STARS —
From Rich Remote Sensing Data to Farmer-relevant Information

UT Data Science Colloquium
Enschede, April 20th, 2015

Rolf A. de By
Raul Zurita-Milla

HIGH TECH
HUMAN TOUCH

UNIVERSITEIT TWENTE.
Background

- Globally, already 1 in 2 people live in big cities; will be 2 in 3 by 2050. The world has a growing food challenge: need to produce 60% more by 2050.
- Global recognition of need for reliable, open and real-time information on the food production systems.
- Critical for agro-policies, stable markets, fight against food crises.
Background

- Globally, 3 in 4 poor people are rural smallholder farmers.
- Smallholders are not usually included in food production chains, let alone in value chains.
- Sub-Saharan Africa has 183 million hectares in use for agriculture; some 451 million hectares of arable land remain unused: this is more than half of the world’s remaining arable land resources.
Project synopsis

Spatial information is revolutionizing agriculture in high-income countries but not in low-income countries.

Here, important adoption barriers exist: heterogeneity.

**STARS:** coordinated effort to

- learn,
- identify opportunities, constraints & risks,
- test hypotheses

*around potential exploitation of RS technology in crop-based production systems and livelihoods of smallholders.*
Project facts

• Five partners: ITC, CSIRO, ICRISAT, University of Maryland, CIMMYT
• 20 months = 1 or 2 crop seasons
• Supported by DigitalGlobe & RapidEye data & knowledge contracts
• Started June 1, 2014
• Various subgrants/-contracts: Uni Sokoine, ESIPPS, Manobi, CEGIS, ...
• Links with other initiatives: VitalSigns, AfSIS, CIP, Cereal Systems for South Asia (CSISA), Drylands Systems CRP, National Food Security Department Tanzania
Objectives

• Better understand why RS has not been taken up for SHA, and which investments are required to unlock potential.

• *Demand-driven* experimental *use cases* where currently poor information is the norm, and where *RS-based workflows can help* improve.

• *Match private sector* partners (satellite data/spatial analysis/web&mobile companies) *with public sector* actors and public good objectives.
Main hypotheses

We think we can

• _monitor crop growth_ within the small farms of sub-Saharan Africa and Southern Asia, _using time-series remote sensing._

If achievable, _this will_

• allow _improved outlooks for crop yields_ throughout the season, _informing policy-makers_, and

• ensure more effective evidence-based advisory services at the farm scale, _informing farmers and agro-business._
Three facets

Stakeholder

Business

Technology

sustainable models of realization

information products that inform & transform agricultural processes

ground-based, airborne, and spaceborne monitoring throughout the crop season
Challenges

Heterogeneity in

• Crops and crop varieties
• Crop systems
• Soils and nutrient content
• Climatic conditions
• Farm field practices and consistency
  – Farm field boundaries
  – Tillage and planting system
  – Rain-fed vs irrigated fields
  – Use of fertilizers
  – Use of mechanization
Challenge of intercropping

**Intercropping** =
Growing multiple crops on the same land, with purpose to create larger yields, or mutual support of crops grown.

- **Principle**: Crops should not compete for space, nutrients, water, or sunlight. Sometimes companion crops can help each other (support, nutrients, pest control).
- **Examples**: deep root + shallow root, tall + short, light-loving + shade-loving
- Crop separation can be done in 2d/3d space and/or in time.
High dimensionality of problem space
Compared to high-income ag

- Smallholder farming in Africa/Asia is a \textit{data-poor} context; there are many facets of the production systems that we do not know.
- Tanzania 2008 maize bumper crop
- Much ground truthing required
- Build an infoconomy with the farmer as active partner?
Where we work

Stakeholder approaches:
- bottom-up
- top-down
- mid-level

two 10×10 km landscapes
four 10×10 km landscapes
six 10×10 km landscapes
STARS Data collection

Satellite data

DigitalGlobe
- Biweekly, WorldView-2 & -3
  8-spectral band, 2m resolution

RapidEye/Blackbridge
- Biweekly, 5-spectral band, 5m resolution

Challenges
- Cloud cover
- Off-nadir angle
- Pre-processing
- Analytical processing
- (Licenses)

