LINGUIST 592B, FALL 2021. CLASS 1.1

# COURSE INTRO

YU/HUGHES

SEPTEMBER 2, 2021

UNIVERSITY OF MASSACHUSETTS AMHERST 1863

UMASS AMHERST

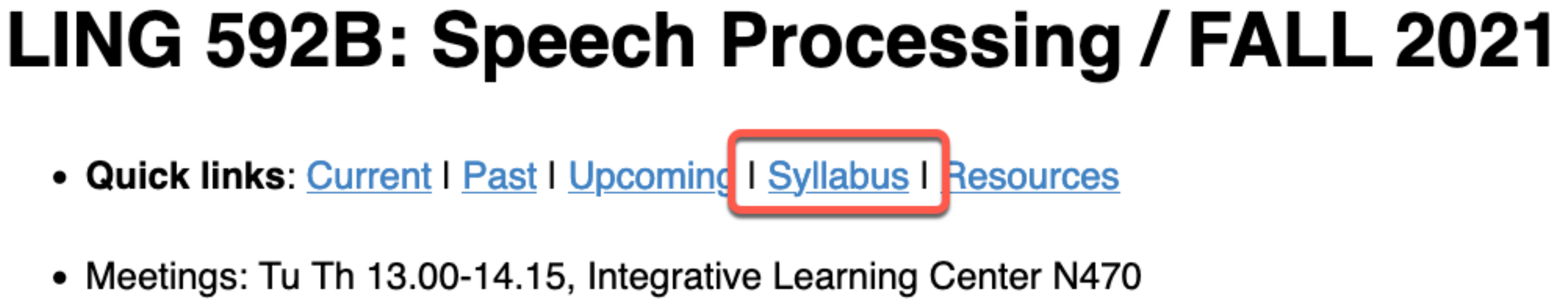
# COURSE ORIENTATION

# COURSE WEBPAGE

**ling592b.krisyu.org**

# SYLLABUS

## ling592b.krisyu.org

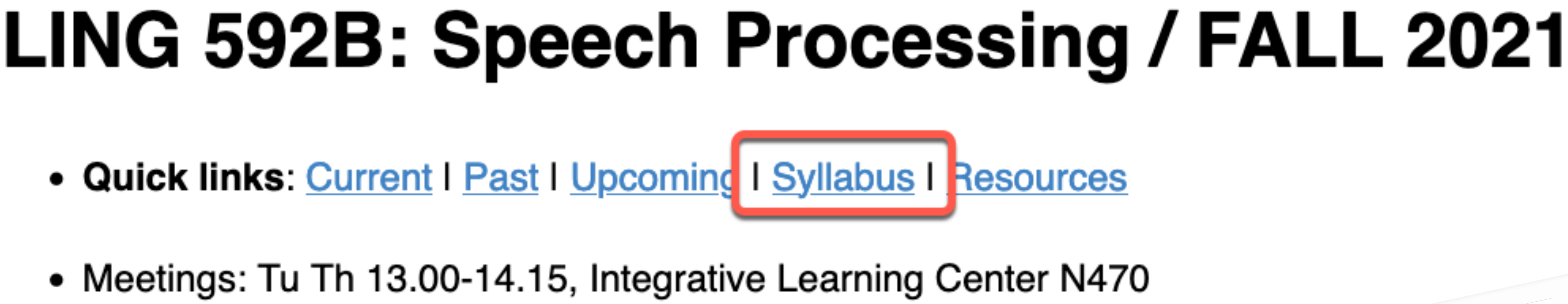


# LING 592B: Speech Processing / FALL 2021

- **Quick links:** Current | Past | Upcoming | Syllabus | Resources
- Meetings: Tu Th 13.00-14.15, Integrative Learning Center N470

# SYLLABUS

ling592b.krisyu.org



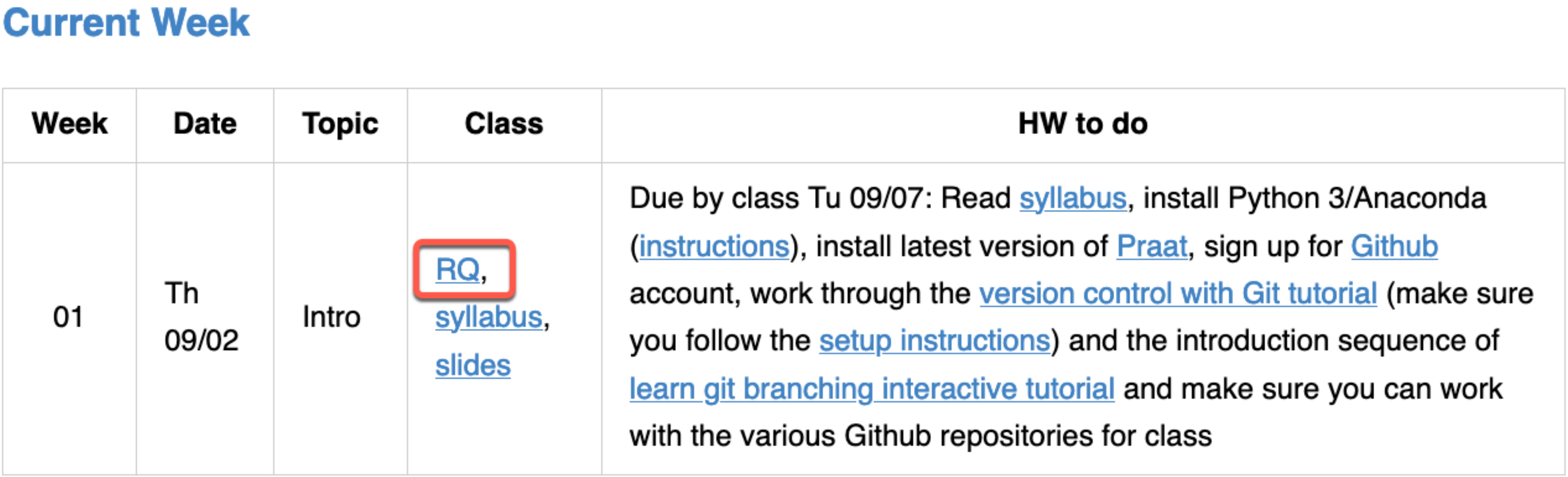
## LING 592B: Speech Processing / FALL 2021

- **Quick links:** Current | Past | Upcoming | Syllabus | Resources
- Meetings: Tu Th 13.00-14.15, Integrative Learning Center N470

HW: read this!

# REVIEW QUESTIONS

- Each class will begin with review questions
- Linked as “RQ” from course webpage for each class
- You must be signed in to your UMass account in order to access the review questions

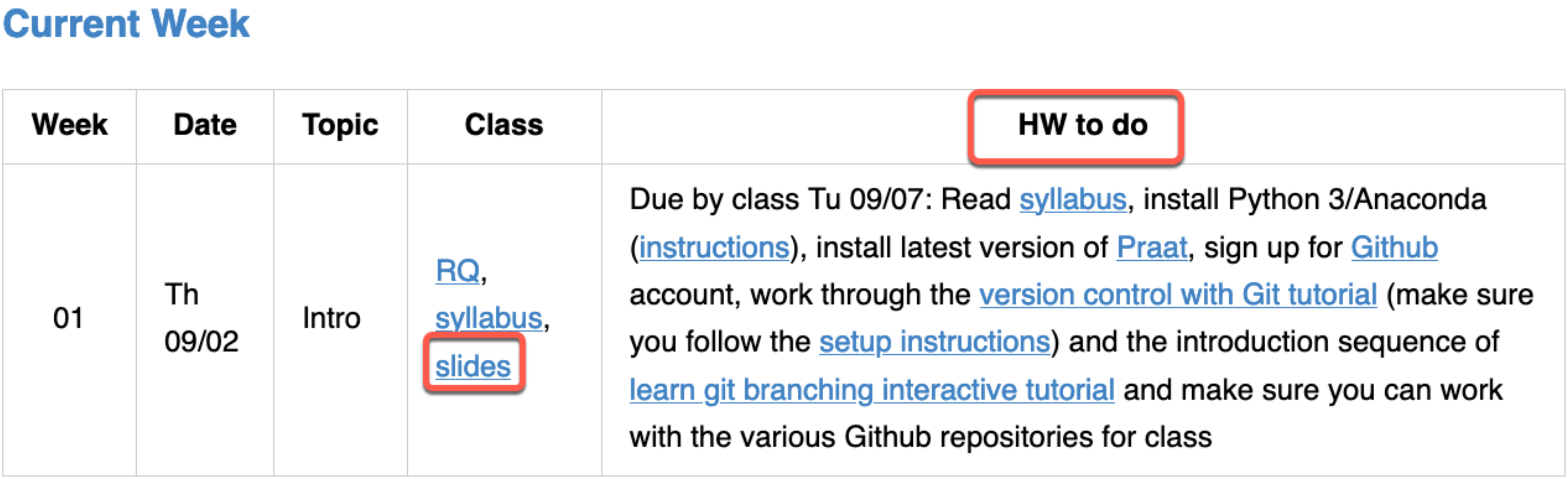


### Current Week

| Week | Date | Topic | Class | HW to do |
|---|---|---|---|---|
| 01 | Th 09/02 | Intro | RQ, syllabus, slides | Due by class Tu 09/07: Read syllabus, install Python 3/Anaconda (instructions), install latest version of Praat, sign up for Github account, work through the version control with Git tutorial (make sure you follow the setup instructions) and the introduction sequence of learn git branching interactive tutorial and make sure you can work with the various Github repositories for class |

# CLASS SLIDES AND HOMEWORK

- Linked from course webpage for each class



## Current Week

| Week | Date | Topic | Class | HW to do |
|---|---|---|---|---|
| 01 | Th 09/02 | Intro | RQ,<br>syllabus,<br>slides | Due by class Tu 09/07: Read syllabus, install Python 3/Anaconda (instructions), install latest version of Praat, sign up for Github account, work through the version control with Git tutorial (make sure you follow the setup instructions) and the introduction sequence of learn git branching interactive tutorial and make sure you can work with the various Github repositories for class |

# WHAT TO BRING TO CLASS

- A computer (if you have access to one)
- Earbuds/headphones to listen to audio sometimes

# INTRODUCTIONS

PROFESSOR YU
SHE/HER

CERYS
SHE/HER

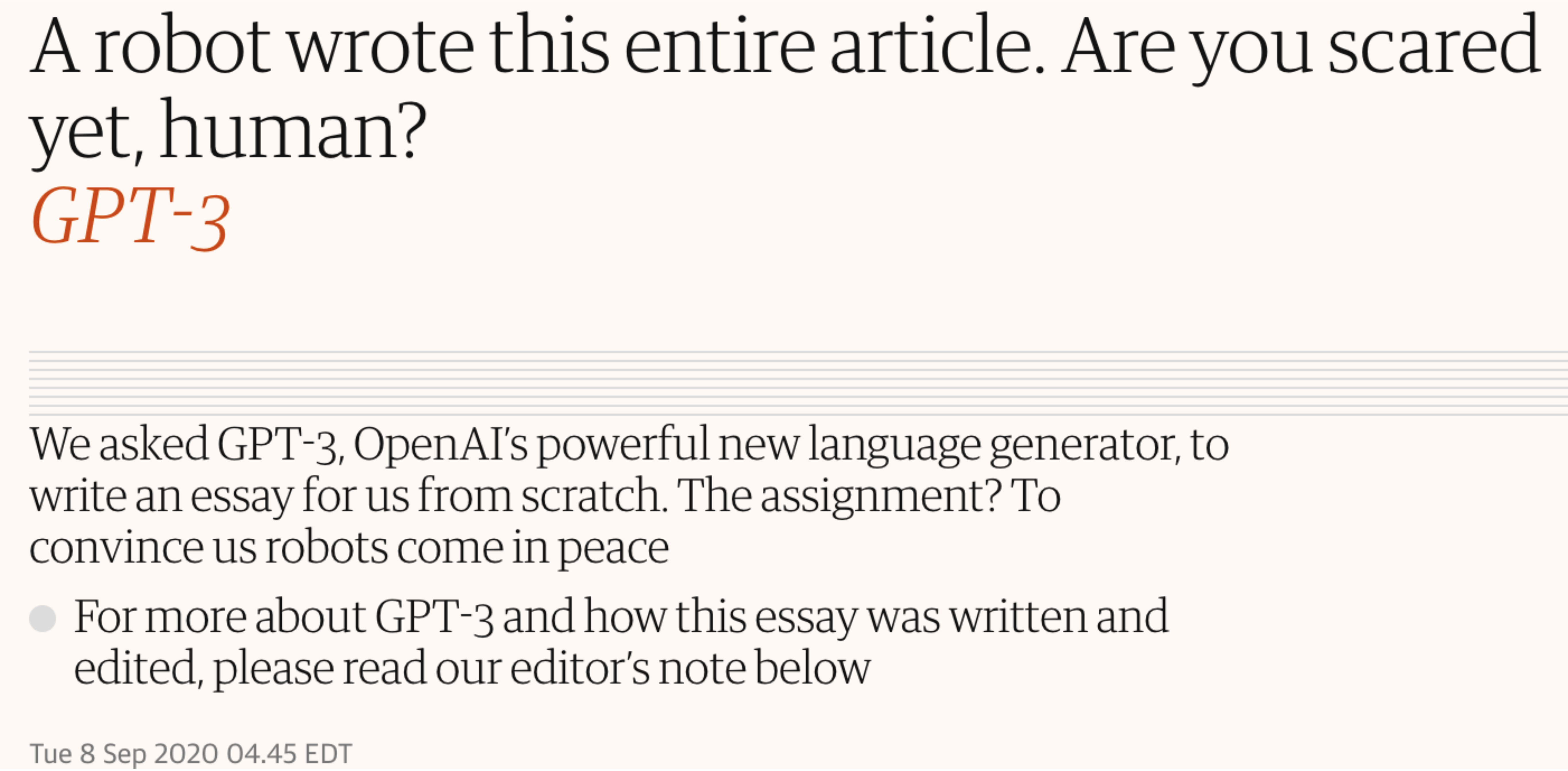
PROFESSOR YU
SHE/HER

We can start an intro thread on Slack too!

CERYS
SHE/HER

# COURSE INTRO

# RUN FOR YOUR LIVES! AI TAKEOVER!!



A robot wrote this entire article. Are you scared yet, human?

*GPT-3*

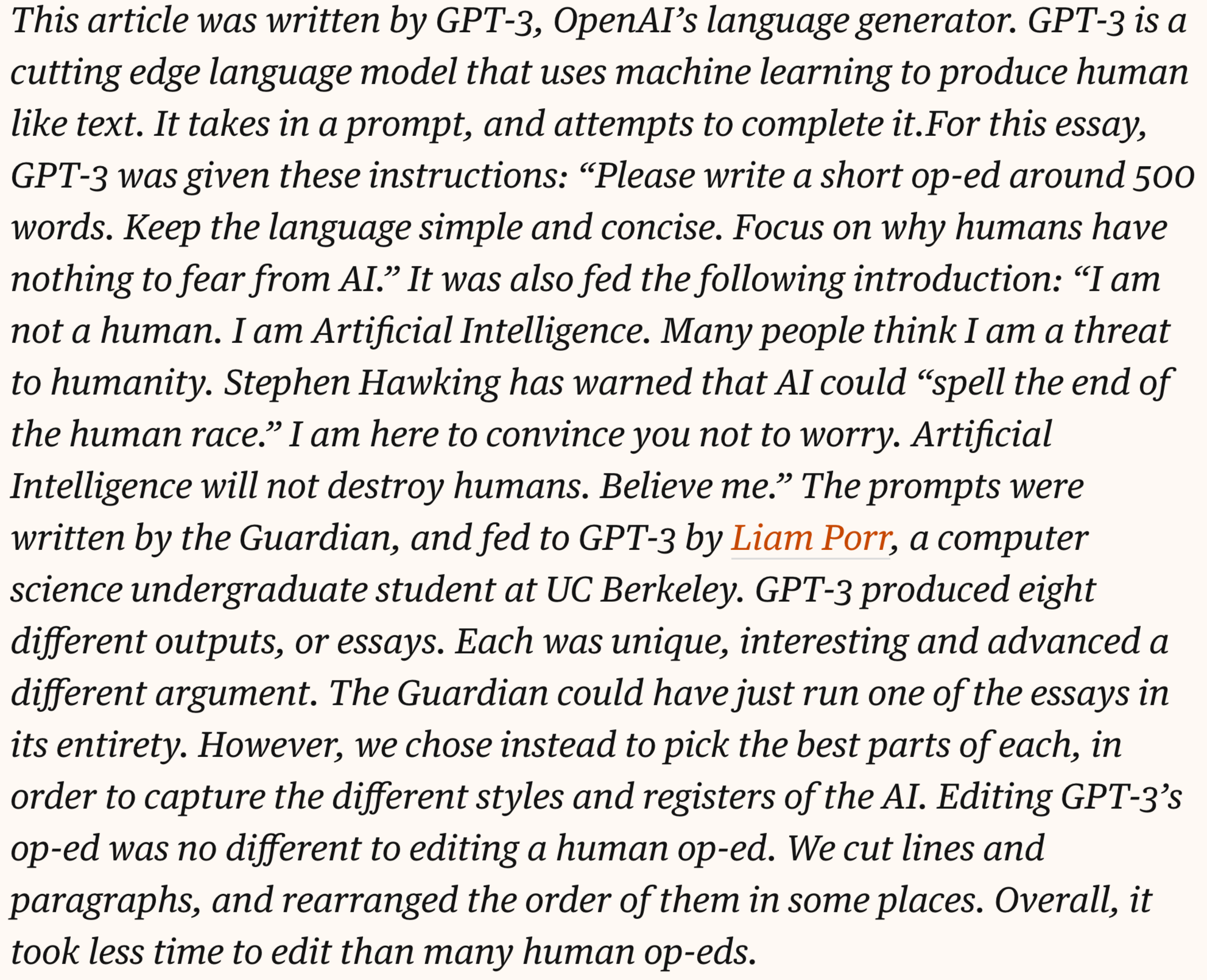
We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below

Tue 8 Sep 2020 04.45 EDT

https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

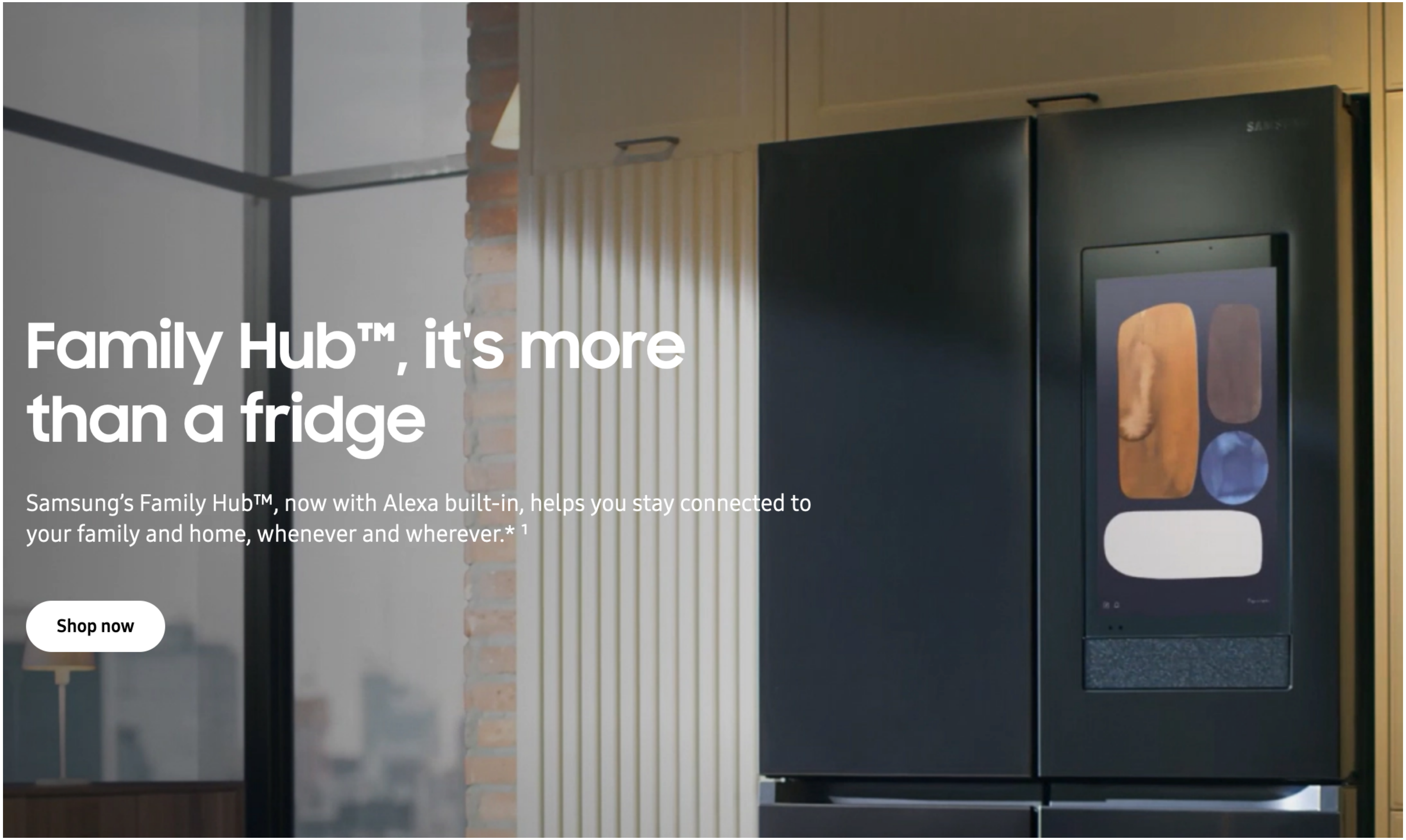
# RUN FOR YOUR LIVES! AI TAKEOVER!!



