Toponym Recognition in Historical Maps by Gazetteer Alignment

Jerod Weinman

Department of Computer Science
Grinnell College

Old Maps Online

www.oldmapsonline.org/

Reader

Old Maps Online

Search

1000 1250 1500 1750 2010

Search Collections Blog About

Instant Search Results: Fulltext

Isothermal chart.
1:95 000 000
1824 - Woodbridge, William C.

Isothermal chart, or, View of climates & productions / drawn from the accounts of Humboldt & others, by W.C. Woodbridge.
1823 -

Inhabited World.
1:93 000 000
1824 - Woodbridge, William C.

Isothermal chart, productions.
1:80 000 000
1837 - Woodbridge, William C.

Animals - World.
1837 - Woodbridge, William C.

Ocean Atlantique et Ocean Indien.
1:50 000 000
1937 - Vivien St Martin, L.

JISC University of Portsmouth KLOKAN TECHNOLOGIES

Media Information

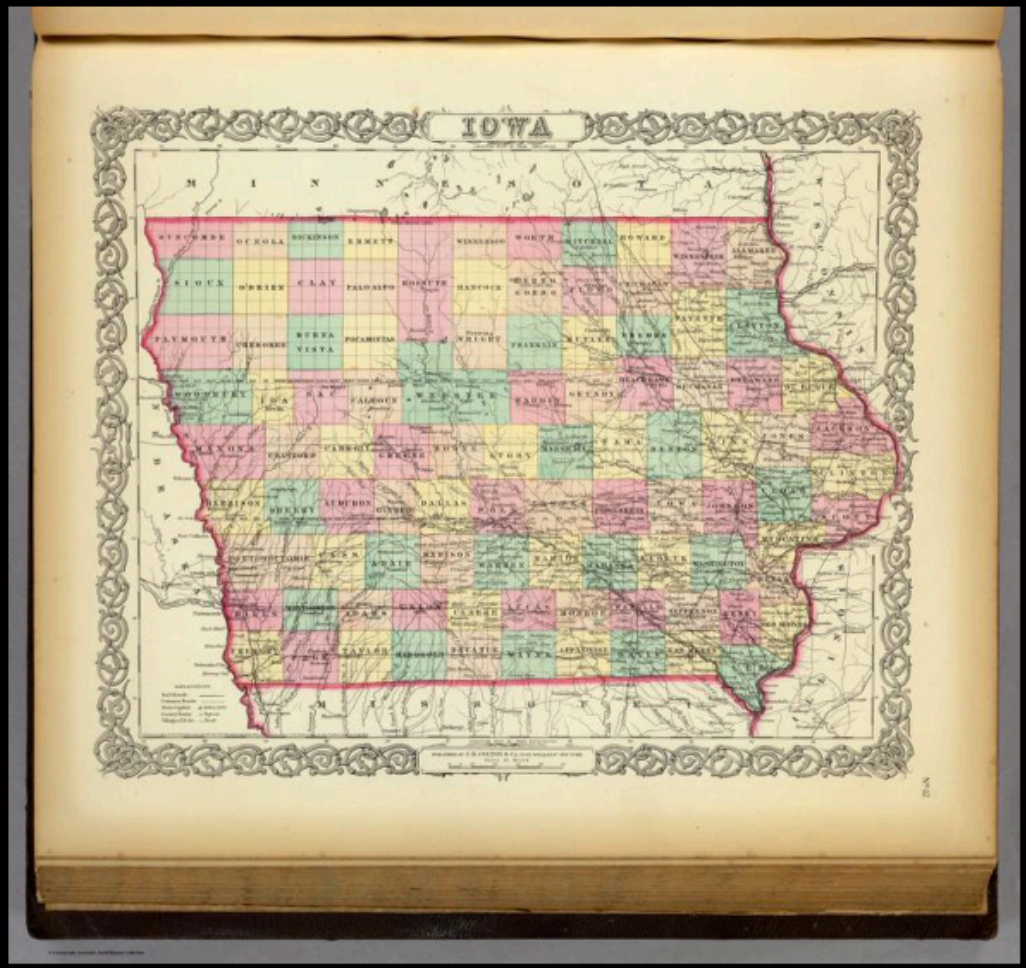
Collection:
David Rumsey Historical Map Collection
Author:
Colton, G.W.
Date:
1856
Short Title:
Iowa.
Publisher:
J.H. Colton
New York
Type:
Atlas Map
Obj Height cm:
34
Obj Width cm:
41
Scale 1:
1,440,000
Note:
In full color by county.
Reference:
P816.
State/Province:
Iowa
Full Title:
Iowa. Published By J.H. Colton & Co. No. 172 William St. New York. Entered ... 1855 by J.H. Colton & Co. ... New York. No. 47.
List No:
0149.052
Series No:
57
Publication Author:
Colton, G.W.
Pub Date:
1856
Pub Title:
Colton's Atlas Of The World, Illustrating Physical And Political Geography.

Iowa. Published By J.H. Colton & Co. No. 172 William St. New York. Entered ... 1855 ...

Printer Friendly

Add To Workspace

Export (Login for hi-res)



Media Information

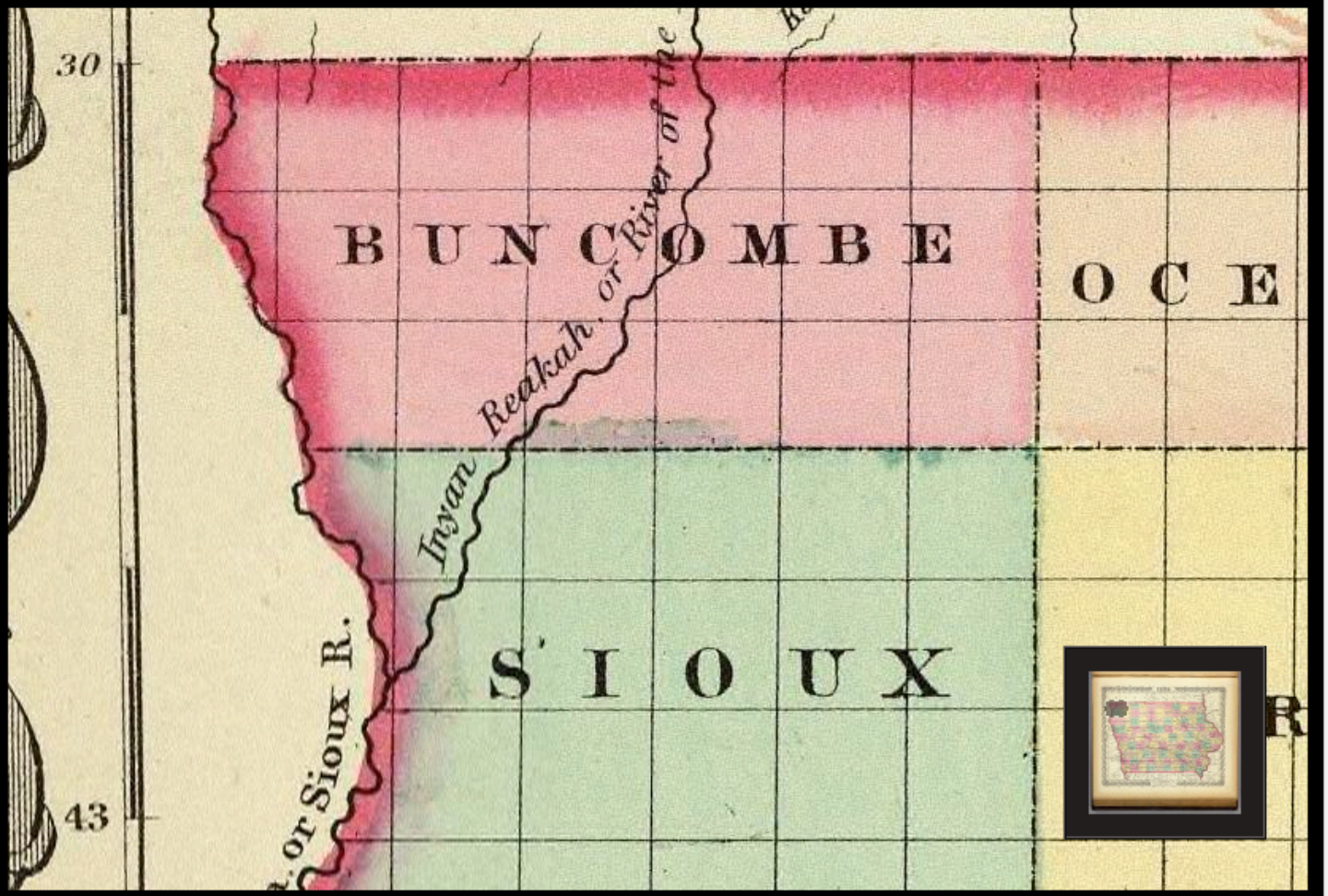
Collection:
David Rumsey Historical Map Collection
Author:
Colton, G.W.
Date:
1856
Short Title:
Iowa.
Publisher:
J.H. Colton
New York
Type:
Atlas Map
Obj Height cm:
34
Obj Width cm:
41
Scale 1:
1,440,000
Note:
In full color by county.
Reference:
P816.
State/Province:
Iowa
Full Title:
Iowa. Published By J.H. Colton & Co. No. 172 William St. New York. Entered ... 1855 by J.H. Colton & Co. ... New York. No. 47.
List No:
0149.052
Series No:
57
Publication Author:
Colton, G.W.
Pub Date:
1856
Pub Title:
Colton's Atlas Of The World, Illustrating Physical And Political Geography.

Iowa. Published By J.H. Colton & Co. No. 172 William St. New York. Entered ... 1855 ...

