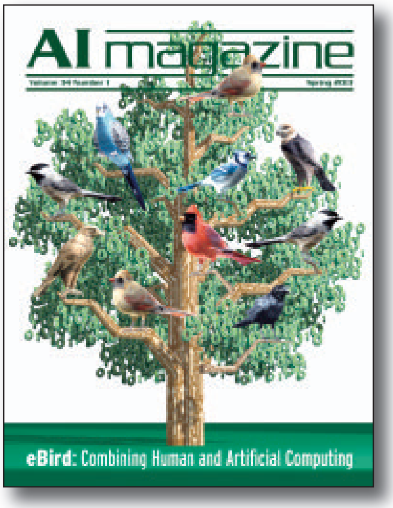
AI magazine

Volume 34 Number 1 Spring 2013

eBird: Combining Human and Artificial Computing

AI magazine

VOLUME 34, NUMBER 1 Spring 2013 ISSN 0738-4602



Cover: *eBird by Giacomo Marchesi, New York, New York.*

The guest editors for the four IAAI articles in this issue are Marcus Fromherz and Héctor Muñoz-Avila

EDITORIAL

IAAI ARTICLES

ARTICLE

COMPETITION REPORT

REPORTS

DEPARTMENTS

Introduction to the IAAI Articles in This Issue

Deployed Innovative Applications of Artificial Intelligence 2012

Markus Fromherz and Héctor Muñoz-Avila

■ *This issue of* AI Magazine *features expanded versions of articles that discuss deployed applications from the 2012 AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI-12).*

The Innovative Applications of Artificial Intelligence conference (IAAI) is the premier venue for documenting the transition of AI technology into applications. In this issue of *AI Magazine,* we continue our presentation of extended versions of papers presented at IAAI-12 (held in Toronto, Ontario, Canada) that were selected for their description of AI technologies that are in practical use.

Our selections for this issue describe *deployed* applications. They explain the context, requirements, and constraints of the application, how the technology was adapted to satisfy those factors, and the impact that this innovation brought to the operation in terms of cost and performance. The articles also supply useful insights into use cases that we hope can also be translated to other work that the AI community is engaged in.

In the first of these deployed application articles, eBird: A Human/Computer Learning Network to Improve Biodiversity Conservation and Research by Steve Kelling, Carl Lagoze, Weng-Keen Wong, Jun Yu, Theodoros Damoulas, Jeff Gerbracht, Daniel Fink, and Carla Gomes, the authors describe an intriguing application that successfully combines the best in human and artificial computing capabilities with an active feedback loop between people and machines.

The next two papers articles describe high-value industrial applications where diagnostic capabilities avoid considerable cost and accidents on a daily basis. A Real-Time Decision Support System for High Cost Oil Well Drilling Operations by Odd Erik Gundersen, Frode Sormo, Agnar Aamodt, and Pal Skalle applies case-based reasoning to the continuous analysis of drilling data to predict potential problems. Statistical Anomaly Detection for Train Fleets by Anders Holst, Markus Bohlin, Jan Ekman, Ola Sellin, Björn Lindström, and Stefan Larsen presents a novel statistical approach to detect anomalies in train operations and again anticipate problems.

A challenging category of applications are those that interact with users that are not aware of the capabilities and limitations of the underlying AI technology. In Applying Automated Language Translation at a Global Enterprise Level, Nestor Rychtyckyj and Craig Plesco describe their experience with adapting language translation technology to the very diverse needs of the employees in a large corporation. Likewise, although on a smaller scale, Stephanie Valentine, Francisco Vides, George Lucchese, David Turner, Hong-hoe Kim, Wenzhe Li, Julie Linsey, and Tracy Hammond present Mechanix: A Sketch-Based Tutoring and Grading System for Free-Body Diagrams, a tool that has been used in several engineering courses to support both teachers and students in explaining and learning the principles of statics.

We hope you enjoy and learn from the articles in this and the previous issue of *AI Magazine*. To learn more about innovative applications of artificial intelligence, we invited you to attend the upcoming conference in Bellingham, Washington, USA (www.aaai.org/iaai13). We also encourage you to submit a description of your AI application to future iterations of the Innovative Applications of Artificial Intelligence Conference, sponsored by AAAI.

Markus Fromherz was chair of the 2012 Innovative Applications of Artificial Intelligence conference. He is the chief innovation officer for healthcare at Xerox and a scientist and former director at the Palo Alto Research Center.

Héctor Muñoz-Avila was the cochair of the 2012 Innovative Applications of Artificial Intelligence conference. He is an associate professor at the Department of Computer Science and Engineering at Lehigh University.

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

Steve Kelling, Carl Lagoze, Weng-Keen Wong, Jun Yu,
Theodoros Damoulas, Jeff Gerbracht, Daniel Fink, Carla Gomes

■ *eBird is a citizen-science project that takes advantage of the human observational capacity to identify birds to species, and uses these observations to accurately represent patterns of bird occurrences across broad spatial and temporal extents. eBird employs artificial intelligence techniques such as machine learning to improve data quality by taking advantage of the synergies between human computation and mechanical computation. We call this 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. In this article we explore how human/computer learning networks can leverage the contributions of human observers and process their contributed data with artificial intelligence algorithms leading to a computational power that far exceeds the sum of the individual parts.*

The transformational power of today's computing, together with information and communication technologies, is providing new opportunities to engage the public to participate in and contribute to a myriad of scientific, business, and technical endeavors. For example, projects such as Galaxy Zoo, eBird, and FoldIt demonstrate the power of engaging the public in the investigation of a variety of 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 (Law and von Ahn 2011).

