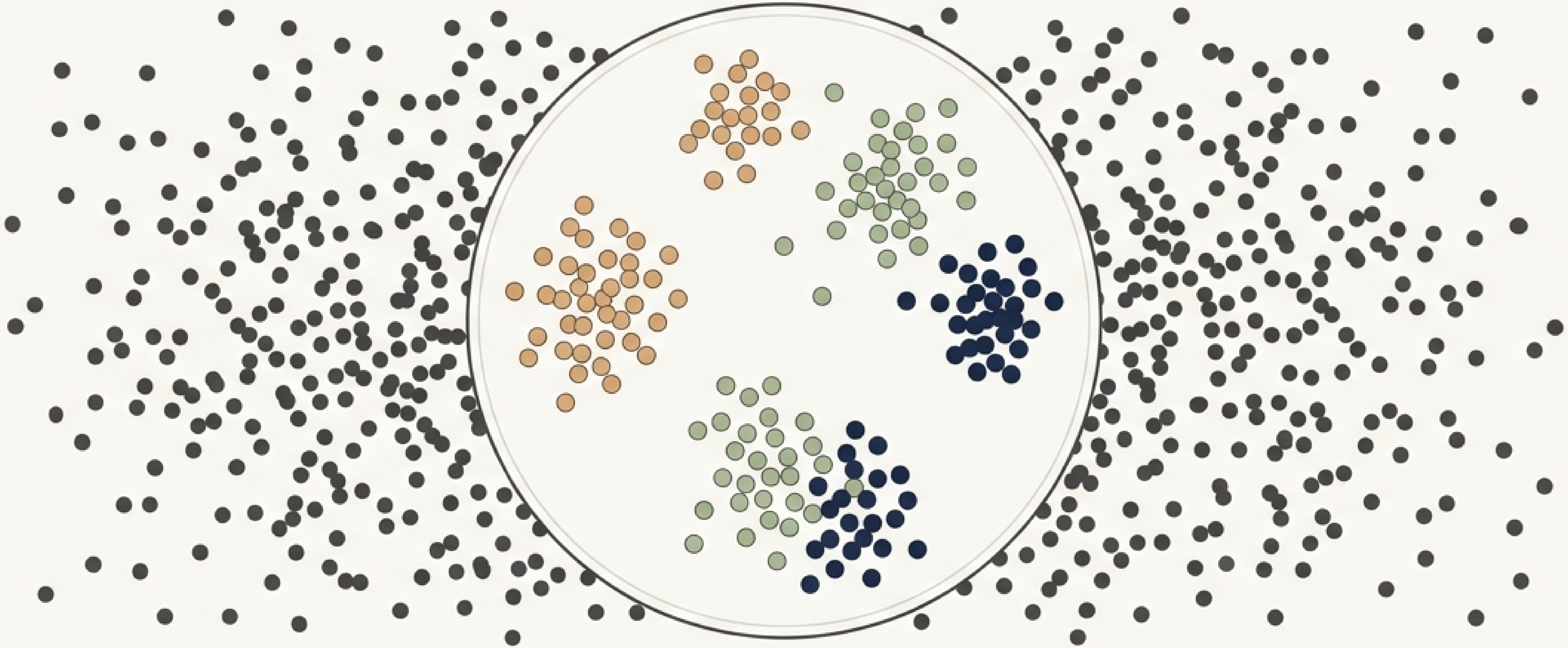
Changing the Lens on Unsupervised Learning

How Social-Context-Aware Distances Reveal Hidden Structure in Data



Vagan Terziyan, Oleksandr Terziyan, Oleksandra Vitko

University of Jyväskylä, FINLAND; Kharkiv National University of Radio Electronics, UKRAINE

