Enlighten Data Quality

Presented by: Sean Purcell, Hector Cordova, Michael Ott

INNOVATIVE SYSTEMS

It's a data driven world. Let us be your guide.

Pittsburgh | London | Dubai | Frankfurt | Mexico City | Sãu Paulo | Singapore

50+
years' experience

1000+
customers

60+
countries

100 Bn+
records per year

Selected Customers

Banking/FinServices

- ally
- Bank of Ireland
- charles SCHWAB
- BMO Bank of Montreal
- Huntington
- BCA
- AL RAYAN BANK

Insurance

- Manulife
- Prudential
- GUGGENHEIM LIFE AND ANNUITY COMPANY
- WPA
- KEMPER
- Bupa بوبا

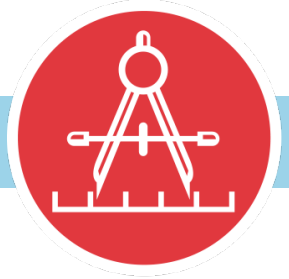
Manufacturing/Retail

- VOLVO Financial Services
- CVS pharmacy
- Hallmark
- United States Steel
- L.L.Bean
- FOKKER TECHNOLOGIES
- verizon

NGOs

- American Red Cross
- World Learning
- CAFOD
- American Diabetes Association
- HelpAge International age helps

Value Proposition



ACCURACY

- Massive AI-based crowdsourced knowledgebases/ grammar files
- Matching algorithm deeply rooted in 50+ years of data quality expertise
- Highest level of data quality – 99.5%+

PERFORMANCE

- Scalability: Processes any volume of records, from thousands to hundreds of millions
- Speed: Meets even the shortest processing windows

IMPLEMENTATION

- ANY deployment options – On-premise, SaaS/Cloud, or Hybrid
- Global data centers processing 100 billion+ records per year
- 50 years of successful integrations & migrations

CUSTOMER SUPPORT

- Customer satisfaction recognized by leading analyst firms
- Timely access to staff with deep technical expertise
- 24/7 follow the sun support

The Data Quality Process (Customer banking example)

Mr. Robert & Mrs. Linda Taylor
(T/A Taylor's)
790 Holiday Drive, Pittsburgh

Bob Taylor ***owner***
790 Holiday Dr, 15220

MR AND MS ROBERT TAYLOR
---- DO NOT MAIL-----
790 HOLIDAY, GREENTREE PA

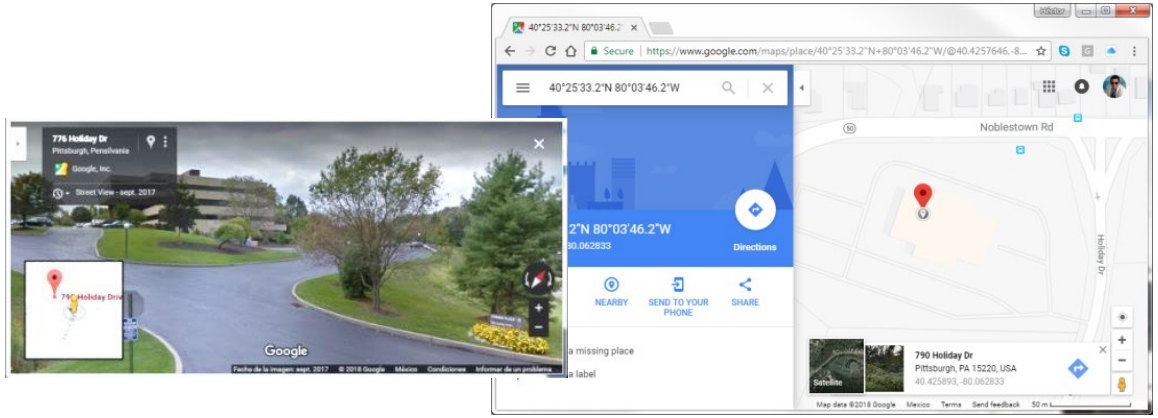
DATA QUALITY

Mr Robert Taylor ***(Male, Individual)***
Checking, Loan, Mortgage

Mrs Linda Taylor ***(Female, Individual)***
Checking, Mortgage

Taylor's ***(Organization)***
Checking

790 Holiday Drive
Pittsburgh
Pennsylvania 15220 – **2750**
Lat: 40.425886°
Long: -80.062835°



Crowdsourced AI Powers Our Technology

Crowdsourcing to build knowledgebases

- Billions of records, thousands of clients
- 4-Eyes quality control by specialists
- Not Personally Identifiable Information (PII)
- Millions of correct and incorrect words, phrases and patterns defined
- Knowledgebases continue to be grown through continuous processing

In-house architecture for AI (knowledgebased) system

- Models human behavior
- Utilizes the crowdsourced knowledgebases as the brains of the system
- Much more accurate, automated and faster than other approaches

Unique scenario-based, match string approach delivers the highest levels of matching accuracy, automation & explainability

