Scolopax: Exploratory Analysis of Scientific Data

Alper Okcan* Mirek Riedewald* Biswanath Panda† Daniel Fink‡

*Northeastern University
Boston, MA, USA
{okcan,mirek}@ccs.neu.edu

†Google Inc.
Mountain View, CA, USA
bpanda@google.com

‡Cornell Lab of Ornithology
Ithaca, NY, USA
daniel.fink@cornell.edu

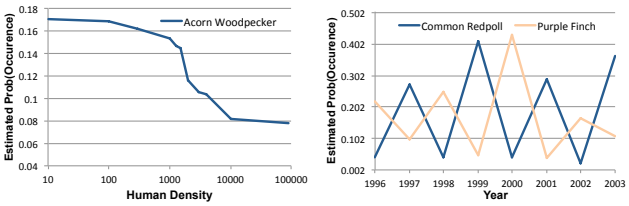
ABSTRACT

The formulation of *hypotheses* based on patterns found in data is an essential component of scientific discovery. As larger and richer data sets become available, new scalable and user-friendly tools for scientific discovery through data analysis are needed. We demonstrate Scolopax, which explores the idea of a search engine for hypotheses. It has an intuitive user interface that supports sophisticated queries. Scolopax can explore a huge space of possible hypotheses, returning a ranked list of those that best match the user preferences. To scale to large and complex data sets, Scolopax relies on parallel data management and mining techniques. These include model training, efficient model summary generation, and novel parallel join techniques that together with traditional approaches such as clustering manipulate massive model-summary collections to find the most interesting hypotheses. This demonstration of Scolopax uses a real observational data set, provided by the Cornell Lab of Ornithology. It contains more than 3.3 million bird sightings reported by citizen scientists and has almost 2500 attributes. Conference attendees have the opportunity to make novel discoveries in this data set, ranging from identifying variables that strongly affect bird populations in specific regions to detecting more sophisticated patterns such as habitat competition and migration.

1. INTRODUCTION

Across many disciplines, scientists have to cope with a flood of data. One of the grand challenges of data-driven science is to find interesting patterns in massive high-dimensional data sets that may lead to new hypotheses. This process is currently limited by the large amount of required human effort and the high computational cost. Scalable exploration tools are required to make pattern search in big data (almost) as easy as searching the Web, while hiding technical data management challenges.

We demonstrate *Scolopax*, a data exploration tool that is based on novel data management techniques and enables scientists to create hypothesis search queries through a user-friendly interface. Scolopax' functionality will be demonstrated for a real data set that was extracted from eBird reports (`http://ebird.org`) in collaboration with the Cornell Lab of Ornithology, one of the world's leading research institutes for birds. eBird is a citizen-science project that takes a crowdsourcing approach, harnessing the power of birders around the world to collect observations of birds with broad scale and fine resolution. It has amassed one of the largest biodiversity data resources in existence and is rapidly growing, at a rate of millions of observations monthly. Since all reports contain location and time, the eBird data is joined with high-quality environmental and geographic data sets, e.g., climate, habitat, and human census data. For our demonstration, we will work with the eBird Reference Data Set version 4.0, which has more than 3.3 million checklists and almost 2500 attributes.



(a) (b)

Figure 1: Example summaries of a prediction model

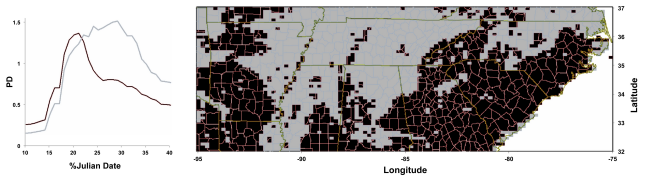
Conference attendees will have the opportunity to make real discoveries in this exciting data set. The following examples illustrate patterns that could be found and the corresponding hypotheses proposed by the domain scientists.

Figure 1a is a 1-dimensional summary that shows how the estimated probability of observing the Acorn Woodpecker in California varies with local human population density. After seeing the strong relationship between increasing human population density and lower woodpecker probability, the scientists hypothesized that the bird's reliance on dying tree branches for storing acorns conflicts with properties of habitats in more densely settled areas.

