Risk, Assurance, and Explainability for Autonomous Systems

Rick Kuhn

NIST Computer Security Division

What is the problem?

- AI systems are good, but sometimes make mistakes, and human users will not trust their decisions without explanation or justification
  → <u>assurance and explainability are closely tied</u>
- There is a tradeoff between AI accuracy and explainability: the most accurate methods, such as convolutional neural nets (CNNs), provide no explanations; understandable methods, such as rule-based, tend to be less accurate
- The black-box nature of these systems that makes explanation difficult also makes assurance and testing even harder
- Life-critical aviation software requires MCDC testing, white-box criterion that cannot be used for neural nets and other non-explainable methods

Testing - can we find a solution?

- Gold standard of assurance and verification of life-critical software <u>can't be used</u> for lots of new life-critical autonomy software

- We can measure "neuron coverage", but not clear how closely related to accuracy and ability to correctly process all of the input space

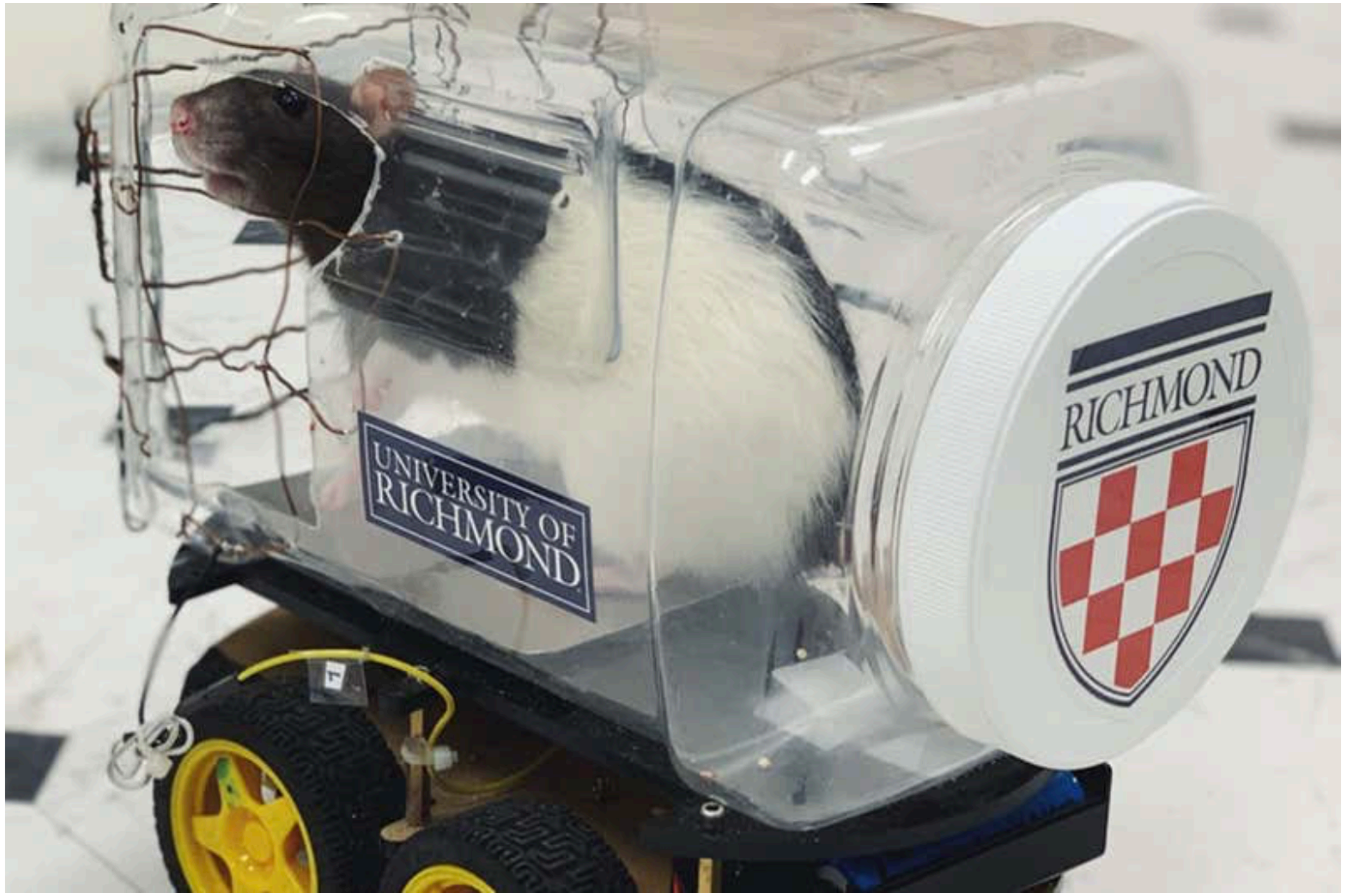
- Why not measure the input space directly?

  Then see if the AI system handles all of it correctly

Scientists have trained rats to drive tiny cars to collect food

LIFE 22 October 2019

By Alice Klein



Yes, but can they do it under all kinds of conditions ….

