Deep Learning for Speech Processing and Information Retrieval

Danna Gurari
University of Colorado Boulder
Fall 2022

https://home.cs.colorado.edu/~DrG/Courses/NeuralNetworksAndDeepLearning/AboutCourse.html

Review

- Previous few lectures:
  - Guest speakers from industry and academia
- Assignments (Canvas):
  - Final project outline due today
  - Final project presentation due in two weeks
    - Poster presentation
    - Video
- Questions?

Today’s Topics

- Speech Processing – Problem and Applications
- Speech Recognition – Evaluation and Models
- Informal Retrieval – Problem and Applications
- Informal Retrieval – Models
- Video making tutorial

Today's Topics

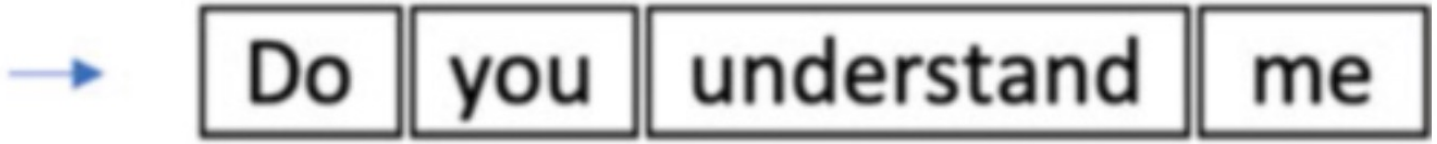
- Speech Processing – Problem and Applications
- Speech Recognition – Evaluation and Models
- Informal Retrieval – Problem and Applications
- Informal Retrieval – Models
- Video making tutorial

Problem Definition

Input: spoken language

Output: machine readable text

Raw Speech Signal



Transcription

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

What Is Speech?

Raw Speech Signal

Compression waves created by pushing air from one's lungs and modulating it using one's tongue, teeth, and lips

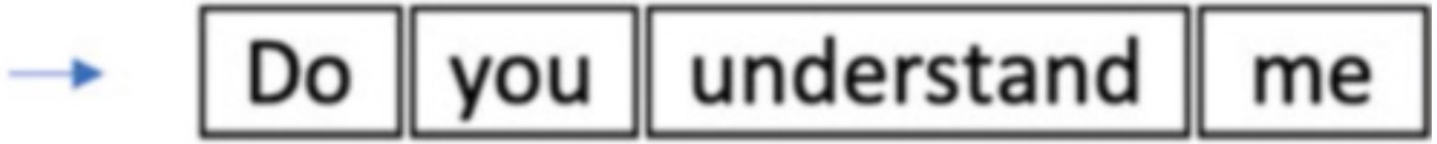
Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Why Is Speech Processing Challenging?

Input can be diverse
including different accents,
volumes, pace, and cadence

Temporal data needs to be
segmented into distinct words

Raw Speech Signal



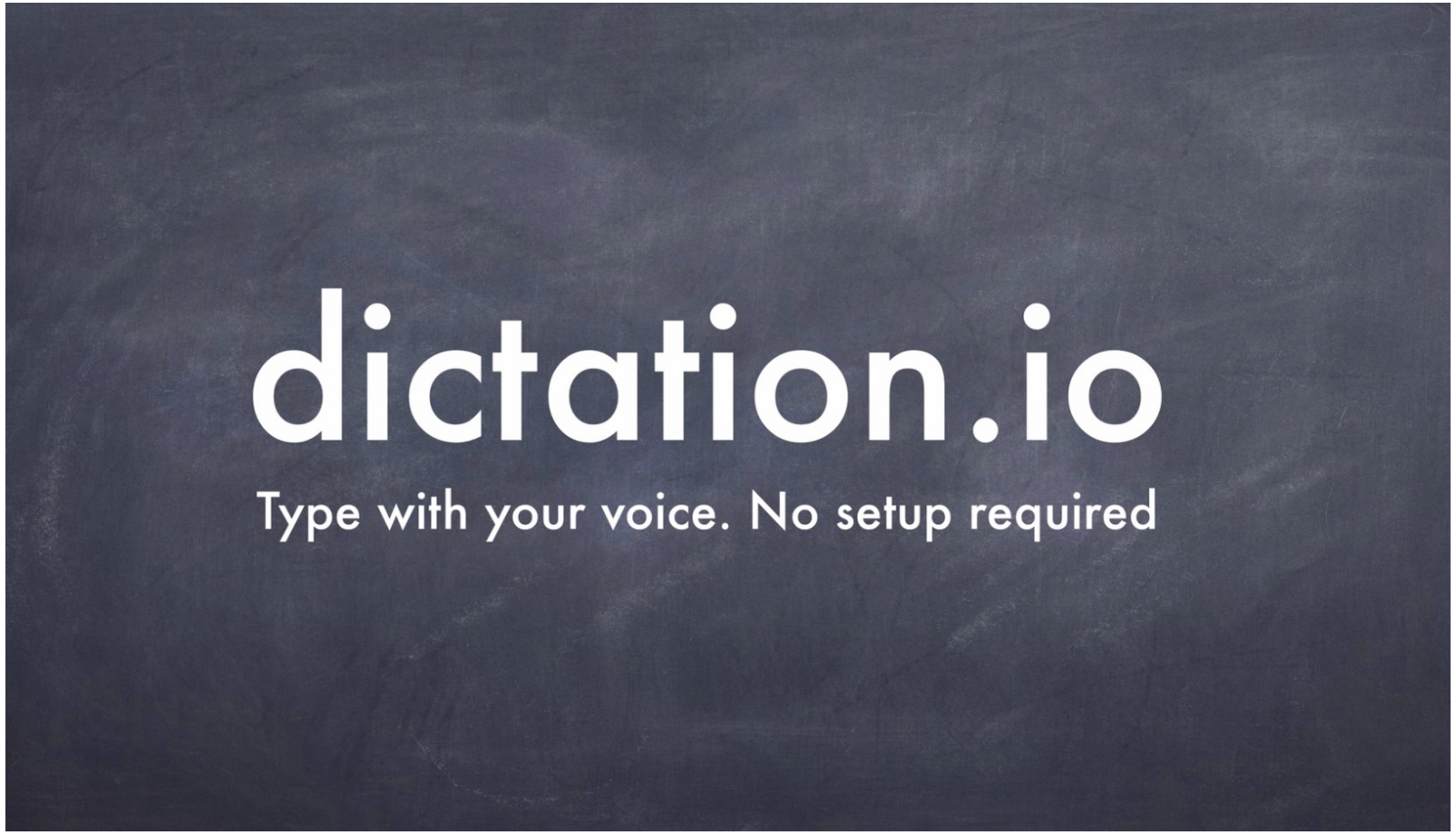
Transcription

Technology can result in many
artifacts including varying quality,
echos, and background noise

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Voice Typing on Mobile Devices

Voice Typing for Productivity Applications



Demo starting at 2:00: https://www.youtube.com/watch?v=5UK4vLzU9co&t=76s

Virtual Assistant

e.g., Amazon’s Echo with Alexa

e.g., Google Home

e.g., Baidu DuerOS

https://dueros.baidu.com/en/html/dueros/index.html

Virtual Assistant

Entertainment Video	Music, movie, television shows, variety show, short clip, audio book, and broadcasting, etc.	Information Inquiry	Cnsultation, weather, stocks, flight, sports (NBA), FAQ, cookbook, images, etc.
Lifestyle Services	Food, movie, take-out, hotel, shopping, taxi, cleaning service, travel, relaxation, and other O2O services.	Travel Conditions	Map, route, road condition, traffic restriction, endorsement, and surrounding environment query, etc.
Utility Tools	Ranslation, time, calculation, exchange rate, and unit conversion, etc.	Personal Assistant	Schedule management, alarm clock, reminder, memo and notepad, etc.
Learning	Encyclopedia, story, nursery rhyme, idiom, parenting, poetry and library, etc.	Chat and Relax	Chat, joke, poetry, idioms, and games, etc.

https://dueros.baidu.com/en/html/dueros/index.html

Audio Transcription (e.g., for Analysis & Situational/Permanent Hearing Impairments)

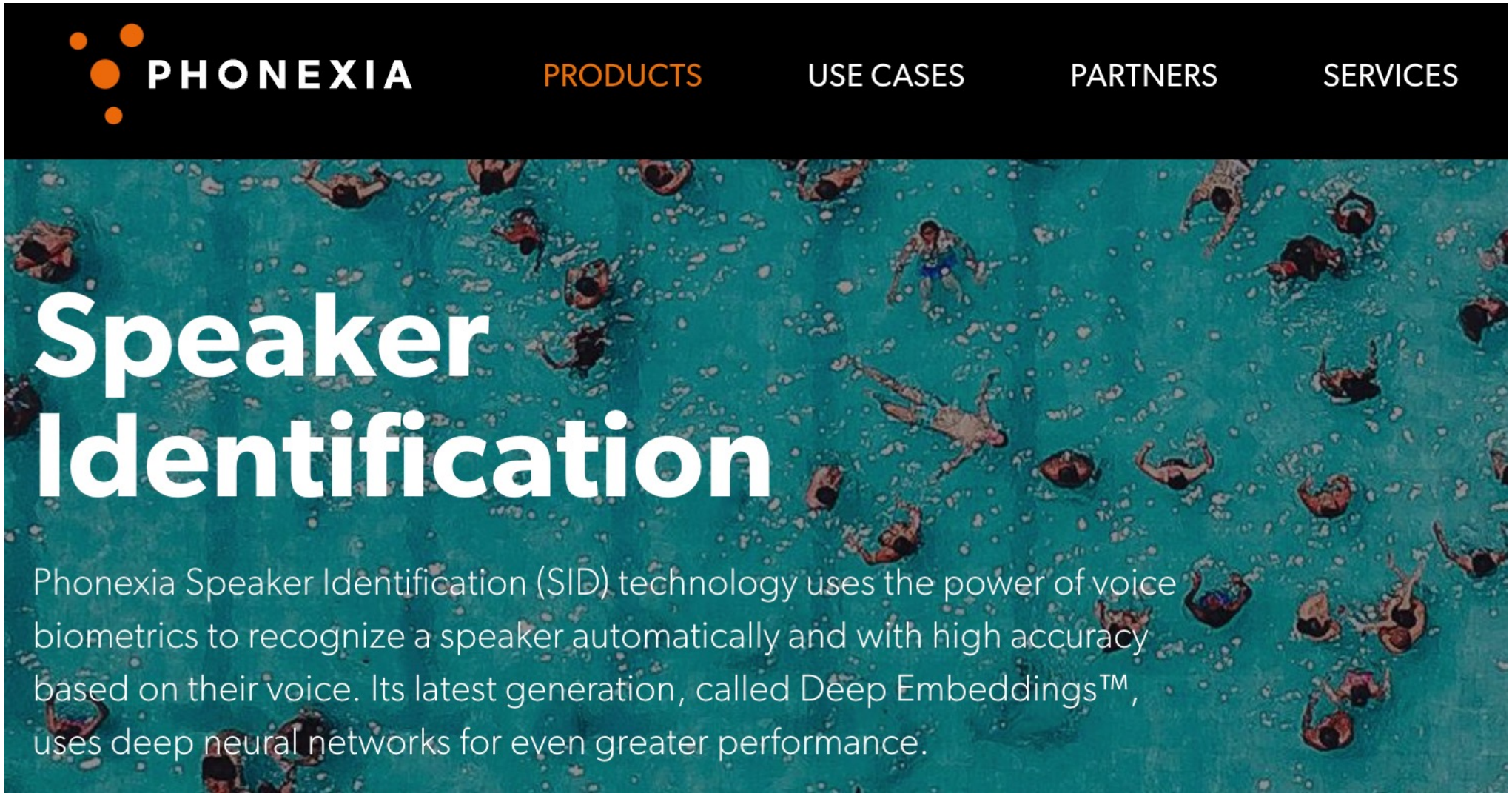
https://www.gmrtranscription.com/blog/podcast-transcription

Video/Movie Captioning (e.g., for Translation & Situational/Permanent Hearing Impairments)

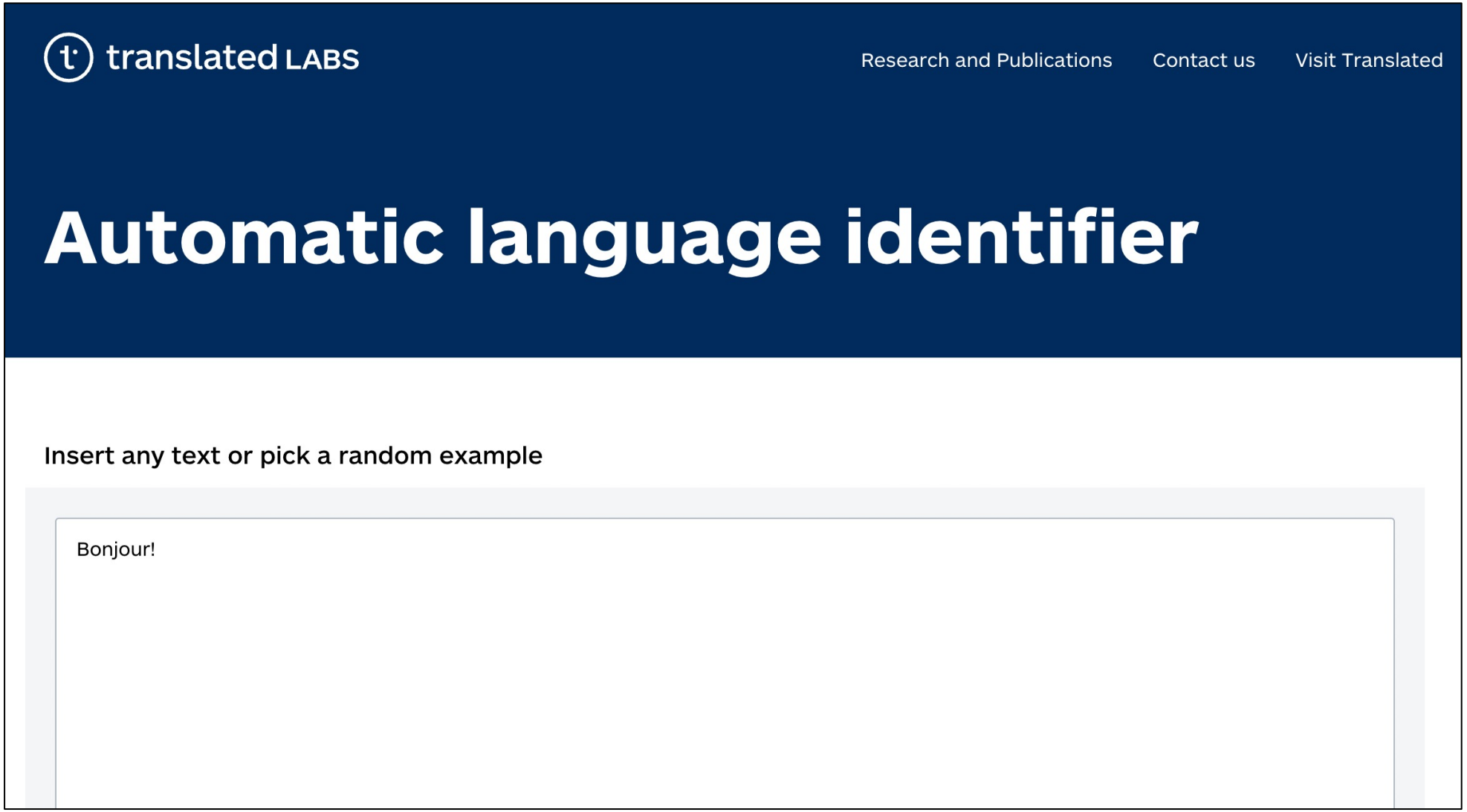
https://www.techsmith.com/blog/add-captions-subtitles-video/

Speech Emotion Recognition (e.g., for Help Desks and Negotiators)

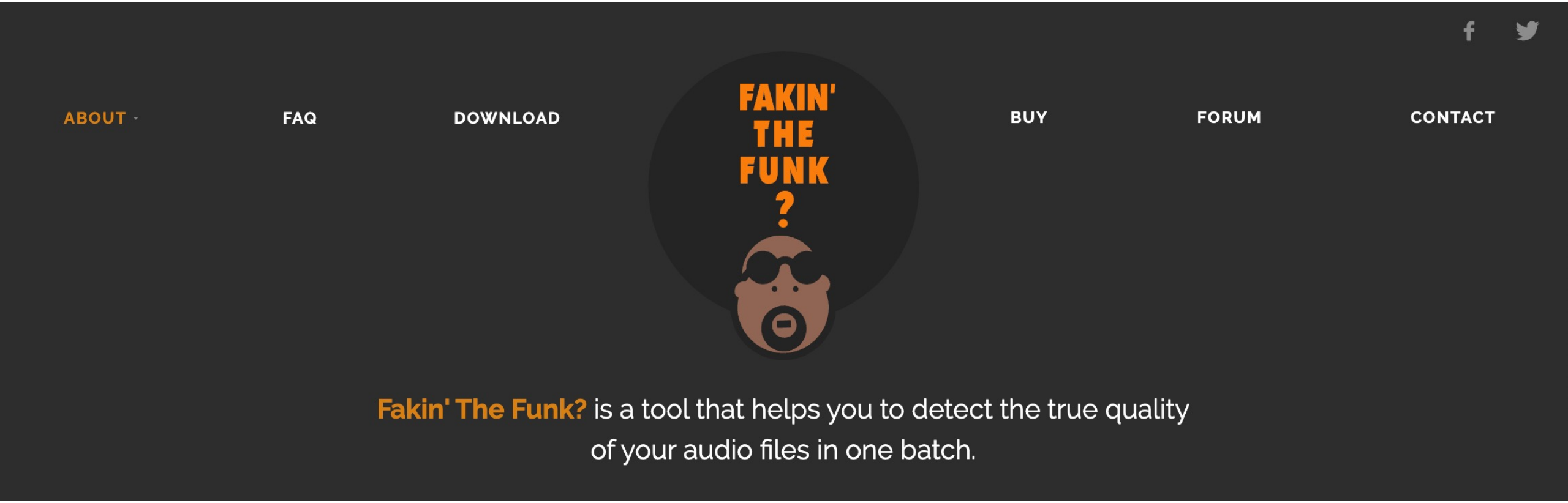
Speaker Identification (e.g., for Security)



Language Identification



Speech Enhancement



What are other potential applications for speech processing?

Today's Topics

- Speech Processing – Problem and Applications
- Speech Recognition – Evaluation and Models
- Informal Retrieval – Problem and Applications
- Informal Retrieval – Models
- Video making tutorial

Spectrum of Tasks

Commands — Discourse

(brief with small vocabulary) (lengthy with large vocabulary)

Word Error Rate

- Indicates edit distance between the prediction and the target as follows:

$$\frac{\overbrace{\text{Substitutions}}^{\text{word replaced (e.g., “goose” transcribed as “choose”)}} + \overbrace{\text{Insertions}}^{\text{word added (e.g., “caboose” transcribed as “the goose”)}} + \overbrace{\text{Deletions}}^{\text{word omitted (e.g., “it is a fun” transcribed as “it is fun”)}}}{\text{Number of Words Spoken}}$$

- What indicates better performance: larger or smaller values?

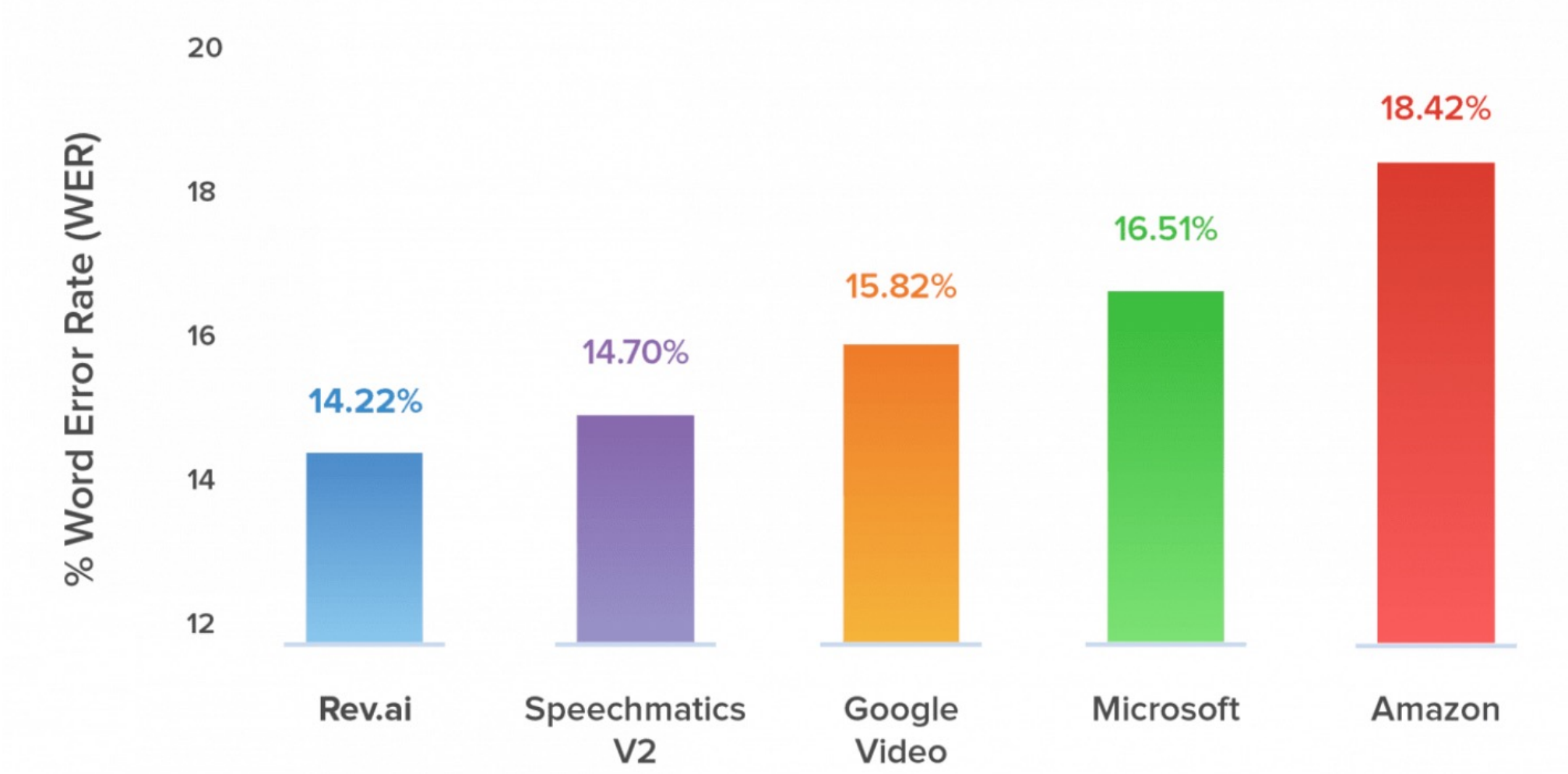
Word Error Rate: Example

- Correct: The sun makes it look like uh a warm, day to go outside to adventure.
- Predicted: The son makes it to bike with a swarm to go outside to Denver today.
- Number of words spoken?
  - 15
- WER?

$$\frac{\text{Substitutions + Insertions + Deletions}}{\text{Number of Words Spoken}}$$

$$\frac{6 + 1 + 1}{15} = 0.53$$

Word Error Rate: Comparison Example



https://www.rev.com/blog/resources/what-is-wer-what-does-word-error-rate-mean

What Are Limitations of Word Error Rate as an Evaluation Metric?

- Does not indicate why errors occur
  - Background noise (e.g., music, other talking)
  - Specialized language (i.e., words reflecting domain expertise)
  - Speaker pronunciations/accent

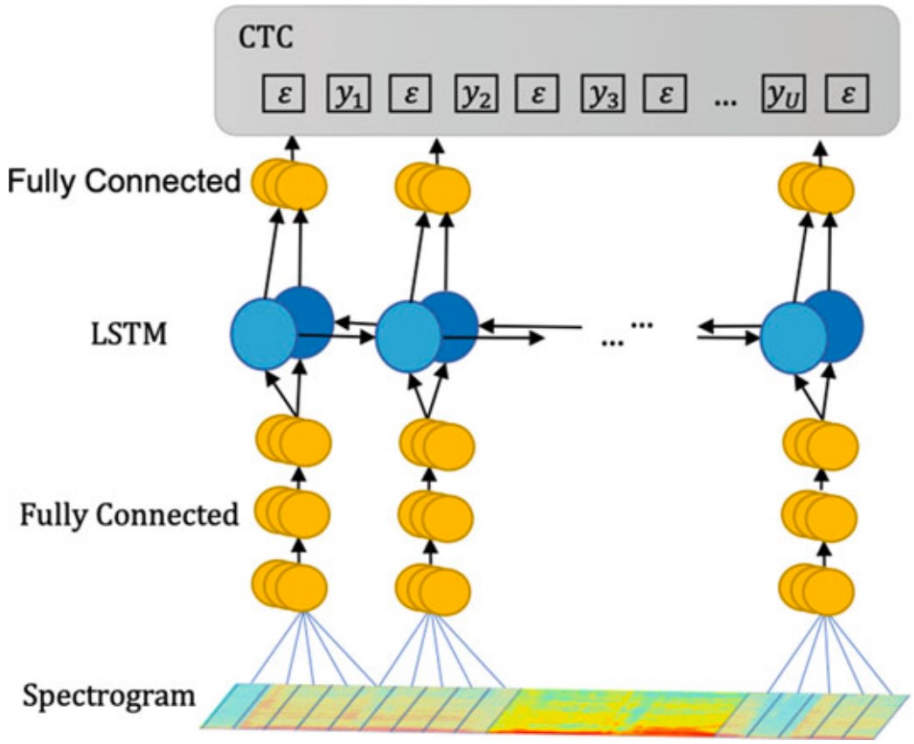
- Does not reflect whether transcription correctly captures:
  - Capitalization
  - Punctuation
  - Numbers
  - Paragraphs

