

Explainable Artificial Intelligence: Probabilistic Approach

Sargur N. Srihari
srihari@buffalo.edu

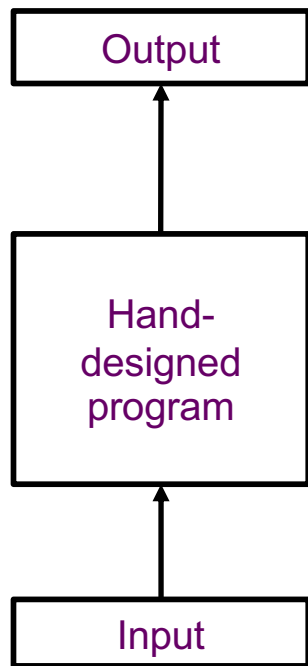
Topics

- Three waves of AI
- Explainable AI
- Forensic comparison
- Most Probable Explanation with PGMs
- Deep Learning + PGM for Forensic Comparison

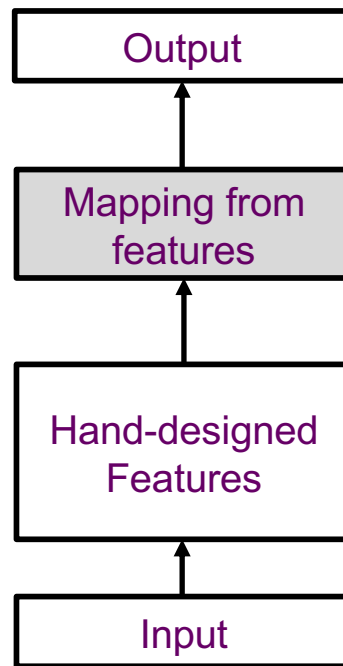
Waves of AI Models

FIRST WAVE

Rule-based System

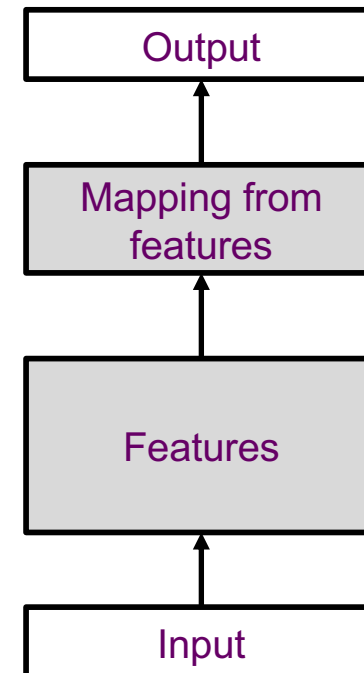


Classic Machine learning

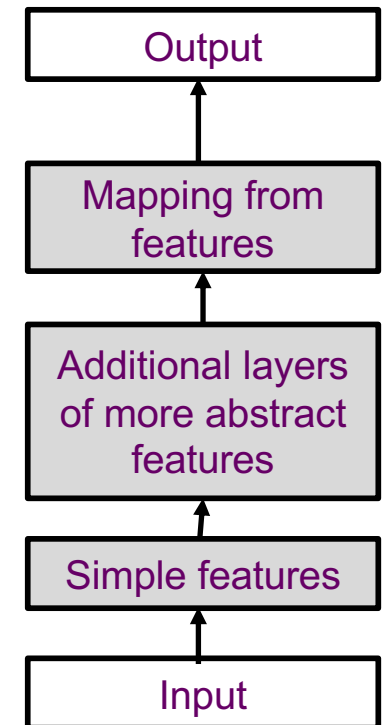


SECOND WAVE

Representation Learning



Deep Learning



Shaded boxes indicate components that can learn from data

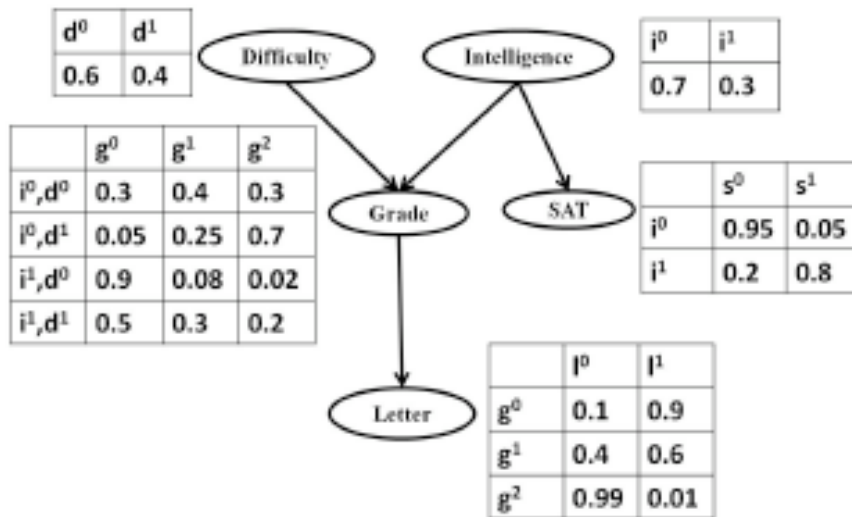
Focus was on discriminative models: supervised learning

Third Wave: Probabilistic AI

- Focus on generative models
 - Construct distributions explicitly that are queried
- Algorithms/Architectures for:
 - Representation
 - Bayesian nets, Markov Random Fields
 - Inference
 - Exact, Approximate, Monte Carlo, Variational
 - Learning

Probabilistic Graphical Models

Bayesian Networks



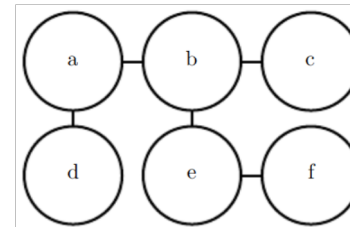
- CPDs: $p(x_i \mid pa(x_i))$
- Joint Distribution $\mathbf{x} = \{x_1, \dots, x_n\}$

$$p(\mathbf{x}) = \prod_{i=1}^N p(x_i \mid pa(x_i))$$

$$P(D, I, G, S, L) =$$

$$P(D)P(I)P(G \mid D, I)P(S \mid I)P(L \mid G)$$

Markov Networks



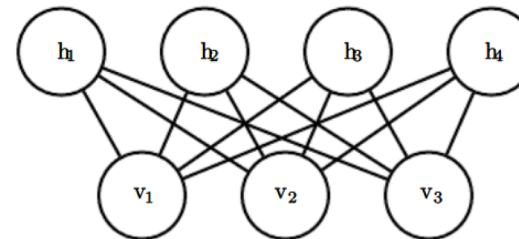
$$p(\mathbf{x}) = \frac{1}{Z} \hat{p}(\mathbf{x}) \quad \tilde{p}(\mathbf{x}) = \prod_{C \in G} \phi(C)$$

$$Z = \int \tilde{p}(\mathbf{x}) d\mathbf{x}$$

$$p(a, b, c, d, e, f) = \frac{1}{Z} \phi_{a,b}(a, b) \phi_{b,c}(b, c) \phi_{a,d}(a, d) \phi_{b,e}(b, e) \phi_{e,f}(e, f)$$