The problem is easier in a constrained environment

Things get tricky as the scene becomes complex

- Multiple conditions involved in accidents
  - "The camera failed to recognize the white truck against a bright sky"
  - "The sensors failed to pick up street signs, lane markings, and even pedestrians due to the angle of the car shifting in rain and the direction of the sun”

- <u>We need to understand what combinations of conditions are included in testing</u>

Understanding combinations tested

- Cover all 2-way, 3-way, as desired
- Measure coverage of the rest
- Or run scenarios and then measure combinations covered
- Or find set difference of covered/not covered
- We have tools for all of these

COMBINATORIAL TEST SUITE OF STRENGTH 2 FOR THE ROAD SECTION ONTOLOGY

Kluck et al., 2019

lane1 _traffic _id	*lane1 _line1 _id*	*lane1 _line2 _id*	*lane1 _surface _condition*	*lane2 _traffic _id*	*lane2 _line1 _id*	*lane2 _line2 _id*	*lane2 _surface _*
’smooth’	’dotted’	’dotted’	’dry’	’smooth’	’dotted’	’dotted’	’dry’
’smooth’	’none’	’none’	’slippery’	’light’	’none’	’none’	’slippery’
’smooth’	’solid’	’solid’	’icy’	’heavy’	’solid’	’solid’	’icy’
’smooth’	’dotted’	’none’	’icy’	’jam’	’dotted’	’none’	’icy’
’light’	’none’	’solid’	’dry’	’smooth’	’none’	’solid’	’dry’
’light’	’solid’	’dotted’	’slippery’	’light’	’solid’	’dotted’	’slippery’
’light’	’dotted’	’none’	’dry’	’heavy’	’solid’	’none’	’dry’
’light’	’none’	’dotted’	’icy’	’jam’	’none’	’dotted’	’icy’
’heavy’	’solid’	’none’	’slippery’	’smooth’	’dotted’	’solid’	’slippery’
’heavy’	’dotted’	’solid’	’icy’	’light’	’none’	’dotted’	’dry’
’heavy’	’none’	’dotted’	’dry’	’heavy’	’solid’	’none’	’slippery’

Combinatorial coverage – what do we mean?

Tests	Variables			
	a	b	c	d
1	0	0	0	0
2	0	1	1	0
3	1	0	0	1
4	0	1	1	1

Variable pairs	Variable-value combinations covered	Coverage
ab	00, 01, 10	.75
ac	00, 01, 10	.75
ad	00, 01, 11	.75
bc	00, 11	.50
bd	00, 01, 10, 11	1.0
cd	00, 01, 10, 11	1.0

100% coverage of 33% of combinations
75% coverage of half of combinations
50% coverage of 16% of combinations

NIST
National Institute of Standards and Technology

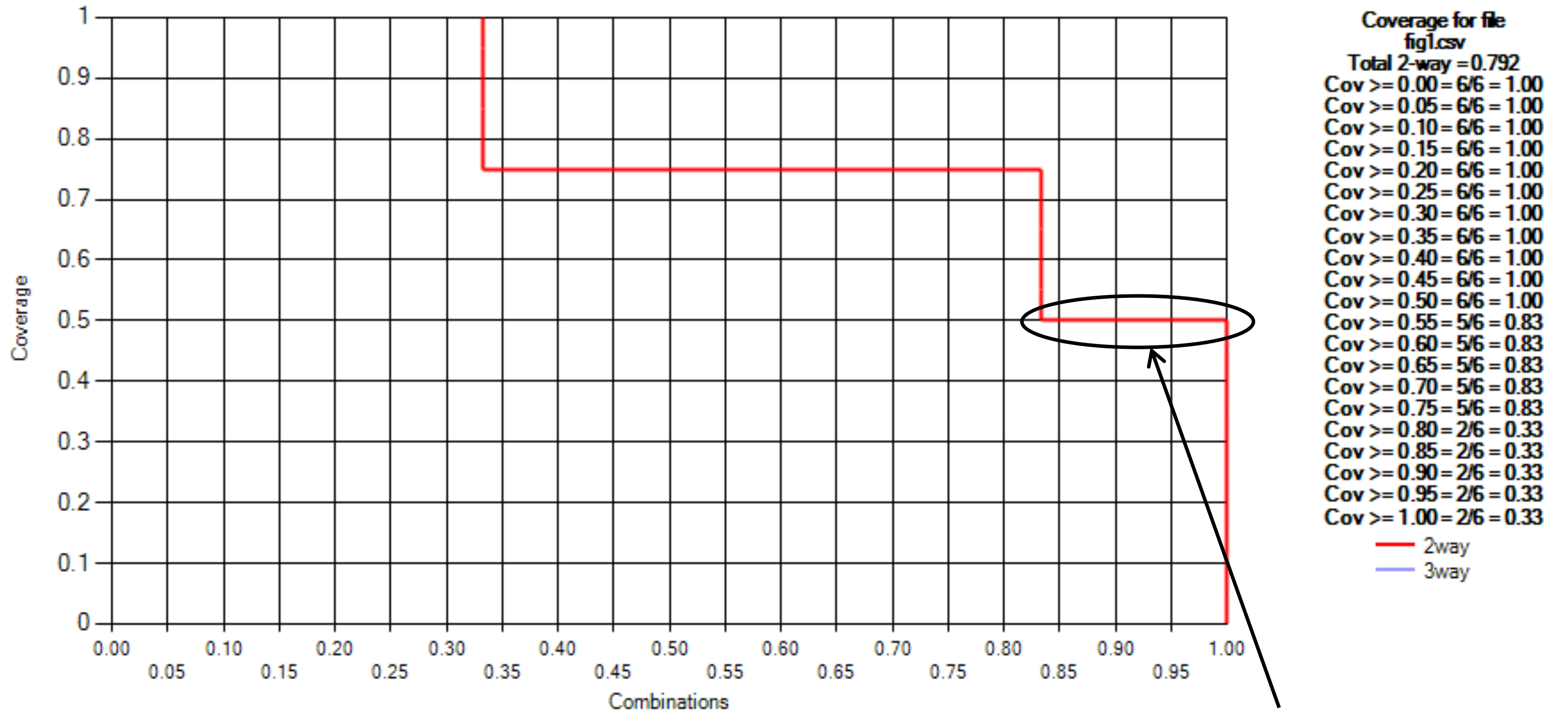
Variable pairs	Variable-value combinations covered	Coverage
ab	00, 01, 10	.75
ac	00, 01, 10	.75
ad	00, 01, 11	.75
bc	00, 11	.50
bd	00, 01, 10, 11	1.0
cd	00, 01, 10, 11	1.0

