An Ensemble of Deep Convolutional Networks for Automatic Film Score Genre Recognition

Toward a Discovery-Rich Recommendation Agent

Problem: Background Track Bottleneck

Video and audio are growing exponentially

Overwhelming choice leads to sub-optimal matches—
a loss for music supervisors, project producers, and recording artists

Solution: Recommender Agent

Automatic tagging means improved search, increased discovery, and more novel background tracks

MIR hears *features*
Humans hear *fit*

Music Genre? Mood-Markers? Film Genre?
What captures 'music fit'?

Explanatory gap: "[Search] users are not able to define their needs in terms of low-level audio parameters (e.g., spectral shape features)"
(Kaminskas and Ricci 2012)

Hypothesis: Film Genre can be modeled directly from soundtracks

Method:
Tag samples with associated film genres

Use the labeled dataset to train a neural network

Predict the appropriate application of unlabeled audio

Timbre = Genre?

Timbre features are work-horse of most Automatic Genre Recognition

Timbre fundamental to actual human perception of genre (+ source ID, phonemes, mood, valence)

If timbre is sufficient for classification, a lot of dimensionality can be discarded from the dataset

Timbre: Fundamental to Music Classification

Subjects can estimate the ***emotional content***, ***style*** and ***decade of release*** of previously unheard recordings significantly better than chance, after exposure of only ***400 ms***

Even subjects with ***amusia***, who are unable to identify any song by name, performed ***equally well*** on judgments of ***style*** and ***emotion***
(Krumhansl 2010)

Timbre: Fundamental to Music Classification

"if genre identification…required the prior classifications of component features like ***melody****,* ***bass****,* ***harmony****, and* ***rhythm****, then it is unlikely that such rapid identification would have been possible. That is, from a single tone one could not infer any reliable information about melody or rhythm."*

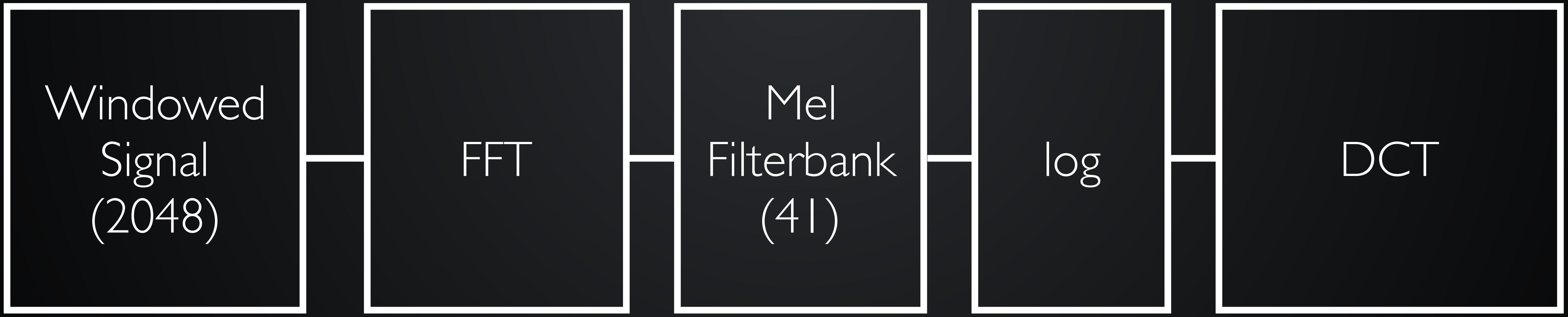
(Gjerdingen and Perrott 2008)

Timbre: Best Feature for Machine Learning

Early research showed highest correlation between timbral features and genre
(Tzanetakis & Cook 2002)

Survey of State-of-the-Art: Best algorithms use MFCCs + spectral stats
(Sturm 2013)

Implemented Features: MFCC


Windowed Signal (2048)

