

#SASGF

# VIRTUAL

SAS® GLOBAL FORUM 2021

# Transcoding: Understand, Troubleshoot, and Resolve a Most Mysterious SAS® Error

Yun (Julie) Zhuo, PRA Health Sciences

Julie is a SAS user of 17 years, holding various roles such as clinical programmer and statistical programmer in the health care and pharmaceutical industries. She actively presents and publishes in recent years. She is a recipient of the Best Contributing Paper Awards from SAS user group conferences such as PharmaSUG and WUSS. She was an invited speaker of the SAS Global Forum 2020.

#SASGF

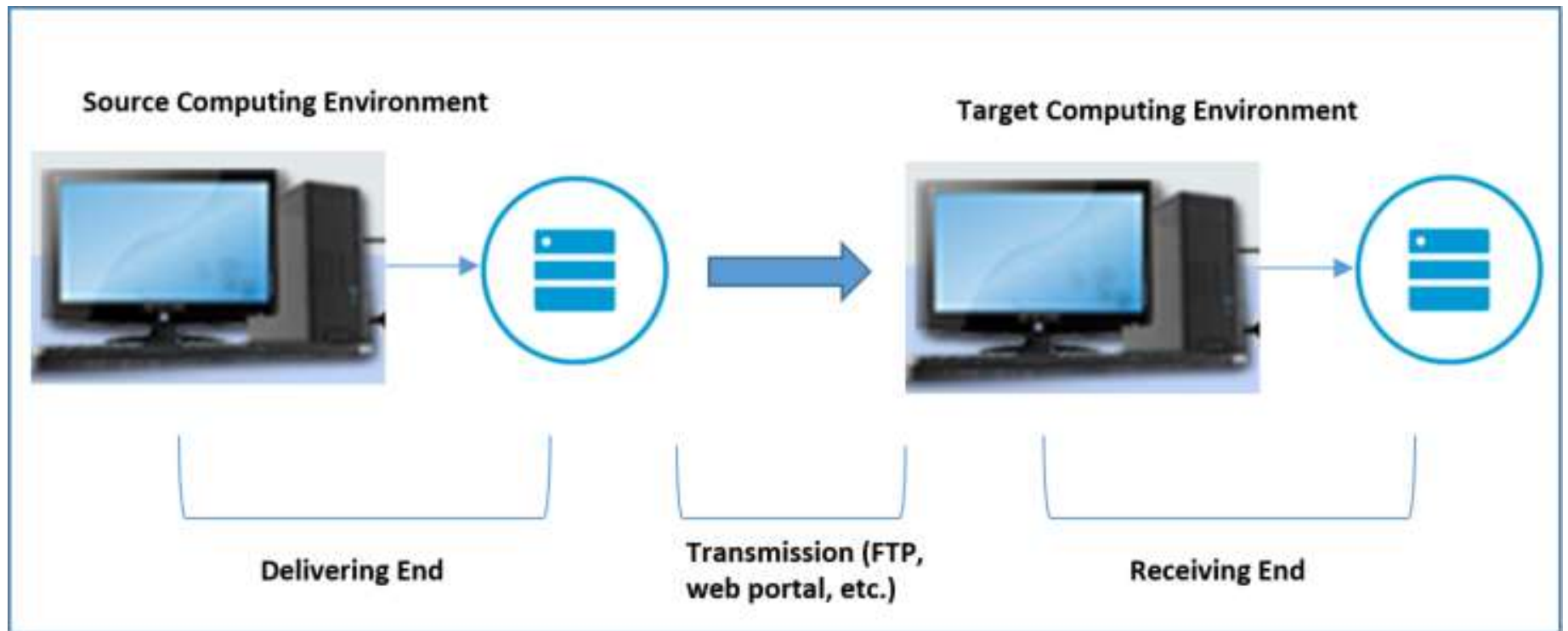
SAS® GLOBAL FORUM 2021

# Outline

- Understand the transcoding error
  - What is encoding? What is transcoding? How does the error occur?
- How to troubleshoot the transcoding error
  - SAS techniques
- How to resolve the transcoding error
  - The use of CVP engine to avoid data truncation
  - Locate and update your SAS session encoding
  - Convert characters

