Data Visualization:
The Missing Link Between Science and Humanity

QED (https://qed.ai)
Michał Łazowik (@mlazowik)

Me - fullstack dev at QED (https://qed.ai) for a few years.

QED - solving medical and agricultural problems in Sub-Saharan Africa and South Asia.

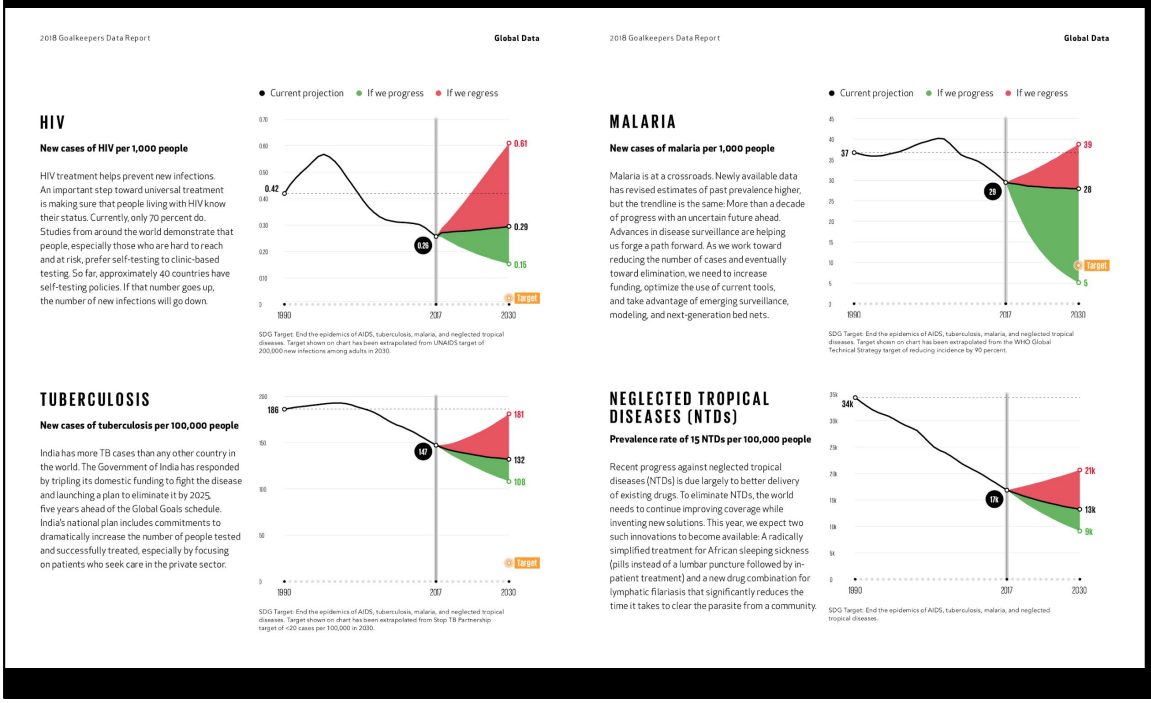
Agenda: to show how important and underappreciated visualisation is in international development work. The lack of good visualizations can cause philanthropic foundations to fail to see the value of computerization, including machine learning. Even if data is collected, it is often not shared and not used well. Proper visualization is often the last step to close the gap in understanding and appreciating epidemics in the developing world.

Problems and Approaches

- Most problems in health and ag are not lending themselves to closed-form solutions, and may never will. (Concrete examples include under-5 mortality gaps, yield gaps, and yield prediction.)
- Machine learning is the current most promising approach toward addressing these questions, because it is capable of ingesting large, multivariate datasets across space and time, and has empirically proven itself capable of breakthroughs in other fields.
- The main bottleneck in the application of machine learning is scarcity of data.

https://sustainabledevelopment.un.org

Our focus is on these three SDGs.



2018 Goalkeepers Data Report **Global Data**

HIV

New cases of HIV per 1,000 people

HIV treatment helps prevent new infections. An important step toward universal treatment is making sure that people living with HIV know their status. Currently, only 70 percent do. Studies from around the world demonstrate that people, especially those who are hard to reach and at risk, prefer self-testing to clinic-based testing. So far, approximately 40 countries have self-testing policies. If that number goes up, the number of new infections will go down.

SDG Target: End the epidemics of AIDS, tuberculosis, malaria, and neglected tropical diseases. Target shown on chart has been extrapolated from UNAIDS target of 200,000 new infections among adults in 2030.

MALARIA

New cases of malaria per 1,000 people

Malaria is at a crossroads. Newly available data has revised estimates of past prevalence higher, but the trendline is the same: More than a decade of progress with an uncertain future ahead. Advances in disease surveillance are helping us forge a path forward. As we work toward reducing the number of cases and eventually toward elimination, we need to increase funding, optimize the use of current tools, and take advantage of emerging surveillance, modeling, and next-generation bed nets.

SDG Target: End the epidemics of AIDS, tuberculosis, malaria, and neglected tropical diseases. Target shown on chart has been extrapolated from the WHO Global Technical Strategy target of reducing incidence by 90 percent.

TUBERCULOSIS

New cases of tuberculosis per 100,000 people

India has more TB cases than any other country in the world. The Government of India has responded by tripling its domestic funding to fight the disease and launching a plan to eliminate it by 2025, five years ahead of the Global Goals schedule. India's national plan includes commitments to dramatically increase the number of people tested and successfully treated, especially by focusing on patients who seek care in the private sector.

SDG Target: End the epidemics of AIDS, tuberculosis, malaria, and neglected tropical diseases. Target shown on chart has been extrapolated from Stop TB Partnership target of <20 cases per 100,000 in 2030.

NEGLECTED TROPICAL DISEASES (NTDs)

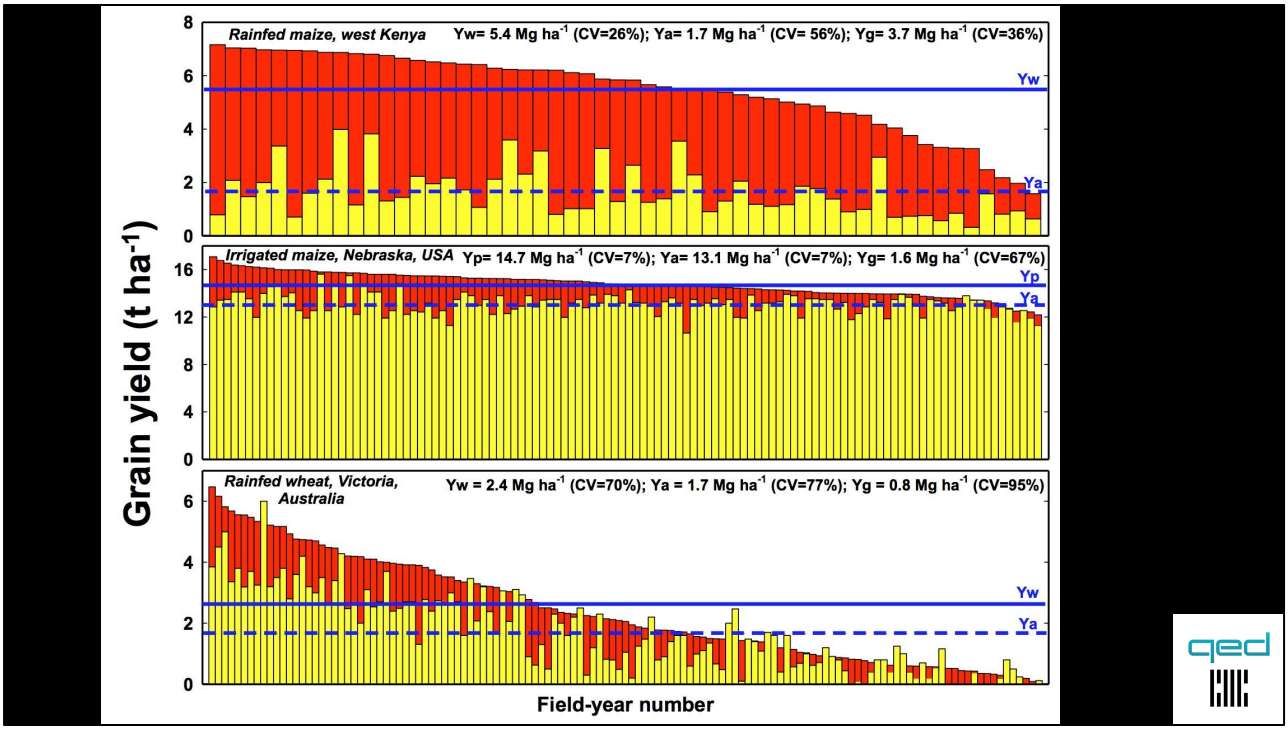
Prevalence rate of 15 NTDs per 100,000 people

Recent progress against neglected tropical diseases (NTDs) is due largely to better delivery of existing drugs. To eliminate NTDs, the world needs to continue improving coverage while inventing new solutions. This year, we expect two such innovations to become available: A radically simplified treatment for African sleeping sickness (pills instead of a lumbar puncture followed by in-patient treatment) and a new drug combination for lymphatic filariasis that significantly reduces the time it takes to clear the parasite from a community.

