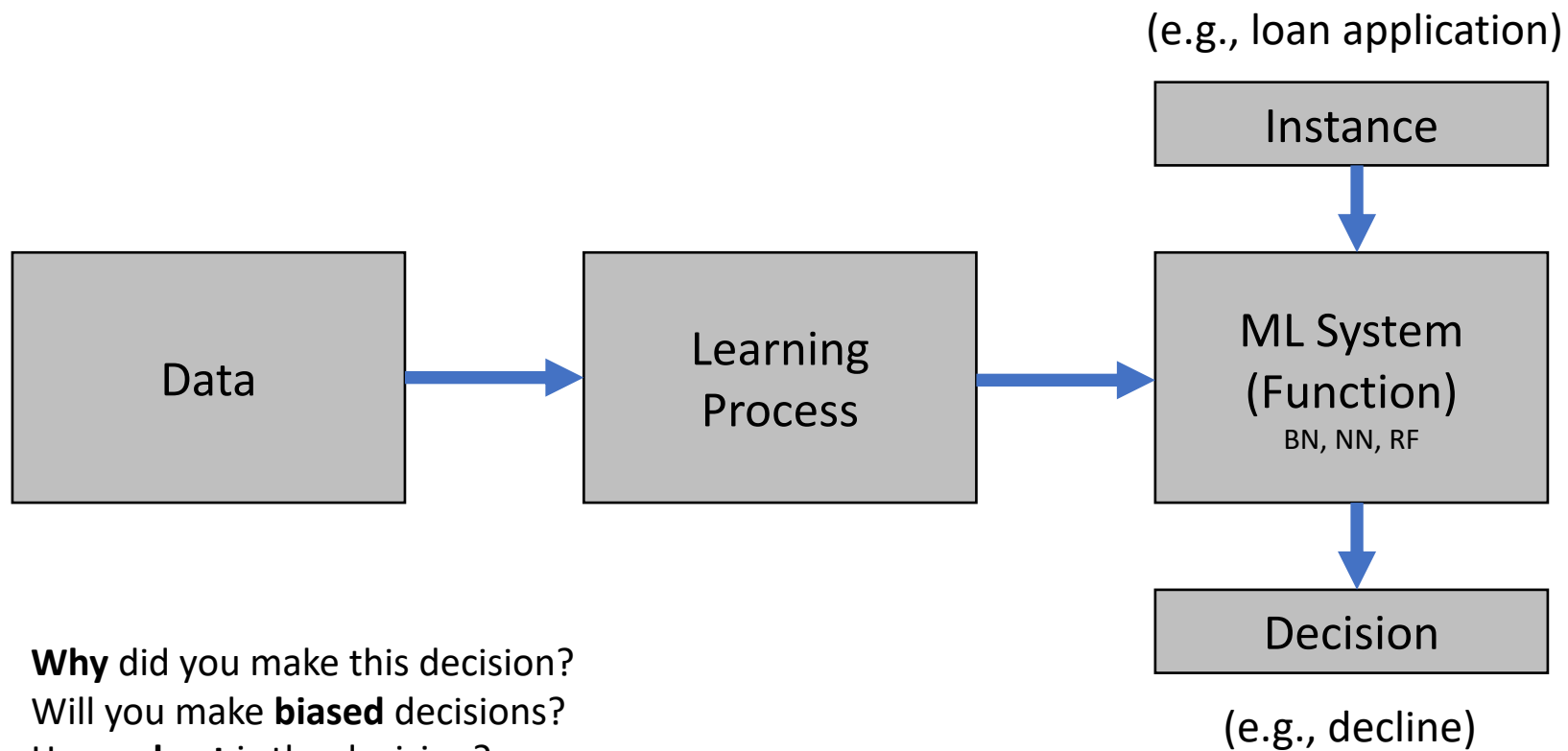


Reasoning About What Was Learned

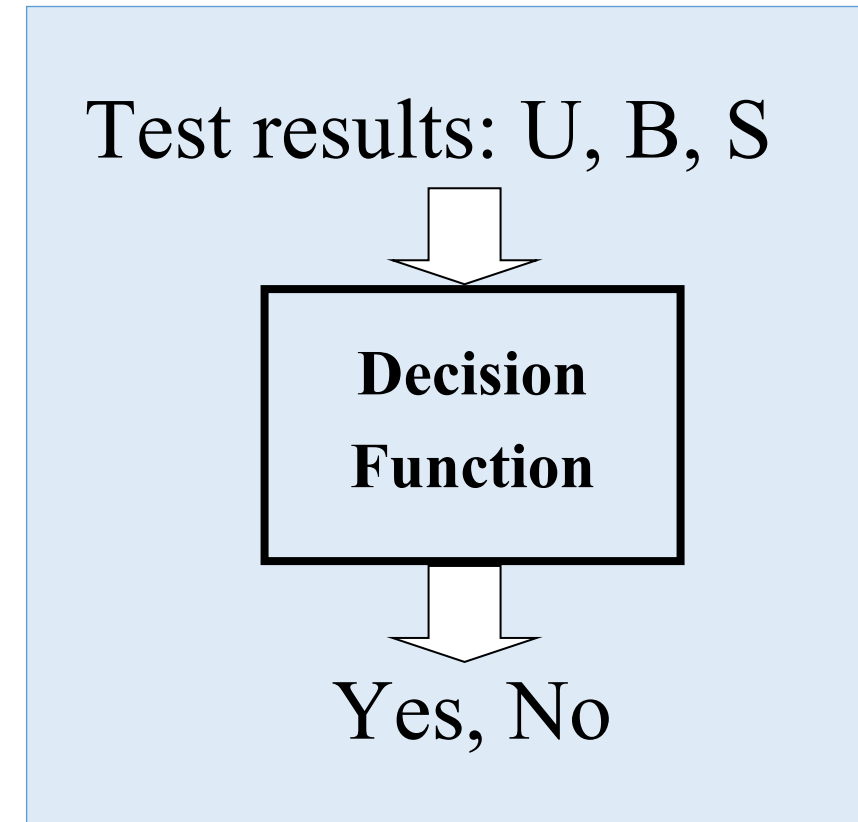
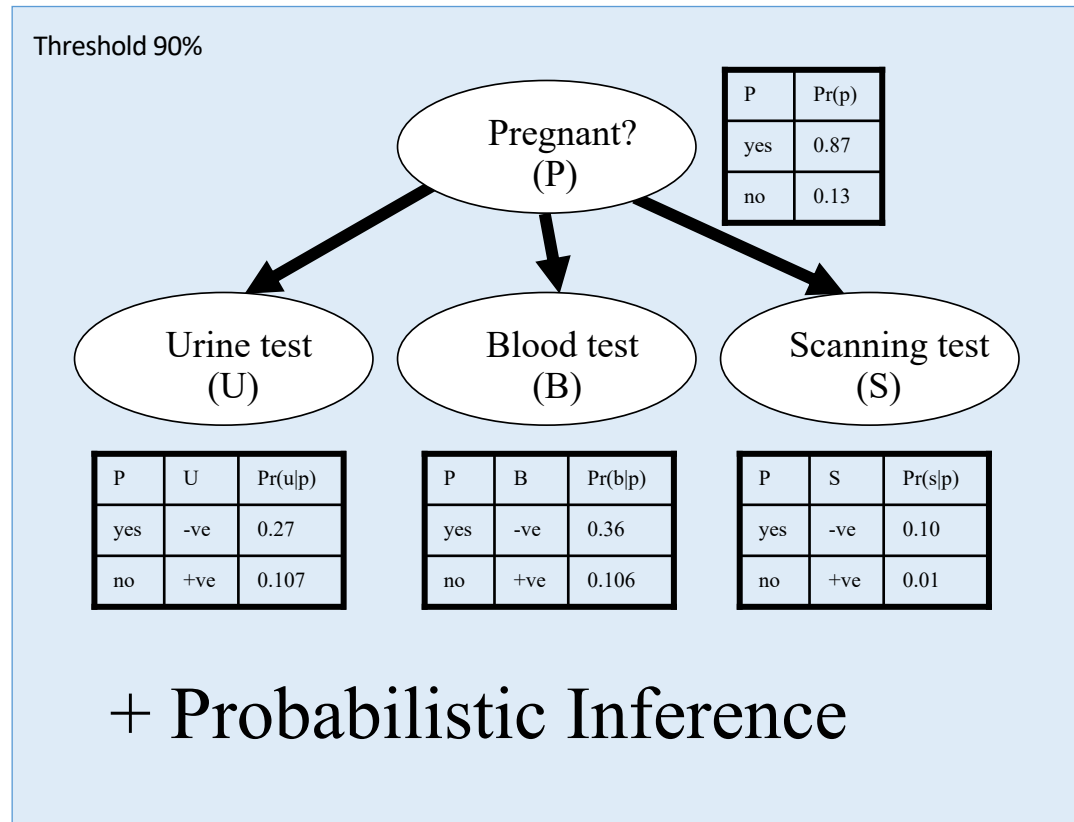
Adnan Darwiche
Computer Science Department
UCLA

ICLA 2021

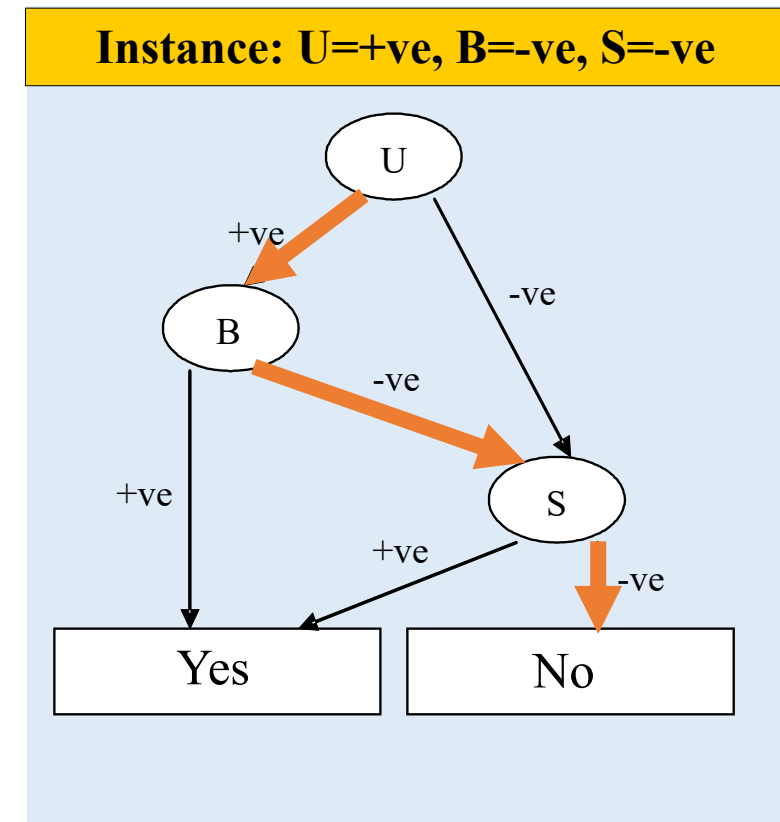
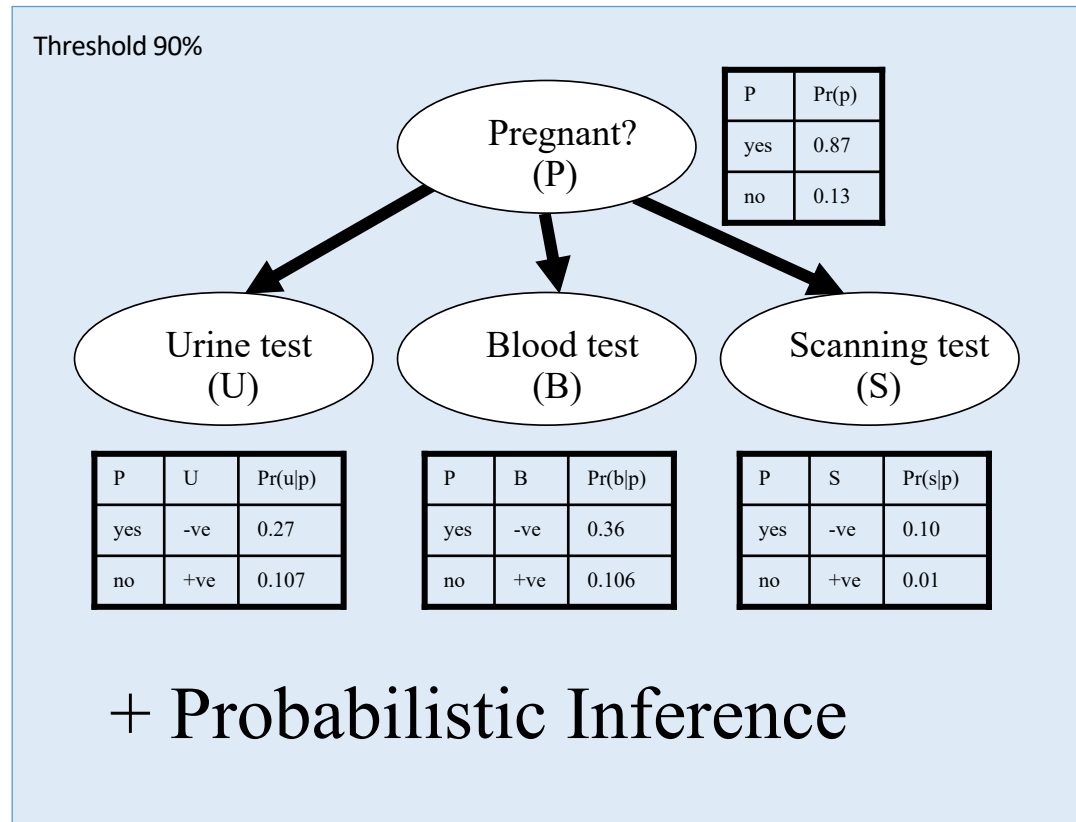


Why did you make this decision?
Will you make **biased** decisions?
How **robust** is the decision?
...

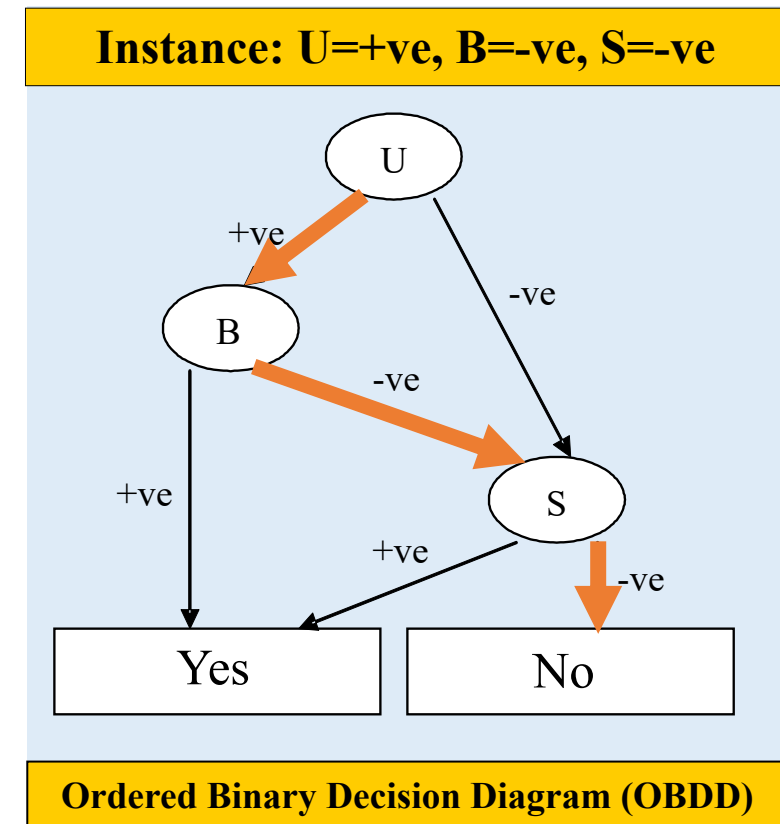
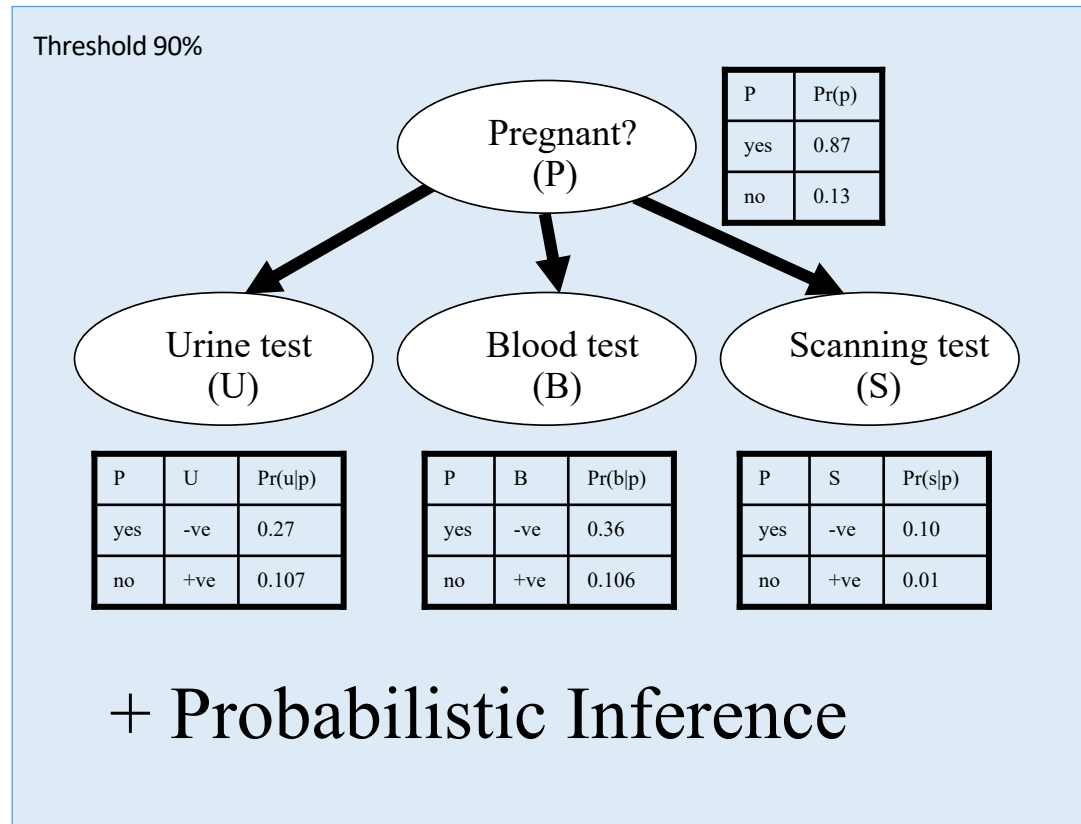
ML Systems as Discrete Functions