# Understand the Transcoding Error

## Understand Encoding and Transcoding



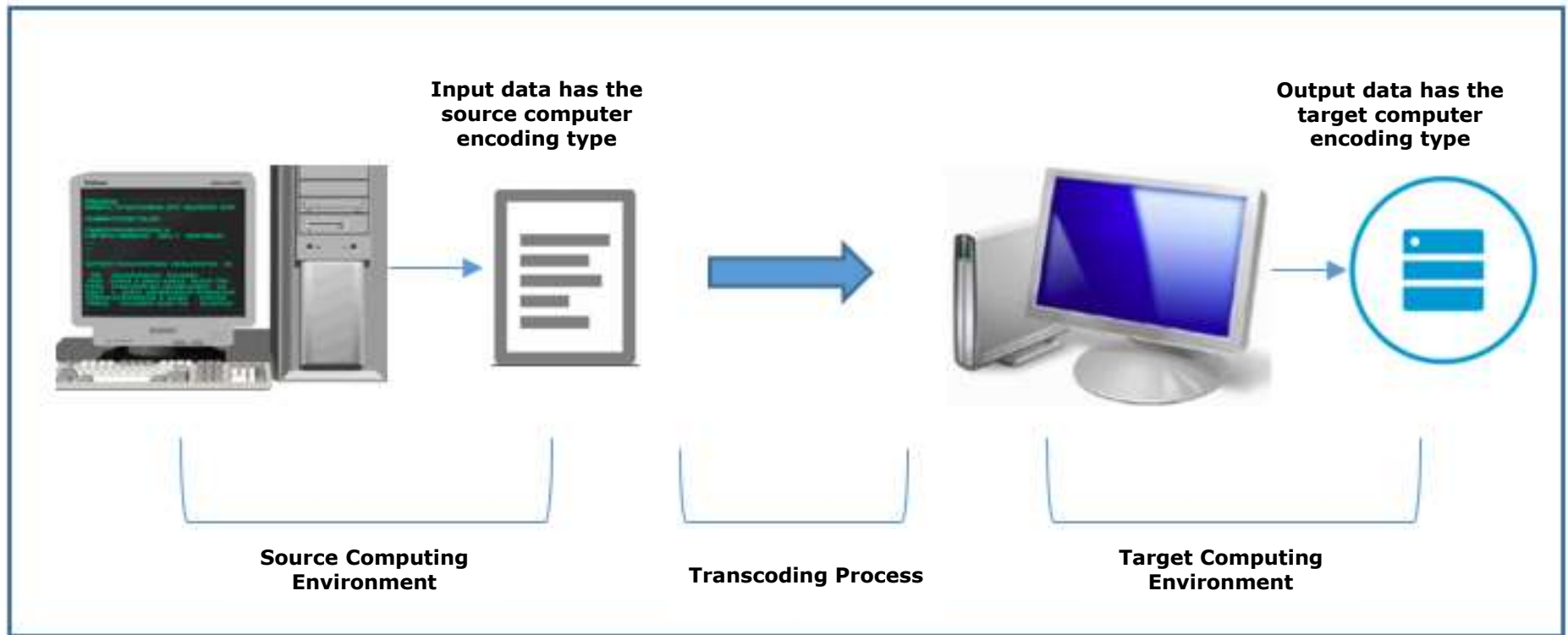
#SASGF

SAS® GLOBAL FORUM 2021

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® Indicates USA registration. Other brand and product names are trademarks of their respective companies.

# Understand the Transcoding Error

## Understand Encoding and Transcoding



#SASGF

SAS® GLOBAL FORUM 2021

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® Indicates USA registration. Other brand and product names are trademarks of their respective companies.

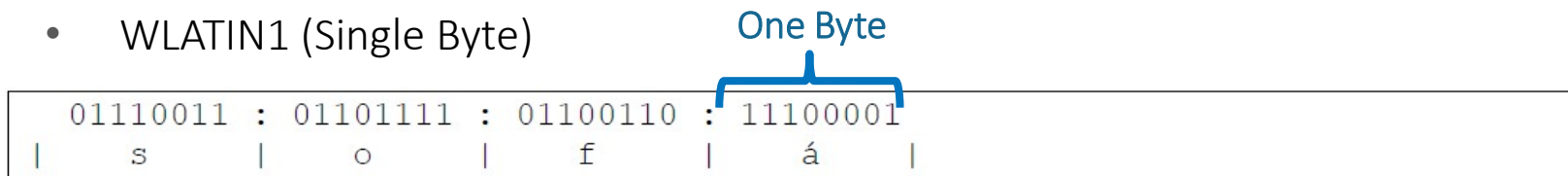
# Understand the Transcoding Error

## Understand Encoding and Transcoding

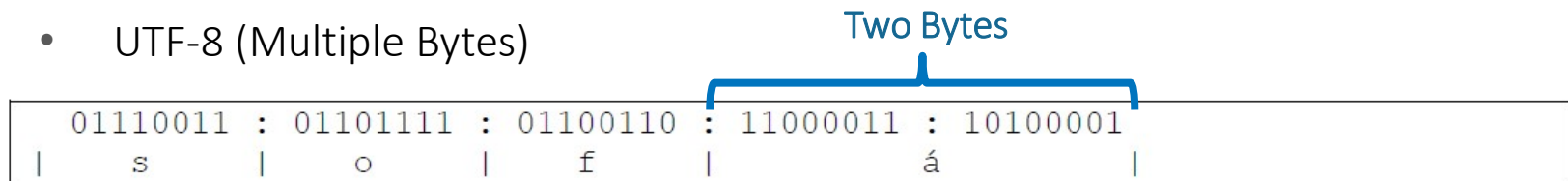
- Encoding

- The way computers represent the character data.

- WLATIN1 (Single Byte)



- UTF-8 (Multiple Bytes)



# Understand the Transcoding Error

## Understand Encoding and Transcoding

- Encoding
- Character Sets
  - WLATIN1
    - 256 Characters
  - UTF-8
    - Over 120,000 characters

