Automatic Music Tagging and Classification

Minz Won
minz.won@upf.edu

MTG
Music Technology
Group

About me

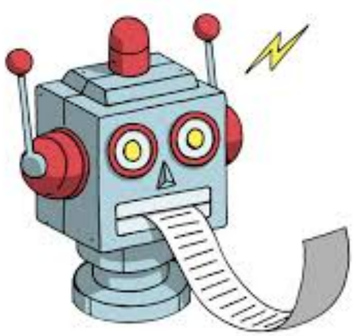
- PhD student at MTG
- Worked at:
    - Pandora (United States)
    - Naver Corp. (South Korea)
    - Kakao Corp. (South Korea)

Contents

- Introduction
- Music tagging models
- Transfer learning
- Limitations
- Lab

Contents

Music Classification



Blues

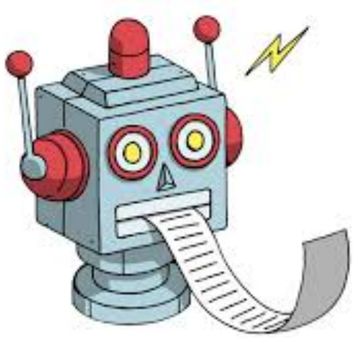
Classical

Hiphop

Jazz

Rock

Automatic Music Tagging



Rock

Guitar

Male vocal

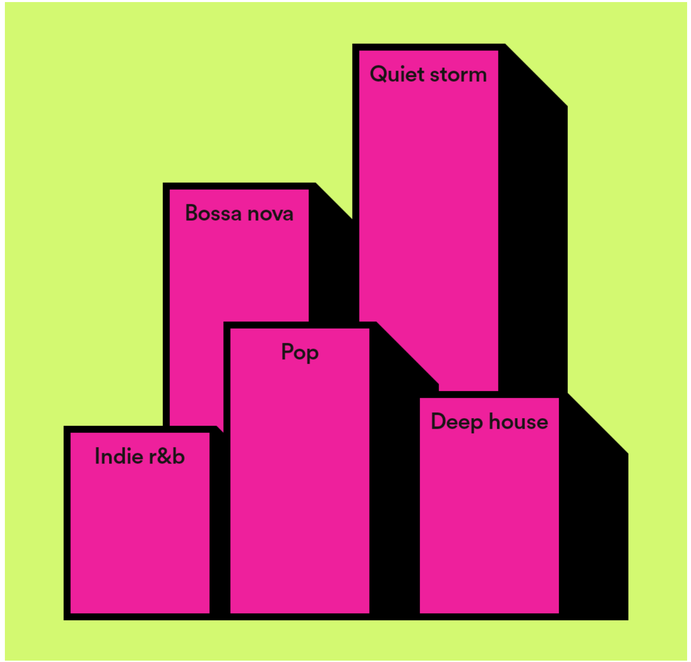
Female vocal

80s

90s

Why do we need this?

Why do we need this?



Recommendation

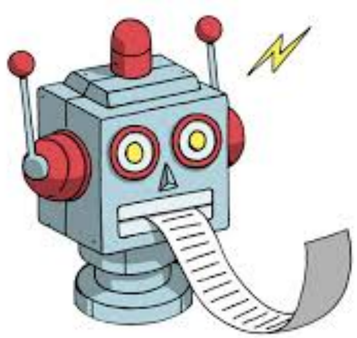
https://www.spotify.com/

Why do we need this?

Retrieval

https://www.beatport.com/

Why do we need this?



Rock

Guitar

Male vocal

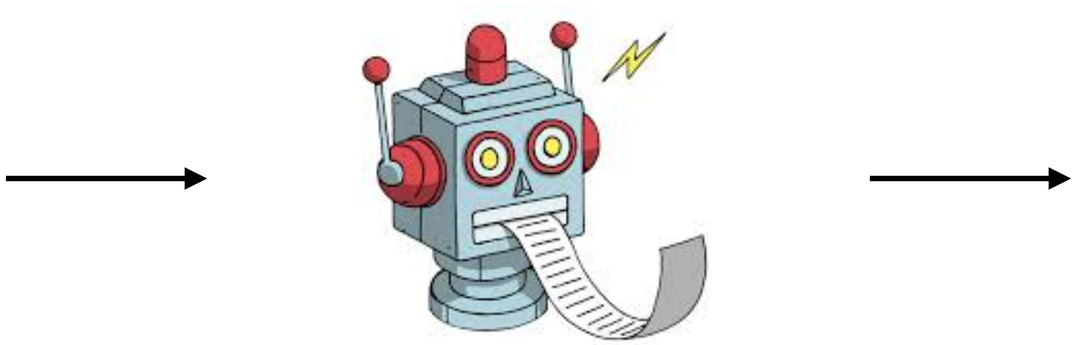
Female vocal

80s

90s

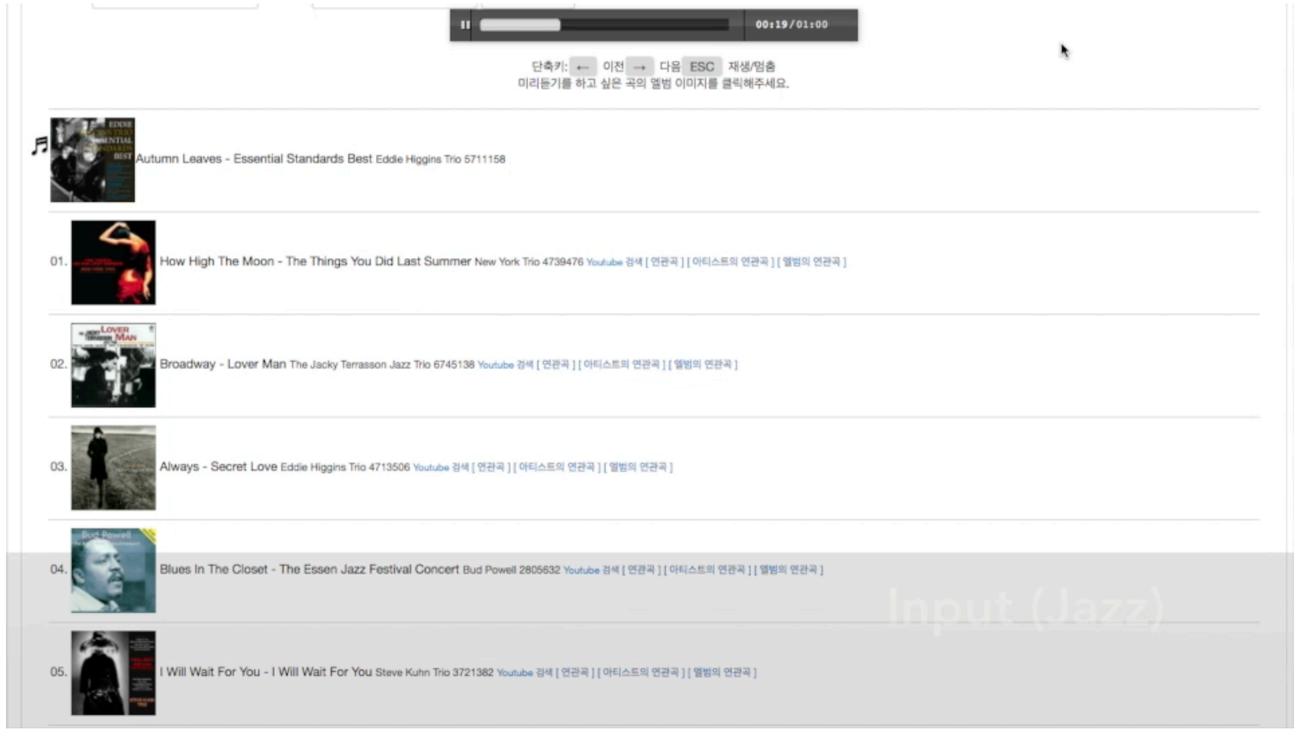
Similarity-based search

Why do we need this?



Similarity-based search

Why do we need this?



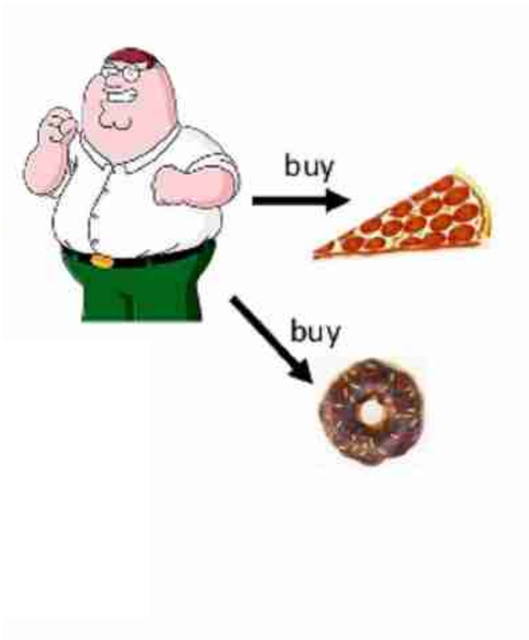
Similarity-based search

https://www.melon.com/

Why do we need this?

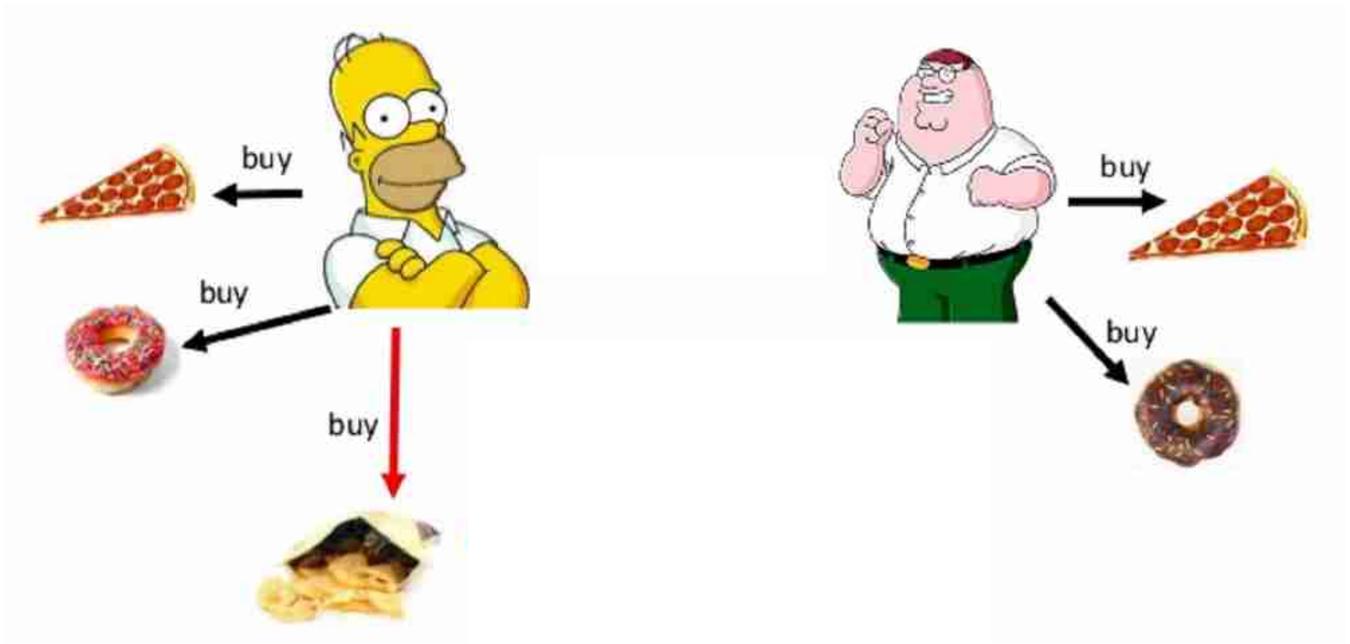
It’s fully content-based!

Why do we need this?



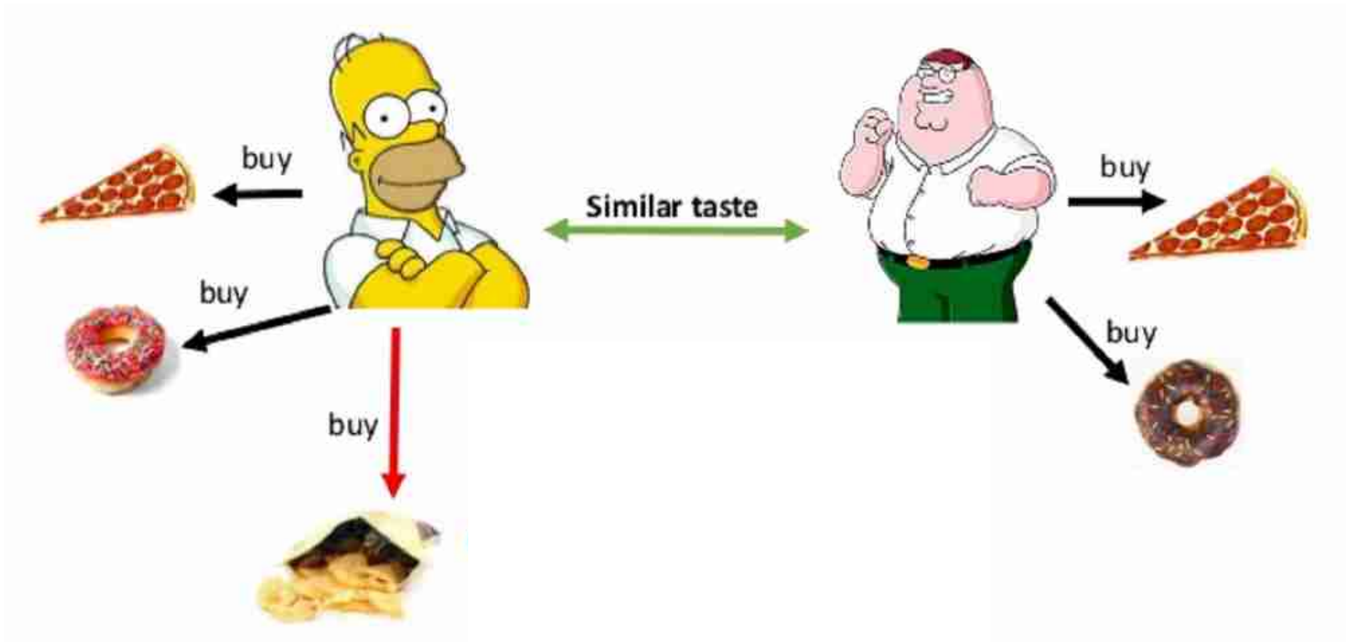
Collaborative filtering

Why do we need this?



