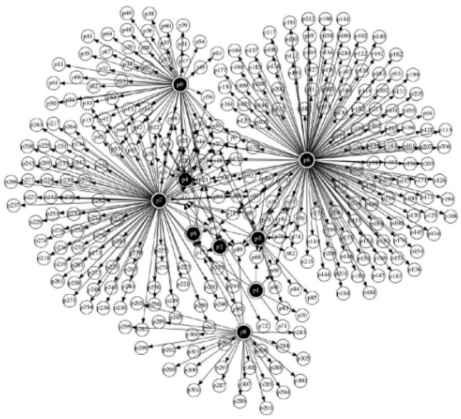
Tamr

White Paper / April 2025

Solving Real-world Entity Resolution at Scale with AI

$$\frac{n \times (n-1)}{2}$$



Abstract
Tamr's AI-driven entity resolution system offers a scalable, efficient solution for unifying large, diverse datasets. By employing specialized AI and patented techniques such as Cluster-Centric Entity Resolution, Blocking and Pre-Grouping, Tamr significantly reduces computational complexity and enhances processing speed. Unlike conventional systems, Tamr supports incremental learning and real-time feedback, enabling continuous adaptation to dynamic data environments. Its modular design makes it suitable for various use cases, including regulatory compliance, fraud detection, and customer insights. Tamr's approach provides a robust framework for achieving high-accuracy data unification at scale.

Contributors
Agarwal, R., Laferriere, N., Li, Q., Pagan, A., Parikh, T., Partakki, E., Slaughter-Katz, K.

1. Introduction

Entity Resolution (ER) is the process of identifying and linking records that refer to the same real-world entity across diverse datasets. As organizations accumulate increasingly large and diverse datasets from multiple sources, the challenge of entity resolution becomes more complex. Accurate and efficient ER is essential for various applications, including analytics, regulatory compliance, customer 360 views, and fraud detection.

Tamr, a leader in data mastering, addresses this challenge by integrating artificial intelligence (AI) into its entity resolution framework. These AI techniques enhance the overall process by autonomously managing data quality, scalability, and adaptive learning. This white paper presents a comprehensive view of how Tamr's AI-driven architecture overcomes the limitations of traditional ER systems, offering a scalable, accurate, and efficient solution.

2. Why Entity Resolution at Scale is Hard

Entity Resolution is inherently challenging due to several factors that compound as datasets grow larger. The most fundamental challenge is the **sheer volume of potential comparisons**. As datasets scale, the complexity of performing pairwise comparisons increases quadratically, making traditional comparison methods impractical for large datasets. As an example, when deduplicating 5 records there are 10 possible comparisons to make. If we double the number of records to 10, we more than quadruple the number of comparisons to 45. The number of comparisons to make quickly becomes challenging as data input size increases.

Additionally, the problem of **imbalanced classification** complicates the ER process. Matches are rare within large datasets, making it difficult to accurately detect duplicates or related entities. Conventional classification models often struggle with this imbalance, leading to higher rates of false positives and false negatives.

Another critical challenge is the **absence of ground truth**. In most cases, there is no definitive, pre-existing set of correct answers against which to compare the results of an ER process. This lack of reliable labels makes training and evaluating ER models particularly difficult.

The quality and diversity of data itself present further obstacles. Real-world data is often messy, incomplete, and inconsistently formatted across sources. Conflicting information is common, particularly when datasets are integrated from different systems.

Moreover, effective ER requires expertise from multiple domains, including data engineering, machine learning, and subject-matter knowledge. Combining these skills to build a robust, scalable solution is a complex and resource-intensive task.

3. Tamr's AI-Driven Approach to Solving ER

Tamr addresses the inherent complexity of entity resolution by employing a sophisticated AI-driven architecture that integrates multiple specialized AI techniques. These techniques enhance efficiency, accuracy, and adaptability by performing various tasks autonomously and collaboratively. The AI is designed to continuously learn from incoming data, interact with human experts, and optimize computational resources.

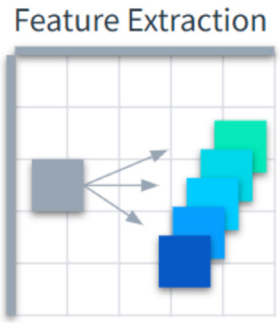
The core components of Tamr's system include:

- **Feature Extraction**
- **Blocking and Pre-Grouping**
- **Enrichment**
- **Pairwise Classification and Clustering**
- **Adaptive Learning and Categorization**
- **Semantic (Embedding) Search**

By embedding AI throughout its workflow, Tamr ensures that computational resources are used efficiently while maintaining high levels of accuracy. The system's modular architecture allows each component to specialize in specific tasks, enabling the overall process to scale effectively.