bd	00, 01, 10, 11
cd	00, 01, 10, 11
ab	00, 01, 10
ac	00, 01, 10
ad	00, 01, 11
bc	00, 11

00, 01, 10, 11	00, 01, 10, 11	00, 01, 10	00, 01, 10	00, 01, 11	00, 11
bd	*cd*	*ab*	*ac*	*ad*	*bc*

Rearranging the table

Graphing Coverage Measurement



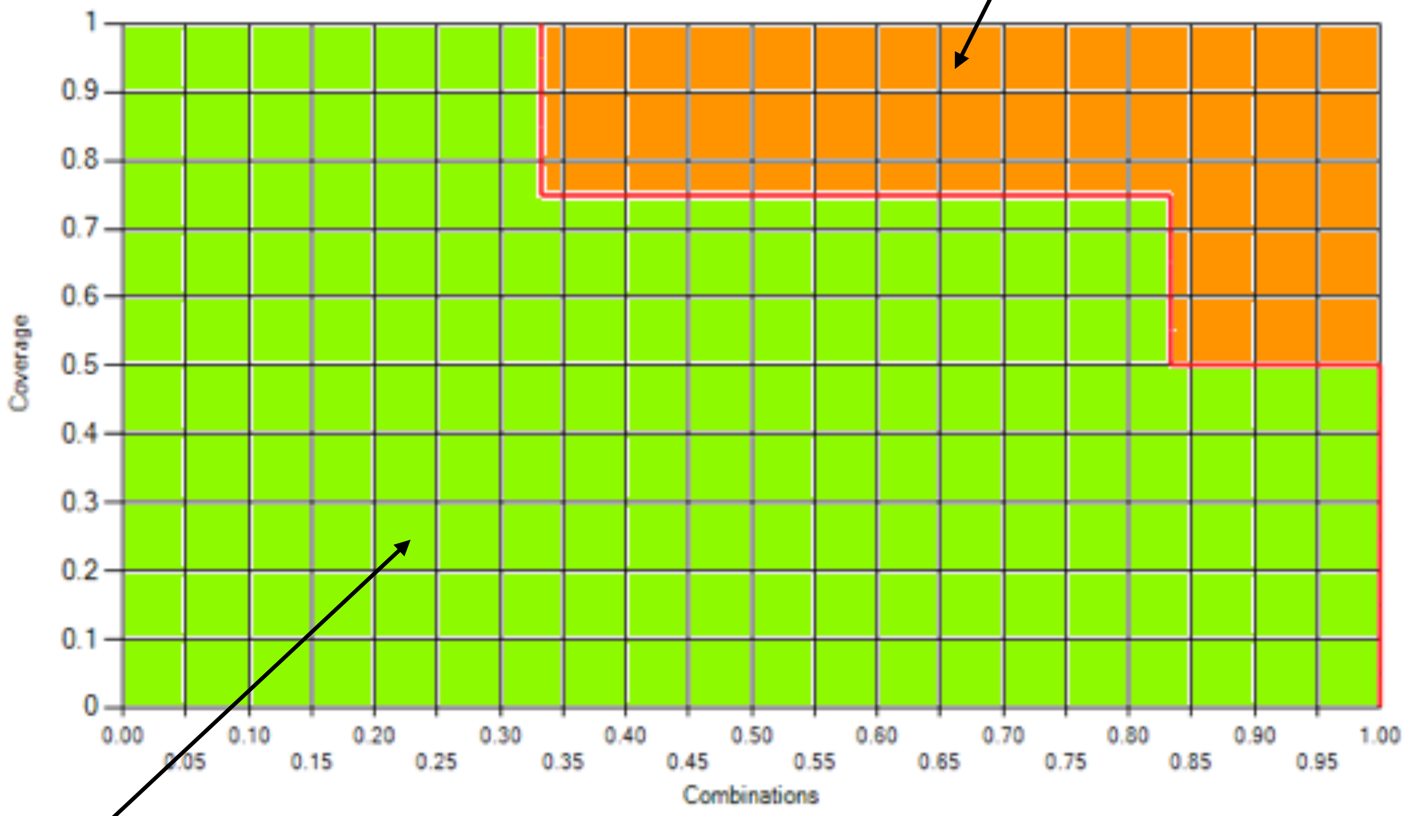
100% coverage of 33% of combinations
75% coverage of half of combinations
50% coverage of 16% of combinations

Bottom line:
All combinations covered to at least 50%

What else does this chart show?

