

# Challenge 4: Crop-Specific and Crop-Independent Questions

Word Frequency Analysis: Visualizing Bi- and Tri-Grams of Questions by Kenyan Farmers in the Producers Direct Dataset of WeFarm SMS

# Producers Direct Dataset

Original CSV file of 20,304,843 rows & 24 columns included 5,865,819 unique questions asked by 1,026,367 farmers

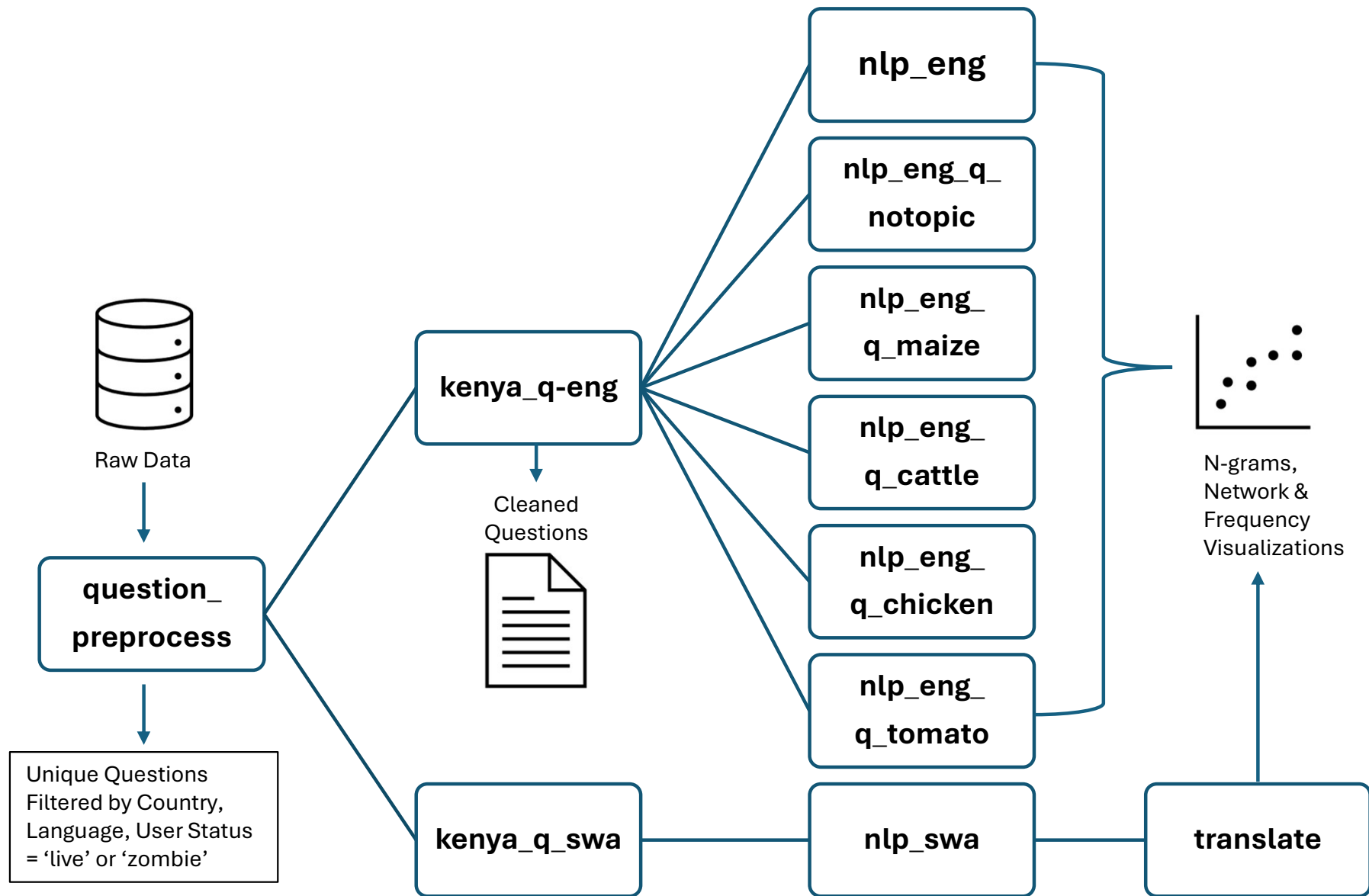
Time to answer questions: mean=4.8 days, s.d. = 4.3 days

## Unique Question Counts by:

| Language |       | Country  |       | Topic (148) |       | User Status |       |
|----------|-------|----------|-------|-------------|-------|-------------|-------|
| eng      | 50.1% | Kenya    | 41.3% | null        | 28.5% | live        | 65.3% |
| swa      | 37.9% | Uganda   | 33.7% | maize       | 9.1%  | zombie      | 19.1% |
| nyn      | 7.4%  | Tanzania | 25.0% | chicken     | 7.9%  | blocked     | 8.2%  |
| lug      | 4.5%  | Gambia   | --    | cattle      | 5.5%  | destroyed   | 7.4%  |
|          |       |          |       | tomato      | 5.6%  |             |       |
|          |       |          |       | cranberry   | --    |             |       |

# Order of Jupyter Notebooks

*Notebooks contain more detail about the steps, inputs, outputs, and dependencies*



# Challenges of Translating Swahili into English

GoogleTranslate has a 5,000 character limit per call, so it's impractical to translate the full text of > 2 mm questions

A possible workaround is to extract and translate the most frequent combination of words in the Swahili questions to derive meaningful insights....

