Opening the black-box: On the interpretability of machine learning models for machine listening

SAUMITRA MISHRA, BOB L. STURM AND SIMON DIXON

HORSES 2017 WORKSHOP, LONDON

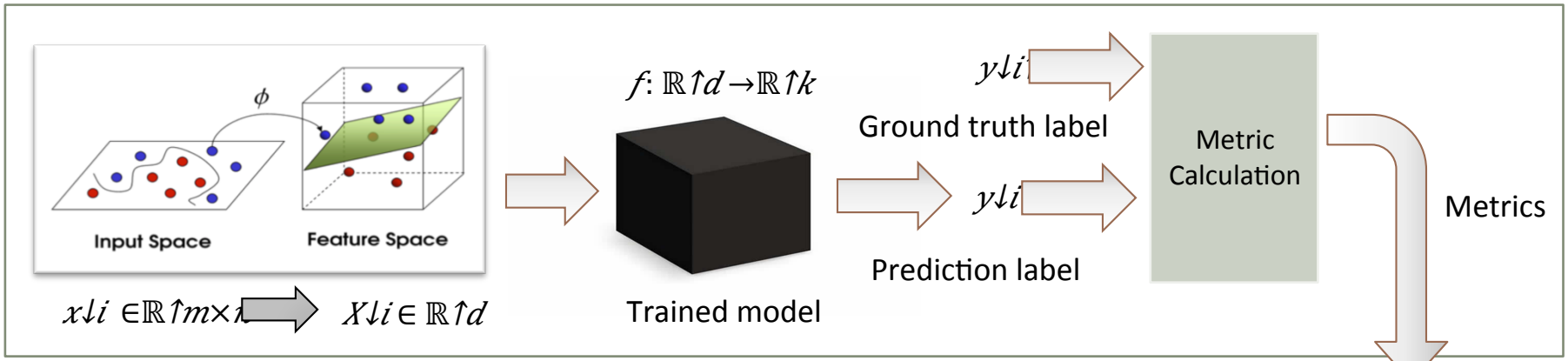
SEPTEMBER 2017

Outline

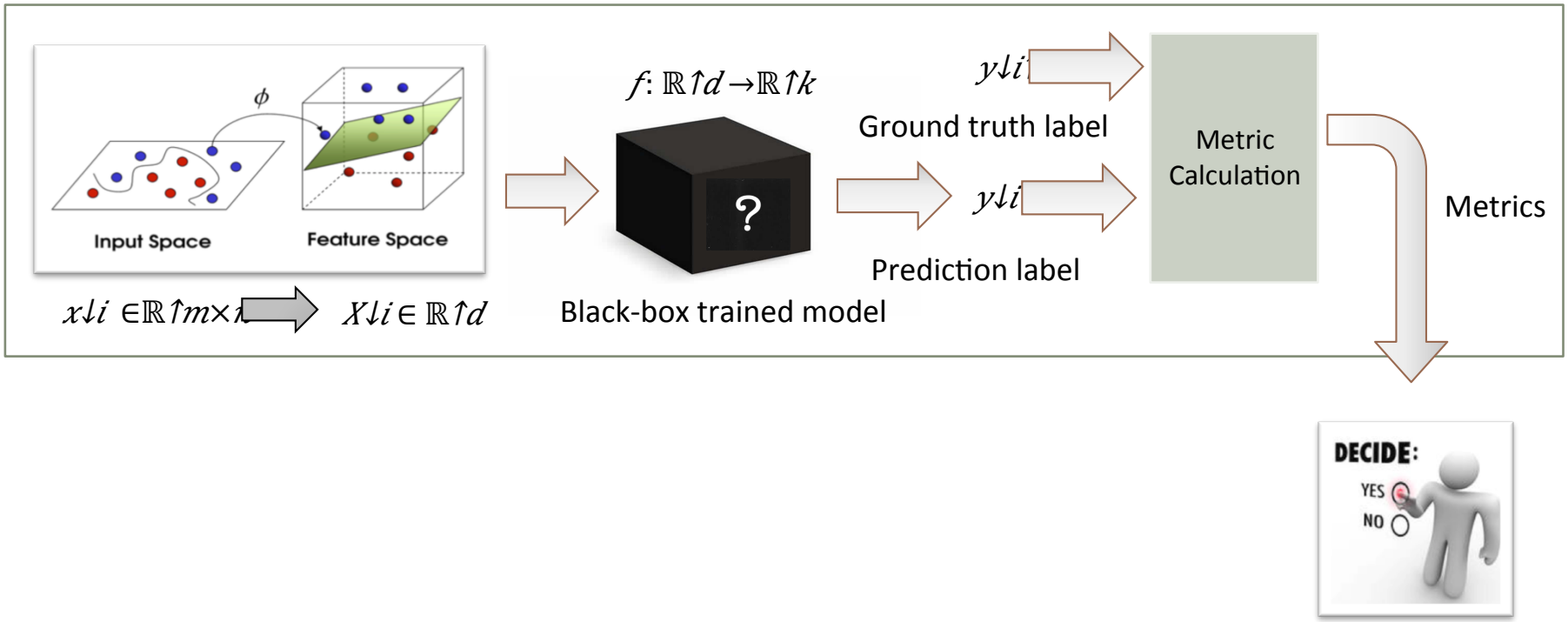
- What is interpretable machine learning?
- Why do we need machine learning models to be interpretable?
- How can we achieve interpretability?

What is interpretable machine learning?

