A Hidden Markov Model for Alphabet-Soup Word Recognition

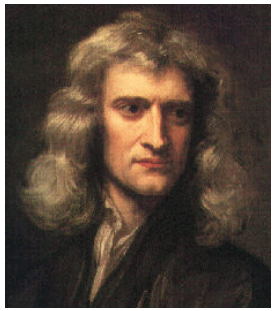
Shaolei Feng[1]
Nicholas R. Howe[2]
R. Manmatha[1]

[1]University of Massachusetts, Amherst
[2]Smith College

Motivation: Inaccessible Treasures

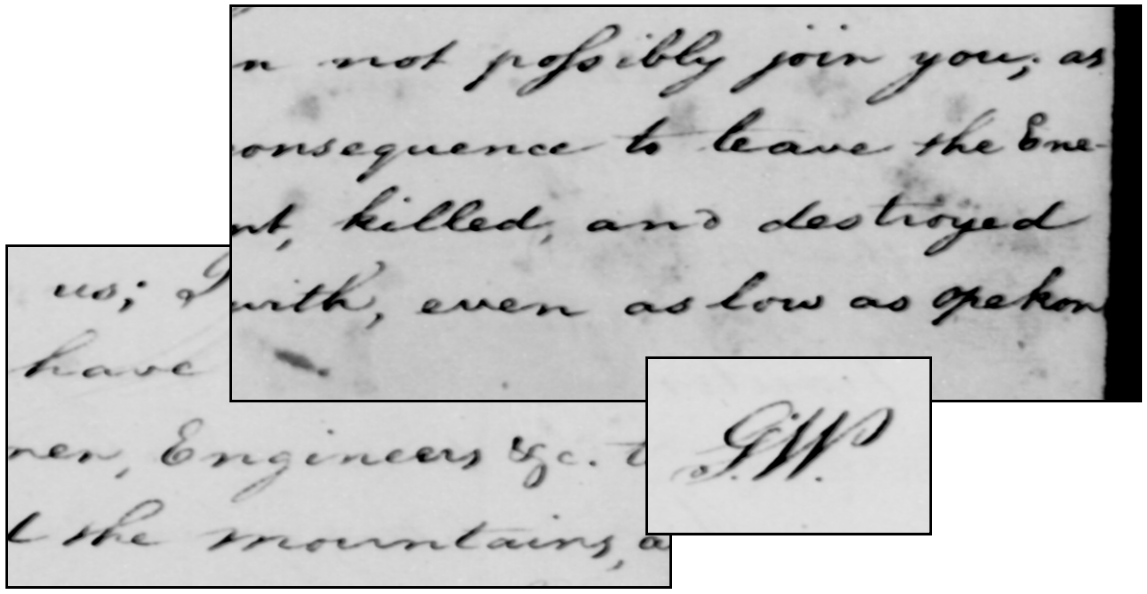
- Historical document collections
  - Scanned images available
  - Transcription often prohibitive ($$$)
  - Unprocessed format limits use
- Many such collections
  - Washington's letters: 140K pages
  - Isaac Newton's manuscripts
  - Scientific field notebooks
  - Antiquities

Goal: automated search/retrieval



Challenges of Historical Documents

- Offline handwriting OCR: success in constrained domains
  - Postal addresses, bank checks, etc.
- Historical documents are much harder
  - Few constraints
  - Fading & stains
  - Hyphenation
  - Misspellings
  - Ink bleed
  - Slant
  - Ornaments

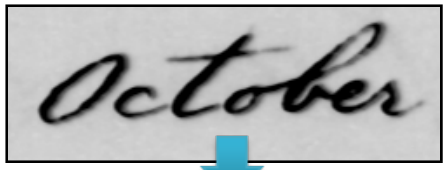


Excerpts from the GW20 collection

APPROACH

Word Recognition & Rare Words

- Most previous work with GW data employs **full-word** recognition.
- Zipf's Law: frequency of i^{th} most common word proportional to i^{-1}



'October'

$\Rightarrow$ Most words appear only rarely

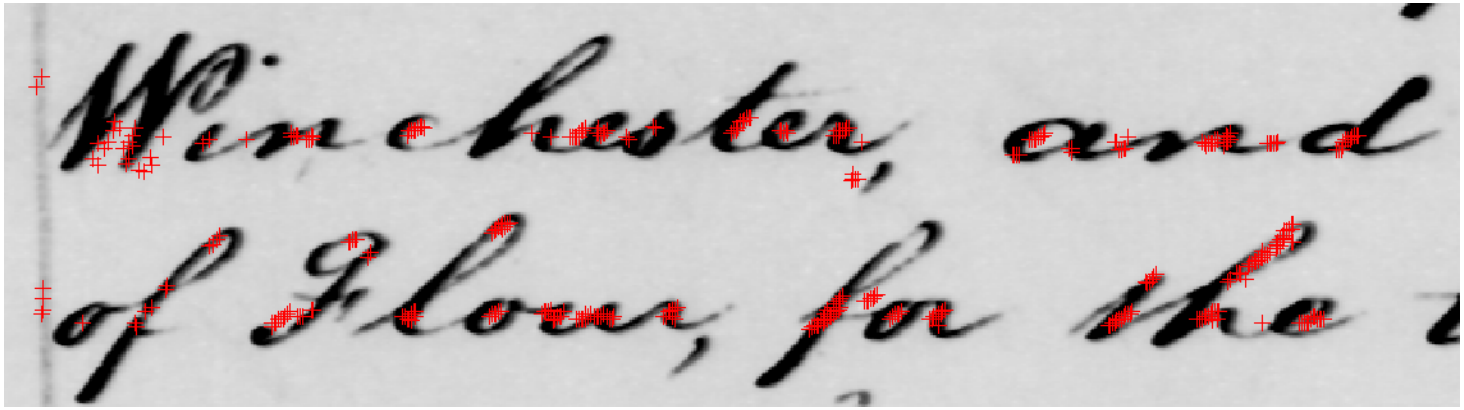
57% of vocabulary: <u>single</u> example

- Hard to learn from one example
- Even harder to learn from zero examples (OOV = out-of-vocabulary)
- <u>Rare words may be most significant</u>!

