
ORACLE®

EDQ

Parsing Essentials

Enterprise Data Quality Product Management

January, 2015

ORACLE

Agenda

- Text Analysis Overview
  - What is parsing and when would we use it?
- Phrase profiling
  - Analyse text fields for content, return most common words and phrases
  - Create reference data
- Parsing
  - Understand, validate and transform data

Note: This deck is concerned only with the capabilities of core EDQ (it does not cover the Product Data Extension)

Text Analysis Overview

What is Parsing?

TITLE	FORENAME	INITIAL	SURNAME	ADDRESS_LINE_1	ADDRESS_LINE_2	ADDRESS_LIN	ADDRESS_LINE_
Mr	& Mrs C	P	Hoskins	21 Railway Terrace	Lindal In Furness	Ulverston	Cumbria
Mr	Roy						Glos
	#						
Mr	Colin						erbyshire
Mrs	Catherine						anchester
Mrs	Katherine						astlefield, Manch

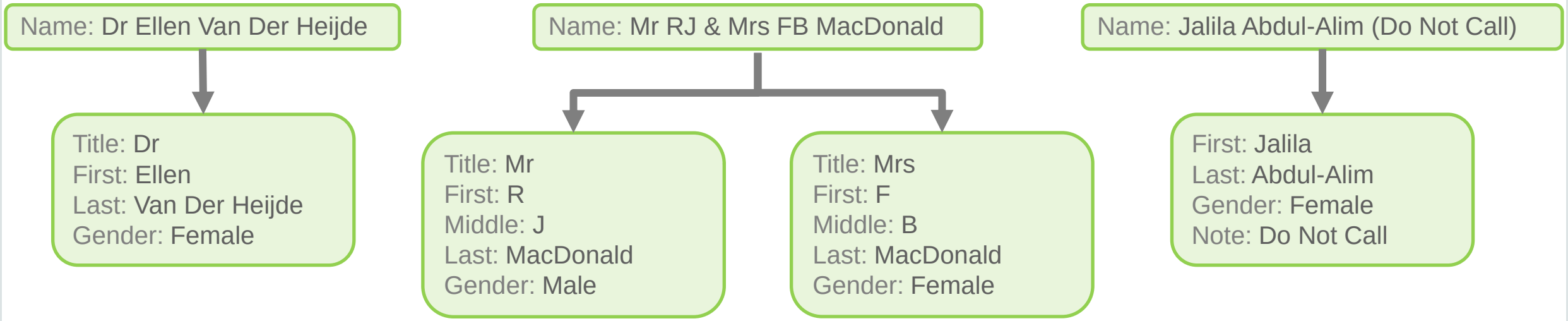
"Parsing is the application of business rules to textual data in order to understand and validate it, and, if required, improve its structure in order to make it fit for purpose."

title	fname							workPhone
Mr	EMMANU							02085948283
Mr and Mrs	PETER							01491572831
Mr	TEST	13090102	TEST	TEST			CF67UU	0201000000
Mr	EMMANUEL	MATTHEWS	OLD BAKEHOUSE	SOUTH STREET	BICESTER	OXFORDSHIRE	OX254NE	02085948283
Mr	COLIN	HOSKINS	21 RAILWAY TERRACE	LINDAL IN FURNE		CUMBRIA	LA120LQ	

Why Parse Free Text Fields?

- Understand and add structure to unstructured data
  - Free format data entry, misplaced data
  - Hidden duplication
- Extract key data into fields
  - e.g. preparation for matching
- Restructuring data
  - Migration
- Contextual auditing
  - Assessing semantic validity across several attributes

Example – Parsing party data



- Split names and name elements
- Identify individuals and businesses
- Derive additional attributes

Typical business problems (1)

We have 10m customer records, in various systems:

TITLE	FORENAME	INITIAL	SURNAME	ADDRESS_LINE_1	ADDRESS_LINE_2	ADDRESS_LIN	ADDRESS_LINE_	F
Mr	& Mrs C	P	Hoskins	21 Railway Terrace	Lindal In Furness	Ulverston	Cumbria	
Mr	Roy		Greenhalgh [DECEASED]	Townwell House	Cromhall	Wotton-Under-E	Glos	
	#	# #	Rock Nominees Limited 292	Granville House	25 Luke Street	London		
Mr	Colin	N	Roberts-Slack	Tintwistle Sunday School	Woodhead Road	Tintwistle	Derbyshire	
Mrs	Catherine	A	Gough	8 Rochdale House	Slate Wharf	Castlefield	Manchester	
Mrs	Katherine		Gough	Flat 8	Rochdale House	15 Slate Wharf	Castlefield, Manch	M

title	fname	lname	addr1	addr2	addr4	addr5	pocode	workPhone	
Mr	EMMANUEL	MATTHEWS	10 GREYS ROAD		HENLEY	OXON	RG91TE	02085948283	
Mr and Mrs	PETER	&_JANE FORSNOR		408 HOWLANDS	WELWYN G	HERTFORDSHIRE	AL74HB	01491572831	0
Mr	TEST	13090102	TEST	TEST			CF67UU	0201000000	0
Mr	EMMANUEL	MATTHEWS	OLD BAKEHOUSE	SOUTH STREET	BICESTER	OXFORDSHIRE	OX254NE	02085948283	0
Mr	COLIN	HOSKINS	21 RAILWAY TERRACE	LINDAL IN FURNE		CUMBRIA	LA120LQ		

Q. How many ***customers*** do we actually have?

Note: Matching may also be required to find duplicate customers.