Human computational systems use the innate abilities of humans to solve certain problems that computers cannot solve (Man-Ching, Ling-Jyh, and King 2009). Now the World Wide Web and wireless handheld devices provide 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 (Cooper et al. 2010); or Galaxy Zoo, which has engaged more than 200,000 participants to classify more than 100 million galaxies (Lintott et al. 2008). Alternatively, the web can be used to engage volunteers to actively collect data and submit it to central data repositories. Projects such as eBird engage a global network of participants to

ISSN 0738-4602

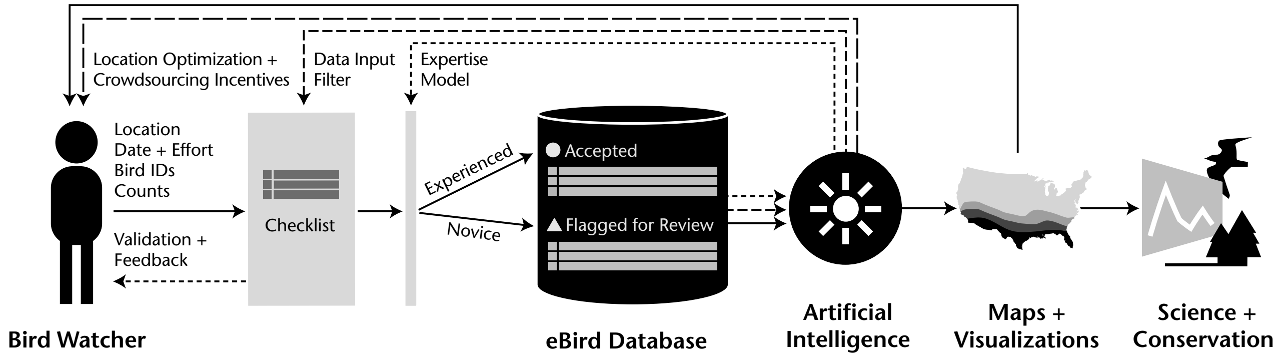


Figure 1. A Human / Computer Learning Network Example.

Human observers and AI processes synergistically improve the overall quality of the entire system. Additionally, AI is used to generate analyses. These analyses also improve as the quantity and quality of the incoming data improves.

report bird observations that are used to generate extremely accurate estimates of species distributions (Sullivan et al. 2009).

Now systems are being developed that employ both human and mechanical computation to solve complex problems through active learning and feedback processes. These human/computer learning networks (HCLNs) 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 an HCLN is illustrated in figure 1. In our example a broad network of volunteers acts as intelligent and trainable sensors to gather observations. AI processes dramatically improve the quality of the observational data the volunteers provide by filtering inputs based on aggregated historical data and observer 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.

A successful HCLN must be able to address the four following challenges. First, a task must be identified that human computational systems can complete but mechanical computational systems cannot (Law and von Ahn 2011). Second, the task must be sufficiently straightforward and incentivized to maximize participation (Wood et al. 2011). 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 (Kelling et al. 2011). Finally novel methods for aggregating and analyzing the noisy and complex data from multiple human computers must be employed (Fink et al. 2010). In this article we use our experience with eBird as a model to address these four 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 identify only 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 (Hochachka et al. 2007) 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 occur at that specific location and date until the species is correctly identified.

eBird[1] (Sullivan et al. 2009) 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 observations of birds to eBird through the web or wireless handheld devices (for example iPhone and Android). To date more than 91,000 individuals have volunteered more than 4 million hours and collected over 100 million bird observations, arguably the largest biodiversity data collection project in existence. These amassed observations provide researchers, scientists, students, educators, and amateur naturalists with data about bird distribution and abundance across a variety of spatiotemporal extents. Dynamic and interactive maps, graphs, and other visualizations are available on the eBird website, and all data are accessible through the Avian Knowledge Network (Iliff et al. 2009). Since 2006 eBird data have been used in more than 60 peer-reviewed publications and reports, from highlighting the importance of public lands in conservation (Kreger and Schmidt 2011) to studies of evolution (McCormack, Zellmer, and Knowles 2009), climate change (Hurlbert and Liang 2012), and biogeography (Klicka et al. 2011).

Challenge 2: Maximizing Participation

eBird uses crowdsourcing techniques to engage a large numbers of people to perform tasks that automated sensors and computers cannot readily accomplish (Howe 2008). This is accomplished through the development of straightforward rules for participation and incentives for contributing. Initial incentives focused on helping scientists study birds. This approach led to disappointing participation in eBird. Recognizing this, the emphasis of the project was changed from having birders help scientists to tools that appealed to the birding community. New features were developed for eBird that allowed participants to (1) keep track of their bird records; (2) sort their personal bird lists by date and region; (3) share their lists with others; and (4) visualize their observations on maps and graphs. By providing these record keeping, exploration, and visualization facilities as a direct reward for participation, eBird participation has grown exponentially (figure 2). eBird appeals to the competitiveness of participants, and through the further development of eBird more interactive and varied tools allowed participants to determine their relative status compared to other participants (such as numbers of species seen) and geographical regions (such as checklists submitted per state and province). Thus, by changing the emphasis of eBird to one that supports the needs and desires of the birding community. For example, more data were gathered in May 2012 than during the first 3 years of the project.

An additional 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 reuse 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 (Kelling 2008) 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.