But Swahili is an under-resourced language in Natural Language Processing:

- Commonly used Python packages such as SpaCy, NLTK, or Gensim do not have inherent Swahili support
- It is an agglutinative language: prefixes, roots, and suffixes are combined into one word. It also has complex noun class structures, that affect verb agreement. These can lead to ineffective lemmatization.

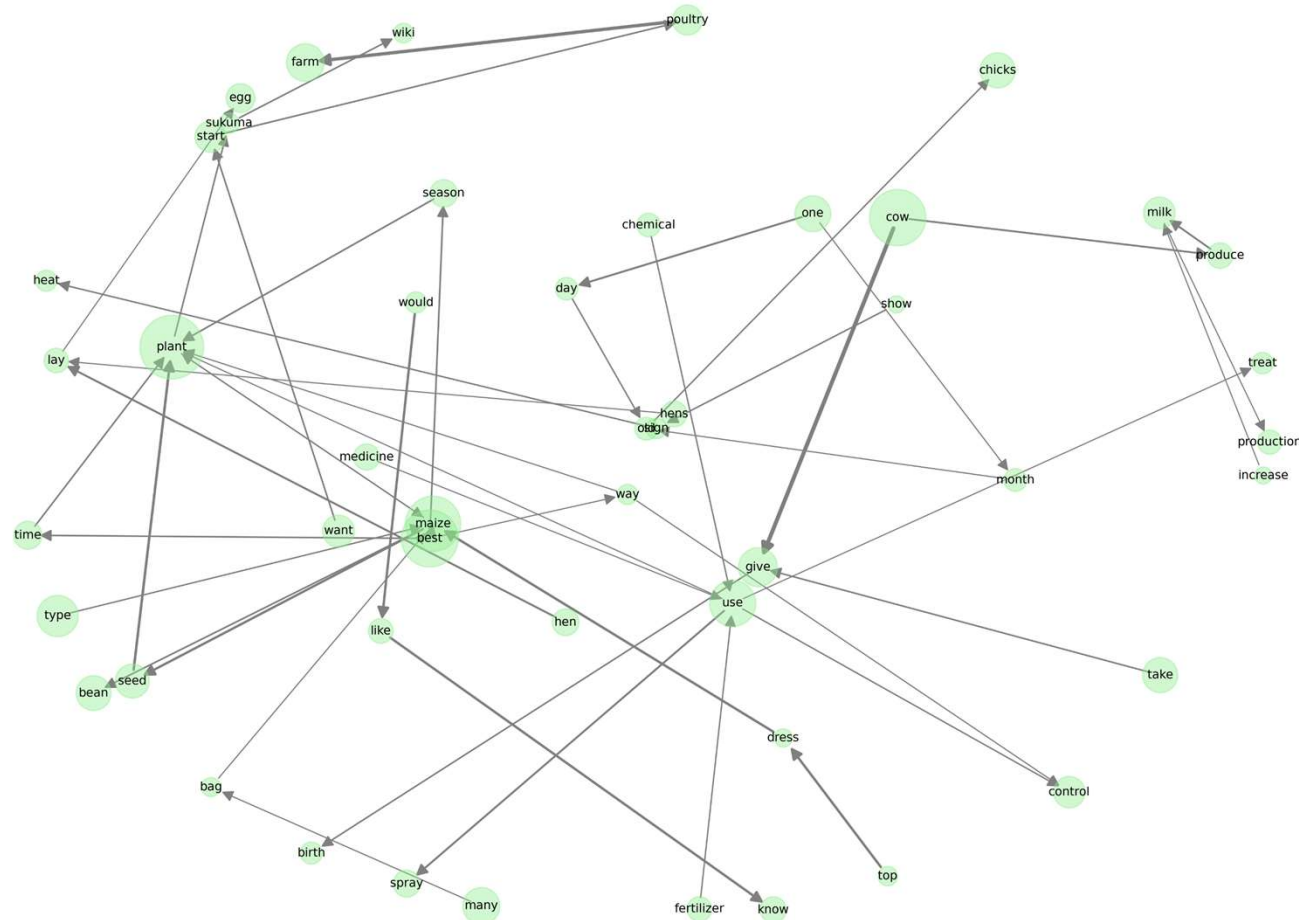
An attempt was made to generate and translate n-grams using Swahili word lists sourced from Menderley, but the translated n-grams was not particularly informative. A more robust analysis requires an agricultural corpus on rural farming in Africa, and custom lists of words and lemma dictionary.

