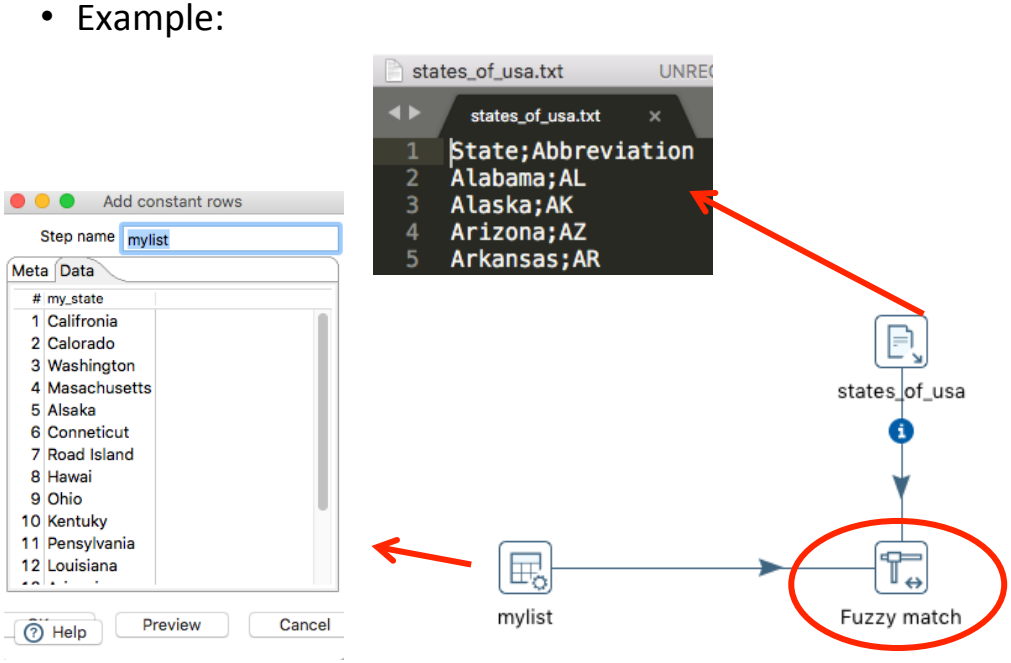
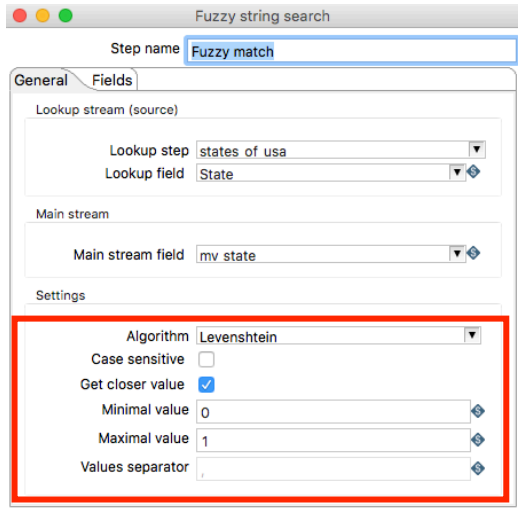
Dealing with non-exact matches

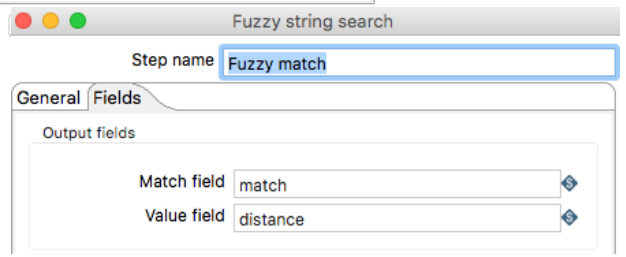
- Common (manual) error: typos, upper/lower case → fuzzy string searching → Fuzzy match step
- Example:



stream of data with incorrect states
Data Grid step

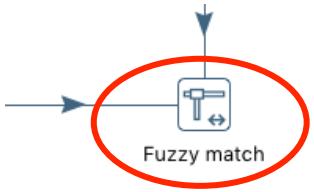


Details of the fuzzy search

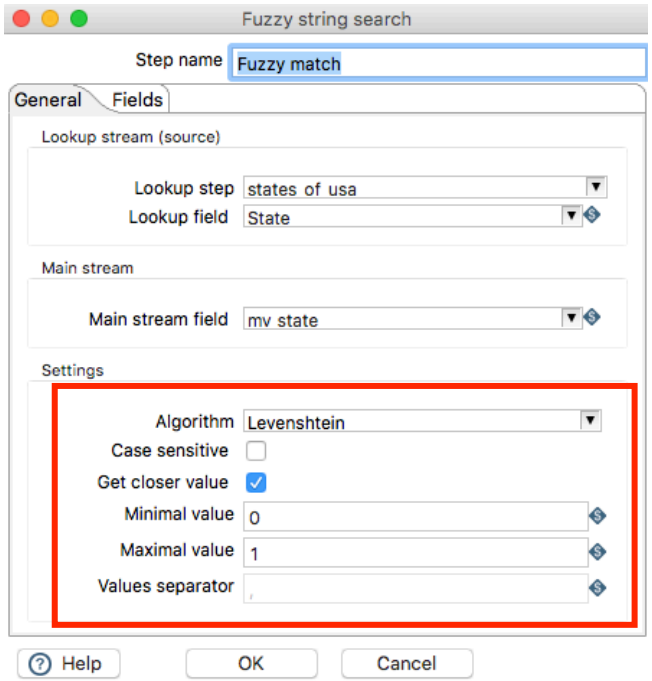


Dealing with non-exact matches

- Common (manual) error: typos, upper/lower case → fuzzy string searching → Fuzzy match step
- Example:



Details of the fuzzy search

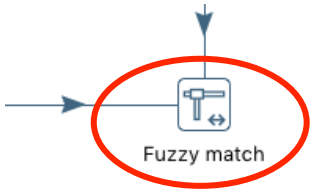


Several matching algorithms:

- Algorithms based on a metric distance, the comparison is based on how the terms are spelt (e.g. Levenshtein: This algorithm calculates the distance between two strings as the number of edit steps (character insertion or deletion, or replacements) needed to get from one string to another
- Phonetic algorithms, the comparison is based on how the terms sound when read in English

Dealing with non-exact matches

- Common (manual) error: typos, upper/lower case → fuzzy string searching → Fuzzy match step
- Example:



Examine preview data

Rows of step: Fuzzy match (17 rows)

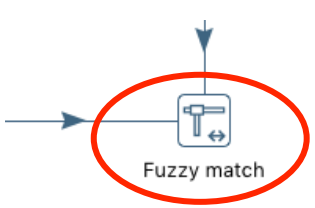
#	my_state	match	distance
1	Califronia	<null>	<null>
2	Calorado	Colorado	1
3	Washington	Washington	0
4	Masachusetts	Massachusetts	1
5	Alsaka	<null>	<null>
6	Conneticut	Connecticut	1
7	Road Island	<null>	<null>
8	Hawai	Hawaii	1

Close

- We used Levenshtein algoritm: This algorithm calculates the distance between two strings as the number of edit steps (character insertion or deletion, or replacements) needed to get from one string to another
- Try the Damerau-Levenshtein algorithm, which is similar to Levenshtein but adds the transposition operation

Deduplicating non-exact matches

- Example: Hawaii, Hawai, and Howaii → We only want a single state: Hawaii



sort

Examine preview data

Rows of step: Fuzzy match (17 rows)

#	my_state	match	distance
16	California	California	0
2	Calorado	Colorado	1
6	Conneticut	Connecticut	1
8	Hawai	Hawaii	1
14	Hawaii	Hawaii	0
17	Howaii	Hawaii	1
10	Kentuky	Kentucky	1