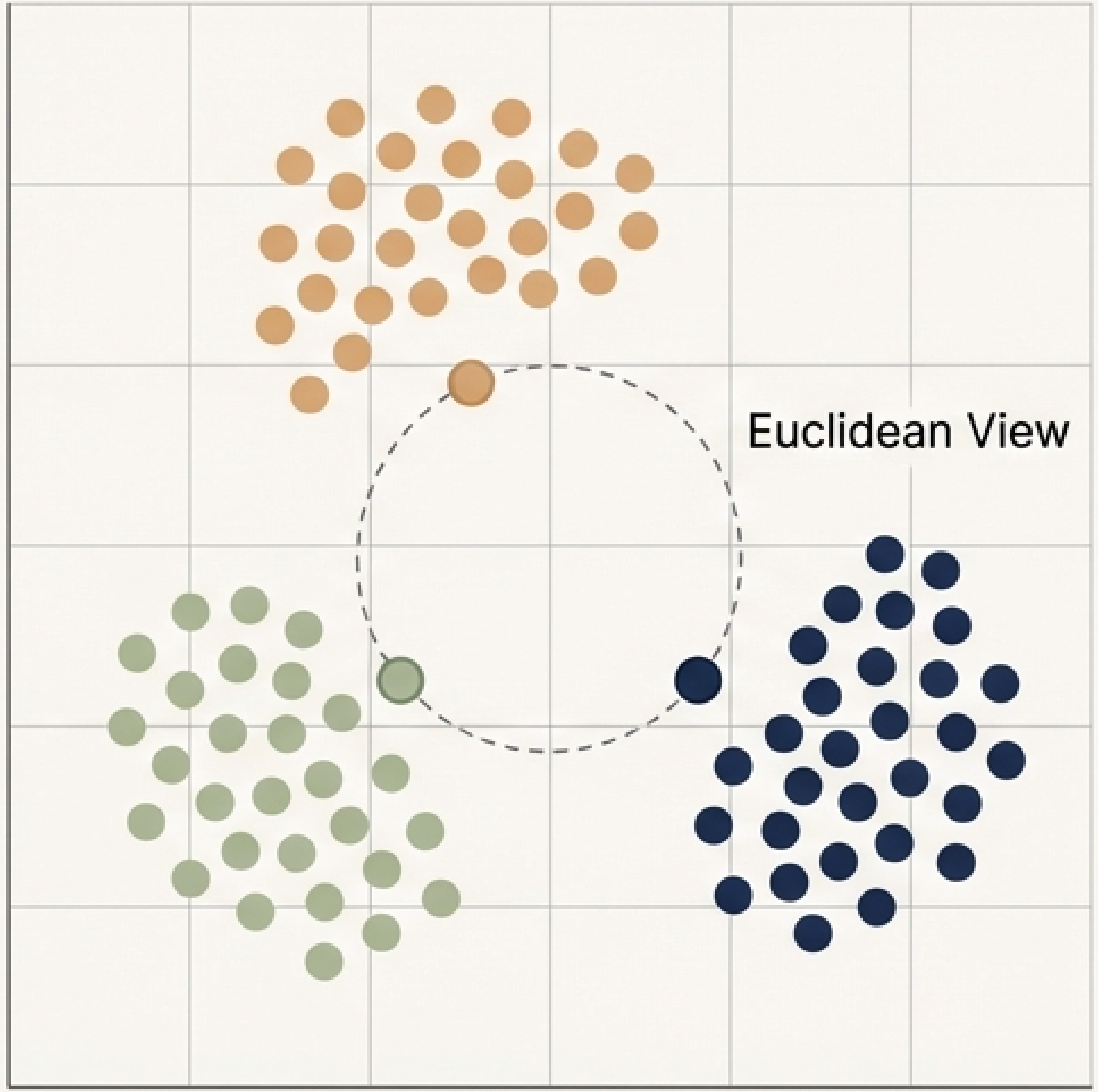
Clustering is Foundational, But Its Foundation is Flawed

A critical yet often underemphasized determinant of clustering performance is the distance or similarity metric used to compare data points.

