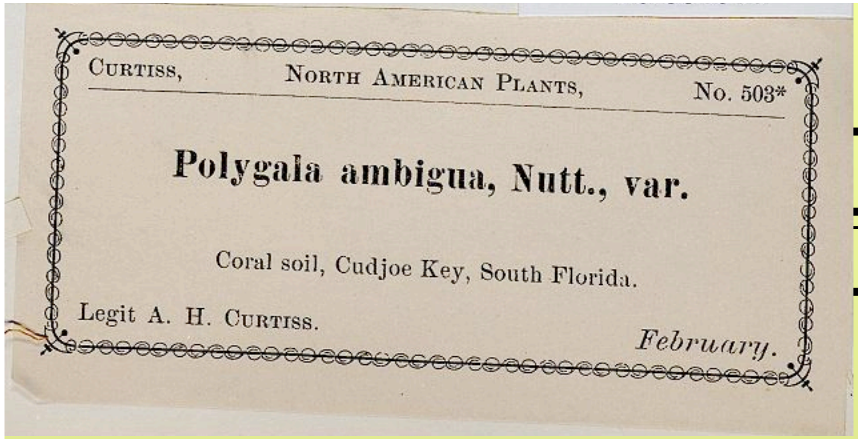


CURTISS, NORTH AMERICAN PLANTS, No. 503*

Polygala ambigua, Nutt., var.

Coral soil, Cudjoe Key, South Florida.

Legit A. H. CURTISS. *February.*

```
…<co> Curtis,  </co><hdlc>          North American Pl
</hdlc><cnl> No.</cnl><cn> 503*</cn>
<gn> Polygala</gn><sp> ambigua,</sp><sa> Nutt.,</
sa><val> var.</val>
<hb> Coral soil,</hb><lc> Cudjoe Key, South Florida.
</lc><col> Legit</col><co> A. H. Curtiss.</
co><dt>February</dt>…
```

Automatic Metadata Extraction (Darwin Core) From Museum Specimen Labels

P. Bryan Heidorn, Qin Wei

University of Illinois at Urbana-Champaign

The problem

- >1 Billion Natural History Specimens
- Collected over 250 years / many languages
- No publishing standards
- Near infinite classes
  - Your high school teacher lied
- 6 min / label * 1B labels = 100M hours
- Saving 1 min = 16.7 Million hours
- $10/hr = $167,000,000
- 1/4790 of U.S. deregulation financial bailout

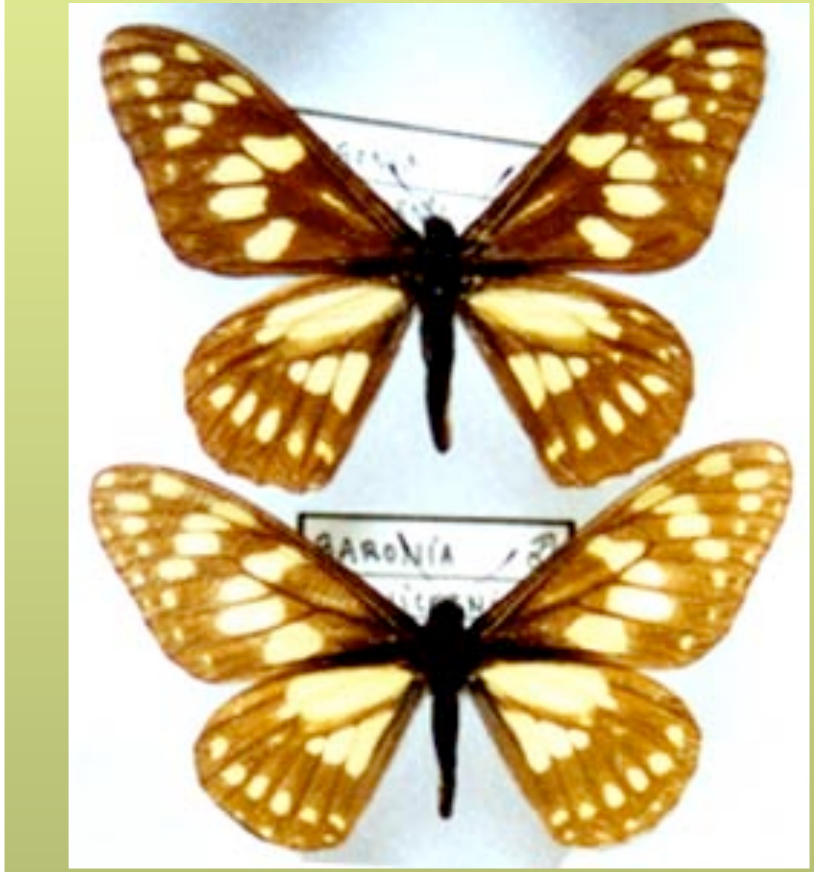
Why care

- Historic distribution of species
- Ecological niche modeling (invasiveness, crop hardiness, pest potential)
- Projections of the impact of climate change
- Where did explorer go? (for error detection)

About the specimens!

- Will I see a Kirkland Warbler here?
- Do tamaracks grow in sand?
- When did Linden trees bloom before the industrial revolution?

A real-life example: *Baronia brevicornis* and its single food plant, *Acacia cochliacantha (Soberon)*



