Explaining & Verifying AI Systems

Adnan Darwiche

UCLA

Explaining AI Systems

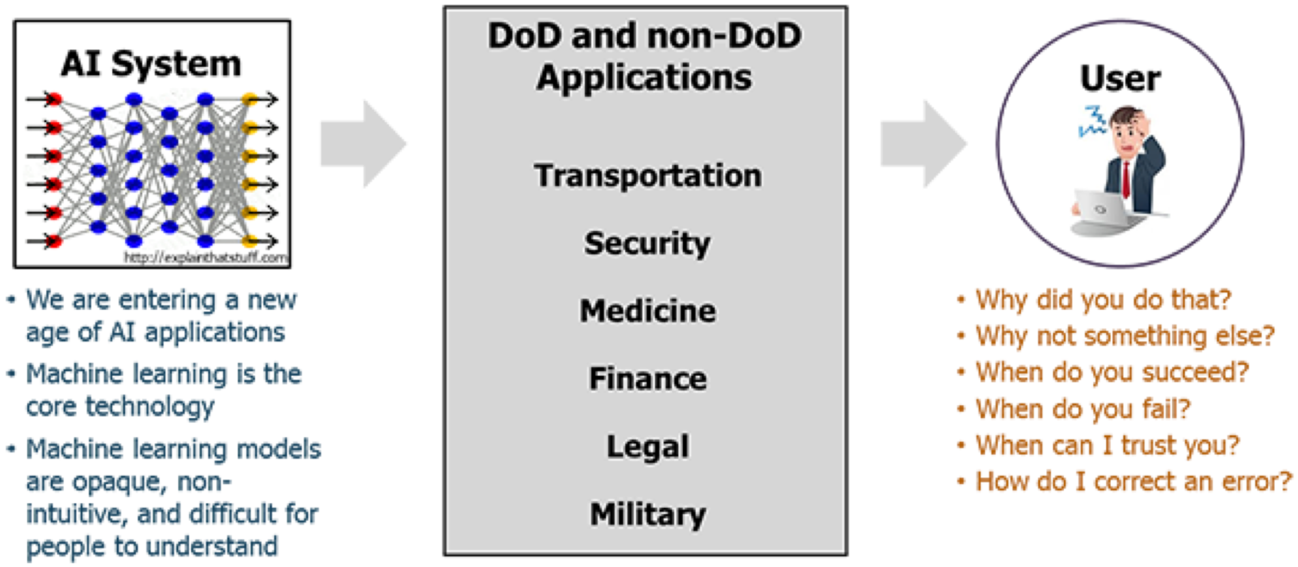
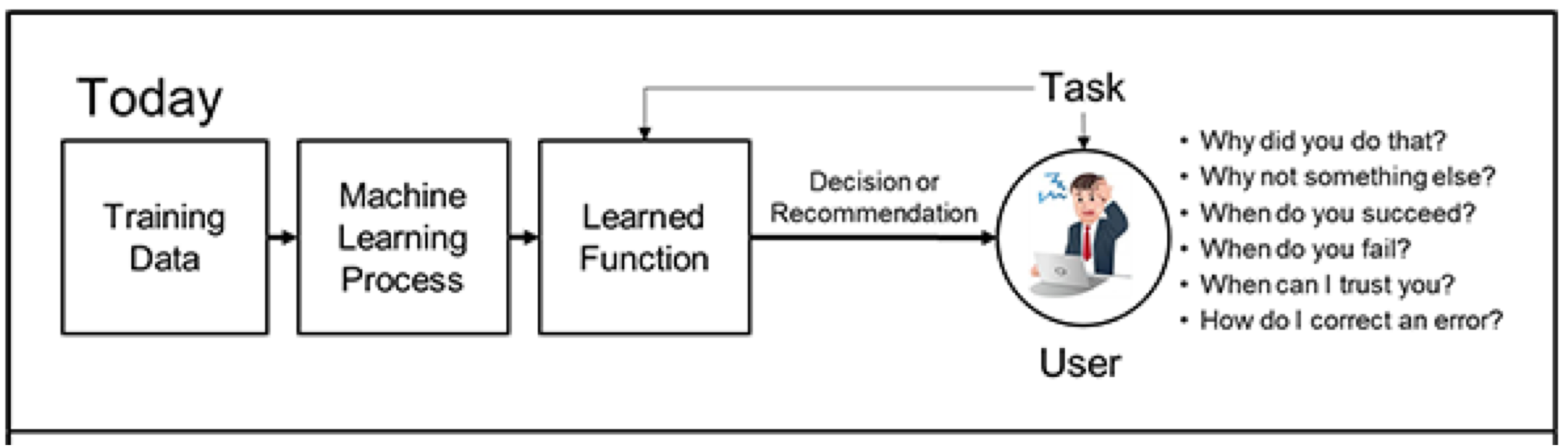