Challenge 3: Identifying the Synergies Between Humans and Machines

While eBird has been successful in engaging a global community of volunteers to contribute large quantities of observations of birds, there are many challenges to using eBird data for analysis. First, observers are bound to misidentify birds, which is the largest source of error in the eBird database. Second, there is much variability in a participant's ability to identify birds, with some eBird contributors being experts in bird identification, while others are novices. Third, participation in eBird is not uniformly distributed in space; most eBird obser-

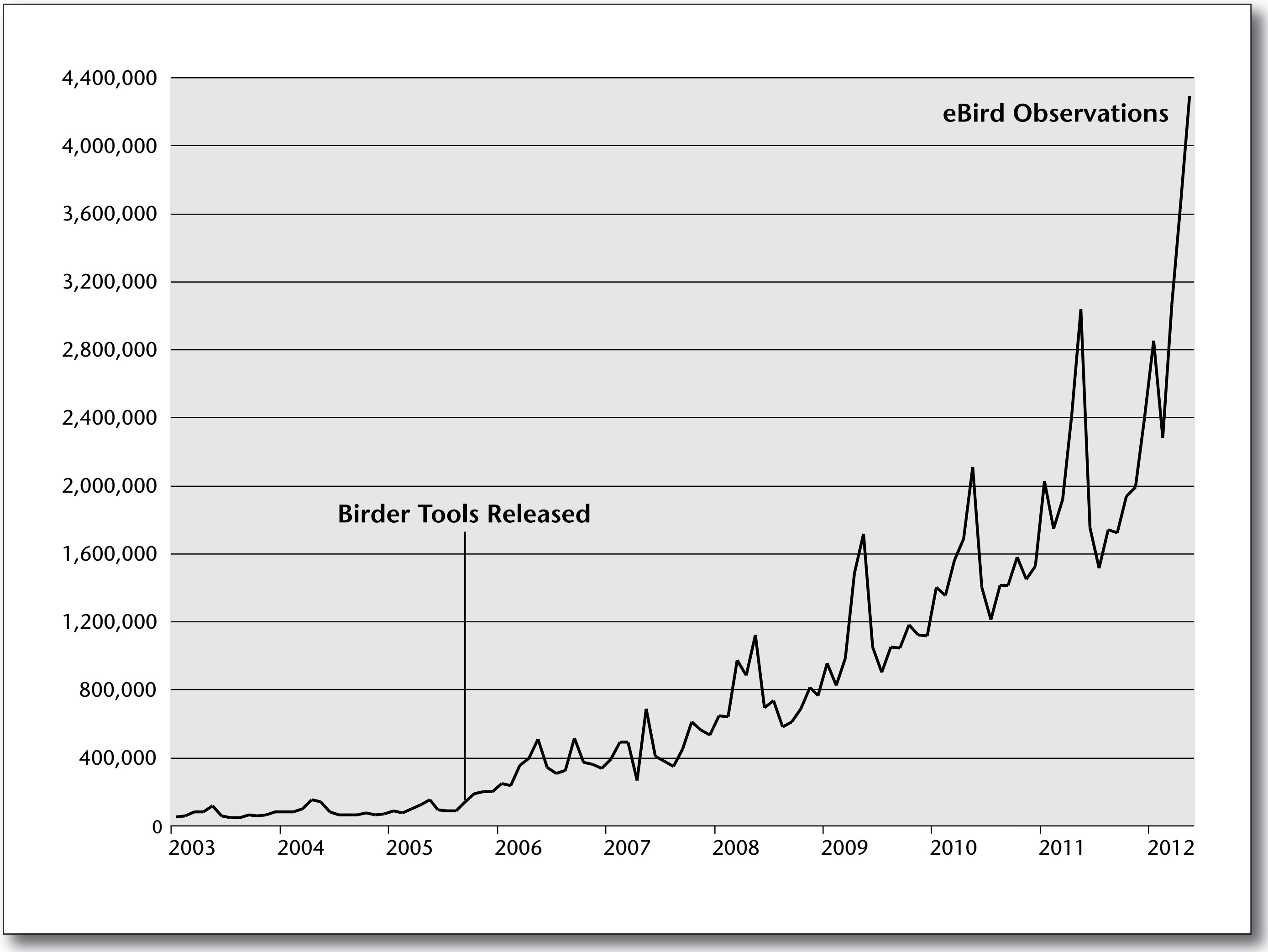


Figure 2. The Number of Observations Submitted Monthly to eBird Since Its Inception in Late 2003.

Tools to better engage the bird watching community were released in mid-2005. Note the annual peaks of submission, which occur each May, when spring migration is at its peak and birders are most active.

vations occur in regions where human population densities are fairly high. Improving eBird data quality is a constant and major effort. This is because as data quality improves the accuracy in estimating patterns of bird occurrence also improves. In this section we describe how the implementation of HCLN processes allow us address these three data-quality issues.

How Can We Efficiently Filter Erroneous Data?

Data quality is a major issue for eBird, particularly as it pertains to an observer's ability to correctly identify birds to the species level. While eBird has motivated tens of thousands of volunteers to collect large amounts of data at relatively little cost, the misidentification of birds is a major concern. To overcome this issue eBird has employed a data-validation system that relies heavily on a network of volunteer experts in bird occurrence to vet the data. However, the sheer volume of data being gathered was beginning to overwhelm the volunteer editor network. Initially the system of record review had been established to filter data regionally (that is, country, state, county), and temporally at the monthly scale. The basic filter mechanism assigned a specific region with a value for a given month, which corresponded to an expert's opinion for a maximum allowable acount for a given region. If a submission exceeded the maximum allowable amount, it was "flagged" for review by one of more than 450 volunteer reviewers. Reviewers contacted those individuals who submitted flagged records to obtain additional information, such as field notes or photographs, in order to con-

