ORACLE

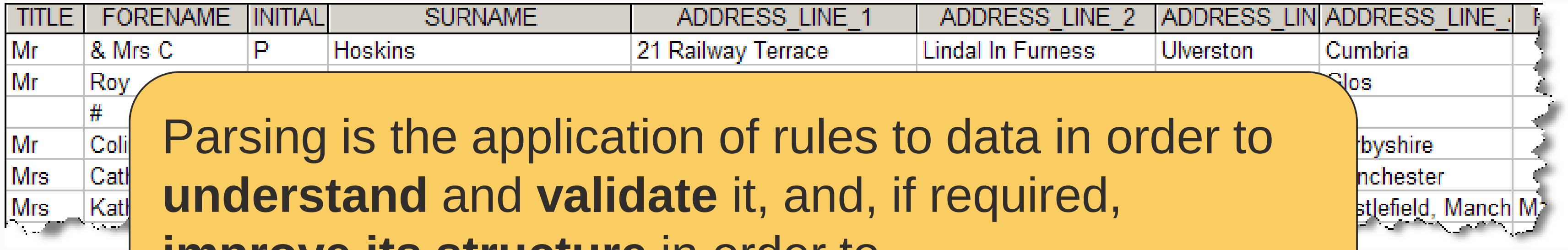
Oracle Enterprise Data Quality

Text Analysis and Parsing Essentials

Product Development

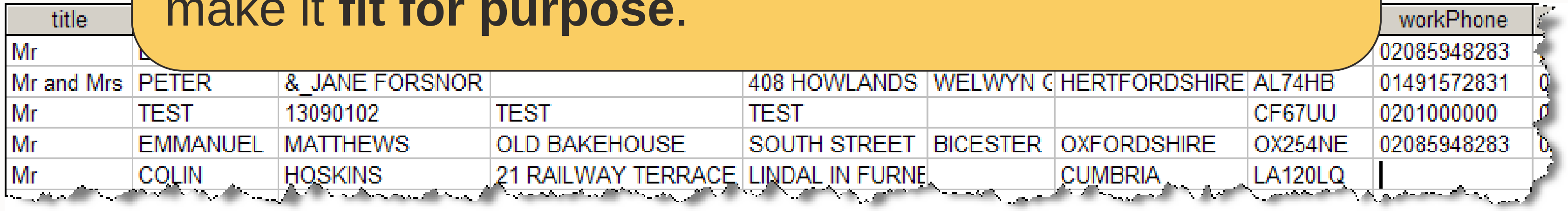
Text Analysis Overview

What is parsing?



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						Glos	
	#							
Mr	Coli						rbyshire	
Mrs	Catl						nchester	
Mrs	Katl						stlefield, Manch	M

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



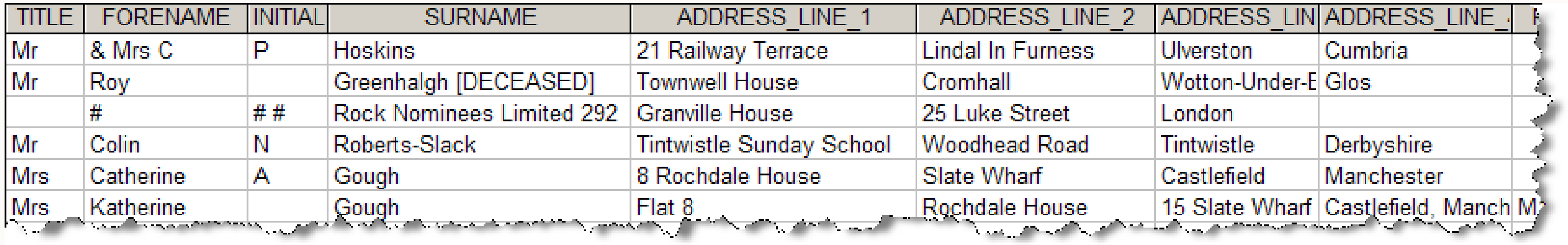
title								workPhone	
Mr	L							02085948283	
Mr and Mrs	PETER	&_JANE FORSNOR		408 HOWLANDS	WELWYN (	HERTFORDSHIRE	AL74HB	01491572831	0
Mr	TEST	13090102	TEST	TEST			CF67UU	0201000000	0
Mr	EMMANUEL	MATTHEWS	OLD BAKEHOUSE	SOUTH STREET	BICESTER	OXFORDSHIRE	OX254NE	02085948283	0
Mr	COLIN	HOSKINS	21 RAILWAY TERRACE	LINDAL IN FURNE		CUMBRIA	LA120LQ		

Why Parse Free Text Fields?

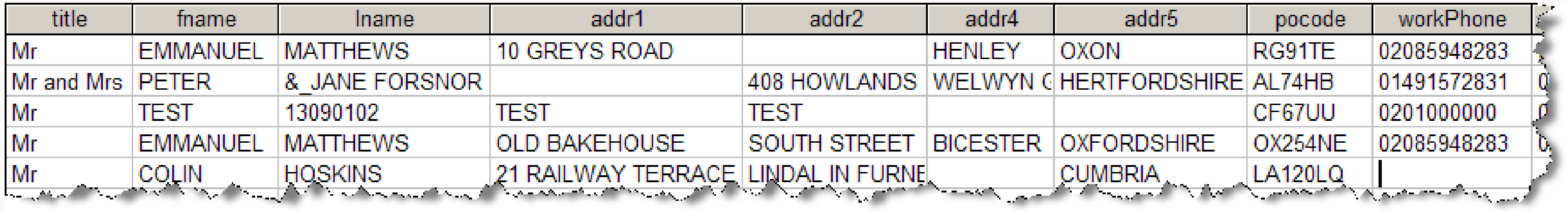
- **Understand and add structure to unstructured data**
  - Free format data entry, misplaced data
  - Hidden duplication
- **Extract key data into fields**
  - For example, to prepare data for matching and improve match efficacy
- **Restructuring data**
  - Migrations
- **Contextual auditing**
  - Assessing semantic validity across several attributes

