PharmaSUG
NORTH CAROLINA • OCT. 21, 2022

A Quick Look at Fuzzy Matching Programming Techniques Using SAS® Software

a presentation by

Kirk Paul Lafler and Stephen B. Sloan

PharmaSUG
NORTH CAROLINA • OCT. 21, 2022

A Quick Look at Fuzzy Matching Programming Techniques Using SAS® Software

Stephen B. Sloan

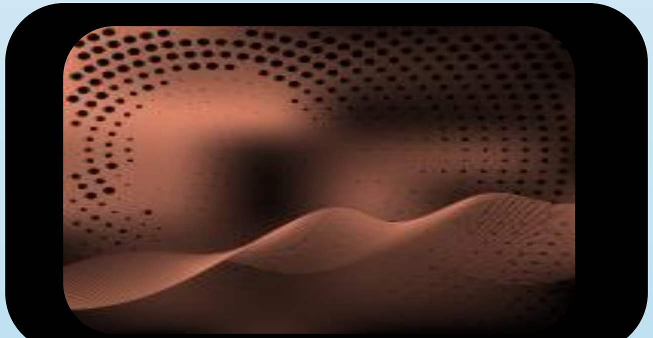
Stephen has worked at Accenture in the Services, Consulting, and Digital groups and is currently a senior manager in the SAS Analytics area. He has worked in a variety of functional areas including Project Management, Data Management, and Statistical Analysis. Stephen has had the good fortune to have worked with many talented people at SAS Institute. Stephen has presented at over 20 SAS conferences and been published in professional journals. Stephen has a B.A. cum laude with Honor in Mathematics from Brandeis University, M.S. degrees in Mathematics and Computer Science from Northern Illinois University, an MBA from Stern Business School at New York University. (1st in class), and a graduate certificate in Financial Analytics from Stevens Institute.

Kirk Paul Lafler

Kirk Paul Lafler is a lecturer and adjunct professor at San Diego State University; an advisor and adjunct professor at the University of California San Diego Extension; and teaches SAS, SQL, Python, R and Excel courses, seminars, workshops, and webinars to students, professionals and users around the world. Kirk has been a SAS user since 1979 and is the author of several books including, PROC SQL: Beyond the Basics Using SAS, Third Edition (SAS Press. 2019) along with papers and articles on a variety of SAS topics. Kirk has also been selected as an Invited speaker, educator, keynote and section leader at SAS conferences; and is the recipient of 25 “Best” contributed paper, hands-on workshop (HOW), and poster awards.

Presentation Objectives

The Fuzzy Matching Process Explained

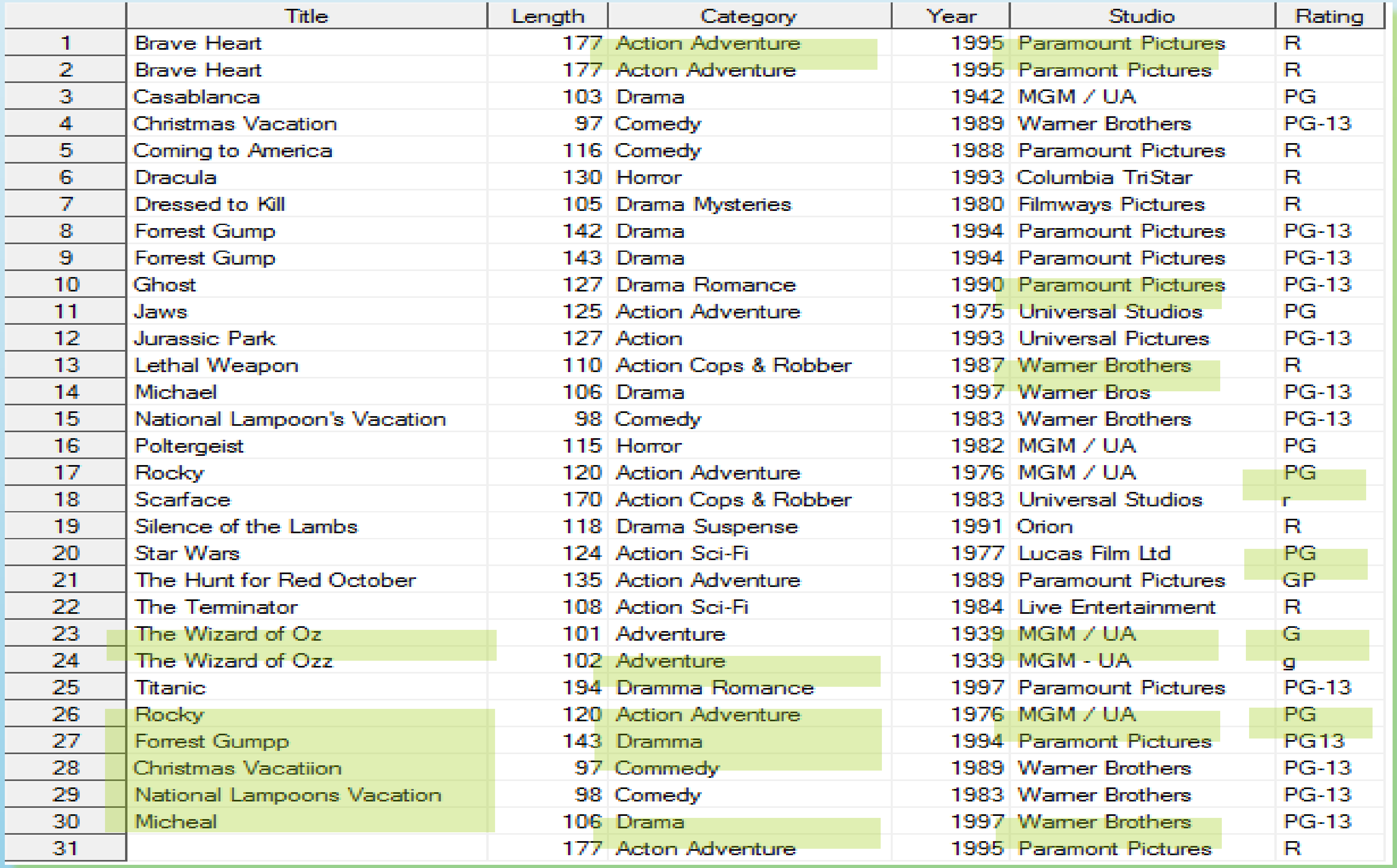


Fuzzy Matching Programming Techniques

Fuzzy Matching Programming Examples

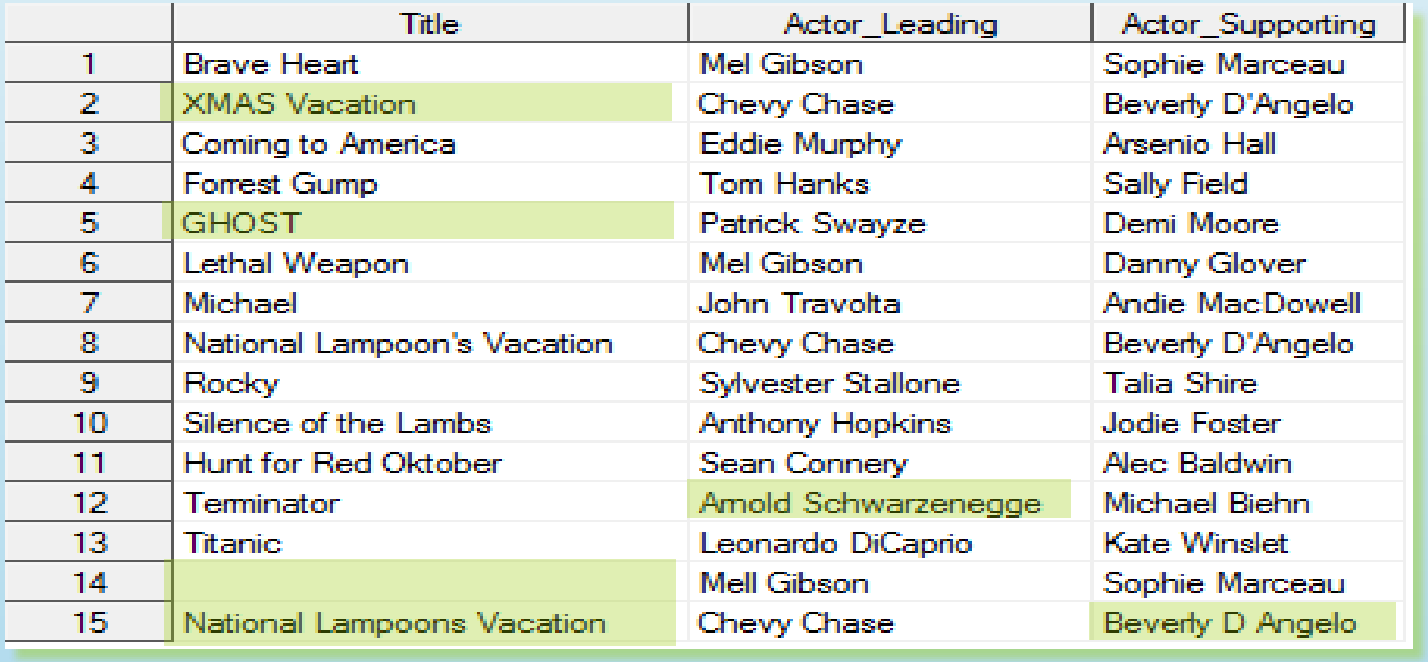
PharmaSUG
NORTH CAROLINA • OCT. 21, 2022