Energy model $E(a, b, c, d, e, f) =$
 $E_{a,b}(a, b) + E_{b,c}(b, c) + E_{a,d}(a, d) + E_{b,e}(b, e) + E_{e,f}(e, f)$
 $\phi_{a,b}(a, b) = \exp(-E(a, b))$

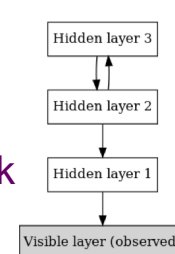
Restricted Boltzmann machine



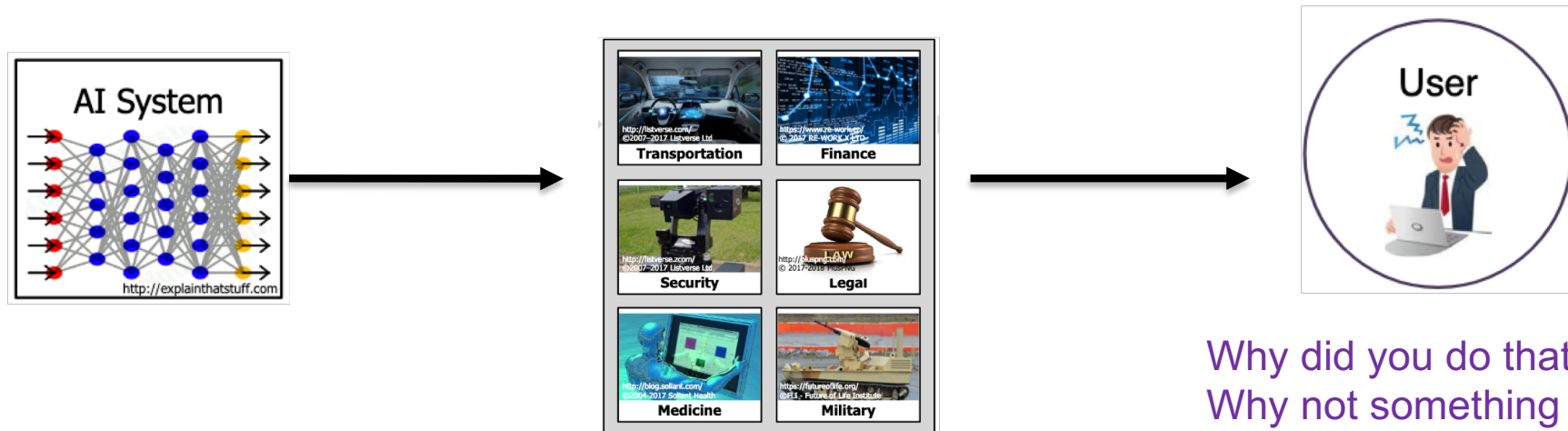
$$E(\mathbf{v}, \mathbf{h}) = -\mathbf{b}^T \mathbf{v} - \mathbf{c}^T \mathbf{h} - \mathbf{v}^T \mathbf{W} \mathbf{h}$$

$$p(\mathbf{h} \mid \mathbf{v}) = \prod_i p(h_i \mid \mathbf{v}) \quad \text{and} \quad p(\mathbf{v} \mid \mathbf{h}) = \prod_i p(v_i \mid \mathbf{h})$$

Deep Belief Network



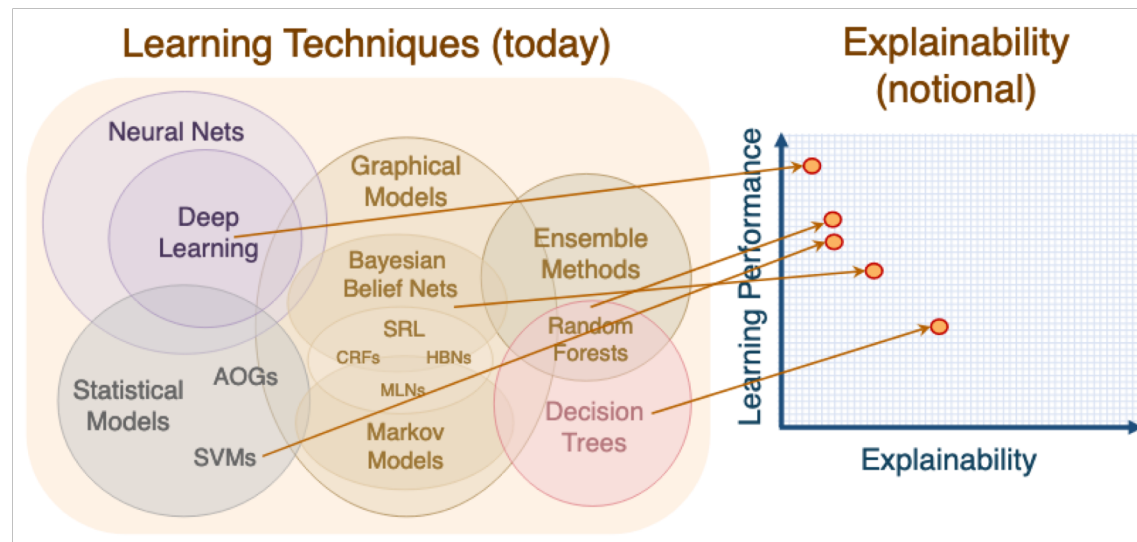
Explainable AI (DARPA)



Anecdote: Medical AI

Decisions can be worse with AI
e.g., Patient discharge to a nursing home

Why did you do that?
Why not something else?
When do you succeed?
When do you fail?
When can I trust you?
How do I correct an error?



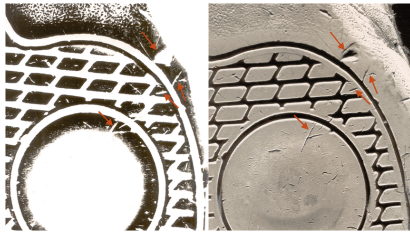
AI Models and XAI

- First wave (Knowledge-based systems such as decision trees were naturally suited for XAI)
- Second wave methods were focused on discrimination
- Third wave is more suited for explainability
 - Inference in PGMs has methods for most probable explanation (MPE)
 - Also called Maximum *A posteriori* Probability assignment

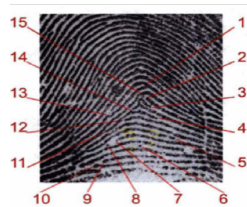
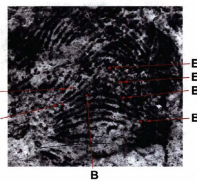
Forensic Comparison

- Impression Evidence:
materials with characteristics of
impressed objects

- Footwear impressions



- latent fingerprints



- Handwriting samples



Forensic Handwriting Examination



Handwriting identification is based on three basic principles –

1. no one writer writes the same way every time (basis of variation),
2. no two writers share the same combination of handwriting features (when given a sufficient amount of comparable writing for comparison)
3. a writer cannot write above his/her current skill level.

