SESSION 2: USES AND CHALLENGES OF 'BIG DATA' FOR AGRICULTURAL DEVELOPMENT

Overview: **Local applications for global data and AI**

Steve Mathews

Gro Intelligence

Abstract

'Big data' has great unrealised potential in most parts of the agricultural value-chain. We can divide that data up into several categories, each with its own good and bad points. Starting in 1972, the US Landsat program collected the original 'big data', but capability to perform meaningful analysis of the photos remained very expensive until recently. Other flows have begun in the past few decades, from private satellites, point-of-sale systems, land-based sensors, and aerial drones. Unlike Landsat, the various newer sources have different ownership statuses. Globally, most smallholders don't generate the revenue to pay for any of the various proprietary data sources or analysis. But we see significant value in the application of machine learning/'big data' techniques to publicly available satellite and other sources. Advances in information technology allow us to disseminate good-quality yield, drought, and other analyses at a much lower cost than previously. As a result, relatively small external contributions can bring the established benefits of modern modelling expertise to a hugely broader and more diverse audience.

In my prior career I was researching and trading commodity futures and options, for agriculture as well as metals and energy. Apart from that background, my experience in 'big data' and modern analysis techniques in agriculture comes almost entirely from my company, Gro Intelligence. I tell you that because inevitably in this talk I am giving you the company's viewpoint on these sorts of analyses. I strongly believe, along with others, that the things that Gro Intelligence is doing and has done are important to enable farmers and smallholders in sub-Saharan Africa to make progress, because for these groups there are many issues that do not seem to be being solved by the data currently available to them.

Agriculture is still at the very early stage of using modern 'big data'. Energy, banking and transportation are way ahead and those fields have many lessons to teach us. Those industries are still working 'full speed ahead' and moving forward at the same speed we are, I would say.

The McKinsey Global Institute built a digitisation index for the US (Figure 1): a matrix of green, red and yellow squares that indicate the level of preparation

---

This paper has been prepared from a transcript and the Powerpoint slides of the presentation.

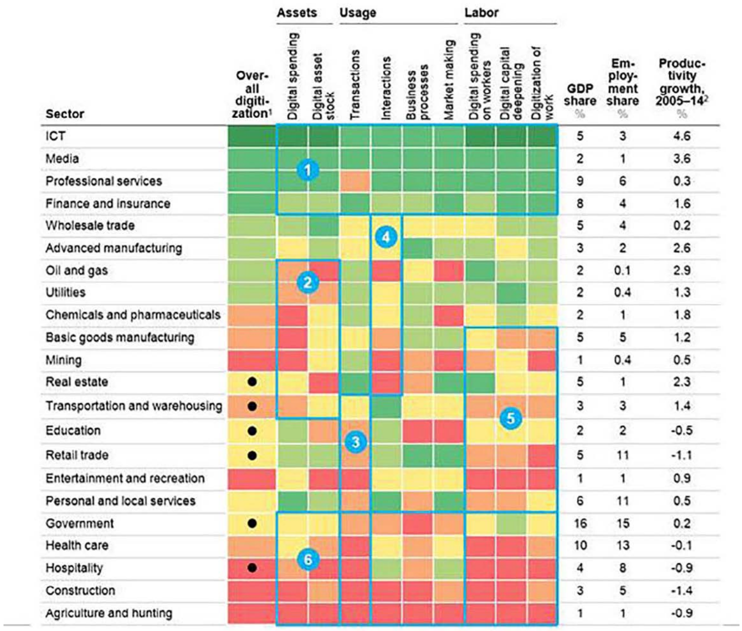


Figure 1. McKinsey Global Institute Digitization Index up to 2014: ICT and media at top, all green; Agriculture & Hunting at bottom, all red.

of each industry in that particular area of 'big data' or digitisation in general. At the bottom is Agriculture (& Hunting), shown in red for all the rated categories (grouped under Assets, Usage and Labour). Agriculture rates lower than mining, government and hospitality, all of which are also fairly far 'behind' (in comparison with other sectors in the US) in their uptake of digital technology.