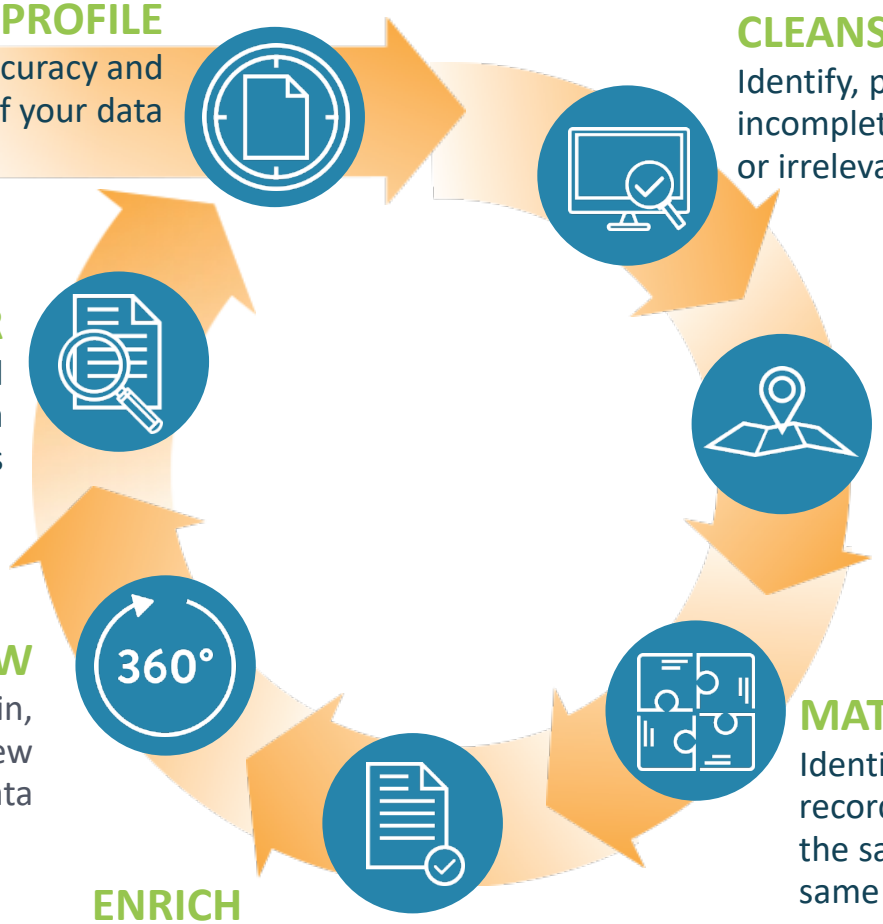
- Applies one specific match string for each scenario (use case)
- Maintains field-level precision
- Much more precise than probabilistic/weighted field or deterministic matching

Enlighten Product Suite – Methodology

Enlighten

The latest generation of our enterprise data quality suite
(formerly i/Lytics)

Enlighten delivers unrivaled accuracy and speed to improved self-service capabilities to provide our clients even more control and flexibility.



PROFILE
Assess the accuracy and completeness of your data

CLEANSE
Identify, parse, and standardize incomplete, inaccurate, incorrect, or irrelevant data fields

ADDRESS VALIDATION & GEOCODING
Verify and correct addresses and add geocoding

MATCH & LINK
Identify, match, and merge records that correspond to the same entities from the same or disparate databases

ENRICH
Augment your existing data with geographic, demographic, and consumer insights

360° VIEW
Integrate, store, maintain, and leverage a single view of your data

MONITOR
Have round-the-clock control over the quality of your data with alerts and audit trails

Demo

Demo – Data Profiler Statistics

Enlighten Data Profiler

Enlighten Data Profiler

Profiling

Relationship Analysis

Export Settings

Exit Project

Test Project A ▸ Customer V1

Run Date	Columns	File Size	Record Count	Read Time
12/01/2021 10:20:14 AM	13	5.72 MB	33456	02h 20m 49s

Profiling Settings

Metric Grid | Chart View | Data Validation | Drilldown

Select and Copy All | Export Grid

No	Name	Uniqueness	Unique Count	Completeness	Row Count	Null Count	Blank Count	Pattern Count	Overall Data Type	Most Common Data Type	Minimum String	M
1	Customer Id	99.71%	33,360	100%	33,456	0	0	2	Alphanumeric	Alphanumeric	A10-E100750	D99-E1
2	Discount Code	0.02%	8	100%	33,456	0	0	1	Alphanumeric	Alphanumeric	A	Z
3	Forename	10.12%	3,386	100%	33,456	0	0	82	Alphanumeric	Alphanumeric	.	vijay
4	Surname	29.87%	9,992	97%	32,452	1,004	0	139	Alphanumeric	Alphanumeric	0'Connor	x
5	Email	99.71%	33,360	100%	33,456	0	0	19,218	Alphanumeric	Alphanumeric	-Fayek@Albany.medical.center.com	vijay.Ca
6	Telephone	99.2%	33,190	100%	33,456	0	0	83	Alphanumeric	Phone	() 312-441-9080	NOT PR
7	Company Name	82.62%	27,640	99.99%	33,451	5	0	11,414	Alphanumeric	Alphanumeric	1-800-Flowers.com, Inc.	yugas
8	First Order Date	40.35%	13,501	99.71%	33,360	96	0	2	Alphanumeric	Alphanumeric	01/01/1969	l6/01/1
9	Sales Last Year	0.97%	324	100%	33,456	0	0	9	Decimal	Integer	-6000	990000
10	Addressline	77.18%	25,823	100%	33,455	1	0	4,942	Alphanumeric	Alphanumeric	# 256, ShenHe Qu QingNian Street	xyz

Showing 10 | 1 of 2

Demo – Data Profiler Drilldown

Enlighten Data Profiler

Enlighten Data Profiler

Profiling

Relationship Analysis

Export Settings

Exit Project

Test Project C ▸ Customer V1

Run Date	Columns	File Size	Record Count	Read Time
12/02/2021 1:02:15 PM	13	5.72 MB	33456	00h 00m 13s

