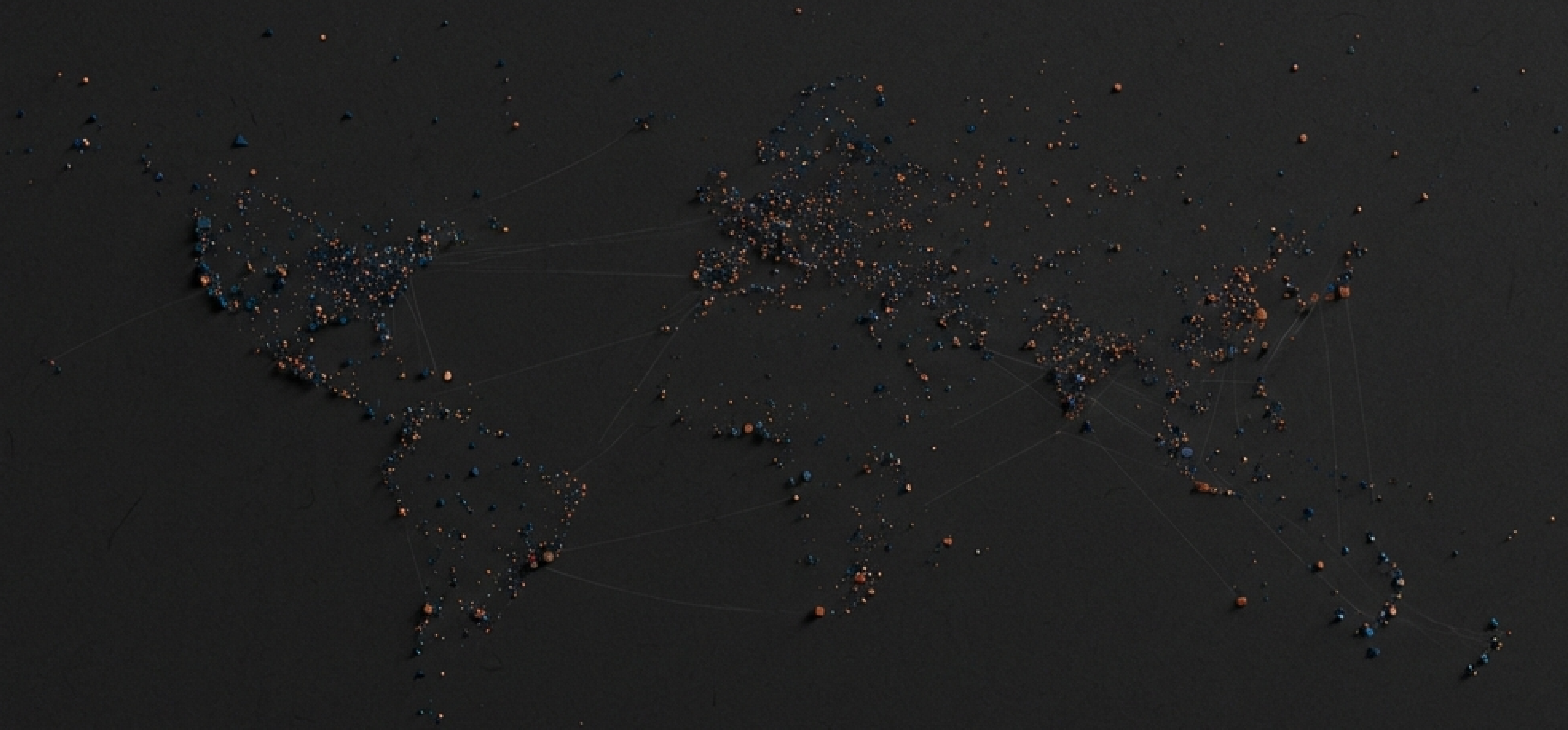
Semantics of Voids: Ignorance-Aware Machine Learning

How understanding what we *don't* know leads to smarter, more efficient models.



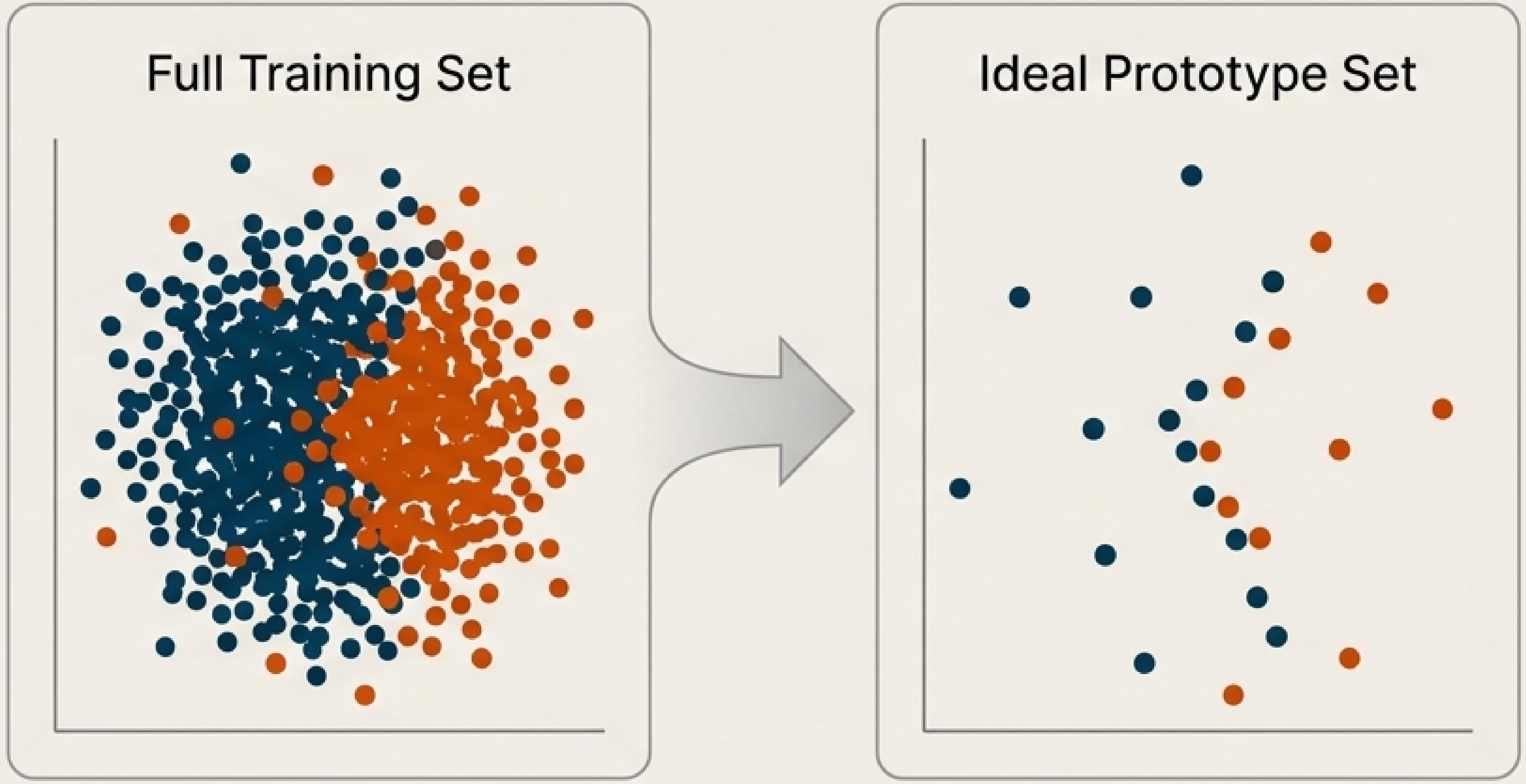
"Empty spaces – what are we living for" - Queen, 1991

