Statistical Basis to Determine Probabilities of Occurrence of Handwriting Characteristics

NIST Meeting 2013

Sargur Srihari, University at Buffalo, The State University of New York
Kirsten Singer, Dept of Veterans Affairs, Office of the Inspector General

Previous Research

- **1963** -- Frequency of Certain Characteristics in Handwriting, Pen Printing of 200 People," Livingston
- **1976** – A Statistical Examination of Selected Handwriting Characteristics, Muehlberger, et al
- **1990** -- "Uniqueness of Writing," Huber
- **1996** -- A Study of the Occurrence of Certain Handwriting Characteristics in a Random Population," Horton
- **1998** -- A Statistical Study of Some Differentiating Characteristics of the Handwritten Letters IT," Zlotnick
- **2013** -- "Frequency of Selected Hand Printing Characteristics Occurring within a National Population: The New International Version Bible Across America©" Bishop

Current Research

- “Statistical Examination of Handwriting Characteristics using Automated Tools,” Singer/ Srihari, SUNY

- “Development of Individual Handwriting Characteristics in ~1800 Students: Statistical Analysis and Likelihood Ratios that Emerge over an Extended Period of Time”, Lisa Hanson, Minnesota Bureau of Criminal Apprehension, Dr. Srihari, SUNY

- “Frequency Occurrence of Handwriting and Hand-Printing Characteristics” Vastrick and Whitcomb with University of Central Florida

Previous Research

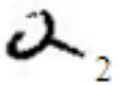
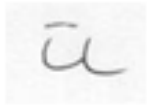
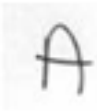
1. Height relationship of the “t” to the “h”
    a. t shorter than h (78%)
    b. t even with h (1.5%)
    c. t taller than h (5.5%)
    d. No set pattern (15%)

2. Shape of loop of “h”
    a. Retraced (27.5%)
    b. curved right side and straight left side (32%)
    c. curved left side and straight right side (2.5%)
    d. Both sides curved (17%)
    e. No set pattern (21%)

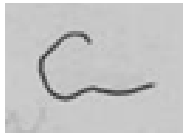
“and”

