



PhUSE 2011: Brighton

TS09

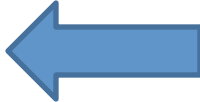
Rectifying Irregular Text Data
a Case for Using Regular Expressions in SAS

Jayshree Garade
Manjusha Gode

Outline

- Problems
- Solutions & Introducing Regular Expressions
- Advantages over SAS String Functions
- Points to note while using Regular Expressions
- References

Outline

- Problems 
- Solutions & Introducing Regular Expressions
- Advantages over SAS String Functions
- Points to note while using Regular Expressions
- References

Problem: Physical abnormalities

SUBJID	TRT	ABNORMALITY
01-011	B	ANEMIA
01-036	D	ANAEMIA
01-026	C	ANEMEA
01-014	B	ANEMIC

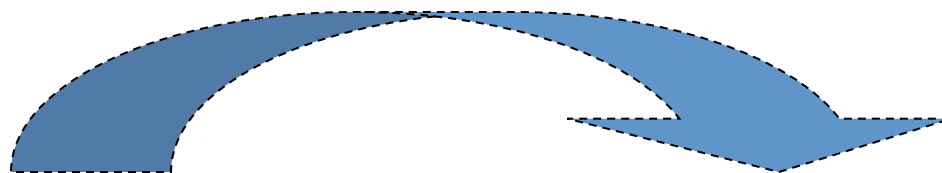
Problem: Time point variable ...

USUBJID	VISIT	VSDT	PRSDTLTM	VNTR_RT	VNTRTUN
1	1	17-Oct-08	Per 1 D01 Predose	47	/min
1	2	3-Nov-08	Per 1 D01 .5 hr	58	/min
1	2	3-Nov-08	Per 1 D 01 01 hr	51	/min
1	2	3-Nov-08	Per 1d01 02hr	49	/min
1	3	4-Nov-08	day1	53	/min
1	90	3-Feb-09	Poststudy	56	/min

...Problem: Time point variable

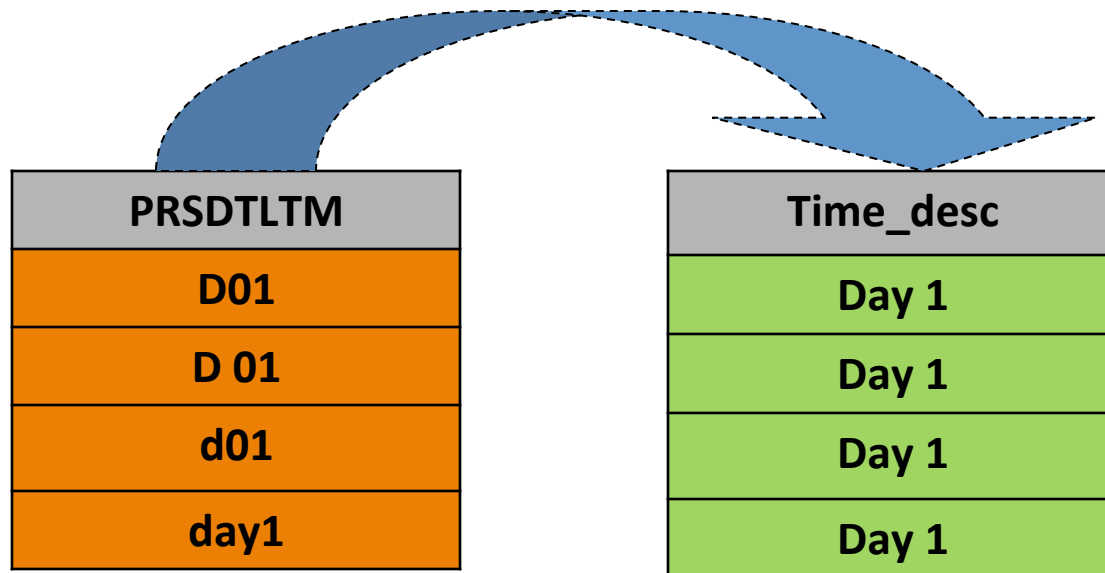
USUBJID	VISIT	VSDT	PRSDTLTM	VNTR_RT	VNTRTUN
1	1	17-Oct-08	Per 1 D01 Predose	47	/min
1	2	3-Nov-08	Per 1 D01 .5 hr	58	/min
1	2	3-Nov-08	Per 1 D 01 01 hr	51	/min
1	2	3-Nov-08	Per 1d01 02hr	49	/min
1	3	4-Nov-08	day1	53	/min
1	90	3-Feb-09	Poststudy	56	/min

...Problem: Time point variable

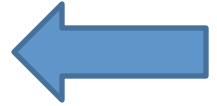


USUBJID	VISIT	VSDT	PRSDTLTM	VNTR_RT	VNTRTUN	Time_desc
1	1	17-Oct-08	Per 1 D01 Predose	47	/min	Predose
1	2	3-Nov-08	Per 1 D01 .5 hr	58	/min	Day 1, 0.5 Hour
1	2	3-Nov-08	Per 1 D 01 01 hr	51	/min	Day 1, 1 Hour
1	2	3-Nov-08	Per 1d01 02hr	49	/min	Day 1, 2 Hours
1	3	4-Nov-08	day1	53	/min	Day 1
1	90	3-Feb-09	Poststudy	56	/min	Poststudy

...Problem: Time point variable



Outline

- Problems
- Solutions & Introducing Regular Expressions 
- Advantages over SAS String Functions
- Points to note while using Regular Expressions
- References

...Ways to approach the problem

- Traditional --- Using SAS String Functions

INDEX TRANWRD SUBSTR ANYALNUM ANYALPHA
 ANYDIGIT ANYSPACE NOTALNUM NOTALPHA ANYALNUM
 NOTUPPER ANYALPHA FIND ANYDIGIT FINDC ANYPUNCT
 ANYSPACE INDEXC NOTALNUM INDEXW NOTALPHA VERIFY
 NOTDIGIT CALL CATS CALL CATT CALL CATX TRANSLATE SCAN
 SCANQ CALL SCAN CALL SCANQ COMPARE COMPLEV CALL
 COMPCOST SOUNDEX COMPGED SPEDIS MISSING RANK
 REPEAT REVERSE.....