NotebookLM

The Challenge of Instance-Based Learning

The Problem: Data Overload

- **High Storage Requirements**: Requires storing the entire training set.
- **High Computational Cost**: Computes distances to all stored prototypes for every new query.
- **Low Noise Tolerance**: Outliers can significantly harm classification accuracy.

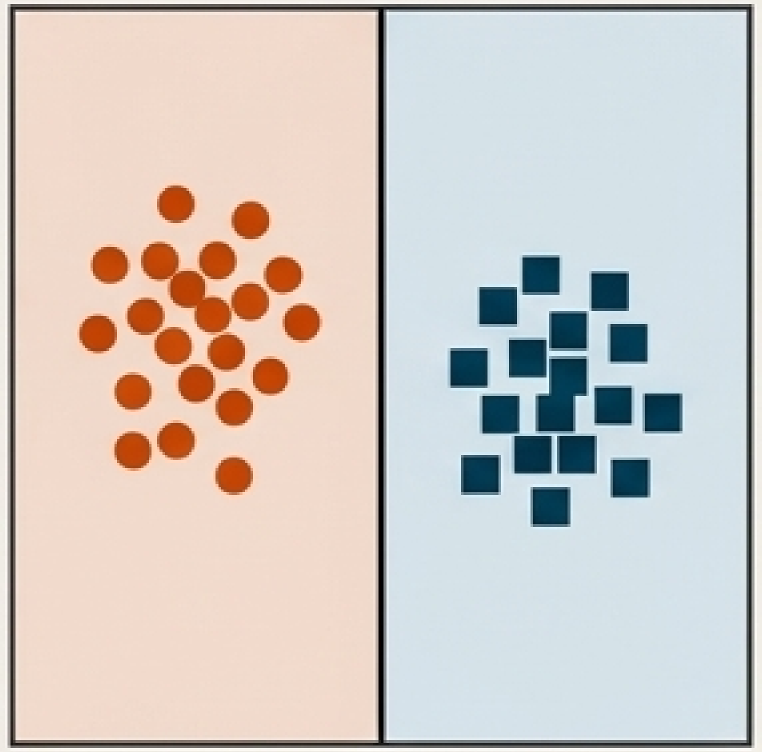


How do we select the few essential prototypes without losing—or even while improving—accuracy?

NotebookLM

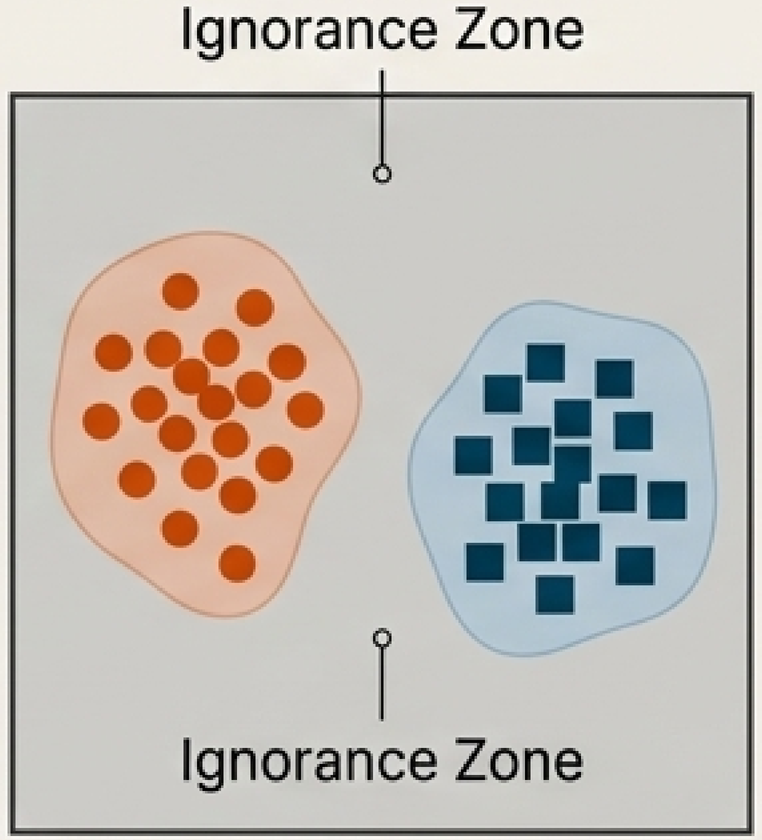
The Answer Isn't in the Data; It's in the Gaps

Traditional machine learning operates under a **Closed World Assumption**: what isn't in the dataset is considered false. This forces a decision for every point in the space.



Closed World: Every point is either 'terracotta' or 'blue'.

We propose an **Open World Assumption**: a lack of information does not imply falsehood. The gaps between known data are not empty; they are zones of **Ignorance.**



Open World: Some points are 'terracotta,' some are 'blue,' and others are 'unknown.'