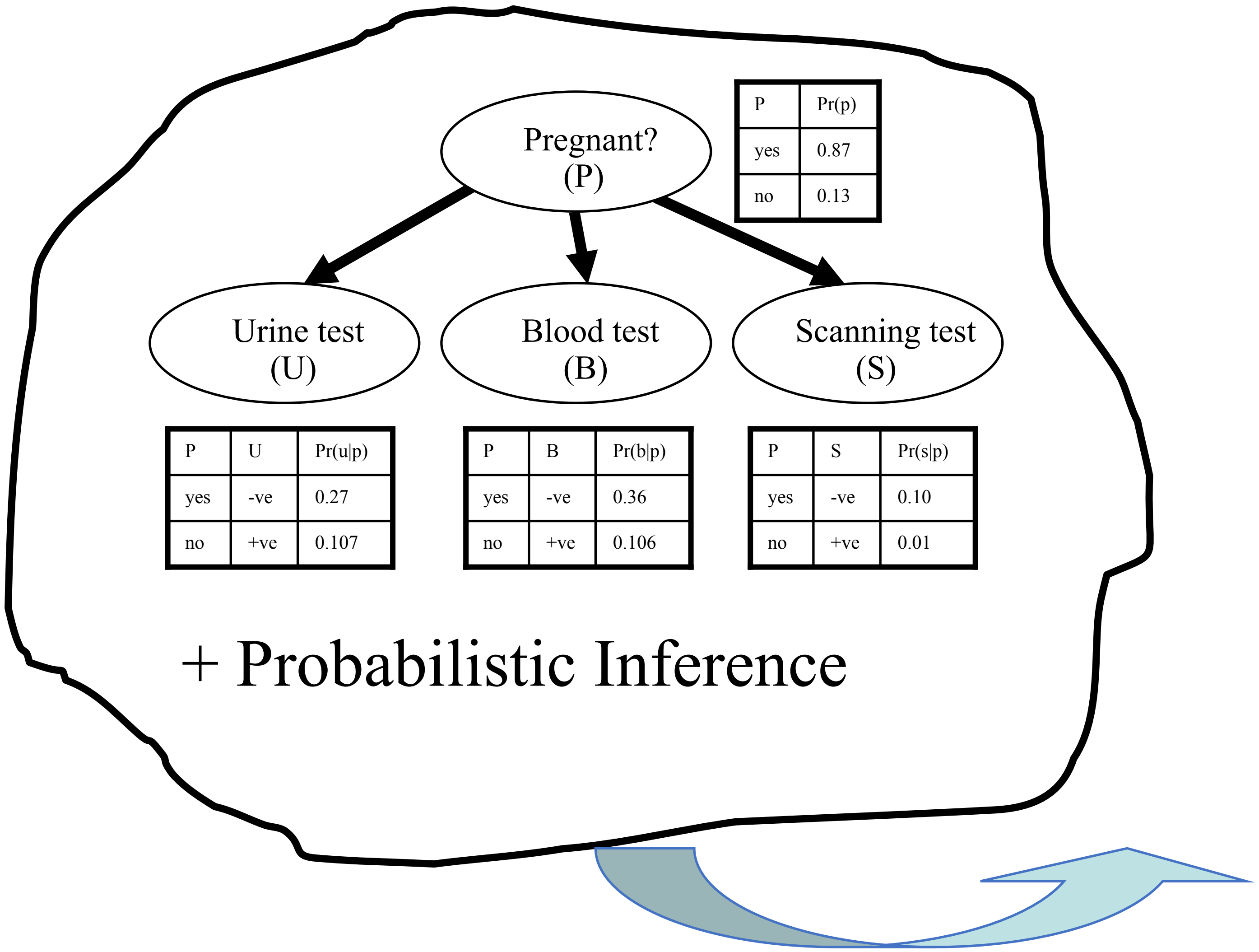
Figure 1. The Need for Explainable AI

XAI DARPA Program

Explaining AI Systems



From Numbers to Decisions

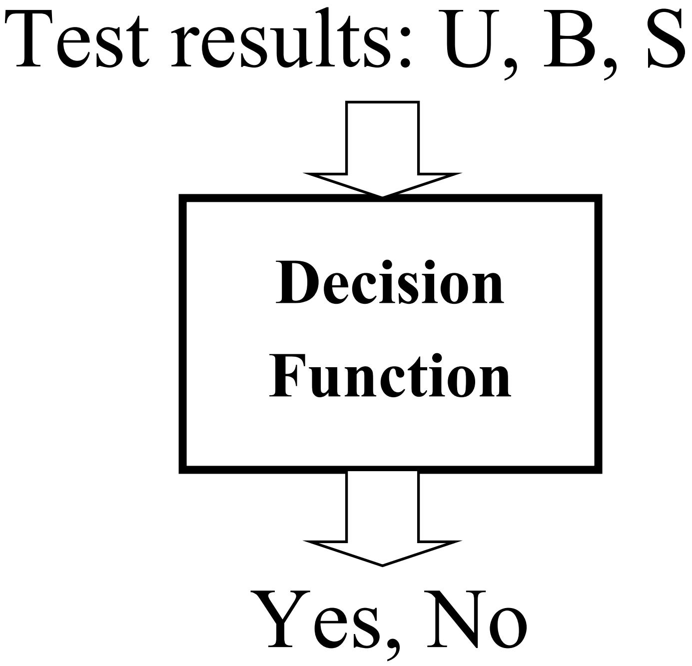


P	Pr(p)
yes	0.87
no	0.13

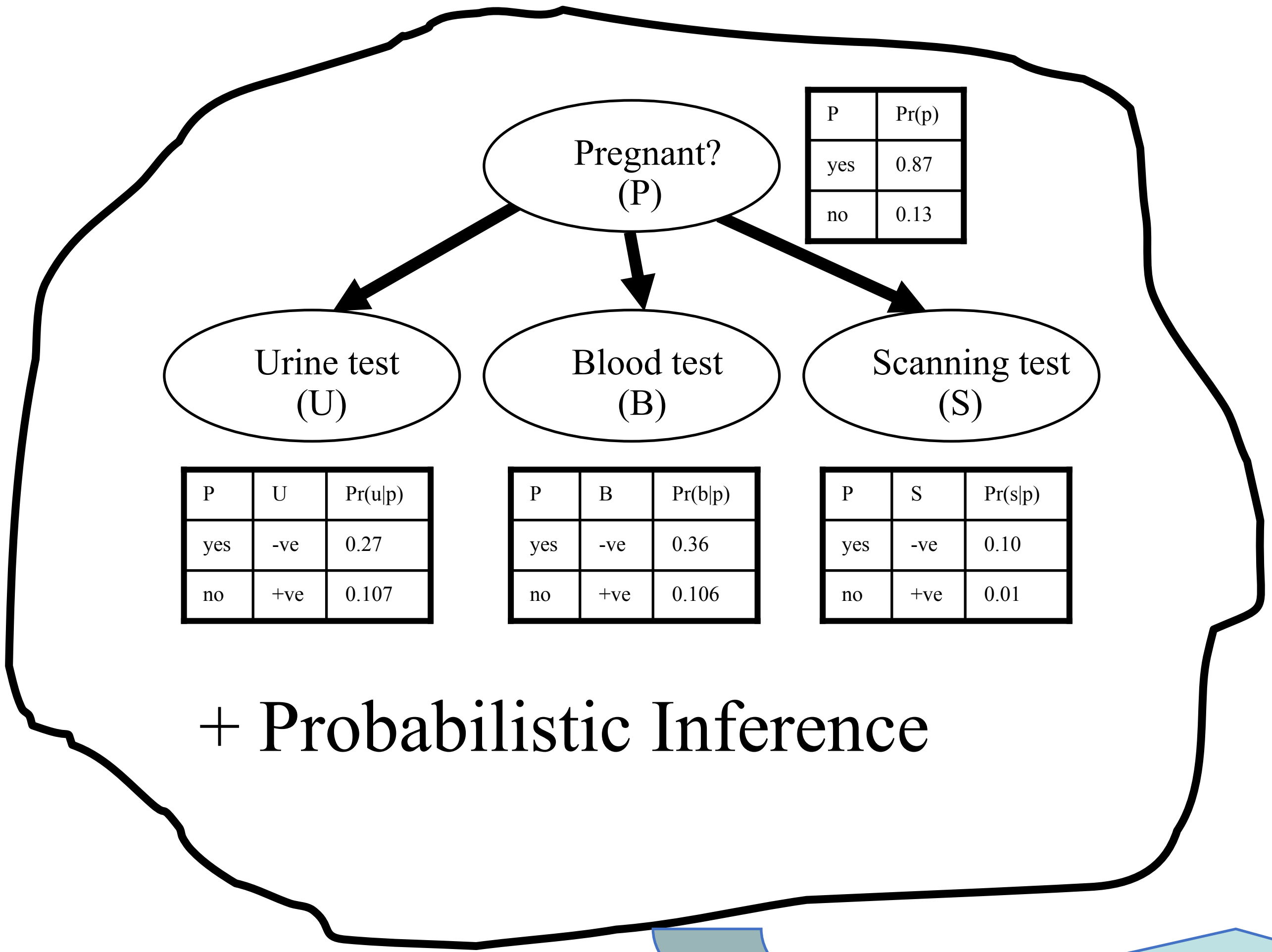
P	U	Pr(u|p)
yes	-ve	0.27
no	+ve	0.107

P	B	Pr(b|p)
yes	-ve	0.36
no	+ve	0.106

P	S	Pr(s|p)
yes	-ve	0.10
no	+ve	0.01



From Numbers to Decisions

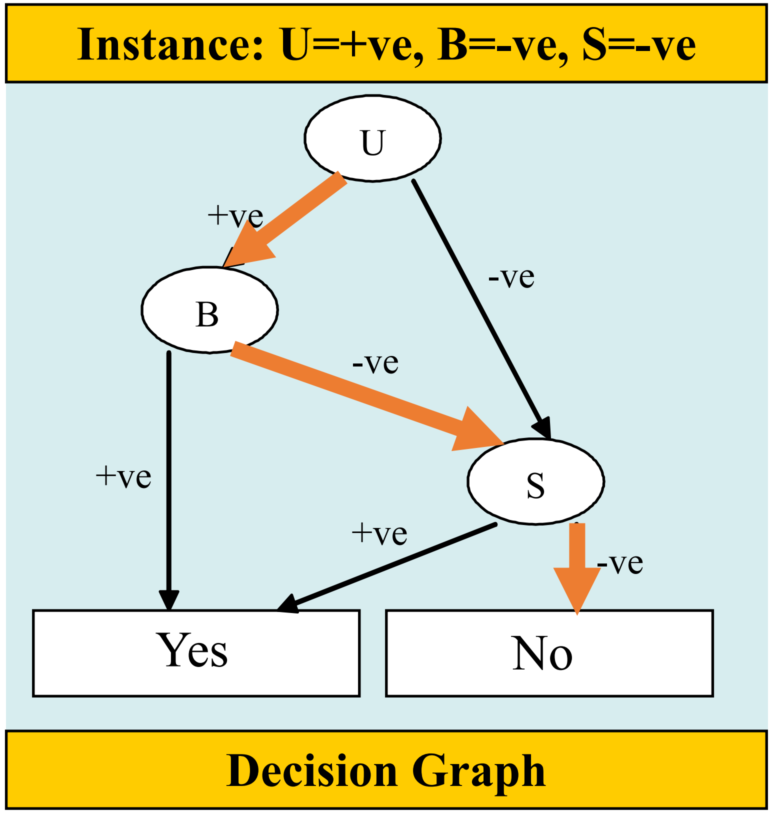


P	Pr(p)
yes	0.87
no	0.13

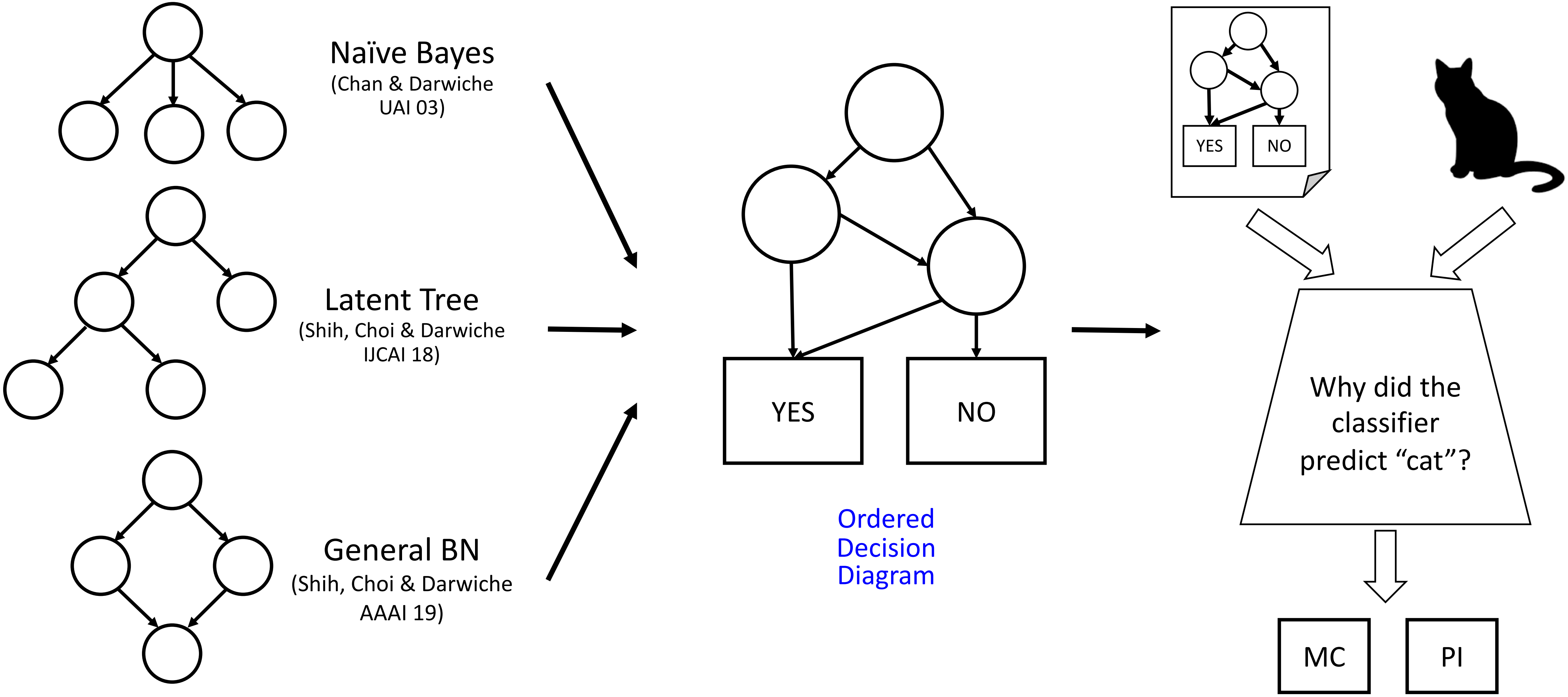
P	U	Pr(u|p)
yes	-ve	0.27
no	+ve	0.107

P	B	Pr(b|p)
yes	-ve	0.36
no	+ve	0.106

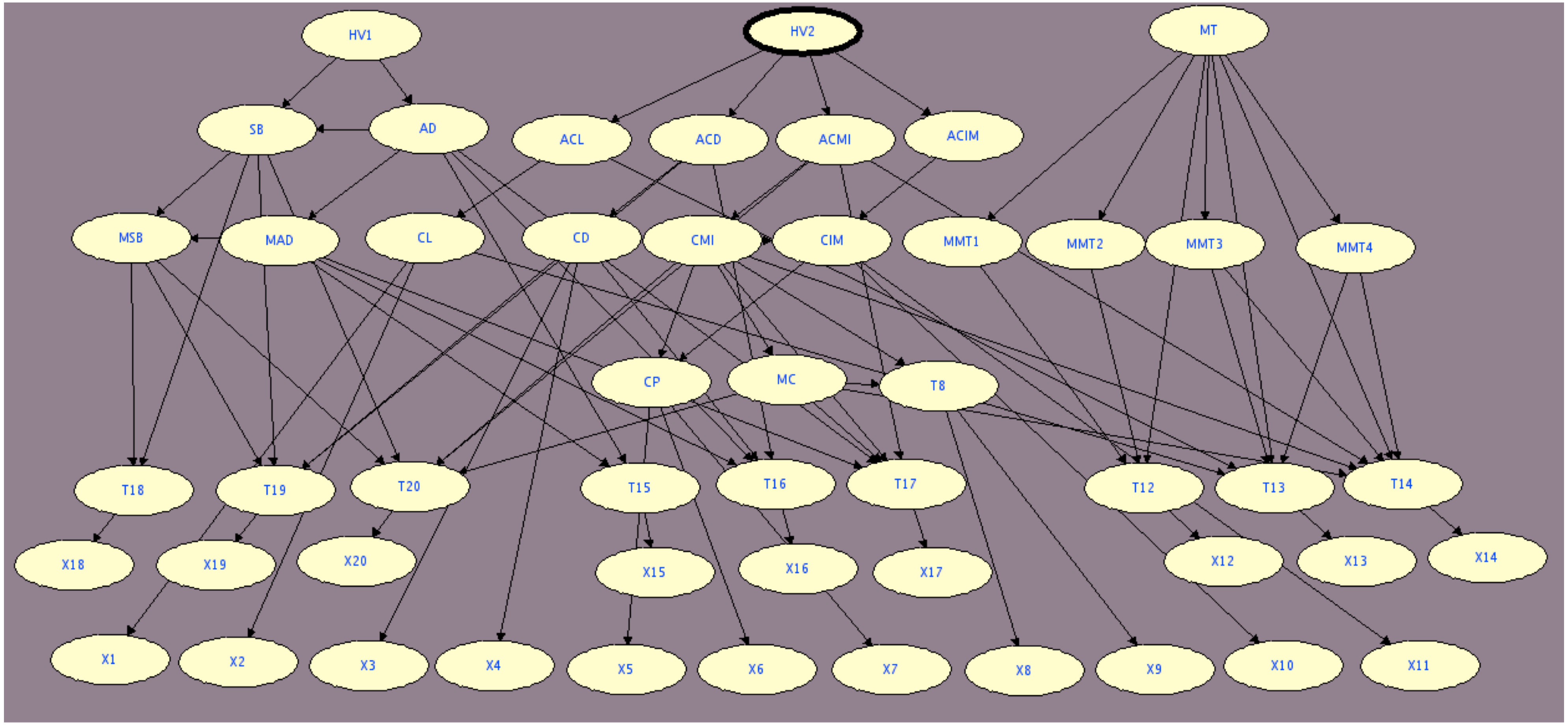
P	S	Pr(s|p)
yes	-ve	0.10
no	+ve	0.01



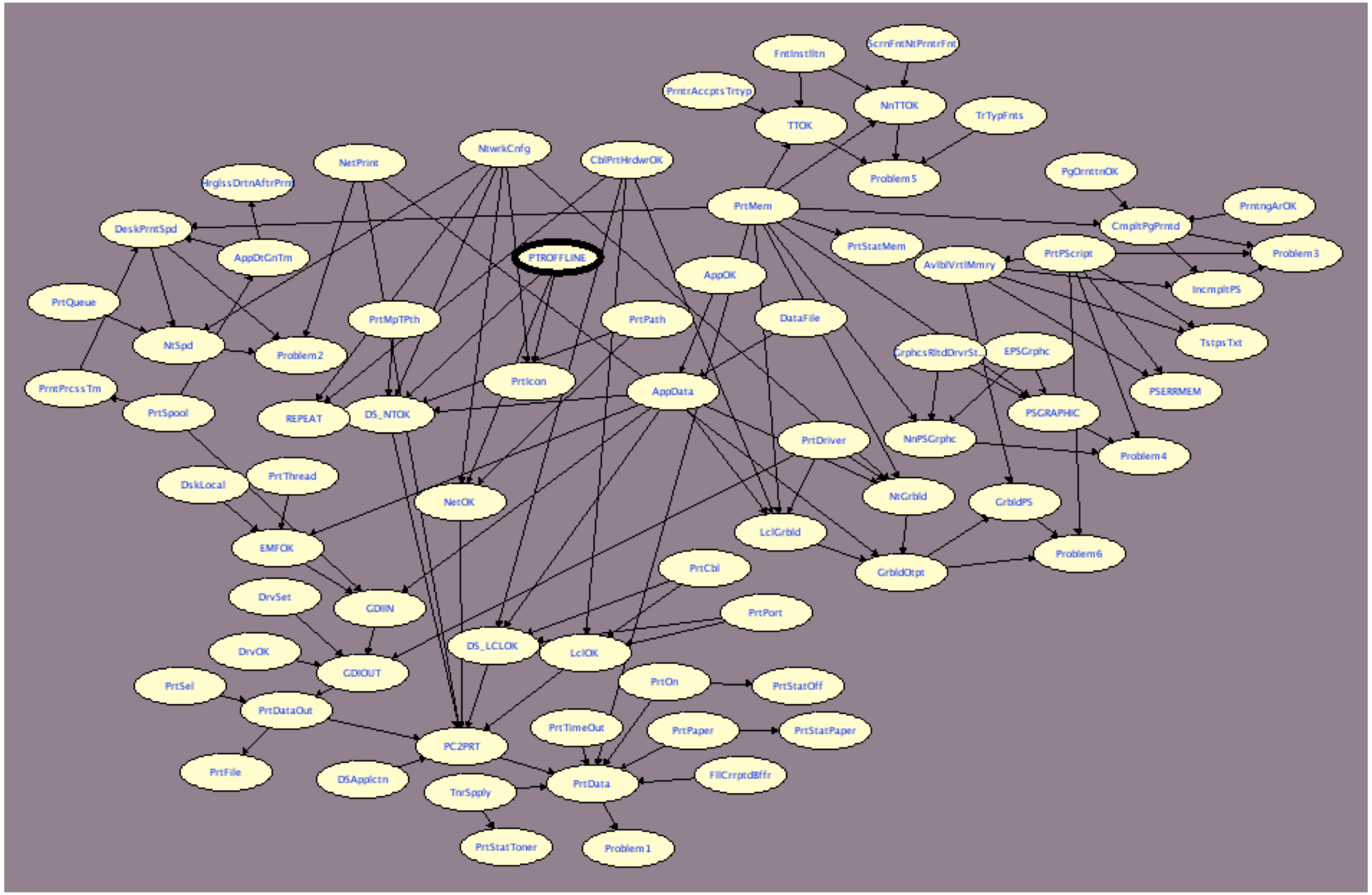
Compiling Classifiers



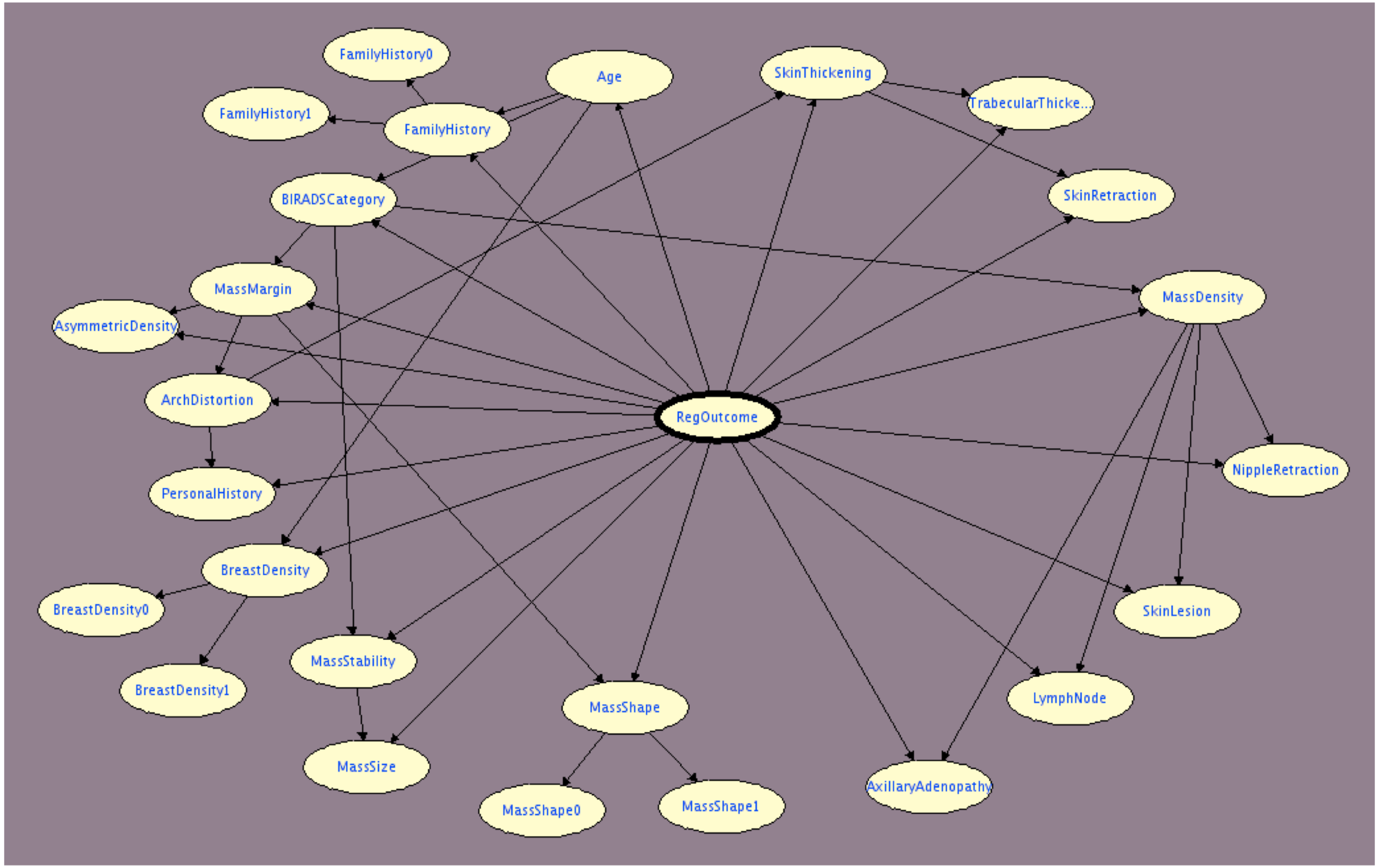
Education Network



Printing Diagnosis Network

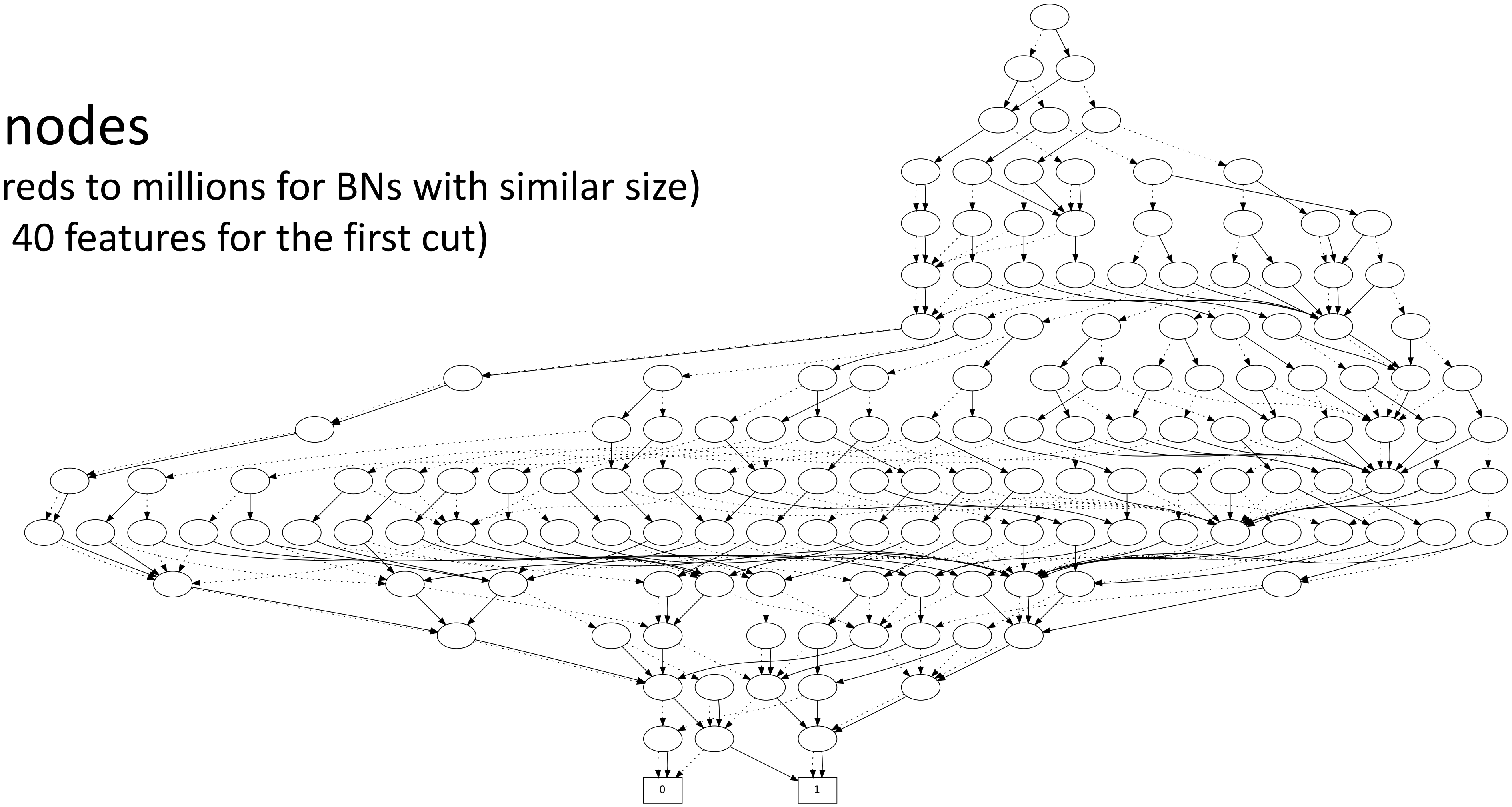


Cancer Network



Cancer Decision Graph

156 nodes
(hundreds to millions for BNs with similar size)
