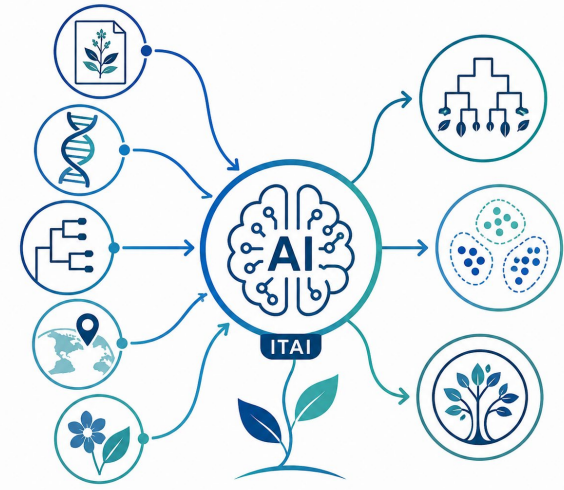


Paradigm Change: From **Identification AI** to **Integrative Taxonomic AI**



Sangtae Kim

Sungshin Women's University

Before ChatGPT

SLOWER. HARDER. MORE ISOLATED.



MANUAL RESEARCH



LONG EMAILS



TRADITIONAL LEARNING



ISOLATED WORK

RESEARCH PLAN
1. Read
2. Take Notes
3. Analyze
4. Write
5. Repeat

Dear Sir/Madam,
I hope this email finds you well.
I am writing to follow up on
regarding...
Best regards,
C...

2022



After ChatGPT

FASTER. SMARTER. MORE CONNECTED.



INSTANT ANSWERS

Get accurate information
in seconds



AI ASSISTED WRITING

Write, edit, and improve
with ease



CODE GENERATION

Build, debug, and ship
faster



TRANSLATION

Break language barriers
instantly



CREATIVITY BOOST

Generate ideas, art,
and content



SMARTER LEARNING

Personalized tutoring
and explanations



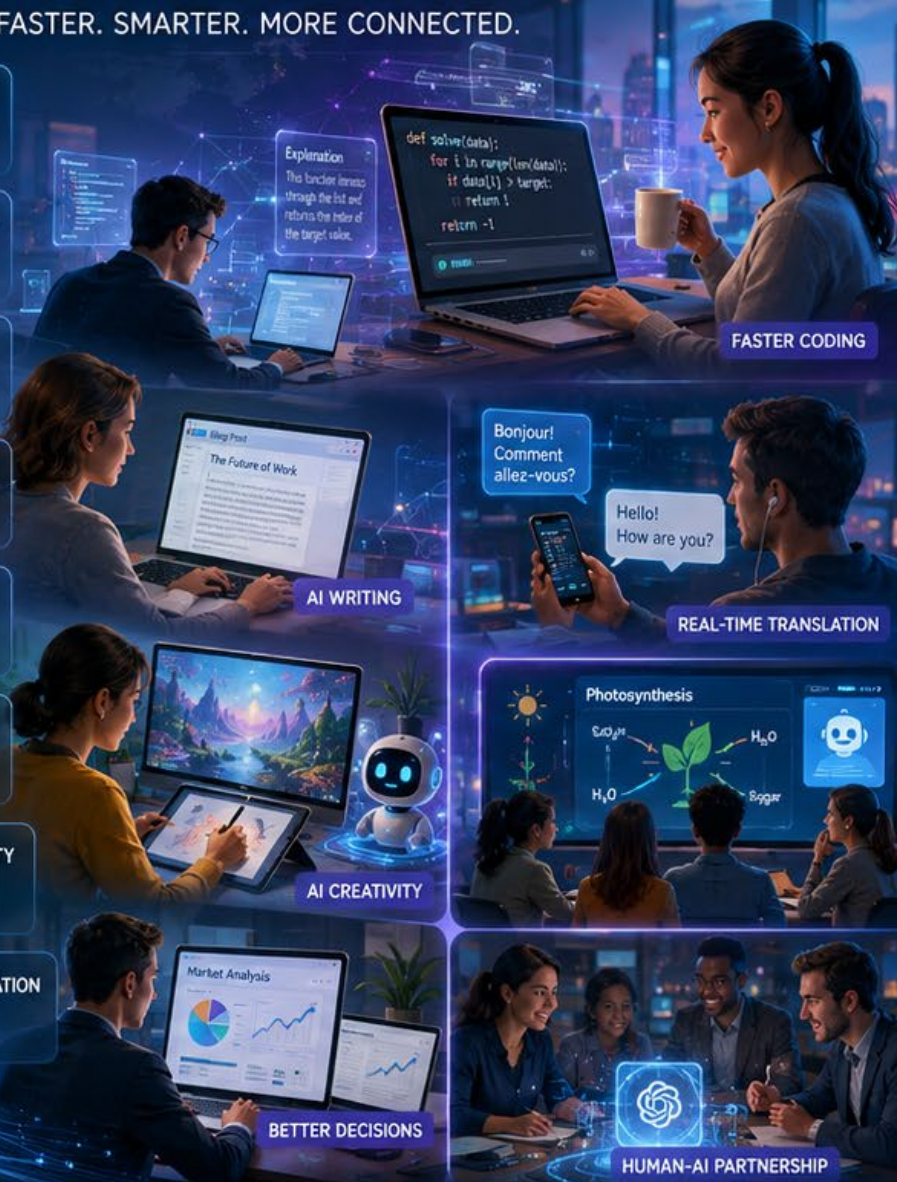
HIGHER PRODUCTIVITY

Automate tasks,
focus on what matters



HUMAN-AI COLLABORATION

AI as a partner,
not a replacement



Explanation
This function iterates
through the list and
returns the index of
the target value.

```
def solve(data):  
    for i in range(len(data)):  
        if data[i] == target:  
            return i  
    return -1
```

FASTER CODING

AI WRITING

Bonjour!
Comment
allez-vous?

Hello!
How are you?

REAL-TIME TRANSLATION

AI CREATIVITY

Photosynthesis
 $6CO_2 + 6H_2O \rightarrow C_6H_{12}O_6 + 6O_2$

BETTER DECISIONS



HUMAN-AI PARTNERSHIP



INFORMATION WAS
HARD TO FIND



TIME-CONSUMING
TASKS



LIMITED ACCESS
TO EXPERTISE



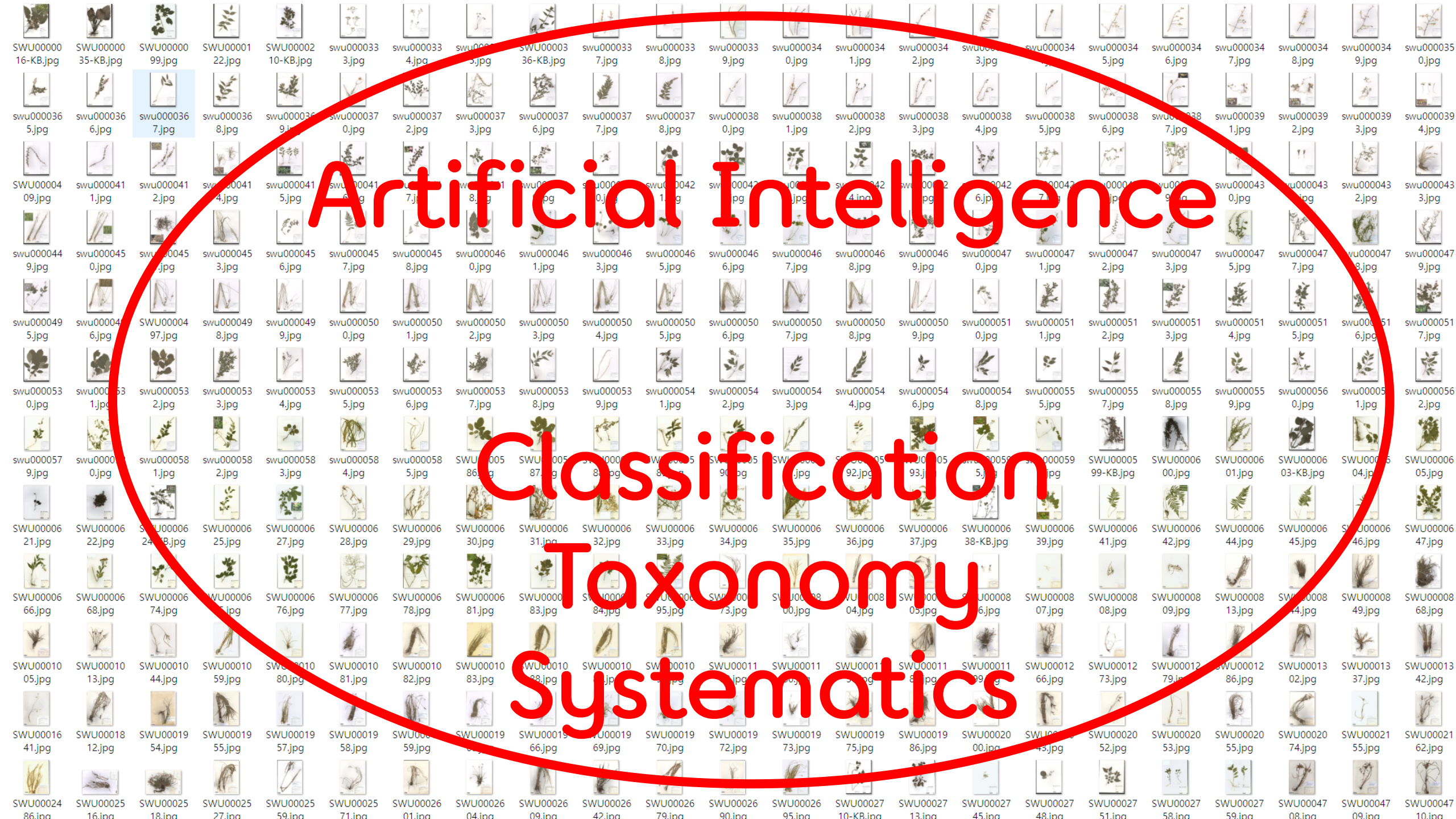
INFORMATION AT
YOUR FINGERTIPS



SAVE TIME,
DO MORE



ACCESS TO GLOBAL
KNOWLEDGE



Artificial Intelligence

Classification

Taxonomy

Systematics

Herbarium specimens in the age of artificial intelligence: from herbarium image identification to Integrative Taxonomic AI

Atsuko Takano^{1,2*}, Hyeonji Moon³, Masato Shirai⁴, Shuichiro Tagane⁵, Diego Tavares Vasques⁶, Chung-Kun Lee^{2,1}, Amjad Khan⁷, Jeongwoo Seo³, Sujeong Han⁸, Chanho Park⁹, Jaesung Lee^{5*}, Takashi Akihiro^{10*}, and Sangtae Kim^{3*}

- ¹ Institute of Natural and Environmental Sciences, University of Hyogo, Kobe, Japan
- ² Museum of Nature and Human Activities, Hyogo, Sanda, Japan
- ³ Department of Biology, Sungshin Women's University, Seoul, Republic of Korea
- ⁴ Interdisciplinary Faculty of Science and Engineering, Shimane University, Matsue, Japan
- ⁵ The Kagoshima University Museum, Kagoshima University, Korimoto, Japan
- ⁶ Graduate School of Science, The University of Tokyo, Tokyo, Japan
- ⁷ Graduate School of Human and Environmental Sciences, University of Hyogo, Kobe, Japan
- ⁸ Department of Artificial Intelligence, Chung-Ang University, Seoul, Republic of Korea
- ⁹ Fisheries Science Institute, Chonnam National University, Gwangju, Republic of Korea
- ¹⁰ Faculty of Life and Environmental Sciences, Shimane University, Matsue, Japan

ORCID

- Atsuko Takano: 0000-0002-8345-5080
- Hyeonji Moon: 0000-0001-8561-755X
- Masato Shirai: 0009-0002-6047-8121
- Shuichiro Tagane: 0000-0002-1974-7329
- Diego Tavares Vasques: 0000-0001-7978-2516
- Chung-Kun Lee: 0000-0003-3080-6905
- Amjad Khan: 0000-0002-5337-7237
- Jeongwoo Seo: 0009-0006-2294-6747
- Sujeong Han: 0009-0002-9444-6240
- Chanho Park: 0000-0003-4047-1918
- Jaesung Lee: 0000-0002-3757-3510
- Takashi Akihiro: 0000-0001-8349-3290
- Sangta Kim: 0000-0002-1821-4707

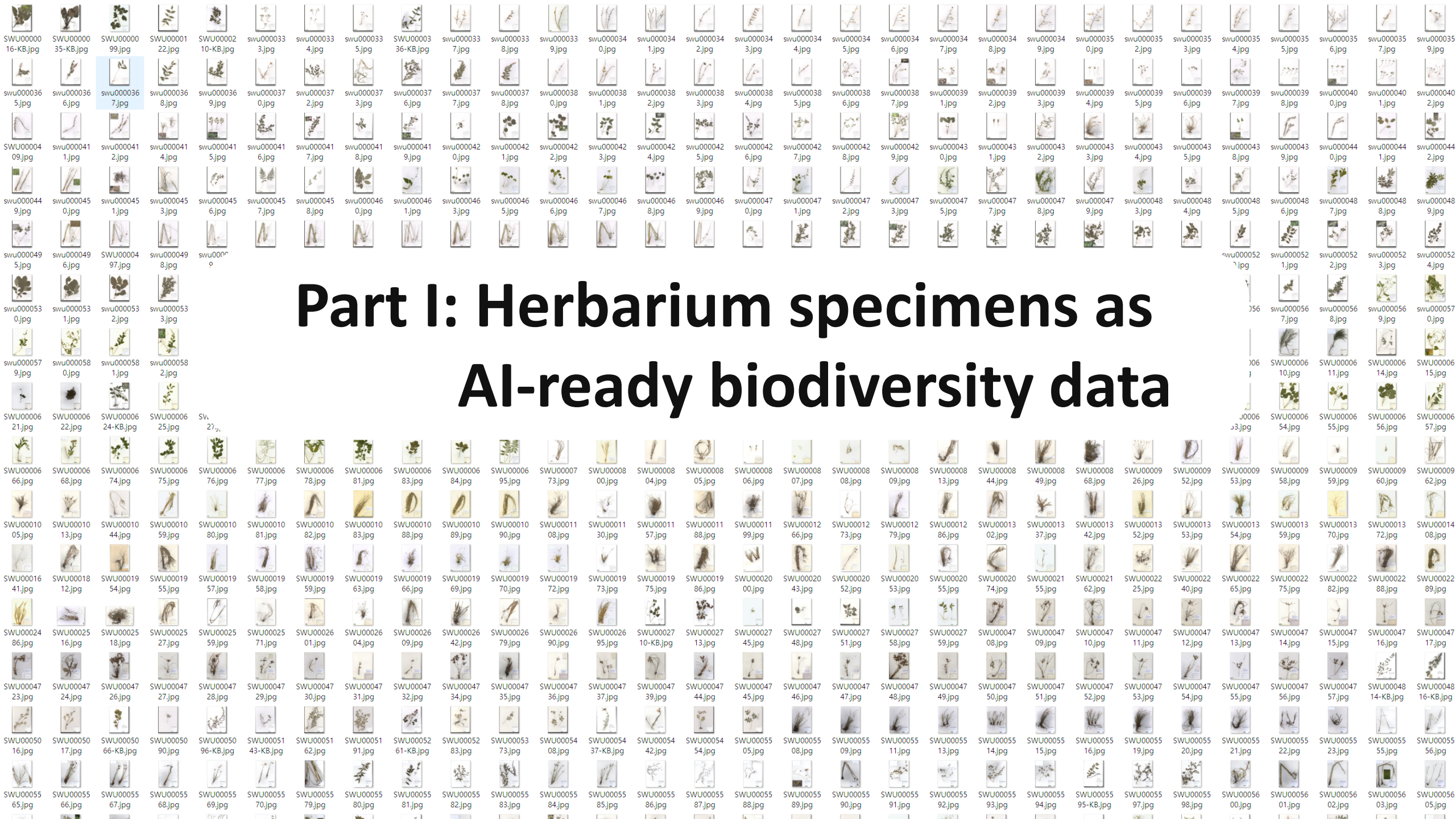
*Corresponding authors:

Sangtae Kim, Takashi Akihiro, Jaesung Lee, and Atsuko Takano

Journal of Plant Research
2026 in press

제 81회 한국생물과학협회 정기학술대회
2026. 7.13-15

Are Linnaean Type Specimens Truly Typical?
An AI-Based Unsupervised Analysis of
Herbarium Images Using Autoencoders:
A Case Study in Korean *Viola*





Herbarium: an hub of botanical information

**What we collect today enables
discoveries tomorrow!**

**Luca Ghini (1490-1556),
an Italian Botanist**

Herbaria enabled the systematic assembly of dispersed observations and records of individual species at a single location, functioning as centralized repositories of biological information, analogous to modern **digital data centers** that integrate distributed data resources.