The plots in Figure 1b represent the annual estimated observation probability of two bird species in Northern New England. The strikingly complementary bi-annual cycles for Purple Finch and Common Redpoll hint at habitat competition, possibly driven by the availability of local food sources.

Broad-scale migratory behavior can be explored through

Proceedings of the VLDB Endowment, Vol. 6, No. 12



(a) (b)

Figure 2: Average cluster trajectories for spring migration and cluster map of Tree Swallow

analysis of similar univariate summaries describing how local occurrence probabilities vary throughout the year. Each summary indicates the time of the year that the species is likely to be observed in the given locality. By clustering local occurrence trajectories across broad extents, the geographic patterns of movement that characterize migrations can be revealed. Figure 2 presents an example for spring migration flyways of the Tree Swallow. In Figure 2a, the average spring cluster trajectories are shown; Figure 2b presents the visualization of the regions and to which of two clusters they belong. Based on this image, scientists hypothesize that an early initial wave of migrants is flying through the dark Piedmont/Southern Appalachian region and the larger wave of migrants moves north a little later through the light region.

Supporting discovery of such patterns is challenging. Complex prediction models need to be trained to address issues such as noise, sparseness, and skew in the raw data. Summaries need to be extracted from the models and raw data in order to find intelligible patterns [1]. An interesting hypothesis could be found in some "slice" or "dice" of the data space, e.g., a certain trend might show only in some small region or a combination of habitat features and elevation ranges. Hence to broadly explore possible hypotheses, a **huge number and variety of such summaries**, both one- and multi-dimensional, needs to be generated. To find correlations, e.g., habitat competition between two species or variables with similar effect on a species, **joins with complex predicates** are needed, e.g., using inequality conditions. As some promising patterns are found, scientists want to **interactively post-process** the result set, e.g., by refining a geographic selection through drill-down or by eliminating groups of uninteresting results.

Scolopax achieves high performance and scalability by relying on novel data management techniques for parallel model-summary generation and parallel processing of theta-joins. Conference attendees will be able to perform real exploratory analysis and also see visualizations of Scolopax' performance.

2. TECHNOLOGY BEHIND SCOLOPAX

Scolopax is designed as a data-driven Web application. The UI runs in any standard Web browser, connecting to an application server. All the heavy lifting is performed in parallel on a cluster running the Hadoop MapReduce system. The Hadoop-powered backend performs tasks such as parallel training of models that accurately predict species observation probabilities, generation of huge collections of model and raw data summaries, join processing for correlation analysis, and clustering. Model-summary generation [4] and theta-join processing [3] showcase our recently proposed data management research techniques and are briefly summarized in this section.

2.1 Model-Summary Generation

As the examples in Section 1 illustrate, intelligible low-dimensional summaries are the basic ingredient for the patterns that motivated a variety of different types of hypotheses. For various reasons, most of them related to data quality and sparseness[1], many scientists prefer to work with sophisticated prediction models that are trained on the collected data. Summaries are then derived from the model, essentially capturing the most important low-dimensional relationships contained in the complex (high-dimensional) model.

Generating a single summary is fairly expensive, requiring a large number of data records to be processed by the model. For broad exploration, scientists want to include many possible summaries. Hence summaries are generated for many combinations of variables, for different geographical regions, but also for a variety of other partitionings, e.g., by combinations of habitat and climate features. (This enables discovery of more complex interactions, e.g., where some habitat change only affects bird populations in certain climate types.)

Based on a careful analysis of workload and model properties, we proposed novel techniques for speeding up computation of massive model-summary collections by one or more orders of magnitude [4]. We model the problem as a database-style query plan and show how it can be decomposed into two much cheaper steps without affecting result quality. Our approach relies on memoization (which short-circuits expensive model evaluation processes), pushing materialized aggregates into model evaluation, and effective bulk-computation that exploits common properties of different summaries. In addition to reducing total cost, we also propose a parallel implementation in MapReduce that shows near-optimal speedup.

