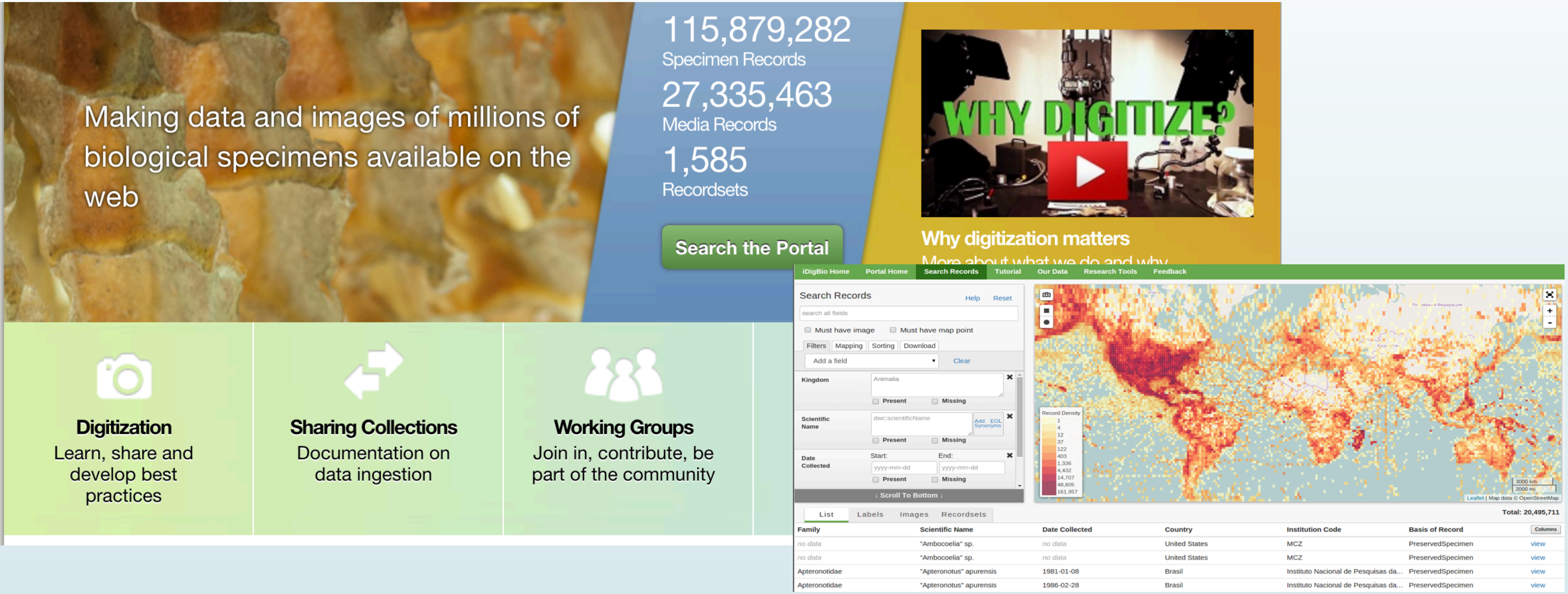
A Pipeline for Processing Images in iDigBio to Train Convolutional Neural Networks

Gaurav Yeole, Matthew Collins, Alex Thompson, Renato Figueiredo

iDigBio
Integrated Digitized Biocollections

ACIS

iDigBio data and portal



Making data and images of millions of biological specimens available on the web

115,879,282 Specimen Records

27,335,463 Media Records

1,585 Recordsets

Why digitization matters

More about what we do and why

Digitization
Learn, share and develop best practices

Sharing Collections
Documentation on data ingestion

Working Groups
Join in, contribute, be part of the community

Search Records

□ Must have image □ Must have map point

Kingdom Animalia
□ Present □ Missing

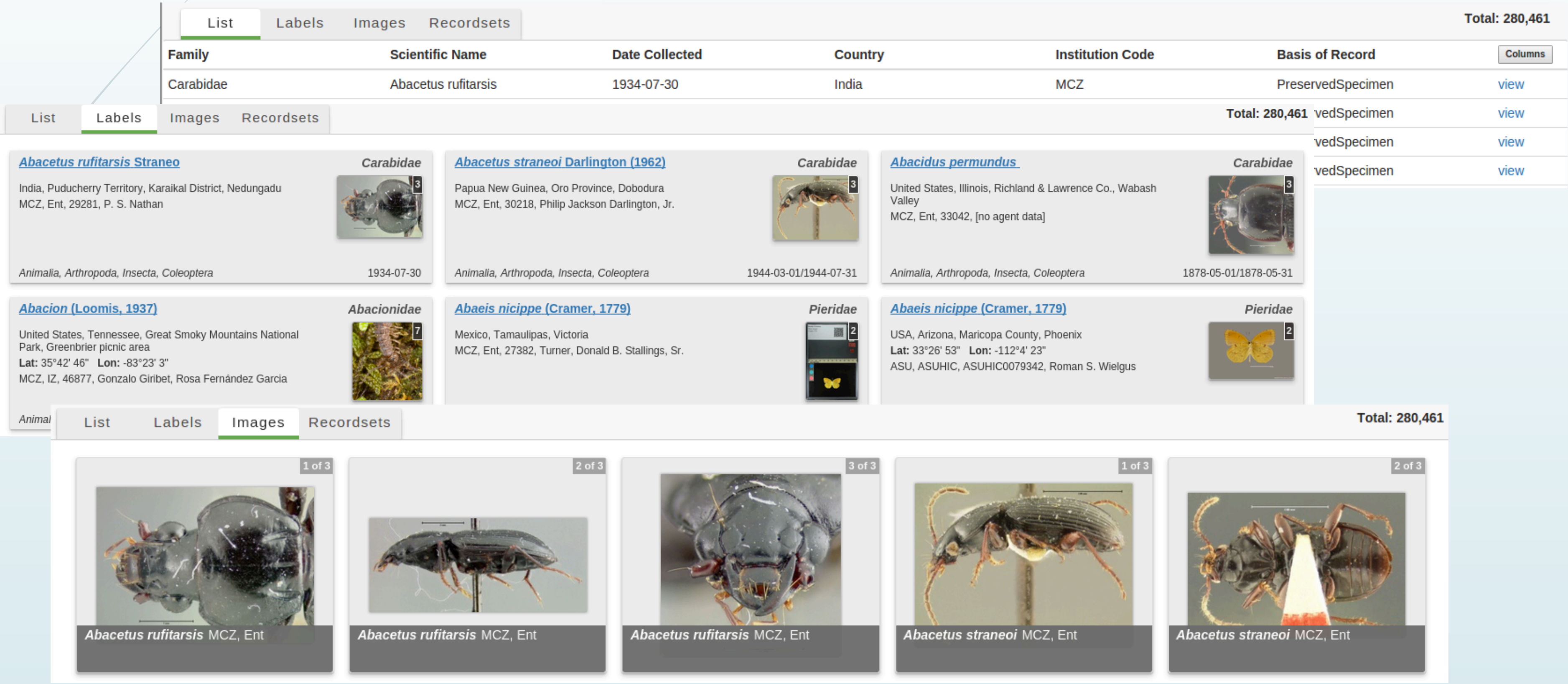
Scientific Name dwc:scientificName
□ Present □ Missing