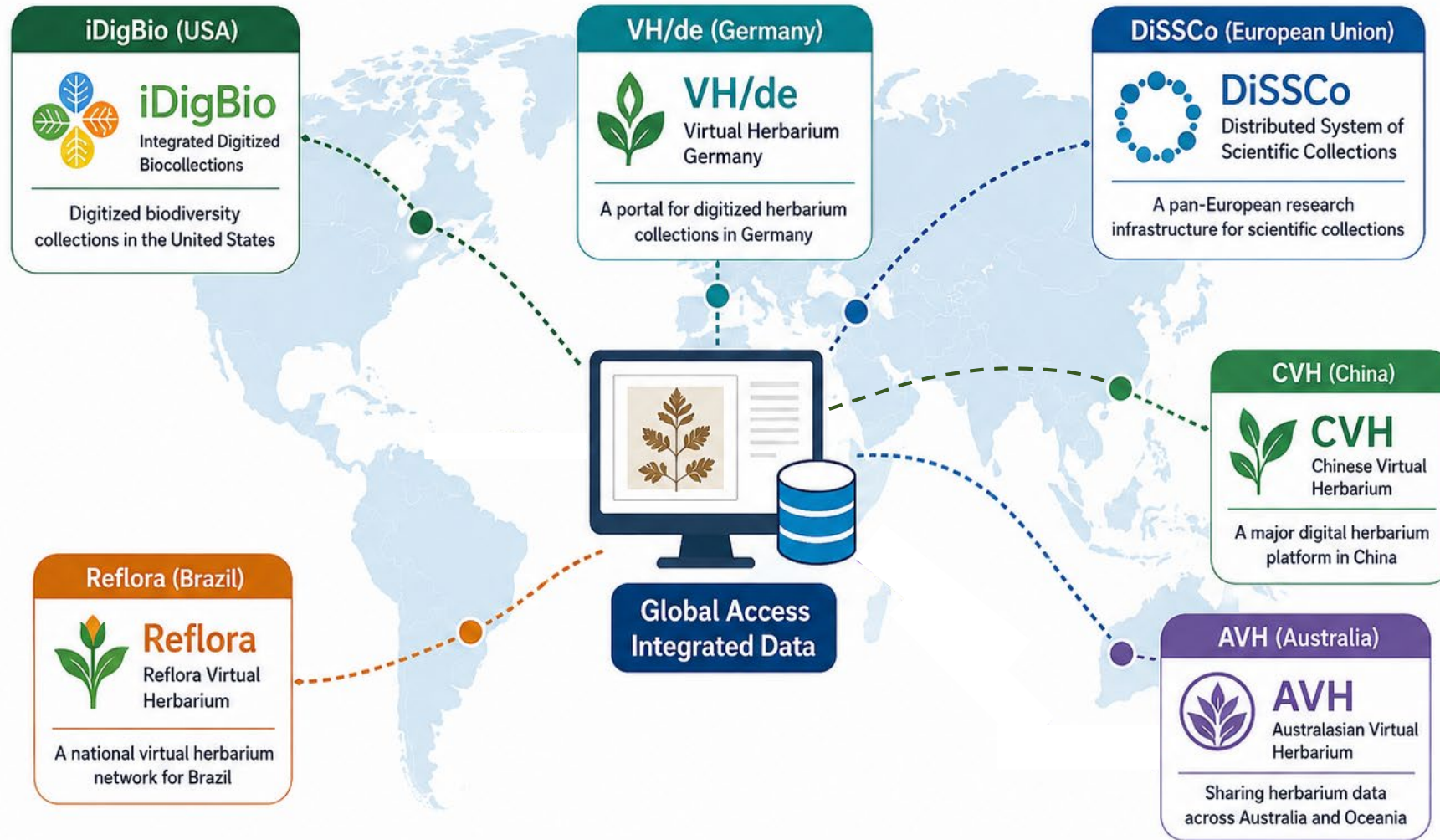
- Luca Ghini created the first herbarium
- **Anything is possible when you collect!**



An “idea” similar to
the current
AI analysis
environment based
on **big data**

Examples of the Major Virtual Herbaria worldwide

Connecting plant biodiversity data across the globe



Millions of Specimens



Rich Biodiversity Data



Global Collaboration



Open Access



Advancing Research
and Conservation

A list of the digital herbarium image portals in Asia (including transcontinental regions)

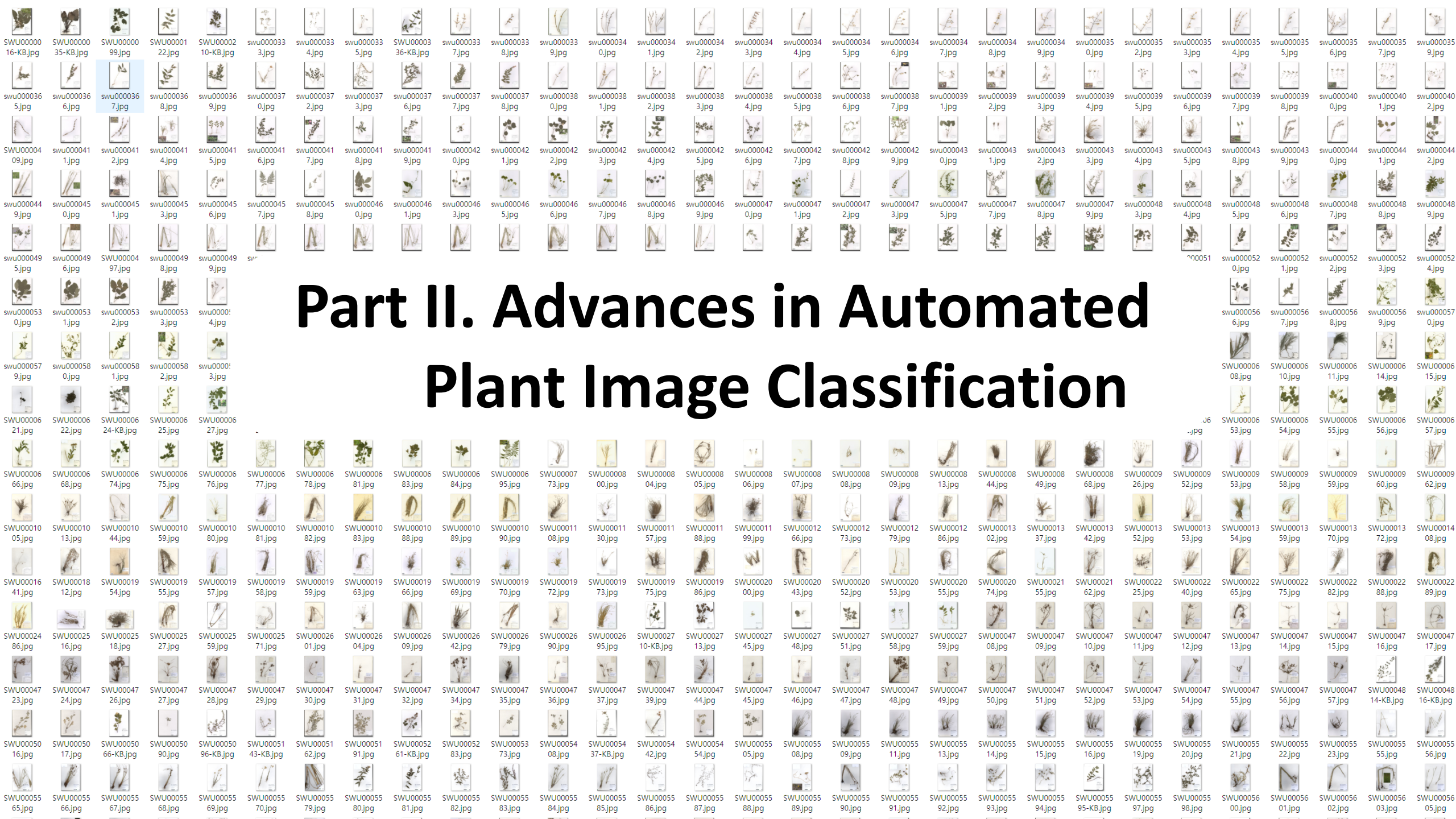
Country	Portal or herbarium name ^a	Number of images (Accessed on Feb., 2026)	URL
China	Chinese Virtual Herbarium (CVH)	8,559,422	https://www.cvh.ac.cn/
Russia	Moscow University Herbarium (MW)	1,403,266	https://plant.depo.msu.ru/
Japan	National Museum of Nature and Science (TNS)	388,916	https://db.kahaku.go.jp/webmuseum_en/
	Shimane University Digital Herbarium	332,278	http://tayousei.life.shimane-u.ac.jp/harbarium/
	Makino Herbarium Type Specimen Image Database (MAK)	ca. 740	http://ameba.i.hosei.ac.jp/BIDP/MakinoCD/makino/html_j/index0.html
	Kobe Gakuin University Botanical Specimens in Saikei Collection	ca. 20,000	https://db.kobegakuin.ac.jp/plant/index.html
	Museum of Nature and Human Activities, Hyogo (HYO)	262,461	https://nathist-net.jp/
	The Kagoshima University Herbarium		http://www.kuh.kyushu-u.ac.jp/hyouthonsitu.html
Taiwan	The University of Tokyo Herbarium		http://www.tshokubu.jp/en/TShokubu.php
	Herbarium of Taiwan		http://www.herbarium.gov.tw/
	Plants of Taiwan, National Herbarium, Academia Sinica		http://www.plants.taiwan.gov.tw/
India	Indian Virtual Herbarium (IVH)	155,531 ^c	https://ivh.bsi.gov.in/
Nepal	National Herbarium and Plant Laboratories (KATH)	164,471	http://plantdatabase.kath.gov.np/
Korea	Korea Biodiversity Information System, Korea National Arboretum (KH)		https://www.nature.go.kr/kbi/plant/smpl/selectPlantSmplGnrIsrcH1.do
	Plant DNA Bank in Korea, Korea University (KUS)	22,000	http://pdbk.korea.ac.kr/index.asp
Singapore	Singapore Botanic Gardens (SING)	46,985	https://herbaria.plants.ox.ac.uk/bol/sing
Thailand	Forest Herbarium (BKF)	13,827	https://botany.dnp.go.th/index_eng.html

Such large-scale access to digital data has made image-based AI analysis possible!

^a the characters in parentheses are abbreviations for the portal or herbarium

^b not indicated on the website

^c including vascular plants as well as algae, bryophytes, lichens, and fungi

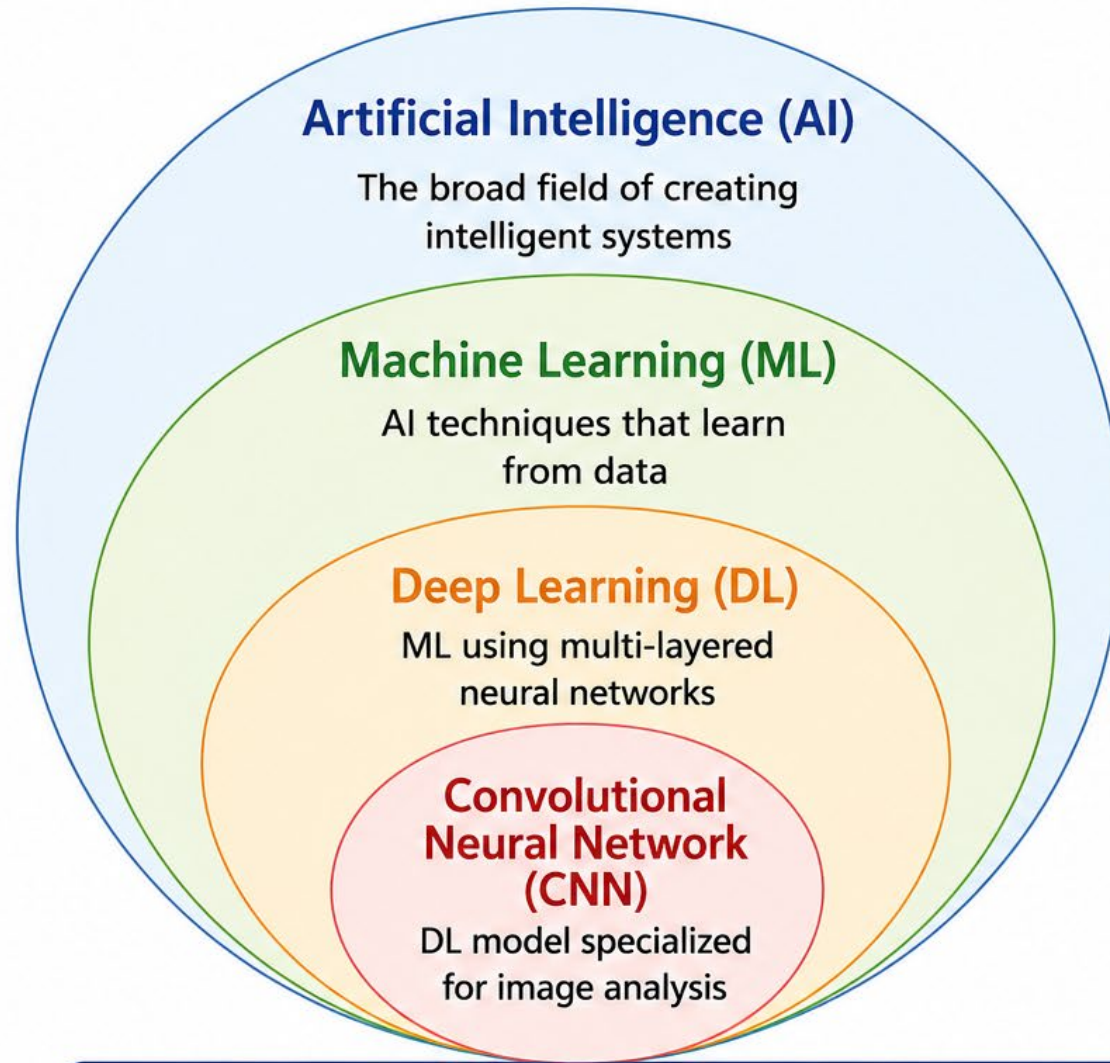


Efforts to automate plant identification

- Dating back to the **1970s**: Bases of the interactive key
Morse LE (1971). Specimen identification and key construction with time-sharing computers. *Taxon* 20:269–282
- Image-Based Automatic Identification: **Began in the 1990s**
e.g., Pankhurst (1991). Practical taxonomic computing. Cambridge University Press, Cambridge
- **ImageNet** Large Scale Visual Recognition Challenge (**ILSVRC**, 2010–2017)

An **international benchmark competition** that greatly accelerated the development of modern computer vision and **deep learning**.
- **ImageCLEF**, part of the Cross Language Evaluation Forum (CLEF)
PlantCLEF (plants; 2011~), BirdCLEF (birds), and AnimalCLEF (animals)
- **LeafSnap** (Kumar et al., 2012)
- **Agriculture**: Detection of pest damage, early detection of disease symptoms, etc.
e.g., grape (Al-Shamayleh and Ibrahim 2025), tea (Wen et al. 2025), rice (Raman and Jayaraman 2025), soybean (Zhang et al. 2025), and date palm (Mir et al. 2025)...

Relationship between Machine Learning, Deep Learning, and CNN



Artificial Intelligence (AI)

Machines that think and act like humans



Machine Learning (ML)

Algorithms that learn patterns from data



Deep Learning (DL)

ML based on deep (multi-layer) neural networks



Convolutional Neural Network (CNN)

Deep learning model designed for image data



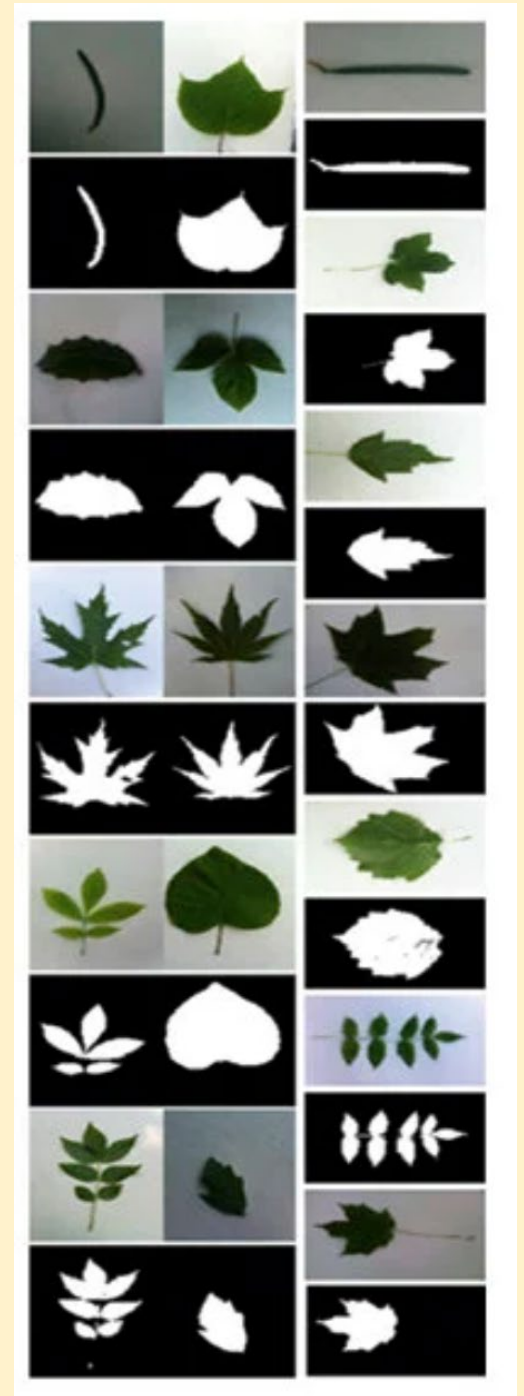
In herbarium image analysis:

CNN learns visual features from herbarium specimen images and uses them for species identification, segmentation, and trait detection.