2.2 Processing Theta-Joins

Once a large collection of summaries is generated, Scolopax explores and ranks patterns by processing these summaries using relational operators such as selection, grouping and join, but also data mining techniques like clustering. The join operator is of particular interest, because scientists are very interested in discovering various types of relationships and correlations. Most of these scenarios require flexible join predicates, not just equality. For example, the following query is executed in Scolopax to search for habitat competition between species (similar to Figure 1b):

```
SELECT S.plotData, T.plotData, S.sumAttr
FROM Summaries AS S, Summaries AS T
WHERE S.species ≠ T.species
      AND S.sumAttr = T.sumAttr
      AND S.region = T.region
      AND |φ(S.plotData, T.plotData)| > ε
```

This query is a self-join to find summaries of different species having the same summary attribute (x-axis of the plot) and complementary trends in the same region where ϕ

[1]Note that sparseness is still a major issue for Big Data due to the large number of attributes.

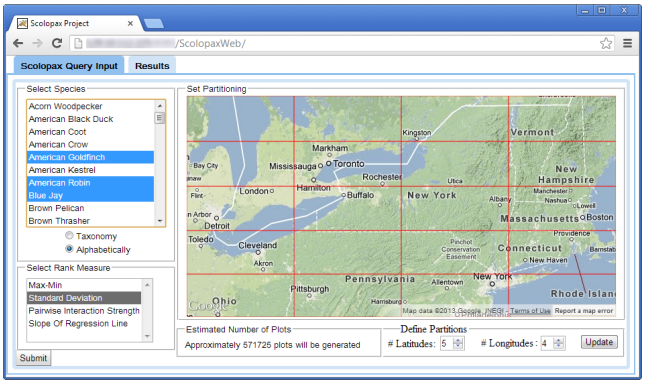


Figure 3: The Scolopax summary generator input screen allows the user to select species of interest, explore areas of interest on the interactive map and create partitions on latitude and longitude.

is a user-defined distance function. A variety of other non-equality conditions are necessary for scenarios where scientists may investigate correlations among summaries with different summary attributes in various regions.

To compute this type of general theta-join, we recently proposed novel techniques that minimize job completion time in shared-nothing parallel architectures [3]. We introduced a join model that allows reasoning about the relationship between record assignments to different processing nodes, the resulting input or output distribution in the system, and its effect on execution time.

For cases where output-related costs dominate execution time, we propose a randomized algorithm that provably achieves a near-optimal job completion time. This algorithm requires minimal knowledge about the input—only the cardinality of both data sets—and works for any theta join, including those used by Scolopax. Unfortunately, due to its reliance on duplication of some input records, this randomized algorithm often is not efficient for joins where input-related costs dominate execution time. For those scenarios we presented techniques that work well for popular join predicates, including equi-join, inequality-join, and band-join. Using histograms of the frequency distribution of the join attributes, these techniques are highly effective in distributing load, even for skewed data. In Scolopax, we deployed the MapReduce implementations of these join algorithms.

3. DEMONSTRATION DESCRIPTION

We will demonstrate how Scolopax can be employed for discovery of interesting patterns in real data about the occurrence of different species of birds. The data set was introduced in Section 1 and is based on a subset of the eBird observations made by citizen scientists in the lower 48 states of the USA between 2000 and 2011 [2]. All scenarios described below will be executed using a 44-core cluster running Hadoop and located at Northeastern University.

The discovery process typically consists of a summary-generation phase, followed by several rounds of interactive result filtering and post-processing. Users will receive guidance to avoid wait times exceeding one minute for summary generation. (This still allows for sufficiently large and interesting summary collections to be produced on-the-fly.)