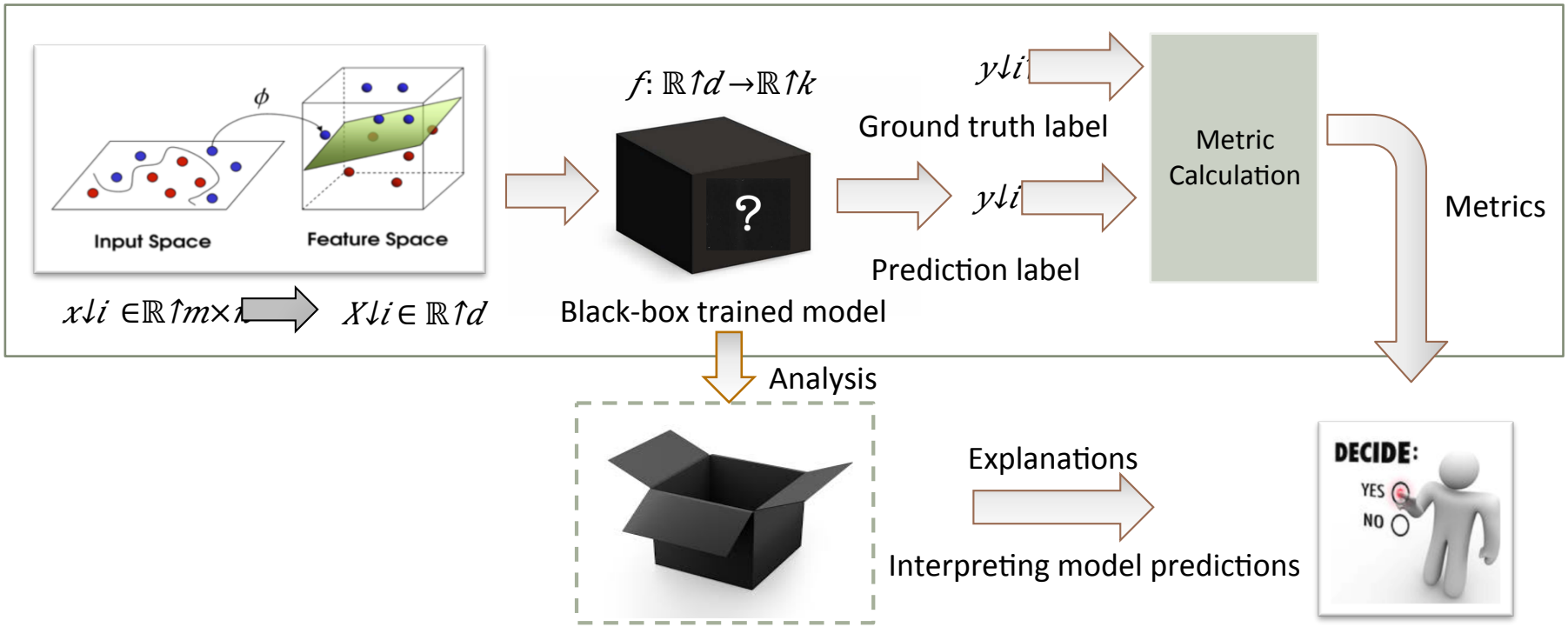
Supervised Learning Pipeline



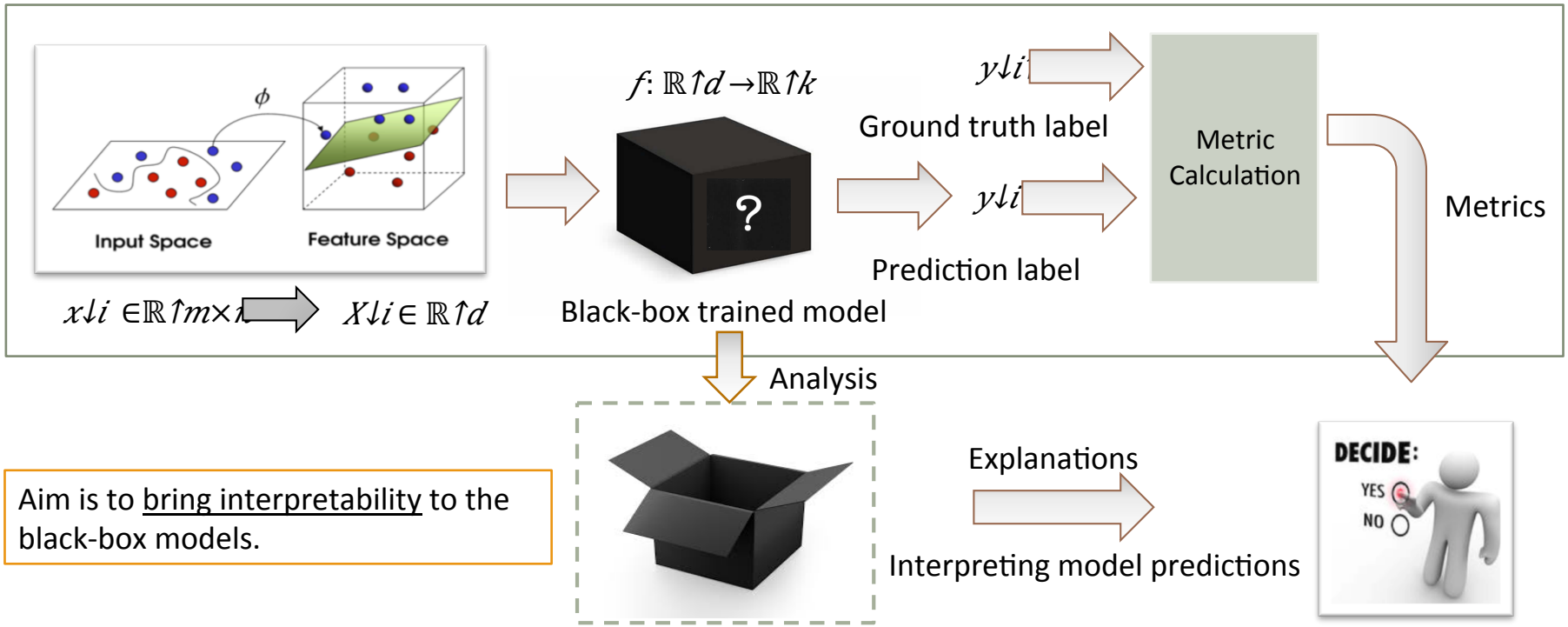
Supervised Learning Pipeline



Supervised Learning Pipeline



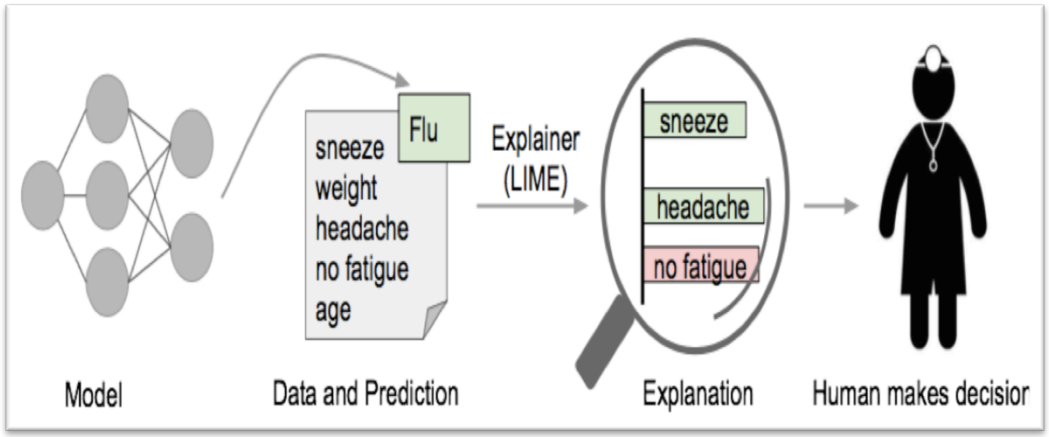
Supervised Learning Pipeline



Aim is to <u>bring interpretability</u> to the black-box models.

An interpretable machine learning model allows a system designer/ end user to understand the mechanism by which it forms its predictions.

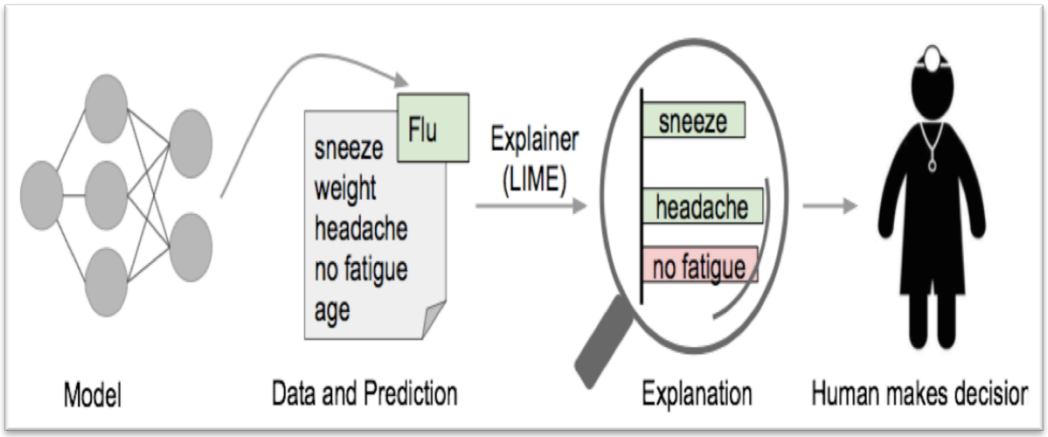
An interpretable machine learning model allows a system designer/ end user to understand the mechanism by which it forms its predictions.



Local

Ribeiro et al., "Why Should I Trust you? : Explaining the Predictions of Any Classifier", in Proc. KDD, 2016.

An interpretable machine learning model allows a system designer/ end user to understand the mechanism by which it forms its predictions.



Local

Global

Ribeiro et al., "Why Should I Trust you? : Explaining the Predictions of Any Classifier", in Proc. KDD, 2016.

Nguyen et al., "Synthesizing the preferred inputs for neurons in neural networks via deep network generators ", in Proc. NIPS, 2016.

Why is model interpretability crucial?

Why we need interpretability?

- **Trust**
  - Applying machine learning and AI within many domains requires transparency and responsibility
    - Health care
    - Finance
    - Autonomous vehicles

- *Ribeiro et al., " Model-Agnostic Interpretability of Machine Learning", in Proc. ICML Workshop on Human Interpretability in Machine Learning, 2016.*
- *Z. C. Lipton, "The Mythos of Interpretability", in Proc. ICML Workshop on Human Interpretability in Machine Learning, 2016.*

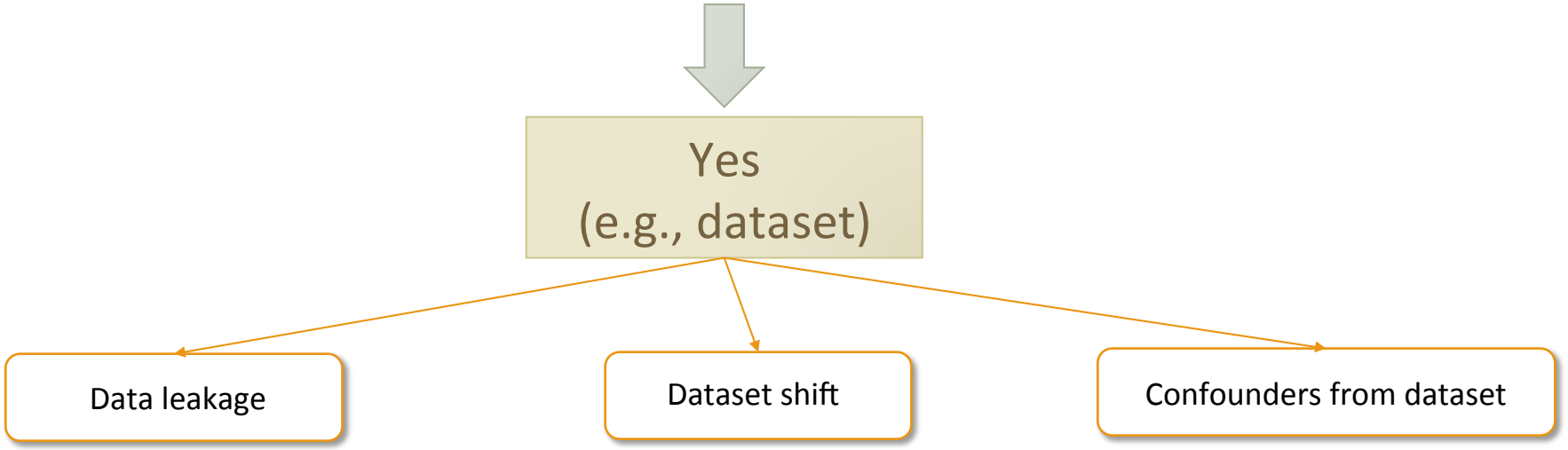
Why we need interpretability?

- **Trust**
  - Applying machine learning and AI within many domains requires transparency and responsibility
    - Health care
    - Finance
    - Autonomous vehicles
  - Is the model giving the right answers for the right reasons?
    - *Sturm, "A simple method to determine if a music information retrieval system is a 'horse'," in IEEE Tran. Mult., 2014.*
    - 'A "horse" is just a system that is not actually addressing the problem it appears to be solving'.