Alternative Approach to Problem

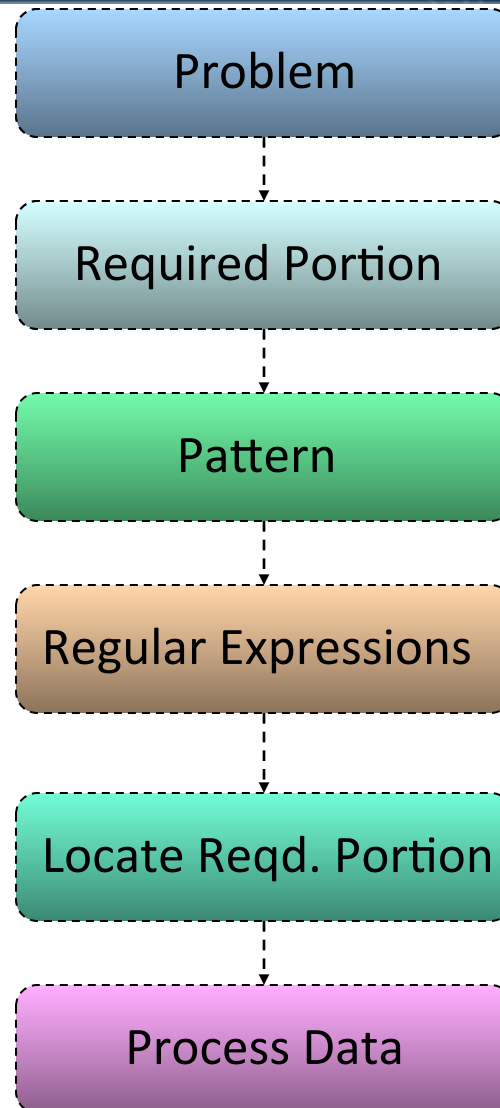


Introducing REGULAR EXPRESSIONS!!

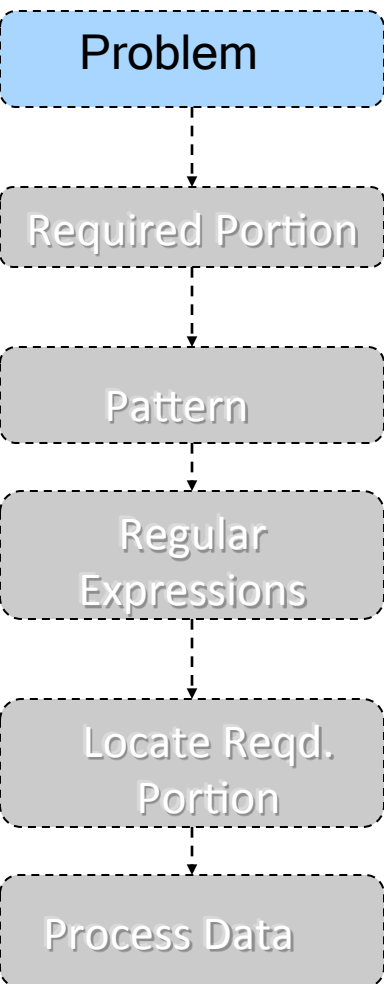
Introduction – Regular Expressions

- Powerful technique for searching and manipulating text data.
- A mini programming language - pattern matching.
- 2 types – pattern matching functions in SAS
 - SAS Regular Expressions – SAS Version 6.12
 - PERL Regular Expressions – SAS Version 9

Steps to use Regular Expressions...

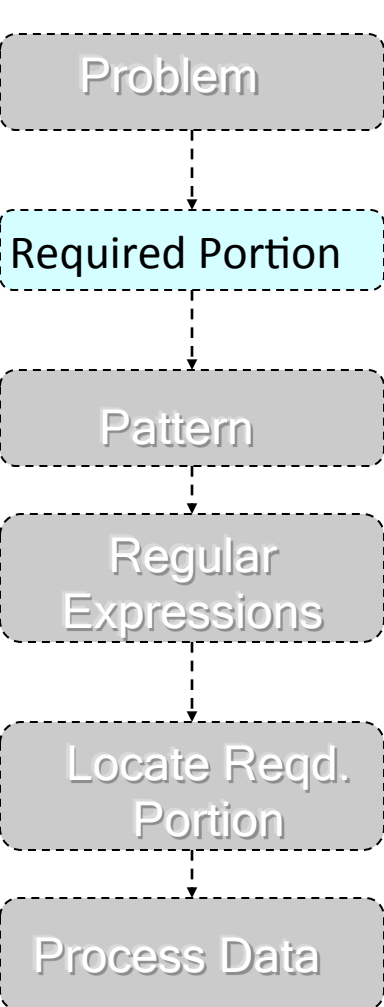


Step1 - Identify the problem ...



