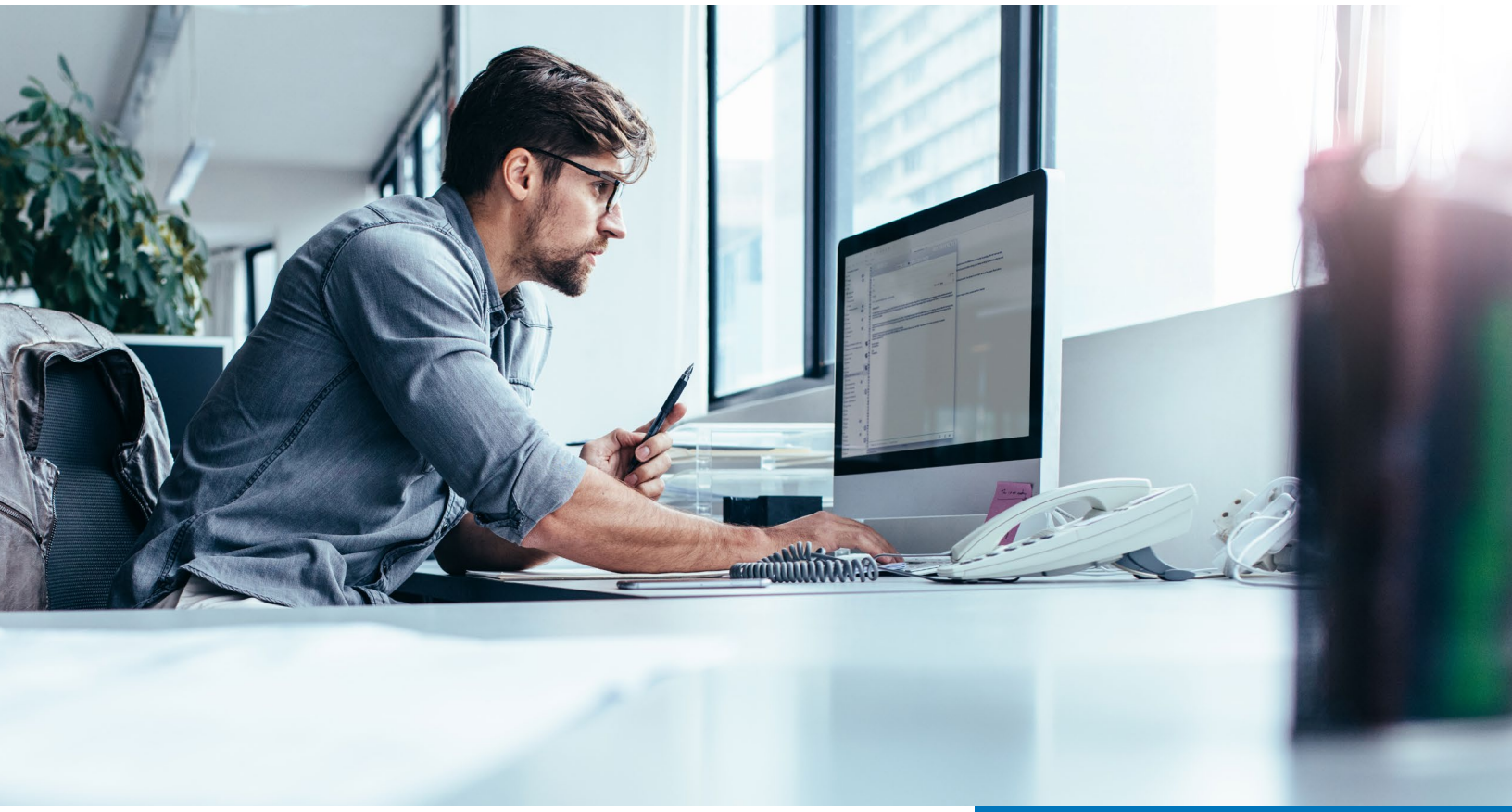
Resultant

Scalable Probabilistic Record Linkage

USING DATA FOR EFFECTIVE DECISION MAKING



Before data can be used to provide actionable insights, the barriers that prevent data from becoming actionable insights must be broken down.

CONTENTS

INTRODUCTION

Today, businesses understand the importance of leveraging data for effective decision making. Before data can be used to provide actionable insights, the barriers that prevent data from becoming actionable insights must be broken down. A primary barrier to actionable insight is combining and analyzing data from disparate sources. Data is often stored in silos within various business departments, across government agencies, or commonly as a result of a merger or acquisition. Even after the sometimes cumbersome procedural, organizational, and legal issues have been addressed, the technical challenges associated with combining data remain. This challenge requires record linkage.

As data continues to grow at an exponential pace, many record linkage solutions fail to scale to hundreds of millions of records or be readily maintained. The systems are incapable of generalizing as new data silos are incorporated and often rely on a complex set of business rules.

In these situations, a unique approach to scalable, probabilistic record linkage is required.

