



# Topics in AI (CPSC 532S): Multimodal Learning with Vision, Language and Sound



A decorative horizontal bar at the bottom of the slide, consisting of five colored segments: light green, medium green, light blue, medium blue, and light purple.

**Lecture 1: Introduction**

# Course logistic

**Times:** Tues & Thurs 11-12:30pm

**Locations:** ICICS 246

**Instructor:** Leonid ~~—~~ Sigal



**E-mail:** [lsigal@cs.ubc.ca](mailto:lsigal@cs.ubc.ca)

**Office:** Hybrid

**Course webpage:** [https://www.cs.ubc.ca/~lsigal/teaching22\\_Term1.html](https://www.cs.ubc.ca/~lsigal/teaching22_Term1.html)

**Discussion:** [piazza.com/ubc.ca/winterterm12022/cpsc532s/home](https://piazza.com/ubc.ca/winterterm12022/cpsc532s/home)

# Course logistic

**Times:** Tues & Thurs 11-12:30pm

**Locations:** ICICS 246

**Instructor:** Leonid ~~—~~ Sigal



**E-mail:** [lsigal@cs.ubc.ca](mailto:lsigal@cs.ubc.ca)

**Office:** Hybrid



This course is oversubscribed  
there is a **waiting list**

Those who are hoping to register are welcome to  
stay, but registered student get seat priority!

**Course webpage:** [https://www.cs.ubc.ca/~lsigal/teaching22\\_Term1.html](https://www.cs.ubc.ca/~lsigal/teaching22_Term1.html)

**Discussion:** [piazza.com/ubc.ca/winterterm12022/cpsc532s/home](https://piazza.com/ubc.ca/winterterm12022/cpsc532s/home)

# Course logistic

**Times:** Tues & Thurs 11-12:30pm

**Locations:** ICICS 246

**Instructor:** Leonid ~~—~~ Sigal



**E-mail:** [lsigal@cs.ubc.ca](mailto:lsigal@cs.ubc.ca)

**Office:** Hybrid

**Course webpage** is the first stop for all materials  
(lecture slides, assignments, readings, etc.)

We will use **Canvas** only for assignment  
submission and grading

**Course webpage:** [https://www.cs.ubc.ca/~lsigal/teaching22\\_Term1.html](https://www.cs.ubc.ca/~lsigal/teaching22_Term1.html)

**Discussion:** [piazza.com/ubc.ca/winterterm12022/cpsc532s/home](https://piazza.com/ubc.ca/winterterm12022/cpsc532s/home)

# Course logistic

**Times:** Tues & Thurs 11-12:30pm

**Locations:** ICICS 246

**Instructor:** Leonid ~~—~~ Sigal



**E-mail:** [lsigal@cs.ubc.ca](mailto:lsigal@cs.ubc.ca)

**Office:** Hybrid

**Piazza** is where discussion and majority of assignment / material help happens

Sign Up Pwd: **McGurk**

(we will learn about this later today)

**Course webpage:** [https://www.cs.ubc.ca/~lsigal/teaching22\\_Term1.html](https://www.cs.ubc.ca/~lsigal/teaching22_Term1.html)

**Discussion:** [piazza.com/ubc.ca/winterterm12022/cpsc532s/home](https://piazza.com/ubc.ca/winterterm12022/cpsc532s/home)

# Course logistic

**Times:** Tues & Thurs 11-12:30pm

**Locations:** ICICS 246

**Instructor:** Leonid Sigal



**E-mail:** [lsigal@cs.ubc.ca](mailto:lsigal@cs.ubc.ca)

**Office:** Hybrid

**TA:** Tanzila Rahman



**E-mail:** [tanzila.himu@gmail.com](mailto:tanzila.himu@gmail.com)

Rayat Hossain



[rayat137@cs.ubc.ca](mailto:rayat137@cs.ubc.ca)

**Course webpage:** [https://www.cs.ubc.ca/~lsigal/teaching22\\_Term1.html](https://www.cs.ubc.ca/~lsigal/teaching22_Term1.html)

**Discussion:** [piazza.com/ubc.ca/winterterm12022/cpsc532s/home](https://piazza.com/ubc.ca/winterterm12022/cpsc532s/home)

# Course logistic

**Times:** Tues & Thurs 11-12:30pm

**Locations:** ICICS 246

Lectures will **not** be recorded.

Lecture material will be delivered in class. If you missed a class, talk to me. I can potentially provide recordings from last offering (will not be 1-1) and/or answer questions.

Slides **will** be made available (directly) before lecture.

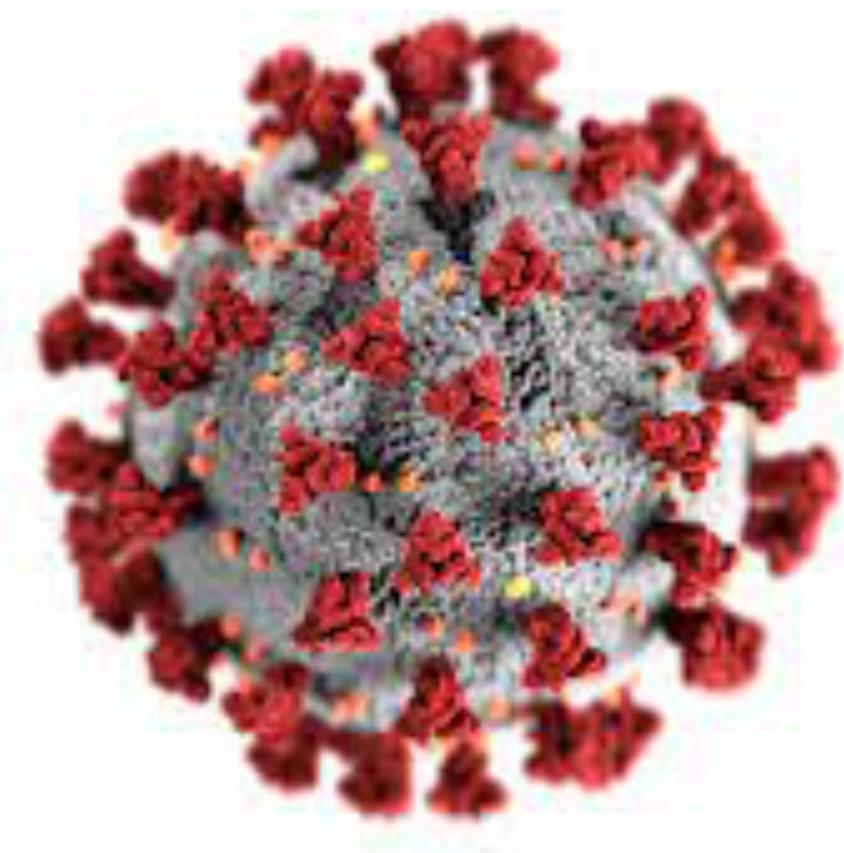
**Course webpage:** [https://www.cs.ubc.ca/~lsigal/teaching22\\_Term1.html](https://www.cs.ubc.ca/~lsigal/teaching22_Term1.html)

**Discussion:** [piazza.com/ubc.ca/winterterm12022/cpsc532s/home](https://piazza.com/ubc.ca/winterterm12022/cpsc532s/home)

# Course logistic

**Times:** Tues & Thurs 11-12:30pm

**Locations:** ICICS 246



Masks are **highly encouraged** as a courtesy to your neighbors  
(but are **not required** as per UBC and provincial policy)

I will not be wearing a mask when lecturing, but  
will put one on in close encounter settings (or as requested)

If you fall ill anytime during the term (whether it is COVID or not) please do stay home

**Course webpage:** [https://www.cs.ubc.ca/~lsigal/teaching22\\_Term1.html](https://www.cs.ubc.ca/~lsigal/teaching22_Term1.html)

**Discussion:** [piazza.com/ubc.ca/winterterm12022/cpsc532s/home](https://piazza.com/ubc.ca/winterterm12022/cpsc532s/home)

# About me ...

# About me ...

**Ugrad, MSc**  
2001 - 2008

**BOSTON**  
**UNIVERSITY**

# About me ...

# About me ...

**Software Engineer**  
1999 - 2001

**COGNEX**

# About me ...

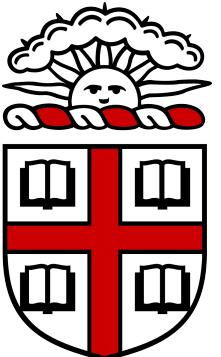


**Software Engineer**  
1999 - 2001

**COGNEX**

# About me ...

**PhD, MSc**  
2001 - 2008



BROWN

**BOSTON  
UNIVERSITY**

**Software Engineer**  
1999 - 2001

**COGNEX**

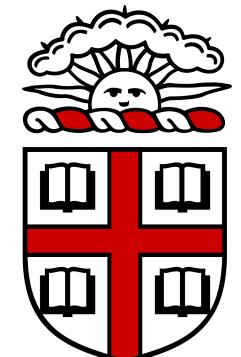
# About me ...

**Postdoctoral Researcher**  
2007 - 2009



UNIVERSITY OF  
**TORONTO**

**PhD, MSc**  
2001 - 2008



BROWN

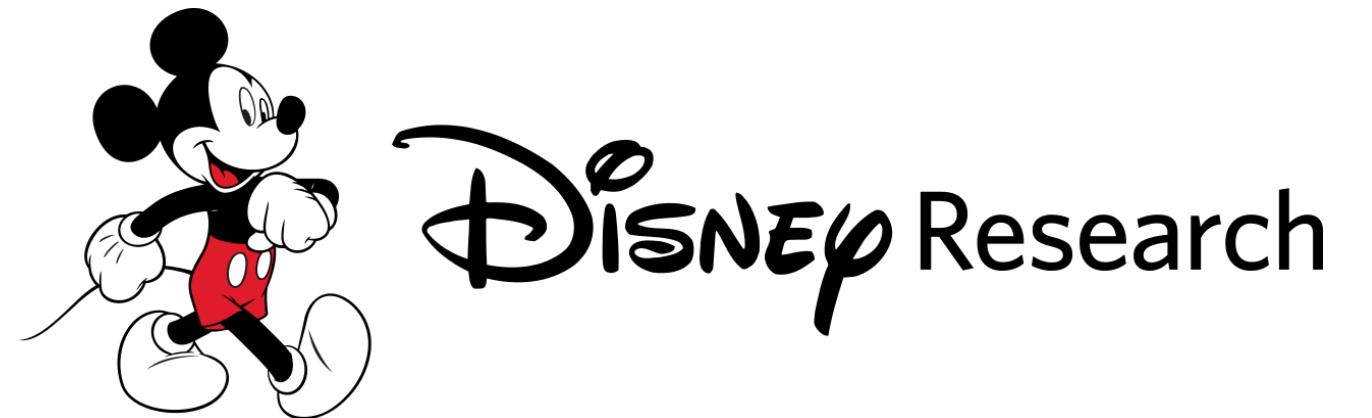
**BOSTON**  
UNIVERSITY

**Software Engineer**  
1999 - 2001

**COGNEX**

# About me ...

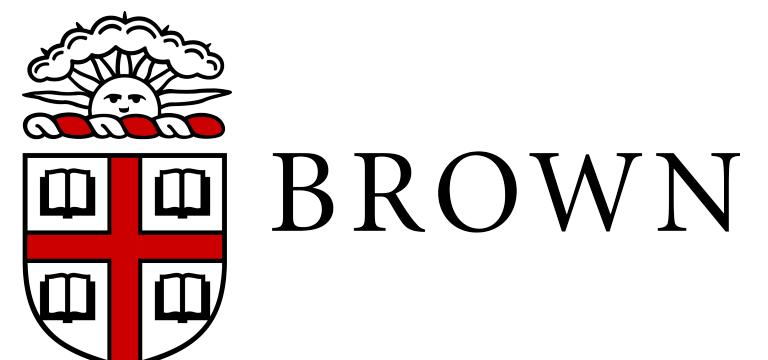
**Senior Research Scientist**  
2009 - 2017



