Humanities Research with Sound: Introduction to Audio Machine Learning

Stephen McLaughlin
& Tanya Clement

The University of Texas at Austin
School of Information

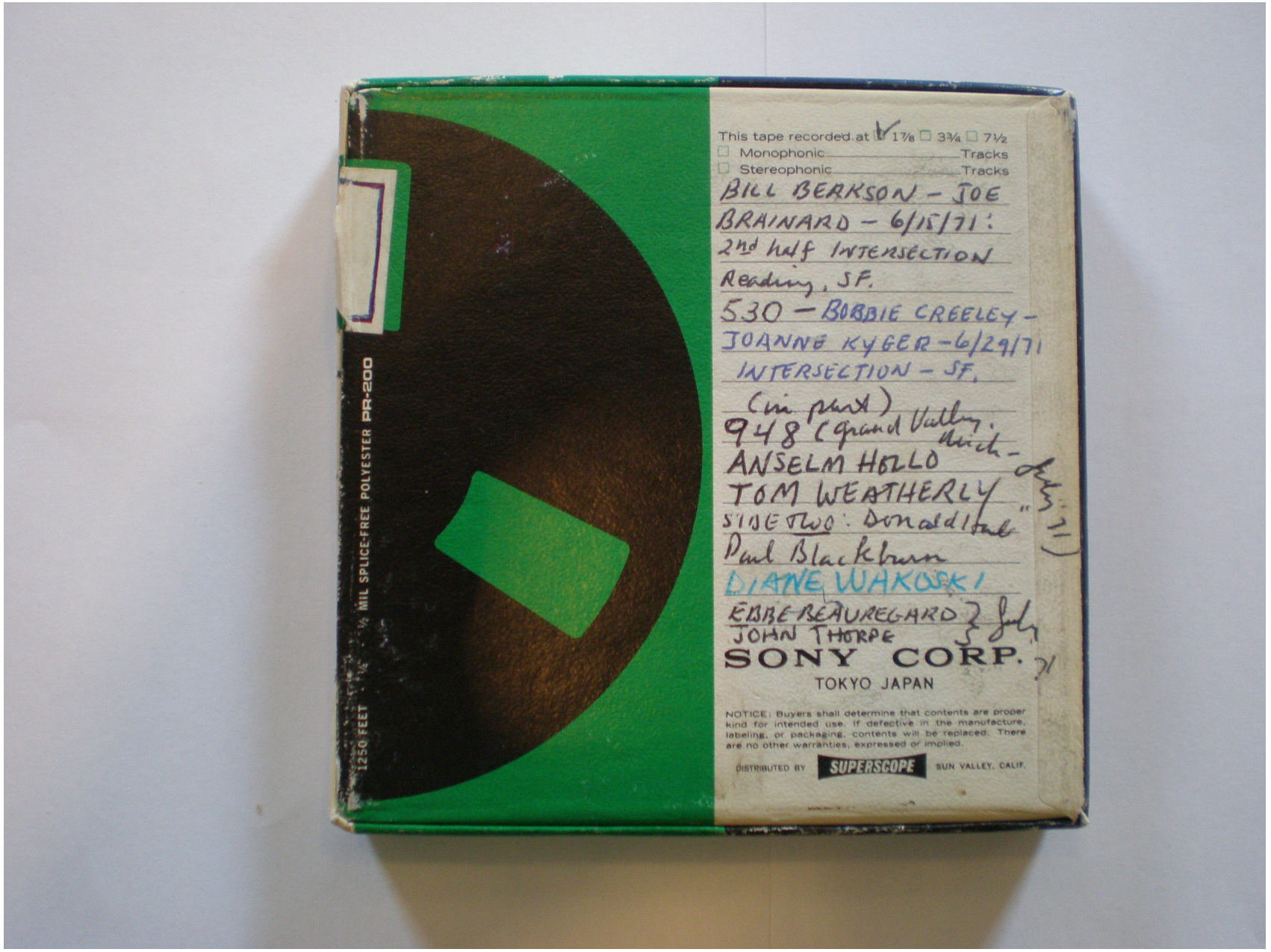

LYM
Alam
MATT MCKEOWN

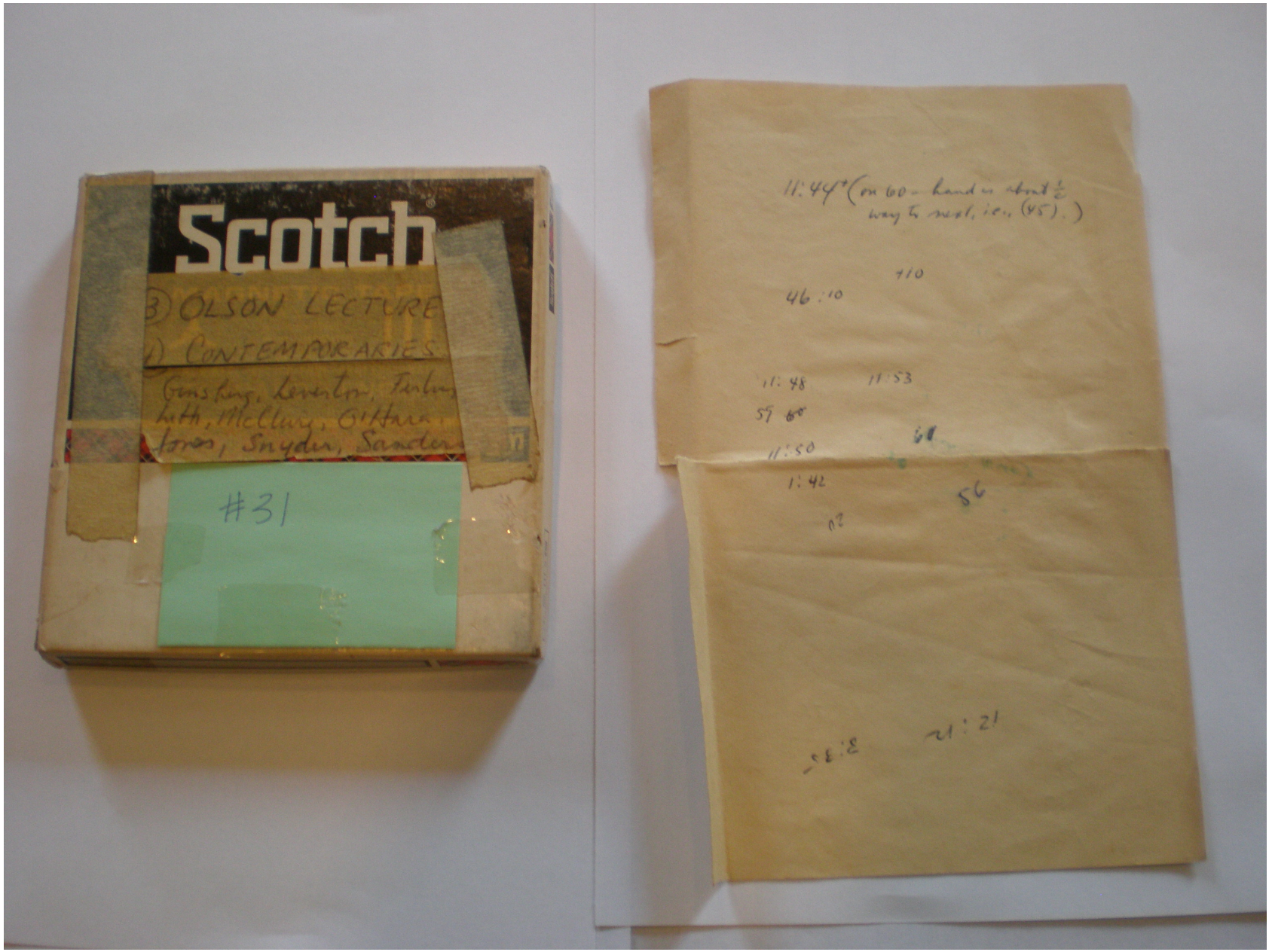
https://github.com/hipstas/shaping-humanities-data

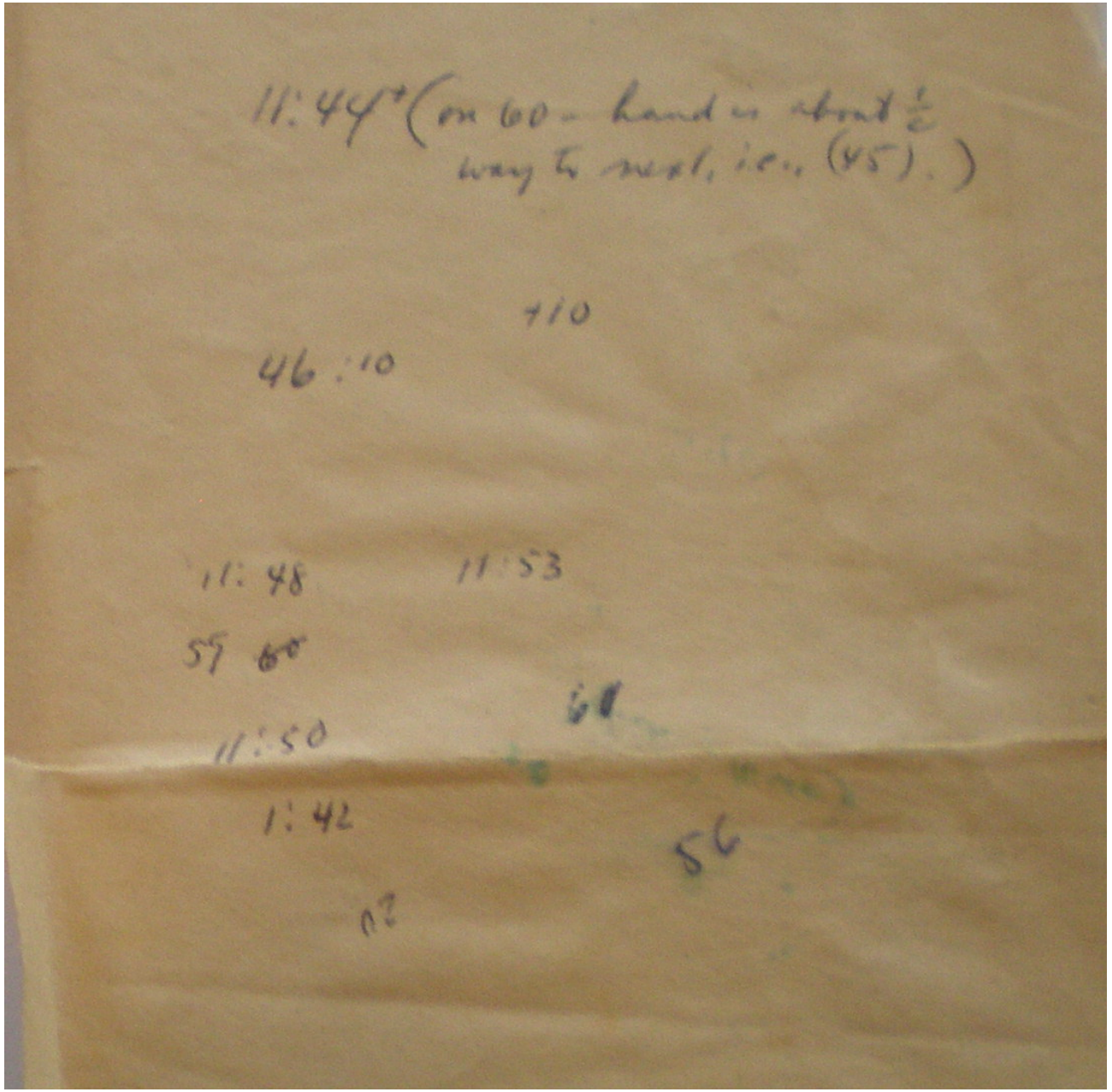
The motivation:

Decaying media must be digitized quickly.

