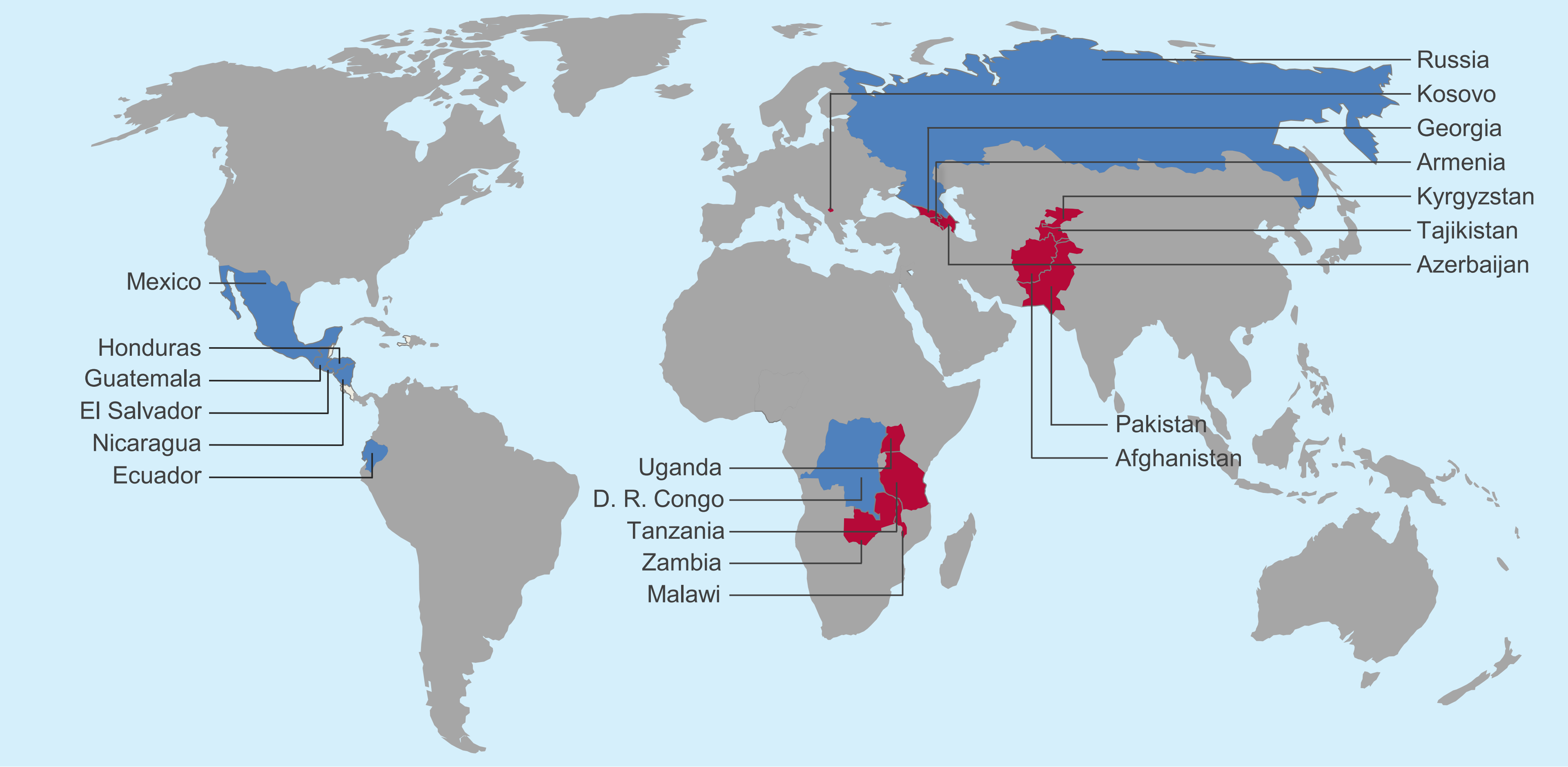
FINCA ValiData

Using Real-Time Algorithms to Improve Field Data Collection

FINCA
ValiData®
(patent pending)

FINCA®

Who are we?



- HH Composition and Consumption
- Living Standards (Health, Education, Basic Services)
- Business Performance, Seasonality
- Employment and Job Creation
- Women's Empowerment
- Client Aspirations

- Demand and Use of Services
- Client Satisfaction
- Competitive Positioning, Loyalty and Brand Awareness
- Branch-level Performance
- Mapping and Efficiency Analysis

Collecting Clean Data with ValiData

3 Review survey statistics.

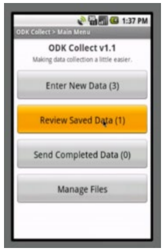
Follow-up on flagged responses.

Edit data.

Automated emails
and reports

Call Center
Field Survey

1 Collect data.



- Input Prompts
- Form Logic
- Field Constraints
- Filed Calculations
- Enumerator Hints
- Data Normalization

Mobile

or Wi-Fi

FINCA

ValiData®

Converted

Data

4 Export and
analyze data

- SPSS
- R
- Stata
- Excel

2 Analyze and flag problematic data.
Observe and report on surveyor behavior.

- Customized Validation and Outlier Check Rules (robust regressions, conditional probabilities, logistic modeling)
- Behavioral Observations (support vector machines and random forests)

