Audio Analytics

Serge Retkowsky
EMEA AI GBB Team
serge.retkowsky@microsoft.com

V1 – 22-03-2022

What is a sound?

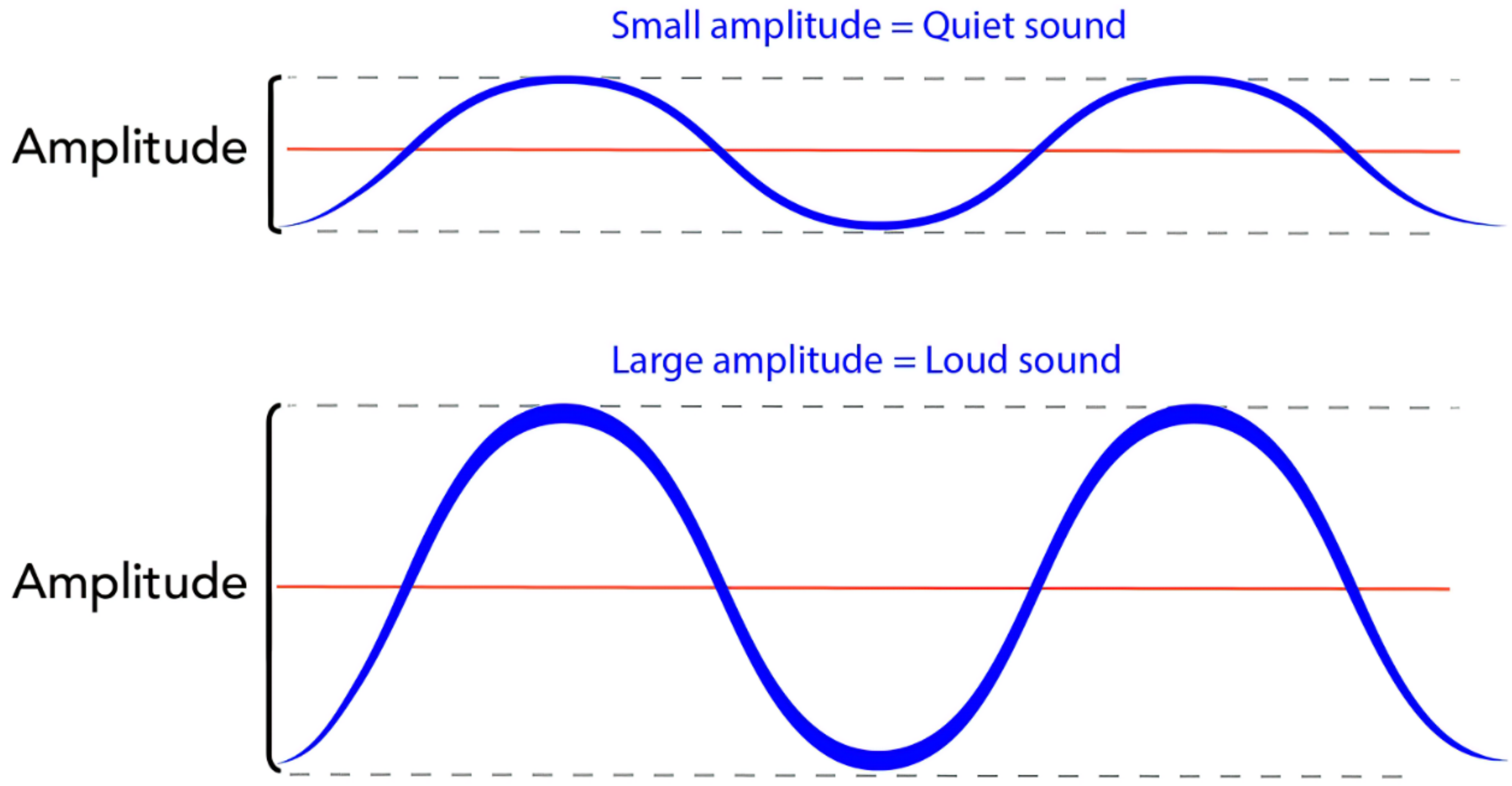
sound

[saʊnd]

NOUN

1. vibrations that travel through the air or another medium and can be heard when they reach a person's or animal's ear.
   "light travels faster than sound"
2. sound produced by continuous and regular vibrations, as opposed to noise.
3. music, speech, and sound effects when recorded and used to accompany a film, video, or broadcast.
   "a sound studio"

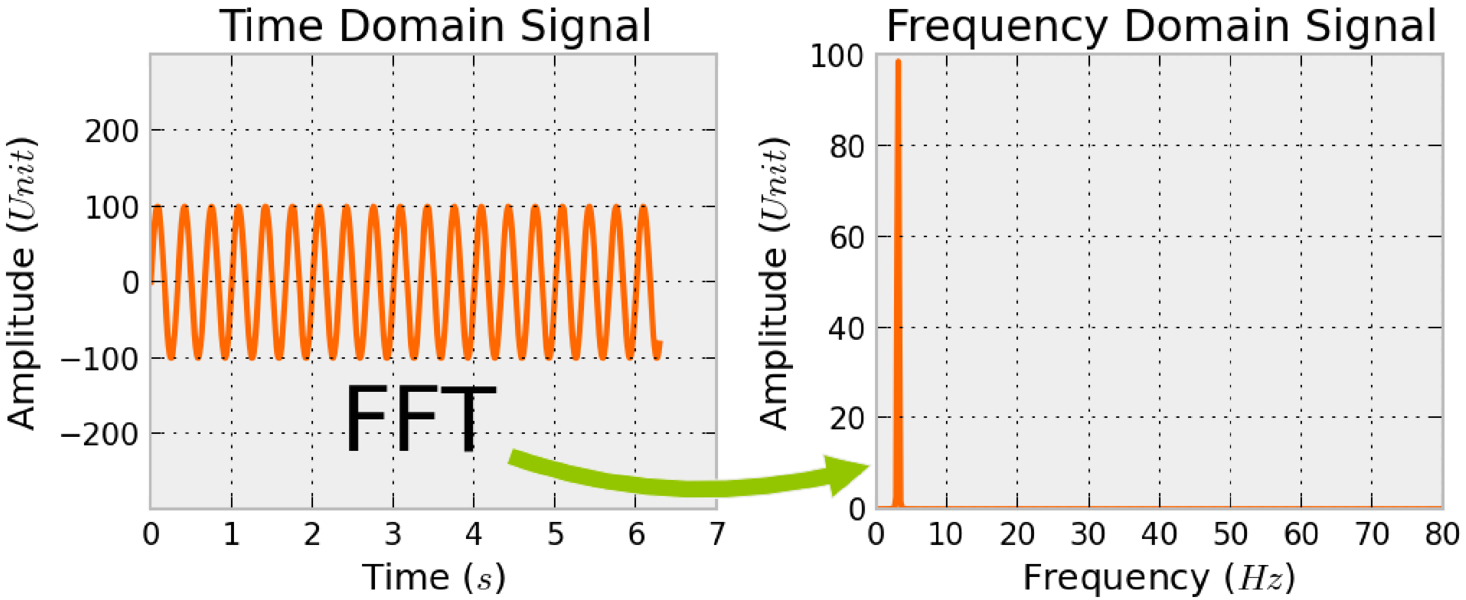
A signal is the sound variation in a certain quantity over time