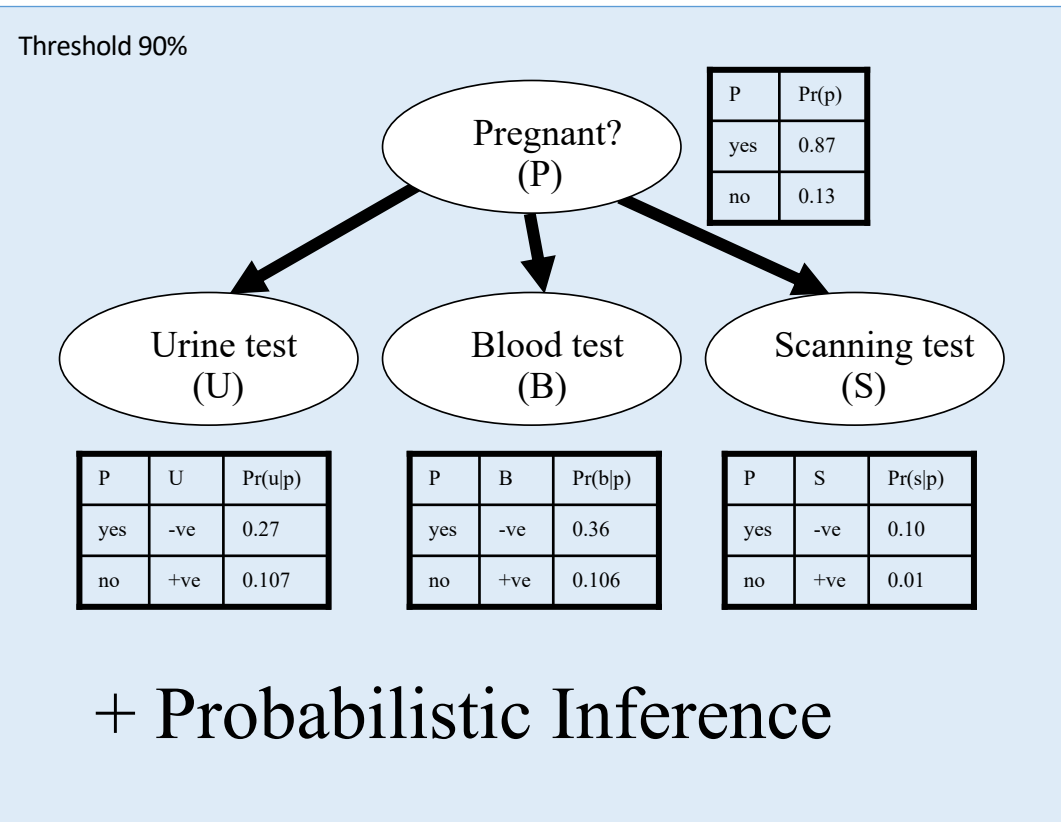
ML Systems as Discrete Functions



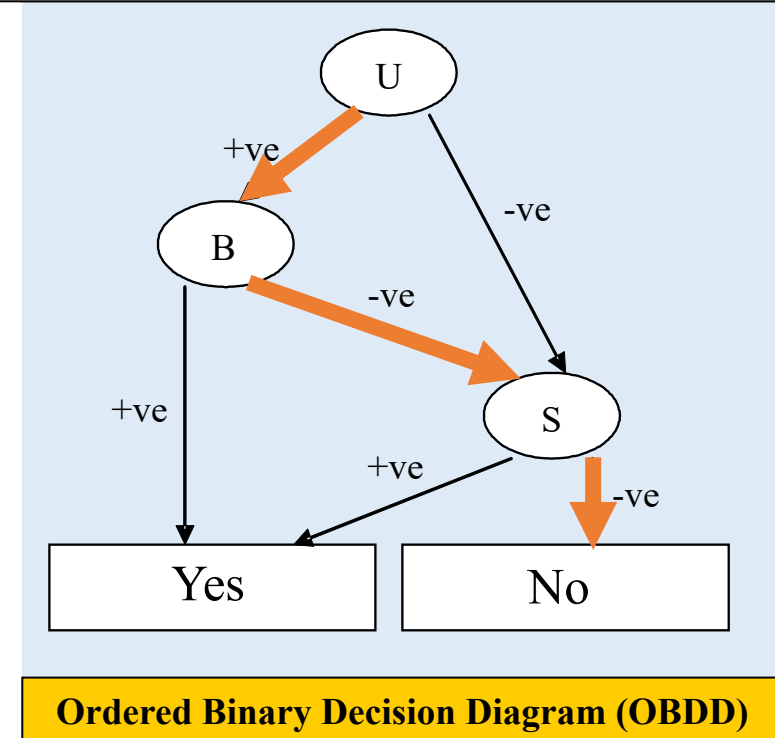
ML Systems as Discrete Functions



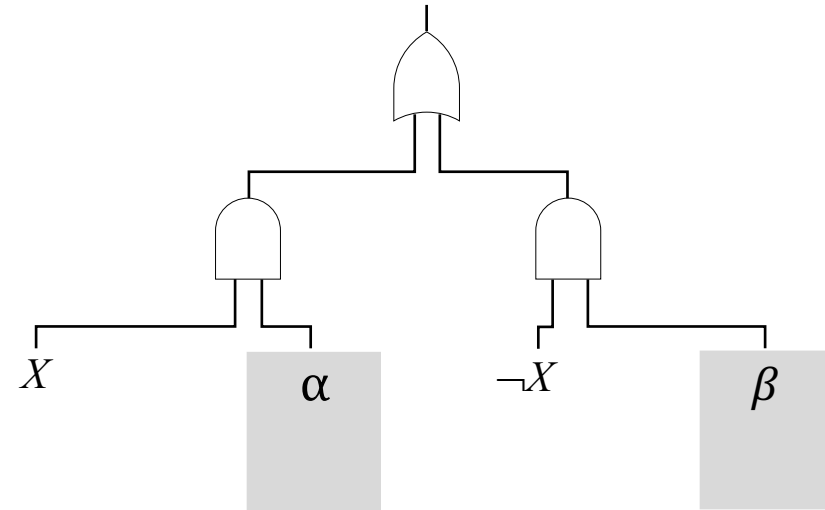
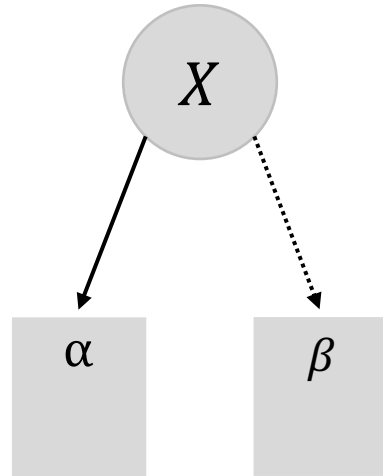
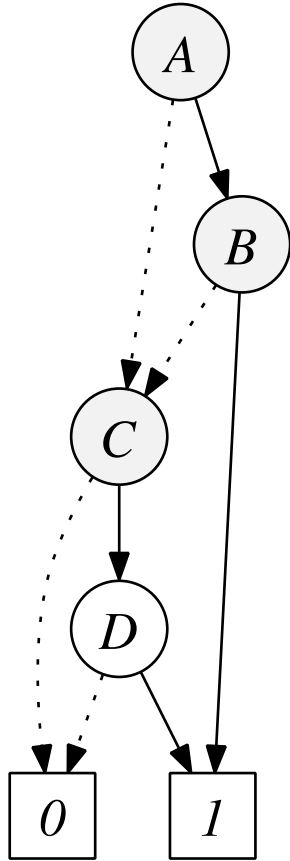
ML Systems as Discrete Functions



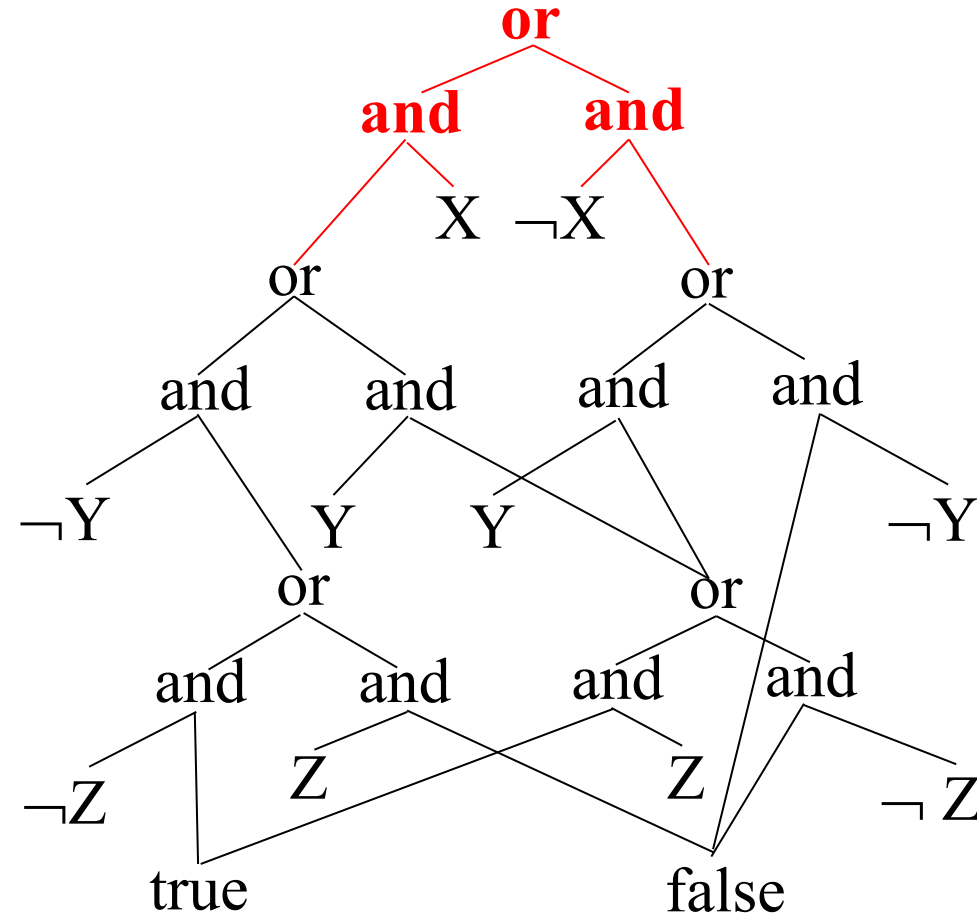
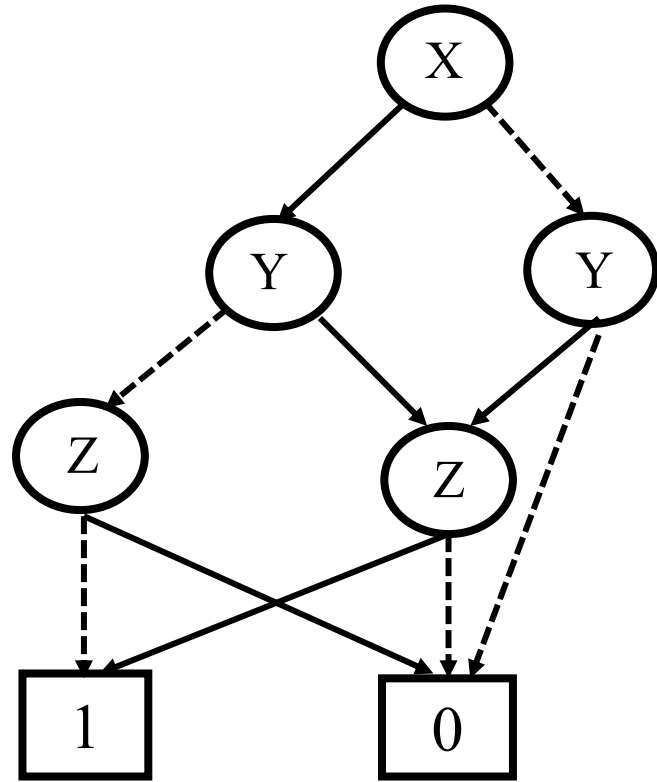
What ML system?
What symbolic representation?
What can we do with the symbolic representation?



Ordered Binary Decision Diagrams (OBDDs)

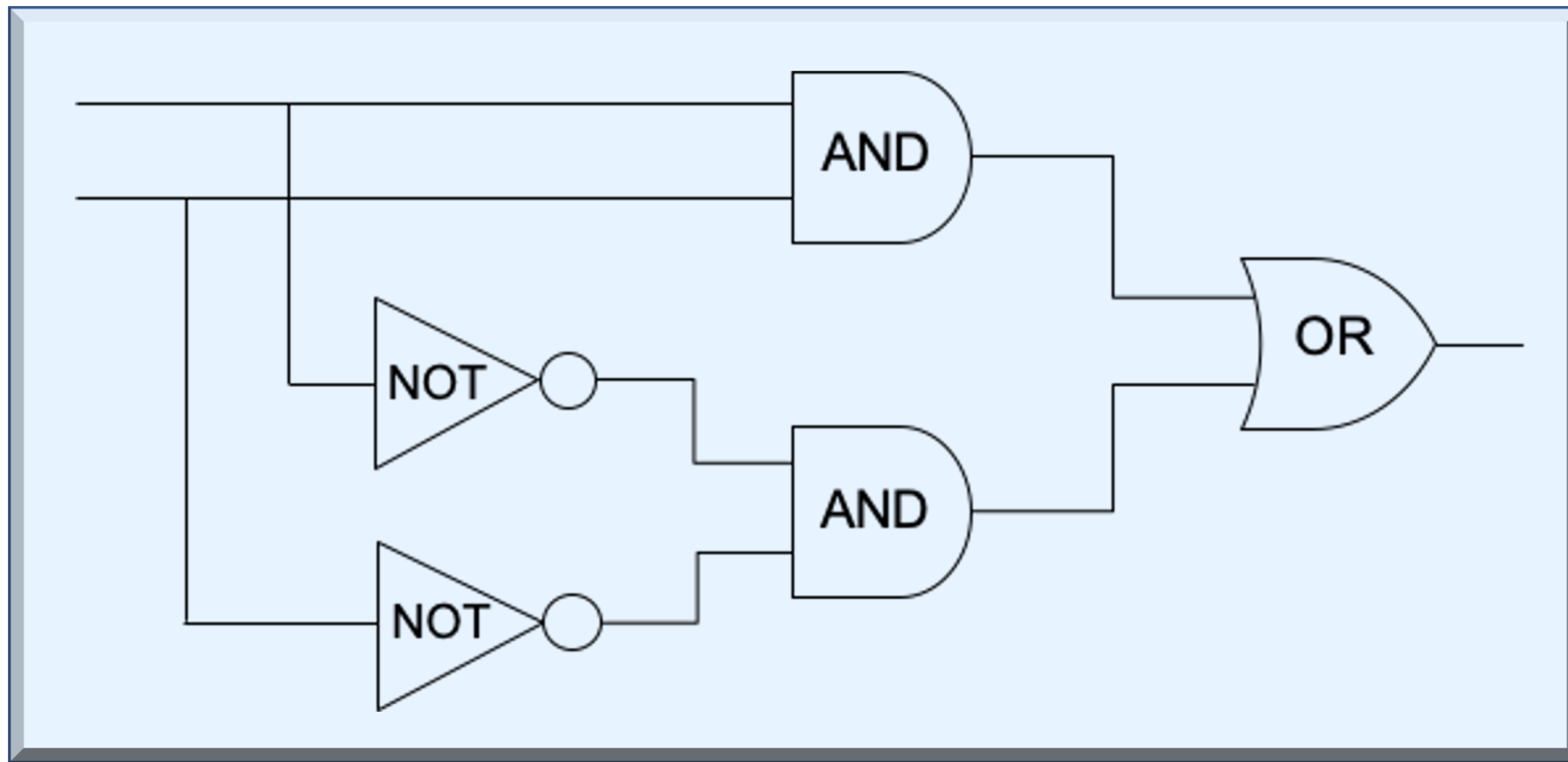


Ordered Binary Decision Diagrams (OBDDs)