Typical business problems (2)

CU_NO	NAME
13861	Roberta R F REYNOLDS
13865	Mr & Mrs J K STEWART
13870	Andrew James SUTHERLAND
15168	Moira BULLIVANT (Do Not Call)
[illegible]	Miss Catherine WALSH

NamePrefix	FirstName	MidName	LastName	NameSuffix
	BERNARD & GUYLENE		ANGRAND	
Mr.	Robert	A	Alvarez	Unknown
	Mark	Duane	Barker	
	SAM JR & LEA		BARR	
	C L		BLANCO	
Mr.	Clayton	J.	Rice	III

Q. How do we migrate these records to a single system (and table structure)?

Address1	Address2	Address3	Address4	PostCode
300/A Annan Road	Dumfries	Dumfriesshire		
300a Annan Road			DUMFRIES	DG1 3JE
304 Annan Road			DUMFRIES	DG1 3JE
[illegible]		[illegible]		

Q. How do we match these records accurately?

Issues to overcome (1)

Invalid data:

title	fname	lname	addr1	addr2	addr4	addr5	pocode	workPhone
Mr	EMMANUEL	MATTHEWS	10 GREYS ROAD		HENLEY	OXON	RG91TE	02085948283
Mr and Mrs	PETER	&_JANE FORSNOR		408 HOWLANDS	WELWYN C	HERTFORDSHIRE	AL74HB	01491572831
Mr	TEST	13090102	TEST	TEST			CF67UU	0201000000
Mr	EMMANUEL	MATTHEWS	OLD BAKEHOUSE	SOUTH STREET	BICESTER	OXFORDSHIRE	OX254NE	02085948283
Mr	COLIN	HOSKINS	21 RAILWAY TERRACE	LINDAL IN FURNE		CUMBRIA	LA120LQ	

Misuse of fields:

Title	Forename	Initials	Surname	Honours
MR	MICHAEL		LEWIS	
MISS	LESLEY	MCLELLAN	SHEILDAIG FARM	
			MISS G CRON	
MISS	SHEILA	L	MANSOUR	
MISS		E	MCDONALD	C/O MS E WILSON

Issues to overcome (2)

Inadequate structure (e.g. for matching):

Address1	Address2	Address3	Address4	PostCode
300/A Annan Road	Dumfries	Dumfriesshire		
300a Annan Road			DUMFRIES	DG1 3JE
304 Annan Road			DUMFRIES	DG1 3JE

Abbreviations, misspellings and truncation:

Building	Thoroughfare Nı	Thoroughfare Name	Locality
GARDEN HSE			LLANARTHNEY
CRONEIL COTAGE		DUNTIBLAE RD	KIRKINTILLOCH
RIVERSIDE HO	103	MONROE ROAD	
		NERSTON INDUSTRIAL ESTAT	E KILBRIDE

Issues to overcome (3)

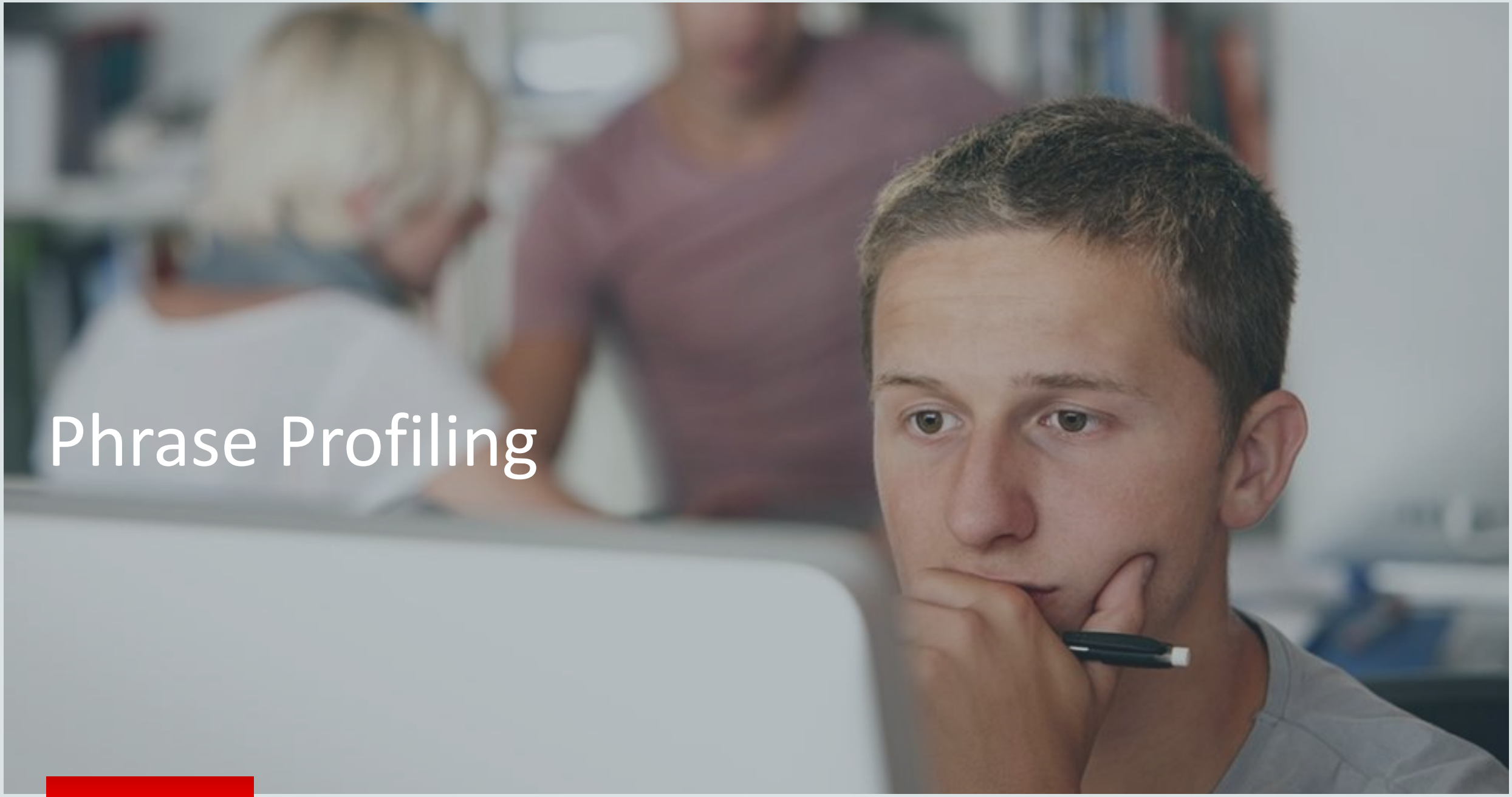
Duplication:

Title	Forename	Initials	Surname	Honours
MRS CHUNG T/A			MRS CHUNG T/A	SUPERWOK

Q. Should these records be split into many?

Title	Forename	Initials	Surname	Honours
MS&MR	P S		COOPE/MILLER	
MR P	&	MRS E	BARRETT	
MR P FERGUSON	MR N MURRAY	&	MRS J THOMAS	COOK SOLICITORS
MR & MRS	D		ROSS	

Phrase Profiling



Phrase Profiling

- Dovetails with Parsing to analyse text fields
  - A quick way of creating the data to build classification lists for parsing
  - Classify key words and phrases in the data
  - Which parsing rules to apply to which attributes
- Once Parsing is configured, use Phrase Profiling to understand ‘unclassified’ data
  - i.e. what the Parser doesn’t understand yet

Common words and phrases

- Example: names and addresses

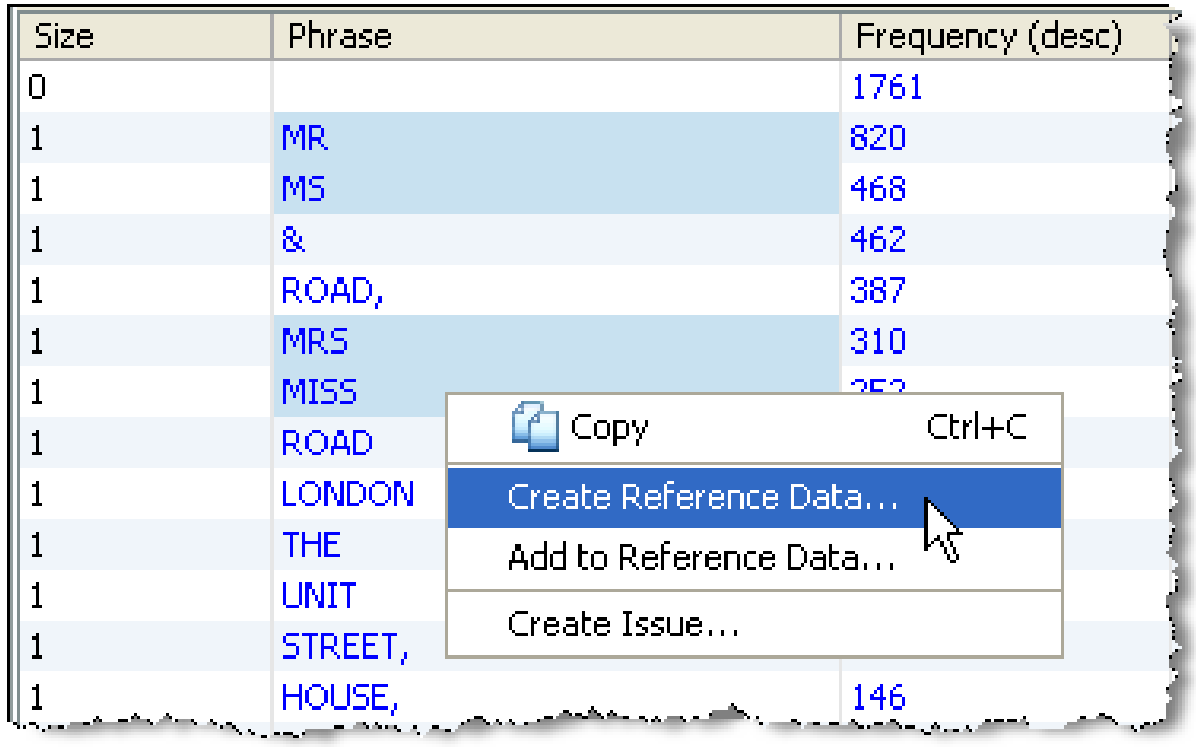
Size	Phrase	Frequency (desc)	TITLE freq.	NAME freq.	BUSINESS freq.	ADDRESS1 freq.	ADDRESS2 freq.	ADDRESS3 fre
0		1761	139	1	337	1	80	970
1	MR	820	819	0	0	1	0	0
1	MS	468	468	0	0	0	0	0
1	&	462	0	0	436	14	1	11
1	ROAD,	387	0	0	0	386	1	0
1	MRS	310	310	0	0	0	0	0
1	MISS	252	252	0	0	0	0	0
1	ROAD	242	0	0	0	231	10	1
1	LONDON	238	0	1	0	20	194	23
1	THE	190	1	0	82	104	3	0
1	UNIT	182	0	0	0	182	0	0
1	STREET	147	0	0	0	47	0	0

Identified words and phrases

Locations of words and phrases

Build Reference Data

- Manage reference data for use in parsing



Identify misplaced data

- Example: misplaced ‘MR’