NotebookLM

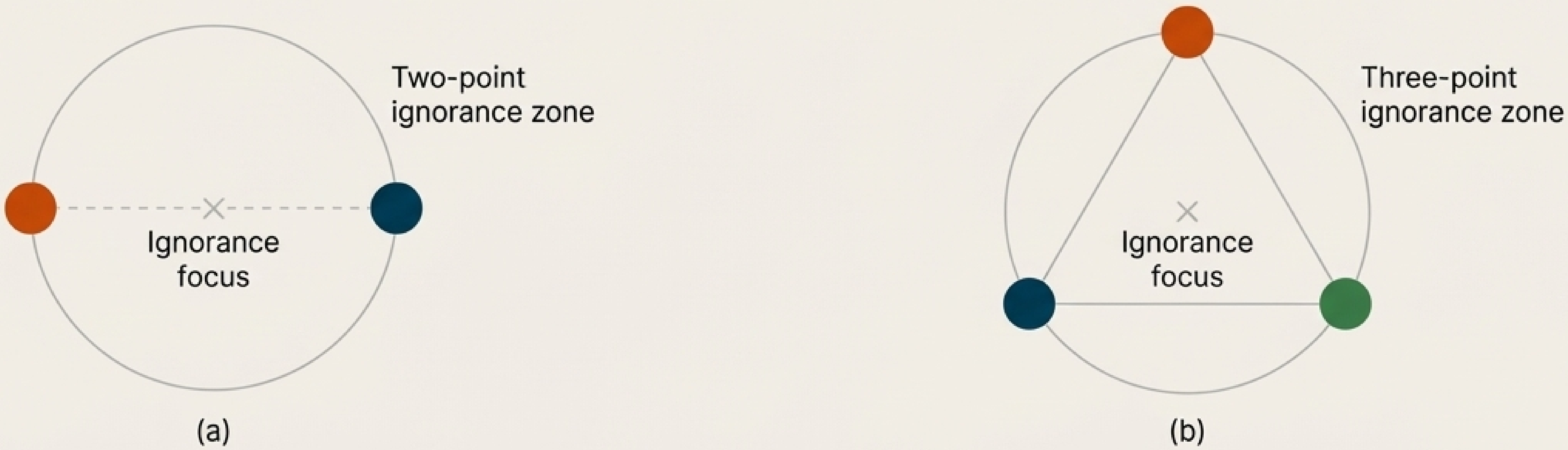
Ignorance Has Contours and Coherence

“Ignorance is not just a blank space on a person’s mental map. It has contours and coherence, and... rules of operation as well. So as a corollary to writing about what we know, maybe we should add getting familiar with our ignorance...”

- Thomas Pynchon, *Slow Learner*

A Simple Model: Discovering Ignorance Zones

We can begin to model these ‘contours’ by identifying the largest empty spaces between data points of different classes. The center of such a space is a ‘focus’ of maximal confusion or ignorance.



NotebookLM

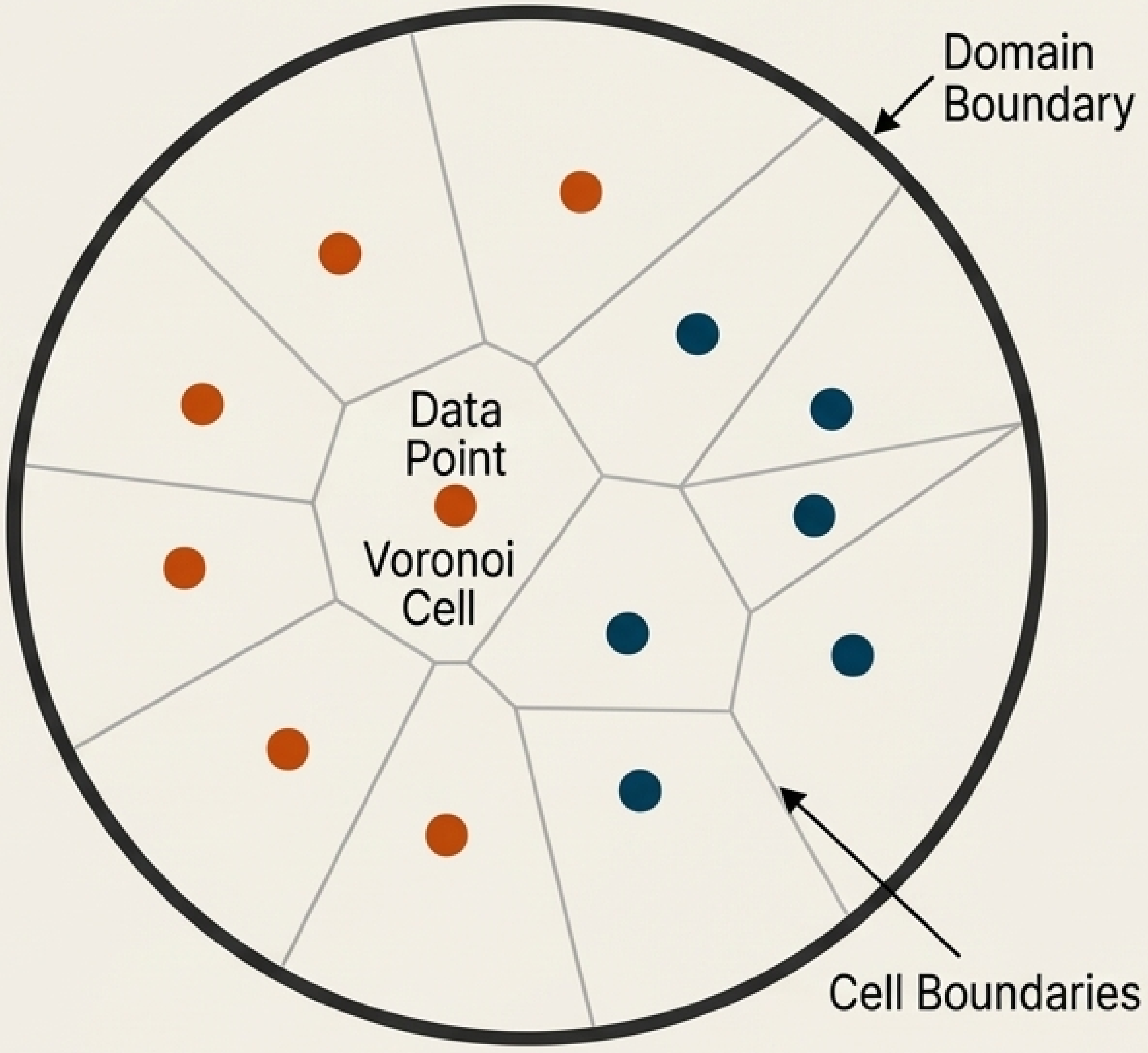
A Generic Model for Ignorance Discovery

Ignorance arises from two fundamental conflicts:

1. **Between Data**: The confusion between points of different classes.
2. **Between Data and Domain**: The uncertainty between a data point and the “rest of the world” at the edge of the known data space.

