

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

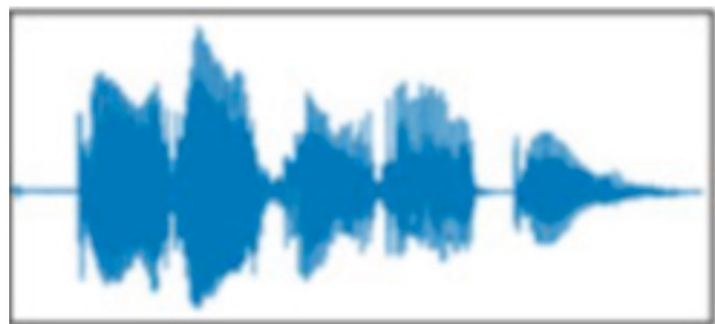
- Speech Processing – Problem and Applications
- Speech Recognition – Evaluation and Models
- Informal Retrieval – Problem and Applications
- Informal Retrieval – Models
- Video making tutorial

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



Raw Speech Signal

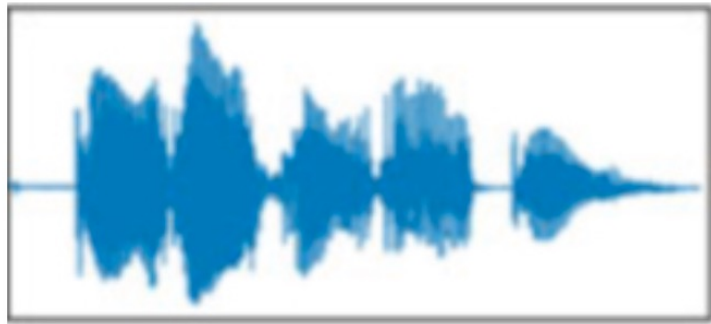


Output: machine readable text

Do you understand me

Transcription

What Is Speech?



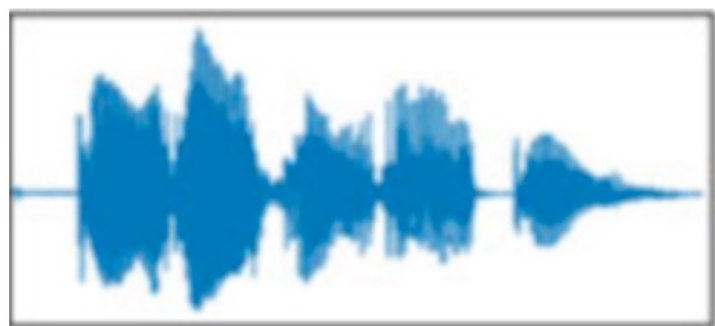
Raw Speech Signal

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

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



Do you understand me

Transcription

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

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

Virtual Assistant



Entertainment
Video

Music, movie, television shows, variety show, short clip, audio book, and broadcasting, etc.



Information
Inquiry

Consultation, 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

Translation, 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.

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



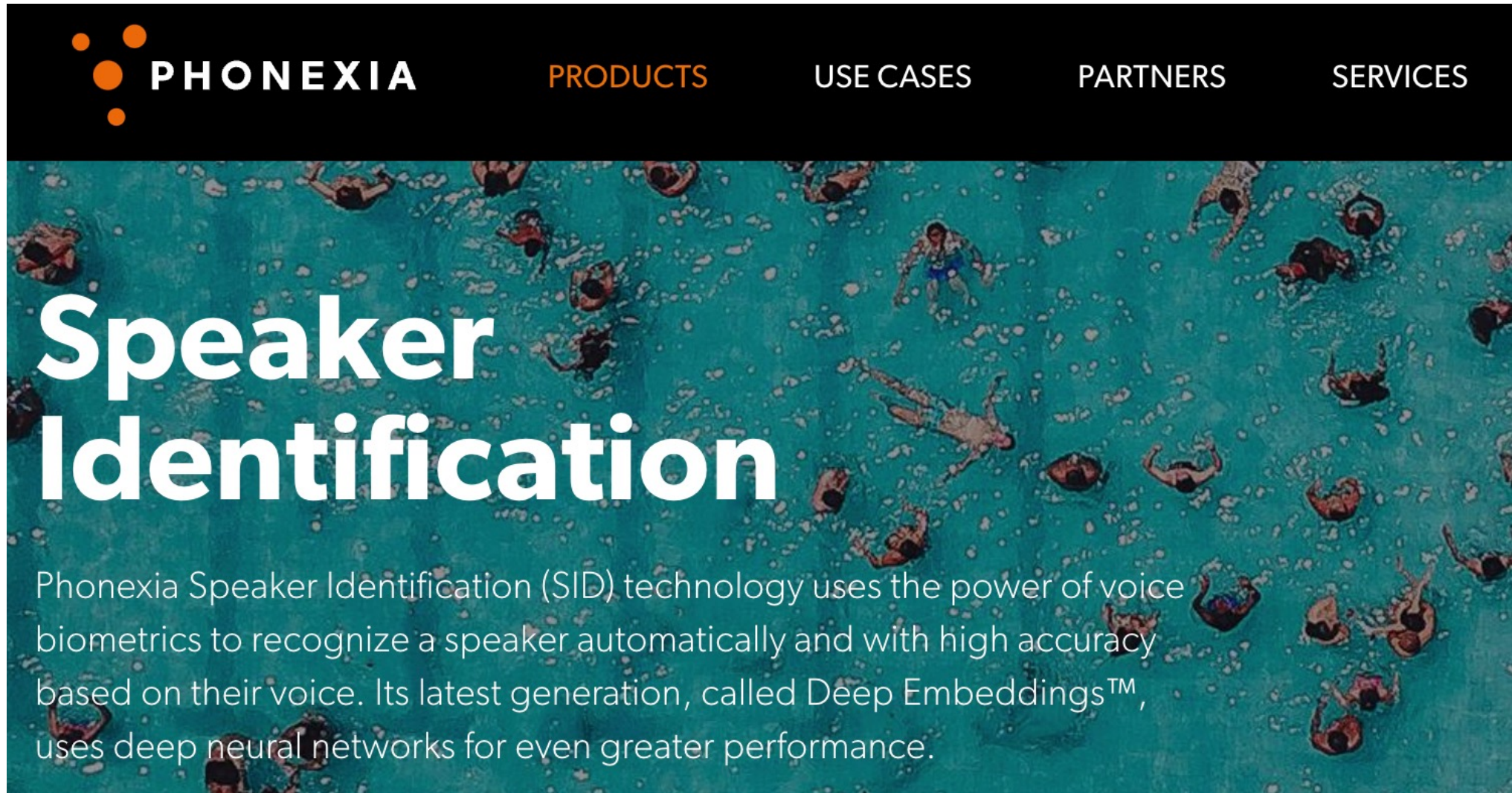
Video/Movie Captioning (e.g., for Translation & Situational/Permanent Hearing Impairments)



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



Speaker Identification (e.g., for Security)

The banner features a dark teal background with a pattern of small, light-colored, irregular shapes. Overlaid on this background are numerous small, stylized human figures in various poses, some appearing to be swimming or floating. The overall aesthetic is modern and tech-oriented.

PHONEXIA PRODUCTS USE CASES PARTNERS SERVICES

Speaker Identification

Phonexia Speaker Identification (SID) technology uses the power of voice biometrics to recognize a speaker automatically and with high accuracy based on their voice. Its latest generation, called Deep Embeddings™, uses deep neural networks for even greater performance.

Language Identification



[Research and Publications](#)

[Contact us](#)

[Visit Translated](#)

Automatic language identifier

Insert any text or pick a random example

Bonjour!

Speech Enhancement



[ABOUT](#) ▾

[FAQ](#)

[DOWNLOAD](#)

**FAKIN'
THE
FUNK
?**



[BUY](#)

[FORUM](#)

[CONTACT](#)

Fakin' The Funk? is a tool that helps you to detect the true quality of your audio files in one batch.

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



Word Error Rate

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

word replaced (e.g., “goose” transcribed as “choose”)	word added (e.g., “caboose” transcribed as “the goose”)	word omitted (e.g., “it is a fun” transcribed as “it is fun”)
		
Substitutions + Insertions + Deletions		
<hr/>		
Number of Words Spoken		

- What indicates better performance: larger or smaller values?

Word Error Rate: Example

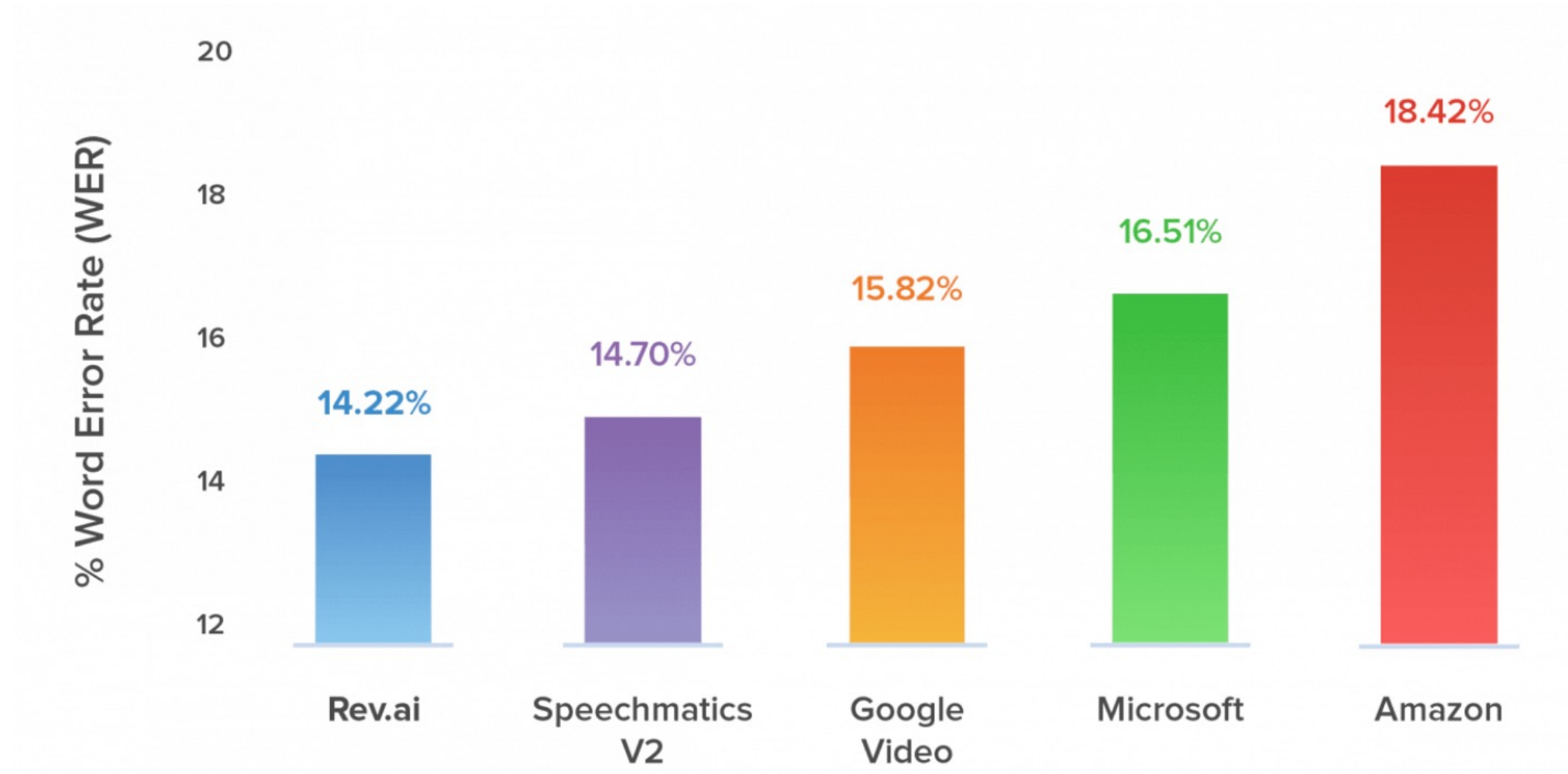
- Correct: The sun makes it look like uh a warm, day to go outside to adventure.
- Predicted: The son makes it to bike with a swarm to go outside to Denver today.
- Number of words spoken?
 - 15
- WER?

Substitutions + Insertions + Deletions

Number of Words Spoken

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

Word Error Rate: Comparison Example

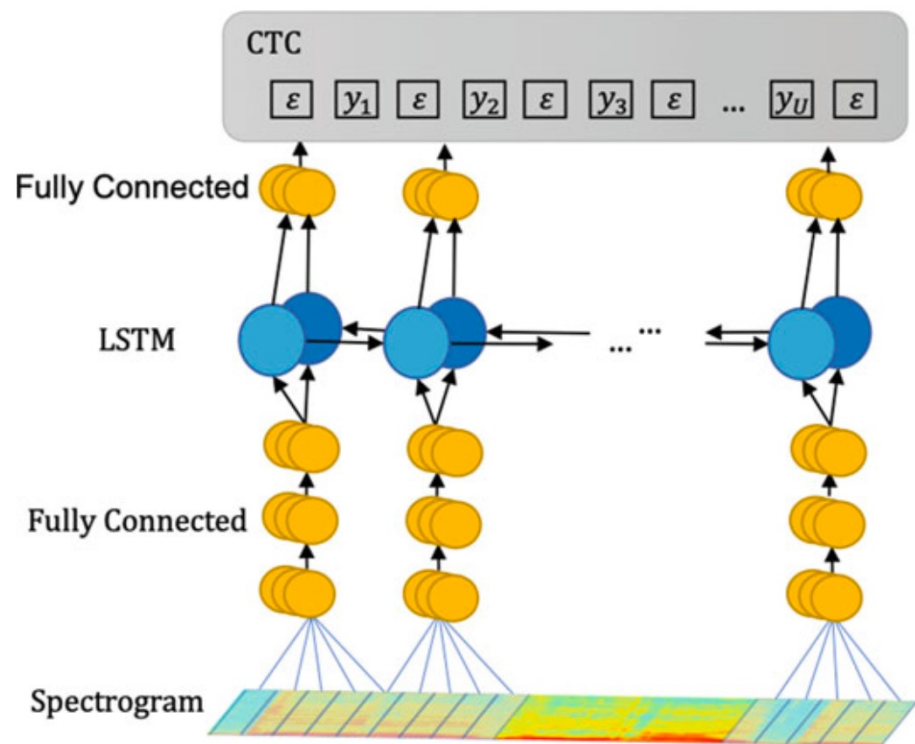


What Are Limitations of Word Error Rate as an Evaluation Metric?

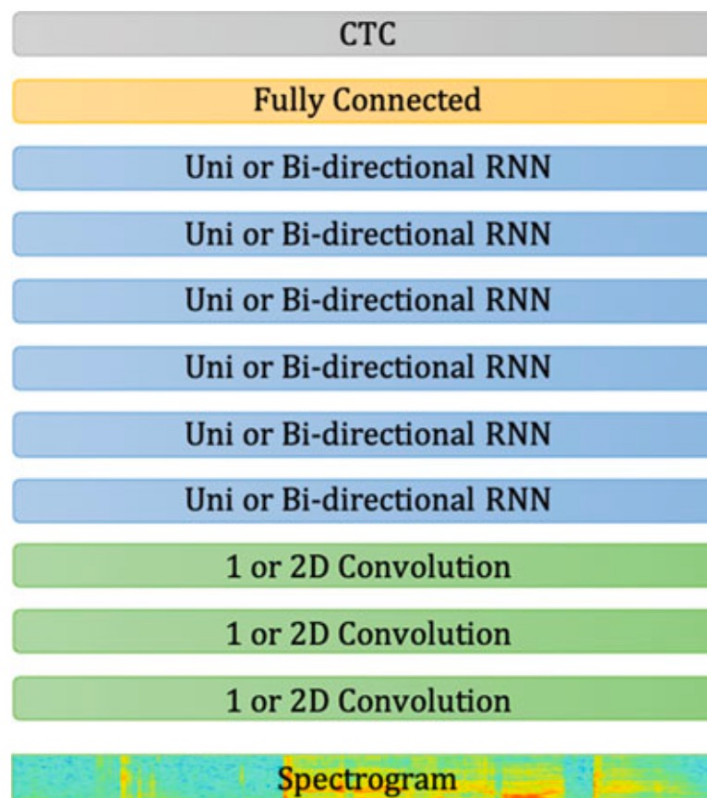
- Does not indicate why errors occur
 - Background noise (e.g., music, other talking)
 - Specialized language (i.e., words reflecting domain expertise)
 - Speaker pronunciations/accent
- Does not reflect whether transcription correctly captures:
 - Capitalization
 - Punctuation
 - Numbers
 - Paragraphs
- May indicate poor quality when humans could understand the content
- Weights all word errors equally

Popular Methods

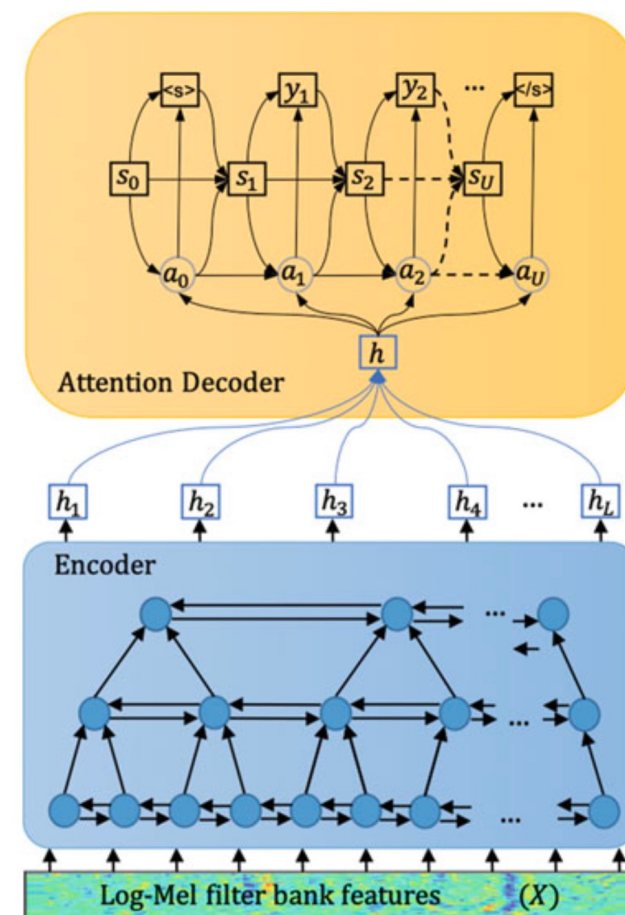
DeepSpeech



DeepSpeech2



Listen, Attend, and Spell



DeepSpeech

Decoder:

Fully Connected

CTC

ϵ y_1 ϵ y_2 ϵ y_3 ϵ ... y_U ϵ

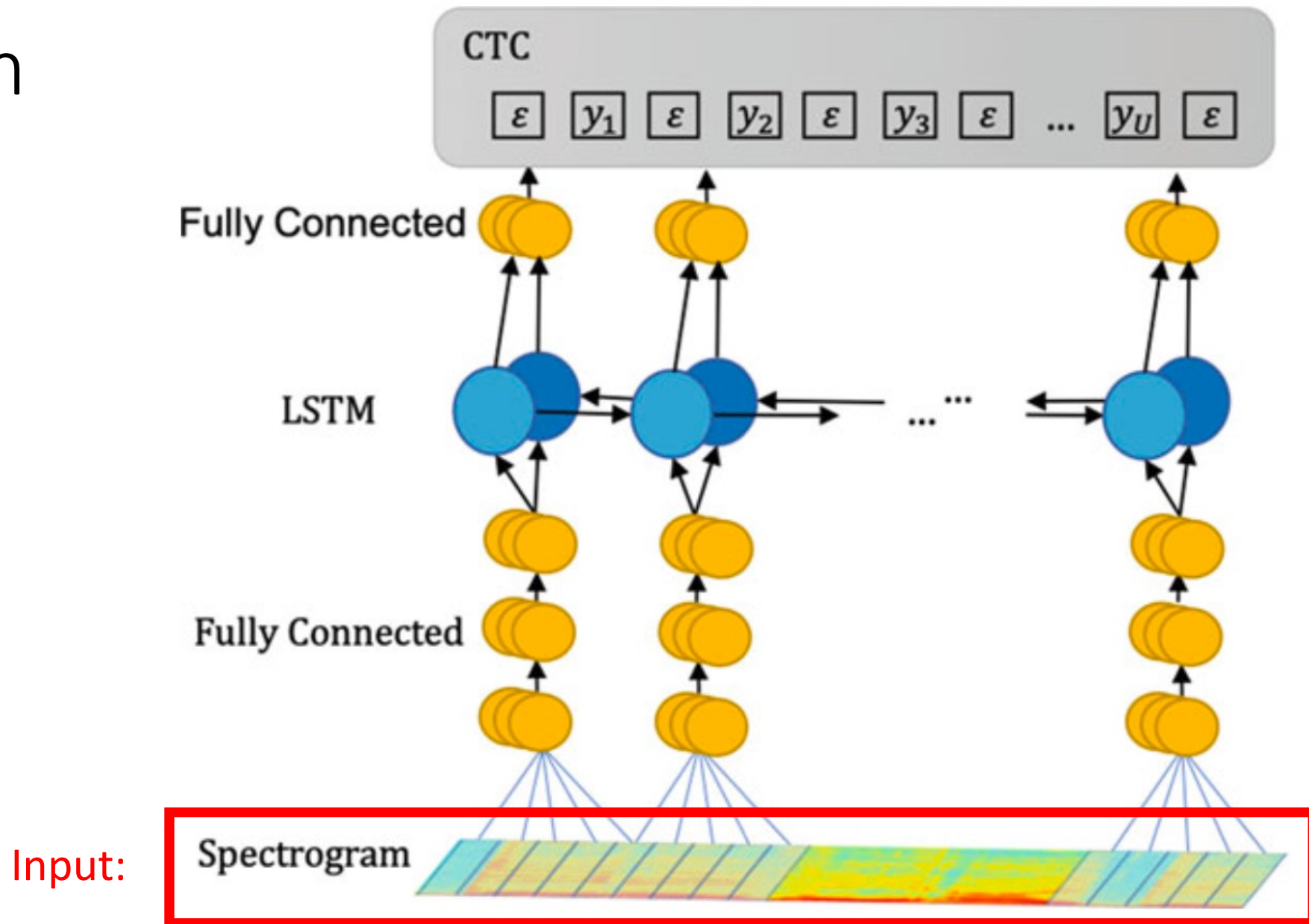
Encoder:

LSTM

Fully Connected

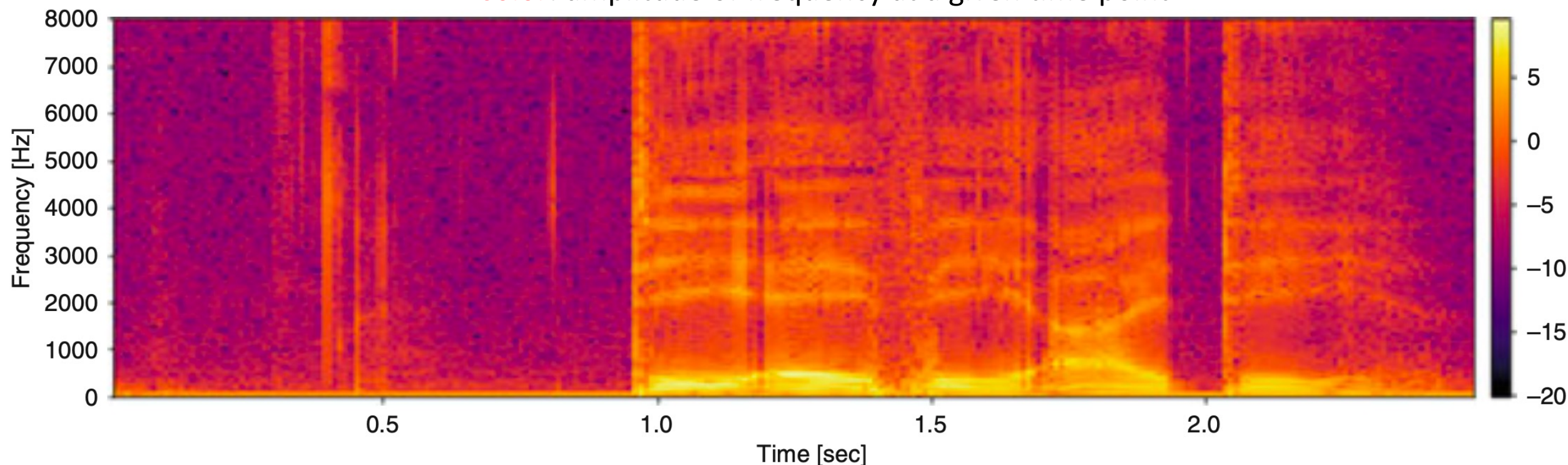
Spectrogram

DeepSpeech



Spectrogram: Visual Representation of Audio

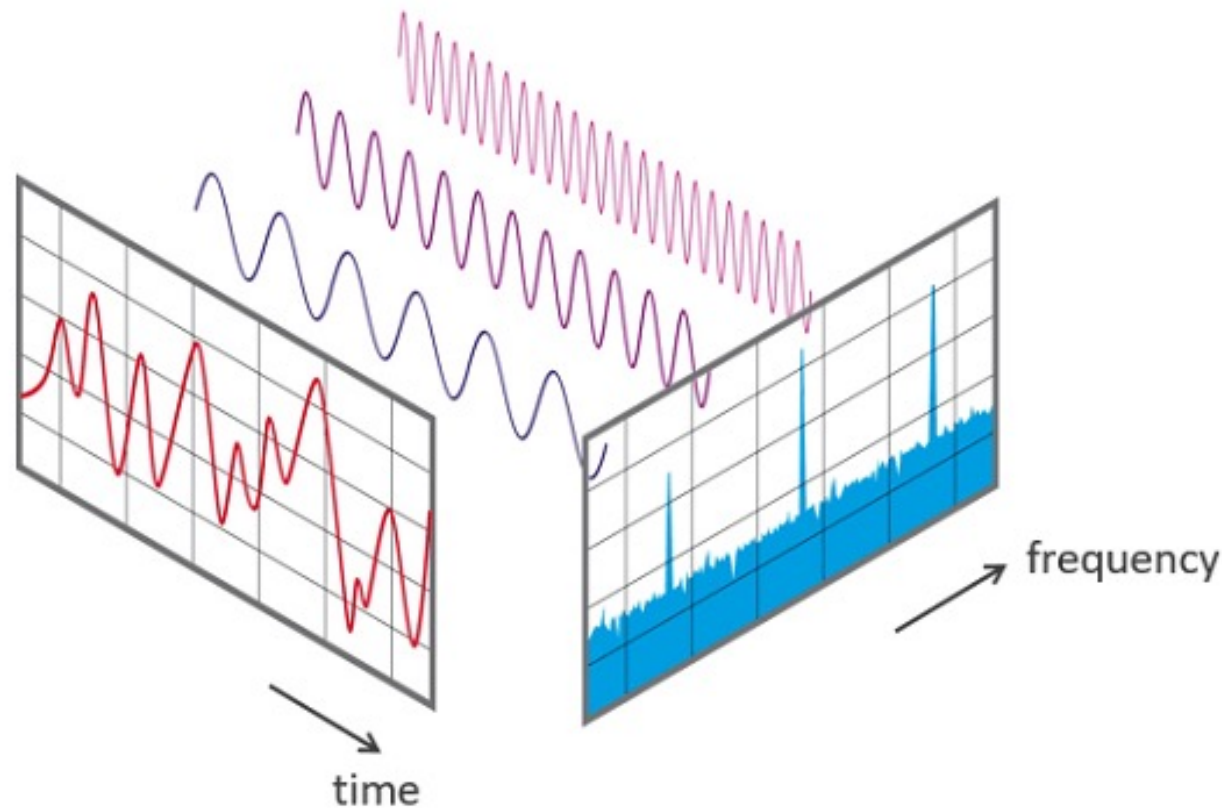
Color: amplitude of frequency at a given time point



Created by sliding a short window across the audio signal and applying a [Fourier transform](#) to each window

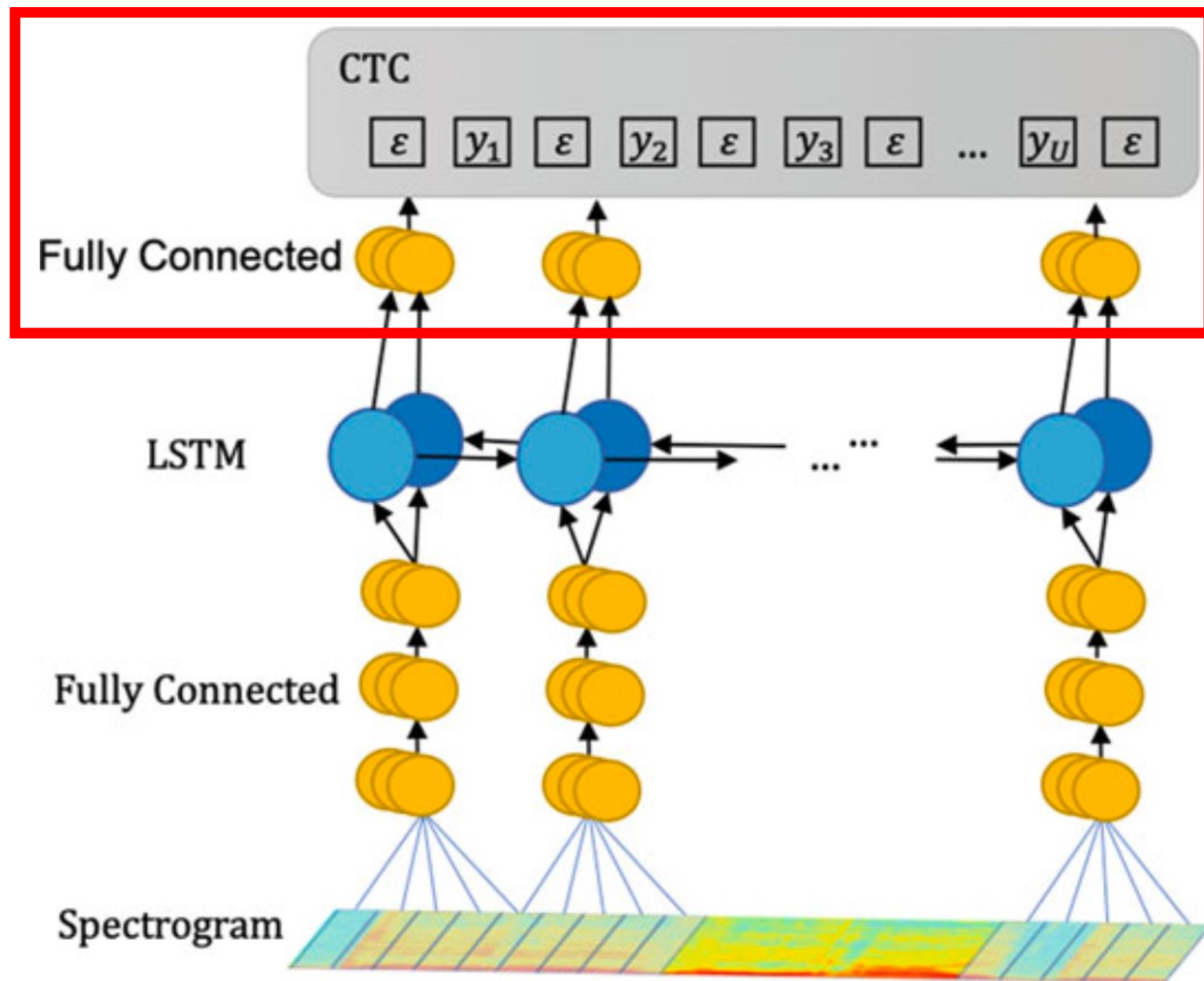
Background: Frequency Analysis of Audio Clip

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

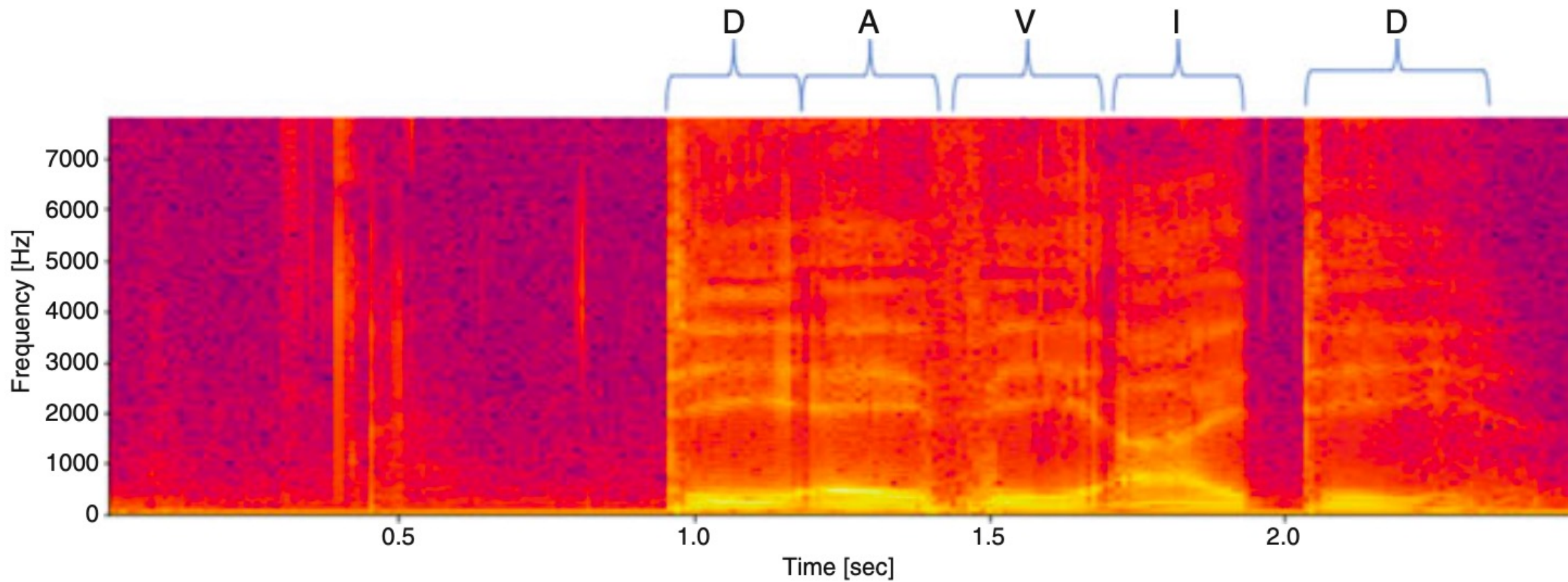


DeepSpeech

Output: character sequence predicted by a softmax layer



CTC: Input-Output Representation



CTC: Input-Output Representation

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

h	h	e	€	€				€			o
---	---	---	---	---	--	--	--	---	--	--	---

First, merge repeat characters.

h	e	€		€		o
---	---	---	--	---	--	---

Then, remove any € tokens.

h	e					o
---	---	--	--	--	--	---

The remaining characters are the output.

h	e			o
---	---	--	--	---

CTC: Input-Output Representation

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

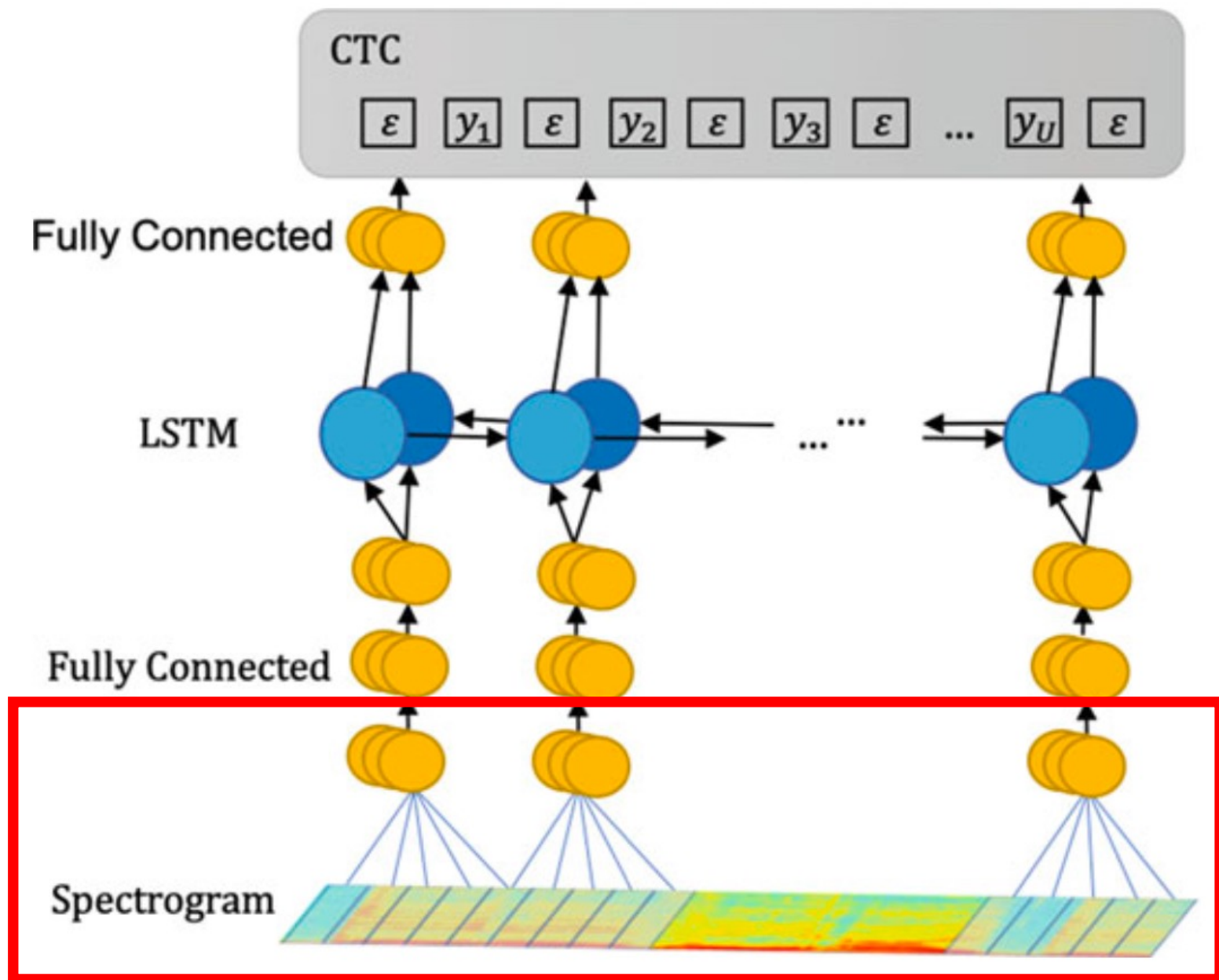
€	c	c	€	a	t
---	---	---	---	---	---

c	c	a	a	t	t
---	---	---	---	---	---

c	a	€	€	€	t
---	---	---	---	---	---

Supports recognizing the same word when spoken differently!

DeepSpeech



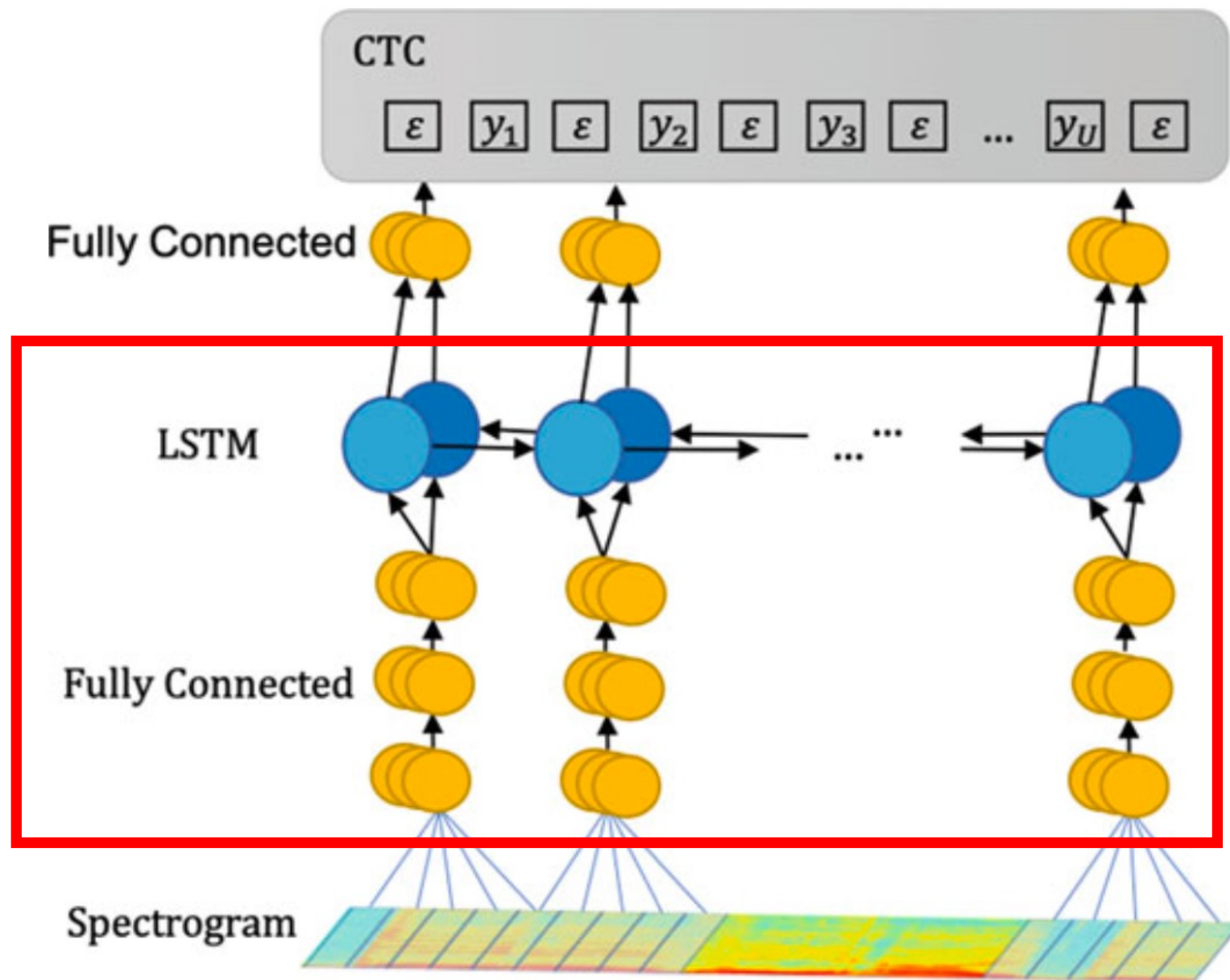
First hidden layer looks
at context around input:

DeepSpeech

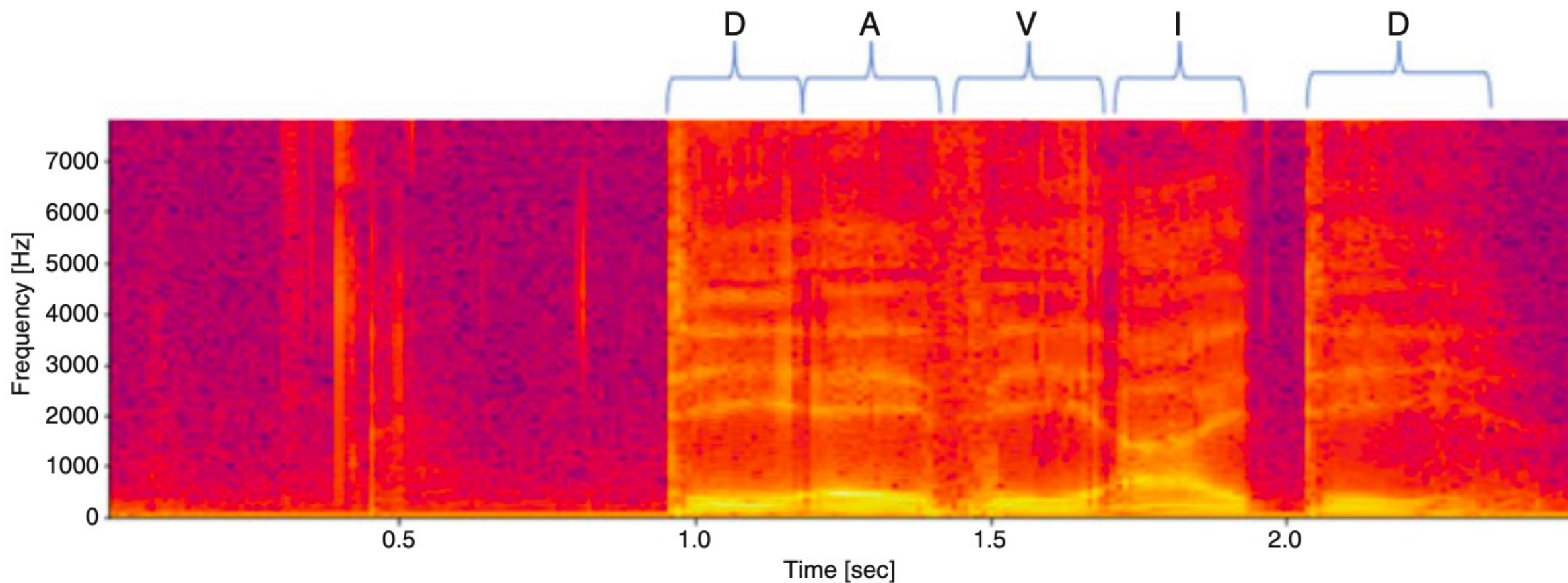
3 fully-connected layers
followed by bidirectional LSTM:

How is a bi-directional LSTM
beneficial?

How is a bi-directional LSTM
limiting?