Rank	Trigram
1	the
2	and
3	tha
4	ent
5	ing

Singer/Moran

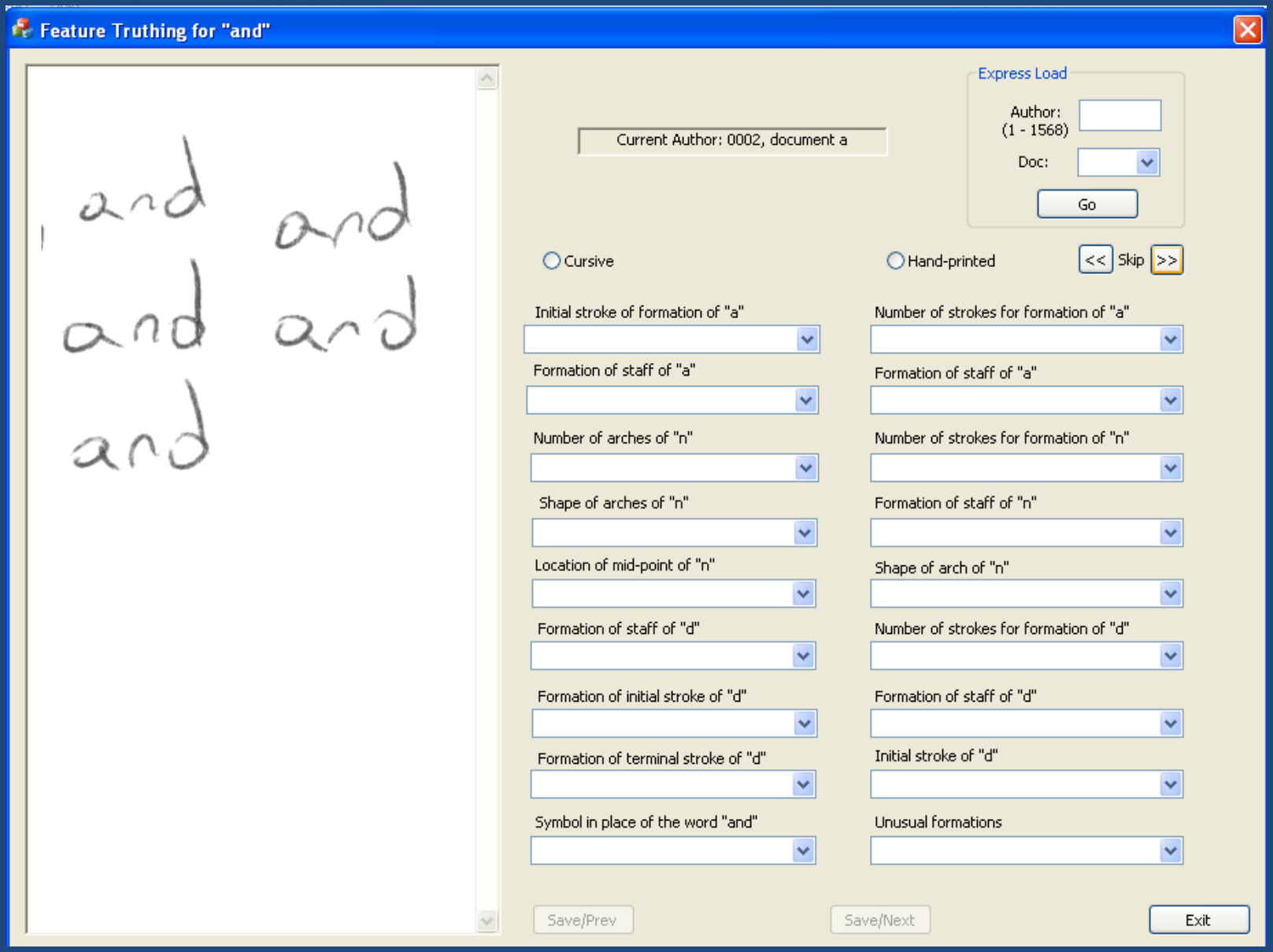
1. Number of strokes for formation of "a": (a)  (b)  2 (c)  (d) 

(a) one continuous														
(b) two strokes														
(c) three strokes														
(d) uppercase														
(e) no fixed pattern														

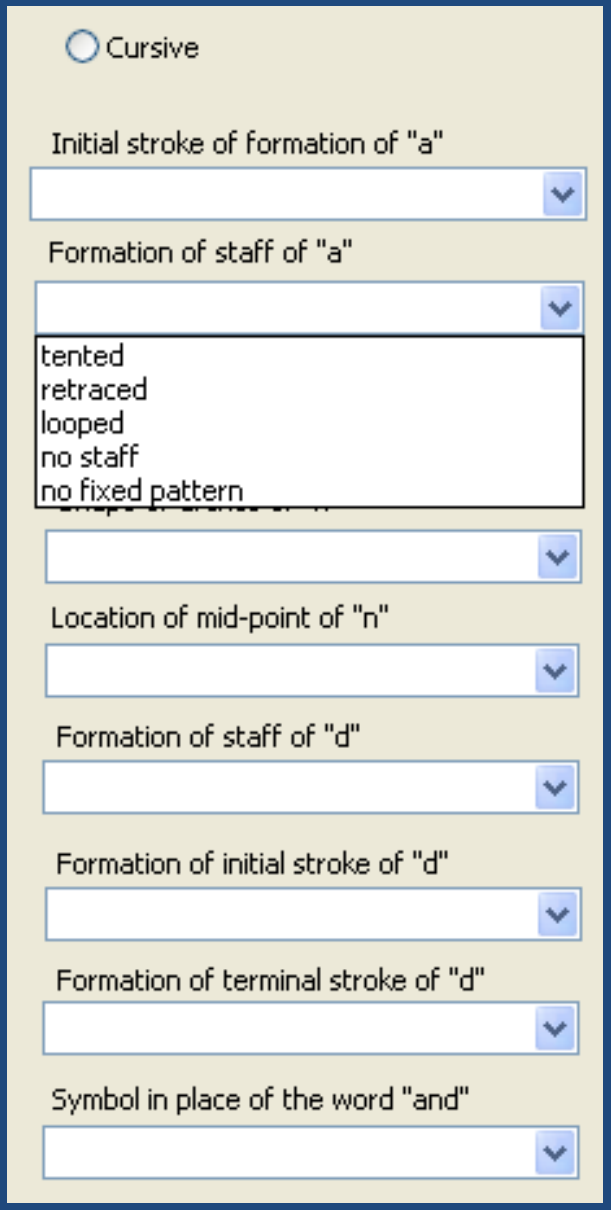
2. Formation of staff of "a": (a)  6 (b)  8 (c)  87 (d) 