- May indicate poor quality when humans could understand the content

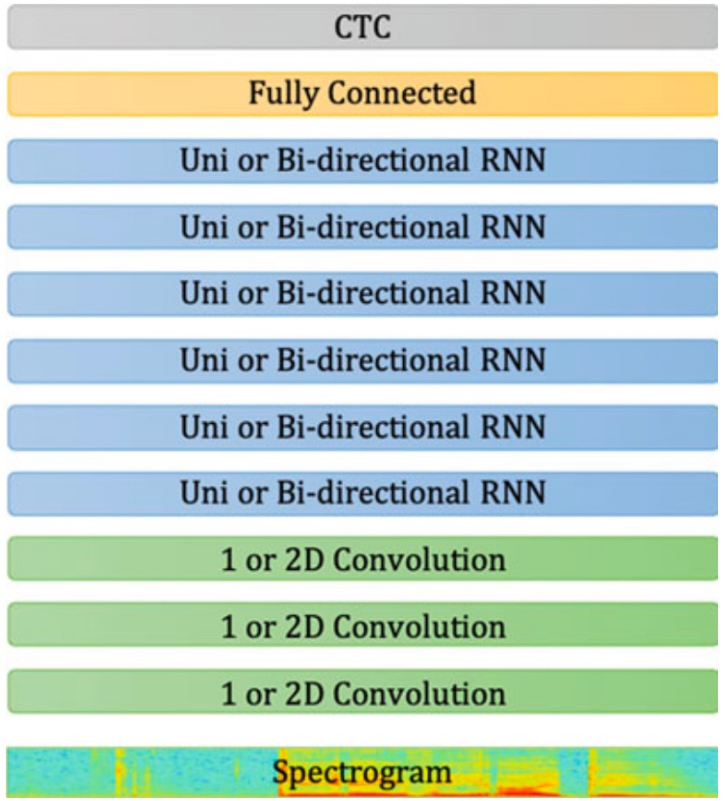
- Weights all word errors equally

Popular Methods

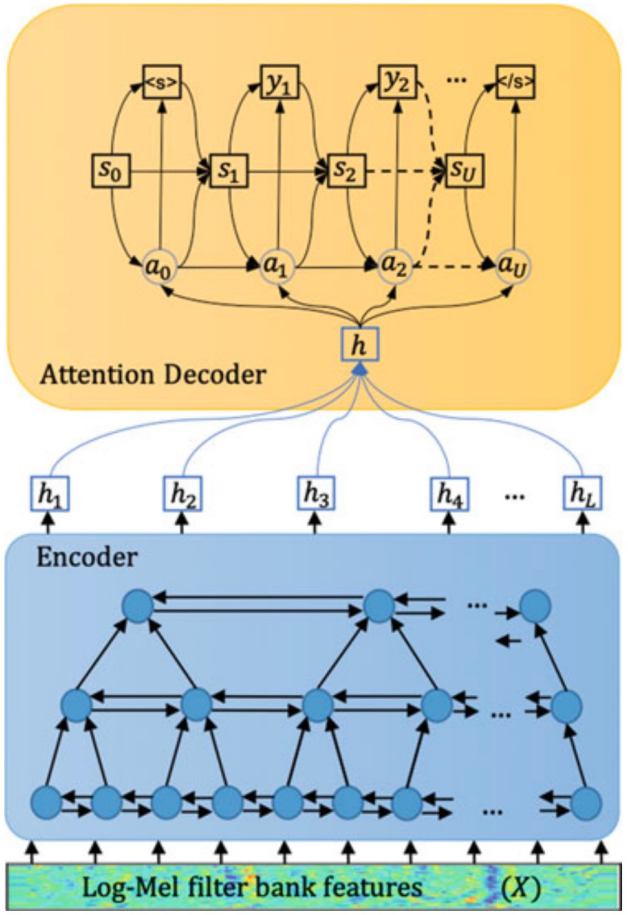
DeepSpeech



DeepSpeech2



Listen, Attend, and Spell

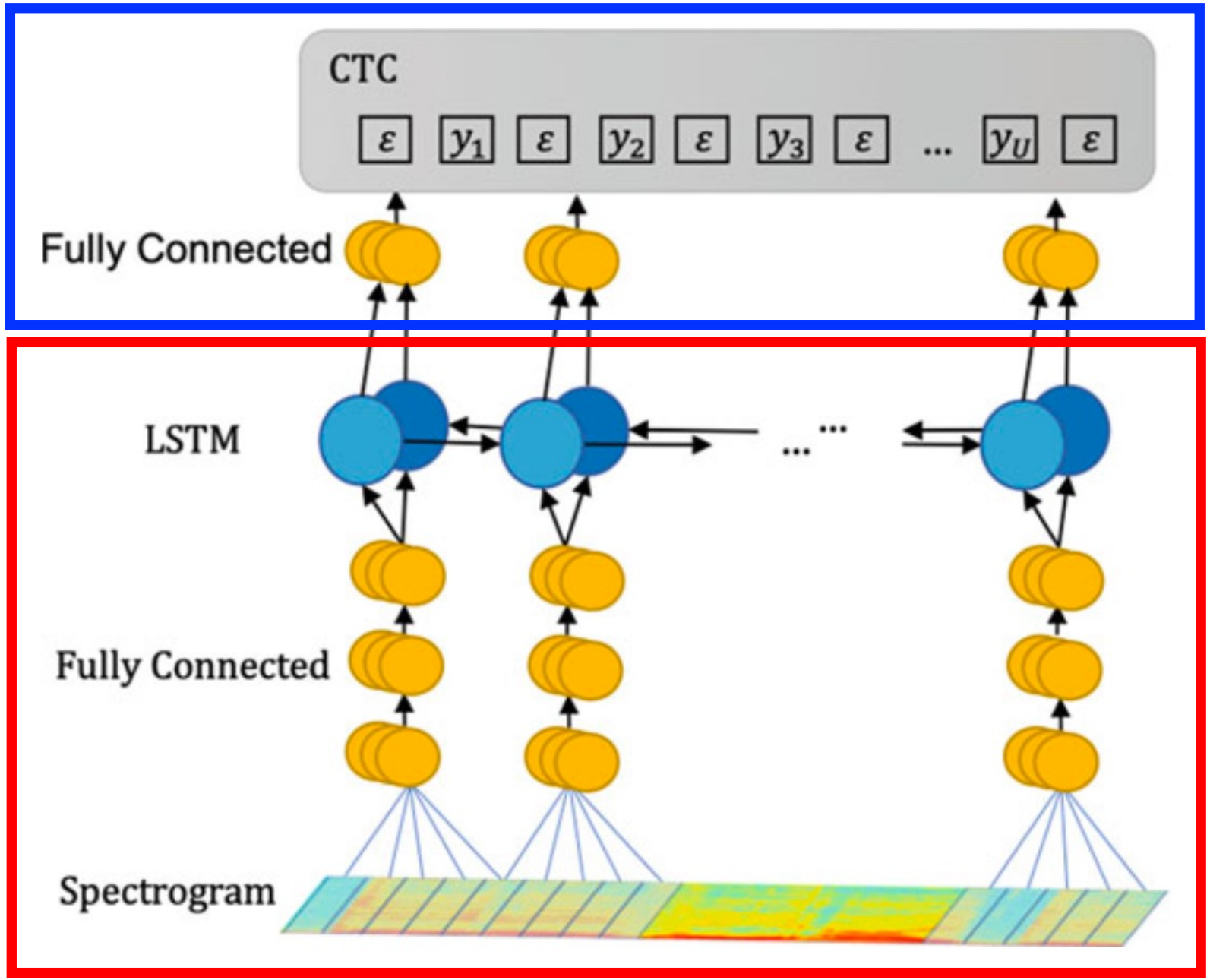


Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech

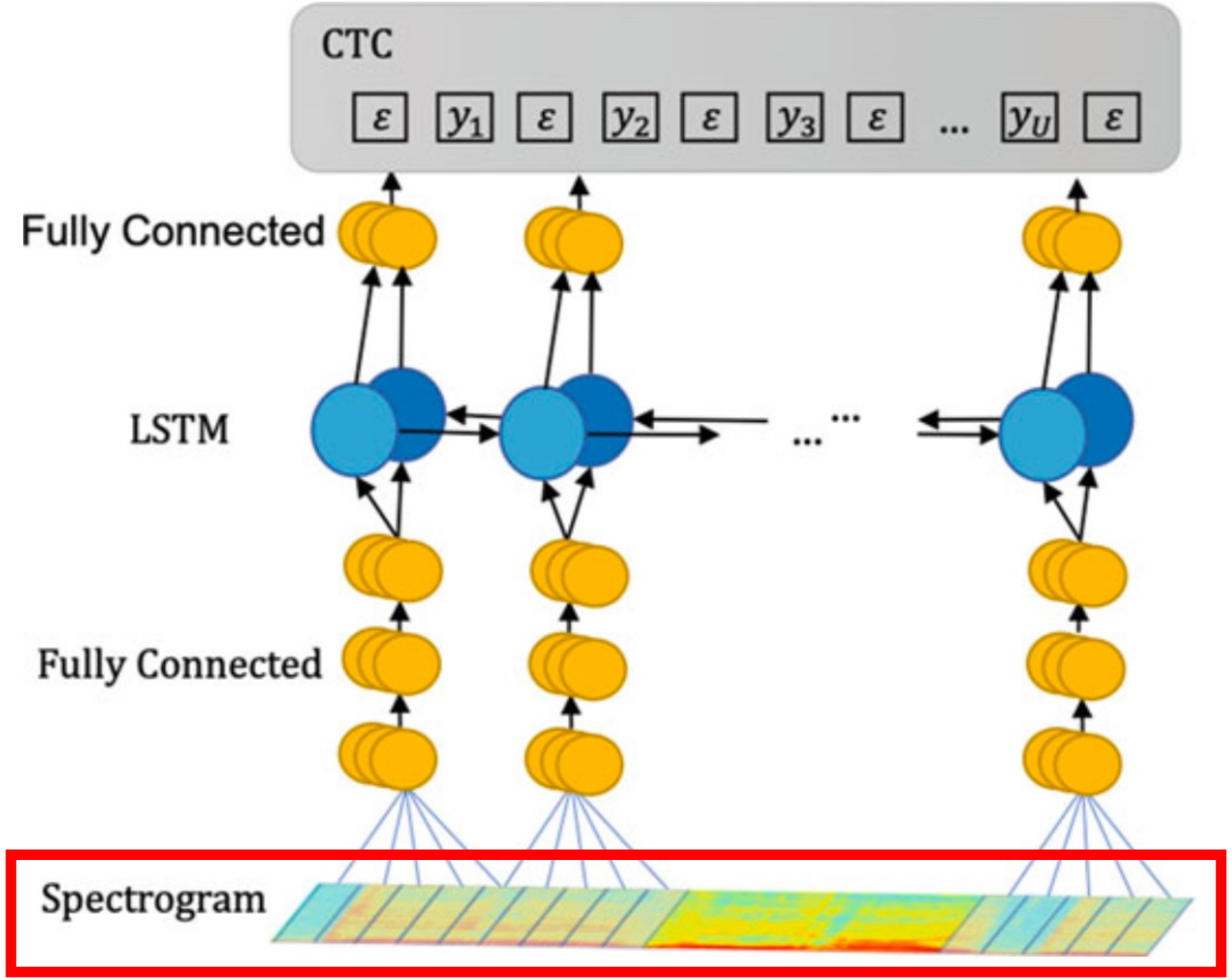
Decoder:

Encoder:



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech

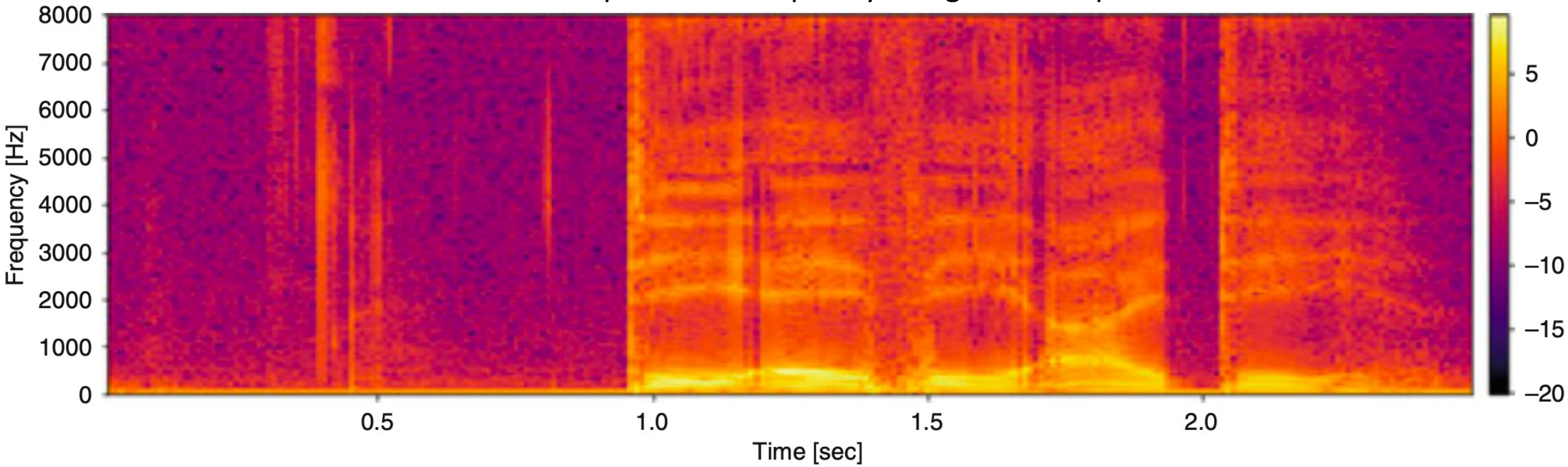


Input:

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Spectrogram: Visual Representation of Audio

Color: amplitude of frequency at a given time point

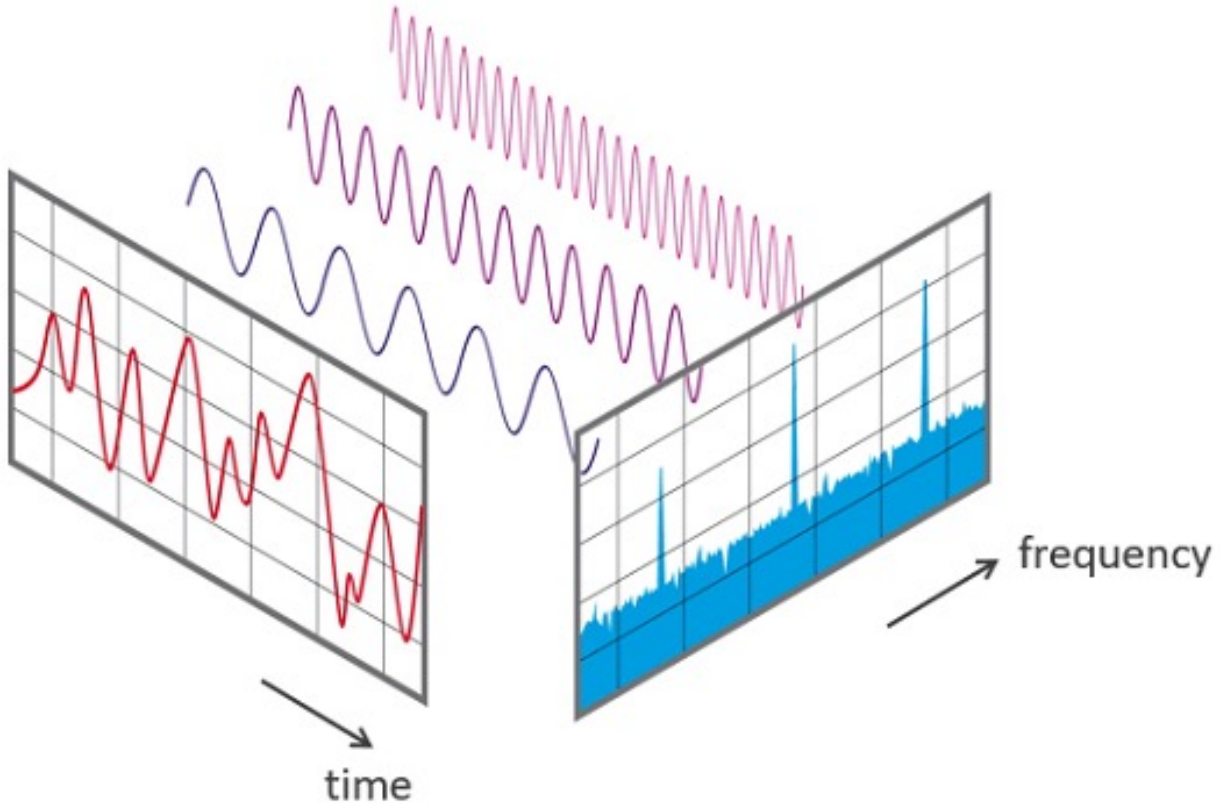


Created by sliding a short window across the audio signal and applying a Fourier transform to each window

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Background: Frequency Analysis of Audio Clip

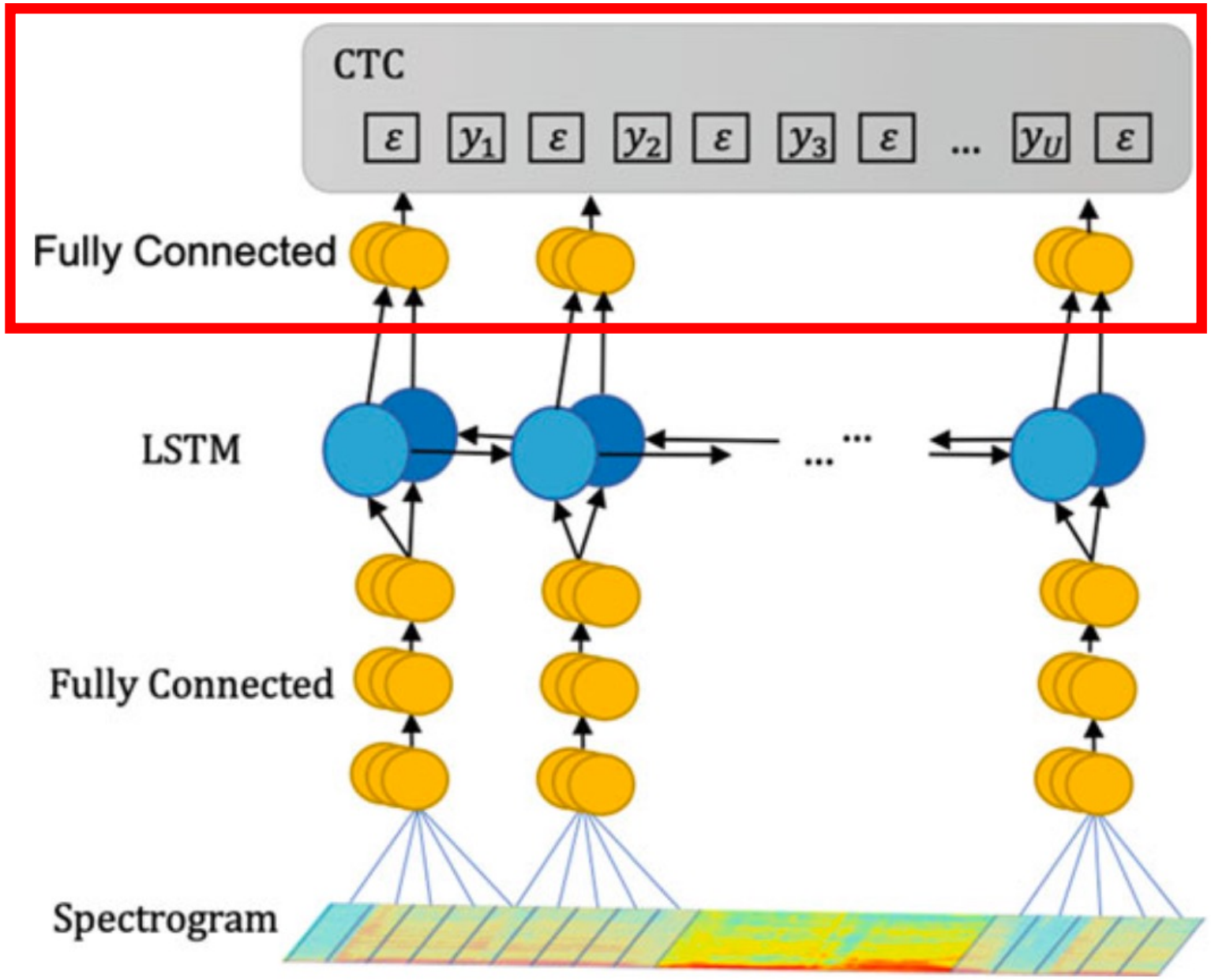
Fourier transform: represents a signal as a sum of sines and cosines (*frequency-domain*):



https://dev.to/trekhleb/playing-with-discrete-fourier-transform-algorithm-in-javascript-53n5

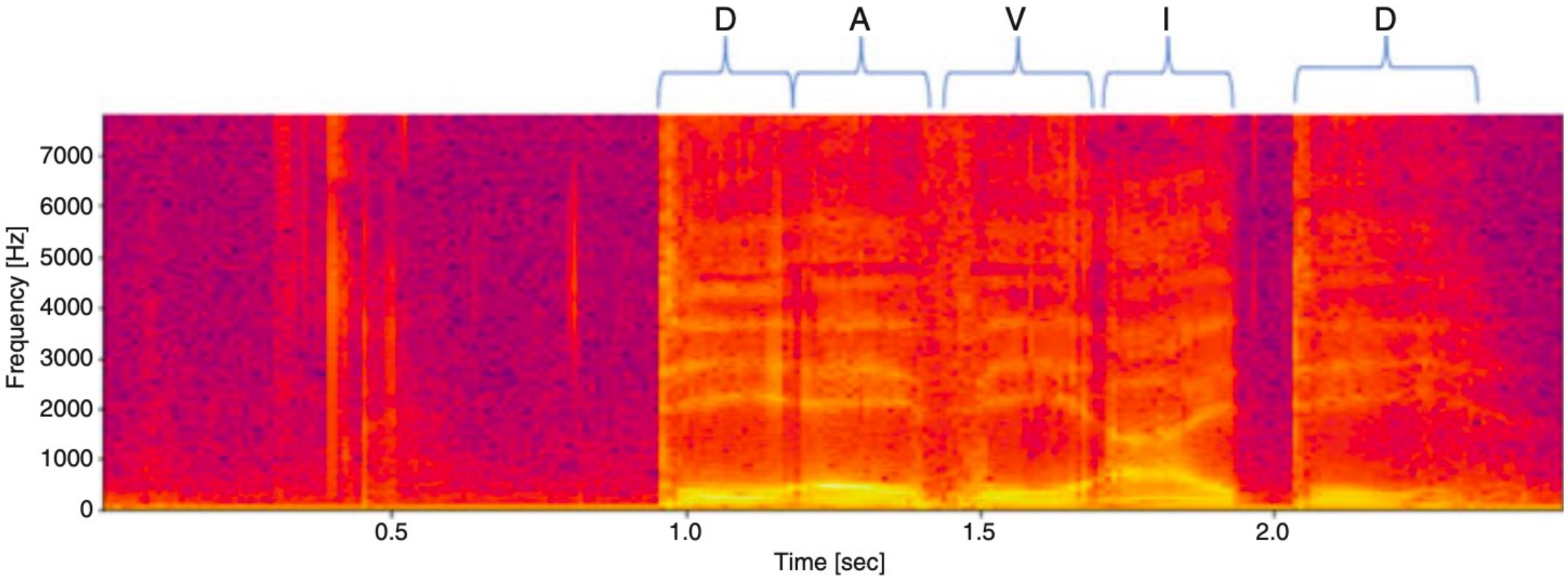
DeepSpeech

Output: character sequence predicted by a softmax layer



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

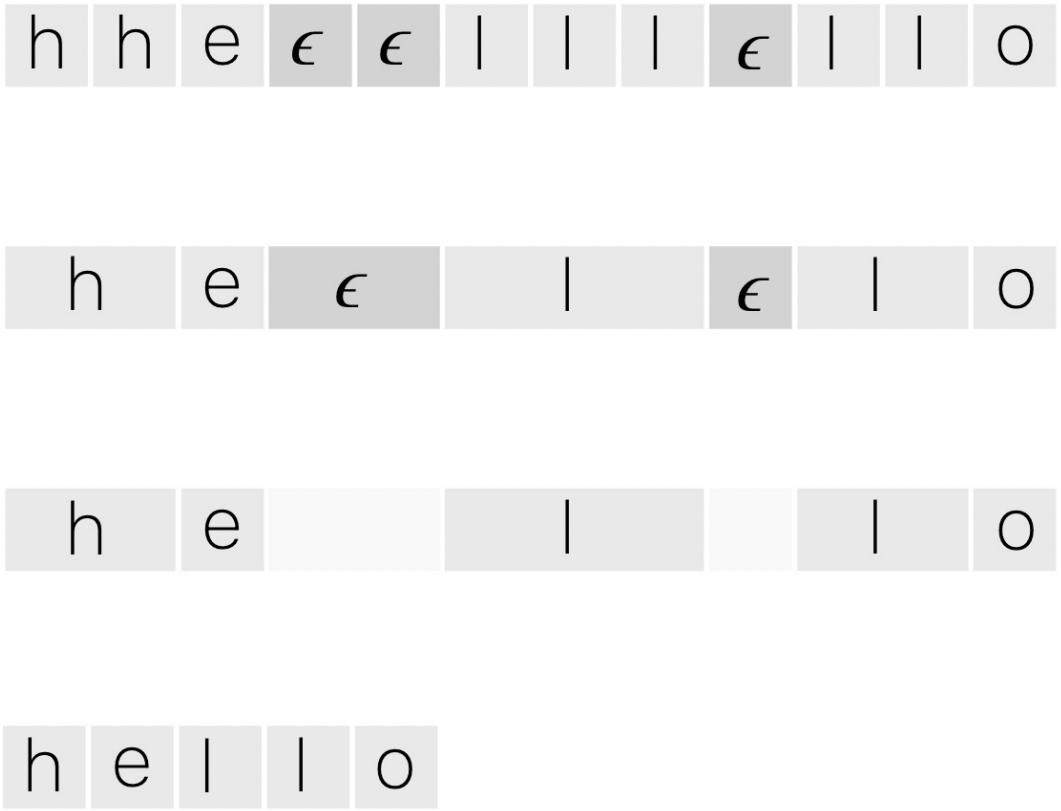
CTC: Input-Output Representation



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

CTC: Input-Output Representation

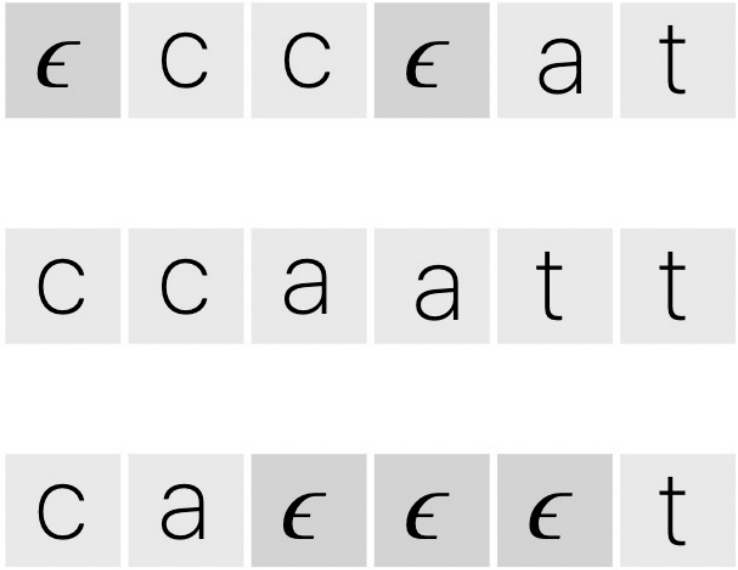
Key idea: blank token supports silent stretches and letter repeats (e.g., “hello” vs “helo”)



https://distill.pub/2017/ctc/

CTC: Input-Output Representation

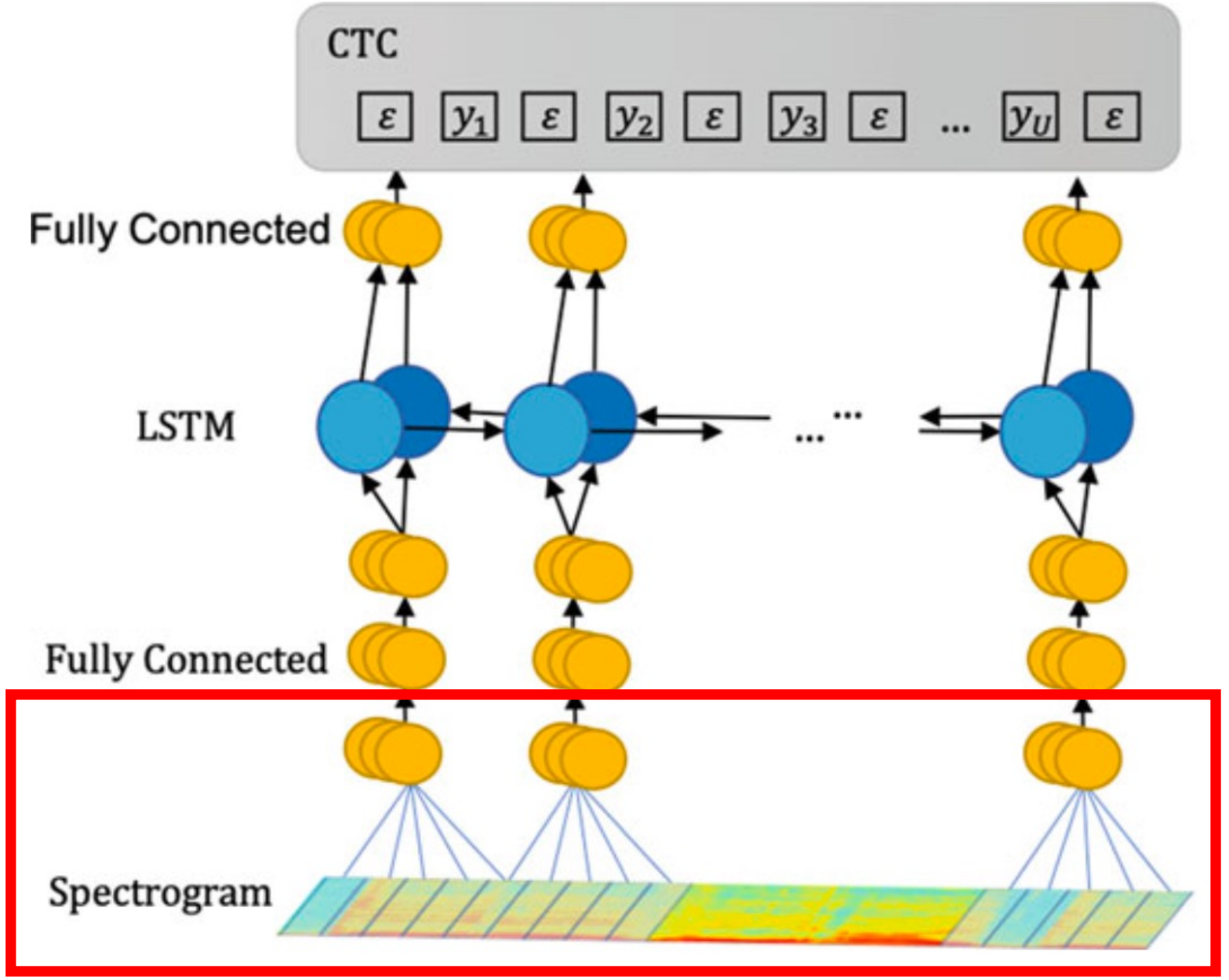
Key idea: blank token supports silent stretches and letter repeats (e.g., “hello” vs “helo”)



Supports recognizing the same word when spoken differently!

https://distill.pub/2017/ctc/

DeepSpeech



First hidden layer looks at context around input:

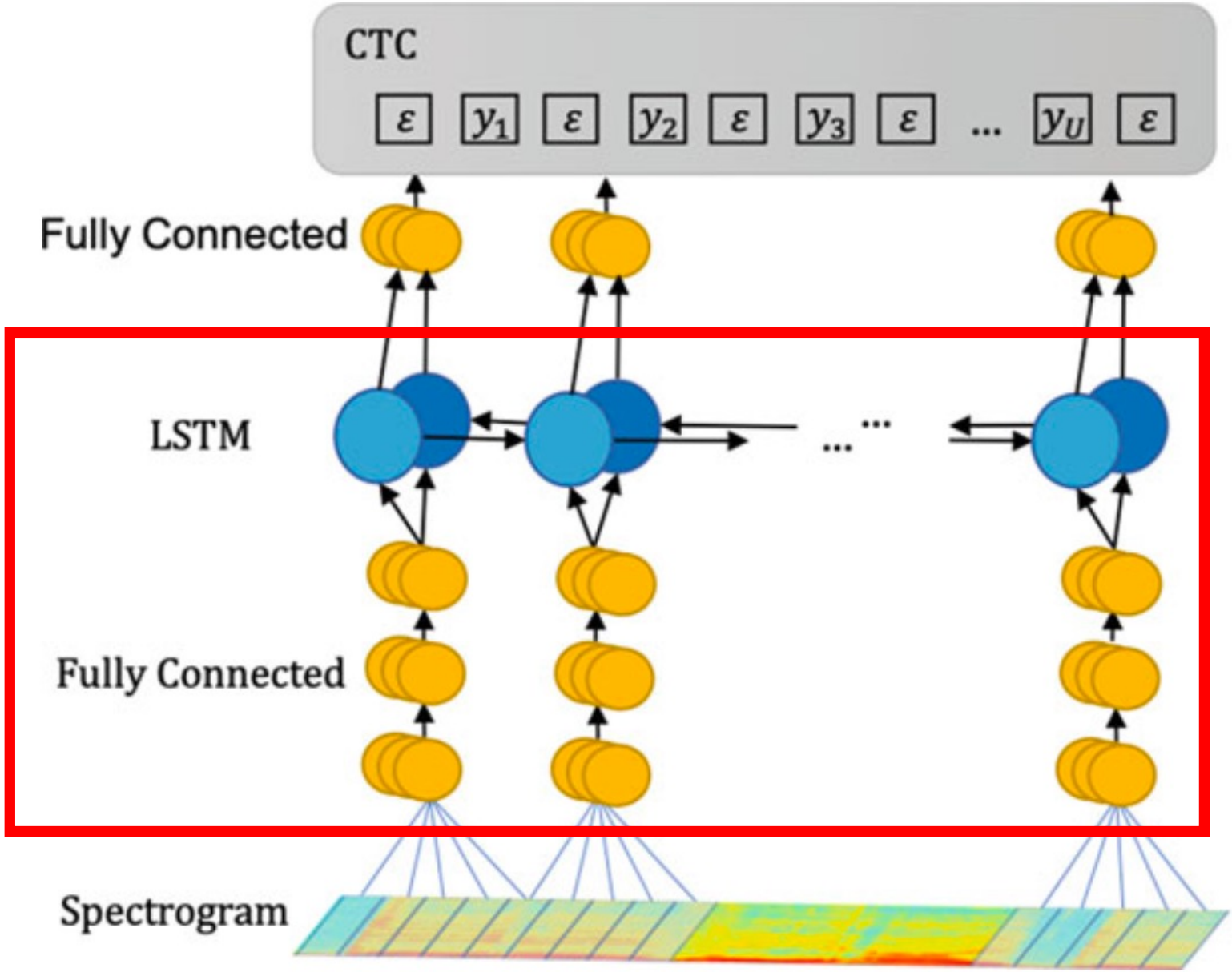
Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech

3 fully-connected layers followed by bidirectional LSTM:

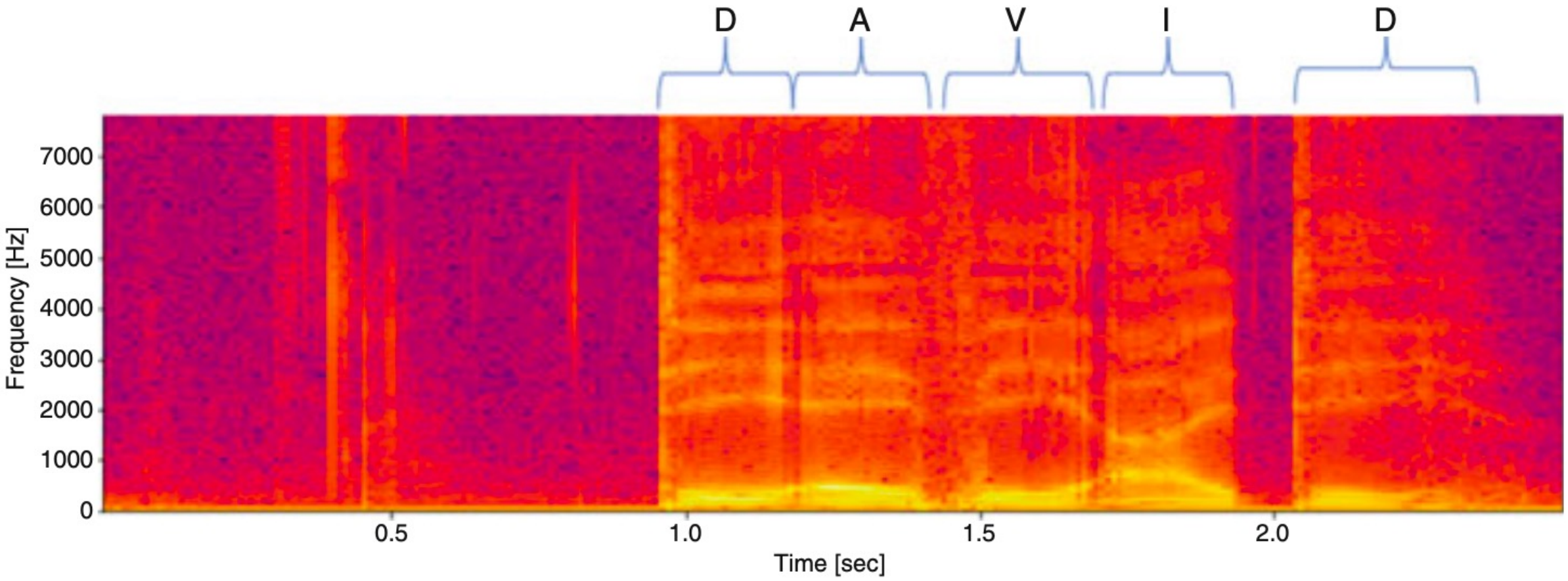
How is a bi-directional LSTM beneficial?

How is a bi-directional LSTM limiting?



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech: Optimization Function (CTC)



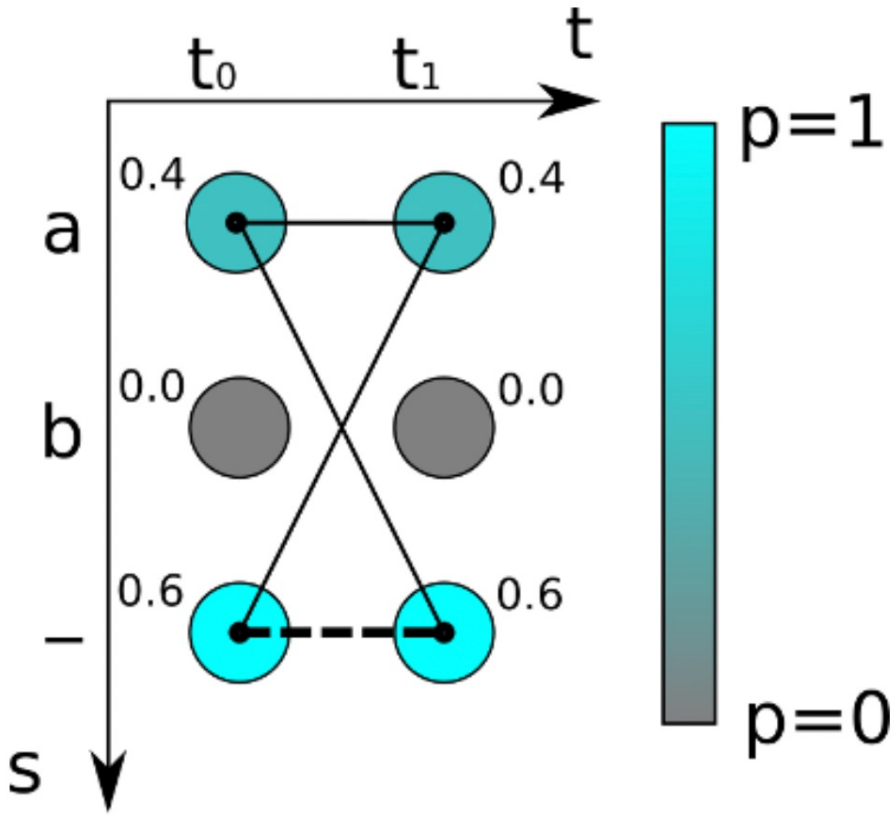
The CTC loss function enables learning output alignment without a per input label

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech: Optimization Function (CTC)

Most plausible from all possible alignments learned; e.g., 2 time steps with 2 potential characters and a blank token (“-”)

- Probability of “a” is sum of all “a” representations
  - Probability of “aa”?
    - 0.4 x 0.4 = 0.16
  - Probability of “a-”?
    - 0.4 x 0.6 = 0.24
  - Probability of “-a”?
    - 0.6 x 0.4 = 0.24
  - Sum: 0.16 + 0.24 + 0.24 = 0.64

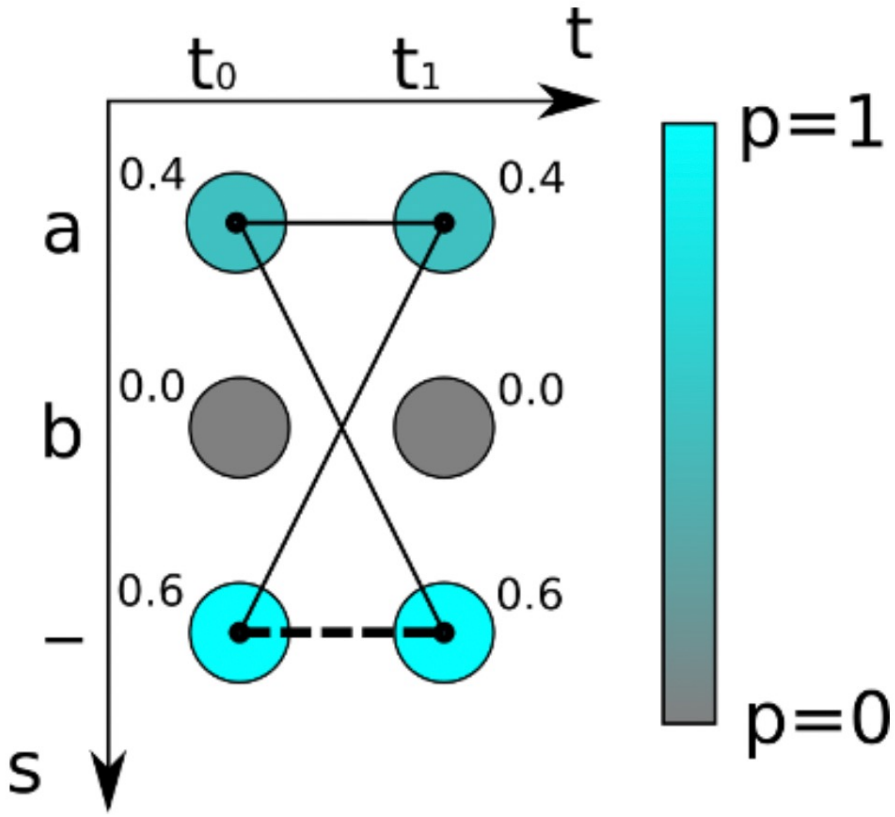


https://towardsdatascience.com/intuitively-understanding-connectionist-temporal-classification-3797e43a86c

DeepSpeech: Optimization Function (CTC)

Most plausible from all possible alignments learned; e.g., 2 time steps with 2 potential characters and a blank token (“-”)

- Probability of “a”: 0.64
- Probability of “” is sum of all “” representations
  - Probability of “--”?
    - 0.6 x 0.6 = 0.36

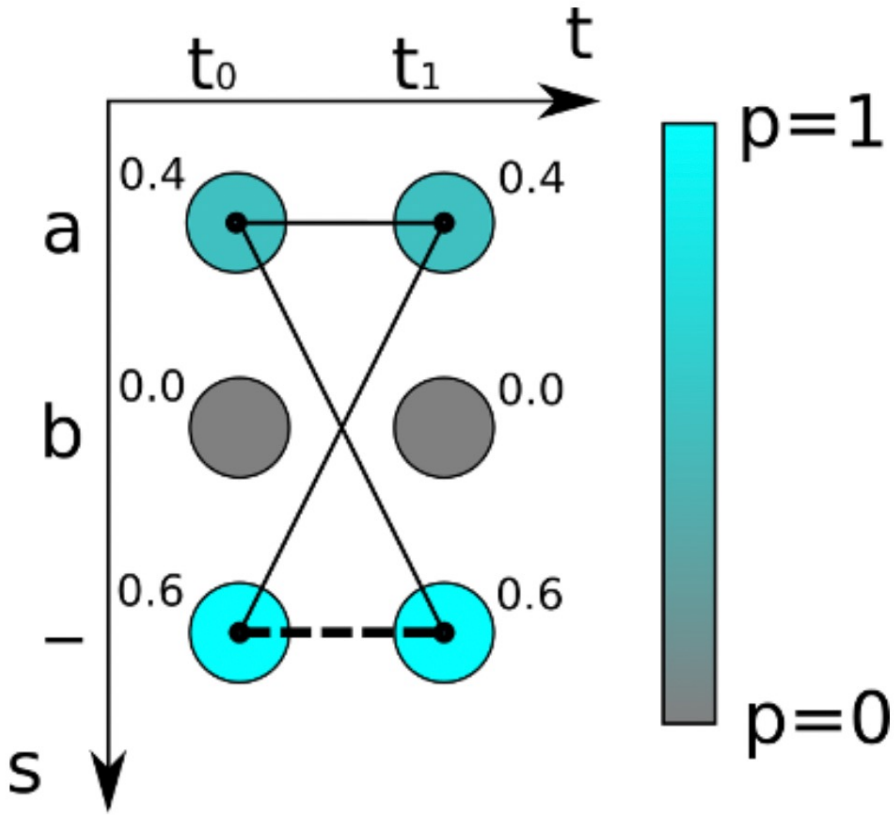


https://towardsdatascience.com/intuitively-understanding-connectionist-temporal-classification-3797e43a86c

DeepSpeech: Optimization Function (CTC)

Most plausible from all possible alignments learned; e.g., 2 time steps with 2 potential characters and a blank token (“-”)

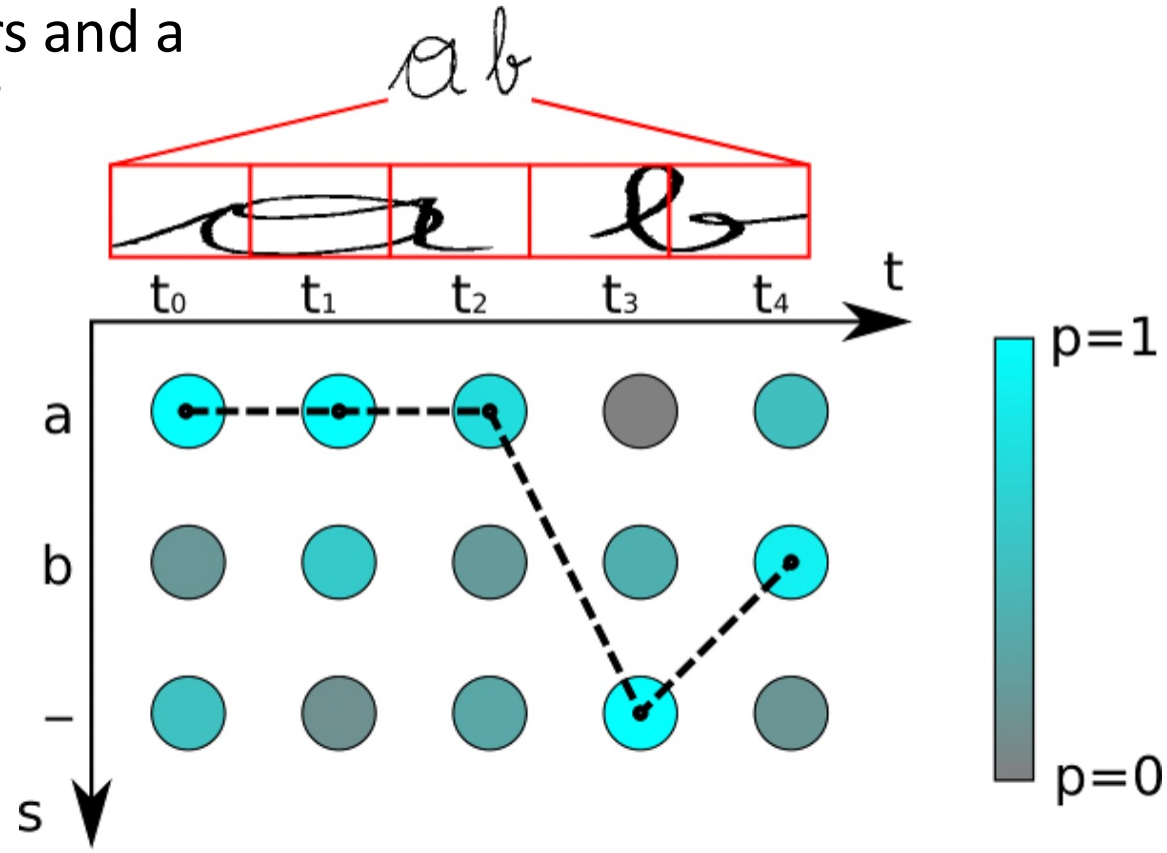
- Probability of “a”: 0.64
- Probability of “”: 0.36
- And so on for all possible alignments…



https://towardsdatascience.com/intuitively-understanding-connectionist-temporal-classification-3797e43a86c

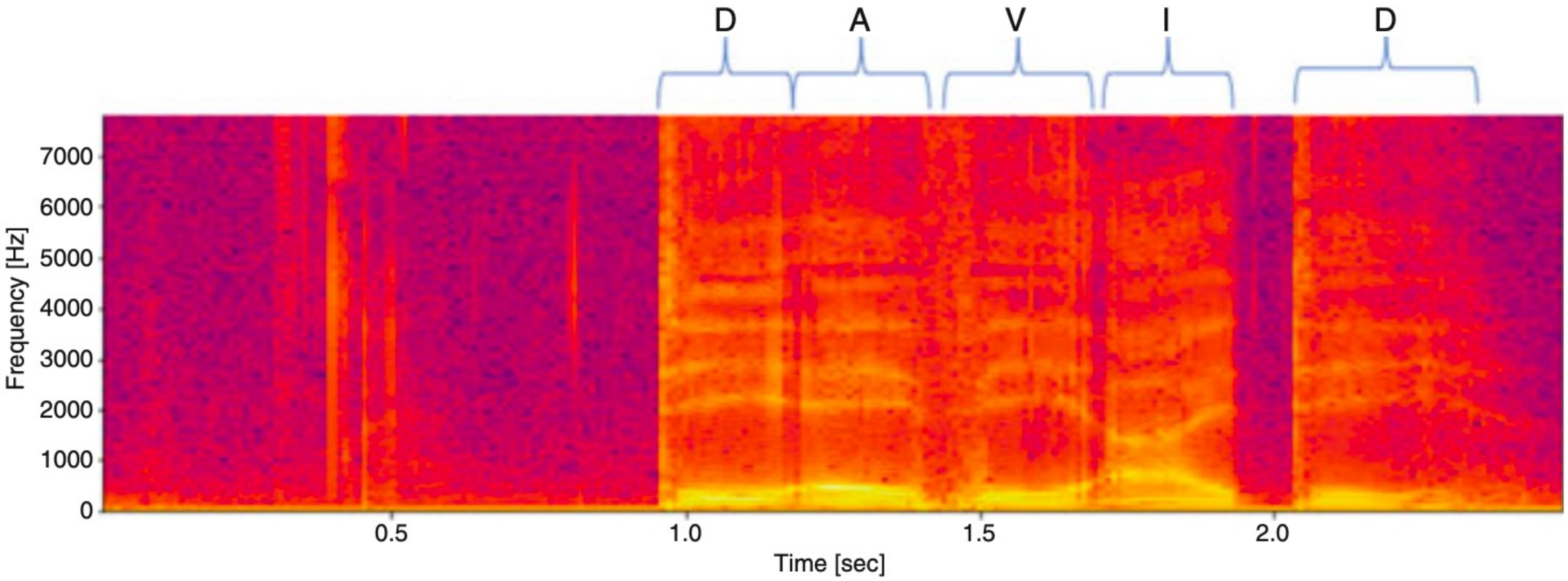
DeepSpeech: Optimization Function (CTC)

Most plausible from all possible alignments learned;
e.g., 2 time steps with 2 potential characters and a blank token ("-") with "best path decoding"



https://towardsdatascience.com/intuitively-understanding-connectionist-temporal-classification-3797e43a86c

DeepSpeech: Optimization Function (CTC)



CTC uses dynamic programming to accelerate computation and is differentiable

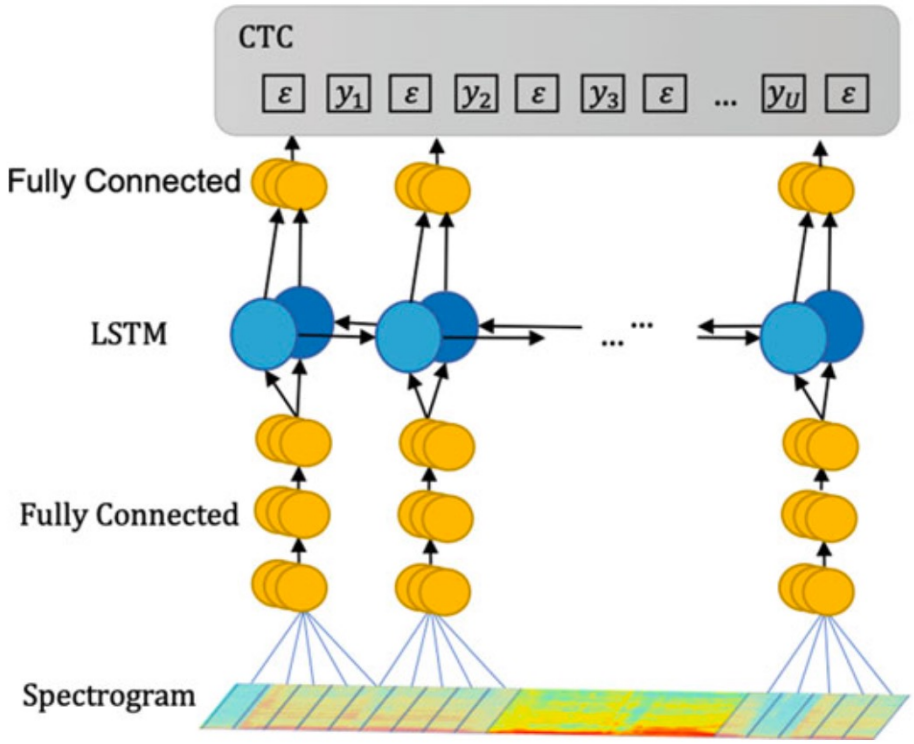
Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech: Training (Key Ideas)

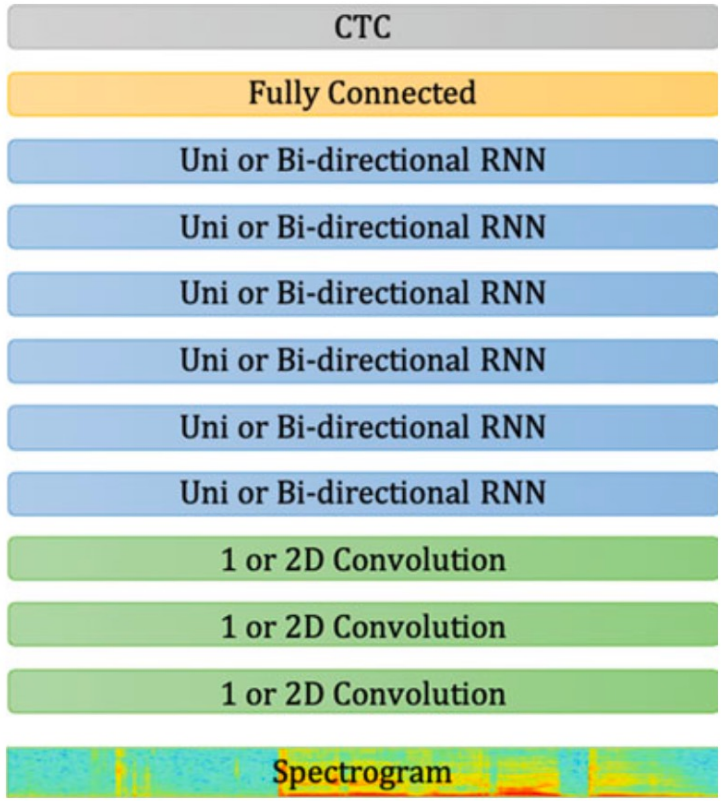
- 5000 hours from 9600 speakers
- Regularization
  - Dropout
  - Data augmentation: audio file translated 5 ms forward and backward
- Results boosted by incorporating a language model

Popular Methods

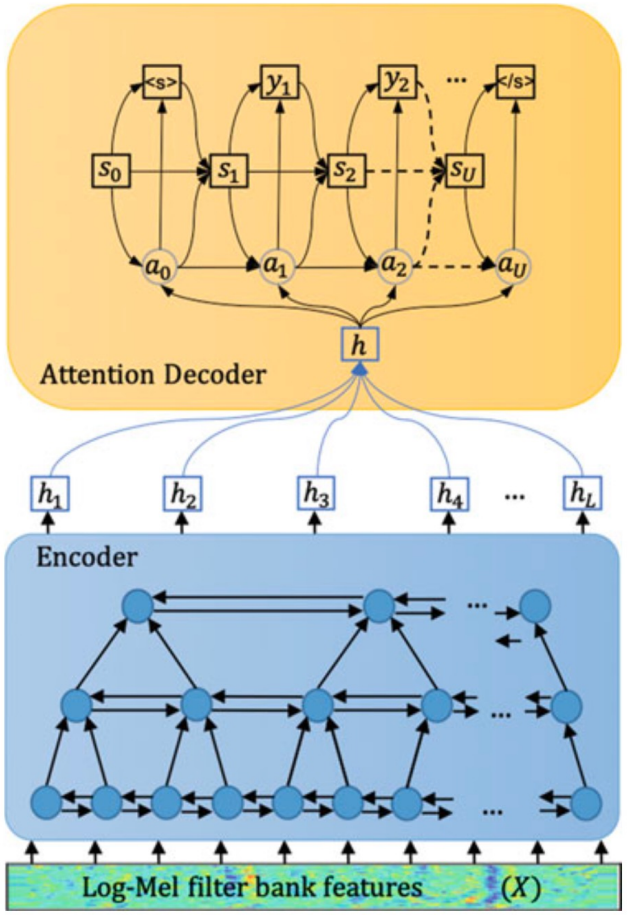
DeepSpeech



DeepSpeech2



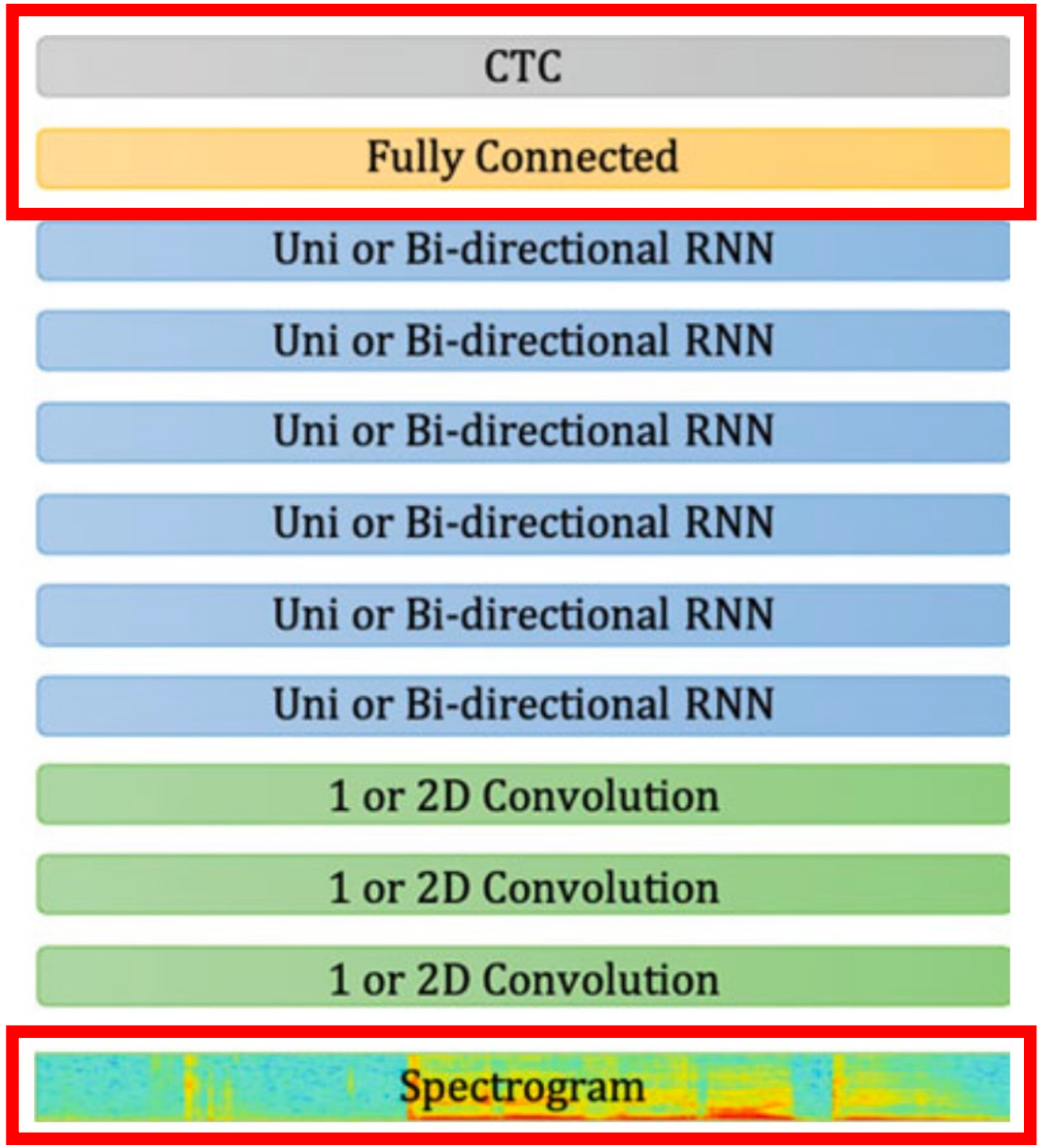
Listen, Attend, and Spell



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech2

Similar output:
(two architectures for
English and Mandarin)



