eBird: A Human/Computer Learning Network for Biodiversity Conservation and Research

Steve Kelling, Jeff Gerbracht, and Daniel Fink
Cornell Lab of Ornithology, Cornell University
stk2@cornell.edu, jag73@cornell.edu, df17@cornell.edu

Carl Lagoze
Information Science, Cornell University
lagoze@cs.cornell.edu

Weng-Keen Wong and Jun Yu
School of EECS, Oregon State University
wong@eecs.oregonstate.edu, juyu@eecs.oregonstate.edu

Theodoros Damoulas and Carla Gomes
Department of Computer Science, Cornell University
Damoulas@cs.cornell.edu, Gomes@cs.cornell.edu

Abstract

In this paper we describe eBird, a citizen-science project that takes advantage of human observational capacity and machine learning methods to explore the synergies between human computation and mechanical computation. We call this model a Human/Computer Learning Network, whose core is an active learning feedback loop between humans and machines that dramatically improves the quality of both, and thereby continually improves the effectiveness of the network as a whole. Human/Computer Learning Networks leverage the contributions of a broad recruitment of human observers and processes their contributed data with Artificial Intelligence algorithms leading to a computational power that far exceeds the sum of the individual parts.

Introduction

The transformational power of todays computing, together with information and communication technologies, are providing new opportunities to engage the public to participate in and contribute to a myriad of scientific, business and technical challenges. For example, citizen-science projects such as Galaxy Zoo, eBird, and Foldit demonstrate the power of crowdsourcing for investigating large-scale scientific problems. These and similar projects leverage emerging techniques that integrate the speed and scalability of *mechanical computation*, using advances in Artificial Intelligence (AI), with the real intelligence of *human computation* to solve computational problems that are beyond the scope of existing algorithms [1].

Human computational systems use the innate abilities of humans to solve certain problems that computers cannot solve [2]. Now the World Wide Web provides the opportunity to engage large numbers of humans to solve these problems. For example, engagement can be game-based such as FoldIt, which attempts to predict the structure of a protein by taking advantage of humans' puzzle solving abilities [3]; or Galaxy Zoo, which has engaged more than 200,000 participants to classify more than 100 million galaxies [4]. Alternatively, the Web can be used to engage large numbers of participants to actively collect data and submit it to central data repositories. Projects such as eBird, engage a global network of volunteers to report bird observations that are used to generate extremely accurate estimates of species distributions [5].

Now systems are being developed that employ both human and mechanical computation to solve complex problems through active learning and feedback. These Human/Computer Learning Networks (HCLN) can leverage the contributions of broad recruitment of human observers and process their contributed data with AI algorithms for a resulting total computational power far exceeding the sum of their individual parts. This combination can be deployed in a variety of domains and holds enormous potential to solve complex computational problems.

A key factor in the power of an HCLN is the manner in which the benefits of active learning are cyclically fed back among the human participants and computational systems. We use "active learning" in both of its commonly used senses: the *machine learning* sense as a form of iterative supervised learning, and the *human* sense in which learners (our volunteers) are actively and dynamically guided to new levels of expertise. The role of active learning in a HCLN is illustrated in figure 1. In our example, broad networks of volunteers act as intelligent and trainable sensors to gather observations. AI processes dramatically improve the quality of the observational data that volunteers provide by filtering inputs based on aggregated historical data and observers' expertise. By guiding observers with immediate feedback on observation accuracy

AI processes contribute to advancing observer expertise. Simultaneously, as observer data quality improves, the training data on which the AI processes make their decisions also improves. This feedback loop increases the accuracy of the analysis, which enhances the general utility of the data for scientific purposes.

cur at that specific location and date, and constantly recalibrate until the species is correctly identified.

eBird (http://ebird.org) [5] is a citizen science project that engages a global network of bird watchers to identify birds to species and report their observations to a centralized database. Anyone can submit their observations of