Rich metadata is often an afterthought.


This tape recorded at ☐ 1⅞ ☐ 3¾ ☐ 7½
☐ Monophonic Tracks
☐ Stereophonic Tracks
BILL BERKSON - JOE
BRAINARD - 6/15/71:
2nd half INTERSECTION
Reading, SF.
530 - BOBBIE CREELEY -
JOANNE KYGER - 6/29/71
INTERSECTION - SF.
(in part)
948 (Grand Valley, Mich.
ANSELM HOLLO
TOM WEATHERLY
SIDE TWO: Donald
Paul Blackburn
DIANE WAKOSKI
EDDIE BEAUREGARD
JOHN THORPE
SONY CORP.
TOKYO JAPAN
NOTICE: Buyers shall determine that contents are proper kind for intended use. If defective in the manufacture, labeling, or packaging, contents will be replaced. There are no other warranties, expressed or implied.
DISTRIBUTED BY SUPERSCOPE SUN VALLEY, CALIF.
1250 FEET ½ MIL SPLICE-FREE POLYESTER PR-200


Scotch
3) OLSON LECTURE
4) CONTEMPORARIES
#31
11:44
46:10
11:48
11:53
11:50
1:42
21:21


11:44 (on 60 - hand is about ½ way to next, i.e., (45).)
+10
46:10
11:48
11:53
11:50
1:42

Possible uses of ML in digital audio collections

- Packaging old material for online distribution

Possible uses of ML in digital audio collections

- Packaging old material for online distribution
- Making collections searchable for scholars

Possible uses of ML in digital audio collections

- Packaging old material for online distribution
- Making collections searchable for scholars
- Helping curators find interesting things

Possible uses of ML in digital audio collections

- Packaging old material for online distribution
- Making collections searchable for scholars
- Helping curators find interesting things
- ???

What do you need to classify sound?

- A little theory

What do you need to classify sound?

- A little theory
- Intuition w/r/t what is tractable

What do you need to classify sound?

- A little theory
- Intuition w/r/t what is tractable
- Patience and skills to manage lots of audio files

What do you need to classify sound?

- A little theory
- Intuition w/r/t what is tractable
- Patience and skills to manage lots of audio files
- Features

What do you need to classify sound?

- A little theory
- Intuition w/r/t what is tractable
- Patience and skills to manage lots of audio files
- Features
- A passable algorithm

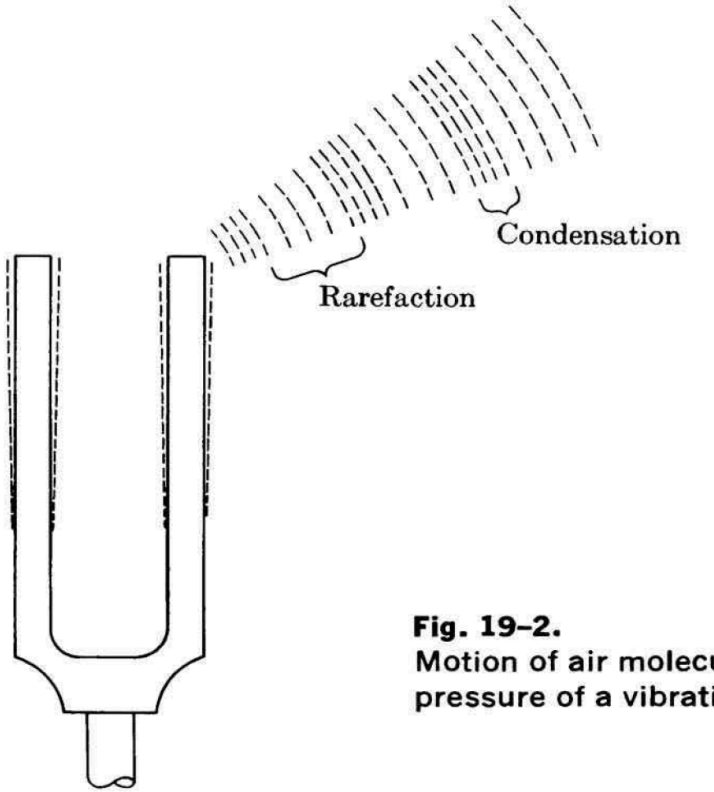
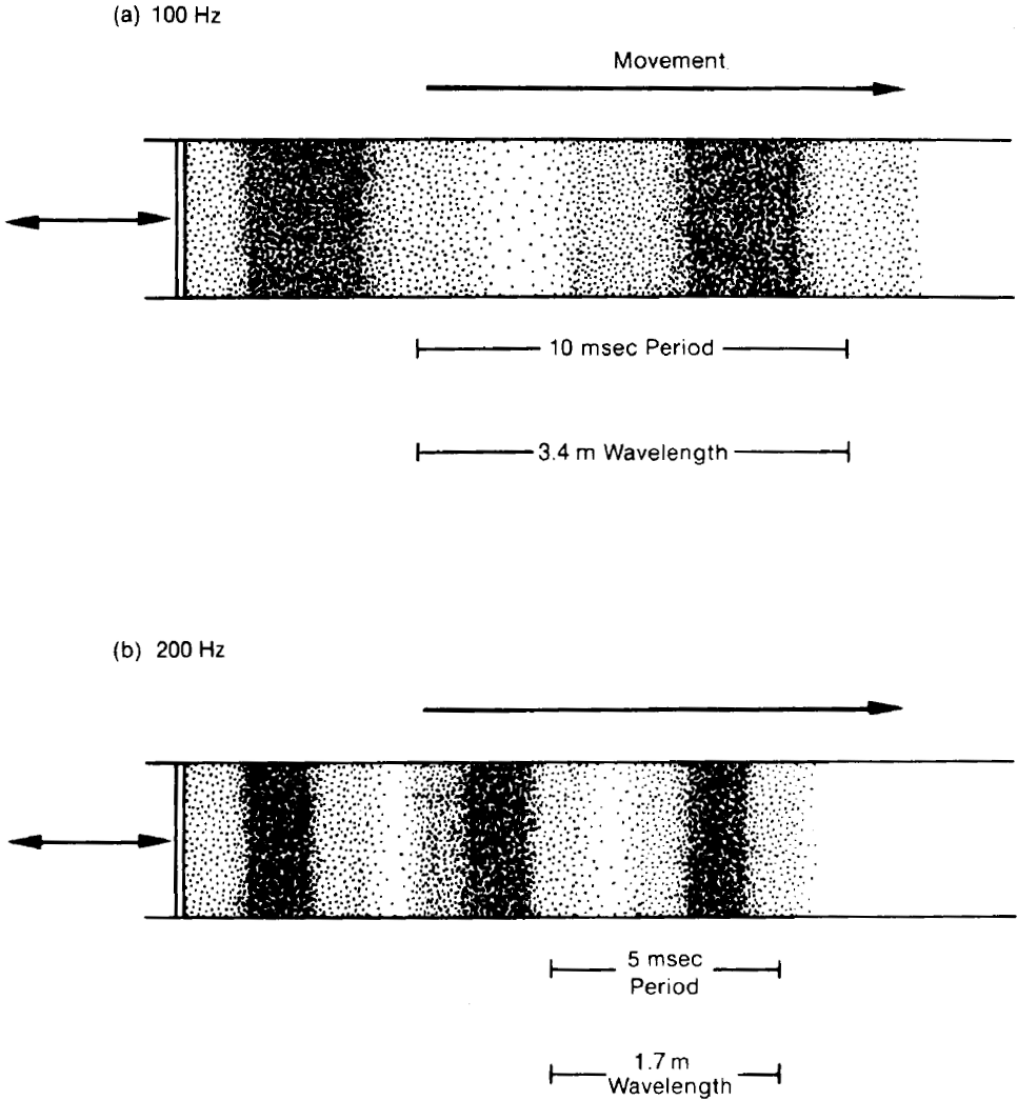
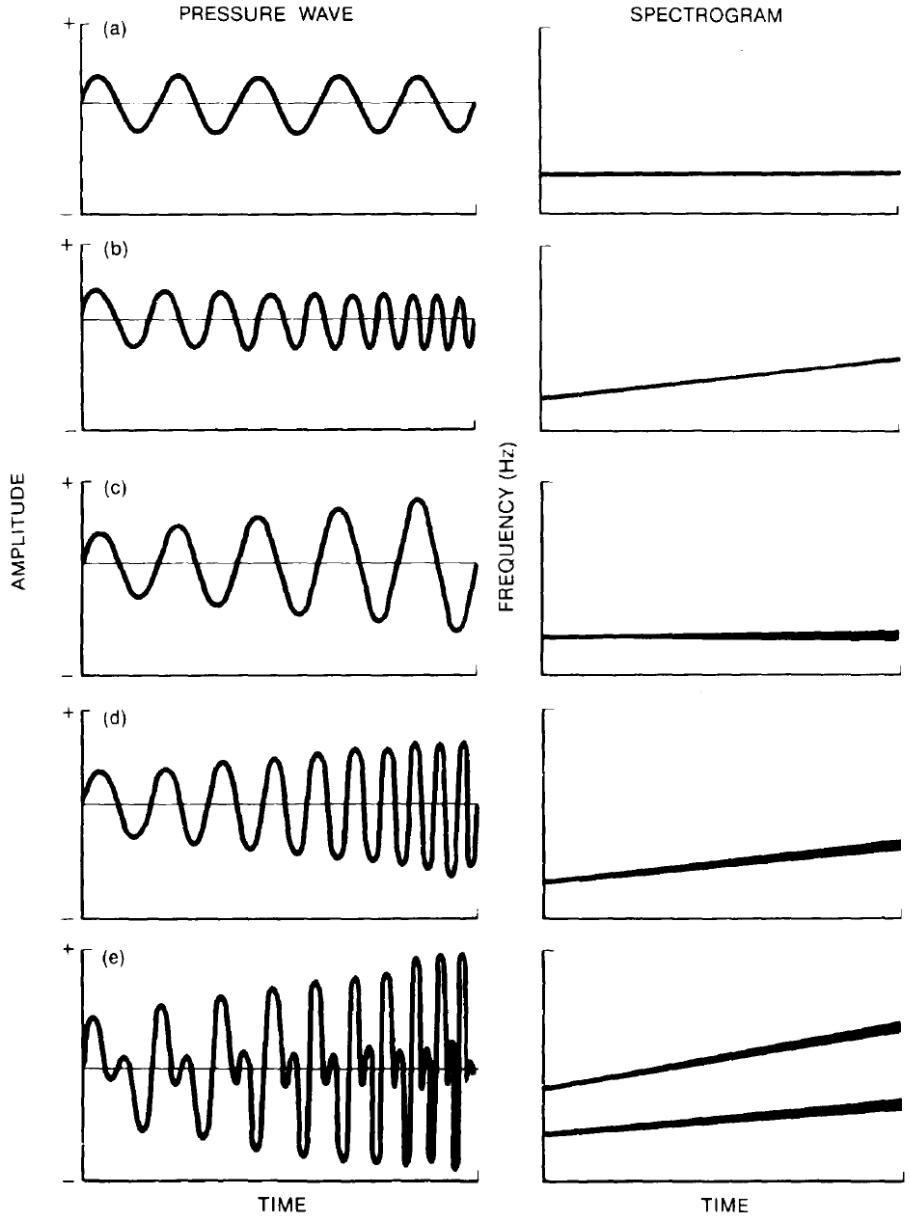
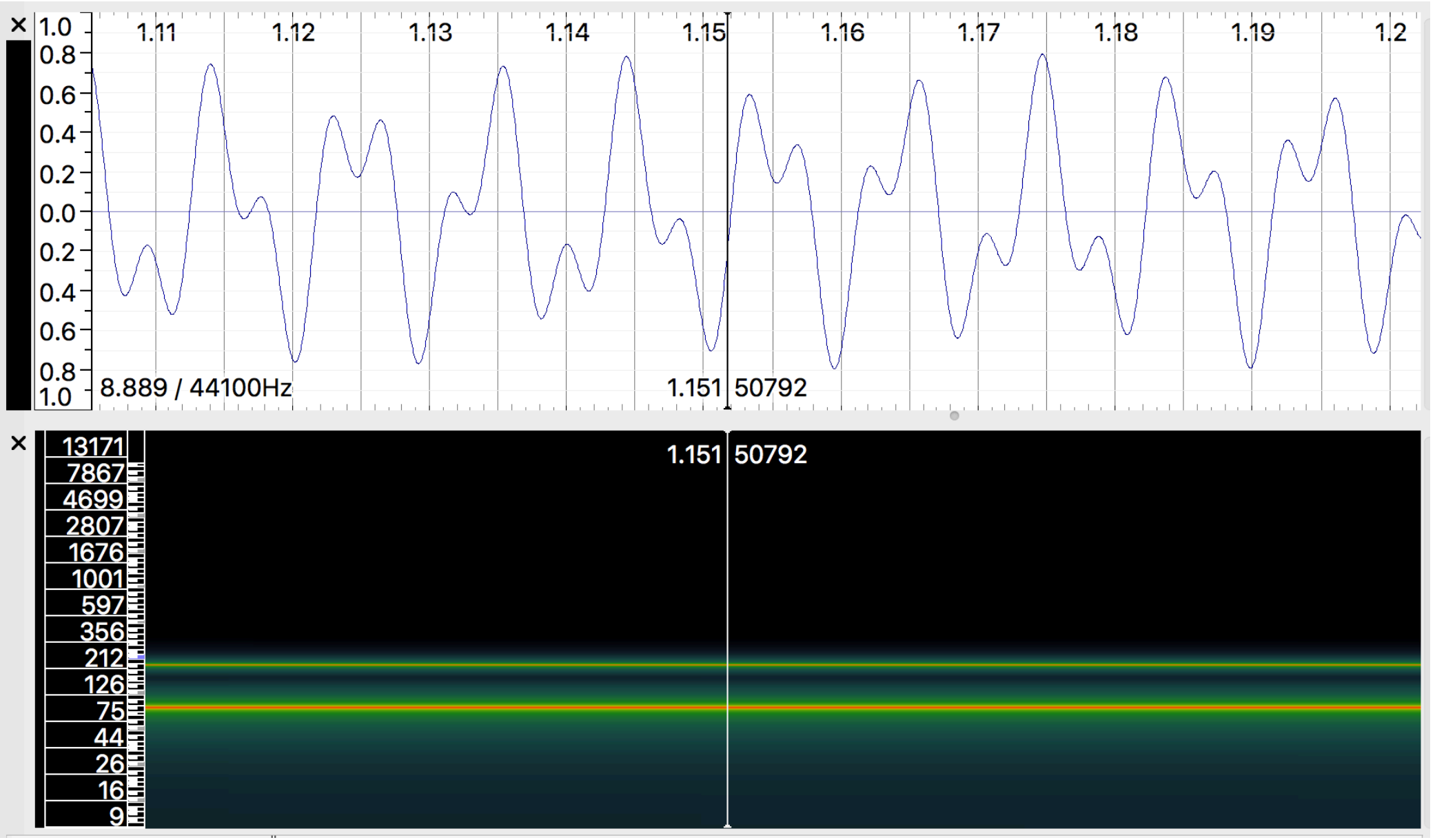


