Style-Based Retrieval for Ancient Syriac Manuscripts

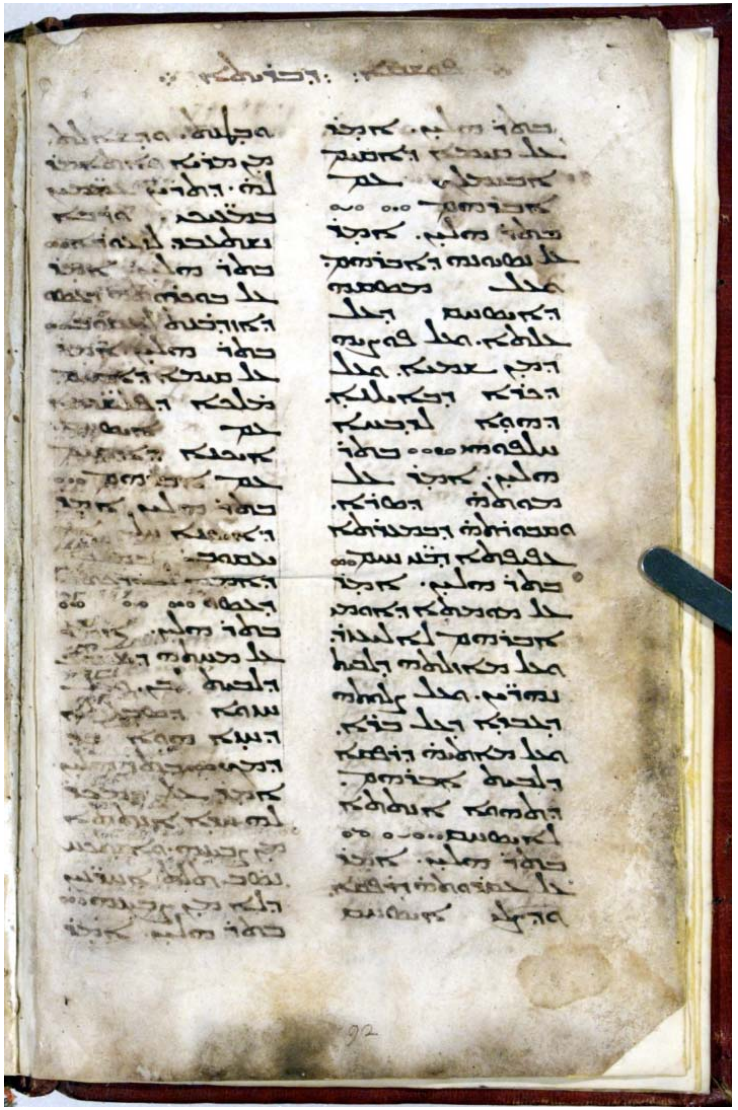
Emma Dalton
Simon Fraser University

Nicholas R. Howe*
Smith College

*prese

What is Syriac?

- Script used in Levant (Syria, Turkey, etc.)
- Over 10K manuscripts produced 400-1200 AD
- Significant early Christian documents
- Variant of Aramaic
  - Reads right-to-left
  - 22 letter alphabet