Typical business problems (1)

We have 10m customer records, in various systems:



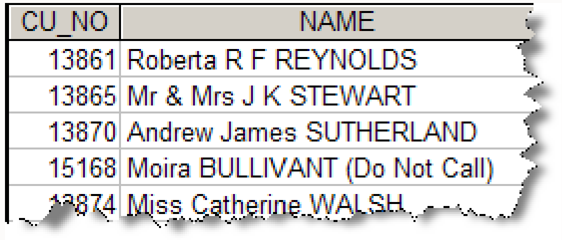
TITLE	FORENAME	INITIAL	SURNAME	ADDRESS_LINE_1	ADDRESS_LINE_2	ADDRESS_LIN	ADDRESS_LINE_	F
Mr	& Mrs C	P	Hoskins	21 Railway Terrace	Lindal In Furness	Ulverston	Cumbria	
Mr	Roy		Greenhalgh [DECEASED]	Townwell House	Cromhall	Wotton-Under-E	Glos	
	#	# #	Rock Nominees Limited 292	Granville House	25 Luke Street	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



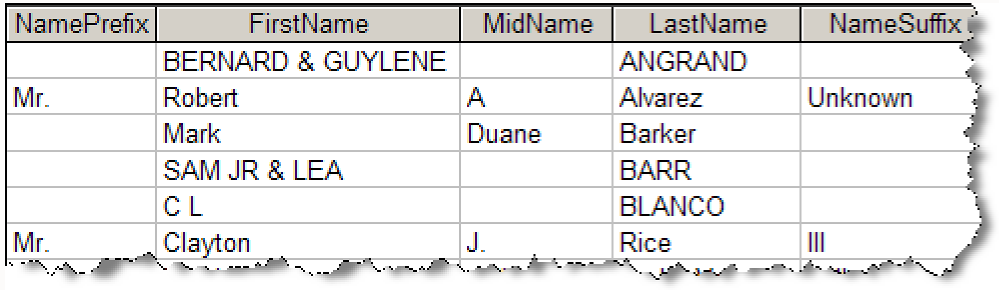
title	fname	lname	addr1	addr2	addr4	addr5	pocode	workPhone	
Mr	EMMANUEL	MATTHEWS	10 GREYS ROAD		HENLEY	OXON	RG91TE	02085948283	
Mr and Mrs	PETER	& JANE FORSNOR		408 HOWLANDS	WELWYN (	HERTFORDSHIRE	AL74HB	01491572831	0
Mr	TEST	13090102	TEST	TEST			CF67UU	0201000000	0
Mr	EMMANUEL	MATTHEWS	OLD BAKEHOUSE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

How many customers do we actually have?

Typical business problems (2)

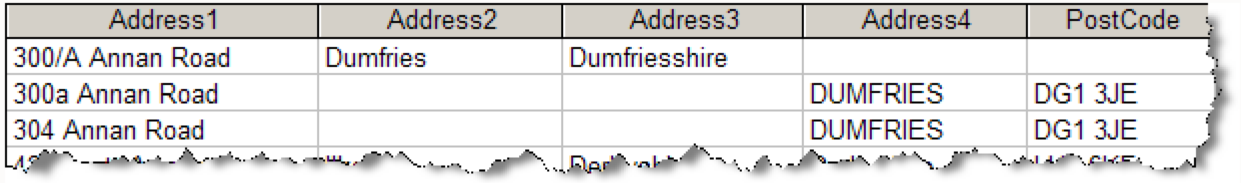


CU_NO	NAME
13861	Roberta R F REYNOLDS
13865	Mr & Mrs J K STEWART
13870	Andrew James SUTHERLAND
15168	Moira BULLIVANT (Do Not Call)
13874	Miss Catherine WALSH



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

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

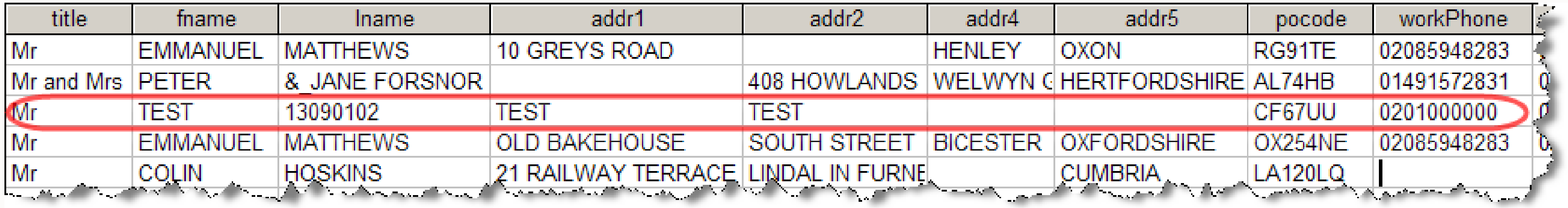


Address1	Address2	Address3	Address4	PostCode
300/A Annan Road	Dumfries	Dumfriesshire		
300a Annan Road			DUMFRIES	DG1 3JE
304 Annan Road			DUMFRIES	DG1 3JE

How do we match these records accurately?