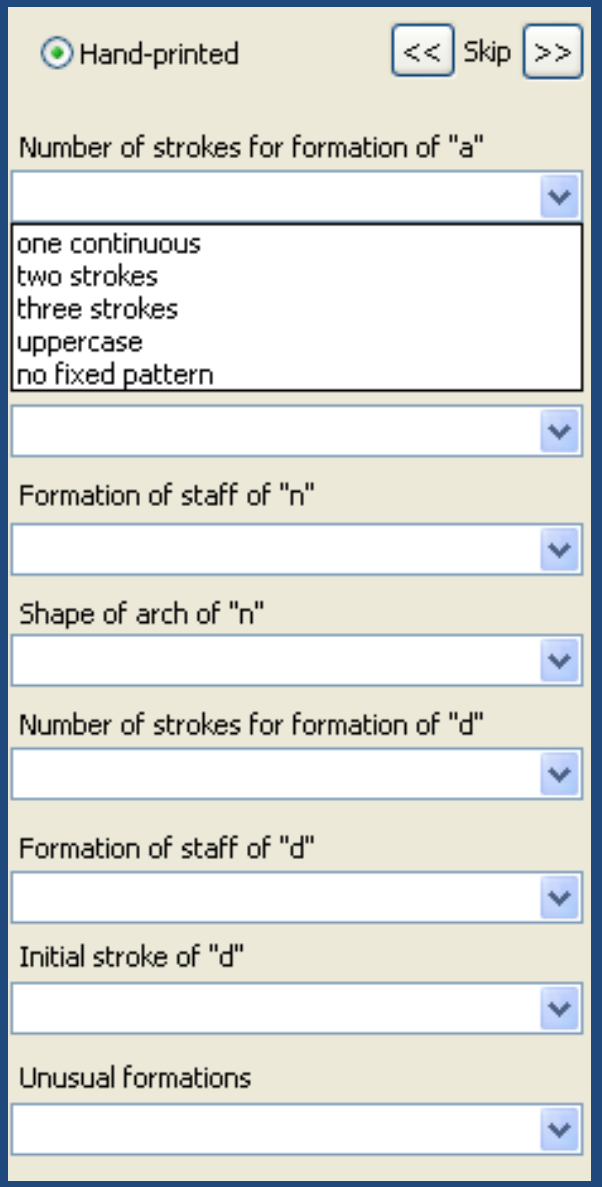
(a) tented												
(b) retraced												
(c) looped												
(d) no staff												
(e) no fixed pattern												