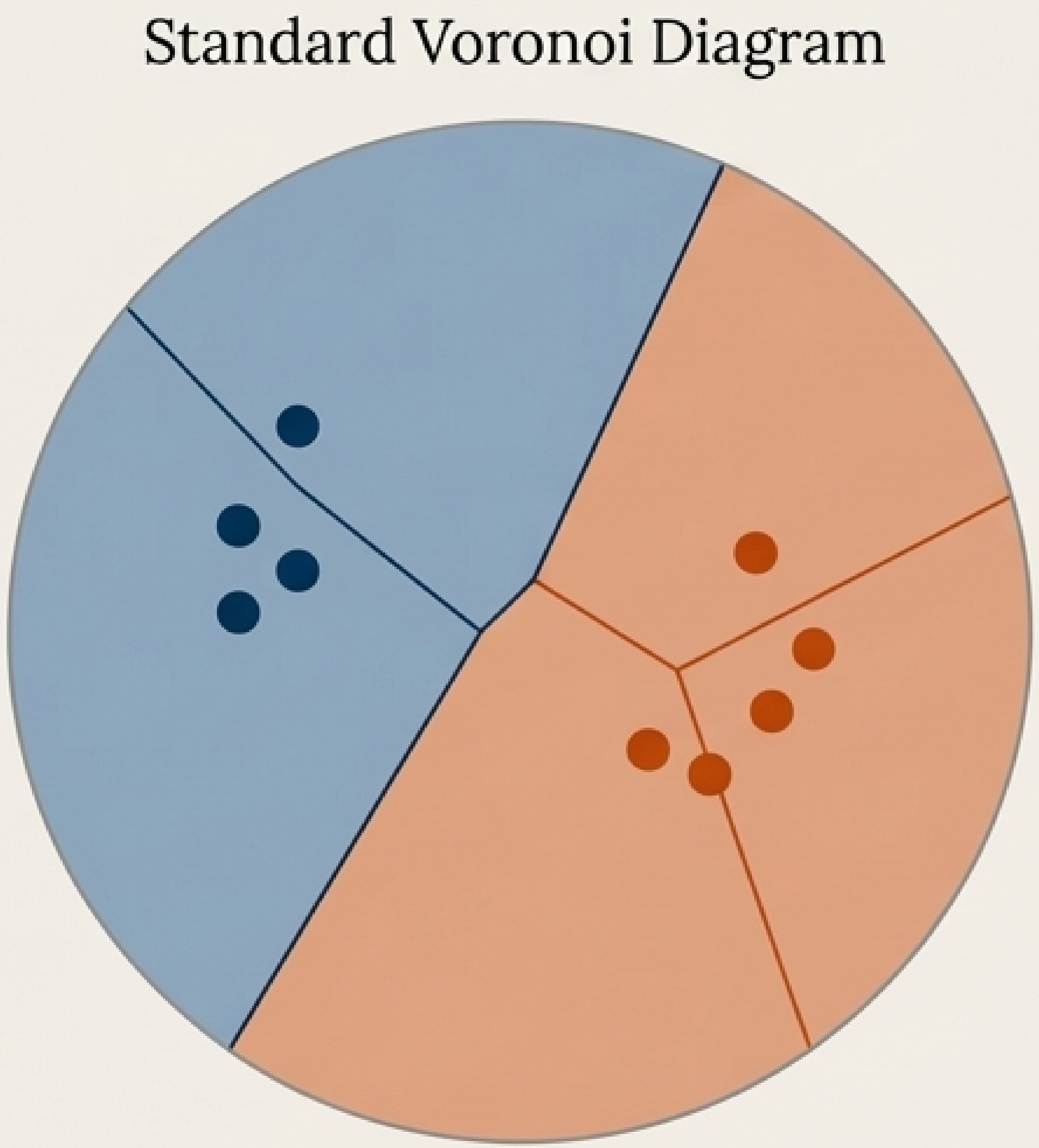
Methodology

1. We first partition the data space using a **Voronoi diagram**. Each data point resides in its own “cell,” where it is the closest point.
2. The boundaries of these cells represent lines of equal influence between neighboring points.
3. We treat these Voronoi cell boundaries—and the overall domain boundary—as the loci of potential ignorance.