- *Ribeiro et al., " Model-Agnostic Interpretability of Machine Learning", in Proc. ICML Workshop on Human Interpretability in Machine Learning, 2016.*
- *Z. C. Lipton, "The Mythos of Interpretability", in Proc. ICML Workshop on Human Interpretability in Machine Learning, 2016.*

Can a model or its evaluation go wrong?

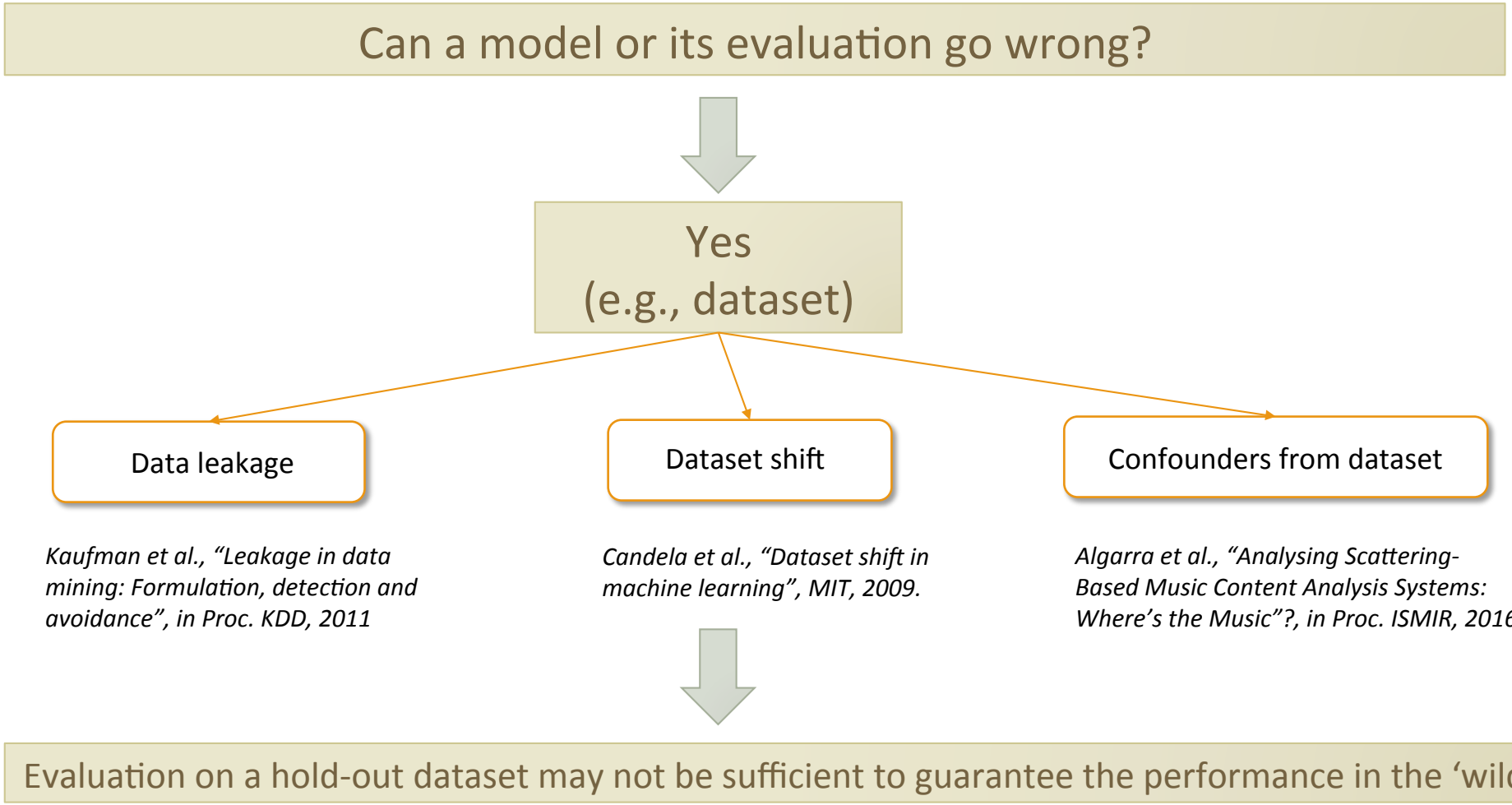
Can a model or its evaluation go wrong?



Kaufman et al., "Leakage in data mining: Formulation, detection and avoidance", in Proc. KDD, 2011

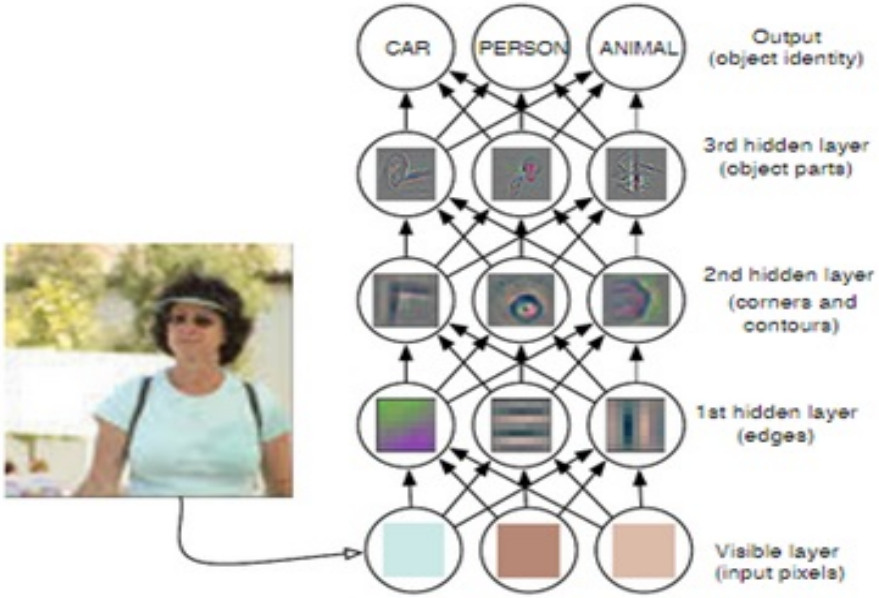
Candela et al., "Dataset shift in machine learning", MIT, 2009.

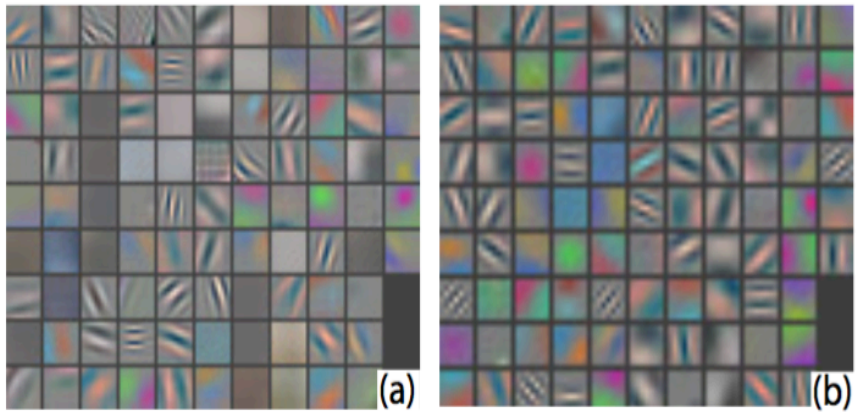
Algarra et al., "Analysing Scattering-Based Music Content Analysis Systems: Where's the Music"?, in Proc. ISMIR, 2016


Can a model or its evaluation go wrong?
Yes
(e.g., dataset)
Data leakage
Dataset shift
Confounders from dataset
Kaufman et al., "Leakage in data mining: Formulation, detection and avoidance", in Proc. KDD, 2011
Candela et al., "Dataset shift in machine learning", MIT, 2009.
Algarra et al., "Analysing Scattering-Based Music Content Analysis Systems: Where's the Music"?, in Proc. ISMIR, 2016
Evaluation on a hold-out dataset may not be sufficient to guarantee the performance in the 'wild'

Why is interpretability important?

- **Improving model architectures**





Layer 1 feature visualization (a) Krizhevsky et al. 2012 (b) Zeiler et al. 2014

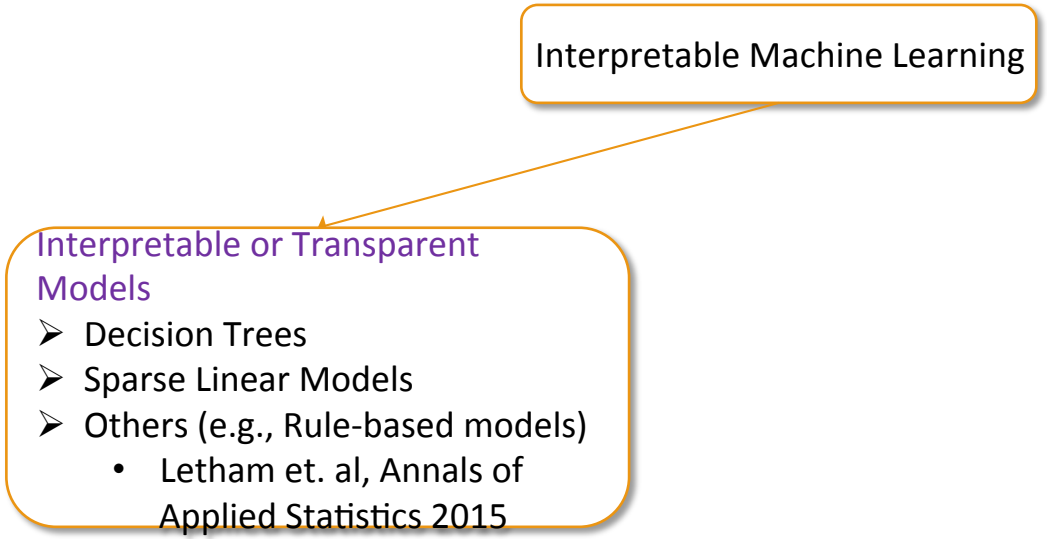
Zeiler et al., "Visualising and Understanding Convolutional Networks", in Proc. ECCV 2014.