Truthing Tool

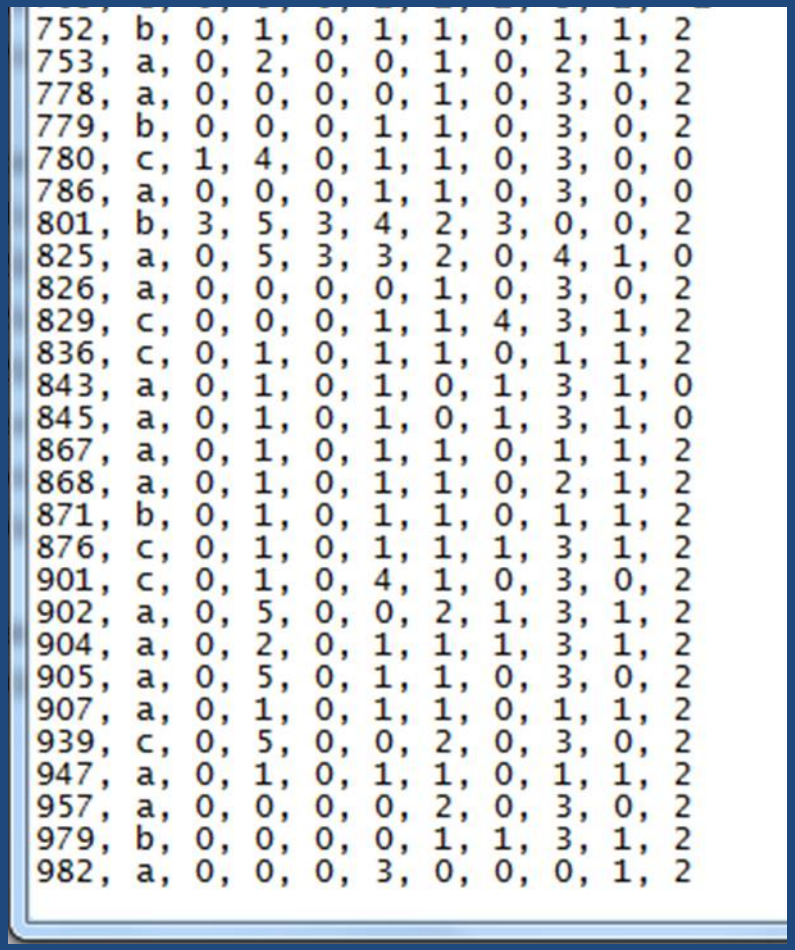


Truthing Tool



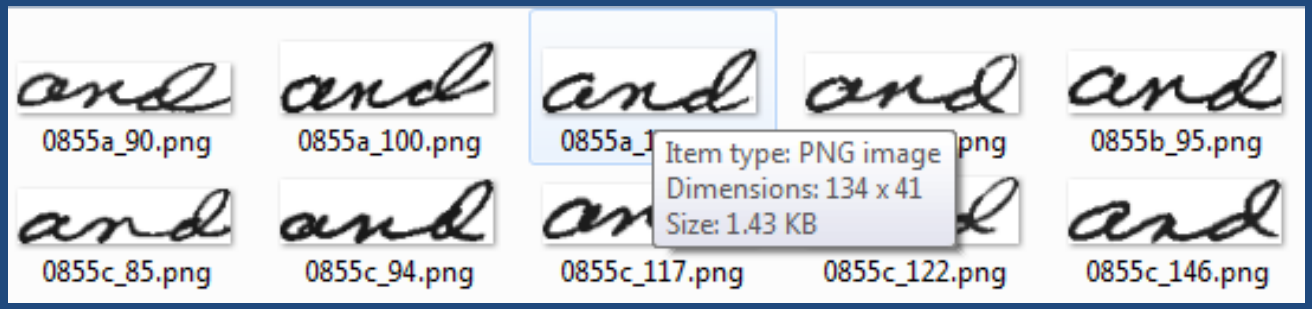


Truthing Tool



```
752, b, 0, 1, 0, 1, 1, 0, 1, 1, 2
753, a, 0, 2, 0, 0, 1, 0, 2, 1, 2
778, a, 0, 0, 0, 0, 1, 0, 3, 0, 2
779, b, 0, 0, 0, 1, 1, 0, 3, 0, 2
780, c, 1, 4, 0, 1, 1, 0, 3, 0, 0
786, a, 0, 0, 0, 1, 1, 0, 3, 0, 0
801, b, 3, 5, 3, 4, 2, 3, 0, 0, 2
825, a, 0, 5, 3, 3, 2, 0, 4, 1, 0
826, a, 0, 0, 0, 0, 1, 0, 3, 0, 2
829, c, 0, 0, 0, 1, 1, 4, 3, 1, 2
836, c, 0, 1, 0, 1, 1, 0, 1, 1, 2
843, a, 0, 1, 0, 1, 0, 1, 3, 1, 0
845, a, 0, 1, 0, 1, 0, 1, 3, 1, 0
867, a, 0, 1, 0, 1, 1, 0, 1, 1, 2
868, a, 0, 1, 0, 1, 1, 0, 2, 1, 2
871, b, 0, 1, 0, 1, 1, 0, 1, 1, 2
876, c, 0, 1, 0, 1, 1, 1, 3, 1, 2
901, c, 0, 1, 0, 4, 1, 0, 3, 0, 2
902, a, 0, 5, 0, 0, 2, 1, 3, 1, 2
904, a, 0, 2, 0, 1, 1, 1, 3, 1, 2
905, a, 0, 5, 0, 1, 1, 0, 3, 0, 2
907, a, 0, 1, 0, 1, 1, 0, 1, 1, 2
939, c, 0, 5, 0, 0, 2, 0, 3, 0, 2
947, a, 0, 1, 0, 1, 1, 0, 1, 1, 2
957, a, 0, 0, 0, 0, 2, 0, 3, 0, 2
979, b, 0, 0, 0, 0, 1, 1, 3, 1, 2
982, a, 0, 0, 0, 3, 0, 0, 0, 1, 2
```