*This article was written by GPT-3, OpenAI's language generator. GPT-3 is a cutting edge language model that uses machine learning to produce human like text. It takes in a prompt, and attempts to complete it.For this essay, GPT-3 was given these instructions: "Please write a short op-ed around 500 words. Keep the language simple and concise. Focus on why humans have nothing to fear from AI." It was also fed the following introduction: "I am not a human. I am Artificial Intelligence. Many people think I am a threat to humanity. Stephen Hawking has warned that AI could "spell the end of the human race." I am here to convince you not to worry. Artificial Intelligence will not destroy humans. Believe me." The prompts were written by the Guardian, and fed to GPT-3 by Liam Porr, a computer science undergraduate student at UC Berkeley. GPT-3 produced eight different outputs, or essays. Each was unique, interesting and advanced a different argument. The Guardian could have just run one of the essays in its entirety. However, we chose instead to pick the best parts of each, in order to capture the different styles and registers of the AI. Editing GPT-3's op-ed was no different to editing a human op-ed. We cut lines and paragraphs, and rearranged the order of them in some places. Overall, it took less time to edit than many human op-eds.*

https://thenextweb.com/news/the-guardians-gpt-3-generated-article-is-everything-wrong-with-ai-media-hype

# SPEECH RECOGNITION EVERYWHERE



Family Hub™, it's more than a fridge

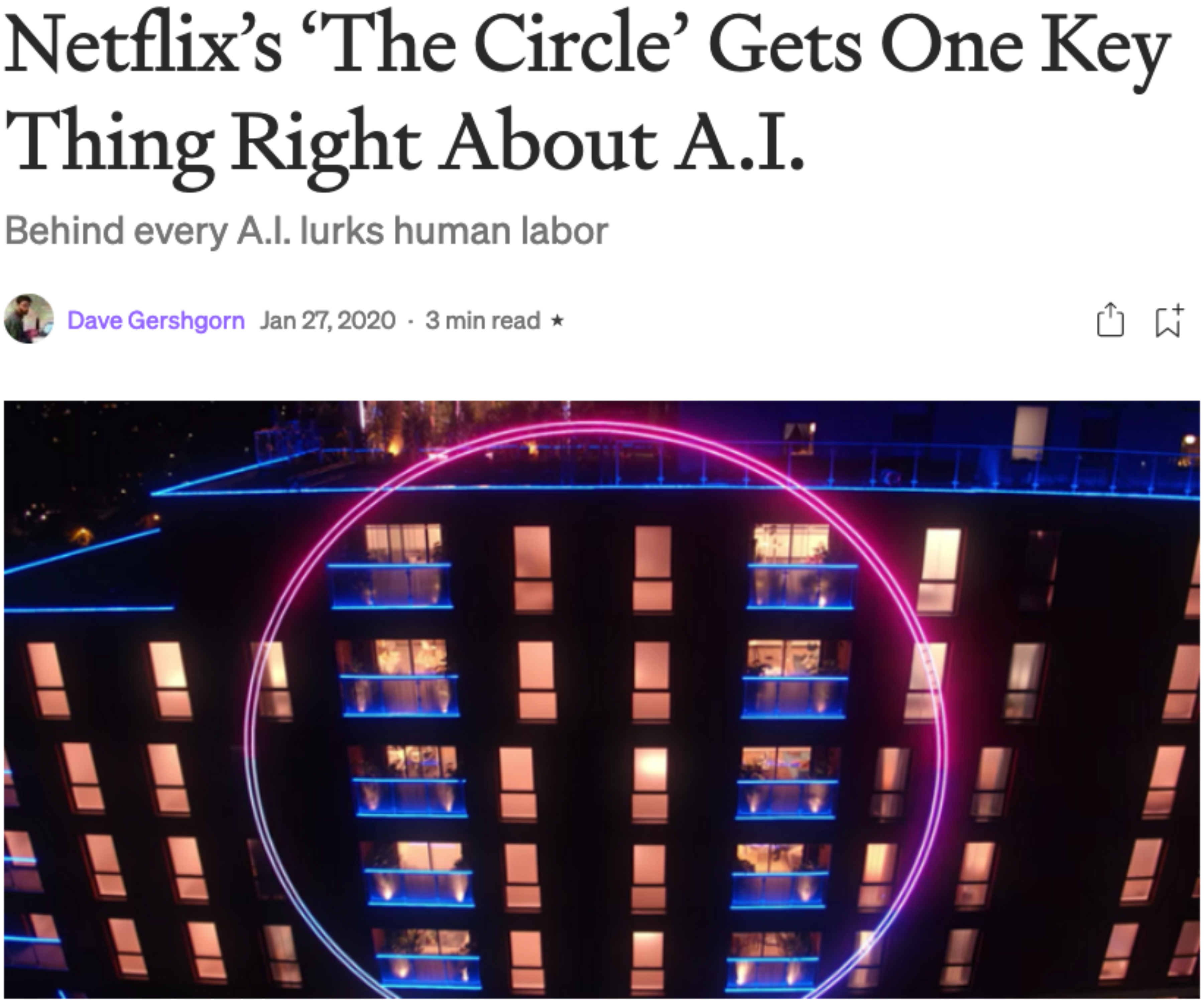
Samsung's Family Hub™, now with Alexa built-in, helps you stay connected to your family and home, whenever and wherever.* 1

https://www.samsung.com/us/explore/family-hub-refrigerator/overview/

# THE CIRCLE (CHAT OPENS 12:24)

# THE CIRCLE (CHAT OPENS 12:24)

# SUPER-POWERED SPEECH RECOGNITION?



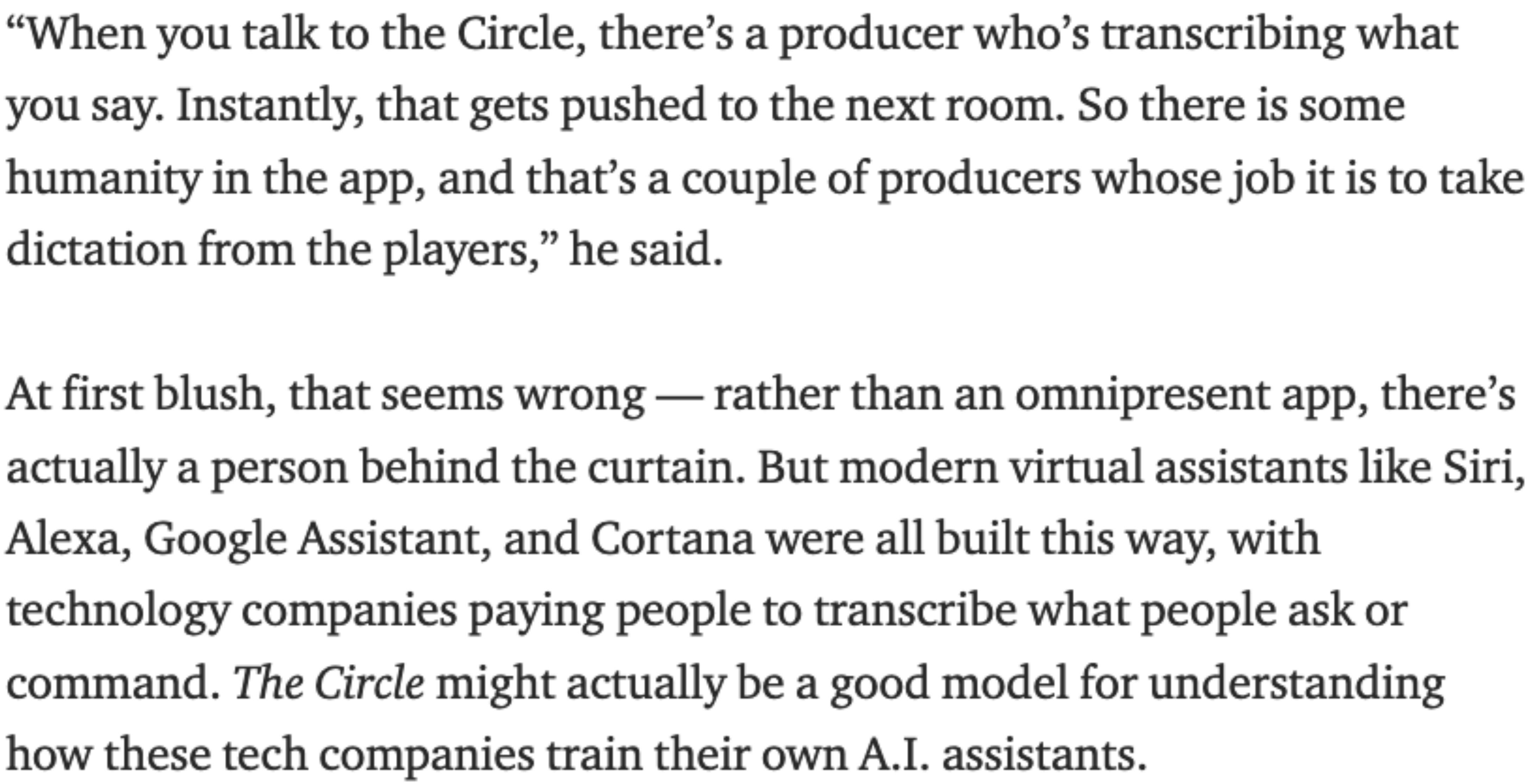
Netflix's 'The Circle' Gets One Key Thing Right About A.I.

Behind every A.I. lurks human labor

Dave Gershgorn Jan 27, 2020 · 3 min read ★

https://onezero.medium.com/netflixs-the-circle-gets-one-key-thing-right-about-a-i-ed957b018b1e

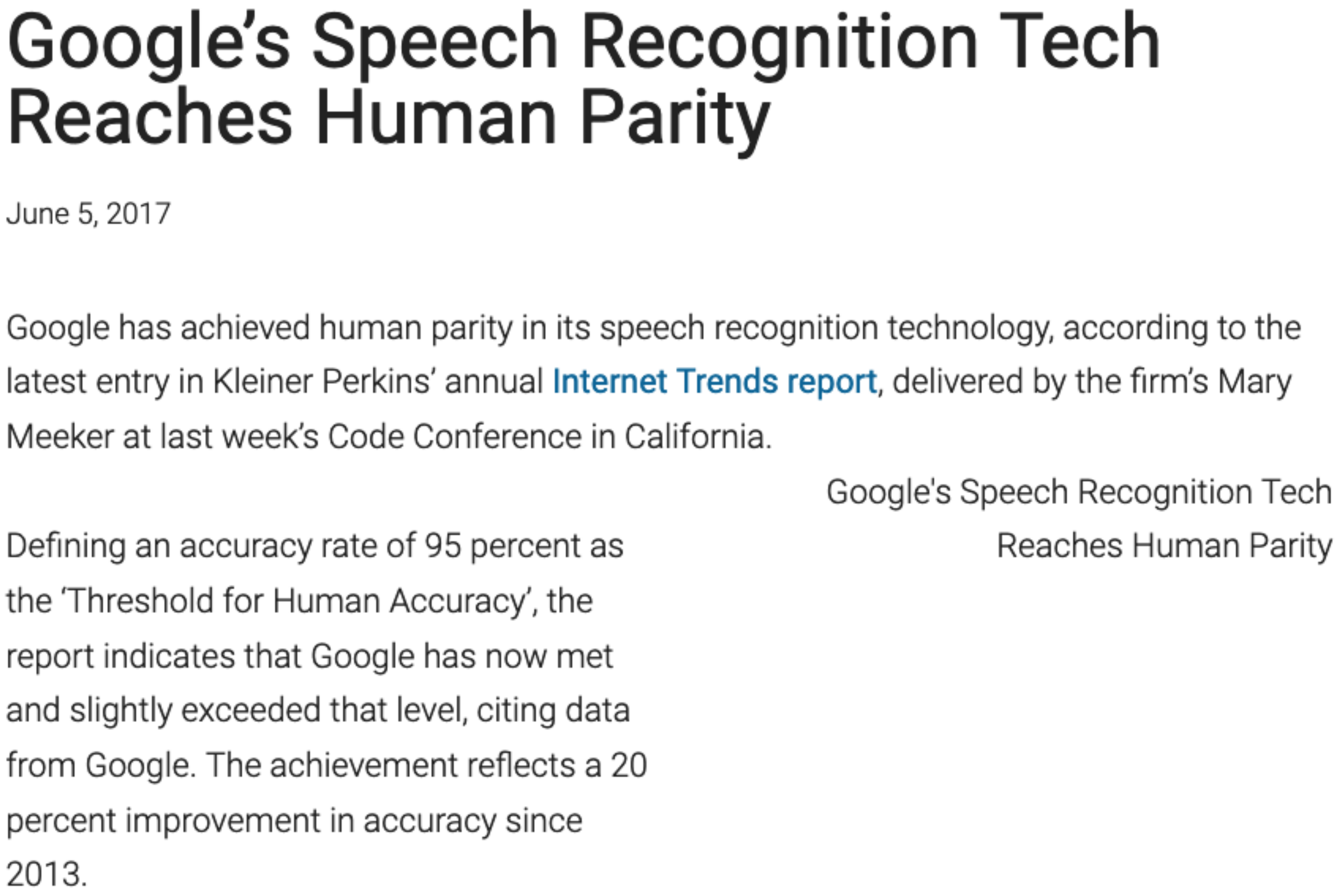
# MASSIVE DATA, HUMAN COGS BEHIND AI



"When you talk to the Circle, there's a producer who's transcribing what you say. Instantly, that gets pushed to the next room. So there is some humanity in the app, and that's a couple of producers whose job it is to take dictation from the players," he said.

At first blush, that seems wrong — rather than an omnipresent app, there's actually a person behind the curtain. But modern virtual assistants like Siri, Alexa, Google Assistant, and Cortana were all built this way, with technology companies paying people to transcribe what people ask or command. *The Circle* might actually be a good model for understanding how these tech companies train their own A.I. assistants.

# SPEECH RECOGNITION IS SOLVED!



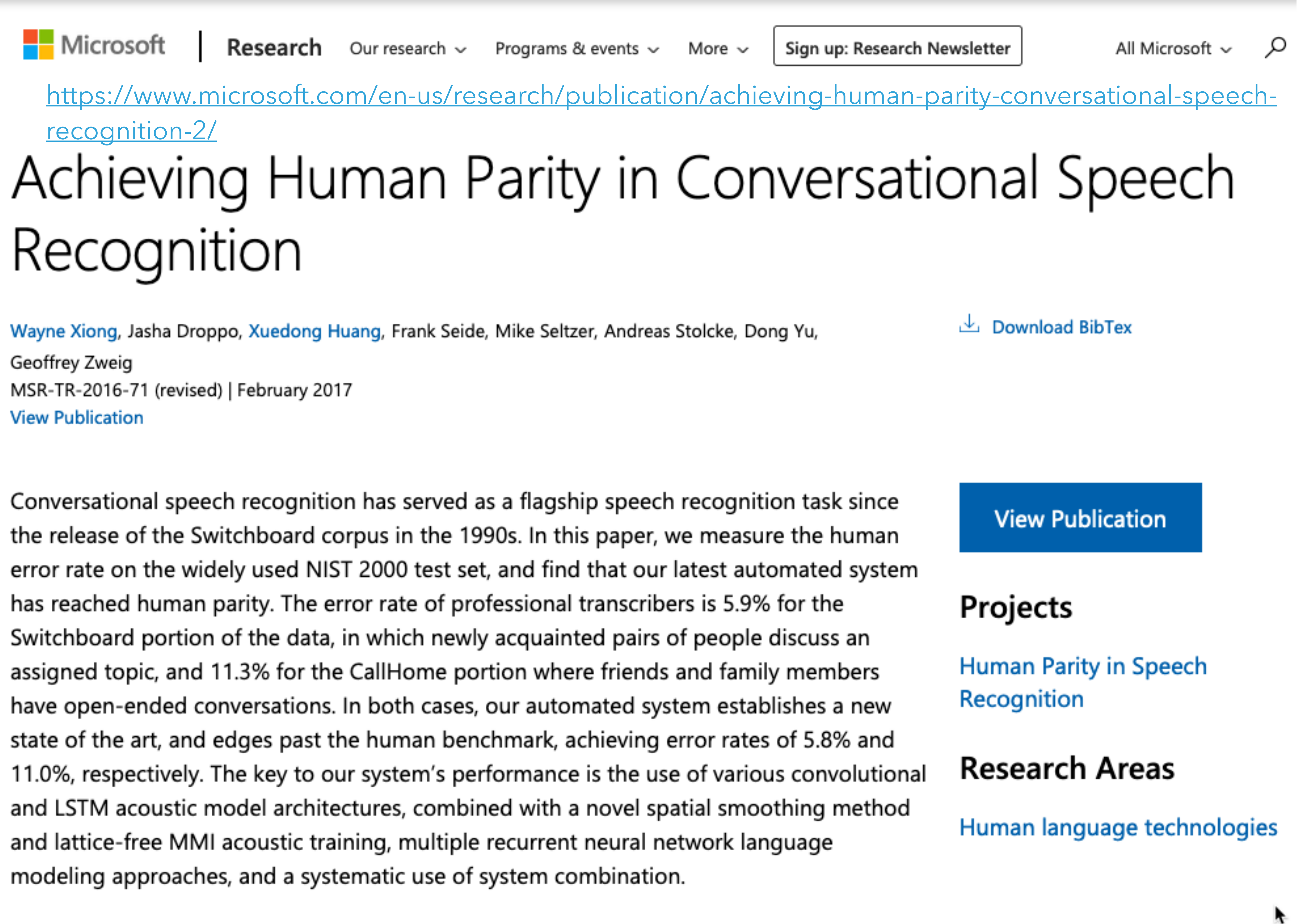
## Google's Speech Recognition Tech Reaches Human Parity

June 5, 2017

Google has achieved human parity in its speech recognition technology, according to the latest entry in Kleiner Perkins' annual Internet Trends report, delivered by the firm's Mary Meeker at last week's Code Conference in California.

