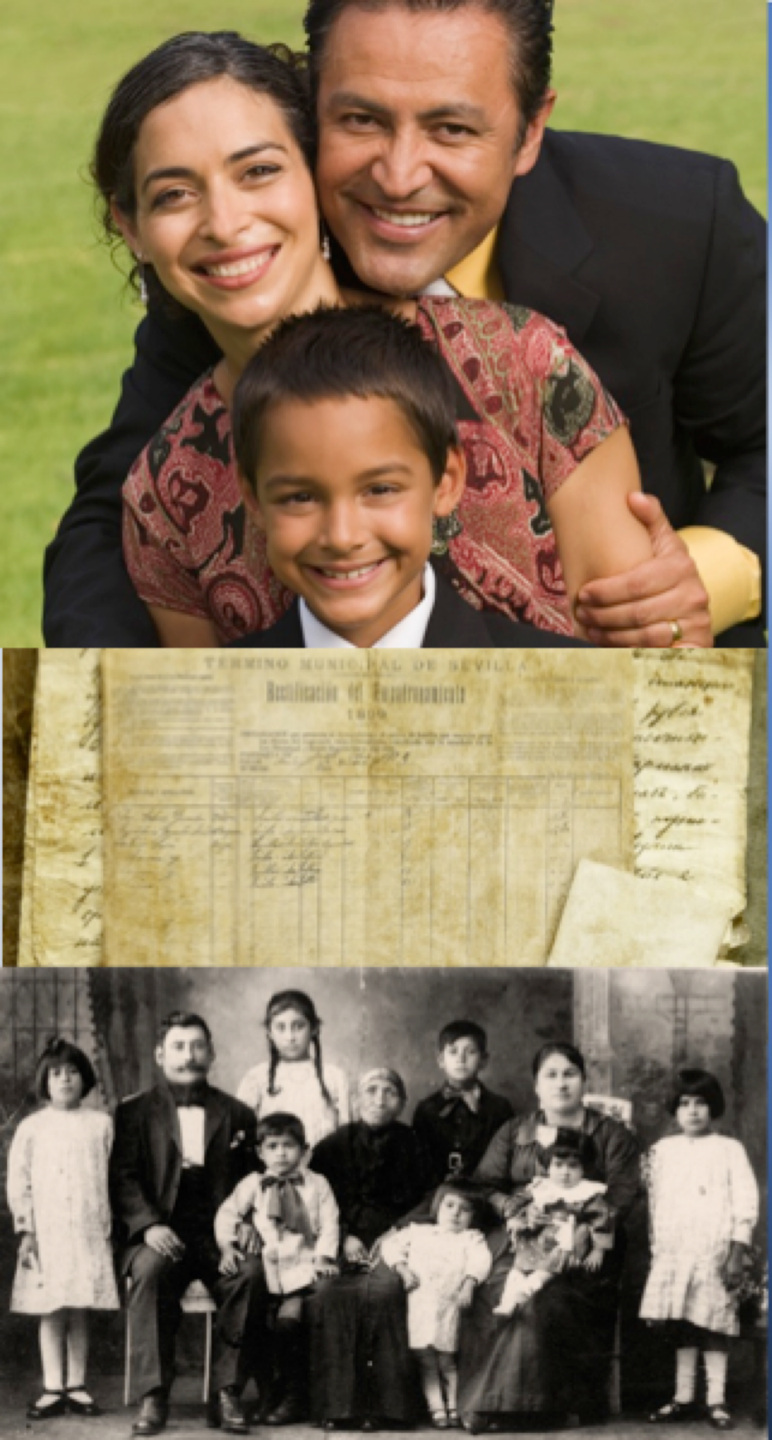


Economical Bimodal Classification of a Massive Heterogeneous Document Collection

Patrick Schone
(patrickjohn.schone@familysearch.org)

24 February 2020

THE CHURCH OF
JESUS CHRIST
OF LATTER-DAY SAINTS

FAMILYSEARCH

Overview

- Timelines (Lead-up)
- Description of the Collections
- Classification Goals for Automation
- Speed-focused System Architectures
- Performance and Outcomes

Timelines (Lead-up)

2015:

FamilySearch was able to auto-*index* 21M born-digital newspapers.
Can auto-indexing work with born-paper? How about handwriting??

2016-2017:

FamilySearch & BYU collaborate on technologies to auto-*transcribe* HW.

2017-2018:

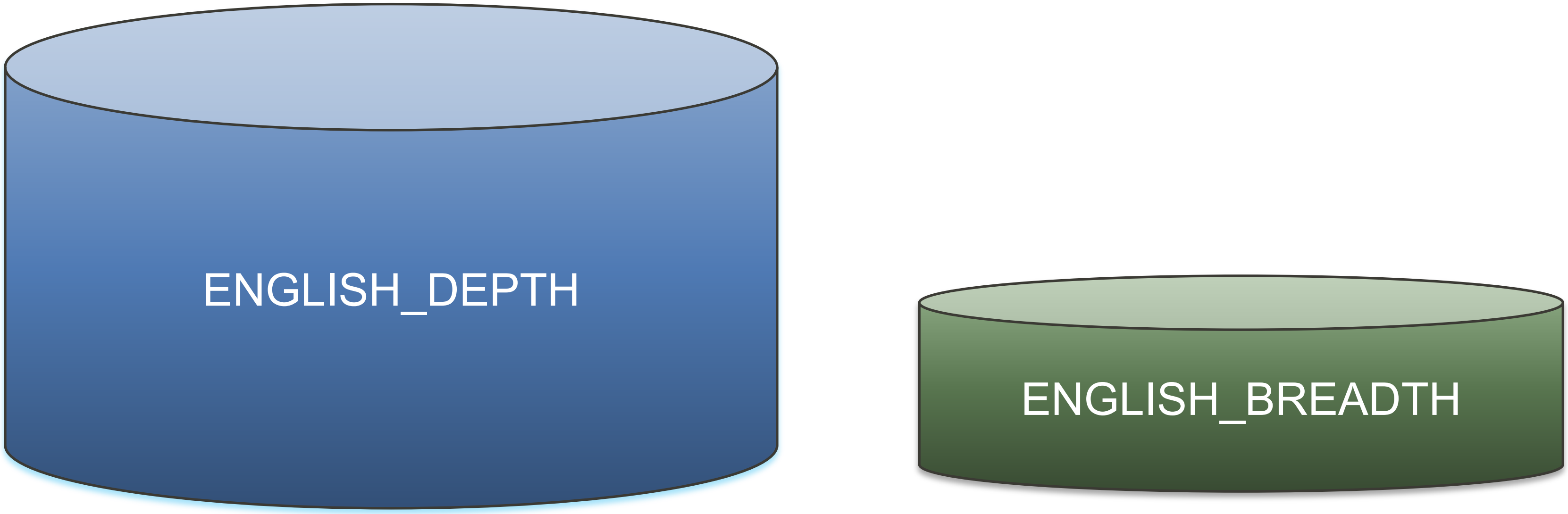
FamilySearch auto-*transcribed* about 33M newspaper stories and over 110M mostly-English handwritten & mixed documents with the goal of auto-*indexing* them.