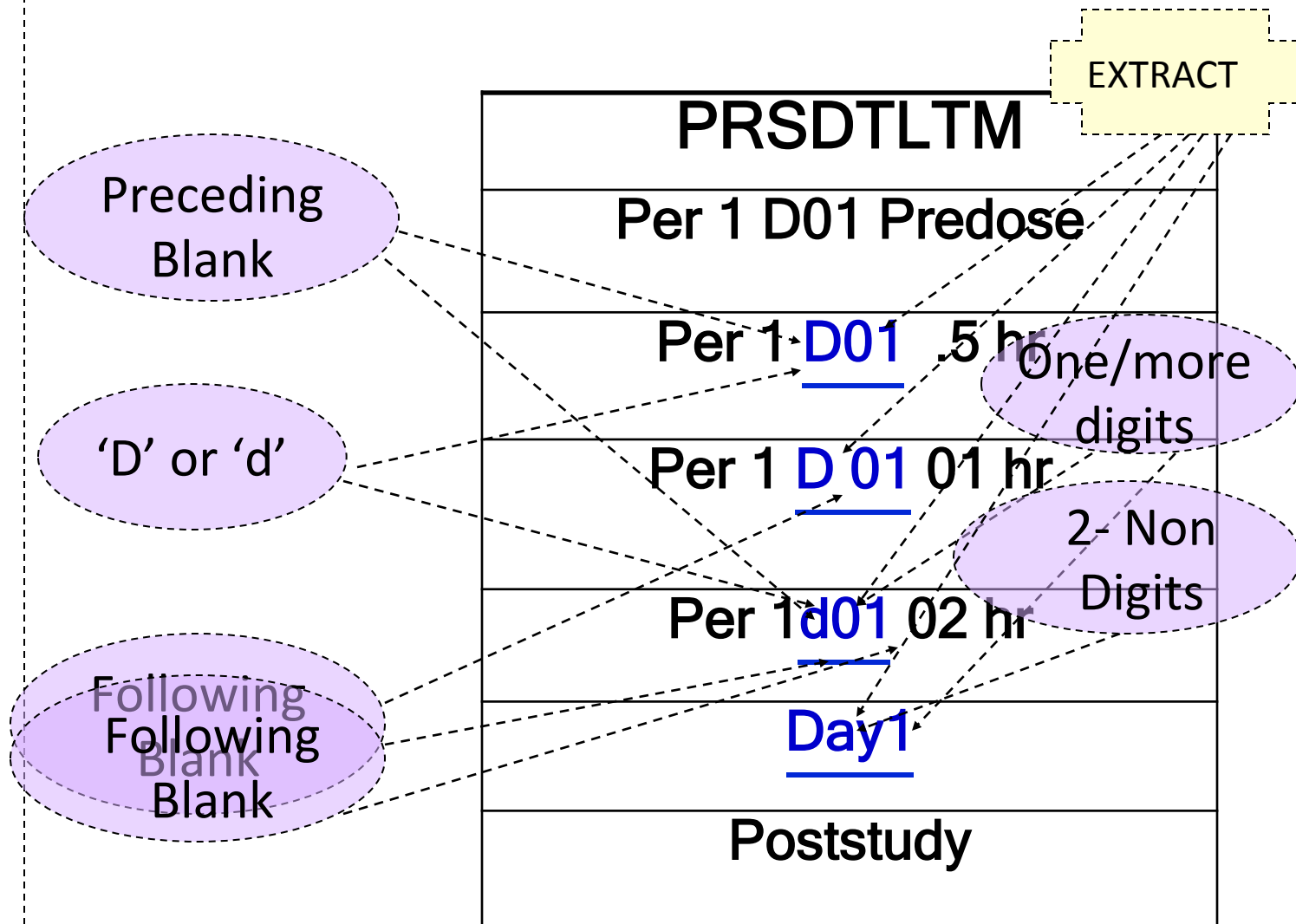
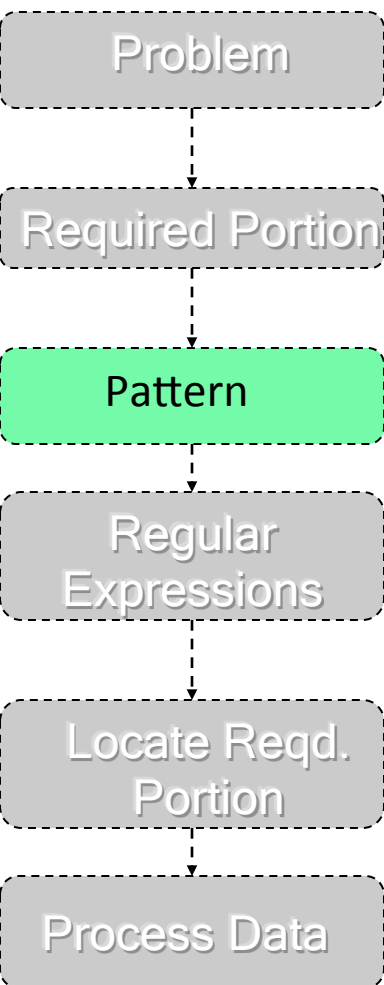
USUB JID	VISIT	VSDT	PRSDTLTM	VNTR_ RT	VNTRT UN	time_desc
1	1	17- Oct-08	Per 1 D01 Predose	47	/min	Predose
1	2	3- Nov-08	Per 1 D01 .5 hr	58	/min	Day 1, 0.5 Hour
1	2	3- Nov-08	Per 1 D 01 01 hr	51	/min	Day 1, 1 Hour
1	2	3- Nov-08	Per 1d01 02 hr	49	/min	Day 1, 2 Hours
1	3	4- Nov-08	Day1	53	/min	Day 1
1	90	3- Feb-09	Poststudy	56	/min	Poststudy

Step2 - Visualize the “Required Portion” within the source text

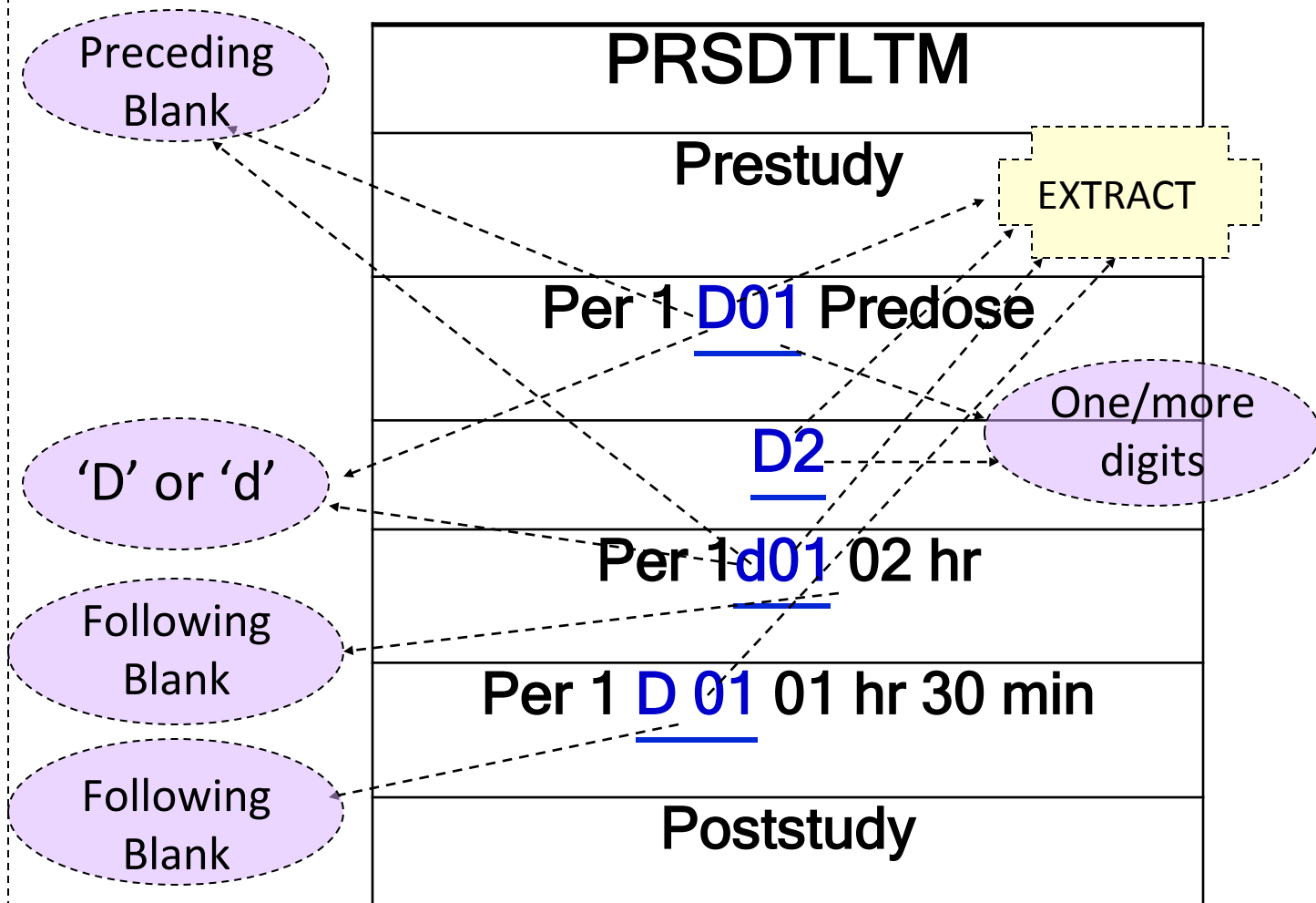
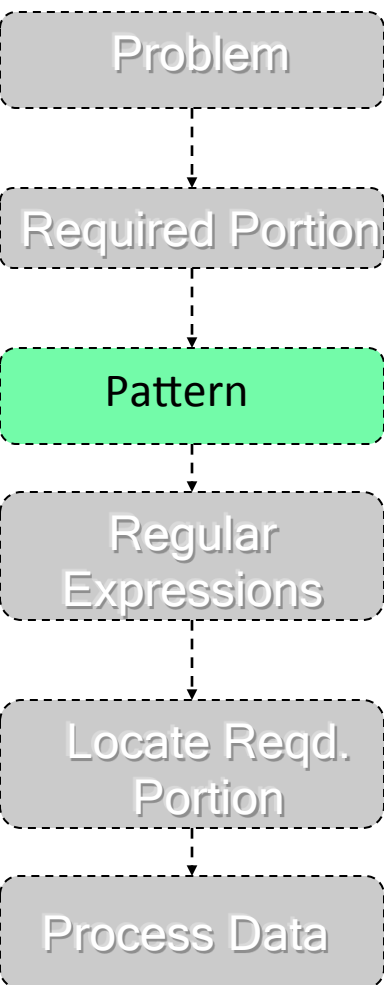