SDG Target: End the epidemics of AIDS, tuberculosis, malaria, and neglected tropical diseases.

https://www.gatesfoundation.org/goalkeepers/report

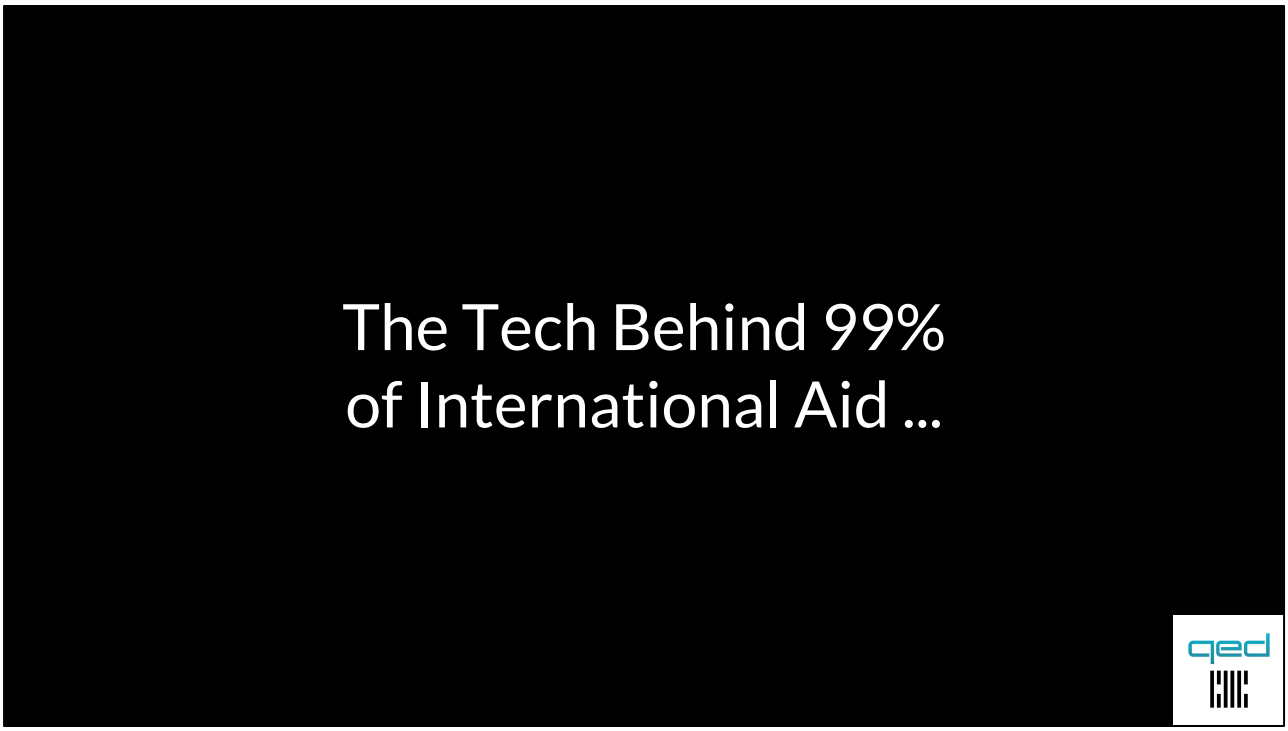
One of very few examples of good visualizations in the aid world. Only very general stats…



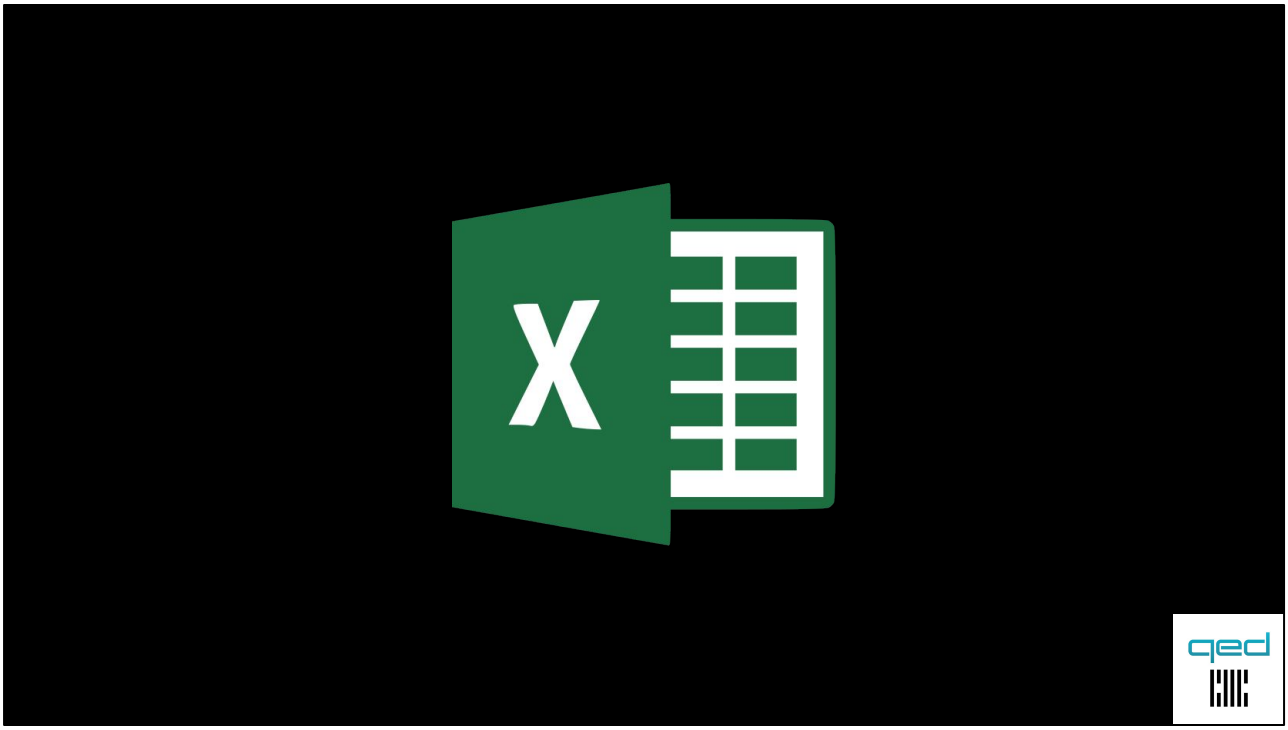
Red - predicted yield
Yellow - actual yield

Many areas in the Global South are not attaining anywhere near their agricultural yield potential.

In the 2018 Goalkeepers Report, the Gates Foundation simply listed “insufficient data” for agriculture.
The world has nothing to show for our progress toward the agricultural SDGs!



- Reasons for scarcity of data:
  - Lack of technologies (software and hardware) for low-cost and reliable collection.
  - Outdated views about the role of technology. (In Silicon Valley, technology is the Holy Grail; in international aid, we are often still just “the help”.)
  - **Inability of most current practitioners to communicate the value of data to higher-level management and general audiences.**

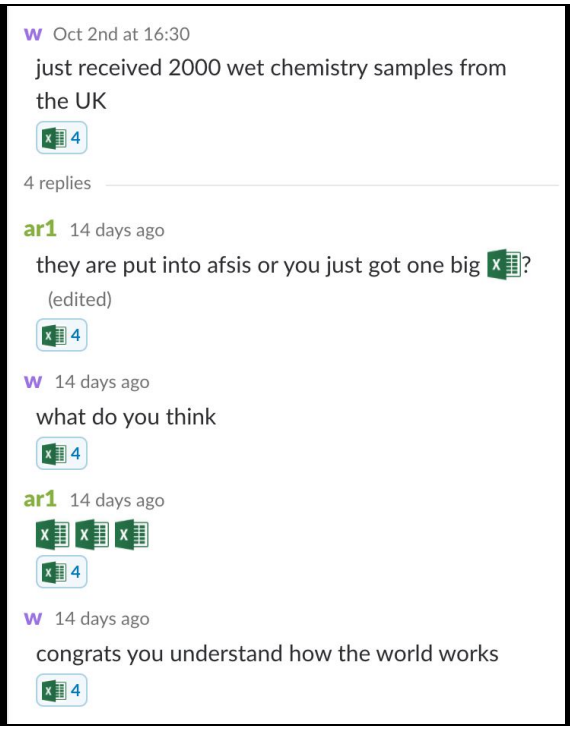


As you've probably guessed, Excel.