4. Key Components of Tamr's AI System

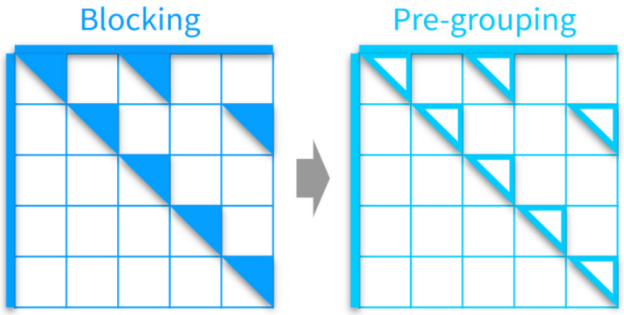
Feature Extraction



The feature extraction process involves transforming raw data into representations suitable for machine learning. Tamr's Feature Extraction uses advanced text processing techniques such as tokenization, lemmatization, stemming, and bigram splitting to improve record comparison. Additionally, these also apply numerical processing techniques, including absolute and relative differences, and handle geographical data using Hausdorff distance metrics.

To enhance efficiency, Feature Extraction converts text tokens to hashed values, reducing computational overhead. It is capable of autonomously detecting schema changes and adjusting extraction processes accordingly.

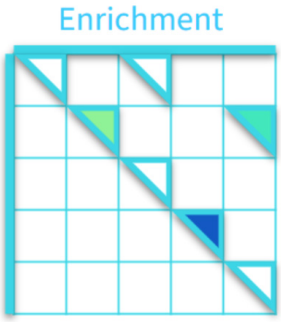
Blocking and Pre-Grouping



Tamr tackles the challenge of quadratic complexity by implementing intelligent blocking and pre-grouping mechanisms managed by specialized AI. Blocking techniques group similar records into overlapping blocks, allowing comparisons to be performed only within these blocks. This approach reduces complexity from $O(N^2)$ to $O(N)$, making large-scale ER feasible.

Pre-Grouping, on the other hand, efficiently identifies functionally identical records and aggregates them before the main resolution process. This approach achieves O(N log N) complexity and significantly reduces the number of comparisons required. It continuously optimizes blocking boundaries based on incoming data characteristics and feedback from classification. Tamr's patented approaches to scalable binning and pre-grouping are central to its ability to efficiently handle massive datasets **(US Patent 10,613,785; US Patent 11,204,707)**.

Enrichment



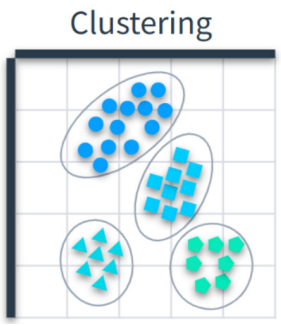
Enrichment is designed to enhance Tamr's entity resolution process by validating, standardizing, and enriching data attributes with high-quality external information. This process improves matching precision by augmenting existing records with contextual data from third-party sources such as publicly available datasets, proprietary business information, and industry-specific directories. The enrichment process begins with validating and standardizing incoming data to ensure consistency and compatibility with Tamr's matching models. Once standardized, external data is integrated to enhance entity profiles, improving accuracy for feature extraction, classification, and clustering.

Enrichment supports Tamr's patented techniques for Cluster-Centric Entity Resolution **(US Patents 10,613,785; 11,204,707)** and Incremental Learning **(US Patent 11,003,636)** by incorporating enriched attributes directly into the entity resolution pipeline. This integration allows AI to continuously improve their performance as new data becomes available, ensuring adaptability and accuracy in dynamic environments.

Pairwise Classification and Clustering

Pairwise Classification is responsible for predicting whether two records refer to the same entity. It uses a Random Forest Classifier, trained on labeled data provided by domain experts. Active learning techniques highlight uncertain cases for expert review, enabling the model to improve over time.