LeafSnap

(Kumar et al., 2012)

It was one of the hot issues at the IBC XVIII meeting in Melbourne (2011) and was the title of the keynote address



Kaggle competition managed by google

A platform for data science
enthusiasts to participate in
machine learning competitions
and challenges.

kaggle

Hi Sangtae Kim (김상태),

Can your model “unlearn” data that was used for training? Outside of simply deleting data used to train a model, can the influence of that data be removed?

This competition considers a realistic scenario in which an age predictor has been trained on face images, and, after training, a certain subset of the training images must be forgotten to protect the privacy or rights of the individuals concerned.

Total Prizes:

\$50,000

Entry Deadline:

November 20, 2023

[Learn More](#)

Your goal is to submit code that takes as input the trained predictor, the forget and retain sets, and outputs the weights of a predictor that has unlearned the designated forget set.

Good luck,

Walter Reade
Kaggle Data Scientist

Kaggle competition (2021) winner’s paper



ORIGINAL RESEARCH
published: 13 January 2022
doi: 10.3389/fpls.2021.806407



Harnessing Large-Scale Herbarium Image Datasets Through Representation Learning

Barnaby E. Walker^{1*}, Allan Tucker² and Nicky Nicolson¹

¹ Royal Botanic Gardens, Kew, Richmond, United Kingdom, ² Department of Computer Science, Brunel University London, Uxbridge, United Kingdom

OPEN ACCESS

Edited by:
Alexis Joly,
Research Centre Inria Sophia
Antipolis Méditerranée, France

Reviewed by:
Mathias Dillen,
Botanic Garden Meise, Belgium
Elizabeth Ellwood,
University of Florida, Gainesville,
United States

***Correspondence:**
Barnaby E. Walker
b.walker@kew.org

Specialty section:
This article was submitted to
Technical Advances in Plant Science,
a section of the journal
Frontiers in Plant Science

Received: 31 October 2021
Accepted: 06 December 2021

INTRODUCTION

- NY provided two million images!



+

Create

Home

Competitions

Datasets

Models

Code

Search

Research Prediction Competition

Herbarium 2022 - FGVC9

Identify plant species of the Americas from herbarium specimens

FGVC Fine-Grained Visual Categorization · 134 teams · a year ago

Overview

DataCodeDiscussionLeaderboardRules



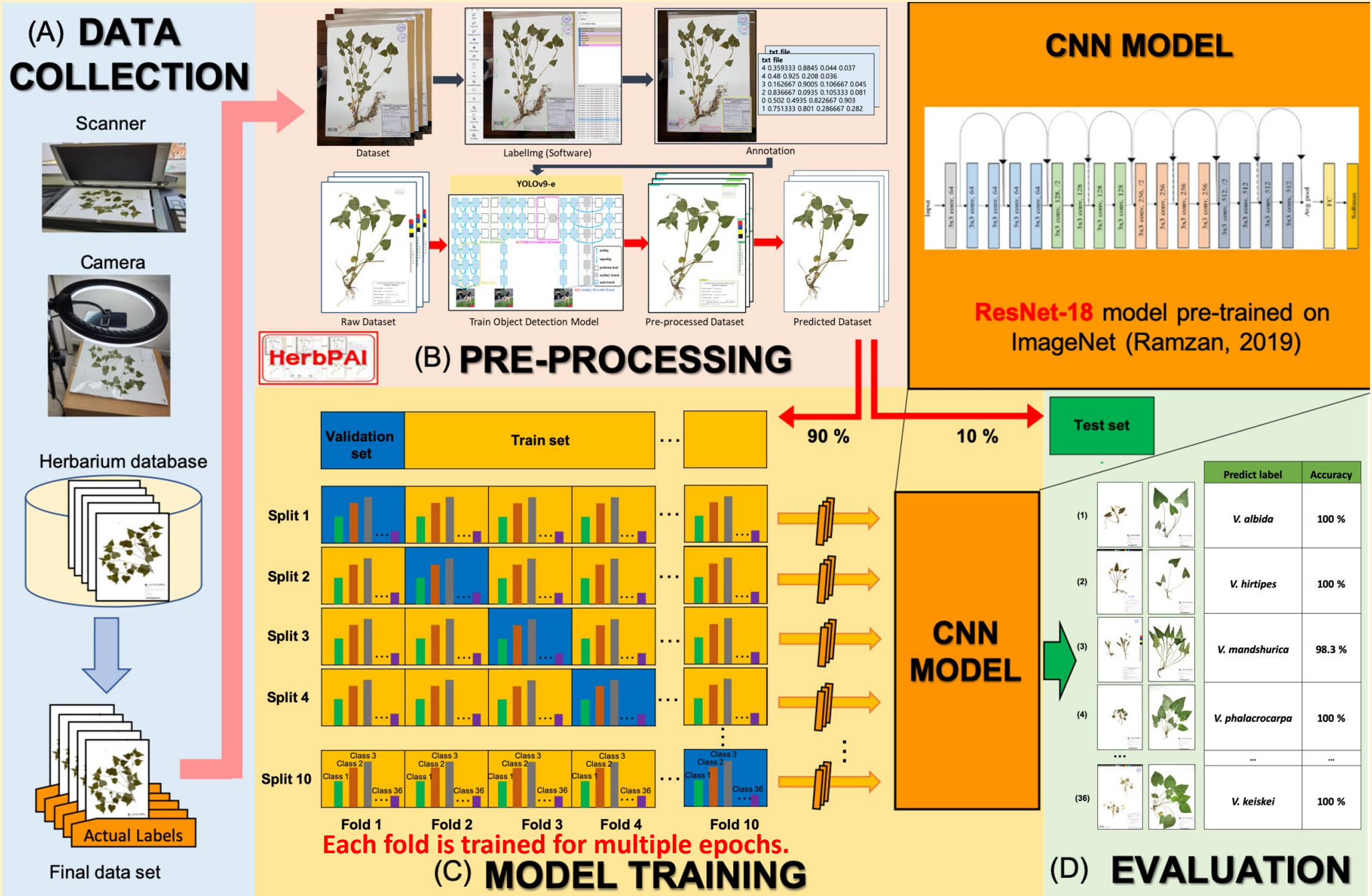
The Herbarium 2022: Flora of North America dataset comprises **1.05 M images** of **15,501 vascular plants**, which constitute more than **90% of the taxa** documented in North America. Our dataset is constrained to include only **vascular land plants** (lycophytes, ferns, gymnosperms, and flowering plants).

Our dataset has a long-tail distribution. **The number of images per taxon** is as few as seven and as many as 100 images. Although more images are available, we capped the maximum number in an attempt to ensure sufficient but manageable training data size for competition participants.

An example of a schematic process showing the development of an automatic herbarium image identification system

An epoch is one complete pass through the training set within that fold.

Moon et al., in prep.



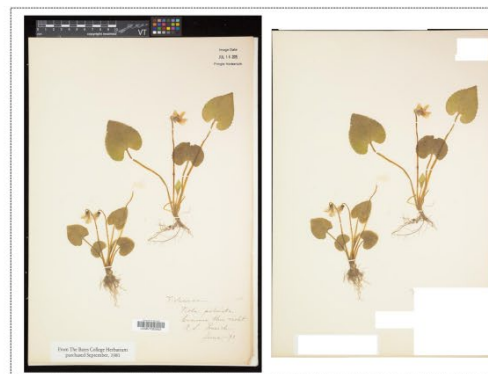


Herbarium image Preprocessing for AI studies

Han et al. *Software X*. in prep.



(a) Scanned herbarium image



(b) Captured with a smartphone



(c) Captured with a digital camera



(d) Low-quality image



(e) Image from a foreign herbarium



(f) Image with a non-white background

An Example of Results from Research on the Automatic Identification of Plant Specimen Images Using Deep Learning

Table 2. Classification performance for different combinations of image sizes and models

Image size	Model	Macro accuracy	Macro F1-score	Macro precision	Macro recall
256 X 256	ResNet-18	0.8409	0.7419	0.7302	0.7766
	ResNet-34	0.8203	0.7225	0.7202	0.7561
256 X 384	ResNet-18	0.8651	0.7703	0.7578	0.8025
	ResNet-34	0.8572	0.7600	0.7487	0.7294

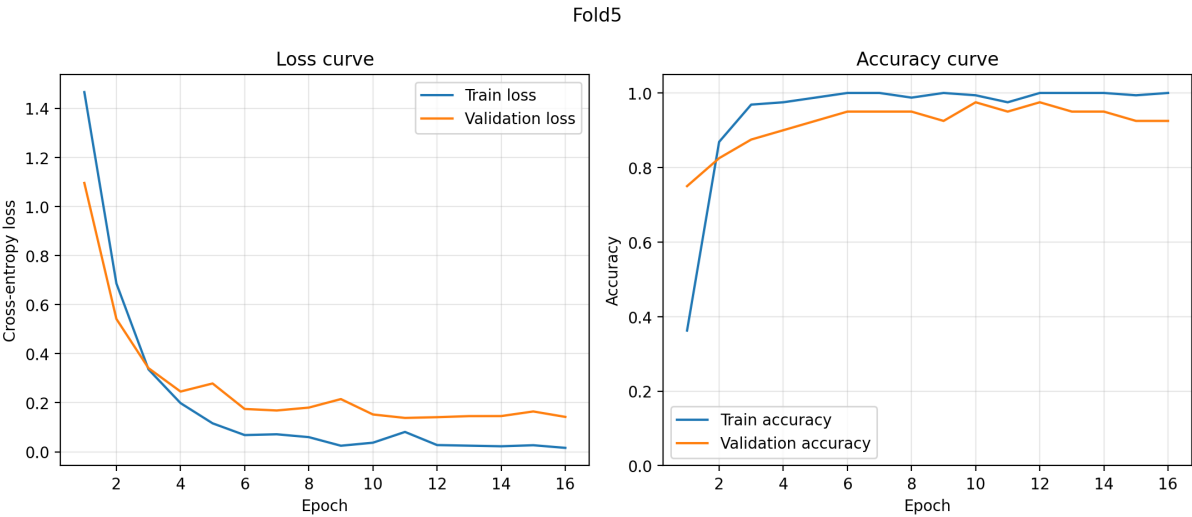


Figure 3. Loss curve graph during model training. Loss indicates the difference between the predicted class and the actual class.

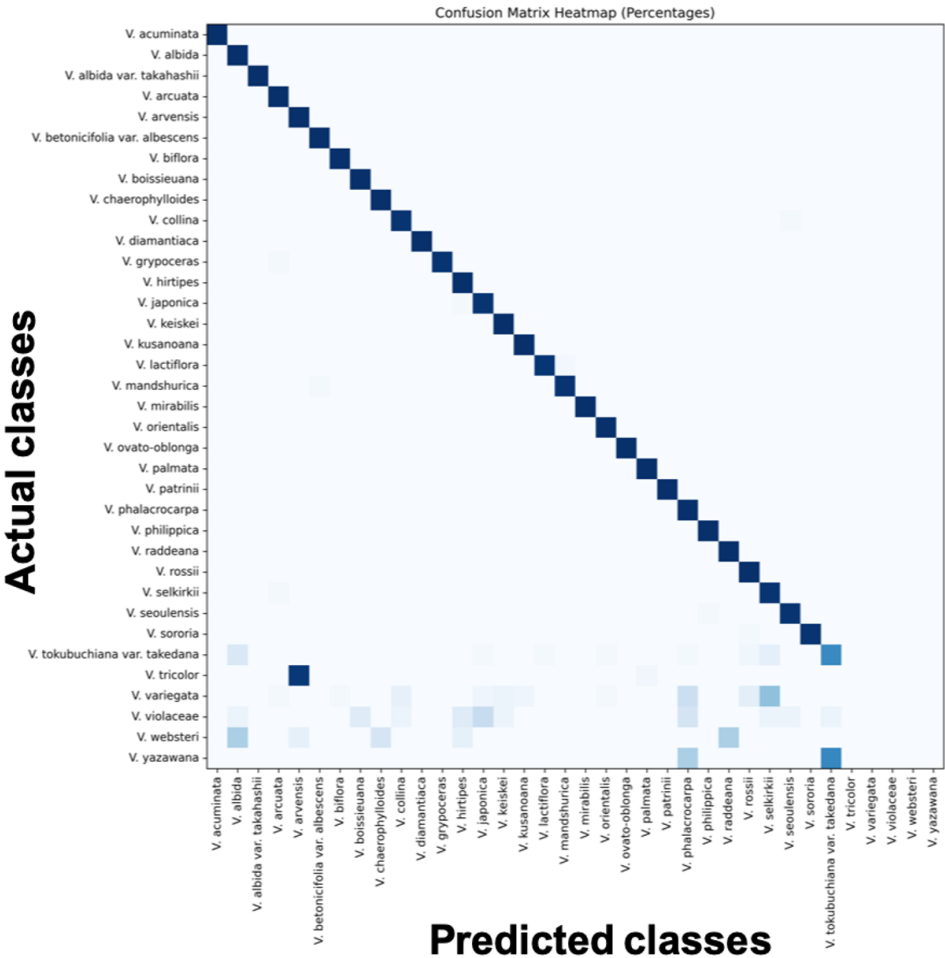
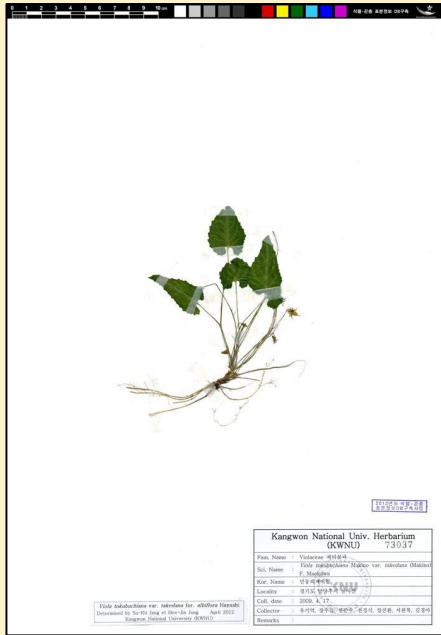


Figure 4. Confusion matrix.

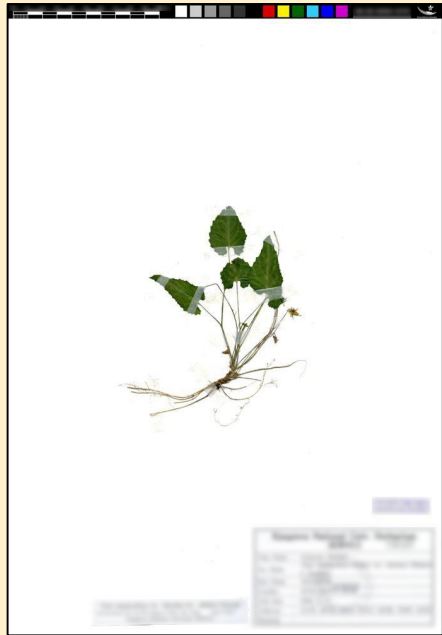
Grad-CAM (Gradient-weighted Class Activation Mapping):

which regions of an image **contributed most strongly** to a deep learning model's prediction.

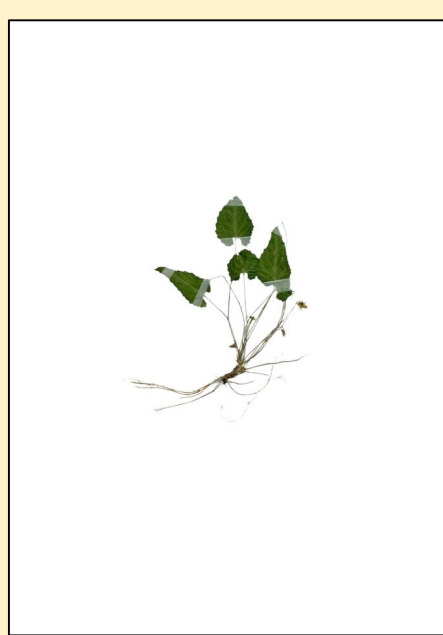
a



b



c



d



e



Image preprocessing strategies and Grad-CAM–based interpretation in herbarium image analysis.

(a) Raw herbarium specimen image with non-plant components.

(b) Text blurring.

(c) Segmentation-based preprocessing.

(d) (e) Grad-CAM visualization highlighting model attention on raw (d) and segmented (e) images.

In the raw image, the classification model often focuses on non-plant regions, whereas in the segmented image, it only focuses on plant structures.