Profiling Settings

Metric Grid | Chart View | Data Validation | Drilldown

Select and Copy All | Export Grid

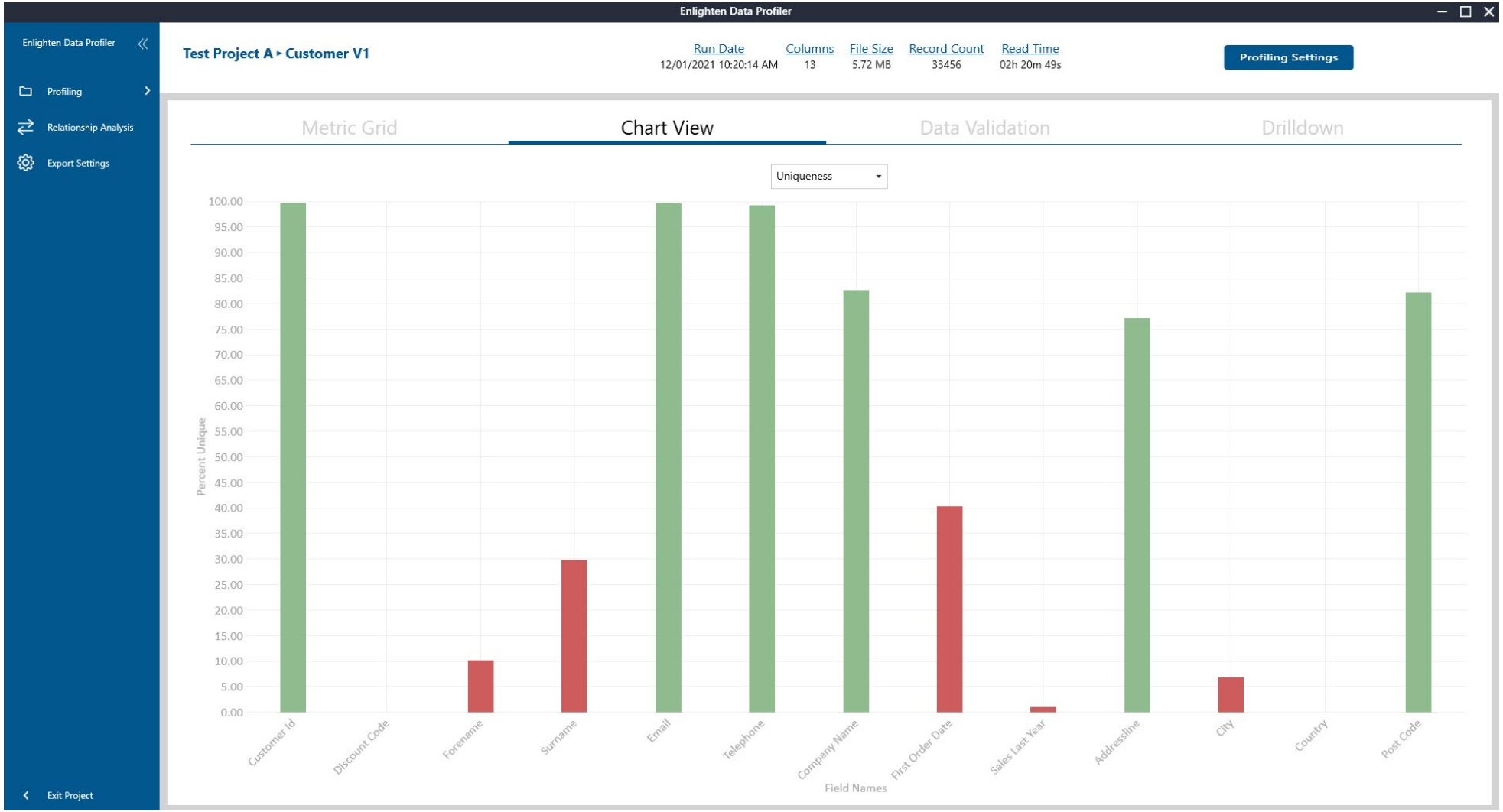
Unique Value Drill Down Summary for Addressline

Addressline	Count
NOT PROVIDED	51
None	35
c/o Logan Britton	15
Postbus 1	9
Postbus 20	8
Postbus 1800	8
148 Grenoble Road	6
135 Bishopthorpe Road	6
145 Annfield Rd	6
184 Well Lane	6

Showing 10

1 of 2583

Demo – Data Profiler Charts



Demo – Cleanse Grammar

Test Parser

Parse | Reset

Grammar: US_Enlighten_grammar_2.0.c

File Type: 825 | Tab Delimited

File To Load

Test Parser Input

```
JOHN SMITH
123 MAIN ST
PITTSBURGH PA 15220
USA
```

Record of -

Token Identification

Token	Dictionaries	Standardizations
JOHN	Alpha2 AlphaWithVowel Name_APPENDAGE_SOFT4_PATTERN Dict_MaleFirstName	
SMITH	Alpha2 AlphaWithVowel Dict_KnownCityNameUS Dict_UnknownGenderLastName	
123	CityUSA_ZIPCODE_INVALID_RULE CityUSA_ZIPCODE_INVALID5 CityUSA_ZIPCODE_INVALID4 StreetUSA_NUMERICORDINAL_RD_NUM DescriptorId_SSN_3DIGIT StreetUSA_APARTMENTNUMBER_LINEBYSELF StreetUSA_APARTMENTNUMBER_LINEBYSELF_NUM13 StreetUSA_FRACTION_3DIGIT Numeric	
MAIN	Alpha2 AlphaWithVowel Name_APPENDAGE_SOFT4_PATTERN Dict_CommonStreetName Dict_KnownCityNameUS Dict_WeakOrgWord	