Fig. 19–2.
Motion of air molecules under pressure of a vibrating tuning fork.





Credit: Handel 1989


1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.11
1.12
1.13
1.14
1.15
1.16
1.17
1.18
1.19
1.2
8.889 / 44100Hz
1.151 50792
13171
7867
4699
2807
1676
1001
597
356
212
126
75
44
26
16
9
1.151 50792

Some tasks are easy.

Some tasks are easy.

- Music/speech classification

Some tasks are easy.

- Music/speech classification
- Applause detection

Some tasks are easy.

- Music/speech classification
- Applause detection
- Identifying test tones and silence

Some tasks are hard.

Some tasks are hard.

- Speech to text

Some tasks are hard.

- Speech to text
- Measuring affect

Some tasks are hard.

- Speech to text
- Measuring affect
- Speaker identification

Humanities research ≠ ML research

Humanities research ≠ ML research

- Our collections are noisy.

Humanities research ≠ ML research

- Our collections are noisy.
- Preprocessing is key; CPU cycles are limited.

Humanities research ≠ ML research

- Our collections are noisy.
- Preprocessing is key; CPU cycles are limited.
- I’m a bad programmer.

Humanities research ≠ ML research

- Our collections are noisy.
- Preprocessing is key; CPU cycles are limited.
- I'm a bad programmer.
- Sub-state-of-the-art systems may be OK.

Bottlenecks to wide application

Bottlenecks to wide application

- Compute time

Bottlenecks to wide application

- Compute time
- Existing metadata quality

Bottlenecks to wide application

- Compute time
- Existing metadata quality
- Human ML training time

Bottlenecks to wide application

- Compute time
- Existing metadata quality
- Human ML training time
- Technical literacy

Bottlenecks to wide application

- Compute time
- Existing metadata quality
- Human ML training time
- Technical literacy
- Copyright restrictions

Changes in the past 5 years

- Processors are faster.
- Storage is cheaper.
- VPSes are friendlier.
- More and better core ML tools.
- More and better high-level wrappers.
- GitHub and Stack Exchange are much larger.

Our options

Our options

- Unwieldy research software
  - bob.bio.spear, SIDEKIT, ALIZE

Our options

- Unwieldy research software
  - bob.bio.spear, SIDEKIT, ALIZE
- 2 buckets & done: pyAudioAnalysis

  https://github.com/tyiannak/pyAudioAnalysis

pyAudioAnalysis

- Created and maintained by Theodoros Giannakopoulos
- Built on scikit-learn and the SciPy stack
- Executes training, classification, silence removal, etc.
- Handles dimension reduction and hyperparameter tuning
- Simple interface, comprehensible code

```
from pyAudioAnalysis import audioTrainTest as aT

aT.featureAndTrain(['Background','Creeley'], 1.0, 1.0, aT.shortTermWindow,
aT.shortTermStep, "svm", "svm_creeley", False)
```

Our options

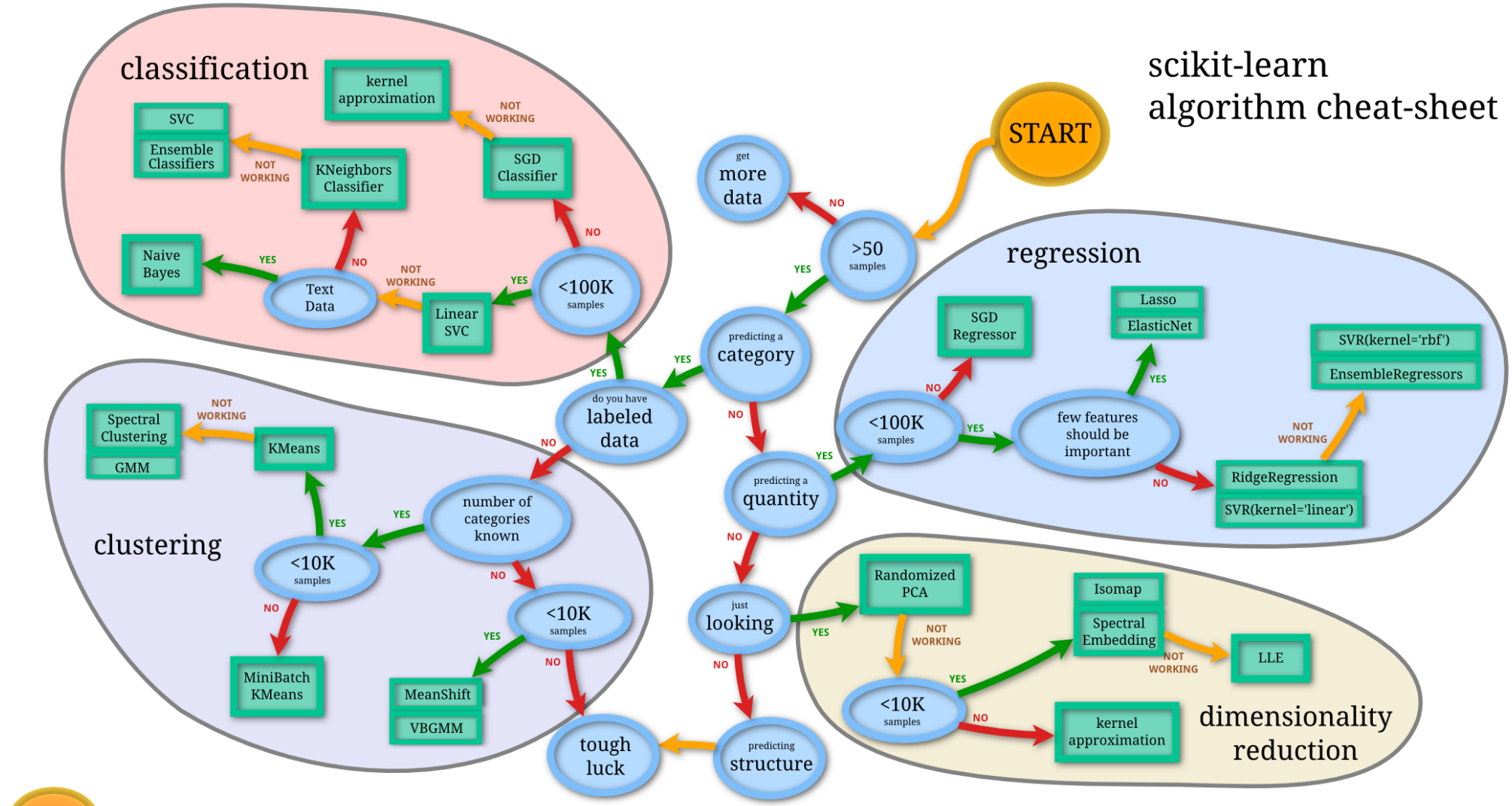
- Unwieldy research software
  - bob.bio.spear, SIDEKIT, ALIZE
- 2 buckets & done: pyAudioAnalysis
  https://github.com/tyiannak/pyAudioAnalysis
- General ML tools
  - scikit-learn, TensorFlow + Keras