Close

Deduplicating non-exact matches

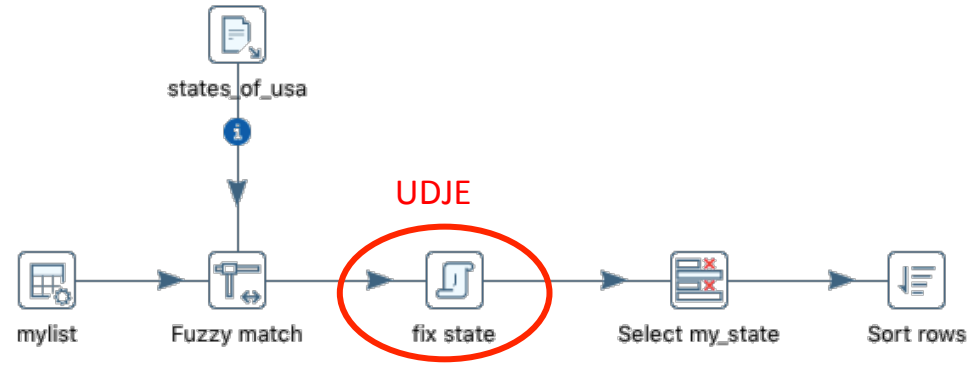
- Example: Hawaii, Hawai, and Howaii → We only want a single state: Hawaii

Examine preview data

Rows of step: fix state (17 rows)

#	my_state	match	distance
5	Alaska	Alaska	1
13	Arizona	Arizona	1
1	California	California	1
16	California	California	0
2	Colorado	Colorado	1
6	Connecticut	Connecticut	1
8	Hawaii	Hawaii	1
14	Hawaii	Hawaii	0
17	Hawaii	Hawaii	1
10	Kentucky	Kentucky	1

Close



User Defined Java Expression

Step name: fix state

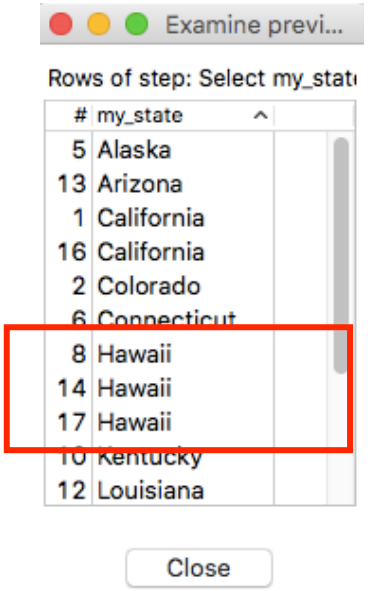
Fields:

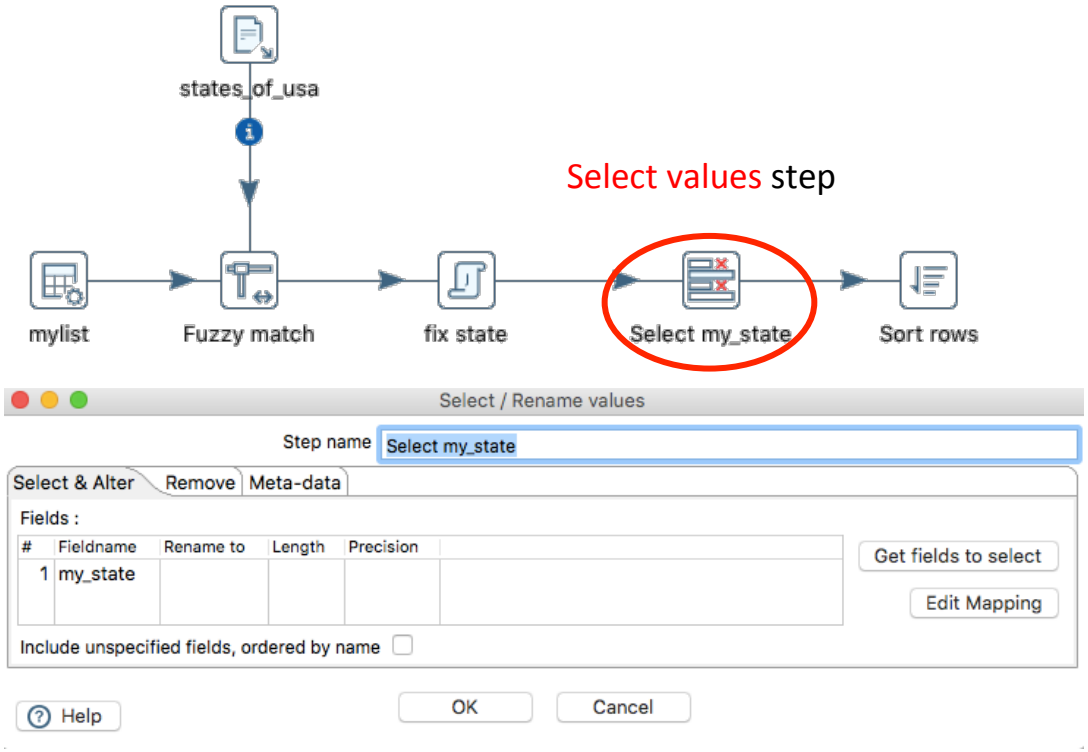
#	New field	Java expression	Value type	Length	Precision	Replace value
1	my_state	(match== null)?my_state:match	String			my_state
2						