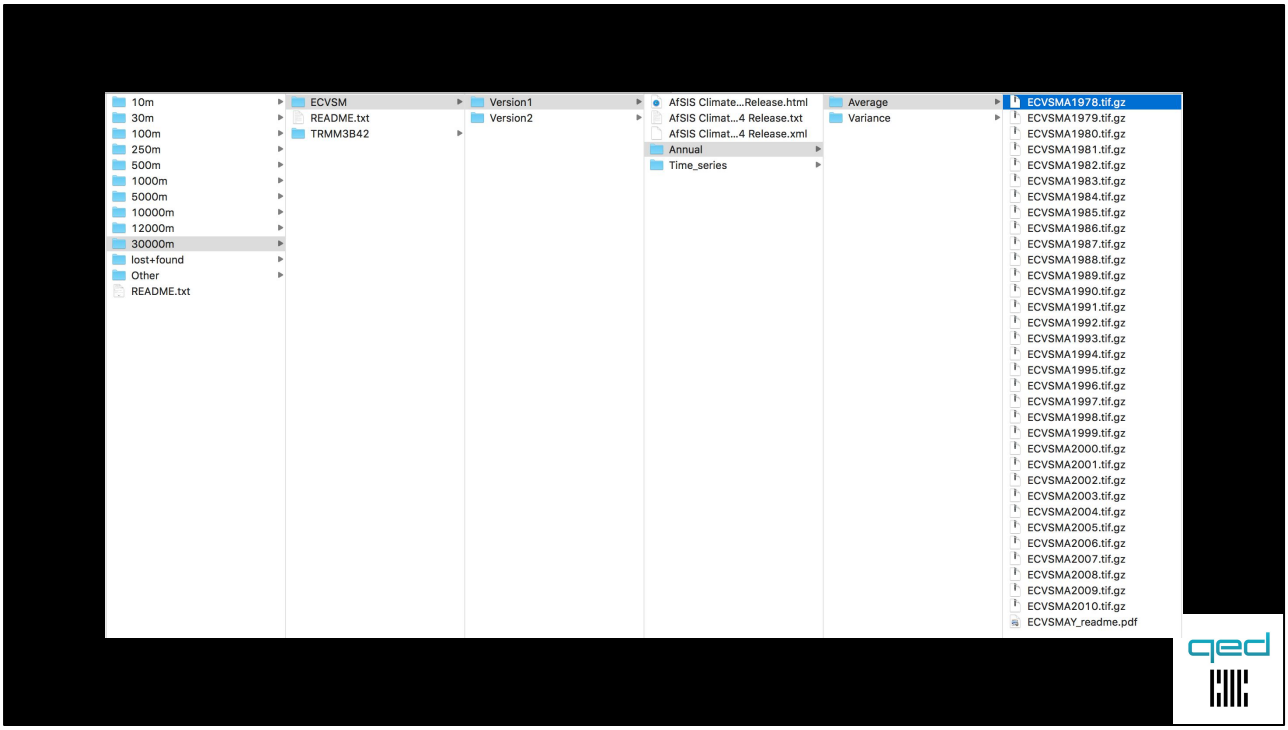
Excel

Don't get me wrong, Excel is a great tool!

Excel



But it's not enough when you care about data quality and integrity, or need a larger scale.



In most aid projects, data “just sits there”. No one interacts with it.

The data is collected by field enumerators using un-expressive systems with little to no visualization of what is going on. If there is a systematic error in the data collection process, it is often discovered at headquarters many months late --- and all data collected to date must either be manually corrected with painstaking effort or even recaptured in the field.

On paper, data that was collected with taxpayer or philanthropic money is supposed to be public and shared. But in practice, scientists are typically making no effort to share data with others. The two most common reasons given are:

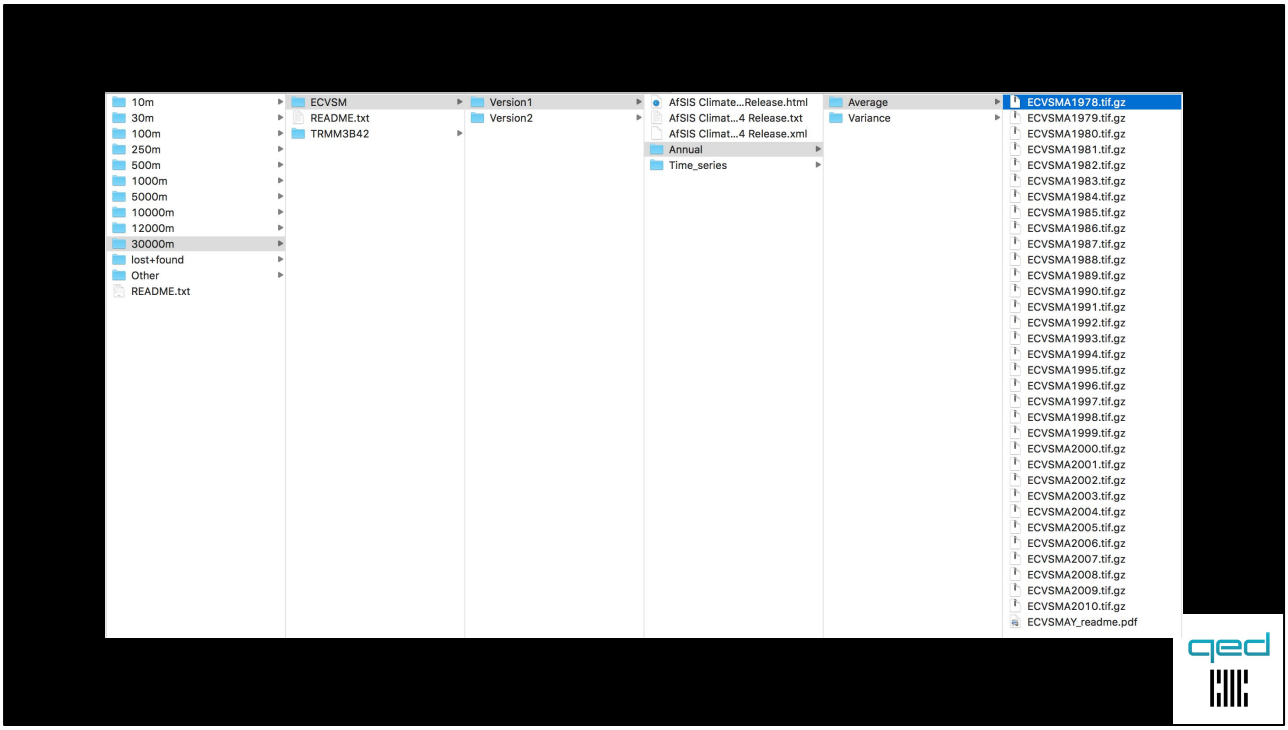
1. **Sanitation:** The data still needs to be cleaned. (Manually.)
2. **Publication:** We are waiting to publish our paper first. (No upper bound as to when this may happen.)

Lastly, if a project manages to collect appreciable data for meaningful inferences, then the data and predictions are presented to stakeholders by scientists and math modelers. Often they have terrible presentation skills, drowning the audience in an emotionless sea of numbers and command-line output. Such presentations fail to communicate what is important, and do not reassure stakeholders about the

importance of big data.

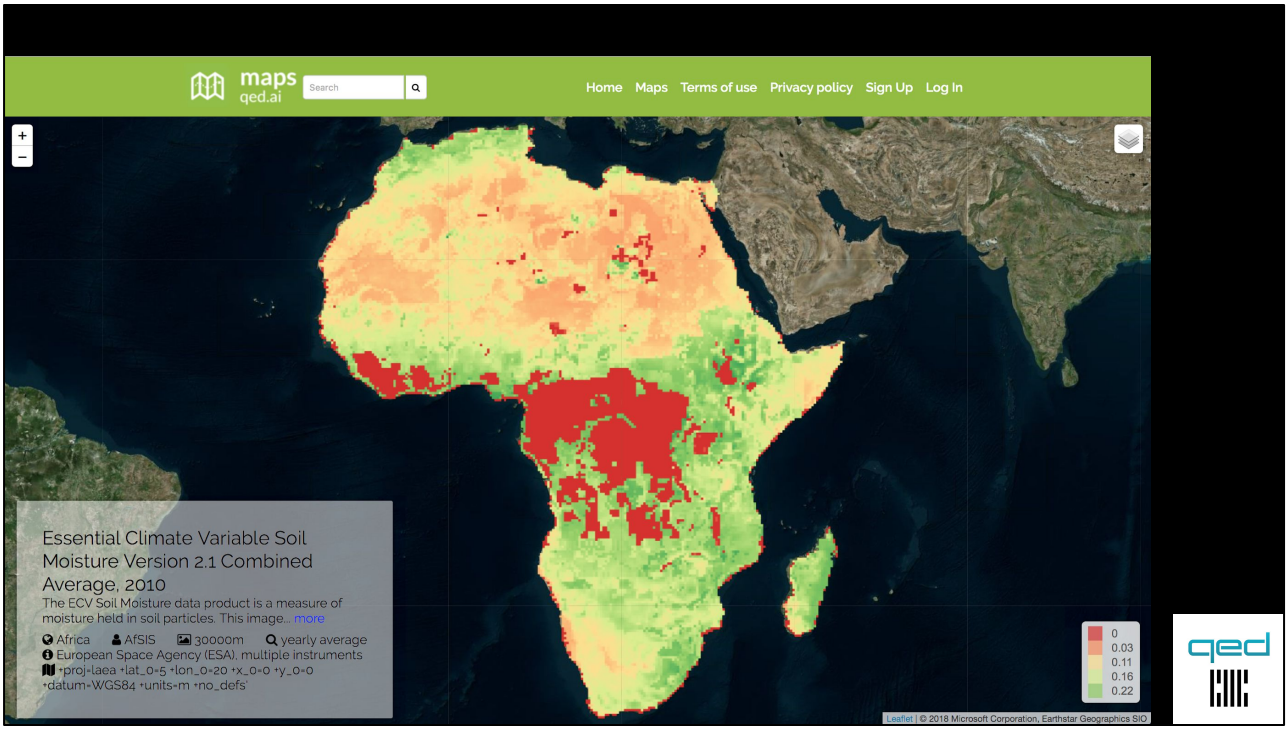
Even if the data is shared, it’s often done in a way that is difficult to consume. For example, here you can see gigabytes of maps on an internet disk, with no previews.