ASCII control characters			ASCII printable characters			Extended ASCII characters										
00	NULL	(Null character)	32	space	64	@	96	`	128	Ç	160	á	192	Ł	224	Ó
01	SOH	(Start of Header)	33	!	65	A	97	a	129	ü	161	í	193	ł	225	õ
02	STX	(Start of Text)	34	"	66	B	98	b	130	ë	162	ó	194	Ł	226	Ö
03	ETX	(End of Text)	35	#	67	C	99	c	131	â	163	ú	195	ł	227	Û
04	EOT	(End of Trans.)	36	\$	68	D	100	d	132	ä	164	ñ	196	—	228	ö
05	ENQ	(Enquiry)	37	%	69	E	101	e	133	à	165	Ñ	197	†	229	Û
06	ACK	(Acknowledgement)	38	&	70	F	102	f	134	ã	166	ª	198	â	230	µ
07	BEL	(Bell)	39	'	71	G	103	g	135	ç	167	º	199	Ã	231	þ
08	BS	(Backspace)	40	(	72	H	104	h	136	ê	168	¸	200	Ł	232	þ
09	HT	(Horizontal Tab)	41	)	73	I	105	i	137	ë	169	©	201	ł	233	Ú
10	LF	(Line feed)	42	^	74	J	106	j	138	è	170	¬	202	Ł	234	Ú
11	VT	(Vertical Tab)	43	+	75	K	107	k	139	ĩ	171	½	203	ł	235	Ú
12	FF	(Form feed)	44	,	76	L	108	l	140	î	172	¾	204	ł	236	ý
13	CR	(Carriage return)	45	-	77	M	109	m	141	ï	173	¿	205	=	237	Ÿ
14	SO	(Shift Out)	46	.	78	N	110	n	142	Ā	174	«	206	≠	238	-
15	SI	(Shift In)	47	/	79	O	111	o	143	Ă	175	»	207	≠	239	'
16	DLE	(Data link escape)	48	0	80	P	112	p	144	Ĕ	176	⋮	208	ð	240	≡
17	DC1	(Device control 1)	49	1	81	Q	113	q	145	æ	177	⋮	209	Ð	241	±
18	DC2	(Device control 2)	50	2	82	R	114	r	146	Æ	178	⋮	210	É	242	≡
19	DC3	(Device control 3)	51	3	83	S	115	s	147	ø	179	⋮	211	Ê	243	¼
20	DC4	(Device control 4)	52	4	84	T	116	t	148	ö	180	⋮	212	Ë	244	¶
21	NAK	(Negative acknowl.)	53	5	85	U	117	u	149	ò	181	À	213	Ì	245	§
22	SYN	(Synchronous idle)	54	6	86	V	118	v	150	û	182	Ā	214	Í	246	÷
23	ETB	(End of trans. block)	55	7	87	W	119	w	151	ù	183	Ă	215	Î	247	°
24	CAN	(Cancel)	56	8	88	X	120	x	152	ÿ	184	©	216	Ï	248	°
25	EM	(End of medium)	57	9	89	Y	121	y	153	Œ	185	⋮	217	Ĵ	249	°
26	SUB	(Substitute)	58	:	90	Z	122	z	154	Ů	186	⋮	218	Ŕ	250	°
27	ESC	(Escape)	59	;	91	[	123	{	155	ø	187	⋮	219	Ŗ	251	°
28	FS	(File separator)	60	<	92	\	124		156	£	188	⋮	220	Ÿ	252	°
29	GS	(Group separator)	61	=	93	]	125	}	157	Ø	189	¢	221	ı	253	°
30	RS	(Record separator)	62	>	94	^	126	~	158	×	190	¥	222	ı	254	■
31	US	(Unit separator)	63	?	95	_			159	f	191	₯	223	ı	255	nbsp
127	DEL	(Delete)														



# Understand the Transcoding Error

## Understand Encoding and Transcoding

- The way computer represents characters with coded binary digits

UTF-8	Coded value	WLATIN1
ü	11000011:10111100	Ã¼

- Distinguishing Differences
  - Storage size: number of bytes used to represent one character
  - Character set: number of characters capable of representing



# Understand the Transcoding Error

## SAS CEDA and Its Limitations

- SAS Strategy – Cross Environment Data Access (CEDA)

- Automatically transcode

NOTE: Data file is in a format that is native to another host, or the file encoding does not match the session encoding. Cross Environment Data Access will be used, which might require additional CPU resources and might reduce performance.

- Limitation

ERROR: Some character data was lost during transcoding in the dataset SOURCE.DM. Either the data contains characters that are not representable in the new encoding or truncation occurred during transcoding.

# Understand the Transcoding Error

## How Does the Error Occur

- Reason for unrepresentable characters

Name of Encoding	Character Set	Size
WLATIN1	ASCII and Extended ASCII character set	All 256 characters are stored in a single byte Single Byte Character Set (SBCS)
SHIFT-JIS	ASCII, Katakana, and other Japanese characters	Characters are stored in either 1 or 2 bytes. Double Byte Character Set (DBCS)
UTF-8	Unicode character set including ASCII, foreign languages, special symbols, and more	Over 120,000 Characters are stored in 1 to 4 bytes. Multi-Byte Character Set (MBCS)

# Understand the Transcoding Error

## How Does the Error Occur

- Reason for truncation

Name of Encoding	Character Set	Size
WLATIN1	ASCII and Extended ASCII character set	All 256 characters are stored in a single byte Single Byte Character Set (SBCS)
SHIFT-JIS	ASCII, Katakana, and other Japanese characters	Characters are stored in either 1 or 2 bytes. Double Byte Character Set (DBCS)
UTF-8	Unicode character set including ASCII, foreign languages, special symbols, and more	Over 120,000 Characters are stored in 1 to 4 bytes. Multi-Byte Character Set (MBCS)

# Troubleshoot the Transcoding Error

## Which is the Reason

ERROR: Some character data was lost during transcoding in the dataset SOURCE.DM. Either the data contains characters that **1** are not representable in the new encoding or **2** truncation occurred during transcoding.