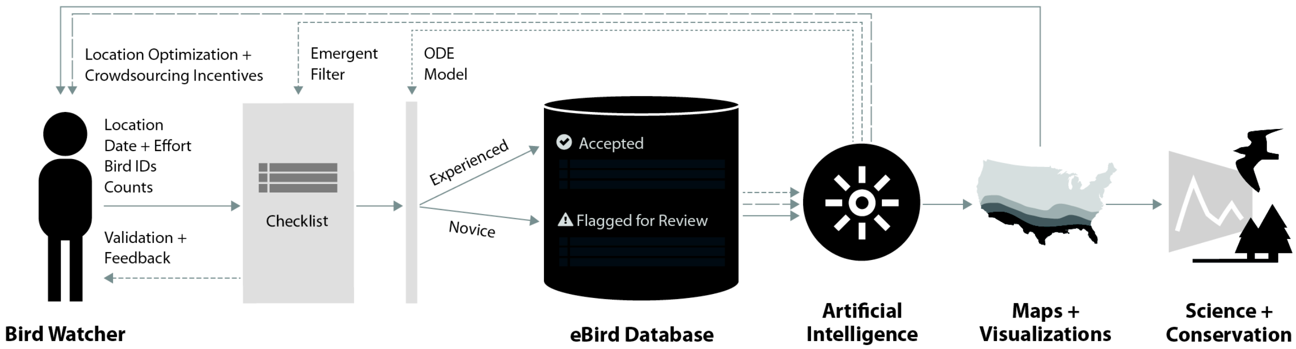

Figure 1. An HCLN example. Human observers and AI processes synergistically improve the overall quality of the entire system. Additionally, AI is used to generate analyses that improve as the quality of the incoming data improves.

A successful HCLN must be able to address the 4 following challenges. First, a task must be identified that human computational systems can complete but mechanical computational systems cannot [1]. Second, the task must be sufficiently straightforward and incentivized to maximize participation [6]. Third, the complimentary abilities of both humans and machines must be clearly identified so that they can be leveraged to increase the accuracy and efficiency of the network [7]. Finally novel methods for extracting biological insights from the noisy and complex outputs provided by multiple human computers must be employed [8]. In this paper we use our experience with eBird as a model to describe a functional HCLN, by explaining how we addressed these 4 primary HCLN challenges.

Challenge 1: Species Identification

Few mechanical computational systems have been developed to classify organisms to the species level. Those that do exist typically can only identify a single or small group of species, and cannot classify a multitude of organisms. Only human observers can reliably identify organisms to the species level [9], and are capable of classifying hundreds of species. This is because identifying a species is a complex task that relies on a combination of factors. First, observers must be able to process impressions of shape, size, and behavior under variable observation conditions. As this process continues, the observer must combine these impressions with a mental list of species most likely to oc-

birds to eBird via the web, and more than 83,000 individuals have volunteered over 4 million hours to collect more than 75 million bird observations; arguably the largest biodiversity data collection project in existence.

The amassed observations from eBird provide researchers, scientists, students, educators, and amateur naturalists with data about bird distribution and abundance across varying spatio-temporal extents. Dynamic and interactive maps, graphs and other visualizations are available on the eBird website, and all data are free and readily accessible through the Avian Knowledge Network [10]. Since 2006 eBird data have been the basis for 56 peer-reviewed publications and reports, from highlighting the importance of public lands in conservation [11], to studies of evolution [12], climate change [13] and biogeography [14].

Challenge 2: Maximizing Participation

eBird is a crowdsourcing activity that engages large numbers of people to perform tasks that automated sensors and computers cannot readily accomplish [15]. This is accomplished through the development of straightforward rules for participation and incentives for contributing. eBird gathers data using protocols that closely match the activities of individuals when they are birding. This maximizes the number of participants in eBird [6]. While eBird requires that participants submit sufficient effort data (see below) to allow the quantitative analysis of the observations, sufficient incentives are provided to reward participation. For example, eBird participants can: (i) keep track of their bird records; (ii) sort their personal bird lists by date and region; (iii) share their lists with others; and (iv) visualize their observations on maps and graphs. By providing these record-keeping facilities as a direct reward for participation eBird appeals to the competitiveness of participants by providing tools for determining relative status of volunteers (e.g. numbers of species seen) and geographical regions (e.g. checklists submitted per state and province). This appeal to competitiveness has been successful in many crowdsourcing projects and citizen-science projects.