Collapse All Nodes

- JOHN SMITH 123 MAIN ST PITTSBURGH PA 15220 USA (Root)
  - JOHN SMITH 123 MAIN ST PITTSBURGH PA 15220 USA (GrammarEnl)
    - JOHN SMITH 123 MAIN ST PITTSBURGH PA 15220 USA (GrammarEnl_R_IDV_BLOCK)
      - JOHN SMITH (GrammarEnl_NAME_IDV_BLOCK_VERYSTRONG)
        - JOHN SMITH (GrammarEnl_LINE_1_IDV)
          - JOHN SMITH (IndividualLines)
            - JOHN SMITH (Name_STRONG)
              - JOHN SMITH (NameWestern_STRONG)
                - JOHN SMITH (NameWestern_STRONG_WRAPPER_SINGLE)
                - JOHN SMITH (NameWestern_STRONG_NAME)
                - JOHN SMITH (NameWestern_STRONG_FIRSTLAST)
                - JOHN SMITH (NameWestern_STRONG_FIRSTLAST_M)
                - JOHN SMITH (NameWestern_STRONG_FIRSTLAST_M_NAME)
                - JOHN (NAME_Full)
                - JOHN (FirstName)
                - JOHN (NameWestern_GIVENNAME_M_RULE)
                - JOHN (DICT_NAME_WESTERN_GIVENNAME_MALE)
                - JOHN (Dict_MaleFirstName)
                - SMITH (NAME_Full)
                - SMITH (LastName)
                - SMITH (NameWestern_SURNAME_HARD_RULE)
                - SMITH (DICT_NAME_WESTERN_SURNAME)
                - SMITH (Dict_UnknownGenderLastName)
        - (EOL)
      - 123 MAIN ST (GrammarEnl_STREET_IDV_BLOCK_VERYSTRONG)
        - 123 MAIN ST (GrammarEnl_STREET_IDV_LINE_VERYSTRONG)
        - (EOL)
      - PITTSBURGH PA 15220 USA (GrammarEnl_CITY_IDV_BLOCK_STRONG)
        - PITTSBURGH PA 15220 (GrammarEnl_CITY_IDV_LINE_STRONG)
        - (EOL)
        - USA (GrammarEnl_CITY_IDV_LINE_STRONG)
        - (EOL)

Demo – Cleanse Project with Foreign Data

Test Cleanse Project

Test | Reset

Cleanse Project

US_RET_CL Working

Manual Entry | Upload File

Test Field Inputs

Field	Occurs	Value
Name Address	1	HÉCTOR (****.,&*^@) MAGA#A-Pérez
Name Address	2	C/O Junior Conceição
Name Address	3	#7 123, MAIN STREET.
Name Address	4	pittsburgh=/((/%% pennsylvania 15220
Name Address	5	USA
Name Address	6	
Name Address	7	
Name Address	8	

In X Ref Number(4) '04'
In X Ref Number(5) '05'
Name Information
Name Line(1) 'HÉCTOR MAGAÑA-PÉREZ'
Name Parsed Elements
First Name(1) 'HÉCTOR'
First Name(2) 'JUNIOR'
Last Name(1) 'MAGAÑA-PÉREZ'
Last Name(2) 'CONCEIÇÃO'
Output Gender(1) 'M'
Output Gender(2) 'M'
Organization Information
Organization Parsed Elements
Street Information
Street Line(3) '# 7 123 MAIN ST'
Street Parsed Elements
House Number(3) '123'
Street Name(3) 'MAIN'
Street Identifier(3) 'ST'
Subaddress Type(3) '#'
Subaddress Number(3) '7'
Urbanization Information
Urbanization Parsed Elements
Municipality Information
Municipality Parsed Elements
City Information
City Line(4) 'PITTSBURGH PA 15220'
City Parsed Elements
City Name(4) 'PITTSBURGH'
State Code(4) 'PA'
Postal Code(4) '15220'
Country(5) 'USA'
Other Information
Other Line(2) 'C/O JUNIOR CONCEIÇÃO'
Other Line(5) 'USA'
User_Group
User Link A Occurs
User Link B Occurs

Demo – Cleanse Project

Test Cleanse Project

Test | Reset

Cleanse Project

US_RET_CL Working

Manual Entry | Upload File

Test Field Inputs

Field	Occurs	Value
Name Address	1	Mr. John And Mrs. Mary A. Smith SAVINGS ACCOUNT 891729381293
Name Address	2	DBA SMITH ASSOCIATES
Name Address	3	CARE OF TOM JACKSON ACCOUNTANT
Name Address	4	BUILDING 18
Name Address	5	2890 north 5th street
Name Address	6	los angeles ca 90134
Name Address	7	
Name Address	8	

Name Line(1) 'JOHN SMITH'
Name Line(2) 'MARY A SMITH'
Name Parsed Elements
Title(1) 'MR'
Title(2) 'MRS'
First Name(1) 'JOHN'
First Name(2) 'MARY A'
First Name(5) 'TOM'
Last Name(1) 'SMITH'
Last Name(2) 'SMITH'
Last Name(5) 'JACKSON'
Name Appendage(5) 'ACCOUNTANT'
Conjunction(1) 'AND'
Output Gender(1) 'M'
Output Gender(2) 'F'
Output Gender(5) 'M'
Organization Information
Organization Line(4) 'SMITH ASSOCIATES'
Organization Parsed Elements
Street Information
Street Line(6) 'BLDG 18'
Street Line(7) '2890 N 5TH ST'
Street Parsed Elements
PreDirectional(7) 'N'
House Number(7) '2890'
Street Name(7) '5TH'
Street Identifier(7) 'ST'
Building Name(6) 'BLDG'
Building Number(6) '18'
Urbanization Information
Urbanization Parsed Elements
Municipality Information
Municipality Parsed Elements
City Information
City Line(8) 'LOS ANGELES CA 90134'
City Parsed Elements
City Name(8) 'LOS ANGELES'
State Code(8) 'CA'
Postal Code(8) '90134'

Demo – Address Validation

PostLocate

Home About Contact

< Back

Input Data

Innovative SysteM

Name of a household, company, location, etc.