FINCA

Avoid Mistakes before they Happen

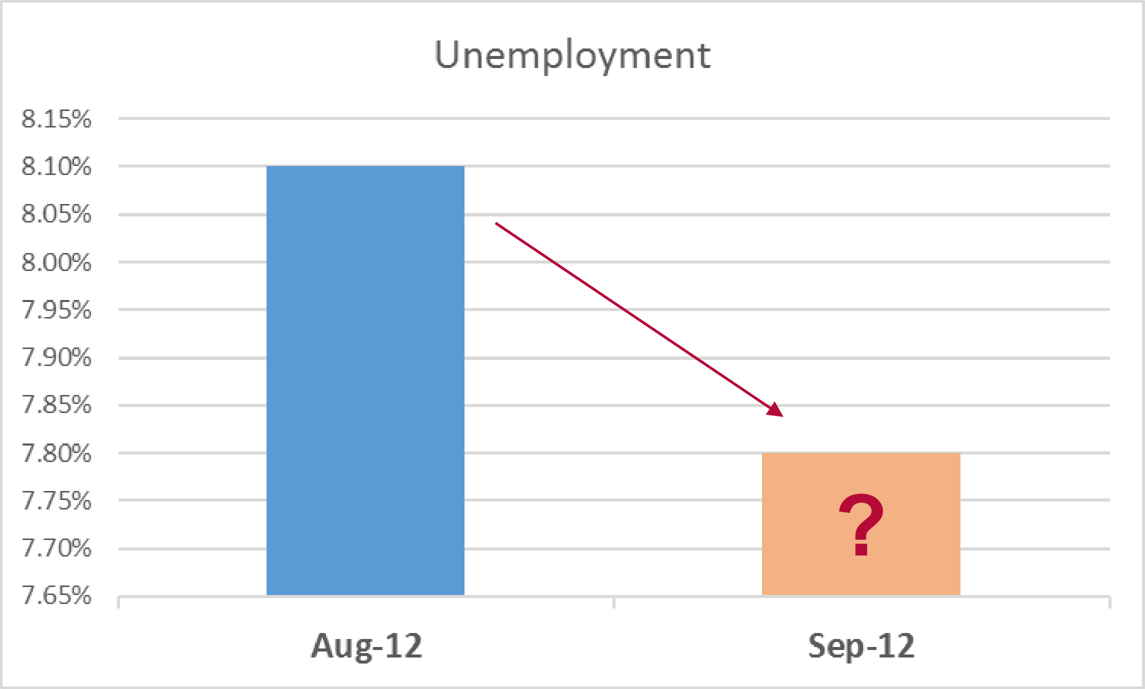
An Unusual Family Tree

Pre-Fieldwork Programming

Technique	PAPI, Basic CAPI	CAPI, ValiData	Examples
in-advance programming	X	Y	Generate new variable (e.g consumption/household size). If per capita consumption is lower than the food poverty line, then populate the questions about food security.
skip logic	X	Y	If the respondent cannot read or write, skip the section on education degree
Constraints	X	Y	Allow the respondent age to be between 18-89
warning messages	X	Y	The respondent for this survey cannot be 2 years old. Please go back and check the response
reminders / hints for each question	X	Y	Enumerator, now please switch on the GPS. Please make sure you are outside under an open sky, but very close to the interview location when capturing the GPS coordinates.

Avoid Mistakes…as they Happen

A Surprising Jobs Report



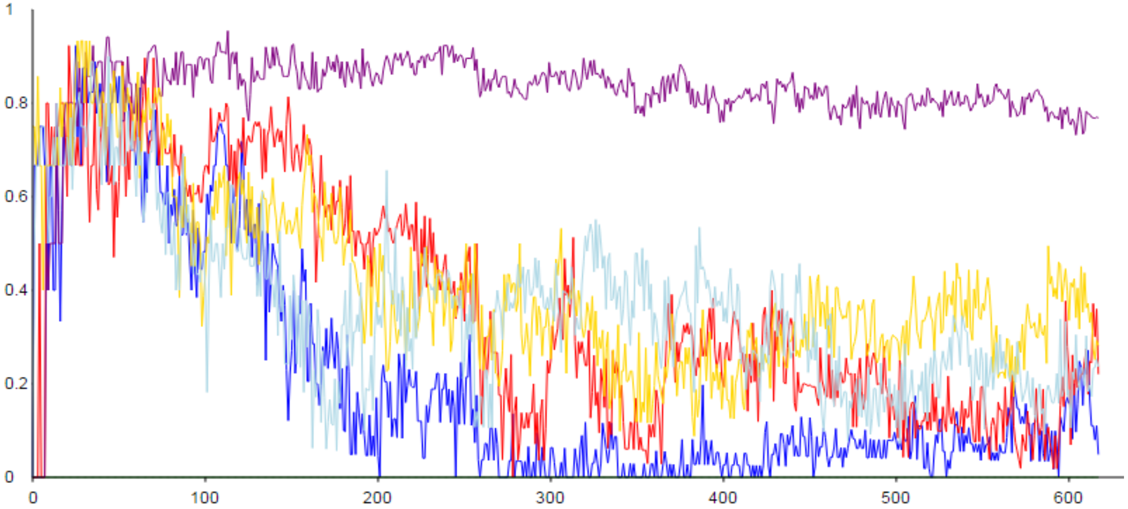
http://nypost.com/2013/11/18/census-faked-2012-election-jobs-report/