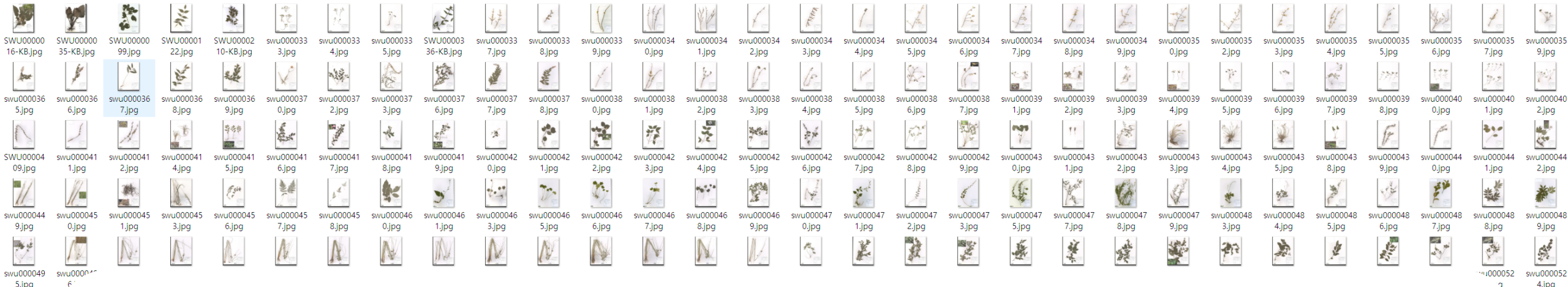
Table 1. Examples of deep learning–based plant classification using herbarium images (multimodal-based methods not included)

Reference	Dataset						Input image size (pixels)	Model architecture		Model performance ^c	
	Taxonomic or geographic category	Source of dataset	# of images	# of taxa (Class)	Taxonomic rank	# of images per class		Architecture category ^b	Baseline architecture	Top-1 accuracy (%)	Macro F1 score
Carranza-Rojas et al. (2017)	-	Herbarium 1K -iDigBio	253,733	1,204	-	-	224 x 224	CNN	Inception v1 (GoogleNet)	79.6	-
Younis et al. (2018)	-	Typical herbarium specimen image: 5-20 MB (jpg) 224 × 224 pixel JPG image: ca. 15-40 KB → ca. 500 times difference!!!						CNN	Modified ResNet	82.4	-
Tan et al. (2019); Little et al. (2020)	Melastomataceae							CNN Ensemble	SeResNeXt-101, ResNet-50	89.8	-
de Lutio et al. (2022)	American, Oceanian, and Pacific species	Herbarium 2021 (FGVC8): Half-Earth	2,500,779	64,500	-	(test set: n≤100)	256×256, 352×352	CNN Ensemble	ResNest-101, ResNeXt-101-IBN-a, ResNeXt-101	84.5	0.757
Park et al. (2024)	90% of North American vascular plants (Canada, Greenland, United States of America)	Herbarium 2022 (FGVC9): NAFLora-1M	105,000	15,501	-	10<n≤100	384 x 384	CNN-ViT Ensemble	ResNeSt-101, EfficientNet-B6, DeiT-III-B, Swin-B, Swin-L, ConvNeXt-B	-	0.877
Shirai et al. (2022)	33.9 % of Japanese vascular plants ^a	Japanese herbaria (author collection, download)	500,554	2,171	192 families, 2064 species, 9 subspecies, 88 varieties, 10 forms	n≥50	299 x 299	CNN	Inception-ResNet-v2	96	0.923
Xie et al. (2025)	53.9% of Chinese vascular plants	Chinese Virtual Herbarium (CVH)	6,500,000	21,589	309 families, 3144 genera, 19723 species, 1866 infraspecific taxa	16≤n≤200	299 x 299	CNN	EfficientNetV2	67.3	-
Kajihara et al. (2025)	Brazilian Piperaceae	speciesLink	10,503	236	1 family, 3 genera	5≤n<240	512 x 512	ViT	ViT	-	0.37
			9,977	160		10≤n<240				-	0.4
			9,237	106		20≤n<240				-	0.45
Arno et al. (2025)	<i>Bulbostylis</i>	GBIF-Royal Botanic Gardens, Kew	4,936	87	1 genus	20≤n≤100	224 x 224	ViT	ViT-H/14	65.38	-
	<i>Ziziphus</i>		1,654	27	1 genus				ViT-B/16	72.61	-

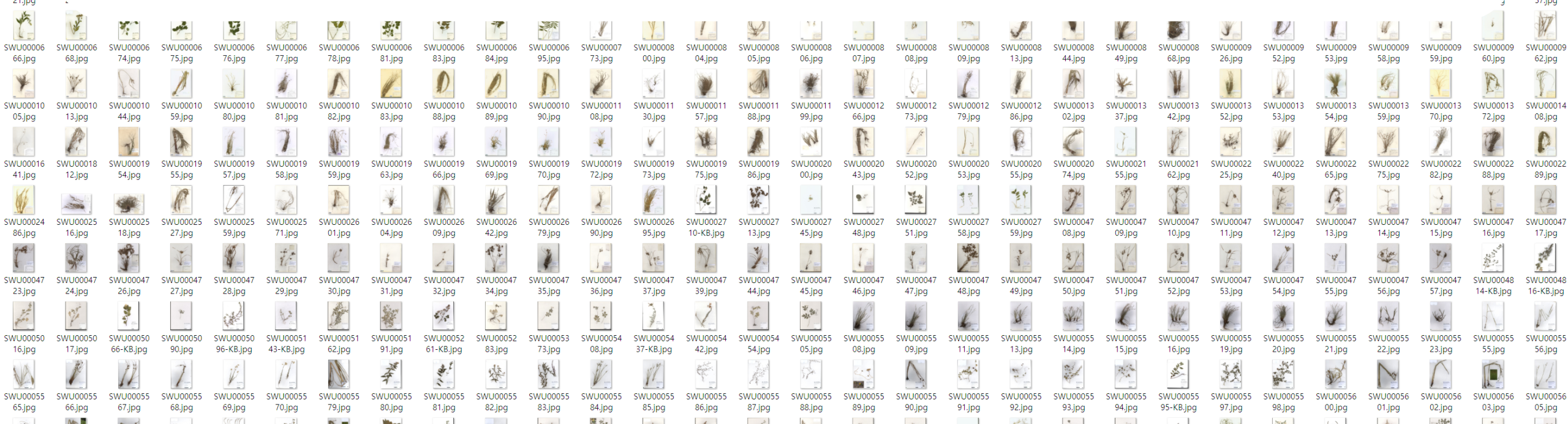
^a % calculation based on the Flora of Japan (reviewed in Wang et al. 2023).

^b CNN, Convolutional Neural Network; ViT, Vision Transformer.

^c Only the best classification results were included.



Part III. Object Detection, Segmentation, Trait Extraction, Label Recognition, and Error Detection



Object Detection

Object detection using **YOLO**. Diwan et al., 2023. *Multimedia Tools and Applications*.

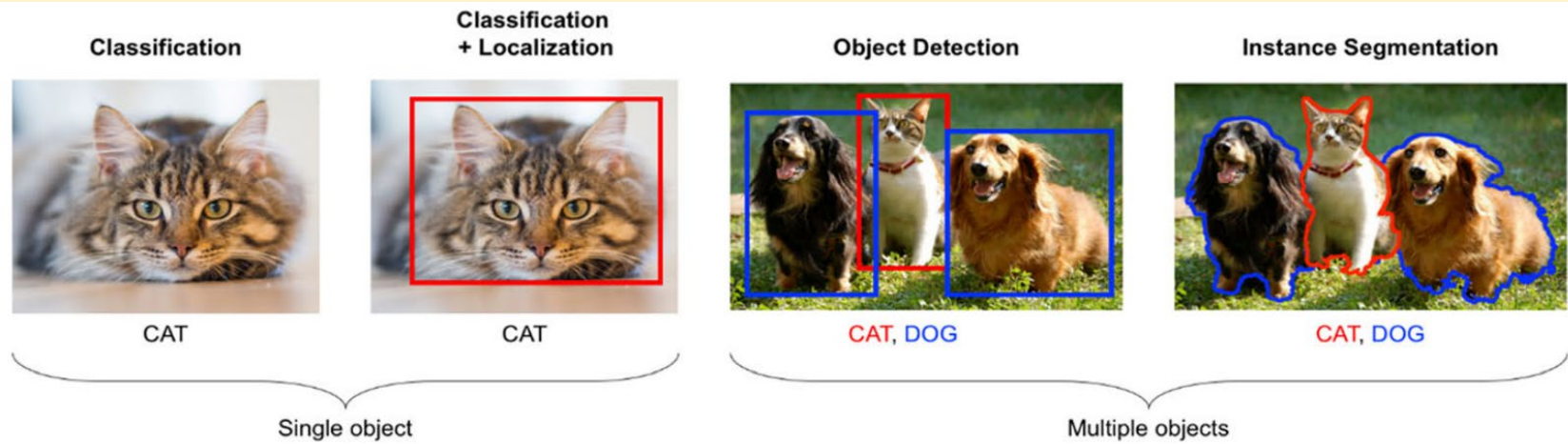


Fig. 1 Classification, Localization, and Segmentation in Single and Multiple Objects image [13]

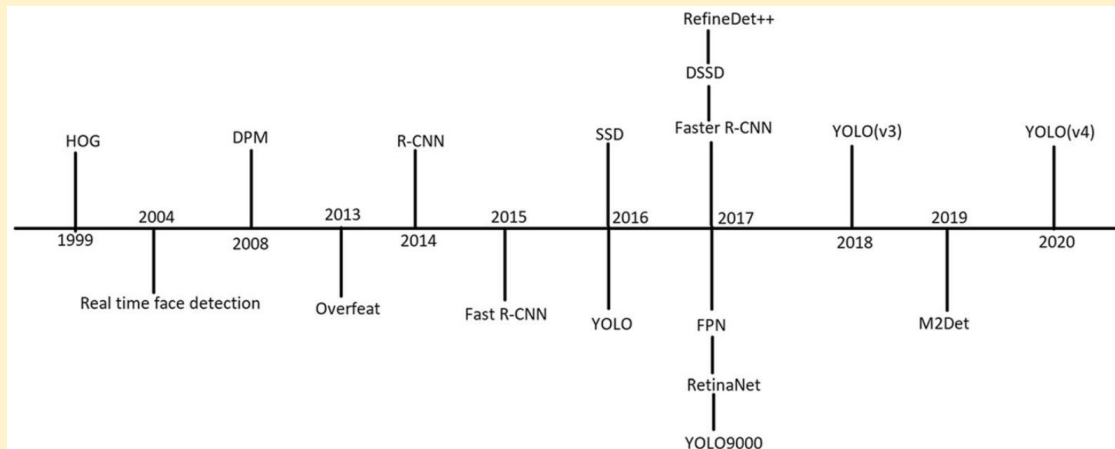


Fig. 3 Year wise evolution of object detection algorithms

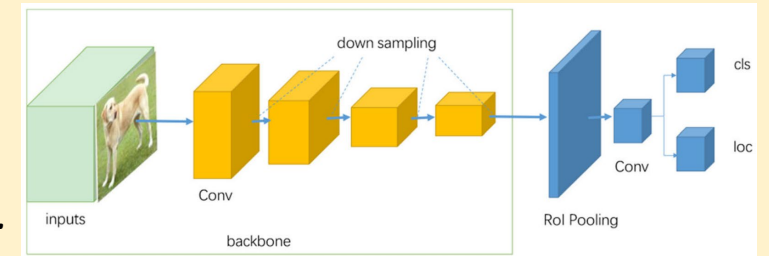


Fig. 2 Generic architecture of single stage object detectors [29]

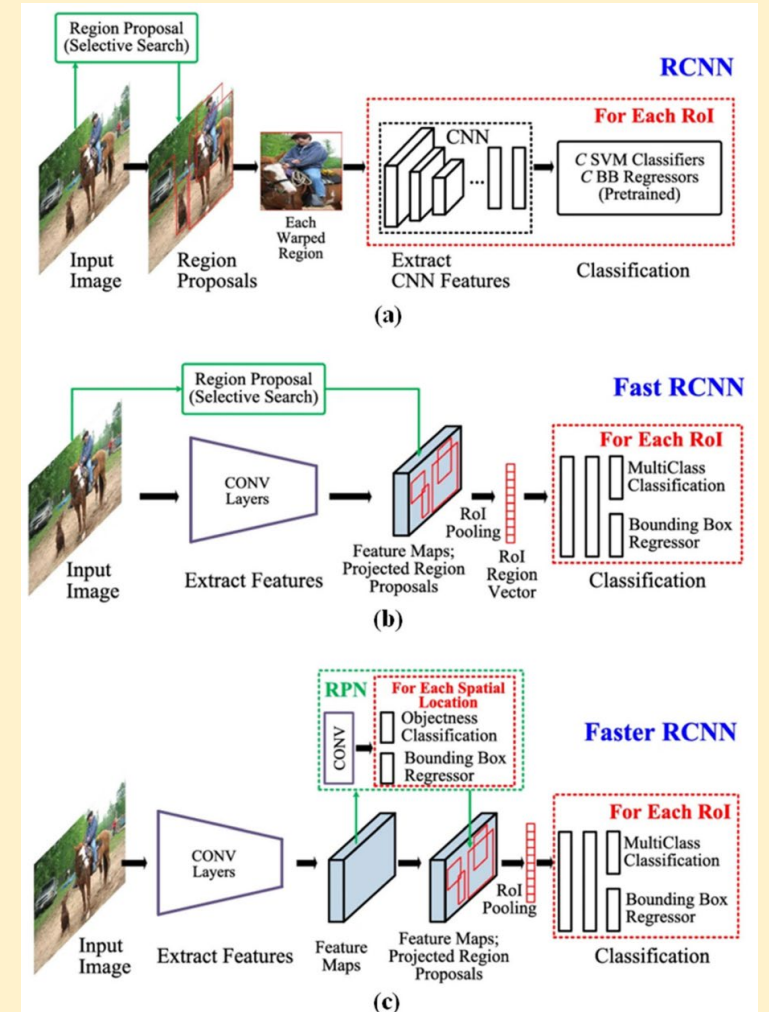
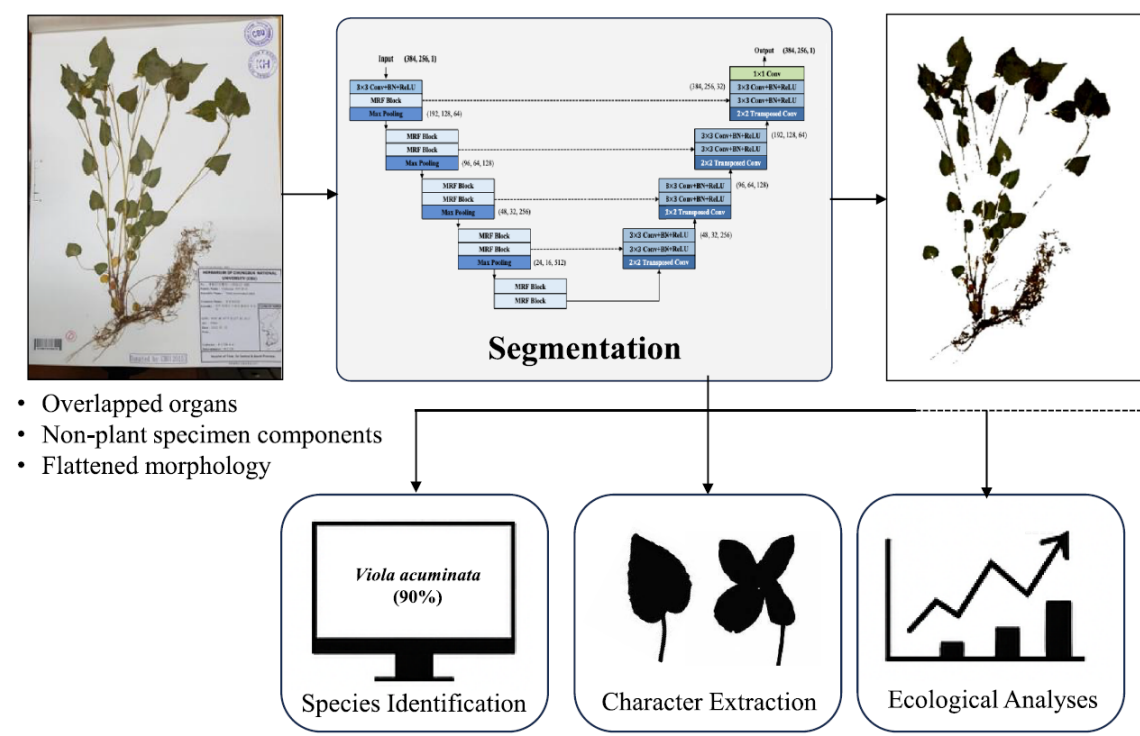


Fig. 5 Two stage object detectors (a) RCNN (b) Fast-RCNN (c) Faster-RCNN [44]

Segmentation of the plant part

Exclude non-plant components such as **labels, barcodes, color palettes, rulers, background areas** outside the specimen, and **digitization artifacts** introduced during curatorial and imaging practices