Size	Phrase	Frequency (desc)	TITLE freq.	NAME freq.	BUSINESS freq.	ADDRESS1 freq.	ADD
0		1761	139	1	337	1	80
1	MR	820	819	0	0	1	0
1	MS	468	468	0	0	0	0
1	&	462	0	0	436	14	1
1	ROAD,	387	0	0	0	386	1
1	MRS	310	310	0	0	0	0
1	MISS	252	252	0	0	0	0
1	ROAD	242	0	0	0	231	10

- Drill down to investigate

TITLE	NAME	BUSINESS	ADDRESS1	ADDRESS2	ADDRESS3	POSTCODE
Mr	Peter CROCKER		Mr Crocker, First Floor Flat	80 Grenville Road,	Plymouth	PL4 9PY

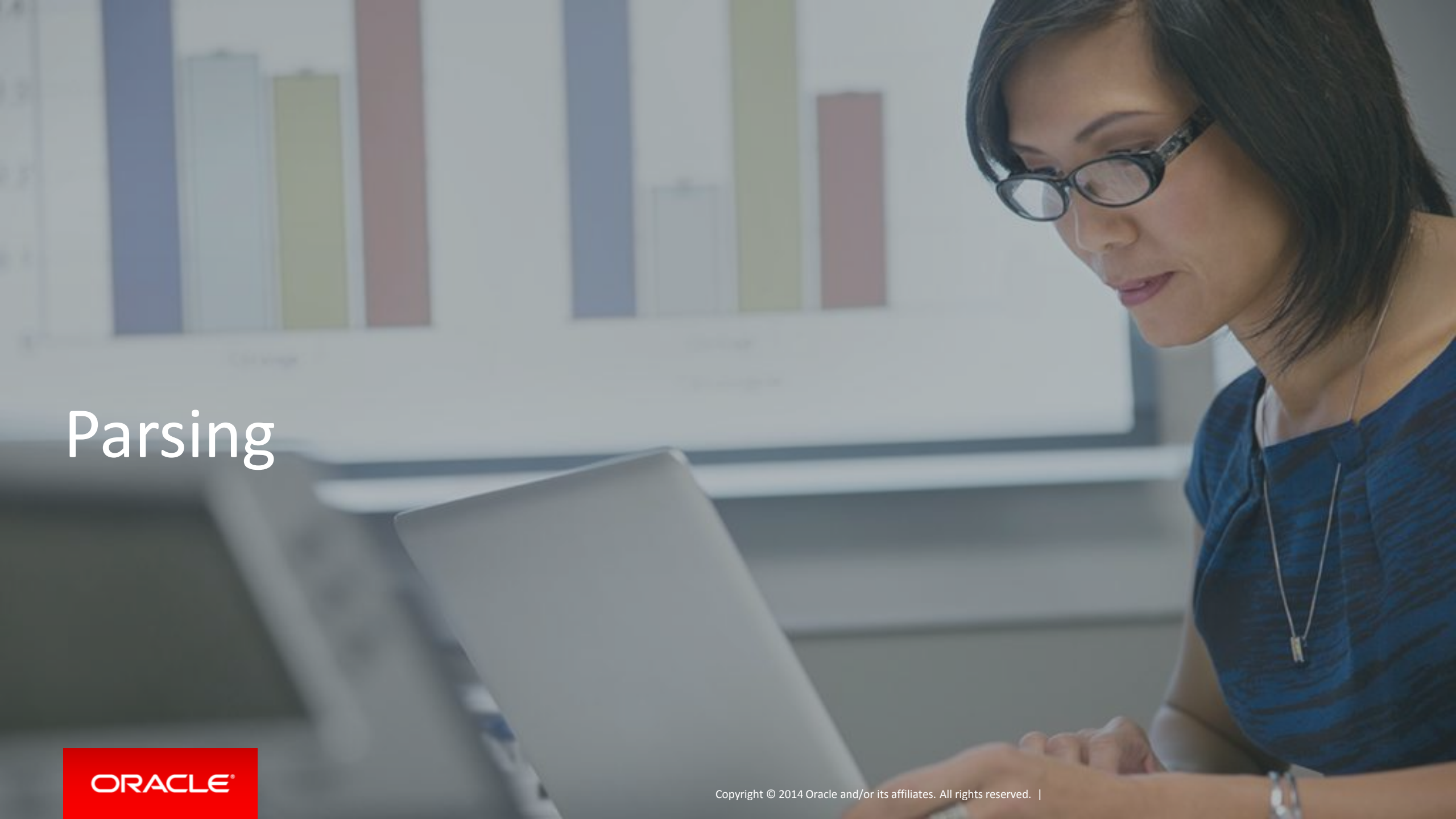
Identify and manage ambiguities

Size	Phrase	Frequency	TITLE freq. ...	NAME freq.	BUSINESS freq.	ADDRESS1 freq.	AD
2	FIRST FLOOR	4	0	0	0	4	0
1	EDWARD	7	0	5	0	2	0
1	BB1	5	0	0	0	0	0
1	BAR	12	0	0	5	7	0
1	BB2	4	0	0	0	0	0
1	BB5	3	0	0	0	0	0
1	BAY	3	0	0	0	2	1
1	BB8	2	0	0	0	0	0
1	LIBRARY,	2	0	0	0	2	0
1	VICTORIA	11	0	1	0	10	0
	BD4	3	0	0	0	0	0

- Example: ‘Victoria’ might be classified as a valid given name, and ‘Victoria Centre’ as a valid building