Methods to understand model behaviour

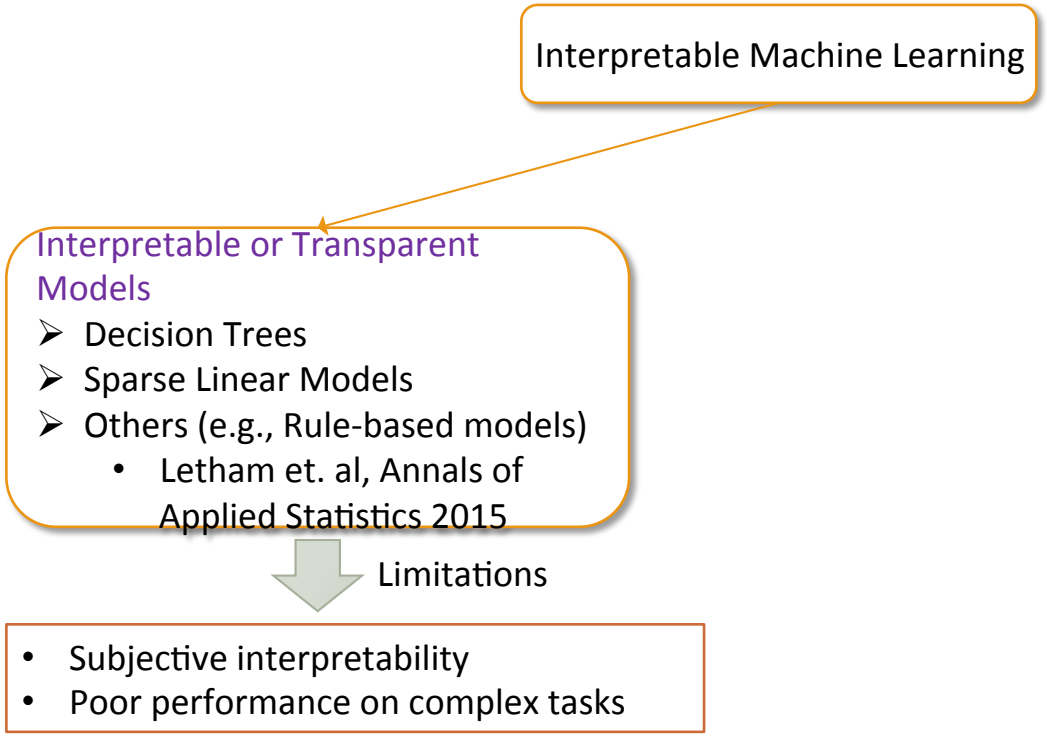
How to open the black-box machine learning models?

Interpretable Machine Learning

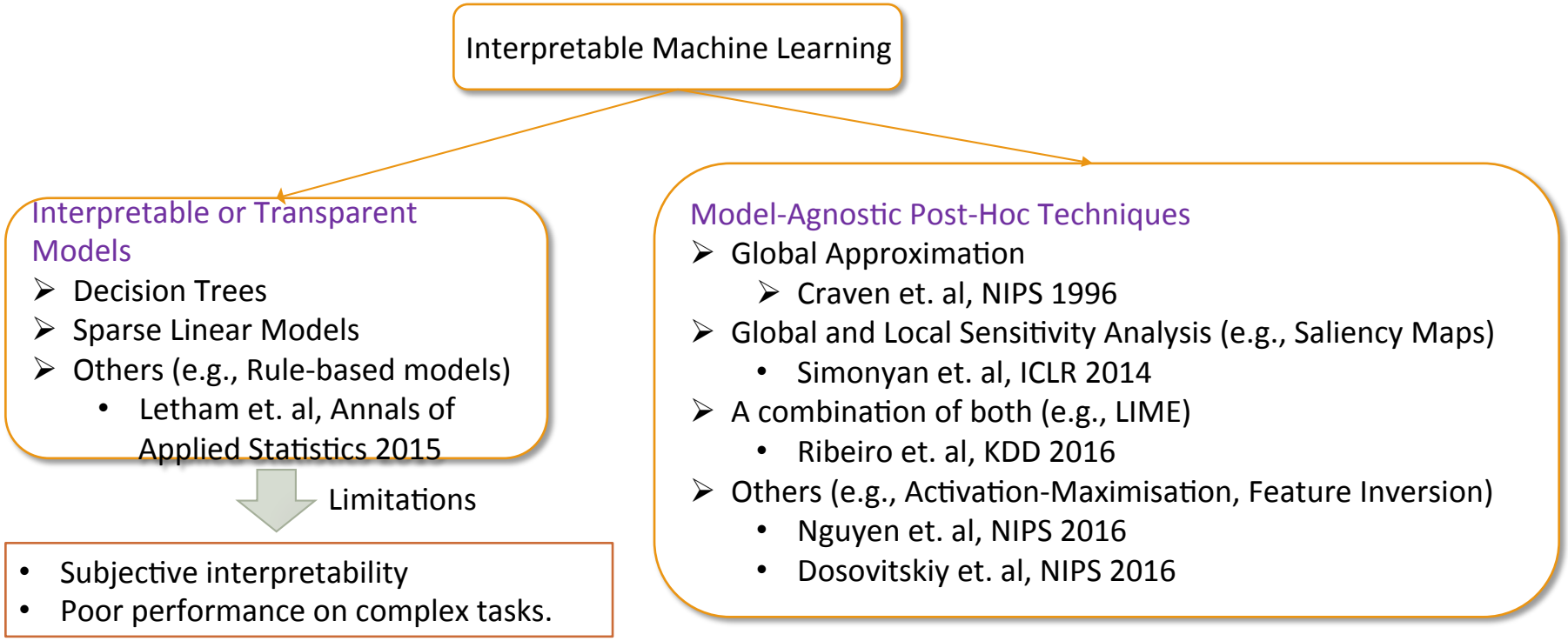
How to open the black-box machine learning models?



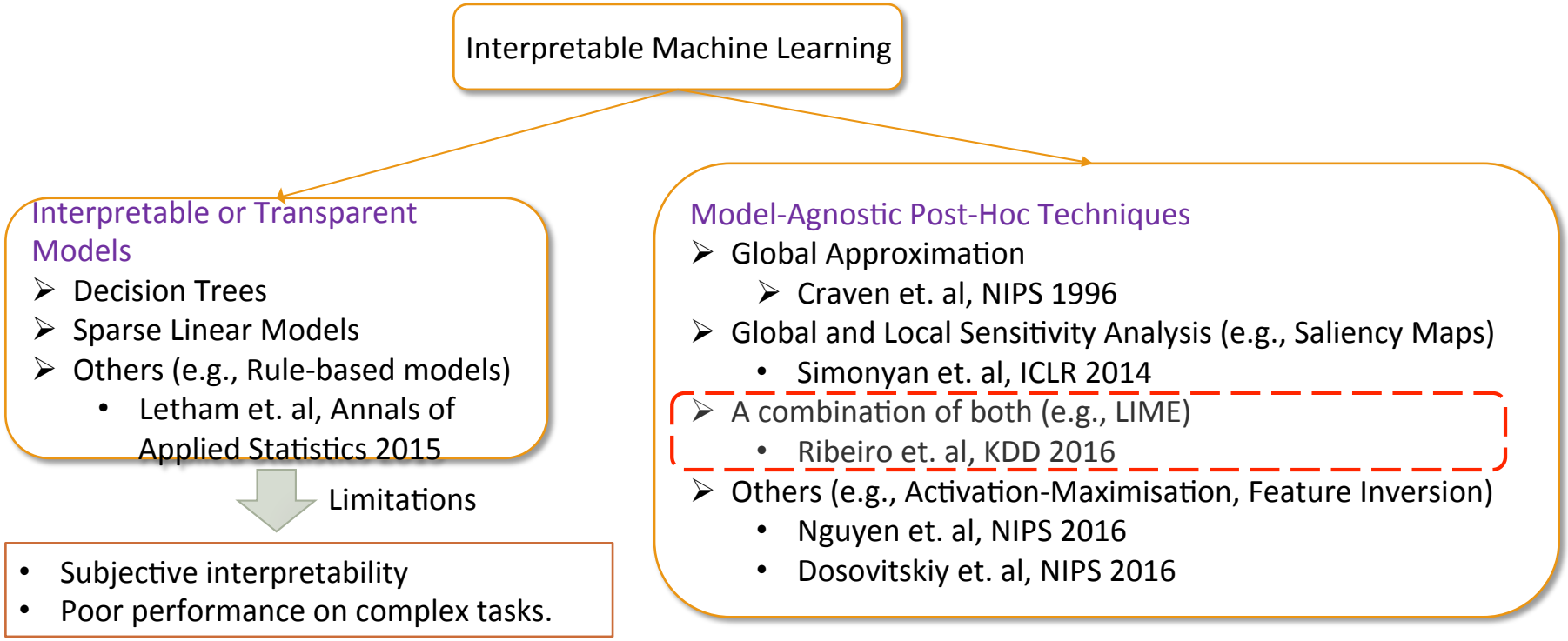
How to open the black-box machine learning models?