Printer Friendly

Add To Workspace

Export (Login for hi-res)



Related Work

- Text/graphics separation in maps
  Myers *et al.* (GREC '96), Luo & Kasturi (GREC '98), …

- Recognition in modern or specialized maps
  Li *et al.* (ICDAR '99), Velázquez & Levachkine (GREC '03),
  Pouderoux *et al.* (ICDAR '07), Chiang & Knoblock (ICPR '10),
  Pezeshk & Tutwiler (Trans. GeoSci & Rem. Sens. '11)

- Recognition using geographical information
  Gelbukh et al. (GREC '04)

Research Questions

1. How accurate is a general robust word recognition system on historical maps?

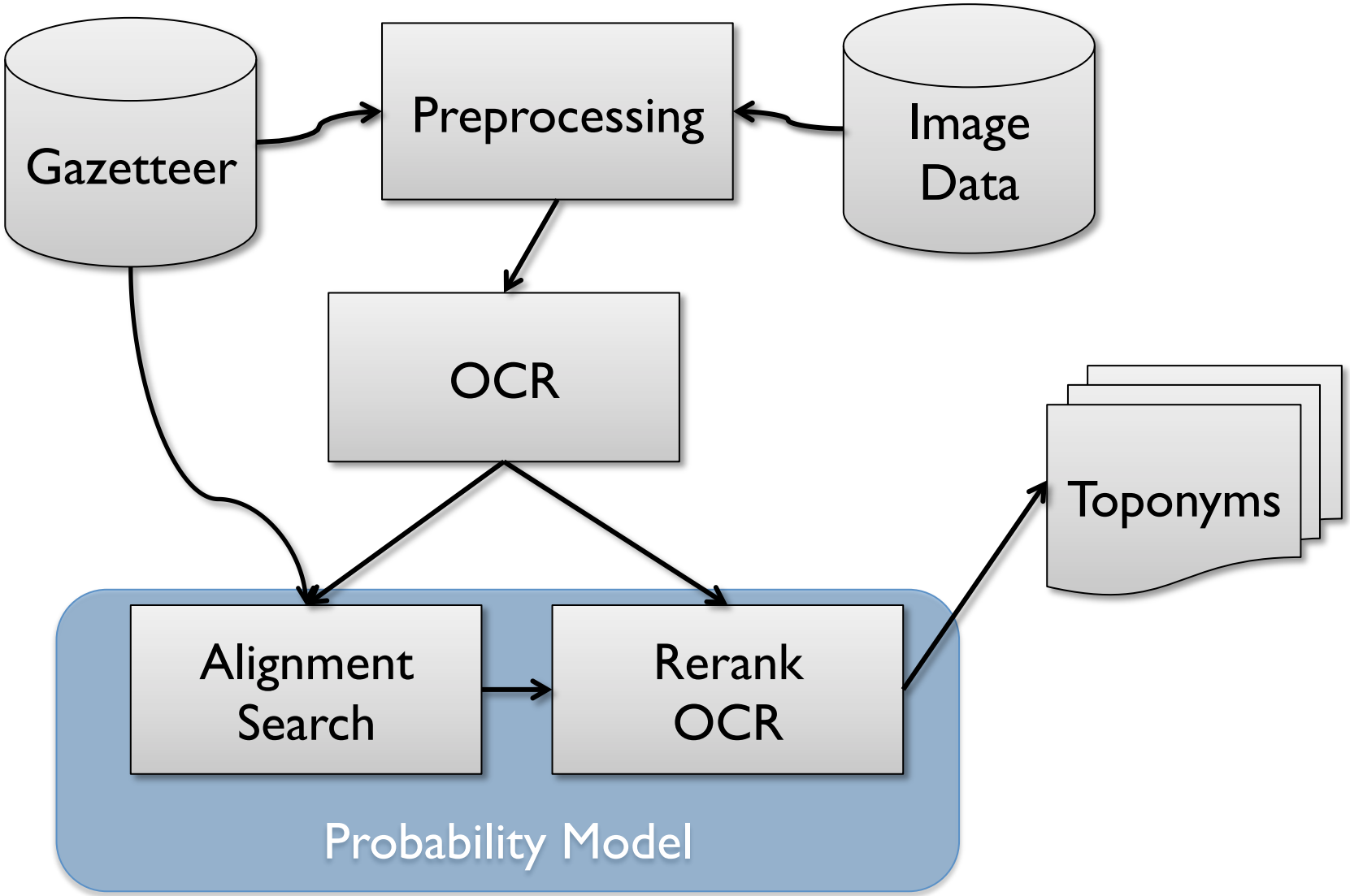
2. How well can we automatically align a map image to known geography using OCR?

3. How much does alignment improve word recognition?

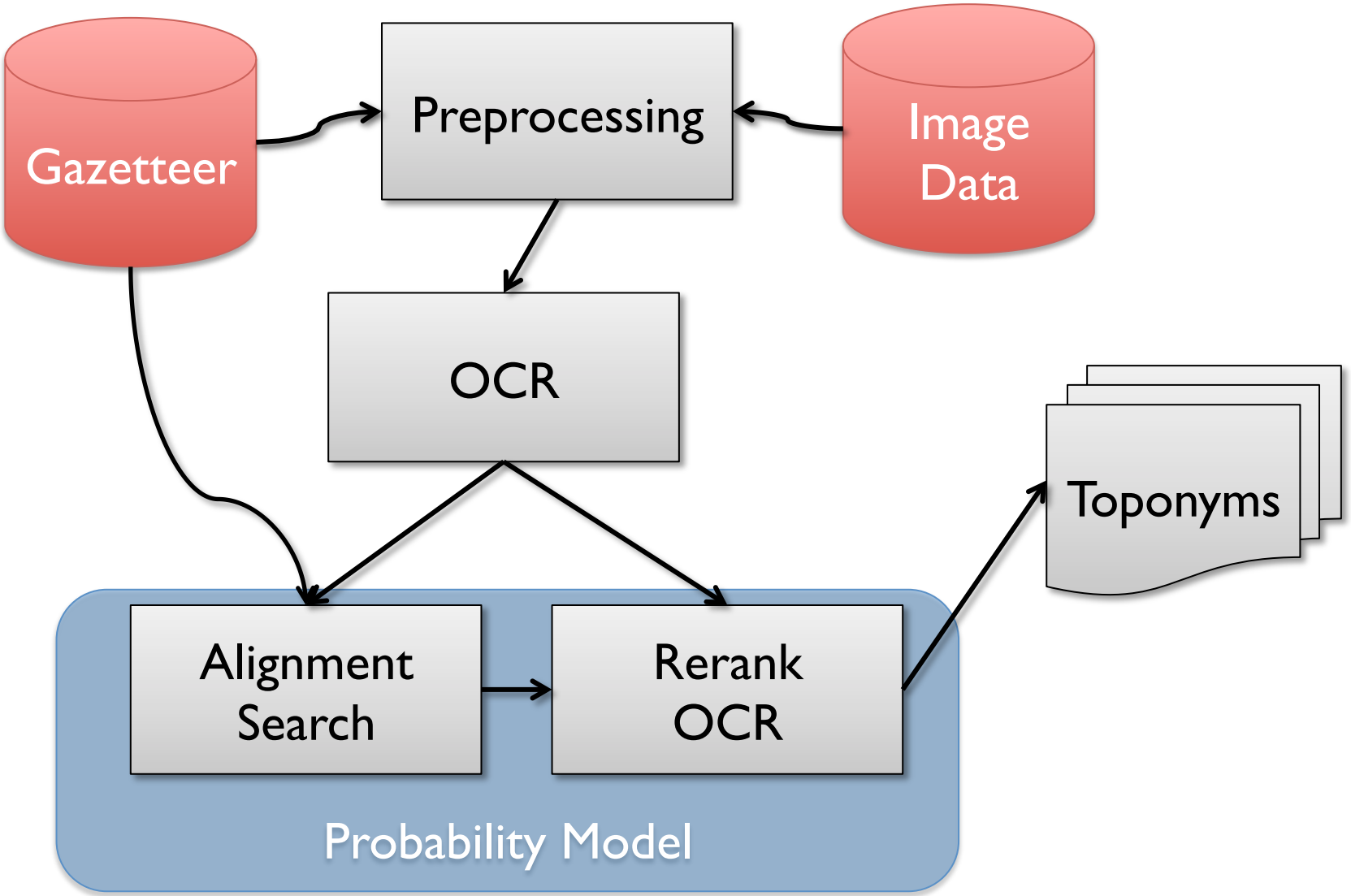
Contributions

- Robust search for coordinate alignment given word hypotheses
- Bayesian update of word hypotheses given alignment
- Experimental data set

Overview


Gazetteer
Preprocessing
Image Data
OCR
Toponyms
Alignment Search
Rerank OCR
Probability Model

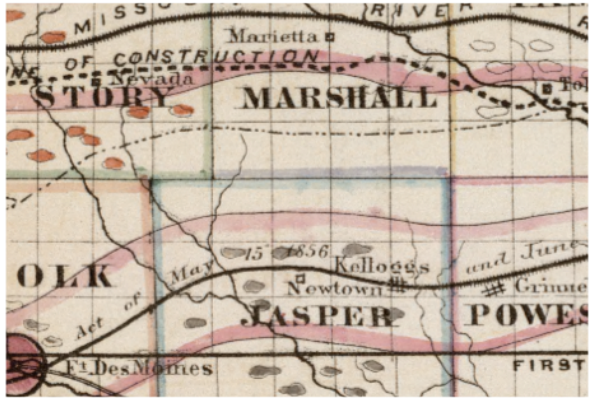
Overview

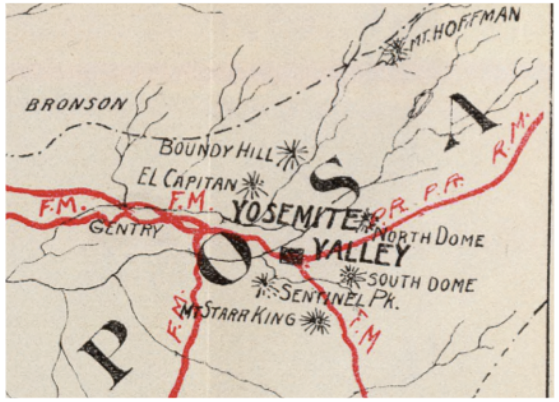

Gazetteer
Preprocessing
Image Data
OCR
Toponyms
Alignment Search
Rerank OCR
Probability Model

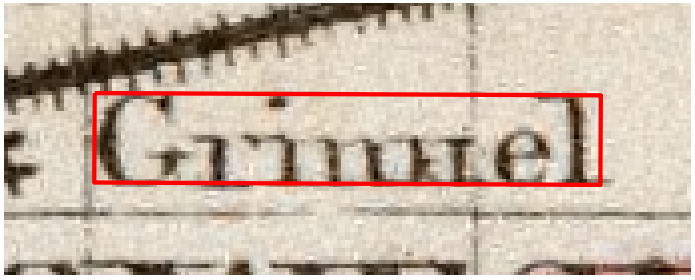
Experimental Data

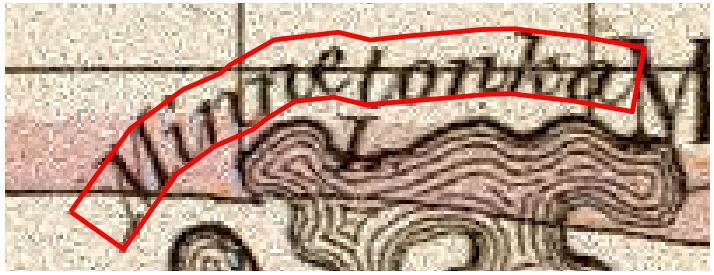
- Map Images
  - Nineteenth century US state and regional maps
  - Tight word bounding polygons, ground truth text
  - `http://hdl.handle.net/11084/3246`