790 holLiday

Address line 1

Address line 2

Address line 3

Wabash, PA

City, State and/or ZIP code

Response Data

INNOVATIVE SYSTEMS INC

Output Line 1

790 HOLIDAY DR

Output Line 2

PITTSBURGH PA 15220-8127

Output Line 3

Full Search Info

Hit Details

Return Code: NormalHit
Country Used: NormalHit
PostalCode: Missing
City: Equal
State: Equal
PreDirection: Equal
Street: Check
StreetType: Missing
PostDirection: Equal
HouseNumber: Exact
Subaddress: Missing
SubaddressType: Missing
Firm: LooseMatch
Locality: Equal

Output Details

Line 0: INNOVATIVE SYSTEMS INC
Line 1: 790 HOLIDAY DR
Line 2: PITTSBURGH PA 15220-8127
Latitude: 40.425852N
Longitude: 80.061619W
GeocodeType: streetlevel

CensusTract: 469000
CensusBlock: 3016
MSA: 6280
CBSA: 38300
MCD: 31256
PlaceCode: 31256
CountyFIPSCode: 003
StateFIPSCode: 42
CountyName: ALLEGHENY

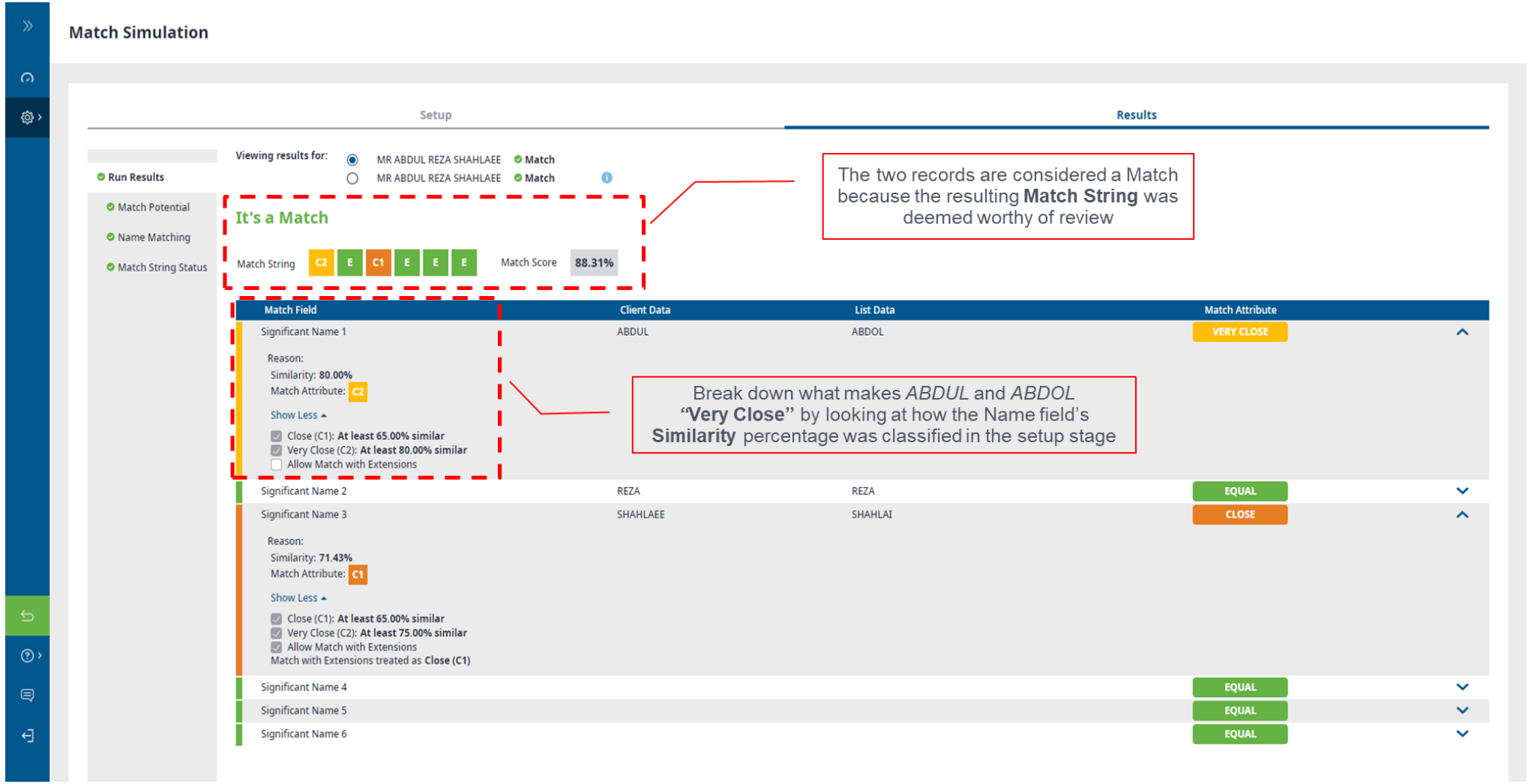
Further Info

Firm: INNOVATIVE SYSTEMS INC
HouseNumber: 790
City: PITTSBURGH
VanityCity: WABASH
State: PA
PostalCode: 152208127
AddressLine1: 50
AddressLine2: 48
CityLine: 51
StreetName: HOLIDAY
StreetType: DR
RecordType: Firm