Help OK Cancel

Deduplicating non-exact matches

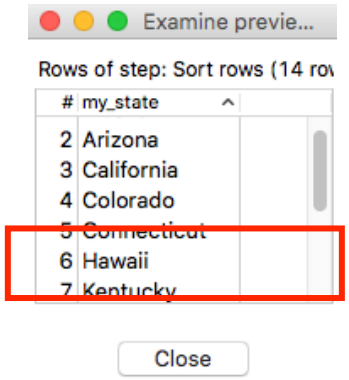
- Example: Hawaii, Hawai, and Howaii → We only want a single state: Hawaii

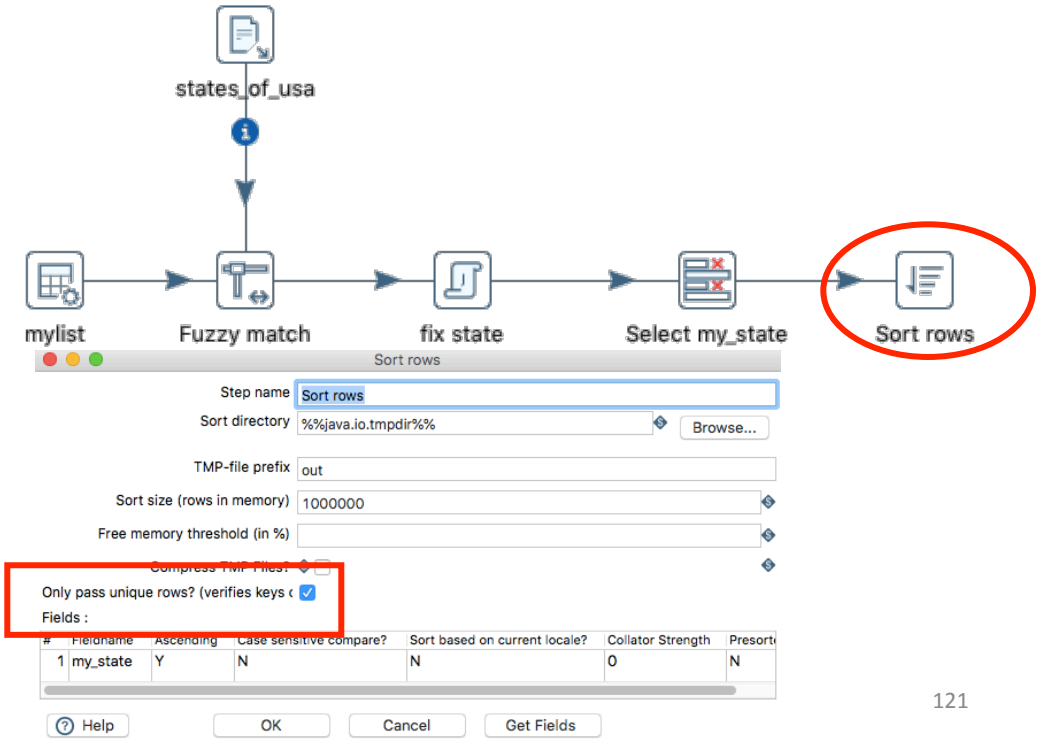




Deduplicating non-exact matches

- Example: Hawaii, Hawai, and Howaii → We only want a single state: Hawaii





Remove duplicates
We eliminated duplicated rows with the Sort rows step. If the data were sorted by state, we could have used a Unique rows step instead.