



Food and Agriculture
Organization of the
United Nations

SUSTAINABLE
DEVELOPMENT
GOALS



AFRICAN COMMISSION ON AGRICULTURAL STATISTICS

28TH SESSION

4–8 December 2023
Johannesburg (South Africa)

AFCAS 28
LEVERAGING
DATA & STATISTICS
FOR AGRIFOOD
SYSTEMS
TRANSFORMATION
IN AFRICA

AGENDA ITEM 10:

NEW DEVELOPMENTS IN THE USE OF ALTERNATIVE DATA SOURCES FOR AGRICULTURAL STATISTICS



AFRICAN
COMMISSION ON
**AGRICULTURAL
STATISTICS**

Data-Driven Strategies: FAO's Data Lab Tools for Filling Data Gaps and Obtaining Timely Insights

Presenter: Christian Mongeau (FAO Statistics Division)





Outline

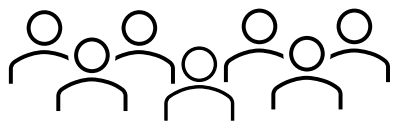
- Introduction
- FAO's Data Lab
- Non-conventional sources
- Data Science
- Data Gaps
- Expanding the Agriculture Production Dataset
- Deriving insights from news articles
- Collecting information on Food Loss and Waste
- Conclusions



FAO's Data Lab

The FAO Data Lab was created at the end of 2019:

- In response to the need to have **timely** information to support decisions
- To use new methods and technologies to extract data from **unstructured sources**
- To find solutions to issues related to the **crisis of traditional data collection systems**



The team:
Seven data scientists



Early Warning
Systems

Main areas of work



Analytics on
Policy
Documents



Hand-in-Hand
Initiative/FAOSTAT



Data Lab's main areas of work and products



Early Warning Systems

- **Daily Food Prices Monitor:** integration between **Nowcasting** Food Prices and **Daily Food Prices Acceleration Monitor**.
- **Banking Sector Monitor:** Banking Sentiment Indicator + Financial Indicators.
- **Automatic Tweet/Article classification** into several topics:
 - **Topic Explorer Dashboard:** sentiment analysis by topic.
 - Data Lab Trends: **Free Search Engine** over 28 million tweet/articles.



Analytics on Policy Documents

- Food Losses and Waste - **FLW database**
 - Standalone application
- **Technology-assisted review** to extract **automatically** metadata from **FAOLEX (legal/policy) documents**.
- Food System Summit - **Pathways Analysis:**
 - Pipeline integrated
 - Dashboard
- **MetroPolicy**
 - Standalone application



Hand-in-Hand Initiative

- Filling in **data gaps:**
 - Data scraped at national and subnational level.
- Vulnerability mapping:
 - Land cover mapping with **costless geospatial data acquisition** and **quicker results** than traditional approaches.
 - **Vulnerability indicator** estimated based on land cover and economic indicators.



What are non-conventional sources?

Non-conventional sources refer to data sources that are typically not used in traditional statistical analysis, usually found in an unstructured way and may be “big data”, e.g.:

- **Social Media Platforms:** Data from Twitter, Facebook, Instagram, and other social media platforms.
- **Web Scraping:** Extracting data from websites, forums, and online news sources.
- **Satellite and Aerial Imagery:** Images and data captured by satellites or drones.
- **Mobile Phone Metadata:** Data from mobile phone networks, including call data records and location information.
- **Internet of Things (IoT) Devices:** Data from sensors and devices connected to the internet.
- **Consumer Transaction Records:** Data from online sales, credit card transactions, and loyalty programs.
- **Open Government Data:** Non-traditional government datasets made publicly available, such as real-time public transport data, health records, or environmental monitoring data.

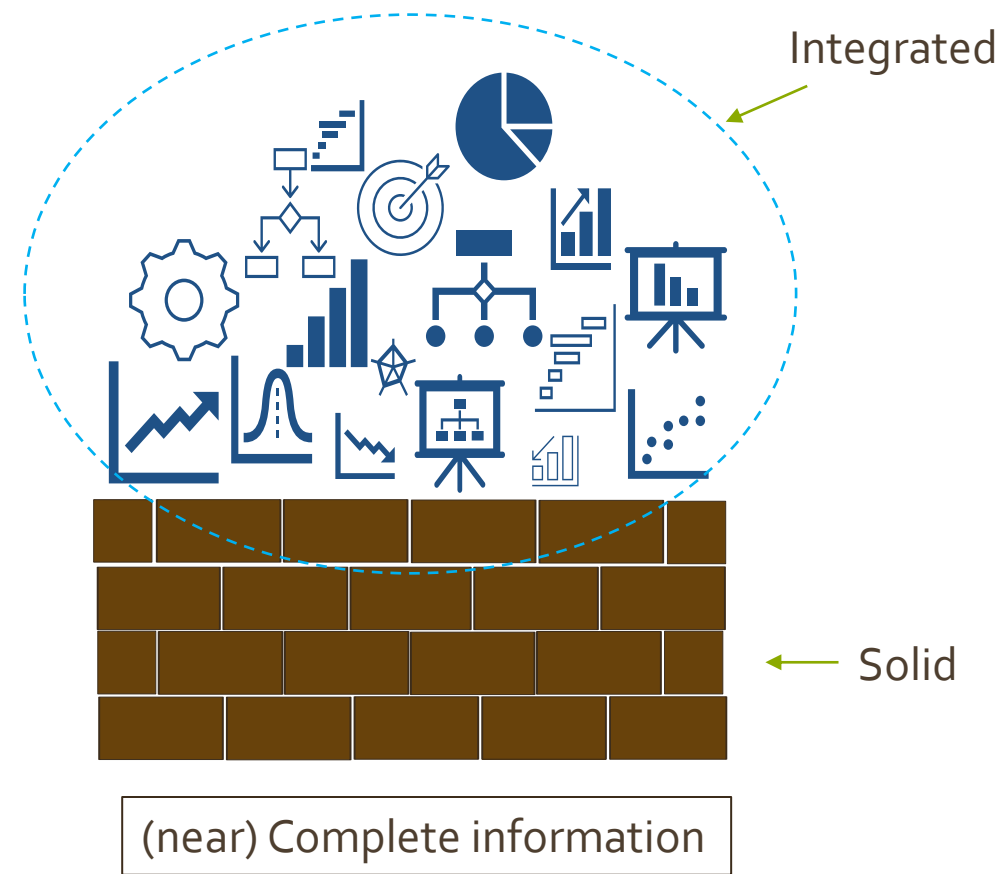
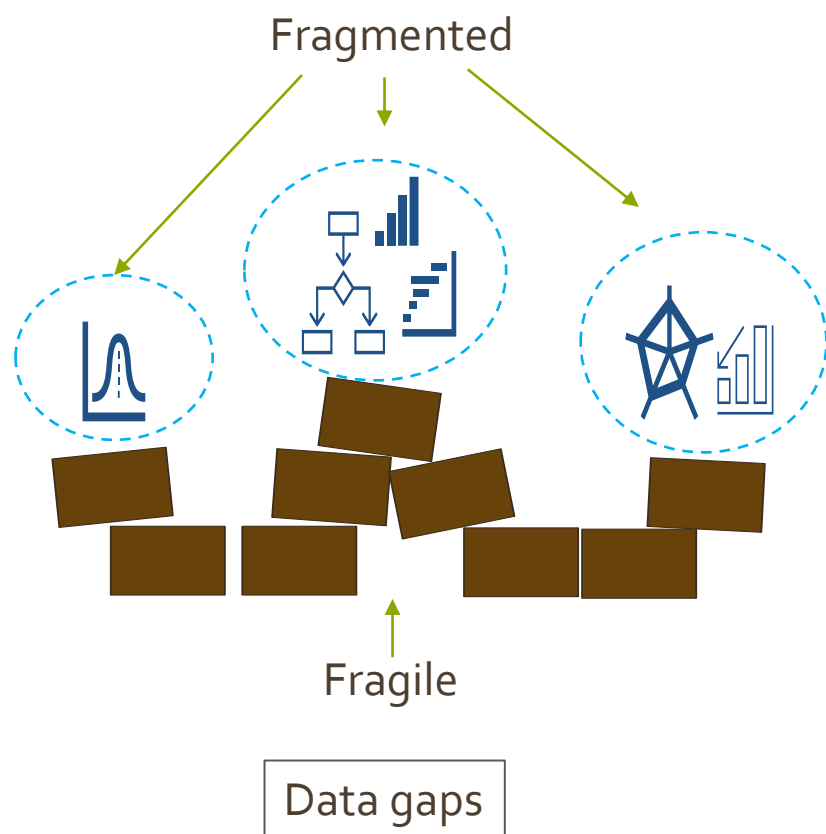