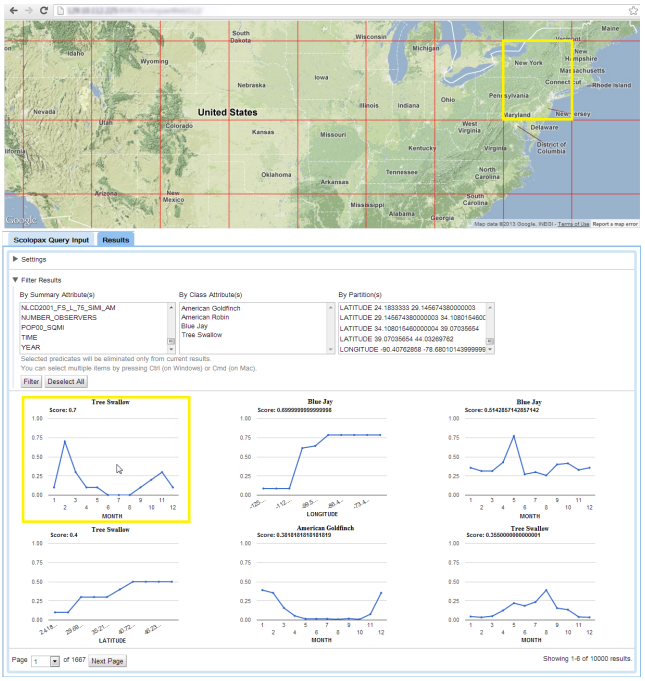


Figure 4: The Scolopax summary generator output screen allows the user to interactively examine the ranked list of summaries, see the corresponding region on the map, filter results for further investigation, and resubmit a refined query.

We will also pre-compute interesting summary collections to both speed up real-time processing and as a backup should network problems at the conference site prevent us from connecting to our cluster. A poster will highlight the Scolopax data management technology (see Section 2). Short wait times for summary generation present ideal opportunities for explaining these techniques using the poster.

Scolopax comes with two different interfaces: "Explorer" and "Performance Monitor", both discussed in the remainder of this section.

3.1 Explorer Interface

The Explorer interface allows conference attendees to make discoveries in a real data set that is actively studied by biologists and ecologists world-wide. It supports search for three broad types of hypotheses.

Important variables. Conference attendees will be able to search for interesting model summaries that may help identify threats to bird species. Users select any set of species and can interactively pick regions of interest by zooming and sub-partitioning on a map as shown in Figure 3. To support broad exploration, Scolopax will then generate all possible one-dimensional summaries (for each of the input data attributes) for each of the regions the user selected. The corresponding summaries are ranked by a user-selected rank measure. While Scolopax can support any function computable on a summary (including its metadata), for simplicity the demo only offers commonly used ones. "Max-Min" and "Standard Deviation" help identify important variables by ranking summaries showing greater changes in output value above those closer to "flat lines".

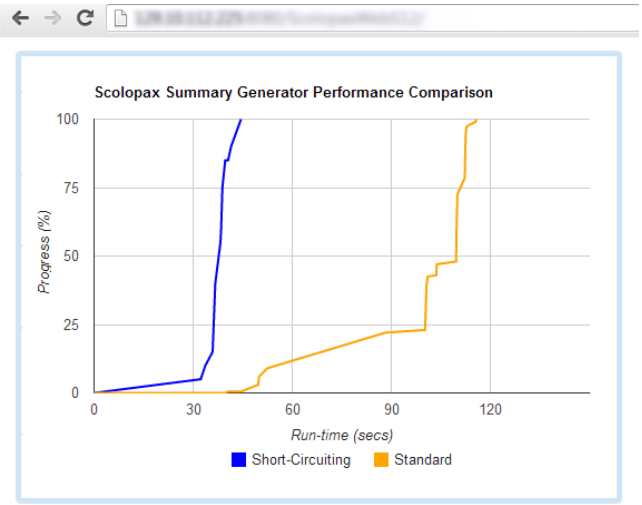


Figure 5: Comparison of Summary Generation Algorithms

Slope-based measures are used to identify summaries showing strong trends, e.g., declining species probability with increasing human presence.