**Postdoctoral Researcher**  
2007 - 2009



**PhD, MSc**  
2001 - 2008



**Software Engineer**  
1999 - 2001



# About me ...

**Associate Professor**

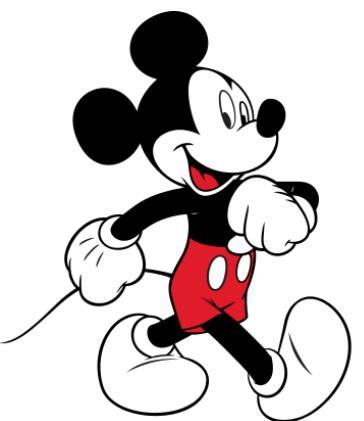
2017 -



THE UNIVERSITY  
OF BRITISH COLUMBIA

**Senior Research Scientist**

2009 - 2017



**Postdoctoral Researcher**

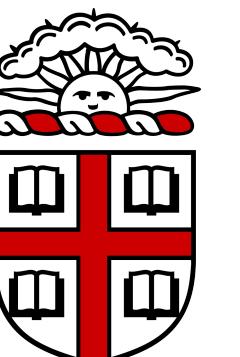
2007 - 2009



UNIVERSITY OF  
**TORONTO**

**PhD, MSc**

2001 - 2008



BROWN

**BOSTON**  
UNIVERSITY

**Software Engineer**

1999 - 2001

**COGNEX**

# About me ...

I have been working  
in **Computer Vision**  
for the last ~20 years

**Associate Professor**

2017 -



THE UNIVERSITY  
OF BRITISH COLUMBIA

**Senior Research Scientist**

2009 - 2017



Disney Research

**Postdoctoral Researcher**

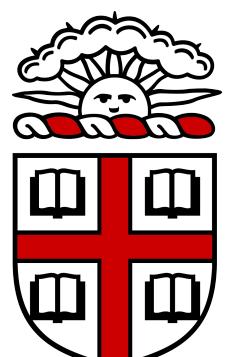
2007 - 2009



UNIVERSITY OF  
**TORONTO**

**PhD, MSc**

2001 - 2008



BROWN

**BOSTON**  
UNIVERSITY

**Software Engineer**

1999 - 2001

**COGNEX**

# About me ...

I have been working  
in **Computer Vision**  
for the last ~20 years

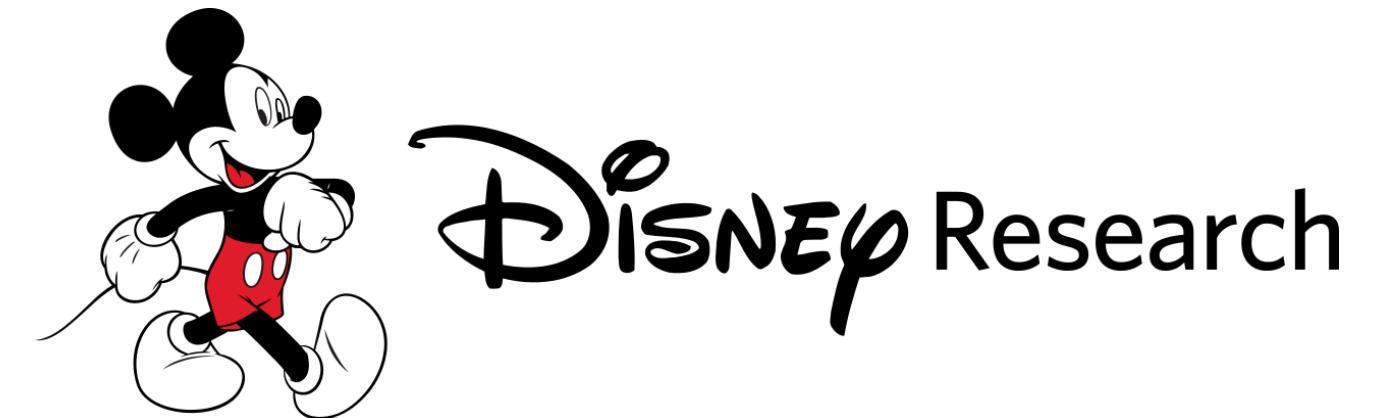
In **multi-modal  
learning** for the last  
~5 year

**Associate Professor**  
2017 -



THE UNIVERSITY  
OF BRITISH COLUMBIA

**Senior Research Scientist**  
2009 - 2017

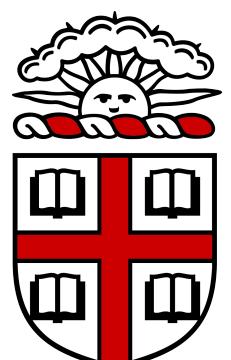


**Postdoctoral Researcher**  
2007 - 2009



UNIVERSITY OF  
**TORONTO**

**PhD, MSc**  
2001 - 2008

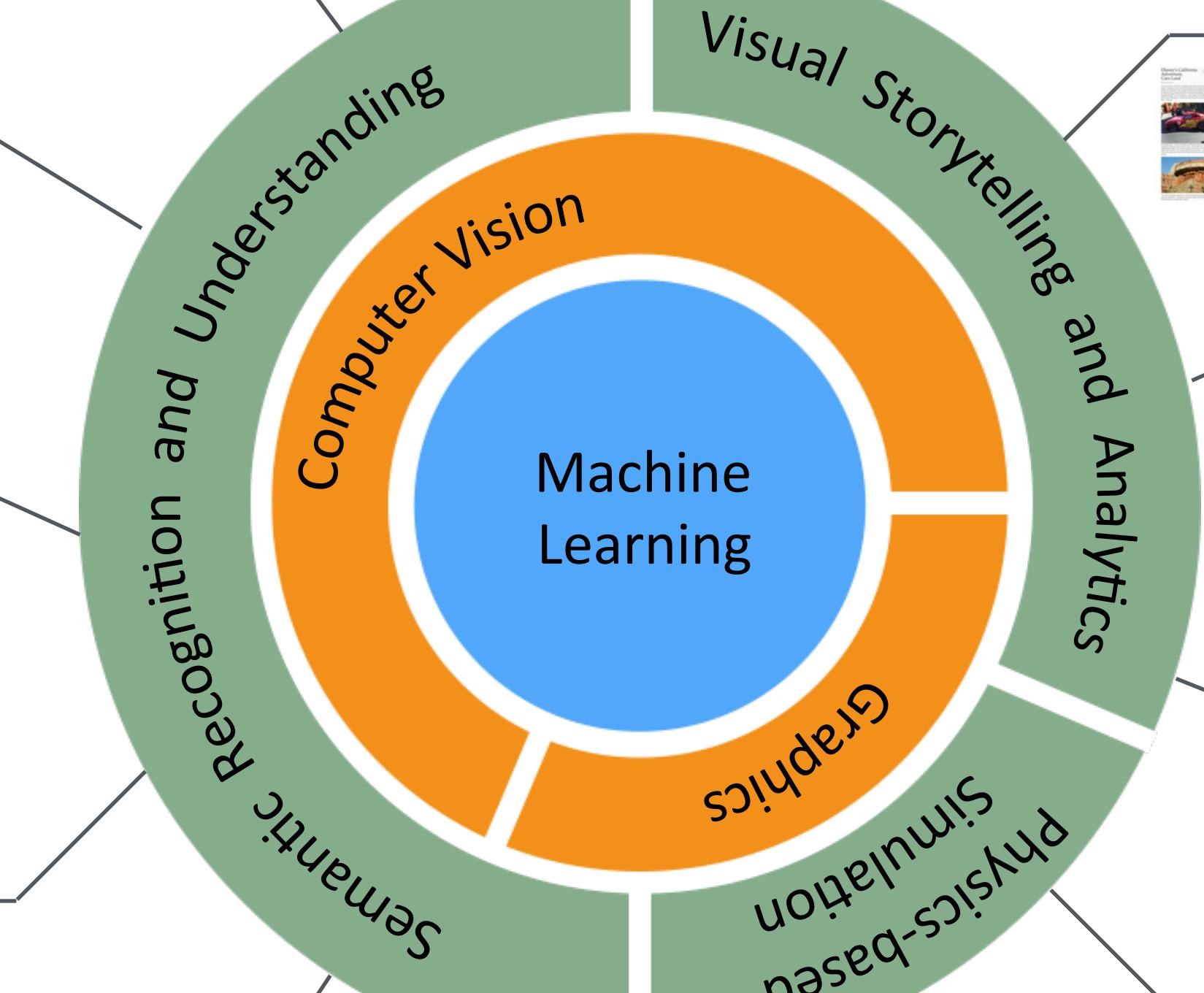
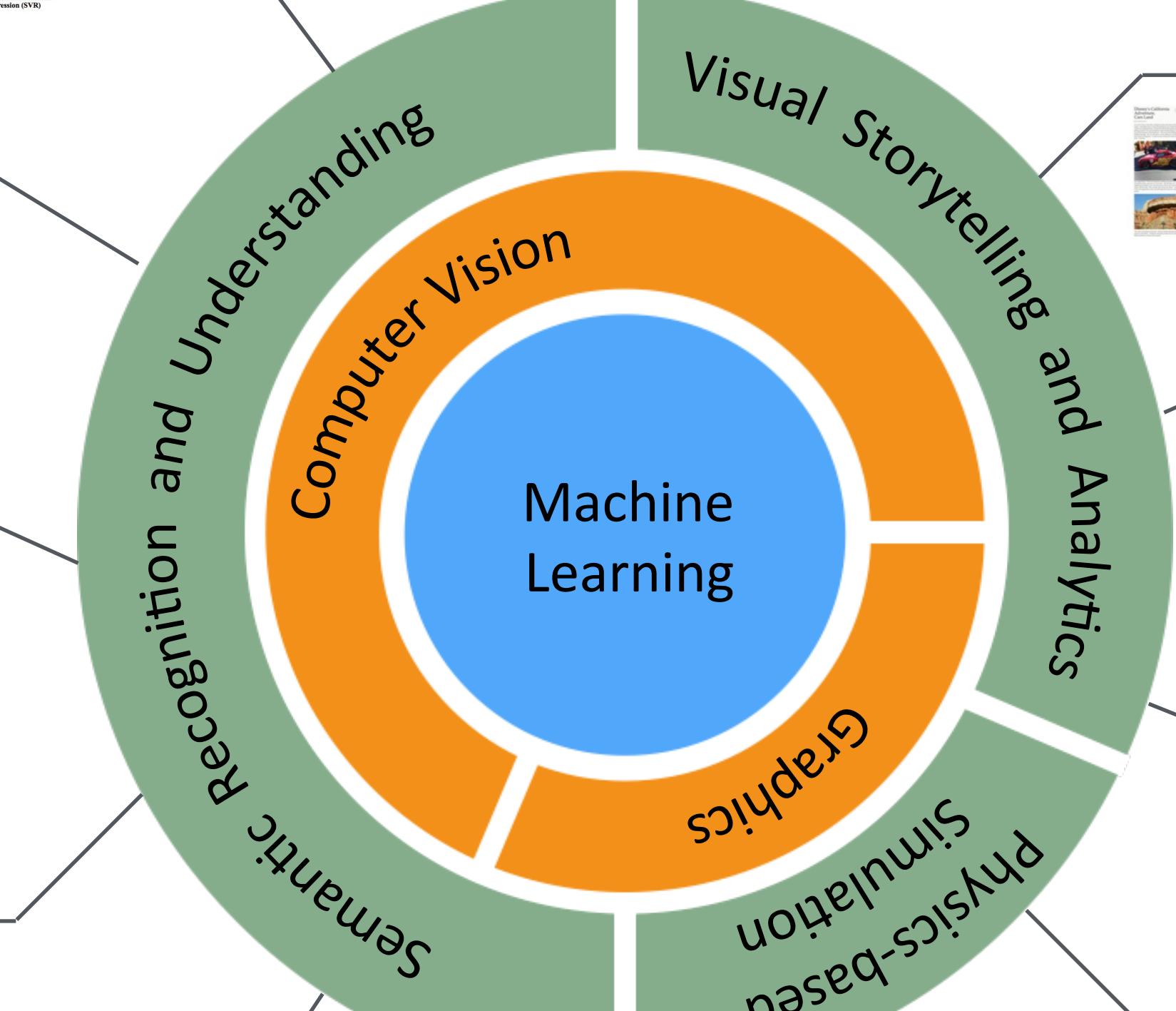