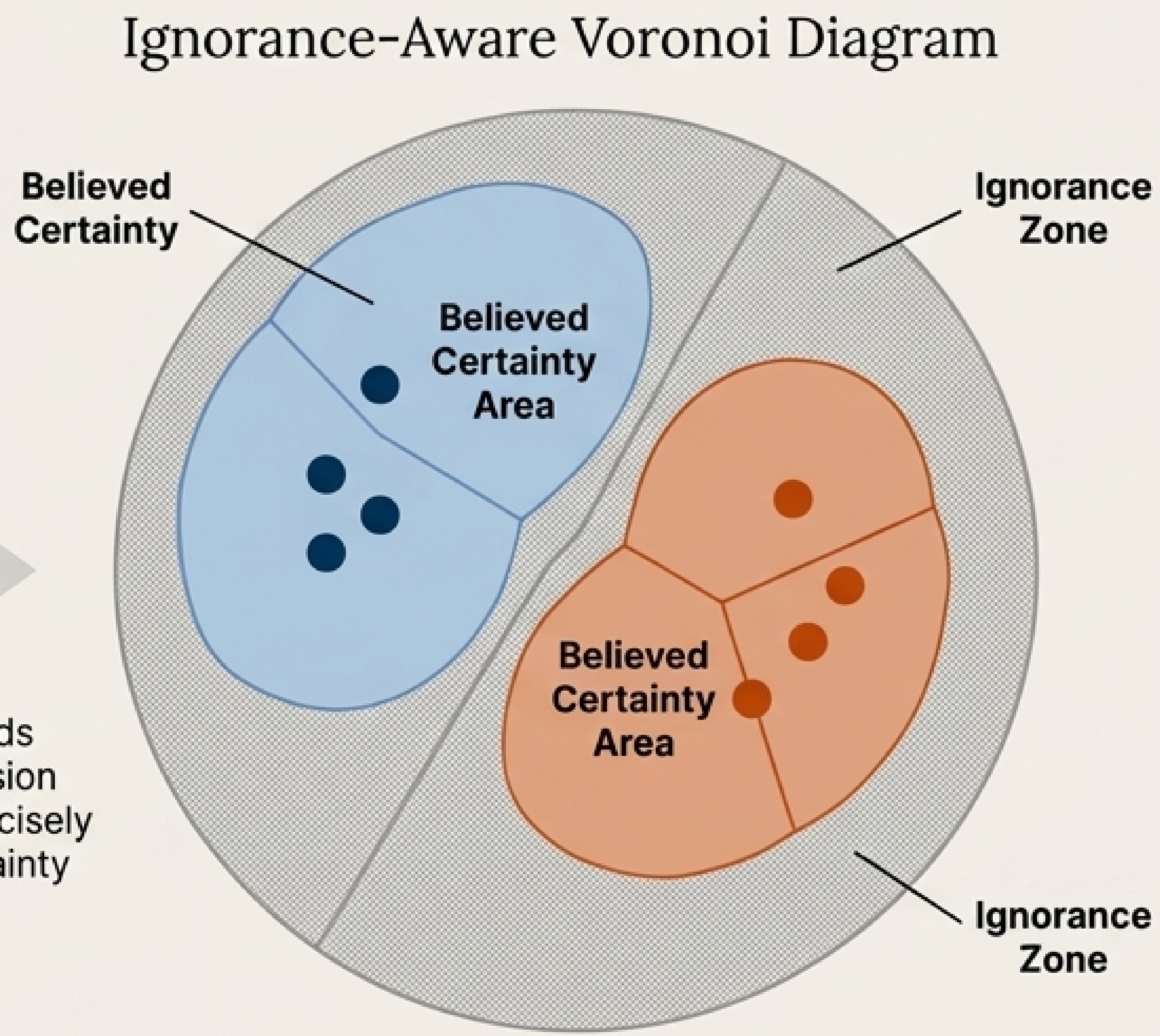
NotebookLM

From Voronoi Partitions to Ignorance-Aware Maps



Each point's influence extends to the cell boundary.

By modeling the voids centered on the decision boundaries, we can precisely map the areas of certainty and ignorance.



Certainty shrinks, and ignorance fills the boundaries of conflict and the edge of the domain.

NotebookLM

IPS: Curiosity-Driven Incremental Prototype Selection

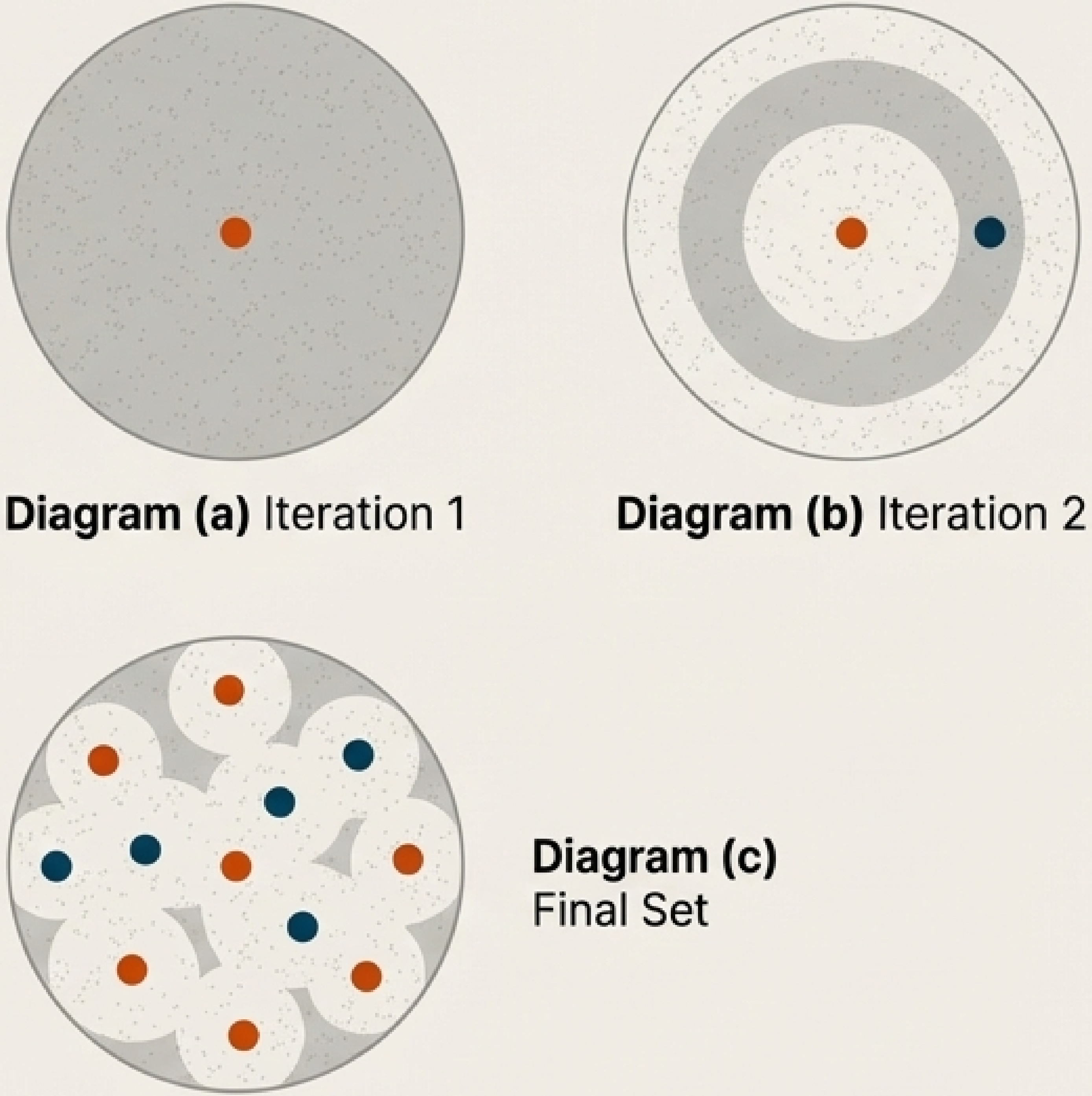
IPS iteratively builds a prototype set by asking:
“Where is our ignorance the greatest?”