## **With These Challenges, Swahili Trigrams Translated in English – All Topics – Were More Informative Than Bi- or Quadgrams.**

1. what\_medicine\_to\_use
2. what\_medicine\_should\_i\_use
3. what\_good\_medicine
4. what\_seed\_india
5. can\_get\_me
6. what\_good\_medicine
7. what\_drug\_to\_use
8. naeza\_pata\_mpi (I can get water)
9. get\_the\_seed
10. where\_is\_the\_problem?
11. to\_lay\_the\_egg
12. how\_long\_take
13. medicine\_can\_i
14. I\_want\_to\_fuga\_ku (I want to raise / domesticate)
15. medicine\_can\_you

# Tri-grams of English Questions Without a Topic, Still Focused on Farming: e.g. Best Practices on Birthing Calves, Planting Maize and Starting a Poultry Farm

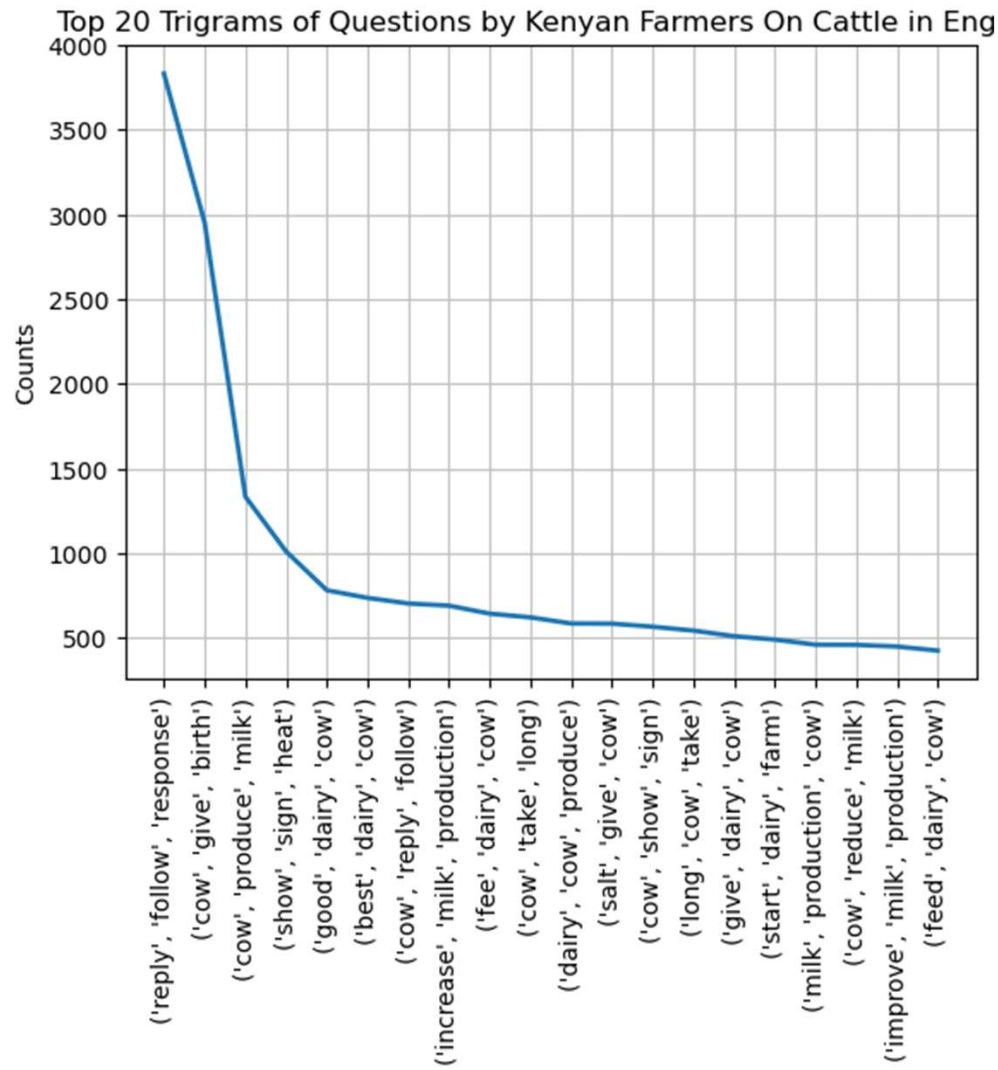
Top 30 Trigrams of Kenyan Farmer English Questions Without Topic (WeFarm, 2022)  
(circle size represents word frequency, arrow width represents trigram frequency)