Real-Time Data Quality Assessment

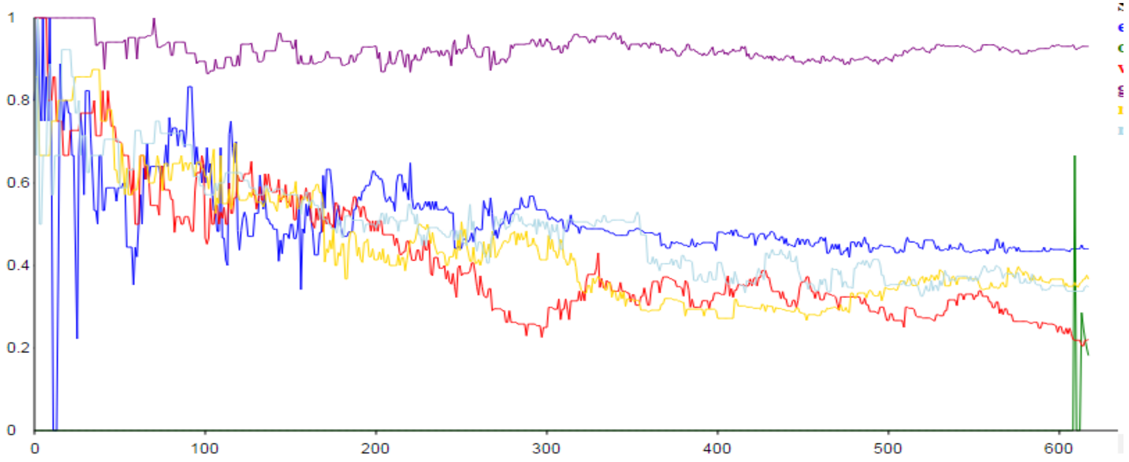
ValiData Techniques	Examples
ongoing behavioral checks using SVM and Random Forests	Monitor the behavior of each data collector constantly, check for possible data fabrications
high quality outlier detection	Robust regression models, logit/probit models,
immediate outlier correction	Reports the outliers and provide an immediate correction abilities as both the surveyor and the respondent are fresh and can verify
random audio audits	Randomly perform audio recording to check the interview quality
real time tracking — frequency, location, duration	Check the duration of each question/section/survey, track the path of each enumerator, monitor the progress per interviewer

ValiData – Detecting Falsified Surveys using SVM and RF Algorithms

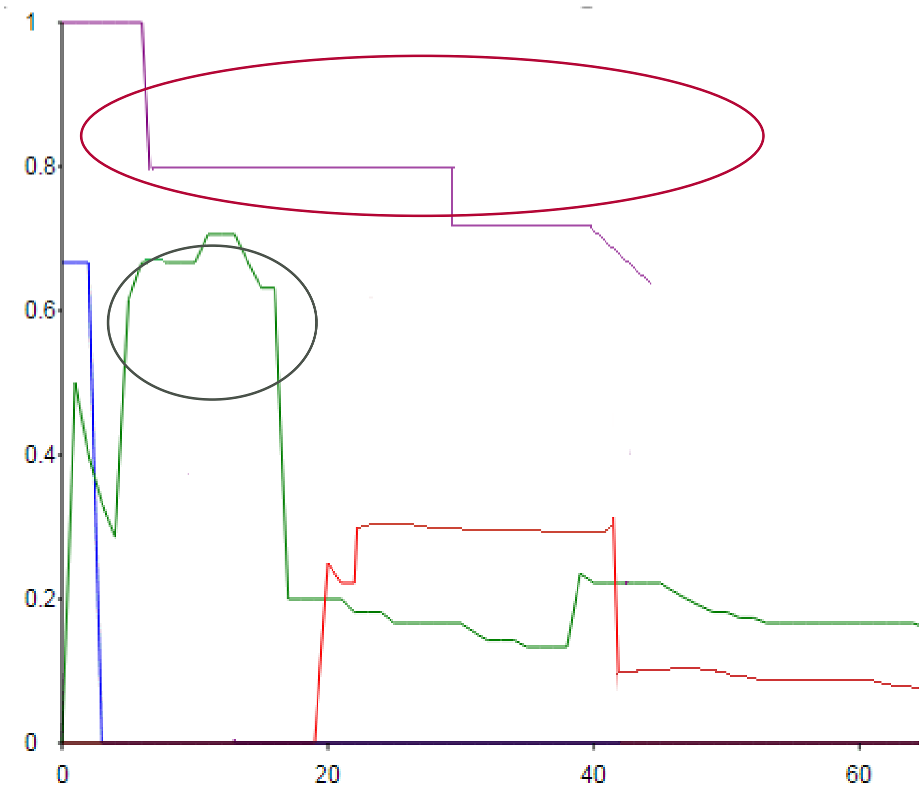
RF



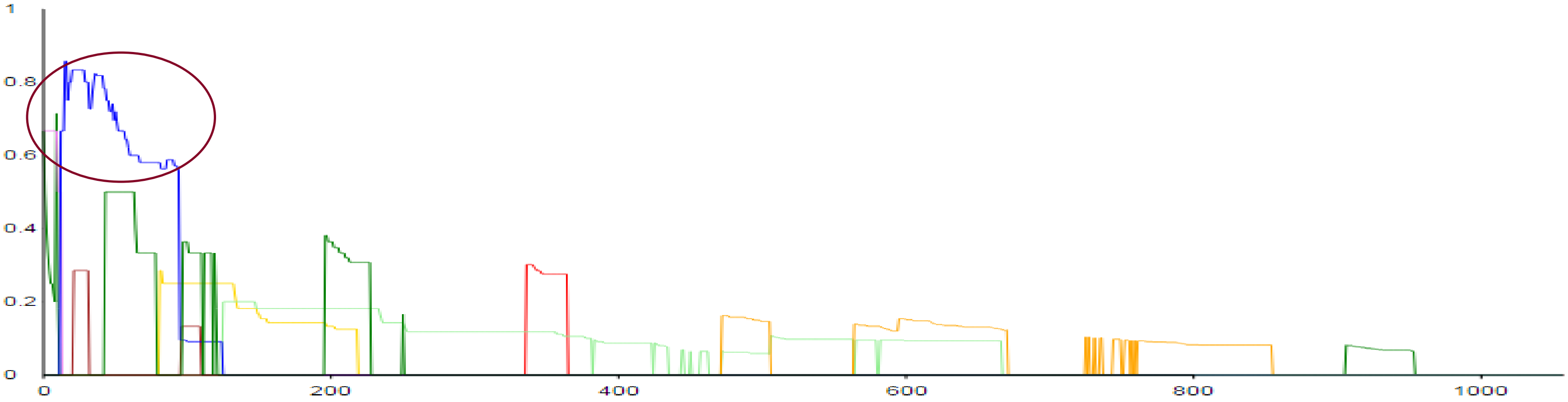
SVM



ValiData – Detecting Falsified and Problematic Surveys



ValiData – Identifying Wrong Routings using SVM and RF Algorithms.