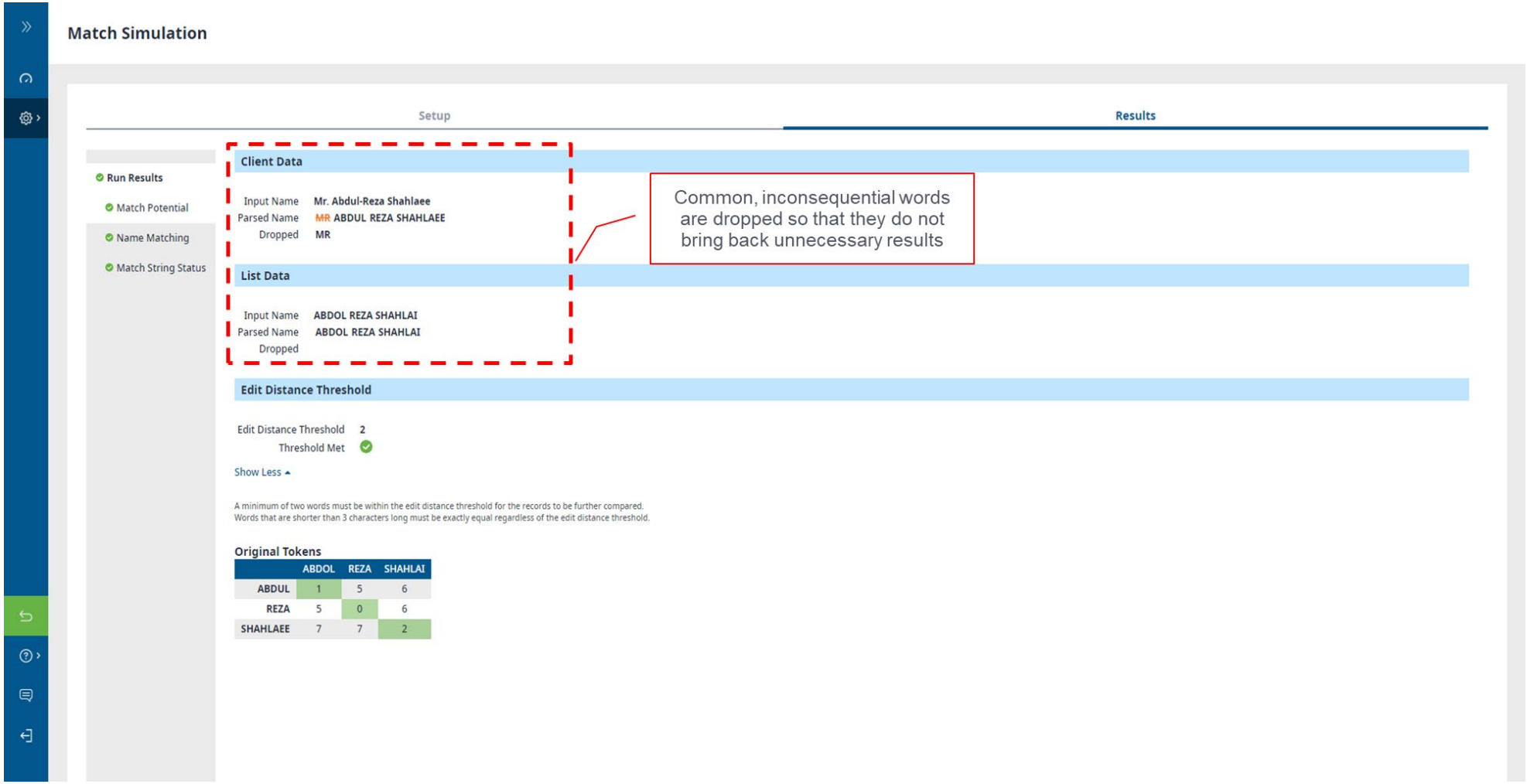
AliasFlag: NotAnAlias
BaseAlternateFlag: BaseMatch
AbbrevPrefFlag: Not-Applicable
UniqueZipFlag: 48
DeliveryPoint: 90
DeliveryPointCheckDigit: 51
CarrierRoute: C056
LacsFlag: 32
LacsLinkCode: Blank
SuiteLinkCode: NotChecked
ELOTAscDesc: Neither
StateFIPSCode: 42

CountyFIPSCode: 003
CountyName: ALLEGHENY
LowStreetNumber: 790
HighStreetNumber: 790
OddEvenFlag: even
LowSubaddressNumber: 11
HighSubaddressNumber: 11
SubaddressFromHouse: false

Demo – Match Simulator



Demo – Match Simulator Details



Demo – Search Example

Live Demo

Match Simulation

Setup | Results

Screening Configuration: DowJonesTest_Final_Continued | v1

Screening Type: Individual / Organization

Simulation Type: Search / Manual

Client Data

Name: Mr. Abdul-Reza Shahlaee

Search Parameters

Lists

- [ ] Select All
- [ ] OFAC Specially Designated Nationals
- [ ] World-Check
- [ ] Dow Jones Watchlist
- [x] Dow Jones - Test File 2020

Search

15 results were brought back using this Screening Configuration. Each result has a corresponding **Match String/Match Score** that show the quality of the Match

Showing 15 results

C2 E C1 E E E	88.31%	ABDOL REZA SHAHLAI
C2 E C1 E E E	88.31%	ABDOL-REZA SHAHLAI
C1 E B E E E	86.41%	REZA SHAHLAI
C1 E B E E E	86.41%	REZA SHAHLAI
C1 E B E E E	86.41%	REZA SHAHLAI