A BRIEF HISTORY: RECORD LINKAGE

The realization of the record linkage problem dates all the way back to 1946 when Halbert L. Dunn first introduced the concept in the article "Record Linkage" published by the American Journal of Public Health1. Upon further exploration by Ivan Fellegi and Alan Sunter in 1969, "A Theory for Record Linkage," was published by the Journal of the American Statistical Association. The article, which remains the foundation for many record linkage applications today, defined record linkage as "a solution to the problem of recognizing those records in two files which represent identical persons, objects, or events (said to be matched)"2. Today, record linkage is characterized as "data matching," "de-duplication," or "record matching," and used by data scientists, statisticians, historians, epidemiologists, computer scientists, and more.

Today, with the proliferation of data collection, historical solutions often fail to meet modern demands.

> "Recognizing those records in two files which represent identical persons, objects, or events said to be matched.
>
> **IVAN FELLEGI AND ALAN SUNTER**
> *A THEORY FOR RECORD LINKAGE*

DEFINING PROBABILISTIC RECORD LINKAGE

The objective of record linkage is to develop a comprehensive view of all relevant information pertaining to the same entity, whether a person, business, or event.

More specifically, when joining data within or across databases, exact matches of table keys are not always possible. However, individuals often have a unique identifier, known as personally identifiable information (PII), such as their social security number or driver's license ID. If two records contain a common key, the records can be joined by exact matching. In the absence of a unique identifier, records are linked based on the similarity of PII and/or other additional information.

In the presence of missing keys, probabilistic record linkage possesses distinct advantages over deterministic matching of PII due to:

- Data quality issues (e.g. typos/misspelling, missing or extra letters)
- Data incompleteness (e.g. last four of social, year of birth, middle initial)
- PII mismatch (e.g. one database collects date of birth while another collects age)
- Lifestyle changes in PII (e.g. marriage, change of address)

Today, systems have been built to aid in record linkage. However, these systems often have issues when confronted with a lack of PII overlap between records. In instances such as this, Resultant has developed a unique methodology which combines a custom record linkage system with refined processes to probabilistically link large-scale data sets. New data silos with varying PII fields populated are automatically incorporated, all while optimizing thresholds within every relationship level, whether at the data silo or table level. When new data silos are added to the system, the algorithm leverages this information to not only form new linkages but add additional power and information to existing matches.

APPLICATIONS OF LARGE-SCALE PROBABILISTIC RECORD LINKAGE

Large-scale record linkage can be used in various use cases when data silos evolve and grow separately, such as state government agencies with separate funding sources, or through business mergers and acquisitions. Below are a few examples of when large-scale record linkage is needed for success.

HEALTHCARE

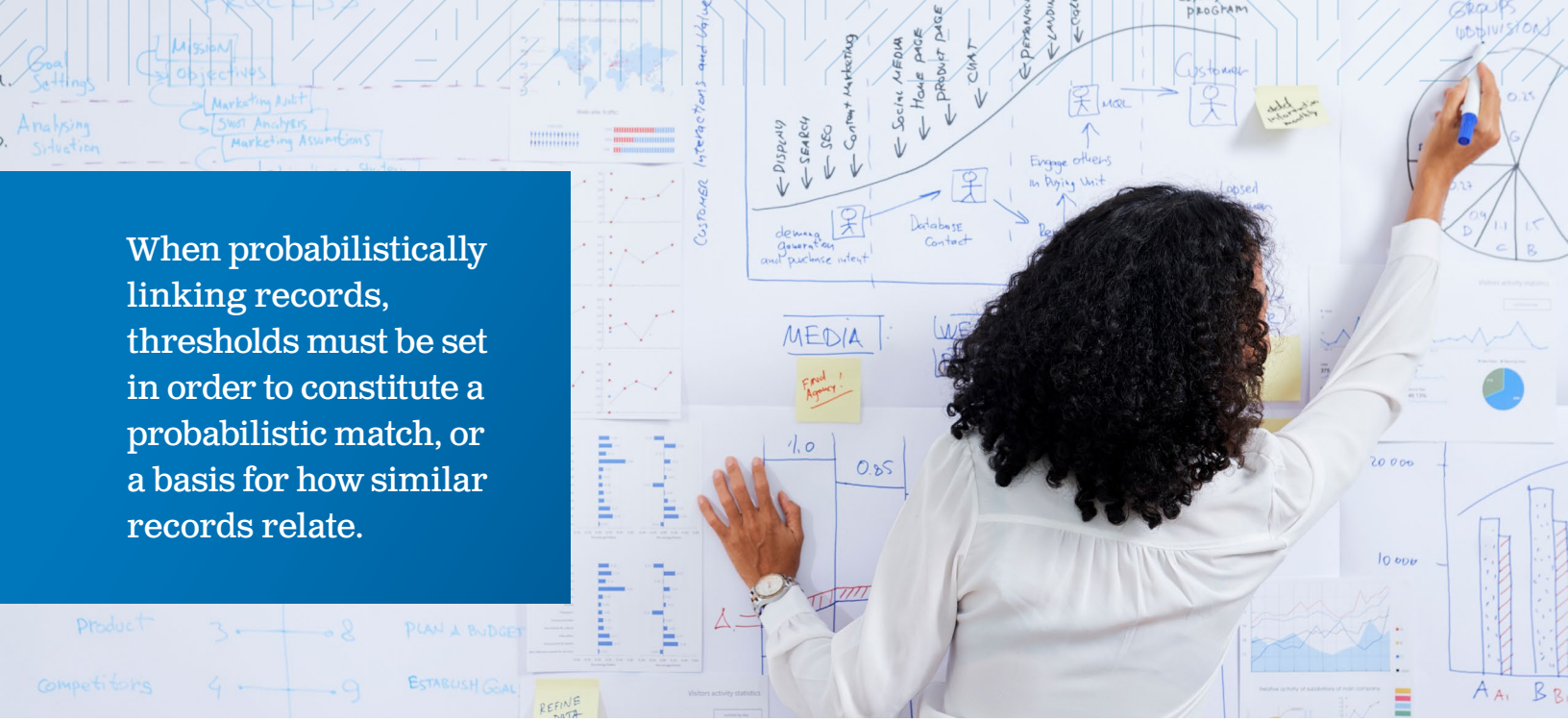
In recent years, hospital mergers can be found across the nation. Each hospital has hundreds of thousands of patients. When two systems are integrated, it is important for healthcare systems and doctors to gain a complete view of a patient's medical history.