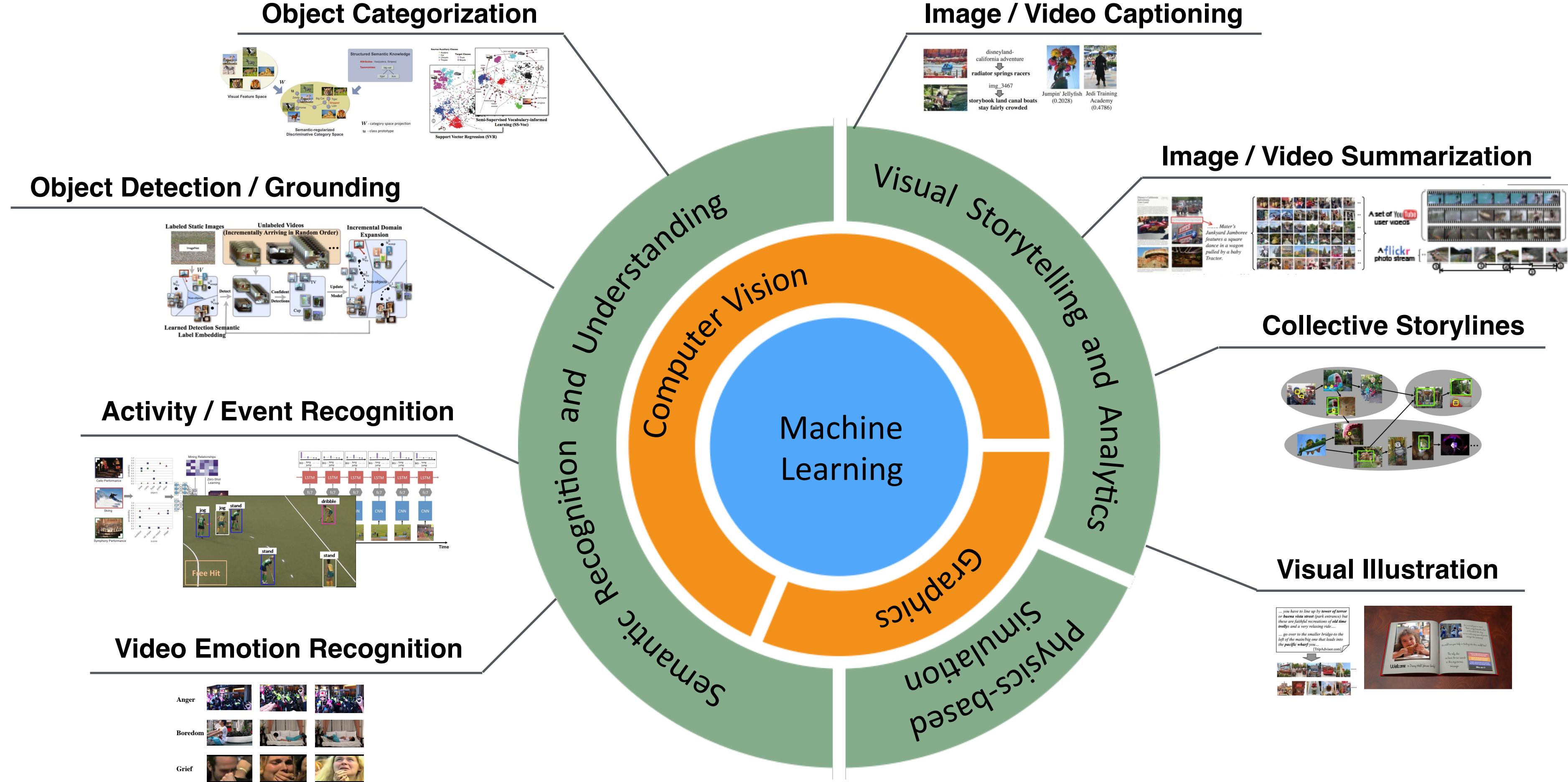
BROWN

**BOSTON  
UNIVERSITY**

**Software Engineer**  
1999 - 2001

**COGNEX**

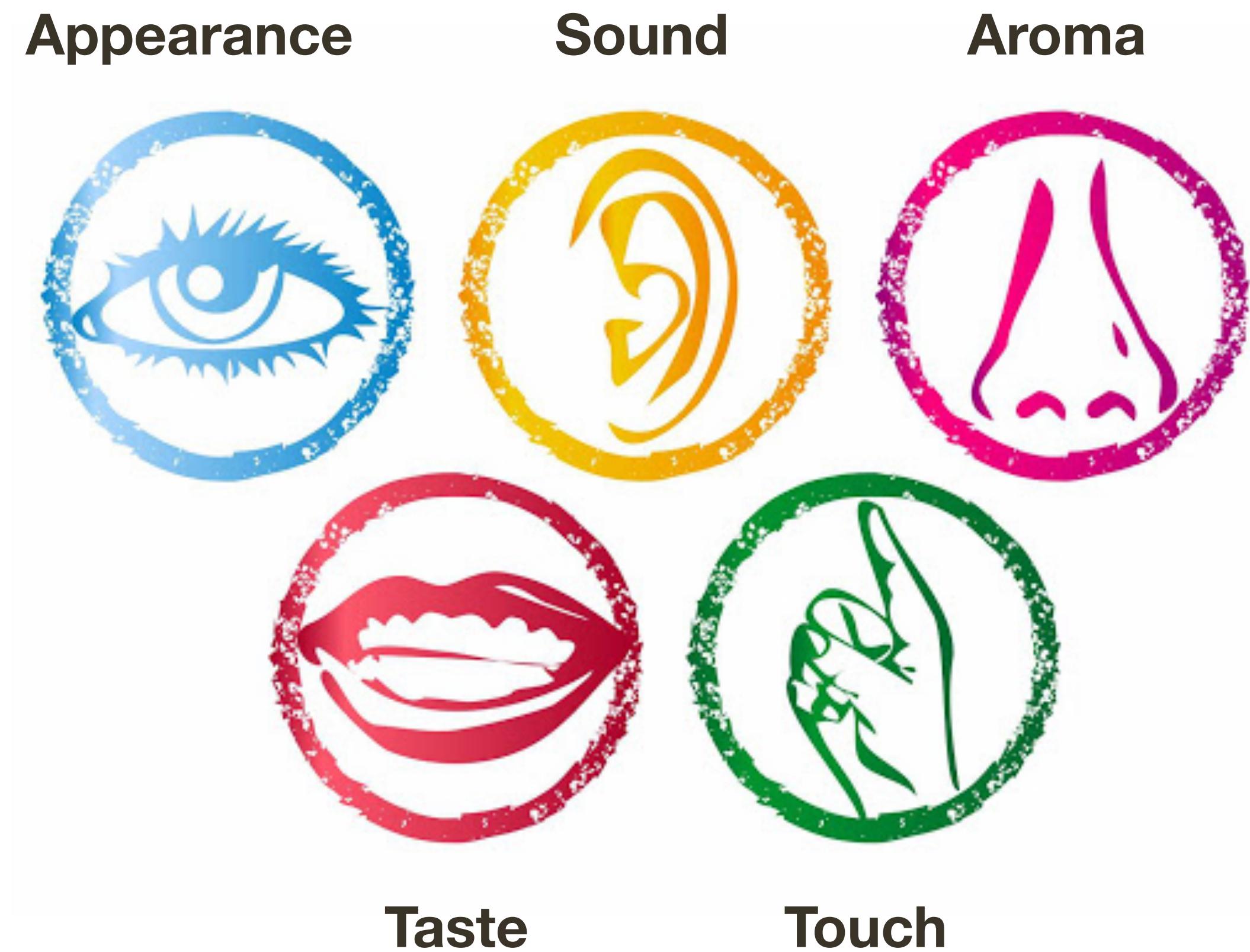




# What is Multi-modal Learning?

# What is Multi-modal Learning?

- **Sensory modality:** one or more primary channels of perception.

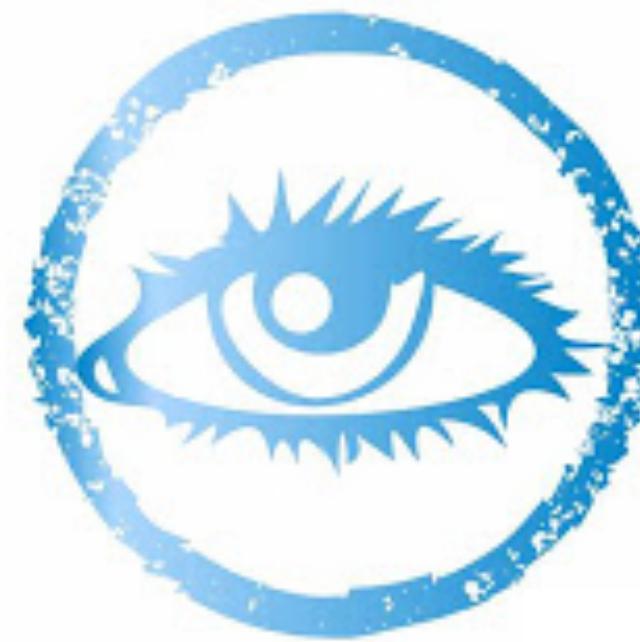


# What is Multi-modal Learning?

- **Sensory modality:** one or more primary channels of perception.

We will focus on modalities that are central to **communication**:

Appearance



Sound



# What is Multi-modal Learning?

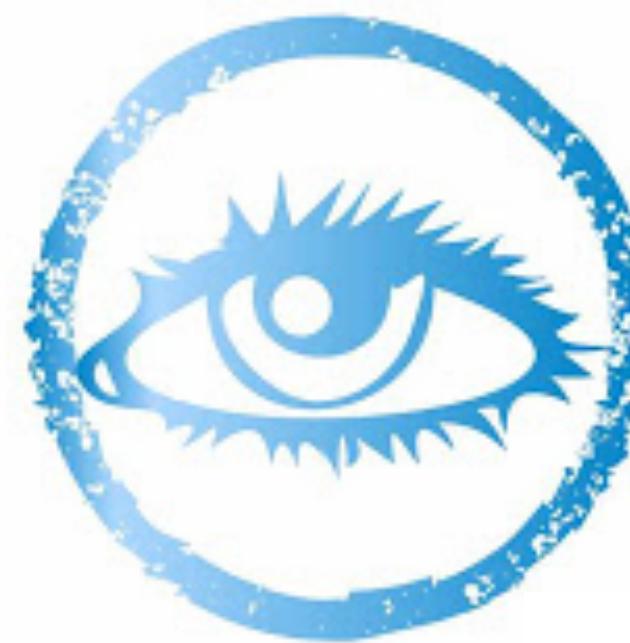
- **Sensory modality:** one or more primary channels of perception.

We will focus on modalities that are central to **communication**:

## Visual

- Perception
- Gestures
- Eye gaze

## Appearance



## Sound



## Verbal

- lexicon — words
- syntax — parts of speech
- pragmatics — discourse

## Vocal

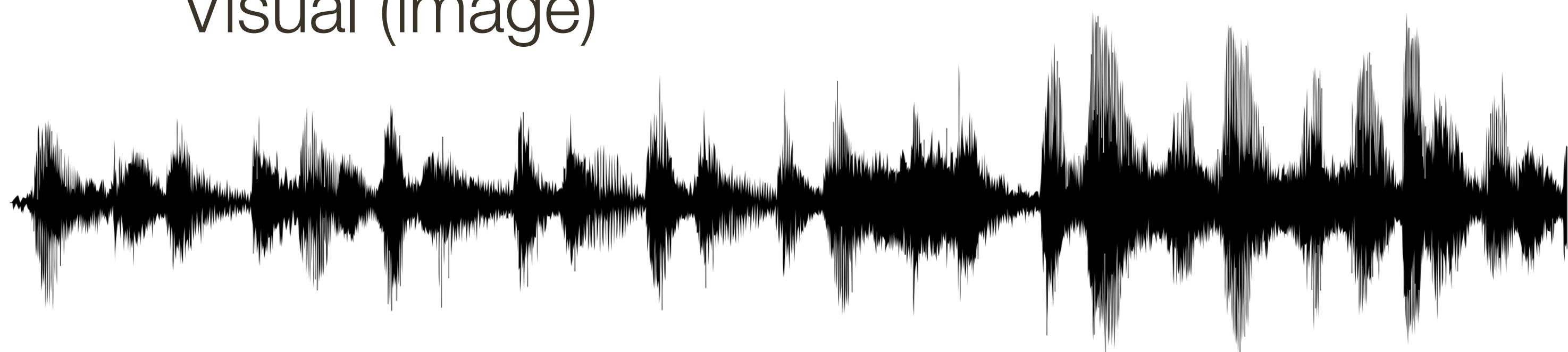
- prosody — intonation
- music — tone, pitch

# What is Multi-modal Learning?

- **Sensory modality:** one or more primary channels of perception/communication.
- **Modality:** refers to a certain type of information and/or representation format in which information is stored.

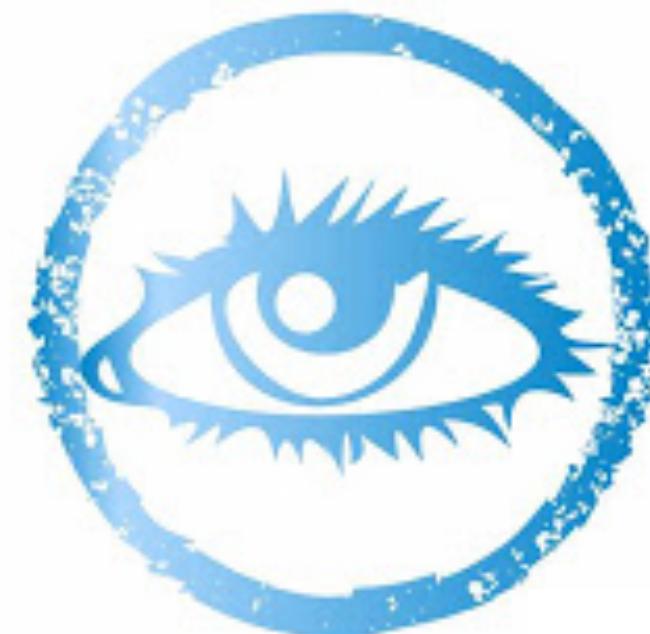


Visual (image)



Auditory (voice / sound)

Appearance



Sound



Natural Language (text)



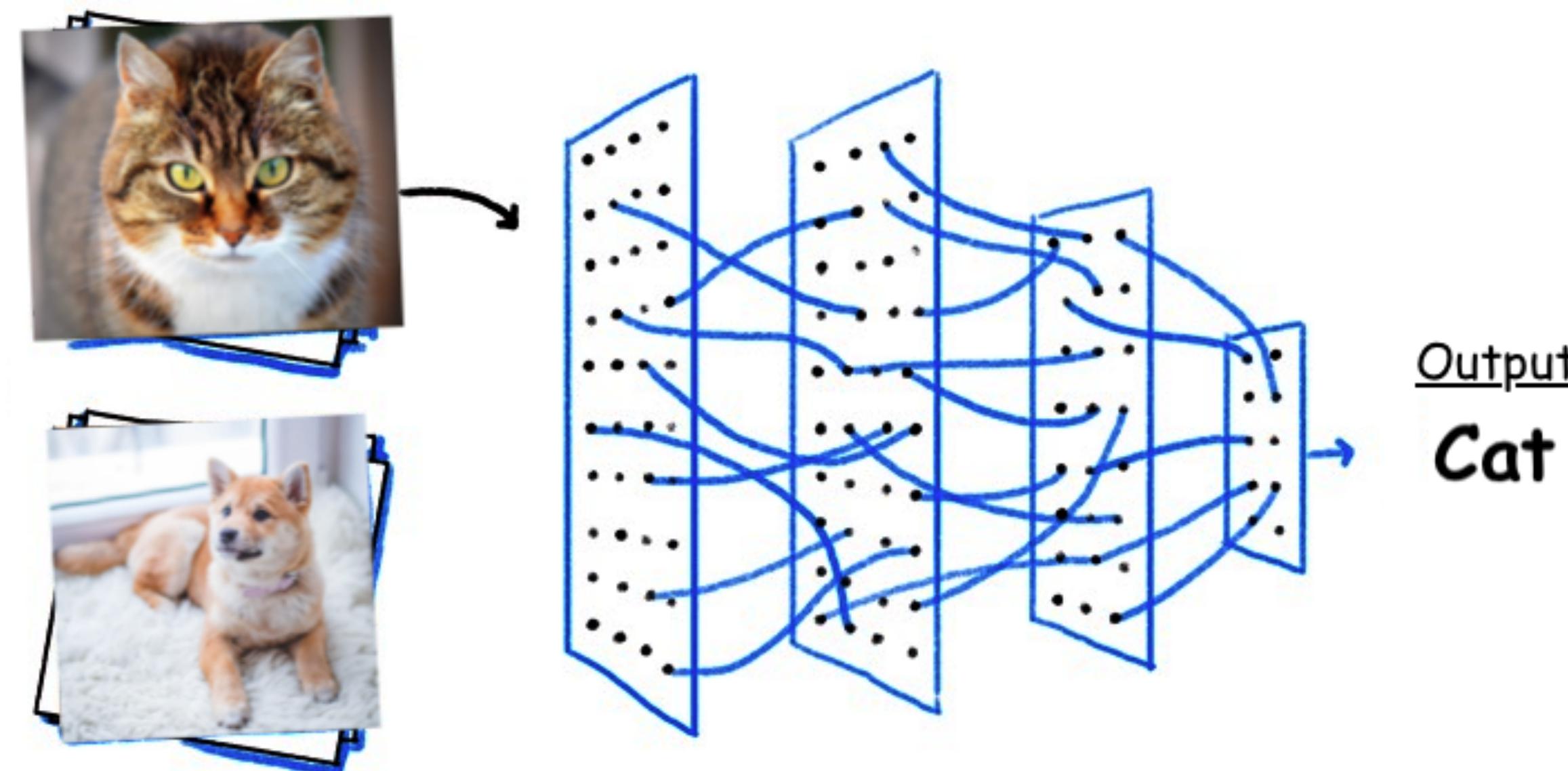
Visual (drawings)

Owl Wisdom, Omens, Vision of the night No bird has as much myth and mystery surrounding it than the owl. Part of this mystical aura is due to the fact that the bird is nocturnal and the night time has always seemed mysterious to humans. The owl is a symbol of the feminine, the moon, and the night. Because of its association with the moon it has ties to fertility and seduction. The owl is bird of magic and darkness of prophecy and wisdom.