Transcode from SBCS (WLATIN1) to MBCS (UTF-8) → Truncation


Transcode from large character set (UTF-8) to small character set (WLATIN1) → Unrepresentable Characters

# Troubleshoot the Transcoding Error

## What is the Encoding of the Input Data

- Proc Contents

**The CONTENTS Procedure**

<b>Data Set Name</b>	SOURCE.DM	<b>Observations</b>	437
<b>Member Type</b>	DATA	<b>Variables</b>	27
<b>Engine</b>	V9	<b>Indexes</b>	0
<b>Created</b>	02/01/2021 08:40:34	<b>Observation Length</b>	360
<b>Last Modified</b>	02/01/2021 08:40:34	<b>Deleted Observations</b>	0
<b>Protection</b>		<b>Compressed</b>	NO
<b>Data Set Type</b>		<b>Sorted</b>	YES
<b>Label</b>	Demographics		
<b>Data Representation</b>	WINDOWS_64		
 <b>Encoding</b>	wlatin1 Western (Windows)		

# Troubleshoot the Transcoding Error

## What is the Encoding of the Input Data

- The ATTRC Function

```
%let dsn=source.DM;  
%let dsid=%sysfunc(open(&dsn,i));  
%put &dsn ENCODING is: %sysfunc(attrc(&dsid,encoding));  
%let rc=%sysfunc(close(&dsid));
```

```
34          %put &dsn ENCODING is: %sysfunc(attrc(&dsid,encoding));
```

 source.DM ENCODING is: wlatin1 Western (Windows)

# Troubleshoot the Transcoding Error

## What is the Encoding of Your Current SAS Session

- Proc Options

```
32      proc options option=encoding;
33      run;

SAS (r) Proprietary Software Release 9.4 TS1M3

ENCODING=WLATIN1 Specifies the default character-set encoding for the SAS session.
NOTE: PROCEDURE OPTIONS used (Total process time):
      real time          0.00 seconds
      cpu time           0.00 seconds
```



# Troubleshoot the Transcoding Error

## What is the Encoding of Your Current SAS Session

- The GETOPTION Function

```
%put encoding=%sysfunc(getoption(encoding));
```

```
31
```

```
32      %put encoding=%sysfunc(getoption(encoding));
```

```
→ encoding=WLATIN1
```

```
33
```

# Resolve the Transcoding Error: Truncation

Avoid Data Truncation with the CVP Engine

Encoding with lower byte (WLATIN1)



Encoding with higher byte (UTF-8)

# Resolve the Transcoding Error: Truncation

## Avoid Data Truncation with the CVP Engine

- The Character Variable Padding (CVP) Engine
  - Expands the length of all character variables
  - An intermediate engine
  - Read-only engine

```
libname source cvp 'Source-data-library';  
libname target 'Target-data-library';  
proc copy noclone in=source out=target;  
run;
```

# Resolve the Transcoding Error: Truncation

## Avoid Data Truncation with the CVP Engine

- An Example
  - Input data set with special (extended ASCII) characters in the WLATIN1 Encoding

	CharVar
1	brød
2	über

Column Name	Type	Length
<del>A</del> CharVar	Text	4

- Transcode to UTF-8: Error Due to Data Truncation

# Resolve the Transcoding Error: Truncation

Avoid Data Truncation with the CVP Engine

```
libname source cvp 'Source-data-library';  
libname target 'Target-data-library';  
proc copy noclone in=source out=target;  
run;
```

Before CVP Engine Processing

After CVP Engine Processing

Alphabetic List of Variables and Attributes			
#	Variable	Type	Len
1	CharVar	Char	4

Alphabetic List of Variables and Attributes			
#	Variable	Type	Len
1	CharVar	Char	6

#SASGF

SAS<sup>®</sup> GLOBAL FORUM 2021

# Resolve the Transcoding Error: Truncation

## Avoid Data Truncation with the CVP Engine

- CVPMULTIPLIER= Option

```
libname source cvp 'Source-data-library' cvpmultiplier=2.0;
```

# Resolve the Transcoding Error: Unrepresentable Characters

## Unrepresentable Characters

Encoding with larger character set (UTF-8)



Encoding with smaller character set (WLATIN1)



# Resolve the Transcoding Error: Unrepresentable Characters

## Locate the Unrepresentable Characters

- Proc Copy

```
INFO: COPY with SELECT performance is in use.
NOTE: Copying RAWDATA.CM4 to INDATA.CM4 (memtype=DATA).
INFO: Data file RAWDATA.CM4.DATA is in a format that is native to another host, or the file
encoding does not match the session encoding. Cross Environment Data Access will be used, which
might require additional CPU resources and might reduce performance.
NOTE: System Options for BUFSIZE and REUSE were used at user's request.
NOTE: Libname and/or system options for compress, pointobs, data representation and encoding
attributes were used at user's request.
INFO: Engine's block-read method cannot be used because:
INFO:   - Cross Environment Data Access is being used
NOTE: There were 800 observations read from the data set RAWDATA.CM4.
ERROR: Some character data was lost during transcoding in the dataset RAWDATA.CM4. Either the data
contains characters that are not representable in the new encoding or truncation occurred
during transcoding.
ERROR: File INDATA.CM4.DATA has not been saved because copy could not be completed.
NOTE: Statements not processed because of errors noted above.
NOTE: PROCEDURE COPY used (Total process time):
      real time           2.41 seconds
      user cpu time       0.37 seconds
      system cpu time     0.06 seconds
      memory              6614 28k
```

#SASGF

SAS® GLOBAL FORUM 2021

# Resolve the Transcoding Error: Unrepresentable Characters

## Locate the Unrepresentable Characters

- Data Step Set Statement

```
7  data indata.cm4;
8  set rawdata.cm4;
INFO: Data file RAWDATA.CM4.DATA is in a format that is native to another host, or the file
encoding does not match the session encoding. Cross Environment Data Access will be used, which
might require additional CPU resources and might reduce performance.
9  run;
```

**ERROR:** Some character data was lost during transcoding in the dataset RAWDATA.CM4. Either the data contains characters that are not representable in the new encoding or truncation occurred during transcoding.

**NOTE:** The DATA step has been abnormally terminated.

**NOTE:** The SAS System stopped processing this step because of errors.

**NOTE:** There were 799 observations read from the data set RAWDATA.CM4.

**WARNING:** The data set INDATA.CM4 may be incomplete. When this step was stopped there were 799 observations and 66 variables.