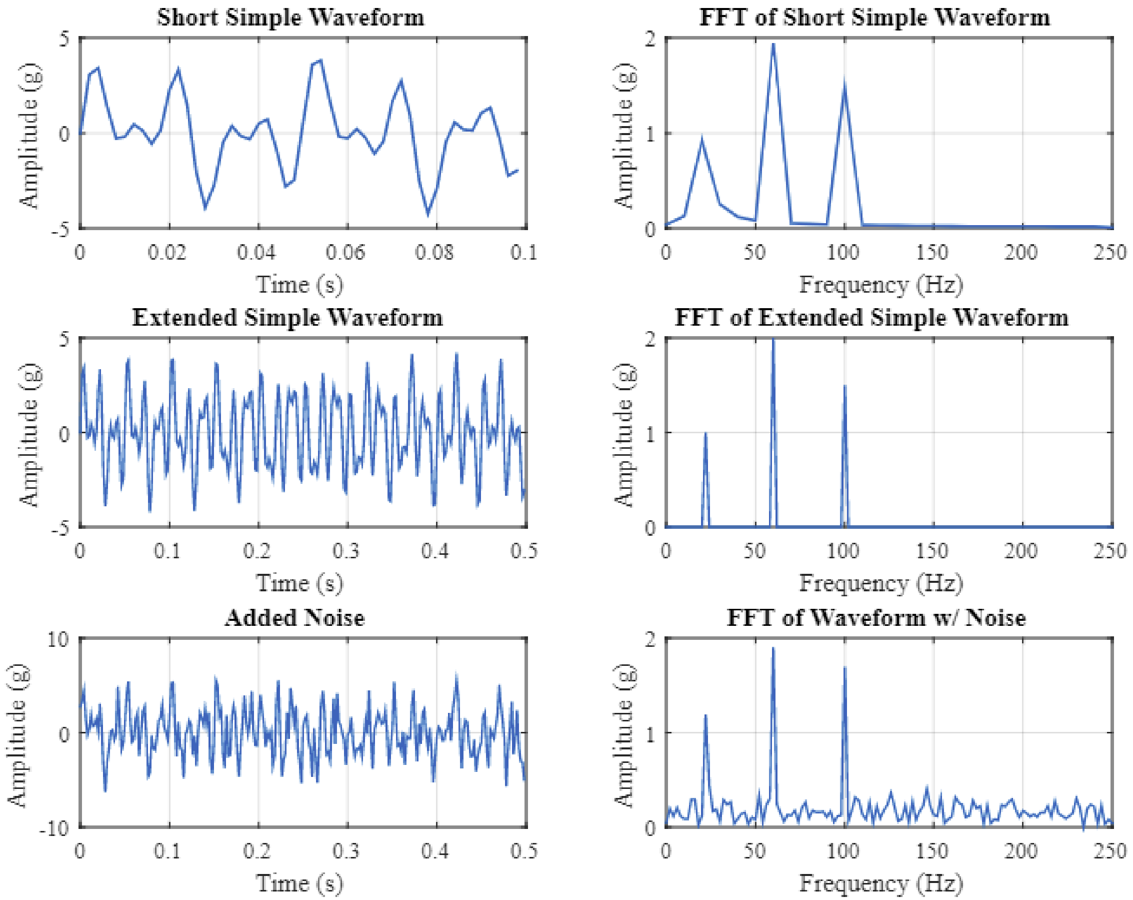
Fast Fourier Transform (FFT)

Jean-Baptiste Joseph Fourier
1768 – 1830
https://en.wikipedia.org/wiki/Joseph_Fourier

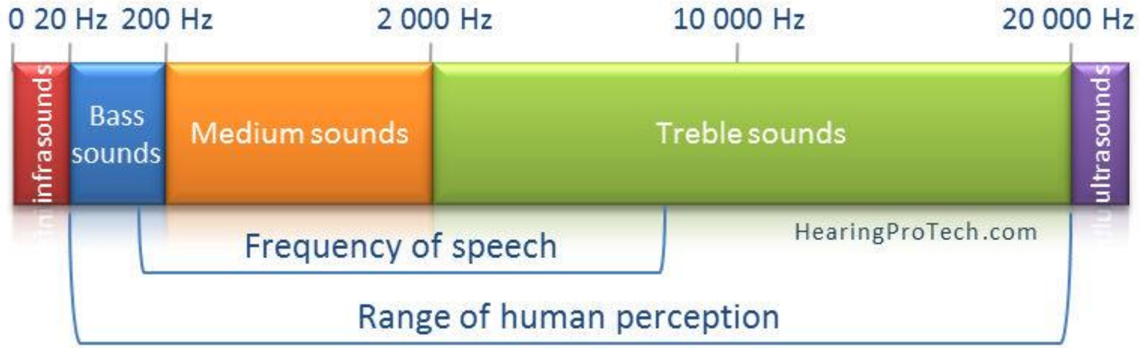
The Fourier transform is a mathematical formula that **converts the signal from the time domain into the frequency domain.**