Courtesy of Kirsten Singer, Forensic Document Examiner

Explanation in Forensic Comparison



This chart represents some of the handwriting features I used to reach my conclusion, and is representative of what I examined but is not meant to replicate my entire examination. There are several handwriting similarities noted using the green arrows/bar, and a dissimilarity noted using the red arrows.

Similarities include the pointed first arch of the letter “n”, the counterclockwise motion of the formation of the front portion of the “d” and the way the stroke connects from the “n” into the middle of the “d”, and the relative height of the staff of the “d” compared to the other letters.

A **dissimilarity** is noted in the connecting stroke from the “a” to the “n”.

Is it below the baseline and has a sharp angle in the questioned “and”, and at or above the baseline and rounded in the known “and”.

Conclusion would likely be “indications one writer wrote both words”.

A more definitive conclusion could not be reached because of the limited amount of writing for comparison, and the dissimilarity noted.

Expert Features

Feature Vector

imagename	0968c_num1.png
pen_pressure	2
letter_spacing	2
size	2
dimension	1
is_lowercase	2
is_continuous	2
slantness	3
tilt	2
entry_stroke_a	1
staff_of_a	2
formation_n	2
staff_of_d	3
exit_stroke_d	2
word_formation	2
constancy	1

Number of Features: 15

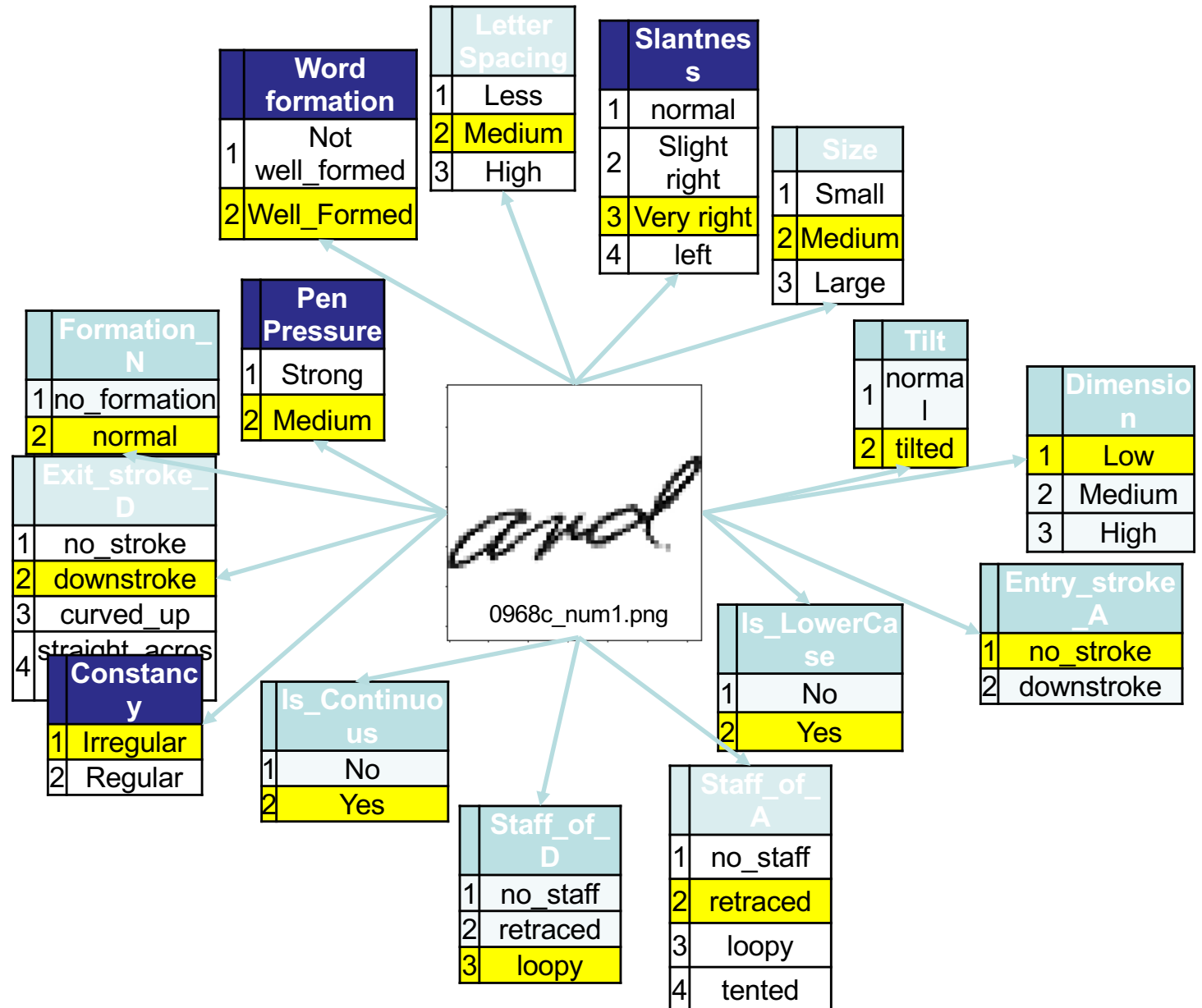


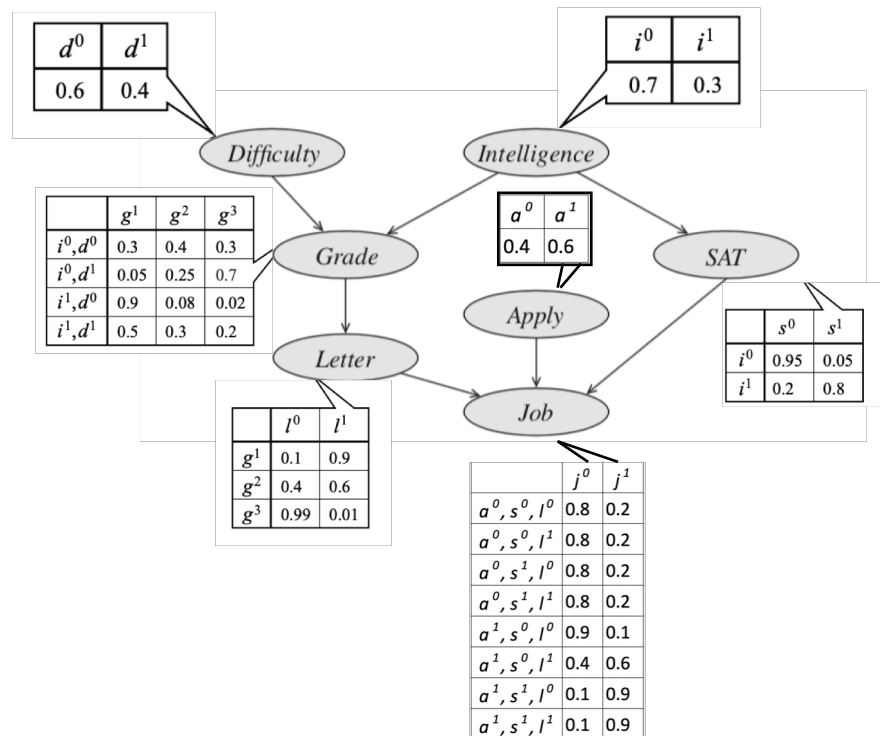
Image 1: Human observed features mapped to image

