# Explainable Machine Learning and its applications to Machine Listening

SAUMITRA MISHRA[1], BOB L. STURM[2] , EMMANOUIL BENETOS[1], AND SIMON DIXON[1]

[1]Centre for Digital Music, School of Electronic Engineering and Computer Science , Queen Mary University of London
[2]School of Electronic Engineering and Computer Science, Royal Institute of Technology KTH, Sweden

# Motivation

Machine Learning (specifically deep learning) + Big data + High computational power are state-of-the-art in many applications. **But**,


TOM SIMONITE BUSINESS 03.09.18 07:00 AM

AI HAS A HALLUCINATION PROBLEM THAT'S PROVING TOUGH TO FIX

https://www.wired.com/story/ai-has-a-hallucination-problem-thats-proving-tough-to-fix/


Technology Intelligence

Gadgets | Innovation | Big Tech | Start-ups | Politics of Tech | Gaming | Podcast

Technology Intelligence

Facebook shuts down robots after they invent their own language

share

Save 55

https://www.telegraph.co.uk/technology/2017/08/01/facebook-shuts-robots-invent-language/

*The black-box nature of algorithms, susceptibility to adversarial attacks, lack of mathematical and empirical understanding has raised concerns.*


REPORT

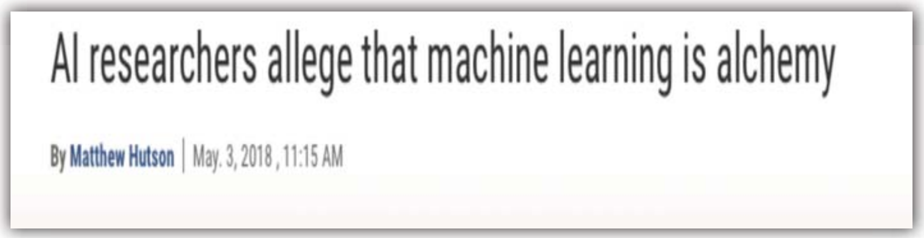
MAGIC AI: THESE ARE THE OPTICAL ILLUSIONS THAT TRICK, FOOL, AND FLUMMOX COMPUTERS

By James Vincent | @jjvincent | Apr 12, 2017, 12:04pm EDT

Illustrations by William Joel

https://www.theverge.com/2017/4/12/15271874/ai-adversarial-images-fooling-attacks-artificial-intelligence

https://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy


AI researchers allege that machine learning is alchemy

By Matthew Hutson | May. 3, 2018 , 11:15 AM

Is a machine learning model **Trustworthy?** **Robust?** **Fair?** **...?**

How can we address the above concerns?

*F. Doshi-Velez and B. Kim. Towards a Rigorous Science of Interpretable Machine Learning. arXiv e- prints, arXiv:1702.08608, 2017.*

Is a machine learning model **Trustworthy?** **Robust?** **Fair?** **...?**

How can we address the above concerns?

By bringing <u>Interpretability</u> to machine learning models

Or

Explainable AI

Or

Interpretable Machine Learning

*F. Doshi-Velez and B. Kim. Towards a Rigorous Science of Interpretable Machine Learning. arXiv e- prints, arXiv:1702.08608, 2017.*

Is a machine learning model **Trustworthy?** **Robust?** **Fair?** **...?**

How can we address the above concerns?

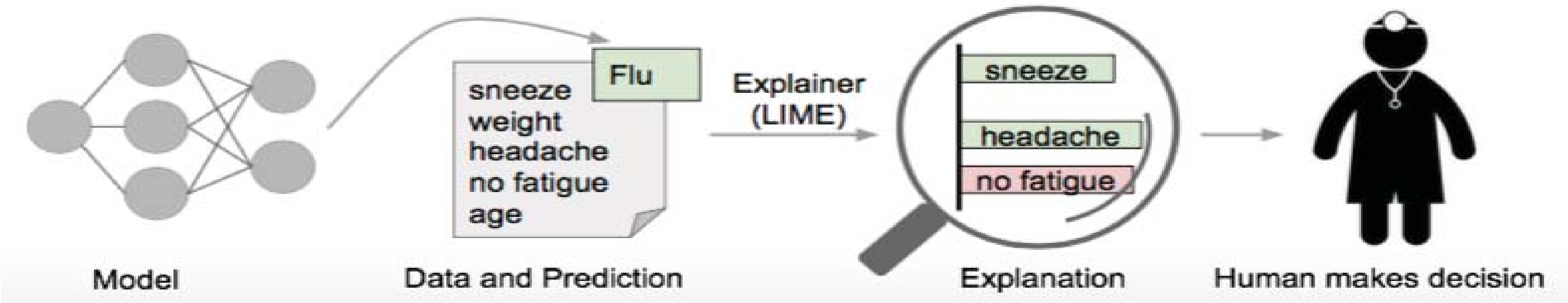
By bringing Interpretability to machine learning models