2019:

Newspaper going forward. But the massively-heterogeneous collection makes auto-*indexing* complex. Need to group & categorize documents, identify ‘gotchas’, and subdivide images.

Collections Of Interest

Two different, but related, kinds of corpora:



163K Rolls of Film, every image
[Abt 110M images]
Represents **EVERY *instance*** of particular types of US Legal documents

~1M Rolls of Film, several ims/roll
[Abt 3-4M images]
Represents **EVERY** 'English' ***roll***

Can We Classify After-the-Fact?

If we could **describe each image** of the Breadth/Depth corpora, we could target sub-collections for auto-indexing based on current capabilities & develop the capability for others.

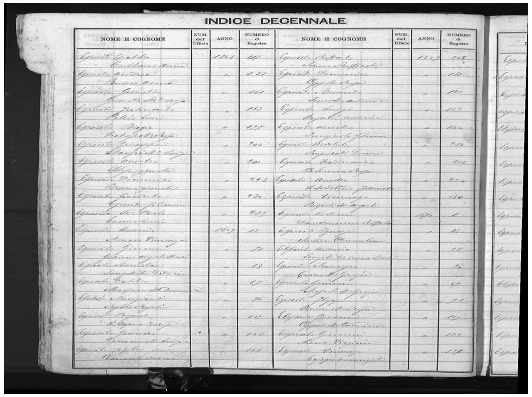
Also, if we could **identify any anomalies**, that might help us do a better job handling them.

But we want to do this **quickly**! We want to finish in a week or so. But if we only took 1 sec/document (typical load time of a full image), it'd take

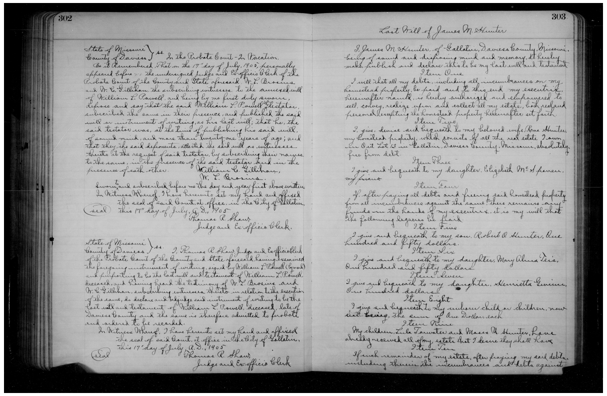
[1.1 x 10^8 images] x [1 sec/image] = 3.5 CPU years !

Classify: Semantic Categories

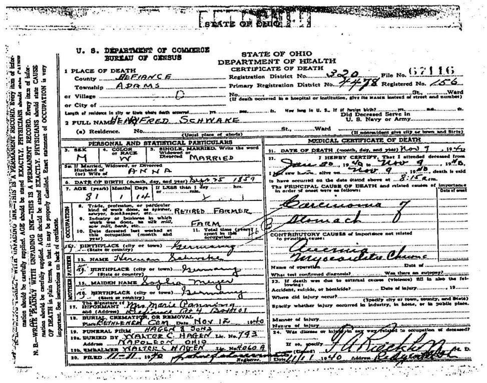
130+ Semantic Categories: What is the *PURPOSE* for the document?