Movies_with_Messy_Data



	Title	Length	Category	Year	Studio	Rating
1	Brave Heart	177	Action Adventure	1995	Paramount Pictures	R
2	Brave Heart	177	Acton Adventure	1995	Paramont Pictures	R
3	Casablanca	103	Drama	1942	MGM / UA	PG
4	Christmas Vacation	97	Comedy	1989	Warner Brothers	PG-13
5	Coming to America	116	Comedy	1988	Paramount Pictures	R
6	Dracula	130	Horror	1993	Columbia TriStar	R
7	Dressed to Kill	105	Drama Mysteries	1980	Filmways Pictures	R
8	Forrest Gump	142	Drama	1994	Paramount Pictures	PG-13
9	Forrest Gump	143	Drama	1994	Paramount Pictures	PG-13
10	Ghost	127	Drama Romance	1990	Paramount Pictures	PG-13
11	Jaws	125	Action Adventure	1975	Universal Studios	PG
12	Jurassic Park	127	Action	1993	Universal Pictures	PG-13
13	Lethal Weapon	110	Action Cops & Robber	1987	Warner Brothers	R
14	Michael	106	Drama	1997	Warner Bros	PG-13
15	National Lampoon's Vacation	98	Comedy	1983	Warner Brothers	PG-13
16	Poltergeist	115	Horror	1982	MGM / UA	PG
17	Rocky	120	Action Adventure	1976	MGM / UA	PG
18	Scarface	170	Action Cops & Robber	1983	Universal Studios	r
19	Silence of the Lambs	118	Drama Suspense	1991	Orion	R
20	Star Wars	124	Action Sci-Fi	1977	Lucas Film Ltd	PG
21	The Hunt for Red October	135	Action Adventure	1989	Paramount Pictures	GP
22	The Terminator	108	Action Sci-Fi	1984	Live Entertainment	R
23	The Wizard of Oz	101	Adventure	1939	MGM / UA	G
24	The Wizard of Ozz	102	Adventure	1939	MGM - UA	g
25	Titanic	194	Dramma Romance	1997	Paramount Pictures	PG-13
26	Rocky	120	Action Adventure	1976	MGM / UA	PG
27	Forrest Gumpp	143	Dramma	1994	Paramont Pictures	PG13
28	Christmas Vacatiion	97	Commedy	1989	Warner Brothers	PG-13
29	National Lampoons Vacation	98	Comedy	1983	Warner Brothers	PG-13
30	Micheal	106	Drama	1997	Warner Brothers	PG-13
31		177	Acton Adventure	1995	Paramont Pictures	R

Actors_with_Messy_Data



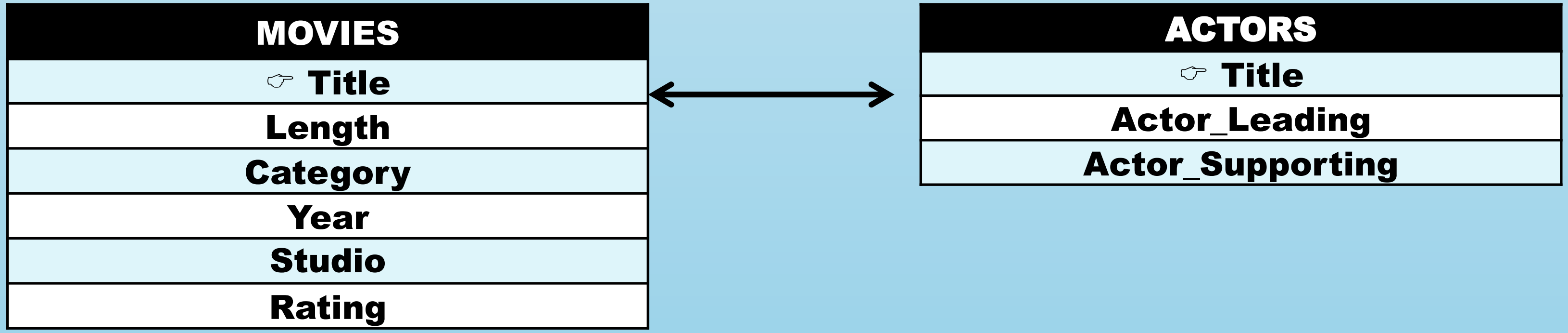
	Title	Actor_Leading	Actor_Supporting
1	Brave Heart	Mel Gibson	Sophie Marceau
2	XMAS Vacation	Chevy Chase	Beverly D'Angelo
3	Coming to America	Eddie Murphy	Arsenio Hall
4	Forrest Gump	Tom Hanks	Sally Field
5	GHOST	Patrick Swayze	Demi Moore
6	Lethal Weapon	Mel Gibson	Danny Glover
7	Michael	John Travolta	Andie MacDowell
8	National Lampoon's Vacation	Chevy Chase	Beverly D'Angelo
9	Rocky	Sylvester Stallone	Talia Shire
10	Silence of the Lambs	Anthony Hopkins	Jodie Foster
11	Hunt for Red Oktober	Sean Connery	Alec Baldwin
12	Terminator	Arnold Schwarzenegge	Michael Biehn
13	Titanic	Leonardo DiCaprio	Kate Winslet
14		Mell Gibson	Sophie Marceau
15	National Lampoons Vacation	Chevy Chase	Beverly D Angelo

The Fuzzy Matching Process Explained

Matching with Common Keys

- Data exists in many forms (text files, JSON, delimited files, CSVs, spreadsheets, datasets, RDBMS) and uses the key(s) in one or more data sources to match and/or create a combined file;

- Using a common and reliable identifier (or key), two or more datasets can be matched, merged or joined.



MOVIES
☞ Title
Length
Category
Year
Studio
Rating

ACTORS
☞ Title
Actor_Leading
Actor_Supporting

But, what happens when a shared and reliable key between data sources is nonexistent, inexact, or unreliable? This can make the matching process more complicated and problematic.

