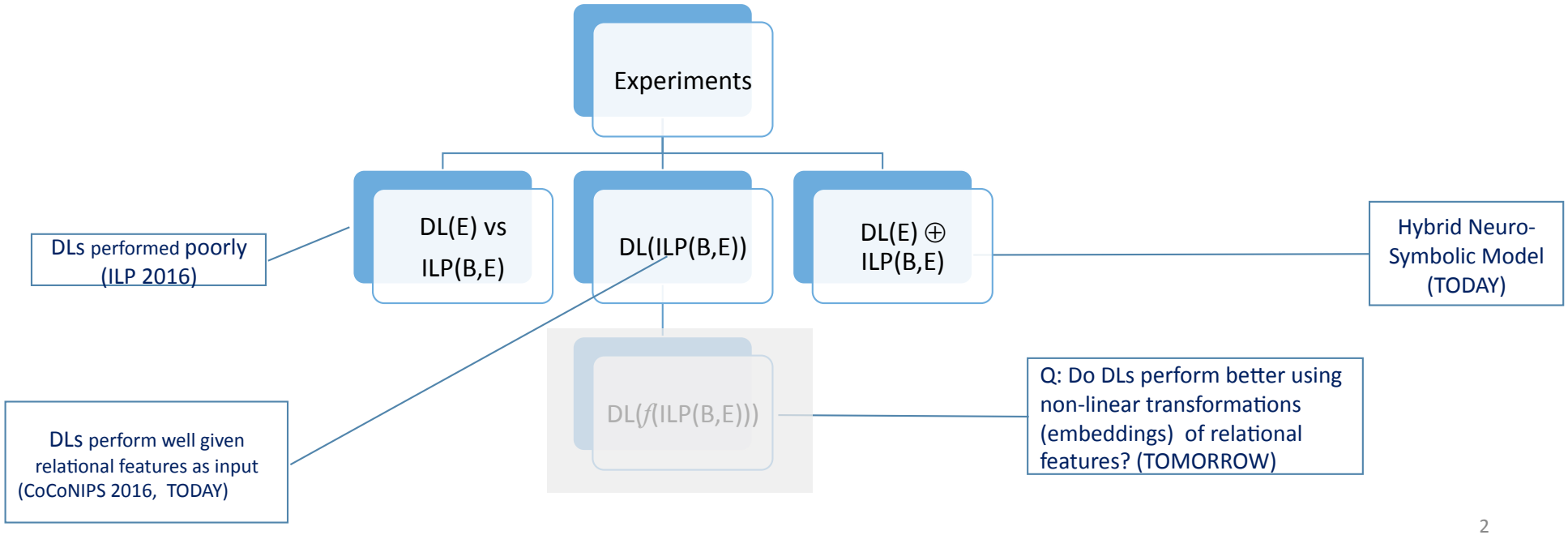
Mode-directed Neural-Symbolic Modelling

Ashwin Srinivasan

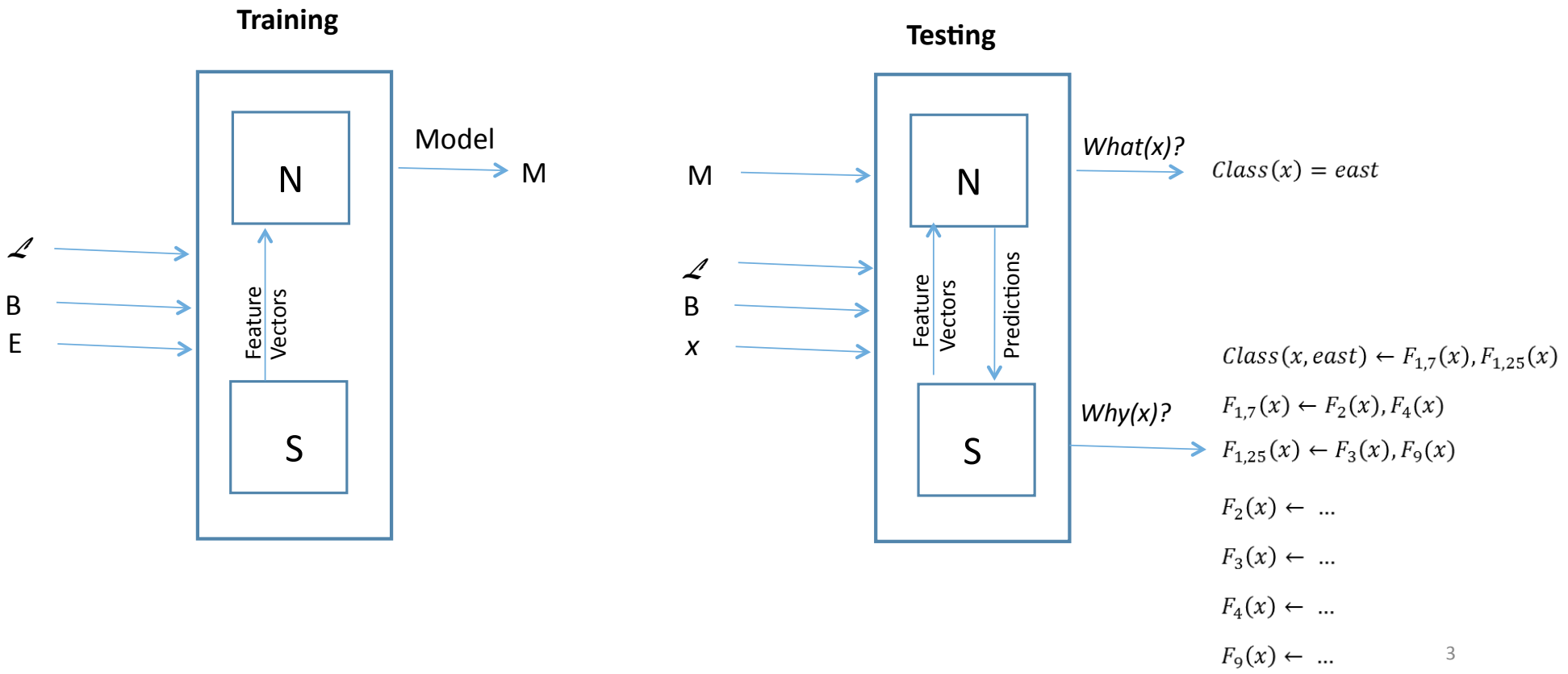
(joint work with Lovekesh Vig and Gautam Shroff)

The Big Picture

- This is part of a broader investigation on Deep Learning and ILP
  - Specific interest: small amounts of (relational) data E, significant domain-knowledge B



