Scalable Information Systems for Ag in Africa and Beyond

November 17, 2016

William Wu
QED | http://qed.ai

QUANTITATIVE ENGINEERING DESIGN

OUTLINE

- **Introduction**
- **Problems: Data Scarcity**
- **Technology Solutions**

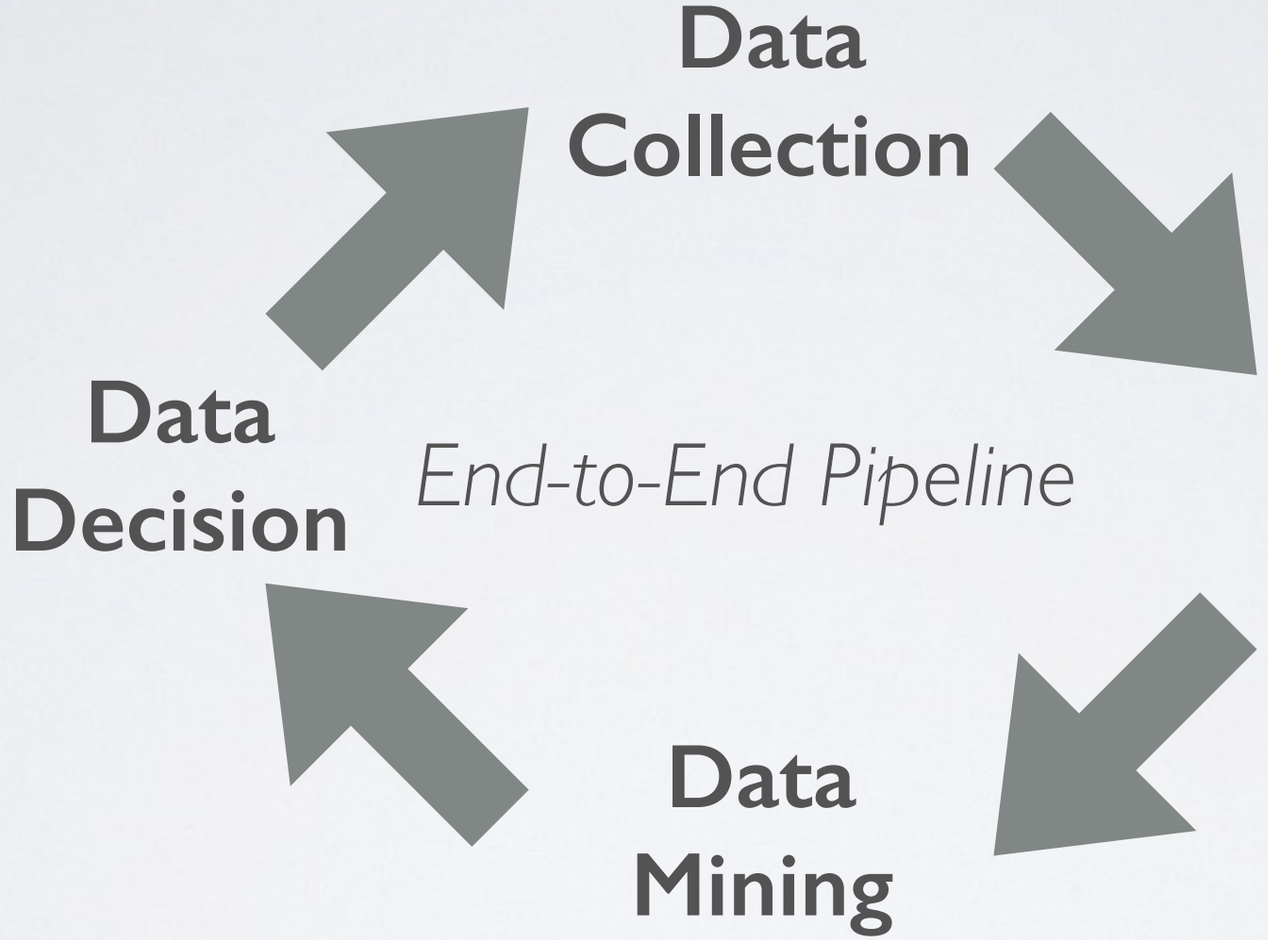
INTRODUCTION

William Wu

Ph.D. EE, M.S. Math (Stanford)
B.S. EECS (Berkeley)
Ex-NASA JPL

Jiehua Chen ("JC")

Ph.D. Statistics, M.S. Econ (Stanford)
B.Sc Math and Physics (Tsinghua)
Ex-Columbia Univ., Earth Institute



QUANTITATIVE ENGINEERING DESIGN

"Deploy or die!" — MIT

***"quod erat demonstrandum"* — Euclid**

Global Development

"Help the world's poorest people lift themselves out of poverty."

Ag Dev

- 3/4 of world's poorest subsist on ~0.5 ha of farmland.
- unproductive soils, plant disease and pests, drought, ill livestock, credit, market access, pricing, land reform
- malnutrition, child mortality

BILL & MELINDA GATES foundation

GLOBAL DEVELOPMENT

Agricultural Development

Emergency Response

Family Planning

Financial Services for the Poor

Global Libraries

Integrated Delivery

Maternal, Newborn & Child Health

Nutrition

Polio

Vaccine Delivery

Water, Sanitation & Hygiene

BILL & MELINDA GATES foundation

AfSIS

soil + landscape info is incomplete and outdated

TAMASA
ACAI
CSISA

agronomy projects for maize, cassava, cereals, etc.

ONE ACRE FUND

inputs and advice to support ~400K SHFs in SSA

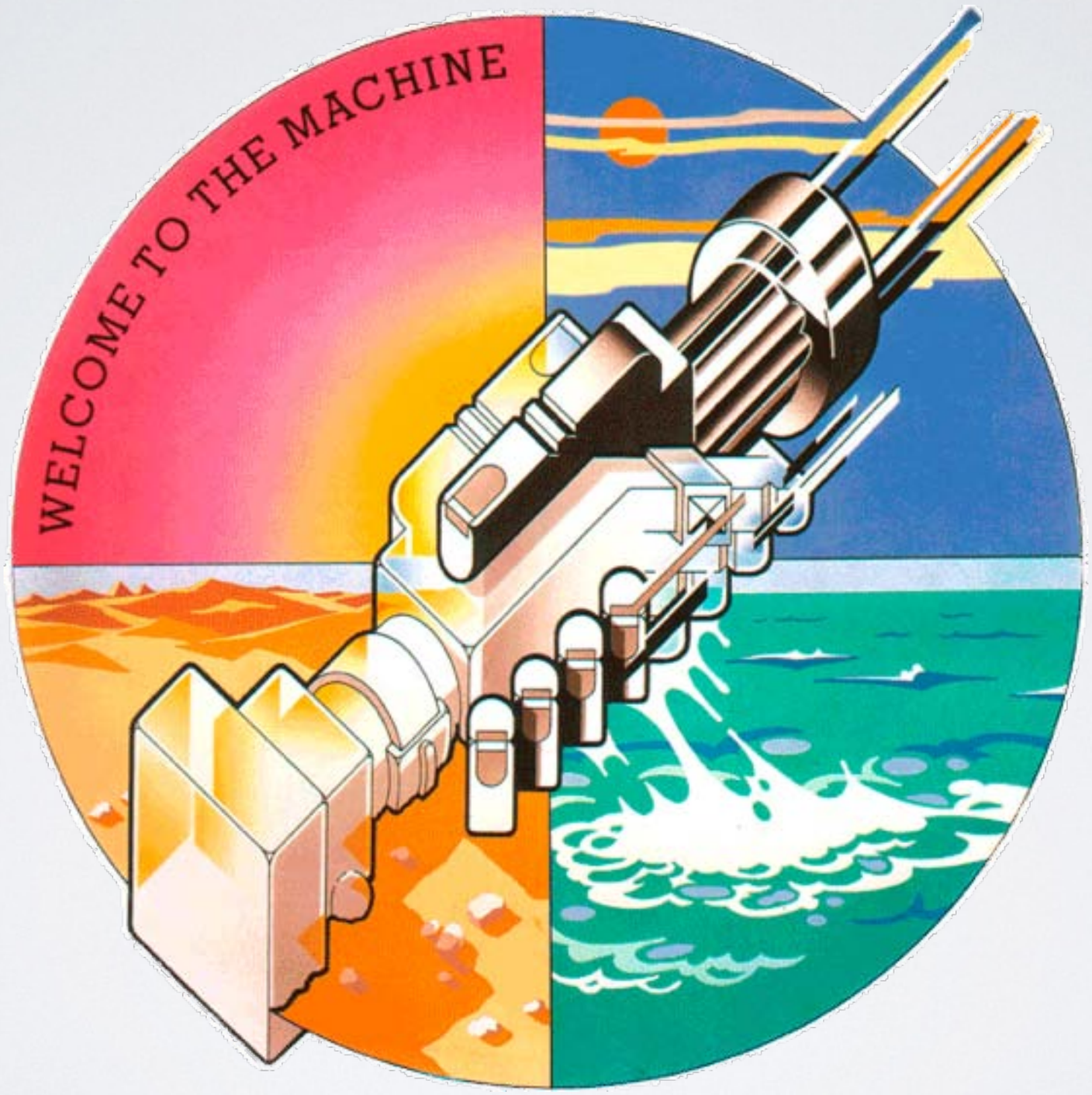
qed

tech consultant

qed

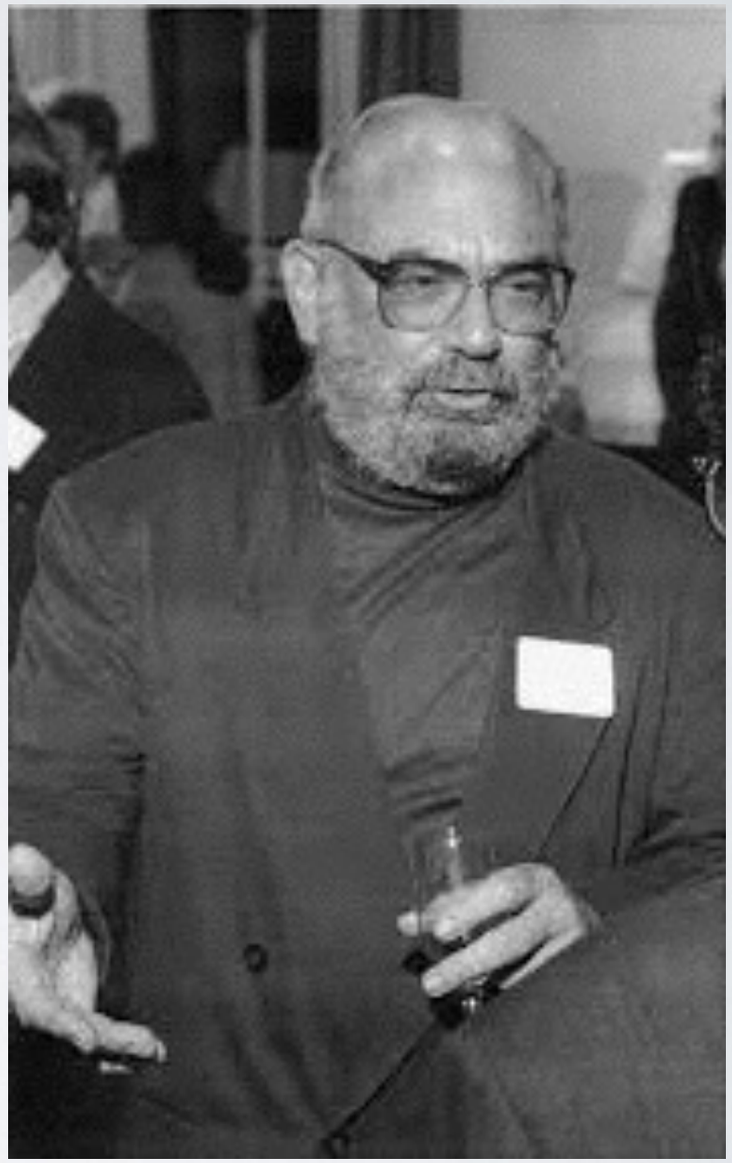
PROBLEMS: DATA SCARCITY

"WELCOME TO THE MACHINE"

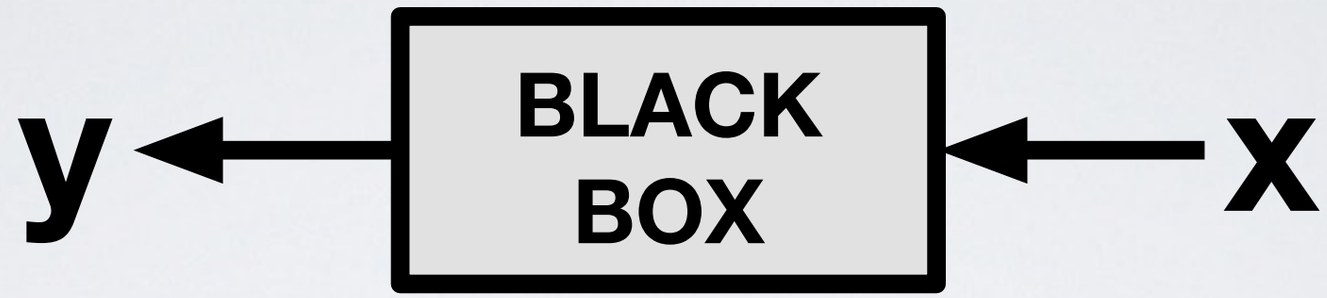


LEO BREIMAN (1928-2005)

- **Left academia to be full-time consultant in industry (13 yrs).**
- **"Statistical Modeling: The Two Cultures" (2001)**



"Goal: Use data to predict and to get information about the underlying data mechanism."



PREDICT

NATURE

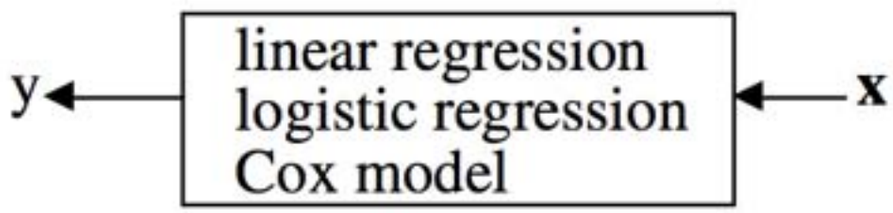
DATA

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = f(predictor variables, random noise, parameters)

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.

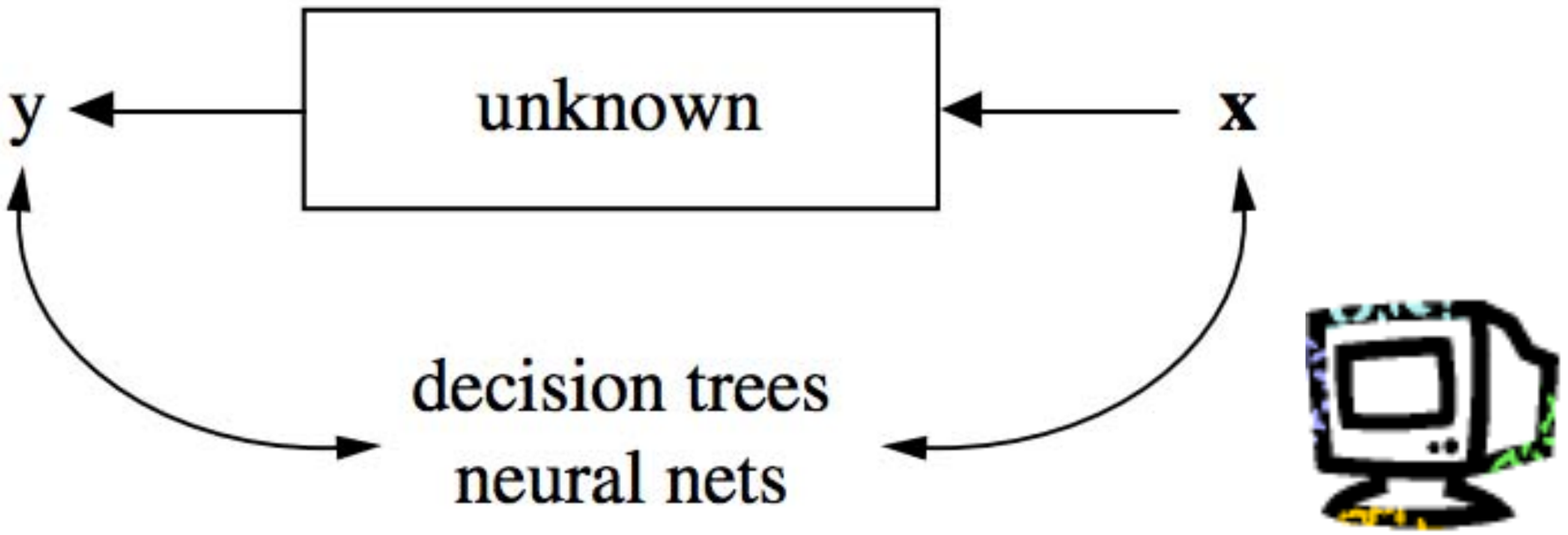
Estimated culture population. 98% of all statisticians.

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

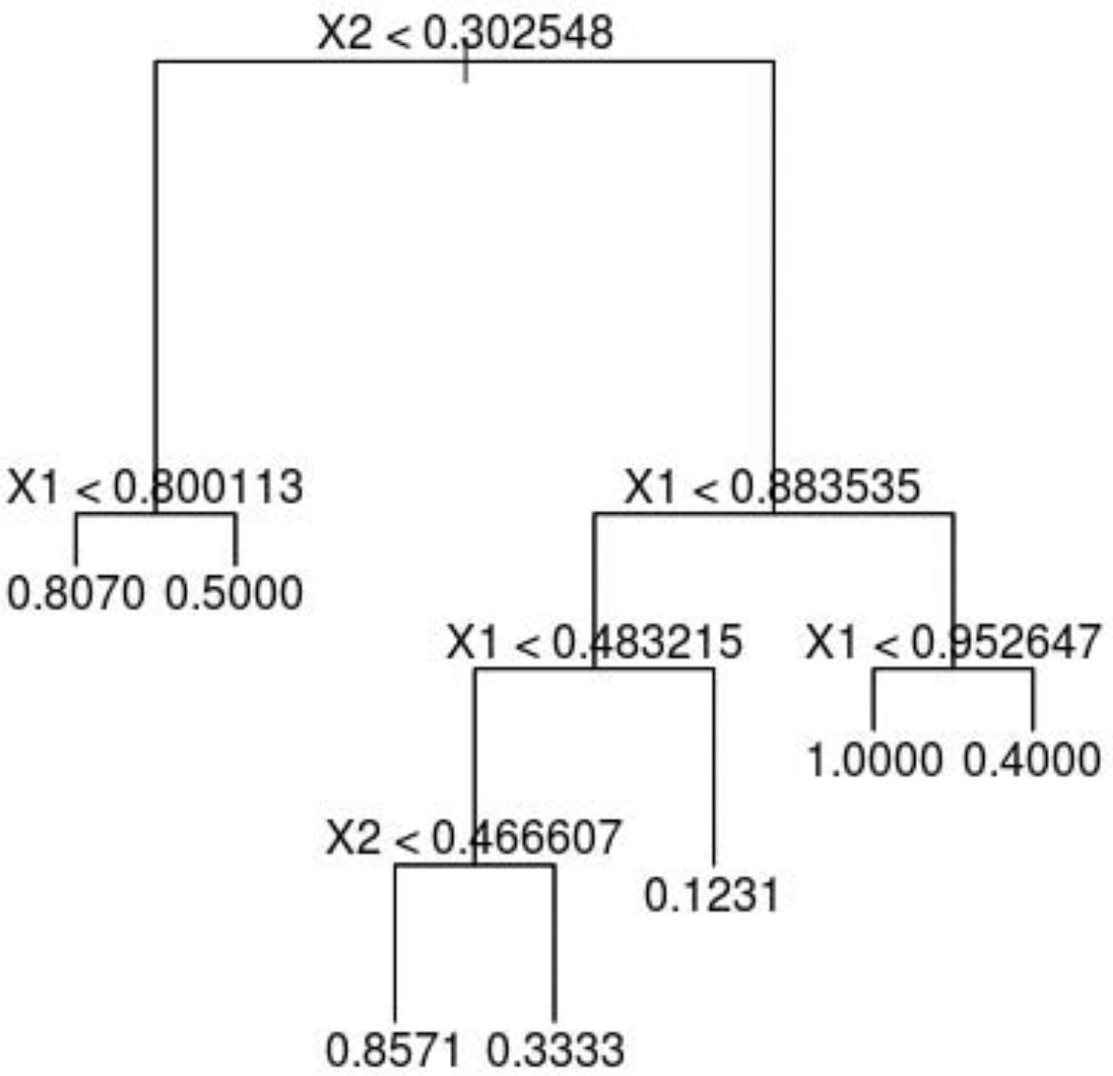
$$\log \frac{p(x)}{1 - p(x)} = \beta_0 + x \cdot \beta$$

