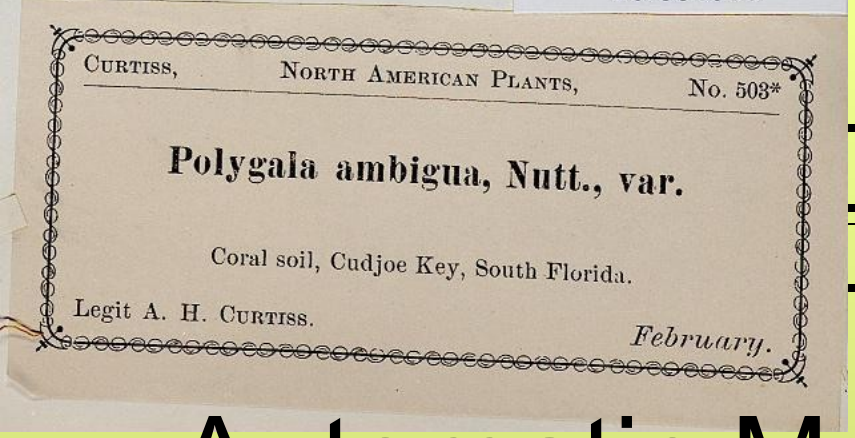


& <co> Curtis, </co><hdlc> North American Pl
</hdlc><cnl> No.</cnl><cn> 503*</cn>
<gn> Polygala</gn><sp> ambigua,</sp><sa>
Nutt.,</sa><val> var.</val>
<hb> Coral soil,</hb><lc> Cudjoe Key, South Florida.
</lc><col> Legit</col><co> A. H.
Curtiss.</co><dt>February</dt>&

Automatic Metadata Extraction (Darwin Core) From Museum Specimen Labels Specimen Specimen

Qin Wei (i), P. Bryan Heidorn,

University of Illinois at Urbana-Champaign, USA USAversity of

Email: qinwei2@illinois.edu

About me

- Phd student in Information Science in UIUC(UI UC( )
  - Research area: information retrieval, natural language processing, text mining l a n g u a g e p l a n g u a g e p r o c e s s i n g
  - Dissertation Topic: Taxonomic Name Recognition from full text r o m f u l l t e x t o p i c : T r o
  - Expected Graduation in Fall 2009
- Master in UIUC
- Bachelor's degree in Information Management in Peking University Ma n a g e Ma n a g e m

Co-author

- Dr. P. Bryan Heidorn
- Professor of Graduate School of Library and Information Science at UIUC
- pheidorn@illinois.edu

The problem

- More than 1 Billion Natural History Specimens Sp e 10 1 0 e c i m
- Collected over 250 years / many languages Co l l e c t e 250 2 5 0 l e
- No publishing standards No p u b l
- Near infinite classes Ne a r i n f i n
- 6 min / label * 1B labels = 100M hours (1 (1 ( )
- Saving 1 min = 16.7 Million hours S 1600 1 6 1 6
- $10/hr = $167,000,000 $ 1 1 6 6 1 0 / h

Why care

- Historic distribution of species Hi s t o H
- Ecological niche modeling (invasiveness, crop hardiness, pest potential) o t e n t
- Projections of the impact of climate change h a n g e

The Project

- Yale University Herbarium
- New York Botanical Garden
- University of Illinois
- Funded by National Science Foundation

Metadata Me t a d

- Data about data a t a a b o u t
  - *Author: James Smith*
  - *Date: August, 14, 2008*

    Compare to:
  - *“Author: James Smith”*
  - *“Date: August, 14, 2008”*
- **The importance of Metadata h**
- **Dublin Core** in library science i n l i b r
- **Darwin Core** in TDWG (More information could be found here f o u n d f o http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome)

Some Elements from Darwin Core

- ”Class”
- ”Order”
- ”Family”
- "Genus"
- "Species"
- "Subspecies"
- "ScientificNameAuthor"
- "IdentifiedBy"
- "YearIdentified"
- "MonthIdentified"
- "DayIdentified"