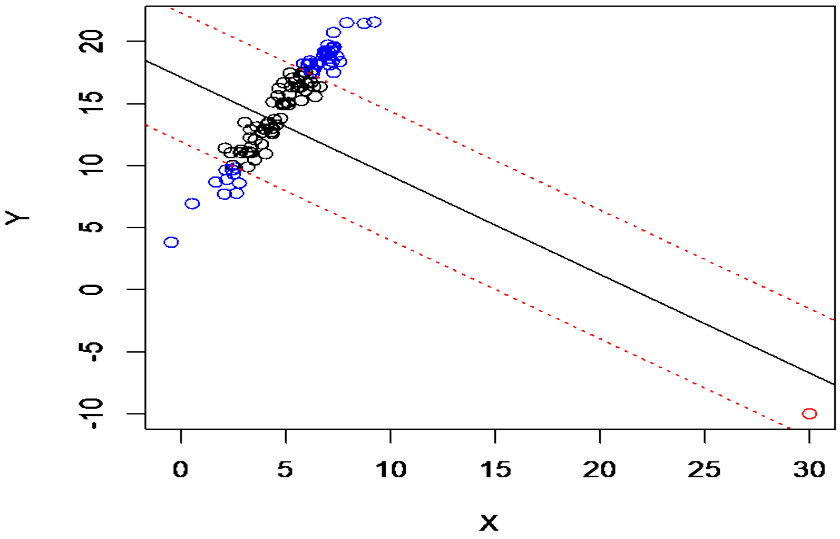


High Quality Outlier Checks

Simple Outlier Check

- Breakdown point of 0%
- Model (over) fits the data, accommodating one incorrect observation.
- Results in Type I (red) and Type II (blue) errors.

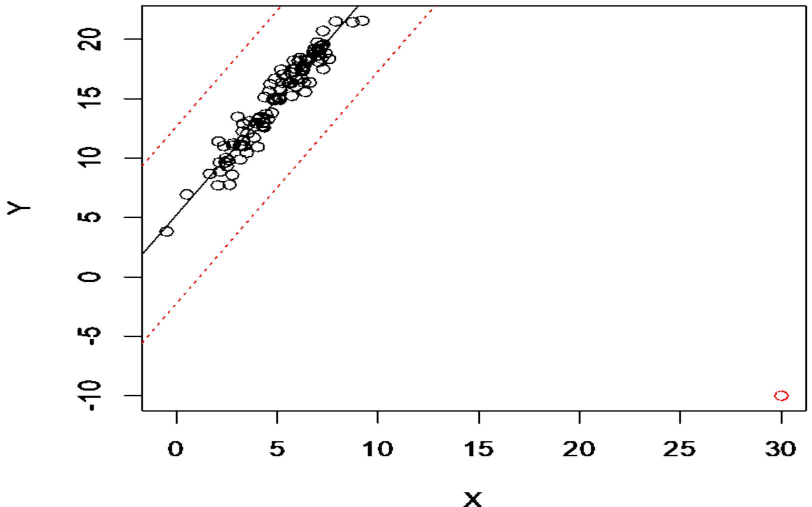
OLS regression skewed by one outlier



Robust Regression

- Inverse weighting on leverage
- Y-error weighted m-estimation
- Correctly identifies the outlier.