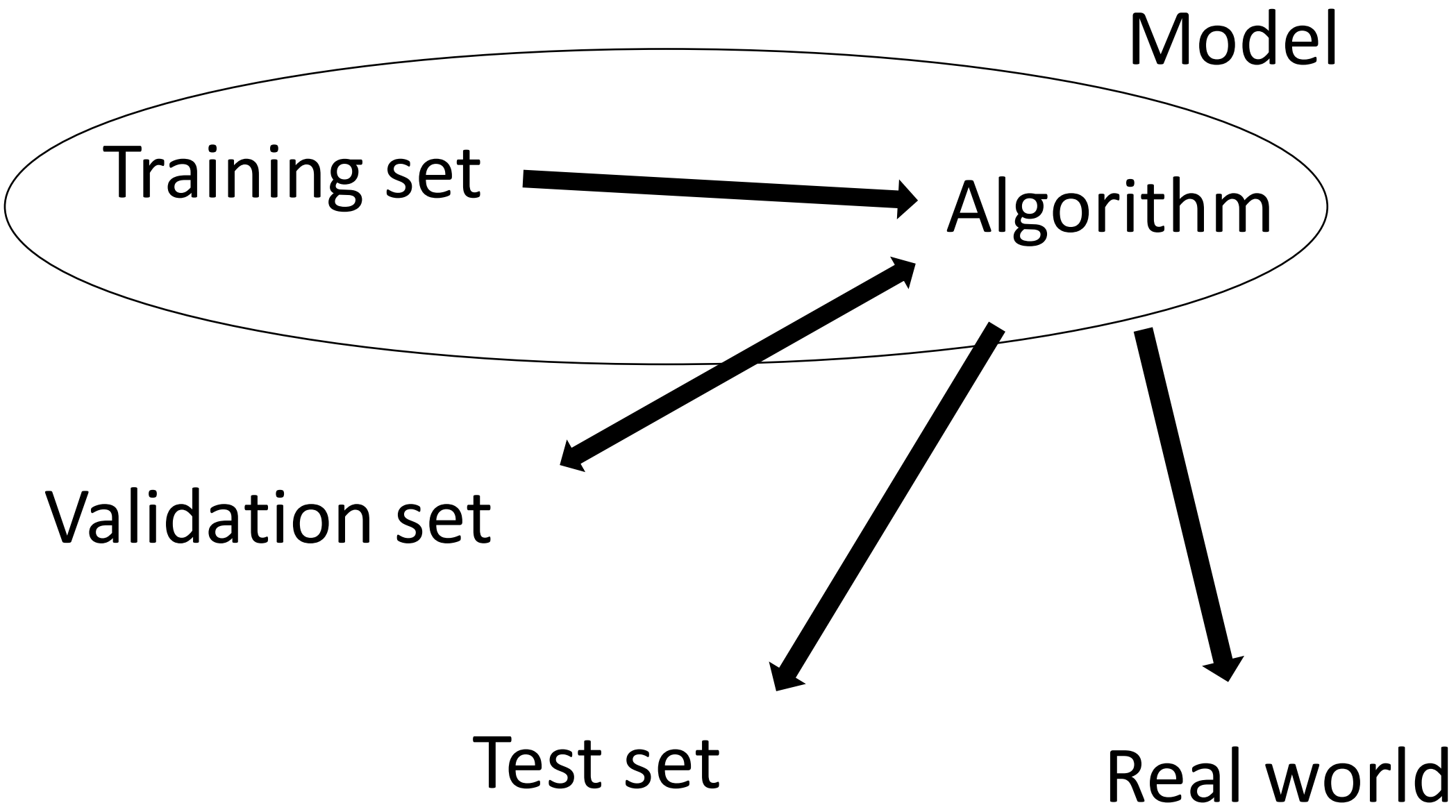
	Feature 1	Feature 2	Feature 3	Feature 4
Label A	11.3166	13.1009	17.1638	1.2876
Label A	0.0023	8.9224	7.3368	12.826
Label A	13.9563	11.0516	8.5414	14.0572
Label A	1.7851	9.1011	10.7928	4.5441
Label B	12.0978	10.4598	5.7786	12.4596
Label B	0.28	17.3005	3.7776	11.1616
Label B	7.3366	13.3285	13.3285	17.5108
Label B	7.1498	13.6181	13.391	7.3366
Label B	10.9546	6.7191	13.0848	3.9394

Types of machine learning

- Supervised learning
  - starting with labeled training data
- Unsupervised learning
  - inferring relationships without class labels
- Reinforcement learning
  - neural networks

scikit-learn algorithm cheat-sheet




Model
Training set
Algorithm
Validation set
Test set
Real world

New tools from HiPSTAS

New tools from HiPSTAS

- Audio Labeler (web app with GUI)
  - github.com/hipstas/audio-labeler

Audio Labeler

Play / Pause | Repeat

/home/audio_labeler/media/20170623full.mp3

2375 seconds

Carol Hills

Apply Label

Background Speaker | Music | Silence | Multiple Speakers | Noise | Not Sure

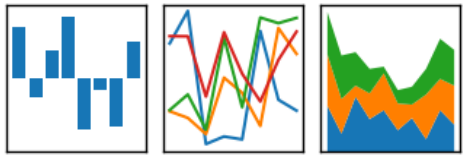
New tools from HiPSTAS

- Audio Labeler (web app with GUI)
  - github.com/hipstas/audio-labeler
- Audio Tagging Toolkit (Python package)
  - github.com/hipstas/audio-tagging-toolkit

pyAudio Analysis

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



scikit learn

FFMPEG

matplotlib

SONIC VISUALISER

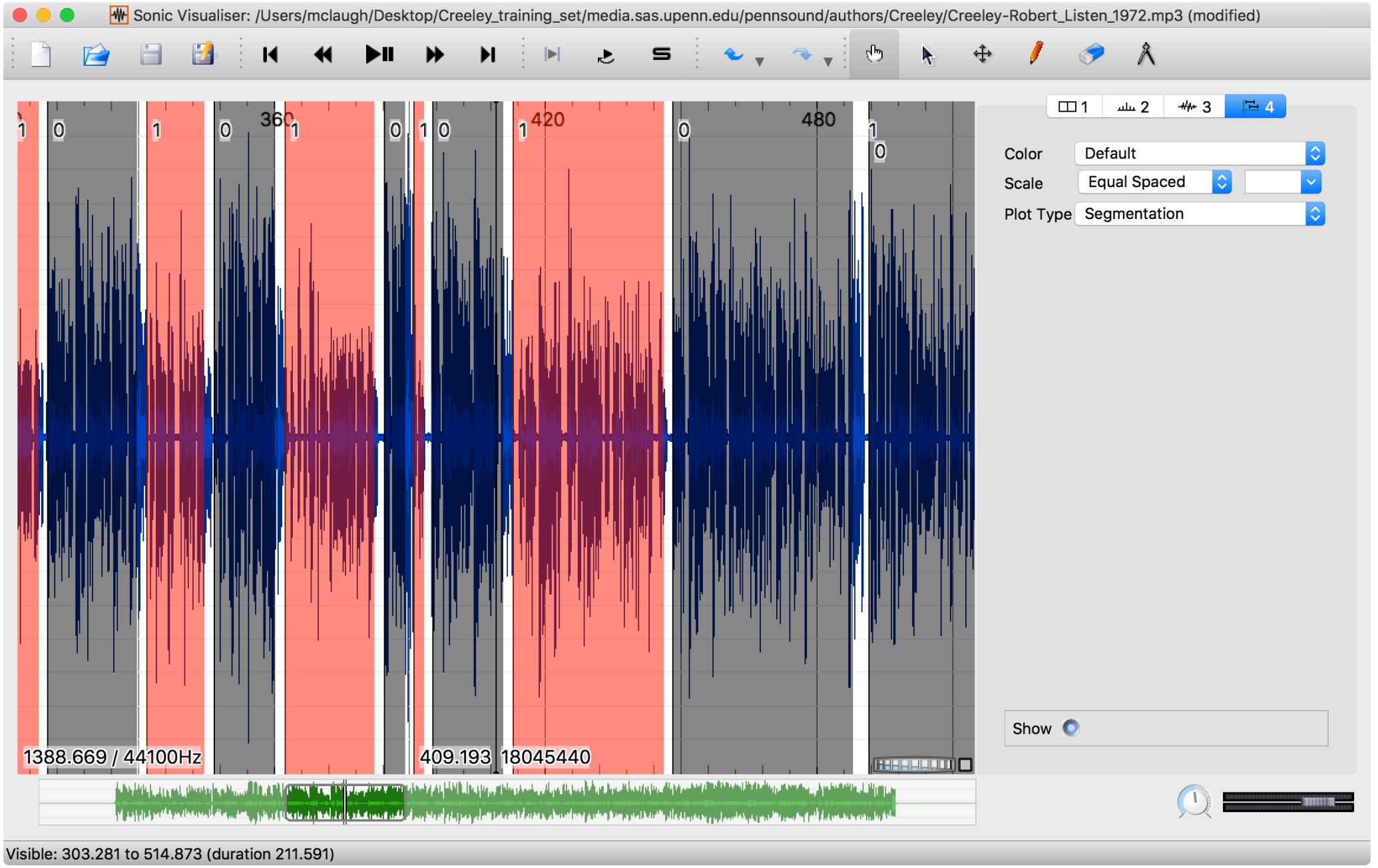
Pydub

Pyperclip

PocketSphinx

New tools from HiPSTAS

- Audio Labeler (web app with GUI)
  - github.com/hipstas/audio-labeler
- Audio Tagging Toolkit (Python package)
  - github.com/hipstas/audio-tagging-toolkit
- Audio ML Lab (Docker container)
  - github.com/hipstas/audio-ml-lab


Sonic Visualiser: /Users/mclaugh/Desktop/Creeley_training_set/media.sas.upenn.edu/pennsound/authors/Creeley/Creeley-Robert_Listen_1972.mp3 (modified)
1
0
1
0
360
1
0
1
0
1
420
0
480
1
0
1388.669 / 44100Hz
409.193
18045440
1
2
3
4
Color
Default
Scale
Equal Spaced
Plot Type
Segmentation
Show
Visible: 303.281 to 514.873 (duration 211.591)

```
142.000000000,0,4.000000000,Applause
498.000000000,0,10.000000000,Applause
512.024761904,1,1406.000000000,Robert Creeley
1937.000000000,0,23.000000000,Applause
```

Sonic Visualiser
CSV file

Additional slides for Q&A

Types of machine learning

- Regression
  - output: a number
- Classification
  - output: a class (i.e., category)

Disciplines in audio ML

- Music information retrieval (MIR)
  - audio fingerprinting (Shazam)
  - automatic transcription
  - audio thumbnailing
- Speech to text
- Speaker recognition
  - speaker identification
  - speaker verification

Some key algorithms

- Linear regression
  - numbers in, numbers out
- Logistic regression
  - numbers and/or classes in, classes out

Some key supervised learning algorithms

- Linear regression
  - numbers in, numbers out
- Logistic regression
  - numbers and/or classes in, classes out

Logistic regression

- prone to overfitting (training data not representative of real world)