# What is Multi-modal Learning?

**Totalitarian view:** nearly everything can be cast as multi-modal learning problem

e.g., image classification is a **translation** from **image** modality to **label** modality



# Multimodal Research: Historical Perspective

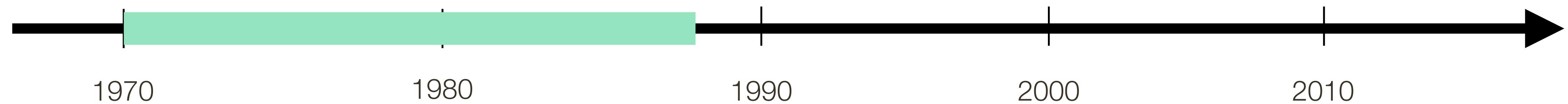
Four **eras** of multimodal research (according to Louis-Philippe (LP) Morency, CMU)

- The “**behavioral**” era (1970s until late 1980s)
- The “**computational**” era (late 1980s until 2000)
- The “**interaction**” era (2000 - 2010)
- The “**deep learning**” era (2010 until now)

# Multimodal Research: Historical Perspective

Studies of multi-sensory integration in **Psychology**

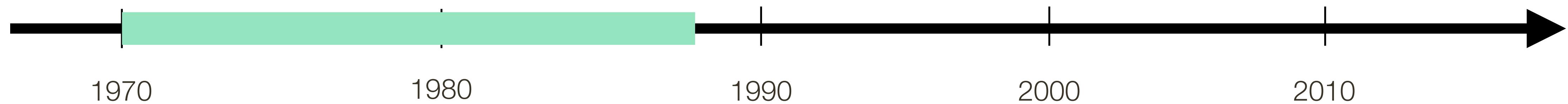
e.g., infant's perception of substance and temporal synchrony in multimodal events



# Multimodal Research: Historical Perspective

Studies of multi-sensory integration in **Psychology**

e.g., infant's perception of substance and temporal synchrony in multimodal events



# Multimodal Research: Historical Perspective

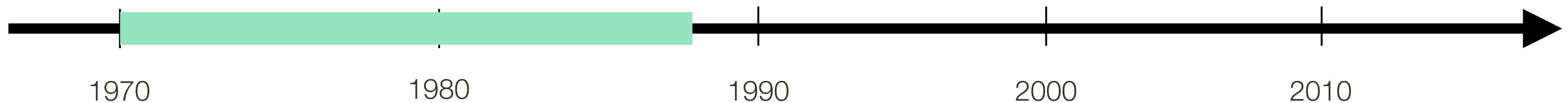
Studies of multi-sensory integration in **Psychology**

e.g., infant's perception of substance and temporal synchrony in multimodal events



**Degeneracy** – observed effect of sensory systems being able to co-educate each other without external teacher (self-supervision)

(e.g., babies randomly moving arms, until they experience a change in haptic state, grabbing something)



# Multimodal Research: Historical Perspective

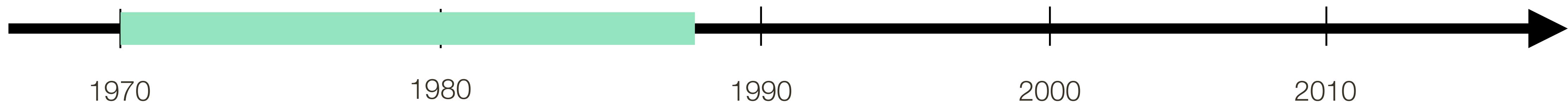
Studies of multi-sensory integration in **Psychology**

e.g., infant's perception of substance and temporal synchrony in multimodal events



**Re-entry** – ability of our brain to inter-relate multiple simultaneous representations across modalities

(e.g., when a person experiences a visual percept of an apple, other modalities are invoked, such that a person can experience its smell, taste, etc.)



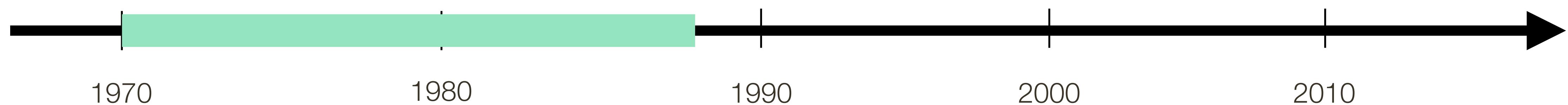
# Multimodal Research: Historical Perspective

Studies of multi-sensory integration in **Psychology**

e.g., infant's perception of substance and temporal synchrony in multimodal events



**Transparency**



# Multimodal Research: Historical Perspective

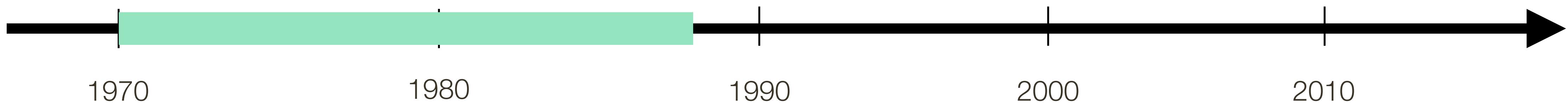
Studies of multi-sensory integration in **Psychology**

e.g., infant's perception of substance and temporal synchrony in multimodal events

## TRIVIA



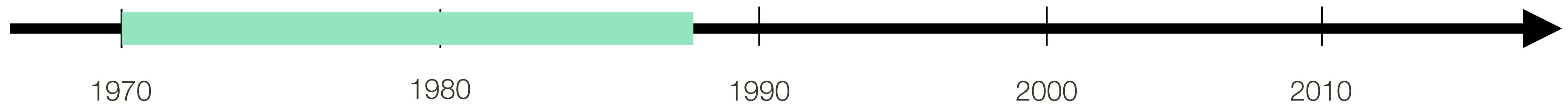
**Geoffrey Hinton** ("father of deep learning") received B.A. in Experimental Psychology from King's College in Cambridge



# Multimodal Research: Historical Perspective

## The McGurk Effect

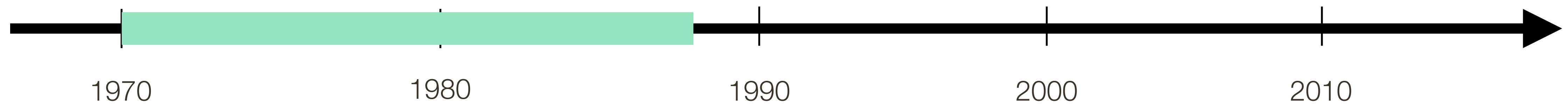
**McGurk Effect (1976)**



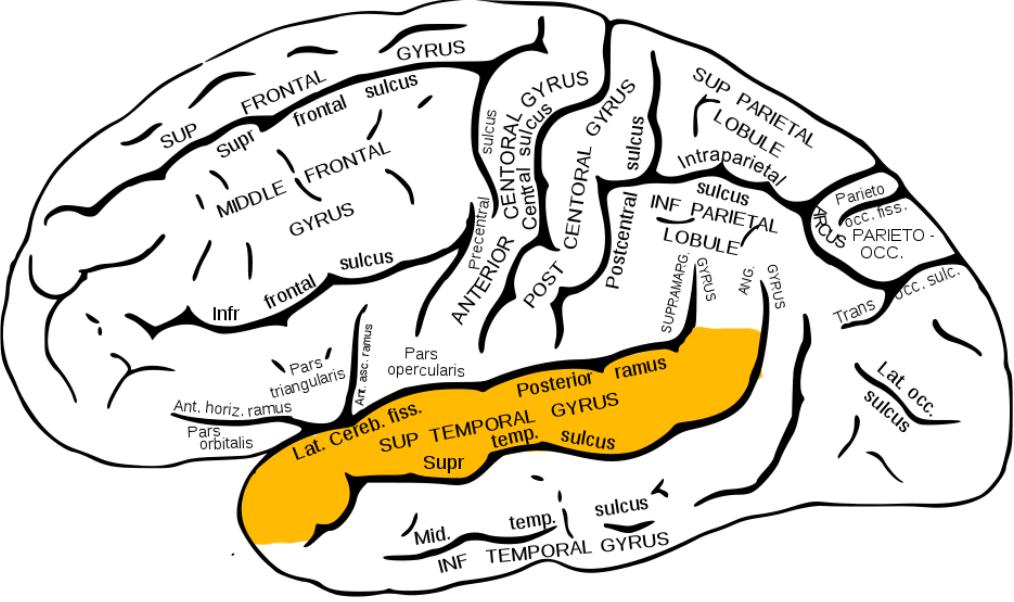
# Multimodal Research: Historical Perspective

## The McGurk Effect

**McGurk Effect (1976)**

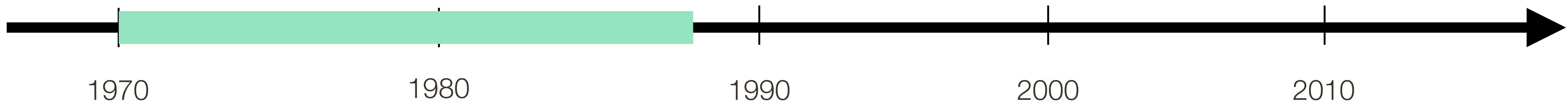


# Multimodal Research: Historical Perspective



**Superior Temporal Sulcus** is responsible for merging visual and auditory signals in the brain [Beauchamp et al. 2010].

## McGurk Effect (1976)



# Multimodal Research: Historical Perspective

Audio-visual speech recognition (motivated by McGurk effect)