WorldView-2 image of the Mali study site, growing season May-November 2014
WorldView2 & 3 VNIR and CAVIS VNIR bands are optimized for different missions.
Soil conditions

Dry soil

Wet soil
Plant health

Plant health
STARS Data collection

UAV data

eBee
• GR+NIR 12 Mp Canon camera, 3.5cm resolution (5cm vertical)
• GRRe+NIR 4-band 1.2Mp multiSPEC camera, 10cm resolution

Geo X-8000 Octocopter
• Tetracam miniMCA 5/6-band 1.3Mp, 10cm resolution
• RGB Sony NEX-7 24 MP, 2cm resolution
• OPTRIS PI 400 thermal camera, 1.1 Mp, 15 cm resolution

Challenges
• Hw robustness, stitching, calibration
STARS data collection

Field campaigns (in Mali)
STARS Data collection

Field data

• Common set of essential variables
  – Plant density
  – Plant height
  – LAI (PocketLAI)
  – FCover (CAN-EYE)
  – Phenology (BBCH scale)

Baret et al. (2010) Agr. For. Met. 150, 1393-1401
STARS Data collection

- F-Cover
Growth Stages in Cereals

- **Stage 1**: One shoot begins
- **Stage 2**: Tillering begins
- **Stage 3**: Tiller formed
- **Stage 4**: Leaf sheaths strengthen
- **Stage 5**: Leaf sheaths strongly erected
- **Stage 6**: First node of stem visible
- **Stage 7**: Second node visible
- **Stage 8**: Last leaf just visible
- **Stage 9**: Ligule of last leaf just visible
- **Stage 10**: In boot
- **Stage 10.1**: Flowering
- **Stage 10.5**: Flowering
- **Stage 11**: Ripening

- **Tillering**
- **Heading**
- **Ripening**
Managing data

• Raw image data as files
• Derived data must live in a database
• CSSL v1.0 in Postgresql/PostGIS
• CSSL v2.0 in MonetDB/SciQL
Public good outcomes

• Landscaping study — CSIRO & partners
  Aim to understand
  – the decision-making environment for key stakeholders
  – the pathways for agricultural development that are likely to emerge
  – the nature of the infrastructure systems needed for delivering the right information to the right stakeholders at the right time.

• Crop Spectrotemporal Signature Library
  – Spectral info for crops followed over time
  – Accompanying farm field data

• Image analysis algorithm repository
  – Data ingestion workflows
  – Analytical workflows
High dimensionality of problem space
Multidimensional RS

- Multi-platform
- Multi-sensor
- Multi-resolution
- Multi-temporal
High dimensionality of problem space (2)

Any sufficiently advanced technology is equivalent to magic
– Arthur C. Clarke
STARS (big) data science

Data science, due to its interdisciplinary nature, requires an intersection of abilities: **hacking skills**, **math and statistics knowledge**, and **substantive expertise** in a field of science.

- **Hacking skills** are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.
- **Math and statistics knowledge** allows a data scientist to choose appropriate methods and tools in order to extract insight from data.
- **Substantive expertise** in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.
- **Traditional research** lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.
- **Machine learning** stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.
- **Danger zone!** Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.
Crop Spectrottemporal Signature Library

\[ X^2_{(\text{wavelength-}\lambda,\text{JulianDay})} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1jd} \\ x_{21} & x_{22} & \cdots & x_{2jd} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\lambda1} & x_{\lambda2} & \cdots & x_{\lambda jd} \end{bmatrix} \]

(Viera et al. 2000)
Crop Spectrotemporal Signature Library (2)

Dealing with the:

- Specifics of the agricultural fields (small, irregular structure and, often, mixed crops)
- Variability in plant, soil & climatic/management factors
- Variability of crop phenological stage, genotype, health
- Variability in the data: multi.sensor (SNR and clouds)
Crop Spectrotemporal Signature Library (3)

UAV/Formosat-2 crop spectral profile, MSc thesis Caroline Gevaert (EOS/ITC) Lund University, 2014
STARS algorithms