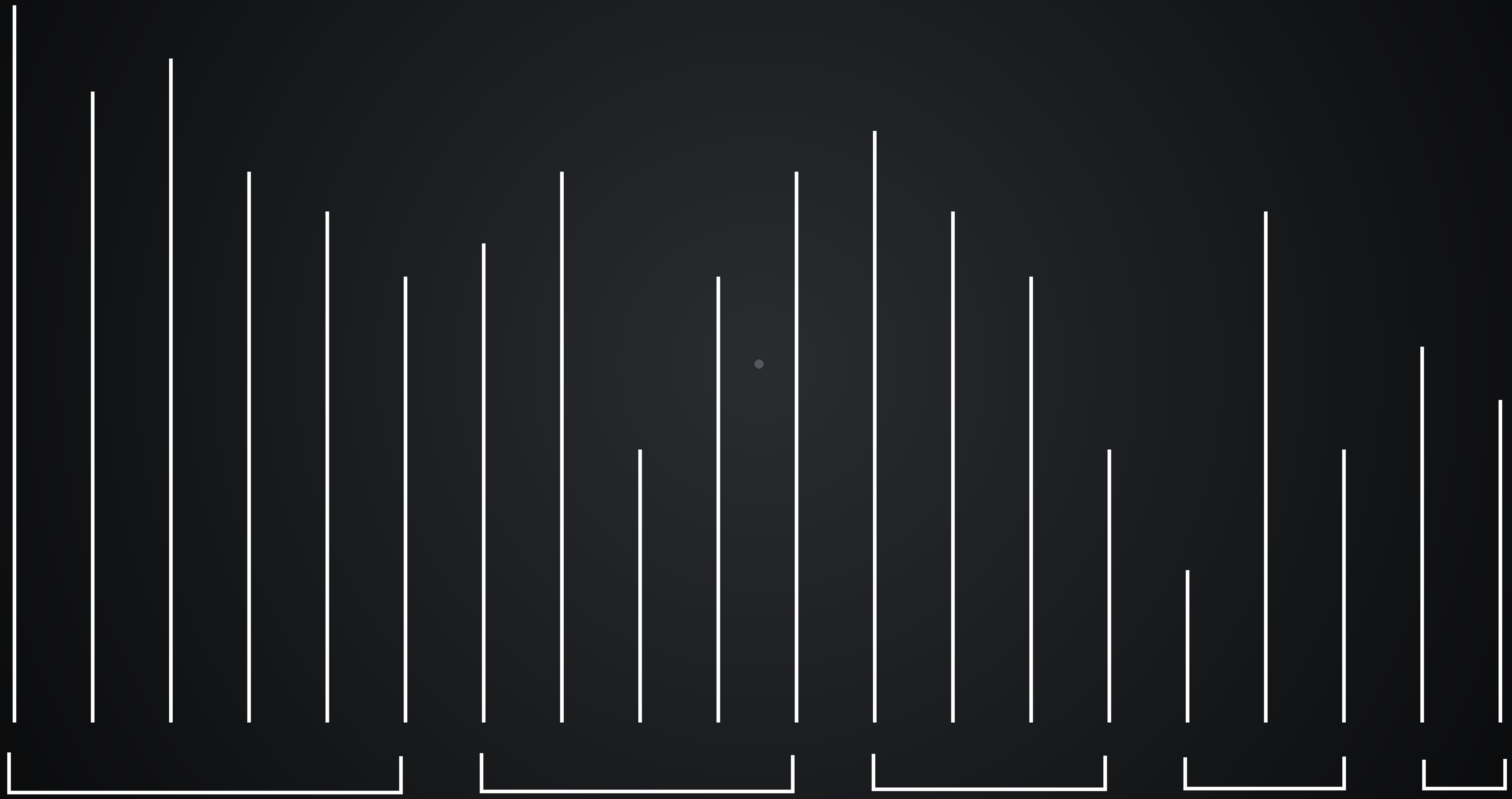
FFT

Mel Filterbank (41)

log

DCT

Linear Bins to Mel-Spaced Bins



$$f = 700(e^{m/1127} - 1)$$

Implemented Feature – MFCC

Pros:

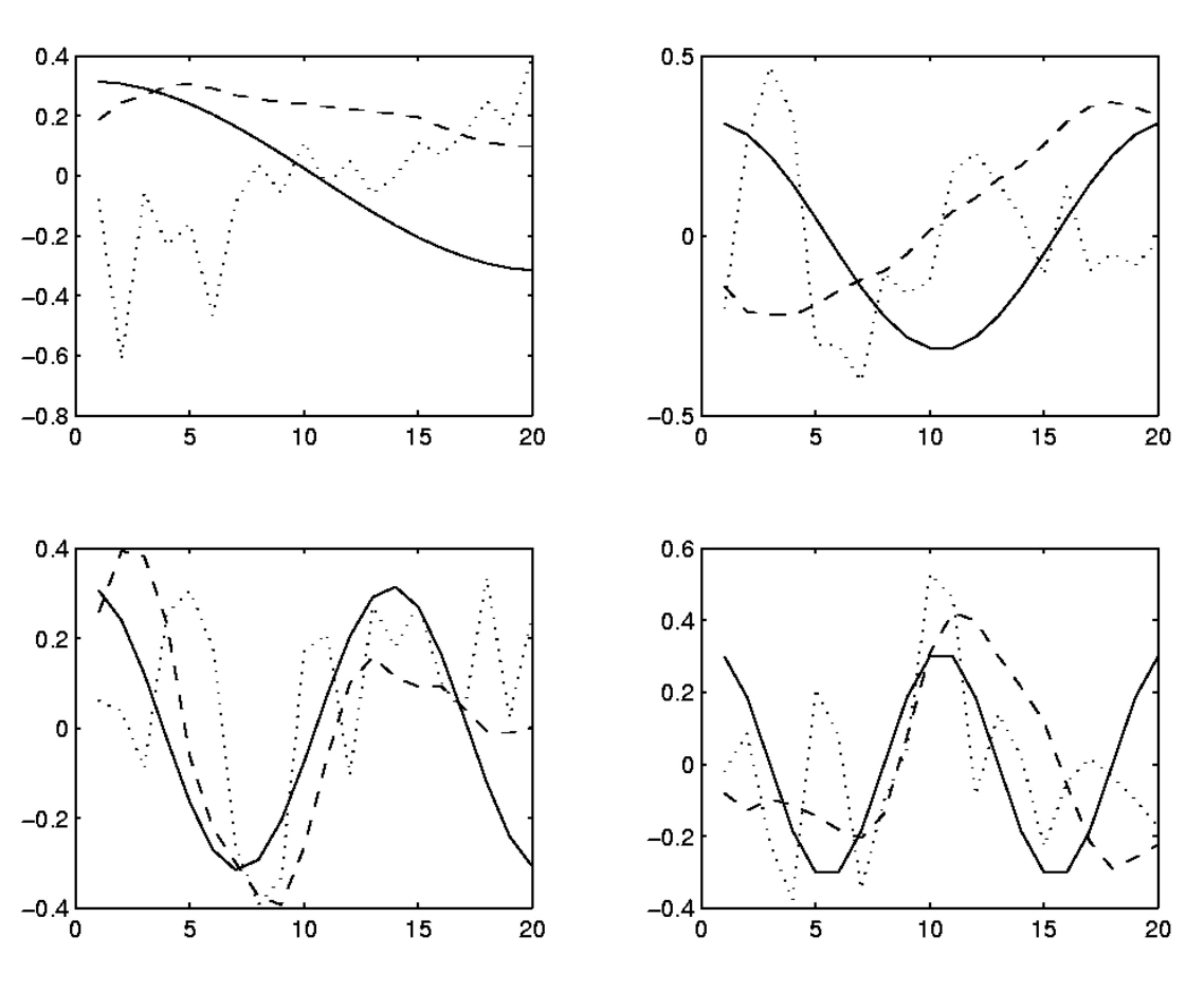
Substantial dimensionality reduction (~10x - 100x)