Robust regression using m-estimation



“Outliers” vs. Outliers

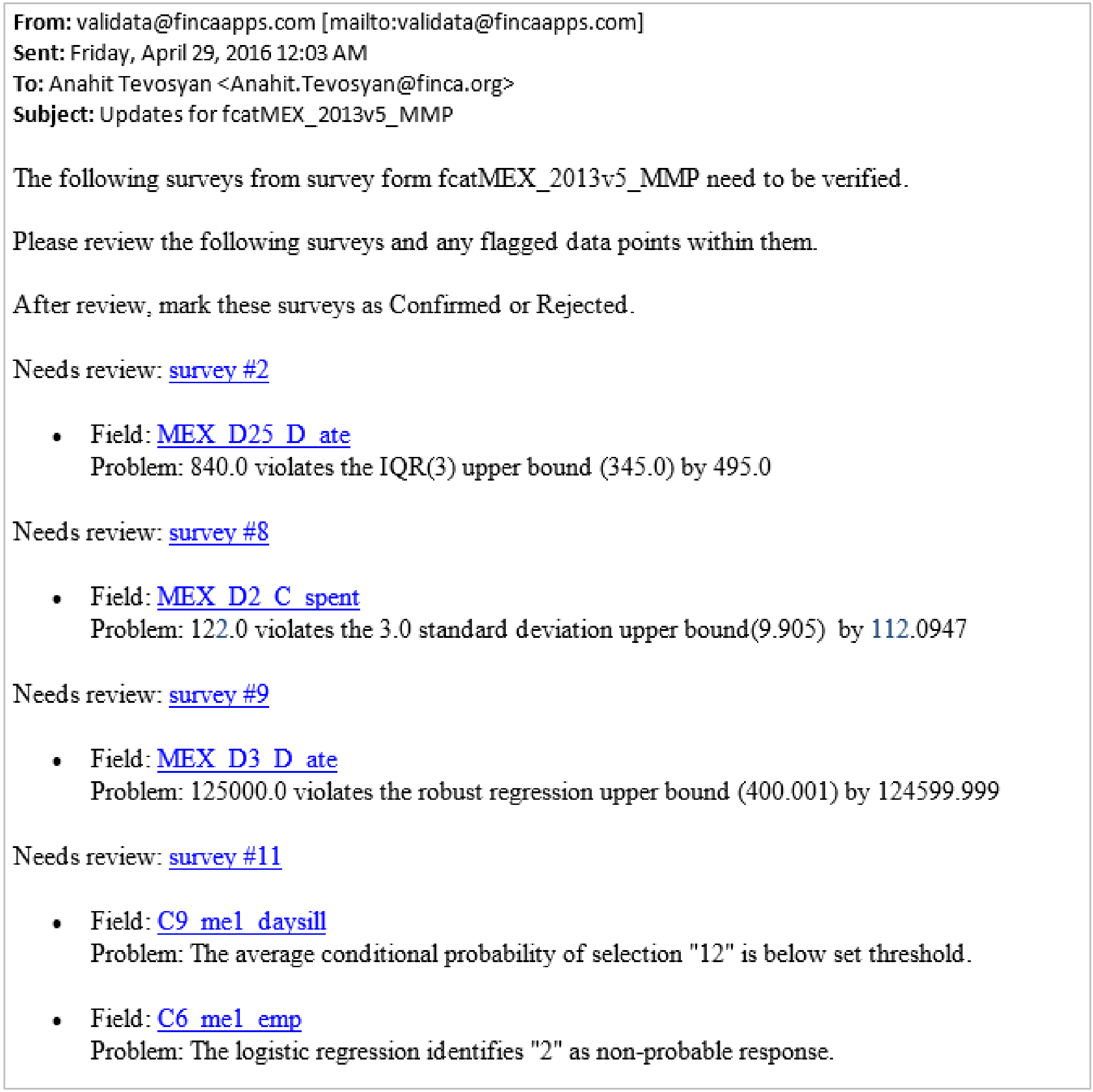
	Identified Outliers	Confirmed Errors	% correctly identified
IQR	1200	50	4%
St dev from mean/median	1050	50	5%
Linear Regression	100	10	10%
Robust Regression	50	48	96%

FINCA *Enterprise Study* in Pakistan

ValiData -- Immediate Outlier Notifications

ValiData sends scheduled emails to survey managers, alerting them to specific data points which need to be reviewed.
The email contains a link both the survey and the specific data field that has been flagged.

The notification includes links to the editing interface, showing the fields which require follow-up, and allowing for corrective action soon as the surveyor and the respondent are fresh and can recall the survey and the questions.



From: validata@fincaapps.com [mailto:validata@fincaapps.com]
Sent: Friday, April 29, 2016 12:03 AM
To: Anahit Tevosyan <Anahit.Tevosyan@finca.org>
Subject: Updates for fcatMEX_2013v5_MMP

The following surveys from survey form fcatMEX_2013v5_MMP need to be verified.

Please review the following surveys and any flagged data points within them.

After review, mark these surveys as Confirmed or Rejected.

Needs review: survey #2

- Field: MEX_D25_D_ate
  Problem: 840.0 violates the IQR(3) upper bound (345.0) by 495.0

Needs review: survey #8

- Field: MEX_D2_C_spent
  Problem: 122.0 violates the 3.0 standard deviation upper bound(9.905) by 112.0947

Needs review: survey #9

- Field: MEX_D3_D_ate
  Problem: 125000.0 violates the robust regression upper bound (400.001) by 124599.999

Needs review: survey #11

- Field: C9_me1_daysill
  Problem: The average conditional probability of selection "12" is below set threshold.

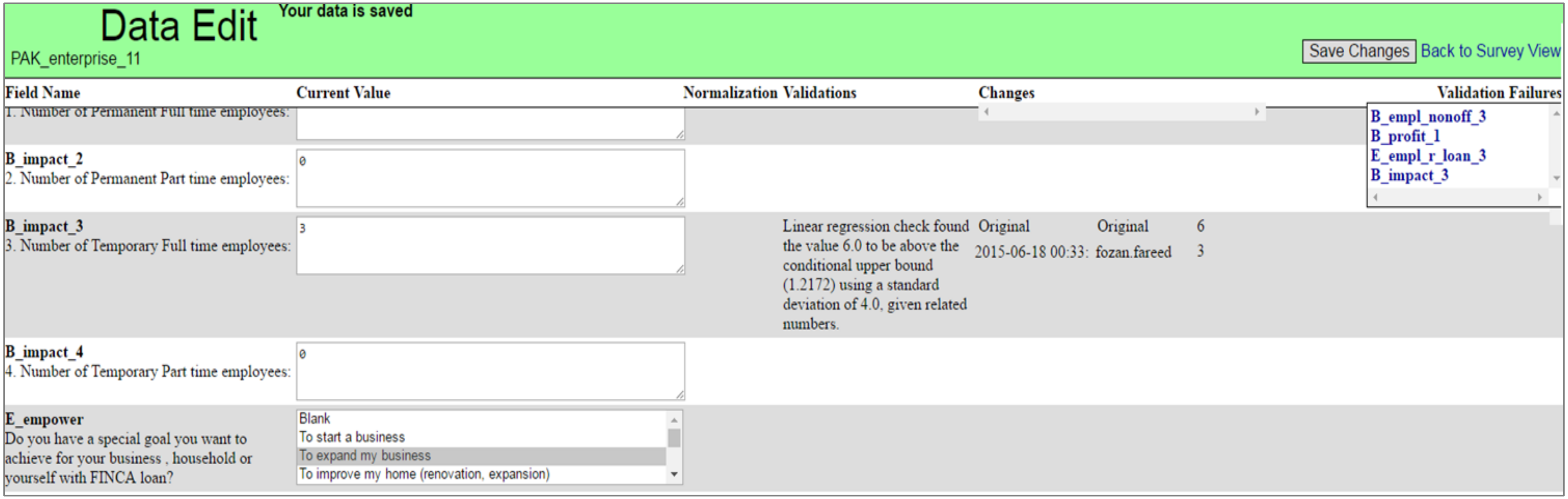
- Field: C6_me1_emp
  Problem: The logistic regression identifies "2" as non-probable response.

