



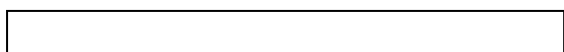
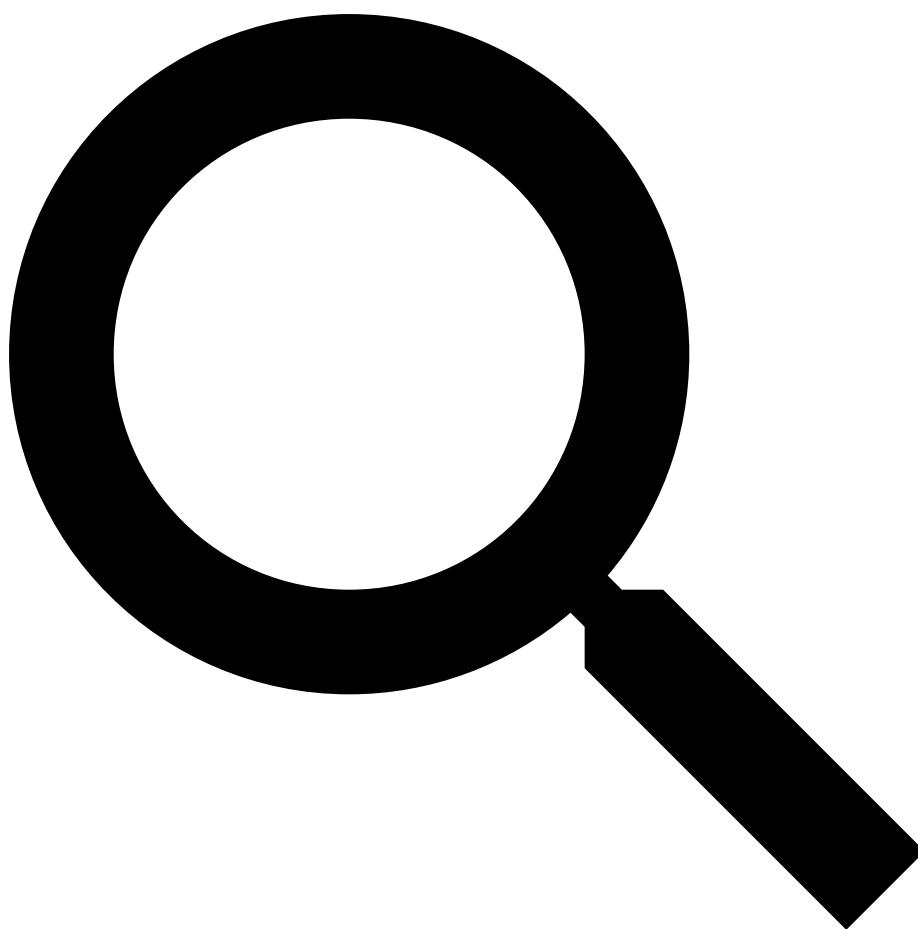
[Skip to content](#)

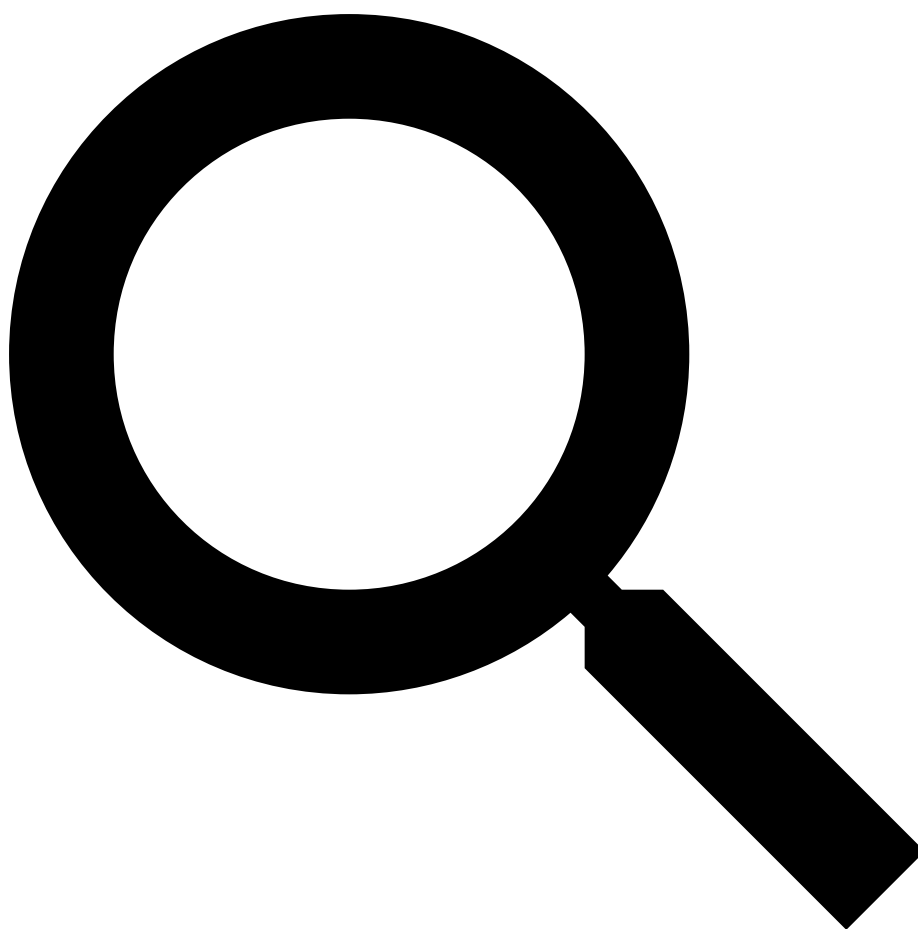
Still using Splink 3 and looking for the old docs? You can find them [here](#)
[logo](#)