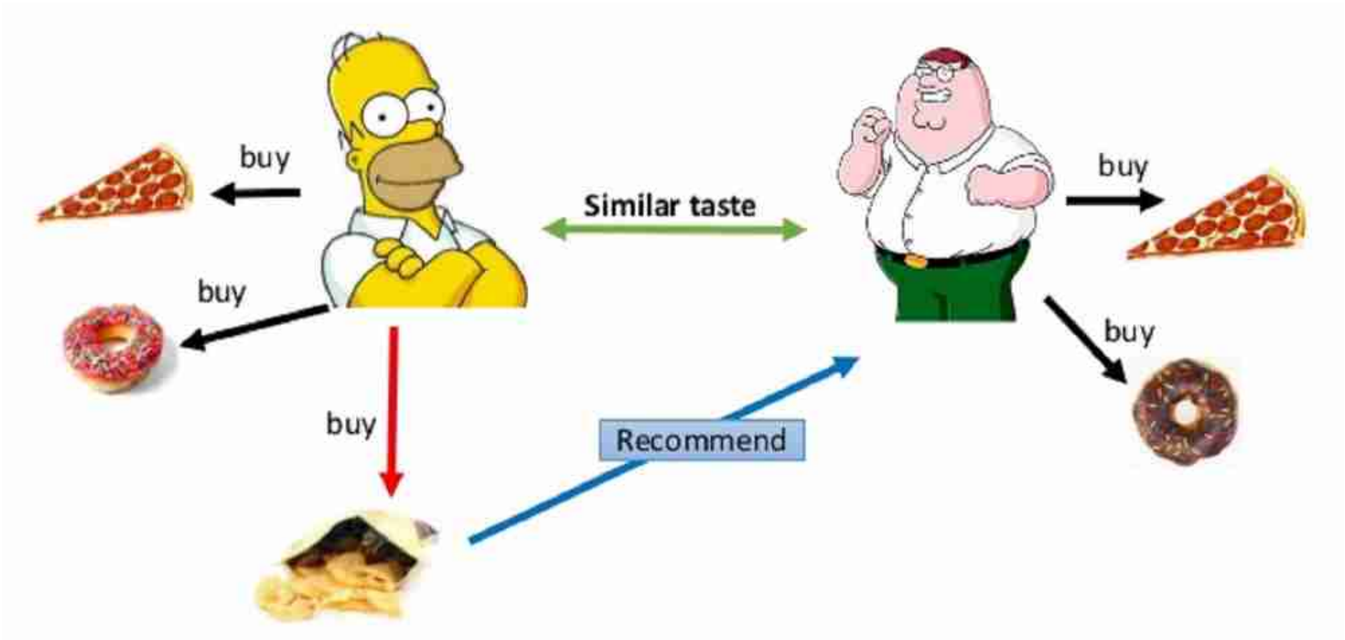
Collaborative filtering

Why do we need this?



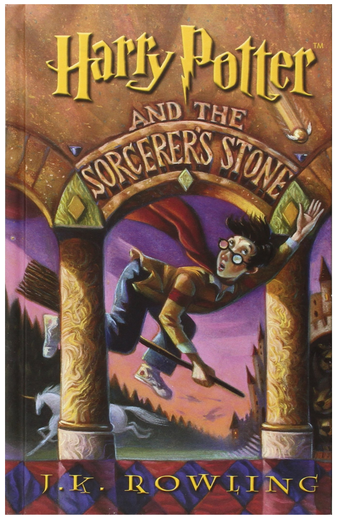
Collaborative filtering

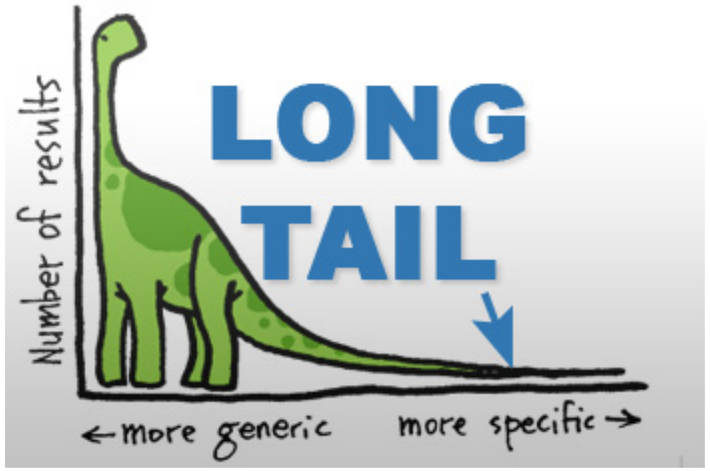
Why do we need this?



Collaborative filtering

Why do we need this?

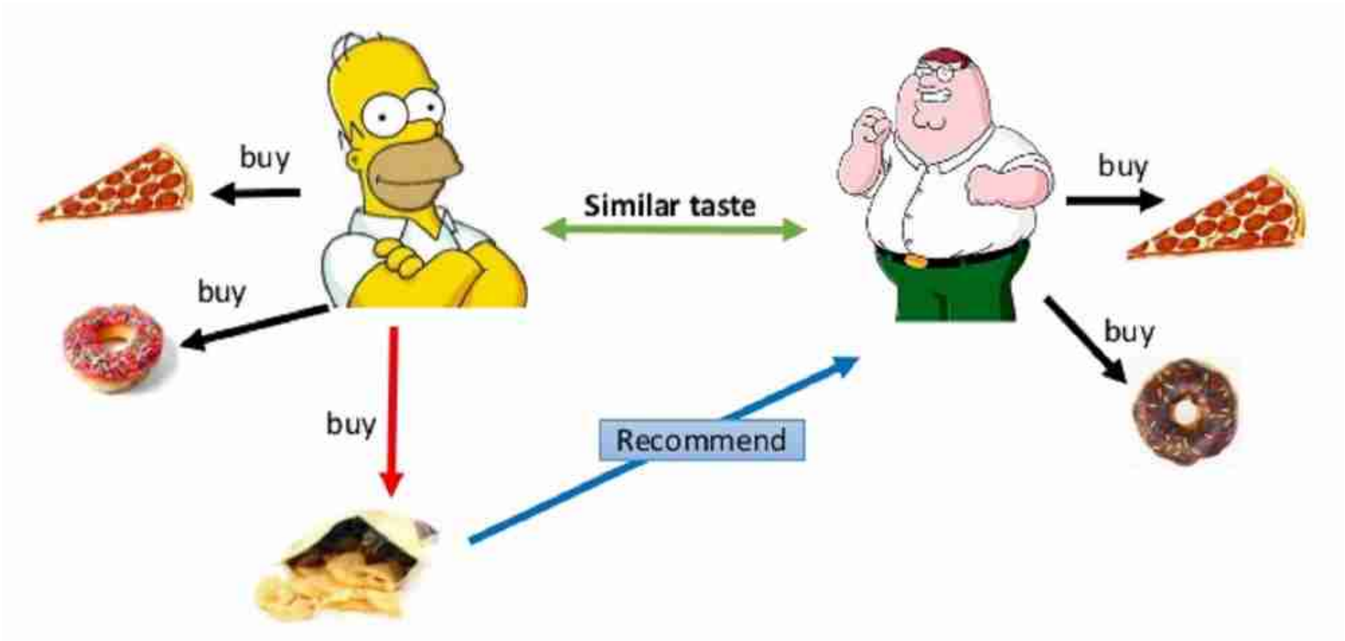




Popularity bias

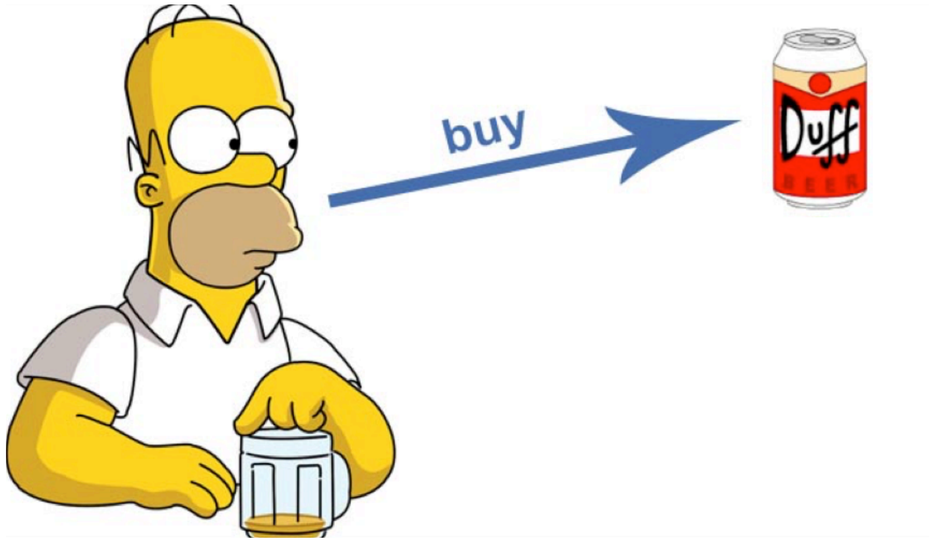
Why do we need this?

New item



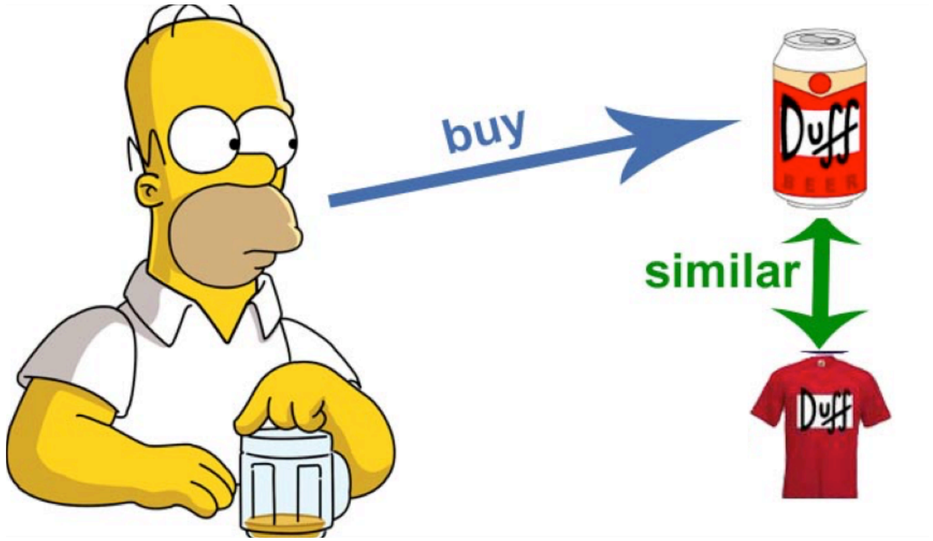
Cold-start problem

Why do we need this?



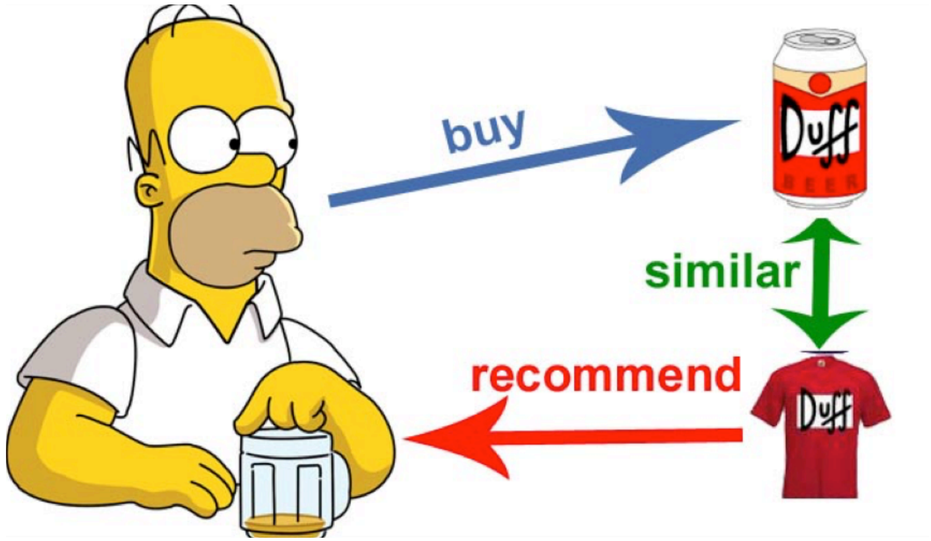
Content-based

Why do we need this?



Content-based

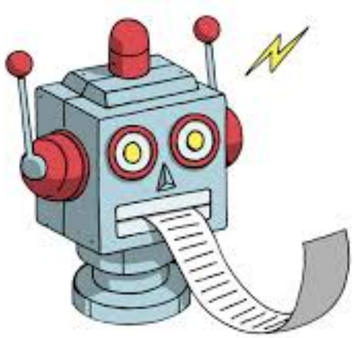
Why do we need this?



Content-based

Contents

Automatic Music Tagging



Rock

Guitar

Male vocal

Female vocal

80s

90s

Traditional approaches

Feature design & extraction

Classifier (e.g., SVM)

Rock

Guitar

Male vocal

Female vocal

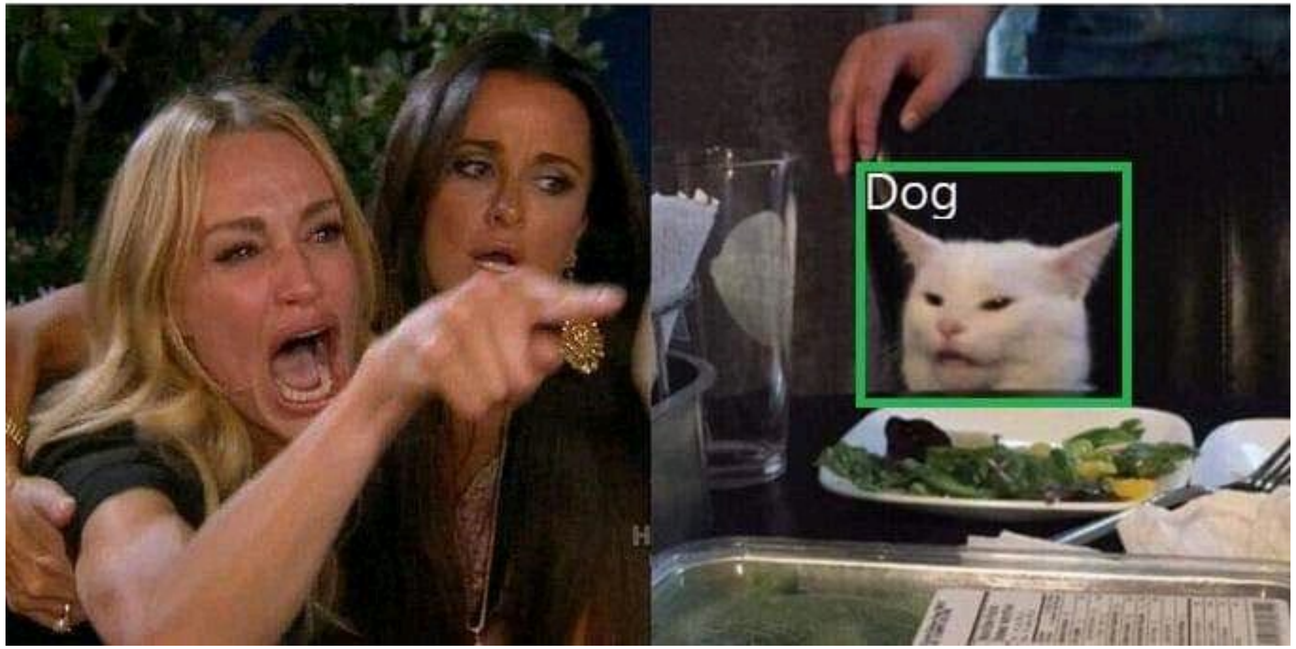
80s

90s

Recent approaches

People with no idea about AI saying it will take over the world:

My Neural Network:



Recent approaches

Deep neural networks

Rock

Guitar

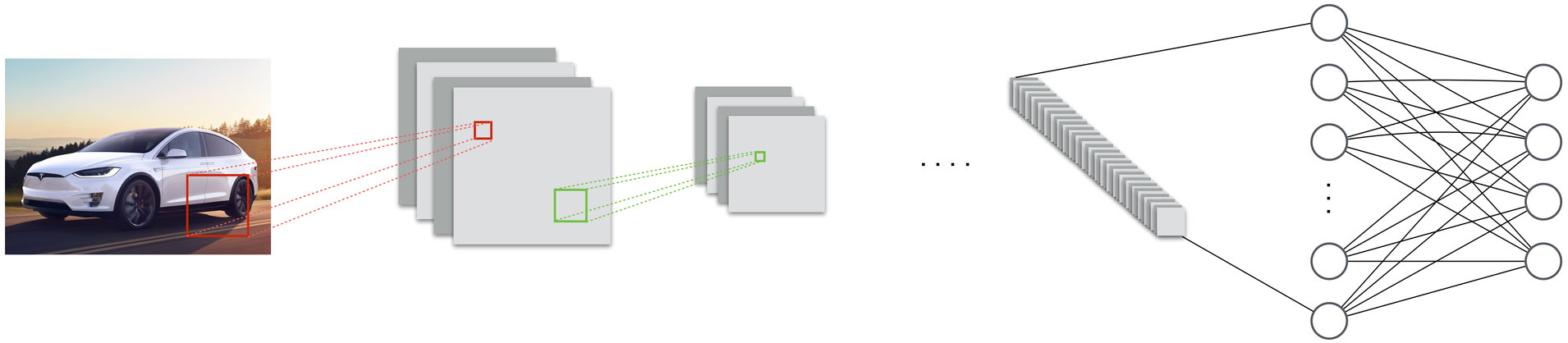
Male vocal

Female vocal

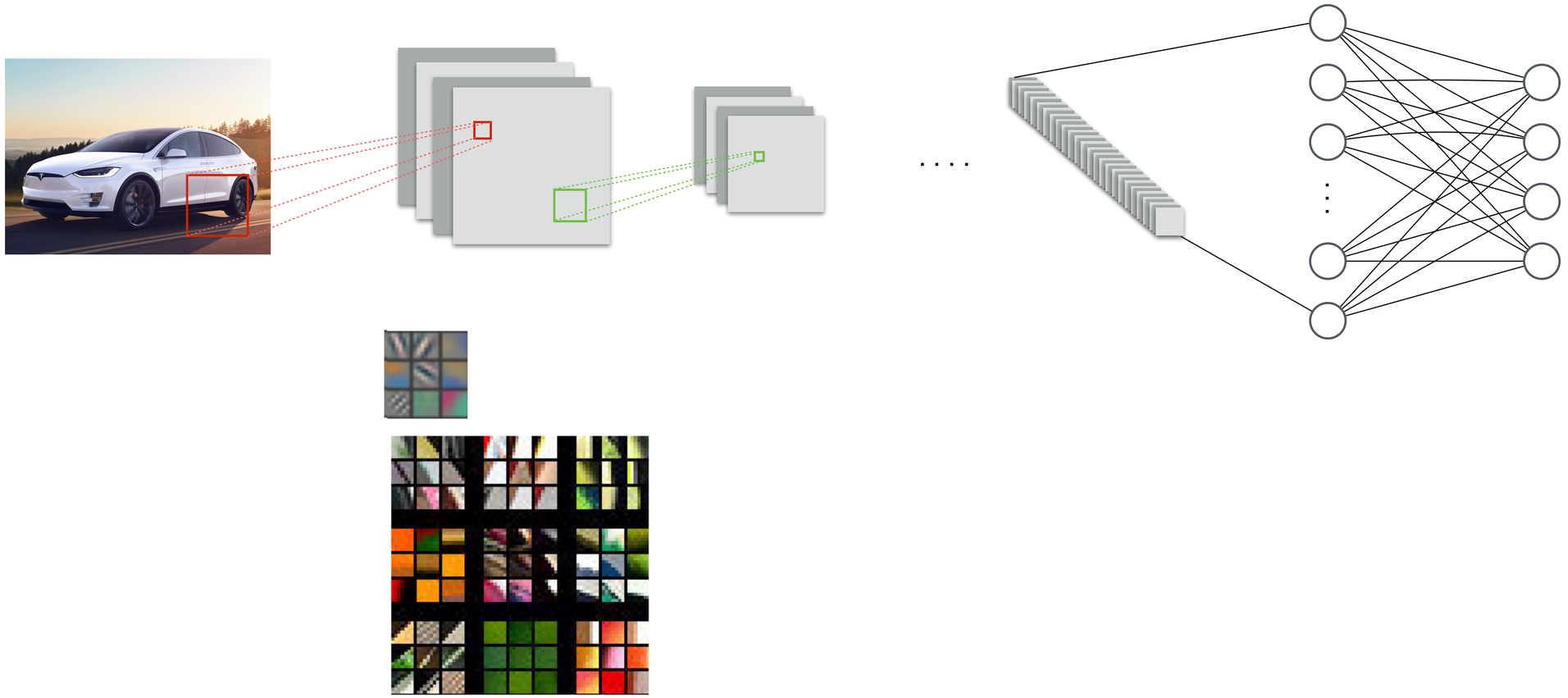
80s

90s

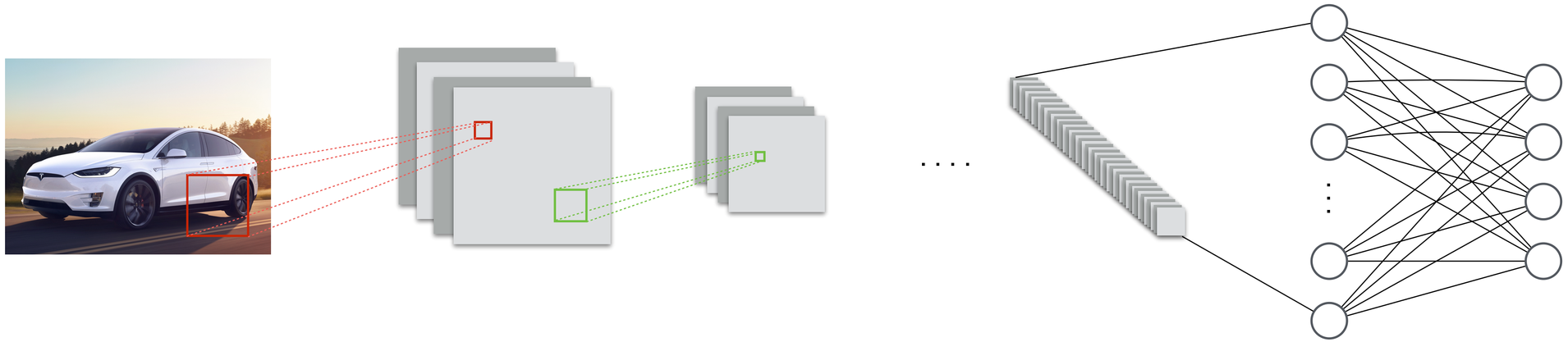
Convolutional Neural Networks

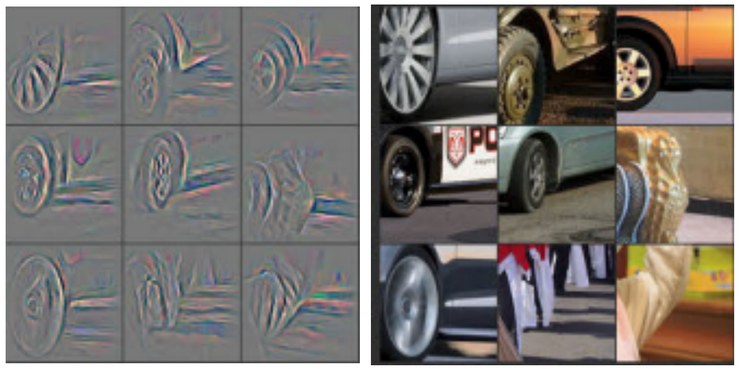


Convolutional Neural Networks

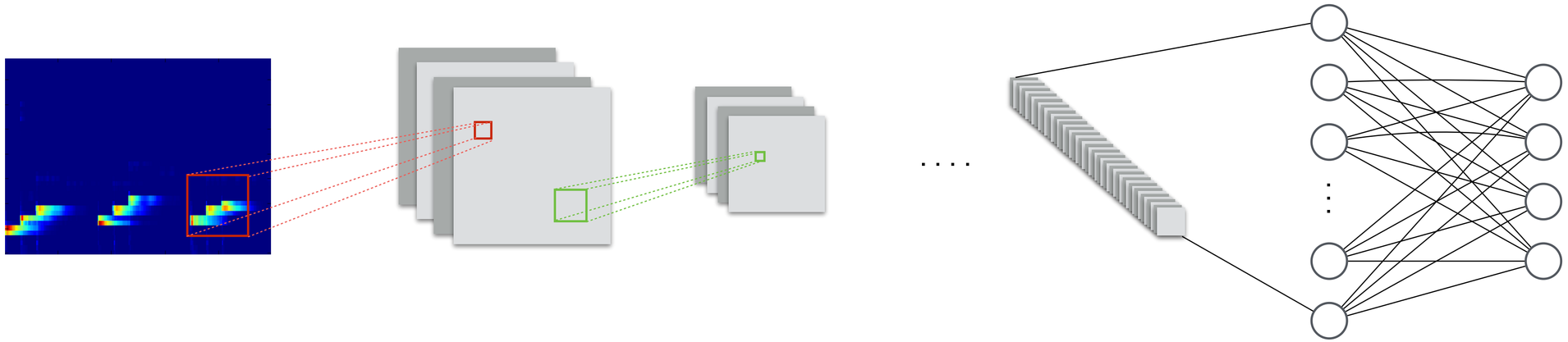


Convolutional Neural Networks





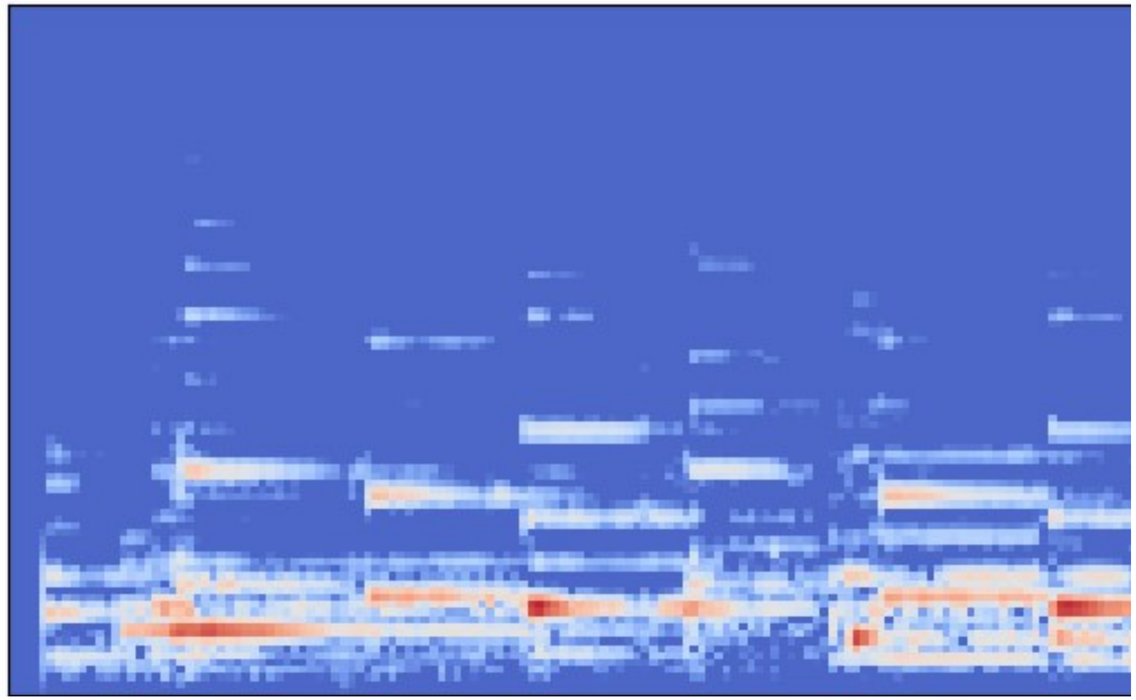
Convolutional Neural Networks



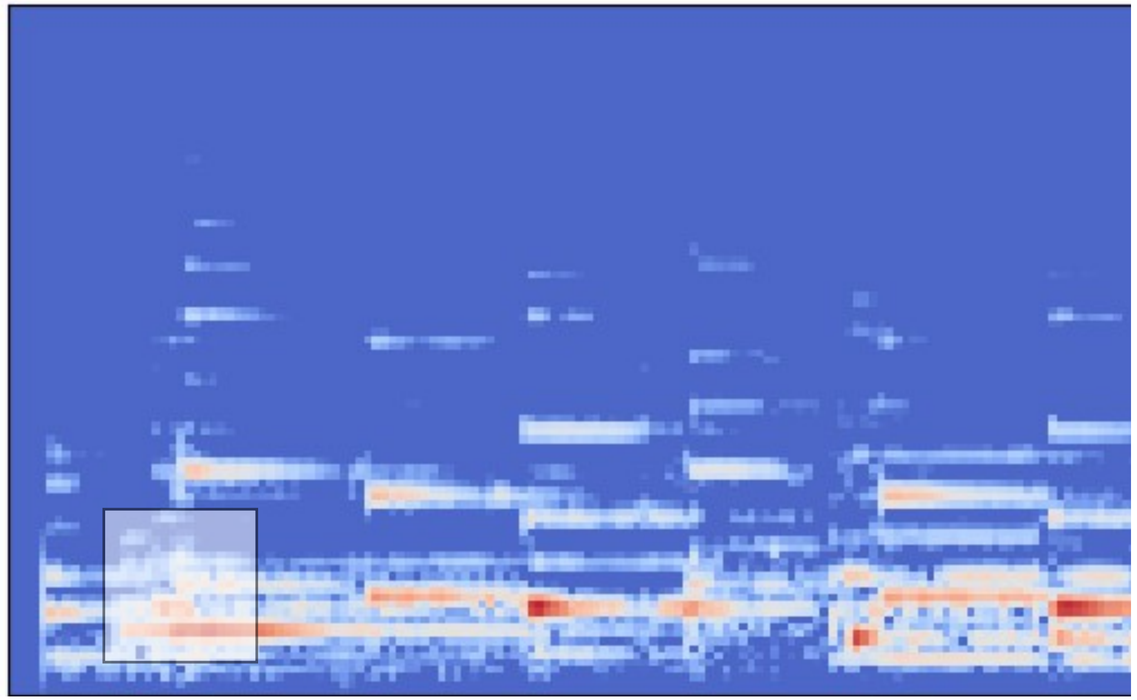
Fully Convolutional Networks (Choi et al., 2016)

Choi, Keunwoo, George Fazekas, and Mark Sandler. "Automatic tagging using deep convolutional neural networks." (2016)