Source: WeFarm 2022 SMS Platform

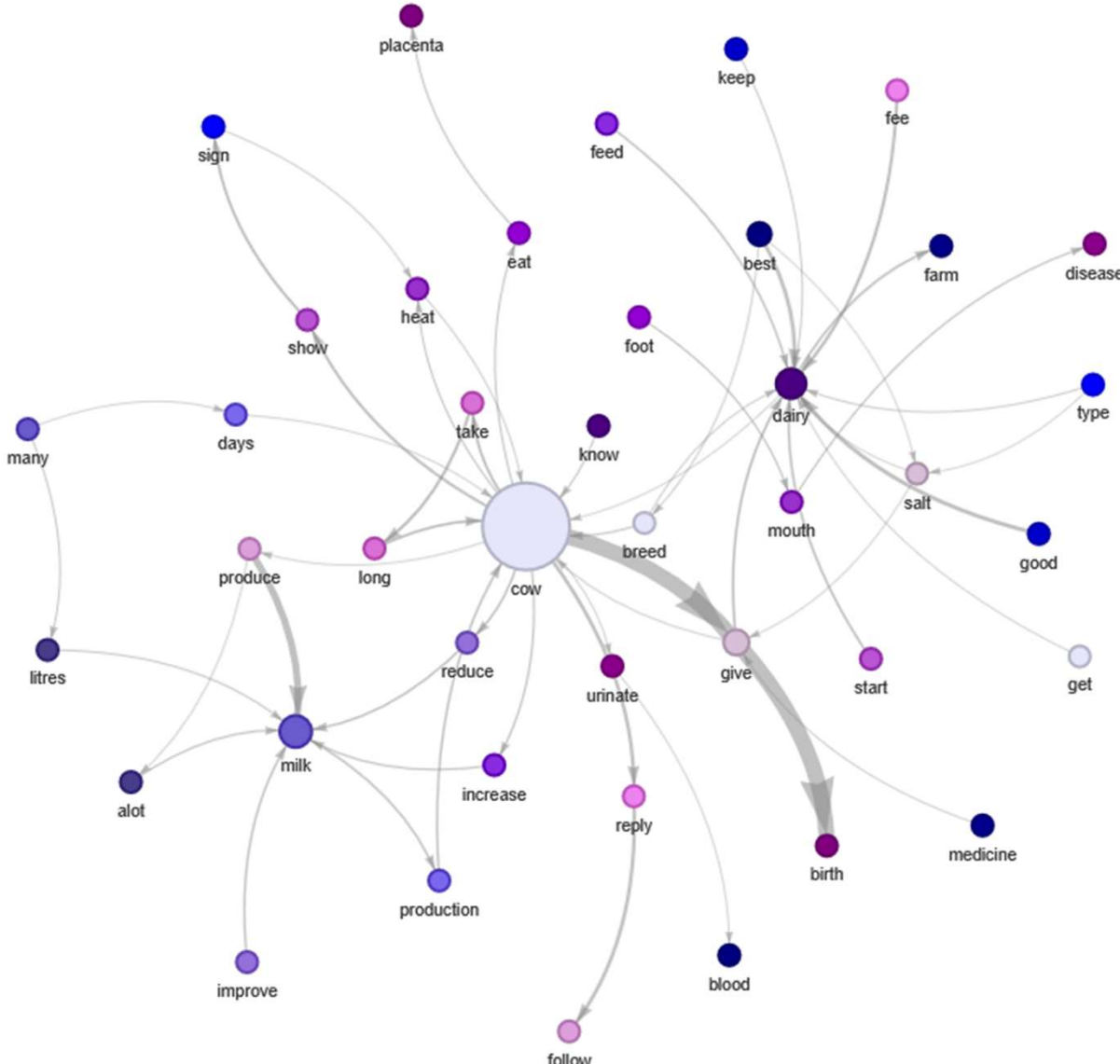
DataKind Producers Direct Challenge -  
December 2025, Beatrice Liu

# Cattle Kenyan Farmers Focused on Giving Birth, Milk Production, and Health



### Network Graph: Top 40 Trigrams from Kenyan Farmers Questions on Cattle in English

Word circle size = word frequency, arrow width = trigram frequency, Source: WeFarm 2022 SMS Platform