Extension of DeepSpeech that achieves a 7x speed-up and 43.4% relative WER improvement with a deeper architecture

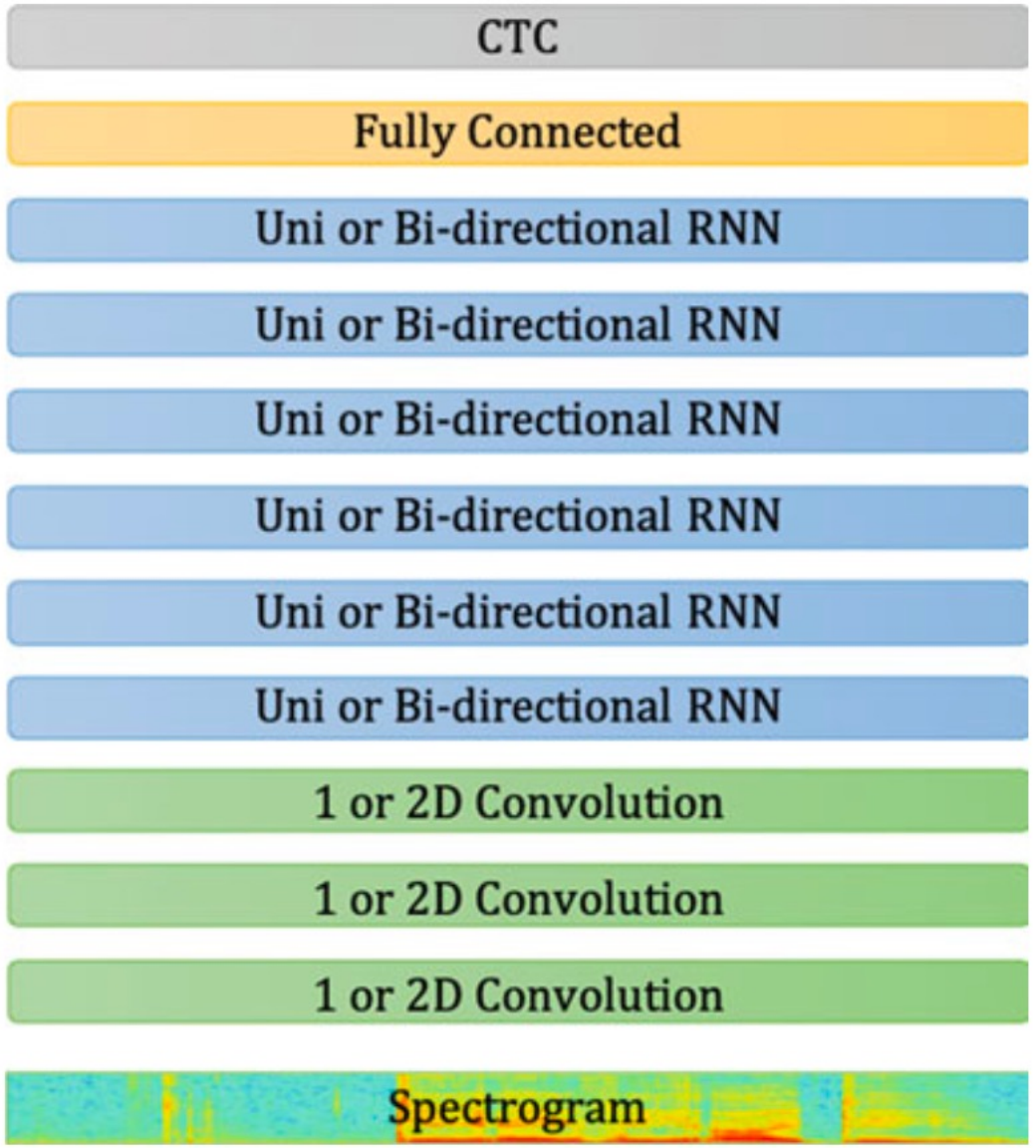
Same input:

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

DeepSpeech2

Training protocol difference from DeepSpeech:

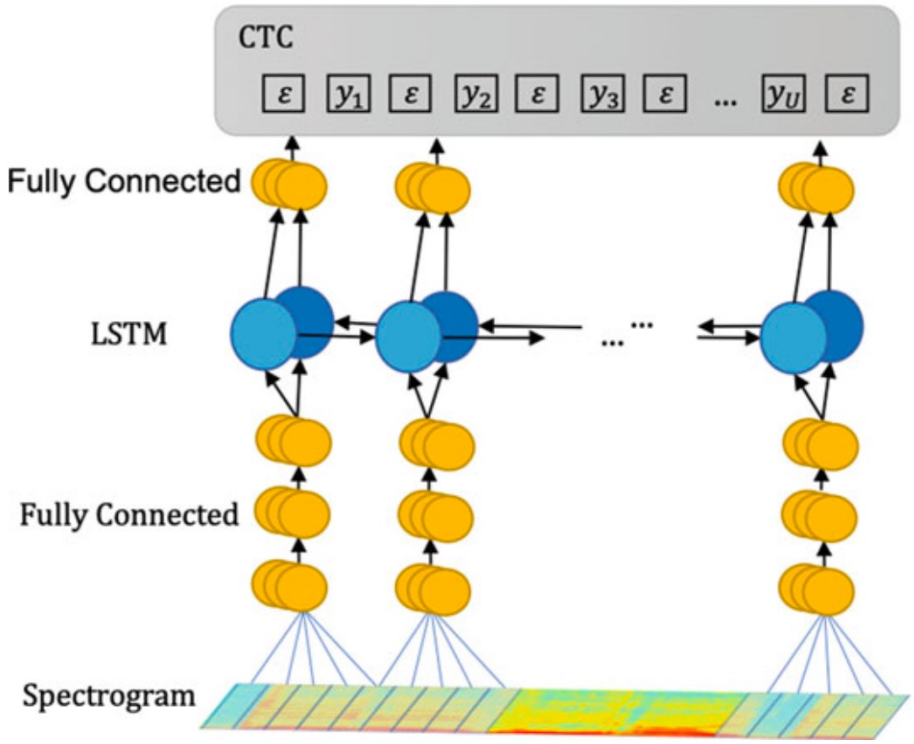
- More training data (11,940 hours for English and 9,400 hours for Mandarin)
- Curriculum learning: trains based on length of utterances for first epoch with shorter ones first (improves WER by over 1 point)



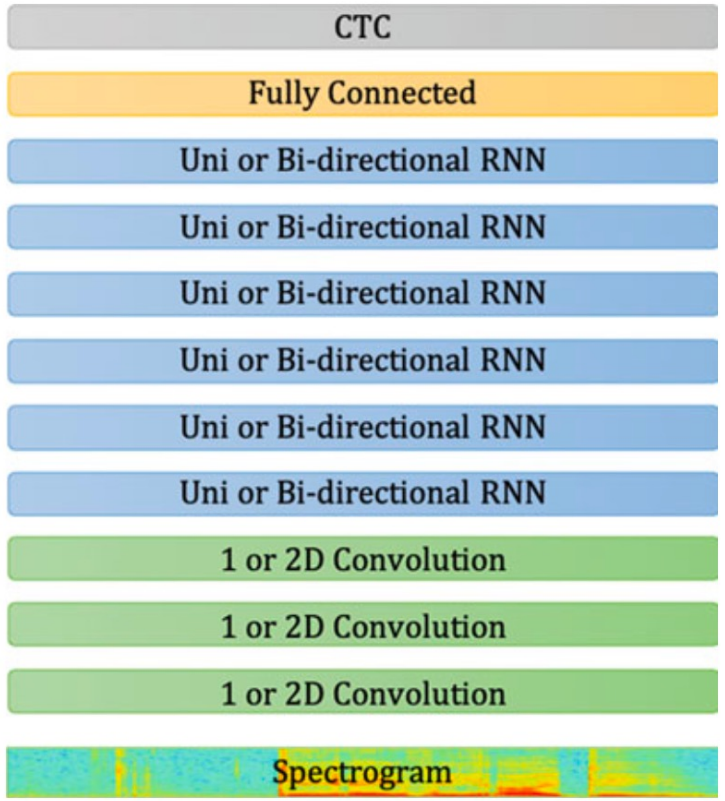
Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Popular Methods

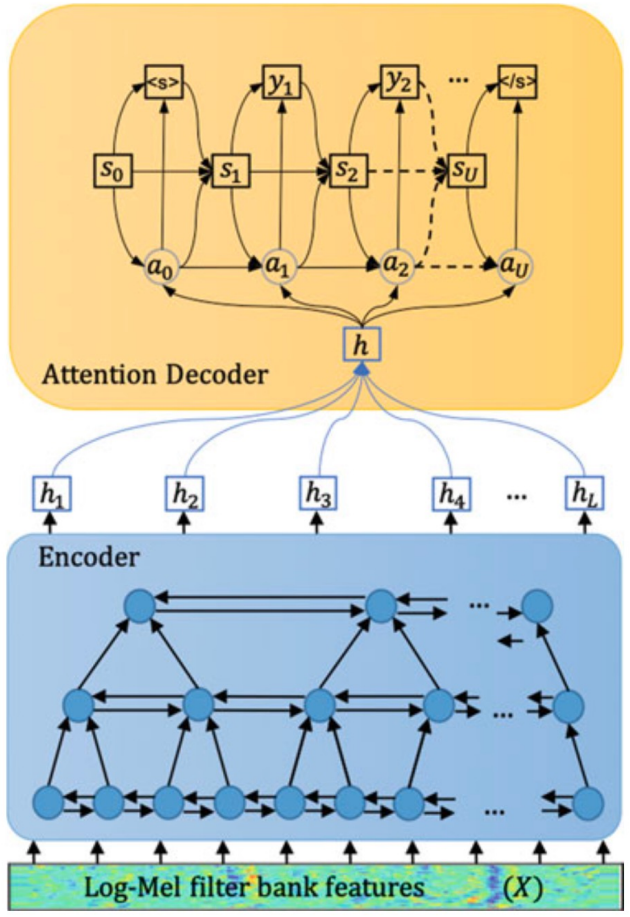
DeepSpeech



DeepSpeech2



Listen, Attend, and Spell

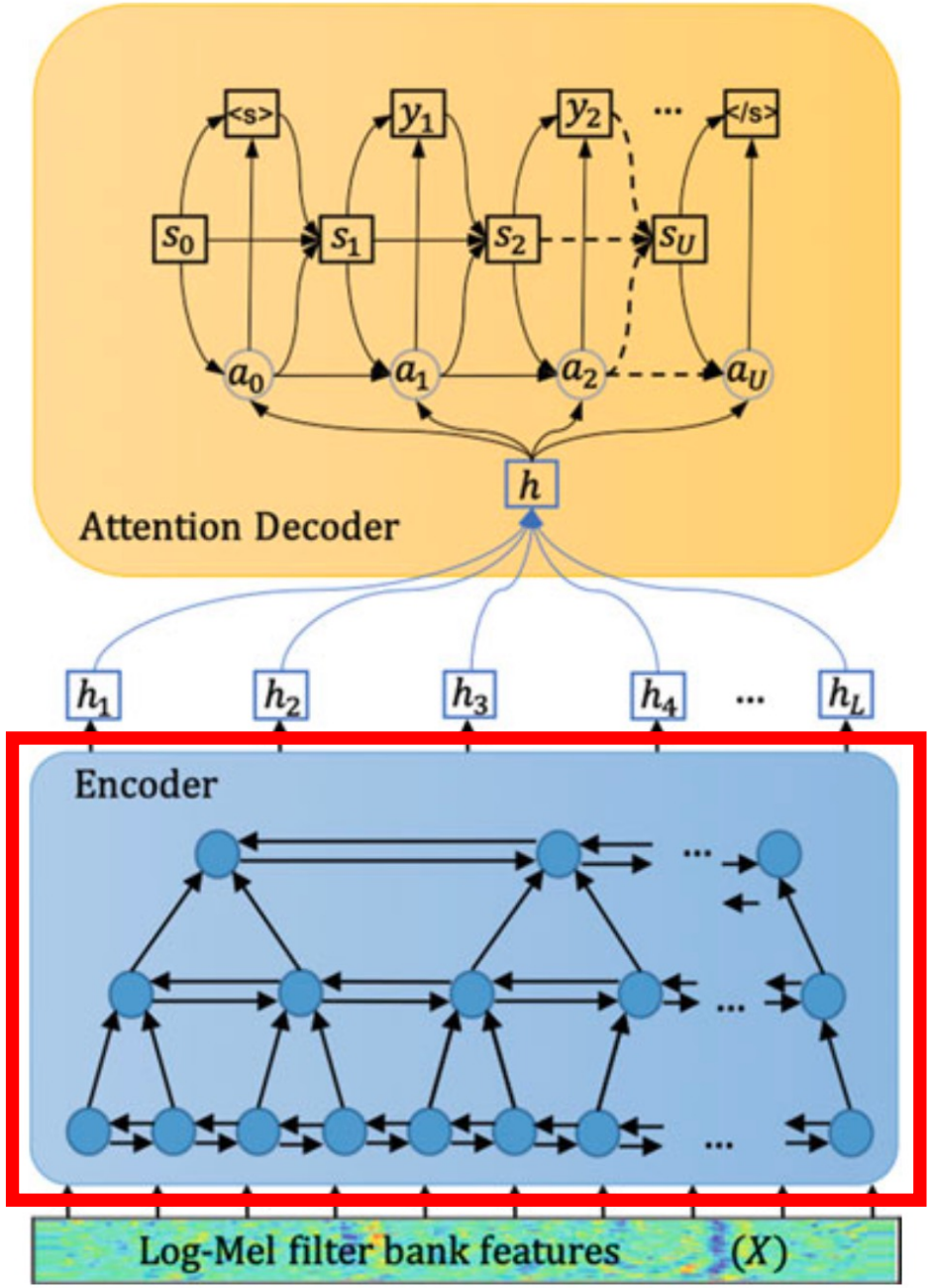


Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Listen, Attend, and Spell

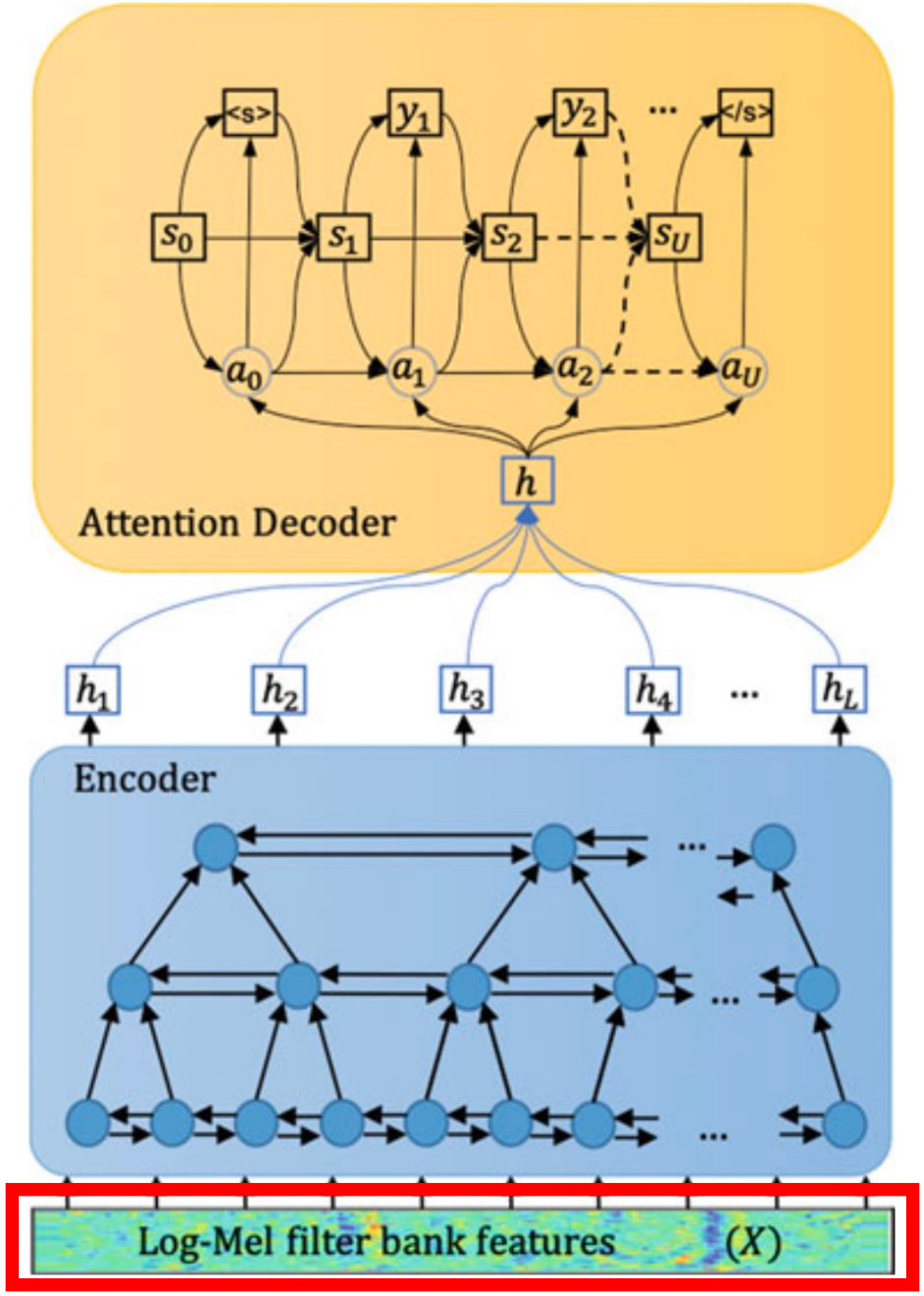
Mimics original paper on sequence to sequence learning with attention where the decoder learns what to attend to in the encoded representation

Pyramid structure reduces number of input time steps



Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Listen, Attend, and Spell

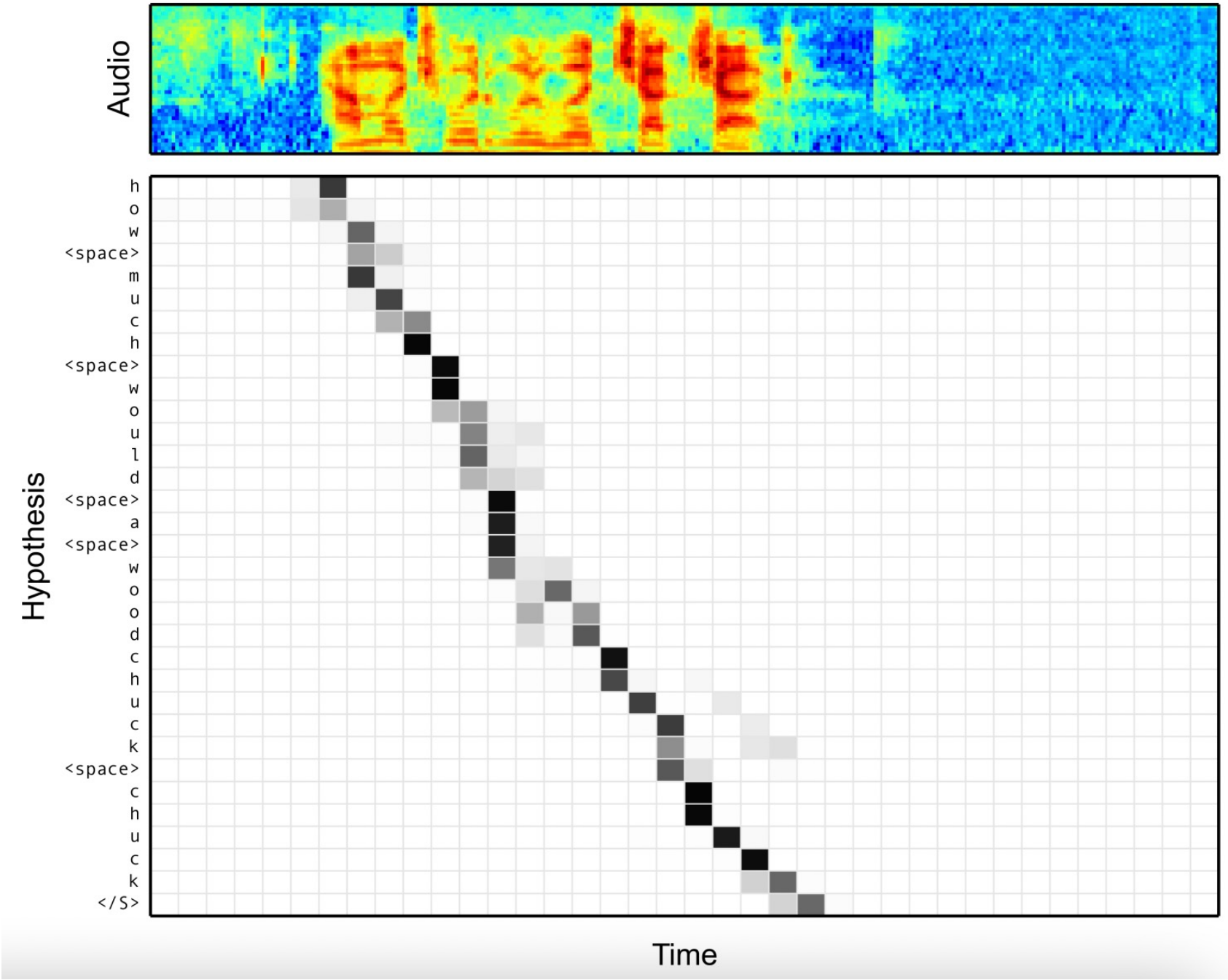


Input: more sophisticated hand-crafted audio representation then spectrogram

Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Result

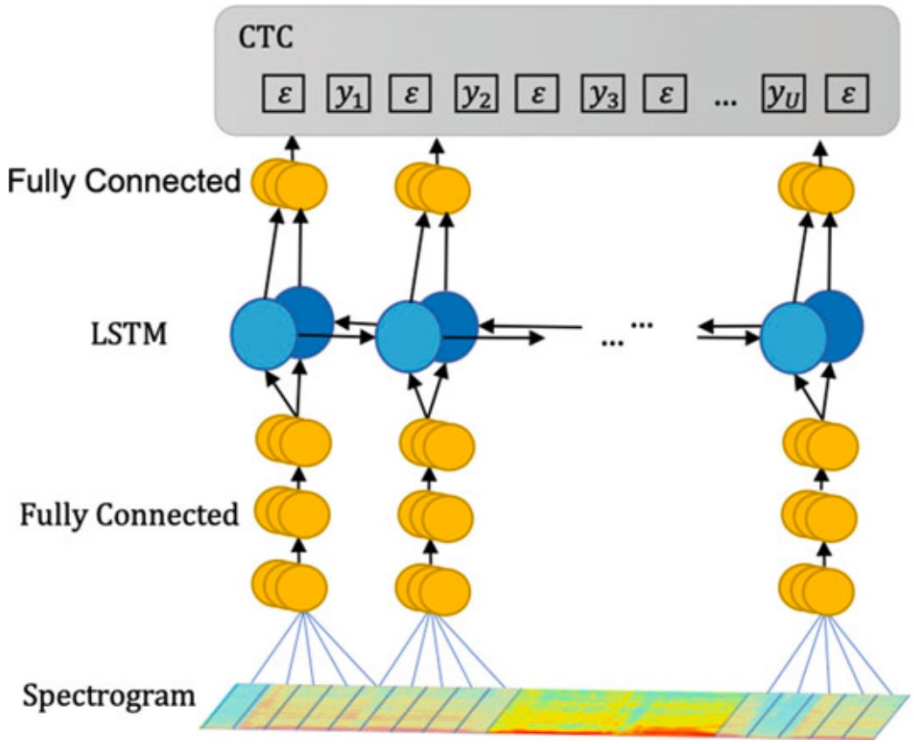
Attention enables visualizing alignment between audio signal and characters



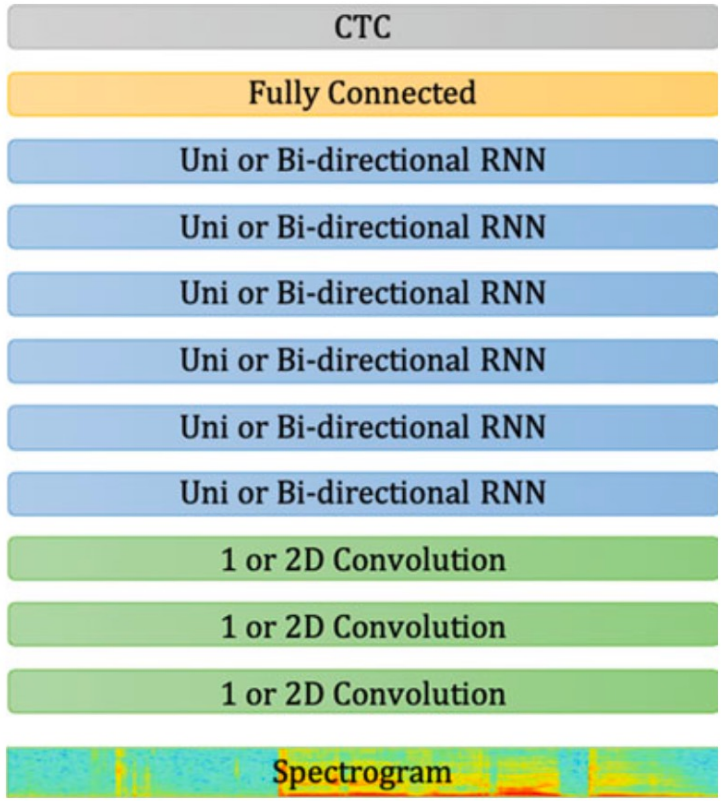
Chan et al. Listen, Attend and Spell. ICASSP 2016.