- Reasons for scarcity of data:
  - Lack of technologies (software and hardware) for low-cost and reliable collection
    - that's what we're working on now, with great progress
  - Outdated views about the role of technology. (In Silicon Valley, technology is the Holy Grail; in international aid, we are often still just "the help".)
  - **Inability of most current practitioners to communicate the value of data to higher-level management and general audiences**
    - **that's where we need help**



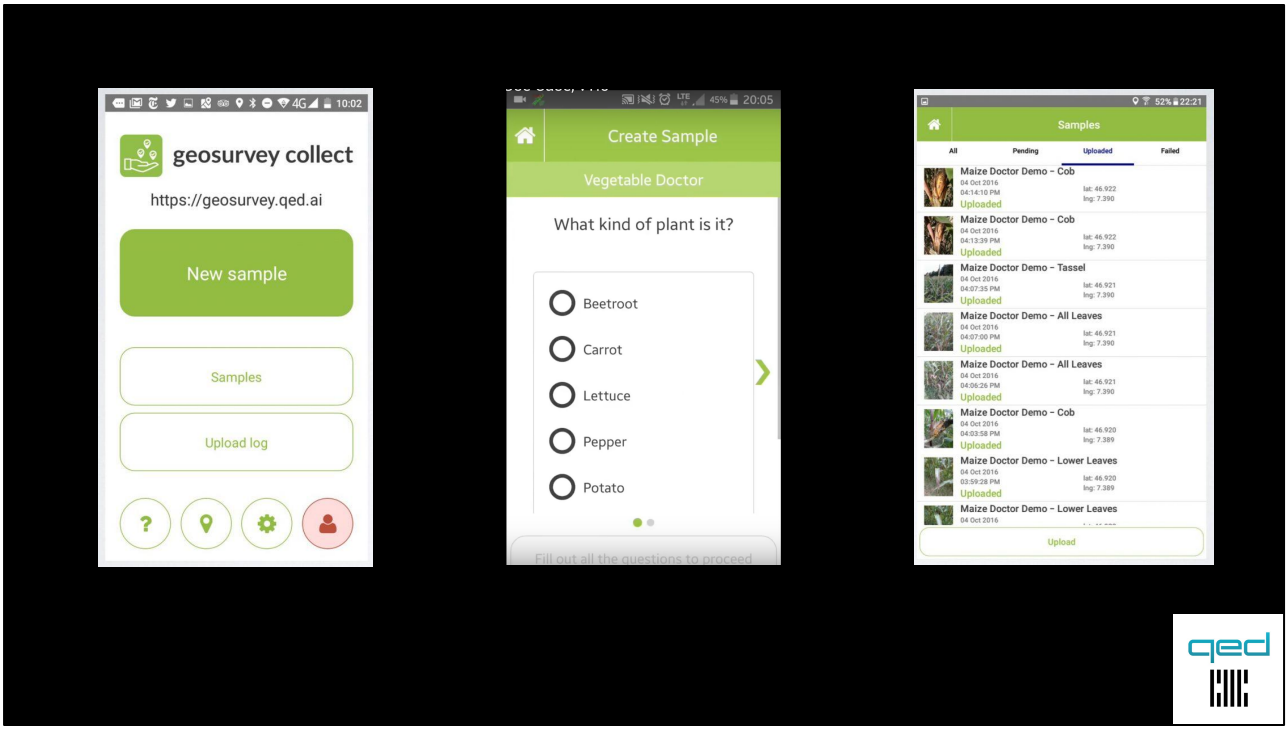
Before: ftp://africagrids.net/

- It has been like this for decades.
- Typical file is several gigabytes. No metadata or preview. Why should I download anything? Internet in some parts of SSA can also be very expensive, a precious resource.



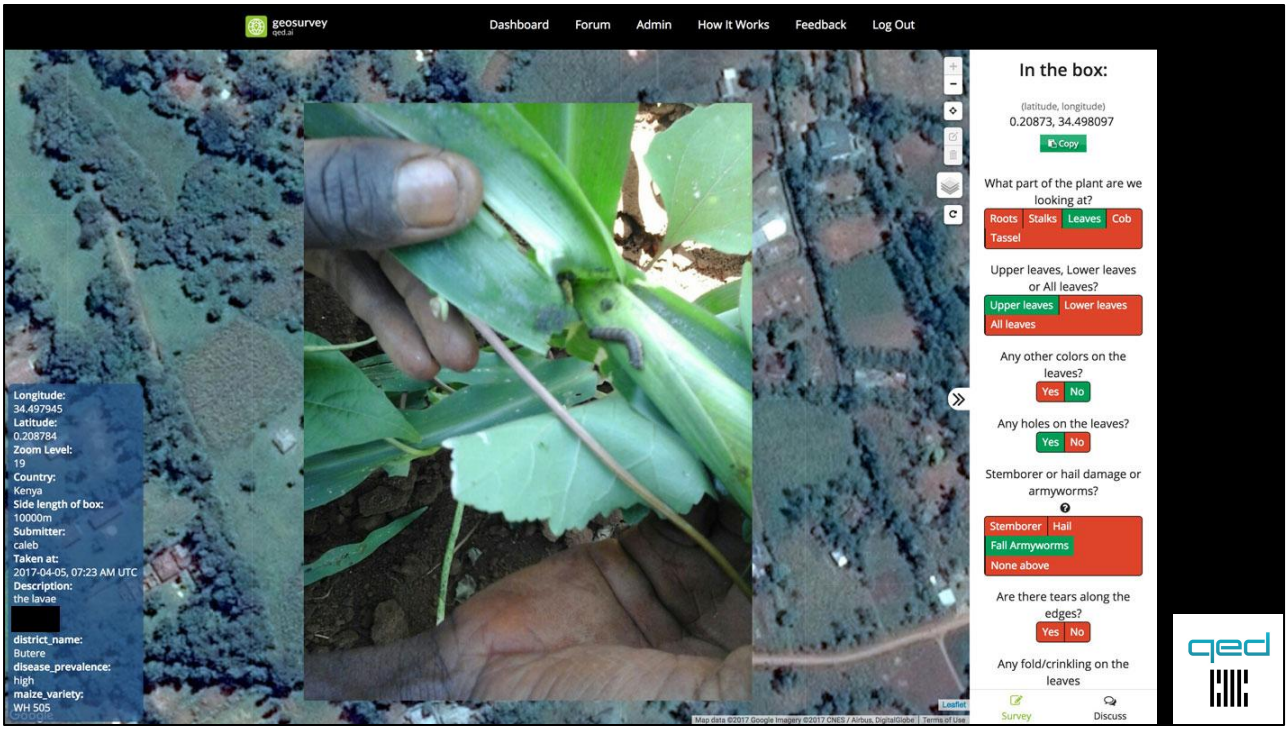
After: https://maps.qed.ai

- Real-time re-rendering of raster data at multiple zoom levels.

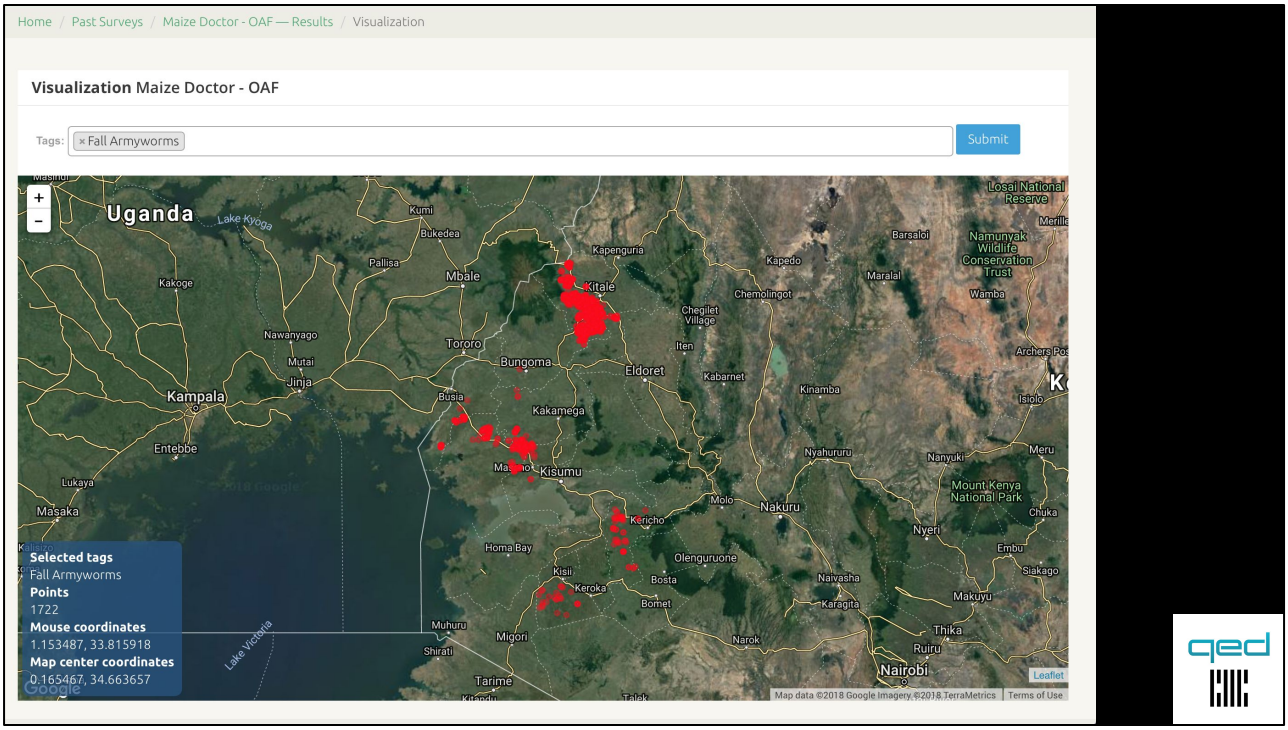


Remote data collection. Extension workers or farmers fill out field data regarding crop conditions, varieties, and farming practices. Georeferenced and timestamped.