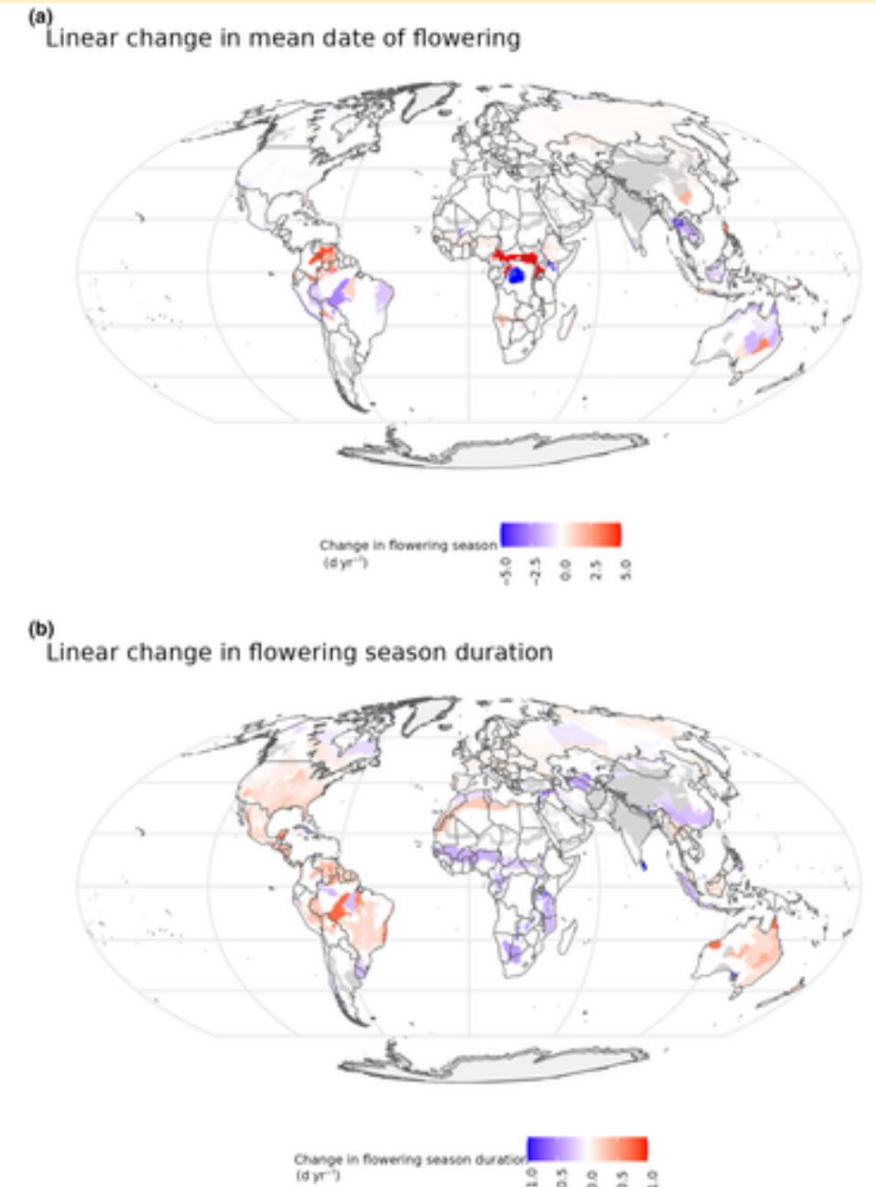
Han, Moon, ... Kim, & Lee (2026) *Expert Systems With Applications*



Structural taxonomy and comparative summary of representative segmentation models.				
Model	Architectural Category	Core Architectural Strategy	Structural Feature	Application Context
CaraNet	CNN + Contextual Attention	Axial reverse attention for small object segmentation	ARA and CFP modules for small object enhancement	Tumor, polyp segmentation
DS-TransUNet	Transformer CNN Hybrid	Dual Swin Transformer encoders with CNN decoder	Combines global attention with local spatial refinement	Medical segmentation
HarDNet-MSEG	CNN Encoder -Decoder	Efficient dense connectivity using HarDNet backbone	HarDBlk modules and cascaded partial decoder	Medical, polyp segmentation
SegFormer	Transformer Encoder	Hierarchical Transformer encoder with lightweight MLP decoder	No positional encoding, efficient MLP decoder	General and remote sensing segmentation
UNet3 +	CNN + Attention	Full-scale skip connections and deep supervision for richer feature integration	Aggregates features from all encoder levels at each decoder stage	Remote sensing, medical segmentation
SETR	Transformer Encoder	Pure Transformer encoder for sequence-to-sequence segmentation	Global context modeling, upsampling decoder	General semantic segmentation
UNet+ +	CNN Encoder -Decoder	Nested and dense skip pathways to improve feature fusion	Redesigned skip connections for multi-scale aggregation	Medical, biomedical segmentation
DeepLabv3 +	CNN Encoder -Decoder	ASPP and encoder-decoder for multi-scale context	Dilated convolutions, spatial pyramid pooling	General, medical segmentation
U-Net	CNN Encoder -Decoder	U-shaped network with skip connections for precise localization	Encoder-decoder with symmetric skip connections	General, medical segmentation

Trait Extraction

Research on Climate Change Response



Long-term trends in global flowering phenology. Williamson et al., 2025. *New Phytologist*

OCR: optical character recognition
NER: named entity recognition

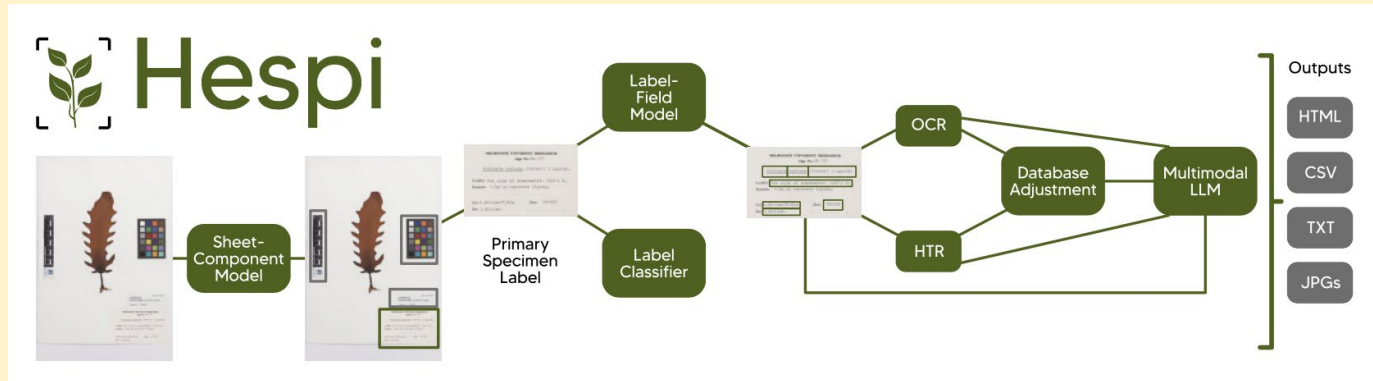
A novel automated label data extraction and database generation system from herbarium specimen images using OCR and NER.

Takano et al. (2024).
Scientific Report 14:112



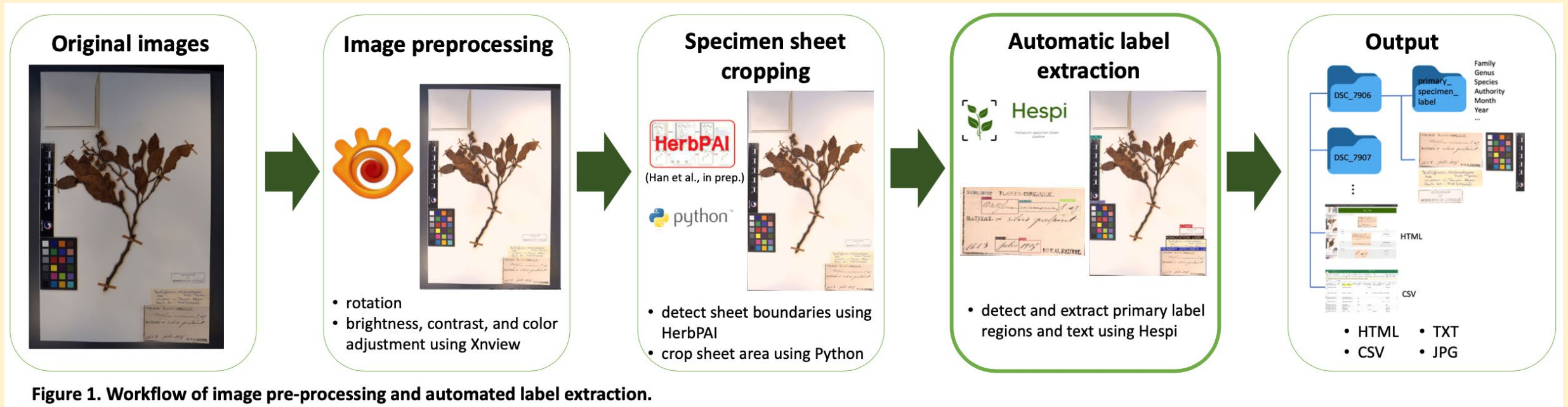
ファイルを選択  C2000003.jpg	OCR ラベリング	Family Suggestion アオサグサ科(Cocculus trilobus (Thunb.) DC.) 入力 科名 Family name (Japanese) ツツジ科
CUTTER MAT OLFA 国産品 人と自然の博物館資料 植物 C2-000003 QPCARD 101 v4 姫短大附属		学名 Scientific name Cocculus trilobus (Thunb.) DC. * Cocculus trilobus (Thunb.) DC. Cocculus trilobus (Thunb.) DC. * 種名 Japanese name アオサグサ アオサグサ アオサグサ 県名 Prefecture (In Japanese) 兵庫県 兵庫県 兵庫県 市町村 City, Town (In Japanese) 神戸市東灘区 神戸市東灘区
HERBARIUM OF SHOEI JR. COLLEGE 標本 標本 アオサグサ Cocculus trilobus (Thunb.) DC. det. N.Kurosaki, Nov. 21, 1997. FLORA OF HYOGO PLANT OF JAPAN Nakano [65] Menispermac.		以下 Streets (In Japanese) 住吉台 高森橋 住吉台 高森橋 県名(英) Prefecture (inEnglish) 市町(英) City, Town (inEnglish)
Loc 標本 アオサグサ ツツジ科 Cocculus trilobus (Thunb.) DC. 採集地 市町村 以下 兵庫県神戸市東灘区住吉台 高森橋 - 渡辺台 (六甲大月地蔵谷)		以下(英) Streets (inEnglish) 国名 Country 年月日 Date of Colletion Sep. 15, 1994 Suggestion suggestion 1994-09-15 入力
(E135° 15' ,N34° 44' 1(300-830 m Alt.) Memo 年 月 日 年月日 Sep. 15, 1994 採集者 備考 Goto Tetsuro No.2409		Do Sep. 15, 1994

HTR (Handwriting Text Recognition)



Hespi: A pipeline for automatically detecting information from herbarium specimen sheets.

Turnbull et al. (2025) *BioScience* 75:637–648



HTR (Handwriting Text Recognition)



성신여자대학교
SUNGSHIN WOMEN'S UNIVERSITY

AI-assisted Extraction of Handwritten Label Information from Approximately 6,000 Historic Korean Peninsula Specimens Collected by Émile Taquet and Urbain Faurie

Hyeonji Moon^{P1}, Kae Sun Chang², Hanna Park² and Sangtae Kim¹

¹Department of Biology, Sungshin Women's University, Seoul 01133;

²Korea National Arboretum, Pocheon 11186

amborella@sungshin.ac.kr



산림청 국립수목원
Korea National Arboretum



<https://irfa.paris/en/missionnaire/1165-faurie-urbain/>

Urbain J. Faurie

(1847–1915)

A French missionary botanist who contributed to the early botanical exploration of Korea through plant collections from Jeju Island and northern Korea in the early 20th century.

Introduction

Historical specimens and digitization challenges

- Early 20th-century herbarium specimens collected from the Korean Peninsula by **Urbain J. Faurie** and **Émile J. Taquet** represent important historical biodiversity records.
- These specimens contain valuable handwritten label data (locality, collection date, collector, and taxonomic identity), but digitizing them remains challenging due to cursive and highly variable handwriting.

AI-based recognition approach (Figure 1)

- Recent advances in Optical Character Recognition (OCR), Handwriting Text Recognition (HTR), and multimodal large language models (LLMs) provide new opportunities for extracting information from historical specimen labels.
- **Hespi** enables automated detection of specimen sheet components and label fields, followed by text extraction and refinement using OCR, HTR, and LLM-based correction (Turnbull et al., 2025).

Aim of this study

- This study applied Hespi to approximately **6,000 digitized historical plant specimens** collected by **Taquet** and **Faurie** from the **Korean Peninsula** and housed at the **Royal Botanic Garden Edinburgh (E)**.
- We evaluated its utility for constructing a structured database of handwritten label information from historical herbarium collections.

Materials & Methods

Dataset

- A total of **5,935 images** of herbarium specimens photographed at the **Royal Botanic Garden Edinburgh (E)** were used in this study.
- The dataset mainly comprised **early 20th-century** collections by **Taquet and Faurie**, including 3,741 Taquet records and 1,708 Faurie records.

Image preprocessing using XnView

- Image orientation was standardized to position specimen primary labels in the lower-right corner.
- Images with low brightness or bluish color casts were corrected by adjusting brightness, contrast, and color temperature using XnView MP (version 1.9.3).

Specimen sheet cropping using HerbPAI

- Specimen sheet boundary coordinates were detected using HerbPAI (<https://github.com/Sujeong-Han/HerbPAI>; Han et al., in prep.), an image preprocessing tool for herbarium specimens.
- Specimen sheet regions were cropped based on the detected coordinates using a custom Python script.

Specimen-label information extraction using Hespi (Figure 2)

- Specimen-label information, including handwritten text, was automatically detected and extracted using Hespi (<https://github.com/rbturnbull/hespi>; Turnbull et al., 2025).
- Hespi was **run with OpenAI API** (gpt-4o) integration to refine the text-recognition results.
- Images were processed in batches of 1,000 to ensure stable analysis and efficient result management.

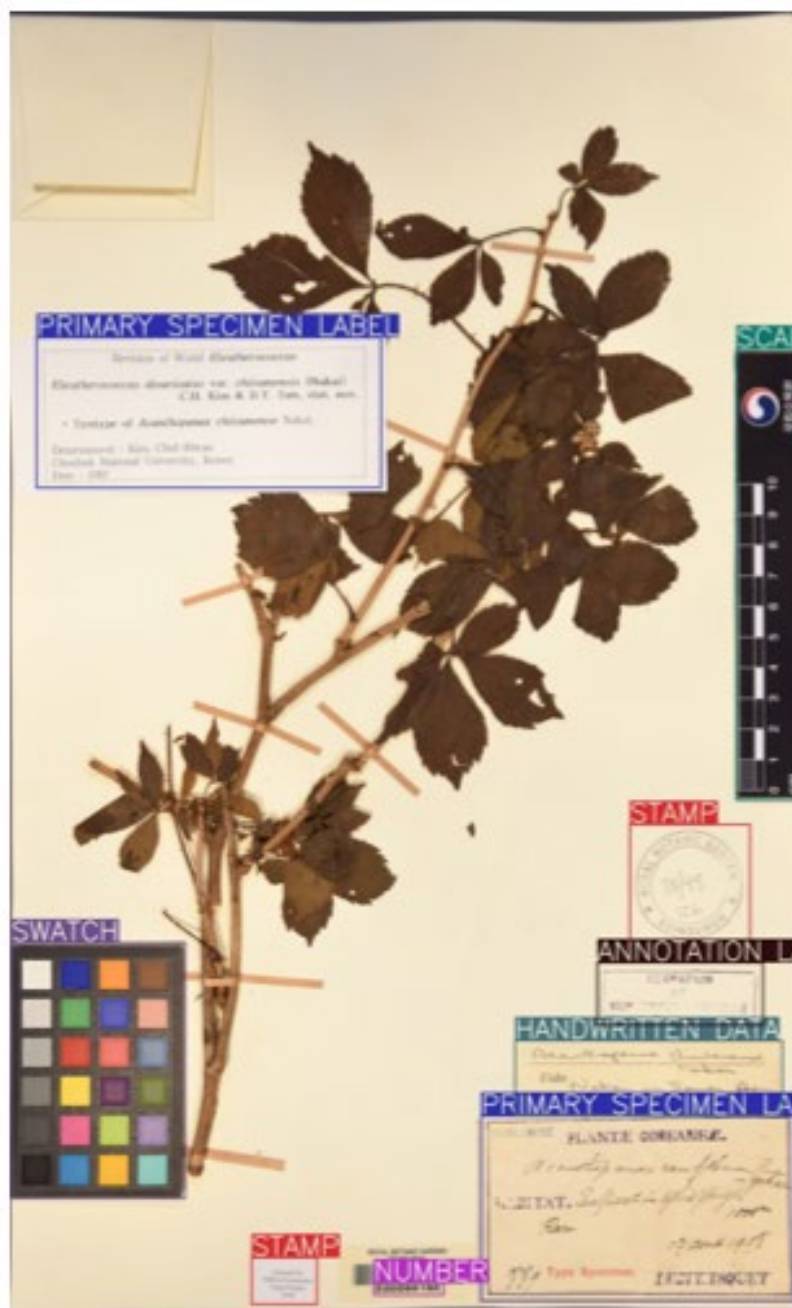


<https://irfa.paris/en/missionnaire/2329-taquet-emile/>

Émile J. Taquet

(1873–1952)

A French missionary botanist based on Jeju Island, where he extensively collected plants, especially around Mt. Halla, providing important materials for the study of Korean flora.