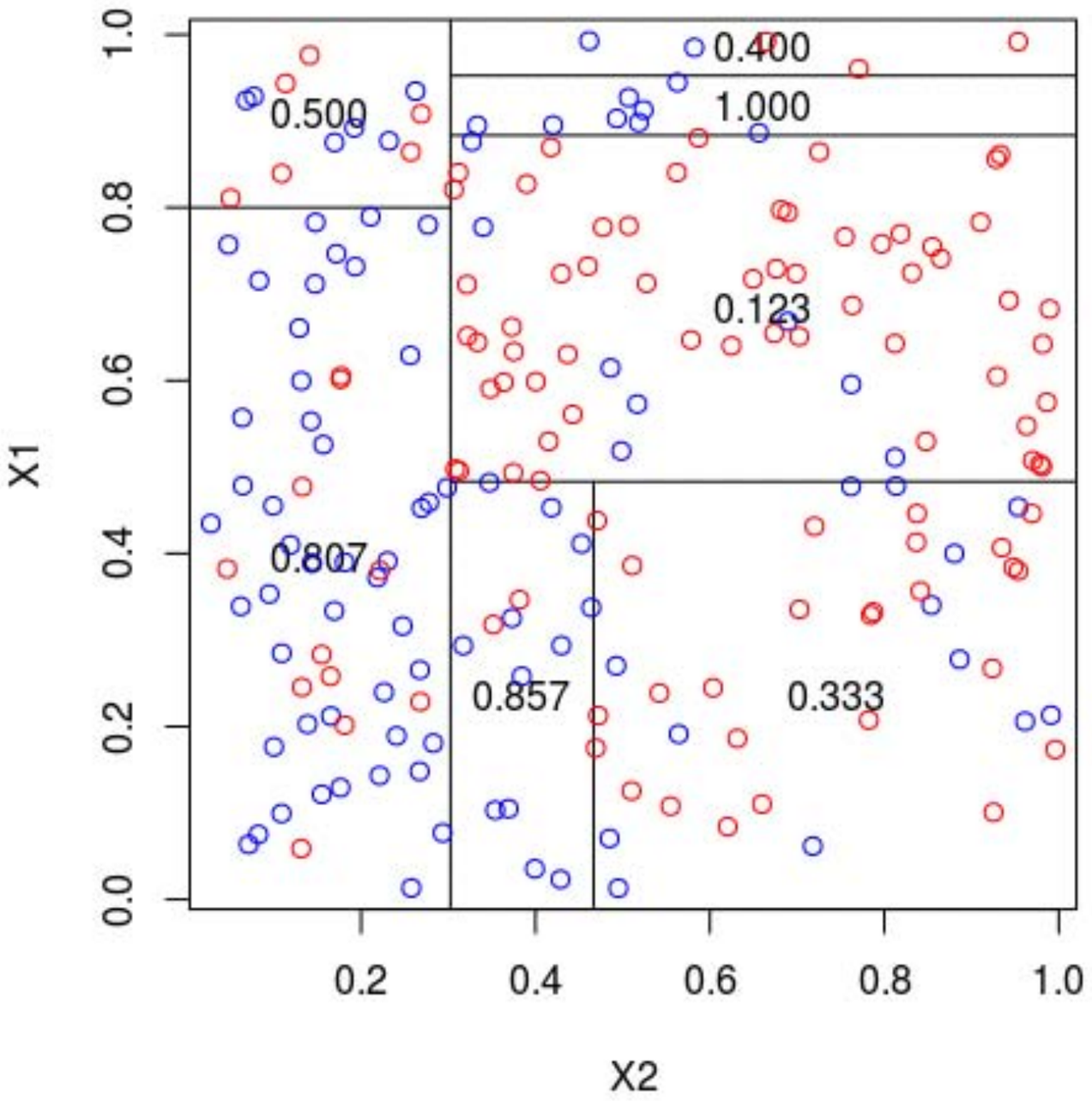
The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on $\mathbf{x}$ to predict the responses $\mathbf{y}$. Their black box looks like this:

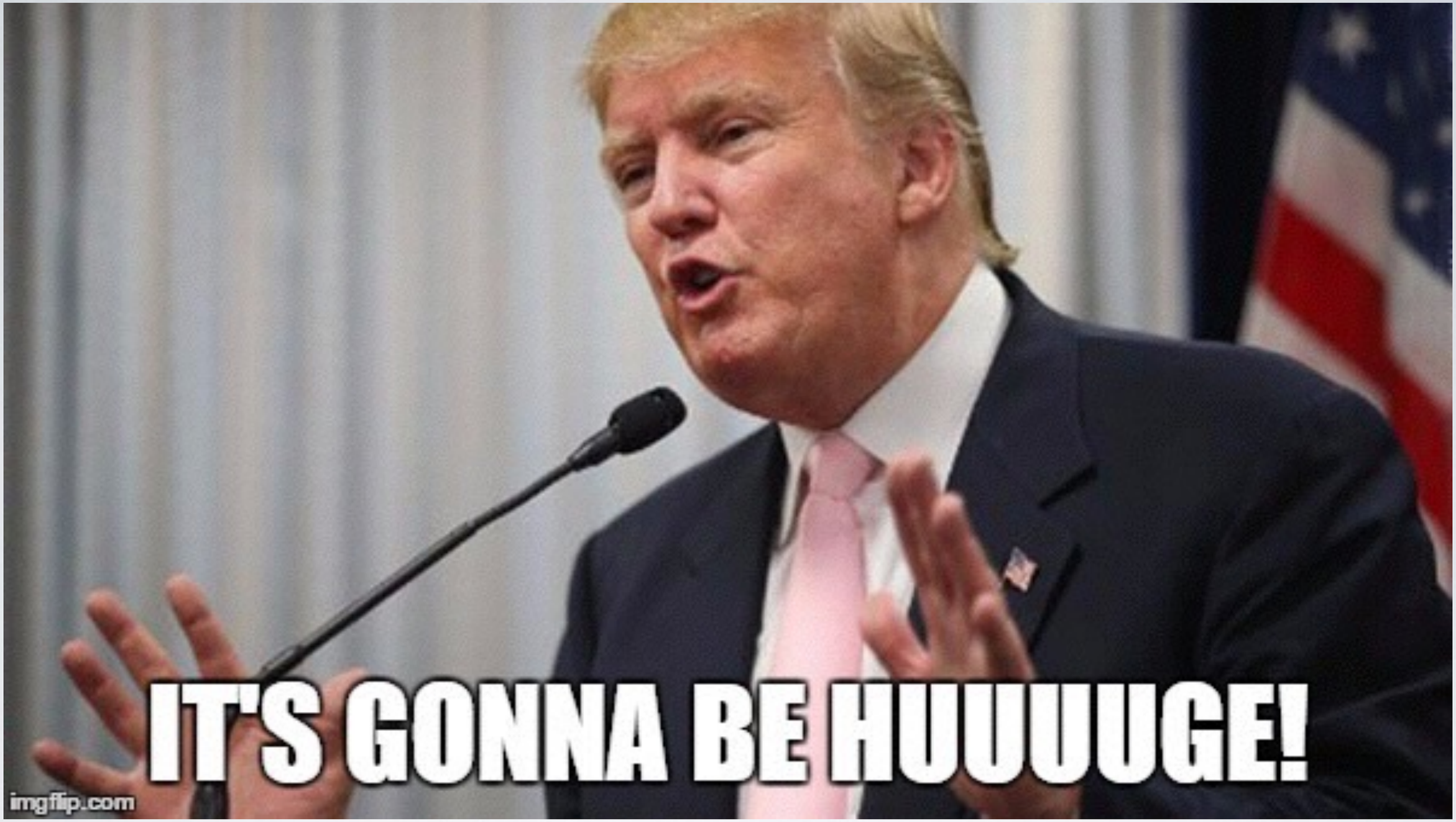


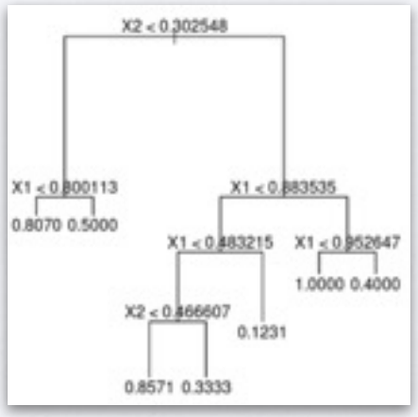
Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

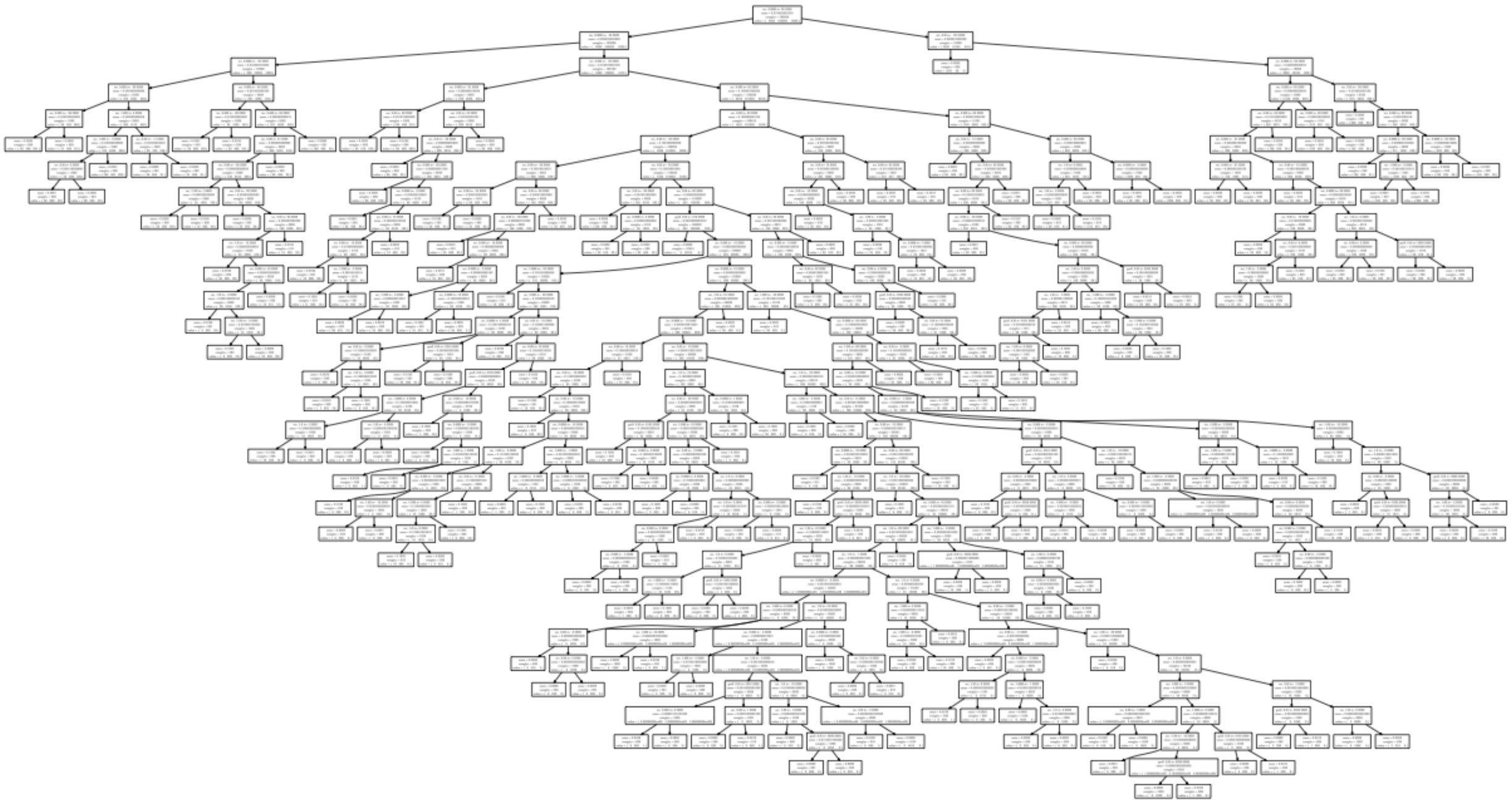




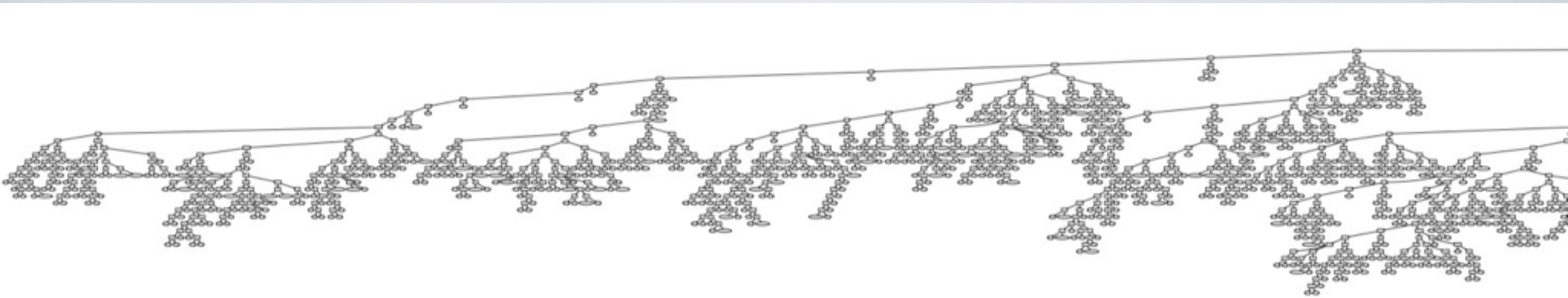
- **Decision Tree: repeatedly partition data into nested classes**
- **Random Forest: make many decision trees from random subsets of the data, and average them**

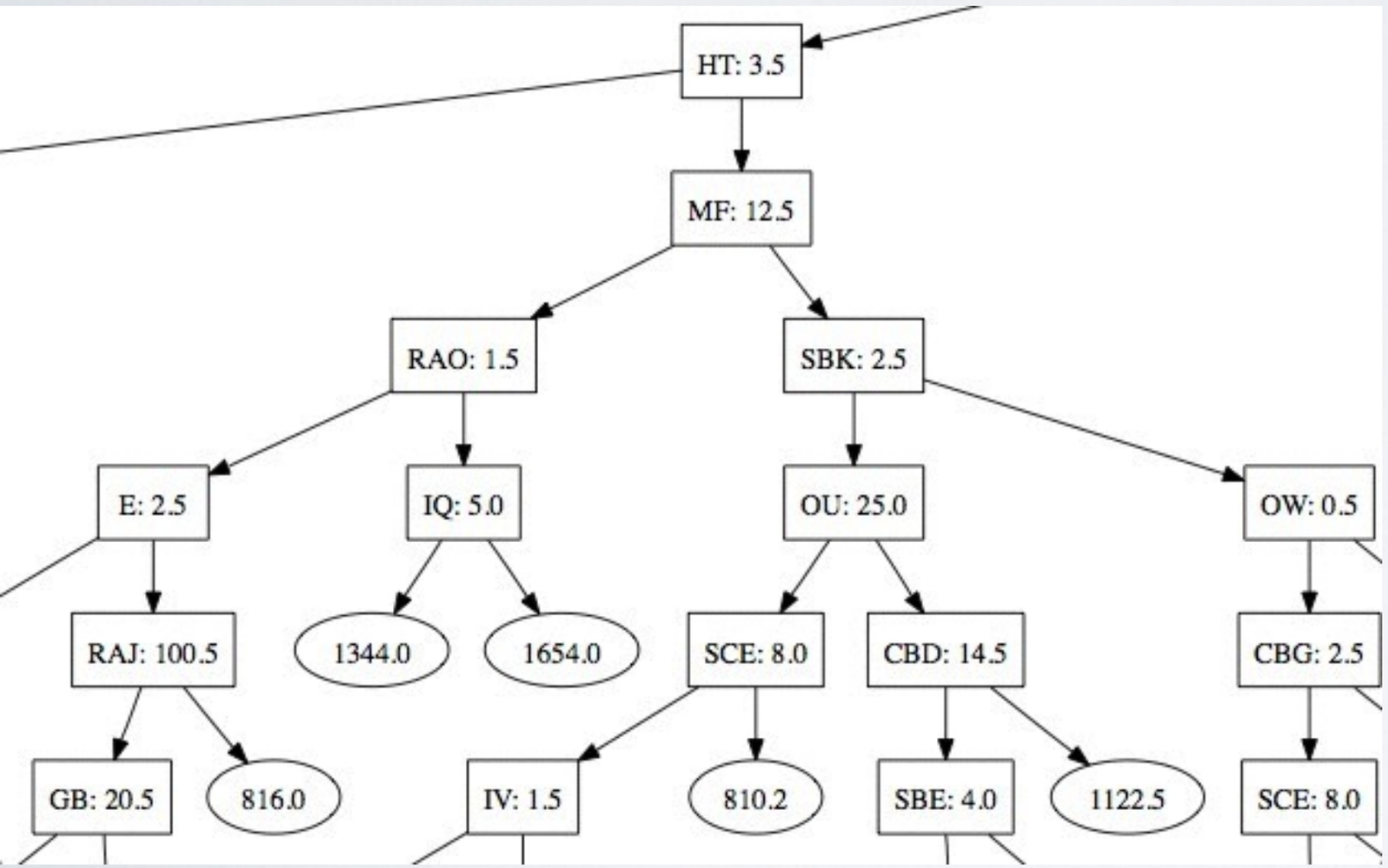






ex: predictive image compression (W.Wu)

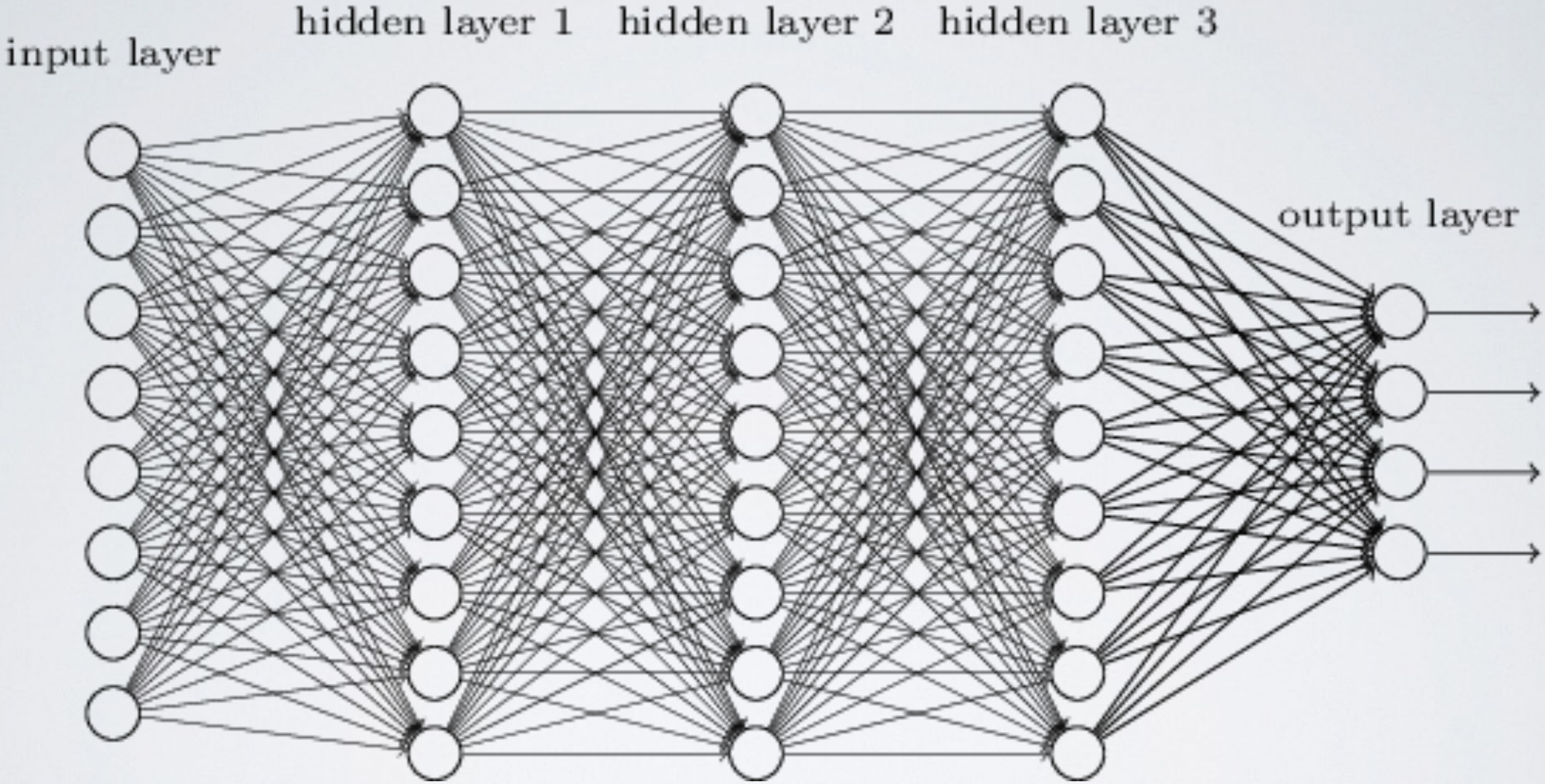




ex: real estate (QED)

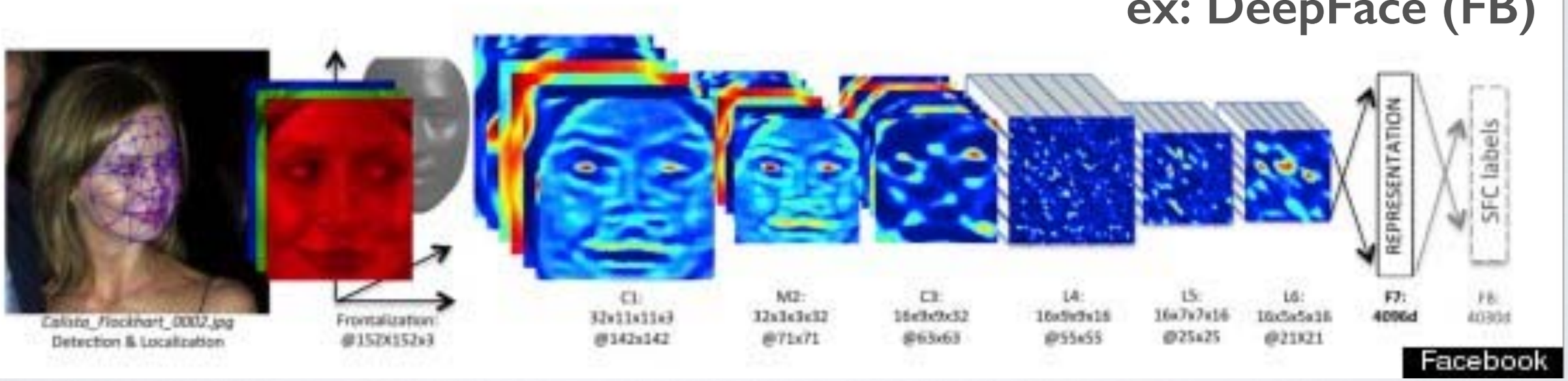
HT: Hotels
MF: Manufacturing
RAJ: Retail trade
SBK: Specialty Food Stores
…

neural networks (1953)