PRSDTLTM
Per 1 D01 Predose
Per 1 <u>D01</u> .5 hr
Per 1 <u>D 01</u> 01 hr
Per 1 <u>d01</u> 02 hr
<u>Day1</u>
Poststudy

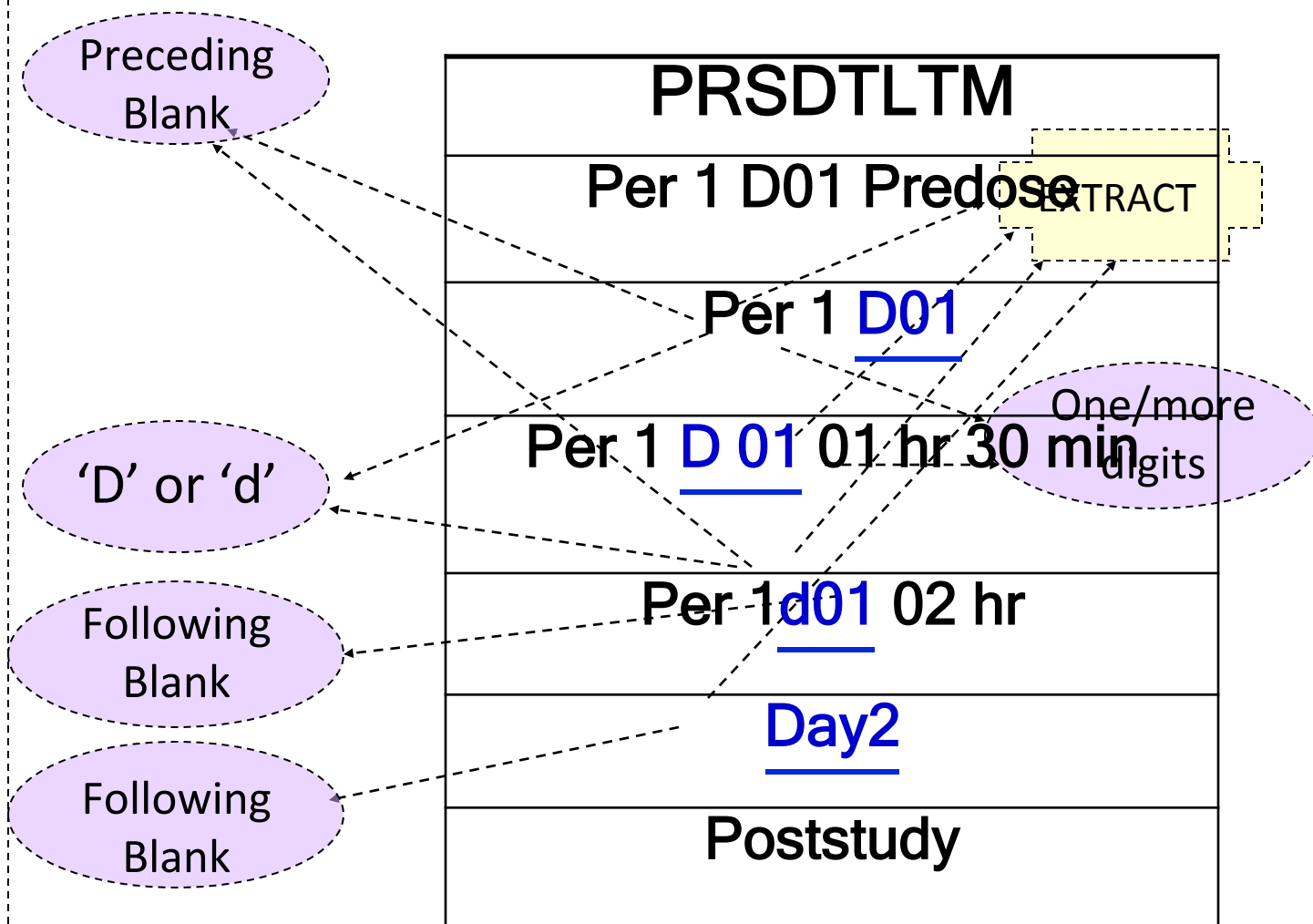
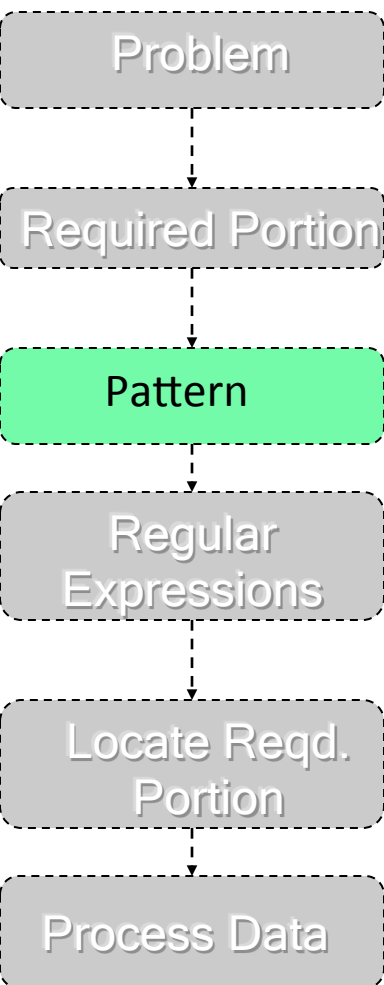
Step 3 – Identify a pattern