A key component of eBird's success has been the implementation of a sound data management strategy, which reduces the risk of data loss and allows for efficient use and re-use of the data. All eBird data contain the following information: observer identification, location, visit, and what was collected. These data form the core observational data model [16] and provide the opportunity for integration, visualization, experimentation and analysis. For example, eBird collects the name and contact information for every observer, which allows each observation to be attributed to a specific person. Location data such as the site name the coordinates where the observations were made and the geographic area represented by the location are stored with every visit to that location. Information about a specific visit consists of data and time of visit, amount of effort expended, such as distance traveled, time spent and area covered, and whether or not all species observed were reported. Species observations consist of a checklist of birds observed and how many individuals of each species were counted. These data fields form the core of the eBird relational database, and the foundation for which all eBird functionality is developed.

Challenge 3: Identifying the Synergies Between Humans and Machines

While eBird is extremely successful in engaging a large community of volunteers to participate, there are many challenges to using eBird data for analysis. A major goal has been to employ HCLN processes to eBird to improve data quality by addressing 3 major questions:

How can we efficiently filter erroneous data before data enter the database?

eBird has motivated thousands of volunteers to collect large amounts of data at relatively little cost. However, the public's ability to identify or classify objects without making errors is highly variable. Misidentification of birds is the major data quality concern of eBird. To address this issue a network of more than 450 volunteers review records in eBird. The reviewers are knowledgeable about bird occurrence for a region, and contact those individuals who submitted questionable (i.e., unusual reports of birds not known to occur in a region) records to obtain additional information, such as field notes or photographs, in order to confirm unusual records. However, our challenge is that eBird's success has generated an enormous volume of observations to be reviewed (e.g., more than 23 million observations were gathered in 2011). This volume is overwhelming the network of volunteer regional reviewer. In order to address this issue we have implemented a data quality filter and screening process that automates the review process, which we now describe.

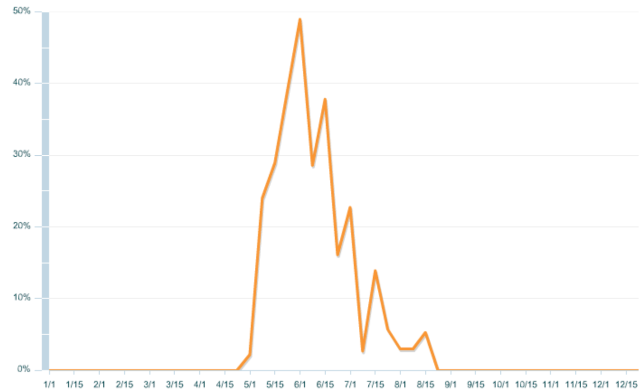

Figure 2. Frequency of occurrence results for Black-billed Cuckoos in upstate New York. The Y-axis is the frequency of eBird checklists that reported this species, and the X-axis is the date. Cuckoos arrive in early May and are detected at high frequencies because they are conspicuous and vocal during their mating season. But after they lay eggs, their detection probability drops dramatically. Most birds leave by mid-August.

One of the most powerful calculations performed on citizen-science data is the frequency of reporting a particular event or organism (Figure 2). Since each observation contains details of where and when a bird was detected, we can estimate the "likelihood" of observing a specific species at any spatial level (e.g., grid, country, state, county, or backyard) and for any date. Frequency of occurrence filters delineates when a species can be reported in a region and determines the validity of an observation.

The eBird database currently holds more than 75 million bird observations. These historical records can be used to filter unusual observations that require review, but allow entry of expected species within the expected times when species should occur. These filters automatically *emerge* from historic eBird data. Through experimentation we have set the emergent filter at 10% of maximum annual frequency of occurrence for every species. This provides a consistent limit that allows expected observations through the filter but flags for review unusual records. For example, if a common species reaches a maximum frequency of

68% then the filter would identify the date at which the filter first crosses the 6.8% threshold. Any record submitted on a date either prior or after the threshold limit, it is flagged for review. Similarly, if a rare species reaches an annual peak of 6.5% frequency, the threshold limit would be .65%. For example, we analyzed eBird data and emergent filter results for 2 counties in New York State, Jefferson Co. and Tompkins Co (Table 1). These 2 counties were selected because Jefferson Co. has relatively sparse year-round data coverage, while Tompkins Co. is one of the most active regions in eBird. Currently, emergent filters are deployed for all counties in the United States.