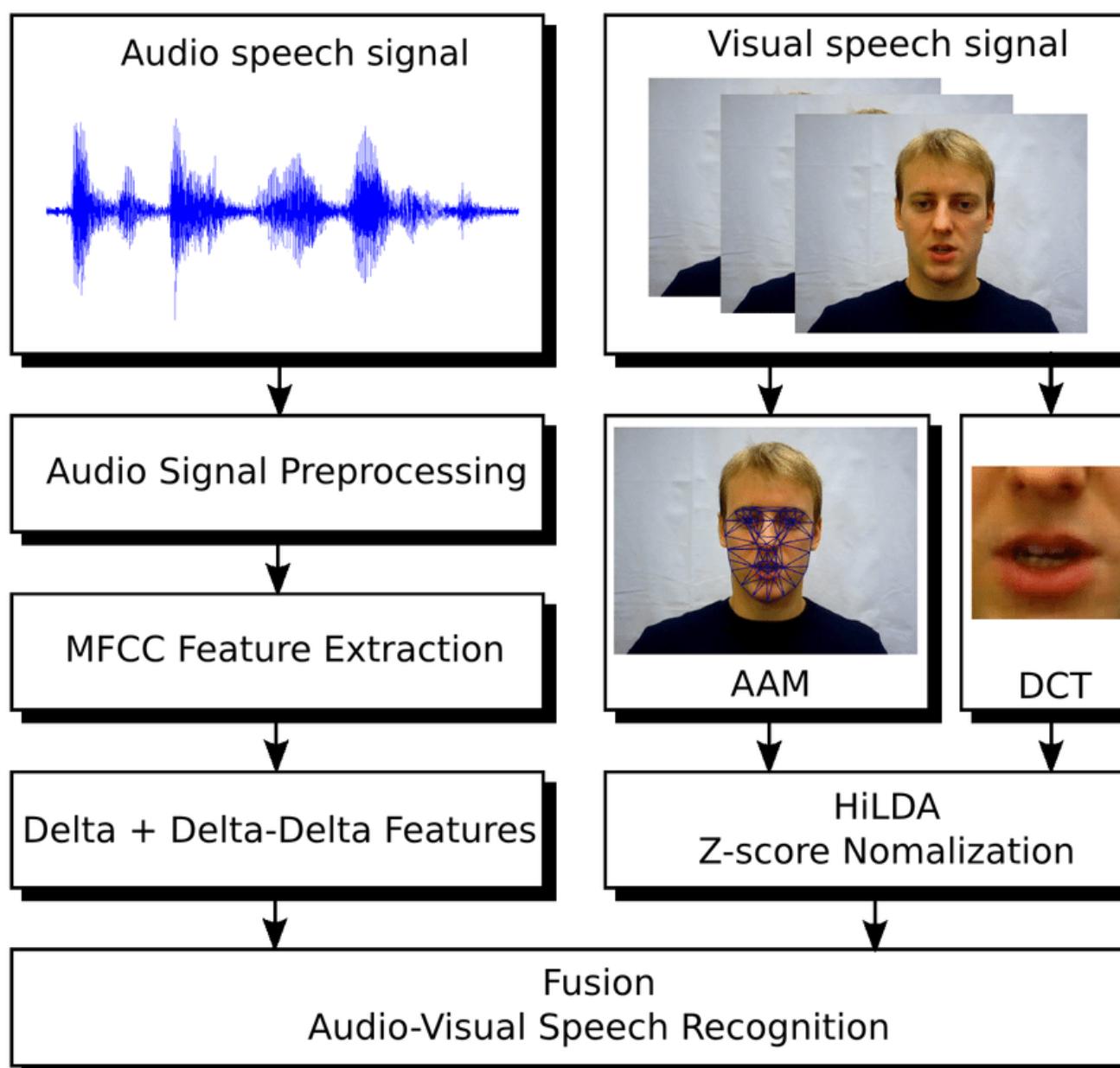
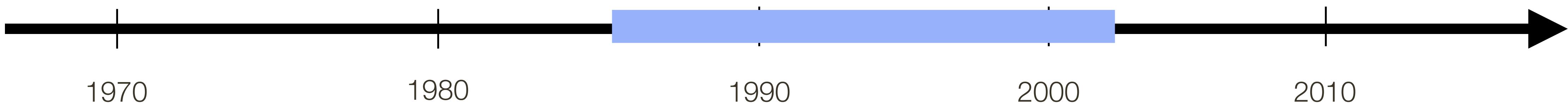
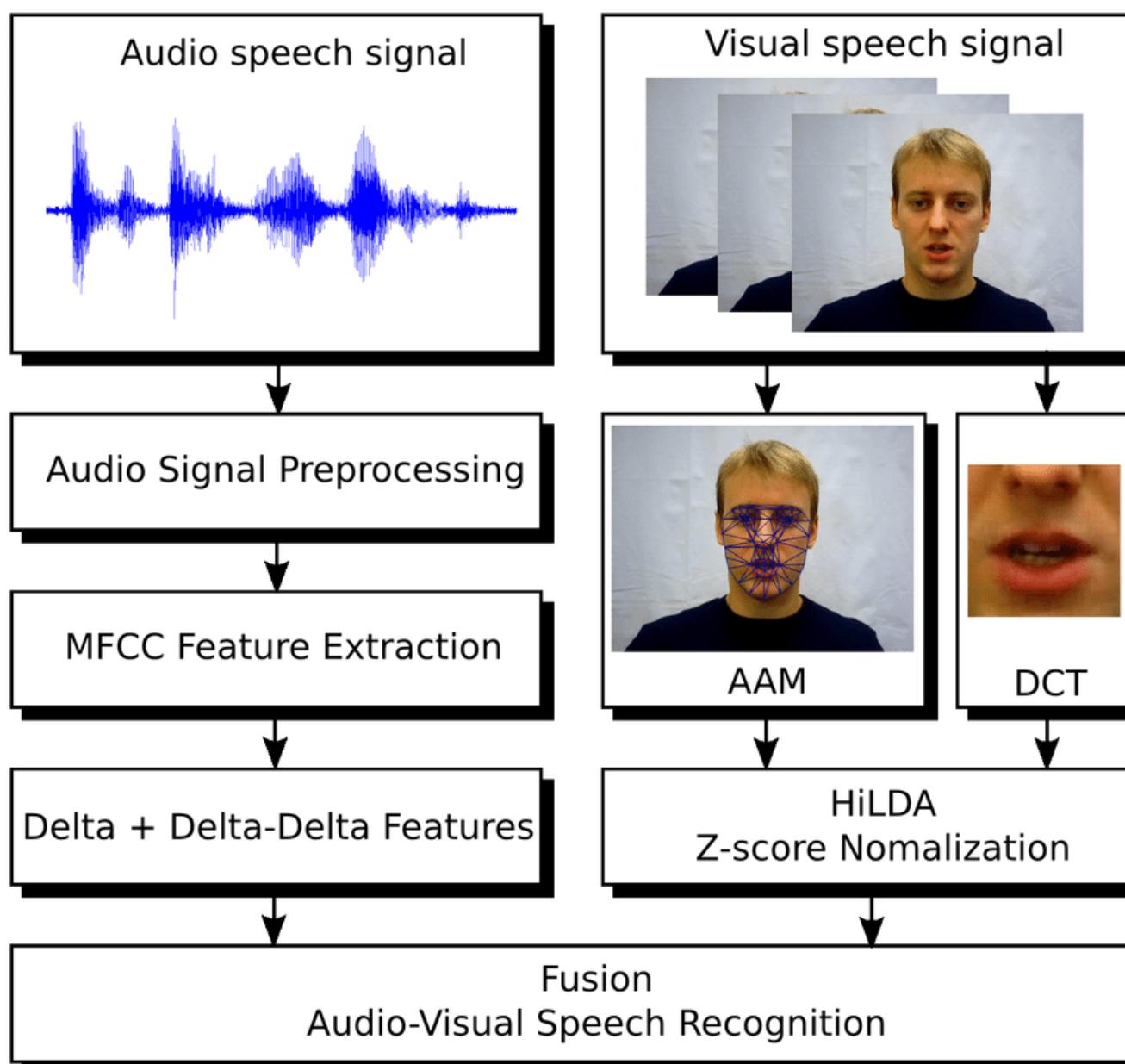


Image from: Palecek and Chaloupka



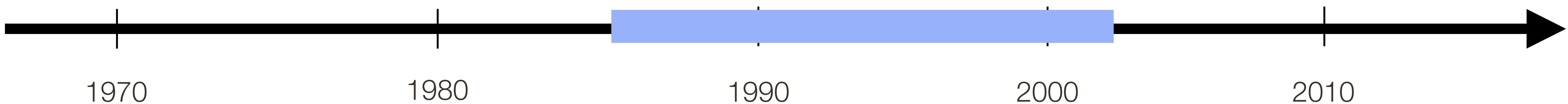
# Multimodal Research: Historical Perspective

Audio-visual speech recognition (motivated by McGurk effect)



Turns out most of the information is **redundant** rather than **complementary** (i.e., improved robustness but not overall quality of speech recognition)

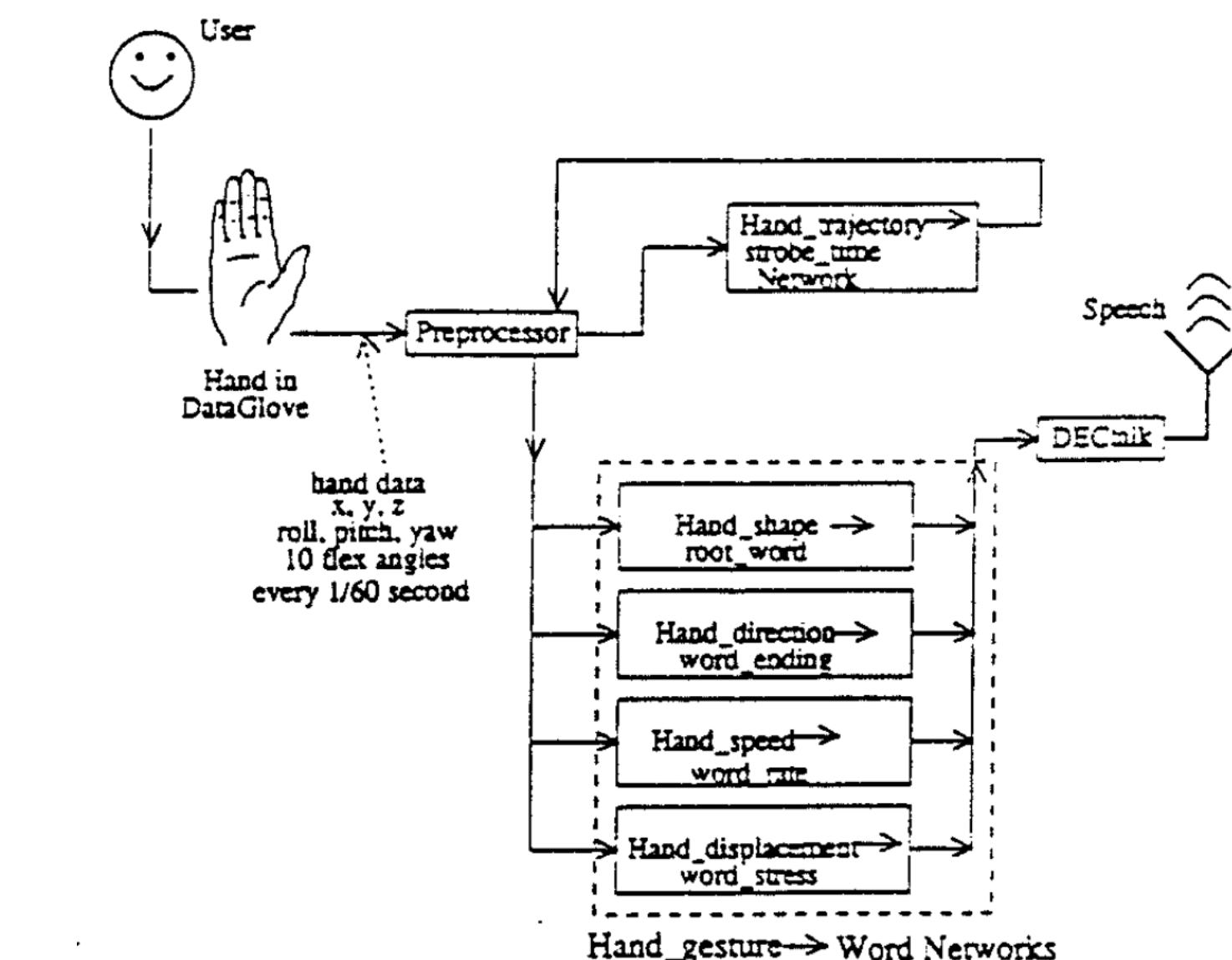
Image from: Palecek and Chaloupka



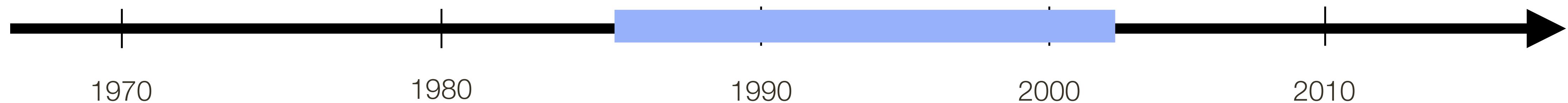
# Multimodal Research: Historical Perspective

Audio-visual speech recognition (motivated by McGurk effect)

Multi-modal and multi-sensory interfaces



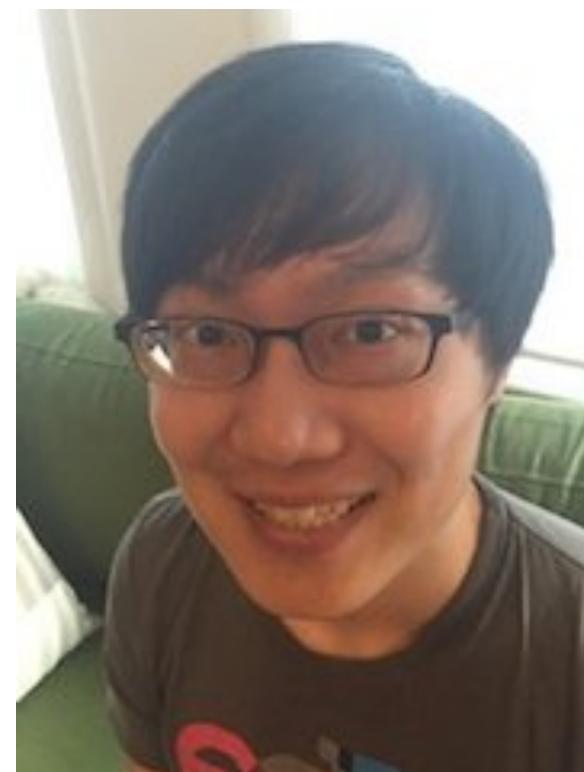
**GloveTalk** by S. Fels and G. Hinton [CHI'95]