DCMA 2008

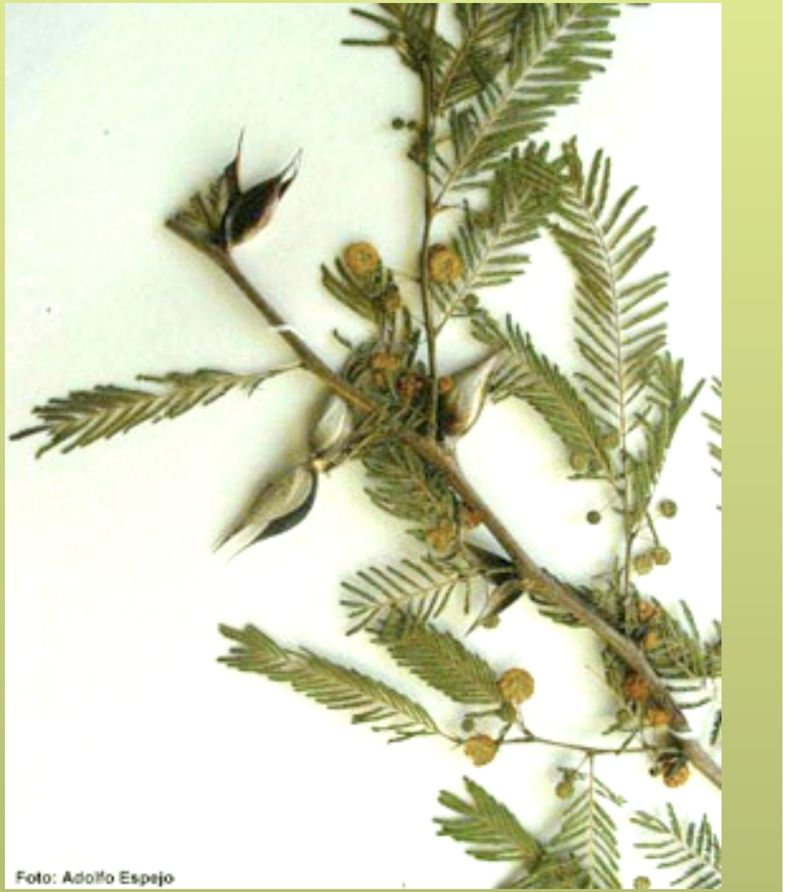
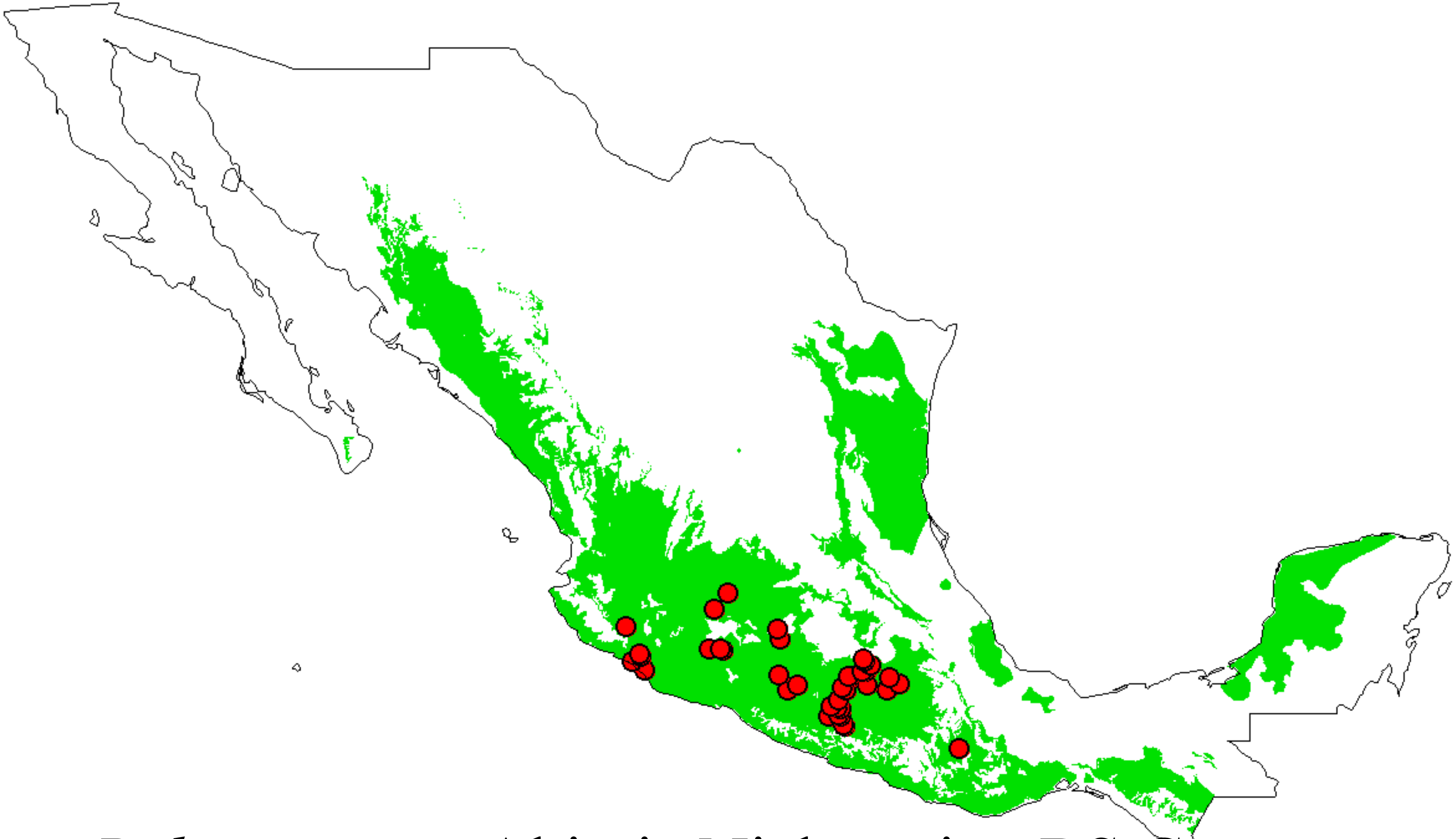