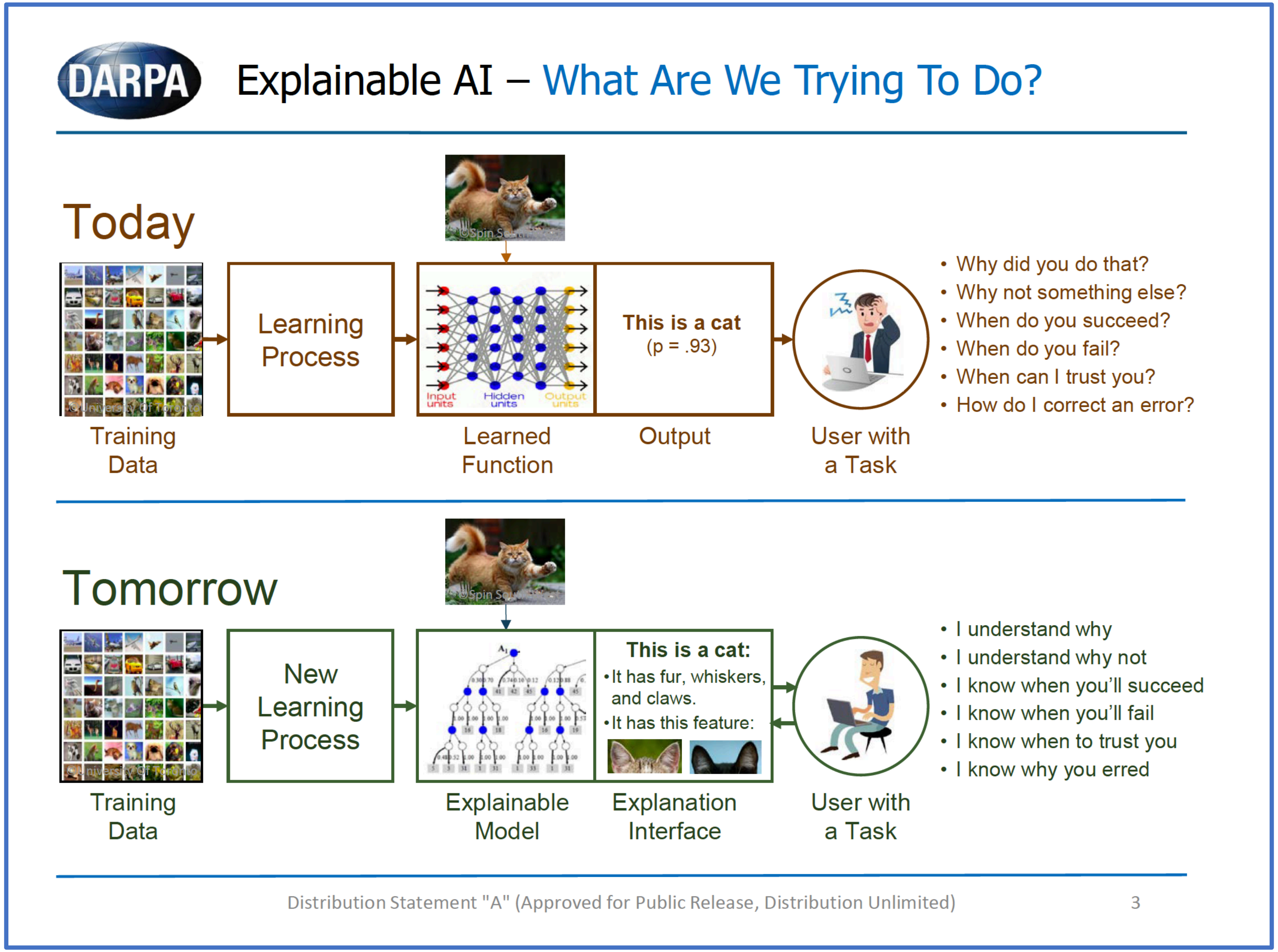
Untested combinations

(look for problems here)



Tested combinations

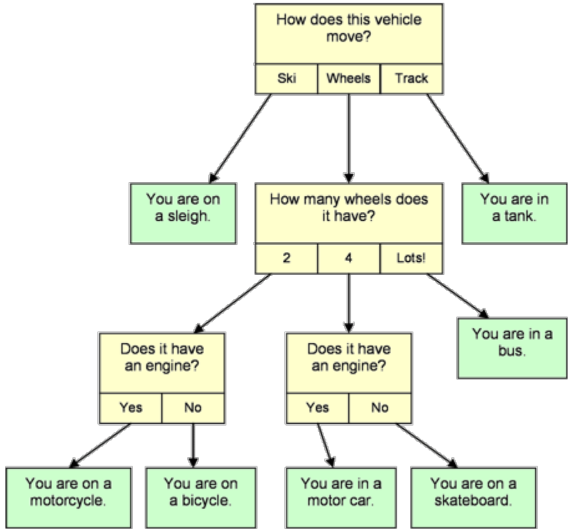
Explainability – what's current state of the art?