Step 3 – Identify a pattern



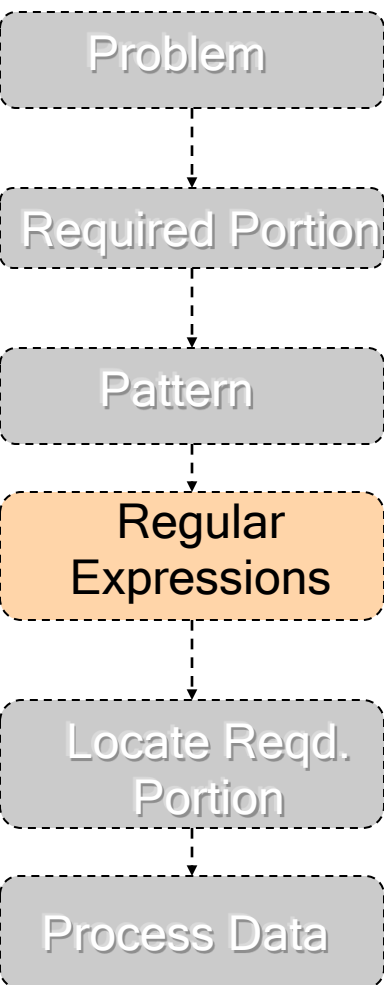
Step 3 – Identify a pattern



Regular Expressions Syntax...at a glance

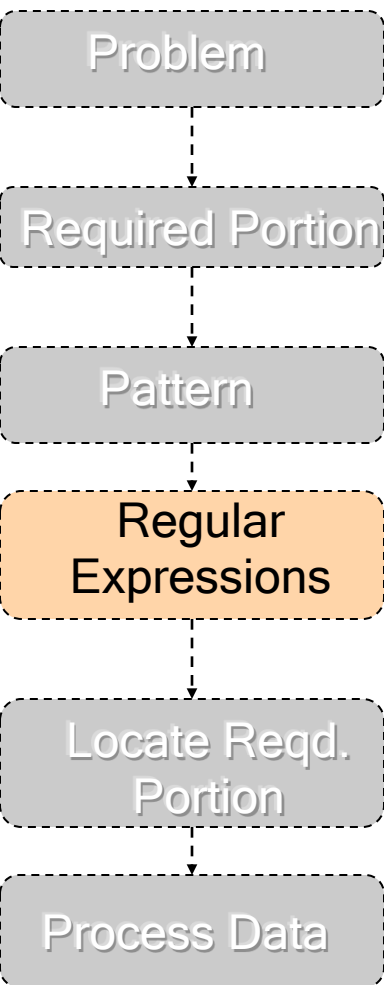
Metacharacter	Description
*	Matches the previous sub expression zero or more times
+	Matches the previous sub expression one or more times
?	Matches the previous sub expression zero or one times
\d	Matches a digit (0-9)
\D	Matches a non-digit
\w	Matches a word character (upper or lower case letter, blank, or underscore)
[abc]	Matches any of the characters in the brackets
\(	Matches (

Step 4 - Write the Regular Expression for the pattern



Preceding Blank	("/ ?	/")
	PRSDTLTM	
	Per 1 D01 Predose	
	Per 1 <u>D01</u> .5 hr	
	Per 1 <u>D 01</u> 01 hr	
	Per 1 <u>d01</u> 02 hr	
	<u>Day1</u>	
	Poststudy	

Step 4 - Write the Regular Expression for the pattern

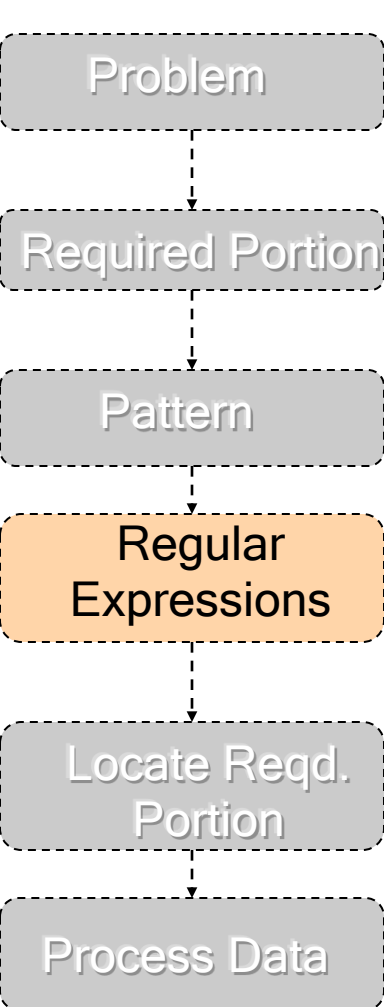


("/ ? [Dd] /")

PRSDTLTM
Per 1 D01 Predose
Per 1 <u>D01</u> .5 hr
Per 1 <u>D 01</u> 01 hr
Per 1 <u>d01</u> 02 hr
<u>Day1</u>
Poststudy

'D' or 'd'

Step 4 - Write the Regular Expression for the pattern

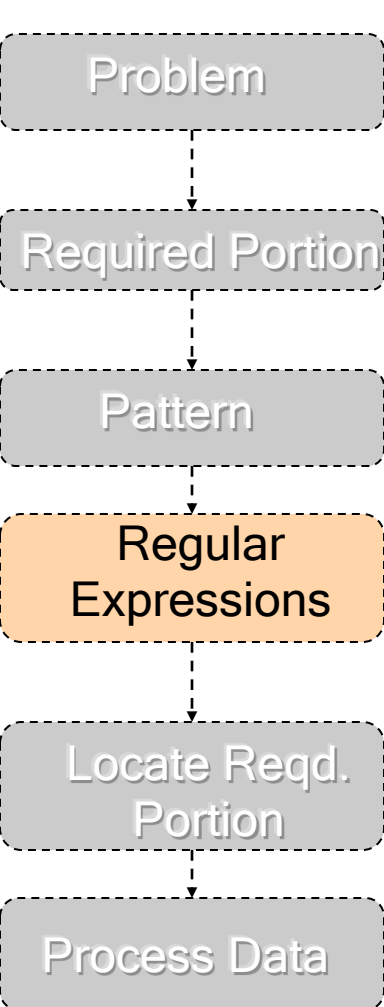


(" / ? [Dd] (\D\D)? /")