Why Machine Learning? Wh y Wh y

- Successfully adopted in other related/similar areas: information retrieval, named entity recognition e c o g n i t i o n o
- Many many tools are already available. (e.g. Weka, D2K) We k a , D2 K) t o o l
- More adaptable to data variability such as spelling variability p e l l i n g v a
- Can be user driven not programmer driven Ca n b e u s e each user may fine tune their own models t h e i r o wn mo d e l s t h e i r

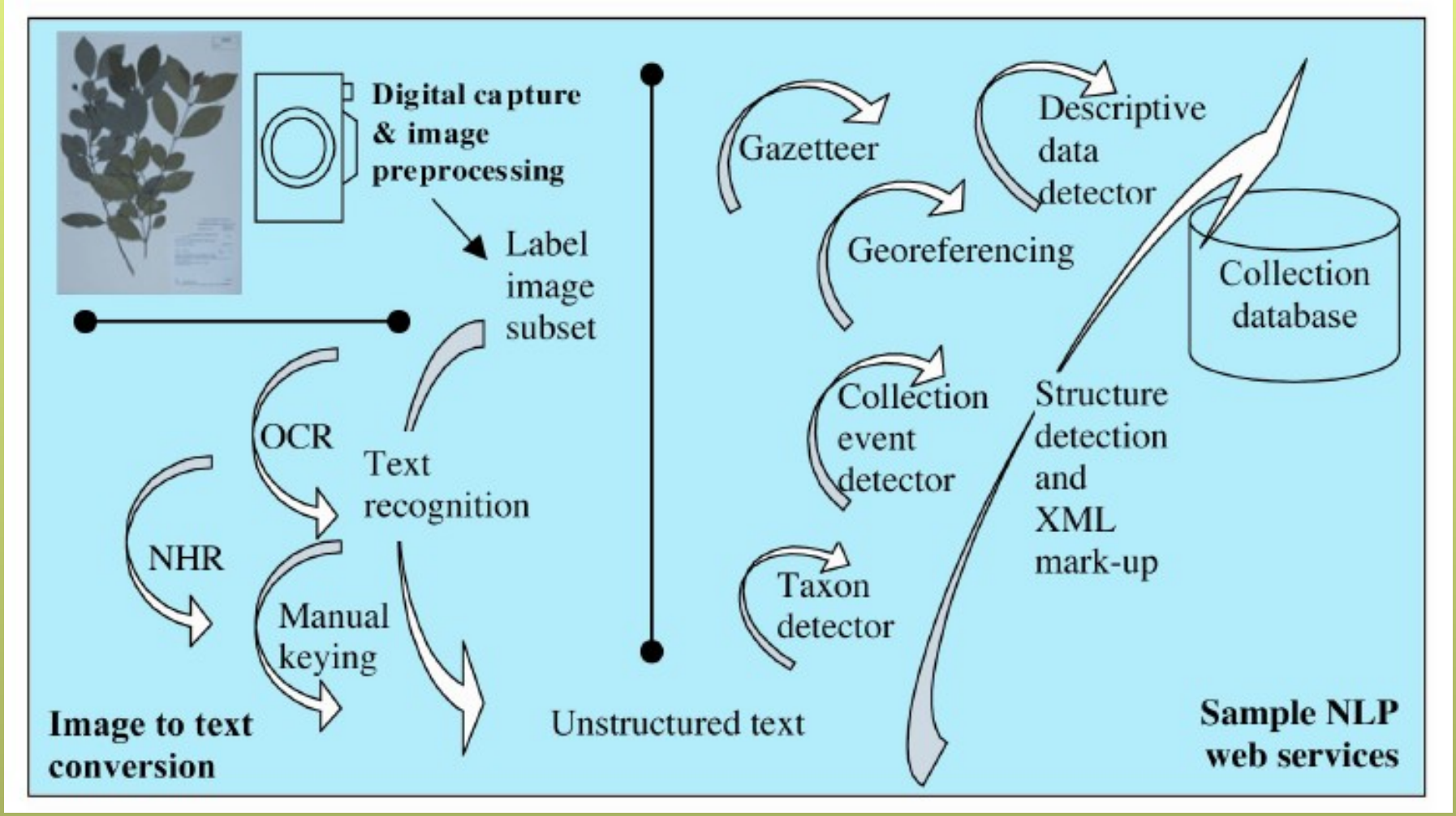
Supervised Machine Learning L e a r n i n g e d