Date Collected Start: yyyy-mm-dd End: yyyy-mm-dd
□ Present □ Missing

Total: 20,495,711

Family	Scientific Name	Date Collected	Country	Institution Code	Basis of Record	
no data	"Ambocoelia" sp.	no data	United States	MCZ	PreservedSpecimen	view
no data	"Ambocoelia" sp.	no data	United States	MCZ	PreservedSpecimen	view
Apteronotidae	"Apteronotus" apurensis	1981-01-08	Brasil	Instituto Nacional de Pesquisas da...	PreservedSpecimen	view
Apteronotidae	"Apteronotus" apurensis	1986-02-28	Brasil	Instituto Nacional de Pesquisas da...	PreservedSpecimen	view

27M+ images stored



List | Labels | Images | Recordsets

Total: 280,461

Family	Scientific Name	Date Collected	Country	Institution Code	Basis of Record	Columns
Carabidae	Abacetus rufitarsis	1934-07-30	India	MCZ	PreservedSpecimen	view
					vedSpecimen	view
					vedSpecimen	view
					vedSpecimen	view

List | Labels | Images | Recordsets

Total: 280,461

***Abacetus rufitarsis* Straneo** — *Carabidae*
India, Puducherry Territory, Karaikal District, Nedungadu
MCZ, Ent, 29281, P. S. Nathan
Animalia, Arthropoda, Insecta, Coleoptera 1934-07-30

***Abacetus straneoi* Darlington (1962)** — *Carabidae*
Papua New Guinea, Oro Province, Dobodura
MCZ, Ent, 30218, Philip Jackson Darlington, Jr.
Animalia, Arthropoda, Insecta, Coleoptera 1944-03-01/1944-07-31

Abacidus permundus — *Carabidae*
United States, Illinois, Richland & Lawrence Co., Wabash Valley
MCZ, Ent, 33042, [no agent data]
Animalia, Arthropoda, Insecta, Coleoptera 1878-05-01/1878-05-31

***Abacion* (Loomis, 1937)** — *Abacionidae*
United States, Tennessee, Great Smoky Mountains National Park, Greenbrier picnic area
Lat: 35°42' 46" **Lon:** -83°23' 3"
MCZ, IZ, 46877, Gonzalo Giribet, Rosa Fernández Garcia
Animal

***Abaeis nicippe* (Cramer, 1779)** — *Pieridae*
Mexico, Tamaulipas, Victoria
MCZ, Ent, 27382, Turner, Donald B. Stallings, Sr.

***Abaeis nicippe* (Cramer, 1779)** — *Pieridae*
USA, Arizona, Maricopa County, Phoenix
Lat: 33°26' 53" **Lon:** -112°4' 23"
ASU, ASUHIC, ASUHIC0079342, Roman S. Wielgus

List | Labels | Images | Recordsets

Total: 280,461

1 of 3 — ***Abacetus rufitarsis*** MCZ, Ent

2 of 3 — ***Abacetus rufitarsis*** MCZ, Ent

3 of 3 — ***Abacetus rufitarsis*** MCZ, Ent

1 of 3 — ***Abacetus straneoi*** MCZ, Ent

2 of 3 — ***Abacetus straneoi*** MCZ, Ent

Wealth of data; information largely untapped
Metadata available; image processing – human eye+brain

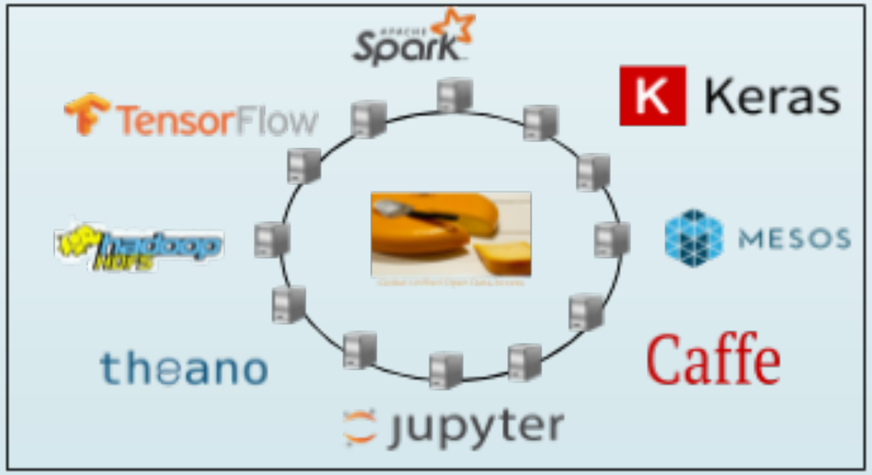
iDigBio storage back-end

- 115+ million specimen records
- 27+ million media records
- 120 terabytes of media files
- Hosted on Ceph Ditributed storage system
- Opportunity: machine learning processing near data

iDigBio Database of Images

GUODA – Global Unified Open Data Access

- Apache Mesos – distributed systems kernel
- Apache Spark – fast and general engine for large-scale data processing
- Jupyter Notebook Interface for Python and R
- Installed frameworks for deep learning –
  - TensorFlow
  - Keras
  - Theono
  - Caffe



GUODA cluster

Applications of Artificial Intelligence (AI)