Mel filterbank preserves perceptually relevant information

Implemented Feature – MFCC

Pros:

Decorrelates features,
approximates Principal Component Analysis projection



Implemented Feature – MFCC

Cons:

Lossy: reconstruction is very sketchy
without prior preservation of pitch and phase

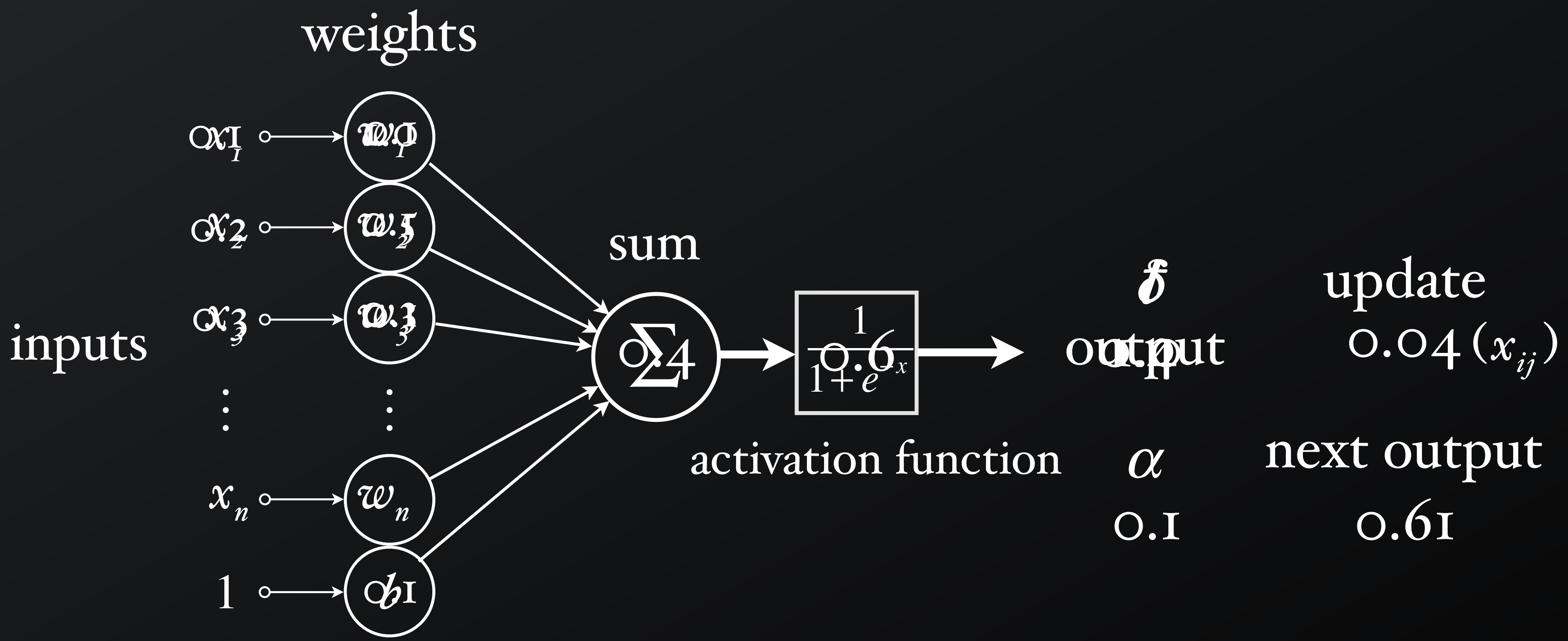
Deep Neural Network: Feature Extractor + Classifier

