# Citizen Science Data Collection with Processing Bottlenecks

Jason Parham, Charles Stewart

## INTRODUCTION & ASSUMPTIONS

The following will detail the process for collecting large amounts of data from independent citizen scientists working together towards a collective goal. This procedure and client-server model will work under the following assumptions:

1. The automatic or manual processing of the collected data is **computationally intensive** or otherwise **time consuming**.
2. The participating citizen scientists expect some level of feedback from the data they personally collected — be that in the form of a document, printout, etc.
3. A small, random sampling of the data collected by the citizen scientist is sufficient to summarize their data as a whole.
4. The processing of the collected data is agnostic to input order *(optional)*.

*The rest of this poster will refer to the procedure used in the Nairobi National Park case study, detailed to the right.*

## STEP 1 - REGISTRATION

Each participating citizen scientist must be registered by being given a unique identifier and their camera used to take pictures must be time localized. In order to achieve this:

1. Each photographer is given a card with a unique, pre-specified code: car number, car color, and person letter. This card's purpose is two fold: to standardize the collected data upon return and also functions as a keepsake for the participant.
2. The current local time of registration is written on the card.
3. Using the participant's camera, **take a picture of the participant's registration card with the local time**.
4. Using the camera's display, write the file name of this newly taken image, $Image0$, on the registration card.
5. Each car is given an i-gotU GPS dongle that will be used upon return to locate all images taken by the car.

$Image0$ functions as a beginning bookend for when images taken belong to the event and also provides a **backup copy** of the registration card in the event it is lost or destroyed.

## STEP 2 - COLLECTION

The citizen scientists are sent out to collect the images. The GPS dongle makes a new record every second (assuming acquired satellite signal) consisting of **latitude, longitude and UTC timestamp**.

Every participant's camera must have a **removable memory card** so that the images can be retrieved easily and in a timely manner. As such, mobile devices are all but excluded. However, this is not much of a limitation considering that most mobile devices do not have sufficient optical capabilities for usable data.

Participants can also be given *(optional)* data collection protocols. For example, our participants were instructed to take pictures of only certain species of animals to eliminate unnecessary noise and to only take pictures in specified areas to ensure coverage.

## CASE STUDY: NAIROBI NATIONAL PARK

The RPI Computer Vision team lead by Dr. Charles Stewart, along with Dr. Tanya Berger-Wolf from UIC and Dr. Daniel Rubenstein from Princeton, administered the

### Great Zebra Count

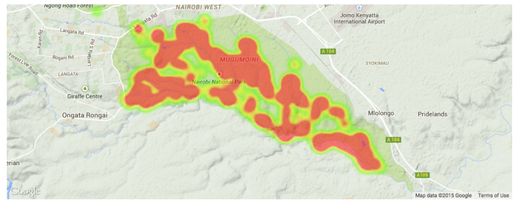
citizen science collection event during the Kenya Wildlife Festival in March 2015 in Nairobi, Kenya. The event was also orchestrated with help from Wildlife Direct in Nairobi, Kenya and support from the Kenyan Wildlife Service. **The count's purpose was to ascertain the number of zebras and giraffes in the park and if it was possible to enlist ordinary citizens in the counting effort.**

The server processing all collected images during the event and the computation of the resulting citizen science statistics was powered by IBEIS.

**See the IBEIS Poster for more details**.

| | |
|---|---|
| **9,406** | Images collected |
| **58** | Citizen scientists |
| **2** | Species (Plains Zebra, Mesai Giraffe) |
| **8,659** | Sightings of Zebra |
| **466** | Sightings of Giraffe |
| **1,258** | Identified Zebras |
| **103** | Identified Giraffes |
| **2,307 ±366** | Estimated Zebras (PL-Index) |
| **119 ± 48** | Estimated Giraffes (PL-Index) |

Image Based Ecological Information System IBEIS




Kenya Wildlife 28th FEB-7th MAR 2015 Festival

WILDLIFEDIRECT

KENYA WILDLIFE SERVICE

## STEP 3 - RETRIEVAL (CLIENT)

Once the participant returns, all images — let's say $N$ number of images — must be processed. However, this is computationally intensive or otherwise time consuming. The solution is **to only process $X$ images**, where $X << N$, and copy all $N$ images for future processing. The number of images, $X$, must therefore be processed to generate feedback.

Using the information on the registration card, the participant's camera memory card, and the car's GPS dongle, all of the needed information can be reconstructed:

1. The registration card provides the **participant's identifying information** and the **file name** of $Image0$,
2. $Image0$ provides the **set of images** to extract from the memory card,
3. The registration card provides **when** $Image0$ was taken in local time (and thus all images taken by that camera),
4. The GPS dongle provides **where** $Image0$ was taken (and thus all) by correlating with the recorded UTC times.

## STEP 4 - PROCESSING (SERVER)

The client submits the $X$ images to the server (potentially after some manual annotation) for processing. The processing algorithm can be treated as a black-box for the intents of the collection. The algorithm outputs desired results that are used to generate all feedback.

If the algorithm requires multiple pieces of information, the server caches the individual pieces until everything required is available for processing. As long as the processing is agnostic to the data input order, and because information is cached for delayed processing, the data submission to the server by the clients can be **asynchronous** and **distributed**.

The final result of the processing can be reviewed one final time and then the document or printout can be generated for the citizen scientist.

## RESULTS

The remaining $N$ images are later processed when computational processing power and/or time constraints are no longer immediate bottlenecks to a (approximate) real-time system. All images are accumulated from all distributed client machines for processing.

The results of the full processing can be made available to the citizen scientists by publicly posting it. The unique identifier on the registration card will allow participants to review their full results while also keeping their contributions anonymous to other participants.

The final full results are used to calculate ecological statistics and other conservation-oriented information, for example: population growth, species distribution, migratory patterns, hotspots for animal congregation, sex distribution, environmental impacts, etc.