

Outline

- Introduction to XAI (Derek Doran, Sargur Srihari, David Gunning, John Sowa)
- Explanation AI – Past, Present and Future (William J. Clancey)
- DARE: Deep Attentional Representations for Explanations (Giedrius T. Buračas)
- Explainable AI: Probabilistic Approach (Sargur N. Srihari)
- Summary

Agenda

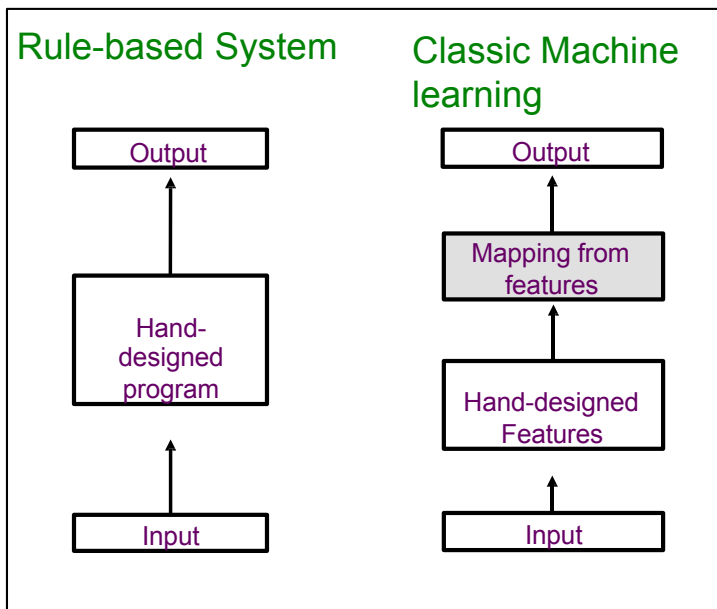
- November 28, 2018: **Derek Doran**, Wright State University, *An Overview of Explainable AI (XAI)*
- February 20, 2019: **William J. Clancey**, Florida Inst. For Human and Machine Cognition, *Explainable AI: Past, Present, and Future – A Scientific Modeling Approach*
- April 3, 2019: **Giedrius T. Buračas**, SRI, *DARE: Deep Attentional Representations for Explanations*
- April 10, 2019: **Sargur Srihari (Hari)**, SUNY Buffalo, *Explainable Artificial Intelligence: The Probabilistic Approach*

Evolution of AI

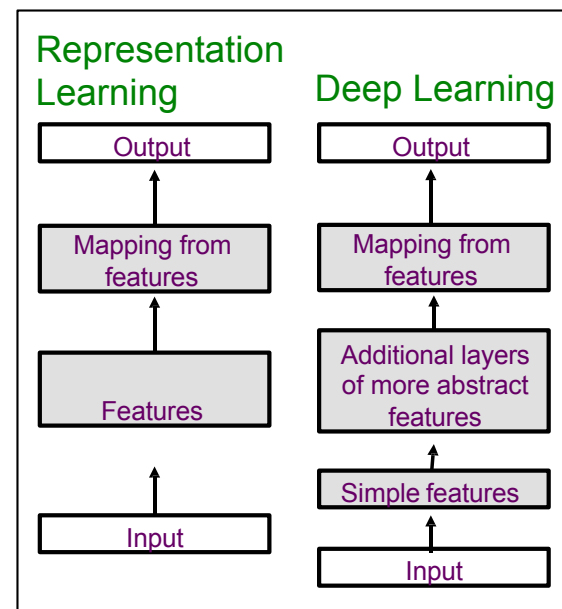
- Prehistory (Paleolithic: Indian and Greek Philosophies; Neolithic: Hume, Russell, Turing)
- The Beginning (1956: AI coined at Dartmouth Conference)
- Early years (1957 – 1980: GPS, Dendral, Hearsay, Macsyma, Mycin, Shrdlu, Shakey, etc.)
- Knowledge is power & Early Neural Networks (1980 -- 1990: The first wave)
- The Silent Period (1990 – 2000: Deep Blue, Rise of Robots, Commercialization of AI technologies, e.g., Speech Recognition)
- Neural Networks to the Fore (2000 – 2015: The second wave: Deep Learning)
- Symbiosis of Neural and Knowledge Networks (2015 – 2025: **Explainable AI**, OKN, etc. – The third wave)
- The Conscious Machine (2025 -- ?? -- The fourth wave: Tsunami)

