4-1-2021

Collections as Data at Florida International University

Jamie Rogers
Florida International University, rogersj@fiu.edu

Collections as Data

Research Possibilities @ FIU

Jamie Rogers
Assistant Director of Digital Collections

What is data?

indvidual facts, figures, signals, measurements, etc. that do not carry inherent meaning until organized, structured, categorized, and calculated.

How can digital collections be used as data?

Types of Digital Collections Data

Metadata

structured and standardized descriptive content

Text

unstructured textual content derived from books, newspapers, transcripts, etc.

Audio/Video

unstructured video and/or audio content in the form of music, oral histories, interviews, home movies, etc.

Images

unstructured visual content in the form of photographs, drawings, maps, etc.

Close vs Distant Reading

in text analysis

But what does that mean?

Distant reading allows us to ask both simple and complex questions of texts.

university
2025
degrees
international
news
strategic
yr
fiu
support
implementation
higher
washington
research
horizon
national
students
metric
success
rate
innovation
academic
ed
bachelor's
taru
public
annual
doctoral
auxiliary
top
current
data
cost
student
alumni
engineering
300m
fy

DH Tools for Data

There are a number of tools and techniques for computational analysis in digital scholarship. These are just a few examples.

Plug and Play

Voyant
Palladio

Supervised Machine Learning

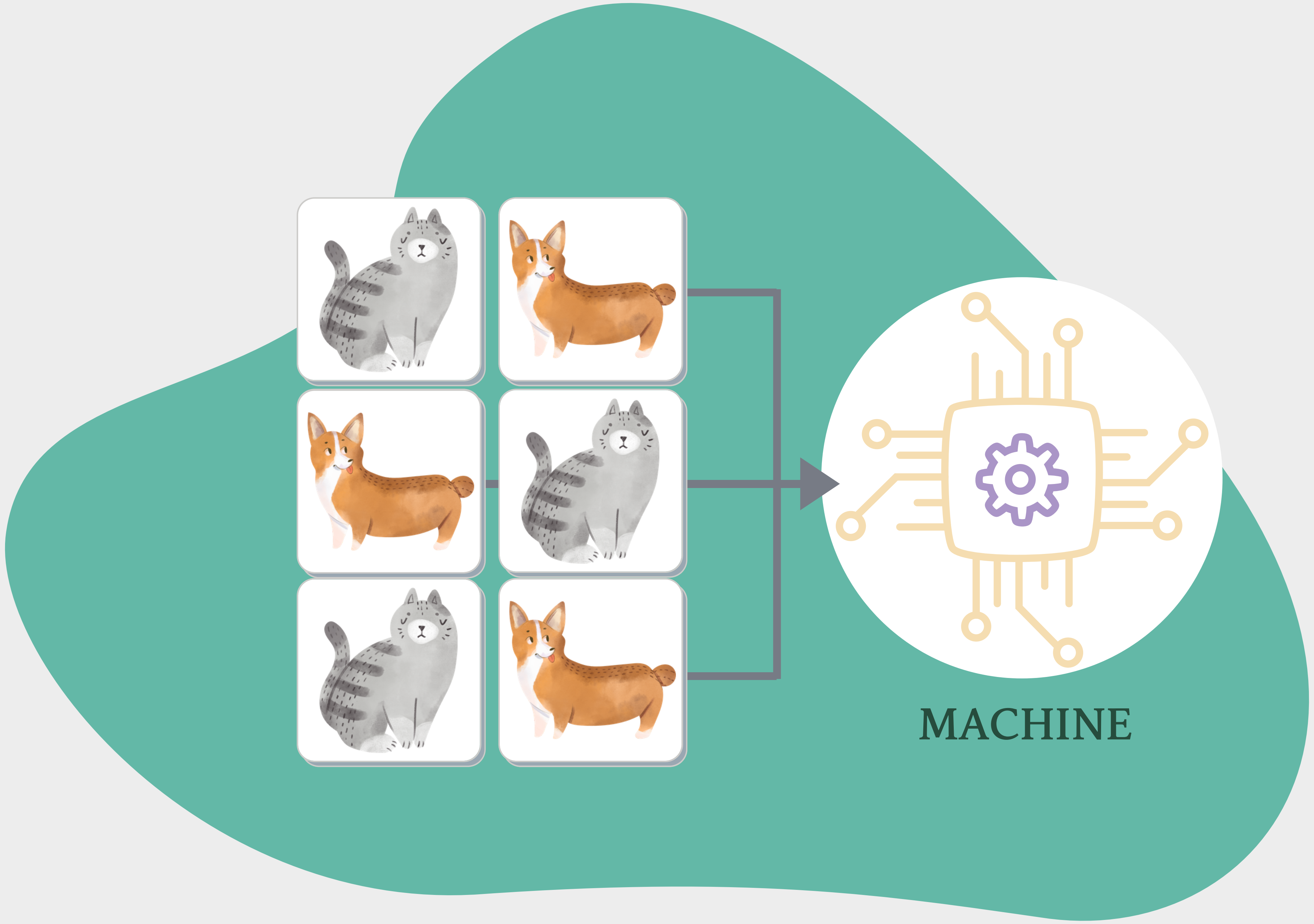
spaCy
Prodigy

Programing Languages

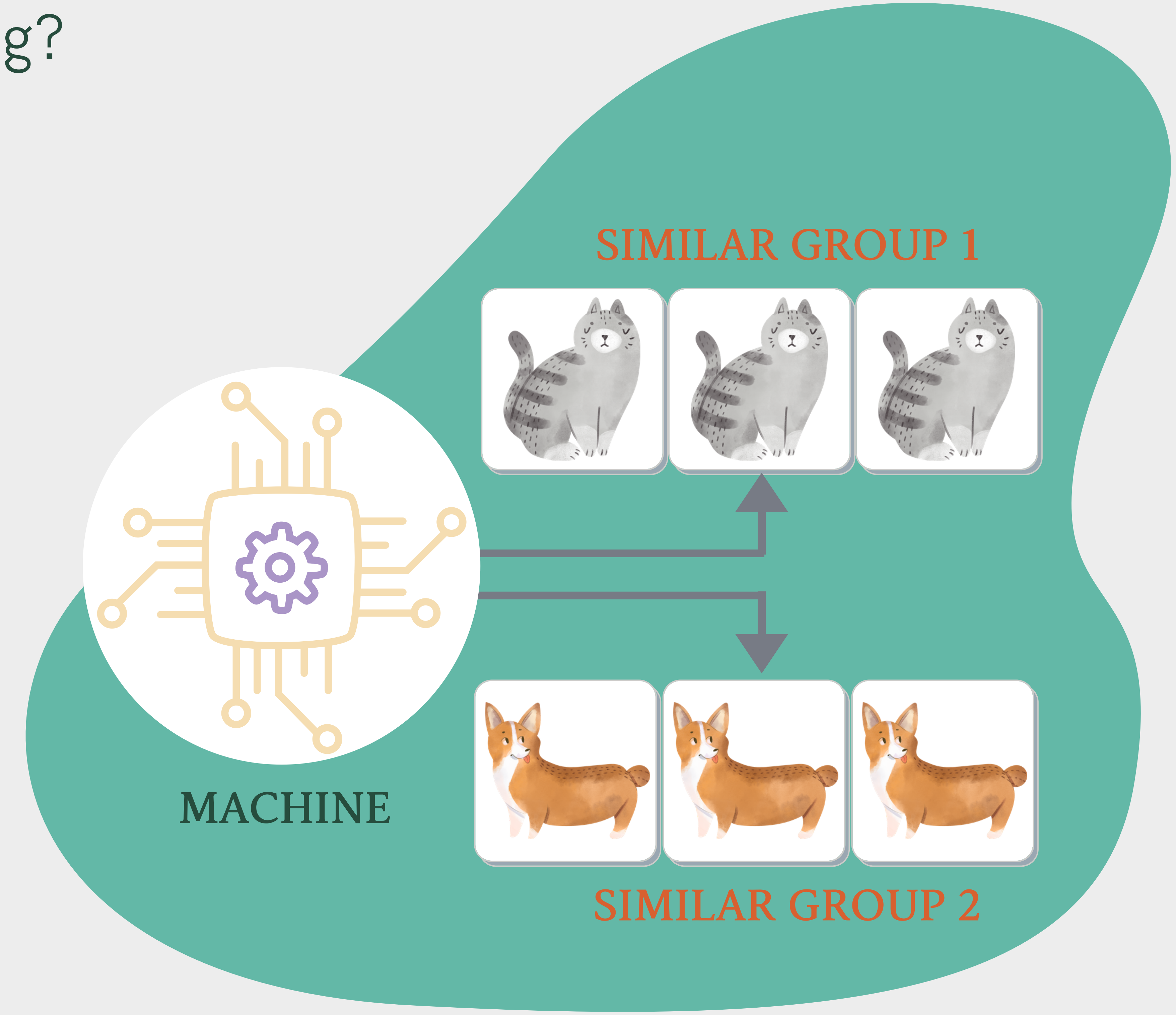
Python
R

What is

Unsupervised Machine Learning?