Negation Normal Form Circuits

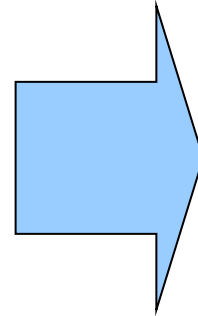
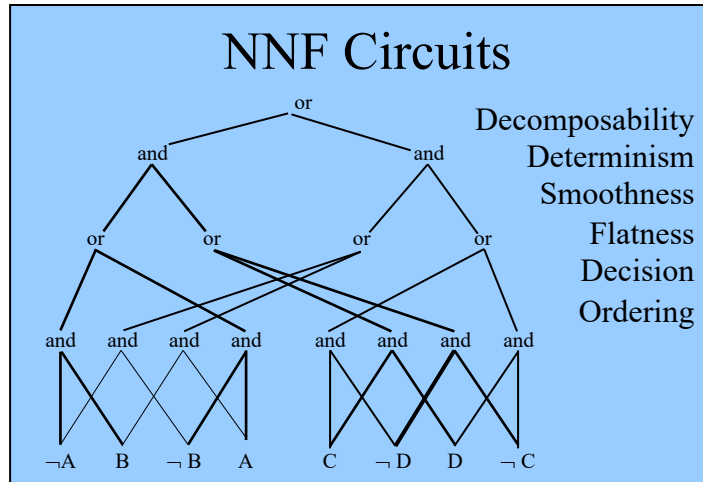
NNF Circuits



Tractable Circuits Knowledge Compilation

Darwiche & Marquis, JAIR 2002

OBDD
SDD
d-DNNF
DNNF
...



Polytime Operations

Consistency (CO)
Validity (VA)
Clausal entailment (CE)
Sentential entailment (SE)
Implicant testing (IP)
Equivalence testing (EQ)
Model Counting (CT)
Model enumeration (ME)

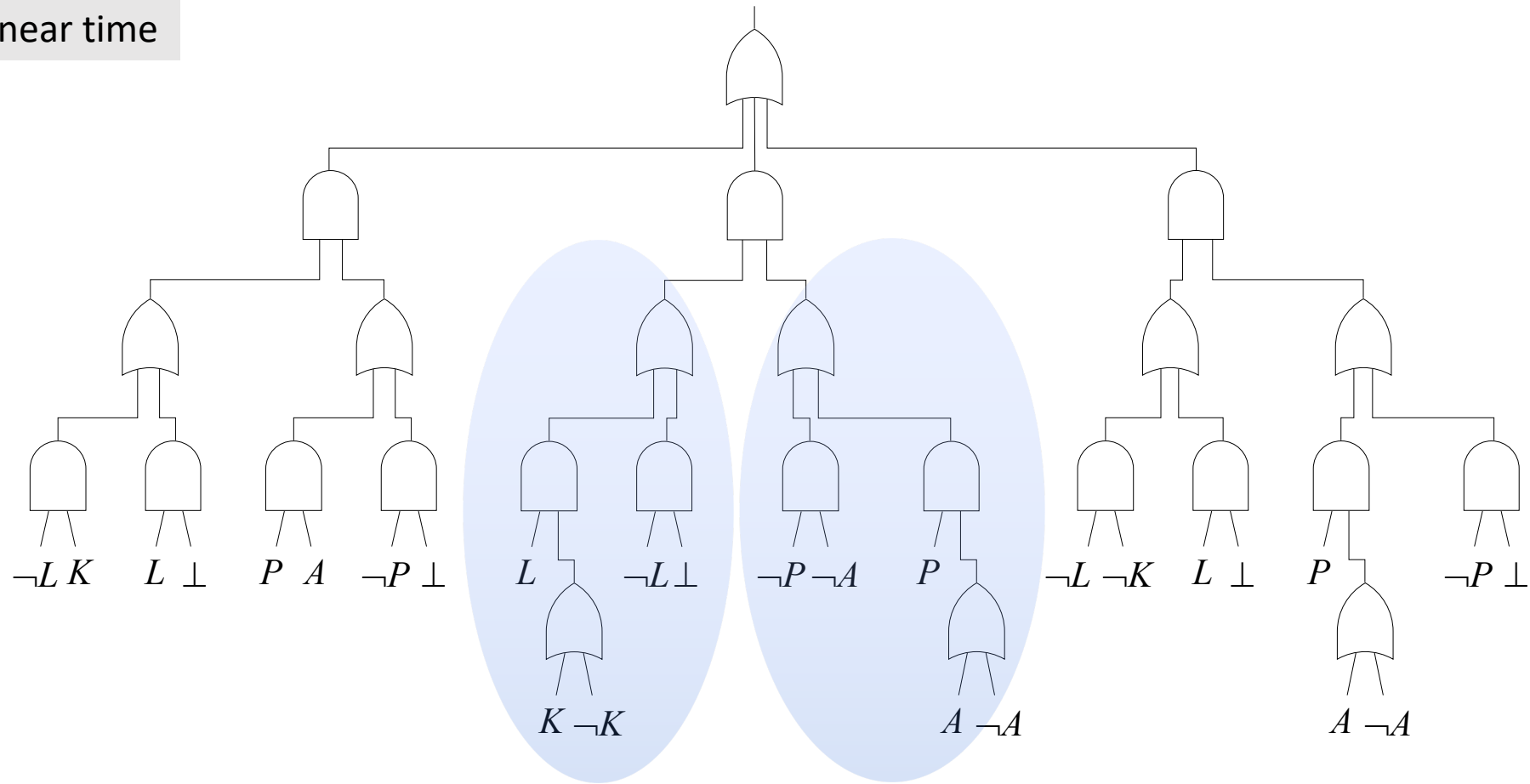
Existential quantification
Conditioning
Conjoin, Disjoin, Negate

Succinctness

Decomposability (DNNF)

Darwiche, JACM 2001

SAT in linear time

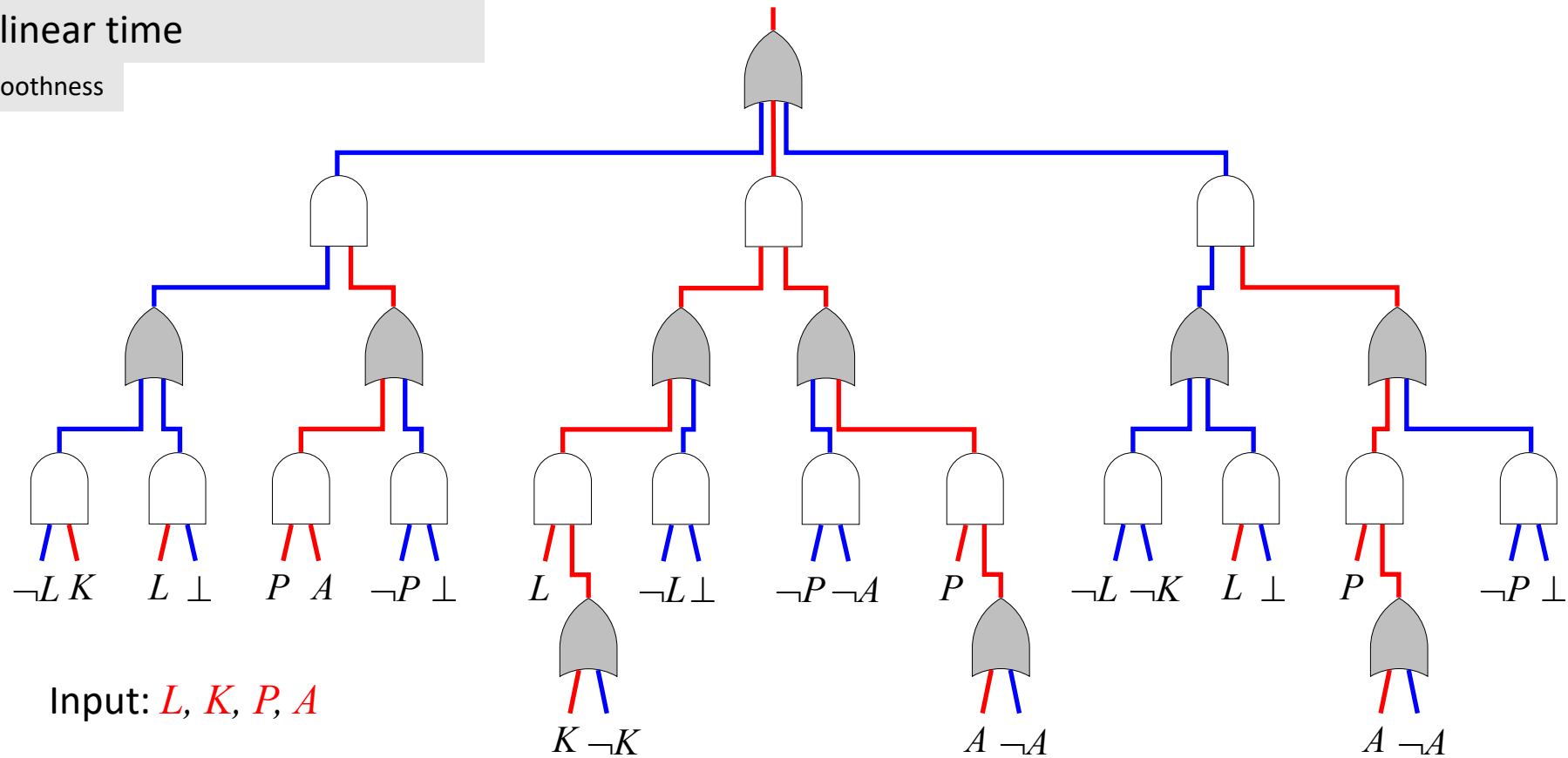


Determinism (d-DNNF)

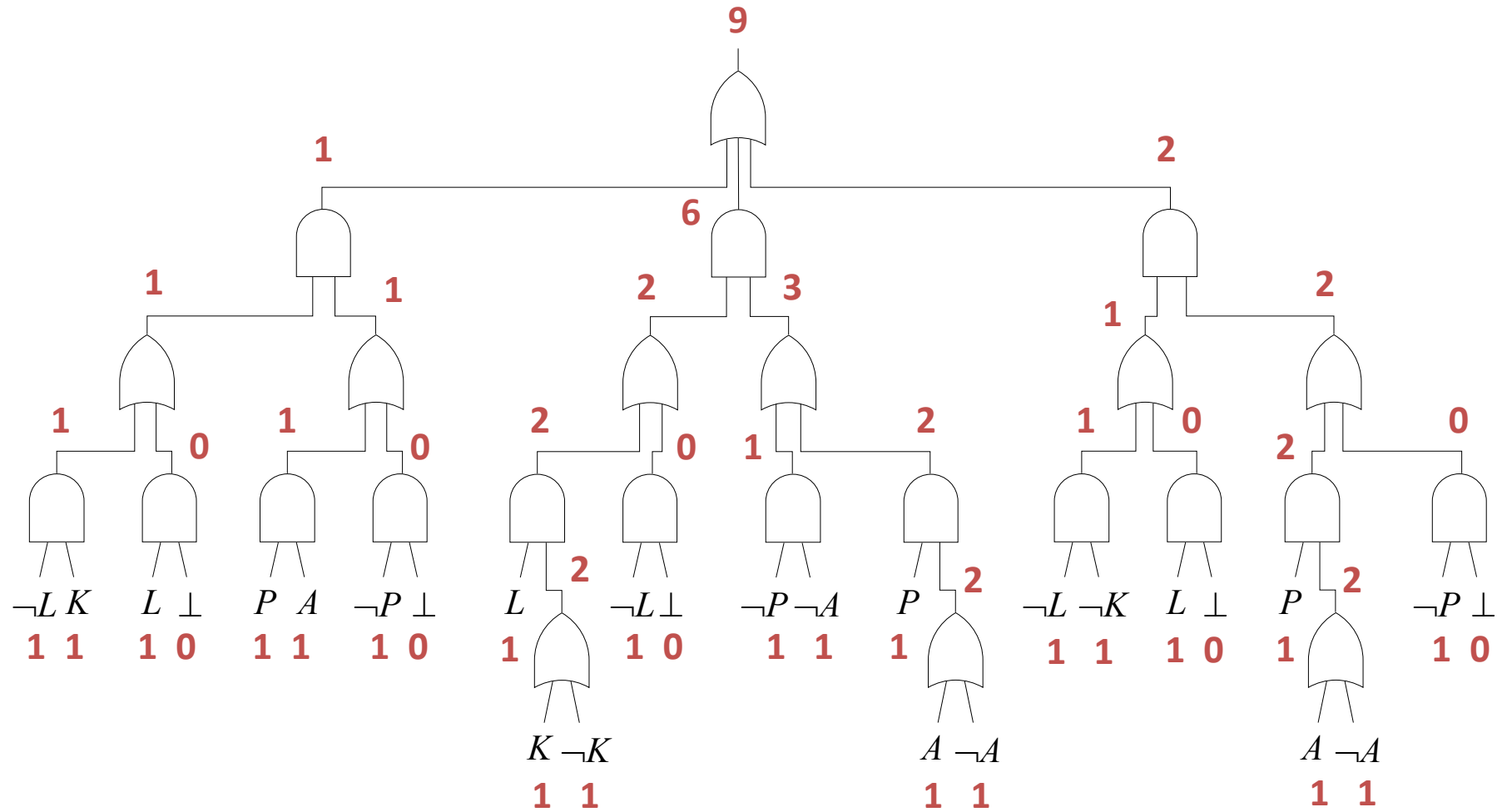
Darwiche, JANCL 2001

#SAT in linear time

requires smoothness



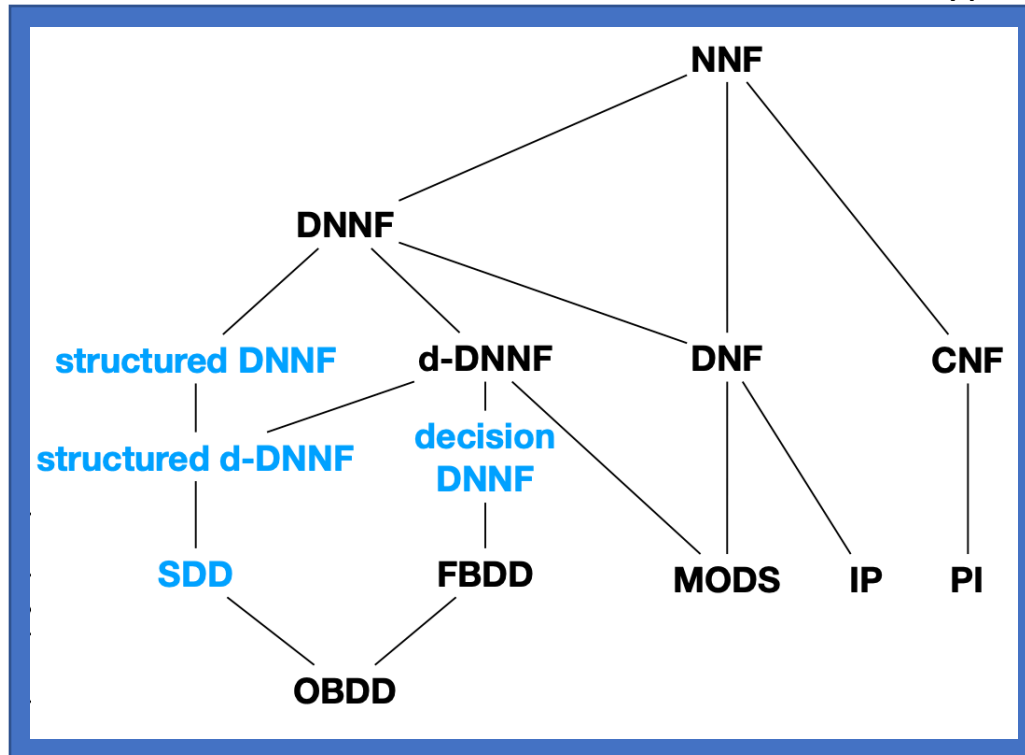
Model Counting (#SAT)