Or

Explainable AI

Or

Interpretable Machine Learning




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

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

# Methods to understand model behaviour

Interpretable Machine Learning

Train Inherently Interpretable Models

- Decision trees
- Sparse linear models
- Rule-based models

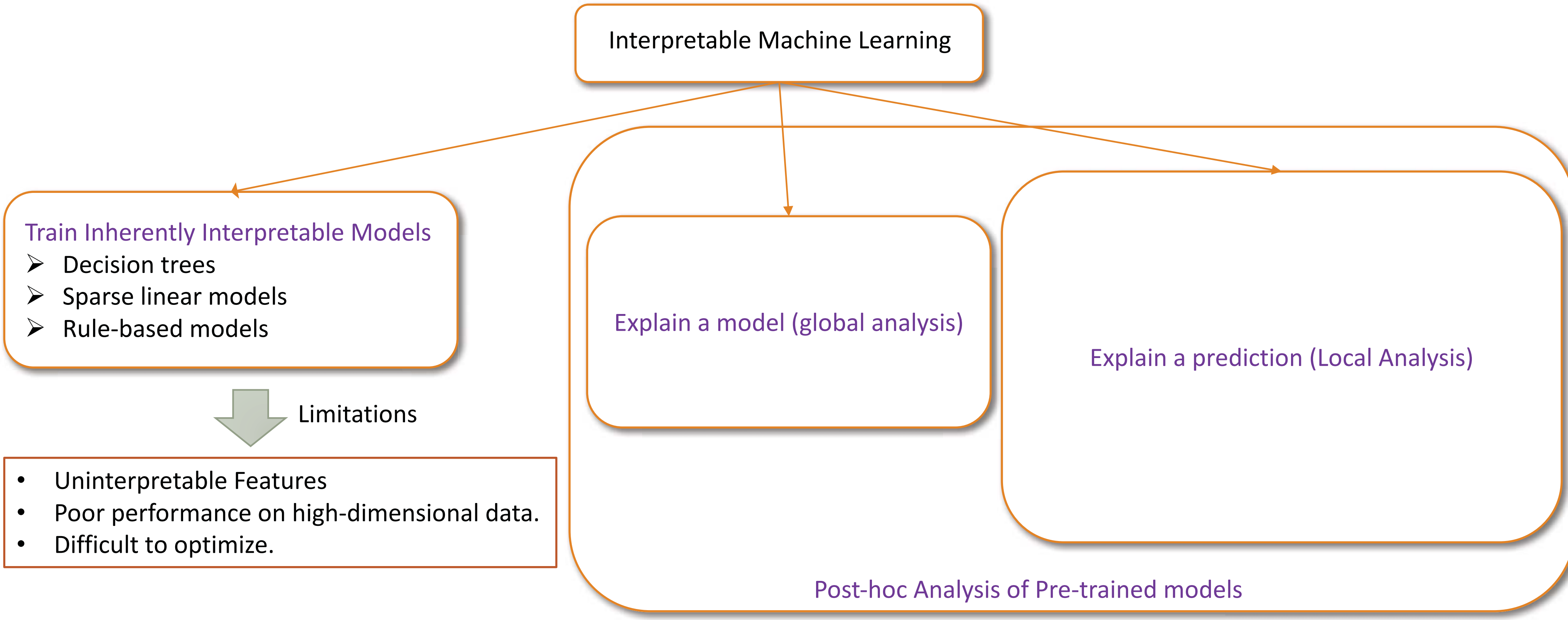
Interpretable Machine Learning

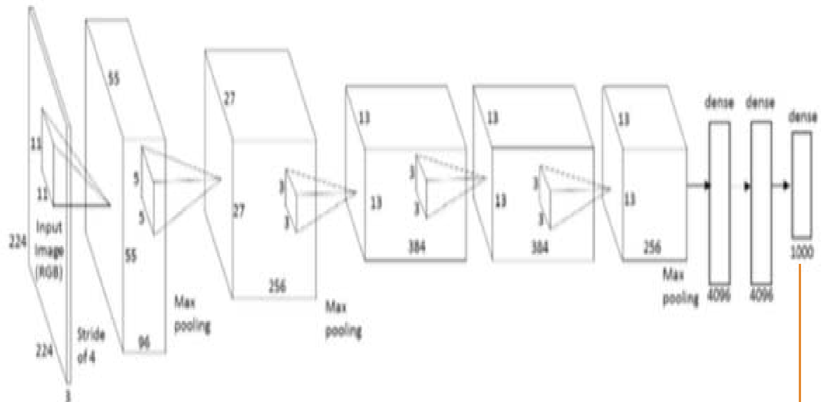
Train Inherently Interpretable Models

- Decision trees
- Sparse linear models
- Rule-based models

Limitations

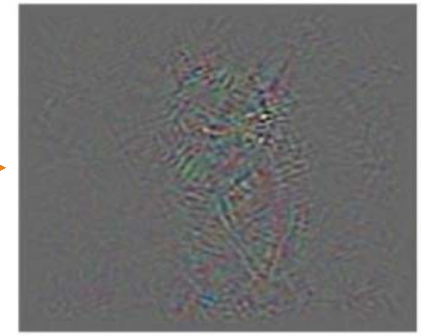
- Uninterpretable Features
- Poor performance on high-dimensional data.
- Difficult to optimize.


Interpretable Machine Learning
Train Inherently Interpretable Models
Decision trees
Sparse linear models
Rule-based models
Limitations
Uninterpretable Features
Poor performance on high-dimensional data.
Difficult to optimize.
Explain a model (global analysis)
Explain a prediction (Local Analysis)
Post-hoc Analysis of Pre-trained models

