MAKING SMITHSONIAN OPEN ACCESS ACCESSIBLE WITH PYTHON AND DASK

Mike Trizna

Smithsonian OCIO Data Science Lab

May 4, 2021 | *csv,conf,v6*

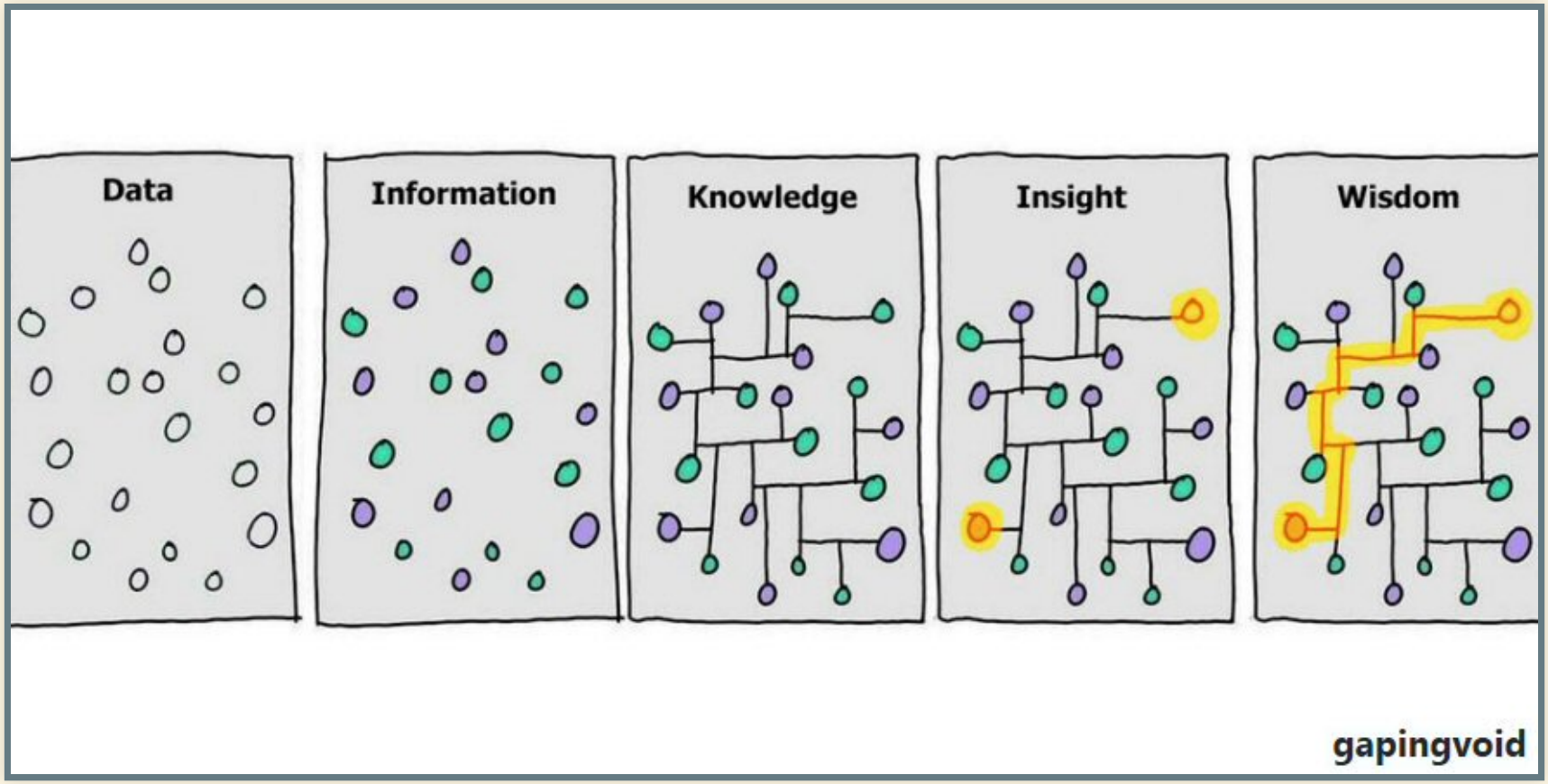
WHAT IS THE SMITHSONIAN INSTITUTION?

Yes, there are the museums (19 of them, mostly in Washington, DC), but we also have 21 libraries, 9 research centers ... and a zoo.