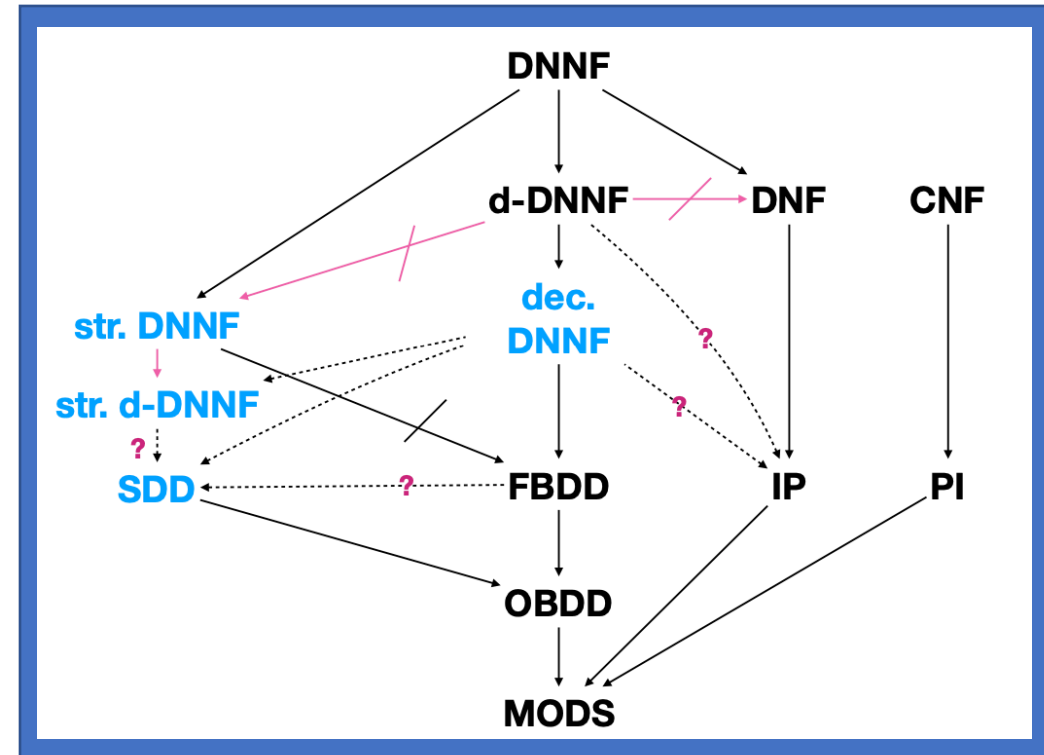
Knowledge Compilation Map

Darwiche & Marquis, JAIR 2002

circuit types



succinctness



KOCOON workshop 2019

Friedrich Slivovsky. An Introduction to Knowledge Compilation [87]

See PODS paper for pointers to:

knowledge compilers, model counters, weighted model counters, reduction tools and other resources

CS264A: Automated Reasoning

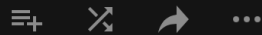
Lecture I-A
Fall 2020

Instructor: Adnan Darwiche

▶ PLAY ALL

Automated Reasoning

35 videos • 10,657 views • Last updated on Dec 17, 2020



Lectures by Adnan Darwiche for his UCLA course on Automated Reasoning. The course is focused on the interplay between logic, probabilistic reasoning and machine learning. The unifying theme is tractable Boolean and Arithmetic circuits (knowledge compilation).



UCLA Automated Reasoning Group

SUBSCRIBE

14

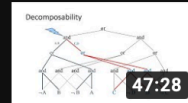


50:34

Lecture 7B: Tractable Circuits & Knowledge Compilation Map

UCLA Automated Reasoning Group

15

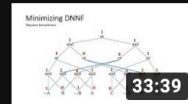


47:28

Lecture 8A: DNNF Circuits (Decomposability)

UCLA Automated Reasoning Group

16

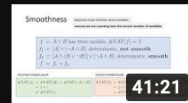


33:39

Lecture 8B: DNNF Circuits (Minimization and Structured Decomposability)

UCLA Automated Reasoning Group

17

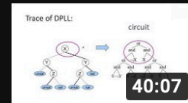


41:21

Lecture 9A: d-DNNF circuits (Determinism and Smoothness)

UCLA Automated Reasoning Group

18

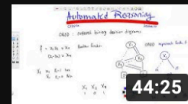


40:07

Lecture 9B: Top-Down Knowledge Compilers

UCLA Automated Reasoning Group

19

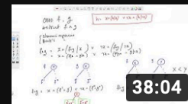


44:25

Lecture 10A: OBDD Circuits (Binary Decision Diagrams)

UCLA Automated Reasoning Group

20

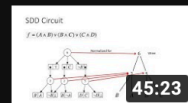


38:04

Lecture 10B: OBDD Circuits (Binary Decision Diagrams)

UCLA Automated Reasoning Group

21

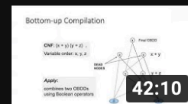


45:23

Lecture 11A: SDD Circuits (Sentential Decision Diagrams)

UCLA Automated Reasoning Group

22



42:10

Lecture 11B: Bottom-Up Knowledge Compilers

UCLA Automated Reasoning Group



Lecture 12A: PSDD Circuits (Probabilistic Sentential Decision Diagrams)

Knowledge Compilation Meets X

Abhay Kumar Jha, Dan Suciu:

Knowledge Compilation Meets Database Theory: Compiling Queries to Decision Diagrams. Theory Comput. Syst. 52(3): 403-440 (2013)

Simone Bova, Florent Capelli, Stefan Mengel, Friedrich Slivovsky:

Knowledge Compilation Meets Communication Complexity. IJCAI 2016: 1008-1014

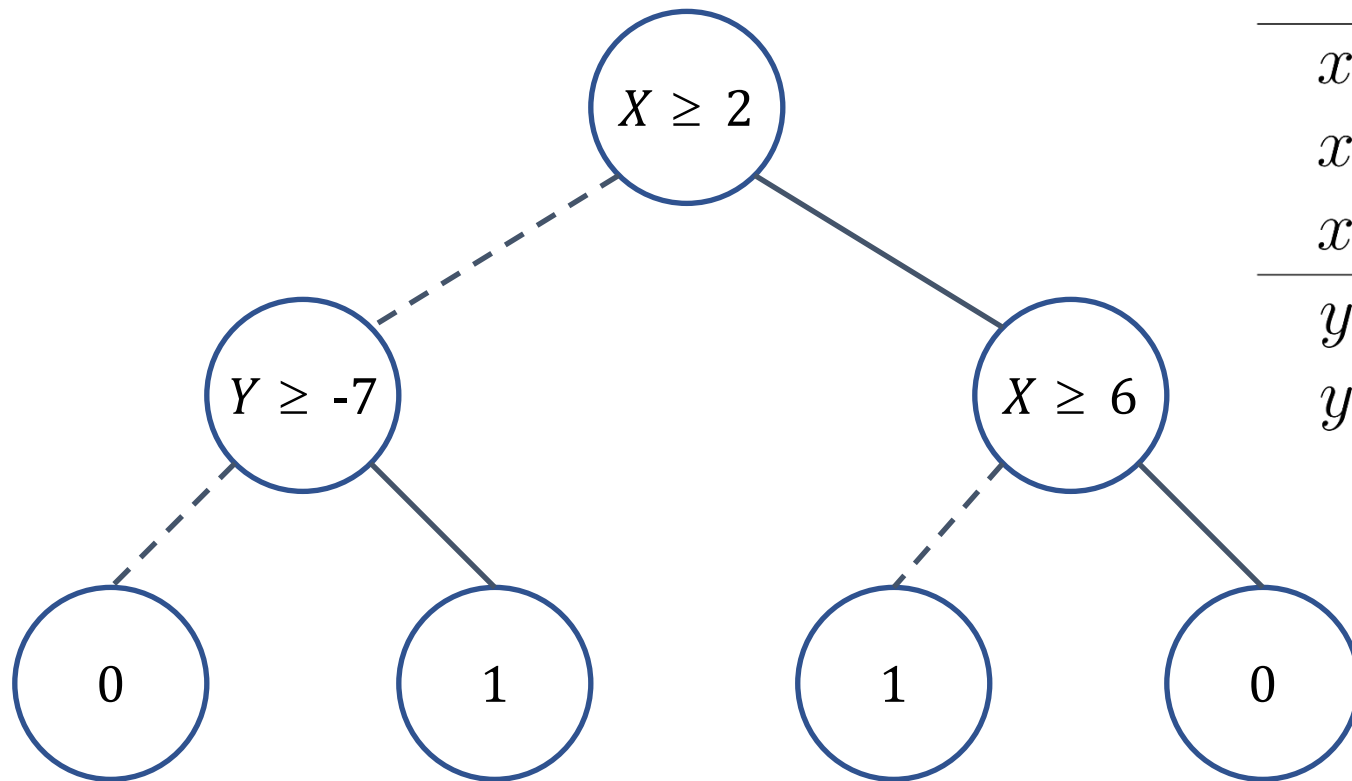
Shubham Sharma, Rahul Gupta, Subhajit Roy, Kuldeep S. Meel:

Knowledge Compilation meets Uniform Sampling. LPAR 2018: 620-636

Compiling ML Classifiers into tractable circuits

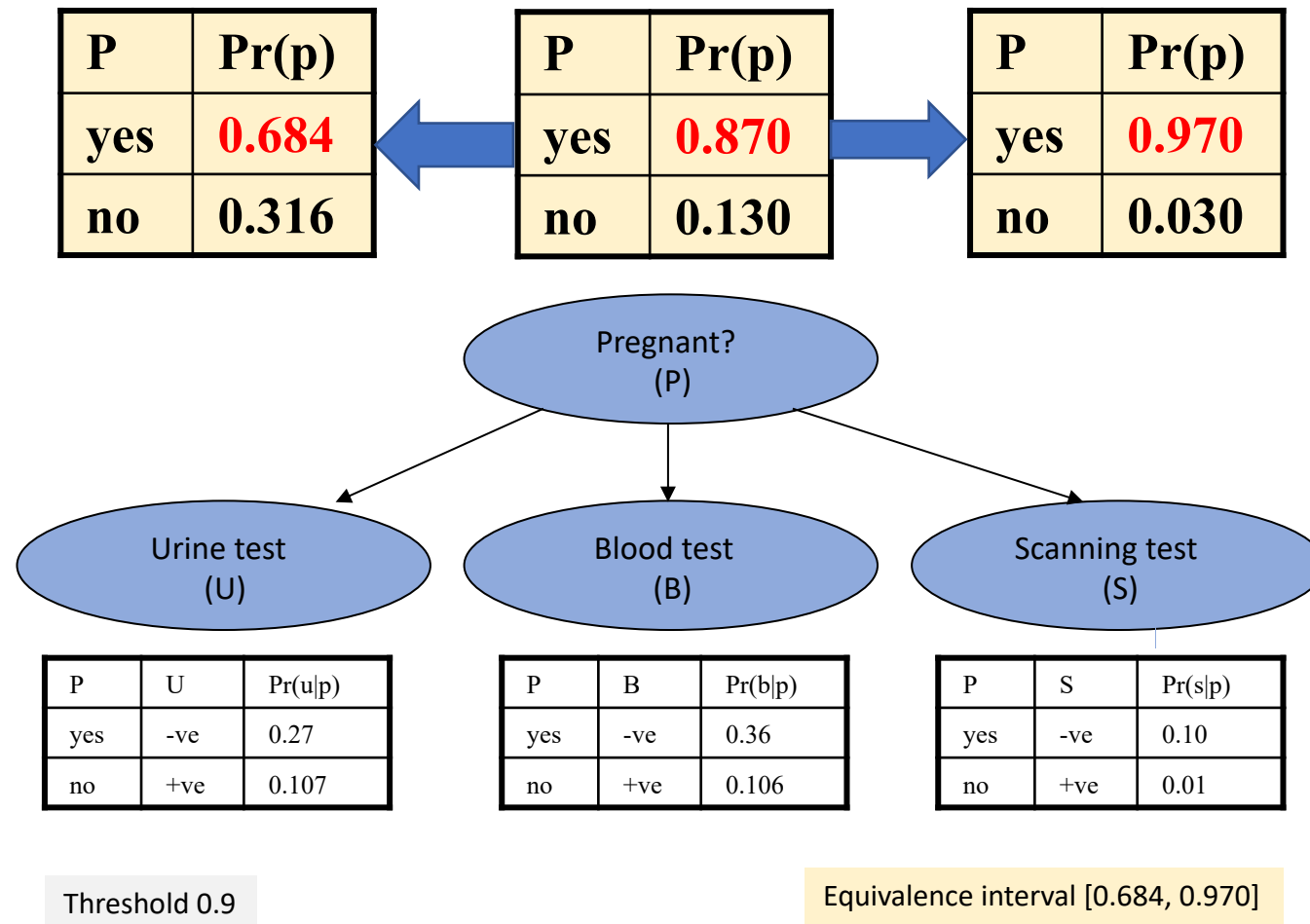
- Bayesian Networks
 - Chan, Darwiche. UAI 03
 - Shih, Choi, Darwiche. IJCAI 18
 - Shih, Choi Darwiche. AAAI 19
- Neural Networks (restricted classes)
 - Choi, Shi, Shih, Darwiche. VNN 2019
 - Shih, Darwiche, Choi. SAT 2019
 - Shi, Shih, Darwiche, Choi. KR 2020
- Decision Trees & Random Forests
 - Choi, Shih, Goyanka, Darwiche. FoMLAS 2020
 - Audemard, Koriche, Marquis. KR 2020

Decision Tree \rightarrow Discrete Function



value	interval	X	Y	$f(X, Y)$
x_1	$(-\infty, 2)$	x_1	y_1	0
x_2	$[2, 6)$	x_1	y_2	1
x_3	$[6, +\infty)$	x_2	y_1	1
y_1	$(-\infty, -7)$	x_2	y_2	1
y_2	$[-7, +\infty)$	x_3	y_1	0
		x_3	y_2	0

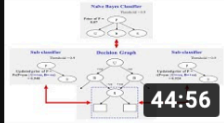
Numbers Don't Matter as Much



Compiling ML Classifiers into tractable circuits

- Bayesian Networks
 - Chan, Darwiche. UAI 03
 - Shih, Choi, Darwiche. IJCAI 18
 - Shih, Choi Darwiche. AAAI 19
- Neural Networks (restricted classes)
 - Choi, Shi, Shih, Darwiche. VNN 2019
 - Shih, Darwiche, Choi. SAT 2019
 - Shi, Shih, Darwiche, Choi. KR 2020
- Decision Trees & Random Forests
 - Choi, Shih, Goyanka, Darwiche. FoMLAS 2020
 - Audemard, Koriche, Marquis. KR 2020

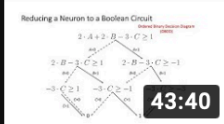
29



44:56

Lecture 15A: Compiling Bayesian Network Classifiers
UCLA Automated Reasoning Group

30



43:40

Lecture 15B: Compiling Neural Network and Random Forest Classifiers
UCLA Automated Reasoning Group

Explaining Decisions

PI-Explanation (Sufficient Reason) Shih, Choi & Darwiche (IJCAI 18)

minimal set of instance characteristics that can trigger the decision
(other features are irrelevant)

Example Explanation

Sally tested negative for
Scanning, **B**lood and **U**rine

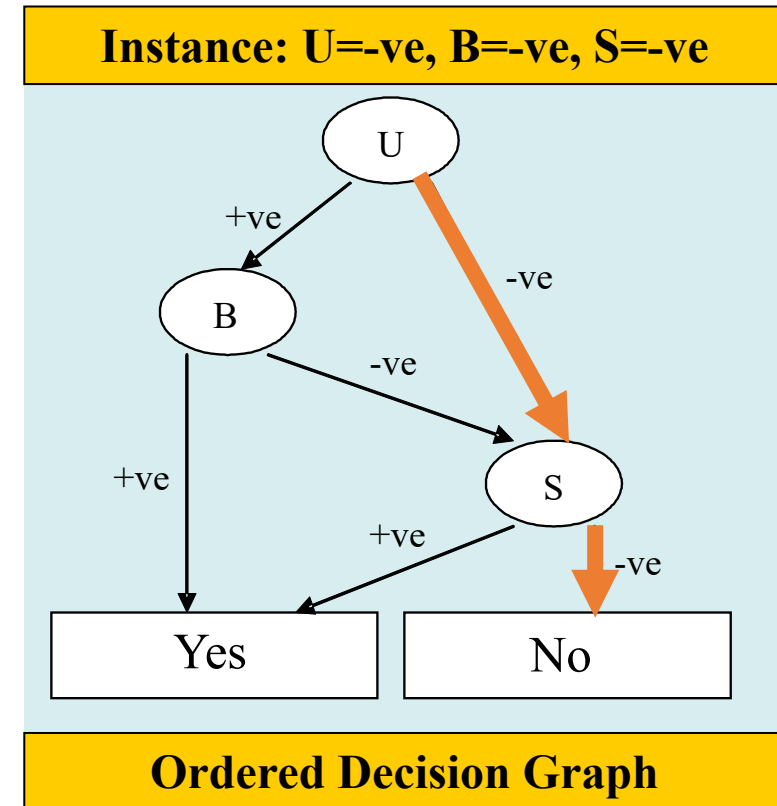
Why did you conclude that Sally
is not pregnant?