Start Empty
Begin with the entire domain as one large ignorance zone.

Find Largest Void
Identify the center of the largest ignorance zone (the “curiosity focus”).

Query & Select
Find the nearest point from the original dataset to this focus and add it to the prototype set.

Recalculate & Repeat
Re-compute the ignorance map with the new prototype. Repeat until a stopping condition is met.



NotebookLM

APS: Adversarial Selection as a “Professor vs. Student” Game

The **Analogy**: Two competing actors select prototypes from the same dataset with conflicting goals.

The Professor

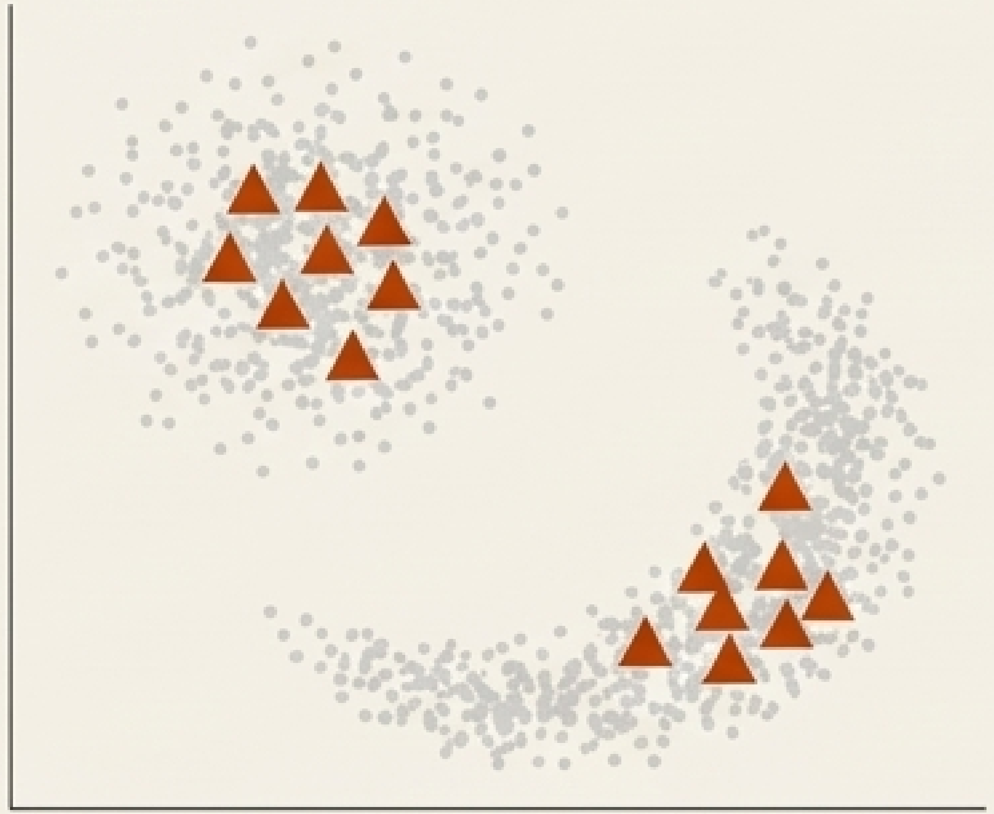
Goal: Find the student’s gaps in knowledge with minimal questions.

Strategy: Selects prototypes in the areas of greatest confusion—the decision boundaries. The professor seeks to find “hard questions.”

The Student

Goal: Cover the most knowledge with minimal study to pass the exam.

Strategy: Selects prototypes in the core areas of certainty to efficiently “answer” the professor’s potential questions. The student seeks to build a “robust knowledge base.”



Combined Outcome: A highly optimized pair of prototype sets for training and testing.

NotebookLM

Putting Ignorance-Aware Selection to the Test

Experimental Methodology

Key Metrics

- **Retention Rate (RR)**: % of original training instances selected as prototypes. (Lower is better).
- **Error Rate (ER)**: % of misclassified instances during testing. (Lower is better).

Process

- 10-fold cross-validation applied to each dataset.
- Performance of a 1-NN classifier is measured using the selected prototypes.