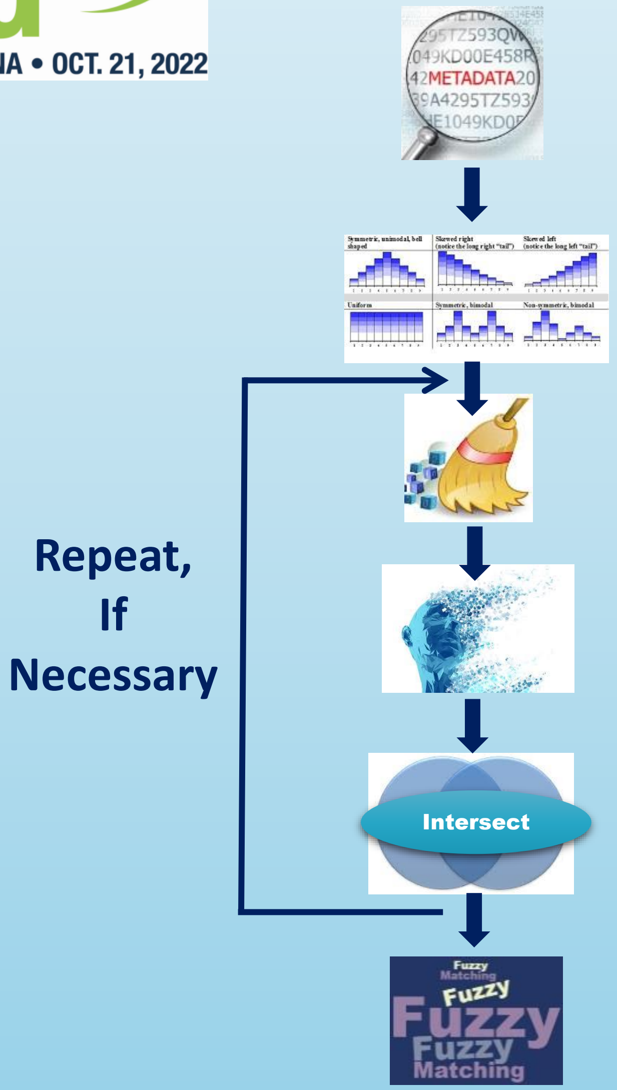
Matching Challenges

Phonetic Similarity
Michael ←→ Micheal
Smith ←→ Smythe
	Missing Spaces & Hyphens
	Mary Ann ←→ MaryAnn
	Mary-Ann ←→ Mary-Anne
		Missing Components
		Mary Frank ←→ Mary Ann Frank
		John Smith ←→ John F. Smith

Spelling Differences
Honor ←→ Honour
Behavior ←→ Behaviour
Labor ←→ Labour
	Titles & Honorifics
	Mr. ←→ Mister
	Ms. ←→ Miss
	Dr. ←→ Ph.D
		Nicknames
		Bill ←→ William
		Dave ←→ David
		Liz ←→ Elizabeth

Truncated Components
Ct. ←→ Court
Ave. ←→ Avenue
Rd. ←→ Road
	Initials & Abbreviations
	J. Smith ←→ John Smith
	Robo ←→ Robo Inc.
		Similar Names
		ABC Co. ←→ ABC Corporation
		Robo LLC ←→ Robo Inc.

6 Step Fuzzy Matching Process



Step 1: Understand Matching Scenarios
Determine the Likely Matching Variables using Metadata.

Step 2: Explore Data Values and Data Types
Understand the Distribution of Data Values.

Step 3: Data Cleaning
Perform Data Cleaning.

Step 4: Data Transformation
Perform Data Transformations.

Step 5: Exact Matching (Inner / Outer Joins)
Process Exact Matches.

Step 6: Fuzzy Matching (Soundex, Spedis, CompLEV, CompGED)
Match Key Fields using Fuzzy Matching Techniques.

Fuzzy Matching Programming Techniques with Examples

Side-by-side Comparison

Soundex	SPEDIS	COMPLEV	COMPGED
Algorithm	Function	Function	Function
Matches words that sound alike (phonetic match)	Translates a word into a smallest distance value	Computes a Levenshtein Edit Distance (LEV) score	Computes a Generalized Edit Distance (GED) score
Ignores case, embedded blanks and punctuations	Computes cost to convert a keyword to query	Computes # of operations to convert between two strings	Computes the minimum cost to convert between two strings
Works best with English-sounding names, not with others	Determines spelling distance between two words	Counts number of insertions, deletions, or replacements	Determines cost associated with each conversion
Assigns a code to each letter and compares code	Matching process can be controlled with logic	Matching process can be controlled with logic	Matching process can be controlled with logic

Fuzzy Matching Using the SOUNDEX Algorithm

SOUNDEX Algorithm

- The SOUNDEX algorithm / function matches character strings in files (or data sets) on words that sound alike;
- Soundex was invented and patented by Margaret K. Odell and Robert C. Russell in 1918 and 1922 to help match surnames that sound alike.

How SOUNDEX Works

- SAS evaluates whether a variable's contents sound alike by converting each word to a code;
- The value assigned consists of the first letter in the word followed by one or more digits;
- Vowels, A, E, I, O and U, along with H, W, Y, and non-alphabetical characters are ignored;
- Double letters (e.g., 'TT') are assigned a single value for both letters.

Derived SOUNDEX Codes

Letter	Value
B, P, F, V	1
C, S, G, J, K, Q, X, Z	2
D, T	3
L	4
M, N	5
R	6

Let’s examine the SOUNDEX algorithm with the movie title, Rocky.

R is assigned a value of 6 but is retained as R; O is ignored; C is assigned a value of 2; K is assigned a value of 2; and Y is ignored. The derived code for “Rocky” is, R22, which is matched against other movie titles.

SOUNDEX Algorithm Function Syntax

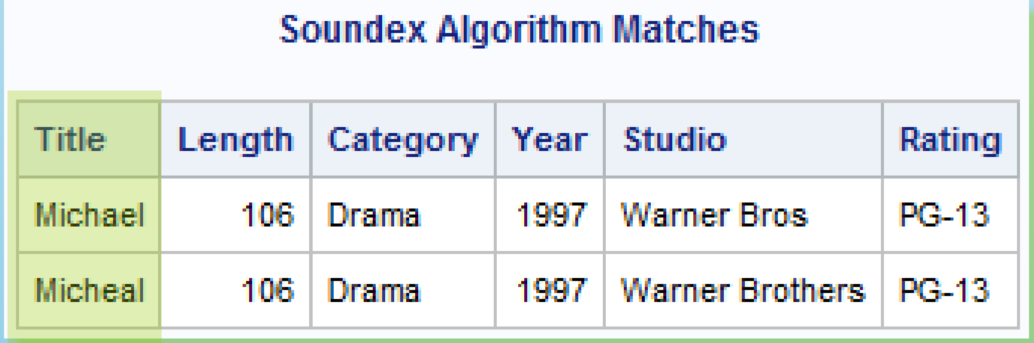
The syntax for the SOUNDEX algorithm is:

Variable =* "character-string"

SOUNDEX(Title) AS SOUNDEX_Value

Soundex Algorithm with PROC PRINT

```
PROC PRINT DATA=mydata.Movies_with_Messy_Data NOOBS ;
  TITLE "Soundex Algorithm Matches" ;
  WHERE Title =* "Michael" ;
RUN ;
```



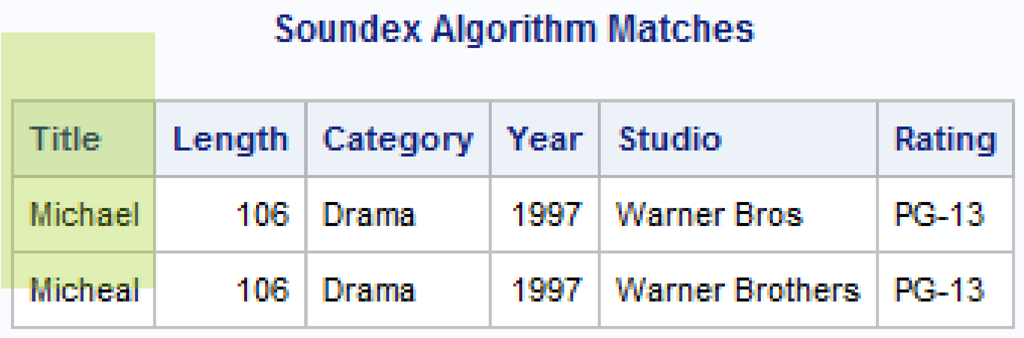
Soundex Algorithm Matches

Title	Length	Category	Year	Studio	Rating
Michael	106	Drama	1997	Warner Bros	PG-13
Micheal	106	Drama	1997	Warner Brothers	PG-13

SOUNDEX Algorithm

Soundex Algorithm with PROC SQL

```
TITLE “Soundex Algorithm Matches” ;
PROC SQL ;
 SELECT *
  FROM mydata.Movies_with_Messy_Data
   WHERE Title =* “Michael” ;
QUIT ;
```