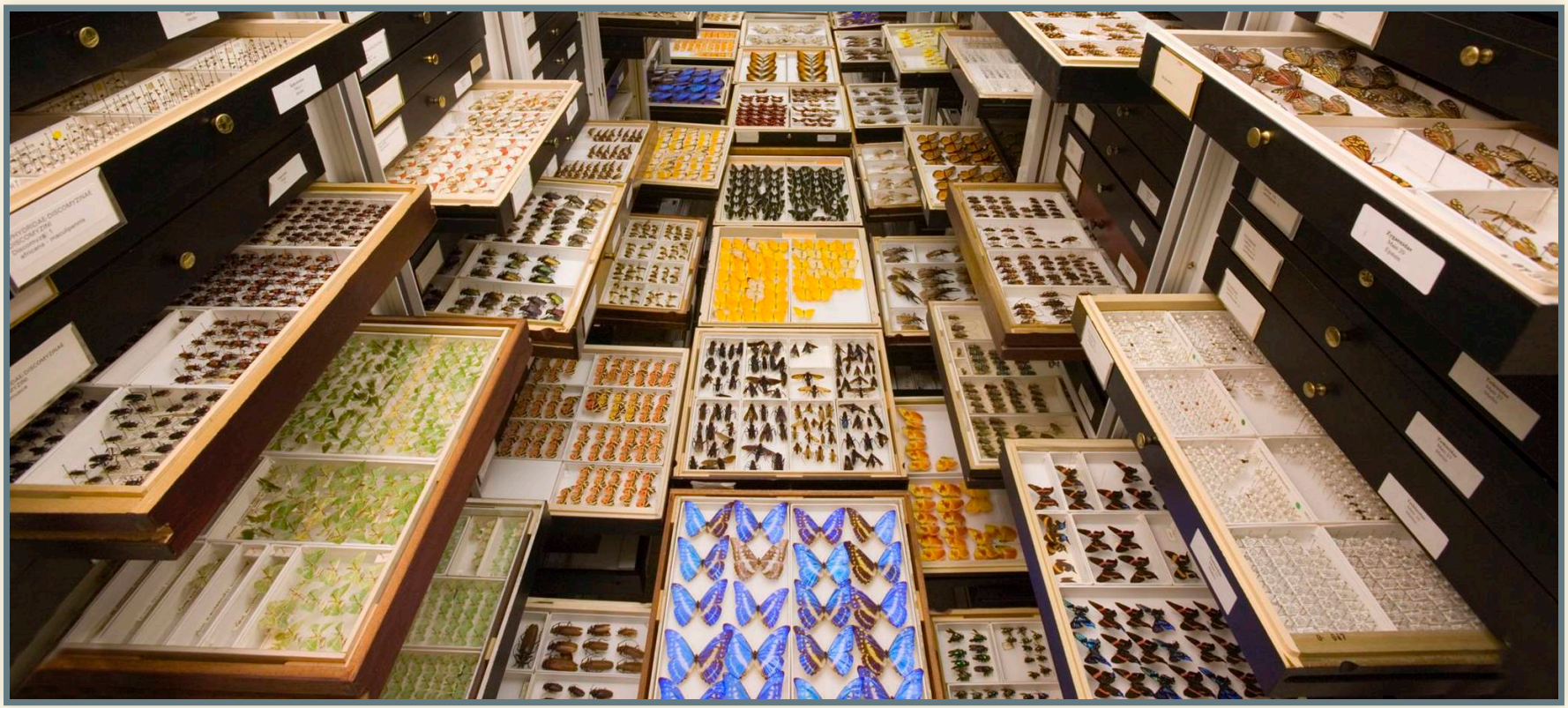
SMITHSONIAN MISSION

Founded in 1846 from the bequest of Englishman James Smithson with the condition:

> *"under the name of the Smithsonian Institution, an establishment for the* ***increase and diffusion of knowledge****."*