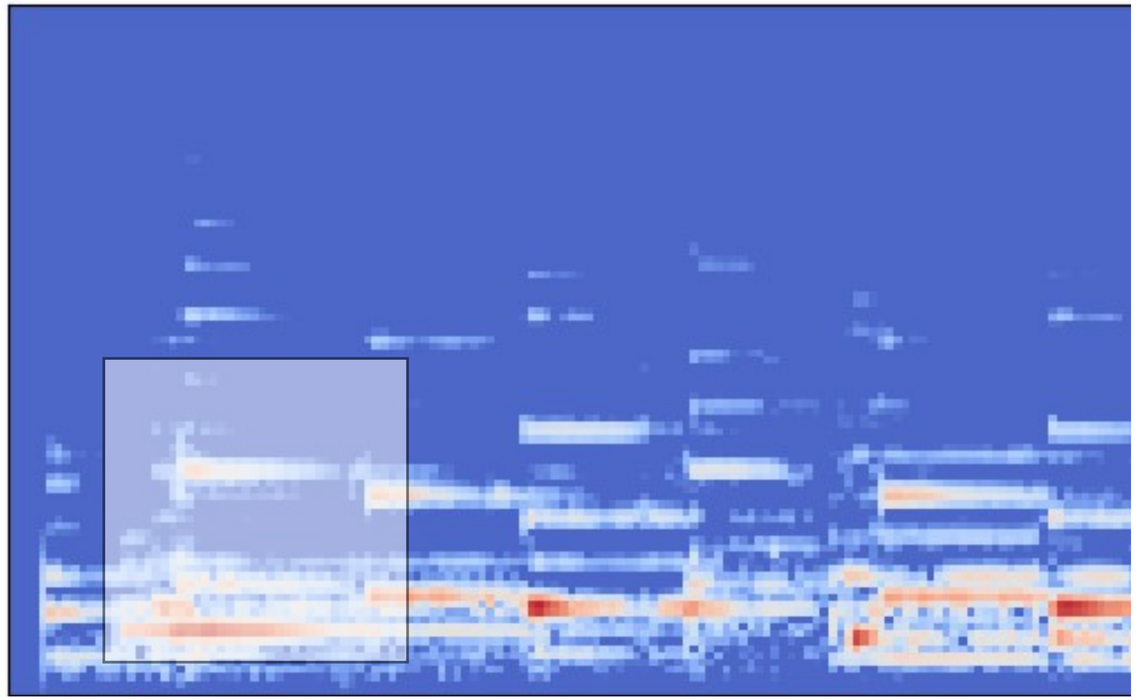
Convolutional Neural Networks



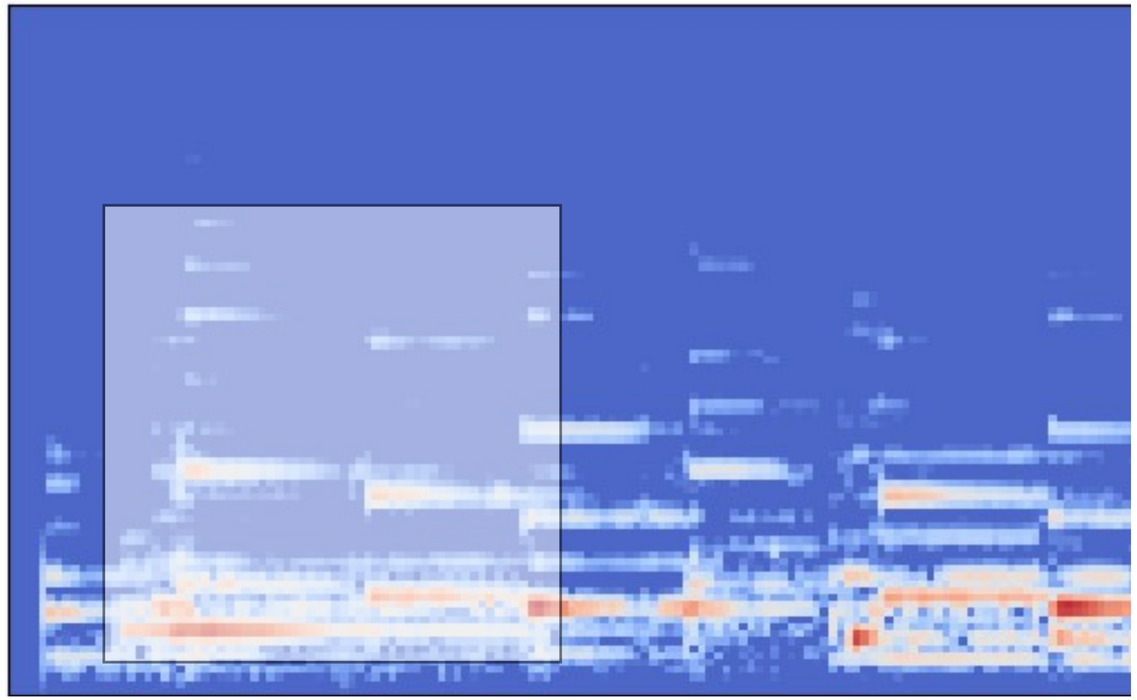
Convolutional Neural Networks



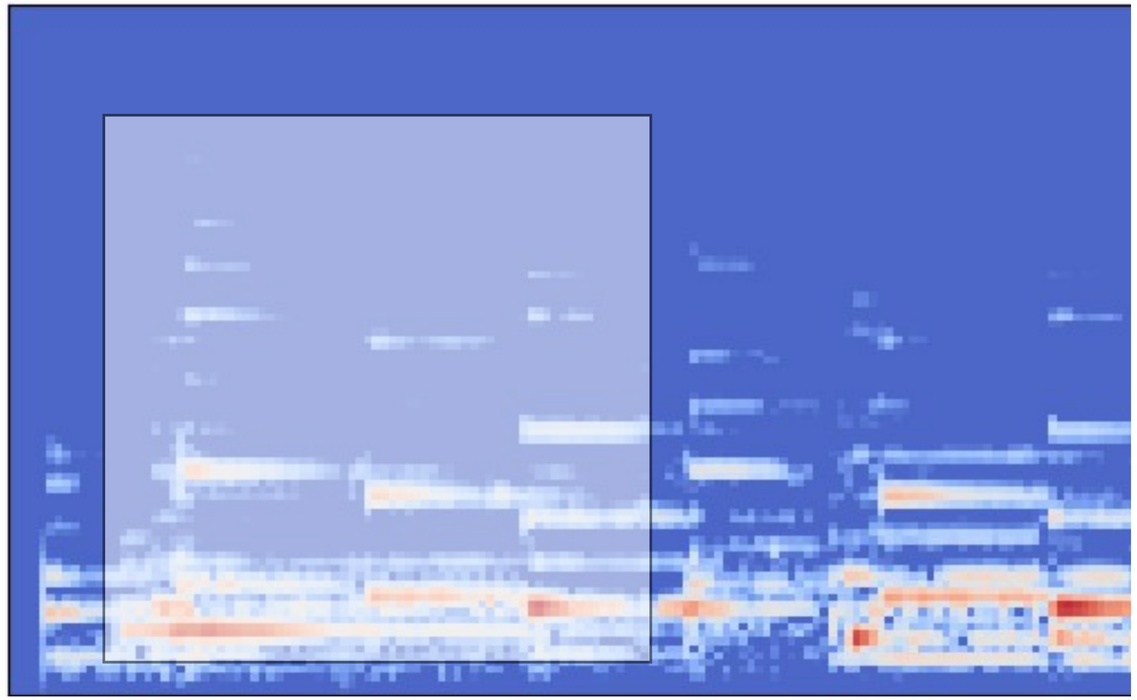
Convolutional Neural Networks



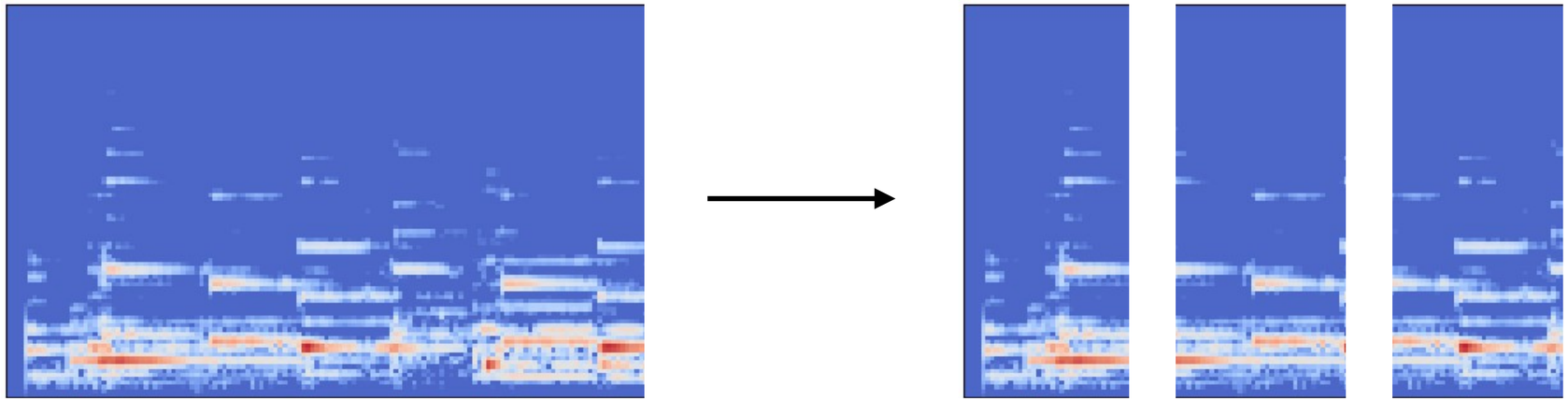
Convolutional Neural Networks



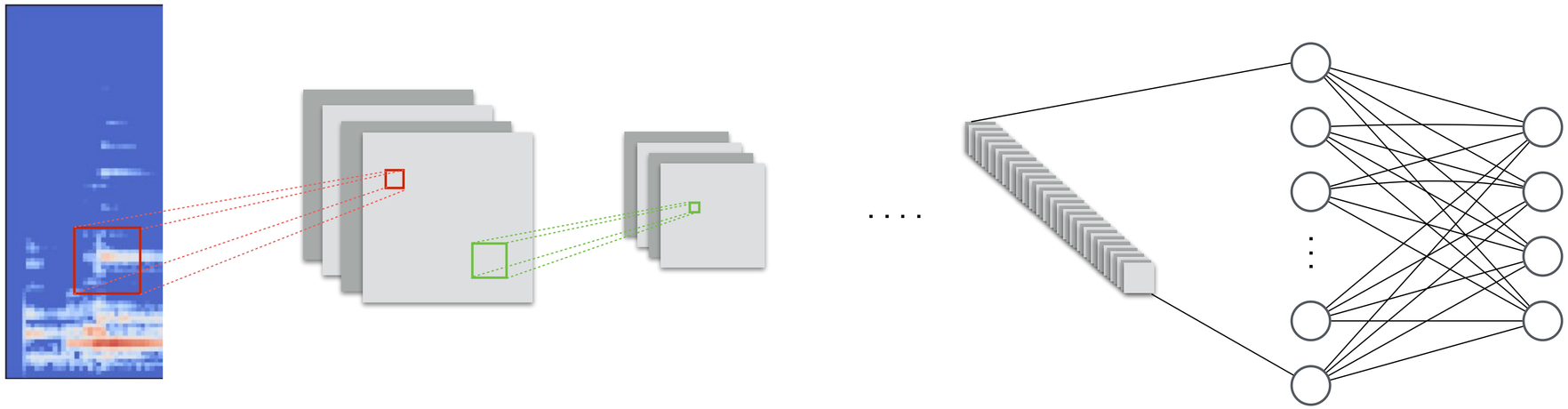
Convolutional Neural Networks



Bag-of-chunk



Bag-of-chunk



Bag-of-chunk

Rock

Guitar

Male vocal

Female vocal

80s

90s

Rock

Guitar

Male vocal

Female vocal

80s

90s

Rock

Guitar

Male vocal

Female vocal

80s

90s

Rock

Guitar

Male vocal

Female vocal

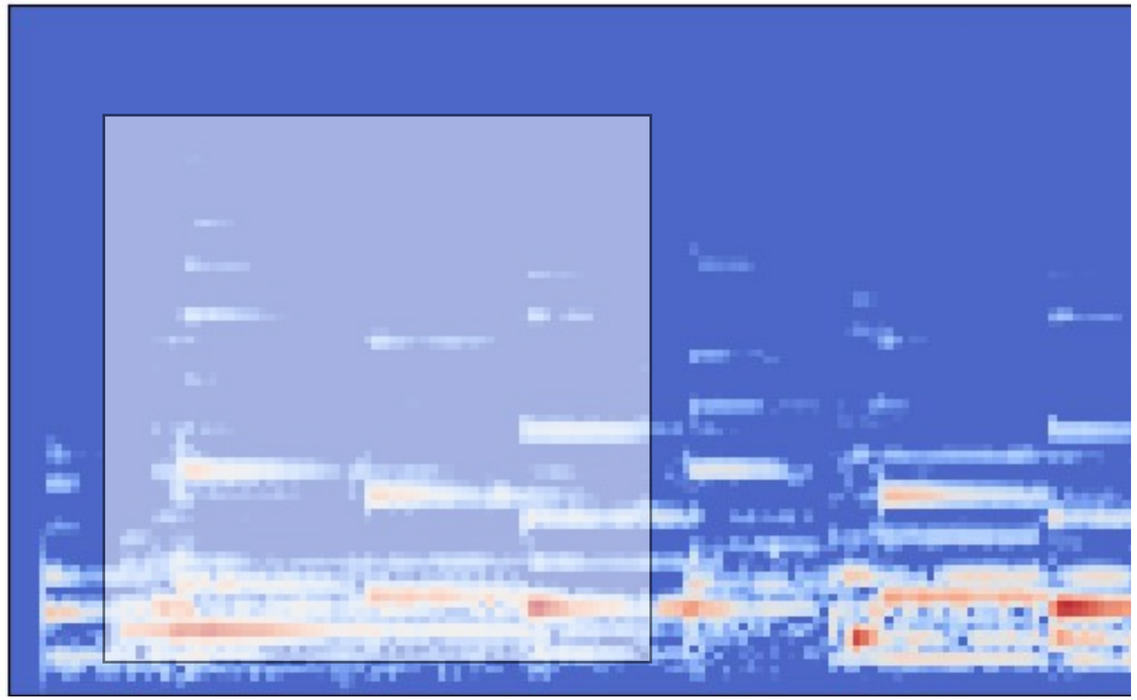
80s

90s

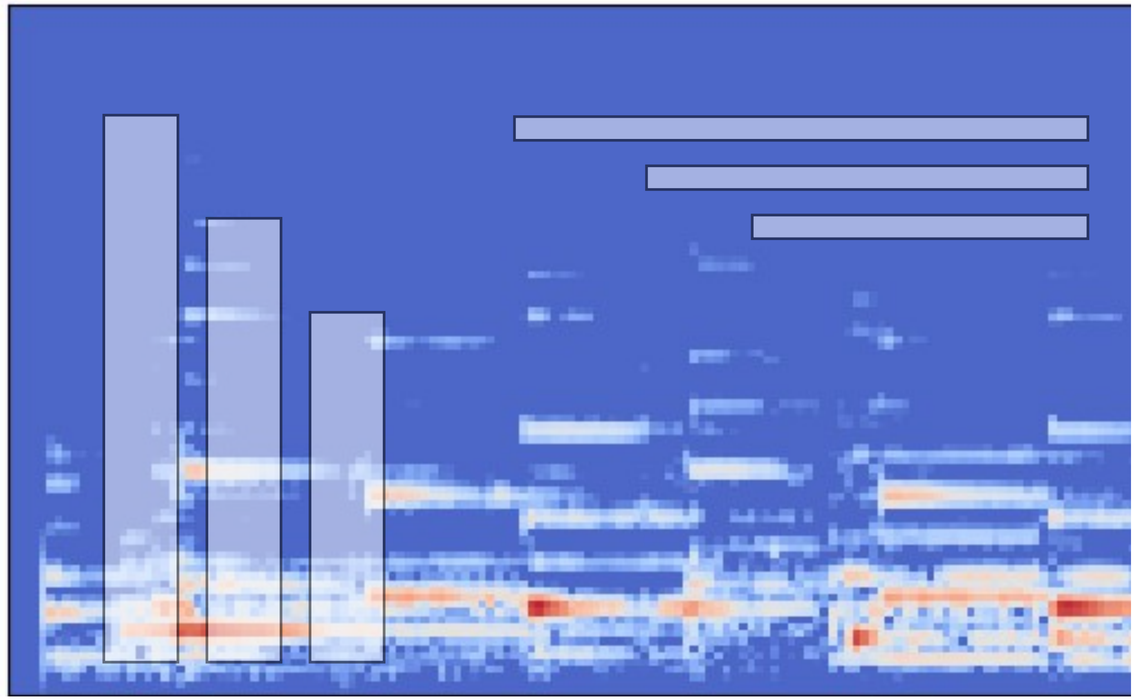
Bag-of-chunk

- Higher performance
- More robust
- A sort of data augmentation
- e.g., 1 million songs x 10 chunks = 10 million chunks

Let’s put some domain knowledge



Let’s put some domain knowledge



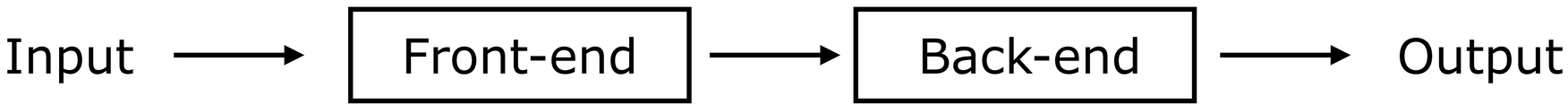
Musicnn (Pons et al., 2018)

Pons, Jordi, et al. "End-to-end learning for music audio tagging at scale." (2017).