Most Probable Explanation (MPE)

Observe evidence: student has received a job offer $J=j^1$.

We want an explanation.

1. Most probable (MAP) setting for all variables I, D, G, S, L, A
2. Most probable setting for I



$$P(I, D, G, S, L, A, J) = P(I)P(D)P(G|I, D)P(S|I)P(L|G)P(A)P(J|A, S, L)$$

MPE for all variables

Conditioning with *Job* with $J=j^1$

Exhaustive listing of 96 Combinations with probabilities

MPE is:

- $I = i^1$: Is *Intelligent*
- $D = d^0$: course is *Difficult*
- $G = g^0$: *Grade is A*
- $S = s^1$: high *SAT*
- $L = l^1$: strong *Letter*
- $A = a^1$: *Apply* for Job

Sr. No.	Difficult	Intelligen	Grade	Letter	SAT	Apply	Job	Joint Probabil
3	0	1	0	1	1	1	1	0.0629856
5	0	0	0	1	0	1	1	0.0387828
10	0	0	1	1	0	1	1	0.0344736
13	1	1	0	1	1	1	1	0.023328
16	1	0	2	0	0	0	1	0.01474704
18	1	0	1	1	0	1	1	0.014364
21	1	0	2	0	0	1	1	0.01106028
22	0	1	0	1	0	1	1	0.0104976
23	1	1	2	0	1	1	1	0.01026432
25	0	0	2	0	0	0	1	0.00948024
27	0	1	0	1	1	0	1	0.0093312
28	1	1	1	1	1	1	1	0.0093312
29	0	0	0	1	0	0	1	0.0086184
31	0	0	1	1	0	0	1	0.0076608
32	0	0	2	0	0	1	1	0.00711018
33	0	1	0	0	1	1	1	0.0069984
37	1	1	1	0	1	1	1	0.0062208

MPE inference:

Conditional Probability:

Numerator:

$$P(I, D, G, S, L, A | J = j^1) = \frac{P(I, D, G, S, L, A, J = j^1)}{P(J = j^1)}$$

$$P(I, D, G, S, L, A, J = j^1) = P(I)P(D)P(G|D, I)P(L|G)P(S|I)P(A)P(J = j^1 | A, S, L)$$

Max of numerator: $P(I, D, G, S, L, A, J) = 0.9 * 0.6 * 0.8 * 0.9 * 0.9 * 0.3 * 0.6 = 0.0629856$

Denominator requires Marginal Inference

$$P(J = j^1) = \sum_I \sum_D \sum_G \sum_S \sum_L \sum_A P(I, D, G, S, L, A, J = j^1) = 0.3679$$

MAX Conditional Probability: $P(I, D, G, S, L, A | J = 1) = 0.0629856 / 0.3679 = 0.1712$

MPE for I given $Job = 1$

The most probable explanation of *Intelligence* given $Job : J=j^1$

Answer: $I = i^1$: Is *Intelligent*

Formula:

$$P(I|J = 1) = P(I, J)/P(J = 1)$$

$$P(I, J) = P(I)P(J|A, S, L)P(A)P(S|I)P(L|G)P(G|D, I)P(D)$$

$$P(I = 0, J = 1) = 0.9 * 0.6 * 0.8 * 0.9 * 0.9 * 0.3 * 0.6 = 0.0629$$

$$P(I = 1, J = 1) = 0.6 * 0.7 * 0.3 * 0.9 * 0.95 * 0.6 * 0.6 = 0.0387$$

$$P(I = 0|J = 1) = 0.0387/(0.0629 + 0.0387) = 0.381$$

$$P(I = 1|J = 1) = 0.0629/(0.0629 + 0.0387) = 0.619$$

MAP of Student
Intelligence

intel_0	0.381
intel_1	0.619

Compare to
Prior of Student
Intelligence

i^0	i^1
0.7	0.3

Sampling from Student BN

Algorithm 12.1 Forward Sampling in a Bayesian network

```

Procedure Forward-Sample (
     $\mathcal{B}$     // Bayesian network over  $\mathcal{X}$ 
)
1  Let  $X_1, \dots, X_n$  be a topological ordering of  $\mathcal{X}$ 
2  for  $i = 1, \dots, n$ 
3       $\mathbf{u}_i \leftarrow \mathbf{x}(\text{Pa}_{X_i})$  // Assignment to  $\text{Pa}_{X_i}$  in  $x_1, \dots, x_{i-1}$ 
4      Sample  $x_i$  from  $P(X_i \mid \mathbf{u}_i)$ 
5  return  $(x_1, \dots, x_n)$ 
    
```

	apply	intel	sat	diff	grade	letter	job
0	1	1	1	1	0	1	1
1	0	0	0	0	0	1	0
2	1	1	1	1	1	1	1
3	1	0	0	1	2	0	0
4	1	1	1	0	2	0	1
5	0	0	0	1	2	0	1
6	1	0	0	0	0	1	0
7	1	0	1	0	1	1	1
8	1	0	0	0	1	1	1
9	1	0	0	1	1	0	0

...

...