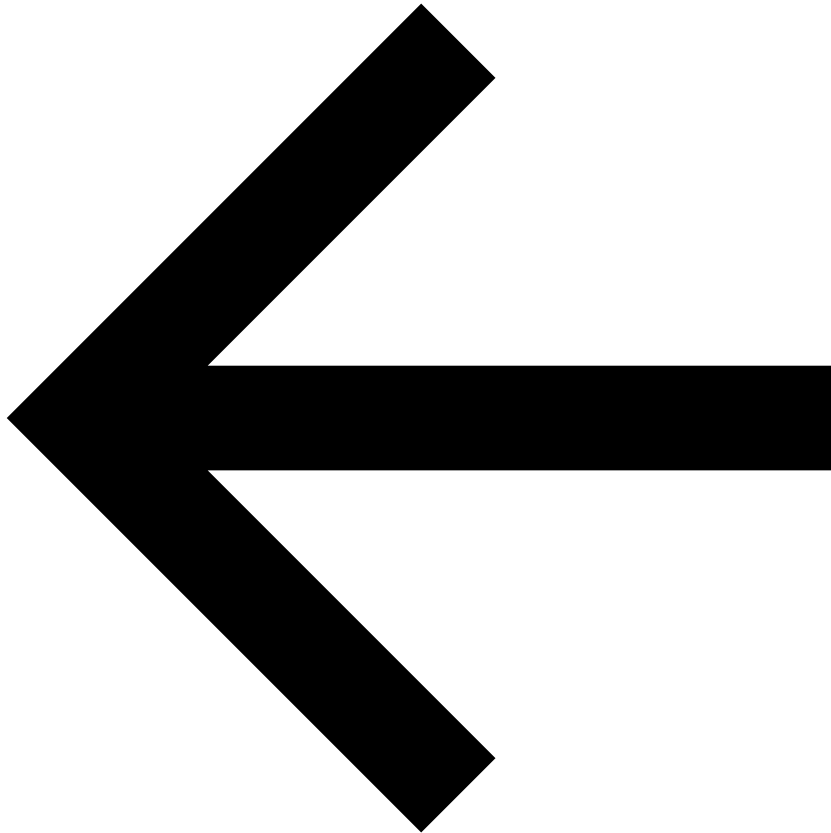


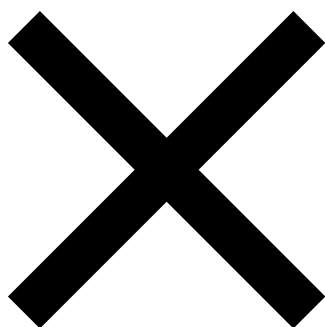
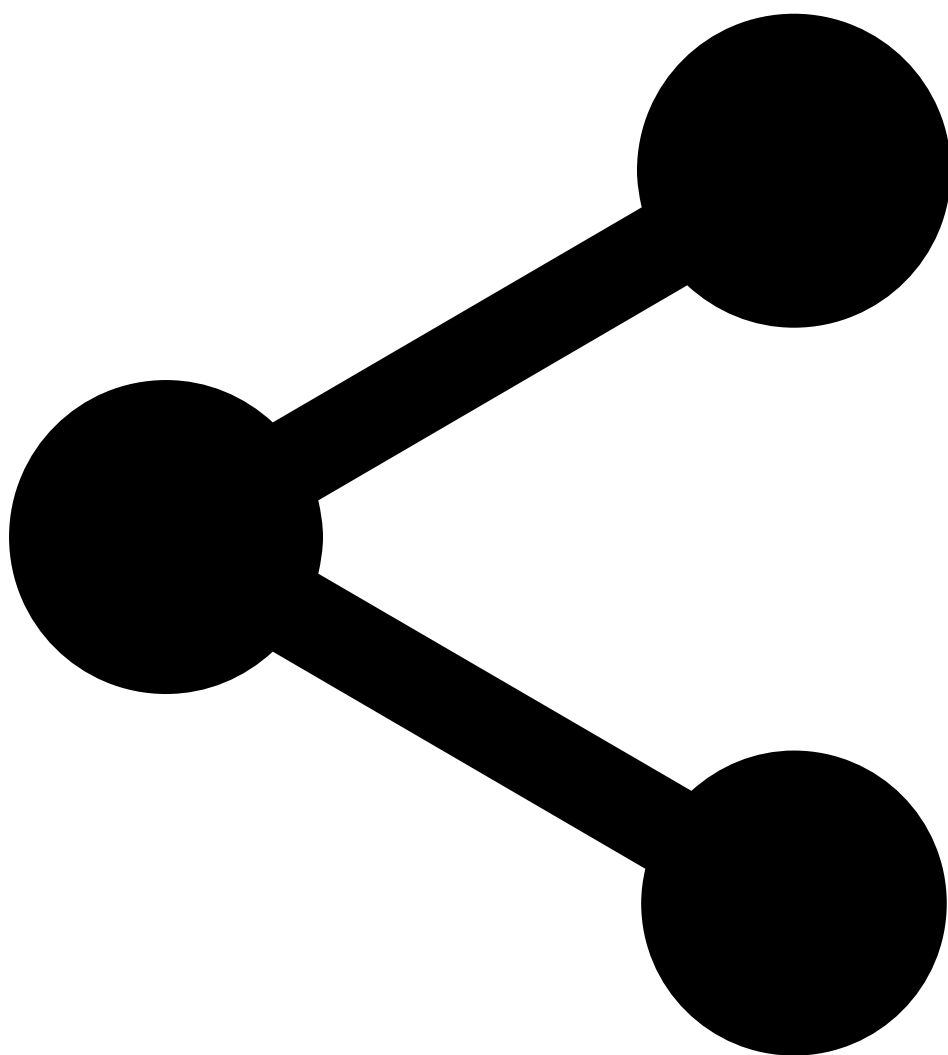
Splink
The Fellegi-Sunter Model