**WARNING:** Data set INDATA.CM4 was not replaced because this step was stopped.

**NOTE:** DATA statement used (Total process time):

real time	1.18 seconds
user cpu time	0.23 seconds
system cpu time	0.06 seconds
memory	1393.06k
OS Memory	15832.00k
Timestamp	03/27/2018 12:02:25 PM
Step Count	3 Switch Count 0

# Resolve the Transcoding Error: Unrepresentable Characters

## Locate the Unrepresentable Characters

- Suppress the transcoding with the ENCODING= option

```
Data indata.cm4;  
  Set rawdata.cm4 (encoding=asciiany);  
Run;
```

```
16  data indata.cm4;  
17  set rawdata.cm4(encoding=asciiany);  
18  run;  
  
NOTE: There were 1163 observations read from the data set RAWDATA.CM4  
NOTE: The data set INDATA.CM4 has 1163 observations and 66 variables.  
NOTE: DATA statement used (Total process time):  
      real time           0.84 seconds  
      user cpu time       0.03 seconds  
      system cpu time     0.15 seconds  
      memory              1032.25k  
      OS Memory          15832.00k  
      Timestamp           03/27/2018 12:06:41 PM  
      Step Count          6  Switch Count  0
```

# Resolve the Transcoding Error: Unrepresentable Characters

## Locate the Unrepresentable Characters

- Suppress the transcoding with the ENCODING= option

	Reported Name of Drug Med or Therapy	Reported Name of Drug Med or TherapyATC
795	REVATIO	ANTIHYPERTENSIVES
796	CALONAL	NERVOUS SYSTEM
797	RIKAVARIN	STOMATOLOGICAL PREPARATIONS
798	XYLOCAINE(LIDOCAINE)	ANESTHETICS FOR TOPICAL USE
799	CARBOCYSTEINE	COUGH AND COLD PREPARATIONS
800	MEDICON <del>8881</del> ¼ DEXTROMETHORPH HYDROBROMIDE HYDRATE <del>1¼%</del>	OPIUM ALKALOIDS AND DERIVATIVES
801	TALION	OTHER ANTIHISTAMINES FOR SYSTEMIC USE
802	KEFRAL	OTHER BETA-LACTAM ANTIBACTERIALS
803	TOCOPHEROL NICOTINATE	CARDIOVASCULAR SYSTEM

#SASGF

SAS® GLOBAL FORUM 2021

# Resolve the Transcoding Error: Unrepresentable Characters

## Locate the Unrepresentable Characters

- Suppress the transcoding with the ENCODING= option
  - How does the ENCODING=ASCIANY option work?

UTF-8	Coded value	WLATIN1
ü	11000011:10111100	Ã¼

- Does not completely resolve the problem

# Resolve the Transcoding Error: Unrepresentable Characters

## Make Encodings Compatible

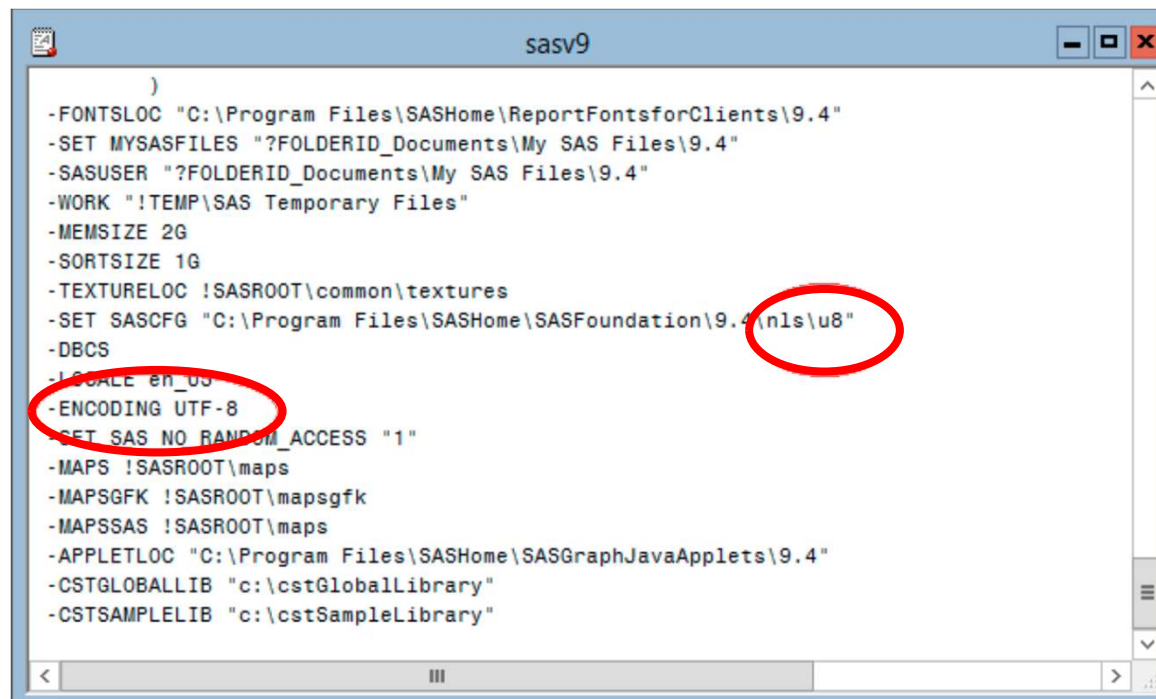
- Update SAS session encoding to
  - UTF-8
  - the same encoding used in the data