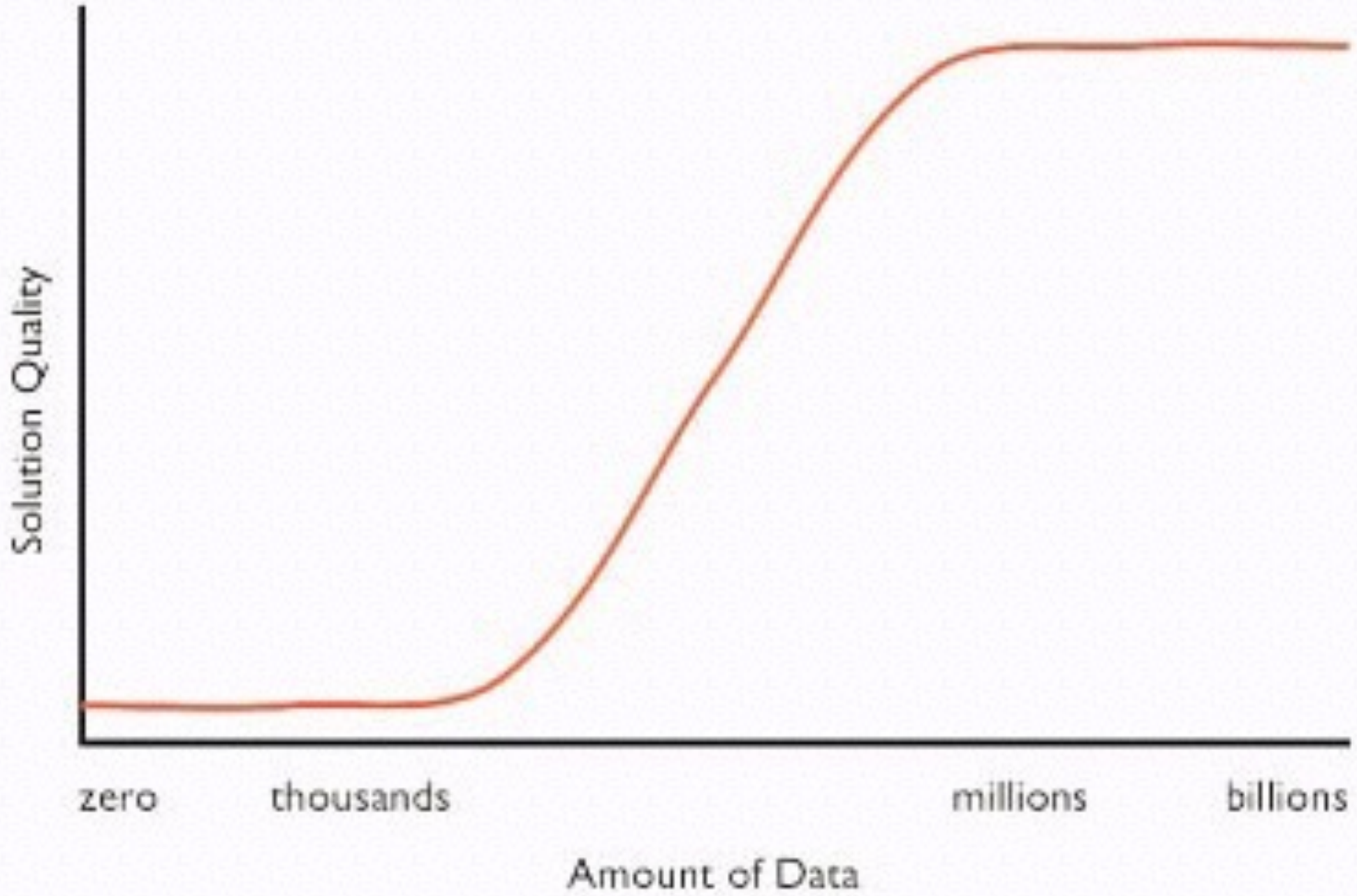
- 2016: Machines beat traditional data models and *humans*:
  - handwriting transcription
  - movie recommendation
  - spam classification
  - online advertising
  - insurance policies
  - breast cancer analysis
  - face detection
  - atari, chess, GO

ex: DeepFace (FB)



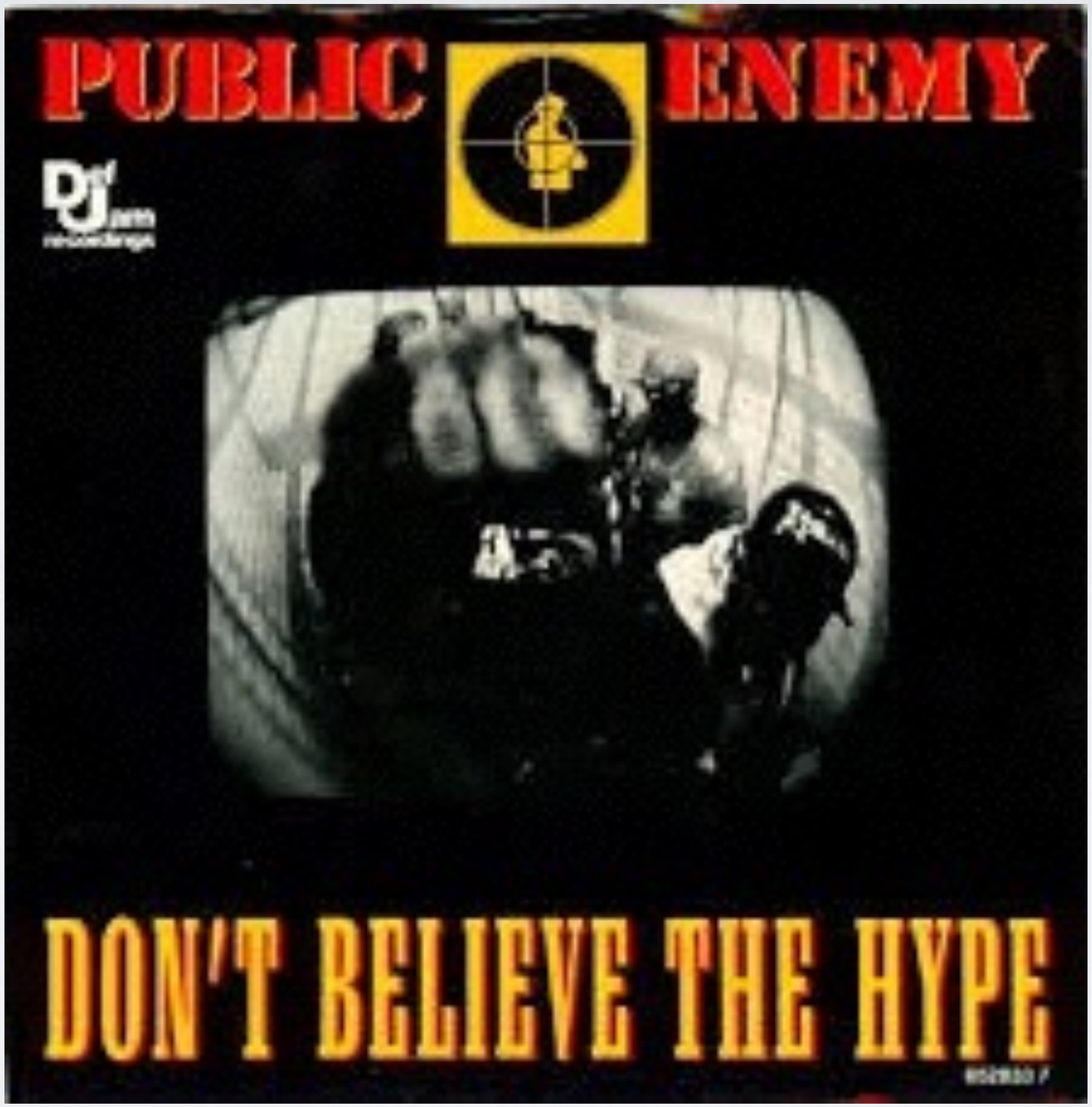
- What's the **catch**?

Peter Norvig (Google Research)



- **Requires ENORMOUS data (TB/PB).**

“DON’T BELIEVE THE HYPE”



AG FANTASY

- **food security (“9B by 2050”)**
- **small-holder farmers**
- **“big data”**
- **machine learning and maps**

AG REALITY

- Hans Jenny:

  s = f(**cl**, **o**, **r**, **p**, **t**, …)

- **yield** = **G** x **E** x **M** x **F**

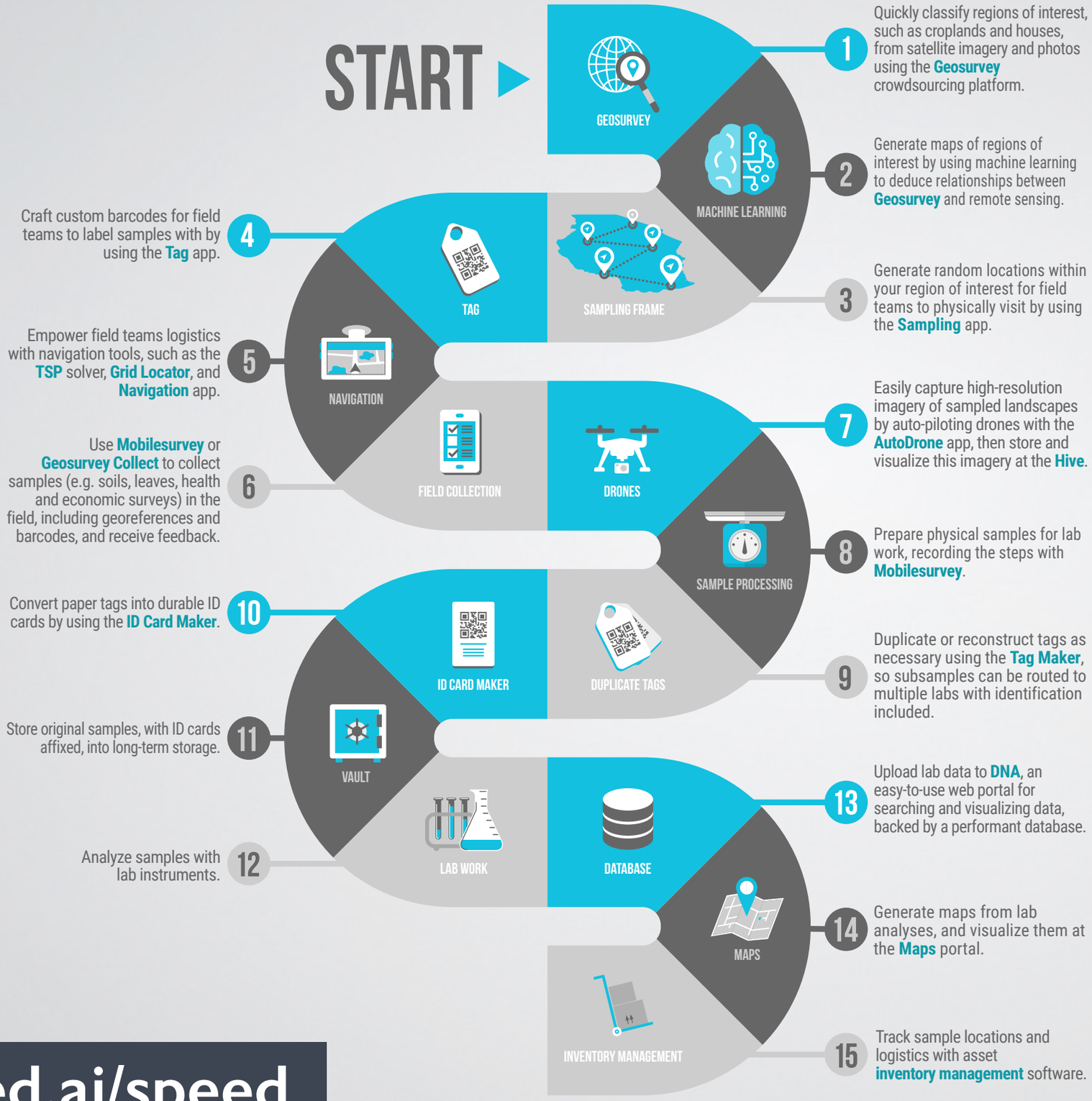
- SHF input = NA
- “databases” = .doc, .xls
- black box = Excel

SOLUTIONS:

SPEED

SCALABLE PIPELINE FOR ENGINEERING OF ENORMOUS DATA

End-to-end technology solutions to support systematic geospatial data collection and analysis.



1 Quickly classify regions of interest, such as croplands and houses, from satellite imagery and photos using the **Geosurvey** crowdsourcing platform.

2 Generate maps of regions of interest by using machine learning to deduce relationships between **Geosurvey** and remote sensing.

3 Generate random locations within your region of interest for field teams to physically visit by using the **Sampling** app.

4 Craft custom barcodes for field teams to label samples with by using the **Tag** app.

5 Empower field teams logistics with navigation tools, such as the **TSP** solver, **Grid Locator**, and **Navigation** app.

6 Use **Mobilesurvey** or **Geosurvey Collect** to collect samples (e.g. soils, leaves, health and economic surveys) in the field, including georeferences and barcodes, and receive feedback.

7 Easily capture high-resolution imagery of sampled landscapes by auto-piloting drones with the **AutoDrone** app, then store and visualize this imagery at the **Hive**.

8 Prepare physical samples for lab work, recording the steps with **Mobilesurvey**.

9 Duplicate or reconstruct tags as necessary using the **Tag Maker**, so subsamples can be routed to multiple labs with identification included.

10 Convert paper tags into durable ID cards by using the **ID Card Maker**.

11 Store original samples, with ID cards affixed, into long-term storage.

12 Analyze samples with lab instruments.

13 Upload lab data to **DNA**, an easy-to-use web portal for searching and visualizing data, backed by a performant database.

14 Generate maps from lab analyses, and visualize them at the **Maps** portal.

15 Track sample locations and logistics with asset **inventory management** software.

QED APP LEGEND

- 1 geosurvey
- 3 sampling
- 4 tag
- 5 navigation
- 5 grid locator
- 5 tsp
- 7 uav:quadcopter
- 7 uav:fixed-wing
- 6 8 mobilesurvey
- 6 8 geosurvey collect
- 9 tag maker
- 10 id card maker
- 13 dna
- 14 maps
- 15 inventory

qed.ai/speed

SUB-PROBLEMS

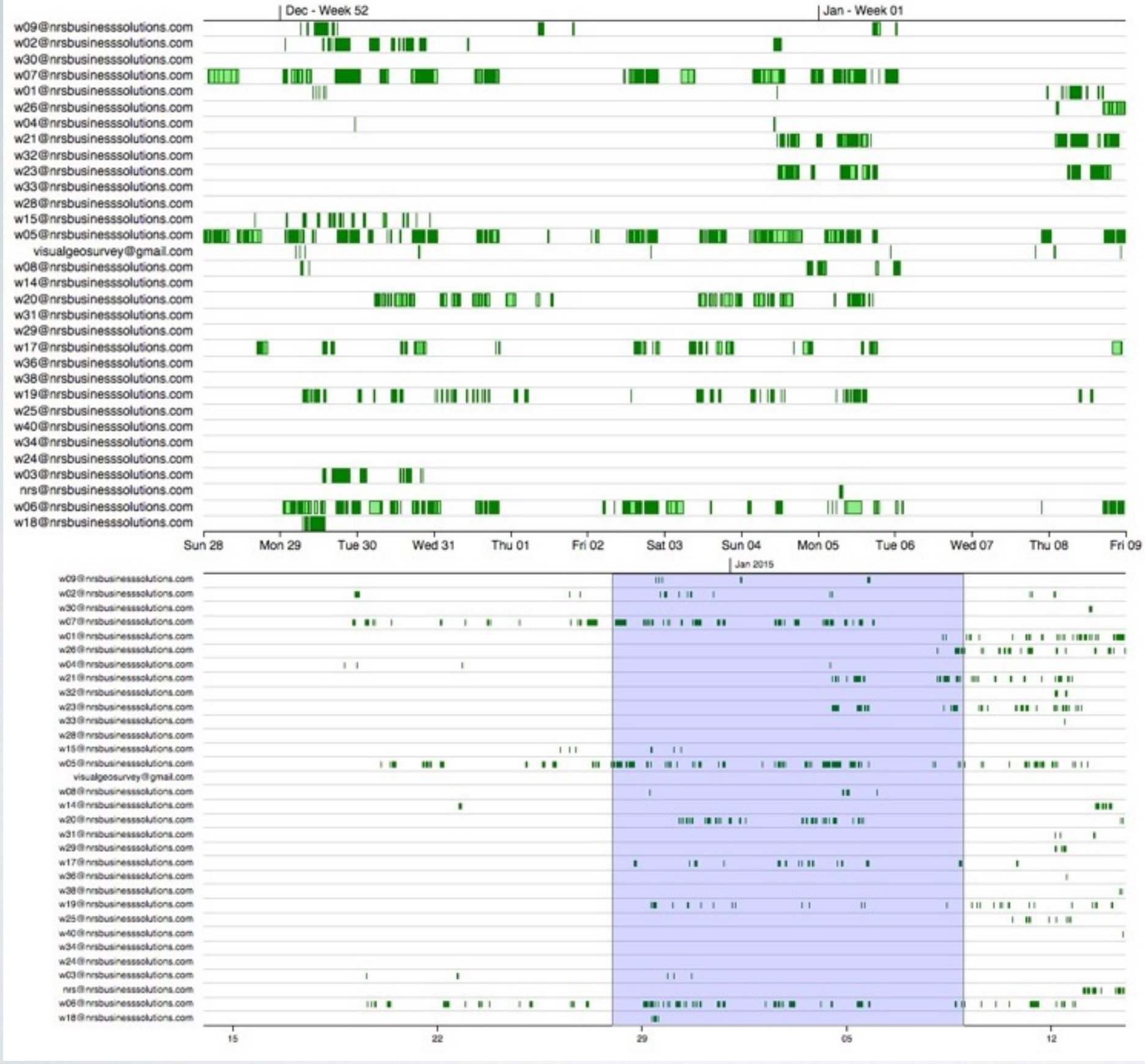
- **determine ROI**
- **sampling frame**
- **navigate to sampling sites**
- **georeferenced extraction**
- **lab analyses**
- **database**
- **mapping (spatial-temporal modeling)**
- **decisions**

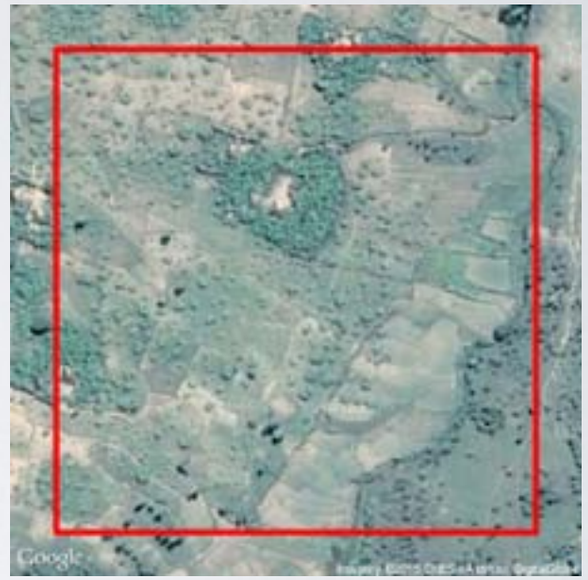
GEOSURVEY

- **problem:**
  - **map croplands and HS for soil sampling**
  - **Govt. Statistics Division story**
- **solution: crowdsourced geospatial analysis**

https://geosurvey.qed.ai/admin/survey/resultchange/839878/

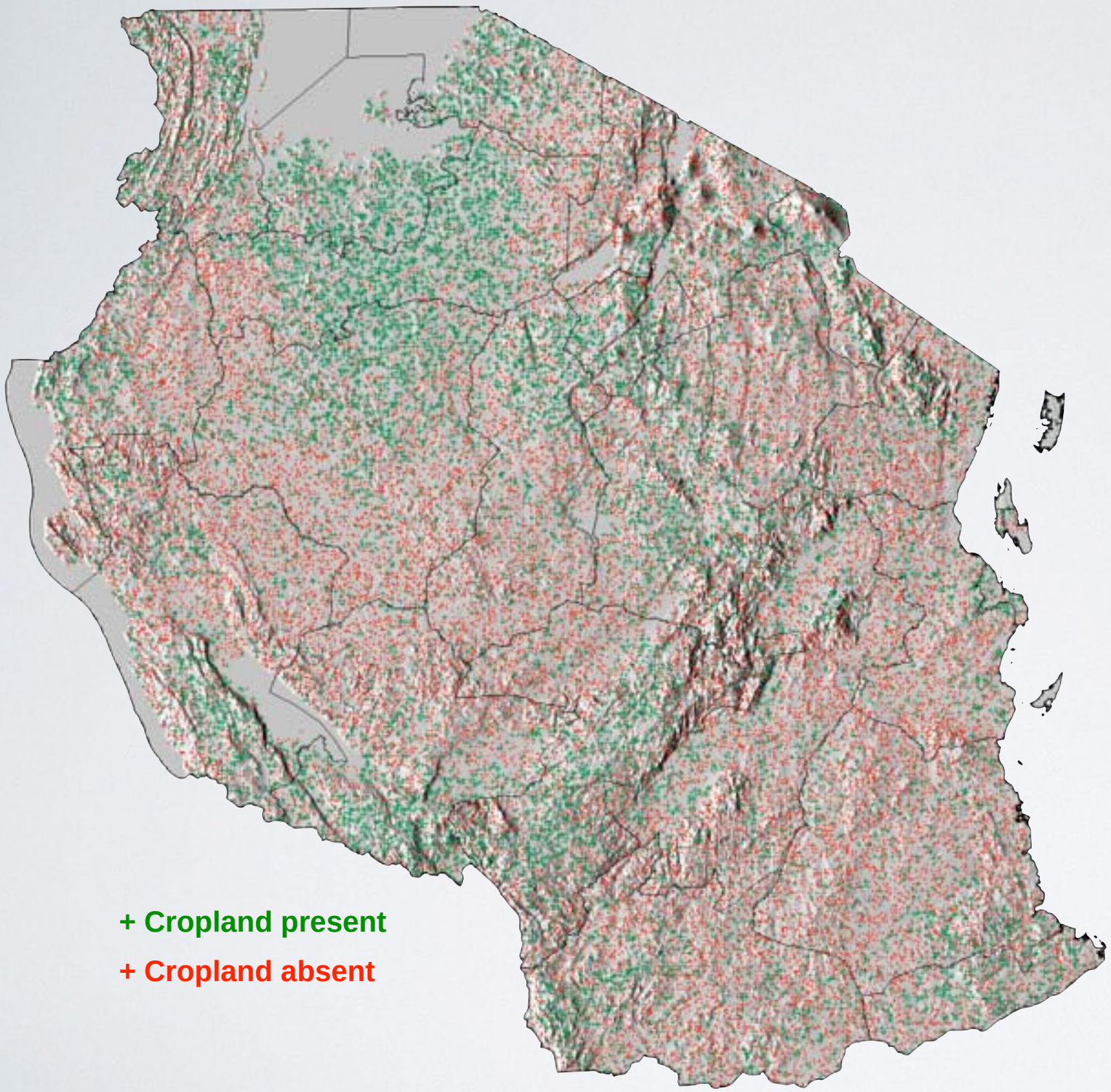
https://geosurvey.qed.ai/admin/survey/resultchange/59472/