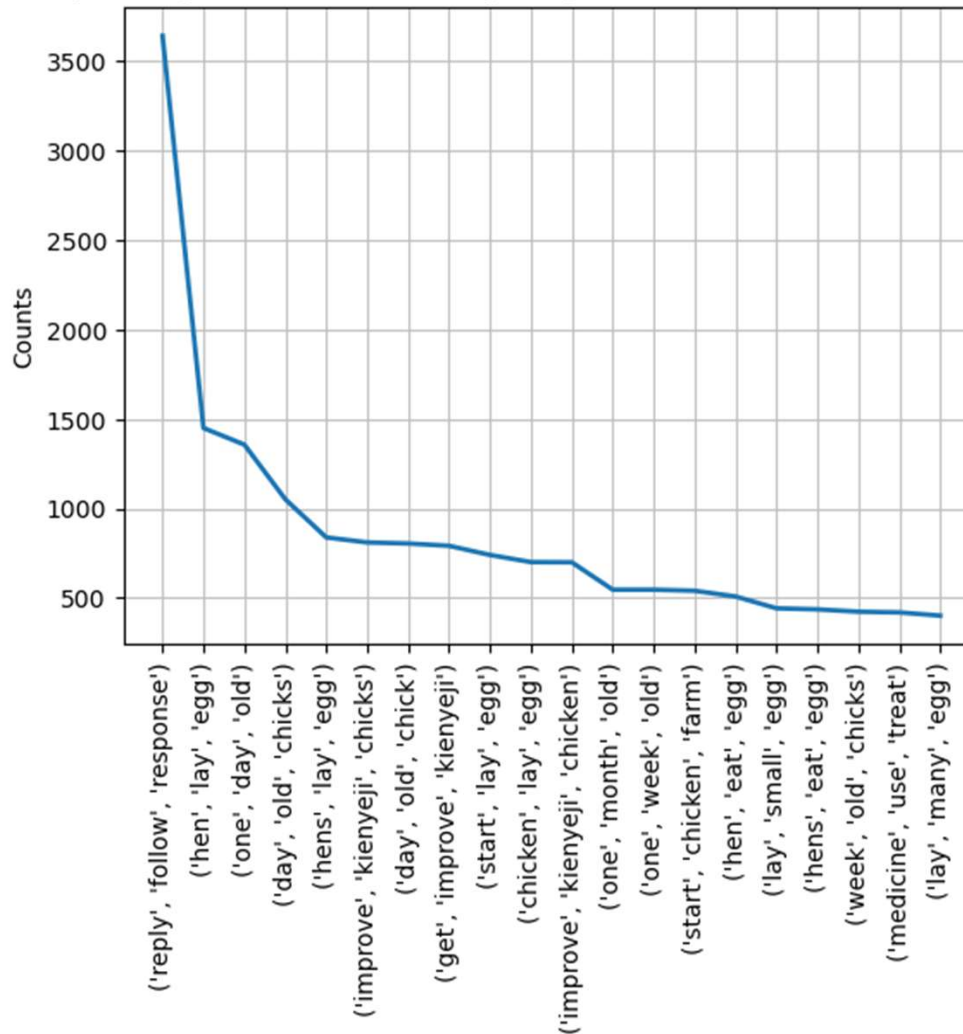


DataKind Producers Direct Challenge -  
December 2025, Beatrice Liu



# Kenyan Chicken Farmers Asked About Young Chicks, Laying Eggs, and Medicines

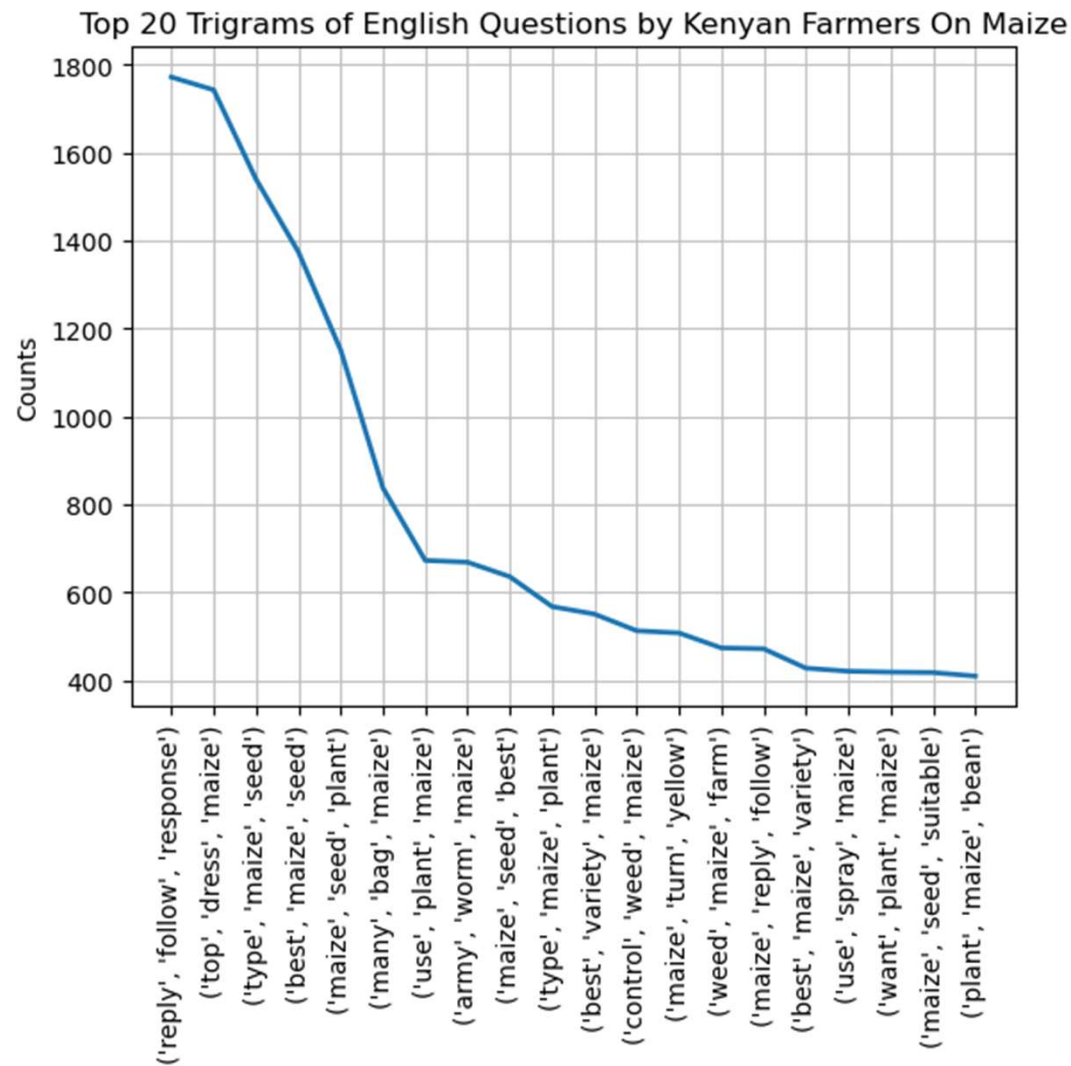
Top 20 Trigrams of Questions by Kenyan Farmers On Chicken in English