(up to 40 features for the first cut)



Size of Decision Diagrams

Name	# Vars	Root	Threshold	ODD Size	Time (s)
Mammography	15	Reg-Outcome	0.02	156	7
Win95pts	16	Printer-Offline	0.50	291	21
Immex	17	Not-Understand	0.50	115	9
Adaptive-Testing	20	HV1	0.50	1164	40
Mooring	22	Environment	0.75	12840	938
Andes	24	Try-Kinematics	0.50	47	11708
Math-Skills	46	S6	0.50	3693629	61088

Explaining

Shih, Choi & Darwiche (IJCAI 18)

Given a decision graph, we can explain the classifier’s decisions.

- MC Explanations
  - “Which positive features are responsible for a *yes* decision?”
  - “Which negative features are responsible for a *no* decision?”

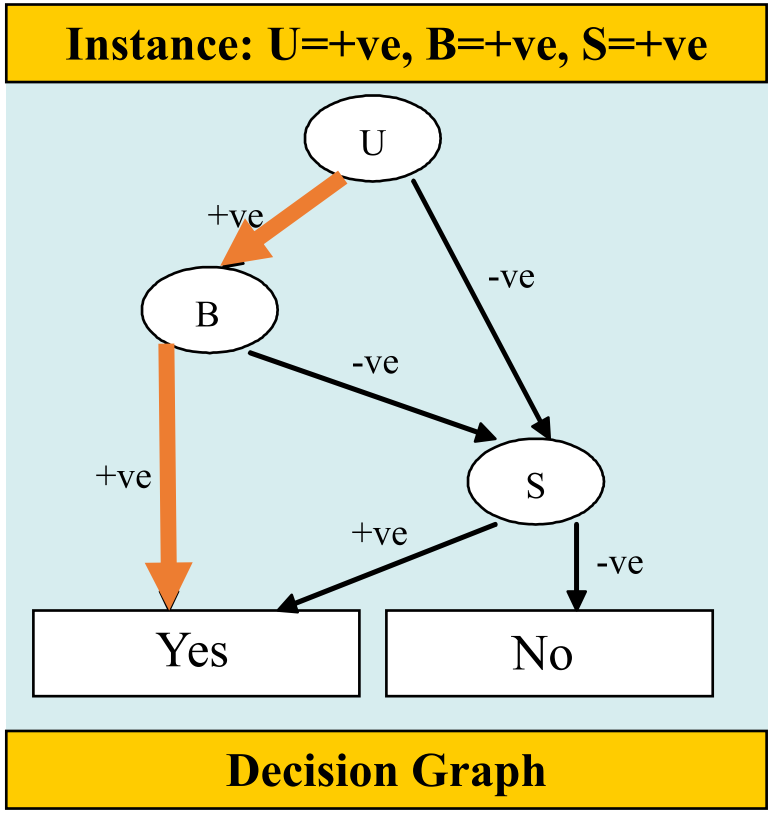
- PI Explanations
  - “Which features (+ or -) make the other features irrelevant?”

Example Explanation

Susan tested positive for **S**canning, **B**lood and **U**rine

Why did you conclude that Susan is pregnant?

Because the Scanning test came out positive