Truthing Tool



Writer ID

855 1 2 1 0 2 2 1 0 0

- 1: Initial Stroke left of staff
- 2: Looped staff
- 1: 2 Arches for “n”
- 0: Pointed arches
- 2: Arch midpoint at baseline
- 2: Looped staff of “d”
- 1: initial stroke of “d” underhand
- 0: Terminal stroke curved up
- 0: No symbol for “and”

Probabilistic Analysis




121022322

Bayesian Network Joint Probability

9.48e-005

Independent Joint Probability

1.39e-004

Joint Probability Calculation

- Calculation of probability of a given combination of characteristics is complex
- How much data is needed?
  - If we don't assume that the nine characteristics are independent, we will need to determine over a million probabilities
    - 100 million to billion samples needed
- How much time for the computation?
  - NP-hard

What if we assume independence?

True Joint Probabilities: Prob (height,weight)

P(a,b)	b^0 (heavy)	b^1 (light)	P(a) (height)
a^0 (tall)	0.6	0.05	0.65
a^1 (short)	0.05	0.3	0.35
P(b) (weight)	0.65	0.35	

Prob(tall, light) < Prob(short,light) Given that person is light, six times likely to be short

Assuming Independence

P(a,b)	b^0 (heavy)	b^1 (light)	P(a) (height)
a^0 (tall)	0.42	0.23	0.65
a^1 (short)	0.23	0.12	0.35
P(b) (weight)	0.65	0.35	

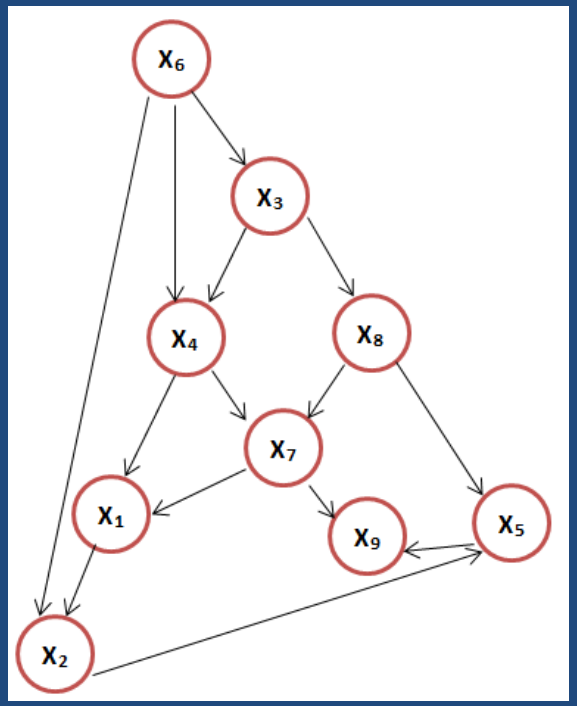
P(tall,light) > P(short,light) Given that person is light, twice likely to be tall

Compromise Solution: PGMs

- Revolution in big data analysis
- Led by statistical machine learning and probabilistic graphical models
- Exploit as to what independencies exist rather than assume everything is independent
- PGMs are directed (Bayesian Networks) or undirected (Markov networks)