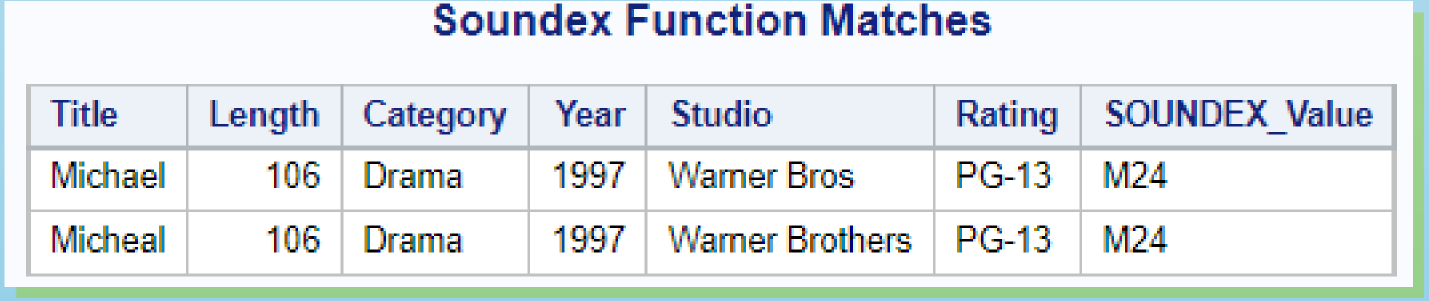
Soundex Algorithm Matches

Title	Length	Category	Year	Studio	Rating
Michael	106	Drama	1997	Warner Bros	PG-13
Micheal	106	Drama	1997	Warner Brothers	PG-13

SOUNDEX Function

Soundex Function with PROC SQL

```
TITLE “Soundex Function Matches” ;
PROC SQL ;
  SELECT *, SOUNDEX(Title,“Michael”) AS SOUNDEX_Value
    FROM mydata.Movies_with_Messy_Data
      WHERE UPCASE(Title) LIKE “MICH%” ;
QUIT ;
```



Soundex Function Matches

Title	Length	Category	Year	Studio	Rating	SOUNDEX_Value
Michael	106	Drama	1997	Warner Bros	PG-13	M24
Micheal	106	Drama	1997	Warner Brothers	PG-13	M24

PharmaSUG NORTH CAROLINA • OCT. 21, 2022

Fuzzy Matching Using the SPEDIS Function

SPEDIS Function

- The SPEDIS (spelling distance) function evaluates matching scenarios by translating a keyword into its smallest distance value;
- The SPEDIS function returns a non-negative value;
- A SPEDIS value of zero is returned when the query and arguments match exactly;
- Users are able to specify spelling distance values greater than zero (e.g., 10, 20, etc.).

How the SPEDIS Function Works

Operation	Cost	Description
Match	0	No change
Singlet	25	Delete one of a double letter
Doublet	50	Double a letter
Swap	50	Reverse the order of two consecutive letters
Truncate	50	Delete a letter from the end
Append	35	Add a letter to the end
Delete	50	Delete a letter from the middle
Insert	100	Insert a letter in the middle
Replace	100	Replace a letter in the middle
Firstdel	100	Delete the first letter
Firstins	200	Insert a letter at the beginning
Firstrep	200	Replace the first letter

The distance is the sum of the costs divided by the length of the query.

Source: http://support.sas.com/documentation/cdl/en/lefunctionsref/69762/HTML/default/viewer.htm#p0vmuxh8ljfn7on164nsgvmdrc5d.htm

SPEDIS Example

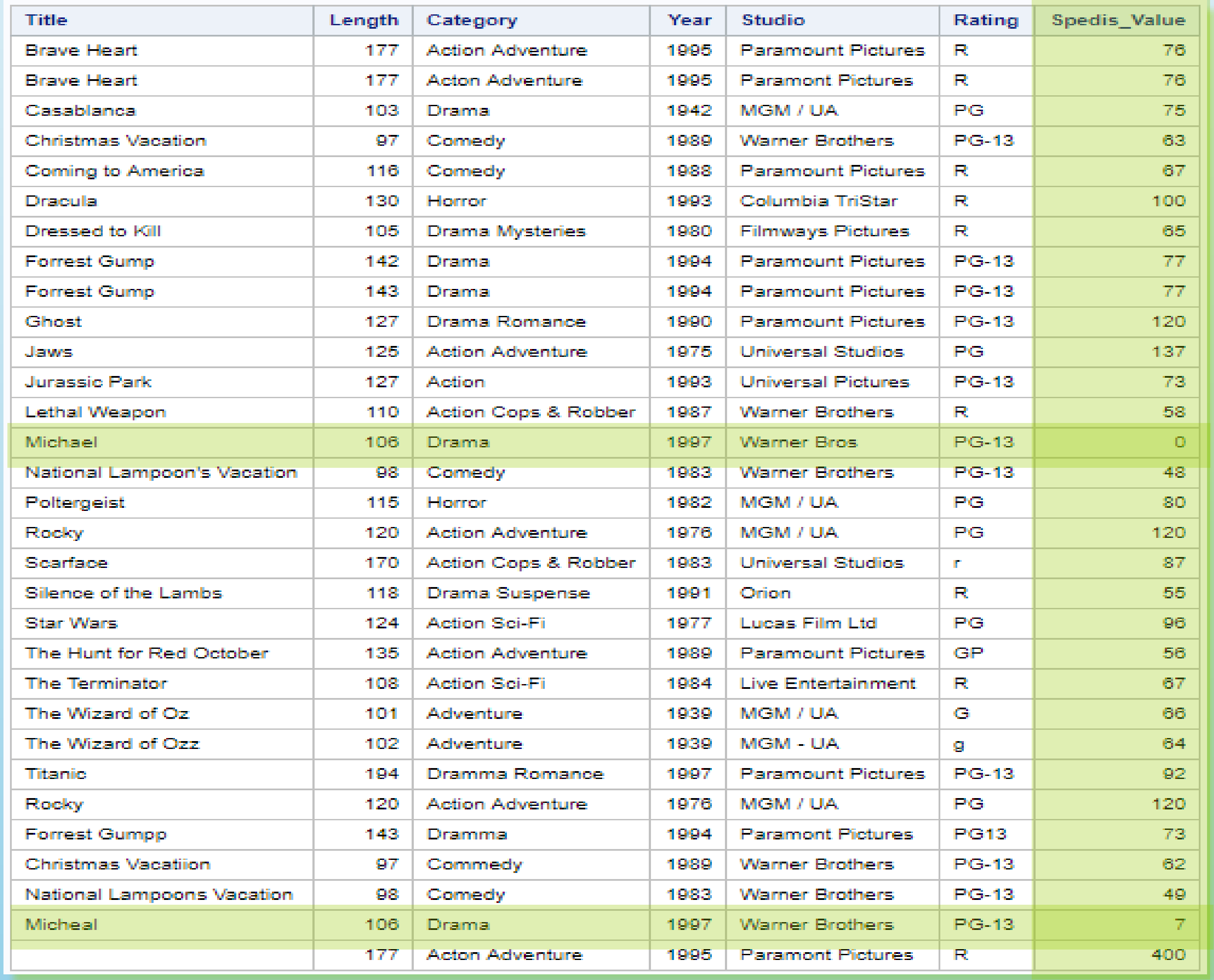
The general syntax for the SPEDIS function is:

SPEDIS (query, keyword)

SPEDIS Function with PROC SQL

```
PROC SQL ;
 SELECT *,
     SPEDIS(Title,"Michael") AS Spedis_Value
  FROM mydata.Movies_with_Messy_Data
   WHERE CALCULATED Spedis_Value GE 0 ;
QUIT ;
```