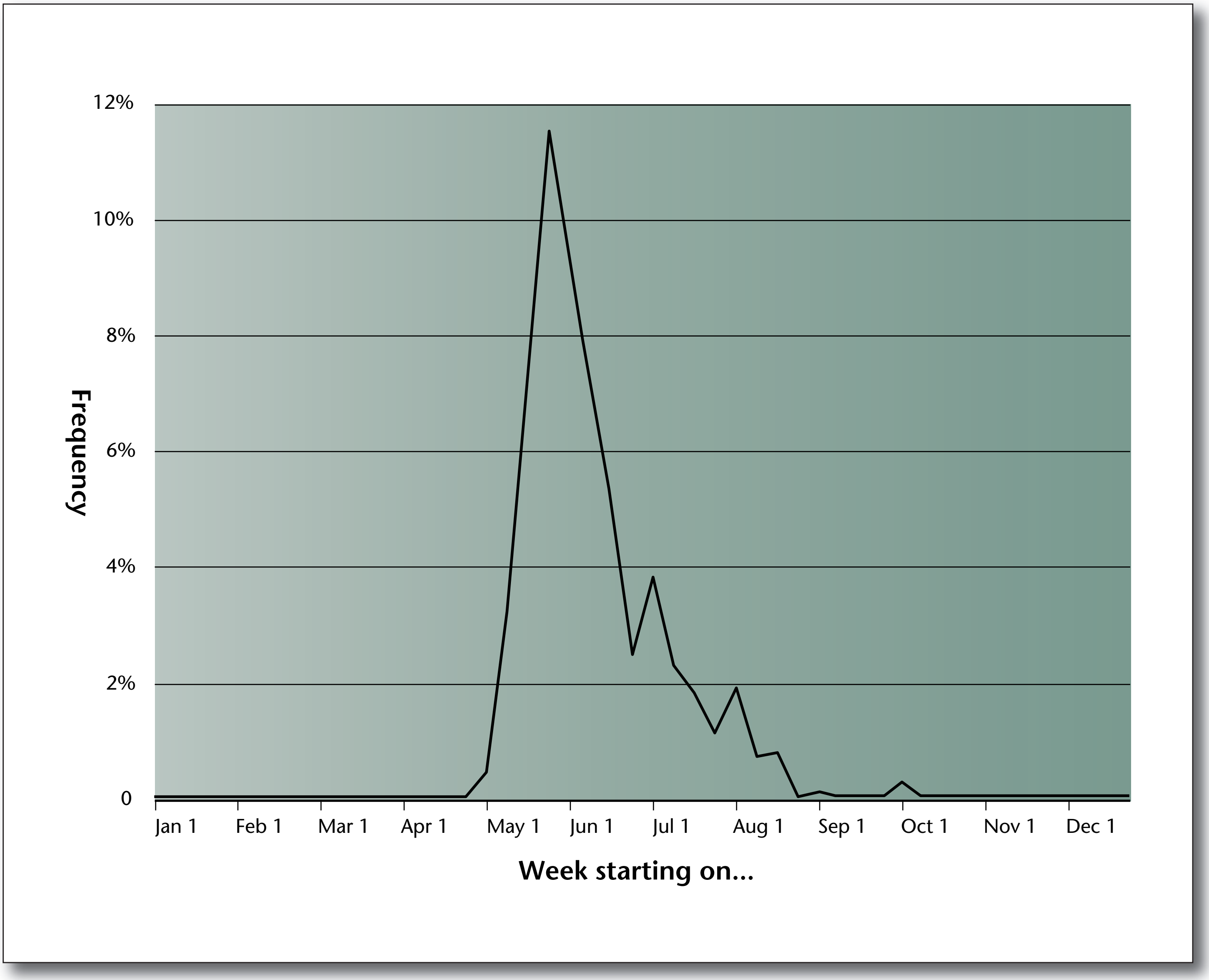


Figure 3. Frequency of Occurrence Results for Black-Billed Cuckoos (Coccyzus erythropthalmus) in Upstate New York.

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.

firm unusual reports. In 2010, 4 percent or 720,000 of the 23 million records submitted to eBird were reviewed. This number put a severe strain on the volunteer network, with many reviewers complaining they were overwhelmed from the sheer volume of records to review.

To decrease the volume of data that needed to be reviewed by the experts, we have implemented a new data-quality filter and screening process that automates much of the review process. This new process was based on one of the most powerful calculations performed on eBird data, which is the frequency in which a particular bird species was reported during a particular period of time (figure 3). 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 (for example, country, state, county, backyard, or any spatial extent of interest) and for any date. Frequency filters delineate when a species can be reported in a region and determines the validity of an observation.

The eBird database currently holds more than 100 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

	Tompkins County	Jefferson County
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

Upper: Applied to two counties in upstate New York. *Lower:* The proportion of flagged records submitted by experts and novices.

from historic eBird data. We have set the emergent filter at 10 percent of maximum annual frequency of occurrence for every species across the United States. 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 percent then the filter would identify the day when the filter first crosses the 6.8 percent threshold. Any record submitted on a date either prior or after the threshold limit is flagged for review. Similarly, if a rare species reaches an annual peak of 6.5 percent frequency, the threshold limit would be .65 percent. Table 1 shows the number of flagged records the emergent filter identifies for two counties in New York State, Jefferson County and Tompkins County. These two counties were selected because Jefferson County has relatively sparse year-round data coverage, while Tompkins County is one of the most active regions in eBird.

When the emergent filter is triggered the submitter gets immediate feedback indicating that this was an unusual observation (figure 1). If submitters confirm that 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.

What is most significant about the emergent filter process is that it identifies key periods during a bird's life history when its patterns of occurrence change (for example during periods of migration when the bird either arrives or departs a specific region). Figure 4 shows those records that are flagged for review by the emergent filter for the 2 New York counties. The Chipping Sparrow *(Spizella passerina)* is a common breeding bird in upstate New York, but departs the region in the fall and 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.

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 interobserver variation must be taken into account during analysis to determine if outlier observations (that is, 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 automatically and 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 (that is, skilled birders like to find rare birds). What is required is an objective measure of observer expertise that would automatically classify unusual observations.