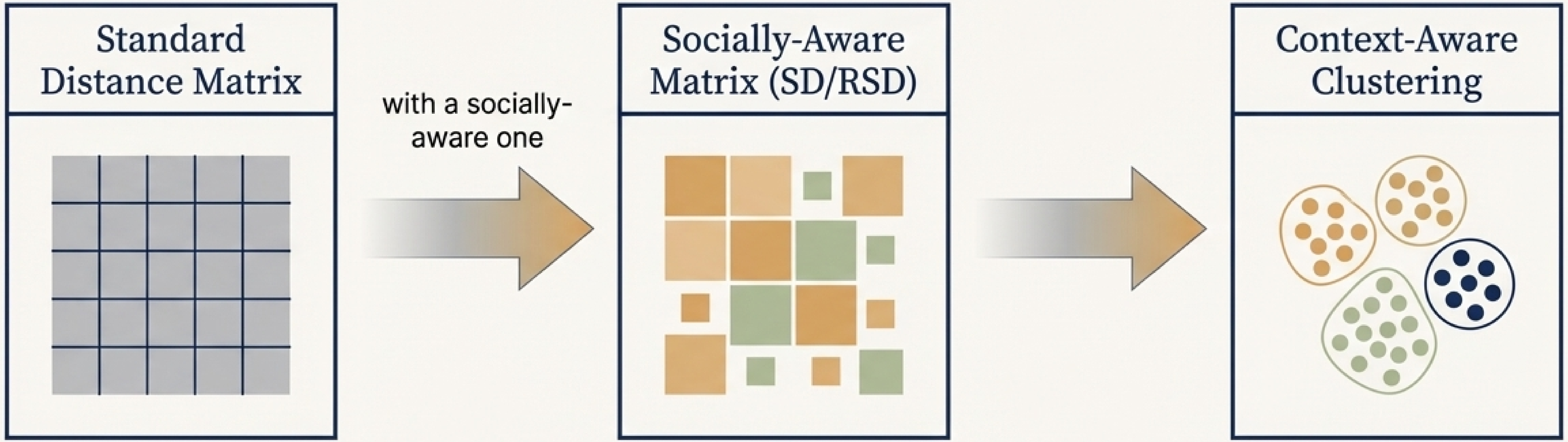
Clustering is a cornerstone of AI, essential for everything from exploratory data analysis to anomaly detection in Industry 4.0/5.0 systems.

However, classical algorithms overwhelmingly rely on standard geometric distances (e.g., Euclidean).

The Flaw: These metrics implicitly assume homogeneous data distributions and often fail to capture contextual relationships, local density variations, or structural dependencies present in real-world data.



The Solution: We Don’t Change the Algorithm; We Change What ‘Distance’ Means



Standard Distance Matrix

with a socially-aware one

Socially-Aware Matrix (SD/RSD)

Context-Aware Clustering

1. We replace the standard distance matrix...

2. The clustering algorithm (e.g., AP, OPTICS) runs as usual, but now operates on this enriched input.

3. The result is a clustering that reflects social embeddedness, not just geometric proximity.

Key Takeaway: This is an algorithm-agnostic enhancement. Any method relying on pairwise distances can be improved.

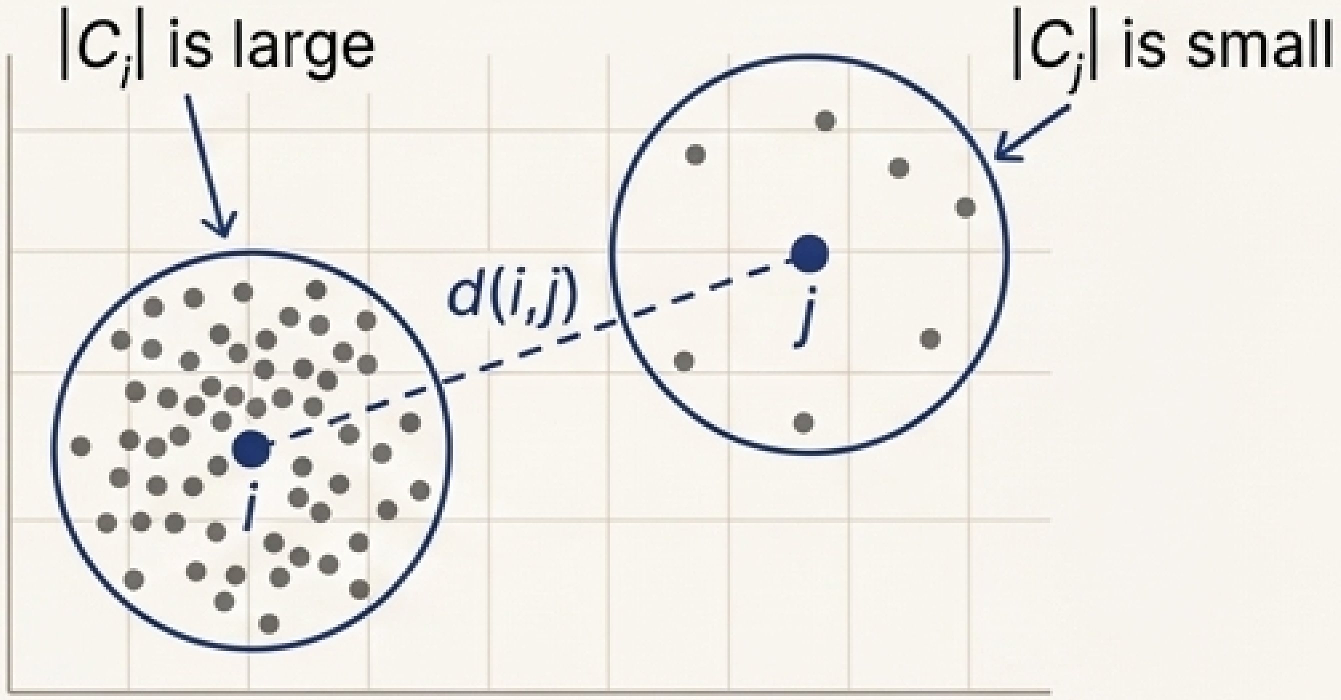
Defining Proximity by Neighborhood Size: Social Distance (SD)

