Data-intensive approaches to digitized museum collections

Rebecca B. Dikow
Data Science Lab
Office of the CIO
Smithsonian Institution

datascience.si.edu
@SIDataScience
A diversity of locations

Heterogeneous digital data

Lack of purpose-built software tools
19 museums, 9 research centers, and a zoo
Smithsonian Collections Holdings

155.5M Objects and Specimens
163.3K Archival Cubic Feet
2.2M Library Volumes

Smithsonian Collections Digitization

32M Objects and Specimens with Digital Record
125K Archival Cubic Feet with Digital Record
1.5M Library Volumes with Digital Record
4.9M Objects and Specimens with Digital Image
56.9K Library Volumes with Digital Image
Digitized collections

photos
taxonomic names
specimen records
genomic sequences
geo-referenced localities

field books
illustrations
observations
scientific publications
taxonomic descriptions
Specimen collecting event → digital data
Research questions add complexity

DATA MANAGEMENT
Repository of all the de novo genome assemblies and projects

DATA VISUALIZATION
Institution and individual project dashboards, genome browsers, summary statistics, and other useful graphs for publication

DATA ANALYSIS
User-friendly web interface for data analysis, with workflows/pipelines that are easy to create and share.

OUTREACH
- Facilitates collaboration and data sharing among researchers
- Creates learning opportunities for students and researchers
- Promotes diffusion of knowledge to the public in general

Assembly base composition
GC (41.6%)
AT (54.6%)

Longest scaffold (15.4 MB)
length (3.6 MB)
length (685.5 kB)

g statistics
GB MB
Adding a research component adds complexity

depth learning

phylogeny

genomics

DATA MANAGEMENT
Repository of all the de novo genome assemblies and projects

DATA VISUALIZATION
Institution and individual project dashboards, genome browsers, summary statistics, and other useful graphs for publication

DATA ANALYSIS
User-friendly web interface for data analysis, with workflows/pipelines that are easy to create and share.

OUTREACH
- Facilitates collaboration and data sharing among researchers
- Creates learning opportunities for students and researchers
- Promotes diffusion of knowledge to the public in general

Assembly base composition
- GC (41.6%)
- AT (54.6%)

Longest scaffold
- (15.4 MB)
- (3.6 MB)
- (685.5 kB)

g statistics
- GB
- MB

60
80
0%
10
20
30
40
50
60
70
80
90
100 kB
1 M

Fragment (2.1%)

NCBI

GBIF

Deep Valley, California
Smithsonian Biodiversity Genome Hub: under construction
DATA MANAGEMENT
Repository of all the de novo genome assemblies and projects

DATA VISUALIZATION
Institution and individual project dashboards, genome browsers, summary statistics, and other useful graphs for publication

DATA ANALYSIS
User-friendly web interface for data analysis, with workflows/pipelines that are easy to create and share.

OUTREACH
- Facilitates collaboration and data sharing among researchers
- Creates learning opportunities for students and researchers
- Promotes diffusion of knowledge to the public in general

Assembly base composition
- GC (41.6%)
- AT (54.6%)

Gene statistics
- Longest scaffold (15.4 MB)
- Length (3.6 MB)
- Length (685.5 kB)
- Frags (2.1%)
- (
- 1)

Galaxy Project
First pilot projects: detecting mercury staining and family ID
**Mercury staining**

- **Predicted**
  - Unstained: 882
  - Stained: 46

- **Actual**
  - Stained: 77
  - Unstained: 682

---

**Family ID**

- **Predicted**
  - Clubmoss: 858
  - Spikemoss: 59

- **Actual**
  - Spikemoss: 23
  - Clubmoss: 901
Bumblebees
Crowdsourced transcription

The Bumblebee Project - Set 2

About the Project

Please help us create digital records for the United States National Entomological Collection! We will be transcribing the labels of specimens of bumblebees. Bumblebees are found in the Bombus genus (Hymenoptera: Apidae). They are social insects that feed on nectar and collect pollen to feed their young. Bumblebees are very important pollinators! Learn how to transcribe this project and get started.

The digitization of this project has been made possible with the generous support of Pixel Acuity, LLC. Please contact Jessica Bird (birdj@si.edu), Department of Entomology, for any questions or comments about the transcriptions and thanks to all of you for your help!

Read More
Models in progress

Training data: 33,347 images with subgenus/species labels
We also have images for >10,000 unidentified specimens

Subgenus model:
15 classes
Overall accuracy: 93.8%

Species model:
178 classes
Overall accuracy: 92.5%

The backbone is a 101 layer deep residual convolutional network (ResNet-101; He et al., 2015)
Activation heat maps allow us to start exploring the models in more detail.
How can we broaden this work to Smithsonian history, art, and culture digital collections and archives?

Morse Daguerreotype Camera

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound Box</td>
<td>86%</td>
</tr>
<tr>
<td>Audio Equipment</td>
<td>79%</td>
</tr>
<tr>
<td>Electronics</td>
<td>79%</td>
</tr>
<tr>
<td>Loudspeaker</td>
<td>75%</td>
</tr>
<tr>
<td>Wood</td>
<td>74%</td>
</tr>
<tr>
<td>Electronic Instrument</td>
<td>63%</td>
</tr>
<tr>
<td>Computer Speaker</td>
<td>59%</td>
</tr>
<tr>
<td>Antique</td>
<td>58%</td>
</tr>
<tr>
<td>Subwoofer</td>
<td>56%</td>
</tr>
</tbody>
</table>
United States Holocaust Memorial Museum

Because of HER Story

Smithsonian American Women’s History Initiative

womenshistory.si.edu
Building capacity across the Smithsonian includes lots of training!

We are building a community of Carpentries instructors across the Smithsonian.

More than 300 Smithsonian researchers have been trained in topics such as Python, R, genome analysis, and data management in the past 3 years.

Workshop materials: github.com/SmithsonianWorkshops

Instructors and schedule: datascience.si.edu/carpentries
We’ve only just begun!

Right now, every application of machine learning tools is a research project given our diverse, unique, incomplete data.

Let’s use these tools to elevate new stories that better represent Smithsonian audiences.
Thank you!

Data Science Lab:
Mike Trizna            Alex Robillard
Mirian Tsuchiya        Maddy Bursell
Alex White             Alejandro Sanchez

Partners:
NMNH Botany
NMNH Entomology
Smithsonian Conservation Biology Institute
OCIO DPO
OCIO DAMS
Smithsonian Institution Archives
American Women’s History Initiative
United States Holocaust Memorial Museum

Funding:
Smithsonian Women’s Committee
Smithsonian Office of the Provost
Smithsonian Office of the CIO