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

sufficient reasons

- S=-ve and B=-ve
- S=-ve and U=-ve

The **complete reason** behind the decision
(S=-ve and (B=-ve or U=-ve))



Prime Implicants

decades old: CS + AI

Boolean function $f = (A + \bar{C})(B + C)(A + B)$

Prime implicants $AB, AC, B\bar{C}$

Instance: $AB\bar{C}$

Decision: 1

Sufficient reasons: $AB, B\bar{C}$

Boolean function $\bar{f} = \overline{(A + \bar{C})(B + C)(A + B)}$

Prime implicants $\bar{A}\bar{C}, \bar{B}\bar{C}, \bar{A}\bar{B}$

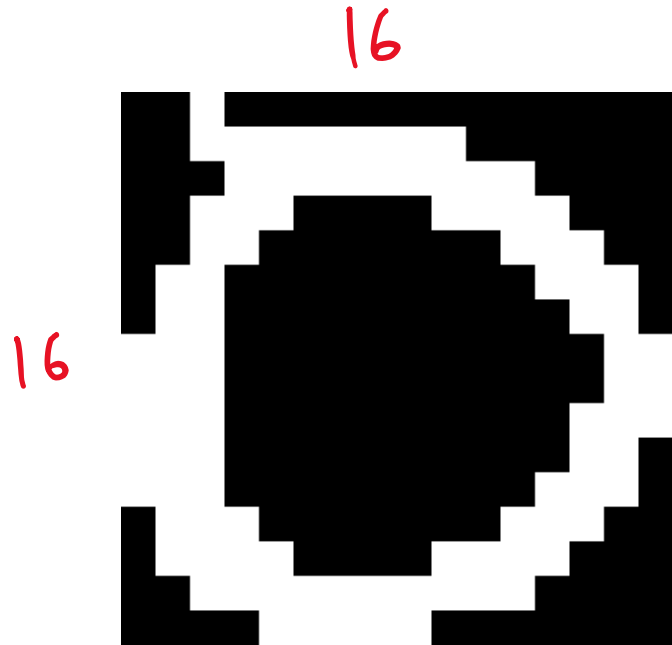
Instance: $\bar{A}BC$

Decision: 0

Sufficient reasons: $\bar{A}\bar{C}$

Issue: we may have an exponential number of prime implicants

Image Classifier (0 vs 1)



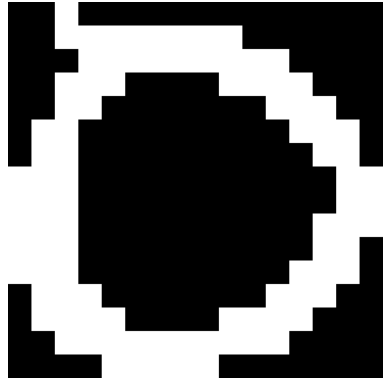
Pixels: $P_{1,1}$ ---- , $P_{16,16}$

Classifier

$f(P_{1,1}, \text{----}, P_{16,16})$

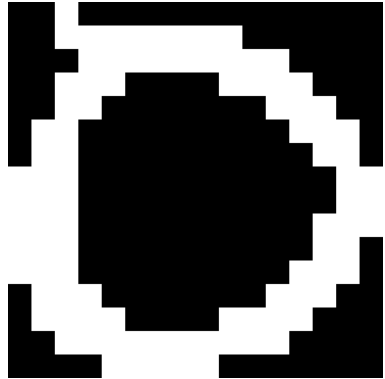
total : $16 \times 16 = 256$ pixels
 2^{256} images

Example: Sufficient Reason

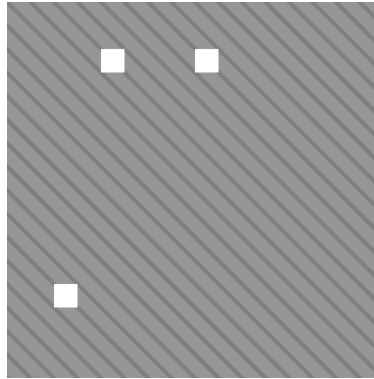


Why did you conclude
that this image is a 0?

Example: Sufficient Reason

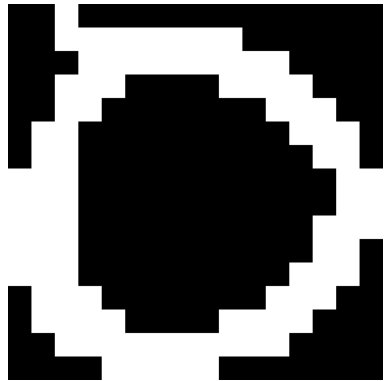


Why did you conclude that this image is a 0?

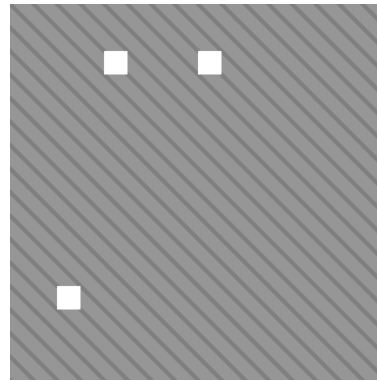


Because these 3 white pixels are sufficient to label the image 0

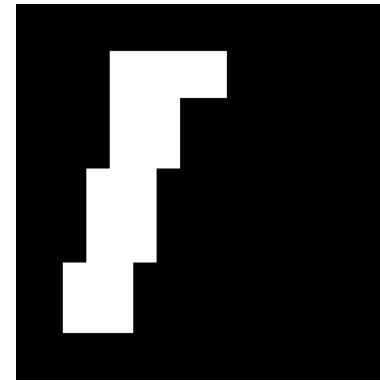
Example: Sufficient Reason



Why did you conclude that this image is a 0?

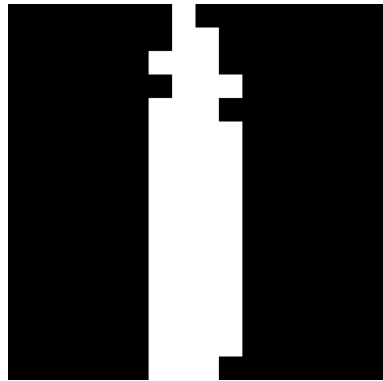


Because these 3 white pixels are sufficient to label the image 0

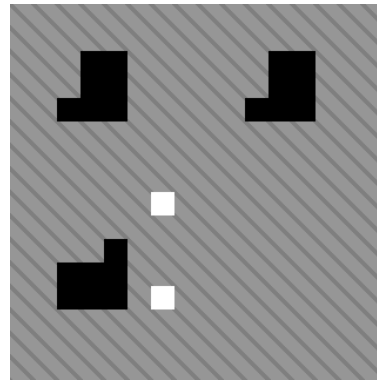


We **know** that this classifier will also label this image as a 0

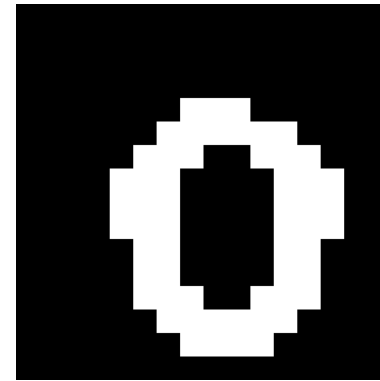
Example: Sufficient Reason



Why did you conclude that this image is a 1?



Because these pixels are sufficient to label the image 1



We **know** that this classifier will also label this image as a 1

The Complete Reason Behind a Decision

Darwiche, Hirth (ECAI 2020)

disjunction of all sufficient reasons

avoids computing sufficient reasons explicitly

Reason Circuit

(for a decision)

tractable circuit representation of the complete reason

permits answering questions about sufficient reasons efficiently

Decision Bias

E: passed the entrance exam

F: first-time applicant

G: has good grades (GPA)

W: has work experience

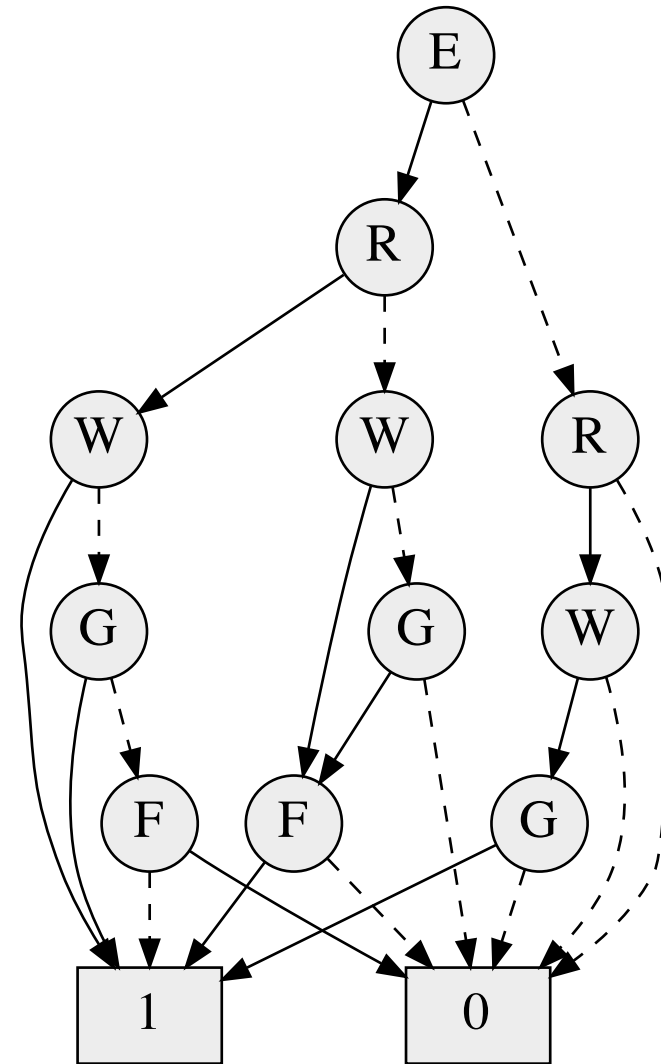
R: comes from a rich hometown

Decision on instance X is **biased** iff it can be different on an instance Y that disagrees with X on protected features only

Theorem: Decision is biased iff each of its sufficient reasons contains at least one protected feature

Classifier is **biased** iff one of its decisions is biased

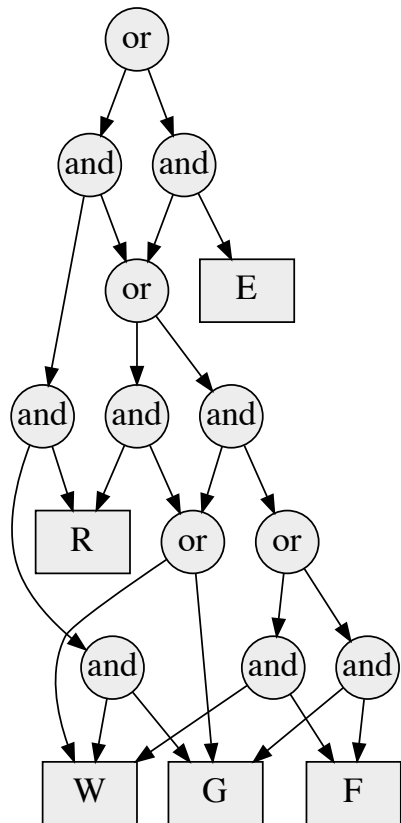
Theorem: Classifier is biased iff one of its decisions has a sufficient reason with at least one protected feature



Admissions classifier compiled into an OBDD

Complete Reason

Robin is admitted. Why?

$$(E, F, G) \quad (E, F, W) \quad (E, G, \underline{R}) \quad (E, \underline{R}, W) \quad (G, \underline{R}, W)$$


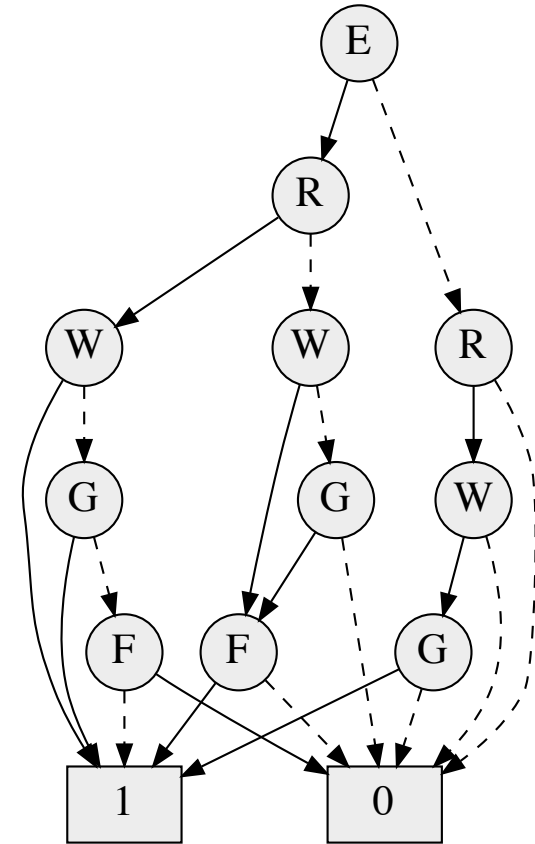
Reason Circuit

obtained in linear time
(OBDD, Decision-DNNF)

monotone circuit

existential quantification in linear time

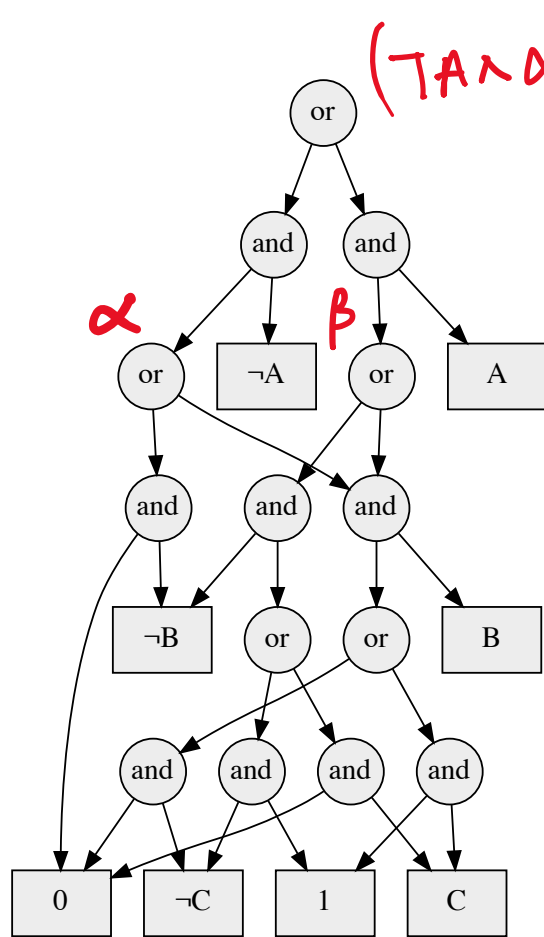
Decision not Biased



Robin

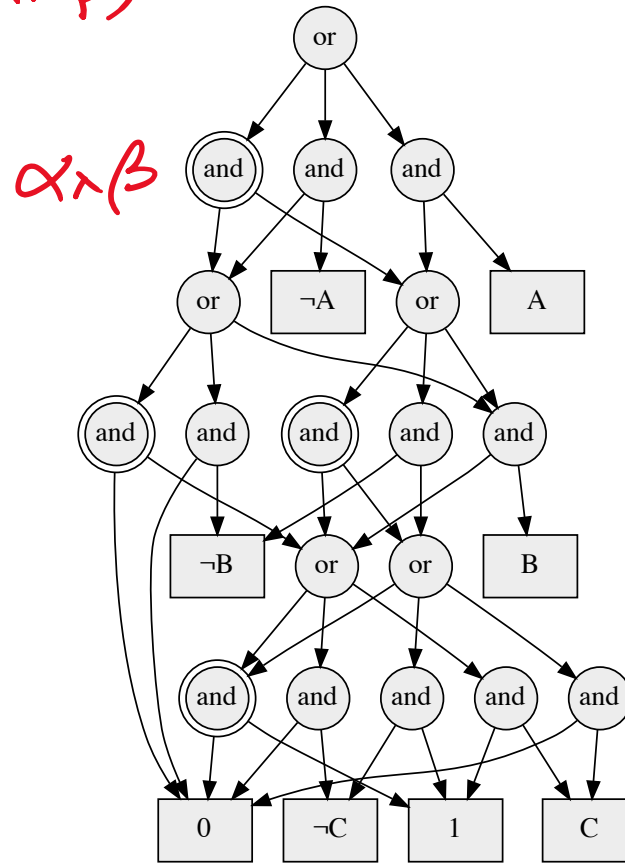
- ✓ **E:** passed the entrance exam
- ✓ **F:** first-time applicant
- ✓ **G:** has good grades (GPA)
- ✓ **W:** has work experience
- ✓ **R:** comes from a rich hometown

Computing the Reason Circuit

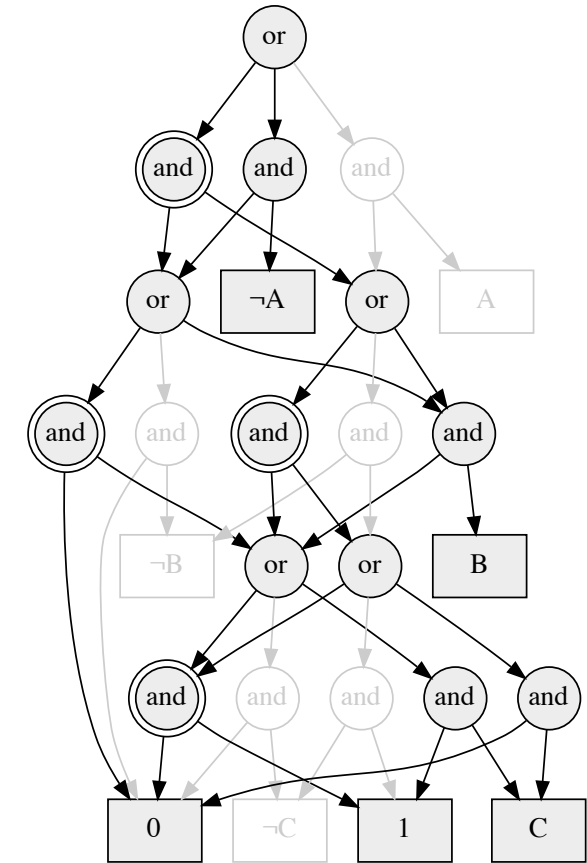


classifier

Instance: not A,B,C



consensus



filtering

Reason Circuit

What else can be done with Reason Circuits?