993	0	0	1	1	1	1	0
994	1	1	1	0	0	1	1
995	1	0	0	0	2	0	0
996	0	0	0	0	0	1	0
997	1	0	0	1	2	0	0
998	0	0	0	0	0	0	1
999	1	0	0	0	2	0	0

1000 rows x 7 columns

PGMPY:

```

from pgmpy.sampling import BayesianModelSampling

forward_sampler = BayesianModelSampling(student)
samples = forward_sampler.forward_sample(size=1000)
    
```

PGM classification over 1000 samples

```
from pgmpy.models import BayesianModel
from pgmpy.factors.discrete import TabularCPD
from pgmpy.inference import VariableElimination

student = BayesianModel()
student.add_nodes_from(['diff', 'intel', 'grade', 'sat', 'letter', 'apply', 'job'])
student.add_edges_from([('diff', 'grade'),
                        ('intel', 'grade'),
                        ('intel', 'sat'),
                        ('grade', 'letter'),
                        ('apply', 'job'),
                        ('letter', 'job'),
                        ('sat', 'job')])

student.fit(train)

inference = VariableElimination(student)
counter = 0
for index, row in test.iterrows():
    mle1 = inference.map_query(variables=['sat', 'letter'],
                              evidence={'job': row['job']})
    mle2 = inference.map_query(variables=['intel'],
                              evidence={'sat': mle1['sat'],
                                       'letter': mle1['letter']})
    if(mle2['intel'] == row['intel']):
        counter += 1

print('Accuracy is: ' + str(counter/test.shape[0]*100))

Accuracy is: 88.125
```

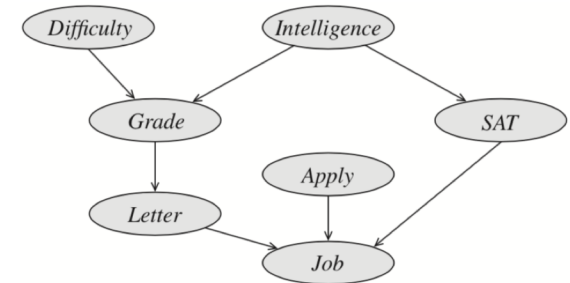


Figure 5.3 The Student example augmented with a *Job* variable

Results:

- Training Accuracy: 91.875%
- **Testing Accuracy: 88.125%**

Logistic Regression for Intelligence

Training Hyperparameters:

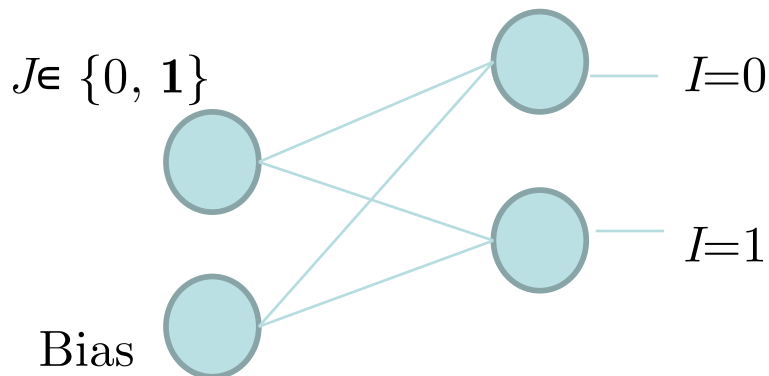
- Number of Epochs: 100
- Batch Size: 100
- Number of samples: 1000
- Number of Training samples: 800
- Number of Validation samples: 200

Results:

- Training Accuracy: 94.62%
- Training Loss: 0.1580
- **Testing Accuracy: 96%**
- **Testing Loss: 0.1534**

Logistic Regression output:

	Intelligence = 0	Intelligence = 1
Job = 1	0.038	0.962



Explanation:

MPE

evidence: student **gets job**

explanation: 96% probability that student is **intelligent**;

Formula:

$$P(I = 0 | J = 1) = 0.498$$

$$P(I = 1 | J = 1) = 0.502$$

Explainability vs Performance

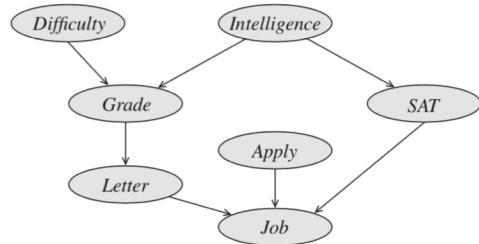


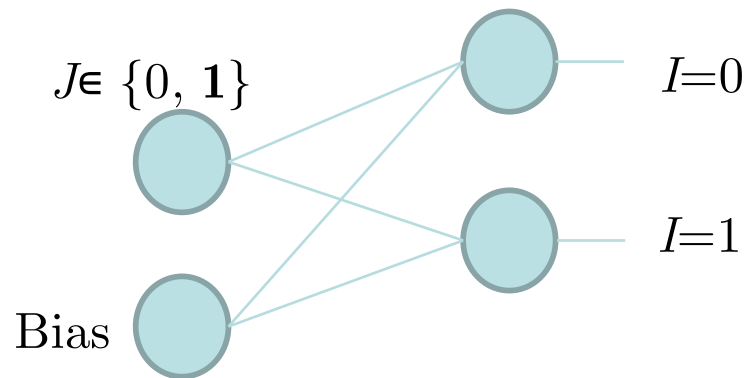
Figure 5.3 The Student example augmented with a *Job* variable

```
inference = VariableElimination(student)
counter = 0
for index, row in test.iterrows():
    mle1 = inference.map_query(variables=['sat', 'letter'],
                              evidence={'job': row['job']})
    mle2 = inference.map_query(variables=['intel'],
                              evidence={'sat': mle1['sat'],
                                       'letter': mle1['letter']})
    if(mle2['intel'] == row['intel']):
        counter += 1
print('Accuracy is: ' + str(counter/test.shape[0]*100))
```

Accuracy is: 88.125

Explainable System Results:

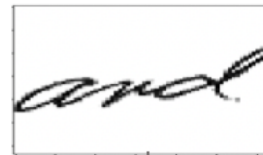
- Training Accuracy: 91.875%
- **Testing Accuracy: 88.125%**



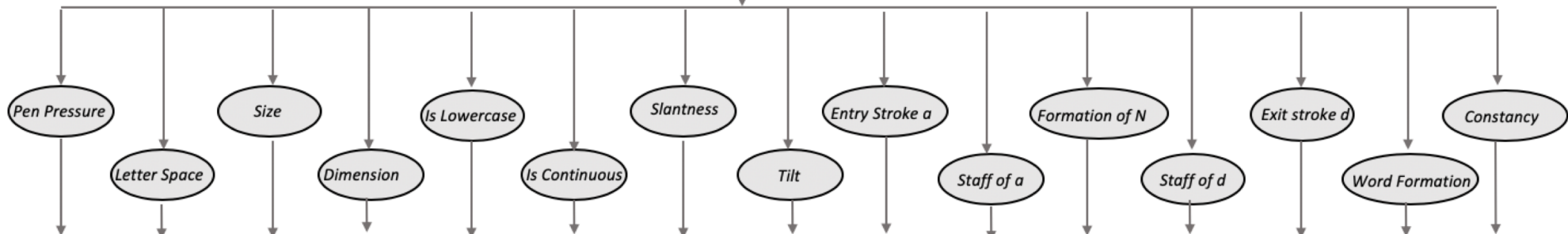
High Performance System Results:

- Training Accuracy: 94.62%
- Training Loss: 0.1580
- **Testing Loss: 0.1534**
- **Testing Accuracy: 96%**