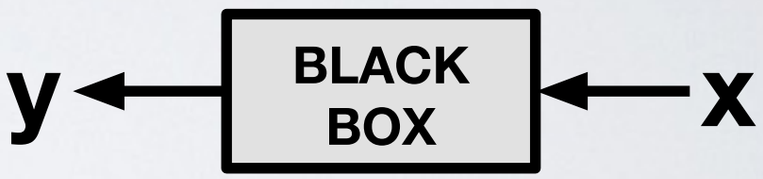




[geosurvey.qed.ai]

RESPONSES



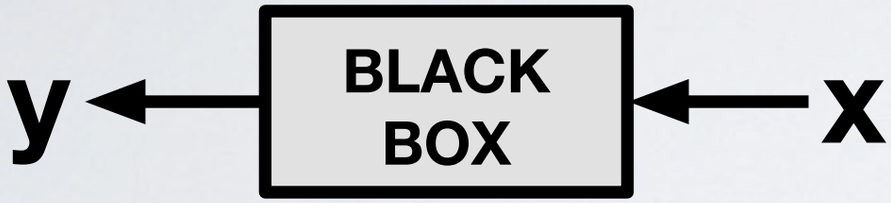


CROPLAND PRESENCE

[geosurvey.qed.ai]

AfSIS

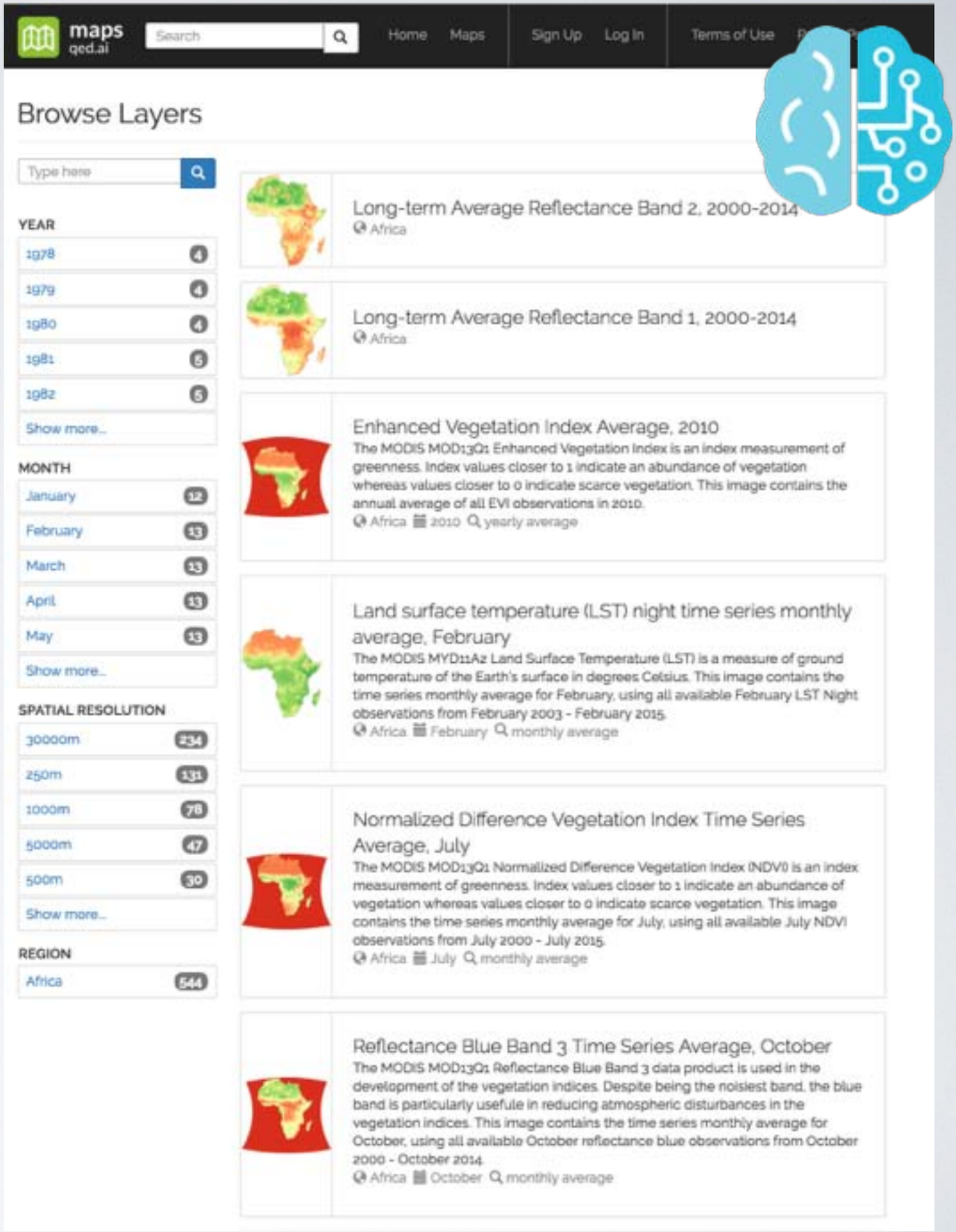
PREDICTORS



CROPLAND PRESENCE

REMOTE SENSING

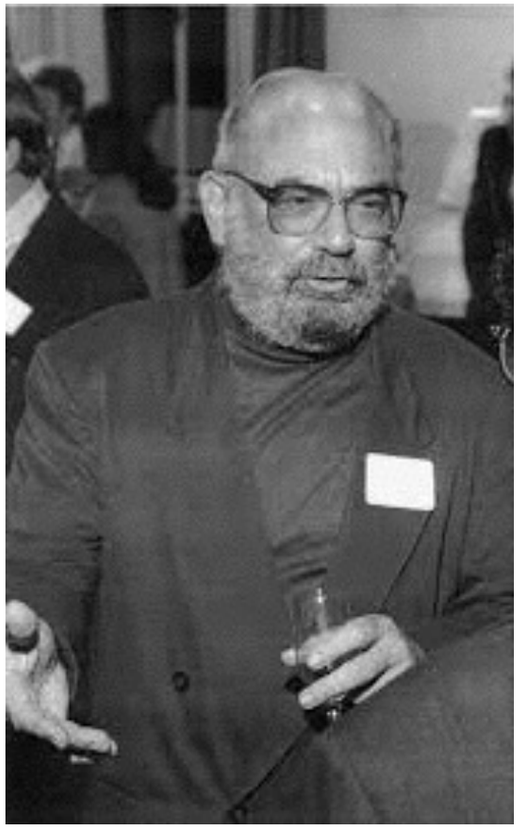
Remote sensing is everywhere.

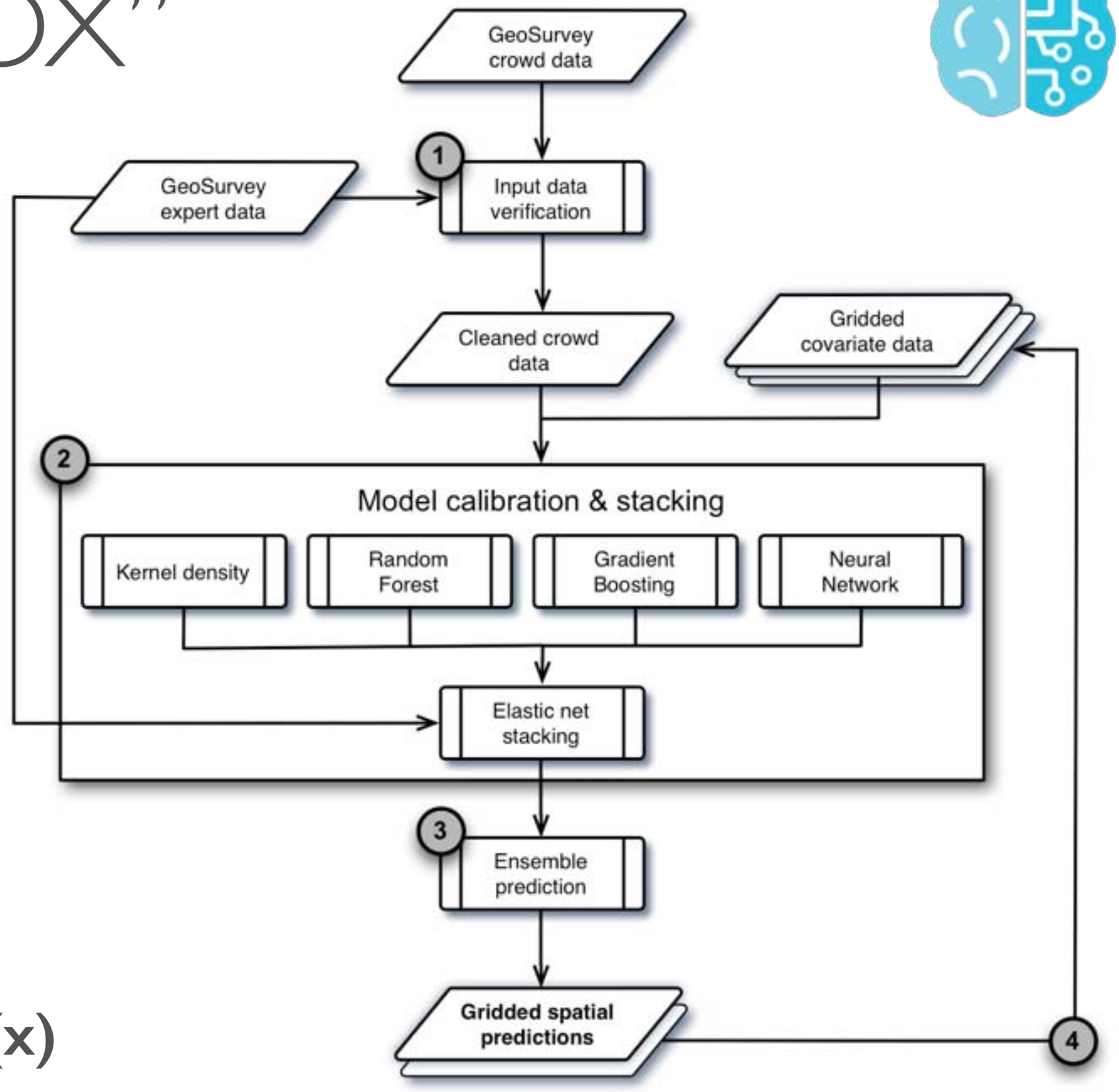


[maps.qed.ai]

AfSIS

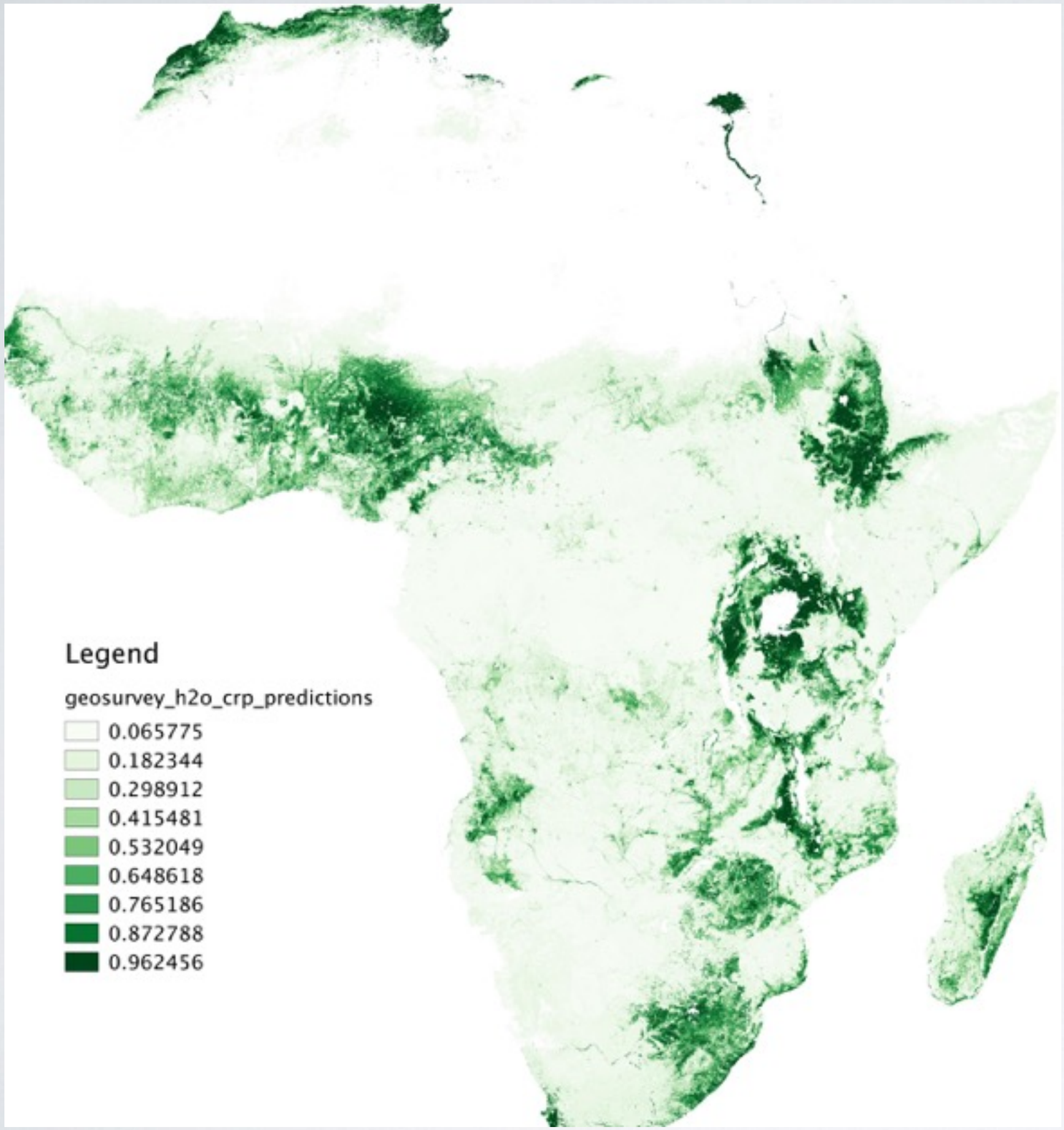
“BLACK BOX”





given (x,y) pairs,
find f such that y ~ f(x)

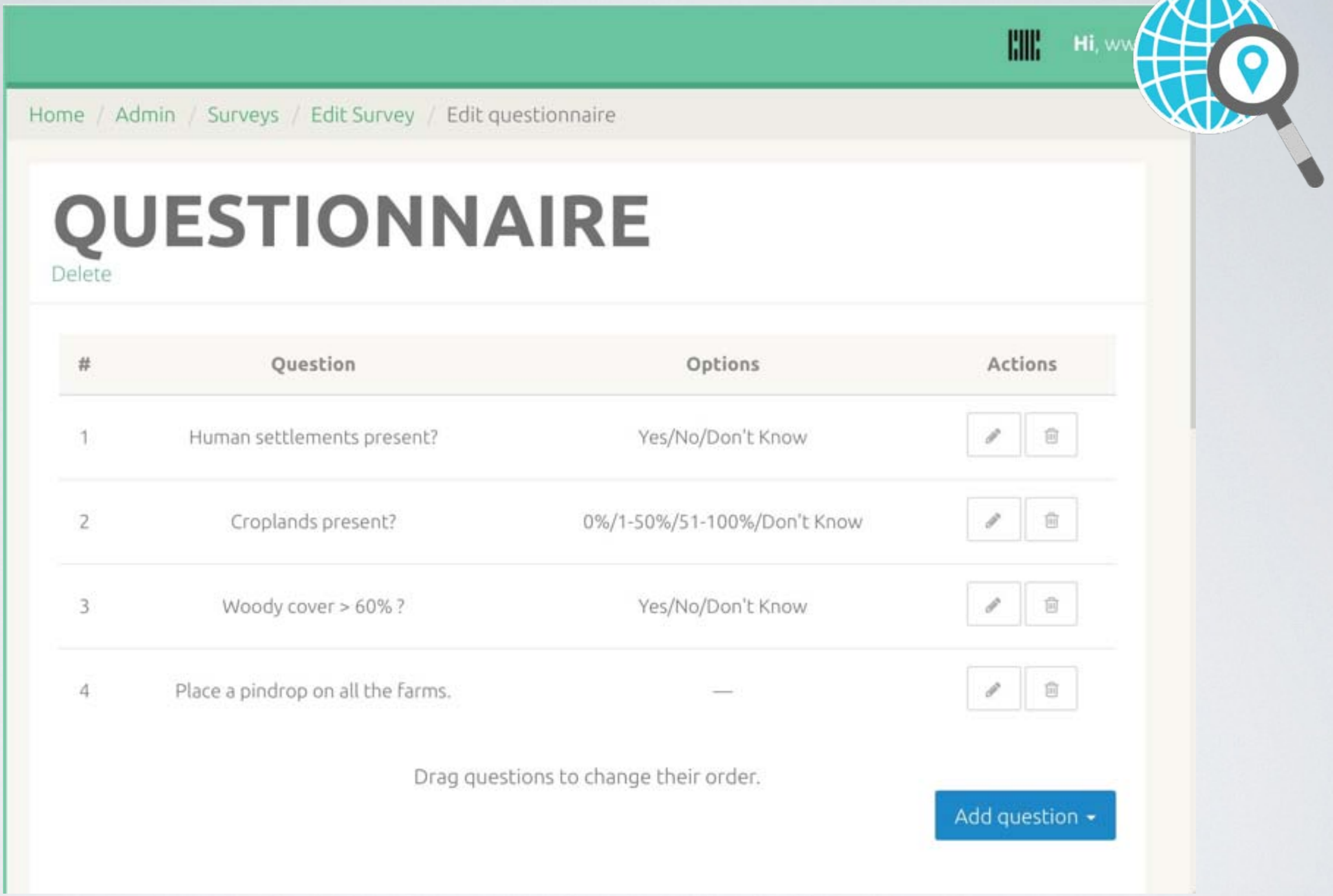
Predicted cropland presence; 1M points



- croplands
- houses
- roads
- phones n’ drones

[geosurvey.qed.ai]

GLOBAL ACCESS



easy to set up arbitrary surveys in arbitrary regions

[geosurvey.qed.ai]

geosurvey.qed.ai

geosurvey
qed.ai

Dashboard Forum Admin Area How It Works Feedback Log Out

Croplands present?
0% | 1-50% | 51-100% | Don't Know

Woody cover > 60% ?
Yes | No | Don't Know

Human settlements present?
Yes | No | Don't Know

Update | Close

Grid
Color: Orange
Show markers
Markers in a row: 4

Submitted by gs-gunrock-09
1 month, 3 weeks ago
(2016-06-30 14:22:16)
Google Maps max zoom: 19
Bing Maps max zoom: 13

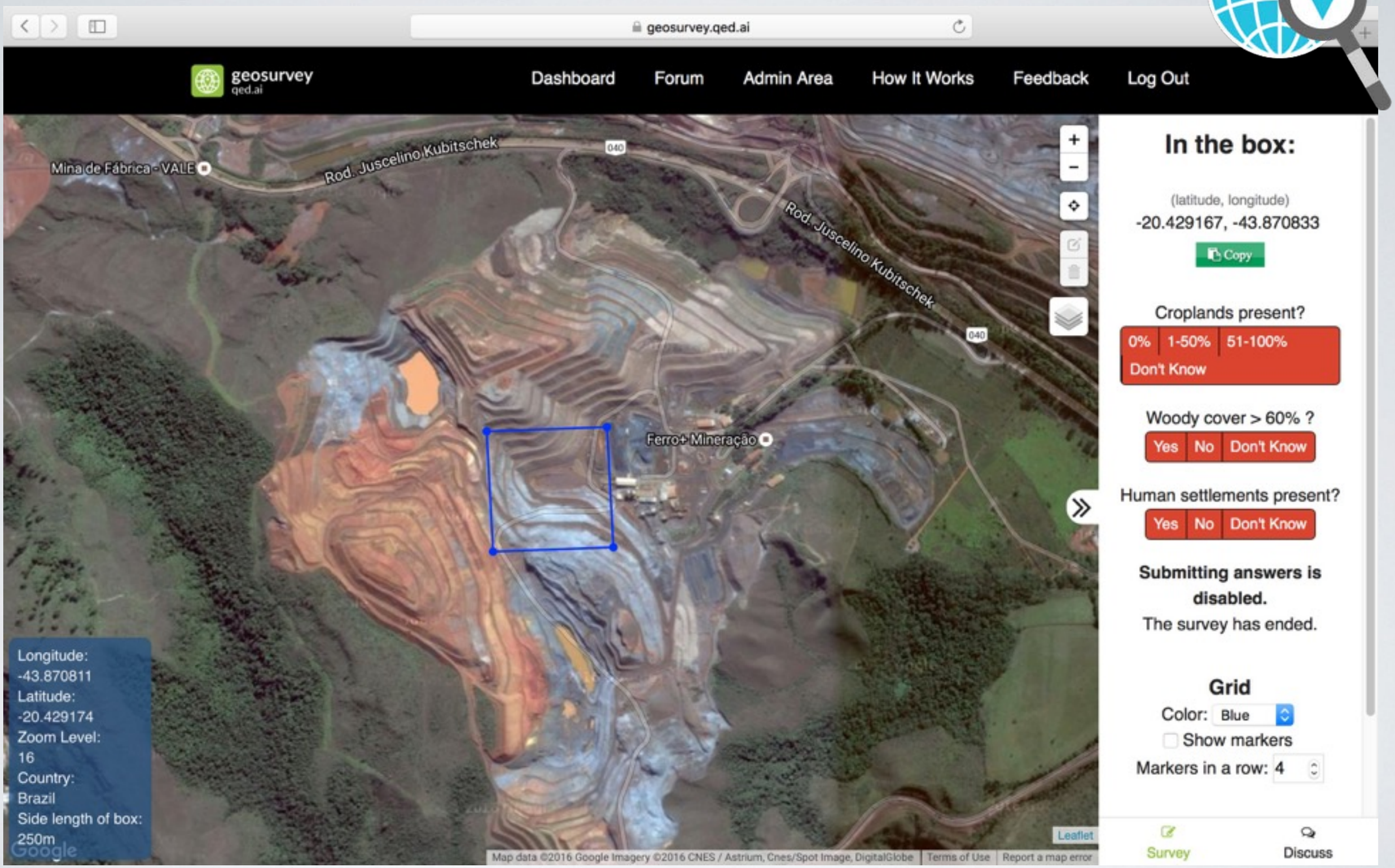
Coordinates
(latitude, longitude)