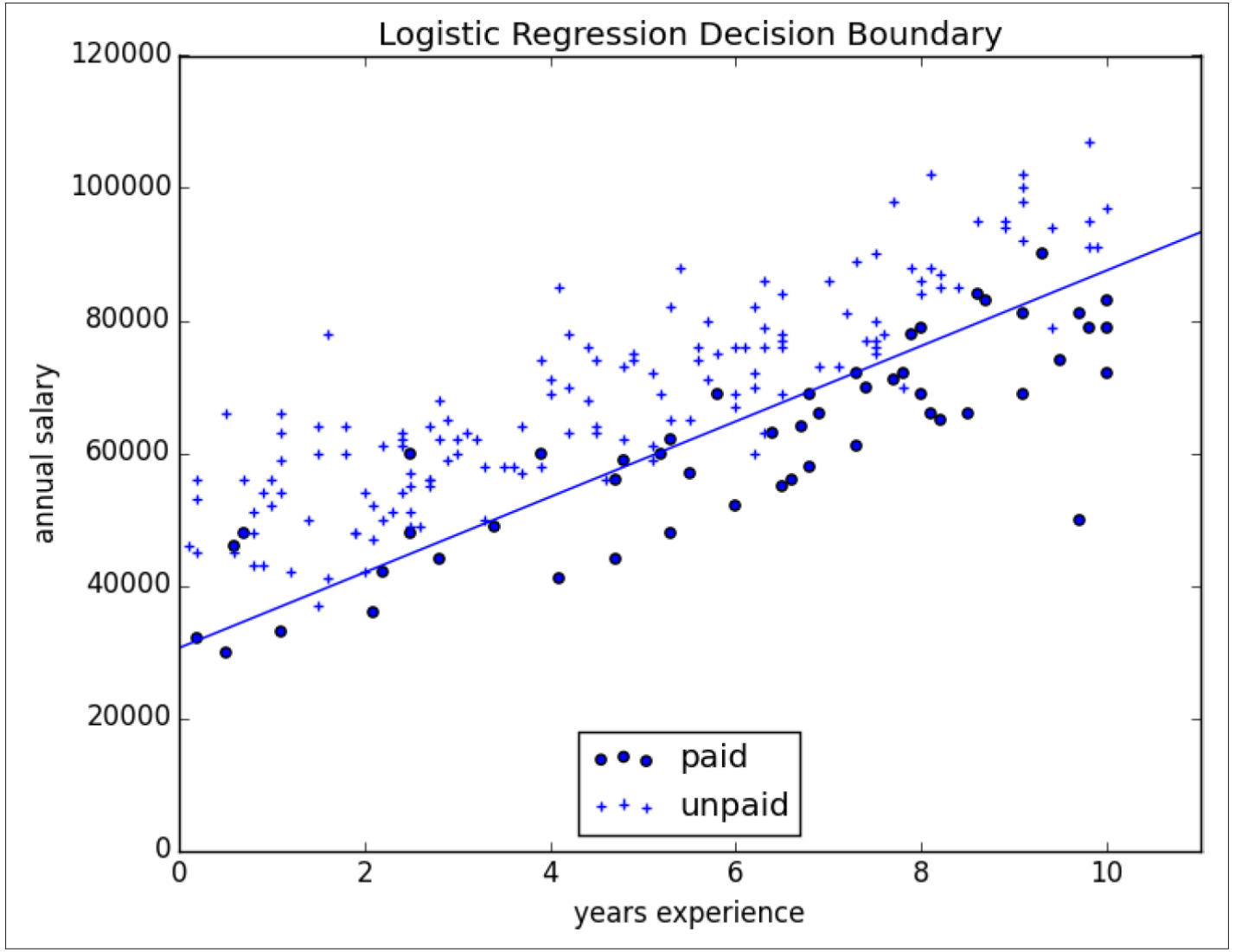
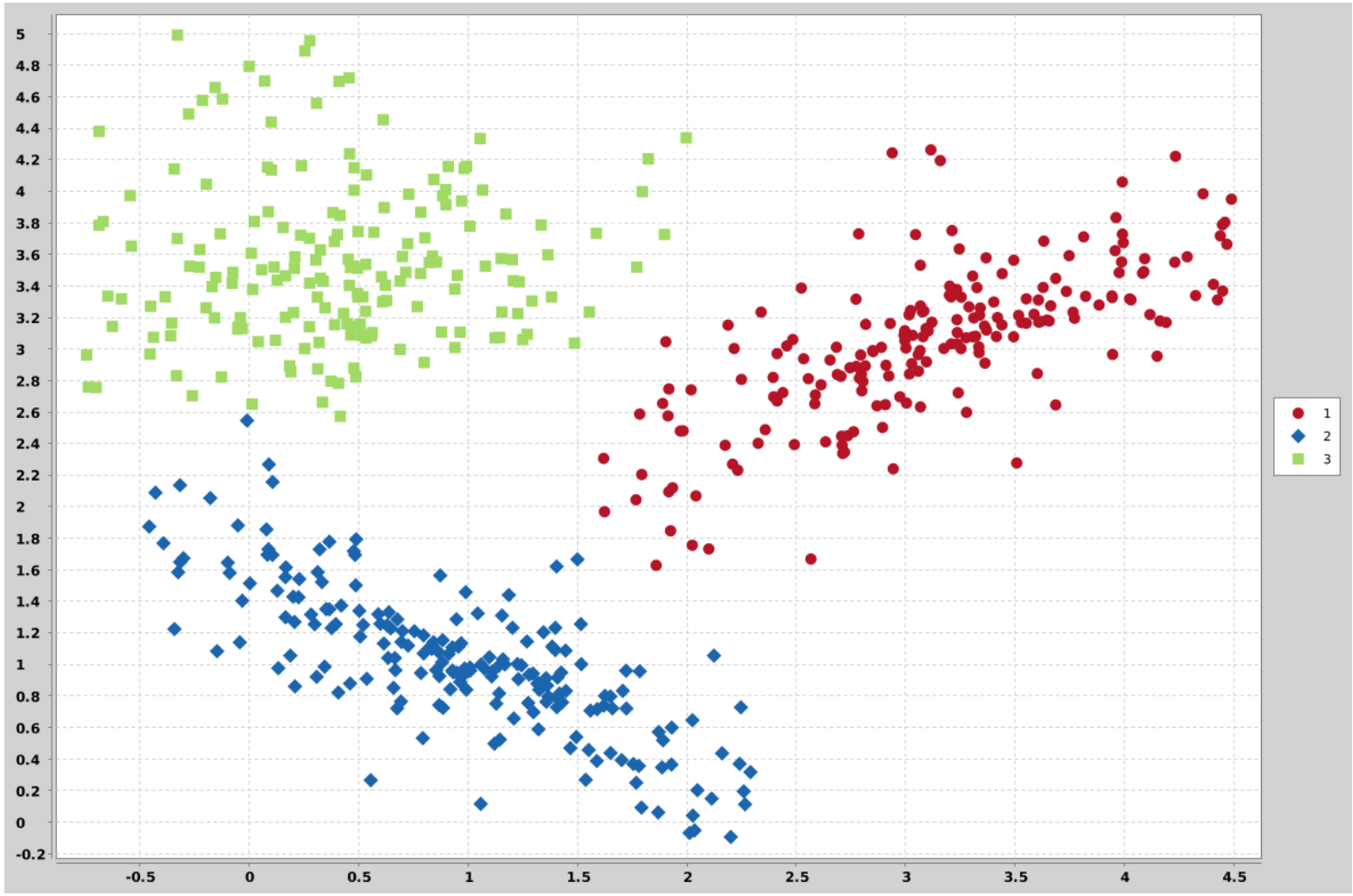


Figure 16-5. Paid and unpaid users with decision boundary

Credit: Grus 2015

K nearest neighbor

- the curse of dimensionality
- intuitive, but not very efficient



Credit: Grus 2015

Support vector machine (SVM)

- Fast, cheap, and pretty good
- Your #1 go-to

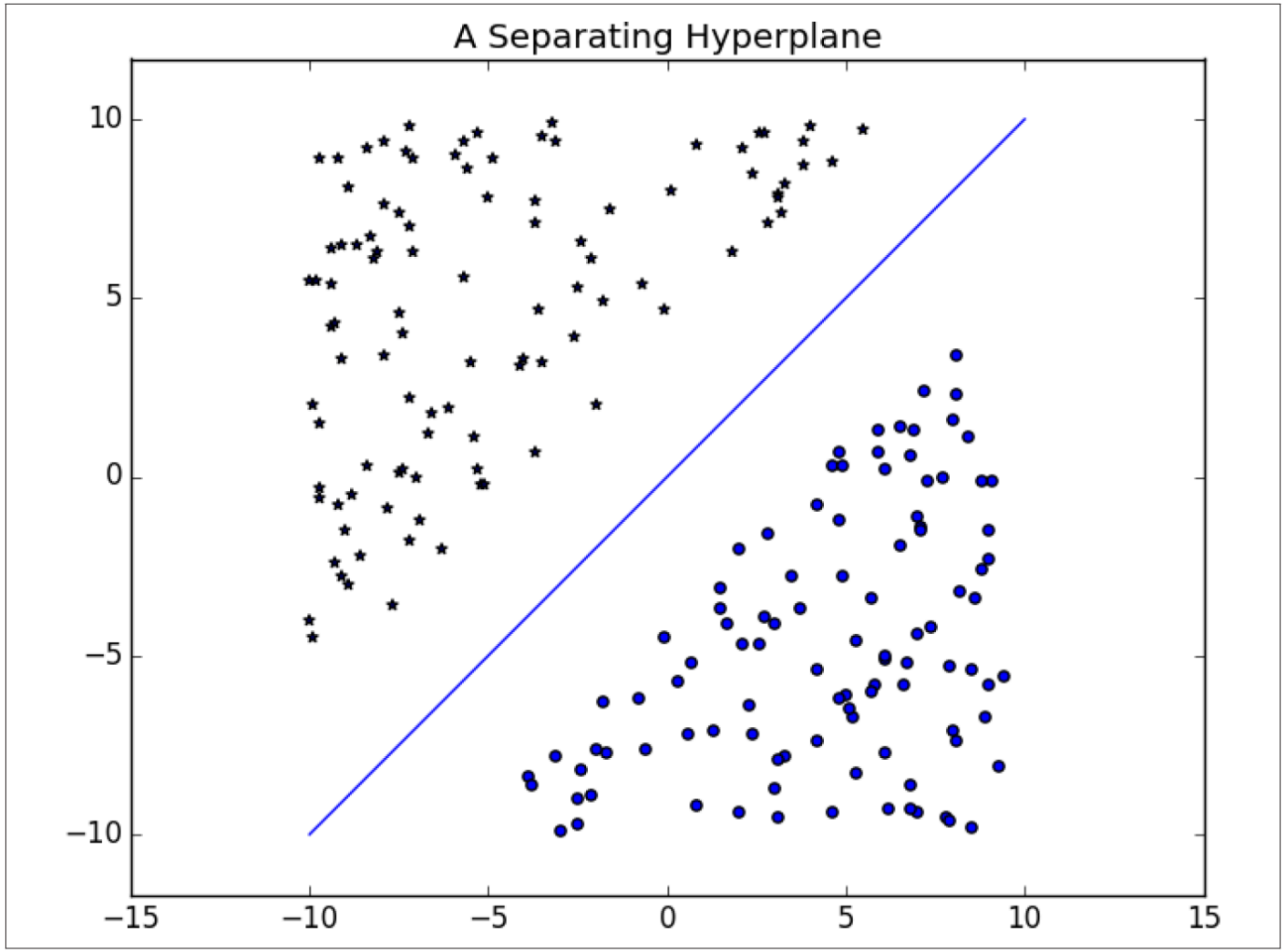
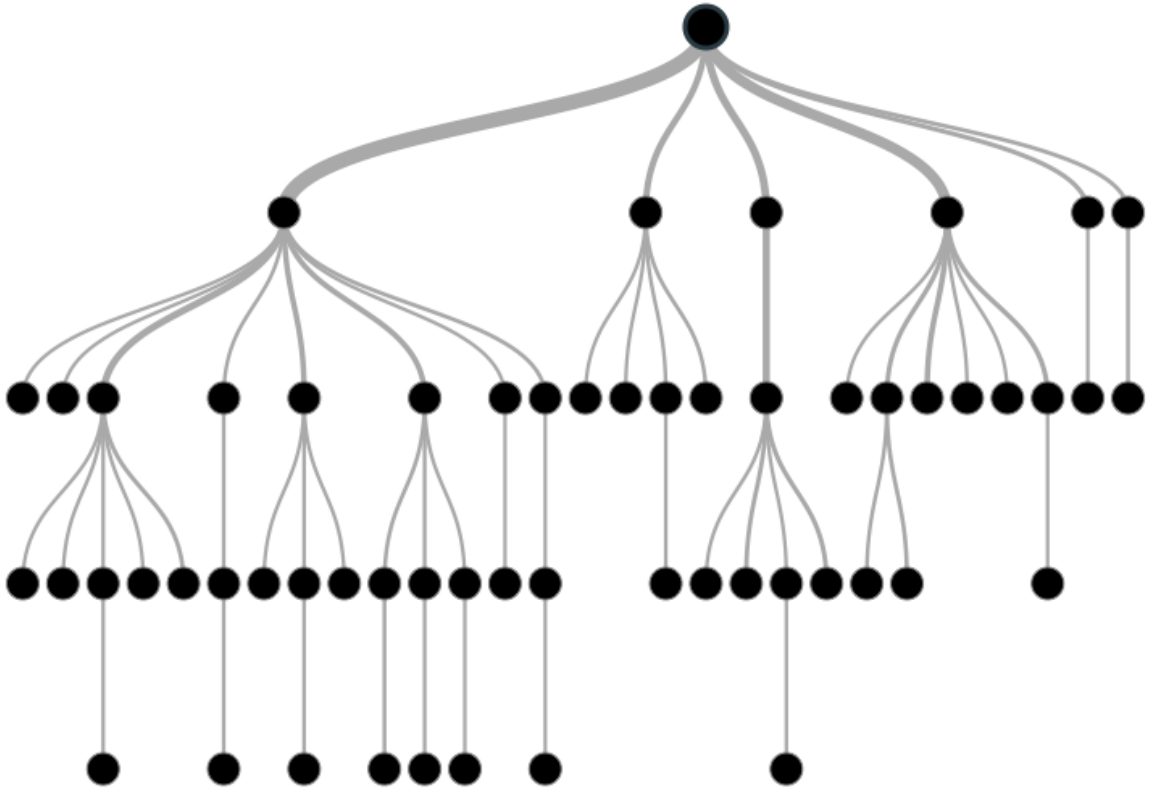