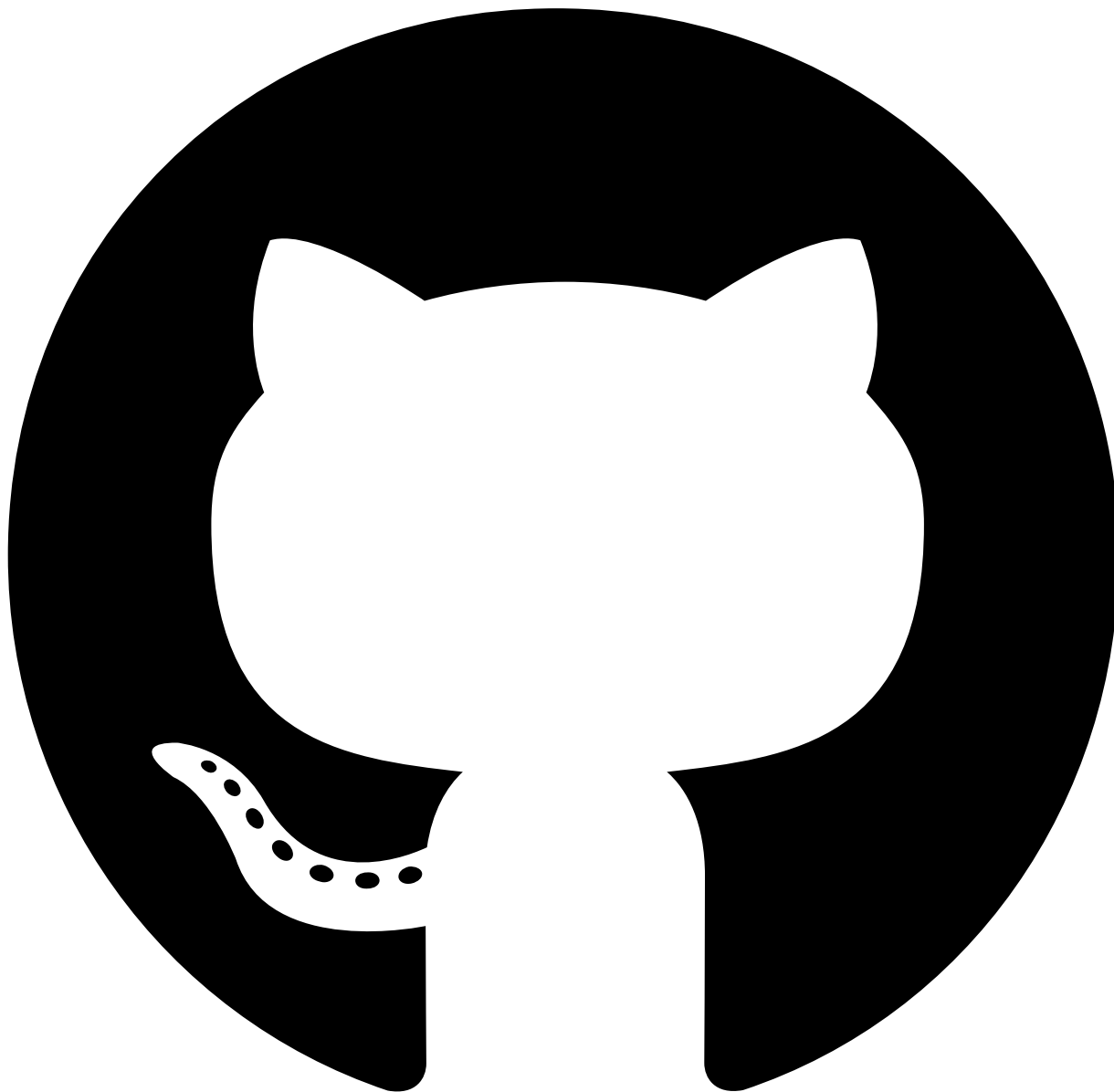








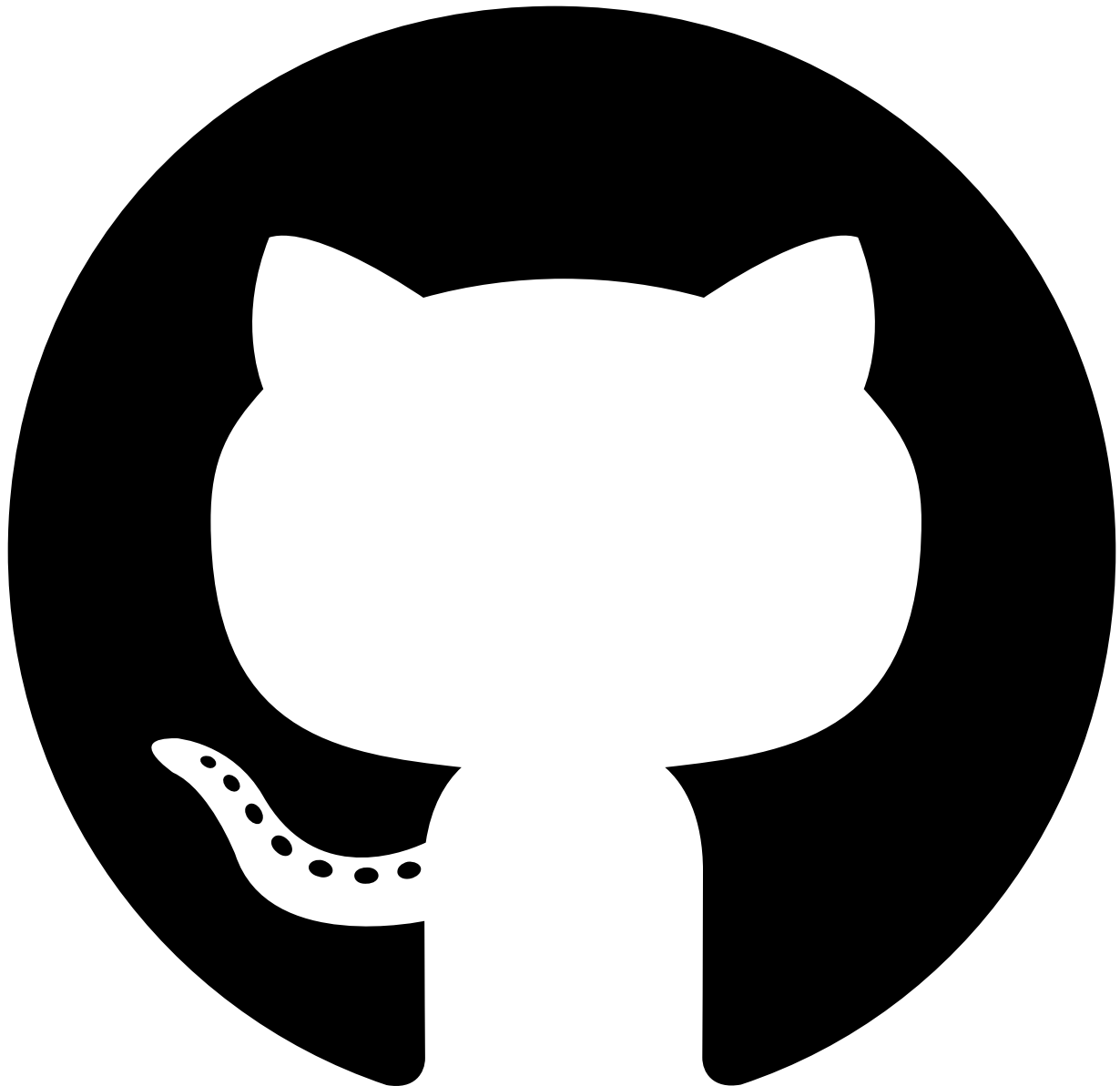
Initializing search



[GitHub](#)

- [Getting Started](#)
- [Tutorial](#)
- [Examples](#)
- [API Docs](#)
- [User Guide](#)
- [Contributing](#)
- [Blog](#)

[logo](#) Splink



[GitHub](#)

- [Getting Started](#)
- ☐ Tutorial
 - Tutorial
 - [Introduction](#)
 - [1. Data prep prerequisites](#)
 - [2. Exploratory analysis](#)
 - [3. Blocking](#)
 - [4. Estimating model parameters](#)
 - [5. Predicting results](#)
 - [6. Visualising predictions](#)
 - [7. Evaluation](#)
 - [8. Tips for building your own model](#)
- ☐ Examples
 - Examples
 - [Introduction](#)

- ☐
 - DuckDB
 - DuckDB
 - [Deduplicate 50k rows historical persons](#)
 - [Linking financial transactions](#)
 - [Linking businesses](#)
 - [Linking two tables of persons](#)
 - [Real time record linkage](#)
 - [Evaluation from ground truth column](#)
 - [Estimating m probabilities from labels](#)
 - [Quick and dirty persons model](#)
 - [Deterministic dedupe](#)
 - [Febrl3 Dedupe](#)
 - [Febrl4 link-only](#)
 - [Cookbook](#)
 - [Investigating Bias in a Splink Model](#)
 - [Comparison playground](#)
 - [Pseudopeople Census to ACS link](#)
- ☐
 - PySpark
 - PySpark
 - [Deduplication using Pyspark](#)
- ☐
 - Athena
 - Athena
 - [Deduplicate 50k rows historical persons](#)
- ☐
 - SQLite
 - SQLite
 - [Deduplicate 50k rows historical persons](#)
- ☐
 - API Docs
 - API Docs
 - [Introduction](#)
 - ☐
 - Linker
 - Linker
 - [Training](#)
 - [Visualisations](#)
 - [Inference](#)
 - [Clustering](#)
 - [Evaluation](#)
 - [Table Management](#)
 - [Miscellaneous functions](#)
 - ☐
 - Comparisons Library
 - Comparisons Library
 - [Comparison Library](#)
 - [Comparison Level Library](#)
 - ☐
 - Other
 - Other
 - [Exploratory](#)
 - [Blocking analysis](#)
 - [Blocking](#)