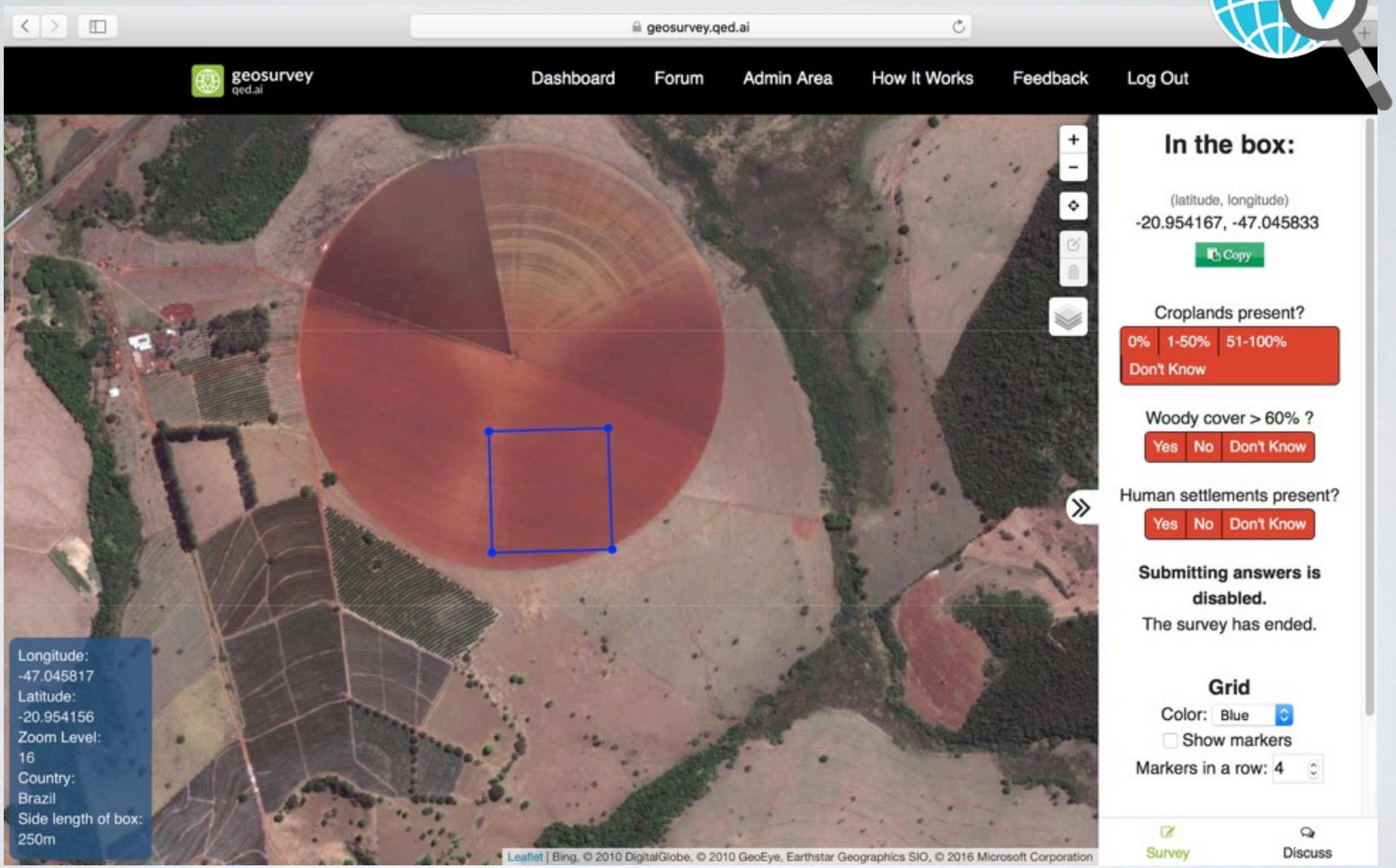
Survey Discuss

Longitude:
64.312506
Latitude:
31.695830
Zoom Level:
16
Country:
Afghanistan
Side length of box:
250m

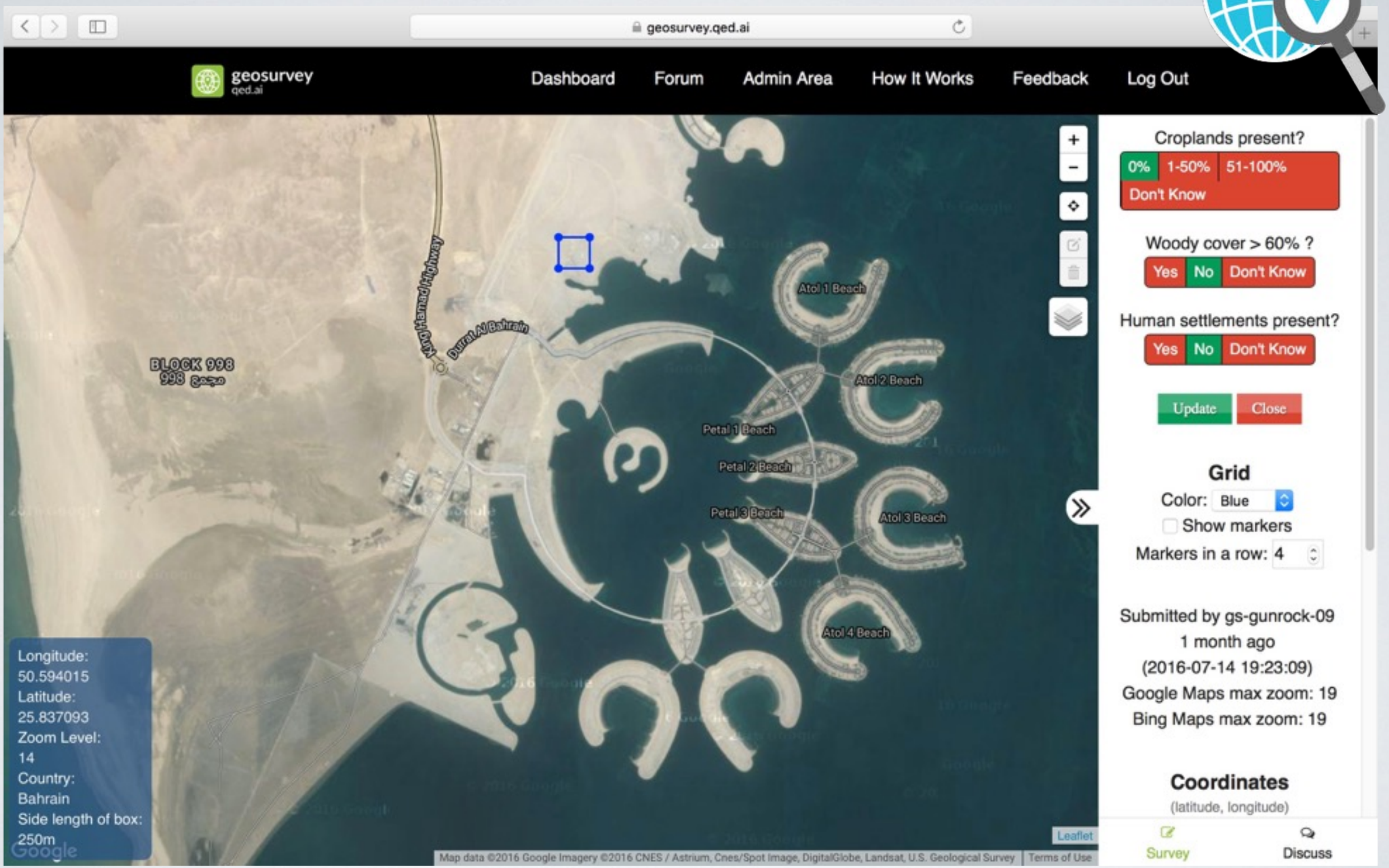
Leaflet | Map data ©2016 Google Imagery ©2016 Cnes/Spot Image, DigitalGlobe | Terms of Use

[geosurvey.qed.ai]

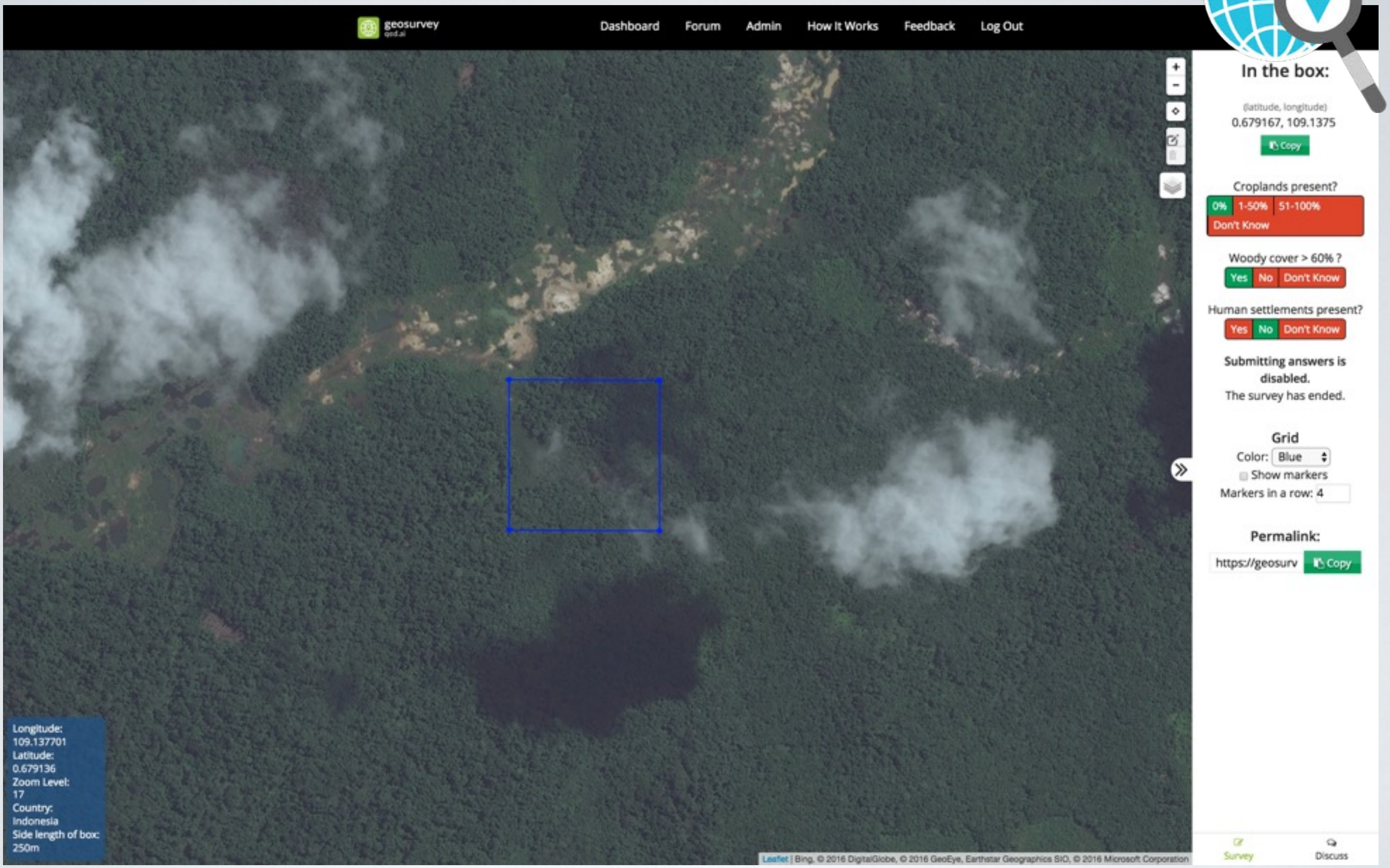




[geosurvey.qed.ai]



Indonesia: 2012-10-15

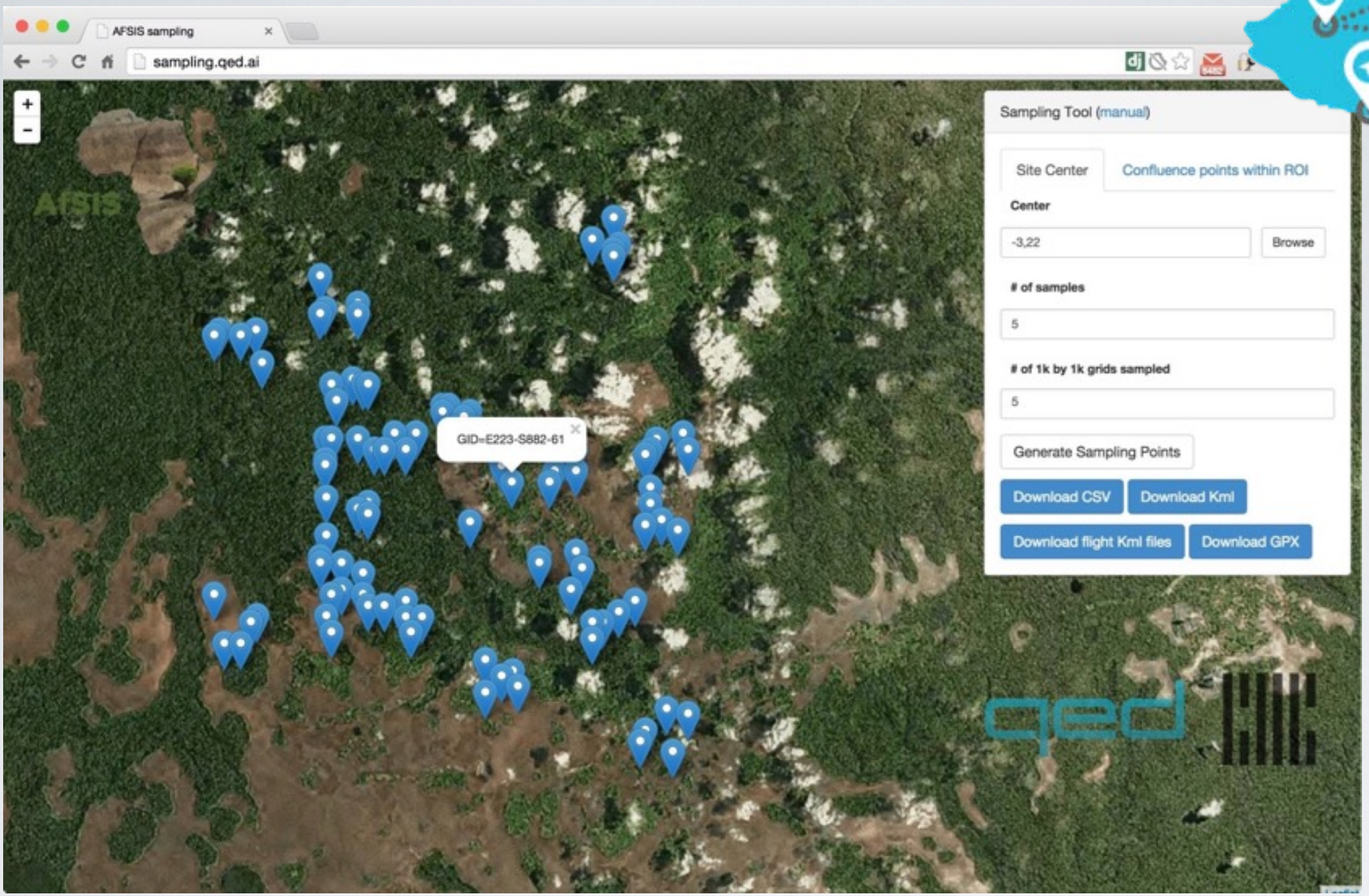


Indonesia: 2015-12-02

[geosurvey.qed.ai]

SUB-PROBLEMS

- **determine ROI**
- **sampling frame**
- **navigate to sampling sites**
- **georeferenced extraction**
- **lab analyses**
- **database**
- **mapping (spatial-temporal modeling)**
- **decisions**

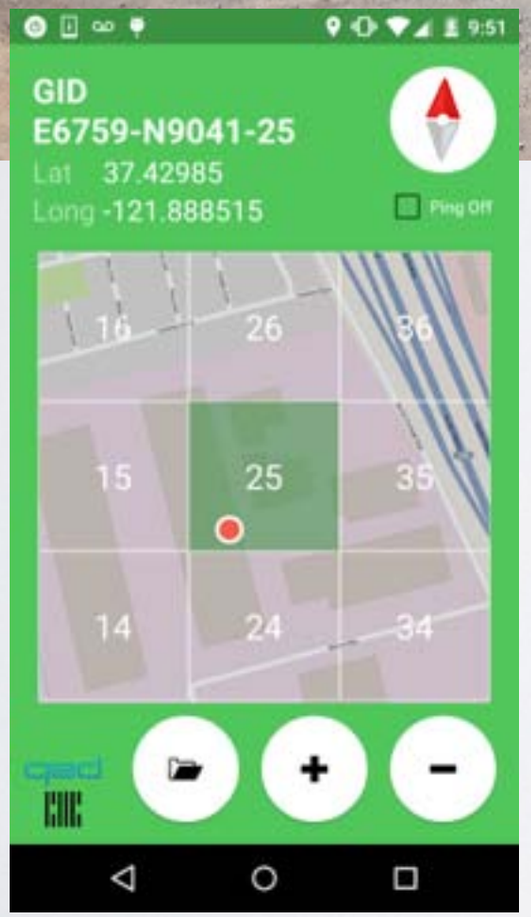


Generate randomized sampling locations that are hierarchically clustered to minimize transportation costs.

[sampling.qed.ai]

SUB-PROBLEMS

- **determine ROI**
- **sampling frame**
- **navigate to sampling sites**
- **georeferenced extraction**
- **lab analyses**
- **database**
- **mapping (spatial-temporal modeling)**
- **decisions**



[fieldwork]

AfSIS

SUB-PROBLEMS

- **determine ROI**
- **sampling frame**
- **navigate to sampling sites**
- **georeferenced extraction**
- **lab analyses**
- **database**
- **mapping (spatial-temporal modeling)**
- **decisions**

tag.kutabiri.com/sheet/

William

tag.kutabiri.com/sheet/

Code sheet generator

Units: mm

Page

Dimensions

Width: 210 mm **Height:** 297 mm

Margins

Top: 0 mm

Left: 0 mm **Right:** 0 mm

Bottom: 0 mm

Codes

Image type: QRCode

Code type: Random String **Length:** 8

Prefix (e.g. country code): max 10 chars

Hidden prefix:

Layout

No. of codes: 1 **No. of copies:** 2

Columns: 3 **Rows:** 4

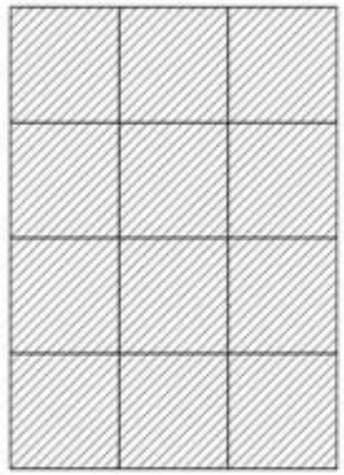
Code padding: 0 **Label:** Show **Number:** Show

Rulers

Horizontal: Show **Vertical:** Show **Font size:** 14

Generate

Preview



Pre-generate customizable QR codes for labeling samples.

