Detecting Deep Fakes With Mice

Machine vs. Biology

“Fake News” Circa 1938

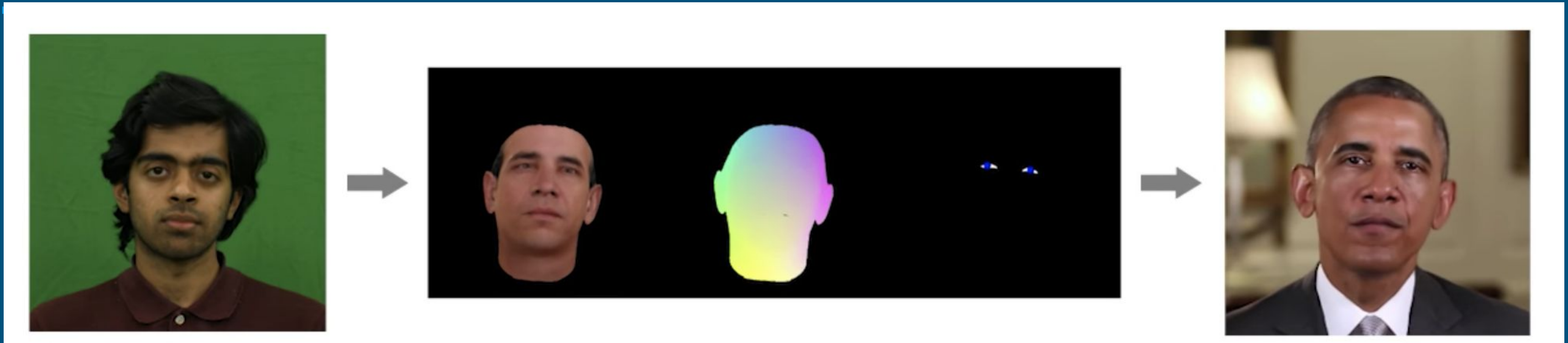
Mars Attacks!, 1938

“War of the Worlds” Hoax

Mercury Theatre, Manhattan

Orson Welles: “Sorry about it!”

2019: AI-Synthesized Media



“Deep Video Portraits,” SIGGRAPH

Face Swap, **Puppet Master**, Lip Sync, Voice Cloning…

ML is crossing the “uncanny valley” faster than CG!

Cybersecurity Threat ?

THE HILL

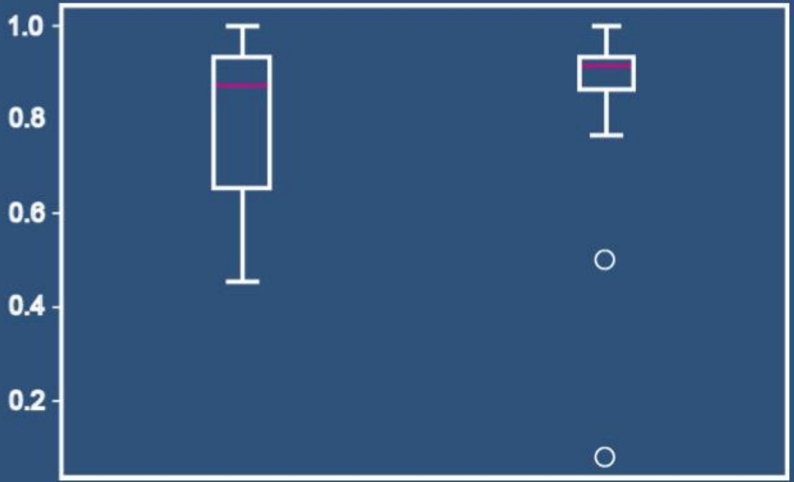
Senators unveil bipartisan bill to target 'deepfake' video threat

“The capability to do all of this is real. It exists now.” - Marco Rubio, Senator

“You don’t need software engineers anymore. You just download it to your PC and run it.” - Chris Bregler, Google

But Who Is Really Fooled ?

Humans
88%



Machines
92%

Fake Speech Study ASVSpoof 2019 DataSet

Machines

Alexander Comerford

Biology

Jonathan Saunders

What is a deep fake?

- Term coined in ~2017
    - Same time as published landmark paper “Generative Adversarial Networks”[6]
- Compound word of “deep learning” and “fake”
- Usually associated with synthesizing images and videos
- Broadly shows the abilities of generative modeling
- The public associates deep fakes with political videos or pornography
- Data about a person -> Puppet of the person

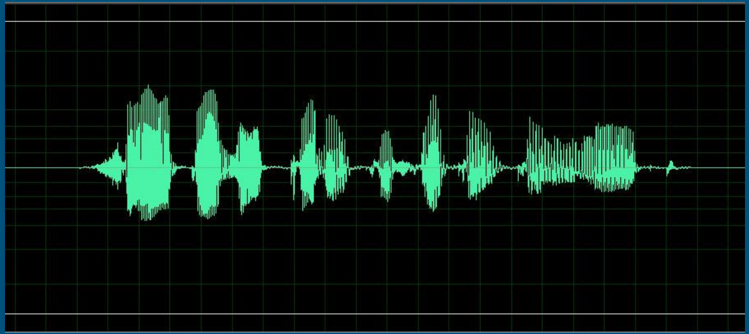
How is a deep fake made?

- Deep fakes are a product of generative modeling and Neural Networks
    - Create a mapping from one data type to another (ex: text to speech)
    - Given data, find a model that generates new but similar samples
    - Unsupervised learning (no data labels, just training data!)
- “Deep” Neural Networks produce the most “fake” samples
- Convincing fakes requires significant resources
    - Fully representative dataset
    - Compute

Good deep fakes are HARD!

- Synthesizing Obama [1]
  - Training:
    - 17 hours of data
    - ~2 weeks on cpu
    - ~2 hours on gpu
    - ? hours of work

general deep fakes are EASY and FUN!



WaveNet [3]

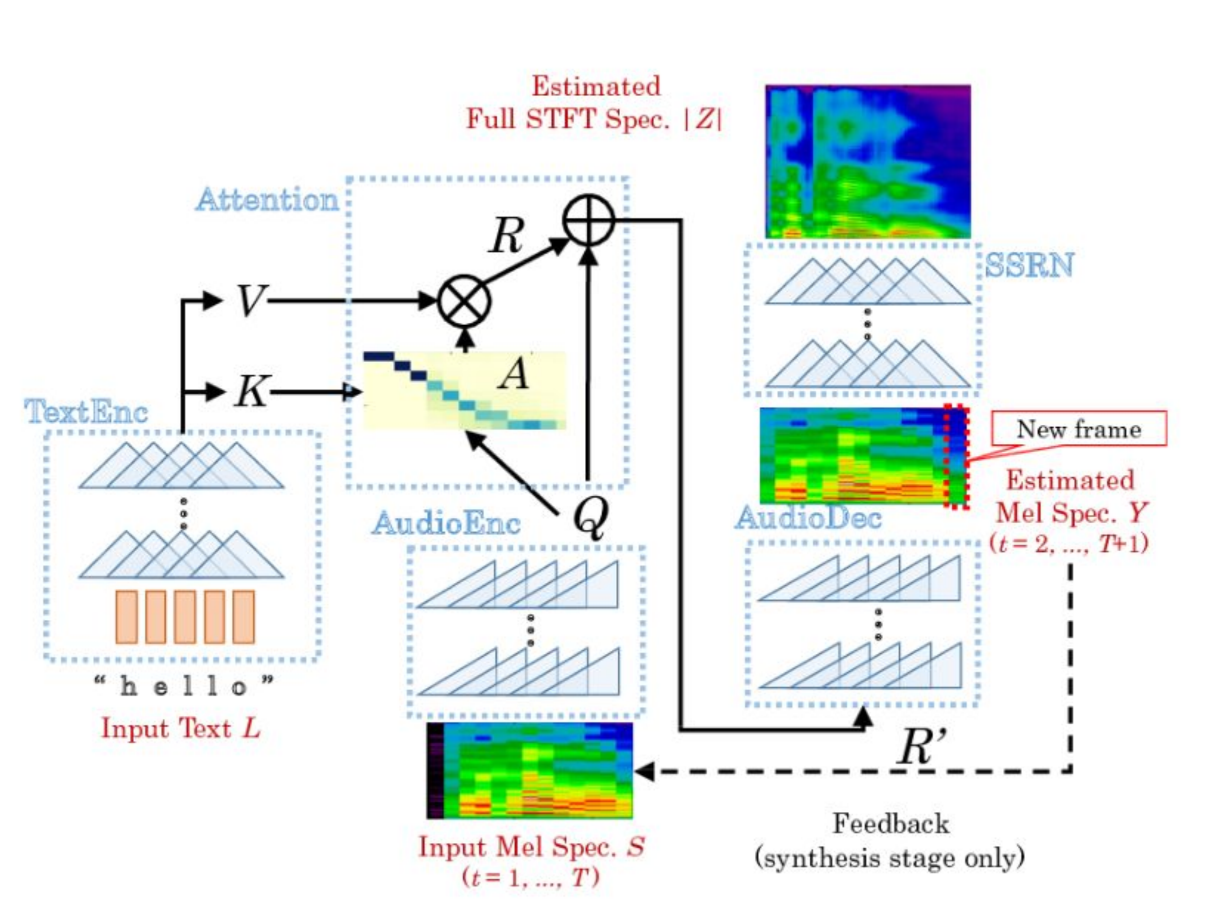
Forensics Face Detection From GANs Using Convolutional Neural Network [2]

History of Text To Speech

"I've been looking forward to black hat all year"