- [Clustering](#)
 - [SplinkDataFrame](#)
 - [EM Training Session API](#)
 - [Column Expressions](#)
 - ☐
 - In-build datasets
 - In-build datasets
 - [SplinkDatasets](#)
 - ☐
 - Splink Settings
 - Splink Settings
 - [Settings Dict](#)
- ☒
 - User Guide
 - User Guide
 - [Introduction](#)
 - ☒
 - Record Linkage Theory
 - Record Linkage Theory
 - [Why do we need record linkage?](#)
 - [Probabilistic vs Deterministic linkage](#)
 - ☐
 - The Fellegi-Sunter Model [The Fellegi-Sunter Model](#)
 - Table of contents
 - [Parameters of the Fellegi-Sunter model](#)
 - [λ probability](#)
 - [m probability](#)
 - [u probability](#)
 - [Interpreting m and u](#)
 - [m probability](#)
 - [u probability](#)
 - [Match Weights](#)
 - [Deriving Match Weights from m and u](#)
 - [Interpreting Match Weights](#)
 - [Match Probability](#)
 - [Deriving Match Probability from Match Weight](#)
 - [Deriving Match Probability from m and u](#)
 - [Further Reading](#)
 - [Linked Data as Graphs](#)
 - ☐
 - Linkage Models in Splink
 - Linkage Models in Splink
 - [Defining Splink models](#)
 - [Retrieving and querying Splink results](#)
 - [Link type - linking vs deduping](#)
 - ☐
 - Splink's SQL backends - Spark, DuckDB etc
 - Splink's SQL backends - Spark, DuckDB etc
 - [Backends overview](#)
 - [PostgreSQL](#)
 - ☐
 - Data Preparation
 - Data Preparation
 - [Feature Engineering](#)
 - ☐

Blocking

Blocking

- [What are Blocking Rules?](#)
- [Computational Performance](#)
- [Model Training Blocking Rules](#)

◦ ☐

Comparing Records

Comparing Records

- [Comparisons and comparison levels](#)
- [Defining and customising comparisons](#)
- [Out-of-the-box comparisons](#)
- [Term frequency adjustments](#)
- ☐

Comparing strings

Comparing strings

- [String comparators](#)
- [Choosing string comparators](#)
- [Phonetic algorithms](#)
- [Regular expressions](#)

◦ ☐

Training

Training

- [Training rationale](#)

◦ ☐

Evaluation

Evaluation

- [Overview](#)
- [Model](#)
- ☐
- Edges (Links)
- Edges (Links)
 - [Overview](#)
 - [Edge Metrics](#)
 - [Clerical Labelling](#)
- ☐

Clusters

Clusters

- [Overview](#)
- [Graph metrics](#)
- [How to compute graph metrics](#)

◦ ☐

Performance

Performance

- [Run times, performance and linking large data](#)
- [Performance of comparison functions](#)
- ☐
- Spark Performance
- Spark Performance
 - [Optimising Spark performance](#)
 - [Salting blocking rules](#)
- ☐

DuckDB Performance

DuckDB Performance

- [Optimising DuckDB performance](#)