[tag.qed.ai]

[fieldwork]

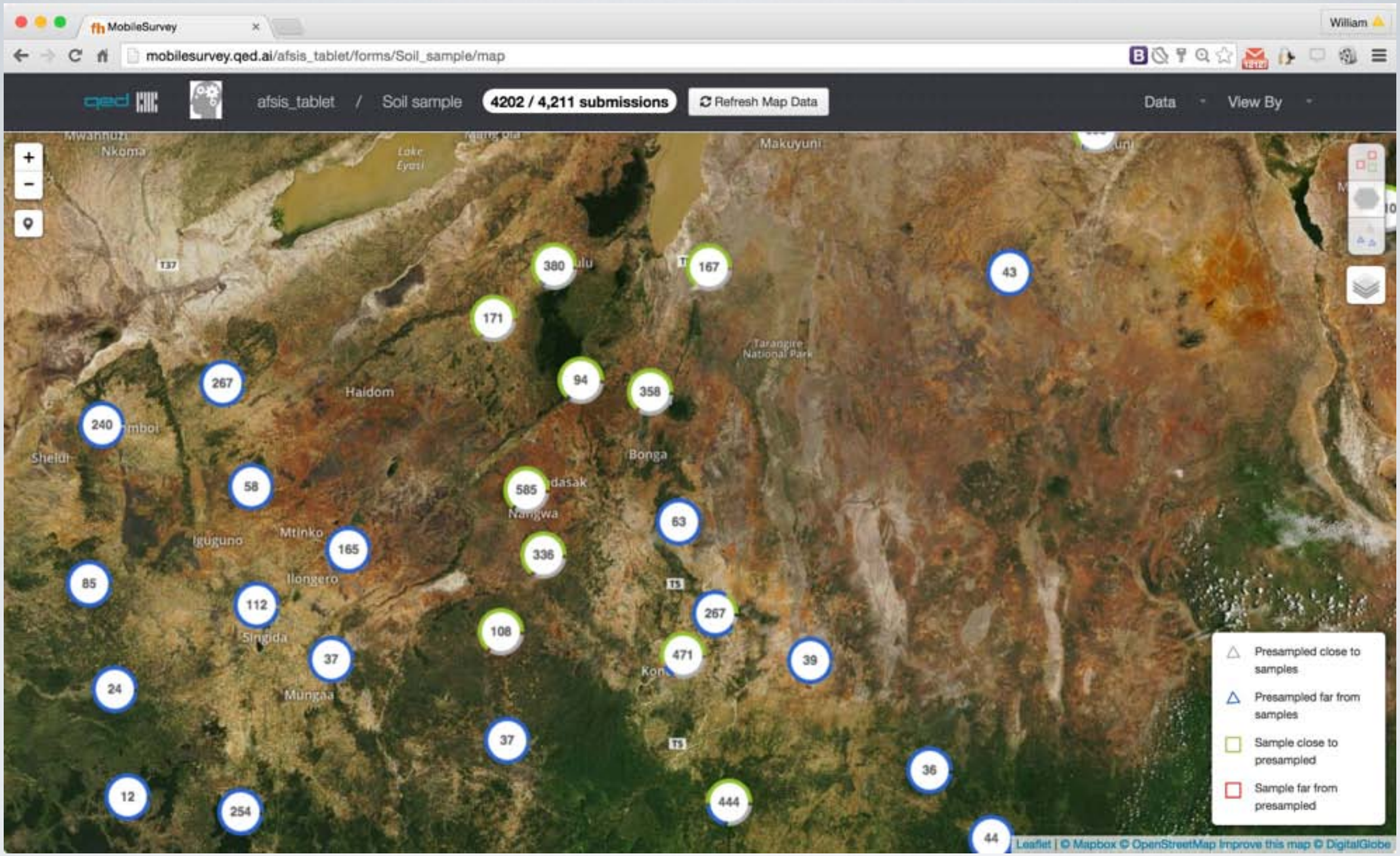
AfSIS

MobileSurvey

- Easily deploy surveys onto Android phones, using Excel spreadsheets.
- Record georeferences, timestamps, pictures, and barcodes.
- Work in zero-wifi areas, caching data until network returns.
- Regular maintenance coupled with scalable geospatial visualizations and fraud detection.

Sign Up!

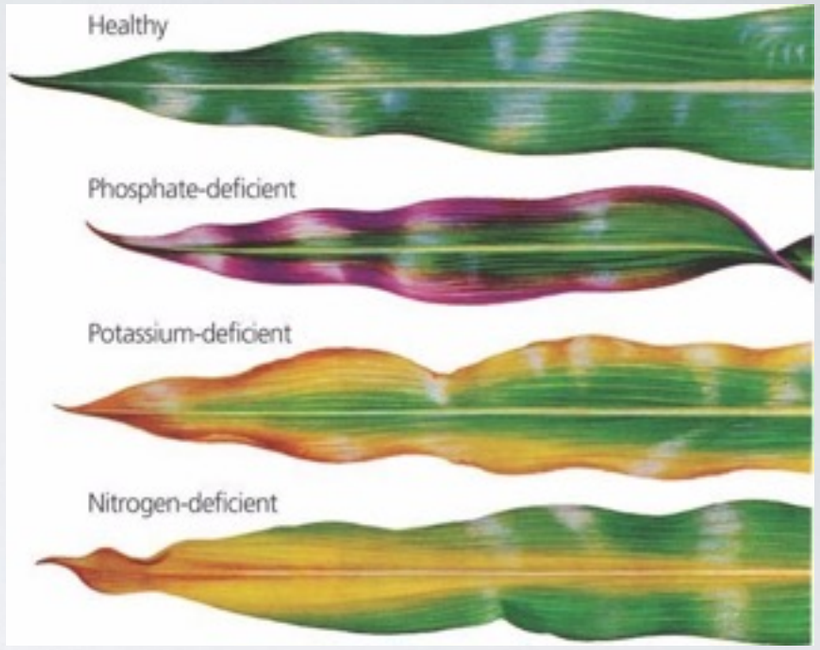
[mobilesurvey.qed.ai]



Build surveys with Excel and deploy to Android. Cache data until internet returns. Fraud detection and geospatial visualization.

[mobilesurvey.qed.ai]

CROP ANALYSIS

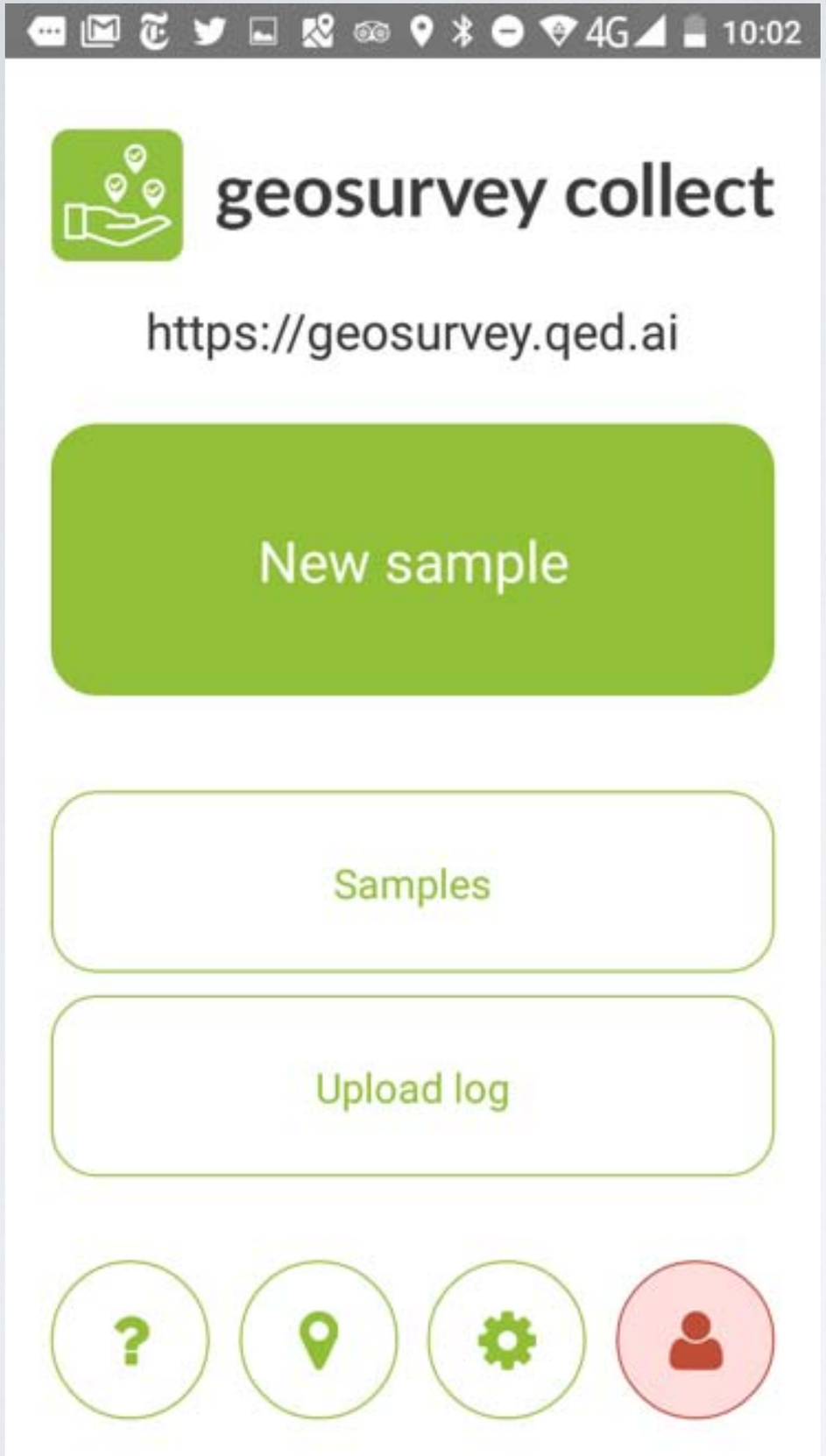


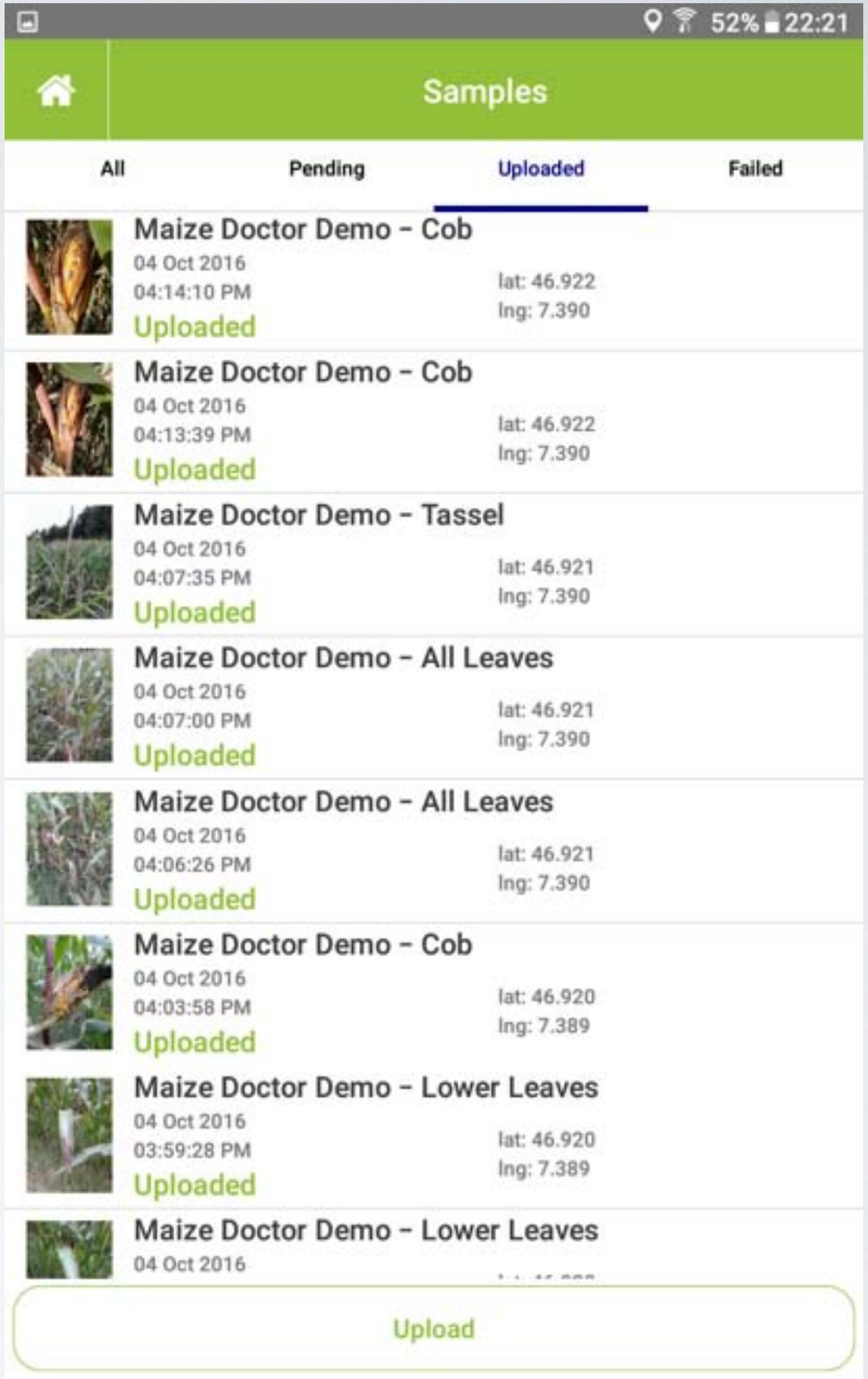
- **Huge yield losses annually from disease and nutrient deficiency. Needs fast intervention.**

GEOSURVEY COLLECT

- **Field operators take pics.**
- **Diagnoses remotely provided on Geosurvey w/ interactive questionnaires.**

[geosurvey:collect]

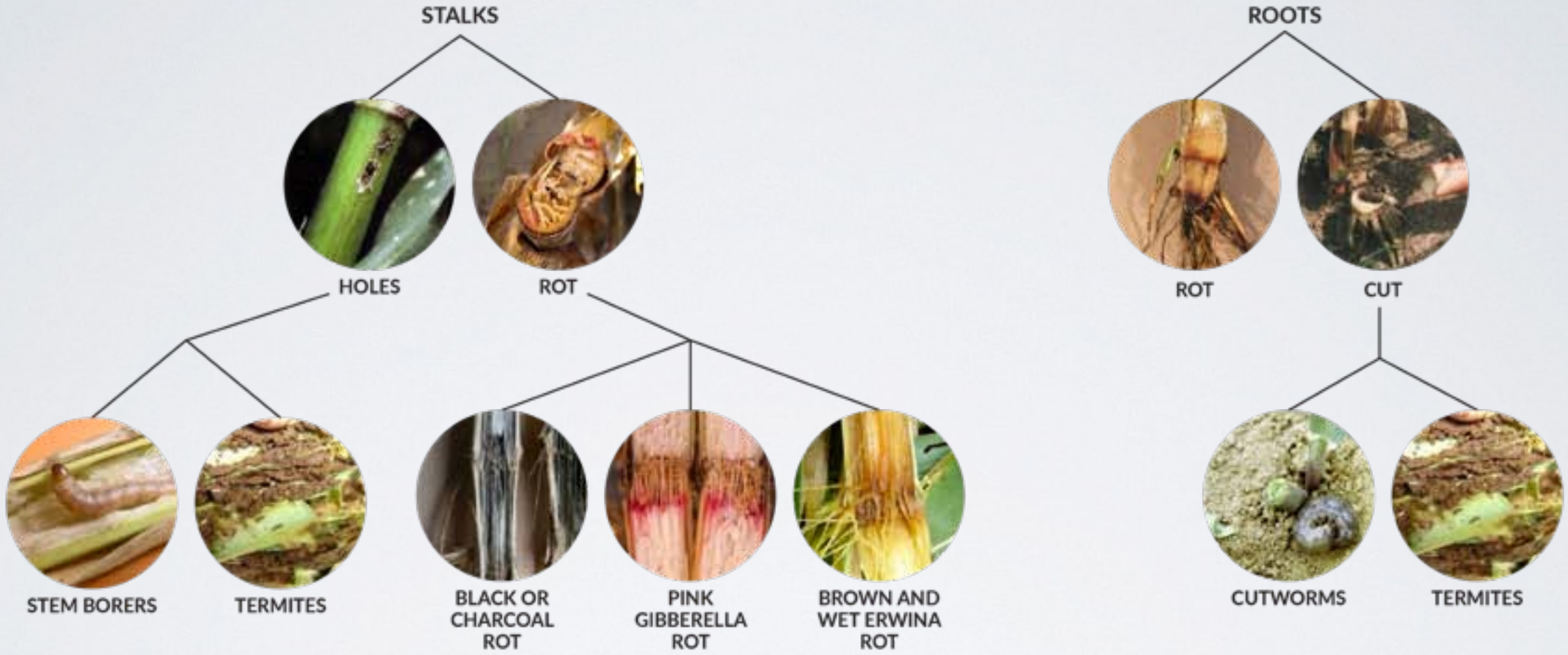




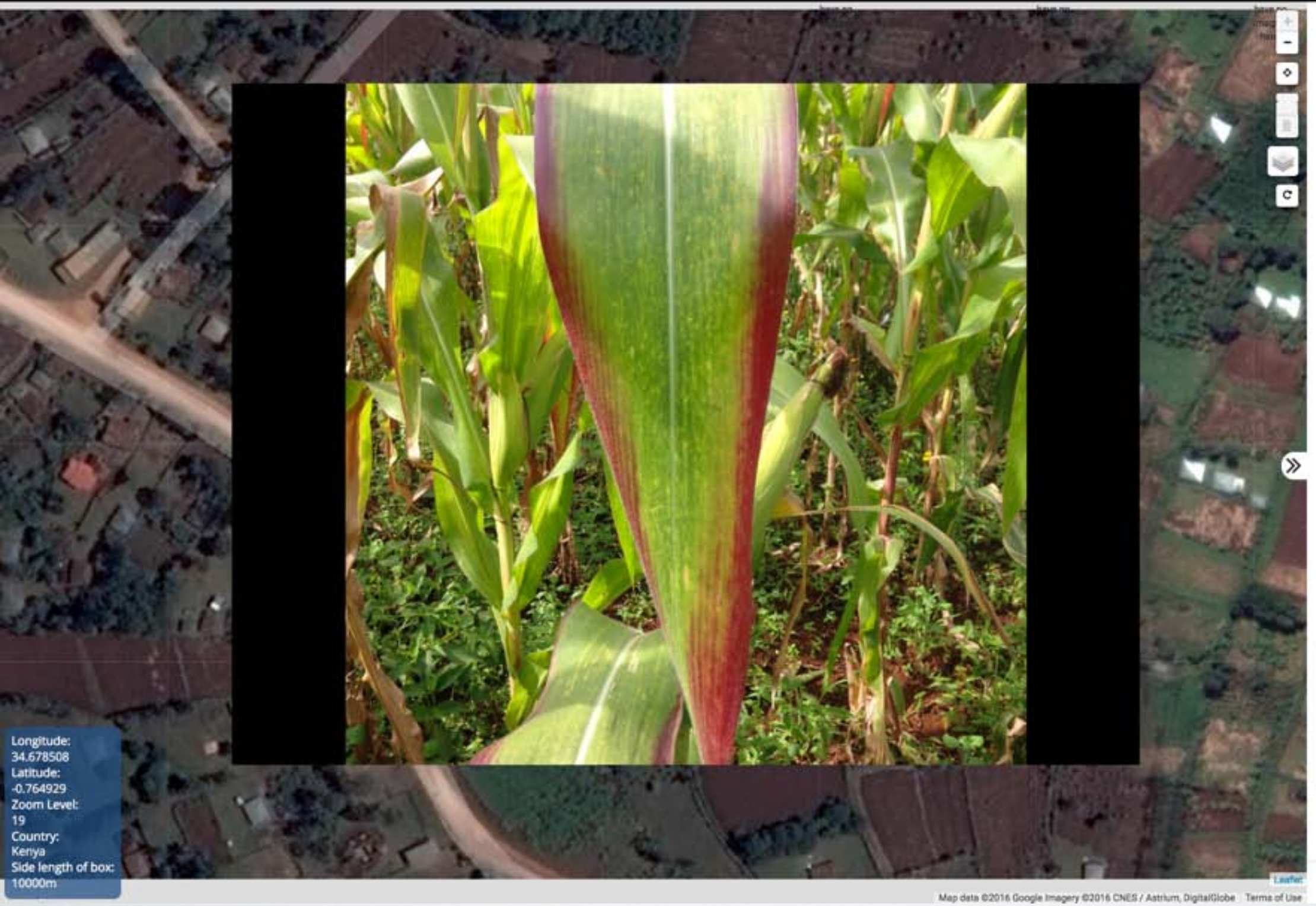
Maize Health Decision Tree v.1.0

Scientific Direction: Dr. David Guerena (1AF)

QED | http://qed.ai



[geosurvey:collect] weblink: maize



Any stripes on the leaves?