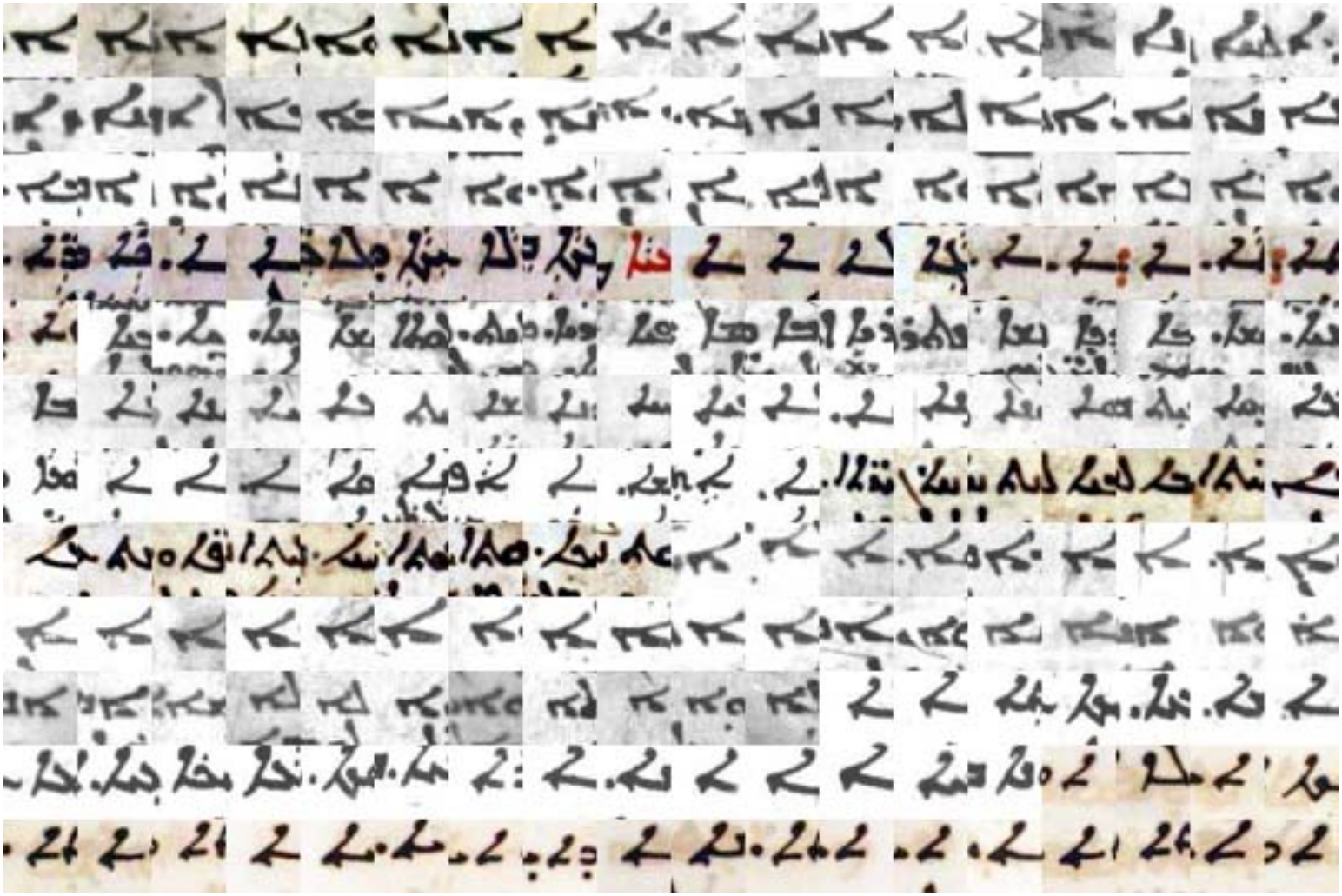
Geographic Extent

Need for Automation

- Few modern scholars read Syriac
- Most documents unexamined
- Need automated tool to guide research
  - 5% of documents identify scribe
  - 95% unattributed