What is Data Science?

- An interdisciplinary field that uses scientific methods, processes, algorithms, and systems to **extract knowledge and insights from structured and unstructured data**.
- It combines aspects of statistics, computer science, information science, and domain-specific knowledge to interpret **complex data**
- Objectives:
 - **Data Exploration and Analysis**
 - **Predictive Analytics**
 - **Machine Learning and Advanced Modeling**
 - **Data Visualization**
 - **Big Data Technologies**
 - **Problem-Solving**
- It is distinct in its emphasis on developing new methods and tools to handle the volume, variety, and velocity of data available today

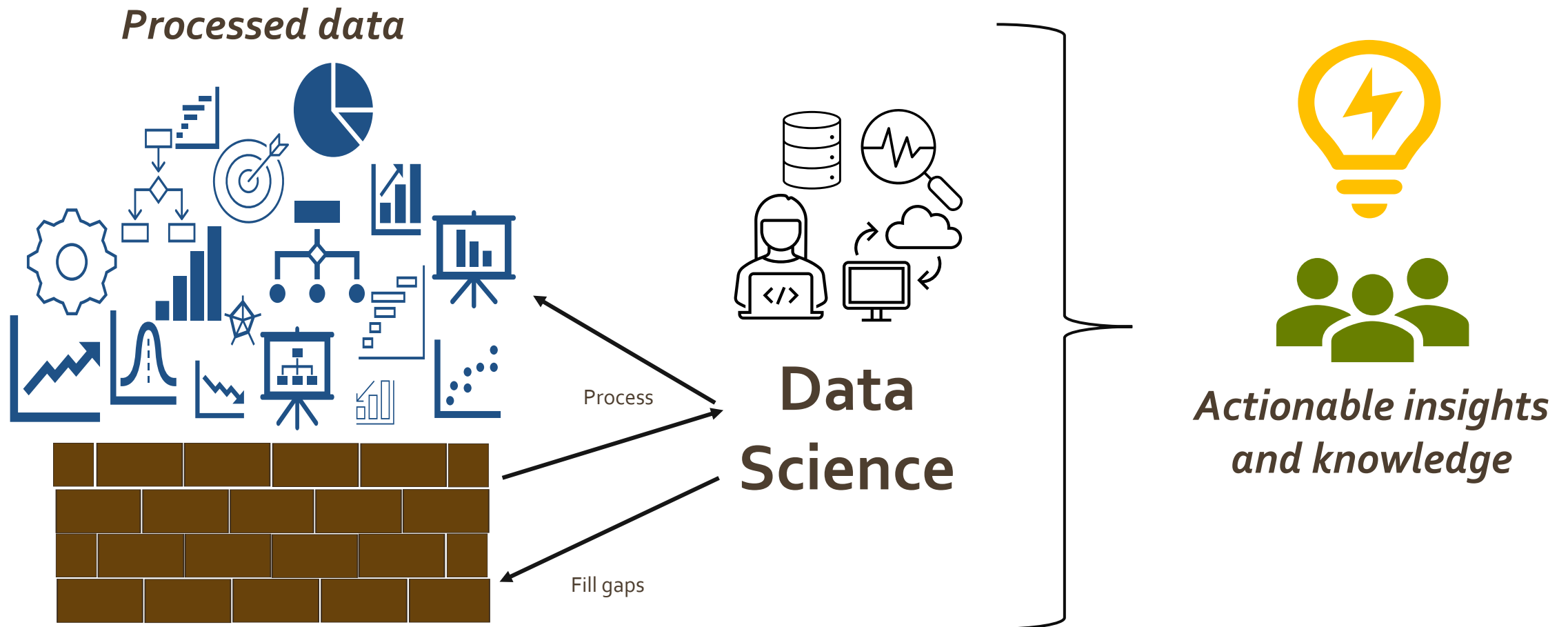


Data gaps vs (near) complete information



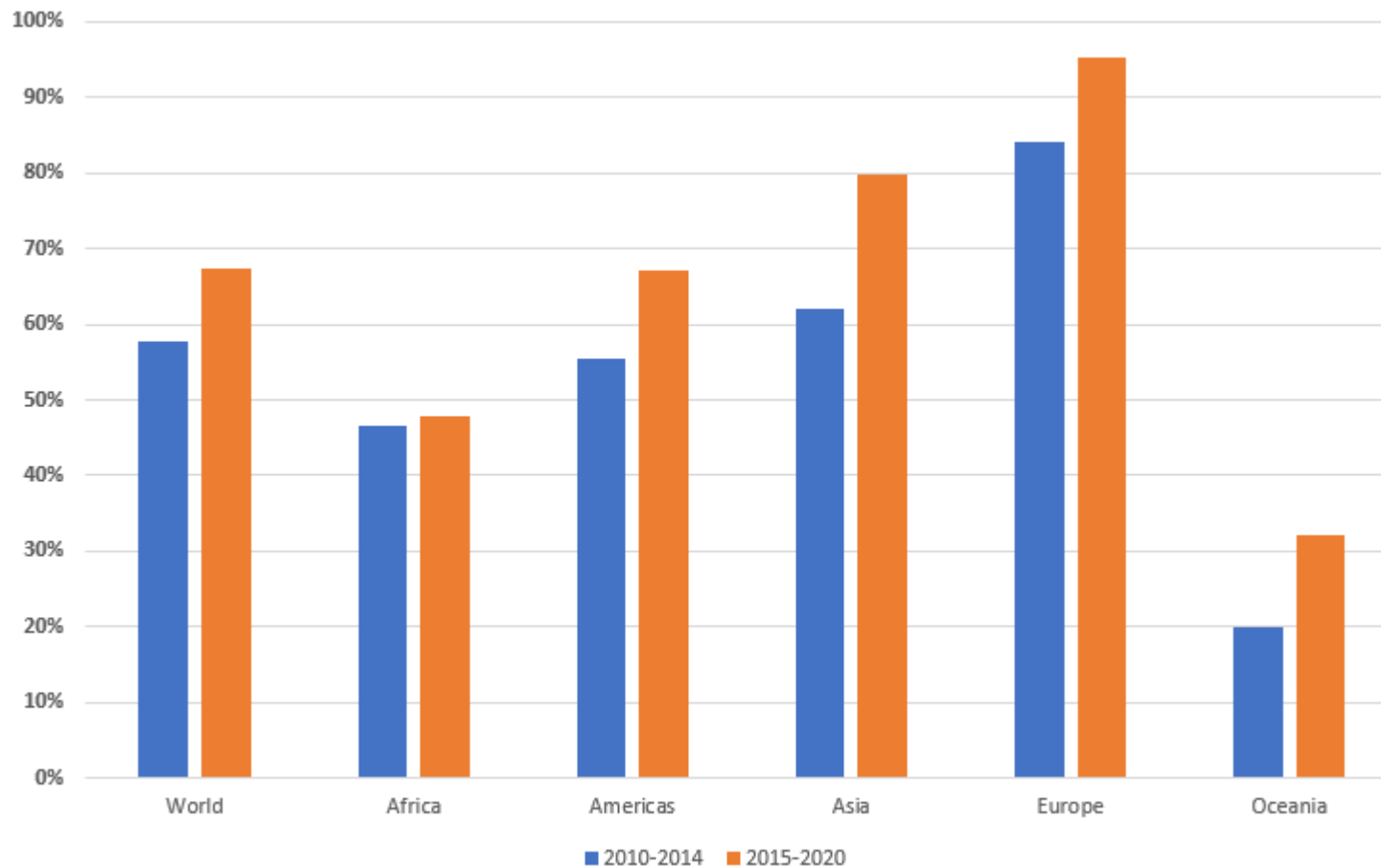


Data science as a tool to derive insights





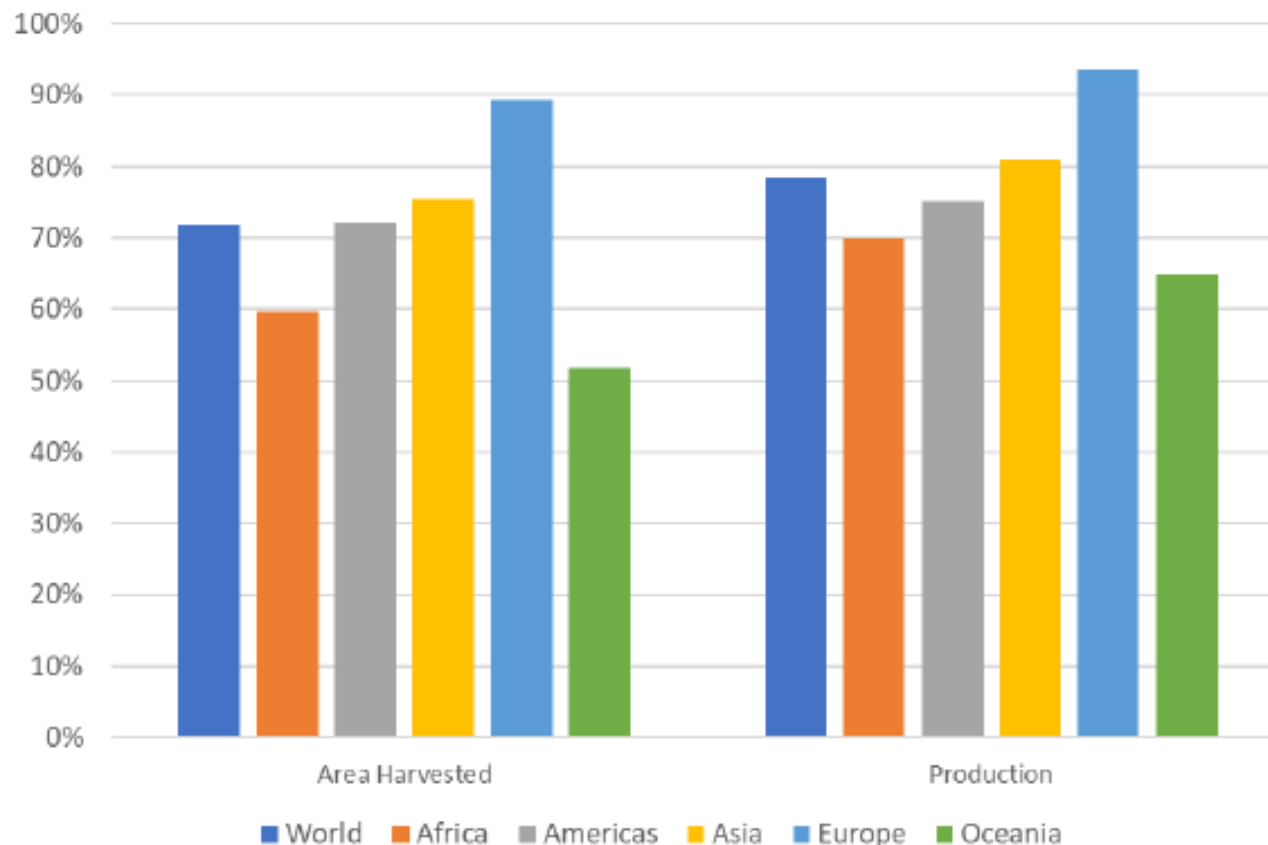
Average response rate to FAO's Agricultural Production Questionnaire



Response rate:
Percentage of received questionnaires over questionnaires sent



Completeness indicator for Agricultural Production Questionnaires



Completeness indicator:

Percentage of official data points over the total number of possible country/commodities combinations for which at least one official data point was ever provided by the country in one APQ