Foto: Adolfo Espejo

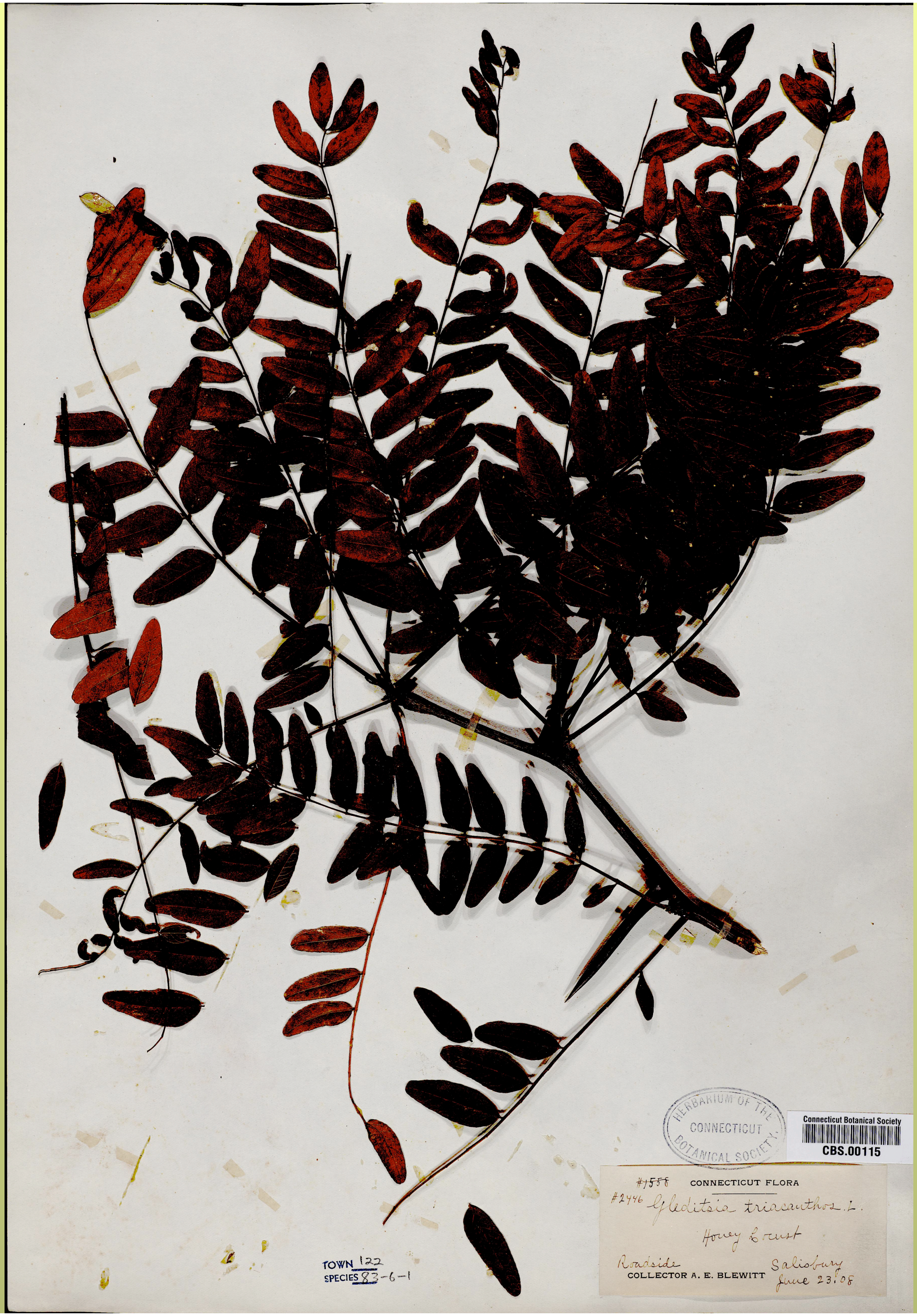


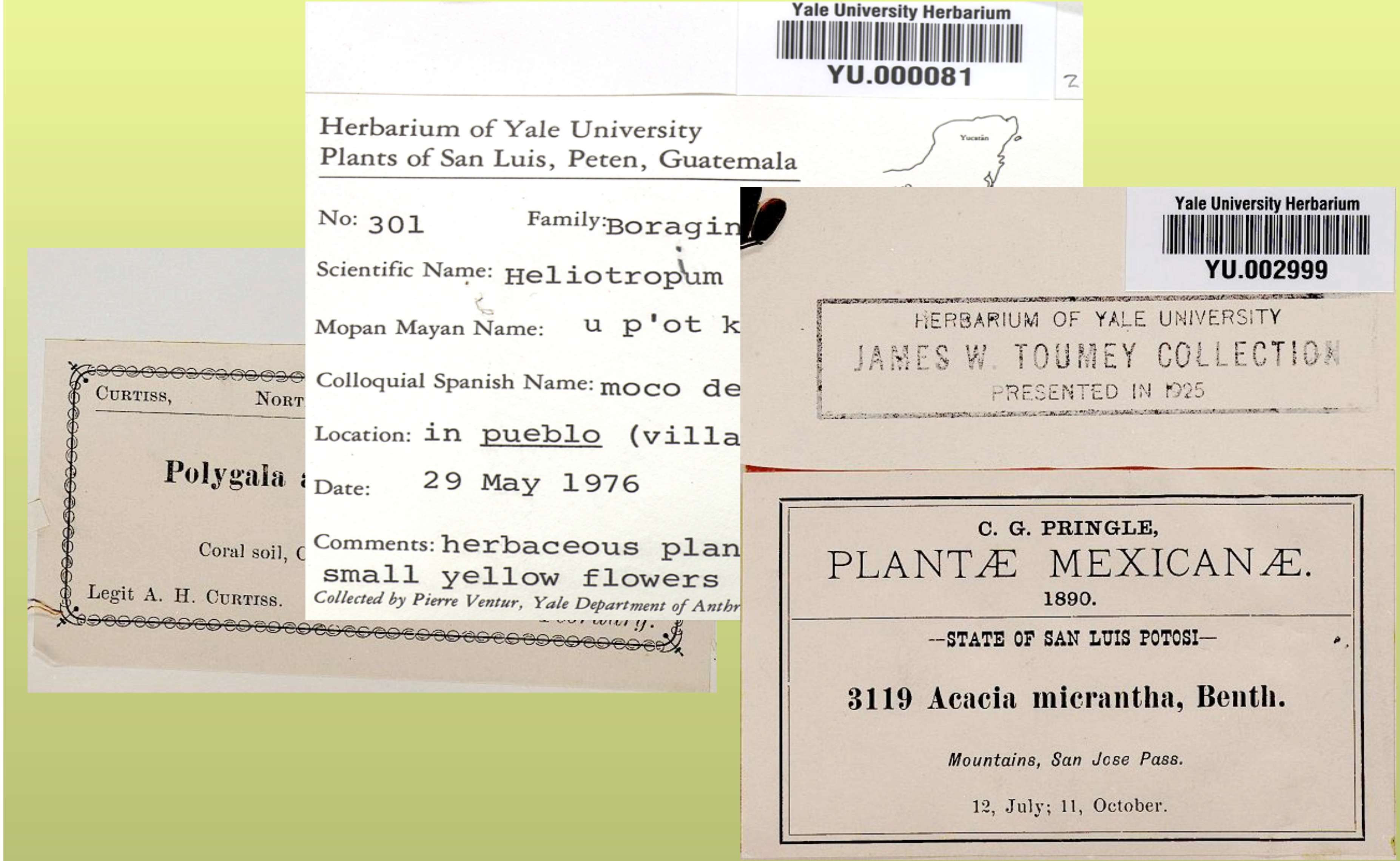
B. brevicornis Abiotic Niche using BS Garp

II: Estimating the “Area of Accesibility” (Soberon)

- From where? What is the initial condition?
- At what scale? In relation to what vagility parameters?
- At certain scales, one can assume that biogeography is a good surrogate for the accesibility areas, this is, we assume that if a species is present in a given biogeographical region, it can reach all of it.