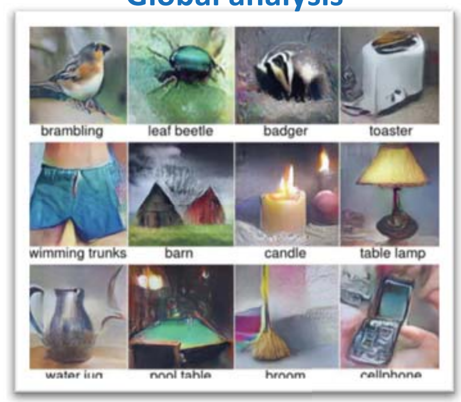




**Local analysis for the 'Cat' Class**



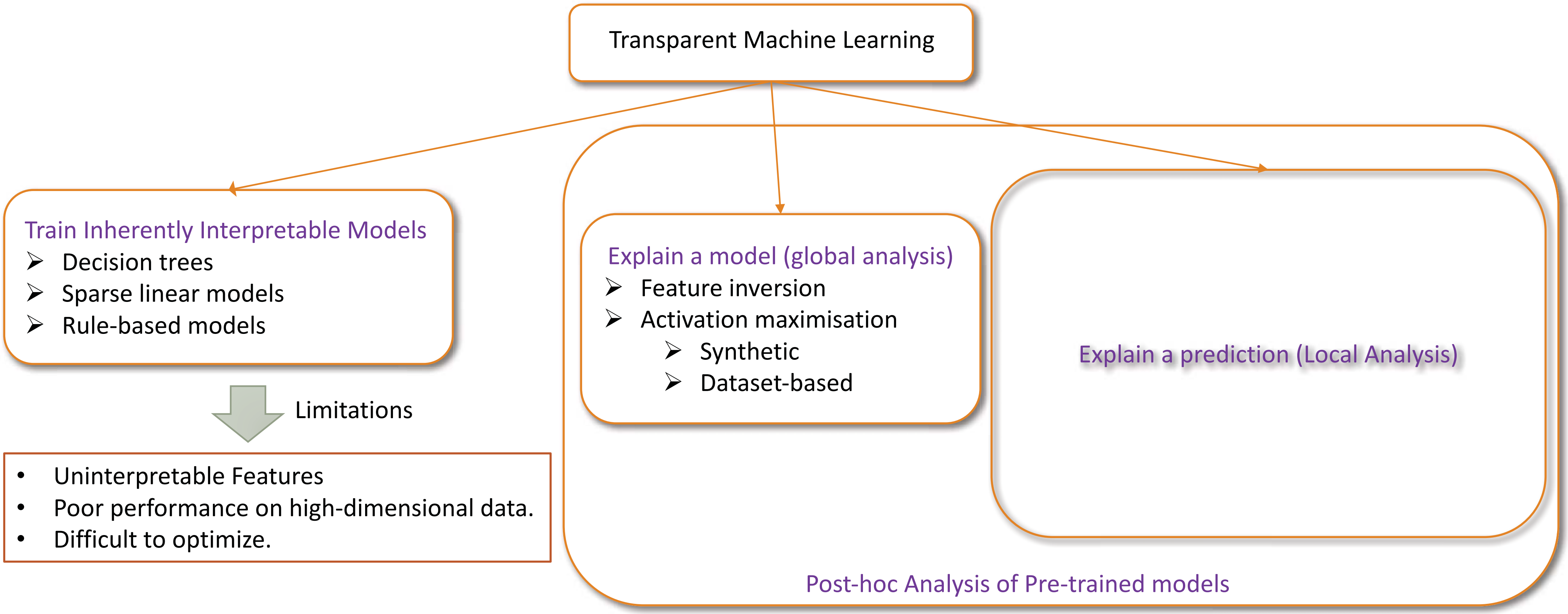
**Global analysis**



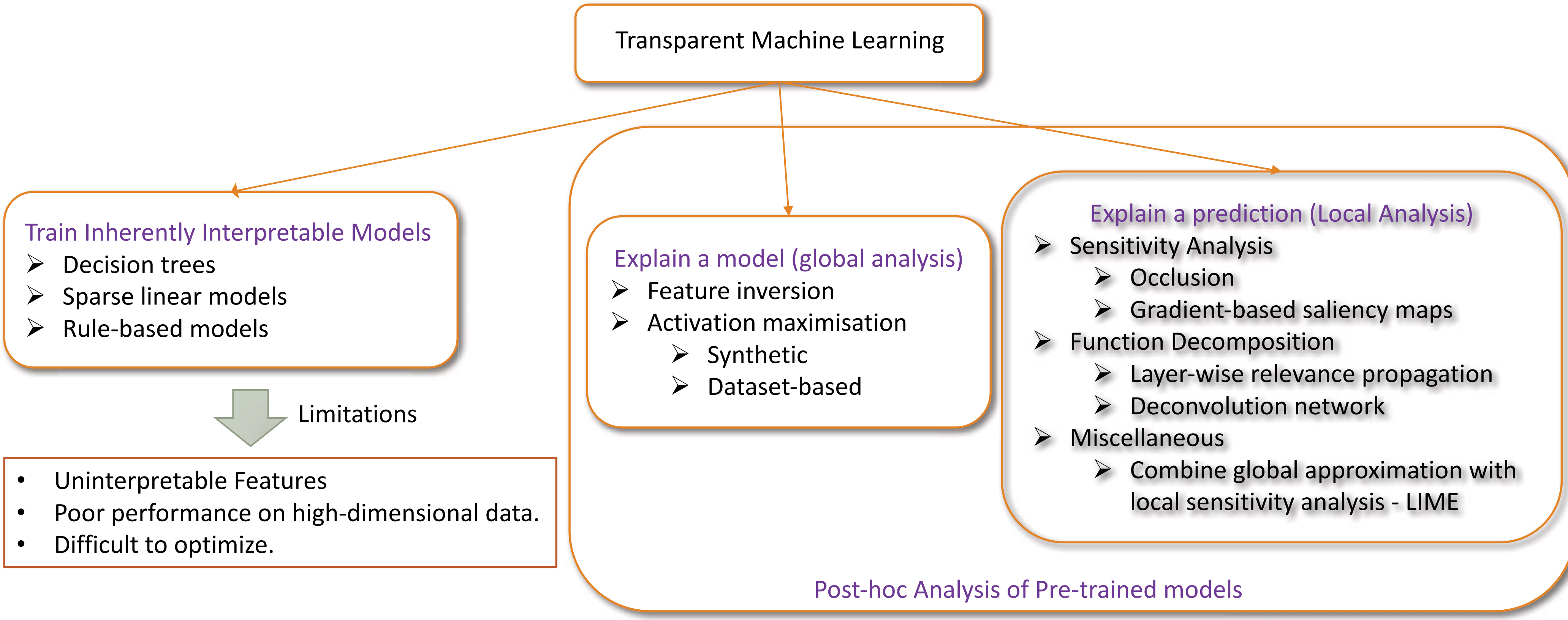


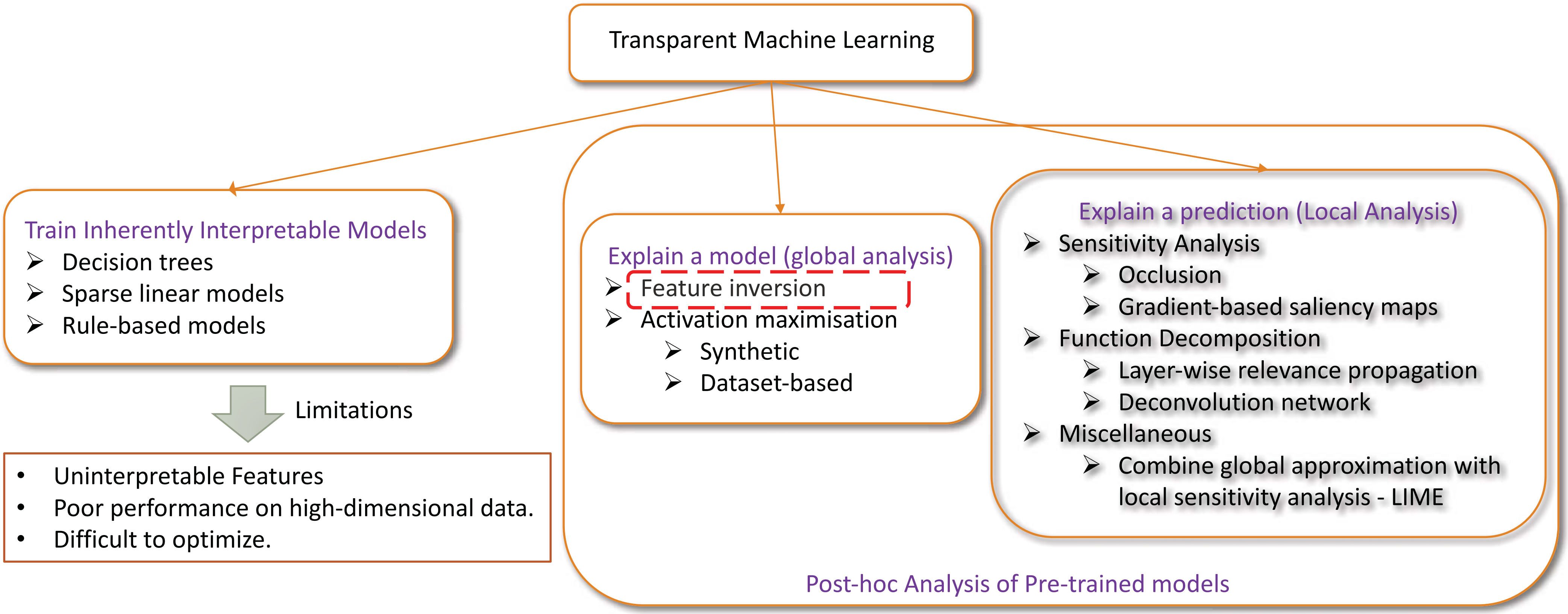