Experimental Data









http://hdl.handle.net/11084/3246

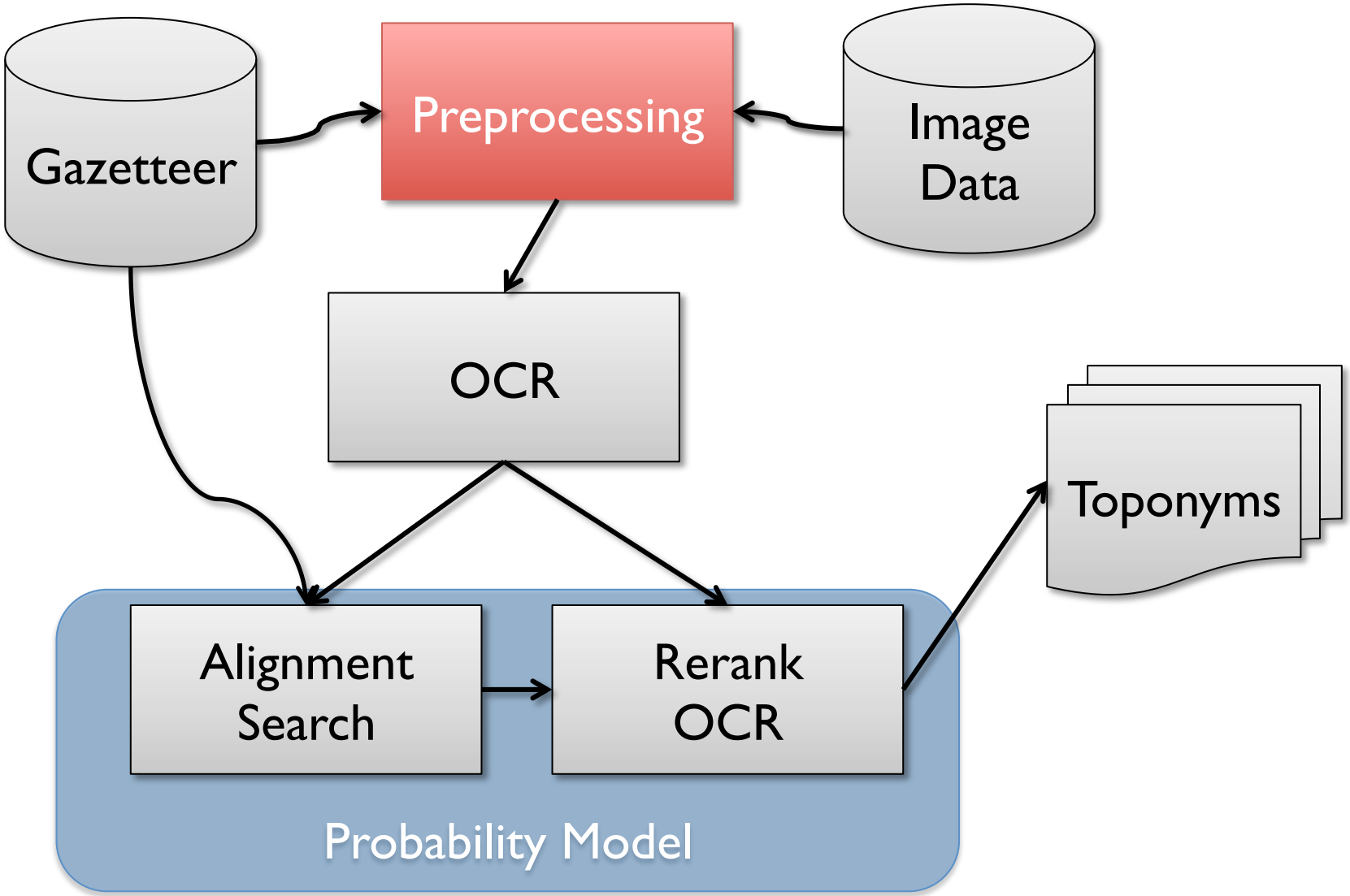
davidrumsey.com

Central Calif. (1896), Geo. W. Blum
Florida (1875), Cram Atlas Company
Iowa (1866), U.S. General Land Office
Minnesota (1866), U.S. General Land Office

Experimental Data

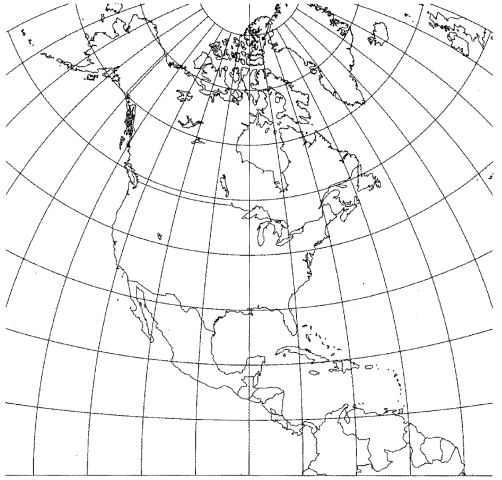
- Map Images
  - Nineteenth century US state and regional maps
  - Tight word bounding polygons, ground truth text
  - `http://hdl.handle.net/11084/3246`
- Gazetteer
  - USGS Board of Geographical Names
  - Two million entries in sixty-five categories
  - Historical features and alternate names/spellings
  - Primary lon/lat coordinates, county, state

Overview

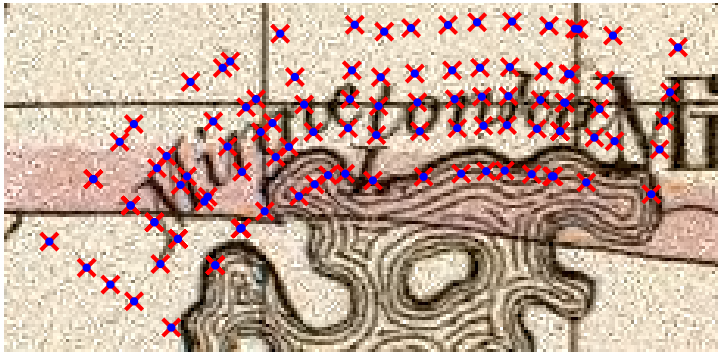

Gazetteer
Preprocessing
Image Data
OCR
Toponyms
Alignment Search
Rerank OCR
Probability Model

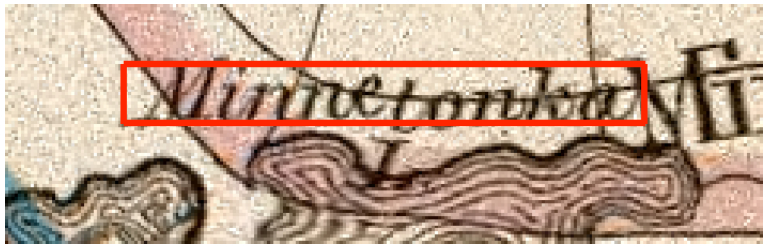
Preprocessing

- Gazetteer restricted to map region
- Spherical → 2D Projection
  - Earth shape: Clarke 1866 geoid
  - Hassler polyconic
- Normalize word images for OCR



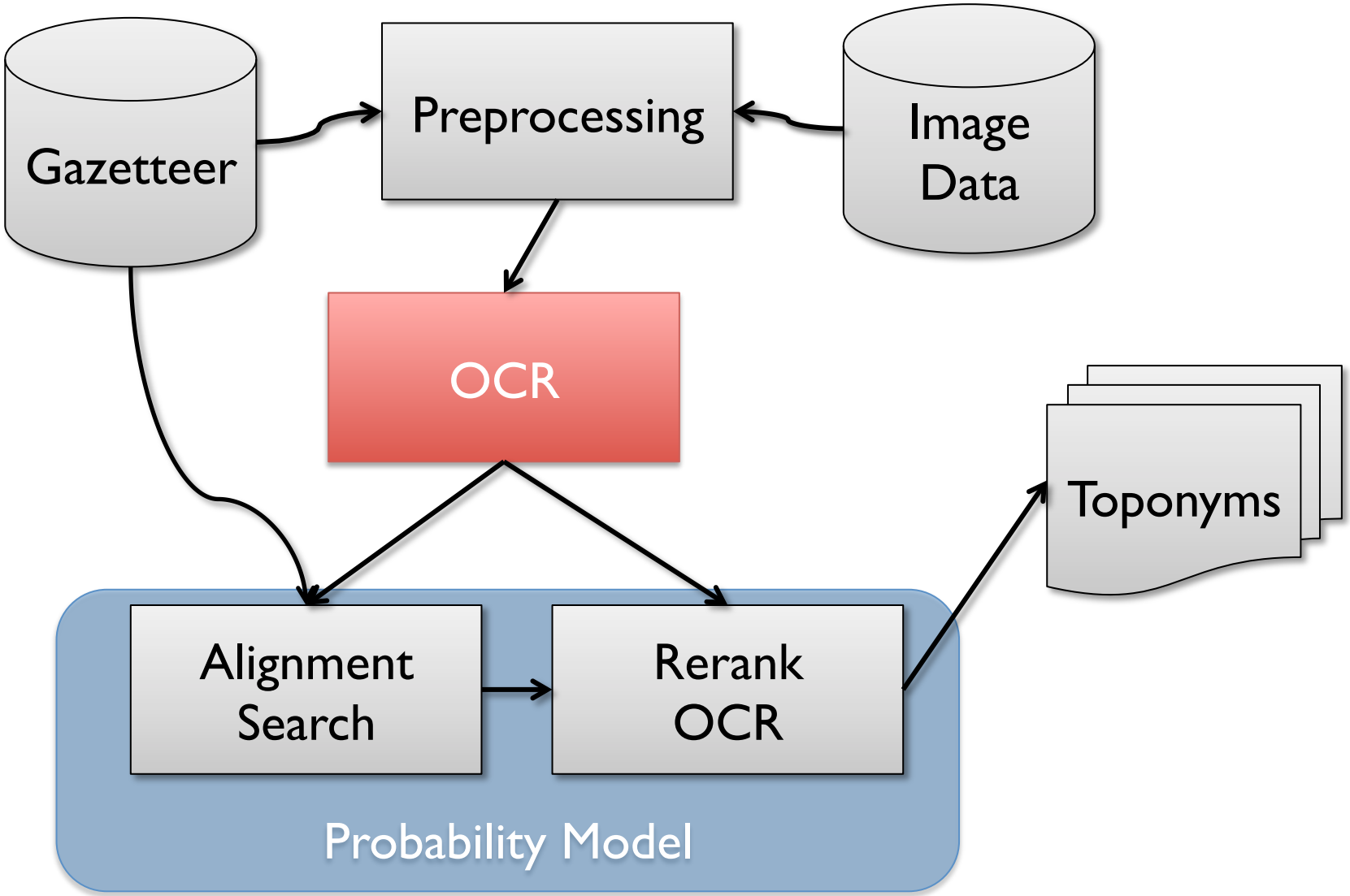
[Snyder 1989, USGS PP 1453]