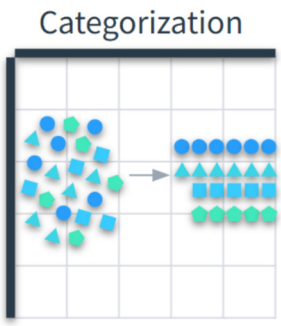
After classification, Clustering utilizes a Bottom-Up Average Linkage Clustering algorithm (WPGMA) optimized to reduce complexity from $O(N^2 \log N)$ to $O(E \log E)$. These techniques ensure efficient clustering through parallel processing, making them suitable for batch, incremental, and streaming workloads. This clustering approach is supported by patented methods focused on cluster-centric entity resolution **(US Patent 11,321,359).**



Adaptive Learning and Categorization

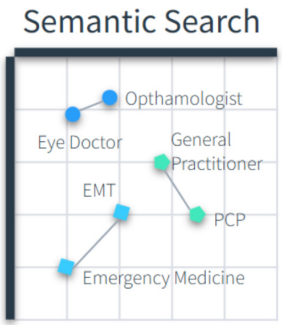
Tamr's AI-driven architecture incorporates Adaptive Learning and Categorization designed to enhance the system's ability to continuously improve and organize data effectively. Adaptive Learning enables Tamr to integrate new data without requiring full retraining, ensuring that models remain accurate and efficient even as datasets evolve. By leveraging incremental learning techniques, protected under **US Patent 11,003,636**, Tamr can rapidly detect changes in data patterns and optimize their predictions in real-time. This approach minimizes downtime and ensures consistent high accuracy across various data integration tasks.

Categorization complements this process by employing hierarchical classification models that intelligently group records into predefined taxonomies. The ability to efficiently categorize data at scale is supported by **Tamr's scalable schema mapping techniques**, covered under **US Patent 10,860,548**. These techniques use logistic regression ensembles and active learning techniques to enhance classification accuracy over time, making them highly effective at processing dynamic and large-scale datasets. Tamr's categorization enables high accuracy feature extraction from long text fields and mis-mapped data. It is further leveraged to select the appropriate Entity Resolution model to apply at the record level. Tamr's integration of adaptive learning and categorization allows the system to seamlessly adapt to evolving requirements, providing a flexible framework for achieving reliable data unification.



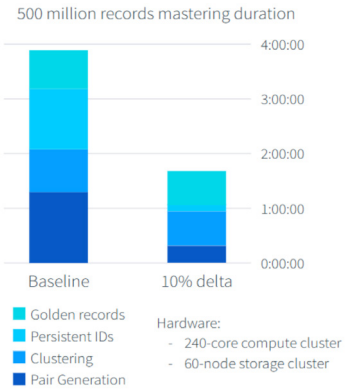
Semantic (Embedding) Search

Tamr incorporates AI-powered semantic search to enhance matching accuracy by providing contextual understanding and improving the resolution process. Unlike traditional keyword-based search mechanisms that rely solely on lexical matches, Tamr's approach leverages semantic search to capture deeper relationships between entities through meaning, intent, and context. By integrating pre-trained machine learning models and contextual data enrichment, Tamr's architecture delivers high-quality matches with minimal duplicates, even when dealing with complex or multilingual data sources. Additionally, the feedback mechanisms within Tamr's system continuously refine the search process, ensuring adaptability and precision as data evolves. This integration of semantic search enhances Tamr's ability to resolve entities accurately and at scale, providing a more robust framework for various use cases such as compliance monitoring, fraud detection, customer intelligence, and product categorization.



5. Efficiency, Scalability, and Adaptability

Tamr's AI-driven architecture demonstrates exceptional efficiency and scalability. For example, using a dataset covering the entire U.S. population, Tamr's AI for blocking, pre-grouping, classification and clustering, reduces the number of comparisons from 250 quadrillion to approximately 800,000 pairs. This approach enables the system to process 500 million records in under four hours.



The adaptability of Tamr's architecture is also critical to its success. The use of adaptive learning and incremental processing allows the system to remain accurate even as new data continuously flows in. Unlike conventional systems that require complete retraining to incorporate new records, Tamr's AI can update their models in real-time. Tamr's patented techniques for continuous learning and scalable schema mapping are critical to this adaptability **(US Patent 11,003,636; US Patent 10,860,548)**.

6. Results, Impact, and Future Development

Tamr's AI-driven approach to entity resolution offers a highly efficient, scalable, and accurate solution for unifying large and diverse datasets. By employing specialized AI throughout the workflow—feature extraction, blocking, classification, clustering, and consolidation—Tamr provides exceptional accuracy and adaptability while maintaining fast processing speeds.