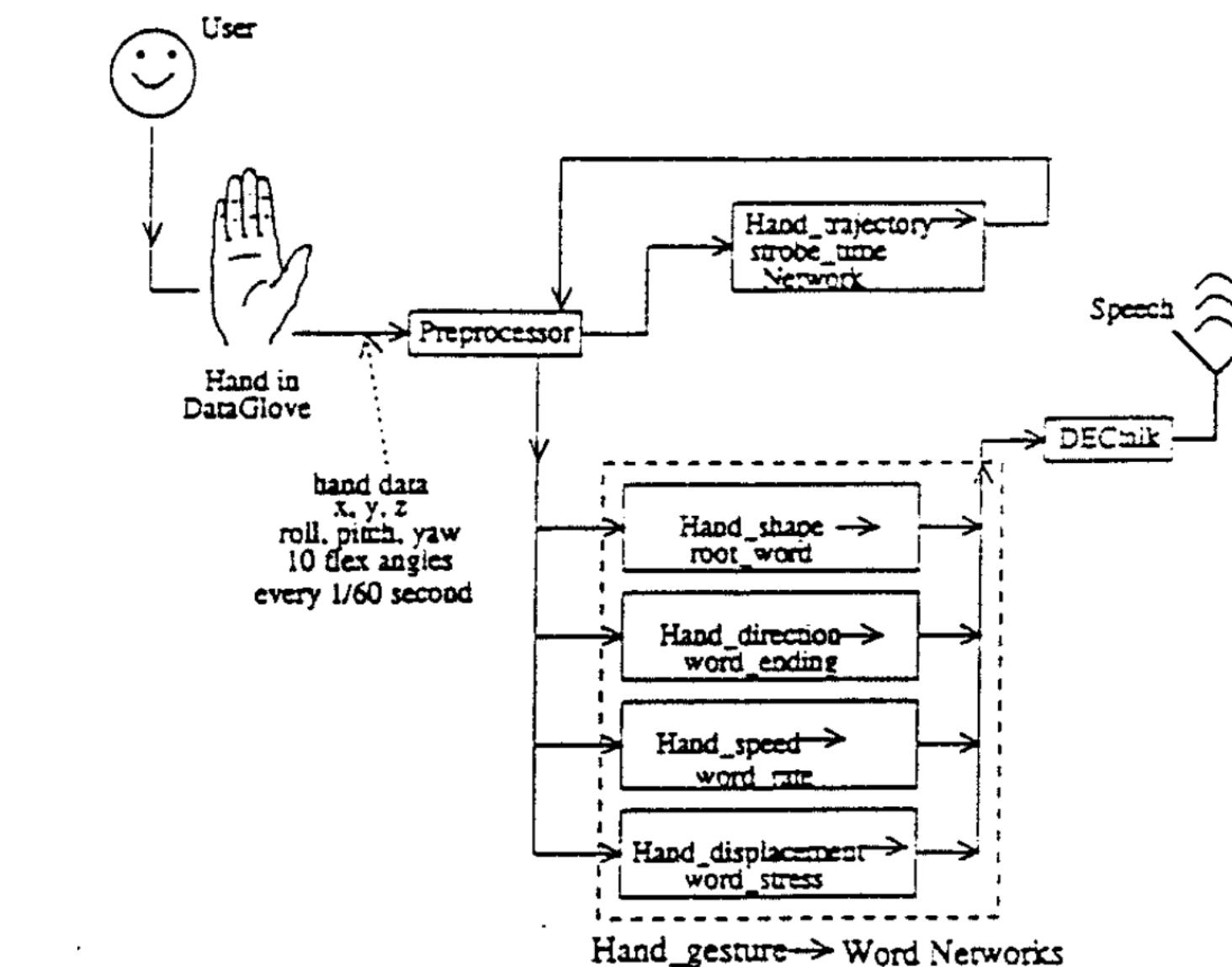
# Multimodal Research: Historical Perspective

Audio-visual speech recognition (motivated by McGurk effect)

Multi-modal and multi-sensory interfaces



Dongwook Yoon



GloveTalk by S. Fels and G. Hinton [CHI'95]



# Multimodal Research: Historical Perspective

Audio-visual speech recognition (motivated by McGurk effect)

Multi-modal and multi-sensory interfaces

Affective Computing



Rosalind Picard



# Multimodal Research: Historical Perspective

## Modeling human multi-modal interactions

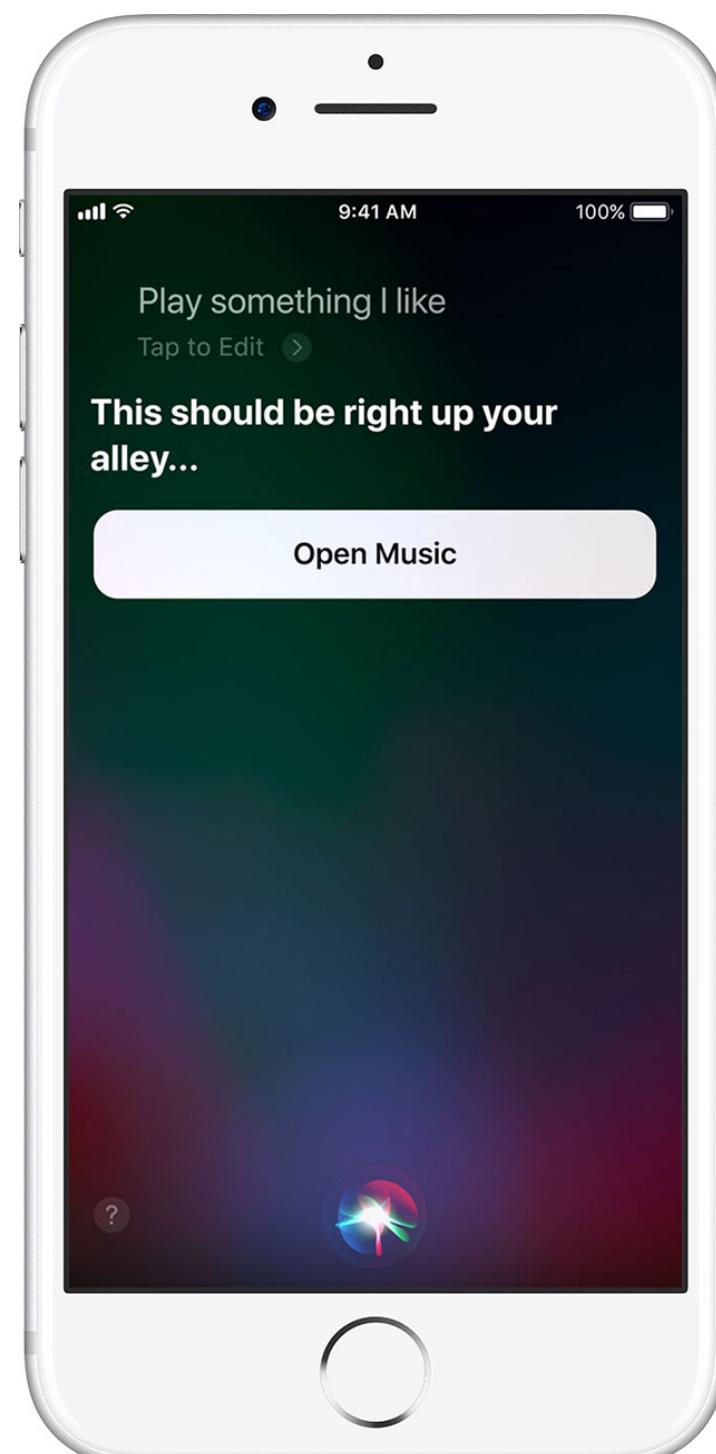
- Huge multi-laboratory efforts

### **AMI Project** [2001-2006, IDIAP]

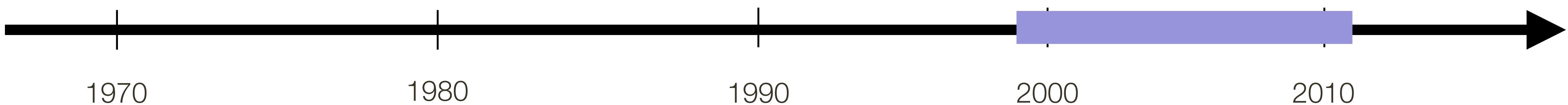
- 100+ hours of meeting recordings
- Synchronized video and audio
- Transcribed and annotated

### **CALO Project** [2003-2008, SRI]

- Cognitive assistant that learns and organizes
- Personalized assistant that learns



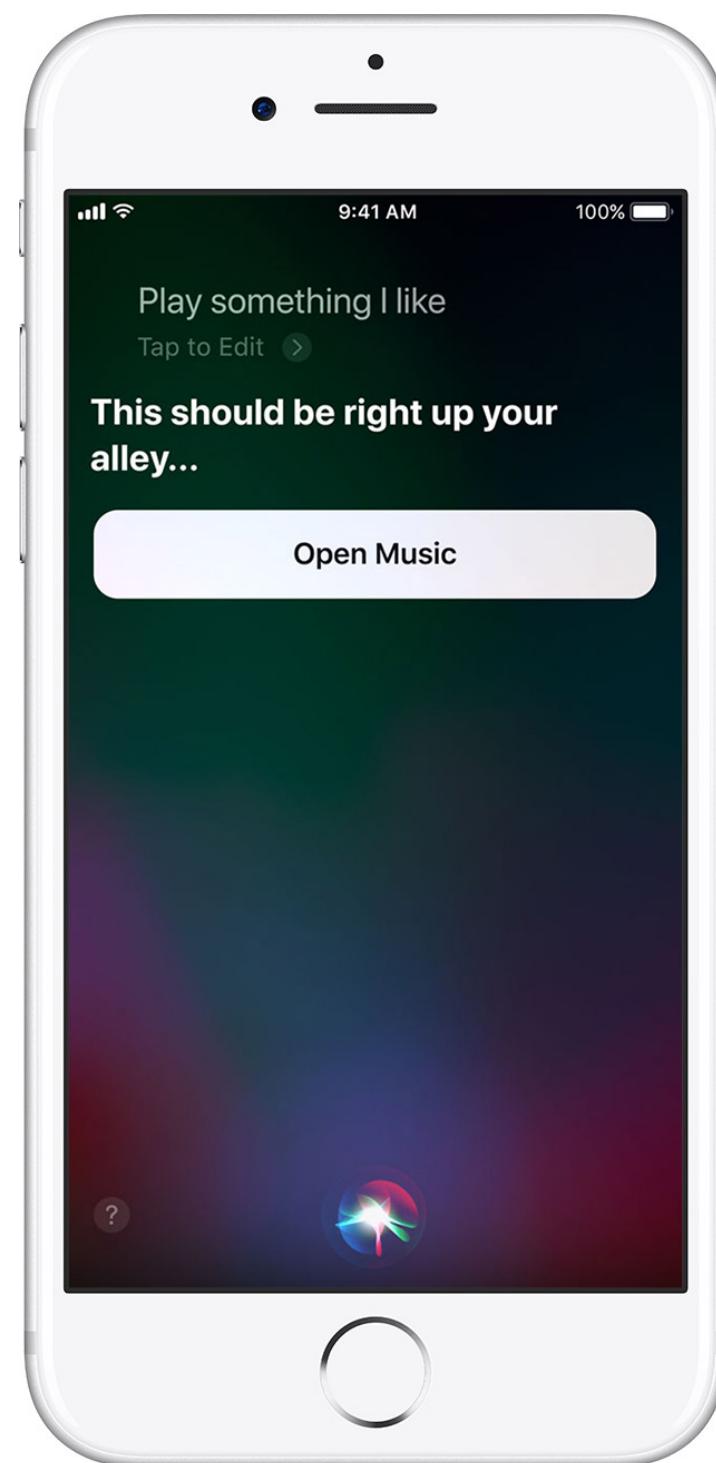
**Siri** was spun as an output of multi-modal interaction projects



# Multimodal Research: Historical Perspective

## Modeling human multi-modal interactions

- Huge multi-laboratory efforts



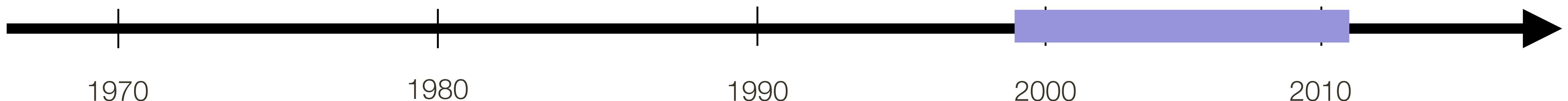
## Multimedia information retrieval

- Lots of challenges and progress

### Research Tasks and Challenges:

- Shot boundary detection, story segmentation, search
- Semantic event, character and object detection

**Siri** was spun as an output of multi-modal interaction projects

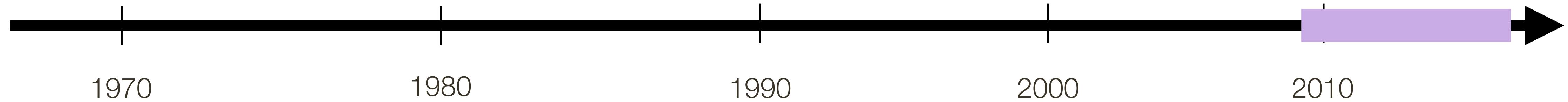


# Multimodal Research: Historical Perspective

## Deep Learning (a.k.a. representation learning)

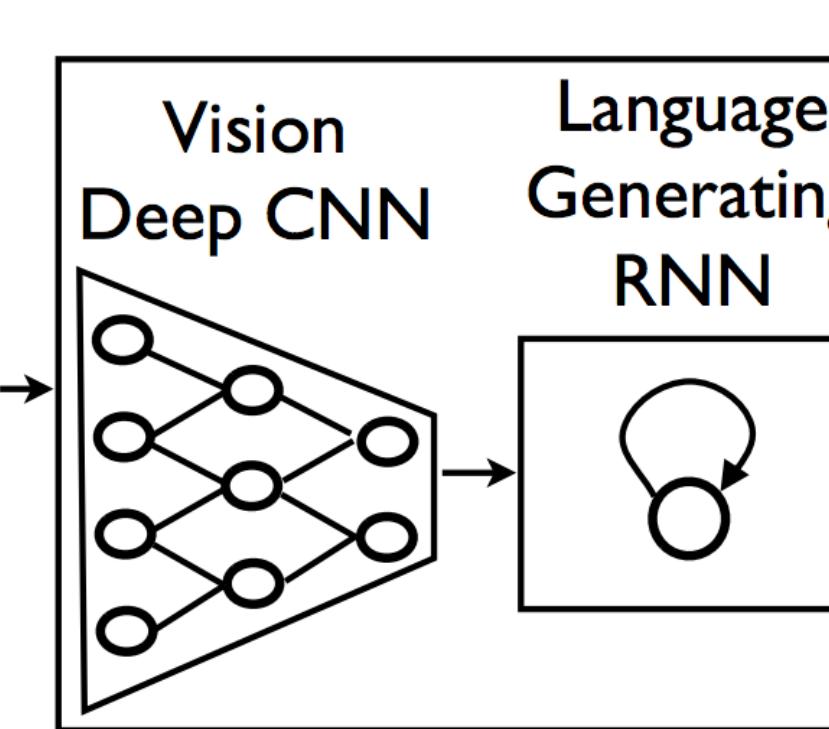
- Better performance
- More interesting problems emerging