The automated emergent filter process significantly reduces the number of records the volunteer observer network had to review. When the emergent filter is triggered the submitter gets immediate feedback indicating that this was an unusual observation (Figure 1). If they confirm they made the observation, their record is flagged for review, and one of the volunteer experts will review the observation. All records, their flags and their review history are retained in the eBird database.

The emergent filter process identifies key periods during a bird's phenology, when their patterns of occurrence change. Figure 3 shows those records that are flagged for review by the emergent filter for the 2 New York Counties. The Chipping Sparrow is a common breeding bird in upstate New York, but rarely occurs in winter. The emergent filter for each county is different, due to the variation in each county's respective historic data. The triangles and circles are all records that are flagged for review by the emergent filter. Without the emergent filter it would be difficult to accurately identify arrival and departure dates of when a bird appears in a county. The threshold of occurrence established by the emergent filter allows the determination of arrival and departure and then accurately flags outlier observation for further processing and review.

	Tompkins Co.	Jefferson Co.
Total Observations	704,053	78,745
Total Flagged	50,743	6,082
Percent Flagged	7	8
Total Flagged Expert	38,574	3,787
Total Flagged Novice	12,170	2,295
Percent Expert	5	5
Percent Novice	2	3

Table 1. Results of the Emergent Filter process applied to 2 counties in Upstate New York (upper), and the proportion of flagged records submitted by experts and novices (lower).

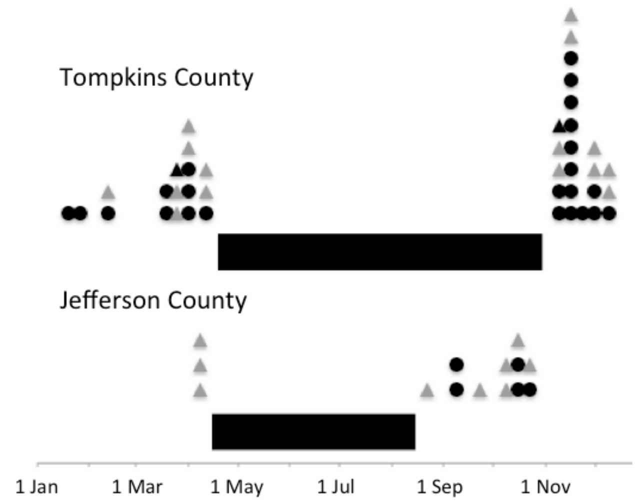

Figure 3. The acceptable date range (dark bars) for the occurrence of Chipping Sparrow in 2 counties in New York. All records that fall outside of the acceptable date range are plotted either as circles (novices) or triangles (experts).

Can we identify observer variability in their ability to detect objects?

eBird data are contributed by observers with a wide range of expertise in identifying birds. At one extreme observers with high identification skill levels contribute "professional grade" observations to eBird, whereas at the other extreme less-skilled participants contribute data of more variable quality. This inter-observer variation must be taken into account during analysis to determine if outlier observations (i.e., those observations that are unusual) are true occurrences of a rare species, or the misidentification of a common species. Since eBird engages a significant number of skilled observers who are motivated to detect rare species or are skilled in detecting elusive and cryptic species, being able to accurately distinguish their observations from those of less-skilled observers is crucial. This is because skilled observers are more likely to submit observations of unusual species that get flagged by the regional emergent filters (i.e., skilled birders like to find rare birds). An objective measure of observer expertise that could classify unusual observations is required.

To better understand observer variability in eBird we have applied a probabilistic approach called the Occupancy-Detection-Experience (ODE) model to provide an objective measure of expertise for all eBird observers [17]. The ODE model extends existing ecological models that measure the viability of a site as suitable habitat for a species, by predicting site occupancy by a particular species.