*http://cs231n.stanford.edu/slides/2016/winter1516_lecture9.pdf*

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



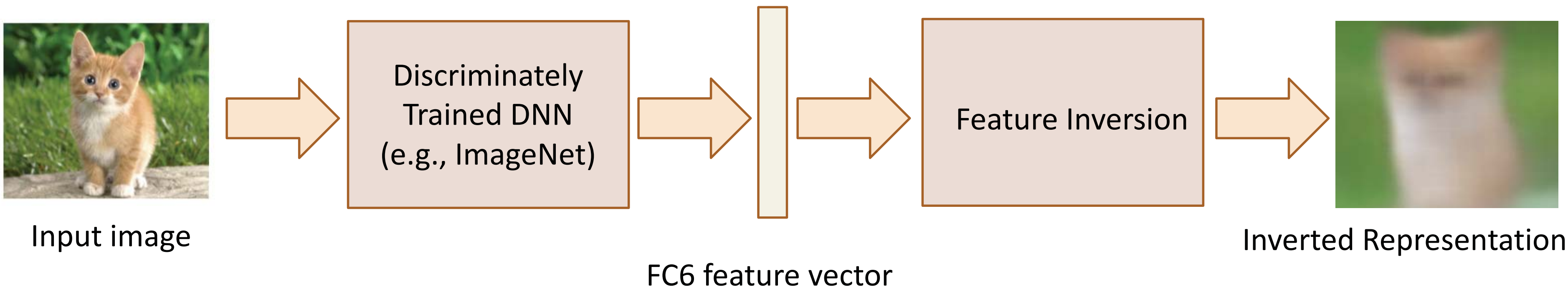
Transparent Machine Learning
Train Inherently Interpretable Models
Decision trees
Sparse linear models
Rule-based models
Limitations
Uninterpretable Features
Poor performance on high-dimensional data.
Difficult to optimize.
Explain a model (global analysis)
Feature inversion
Activation maximisation
Synthetic
Dataset-based
Explain a prediction (Local Analysis)
Post-hoc Analysis of Pre-trained models