Registration/Civil

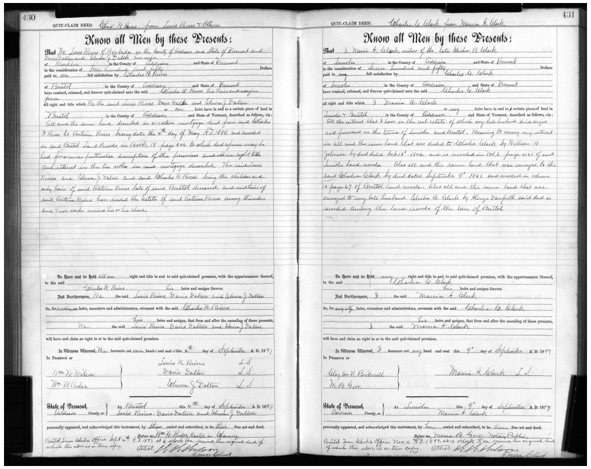


Probate/Will

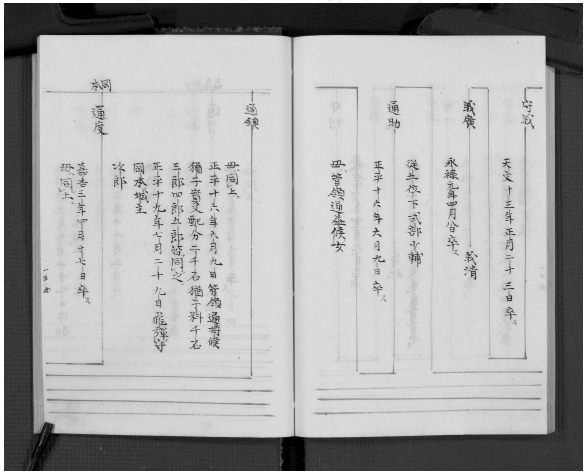


Vital/Death/Legal

General/Newspaper



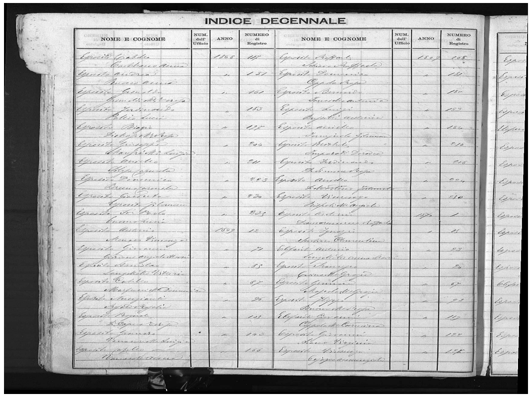
Land/Deed



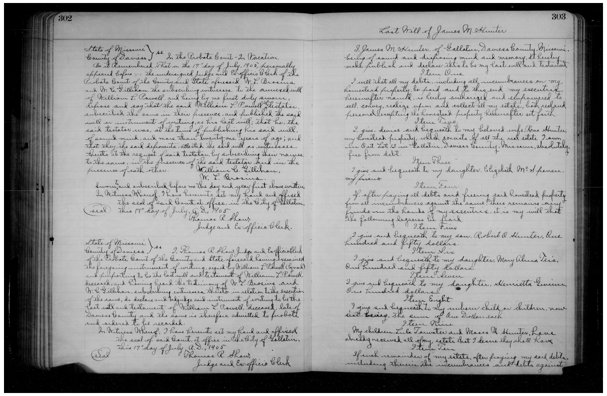
Family/Pedigree

Classify: Layout Categories

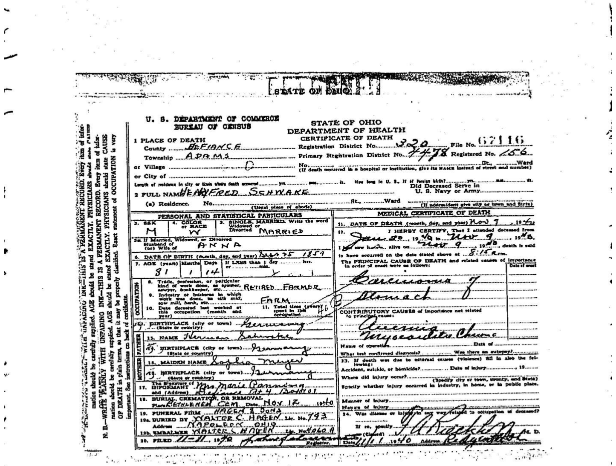
<u>~12 Layout Categories</u>: What is the *STRUCTURE of* the document?


INDICE DECENNALE
NOME E COGNOME
ANNO
NUMERO

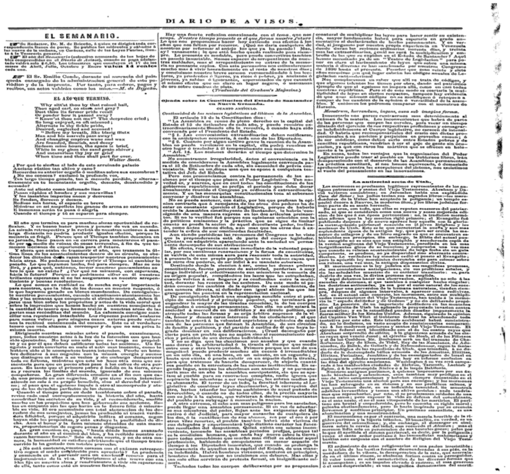
Table/1 Line Per Row



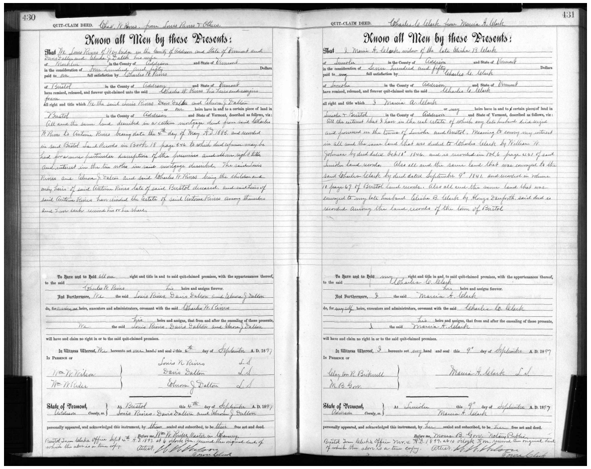
Freeform


STATE OF OHIO
U. S. DEPARTMENT OF COMMERCE BUREAU OF CENSUS
STATE OF OHIO DEPARTMENT OF HEALTH CERTIFICATE OF DEATH

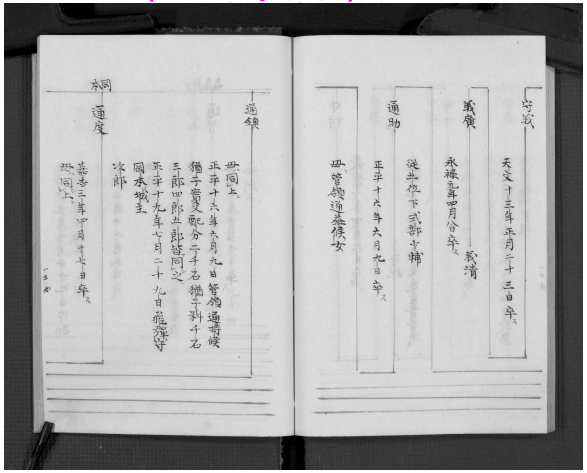
(Complex) Form


DIARIO DE AVISOS.
EL SEMANARIO.

Multicolumn


Know all Men by these Presents:
Know all Men by these Presents:

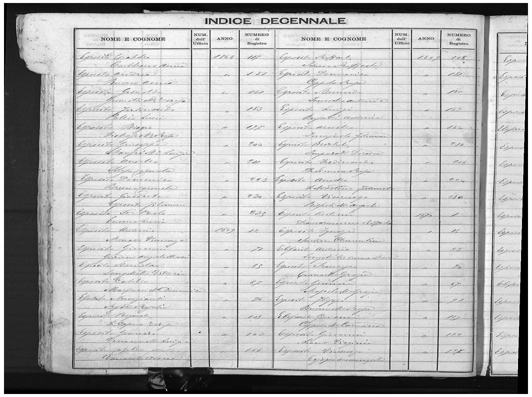
Fill in the Blank



Graphical

Classify: Story Count

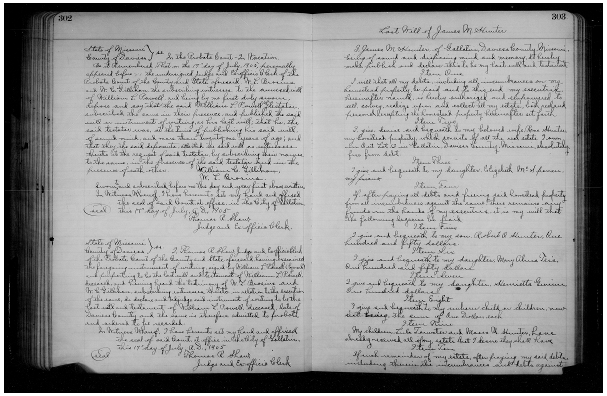
<u>~12 Story Classification</u>: How many unique 'stories' are in the document?