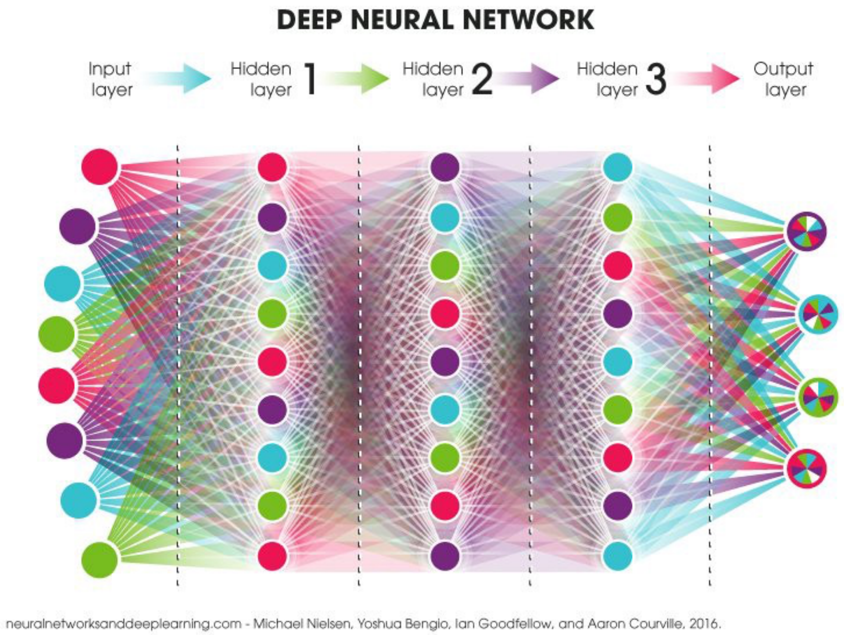
Black-box statistical predictions are inadequate

Explanations must be understandable to non-specialist

Tradeoff:



- OR -



Expert system:
Good for explanations,
not so good for accuracy

Neural nets:
Good for accuracy,
not so good for explanations

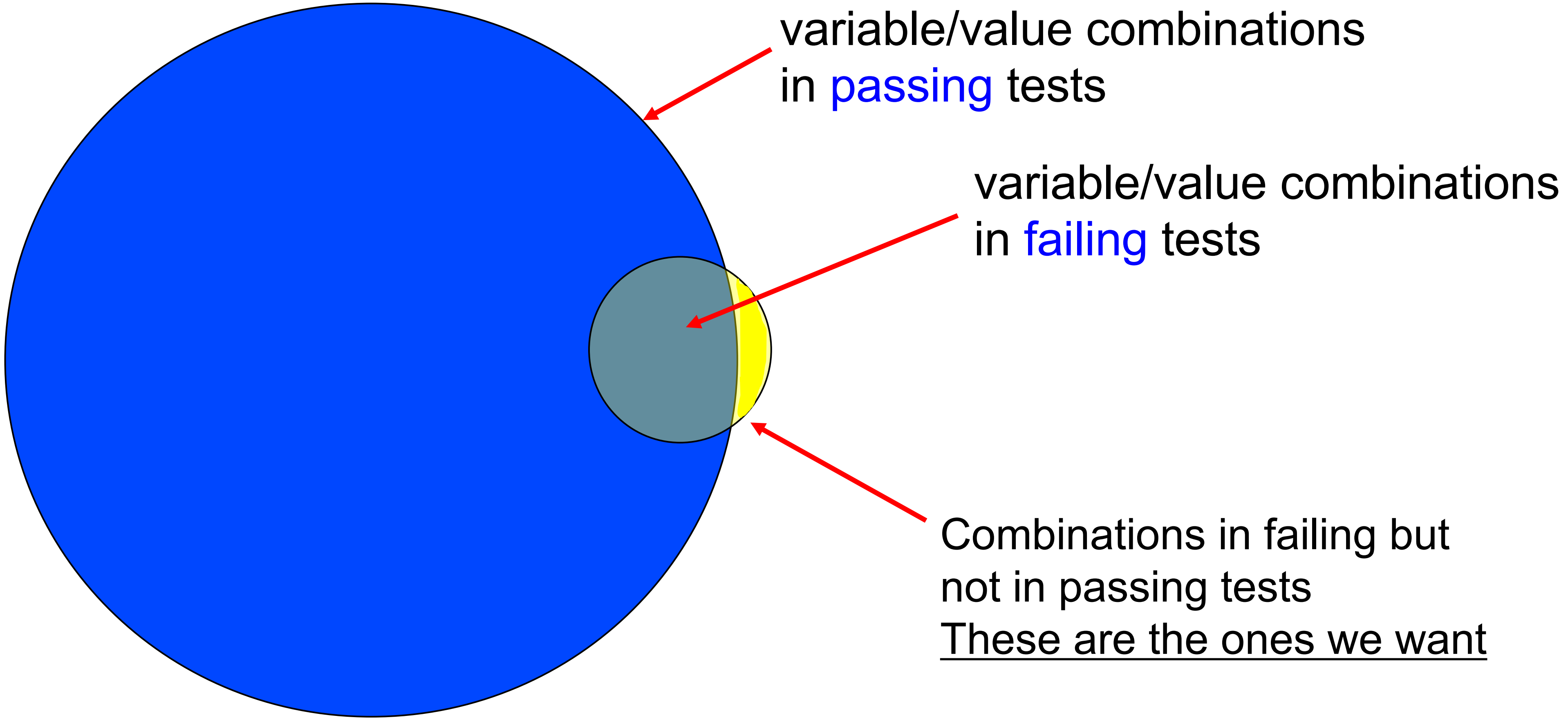
How do we get the best of both worlds?

What has been tried?