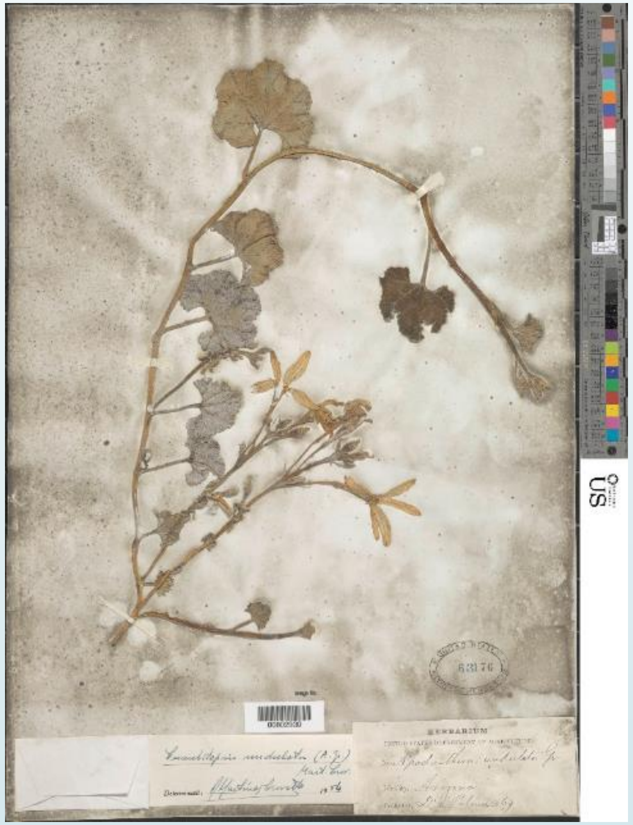
- Decision-making
- System Identification and Control
  - Vehicle control, process control, natural resource management
- Pattern Recognition
  - Face identification, object recognition, etc.
- Classification
  - Image classification, e-mail spam filtering

- **Emerging applications of AI for specimen images?**

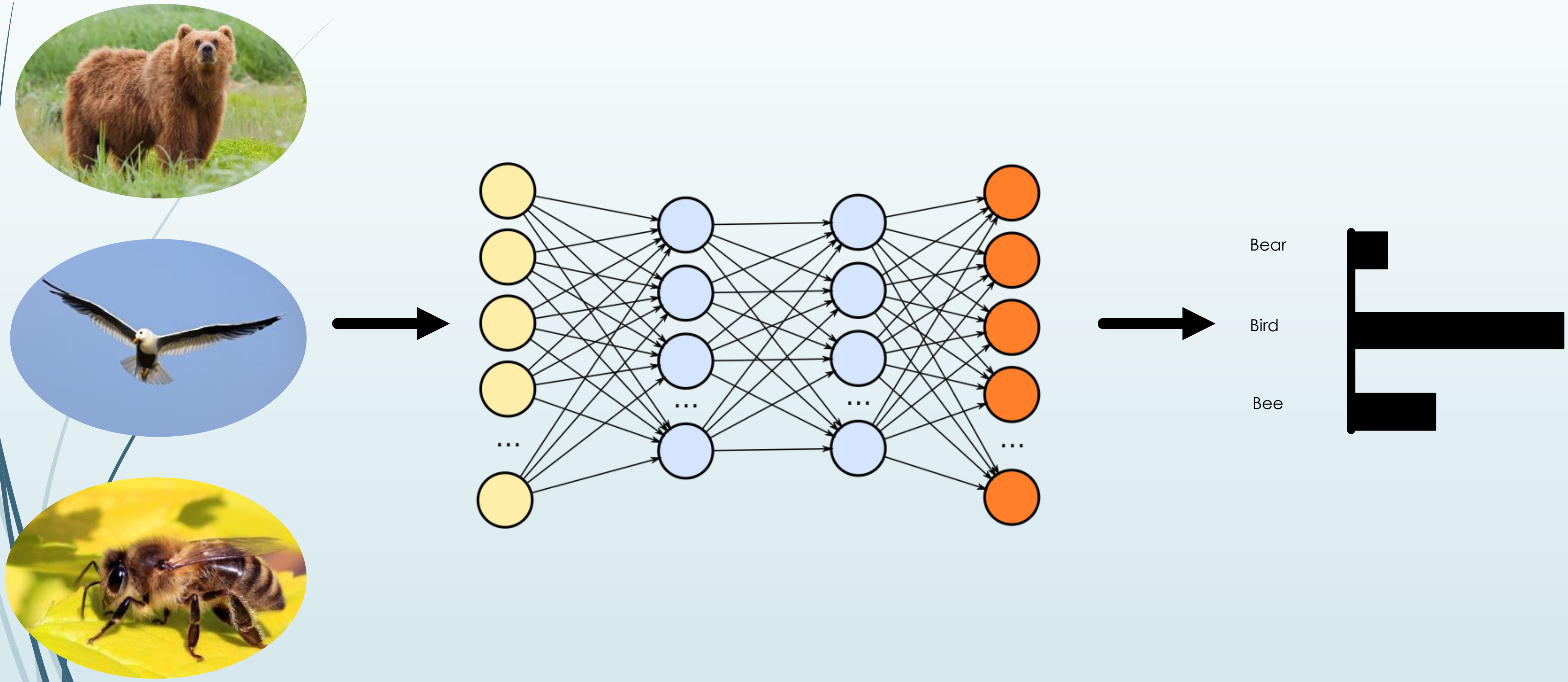
Use Case – Find Mercury Contamination on Herbarium Specimens

- Mercury salt used to specimens as insecticides
- Mercury marking is visible on the specimens
- Mercury vapor and mercury compounds pose threat to human health.
- Mercury contaminated images can be classified using convolutional neural networks.