Word circle size = word frequency, arrow width = trigram frequency, Source: WeFarm 2022 SMS Platform

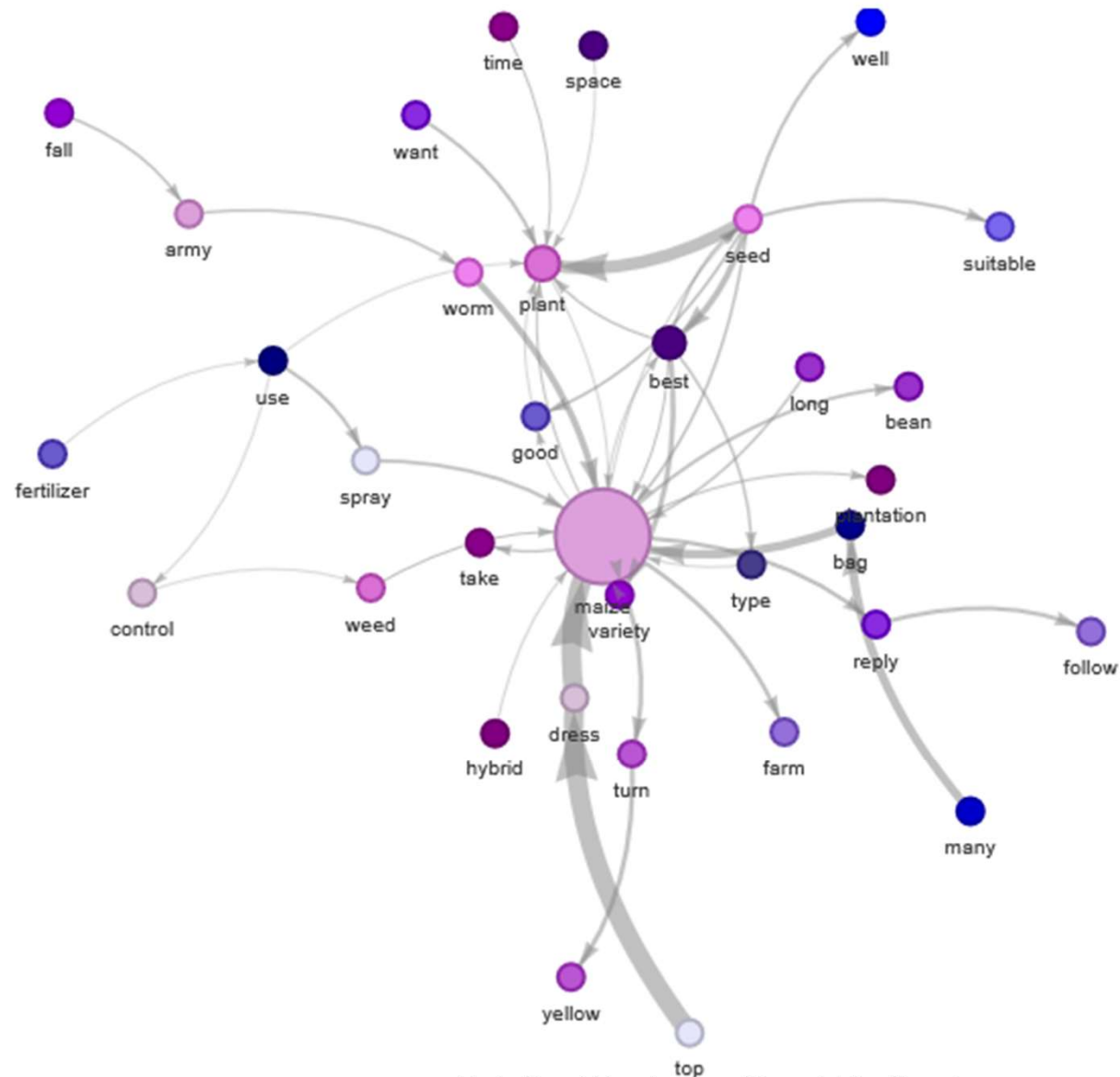


# Maize Farmers in Kenya Asked About the Best Fertilizer and Weed Control, Best Seeds and