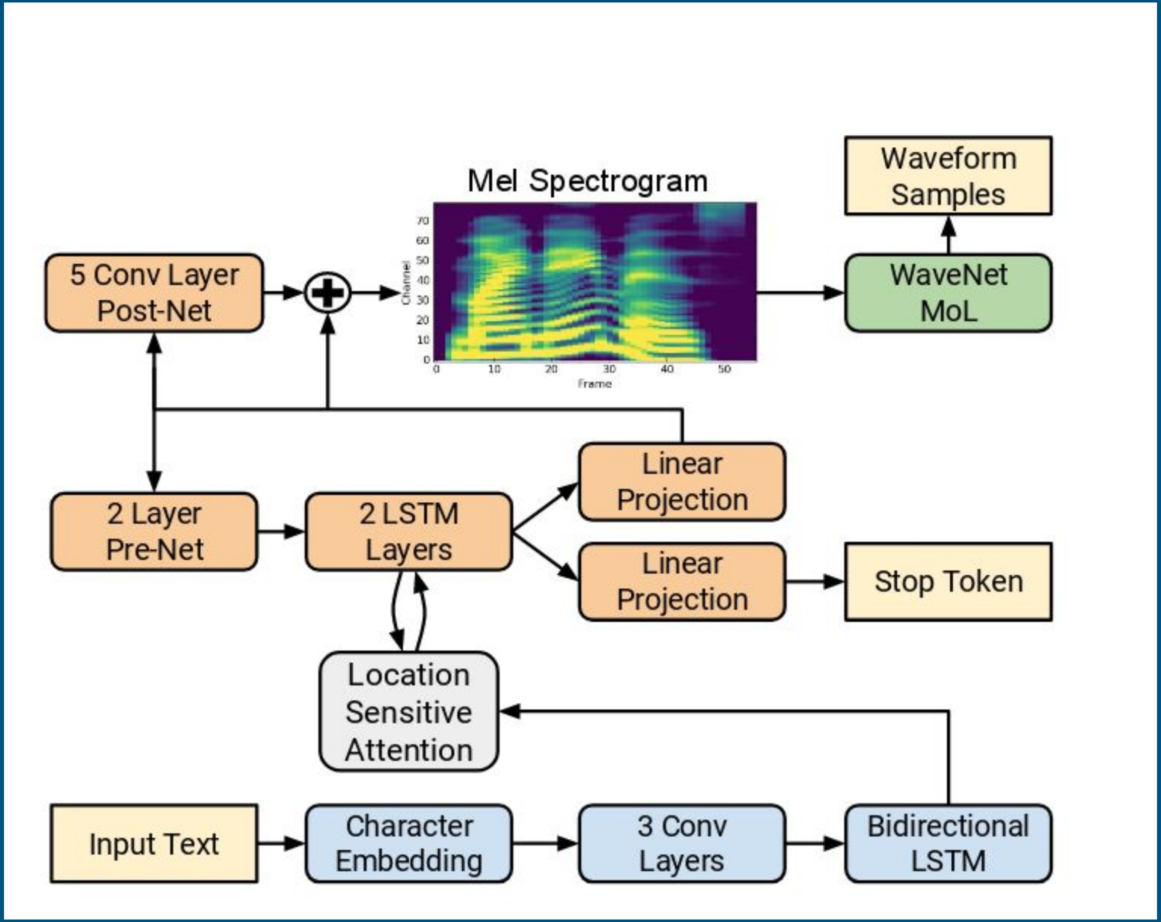
DC-TTS

https://github.com/Kyubyong/dc_tts



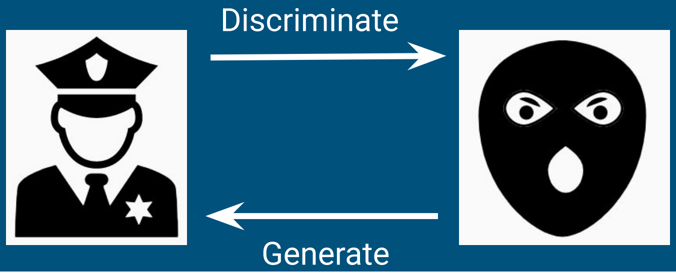
Tacotron2

https://github.com/NVIDIA/tacotron2



Taking advantage of GAN[6] discriminators

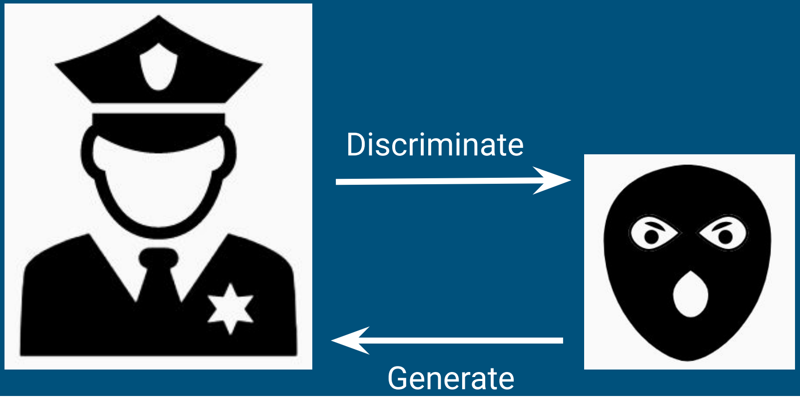
- GANs are Generative Models
- Generative and Discriminative component
  - Creates samples (Audio, Images, Videos)
  - Classifies samples as “real” or “fake”
- Components train by playing a “game” to trick the other
- We want a powerful discriminator
- Train WaveGAN on asv-spoof data
  - Epochs: 5k
  - Parameter combinations: 300



$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

Taking advantage of GAN[6] discriminators

- GANs are Generative Models
- Generative and Discriminative component
  - Creates samples (Audio, Images, Videos)
  - Classifies samples as “real” or “fake”
- Components train by playing a “game” to
- We want a powerful discriminator
- Train WaveGAN on asv-spoof data
  - Epochs: 5k
  - Parameter combinations: 300



$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

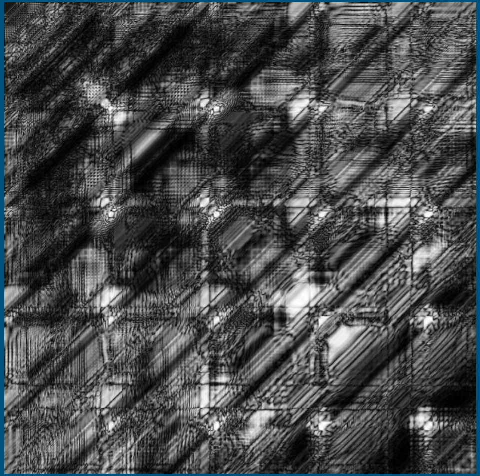
Approach 1: GAN[6] discriminators

- Discriminator is not powerful enough to generalize
- Future directions
  - Train discriminator on non generator samples
  - Richer features
  - Train discriminator separately after convergence

Approach 2: Bispectral Analysis

- Use the bispectrum of the raw audio as the evaluating feature[8]
- Bicoherence (normalized bispectrum) of a signal represents higher-order correlations in the Fourier domain

"There are different cultures in different departments"



Approach 2: Bispectral Analysis

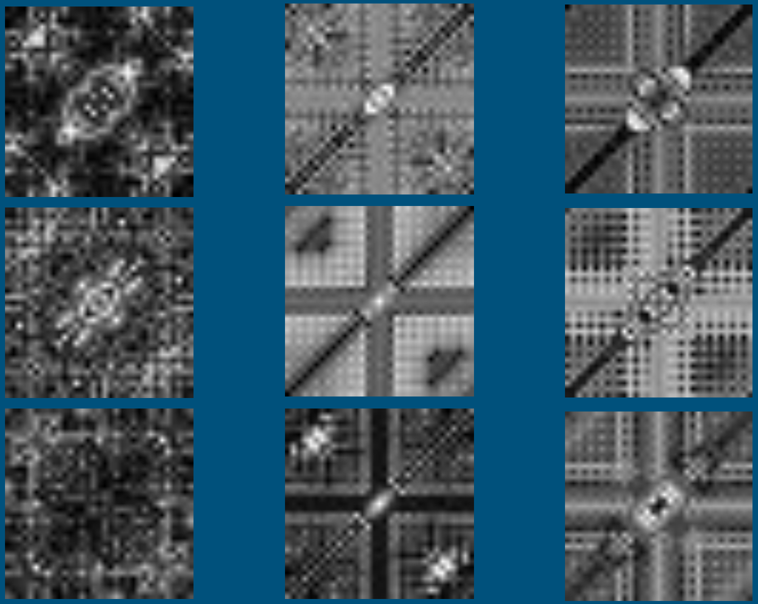
- The averaged bicoherent magnitude across segments of a waveform produces a signature

DC-TTS | Tacotron2 | Human

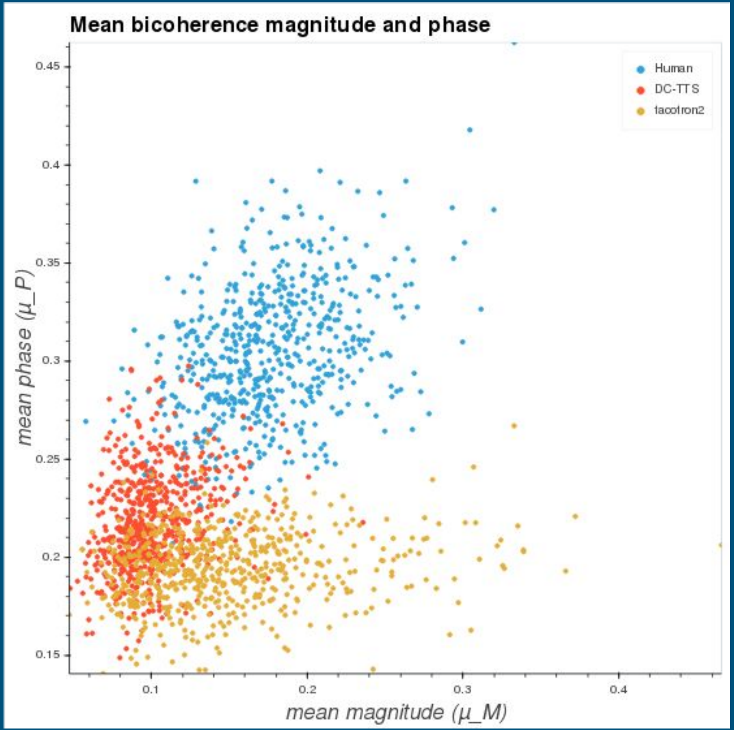
“There are different cultures in different departments.”

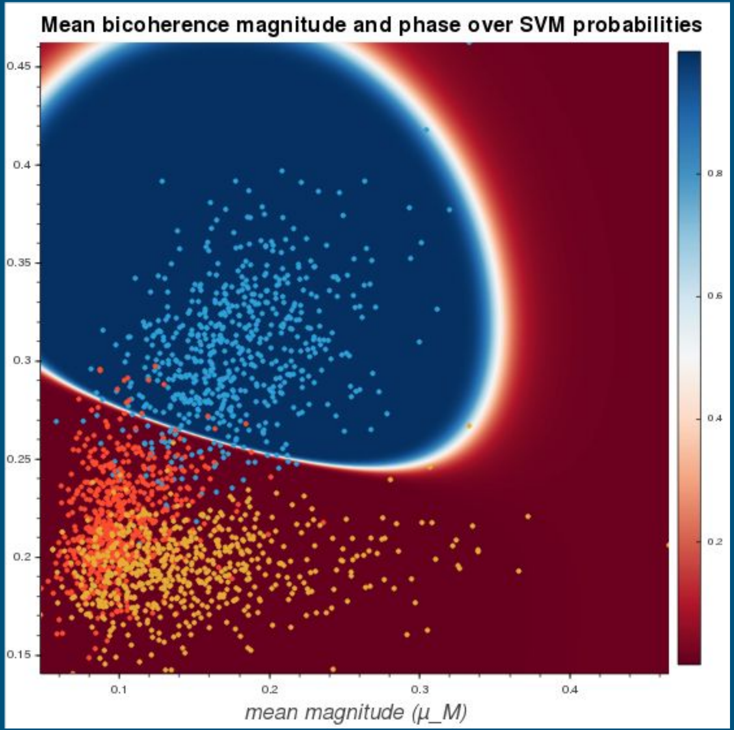
“Don’t you think it was a fine performance.”

“Where do we go from here.”



Approach 2: Bispectral Analysis





Accuracy:	0.95
Precision:	0.95
Recall:	0.94
Samples:	1800

* Samples from LJ speech dataset

References

[1] Suwajanakorn, Supasorn, et al. “Synthesizing Obama.” *ACM Transactions on Graphics*, vol. 36, no. 4, 2017, pp. 1–13., doi:10.1145/3072959.3073640.
[2] Do Nhu, Tai & Na, In & Kim, S.H.. (2018). Forensics Face Detection From GANs Using Convolutional Neural Network.
[3] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, “WaveNet: A Generative Model for Raw Audio” arXiv:1609.03499 [cs], Sep. 2016.
[4] Chris Donahue, Julian McAuley, Miller Puckette, “Adversarial Audio Synthesis" arXiv:1802.04208v3 [cs] Feb. 2019
[5] Shan Yang, Lei Xie, Xiao Chen, Xiao Lou, Xuan Zhu, Dongyan Huang, Haizhou Li, “Statistical Parametric Speech Using Generative Adversarial Networks Under A Multi-Task Learning Framework” arXiv:1707.01670v2 [cs] Jul. 2017
[6] Generative Adversarial Networks “Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio” 1406.2661 [cs] Jun. 2014
[7] Marc Schröder. Interpolating Expessions in Unit Selection. In *Proc. 2nd ACII*, Lisbon, Portugal, 2007
[8] Albadawy, Ehab & Lyu, Siwei & Farid, Hany. (2019). Detecting AI-Synthesized Speech Using Bispectral Analysis.
[9] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. arXiv:1808.07371 2018

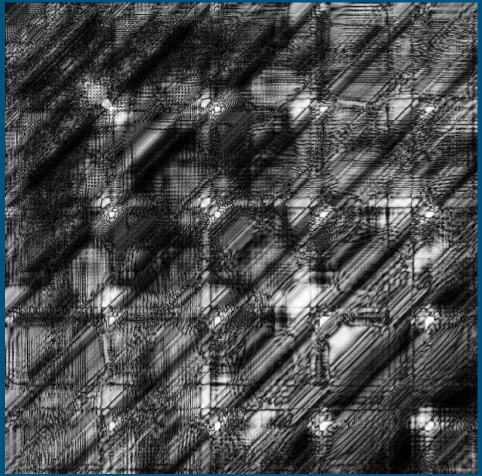
Detecting Deep Fakes: Insights from Biological Neural Nets

Jonathan Saunders, University of Oregon

What kind of deepfake detection do we want?