SPEDIS Example Results

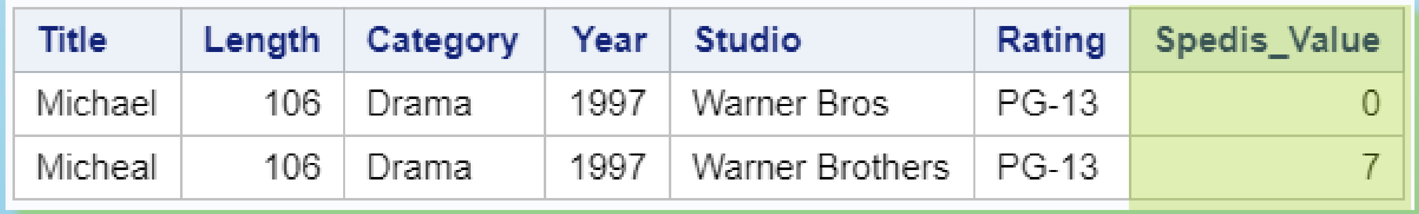


Title	Length	Category	Year	Studio	Rating	Spedis_Value
Brave Heart	177	Action Adventure	1995	Paramount Pictures	R	76
Brave Heart	177	Acton Adventure	1995	Paramont Pictures	R	76
Casablanca	103	Drama	1942	MGM / UA	PG	75
Christmas Vacation	97	Comedy	1989	Warner Brothers	PG-13	63
Coming to America	116	Comedy	1988	Paramount Pictures	R	67
Dracula	130	Horror	1993	Columbia TriStar	R	100
Dressed to Kill	105	Drama Mysteries	1980	Filmways Pictures	R	65
Forrest Gump	142	Drama	1994	Paramount Pictures	PG-13	77
Forrest Gump	143	Drama	1994	Paramount Pictures	PG-13	77
Ghost	127	Drama Romance	1990	Paramount Pictures	PG-13	120
Jaws	125	Action Adventure	1975	Universal Studios	PG	137
Jurassic Park	127	Action	1993	Universal Pictures	PG-13	73
Lethal Weapon	110	Action Cops & Robber	1987	Warner Brothers	R	58
Michael	106	Drama	1997	Warner Bros	PG-13	0
National Lampoon's Vacation	98	Comedy	1983	Warner Brothers	PG-13	48
Poltergeist	115	Horror	1982	MGM / UA	PG	80
Rocky	120	Action Adventure	1976	MGM / UA	PG	120
Scarface	170	Action Cops & Robber	1983	Universal Studios	r	87
Silence of the Lambs	118	Drama Suspense	1991	Orion	R	55
Star Wars	124	Action Sci-Fi	1977	Lucas Film Ltd	PG	96
The Hunt for Red October	135	Action Adventure	1989	Paramount Pictures	GP	56
The Terminator	108	Action Sci-Fi	1984	Live Entertainment	R	67
The Wizard of Oz	101	Adventure	1939	MGM / UA	G	66
The Wizard of Ozz	102	Adventure	1939	MGM - UA	g	64
Titanic	194	Dramma Romance	1997	Paramount Pictures	PG-13	92
Rocky	120	Action Adventure	1976	MGM / UA	PG	120
Forrest Gumpp	143	Dramma	1994	Paramont Pictures	PG13	73
Christmas Vacatiion	97	Commedy	1989	Warner Brothers	PG-13	62
National Lampoons Vacation	98	Comedy	1983	Warner Brothers	PG-13	49
Micheal	106	Drama	1997	Warner Brothers	PG-13	7
	177	Acton Adventure	1995	Paramont Pictures	R	400

SPEDIS Example #2

SPEDIS Function with PROC SQL

```
PROC SQL ;
 SELECT *,
        SPEDIS(Title,"Michael") AS Spedis_Value
  FROM mydata.Movies_with_Messy_Data
   WHERE CALCULATED Spedis_Value LE 7 ;
QUIT ;
```



Title	Length	Category	Year	Studio	Rating	Spedis_Value
Michael	106	Drama	1997	Warner Bros	PG-13	0
Micheal	106	Drama	1997	Warner Brothers	PG-13	7

Fuzzy Matching Using the COMPLEV Function

COMPLEV Function

- The COMPLEV function stands for Levenshtein Edit Distance;
- The COMPLEV function provides an indication of how close two strings are;
- It returns the number of operations that have been performed;
- The lower the number of operations the better the match (e.g., 0=Best match, 1=Next Best match, etc.).

COMPLEV Function Arguments

The general syntax for the COMPLEV function is:

COMPLEV (string-1, string-2 <,cutoff-value> <,modifier>)

Optional Arguments:

Cutoff-value specifies a numeric variable, constant or expression.

Modifier specifies a value that alters the action of the COMPLEV function. Valid modifier values are:

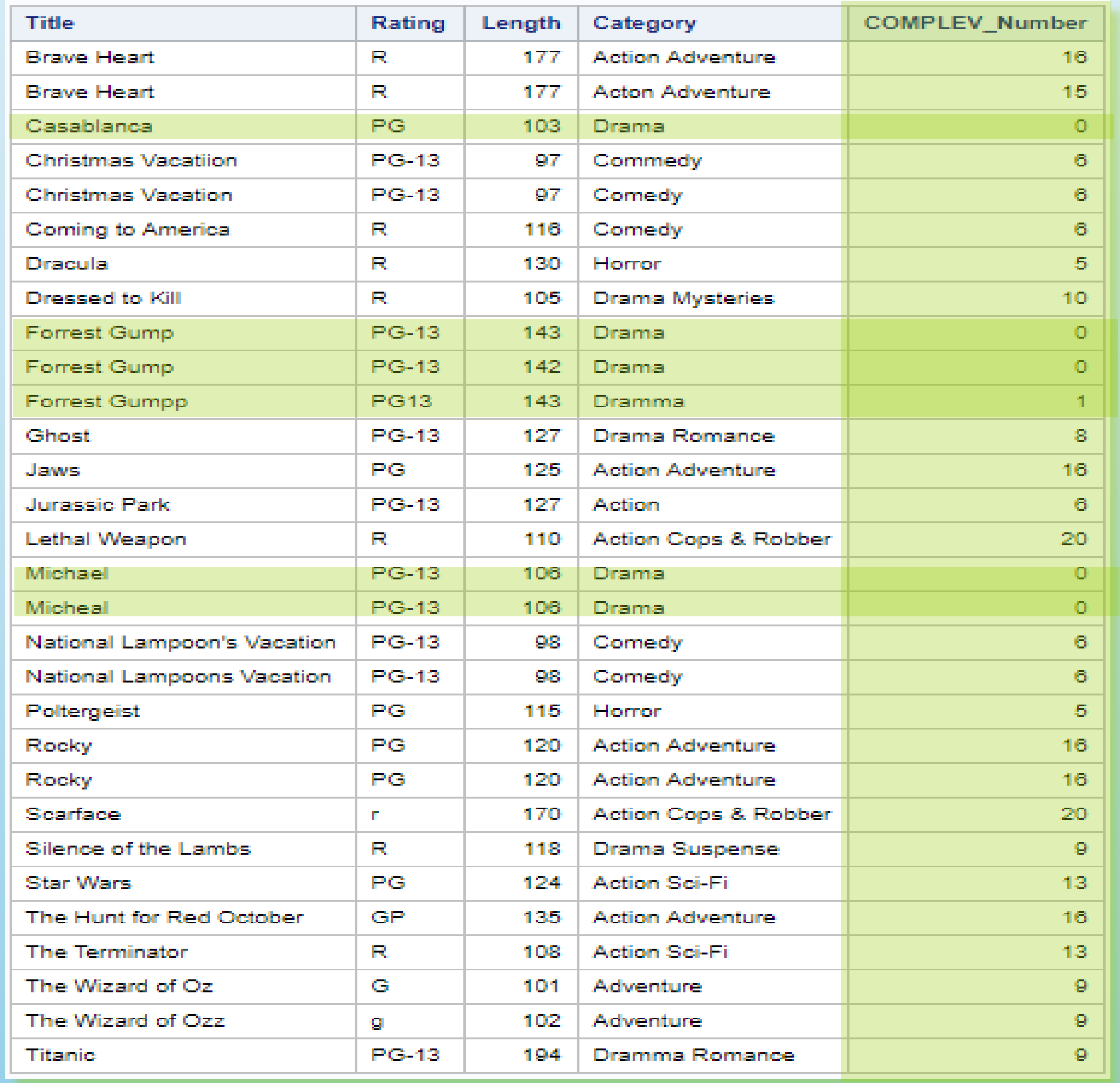
- ✓ i or I Ignores the case (case insensitive) in string-1 and string-2.
- ✓ l or L Removes leading blanks before comparing the values in string-1 or string-2.
- ✓ n or N Ignores quotation marks around string-1 or string-2.
- ✓ : (colon) Truncates the longer of string-1 or string-2 to the shortest string length.