◦ ☐

Charts Gallery

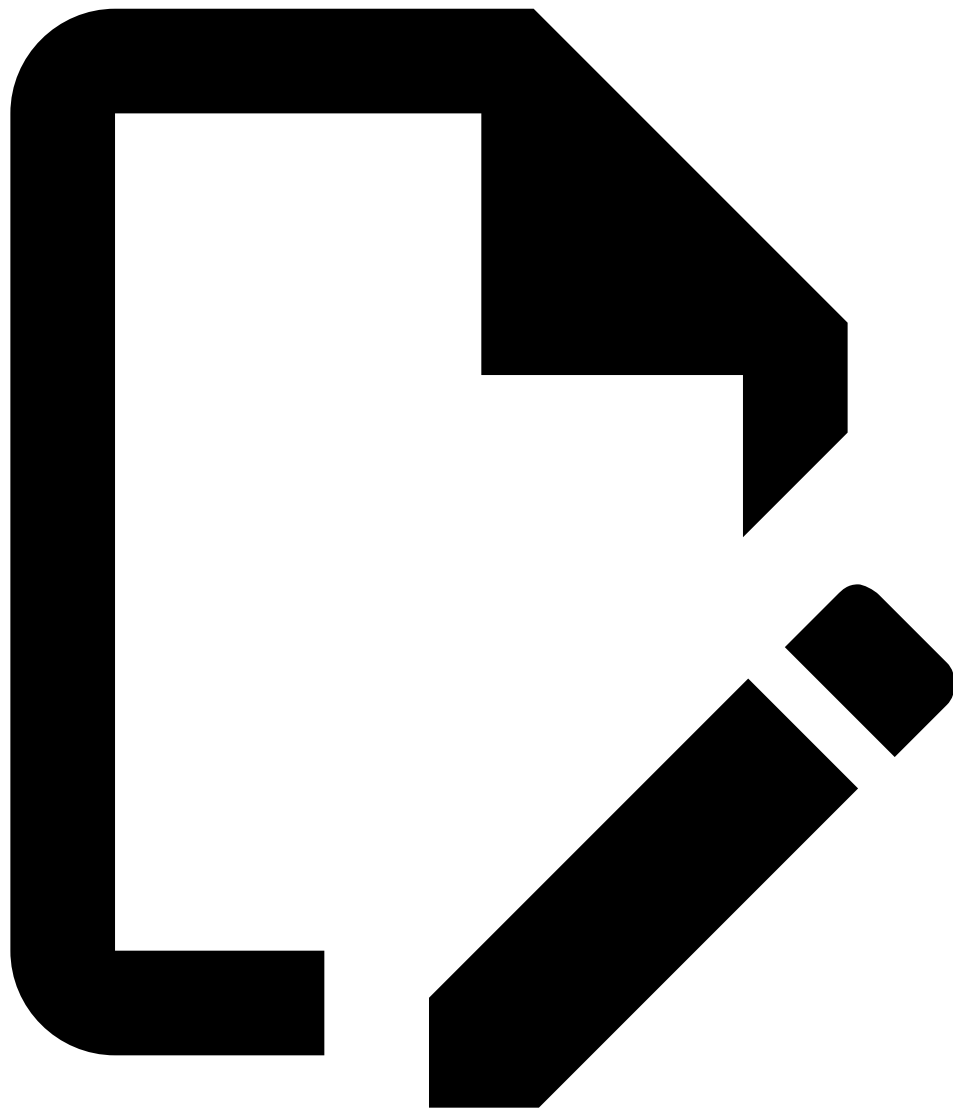
Charts Gallery

- ☐
 - Exploratory Analysis
 - Exploratory Analysis
 - [completeness chart](#)
 - [profile columns](#)
 - ☐
 - Blocking
 - Blocking
 - [cumulative num comparisons from blocking rules chart](#)
 - ☐
 - Similarity analysis
 - Similarity analysis
 - [Comparator score chart](#)
 - [Comparator score threshold chart](#)
 - [Phonetic match chart](#)
 - ☐
 - Model Training
 - Model Training
 - [comparison viewer dashboard](#)
 - [match weights chart](#)
 - [m u parameters chart](#)
 - [parameter estimate comparisons chart](#)
 - [tf adjustment chart](#)
 - [unlinkables chart](#)
 - [waterfall chart](#)
 - ☐
 - Clustering
 - Clustering
 - [cluster studio dashboard](#)
 - ☐
 - Model Evaluation
 - Model Evaluation
 - [accuracy chart from labels table](#)
 - [threshold selection tool](#)
 - [LLM prompts](#)
- ☐
 - [Contributing](#)
 - Contributing
 - ☐
 - Contributing to Splink
 - Contributing to Splink
 - [Contributor Guide](#)
 - [Development Quickstart](#)
 - [Linting and Formatting](#)
 - [Testing](#)
 - [Contributing to Documentation](#)
 - [Managing Environment and Dependencies](#)
 - [Releasing a Package Version](#)
 - [Contributing to the Splink Blog](#)
 - ☐
 - How Splink works
 - How Splink works
 - [Understanding and debugging Splink](#)

- [Transpilation using sqlglot](#)
 - ☐
 - Performance and caching
 - Performance and caching
 - [Caching and pipelining](#)
 - [Spark caching](#)
 - ☐
 - Charts
 - Charts
 - [Understanding and editing charts](#)
 - [Building new charts](#)
 - [User-Defined Functions](#)
 - [Dependency Compatibility Policy](#)
- ☐
 - [Blog](#)
 - Blog
 - ☐
 - Categories
 - Categories
 - [Bias](#)
 - [Ethics](#)
 - [Feature Updates](#)
 - [Production Splink pipelines](#)

Table of contents

- [Parameters of the Fellegi-Sunter model](#)
 - [λ probability](#)
 - [m probability](#)
 - [u probability](#)
- [Interpreting m and u](#)
 - [m probability](#)
 - [u probability](#)
- [Match Weights](#)
 - [Deriving Match Weights from m and u](#)
 - [Interpreting Match Weights](#)
- [Match Probability](#)
 - [Deriving Match Probability from Match Weight](#)
 - [Deriving Match Probability from m and u](#)
- [Further Reading](#)



The Fellegi-Sunter model

This topic guide gives a high-level introduction to the Fellegi Sunter model, the statistical model that underlies Splink's methodology.

For a more detailed interactive guide that aligns to Splink's methodology see Robin Linacre's [interactive introduction to probabilistic linkage](#).