Table 2 Terms and notations used for the ODE models	
$z_l \in (0, 1)$	The occupancy of location l by the species of interest.
$y_i \in (0, 1)$	The detection/non-detection of the species of interest in observation i.
$e_b \in (nov., exp.)$	The expertise of the observer b
$\boldsymbol{v}_l$	A vector of environmental covariates for location l.
$\boldsymbol{w}_i$	A vector of covariates describing the observation process for observation i.
$\boldsymbol{u}_o$	A vector of expertise covariates for observer o.
$l(i)$	The location of observation i.
$o(i)$	The observer that recorded observation i

A general form for Occupancy Detection models is shown in Equation 1, where $\boldsymbol{v}_l$ is a set of environmental covariates for location l, $\boldsymbol{z}_l$ represents the occupancy of location l and $f^{occ}(v_l)$ is the function capturing the occupancy model (see Table 1 for notation description).

$$\Pr(z_l = 1) = f^{\text{occ}}(\boldsymbol{v}_l) \quad (1)$$

If a species is erroneously reported to be absent at a site when it was in fact present at that site, then species distribution models built from such data will underestimate the true occupancy of that species for that site. To address this issue, Mackenzie et al. [18] proposed an Occupancy-Detection (OD) model where true occupancy of a site l is represented as a latent variable $\boldsymbol{z}_l$. Under the OD model, a site is visited multiple times. Each visit $\boldsymbol{i}$ results in an observation $\boldsymbol{y}_i$, where the observation process is influenced by the true occupancy of the site and by a function $f^{\text{det}}(\boldsymbol{w}_i)$, where $\boldsymbol{w}_i$ are detection covariates (under the notation of Mackenzie et al. [21], $\psi = f^{\text{occ}}(\boldsymbol{v}_l)$ and $p = f^{\text{obs}}(\boldsymbol{w}_i)$). Equation 2 summarizes the process:

$$\Pr(y_i = 1) = z_{l(i)} \cdot f^{\text{obs}}(\boldsymbol{w}_i) \quad (2)$$

The OD model makes two key assumptions. First, it assumes population closure in which the true occupancy of a site $\boldsymbol{z}_l$ remains unchanged over the multiple visits to that site. Second, the OD model assumes that observers do not report false positives (i.e., an observer does not mistakenly report a species to be present when it is in fact absent).

The eBird experience level of an observer, which is the combination of their ability in identifying birds and their level of participation in eBird, can also influence the observation process. As a result we extended the OD model with an eBird experience component resulting in the Occupancy-Detection-Experience (ODE) model. In this extension, we add a new latent variable $E_{o(i)}$ and associated function $f^{\text{exp}}(\boldsymbol{u}_{o(i)})$ which capture the experience level (ie. eBird experience rated as *high* or *low*) of the observer $o(i)$ that recorded observation i.

As shown in Equation 3, this experience variable is a function of a set of covariates $\boldsymbol{u}_{o(i)}$ that include characteristics of the observer such as the total number of checklists submitted and relative to the total number of species reported, and the total number of flagged records rejected. As shown in Equation 4, the observation process is now influenced by the true occupancy of a site and by the function $f^{obs}(\boldsymbol{w}_i, e_{o(i)})$, which is now a function of the observation covariates.

$$\Pr\left(e_{o(i)} = 1\right) = f^{\text{exp}}\left(\boldsymbol{u}_{o(i)}\right) \quad (3)$$

$$\Pr(y_i = 1) = z_{l(i)} \cdot f^{\text{obs}}\left(\boldsymbol{w}_i, e_{b(i)}\right) \quad (4)$$

The ODE model relaxes the assumptions of the OD model by allowing false positives by the observers, for both levels of expertise. More details about the ODE model can be found in [17].

We can use the ODE model to distinguish the difference between expert observers, who will find more birds and are more likely to find them outside of the emergent filter limits, and novice birders, who are more likely to misidentify common birds. Table 1 (bottom) shows the total number of observations by experts and novices that are flagged. As expected, expert observers had a greater number of flagged records, because of their enhanced bird identification skills, and their desire to find unusual birds. We can use the ODE model results for experts in the data filtering process by automatically accepting their expert observations, which dramatically reduces the total number of flagged records that need to be reviewed. Finally, to test the accuracy of the ODE model we analyzed all observations that fell outside of the emergent filter for more than a dozen species that easily confuse novices, and show results for Chipping Sparrow (Figure 2). For all species, reviewers as valid observations accepted more than 95% of the expert observations that fell outside of the emergent filters.