George K. Zipf

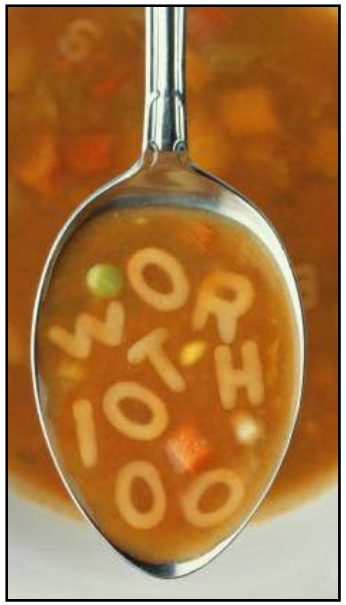
Character-Based Recognition: How?

- Character segmentation is hard & error-prone
- Easier to locate putative letters without segmentation
- Borrow techniques from object recognition



Alphabet Soup

- Letter detection sounds good, but how do we make whole words?
- Employed new inference model (or new twist on good old HMM)
- Remainder of talk:
  I. Letter Detection
  II. Inference Model
  III. Experimental Results

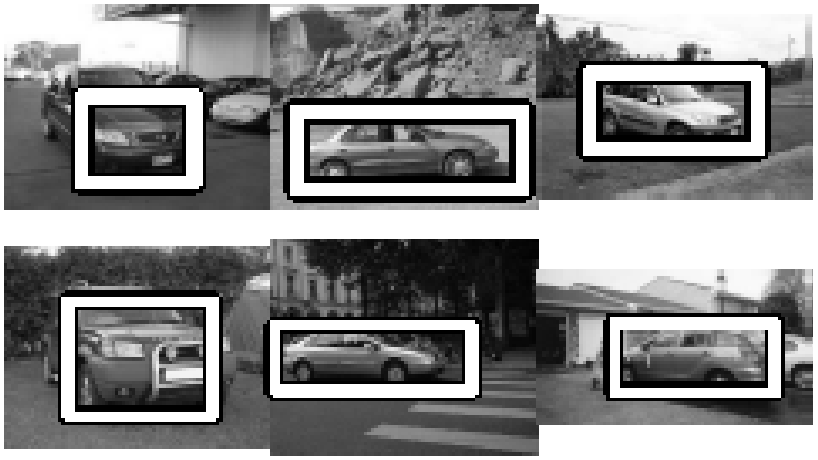


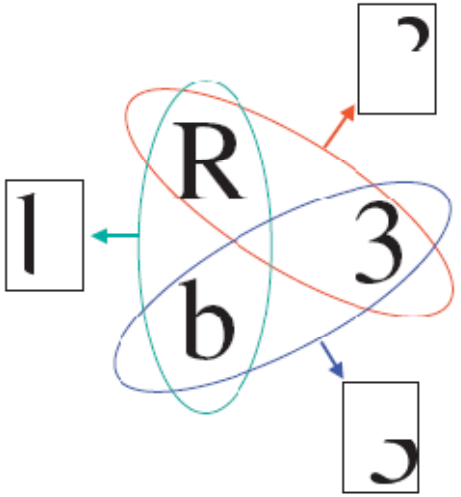
LETTER DETECTION

What are Latest Detection Results?

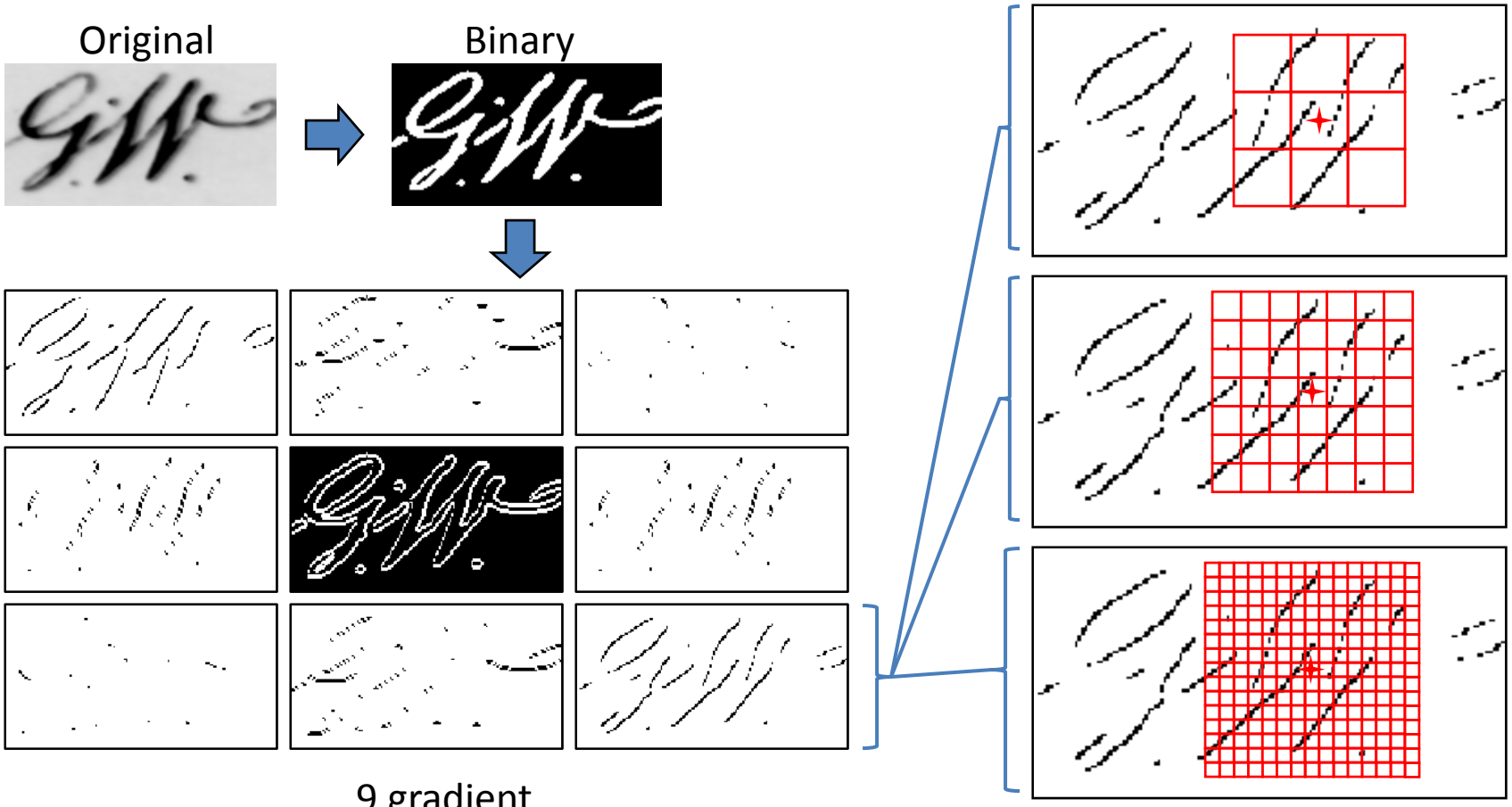
- Object detection
  - Use many *features*
  - Statistical methods pick indicative combinations
  - Torralba, Murphy & Freeman: *joint boosting*