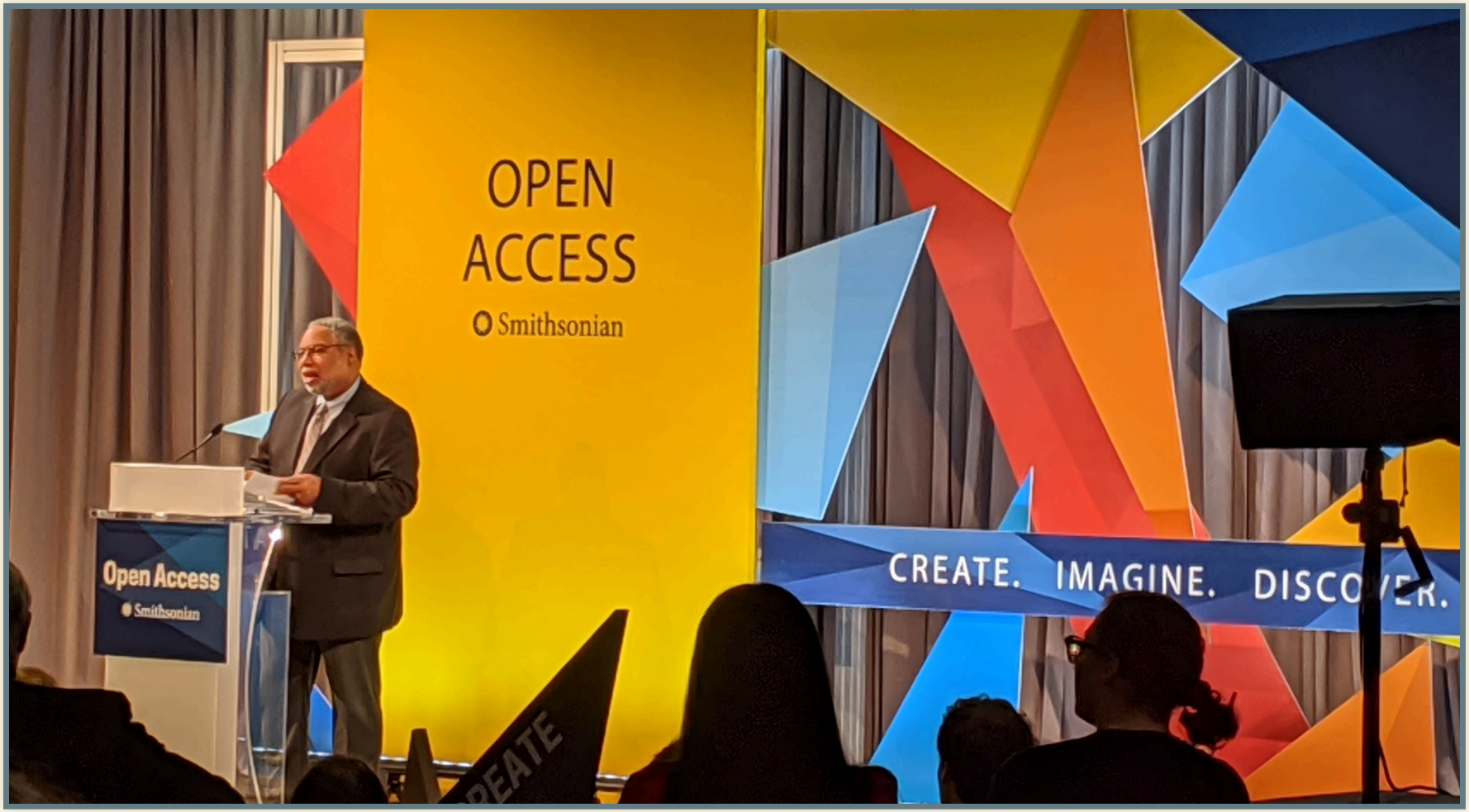
The Smithsonian has been increasing and diffusing *Knowledge* since 1846, but what about all of that *Data*?



All of that *data* and *info* that fed into *knowledge*, *insight*, *and wisdom* were dutifully cataloged and stored.

LAUNCH EVENT

February 2020


OPEN
ACCESS
Smithsonian
Open Access
Smithsonian
CREATE. IMAGINE. DISCO ER.

SI OPEN ACCESS RELEASE

Of the Smithsonian's 155 million objects, 2.1 million library volumes and 156,000 cubic feet of archival collections:

- 2.8 million 2-D and 3-D images
- X million collection metadata objects

TERMS OF USE

Before February 2020, all Smithsonian museums and units made their data searchable and sometimes able to download, but through each individual unit. Many different use agreements.

SI Open Access put all media and metadata in one place, and all Open Access media is CC0.

SO HOW CAN ALL OF THIS DATA BE ACCESSED?

I will cover 3 different ways.

All 3 share metadata records in the same deeply-nested JSON structure.