Example Explanation

Susan tested positive for **S**canning, **B**lood and **U**rine

Why did you conclude that Susan is pregnant?

Because the Scanning test came out positive

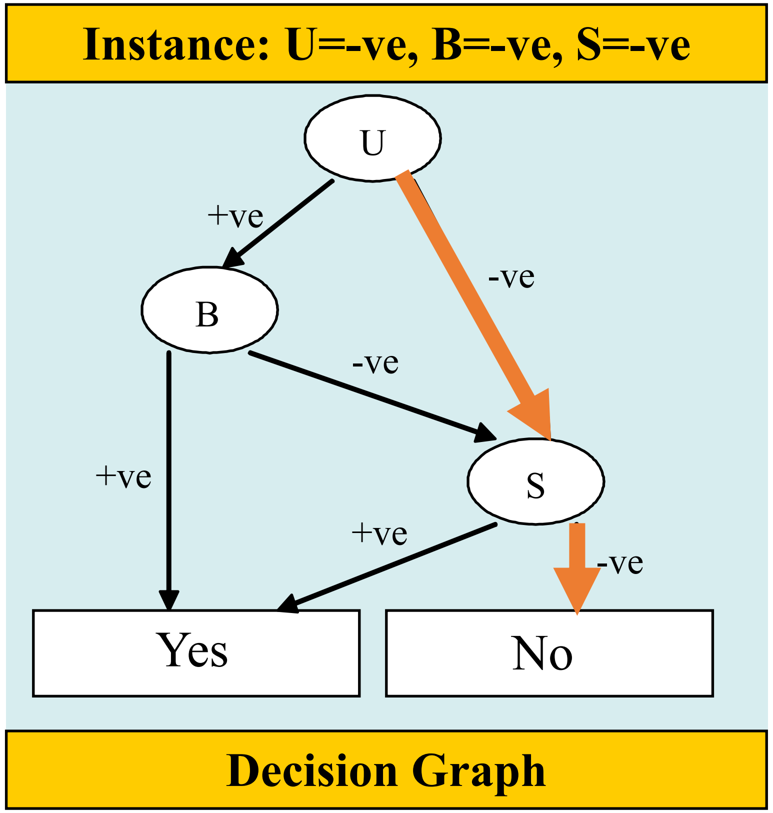
U	B	S		**Explanations**
+ve	+ve	+ve	**Yes**	**(-,-,+)**
+ve	+ve	-ve	**Yes**	**(+,+,-)**
+ve	-ve	+ve	**Yes**	**(-,-,+)**
+ve	-ve	-ve	**No**	**(+,-,-)**
-ve	+ve	+ve	**Yes**	**(-,-,+)**
-ve	+ve	-ve	**No**	**(-,+,-)**
-ve	-ve	+ve	**Yes**	**(-,-,+)**
-ve	-ve	-ve	**No**	**(+,-,-), (-,+,-)**

Example Explanation

Sally tested negative for **S**canning, **B**lood and **U**rine

Why did you conclude that Sally is not pregnant?

Because the Scanning test, and one of the Blood and Urine tests came out negative



Explanations can be computed in linear time

Example Explanation

Sally tested negative for **S**canning, **B**lood and **U**rine

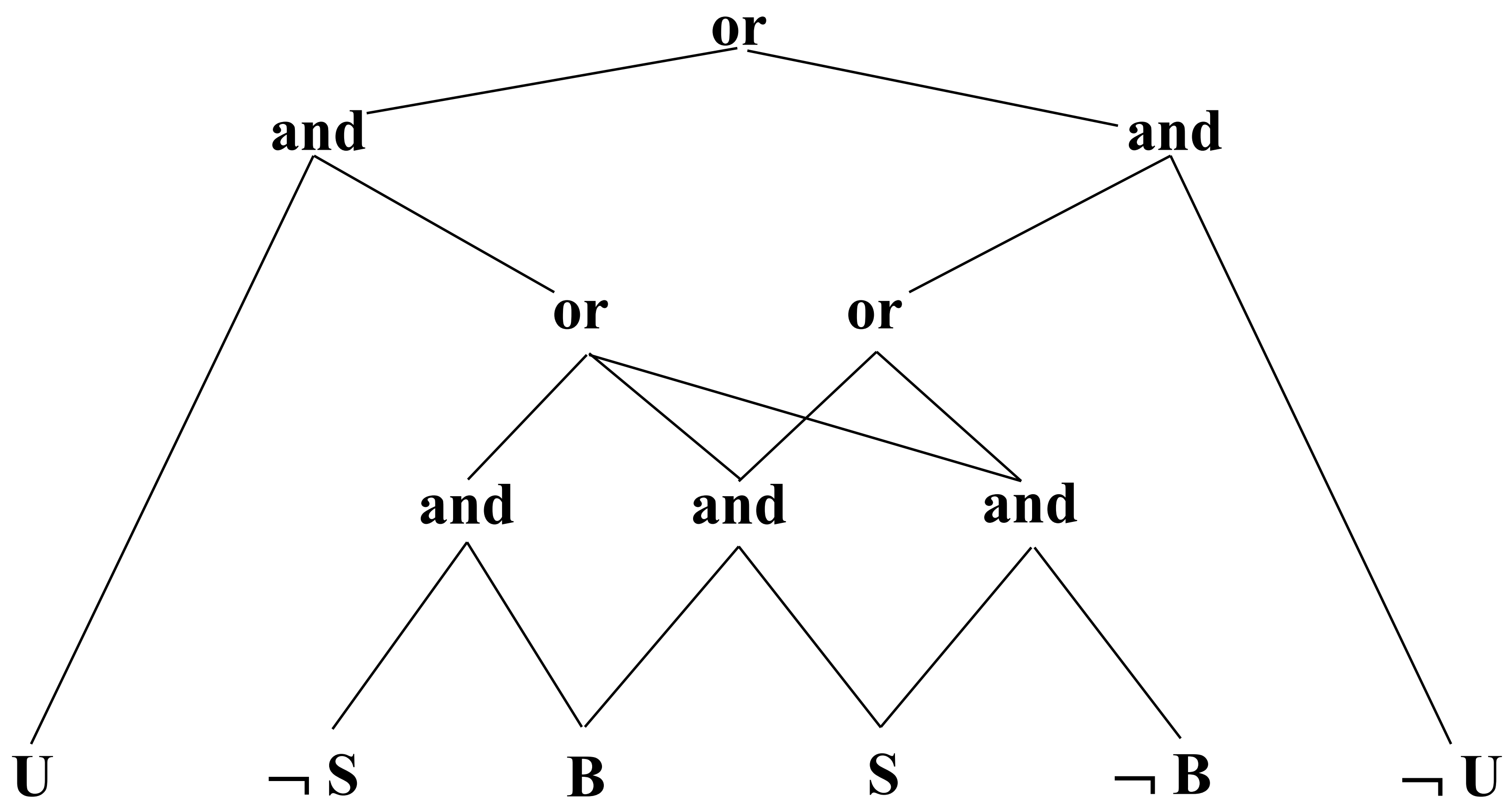
Why did you conclude that Sally is not pregnant?

