# How Machine Learning Can Help Turn Evidence Into Policy

## If we want an evidence based SDG 2 by 2030, we need transformative approaches to generating evidence—today.

Agriculture has seen the total volume of research double in the past ten years from two to more than four million articles. The trend is mirrored across the sciences where, every seven seconds, a new scientific study is published. And in addition to thousands of academic journals, there are hundreds of organizations doing important research published outside of academic journals.

Before we even begin to look at what the research says on a particular issue, it may take us a year just to figure out where all the relevant research is. There is no system—not even Google—that can connect all this information in a way that enables us to synthesize it into evidence, quickly and effectively.

And finding the relevant research is just the first step.

How do we search for meaning among thousands of studies without cherry picking a few of our favorites? How do we decide whether the evidence we have found is weak or strong—or that we have enough to act now?

These are questions we are answering in Ceres2030: Sustainable Solutions to End Hunger, a partnership of organizations working with researchers to figure out what science can tell us to help achieve zero hunger—and to figure out what these interventions might cost to implement.

Our initial study of 50,000 papers on small-scale food producers between 2008-2018 found that the word “intervention” only occurred in 5 percent of the articles.

How do we generate evidence-based consensus for effective interventions in agriculture if we can’t locate interventions in the research?

---

1 We needed to search for research on interventions scattered across at least 60 different repositories and thousands of journals. But while there are many interventions designed to tackle agricultural and food security problems, many are not not described as “interventions” in the literature.

A simple problem in classification thus presents a huge problem for researchers, and one that conventional search engines can do little to solve. Even if they could access all the journals, websites, and databases we needed to look through, how would they help us discover interventions that weren’t described as interventions?

This is why we turned to machine learning for help.

Ceres2030 information science meeting, November 2018; participants including Nature Plants, Campbell Collaborative.

Ceres2030
Sustainable Solutions to End Hunger

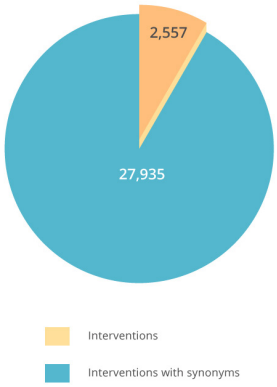
Using this approach, we connected previously unconnected research. This saves researchers weeks, even months, of time, and provides fast and accurate searching.

Machine learning models help us explore text. We train these models to understand the way we use language by exposing them to millions of words and the way these words relate to other words in a vast range of contexts.

The primary training tools for these open-source machine learning models are Wikipedia and Google News—the largest repositories of digital text available—and then they are refined by having them explore more specialized collections of text.

We trained our model on a sample of 50,000 research papers selected on the basis that they addressed a target population of SDG 2: Small scale food producers. First, we found all the synonyms for intervention. These turned out to include words like "targeting" and phrases like "capacity building."

Why our words matter




We increased discovery from 5 to 55 percent when we used machine learning to help us search for "intervention(s)."

| Word | Articles using the word | | |
|---|---|---|---|
| Intervention | 2561 | Targeting | 674 |
| Policy | 7234 | Capacity building | 428 |
| Strategy | 5752 | Participatory approach | 393 |
| Measure | 2822 | Programming | 323 |
| Program | 2785 | Social protection | 263 |
| Project | 2609 | Entry point | 166 |
| Programme | 1961 | Policy option | 138 |
| Outcome | 1773 | Nutrition education | 62 |
| Recommendation | 1180 | Multi-sectoral approach | 14 |
| Initiative | 1085 | | |

Natural language processing enabled us to discover how scientists refer to "interventions" in the research literature.

2 We then used machine learning to explore the text "as a collection," asking the machine to 'tag' each paper based on what was in the text.

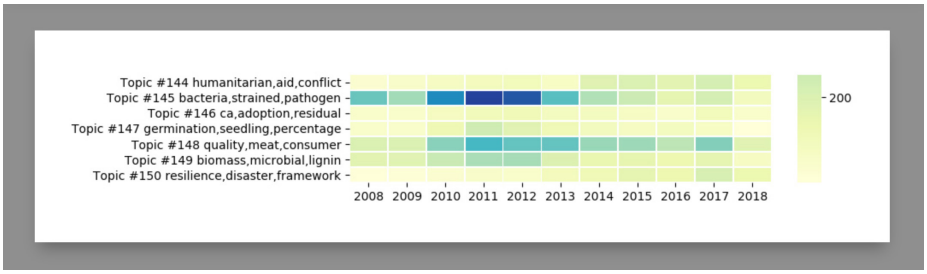
Normally, when a research paper is accepted or uploaded to a database, it is 'tagged' with keywords from a pre-defined list. We flipped this around so that we could 'ask' the research 'tell' us about itself first.