COMPLEV Function Example #1

COMPLEV Function with PROC SQL

```
PROC SQL ;
  SELECT Title, Rating, Length, Category,
        COMPLEV(Category,"Drama") AS COMPLEV_Number
   FROM mydata.Movies_with_Messy_Data
     WHERE Title NE ""
       ORDER BY Title ;
QUIT ;
```

COMPLEV Function Example #1



Title	Rating	Length	Category	COMPLEV_Number
Brave Heart	R	177	Action Adventure	16
Brave Heart	R	177	Acton Adventure	15
Casablanca	PG	103	Drama	0
Christmas Vacatiion	PG-13	97	Commedy	6
Christmas Vacation	PG-13	97	Comedy	6
Coming to America	R	116	Comedy	6
Dracula	R	130	Horror	5
Dressed to Kill	R	105	Drama Mysteries	10
Forrest Gump	PG-13	143	Drama	0
Forrest Gump	PG-13	142	Drama	0
Forrest Gumpp	PG13	143	Dramma	1
Ghost	PG-13	127	Drama Romance	8
Jaws	PG	125	Action Adventure	16
Jurassic Park	PG-13	127	Action	6
Lethal Weapon	R	110	Action Cops & Robber	20
Michael	PG-13	106	Drama	0
Micheal	PG-13	106	Drama	0
National Lampoon's Vacation	PG-13	98	Comedy	6
National Lampoons Vacation	PG-13	98	Comedy	6
Poltergeist	PG	115	Horror	5
Rocky	PG	120	Action Adventure	16
Rocky	PG	120	Action Adventure	16
Scarface	r	170	Action Cops & Robber	20
Silence of the Lambs	R	118	Drama Suspense	9
Star Wars	PG	124	Action Sci-Fi	13
The Hunt for Red October	GP	135	Action Adventure	16
The Terminator	R	108	Action Sci-Fi	13
The Wizard of Oz	G	101	Adventure	9
The Wizard of Ozz	g	102	Adventure	9
Titanic	PG-13	194	Dramma Romance	9

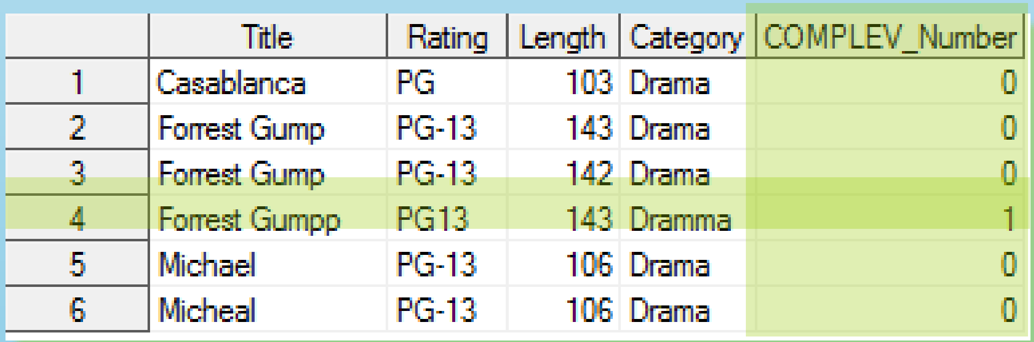
Drama 0

Dramma 1

COMPLEV Function Example #2

COMPLEV Function with PROC SQL

```
PROC SQL ;
  SELECT Title, Rating, Length, Category,
        COMPLEV(Category,"Drama") AS COMPLEV_Number
   FROM mydata.Movies_with_Messy_Data
    WHERE Title NE "" AND CALCULATED COMPLEV_Number LE 1
     ORDER BY Title ;
QUIT ;
```



	Title	Rating	Length	Category	COMPLEV_Number
1	Casablanca	PG	103	Drama	0
2	Forrest Gump	PG-13	143	Drama	0
3	Forrest Gump	PG-13	142	Drama	0
4	Forrest Gumpp	PG13	143	Dramma	1
5	Michael	PG-13	106	Drama	0
6	Micheal	PG-13	106	Drama	0

Fuzzy Matching Using the COMPGED Function

COMPGED Function

- The COMPGED function computes a Generalized Edit Distance (GED) score when comparing two text strings;
- The GED score acts as a measure of dissimilarity between two strings;
- The higher the GED score – the less likely the two strings match;
- Users should seek the lowest derived GED score for the greatest likelihood of a match (e.g., 0=Best match, 10, 20, 30, etc.).

COMPGED Function Arguments

The general syntax for the COMPGED function is:

COMPGED (string-1, string-2 <,cutoff-value> <,modifier>)

Optional Arguments:

Cutoff-value specifies a numeric variable, constant or expression.

Modifier specifies a value that alters the action of the COMPGED function. Valid modifier values are:

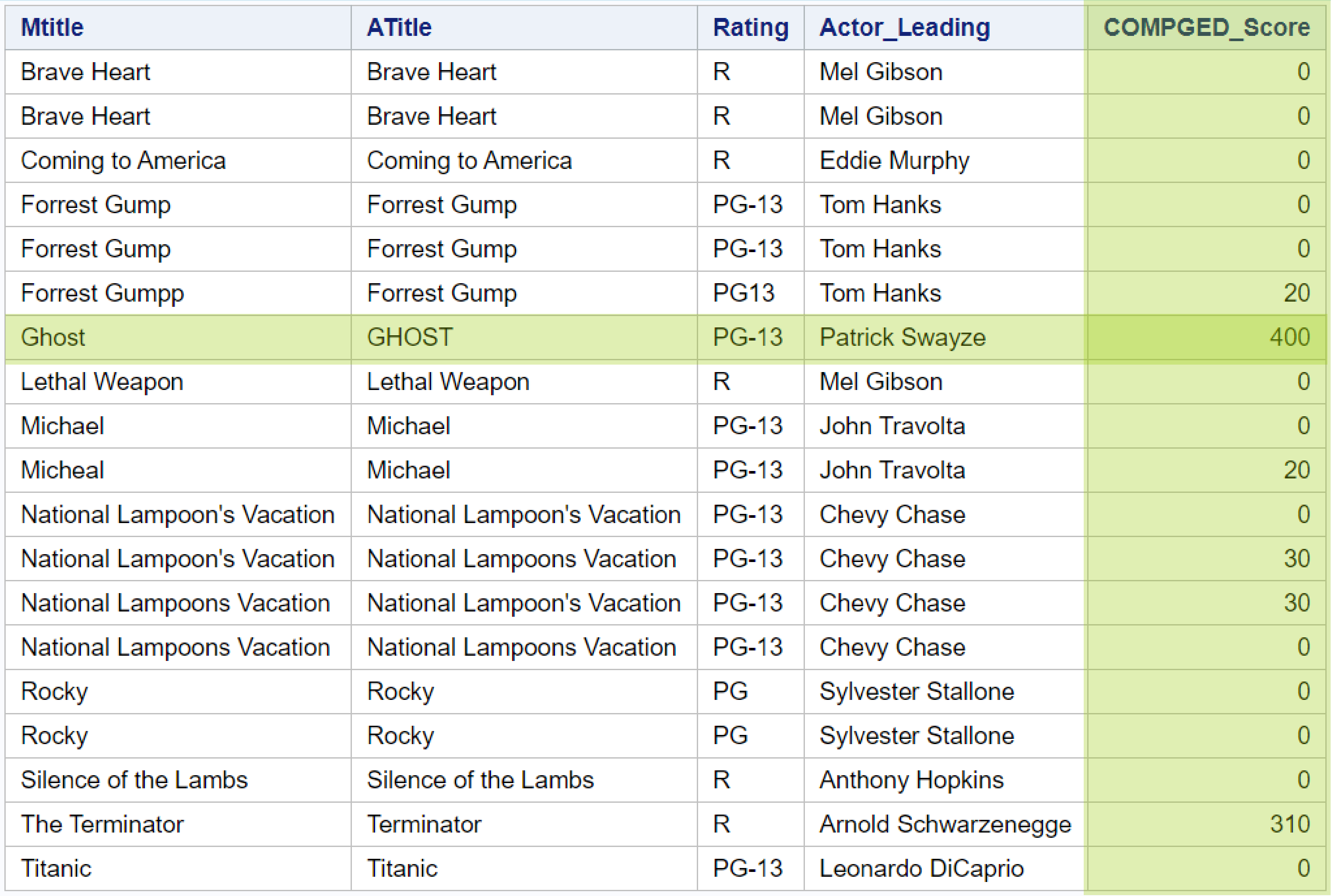
✓	i or I	Ignores the case (case insensitive) in string-1 and string-2.
✓	l or L	Removes leading blanks before comparing the values in string-1 or string-2.
✓	n or N	Ignores quotation marks around string-1 or string-2.
✓	: (colon)	Truncates the longer of string-1 or string-2 to the shortest string length.