Handwriting comparison– expert features

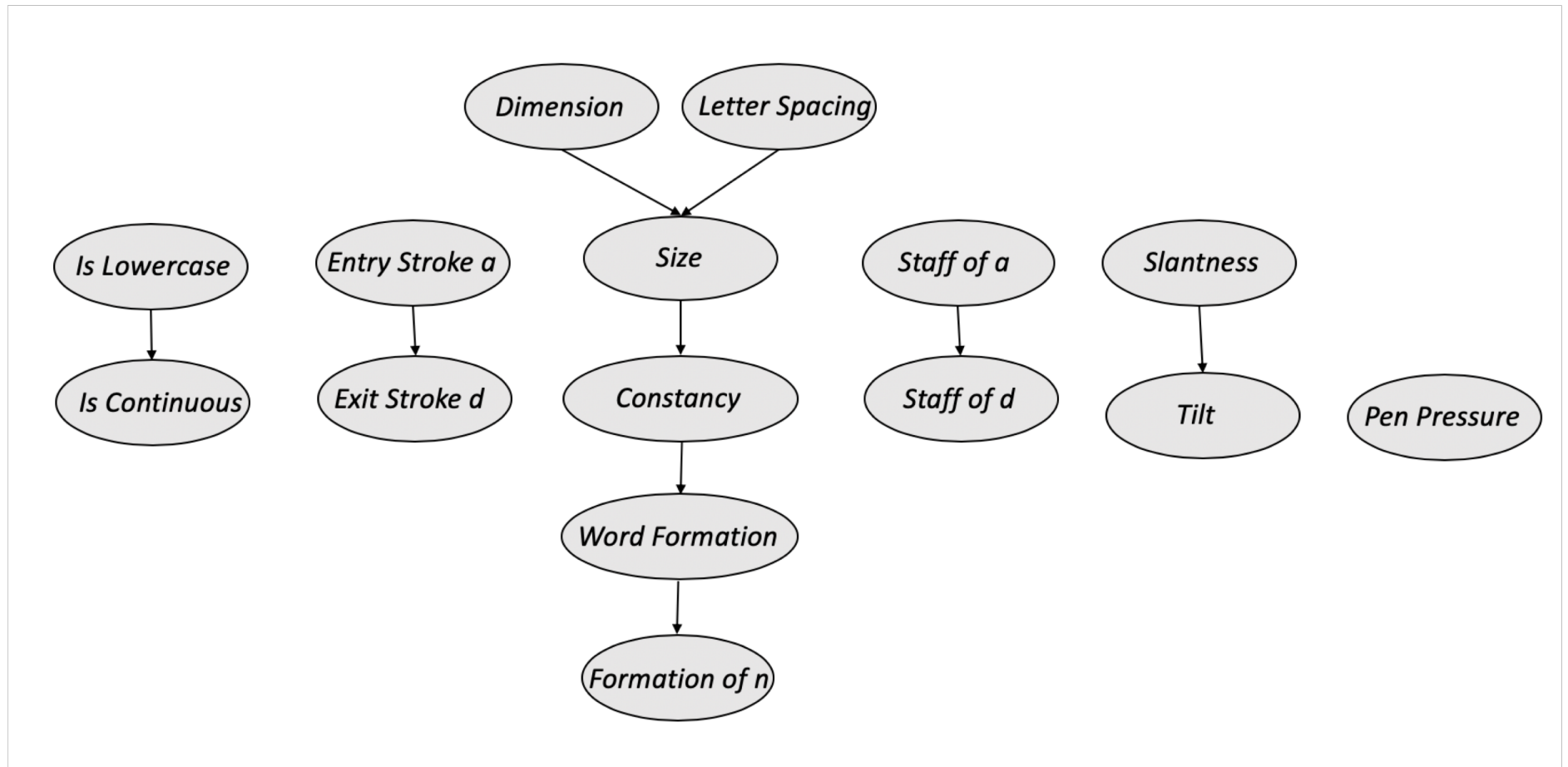


Document Examiner (DQE)

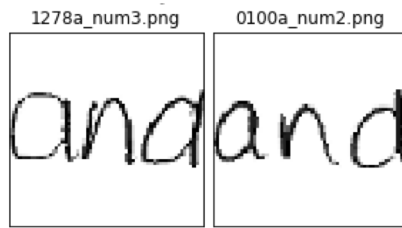


	Pen Pressure	Letter Spacing	Size	Dimension	Is LowerCase	Is Continuous	Slantness	Tilt	Entry stroke "A"	Staff of "A"	Formation "N"	Staff of "D"	Exit stroke "D"	Word forma	Constancy
1	Strong	Less	Small	Low	No	No	normal	normal	no stroke	no staff	no formation	no staff	no stroke	Not well form	Irregular
2	Medium	Medium	Medium	Medium	Yes	Yes	slight right	tilted	downstroke	retraced	normal	retraced	downstroke	Well Formed	Regular
3		High	Large	High			very right			loopy		loopy	curved up		
4							left			tented			straight across		

PGM for human observed 'AND' dataset

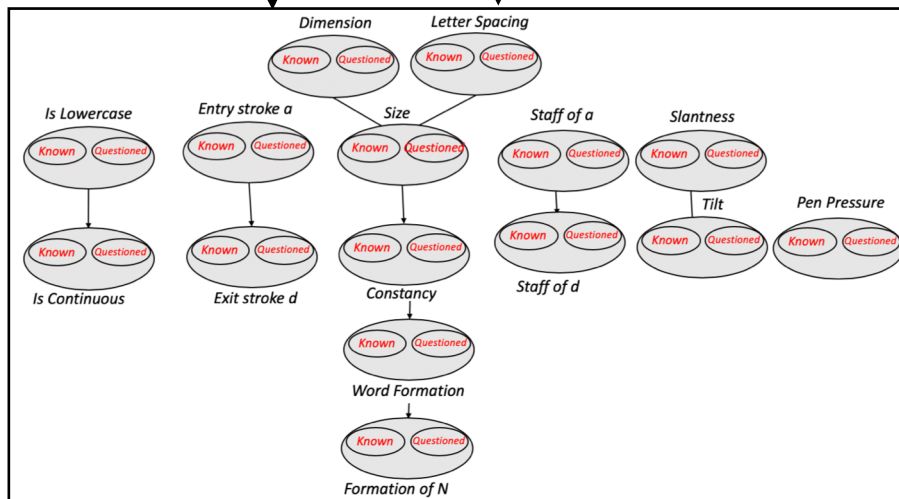


PGM for Writer Verification



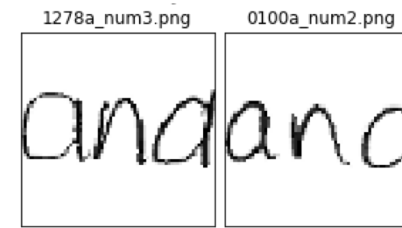
Known Questioned

↓ (k) ↓ (q)