Deep Networks converge on lower-dimensional projection that minimizes cost function

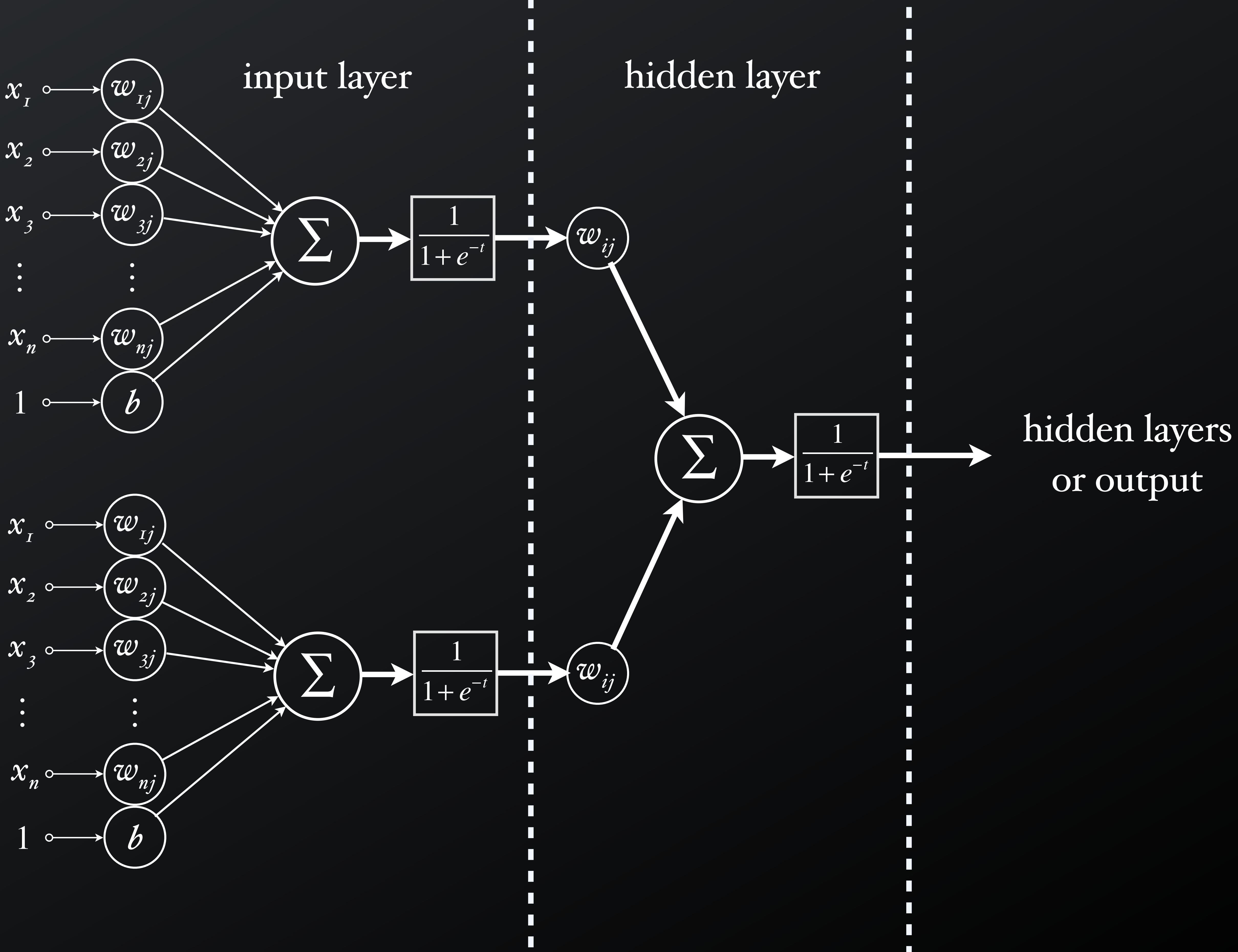
DNN can extract best features from raw spectrogram or further reduce dimensions of large MFCC tile

Softmax on k outputs can do multi-class

McCulloch-Pitt Neuron Model



Feed-Forward Training



input layer

hidden layer

hidden layers or output

Back-propagation of Error

$$\frac{\partial E}{\partial w_{ji}} = -\overbrace{(t_j - y_j)}^{\text{cost}} \quad \underbrace{g'(h_j)}_{\substack{\text{deriv. of activation function} \\ \text{w/ respect to summation}}} \quad \overbrace{x_i}^{\text{input i}}$$

$$\text{update} = \alpha(t_j - y_j)g'(h_j)x_i$$

Deep Neural Network

Pros:

Powerful: automatically extracts the best features for the task

Flexible: trivial to add increased depth and complexity to the model

Deep Neural Network

Cons:

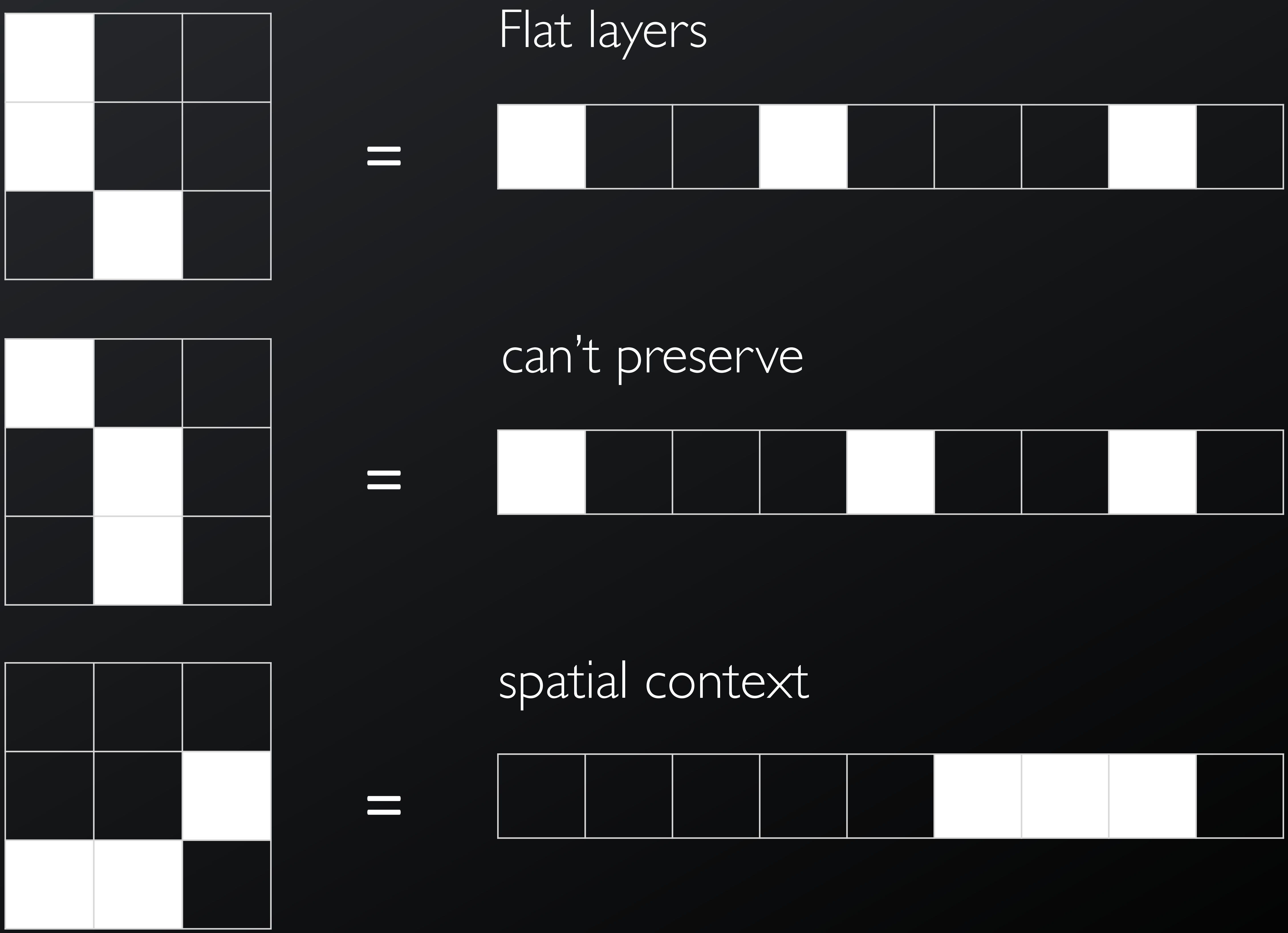
“Black Box”: Parameters are inscrutable;
‘reasoning’ can only be understood empirically