Deng et. al., CVPR 2007



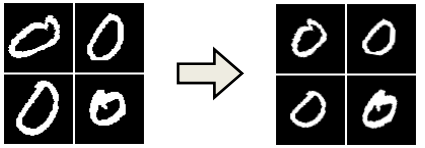


Histograms of Gradient Orientations (HoG)



Training a Letter Detector

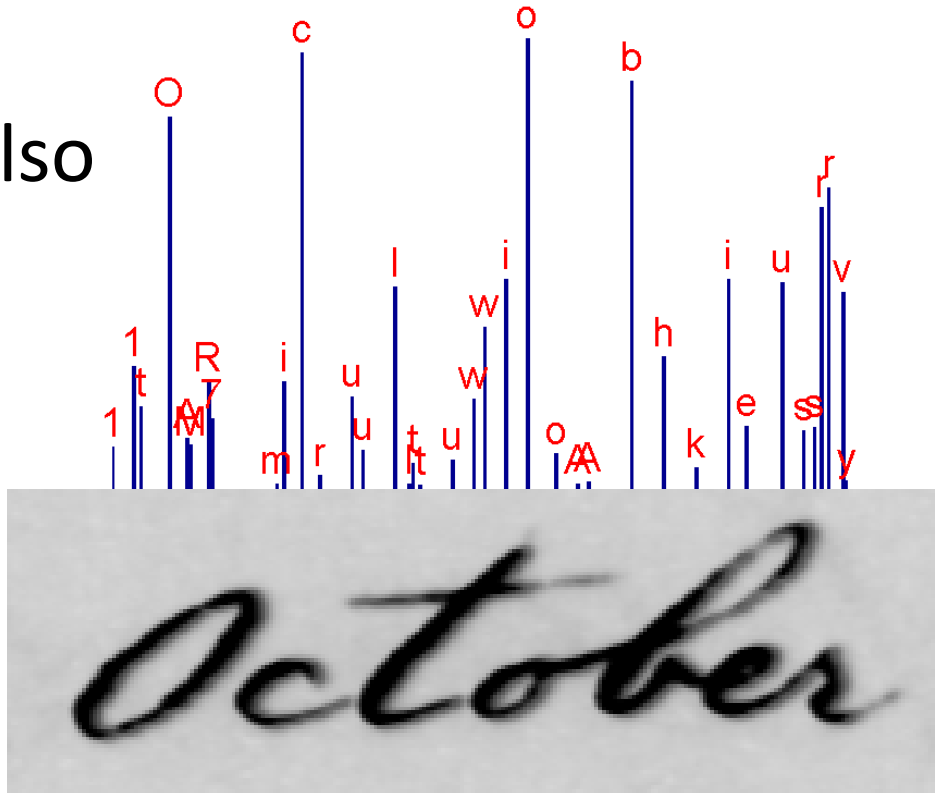
- Human identifies ~16 samples per character
- Samples are aligned



- Additional samples found automatically
- HoG feature vector created for each
- Joint boosting trains classifier on all characters
- Classifier looks at all points on midline of unknown word

Letter Detections

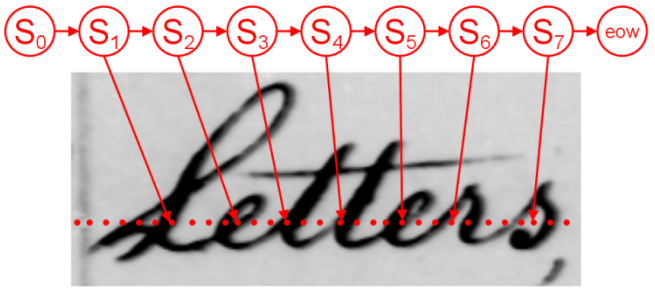
- Candidates include false positive detections
- Correct detections also
- Choice of many possible sequences
- Helpful hints:
  - Detection score
  - Letter sequence
  - Spatial separation



INFERENCE

Inference Model

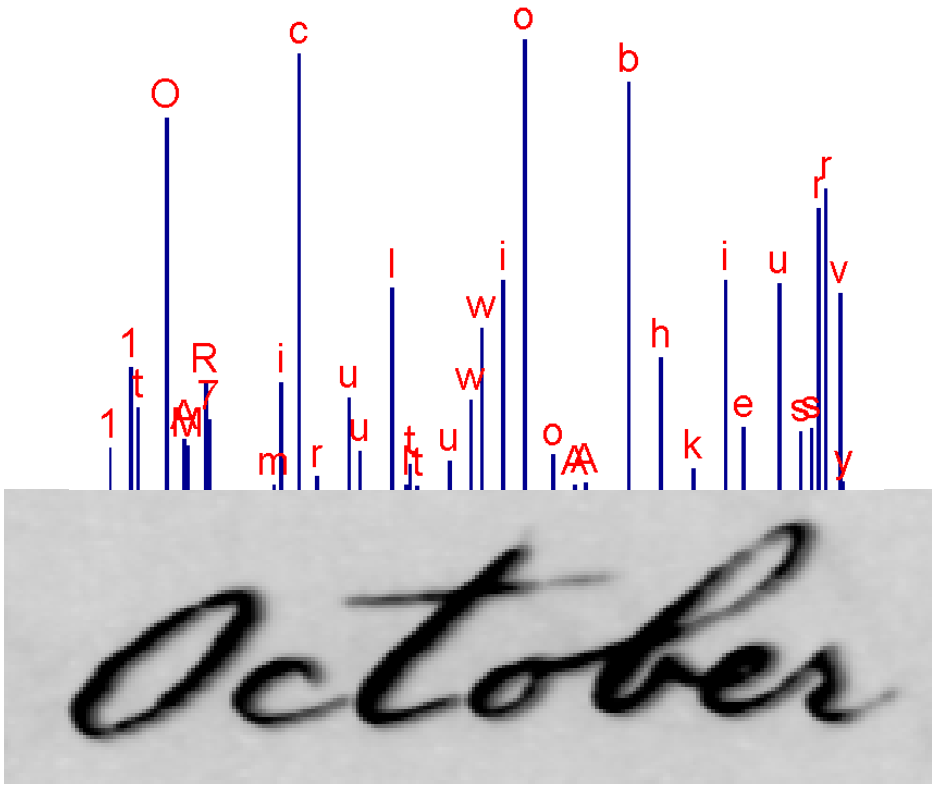
- State-per-slice or state-per-detection leads to complex HMM
- If number of letters in word known, can make small HMM with one state per letter
  - We don't know, so make multiple HMMs, one for each length
  - Try **all** lengths
  - Observations are detections