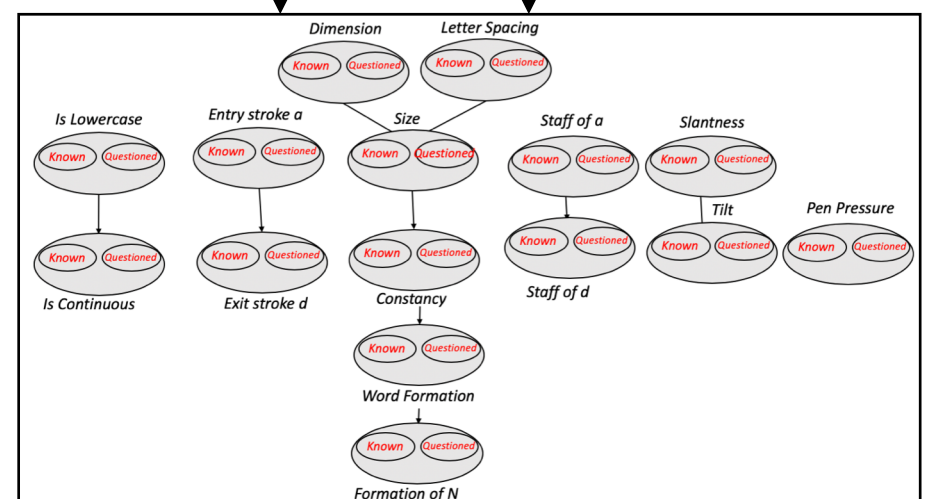
PGM for same writers ($h = 0$)

$$P(k, q | h = 0)$$



Known Questioned

↓ (k) ↓ (q)



PGM for different writers ($h = 1$)

$$P(k, q | h = 1)$$

$$LLR = \ln \left(\frac{P(k, q | h = 0)}{P(k, q | h = 1)} \right)$$

Finding LLR

$$LLR = \ln \left(\frac{P(k, q | h = 0)}{P(k, q | h = 1)} \right)$$

Approach 1

· A simple **Distance based LLR** approach defines scalar distance between known and questioned samples defining loglikelihood ratio as follows:

$$LLR_D = \ln \left(\frac{P(d(k, q) | h = 0)}{P(d(k, q) | h = 1)} \right)$$

Disadvantage: Distribution of N variables is mapped into a scalar thereby incurring severe loss of information (many pairs of k and q can have the same distance).

Approach 2

· **Rarity with Distance based LLR** approach uses · strength of the features between the known and the questioned samples defining loglikelihood ratio as follows:

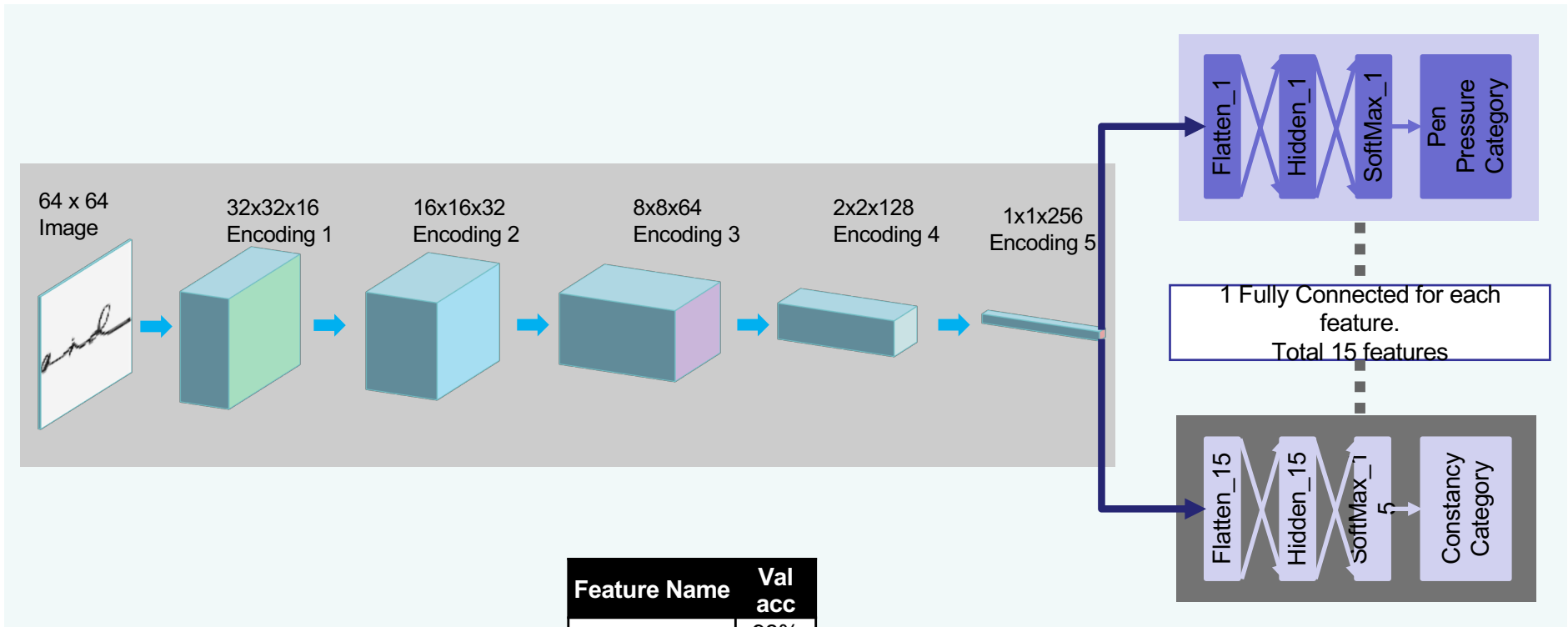
$$LLR_{RD} = \ln \left(\frac{\frac{1}{\sqrt{2}\sigma} e^{-\frac{(k-q)^2}{4\sigma^2}}}{\frac{1}{\tau} e^{-\frac{(m-\mu)^2}{4\tau^2}}} \right)$$

← Distance

← Rarity

Disadvantage: Not always computationally tractable

Feature extraction architecture/performance



Feature Name	Val acc
pen_pressure	98%
letter_spacing	78%
size	92%
dimension	89%
is_lowercase	99%
is_continuous	94%
slantness	72%
tilt	98%
entry_stroke_a	97%
staff_of_a	84%
formation_n	95%
staff_of_d	86%
exit_stroke_d	66%
word_formation	89%
constancy	84%

Comparison performance

Performance based on Overall Similarity

Explainability:

- Compare corresponding 'SoftMax_x' ($x : \{1 \text{ to } 15\}$) of two images using cosine similarity to obtain feature wise explanation