→ There are still data gaps to be filled.

How to deal with this?

↓
Web scraping + text mining



AFCAS 28

Filling Data Gaps in official production and area statistics

Table: 3.1.2.6 Area and Production of Maize (Rabi & Kharif) by District, 2016-17 to 2018-19.

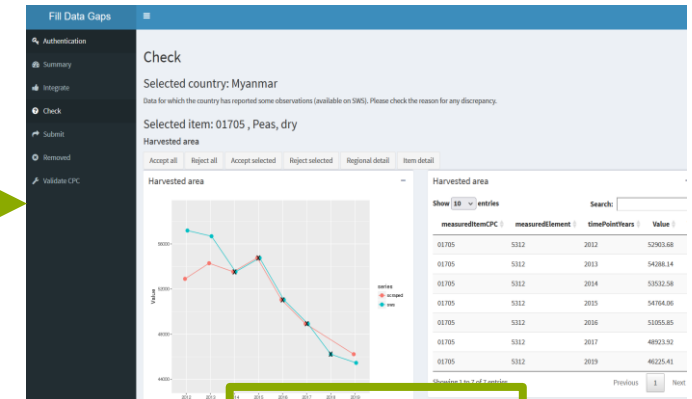
Zila/Division	2016-17		2017-18		2018-19	
	Area (acres)	Production (M. Ton)	Area (acres)	Production (M. Ton)	Area (acres)	Production (M. Ton)
1 Barguna	125	170	128	228	124	160
2 Barishal	322	557	322	559	320	554
3 Bhola	858	1011	390	633	1005	1528
4 Jhalokati	176	257	140	140		248
5 Patuakhali	111	149				
6 Projpur	619	1123				
7 Barisal	2211	3267				
8 Bandarban	741	575				
9 Brahmanbaria	0	0				
10 Chandpur	15130	38570				
11 Chattogram	226	618				
12 Cumilla	22207	38589				
13 Cox's Bazar	205	425				
14 Feni	7	4				
15 Khagrachhari	1332	1495				
16 Lakshmipur	0	0				
17 Noakhali	3	3				
18 Rangamati	1500	2612				
19 Chattogram	41381	82891				
20 Dhaka	11433	32362				
21 Faridpur	960	3516				
22 Gazipur	130	184				
23 Gopalganj	16	12				
24 Kishorganj	9871	28923				
25 Madaripur	156	349				
26 Manikganj	49345	126457				