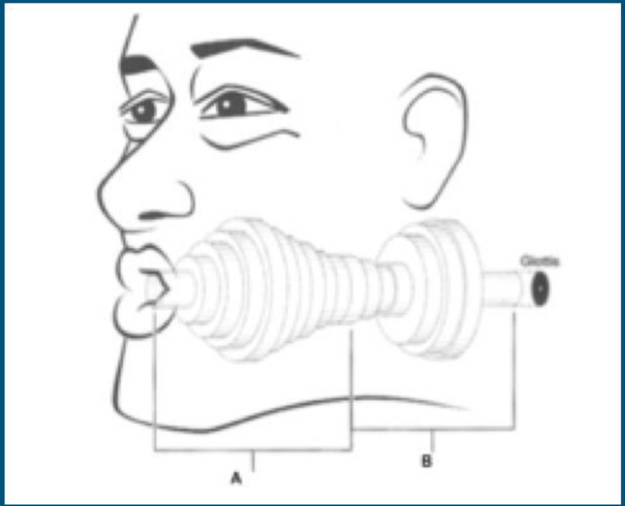
Generation Algorithm Dependent

- Throw data at it
- Always vulnerable to new algorithm
  - Eg. Phase-based detection defeated if complex spectra used in generation



Generation Algorithm Independent

- Requires phonetics & neuroscience
- General solution



ned2 look closer in2 this slimy clarinet

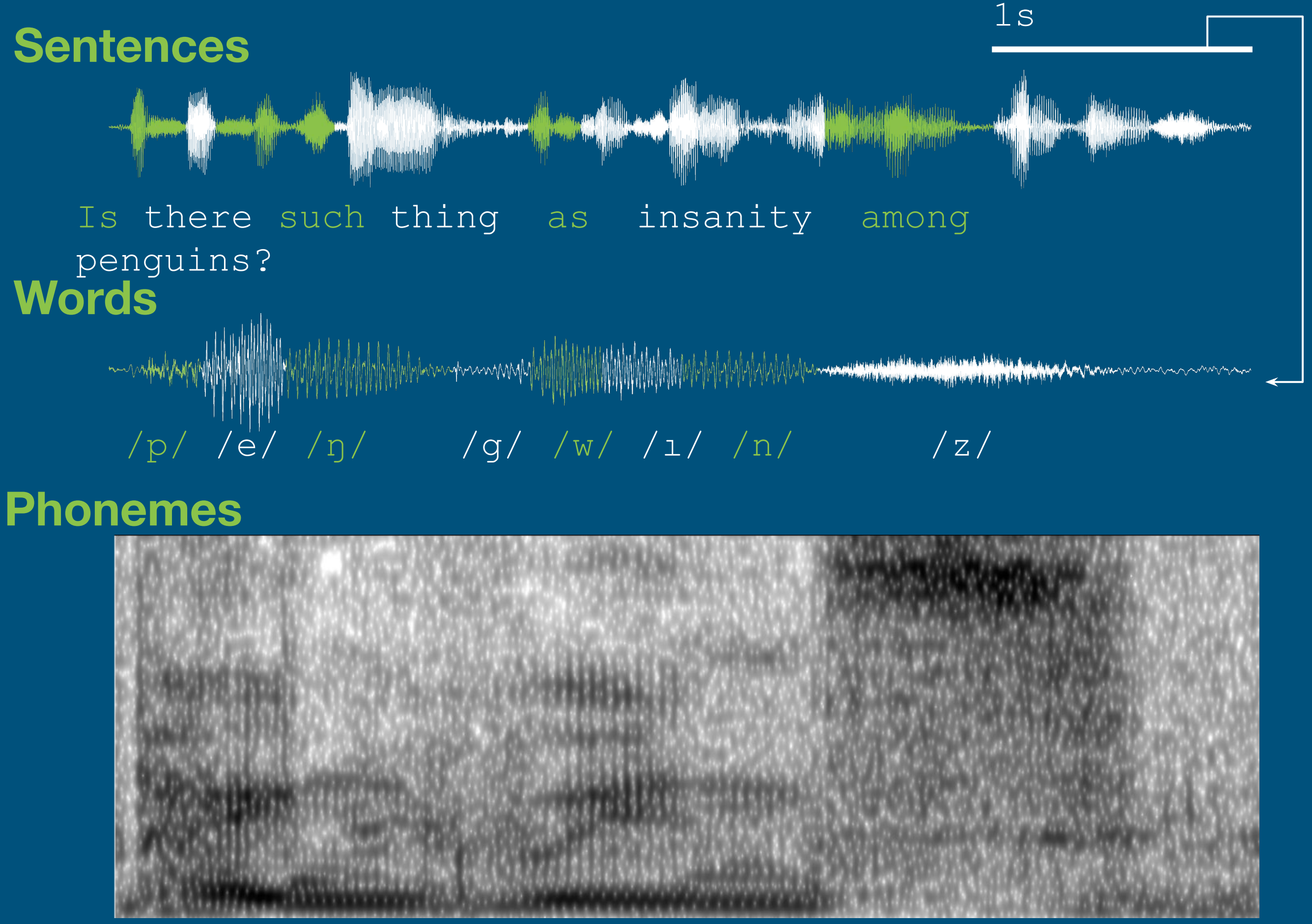
Listening to people talk is hard

Speech is...

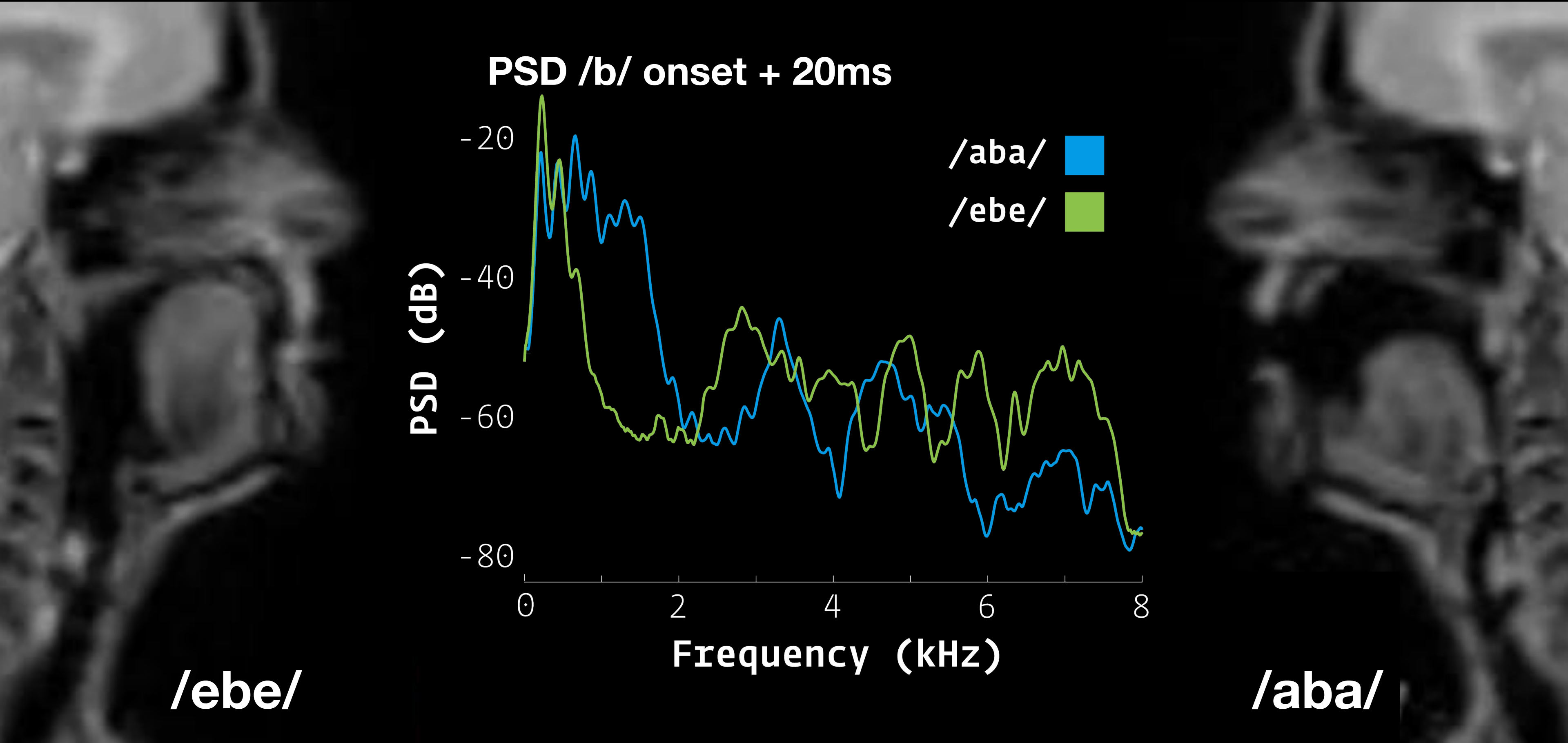
- Hierarchical
- ***Fast:***
  - 10-30 phonemes/s

To detect phonemes, have to normalize...

- Voice Timbre
- Rate
- Prosody
- Accent



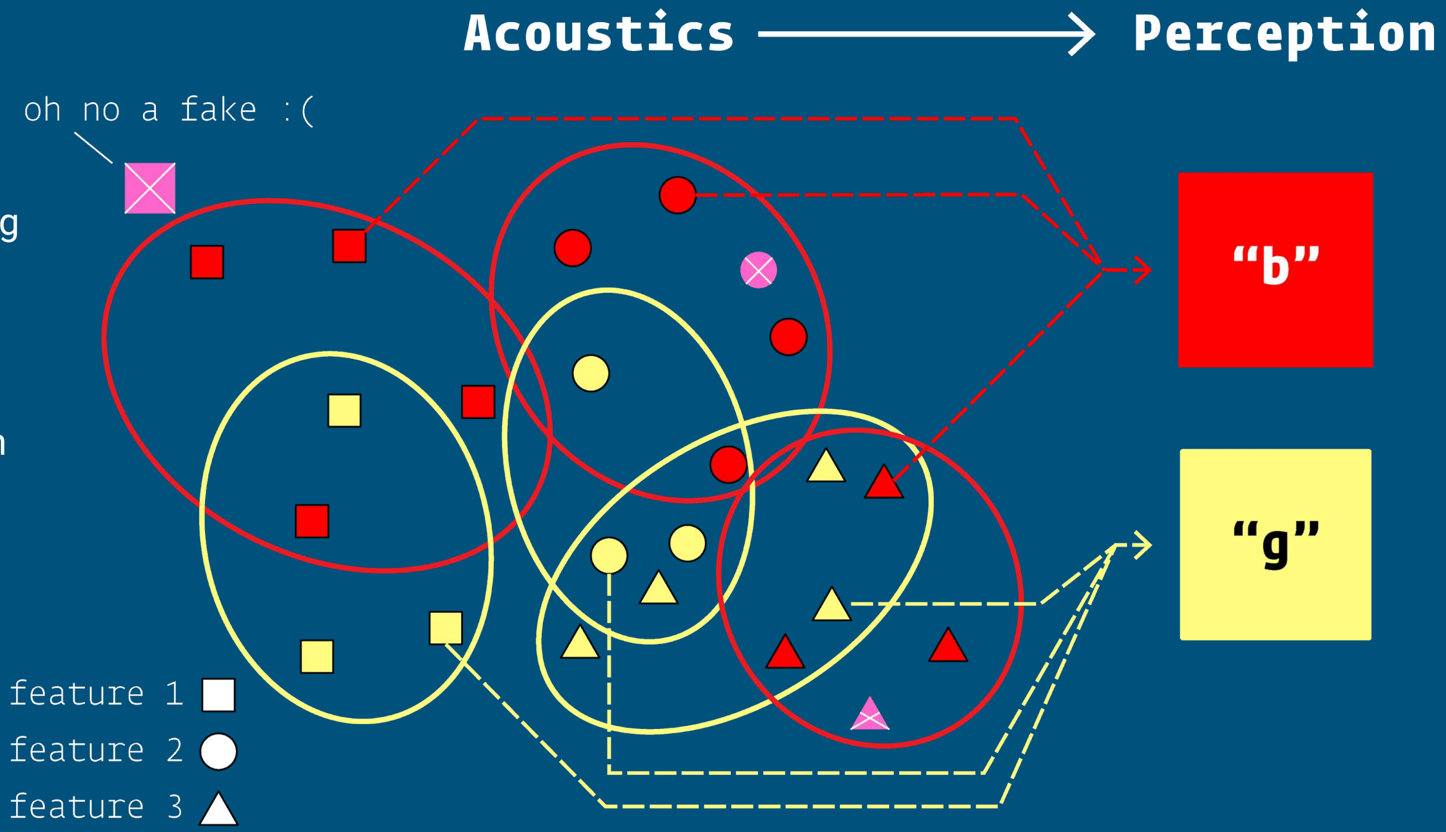
Coarticulation: No unique acoustic structure for phonemes