DSC_0793

Primary Specimen Label

Primary Specimen Label 2

Sheet Components



Classification handwritten

Field

Image

Text

family

genus

Acanthopanax

Acanthopanax

Thesarak: Acanthopanax, acnos → Acanthopanax, acnos

TrOCR: Acanthopanax → Acanthopanax

|| M: Acanthopanax

species

senticosum

senticosum

TrOCR: assiflorum → cressiflorum [0.857]

LLM: senticosum

genus

species

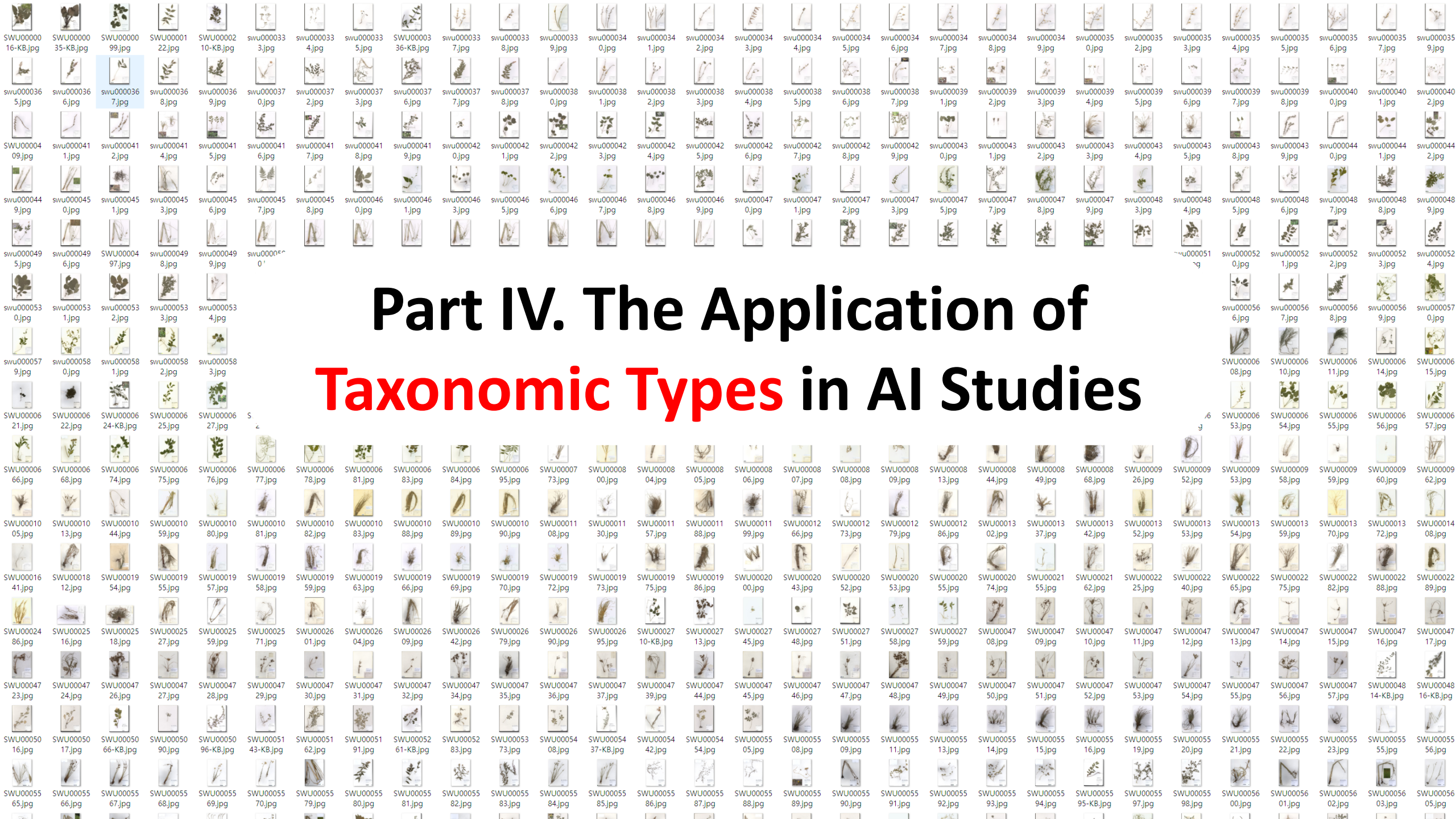
Acanthopanax

senticosum

Detection of mis-Identified specimen

Check out the presentation by Prof. Takashi Akihiro today





Part IV. The Application of

Taxonomic Types in AI Studies

The **concept of species** is one of the most fundamental and still unresolved questions in biology.



Type

- Morphological (typological) species concept
- Biological species concept
- Evolutionary species concept



Plato

BC 828-347, Greek philosopher

From Platonic ideal form to **typological species concept**

- Ideal Form / **Archetype**
 - Imperfect real-world manifestations of the ideal form
 - Recognition of **representative forms**
 - **Typological species concept**
- Species were historically often interpreted through the idea of a representative form, but modern type specimens are **nomenclatural anchors**, not ideal archetypes.

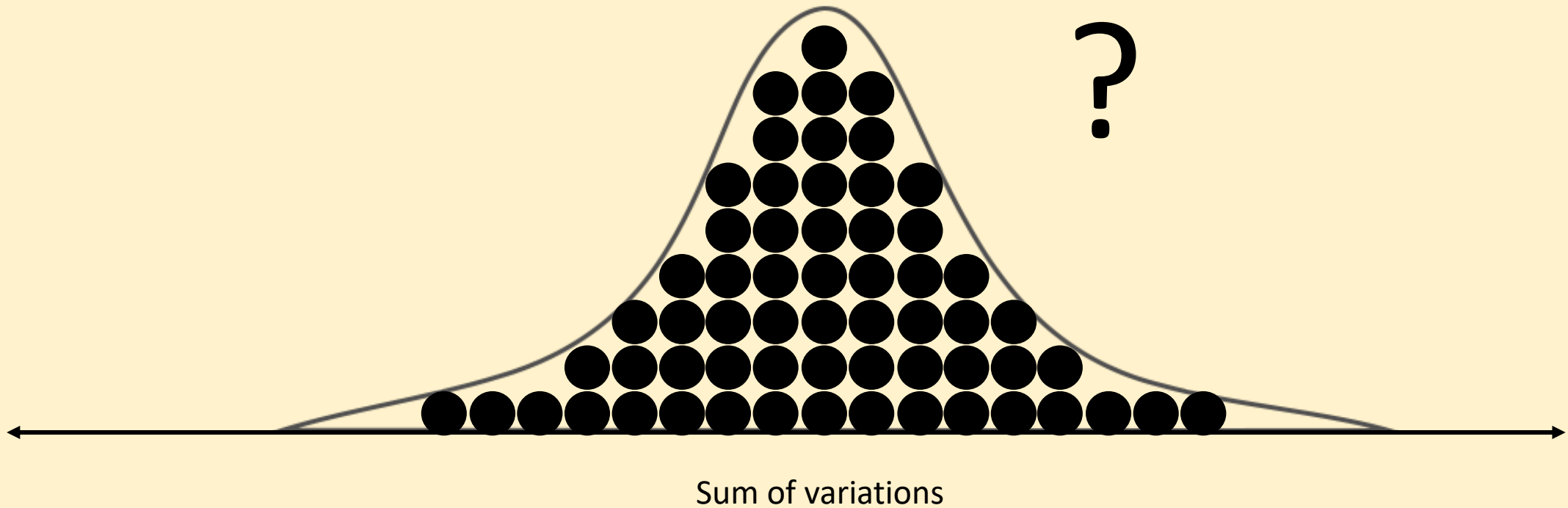
Core taxonomic dilemma

Species are variable, but taxonomy depends on a **single type**

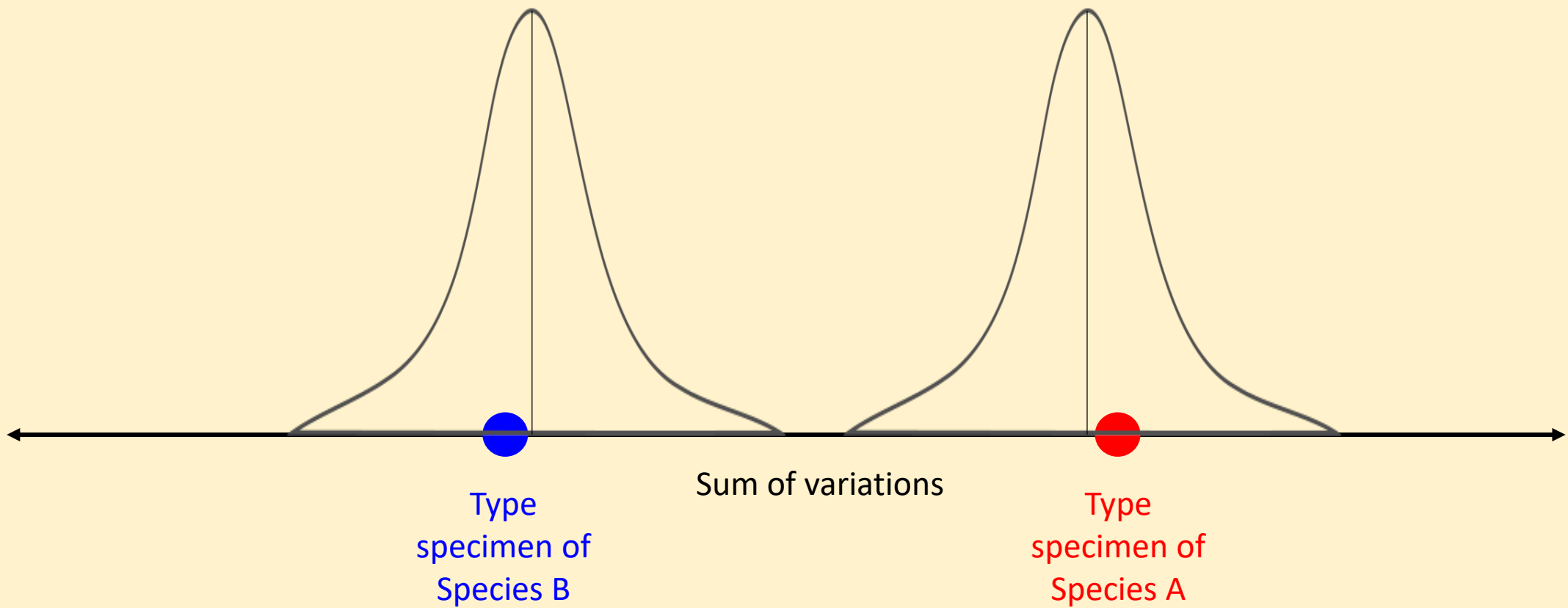
- A natural species is not a single ideal form.
- Individuals vary in morphology, geography, ecology, phenology, and genetic background.
- **No single specimen can fully represent this multi-dimensional variation.**
- Nevertheless, taxonomy requires a fixed reference for each name.
- Central dilemma:

The type stabilizes the name, but the species extends beyond the type.

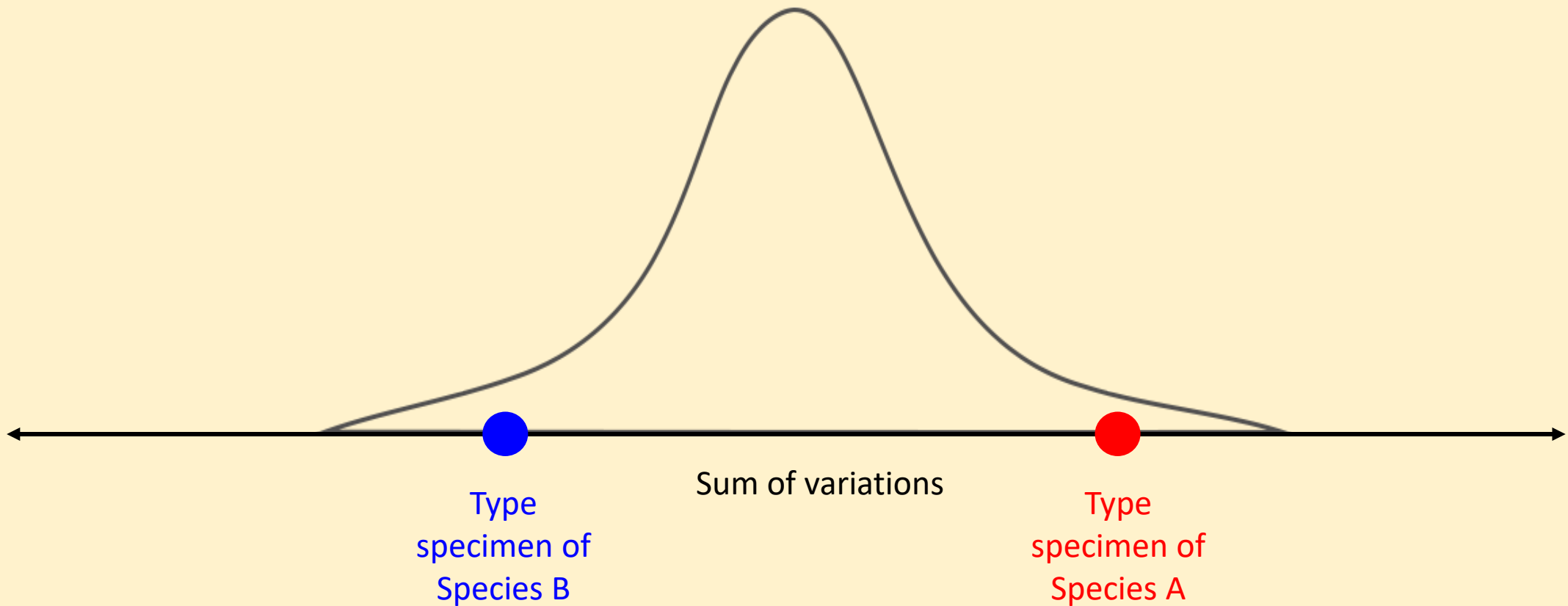
Normal distribution of species variation



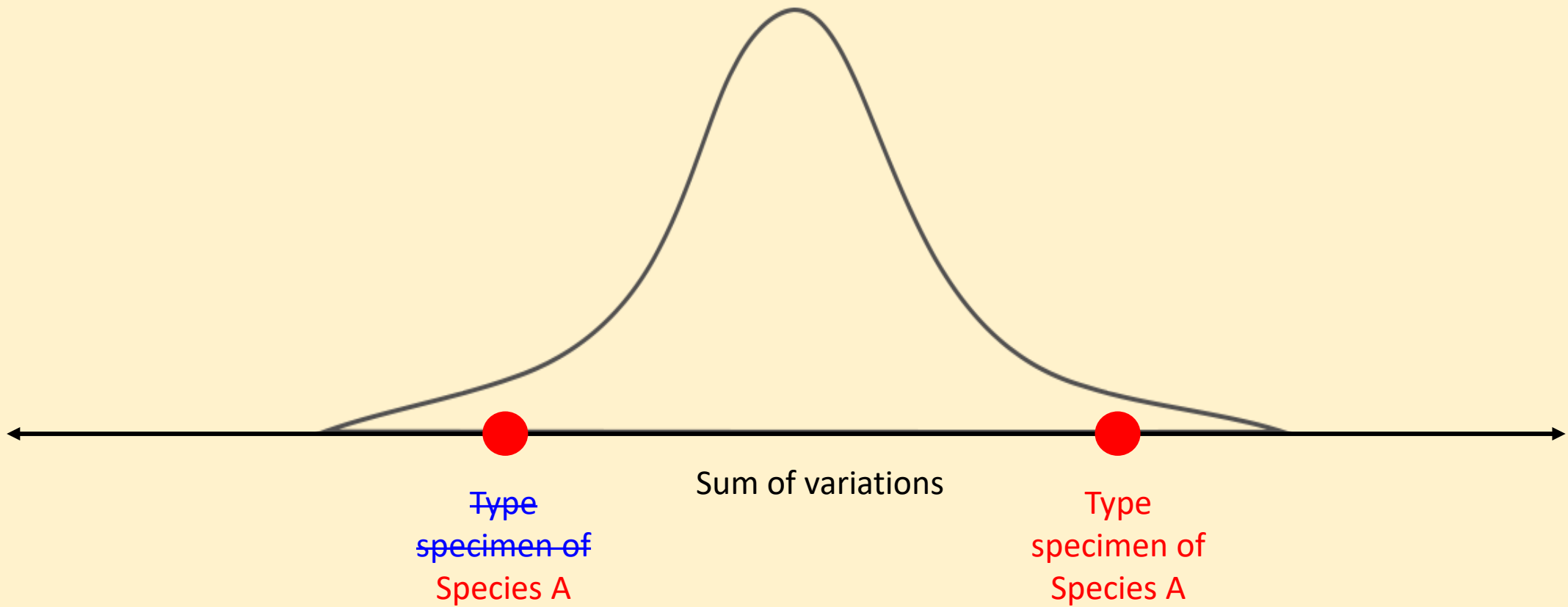
Normal distribution of species variation



Normal distribution of species variation



Normal distribution of species variation



Where is the **type specimen** located within the species-level **morphospace**?