FINCA®

ValiData – Timely Editing of Data

ValiData enables correcting the outlying data points immediately and track the date, time and the object of correction.

The clean data downloads are available at any point during and after the interview.



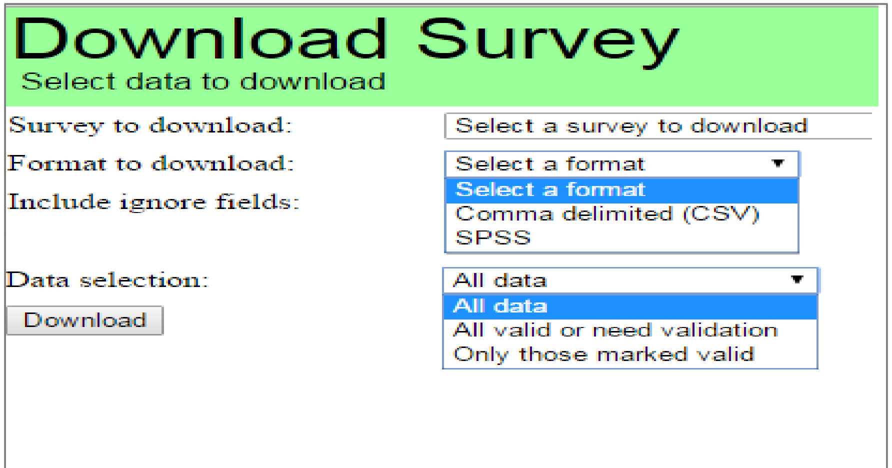
Data Edit

Your data is saved

PAK_enterprise_11

Save Changes Back to Survey View

Field Name	Current Value	Normalization Validations	Changes	Validation Failures
1. Number of Permanent Full time employees:				B_empl_nonoff_3 B_profit_1 E_empl_r_loan_3 B_impact_3
B_impact_2 2. Number of Permanent Part time employees:	0			
B_impact_3 3. Number of Temporary Full time employees:	3	Linear regression check found the value 6.0 to be above the conditional upper bound (1.2172) using a standard deviation of 4.0, given related numbers.	Original Original 6 2015-06-18 00:33: fozan.fareed 3	
B_impact_4 4. Number of Temporary Part time employees:	0			
E_empower Do you have a special goal you want to achieve for your business , household or yourself with FINCA loan?	Blank To start a business To expand my business To improve my home (renovation, expansion)			



Download Survey

Select data to download

Survey to download: Select a survey to download

Format to download: Select a format

Select a format

Comma delimited (CSV)

SPSS

Include ignore fields:

Data selection: All data

All data

All valid or need validation

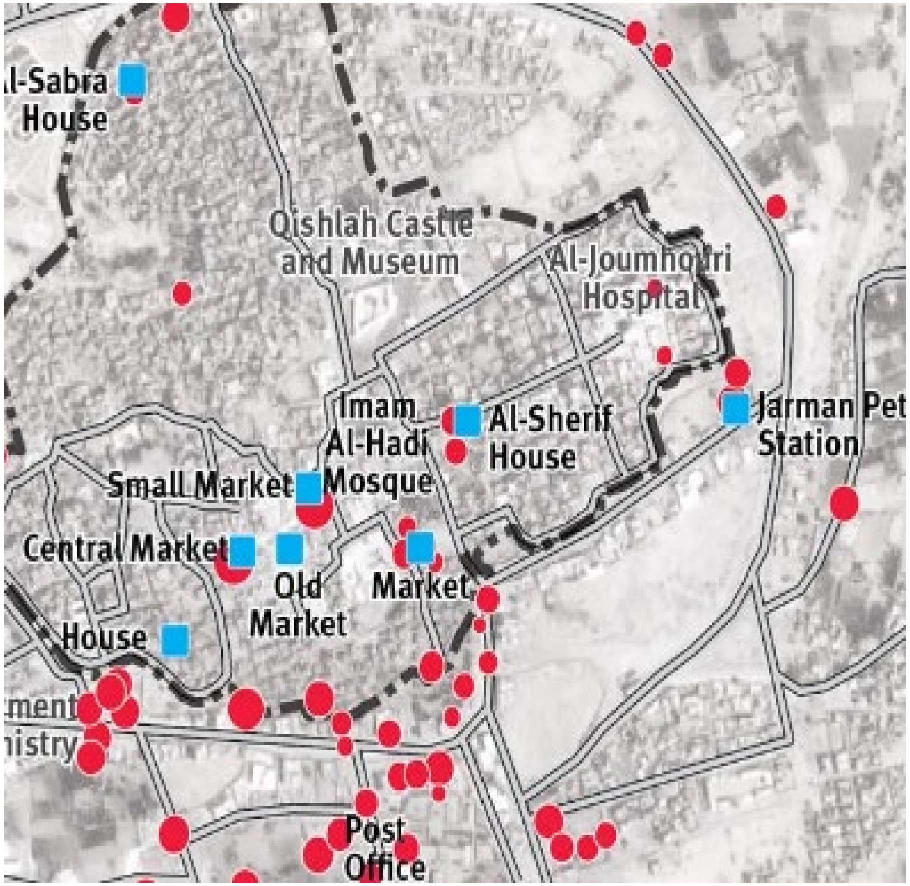
Only those marked valid

Download

Geographic Monitoring

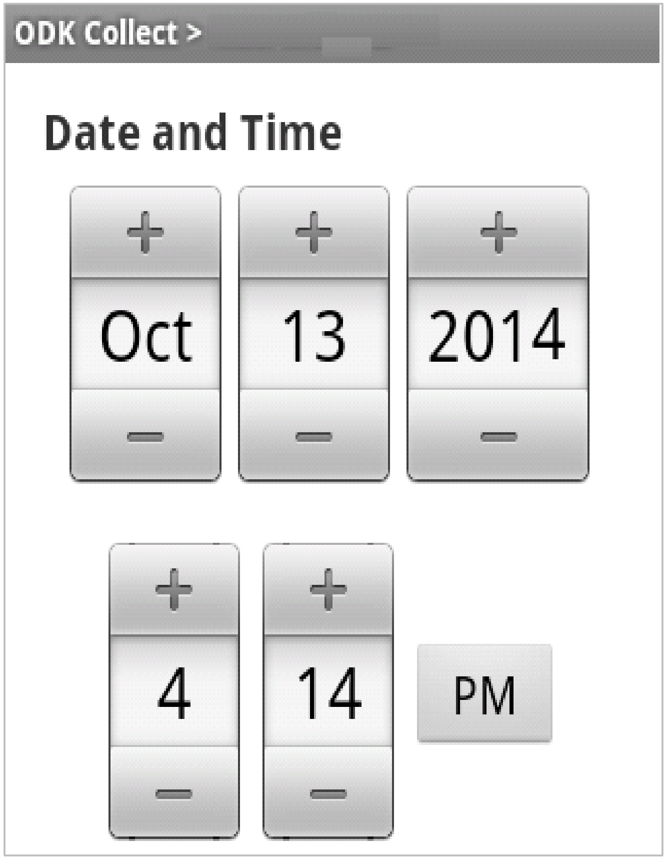
Track survey locations vs. sampling plan.

- Use GPS records to track the geo coverage of the field activities.
- Ensure alignment with the sampling plan
- Minimize the chances of data faking
- Enhance the enumerators' awareness and care for their work.



Time Tracking and other Tricks

- ValiData also uses rich set of time metadata to model and crosscheck the survey accuracy including automatically generated time stamps after each question/section/module.
- Time stamp information is then used in modeling the behavior of each enumerator as well as a key variable identifying the norm duration of each question/section/module.



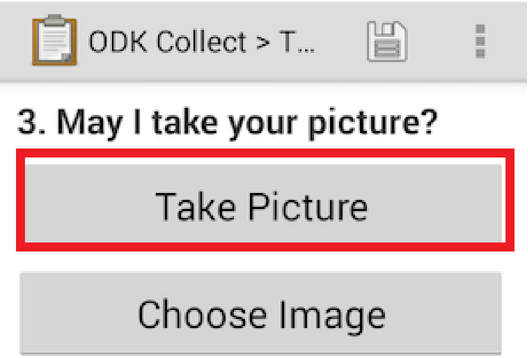
ODK Collect >

Date and Time

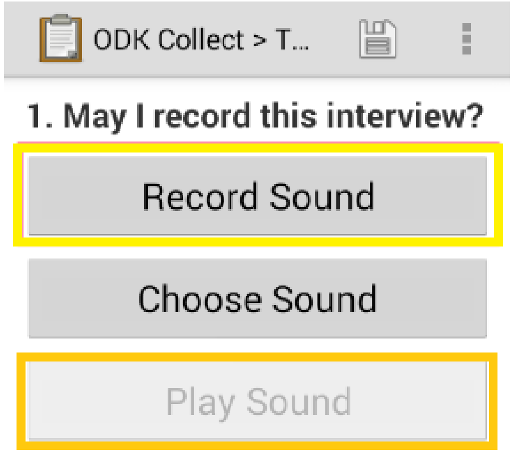
Oct 13 2014

4 14 PM

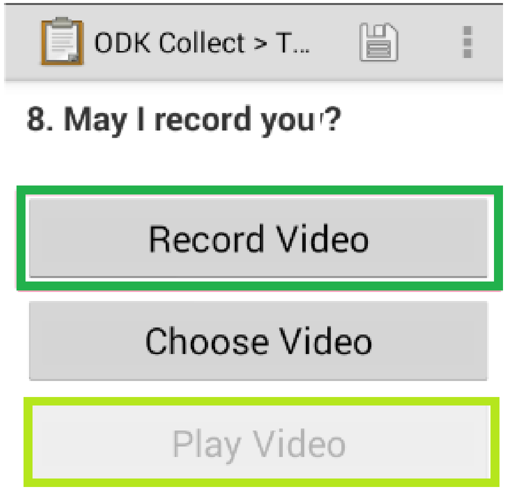
Real Time Quality Controls


ODK Collect > T...