Waves of AI Models

FIRST WAVE



SECOND WAVE



■ Shaded boxes indicate components that can learn from data

Focus was on discriminative models: supervised learning

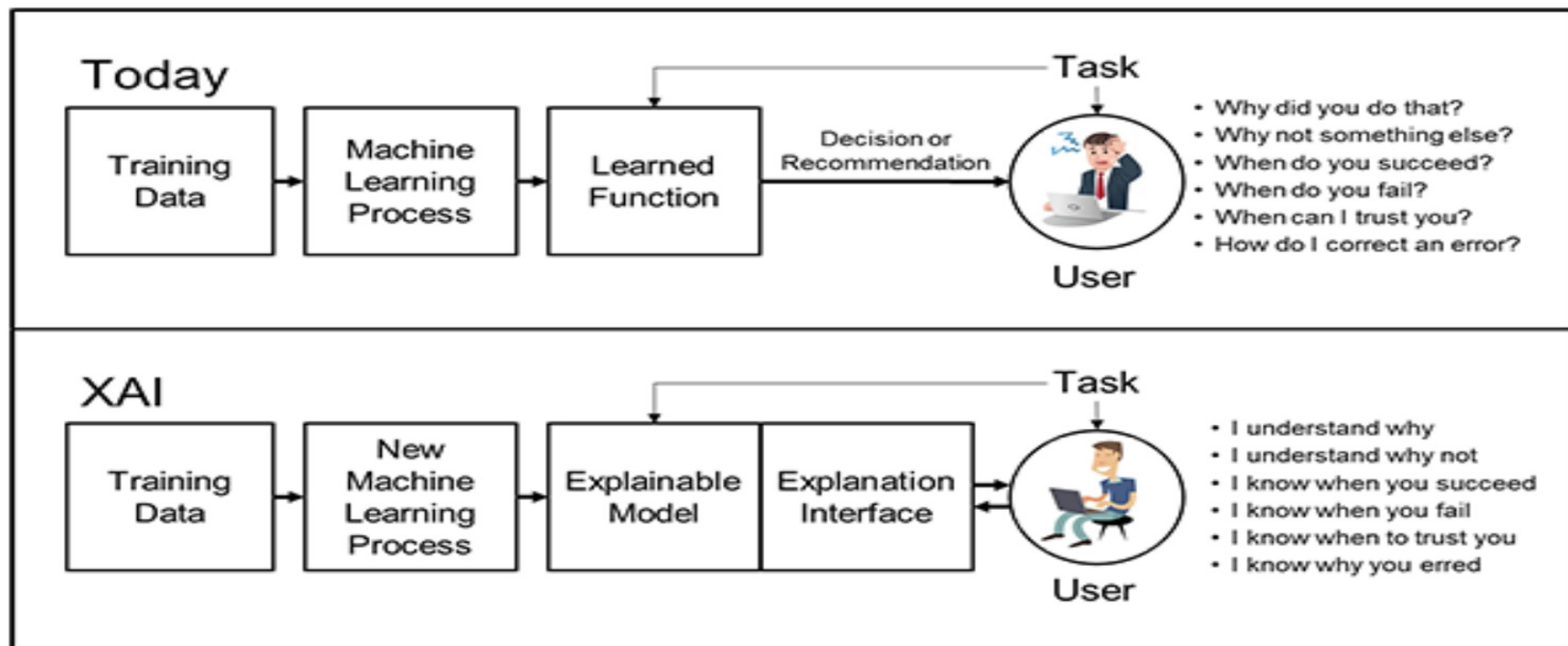
Third Wave: Probabilistic AI

- 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

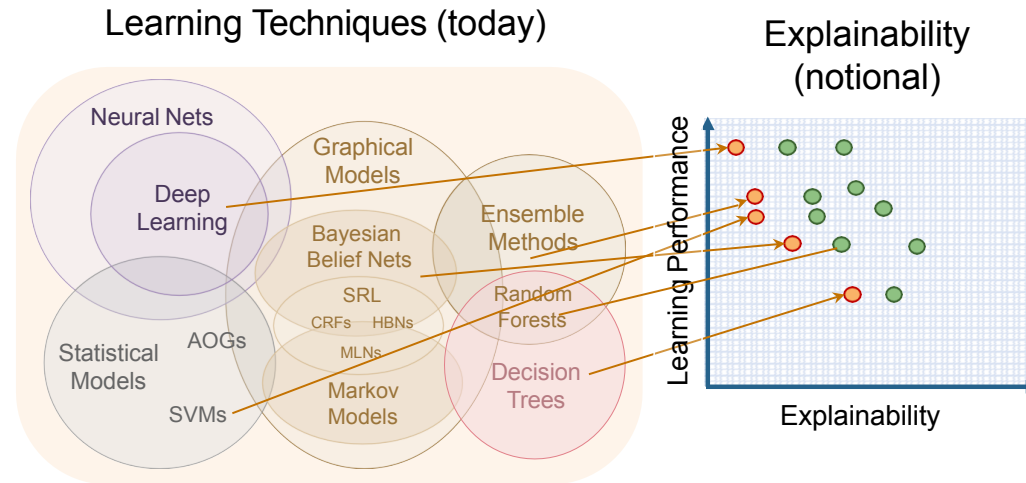
What is an Explanation

- An explanation is the ability to answer how or why question and to answer a follow up question to clarify in a particular context.