Doesn’t work “out-of-the-box”; search of hyper-parameters required

Too Powerful: without careful regularization,
prone to overfitting

Fully-connected layers confused by translation and scaling

Lost in Translation

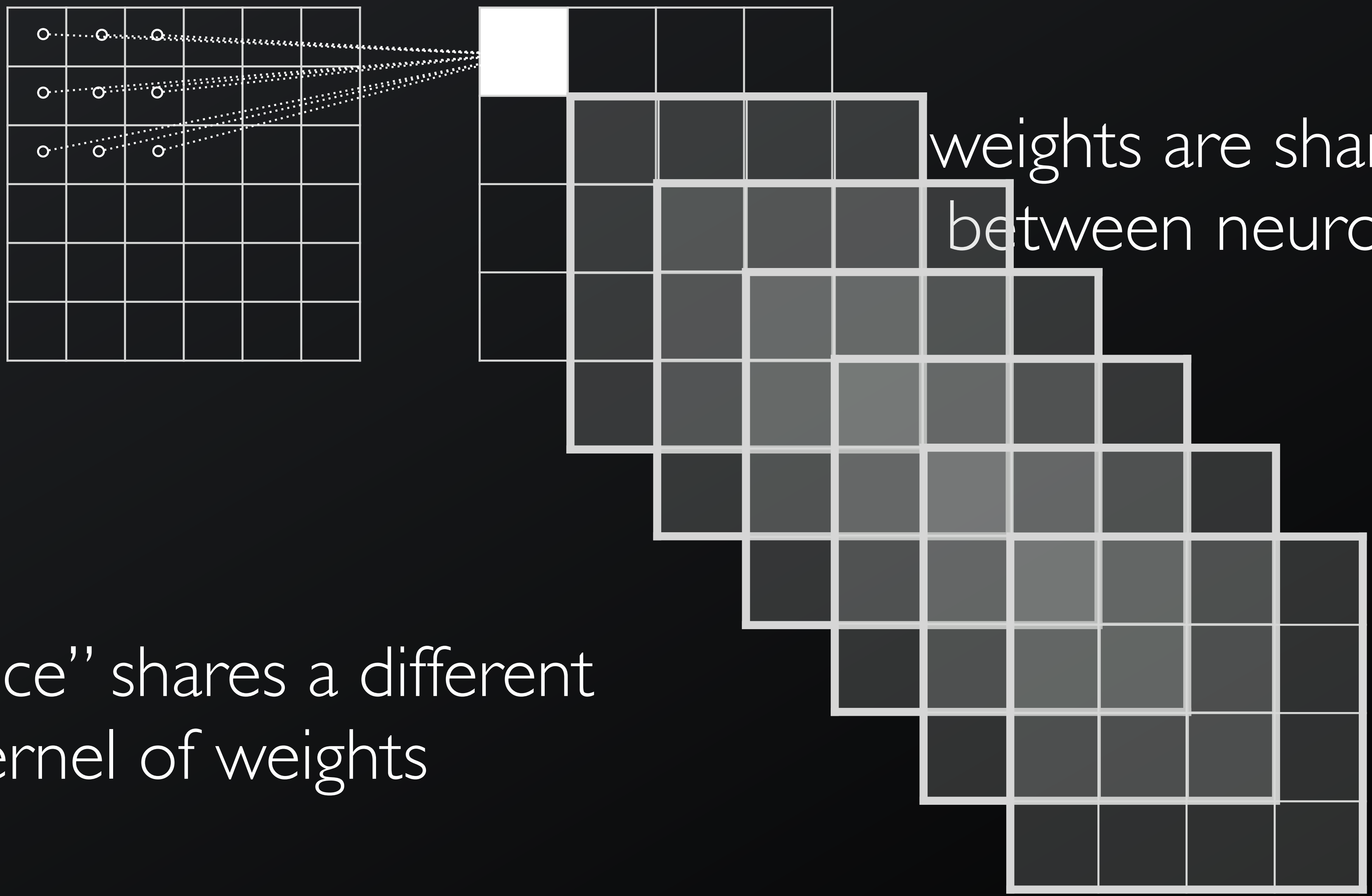


Flat layers

can’t preserve

spatial context

Convolutional Layers



locally connected
input

weights are shared
between neurons

each "slice" shares a different
kernel of weights

Dataset

700+ soundtracks with IMDB genre tags

130,000 samples, 9.2 seconds each

40 MFCCs x 100 frames

13,000 samples for each of 10 classes:

Action, Adventure, Comedy, Crime, Drama,
Fantasy, Musical, Romance, Sci-Fi , Thriller

Network Hyper-Parameters

Input Dimension	1 x 40 x 100
Convolutional Layer	32 filters 5 x 5
Max Pool Layer	2 x 2
Dropout	0.1
Convolutional Layer	64 filters 3 x 3
Max Pool Layer	2 x 2
Dropout	0.2
Convolutional Layer	128 filters 2 x 2
Max Pool Layer	2 x 2
Dropout	0.3
Convolutional Layer	256 filters 2 x 2
Max Pool Layer	2 x 2
Dropout	0.4
Dense Layer	50
Dropout	0.5
Dense Layer	50
Output Layer	10 (softmax)

Validation Size	0.2
Learning Rate	0.01 (linear decay)
Training Epochs	655

F1 Scores by Class for Fully-Connected and Convolutional Models:

Class	Fully-Connected Model	Convolutional Model	Change
Sci-Fi	0.008	0.220	**+0.212**
Romance	0.027	0.176	+0.149
Musical	**0.356**	**0.483**	+0.127
Thriller	0.042	0.137	+0.095
Crime	0.193	0.202	+0.09
Fantasy	0.113	0.202	+0.089
Drama	0.014	0.101	+0.085
Comedy	0.022	0.165	+.143
Action	0.233	0.219	-0.014
Adventure	0.204	0.196	-0.08

Shortfalls:

Even the best class (Musical) is still far short of perfect

Softmax output only assigns a single label; not a good representation of the actual data

Bagged Ensemble

Replace single multi-class network with 10 binary classifiers

Train each on full ‘positive’ subset
and a random sampling of ‘negative’ subset

Logistic output delivers value between 0.0 and 1.0

Rank predictions

Multi-Label Metrics:

Coverage Error: Average number of predicted labels needed to recall all ground-truth labels