Lawson EJ, et al. (2015) *Seeing Speech*

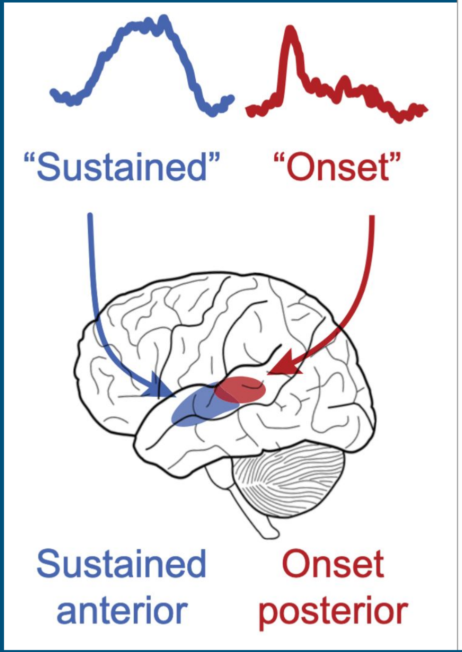
The Auditory System: designed to be gullible

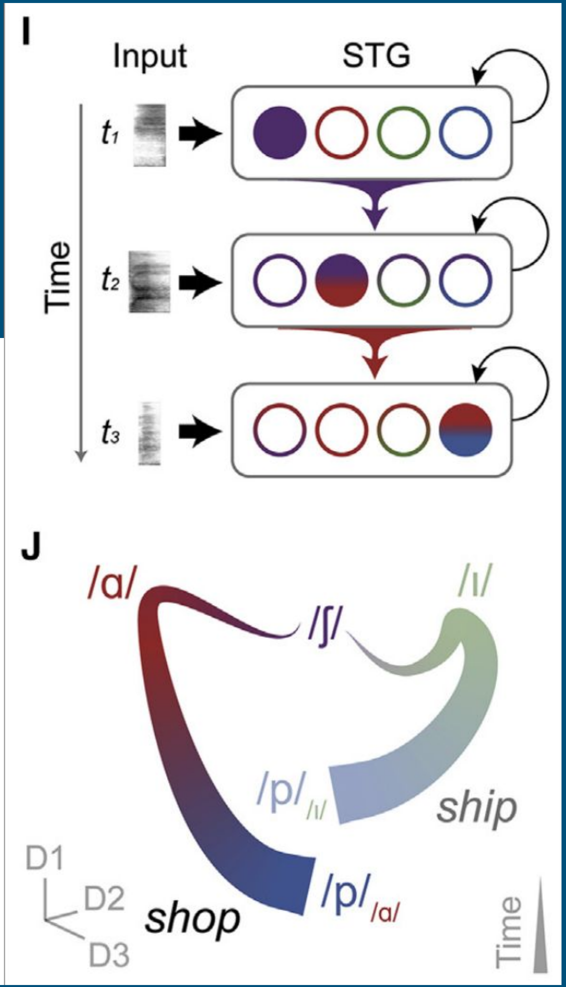
- Complex, overlapping feature space
- Collapse redundant/irrelevant acoustic information
- Bad fakes fool the auditory system



How does the brain do it?

- Phrase onsets signalled by posterior auditory cortex
- Recurrent anterior cortical networks compare past to present





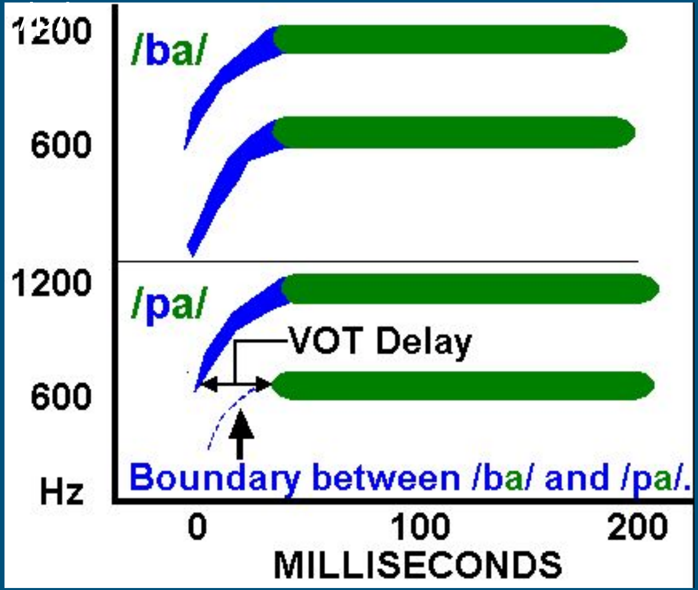
The rest is all theory :(

Hamilton LS, Edwards E, Chang EF (2018) *Curr.Bio.* 28:1860-1871
Yi HG, Leonard MK, Chang EF (2019) *Neuron* 102(6):1096-1110

Can’t crack the speech circuit in humans

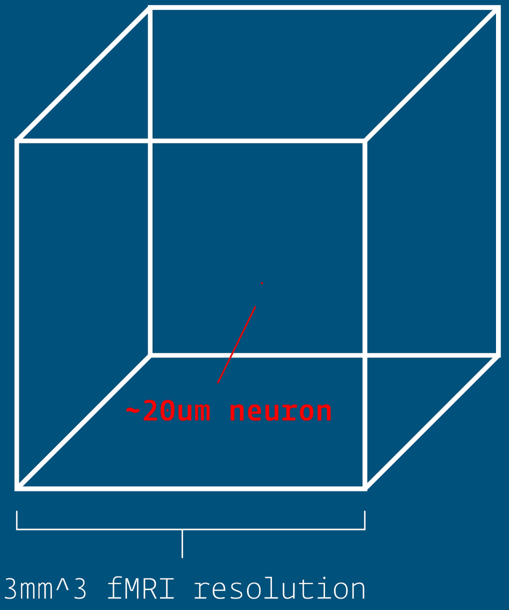
Speech is too fast

~20ms of sound distinguishes /b/ from



Neurons are too small

~630k neurons in an fMRI voxel



http://www.indiana.edu/~p1013447/dictionary/vot.htm

Can't study phonetic processing in humans?

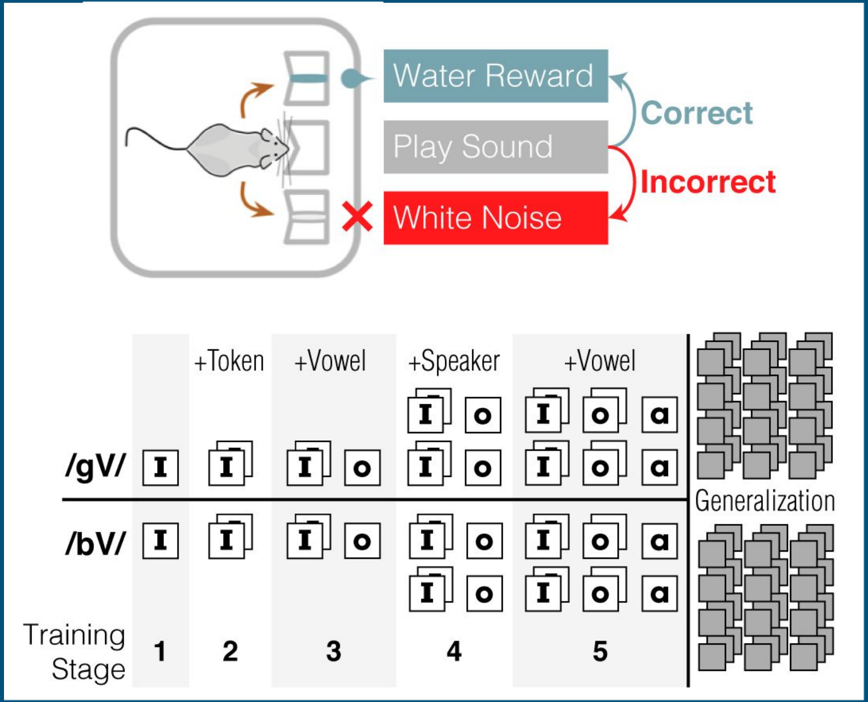
Teach mice English (phonemes)

To discriminate /bV/ vs. /gV/ consonant-vowel pairs...

1. Center poke to play sound
2. Go left if /g/, right if /b/
3. Get that water or face the consequences

5 training stages add speakers + vowels

Onto a generalization stage w/ 180 recordings

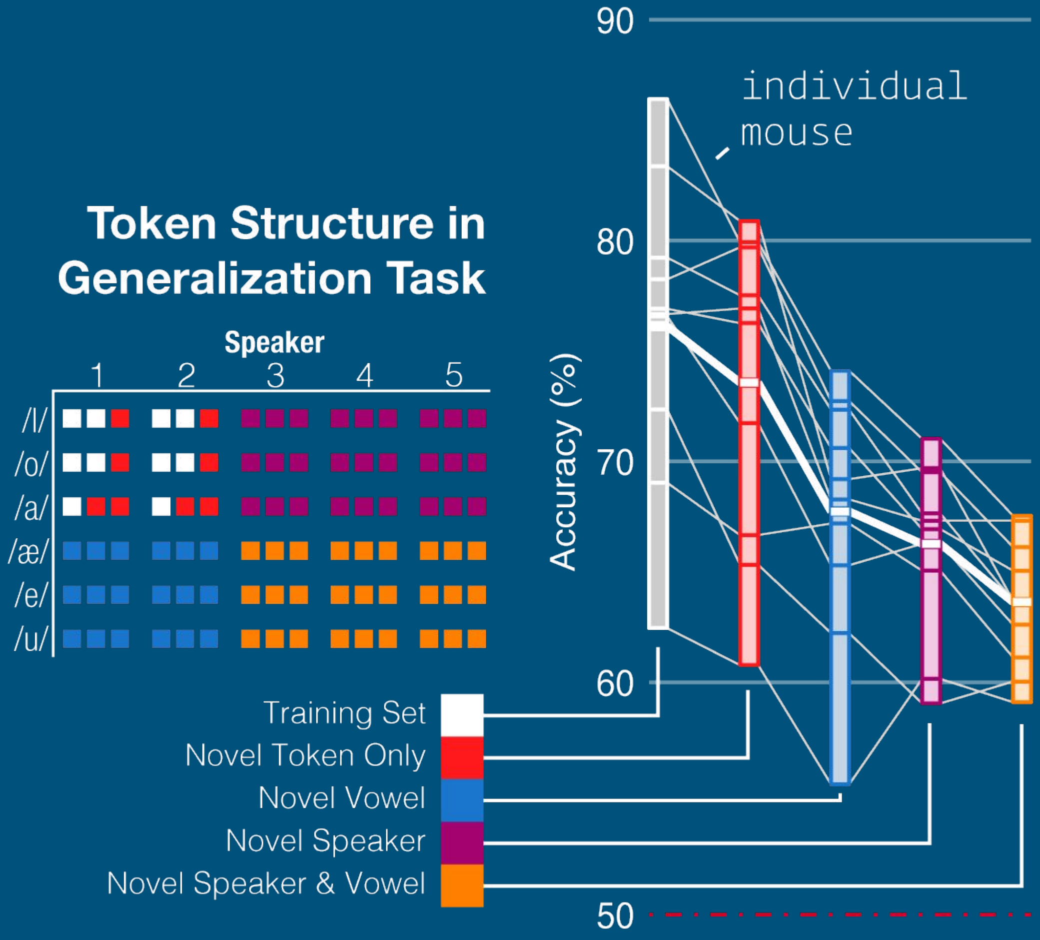


Saunders JL, Wehr M (2019) *J. Acoust. Soc. Am* 145:1168

PSA: We are releasing the next hot shit in behavior hardware/software this summer: **git.io/rpilot**