- *Can we identify related documents by the writing style?*

216 Alaphs

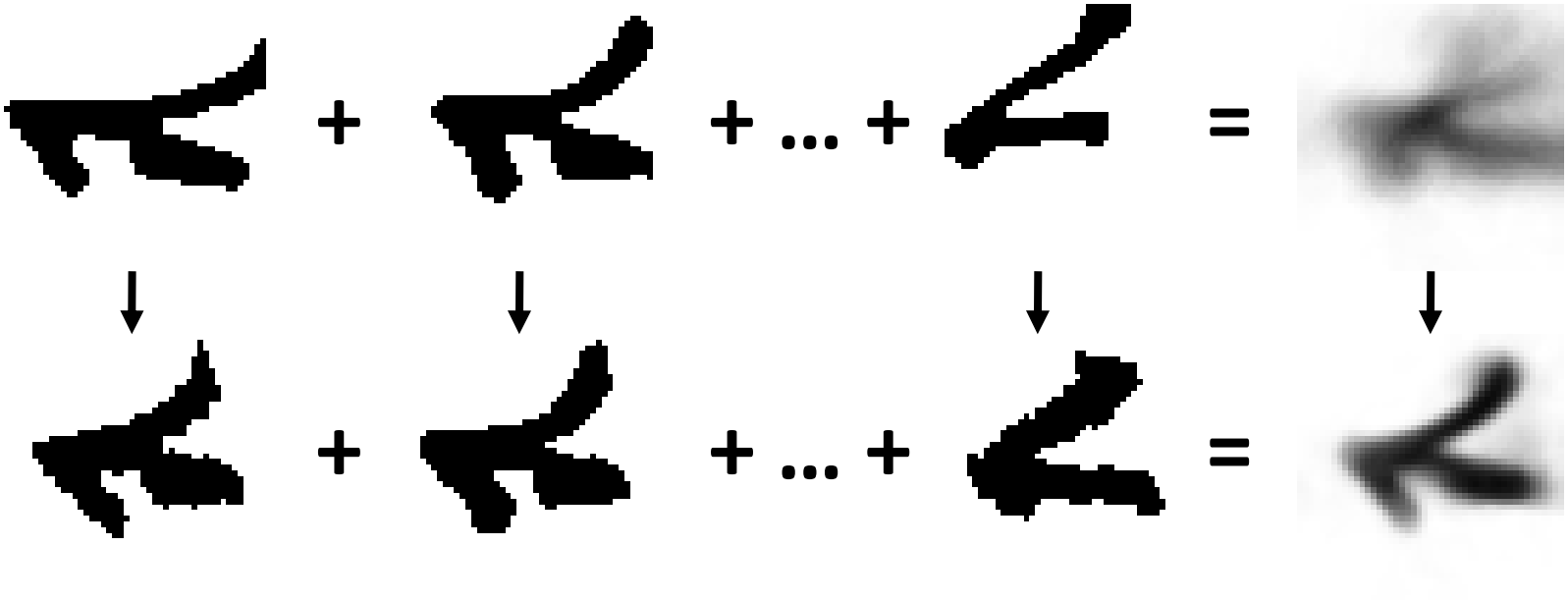


Style Variations

- Axiom: Each letter has a Platonic ideal form
  - Individual letter instances are variants of this
  - Ideal form some kind of “average” of all instances
- Variations from ideal may be global or local
  - Global: tall & skinny
  - Local: enhanced curve on ascender

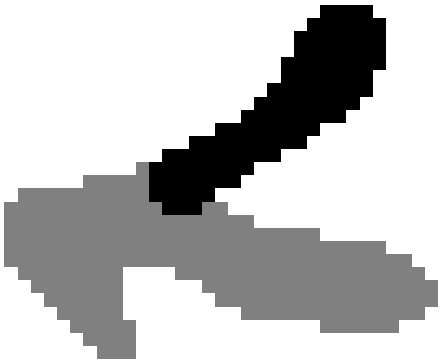
Exploring Plato's Cave

- **Congealing** [Learned-Miller '05]
  - Aligns set of character samples via entropy-minimizing affine transforms
  - Mean aligned image is ideal form



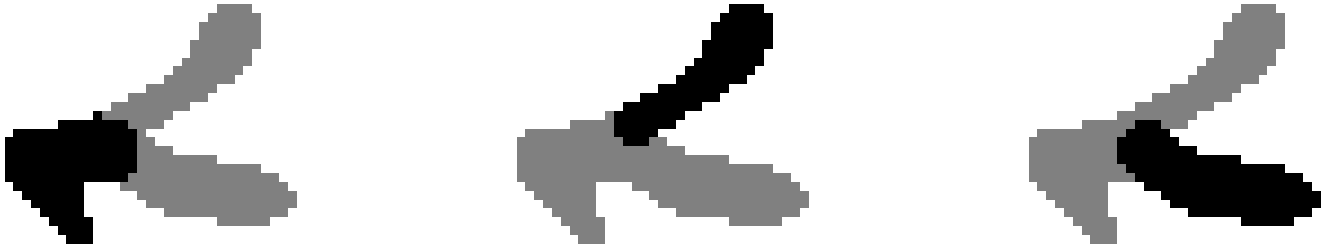
Exploring Plato's Cave (2)

- Standard congealing finds global variations
  - Scale, translation, rotation all ignored
  - Shear & aspect ratio significant
- Need way to capture local variations also
- Solution: describe isolated letter components
  - Use second round of congealing
  - All affine parameters used