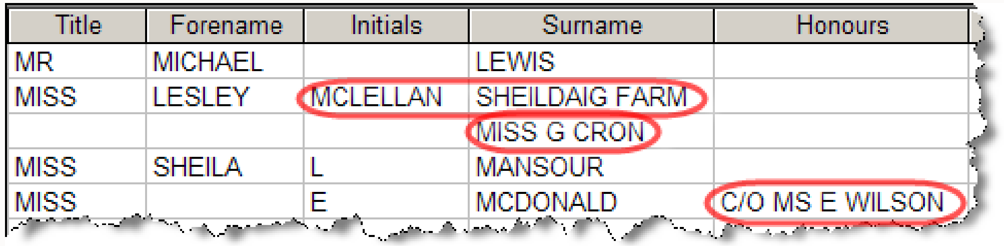
Issues to overcome (1)

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 (	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	

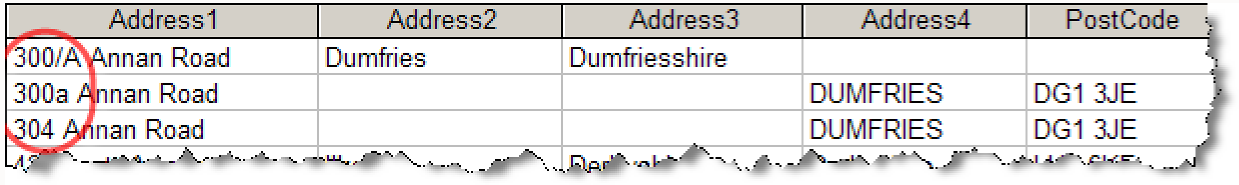
Misuse of fields:



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

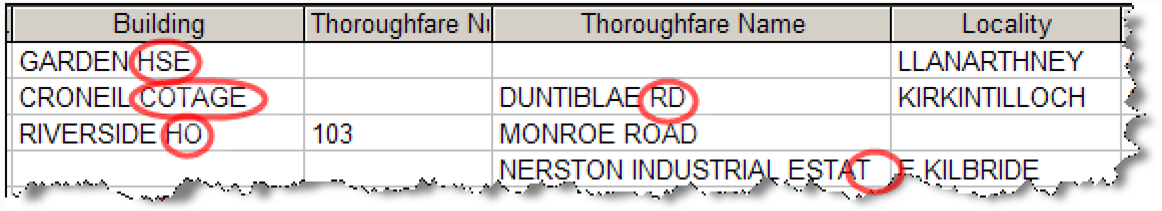
Issues to overcome (2)

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

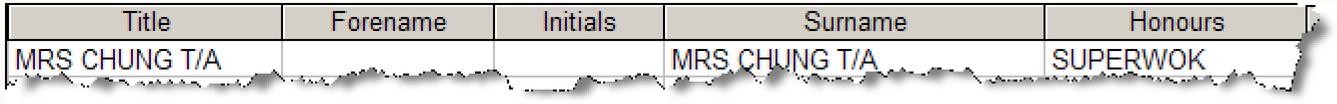
Abbreviations, mis-spellings and truncation:



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

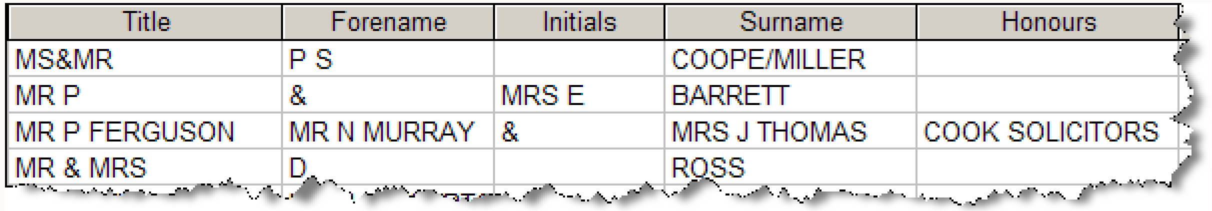
Issues to overcome (3)

Duplication:



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

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	

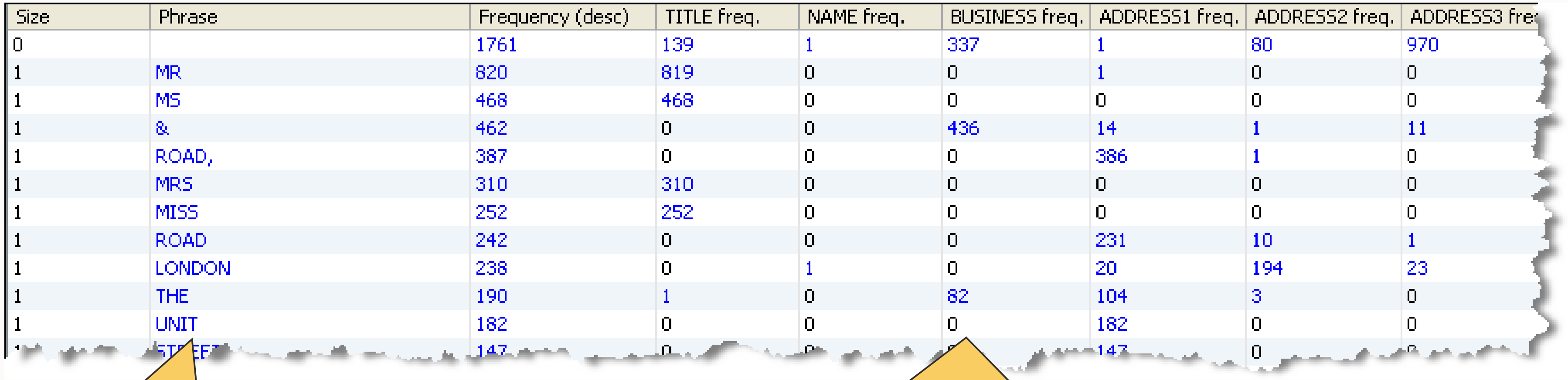
Phrase Profiling

Phrase Profiling

- **Dovetails with Parsing to analyse text fields**
  - A quick way of creating the data to build classification reference data for parsing
  - Find and classify key words and phrases in the data
  - Understand 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**