To better understand observer variability in

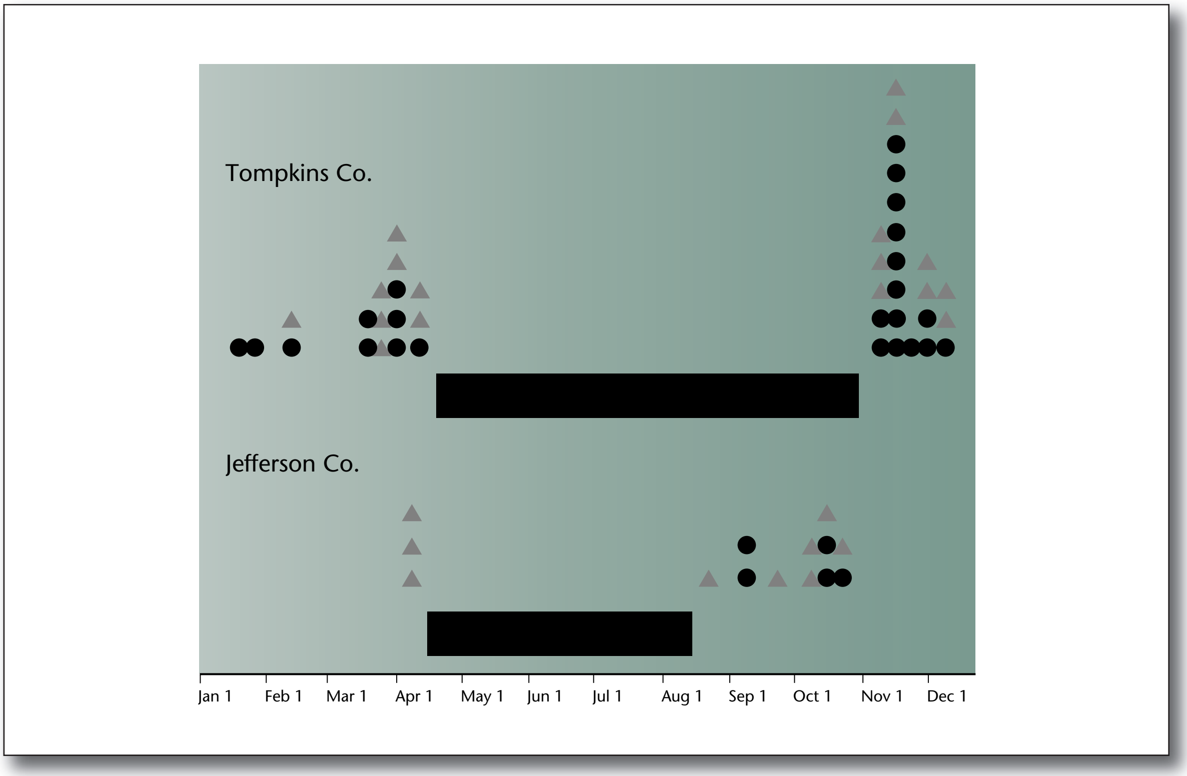


Figure 4. The Acceptable Date Range for the Occurrence of Chipping Sparrow in Two Counties in New York.

Date ranges are the dark bars. All records that fall outside of the acceptable date range are plotted either as circles (novices) or triangles (experts).

eBird we have applied a probabilistic machine-learning approach called the occupancy-detection-experience (ODE) model to provide an objective measure of expertise for all eBird observers (Yu, Wong, and Hutchinson 2010). 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.

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 (table 1 bottom). 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 we show results for the Chipping Sparrow (figure 4). We did this by engaging the current reviewers for the two counties in New York, who confirmed that the ODE model properly categorized the observer as either an expert or novice and validated more than 95 percent 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 both improving data quality and streamlining the review process in eBird. This two-step approach, where the emergent data filters are used to identify outliers and the ODE model allows us to identify valid outliers, identifies unusual records more accurately than previous methods. The result is that we can now provide accurate occurrence probabilities, which are based on existing eBird data to allow 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 toward 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