Concept

For any two points *i* and *j*, their “social distance” is determined by comparing the number of other points in their immediate vicinity.

Methodology

1. **Define context areas**: For points *i* and *j*, create hyperspheres with radius $r = d(i,j)$.
2. **Count neighbors**: Let $|C_i|$ and $|C_j|$ be the number of points within each respective hypersphere.
3. **Aggregate counts**: Use a generalized mean to combine $|C_i|$ and $|C_j|$ into a single SD value.



$$SD(i,j) = \text{Lehmer mean}(|C_i|, |C_j|) = \frac{|C_i|^{k+1} + |C_j|^{k+1}}{|C_i|^k + |C_j|^k}$$

(An alternative Power mean can also be used).

The parameter *k* controls the bias:

- $k > 0$: Gives more weight to larger, denser neighborhoods.
- $k < 0$: Emphasizes smaller, sparser neighborhoods.

A Richer Context: Real Social Distance (RSD)

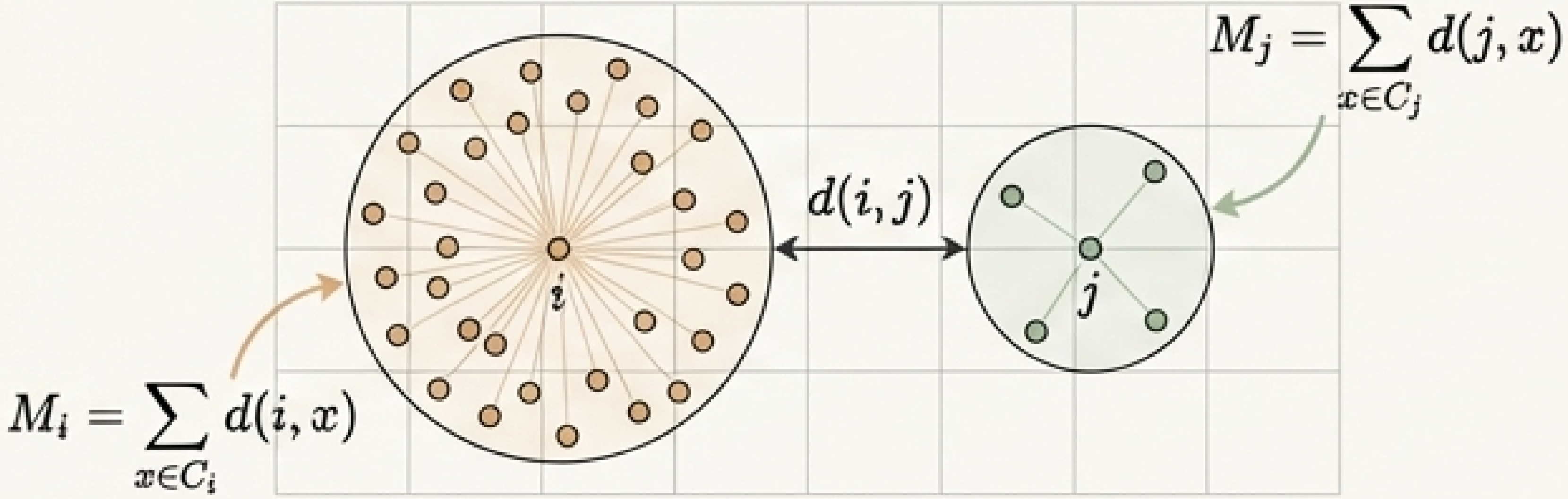
Concept

RSD generalizes SD by aggregating the *sum of distances* to neighbors within the context area, not just the count. This is called the “social mass.”

Key Insight: RSD accounts for both local density (the count) and the internal spatial arrangement of the neighborhood (the spread).