مساحة وإنتاج وفئة التغير حسب المحافظات لعام 2018 وتغيرها على مستوى القطر خلال الفترة (2016-2019)

Area, Production & Yield of Grazing barley by Governorate for 2018 & their Development at the Country Level during (2009-2018)

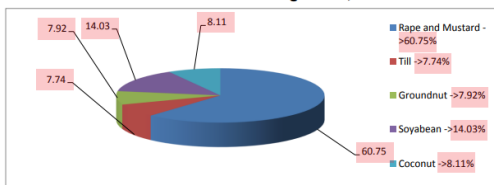
Year	Area (Ha)	Production (M.Ton)	Yield (kg/ha)
2009	37750	465197	12323
2010	13113	565304	43111
2011	18458	621875	59389
2012	12236	795508	65146
2013	4610	1181751	25837
2014	5216	1164561	22350
2015	9668	439182	45333
2016	4993	125180	17926
2017	11130	396684	35606
2018	9445	186390	19734

iso3	region_name	product_name	year	value	unit
PAK	punjab	dates	2007.00	44360.00	t
PAK	sindh	dates	2007.00	253090.00	t
PAK	khyber pakhtoonkhw	dates	2007.00	10380.00	t
PAK	balochistan	dates	2007.00	249690.00	t
PAK	punjab	dates	2008.00	44610.00	t
PAK	sindh	dates	2008.00	261950.00	t
PAK	khyber pakhtoonkhw	dates	2008.00	11340.00	t
PAK	balochistan	dates	2008.00	248590.00	t
PAK	punjab	dates	2009.00	44700.00	t
PAK	sindh	dates	2009.00	265300.00	t
PAK	khyber pakhtoonkhw	dates	2009.00	11300.00	t
PAK	balochistan	dates	2009.00	209900.00	t
PAK	punjab	dates	2010.00	42500.00	t
PAK	sindh	dates	2010.00	268600.00	t
PAK	khyber pakhtoonkhw	dates	2010.00	6800.00	t
PAK	balochistan	dates	2010.00	204300.00	t
PAK	punjab	dates	2011.00	44200.00	t
PAK	sindh	dates	2011.00	299800.00	t
PAK	khyber pakhtoonkhw	dates	2011.00	11800.00	t
PAK	balochistan	dates	2011.00	202300.00	t
PAK	punjab	dates	2012.00	43600.00	t
PAK	sindh	dates	2012.00	268900.00	t
PAK	khyber pakhtoonkhw	dates	2012.00	11800.00	t
PAK	balochistan	dates	2012.00	201100.00	t
PAK	punjab	dates	2013.00	43900.00	t
PAK	sindh	dates	2013.00	270500.00	t
PAK	khyber pakhtoonkhw	dates	2013.00	12400.00	t
PAK	balochistan	dates	2013.00	200000.00	t
PAK	punjab	dates	2014.00	44800.00	t
PAK	sindh	dates	2014.00	280800.00	t
PAK	khyber pakhtoonkhw	dates	2014.00	13800.00	t
PAK	balochistan	dates	2014.00	200000.00	t
PAK	punjab	dates	2015.00	42930.00	t
PAK	sindh	dates	2015.00	201170.00	t
PAK	khyber pakhtoonkhw	dates	2015.00	12320.00	t
PAK	balochistan	dates	2015.00	211340.00	t
PAK	punjab	dates	2016.00	43550.00	t
PAK	sindh	dates	2016.00	202300.00	t
PAK	khyber pakhtoonkhw	dates	2016.00	12380.00	t
PAK	balochistan	dates	2016.00	180760.00	t
PAK	punjab	dates	2017.00	37800.00	t
PAK	sindh	dates	2017.00	309700.00	t
PAK	khyber pakhtoonkhw	dates	2017.00	12400.00	t



Data validation

Area Under oil seeds Cultivation in Bangladesh, 2018-2019



Scraped and/or unstructured data

Structured database



Dissemination: <https://www.fao.org/faostat/>

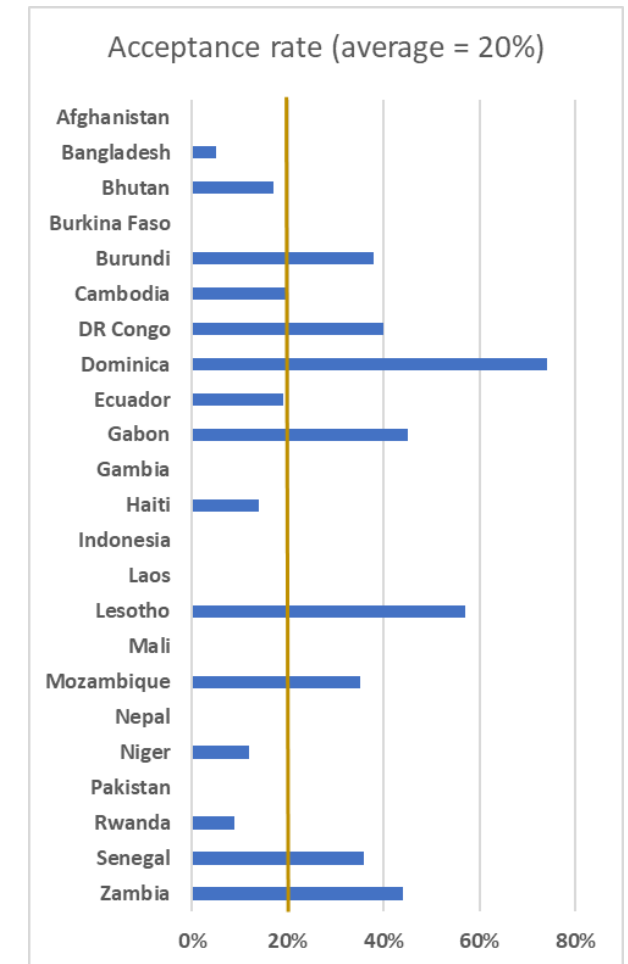


Data scraping results on FAO's 2023 production cycle

- Scraped data has been shared with the Crop, Livestock and Food Statistics team
- This year's results are summarised below

Country	Products scraped	Accepted		
		Products	Data points	% products
Afghanistan	20	0	0	0%
Bangladesh	85	4	15	5%
Bhutan	63	11	35	17%
Burkina Faso	13	0	0	0%
Burundi	13	5	5	38%
Cambodia	44	9	121	20%
DR Congo	15	6	10	40%
Dominica	39	29	200	74%
Ecuador	32	6	141	19%
Gabon	20	9	39	45%
Gambia	15	0	0	0%
Haiti	14	2	3	14%