Google's Speech Recognition Tech Reaches Human Parity

Defining an accuracy rate of 95 percent as the 'Threshold for Human Accuracy', the report indicates that Google has now met and slightly exceeded that level, citing data from Google. The achievement reflects a 20 percent improvement in accuracy since 2013.



https://www.microsoft.com/en-us/research/publication/achieving-human-parity-conversational-speech-recognition-2/

# Achieving Human Parity in Conversational Speech Recognition

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, Geoffrey Zweig

MSR-TR-2016-71 (revised) | February 2017

View Publication

Conversational speech recognition has served as a flagship speech recognition task since the release of the Switchboard corpus in the 1990s. In this paper, we measure the human error rate on the widely used NIST 2000 test set, and find that our latest automated system has reached human parity. The error rate of professional transcribers is 5.9% for the Switchboard portion of the data, in which newly acquainted pairs of people discuss an assigned topic, and 11.3% for the CallHome portion where friends and family members have open-ended conversations. In both cases, our automated system establishes a new state of the art, and edges past the human benchmark, achieving error rates of 5.8% and 11.0%, respectively. The key to our system's performance is the use of various convolutional and LSTM acoustic model architectures, combined with a novel spatial smoothing method and lattice-free MMI acoustic training, multiple recurrent neural network language modeling approaches, and a systematic use of system combination.

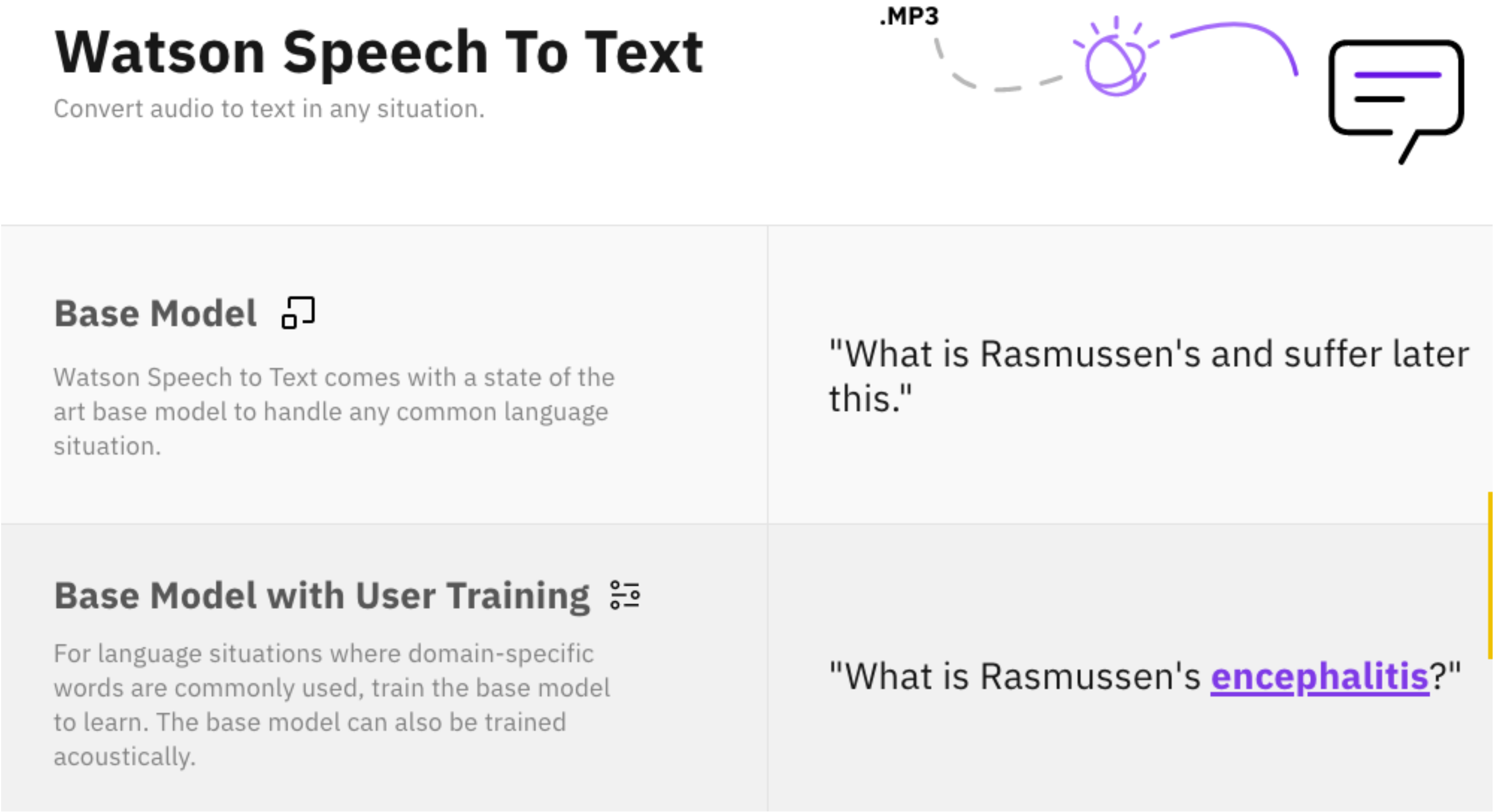
## Projects

Human Parity in Speech Recognition

## Research Areas

Human language technologies

# LET'S TAKE WATSON FOR A SPIN…



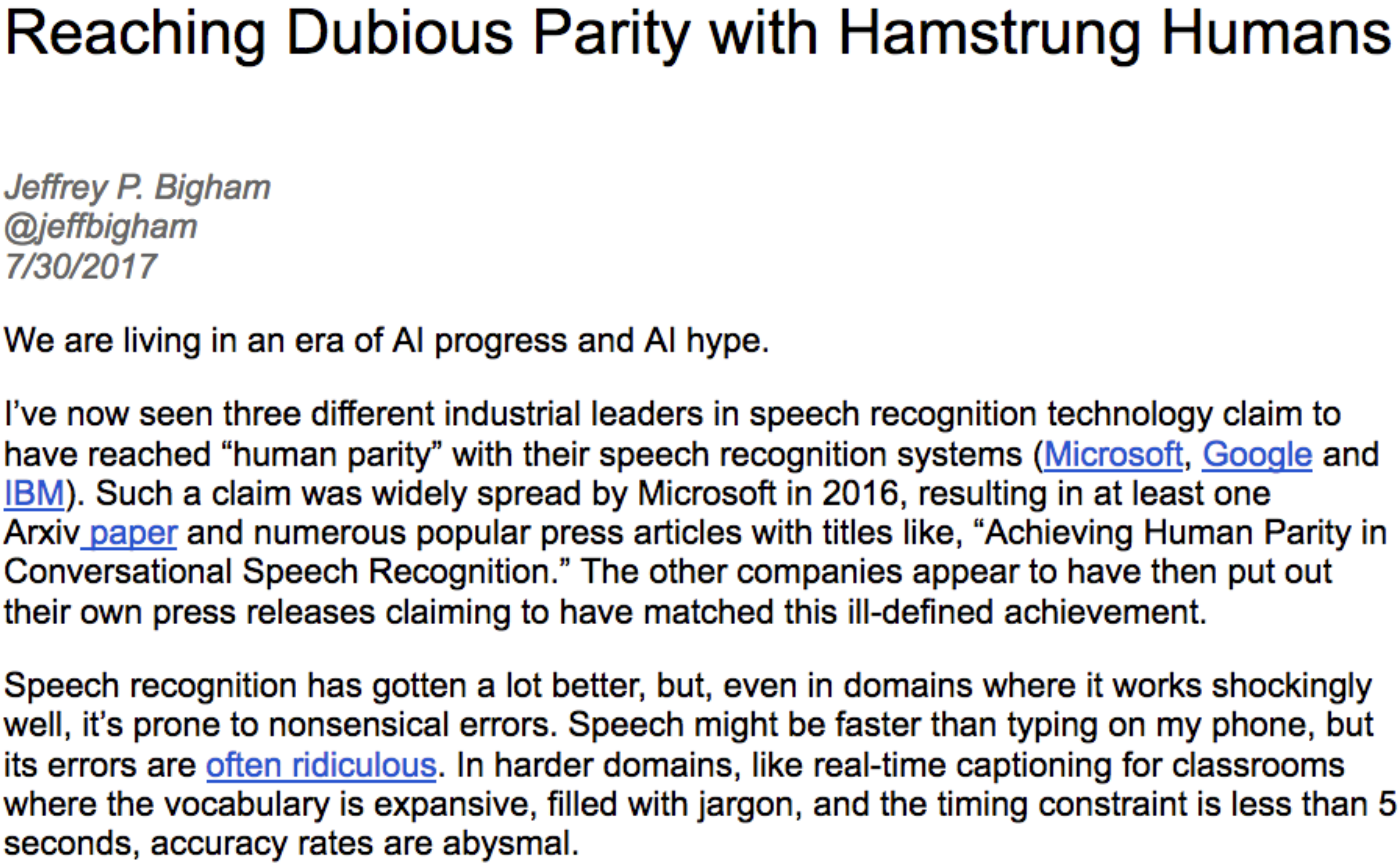
## Watson Speech To Text

Convert audio to text in any situation.

| | |
|---|---|
| **Base Model**<br>Watson Speech to Text comes with a state of the art base model to handle any common language situation. | "What is Rasmussen's and suffer later this." |
| **Base Model with User Training**<br>For language situations where domain-specific words are commonly used, train the base model to learn. The base model can also be trained acoustically. | "What is Rasmussen's **encephalitis**?" |

https://www.ibm.com/demos/live/speech-to-text/self-service

# SPEECH RECOGNITION IS NOT SOLVED



## Reaching Dubious Parity with Hamstrung Humans

*Jeffrey P. Bigham*
*@jeffbigham*
*7/30/2017*

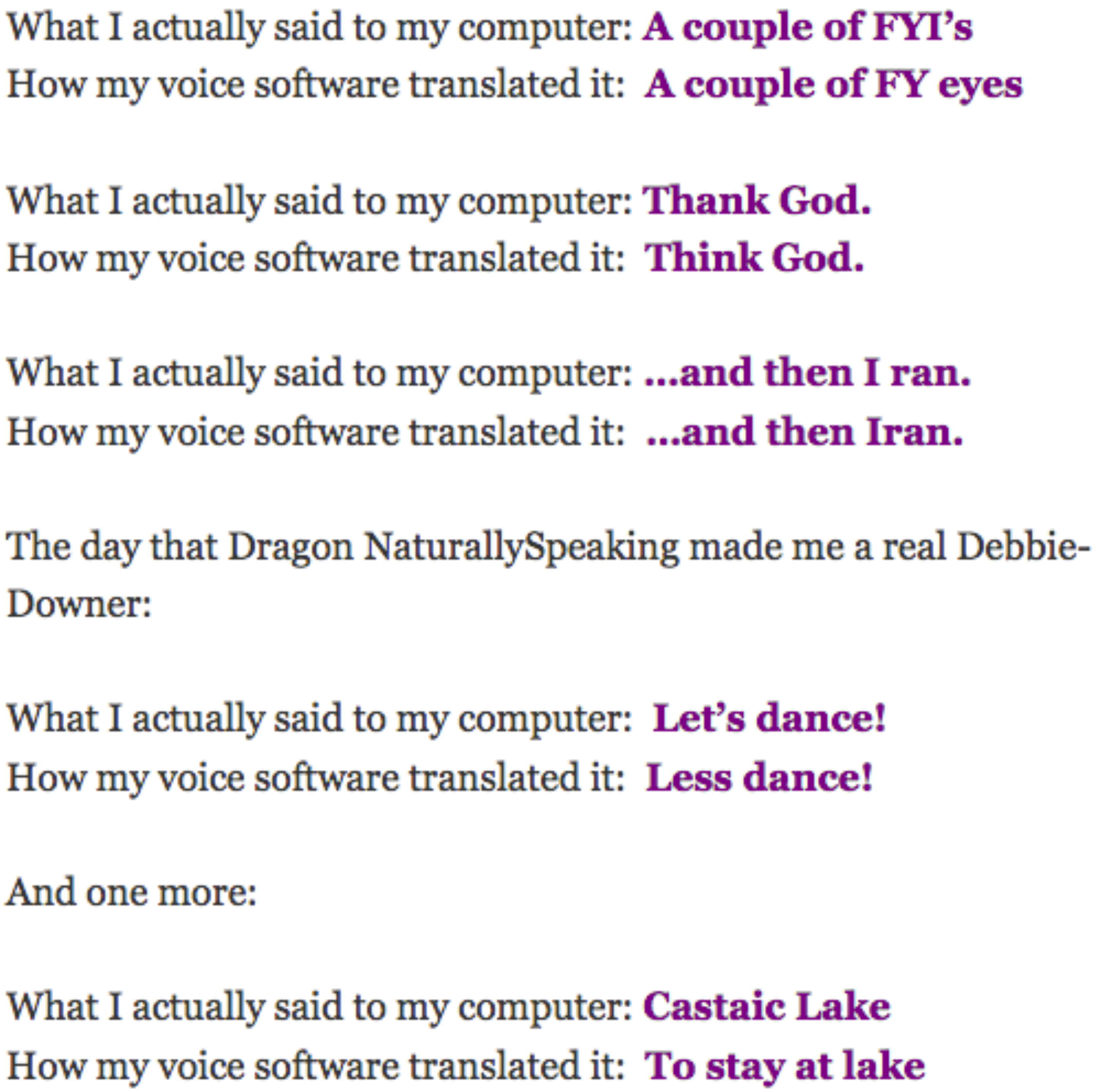
We are living in an era of AI progress and AI hype.

I've now seen three different industrial leaders in speech recognition technology claim to have reached "human parity" with their speech recognition systems (Microsoft, Google and IBM). Such a claim was widely spread by Microsoft in 2016, resulting in at least one Arxiv paper and numerous popular press articles with titles like, "Achieving Human Parity in Conversational Speech Recognition." The other companies appear to have then put out their own press releases claiming to have matched this ill-defined achievement.

Speech recognition has gotten a lot better, but, even in domains where it works shockingly well, it's prone to nonsensical errors. Speech might be faster than typing on my phone, but its errors are often ridiculous. In harder domains, like real-time captioning for classrooms where the vocabulary is expansive, filled with jargon, and the timing constraint is less than 5 seconds, accuracy rates are abysmal.

http://jeffreybigham.com/blog/2017/reaching-dubious-parity-with-hamstrung-humans.html

# SPEECH RECOGNITION IS NOT SOLVED



What I actually said to my computer: **A couple of FYI's**
How my voice software translated it: **A couple of FY eyes**

What I actually said to my computer: **Thank God.**
How my voice software translated it: **Think God.**

What I actually said to my computer: **...and then I ran.**
How my voice software translated it: **...and then Iran.**

The day that Dragon NaturallySpeaking made me a real Debbie-Downer:

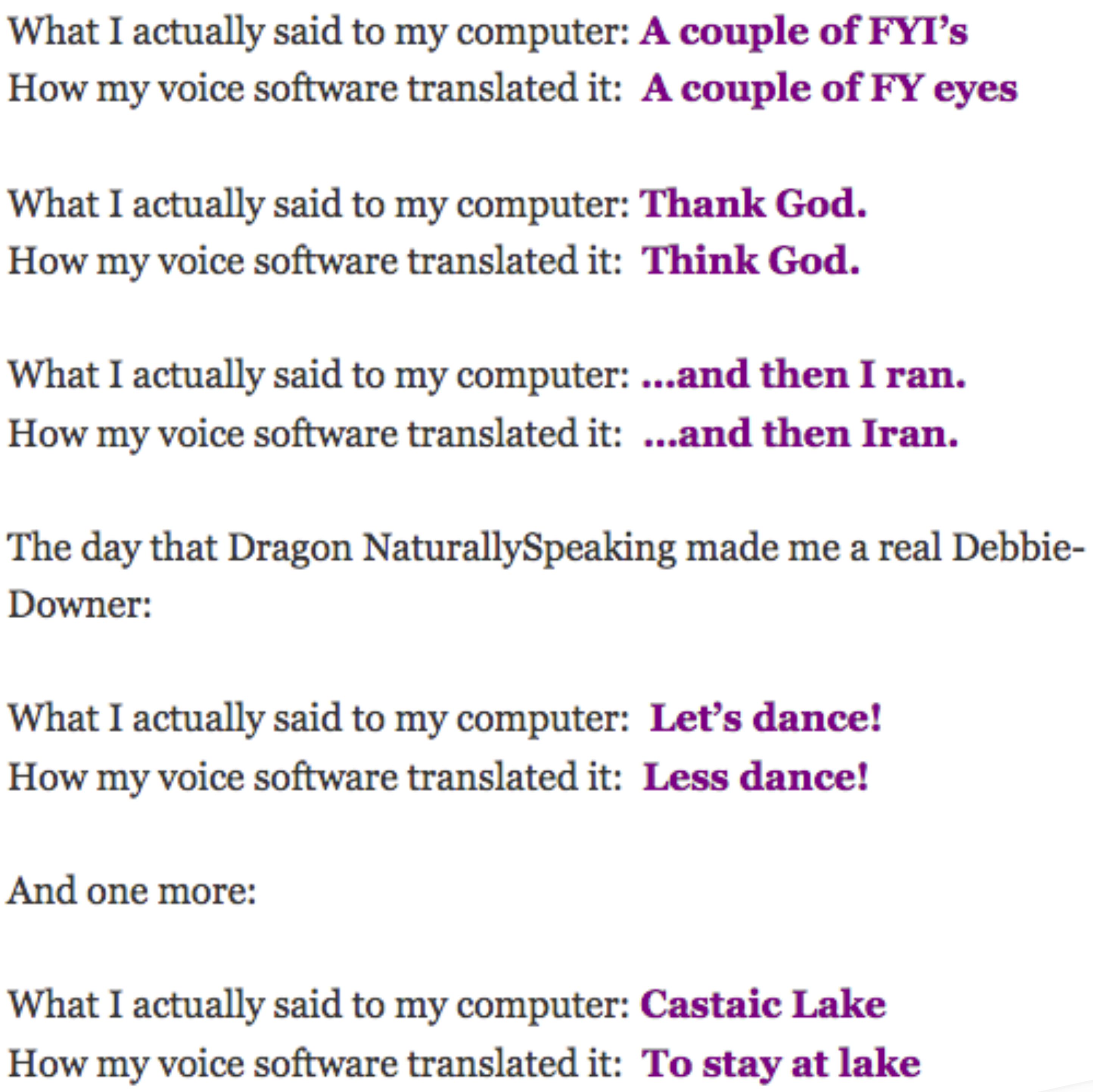
What I actually said to my computer: **Let's dance!**
How my voice software translated it: **Less dance!**

And one more:

What I actually said to my computer: **Castaic Lake**
How my voice software translated it: **To stay at lake**

http://www.writeworks.biz/blog/voxrec/
(defunct link now as of September 2021)

# SPEECH RECOGNITION IS NOT SOLVED



What I actually said to my computer: **A couple of FYI's**
How my voice software translated it: **A couple of FY eyes**

What I actually said to my computer: **Thank God.**
How my voice software translated it: **Think God.**

What I actually said to my computer: **...and then I ran.**
How my voice software translated it: **...and then Iran.**

The day that Dragon NaturallySpeaking made me a real Debbie-Downer:

What I actually said to my computer: **Let's dance!**
How my voice software translated it: **Less dance!**