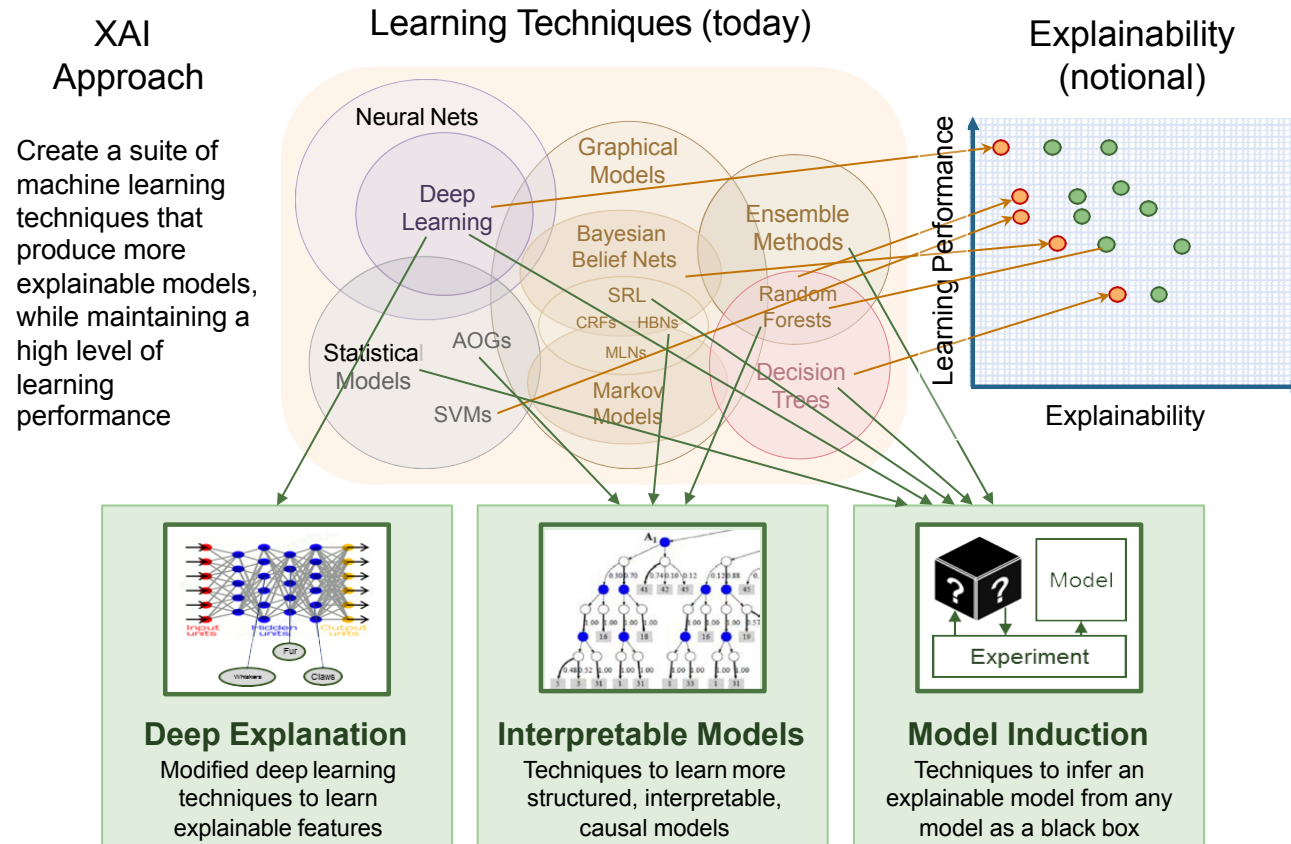
Toward XAI



Performance vs. Explainability: DARPA XAI Program



Performance vs. Explainability: DARPA XAI Program



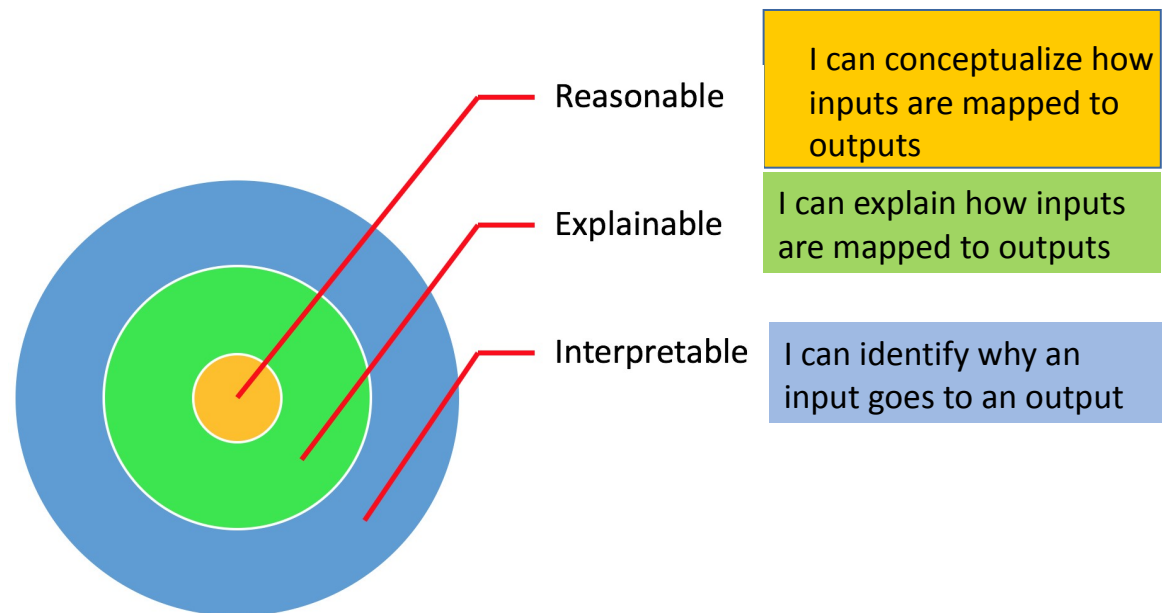
Types of AI Systems

- Symbolic
- Sub-symbolic
 - White-box – You can see model mechanisms simple enough to be able to trace how inputs map to outputs (Regressions, Decision Trees, Rule mining, SVMs)
 - Grey-box – You have some vision mechanisms, but parameters are numerous, decisions are probabilistic, or inputs gets lost (Clustering, Bayesian nets, Genetic algorithms, ..)
 - Black-box – The model is so complex with so large a parameter space that is not easily decipherable (Neural Networks)