INDICE DECENNALE

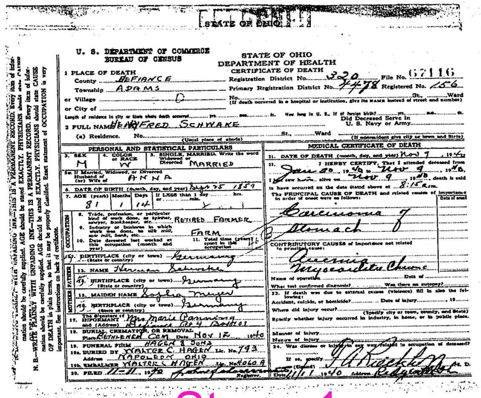
NOME E COGNOME | NUM. dell' Ufficio | ANNO | NUMERO di Registro

Story=1n



Last Will of James M. Hunter

Story=E&S

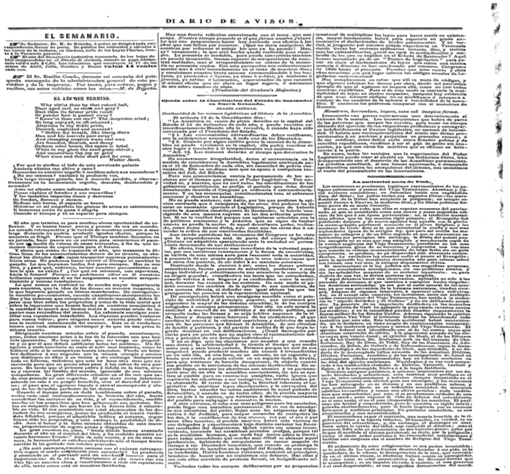


STATE OF OHIO

U. S. DEPARTMENT OF COMMERCE
BUREAU OF CENSUS

STATE OF OHIO
DEPARTMENT OF HEALTH
CERTIFICATE OF DEATH

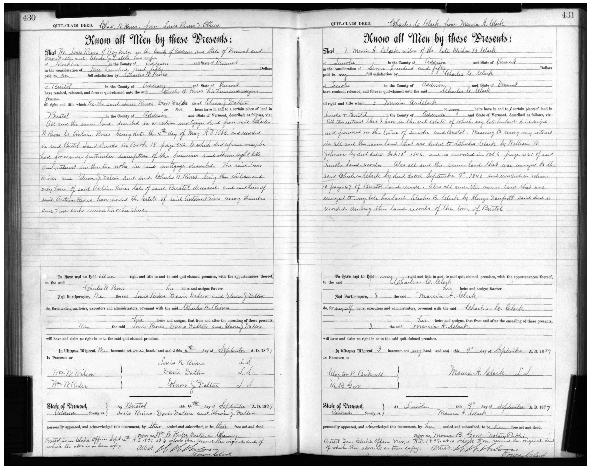
Story=1



DIARIO DE AVISOS.

EL SEMANARIO.

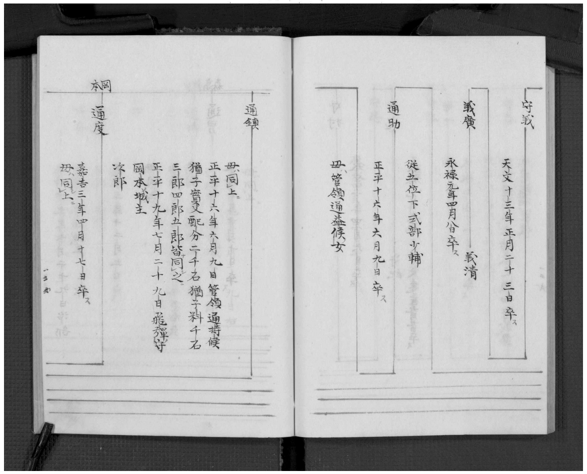
Story=many



QUIT-CLAIM DEED.

Know all Men by these Presents:

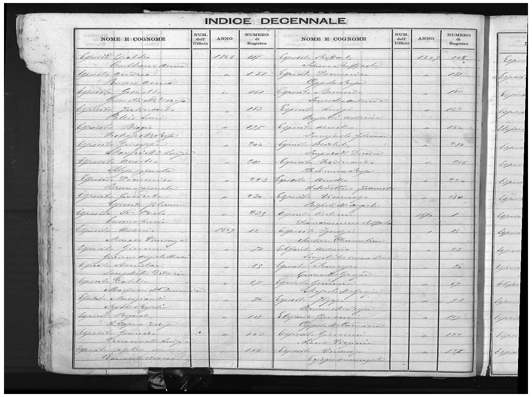
Story=2



Story=0p

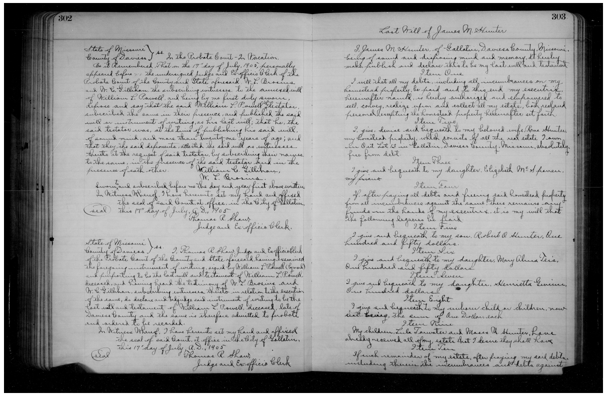
Classify: Language Info

Linguistics: What are the Unicode scripts, language, countries, writing style?