Figure 16-6. A separating hyperplane

Credit: Grus 2015

Decision trees

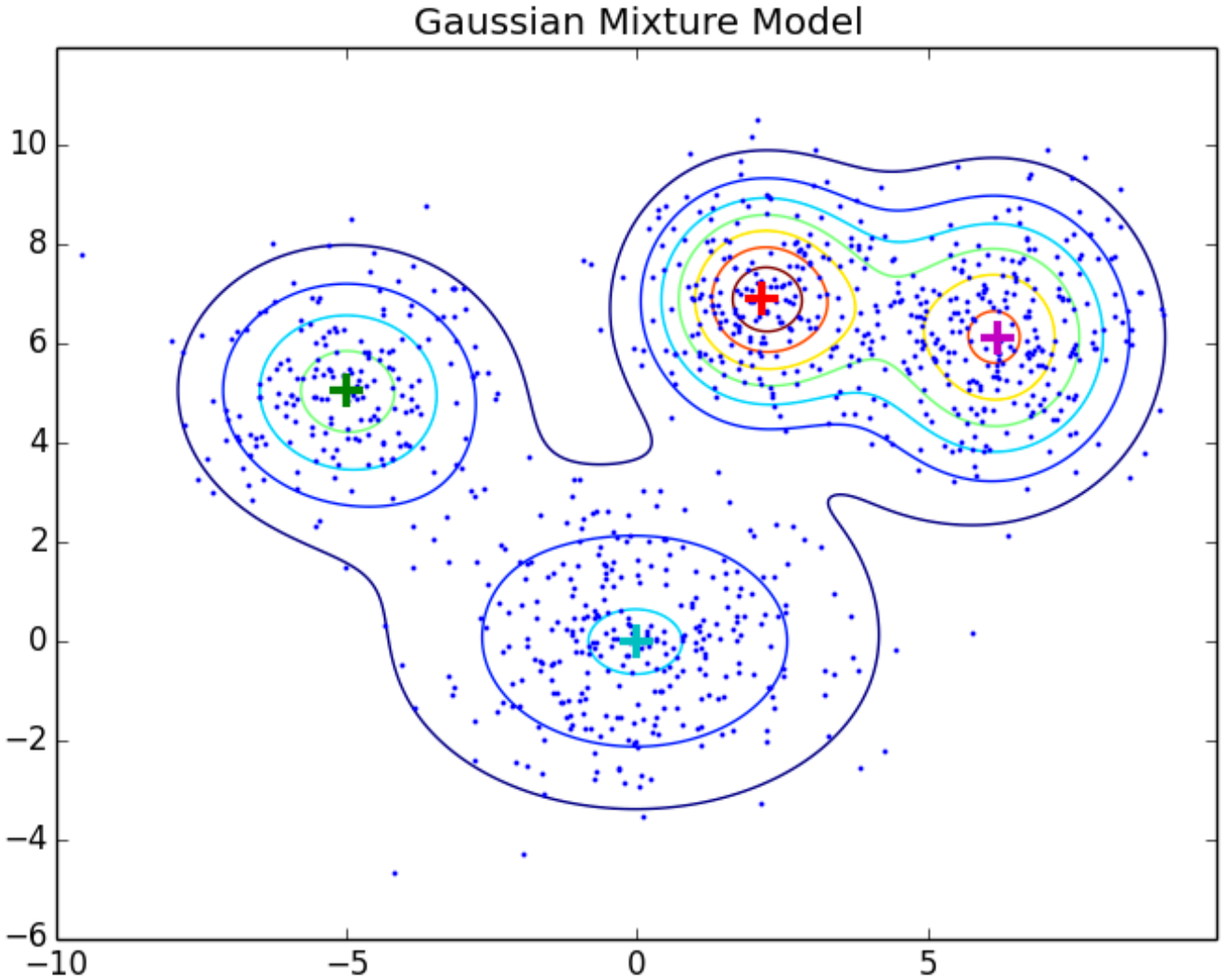
- Slower than SVMs
- Highly interpretable results
- Random forests and gradient boosting are improved variations



Credit: Grus 2015

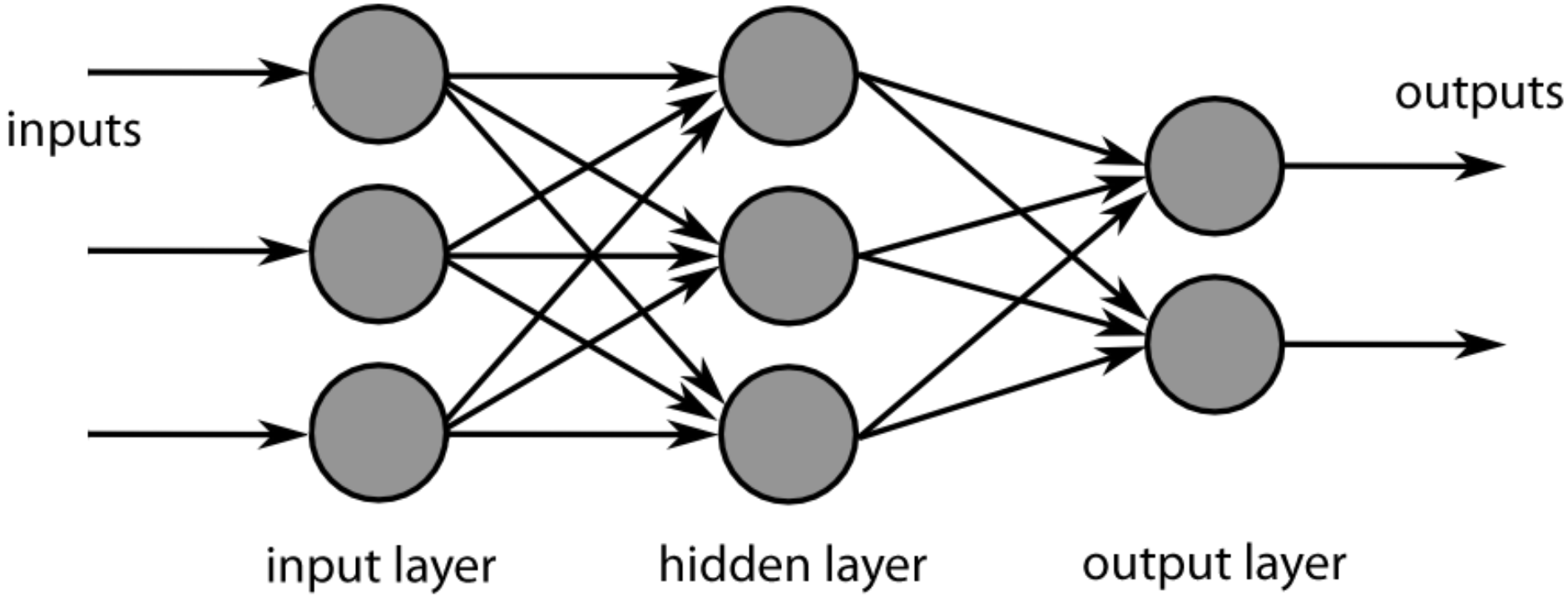
Gaussian mixture models

- OK with shades of gray
- Well-suited for speaker identification



Credit: Yu Zhu

Neural networks



- "The future"
- Output not interpretable
- Slow, expensive, steep learning curve

Credit: Wikimedia Commons

SPEECH TRAX

PRESENTATION DEMONSTRATION CONTACT FR / EN

00:02:04 / 00:15:00

04:30 m 05:00 m 05:30 m 06:00 m 06:30 m 07:00 m 07:30 m 08:00 m 08:30 m 09:00 m 09:30 m 10:00 m 10:30 m

ELISABETH ALLAIN (67.6%)

SPEAKER 2

FRANÇOIS HOLLANDE (59.5%)