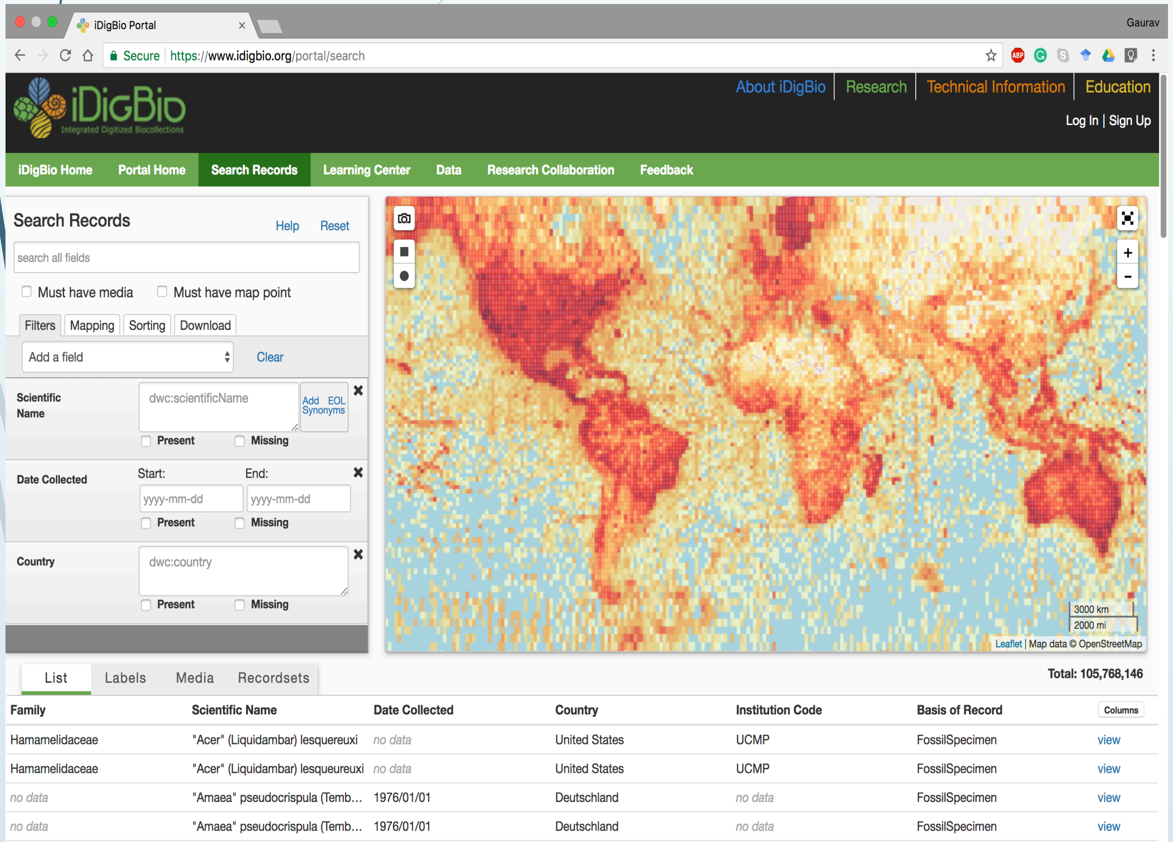
- E. Schuettpelz, P. Frandsen, R. Dikow, A. Brown, S. Orli, M. Peters, A. Metallo, V. Funk, L. Dorr. (2017). Applications of deep convolutional neural networks to digitized natural history collections. Biodiversity Data Journal. 5. e21139. 10.3897/BDJ.5.e21139.



Convolutional Neural Networks (CNN)

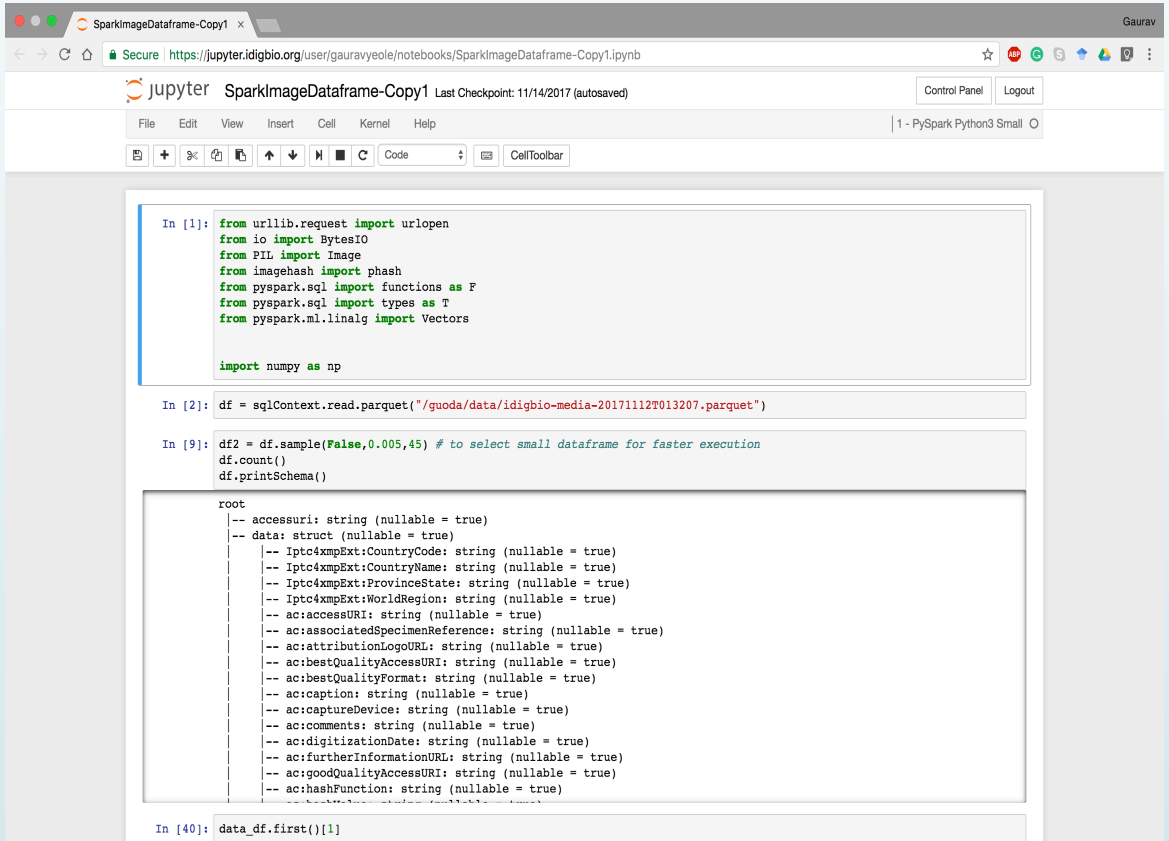


Accessing iDigBio Images



Total: 105,768,146

Family	Scientific Name	Date Collected	Country	Institution Code	Basis of Record	
Hamamelidaceae	"Acer" (Liquidambar) lesquereuxi	*no data*	United States	UCMP	FossilSpecimen	view
Hamamelidaceae	"Acer" (Liquidambar) lesqueureuxi	*no data*	United States	UCMP	FossilSpecimen	view
no data	"Amaea" pseudocrispula (Temb...	1976/01/01	Deutschland	*no data*	FossilSpecimen	view
no data	"Amaea" pseudocrispula (Temb...	1976/01/01	Deutschland	*no data*	FossilSpecimen	view