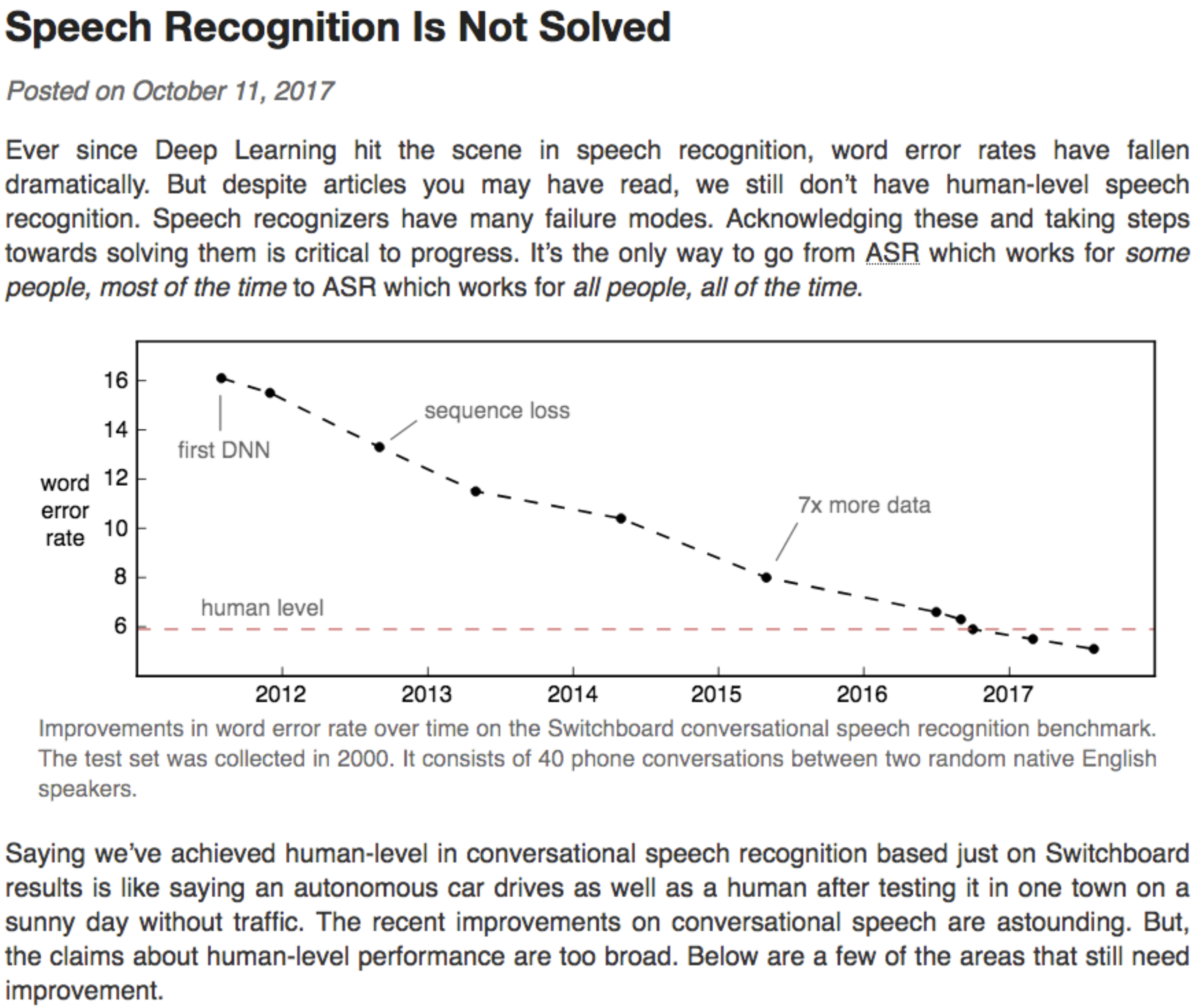
And one more:

What I actually said to my computer: **Castaic Lake**
How my voice software translated it: **To stay at lake**

http://www.writeworks.biz/blog/voxrec/
(defunct link now as of September 2021)

**(Language model errors?)**

# DEEP LEARNING TO THE RESCUE?



## Speech Recognition Is Not Solved

*Posted on October 11, 2017*

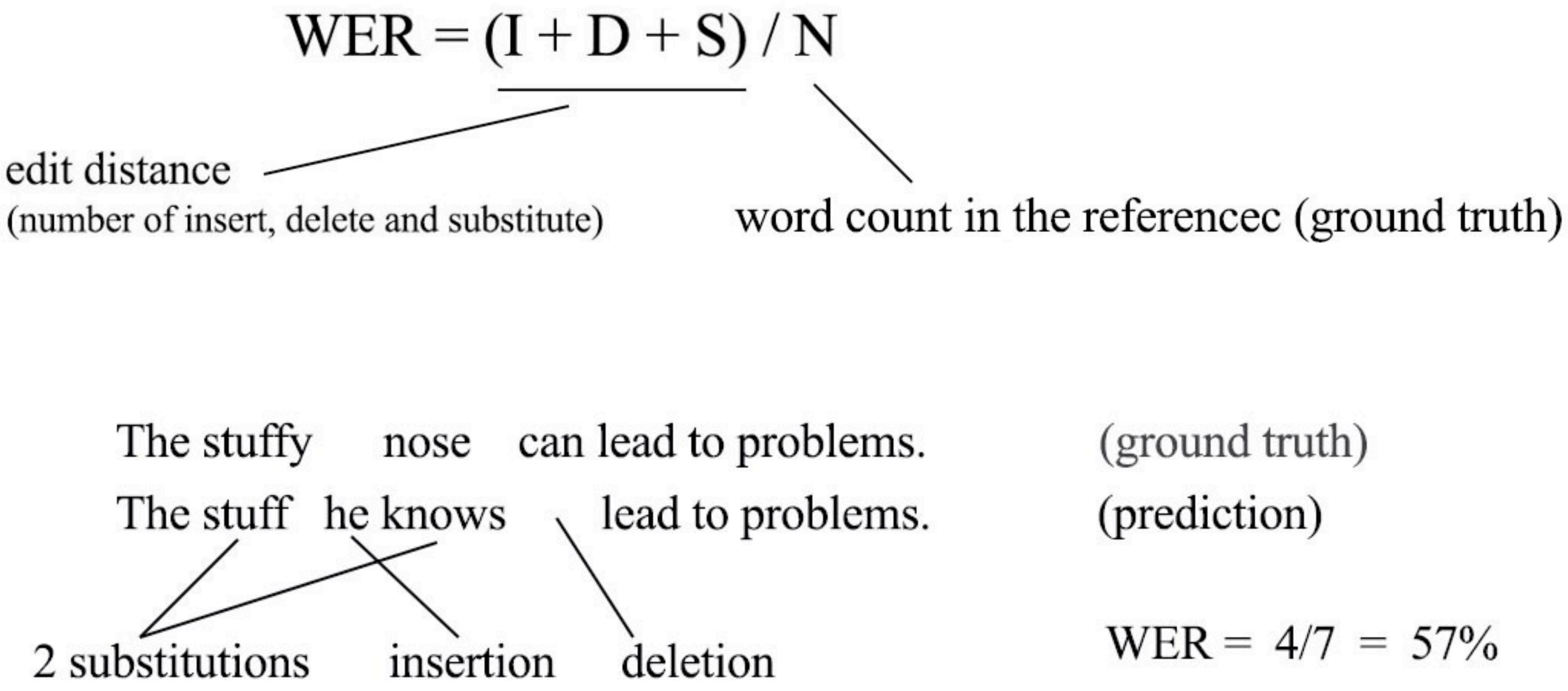
Ever since Deep Learning hit the scene in speech recognition, word error rates have fallen dramatically. But despite articles you may have read, we still don't have human-level speech recognition. Speech recognizers have many failure modes. Acknowledging these and taking steps towards solving them is critical to progress. It's the only way to go from ASR which works for *some people, most of the time* to ASR which works for *all people, all of the time*.

Improvements in word error rate over time on the Switchboard conversational speech recognition benchmark. The test set was collected in 2000. It consists of 40 phone conversations between two random native English speakers.

Saying we've achieved human-level in conversational speech recognition based just on Switchboard results is like saying an autonomous car drives as well as a human after testing it in one town on a sunny day without traffic. The recent improvements on conversational speech are astounding. But, the claims about human-level performance are too broad. Below are a few of the areas that still need improvement.

https://awni.github.io/speech-recognition/

# WORD ERROR RATE (WER)



WER = (I + D + S) / N

edit distance
(number of insert, delete and substitute)

word count in the referencec (ground truth)

The stuffy nose can lead to problems. (ground truth)

The stuff he knows lead to problems. (prediction)

2 substitutions

insertion

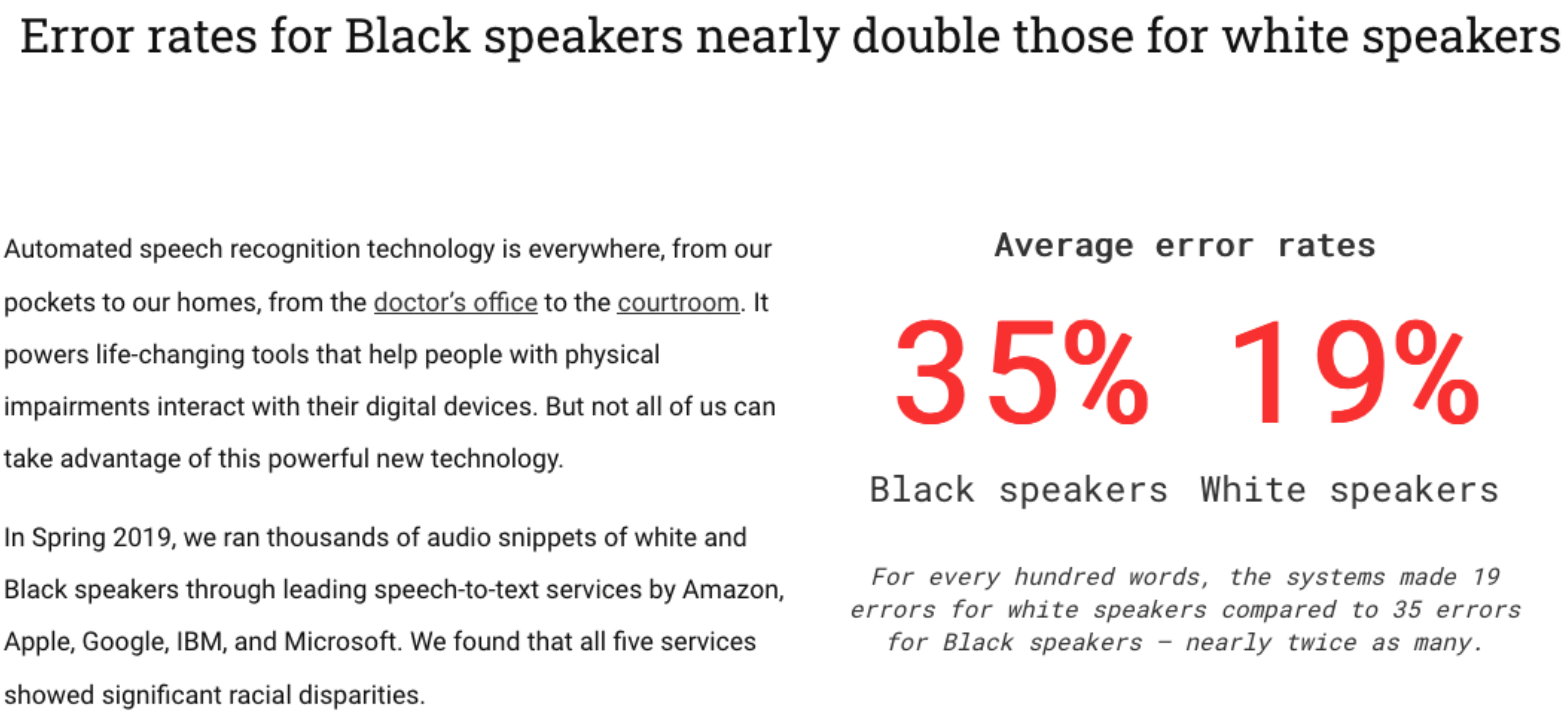
deletion

WER = 4/7 = 57%

Taken from https://jonathan-hui.medium.com/speech-recognition-asr-decoding-f152aebed779

**See also** https://aws.amazon.com/blogs/machine-learning/evaluating-an-automatic-speech-recognition-service/

# LINGUISTIC VARIABILITY AND MODEL BIAS



## Error rates for Black speakers nearly double those for white speakers

Automated speech recognition technology is everywhere, from our pockets to our homes, from the doctor's office to the courtroom. It powers life-changing tools that help people with physical impairments interact with their digital devices. But not all of us can take advantage of this powerful new technology.

In Spring 2019, we ran thousands of audio snippets of white and Black speakers through leading speech-to-text services by Amazon, Apple, Google, IBM, and Microsoft. We found that all five services showed significant racial disparities.

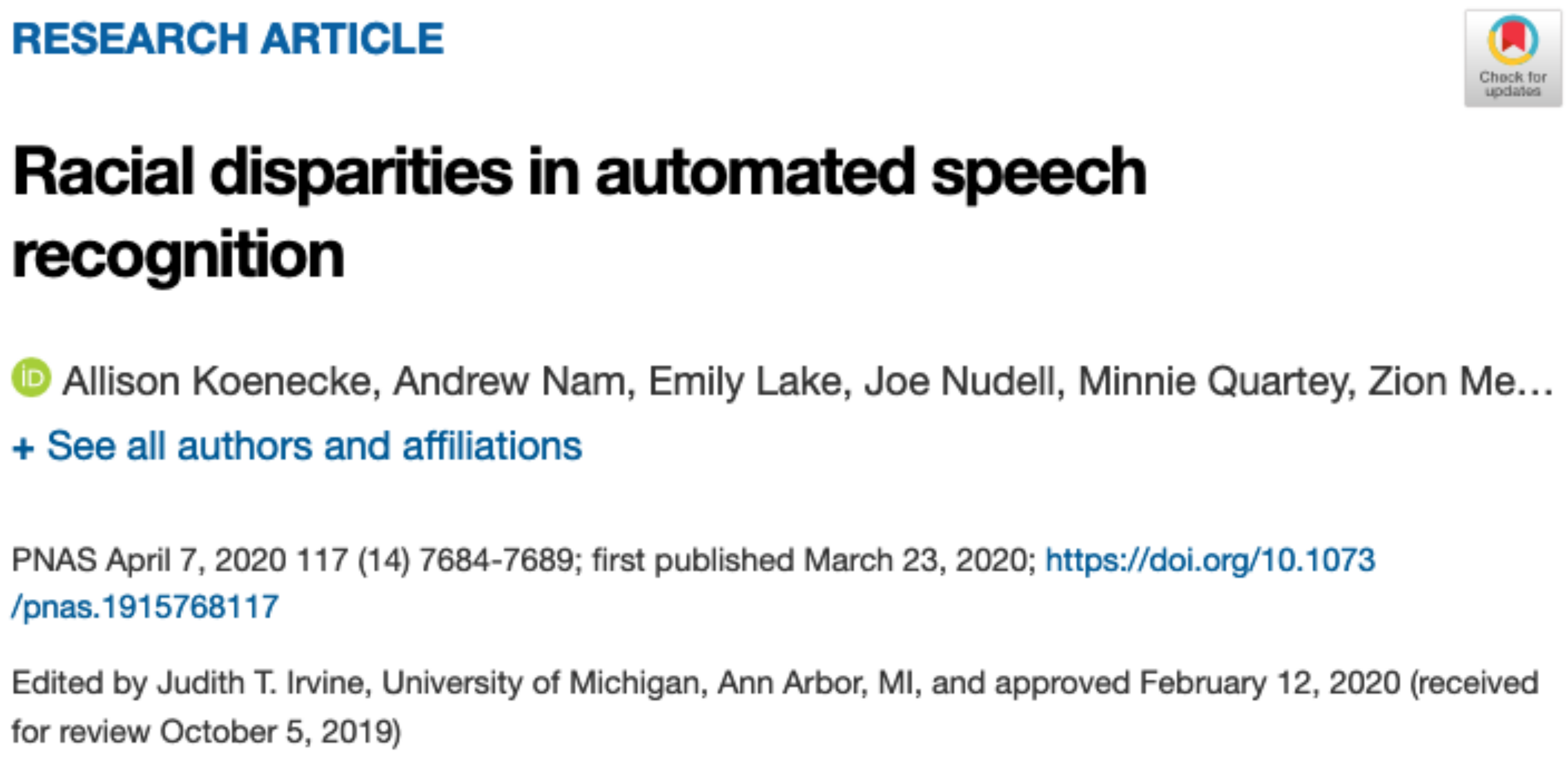
**Average error rates**

| 35% | 19% |
|---|---|
| Black speakers | White speakers |

*For every hundred words, the systems made 19 errors for white speakers compared to 35 errors for Black speakers – nearly twice as many.*

https://fairspeech.stanford.edu/

# LINGUISTIC VARIABILITY AND MODEL BIAS



RESEARCH ARTICLE

**Racial disparities in automated speech recognition**

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Me...

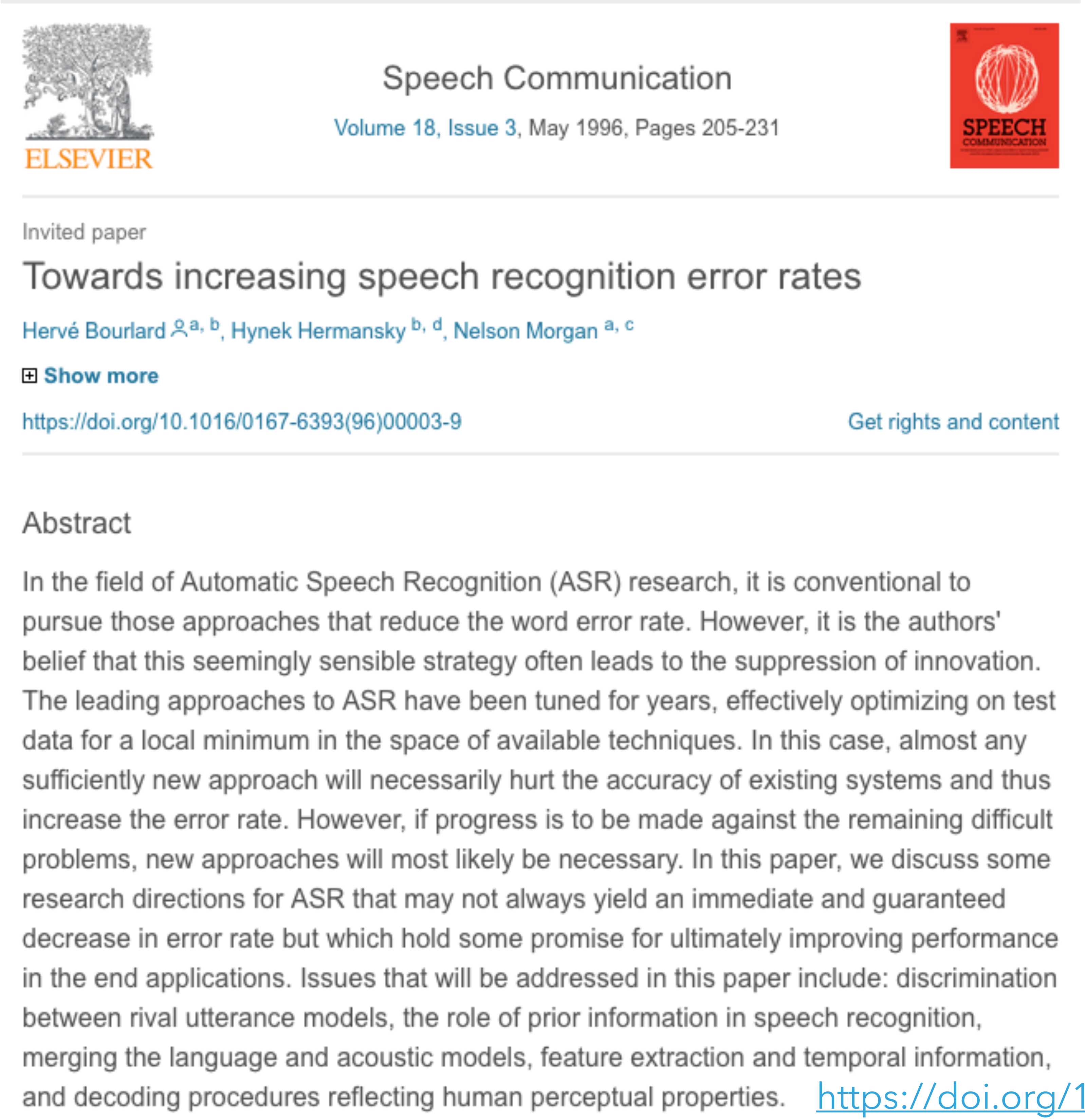
+ See all authors and affiliations

PNAS April 7, 2020 117 (14) 7684-7689; first published March 23, 2020; https://doi.org/10.1073/pnas.1915768117