ADDRESS1	AD
Victoria Corn Mills, Denby Dale	H
124 Victoria Road,	F
The Marine Laboratory, Victoria Road	A
Victoria Road South,	
Victoria Rd,	H
308c Victoria Centre	
Victoria St,	E
Unit A, Victoria Centre	C
10-12 Victoria Lane,	H
10/22 Victoria Street,	B

Parsing



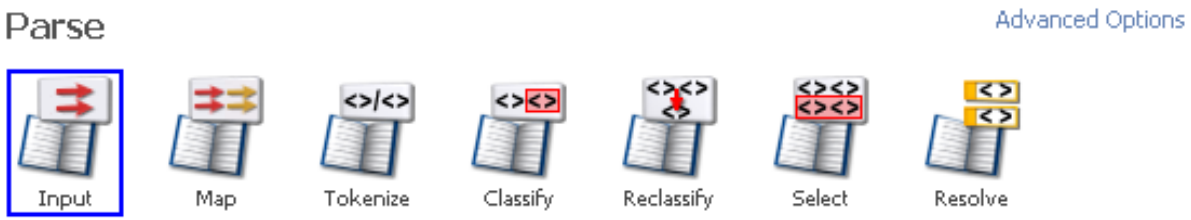
Parsing overview

- Analyse and understand the meaning of data
  - Lists of values – dictionaries or syntax
  - Patterns
  - Grammar of the ‘language’ used
  - Rules
- Validate and structurally improve data
  - e.g. identify a name in an address column and map it to a new column in a different structure