$$P(O,S) = \prod_{i=1}^{m} P(s_i|s_{i-1})P(o_i|s_i)$$

Generative Probabilities $P(o_i/s_i)$

- $P(o_i/s_i)$ taken as exponential of detection score (times some very small constant)
- More complex modeling didn't work very well



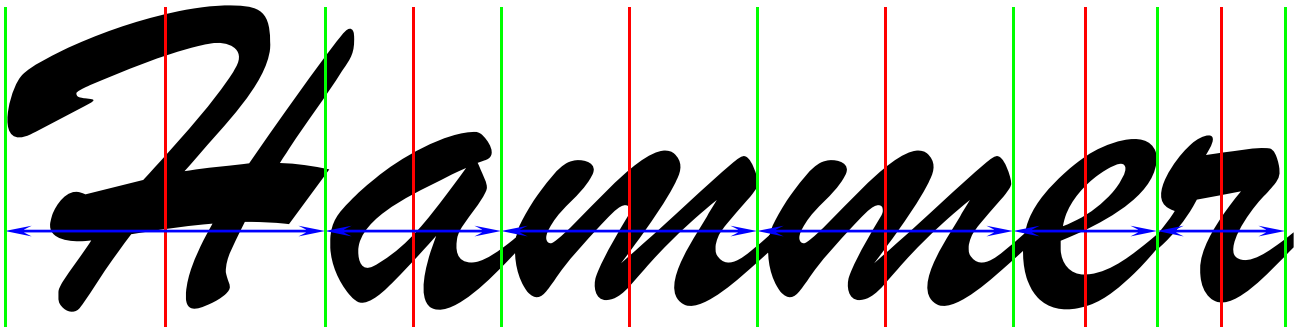
Transition Probabilities $P(s_i/s_{i-1})$

$P(s_i/s_{i-1})$ estimate has two components:

- Character transitions
  - Bigram or trigram
  - Estimated on training corpus using smoothing
- Spatial separation
  - Mean separation assumed dependent on characters at s_i and s_{i-1}
  - Variation assumed normal around mean

Character Separation Model

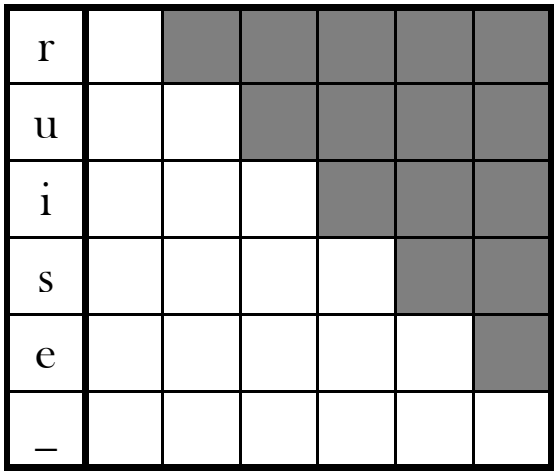
- Missing data problem for mean separations
- Model: $S_{ij} = \frac{1}{2}(w_i + w_j)$

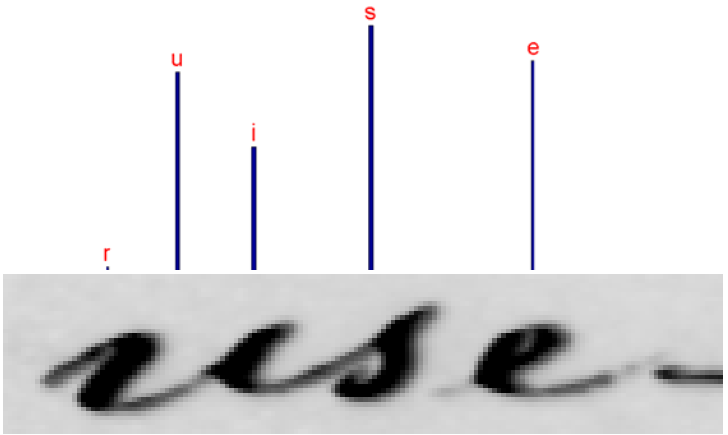


- Observed separations overconstrain w_i
  - Use least squares solution
- Assume normal variation; estimate variance

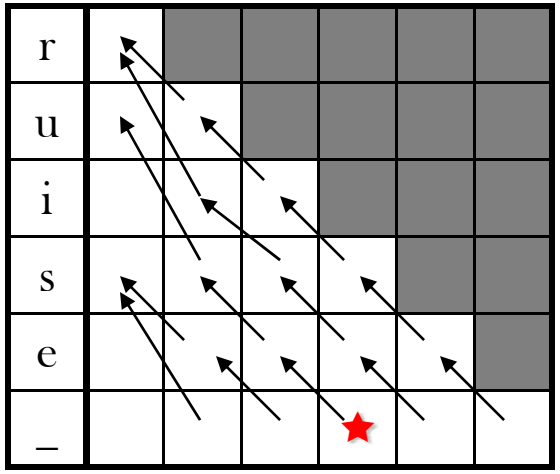
Dynamic Programming

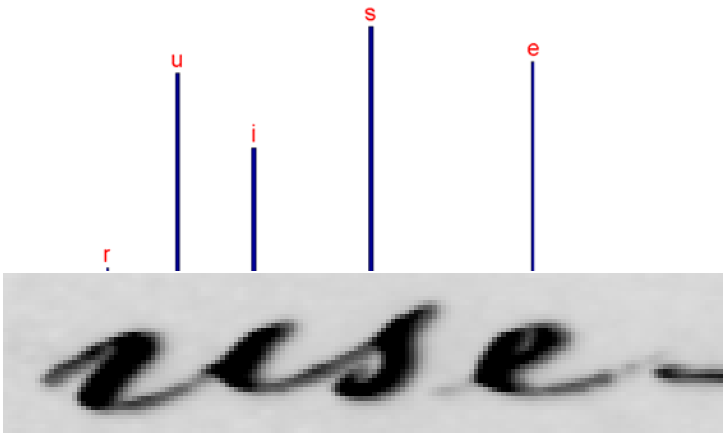
- Run Viterbi for HMM of each length
  - Reuse partial results for efficiency
- Dynamic programming computes likelihood of i^{th} detection in j^{th} word position ($i \geq j$)