Natural History Specimens


HERBARIUM OF THE CONNECTICUT BOTANICAL SOCIETY.
Connecticut Botanical Society
CBS.00115
~~#1558~~ CONNECTICUT FLORA
#2446 Gleditsia triacanthos L.
Honey Locust
Roadside
Salisbury
COLLECTOR A. E. BLEWITT
June 23.08
TOWN 122
SPECIES 83-6-1



Curtiss, Nort

Polygala a

Coral soil, C

Legit A. H. Curtiss.

ruary.

Yale University Herbarium

YU.000081

z

Herbarium of Yale University
Plants of San Luis, Peten, Guatemala

No: 301 Family: Boragin

Scientific Name: Heliotropum

Mopan Mayan Name: u p'ot k

Colloquial Spanish Name: moco de

Location: in pueblo (villa

Date: 29 May 1976

Comments: herbaceous plan
small yellow flowers

Collected by Pierre Ventur, Yale Department of Anthr

Yale University Herbarium

YU.002999

HERBARIUM OF YALE UNIVERSITY
JAMES W. TOUMEY COLLECTION
PRESENTED IN 1925

C. G. PRINGLE,
PLANTÆ MEXICANÆ.
1890.

--STATE OF SAN LUIS POTOSI--

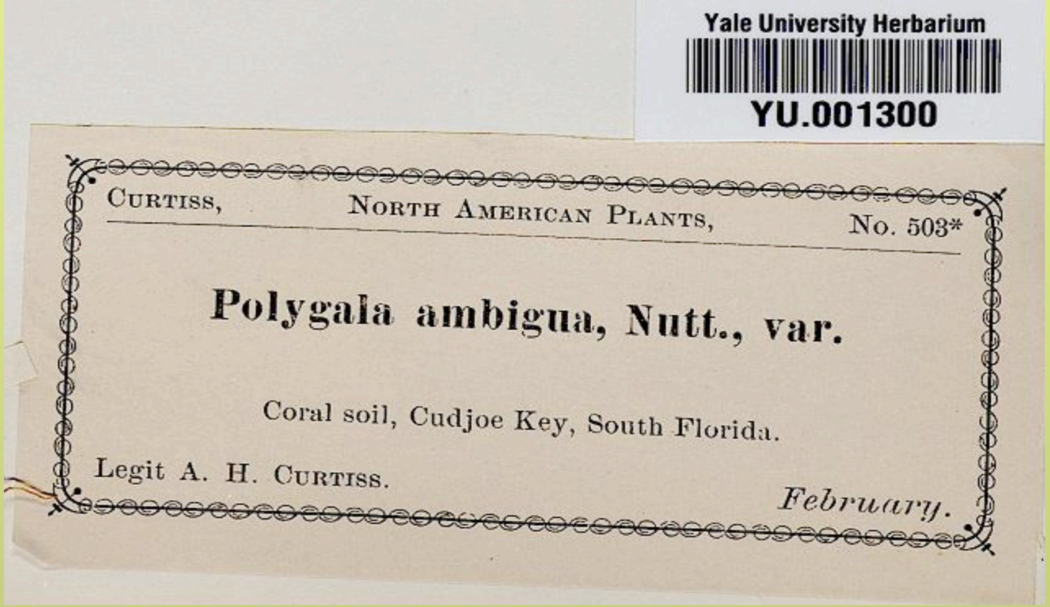
3119 Acacia micrantha, Benth.

Mountains, San Jose Pass.

12, July; 11, October.

Sample OCR Output

Yale University Herbarium
~r-^""" r-n-------
YU.001300
Curtisb, North American Pl
C^o.nr r^-n
ANTS,
No. 503* "^
Polygala ambigna, Nntt., var.
Coral soil, Cudjoe Key, South Florida.
Legit A. H. Curtiss.



Yale University Herbarium
YU.001300

CURTISS, NORTH AMERICAN PLANTS, No. 503*

Polygala ambigua, Nutt., var.

Coral soil, Cudjoe Key, South Florida.

Legit A. H. CURTISS.

February.

Label Labels

- bc - barcode
- bt - barcode text
- cm - common/colloquial name
- cn - collection number
- co - collector
- cd - collection date
- fm - family name
- ft - footer info