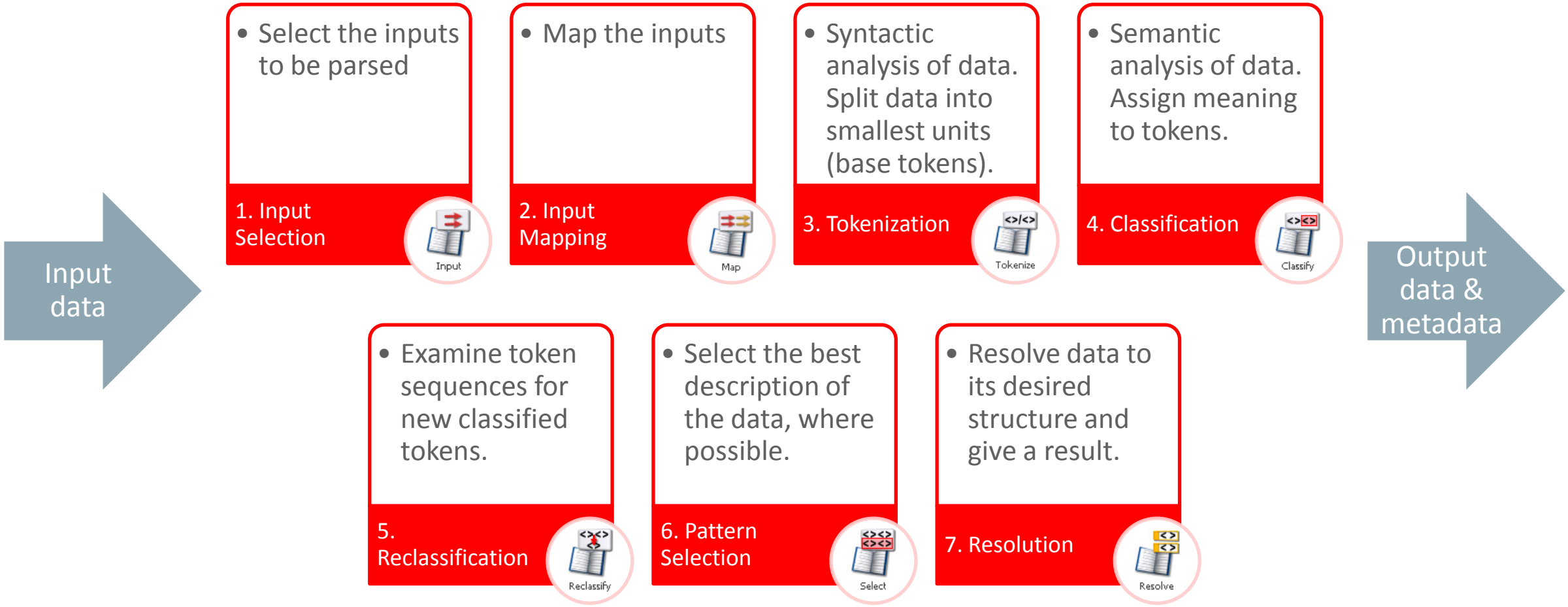
The EDQ Parse processor

- Parse Processor

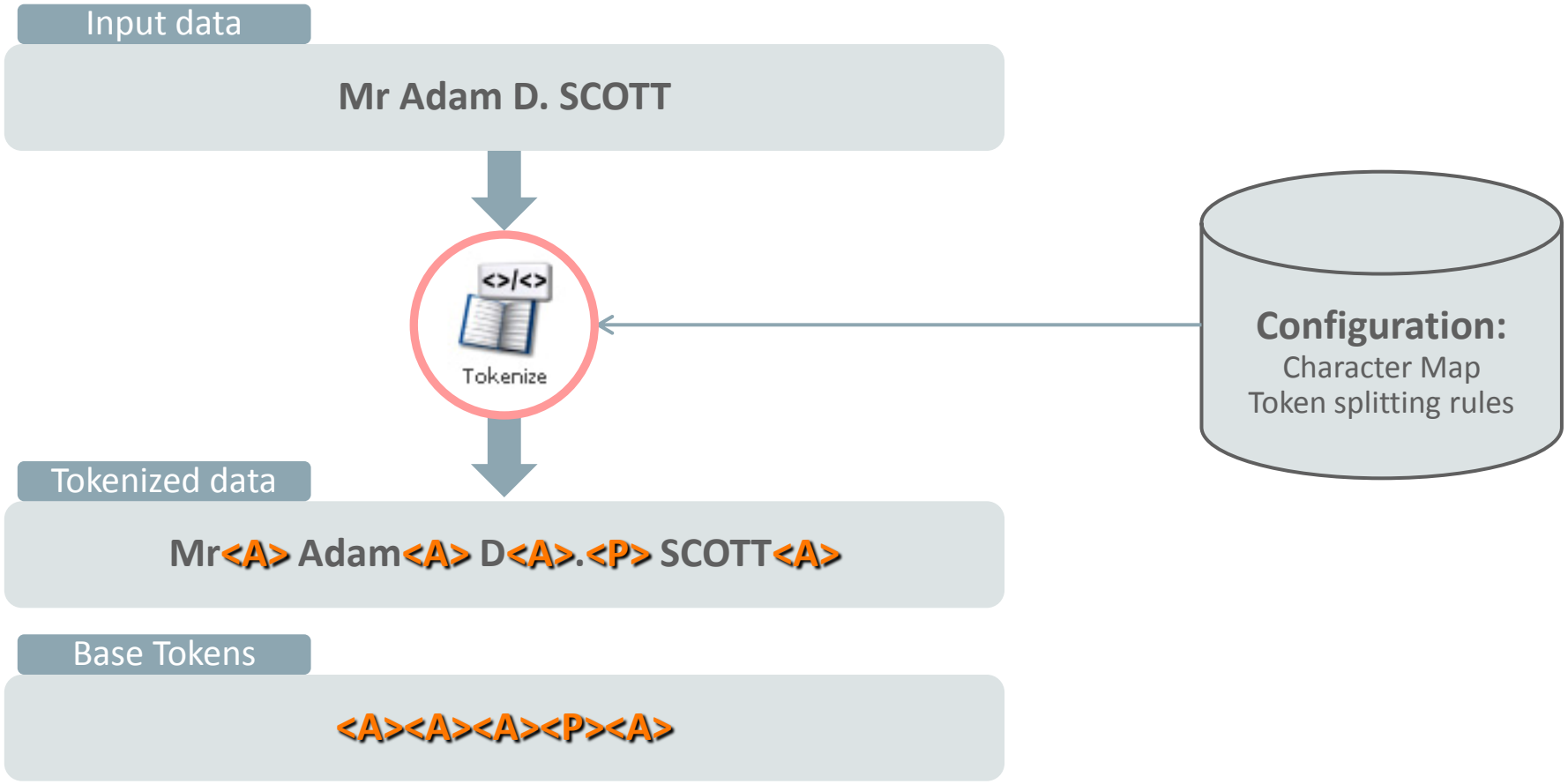
- Seven sub-processors



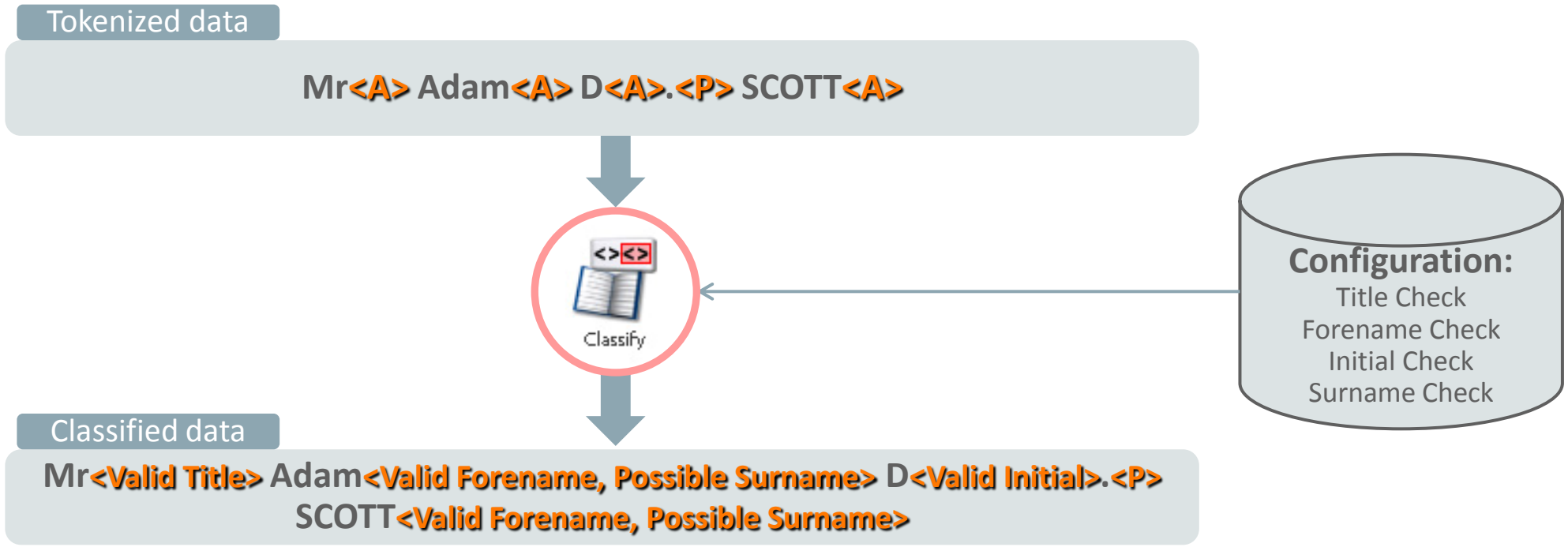
The EDQ Parse processor



Tokenization

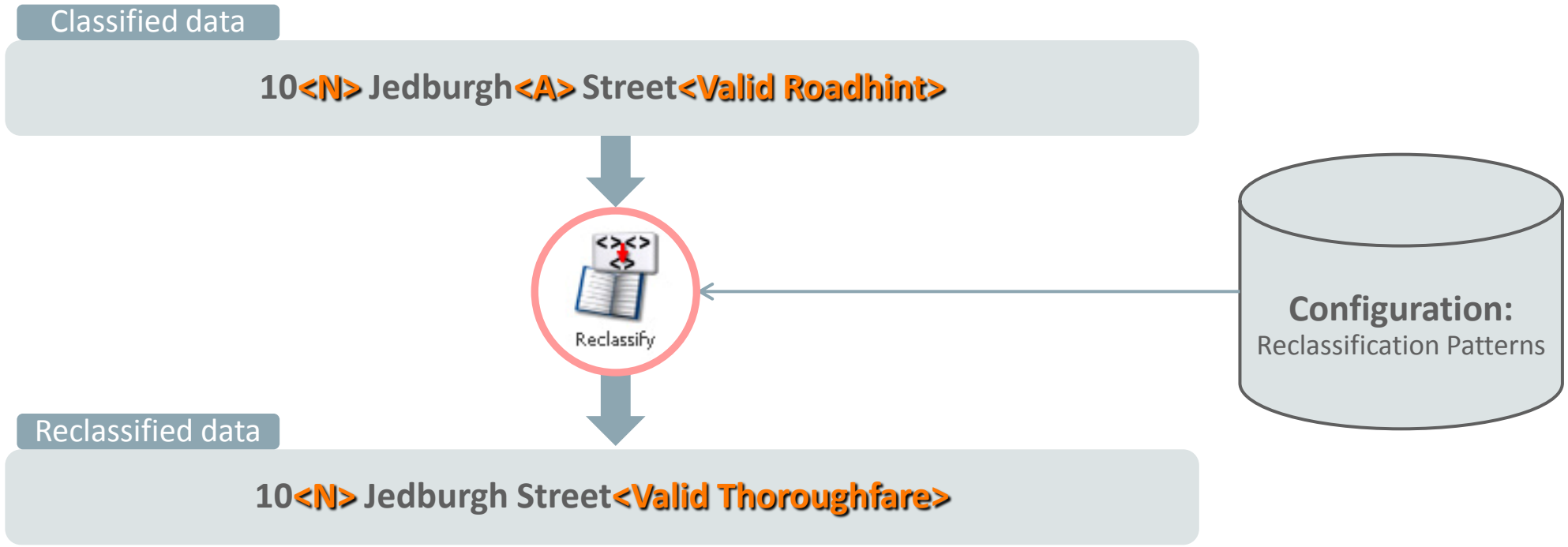