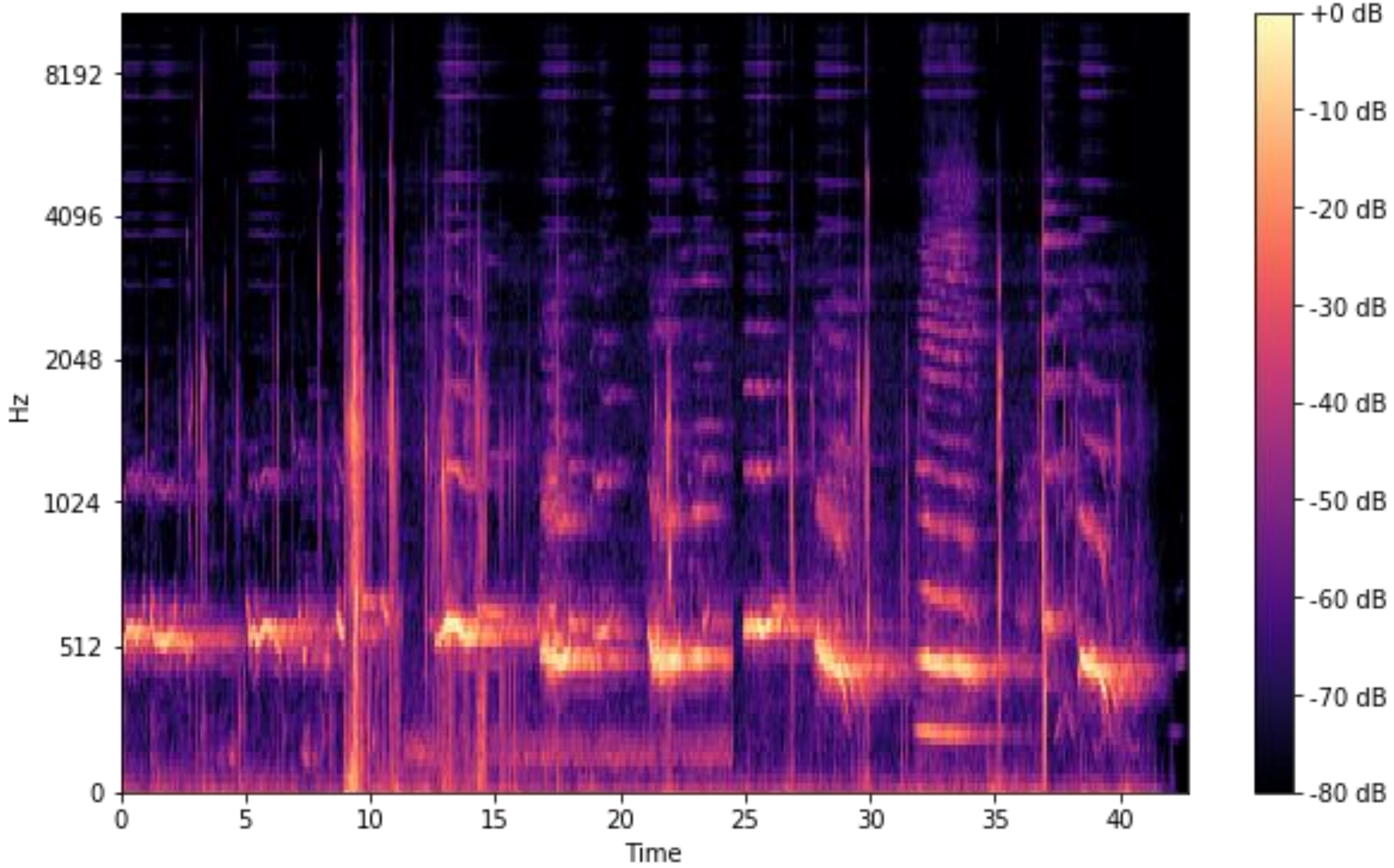
Spectrum



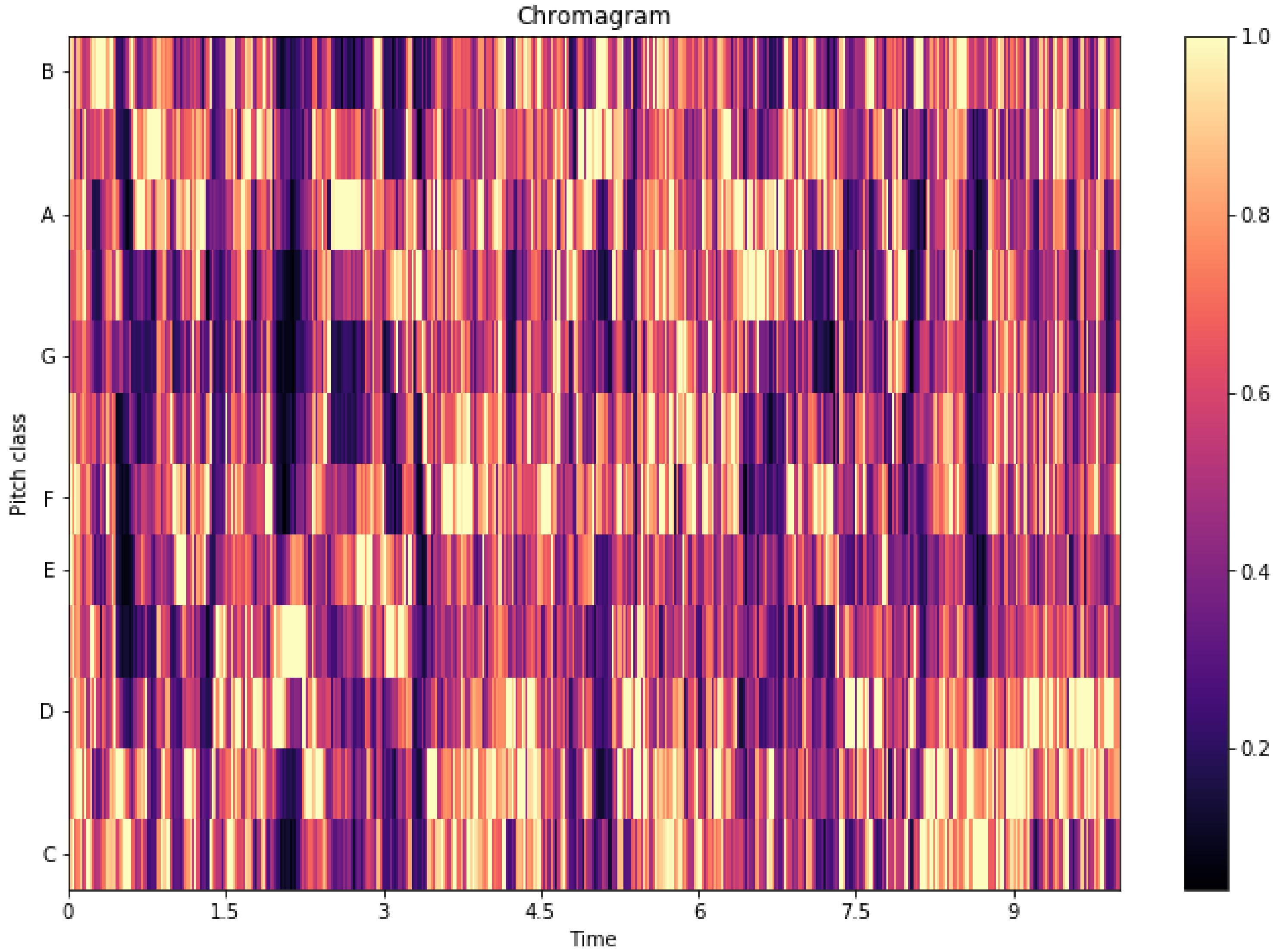
The **spectrum** is the distribution of amplitudes for each