On The Reasons Behind Decisions.
Darwiche & Hirth (ECAI 2020)

8.3 Computing Queries

Susan would still be admitted **even if** she did not have a high GPA **because** she passed the entrance exam and comes from a rich hometown.

Sufficient Reasons. By Theorems 2 and 12, the call $\text{PI}(\Delta_\alpha, \alpha)$ to Algorithm 1 will return all sufficient reasons for decision $\Delta(\alpha)$, assuming Δ_α is a Decision-DNNF circuit. The number of sufficient reasons can be exponential, but we can actually answer many questions about them without enumerating them directly as shown below.

Necessary Property. By Proposition 4, characteristic (literal) l is necessary for decision $\Delta(\alpha)$ iff $\mathcal{R} \models l$. This is equivalent to $\mathcal{R}|\neg l$ being unsatisfiable, which can be decided in $O(n)$ time given Theorem 10. The necessary property (all necessary characteristics) can then be computed in $O(n \cdot m)$ time.

Necessary Reason. To compute the necessary reason (if any) we compute the necessary property and check whether it satisfies the complete reason. This can be done in $O(n \cdot m)$ time.

Because Statements. To decide whether decision $\Delta(\alpha)$ was made “because τ ” we check whether property τ is the complete reason for the decision (Definition 5): $\tau \models \mathcal{R}$ and $\mathcal{R} \models \tau$. We have $\tau \models \mathcal{R}$ iff $(\neg \mathcal{R})|\tau$ is unsatisfiable. Moreover, $\mathcal{R} \models \tau$ iff $\mathcal{R}|\neg l$ is unsatisfiable for every literal l in τ . All of this can be done in $O(n \cdot |\tau|)$ time.

Even if, Because Statements. To decide whether decision $\Delta(\alpha)$ would stick “even if $\bar{\rho}$ because τ ” we replace property ρ with $\bar{\rho}$ in instance α to yield instance β (Definition 6). We then compute the complete reason for decision $\Delta(\beta)$ and check whether it is equivalent to τ . All of this can be done $O(n \cdot |\tau|)$ time.

Decision Bias. To decide whether decision $\Delta(\alpha)$ is biased we existentially quantify all unprotected features from circuit \mathcal{R} and then check the validity of the result (Theorem 6). All of this can be done in $O(n)$ time given Theorems 10 and 11.

Decision and Classifier Robustness

Hamming Distance

1. Instance $x=1, y=0, z=1, w=1$
2. Instance $x=1, y=0, z=0, w=0$

$$d(\text{Instance 1}, \text{Instance 2}) = 2$$

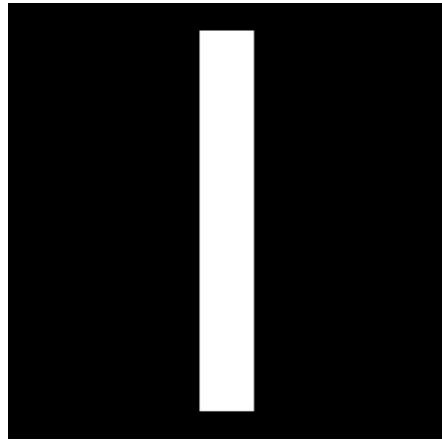
Decision Robustness

Shih, Choi, Darwiche (PGM 18)

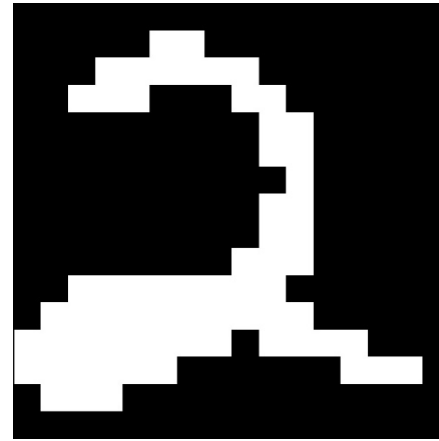
$$\text{robustness}_f(\mathbf{x}) = \min_{\mathbf{x}': f(\mathbf{x}') \neq f(\mathbf{x})} d(\mathbf{x}, \mathbf{x}')$$

- How many features do we need to flip, to flip the classifier's decision?
- Decision robustness is coNP-complete
- Linear time in an OBDD

Example: Robustness



most robust 1:
robustness 3



most robust 2:
robustness 13

Classifier Robustness

Shi, Shih, Darwiche, Choi. KR 2020

$$\text{model-robustness}(f) = \frac{1}{2^n} \sum_{\mathbf{x}} \text{robustness}_f(\mathbf{x})$$

the expected robustness, averaged over all possible 2^n inputs

Classifier Robustness

$$\text{model-robustness}(f) = \frac{1}{2^n} \sum_{\mathbf{x}} \text{robustness}_f(\mathbf{x})$$

$$\text{model-robustness}(f_{\text{parity}}) = 1$$

$$\text{model-robustness}(f_{\text{or}}) = \frac{n}{2} + \frac{1}{2^n}$$

Classifier Robustness

Shi, Shih, Darwiche, Choi. KR 2020

NN1:

98.18% accuracy

1,298 nodes

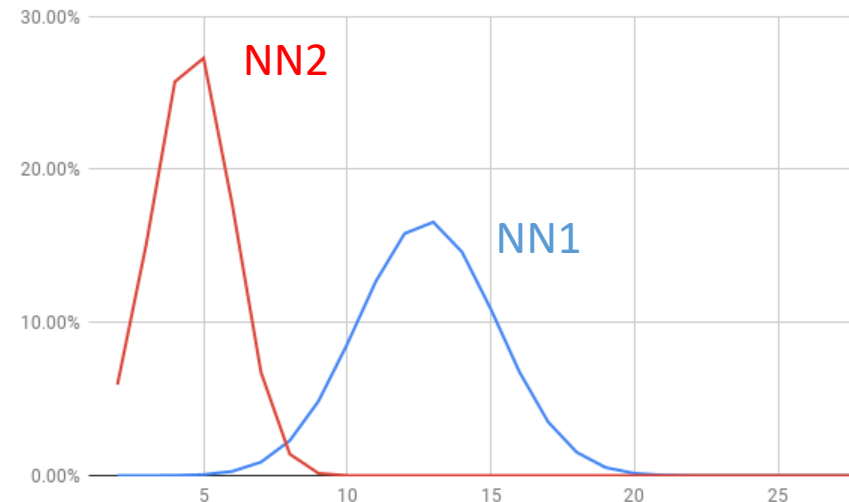
3,653 edges

11.77

average robustness

27

max robustness



NN2:

96.93% accuracy

203 nodes

440 edges

3.62

average robustness

13

max robustness

<http://reasoning.cs.ucla.edu/xai>

Verification: Monotone Classifiers

Shih, Choi, Darwiche (PGM 18)

Positive decision remains positive if we flip some features from $-$ to $+$.

If instance $(+,-,-,+)$ gives positive decision, these also give positive decision:

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

Educational Testing:

Susan's correct answers include Jack's correct answers

Susan should pass if Jack passed

Credit Application:

Susan and Jack have the same characteristics, except that Susan has a higher income

Susan should be approved if Jack is approved

Verification: Monotone Classifiers

Shih, Choi, Darwiche (PGM 18)

- Quadratic complexity on OBDDs
- Educational assessment classifier not monotone (threshold $\frac{1}{2}$)
- 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

Reasoning about the Behavior of ML Systems

new role for symbolic AI & CS methods

Reason About What Was Learned

Systems 1 / 2 (thinking fast and slow), reflection, meta-reasoning

compile-then-reason paradigm

(VNN community: other techniques including SAT)

Three Modern Roles for Logic in AI

Adnan Darwiche
Computer Science Department
University of California, Los Angeles
darwiche@cs.ucla.edu

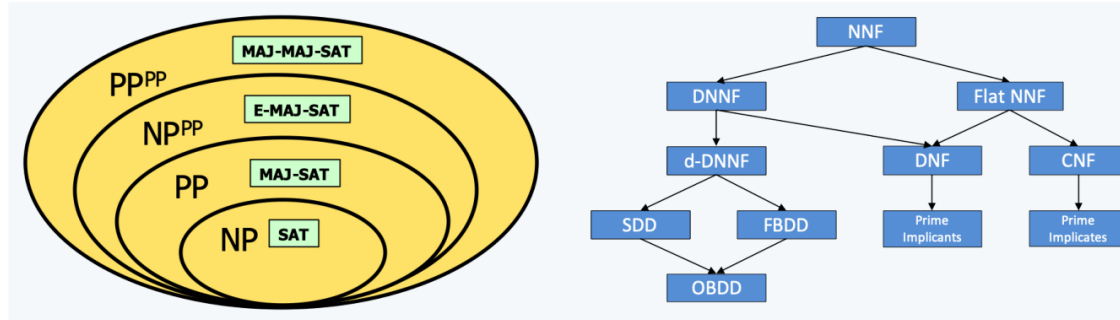


Figure 1: Tractable Boolean circuits as a basis for computation.

ABSTRACT

We consider three modern roles for logic in artificial intelligence, which are based on the theory of tractable Boolean circuits: (1) logic as a basis for computation, (2) logic for learning from a combination of data and knowledge, and (3) logic for reasoning about the behavior of machine learning systems.

CCS CONCEPTS

• Computing methodologies → Learning in probabilistic graphical models; Logical and relational learning; • Theory of computation → Automated reasoning; Complexity classes; Problems, reductions and completeness.

KEYWORDS

tractable circuits, knowledge compilation, explainable AI

ACM Reference Format:

Adnan Darwiche. 2020. Three Modern Roles for Logic in AI. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS’20)*, June 14–19, 2020, Portland, OR, USA. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3375395.3389131>

1 INTRODUCTION

Logic has played a fundamental role in artificial intelligence since the field was inception [52]. This role has been mostly in the area of knowledge representation and reasoning, where logic is used to

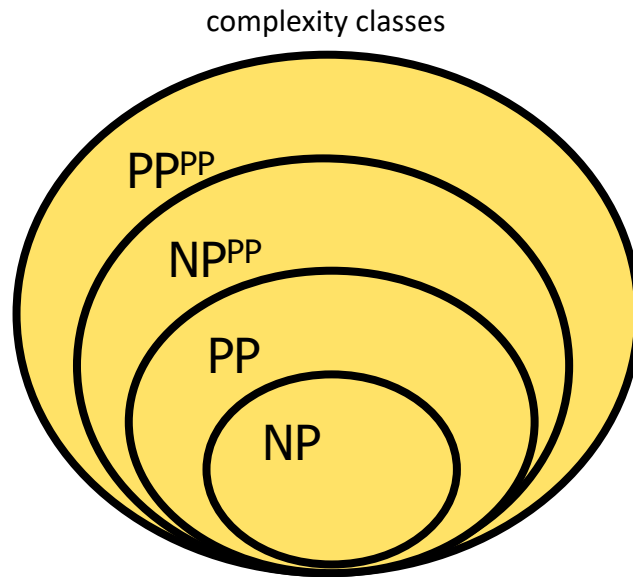
represent categorical knowledge and then draw conclusions based on deduction and other more advanced forms of reasoning. Starting with [59], logic also formed the basis for drawing conclusions from a mixture of categorical and probabilistic knowledge.

In this paper, we review three modern roles for propositional logic in artificial intelligence, which are based on the theory of tractable Boolean circuits. This theory, which matured considerably during the last two decades, is based on Boolean circuits in Negation Normal Form (NNF) form. NNF circuits are not tractable, but they become tractable once we impose certain properties on them [34]. Over the last two decades, this class of circuits has been studied systematically across three dimensions. The first dimension concerns a synthesis of NNF circuits that have varying degrees of tractability (the polytime queries they support). The second dimension concerns the relative succinctness of different classes of tractable NNF circuits (the optimal size circuits can attain). The third dimension concerns the development of algorithms for compiling Boolean formula into tractable NNF circuits.

The first modern role for logic we consider is in using tractable circuits as a basis for computation, where we show how problems in the complexity classes NP, PP, NP^{PP} and P^{PP} can be solved by compiling Boolean formula into corresponding tractable circuits. These are rich complexity classes, which include some commonly utilized problems from probabilistic reasoning and machine learning. We discuss this first role in two steps. In Section 2, we discuss the prototypical problems that are complete for these complexity classes, which are all problems on Boolean formula. We also discuss problems from probabilistic reasoning which are complete for these classes and their reduction to prototypical problems. In Section 3, we introduce the theory of tractable circuits with exposure to circuit types that can be used to efficiently solve problems in these complexity classes (if compiled successfully).