# THIS IS OUR COURSE



# Multimodal Research: Historical Perspective

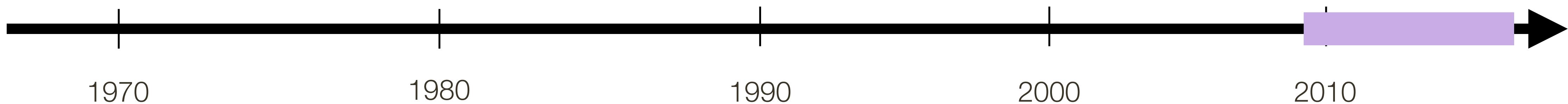
## Deep Learning (a.k.a. representation learning)



**A group of people shopping at an outdoor market.**  
**There are many vegetables at the fruit stand.**

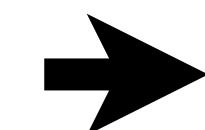
[ Vinyals et al., 2015 ]

Natural language description generation



# Multimodal Research: Historical Perspective

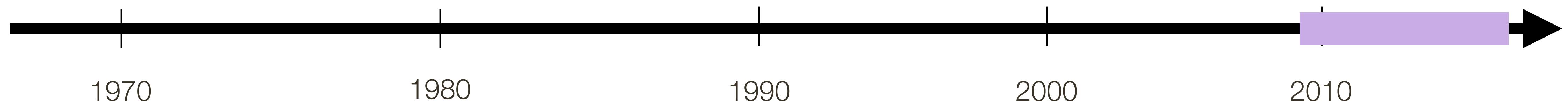
## Deep Learning (a.k.a. representation learning)



A few miles before tioga road reached highway 395 and the town of lee vining, smith turned onto a narrow blacktop road. On either side were parched, grassy open slopes with barbed-wire fences marking property lines. Cattle and horses grazed under trees whose black silhouettes stood stark against the gold-velvet mountains. Marty burst into song: “ home , home on the range, where the deer and the antelope play! Where seldom is heard a discouraging word and the skies are not cloudy all day!”

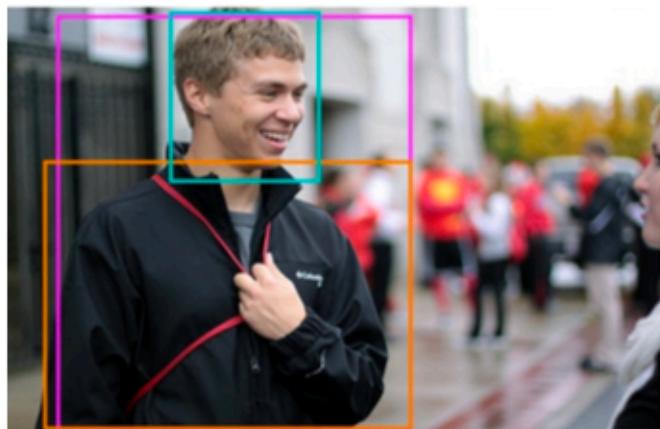
[ Zhu et al, ICCV 2015 ]

Story generation

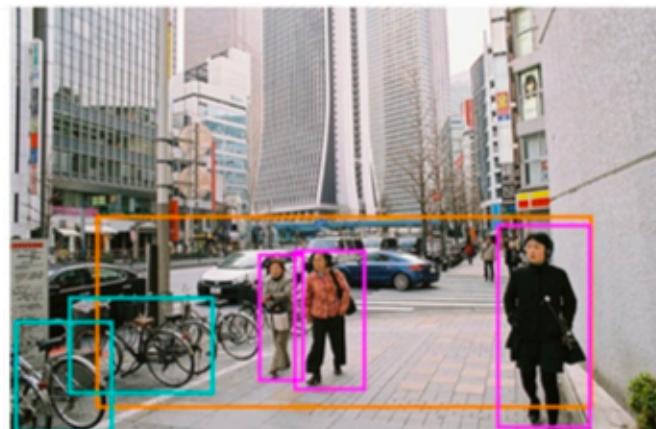


# Multimodal Research: Historical Perspective

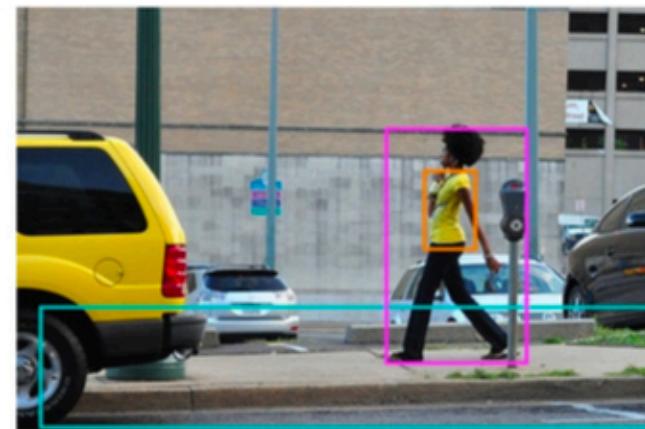
## Deep Learning (a.k.a. representation learning)



(a) A man wearing a black-jacket has a smile on his face.



(b) People are walking on the street, with bikes parked up to the left of the picture.



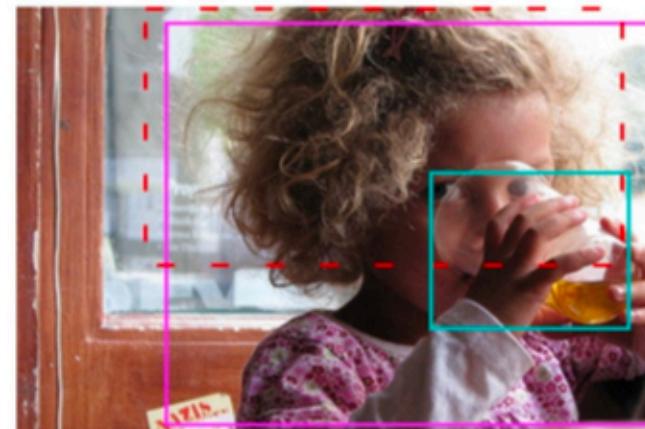
(c) A woman in a yellow shirt is walking down the sidewalk.



(f) Two women in colorful clothing are dancing inside a circle of other women.



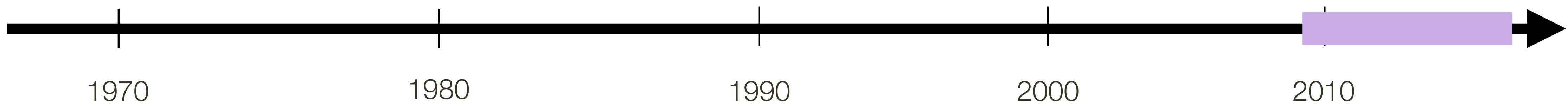
(g) Lady wearing white shirt with blue umbrella in the rain.



(h) Young girl with curly hair is drinking out of a plastic cup.

[ Bajaj *et al.*, ICCV 2019 ]

Language grounding



# Multimodal Research: Historical Perspective

## Deep Learning (a.k.a. representation learning)

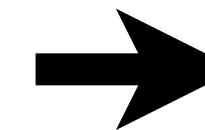
### Corn Poppy

*Papaver rhoeas* (common names include corn poppy, corn rose, field poppy, Flanders poppy, red poppy, red weed, coquelicot, and, due to its odour, which is said to cause them, as headache and headwark) is a species of flowering plant in the poppy family, Papaveraceae. This poppy, a native of Europe, is notable as an agricultural weed (hence the "corn" and "field") and as a symbol of fallen soldiers.

*P. rhoeas* sometimes is so abundant in agricultural fields that it may be mistaken for a crop. The only species of Papaveraceae grown as a field crop on a large scale is *Papaver somniferum*, the opium poppy.

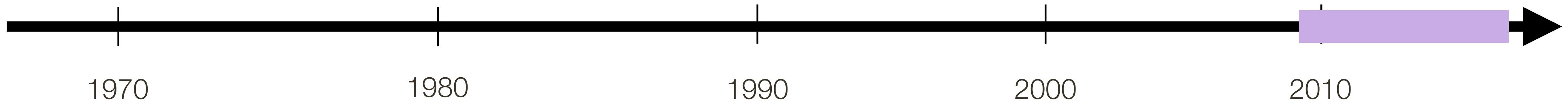
The plant is a variable annual, forming a long-lived soil seed bank that can germinate when the soil is disturbed. In the northern hemisphere it generally flowers in late spring, but if the weather is warm enough other flowers frequently appear at the beginning of autumn. The flower is large and showy, with four petals that are vivid red, most commonly with a black spot at their base. Like many other species of *Papaver*, it exudes a white latex when the tissues are broken.

.....



[ Ba et al., ICCV 2015 ]

Detecting objects based on linguistic descriptions



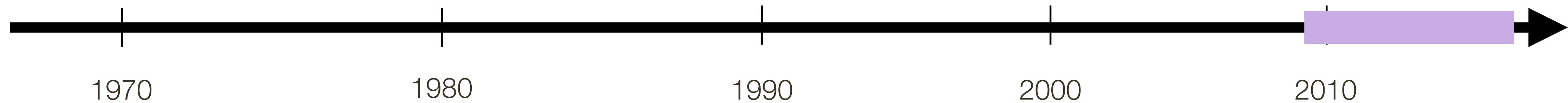
# Multimodal Research: Historical Perspective

**Deep Learning** (a.k.a. representation learning)



[ Zhu et al, ICCV 2015 ]

Book-to-Movie alignment



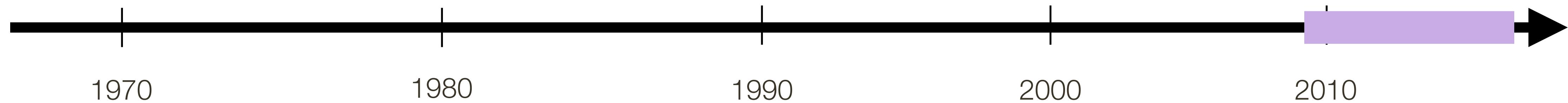
# Multimodal Research: Historical Perspective

**Deep Learning** (a.k.a. representation learning)



[ Zhu et al, ICCV 2015 ]

Book-to-Movie alignment



# Multimodal Research: Historical Perspective

## Deep Learning (a.k.a. representation learning)

*... you have to line up by **tower of terror** or **buena vista street** (park entrance) but these are faithful recreations of **old time trollys** and a very relaxing ride....*

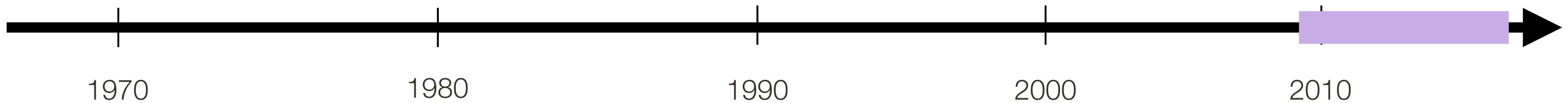
*... go over to the smaller bridge-to the left of the main/big one that leads into the **pacific wharf**-you...*

[TripAdvisor.com]



[ Kim & Sigal, CVPR 2015 ]

Auto illustration



# Multimodal Research: Historical Perspective

## Deep Learning (a.k.a. representation learning)

**Q1:** *What color is the bowl ?*

GT answer: *White*

Predicted answer: *White*

Rank of GT: *1*



[ Seo et al., NIPS 2017 ]