Minnesota (1866), U.S. General Land Office, davidrumsey.com

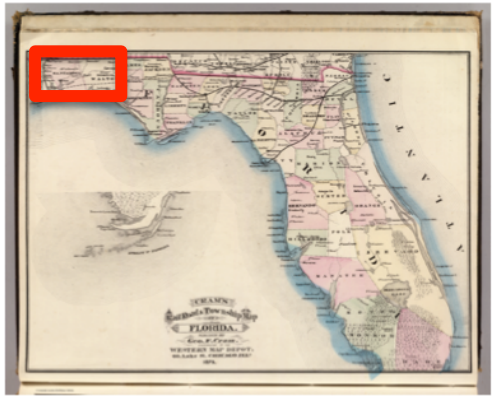
Overview

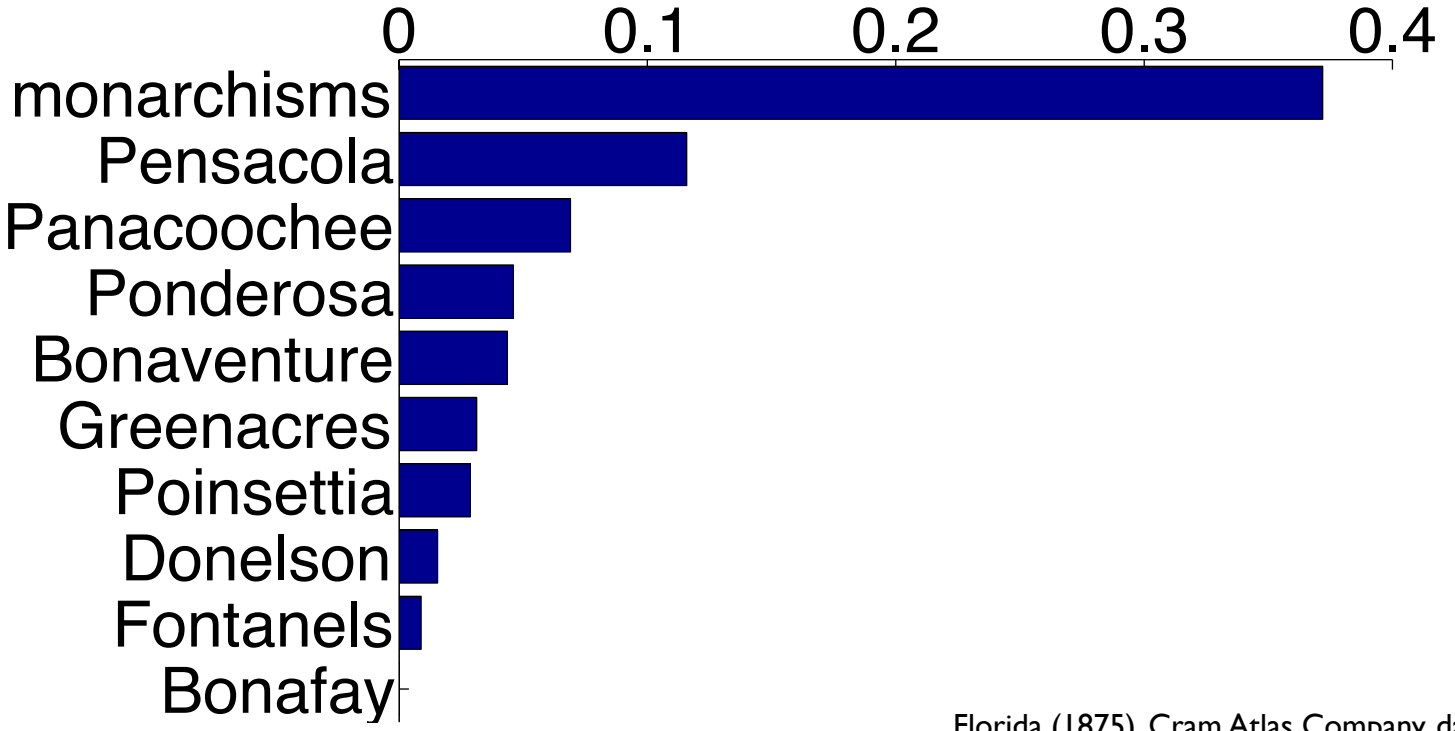

Gazetteer
Preprocessing
Image Data
OCR
Toponyms
Alignment Search
Rerank OCR
Probability Model

Text Recognition

- Discriminative Semi-Markov Model
  Weinman *et al.* Toward Integrated Scene Text Reading (PAMI 2013)
  - Unified segmentation/recognition
  - Lexicon (82,000 English words; Gazetteer words)
- Trained for robust reading competition
- ★ Produces a list of top word probabilities

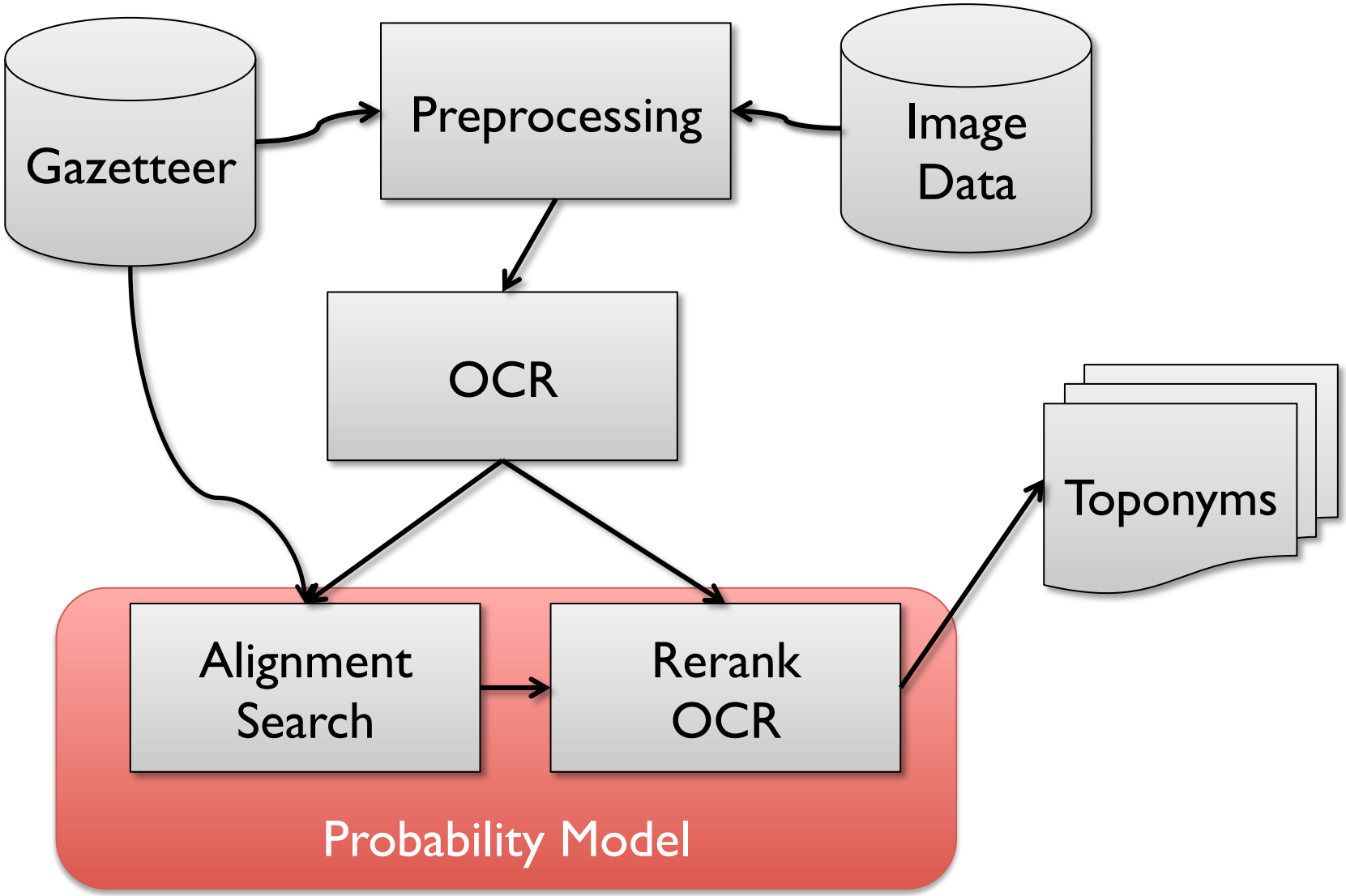
Text Recognition



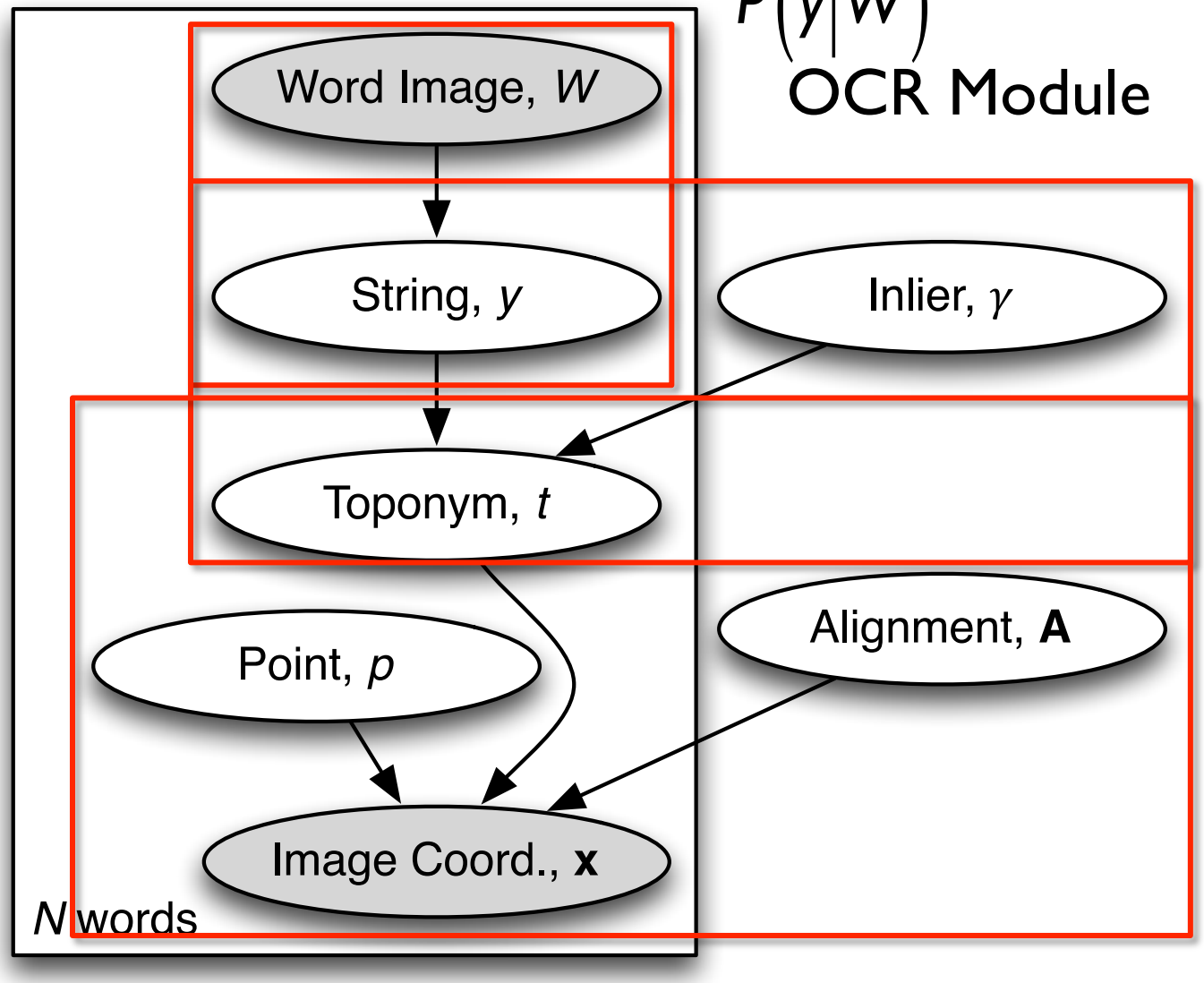


Florida (1875), Cram Atlas Company, davidrumsey.com

Overview


Gazetteer
Preprocessing
Image Data
OCR
Toponyms
Alignment Search
Rerank OCR
Probability Model