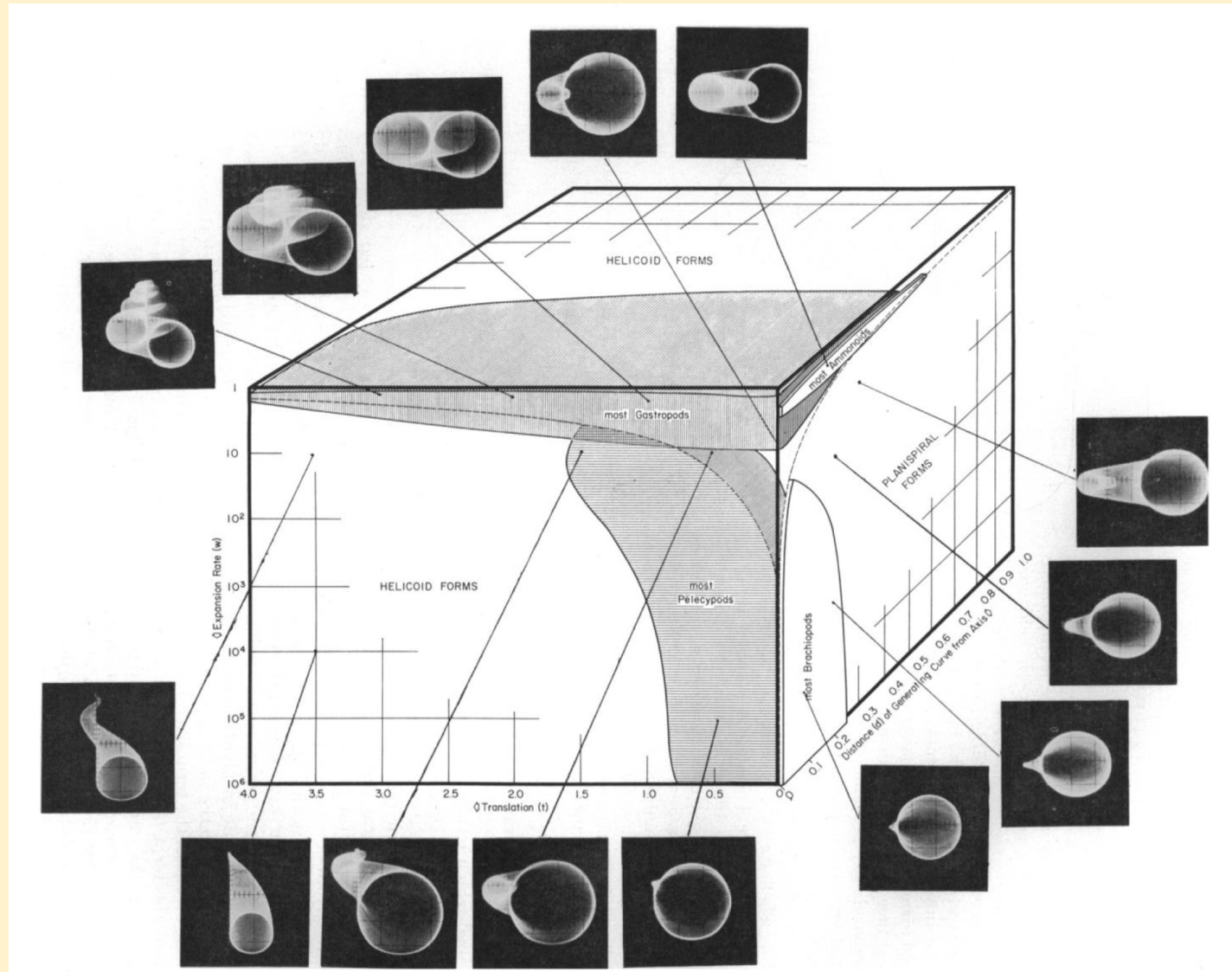
Morphological representation space

(**morphospace** s.l.: Raup, 1966; Budd 2021; Cavanagh et al., 2021)

Viewing species as morphospaces

- Each specimen can be represented as a point in a **multi-dimensional morphospace**.
- A species can be viewed as a distribution of specimen points.
- The center of this distribution can be interpreted as the **global mean (GM)**.
- A taxonomic type may occur near the center, at the margin, or in a distinct region.

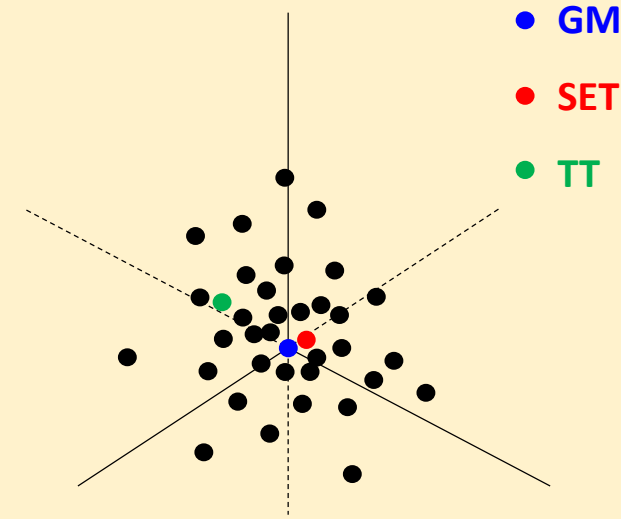
Morphospace



morphospace s.l.: Raup, 1966; Budd 2021; Cavanagh et al., 2021

Terms for AI-Derived Morphospace study

- **Global Mean (GM):** The center of intraspecific variation.
- **Taxonomic Type (TT):** The Linnean type specimen.



NEWLY PROPOSED TERMS IN THIS STUDY

- **Selected Encoding Type (SET):** The real specimen closest to GM.
- **Reconstructed Decoding Type (RDT):** A hypothetical image decoded from GM.

⌘ **Remember:** SET and RDT do not replace the taxonomic type.
They help **evaluate** its morphological position.

In this study about the Korean *Viola*:

- **Target group:** 17 Korean species of *Viola*
- **Data:** More than 500 digitized herbarium images per species
- **Method:** Autoencoder-based unsupervised image analysis
- **Main comparison:** GM-TT, GM-SET, and SET-TT
- RDT? Next time!

Materials

Summary of digitized herbarium images of 17 Korean *Viola* species used for MIWAE-based morphospace analysis

Taxon	# of image							Total	Type specimen		
	KH	NIBR	SWU	SKK	KWNU	iDigBio	CVH		category	Herbarium	Specimen #
<i>Viola acuminata</i>	537	0	63	0	0	0	0	600	lectotype	LE	LE 1011514
<i>Viola albida</i>	473	37	87	3	0	0	0	600	isotype	K	K000254246
<i>Viola hamiltoniana</i>	559	4	29	0	0	0	0	592	probable syntype	BM	BM000521643
<i>Viola biflora</i>	25	0	0	2	0	500	0	527	lectotype	LINN	LINN 1052.16
<i>Viola chaerophylloides</i>	578	0	22	0	0	0	0	600	lectotype	G	G00189226
<i>Viola collina</i>	483	16	34	0	0	0	0	533	probable syntype	S	S11-33244
<i>Viola keiskei</i>	440	22	86	0	0	0	0	548	holotype	L	L0175365
<i>Viola mandshurica</i>	302	85	200	5	8	0	0	600	Syntype (isolectotype or paralectotype)	P	P04672354
<i>Viola mirabilis</i>	4	0	0	9	0	500	0	513	lectotype	LINN	LINN 1052.15
<i>Viola orientalis</i>	523	0	25	0	0	0	0	548	paratype	M	M-0112818
<i>Viola phalacrocarpa</i>	306	92	78	124	0	0	0	600	lectotype	LE	LE 01017336
<i>Viola philippica</i>	219	177	76	99	1	0	0	572	isolectotype	MA	MA82816
<i>Viola rossii</i>	472	38	88	2	0	0	0	600	syntype	K	K000254283
<i>Viola selkirkii</i>	125	132	14	94	0	170	0	535	lectotype	K	K000370205
<i>Viola sororia</i>	15	37	10	15	0	473	0	550	lectotype	BM	B-W4898
<i>Viola tokubuchiana</i> var. <i>takedana</i> *	164	175	21	194	0	0	0	554	Syntype (isolectotype or paralectotype)	TI	TI00082947
<i>Viola variegata</i>	290	199	17	0	0	0	0	506	Syntype (isolectotype or paralectotype)	G	G00208825

* Used *V. tokubuchiana*
type

Total ca. 10,000 images

V. acuminata
줄방제비꽃



V. albida
태백제비꽃



V. hamiltoniana
콩제비꽃



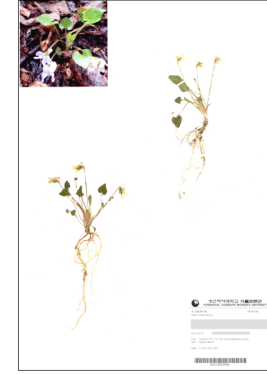
V. biflora
장백제비꽃



V. chaerophylloides
남산제비꽃



V. collina
등근털제비꽃



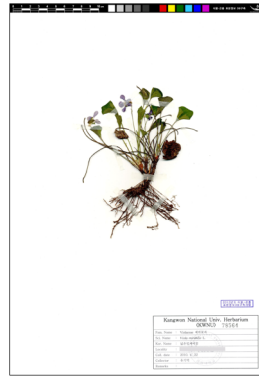
V. keiskei
잔털제비꽃



V. mandshurica
제비꽃



V. mirabilis
넓은잎제비꽃



V. orientalis
노랑제비꽃



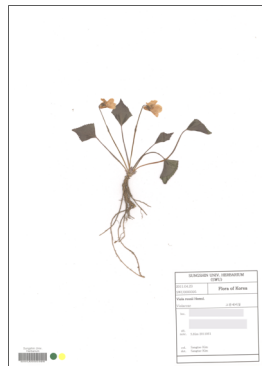
V. phalacrocarpa
털제비꽃



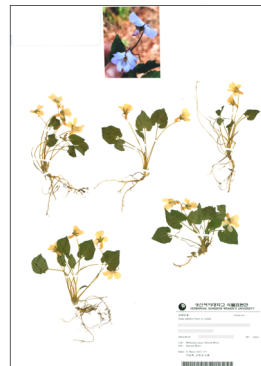
V. philippica
호제비꽃



V. rossii
고깔제비꽃



V. selkirkii
외제비꽃



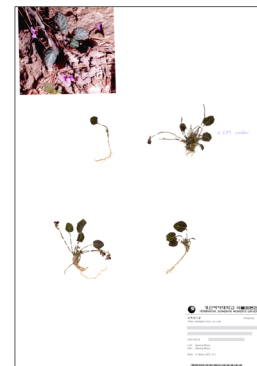
V. sororia
종지나물



V. tokubuchiana var. *takedana*
민둥외제비꽃



V. variegata
알록제비꽃



Materials

2025. 2. 한국식물분류학회 정기학술대회 발표



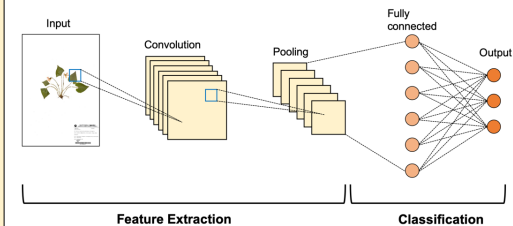
Development of an AI-powered herbarium image pre-identification system in Violaceae: focusing on Korean species



Hyeonji Moon^{p1}, Su-Jeong Han², Jaesung Lee², and Sangtae Kim¹

¹Department of Biology, Sungshin Women's University, Seoul, Rep. of Korea;

²Department of Artificial Intelligence, Chung-Ang University, Seoul, Rep. of Korea



박사과정 문현지

Resampling methods tested

- 1) Photoshop (model unspecified)
- 2) Lanczos model (XnView)
- 3) Lanczos-gamma model (XnView)
- 4) Bilinear model (XnView)
- 5) MitChell (XnView)
- 6) Lanczos2 model (XnView)
- 7) Lanczos model (Python script)



Lanczos model (XnView)



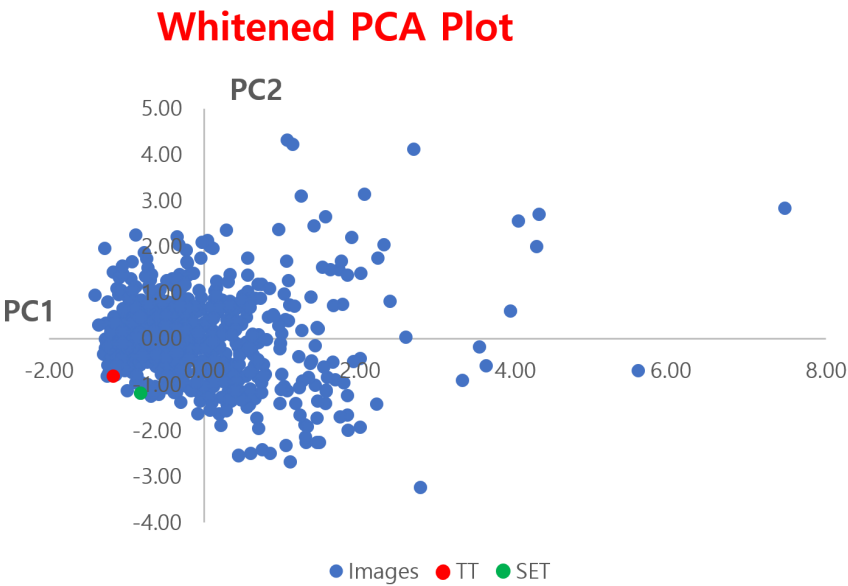
Lanczos model (Python script)

→ **Lanczos resampling implemented in a Python script is selected**

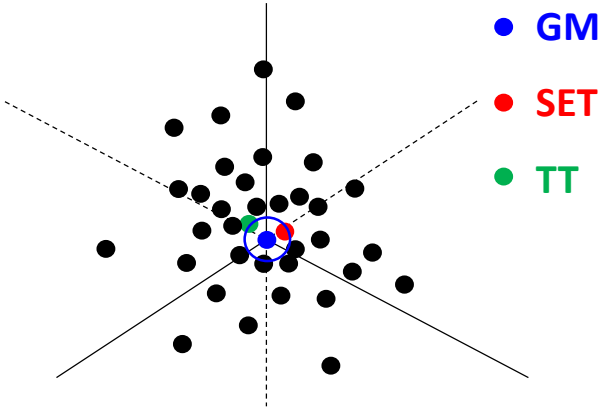
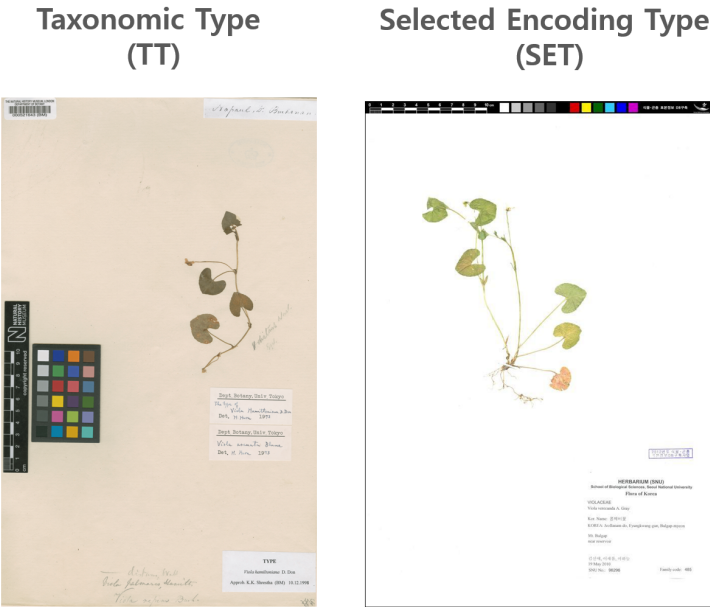
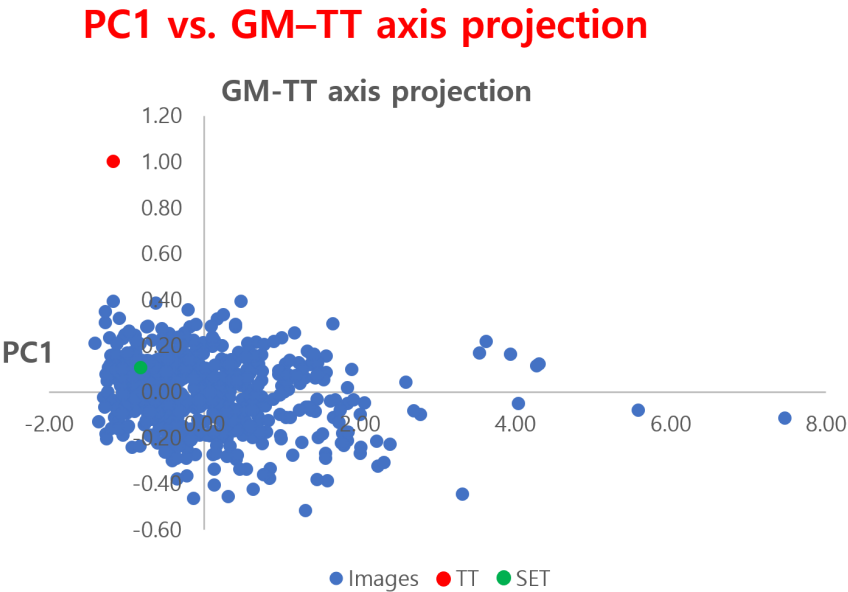
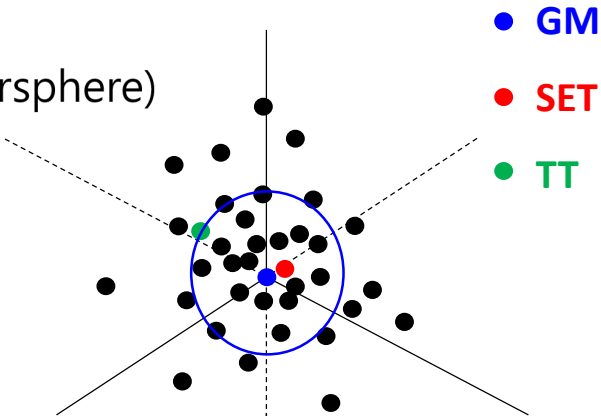
- It produces minimum black-dot artifacts around the segmented plant image.
- Photoshop also produced good results, but it was not suitable for automated processing of large image datasets.