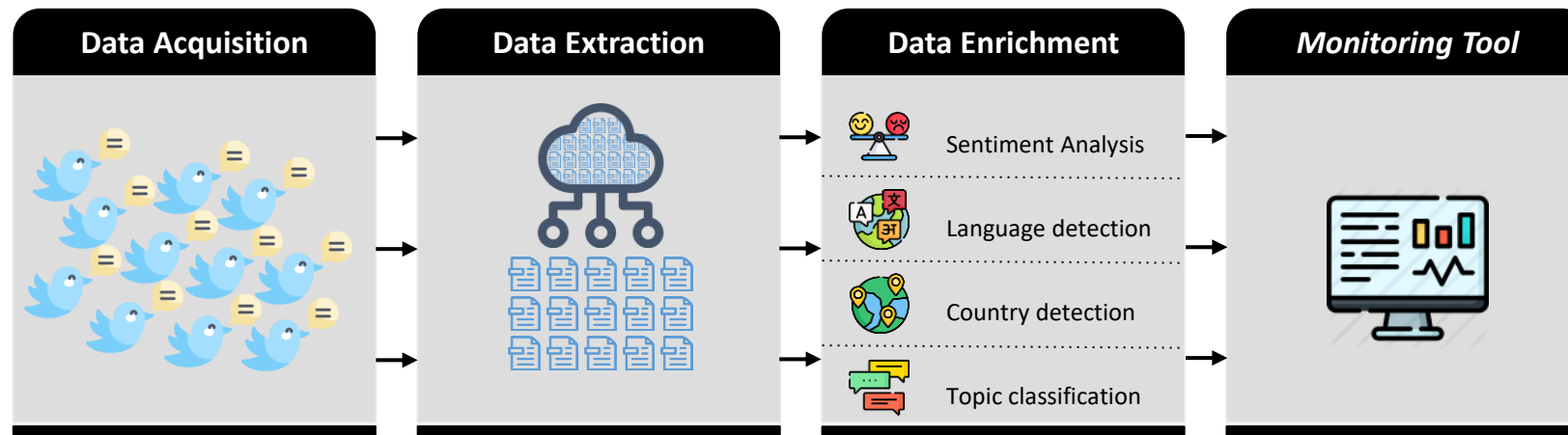
Country	Products scraped	Accepted		
		Products	Data points	% products
Indonesia	22	0	0	0%
Laos	12	0	0	0%
Lesotho	7	4	13	57%
Mali	6	0	0	0%
Mozambique	20	7	11	35%
Nepal	9	0	0	0%
Niger	17	2	13	12%
Pakistan	33	0	0	0%
Rwanda	65	6	12	9%
Senegal	11	4	43	36%
Zambia	27	12	50	44%





Obtaining insights from news articles

- Started scraping news articles through Twitter in **January 2020** → Contains **28 million articles today**
- After Twitter's became X we started scraping directly newspapers websites → from July 2023 **~3 million articles**
- A daily procedure** that scrapes articles from **more ~500 newspapers: 188 areas, 8 languages**
- Depending on the use case, relevant articles are identified and used to derive insights into a particular topic





From data complexity to knowledge

- We have 30+ million news articles: how to use this **huge** dataset?
- Data science → We have trained machine learning models on topics relevant to FAO’s mandate:

Agriculture Prices

Animal Health

Climate Change

Covid-19

Financial Sector

Food Loss and Waste

Food Prices

Food Security

Healthy Diets

Natural Disasters

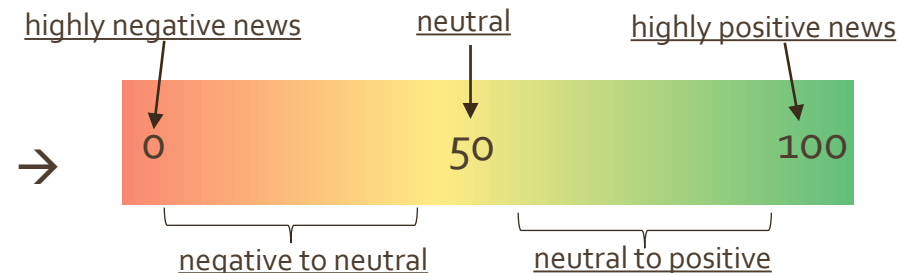
Social Unrest

Socio-Economic Development

Food Value Chains

- Additionally, we compute a “sentiment” score (w+ and w- are positive and negative words, respectively)

$$S = \frac{\#w^+}{\#w^+ + \#w^-} \times 100 \rightarrow$$

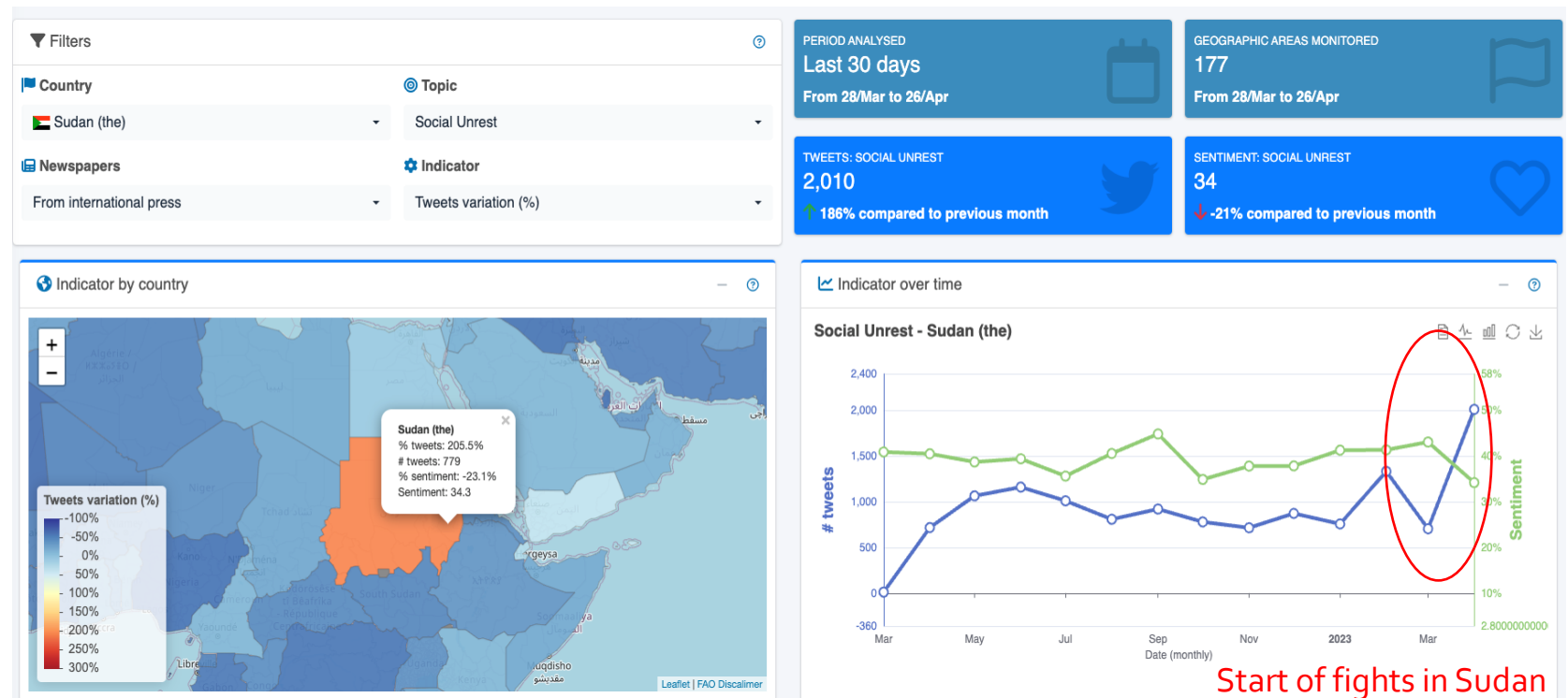




Topics explorer

- Topic Selected: Social Unrest
- Country: Sudan
- Map: International Press

- The popularity (i.e., number of articles) of the Social Unrest topic related to Sudan **increased by 186 %** in the international press
- The associated **sentiment fell by 21 %**