Step 1

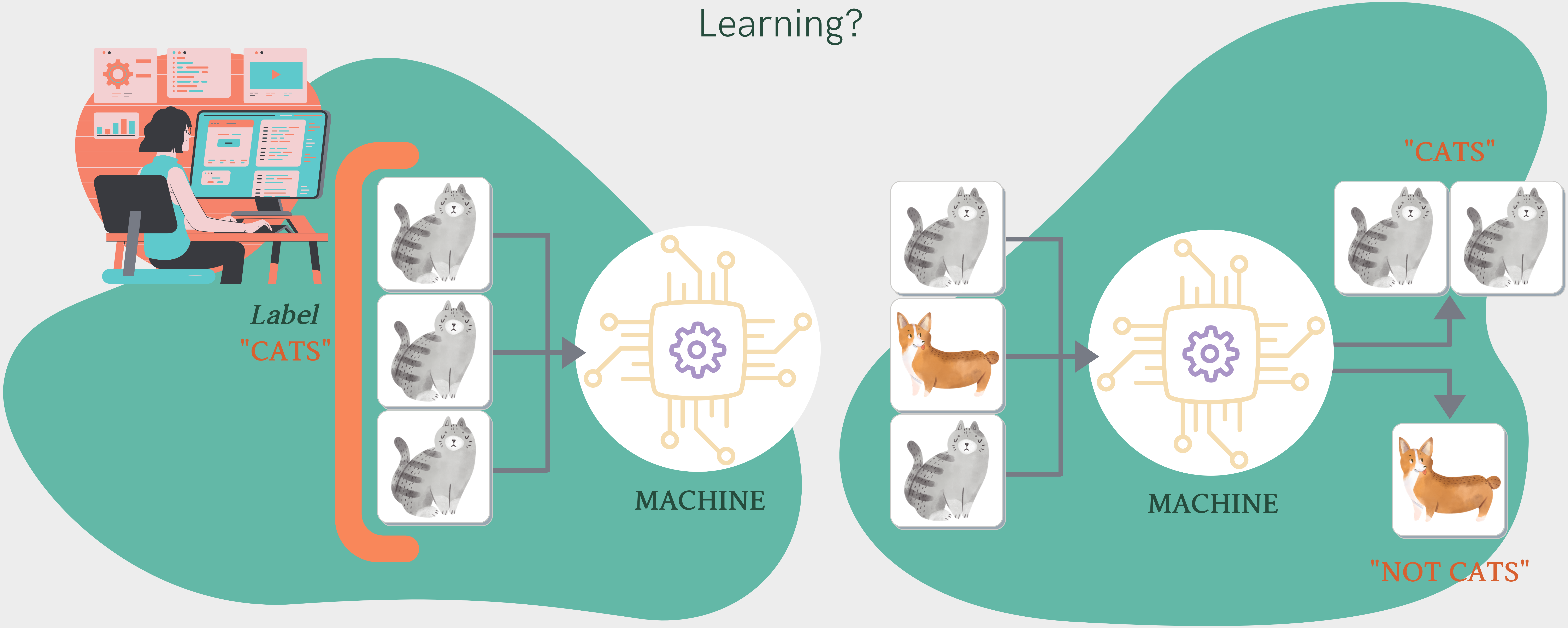


Step 2

*adapted from Hussnain Fareed's "Machine Learning for Dummies" in *towards data science*, 2018

What is

Supervised Machine Learning?



Label
"CATS"

MACHINE

"CATS"

MACHINE

"NOT CATS"

Step 1

Step 2

*adapted from Hussnain Fareed's "Machine Learning for Dummies" in *towards data science*, 2018

dLOC as Data

https://dlocasdata.domains.uflib.ufl.edu

Goal 1

Provide access to previously digitized Caribbean newspapers for bulk download of text and images

Goal 2

Develop a thematic toolkit for text analysis focusing on the history of hurricanes and tropical cyclones and their impact on the region

Goal 3

Develop outreach strategies that emphasizes training opportunities, supporting our local & dLOC community

THE ANDREW W. MELLON FOUNDATION

dLOC as Data Team

Project Leads

Miguel Asencio, Executive Director, Digital Library of the Caribbean (dLOC), FIU
(Senior Administrative Lead)
Jamie Rogers, Assistant Director of Digital Collections, FIU
(Senior Administrative Lead)
Perry Collins, *Scholarly Communications Librarian, UF*
(Project Lead)
Hadassah St. Hubert, Program Officer at National Endowment for the Humanities
(Scholarly Lead)