DeepSpeech: Optimization Function (CTC)

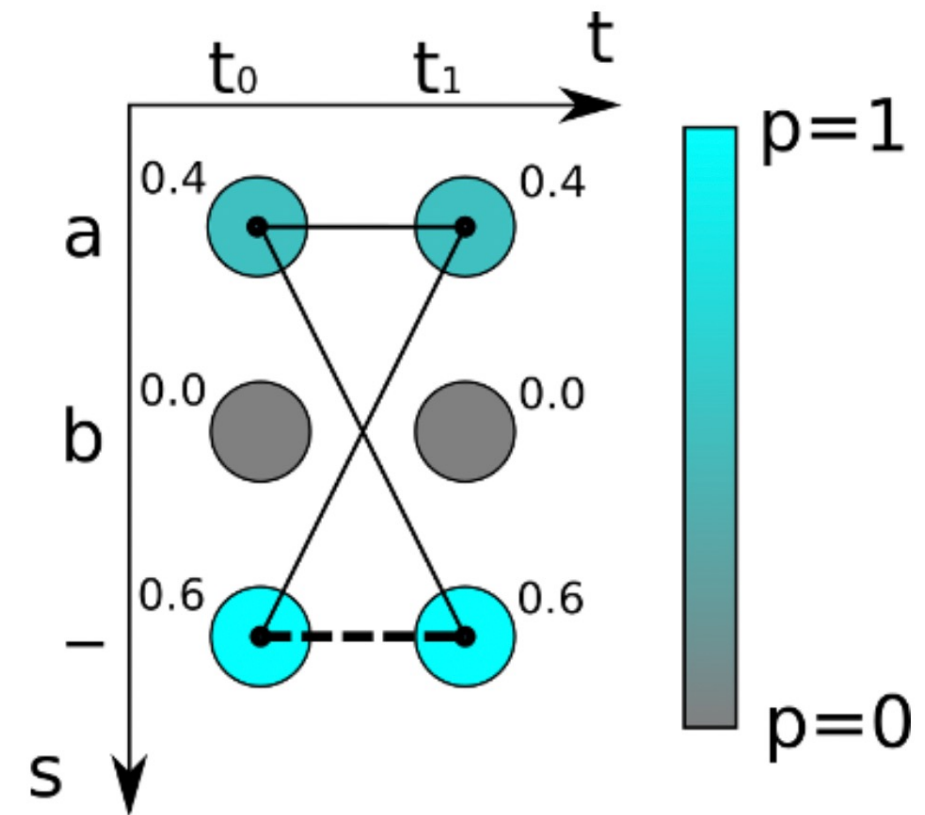


The CTC loss function enables learning output alignment without a per input label

DeepSpeech: Optimization Function (CTC)

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

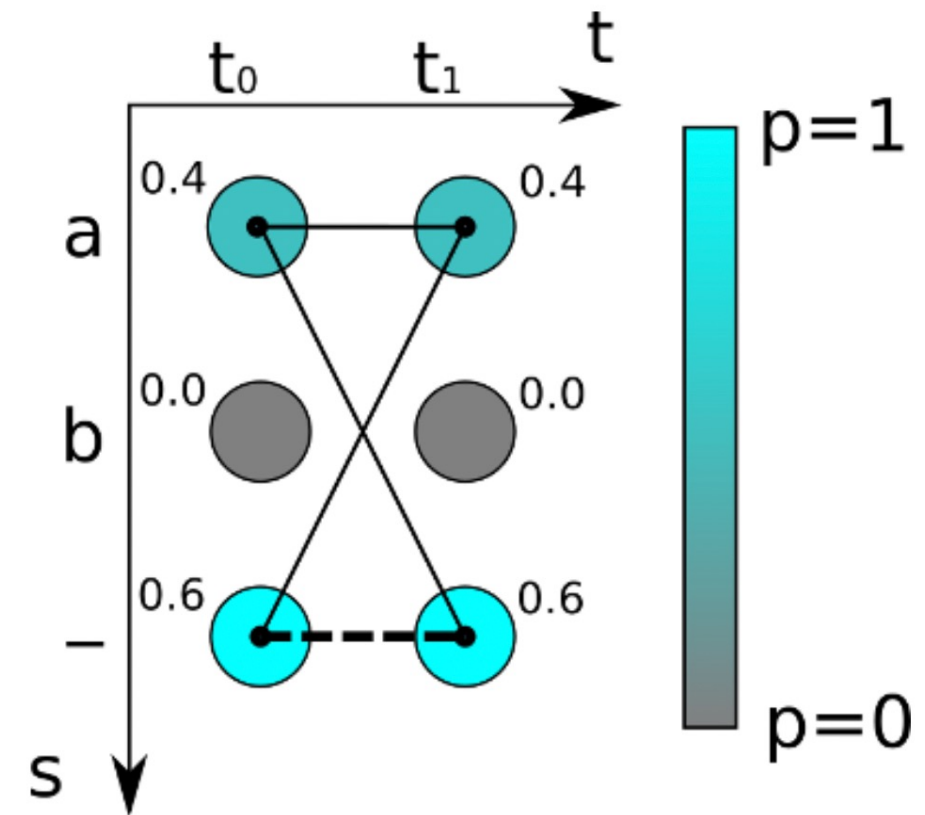
- Probability of "a" is sum of all "a" representations
 - Probability of "aa"?
 - $0.4 \times 0.4 = 0.16$
 - Probability of "a-"?
 - $0.4 \times 0.6 = 0.24$
 - Probability of "-a"?
 - $0.6 \times 0.4 = 0.24$
 - Sum: $0.16 + 0.24 + 0.24 = 0.64$



DeepSpeech: Optimization Function (CTC)

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

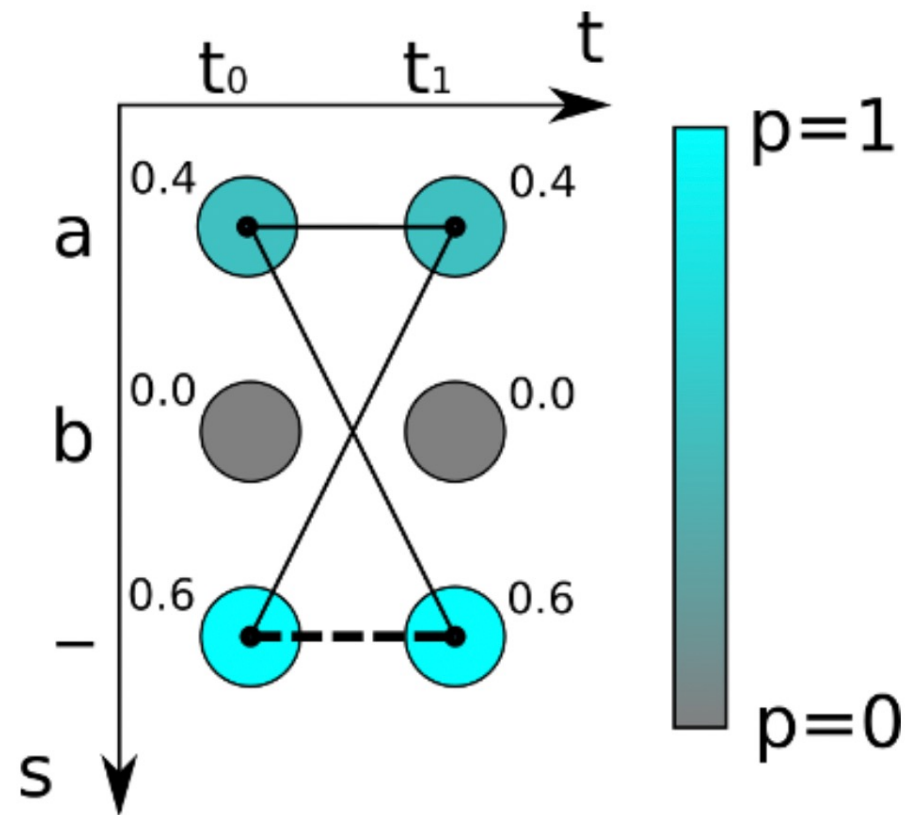
- Probability of "a": 0.64
- Probability of "" is sum of all "" representations
 - Probability of "--"?
 - $0.6 \times 0.6 = 0.36$



DeepSpeech: Optimization Function (CTC)

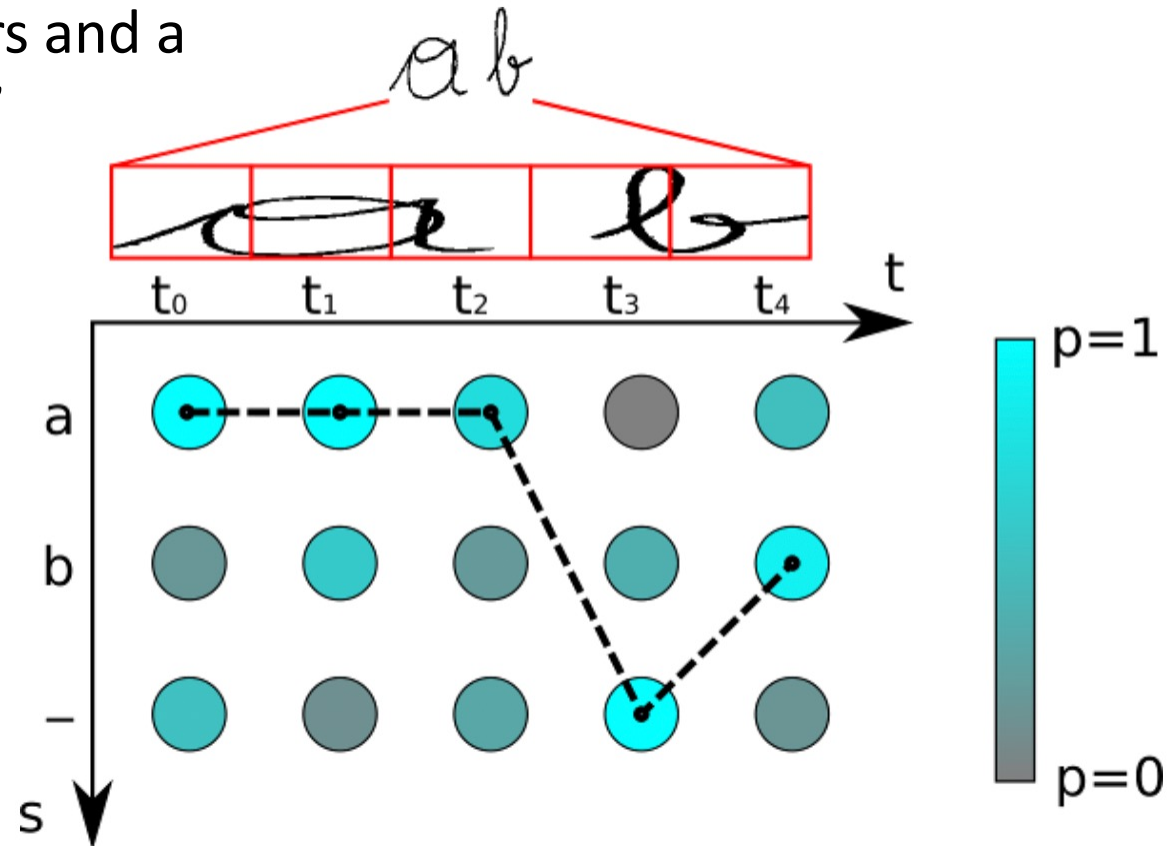
Most plausible from all possible alignments learned;
e.g., 2 time steps with 2 potential characters and a
blank token ("-")

- Probability of "a": 0.64
- Probability of "": 0.36
- And so on for all possible alignments...

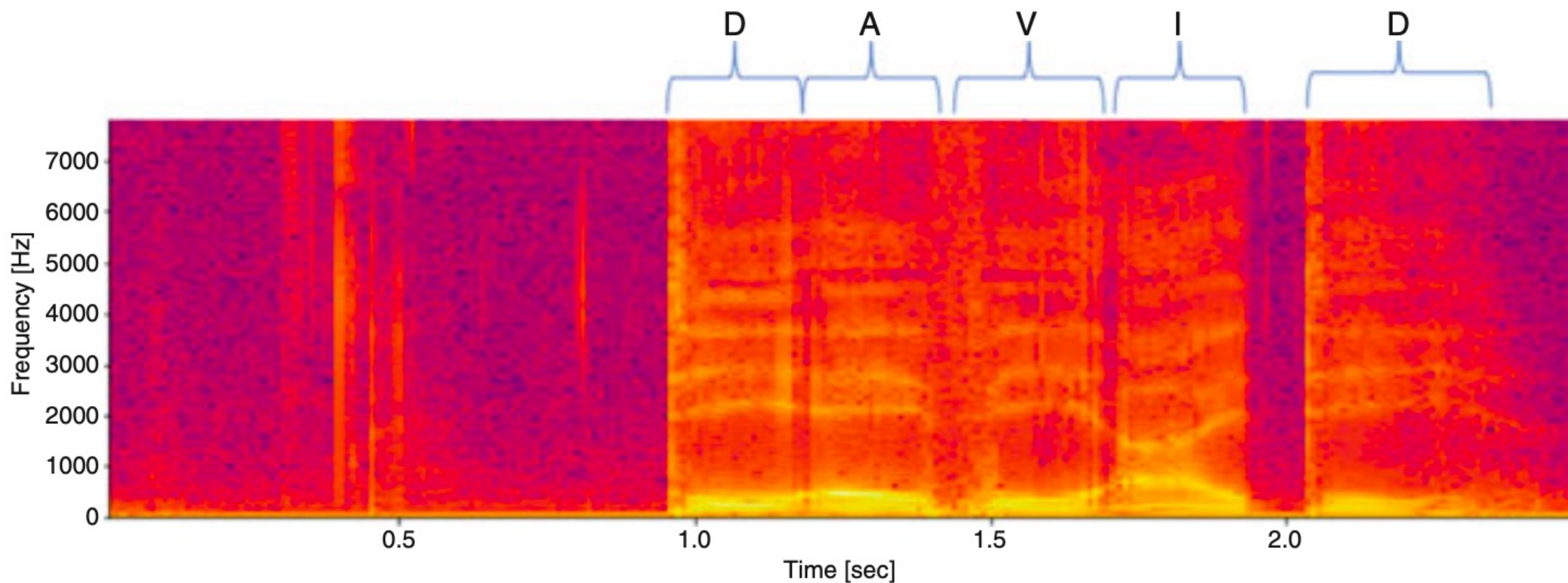


DeepSpeech: Optimization Function (CTC)

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



DeepSpeech: Optimization Function (CTC)



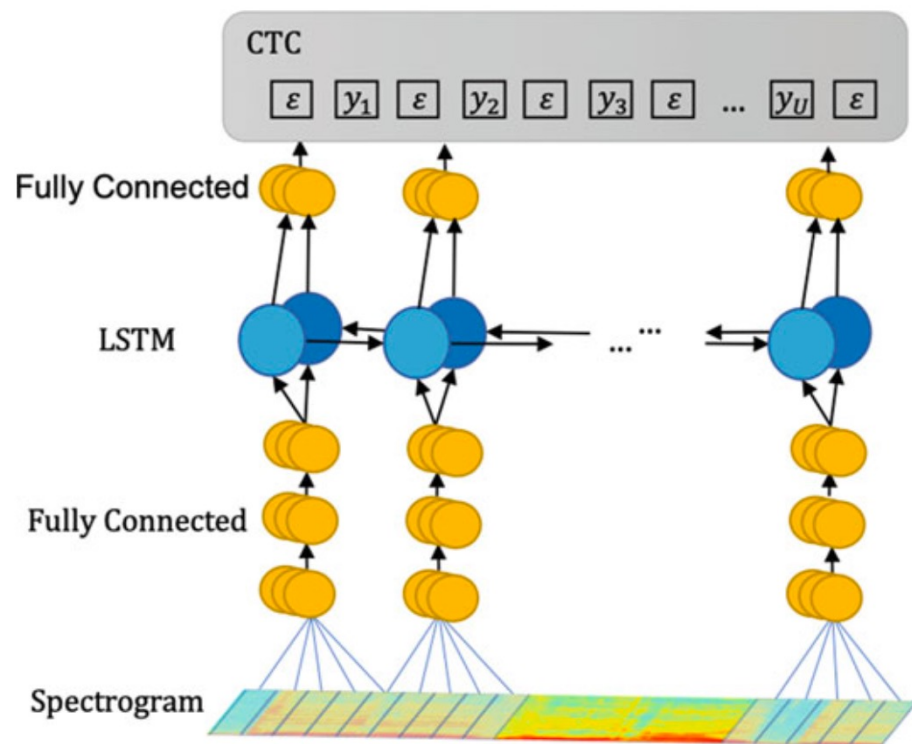
CTC uses dynamic programming to accelerate computation and is differentiable

DeepSpeech: Training (Key Ideas)

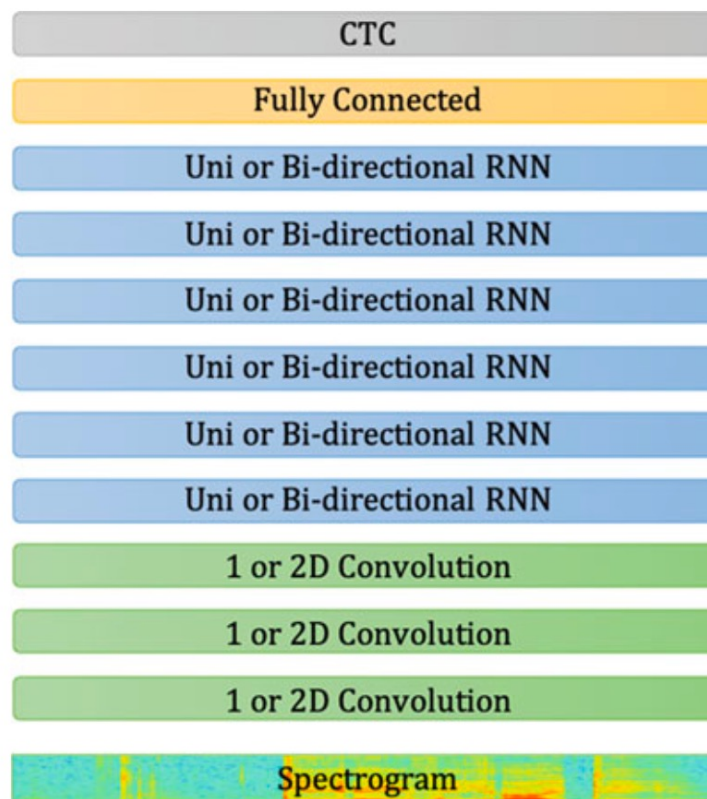
- 5000 hours from 9600 speakers
- Regularization
 - Dropout
 - Data augmentation: audio file translated 5 ms forward and backward
- Results boosted by incorporating a language model

Popular Methods

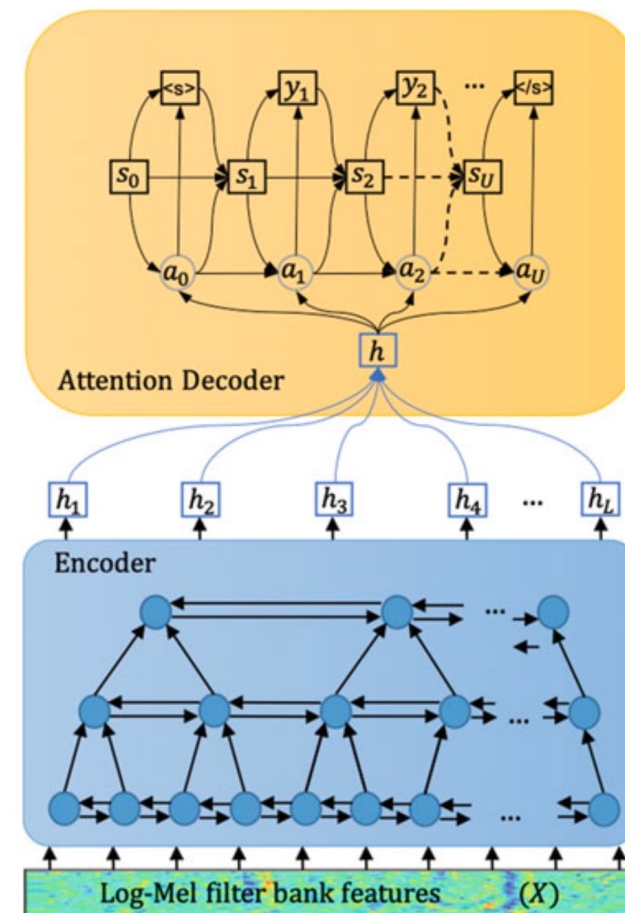
DeepSpeech



DeepSpeech2



Listen, Attend, and Spell

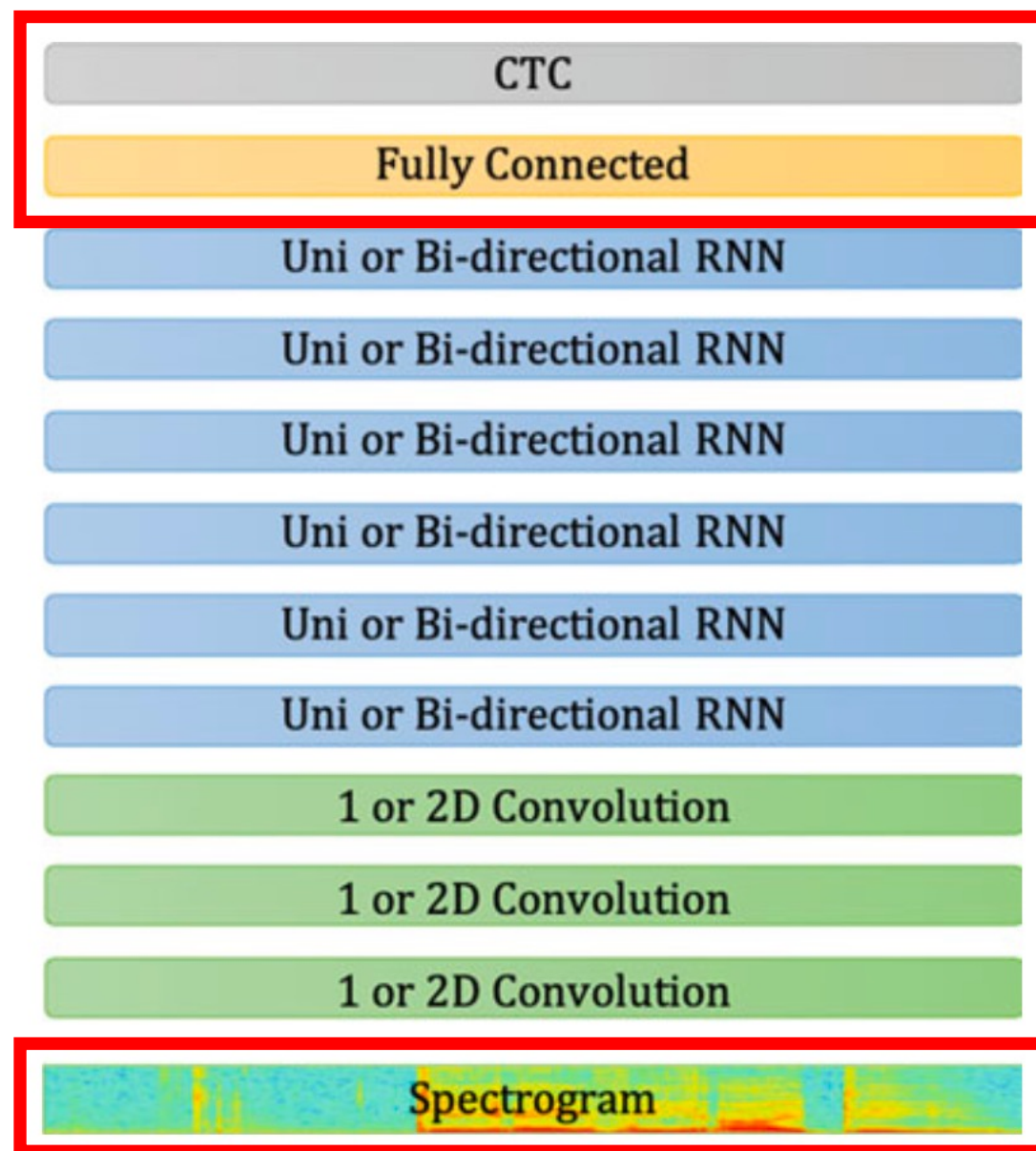


DeepSpeech2

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

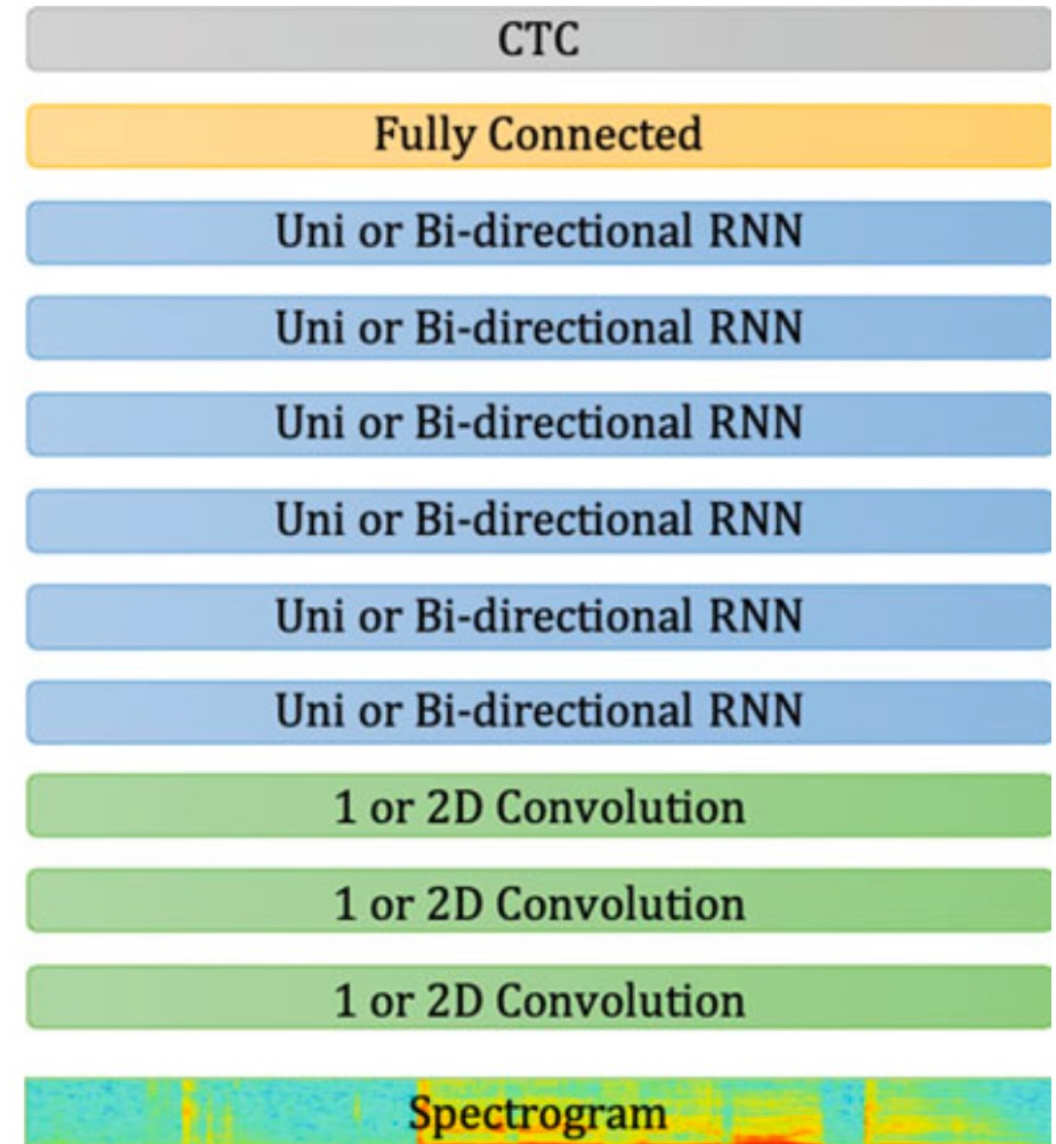
Same input:



DeepSpeech2

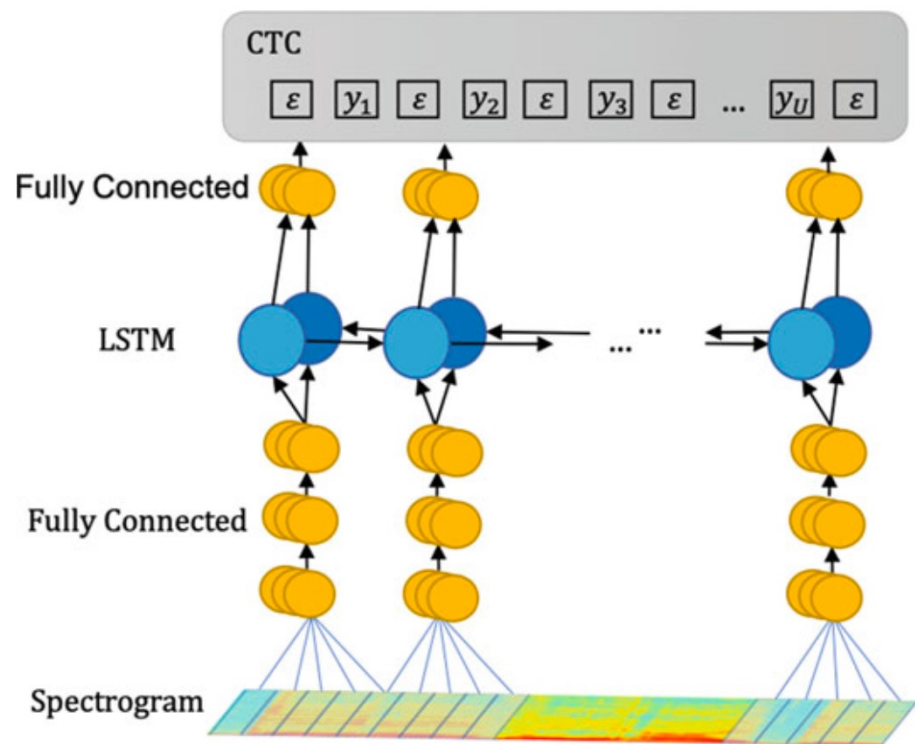
Training protocol difference from DeepSpeech:

- 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)