Four Theses of XAI

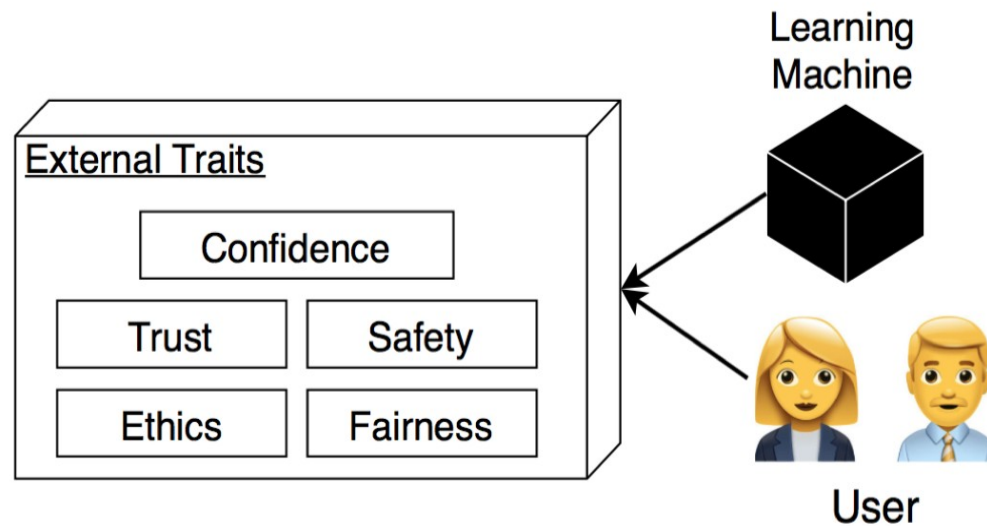
1. Explainable AI defines a field of research, not define a particular “type” of AI.
2. There exist unique notions of ‘explanations’, making problem or context-independent explainable AI work inappropriate.
3. “Strong” and “Ultrastrong” AI systems (by Michie's definition) demands the integration of symbolic systems and reasoning to provide operationally effective communications of the internal state and operation of an AI.
4. Meaningful ontologies and knowledge bases, and Symbolic AI methods are a fundamental (but right now, mostly ignored) to the most important XAI use-case: when in practice, within some context, a layman must understand, trust, and be responsible for the conclusions an AI draws.

Hierarchy of Explanations



Traits of Reasoning XAI Systems

- Typical desiderata of XAI traits are only enabled through reasoning



Explanation AI – Past, Present and Future: A Scientific Modeling Approach

William J. Clancey

MYCIN – First AI System to Explain Reasoning

- MYCIN was developed at Stanford in the 1970s as a diagnostic tool for infectious diseases
- MYCIN had an Explanation Module which provided explanation on why certain decisions were taken and why certain questions were not asked.
- EMYCIN is a domain independent version of MYCIN
- EMYCIN later influenced development of knowledge-based expert systems in many domains (SACON – Structural, CASTER – Mechanical)
- Ontologies were implicit in above systems

Therapy Explanation in MYCIN

**** WHY DID YOU GIVE CHLORAMPHENICOL FOR E.COLI IN REC-1?**

CHLORAMPHENICOL was prescribed for ITEM-2 in RECOMMENDATION-1:

Since

- CHLORAMPHENICOL is a treatment of choice for e.coli in meningitis
- ITEM-2 is sensitive to CHLORAMPHENICOL
- there were no contraindications for it

CHLORAMPHENICOL was prescribed because it was part of the recommendation that covers for all of the items, using the fewest number of drugs.

Replacing original therapy rules by table-driven generate & test method made the therapy selection process formal & perspicuous, significantly improving maintainability, and provided a new capability for explanation and instruction.

“Why Not?” Therapy Explanation in MYCIN

**** WHY DIDN'T YOU SUGGEST PENICILLIN IN REC-1 FOR STAPH-COAG + ?**

PENICILLIN was not prescribed for ITEM-1 in RECOMMENDATION-1:

PENICILLIN was discounted for ITEM-1 because it is NOT DEFINITE that the item is sensitive to this drug.
There are other potential therapies under consideration which are much more desirable, viz., current therapies or drugs to which the item is definitely sensitive.

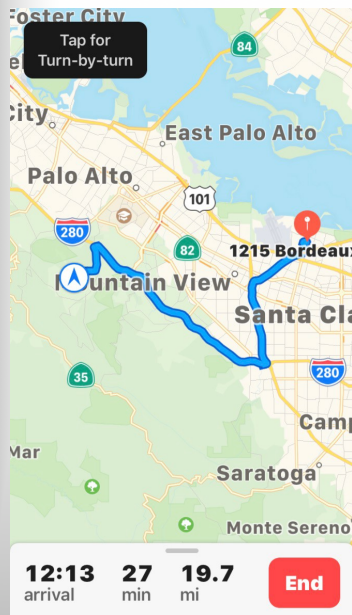
Would you like to see some details? **** YES**