Methodology

1. **Define context areas as before** (hyperspheres of radius $d(i, j)$).
2. **Calculate social mass:**

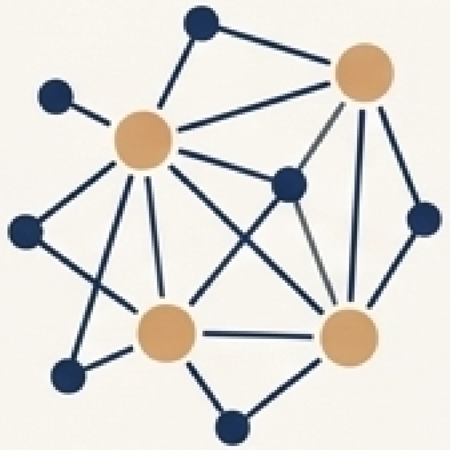

$M_j = \sum_{x \in C_j} d(j, x)$
$d(i, j)$
i
j
$M_i = \sum_{x \in C_i} d(i, x)$

3. **Aggregate masses**: Use the same generalized mean framework (Lehmer or Power) on M_i and M_j.

$$RSD(i, j) = \text{Lehmer mean}(M_i, M_j) = \frac{M_i^{k+1} + M_j^{k+1}}{M_i^k + M_j^k}$$

A Robust Testbed: Evaluating Across Two Complementary Paradigms

Affinity Propagation (AP)



Paradigm: Exemplar-based, global, similarity-driven.

Mechanism: Iteratively exchanges “responsibility” and “availability” messages to identify representative exemplars.

Why it’s a good test: AP's outcome is driven *entirely* by the pairwise similarity matrix. Changing the distance metric directly and transparently alters its behavior. With SD/RSD, exemplars become “socially central,” not just geometrically central.

OPTICS

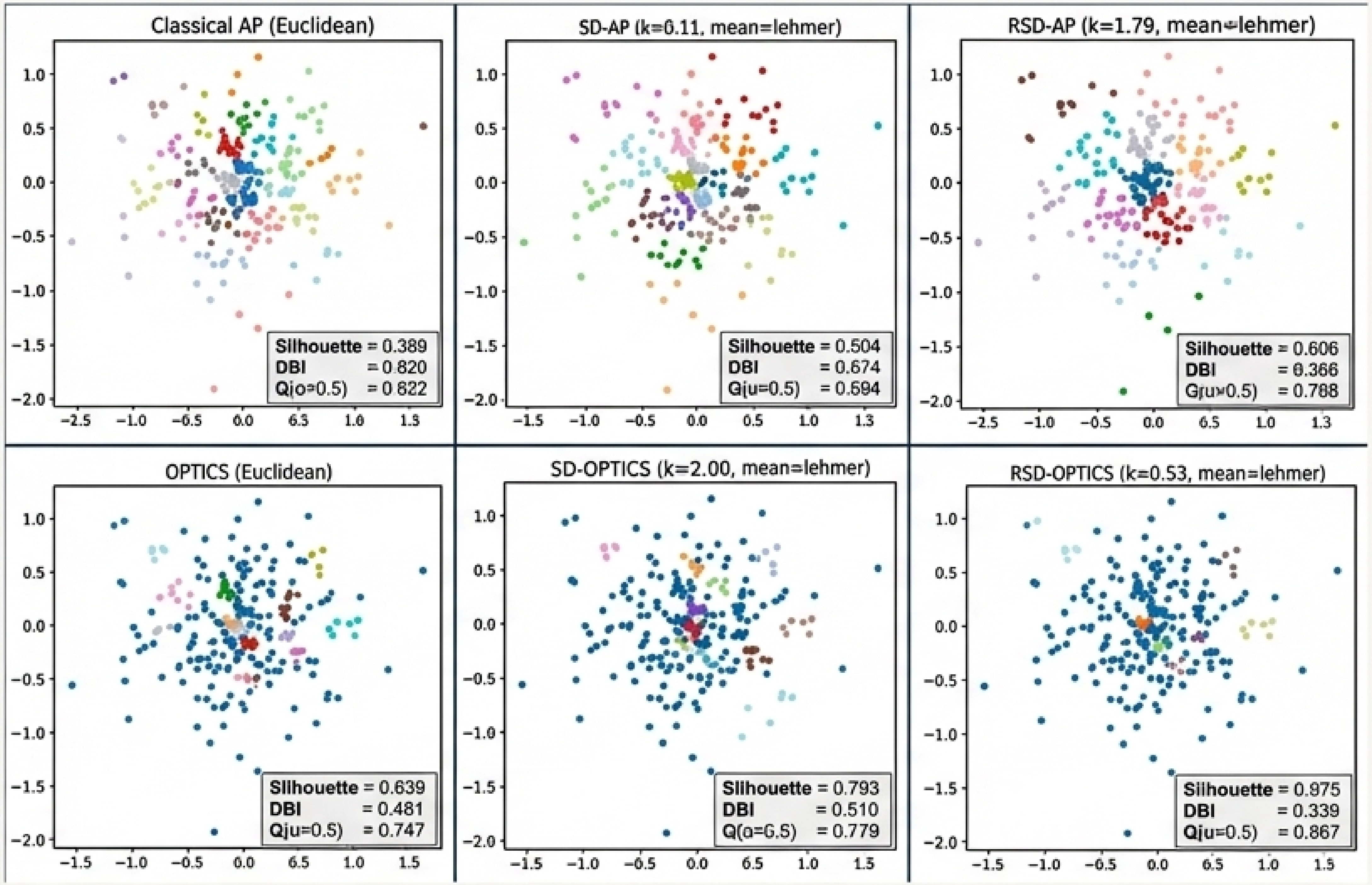
Paradigm: Density-based, local, hierarchical.