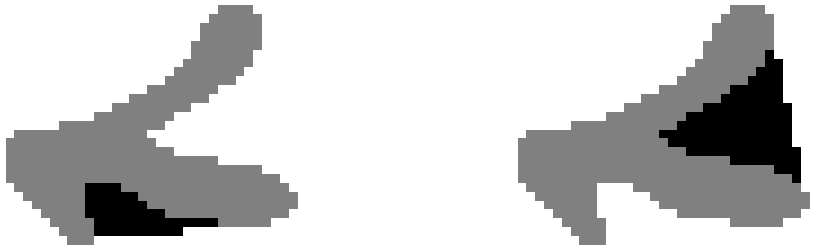


Letter Components

- Break letter form at skeleton junction points
- Reconstitute each segment

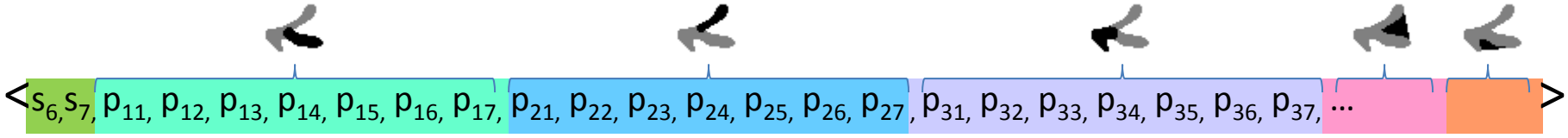


- Also use major letter cavities [Bar-Yosef ‘07]



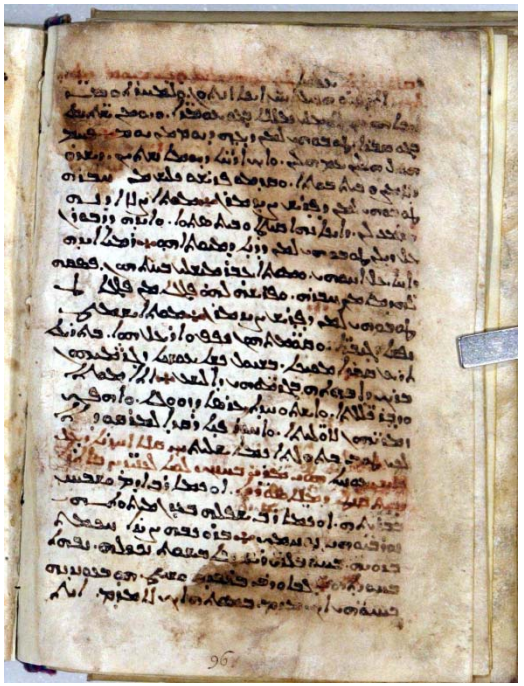
Outline of Method

- Prepare pages (scale, orient, binarize)
- Human annotator identifies character samples
- Samples for each letter congealed to find ideal
- Parts & cavities identified in ideal form
  - Corresponding component found in each sample
  - Second congealing over each sample component
- Sample descriptor composed of select global & all local affine parameters



Comparing Documents

- We have ***n*** documents, each with ***m*** samples
  How do we compare them?



Comparing Documents (2)