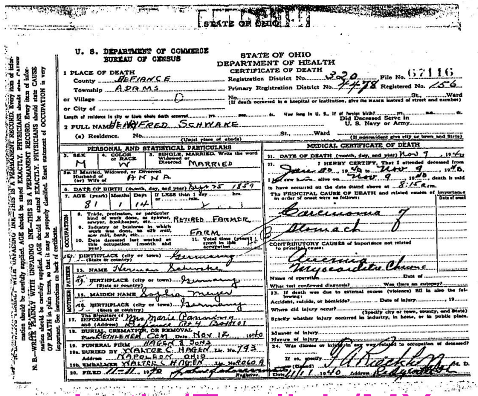
INDICE DECENNALE

Latin/Italian/MX



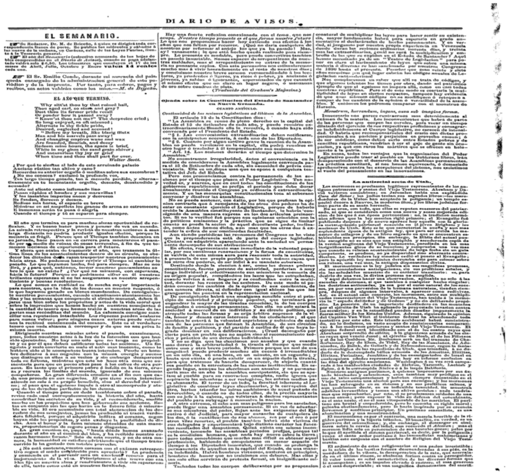
Last Will of James M Hunter

Latin/English/HW



STATE OF OHIO DEPARTMENT OF HEALTH CERTIFICATE OF DEATH

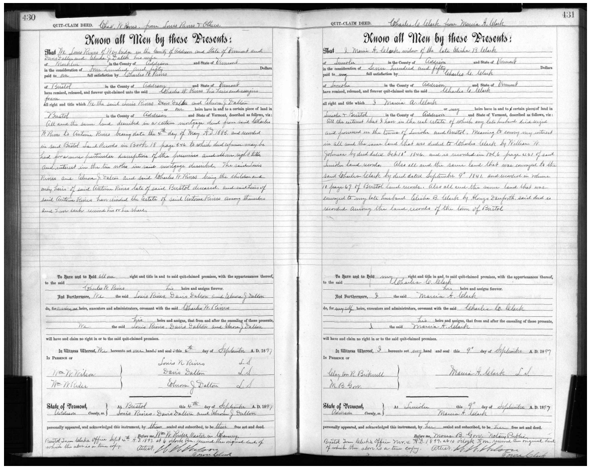
Latin/English/MX



DIARIO DE AVISOS

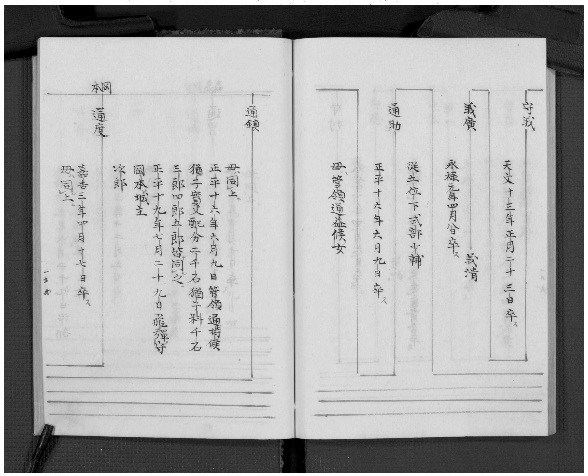
EL SEMANARIO.

Latin/Spanish/PR



Know all Men by these Presents:

Latin/English/MX

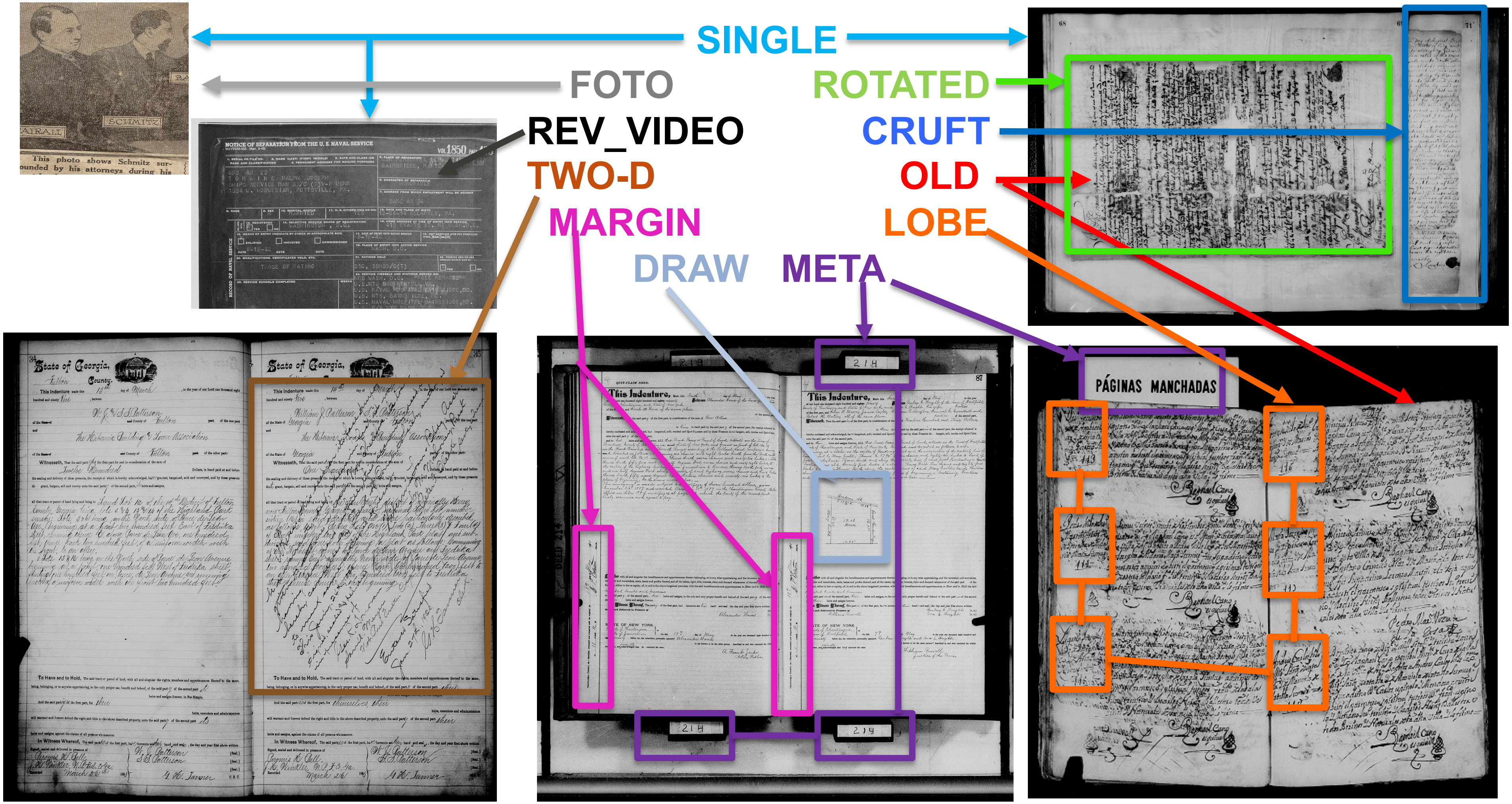


Chinese/Japanese/HP

THE CHURCH OF JESUS CHRIST OF LATTER-DAY SAINTS

FAMILYSEARCH

Anomalies: Binary Properties



SINGLE

FOTO

REV_VIDEO

TWO-D

MARGIN

DRAW

META

ROTATED

CRUFT

OLD

LOBE

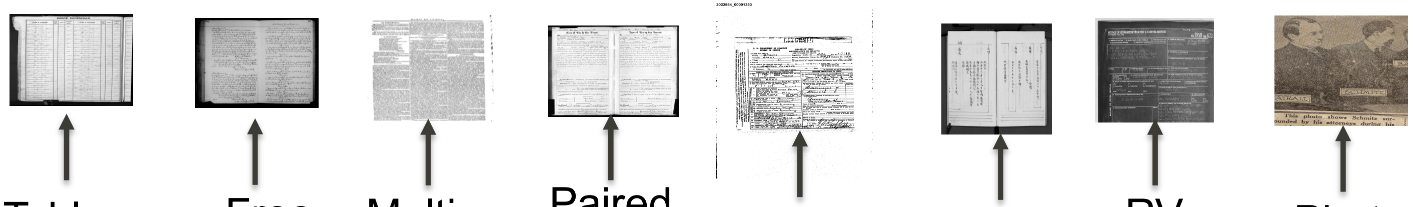
SCHMITZ
This photo shows Schmitz surrounded by his attorneys during his
NOTICE OF SEPARATION FROM THE U. S. NAVAL SERVICE
State of Georgia,
This Indenture
214
PÁGINAS MANCHADAS

Speedy Classification?

One Option: Use **thumbnail images** and do image-level classification.

Definite 'Wins':

- FamilySearch automatically stores 200x200 thumbnails of each image.
- Thumbnails for an entire roll of film (1000 images) occupy about the same storage space as 3 images [so, over 99% compression].
- Since these are small, load time and subsequent processing time is short.
- Can see color, periphery, two-up-ness, photos, & line patterns



Table

Free

Multi-column

Paired Forms

Form

Vertical

RV

Photo

Drawbacks:

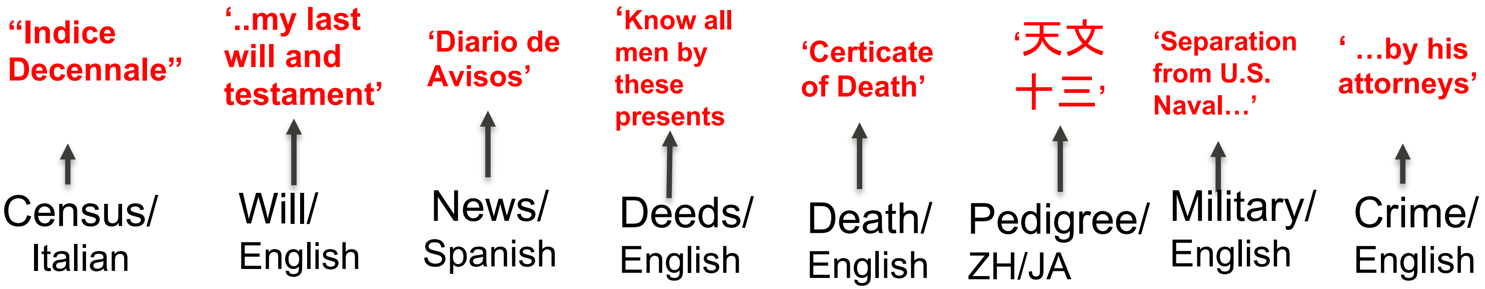
- Their amount of detail is limited, so it's hard to assess the true semantics. Have to guess the semantics based on 'this is a paired form, and that's what deeds look like, so I'll guess it's a deed."

Speedy Classification?

<u>Another Option</u>: Use **transcripts with bounding boxes** & do text-level classification.

Definite Wins:

- Processing transcript is *orders of magnitude faster* than thumbnails.
- Semantic information is often very clear at the textual level.
- Language, script, country, writing style – should all be straightforward to note.



“Indice Decennale”	**‘..my last will and testament’**	**‘Diario de Avisos’**	**‘Know all men by these presents**	**‘Certicate of Death’**	**‘天文十三’**	**‘Separation from U.S. Naval…’**	**‘ …by his attorneys’**
↑	↑	↑	↑	↑	↑	↑	↑
Census/ Italian	Will/ English	News/ Spanish	Deeds/ English	Death/ English	Pedigree/ ZH/JA	Military/ English	Crime/ English

Serious Drawbacks:

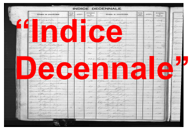
- Color is gone; borders are likely gone; photos are gone. How can one even tell if an image was reverse video if all you have is the transcript? How can you tell if it was complicated form or if it was nicely laid out?
- One needs to have the transcripts already.

Speedy Classification?

BEST Option: Use **BOTH snapshots AND transcripts+bounding boxes**.

Definite Wins:

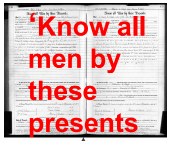
- Get the best of both worlds: semantics from text, visuals from thumbnail.
- Not much more expensive than JUST thumbnails when using both.
- Can toggle and use text-based or image-based models if that's all one has.