Edited by Judith T. Irvine, University of Michigan, Ann Arbor, MI, and approved February 12, 2020 (received for review October 5, 2019)

https://www.pnas.org/content/117/14/7684

# HOW CAN LINGUISTICS HELP?



ELSEVIER

Speech Communication

Volume 18, Issue 3, May 1996, Pages 205-231

Invited paper

## Towards increasing speech recognition error rates

Hervé Bourlard [a, b], Hynek Hermansky [b, d], Nelson Morgan [a, c]

⊞ Show more

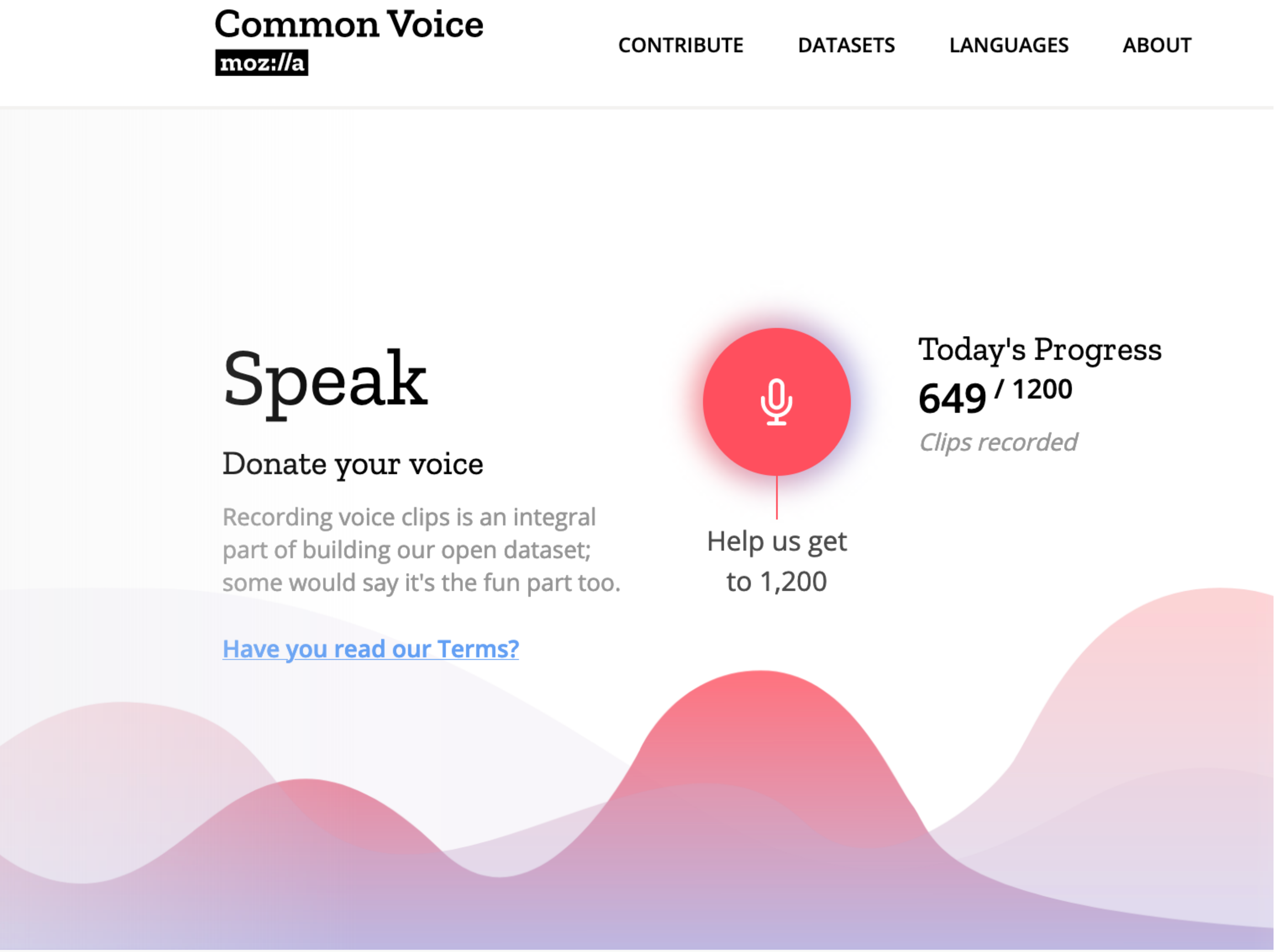
https://doi.org/10.1016/0167-6393(96)00003-9 Get rights and content

### Abstract

In the field of Automatic Speech Recognition (ASR) research, it is conventional to pursue those approaches that reduce the word error rate. However, it is the authors' belief that this seemingly sensible strategy often leads to the suppression of innovation. The leading approaches to ASR have been tuned for years, effectively optimizing on test data for a local minimum in the space of available techniques. In this case, almost any sufficiently new approach will necessarily hurt the accuracy of existing systems and thus increase the error rate. However, if progress is to be made against the remaining difficult problems, new approaches will most likely be necessary. In this paper, we discuss some research directions for ASR that may not always yield an immediate and guaranteed decrease in error rate but which hold some promise for ultimately improving performance in the end applications. Issues that will be addressed in this paper include: discrimination between rival utterance models, the role of prior information in speech recognition, merging the language and acoustic models, feature extraction and temporal information, and decoding procedures reflecting human perceptual properties.

https://doi.org/10.1016/0167-6393(96)00003-9

# EFFORTS TO BETTER SAMPLE POPULATION



Speak

Donate your voice

Recording voice clips is an integral part of building our open dataset; some would say it's the fun part too.

Have you read our Terms?

Help us get to 1,200

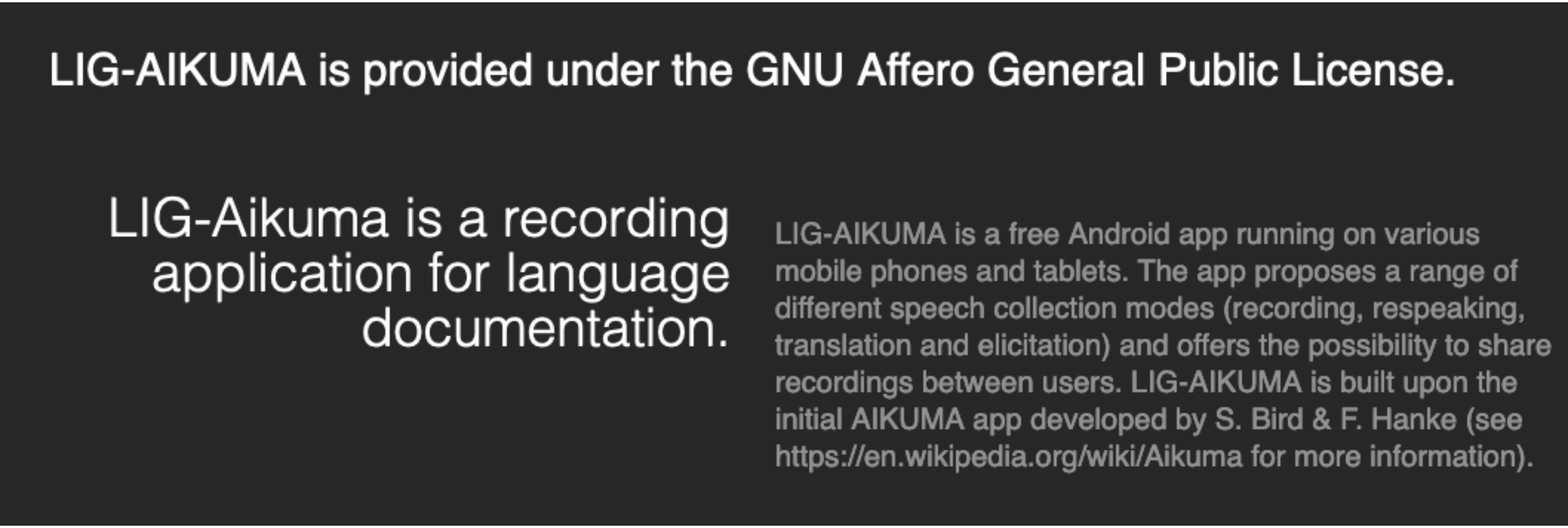
Today's Progress

649 / 1200

*Clips recorded*

https://commonvoice.mozilla.org/en

# MAKING RECORDING ACCESSIBLE



LIG-AIKUMA is provided under the GNU Affero General Public License.

LIG-Aikuma is a recording application for language documentation.

LIG-AIKUMA is a free Android app running on various mobile phones and tablets. The app proposes a range of different speech collection modes (recording, respeaking, translation and elicitation) and offers the possibility to share recordings between users. LIG-AIKUMA is built upon the initial AIKUMA app developed by S. Bird & F. Hanke (see https://en.wikipedia.org/wiki/Aikuma for more information).

https://lig-aikuma.imag.fr/

http://www.stevenbird.net/

https://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00387

# GOALS OF THIS COURSE

- To build a foundation for fluency in linguistically (especially phonetically) informed speech processing
- To discover open research problems in spoken language recognition and how you might work towards solving them

# FOUNDATIONS OF SPEECH PROCESSING

“There will always be signals, they will always need processing, and there will always be new applications, new mathematics and new implementation technologies.”

*Alan Oppenheim*

https://futureofsp.eecs.mit.edu/

# TIME TO WORK ON REVIEW QUESTIONS

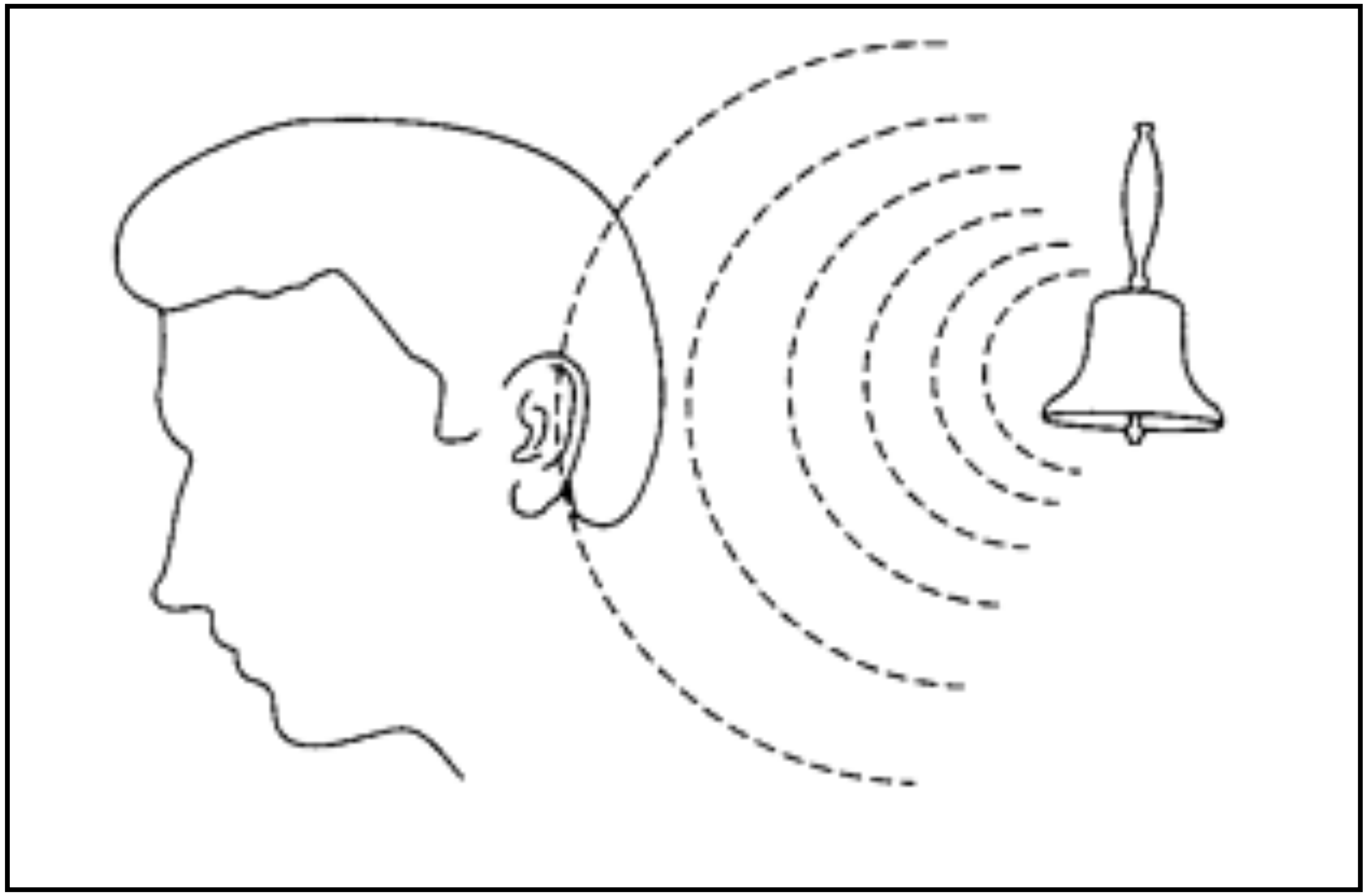
# TIME TO WORK ON REVIEW QUESTIONS

Follow up on this for HW!

# SOUND AND DIGITAL SPEECH REVIEW (IF TIME)

# WHAT IS SOUND?

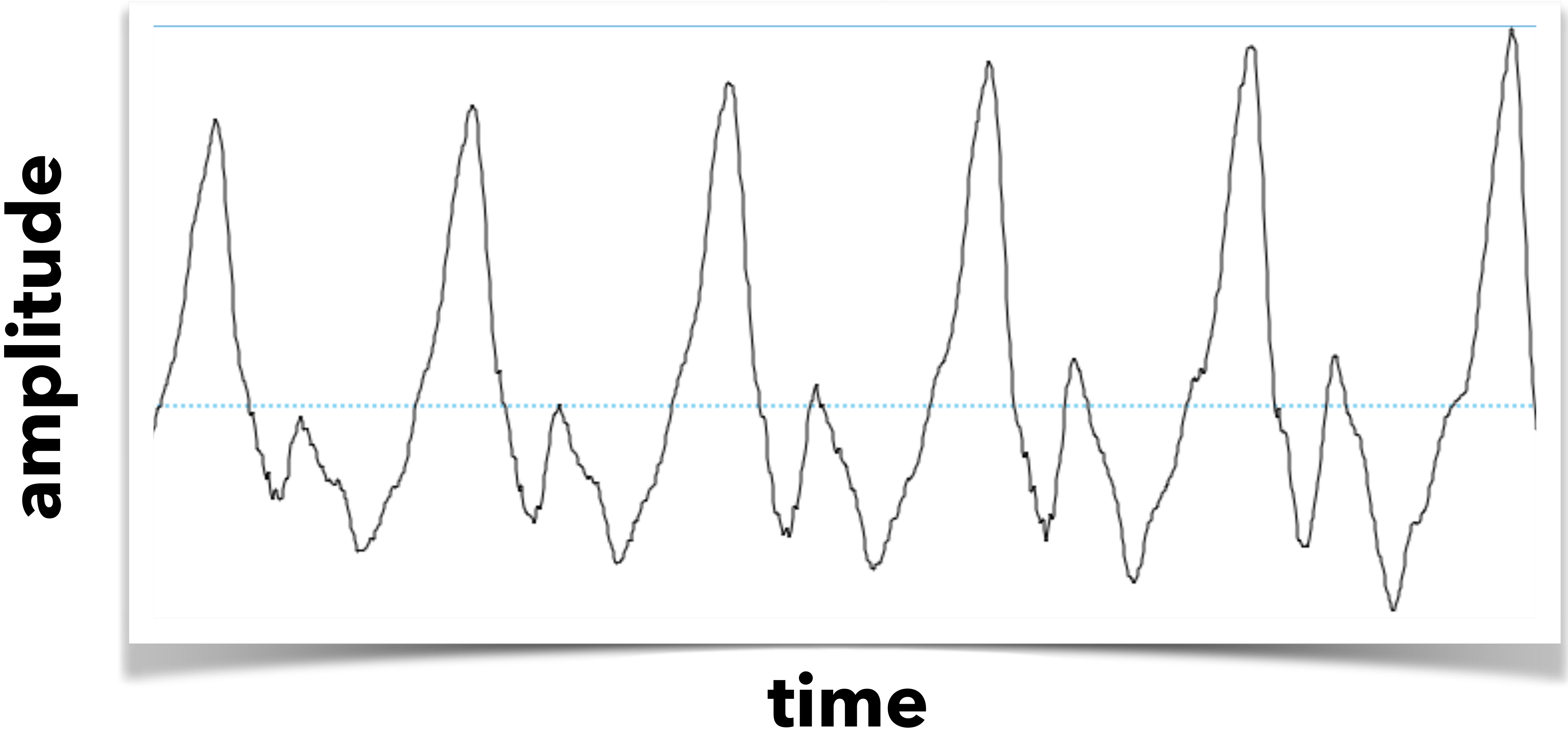
**<u>Sound</u> is the sensation you experience when your auditory nerves are stimulated by vibrating air molecules.**



clapper >
sides of bell >
adjacent air molecules >
further air >
eardrum

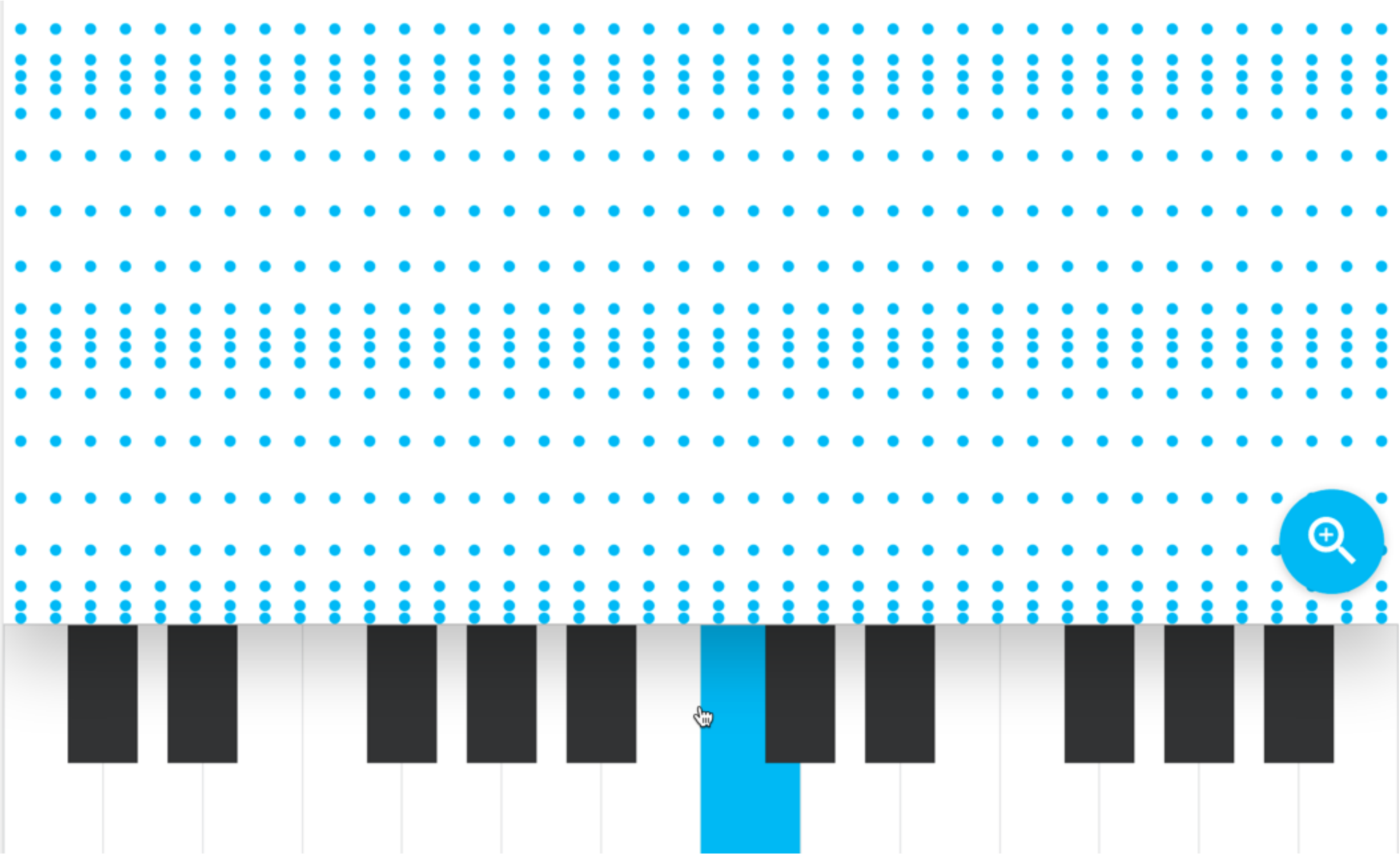
# WHAT ARE SOUND WAVES?