Agriculture is only just starting to look at large-scale low-cost parallel processing, distributed storage of data, access to better data from private satellites, point-of-sale systems, land-based sensors, aerial drones and individual farmer smartphones. These can be transformational in the long-term, encouraging private sector activity and attracting the types of agricultural investments that can help make the sector more resilient.

Making agricultural data valuable to business

Gro's insight was that, although poor farmers in Africa know a lot about their farms and are not always easy to convince that computer tools can help them, there are people affiliated with them who do understand, or may understand, the role of data, and may have the capability to pay for that. Gro Intelligence is a 'for profit' company but we are working on this in parallel with non-profit and NGO-type activities, and other companies in our area. We think that more sophisticated analytics that respond to the marketplace and respond to customer demand definitely have a place. That is where solutions like ours can help agriculture to make progress.

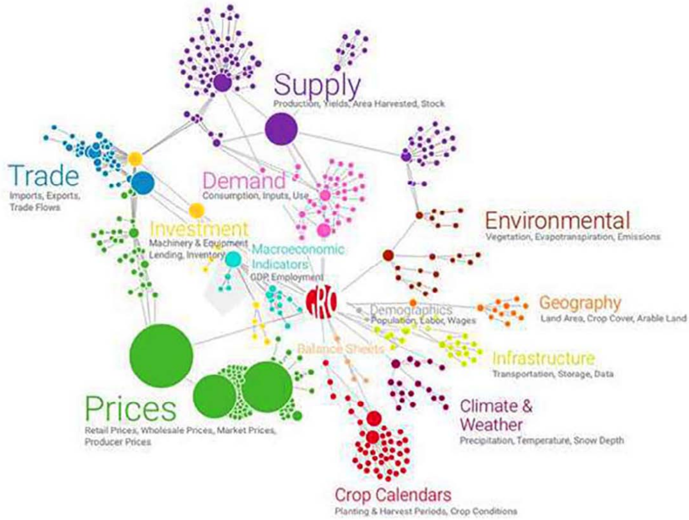


Figure 2. Gro Intelligence's proprietary ontology.

At a high level, Gro Intelligence is a fairly simple and straightforward company: we take in data from disparate sources, manipulate it and disseminate it to disparate clients. In my opinion, we add a lot of value in that middle phase of manipulation.

We have used a cadre of experts in agronomy, geospatial data, cartography, statistics, computer science and finance to create an ontology. (An ontology is a taxonomy without the requirement for hierarchy; in biology, for example, the taxonomy of every known species of life on the planet places them in a strict hierarchy.) The ontology Gro Intelligence has built is designed to help agribusiness understand the meaning in the agricultural data we collect.

Gro's proprietary ontology avoids the concept of hierarchy so as to create greater value for commerce. Building an ontology of that nature takes a lot of sophisticated work and the outcome is hard to picture (illustrated in Figure 2 in two dimensions).

We are in contact with agricultural trade and statistical agencies all over the world and the data we routinely receive is in a range of formats often derived from collection or processing in local units or standards, and in some cases using archaic methods. Bushels per acre versus kilograms per hectare is just the start. Different crop years, even different calendar years, changing borders and nations – it's a mess! We have done and continue to do a lot of work to make the data accessible and useful. Furthermore, we constantly assimilate more sources and data into the framework. As well, Gro has integrated geospatial information from satellites and other sources.

Landsat images of Earth were the world's first data that could be really qualified as 'big', and they began to be collected in the early 1970s before there was widespread ability and capable equipment to handle them. Now, the geospatial

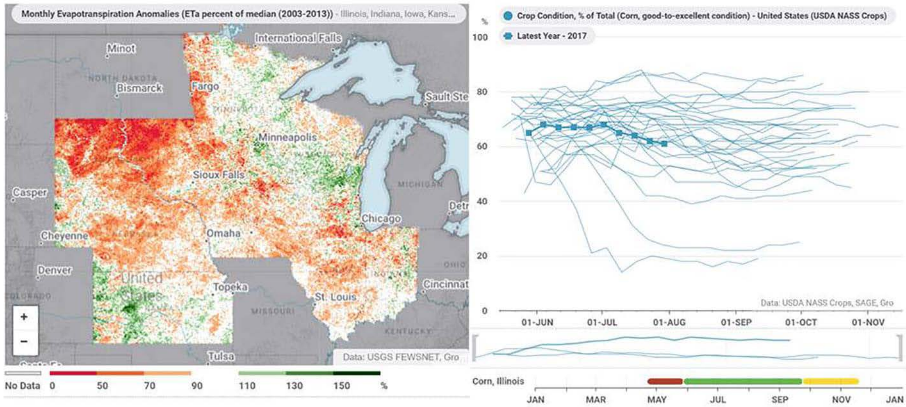


Figure 3. Satellite pictures (left) are converted to actionable indices (right).