Probability Model



$P(y|W)$

OCR Module

$P(t|y,\gamma) \propto$

Match(y,t)

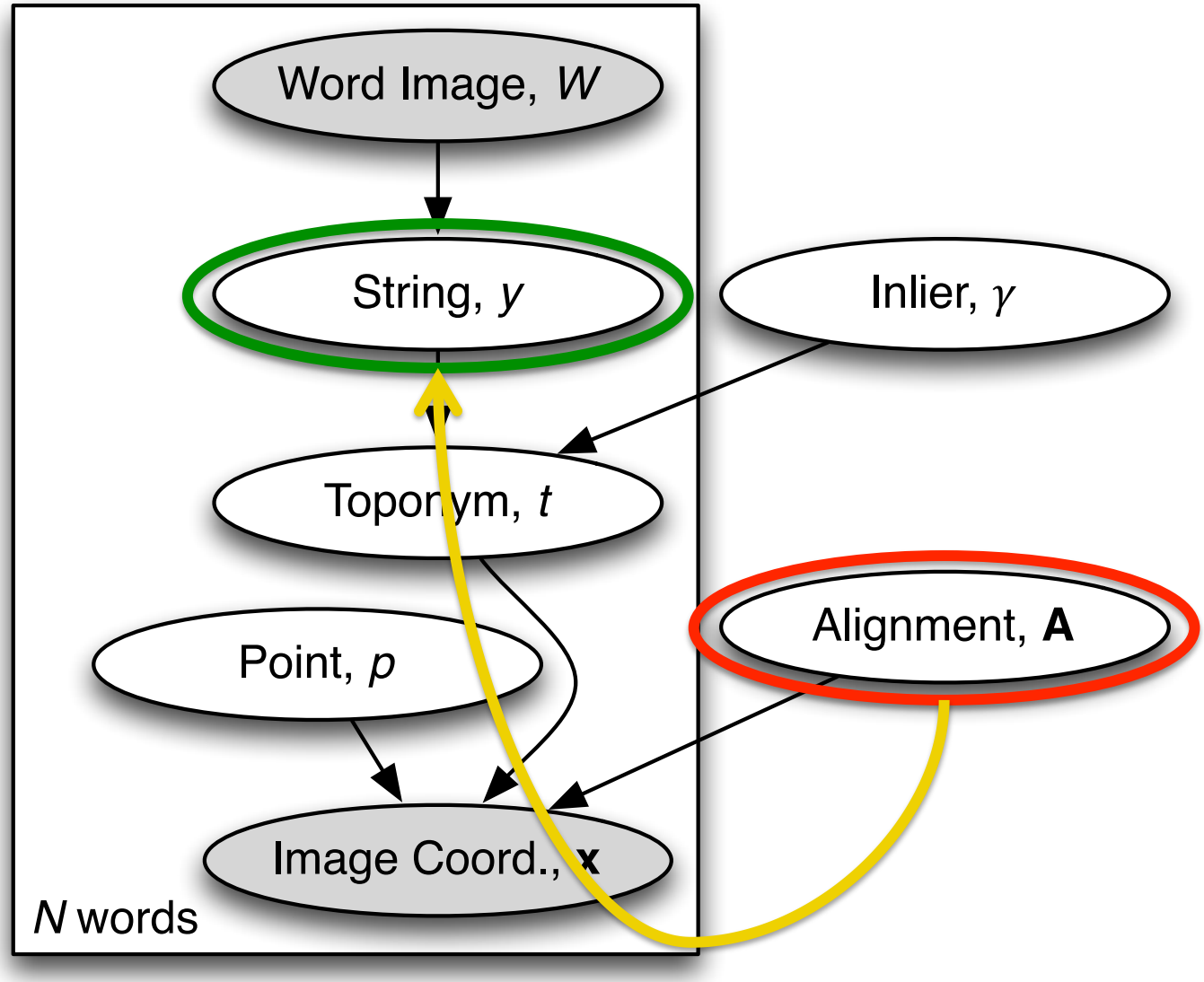
$P(t_0|y,\gamma) = 1-\gamma$

$P(\vec{x}_p|p,\mathbf{A},t) =$

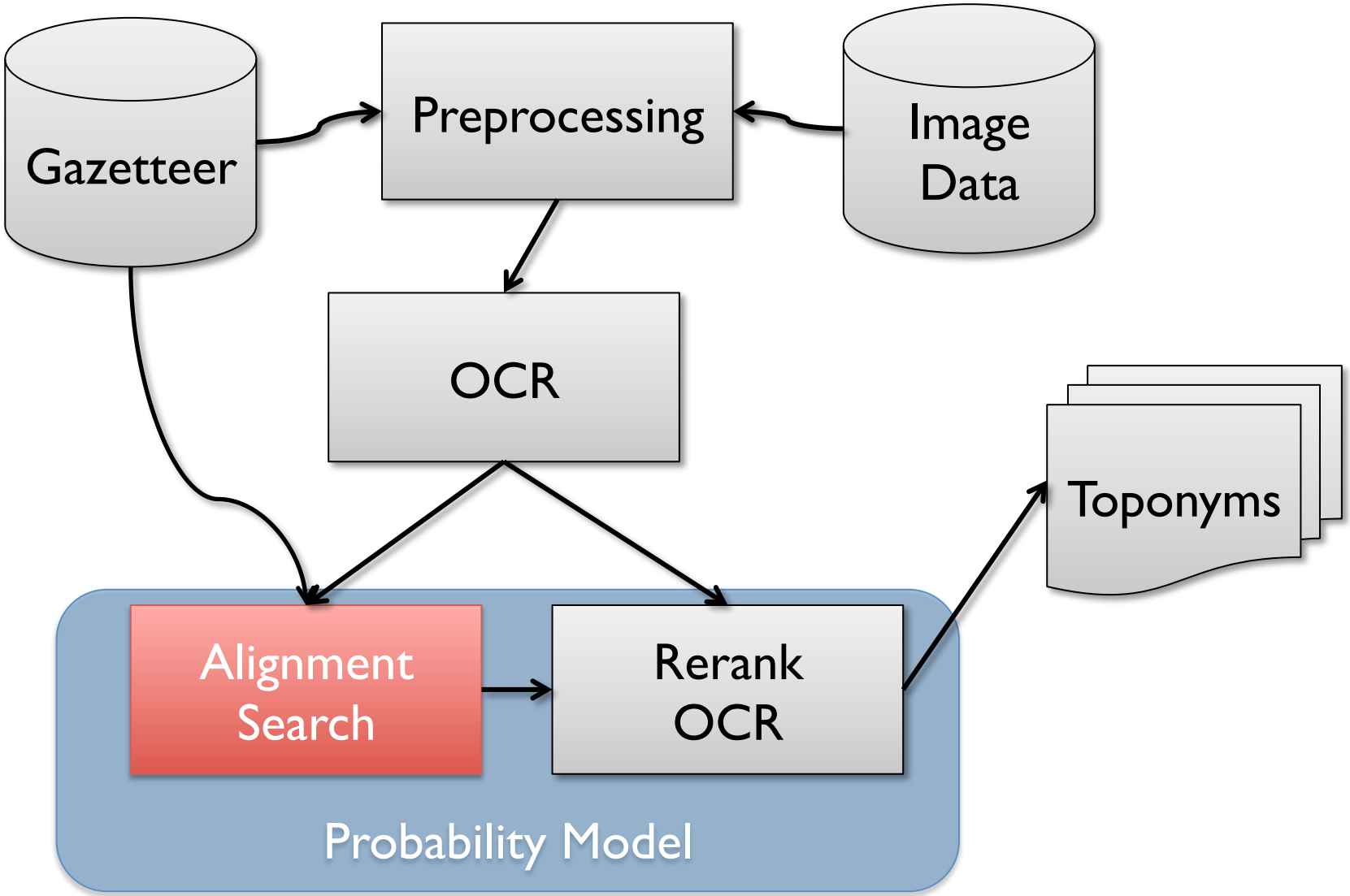
$\mathcal{N}(\vec{x}_p;\mathbf{A}\vec{c}_t,\sigma^2)$

$P(\vec{x}_p|p,\mathbf{A},t_0) \propto 1$

Probability Model



Overview


Gazetteer
Preprocessing
Image Data
OCR
Toponyms
Alignment Search
Rerank OCR
Probability Model

Alignment Search

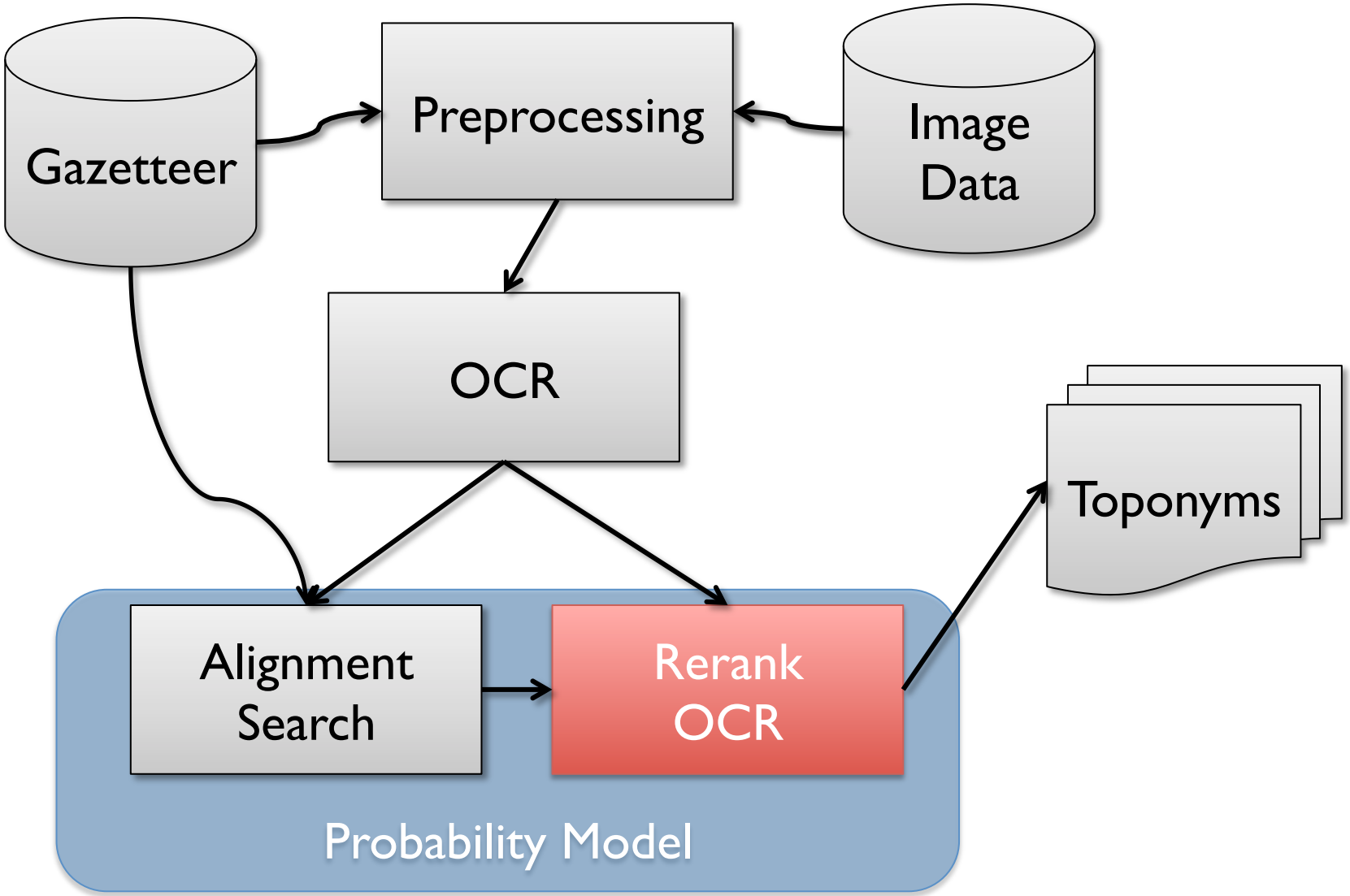
- Generalized RANSAC / MLESAC Variant
  Fischler & Bolles (CACM '81), Torr & Zisserman (CVIU '00),
  Zhang & Kosecká (3DDPVT '06)
- Repeat
  - Choose three words
  - Sample image point, toponym correspondences
  - Calculate affine transform $\mathbf{A}$
  - Estimate inlier bias γ
  - Score marginal probability $P(\mathbf{x} \mid \mathbf{A}, \gamma, W)$
- Return best-scoring parameters $\hat{\gamma}, \hat{\mathbf{A}}$