Source: FAO Data Lab, 2023



Data Lab Trends

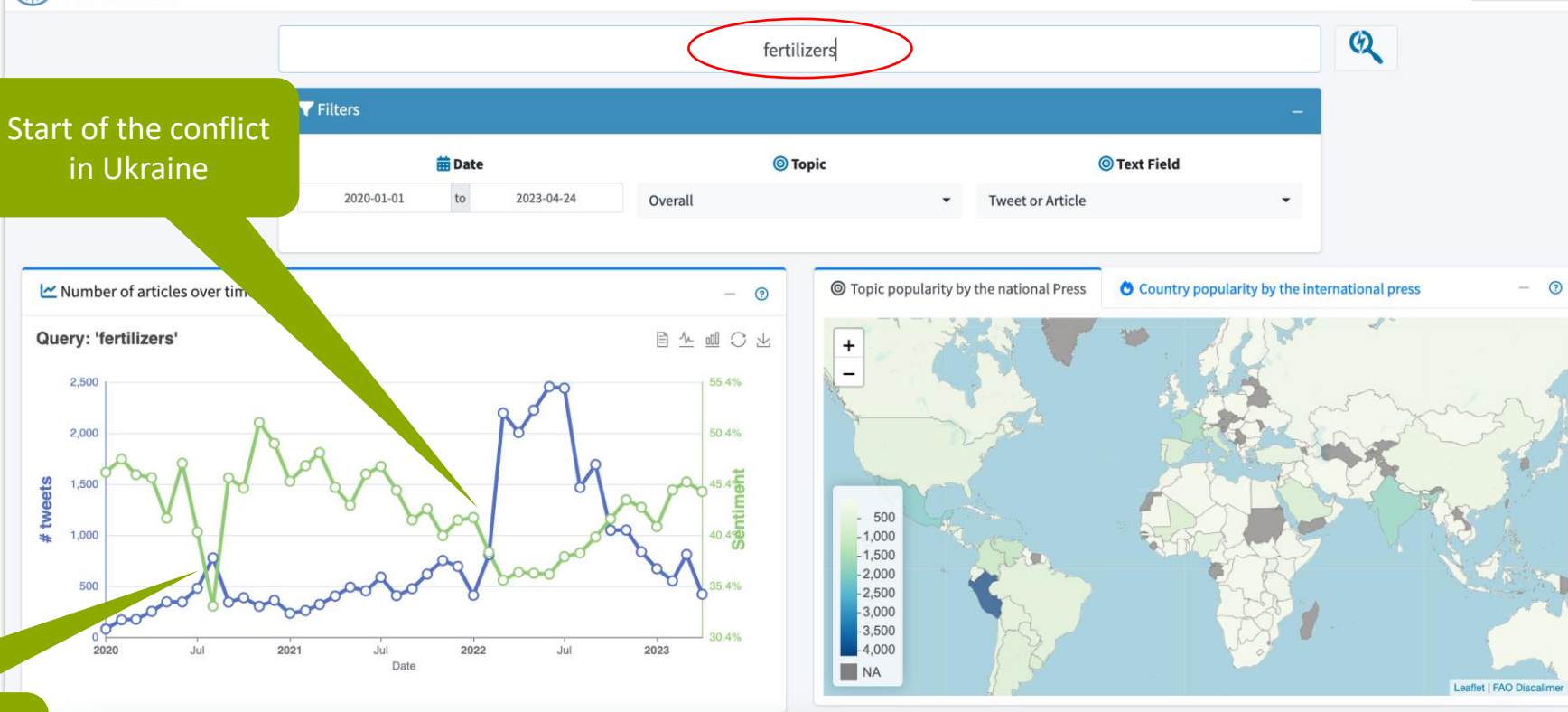


Show intro

- Stepping away from pre-defined topics, the tool analyzes the **popularity** of **any search query** over the articles gathered from media outlets across various countries and languages
- Example:
 - Search query: **"fertilizers"**
 - Jan/2020 - Apr/2023

Start of the conflict in Ukraine

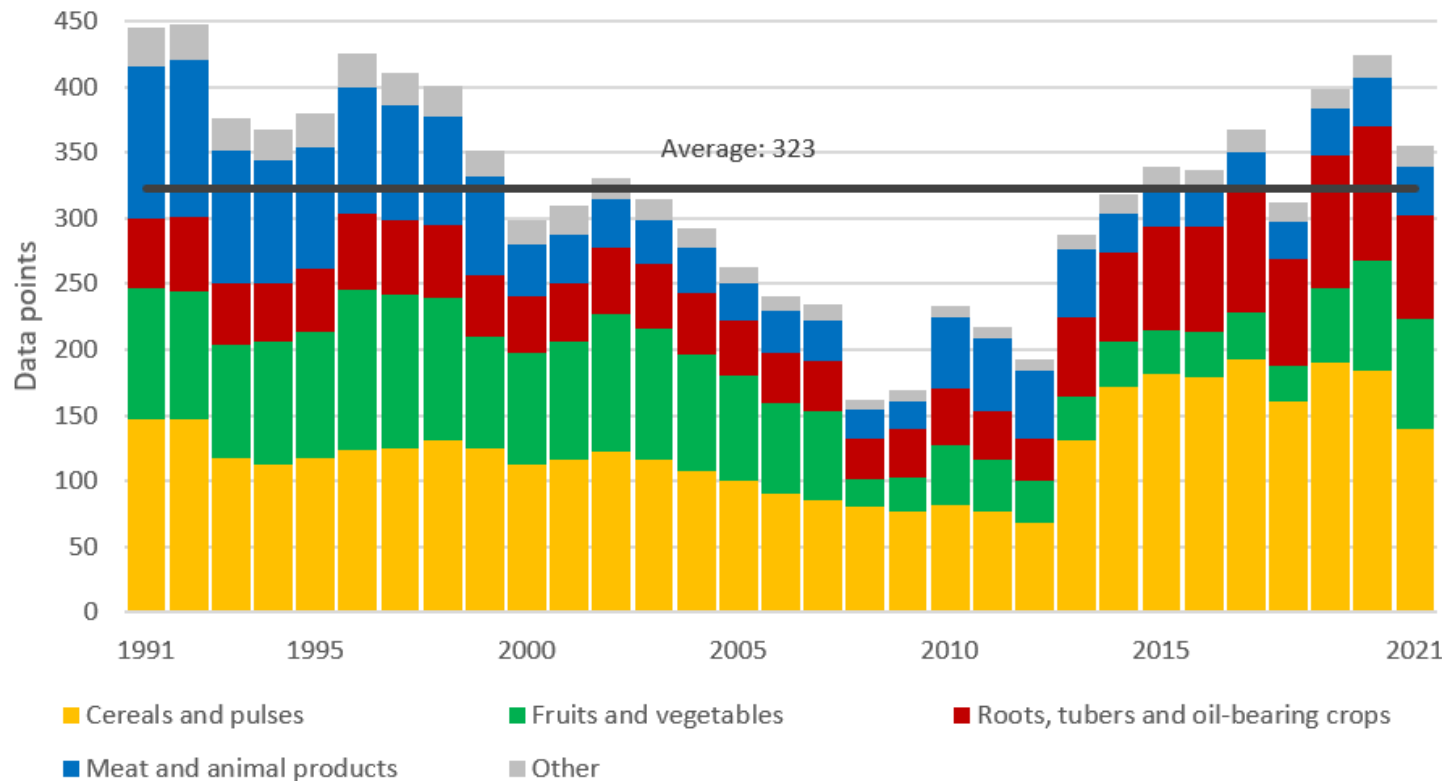
Ammonium nitrate stored at the Port of Beirut in the capital city of Lebanon exploded



Source: FAO Data Lab, 2023



Available information on Food Losses



On average, FAO collects each year 323 data points from countries (Figure 2), which is only around 4 percent of the data collected for agriculture production



“...reduce food losses along production and supply chains, including post-harvest losses.”