COMPGED Function Example #1

COMPGED Function with PROC SQL

```
PROC SQL ;
  SELECT M.Title AS Mtitle, A.Title AS ATitle,
        Rating, Actor_Leading,
        COMPGED(M.Title,A.Title) AS COMPGED_Score
    FROM mydata.Movies_with_Messy_Data M,
        mydata.Actors_with_Messy_Data A
      WHERE M.Title NE "“ AND CALCULATED COMPGED_Score LE 400
        ORDER BY M.Title ;
QUIT ;
```

COMPGED Function Example #1



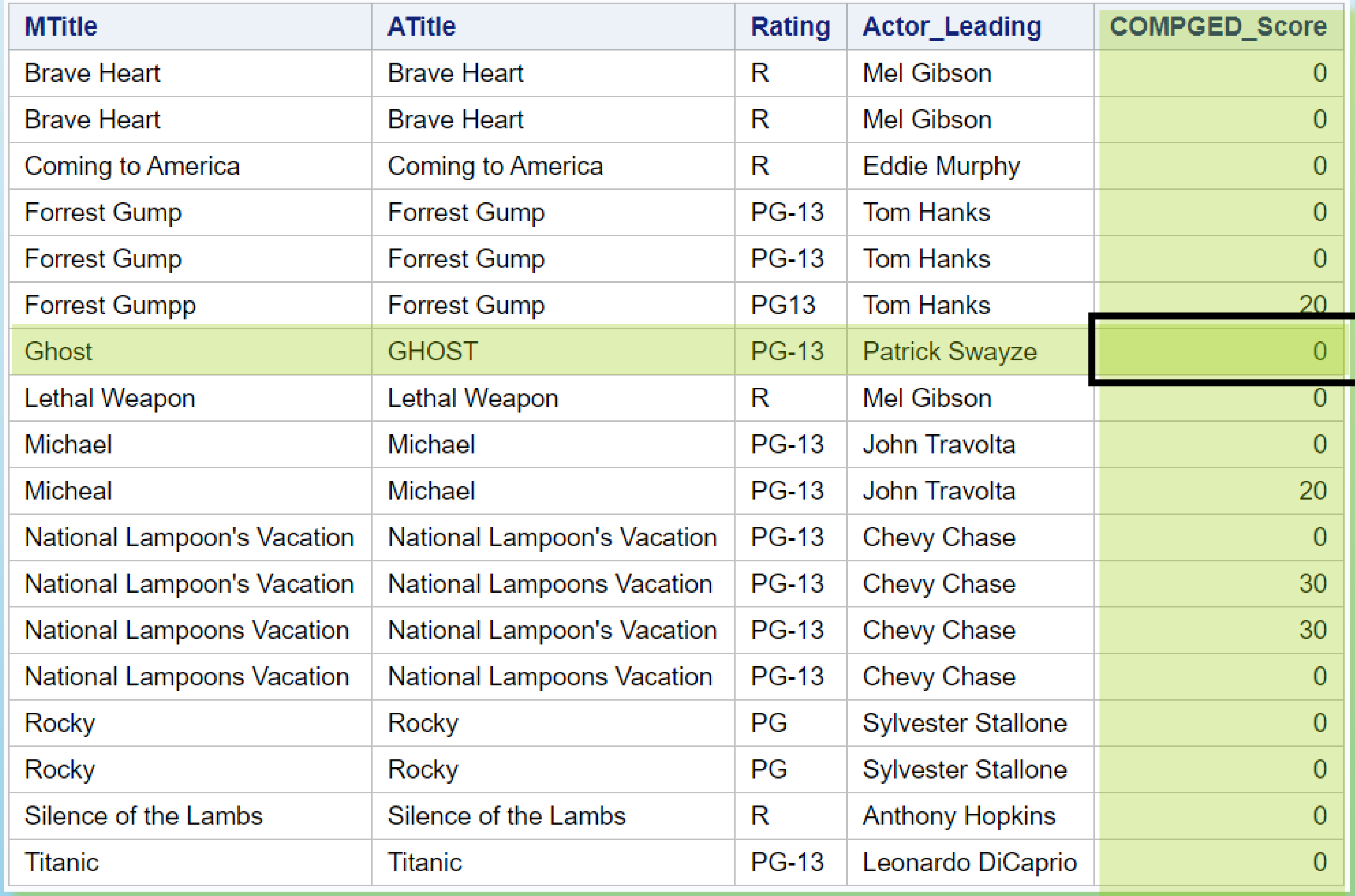
Mtitle	ATitle	Rating	Actor_Leading	COMPGED_Score
Brave Heart	Brave Heart	R	Mel Gibson	0
Brave Heart	Brave Heart	R	Mel Gibson	0
Coming to America	Coming to America	R	Eddie Murphy	0
Forrest Gump	Forrest Gump	PG-13	Tom Hanks	0
Forrest Gump	Forrest Gump	PG-13	Tom Hanks	0
Forrest Gumpp	Forrest Gump	PG13	Tom Hanks	20
Ghost	GHOST	PG-13	Patrick Swayze	400
Lethal Weapon	Lethal Weapon	R	Mel Gibson	0
Michael	Michael	PG-13	John Travolta	0
Micheal	Michael	PG-13	John Travolta	20
National Lampoon's Vacation	National Lampoon's Vacation	PG-13	Chevy Chase	0
National Lampoon's Vacation	National Lampoons Vacation	PG-13	Chevy Chase	30
National Lampoons Vacation	National Lampoon's Vacation	PG-13	Chevy Chase	30
National Lampoons Vacation	National Lampoons Vacation	PG-13	Chevy Chase	0
Rocky	Rocky	PG	Sylvester Stallone	0
Rocky	Rocky	PG	Sylvester Stallone	0
Silence of the Lambs	Silence of the Lambs	R	Anthony Hopkins	0
The Terminator	Terminator	R	Arnold Schwarzenegge	310
Titanic	Titanic	PG-13	Leonardo DiCaprio	0

COMPGED Function Example #2

COMPGED Function with PROC SQL

```
PROC SQL ;
  SELECT M.Title AS MTitle, A.Title AS ATitle,
         Rating, Actor_Leading,
         COMPGED(M.Title,A.Title,'I') AS COMPGED_Score
    FROM mydata.Movies_with_Messy_Data M,
         mydata.Actors_with_Messy_Data A
      WHERE M.Title NE "" AND CALCULATED COMPGED_Score LE 30
        ORDER BY M.Title ;
QUIT ;
```

COMPGED Function Example #2



MTitle	ATitle	Rating	Actor_Leading	COMPGED_Score
Brave Heart	Brave Heart	R	Mel Gibson	0
Brave Heart	Brave Heart	R	Mel Gibson	0
Coming to America	Coming to America	R	Eddie Murphy	0
Forrest Gump	Forrest Gump	PG-13	Tom Hanks	0
Forrest Gump	Forrest Gump	PG-13	Tom Hanks	0
Forrest Gumpp	Forrest Gump	PG13	Tom Hanks	20
Ghost	GHOST	PG-13	Patrick Swayze	0
Lethal Weapon	Lethal Weapon	R	Mel Gibson	0
Michael	Michael	PG-13	John Travolta	0
Micheal	Michael	PG-13	John Travolta	20
National Lampoon's Vacation	National Lampoon's Vacation	PG-13	Chevy Chase	0
National Lampoon's Vacation	National Lampoons Vacation	PG-13	Chevy Chase	30
National Lampoons Vacation	National Lampoon's Vacation	PG-13	Chevy Chase	30
National Lampoons Vacation	National Lampoons Vacation	PG-13	Chevy Chase	0
Rocky	Rocky	PG	Sylvester Stallone	0
Rocky	Rocky	PG	Sylvester Stallone	0
Silence of the Lambs	Silence of the Lambs	R	Anthony Hopkins	0
Titanic	Titanic	PG-13	Leonardo DiCaprio	0