Metric	Seen	Unseen	Shuffled
Intra Writer Accuracy	88%	71%	81%
Inter Writer Accuracy	96%	93%	95%
Average Accuracy	96%	94%	95%

Explanation of Comparison

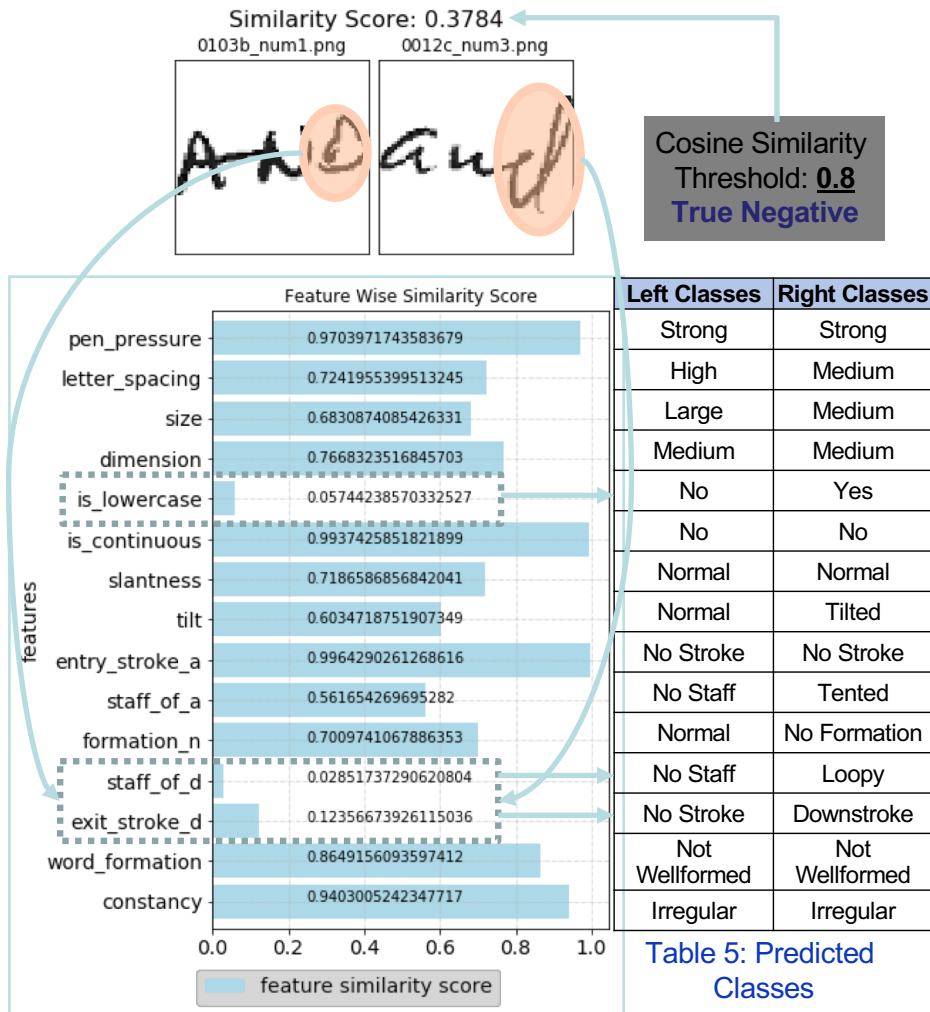


Image 6: Graph of Similarity Score of Left & Right Image Features
Different Writer Samples

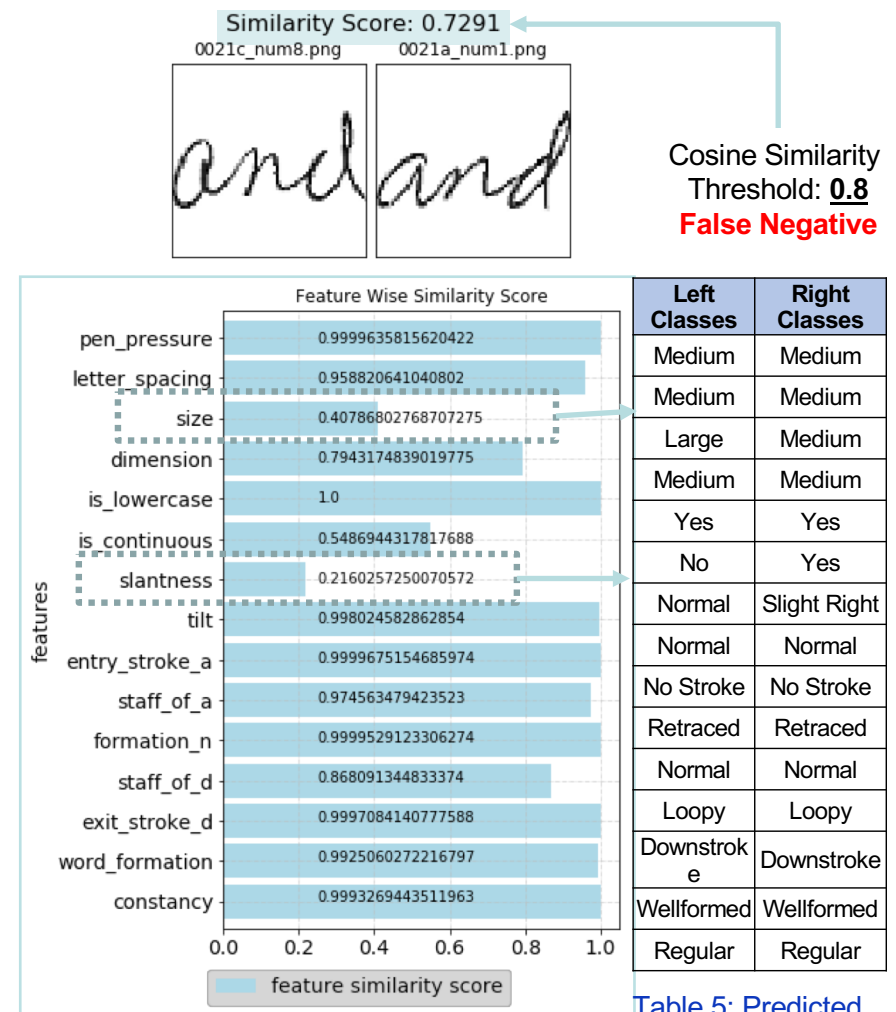


Image 7: Graph of Similarity Score of Left & Right Image Features
Same Writer Samples

Report Card of Comparing two Samples

Summary and Conclusion

- Next wave of AI is probabilistic AI (Intel)
- Most Probable Explanation (MPE) of PGMs is useful for XAI
- Forensics demands explainability
- Forensic Impression evidence lends itself to a combination of deep learning and probabilistic explanation