LAURENT FABIUS (58.3%)

elisabeth allain

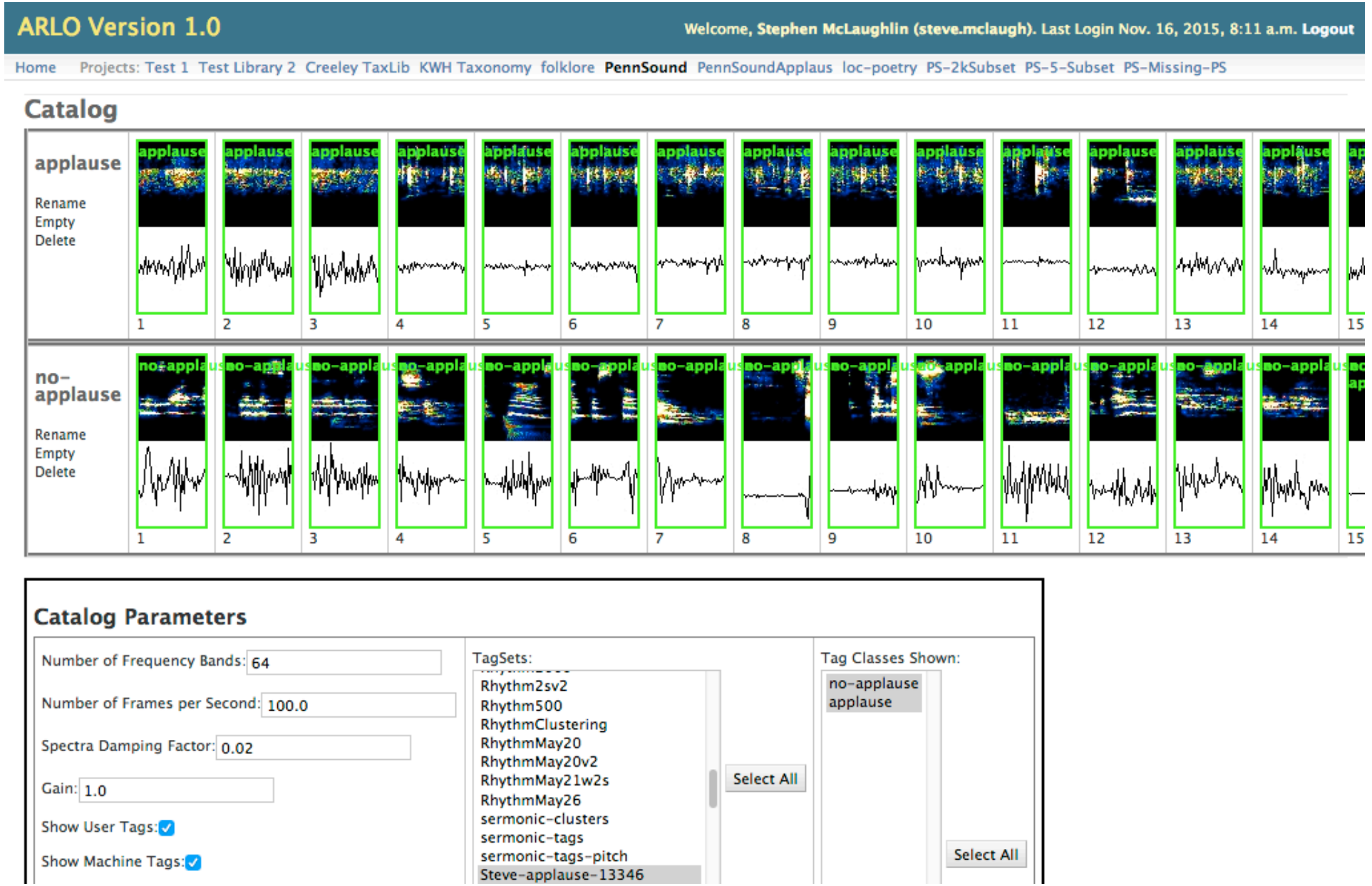
00:01:36
00:02:04

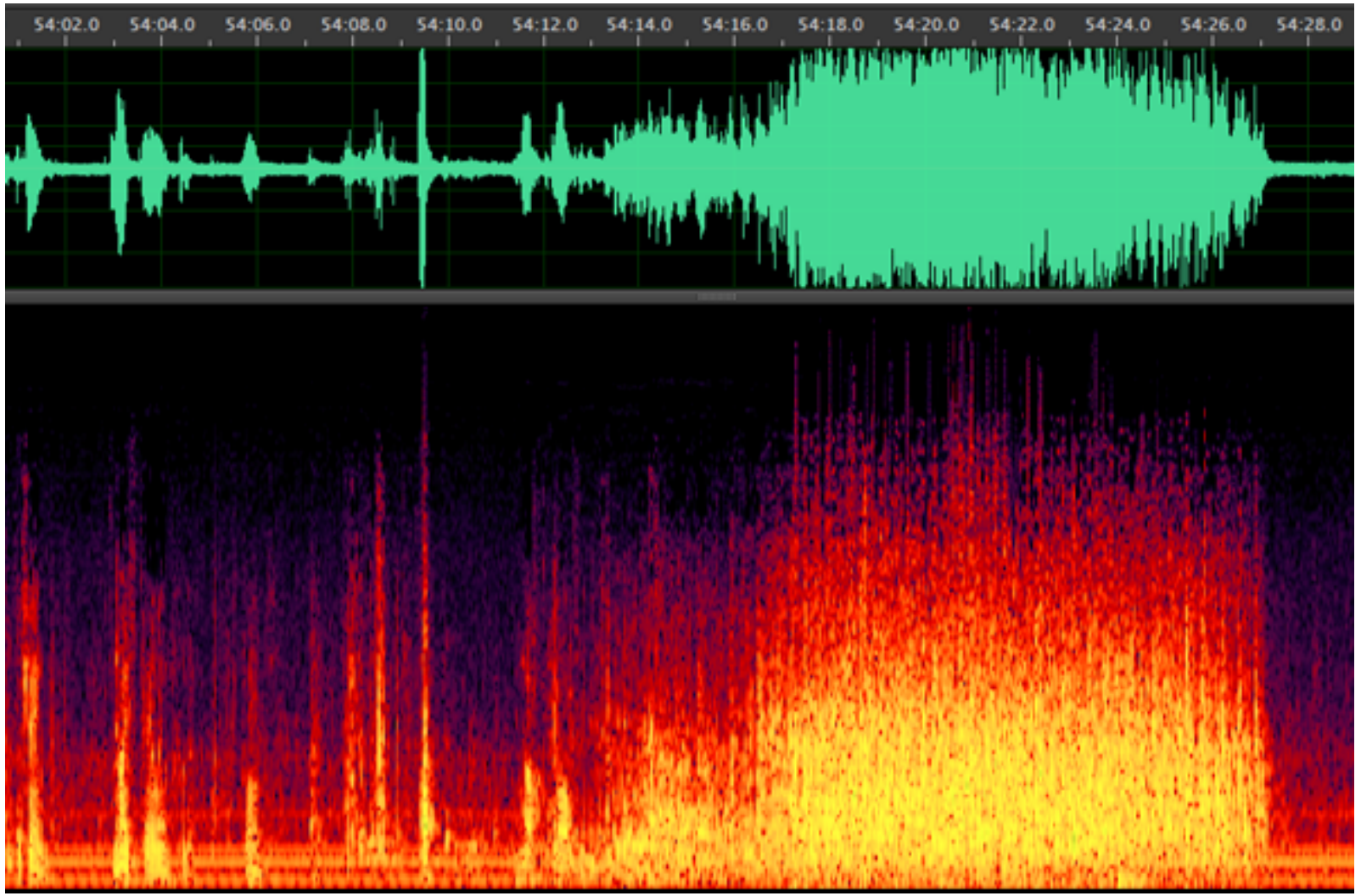
Le journal est cette réunion tout d'abord réunion internationale sur le Liban, François Hollande, accueille aujourd'hui son homologue Michel Sleimane à l'Élysée mais aussi une vingtaine de convives, six mois après la création d'un groupe international de soutien, il s'agit de faire le point sur les conséquences de la crise syrienne sur le pays du sel, chaque mois, selon le HCR, le pays verra débarquer 60000 nouveaux par réfugiés, la situation est devenue plus que critique.