# Resolve the Transcoding Error: Unrepresentable Characters

## Make Encodings Compatible

- What is a configuration file (SASV9.CFG)



```
)  
-FONTSLOC "C:\Program Files\SASHome\ReportFontsforClients\9.4"  
-SET MYSASFILES "?FOLDERID_Documents\My SAS Files\9.4"  
-SASUSER "?FOLDERID_Documents\My SAS Files\9.4"  
-WORK "!TEMP\SAS Temporary Files"  
-MEMSIZE 2G  
-SORTSIZE 1G  
-TEXTURELOC !SASROOT\common\textures  
-SET SASCFG "C:\Program Files\SASHome\SASFoundation\9.4\nls\u8"  
-DBCS  
-LSCALE en_US  
-ENCODING UTF-8  
-SET SAS NO_RANDOM_ACCESS "1"  
-MAPS !SASROOT\maps  
-MAPSGFK !SASROOT\mapsgfk  
-MAPSSAS !SASROOT\maps  
-APPLETLOC "C:\Program Files\SASHome\SASGraphJavaApplets\9.4"  
-CSTGLOBALLIB "c:\cstGlobalLibrary"  
-CSTSAMPLELIB "c:\cstSampleLibrary"
```

#SASGF

SAS<sup>®</sup> GLOBAL FORUM 2021



# Resolve the Transcoding Error: Unrepresentable Characters

## Make Encodings Compatible

- Where is your configuration file

```
1  proc options option=config;  
2  run;
```

SAS (r) Proprietary Software Release 9.4 TS1M2

```
CONFIG=C:\Program Files\SASHome\SASFoundation\9.4\nls\en\SASV9.CFG
```

Specifies the configuration file that is used when initializing or overriding the values of SAS system options.

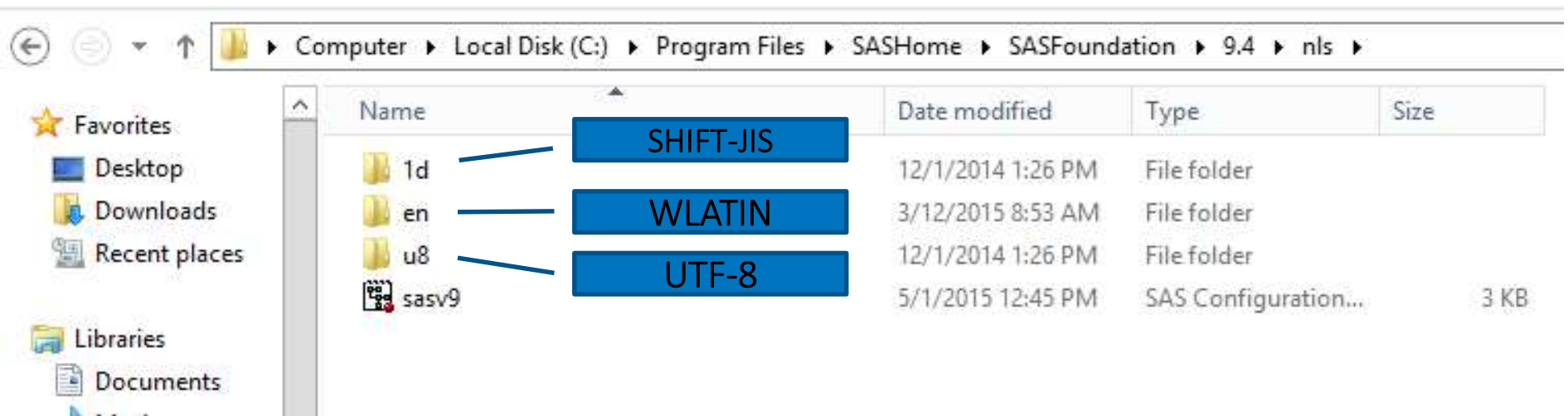
NOTE: PROCEDURE OPTIONS used (Total process time):

real time	0.06 seconds
user cpu time	0.00 seconds
system cpu time	0.00 seconds
memory	19.09k
OS Memory	8672.00k
Timestamp	03/26/2018 01:30:49 PM
Step Count	2 Switch Count 0