We have found that the combination of the emergent checklist filters with the ODE model provides the best strategy for improving data quality in eBird. This two-step approach, where the emergent data filters are used to identify outliers, and the ODE model allowed us to identify valid outliers, identifies unusual records more accurately than previous methods. This approach establishes accurate occurrence probabilities and allows the quick identification and classification of outliers.

How can we address the spatial bias in citizen-science projects?

An inherent liability with many citizen-science projects is that observation locations are highly biased towards regions with high human populations. If this inequity is ignored, the spatial bias will produce results in which regions with the most data have excessive influence on the overall results accuracy and regions with the least data are under represented [8]. We address this issue using an AI mediated optimization strategy to identify areas that if sampled would most improve eBird spatial coverage.

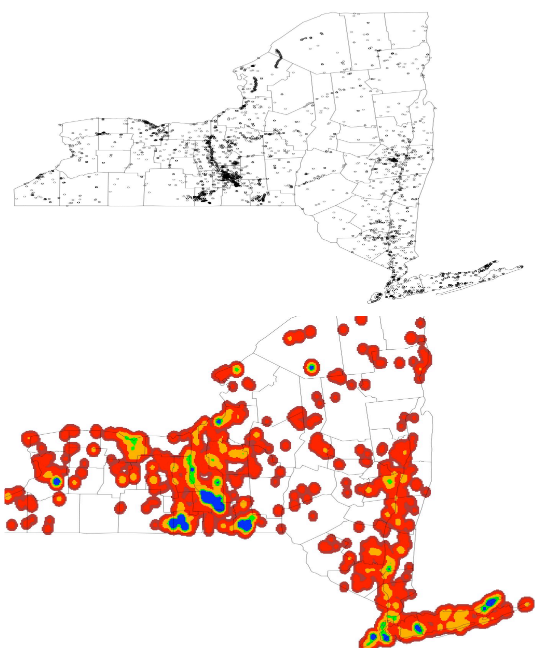

Figure 4. Top: locations in New York where submissions were made in eBird. Bottom: Results showing areas with sufficient data density (colored regions) and those requiring more data (white regions).

Machine learning algorithms can improve the predictive performance of eBird by guiding the sampling process. Consider eBird observations for New York (Figure 3). It is clear that spatial sampling biases are present as the majority of the observations come from a small subset of geographical locations. Active learning applied to eBird improve the resulting predictive models by providing a context to advise participants *where to sample next*. A first strategy, as displayed in figure 3, has been to aim for a uniform sampling coverage in geographical space, by concentrating data collection efforts to the areas of highest model uncertainty and low density. This is accomplished through a novel active learning approach that combines density information and information-theoretic measures [19].

Already, our research in offering optimal sampling strategies is paying off. We display maps similar to Figure 4 (bottom) on the eBird website, and provide rewards for individuals who report checklists from under sampled regions. Eventually, such sampling trajectories will be employed within eBird, to enhance the overall birding experience. For example, it is straightforward to propose paths that have the highest probability of detecting birds. Hence one can envision educating observers by proposing appropriate paths that trains their detection capabilities on specific species or increases the probability of them recording a species they have never observed before.

Challenge 4: Species Distribution Models

The motivation for eBird is to explore the continent-wide inter-annual patterns of occurrence of North American birds. To do this we have developed new Spatial-temporal Exploratory Models (STEM) of species distributions, that allow us to automatically discover patterns in spatiotemporal data [8]. We designed our statistical models specifically to discover seasonally- and regionally-varying patterns in eBird data. Spatiotemporal variation in habitat associations are captured by combining a series of separate submodels, each describing the distribution within a relatively small area and time window. The approach is semi-parametric, yielding a highly automated predictive methodology that allows an analyst to produce accurate predictions without requiring a detailed understanding of the underlying dynamic processes. This makes STEMs especially well suited for exploring distributional dynamics arising from a variety of complex dynamic ecological and anthropogenic processes. STEMs can be used to study how spatial distributions of populations respond over time to broad-scale changes in their environments, for example, changes in land-use patterns, pollution patterns, or climate change (Figure 5).