Alignment Search

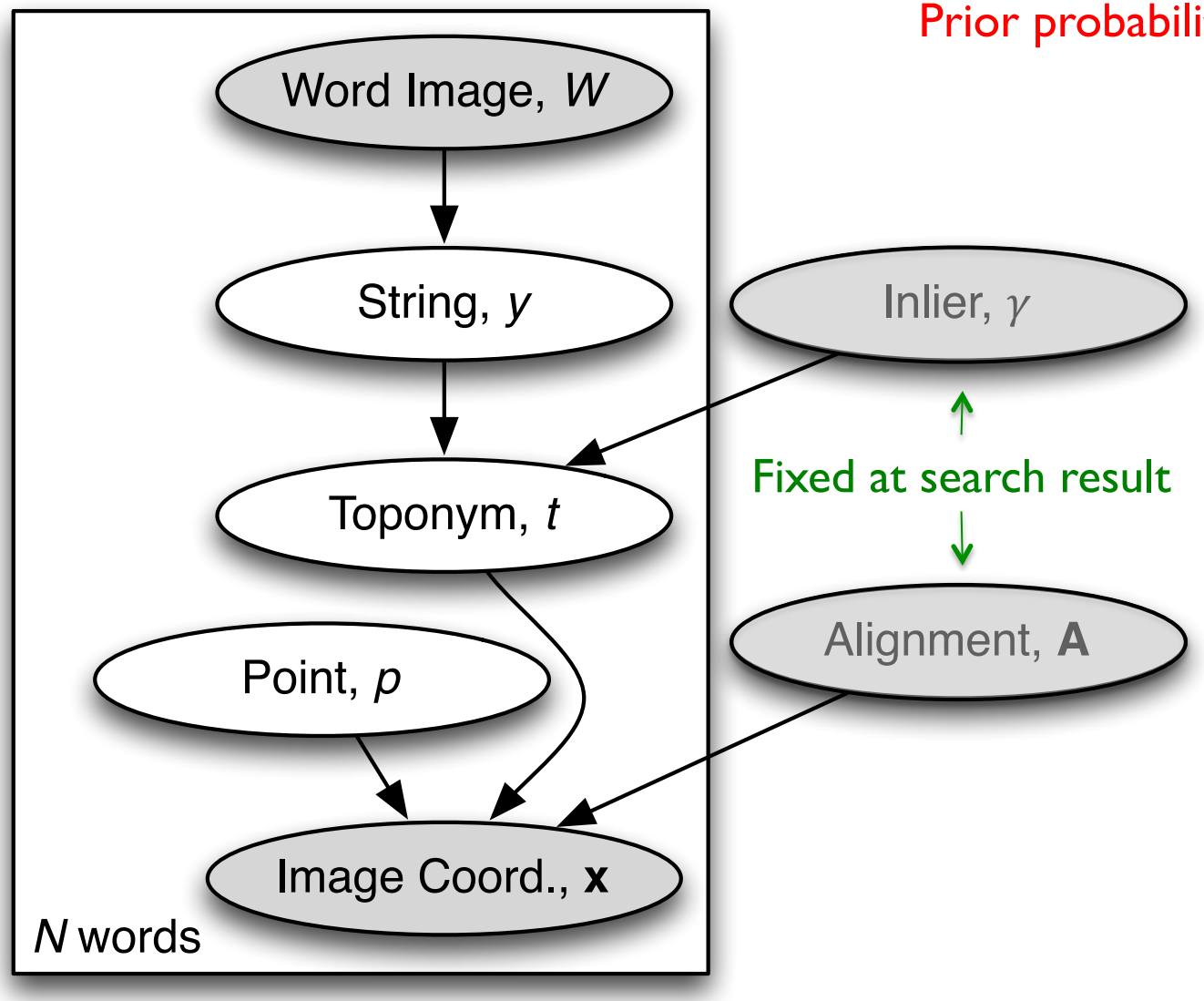
Iter.	Log Lik.	Inliers
40	-4487	3
86	-4403	7
156	-4244	12
187	-4120	16
628	-3192	19
721	-3693	29
1352	-3523	44

Ohio (1866), U.S. General Land Office, davidrumsey.com

Overview


Gazetteer
Preprocessing
Image Data
OCR
Toponyms
Alignment Search
Rerank OCR
Probability Model

Rerank OCR

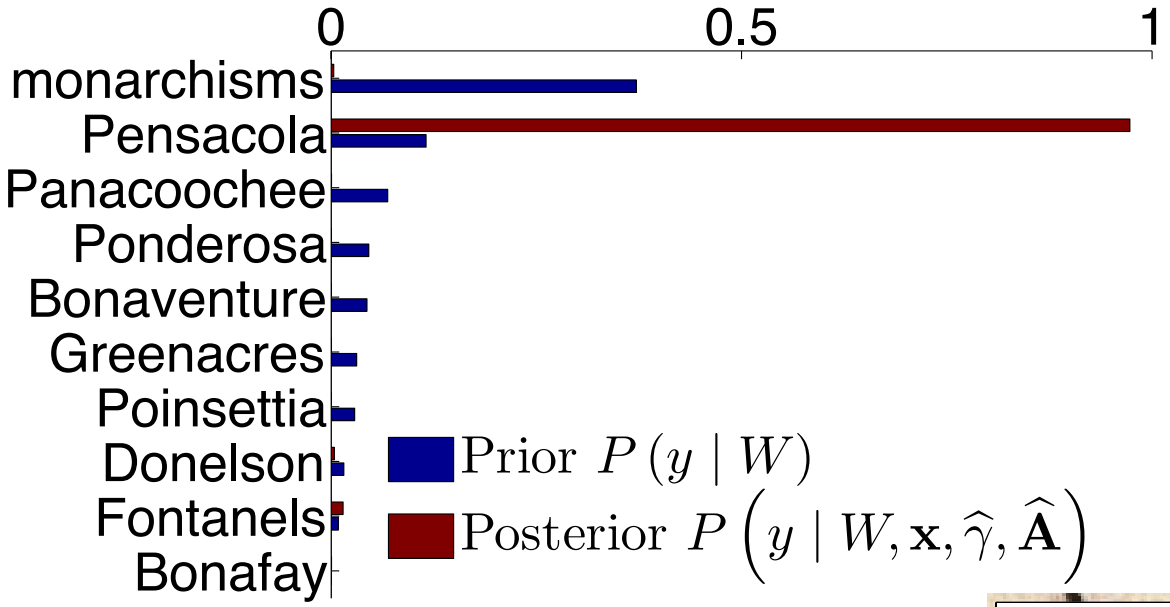


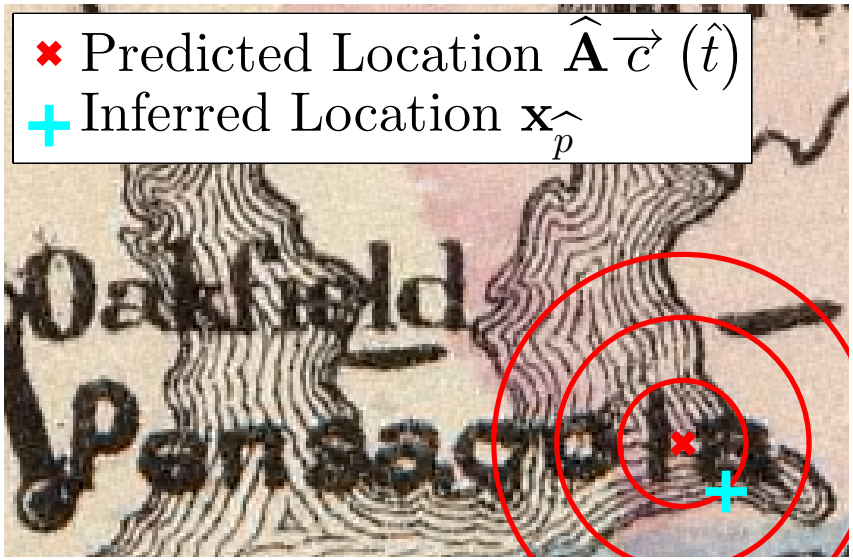
Prior probability $P(y|W)$

Calculate posterior probabilities

$P(y|\mathbf{x}, \hat{\gamma}, \hat{\mathbf{A}}, W)$

Text Recognition





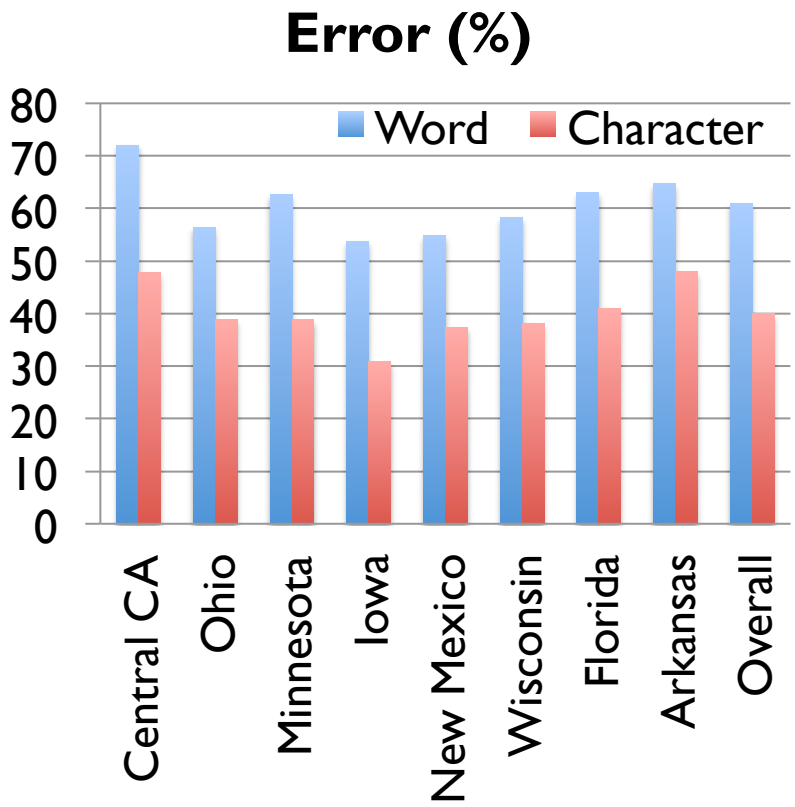
Florida (1875), Cram Atlas Company, davidrumsey.com