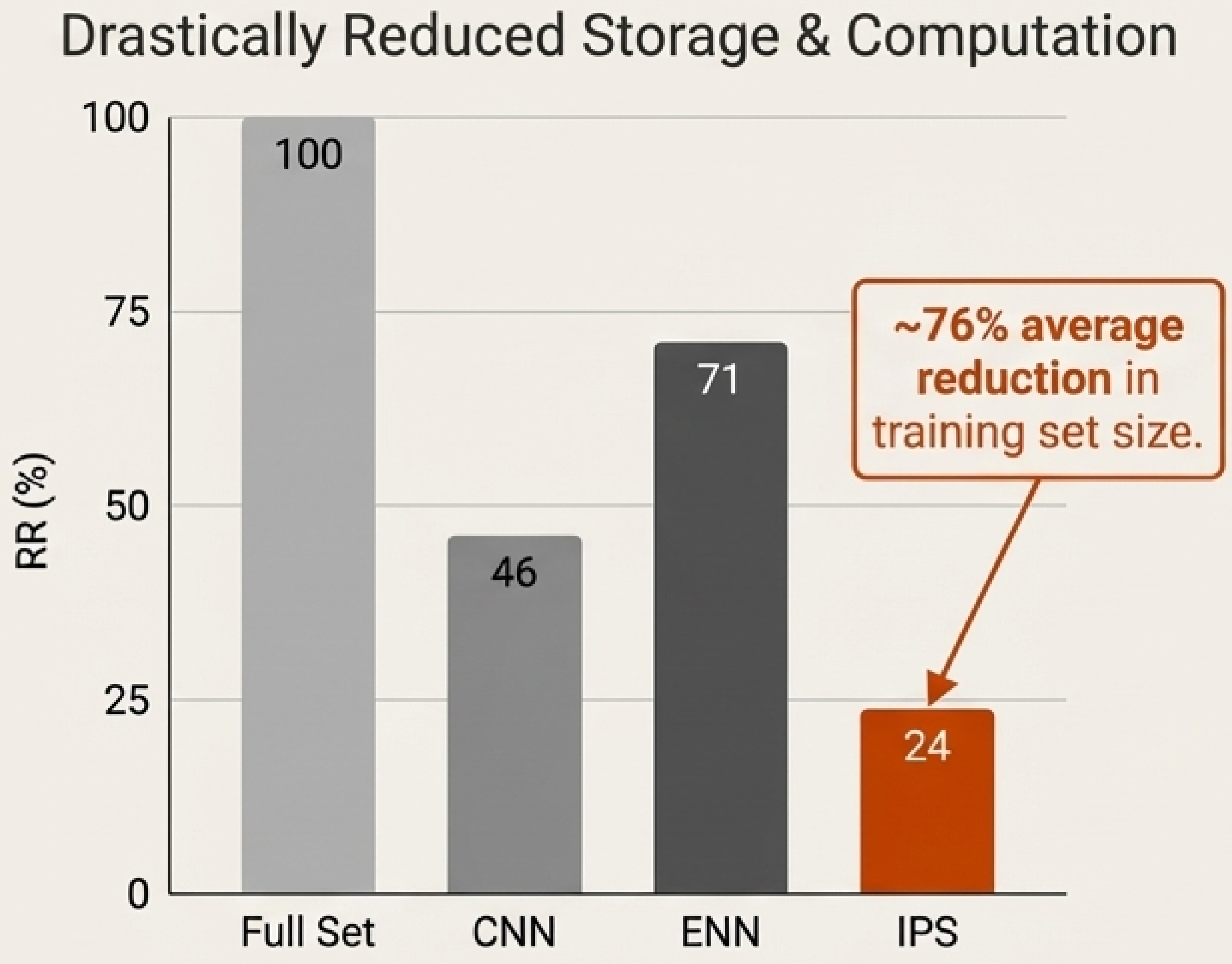
Diversity of the Testbed

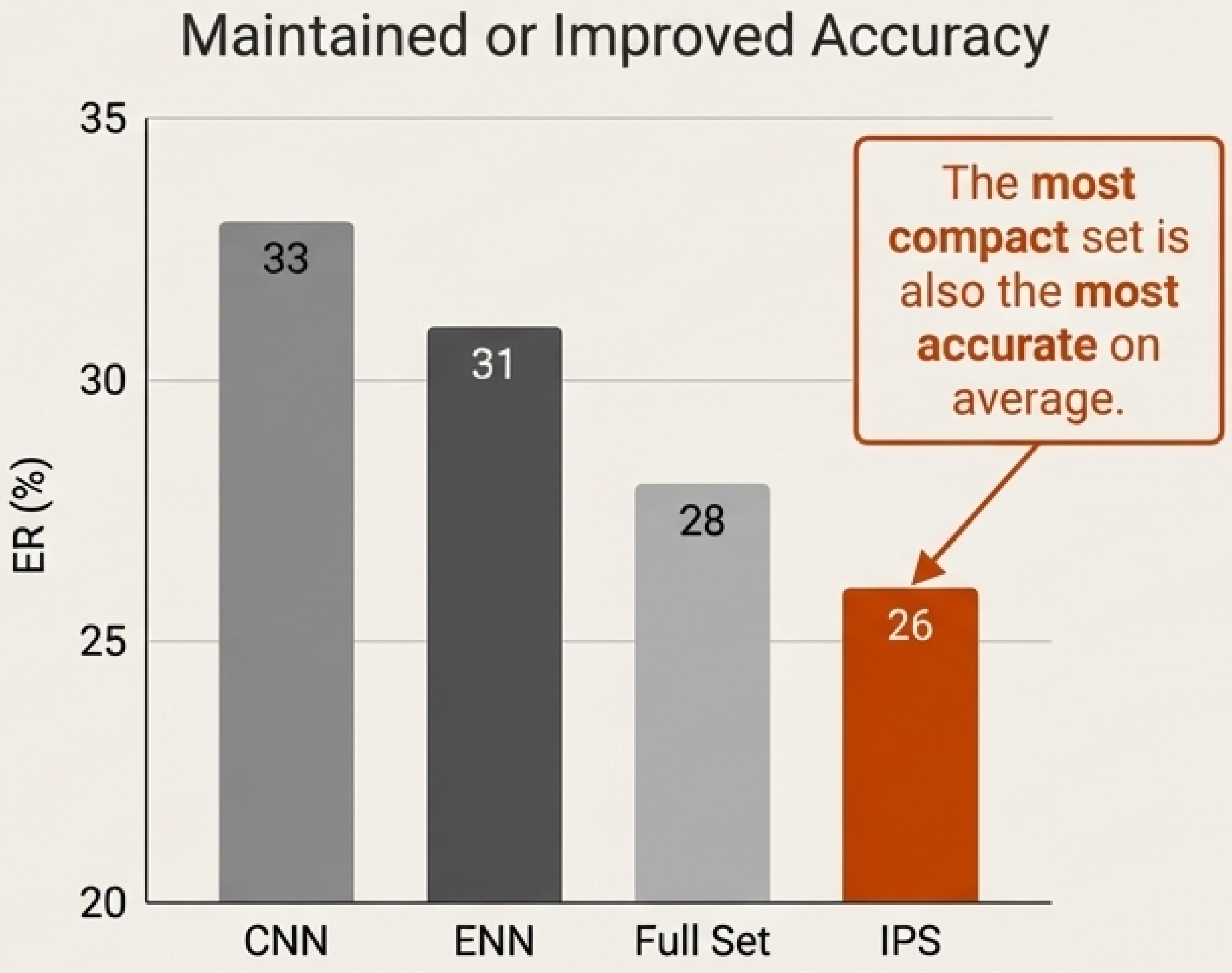
Dataset	# Exemplars	# Attributes	# Classes
Iris	150	4	3
Wine	178	13	3
Pima	768	8	2
Breast Cancer	699	9	2
Ionosphere	351	34	2
Glass	214	9	7
Bupa	345	6	2
Transfusion	748	4	2