The drugs to which the staphylococcus-coag-pos is sensitive are: cephalothin (1.0) vancomycin (1.0)
gentamycin (1.0) tobramycin (1.0) erythromycin-and-tetracycline (1.0) chloramphenicol-and-erythromycin (1.0)
[RULE098 RULE445]

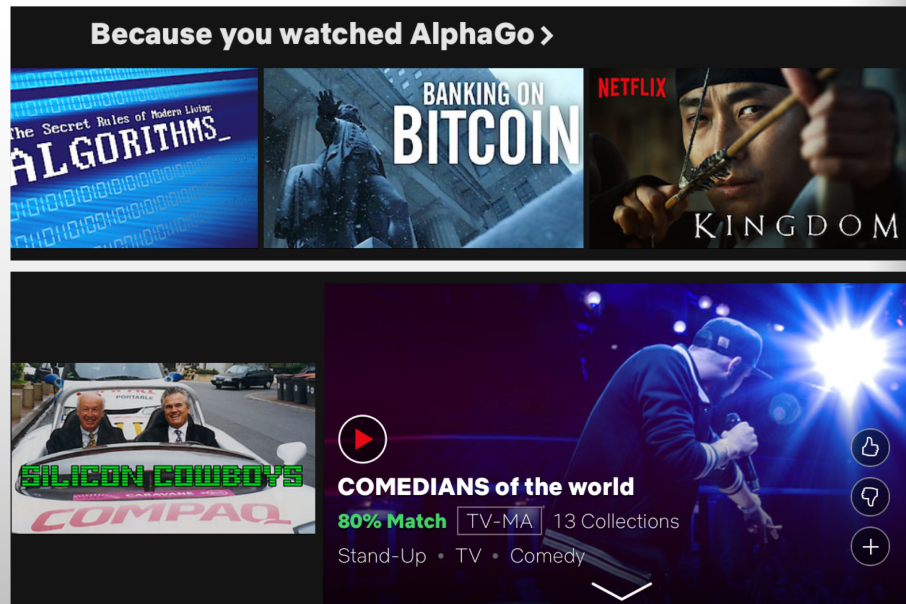
Would you like to know about the history of PENICILLIN in the decision
process up to this point? **** YES**

-- PENICILLIN is a treatment of choice for staphylococcus-coag-pos in meningitis. But as explained above,
PENICILLIN was discounted.

Common Interactive Systems Today Cannot Explain Behavior or Advice



Why not Page Mill & 101?
What comes after 280?



How are these Netflix suggestions related to AlphaGo?
Show me the highest “match” ones first...



Explanation Might Address Shortcomings of Symbolic AI



- Identifying domain representations as “knowledge” obscured system-modeling methods & hence the domain-general scientific accomplishment
→ *abstract the modeling frameworks & operational tasks*
- Ongoing tuning and extension required “Knowledge Engineers”
→ *e.g., need principles for appropriate/sufficient data sets*
- Brittle – boundaries not tested; system not reflective
→ *recognize operating outside designed & tested situations*
- Not integrated with legacy systems & work practice
→ *develop tools for work systems; iterative experiments with prototypes*

DARE: Deep Attentional Representations for Explanations

Giedrius T. Buračas

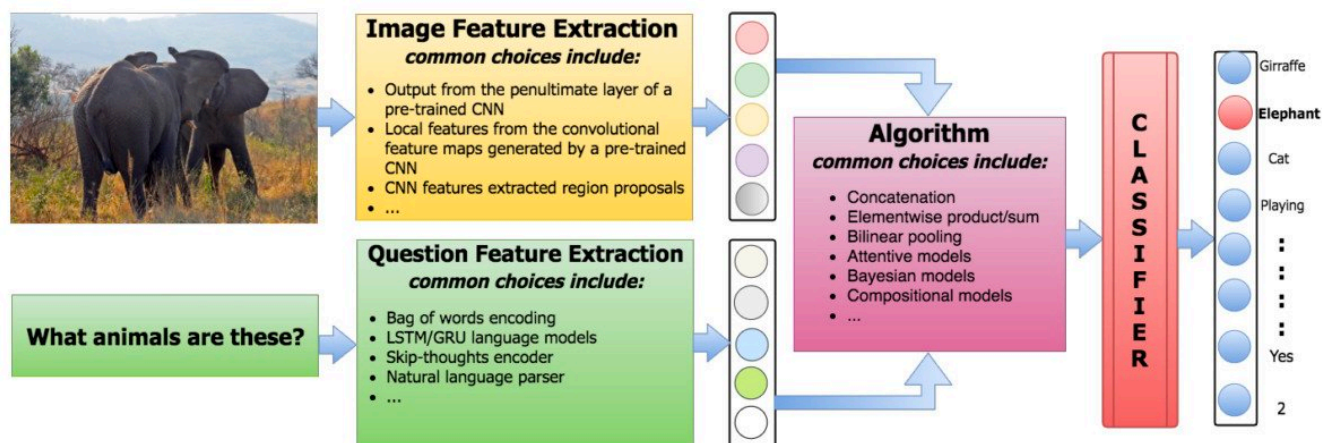
Visual Question Answering (VQA)

- Given an Image, VQA system could answer the following:
 - What is in the image?
 - Are there any humans?
 - What sport is being played?
 - Who has the ball?
 - How many players are in the image?
 - Who are the teams?
 - Is it raining?



