Improving the Quality Of Citizen Science Data

Carl Lagoze
University of Michigan School of Information
October 9, 2012
Microsoft eScience Workshop
Acknowledgments to:

- Steve Kelling (Cornell Lab of Ornithology)
- Weng-Keen Wong (Oregon State CS)
- Theo Damoulas (Cornell CS)
- The National Science Foundation
- All the eBirders
Harnessing human intelligence to solve computational problems that are beyond the scope of existing machine intelligence:

- ESP Game – image labels
- FoldIt – protein structure
- Galaxy Zoo – galaxy classification
- reCAPTCHA – OCR of old print material
• More than ...
  • 100 million observations submitted
  • 7.5 million checklists entered
  • 85 thousand contributors
  • 9200 species
  • 220 countries
Over 7,500,000 hours in the field
Growth in eBird Observations and Checklists

eBird 2.0 launch

Observations

Checklists
Birders
Keys to eBird success among birders

• Ease of use
• Tools for improving personal skills and expertise
• Appealing to birders “benevolent competitiveness”
View and Explore Data

**Range and Point Maps**
Explore interactive range maps by species or subspecies — zoom in for details.

**Bar Charts**
Find out what birds to expect throughout the year in a region or location.

**Line Graphs**
Explore different metrics of species occurrence in a region or location.

**Your Totals**
Track your totals and compare with other eBirders.

- **Yard Totals**
  How many species and checklists have you submitted for your yard?

- **Patch Totals**
  How many have you submitted for your favorite birding patches?

- **Top 100**
  Compare with the top eBirders in your region.

**Arrivals and Departures**
Arrivals and departures for a country, state/province, county, or hotspot

**All-Time First/Last Records**
All-time records for species arrival and departure in a region

**High Counts**
Species high counts for a region

**Alerts**
Reports and email alerts for rarities and species you haven't seen

**Summary Tables**
Observations summarized by week, month, or year

©2011 Audubon and Cornell Lab of Ornithology | Affiliates and Sponsors | Privacy Policy | Contact | FAQ
# Bird Observations

**For** [Washtenaw]

281 species (+45 other taxa)

<table>
<thead>
<tr>
<th>Species</th>
<th>MAP</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater White-fronted Goose</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Snow Goose</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ross's Goose</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Snow X Ross's Goose (hybrid)</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Snow/Ross's Goose</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cackling Goose</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canada Goose</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cackling/Canada Goose</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goose sp.</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mute Swan</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trumpeter Swan</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tundra Swan</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trumpeter/Tundra Swan</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mute Swan sp.</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wood Duck</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gadwall</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Wigeon</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Black Duck</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mallard</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mallard (Domestic type)</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Black Duck x Mallard (hybrid)</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blue-winged Teal</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northern Shoveler</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northern Pintail</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Green-winged Teal</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canvasback</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Redhead</td>
<td>MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Where can I see a Golden-winged Warbler?
Year-Round, All Years, Full Range
Kenn Kaufman’s 6 May 2011 list

<table>
<thead>
<tr>
<th>Species</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ovenbird</td>
<td>1</td>
</tr>
<tr>
<td>Northern Waterthrush</td>
<td>4</td>
</tr>
<tr>
<td>Blue-winged Warbler</td>
<td>1</td>
</tr>
<tr>
<td>Golden-winged Warbler</td>
<td>1</td>
</tr>
<tr>
<td>Black-and-white Warbler</td>
<td>15</td>
</tr>
<tr>
<td>Tennessee Warbler</td>
<td>3</td>
</tr>
<tr>
<td>Nashville Warbler</td>
<td>15</td>
</tr>
<tr>
<td>Common Yellowthroat</td>
<td>3</td>
</tr>
<tr>
<td>American Redstart</td>
<td>10</td>
</tr>
<tr>
<td>Cape May Warbler</td>
<td>5</td>
</tr>
<tr>
<td>Northern Parula</td>
<td>20</td>
</tr>
<tr>
<td>Magnolia Warbler</td>
<td>4</td>
</tr>
<tr>
<td>Bay-breasted Warbler</td>
<td>1</td>
</tr>
<tr>
<td>Blackburnian Warbler</td>
<td>6</td>
</tr>
<tr>
<td>Yellow Warbler</td>
<td>24</td>
</tr>
<tr>
<td>Chestnut-sided Warbler</td>
<td>8</td>
</tr>
<tr>
<td>Black-throated Blue Warbler</td>
<td>7</td>
</tr>
<tr>
<td>Palm Warbler</td>
<td>22</td>
</tr>
<tr>
<td>Canada Goose</td>
<td>15</td>
</tr>
<tr>
<td>Wood Duck</td>
<td>6</td>
</tr>
<tr>
<td>Mallard</td>
<td>1</td>
</tr>
<tr>
<td>Double-crested Cormorant</td>
<td>8</td>
</tr>
<tr>
<td>Great Blue Heron</td>
<td>5</td>
</tr>
<tr>
<td>Great Egret</td>
<td>2</td>
</tr>
<tr>
<td>Osprey</td>
<td>1</td>
</tr>
<tr>
<td>Bald Eagle</td>
<td>1</td>
</tr>
<tr>
<td>Sora</td>
<td>1</td>
</tr>
<tr>
<td>American Coot</td>
<td>5</td>
</tr>
<tr>
<td>Kildeer</td>
<td>1</td>
</tr>
</tbody>
</table>
Science
## Observational Data Model