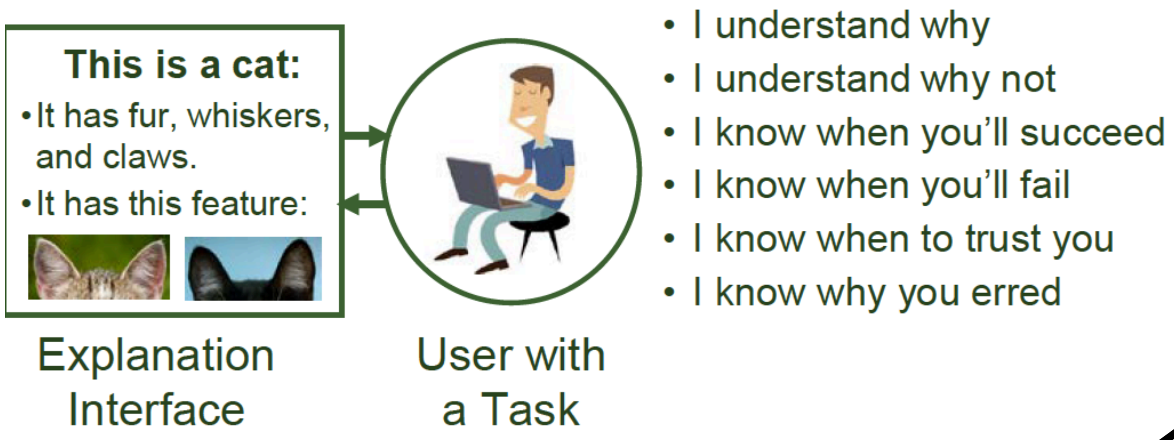
- Interpretable models – e.g. rule-based expert systems: “if patient has symptoms A and B, or has B with C and D, then illness is X”
  - best for explanations
  - hard to find rules
  - less accurate than other approaches
- Modify neural nets etc. to add explanations
  - reduces accuracy, complicates the system
  - explanations still not very understandable
- Model induction - infer explainable model from black-box
  - flexible for application, good explanations using only input, output
  - hard to produce the explainable model
- Our approach – derive rule predicates from inputs and outputs to CNNs and other black-box functions

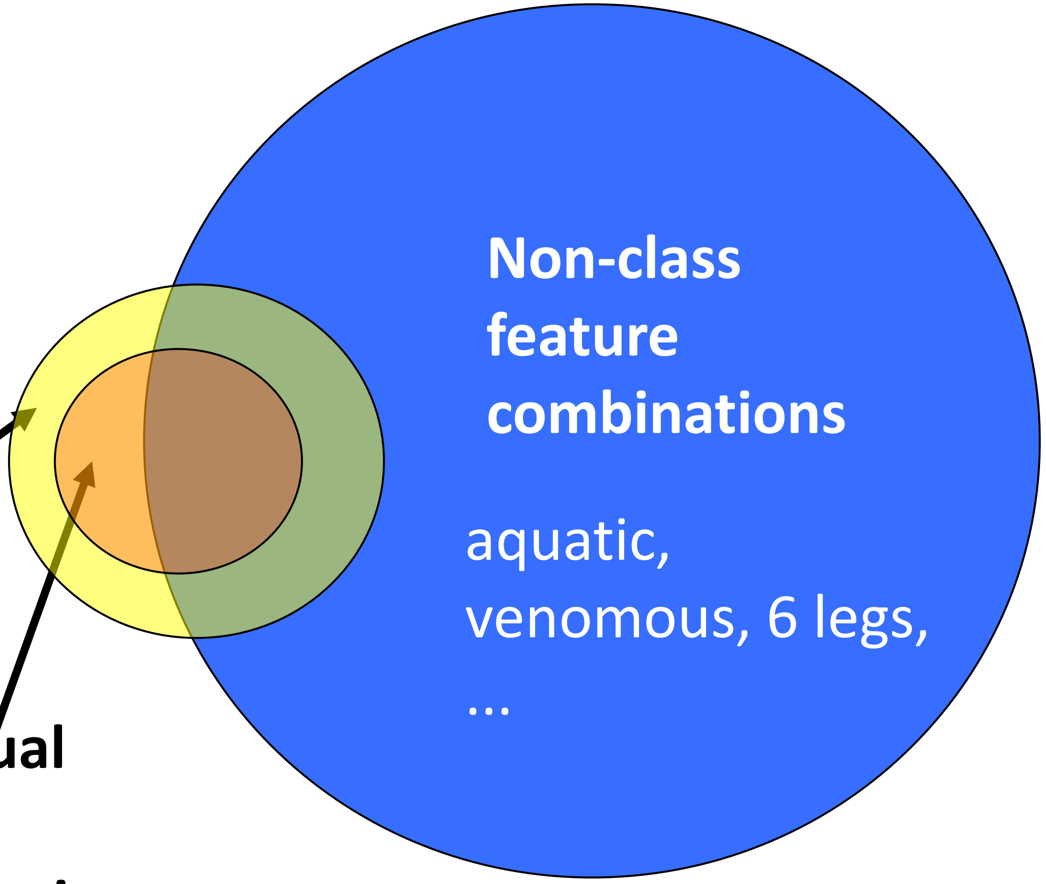
Fault location

Given: a set of tests that the SUT fails, which combinations of variables/values triggered the failure?



Relevance to explainable AI





Class feature combinations - brown & furry, black & furry, whiskers, claws, **...** not aquatic, not venomous, not 6 legs,

Individual feature combinations – brown & furry, whiskers, claws, not aquatic, not venomous, not 6 legs, **...**

Animal shares features with <u>cat</u> class

Animal does not share features with <u>non-cat</u> classes

Why is this creature recognized as a reptile?