Let’s put some domain knowledge

- More efficient
- Good performance with a limited size of data
- However, with sufficient amount of data and computing power, assumption-free models still work better

Music is sequential!



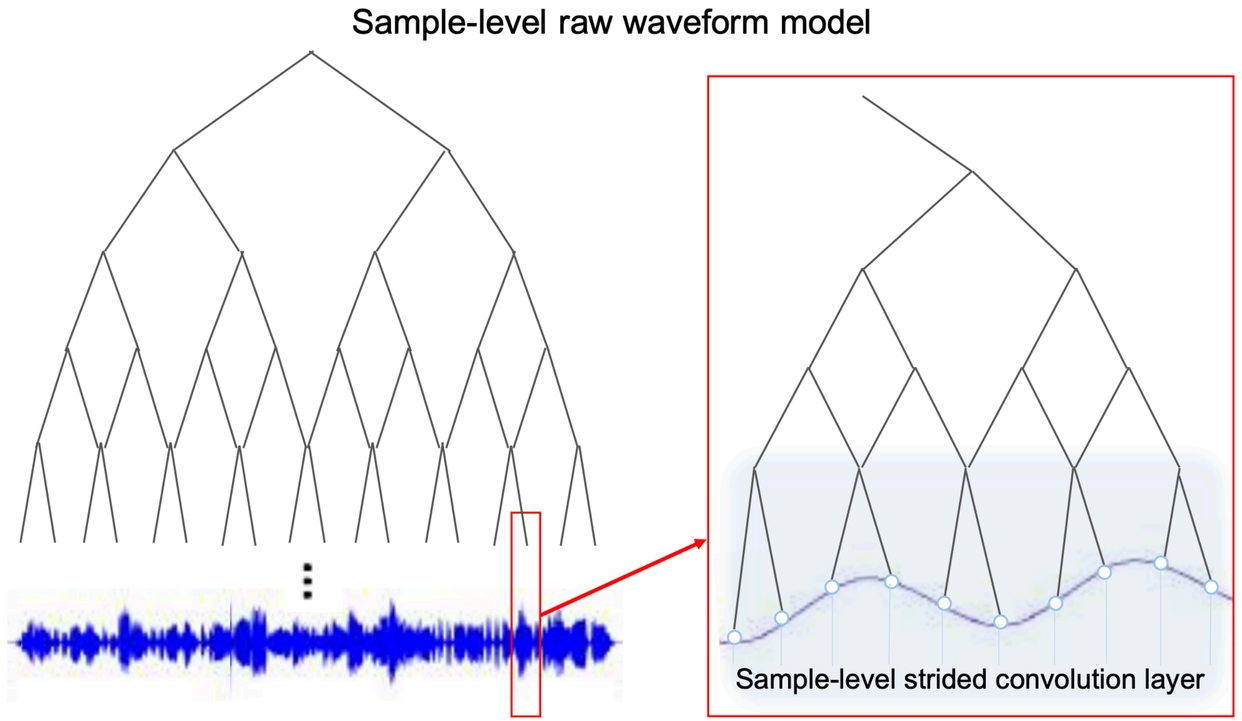
Music is sequential!

- Front-end: capture acoustic features with CNN
- Back-end: temporally summarize extracted features
- CNN front-end + RNN back-end (Choi et al. 2017)
- CNN front-end + self-attention back-end (Won et al. 2019)

Choi, Keunwoo, et al. "Convolutional recurrent neural networks for music classification.” (2017)

Won, Minz, Sanghyuk Chun, and Xavier Serra. "Toward interpretable music tagging with self-attention." (2019).

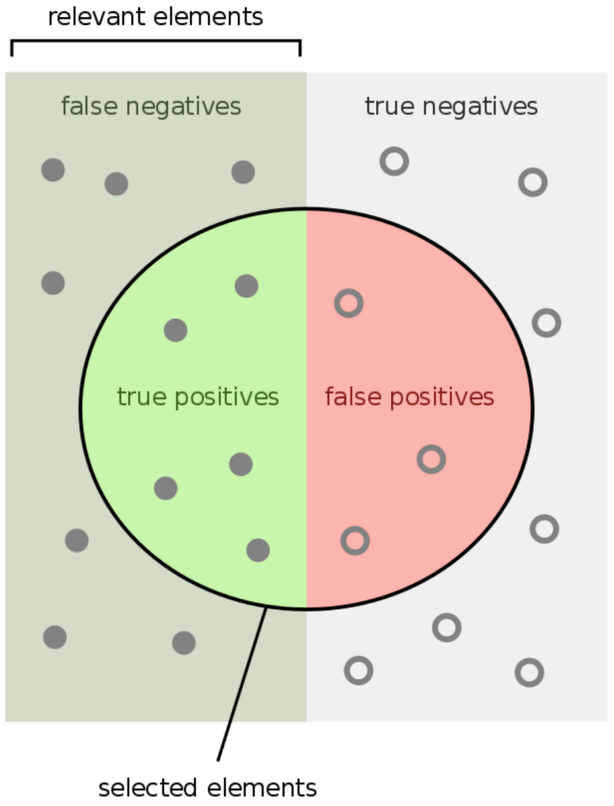
Let’s make it end-to-end



Sample-level CNN (Lee et al., 2017)

Lee, Jongpil, et al. "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms." (2017).

Evaluation



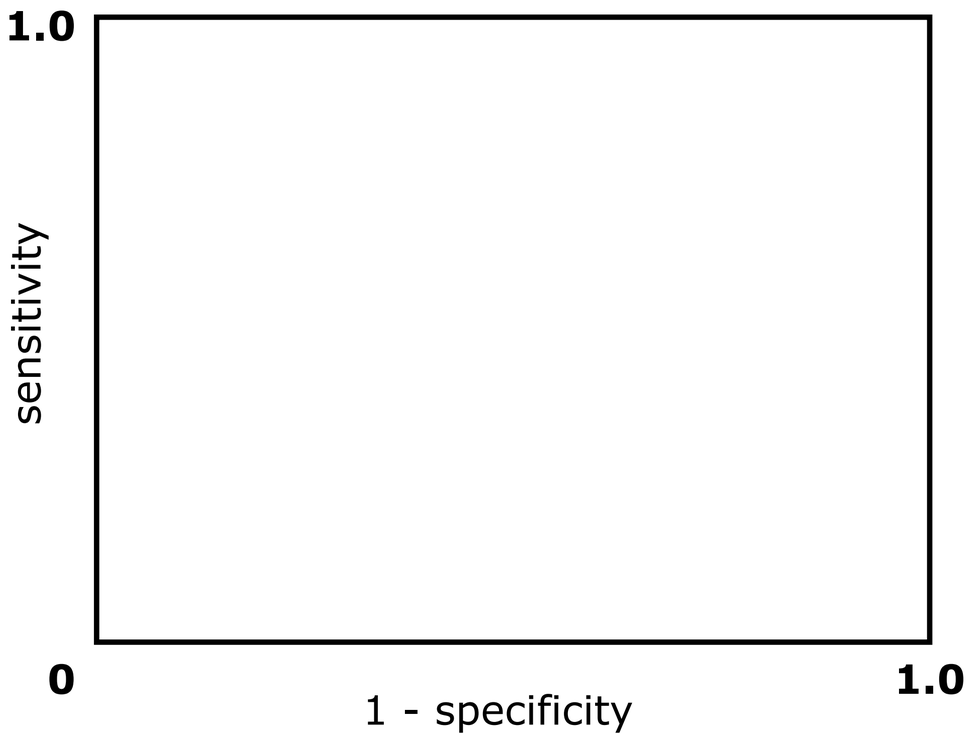
How many relevant items are selected? e.g. How many sick people are correctly identified as having the condition.

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{relevant elements}}$$

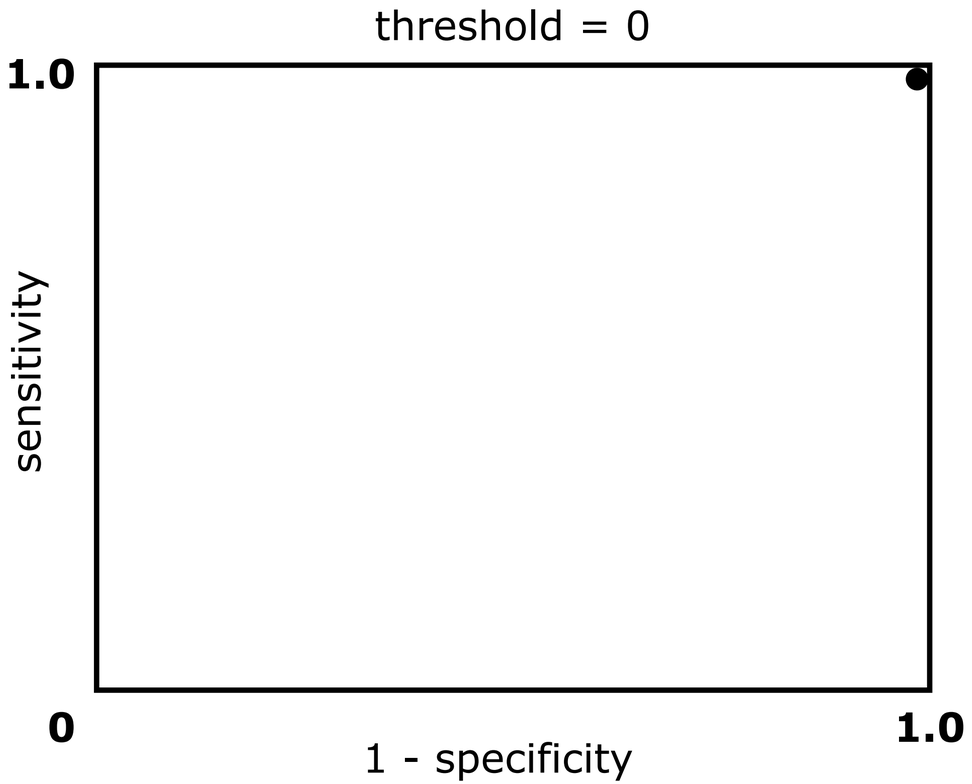
How many negative selected elements are truly negative? e.g. How many healthy peple are identified as not having the condition.