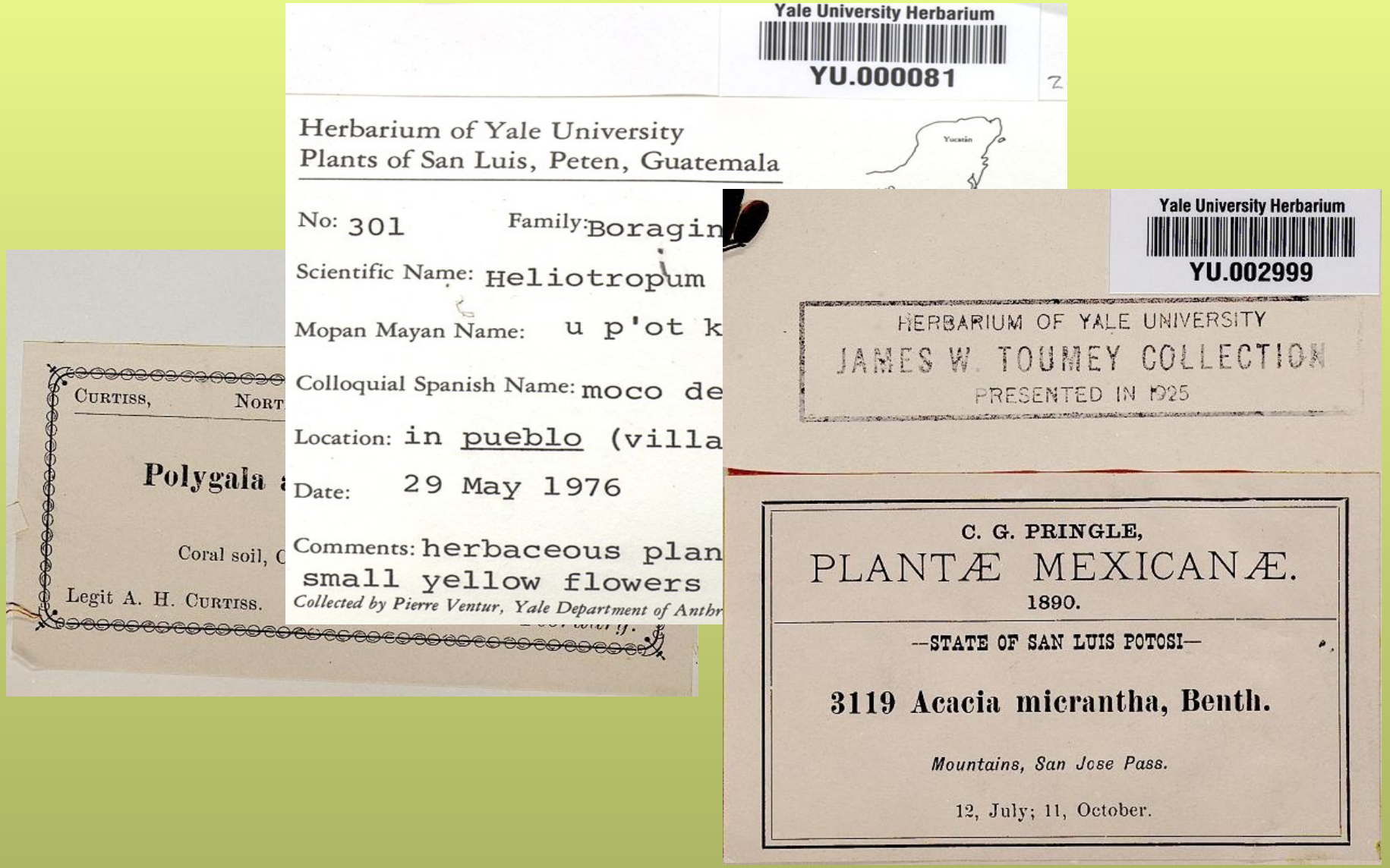
- “The method operates under supervision by being provided with the actual outcome for each of the training examples” (Witten, 2005)

e a c h o

- In another words, the learner gets the knowledge from the examples and then use the knowledge to classify new examples. t h t h e k n