Transparent Machine Learning
Train Inherently Interpretable Models
Decision trees
Sparse linear models
Rule-based models
Limitations
Uninterpretable Features
Poor performance on high-dimensional data.
Difficult to optimize.
Explain a model (global analysis)
Feature inversion
Activation maximisation
Synthetic
Dataset-based
Explain a prediction (Local Analysis)
Sensitivity Analysis
Occlusion
Gradient-based saliency maps
Function Decomposition
Layer-wise relevance propagation
Deconvolution network
Miscellaneous
Combine global approximation with local sensitivity analysis - LIME
Post-hoc Analysis of Pre-trained models


Transparent Machine Learning
Train Inherently Interpretable Models
Decision trees
Sparse linear models
Rule-based models
Limitations
Uninterpretable Features
Poor performance on high-dimensional data.
Difficult to optimize.
Explain a model (global analysis)
Feature inversion
Activation maximisation
Synthetic
Dataset-based
Explain a prediction (Local Analysis)
Sensitivity Analysis
Occlusion
Gradient-based saliency maps
Function Decomposition
Layer-wise relevance propagation
Deconvolution network
Miscellaneous
Combine global approximation with local sensitivity analysis - LIME
Post-hoc Analysis of Pre-trained models

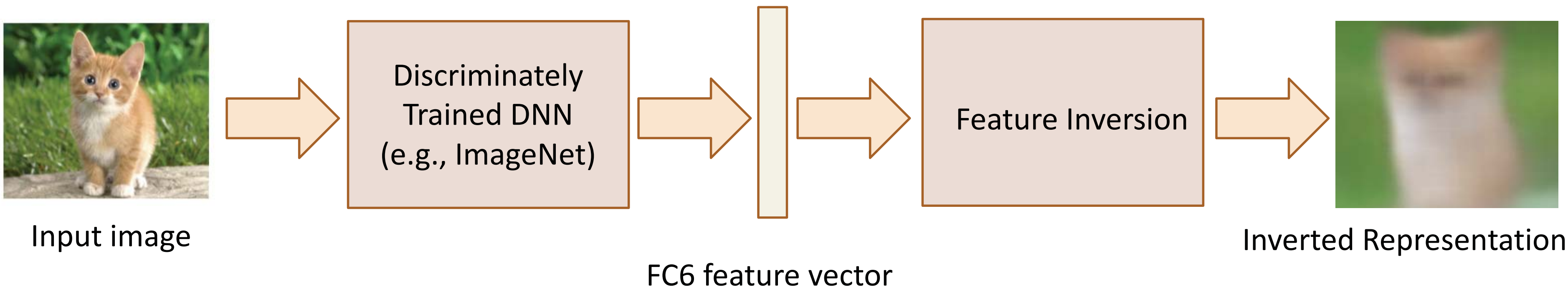
# Feature Inversion

Feature inversion ( or inverting a feature vector) involves mapping (in some way) a feature vector from a layer, back to the input space (e.g., image, time-frequency spectrogram).



# Feature Inversion

Feature inversion ( or inverting a feature vector) involves mapping (in some way) a feature vector from a layer, back to the input space (e.g., image, time-frequency spectrogram).