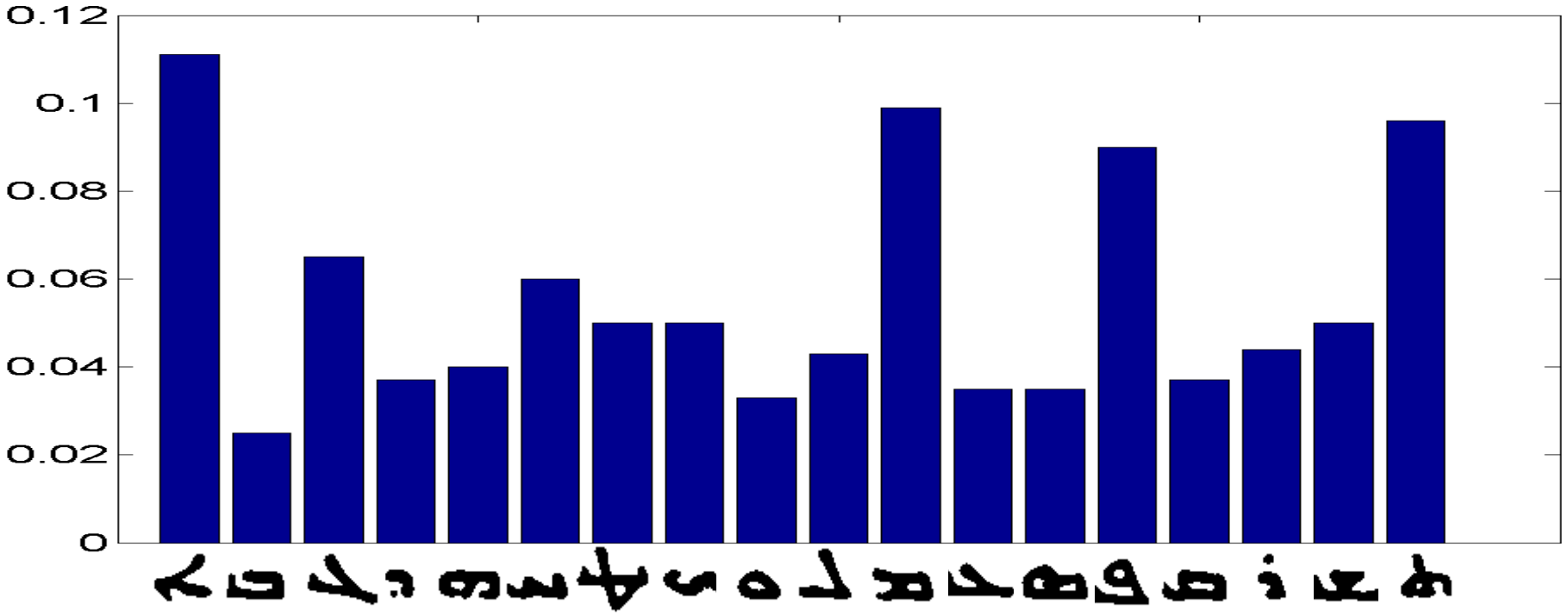
Simple Voting: All samples vote for the document containing their most similar match

Rank Voting: Each sample votes for documents based on the mean inverse ranks of its matches.

Weighted Rank Voting: Like rank voting, but some letter types count more (learned)

Weighted Rank Voting

- Greedy method: select best feature on training set, repeat until precision plateaus
- Folds show some variance