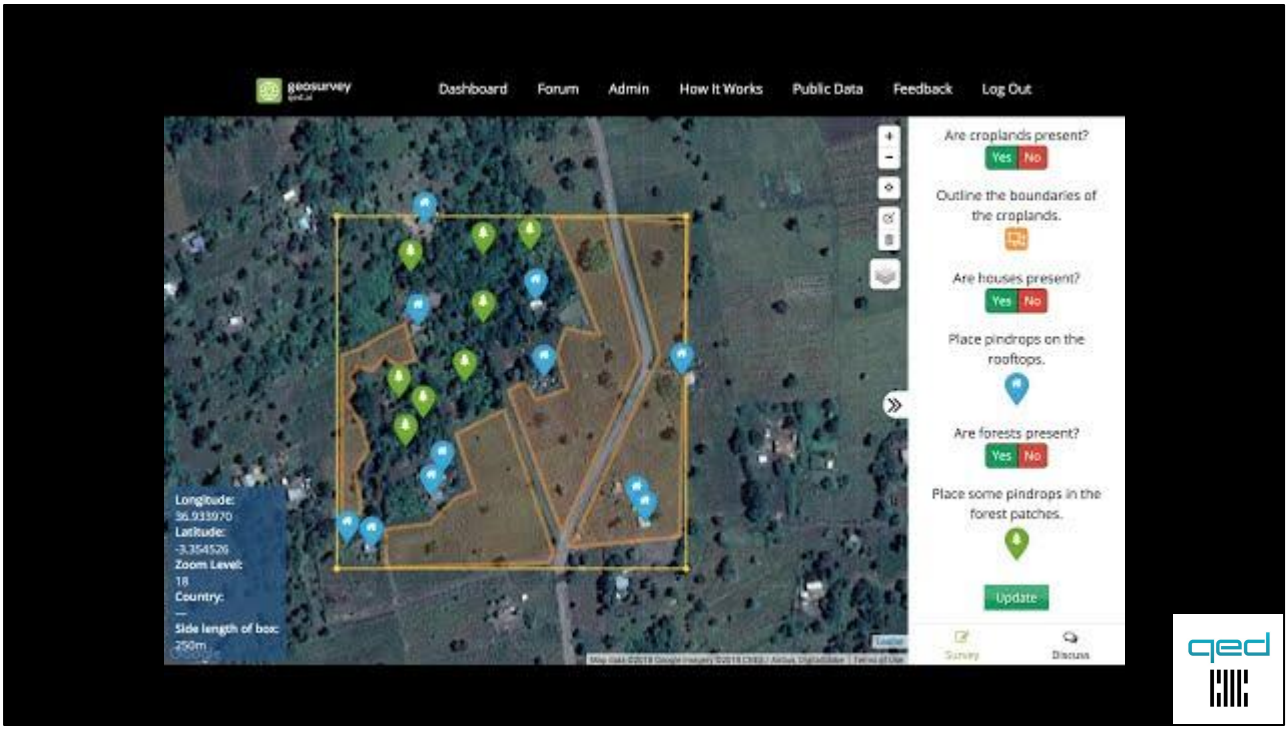
https://qed.ai/geosurvey-collect/



Remote plant health assessment. Experts can make immediate judgments or be guided by interactive decision trees on the right.
https://qed.ai/geosurvey-collect/



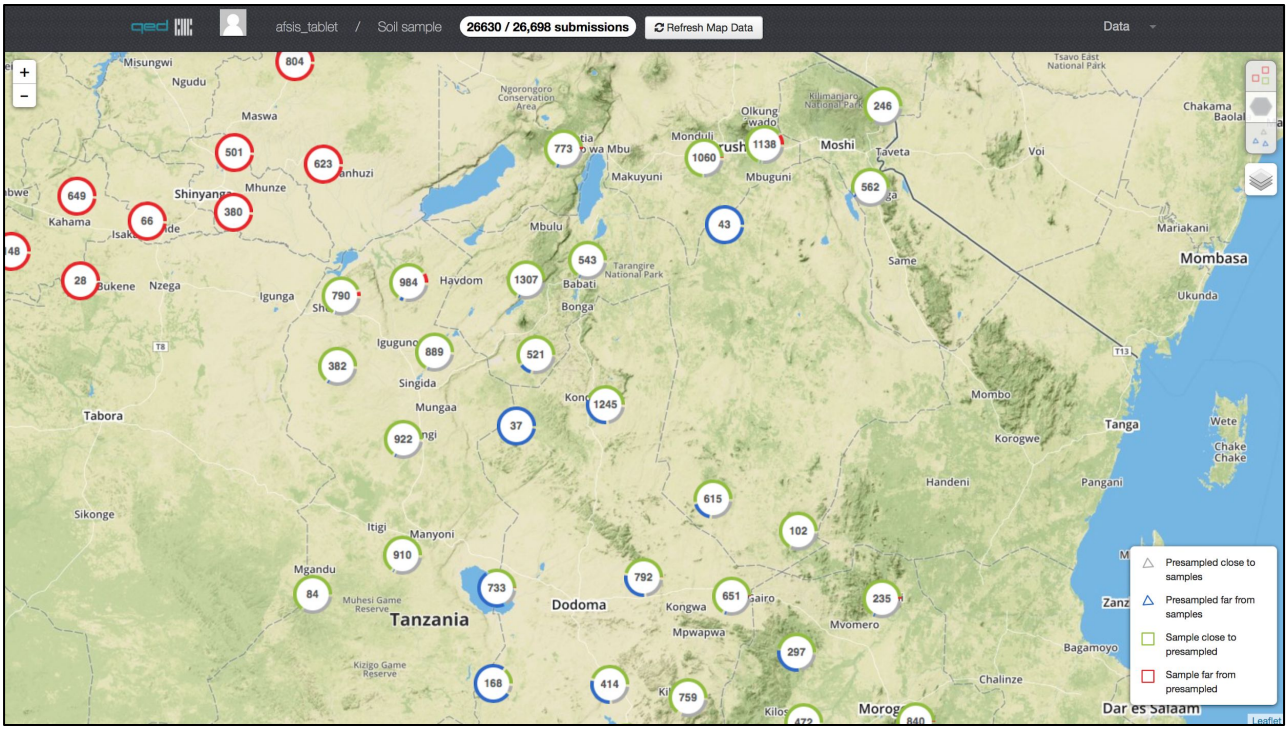
- Geosurvey Collect: Maize Doctor OAF
  - This is real-time rendering of maize illnesses, direct from the field. We are not aware of any other farming cooperative in the world with this kind of functionality.
  - Demonstrate usage of tags to filter incidences of FAW,
  - Improves geospatial strategy: what problems are cropping up and where; what varieties are most susceptible to these problems. Consider changing varieties in certain districts next season.



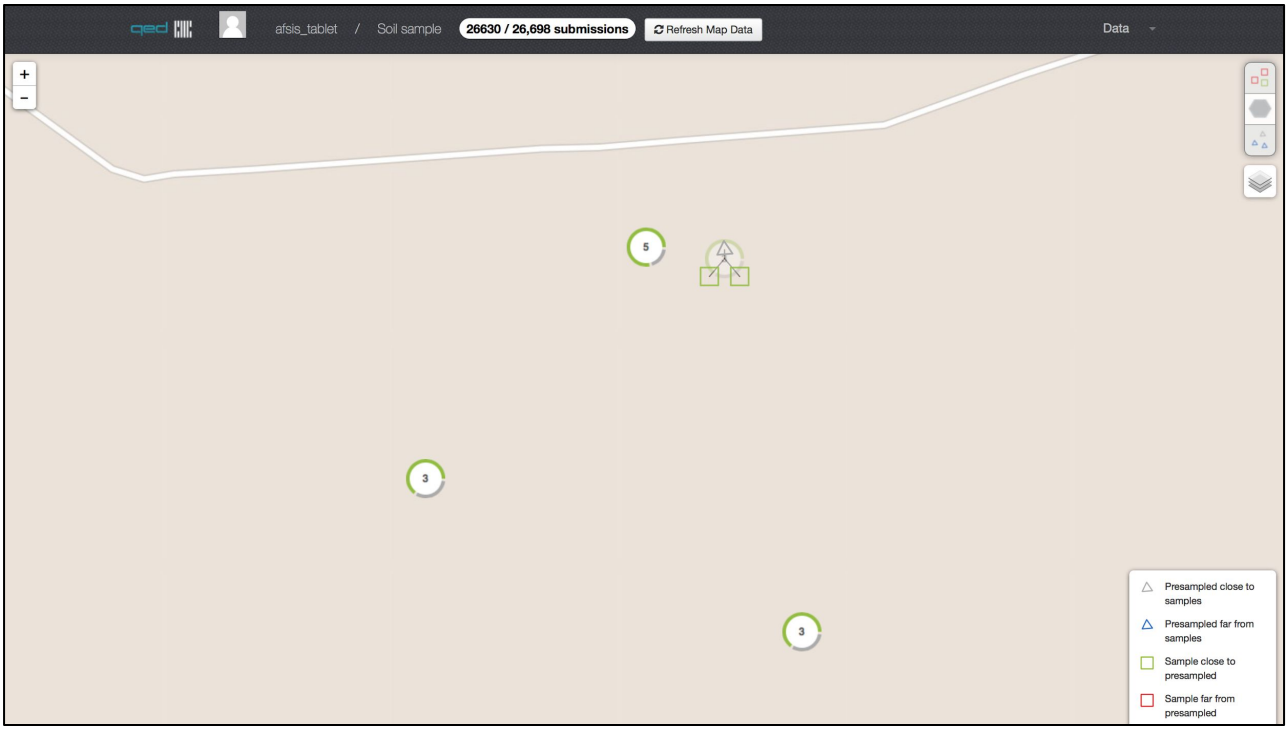
Tagging shapes and points:
https://qed.ai/geosurvey/
https://www.youtube.com/watch?v=IuIXR9t71Mg

Allowed us to train machine learning to go from this…

…to this. You can go to https://maps-mini.qed.ai to play with this yourself or download the data.