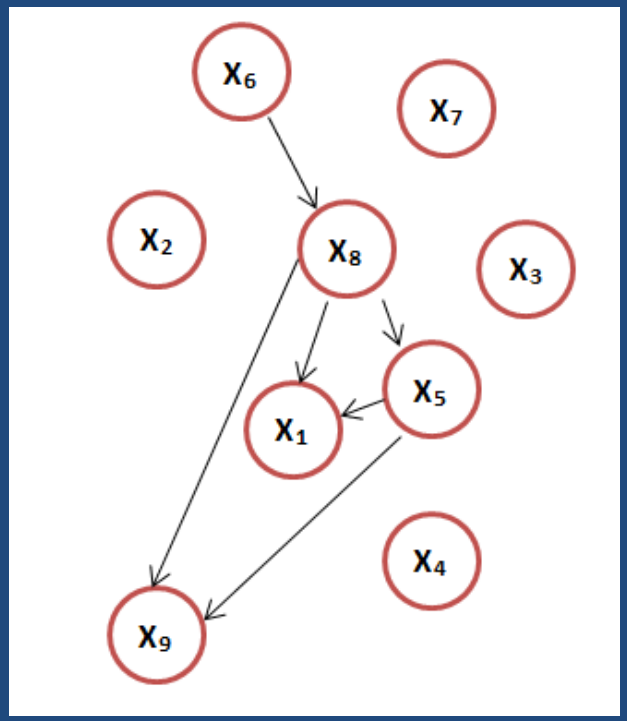
Bayesian Networks for *and*

Cursive



99 parameters

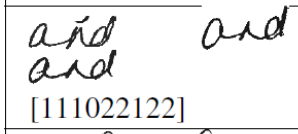
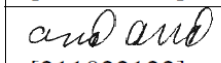
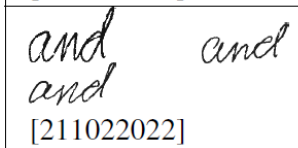
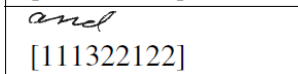
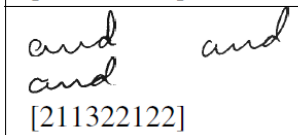
Handprint



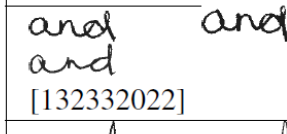
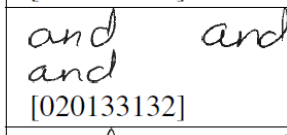
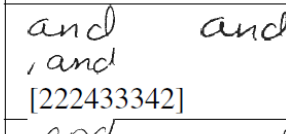
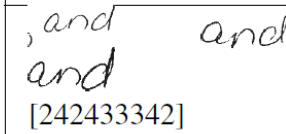
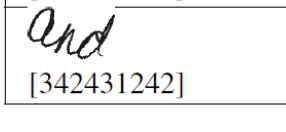
77 parameters

Common and Rare *and*

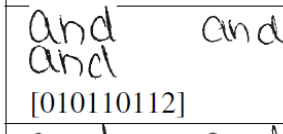
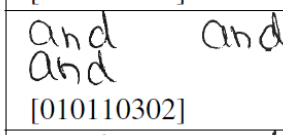
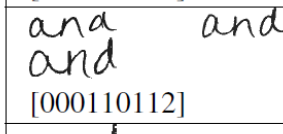
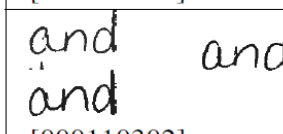
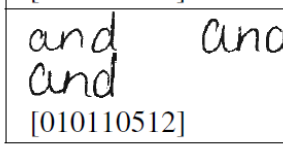
(a) Cursive-Common

Samples with Characteristics	Probability
 [111022122]	5.46×10^{-3}
 [211022122]	5.39×10^{-3}
 [211022022]	4.86×10^{-3}
 [111322122]	4.52×10^{-3}
 [211322122]	4.46×10^{-3}

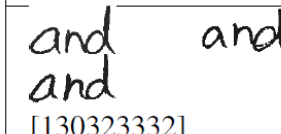
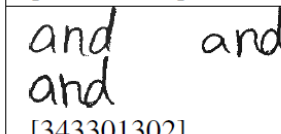
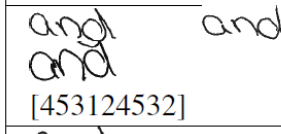
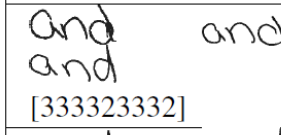
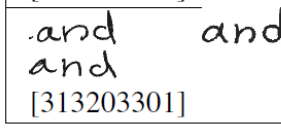
(b) Cursive-Rare