Because the Scanning test, and one of the Blood and Urine tests came out negative

U	B	S		**Explanations**
+ve	+ve	+ve	**Yes**	**(-,-,+)**
+ve	+ve	-ve	**Yes**	**(+,+,-)**
+ve	-ve	+ve	**Yes**	**(-,-,+)**
+ve	-ve	-ve	**No**	(+,-,-)
-ve	+ve	+ve	**Yes**	**(-,-,+)**
-ve	+ve	-ve	**No**	(-,+,-)
-ve	-ve	+ve	**Yes**	**(-,-,+)**
-ve	-ve	-ve	**No**	**(+,-,-), (-,+,-)**

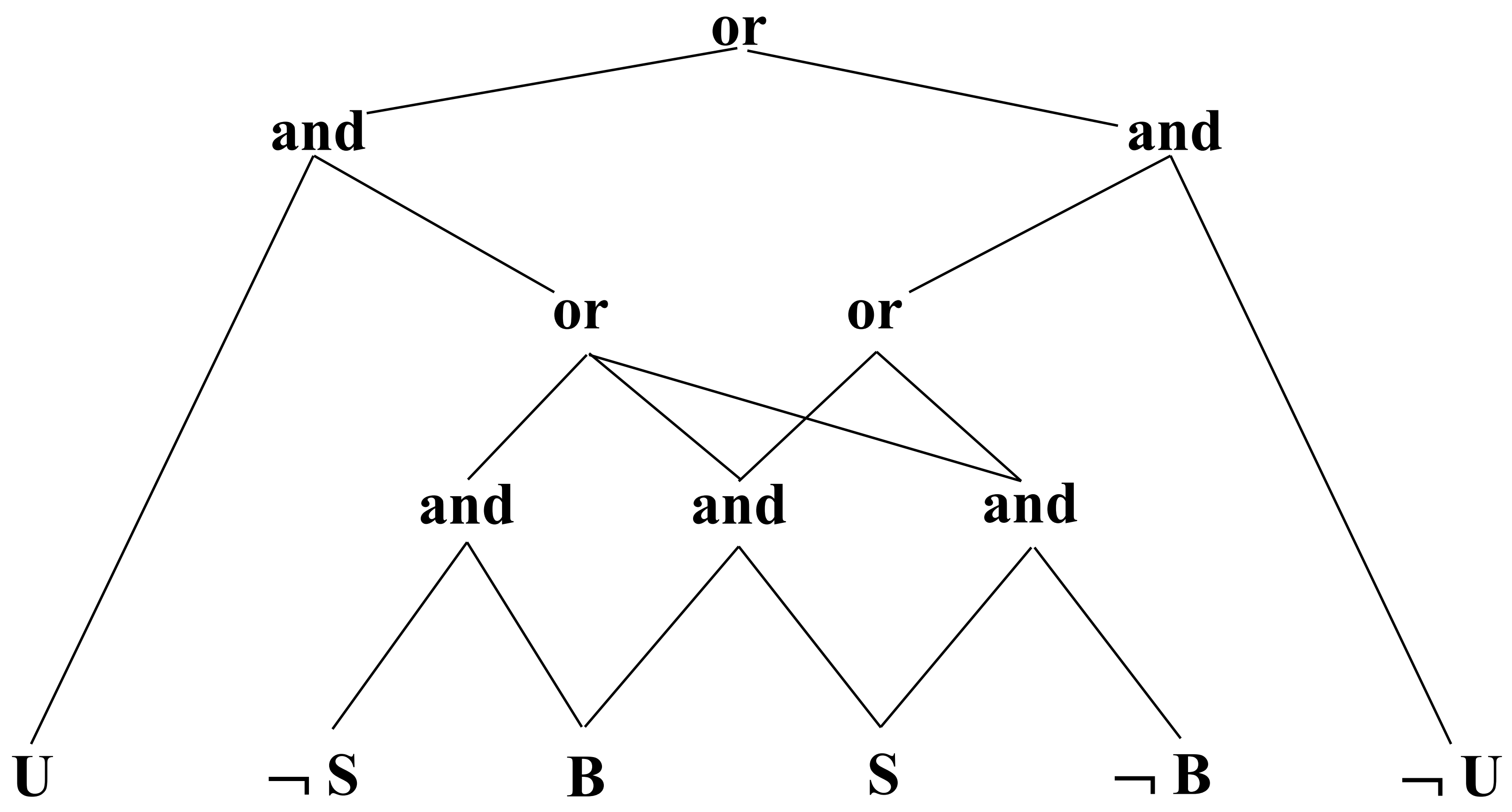
Explaining in Linear Time

positive instance: U,¬B, S