Parameters of the Fellegi-Sunter model

The Fellegi-Sunter model has three main parameters that need to be considered to generate a match probability between two records:

- λ - probability that any two records match
- m - probability of a given observation *given* the records are a match

- u - probability of a given observation *given* the records are **not** a match
-

λ probability

The λ parameter is the prior probability that any two records match. I.e. assuming no other knowledge of the data, how likely is a match? Or, as a formula:

$$\lambda = \Pr(\text{Records match})$$

This is the same for all records comparisons, but is highly dependent on:

- The total number of records
 - The number of duplicate records (more duplicates increases λ)
 - The overlap between datasets
 - Two datasets covering the same cohort (high overlap, high λ)
 - Two entirely independent datasets (low overlap, low λ)
-

m probability

The m probability is the probability of a given observation *given the records are a match*. Or, as a formula:

$$m = \Pr(\text{Observation} \mid \text{Records match})$$

For example, consider the m probability of a match on Date of Birth (DOB). For two records that are a match, what is the probability that:

- **DOB is the same:**
 - Almost 100%, say 98% ($m \approx 0.98$)
- **DOB is different:**
 - Maybe a 2% chance of a data error? ($m \approx 0.02$)

The m probability is largely a measure of data quality - if DOB is poorly collected, it may only match exactly for 50% of true matches.

u probability

The u probability is the probability of a given observation *given the records are not a match*. Or, as a formula:

$$u = \Pr(\text{Observation} \mid \text{Records do not match})$$

For example, consider the u probability of a match on Surname. For two records that are not a match, what is the probability that:

- **Surname is the same:**
 - Depending on the surname, $<1\%$? ($u \approx 0.005$)
- **Surname is different:**
 - Almost 100% ($u \approx 0.995$)

The u probability is a measure of coincidence. As there are so many possible surnames, the chance of sharing the same surname with a randomly-selected person is small.

Interpreting m and u

In the case of a perfect unique identifier:

- A person is only assigned one such value - $(m = 1)$ (match) or $(m = 0)$ (non-match)
- A value is only ever assigned to one person - $(u = 0)$ (match) or $(u = 1)$ (non-match)

Where (m) and (u) deviate from these ideals can usually be intuitively explained:

m probability

A measure of **data quality/reliability**.

How often might a person's information change legitimately or through data error?

- **Names:** typos, aliases, nicknames, middle names, married names etc.
- **DOB:** typos, estimates (e.g. 1st Jan YYYY where date not known)
- **Address:** formatting issues, moving house, multiple addresses, temporary addresses

u probability

A measure of **coincidence/cardinality**.

How many different people might share a given identifier?

- **DOB** (high cardinality) – for a flat age distribution spanning ~30 years, there are ~10,000 DOBs (0.01% chance of a match)
- **Sex** (low cardinality) – only 2 potential values (~50% chance of a match)

Match Weights

One of the key measures of evidence of a match between records is the match weight.

Deriving Match Weights from m and u

The match weight is a measure of the relative size of (m) and (u) :

$$M = \log_2 \left(\frac{\lambda}{1-\lambda} \right) + \log_2 \frac{K}{m} \\ K = \log_2 \left(\frac{\lambda}{1-\lambda} \right) + \log_2 u$$

where (λ) is the probability that two random records match and $(K=m/u)$ is the Bayes factor.

A key assumption of the Fellegi Sunter model is that observations from different column/ comparisons are independent of one another. This means that the Bayes factor for two records is the products of the Bayes factor for each column/comparison:

$$K_{\text{features}} = K_{\text{forename}} \cdot K_{\text{surname}} \cdot K_{\text{dob}} \\ \cdot K_{\text{city}} \cdot K_{\text{email}}$$

This, in turn, means that match weights are additive:

$$M_{\text{obs}} = M_{\text{prior}} + M_{\text{features}}$$

where $(M_{\text{prior}} = \log_2\left(\frac{\lambda}{1-\lambda}\right))$ and $(M_{\text{features}} = M_{\text{forename}} + M_{\text{surname}} + M_{\text{dob}} + M_{\text{city}} + M_{\text{email}})$.

So, considering these properties, the total *match weight* for two observed records can be rewritten as:

$$\begin{aligned} M_{\text{obs}} &= \log_2\left(\frac{\lambda}{1-\lambda}\right) + \sum_i \log_2\left(\frac{m_i}{u_i}\right) \\ &= \log_2\left(\frac{\lambda}{1-\lambda}\right) + \log_2\left(\prod_i \log_2\left(\frac{m_i}{u_i}\right)\right) \end{aligned}$$

Interpreting Match Weights

The *match weight* is the central metric showing the amount of evidence of a match is provided by each of the features in a model. The is most easily shown through Splink's Waterfall Chart:

- 1 are the two records being compared
- 2 is the *match weight* of the **prior**, $(M_{\text{prior}} = \log_2\left(\frac{\lambda}{1-\lambda}\right))$. This is the *match weight* if no additional knowledge of features is taken into account, and can be thought of as similar to the y-intercept in a simple regression.
- 3 are the *match weights* of **each feature**, (M_{forename}) , (M_{surname}) , (M_{dob}) , (M_{city}) and (M_{email}) respectively.
- 4 is the **total match weight** for two observed records, combining 2 and 3:

$$\begin{aligned} M_{\text{obs}} &= M_{\text{prior}} + M_{\text{forename}} + M_{\text{surname}} + M_{\text{dob}} + M_{\text{city}} + M_{\text{email}} \\ &= -6.67 + 4.74 + 6.49 - 1.97 - 1.12 + 8.00 \\ &= 9.48 \end{aligned}$$

- 5 is an axis representing the $(\text{match weight} = \log_2(\text{Bayes factor}))$
- 6 is an axis representing the equivalent *match probability* (noting the non-linear scale). For more on the relationship between *match weight* and *probability*, see the [sections below](#)

Match Probability

Match probability is a more intuitive measure of similarity than *match weight*, and is, generally, used when choosing a similarity threshold for record matching.

