How is the next work predicted ?

Whenever you give the algorithm a set of words, it asks this simple question – "Whenever people have said the given words, what the is next word they have used the most."
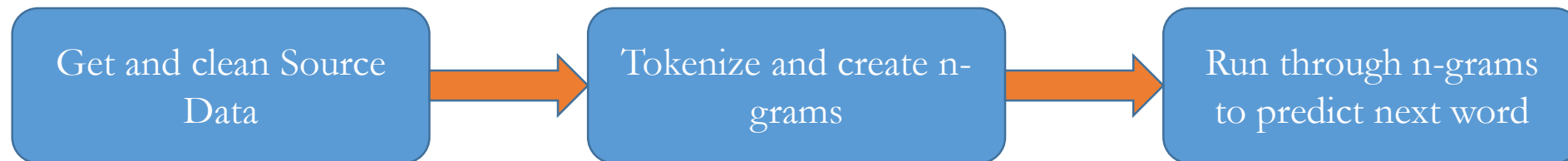
On what data does the machine learn?

The algorithm uses data scrapped and dumped from twitter, various blogs and news websites. The data is provided by [SwiftKey Corp.](#)

Does this Algorithm use pre-set statistical regression techniques?

No. The algorithm is completely custom built by Logic. No statistical methods are used.

Can you explain the entire algorithm in a snapshot?

| Get and clean Source Data | → | Tokenize and create n-grams | → | Run through n-grams to predict next word |

Tokenization and n-grams

"The weather is really good today."

If word 1 is "weather"
and word 2 is "is"
and word 3 is "really"
then predict word 4 as "good"

Four Grams
- The weather is really
- Weather is really good
- Is Really Good today

Three Grams
- The weather is
- Weather is really
- Is Really Good
- Really good today

Two Grams (first four)
- The weather
- Weather is
- Is really
- Really good

Sample 4-gram frequency table when the user entered "I am going to"

Word 1	Word 2	Word 3	Word 4	Frequency
I am	Going	To	Be	1190
I am	Going	To	Bed	1000
I am	Going	To	Sleep	800
I am	Going	To	The	200

Prediction Algorithm