team at Gro Intelligence, comprising five fulltime scientists in New York and Colorado, extract more useable insights from satellite pictures each day. The data rolls in daily in gigabytes and is stored in our multiple petabyte database.

Examples

Data is only as useful as what you can do with it. Visualisations, charts, maps, calendars – things of that nature – allow users to access the information in unanticipated ways that mean something to them.

Some of our users now need only minutes to access data that took them days, weeks or even months to partially obtain through manual data research projects. Another set of our users is accessing data at large-scale for the first time, because previously it was only available to institutions that could afford to employ entire data teams. With new technology we can pool and analyse trillions of agricultural data points from a variety of sources, such as government reports, satellite imagery and weather forecasts, to give them universal meaning and to give users insight.

Geographic information systems (GIS) link satellite imagery to ground-collected information. Geospatial knowhow means we can distil a complicated and colourful map into a single index which summarises the whole picture (e.g. Figure 3). You can see a crop calendar there on the bottom of the line chart. The GIS expertise in our team allows us to take those geospatial data points that you see on the map on the left and sum them within almost arbitrary geographic regions – down to state, county, district level, whatever you require.

Machine learning techniques applied to those various data series, which include ground-collected data, satellite data or other types, can then generate predictive models. Machine learning can seem like a 'black box', and I don't want to 'get into the weeds' on the subject. But great advances made recently mean that we can now use machine learning very inexpensively, at a low level, to generate models which are better at prediction than previously.

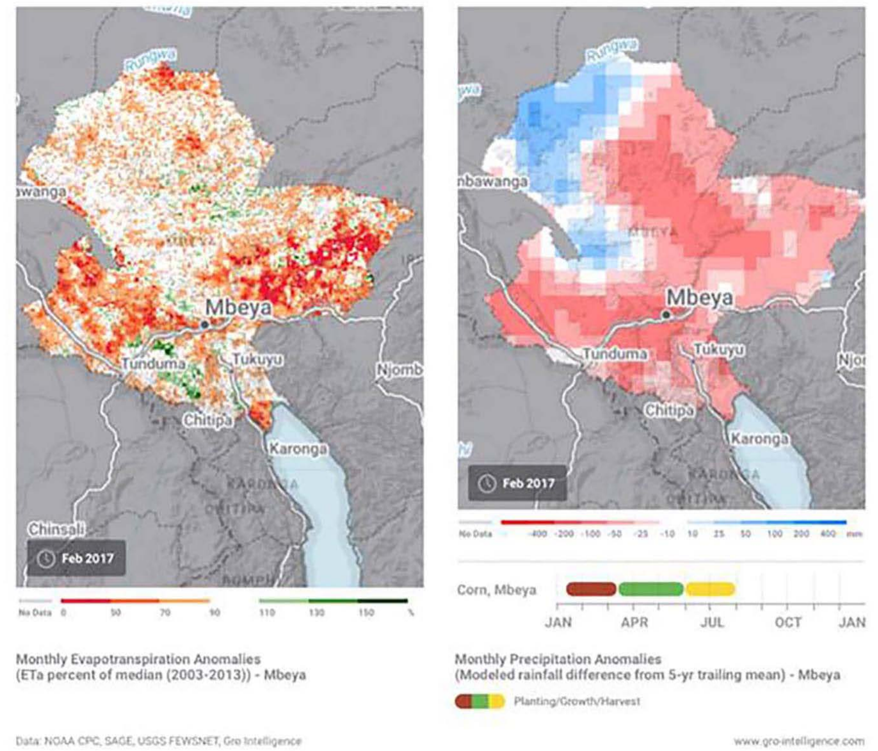
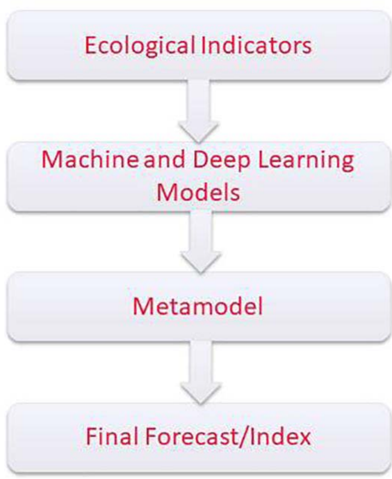