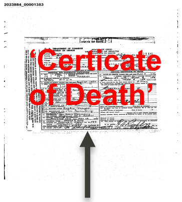
Census/
Italian/
Table

Will/
English/
Free

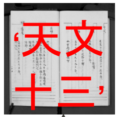
News/
Spanish
Multicol



Deeds/
English/
PairForm

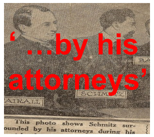


Death/
English/
Form



Pedigree/
ZH/JA
Vertical

Military/
English/
RV

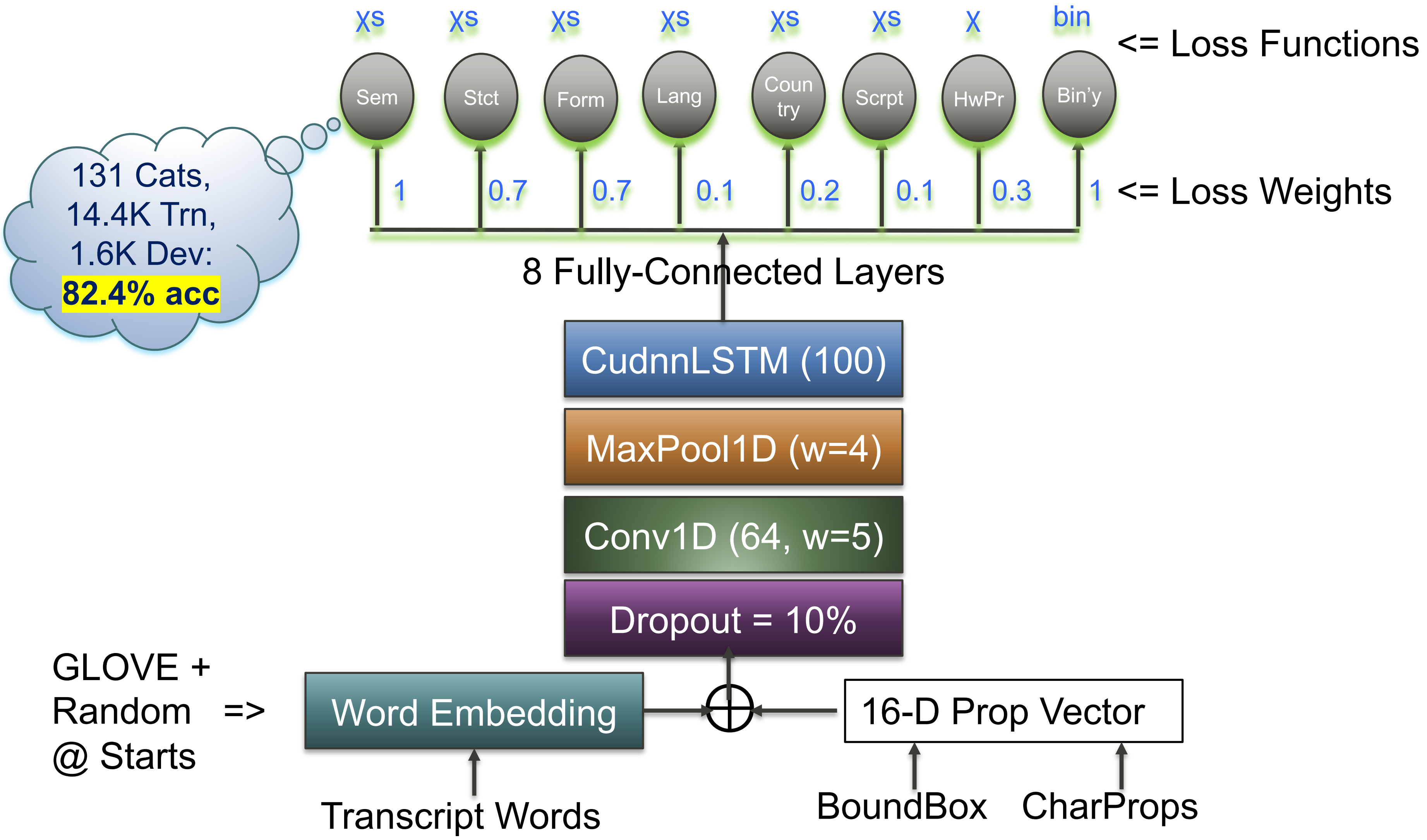

'…by his attorneys'
This photo shows Schmitz sur-
rounded by his attorneys during his

Crime/
English/
Newsclip
w/photo

Drawbacks:

- Model management is slightly more complex.