Popular Methods

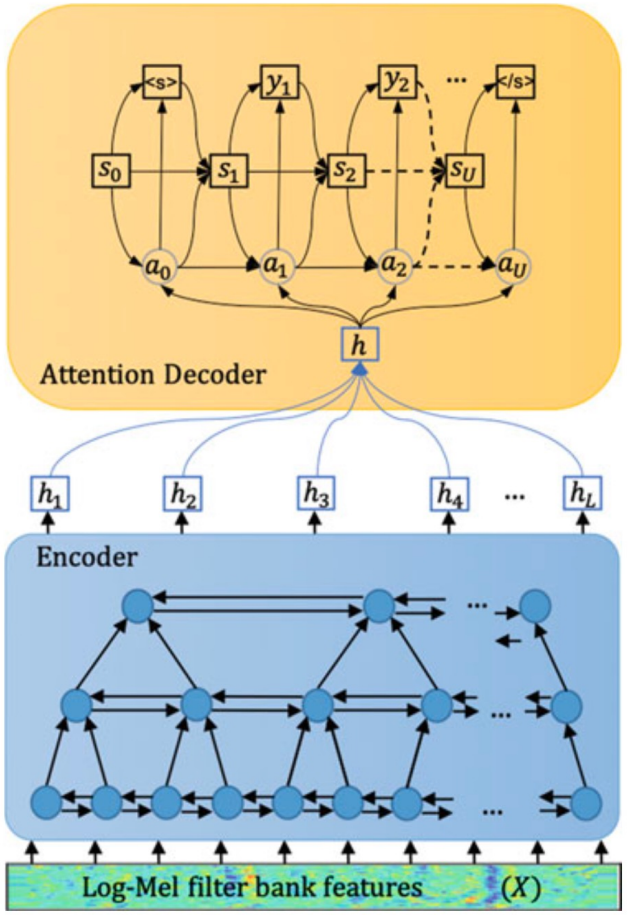
DeepSpeech



DeepSpeech2



Listen, Attend, and Spell



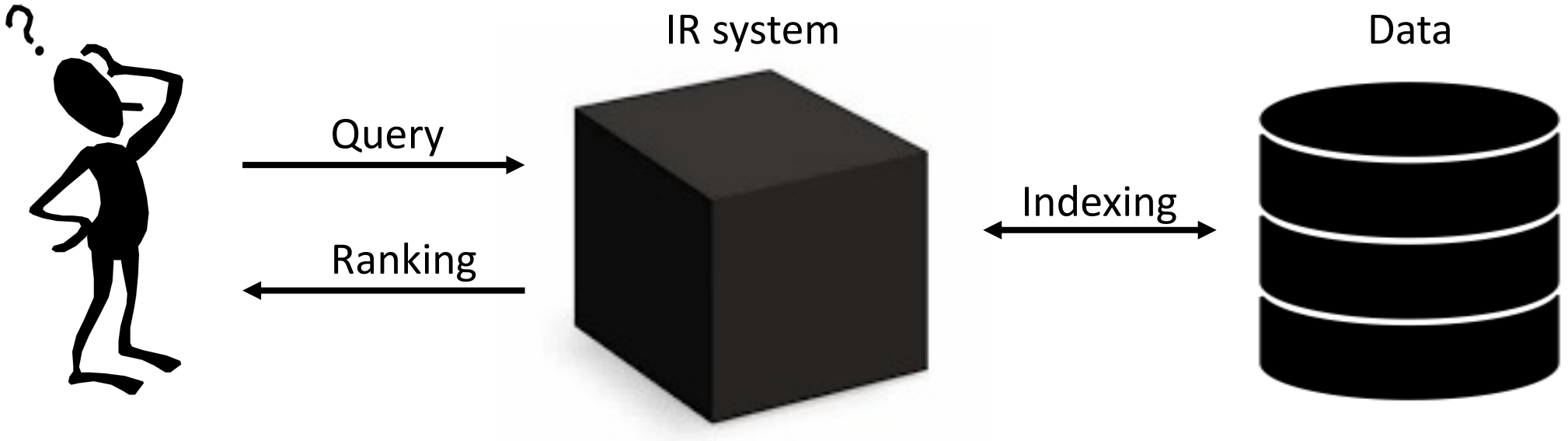
Kamath, Liu, and Whitaker. Deep Learning for NLP and Speech Recognition. 2019.

Today's Topics

- Speech Processing – Problem and Applications
- Speech Recognition – Evaluation and Models
- Informal Retrieval – Problem and Applications
- Informal Retrieval – Models
- Video making tutorial

Problem Definition

- **Neural search**: coined in 2016 SIGIR workshop for deep neural networks applied to information retrieval (IR)



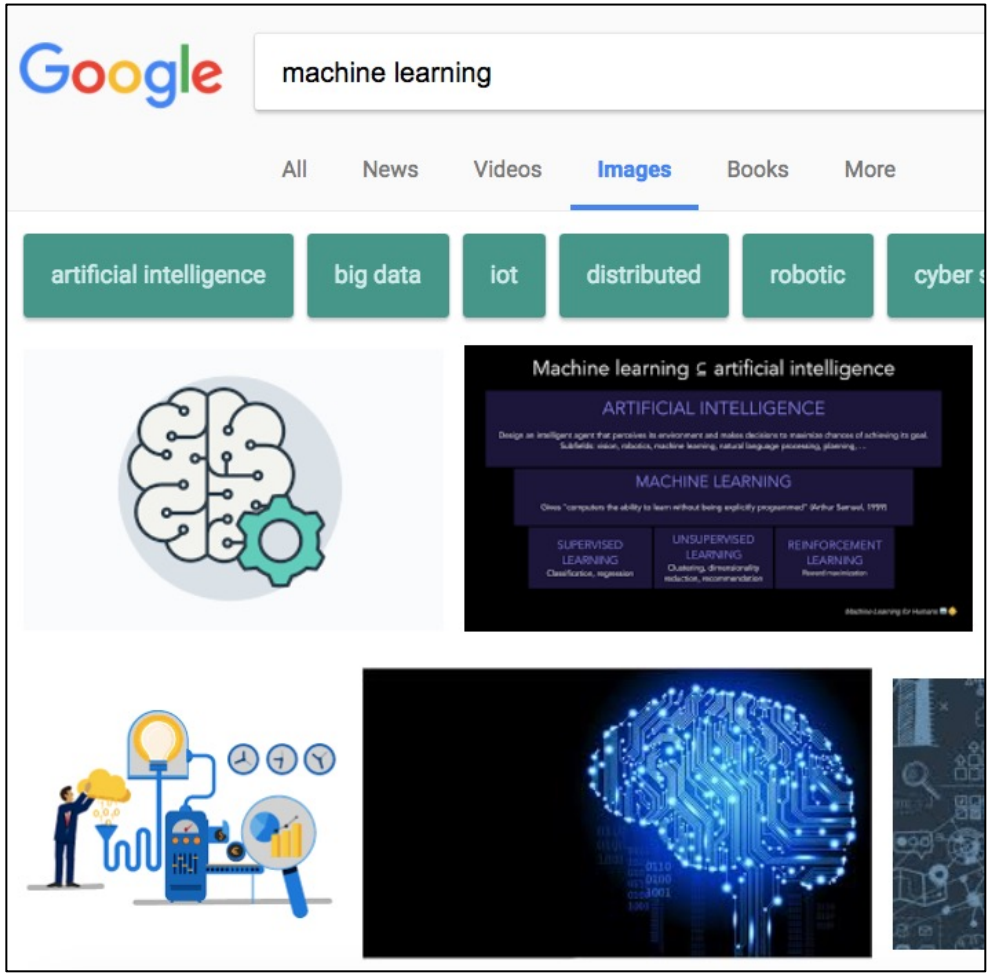
Historical Context: Libraries

Search supported by humans (i.e., librarians), who typically were trained in a “School of Information” or “Information Science” program

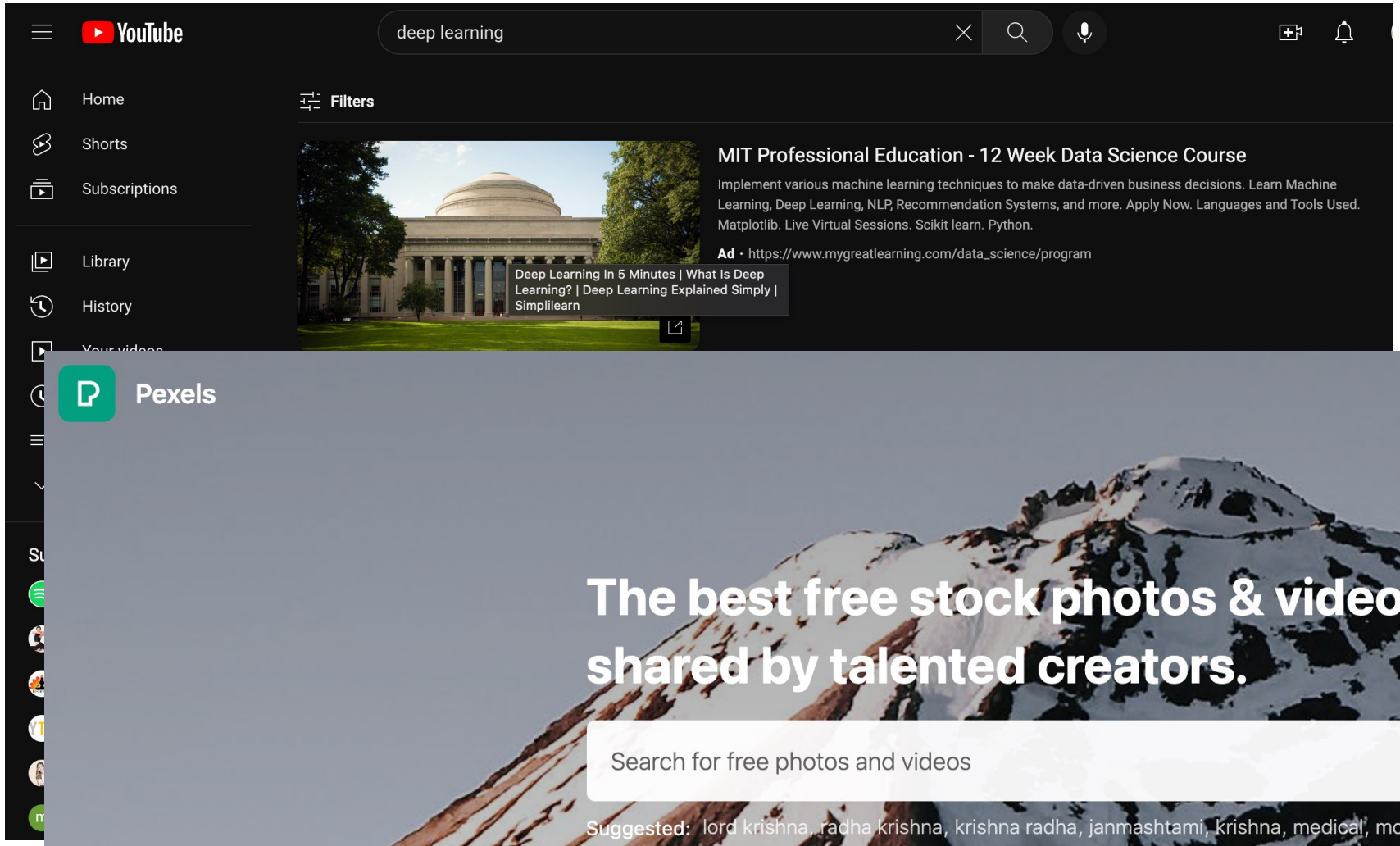
Why Is Search Challenging?

- Should **rank** results by relevance to accelerate locating target information
- Must be **fast** to retain users
- Should **tailor** results to each user

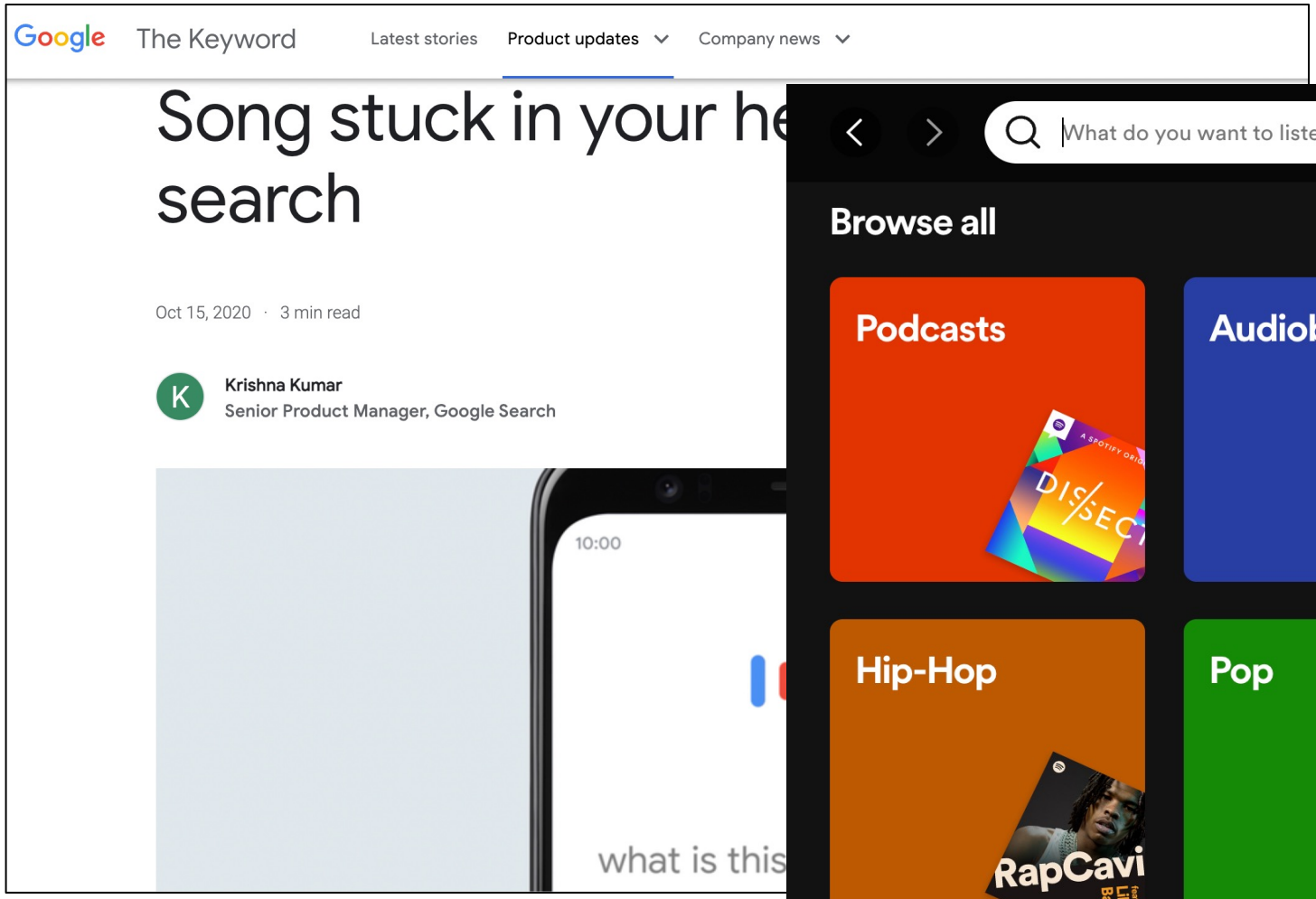
Internet Search

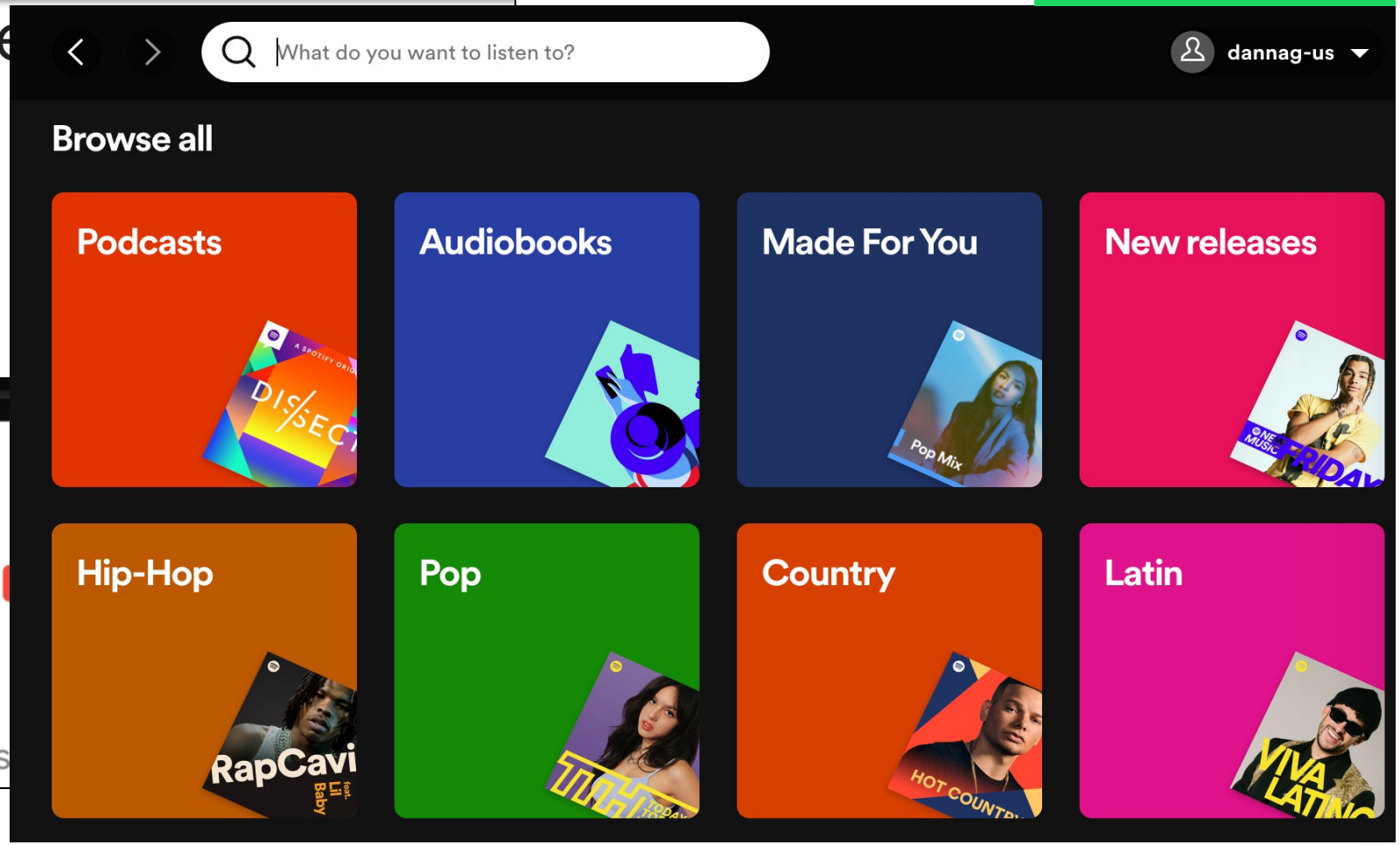


Visual Search

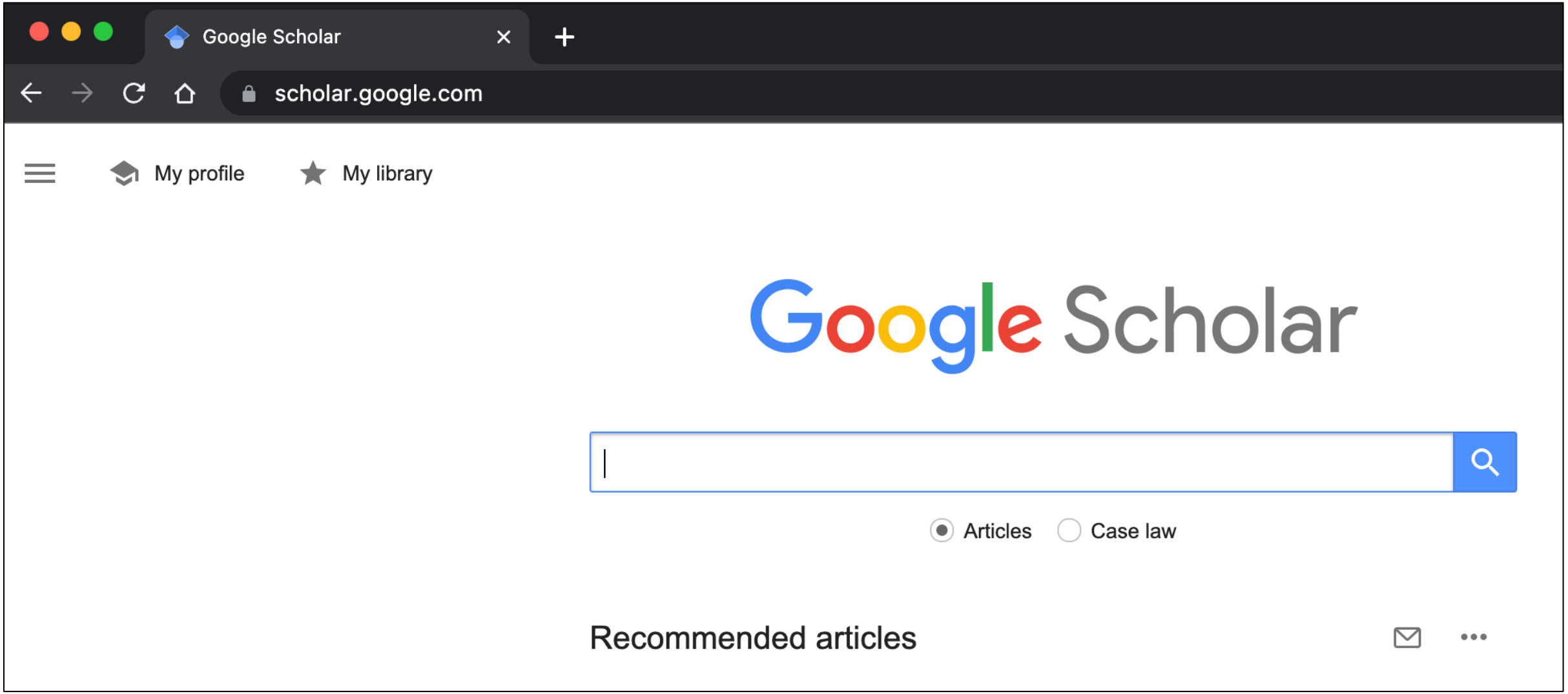


Song Search

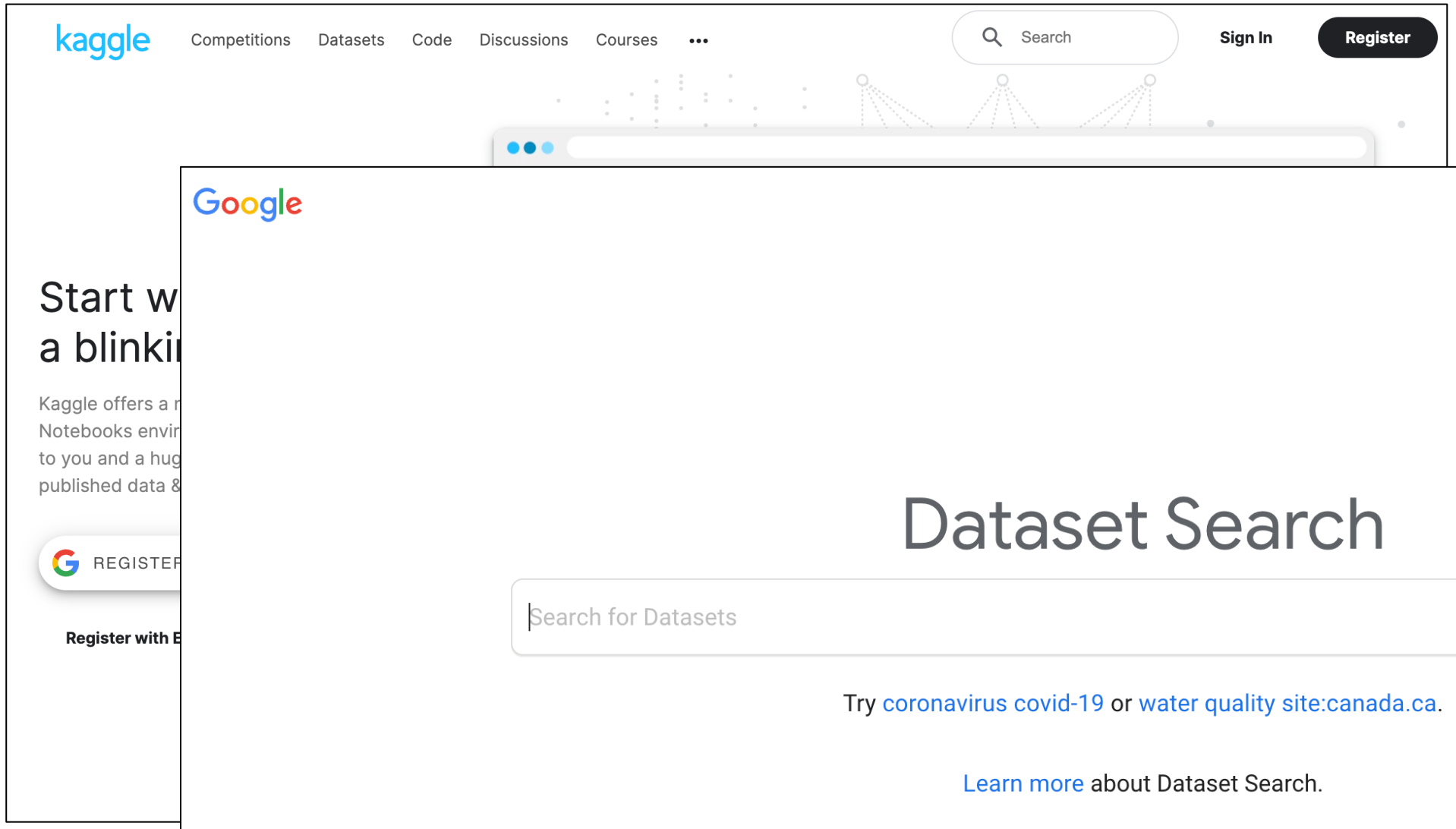

Google
The Keyword
Latest stories
Product updates
Company news
Song stuck in your he
search
Oct 15, 2020 · 3 min read
Krishna Kumar
Senior Product Manager, Google Search
10:00
what is this


Spotify
What do you want to listen to?
dannag-us
Browse all
Podcasts
Audiobooks
Made For You
New releases
Pop Mix
Hip-Hop
Pop
Country
Latin
RapCavi
feat. Lil Baby
HOT COUNTRY
VIVA LATINO

