Business School
UNIVERSITY OF COLORADO DENVER

Information Systems Program

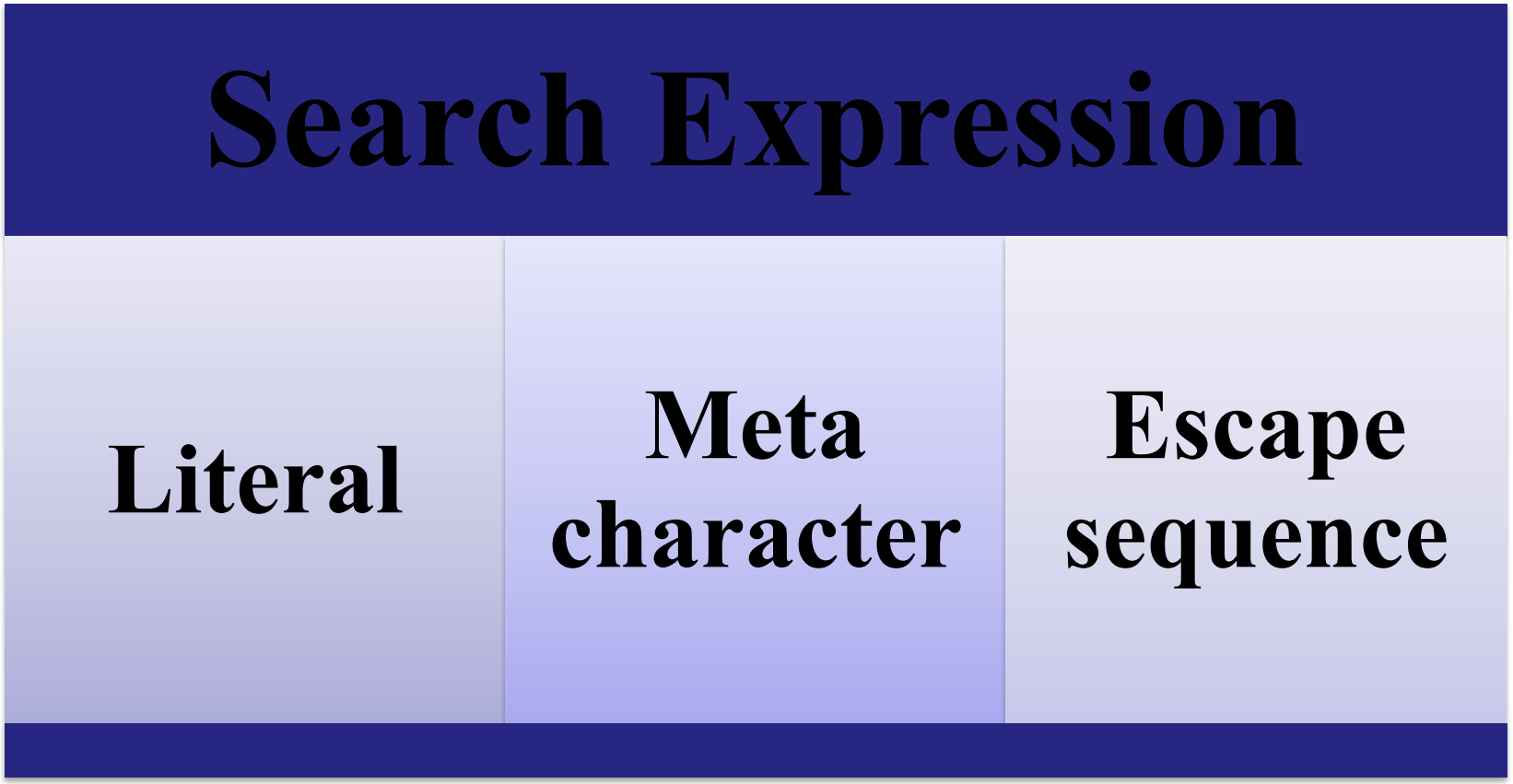
Module 4
Data Integration Concepts, Processes, and Techniques

Lesson 4: Pattern Matching with Regular Expressions

Lesson Objectives

- Explain the three major elements of regular expressions
- Practice with regular expressions
- Reflect on the complexity and limitations of regular expressions

Regular Expressions (regex)



Pattern Matching

Search expression	Target string	Match result
^[a-z]+\.com$	abc.com	abc.com

Meta characters
- ^
- [
-]
- +
- -
- \
- $

Literals
- c
- o
- m
- a
- z
- .

Escape sequence
- \.

Common meta characters

Iteration or quantifier				Position			Other		
?	*	+	{n}, {n,m}	.	^	$	[ ], [^]	\	|

Search expression

^[a-z]+\.com$

Meta Character Summary

Metacharacter	Type	Meaning
?	Iteration	Matches preceding character 0 or 1 time
*	Iteration	Matches preceding character 0 or more times
+	Iteration	Matches preceding character 1 or more times
{n}	Iteration	Matches preceding character exactly n times
{n,m}	Iteration	Matches preceding character at least n times and at most m times
[]	Range	Matches one of enclosed characters one time
^	Position	Matches at the beginning of the target string; only has meaning as the first character in a regular expression
^	Range	Negation of search pattern if ^ is inside []. Hyphen inside square brackets defines a range of characters.
$	Position	Matches at the end of the target string; only has meaning as the last character in a regular expression.
.	Position	Matches any character except a newline character at the specified position only
|	Alteration	Matches either pattern to the left or right of the | character.

Meta Character Examples I

Search Expression	Target Strings	Evaluation
“colou?r”	“color”, “colour”	Matches both target strings
“tre*”	“tree”, “tread”, “trough”	Matches all three target strings; Matches preceding character 0 times in third target string
“tre+”	“tree”, “tread”, “trough”	Does not match the third target string
“[abcd]”	“dog”, “fond” , “pen”	Matches first two strings but not the third string
“[0-9]{3}-[0-9]{4}”	“123-4567”, “1234-567”	Matches first string but not the second string
“ba{2,3}b”	“baab”, “baaab”, “bab”, “baaaab”	Matches first two strings but not the last two strings

Meta Characters II

Search Expression	Target Strings	Evaluation
“^win”	“erwin”, “window”	Second string but not first string
“win$”	“erwin”, “window”	First string but not second string
“[^0-9]+”	“123”, “abc”, “a456”	Matches the second and third target strings
“abc.e*”	“fabc”, “fabcd”, “fabcee”	Matches the second and third target strings
“dog|cat|frog”	“a dog”, “cat friend”, “frogman”	Matches all three target strings

More Complex Examples

Field	Search Expression
User name	^[a-z0-9_-]{3,16}$
Hex value	^#?([a-f0-9]{6}|[a-f0-9]{3})$
Email address	^([a-z0-9_\.-]+)@([\da-z\.-]+)\.([a-z\.]{2,6})$
Web address	^(https?:\/\/)?([\da-z\.-]+)\.([a-z\.]{2,6})([\/\w \.-]*)*\/?$

Summary

- Powerful pattern matching for text fields with multiple components
- Wide availability of regular expression parsing
- Literals, meta characters, and escape sequences