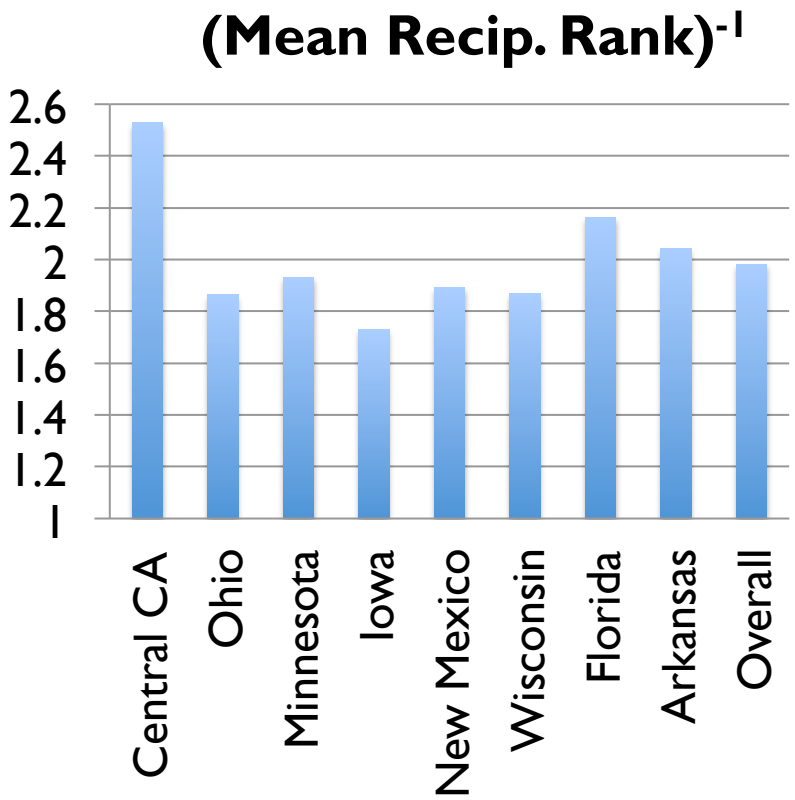
Experimental Results

- 12 carefully annotated maps (+18 to verify)
  - Four for training/tuning
  - Eight for evaluation
- 2325 test words (291 ± 30 / map)
  - 1800 (77%) appear in gazetteer
  - 1585 (68%) appear in gazetteer and OCR list

Question 1

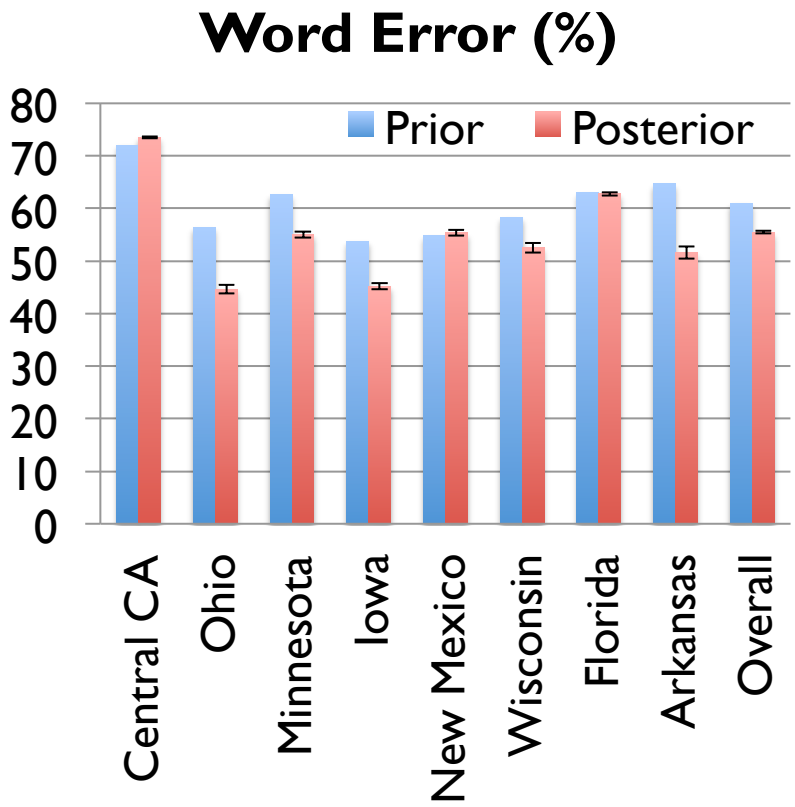
How accurate is a general robust word recognition system on historical maps?

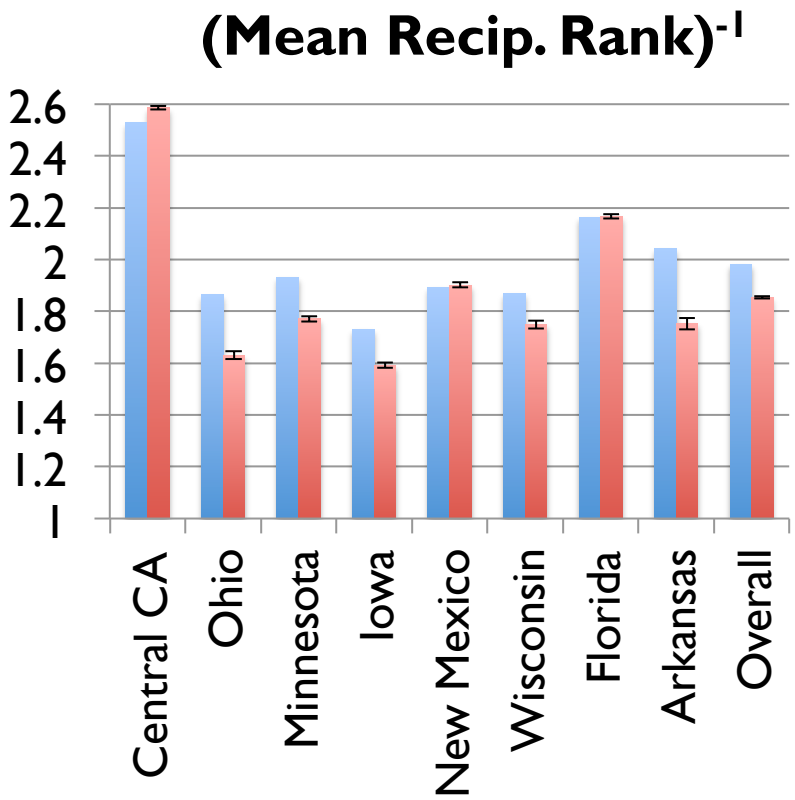




Question 3

How much does alignment improve word recognition?

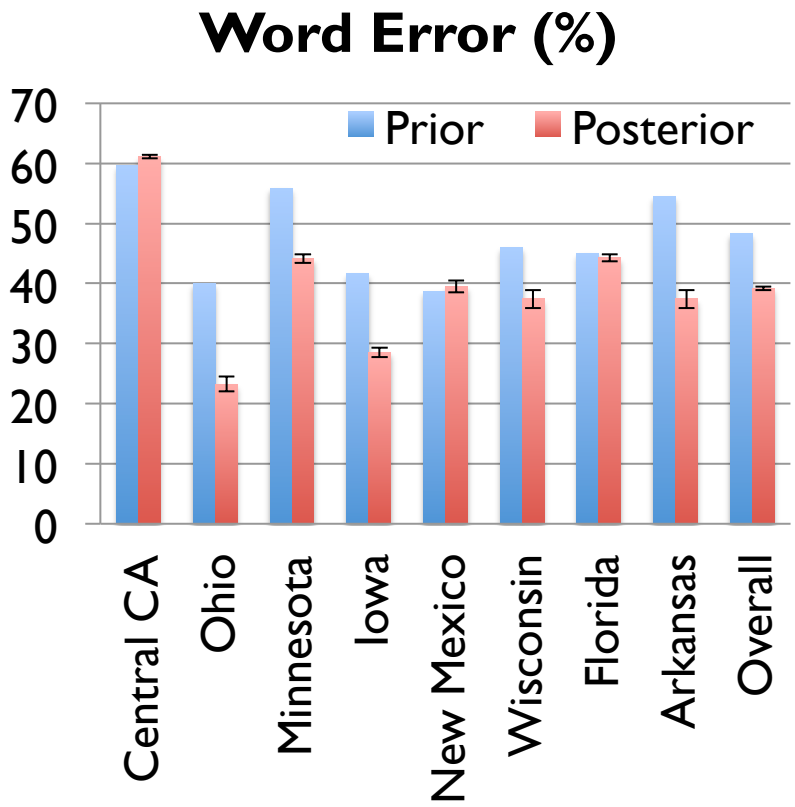


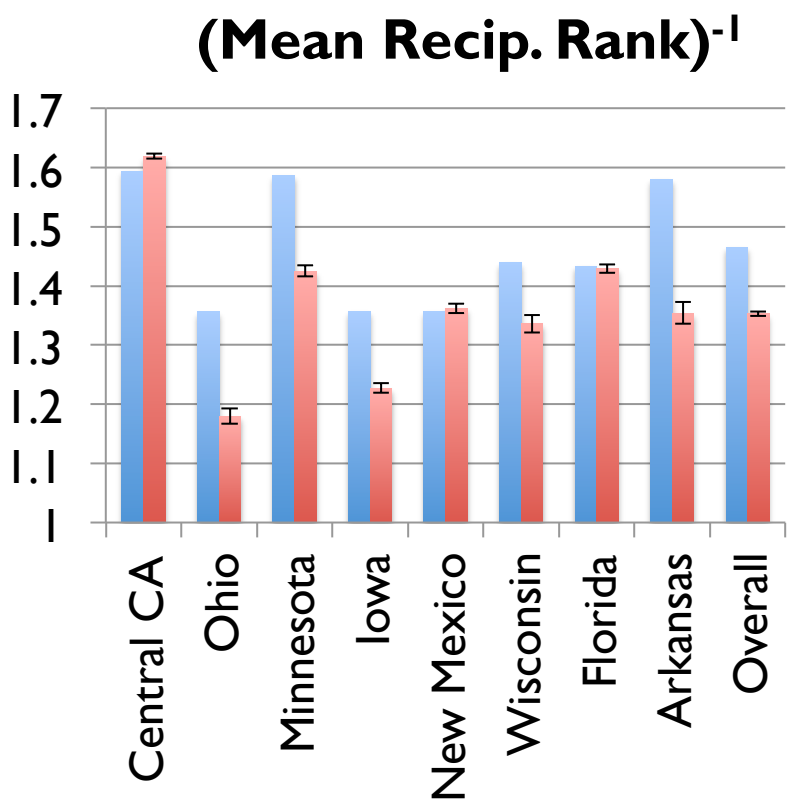


Question 3

How much does alignment improve word recognition?