WEB API

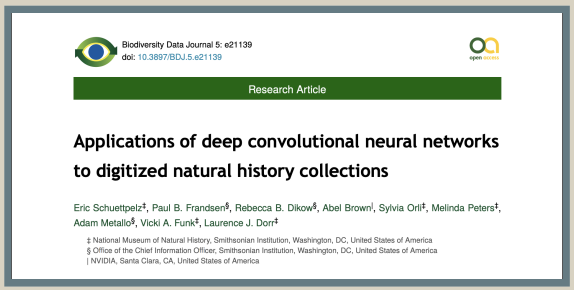
🔗: http://edan.si.edu/openaccess/apidocs/

- API Key needed (but free and painless to register)
- Great for getting a feel for record structure

HOWEVER YOU WILL QUICKLY RUN INTO LIMITATIONS

- Records are extensively indexed, but can only search indexed fields.
- Row limit of 1000 per API call

EXAMPLE: HERBARIUM SHEET MERCURY DETECTOR



Biodiversity Data Journal 5: e21139
doi: 10.3897/BDJ.5.e21139

Research Article

Applications of deep convolutional neural networks to digitized natural history collections

Eric Schuettpelz‡, Paul B. Frandsen§, Rebecca B. Dikow§, Abel Brown|, Sylvia Orli‡, Melinda Peters‡, Adam Metallo§, Vicki A. Funk‡, Laurence J. Dorr‡

‡ National Museum of Natural History, Smithsonian Institution, Washington, DC, United States of America
§ Office of the Chief Information Officer, Smithsonian Institution, Washington, DC, United States of America
| NVIDIA, Santa Clara, CA, United States of America

2017 paper that described building a machine learning model to detect herbarium sheets that had been stained with mercury.

https://doi.org/10.3897/bdj.5.e21139

EXAMPLE: HERBARIUM SHEET MERCURY DETECTOR

I wanted to create a new model on same dataset (2017 is ancient history in Machine Learning)

All training images are shared on Figshare, but photos are resized and I wanted original metadata

EXAMPLE: HERBARIUM SHEET MERCURY DETECTOR

[Screenshot of Barcode among other metadata]