frequency component of the signal.



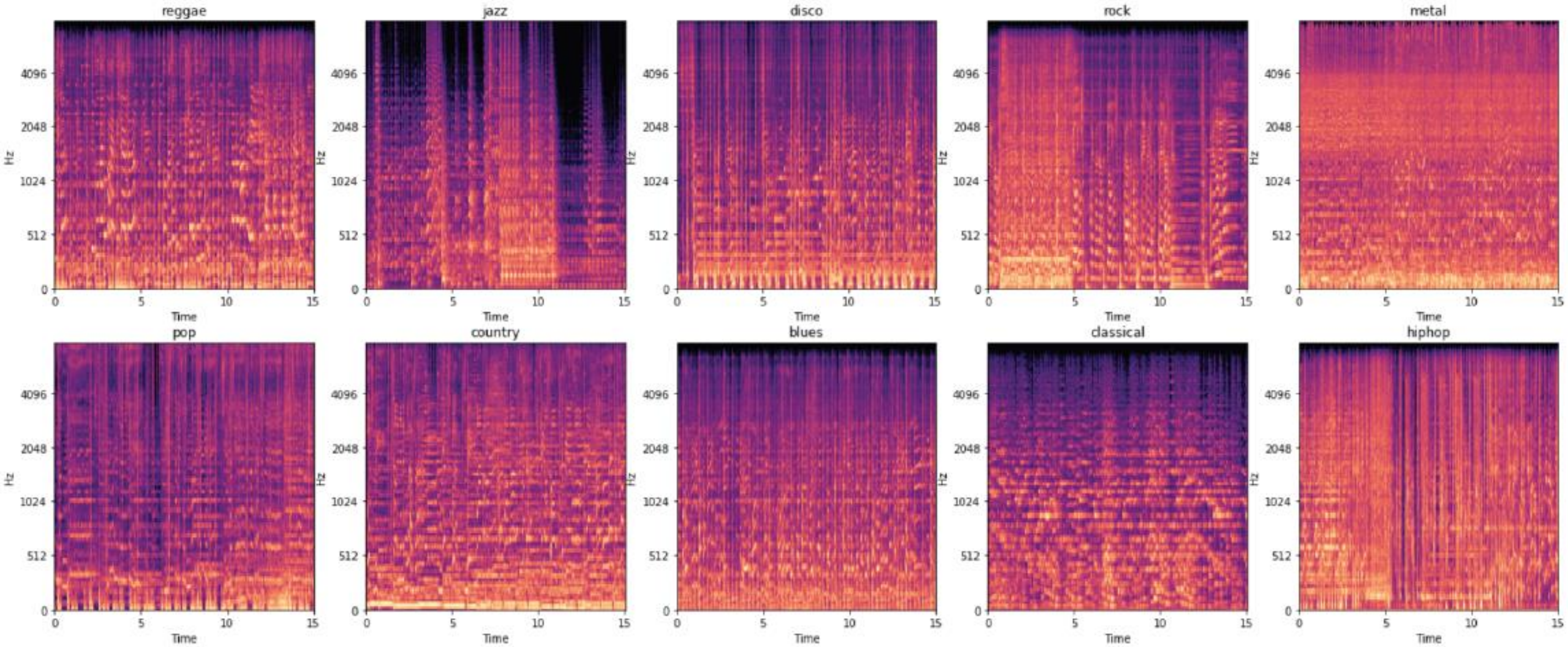
The spectrogram represents how the spectrum of frequencies vary over time