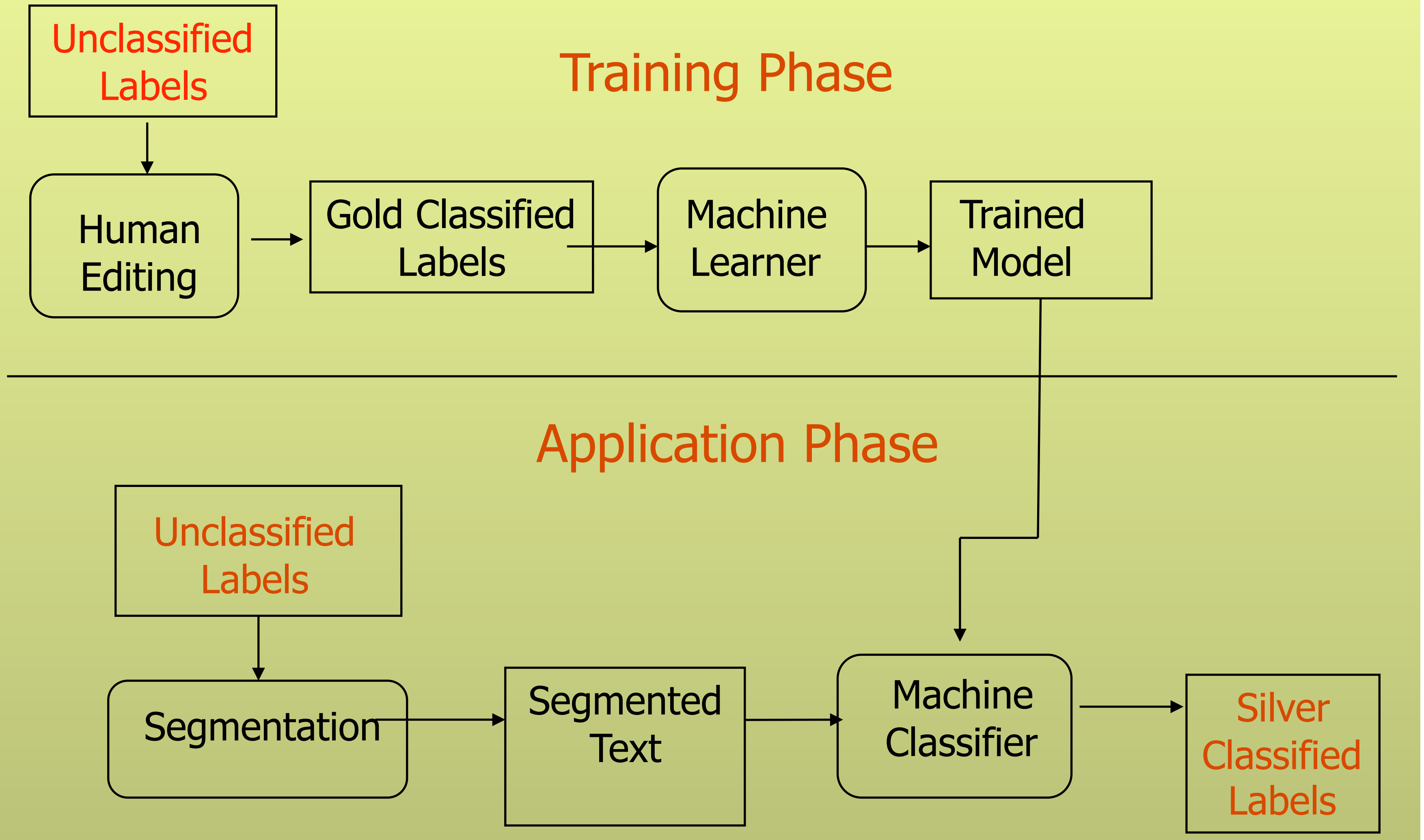
Label Labels

- gn - genus name
- hd - header info
- in - infra name
- ina - infra name author
- lc - location
- pd - plant description
- sa - scientific name author
- sp - species name

Example Training Record

```
<?xml version="1.0" encoding="UTF-8"?>
<?oxygen RNGSchema="http://www3.isrl.uiuc.edu/~TeleNature/Herbis/semanticrelax.rng"
      type="xml"?>
<labeldata>
<bt>Yale University Herbarium
</bt><ns> ~r-^"" r-n------</ns><bc> YU.001300
</bc><co cc="Curtiss"> Curtisb,  </co><hdlc cc="North American Plants">           North
      American Pl
</hdlc><ns>C^o.nr r^-n
ANTS,</ns>
<cnl> No.</cnl><cn> 503*</cn><ns> "^</ns>
<gn> Polygala</gn><sp> ambigna,</sp><sa> Nntt.,</sa><val> var.</val>
<hb> Coral soil,</hb><lc> Cudjoe Key, South Florida.
</lc><col> Legit</col><co> A. H. Curtiss.</co>
</labeldata>
```