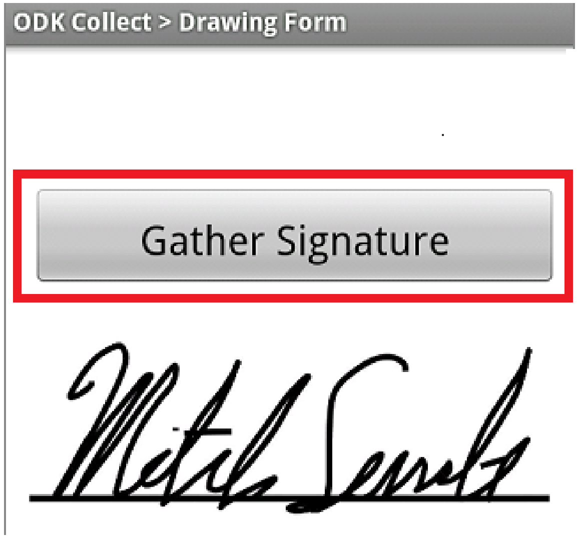
3. May I take your picture?

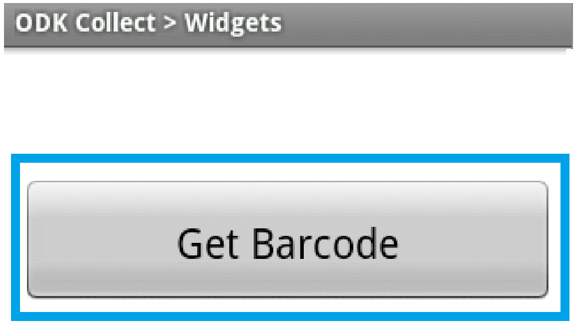

ODK Collect > T...

1. May I record this interview?


ODK Collect > T...

8. May I record you?


ODK Collect > Drawing Form

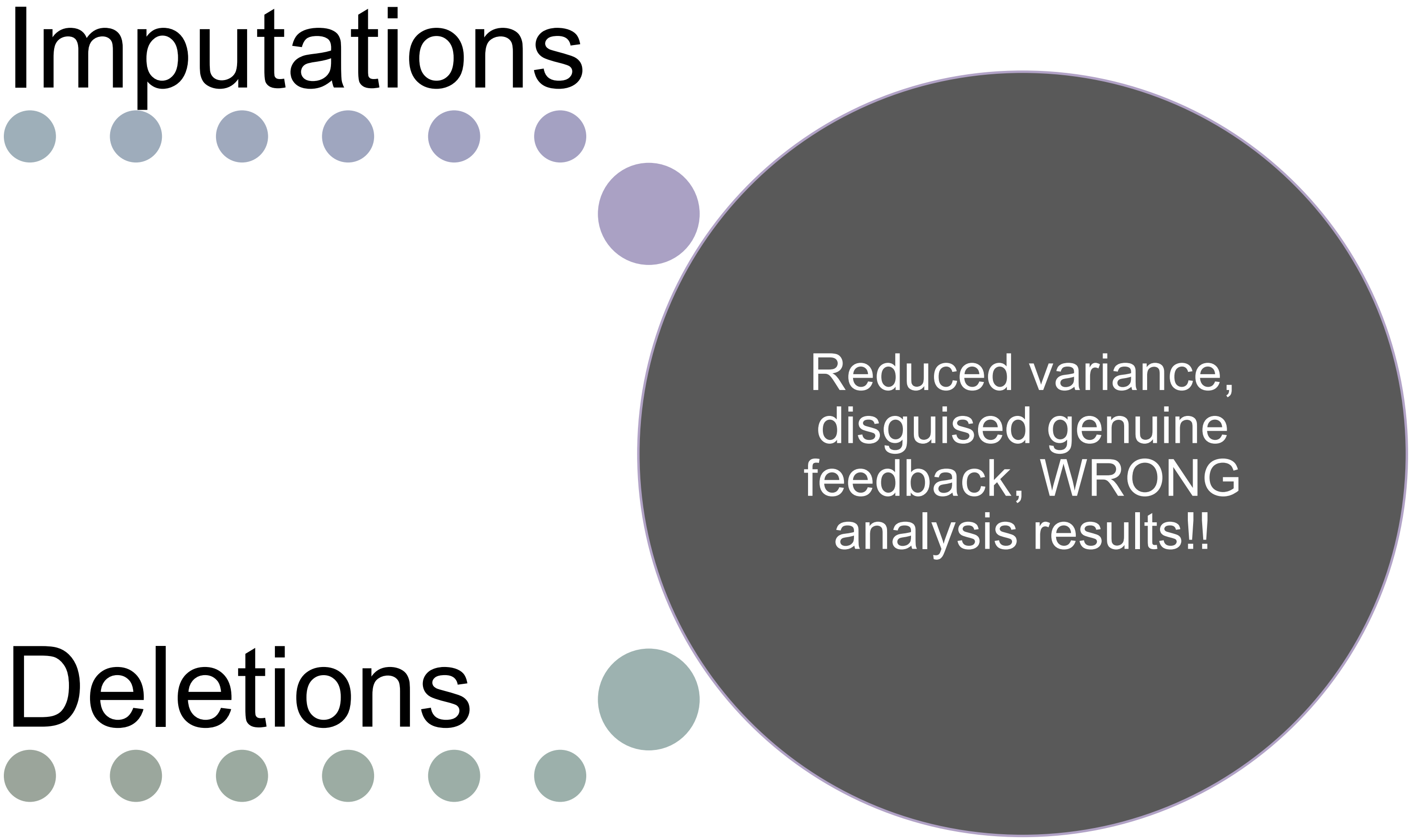

ODK Collect > Widgets

Fix Mistakes Quickly

An Uncommon Instance of Poverty

STOP Throwing Away Your Variance

The Cost Of Post Survey Outlier Cleaning



Imputations

Deletions

Reduced variance, disguised genuine feedback, WRONG analysis results!!