NotebookLM

The Result: Radically Smaller Sets, Improved Accuracy

Performance Comparison: IPS vs. Standard Methods

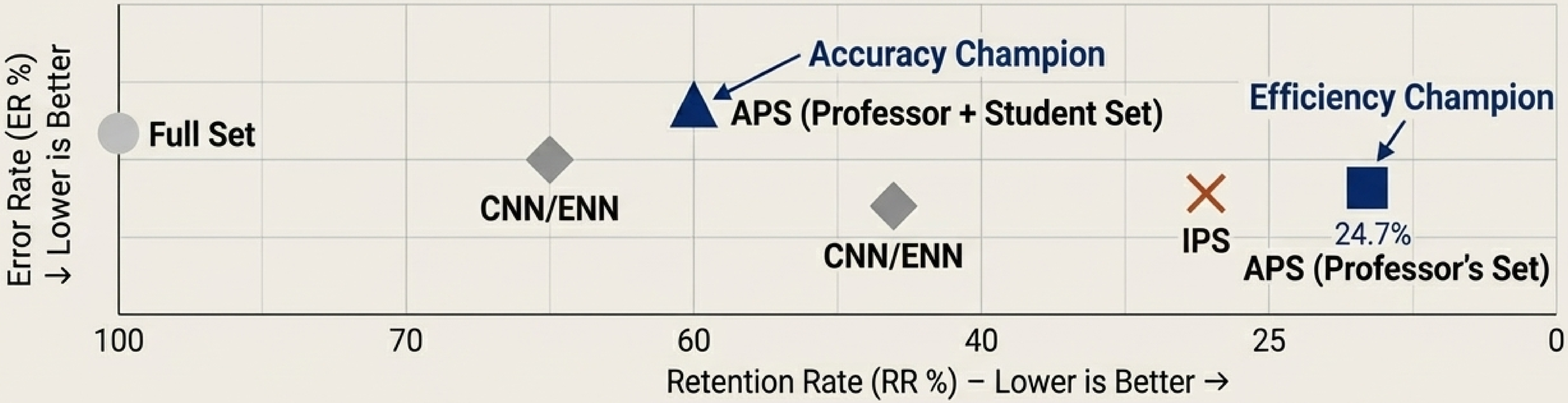




NotebookLM

The Adversarial Advantage: A Spectrum of Performance

APS Performance: Professor vs. Combined Sets



The Professor's selection yields the most compact training sets possible, reducing the original data by an average of **84%** while improving accuracy.

The combined Student & Professor set delivers the highest accuracy of any method tested, while still being less than half the size of the original dataset.

NotebookLM

Wiser Models Come from Understanding Ignorance

By explicitly modeling the voids within our data, we can guide machine learning algorithms to focus only on the most essential information.

Smaller

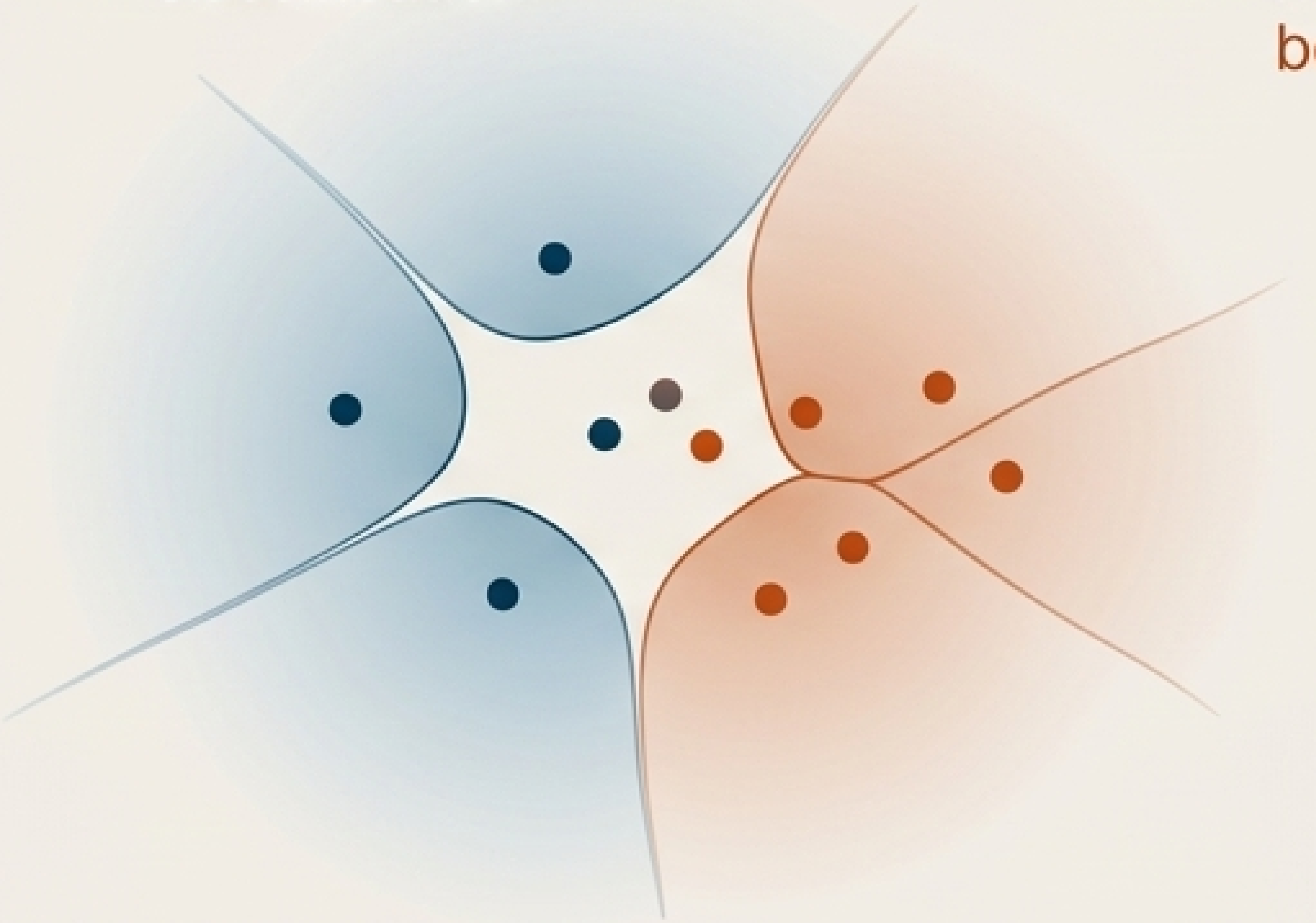
Dramatically reduced prototype sets (up to **84% smaller**).

Faster

Lower computational cost for classification.

Smarter

Higher accuracy by eliminating noise and focusing on decision boundaries.



The geometry of what we don't know is a powerful piece of knowledge. By learning to navigate the contours of our ignorance, we build not just more intelligent, but wiser, systems.

NotebookLM