### Location
- LOC_ID (PK)
- NAME
- DESCRIPTION
- LATITUDE
- LONGITUDE
- OWNER_USER_ID
- SUBNATIONAL1_CODE (FK)
- SUBNATIONAL2_CODE (FK)
- IS_BIRDING_HOTSPOT
- GEOM
- CREATION_DT
- LAST_EDITED_DT

### Collection Event
- SUB_ID (PK)
- PROJ_ID (FK)
- PROTOCOL_ID (FK)
- LOC_ID (FK)
- USER_ID (FK)
- OBS_DT
- OBS_TIME_VALID
- ALL_OBS_REPORTED
- COMMENTS
- EFFORT_HRS
- EFFORT_PT_RADIUS_M
- EFFORT_DISTANCE_KM
- EFFORT_AREA_HA
- FILTER_ID (FK)
- GROUP_ID
- CREATION_DT
- LAST_EDITED_DT

### Species Observation
- OBS_ID (PK)
- SPECIES_CODE (FK)
- PROJ_ID (FK)
- SUB_ID (FK)
- HOW_MANY
- COMMENTS
- VALID
- REVIEWED
- CREATION_DT
- LAST_EDITED_DT

### Observer
- USER_ID (PK)
- EMAIL
- LAST_NAME
- FIRST_NAME
- ADDRESS...
- LAST_LOGIN_DT
Conservation
The State of the Birds 2011
Report on Public Lands and Waters
United States of America
U.S. Public Lands

- Protected Areas Database for the U.S. (PAD-US 1.1; USGS-GAP)
- Includes Federal and State agencies
- Biodiversity protection status
- Bird distributions overlaid on public land ownership map

Maps produced by National Gap Analysis Program
ARIDLANDS and Bureau of Land Management

• Public lands support more than half the distribution of aridland birds (36 obligate species)
• Declining aridland birds affected by grazing, invasive species, energy development

79% of Gunnison Sage-Grouse distribution on public lands
BLM manages 245 million acres, including more aridland habitat than any other agency
Stewardship Responsibilities for Birds

Mountain Bluebird

Willow Flycatcher

Agency Breadown
Mapping Migratory Bird Stopover Habitat to Inform Wind Energy Siting

Ron Rohrbaugh, Frank La Sorte, Daniel Fink, Andrew Farnsworth, and Steve Kelling
Observer/Data Quality

About eBird

eBird Data Quality

Did you know that every record submitted to eBird goes through the eBird data verification process? Using a combination of automated data filters and a network of local experts, eBird tackles the issue of data quality in Citizen-Science. In order for us to maintain the integrity of the database, and for it to be used fully by the science and conservation community, we as observers must fully understand and strive to reach the highest level of data quality. Therefore, we’ve developed procedures to facilitate communication between eBird observers and scientists, including some new and improved review tools for our editors. Through our combined effort to maintain high data quality, eBird will take its place among the most valuable large-scale data sets on bird distribution and abundance in the world. Read more about our data verification process....

What do we do when you report birds as rare or far out of range as this Siberian Taiga Flycatcher at Putah Creek, California, found by eBider John Sterling, or the Ivory Gull found near New York City? Even more complex is the issue of how to deal with records of early/late migrants, or out of season records of lingering birds at local scales. Issues surrounding how to verify data like these certainly come into play when maintaining a database of records that is meant to become part of the scientific record. As with any large-scale citizen-science project open to the public, there is the possibility that erroneous data will be submitted. At eBird we consider data quality to be paramount, and we’re taking every step possible to ensure that our data are the best they can be. Using advanced data vetting technology, we’ve developed a combination of automated filters and a network of regional editors that work together to verify eBird data. Each eBird submission, regardless of observer or location, is checked for data quality in exactly the same way.

Why worry about data quality?

A database is only as good as its weakest record. If even a few records can be deemed questionable, then the entire data set can be labeled as such. With that in mind, we should all strive to keep the eBird data as clean as possible. You can do your part by being conservative in the field and meticulous with your data entry, and we can do ours by building better connections between the eBird community and scientists.
Enhance the effectiveness of citizen science

- build accurate data input filters to limit erroneous data submission
- identify variation in observer ability
- address the spatial bias of where observers make their observations
Human/Computer Learning Network
Active Learning

Bird Watcher

Location
Optimization +
Crowdsourcing
Incentives

Emergent
Filter

ODE
Model

Checklist

Location
Date+Effort
Bird IDs
Counts

Validation + Feedback

eBird Database

Experienced
Accepted

Novice

Flagged for Review

Artificial Intelligence

Active Learning diagram.
Emergent filters to limit erroneous observations

• Replace unscalable system of volunteer regional experts
• Apply frequency of occurrence filters, generated from eBird data, to delineate when a species can be reported in a region
• Automatically flag unusual observations
Addressing observer variability by modeling observer ability

- Can’t establish “ground truth” through multiple observers
- Occupancy-Detection-Experience Model (ODE)
  - probability of the occupancy of the site by a particular species given a set of environmental covariates describing that site
  - Factor in “ease of detection” covariates
  - Quantify individual observer expertise variation based on covariates such as number of checklists submitted, number of records flagged by experts, number of species reported, etc.
  - Improved expertise by suggesting species they “should” see
Improving spatial coverage in citizen science

Improve the predictive performance of the machine learning algorithms by guiding the sampling process.
Measuring HCLN impact on citizen science participants

- Does the system and expertise ranking provide incentives for greater participation?
- Does the system encourage greater geographic distribution of observations?
- Is the system “educating” the observers?
- What is the nature of the detection curve for various types of observers and how does it change over time?
Questions?