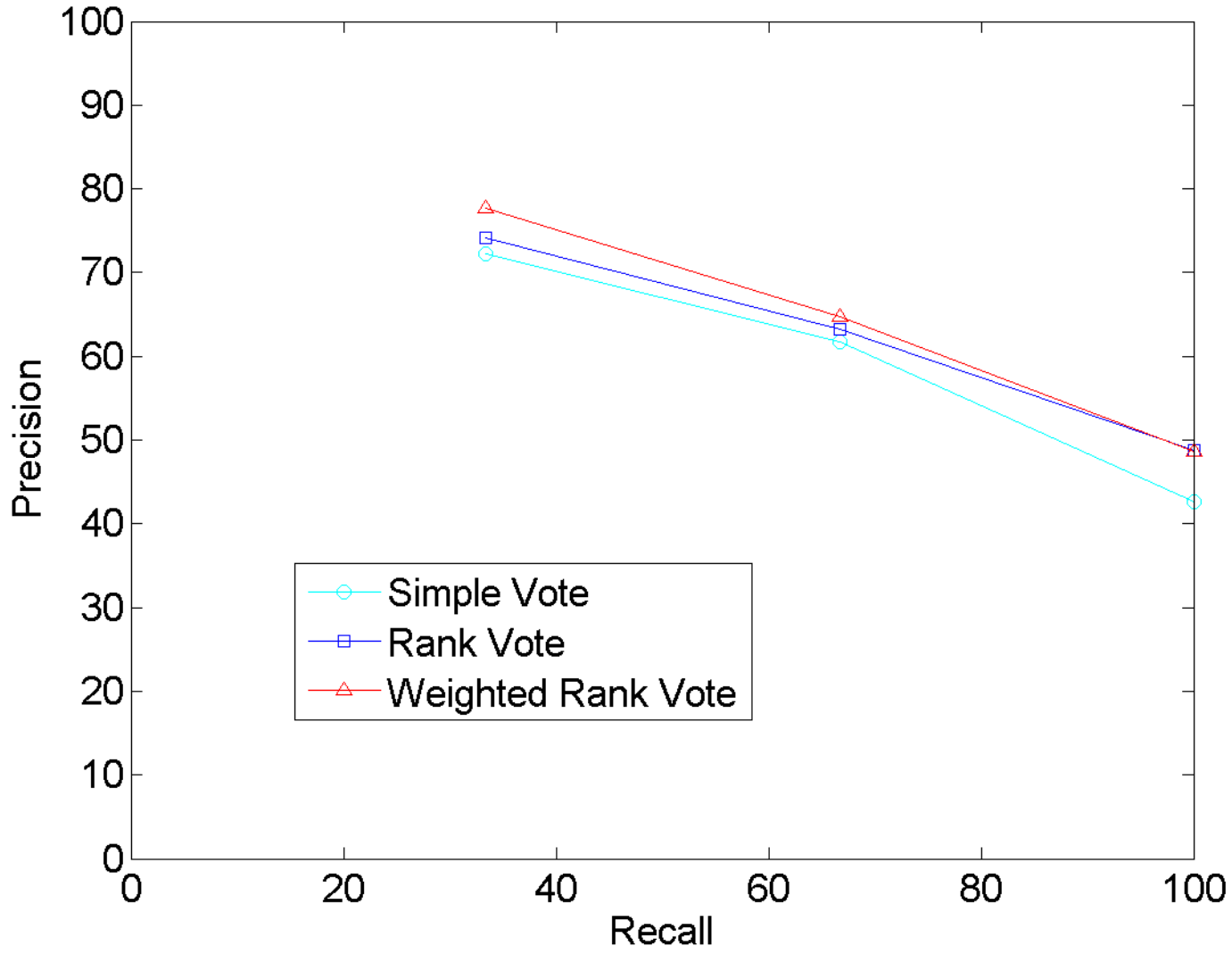
Experiment

- 19 manuscripts x 4 pages from each
- Four-fold cross validation
  - Each query page has three targets
  - Measure precision at 33%, 67%, 100% recall