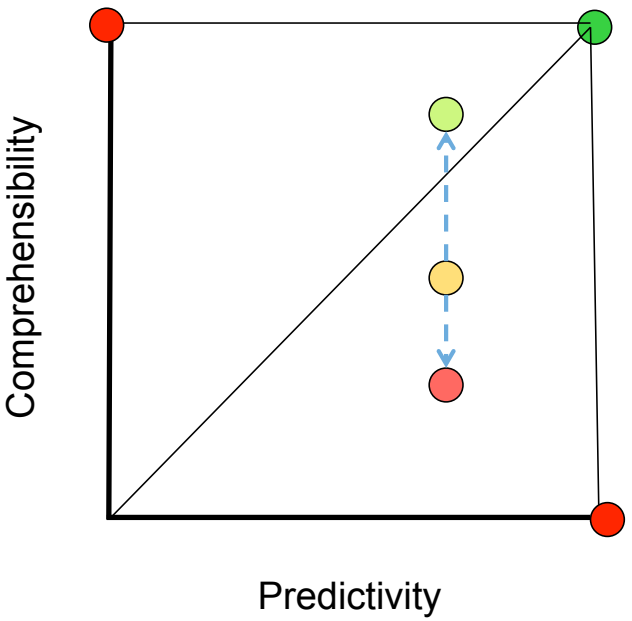
Almost The Full Story



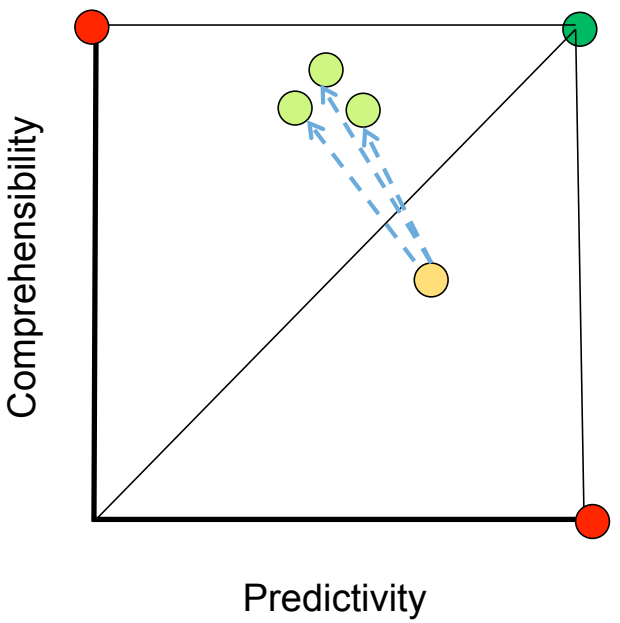
An explanatory model

Explanatory Models from Predictive Ones

Globally Consistent Explanatory Model

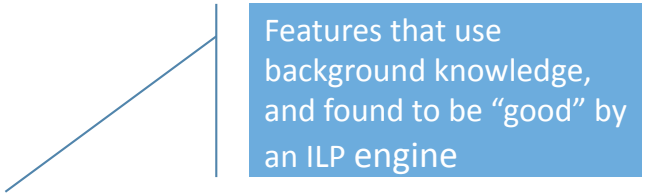


Locally Consistent Explanatory Models



Step 1: Getting a Good Predictive Model

- We have seen that:
  - A deep network with ILP-features in the last layer predicts better than a deep network that does not have such features (CoCoNIPS 2016)
  - So, can't we use this "ILP-assisted deep network" as the predictive model?
- Yes, BUT:
  - In our experiments ILP-assisted deep network model in fact were worse predictors than vanilla ILP models (WHY?)
  - All we could say: ILP-assisted deep networks better than non-assisted deep networks
    - Q: "Can DL benefit from ILP?. A: YES
- So, are the ILP models the best we can do?



What do the relational features look like?

$F_1(x) = 1$ if $(Train(x) \wedge HasCar(x, y) \wedge Short(y))$ and 0 otherwise

$F_2(\mathrm{x}) = 1$ if $\big(Train(x) \wedge HasCar(x, y) \wedge \mathrm{Short(y)} \wedge Closed(y)\big)$ and 0 otherwise

$F_3(x) = 1$ if $\big(Train(x) \wedge HasCar(x, y) \wedge HasCar(x, z) \wedge InFront(y, z) \wedge Short(\mathrm{y}) \wedge Long(z)\big)$ and 0 otherwise

We will be giving values of such features for instances ("propositionalising")

Step 1 (contd.): DNs with relational features

- A DN with relational features as input:
  - Interesting features using an ILP engine: input layer of a DN or in the top-layer of a DN
  - HERE: Do not pre-select using ILP. Give all possible first-order features and let DNs work out what's useful
- Small print:
  - Use a depth-bounded mode language
  - Draw randomly from feature-space
  - Do not want to draw irrelevant features

    $(Train(x), HasCar(x,y), Short(y), Long(y))$
  - Avoid drawing redundant features

Repeat:

1. Draw example
2. Construct bottom clause
3. Draw clause
4. Check (subsumption) equivalence
5. Construct feature

How good are the predictive models?