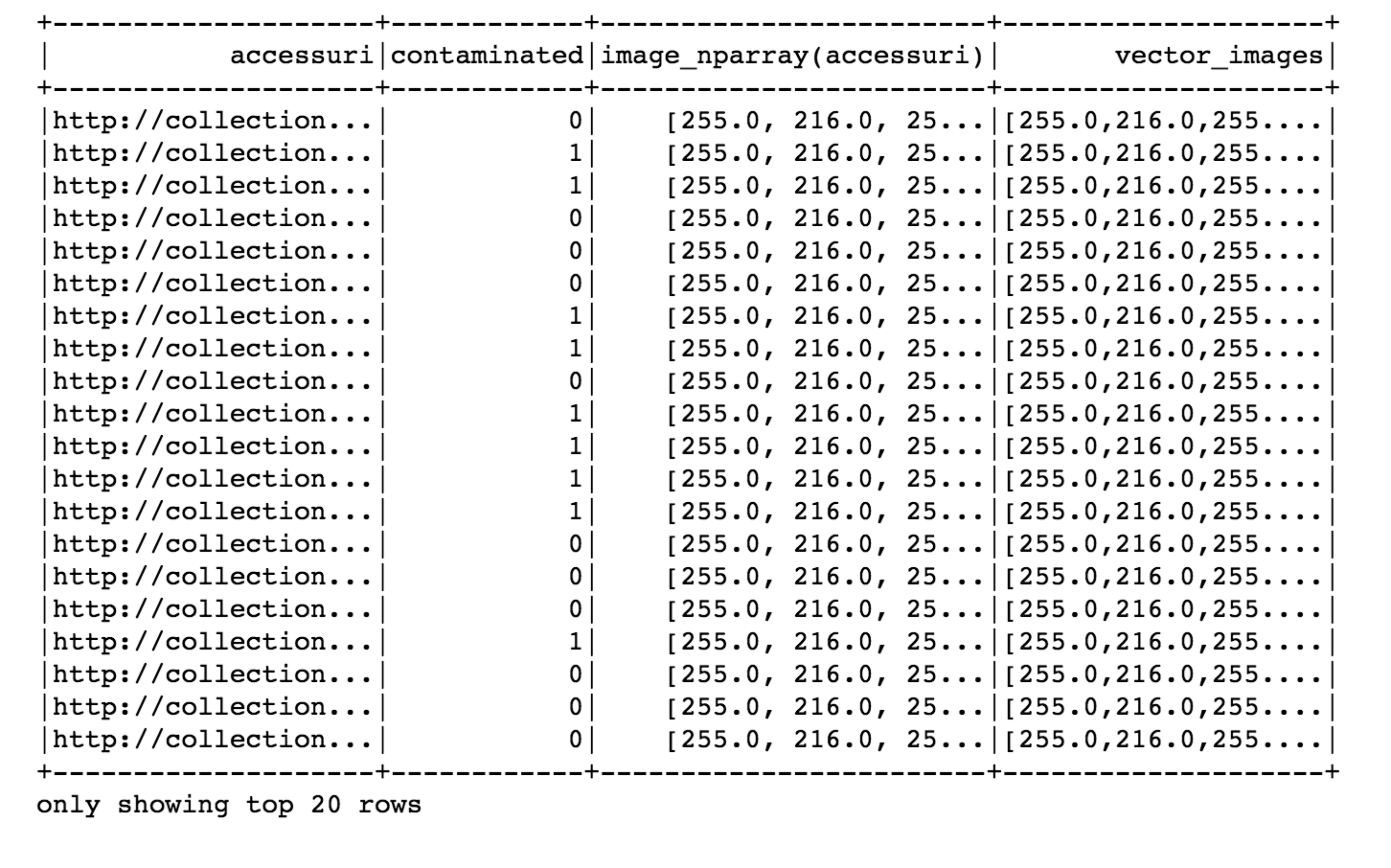

SparkImageDataframe-Copy1
Gaurav
Secure https://jupyter.idigbio.org/user/gauravyeole/notebooks/SparkImageDataframe-Copy1.ipynb
jupyter
SparkImageDataframe-Copy1 Last Checkpoint: 11/14/2017 (autosaved)
Control Panel
Logout
File
Edit
View
Insert
Cell
Kernel
Help
1 - PySpark Python3 Small
Code
CellToolbar

```
In [1]: from urllib.request import urlopen
        from io import BytesIO
        from PIL import Image
        from imagehash import phash
        from pyspark.sql import functions as F
        from pyspark.sql import types as T
        from pyspark.ml.linalg import Vectors


        import numpy as np

In [2]: df = sqlContext.read.parquet("/guoda/data/idigbio-media-20171112T013207.parquet")

In [9]: df2 = df.sample(False,0.005,45) # to select small dataframe for faster execution
        df.count()
        df.printSchema()
```

```
root
 |-- accessuri: string (nullable = true)
 |-- data: struct (nullable = true)
 |    |-- Iptc4xmpExt:CountryCode: string (nullable = true)
 |    |-- Iptc4xmpExt:CountryName: string (nullable = true)
 |    |-- Iptc4xmpExt:ProvinceState: string (nullable = true)
 |    |-- Iptc4xmpExt:WorldRegion: string (nullable = true)
 |    |-- ac:accessURI: string (nullable = true)
 |    |-- ac:associatedSpecimenReference: string (nullable = true)
 |    |-- ac:attributionLogoURL: string (nullable = true)
 |    |-- ac:bestQualityAccessURI: string (nullable = true)
 |    |-- ac:bestQualityFormat: string (nullable = true)
 |    |-- ac:caption: string (nullable = true)
 |    |-- ac:captureDevice: string (nullable = true)
 |    |-- ac:comments: string (nullable = true)
 |    |-- ac:digitizationDate: string (nullable = true)
 |    |-- ac:furtherInformationURL: string (nullable = true)
 |    |-- ac:goodQualityAccessURI: string (nullable = true)
 |    |-- ac:hashFunction: string (nullable = true)
```