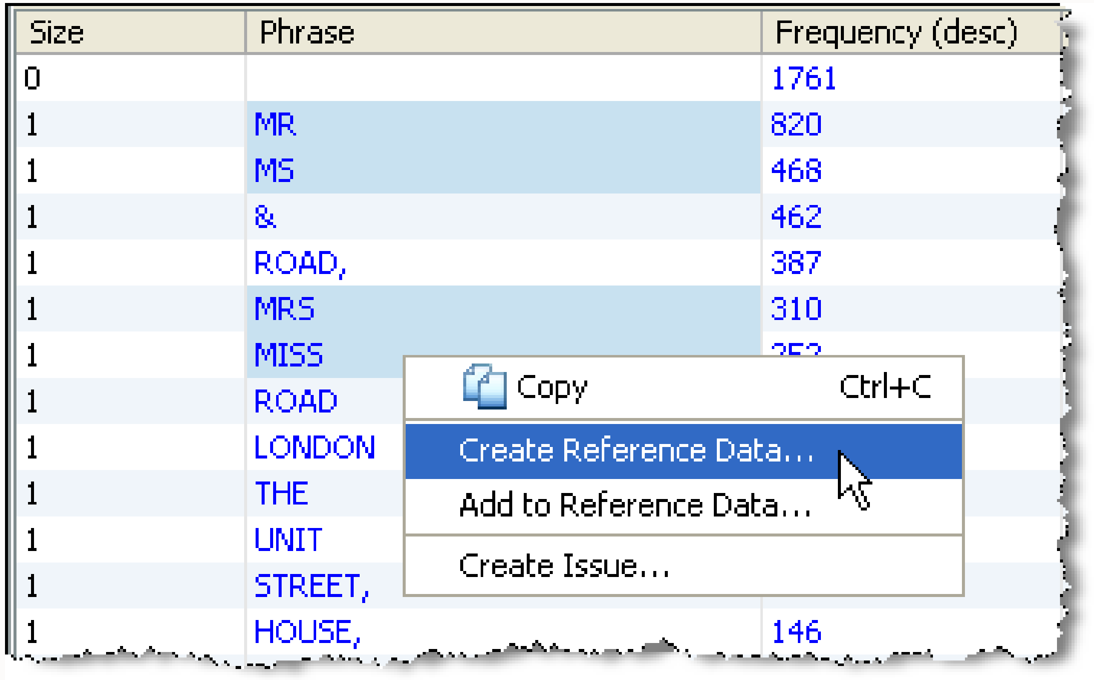
Size	Phrase	Frequency (desc)	TITLE freq.	NAME freq.	BUSINESS freq.	ADDRESS1 freq.	ADDRESS2 freq.	ADDRESS3 fre
0		1761	139	1	337	1	80	970
1	MR	820	819	0	0	1	0	0
1	MS	468	468	0	0	0	0	0
1	&	462	0	0	436	14	1	11
1	ROAD,	387	0	0	0	386	1	0
1	MRS	310	310	0	0	0	0	0
1	MISS	252	252	0	0	0	0	0
1	ROAD	242	0	0	0	231	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	147	0	0

Identified words and phrases

Locations of words and phrases

Reference Data

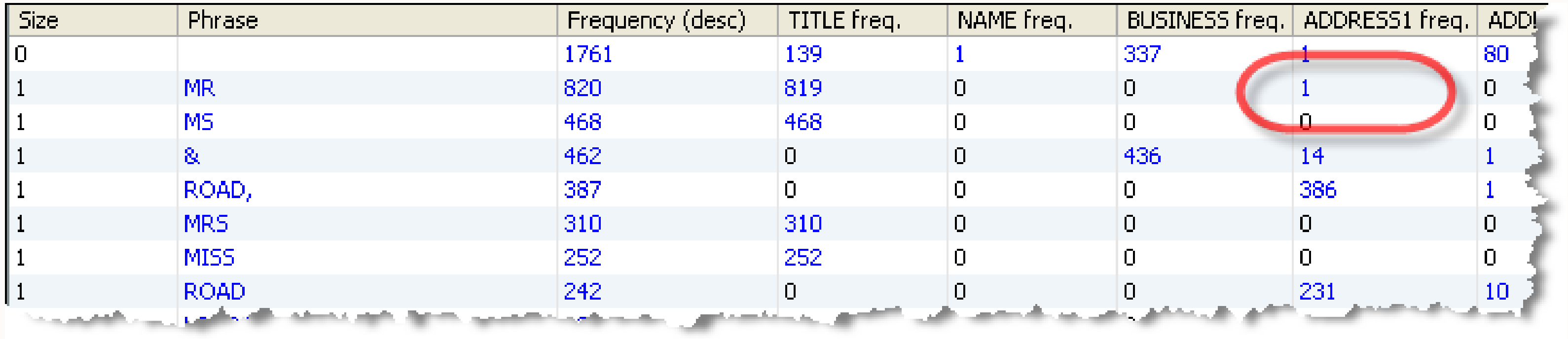
- **Create and update reference data for use in parsing from profiling results**



Size	Phrase	Frequency (desc)
0		1761
1	MR	820
1	MS	468
1	&	462
1	ROAD,	387
1	MRS	310
1	MISS	253
1	ROAD	
1	LONDON	
1	THE	
1	UNIT	
1	STREET,	
1	HOUSE,	146