Deriving Match Probability from Match Weight

Probability of two records being a match can be derived from the total *match weight*:

$$\Pr(\text{Match} \mid \text{Observation}) = \frac{2^{M_{\text{obs}}}}{1 + 2^{M_{\text{obs}}}}$$

Example

Consider the example in the [Interpreting Match Weights](#) section. The total *match weight*, $(M_{\text{obs}} = 9.48)$. Therefore,

$$\Pr(\text{Match} \mid \text{Observation}) = \frac{2^{9.48}}{1 + 2^{9.48}} \approx 0.999$$

Understanding the relationship between Match Probability and Match Weight

It can be helpful to build up some intuition for how *match weight* translates into *match probability*.

Plotting *match probability* versus *match weight* gives the following chart:

Some observations from this chart:

- $(\text{Match weight} = 0 \rightarrow \text{Match probability} = 0.5)$
- $(\text{Match weight} = 2 \rightarrow \text{Match probability} = 0.8)$
- $(\text{Match weight} = 3 \rightarrow \text{Match probability} = 0.9)$
- $(\text{Match weight} = 4 \rightarrow \text{Match probability} = 0.95)$
- $(\text{Match weight} = 7 \rightarrow \text{Match probability} = 0.99)$

So, the impact of any additional *match weight* on *match probability* gets smaller as the total *match weight* increases. This makes intuitive sense as, when comparing two records, after you already have a lot of evidence/features indicating a match, adding more evidence/features will not have much of an impact on the probability of a match.

Similarly, if you already have a lot of negative evidence/features indicating a match, adding more evidence/features will not have much of an impact on the probability of a match.

Deriving Match Probability from m and u

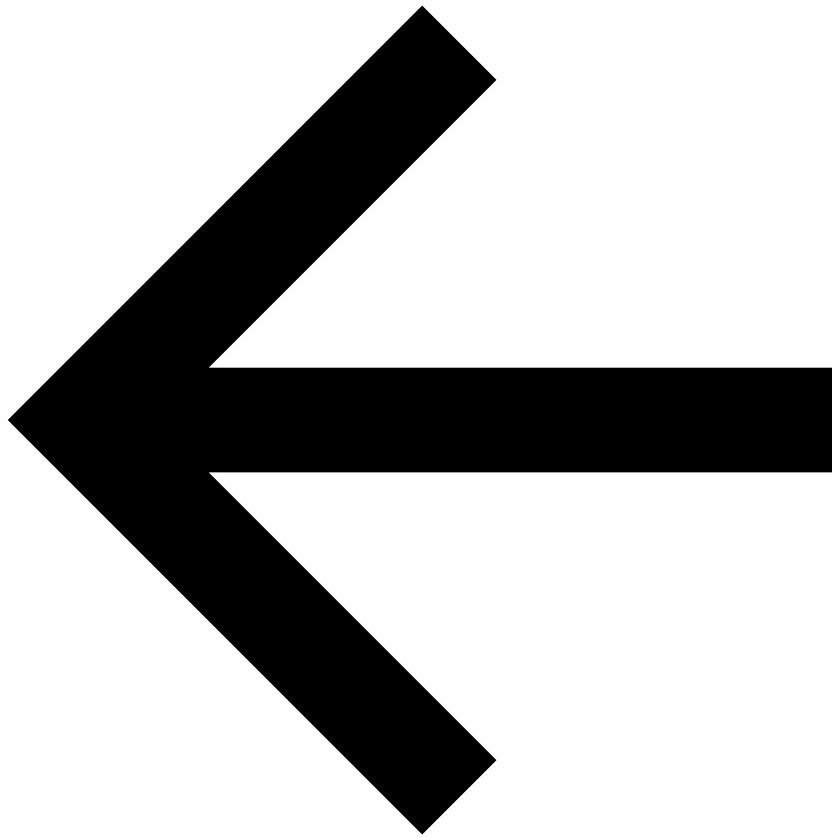
Given the definitions for *match probability* and *match weight* above, we can rewrite the probability in terms of (m) and (u) .

$$\begin{aligned} \Pr(\text{Match} \mid \text{Observation}) &= \frac{2^{\log_2\left(\frac{\lambda}{1-\lambda}\right) + \log_2\left(\prod_i^{\text{features}} \frac{m_i}{u_i}\right)}}{1 + 2^{\log_2\left(\frac{\lambda}{1-\lambda}\right) + \log_2\left(\prod_i^{\text{features}} \frac{m_i}{u_i}\right)}} \\ &= \frac{\left(\frac{\lambda}{1-\lambda}\right) \prod_i^{\text{features}} \frac{m_i}{u_i}}{1 + \left(\frac{\lambda}{1-\lambda}\right) \prod_i^{\text{features}} \frac{m_i}{u_i}} \\ &= 1 - \frac{1}{1 + \left(\frac{\lambda}{1-\lambda}\right) \prod_i^{\text{features}} \frac{m_i}{u_i}} \end{aligned}$$

Further Reading

[This academic paper](#) provides a detailed mathematical description of the model used by R [fastLink package](#). The mathematics used by Splink is very similar.

1. Cardinality is the the number of items in a set. In record linkage, cardinality refers to the number of possible values a feature could have. This is important in record linkage, as the number of possible options for e.g. date of birth has a significant impact on the amount of evidence that a match on date of birth provides for two records being a match. ↩



[Previous](#)
[Probabilistic vs Deterministic linkage](#)
[Next](#)
[Linked Data as Graphs](#)