Proven Impact

Tamr's patented techniques, including Cluster-Centric Entity Resolution, Cluster-Centric Accuracy Metrics, and pre-grouping methods **(US Patents 10,613,785; 11,204,707)**, drastically reduce the number of comparisons from hundreds of trillions to manageable levels. Efficient pre-grouping of "obvious matches" further enhances speed and accuracy by streamlining the resolution process.

The architecture's incremental learning capabilities, supported by scalable schema mapping and real-time learning patents **(US Patents 11,003,636; 10,860,548)**, allow Tamr to continually adapt and optimize its models based on evolving datasets. Unlike traditional systems that require periodic retraining, Tamr's AI autonomously detects changes, applies real-time feedback, and enhances prediction accuracy through continuous learning.

This modular design is applicable across various industries, including regulatory compliance, fraud detection, customer insights, and supply chain management. Organizations can customize workflows to meet specific requirements, enhancing efficiency and reliability.

Future Opportunities

The future of entity resolution will be shaped by continued advances in AI, particularly in the areas of semantic understanding, adaptive workflows, and autonomous decision-making. Tamr's AI-native architecture is well-positioned to incorporate these trends, expanding the boundaries of what's possible in data mastering.

One major area of innovation is the integration of Agentic AI—autonomous agents capable of making proactive decisions about how and when to resolve data. These agents will drive more intelligent orchestration across workflows, identifying anomalies, retraining models as needed, and optimizing deduplication strategies in real time.

The increasing use of Large Language Models (LLMs) will further augment semantic understanding, allowing entity resolution systems to recognize equivalence across different languages, formats, and business contexts. Combined with Tamr's current strengths in enrichment and categorization, LLMs can help move beyond structured matching toward deeper contextual linking.

Likewise, advances in clustering algorithms and incremental learning will enable real-time updates to entity groupings as new data arrives, without compromising performance. Tamr's architecture already supports these capabilities, and continued refinement will drive faster, more accurate decisions at scale.

To support the emergence of intelligent, agent-based systems, Tamr sees strong potential in adopting frameworks like the **Model Context Protocol (MCP)**. MCP defines a shared interface between AI models and the context in which they operate—such as domain-specific rules, human feedback, and business logic. This enables models to reason not just about the data, but about **how they are being used**.

By integrating MCP into Tamr's AI system, each agent can access real-time context about the broader resolution task it's contributing to, improving decision quality and coordination. Whether resolving entities, categorizing records, or enriching fields, MCP allows models to dynamically adapt to evolving business needs—without brittle, hardcoded logic.

7. Conclusion

Tamr's AI-driven architecture offers a comprehensive solution to the complex challenges of entity resolution. Its innovative use of patented techniques for clustering, pre-grouping, and incremental learning ensures high accuracy and efficiency even when processing vast amounts of data.

Through continuous learning and real-time feedback, Tamr maintains the relevance and effectiveness of its models over time. The architecture's flexibility and scalability make it suitable for various industries and use cases, including regulatory compliance, fraud detection, and customer insights.

By embracing advancements such as Model Context Protocol, improved clustering algorithms, Large Language Models, and Agentic AI, Tamr is well-positioned to remain a leader in the field of entity resolution. Its architecture provides a robust foundation for delivering reliable, high-quality data unification at scale.

8. References

- Dong, X. L., & Srivastava, D. (2013). **Big Data Integration**. Morgan & Claypool Publishers.
- Christen, P. (2012). **Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.** Springer Science & Business Media.
- Papadakis, G., et al. (2020). **The Four Generations of Entity Resolution: A Survey from Big Data to AI.** ACM Computing Surveys (CSUR), 53(6).
- US Patents: 10,613,785; 11,204,707; 11,321,359; 11,003,636; 10,860,548.

About Tamr, Inc.

Tamr provides the only AI-native master data management (MDM) solution that delivers real-time master data for every dashboard, application, and person in your business. Tamr accelerates the discovery, enrichment, and maintenance of Golden Records, enabling informed decision-making, improved revenue growth, and better customer experiences. Tamr's patented, AI-centric approach – with human refinement and oversight – delivers value in days, not months or years like traditional rules-based MDM and DIY solutions. For more information on Tamr, please visit https://www.tamr.com