$$\text{Specificity} = \frac{\text{true negatives}}{\text{false positives}}$$

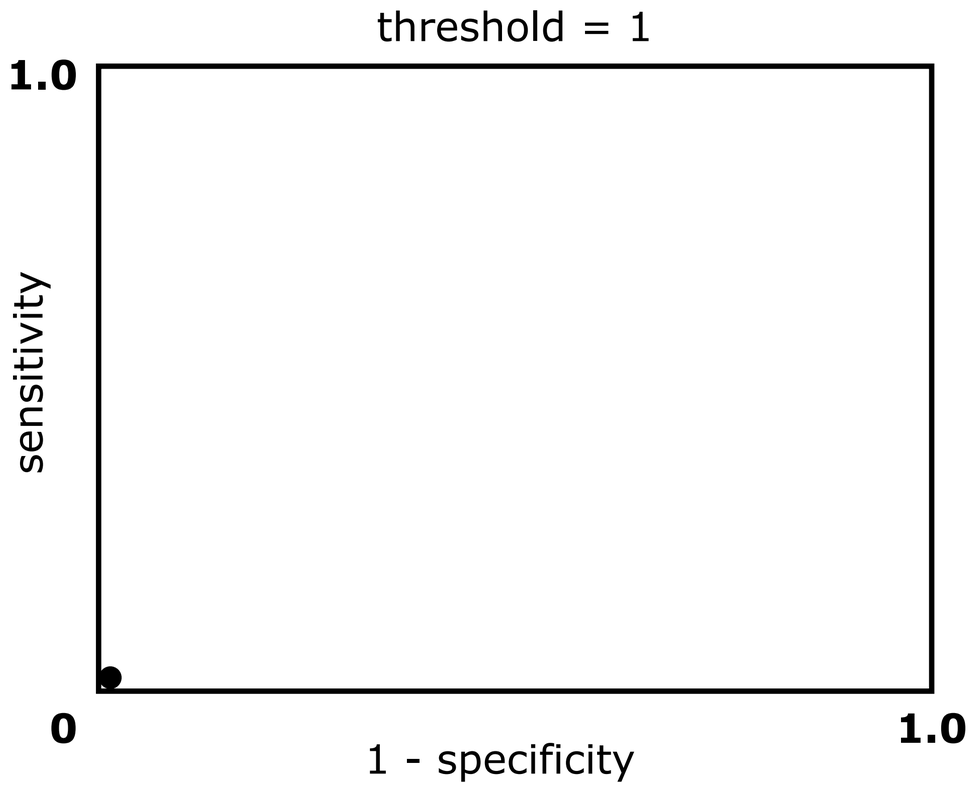
Evaluation



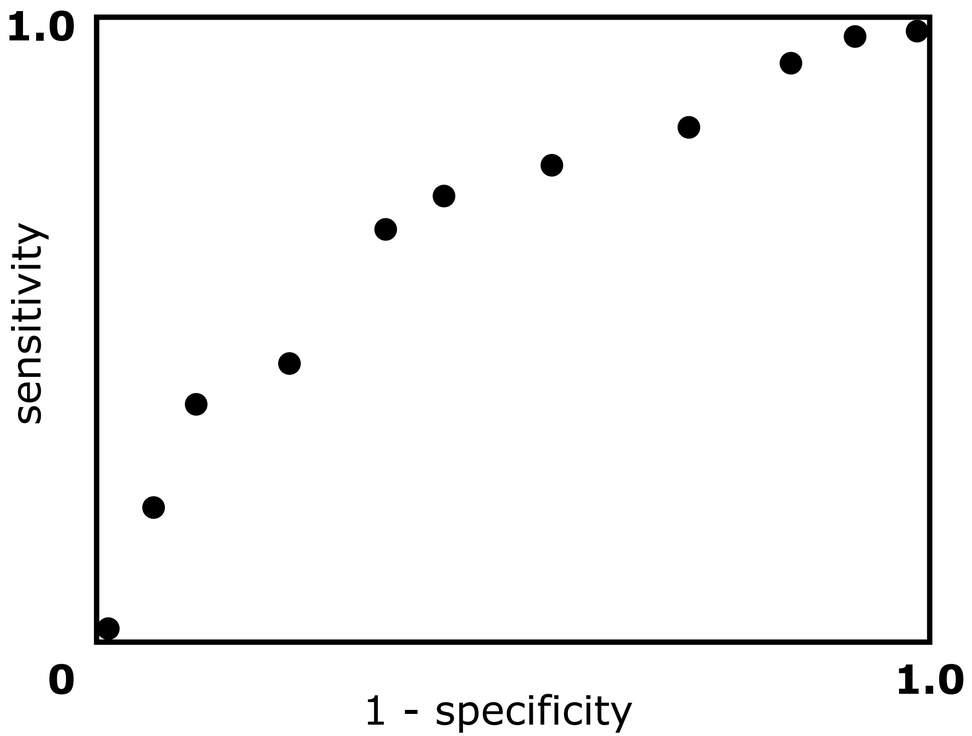
Evaluation



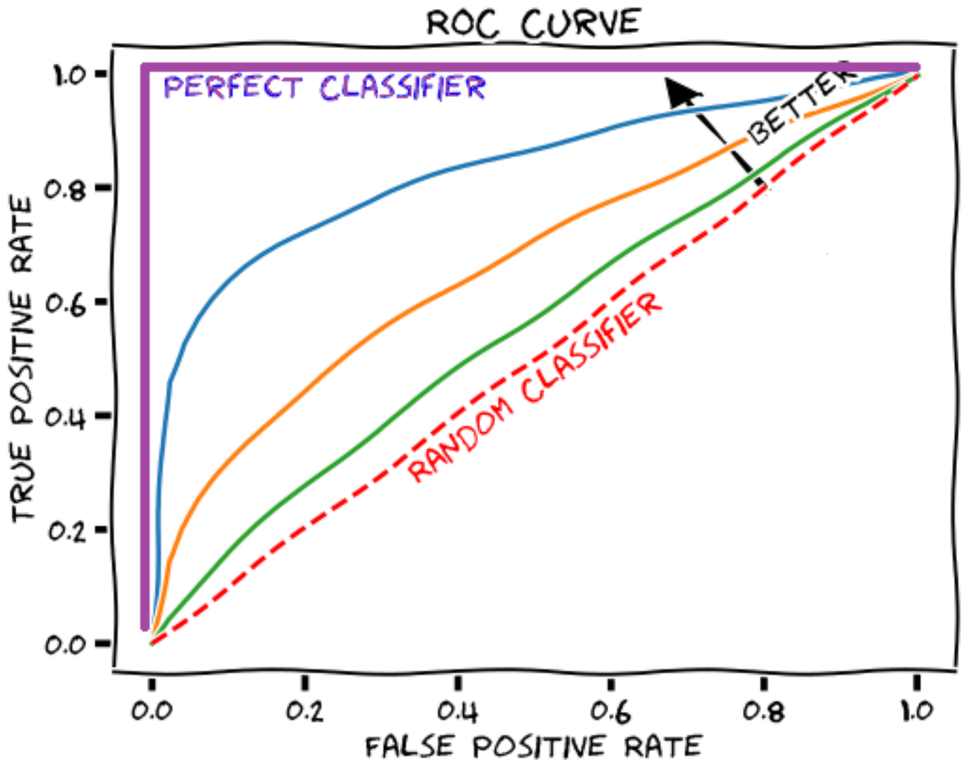
Evaluation



Evaluation

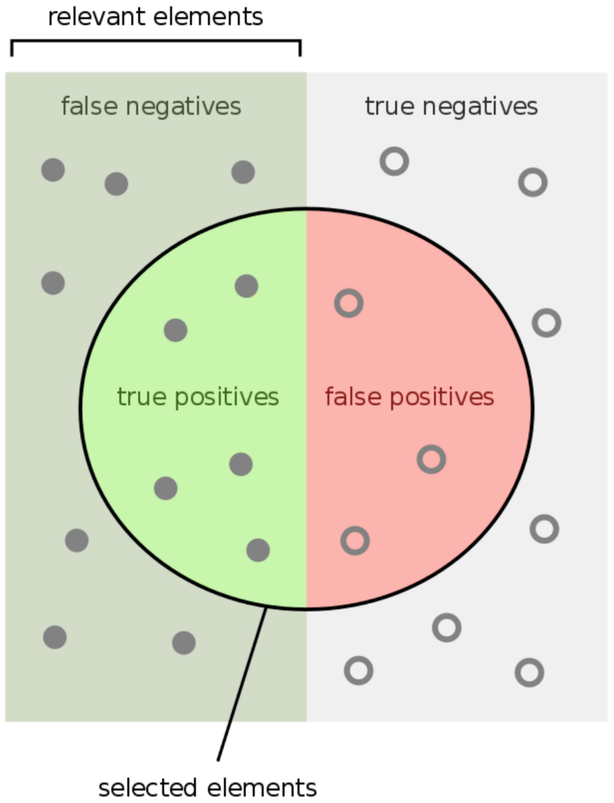


Evaluation



Receiver operating characteristic curve

Evaluation



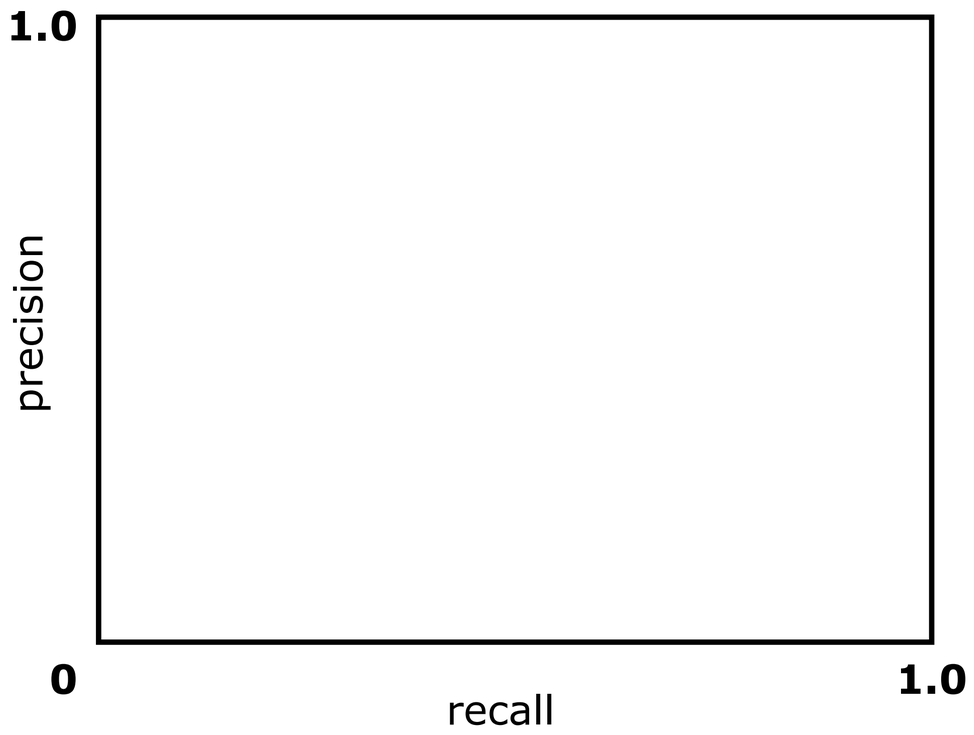
How many selected items are relevant?

Precision = $\frac{\text{true positives}}{\text{selected elements}}$

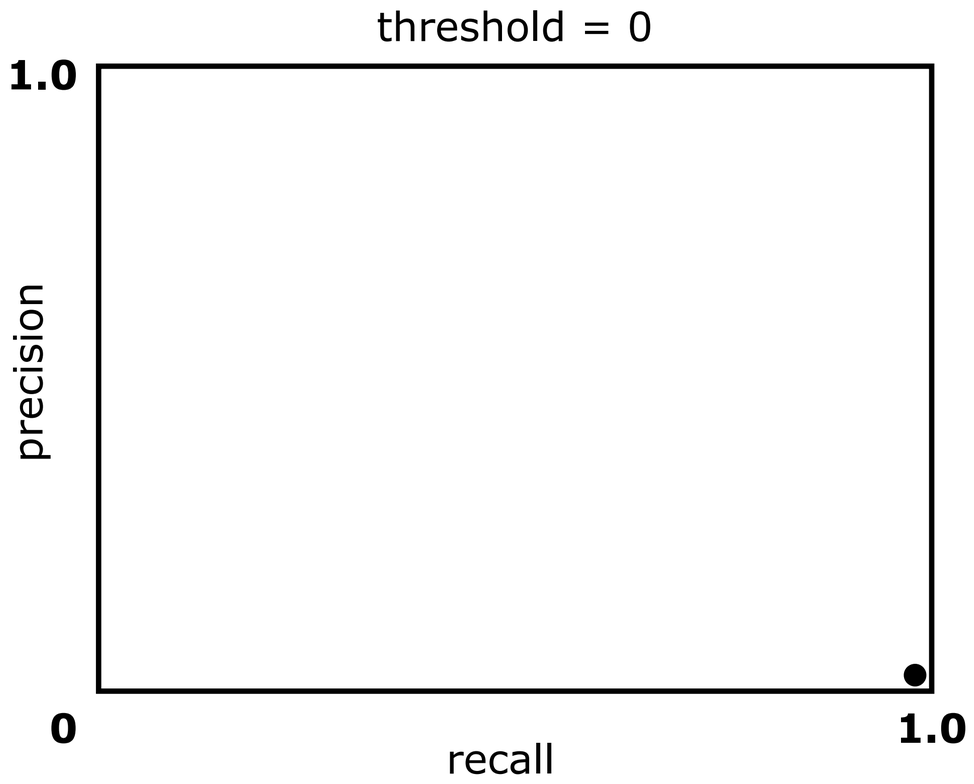
How many relevant items are selected?

Recall = $\frac{\text{true positives}}{\text{relevant elements}}$

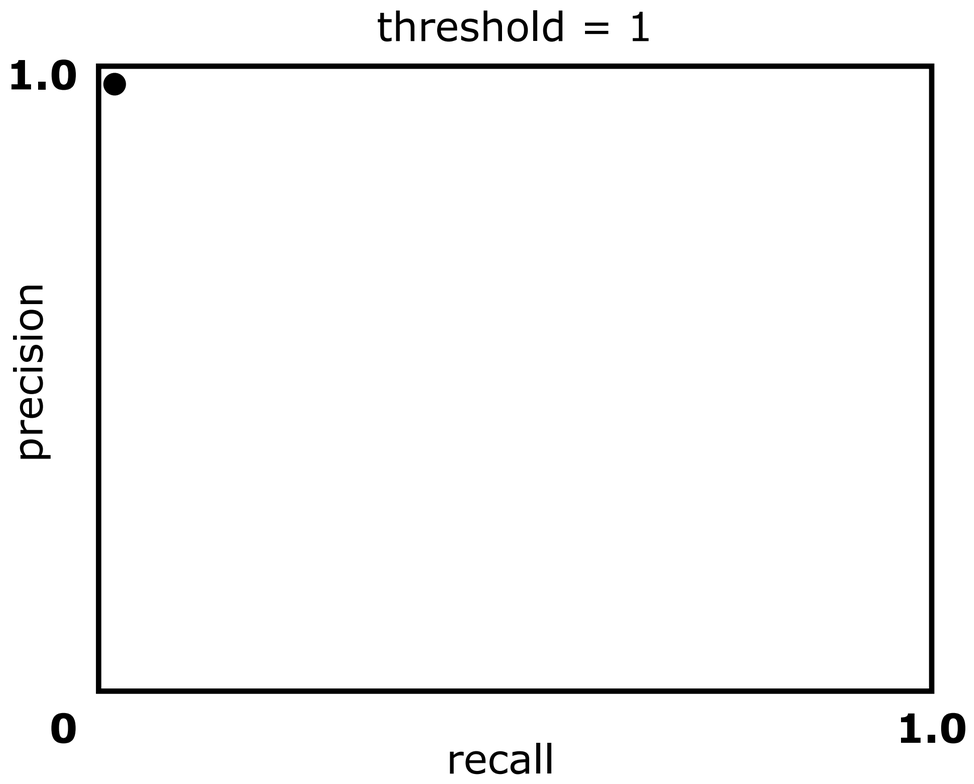
Evaluation



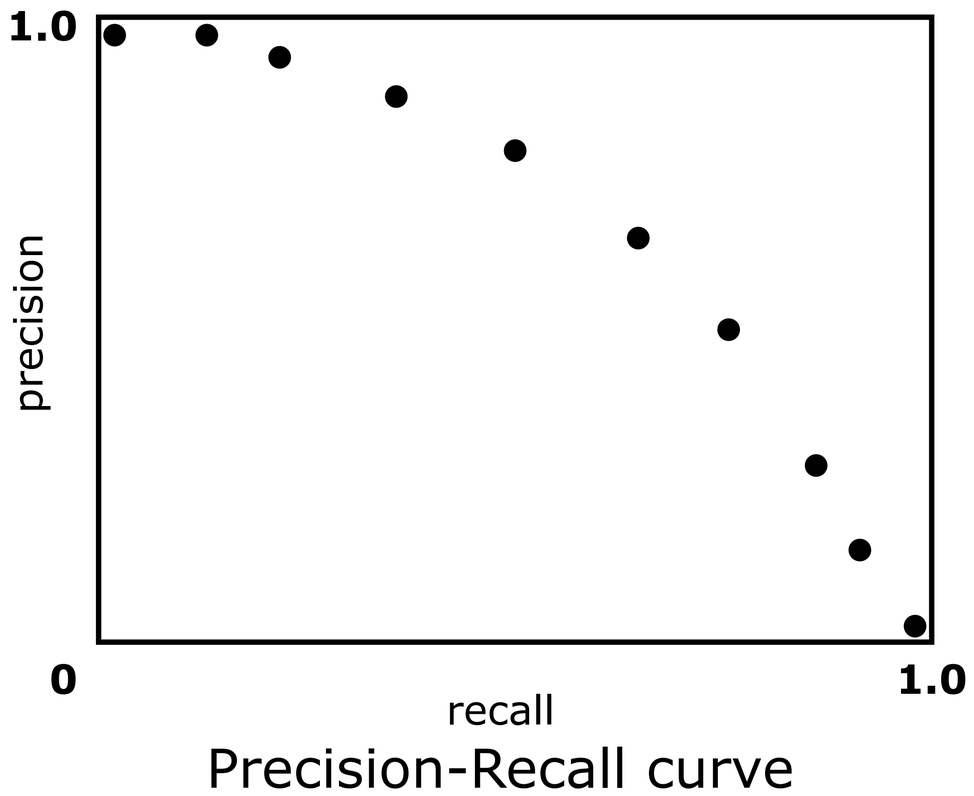
Evaluation



Evaluation



Evaluation



Precision-Recall curve

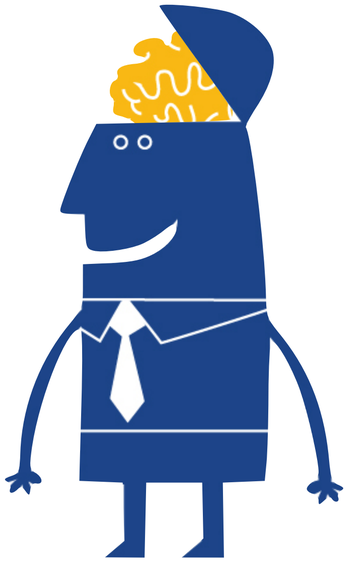
Benchmark

Methods	MTAT		MSD		MTG-Jamendo	
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
FCN [1]	0.9020	0.4367	0.8744	0.2970	0.8255	0.2801
Musicnn [2]	0.9107	0.4511	0.8803	0.2983	0.8226	0.2713
Sample-level [3]	0.9065	0.4428	0.8789	0.2959	0.8208	0.2742
Sample-level + SE [4]	0.9097	0.4540	0.8838	0.3109	0.8233	0.2784
CRNN [6]	0.8729	0.3599	0.8499	0.2469	0.7978	0.2358
Self-attention [7]	0.9096	0.4485	0.8810	0.3103	0.8261	0.2883
BoC-CNN + Res	0.9129	0.4614	0.8898	0.3280	0.8316	0.2951

Won, Minz, Andres Ferraro, Dmitry Bogdanov, Xavier Serra. “A benchmark for music tagging." (not published yet).