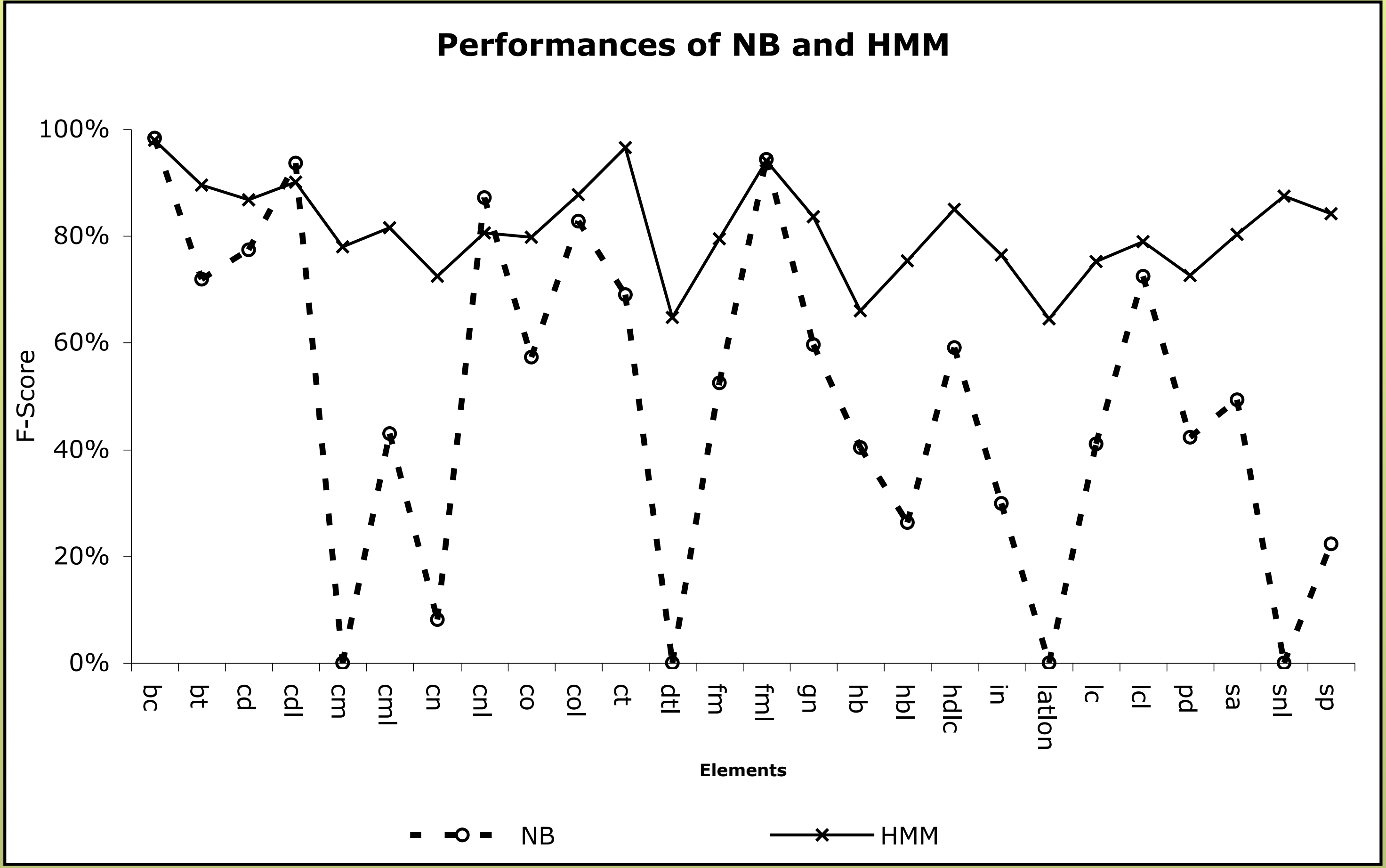
Supervised Learning Framework

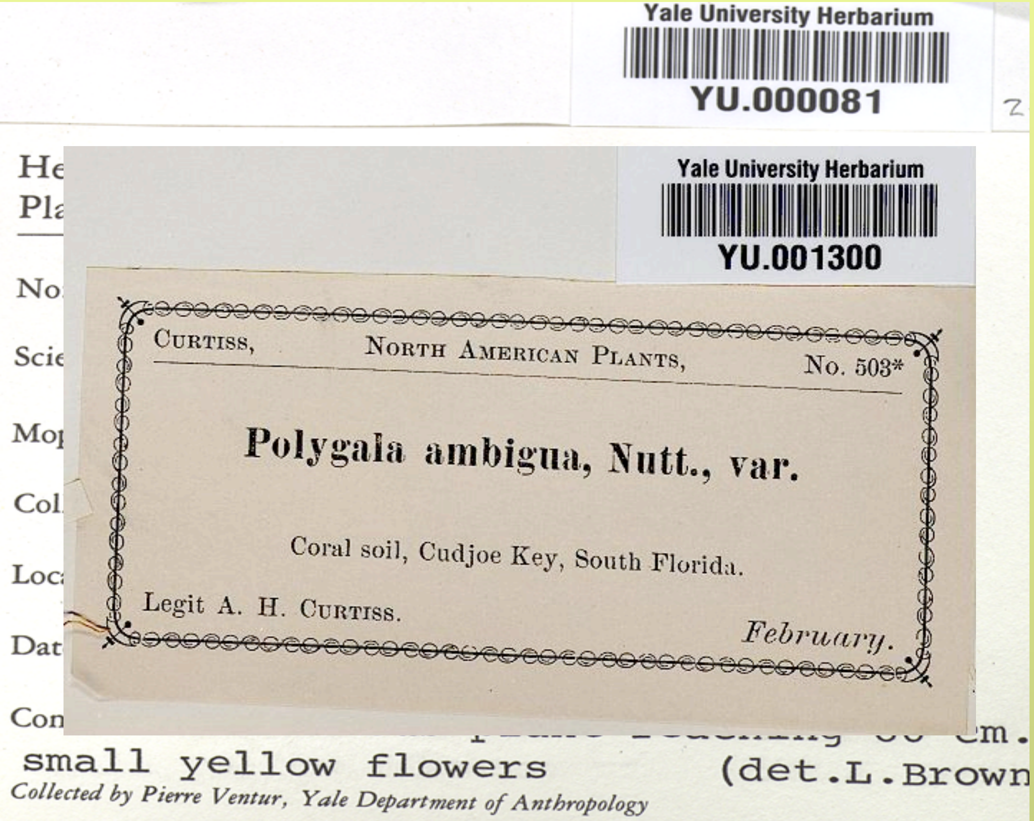


Herbis Experimental Data

- 295 marked up records
- 74 label states
- 5-fold cross-validation

Performances of NB and HMM





Yale University Herbarium

YU.000081

Z

He
Pla

No

Scie

Mo

Col

Loc

Dat

Con

Yale University Herbarium

YU.001300

CURTISS, NORTH AMERICAN PLANTS, No. 503*

Polygala ambigua, Nutt., var.

Coral soil, Cudjoe Key, South Florida.

Legit A. H. CURTISS.

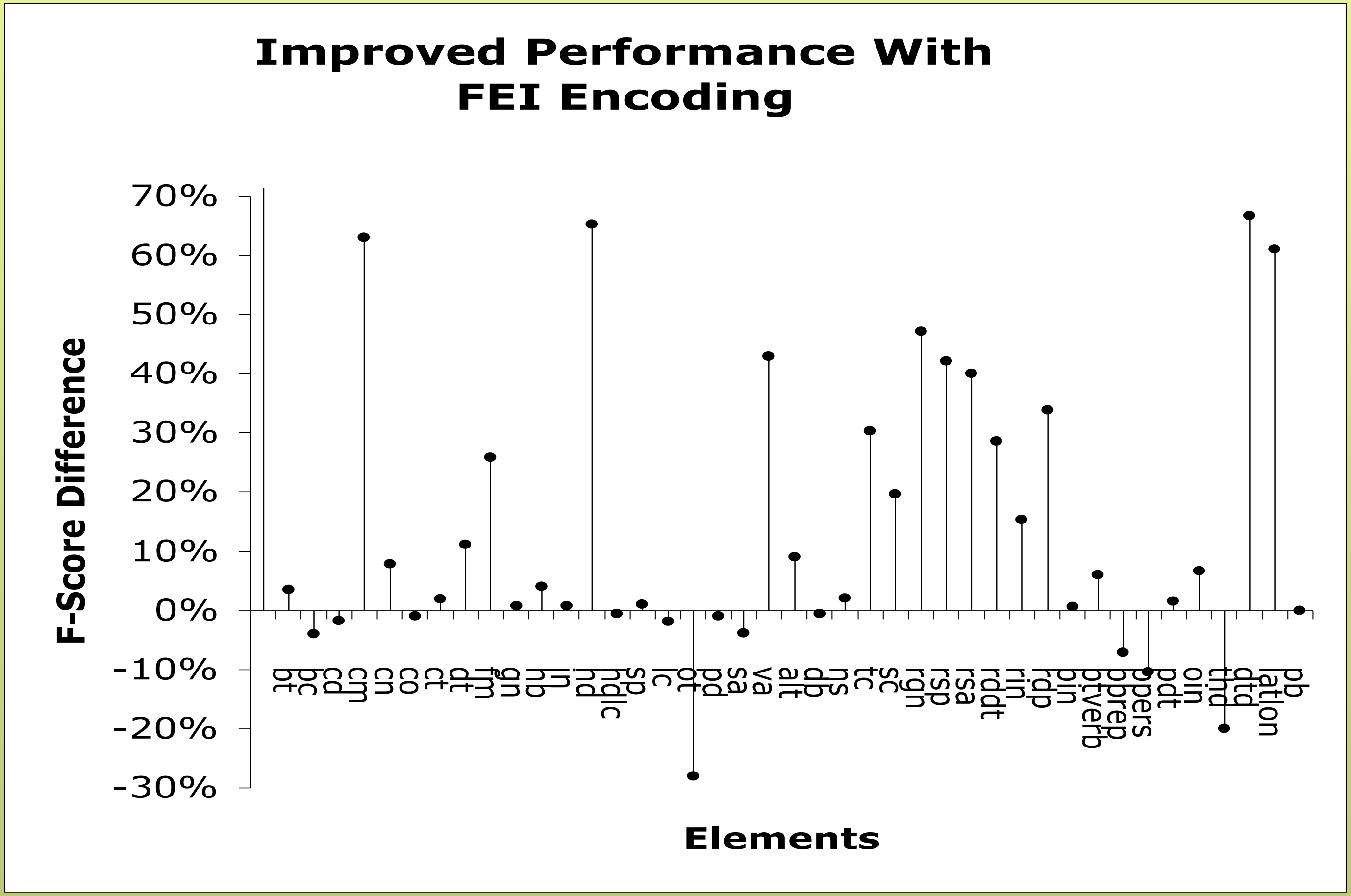
February.