*How can we use this idea of feature inversion to understand a model?*

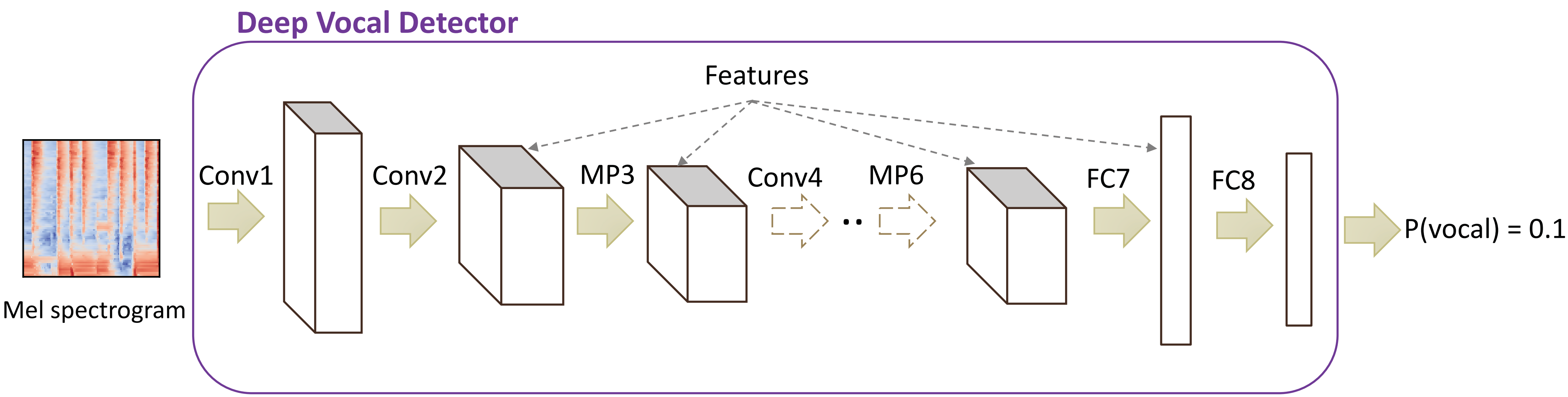
# Key Idea

Discriminative training forces each hidden layer of a deep discriminative model to only preserve information relevant to the discrimination task.