Planting



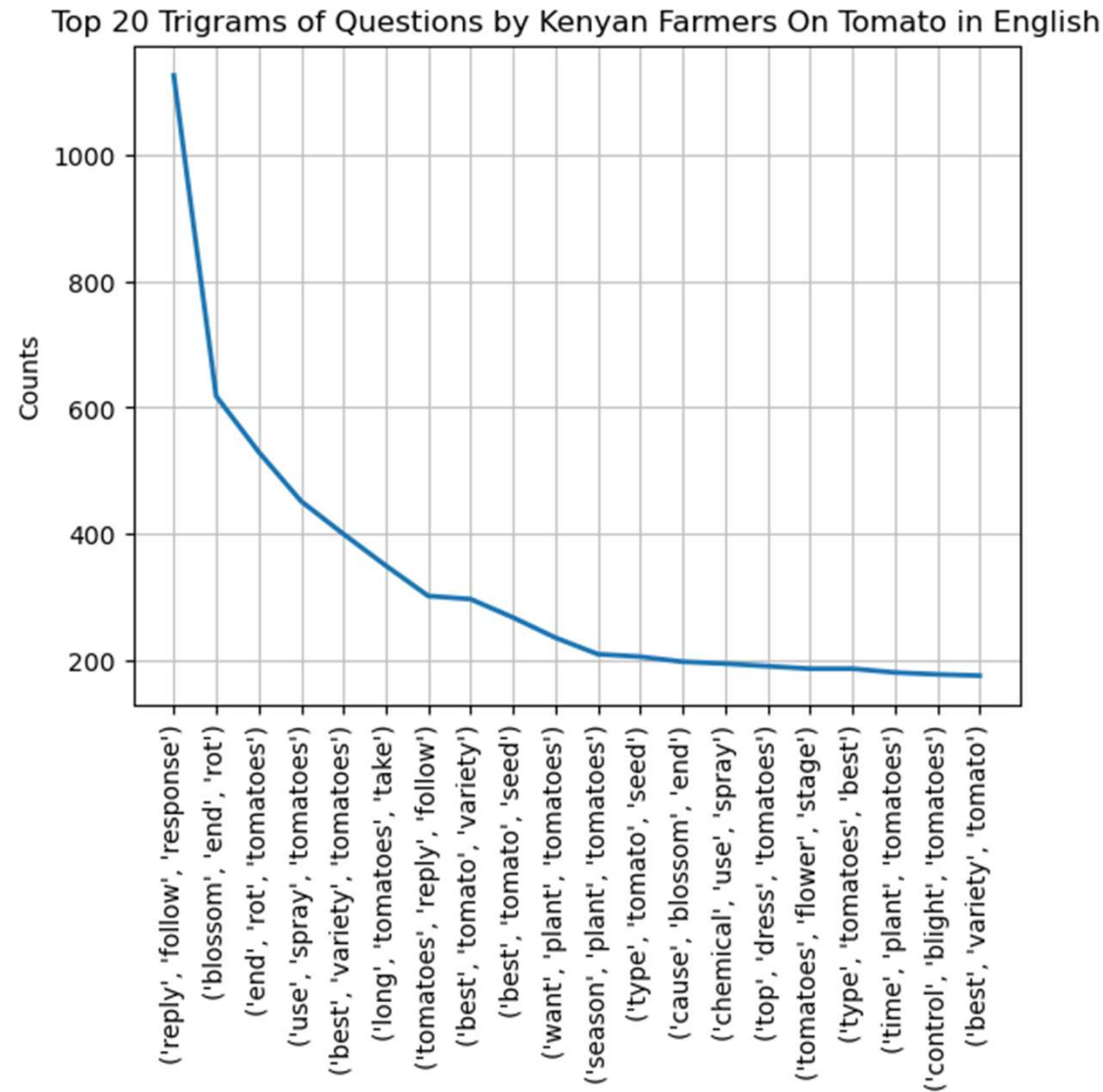
# Network Graph: Top 40 Trigrams from Kenyan Farmer English Questions on Maize

Word circle size = word frequency, arrow width = trigram frequency, Source: WeFarm 2022 SMS Platform