Pre-sampled points (where sample should have been collected) vs. sampled points (where they've been actually collected).



True story from QED: Enumerators going from bar to bar along the roads of Ethiopia rather than visiting the intended destinations for soil sampling.

Soil samples and processing facility

Table 1. CHAMPS Implementation Overview

	Stillbirths	Neonates	Infants	Children 1-5 years	TOTAL
All cases					
All Initial Death Notification received	**74**	**83**	**79**	**98**	**334**
Unique notifications received (de-duplicated notifications)	74	83	79	98	334
Deaths that proceeded to eligibility	68 (91.9%)	73 (88.0%)	68 (86.1%)	75 (76.5%)	284 (85.0%)
Able to contact family to conduct eligiblilty screening	68 (91.9%)	73 (88.0%)	68 (86.1%)	74 (75.5%)	283 (84.7%)
CHAMPS-eligible cases†	60 (81.1%)	69 (83.1%)	61 (77.2%)	67 (68.4%)	257 (76.9%)
MITS only					
MITS-eligible cases	**36 (60.0%)**	**47 (68.1%)**	**54 (88.5%)**	**49 (73.1%)**	**186 (72.4%)**
MITS consent requested	36 (60.0%)	47 (68.1%)	53 (86.9%)	48 (71.6%)	184 (71.6%)
Refused MITS Consent	5 (8.3%)	5 (7.2%)	11 (18.0%)	11 (16.4%)	32 (12.5%)
Family consented to MITS	31 (51.7%)	42 (60.9%)	42 (68.9%)	37 (55.2%)	152 (59.1%)
MITS procedure conducted	30 (50.0%)	39 (56.5%)	42 (68.9%)	37 (55.2%)	148 (57.6%)
HIV testing information received‡	*27 (45.0%)*	*35 (50.7%)*	*42 (68.9%)*	*33 (49.3%)*	*137 (53.3%)*
Malaria testing information received‡	*29 (48.3%)*	*38 (55.1%)*	*41 (67.2%)*	*33 (49.3%)*	*141 (54.9%)*
Bacteriology testing information received‡	*29 (48.3%)*	*38 (55.1%)*	*42 (68.9%)*	*35 (52.2%)*	*144 (56.0%)*
Tuberculosis testing information received‡	*29 (48.3%)*	*38 (55.1%)*	*42 (68.9%)*	*34 (50.7%)*	*143 (55.6%)*
TAC testing information received‡	NA	*NA*	*NA*	*NA*	NA
Site pathology information received‡	*26 (43.3%)*	*35 (50.7%)*	*40 (65.6%)*	*35 (52.2%)*	*136 (52.9%)*
Clinical records abstracted	5 (8.3%)	35 (50.7%)	37 (60.7%)	32 (47.8%)	109 (42.4%)
VA completed	NA	NA	NA	NA	NA
CPL pathology report completed	NA	NA	NA	NA	NA
DeCoDe completed	18 (30.0%)	14 (20.3%)	30 (49.2%)	24 (35.8%)	86 (33.5%)
Final family follow-up completed	NA	NA	NA	NA	NA
Non-MITS only					
MITS ineligible cases	**27 (45.0%)**	**25 (36.2%)**	**12 (19.7%)**	**22 (32.8%)**	**86 (33.5%)**
Non-MITS consent requested	31 (51.7%)	28 (40.6%)	21 (34.4%)	34 (50.7%)	114 (44.4%)
Refused Non-MITS Consent	*2 (3.3%)*	*1 (1.4%)*	*2 (3.3%)*	*4 (6.0%)*	*9 (3.5%)*
Family consented to Non-MITS	29 (48.3%)	27 (39.1%)	19 (31.1%)	30 (44.8%)	105 (40.9%)

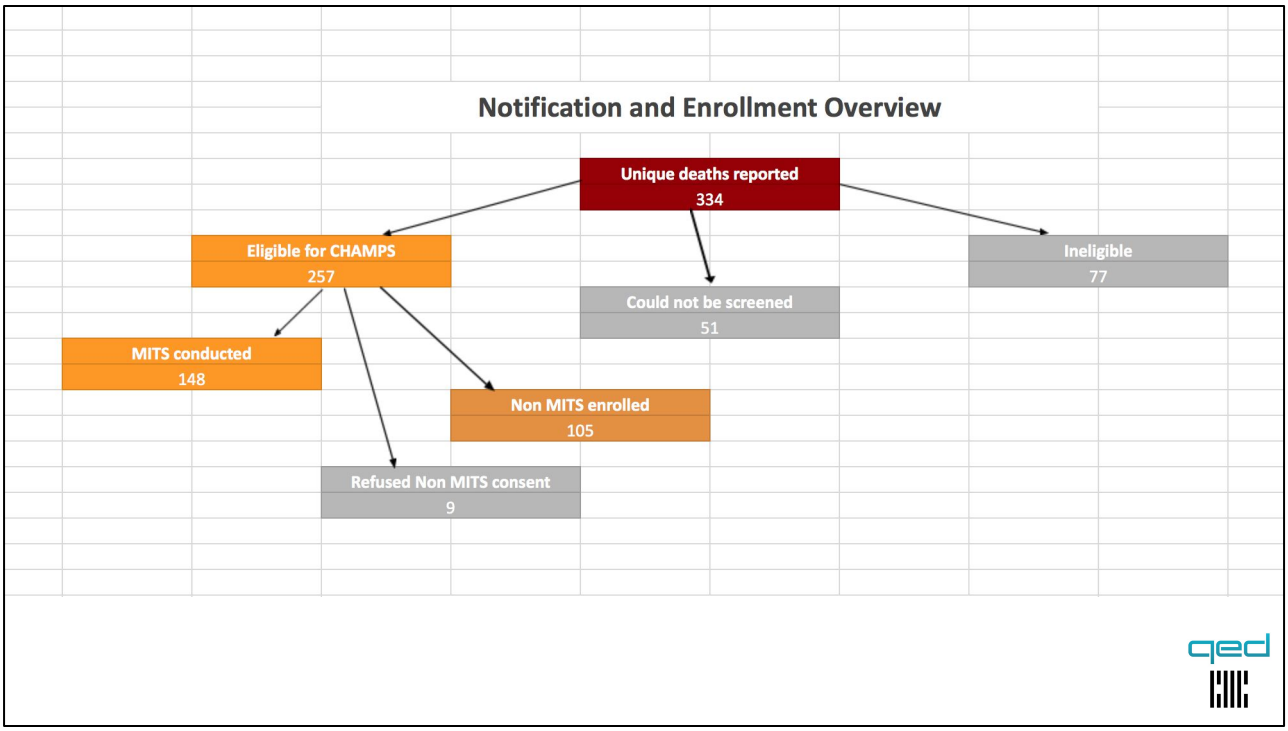
Denominator used for yield percentages: Unique notifications received (de-duplicated notifications)

Denominator used for yield percentages: CHAMPS-eligible cases†

Denominator used for yield percentages: CHAMPS-eligible cases†

Champs

- Auto updated
- 20 year long study
- Monitoring of progress
- Already 2 years in
- Graphs poor
- First outcomes ()

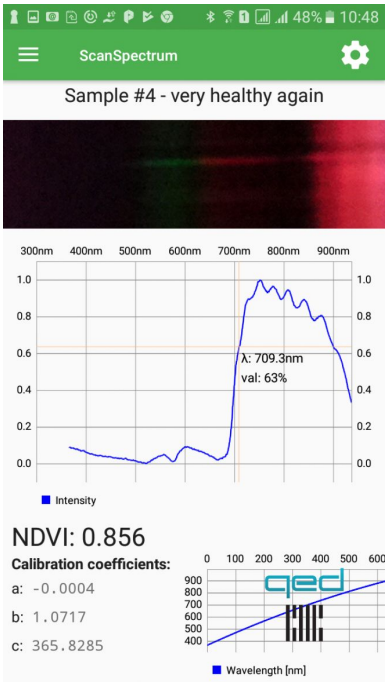


Champs

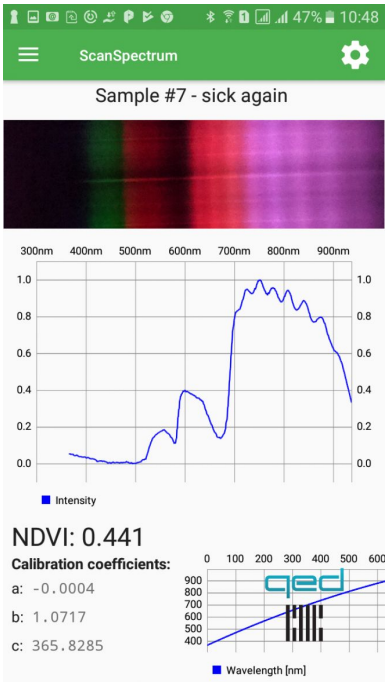
- Auto updated
- 20 year long study
- Monitoring of progress
- Already 2 years in
- Graphs poor
- First outcomes ()
-