How to open the black-box machine learning models?

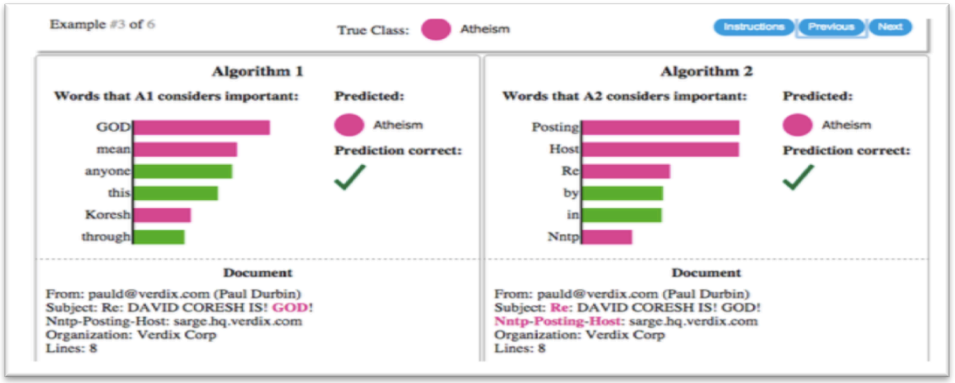


How to open the black-box machine learning models?



Combining global approximation with local sensitivity analysis

Local Interpretable Model-Agnostic Explanations (LIME)

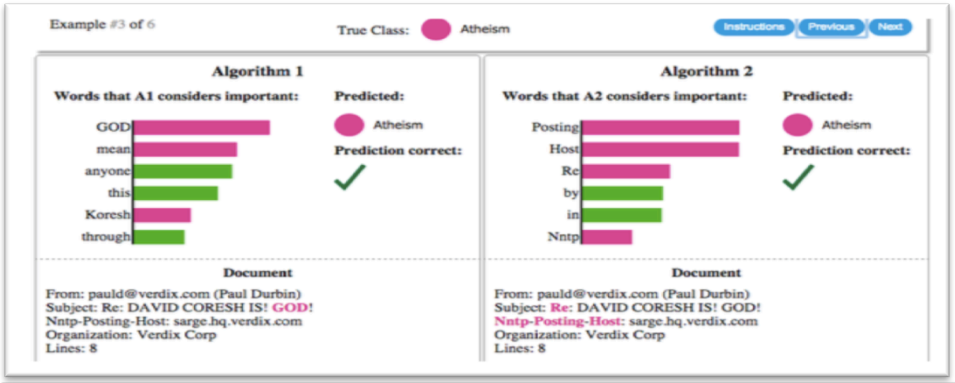


(a)

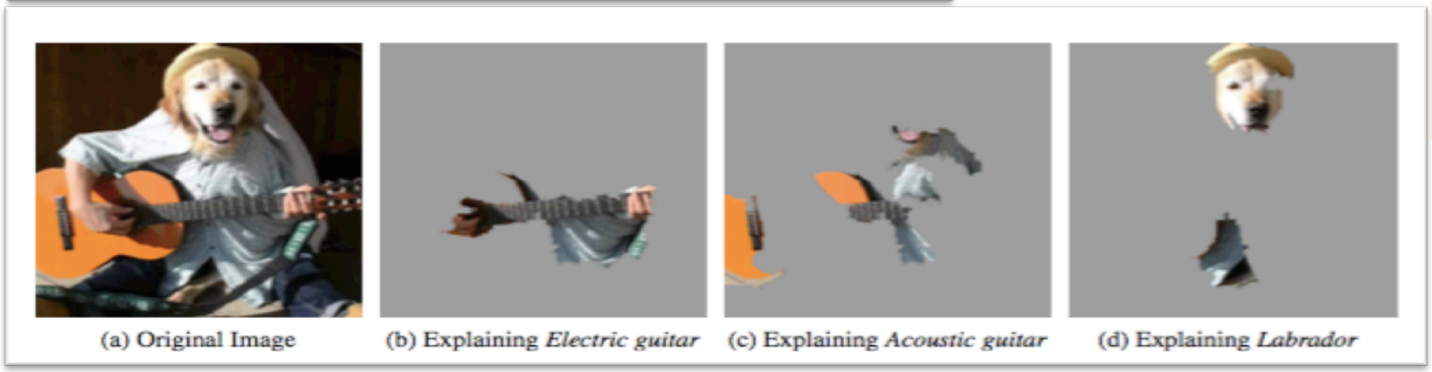
Interpretable explanations generated by LIME for (a) email classification

Ribeiro et al., "Why Should I Trust you? : Explaining the Predictions of Any Classifier", in Proc. KDD, 2016.

Local Interpretable Model-Agnostic Explanations (LIME)



(a)

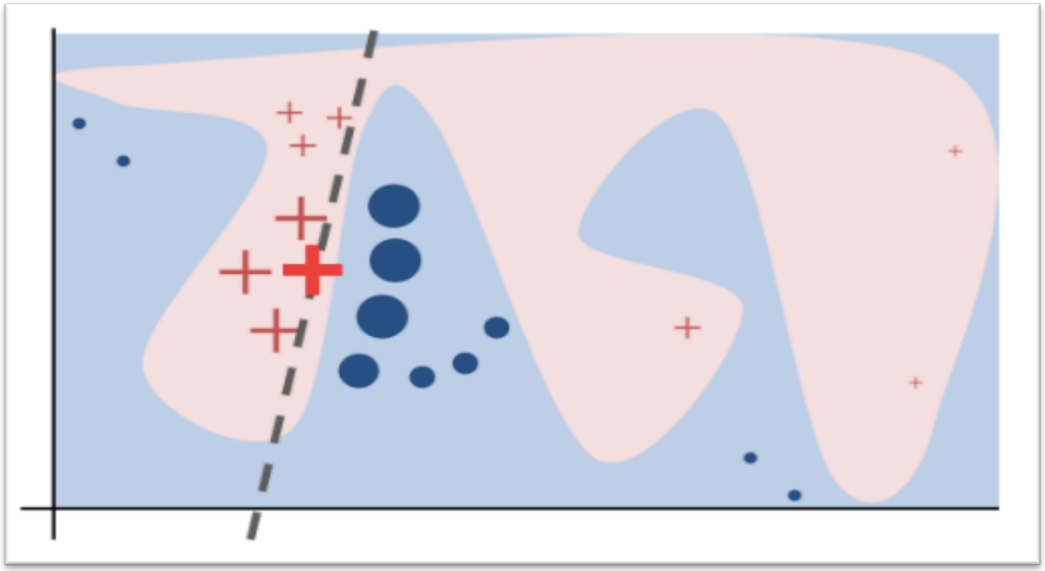


(b)

Interpretable explanations generated by LIME for (a) email classification (b) object recognition task.

Ribeiro et al., "Why Should I Trust you? : Explaining the Predictions of Any Classifier", in Proc. KDD, 2016.

Local Interpretable Model-Agnostic Explanations (LIME)



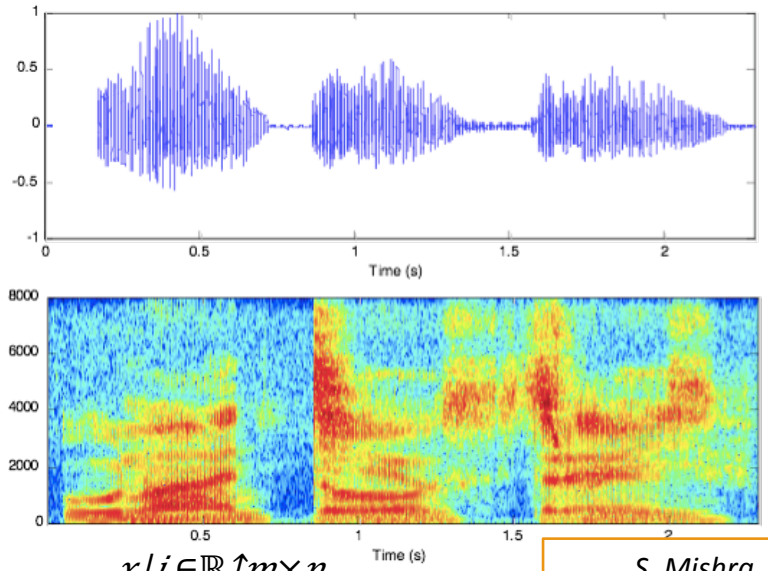
Visual intuition of LIME's methodology.

Ribeiro et al., "Why Should I Trust you? : Explaining the Predictions of Any Classifier", in Proc. KDD, 2016.