GOVERNMENT

Within state government, many agencies, such as the Department of Workforce Development, Department of Corrections, or the Department of Education, retain separate records. To solve large- scale challenges, such as infant mortality, child well-being, unemployment, and criminal recidivism, disparate data sets must be combined to aggregate the data needed to enable data-driven decision making.

ENTERPRISES

Large enterprises often grow via mergers and acquisitions. When these mergers and acquisitions take place, it becomes necessary to merge company data. For instance, if two major air carriers merge, they may have loyalty programs in place in which varying customer information has been obtained. To link individuals across the programs into one database, record linkage must be used.



RECORD LINKAGE CONSIDERATIONS

When linking records, there are three common problems that arise: scalability, thresholds, and flexibility.

SCALABILITY

When combining data silos, the goal is to link tens of millions of records that live in numerous tables and span multiple silos. In probabilistic record linkage, organizations desire to combine large data sets, but fail to possess a single identifier that links records together.

Solution: The system developed by Resultant has the ability to scale to incorporate data records all while maintaining robust record linkages.

SETTING THRESHOLDS

When probabilistically linking records, thresholds must be set in order to constitute a probabilistic match, or a basis for how similar records relate. It is imperative to have an automated process in place to minimize error both within and across data silos.

Solution: Probabilistically linking within a data silo, called de-duplication, is separate from matching across data sets, but each technique can be performed using a similar process. Resultant has developed an automated process that minimizes the error of misclassification of false positives and false negatives. Furthermore, the Resultant record linkage solution conforms to all data environments as thresholds are optimized at the unique table relationship level.

FLEXIBILITY

Regardless of the record linkage tool selected, the tool should be flexible to changes in the existing data or additions of new data. For instance, if additional fields or silos are added, the record linkage algorithm should automatically detect and update for this new situation. That is, the selected algorithm must be flexible and functional when adapting to the addition or removal of new data tables and PII. Furthermore, the system should use the additional information provided by new data sources to further refine and improve the existing matches.

Solution: Resultant has developed a flexible system that automatically updates to reflect additional or removed data.

SCALABLE PROBABILISTIC RECORD LINKAGE METHODOLOGY

The Resultant data analytics team, finding current solutions were ineffective for today's complex data records when reconciling commonality across data sets, developed a unique system to probabilistically link records. The system, combined with Resultant processes, reflects the methodology the team has refined over time.

This methodology allows businesses with disparate data and varying subsets of PII to find relationships across the data by refining the system to accurately match typos, transpositions, and missing information. Additionally, the system has the power to match a certain subset of PII with an individual with an entirely different PII, and then use the existing record as a potential link to a third person with a completely different subset of PII. Even with multiple data sets that do not contain the same PII, the system is capable of linking records. As new data is brought into the system it updates automatically to minimize error within the system as well as identify duplicates within data silos.

Resultant has experience assisting enterprises with breaking down data silos through scalable probabilistic record linkage to help turn data into actionable information.

[1] American Journal of Public Health: Record Linkage, 1946

[2] Journal of the American Statistical Association: A theory for record linkage, 1969

ABOUT RESULTANT

Our team believes solutions are more valuable, transformative, and meaningful when reached together. Through outcomes built on solutions rooted in data analytics, technology, and management consulting, Resultant serves as a true partner by solving problems with our clients, rather than for them.

DATA ANALYTICS

We help organizations understand their data landscape and solve problems by turning data into insight. While data can be dense, our team's empathetic approach to problem solving creates meaningful solutions with deep technical foundations.

Learn more about Resultant data analytics services.

VISIT **RESULTANT.COM**