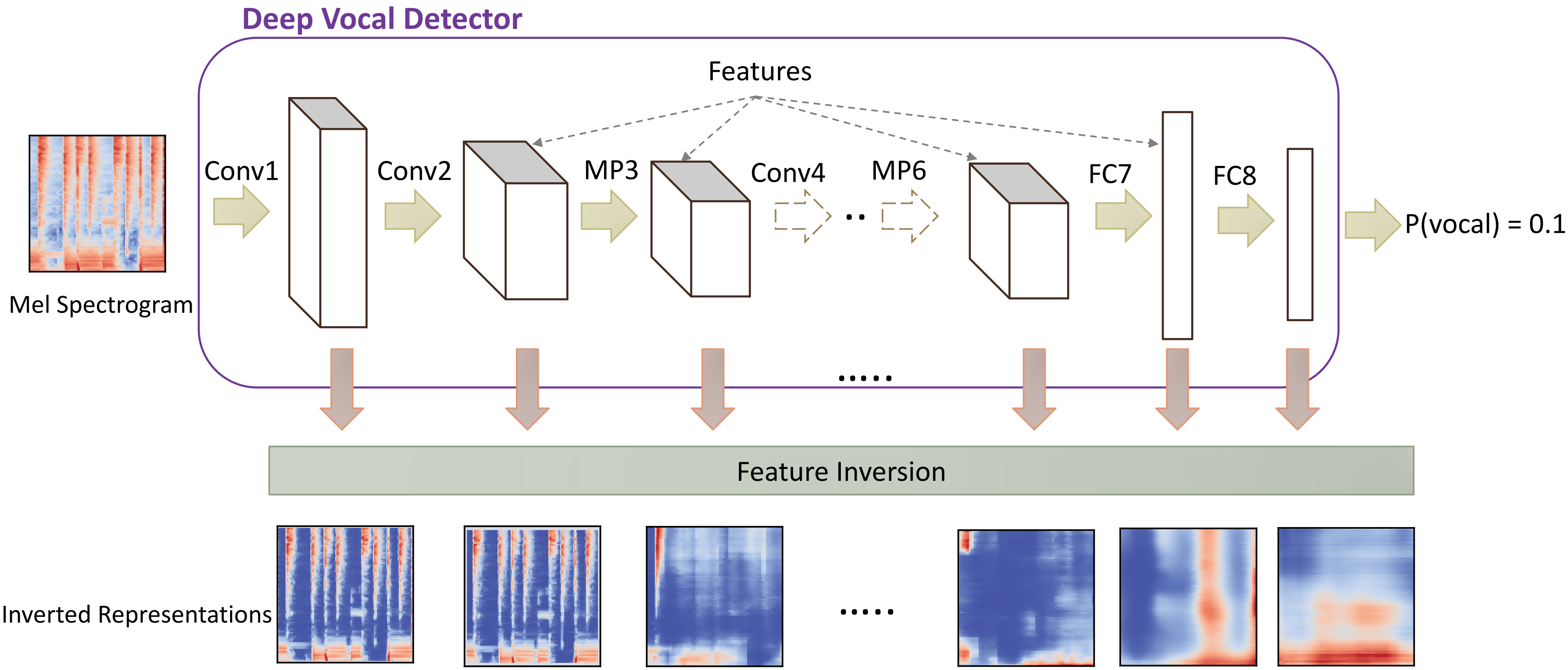
# Key Idea

Discriminative training forces each hidden layer of a deep discriminative model to only preserve information relevant to the discrimination task.

*Inversion of features (e.g., features maps) generated at a layer (e.g., convolutional layer) back to the input space (e.g., pixel space) will assist in understanding information a model preserves at that layer.*



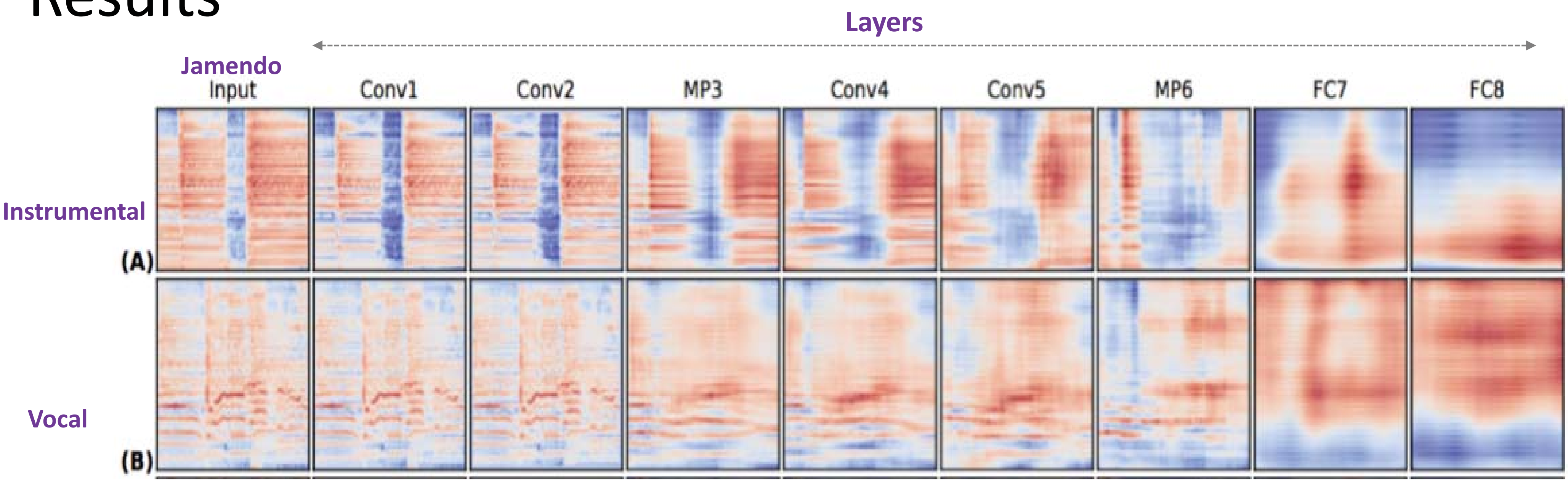

*J. Schlu¨ter et. al, "Exploring data augmentation for improved singing voice detection with neural networks", in Proc. ISMIR, 2015.*




*S. Mishra, B.L. Sturm, S. Dixon, "Understanding a Deep Machine Listening Model Through Feature Inversion", in Proc. ISMIR 2018.*

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

# Results




- FC8 does not preserve any temporal and harmonic information, but the reconstructions from shallower layers are visually similar to the input.
- Inverted representations from FC8 suggest that the SVD model learns a class-decision function in this layer.
- Deeper layers capture more invariances from data than shallow layers.
- The above results generalize across datasets.

# 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.
- There exist several ways to analyse the behaviour of machine learning models
  - Using inherently Interpretable models
  - Using post-hoc methods to analyse a pre-trained model.

THANK YOU