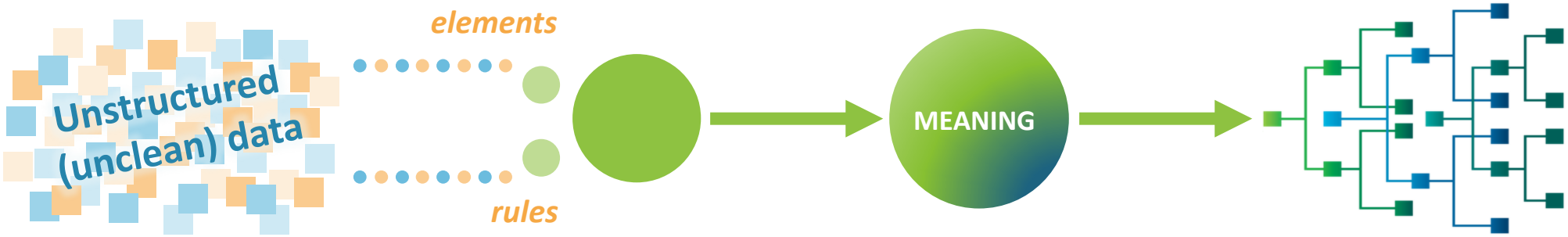
Simulate

Simulate, step-by-step, how the final Match/No-Match determination was made

Appendix

Crowdsourced Knowledgebases (a.k.a. Grammars)

- **Grammars contain the rules that specify how to identify, cleanse and standardize your data**
  1. They control how each of the words, numbers, and data **elements** are classified and choose whether any standardizations are applied
  2. These classifications, combined with **rules** are used to identify the **meaning** and correct order of the data



Lorem ipsum dolor sit amet, consectetuer adipiscing elit. **John Smith** Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus **123 Main St** et magnis dis parturient montes, nascetur ridiculus mus. **15220** Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, **4129379300** ut, imperdiet a, venenatis vitae, justo. Nullam.

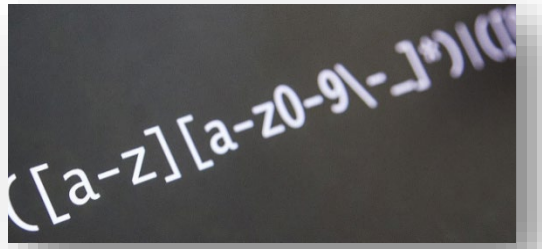
John Smith *(Male, Individual)*
123 Main Street
Pittsburgh Pennsylvania 15220
(412) 937-9300

What Type of Data Can We Use in Grammars?

- **Dictionaries**: first names, last names, common organizational and address terms. (any collection of words in any language)

- Country-specific or regional **postal standards** from any available database (public, license or self-owned)

- **Regular expressions** to understand and identify any data. E.g. e-mail (RFC 5322 standard), Telephone, National Identification numbers, part numbers (any data with a pre-defined structure).



- **Common domains**
  Public (e-mail, website, countries, area codes)
  Organization-specific (invalid or dummy data – XXXX, 000, 111, etc.)

ISO

Federal Communications Commission

…etcetera…

*Anything that can be represented either as a **collection of elements** or as a **pattern of text characters** in any language*

What Does It Deliver?

- Cleansing and Matching in **any language**
  - Support for special characters, accent marks, etc. (Unicode-ready)
  - The ability to use any field and any number of fields in matching
- **Custom parsing** to meet unique data needs
  - Specific requirements by country / by industry / by client / by type of data
- Ability to implement cleansing processes for **data beyond name and address**
  - Telephone, e-mail, date of birth, tax number, national identification number, etc.
- … and **beyond the Customer/Party domain**
  - e.g., Product, Inventory
- **Out-of-the-box** "starter" set of grammars and configuration options for rapid deployment

漢

דָנִיאֵלָה

Núñez

привет

ΤΥΦΧΨΩ

שלום

和

Matching Algorithms – Comparison

Comparison of Matching Technologies & Implications

Percent-based, or Weighted Field Scoring approach

85%

Granular, Pattern-based approach

Weighted Field Scoring – Most Commonly Used

Limitations of this Approach:

Match 1: Turns out to be a True Match

Charles	Taylor	Male	19480128	Liberia	321-45-9876	SCORE 85
Charles	Taylor		19480128	USA	321-45-9876	

Match 2: Turns out to be a False Match

Charles	Taylor	Male	19480128	Liberia	321-45-9876	SCORE 85
Charles	T	Male	19480128	Liberia	211-45-9878	

- Matches can receive the same score, regardless of the **reason** for the match
  - Both True Match and False Match received the same score
- No transparency to the reasons (you only see the score) → Harder to tune the matching criteria

Weighted Field Scoring – Most Commonly Used

Limitations of this Approach:

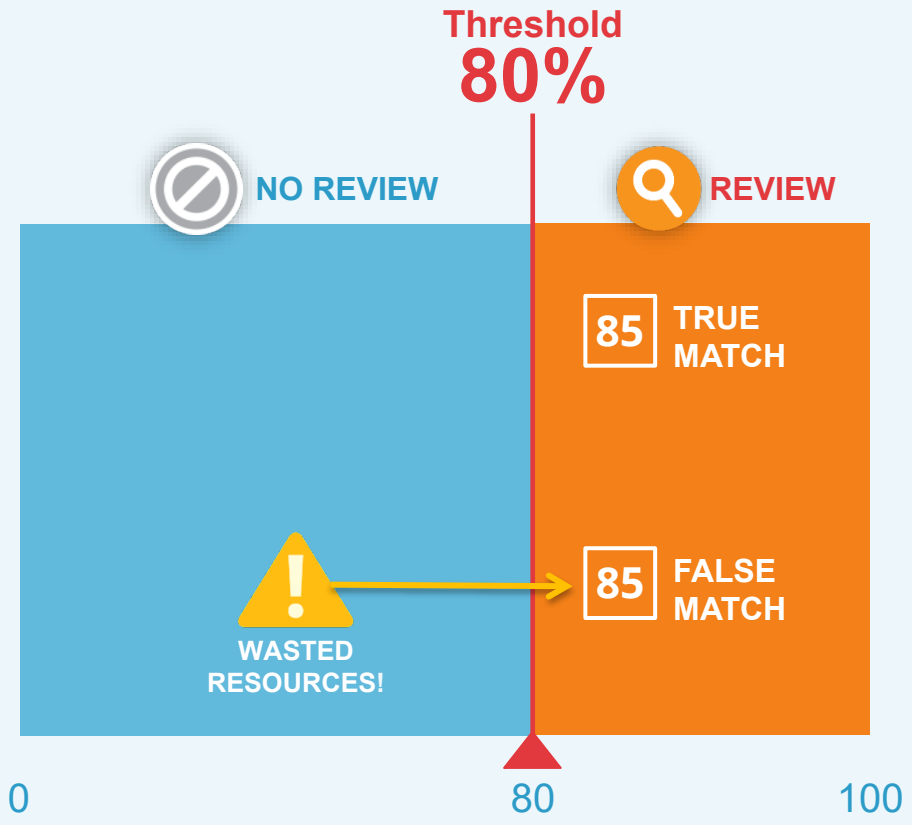
Charles	Taylor	Male	19480128	Liberia	321-45-9876	SCORE 85
Charles	Taylor		19480128	USA	321-45-9876	

True Match

Charles	Taylor	Male	19480128	Liberia	321-45-9876	SCORE 85
Charles	T	Male	19480128	Liberia	211-45-9878	

False Match

- When setting the threshold, you can make only broad, sweeping adjustments
- This creates unnecessary review work



Weighted Field Scoring – Most Commonly Used

Limitations of this Approach:

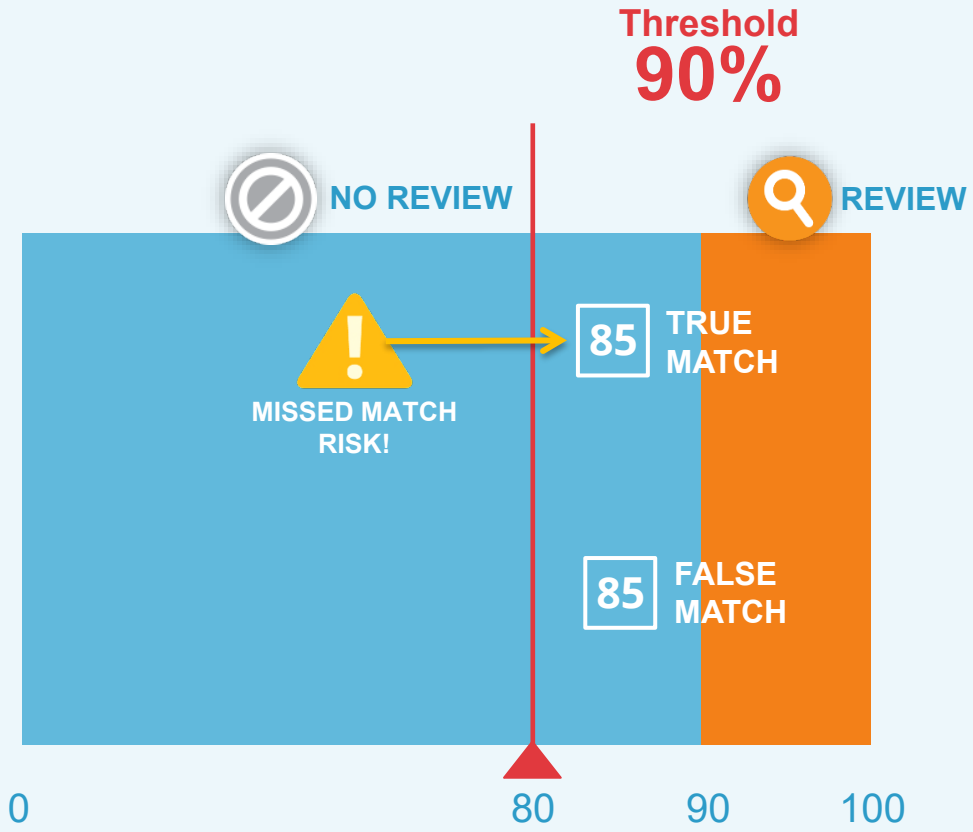
Charles	Taylor	Male	19480128	Liberia	321-45-9876	SCORE 85
Charles	Taylor		19480128	USA	321-45-9876	

True Match

Charles	Taylor	Male	19480128	Liberia	321-45-9876	SCORE 85
Charles	T	Male	19480128	Liberia	211-45-9878	

False Match

- Not granular enough – as you raise your threshold, you manage to avoid the False Match but you miss the True Match
- Hidden risk as you change the threshold
  - i.e., more missed matches with a higher threshold
- Results in under- or over-matching, the "seesaw" effect between too many false positives and the risk of missing a true match
- Therefore, hard to tune the rules – cannot specify the exact type of matches you want



Synchronos®

Enlighten®
Data Quality

Instant 360°™
Customer view

FinScan® Compliance
Sanctions and PEP Screening

Synchronos® Platform

Instant 360° View

- Enterprise Data Model with Customer as Key Entity
- Ability to Add to the Model and User Interface
- Data Analysis and Monitoring
- Integrated Data Quality Software Suite
- 360 Degree View of Customer
- Builds Customer Golden Record – Intelligent Combine Rules
- Customer-to-Customer Relationship Types
  - Originated from source, e.g., payor
  - Derived, e.g., possible customer match
- Household View of Customer
- Customer Information User Interface and Reports
- Supports DQ and Associated Policies and Processes