Figure 4. Mbeya, Kenya, 2017 drought

Tracking environmental indicators as they arise and evolve is critical in developing countries where subsistence farming prevails and farming families are vulnerable to even slight or short-term changes in weather or other factors that can have devastating effects. Figure 4 relates to the drought earlier this year in part of the corn belt in Kenya. We were monitoring that very closely for reasons which you'll see later on. Advances in information technology allow us to disseminate high-quality analyses of yield, drought and other parameters at much lower cost than previously possible, so we can bring the established benefits of modern modelling expertise to a hugely broader and more diverse audience.



The flow chart (right) shows our rubric for approaching the modelling task when we want to make a forecast. We feed ecological indicators into our system and its deep learning models, build a meta-model – which is an algorithm for picking which local model to use at any given time of year in any given place – and then generate a final forecast and an index.

We built a state-of-the-art US corn-yield model using our data platform and we have been giving it away for free, which is a first, at least for a private company, as far as I know. My guess is that this is at least as high in quality as anything else that is available, and certainly a lot cheaper than earlier models, which were priced at hundreds of thousands of dollars per year simply to provide the results. Our result is posted on our website every week. We currently (August) predict a national crop yield of around 164 bushels per acre this year (Figure 5), while the USDA's current estimate for the US corn crop is around 170 bushels per acre. To a corn-futures trader, that is a very significant difference which could cause significant price-movement.

About the company

Our company started in Nairobi, so we have a lot of contacts in Africa. We've been asking for various types of information to be collected or to be assimilated somehow into our database (Table 1), and some of that is happening. The more data and the better data you have, the better the models. Ultimately, the objective is to build models that help people to make decisions.

A survey of our users late last year shows that 35% of our users are still in Africa. The number of users in North America has been rising and at 38% is now similar to the number in Africa. We also have users in Europe (16%), Asia, South America and Oceania (5% or fewer). This is a very unusual beginning for an agricultural research firm. Most of them focus on either the US market which is where the global trading hubs are located, or the European or some other developed-country market. We hope interest in Africa will remain strong, and we maintain an office in Nairobi to support that.

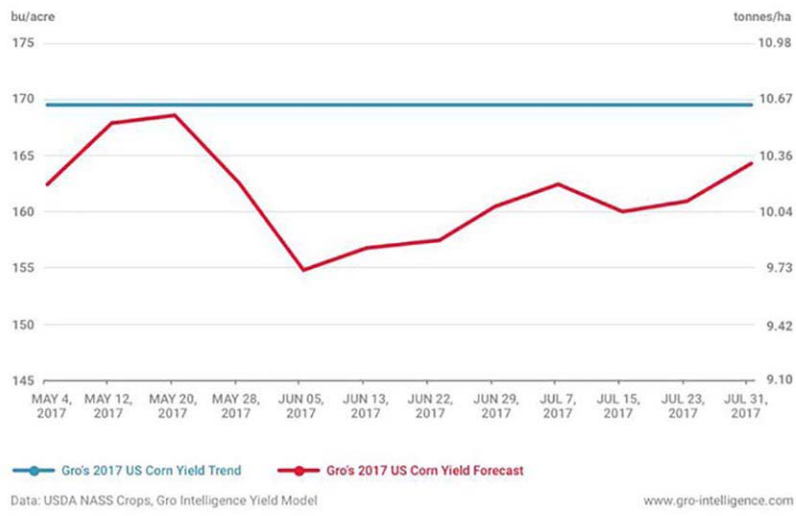


Figure 5. Gro's 2017 US corn yield forecast (lower, red, line) breaks higher

Table 1. Sample of data categories Gro believes would be critical to collect for African countries at a more frequent and detailed level