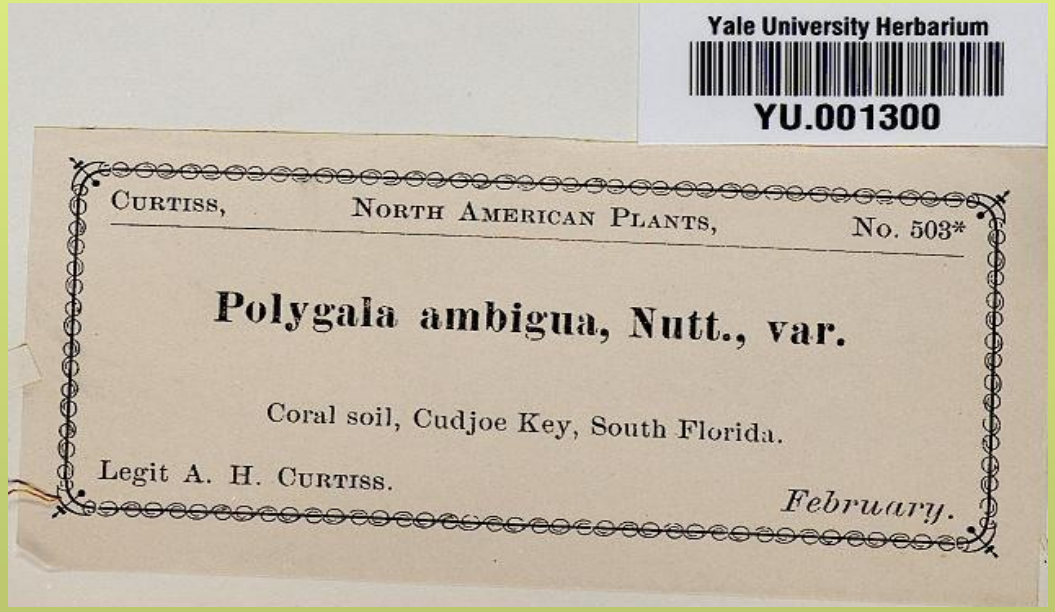
Work flow


Digital capture & image preprocessing
Label image subset
OCR
NHR
Text recognition
Manual keying
Image to text conversion
Unstructured text
Gazetteer
Descriptive data detector
Georeferencing
Collection database
Collection event detector
Structure detection and XML mark-up
Taxon detector
Sample NLP web services


Yale University Herbarium
YU.000081
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
Curtiss,
Polygala
Coral soil,
Legit A. H. Curtiss.
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