Predictivity is usually higher than:

- ILP-only models; and
- DNs with ILP-selected features as input

Use first-order features that require domain-knowledge to compute their values

Problem	Accuracy			
	OptILP [26]	*Stat* [20]	*DRM* [12]	*KRDN*
*Mut*188	0.88(0.02)	0.85(0.05)	0.90(0.06)	0.91(0.06)
*Canc*330	0.58(0.03)	0.60(0.02)	–	0.68(0.03)
DssTox	0.73(0.02)	0.72(0.01)	0.66(0.02)	0.70(06)
Amine	0.80(0.02)	0.81(0.00)	–	0.89(0.04)
Choline	0.77(0.01)	0.74(0.00)	–	0.81(0.03)
Scop	0.67(0.02)	0.72(0.02)	–	0.80(0.05)
Toxic	0.87(0.01)	0.84(0.01)	–	0.93(0.03)

Step 2: Explanatory models

- Explanatory Model = Symbolic Model
  - Not necessary that Symbolic Model = Comprehensible Model

- **Globally Consistent**: Symbolic model that results in the same output as the predictive model on all training instances

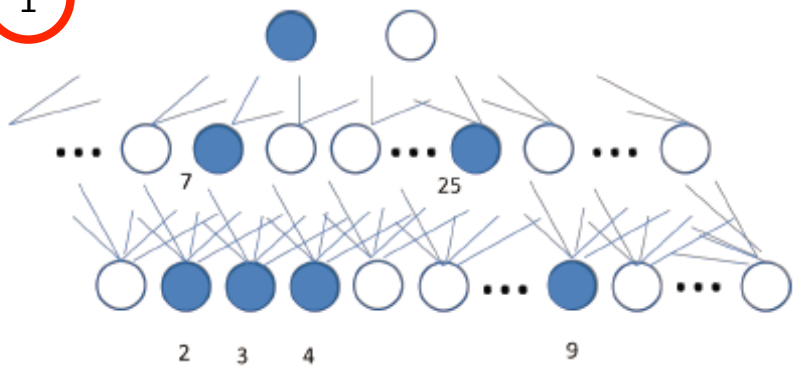
- **Locally Consistent**: Symbolic model that makes the same prediction as the predictive model on a query instance (and it's friends)

- BUT:
  - It may not be possible to obtain a single symbolic model that is either globally or locally consistent for all instances
  - Even if such a single symbolic model exists, it may be too complex to be comprehensible

Locally consistent explanations: Why?

- We think a locally consistent model will be simpler, and therefore more comprehensible

1



For any given instance, only a small number of units are “active”

2

$$F_{1,7}(x) \leftarrow F_2(x), F_4(x)$$
$$F_{1,25}(x) \leftarrow F_3(x), F_9(x)$$

Hidden active units can represent intermediate predicates (based on activations)

3

$$Class(x, east) \leftarrow F_{1,7}(x), F_{1,25}(x)$$
$$F_{1,7}(x) \leftarrow F_2(x), F_4(x)$$
$$F_{1,25}(x) \leftarrow F_3(x), F_9(x)$$

An explanation

Locally consistent explanations: How?

(Like Ribeiero *et al.*'s LIME for linear models)

1. Get the feature-vector representation of the query instance *x*
2. Get the feature-vector representation of nearby (training) instances
3. Find the predictions made by the predictive model for all these instances
4. Find the symbolic model that is consistent with these predictions (at least consistent with the prediction for *x* and as many of the nearby predictions)

- Step 4 can be tackled like a usual Progol-like search, with a most specific clause for *x* from Step 1
- Using a deep network for prediction allows us a further step:

5. Transform the model in (4) into a structured form (using active hidden nodes)

$$Class(x, east) \leftarrow F_2(x), F_3(x), F_4(x), F_9(x)$$

Folding ⇢

⇠ Unfolding

$$Class(x, east) \leftarrow F_{1,7}(x), F_{1,25}(x)$$
$$F_{1,7}(x) \leftarrow F_2(x), F_4(x)$$
$$F_{1,25}(x) \leftarrow F_3(x), F_9(x)$$

How good are the explanatory models?