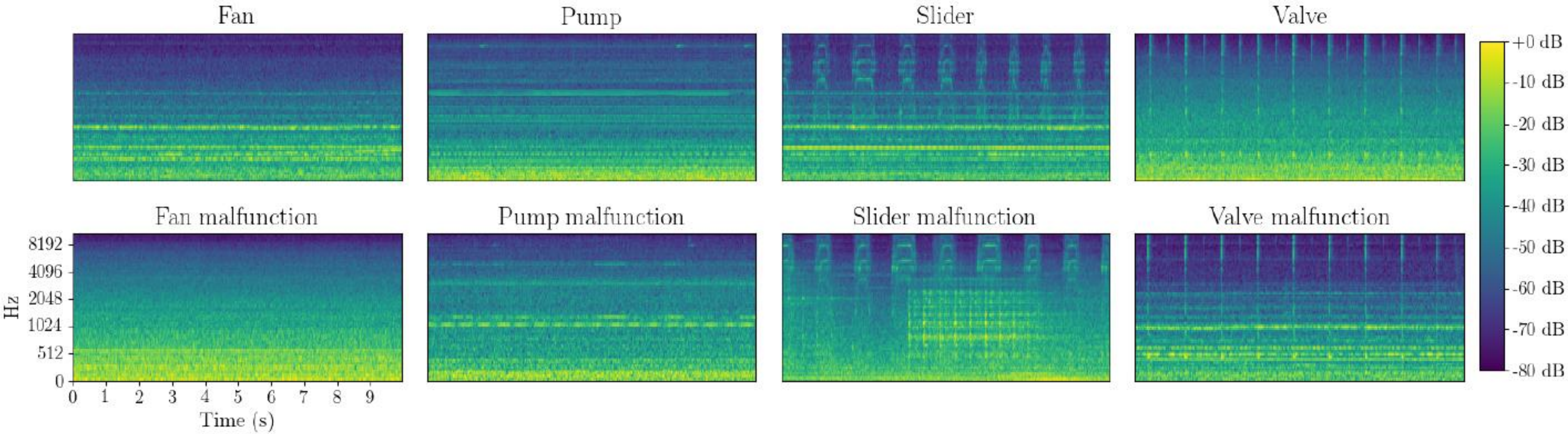
Chromagram display the intensity of each pitch for each time interval



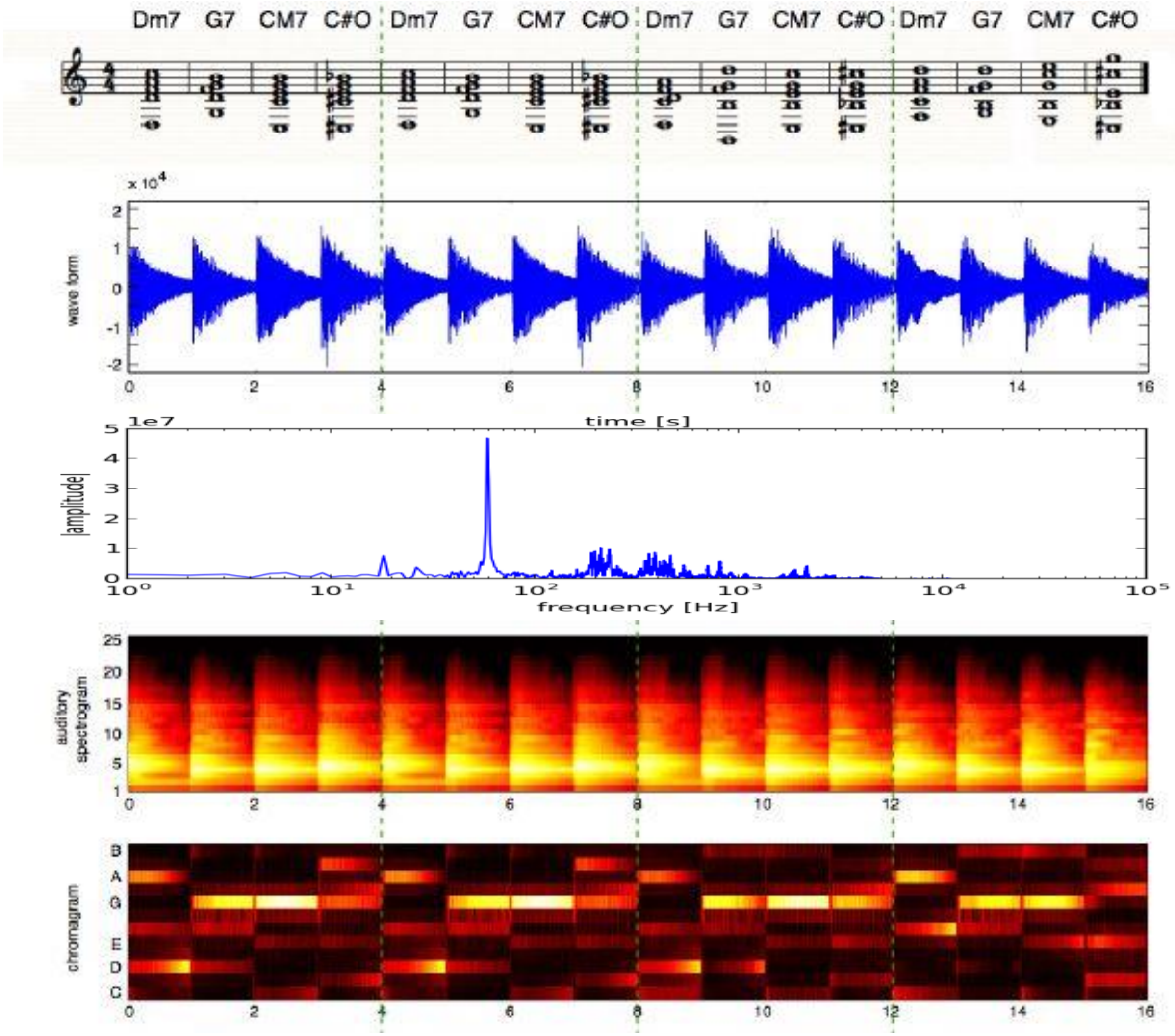
Using spectrograms images as features, we can train a computer vision model to classify audio files