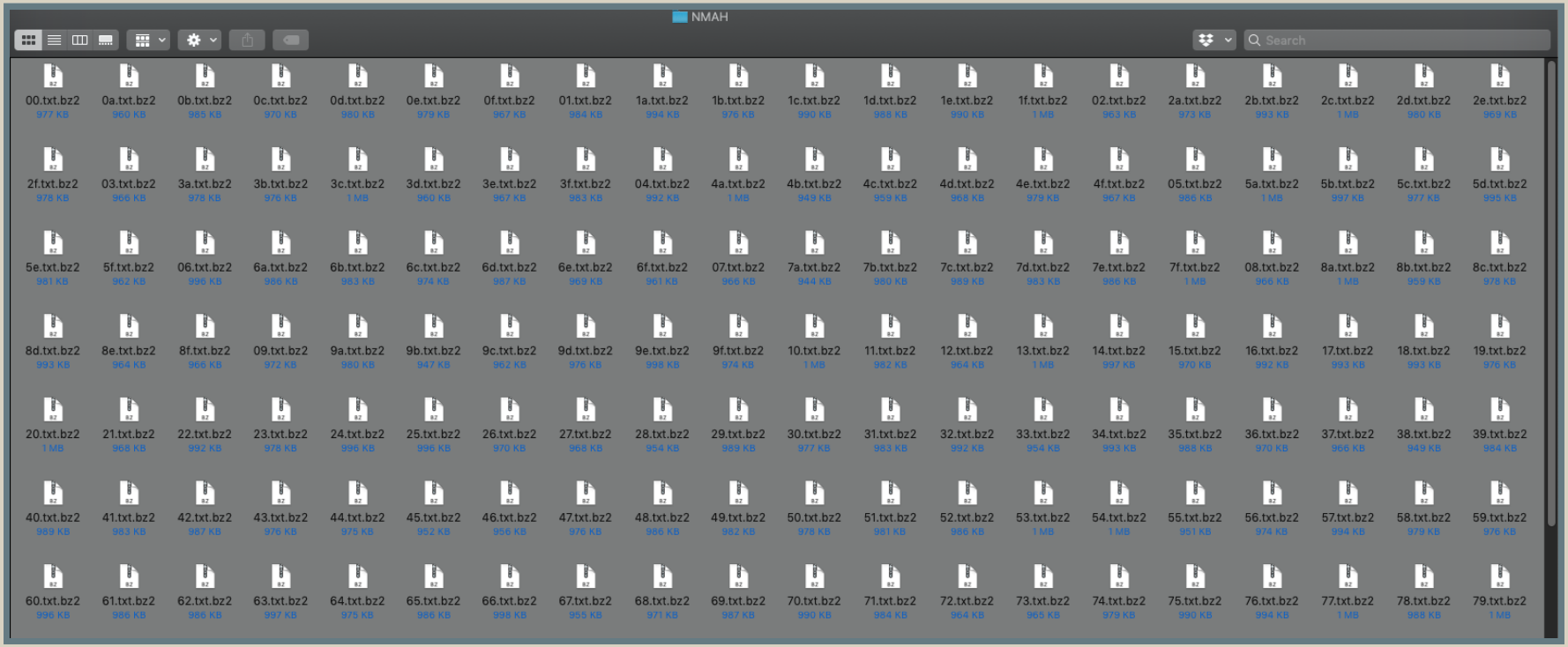
FULL DATASET SOURCES: GITHUB AND AWS

- AWS S3 for all metadata and images
- GitHub for versioned metadata

HOW IT'S PACKAGED

- Files are serialized as line-delimited JSON and compressed with bzip2.
- Directories are organized by owning unit and files are distributed by first two characters of content serialization hash.

HOW IT'S PACKAGED



BUMMER, WE NEED TO GO THROUGH EVERY SINGLE FILE ONE AT A TIME?

Example using the S3 interface

BUT THERE'S ACTUALLY A BENEFIT TO 256 FILES TO SEACH -- WE CAN MULTITASK

ENTER DASK

DASK LETS YOU SET UP A MINI CLUSTER ON YOUR MACHINE ... OR ON AN ACTUAL COMPUTE CLUSTER

WHAT IS DASK?

Dask is more well-known for parallel processing of DataFrames, but it also contains a really useful catch-all "Bag" type.

[[Show code for both AWS and GitHub]]