2-Non
Digits

PRSDTLTM
Per 1 D01 Predose
Per 1 <u>D01</u> .5 hr
Per 1 <u>D 01</u> 01 hr
Per 1 <u>d01</u> 02 hr
<u>Day1</u>
Poststudy

Step 4 - Write the Regular Expression for the pattern

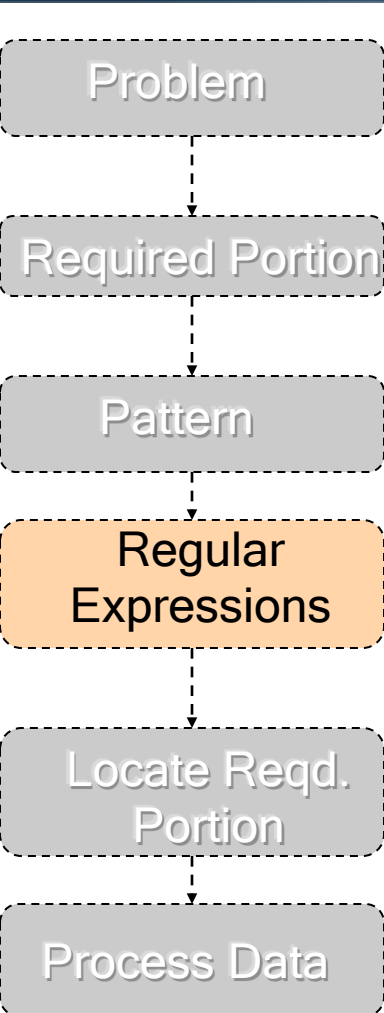


Following
Blank

(" / ? [Dd] (\D\D)? ? /")

PRSDTLTM
Per 1 D01 Predose
Per 1 <u>D01</u> .5 hr
Per 1 <u>D 01</u> 01 hr
Per 1 <u>d01</u> 02 hr
<u>Day1</u>
Poststudy

Step 4 - Write the Regular Expression for the pattern

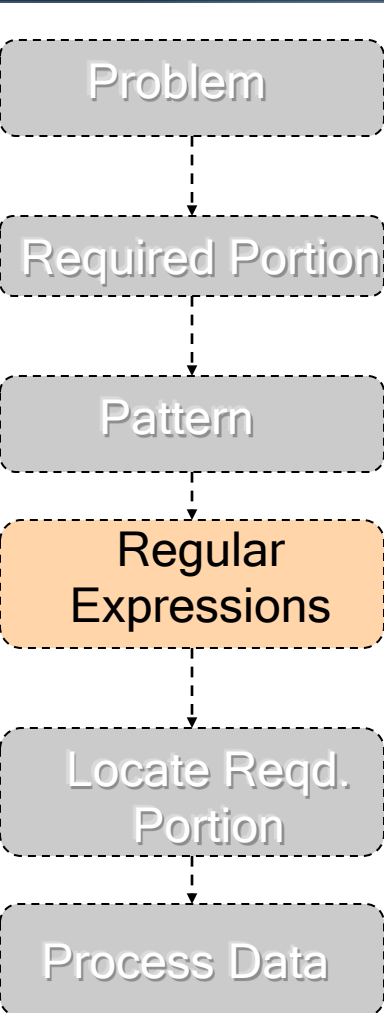


(" / ? [Dd] (\D\D)? ? \d+ /")

One/more digits

PRSDTLTM
Per 1 D01 Predose
Per 1 <u>D01</u> .5 hr
Per 1 <u>D 01</u> 01 hr
Per 1 <u>d01</u> 02 hr
<u>Day1</u>
Poststudy

Step 4 - Write the Regular Expression for the pattern

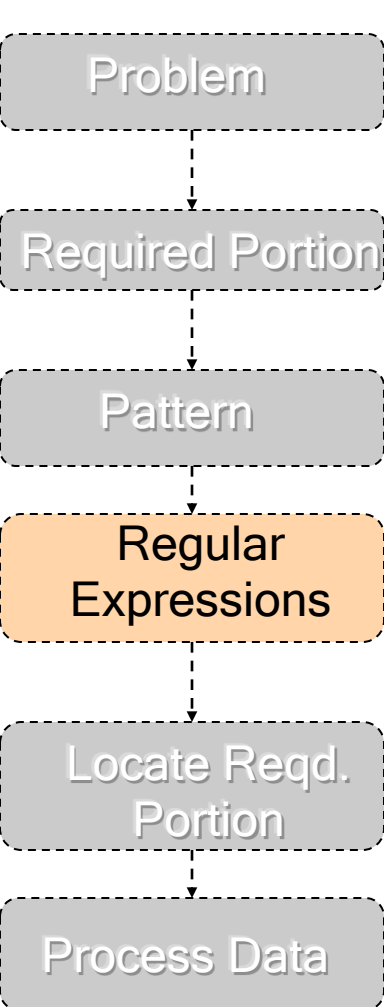


(" / ? [Dd] (\D\D)? ? \d+ + /")

Following blank

PRSDTLTM
Per 1 D01 Predose
Per 1 <u>D01</u> .5 hr
Per 1 <u>D 01</u> 01 hr
Per 1 <u>d01</u> 02 hr
<u>Day1</u>
Poststudy

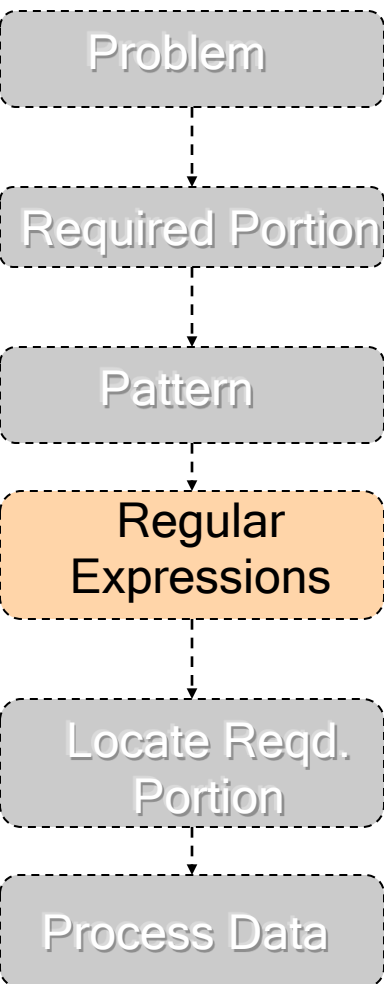
Step 4 - Write the Regular Expression for the pattern



("/ ?[Dd] (\D\D)? ?\d+ +/')

PRSDTLTM
Per 1 D01 Predose
Per 1 <u>D01</u> .5 hr
Per 1 <u>D 01</u> 01 hr
Per 1 <u>d01</u> 02 hr
<u>Day1</u>
Poststudy