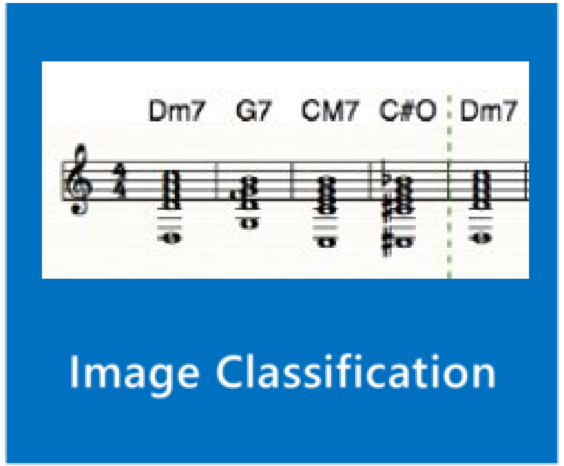
Or to predict an acoustic anomaly from the sound of a machine



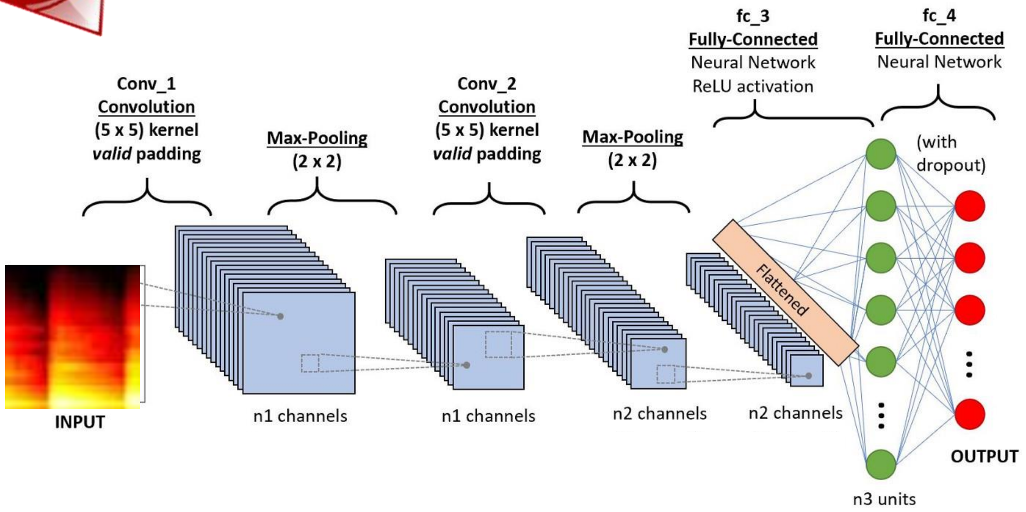
In summary



Dm7 G7 CM7 C#O Dm7 G7 CM7 C#O Dm7 G7 CM7 C#O Dm7 G7 CM7 C#O
x 10^4
wave form
time [s]
1e7
|amplitude|
frequency [Hz]
auditory spectrogram
chromagram



Dm7 G7 CM7 C#O Dm7
Image Classification



Conv_1
Convolution
(5 x 5) kernel
valid padding
Max-Pooling
(2 x 2)
Conv_2
Convolution
(5 x 5) kernel
valid padding
Max-Pooling
(2 x 2)
fc_3
Fully-Connected
Neural Network
ReLU activation
fc_4
Fully-Connected
Neural Network
(with dropout)
INPUT
n1 channels
n1 channels
n2 channels
n2 channels
Flattened
n3 units
OUTPUT

Azure AutoML for Images Algorithm

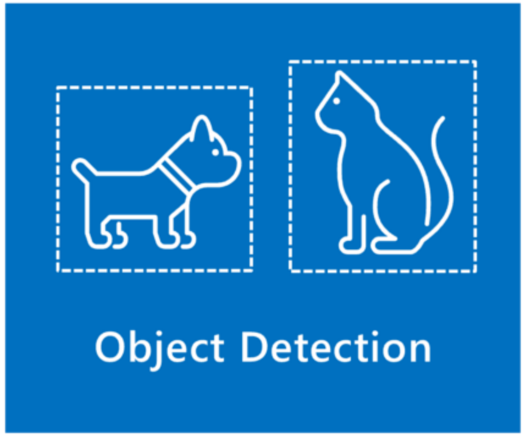
MobileNet: Light models for mobile applications