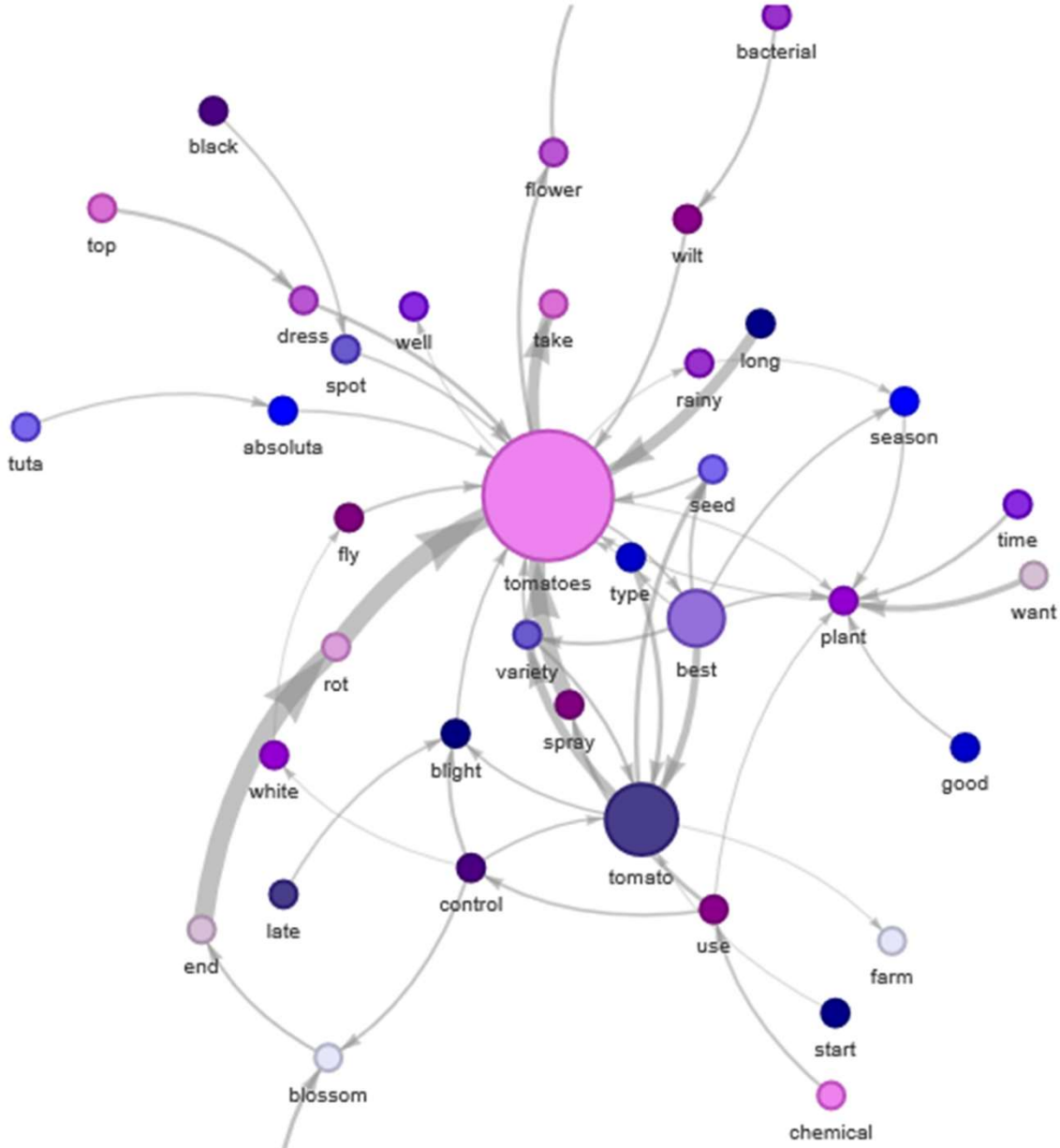
DataKind Producers Direct Challenge -  
December 2025, Beatrice Liu

# Tomato Kenyan Farmers Asked About Planting, Best Seeds, and Keeping Tomatoes Disease-Free



### Network Graph: Top 40 Trigrams from Kenyan Farmers Questions on Tomato in English

Word circle size = word frequency, arrow width = trigram frequency, Source: WeFarm 2022 SMS Platform



December 2025, Beatrice Liu

# Additional Resources

- Data files and visualizations created by these notebooks in Google Drive:  
[https://drive.google.com/drive/folders/1tpwqTqoFfZCWvDvncJjaSbzzua0Y6Q\\_i?usp=sharing](https://drive.google.com/drive/folders/1tpwqTqoFfZCWvDvncJjaSbzzua0Y6Q_i?usp=sharing)
- Common Swahili Stop-Words;  
<https://data.mendeley.com/datasets/mmf4hnsn2n/1>
- Swahili Agriculture Corpus: KILIMO:  
<https://data.mendeley.com/datasets/d4yhn5b9n6/2/files/cfd0108d-863d-460d-b52c-a51ce4101f79>
- Swahili Verb Conjugation Dataset for lemmatization:  
<https://data.mendeley.com/datasets/rvt89578g5/1>