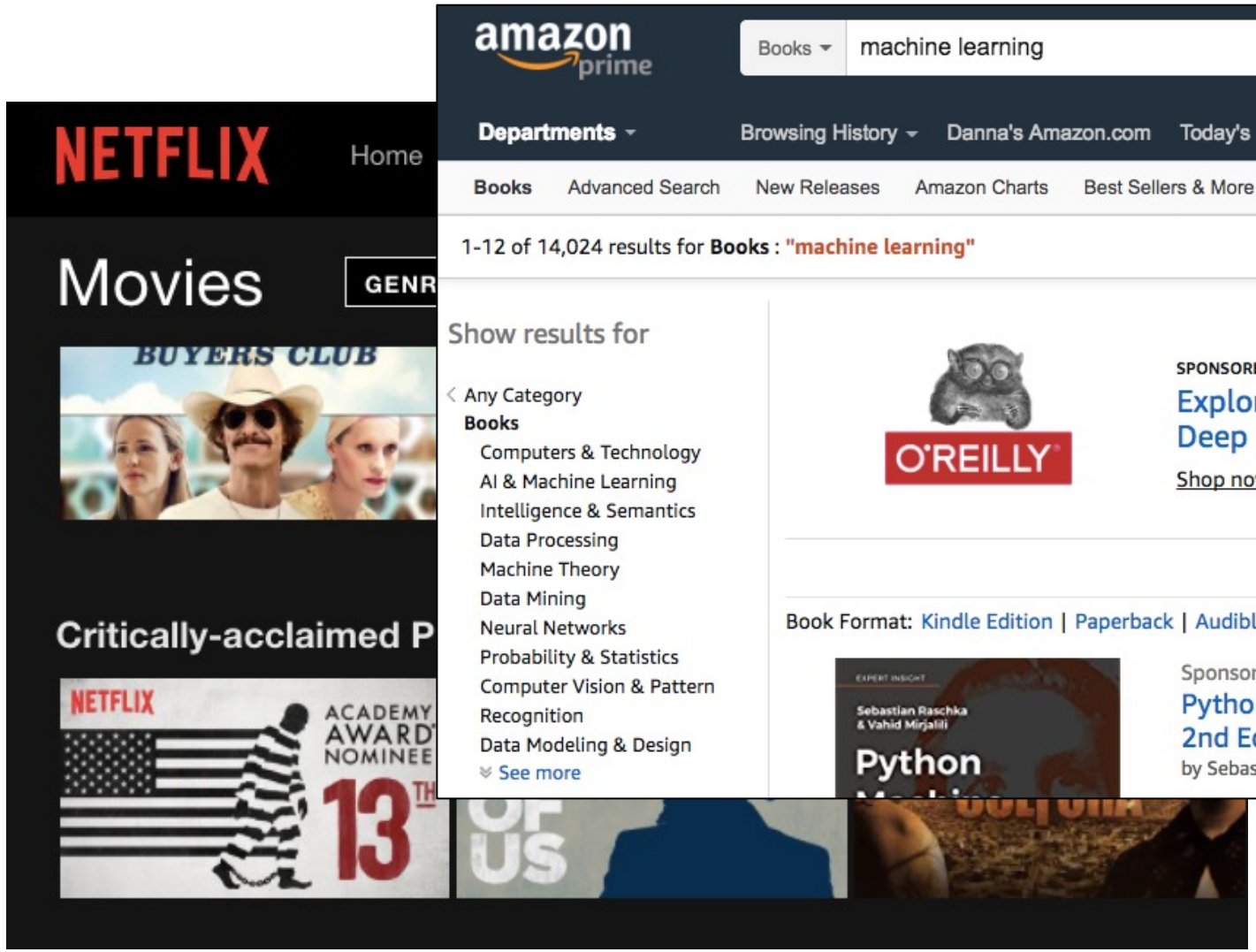
Academic and Legal Search



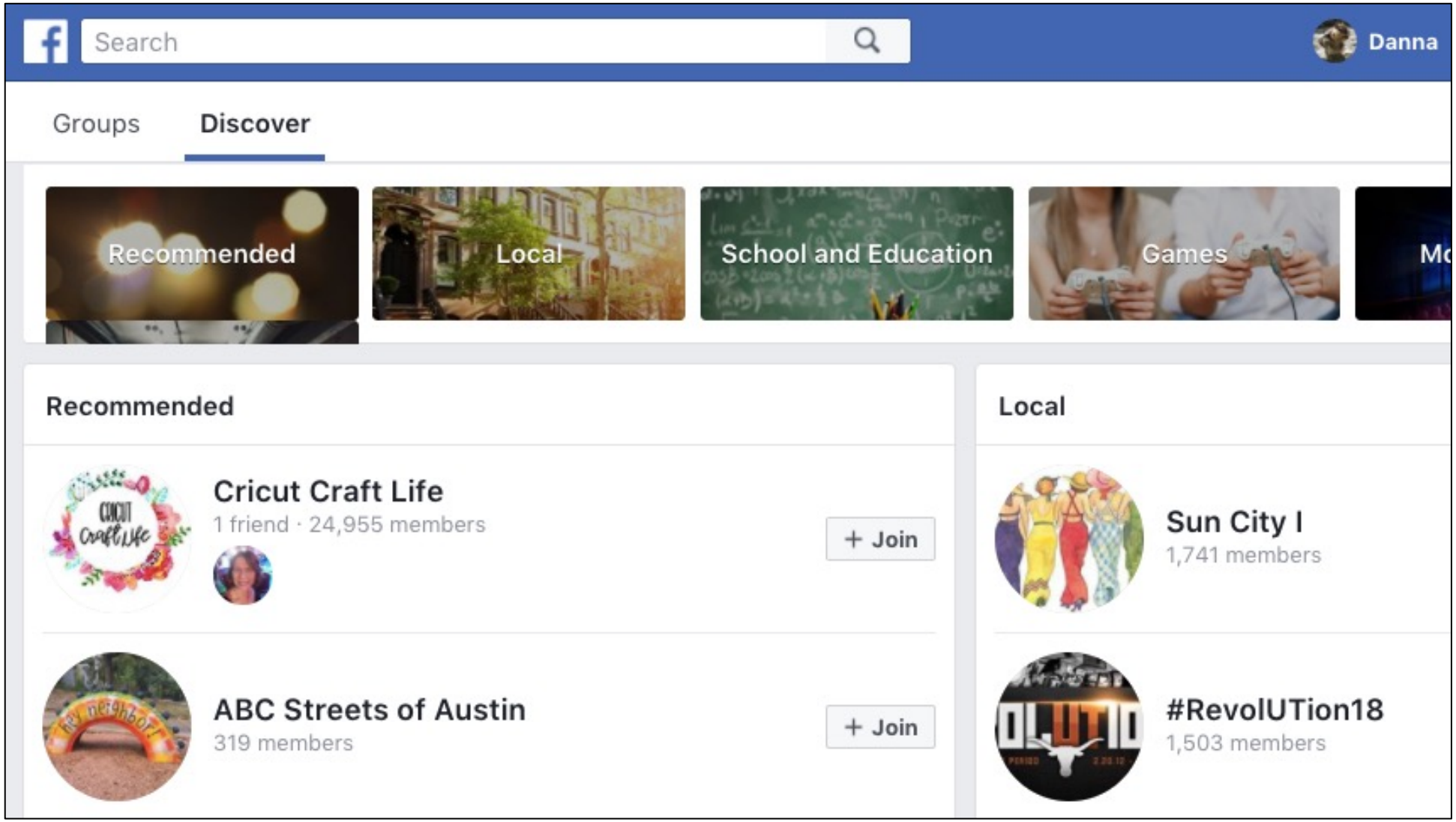
AI Challenge/Dataset Search



Entertainment and Shopping Search



Social Media Search



What are other potential applications for neural search?

Today's Topics

- Speech Processing – Problem and Applications
- Speech Recognition – Evaluation and Models
- Informal Retrieval – Problem and Applications
- Informal Retrieval – Models
- Video making tutorial

NN Popularity at Top IR Conference

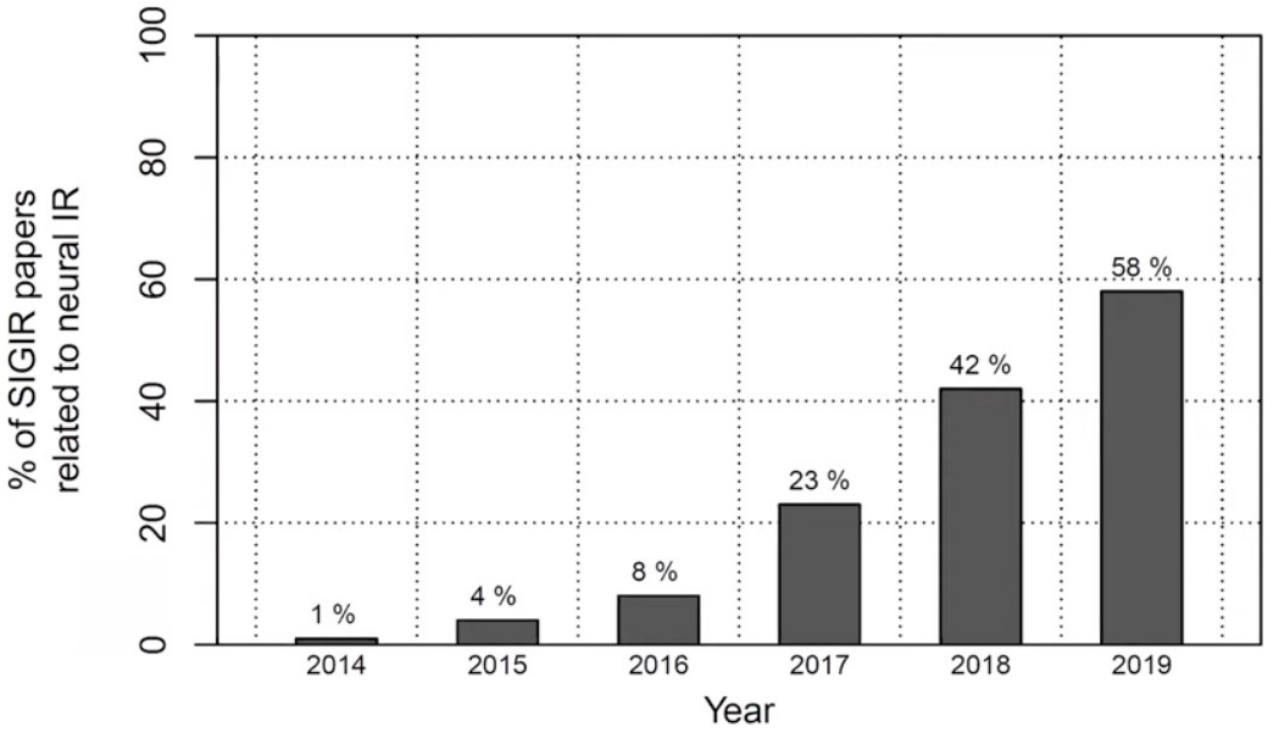
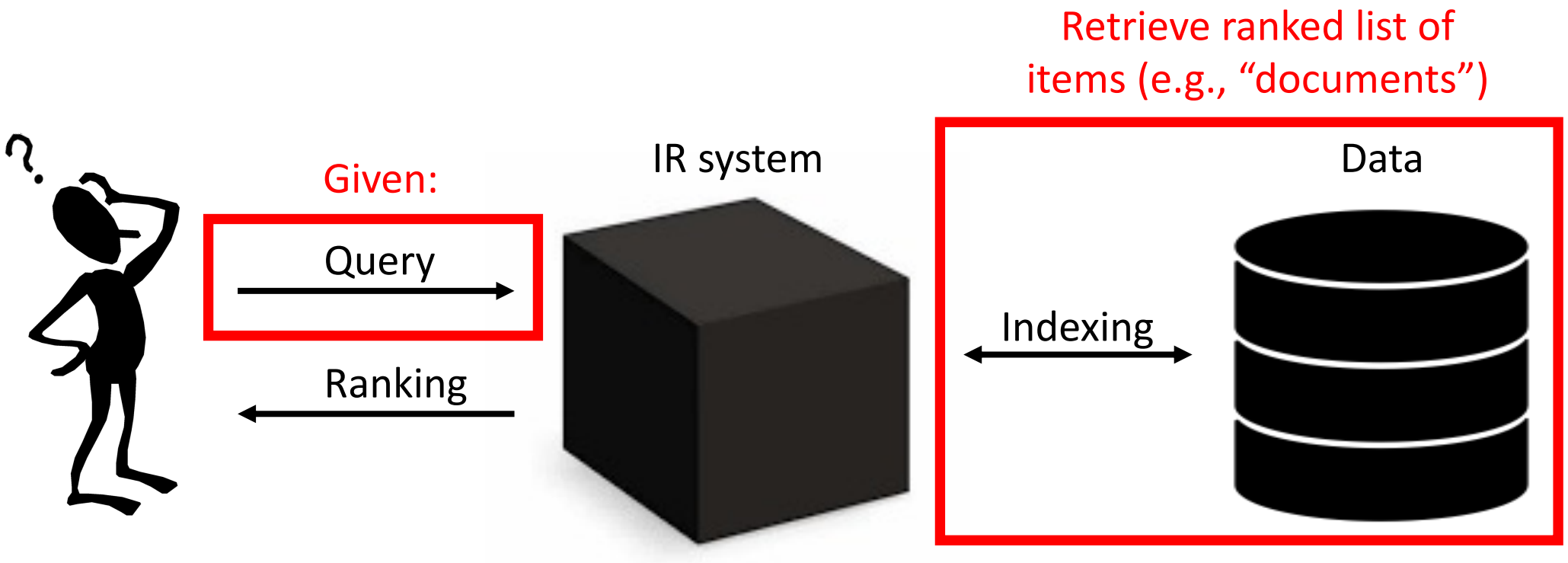


Figure 1.1: The percentage of neural IR papers at the ACM SIGIR conference—as determined by a manual inspection of the papers—shows a clear trend in the growing popularity of the field.

Source: https://www.youtube.com/watch?v=y-6OJzLZgEE&t=185s

Excellent talk summarizing progress: https://www.microsoft.com/en-us/research/video/neural-information-retrieval-in-search-of-meaningful-progress-ciir-talk-series-university-of-massachusetts-amherst/

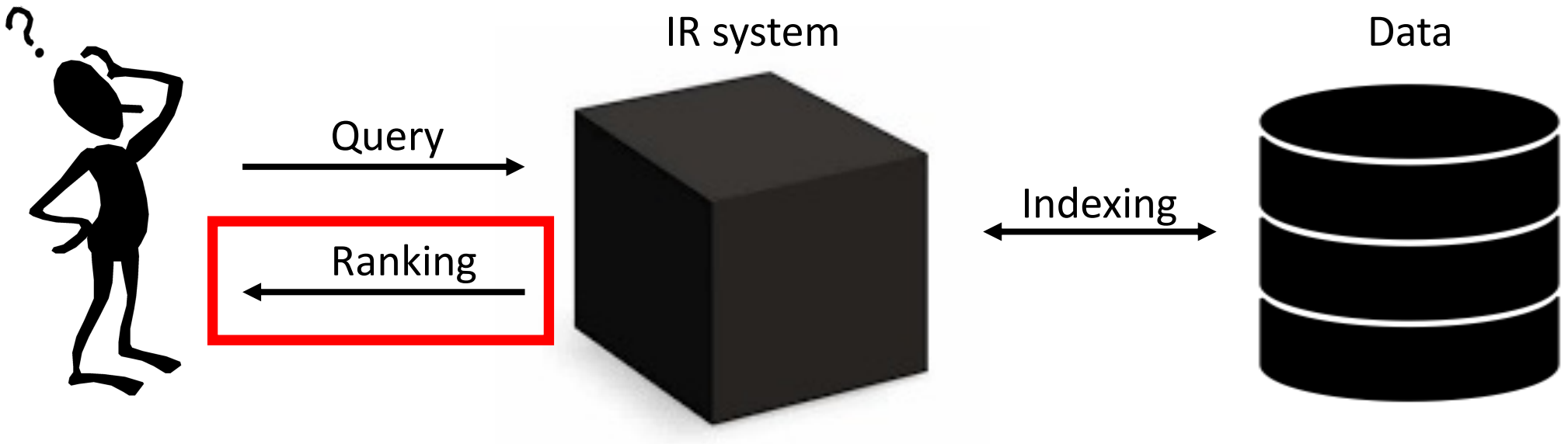
General Approach



1. Establish compact vector representations for query and items in database
2. Establish similarity measure to indicate proximity between query and items
3. Rank similarity of items in database to query

Relevant tutorial: https://www.youtube.com/watch?v=OIsgoC9uWDc

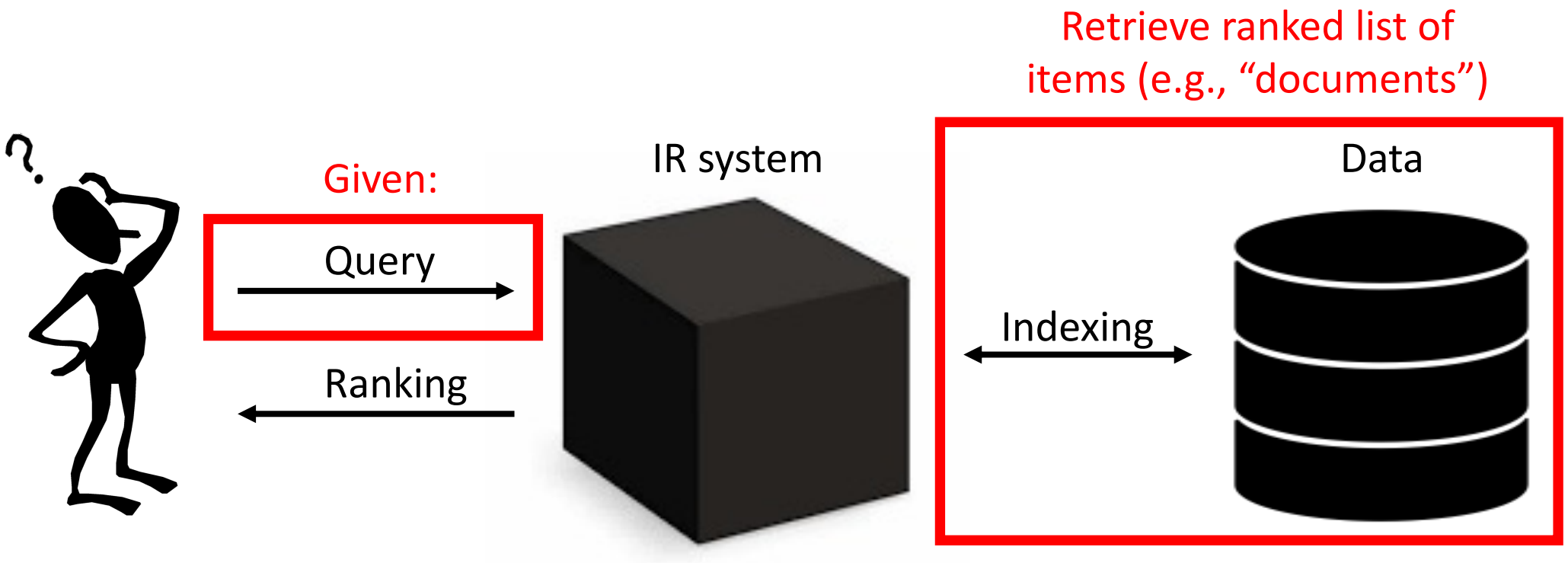
General Approach



How many examples do you want to see when searching?

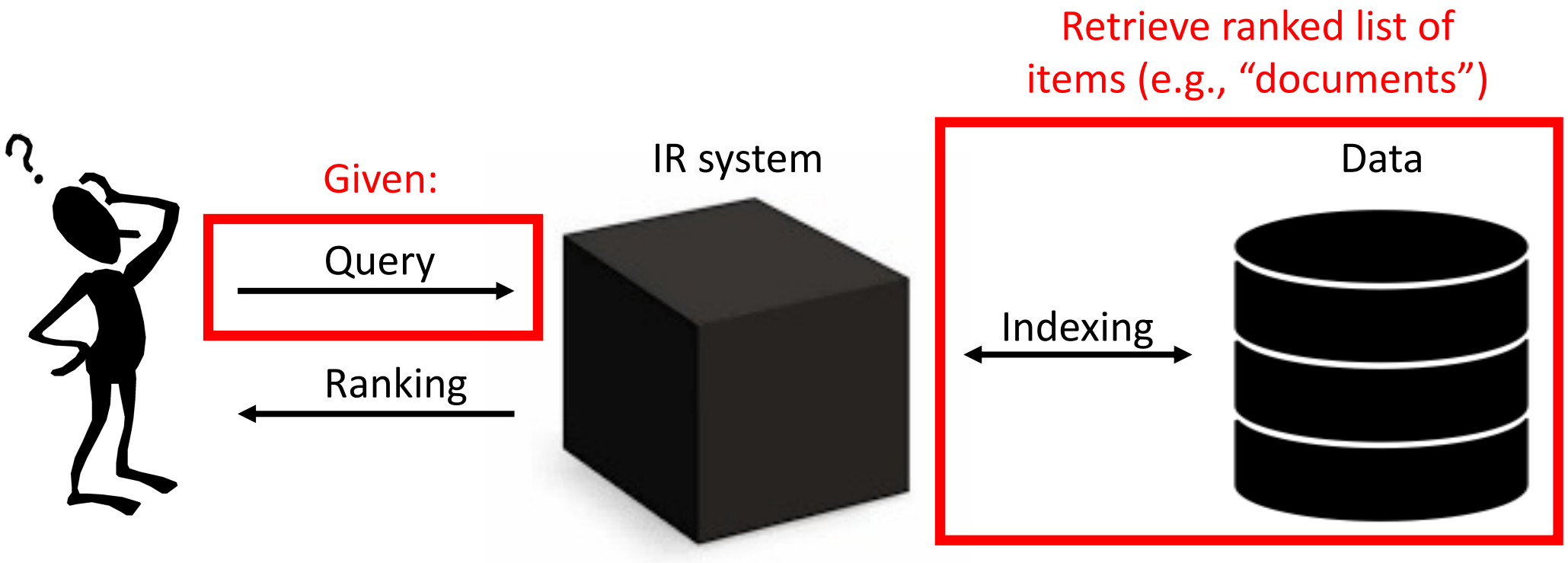
a. 1, b. 2-5, c. 6-10, d. More than 10, e. Undecided

General Approach



Due to scale of many search problems, often a 2-stage approach is used: (1) fast, simple approach generates candidate items and (2) slower, high-quality approach ranks most relevant candidates (e.g., NNs)

General Approach

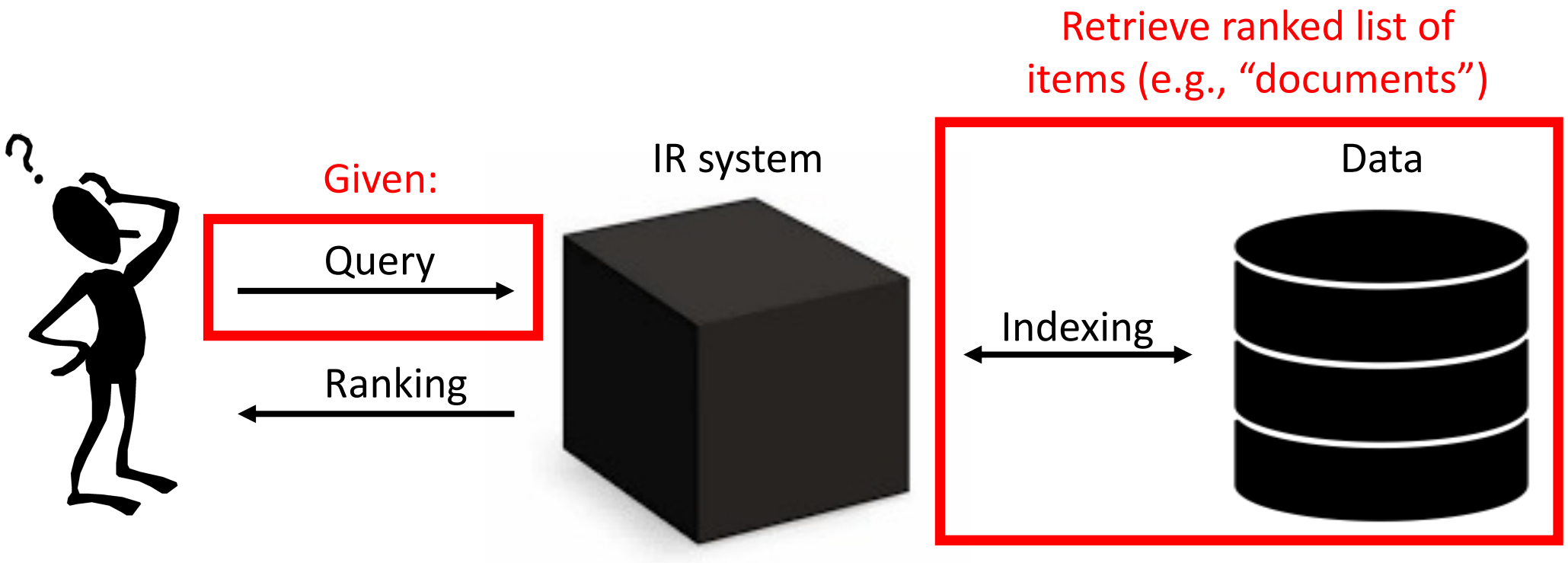


Need suitable objective function for learning to rank items (what to optimize?)

- e.g., likelihood of user favoriting, clicking, buying, adding to cart, etc
- can use standard objective functions for regression and classification

Relevant tutorial: https://www.youtube.com/watch?v=1uut7IGStw8

e.g., Transformer–Based Approaches



Can sort items based on probability each item belongs to the query category (e.g., using probability of relevance from cross entropy score of [CLS] token when feeding fine-tuned BERT a query followed by item)

Nogueira and Cho. Passage Re-Ranking with BERT. arXiv 2019.
Yates, Nogueira, and Lin. Pretrained Transformers for Text Ranking: BERT and Beyond. WSDM 2021.

e.g., Transformer–Based Approaches

SEJ Search Engine Journal LATEST SEO PAID MEDIA CONTENT SOCIAL DIGITA

SEJ · News

Bing is Now Utilizing BERT at a Larger Scale Than Google

SEJ STAFF
Matt G. Southern
November 19, 2019 · 2 min read

266 SHARES 4.4K READS

Search Engine Land SEO PPC Platforms SMX Awards Webinars Intelligence Reports White Papers About

Google: BERT now used on almost every English query

Google announced numerous improvements made to search over the year and some new features coming soon.

Barry Schwartz on October 15, 2020 at 3:17 pm | Reading time: 3 minutes

BERT powers almost every single English based query done on Google Search, the company said during its virtual Search on 2020 event Thursday. That's up from just 10% of English queries when Google first announced the use of the BERT algorithm in Search last October.

In December 2019, Google expanded the use of BERT to over 70 languages.

Note, this wouldn't impact a site's ranking exactly. SEOs cannot optimize for BERT per se. Instead, BERT is designed to improve the relevancy of search results by better understanding the content on web pages.

To learn more about how the algorithm works, see our deep dive on Google BERT.

Google also explained that it has improved results on "specific searches" by 7%. Google did this through different AI and machine learning techniques. Google said it is also improving search results and answers on more "broader searches."