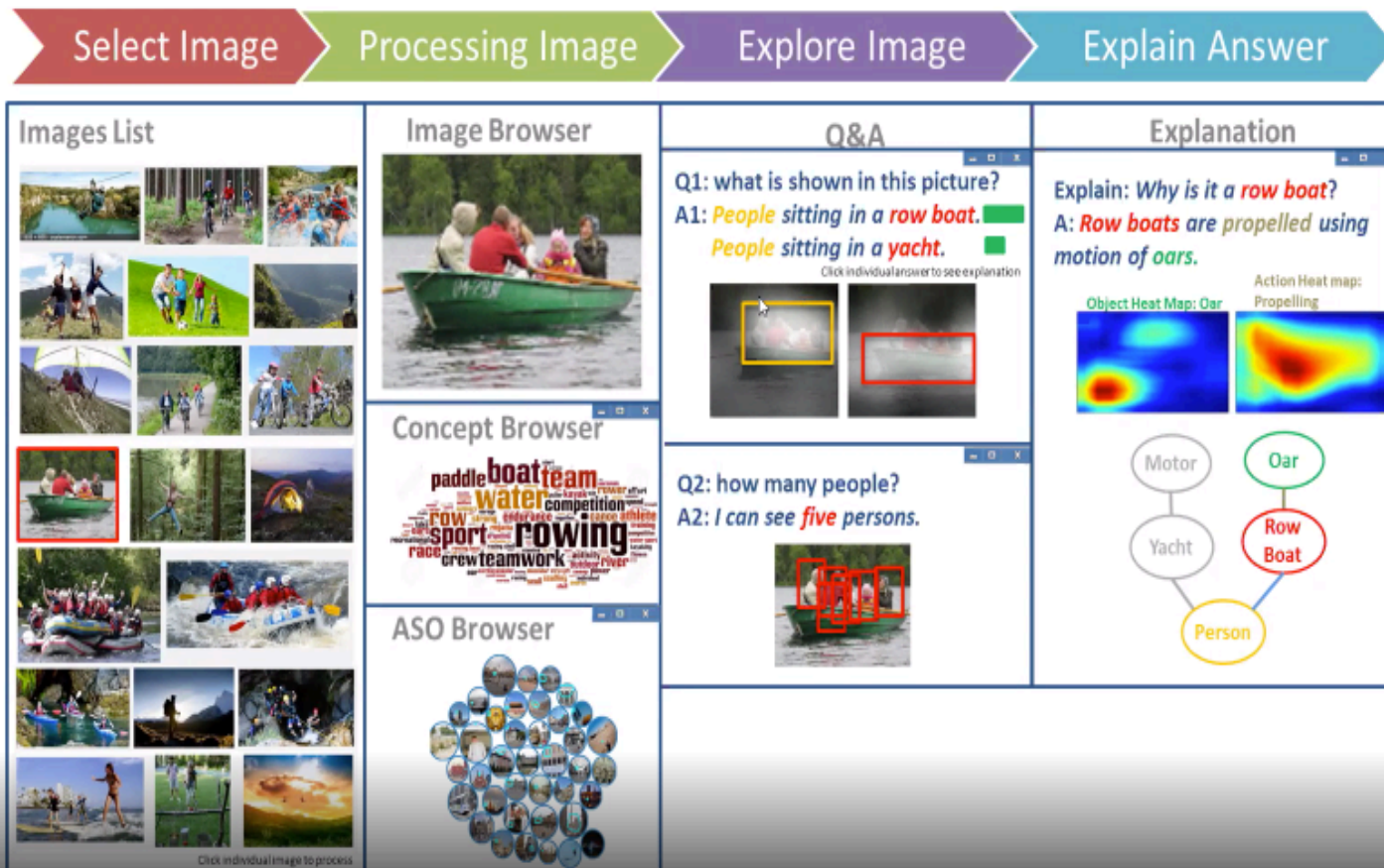
Current Approaches

- Extract features from the question
- Extract features from the image
- Combine features to generate and answer



DARE: Deep Attentional Representations for Explanations

Interpretable and Explainable

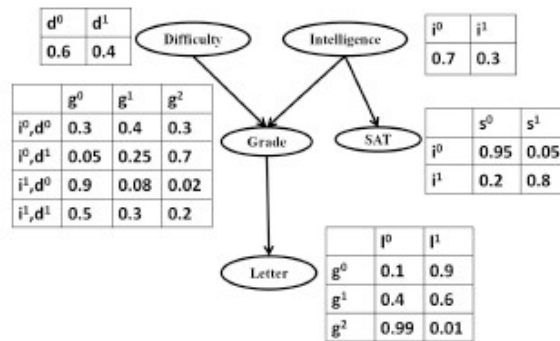


Explainable AI: Probabilistic Approach

Sargur N. Srihari

Probabilistic Graphical Models

Bayesian Networks

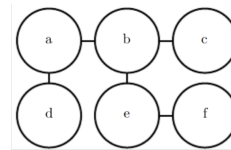


- CPDs: $p(x_i | pa(x_i))$
- Joint Distribution $\mathbf{x} = \{x_1, \dots, x_n\}$

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

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

Markov Networks



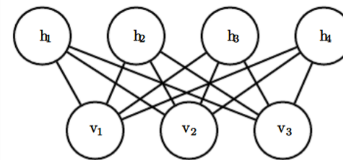
$$p(\mathbf{x}) = \frac{1}{Z} p^*(\mathbf{x}) \quad p^*(\mathbf{x}) = \prod_{C \in \mathcal{C}} \phi(C)$$