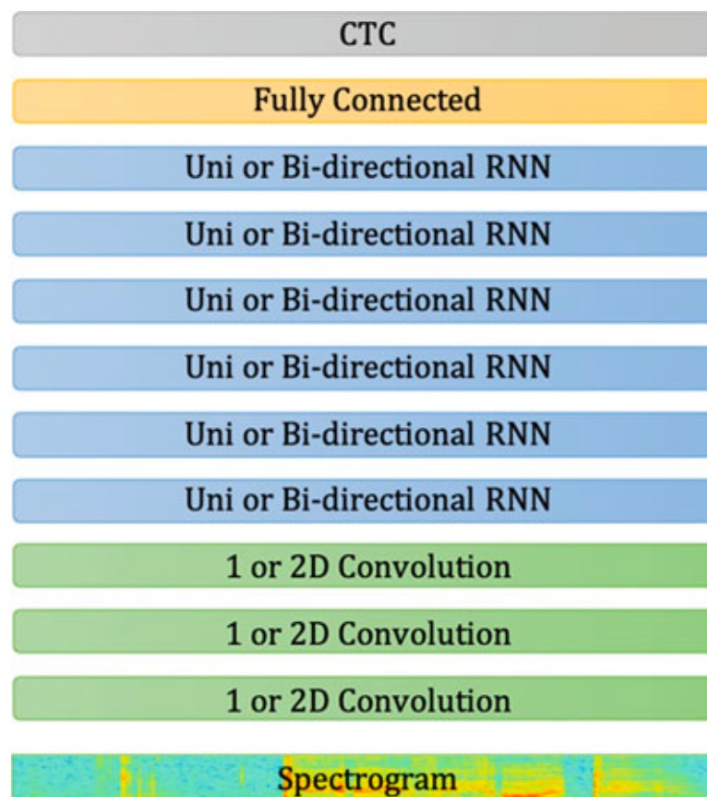


Popular Methods

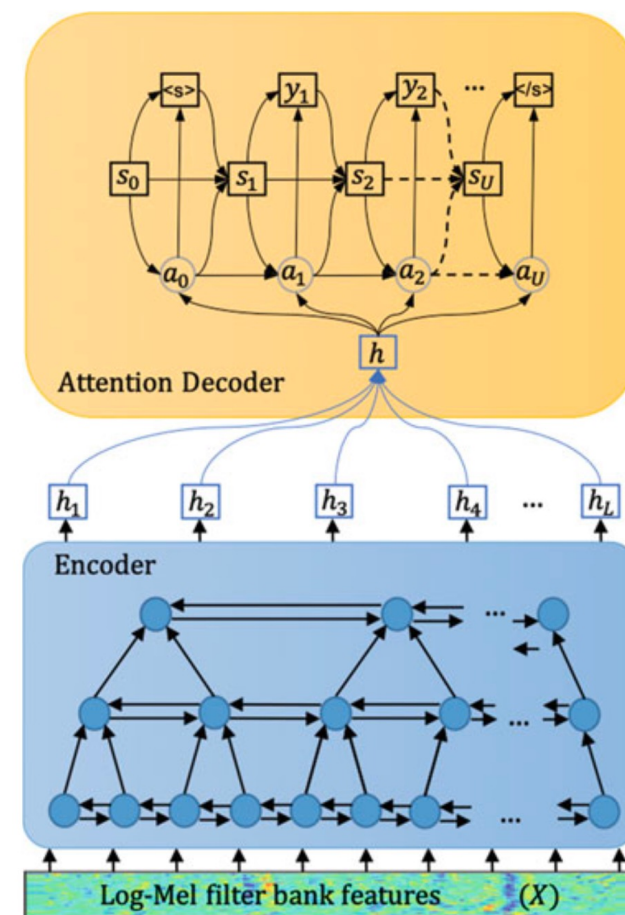
DeepSpeech



DeepSpeech2



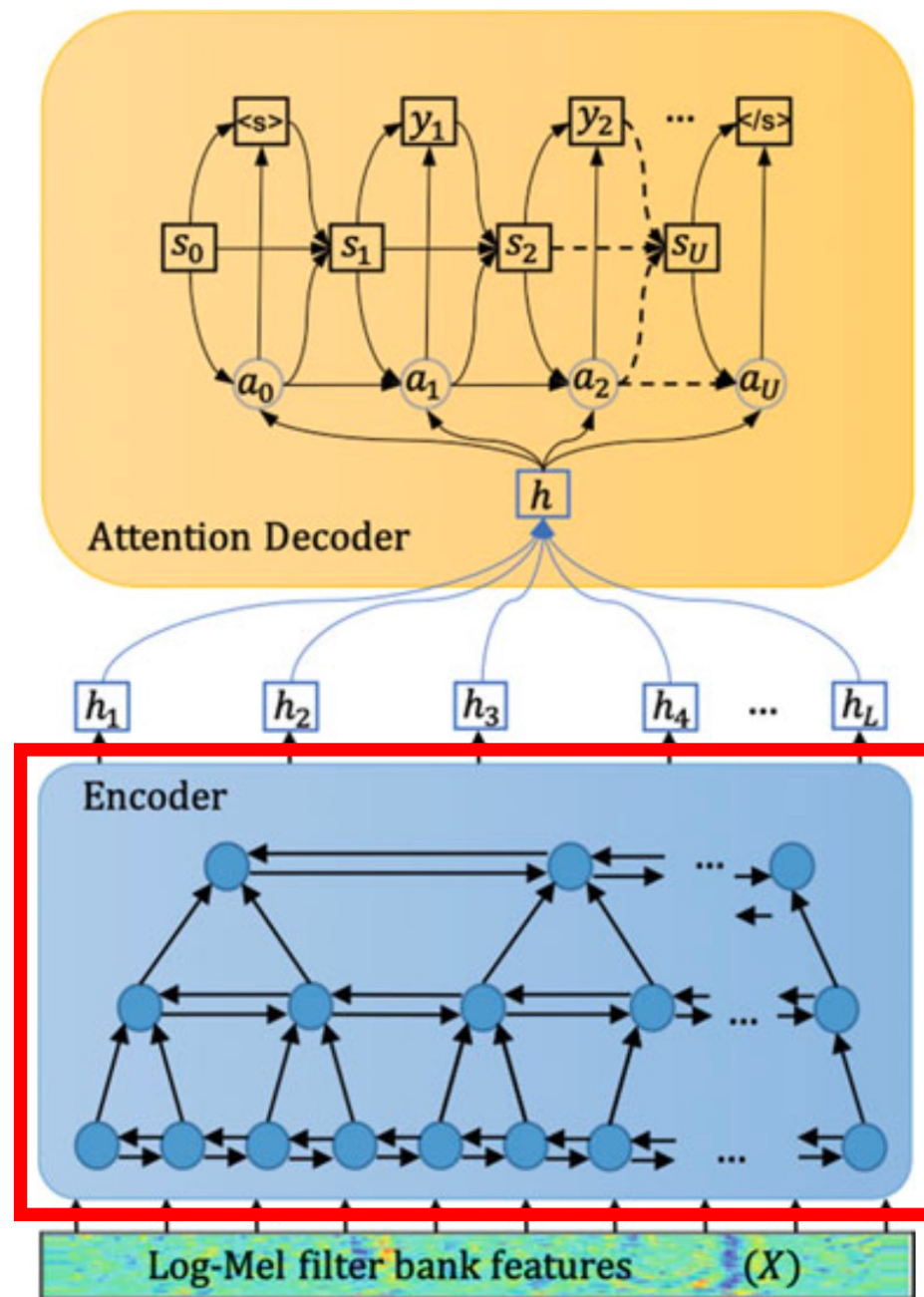
Listen, Attend, and Spell



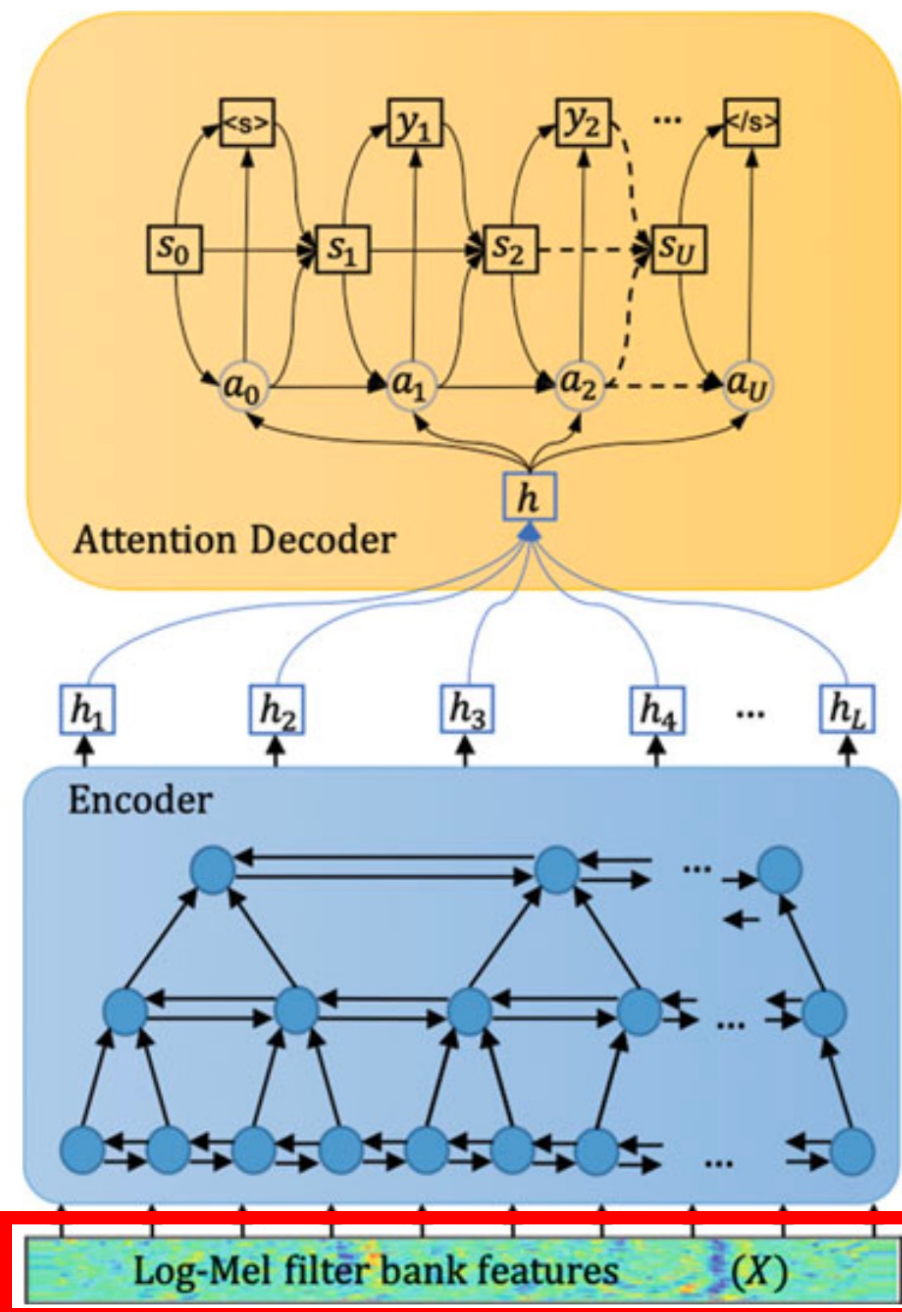
Listen, Attend, and Spell

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



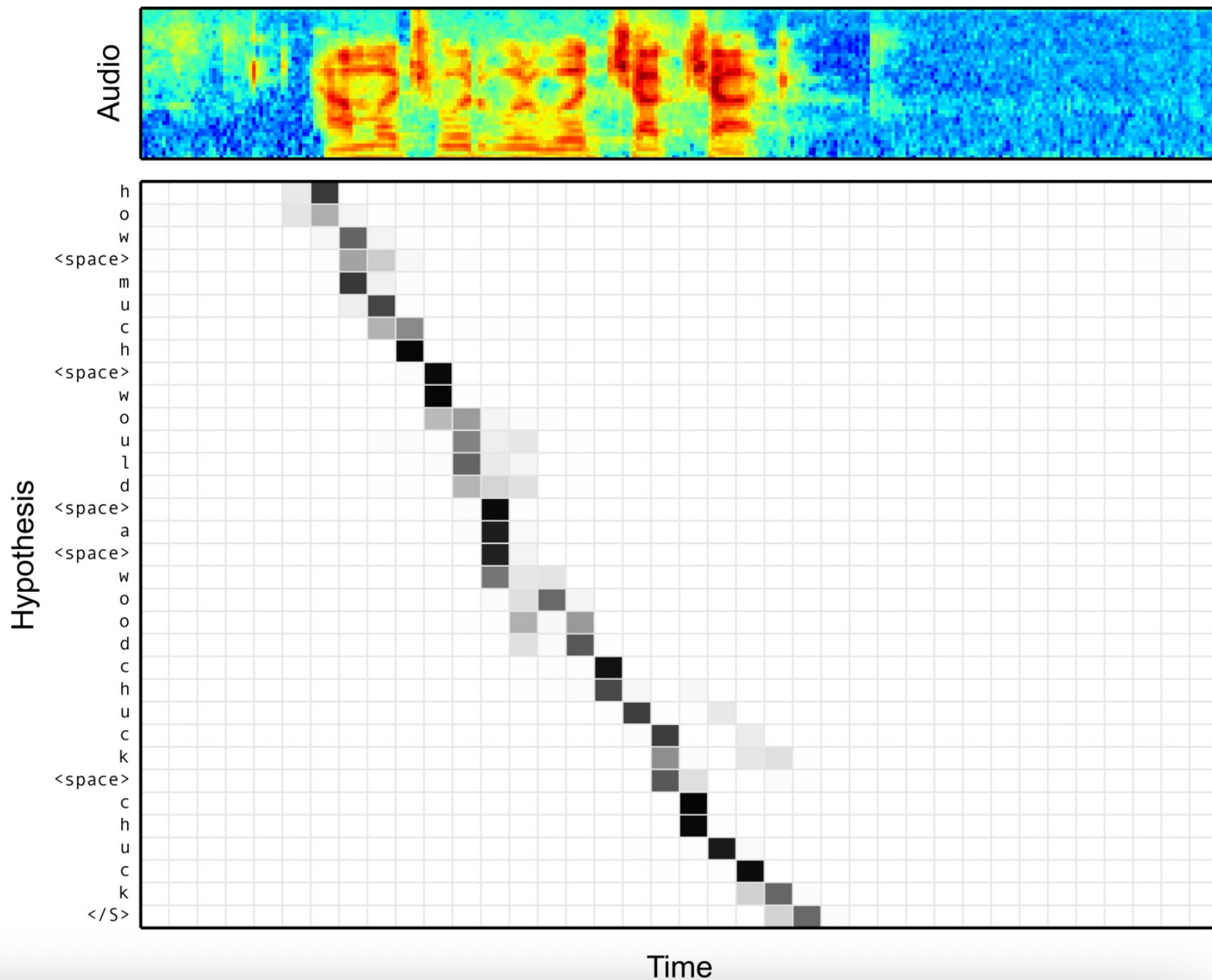
Listen, Attend, and Spell



Input: more sophisticated hand-crafted
audio representation then spectrogram

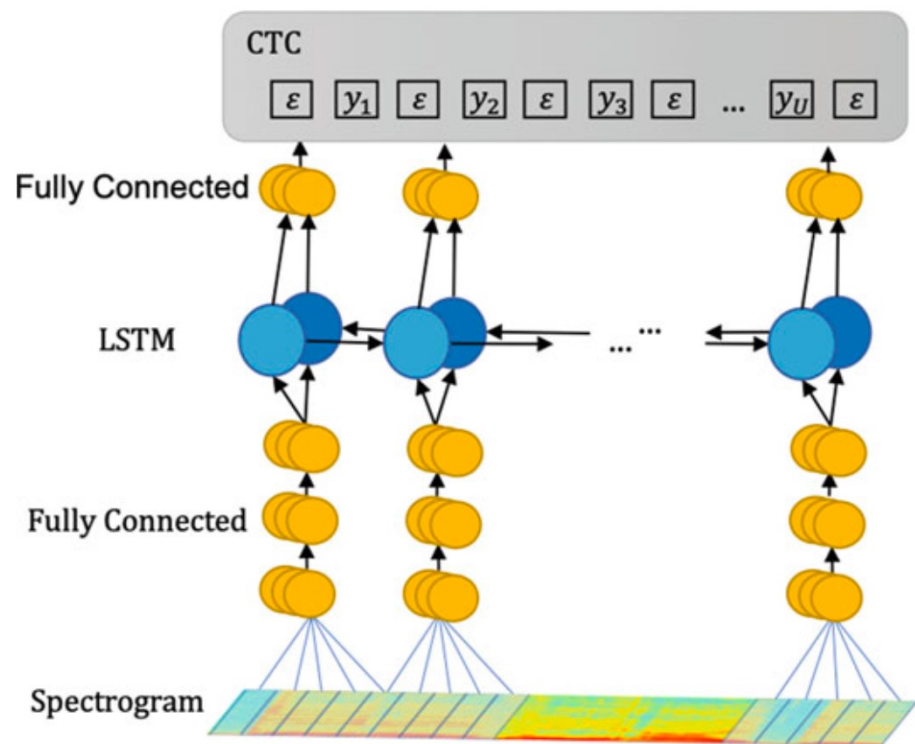
Result

Attention enables
visualizing alignment
between audio signal
and characters

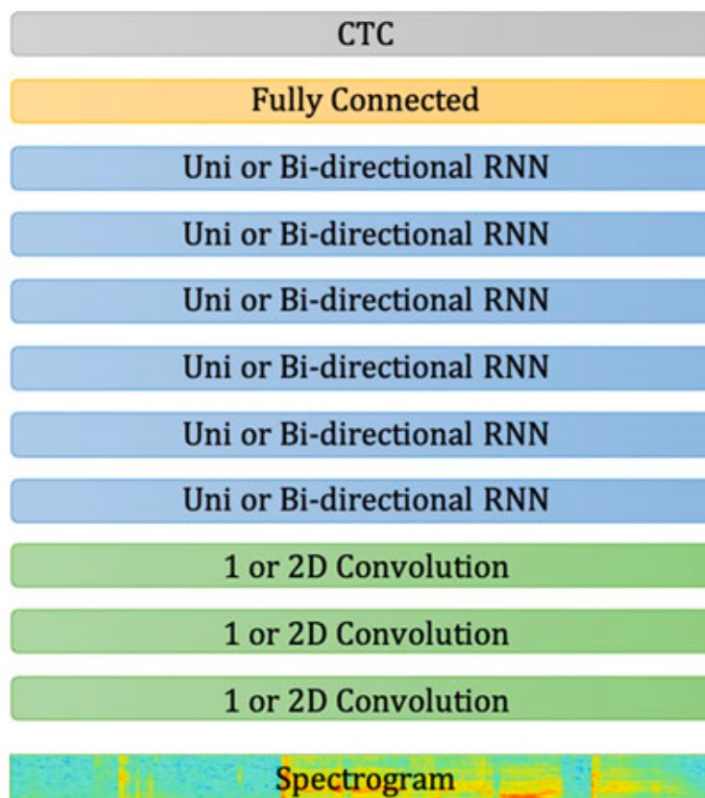


Popular Methods

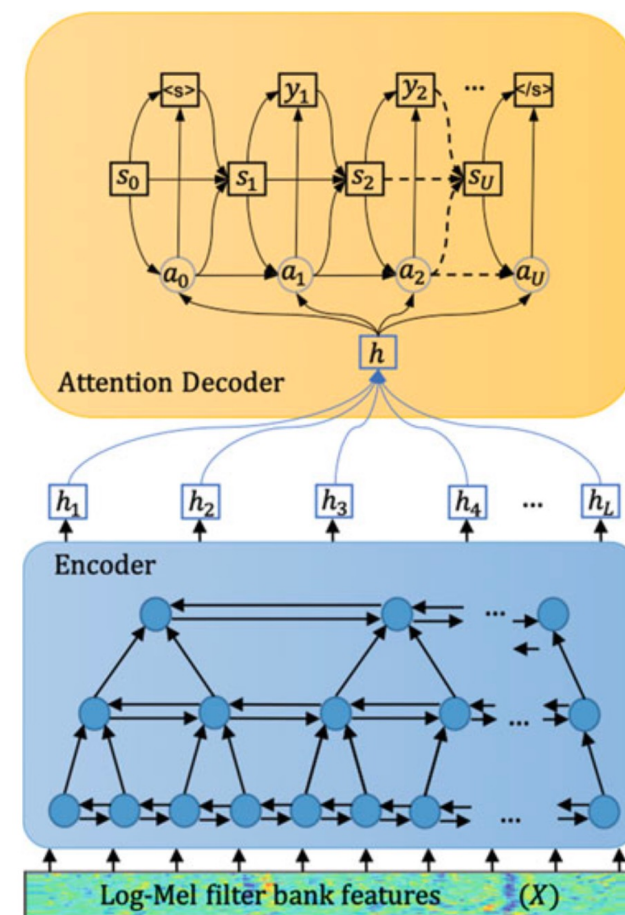
DeepSpeech



DeepSpeech2



Listen, Attend, and Spell

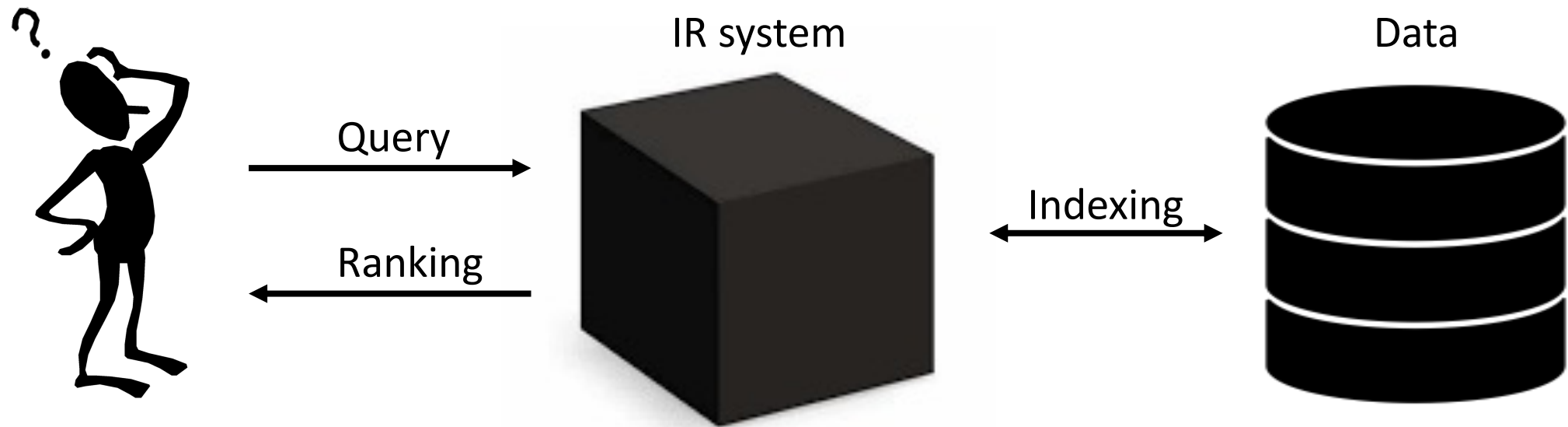


Today's Topics

- 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


Baidu 新闻 machine learning **百度一下**

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

找到相关新闻92篇 ☒ 新闻全文 ☐ 新闻标题 | 按焦点排序 ▾

[...GWAS summary statistics for data mining and machine learning](#)
中国矿业大学 2017年12月26日 16:58
报告题目:Using GWAS summary statistics for data mining and machine learning 时间:12月29日上午9:00 地点:文昌校区逸夫楼邵206 主办单位:中国矿业大学信息与... [百度快照](#)

[\[Machine Learning\] 深度学习中消失的梯度](#)
深圳热线 2017年12月07日 09:46
原标题:[Machine Learning] 深度学习中消失的梯度 好久没有更新blog了,最近抽时间看了Nielsen的《Neural Net [百度快照](#)




[Machine learning can help enhance drug trials: study](#)
新华网 2017年11月16日 01:40
15 (Xinhua) -- A team of researchers have demonstrated that machine learning could be an effective tool in determining whether a new drug works in the... [百度快照](#)

Google machine learning

All News Videos **Images** Books More

artificial intelligence big data iot distributed robotic cyber s



Machine learning ⊆ artificial intelligence


ARTIFICIAL INTELLIGENCE
Design an intelligent agent that perceives its environment and makes decisions to maximize chances of achieving its goal. Subfields: vision, robotics, machine learning, natural language processing, planning,...


MACHINE LEARNING
Often "computers the ability to learn without being explicitly programmed" (Arthur Samuel, 1959)


SUPERVISED LEARNING
Classification, regression

UNSUPERVISED LEARNING
Clustering, dimensionality reduction, recommendation

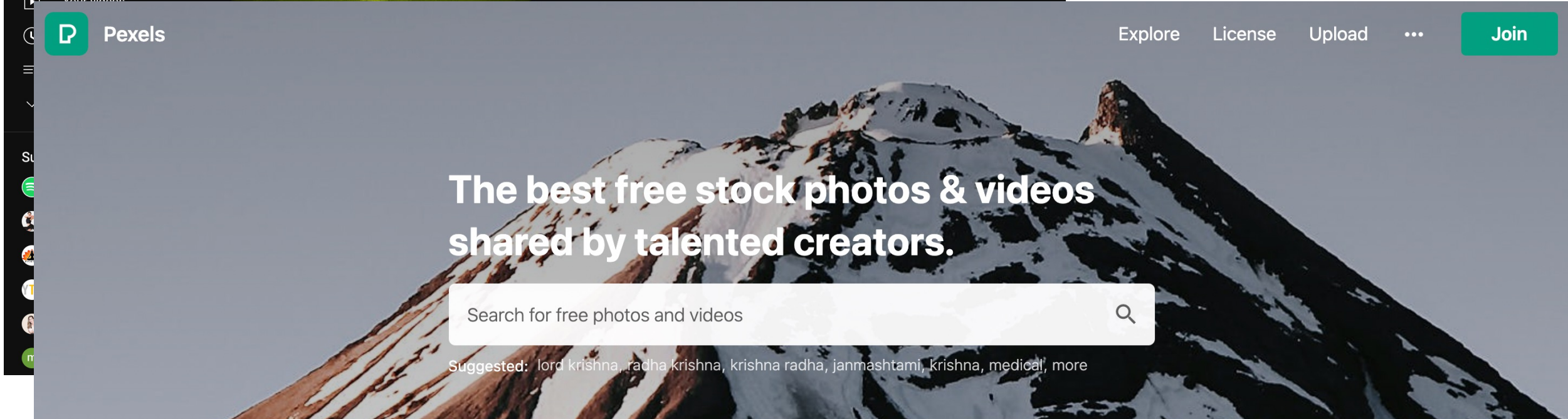
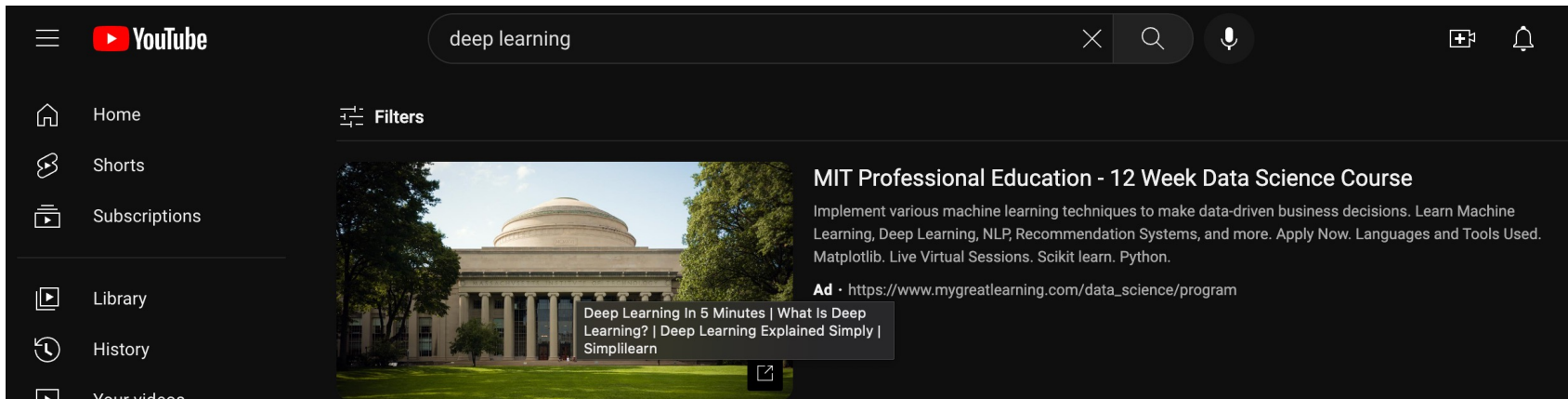
REINFORCEMENT LEARNING
Reward maximization



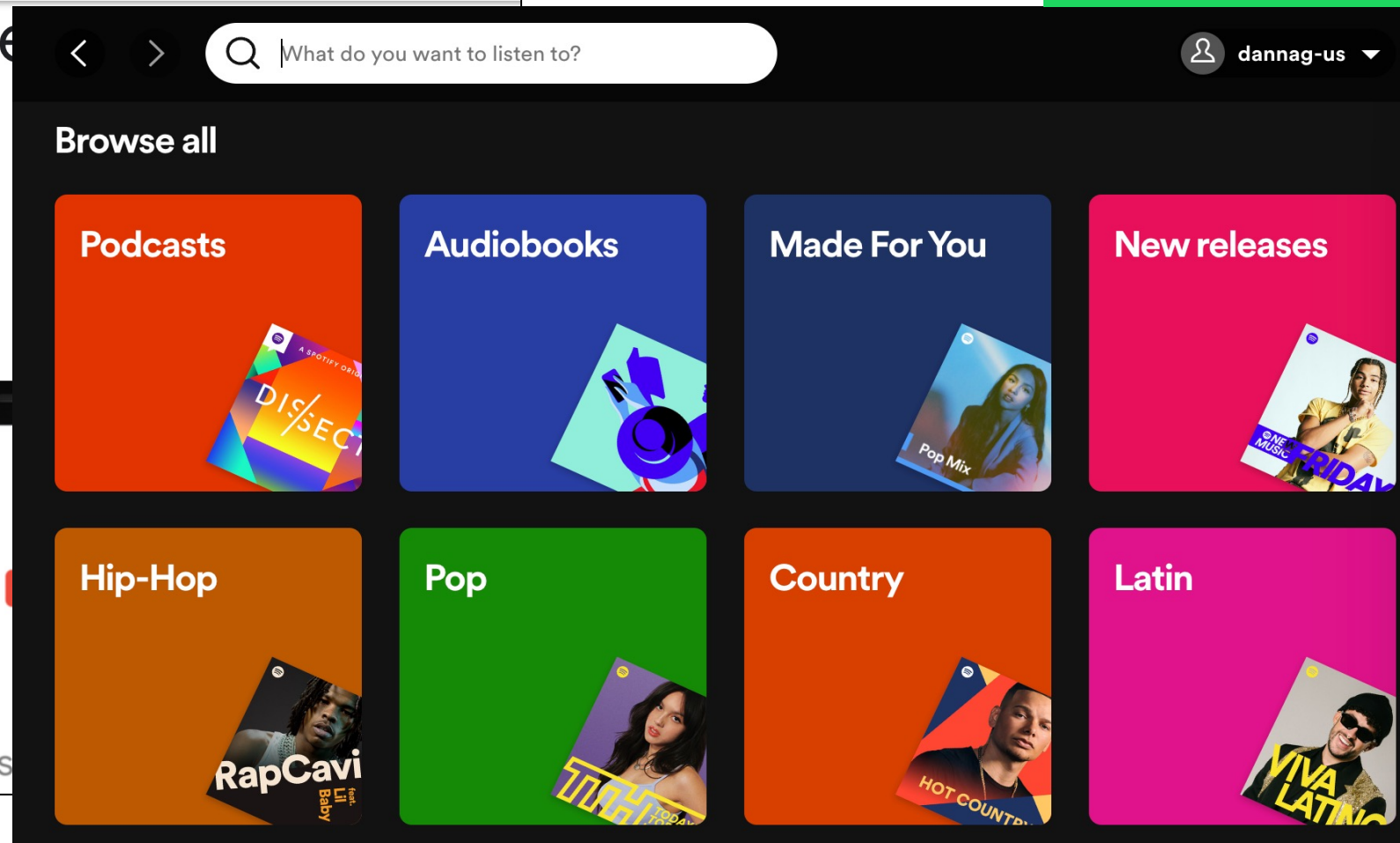
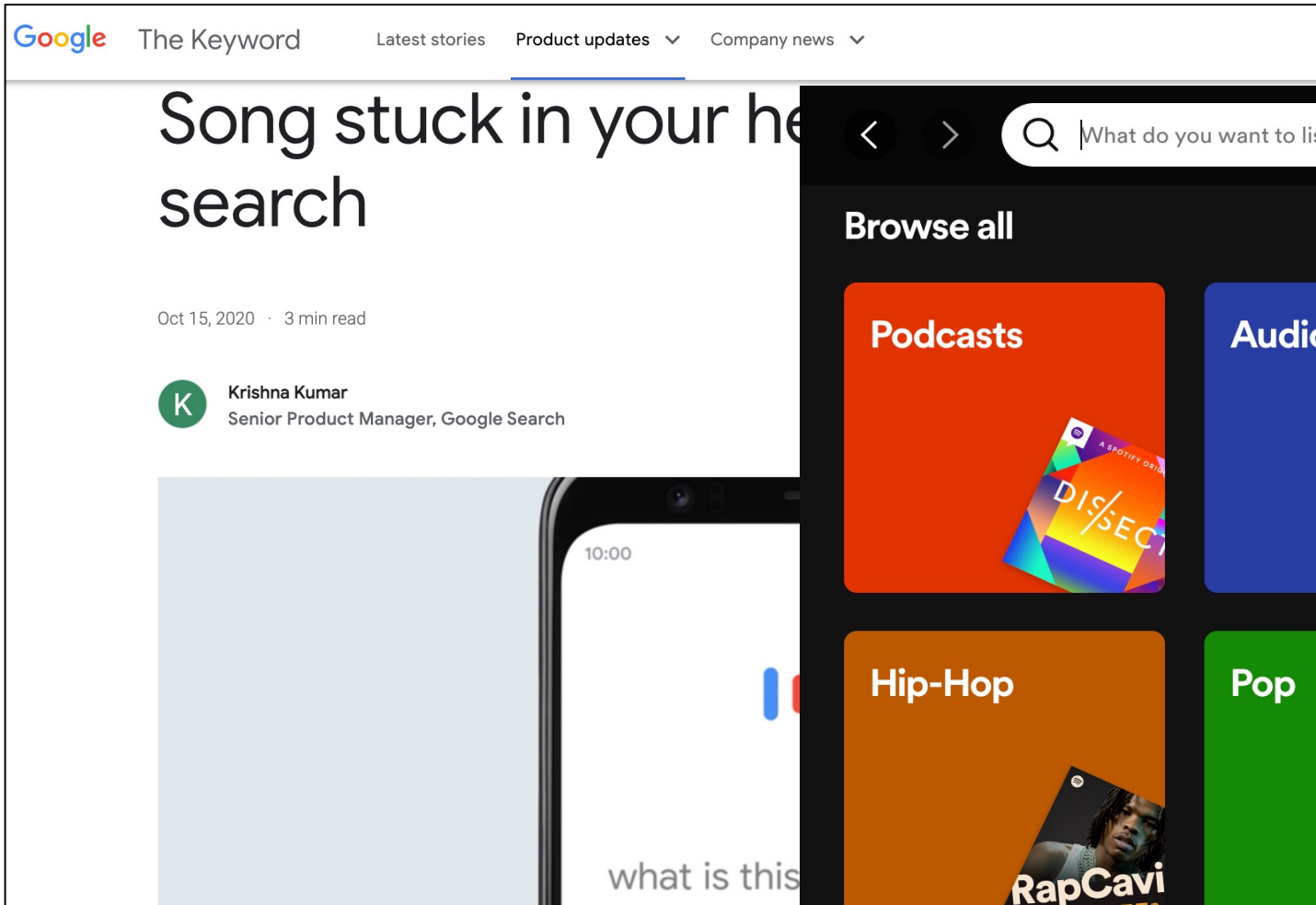




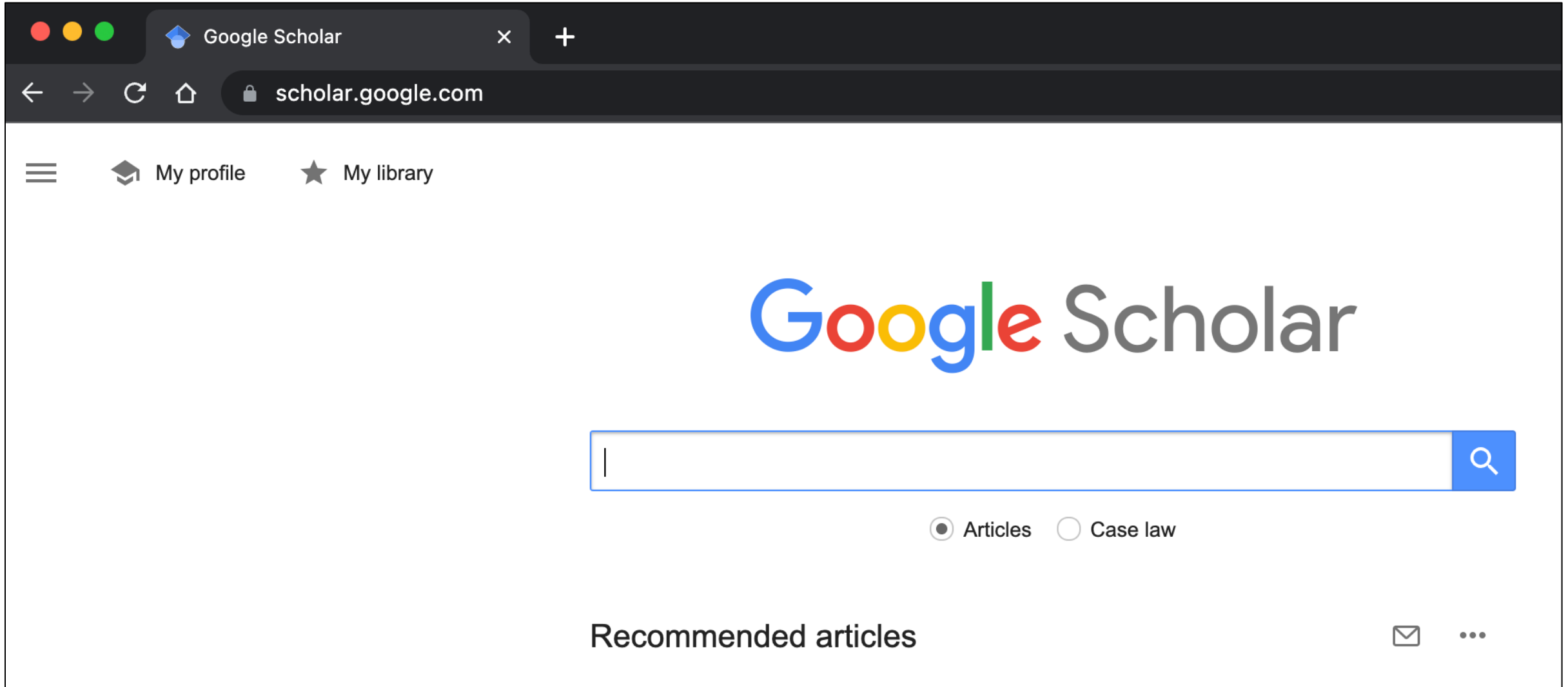
Visual Search



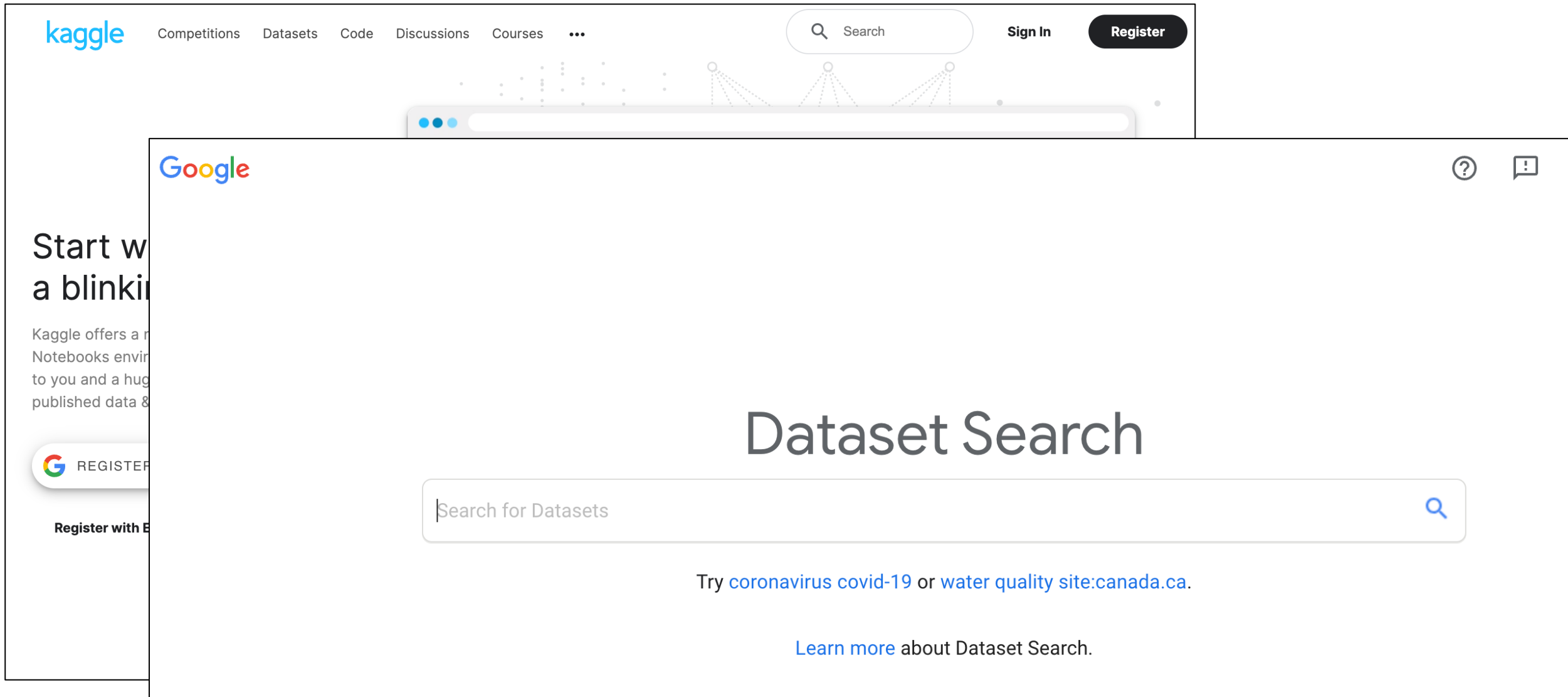
Song Search



Academic and Legal Search



AI Challenge/Dataset Search



The image shows a screenshot of the Kaggle website's Dataset Search page. The top navigation bar includes the Kaggle logo, links for Competitions, Datasets, Code, Discussions, and Courses, a search bar, and buttons for Sign In and Register. The main content area features the Google logo in the top left, a large 'Dataset Search' heading, a search input field with the placeholder text 'Search for Datasets', and a search button. Below the search field, there is a suggestion: 'Try coronavirus covid-19 or water quality site:canada.ca.' and a link to 'Learn more about Dataset Search.' On the left side, there is a sidebar with the text 'Start with a blink' and 'Kaggle offers a r... Notebooks envir... to you and a hug... published data &...' along with a 'REGISTER' button and the text 'Register with B...'. The background of the main content area has a faint illustration of a neural network.