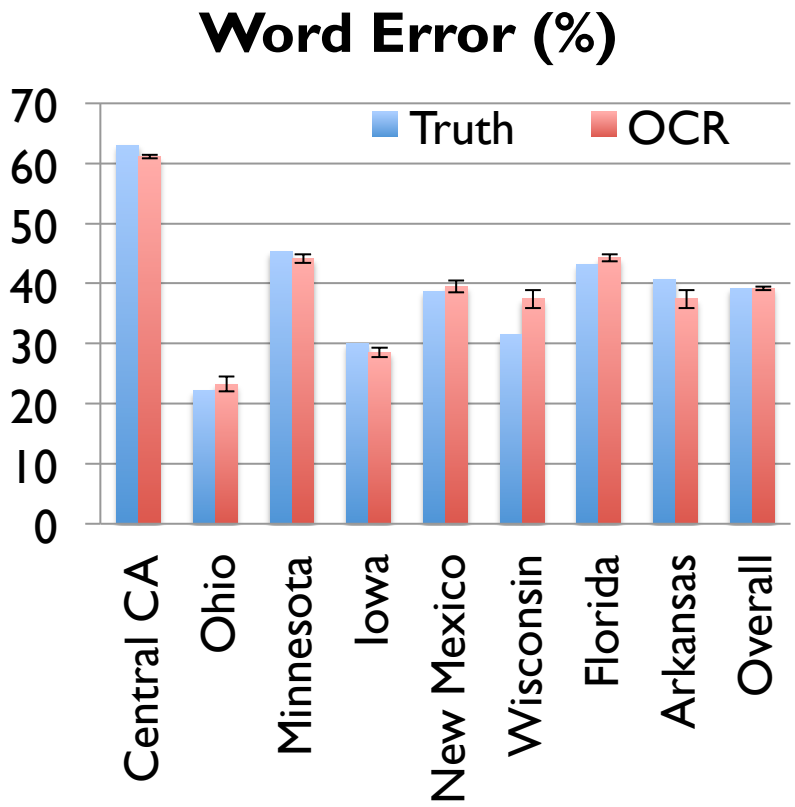
improveable

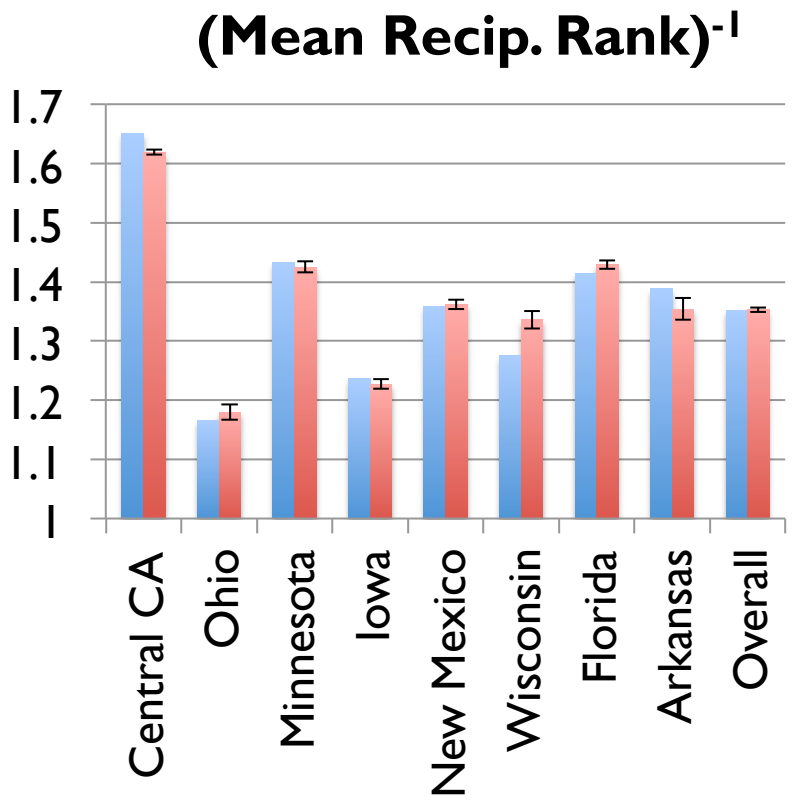




Question 2

How well can we automatically align a map image to known geography using OCR?

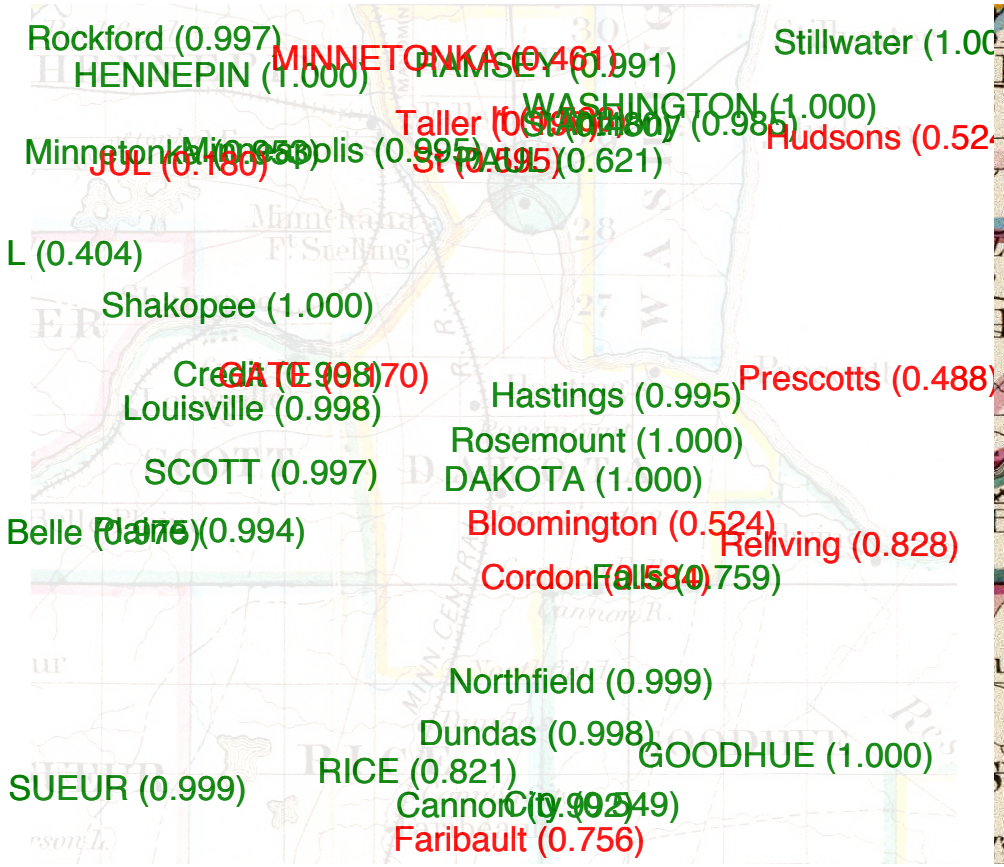


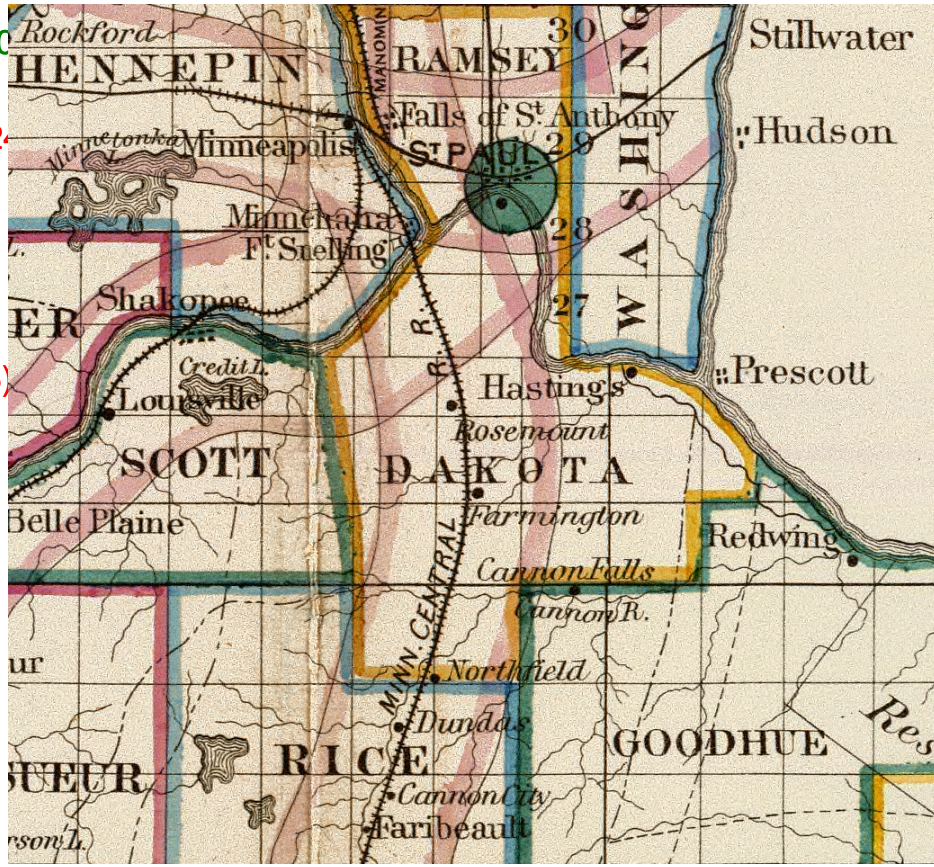


Conclusions

- Robust word recognition works well enough
- RANSAC search usually finds good alignments
- Gazetteer improves word accuracy by ≈20%
- Future work
  - Expectation-maximization; adapt error scale
  - Restrict OCR vocabulary using initial alignment
  - Expand feature classes
  - Incorporate regional and linear toponym likelihoods
  - Integrate word detection

Discussion





Minnesota (1866), U.S. General Land Office, davidrumsey.com