EXAMPLE: HERBARIUM SHEET MERCURY DETECTOR

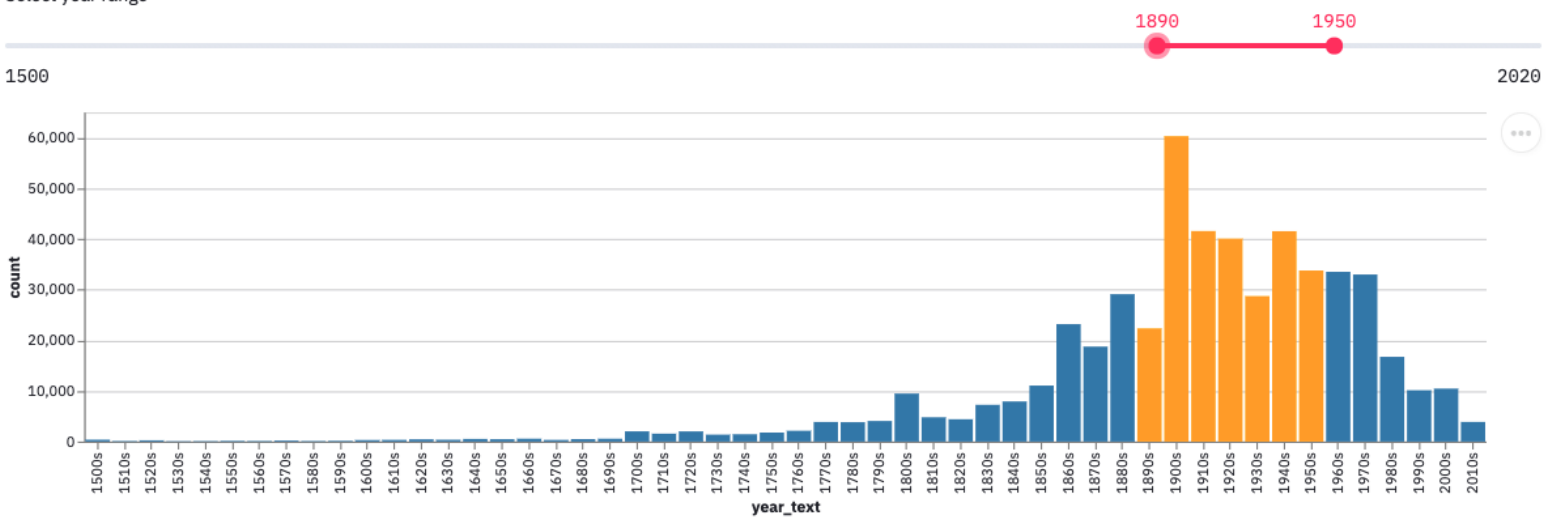
Creat

DIGITAL HUMANITIES INTERNSHIP PROJECT

PATRICK MCMANUS FROM GEORGE MASON UNIVERSITY

National Museum of American History Date Breakdown

Select year range



Top Object Types

	object_type	count
0	Certified Proof	35832
1	Money	35811
2	certified proof	35809
3	Exchange Medium	35808
4	Face	35808

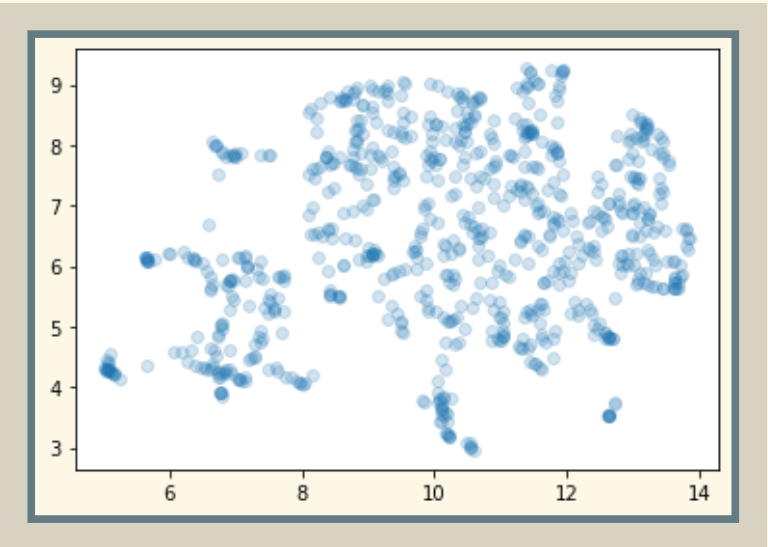
Top Places

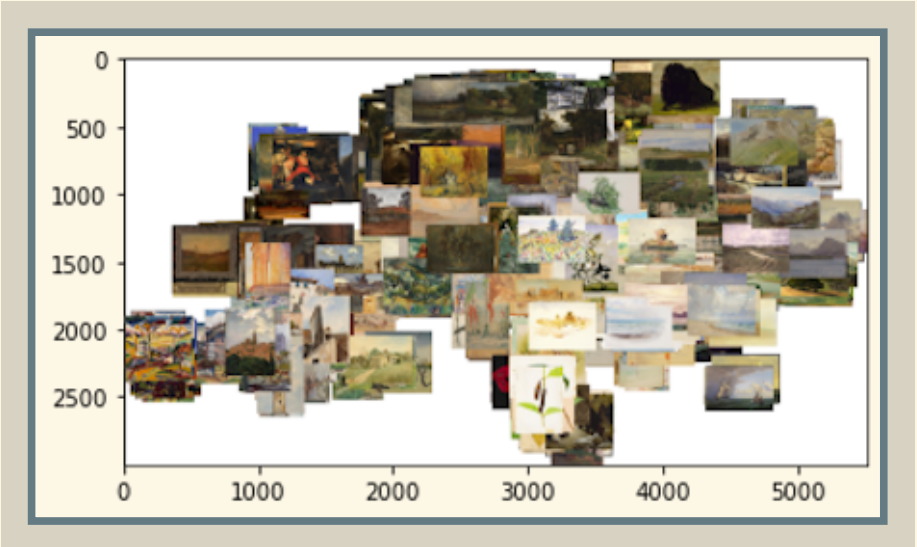
	place	count
0	United States	101491
1	New York	17772
2	Germany	10526
3	New York City	8516
4	New Jersey	7709

Top Topics

	topic	count
0	Work and Industry: Natio…	83039
1	Coins, Currency and Meda…	36109
2	Cultural and Community L…	28091
3	Political and Military H…	18210
4	Work and Industry: Photo…	11388

EXAMPLE: SEMANTIC CLUSTERING AMERICAN ART PAINTINGS





Full interactive notebook (through Binder) available at https://github.com/sidatasciencelab/siopenaccess.

QUESTIONS?