ResNet: Residual networks

ResNeSt: Split attention networks

SE-ResNeXt50: Squeeze-and-Excitation networks

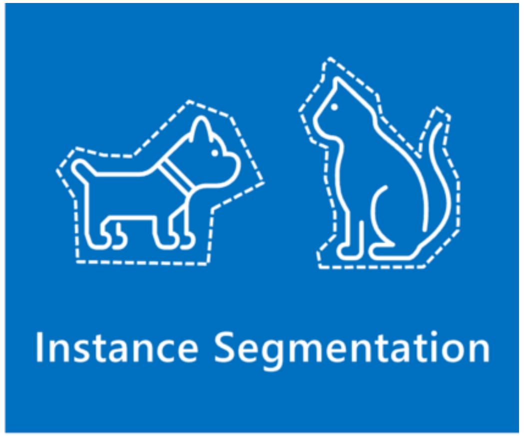
ViT: Vision transformer networks



YOLOv5: One stage object detection

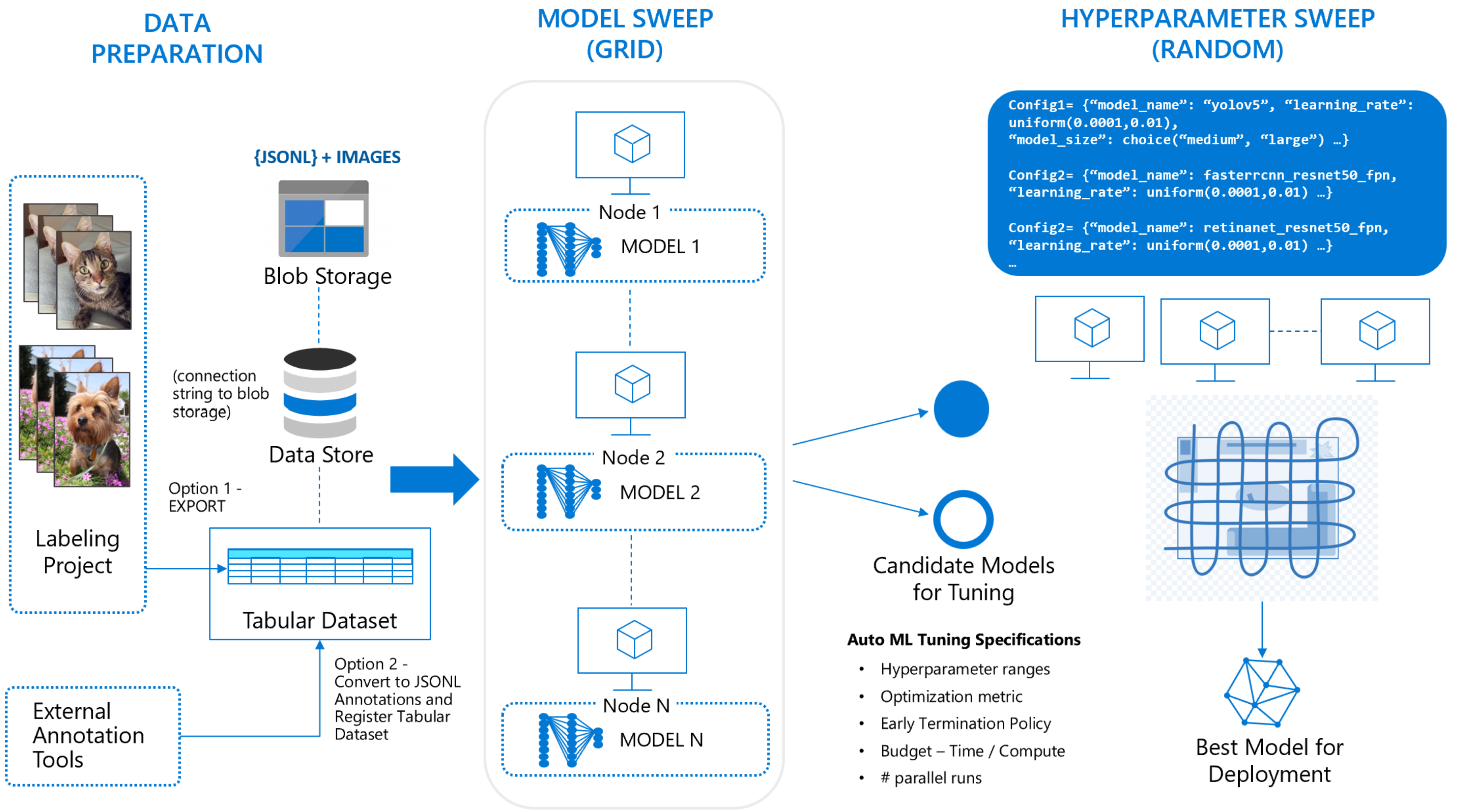
Faster RCNN ResNet FPN: Two stage object detection

RetinaNet ResNet FPN: address class imbalance with focal loss



MaskRCNN ResNet FPN

Azure AutoML for Images



We can generate audio features from audio files and use a generic classification algorithm

Spectral features

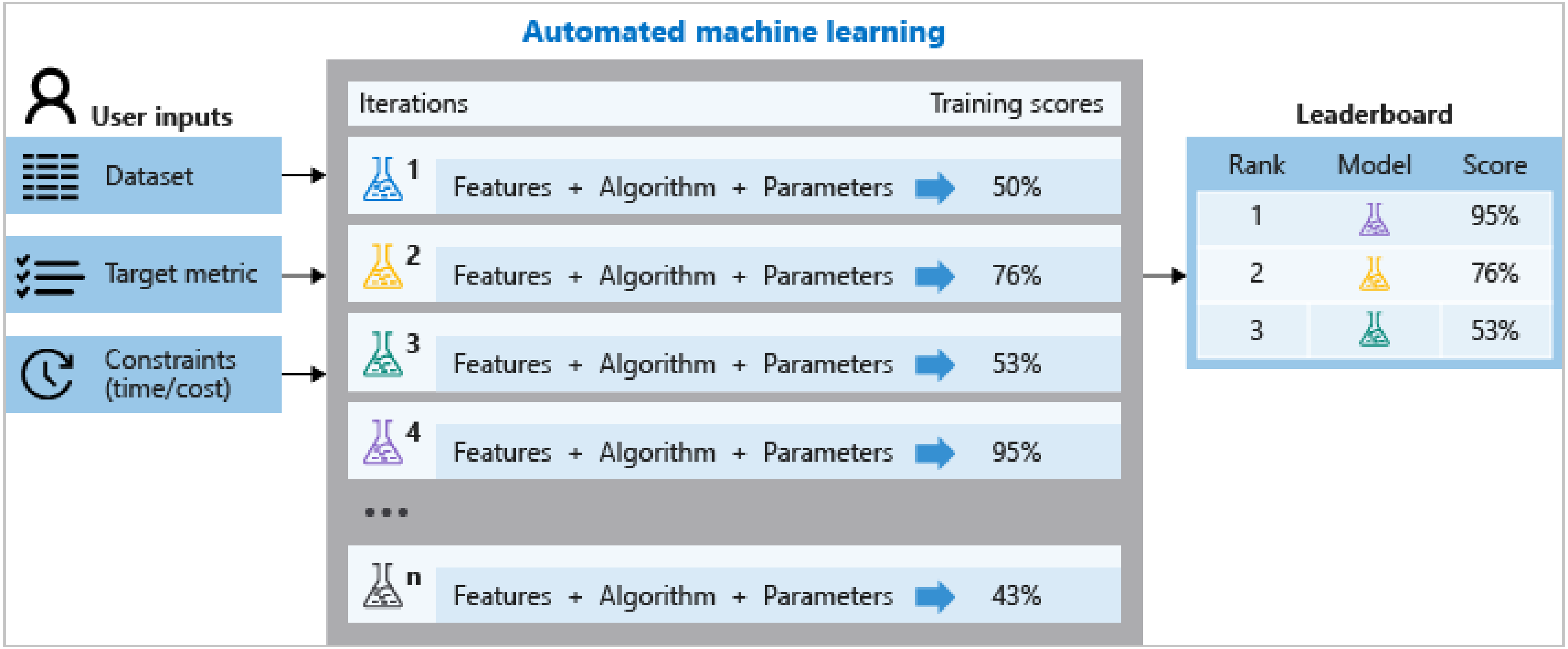
- Mel-Frequency Cepstral Coefficients
  Spectral Centroid
- Zero Crossing Rate
- Chroma Frequencies
- Spectral Roll-off
- Spectral contrast
- Spectral flatness
- Nth order polynomial of spectrogram
- Tonal centroid features