Extending LIME to Machine Listening

SoundLIME (SLIME)

- **Interpretable space (IS) and Interpretable representation (IR)**



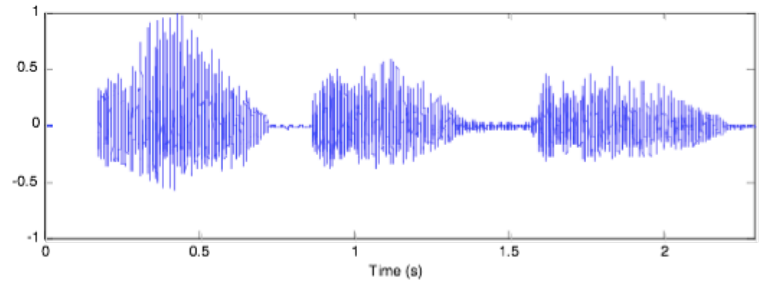
$x \downarrow i \in \mathbb{R} \uparrow m \times n$

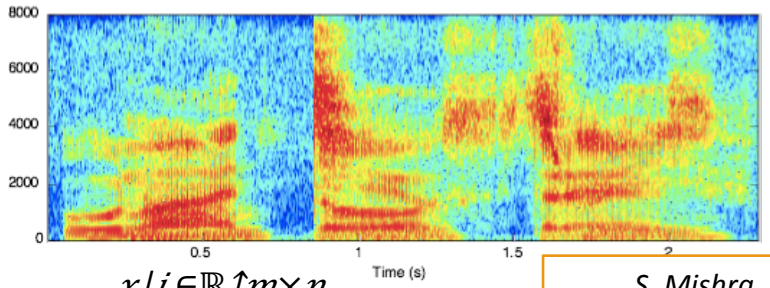
Input

S. Mishra, B. L. Sturm and S. Dixon, "Local Interpretable Model-Agnostic Explanations for Music Content Analysis", in Proc. ISMIR, 2017.

SoundLIME

- **Interpretable space (IS) and Interpretable representation (IR)**

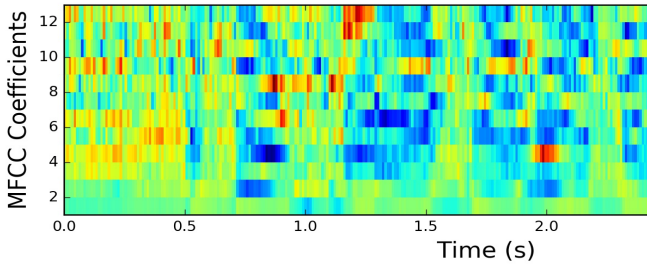




$x \downarrow i \in \mathbb{R} \uparrow m \times n$

Input

$\mathbf{x} \downarrow \boldsymbol{i} \in \mathbb{R} \uparrow m \times 13$

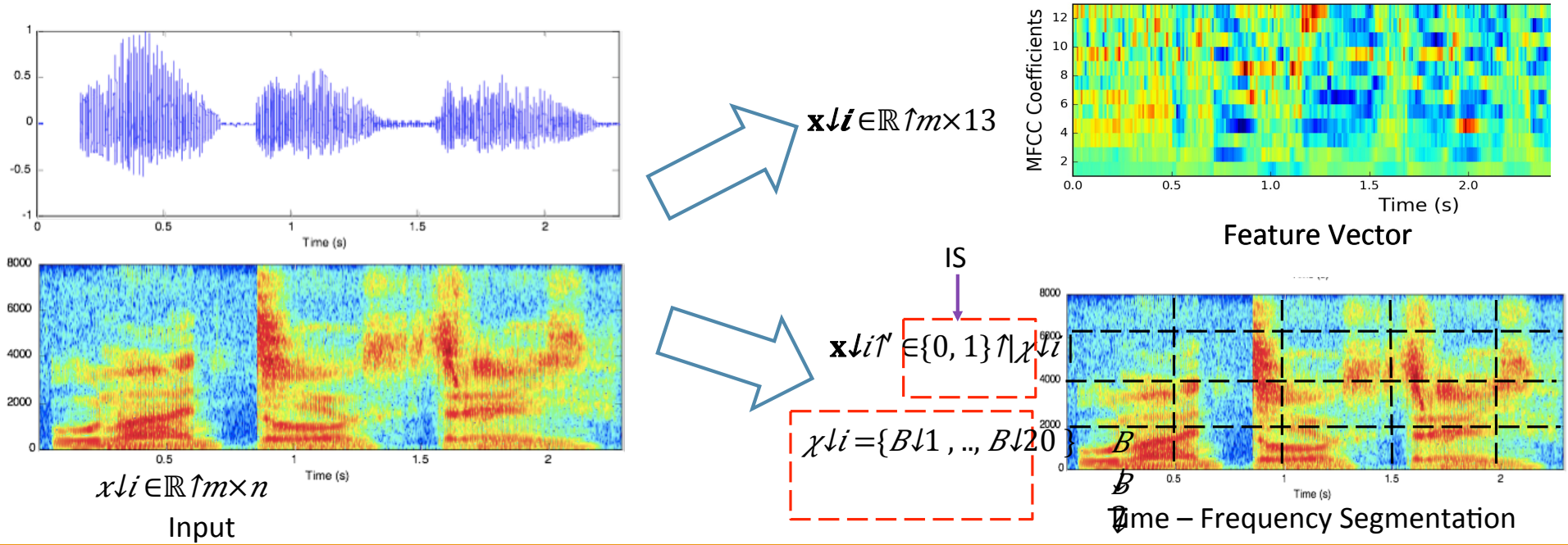


Feature Vector

S. Mishra, B. L. Sturm and S. Dixon, "Local Interpretable Model-Agnostic Explanations for Music Content Analysis", in Proc. ISMIR, 2017.

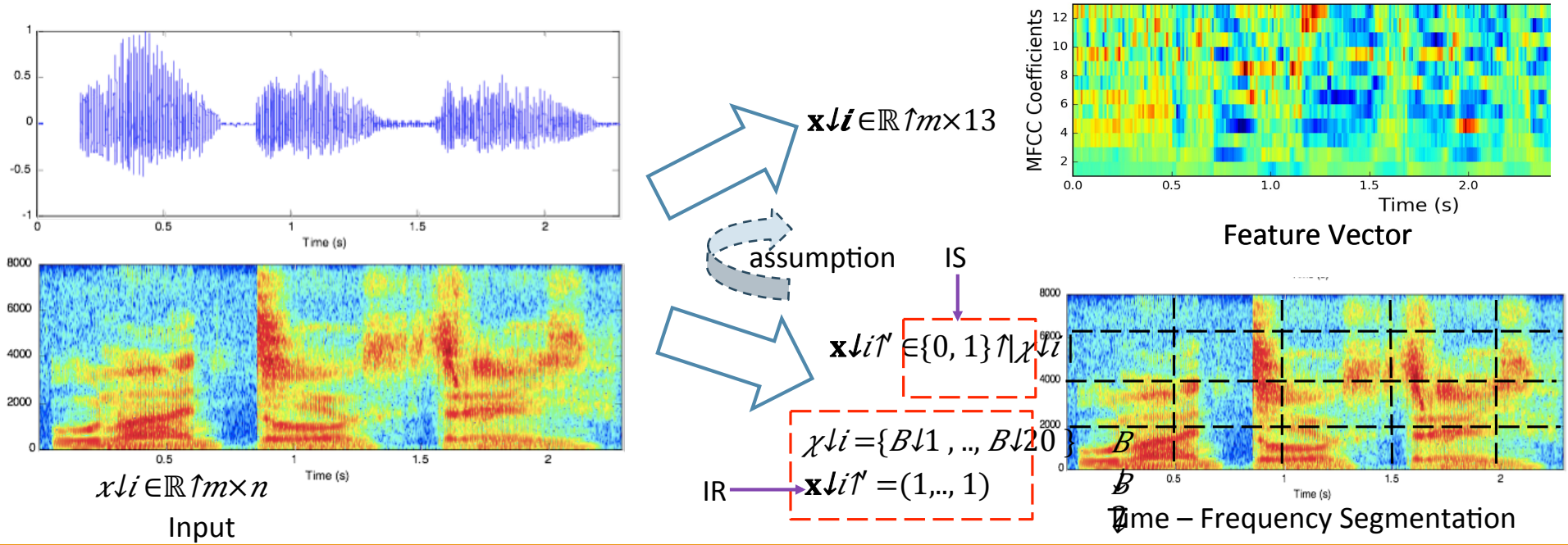
SoundLIME

- **Interpretable space (IS) and Interpretable representation (IR)**



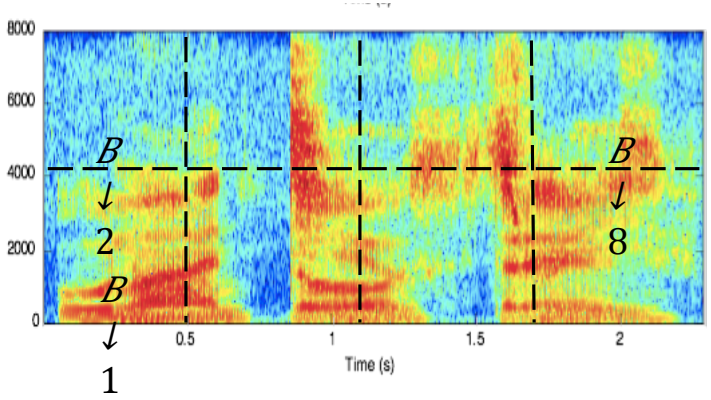
SoundLIME

- **Interpretable space (IS) and Interpretable representation (IR)**



SoundLIME

- **Sampling from the interpretable space by randomly setting dimensions of the input instance to zero.**