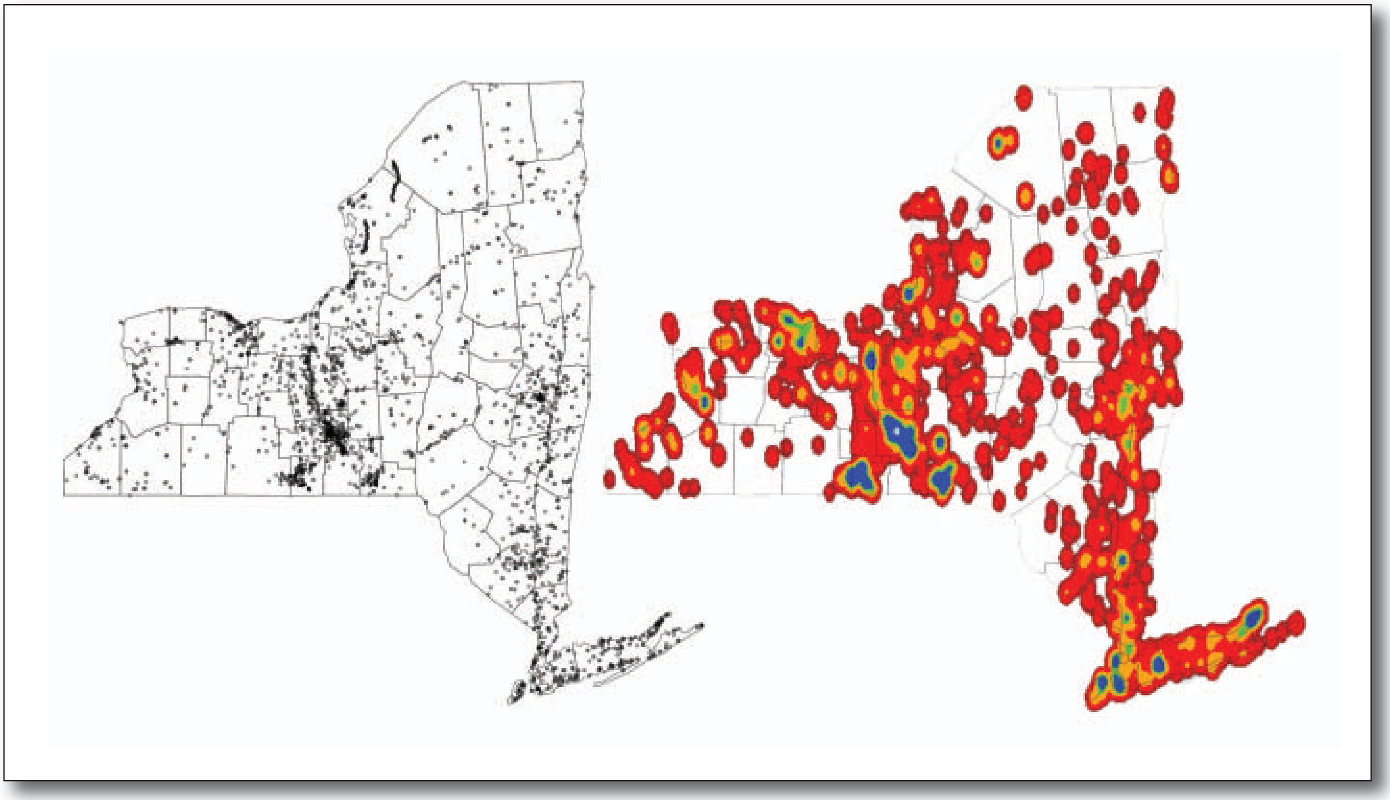


Figure 5. Locations and Results.

Top: Locations in New York where submissions were made to eBird in 2009. *Bottom:* Results showing areas with sufficient data density (shaded regions) and those requiring more data (white regions).

result's accuracy and regions with the least data are underrepresented (Fink et al. 2010). We address this issue using a mediated optimization strategy to identify areas that if sampled would most improve eBird spatial coverage and improve analysis results.

Machine-learning algorithms can improve the predictive performance of eBird by guiding the sampling process. Consider the locations where eBird observations were made in New York (figure 5 top). 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 improves the resulting predictive models by providing a context to advise participants where to sample next. A first strategy, as displayed in figure 5 (bottom), 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 (Dilkina, Lai, and Gomes 2011).

Already, our research in offering optimal sampling strategies is paying off. We display maps similar to figure 5 (bottom) on the eBird website and provide rewards for individuals who report checklists from undersampled regions. Eventually, such sampling trajectories will be employed throughout 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 train their detection capabilities on specific species or increase the probability of them recording a species they have never observed before.

Challenge 4: Species Distribution Models

The effective management and conservation of biodiversity requires knowledge of a species' geographic distribution throughout the year. Until the inception of eBird, detailed data documenting a species' distribution were often not available for the entire species' range, particularly for widely distributed species or species not closely studied. eBird provides broad-scale survey data that allows researchers to analyze and interpret a species' distribution across broad spatial extents and for any time of year.

One major area of analysis of eBird data is to explore the continentwide interannual patterns of occurrence of North American birds. To do this we have developed new spatiotemporal exploratory models (STEMs) of species distributions that allow us to automatically discover patterns in spatiotemporal data (Fink et al. 2010).

We designed our statistical models specifically to discover seasonally and regionally varying patterns in eBird data. Spatiotemporal variation in habitat associations is captured by combining a series of separate submodels, each describing the distribution within a relatively small area and time window. The approach is semiparametric, 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, both seasonally (figure 6) as well as to broad-scale changes in their environments (that is, changes in land-use patterns, pollution patterns, or climate change).

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 prepared for the U. S. Department of Interior by the North American Bird Conservation Initiative (NABCI), was based on STEM model results. Additionally, other federal (that is, Bureau of Land Management and U.S. Forest Service) and nongovernmental agencies (that is, 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.

Conclusion

In this article, we have demonstrated the implementation of a novel network that links machine-learning methods and human observational capacity to address several 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.

eBird uses a broad-scale survey design to maximally engage volunteers to gather bird observations following a basic protocol for data collection. Designing such broad-scale surveys to maximize the information obtained for use in analysis depends on finding the proper balance between data quantity and data quality. If we can engage a large number of participants to collect data through eBird's very basic protocols a sufficiently large volume of data can be gathered and effectively analyzed. While eBird data has relatively lower per datum information content, we have found that eBird data can contain more information for broad-scale distribution estimates than a smaller amount of data with higher per datum quality.

The appropriate design of data input and management procedures is critical to maintain the balance between data quantity and data quality in broad-scale citizen-science projects. The additional implementation of novel AI functionality provides incentives for encouraging surveyors to contribute even more data while simultaneously limiting errors and providing opportunities for dramatically improved data review and validation procedures.

Although our discussion has focused on one citizen-science project, eBird, the general HCLN approach is more widely applicable. Specifically, by implementing an uncomplicated protocol through web-based and wireless handheld devices and providing appropriate rewards for participation, citizen-science projects can recruit large numbers of participants to submit massive quantities of meaningful data. By taking an adaptive learning approach 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 nonparametric 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 that gather observations across broad spatial (for example, globally) and temporal (for example, any time) extents. 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 HCLN process contributes to advancing expertise of the observers, while simultaneously improving the quality of the training data on which the machine-learning processes make their decisions. The outcome is improved data quality that can be used for research and analysis.