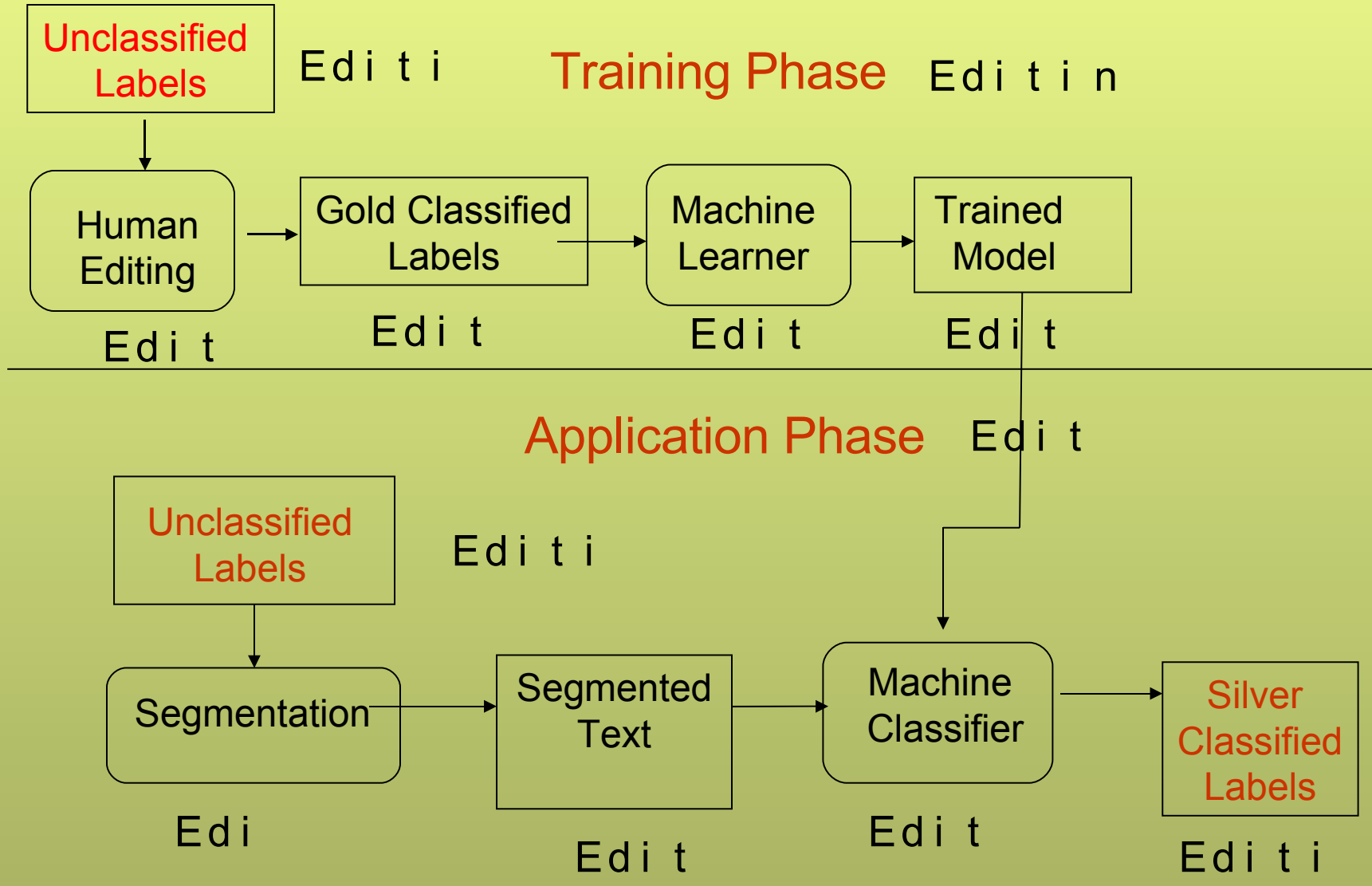
Yale University Herbarium
~r-^"""" r-n-------
YU.001300
Curtisb, North American Pl
C^o.nr r^-n
ANTS,
No. 503* "^
Polygala ambigna, Nntt., var.
Coral soil, Cudjoe Key, South Florida.
Legit A. H. Curtiss.