Kaggle

Competitions Datasets Code Discussions Courses ...

Search

Sign In Register

Google

Start with a blink

Kaggle offers a r... Notebooks envir... to you and a hug... published data &...

REGISTER

Register with B...

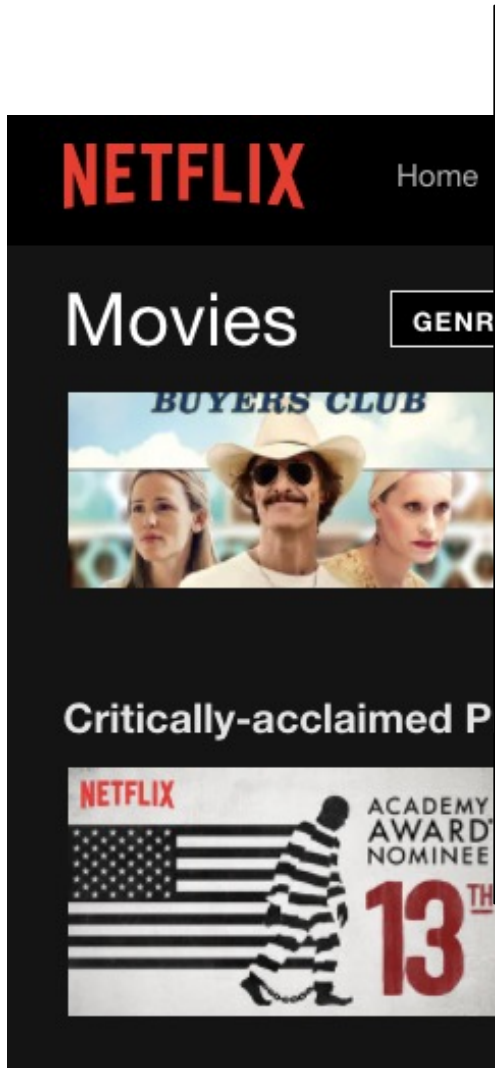
Dataset Search

Search for Datasets

Try [coronavirus covid-19](#) or [water quality site:canada.ca.](#)

[Learn more](#) about Dataset Search.

Entertainment and Shopping Search



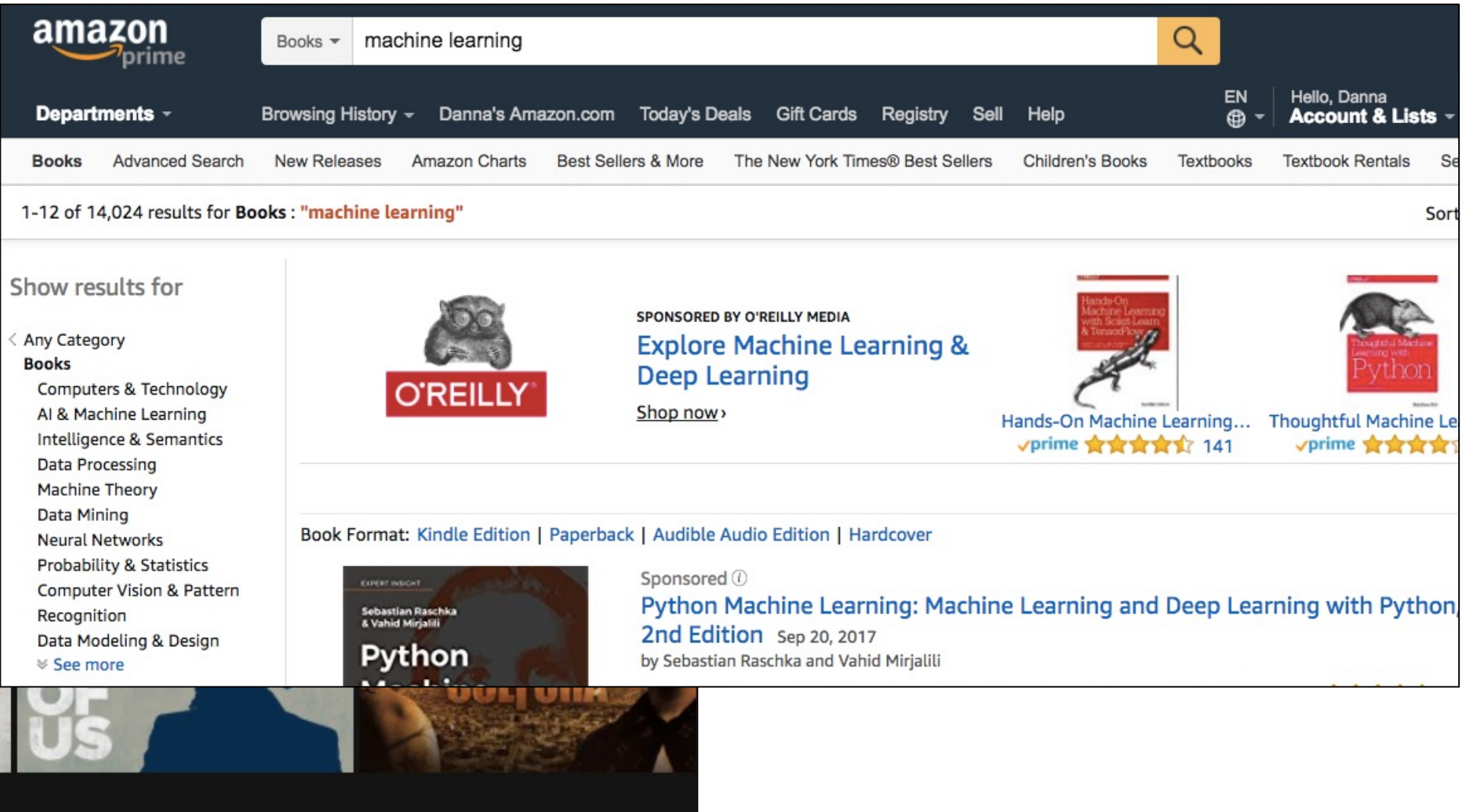
NETFLIX Home

Movies GENRE

BUYERS CLUB

Critically-acclaimed P

NETFLIX ACADEMY AWARD NOMINEE **13TH**



amazon prime Books machine learning

Departments Browsing History Danna's Amazon.com Today's Deals Gift Cards Registry Sell Help

Books Advanced Search New Releases Amazon Charts Best Sellers & More The New York Times® Best Sellers Children's Books Textbooks Textbook Rentals

1-12 of 14,024 results for Books : "machine learning" Sort

Show results for

< Any Category

Books

- Computers & Technology
- AI & Machine Learning
- Intelligence & Semantics
- Data Processing
- Machine Theory
- Data Mining
- Neural Networks
- Probability & Statistics
- Computer Vision & Pattern Recognition
- Data Modeling & Design
- See more

O'REILLY

SPONSORED BY O'REILLY MEDIA

Explore Machine Learning & Deep Learning

[Shop now](#)

Hands-On Machine Learning... Thoughtful Machine Learning with Python

Book Format: [Kindle Edition](#) | [Paperback](#) | [Audible Audio Edition](#) | [Hardcover](#)



Python Machine Learning


Sponsored

Python Machine Learning: Machine Learning and Deep Learning with Python, 2nd Edition Sep 20, 2017

by Sebastian Raschka and Vahid Mirjalili

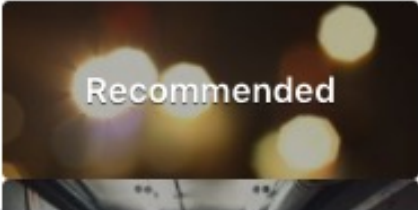
Social Media Search





 Danna


Groups


Discover

 Recommended


 Local

 School and Education

 Games


 More

Recommended




Cricut Craft Life

1 friend · 24,955 members



+ Join




ABC Streets of Austin

319 members


+ Join

Local



Sun City I

1,741 members



#Revolution18

1,503 members

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

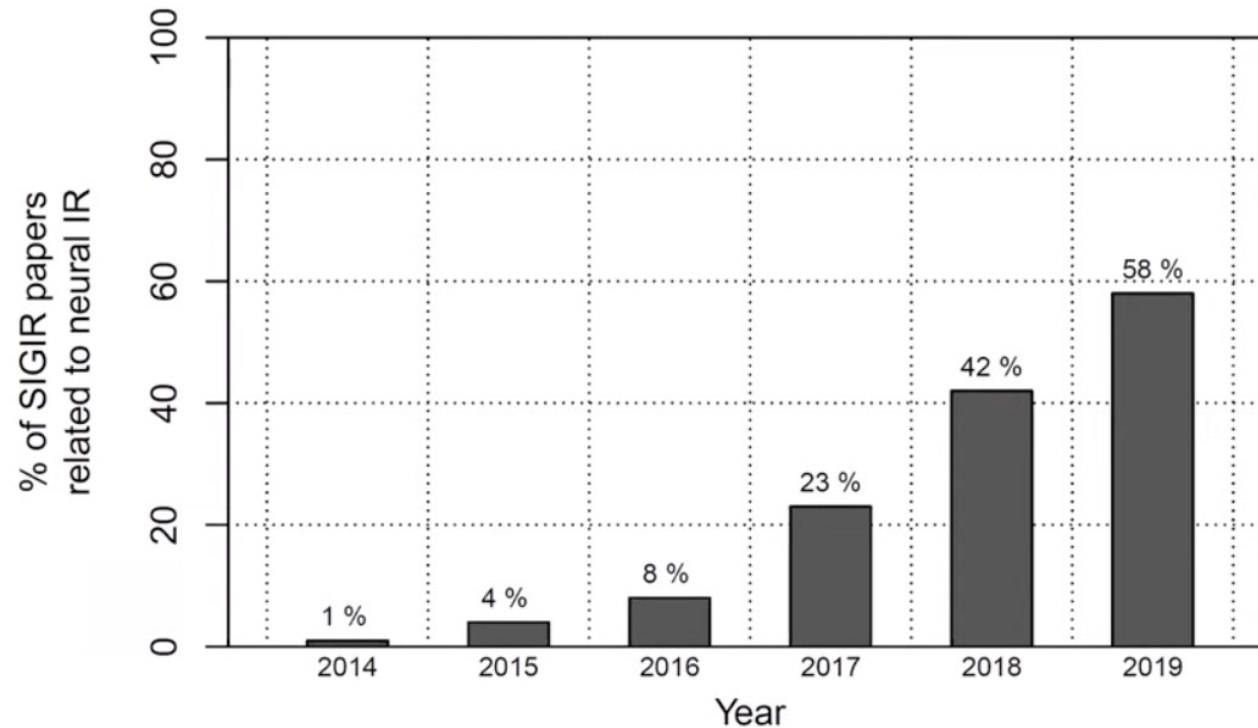
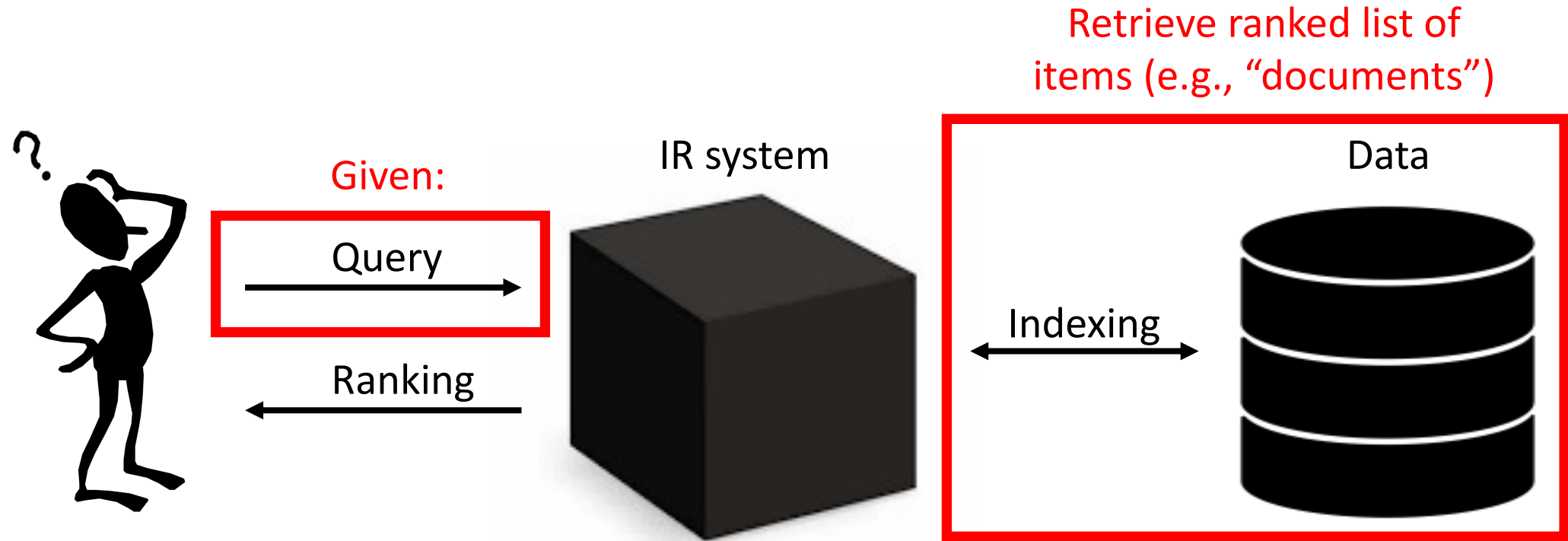


Figure 1.1: The percentage of neural IR papers at the ACM SIGIR conference—as determined by a manual inspection of the papers—shows a clear trend in the growing popularity of the field.

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

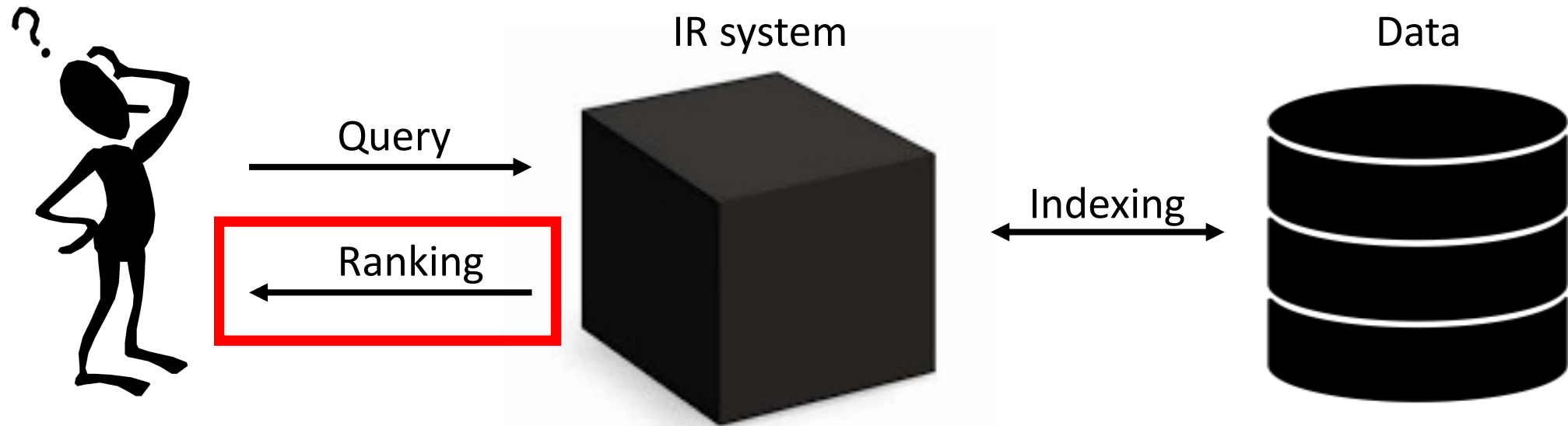
Excellent talk summarizing progress: <https://www.microsoft.com/en-us/research/video/neural-information-retrieval-in-search-of-meaningful-progress-ciir-talk-series-university-of-massachusetts-amherst/>

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

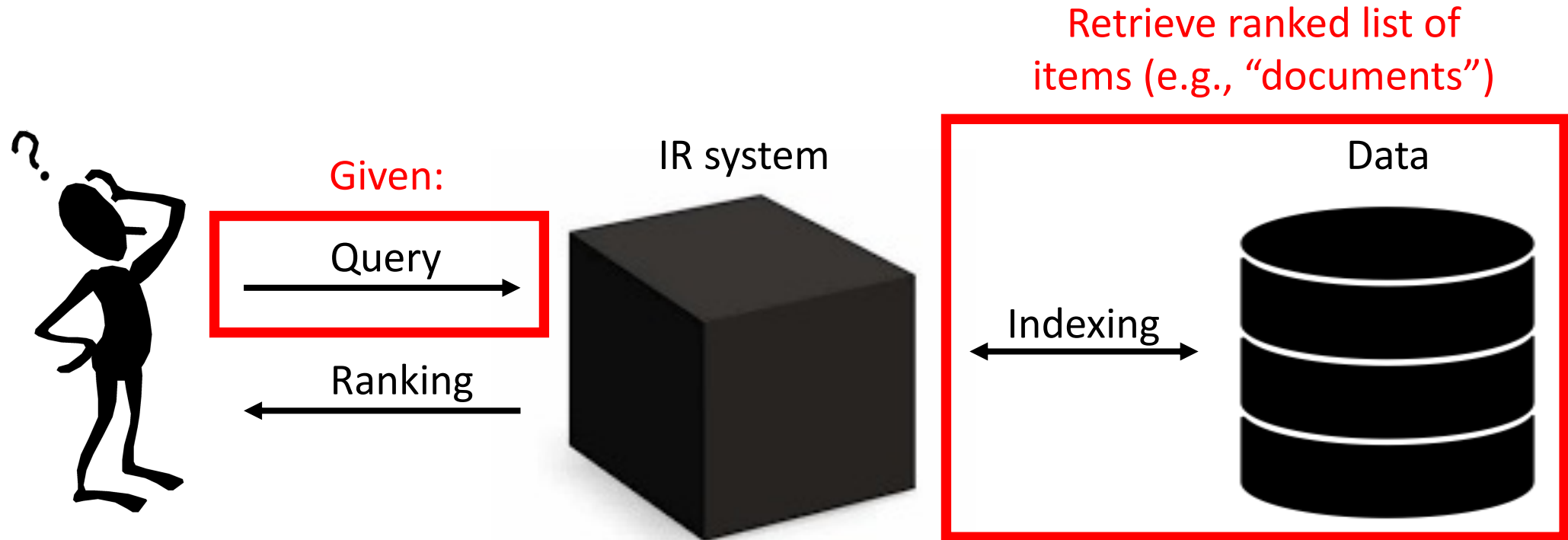
General Approach



How many examples do you want to see when searching?

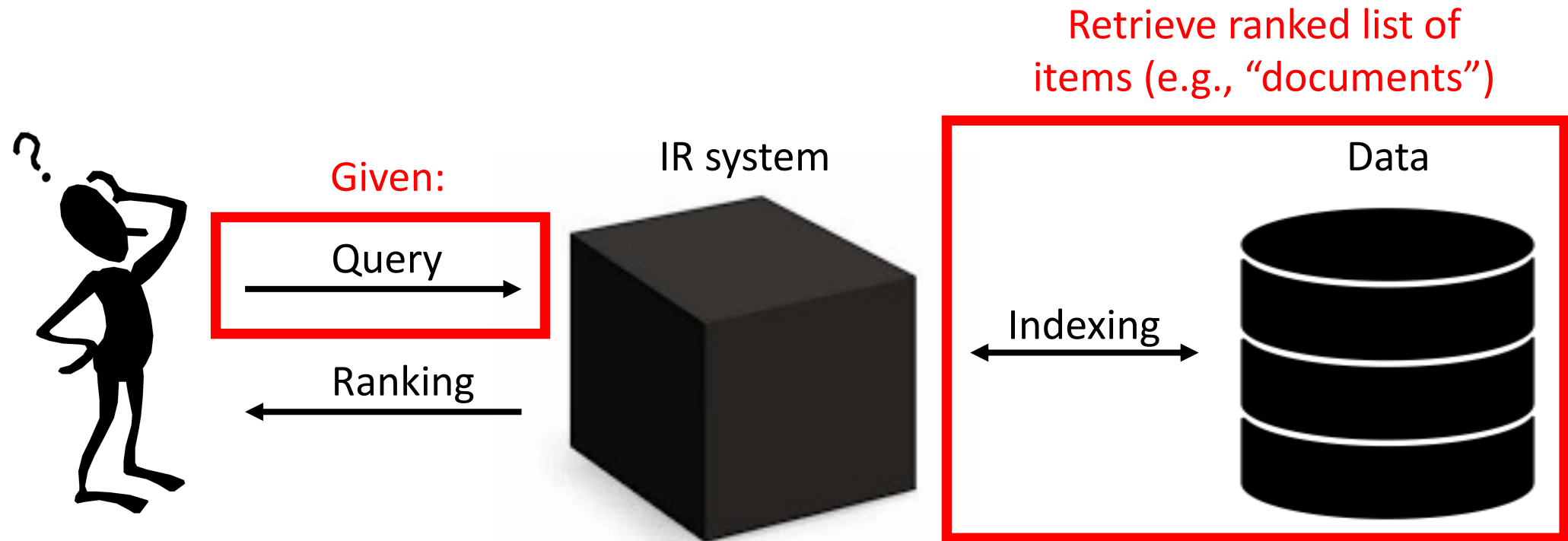
a. 1, b. 2-5, c. 6-10, d. More than 10, e. Undecided

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)

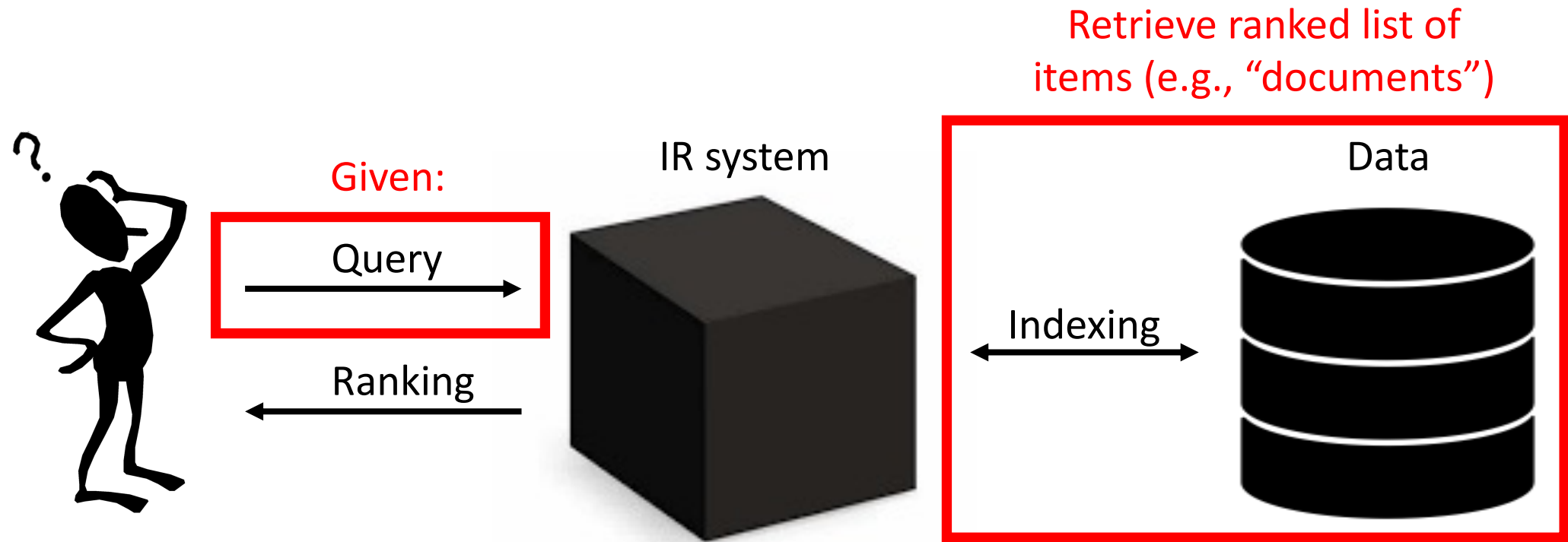
General Approach



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

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 · 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