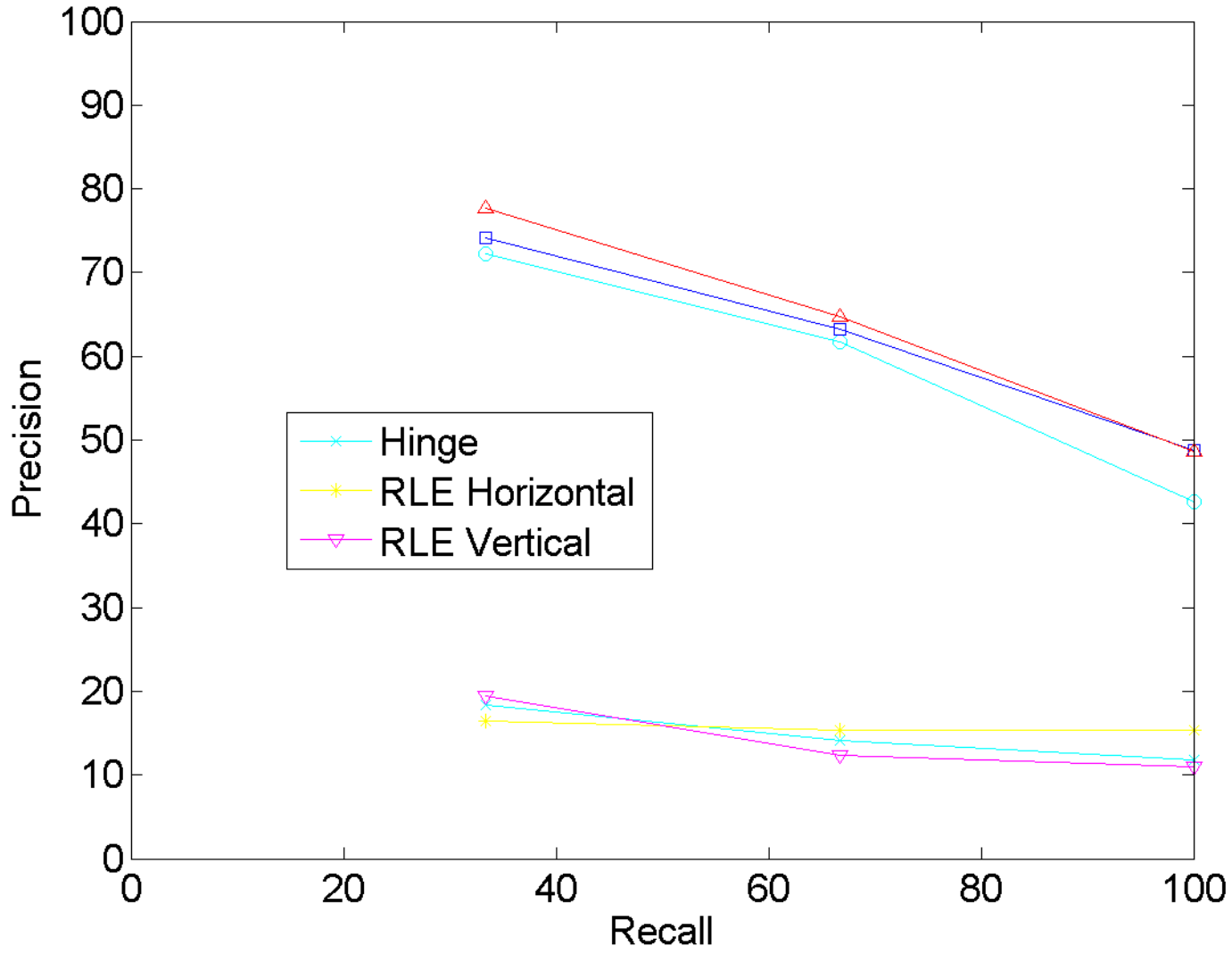
Computational Digression

- Congealing computation takes time:
  Not practical to complete for every query!
- Must align query sample to pre-congealed set
  - Record mean sample image and stabilization correction after each congealing round
  - Query congealing uses stored values for 2 & 3

Results


Precision
Recall
0
10
20
30
40
50
60
70
80
90
100
0
20
40
60
80
100
Simple Vote
Rank Vote
Weighted Rank Vote

Baseline Comparison



Conclusion

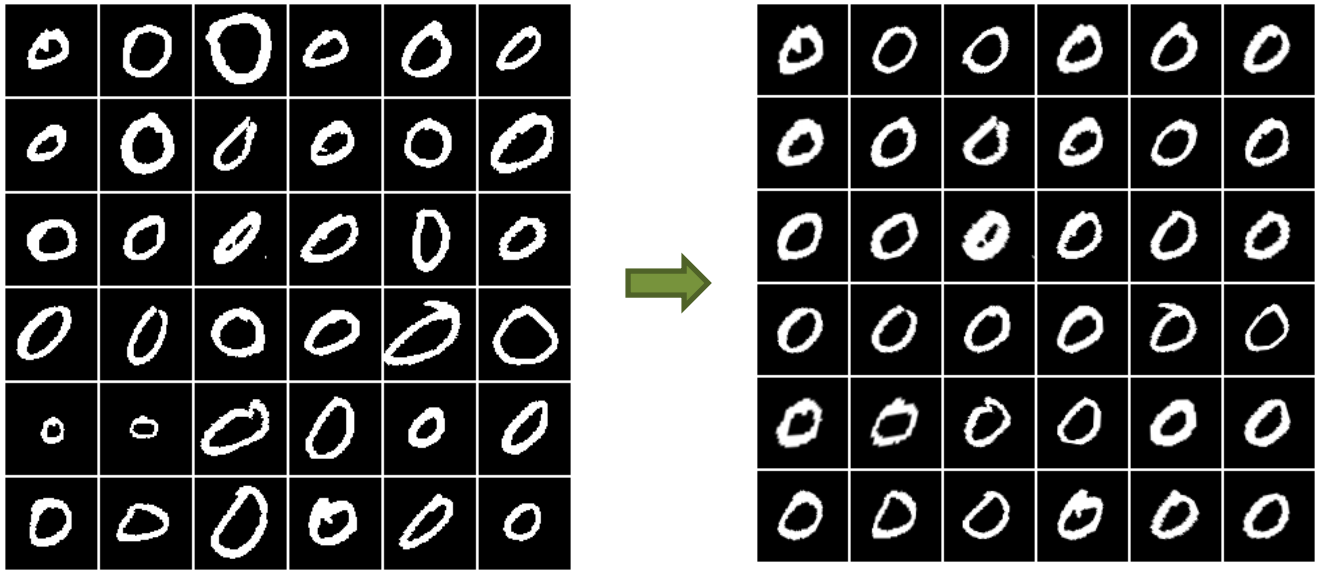
- Promising approach to style-based retrieval
- Needs testing on more data (coming!)
- Goal: Build working system to serve community of Syriac scholars