Acknowledgements

This work was funded by the Leon Levy Foundation, Wolf Creek Foundation, and the National Science Foun-

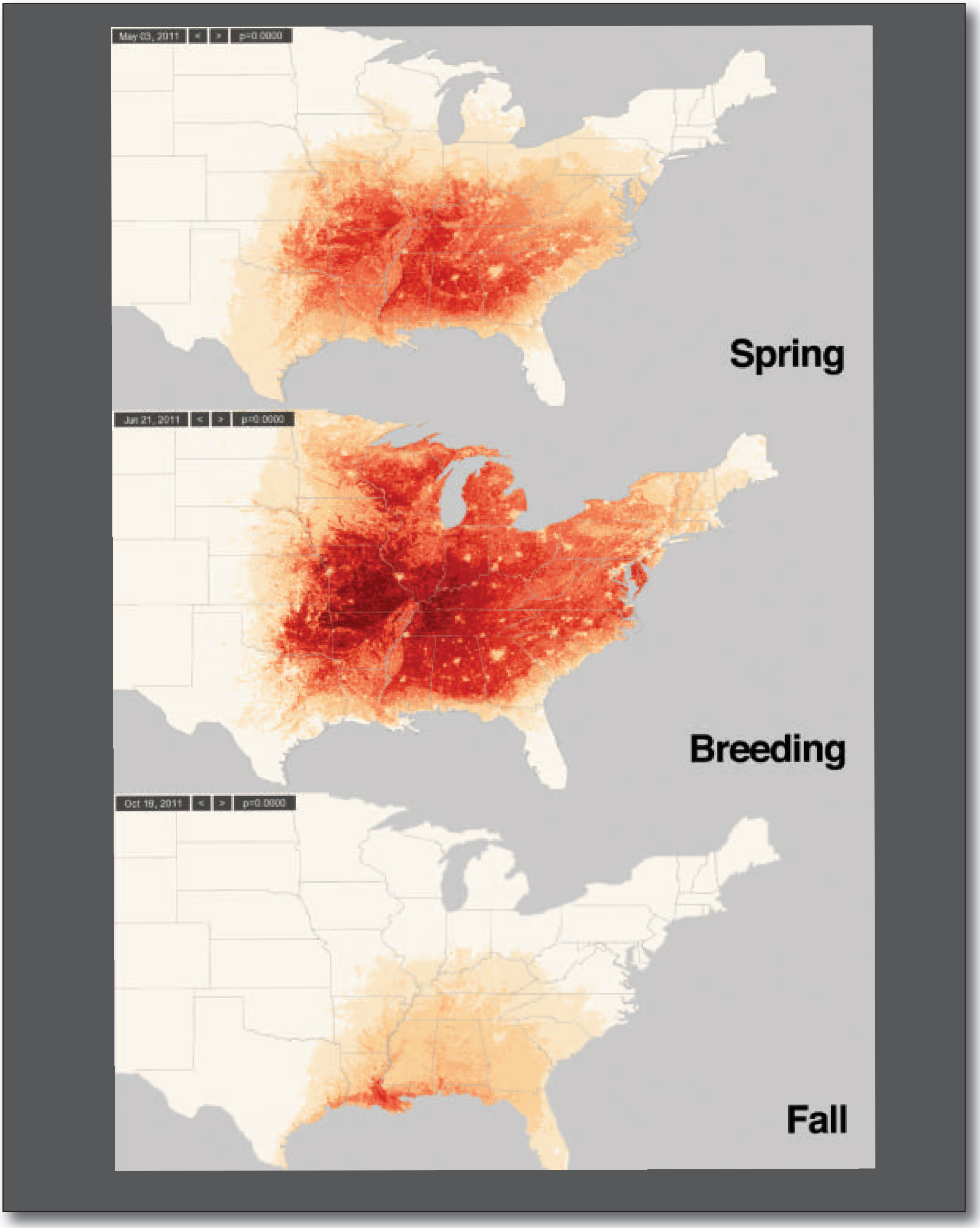


Figure 6. Maps Illustrating the Seasonal Patterns of Occurrence of the Indigo Bunting (Passerina cyanea) Throughout the United States.

The maps illustrate the seasonal distribution estimates from a STEM during spring migration (top), the breeding season (middle), and during fall migration (bottom). Indigo Buntings are neotropical migrants, wintering in Central America and returning to the United States annually to breed. The occurrence maps show the probability of encountering the species on an early morning 1-hour birding walk, with darker colors indicating higher probabilities. These maps provide continental-scale distribution estimates that allow the quick assessment of the rate of arrival and departure from breeding grounds, and over time will allow researchers to identify and quantify changes in bird populations. More STEM maps can be viewed on the eBird website.[2]

dation (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, IIS-1209589, AGS-0835821, CNS-0751152, CNS-0855167).

Notes

1. See ebird.org.

2. www.ebird.org.

References

Cooper, S.; Khatib, F.; Treuille, A.; Barbero, J.; Lee, J.; Beenen, M.; Leaver-Fay, A.; Baker, D.; Popovic, Z.; and Players, F. 2010. Predicting Protein Structures with a Multiplayer Online Game. *Nature* 466(7307): 756–760.

Dilkina, B.; Lai, K. J.; and Gomes, C. P. 2011. Upgrading Shortest Paths in Networks. In *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems: Proceedings of the 8th International Conference*. Lecture Notes in Computer Science 6697, 76–91. Berlin: Springer.

Fink, D.; Hochachka, W. M.; Zuckerberg, B.; Winkler, D. W.; Shaby, B.; Munson, M. A.; Hooker, G.; Riedewald, M.; Sheldon, D.; and Kelling, S. 2010. Spatiotemporal Exploratory Models for Large-Scale Survey Data. *Ecological Applications* 20(8): 2131–2147.