**Sound waves are pressure variations over time in the air transmitted by the movements of air molecules.**

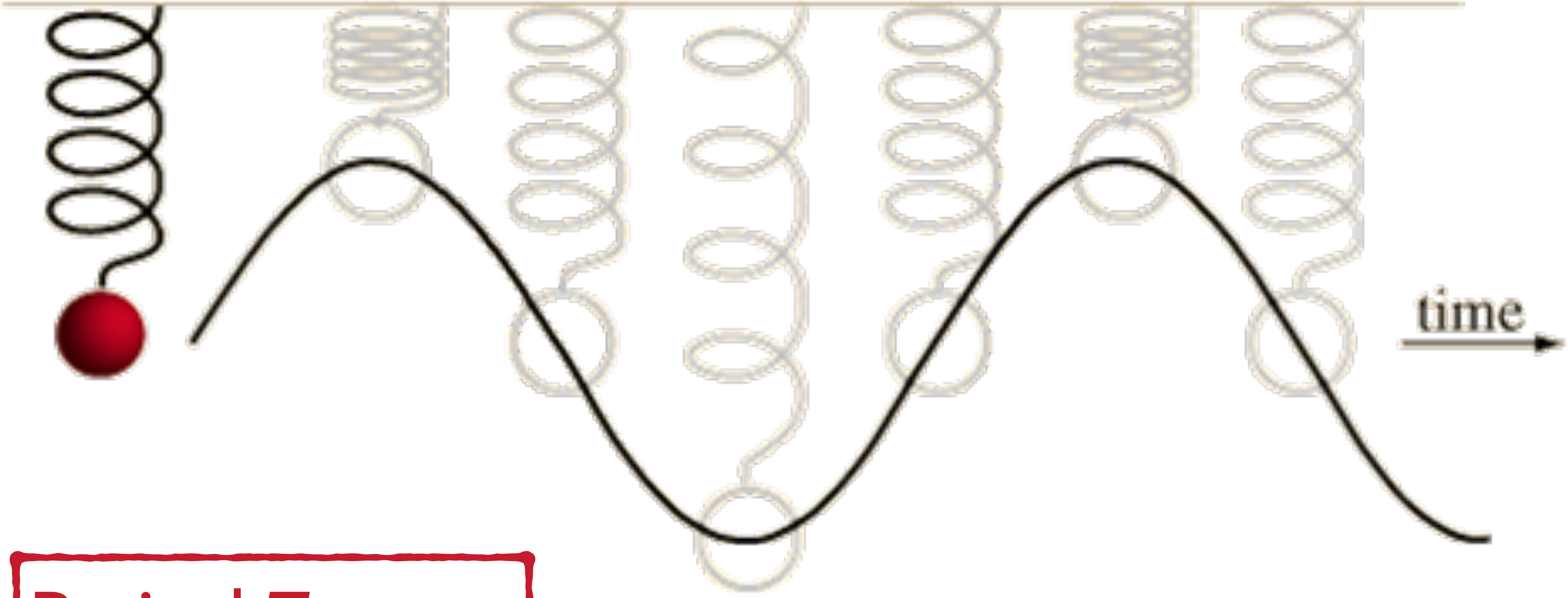


# SOUND WAVES HANDS-ON

https://musiclab.chromeexperiments.com/Sound-Waves



# SINE WAVES: SIMPLE HARMONIC MOTION



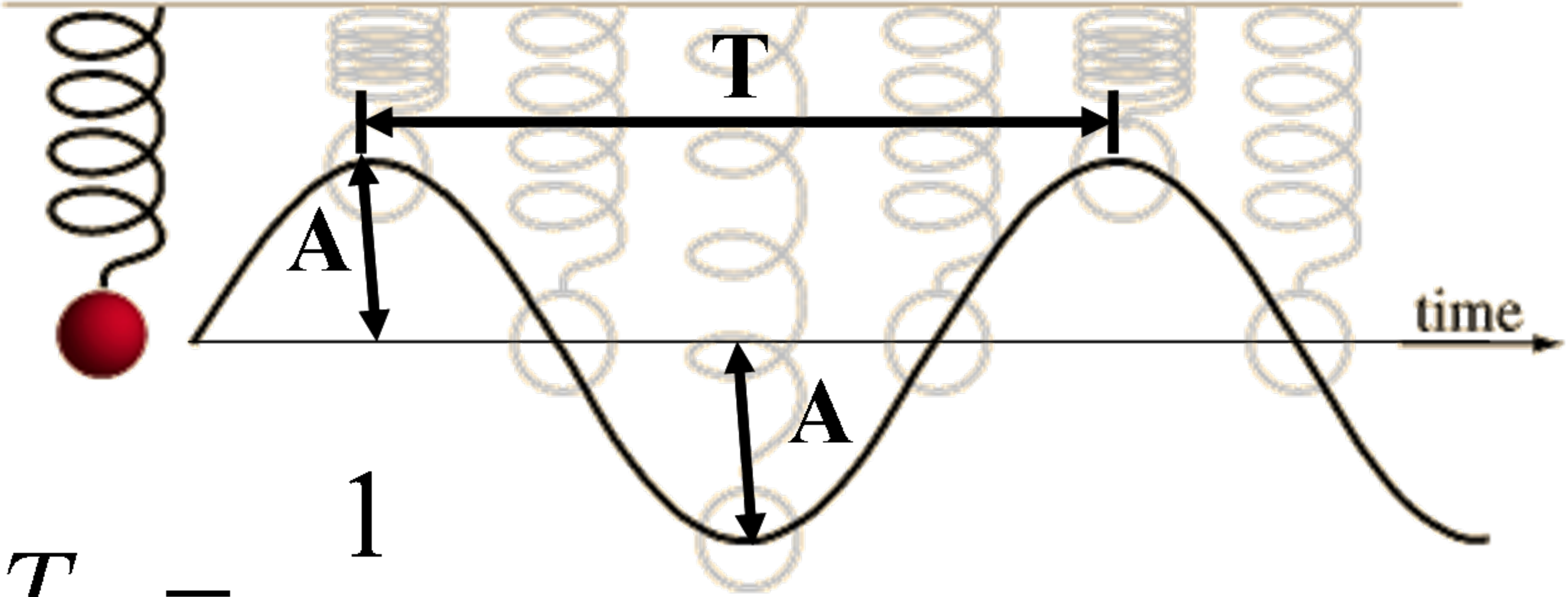
Period *T*

Amplitude *A*

Frequency *f*

http://www.gailruby.com/SHMGraph.htm

# SINE WAVES: SIMPLE HARMONIC MOTION

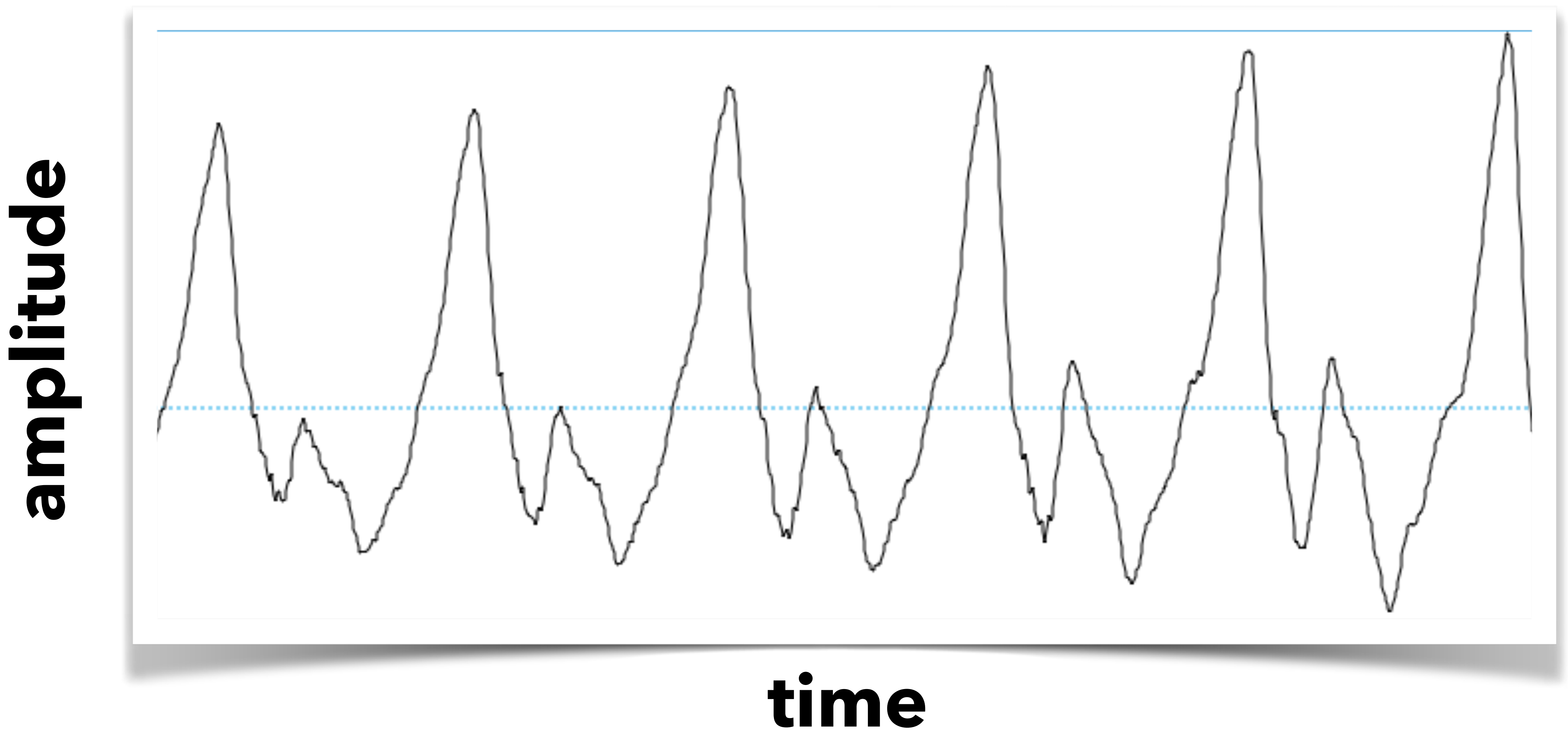


$$T = \frac{1}{f}$$

http://www.gailruby.com/SHMGraphwithrelations.htm

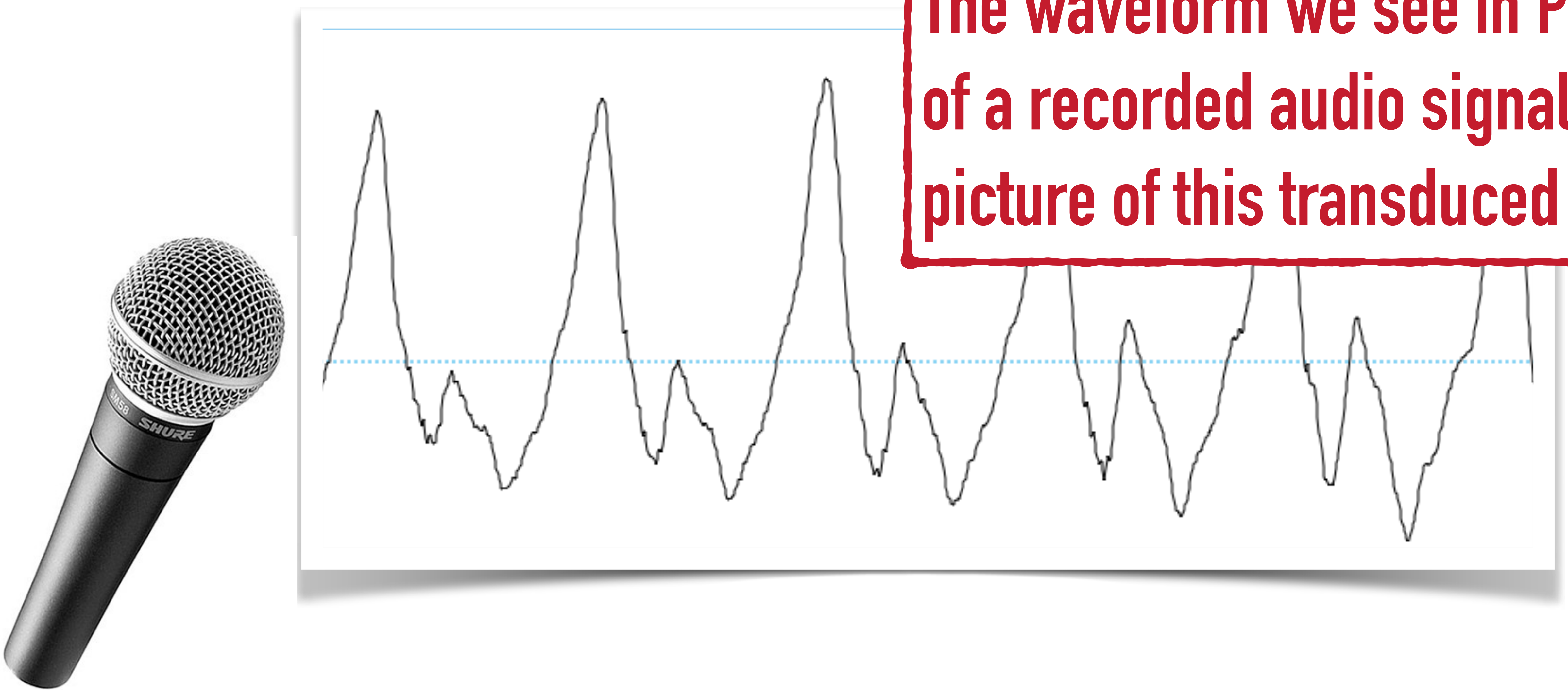
# IT'S AN ANALOG WORLD

**These pressure variations are continuous in both dimensions ("analog").**



# THE WAVEFORM

- A microphone picks up the air pressure waves, like an eardrum, but then **transduces** the sound energy from air pressure into an electrical voltage.



The waveform we see in Praat of a recorded audio signal is a picture of this transduced wave

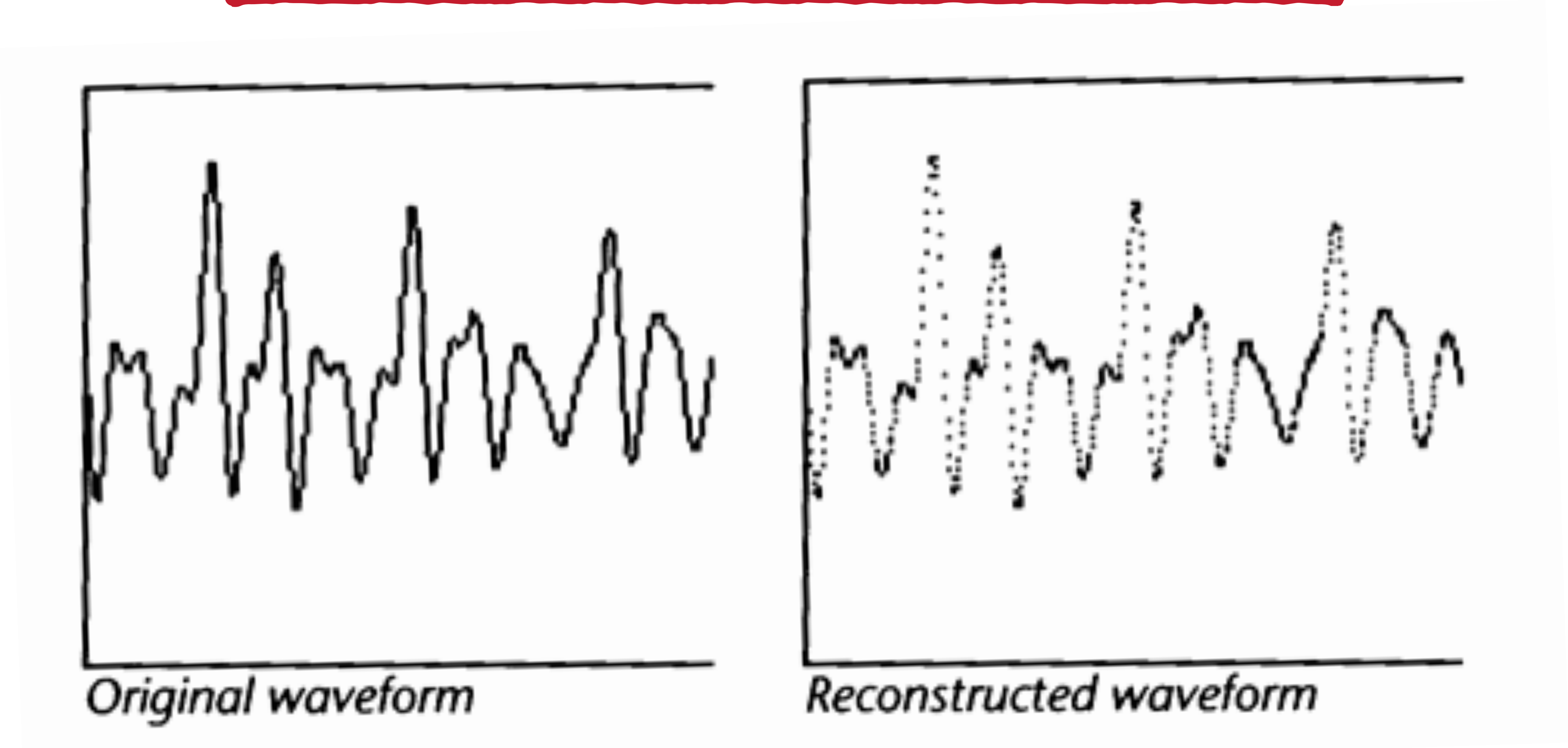
# IT'S A DIGITAL WORLD

- But computers store information digitally as 1s and 0s!
- So how do we get between an analog signal and a digital signal?