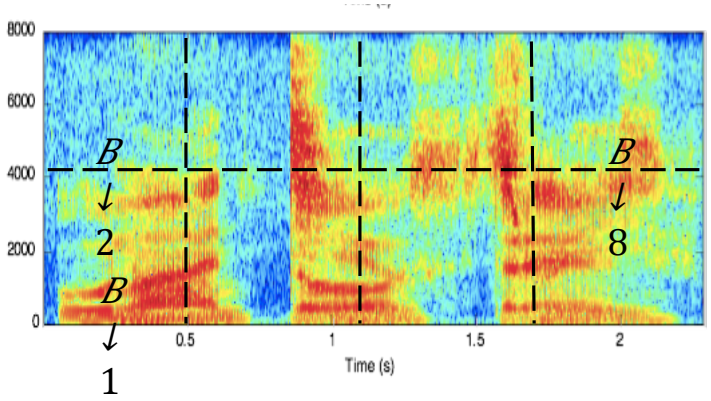


$$\chi\downarrow i = \{B\downarrow 1, .., B\downarrow 8\}$$

IR → $\mathbf{x}\downarrow i\prime = (1, 1, 1, 1, 1, 1, 1, 1)$

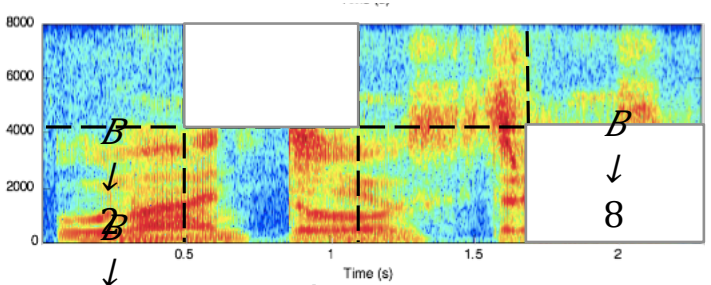
SoundLIME

- **Sampling from the interpretable space by randomly setting dimensions of the input instance to zero.**



$$\chi\downarrow i = \{B\downarrow 1, .., B\downarrow 8\}$$

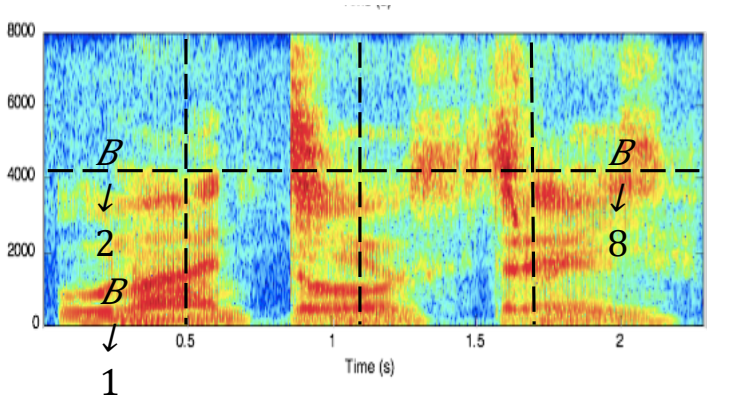
IR → $\mathbf{x}\downarrow i\prime = (1, 1, 1, 1, 1, 1, 1, 1)$



IR → $\mathbf{z}\downarrow i\prime = (1, 1, 1, 0, 1, 1, 0, 1)$

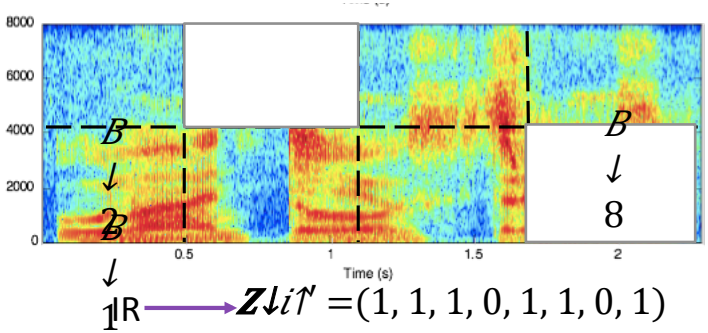
SoundLIME

- **Sampling from the interpretable space by randomly setting dimensions of the input instance to zero.**

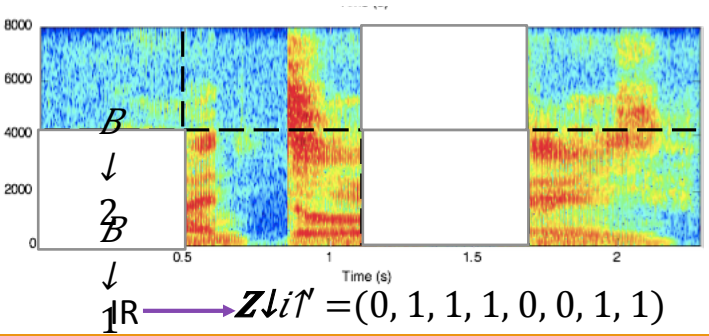


$$\chi\downarrow i = \{B\downarrow 1, .., B\downarrow 8\}$$

IR → $\mathbf{x}\downarrow i' = (1, 1, 1, 1, 1, 1, 1, 1)$

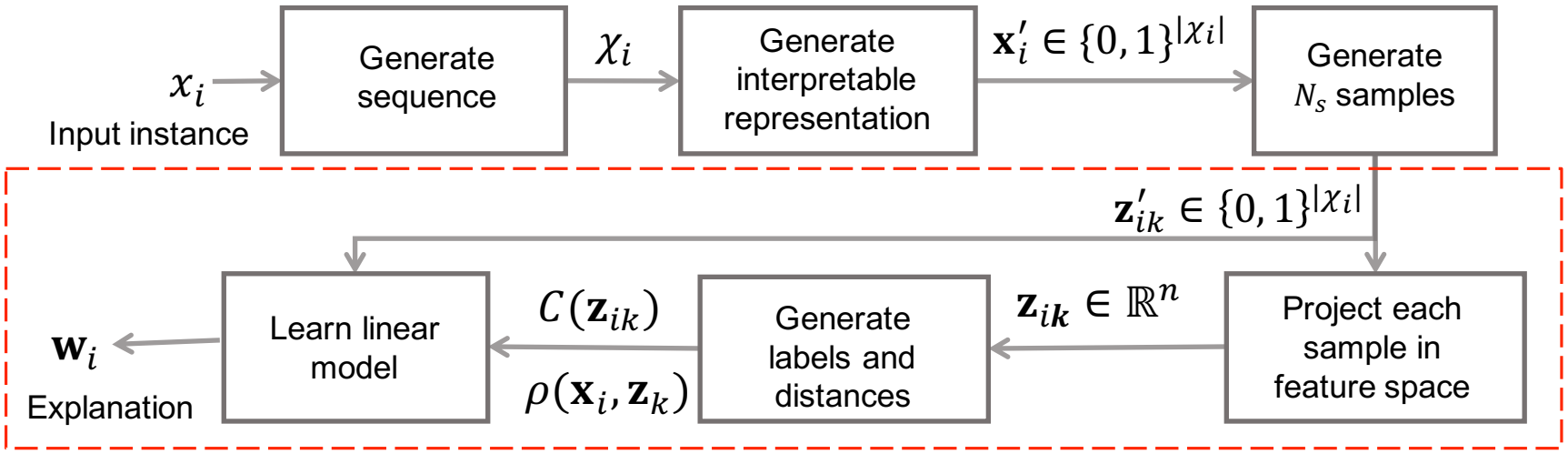


IR → $\mathbf{z}\downarrow i' = (1, 1, 1, 0, 1, 1, 0, 1)$



IR → $\mathbf{z}\downarrow i' = (0, 1, 1, 1, 0, 0, 1, 1)$

SoundLIME



Functional block diagram of SLIME

S. Mishra, B. L. Sturm and S. Dixon, "Local Interpretable Model-Agnostic Explanations for Music Content Analysis", in Proc. ISMIR, 2017.

SoundLIME

- Formally, we define an explanation as a model $g \in G$, where G denotes a class of interpretable models (e.g., linear models, decision trees).