françois hollande

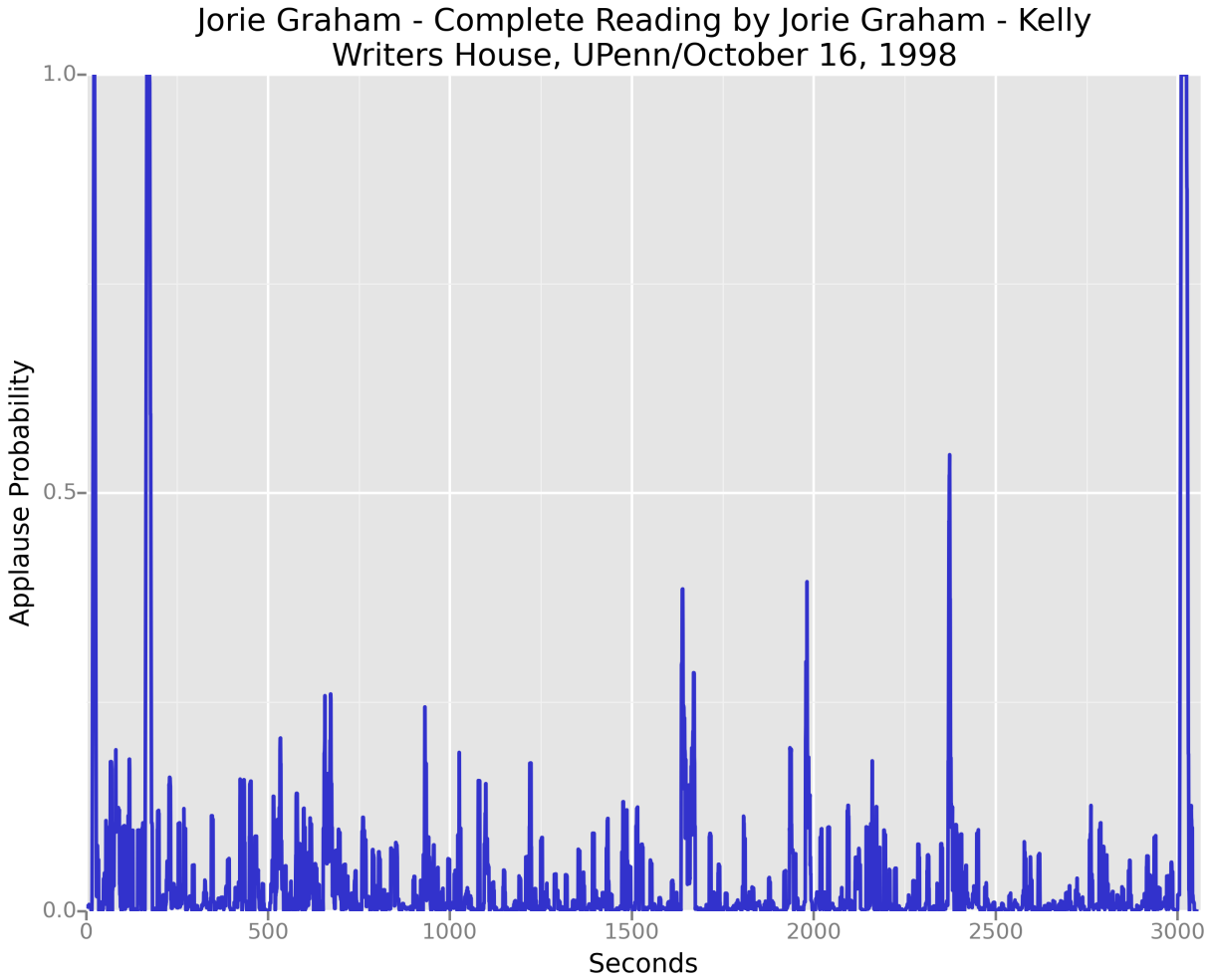
00:02:04
00:02:47

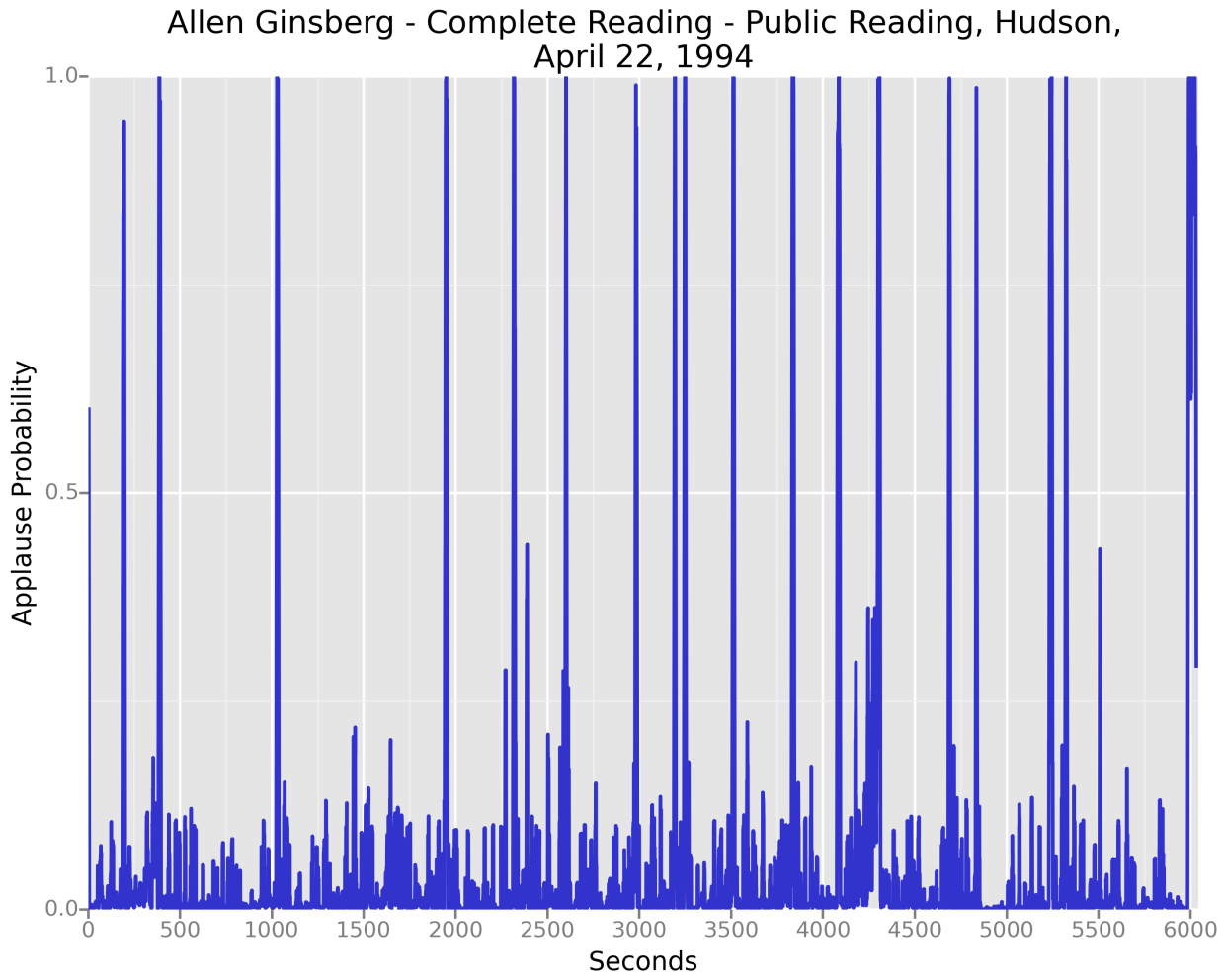
La communauté internationale, elle doit se combiner, et c'est ce que fait aujourd'hui dans trois directions : d'abord, l'aide aux réfugiés, il y a des montants financiers qui ont été dégagés, ils doivent être complétés et ils doivent être amplifié la seconde priorité, c'est le soutien à l'économie libanaise la dernière priorité c'est de garantir la sécurité du Liban et de permettre à l'armée libanaise d'avoir des équipements indispensables, la France,


ARLO Version 1.0
Welcome, Stephen McLaughlin (steve.mclaugh). Last Login Nov. 16, 2015, 8:11 a.m. Logout
Home Projects: Test 1 Test Library 2 Creeley TaxLib KWH Taxonomy folklore PennSound PennSoundApplaus loc-poetry PS-2kSubset PS-5-Subset PS-Missing-PS
Catalog
applause
Rename
Empty
Delete
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
no-applause
Rename
Empty
Delete
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Catalog Parameters
Number of Frequency Bands: 64
Number of Frames per Second: 100.0
Spectra Damping Factor: 0.02
Gain: 1.0
Show User Tags:
Show Machine Tags:
TagSets:
Rhythm2sv2
Rhythm500
RhythmClustering
RhythmMay20
RhythmMay20v2
RhythmMay21w2s
RhythmMay26
sermonic-clusters
sermonic-tags
sermonic-tags-pitch
Steve-applause-13346
Select All
Tag Classes Shown:
no-applause
applause
Select All



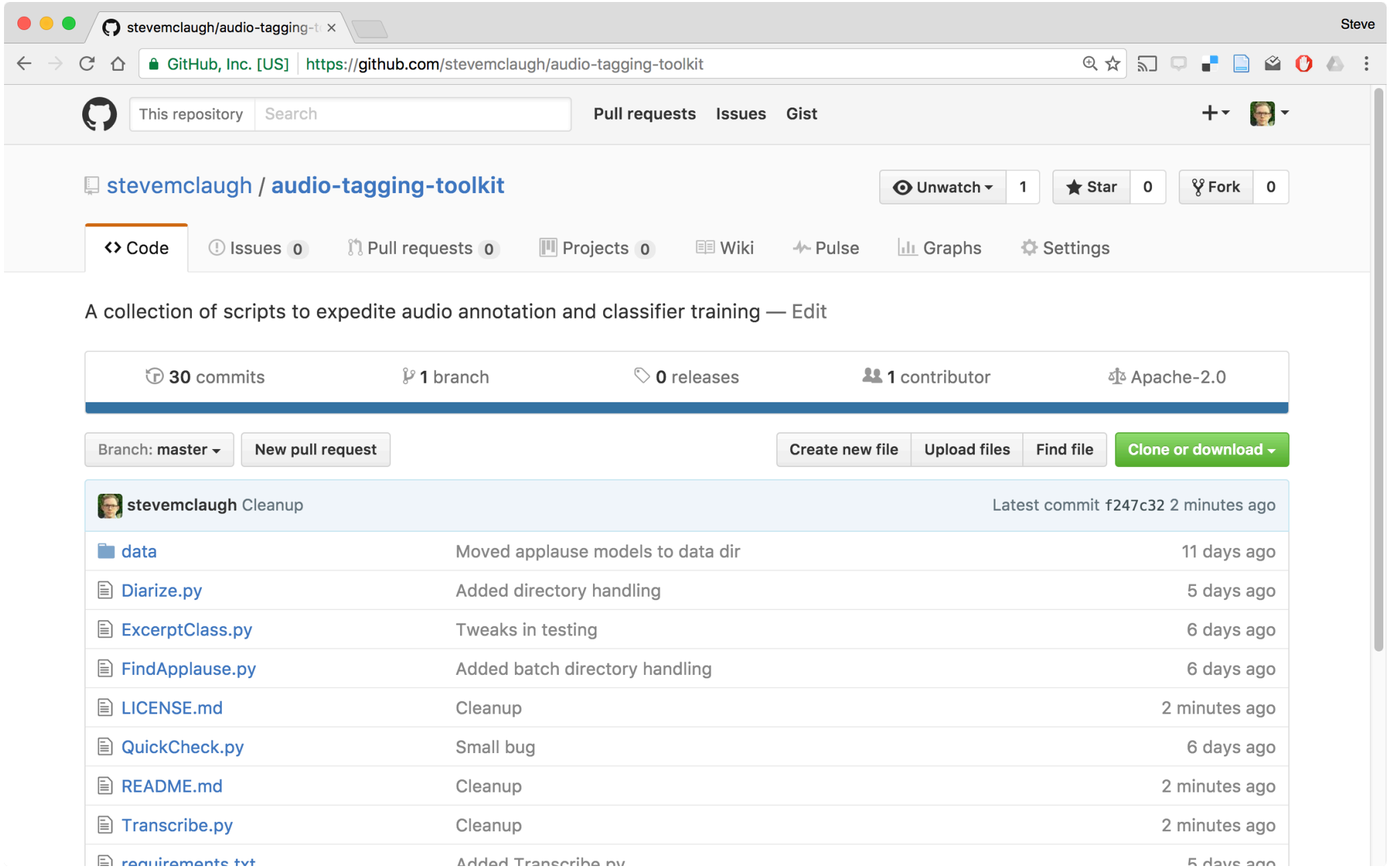
Reading by Charles Olson, Vancouver Poetry Conference, 1963


Jorie Graham - Complete Reading by Jorie Graham - Kelly Writers House, UPenn/October 16, 1998
Applause Probability
1.0
0.5
0.0
0
500
1000
1500
2000
2500
3000
Seconds


Allen Ginsberg - Complete Reading - Public Reading, Hudson, April 22, 1994
Applause Probability
1.0
0.5
0.0
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
Seconds

	Bernstein	Creeley	Armantrout	DuPlessis	Andrews	Watten	Mean Applause (seconds)	No. Readings
Bernstein		0.3092	0.4657	**0.0496**	**0.0069**	**0.0021**	31.97	29
Creeley	0.6908		0.4815	0.1416	0.0600	**0.0356**	29.0	16
Armantrout	0.5343	0.5185		0.1239	**0.0089**	**0.0076**	25.67	12
DuPlessis	0.9504	0.8584	0.8761		0.3752	0.2018	16.88	12
Andrews	0.9931	0.9400	0.9911	0.6248		0.1607	12.39	18
Watten	0.9979	0.9644	0.9924	0.7982	0.8393		12.07	14

P-values for Pairwise Directional Mann-Whitney *U* Tests Between Six Poets' Applause Durations


stevemclaugh/audio-tagging-t
Steve
GitHub, Inc. [US]
https://github.com/stevemclaugh/audio-tagging-toolkit
This repository
Search
Pull requests
Issues
Gist
stevemclaugh / audio-tagging-toolkit
Unwatch 1
Star 0
Fork 0
Code
Issues 0
Pull requests 0
Projects 0
Wiki
Pulse
Graphs
Settings
A collection of scripts to expedite audio annotation and classifier training — Edit
30 commits
1 branch
0 releases
1 contributor
Apache-2.0
Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
stevemclaugh Cleanup
Latest commit f247c32 2 minutes ago
data Moved applause models to data dir 11 days ago
Diarize.py Added directory handling 5 days ago
ExcerptClass.py Tweaks in testing 6 days ago
FindApplause.py Added batch directory handling 6 days ago
LICENSE.md Cleanup 2 minutes ago
QuickCheck.py Small bug 6 days ago
README.md Cleanup 2 minutes ago
Transcribe.py Cleanup 2 minutes ago

 ⟶ applause detection and/or diarization ⟶ manual validation ⟶

speaker recognition ⟶ manual validation ⟶ annotated corpus ⟶

retrieval, segmentation, new ML training, etc.

General metadata goals

- Segmentation by speaker
- Segmentation by event structure
- High-level labels
  - Language
  - Genre
  - Topic/theme
- Low-level labels
  - Notable passages
  - Material features/quirks

General search goals

- Acoustic fingerprint search (à la Shazam)
- Locate speaker of interest
- Locate sound classes of interest
- Full-text search