- By construction, an explanation will be consistent on any query instance *x*.
- But, how consistent is it on nearby instances *Nbd(x)* ?

$$Fidelity(H) = \frac{|AgreePos\big(Nbd(x)\big) + AgreeNeg\big(Nbd(x)\big)|}{|PredPos(Nbd(x) + PredNeg(Nbd(x))|}$$

Problem	Fidelity	
	H_5	H_{10}
$Mut188$	0.86(0.05)	0.86(0.04)
$Canc330$	0.73(0.09)	0.73(0.09)
$DssTox$	0.92(0.04)	0.85(0.03)
$Amine$	0.87(0.03)	0.83(0.04)
$Choline$	0.79*0.02)	0.73(0.03)
$Scop$	0.75(0.03)	0.70(0.02)
$Toxic$	0.81(0.02)	0.77(0.02)

- About 10 training instances in H5
- About 30 training instances in H10

As neigbourhood shrinks fidelity increases

How to get a structured explanation

Nodes activated in Layer 0
F7, F39, F62, F63

Weights from Layer 0 to Layer 1

	F1,1	F1,3	F1,4	F1,5
F7	0.17	0.06	-0.06	-0.08
F39	-0.14	0.17	-0.04	0.00
F62	0.14	0.17	0.03	0.05
F63	-0.03	0.02	0.06	-0.12

Weights from Layer 1 to Layer 2

	F2,2	F2,4	F2,6
F1,1	0.38	0.33	0.06
F1,3	0.34	0.35	0.04
F1,4	0.16	-0.63	0.02
F1,5	0.09	-0.22	0.25

(and so on)

Basic principles:

1. In any layer, any node with activation > α is an "active" node
2. Between layers, only retain edges between active nodes with weights above some threshold β
3. Active nodes in layer (i) "partitioned" to active nodes in layer (i+1)
4. Single-clause definition for each predicate
5. Fold, check for redundancies

Example: $\alpha = 0$ $\beta = 0.1$

$F_{1,1} \leftarrow F_7, F_{62}$

$F_{1,3} \leftarrow F_7, F_{39}, F_{62}, F_{63}$ $\quad F_{1,3} \leftarrow F_{1,1}, F_{1,4}$

$F_{1,4} \leftarrow F_{62}, F_{63}$

$F_{1,5} \leftarrow F_{39}, F_{62}$

$F_{2,2} \leftarrow F_{1,1}, F_{1,3}, F_{1,4}, F_{1,5}$ $\quad F_{2,2} \leftarrow F_{1,1}, F_{1,4}, F_{1,5}$

$F_{2,4} \leftarrow F_{1,1}, F_{1,3}$

$F_{2,6} \leftarrow F_{1,1}, F_{1,3}, F_{1,4}, F_{1,5}$ $\quad F_{2,6} \leftarrow F_{2,2}$

Concluding Remarks

- "It's life, Jim, but not as we know it"
  - Neural models aren't quite like the usual DN models
  - Symbolic models aren't quite like the usual ILP models

- BUT:
  - We are able to get highly predictive models by using a DN with randomly drawn features that are: (a) defined within a depth-bounded mode language; and (b) use domain knowledge
  - We are able to get high-fidelity explanatory models using the predictions of the DN, which can be rewritten to contain intermediate predicates guided by activations in the DN

- Interesting echoes to work by Michie on behavioural cloning, where executable models and explanatory models need not be the same

Acknowledgements

- TCS Research Lab
- SERB EMR/2016/002766

- Michael Bain, UNSW