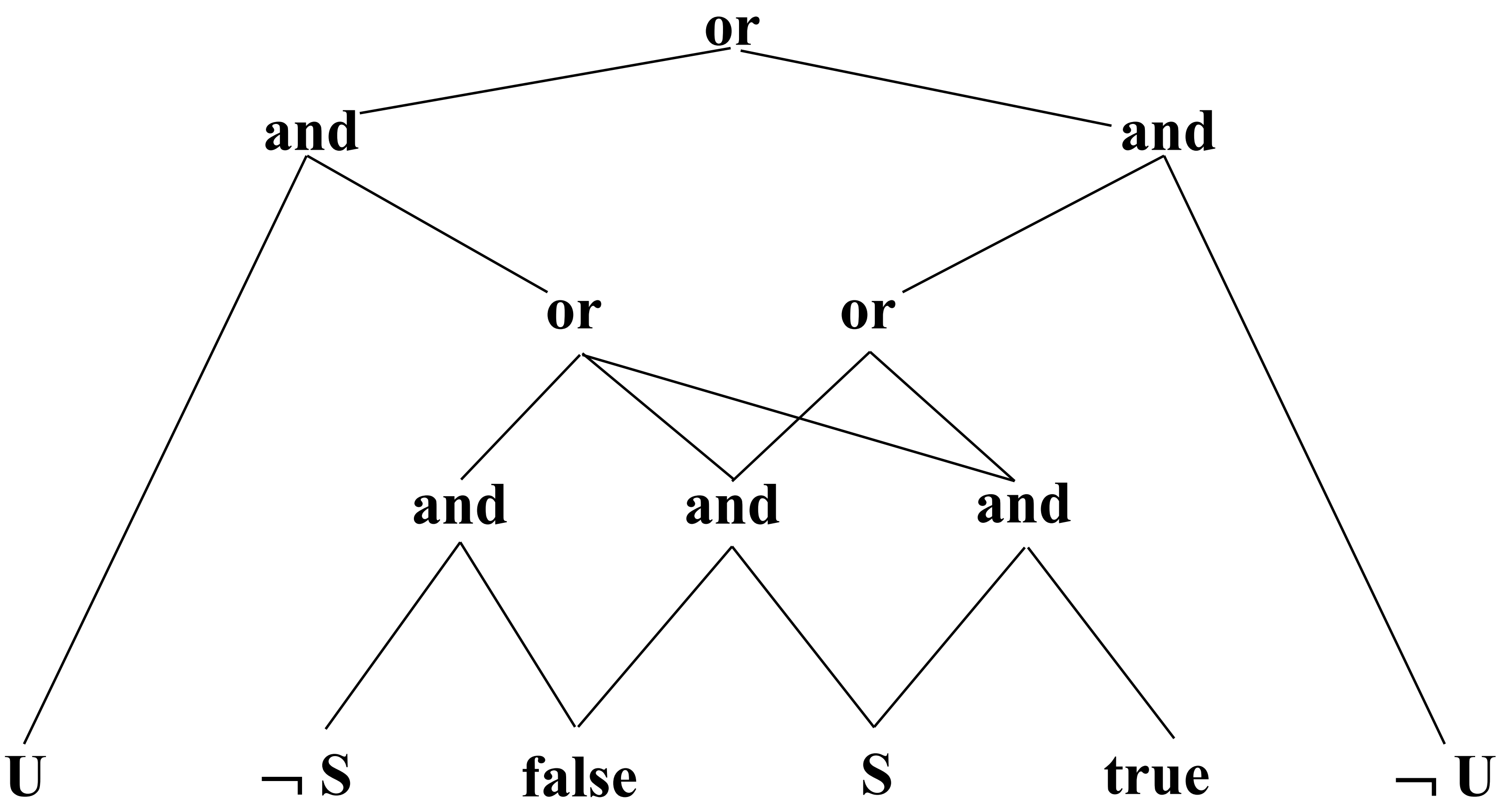
1. Condition on $\neg$B

positive instance: U,¬B, S



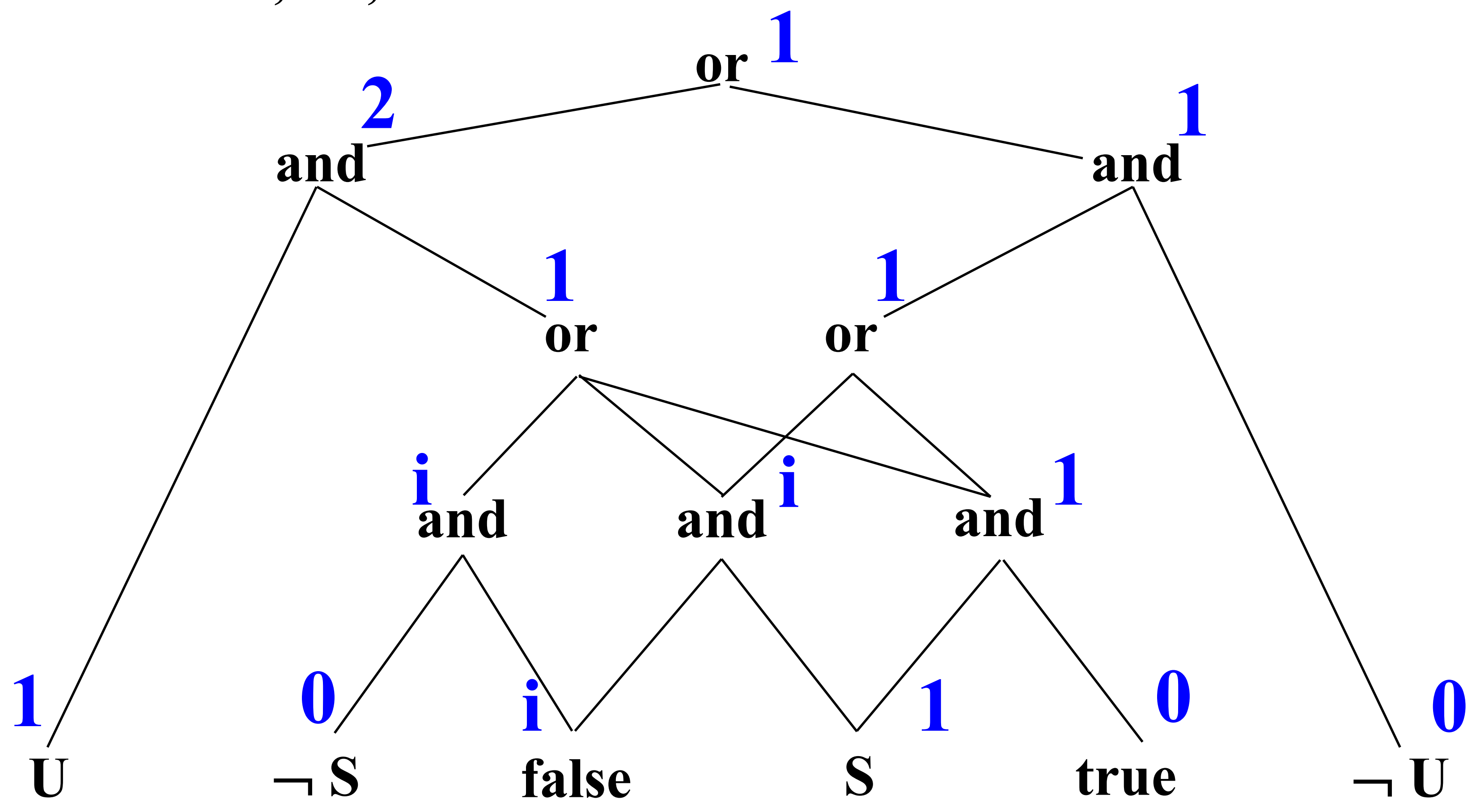
1. Condition on ¬B

positive instance: U, ¬B, S



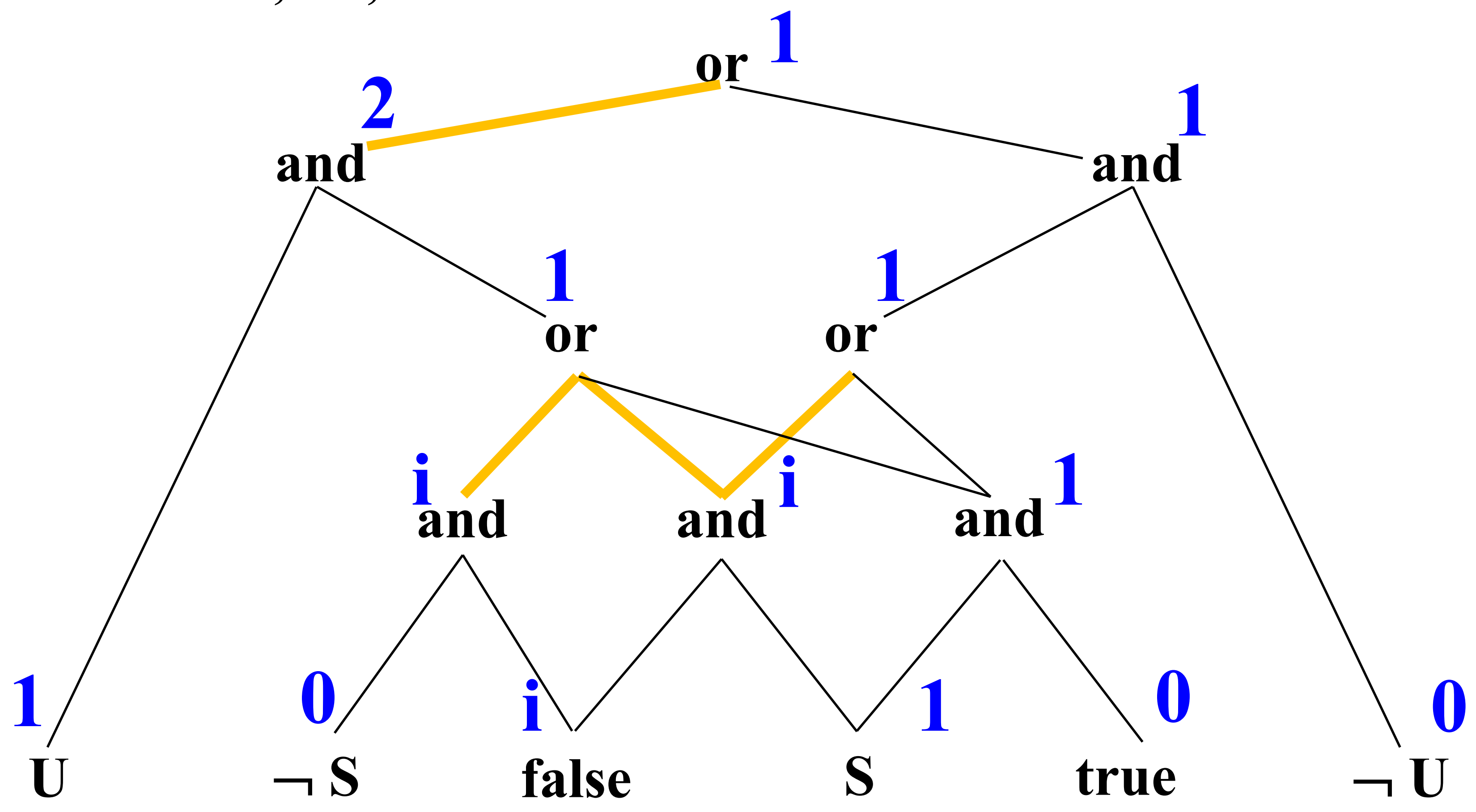
2. Compute Minimum Cardinality

positive instance: U,¬B, S



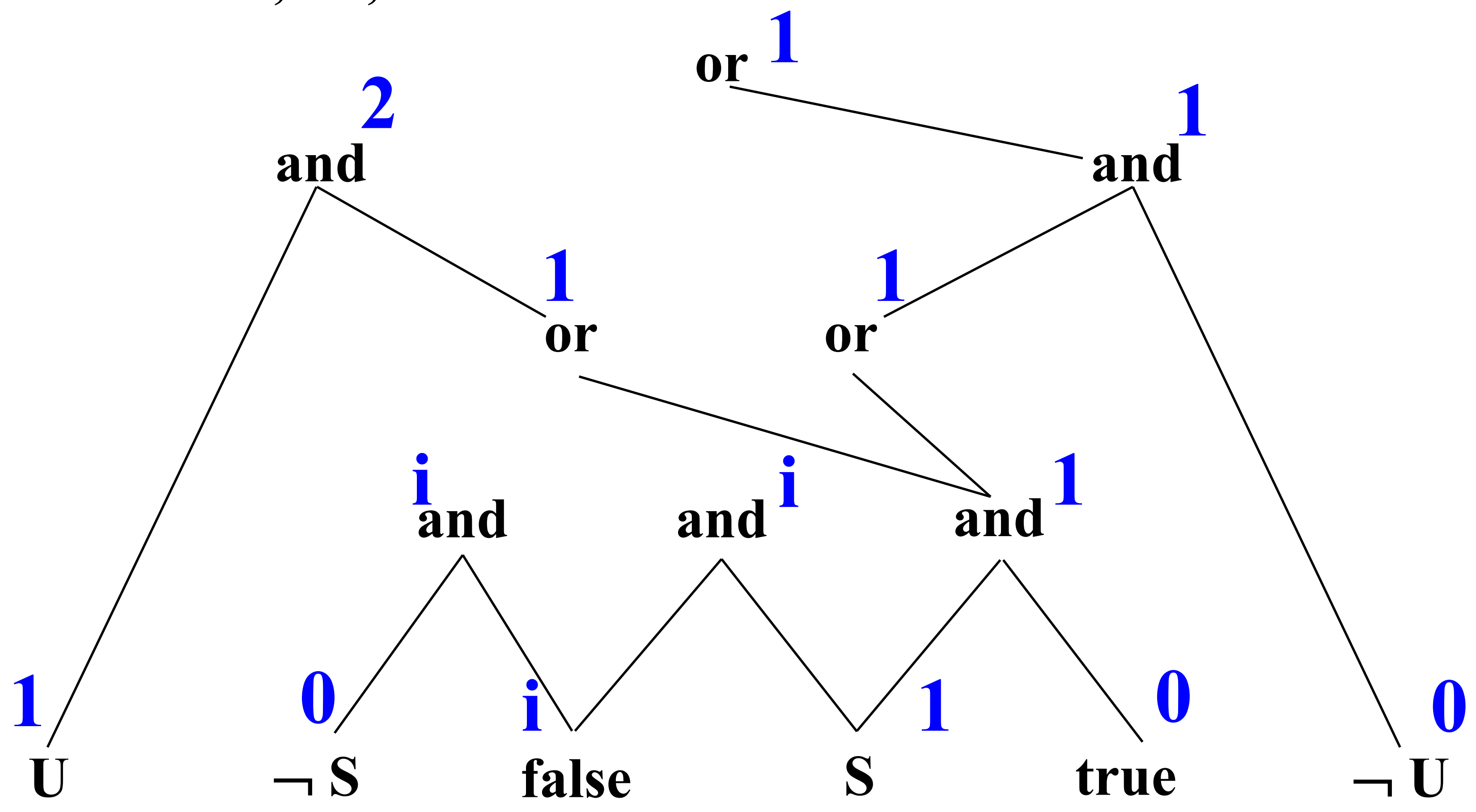
3. Minimize

positive instance: U,¬B, S



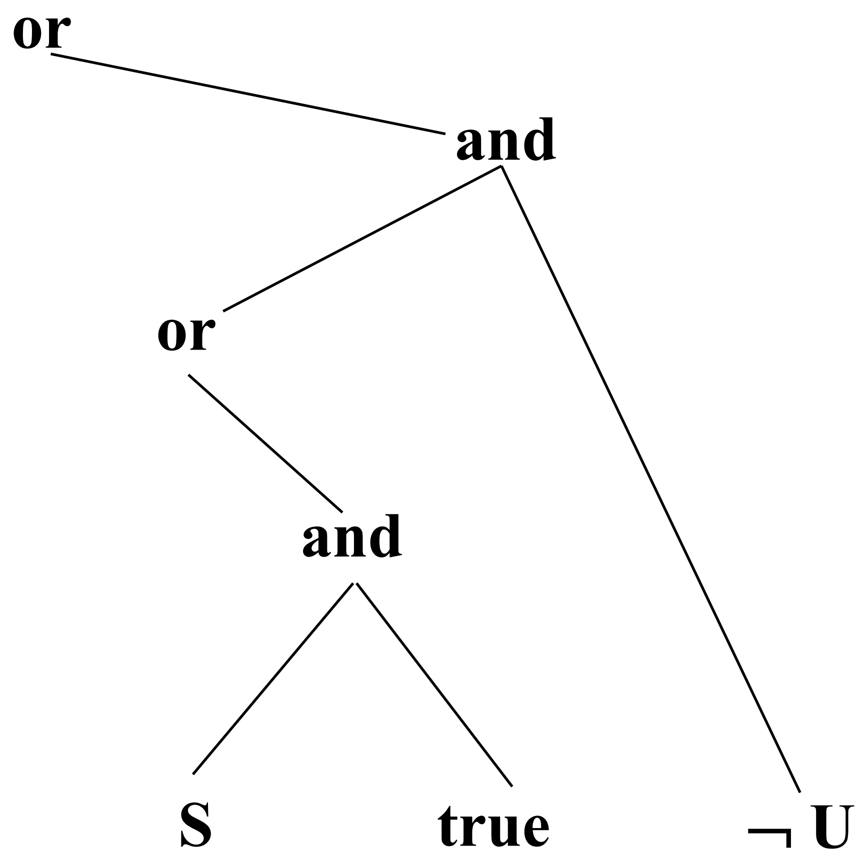
3. Minimize

positive instance: U,¬B, S



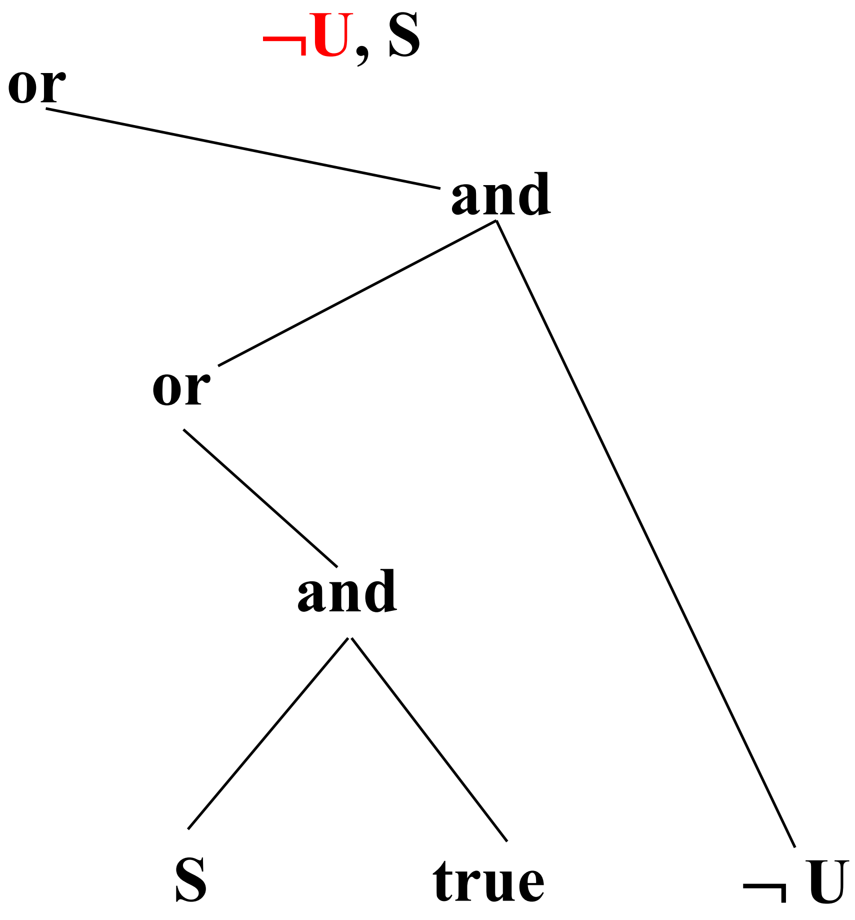