Mechanism: Generates an ordering of points based on “reachability distance,” encoding cluster structure across multiple density levels.

Why it’s a good test: OPTICS is critically dependent on the notion of distance and neighborhood. SD/RSD redefines density itself, leading to socially-mediated reachability and clearer separation across density gradients.

NotebookLM

The Impact in Focus: A Visual Comparison on a Complex Dataset



Key Observation: Both visually and quantitatively, SD and RSD enhancements systematically improve the clustering results for both AP and OPTICS. The improvement from RSD is particularly profound, achieving nearly perfect separation.

Social Distances Systematically Outperform the Baseline Across Datasets

Performance was aggregated over three synthetic datasets of increasing structural difficulty. The best configuration for SD/RSD was selected after a uniform sweep of the parameter k.

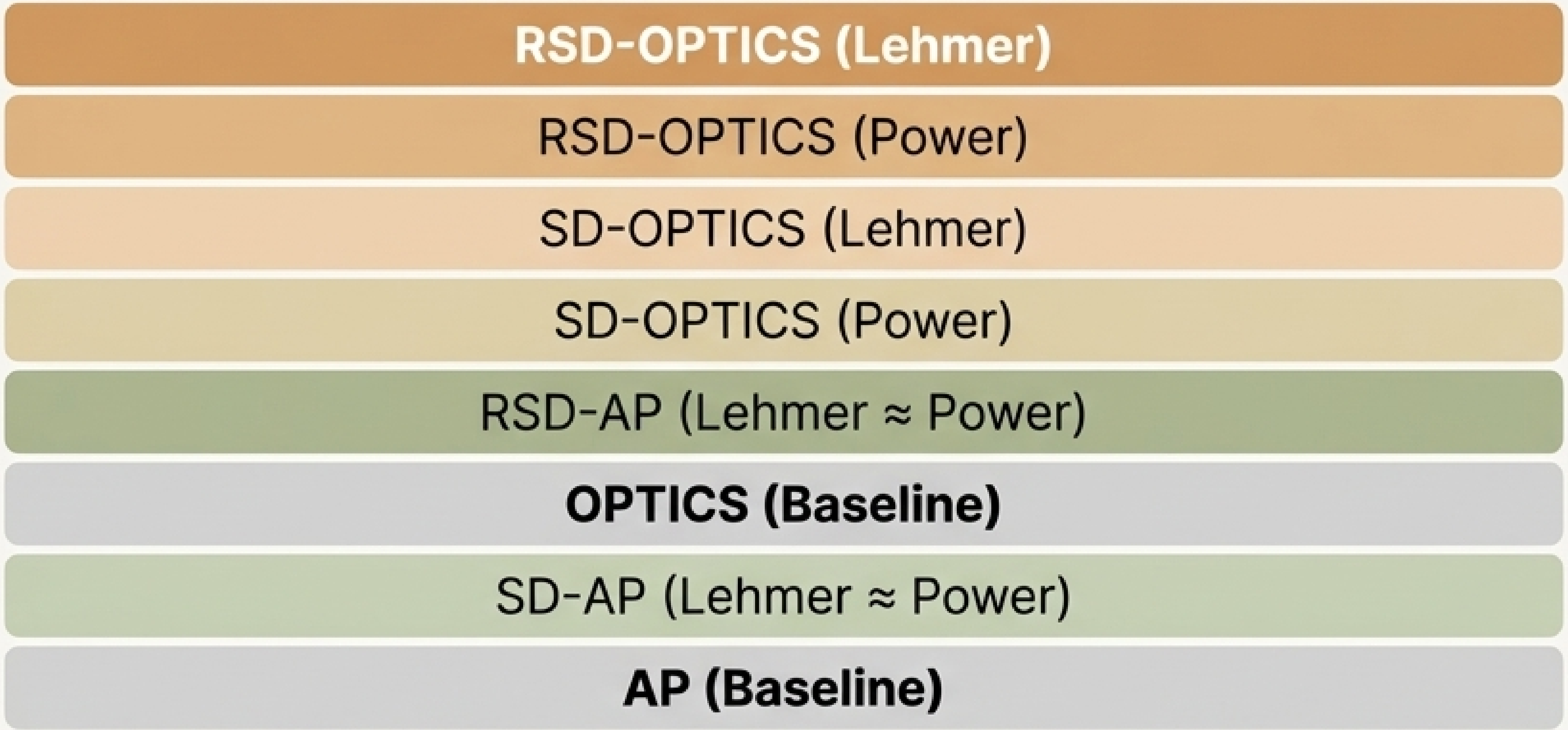
Method	Mean Option	Avg. Silhouette ↑	Avg. DBI ↓	Avg. Q(α) ↑
RSD-OPTICS	**Lehmer**	**0.955**	**0.349**	**0.859**
RSD-OPTICS	Power	0.908	0.447	0.823
SD-OPTICS	Lehmer	0.811	0.451	0.798
RSD-AP	Lehmer	0.595	0.361	0.766
OPTICS (Baseline)	-	*0.608*	*0.468*	*0.742*
AP (Baseline)	-	*0.371*	*0.726*	*0.626*