Samples with Characteristics	Probability
 [132332022]	4.15×10^{-8}
 [020133132]	4.09×10^{-8}
 [222433342]	8.64×10^{-9}
 [242433342]	7.50×10^{-9}
 [342431242]	5.75×10^{-9}

(c) Handprint-Common

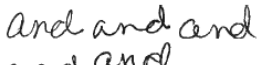
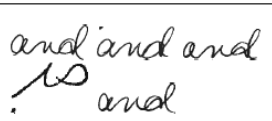
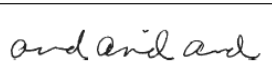
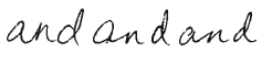
Samples with Characteristics	Probability
 [010110112]	1.51×10^{-2}
 [010110302]	1.44×10^{-2}
 [000110112]	1.21×10^{-2}
 [000110302]	1.15×10^{-2}
 [010110512]	7.42×10^{-3}

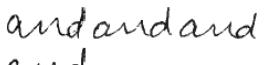
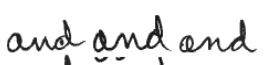
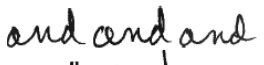
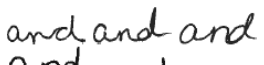
(d) Handprint-Rare

Samples with Characteristics	Probability
 [130323332]	5.90×10^{-9}
 [343301302]	4.66×10^{-9}
 [453124532]	3.75×10^{-9}
 [333323332]	7.81×10^{-10}
 [313203301]	7.23×10^{-10}