3. Minimize

positive instance: U,¬B, S



4. Enumerate

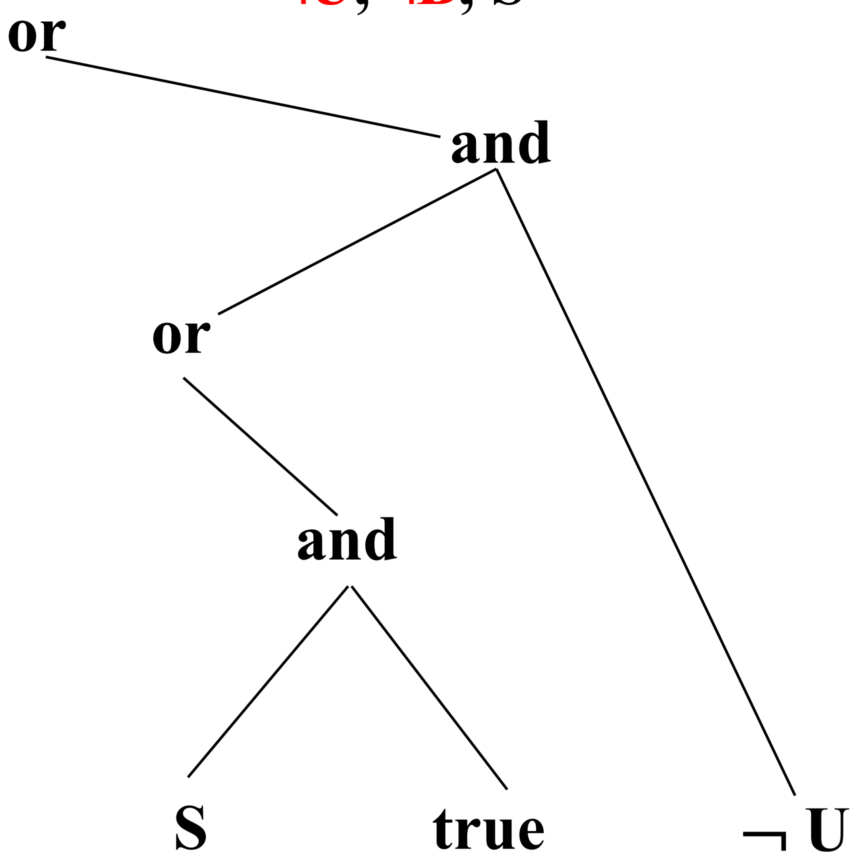
positive instance: U,¬B, S



Explanation

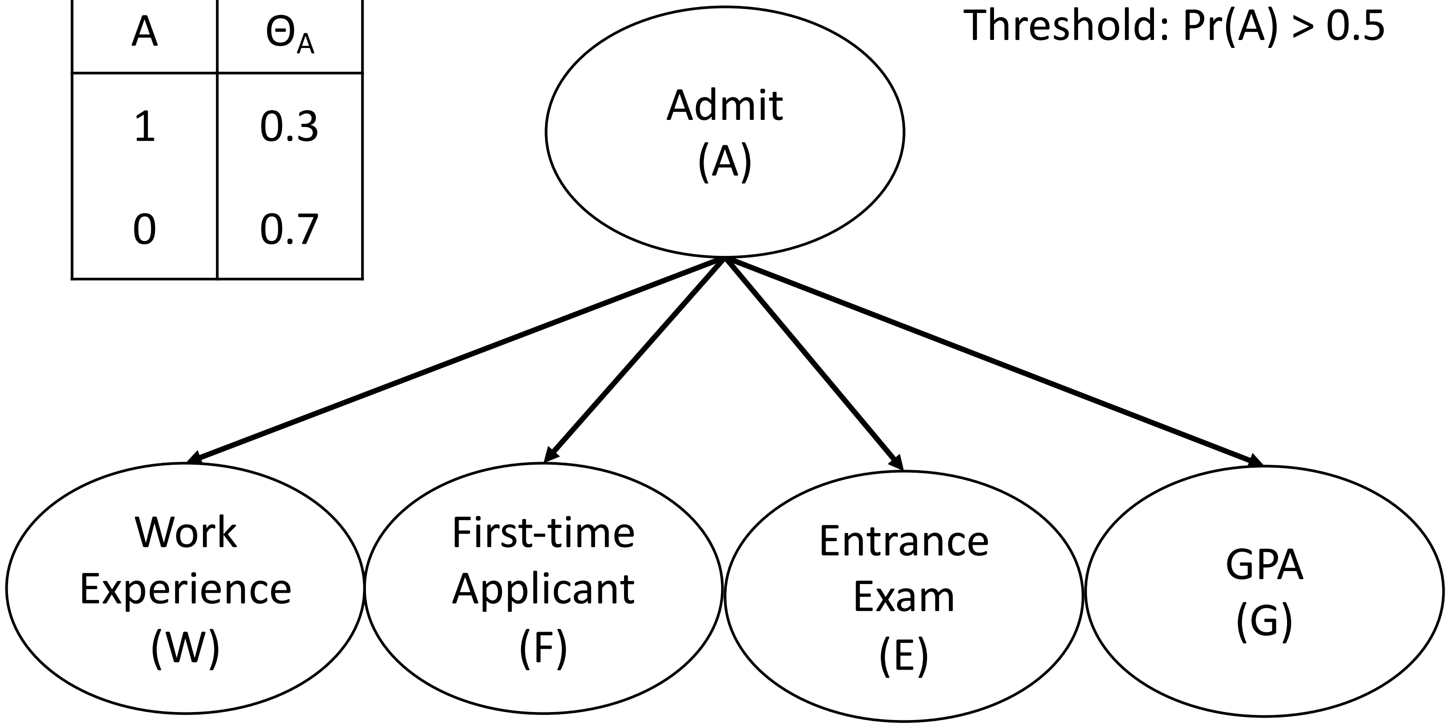
positive instance: U,¬B, S

¬U,¬B, S



Example Classifier

A	Θ_A
1	0.3
0	0.7



Threshold: Pr(A) > 0.5

W	
f_n	0.04
f_p	0.10

F	
f_n	0.30
f_p	0.20

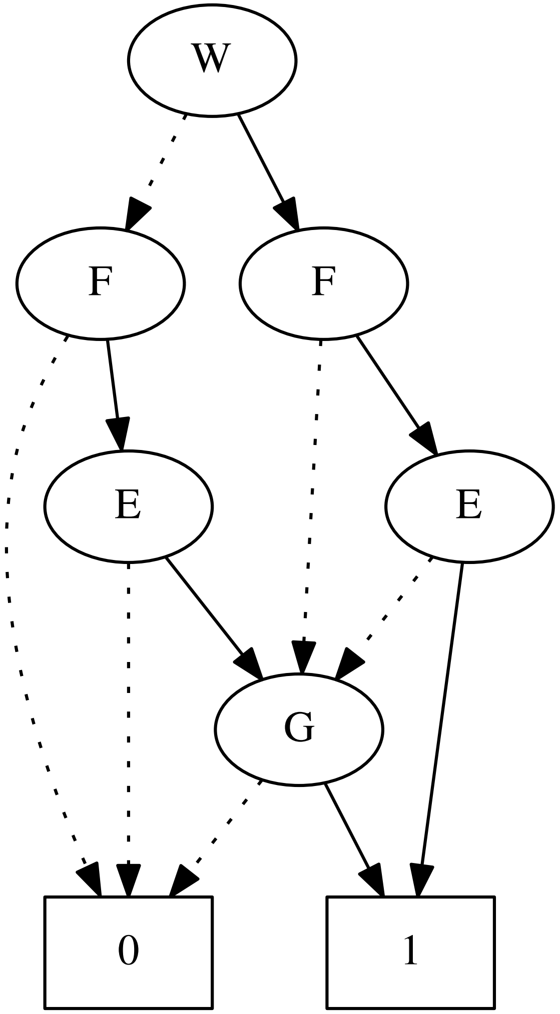
E	
f_n	0.60
f_p	0.15

G	
f_n	0.03
f_p	0.11

Example Explanation

Why did you admit Sally (+,+,+,+)?