Hochachka, W. M.; Caruana, R.; Fink, D.; Munson, A.; Riedewald, M.; Sorokina, D.; and Kelling, S. 2007. Data-Mining Discovery of Pattern and Process in Ecological Systems. *Journal of Wildlife Management* 71(7): 2427–2437.

Howe, J. 2008. Crowdsourcing. W*hy The Power of the Crowd Is Driving the Future of Business*. New York: Crown Business.

Hurlbert, A. H., and Liang, Z. 2012. Spatiotemporal Variation in Avian Migration Phenology: Citizen Science Reveals Effects of Climate Change. *PLoS ONE* 7(2): e31662.

Iliff, M.; Salas, L.; Inzunza, E. R.; Ballard, G.; Lepage, D.; and Kelling, S. 2009. The Avian Knowledge Network: A Partnership to Organize, Analyze, and Visualize Bird Observation Data for Education, Conservation, Research, and Land Management. Paper presented at the Fourth International Partners in Flight Conference: Tundra to Tropics: Connecting Birds, Habitats, and People. McAllen, Texas, 13–16 Feb.

Kelling, S. 2008. *The Significance of Observations in Biodiversity Studies*. Copenhagen, Denmark: Global Biodiversity Information Facility (www.gbif.org/communications/news-and-events/showsingle/article/now-available-white-paper-on-significance-of-organism-observations).

Kelling, S.; Yu, J.; Gerbracht, J.; and Wong, W.-K. 2011. Emergent Filters: Automated Data Verification in a Large-Scale Citizen Science Project. Paper presented at the 2011 IEEE Seventh International Conference on e-Science Workshop, Stockholm, 5–8 December.

Klicka, J., Spellman, G. M.; Winker, K.; Chua, V.; and Smith, B. T. 2011. A Phylogeographic and Population Genetic Analysis of a Widespread, Sedentary North American Bird: The Hairy Woodpecker *(Picoides villosus)*. *The Auk* 128(2): 346–362.

Kreger, M., and Schmidt, P., eds. 2011. *U.S. State of the Birds 2011 Report on Public Lands and Waters: 48*. Washington, D.C.: United States Department of the Interior.

Law, E., and von Ahn, L. 2011. Human Computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5(3): 1–121.

Lintott, C. J.; Schawinski, K.; Slosar, A.; Land, K.; Bamford, S.; Thomas, D.; Raddick, M. J.; Nichol, R. C.; Szalay, A.; Andreescu, D.; Murray, P.; and van den Berg, J. 2008. Galaxy Zoo: Morphologies Derived from Visual Inspection of Galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389(3): 1179–1189.

Man-Ching, Y.; Ling-Jyh, C.; and King, I. 2009. A Survey of Human Computation Systems. In *Proceedings of the 2009 IEEE International Conference on Compuational Science and Engineering*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

McCormack, J. E.; Zellmer, A. J.; and Knowles, L. L. 2009. Does Niche Divergence Accompany Allopatric Divergence In *Aphelocoma* Jays as Predicted Under Ecological Speciation?: Insights From Tests With Niche Models. *Evolution* 64(5): 1231–1244.

Sullivan, B. L.; Wood, C. L.; Iliff, M. J.; Bonney, R. E.; Fink, D.; and Kelling, S. 2009. eBird: A Citizen-Based Bird Observation Network in the Biological Sciences. *Biological Conservation* 142(10): 2282–2292.

Wood, C.; Sullivan, B.; Iliff, M.; Fink, D.; and Kelling S. 2011. eBird: Engaging Birders in Science and Conservation. *PLoS Biology* 9(12): e1001220.

Yu, J.; Wong, W. K.; and Hutchinson, R. A. 2010. Modeling Experts and Novices in Citizen Science Data for Species Distribution Modeling. In *Proceedings of the 2010 IEEE International Conference on Data Mining (ICDM),* 1157–1162. New York: Institute of Electrical and Electronics Engineers.

Steve Kelling, director of information science at the Cornell Lab of Ornithology, leads a team of ornithologists, computer scientists, statisticians, application developers, and data managers to develop programs, tools, and analyses to gather, understand, and disseminate information on birds and the environments they inhabit.

Carl Lagoze is an associate professor at the University of Michigan School of Information. His current research focuses on sociotechnical aspects of infrastructure for eScience.

Weng-Keen Wong is an associate professor of computer science at Oregon State University. His research areas are in data mining and machine learning, with specific interests in anomaly detection, ecosystem informatics, and “human in the loop” machine learning.

Jun Yu is a Ph.D. candidate at Oregon State University, working under the guidance of Weng-Keen Wong. His current research involves machine learning in citizen science and species distribution modeling, and multilabel classification.

Theo Damoulas is a research associate in the Department of Computer Science at Cornell University and also a member of the Institute for Computational Sustainability and affiliated researcher of the Cornell Lab of Ornithology. His research interests are in machine learning and artificial intelligence, with applications to biology and sustainability.

Jeff Gerbracht is a senior software engineer at the Cornell Lab of Ornithology where he is the technical lead for eBird. His current focus is implementing data-quality processes that utilize the current dataset to identify outliers across multiple dimensions.

Daniel Fink is a research associate at the Cornell Lab of Ornithology. His current research focuses on developing statistical methods to discover and quantify multiscale spatiotemporal patterns using citizen science data.

Carla Gomes is a professor of computer science and the director of the Institute for Computational Sustainability at Cornell University. Her research themes include constraint reasoning, mathematical programming, and machine learning, for large scale combinatorial problems. Recently, Gomes has helped found the new field of computational sustainability, which is her current main research focus.