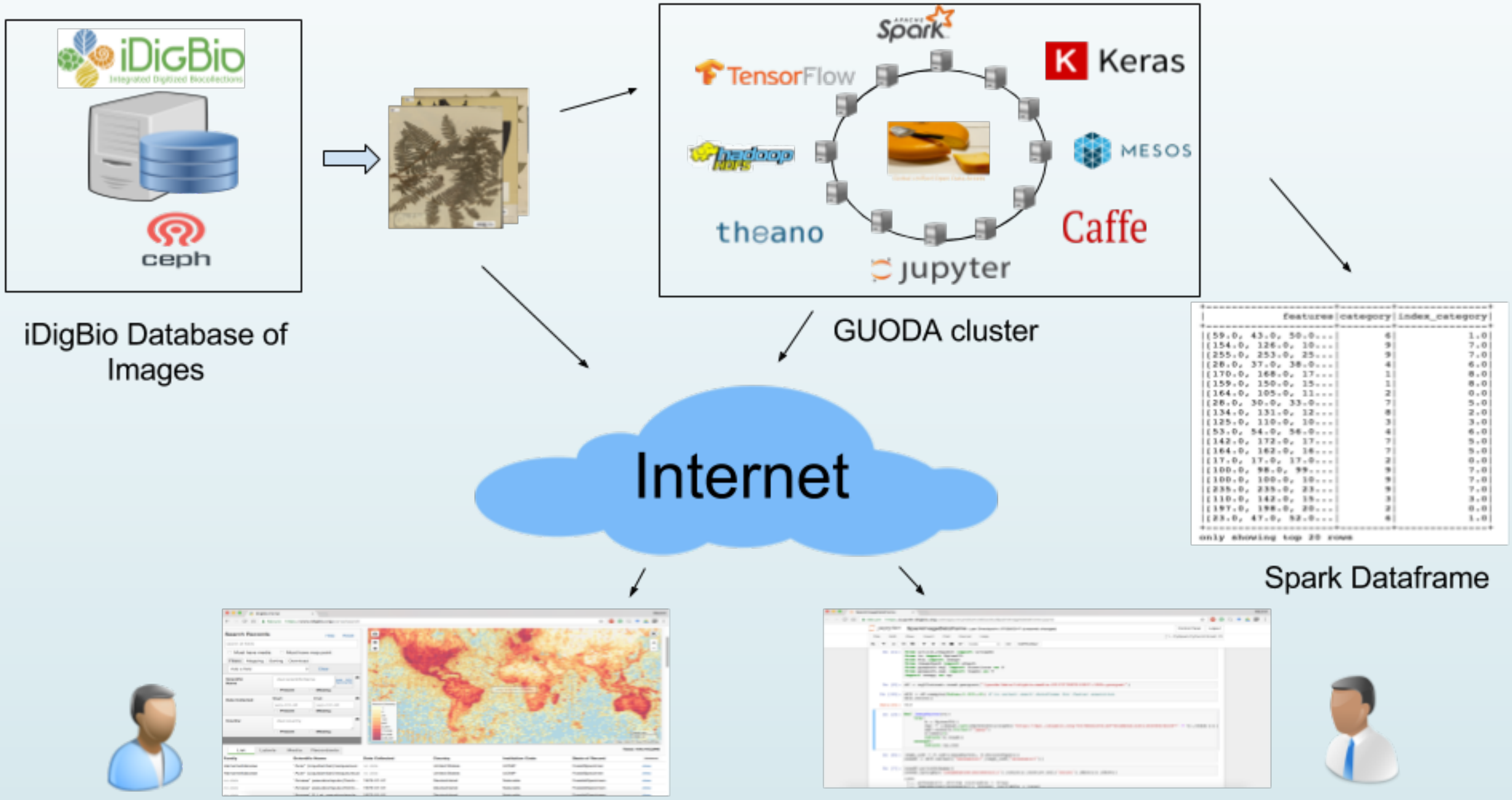
```
In [40]: data_df.first()[1]
```

Spark dataframe of iDigBio media



```
+--------------------+------------+------------------------+--------------------+
|           accessuri|contaminated|image_nparray(accessuri)|       vector_images|
+--------------------+------------+------------------------+--------------------+
|http://collection...|           0|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           1|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           1|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           0|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           0|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           0|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           1|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           1|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           0|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           1|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           1|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           1|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           1|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           0|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           0|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           0|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           1|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           0|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           0|    [255.0, 216.0, 25...|[255.0,216.0,255....|
|http://collection...|           0|    [255.0, 216.0, 25...|[255.0,216.0,255....|
+--------------------+------------+------------------------+--------------------+
only showing top 20 rows
```

Image Processing Pipeline



iDigBio Database of Images

GUODA cluster

Internet

```
+--------------------+--------+--------------+
|            features|category|index_category|
+--------------------+--------+--------------+
|[59.0, 43.0, 50.0...|       6|           1.0|
|[154.0, 126.0, 10...|       9|           7.0|
|[255.0, 253.0, 25...|       9|           7.0|
|[28.0, 37.0, 38.0...|       4|           6.0|
|[170.0, 168.0, 17...|       1|           8.0|
|[159.0, 150.0, 15...|       1|           8.0|
|[164.0, 105.0, 11...|       2|           0.0|
|[28.0, 30.0, 33.0...|       7|           5.0|
|[134.0, 131.0, 12...|       8|           2.0|
|[125.0, 110.0, 10...|       3|           3.0|
|[53.0, 54.0, 56.0...|       4|           6.0|
|[142.0, 172.0, 17...|       7|           5.0|
|[164.0, 162.0, 16...|       7|           5.0|
|[17.0, 17.0, 17.0...|       2|           0.0|
|[100.0, 98.0, 99....|       9|           7.0|
|[100.0, 100.0, 10...|       9|           7.0|
|[235.0, 235.0, 23...|       9|           7.0|
|[110.0, 142.0, 15...|       3|           3.0|
|[197.0, 198.0, 20...|       2|           0.0|
|[23.0, 47.0, 52.0...|       6|           1.0|
+--------------------+--------+--------------+
only showing top 20 rows
```