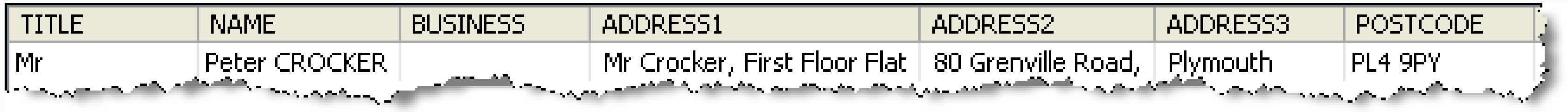
Identify misplaced data

- Example: misplaced ‘MR’



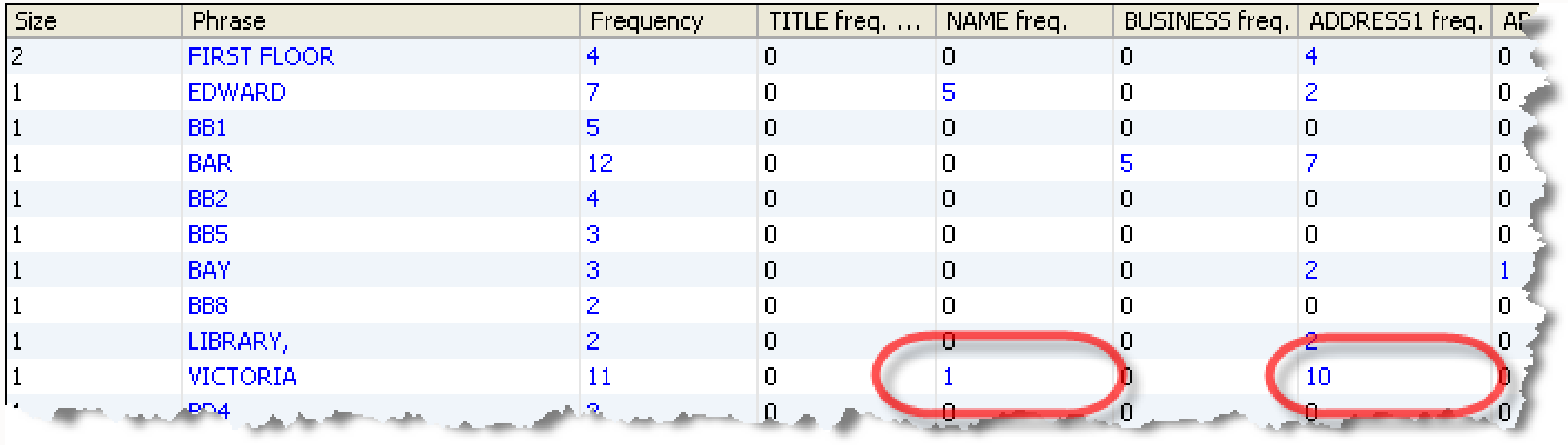
Size	Phrase	Frequency (desc)	TITLE freq.	NAME freq.	BUSINESS freq.	ADDRESS1 freq.	ADD
0		1761	139	1	337	1	80
1	MR	820	819	0	0	1	0
1	MS	468	468	0	0	0	0
1	&	462	0	0	436	14	1
1	ROAD,	387	0	0	0	386	1
1	MRS	310	310	0	0	0	0
1	MISS	252	252	0	0	0	0
1	ROAD	242	0	0	0	231	10

- Drill down to investigate



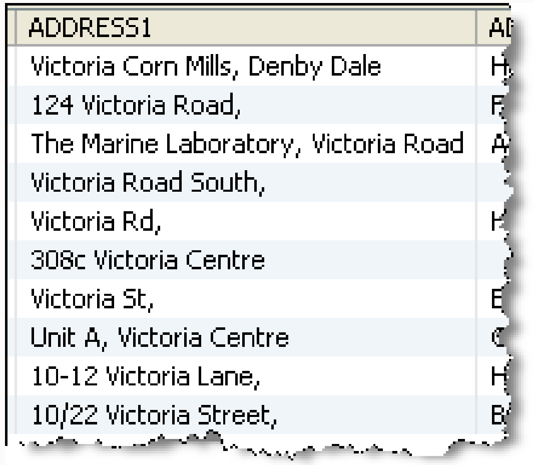
TITLE	NAME	BUSINESS	ADDRESS1	ADDRESS2	ADDRESS3	POSTCODE
Mr	Peter CROCKER		Mr Crocker, First Floor Flat	80 Grenville Road,	Plymouth	PL4 9PY

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
1	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

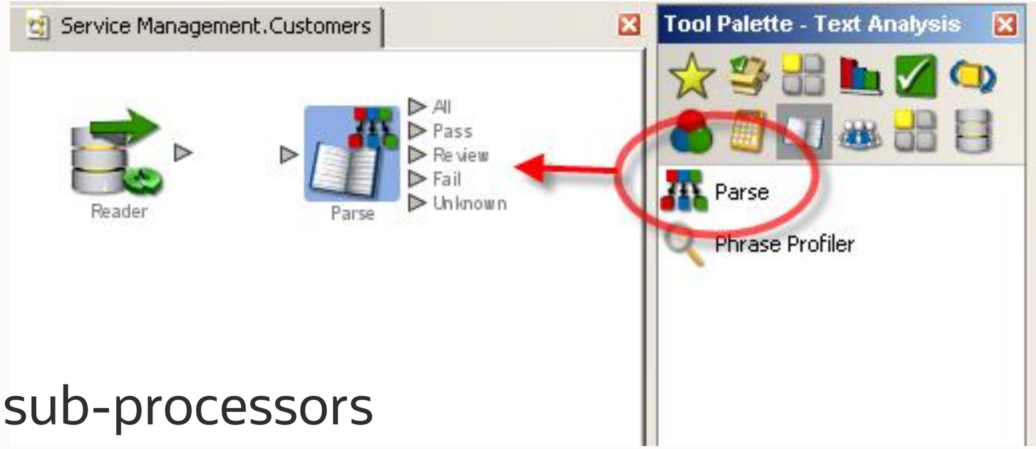
Parsing overview

- **Analyse and understand the meaning of data**
  - Lists of values – dictionaries or syntax
  - Patterns
  - Grammar of the ‘language’ used
  - Rules