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

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

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

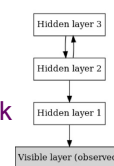
Restricted Boltzmann machine



$$E(v, h) = -b^T v - c^T h - v^T W h$$

$$p(h|v) = \prod_i p(h_i|v) \text{ and } p(v|h) = \prod_i p(v_i|h)$$

Deep Belief Network



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

image	name	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

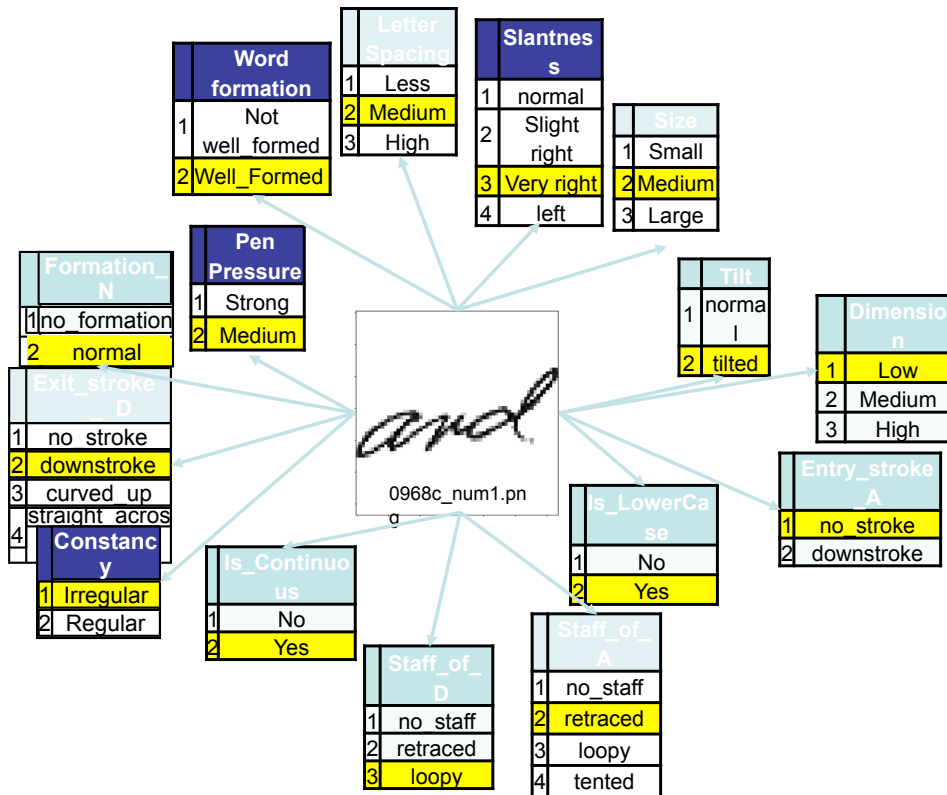
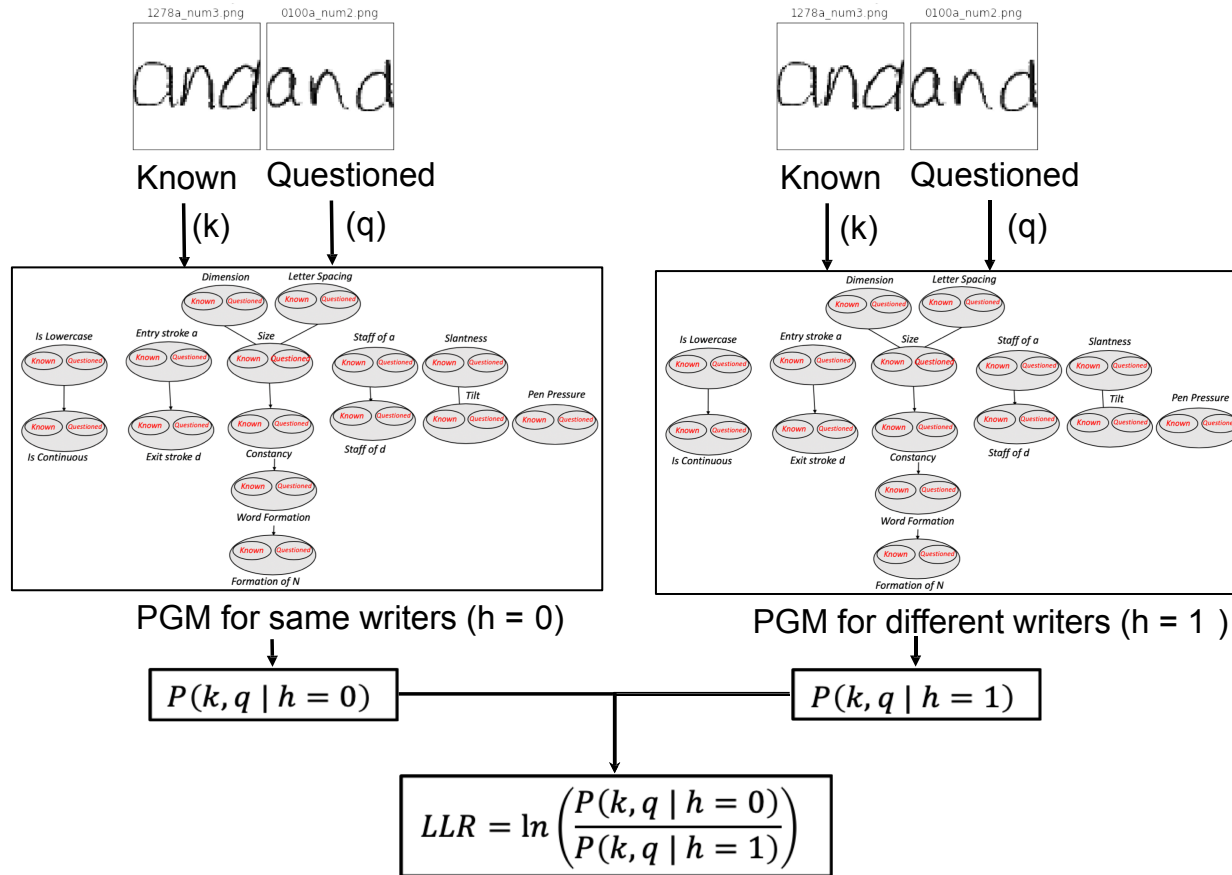
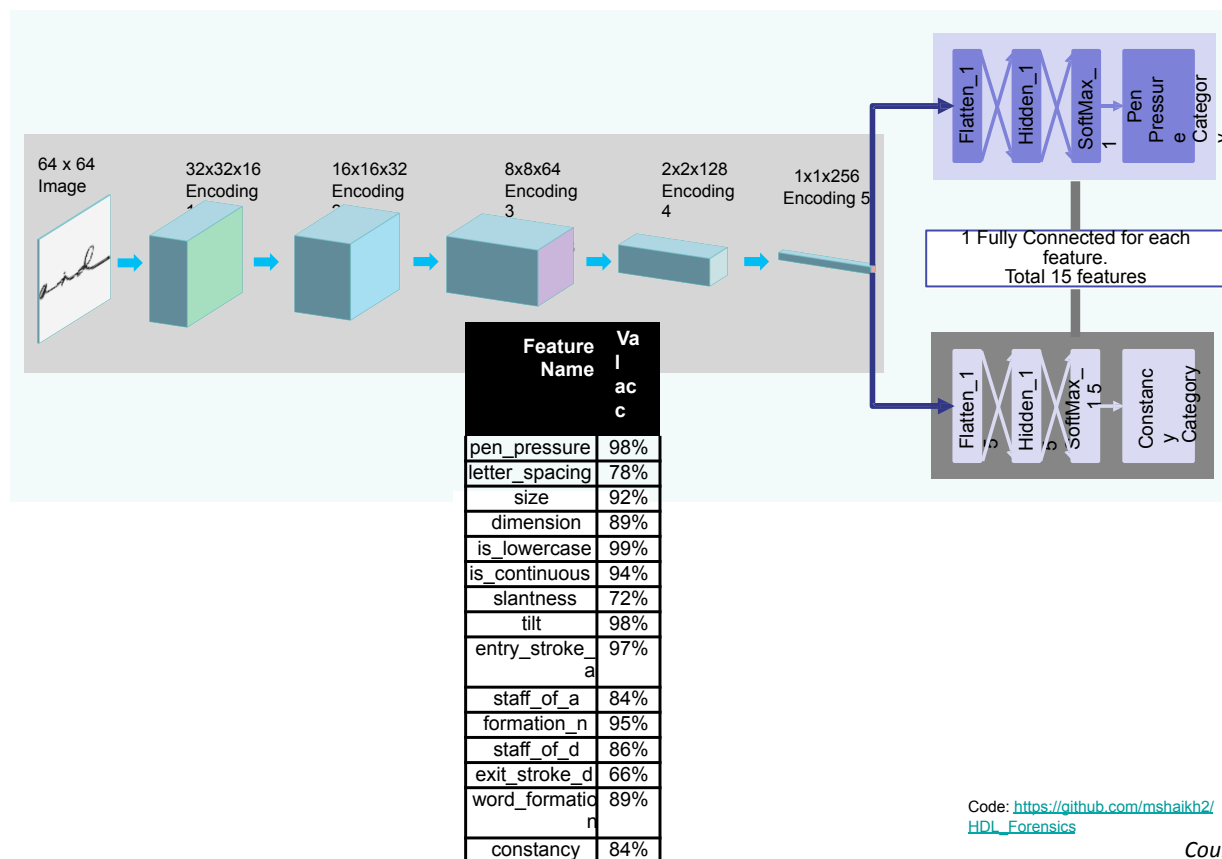


Image 1: Human observed features mapped to image

PGM for Writer Verification



Feature extraction architecture/performance



Code: https://github.com/mshaikh2/HDL_Forensics

Courtesy: Sargur N. Srihari

Srihari's Conclusions

- 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 (Enhanced Third Wave)

Outline

- Introduction to XAI (Derek Doran, Sargur Srihari, David Gunning, John Sowa)
- Explanation AI – Past, Present and Future (William J. Clancey)
- DARE: Deep Attentional Representations for Explanations (Giedrius T. Buračas)
- Explainable AI: Probabilistic Approach (Sargur N. Srihari)
- Summary

Summary

- Several AI waves
- Current wave will use deep learning and knowledge networks (PGMs are an example)
- Explanation will provide trust in AI
- Ontologies will play a major role in XAI
 - Knowledge structure can help in creating explanations
- Lots of synergy across tracks