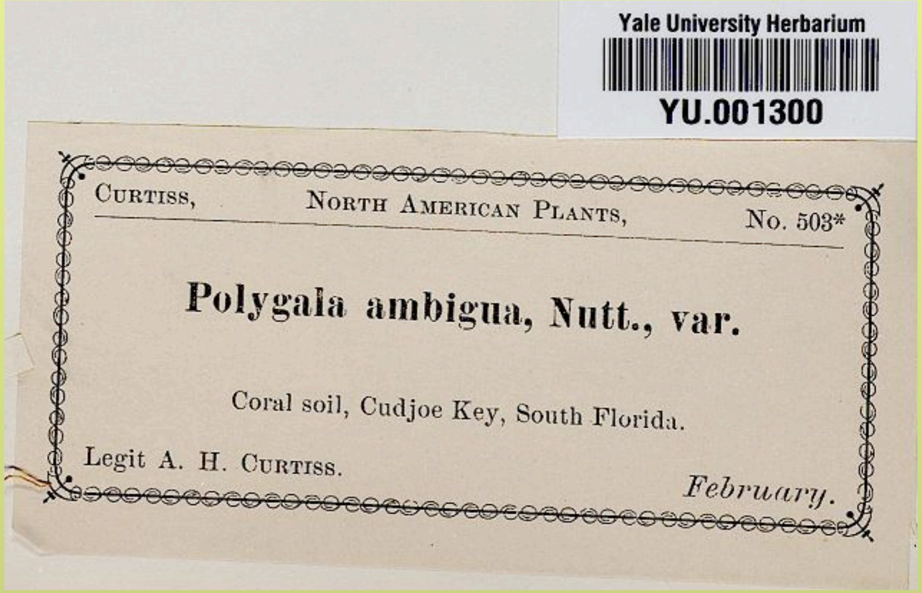
cm.

small yellow flowers (det.L.Brown

Collected by Pierre Ventur, Yale Department of Anthropology

Improved Performance With Field Element Identifiers





Yale University Herbarium

YU.001300

CURTISS, NORTH AMERICAN PLANTS, No. 503*

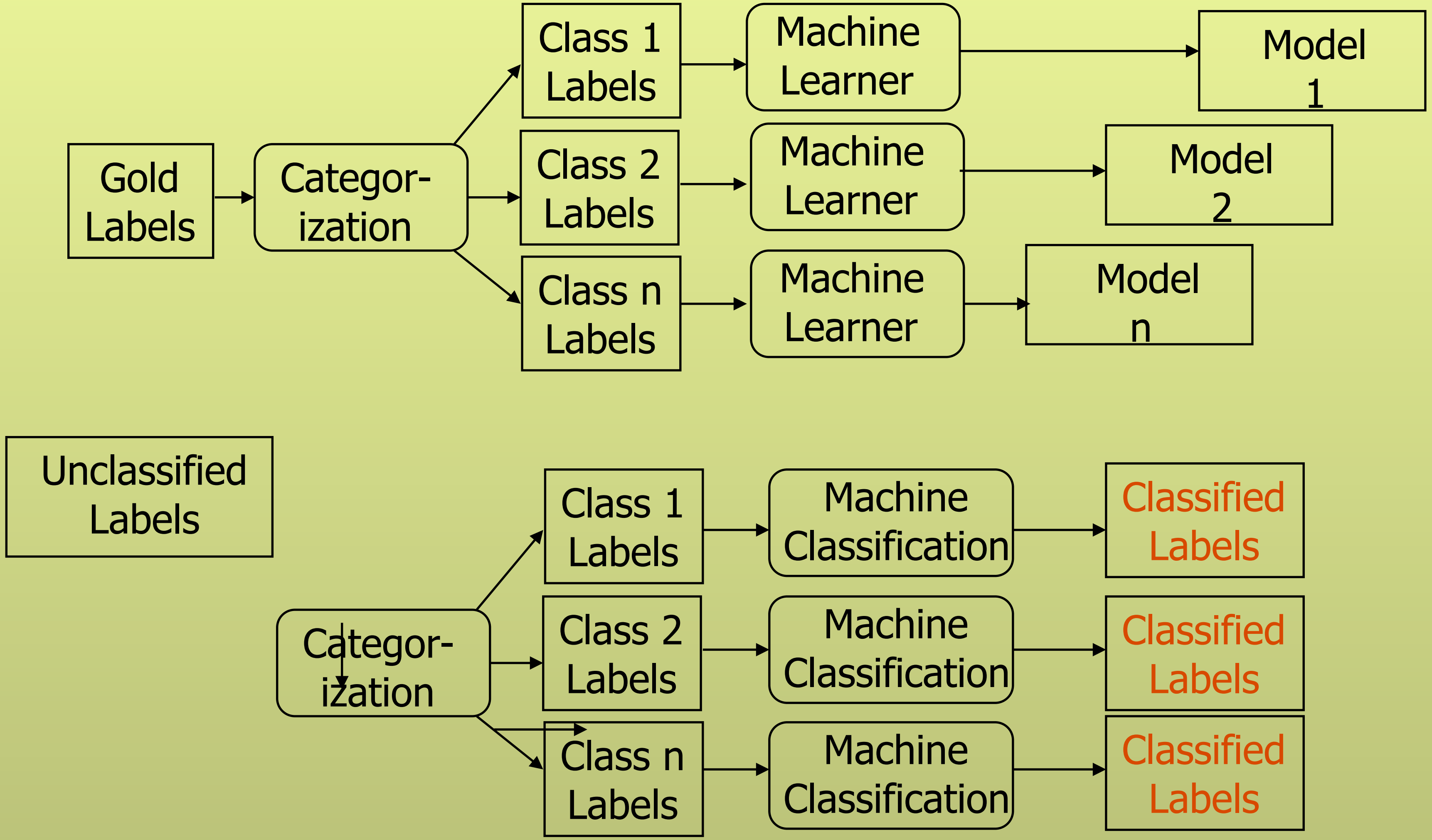
Polygala ambigua, Nutt., var.

Coral soil, Cudjoe Key, South Florida.

Legit A. H. CURTISS.