Yes No

Update Close

View full-size image

Grid

Color: Orange

Show markers

Markers in a row: 2

Submitted by

david.guerena

2 months, 1 week ago

Max zoom

Google: 19

Bing: 19

Coordinates

(latitude, longitude)

-0.764838, 34.678611

Copy

Send a message to the submitter

Save Delete

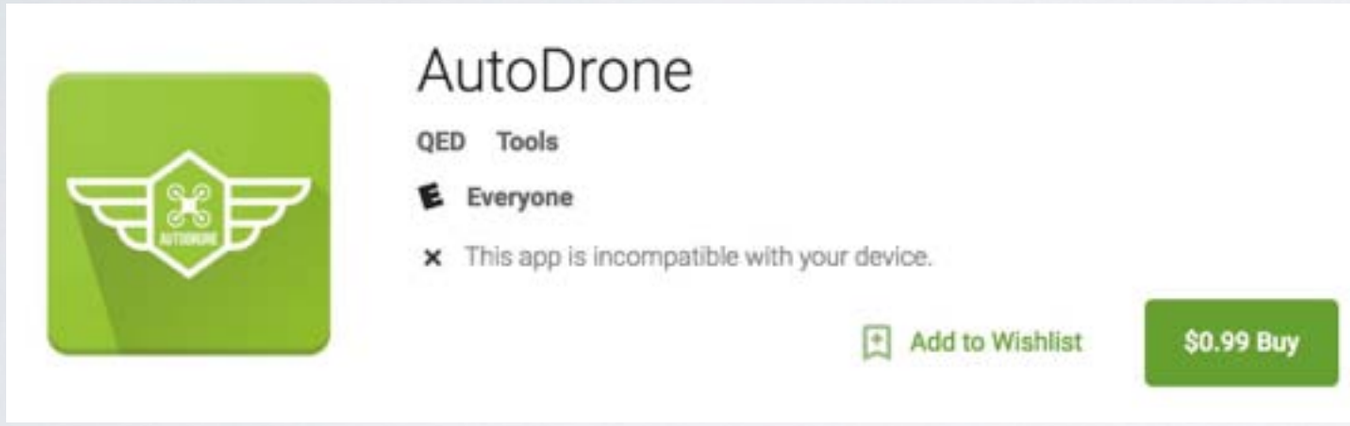
Survey Discuss

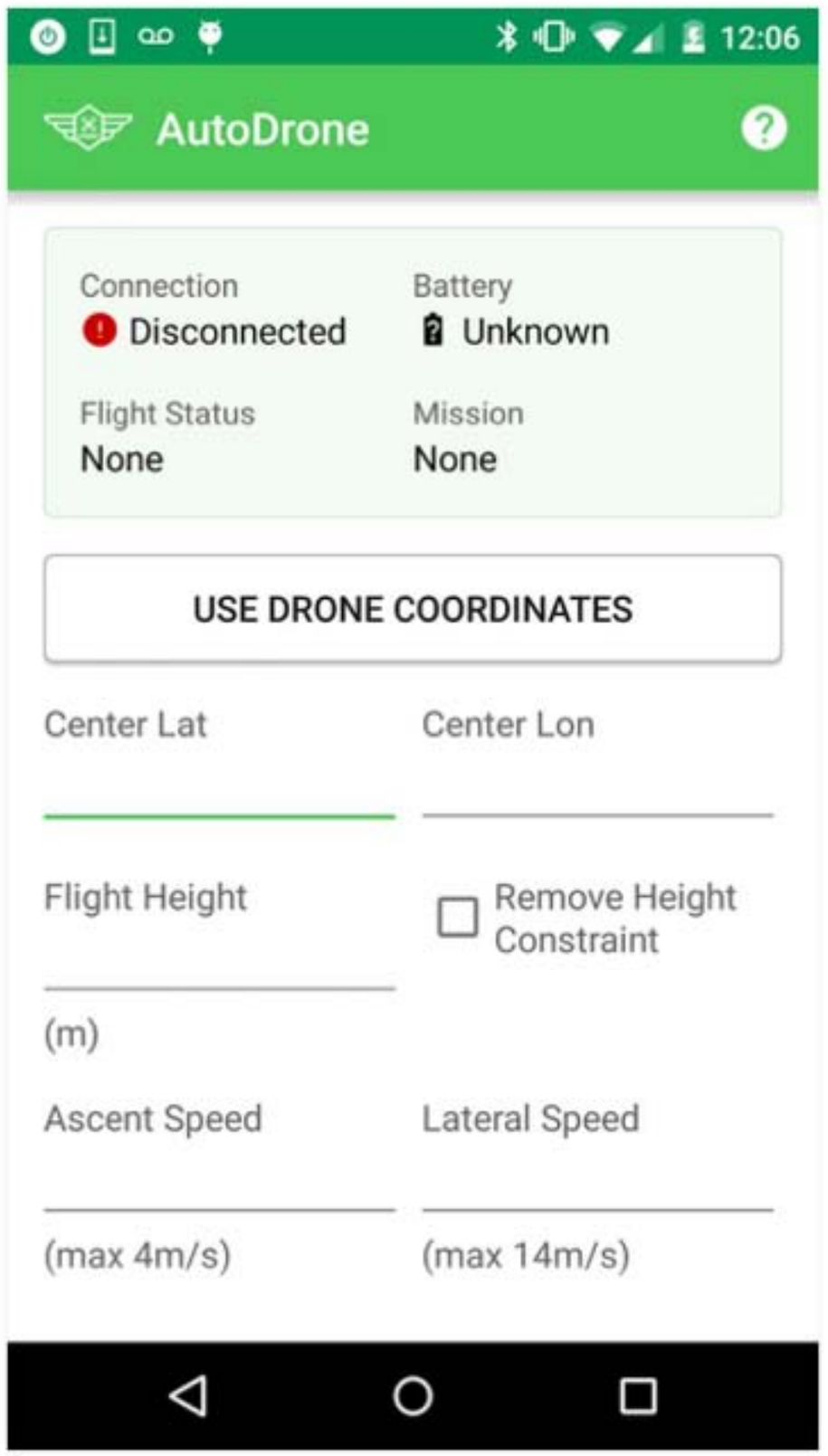
https://youtu.be/SP0HTqImv-Q

weblink: uav-orthomosaic

[uav.qed.ai]

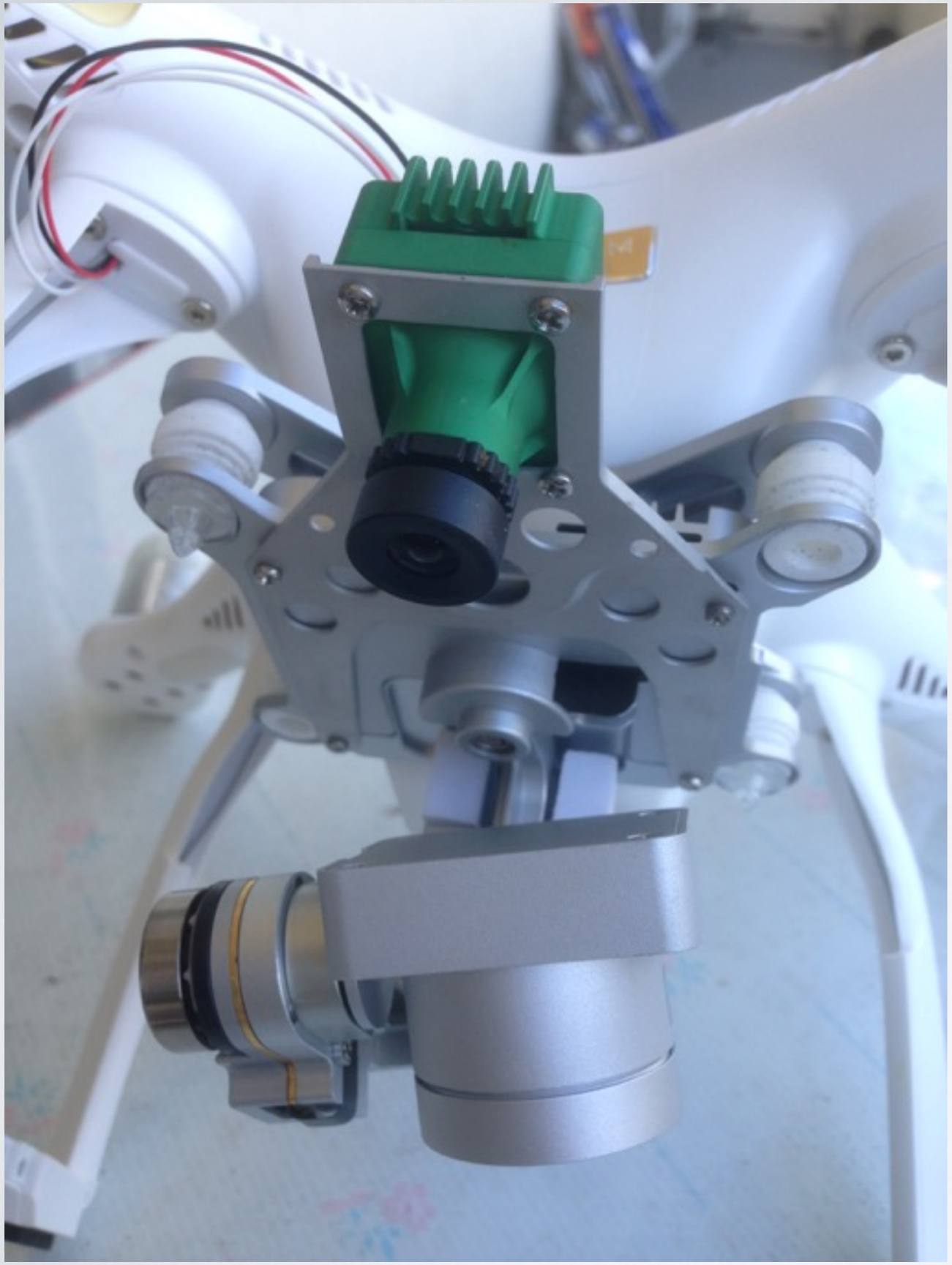
AfSIS

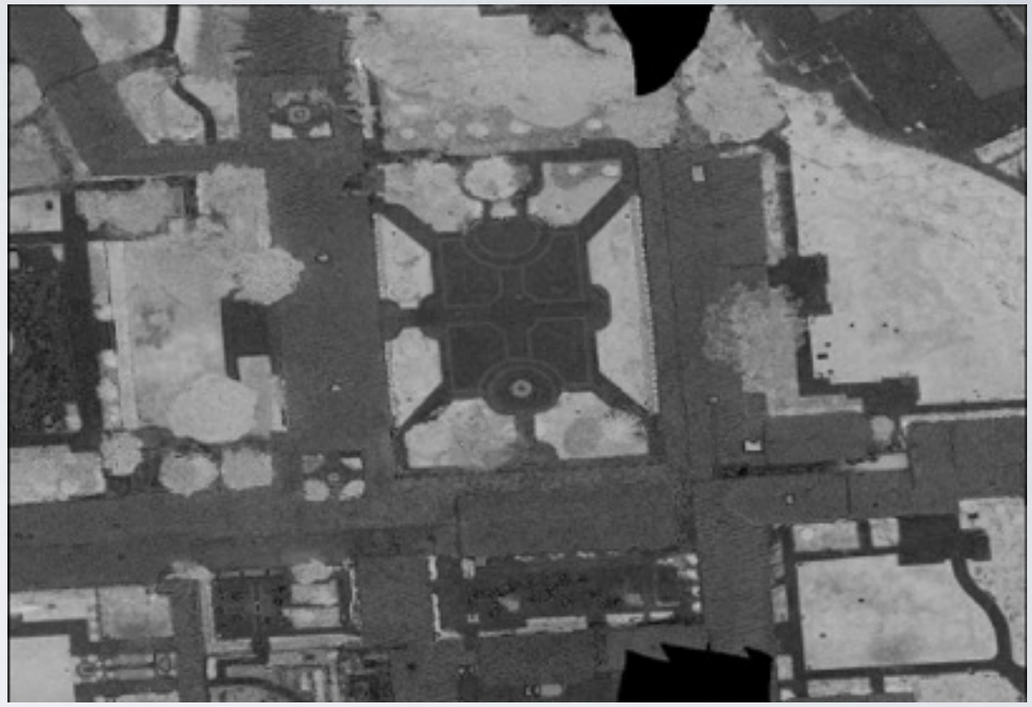




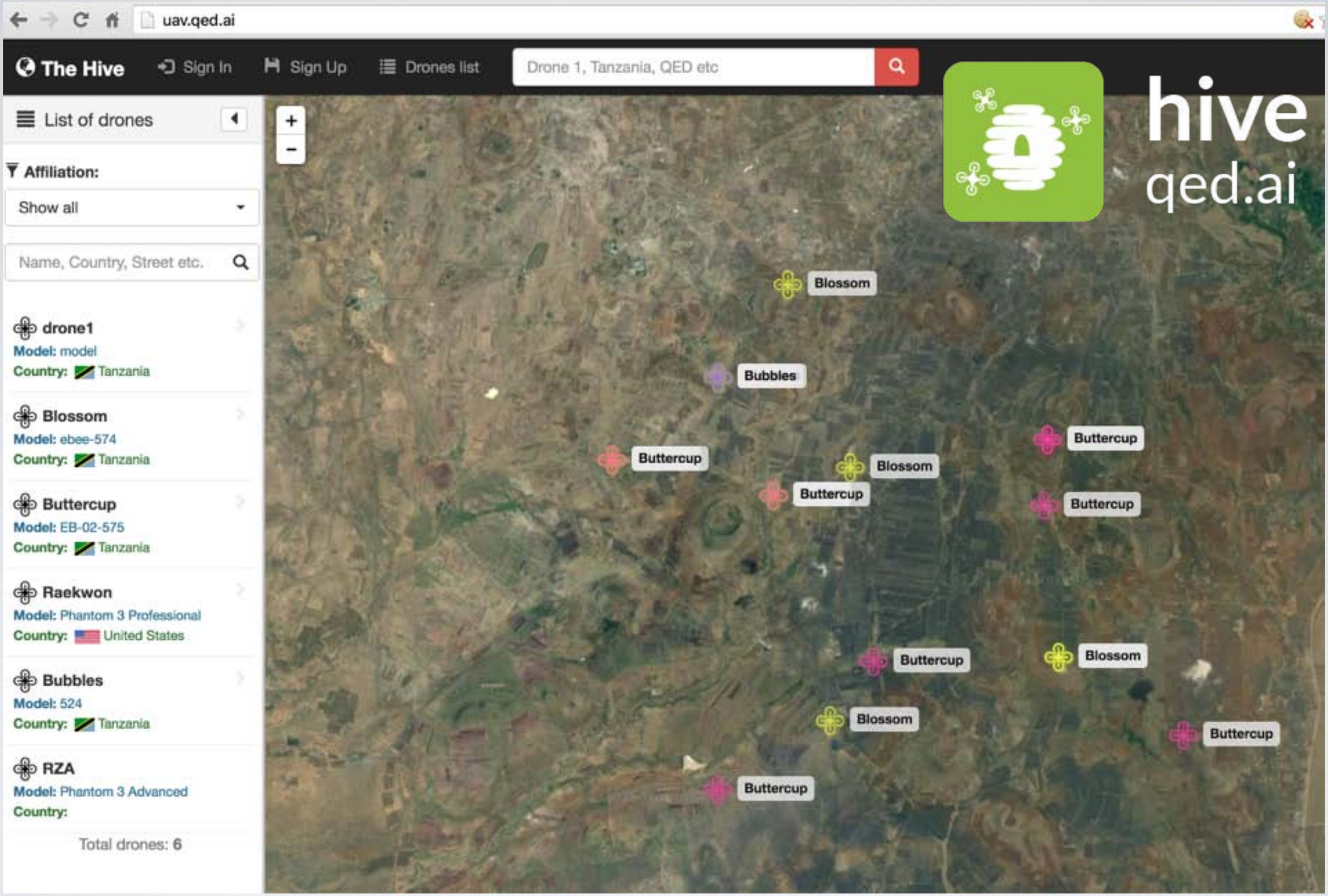
Automated orthomosaic flight control. Works offline. 99 cents.

[uav:auto-drone]

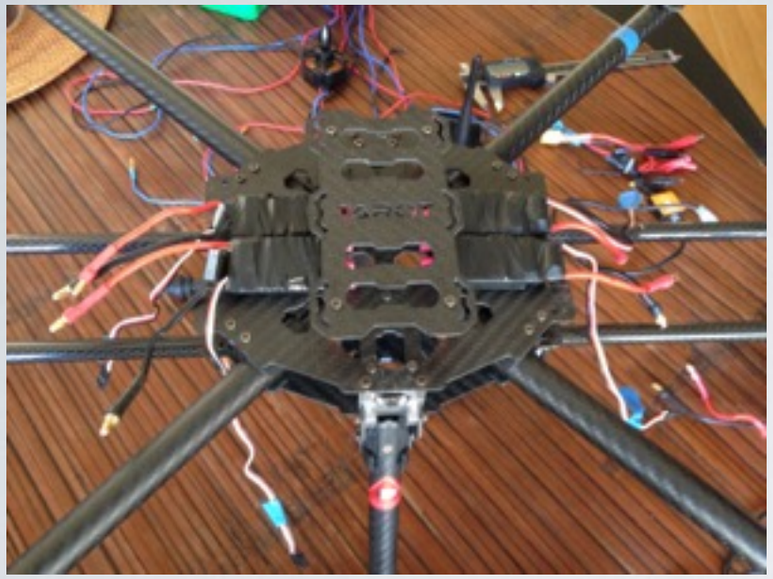




[uav:auto-drone]



Application: Store and visualize orthomosaics of UAVs in the cloud.



https://youtu.be/3wj6pvXsgOA

[uav:methodman]

SUB-PROBLEMS

- **determine ROI**
- **sampling frame**
- **navigate to sampling sites**
- **georeferenced extraction**
- **lab analyses**
- **database**
- **mapping (spatial-temporal modeling)**
- **decisions**

[soil:preparation]

AfSIS

SPECTROSCOPY

y = wet chemistry ($$$) x = spectroscopy ($)

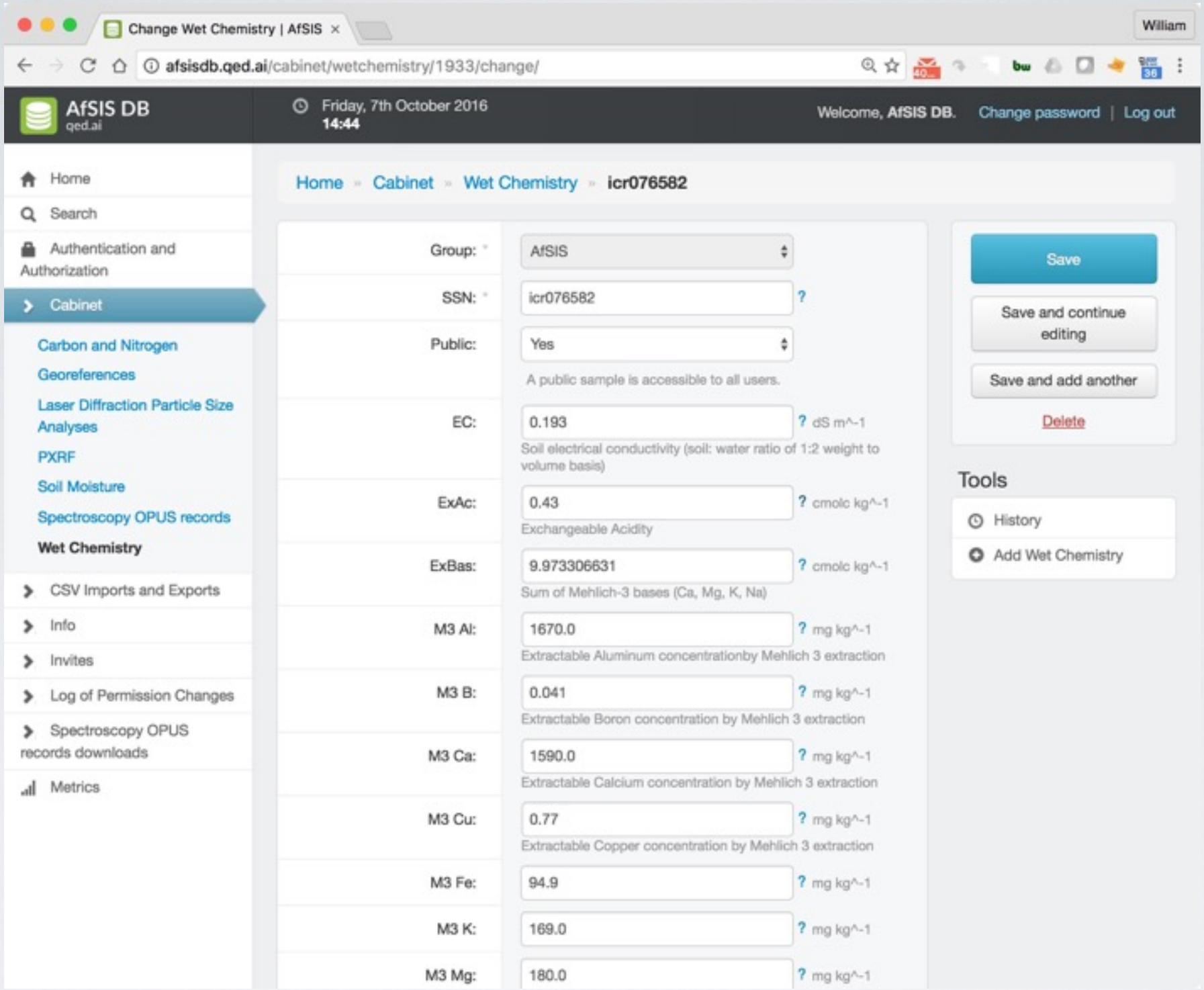
given lots of (x,y) pairs, find f such that y = f(x)

spectpred
qed.ai

https://www.kaggle.com/c/afsis-soil-properties

SUB-PROBLEMS

- **determine ROI**
- **sampling frame**
- **navigate to sampling sites**
- **georeferenced extraction**
- **lab analyses**
- **database**
- **mapping (spatial-temporal modeling)**
- **decisions**



AFSIS DATABASE

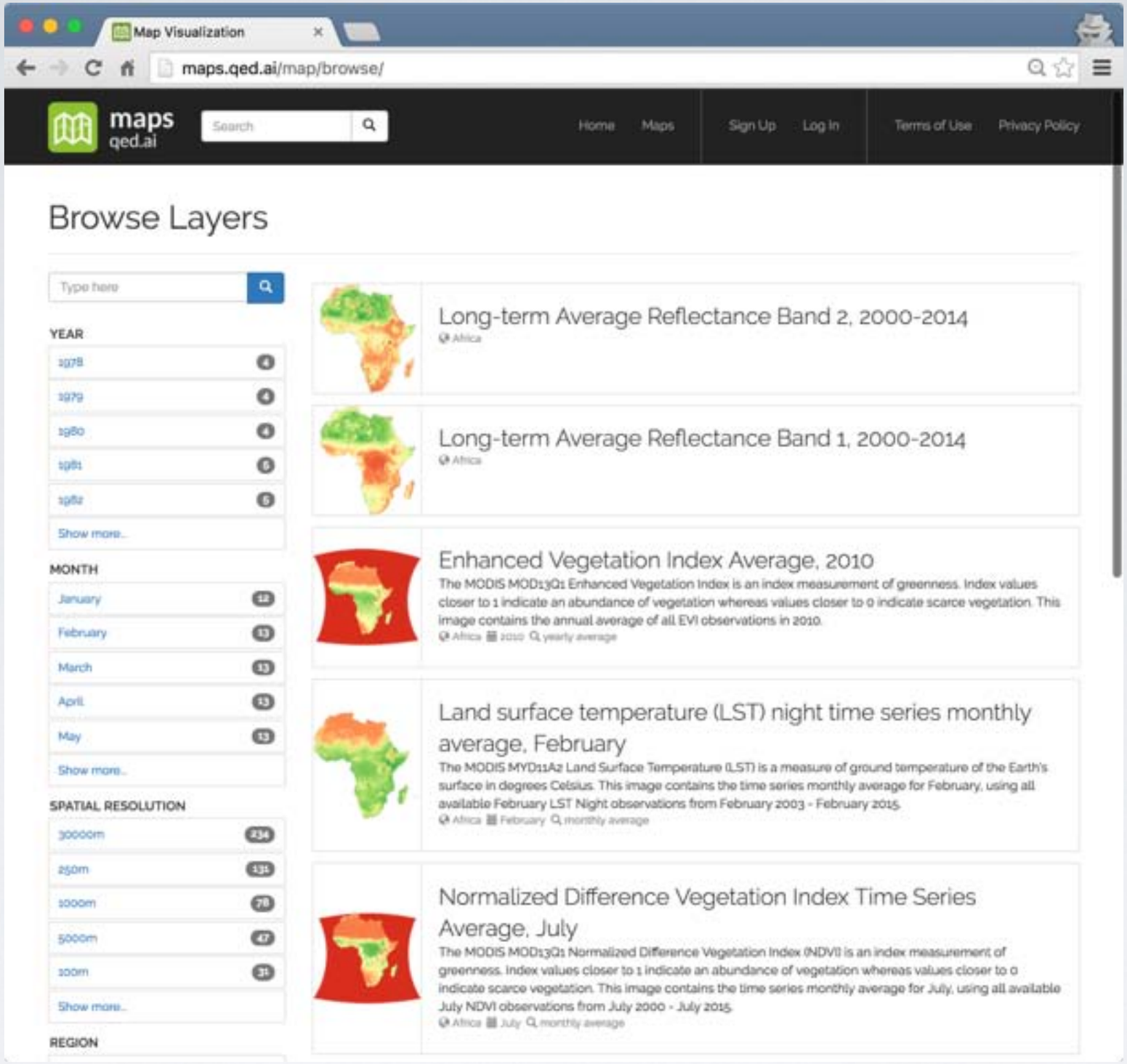
Application: Cloud DB of soil information. RDBMS performance, off-site backup, batch upload, input validation, geospatial search, API support.