Input configuration $2^{15}6^1$

Class File:	Class file rep1.csv; rows=1; cols=16
Nominal File:	Nominal file notreptile.csv; rows=96; cols=16 || 2-way: 120 3-way: 560 4-way: 1,820 5-way: 4,368 6-way: 8,008
Class File Contents:	

hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	nlegs	tail	domestic	catsize
0	0	1	0	0	0	0	0	1	1	0	0	4	1	0	1

```
----------------
0053 occurrences = 0.552 of cases, hair = 0
0076 occurrences = 0.792 of cases, feathers = 0
0055 occurrences = 0.573 of cases, eggs = 1
0055 occurrences = 0.573 of cases, milk = 0
0072 occurrences = 0.750 of cases, airborne = 0
0061 occurrences = 0.635 of cases, aquatic = 0
0044 occurrences = 0.458 of cases, predator = 0
0039 occurrences = 0.406 of cases, toothed = 0
0078 occurrences = 0.813 of cases, backbone = 1
0076 occurrences = 0.792 of cases, breathes = 1
0090 occurrences = 0.938 of cases, venomous = 0
0079 occurrences = 0.823 of cases, fins = 0
0036 occurrences = 0.375 of cases, nlegs = 4
0070 occurrences = 0.729 of cases, tail = 1
0083 occurrences = 0.865 of cases, domestic = 0
0043 occurrences = 0.448 of cases, catsize = 1
```

No single feature is sufficient explanation – shares features with non-reptiles

No pair of features sufficient – shares 2-way combinations w/ non-reptiles

```
0002 occurrences = 0.021 of cases, toothed,nlegs = 0,4
0005 occurrences = 0.052 of cases, hair,nlegs = 0,4
0005 occurrences = 0.052 of cases, milk,nlegs = 0,4
0006 occurrences = 0.063 of cases, eggs,nlegs = 1,4
0008 occurrences = 0.083 of cases, toothed,catsize = 0,1
0011 occurrences = 0.115 of cases, milk,catsize = 0,1
0012 occurrences = 0.125 of cases, eggs,catsize = 1,1
0013 occurrences = 0.135 of cases, hair,catsize = 0,1
0015 occurrences = 0.156 of cases, predator,catsize = 0,1
```

3-way combinations produce rules to explain recognition of Testudo as a reptile

```
00000 occurrences = 0.000 of cases, aquatic,toothed,nlegs = 0,0,4
00000 occurrences = 0.000 of cases, eggs,aquatic,nlegs = 1,0,4
00000 occurrences = 0.000 of cases, hair,aquatic,nlegs = 0,0,4
00000 occurrences = 0.000 of cases, hair,nlegs,catsize = 0,4,1
00000 occurrences = 0.000 of cases, milk,aquatic,nlegs = 0,0,4
00000 occurrences = 0.000 of cases, milk,nlegs,catsize = 0,4,1
00000 occurrences = 0.000 of cases, predator,toothed,nlegs = 0,0,4
00001 occurrences = 0.010 of cases, eggs,nlegs,catsize = 1,4,1
00001 occurrences = 0.010 of cases, eggs,predator,nlegs = 1,0,4
00001 occurrences = 0.010 of cases, feathers,toothed,backbone = 0,0,1
```

Non-reptiles in the database do not have these 3-way combinations

Only reptiles have these combinations of features:

not aquatic AND not toothed AND four legs

egg-laying AND not aquatic AND four legs

not hairy AND four legs AND cat size

not milk-producing AND not aquatic AND four legs

not milk-producing AND four legs AND cat size

not predator AND not toothed AND four legs

Mapping combinations to expressions

- Report identifies t-way combinations that distinguish the predicted class from others
- Combinations can be mapped to expressions to produce a rule-based type of explanation

if (not aquatic AND not toothed AND four legs)
OR (egg-laying AND not aquatic AND four legs)
OR (not hairy AND four legs AND cat size)
OR (not milk-producing AND not aquatic AND four legs)
OR (not milk-producing AND four legs AND cat size)
OR (not predator AND not toothed AND four legs)
then reptile;
else not reptile;

As noted, none of the single factors above is sufficient for explanation

Example: empty vs. occupied rooms, using sensor data

File Information

Class File:	Class file o1.csv; rows=1; cols=5
Nominal File:	Nominal file empty.csv; rows=7703; cols=5 || 2-way: 10 3-way: 10 4-way: 5 5-way: 1 6-way: 0
Class File Contents:	Temperature Humidity Light CO2 HumidityRatio B3 B3 B2 B2 B4

2-Way | 3-Way | 4-Way | 5-Way | 6-Way

Enabled

```
Combinations = 10, Settings = 210

0016 occurrences = 0.002 of cases, Humidity,Light = B3,B2
0016 occurrences = 0.002 of cases, Light,CO2 = B2,B2
0036 occurrences = 0.005 of cases, Temperature,Light = B3,B2
0040 occurrences = 0.005 of cases, CO2,HumidityRatio = B2,B4
0043 occurrences = 0.006 of cases, Light,HumidityRatio = B2,B4
0054 occurrences = 0.007 of cases, Temperature,CO2 = B3,B2
0078 occurrences = 0.010 of cases, Humidity,CO2 = B3,B2
0205 occurrences = 0.027 of cases, Temperature,HumidityRatio = B3,B4
0247 occurrences = 0.032 of cases, Temperature,Humidity = B3,B3
0495 occurrences = 0.064 of cases, Humidity,HumidityRatio = B3,B4
---------------
0523 occurrences = 0.068 of cases, Temperature = B3
2415 occurrences = 0.314 of cases, Humidity = B3
0085 occurrences = 0.011 of cases, Light = B2
0534 occurrences = 0.069 of cases, CO2 = B2
2190 occurrences = 0.284 of cases, HumidityRatio = B4
```

Why do we conclude this room is occupied?