Logic For Computation

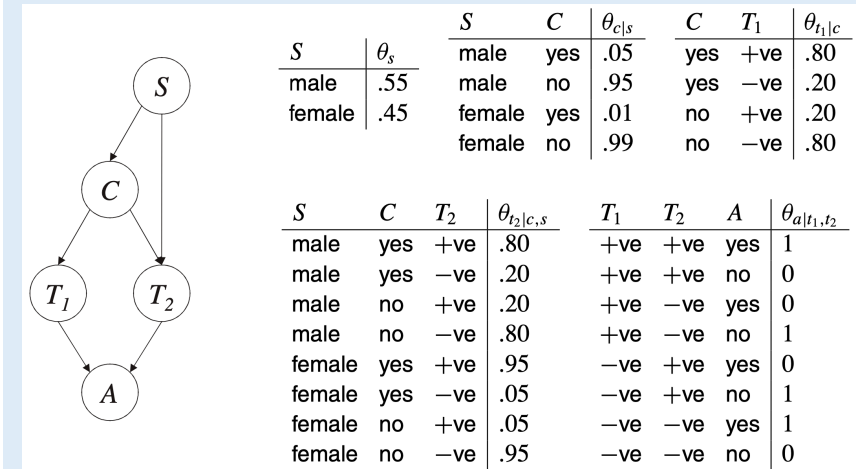
reducing NP & 'Beyond NP' problems to logical reasoning



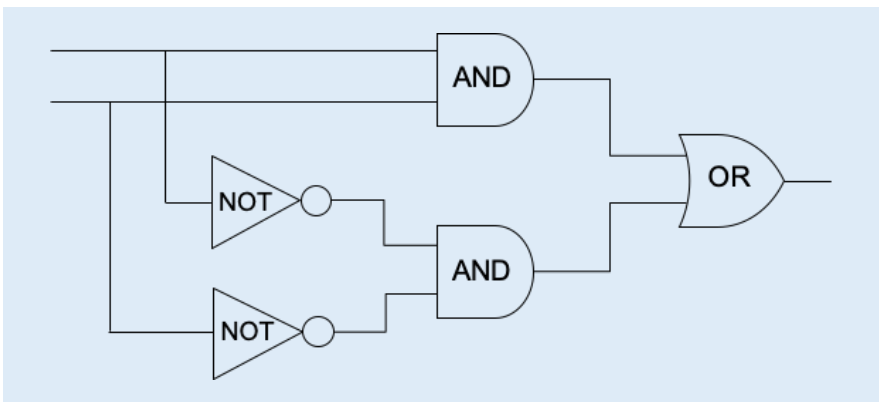
prototypical problems (on Boolean formula)

$((A \text{ or } B) \text{ and } (\text{not } C)) \text{ or } (\text{not } B \text{ and } D)$

complete problems (probabilistic reasoning & ML)

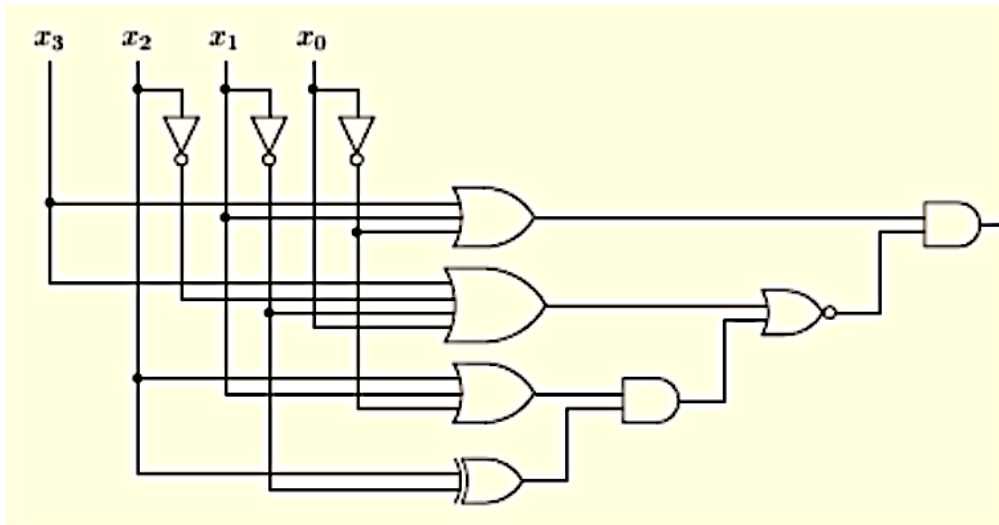


tractable Boolean circuits (essence of computation)

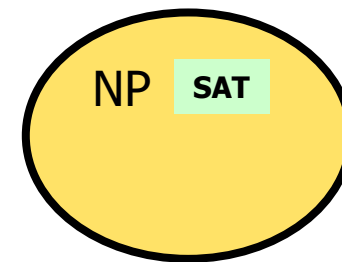


Boolean Circuits

complexity classes

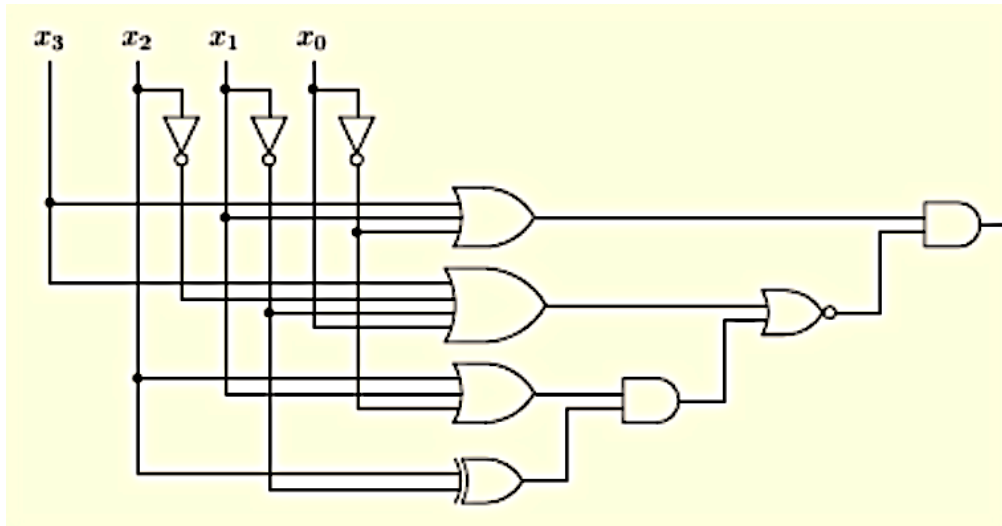


Circuit input that generates 1-output?

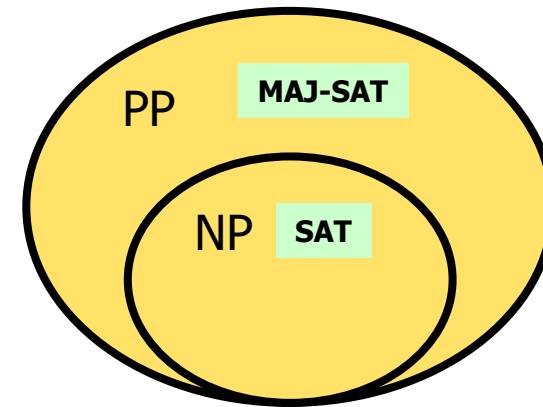


Boolean Circuits

complexity classes

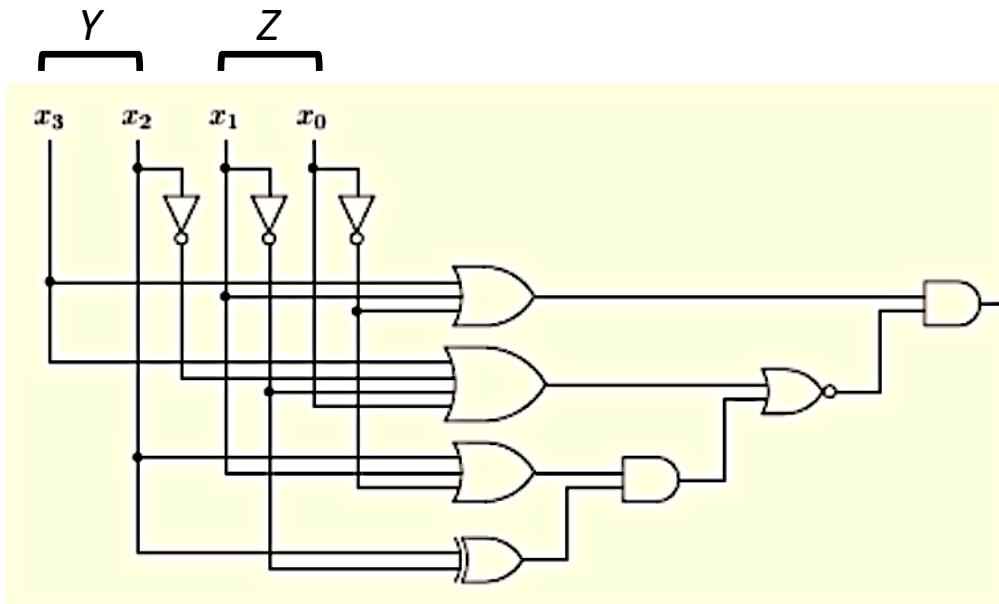


Majority of circuit inputs generate 1-output?

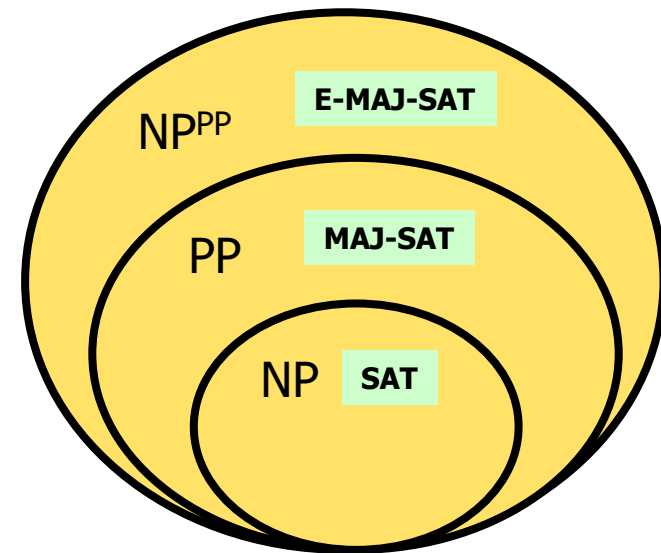


Boolean Circuits

complexity classes



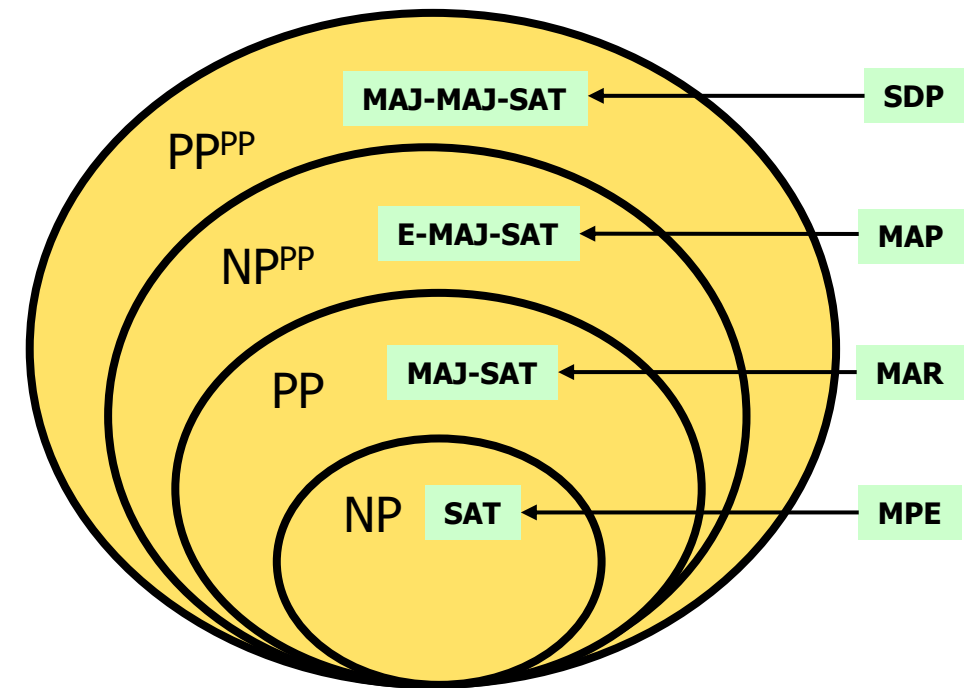
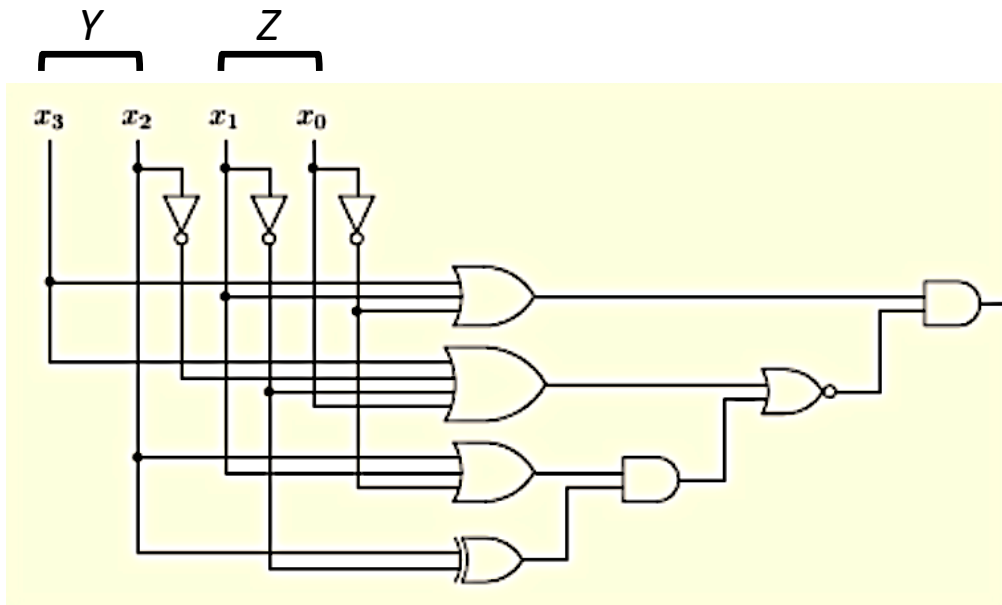
Y-input under which majority of Z-inputs generate 1-output?



Boolean Circuits

complexity classes

probabilistic reasoning



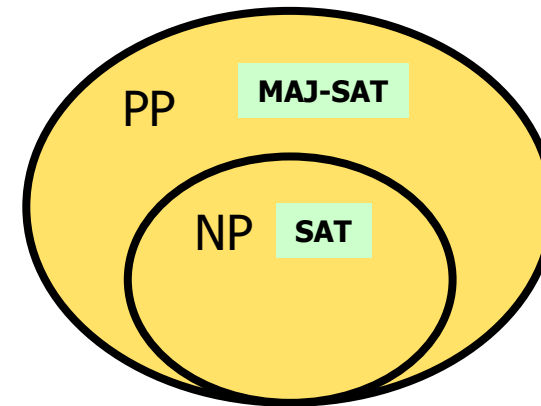
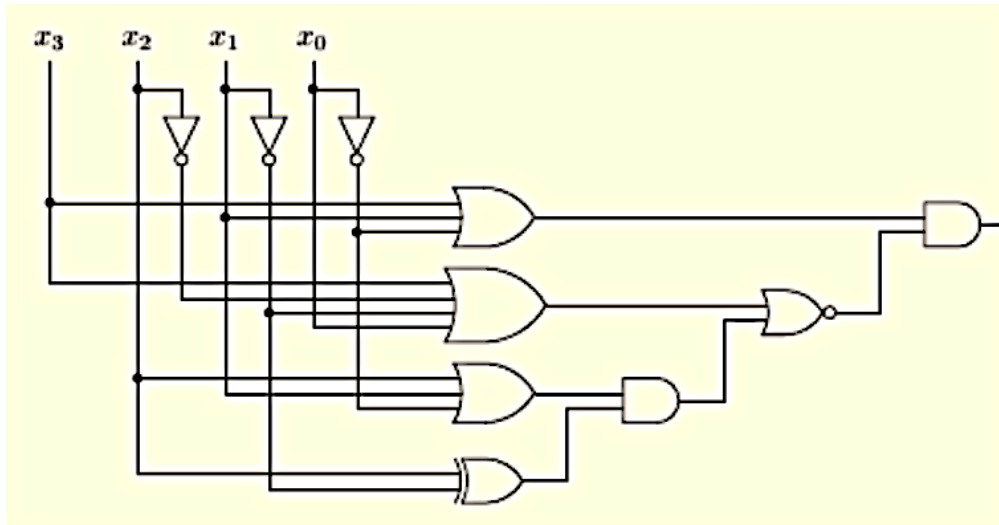
Majority of Y-inputs under which majority of Z-inputs generate 1-output?

Boolean Circuits

complexity classes

weights

$w(x_0=0), w(x_0=1) \dots w(x_3=0), w(x_3=1)$



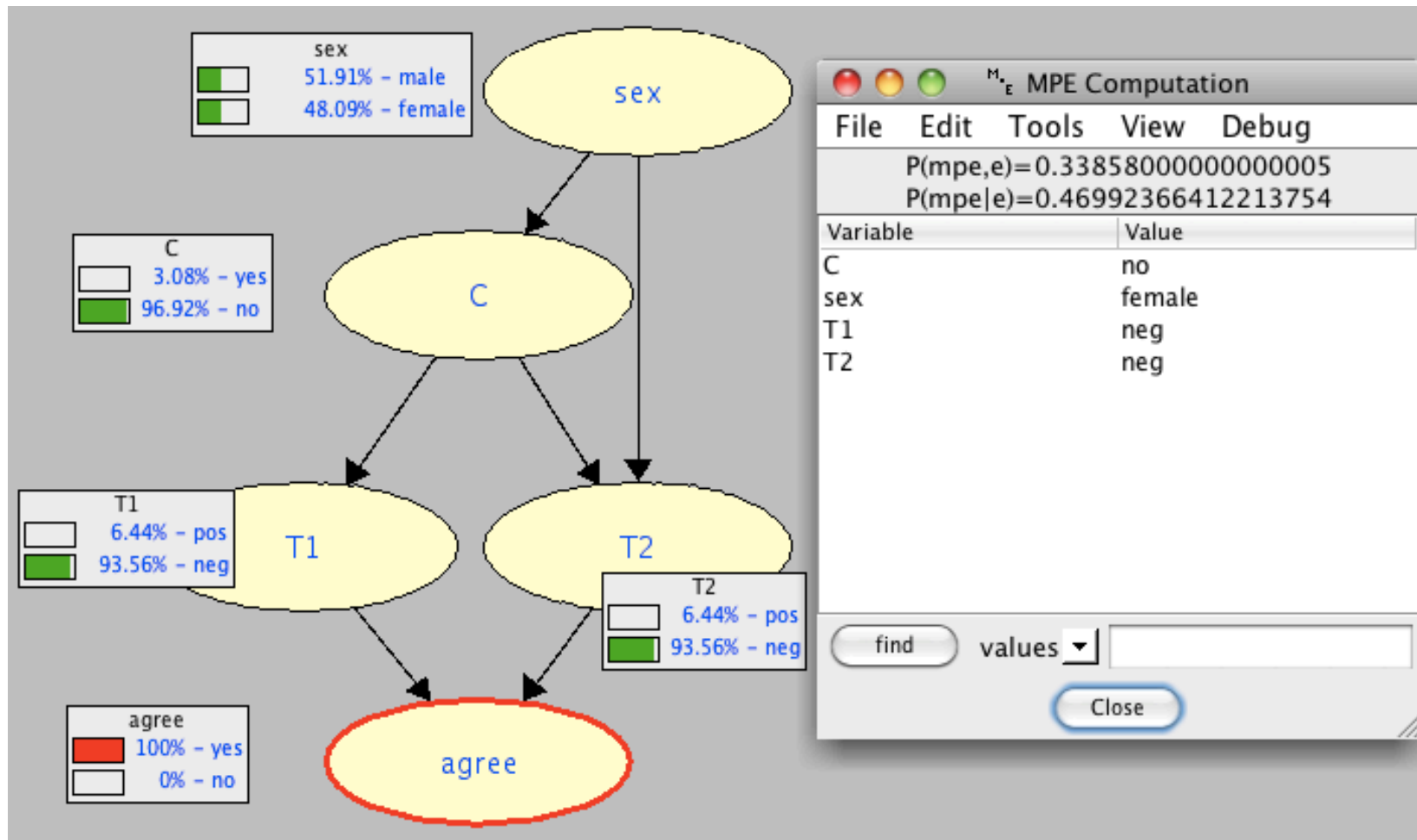
Majority of circuit inputs generate 1-output?