*(Showing top performers and baselines for clarity)

Key Findings:

- **Dominance**: RSD-OPTICS is the clear top performer.
- **Advantage of RSD**: RSD variants consistently outperform their SD counterparts for both algorithms.
- **Universal Improvement**: All SD/RSD variants outperform their respective classical baselines.

The Empirical Verdict: A Clear Performance Hierarchy Emerges



- **RSD-OPTICS (Lehmer)**
- RSD-OPTICS (Power)
- SD-OPTICS (Lehmer)
- SD-OPTICS (Power)
- RSD-AP (Lehmer ≈ Power)
- **OPTICS (Baseline)**
- SD-AP (Lehmer ≈ Power)
- **AP (Baseline)**

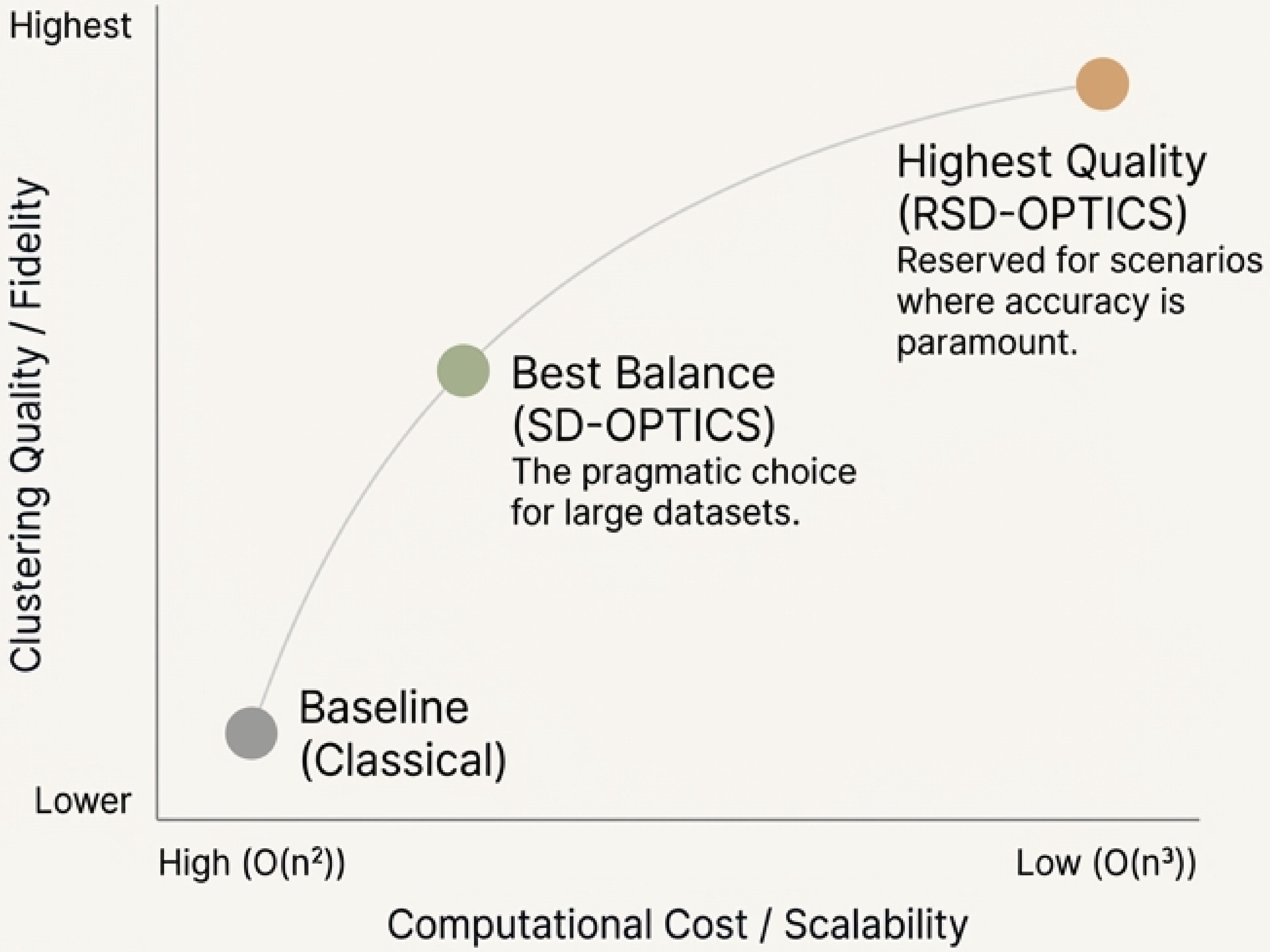
Social-context-aware distance semantics, particularly RSD, are the primary driver of clustering improvement—more influential than the choice of clustering algorithm itself.

The Practitioner’s View: The Effectiveness vs. Efficiency Trade-Off

How much computational overhead do these improvements introduce?

Complexity Analysis (for n samples)

- Baseline Clustering (AP/OPTICS): $O(n^2)$
- SD Construction: Adds an $O(n^2)$ overhead. Total complexity remains **$O(n^2)$**.
- RSD Construction: Adds an $O(n^3)$ overhead. Total complexity becomes **$O(n^3)$**.



The pragmatic choice for large datasets.

Reserved for scenarios where accuracy is paramount.

A Unique Contribution in Context-Aware Distances

How SD/RSD Differs from...

Metric Learning

SD/RSD are **unsupervised and deterministic**. They are constructed from intrinsic neighborhood relations, not learned from labeled data or pairwise constraints.

Algorithm-Specific Enhancements

SD/RSD are **algorithm-agnostic plug-ins**. Unlike methods that embed density into a specific clustering procedure, they redefine distance at a fundamental level, making them reusable.

Heuristic Weighting

SD/RSD use **principled aggregation theory**. The parameter *k* in the generalized means provides an explicit, interpretable control for biasing the metric, grounded in mathematics.

This work bridges the gap between theoretical distance design and practical clustering performance by empirically establishing the value of SD/RSD as general-purpose semantic amplifiers.

Conclusion: Context is the Key to More Meaningful Structure Discovery

Summary of Findings

- Enriching distance semantics with neighborhood context (SD/RSD) yields consistent, measurable improvements in clustering quality.
- This approach acts as a principled, interpretable, and algorithm-agnostic enhancement to classical methods.
- A clear efficiency-accuracy spectrum exists, with SD offering a scalable improvement and RSD providing maximum fidelity.

Future Outlook

- This work provides a practical pathway toward more context-sensitive unsupervised learning.
- Future research will focus on scalability (e.g., approximate RSD) and application to new domains like anomaly detection and representation learning.

As AI systems operate in increasingly complex and high-stakes environments, such distance semantics offer a promising foundation for more robust and meaningful data-driven discovery.