The STEM visualizations are now being employed in a number of research and conservation initiatives. For example, bird distribution information used in the *2011 State of the Birds Report* by the U. S. Department of Interior, was based on STEM model results. Additionally, other federal (i.e., Bureau of Land Management and U.S. Forest Service) and non-governmental agencies (i.e., The Nature

Conservancy) are using STEM distribution estimates to study placement of wind farms for sustainable energy production, identifying and prioritizing areas for avian conservation in the southwestern United States and the Pacific Northwest.

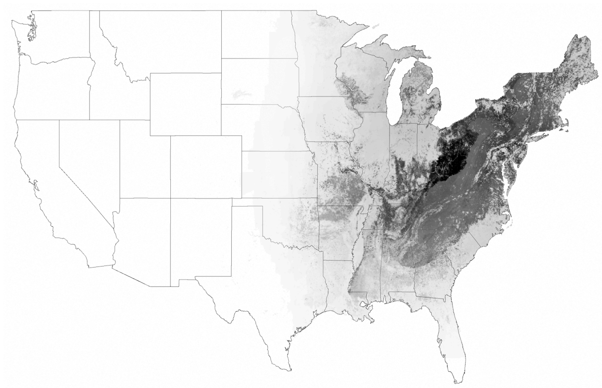

Figure 5. This map illustrates a STEM distribution estimate for Wood Thrush, a migratory songbird that winters in the the tropics and breeds in the northeastern U.S. and eastern Canada. The occurrence map shows the probability of encountering the species, with darker colors indicating higher probabilities. More STEM maps can be viewed on the eBird website (http://www.eBird.org).

Conclusion

In this paper, we have demonstrated the implementation of a novel network that links machine learning methods and human observational capacity to address several of the unique challenges inherent in a broad-scale citizen-science project. By exploring the synergies between mechanical computation and human computation, which we call a Human/Computer Learning Network we can leverage emerging technologies that integrate the speed and scalability of AI, with human computation to solve computational problems that are currently beyond the scope of existing AI algorithms.

Although our discussion has focused on one citizen-science project, eBird, the general HCLN approach are more widely applicable. Specifically, implementing an uncomplicated protocol and providing appropriate rewards for participation can recruit large numbers of participants. Then by using adaptive learning techniques for both humans and computers we can improve the quality and scope of the data that the volunteers provide. Finally, new analysis techniques that bridge the gap between parametric and non-parametric processes provide extremely accurate estimates of species occurrence at continental levels.

In conclusion, broad-scale citizen-science projects can recruit extensive networks of volunteers, who act as intelligent and trainable sensors in the environment to gather observations. However, there is much variability in the observations volunteers make. Artificial Intelligence processes can dramatically improve the quality of the observational data by filtering inputs using emergent filters based on aggregated historical data, and on the observers' expertise. By guiding the observers with immediate feedback on observation accuracy, the Artificial Intelligence processes contribute to advancing expertise of the observers, while simultaneously improving the quality of the training data on which the Artificial Intelligence processes make their decisions. The outcome is improved data quality that can be used for research and analysis.

Acknowledgments

This work was funded by the Leon Levy Foundation, Wolf Creek Foundation and the National Science Foundation (Grant Numbers OCI-0830944, CCF-0832782, ITR-0427914, DBI-1049363, DBI-0542868, DUE- 0734857, IIS-0748626, IIS-0844546, IIS-0612031, IIS-1050422, IIS-0905385, IIS-0746500, AGS-0835821, CNS-0751152, CNS-0855167).

References

[1] E. Law and L. v. Ahn, "Human Computation," *Synthesis Lectures on Artificial Intelligence and Machine Learning,* vol. 5, pp. 1-121, 2011/06/30 2011.