# Resolve the Transcoding Error: Unrepresentable Characters

## Make Encodings Compatible

- Where is your configuration file

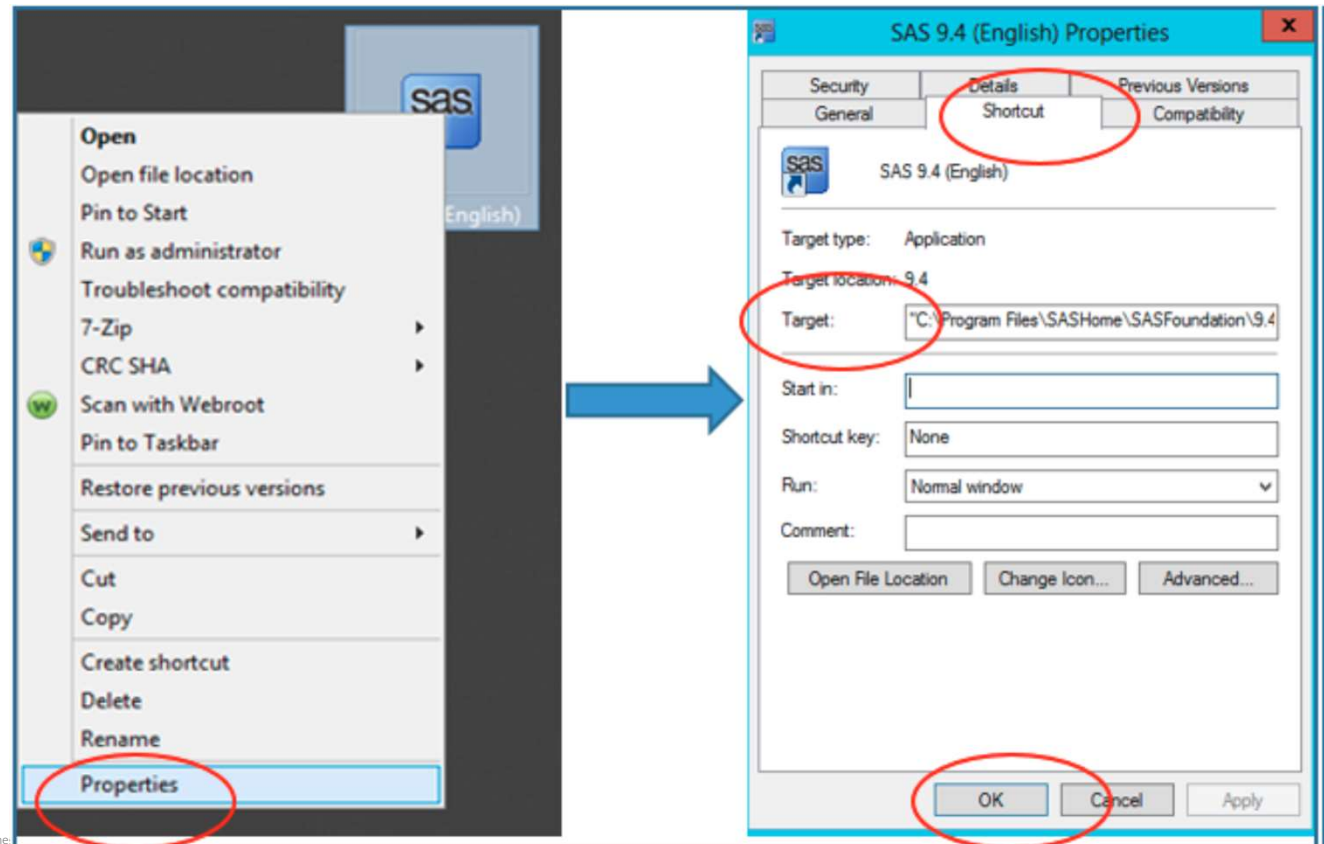


- Use the configuration file with the right encoding

# Resolve the Transcoding Error: Unrepresentable Characters

## Make Encodings Compatible

- Update your SAS session encoding
- Method 1: Customizing SAS Shortcut Icon



#SASGF

SAS and all other SAS Institute Inc. product or service names

# Resolve the Transcoding Error: Unrepresentable Characters

## Make Encodings Compatible

- Update your SAS session encoding
  - Method 1: Customizing SAS Shortcut Icon

"C:\Program Files\SASHome\SASFoundation\9.4\sas.exe" -CONFIG "C:\Program Files\SASHome\SASFoundation\9.4\nls\en\SASV9.CFG"

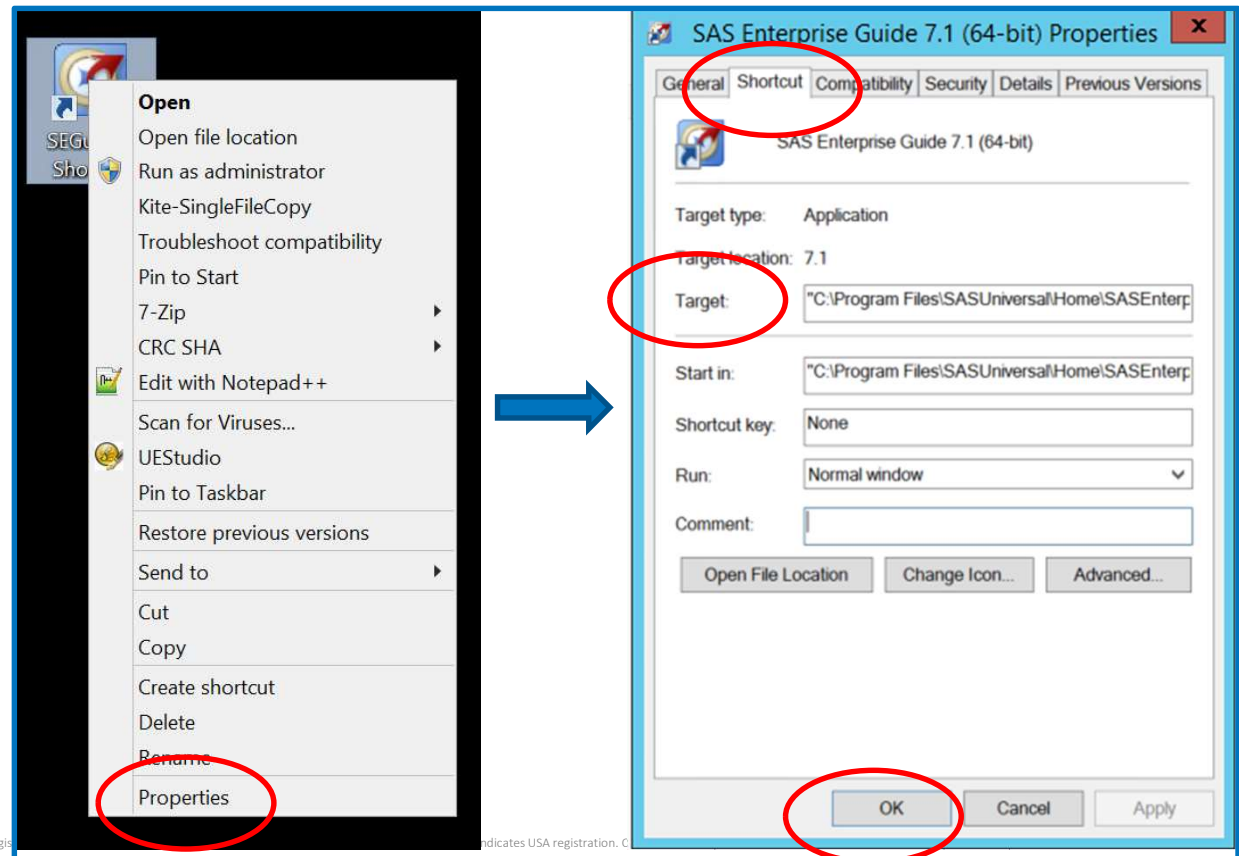


"C:\Program Files\SASHome\SASFoundation\9.4\sas.exe" -CONFIG "C:\Program Files\SASHome\SASFoundation\9.4\nls\u8\SASV9.CFG"