After the query is submitted, Scolopax returns a ranked list of patterns that best match the users' interest as shown in Figure 4. Hovering over each plot reveals the region the summary belongs to on the map. Since the initial summary generation is intentionally broad to include many potentially interesting patterns, the result usually contains highly-ranked summaries that turn out to be of little interest. The user can then interactively filter results by selecting attributes she wants to further investigate or remove. At any point, the user can go back and modify the initial selections to execute another query. Scolopax automatically stores executed query results in order to be able to respond rapidly to previously executed queries.

Correlations. Users will be able to search for habitat competition and other correlations, e.g., variables with similar effects on a single species. Similar to the previous scenario, users first select species and region(s) of interest on the map. They also select on which data attributes and region properties an equality or inequality condition should be enforced. Furthermore, the user can specify if she is interested in correlation or anti-correlation and choose from a variety of data manipulations that should be applied before computing the correlation score, e.g., shifting or scaling. This allows a great variety of possible join conditions to be expressed. Similar to the previous scenario shown in Figure 4, the join results are presented ranked by strength of correlation and the user can filter and post-process them to home in on the most interesting ones.

Spatial patterns. Users can search for spatial patterns, including migration. Similar to the previous scenarios, a set of species and regions of interest are selected interactively. In addition, attributes of interest are selected, e.g., the day of the year for migration patterns or human population density for exploring regional effects on the association between bird occurrence and human population density. Scolopax computes the corresponding set of summaries and clusters them based on a similarity function, cluster algorithm, and clustering parameters also selected by the user. These clusters are visualized on the map together with the average cluster trajectories (see Figure 2). The user can also examine the individual summaries generated by hovering over the corresponding area on the map. If the user suspects an interesting migration pattern on the map, she can further explore the area by zooming in and re-partitioning the geographic space.

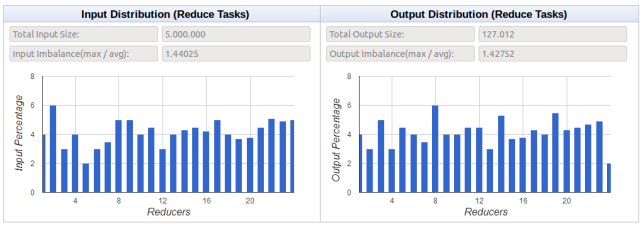


Figure 6: Workload Distribution of Join Queries

3.2 Performance Monitor Interface

The Performance Monitor interface allows the conference attendees to experience the effectiveness of our recently proposed data management techniques [3, 4]. For summary generation, Scolopax will execute our algorithm and the previous state-of-the-art technique side-by-side. The system continuously updates the progress of both methods as shown in Figure 5. While running a user-specified join for correlation search, Scolopax dynamically visualizes the real-time input/output workload distribution among reducer nodes as shown in Figure 6.

4. ACKNOWLEDGMENTS

We would like to thank Priyank Desai, Pratik Gawande, Shweta Memane, Mathi Ramakrishnan, Baturalp Torun, and Yeshwanth Venkatesh for their contributions to the Scolopax implementation; and the Cornell Lab of Ornithology team, in particular Steve Kelling, Wes Hochachka, and Kevin Webb, for their frequent feedback and advice.

This work was supported by the National Science Foundation under Grant Nos. IIS-1017793 and DRL-1010818. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additional support came from the Leon Levy Foundation.

5. REFERENCES

[1] S. Kelling, W. M. Hochachka, D. Fink, M. Riedewald, R. Caruana, G. Ballard, and G. Hooker. Data intensive science: A new paradigm for biodiversity studies. *Bioscience*, 57(7):613–620, 2009.

[2] A. M. Munson, K. Webb, D. Sheldon, D. Fink, W. M. Hochachka, M. Iliff, M. Riedewald, D. Sorokina, B. Sullivan, C. Wood, and S. Kelling. The ebird reference dataset, version 4.0. *Cornell Lab of Ornithology and National Audubon Society*, Ithaca, NY, January 2012.

[3] A. Okcan and M. Riedewald. Processing theta-joins using mapreduce. In *SIGMOD*, pages 949–960, 2011.

[4] B. Panda, M. Riedewald, and D. Fink. The model-summary problem and a solution for trees. In *ICDE*, pages 449–460, 2010.