PharmaSUG
NORTH CAROLINA • OCT. 21, 2022

Summary of Fuzzy Matching Techniques

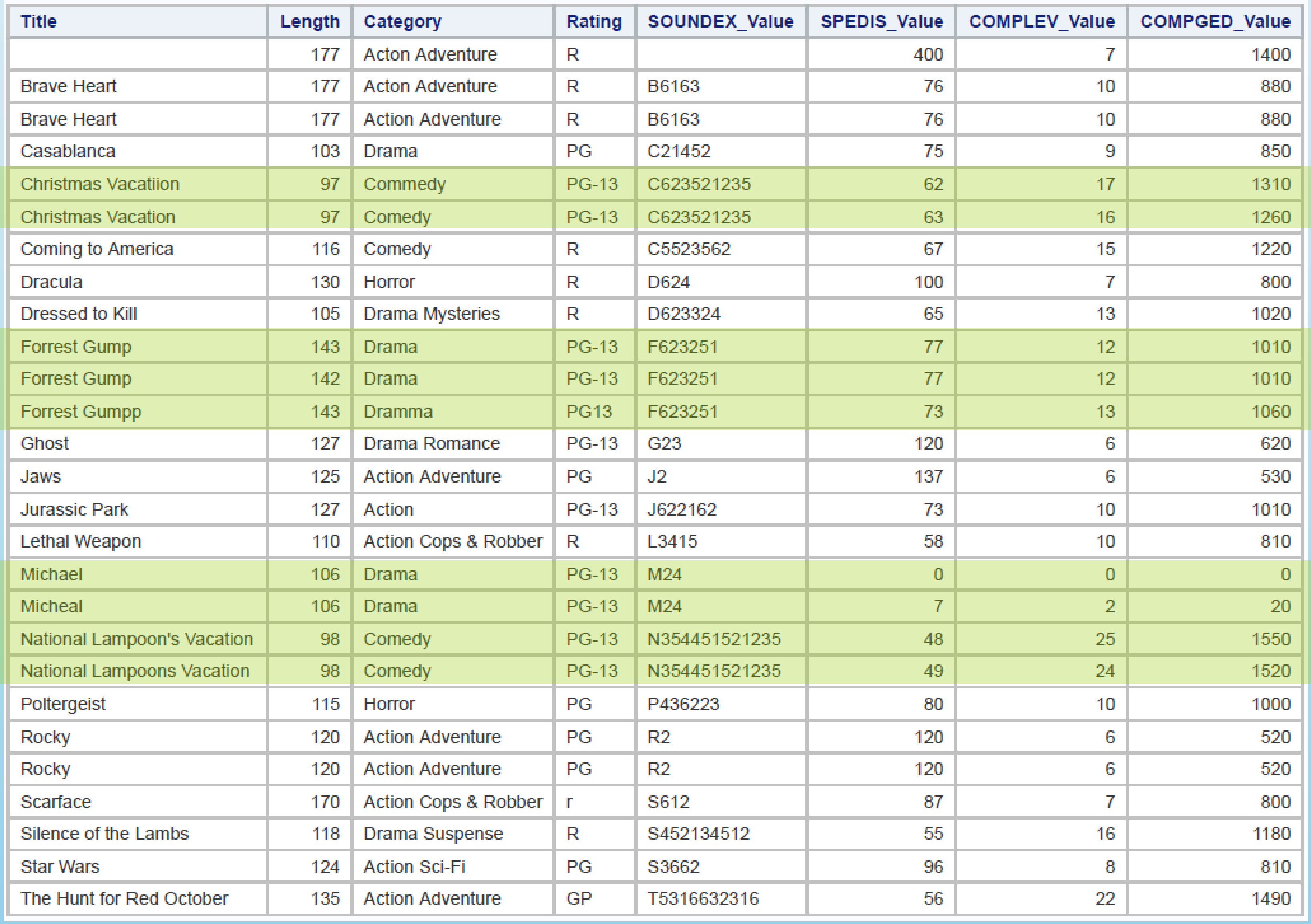
Summary of Fuzzy Matching Techniques

Comparison of the Different Techniques

```
PROC SQL ;
 SELECT Title
      , Length
      , Category
      , Rating
      , SOUNDEX(Title)          AS SOUNDEX_Value
      , SPEDIS(Title,'Michael')  AS SPEDIS_Value
      , COMPLEV(Title,'Michael') AS COMPLEV_Value
      , COMPGED(Title,'Michael') AS COMPGED_Value
  FROM MYDATA.Movies_with_Messy_Data
   WHERE CALCULATED SPEDIS_Value   GE 0
     AND CALCULATED COMPLEV_Value  GE 0
     AND CALCULATED COMPGED_Value  GE 0
    ORDER BY Title ;
QUIT ;
```

PharmaSUG
NORTH CAROLINA • OCT. 21, 2022

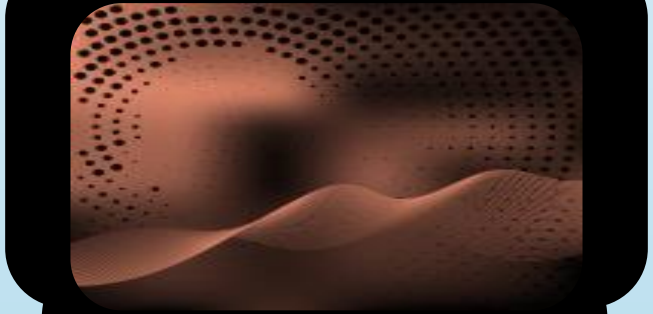
Fuzzy Matching Techniques on Title



Title	Length	Category	Rating	SOUNDEX_Value	SPEDIS_Value	COMPLEV_Value	COMPGED_Value
	177	Acton Adventure	R		400	7	1400
Brave Heart	177	Acton Adventure	R	B6163	76	10	880
Brave Heart	177	Action Adventure	R	B6163	76	10	880
Casablanca	103	Drama	PG	C21452	75	9	850
Christmas Vacatiion	97	Commedy	PG-13	C623521235	62	17	1310
Christmas Vacation	97	Comedy	PG-13	C623521235	63	16	1260
Coming to America	116	Comedy	R	C5523562	67	15	1220
Dracula	130	Horror	R	D624	100	7	800
Dressed to Kill	105	Drama Mysteries	R	D623324	65	13	1020
Forrest Gump	143	Drama	PG-13	F623251	77	12	1010
Forrest Gump	142	Drama	PG-13	F623251	77	12	1010
Forrest Gumpp	143	Dramma	PG13	F623251	73	13	1060
Ghost	127	Drama Romance	PG-13	G23	120	6	620
Jaws	125	Action Adventure	PG	J2	137	6	530
Jurassic Park	127	Action	PG-13	J622162	73	10	1010
Lethal Weapon	110	Action Cops & Robber	R	L3415	58	10	810
Michael	106	Drama	PG-13	M24	0	0	0
Micheal	106	Drama	PG-13	M24	7	2	20
National Lampoon's Vacation	98	Comedy	PG-13	N354451521235	48	25	1550
National Lampoons Vacation	98	Comedy	PG-13	N354451521235	49	24	1520
Poltergeist	115	Horror	PG	P436223	80	10	1000
Rocky	120	Action Adventure	PG	R2	120	6	520
Rocky	120	Action Adventure	PG	R2	120	6	520
Scarface	170	Action Cops & Robber	r	S612	87	7	800
Silence of the Lambs	118	Drama Suspense	R	S452134512	55	16	1180
Star Wars	124	Action Sci-Fi	PG	S3662	96	8	810
The Hunt for Red October	135	Action Adventure	GP	T5316632316	56	22	1490

Conclusion

The Fuzzy Matching Process Explained



Fuzzy Matching Programming Techniques

Fuzzy Matching Programming Examples

Please read our paper for detailed fuzzy matching techniques using SAS.

Thank you for attending!

Questions?

a presentation by

Kirk Paul Lafler
KirkLafler@cs.com
@sasNerd

Stephen B. Sloan
Stephen.B.Sloan@accenture.com
917-375-2937
https://www.linkedin.com/in/stephen-sloan-1949423/
@StephenBaileySl