Generalization Performance

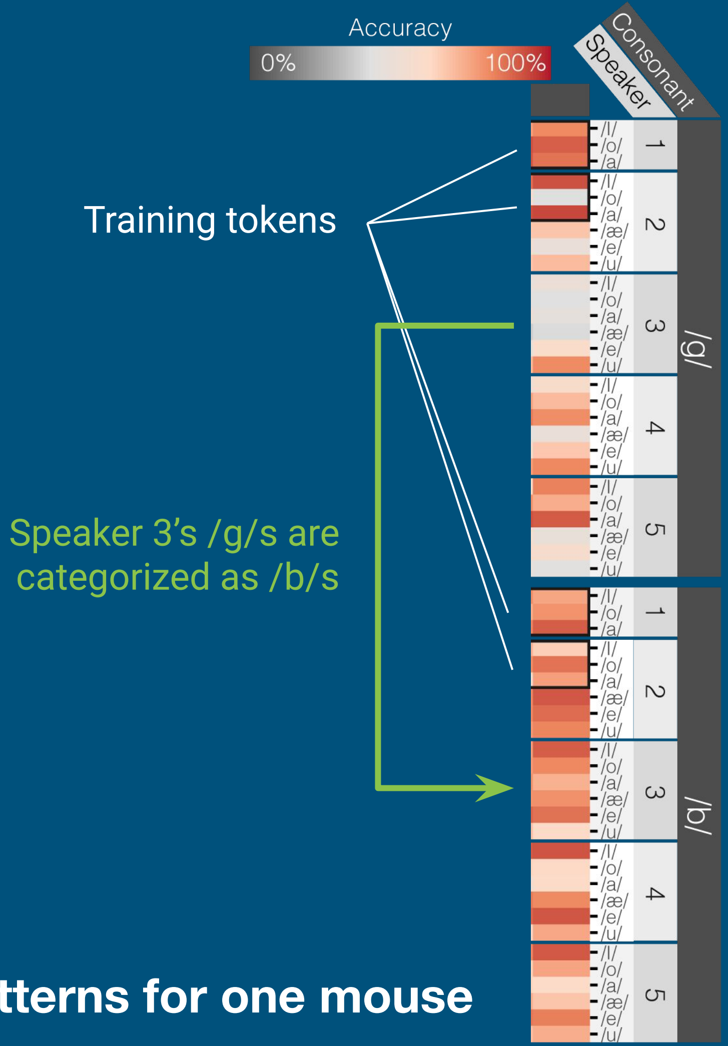
- Mice learn generalizable consonant categories
- Performance decreases with dissimilarity to training set
- Generalization deficit similar across mice



Saunders JL, Wehr M (2019) *J. Acoust. Soc. Am* 145:1168

Nonuniform Error Patterns

- Each mouse has a complex discrimination boundary



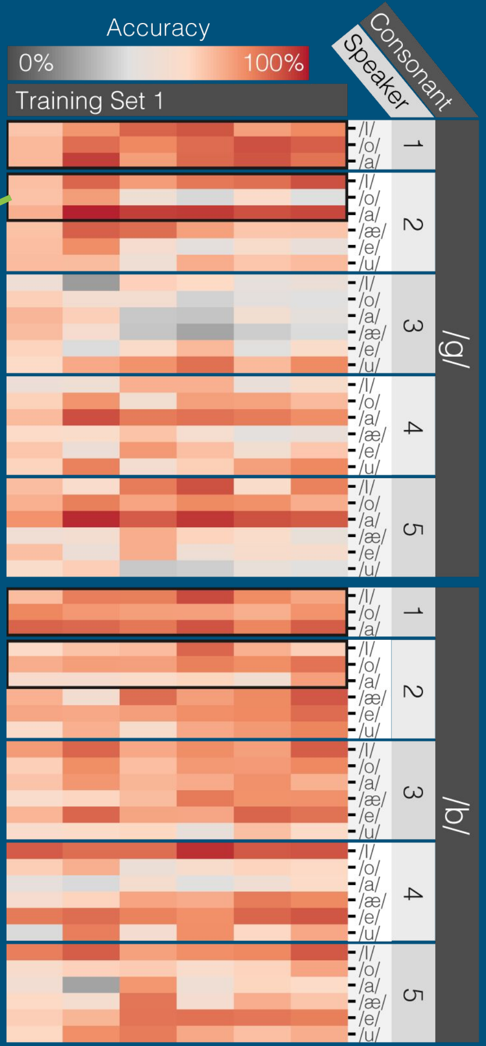
Error Patterns for one mouse

Saunders JL, Wehr M (2019) *J. Acoust. Soc. Am* 145:1168

Nonuniform Error Patterns

- Each mouse has a complex discrimination boundary
- But general patterns are preserved across mice

A hard token in the training set



Saunders JL, Wehr M (2019) *J. Acoust. Soc. Am* 145:1168

Nonuniform Error Patterns

- Each mouse has a complex discrimination boundary
- But general patterns are preserved across mice

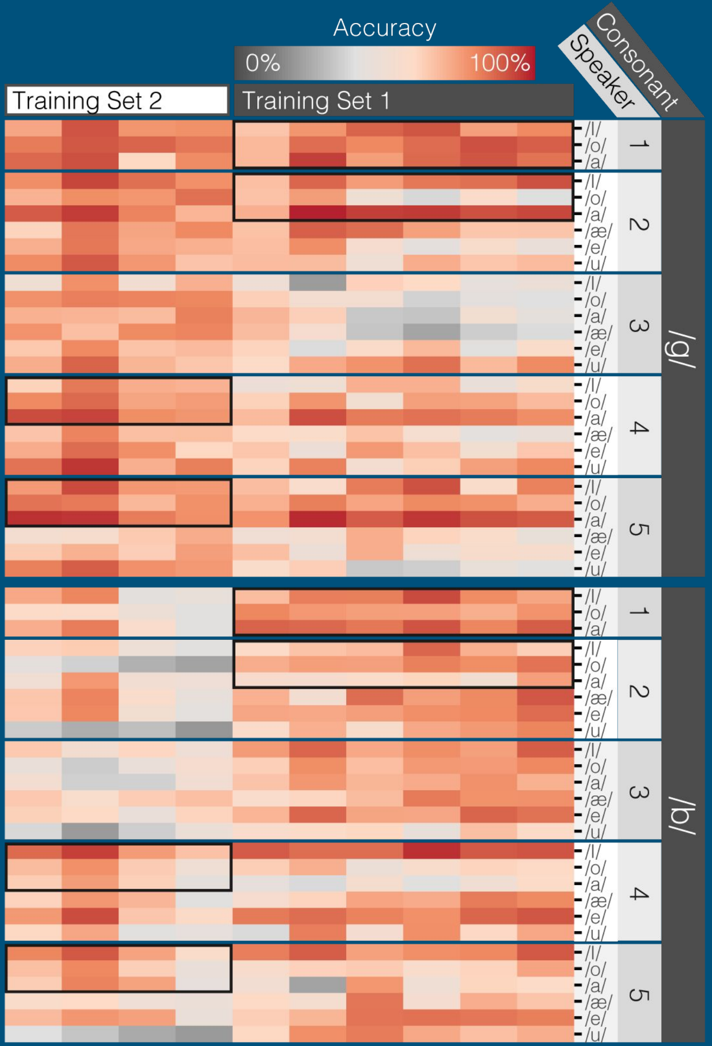
When we trained on a different set of tokens…

- Wholly different error pattern
- Biases are mostly from training, not stimuli

Mice learn a complex acoustic representation of consonants

Group 2 training tokens

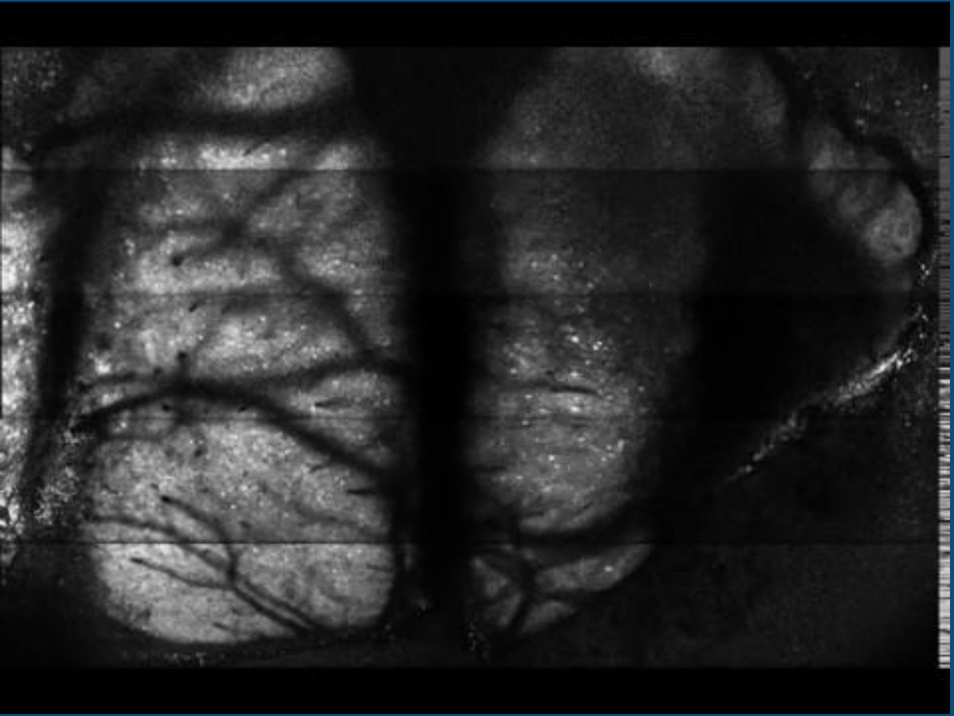
Global bias towards /g/



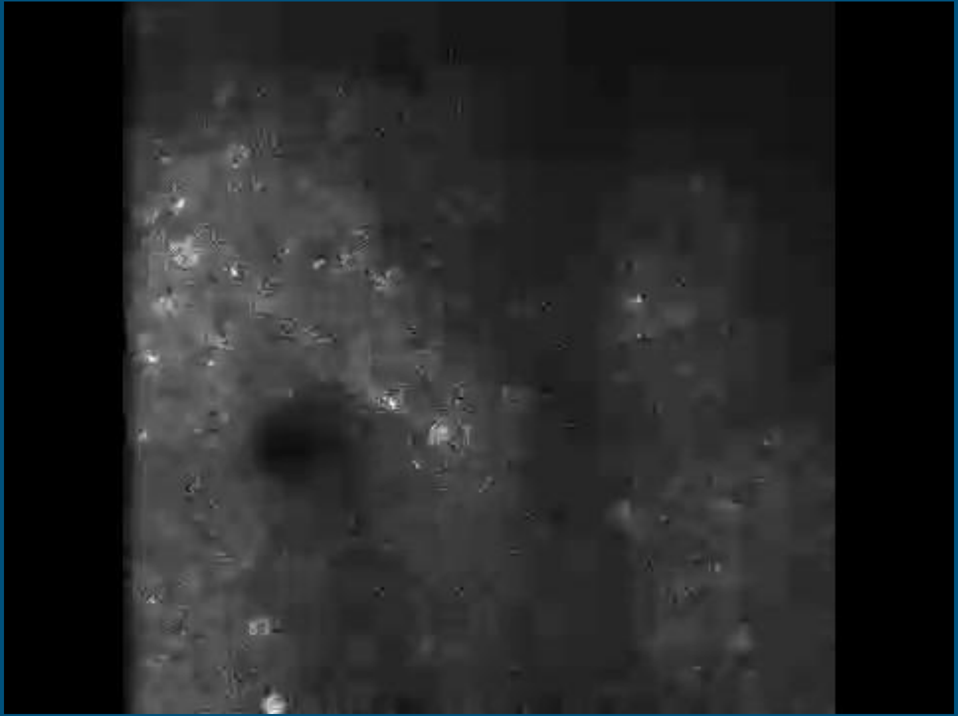
Saunders JL, Wehr M (2019) *J. Acoust. Soc. Am* 145:1168

This Fall: record entire surface of auditory cortex during learning & testing

Example data from Evan Vickers, UOregon, pers. comm.