Now we could see exactly what the interventions were and how they were related to each other.

For example, if you were to do a key-word search of our dataset for associations between "interventions" and "greenhouse gas emissions," you would return about 10 percent of the dataset.

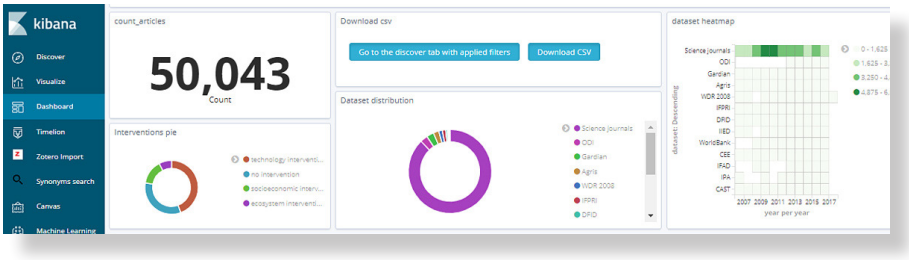
But if you did a search looking for associations between synonyms for intervention and greenhouse gas emissions, you would end up finding a lot more relevant material—approxi-mately 60 percent of the dataset.

Our approach enabled us to see trends in agricultural research in our dataset

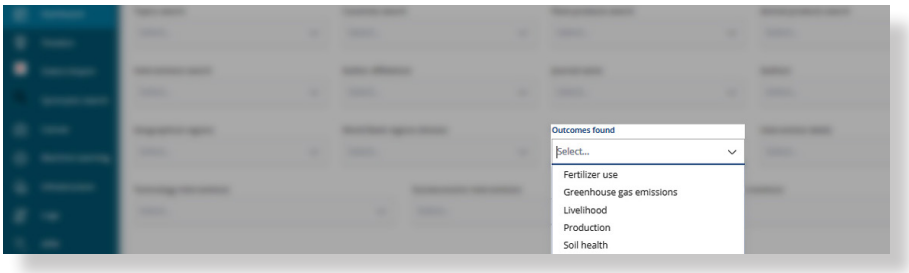



Machine learning enabled us to identify over 2,000 interventions and classify agricultural reserarch into 150 topics. This density map shows where research has been concentrated (the darker the blue, the more research papers), and where there are gaps in research.

Ceres2030
Sustainable Solutions to End Hunger




3 Our database currently contains a dataset of 50,043 articles, compiled from a search for research on small scale food producers. We use a Kibana dashboard to access and visualize the dataset. Machine learning allows us to classify information in multiple ways, allowing a richer and more informative search experience. For example, we'll select an outcome from the "outcomes found" search box, in this case "Greenhouse gas emissions.




We then build a search using different filters to select for the specific information we want to find—in this case, we also want to know about greenhouse gas emissions in terms of rural infrastructure and socioeconomic interventions. Our search turns up 201 articles.

We can analyze the search data through Kibana's visualizations, and quickly see topic trends, authors, countries covered. We also built an app so that data can be downloaded as a CSV and shared through the open source reference management software, Zotero.


These kind of tools are widely used in the commercial world but they are new for exploring research literature.

They put research quickly in reach of policy and decision makers.

"A well-designed search strategy is a critical to achieving high-quality systematic reviews, and success hinges on finding every relevant synonym for key concepts in the literature. This is an incredibly helpful resource for researchers and librarians. It will help us design better searches and it will enable us to do those searches quickly—with the confidence we are able to find all the relevant evidence."

— Kate Ghezzi-Kopel, Health Sciences and Evidence Synthesis Librarian and lead of the research synthesis team for Ceres2030.

**Jaron Porciello,** Co-Director Ceres2030, Research Faculty, Primary Investigator, and Associate Department Director, Cornell University

**Trevor Butterworth,** Communications, Ceres2030; Exec Director Sense About Science USA

**For more details, contact:**
ceres2030@cornell.edu

**Ceres 2030** is a partnership between **Cornell University**, the **International Food Policy Research Institute** (IFPRI), and the **International Institute of Sustainable Development** (IISD)

www.Ceres2030.org

Ceres2030
Sustainable Solutions to End Hunger