Label Rank Average Precision: Average ratio of relevant labels predicted/total labels predicted

Coverage Error: **5.52**
Best possible CE: **2.33**

LRAP: **0.59** out of **1.00**

Average predictions over full track

Coverage Error: **5.02**
Best possible CE: **2.33**

LRAP: **0.61** out of **1.00**

F1 Scores by Class for Single Model and Ensemble (Bagged and Averaged):

Class	Single	Ensemble	Change
Drama	0.10	0.55	**+0.45**
Comedy	0.17	**0.61**	+0.44
Adventure	0.20	0.54	+0.34
Thriller	0.14	0.48	+0.34
Action	0.22	0.55	+0.32
Romance	0.18	0.44	+0.26
Sci-Fi	0.22	0.46	+0.24
Fantasy	0.20	0.38	+0.18
Crime	0.20	0.37	+0.17
Musical	0.48	**0.61**	+0.13

Q: Is accuracy necessary/sufficient?

"Good enough" can be more productive for discovery than "perfect".
(Lopresti 2001) (Kaminskas and Ricci 2012)

Even a highly accurate classifier may not necessarily be modeling the intended pattern.
(Sturm 2013)

Q: Where did the mis-labelings come from?

Poorly labeled ground-truth
Overdetermination of classes
Model weakness

Q: Are some ‘mis-labelings’ actually better than the ‘accurate’ labels?

Track	Predictions	Targets
Sneakers: The Hand-Off	Action, Sci-Fi, Adventure	Comedy, Crime, Drama
The Truman Show: Powaqqatsi — Anthem	Action, Sci-Fi, Thriller	Comedy, Sci-Fi
Jaws: One Barrel Chase	Comedy, Romance, Fantasy	Drama, Thriller
O Brother, Where Art Thou?: Down in the River	Drama, Fantasy, Crime	Crime

Conclusions

Film score genres can be modeled successfully
without establishing high-level features like music genre, harmony, melody

Overlap between film genre and film score genre is
significant enough to be useful

Ensemble of convolutional models is clear winner out of algorithms evaluated

Future Research

Raw spectrograms: preserve pitch and phase information, make it possible to 'hear' the filters and parse the reasoning of the network

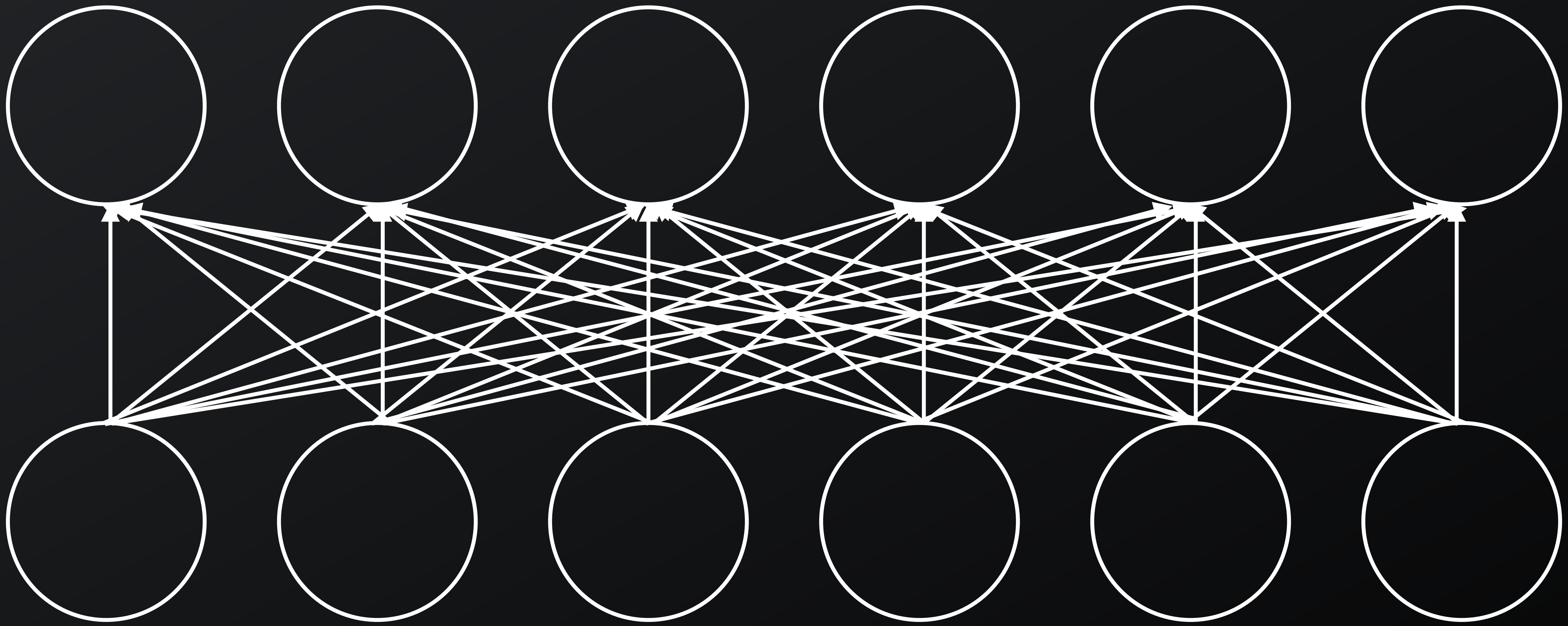
Better labeling on the dataset: more accurate and fine-grained labels, labels at the track level, multiple labels for training

Recurrent Network: RNN can preserve feature interaction over time, whereas CNN merely recognizes presence of complex features

Thanks!