Contents

Transfer Learning



Rock

Guitar

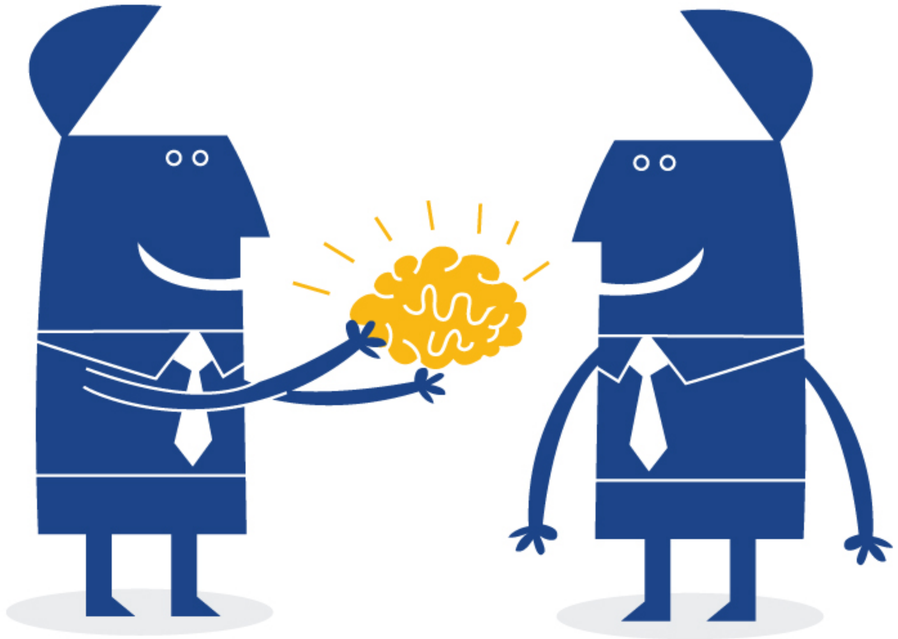
Male vocal

Female vocal

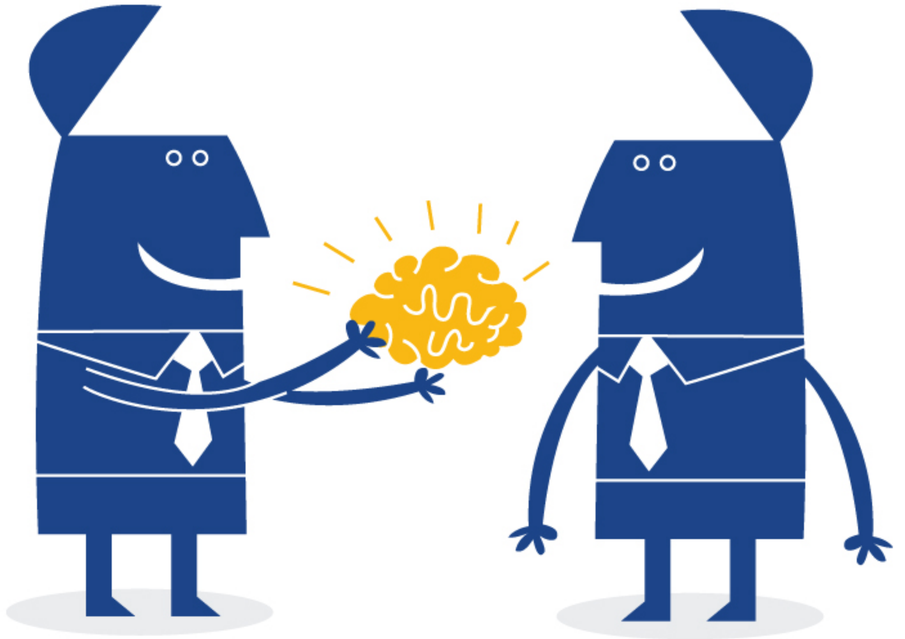
80s

90s

Transfer Learning

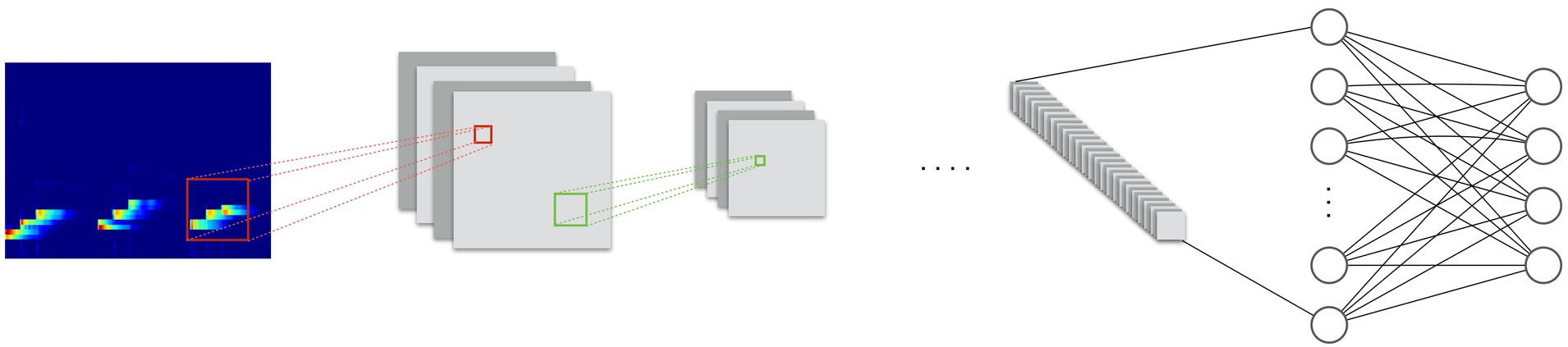


Transfer Learning

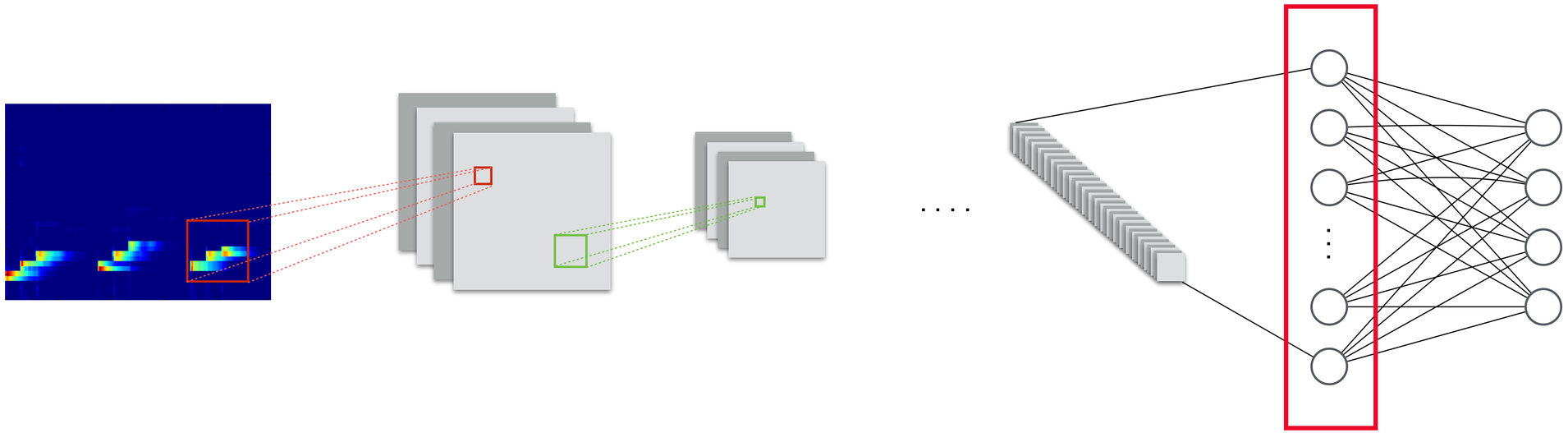


Genre classification ⟶ Music similarity

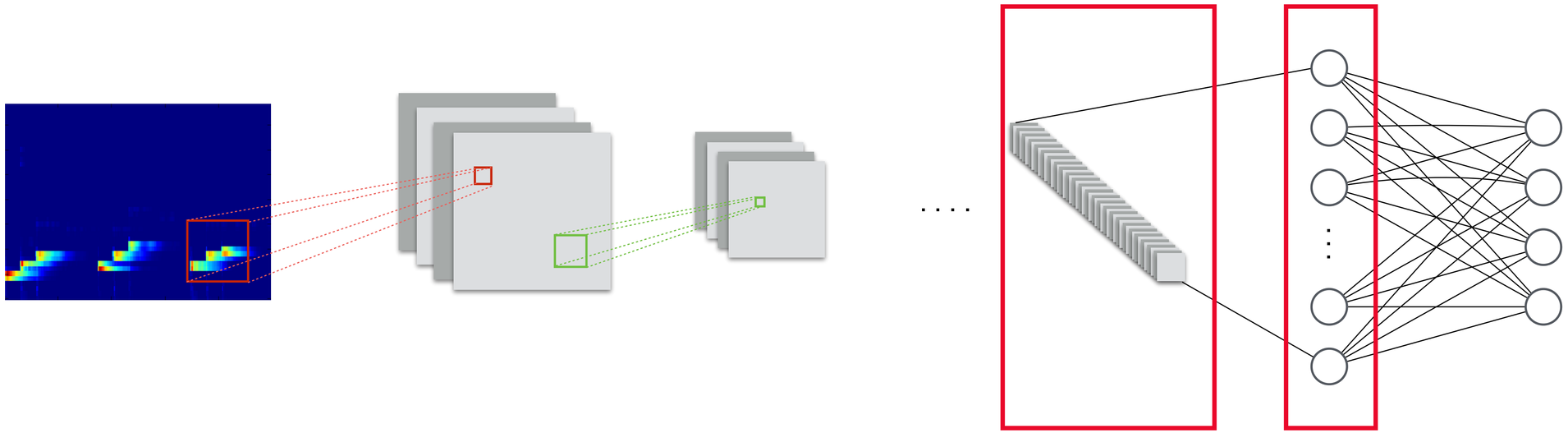
Transfer Learning



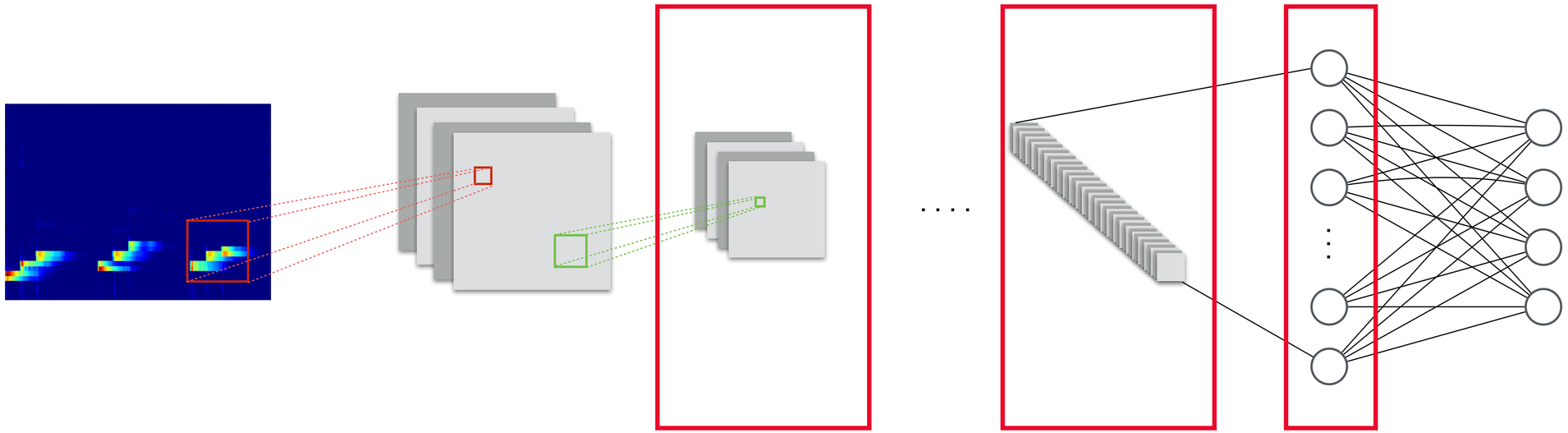
Transfer Learning



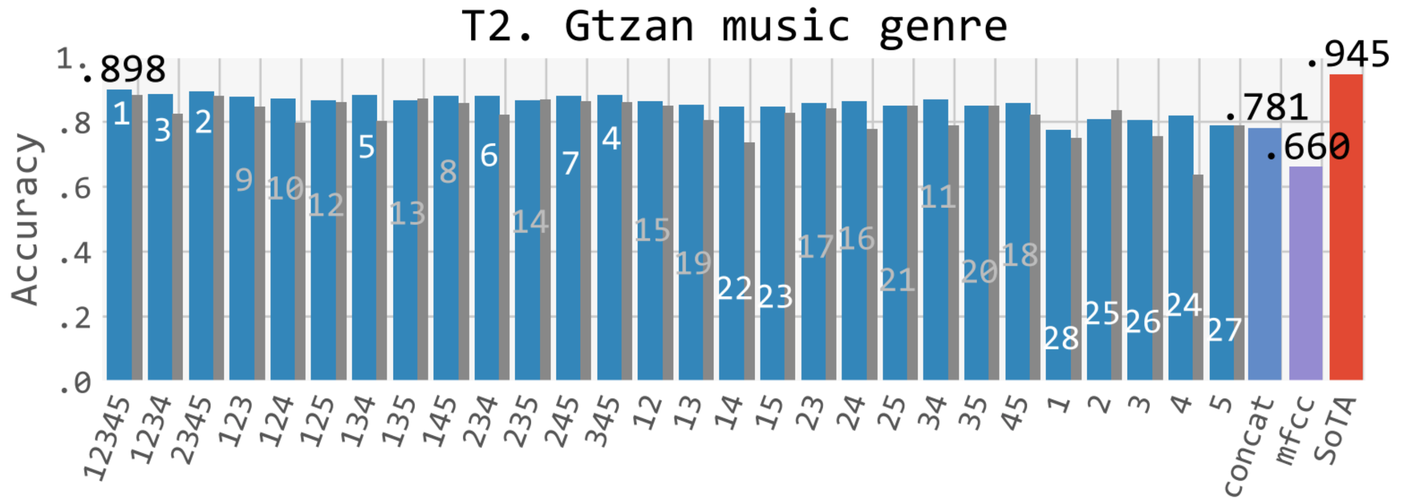
Transfer Learning



Transfer Learning



Transfer Learning



Choi, Keunwoo, et al. "Transfer learning for music classification and regression tasks." (2017).

Contents

Limitations

(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

(b) Content image

71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

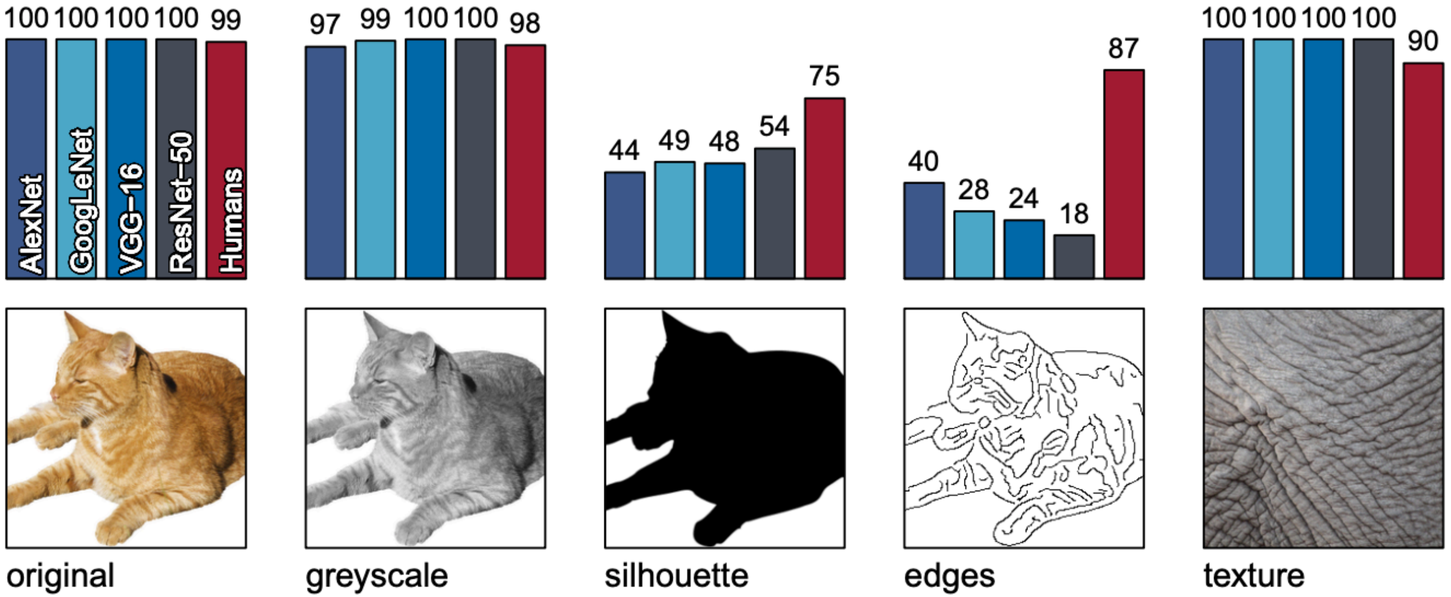
(c) Texture-shape cue conflict

63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

Inherent texture bias in CNN

Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." (2018).