Word Decoding

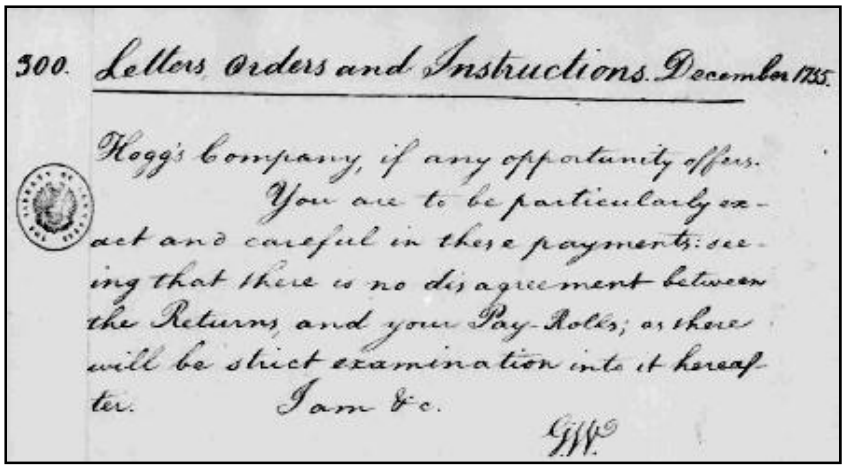




- Scores in bottom row correspond to HMM solutions for each length word
- Normalize by word length & choose highest
- Backpointers allow word decoding

EXPERIMENTS

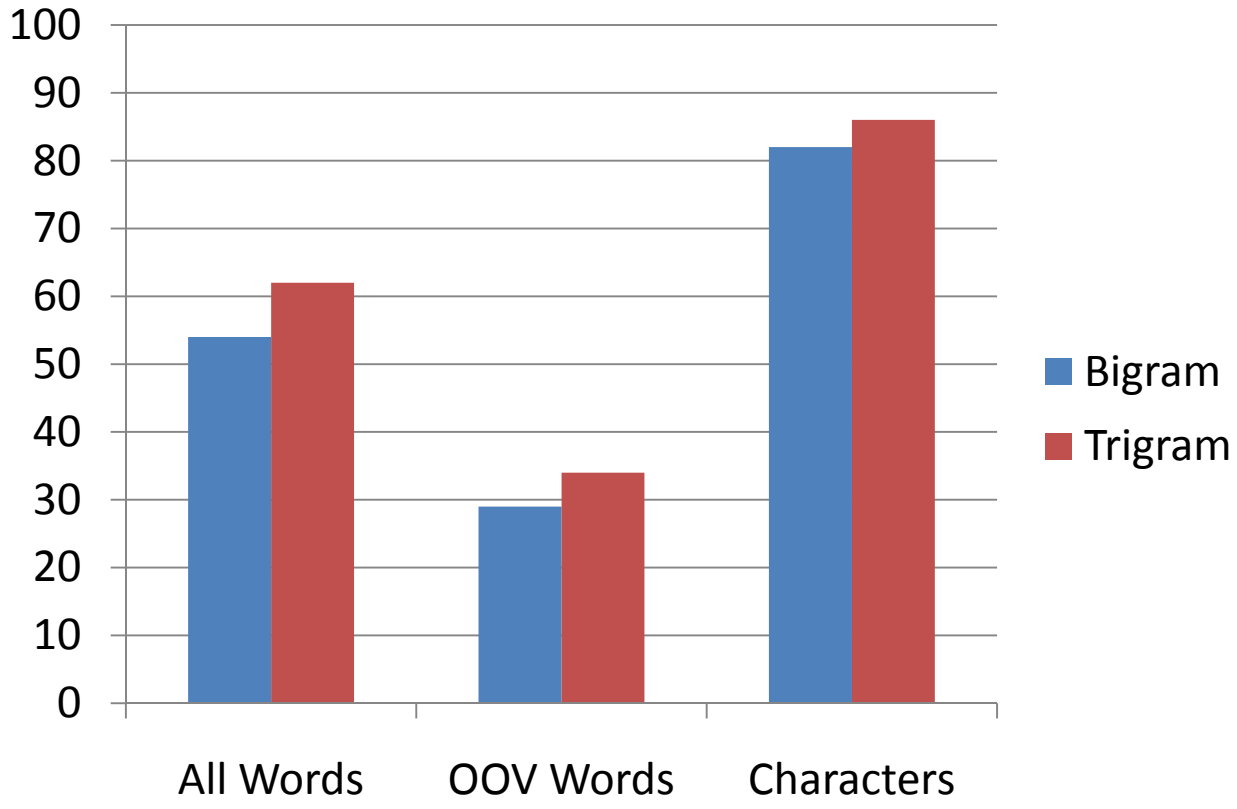
GW20 Corpus


300. Letters Orders and Instructions. December 1755.

Hogg's company, if any opportunity offers.
You are to be particularly ex-
act and careful in these payments: see-
ing that there is no disagreement between
the Returns, and your Pay-Rolls; as there
will be strict examination into it hereaf-
ter. I am &c.
GW.