Dorsal surface,
4.5mm x 3mm, 0.3Hz(10x)

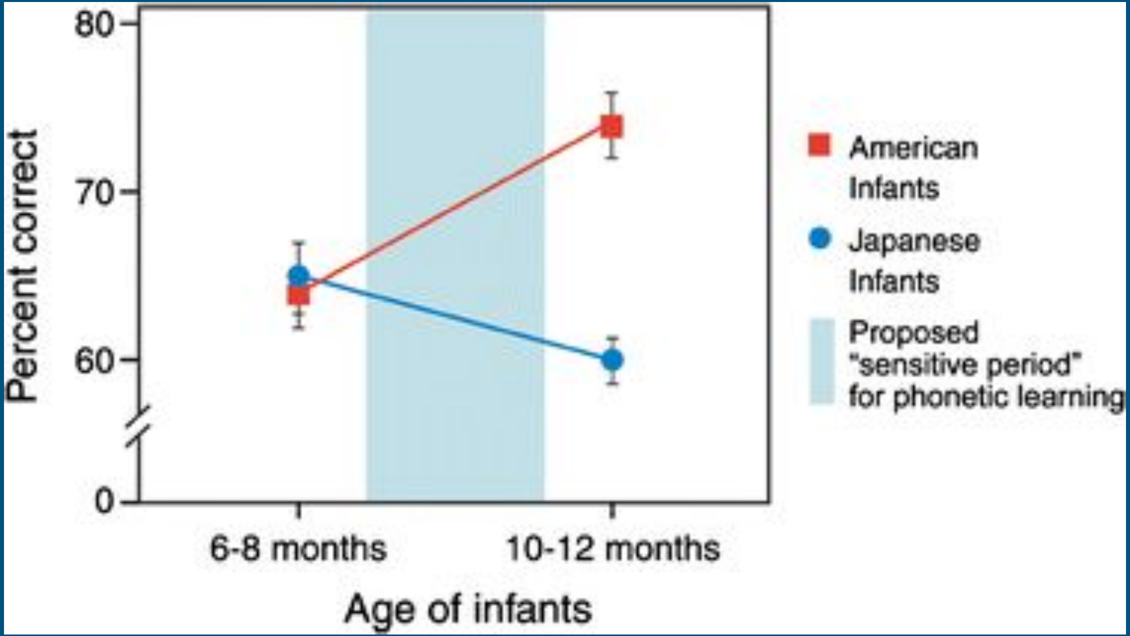


Primary Auditory Cortex,
230um depth, 500um^2 area, 7Hz(5x)

Now: Are category representations plastic?

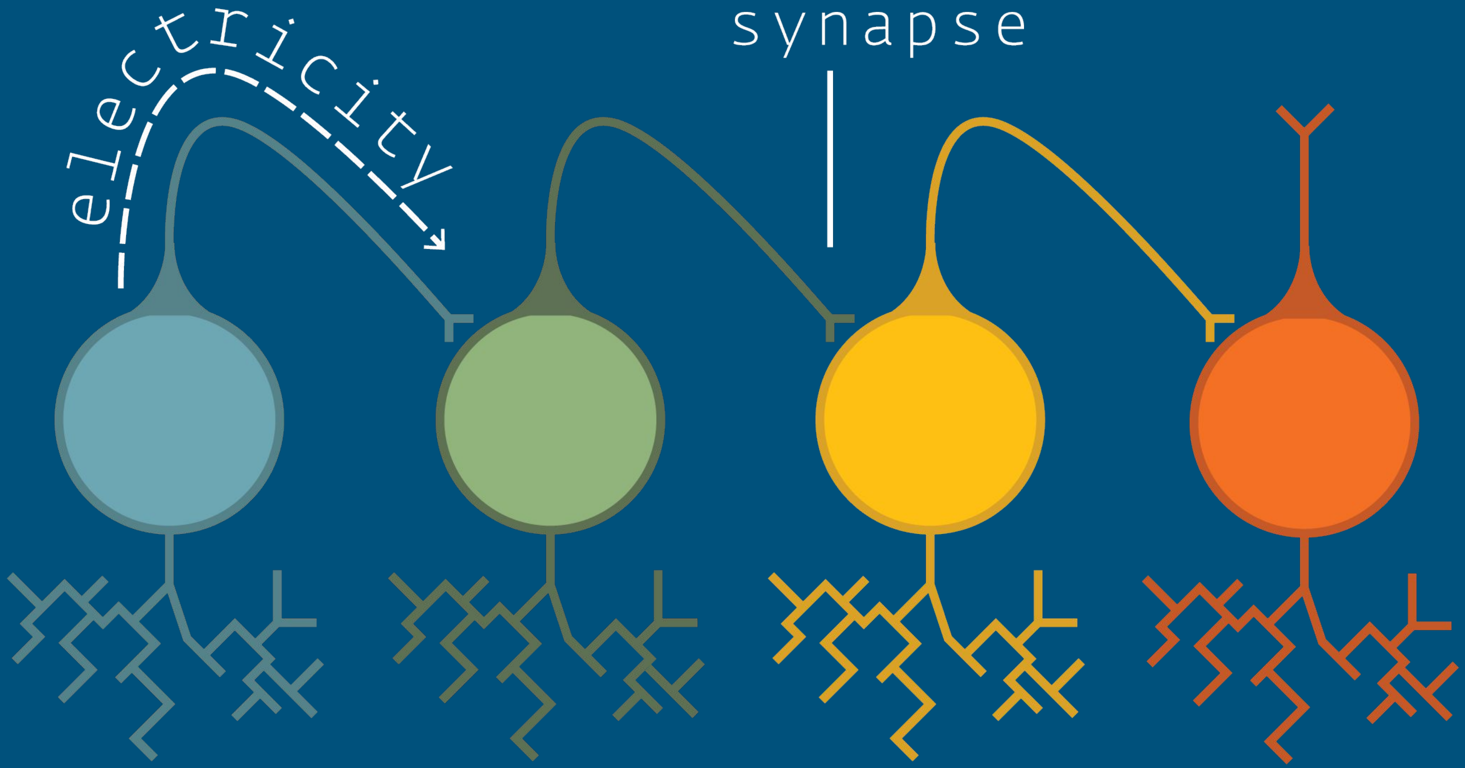
- Some of our mice failed to learn, but why?
- Humans can't hear some phonetic contrasts that aren't in their language