Because of her Work Experience and Good GPA



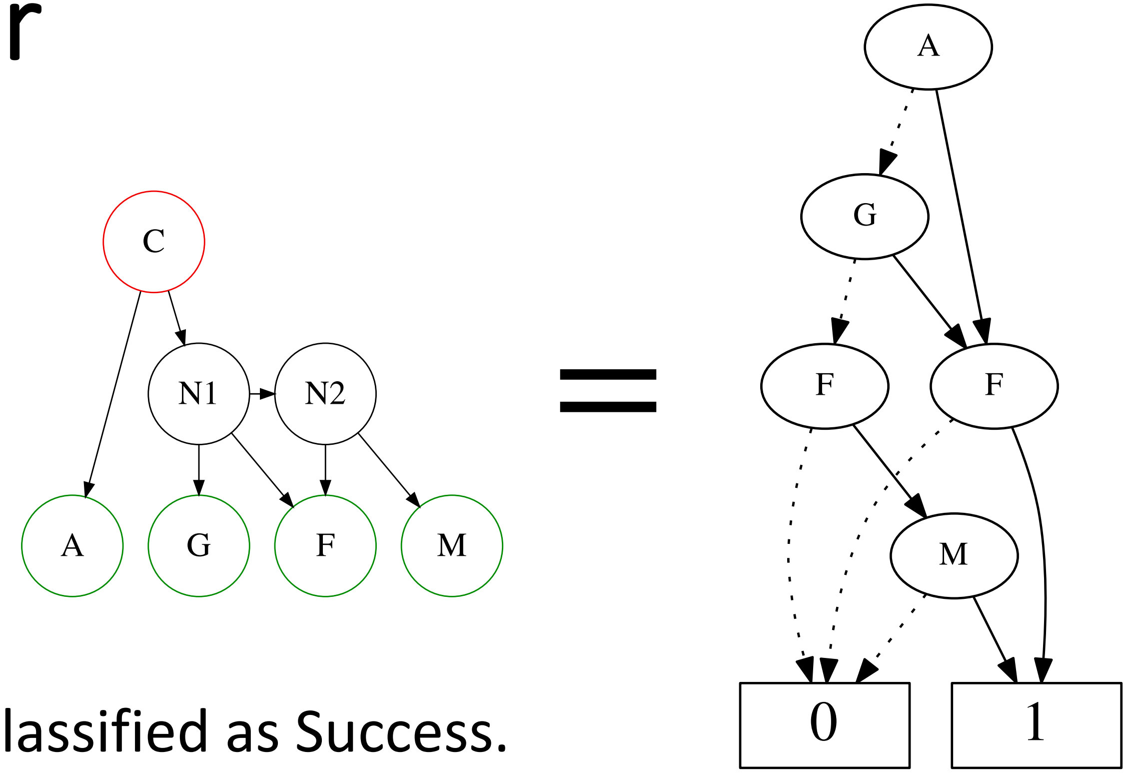
(+,+,+,+)	yes	(+,-,-,+)

Example: Movie Classifier

Class: Box Office Success (1) or Failure (0)

Features:

- A – Adapted Screenplay
- G – Great Cinematography
- F – Famous Cast
- M – Marketing



Consider a movie with {A=1,G=1,F=1,M=1}, classified as Success.

- Why? Because partial instantiation {A=1,F=1} is enough to guarantee Success
- Remaining features do not matter

Classifier is monotonic

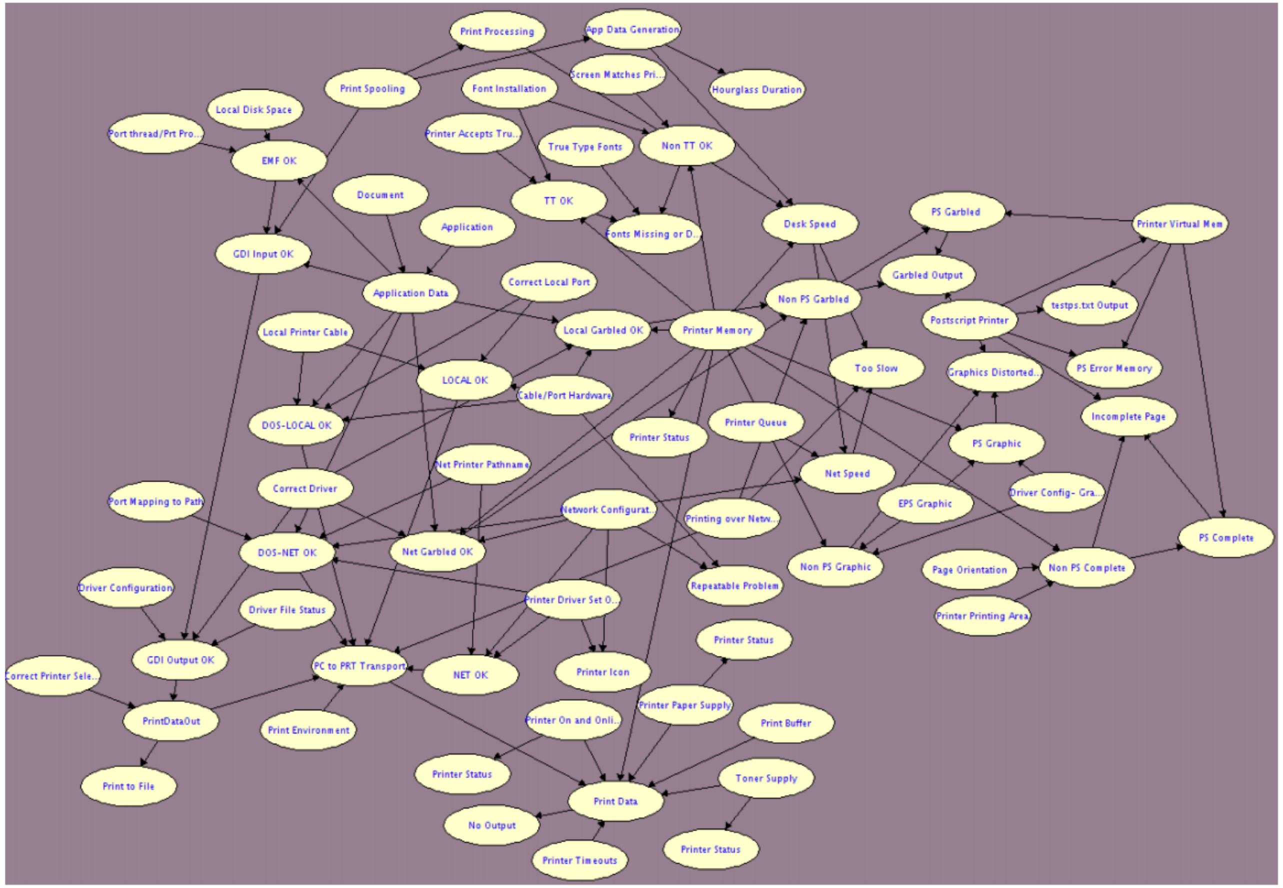
- instance classified as success remains success if features are flipped from 0 to 1.

Case Study: win95pts

Network diagnoses a printing failure

Details

- 76 nodes, treewidth 9
- state space = 65,536
- root: PtrOffline
  - checks if printer driver set is offline
- threshold: 0.5
- # features: 16



Case Study: win95pts

Consider a printer with the symptoms

slow printing, low toner, printer status off, printer icon greyed out, hourglass display too slow, low memory, and ***font missing***

MC: Having the symptom of ***printer icon greyed out*** and <u>*nothing else*</u> leads to a classification of offline

PI: Fixing three symptoms guarantees classification of printer driver set as offline

repeatable problem, printer icon greyed out, no distorted graphics

Verification: Monotone Classifiers

Shih, Choi, Darwiche (PGM 18 + JMLR)

Positive instance remains positive even if we flip some features from – to +.

If (+,-,-,+) is a positive instance, these instances are also positive

- (+,+,-,+)
- (+,-,+,+)
- (+,+,+,+)

Educational Testing:

Susan’s correct answers include Jack’s correct answers

Susan should pass if Jack passed

Verification: Monotone Classifiers

Shih, Choi, Darwiche (PGM 18 + JMLR)

- Educational assessment classifier not monotone
  (threshold ½)

- Cancer classifier not monotone
  (threshold .02 based on BI-RADS assessment scale)
- Two patients, same mammography report except for personal history.
  - One with personal history → Benign
  - One with no personal history → Malignant

- Classification robustness, …

Reasoning about the Behavior of AI Systems

Keynote at IJCAI-19

Current Focus:

Explanation, Robustness, Verification

NN, Graphical Models, Random Forests

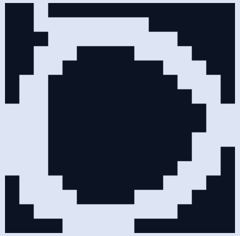
How?

Compile into tractable circuits that make same decisions

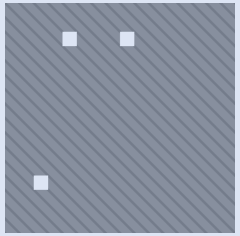
A wealth of AI and CS tools become immediately relevant
(e.g., knowledge compilation and formal verification)

CNN (0-vs-1 CNN) with 98.74% accuracy

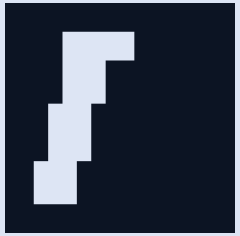
Why 0?



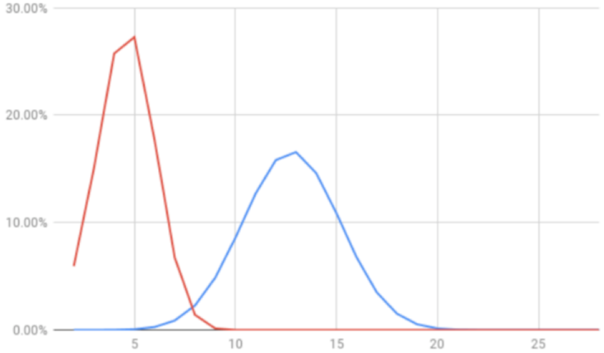
Because



I can fool you!



two CNNs with almost same accuracy
one is significantly more **robust**
plots for 2^{256} instances!



integrate robustness with accuracy
(perhaps into loss function)

Choi, Shi, Shih, Darwiche VNN 19

- New role for symbolic AI & CS methods: Reason about what was learned
- Systems 1/2 (thinking fast and slow), reflection, meta-reasoning

Thank You