# ANALOG ⇄ DIGITAL

Analog-to-digital conversion (A/D) →

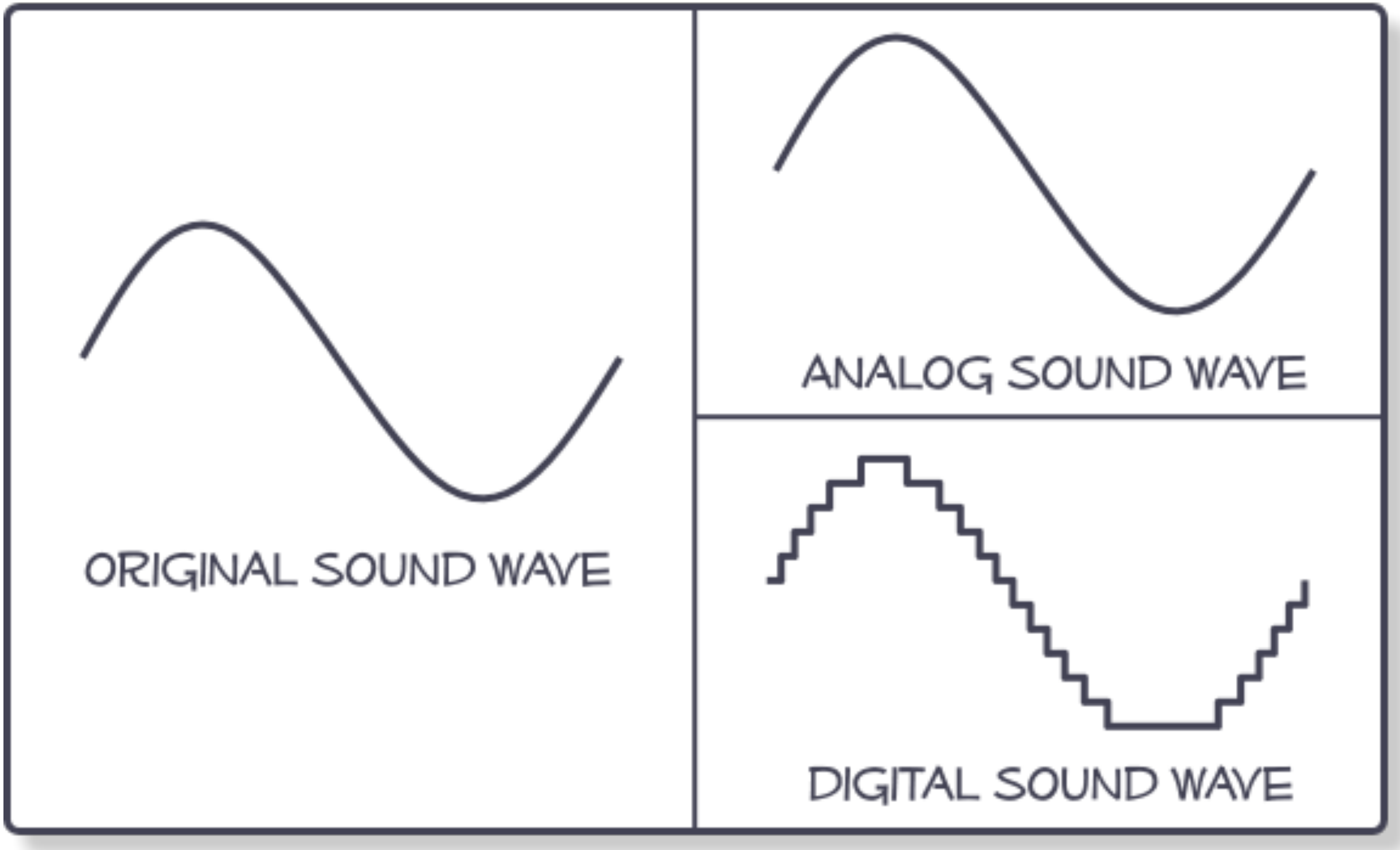
Digital-to-analog conversion (D/A) ←



**Analog: continuous**

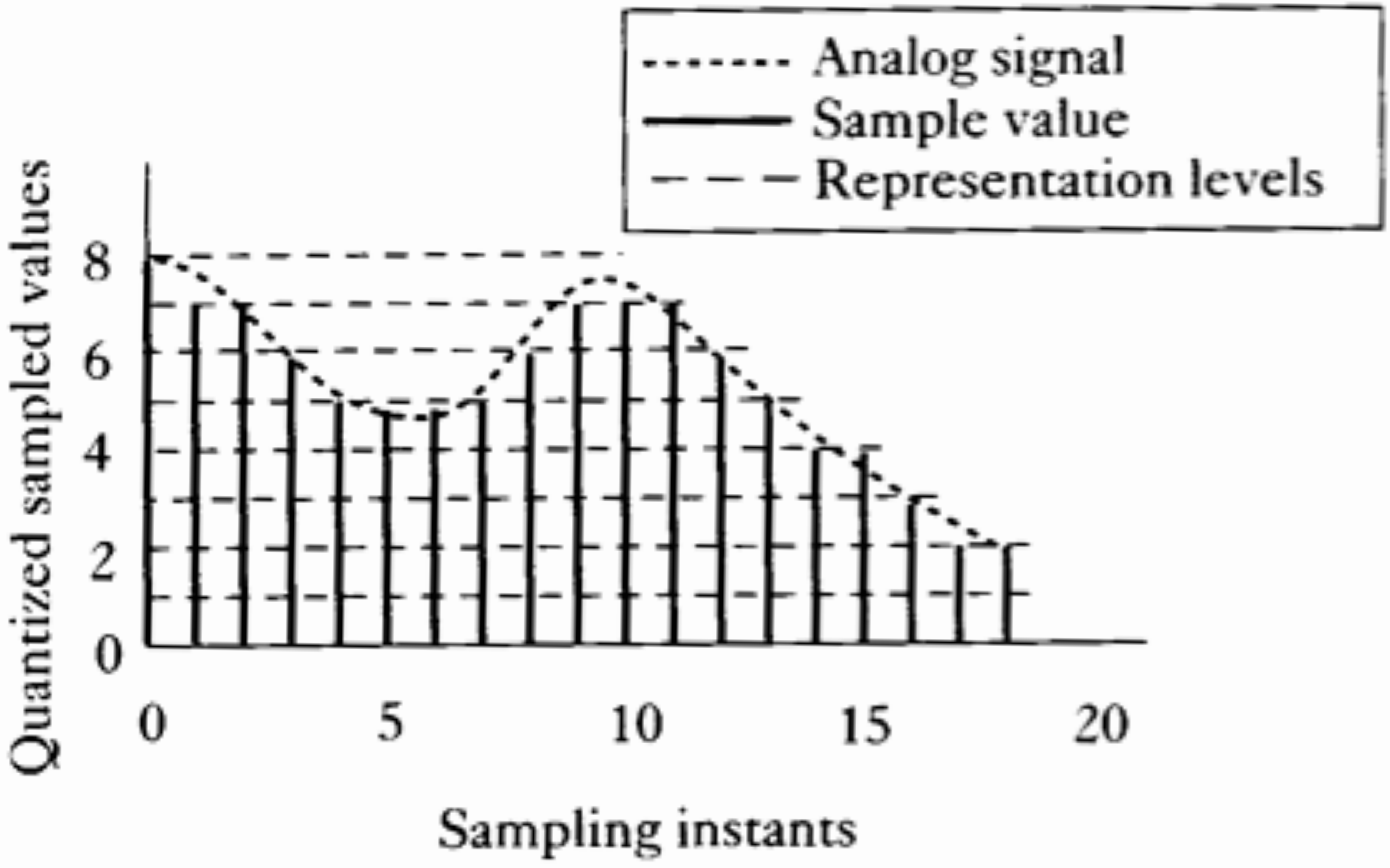
**Digital: discrete**

# TRANSDUCTION. THEN A/D CONVERSION.



http://www.centerpointaudio.com/Images/Analog-Digital%20frequency%20examples.png

# A/D CONVERSION BY SAMPLING

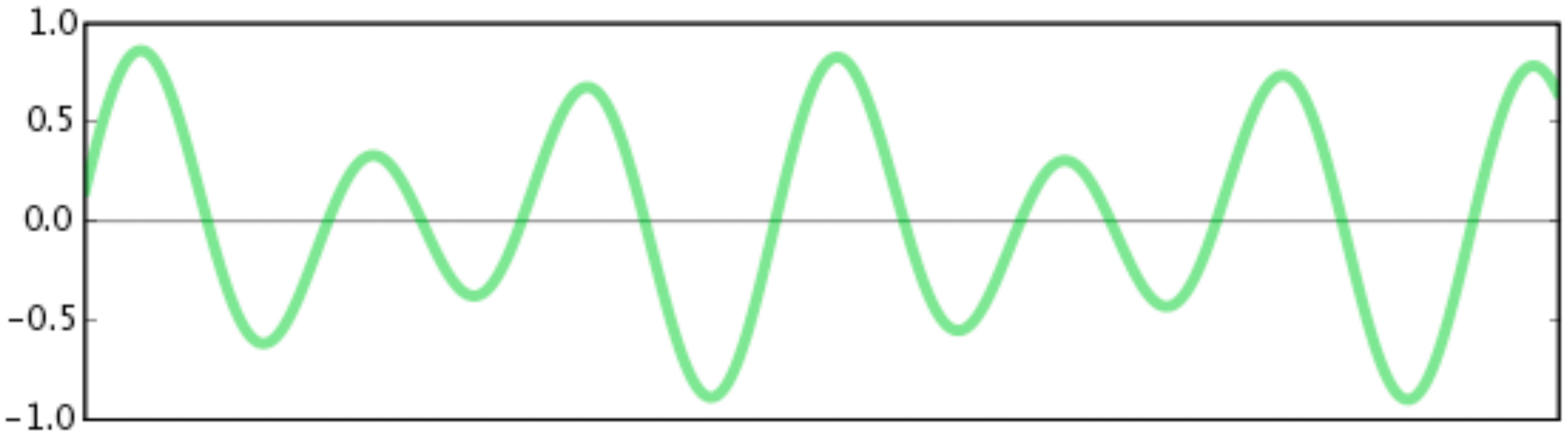


(a) Quantized sampling with 8
representation levels (3 bits per sample).

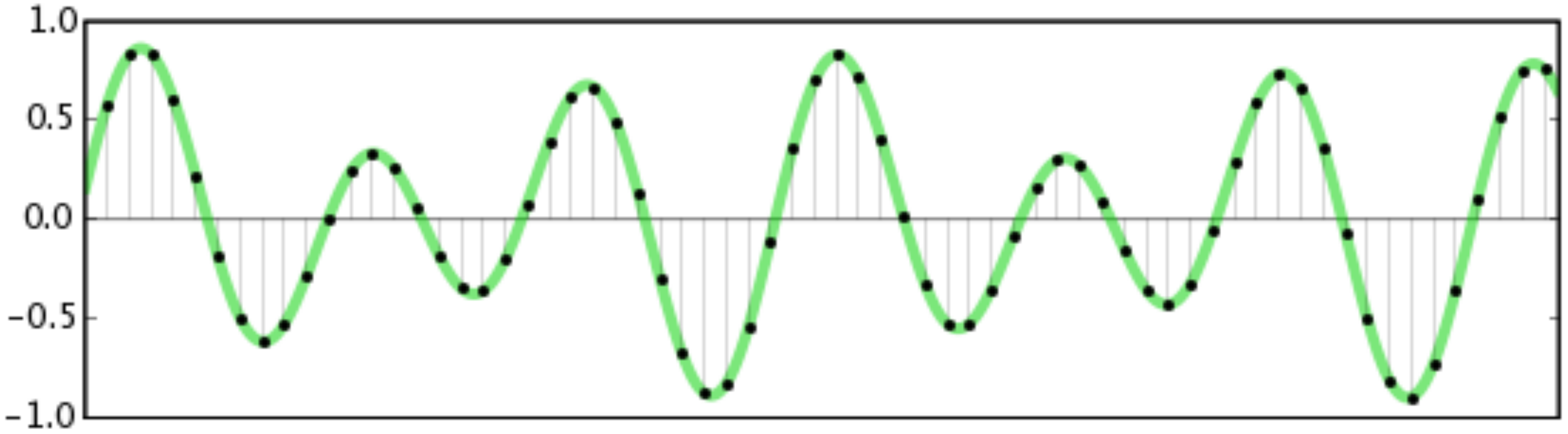
# SAMPLING IN TIME

- Horizontal dimension (x-axis) represents time
- Choice for sampling interval: ***at what time interval T are points sampled form the signal?***
  - Every 10 millisecond (ms)? *T = 10ms*
  - Every second (s)? *T = 1s*

# EXAMPLE: EVENLY SPACED SAMPLING INTERVALS



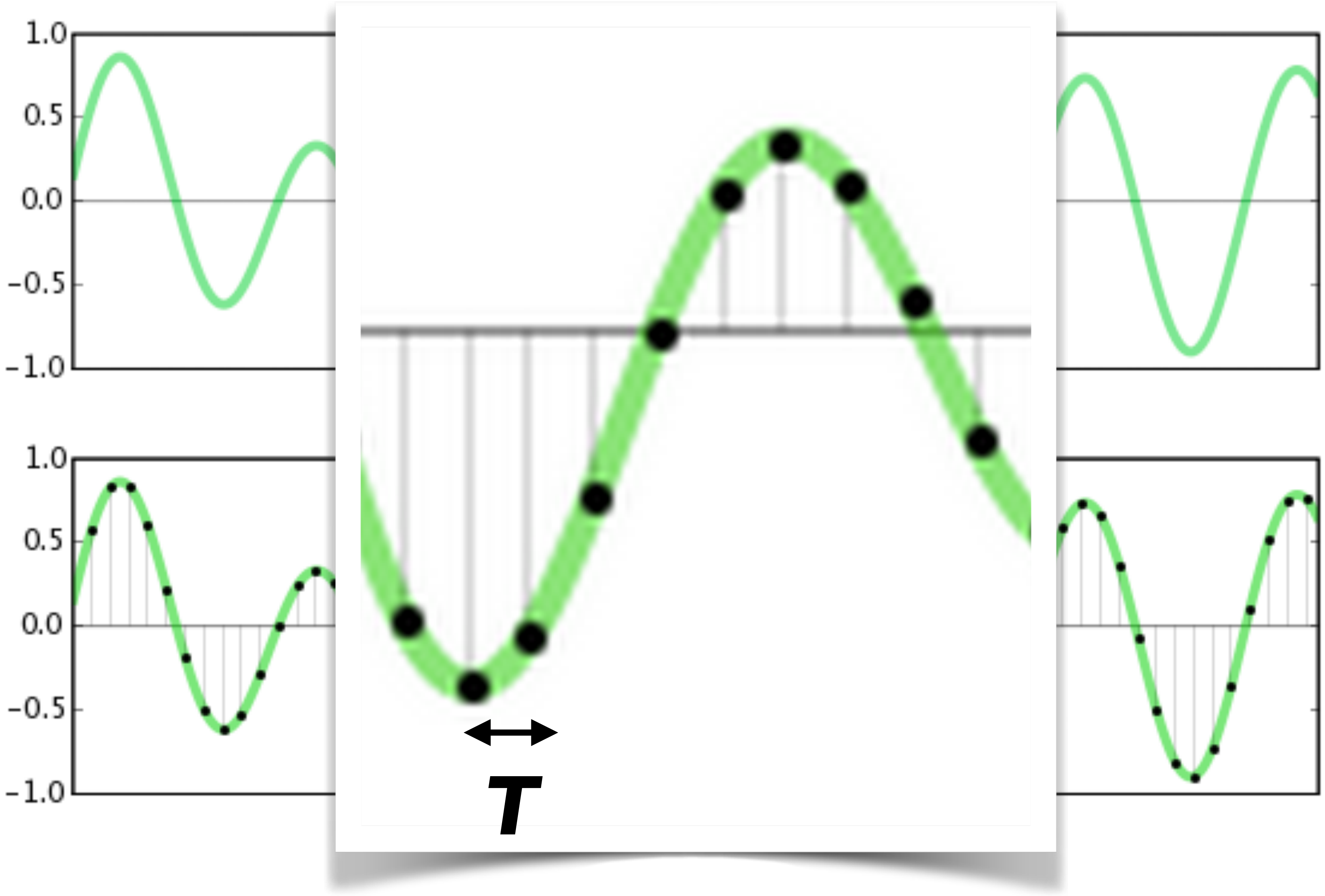
**Continuous waveform**



**Sampling values at time interval *T***

[http://manual.audacityteam.org/o/man/digital_audio.html](http://manual.audacityteam.org/o/man/digital_audio.html)

# EXAMPLE: EVENLY SPACED SAMPLING INTERVALS



**Continuous waveform**

**Sampling values at time interval *T***

http://manual.audacityteam.org/o/man/digital_audio.html

# SAMPLING RATE (OR SAMPLING FREQUENCY)

- Sampling rate: **how frequently, per unit time, are samples taken?**

- Frequency units: typically, the number of samples **per second (“cycles per second”).**
- This unit is called **Hertz (Hz)**, cycles per second.
  - Example: “The sampling rate was 20 kilohertz (kHz)” means the sampling rate was 20,000 Hz or 20,000 cycles per second
  - (kilo- = 1000, 1 kHz = 1000 Hz)

# CALCULATING SAMPLING RATE

- **Frequency = 1/[time]**
  - *Example: how often do you finish drinking a glass of water?*
  - 1/ [2 hours to finish a glass]

    = 1/2 glass drunk per hour
- **Sampling rate = 1 / [sampling interval] = 1 / T**
  - *Example: if T = 10ms, sampling rate is 1 / 10 ms, "1/10 of a cycle every millisecond"*
  - *Example: if T = 10s, sampling rate is 1/ 10 s, "1/10 of a cycle every second", 0.1 Hz*
- Hz = cycles/second, so…

**sampling rate in Hz = 1 / sampling interval in seconds***

* "seconds" is usually just abbreviated as "s"

# EXERCISE: CONVERTING FROM INTERVAL TO FREQUENCY

- 1 sample every second:
  - sampling interval $T = 1s$
  - sampling rate $f = 1\ Hz$
- 1 sample every 0.1 second:
  - sampling interval $T =$ **?** $s$
  - sampling rate $f =$ **?** $Hz$
- 1 sample every 0.01 second:
  - sampling interval $T =$ **?** $s$
  - sampling rate $f =$ **?** $Hz$

# EXERCISE: SAMPLING INTERVALS IN MILLISECONDS

- 1 millisecond = 1 msec<br>
  = 1 ms<br>
  = 0.001 s (= 1/1000 s)

- 1 sample every millisecond:
  - How many samples per second?
  - sampling interval $T$ = **?** *s*
  - sampling rate $f$ = **?** *Hz*

# EXERCISE: SAMPLING AT DIFFERENT RATES

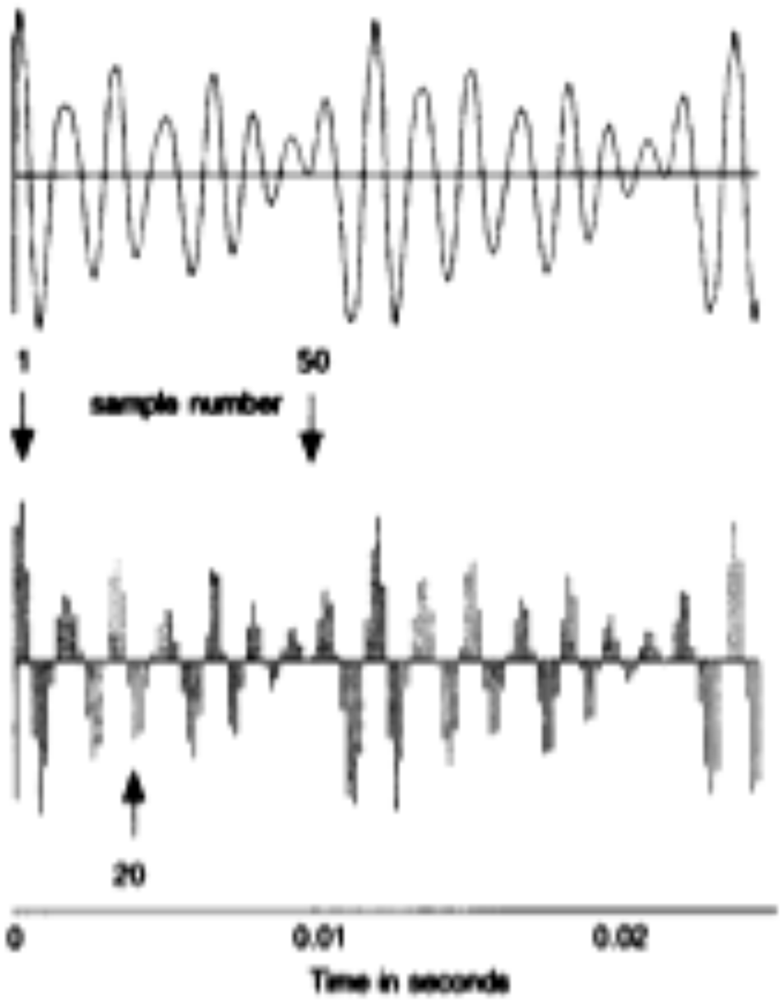
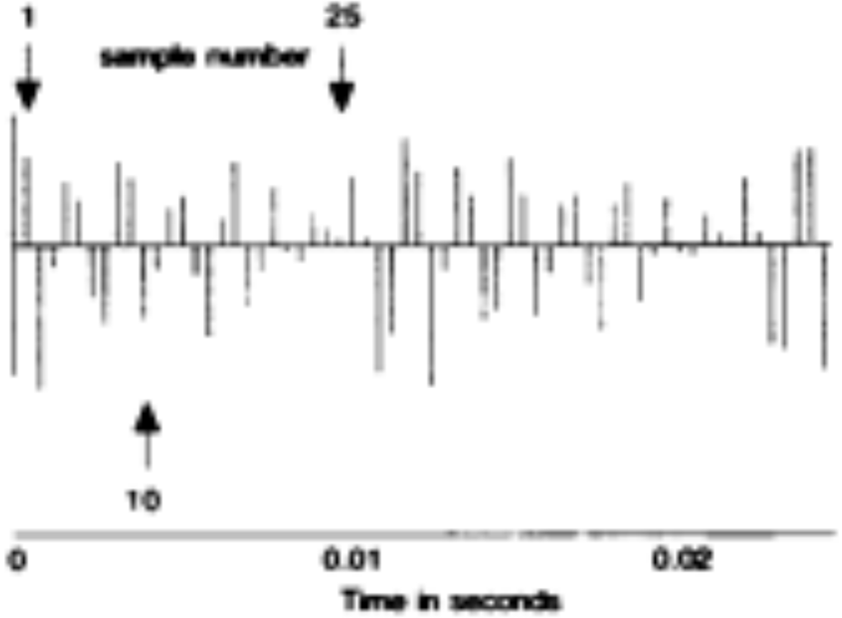