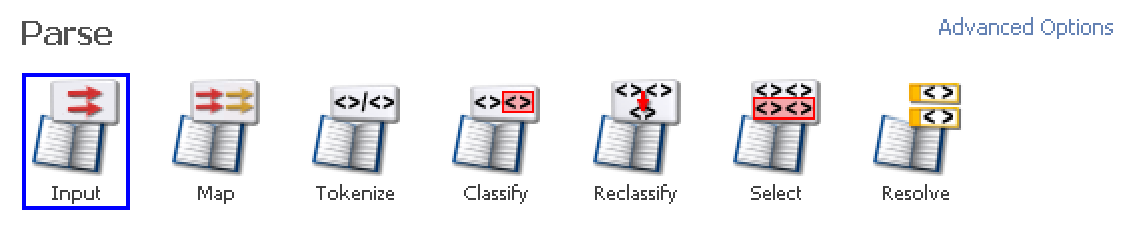
- **Validate and structurally improve data**
  - For example, identify a name in an address column and map it to a new column in a different structure

The EDQ Parse processor

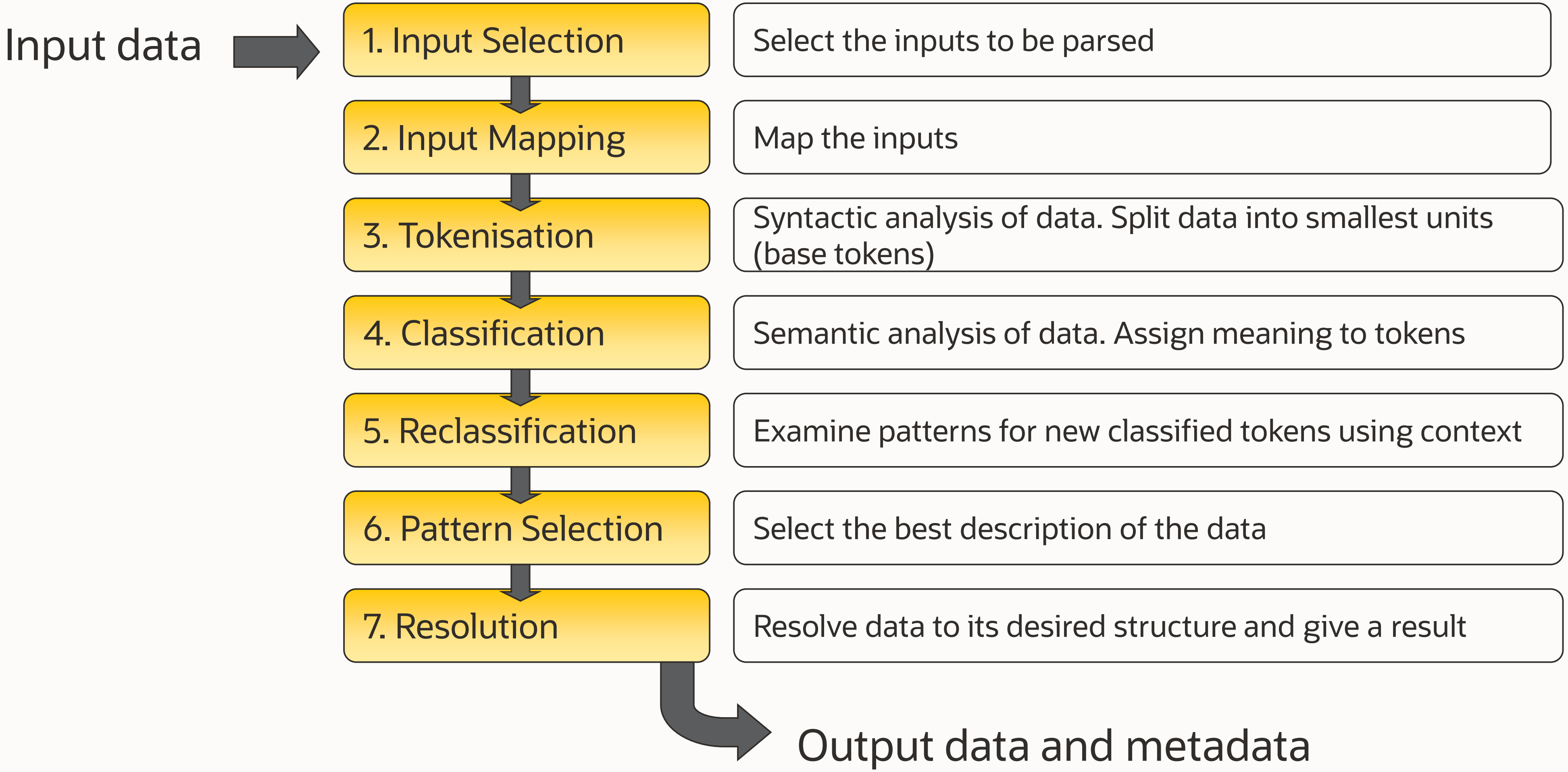
- Parse processor



- Seven sub-processors



The EDQ Parse processor

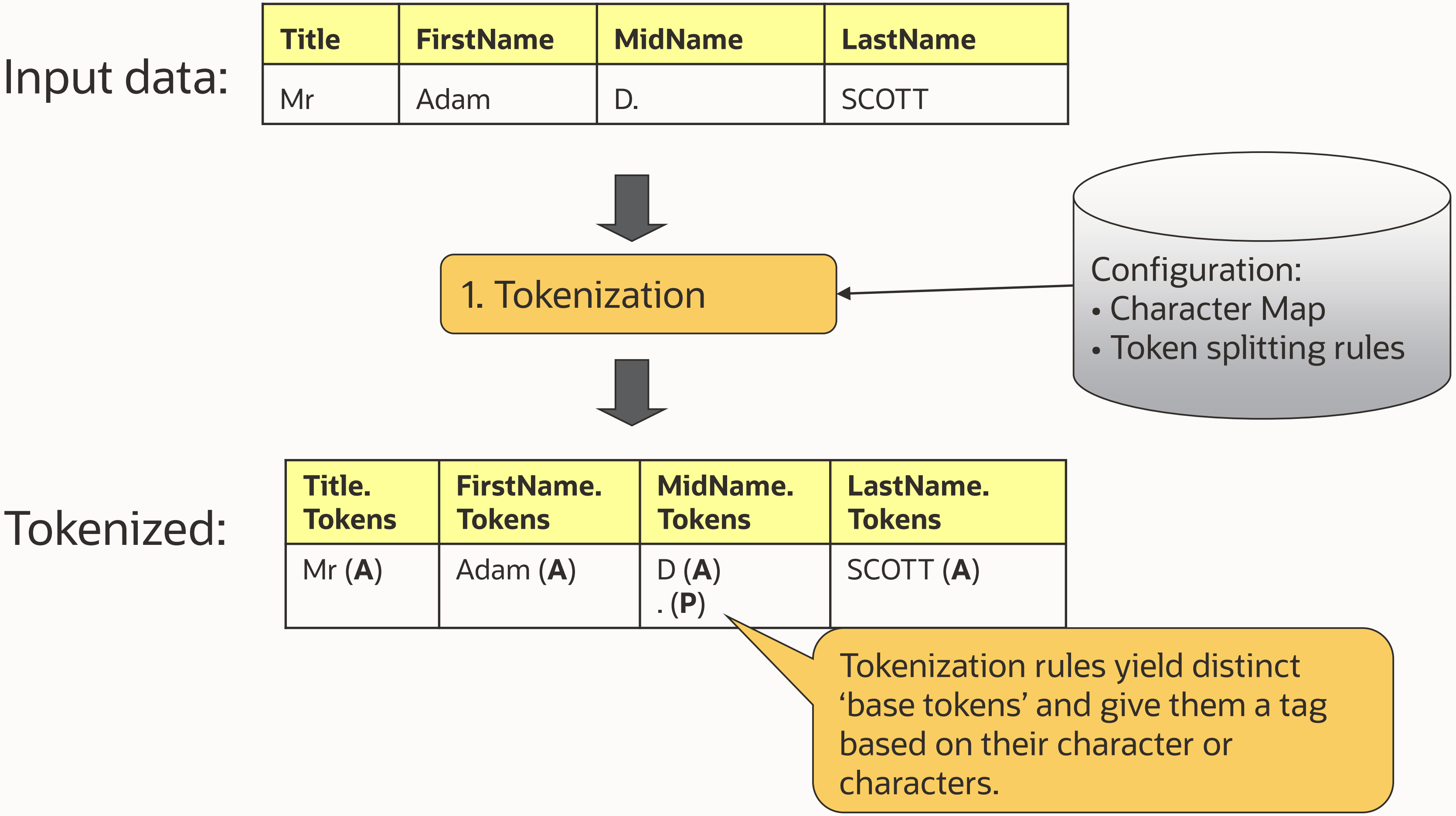


Input data

1. Input Selection	Select the inputs to be parsed
2. Input Mapping	Map the inputs
3. Tokenisation	Syntactic analysis of data. Split data into smallest units (base tokens)
4. Classification	Semantic analysis of data. Assign meaning to tokens
5. Reclassification	Examine patterns for new classified tokens using context