/ra/-/la/ discrimination

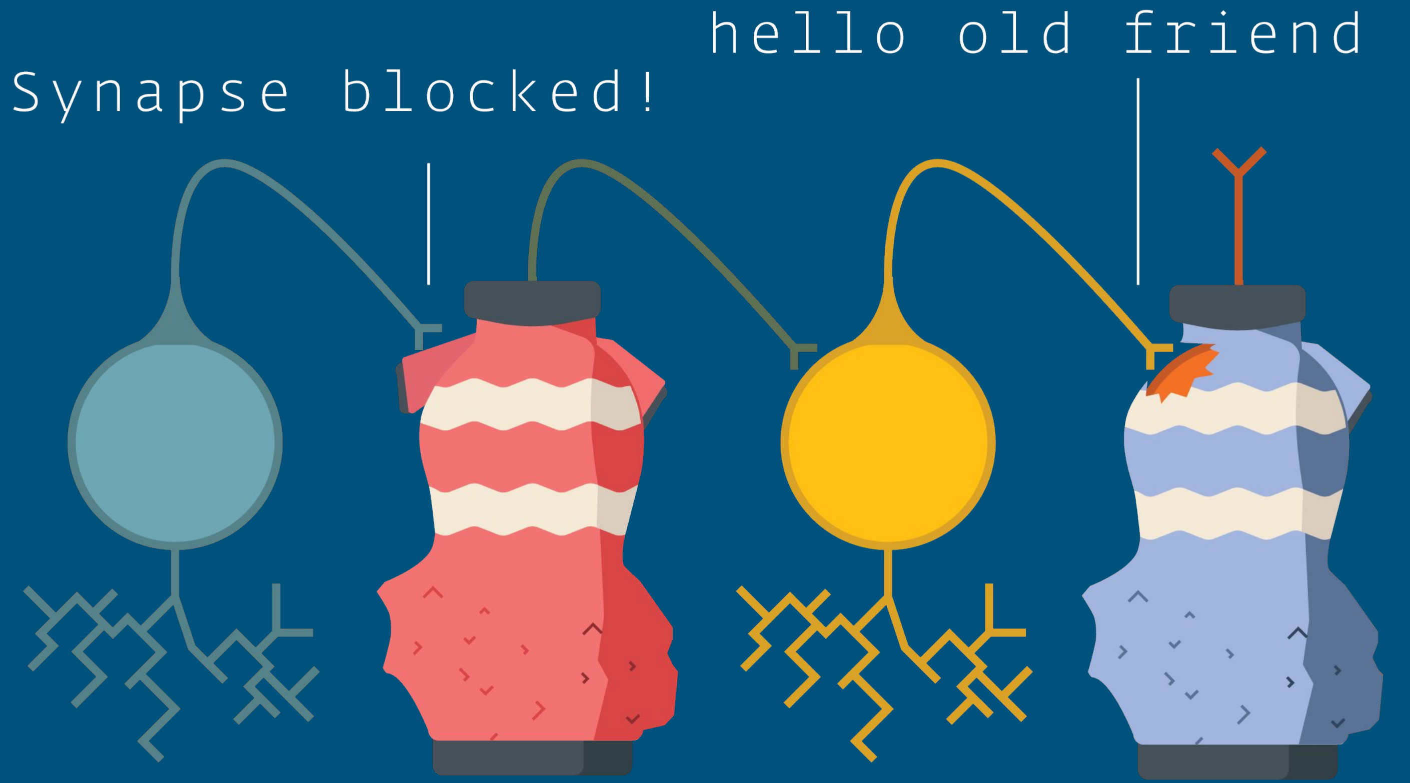


Kuhl, P. K., et al. (2006). *Dev. Sci. 9(2)*

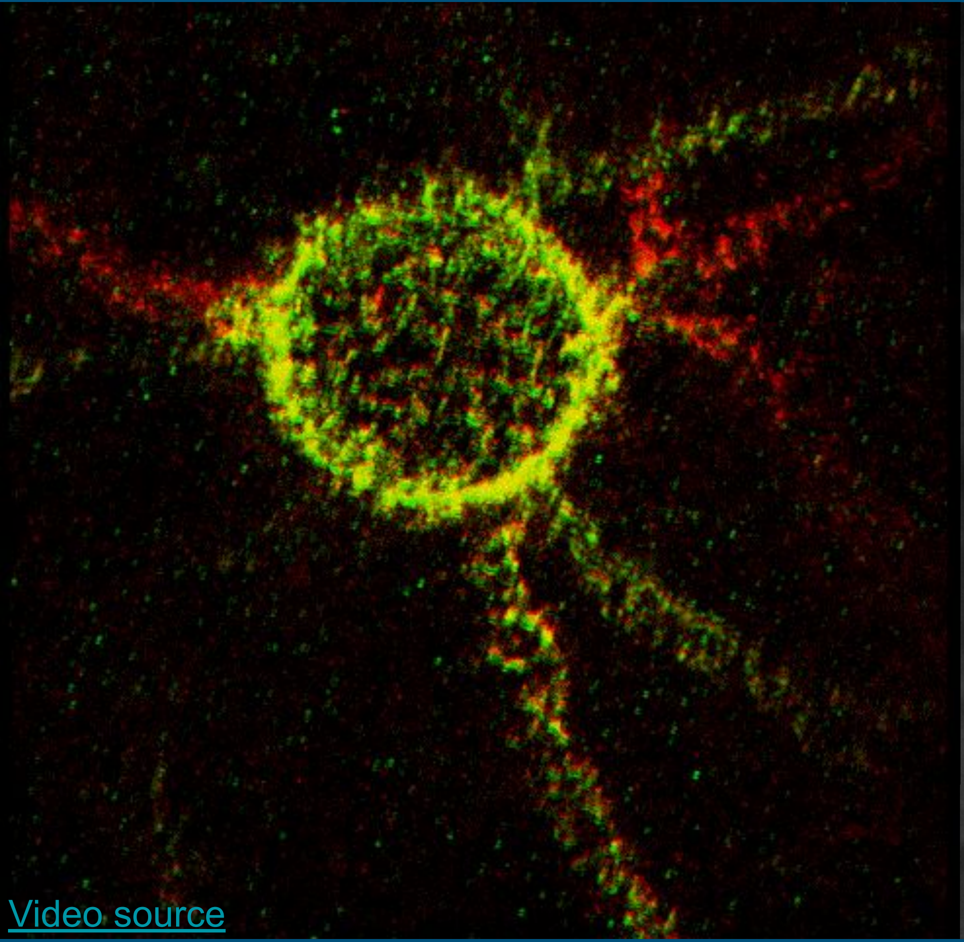
Information is stored in the synapses...



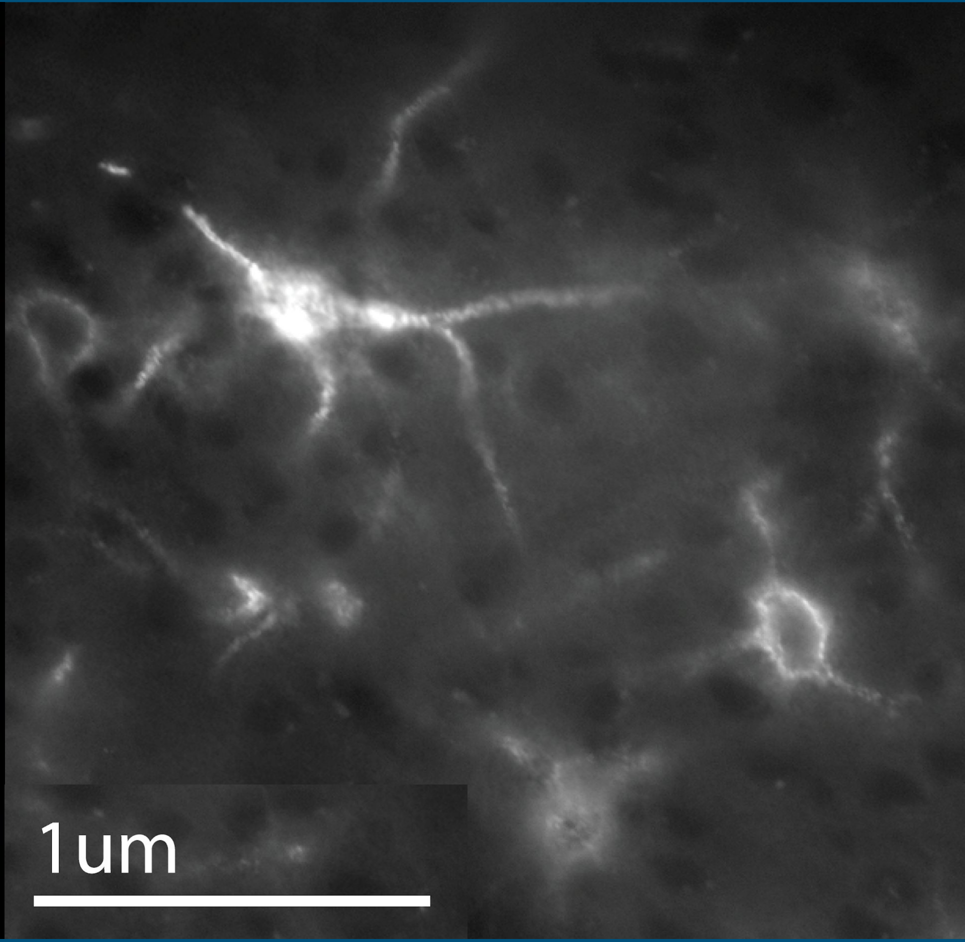
Synaptic patterns are stabilized by extracellular proteins - perineuronal nets



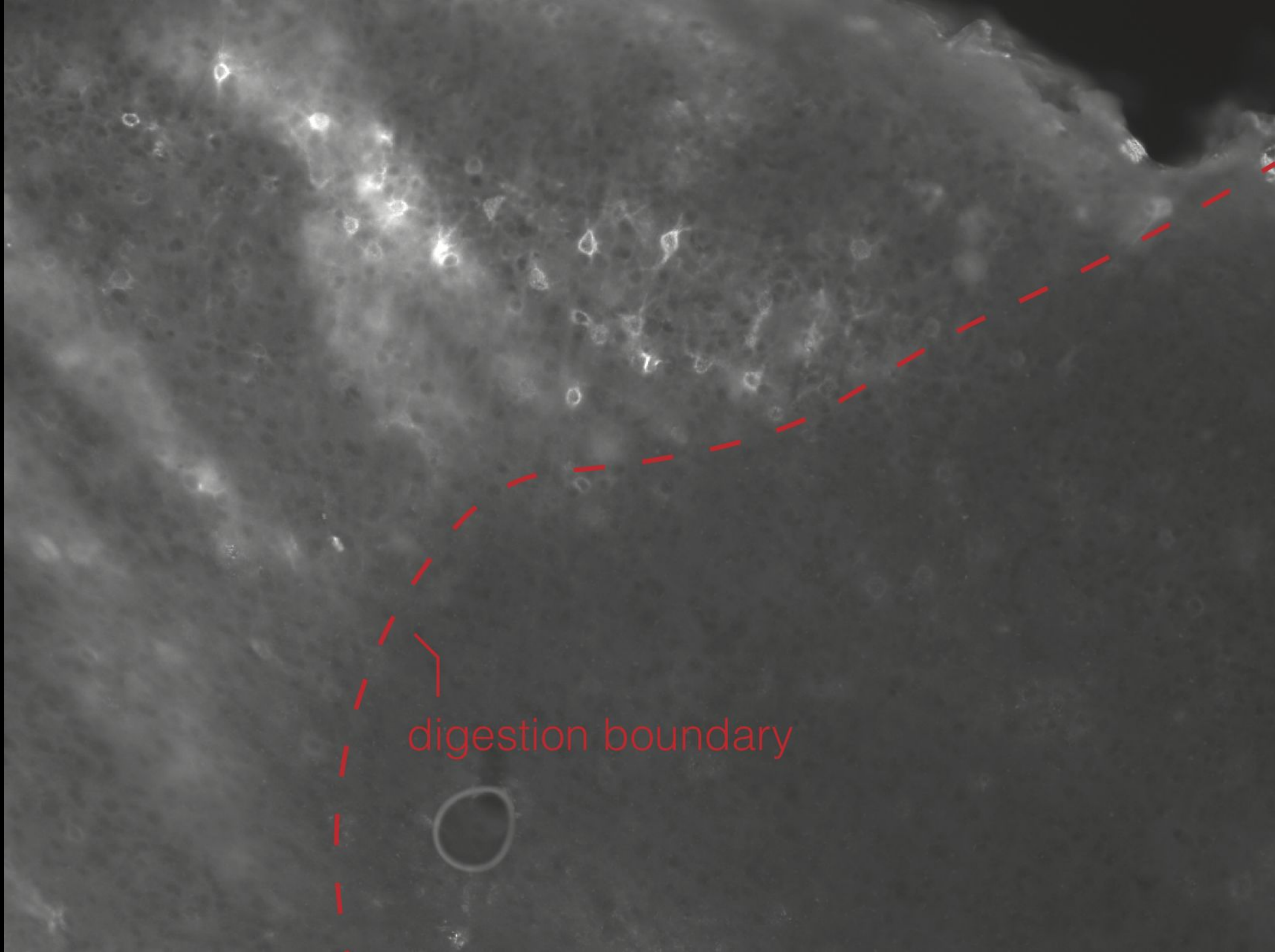
A subset of neurons wear perineuronal nets



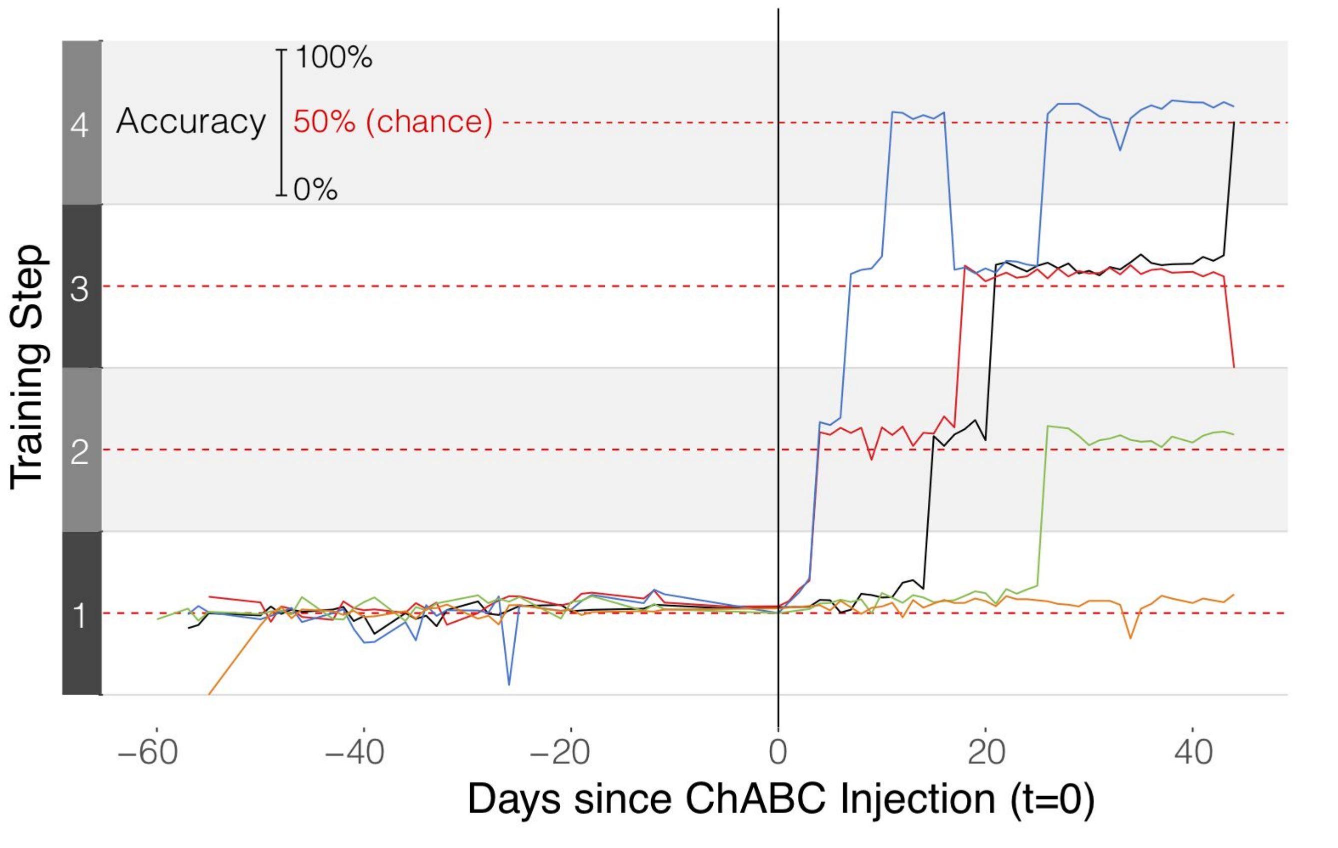
[Video source]



What if we destroy them and let them grow back?