Agronomy	Production	Supply	Use	Prices	Economics
Ag equipment in use (c)	Seed production (a)	Imports (m)	Upstream – oilseed processing (m)	Inputs – fertilisers (a)	Land rent/ water rights (a)
Fertiliser use per acre (a)	Acres planted – county (a)	Stocks – province/ state level (q)	By type – maize, cornmeal (a)	Ag labour rates (a)	Total input costs (m)
Pesticide use per acre (a)	Acres harvested – county (a)	Stocks by type – grain, silage (q)	By use – seed, food, industrial (a)	Farm-gate (w)	Producer margin – by crop (a)
GMO seed use (a)	Yield – county (a)	Grain – cold storage capacity (a)	Per capita consumption (a)	Export prices (m)	Avg farm size (c)
Soil conditions (a)	Crop progress reports (m)	Grain shipments – deliveries (q)	Domestic gov't tenders (w)	Domestic freight rates (m)	Avg farm income (a)
Fallowed acres (a)	Harvest grading: ergot levels (a)	Grain handling capacity – port level (a)	Exports (m)	By-products – wholesale (d)	Avg farm debt (a)
Irrigated acres (a)	By-product = meat, feed (m)	Obsolescence – loss rates (a)	Retail expenditures (m)	Processed – retail (d)	Storage, insurance rates (m)

c = census. a = annual. m = monthly. w = weekly. d = daily. q = quarterly.

Eventually we aim to develop 'parametric risk measures' for the insurance market. These could be the basis of very precise credit and insurance products that will be critical to transforming agricultural markets around the world: for instance, by lowering the cost of capital and expanding access to credit. Better credit models can be developed by incorporating 'market risk' data into the models. Gro has access to data series that add information well beyond the existing credit history of a borrower:

- growing cycles
- climate and growing conditions
- perils (fire, flood, infestations, predation)
- prices of crops
- prices of other commodities
- national and regional programs and restrictions
- trade policies.

We all know that capital shortage is a problem for many smallholders and larger cultivation operations. Good information about risk and financial services, based on data and models that are easily accessible and understandable, should be of much greater value to smallholders than physical gadgets they buy that only work for a short time.

Fundamentally, agricultural data is like any form of critical infrastructure: it should be robust and meticulously well-maintained. While that will require effort and investment, good data, like good infrastructure, helps societies to thrive. Gro Intelligence economists, scientists and engineers are working to get more data, with more precision for more of agriculture.

Summary

In summary, by imposing structure on, and building models with, the increasingly bigger data that is available to the farm sector in Africa and elsewhere, we believe that we can bring the African agricultural community the data tools it needs to improve its own performance itself.

We believe history supports our belief that a set of solutions and methods, arrived at on the ground, locally, with the benefit of world-class data and models, will succeed well beyond any externally formulated program.

We believe that the best solution to the data problem in sub-Saharan Africa is not limited to sub-Saharan Africa: it's actually a global solution, and we believe that the free market has a significant role to play in helping people to make the decisions that only they know the various parameters of.

They can use a product, like Gro, not necessarily Gro, but something like that, to get the data that they need without us saying what they should be interested in, and I think that's of great value.

Steve Mathews is the Head of Strategy at Gro Intelligence, a software company focused on the global food and agriculture markets. Before joining Gro, Steve worked as a portfolio manager and the head of commodities research at Tudor Investment Corporation. During his tenure there, he developed extensive commodities analysis software and conducted practical study of agriculture, energy, and metals. He is a crop scout with the Pro Farmer Crop Tour each summer, and teaches agricultural hedging at the University of Memphis. Prior to finance, Steve commanded a tank company in the US Army. Steve holds a BS in Operations Research from USMA (West Point) and an MBA in Statistics from the Stern School of Business at NYU. He's currently about halfway through an MS degree in Agronomy from Iowa State University. He's also a holder of the Chartered Financial Analyst designation.