6. Pattern Selection	Select the best description of the data
7. Resolution	Resolve data to its desired structure and give a result

Output data and metadata

Tokenization



Input data:

Title	FirstName	MidName	LastName
Mr	Adam	D.	SCOTT

1. Tokenization

Configuration:
- Character Map
- Token splitting rules

Tokenized:

Title. Tokens	FirstName. Tokens	MidName. Tokens	LastName. Tokens
Mr (**A**)	Adam (**A**)	D (**A**) . (**P**)	SCOTT (**A**)

Tokenization rules yield distinct ‘base tokens’ and give them a tag based on their character or characters.

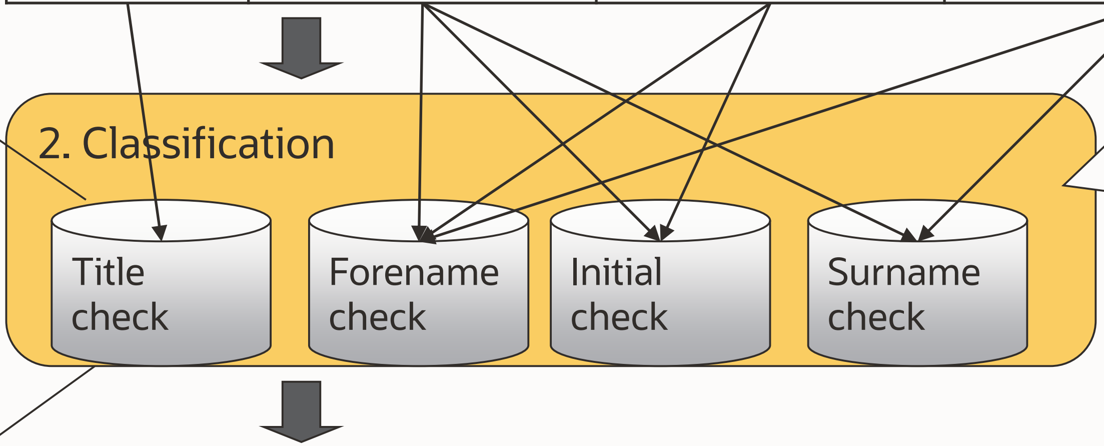
Classification

Input tokens:

Title. Tokens	**FirstName. Tokens**	**MidName. Tokens**	**LastName. Tokens**
Mr (**A**)	Adam (**A**)	D (**A**) . (**p**)	SCOTT (**A**)

Title token check (example)

Condition	**Result**
1. Matches list of valid Titles	Valid
2. Has base token A	Possible



Rules and their application to attributes completely configurable.

Classified tokens:

Title. Tokens	**FirstName. Tokens**	**MidName. Tokens**	**LastName. Tokens**
Mr (**Valid Title**)	Adam (**Valid Forename, Possible Surname**)	D (**Valid Initial**) . (**P**)	SCOTT (**Possible Surname, Valid Forename**)

Reclassification

Input tokens:

Address1. Tokens	Address2. Tokens	Address3. Tokens	Postcode. Tokens
James (A) House (Valid BuildHint)	10 (N) Jedburgh (A) Street (Valid Roadhint)	London (Valid Town)	SW11 5QB (Valid Postcode)

Reclassification is an optional way of creating new tokens from sequences of other tokens.

3. Reclassification

Rule	Match sequence	Reclassify as
1	N(1)A(1-2)Valid Roadhint(1)	Valid Thoroughfare

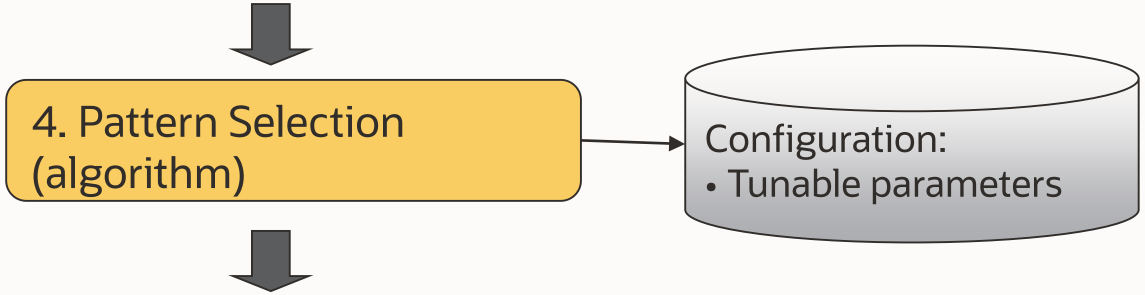
Reclassified tokens:

Address1. Tokens	Address2. Tokens	Address3. Tokens	Postcode. Tokens
James (A) House (Valid BuildHint)	10 (N) Jedburgh (A) Street (Valid Roadhint) **10 Jedburgh Street (Valid Thoroughfare)**	London (Valid Town)	SW11 5QB (Valid Postcode)

Pattern Selection

Possible token patterns:

Pattern	Title	FirstName	MidName	LastName
	Mr	Adam	D.	SCOTT
1	**<Title>**	<Surname>	**<Initial>**<P>	**<Forename>**
2	**<Title>**	**<Forename>**	**<Initial>**<P>	<Surname>
3	<A>	**<Forename>**	<A><P>	<Surname>
Etc.				



4. Pattern Selection (algorithm)

Configuration:
- Tunable parameters

Selected pattern:

Pattern	Title	FirstName	MidName	LastName
1	**<Title>**	<Surname>	**<Initial>**<P>	**<Forename>**
2	**<Title>**	**<Forename>**	**<Initial>**<P>	<Surname>
3	<A>	**<Forename>**	<A><P>	<Surname>

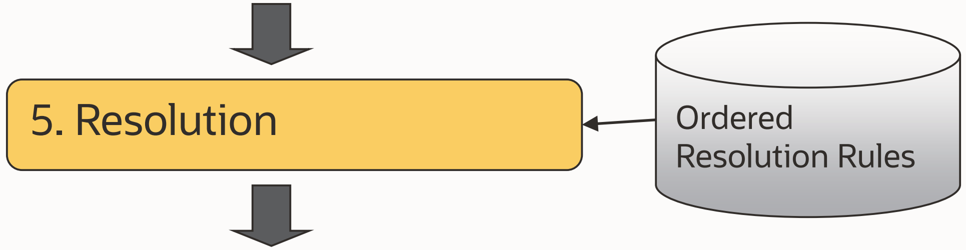
In this case, pattern 3 is ruled out because it has more unclassified tokens than patterns 1 and 2. Pattern 2 is selected because it occurs much more often across the data set.

Resolution

Based on the selected pattern for each record…

Title	FirstName	MidName	LastName
Mr	Adam	D.	Scott
<Title>	<Forename>	<Initial><P>	<Surname>
	SMITH	John	Richard
	<Surname>	<Forename.1>	<Forename.2>

…match a Resolution Rule (may be exact or inexact)…



…and assign a Result, optional Comment, and Output Format:

NewTitle	Forenames	Mid Initial	Surname	Result	Comment
Mr	Adam	D.	Scott	Pass	
	John Richard		SMITH	Review	Misfielded names

Resolution rules may be used simply for getting a result from the parsing process (for understanding) or may include transformation rules, where tokens are mapped to output attributes.

Our mission is to help people see data in new ways, discover insights, unlock endless possibilities.