# Resolve the Transcoding Error: Unrepresentable Characters

## Make Encodings Compatible

- Update your SAS session encoding
- Method 1:  
Customizing SAS  
Shortcut Icon



#SASGF

SAS and all other SAS Institute Inc. product or service names are registered trademarks of SAS Institute Inc. in the USA and other countries.

indicates USA registration. ©

# Resolve the Transcoding Error: Unrepresentable Characters

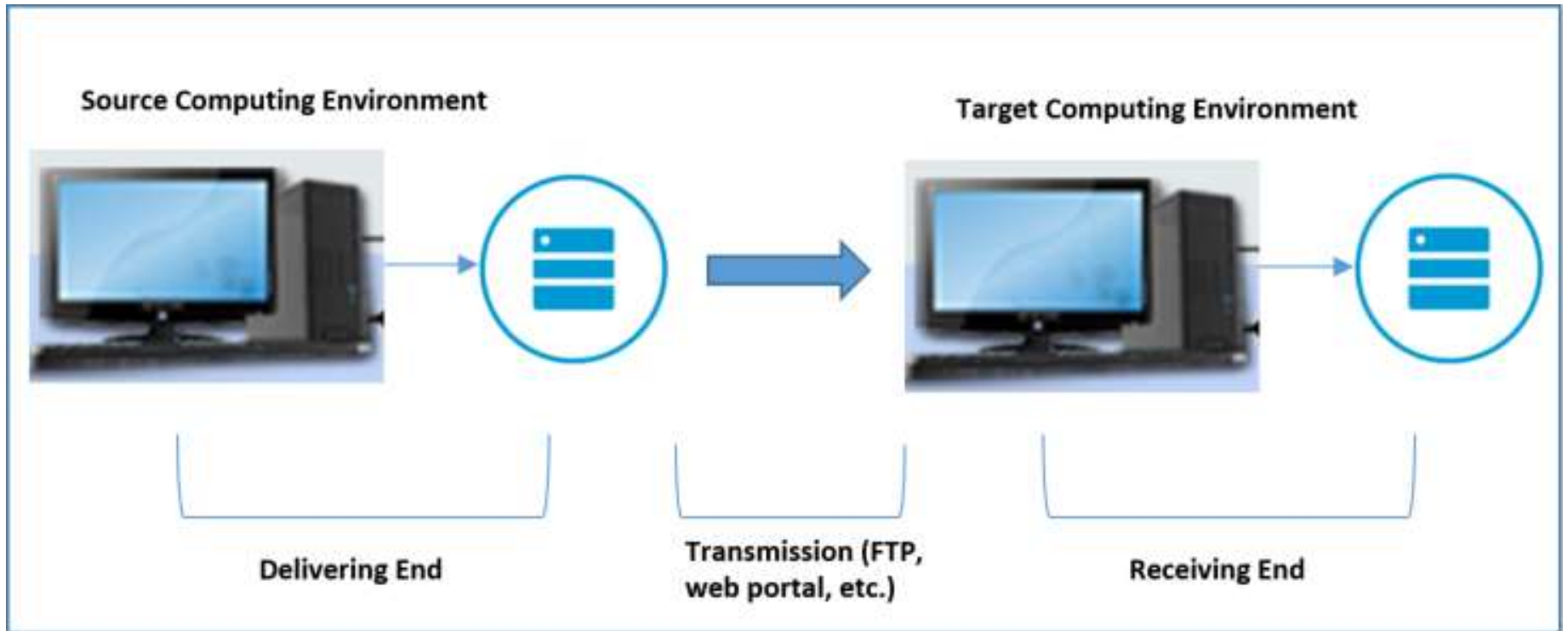
## Make Encodings Compatible

- Update your SAS session encoding
  - Method 2: the CONFIG option for SAS command line invocation

```
C:\Progra~1\SASHome\SASFou~1\9.4\sas.exe  
    -sysin ProgramName.sas  
    -config C:\Progra~1\SASHome\SASFou~1\9.4\nls\u8\SASV9.CFG
```

# Resolve the Transcoding Error: Unrepresentable Characters

Compatible Encoding = No More Transcoding Error



#SASGF

SAS® GLOBAL FORUM 2021

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® Indicates USA registration. Other brand and product names are trademarks of their respective companies.



# Conclusions

ERROR: Some character data was lost during transcoding in the dataset SOURCE.DM. Either the data contains characters that are not representable in the new encoding or truncation occurred during transcoding.

- Use the CVP engine if truncation is the problem
- Convert your SAS session encoding if the problem is unrepresentable characters

The background is a solid blue color with several faint, overlapping circular patterns of varying shades of blue. These patterns consist of concentric arcs and dots, creating a sense of depth and movement.

# Thank you!

Contact Information  
[ZhuoYun@PRAHS.com](mailto:ZhuoYun@PRAHS.com)

#SASGF

SAS<sup>®</sup> GLOBAL FORUM 2021

#SASGF

# SAS® GLOBAL FORUM 2021

[sasglobalforum.com](https://sasglobalforum.com)