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>
```

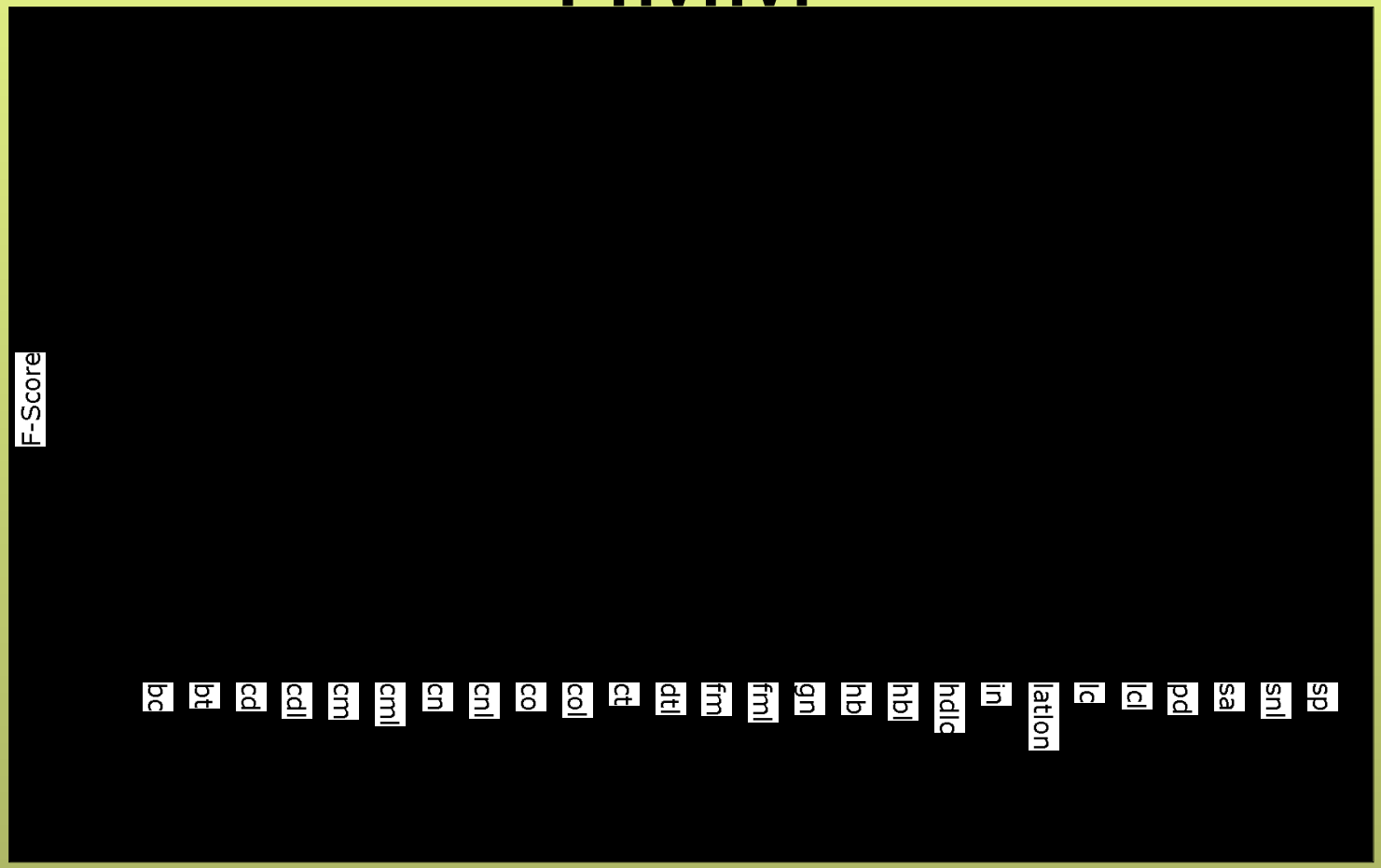
Supervised Learning Framework



Experimental Data E x p e r i

- 295 marked up records m 295 2 9 5 k e d
- printed labels, no handwriting p r i n t e d l p r i n t
- 74 label states 7 74 7 4 l
- NaiveBayes classifier VS. Hidden Markov Model Mo d e l a y e s c l a s s i f i
- 5-fold cross-validation 5 5 5 - f o l d

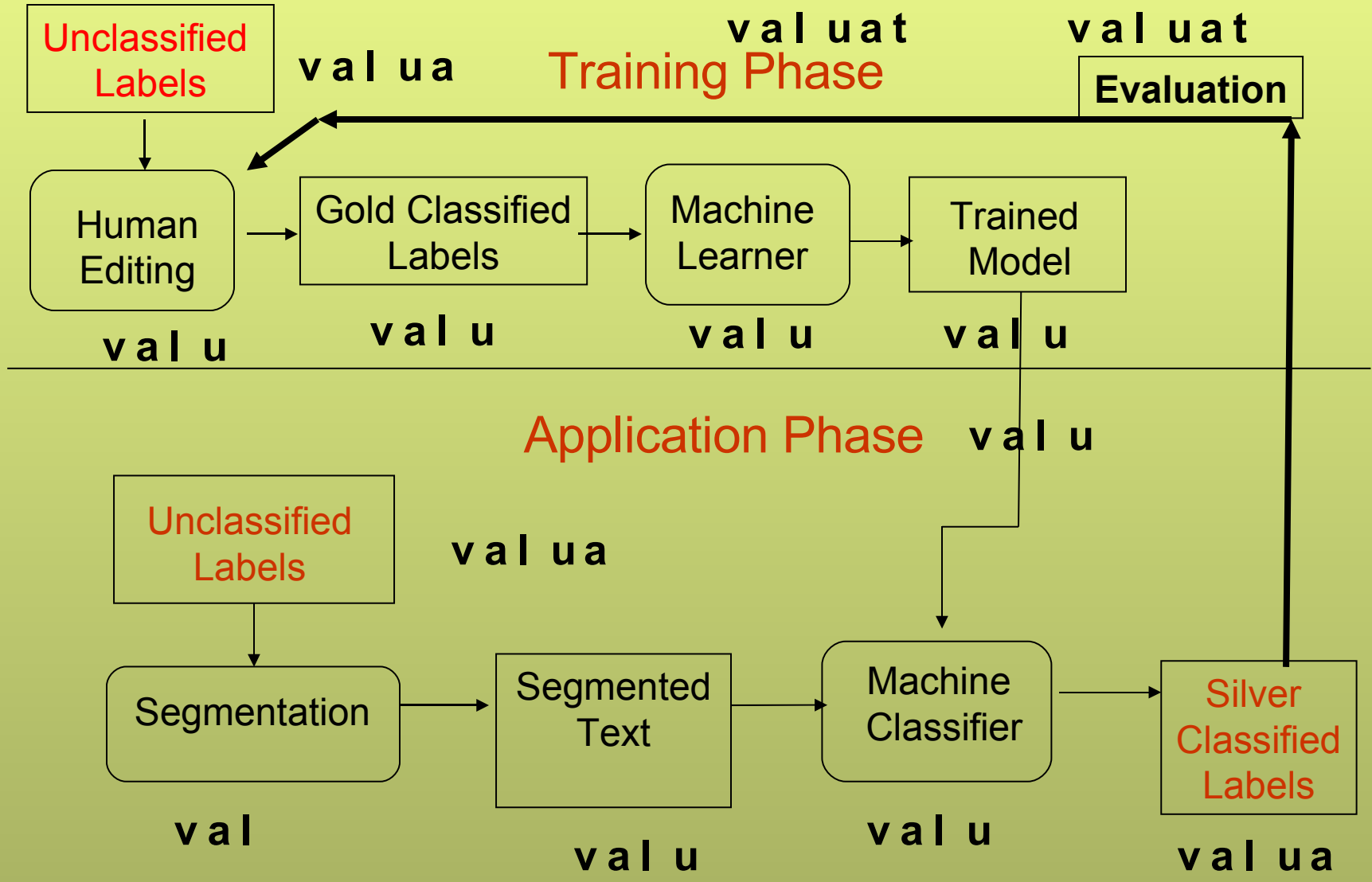
Performances of NB and HMM



Future Work ((t u r e)

- Community Learning Models C o mmu n i t y
- Label records might be processed in different orders to maximize learning and minimize error rate a x i mi z e a x i
- 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. mu c h mu c h l a r

Community Learning Models




Listen
Human Learning
Ask Questions

References

- Witten, I. H., and Frank, E. (2005). *Data mining: practical machine learning tools and techniques* (2 ed.). Boston, MA: Morgan Kaufmann Publishers.
- Cui, H., and Heidorn, P. B. (2007). The reusability of induced knowledge for the automatic semantic markup of Taxonomic Descriptions. *Journal of the American Society for Information Science and Technolog, 58*(1), 133-149.
- Bluma, A. L., and Langley, P. (1997).Selection of relevant features and examples in machine learning. *Artificial Intelligence,* 97, 245-271.
- Melville, P., and Mooney, R. J. (2003). Constructing diverse classifier ensembles using artificial training examples. In *Proceedings of the IJCAI-2003*, 505-510.