Spark Dataframe

iDigBio Portal

GUODA Jupyter Notebook

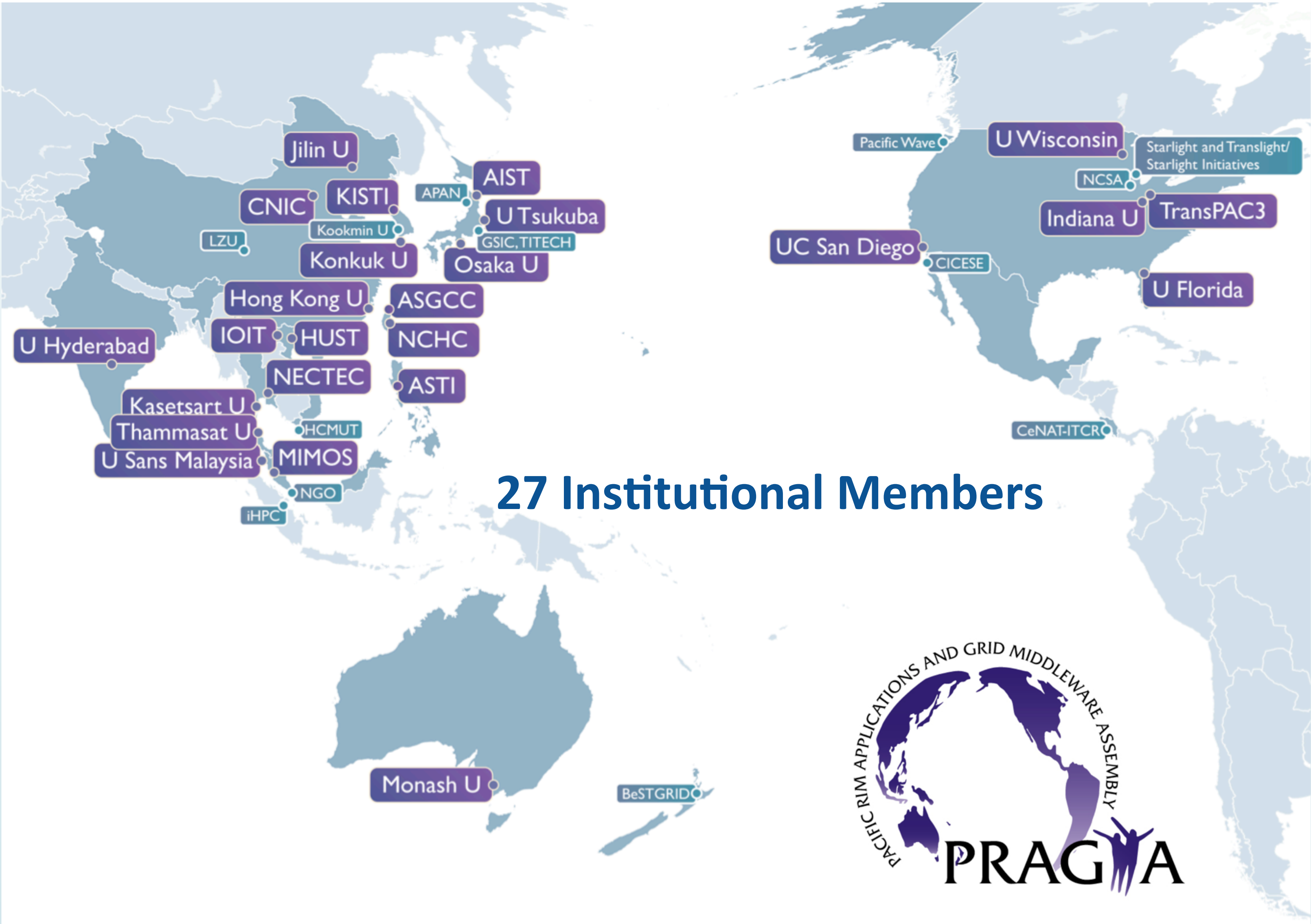
Opportunities for collaboration

Founded in 2002

Focused on researchers and institutions on the Pacific Rim

Open Community of Practice

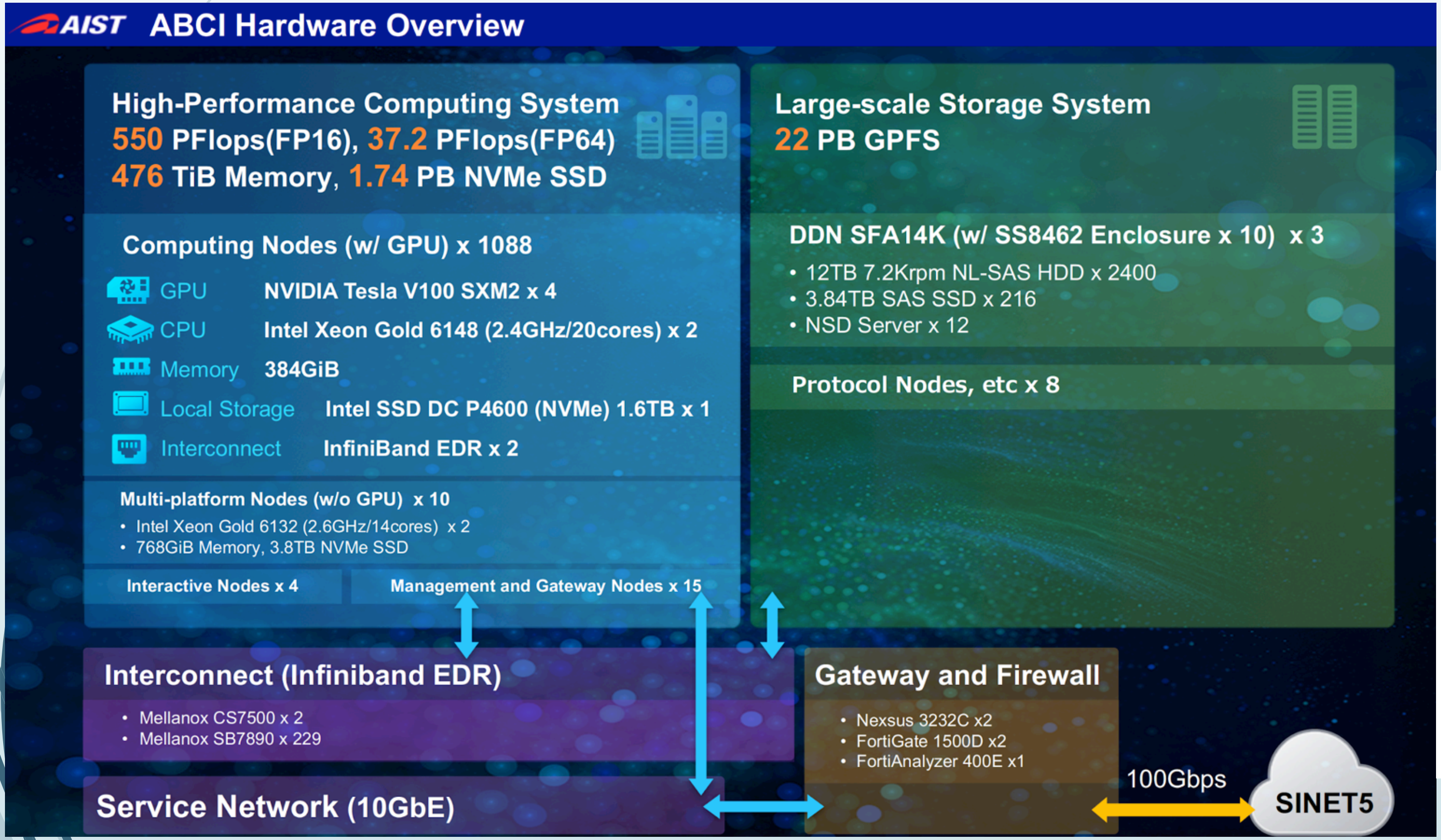
Engages "Long Tail" science communities



27 Institutional Members

AI-focused high-performance clusters

ABCI – top #5 in the world; similar system at NCHC, Taiwan



AIST ABCI Hardware Overview

High-Performance Computing System
550 PFlops(FP16), 37.2 PFlops(FP64)
476 TiB Memory, 1.74 PB NVMe SSD

Computing Nodes (w/ GPU) x 1088

GPU	NVIDIA Tesla V100 SXM2 x 4
CPU	Intel Xeon Gold 6148 (2.4GHz/20cores) x 2
Memory	384GiB
Local Storage	Intel SSD DC P4600 (NVMe) 1.6TB x 1
Interconnect	InfiniBand EDR x 2

Multi-platform Nodes (w/o GPU) x 10
- Intel Xeon Gold 6132 (2.6GHz/14cores) x 2
- 768GiB Memory, 3.8TB NVMe SSD

Interactive Nodes x 4

Management and Gateway Nodes x 15

Large-scale Storage System
22 PB GPFS

DDN SFA14K (w/ SS8462 Enclosure x 10) x 3
- 12TB 7.2Krpm NL-SAS HDD x 2400
- 3.84TB SAS SSD x 216
- NSD Server x 12

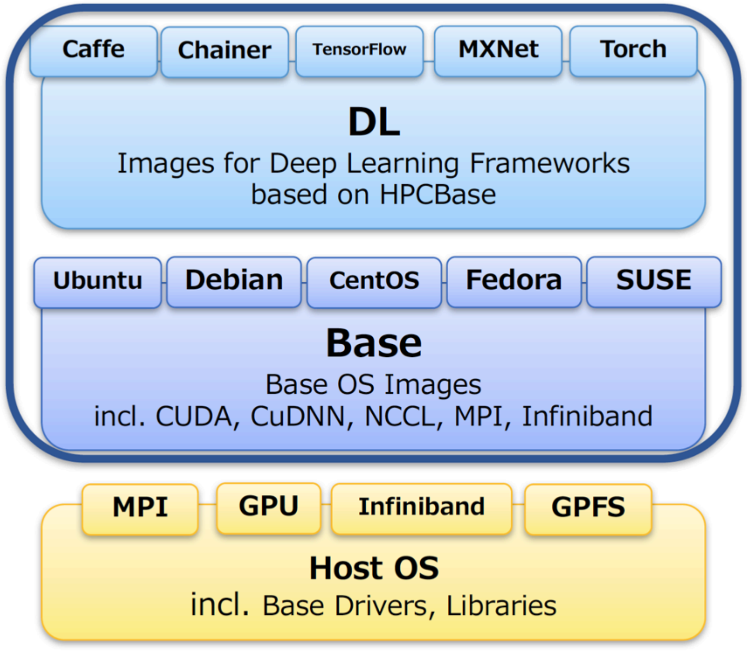
Protocol Nodes, etc x 8