AFCAS 28

Food Loss and Waste Extraction via Text-Mining

Harvested scientific articles, reports, etc.

Post-harvest dip of enhanced freshness formulation to extend the shelf life of banana (Musa acuminata cv. Grand Naine) in India

Kammar Venkateshbabu¹, Jayaprakash Muthuraj¹, Selvakumaran Sundaresan¹, Kishore K. Subramanian², Jeyaraj Ganapagan Jayakumar¹, Alan Sullivan¹, Gopinathan Polysaiah¹ and Jayashanker Subramanian¹

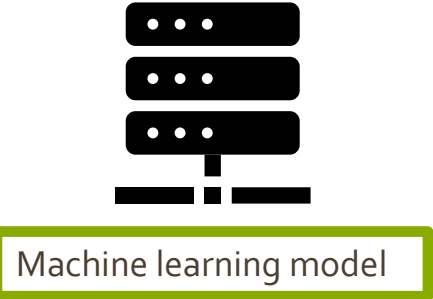
¹The Horticultural College and Research Station, Tamil Nadu Agricultural University, Periyaralvar, Tamil Nadu, India; ²Department of Food Science and Technology, Tamil Nadu Agricultural University, Combarakere, Tamil Nadu, India.

³Department of Plant Agriculture, University of Guelph, Guelph, Ontario, Canada; ⁴Corresponding author email: Jayashankar@rediffmail.com

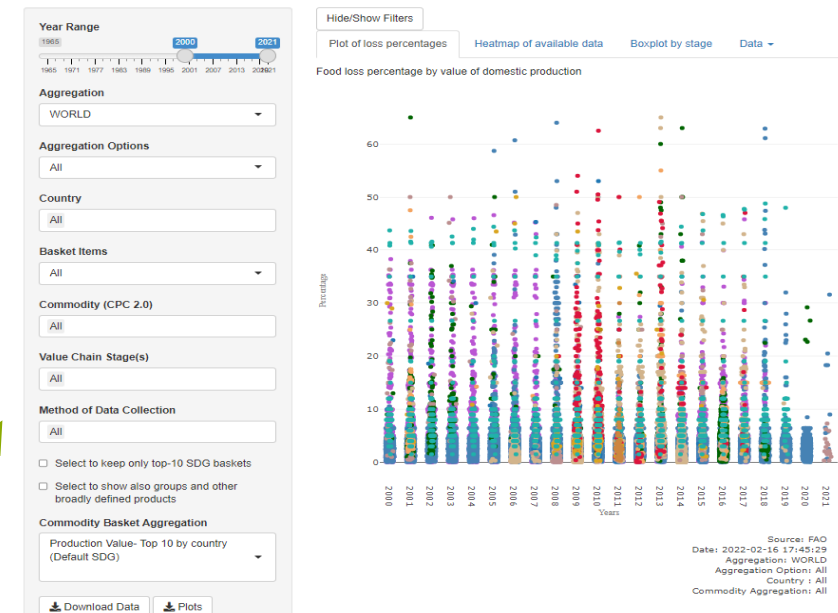
Abstract
 A laboratory study was undertaken to determine the effect of a novel antimicrobial preservative formulation (EPF) on a post-harvest dip technique to minimize the post-harvest losses and to extend the shelf life of banana. The dipping solution was prepared by mixing 0.05% (v/v) of EPF with 0.5% (v/v) of sodium hypochlorite solution. The results of the study indicated that the dipping solution significantly reduced the microbial load and the ripening rate of banana. The EPF showed a higher weight and higher firmness throughout the study period, regardless of maturity level at harvest. Treated fruit had higher leaf and soluble solids, total acid, and soluble solids content than the control. The EPF showed a higher weight and higher firmness throughout the study period, regardless of maturity level at harvest. Treated fruit had higher leaf and soluble solids, total acid, and soluble solids content than the control. The EPF showed a higher weight and higher firmness throughout the study period, regardless of maturity level at harvest. Treated fruit had higher leaf and soluble solids, total acid, and soluble solids content than the control.

Figure 1: Physiological weight loss in sweet orange in ambient (left; 28±2°C) and cold (right; 18±2°C) conditions. Top panel depicts cv 'Jaffa' and the bottom panel depicts cv 'Mesa'. Values are mean ± Standard deviation.

Effects of post-harvest treatments, storage duration and storage conditions on titratable acidity.