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."

in

tw

f

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

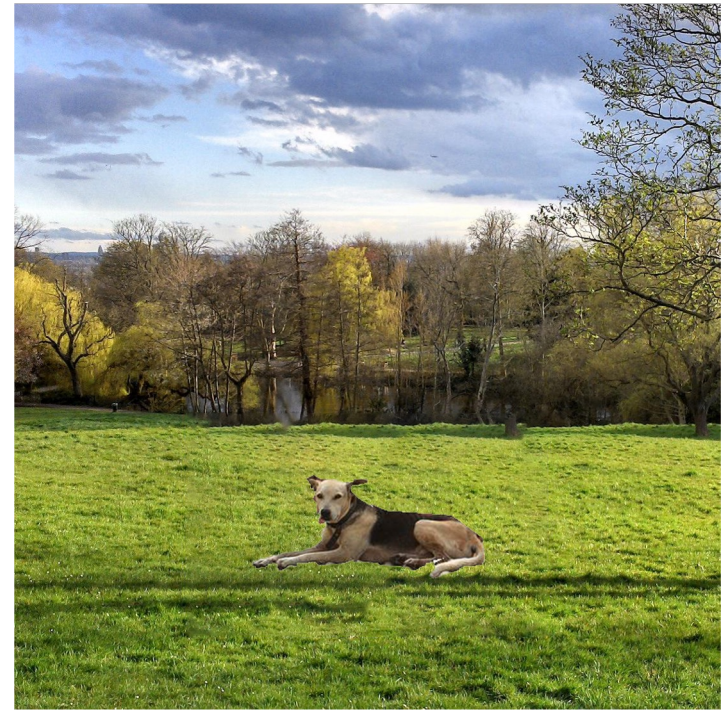
Hole-filling approach: find object, position it, and **fill remainder of hole**



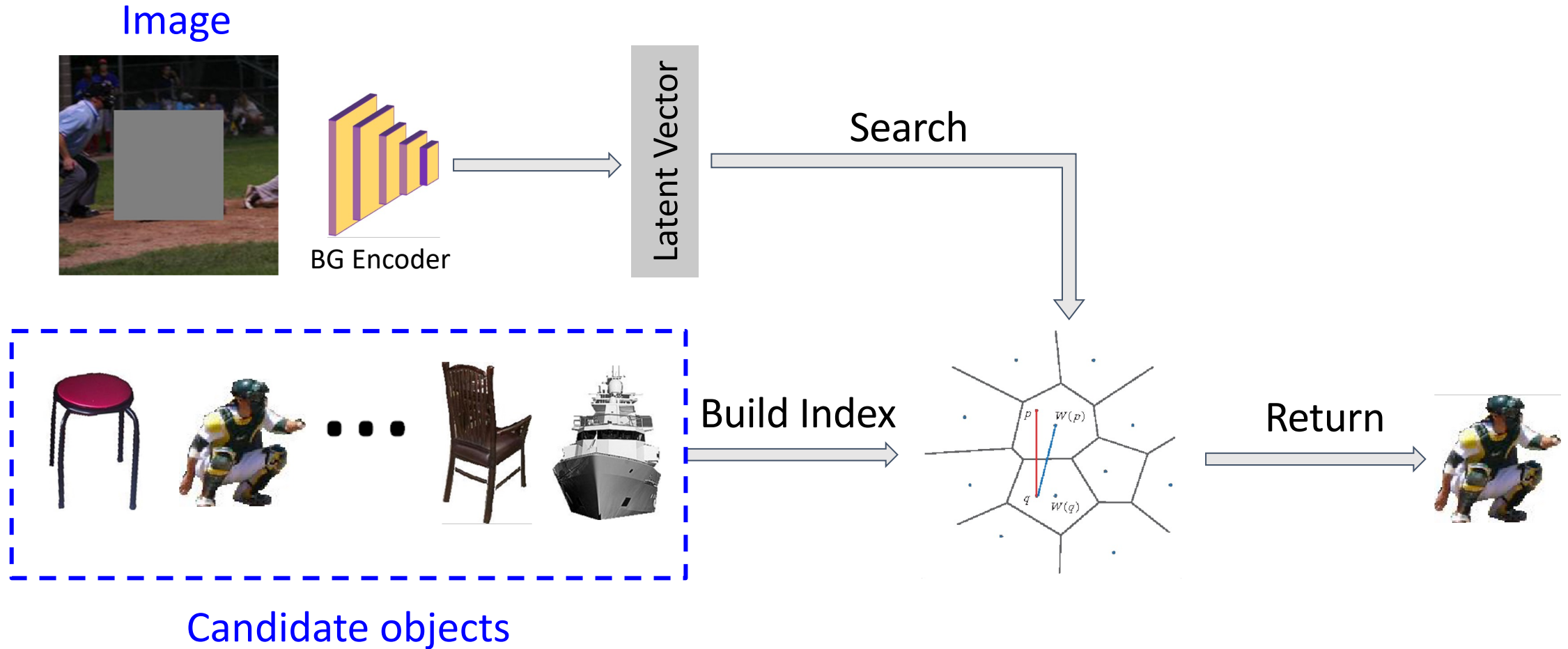
Synthesize

e.g., Searching for Objects to Inpaint

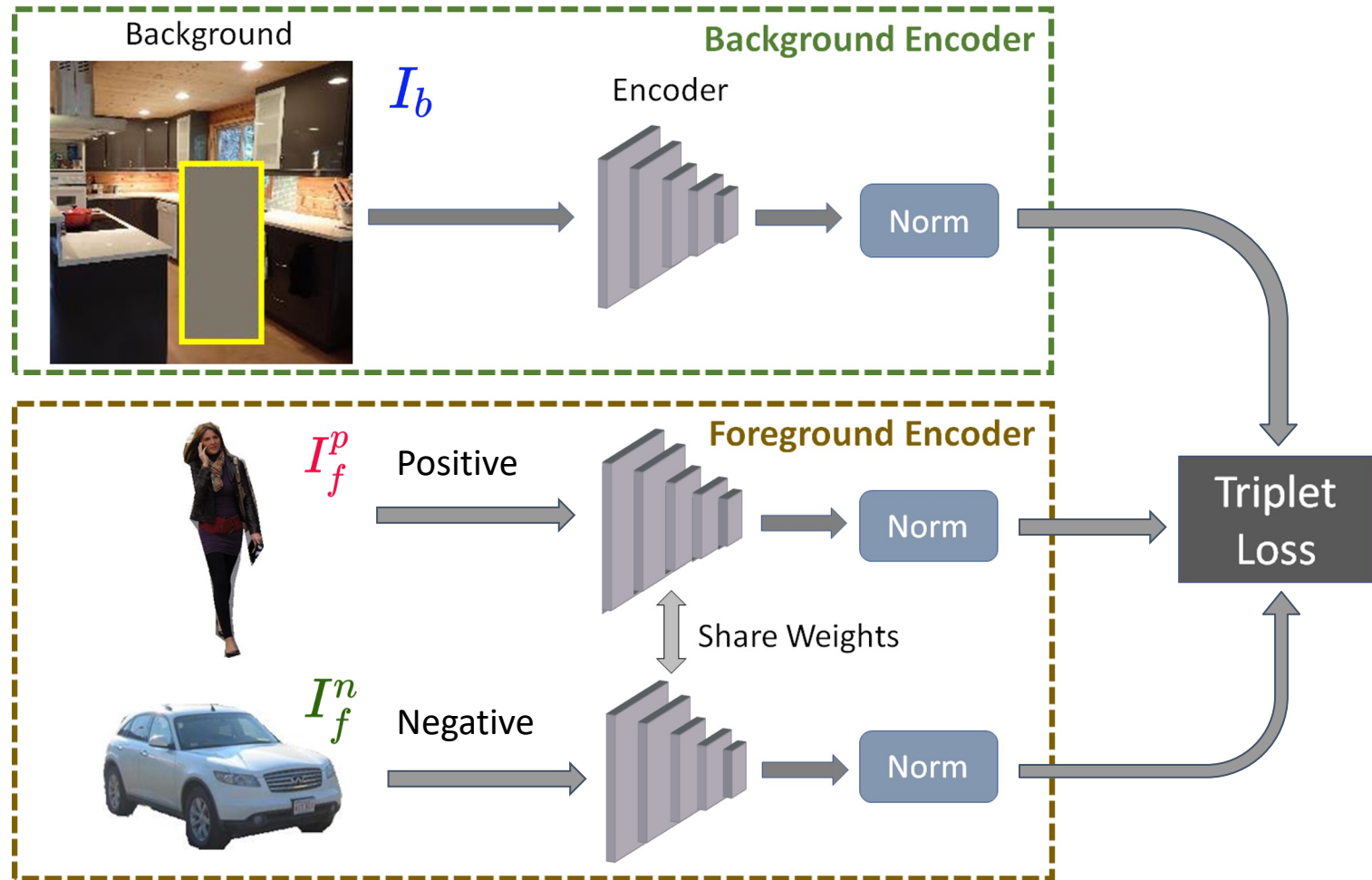
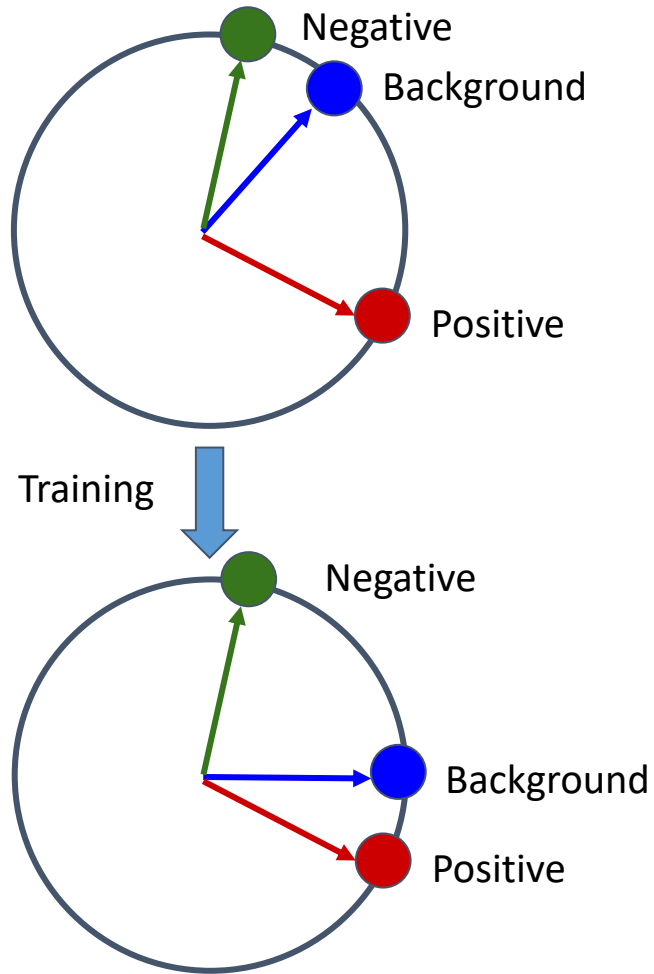
Hole-filling approach: find object, position it, and **fill remainder of hole**



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



UFO Search Training

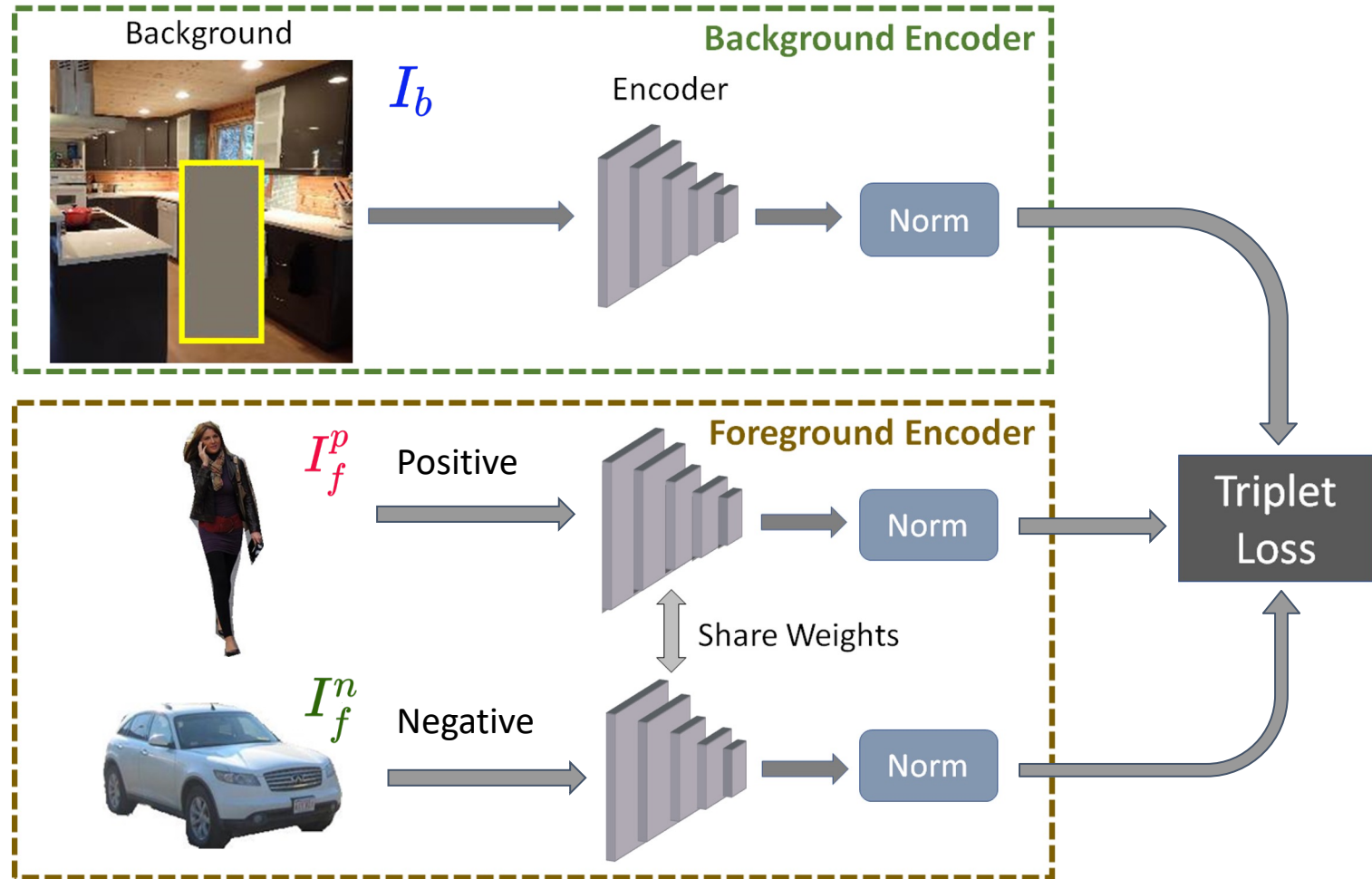


Positive objects are pushed closer to the background image than negative ones

Method: Base Architecture

We want:

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



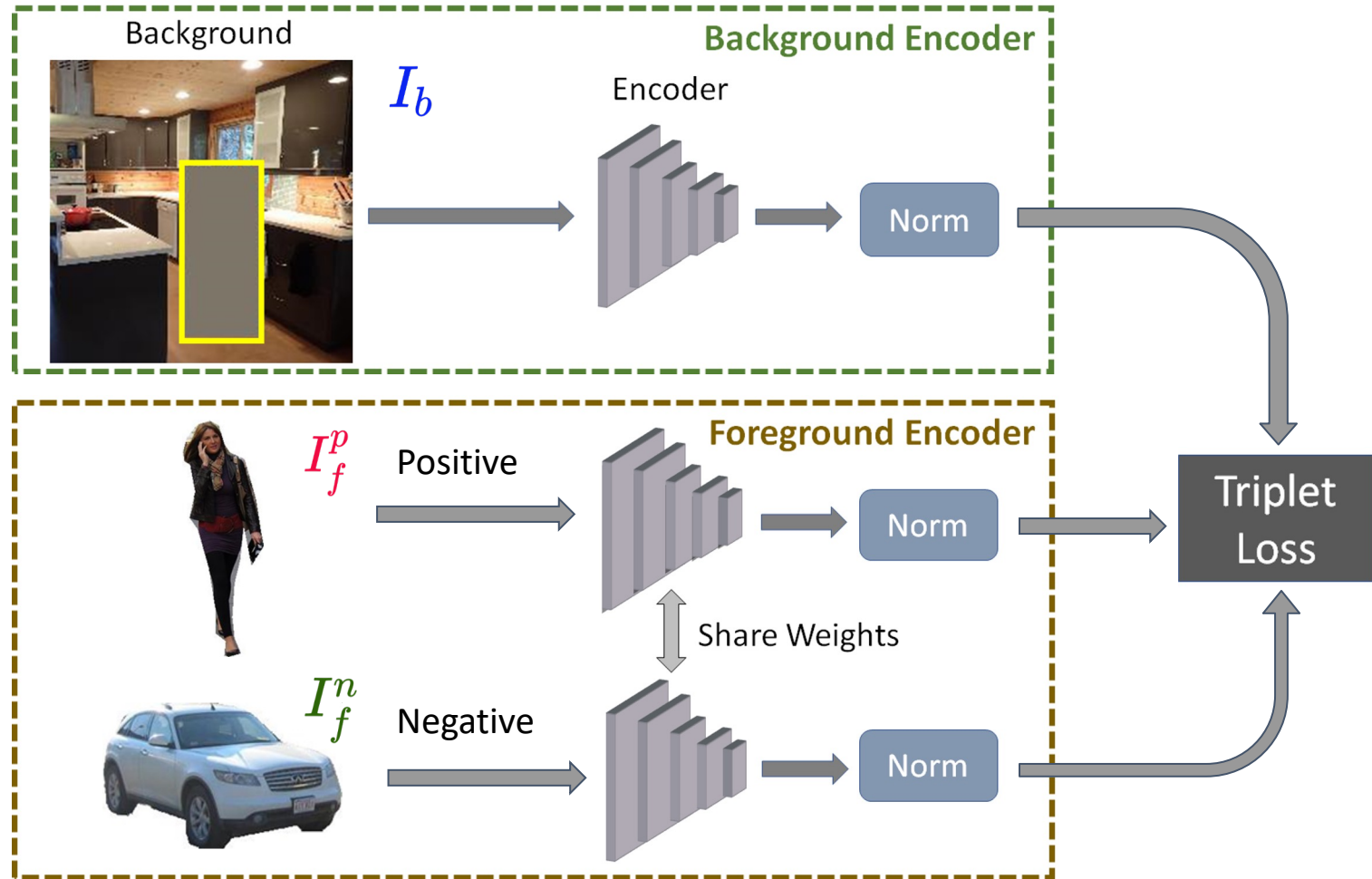
Method: Base Architecture

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:

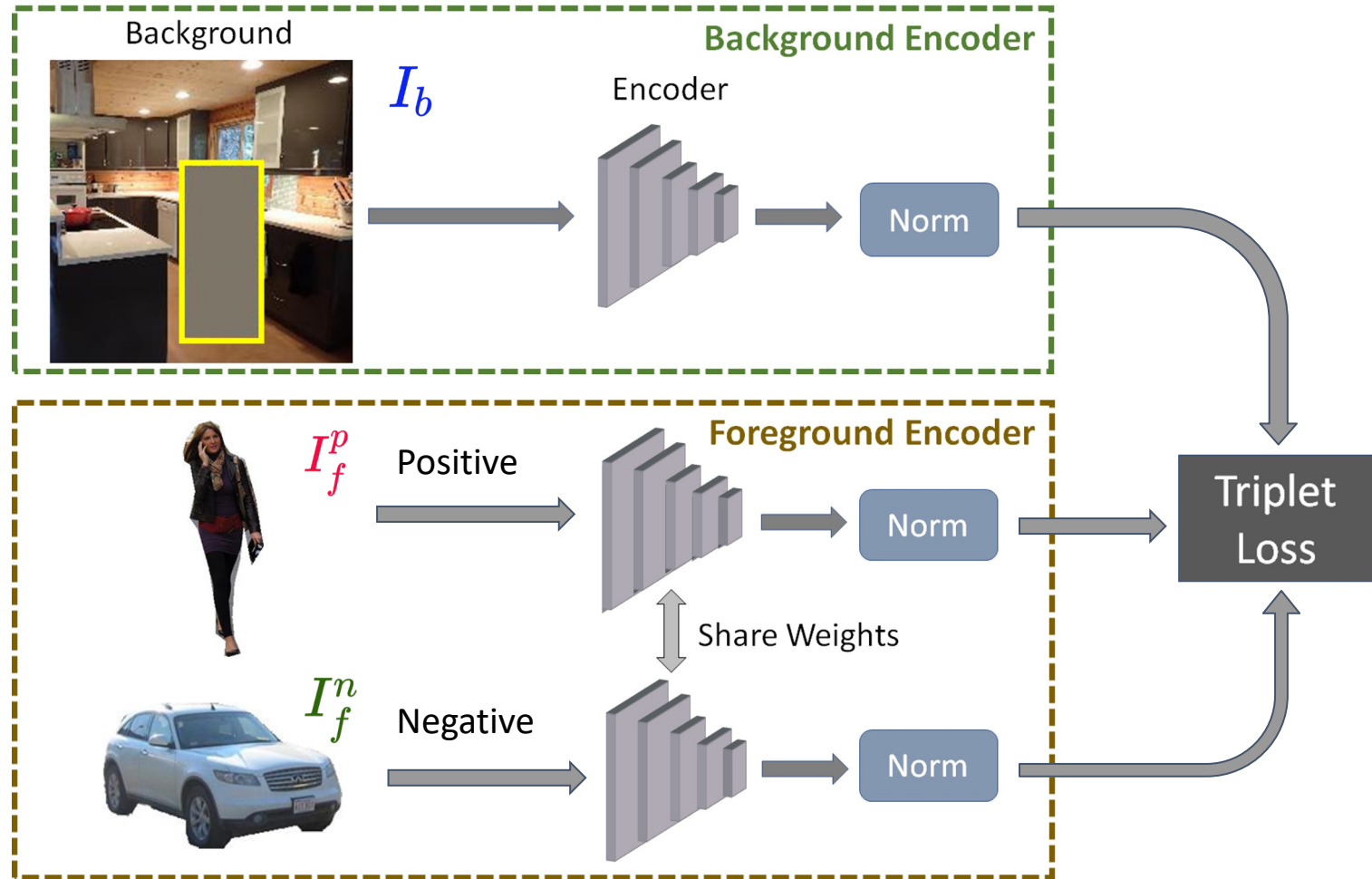
$$C(I_b, I_f^p) > C(I_b, I_f^n) + M$$

We use the triplet loss:

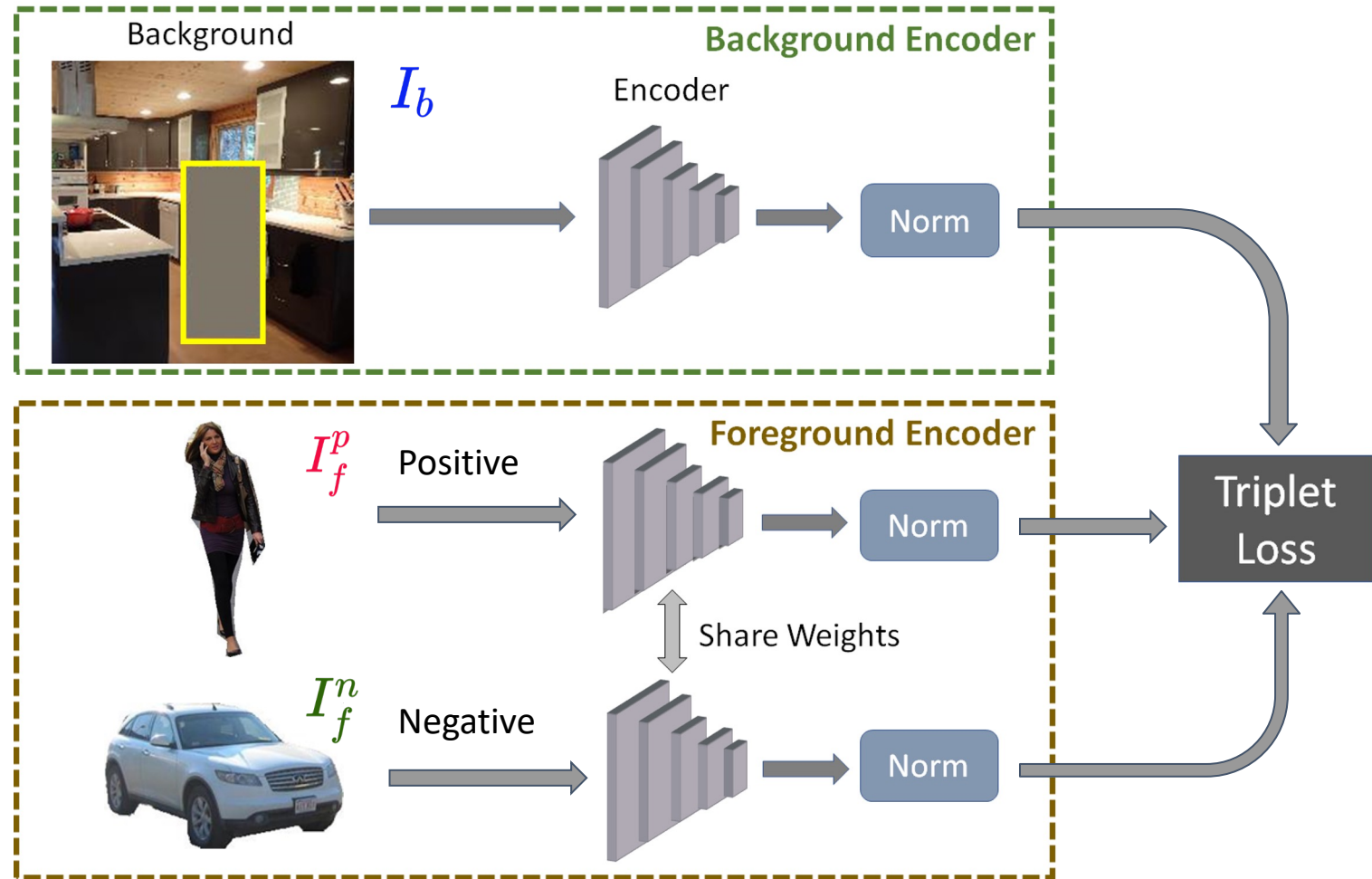
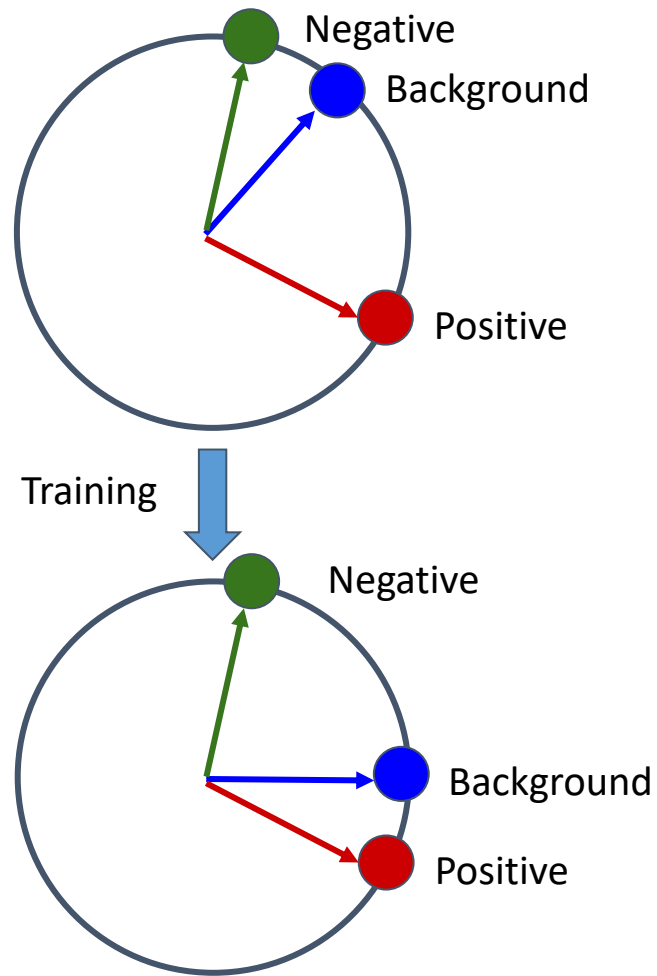
$$\max(0, C(I_b, I_f^n) + M - C(I_b, I_f^p))$$

We define compatibility as:

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



UFO Search Training

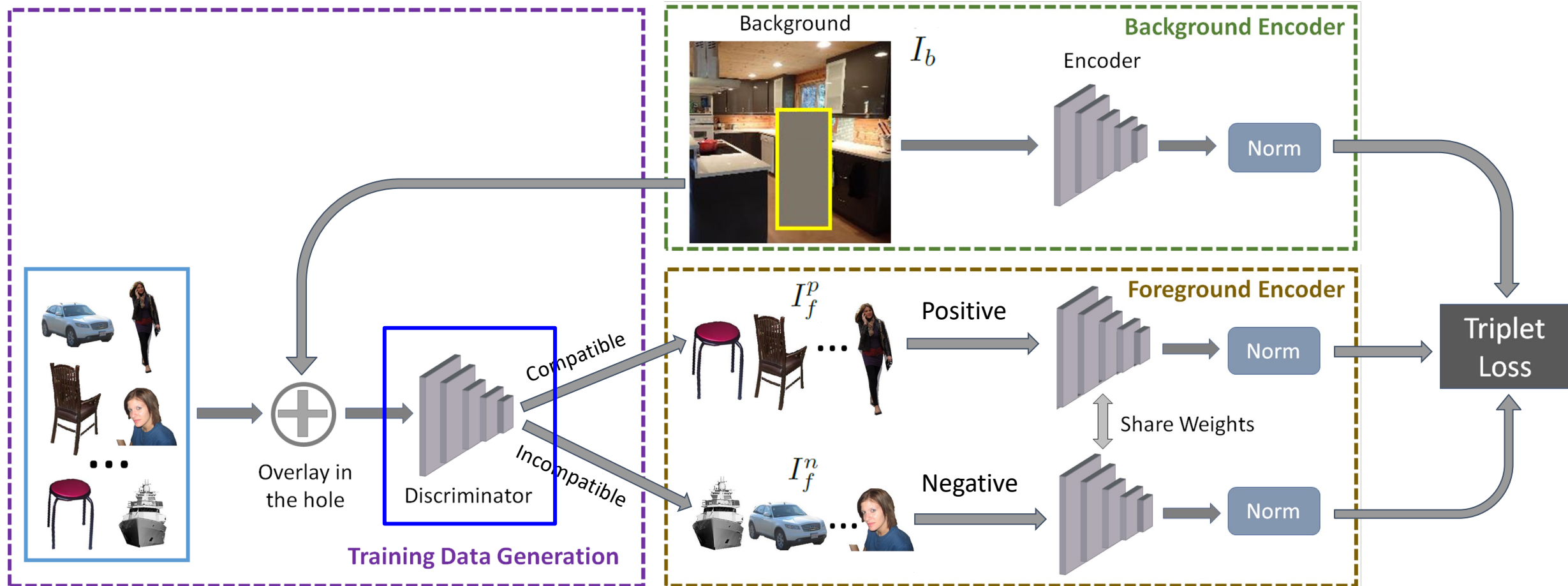


Key challenge: how to obtain training data for positive objects?

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

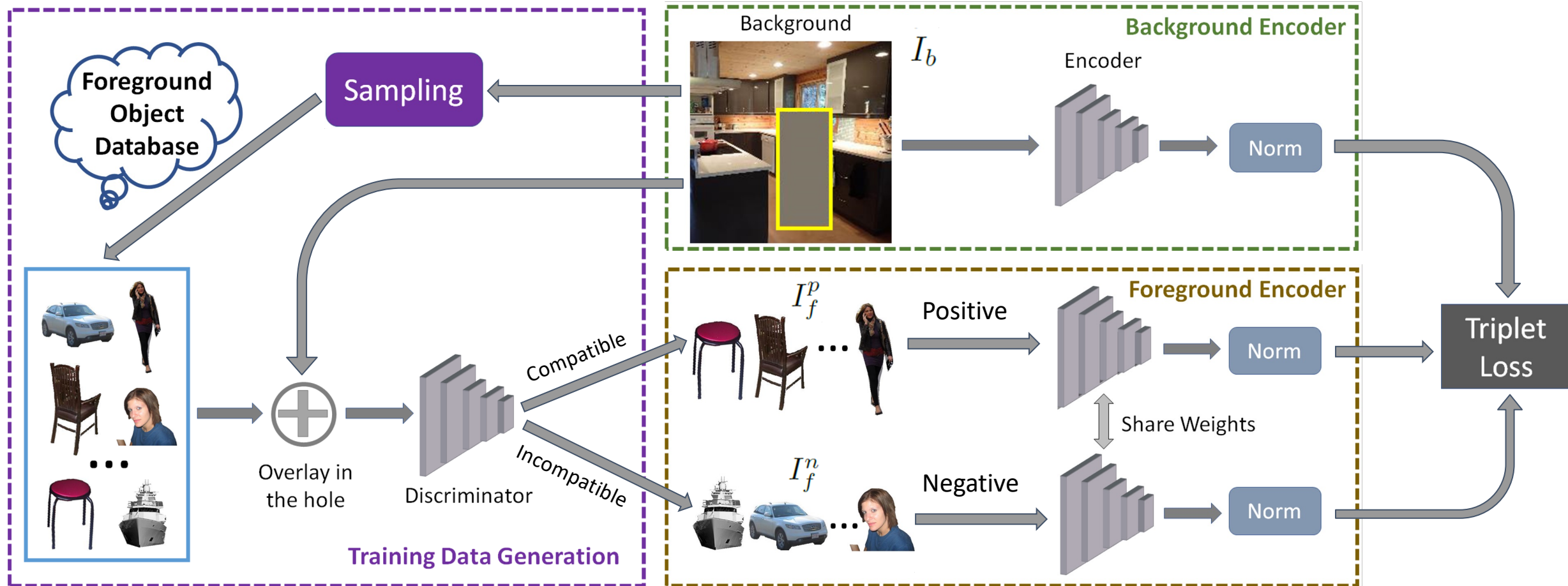
[Zhao et al. ICCV 2019]

UFO Search Training



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

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

UFO Search Training

- Similar to the original object



UFO Search Training

- Similar to the original object



UFO Search Training

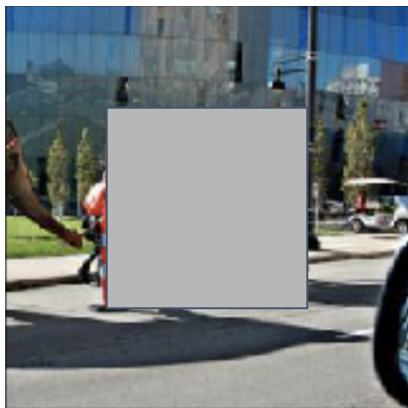


Similar Objects



UFO Search Training

- Similar Background



UFO Search Training

- Similar Background



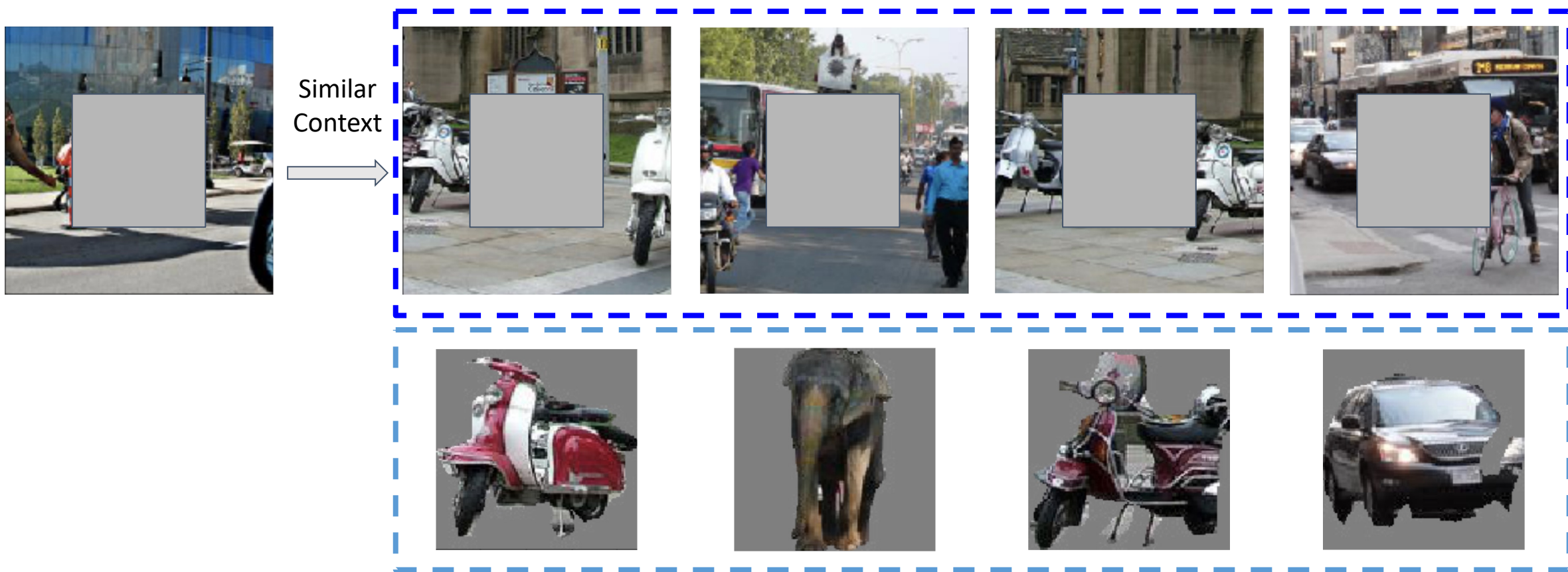
UFO Search Training

- Similar Background



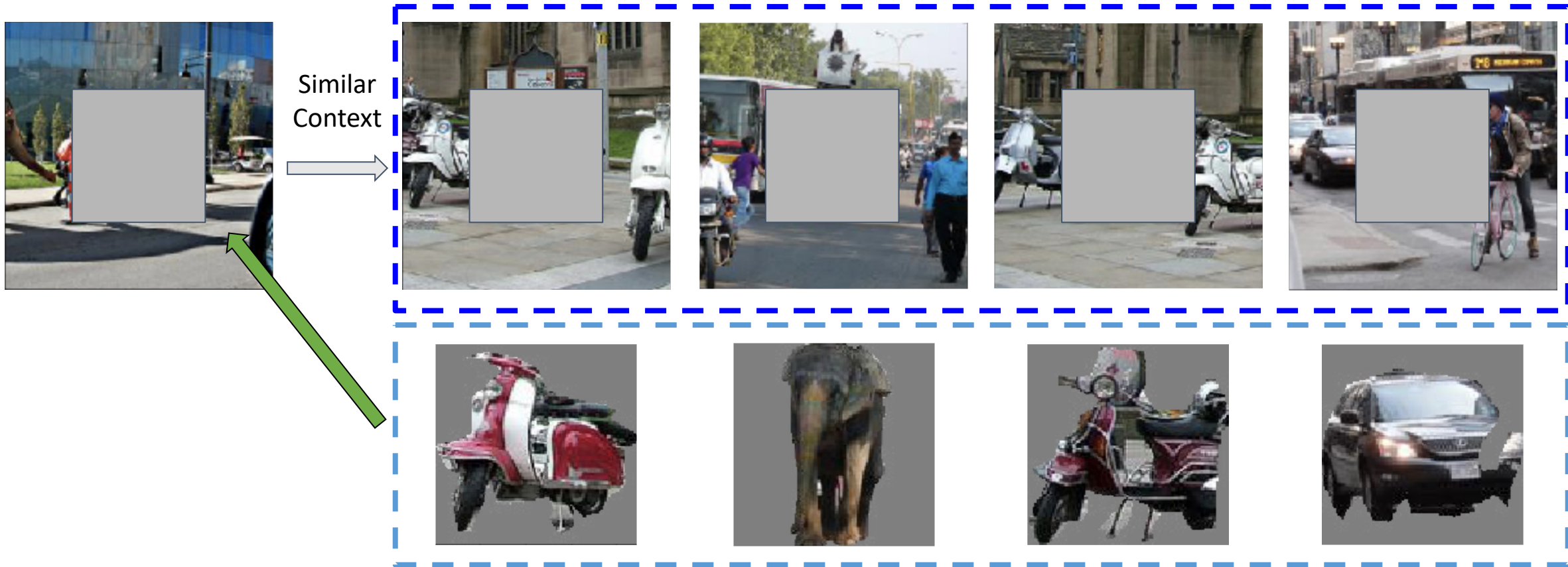
UFO Search Training

- Similar Background



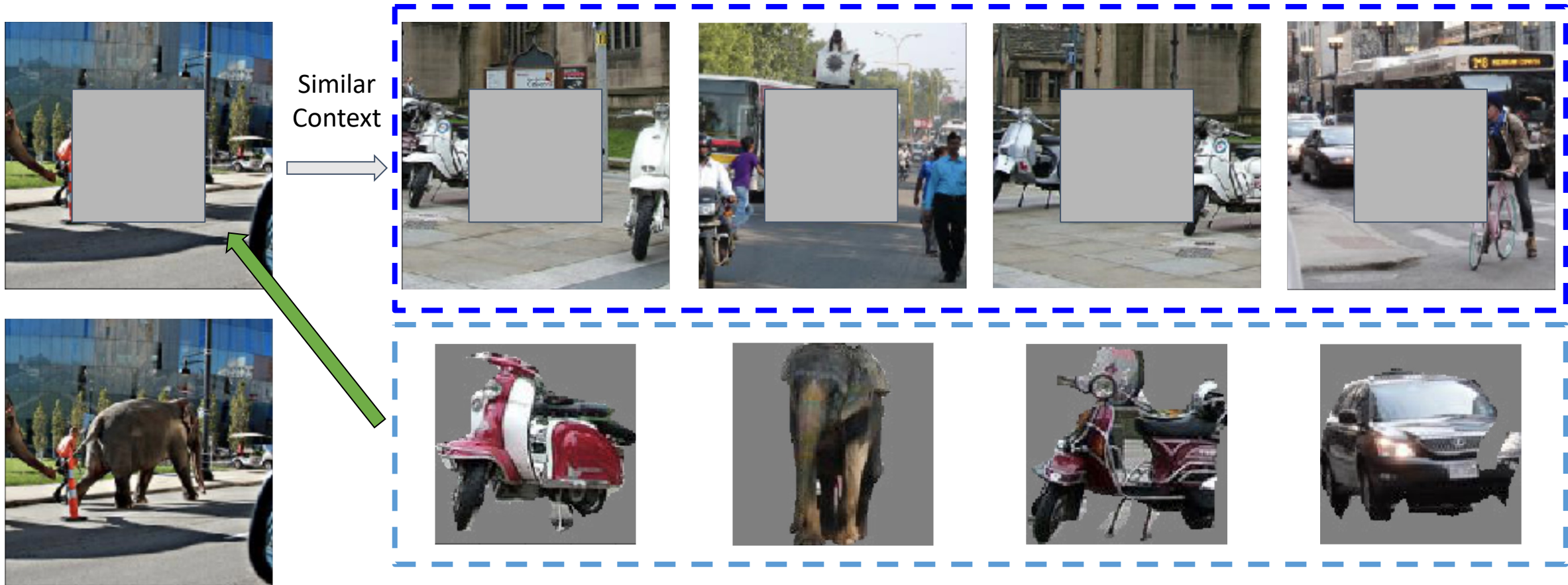
UFO Search Training

- Similar Background

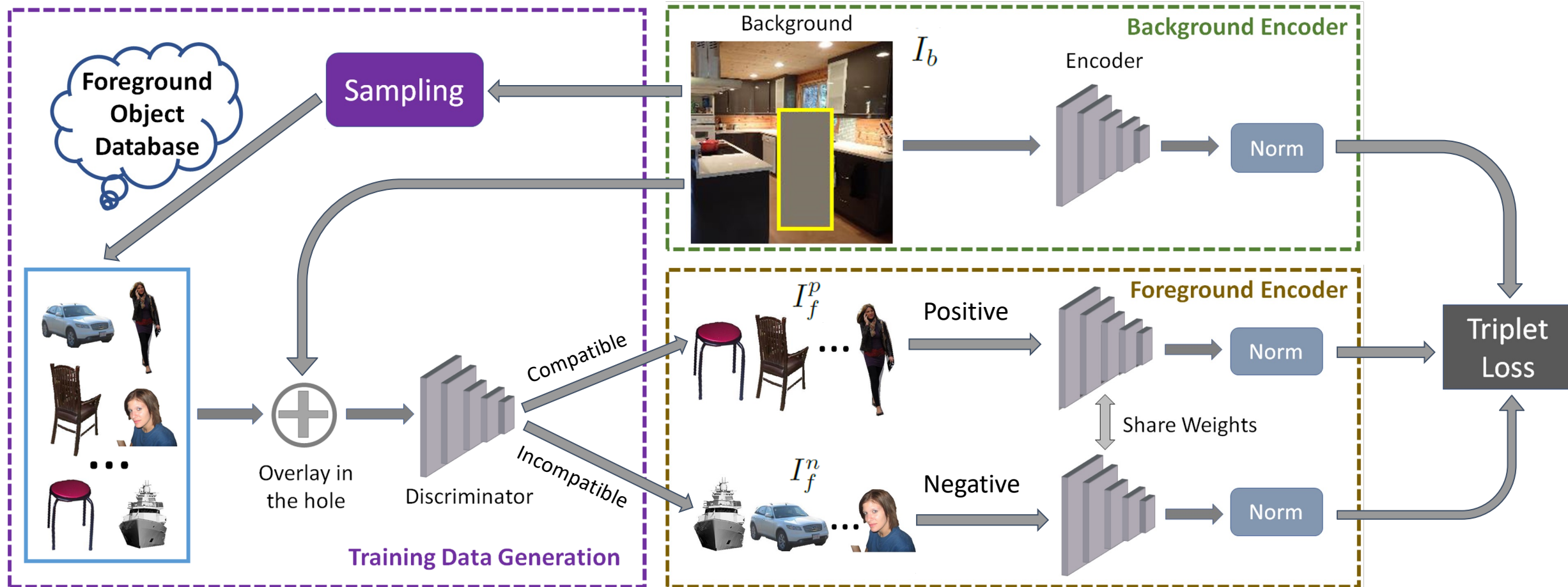


UFO Search Training

- Similar Background



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



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>)

Symbolic	Neural
<ul style="list-style-type: none">• Rules-based• Mature technology• Predictable• Requires maintenance and tuning• Can only find what is in the rules	<ul style="list-style-type: none">• Data-based• Evolving• Variable• Resilient, automatic• Neither requires nor respects rules

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