Classification



Note: Rules and their application to attributes completely configurable.

Reclassification



Reclassification is an optional way of creating new tokens from sequence of tokens

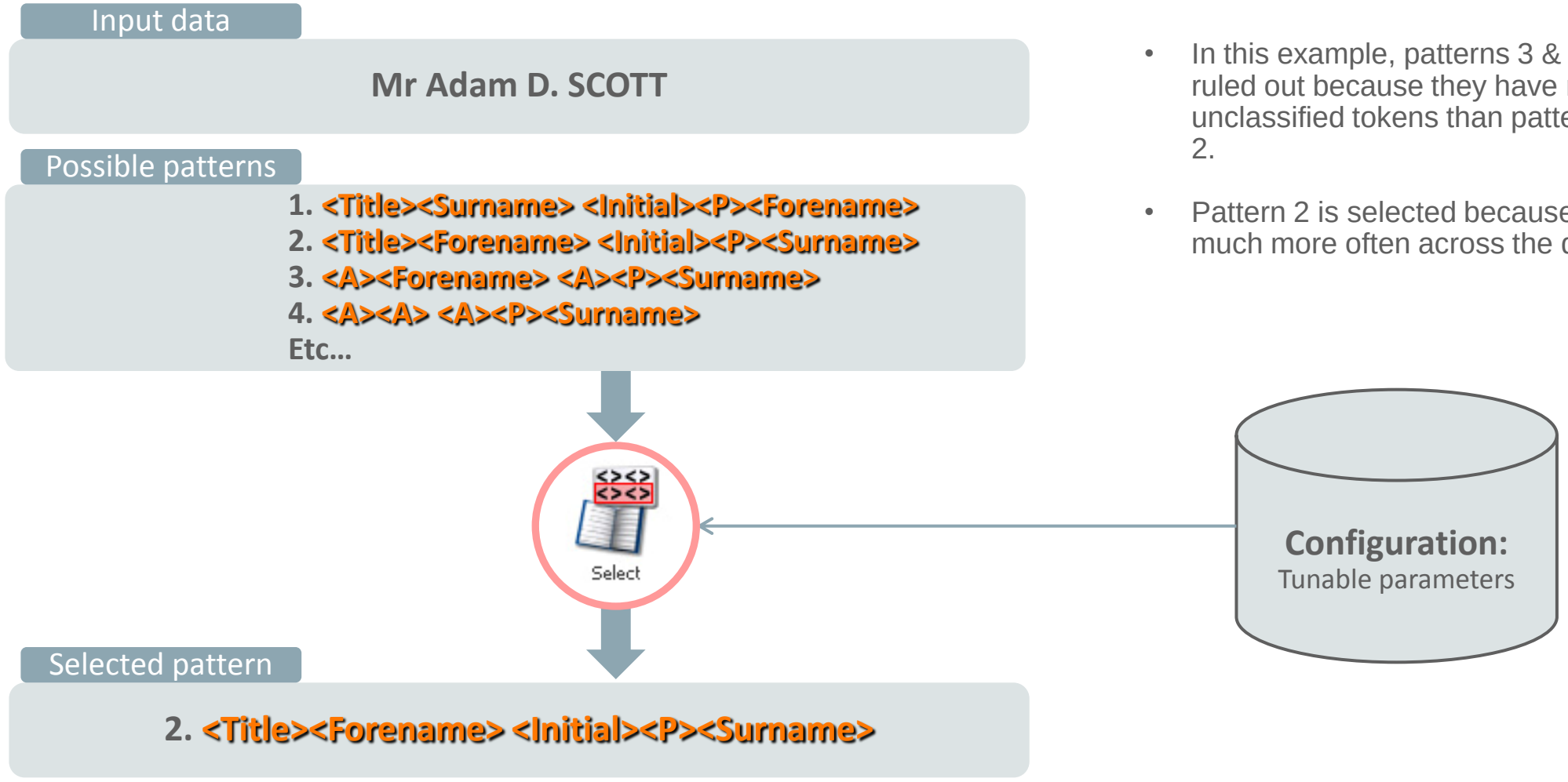
Example Rule:

Match pattern: <N>([<A>]{1,2}<Roadhint>)

Reclassify as: <Thoroughfare>

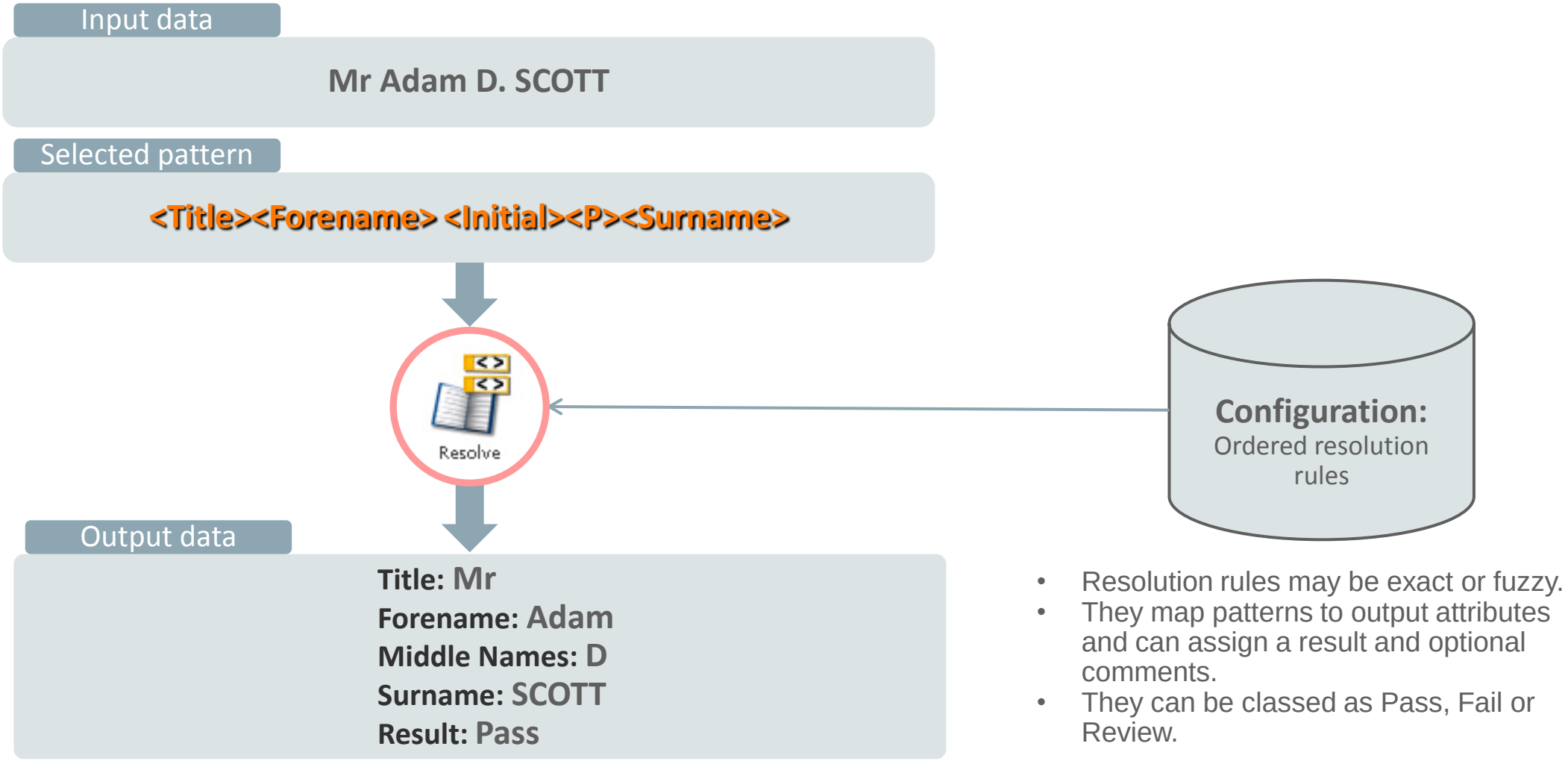
Note: Rules and their application to attributes completely configurable.

Pattern Selection



- In this example, patterns 3 & 4 are ruled out because they have more unclassified tokens than patterns 1 and 2.
- Pattern 2 is selected because it occurs much more often across the data set.

Resolution



- Resolution rules may be exact or fuzzy.
- They map patterns to output attributes and can assign a result and optional comments.
- They can be classed as Pass, Fail or Review.


ORACLE®