Acknowledgement

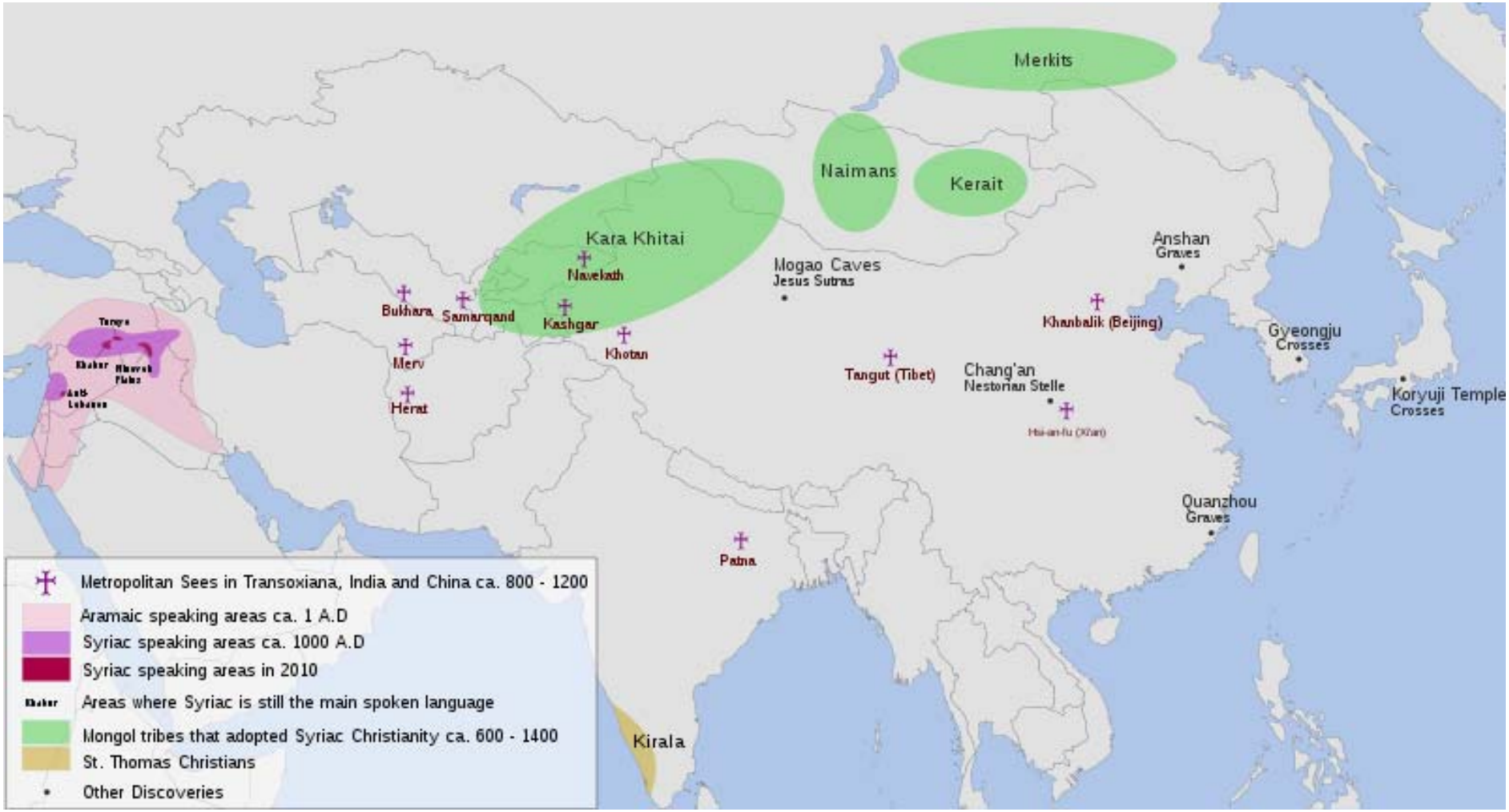
- Emma Dalton laid most of the groundwork in her 2010 senior thesis
- Michael Penn introduced me to Syriac and provided help in myriad ways
- Andreea Bancila is working on the interface and implementation for th working system

Exploring Plato's Cave

- **Congealing** [Learned-Miller '05]
  - Aligns set of character samples via entropy-minimizing affine transforms
  - Mean aligned image is ideal form



Geographic Extent



Letter Weights