Interconnect (Infiniband EDR)
- Mellanox CS7500 x 2
- Mellanox SB7890 x 229

Gateway and Firewall
- Nexsus 3232C x2
- FortiGate 1500D x2
- FortiAnalyzer 400E x1

Service Network (10GbE)

東京大学 THE UNIVERSITY OF TOKYO
AIST
Univ. Tokyo / Kashiwa II Campus


Caffe
Chainer
TensorFlow
MXNet
Torch
DL
Images for Deep Learning Frameworks based on HPCBase
Ubuntu
Debian
CentOS
Fedora
SUSE
Base
Base OS Images
incl. CUDA, CuDNN, NCCL, MPI, Infiniband
MPI
GPU
Infiniband
GPFS
Host OS
incl. Base Drivers, Libraries

Future Work

- Join specimen metadata from iDigBio with images in dataframes which can serve as image and label pairs to train more exciting deep learning models
- Build another audio signal processing pipeline for existing audio data in iDigBio
- Make this pipeline portable and make it compatible to run on infrastructures such as NSF's XSEDE, AIST's ABCI

Acknowledgements

- Sylvia Orli, Paul Frandsen, Rebecca Dikow, Smithsonian Institution

- iDigBio is funded by grants from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program [DBI-1115210 (2011-2018) and DBI-1547229 (2016-2021)]. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

iDigBio
Integrated Digitized Biocollections

ACIS