System Architecture: Text Input



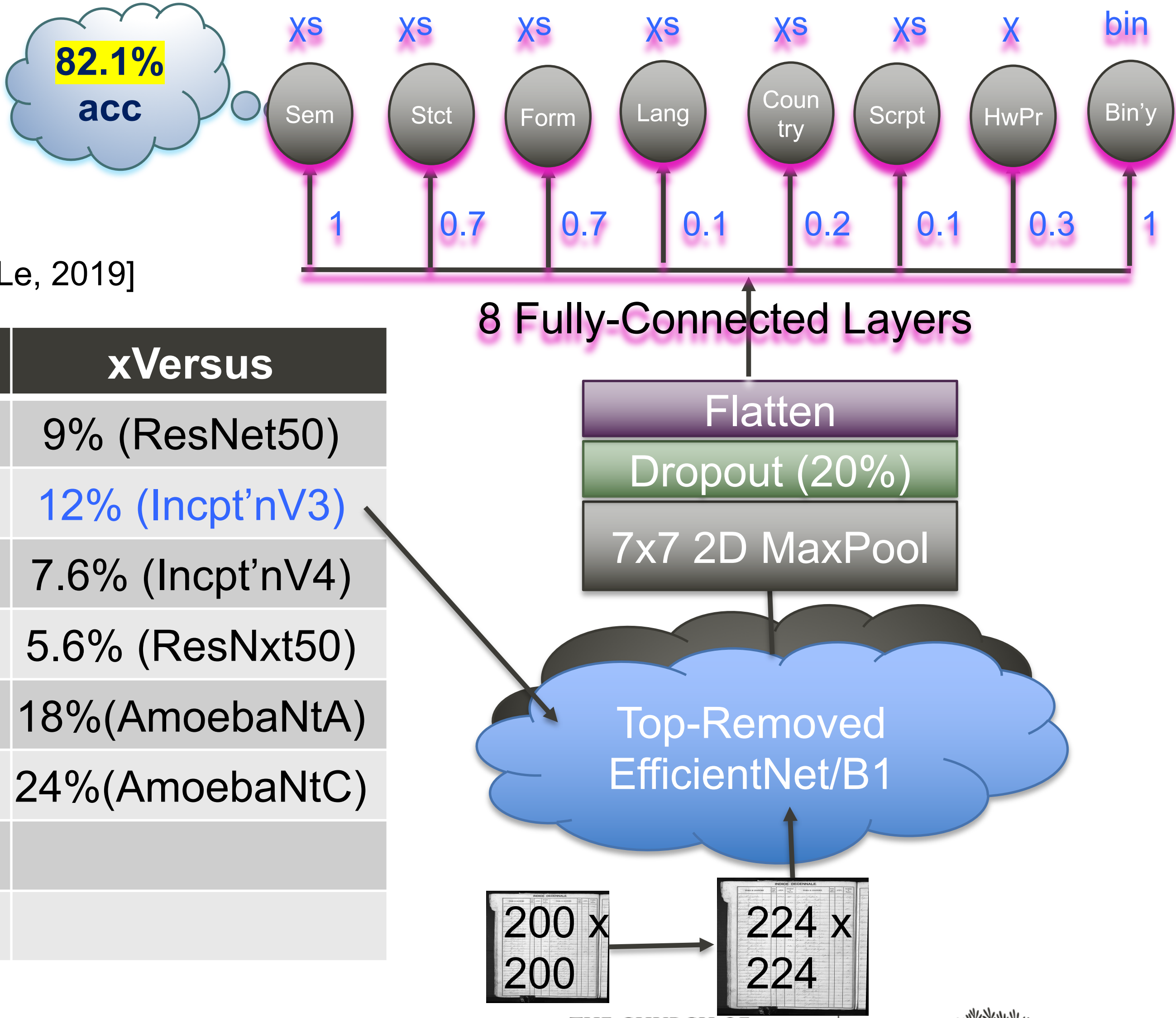
THE CHURCH OF JESUS CHRIST OF LATTER-DAY SAINTS | FAMILYSEARCH

System Design: Image Input

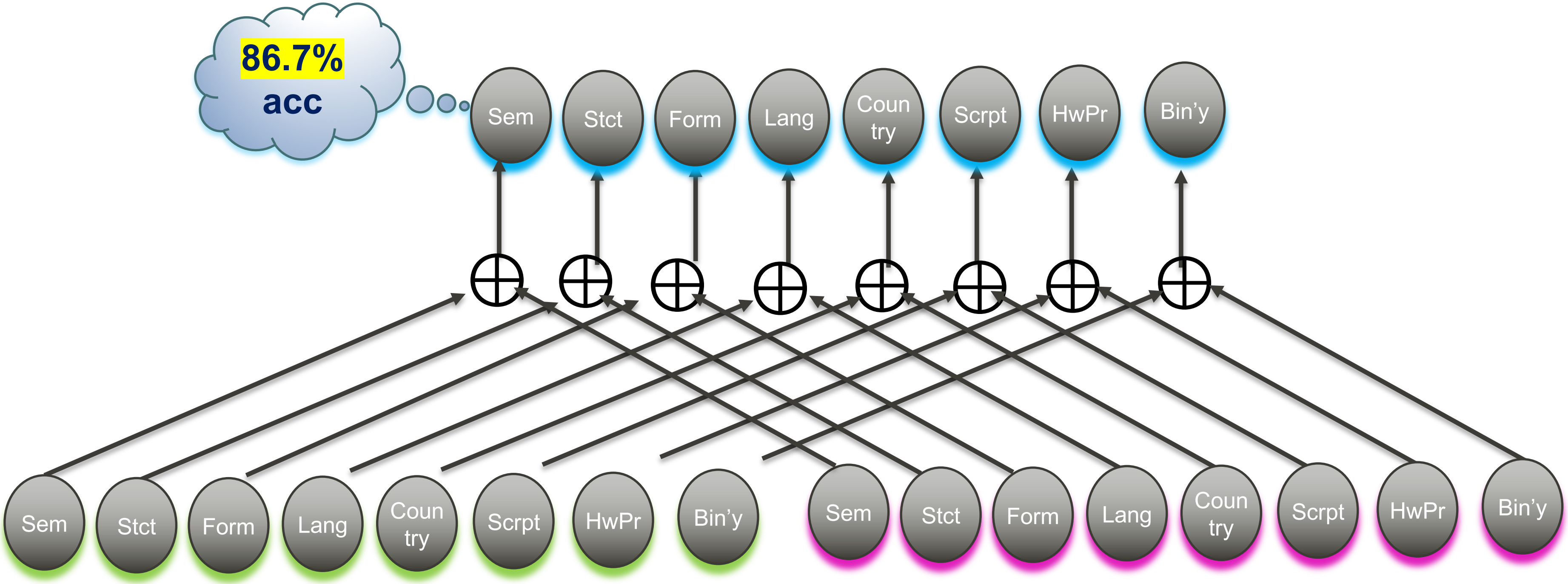
EfficientNet [M. Tan, Q. Le, 2019]

Net	#Param	#Flops	xVersus
B0	5.3M	0.39B	9% (ResNet50)
B1	7.8M	0.70B	12% (Incpt'nV3)
B2	9.2M	1.0 B	7.6% (Incpt'nV4)
B3	12M	1.8 B	5.6% (ResNxt50)
B4	19M	4.2 B	18%(AmoebaNtA)
B5	30M	9.9 B	24%(AmoebaNtC)
B6	43M	19 B	
B7	66M	37 B	

Results reported by Tan&Le.



System Design: Fused Input



For fully-connected weights at start, assume near-50% weights for class C from text(or image) going to class C in final, and near-zero weights for all other connections.

Outcomes: Timings

115,973,482 Images

Ran TWO trials. First was TEXT ONLY, second was FULL.

TextOnly:

Ran on one box (Dual-Gpu System).
Three jobs/Gpu (but lock around Gpu process)
Took 3.5 days.

FullSystem:

Re-Ran on 3 diff't machines, with variable number of Gpus.
But would have taken ~20 days on system of 'TextOnly' (with bulk of the additional cost going to thumbnail processing).

Outcomes: Results

115,973,482 Images

Semantics	%
Deeds	52.6
Land Index	11.6
Gen.Legal	8.3
Gen.Probate	5.6
Will	4.0
Inventory	3.4
Recpt/Check	1.1

Layouts	%
Freeform	68.1
Fill-in	18.2
Table/1line	10.4
Form	1.7

Recording	%
Handwrit’n	59.1
Mixed	22.0
PrintOnly	18.3
Blank	0.7

Anomalies	%
One-ups	52.4
Old (<1800)	3.7
HasMeta	2.0
HasLobes	1.5
ReverseVid	0.6
BleedThru	0.5

#Stories	%
Exactly 1	35.0
EndOrStrt	19.3
>1, but <2	9.3
End&Start	8.4
1-∞ Index	7.7
Exactly 2	7.2
Many	7.0

Summary

- Identified deep neural networks to mine text and image content, with sparse network combiner
- 86.7% acc on 131 category determination, plus generates multiple other kinds of classifications simultaneously
- Demonstrated result on large collection of >110 images

QUESTIONS?