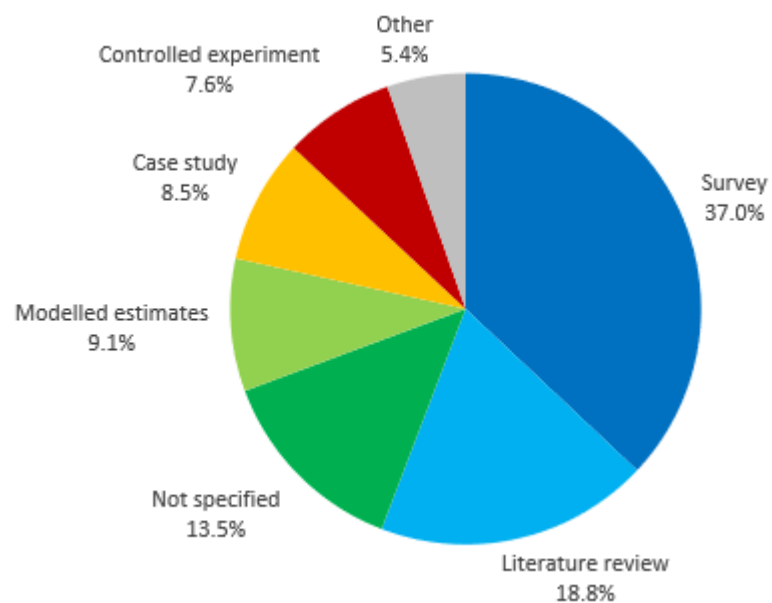


Dissemination

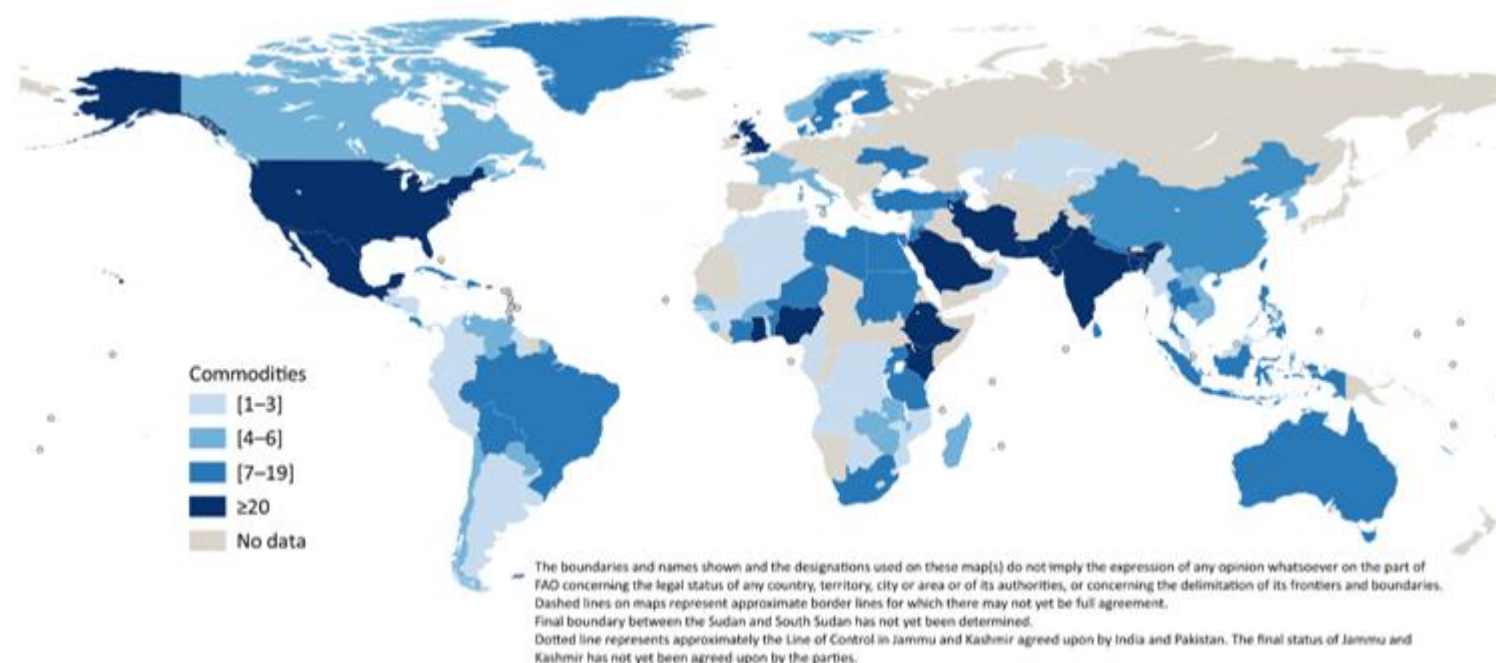




Results of Food Loss Data Extraction



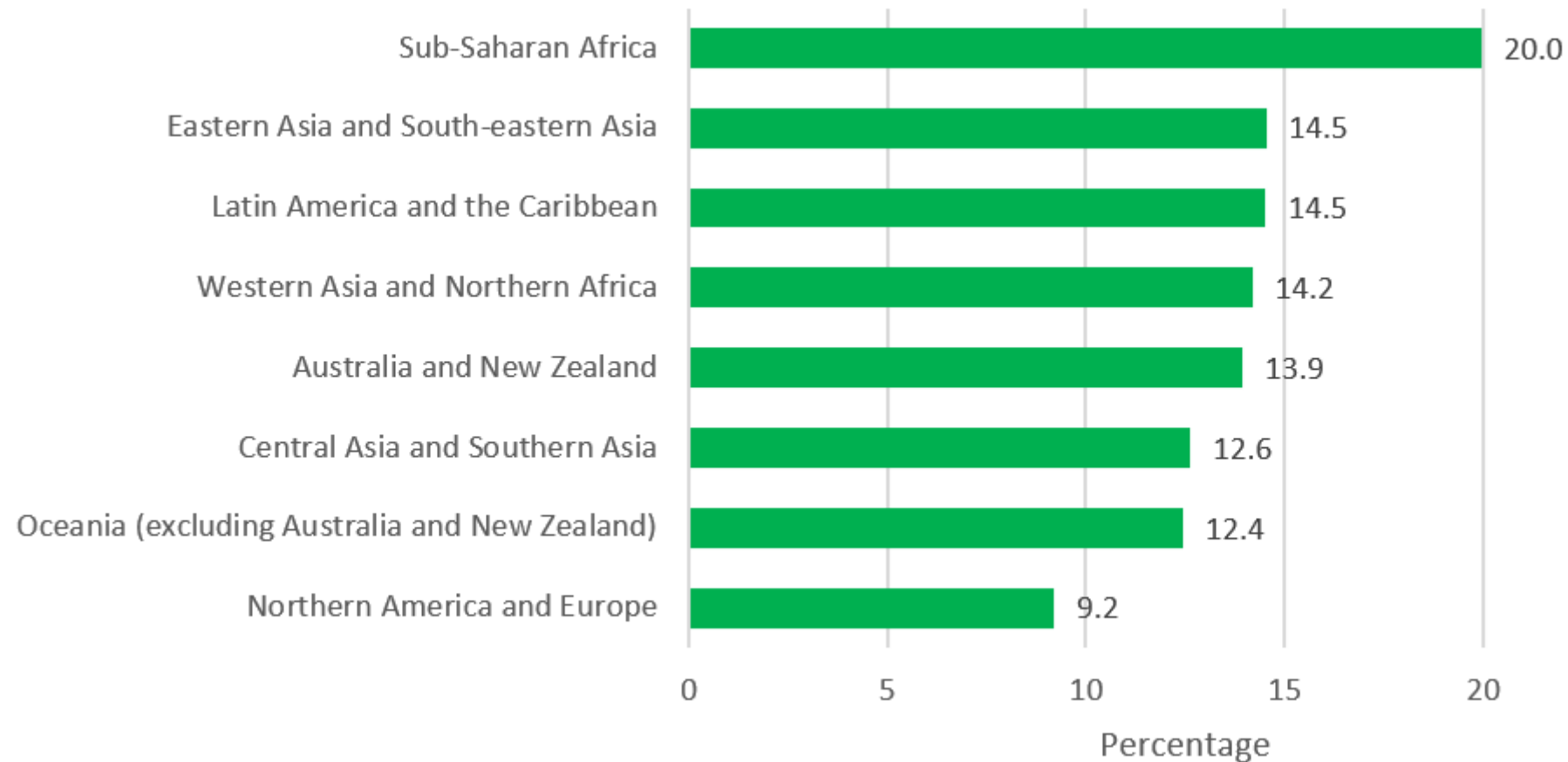
Methods of data collection used from reviewed literature



Number of products available from reviewed literature



Food loss percentages by at the SDG region (2021)





Conclusions

- There is a need to overcome traditional data collection challenges
- Non-traditional data sources and methods like social media, IoT devices, web scraping, and text-mining are revolutionizing data collection and analysis
- Big data, data science and artificial intelligence have a growing role in national and international data systems
- FAO's Data Lab is using advanced technologies and non-conventional data to enhance data quality and coverage, e.g. in agricultural statistics and food losses, and providing global insights
- Organizations and countries should embrace modern data science techniques and non-conventional data sources to stay relevant and efficient in a data-driven world

The boundaries and names shown and the designations used on this/these map(s) do not imply the expression of any opinion whatsoever on the part of FAO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers and boundaries. Dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

Thank you for your attention!

For more information, please visit:

<https://www.fao.org/food-agriculture-statistics/resources/events/afcas/en/>

Contact: Christian.Mongeau@fao.org

AFCAS 28
LEVERAGING
DATA & STATISTICS
FOR AGRIFOOD
SYSTEMS
TRANSFORMATION
IN AFRICA