Count of circuit inputs that generate 1-output? (**#SAT**)

Weighted count of circuit inputs that generate 1-output? (**WMC**)

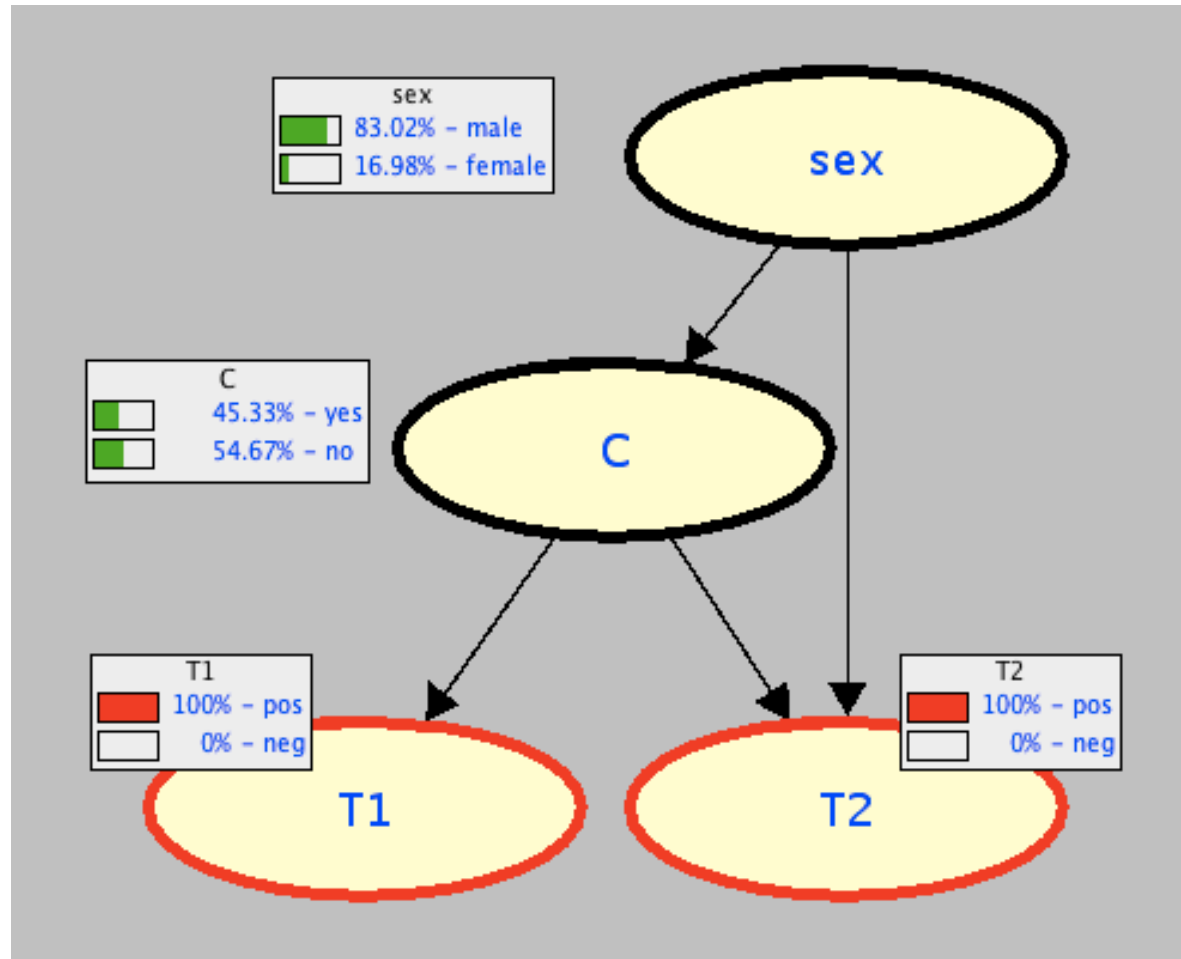
NP-complete query

Most Probable Explanation (MPE)



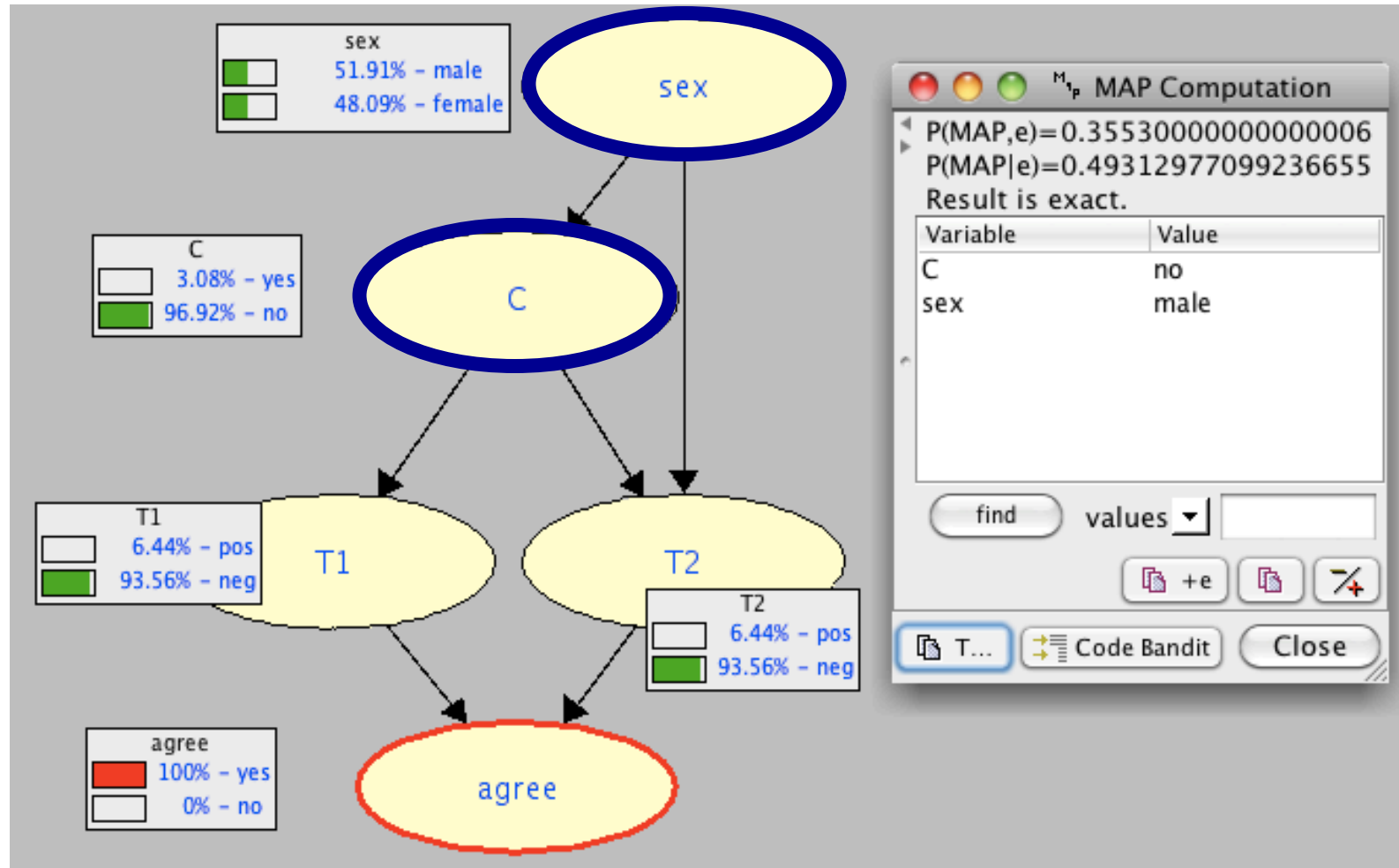
PP-complete query

Marginal Probabilities (MAR)



NP^{PP}-complete query

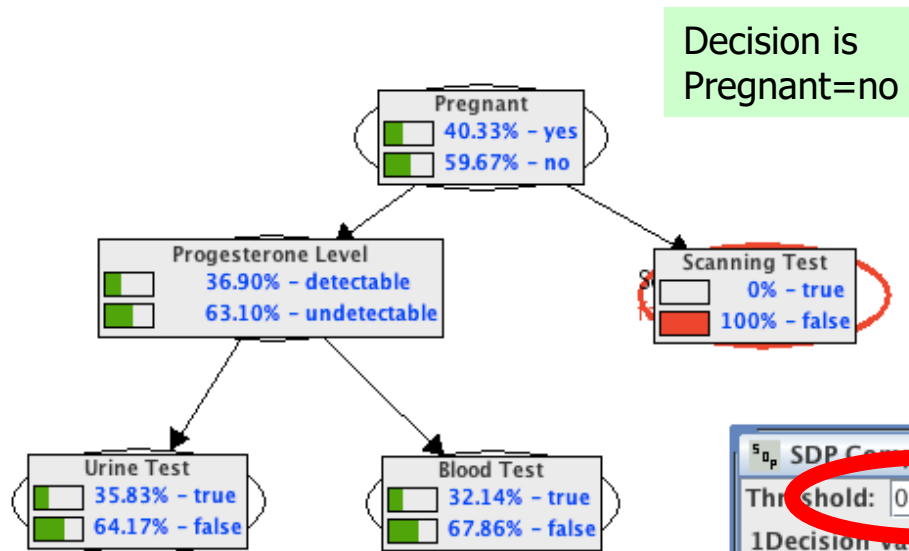
Maximum a Posterior Hypothesis (MAP)



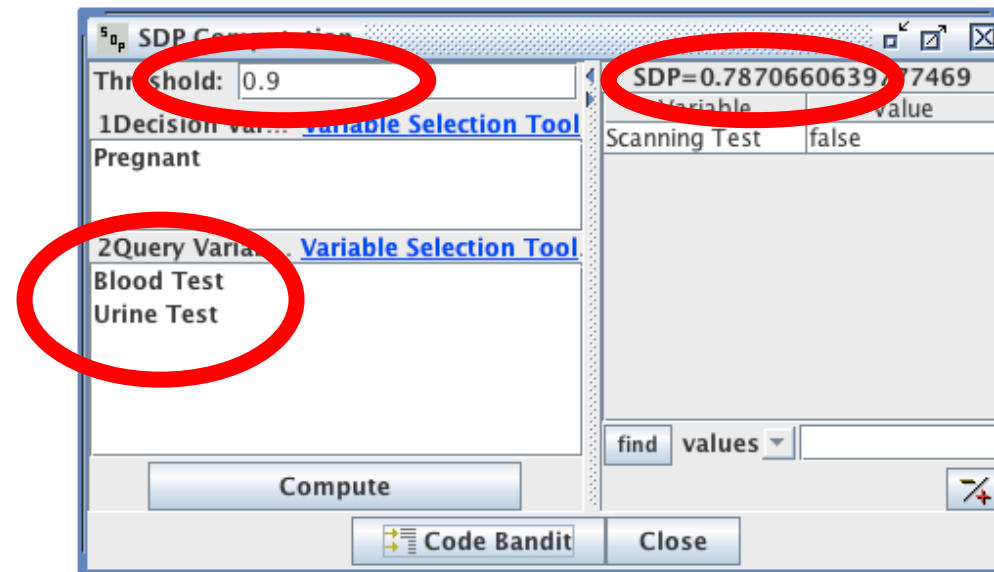
PP^{PP}-complete query

Same-Decision Probability (SDP)

Darwiche & Choi, PGM 2010



78.7% chance you will still make the same decision after collecting the blood and urine tests.

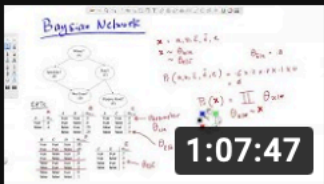


$w(A) = w(\neg A) = w(B) = w(\neg B) = w(C) = w(\neg C) = 1$
 $w(\neg P_{\alpha|\beta}) = 1$
 $w(P_{\alpha|\beta}) = \theta_{\alpha|\beta}$

$A \wedge B \Leftrightarrow P_{A|B}$

A	B	C	Pr(.)
T	T	T	$\theta_A \theta_{B A} \theta_{C A}$
T	T	F	$\theta_A \theta_{B A} \theta_{\neg C A}$

31



Lecture 16: Reducing Probabilistic Reasoning (MPE) to Weighted MAX-SAT

UCLA Automated Reasoning Group

32



Lecture 17A: Reducing Probabilistic Reasoning (MAR) to Weighted Model Counting

UCLA Automated Reasoning Group

$A \wedge B \wedge \neg C$

F	F	F	$\theta_{\neg A} \theta_{\neg B \neg A} \theta_{\neg C \neg A}$
---	---	---	---

$m = \overbrace{A, B, \neg C}, \overbrace{P_A, P_{B|A}, P_{\neg C|A}}, \overbrace{\neg P_{\neg A}, \neg P_{\neg B|A}, \neg P_{B|\neg A}, \neg P_{\neg B|\neg A}, \neg P_{C|A}, \neg P_{C|\neg A}, \neg P_{\neg C|\neg A}}$

$w(m) = \theta_A \theta_{B|A} \theta_{\neg C|A}$
 $Pr(\delta) = wmc(\Delta \wedge \delta)$

UCLA Automated Reasoning Group

 Publications & Software



UCLA Automated Reasoning Group

1.15K subscribers

SUBSCRIBE

HOME

VIDEOS

PLAYLISTS

COMMUNITY

CHANNELS

ABOUT



Three Modern Roles for Logic in AI | ...



Three Modern Roles for Logic in AI

Adnan Darwiche
UCLA



PODS (June 16, 2020)



0:00 / 1:26:26



Three Modern Roles for Logic in AI | Adnan Darwiche | PODS 2...

1,216 views • 8 months ago

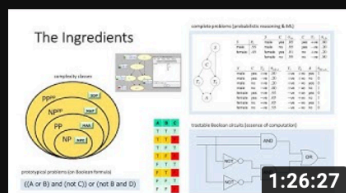
Invited tutorial given at the database theory conference (PODS) on June 16, 2020. The associated PODS paper can be found at: <https://dl.acm.org/doi/abs/10.1145/33...>

The tutorial considers three modern roles for logic in artificial intelligence, which are based on the theory of tractable Boolean (and Arithmetic) circuits:

READ MORE

Invited Talks

▶ PLAY ALL



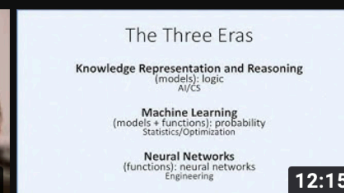
Three Modern Roles for Logic in AI | Adnan Darwiche | ...



Reasoning about the Behavior of AI Systems – ...



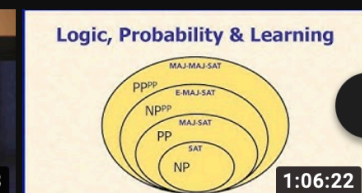
CACM Oct. 2018 - Human-Level Intelligence or Animal...



On AI Education – Adnan Darwiche



2017 WCE: On Model-Based versus Model-Blind...



On the Role of Logic in Probabilistic Inference and...