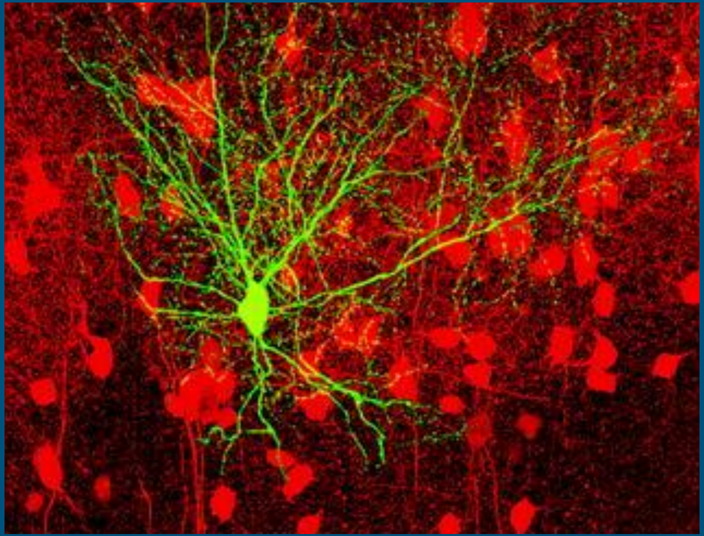
PNN Digestion reopens speech learning

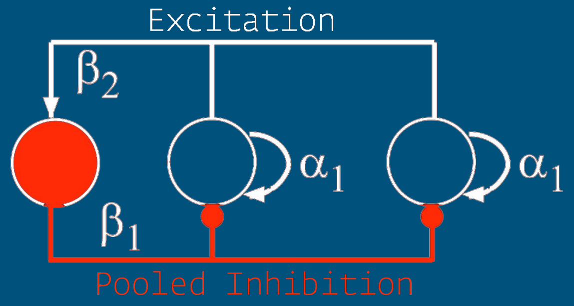


*warning, unpublished pilot experiment, replication in progress

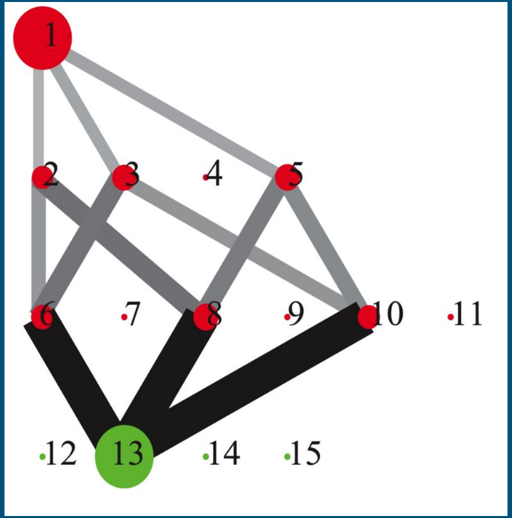
ANNs Need Inhibition

- PNNs are worn by neurons with strong local inhibition
- Local Inhibitory neurons
  - Integrate recent past to steer recurrent computation
  - Store long-term auditory percepts (?)





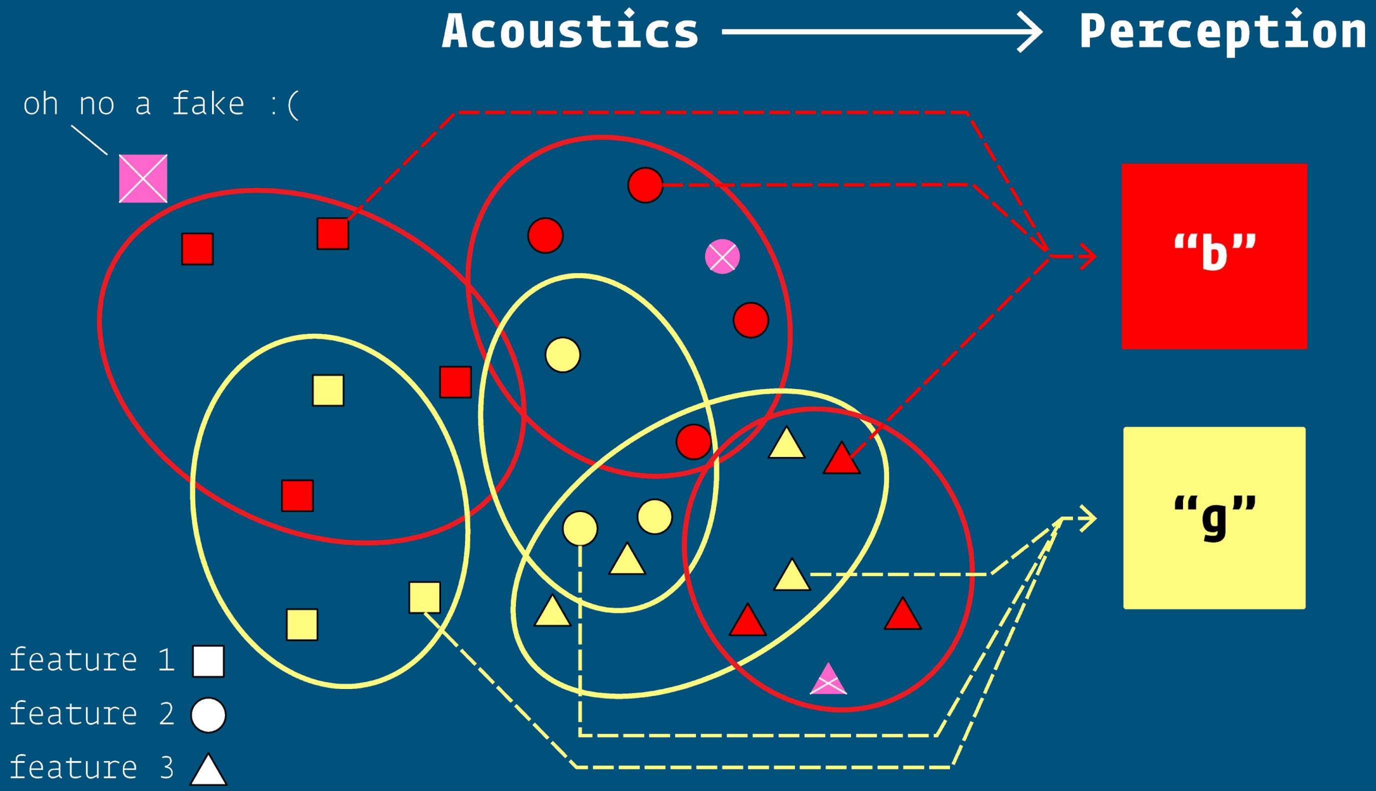
Inhibition 'forbids' some state transitions to steer computation



Staiger JF, et al (2009) *Brain Structure and Function* 214:1
Rutishauser U, et al (2015) *PLOS Comp. Biol.* 11(1): e1004039

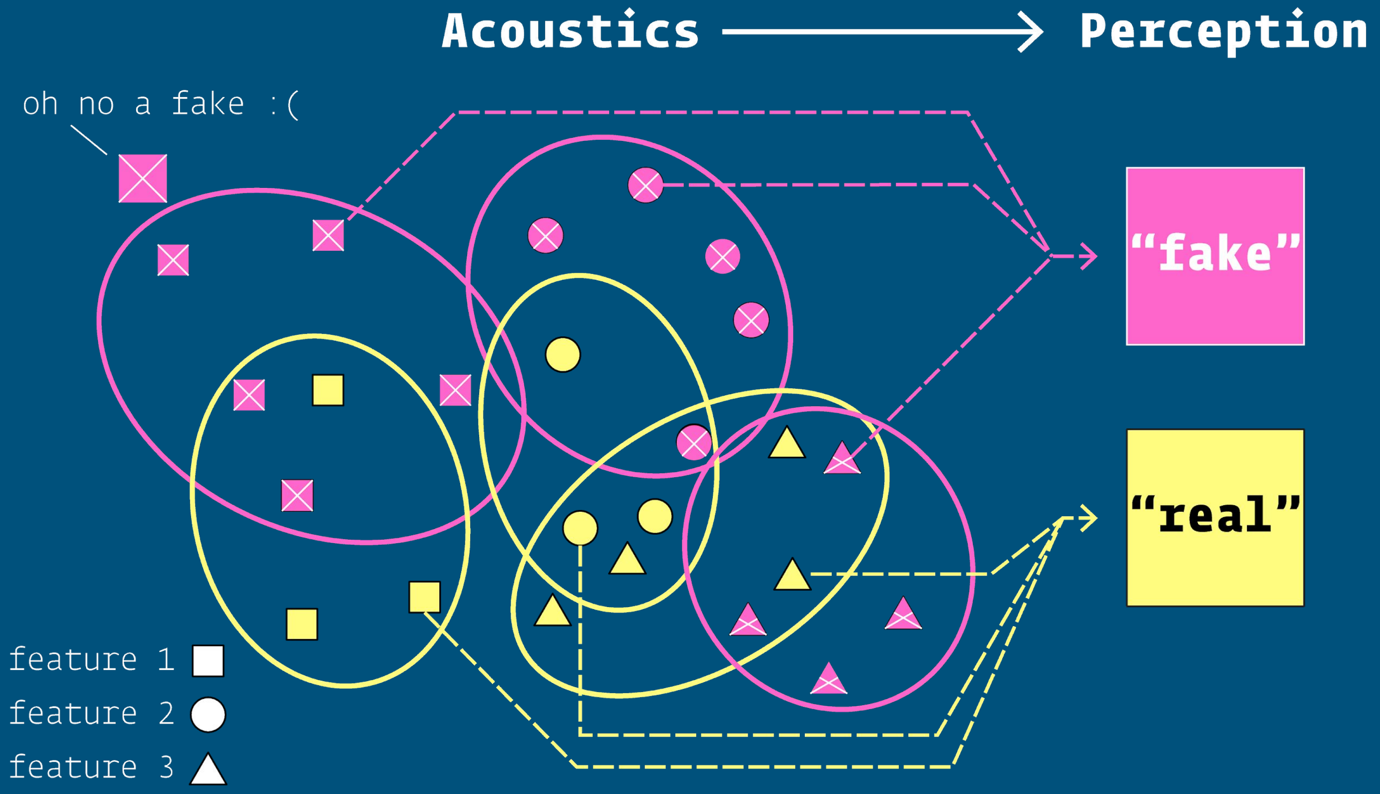
Detecting deep fakes like the brain?

Training mice to detect fakes could inform better detection algorithms



Detecting deep fakes like the brain?

Training mice to detect fakes could inform better detection algorithms



Thank you, BlackHat !

Jonathan Saunders, University of Oregon

- Email: jsaunder@uoregon.edu

George Williams, GSI Technology

- Twitter: @cgeorgewilliams

Alex Comerford, Data Scientist

- Github: @cmrfrd

Participate in our Deep Fake Study at: **https://blackhat.deepfakequiz.com**