CHAMPS labs and clinicians

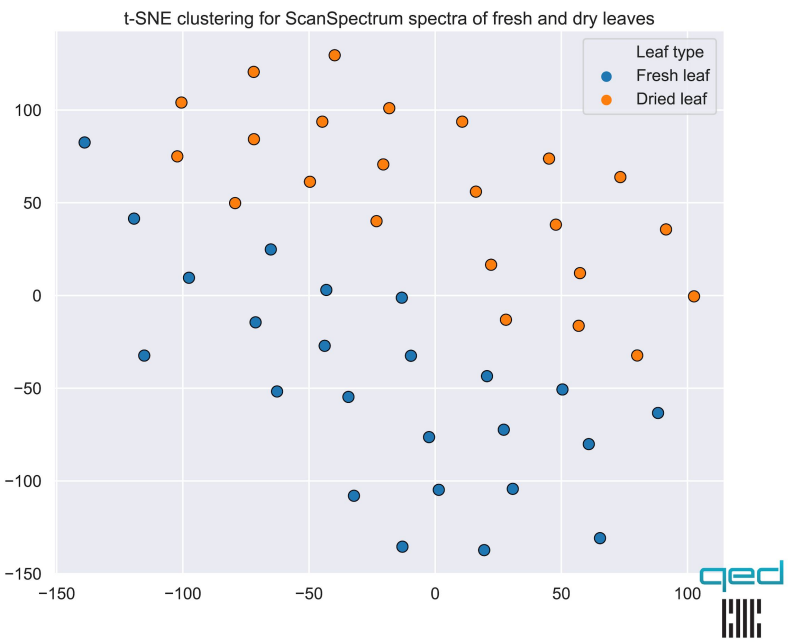
Low-cost handheld spectrometer - https://qed.ai/scanspectrum/
(This is an NIR device)



Work in progress, but promising

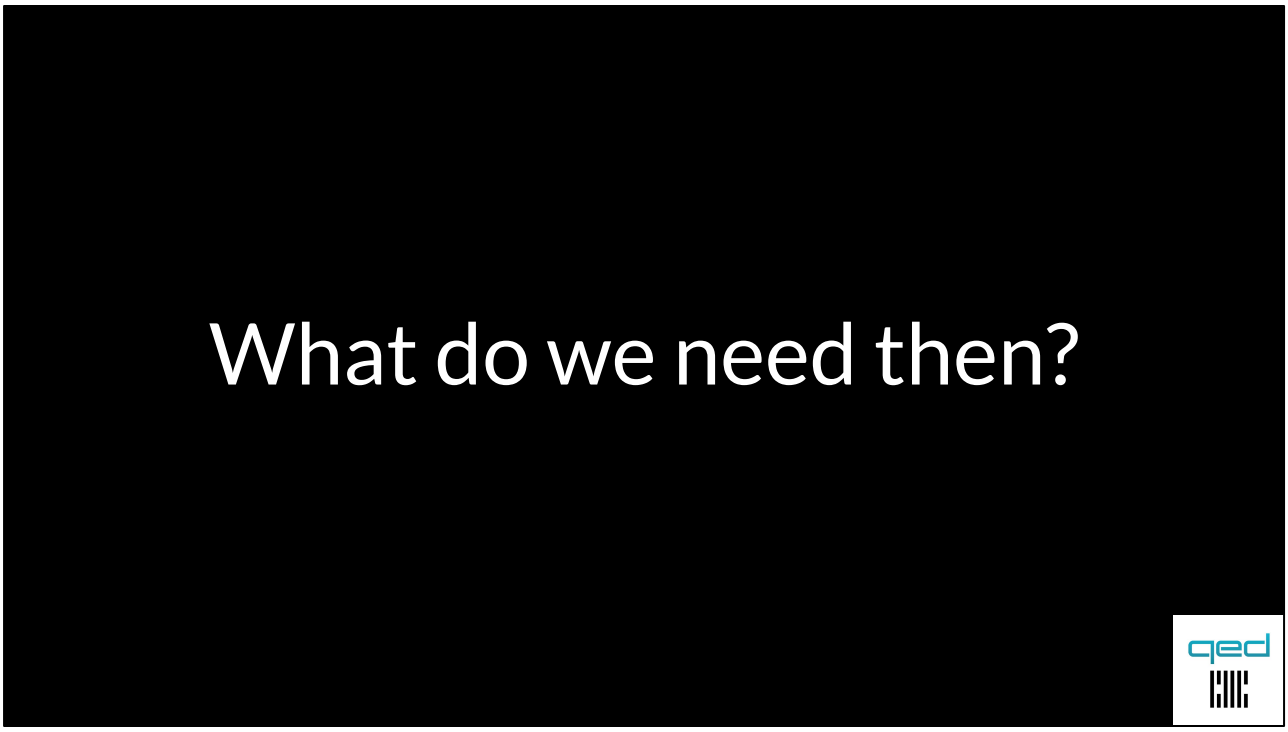


Work in progress, but promising

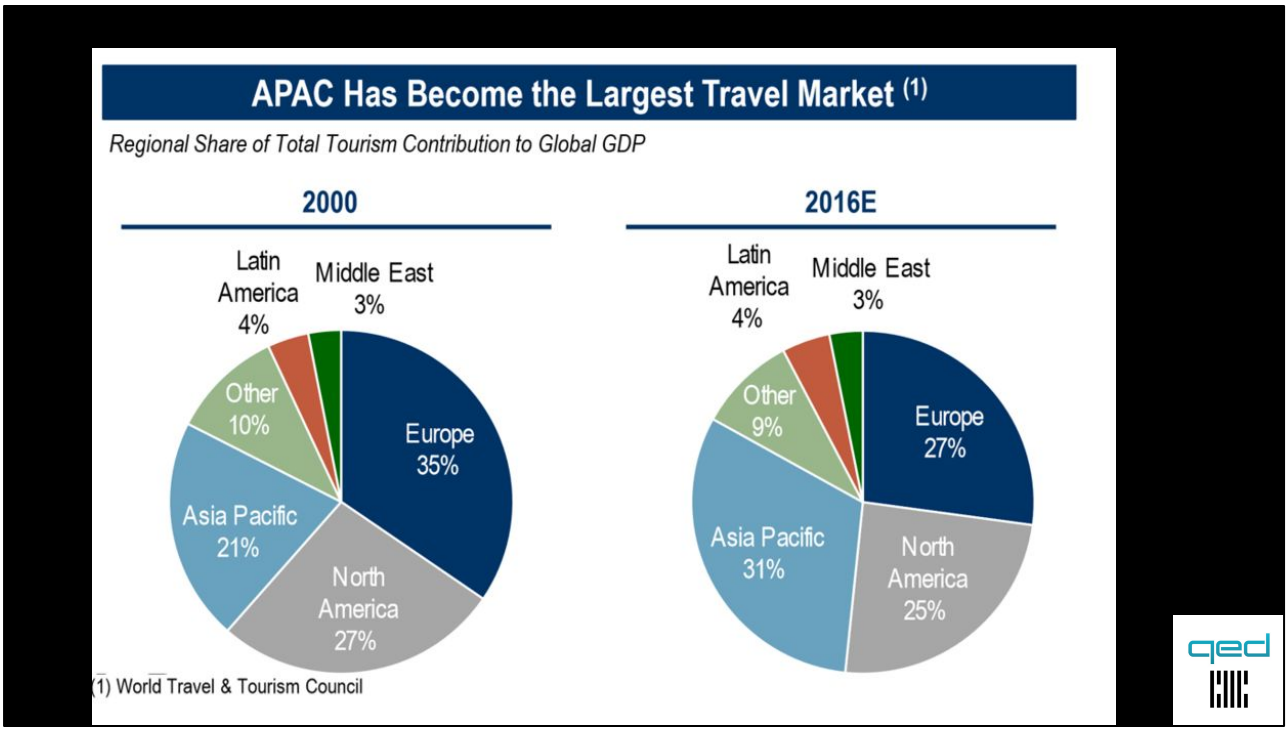


t-distributed Stochastic Neighbor Embedding (t-SNE) visualization of moist leaves vs. dry leaves, successfully discriminated by applying this clustering algorithm to ScanSpectrum spectra.