SoundLIME

- Formally, we define an explanation as a model $g \in G$, where G denotes a class of interpretable models (e.g., linear models, decision trees).
- SLIME learns a model g over the interpretable space by the optimisation:

$$\min_{g \in G} \underbrace{L(C, g, \rho_{x_i})}_{\text{Locally-weighted loss}} + \underbrace{\Delta(g)}_{\text{Model complexity (e.g., sparsity in linear models)}}$$

SoundLIME

- Formally, we define an explanation as a model $g \in G$, where G denotes a class of interpretable models (e.g., linear models, decision trees).
- SLIME learns a model g over the interpretable space by the optimisation:

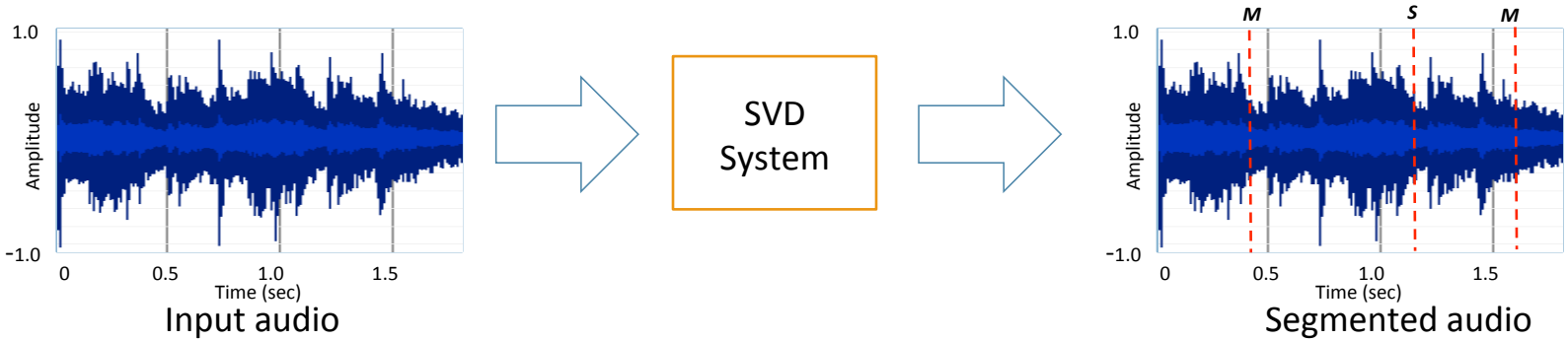
$$\min_{g \in G} \underbrace{L(C, g, \rho_{x_i})}_{\text{Locally-weighted loss}} + \underbrace{\Delta(g)}_{\text{Model complexity (e.g., sparsity in linear models)}}$$

$$L(C, g, \rho_{x_i}) = \sum_{(\mathbf{z}'_k, \mathbf{z}_k) \in Z} \underbrace{\rho(\mathbf{x}_i, \mathbf{z}_k)}_{\text{Weight function}} [C(\mathbf{z}_k) - g(\mathbf{z}'_k)]^2$$

Preliminary experiments to explain predictions of singing voice detectors

What is a Singing Voice Detector (SVD)?

- A singing voice detection system classifies an input audio frame/excerpt into two categories: music without singing voice (*M*), and music with singing voice (*S*).



J. Schlu¨ter et. al, "Exploring data augmentation for improved singing voice detection with neural networks", in Proc. ISMIR, 2015.
B. Lehner et. al, " Towards light-weight, real-time-capable singing voice detection", in Proc. ISMIR, 2013.

Two shallow singing voice detection systems

- Features vector: Statistical measures of MFCC vectors
- Classification is done over a 1 sec excerpt
- System *S1* trains a binary decision tree, and system *S2* trains a Random forest classifier.

Classifier	*Acc[%]*	*Prec.*	*Recall*	*F- score*
Decision tree	71.4	0.72	0.81	0.75
Random forest	76.3	0.75	0.88	0.79

Table 1: Singing voice class evaluation results for the two selected shallow SVD systems (a) Binary decision tree of depth 8 and information gain as the split criterion (b) Random forest of 64 trees, each with depth 16.

Do these systems really know how to differentiate between music and singing voice?

Do these systems really know how to differentiate between music and singing voice?

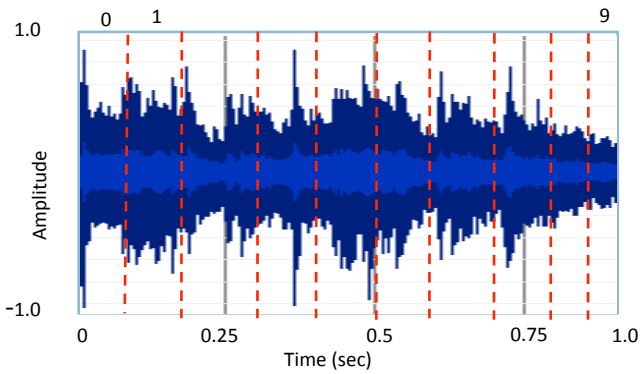
Are the vocal predictions caused by the presence of singing voice?

Do these systems really know how to differentiate between music and singing voice?

Are the vocal predictions caused by the presence of singing voice?

Apply SLIME to generate temporal explanations for each classifier.

Temporal explanations by SLIME

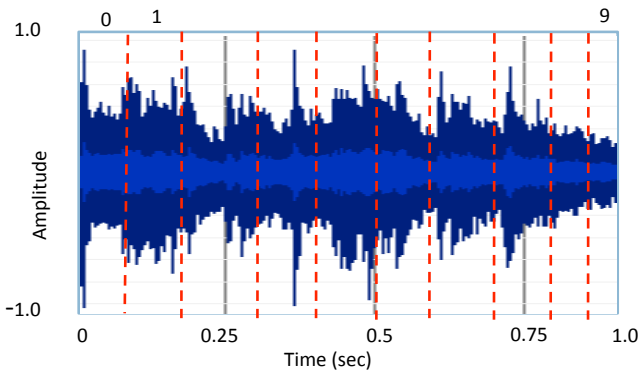


Temporal Segmentation of the Input audio

Id.	*Dur. (s)*	*Prob-Vocal*		*SS-Pred.*		*SS-True*
		BDT	*RF*	*BDT*	*RF*	
41	1.0	0.97	0.85	6,7,9	2,0,7	0-9
178	1.0	0.86	0.86	9,8,4	9,6,0	0-9
58	0.4	0.80	0.76	6,5,3	0,2,6	0-3
124	0.4	0.92	0.84	0,4,6	6,9,8	6-9

Table 2. Instance-based temporal explanations generated by SLIME

Temporal explanations by SLIME



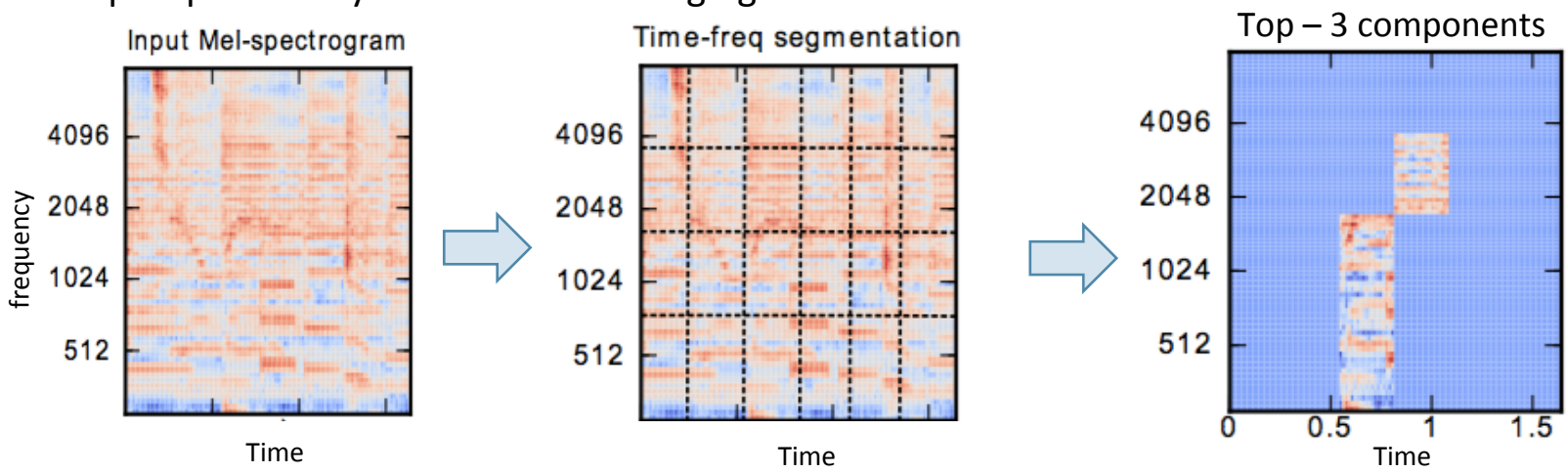
Temporal Segmentation of the Input audio

Id.	*Dur. (s)*	*Prob-Vocal*		*SS-Pred.*		*SS-True*
		BDT	*RF*	*BDT*	*RF*	
41	1.0	0.97	0.85	6,7,9	2,0,7	0-9
178	1.0	0.86	0.86	9,8,4	9,6,0	0-9
58	0.4	0.80	0.76	6,5,3	0,2,6	0-3
124	0.4	0.92	0.84	0,4,6	6,9,8	6-9

Table 2. Instance-based temporal explanations generated by SLIME

Explaining predictions of a deep vocal detector

- Used a state-of-the-art, pre-trained convolutional neural network based system.
- Input: a Mel-spectrogram representation of a 1.6 s audio excerpt.
- Output: probability that it contains singing voice.



J. Schlu¨ter et. al, "Exploring data augmentation for improved singing voice detection with neural networks", in Proc. ISMIR, 2015.

Take Away Points

- Relying just on performance metrics for model selection may lead to selection of suboptimal models.
- Combining performance metrics with interpretable explanations may provide more insight into model behaviour, leading to the development and selection of trustworthy models.
- Several methods exist to understand a trained model behaviour
  - Interpretable models
  - Model-agnostic post-hoc methods.

THANK YOU!!

Any question/ criticism/ suggestion?

Need more time to think? Mail you queries to
saumitra.mishra@qmul.ac.uk