An example of GM-TT distance evaluation

Viola hamiltoniana 콩제비꽃



- GM-TT: 6.26
- GM-SET: 5.72
- GM-TT # of images: 7
(No. of images within the GM-TT hypersphere)
- Total images: 592
- **GT-index: 2.7 %**



Summary of MIWAE autoencoder analyses

Class_ID	Scientific Name	Comm. Name	^a GM-TT	GM-SET	GM-TT # of images	Total images	GT-index (%)	SET = TT
V01	<i>Viola acuminata</i>	줄방제비꽃	5.26	5.26	0	600	0.0	O
V02	<i>Viola albida</i>	태백제비꽃	4.89	4.89	0	600	0.0	O
V04	<i>Viola hamiltoniana</i>	콩제비꽃	6.26	5.72	7	592	1.2	
V07	<i>Viola biflora</i>	장백제비꽃	7.98	5.18	69	527	13.1	
V10	<i>Viola chaerophylloides</i>	남산제비꽃	4.27	4.27	0	600	0.0	O
V11	<i>Viola collina</i>	둥근털제비꽃	5.18	5.18	0	533	0.0	O
V17	<i>Viola keiskei</i>	잔털제비꽃	6.55	5.03	8	548	1.5	
V20	<i>Viola mandshurica</i>	제비꽃	7.83	5.22	61	600	10.2	
V21	<i>Viola mirabilis</i>	넓은잎제비꽃	8.19	5.51	42	513	8.2	
V22	<i>Viola orientalis</i>	노랑제비꽃	10.39	4.58	255	548	46.5	
V26	<i>Viola phalacrocarpa</i>	털제비꽃	6.46	6.18	3	600	0.5	
V27	<i>Viola philippica</i>	호제비꽃	8.74	4.66	100	572	17.5	
V30	<i>Viola rossii</i>	고깔제비꽃	5.03	5.03	0	600	0.0	O
V31	<i>Viola selkirkii</i>	외제비꽃	6.15	4.86	10	535	1.9	
V33	<i>Viola sororia</i>	종지나물	6.15	5.67	5	550	0.9	
V34	<i>Viola tokubuchiana</i> var. <i>takedana</i>	민둥외제비꽃	7.11	5.46	15	554	2.7	
V36	<i>Viola variegata</i>	알록제비꽃	6.36	4.67	27	506	5.3	

^a Multiple TT: the nearest one selected

GT-index

- Top five closest: 줄방제비꽃(0), 태백제비꽃(0), 남산제비꽃(0), 둥근털제비꽃(0), 고깔제비꽃(0) GM-TT
- Bottom three most distant: 노랑제비꽃(46.5), 호제비꽃(17.5), 장백제비꽃(13.1) GM-----TT

Potential Applications of GT-index Analysis + SET

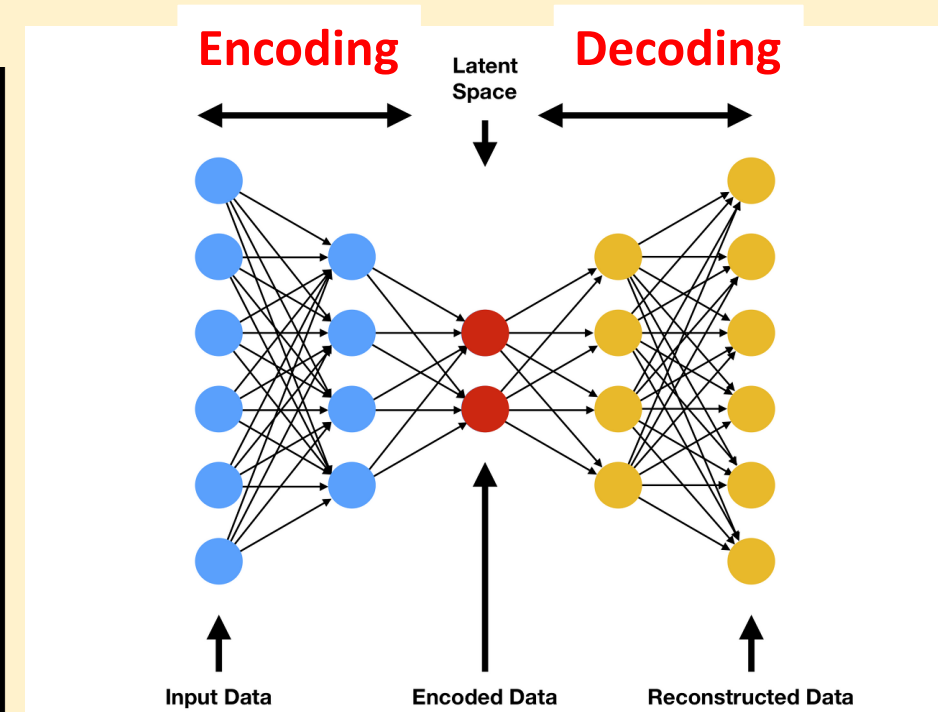
- Supporting **lectotypification** decisions
- Assisting **neotype** or **epitype** designation
- Identifying morphologically **atypical type specimens**
- Prioritizing taxa for further **taxonomic revision**
- Detecting possible **hidden intraspecific structure**
- **Selecting** representative specimens for **AI training datasets**
- Identifying representative specimens for **education** and **public outreach**

⌘ Note: **GT-index** and **SET** provide **supplementary** morphological **evidence**, not a substitute **for** nomenclatural and taxonomic **judgment**

Further Study

- Address **Reconstructed Decoding Type (RDT)**

Reconstruction of an Average Korean Face



Autodecoding test using LLM AIs

Original images



Prompt: 위의 네 이미지에서 식물의 각각의 부분들을 인식하여 이들의 평균적인 모습을 재구성하여 하나의 식물이미지를 만들어 주시오.

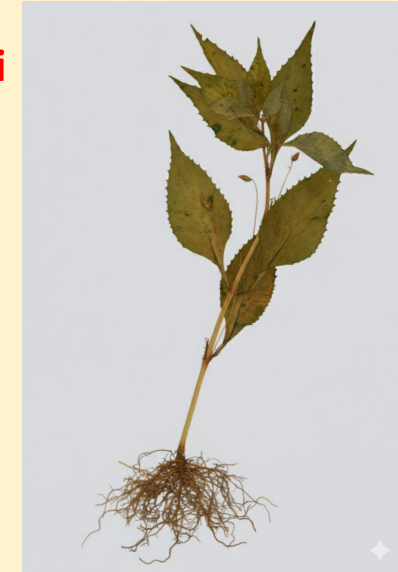


식물을
감싼
밴드는
제거해
주세요



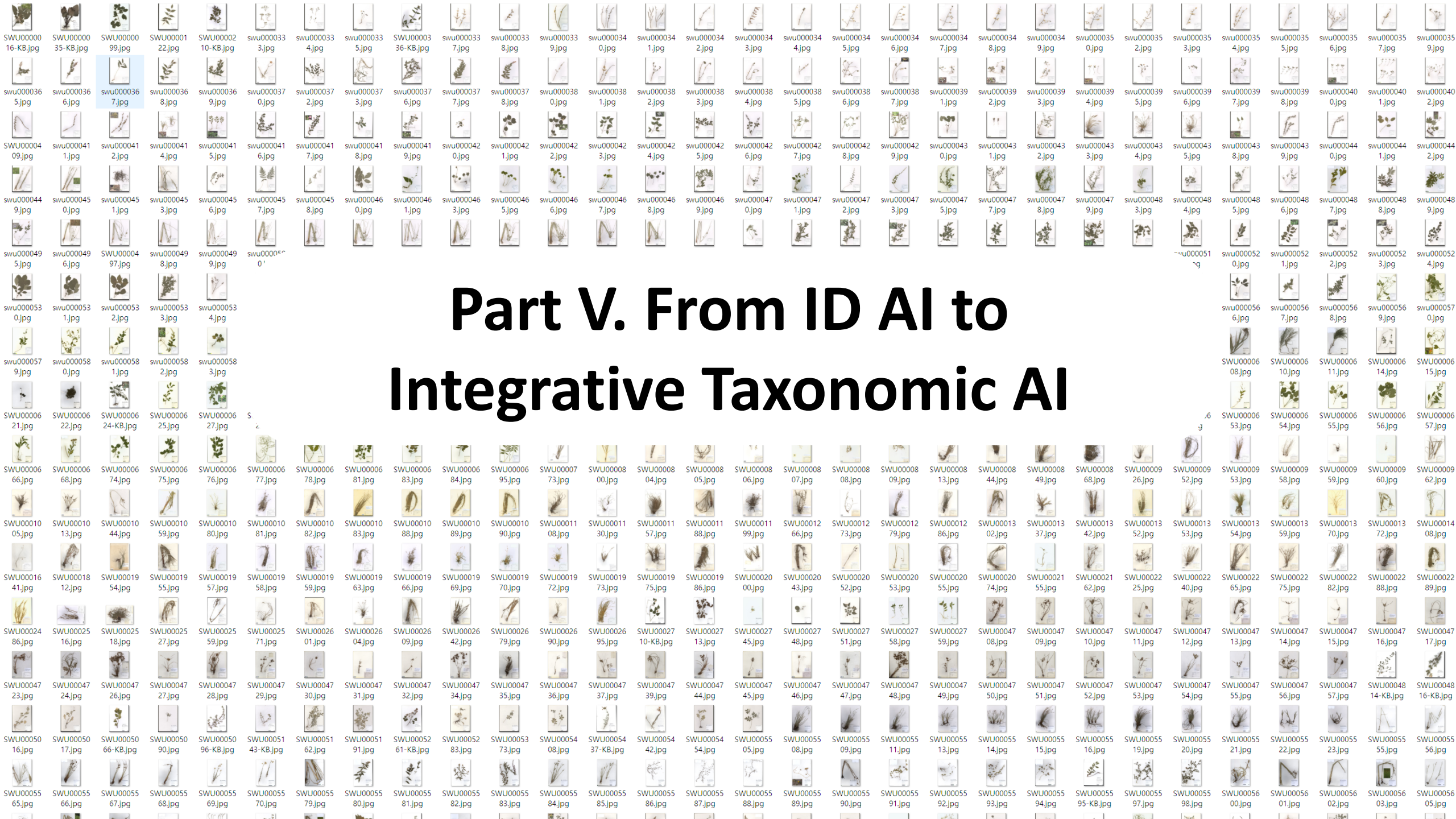
ChatGPT5

Gemini

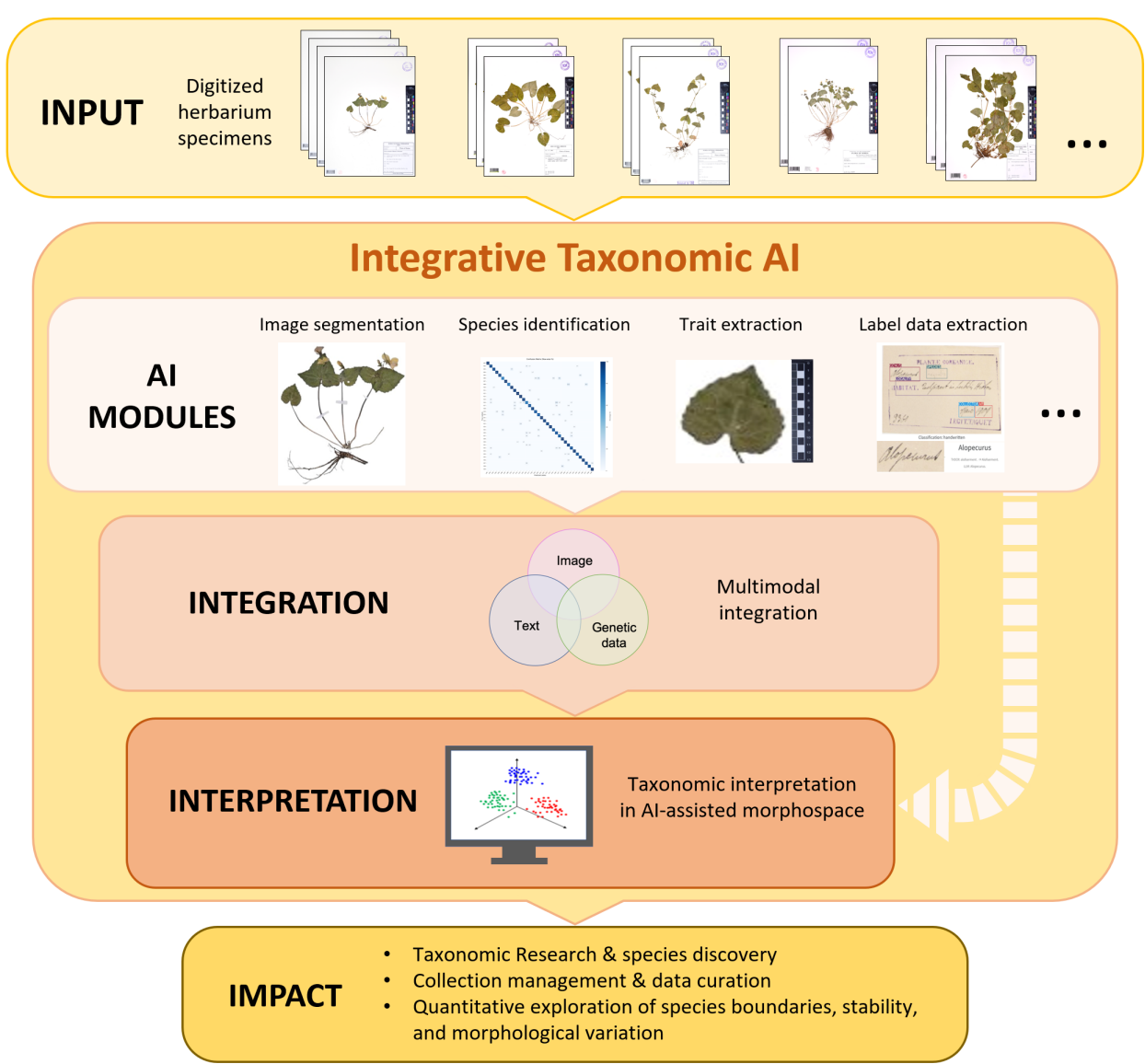


Key Taking-Home Messages from the GT-index Study

**The type fixes the name;
the morphospace shows
how typical the type is.**



The Age of AI: Even in Plant Systematics!



Category	ID AI (Traditional species identification AI)	Integrative Taxonomic AI (Paradigm of the present study)
Definition	Species identification assisted by AI	<u>Taxonomic interpretation within an AI-assisted morphospace</u> , integrating AI-driven taxonomic data curation
Core question	How accurately can a species label be predicted from a given specimen image?	How can species boundaries, stability, and morphological variation be quantitatively represented, explored, and validated within a morphological representation space?
Role of AI	An automated decision-making system that directly outputs the final species identification	A <u>higher-level decision-support framework</u> that integrates identification outputs with AI-driven taxonomic data curation to interpret morphological structure and evaluate species boundaries
Scale of inference	Individual specimen-level inference	Multi-specimen and species-level inference within a morphospace

Conceptual framework of **Integrative Taxonomic AI**

Takano, Moon, ... Kim et al. (In press). *Journal of Plant Research*

Multimodal Approach

Check out the presentation by Prof. Atsuko Takano today



The central thesis

핵심 논지 主な論点

Identification AI predicts labels.

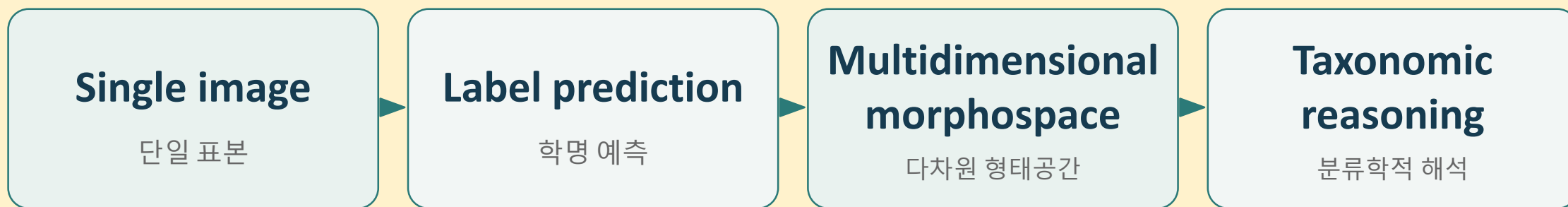
동정 AI는 표본의 이름을 예측한다.

識別AIはラベルを予測します。

Integrative Taxonomic AI interprets variation and taxonomic meaning.

통합 분류학적 AI는 변이와 분류학적 의미를 해석한다.

統合的分類学AIは、多様性を解釈する。



Paradigm change: from final-answer AI to expert-guided decision support

패러다임 변화: 최종 답을 내는 AI에서 전문가 주도 의사결정 지원 도구로

パラダイムシフト:「最終的な答えを出すAI」から「専門家の指導に基づく意思決定支援」へ

Thanks to: Members of the Korea-Japan taxonomic AI team and