- 20 pages of George Washington's letters
  - Written by multiple (30) secretaries
  - Available from UMass CIIR web site
- Cross-validation format
  - Train on 19 pages, test on 1
  - Rotate through all pages

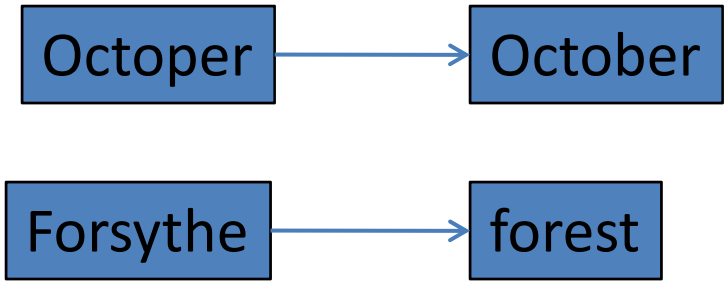
Accuracy: Base Results



Observation: Choice of word length could be improved
- Results improve ~10% when length is given

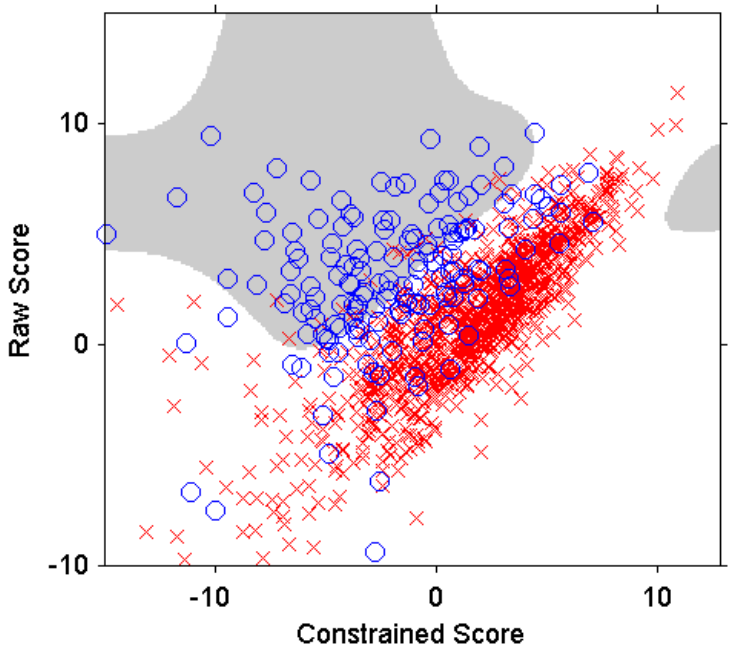
Using Lexicon Constraints

- Some bad predictions are not words: Octoper
- Restricted technique: constrain prediction to top-scoring word from training lexicon
  - OOV words not handled



Hybrid Prediction

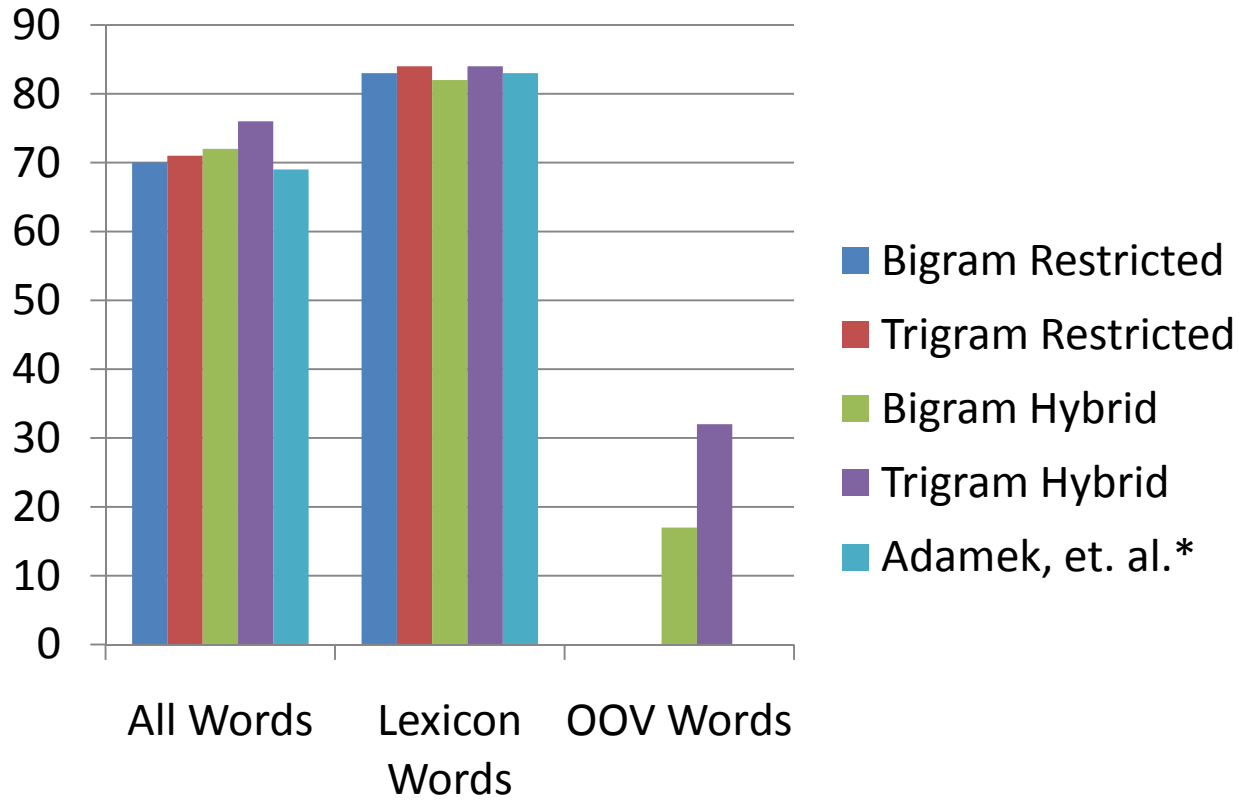
- Idea: Use relative scores to choose between original and restricted predictions