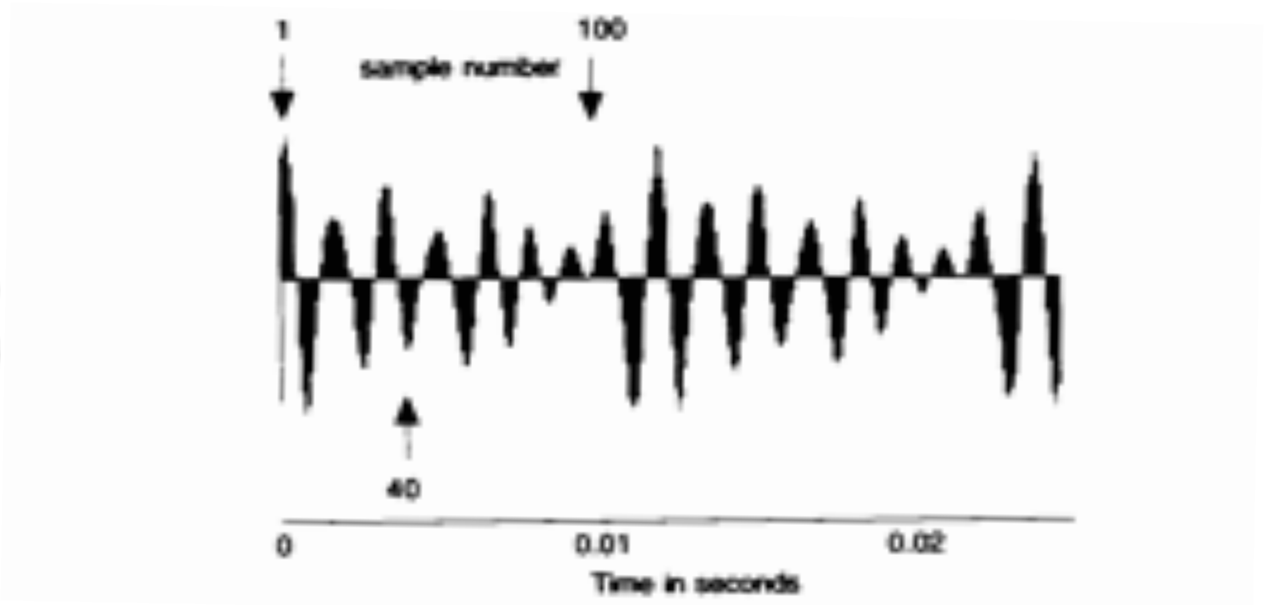
Fig. 9.1. Top: a continuously varying voltage recorded during the production of the vowel [ɑ]; below: lines corresponding to the size of this voltage (the amplitude) at regular intervals of time.

(A/D) →

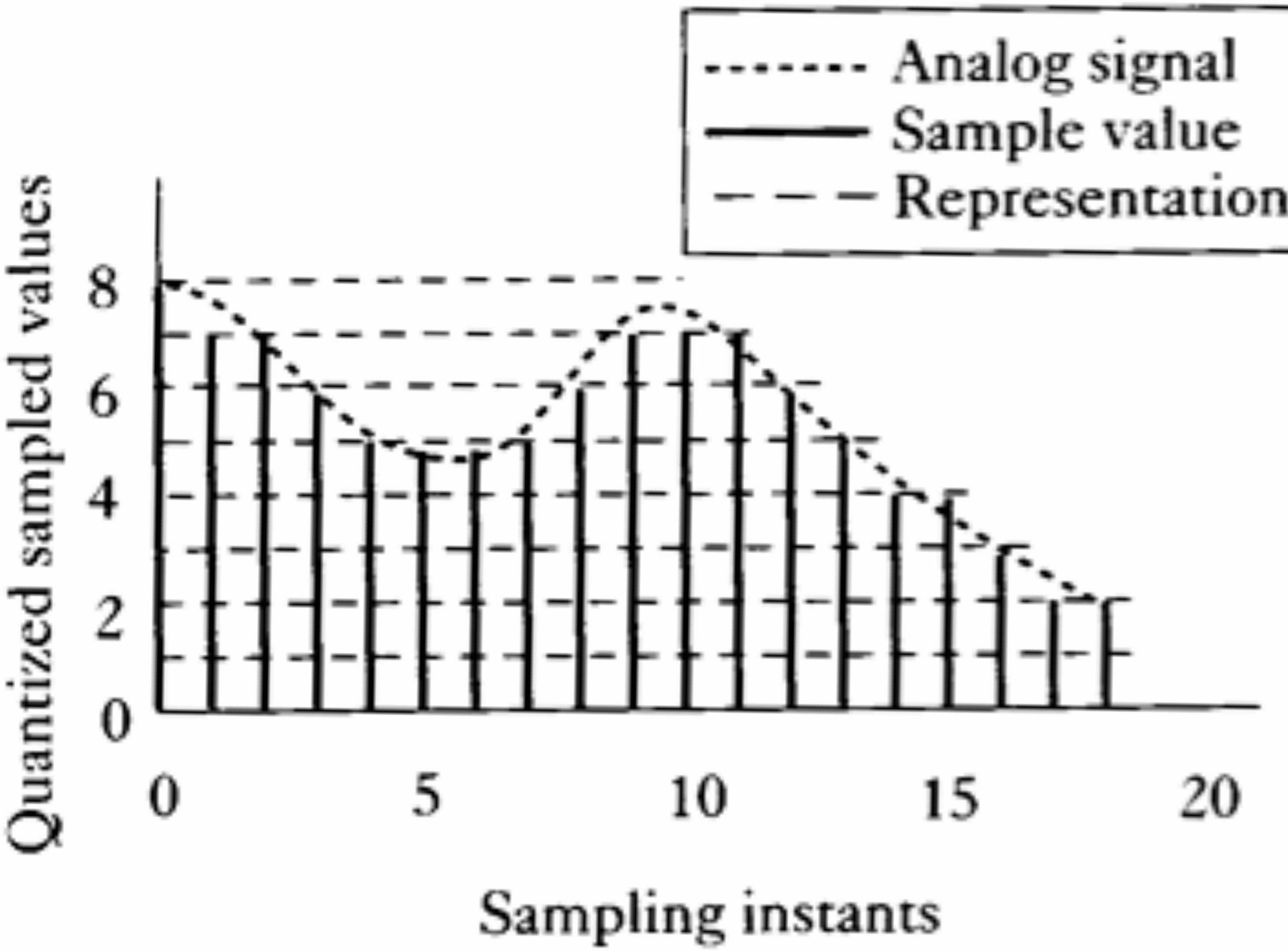
2,500 Hz (? kHz)



10,000 Hz (? kHz)



# QUANTIZATION: “SAMPLING” AMPLITUDE



(a) Quantized sampling with 8 representation levels (3 bits per sample).

Vertical dimension (y-axis) is divided into some number of values: here, 8 values ($2^3$)

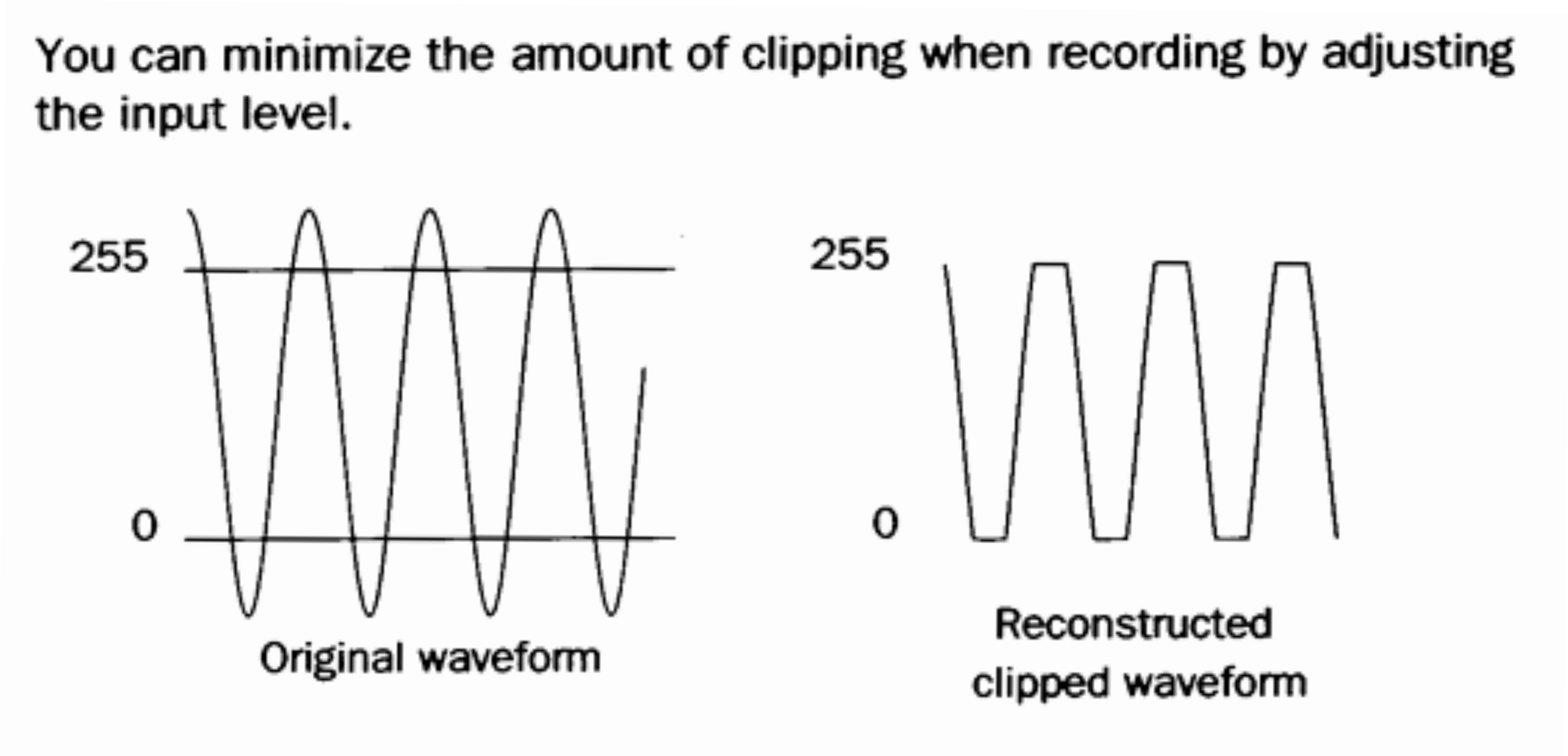
Quantization rate (bit depth) = 3 bits

# EXERCISE: QUANTIZATION RATE (BIT DEPTH)

- Divide amplitude range into 65,536 values
  - 65,536 = $2^{16}$
  - Quantization rate = **?** bits
- Divide amplitude range into 4096 values
  - 4096 = $2^{?}$
  - Quantization rate = **?** bits
- Quantization rate = 4 bits
  - Amplitude range divided into **?** values

# CLIPPING

**Clipping: the peak amplitude of the sample is greater than the largest amplitude allowed by the quantization.**



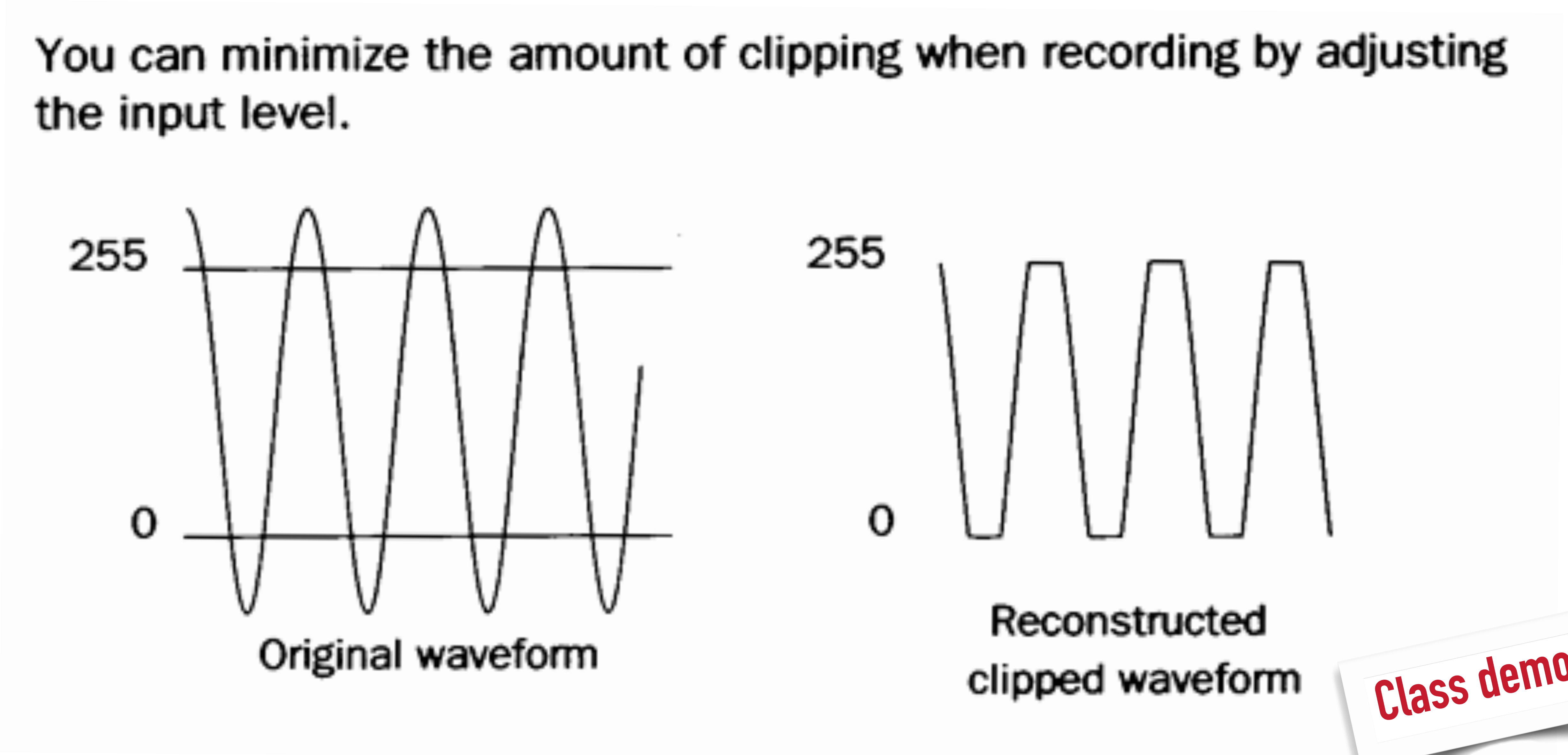
You can minimize the amount of clipping when recording by adjusting the input level.

Original waveform

Reconstructed clipped waveform

# CLIPPING

**<u>Clipping</u>: the peak amplitude of the sample is greater than the largest amplitude allowed by the quantization.**



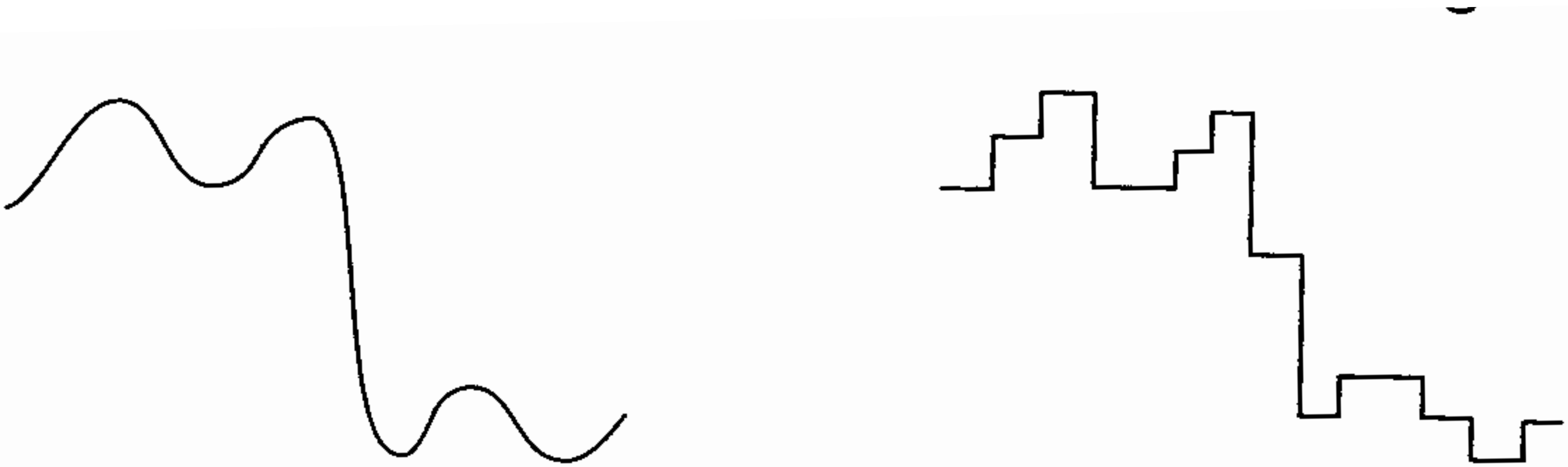
You can minimize the amount of clipping when recording by adjusting the input level.

Original waveform

Reconstructed clipped waveform

Class demo!

# TOO FEW AMPLITUDE LEVELS?



Original waveform

Reconstructed quantized waveform

# MORE SAMPLES ⇒ BETTER AUDIO QUALITY

- Higher **sampling** rate gives better **time** resolution

- Higher **quantization** rate gives better **amplitude** resolution

# WHY NOT EVEN HIGHER RATES?

**Trade-off between fidelity to original signal and demands on memory and processing of the stored string of numbers**

- Higher sampling rate: more numbers to be stored, retrieved and processed
- Higher quantization rate: bigger numbers to be stored, retrieved, and processed