February.

Learning w/ pre categorization



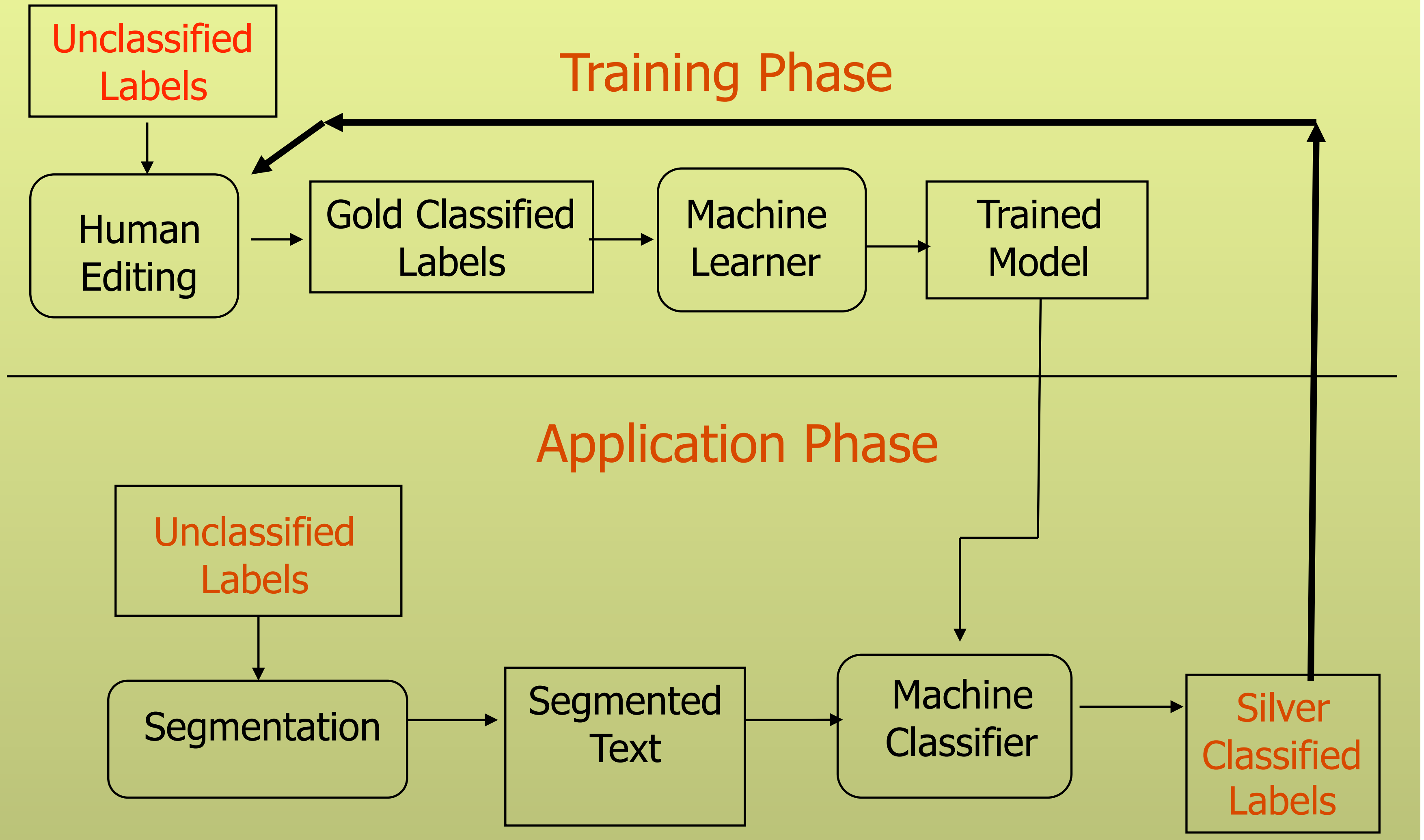
Specialist100 Curtiss VS 100 General

FIG. 5. Improved Performance of Specialist Model

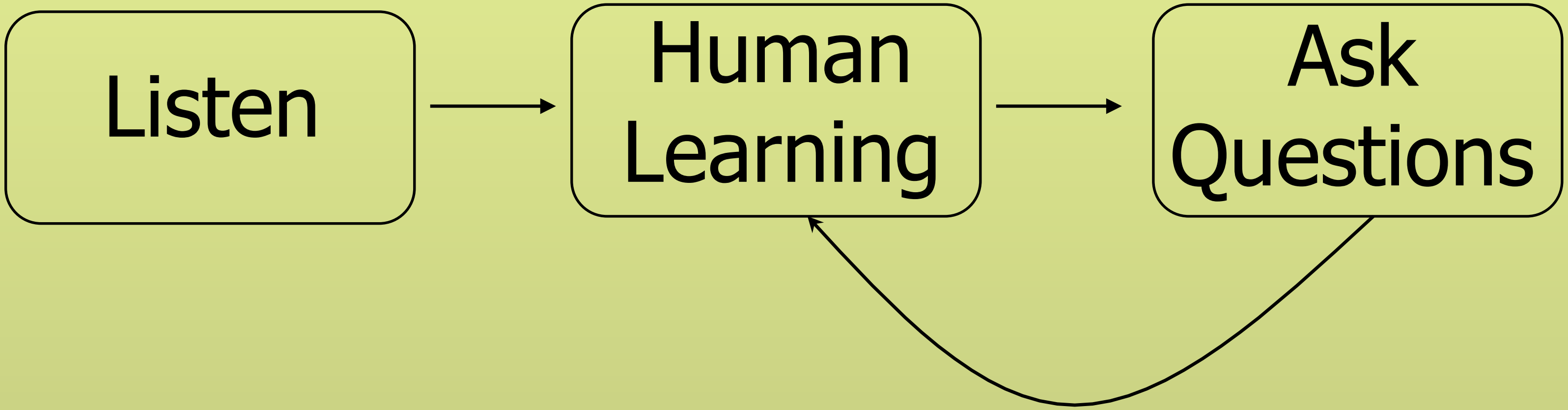
Future Work

- Community Learning Models
- Label records might be processed in different orders to maximize learning and minimize error rate.
- OCR correction might be improved using context dependent information. Context dependent correction means conducting the correct after knowing the word's class. For example, word "Ourtiss" should be corrected as "Curtiss". If the system already identified "Ourtiss" as collector, we can use the smaller collector dictionary instead of using a much larger general dictionary to do the correction.

Community Learning Models



Many thanks to Qin Wei



Listen

Human Learning

Ask Questions