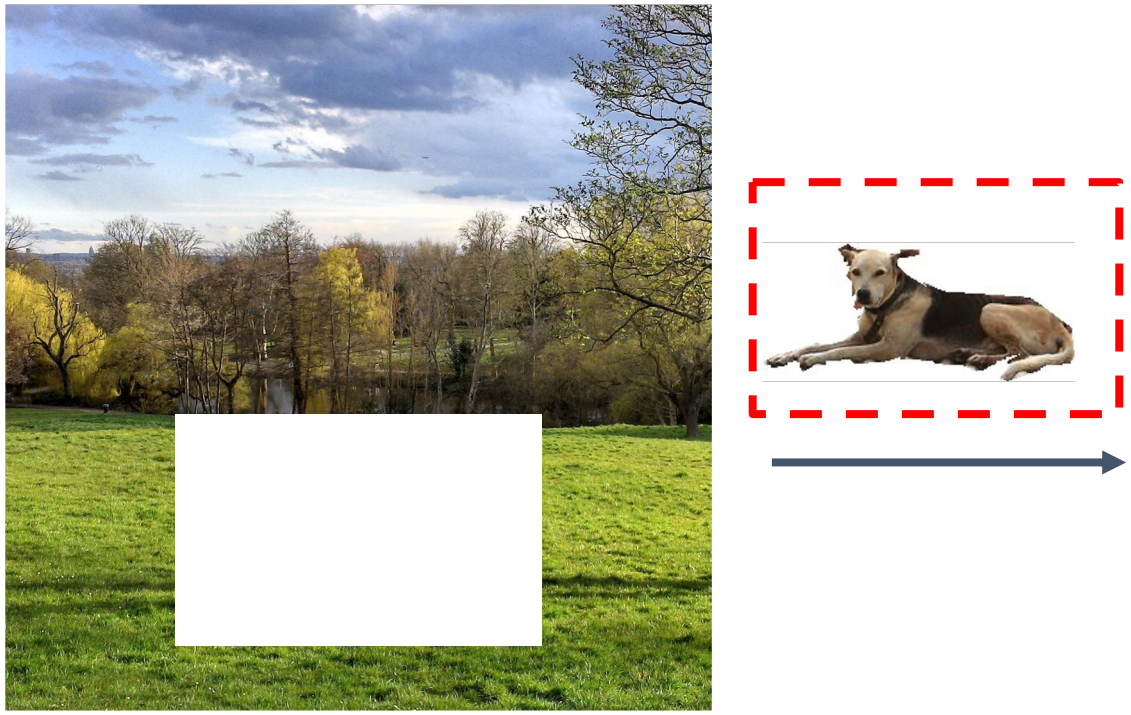
e.g., Searching for Objects to Inpaint

Hole-filling approach: find object, position it, and fill remainder of hole

Search Automatically

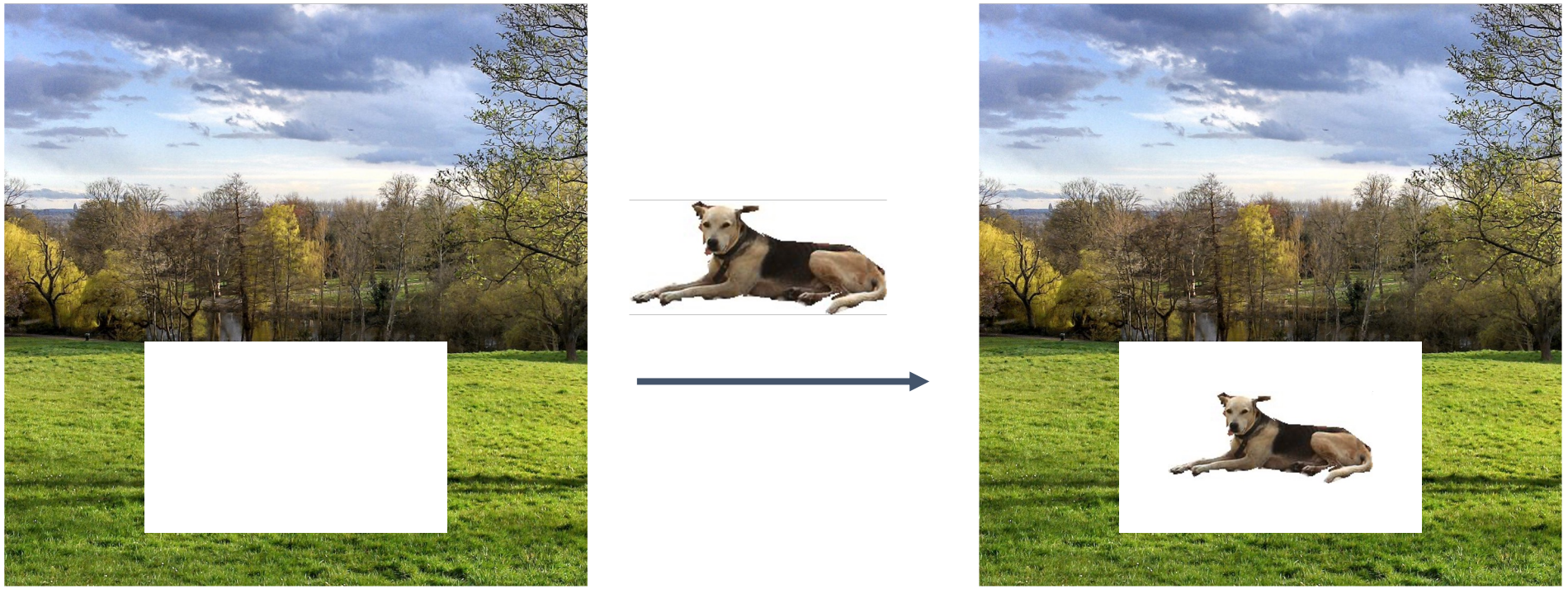
e.g., Searching for Objects to Inpaint

Hole-filling approach: find object, position it, and fill remainder of hole



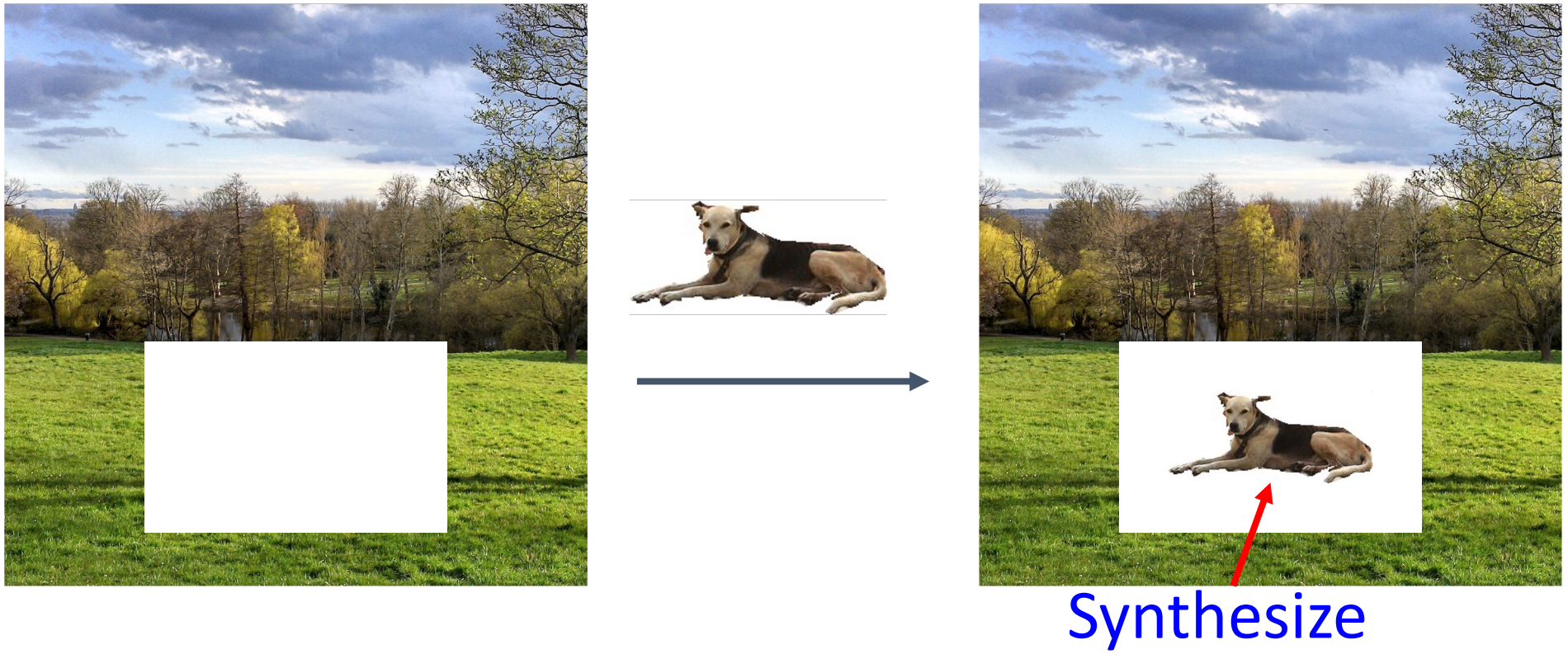
e.g., Searching for Objects to Inpaint

Hole-filling approach: find object, position it, and fill remainder of hole



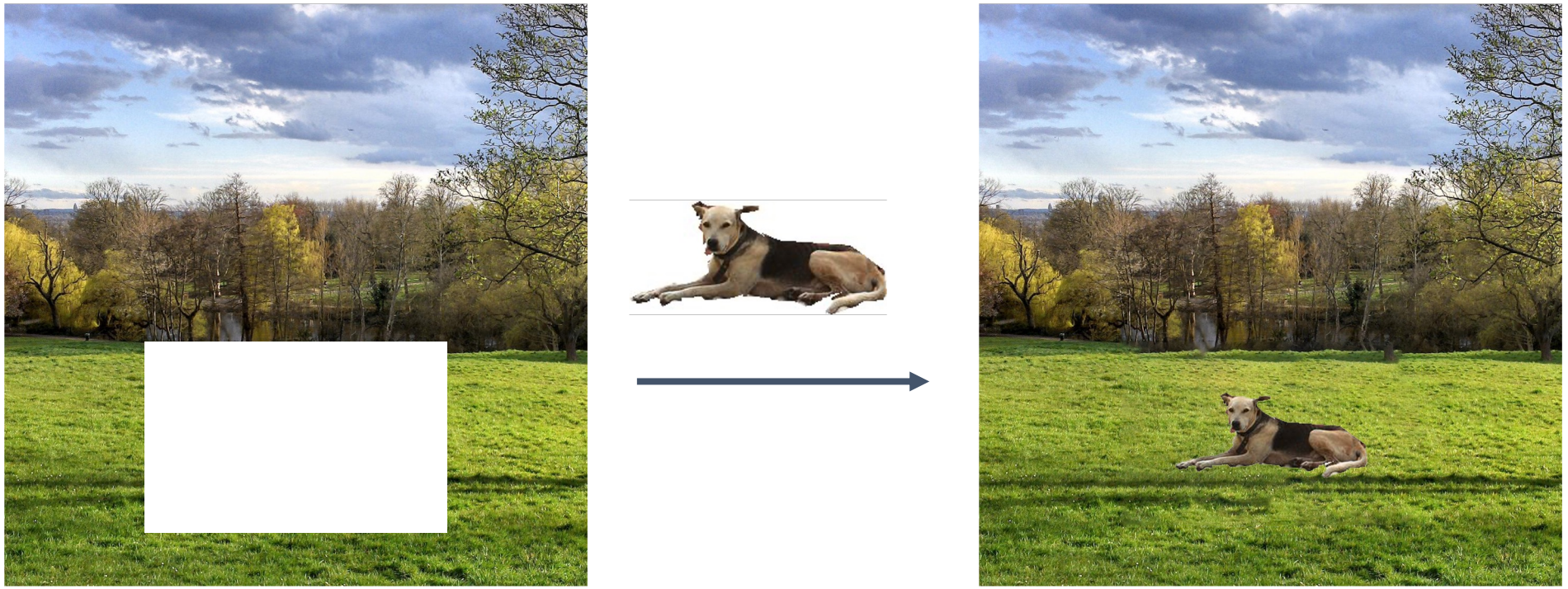
e.g., Searching for Objects to Inpaint

Hole-filling approach: find object, position it, and fill remainder of hole

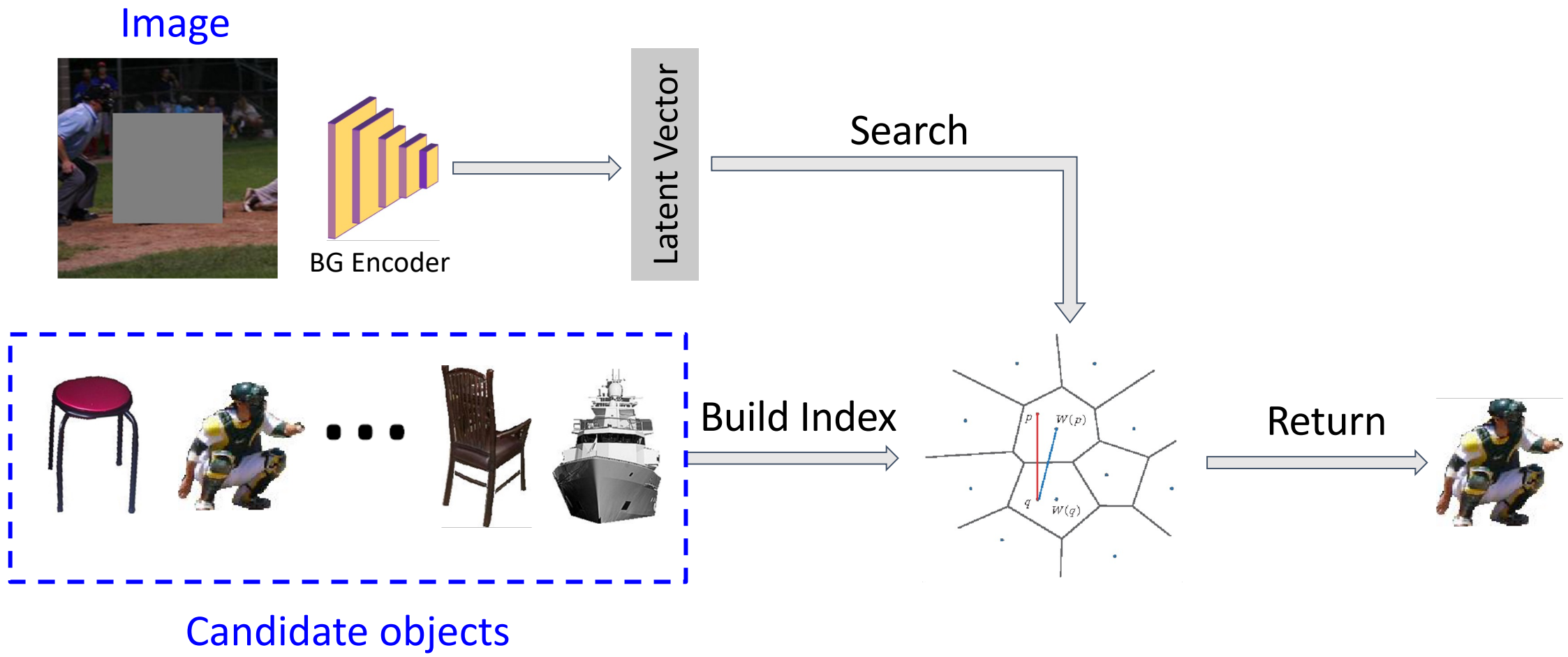


e.g., Searching for Objects to Inpaint

Hole-filling approach: find object, position it, and fill remainder of hole

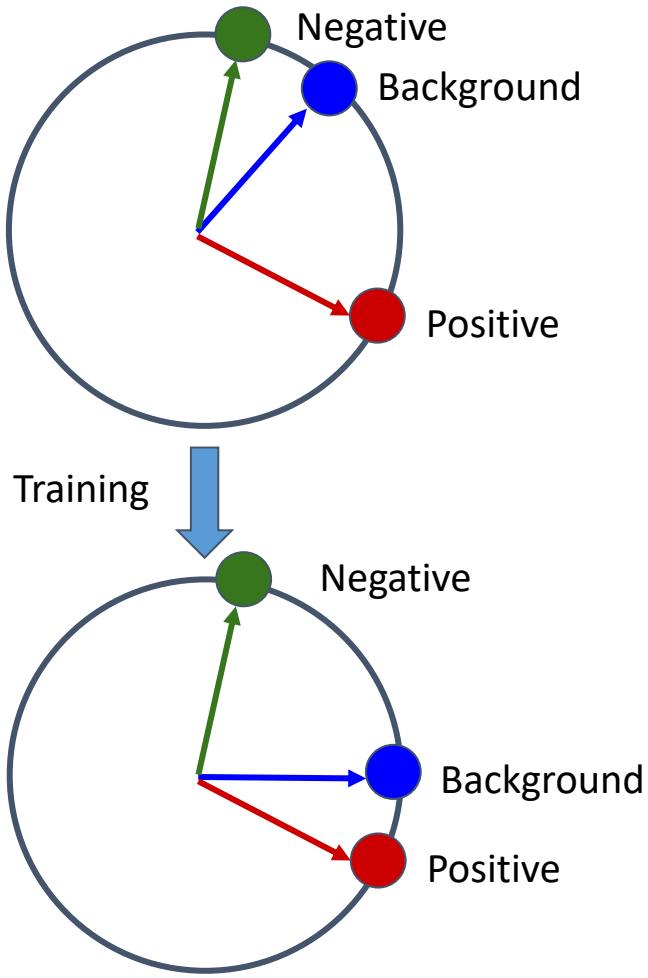


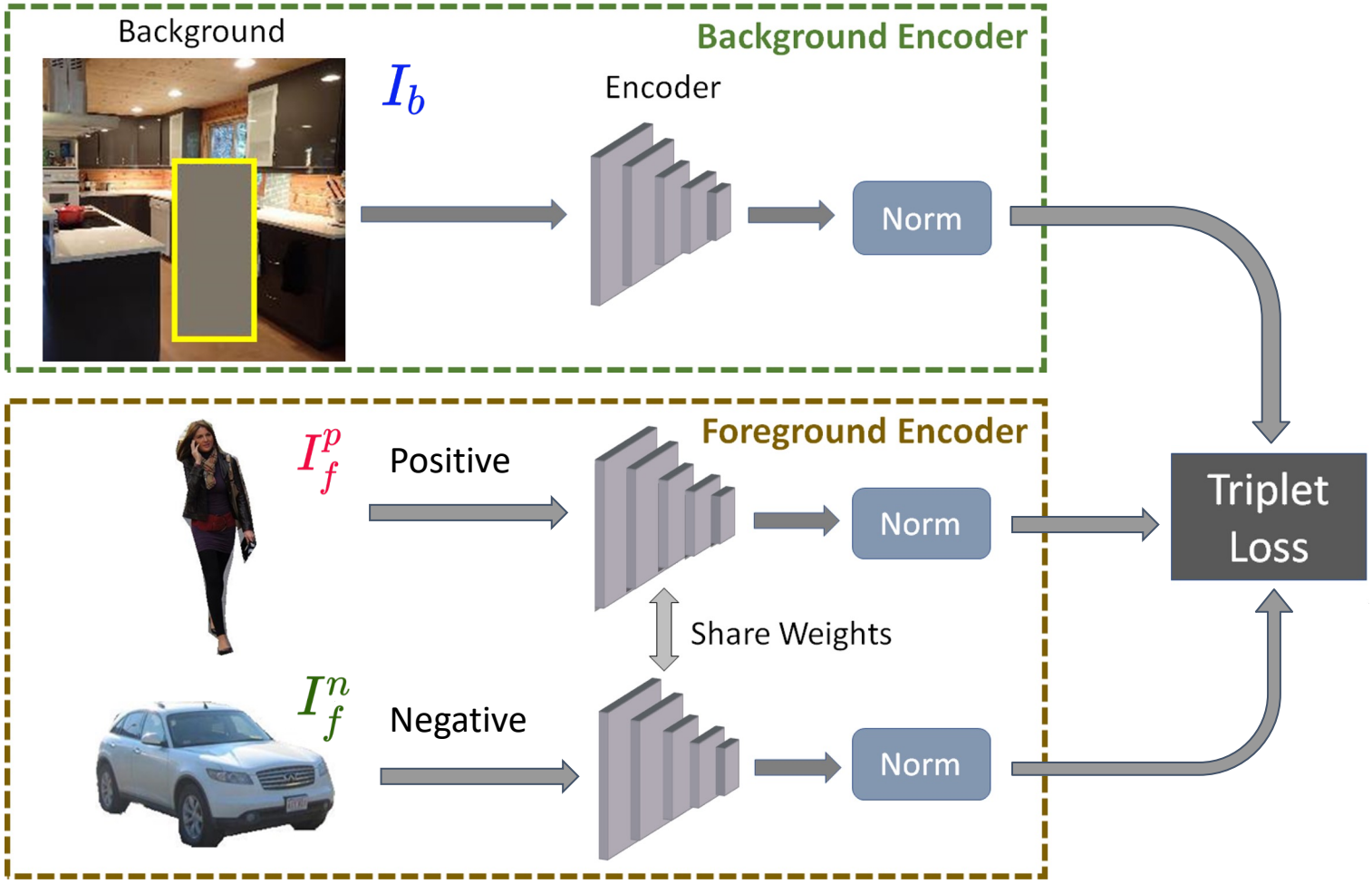
Solution: CNN Encodes the Image and Candidate Objects in the Same Latent Space So Compatible Ones are Closer



[Zhao et al. ICCV 2019]

UFO Search Training





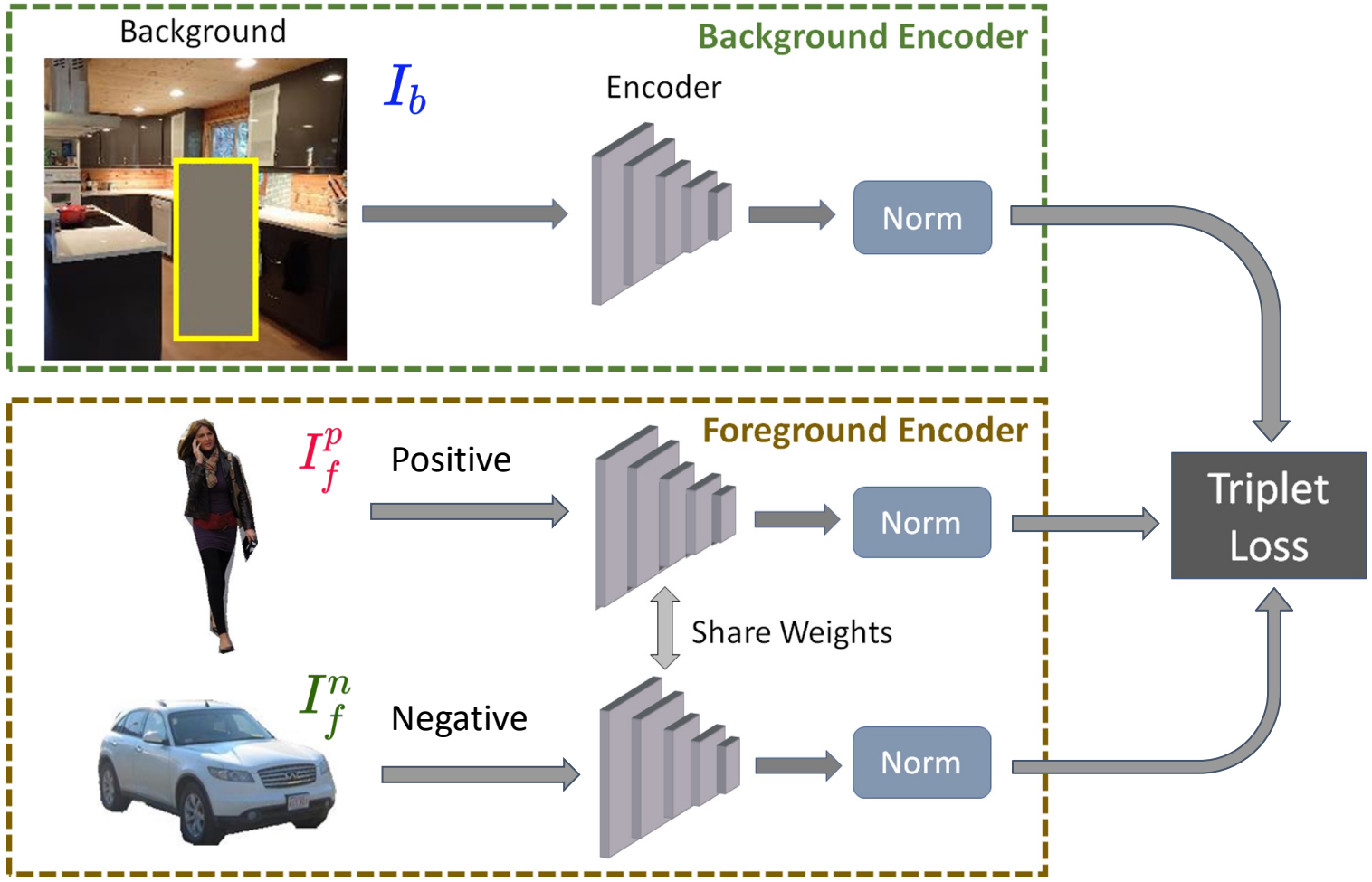
Positive objects are pushed closer to the background image than negative ones

[Zhao et al. ICCV 2019]

Method: Base Architecture

We want:

$$C(I_b, I_f^p) > C(I_b, I_f^n) + M$$



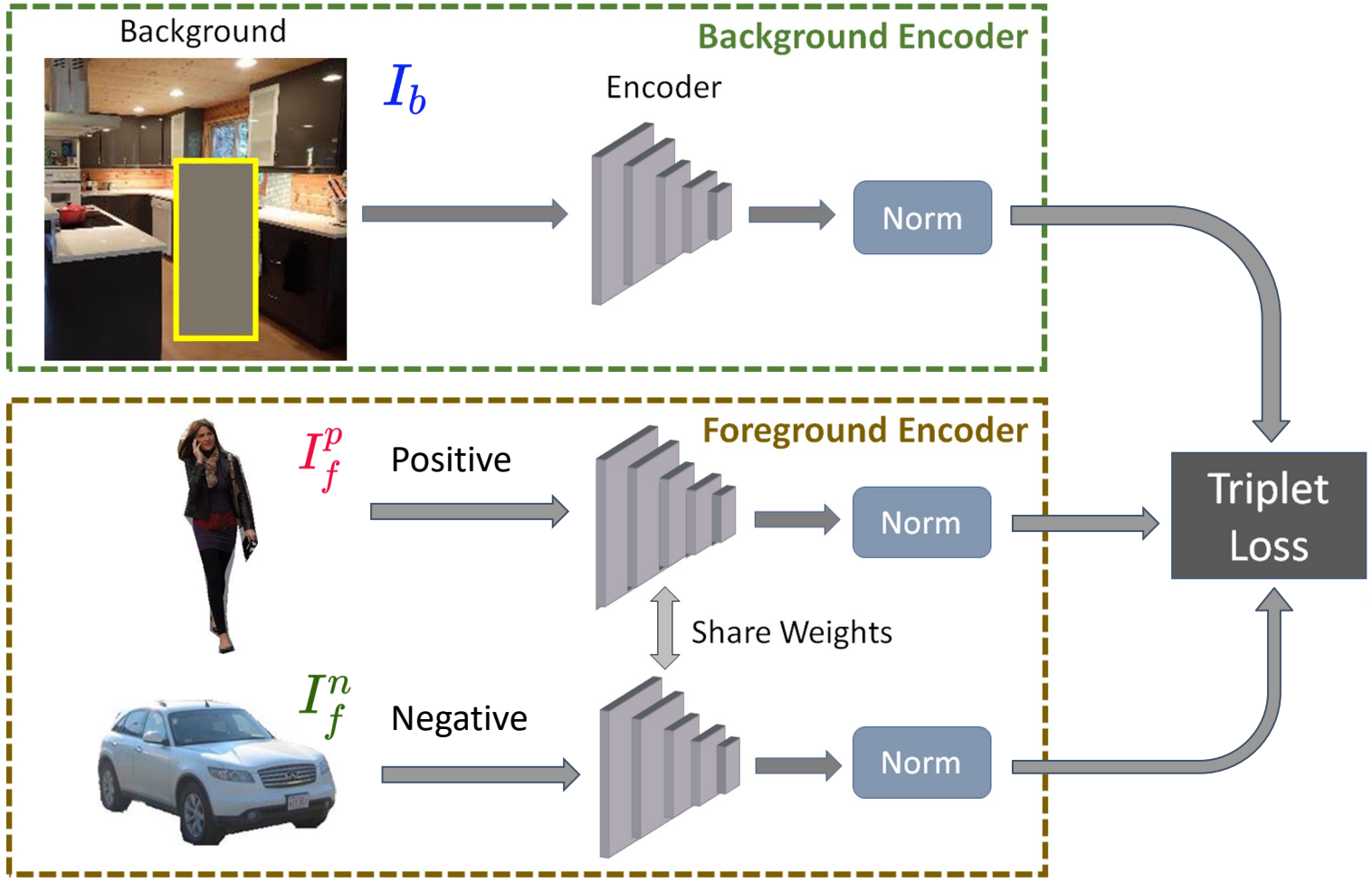
Method: Base Architecture

We want:

$$C(I_b, I_f^p) > C(I_b, I_f^n) + M$$

We use the triplet loss:

$$max(0, C(I_b, I_f^n) + M - C(I_b, I_f^p))$$



Method: Base Architecture

We want:

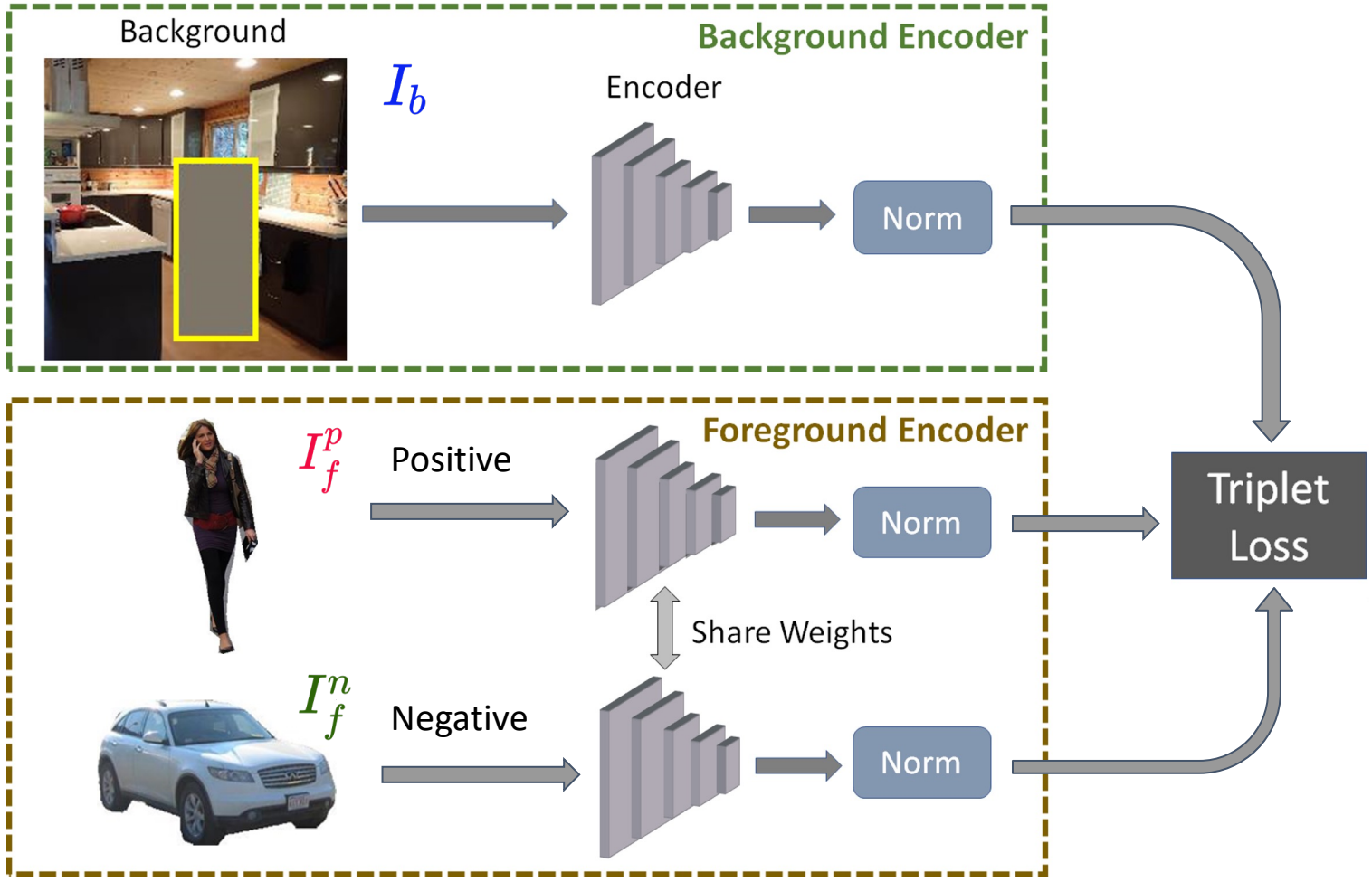
$$C(I_b, I_f^p) > C(I_b, I_f^n) + M$$

We use the triplet loss:

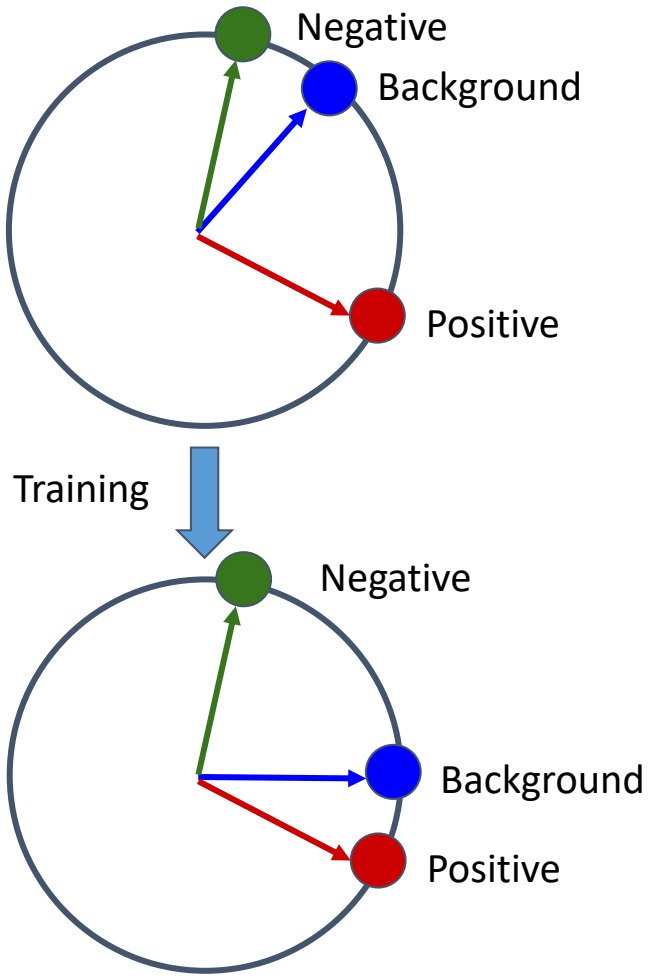
$$max(0, C(I_b, I_f^n) + M - C(I_b, I_f^p))$$

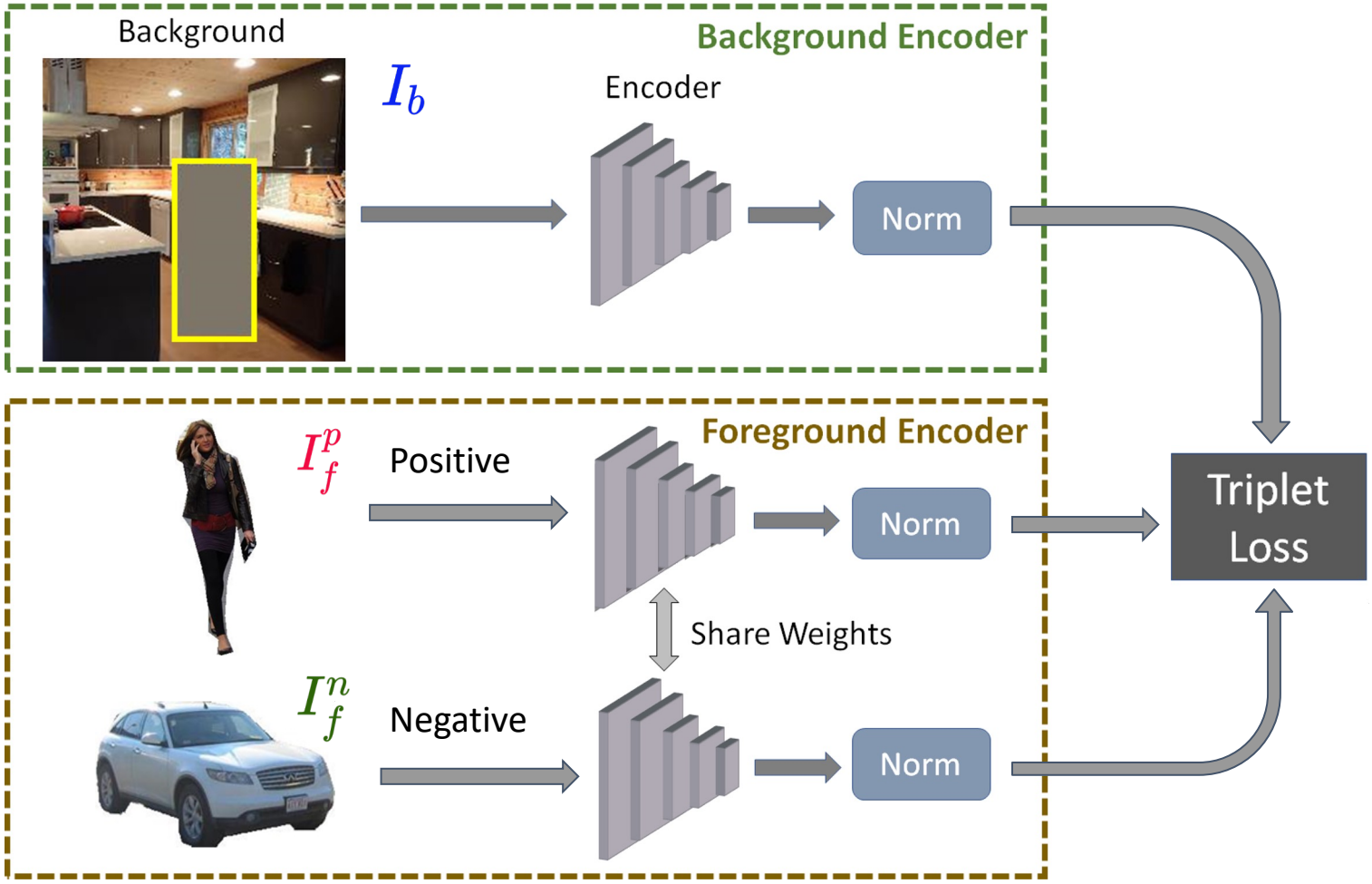
We define compatibility as:

$$C(I_b, I_f^p) = cos(f_b, f_f^p)$$



UFO Search Training



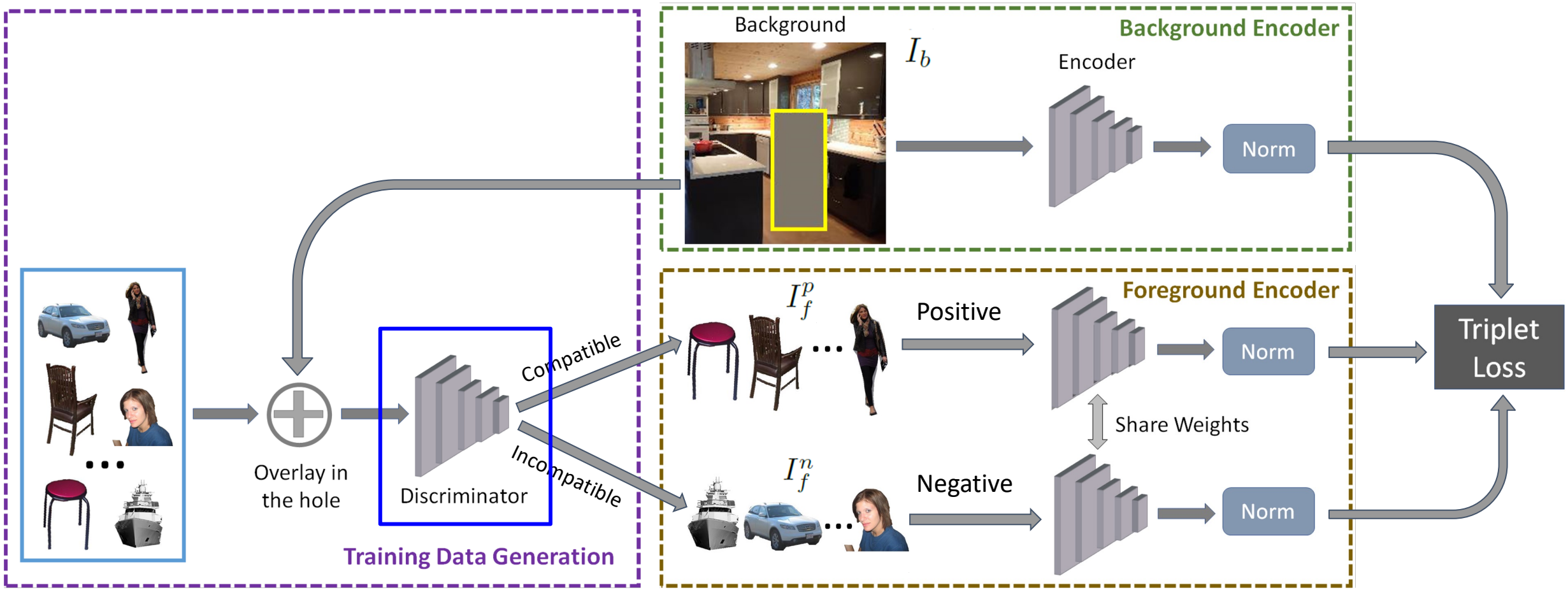


Key challenge: how to obtain training data for positive objects?

(only one positive object known for each background image!)

[Zhao et al. ICCV 2019]

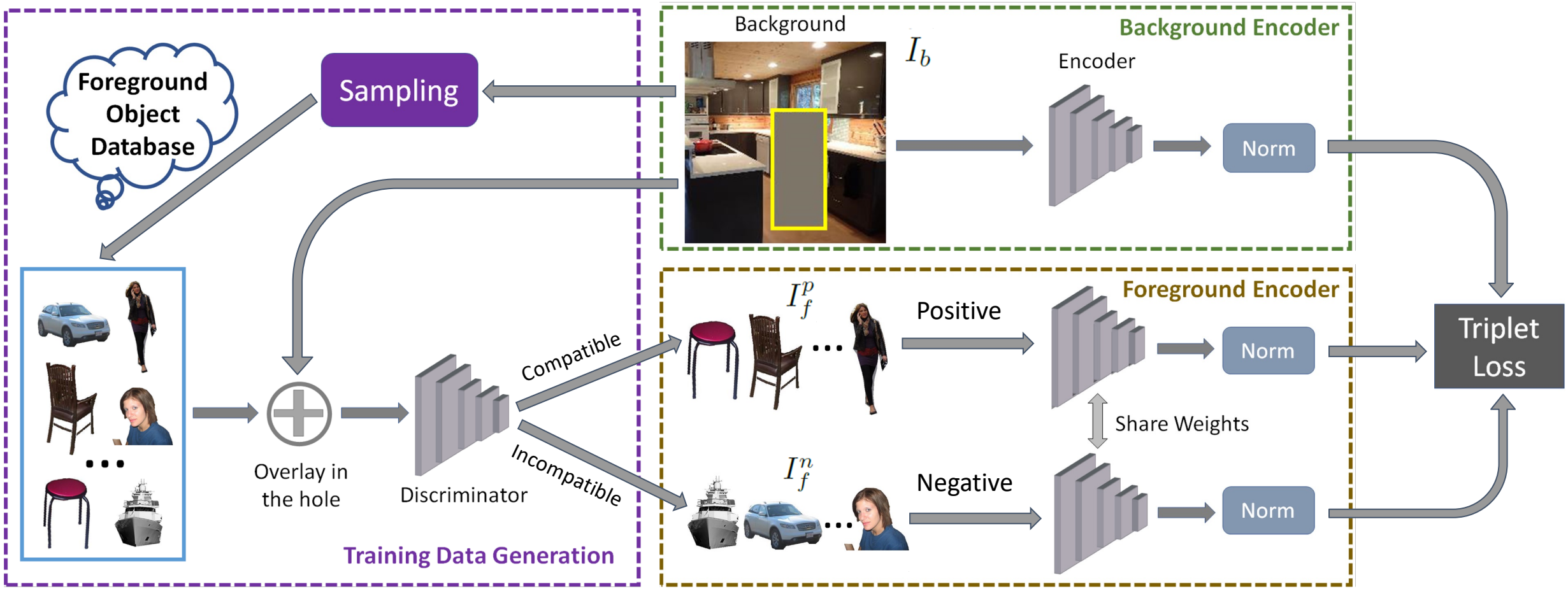
UFO Search Training



Key idea: use a noisy discriminator to offer a richer training set of positive objects

[Zhao et al. ICCV 2019]

UFO Search Training



Key idea: use a noisy discriminator to offer a richer training set of positive objects with a sampling heuristic to accelerate finding examples

[Zhao et al. ICCV 2019]

UFO Search Training

- Similar to the original object

[Zhao et al. ICCV 2019]

UFO Search Training

- Similar to the original object

[Zhao et al. ICCV 2019]

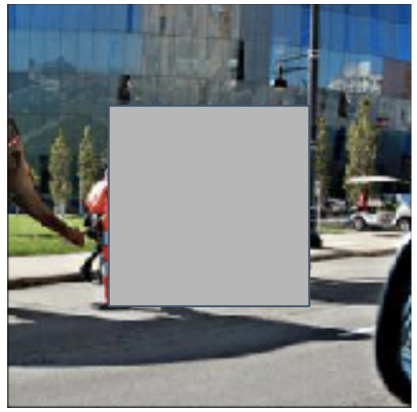
UFO Search Training

Similar Objects

[Zhao et al. ICCV 2019]

UFO Search Training

- Similar Background



[Zhao et al. ICCV 2019]

UFO Search Training

- Similar Background

[Zhao et al. ICCV 2019]

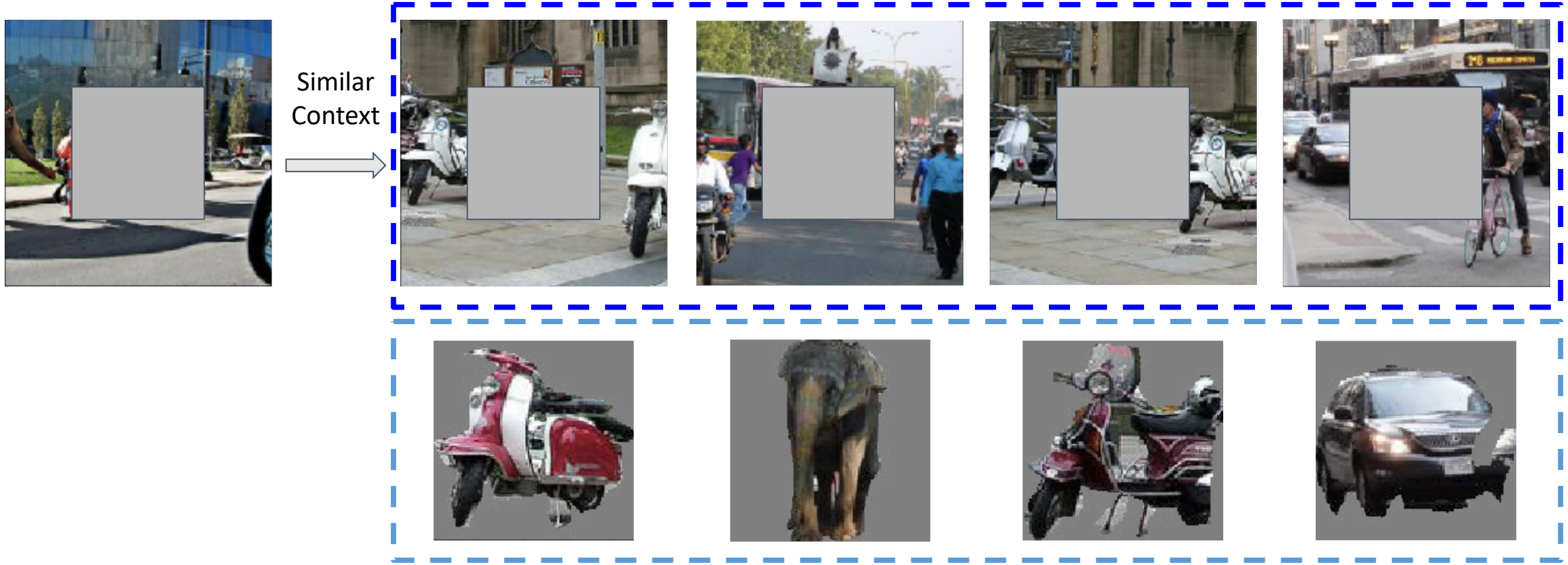
UFO Search Training

- Similar Background

[Zhao et al. ICCV 2019]

UFO Search Training

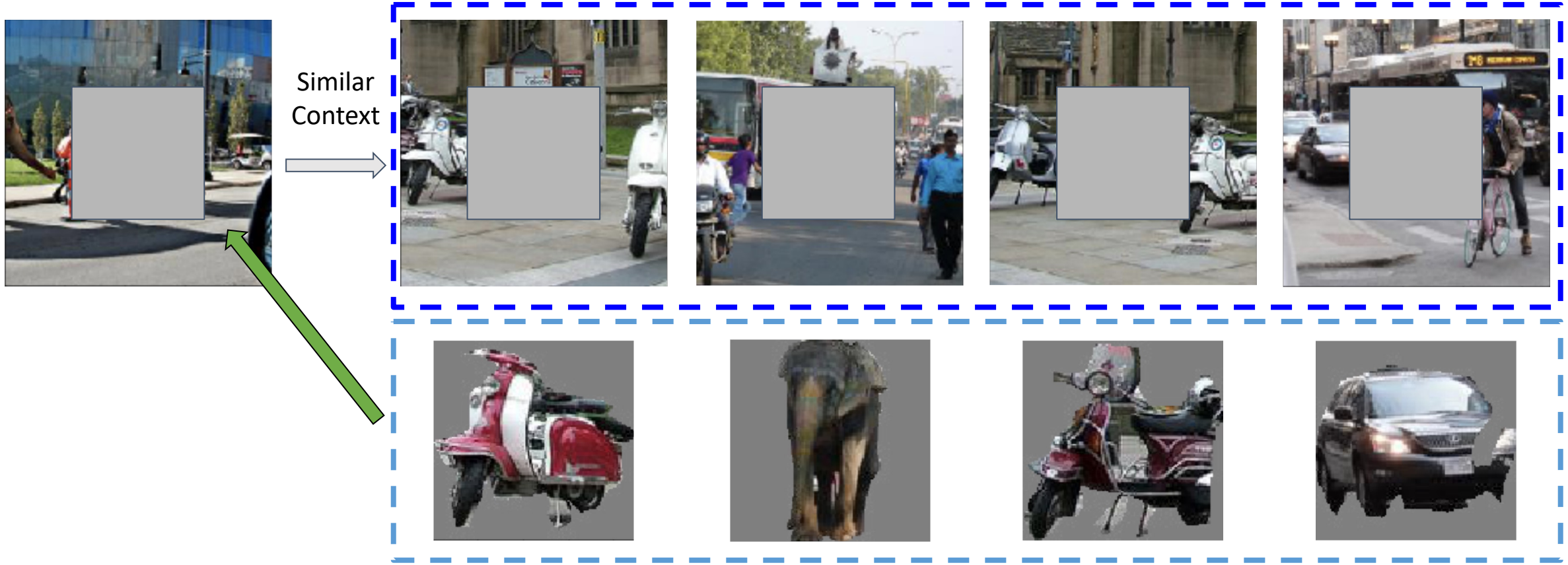
- Similar Background



[Zhao et al. ICCV 2019]

UFO Search Training

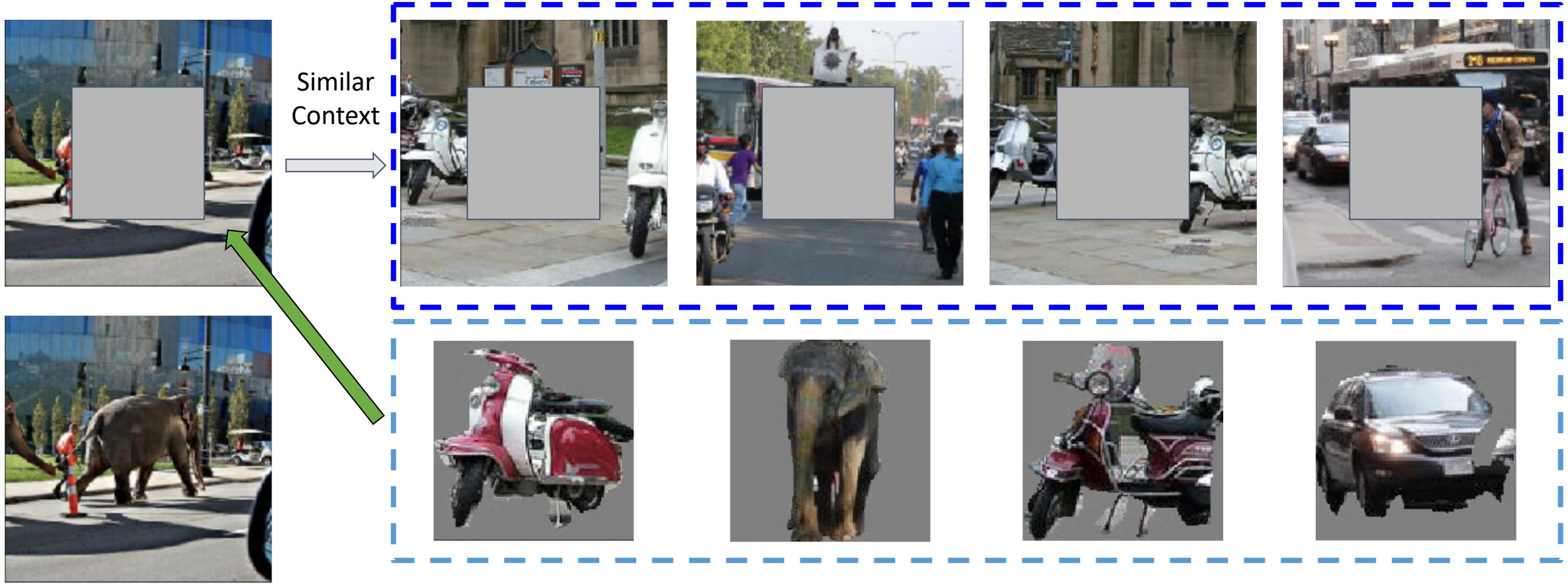
- Similar Background



[Zhao et al. ICCV 2019]

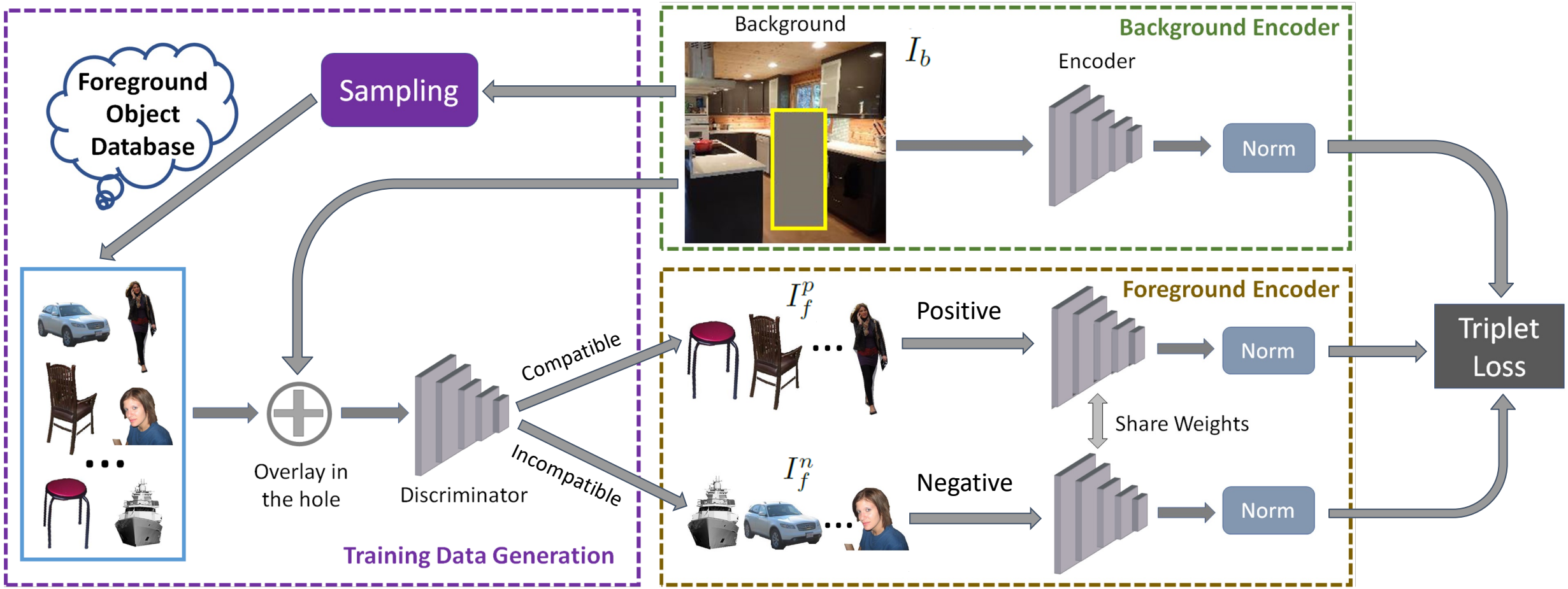
UFO Search Training

- Similar Background



[Zhao et al. ICCV 2019]

UFO Search Training



Note: while the discriminator takes a background image with overlaid object to predict "yes/no", our final method de-couples the images and objects and indicates their distance

UFO Search Performance

Recognizes when to return no diversity: only catchers

Quantitative results demonstrate the advantage of this new approach over existing baselines!

Recognizes when to return diversity: various foods

[Zhao et al. ICCV 2019]

Note: DL Is Not Always Appropriate for Search

- e.g., Legal research by Judicata (https://blog.judicata.com/contextualizing-ai-the-cat-and-the-mistaken-hat-f3c445e819ce)

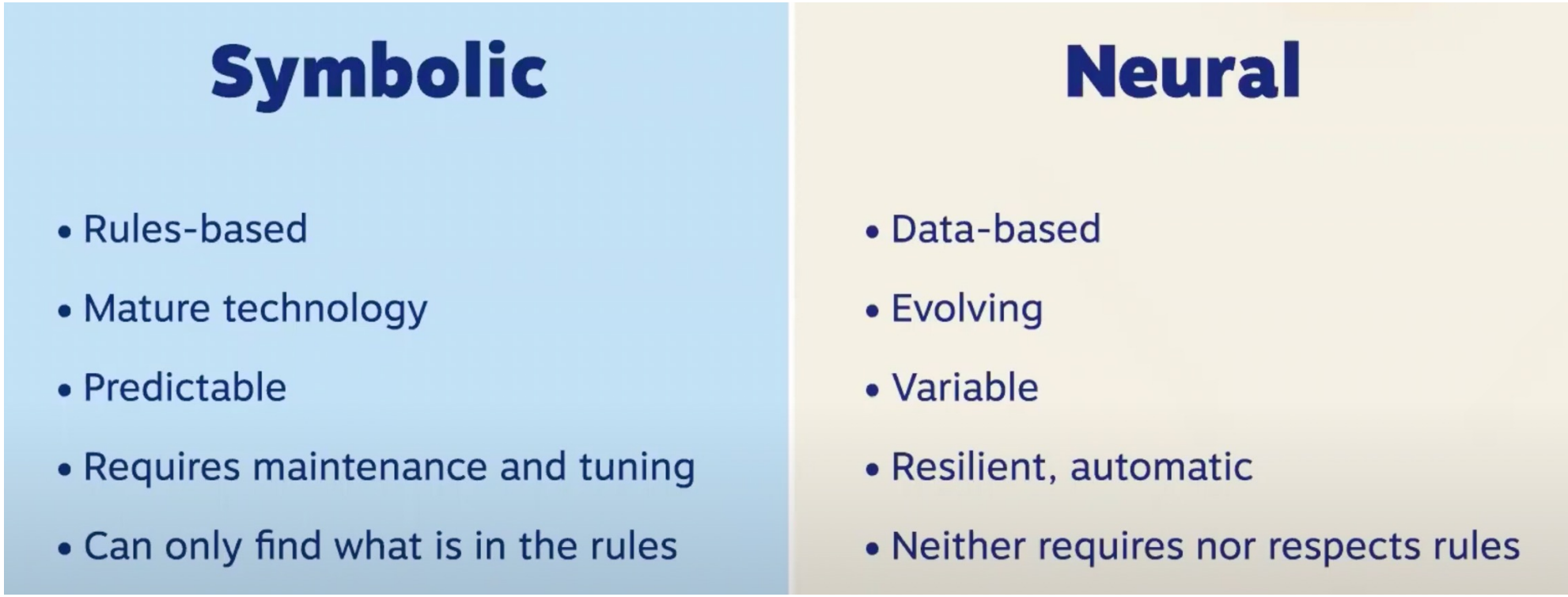


Image Source: https://www.youtube.com/watch?v=OIsgoC9uWDc

Today's Topics

- Speech Processing – Problem and Applications
- Speech Recognition – Evaluation and Models
- Informal Retrieval – Problem and Applications
- Informal Retrieval – Evaluation and Models
- Video making tutorial


The End