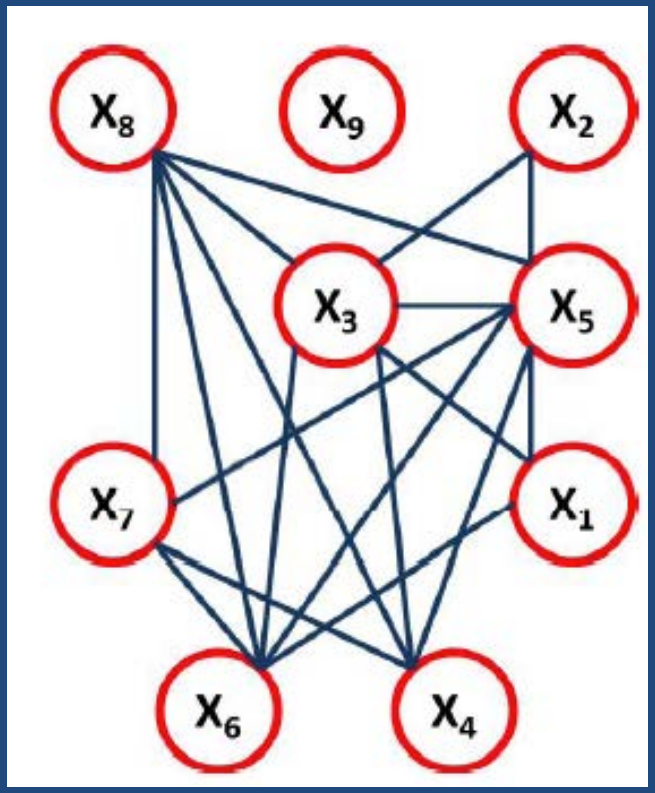
Probabilities available online

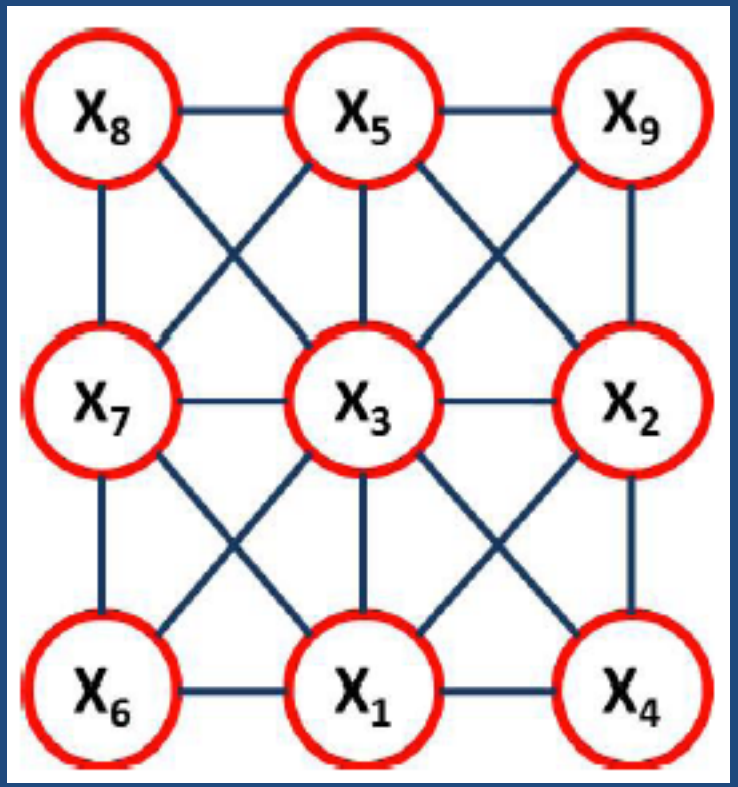
Cursive data: http://www.cedar.buffalo.edu/~srihari/HW-Stats/cursive-and
Handprint data: http:/www.cedar.buffalo.edu/~srihari/HW-Stats/handprint-and

#	Samples	ID	Characteristics	BN Joint Prob	Indep Joint Prob
986		0387b	220101021	4.70e-007	4.57e-007
987		0522c	312402032	4.68e-007	7.20e-006
988		1123a	212101121	4.16e-007	6.34e-007
989		1198a	302422042	3.76e-007	9.24e-006
990		1198b	101320221	3.65e-007	9.33e-007
991		1098a	010101111	3.38e-007	1.17e-006

#	Samples	ID	Characteristics	BN Joint Prob	Indep Joint Prob
1000		1271c	012422222	2.50e-007	3.53e-006
1001		0354c	100101322	2.49e-007	5.85e-007
1002		1091a	312422342	2.22e-007	1.12e-005
1003		0556b	020102111	2.07e-007	5.45e-007
1004		0556c	010103101	2.05e-007	3.66e-007
1005		0387a	010201222	1.87e-007	3.50e-006

Markov networks for *and*



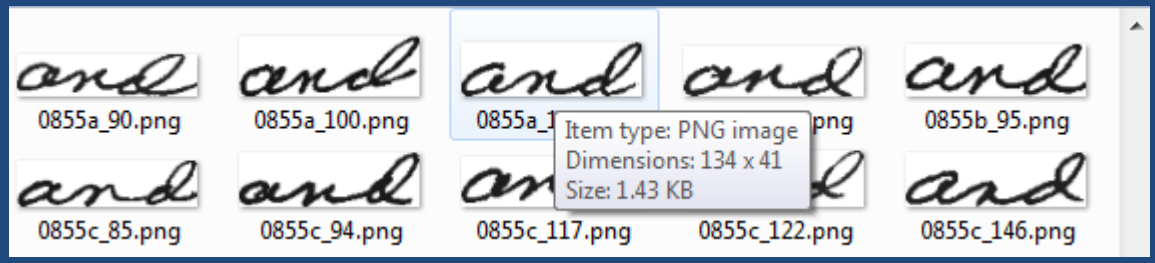


Conclusions

- FDEs defined a set of characteristics for a common word *and*
- Developed a truthing interface
- FDEs entered data using interface
- Developed learning algorithms to create statistical models
- Models used to infer probability of characteristics

Future Research

- Continue to mine existing data for information, add more individualizing characteristics
- Continue to research “th” combination
- Study the same characteristics with a more homogenous population (e.g. Durina research, twins)



855, a, 1, 2, 1, 0, 2, 2, 1, 0, -1