Octoper → October

Forsythe ✗ forest

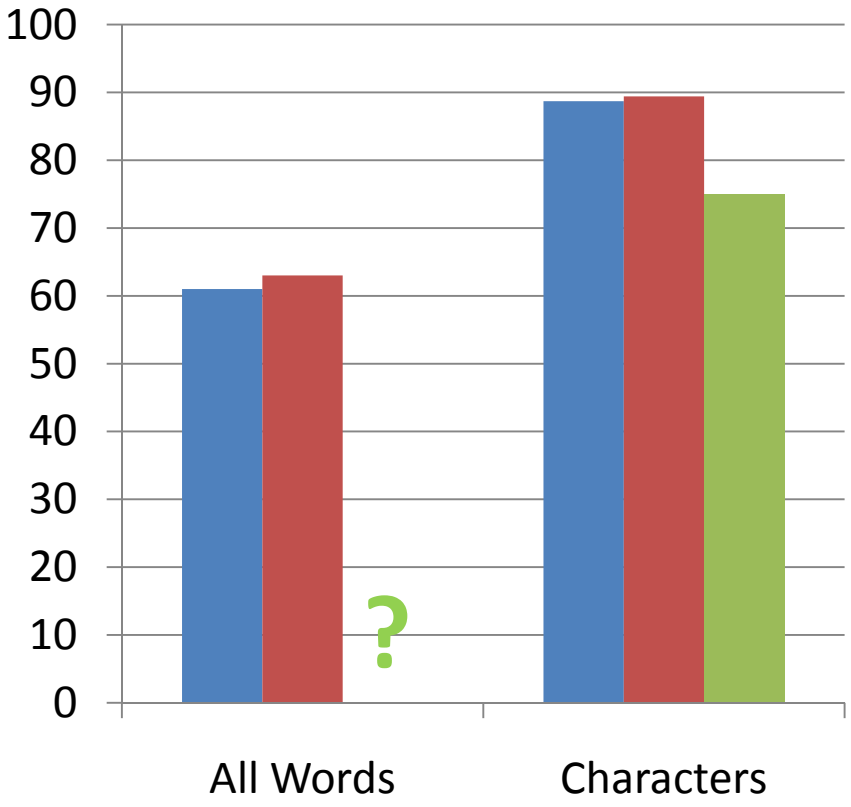
Results: Lexicon Restriction & Hybrid

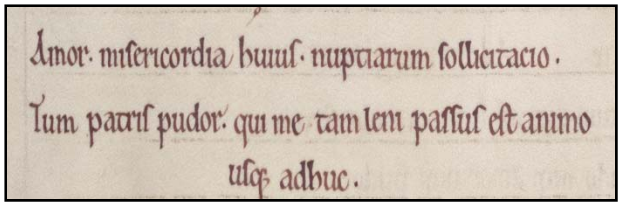


*Best prior result

Medieval Latin

- Results for Terence's *Comedies*





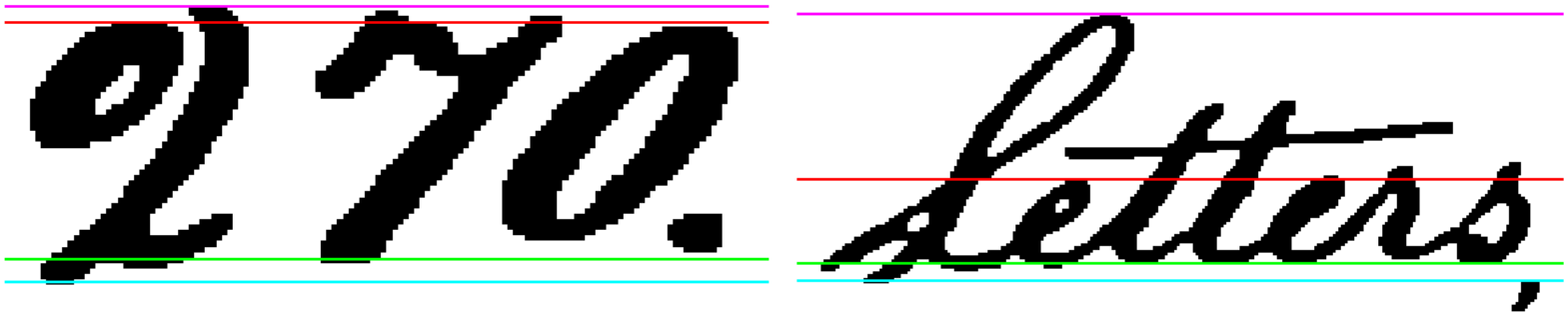
Final Remarks

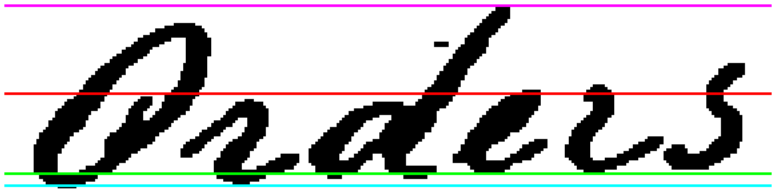
- All components of inference are important
  - Detection score
  - Character bigram/trigram
  - Physical separation
- Is HoG + joint boosting the best? Maybe…

  Any detector *may be used!*
- Try some alphabet soup for yourself!

And suddenly there it was, the perfect opening for Tommy's novel, lying at the bottom of his bowl of Alphabet Soup.

Finding Baselines





provide all other necessaries for the Expedition, which you know will be wanted

As there are several Contracts made by me to have Cattle delivered here &c. by the 1st. of next month, — I desire that for such as you receive upon that account, if you have money in your hands, you make immediate payment.

Given &c.

Winchester October 29th. 1755. G.W.

29th Winchester October 29th 1755.

Parole Williamsburgh.

One Subaltern, one Sergeant, one Corporal, one Drummer and twenty five private men, the Guard to-day — Captain Peachy is ordered to take upon him the command of the Recruits which arrived here under Lieutenant Hall and Ensign Price, who are also ordered to act under him, until further orders — Ensign Hedgeman, and the Recruits which arrived with him, are ordered to join Lieutenant King, and be under his command until further orders — Lieutenant Eustace, and the eight men with him are to join (as soon as they arrive at Fort Cumberland) the Company which Captain Waggoner commands at present; and the Party left with Sergeant Shaw, is to return to their respective Companies, so soon as they reach the Fort — The Commissary is to see that the Magazine is secured by fastening up the windows &c. better than they now are.

The Officers are to see that the men are clo-

compt. If that quantity can not be procured, send any lesser quantity that can be got.

I beg you will lose no time herein; by which you will oblige

Yours G.W.

April 21st. 1756.

To Ensign Hubbard.

Commanding at Enocks Fort.