Project Team

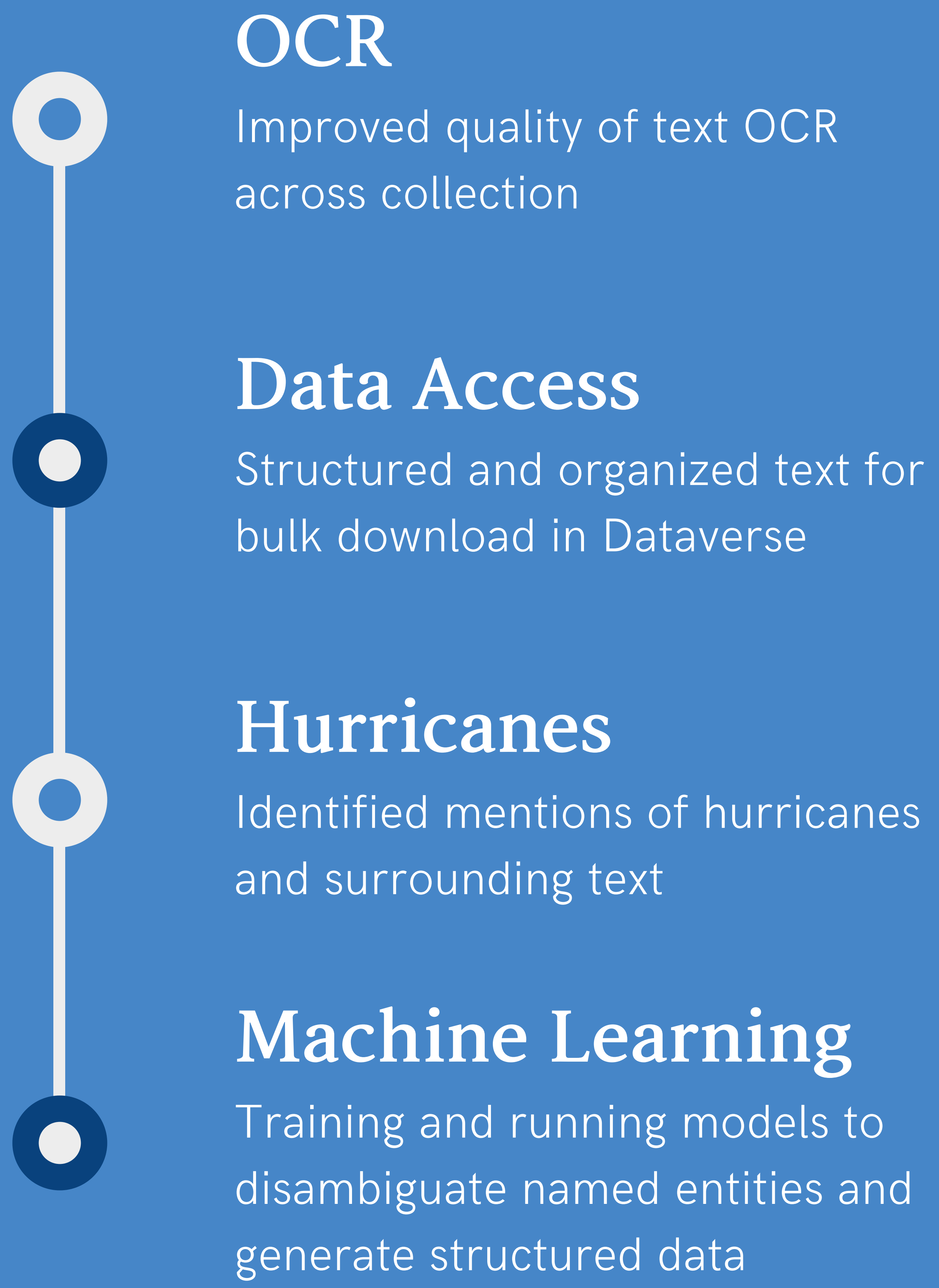
Rebecca Bakker, Digital Collections Librarian
Molly Castro, Digital Humanities Librarian
Boyuan Guan, Lead Developer, GIS Center
Jill Krefft, Institutional Repository Coordinator
Chelsea Dinsmore, Director, Digital Support Services
Laura Perry, Digital Production Manager, Digital Support Services
Laurie Taylor, Chair, Digital Partnerships & Strategies
Ivanna Moreno, Caribbean Data Curation Graduate Intern
Matthew Davidson, Caribbean Data Curation Graduate Intern

Advisory Board

Julio Capo Jr., Associate Professor of History, FIU
Fletcher Durant, Head of Conservation and Preservation, UF
Alex Gil, Digital Humanities Librarian, Columbia University
Melissa Jerome, Project Coordinator for the Florida & Puerto Rico Digital Newspaper Project, UF
Amalia Levi, Archivist and Cultural Heritage Professional, HeritEdge Connection in Barbados
Preeya Mohan, Fellow, Sir Arthur Lewis Institute of Social and Economic Studies, University of the West Indies, St. Augustine
Leah Rosenberg, Professor of English, UF

Project Activities

What we have accomplished so far…



OCR

Improved quality of text OCR across collection

Data Access

Structured and organized text for bulk download in Dataverse

Hurricanes

Identified mentions of hurricanes and surrounding text

Machine Learning

Training and running models to disambiguate named entities and generate structured data

Lesson 1

Good quality OCR is VERY important

Lesson 2

It is helpful to envision how data will be used in order to organize and make availble

Lesson 3

Having a research question in mind is VERY useful before you begin data analysis

Lesson 4

Even computer aided analysis requires significant human input and understanding of the data

Discussion

What FIU collections do you think would be useful as data?

Other project examples

Collections as Data: Part to Whole

https://collectionsasdata.github.io/part2whole/

Thank you!

Jamie Rogers

rogersj@fiu.edu