Preventing Overfitting with Dropout

Neurons tend to split up the work and “co-adapt”, converging on a low-cost solution that generalizes poorly

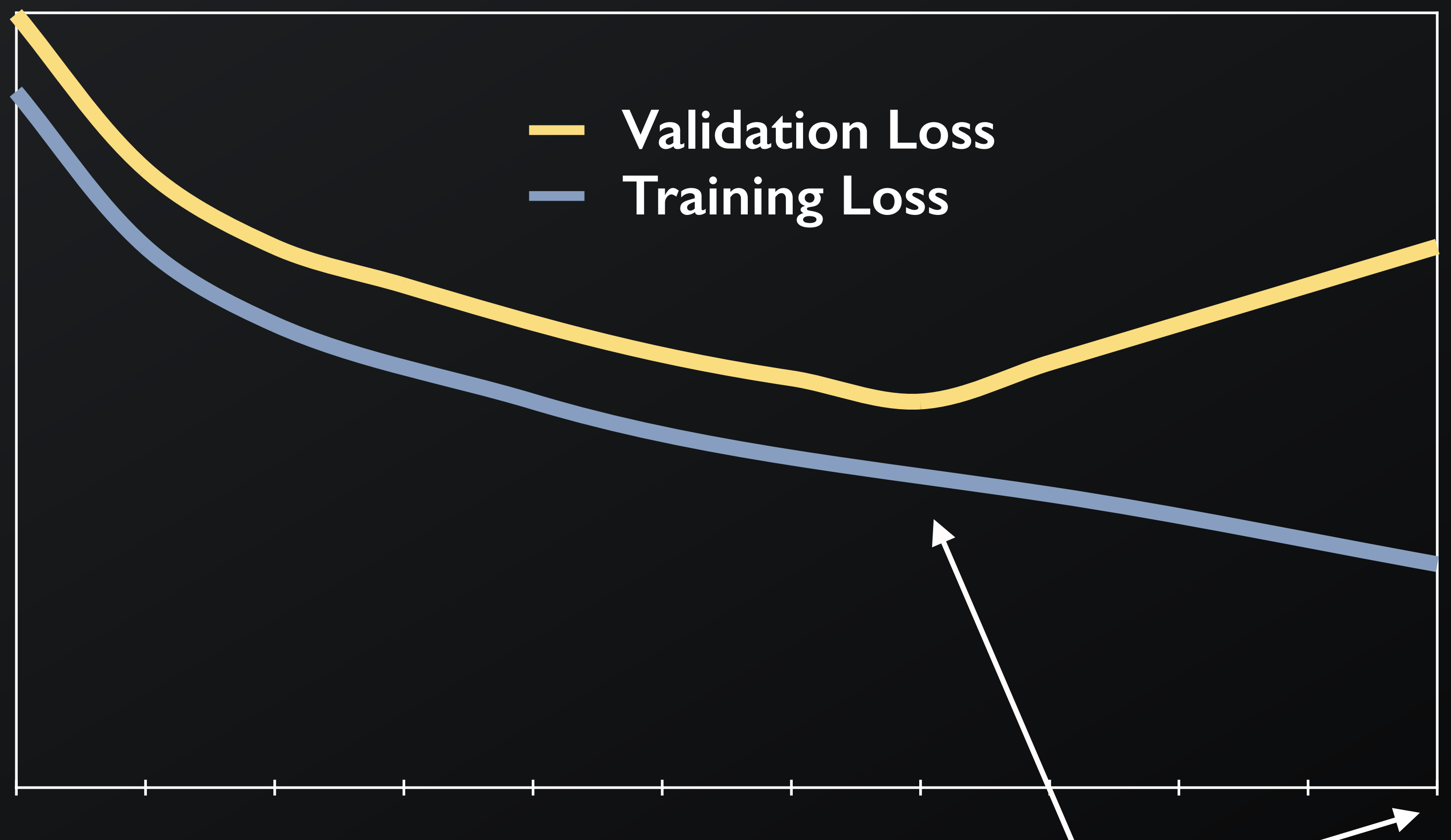


Random dropout forces individual neurons to be more ‘descriptive’

Preventing Overfitting with Validation and Early Stopping

Validation set is held out from training and test set

Stop training after validation error increases for given number of epochs, set parameters to best model



stop here

use these parameters

Exhaustive Search

Evaluate all combinations of a pre-set dictionary of hyper-parameters, e.g.,

```
{‘learning_rate’: (0.1,0.01,0.001),
’n_hidden’: (50,100,200),
‘batch_size’: (200,400,800)}
```

Combinatorial, random search sampled from a range can be better