Limitations



Inherent texture bias in CNN

Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." (2018).

Limitations

- Data augmentation can be useful
- e.g., Time stretch, Pitch shift, Dynamic range compression, Add noise
- But still it cannot overcome inherent texture bias. For example, tasks like cover song identification rely more on chroma features.

Salamon, Justin, and Juan Pablo Bello. "Deep convolutional neural networks and data augmentation for environmental sound classification."(2017)

Limitations

common images

turtle in grass
baseline: turtle
ours: turtle

uncommon images

crab in grass
baseline: turtle
ours: crab

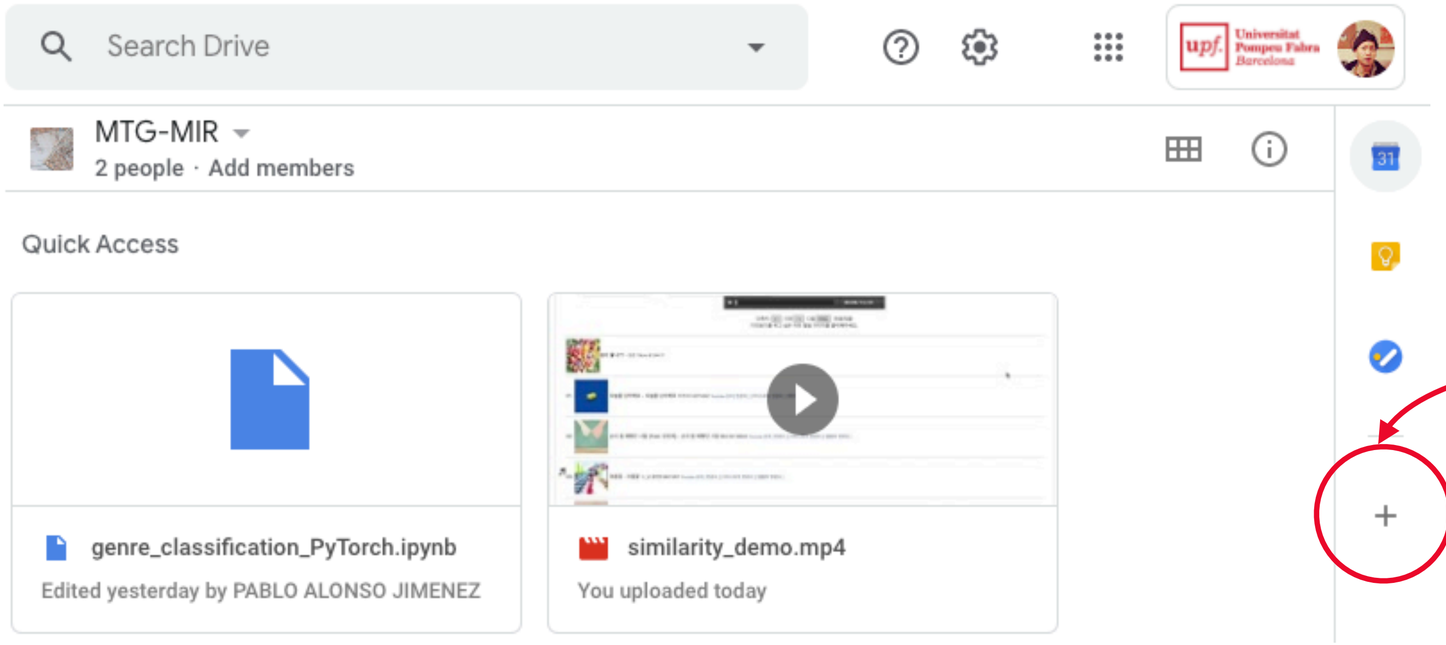
Bias in dataset

Bahng, Hyojin, et al. "Learning De-biased Representations with Biased Representations." (2019).

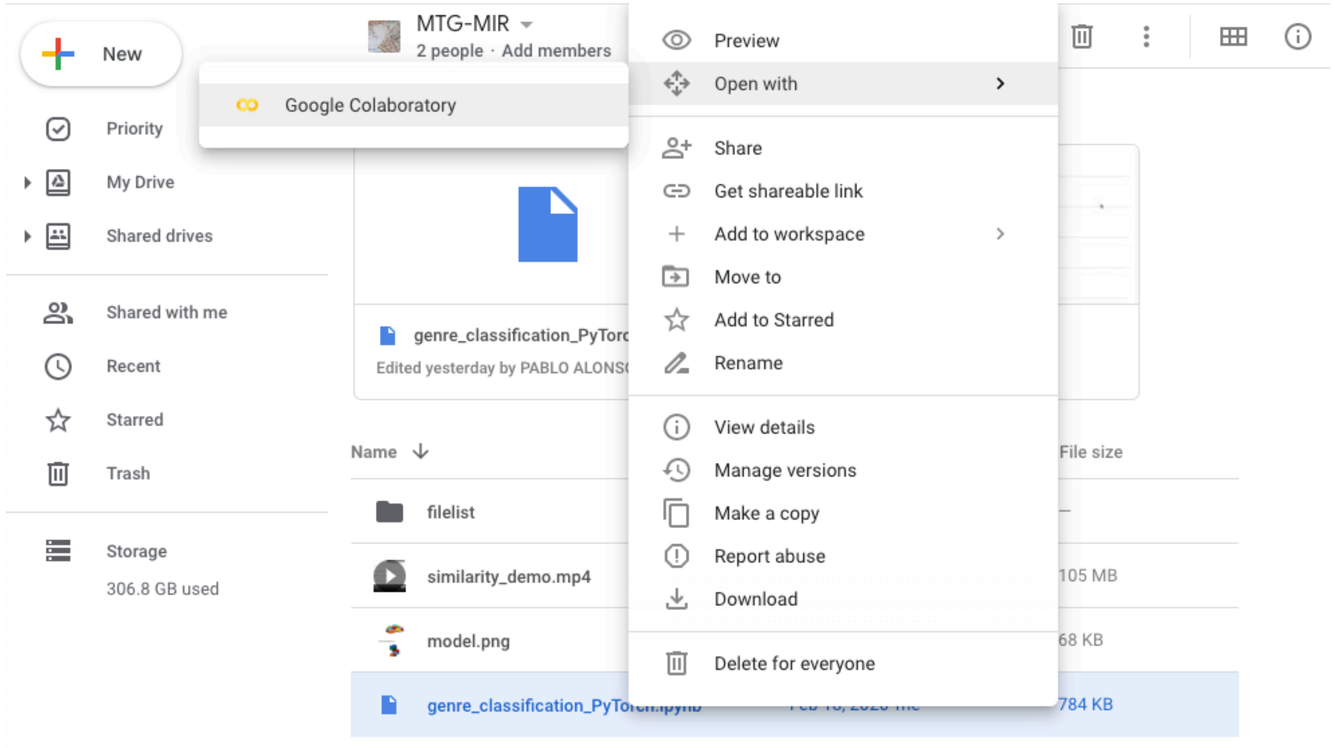
Contents

- Introduction
- Music tagging models
- Transfer learning
- Limitations
- **Lab**

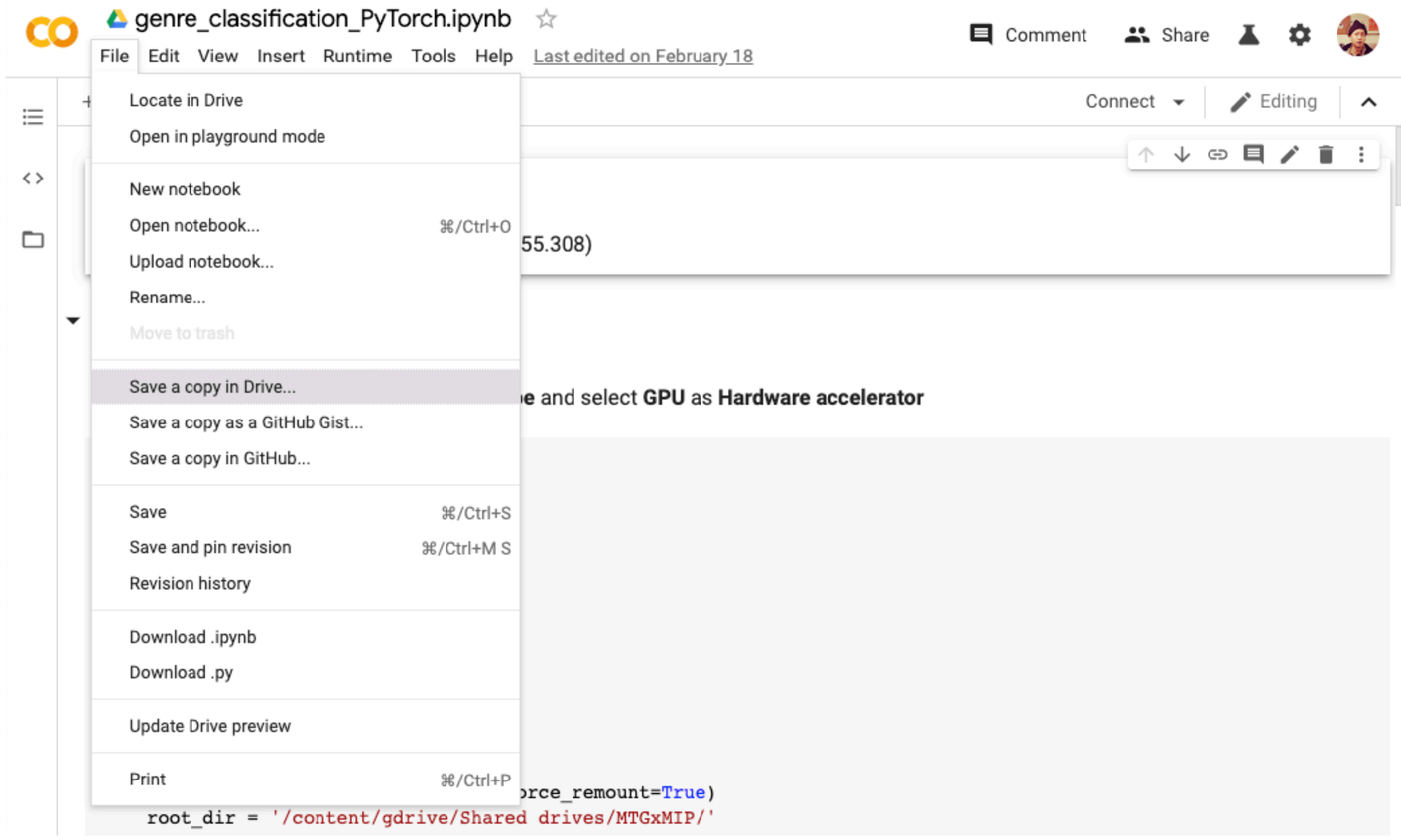
Lab


Search Drive
MTG-MIR
2 people · Add members
Quick Access
genre_classification_PyTorch.ipynb
Edited yesterday by PABLO ALONSO JIMENEZ
similarity_demo.mp4
You uploaded today
Click!
Add Colaboratory

Lab


New
MTG-MIR
2 people · Add members
Google Colaboratory
Priority
My Drive
Shared drives
Shared with me
Recent
Starred
Trash
Storage
306.8 GB used
Edited yesterday by PABLO ALONS
Name
File size
filelist
similarity_demo.mp4
105 MB
model.png
68 KB
784 KB
Preview
Open with
Share
Get shareable link
Add to workspace
Move to
Add to Starred
Rename
View details
Manage versions
Make a copy
Report abuse
Download
Delete for everyone

Lab

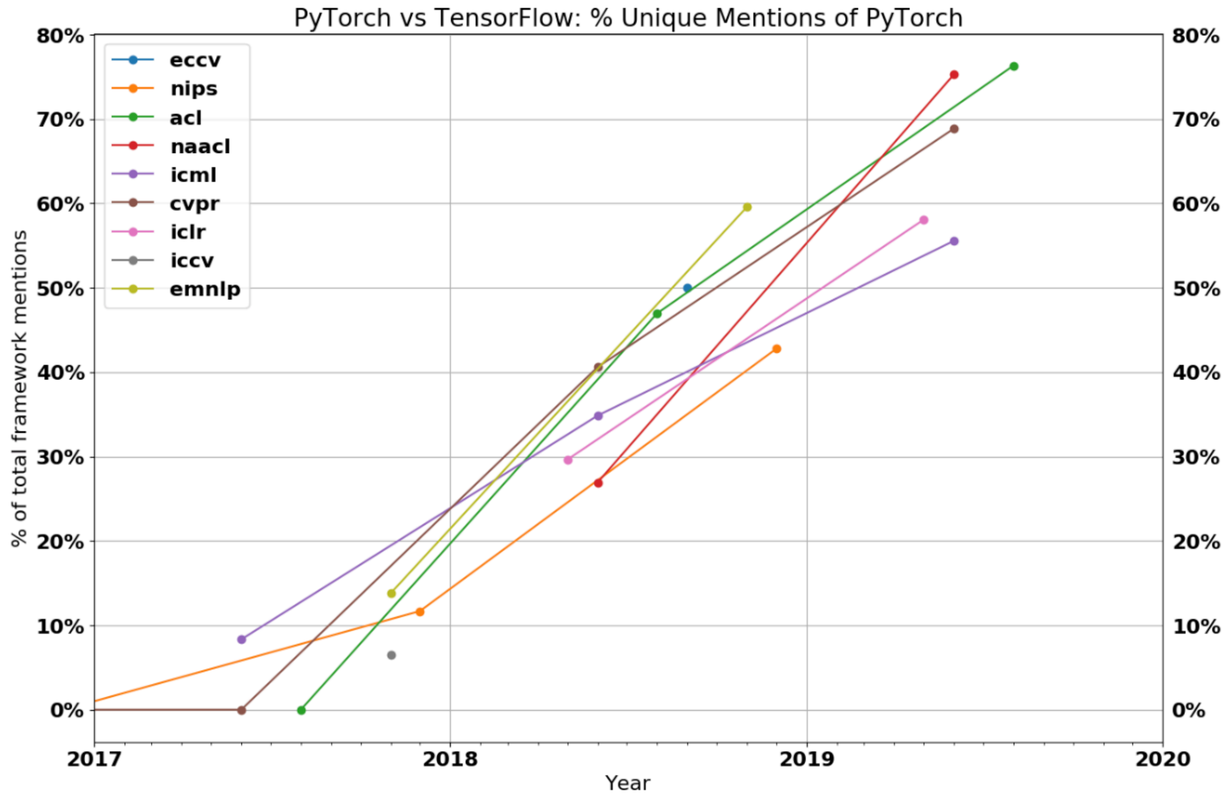

genre_classification_PyTorch.ipynb
File Edit View Insert Runtime Tools Help
Last edited on February 18
Comment
Share
Connect
Editing
Locate in Drive
Open in playground mode
New notebook
Open notebook... ⌘/Ctrl+O
Upload notebook...
Rename...
Move to trash
Save a copy in Drive...
Save a copy as a GitHub Gist...
Save a copy in GitHub...
Save ⌘/Ctrl+S
Save and pin revision ⌘/Ctrl+M S
Revision history
Download .ipynb
Download .py
Update Drive preview
Print ⌘/Ctrl+P
55.308)
e and select GPU as Hardware accelerator
orce_remount=True)
root_dir = '/content/gdrive/Shared drives/MTGxMIP/'

PyTorch

- Easy to install
- Easy to prototype
- Easy to debug
- More pythonic
- Dynamic graph

PyTorch

- Easy to install
- Easy to prototype
- Easy to debug
- More pythonic
- Dynamic graph



PyTorch

```
class MyModel(nn.Module):

    def __init__(self):
```

```
    def forward(self, x):
```