Sir,

You are hereby desired if possible, to retreat with what men and provision you have to Edwards's; and to Escort what families have put themselves under your protection — But if you find this impracticable without a reinforcement, on your applying to Captain Harrison at Edwards's, a Detachment will be sent to assist you. You are not to fail in bringing off all the Stores you can.

I am &c.

G.W.

April 21st. 1756.

To Captain Harrison,

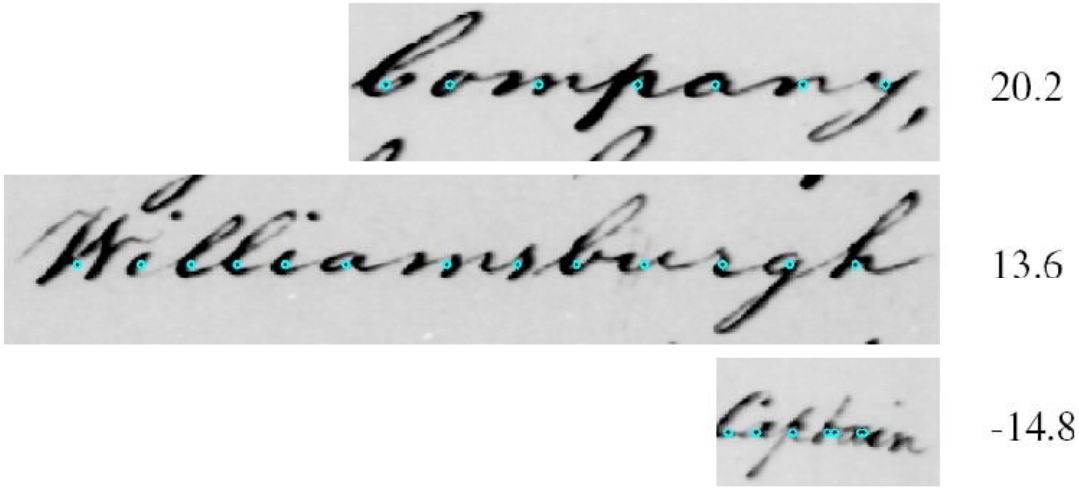
Commanding at Edwards's.

Sir,

It is out of my power at this juncture to supply you with any Provision. Therefore I would have you apply to Edwards, to whom I write. Acquaint him, that whatever he expends, he shall receive a reasonable satisfaction for: and hint to him, that without his compliance

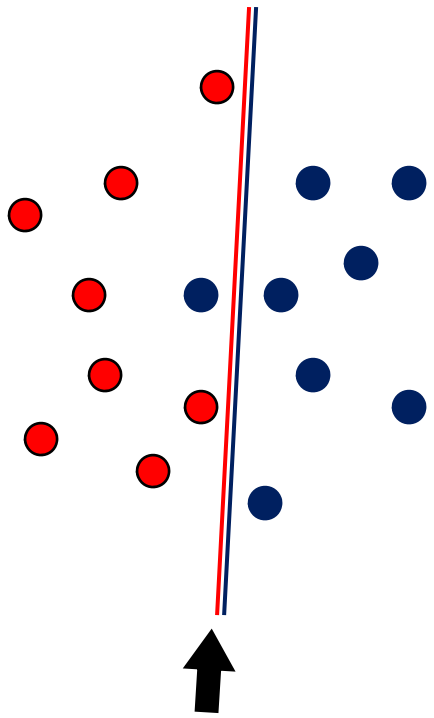
Locating Letters

- Easier to locate known letters than unknown
  - Only allow correct letter transitions
  - Use all possible detections
  - Gives position data for estimating separations



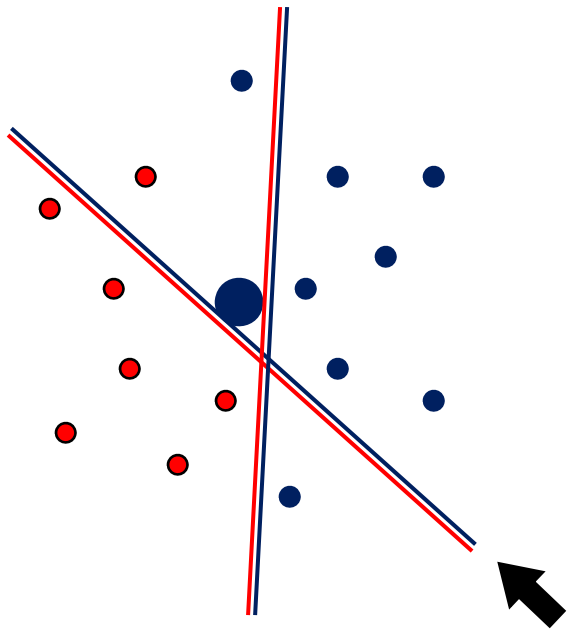
Example: Boosting

- Base rule must classify at least half of examples correctly.



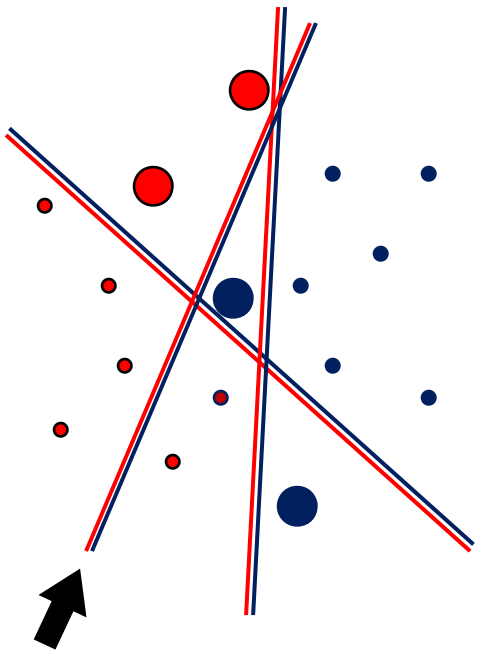
Example: Boosting

- Base rule must classify at least half of examples correctly.
- Reweight data before training new rule (focus on errors)



Example: Boosting

- Base rule must classify at least half of examples correctly.
- Reweight data before training new rule (focus on errors)
- Each new rule has different viewpoint



Example: Boosting

- Base rule must classify at least half of examples correctly.
- Reweight data before training new rule (focus on errors)
- Each new rule has different viewpoint
- Combined predictions are better than single classifier alone.