- Physical retrievals: Radiative transfer models
- Statistical retrievals: (linear) regression

Dorigo, Zurita-Milla et al., 2007
STARS algorithms (2)

Machine learning
• Unsupervised
• Supervised
• Semi-supervised
Supervised learning

• Classification
  – Crop type = f(RS, Env. data)

• Regression
  – Crop status = f(RS, Env. data)

• Field data is key
  – RS is not a miracle tech
  – Crowdsourcing (farmers, citizens, scientists)

(Walsh et al., in prep)
Supervised learning: classification

Abstract
We evaluate 179 classifiers arising from 17 families (discriminant analysis, Bayesian, neural networks, support vector machines, decision trees, rule-based classifiers, boosting, bagging, stacking, random forests and other ensembles, generalized linear models, nearest-neighbors, partial least squares and principal component regression, logistic and multinomial regression, multiple adaptive regression splines and other methods), implemented in Weka, R (with and without the caret package), C and Matlab, including all the relevant classifiers available today. We use 121 data sets, which represent the whole UCI data base (excluding the large-scale problems) and other own real problems, in order to achieve significant conclusions about the classifier behavior, not dependent on the data set collection. The classifiers most likely to be the bests are the random forest (RF) versions, the best of which (implemented in R and accessed via caret) achieves 94.1% of the maximum accuracy overcoming 90% in the 84.3% of the data sets. However, the difference is not statistically significant with the second best, the SVM with Gaussian kernel implemented in C using LibSVM, which achieves 92.3% of the maximum accuracy. A few models are clearly better than the remaining ones: random forest, SVM with Gaussian and polynomial kernels, extreme learning machine with Gaussian kernel, C5.0 and avNNet (a committee of multi-layer perceptrons implemented in R with the caret package). The random forest is clearly the best family of classifiers (3 out of 5 bests classifiers are RF), followed by SVM (4 classifiers in the top-10), neural networks and boosting ensembles (5 and 3 members in the top-20, respectively).
Random forest

• “Big brother” of decision trees
• Leo Breiman discovered (~2000) that accuracy improves by using ensembles of trees
• Each tree grown in a “random” fashion

“Taming E-mail” Decision Tree

Receive New E-mail

3-minutes or less?
DO IT!!

Longer than 3-minutes?
TASK IT!!

Once DONE or TASKED, FILE or DELETE!!

Going Paperless? Print to PDF?

If no file folder? CREATE IT!!

Can be done by Printing or “Drag & Drop

Every e-mail received can be handled this way!
Random forest (2)

- Input data: N training cases each with M variables
- n out of N samples are chosen with replacement (bagging).
- Rest of the samples to estimate the error of the tree (out of bag)
- m << M variables are used to determine the decision at a node of the tree
- Each tree is fully grown and not pruned
Random forest (3)

- RF are fast and easy to implement.
- They yield highly accurate predictions (even if the input data has a high dimensionality)
- No overfitting
- Provide insight on the importance of each attribute/feature/dimension
- They are easily parallelizable
- Data does not need pre-processing
SVM

• A classifier derived from statistical learning theory by Vapnik et al. in 1992

Kernel’s trick

$$\Phi: x \rightarrow \varphi(x)$$
SVM (2)
Semi-supervised learning
STARS algorithms
Ill-posed problem

1. The problem must have a solution
2. The solution must be unique
3. The solution must be stable under small changes to the data

(Hadamard, 1923)
Ill-posed problem (2)

(Zurita-Milla et al., 2015)
Ill-posed problem (3)

(Zurita-Milla et al., 2015)
Ill-posed problem

STARS data collections

• Multi-RS
• Field data
Bonus: cloud computing

Google Green-Wave project
Modelling Spring onset at continental scales
May, 22 + June, 26 + July, 29 + Sept, 29 + Oct, 18 + Nov, 14
Take home message

• Moving from rich RS data to farmer-relevant information requires solving a high-dimensional problem

• STARS has a high-dimensional dataset
• The STARS team is multi-disciplinary