[2] Y. Man-Ching, C. Ling-Jyh, and I. King, "A Survey of Human Computation Systems," in *2009 International Conference on Computational Science and Engineering*, 2009, pp. 723-728.

[3] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popovic, and F. players, "Predicting protein structures with a multiplayer online game," *Nature,* vol. 466, pp. 756-760, 2010.

[4] C. J. Lintott, K. Schawinski, S. Anze, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg, "Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey," *Monthly Notices of*

the Royal Astronomical Society, vol. 389, pp. 1179-1189, 2008.

[5] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling, "eBird: A citizen-based bird observation network in the biological sciences," *Biological Conservation,* vol. 142, pp. 2282-2292, 2009.

[6] C. Wood, B. Sullivan, M. Iliff, D. Fink, and S. Kelling, "eBird: Engaging Birders in Science and Conservation," *PLoS Biol,* vol. 9, p. e1001220, 2011.

[7] S. Kelling, J. Yu, J. Gerbracht, and W. K. Wong, "Emergent Filters: Automated Data Verification in a Large-scale Citizen Science Project," *Proceedings of the IEEE eScience 2011 Computing for Citizen Science Workshop. To Appear.,* 2011.

[8] D. Fink, W. M. Hochachka, D. Winkler, B. Shaby, G. Hooker, B. Zuckerberg, M. A. Munson, D. Sheldon, M. Riedewald, and S. Kelling, "Spatiotemporal Exploratory models for Large-scale Survey Data," *Ecological Applications,* vol. 20, pp. 2131-2147, 2010.

[9] W. M. Hochachka, R. Caruana, D. Fink, A. Munson, M. Riedewald, D. Sorokina, and S. Kelling, "Data-mining discovery of pattern and process in ecological systems," *Journal of Wildlife Management,* vol. 71, pp. 2427-2437, 2007.

[10] M. J. Iliff, L. Salas, E. R. Inzunza, G. Ballard, D. Lepage, and S. Kelling, "The Avian Knowledge Network: A Partnership to Organize, Analyze, and Visualize Bird Observation Data for Education, Conservation, Research, and Land Management," presented at the Proceedings of the Fourth International Partners in Flight Conference: Tundra to Tropics, McAllen Texas, USA, 2009.

[11] N. A. B. C. I. U.S., "The State of the Birds 2011 Report on Public Lands and Waters," Washington, D.C.2011.

[12] J. E. McCormack, A.J. Zellmer, and L. L. Knowles, "Does niche divergence accompany allopatric divergence in Aphelocoma jays as predicted under ecological speciation?: insights from tests with niche models.," *Evolution,* pp. 1-13, 2009.

[13] A. H. Hurlbert and Z. Liang, "Spatiotemporal Variation in Avian Migration Phenology: Citizen Science Reveals Effects of Climate Change," *PLoS ONE,* vol. 7, p. e31662, 2012.

[14] J. Klicka, G. M. Spellman, K. Winker, V. Chua, and B. T. Smith, "A Phylogeographic and Population Genetic Analysis of a Widespread, Sedentary North American Bird: The Hairy Woodpecker (Picoides villosus)," *The Auk,* vol. 128, pp. 346-362, 2011/04/01 2011.

[15] J. Howe, *Crowdsourcing. Why The Power of the Crowd is Driving the Future of Business*. New York: Crown Business, 2008.

[16] S. Kelling, "The Significance of Observations in Biodiversity Studies," *GBIF Best Practices Series (*http://www.gbif.org/communications/news-and-events/showsingle/article/now-available-white-paper-on-significance-of-organism-observations/)*,* 2008.

[17] J. Yu, W. K. Wong, and R. Hutchinson, "Modeling experts and novices in citizen science data for species distribution modeling," presented at the IEEE International Conference on Data Mining, Sydney, Australia, 2010.

[18] D. I. MacKenzie, J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines, *Occupancy Estimation and Modeling*. Amsterdam: Elsevier, 2006.

[19] B. Dilkina, K. Lai, and C. Gomes, "Upgrading Shortest Paths in Networks.
Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems." vol. 6697, T. Achterberg and J. Beck, Eds., ed: Springer Berlin / Heidelberg, 2011, pp. 76-91.