Step 4 - Write the Regular Expression for the pattern



```
/* Extracting the Day Text portion*/
data day_txt;
set lb.ecg(keep = PRSDTLTM) ;
retain day_exp;

if _n_ = 1 then
do ;

* defined to describe the day text
pattern;

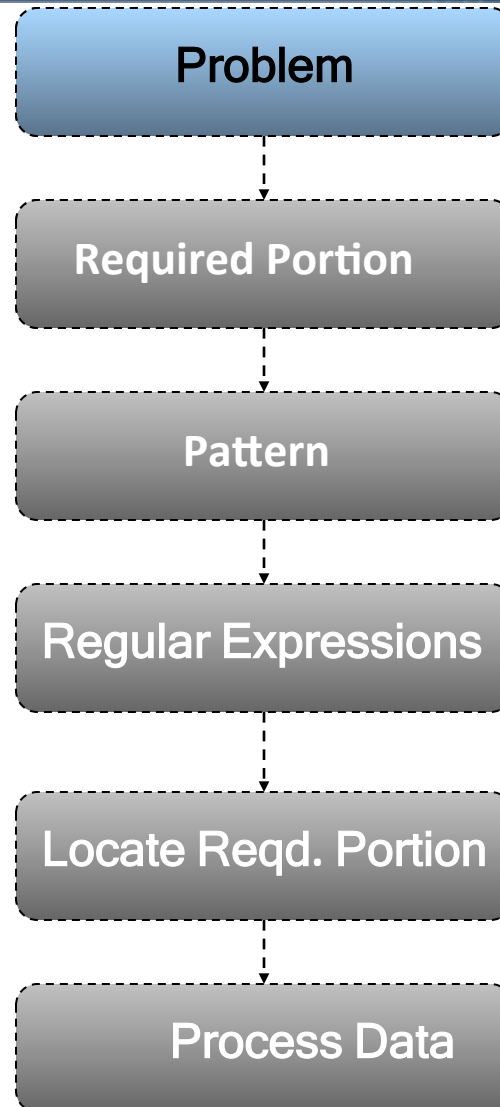
day_exp
=PRXPARSE ("/ ?[Dd] (\D\D)? ?\d+ +/");

end;

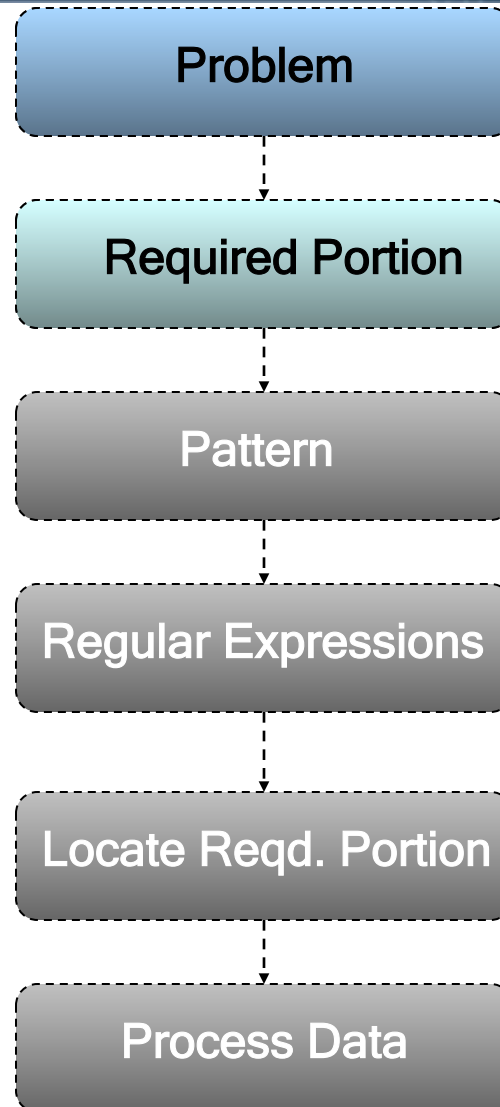
run;
```

Metacharacters

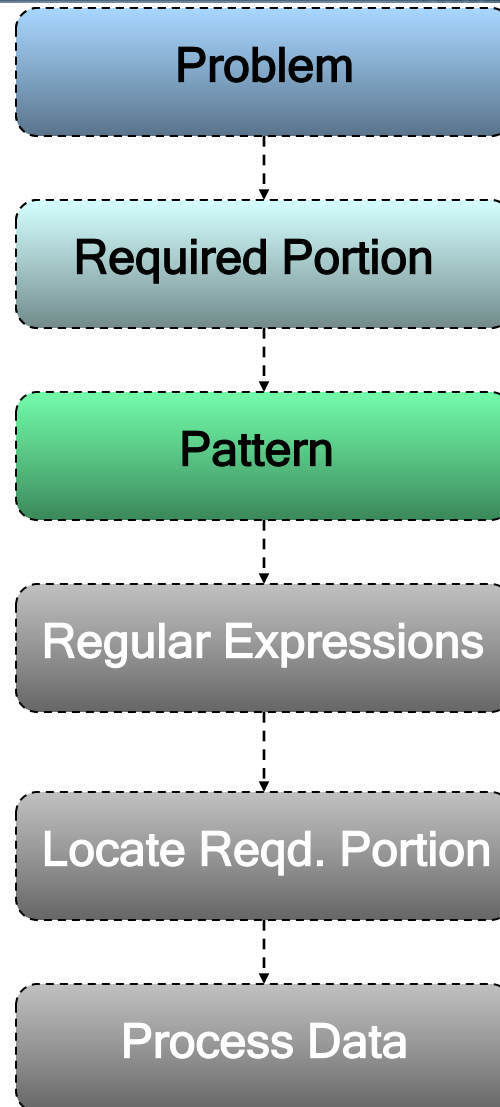
Recap... Steps to use Regular Expressions...