Rhythm features

- Tempogram
- Fourier tempogram

	E	F	G	H	I	J	K	L	
1	filename	sampling_rate	total_samples	duration	chroma_mean_0	chroma_mean_1	chroma_mean_2	chroma_mean_3	chrom
2	blues.00000.wav	22050	661794	30,0133333	0,26800678	0,251611333	0,226004798	0,095506201	0,2
3	blues.00001.wav	22050	661794	30,0133333	0,245332417	0,119800933	0,153988396	0,053337448	0,1
4	blues.00002.wav	22050	661794	30,0133333	0,044974575	0,053335268	0,213918681	0,299785802	0,4
5	blues.00003.wav	22050	661794	30,0133333	0,149017047	0,095013408	0,256969357	0,361600207	0,5
6	blues.00004.wav	22050	661794	30,0133333	0,149890471	0,324409668	0,231447684	0,049297483	0,3
7	blues.00005.wav	22050	661794	30,0133333	0,174924776	0,297632319	0,490159317	0,029940231	0,0
8	blues.00006.wav	22050	661794	30,0133333	0,303592247	0,044819245	0,033725801	0,071567812	0,2
9	blues.00007.wav	22050	661794	30,0133333	0,133722862	0,160319298	0,161420863	0,115934335	0,1
10	blues.00008.wav	22050	661794	30,0133333	0,369537793	0,167340323	0,241840856	0,198154131	0,4
11	blues.00009.wav	22050	661794	30,0133333	0,214746992	0,246957063	0,431365761	0,0903425	0,1
12	blues.00010.wav	22050	661794	30,0133333	0,269612392	0,065305786	0,187212141	0,257468495	0,3
13	blues.00011.wav	22050	661794	30,0133333	0,056971933	0,033895528	0,112376054	0,333991449	0,6
14	blues.00012.wav	22050	661794	30,0133333	0,18643648	0,164015898	0,249181222	0,168915739	0,1
15	blues.00013.wav	22050	661794	30,0133333	0,175153815	0,174307514	0,185878742	0,143578468	0,0
16	blues.00014.wav	22050	661794	30,0133333	0,308505654	0,298338033	0,372414037	0,185313681	0,2
17	blues.00015.wav	22050	661794	30,0133333	0,272835729	0,260926366	0,276364656	0,229465334	0,2
18	blues.00016.wav	22050	661794	30,0133333	0,254976744	0,301687379	0,172817155	0,261988886	0,3
19	blues.00017.wav	22050	661794	30,0133333	0,263373784	0,193970588	0,129297141	0,167151084	0
20	blues.00018.wav	22050	661794	30,0133333	0,182916913	0,225288392	0,222885971	0,159066622	0,1
21	blues.00019.wav	22050	661794	30,0133333	0,2424882	0,295897792	0,204931068	0,227288107	0,3
22	blues.00020.wav	22050	661794	30,0133333	0,299063117	0,2324963	0,185303475	0,216888446	0,4
23	blues.00021.wav	22050	661794	30,0133333	0,207339627	0,167177949	0,11422442	0,117166118	0,1
24	blues.00022.wav	22050	661794	30,0133333	0,384442008	0,29263856	0,194021396	0,203634849	0,1
25	blues.00023.wav	22050	661794	30,0133333	0,17348474	0,192353927	0,105343125	0,209032257	0,3
26	blues.00024.wav	22050	661794	30,0133333	0,203956906	0,184026079	0,177608028	0,115847307	0
27	blues.00025.wav	22050	661794	30,0133333	0,203488798	0,154098805	0,236616226	0,093262482	0,1
28	blues.00026.wav	22050	661794	30,0133333	0,220660965	0,138306133	0,138505317	0,084929196	0,1
29	blues.00027.wav	22050	661794	30,0133333	0,336735659	0,142619544	0,19476544	0,105940677	0,3
30	blues.00028.wav	22050	661794	30,0133333	0,193278268	0,12818317	0,168773284	0,089887121	0,1

AutoML for Classification



Audio Processing with Azure ML

Audio processing can consist of extracting audio signal information into **spectrograms** (time vs frequency vs Db) **images** that we can use to build a custom vision model with **Azure using AutoML for Images.**

We can as well extract some audio components and use a generic **classification model with Azure ML and its AutoML features.**

Demo1: Music Genre Prediction

- **Problem:**
  - Is it possible to predict the music genre of an audio file?
- **Solution:**
  1. We will build spectrograms for all the training music files
  2. Then we will use these images to build, train and deploy an Image Computer Vision model with AutoML for Images
  3. We will test the model to predict the genre based on an audio file

Demo2: Acoustic Anomaly Detection for Machine Sounds based on Images

- **Problem:**
  - Is it possible to detect an anomaly (not normal noise) using a machine sound file?

- **Solution:**
  1. We will collect some normal and anomaly sounds files
  2. We will generate spectrograms for all the files
  3. We will build and train a two-class classification model (Anomaly vs no anomaly)
  4. We will test the anomaly detection model

Links

- Azure ML

https://aka.ms/AIShow/AutoML/AzureML

- AutoML for Images

http://aka.ms/AutoMLforImagesDoc

- AutoML for Images Algorithms

http://aka.ms/AutoMLforImagesAlgorithms

- AutoML for Images tutorial

http://aka.ms/AutoMLforImagesTutorial