[afsisdb.qed.ai]

AfSIS

SUB-PROBLEMS

- **determine ROI**
- **sampling frame**
- **navigate to sampling sites**
- **georeferenced extraction**
- **lab analyses**
- **database**
- **mapping (spatial-temporal modeling)**
- **decisions**

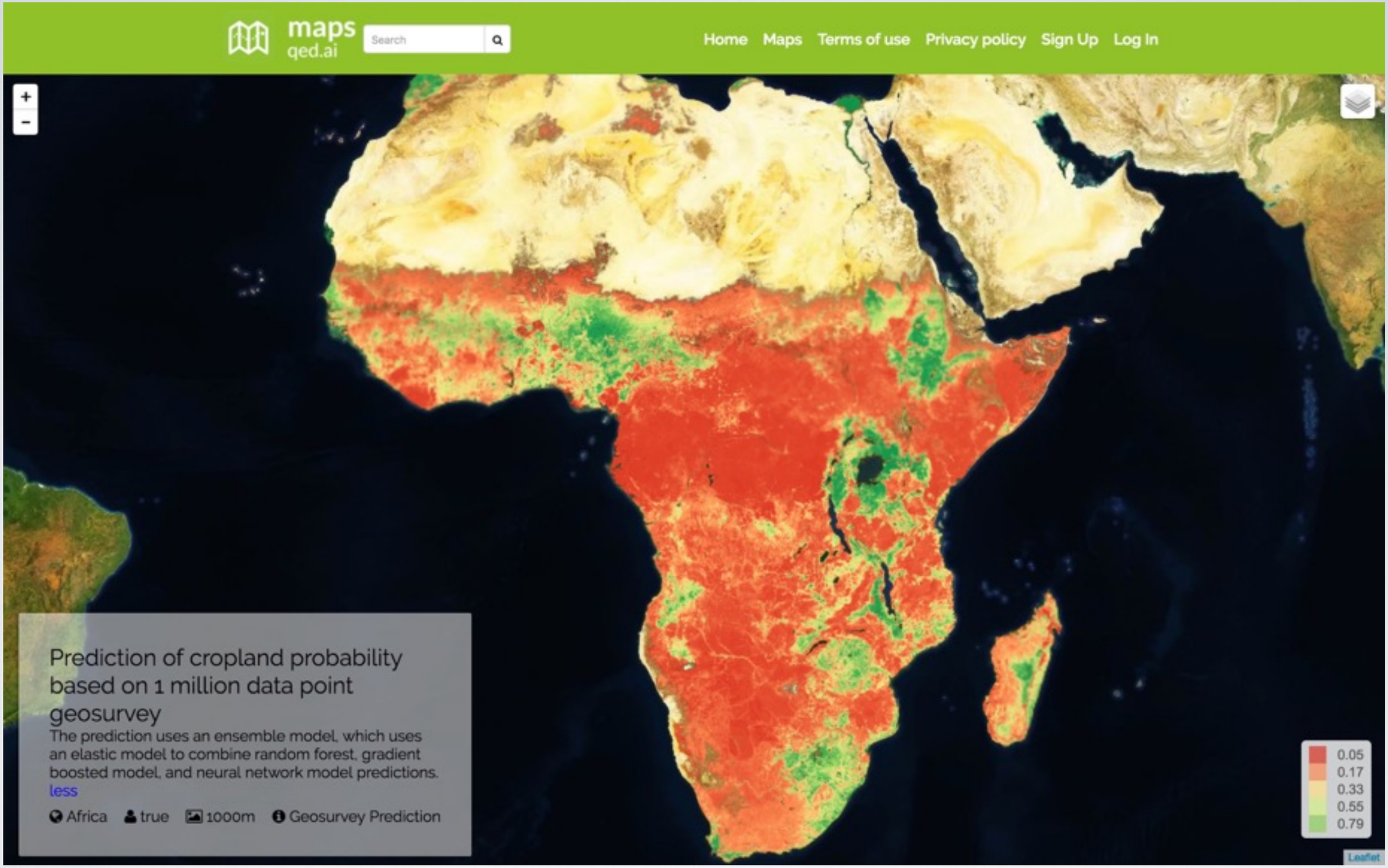


MAPS

Scalable viz. and mgmt. of remote sensing layers and geospatial predictions.

[maps.qed.ai]

AfSIS



Scalable viz. and mgmt. of remote sensing layers and geospatial predictions.

[maps.qed.ai]

AfSIS

SUB-PROBLEMS

- **determine ROI**
- **sampling frame**
- **navigate to sampling sites**
- **georeferenced extraction**
- **lab analyses**
- **database**
- **mapping (spatial-temporal modeling)**
- **decisions …**

CONCLUSION

BILL & MELINDA GATES foundation

AfSIS

ONE ACRE FUND

Soil Erosion

Nutrient Imbalance

Plant Disease

Financial Services

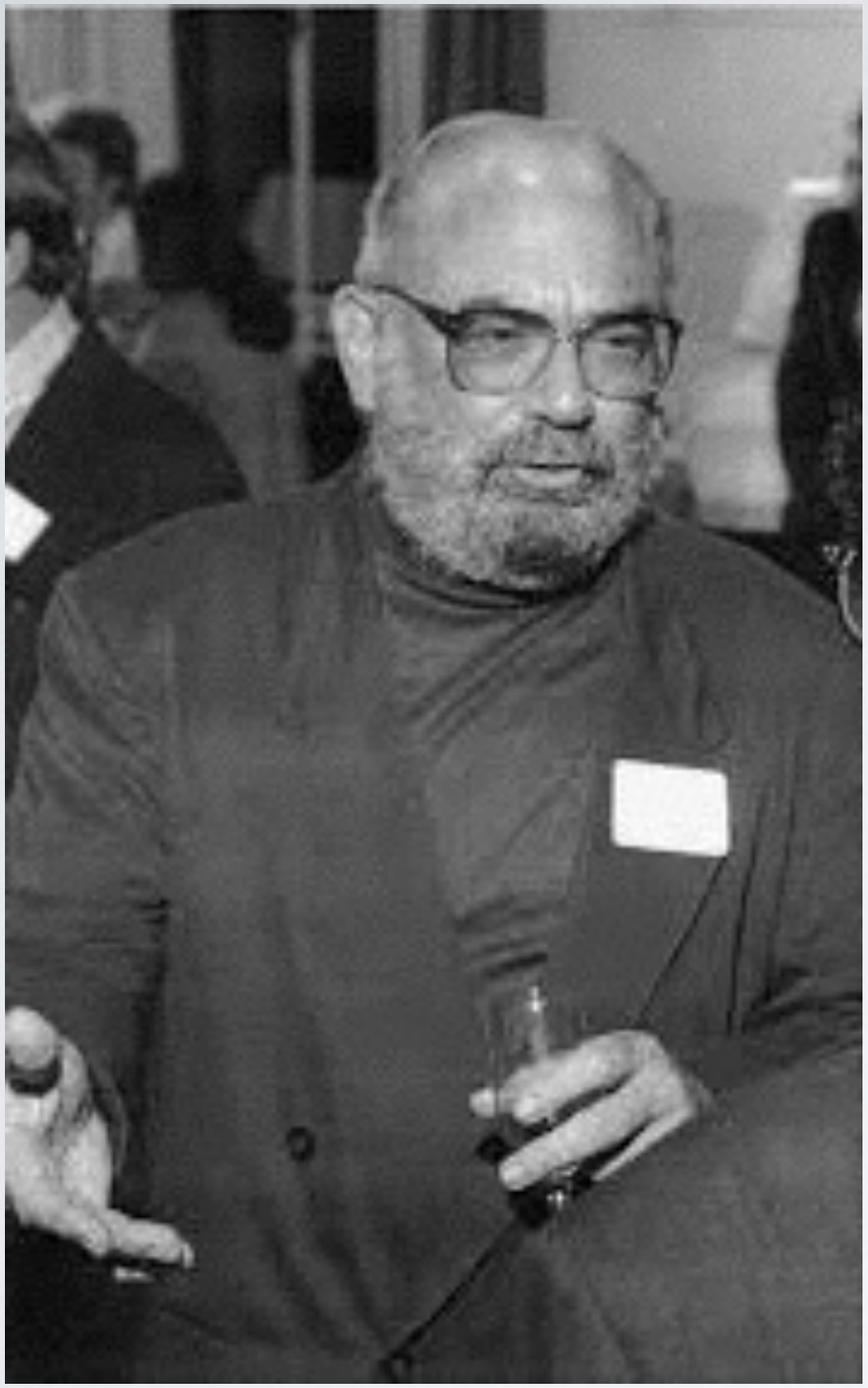
Market Access

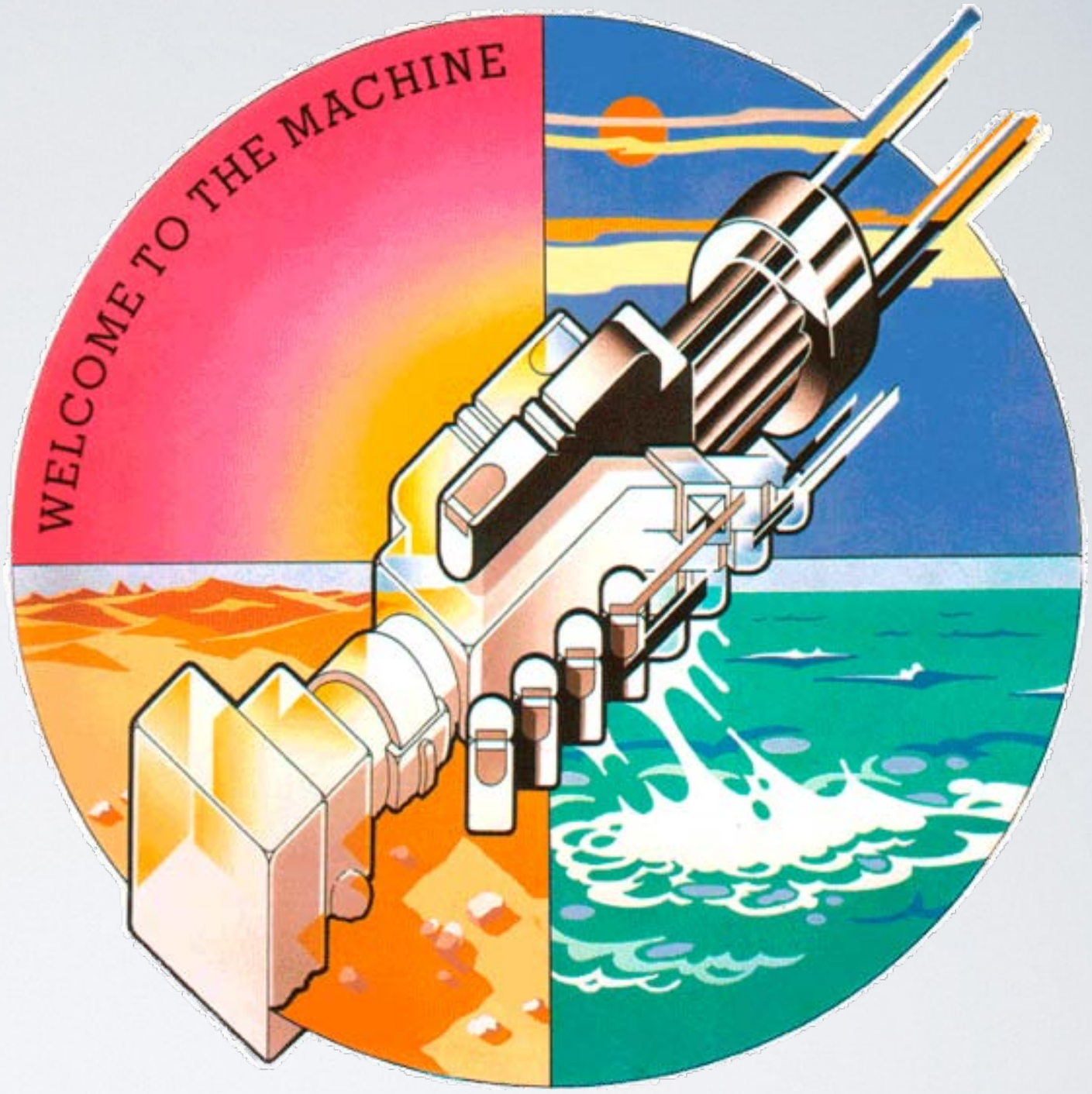
Yield Gaps

Child Mortality

Malnutrition

Climate Change

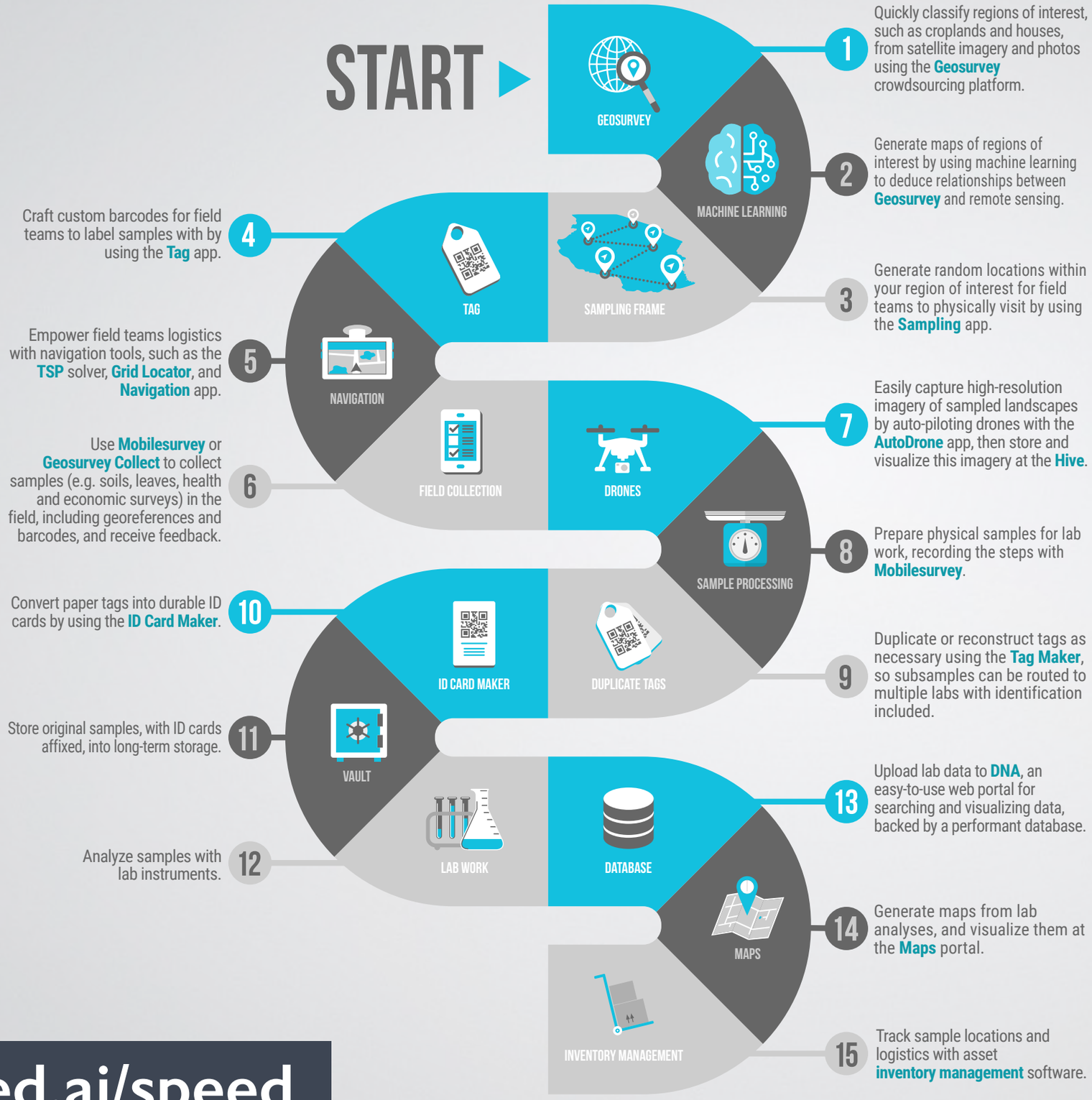




SPEED

SCALABLE PIPELINE FOR ENGINEERING OF ENORMOUS DATA

End-to-end technology solutions to support systematic geospatial data collection and analysis.



1 Quickly classify regions of interest, such as croplands and houses, from satellite imagery and photos using the **Geosurvey** crowdsourcing platform.

2 Generate maps of regions of interest by using machine learning to deduce relationships between **Geosurvey** and remote sensing.

3 Generate random locations within your region of interest for field teams to physically visit by using the **Sampling** app.

4 Craft custom barcodes for field teams to label samples with by using the **Tag** app.

5 Empower field teams logistics with navigation tools, such as the **TSP** solver, **Grid Locator**, and **Navigation** app.

6 Use **Mobilesurvey** or **Geosurvey Collect** to collect samples (e.g. soils, leaves, health and economic surveys) in the field, including georeferences and barcodes, and receive feedback.

7 Easily capture high-resolution imagery of sampled landscapes by auto-piloting drones with the **AutoDrone** app, then store and visualize this imagery at the **Hive**.

8 Prepare physical samples for lab work, recording the steps with **Mobilesurvey**.

9 Duplicate or reconstruct tags as necessary using the **Tag Maker**, so subsamples can be routed to multiple labs with identification included.

10 Convert paper tags into durable ID cards by using the **ID Card Maker**.

11 Store original samples, with ID cards affixed, into long-term storage.

12 Analyze samples with lab instruments.

13 Upload lab data to **DNA**, an easy-to-use web portal for searching and visualizing data, backed by a performant database.

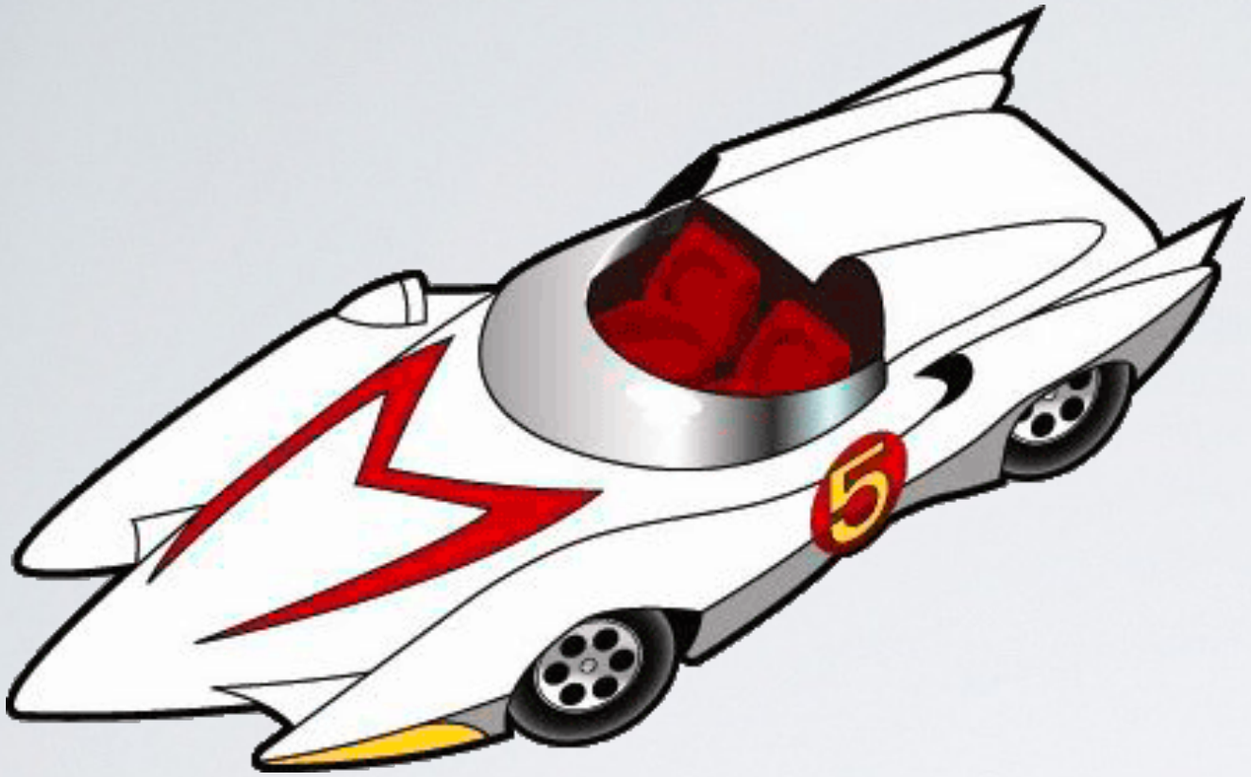
14 Generate maps from lab analyses, and visualize them at the **Maps** portal.

15 Track sample locations and logistics with asset **inventory management** software.

qed.ai/speed

QED APP LEGEND

- 1 geosurvey
- 3 sampling
- 4 tag
- 5 navigation
- 5 grid locator
- 5 tsp
- 7 uav:quadcopter
- 7 uav:fixed-wing
- 6 8 mobilesurvey
- 6 8 geosurvey collect
- 9 tag maker
- 10 id card maker
- 13 dna
- 14 maps
- 15 inventory



+

Technology **Driver = ?**

"War is ninety percent information."
- Napoleon Bonaparte

"In God we trust; all others must bring data!"
- William Edwards Deming

THANKS FOR LISTENING!

To get in touch:

info@qed.ai

http://qed.ai

#TweetQED