Exploratory / Explanatory

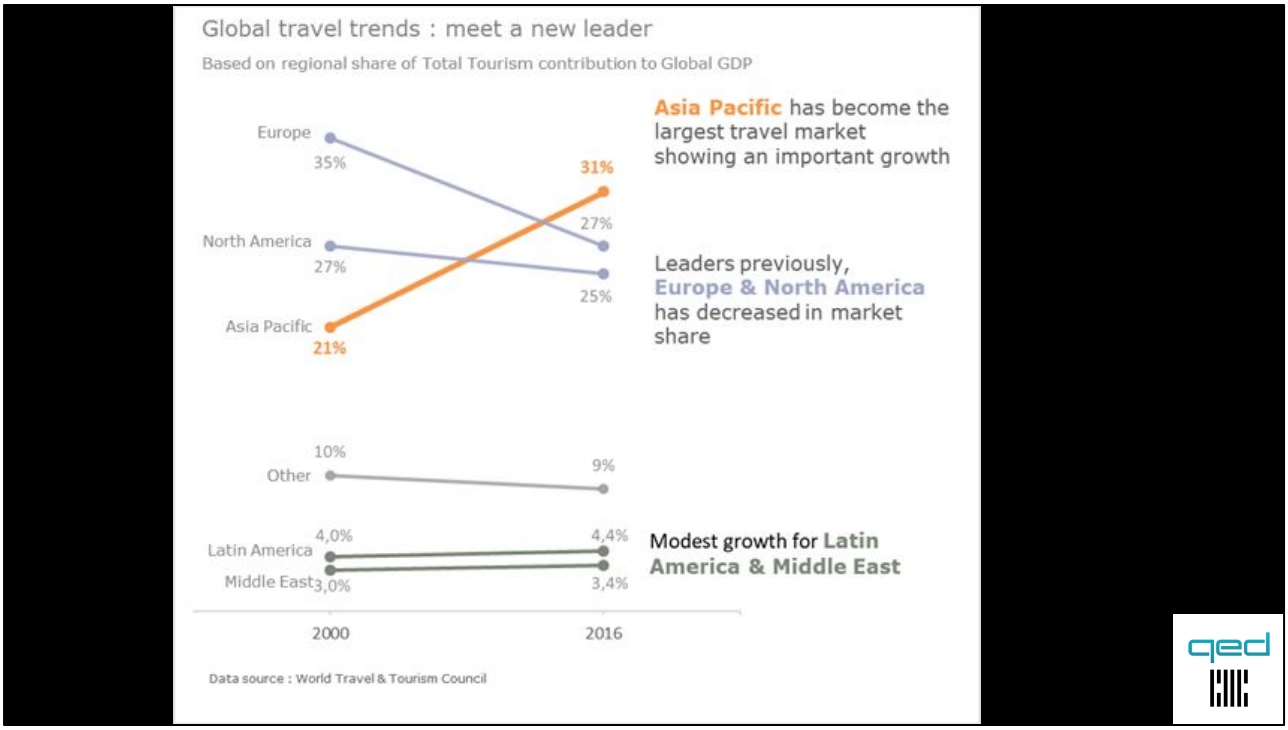


Speaking more personally, data visualization is also important to the success of our own company. If projects do not secure high quality visualizations --- which are often the final deliverables in international summary presentations --- then that sabotages lifetimes of effort by everyone involved.



We need

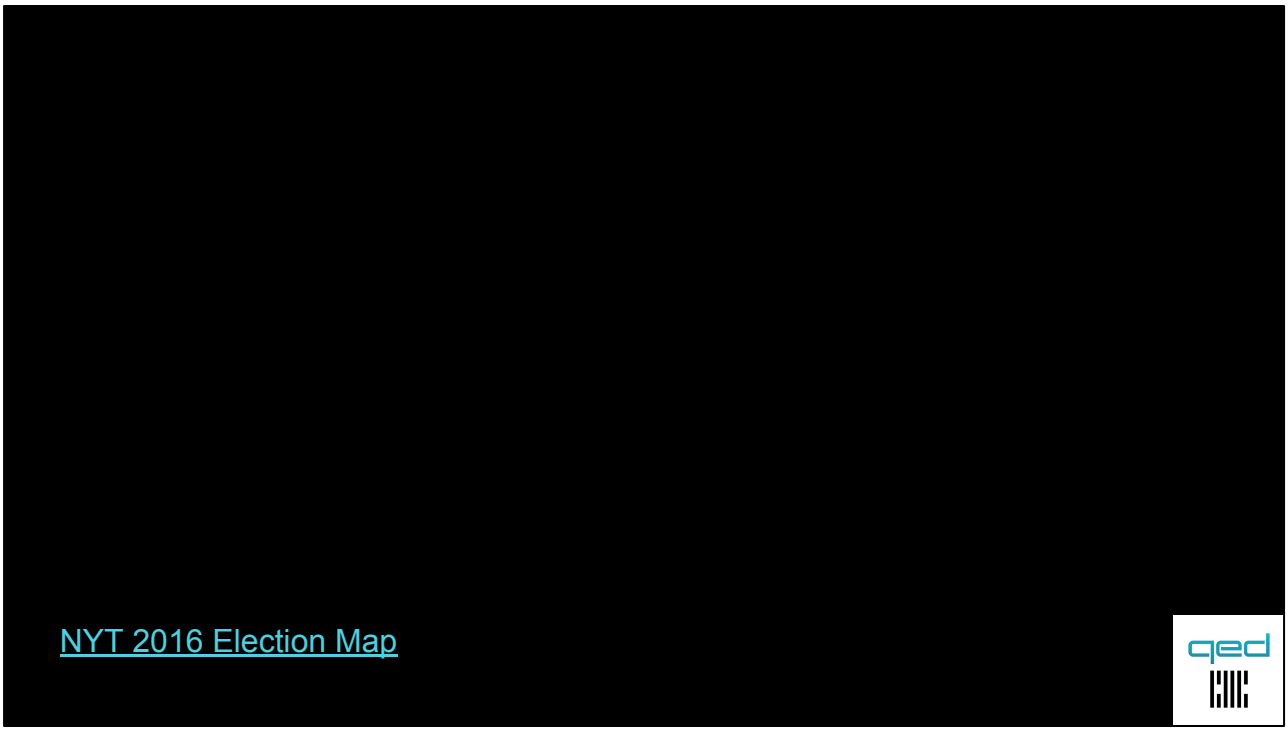
- knowledge of chart types, their properties
- understanding about how the brain works to make information more digestable
- aesthetic sense in data representation
- but we need not only aesthetic sense, but also statistical sense.



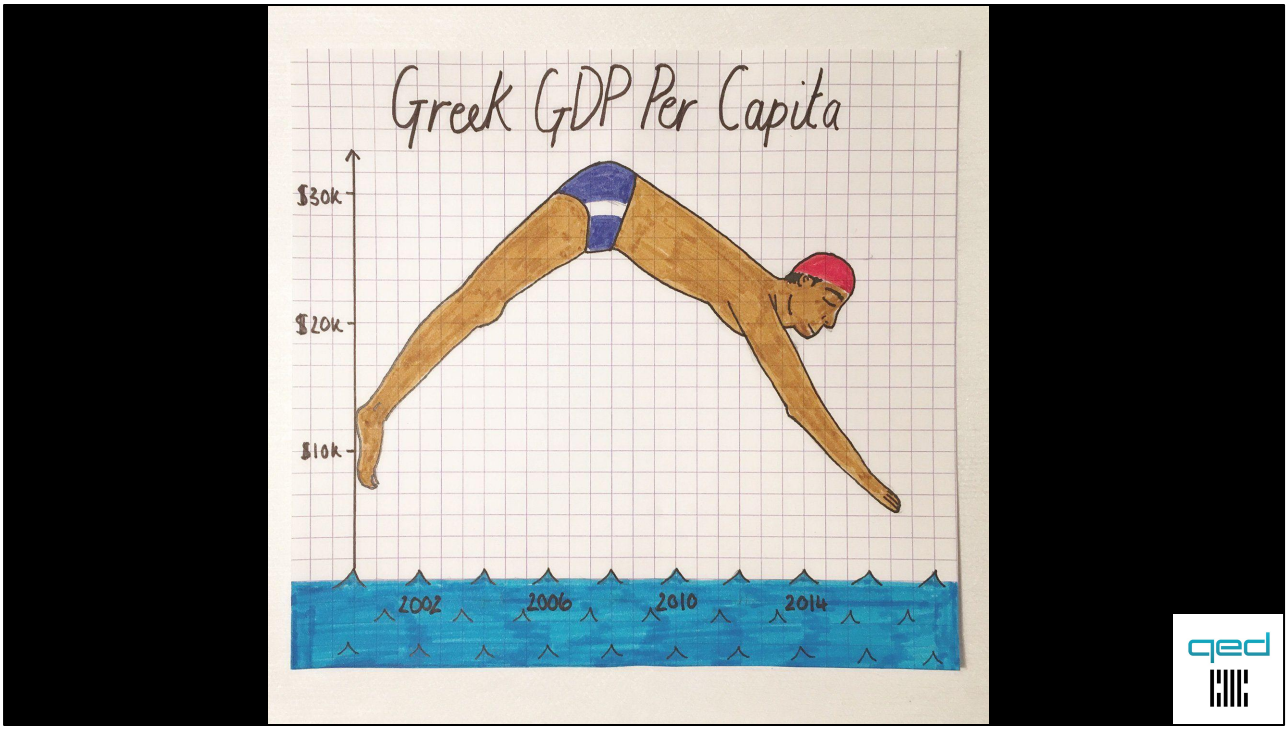
We need

- knowledge of chart types, their properties
- understanding about how the brain works to make information more digestable
- aesthetic sense in data representation
- but we need not only aesthetic sense, but also statistical sense.


animating
data
a quick short video from
storytelling WITH data
qed



Hey, we need not only aesthetics and statistics, but also algorithmics.



Yet we need not only aesthetics, statistics, and algorithmics, but also emotion, creativity, and humor!

qed.ai/jobs

@TweetQED
@mlazowik

info@qed.ai

Join us!