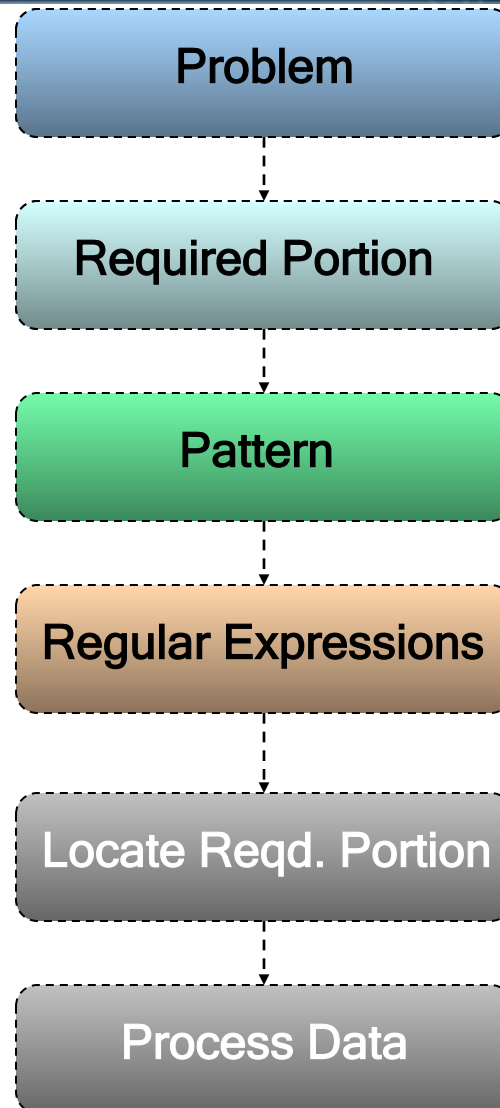
Recap... Steps to use Regular Expressions...



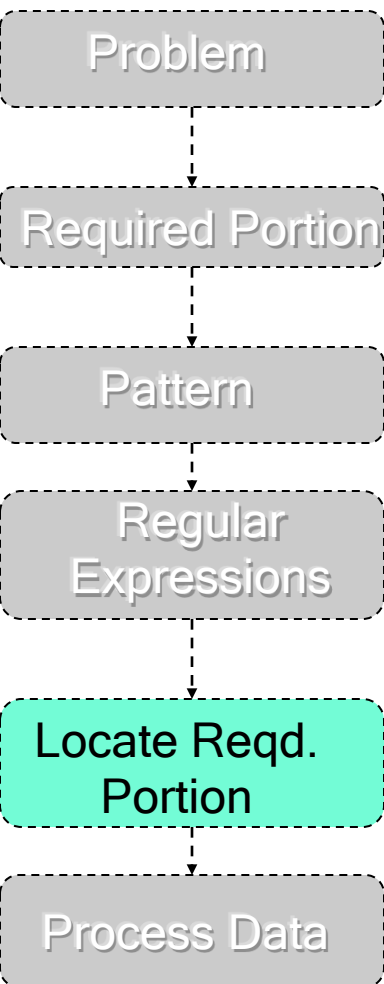
Recap... Steps to use Regular Expressions...



Recap... Steps to use Regular Expressions...



Step 5 - Locate the "Required Portion"



/* Extracting the Day Text portion*/

data day_txt;

set lb.ecg(keep = PRSDTLTM);

retain day_exp day_nexp;

if _n_ = 1 then

do ;

* defined to describe the day text pattern;

day_exp = PRXPARSE("/ ?[Dd](\D\D)? ?\d+ +/");

end;

***Locating the day text pattern in the PRSDTLTM**

var;

CALL

PRXSUBSTR(day_exp, PRSDTLTM, dayst, dayln) ;

run;

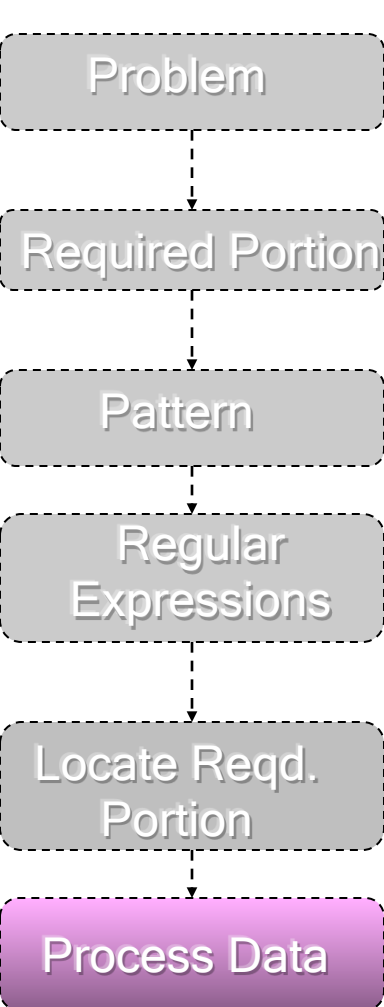
Pattern defn

Source Variable

Stores Start position of matched string

Stores length of matched string

Step 6 - Use other SAS text functions to further process data



/* Extracting the Day Text portion*/

data day_txt;

```
set lb.ecg(keep = PRSDTLTM);  
retain day_exp day_nexp;
```

```
if _n_ = 1 then  
do ;
```

```
    * defined to describe the day text pattern;  
    day_exp = PRXPARSE("/ ?[Dd](\D\D)? ?\d+ +/");  
end;
```

```
    * Locating the day text pattern in the PRSDTLTM var;  
CALL PRXSUBSTR(day_exp, PRSDTLTM, dayst, dayln);
```

*** Extracting the day text pattern;**

```
day txt = substrn  
(PRSDTLTM, dayst, dayln) ;
```

run;

Source
Variable

Starting
Position

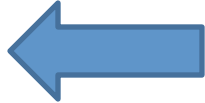
Length of
matched pattern

...Output

PRSDTLTM	day_txt
Per 1 D01 Predose	
Per 1 D01 .5 hr	D01
Per 1 D 01 01 hr	D 01
Per 1d01 02 hr	d01
Day1	Day1
Poststudy	

Extracted string

Outline

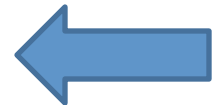
- Problems
- Solutions & Introducing Regular Expressions
- Advantages over SAS String Functions 
- Points to note while using Regular Expressions
- References

Advantages...

- Compact solution
- Tremendous flexibility
 - Concise description.
 - Highly unstructured data streams.
 - Multiple matching patterns in one step.

Outline

- Problems
- Solutions & Introducing Regular Expressions
- Advantages over SAS String Functions
- Points to note while using Regular Expressions
- References



Look before you leap

- Document thoroughly.
- Understand patterns.
- Define before use.
- Define only once in a data step.

Outline

- Problems
- Solutions & Introducing Regular Expressions
- Advantages over SAS String Functions
- Points to note while using Regular Expressions
- References 

...References

- ❖ [Support.sas.com](http://support.sas.com)
- ❖ Paper TU02-
An Introduction to Regular Expressions with Examples from Clinical Data - Richard F. Pless, Ovation Research Group, Highland Park, IL
- ❖ SUGI 29-Tutorials - Paper 265-29
An Introduction to Perl Regular Expressions in SAS 9
Ron Cody, Robert Wood Johnson Medical School, Piscataway, NJ
- ❖ An Introduction to PERL Regular Expression in SAS®
James J. Van Campen, SRI International, Menlo Park, CA

Q & A

Contact :

jayshree.garade@cytel.com

manjusha.gode@cytel.com



Thank you