These levels of humidity and lighting are strong indication

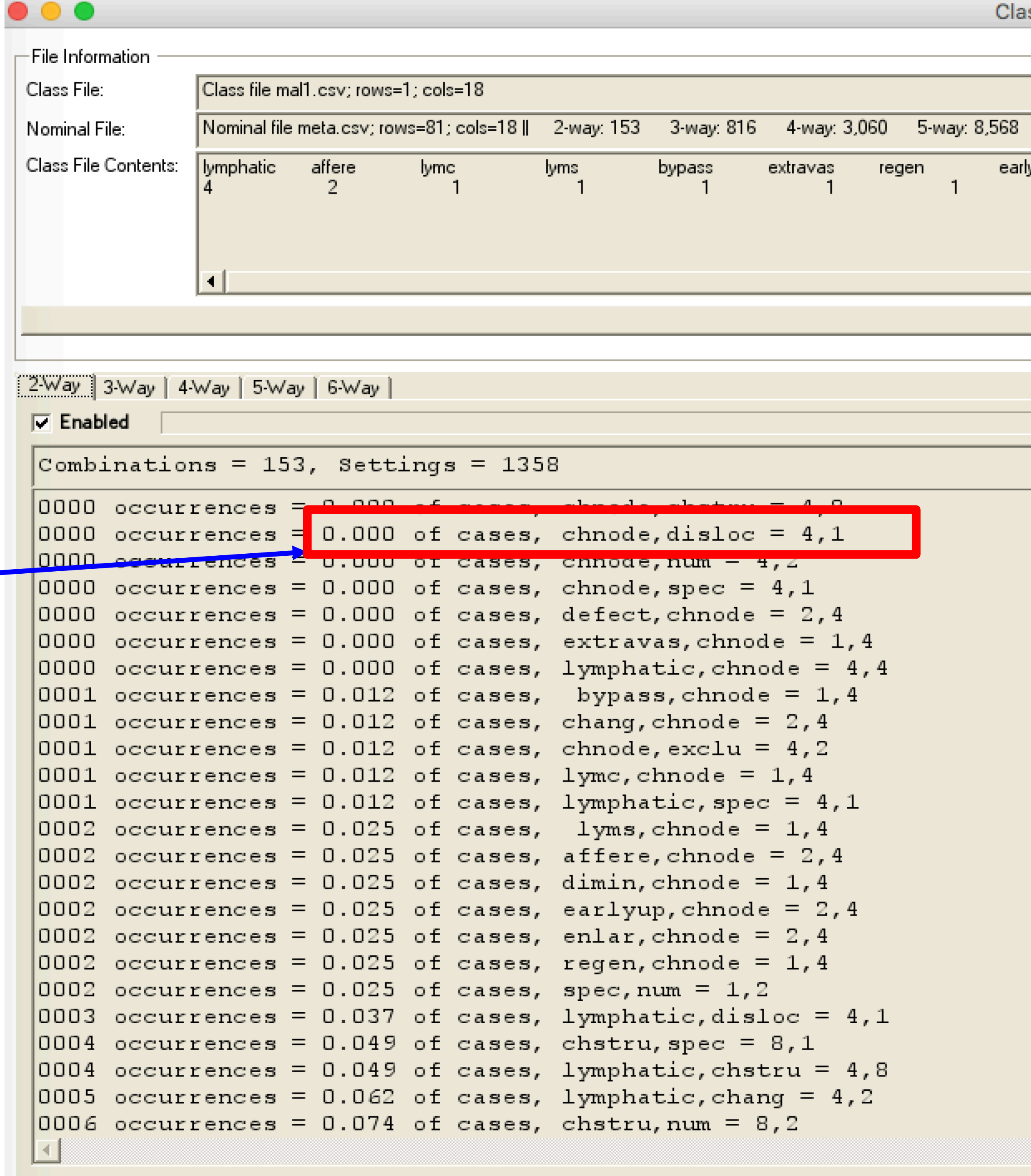
Considering levels of lighting, CO2, and humidity ratio provide even stronger evidence:

<u>Empty rooms don’t have these levels</u>

```
00003 occurrences = 0.000 of cases, Light,CO2,HumidityRatio = B2,B2,B4
00005 occurrences = 0.001 of cases, Humidity,Light,CO2 = B3,B2,B2
00008 occurrences = 0.001 of cases, Temperature,Light,CO2 = B3,B2,B2
00011 occurrences = 0.001 of cases, Humidity,Light,HumidityRatio = B3,B2,B4
```

A different example: lymph node pathology – why is this classified as malignant not metastatic?

- These combinations are characteristic of lymphoma that arises in lymph node instead of metastatic that spread to node from somewhere else



Obvious question – Can we use these methods for prediction as well as explanation?

- Maybe, but consider:

Summary

- Combinatorial methods can provide explainable AI
- We have prototype that applies this approach
  - Determine combinations of variable values that differentiate an example from other possible conclusions
    - ➔ Feature combinations present shared with class
    - ➔ Feature combinations not shared with class not present
- Method can be applied to black-box functions such as CNNs
- Present explanation in the preferred form of rules,
  "if A & B, or C with D & E, then conclusion is X"

Please contact us if you're interested!

http://csrc.nist.gov/acts

Rick Kuhn
kuhn@nist.gov

Raghu Kacker
raghu.kacker@nist.gov

Jeff Lei, University of Texas at Arlington
Dimitris Simos, SBA Research
M S Raunak, Loyola University