**Q2:** *Do you see any people?*

GT answer: *No*

Predicted answer: *No, just the cat*

Rank of GT: *2*



**Q3:** *What color is the cat ?*

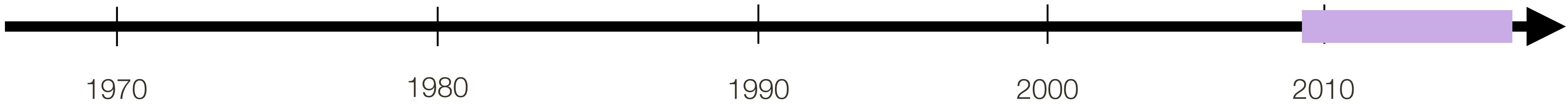
GT answer: *Grey, white, and black*

Predicted answer: *Grey, black and white*

Rank of GT: *6*

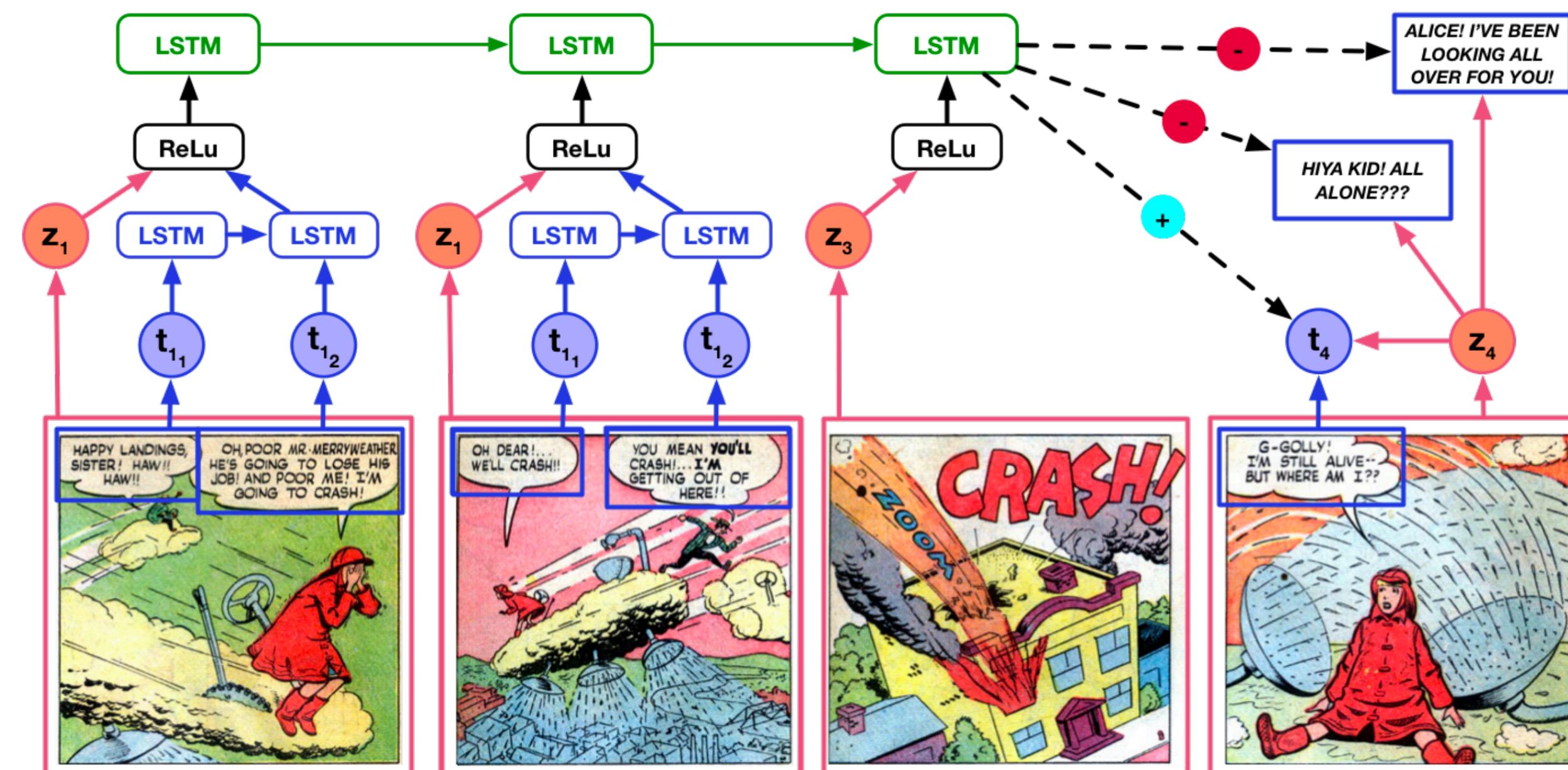


Visual question answering / dialog



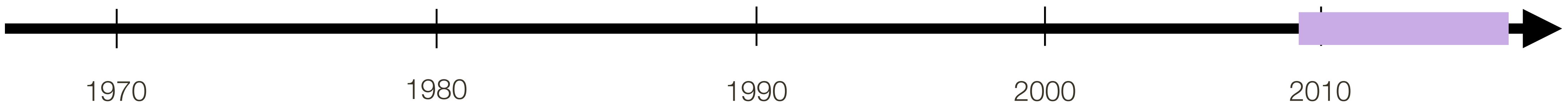
# Multimodal Research: Historical Perspective

## Deep Learning (a.k.a. representation learning)



[ Iyyer et al., CVPR 2017 ]

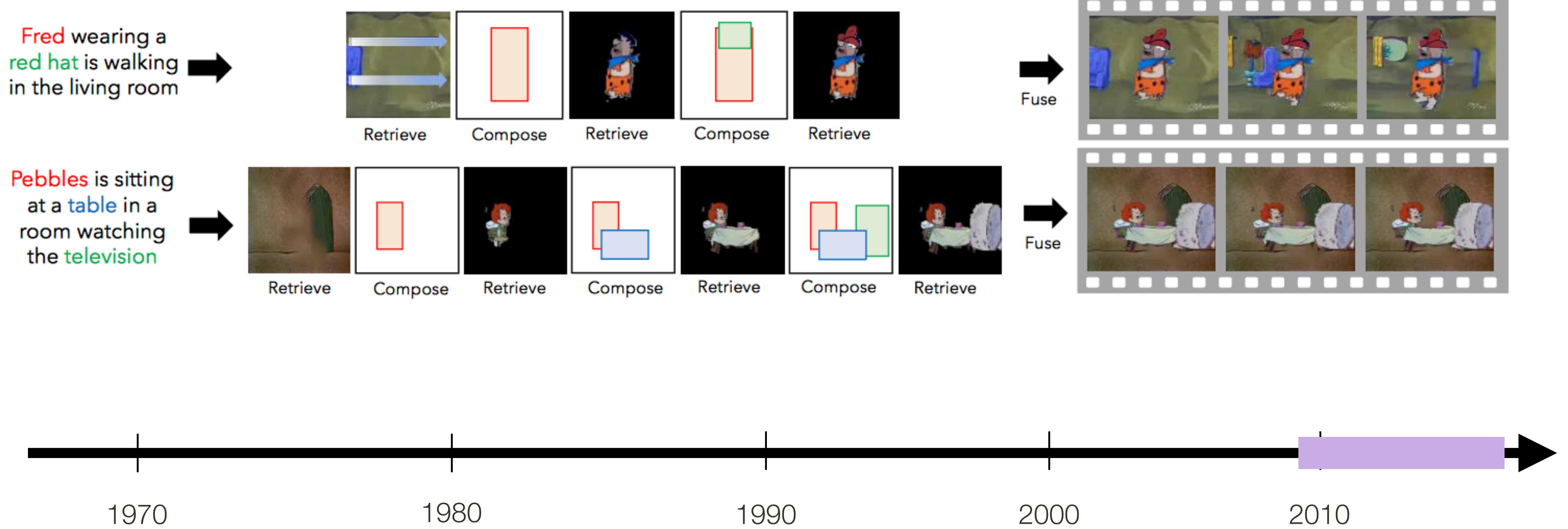
Narrative plot understanding



# Multimodal Research: Historical Perspective

## Deep Learning (a.k.a. representation learning)

[ Gupta et al., ECCV 2018 ]



# Multimodal Research: Historical Perspective

## Deep Learning (a.k.a. representation learning)

TEXT DESCRIPTION

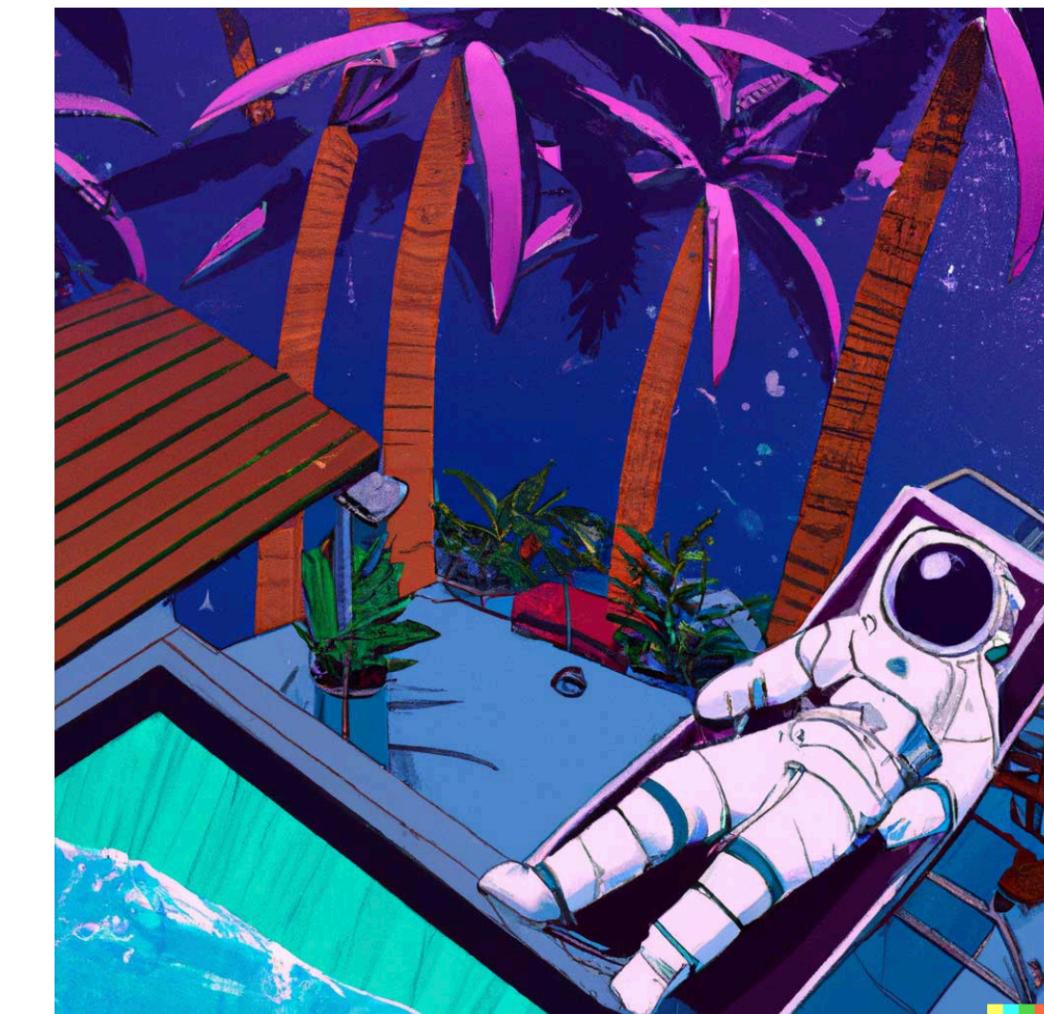
An astronaut Teddy bears A bowl of soup

riding a horse lounging in a tropical resort in space playing basketball with cats in space

in a vaporwave style as pixel art in a photorealistic style

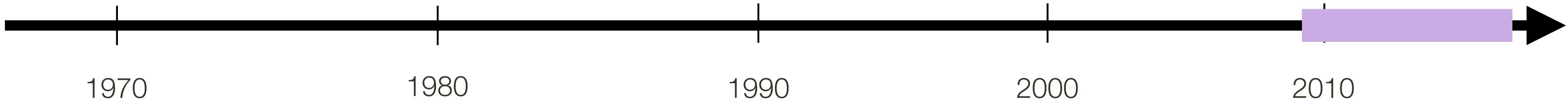
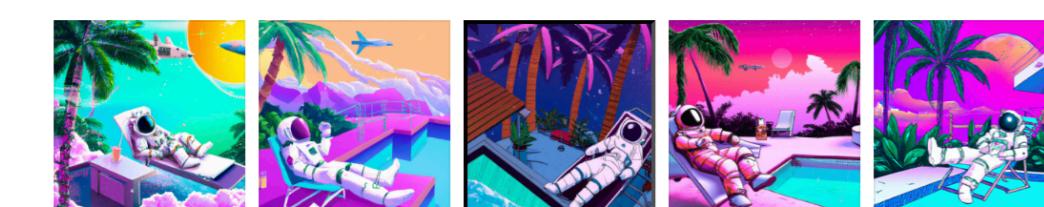


DALL-E 2



[ Ramesh et al., CVPR 2022 ]

<https://openai.com/dall-e-2/>



# Multimodal Research: Historical Perspective

## Deep Learning (a.k.a. representation learning)

[ Zhu et al., ICCV 2017 ]

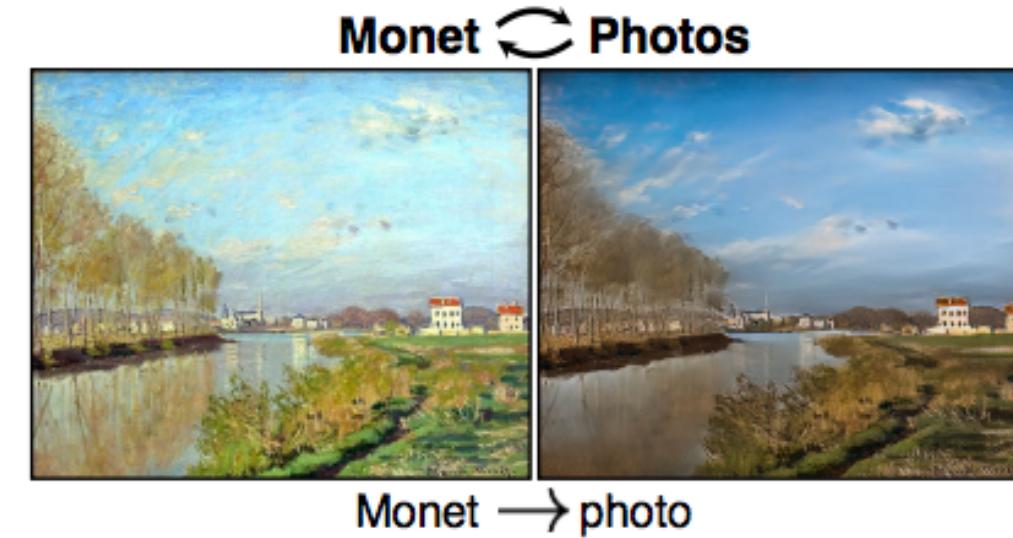
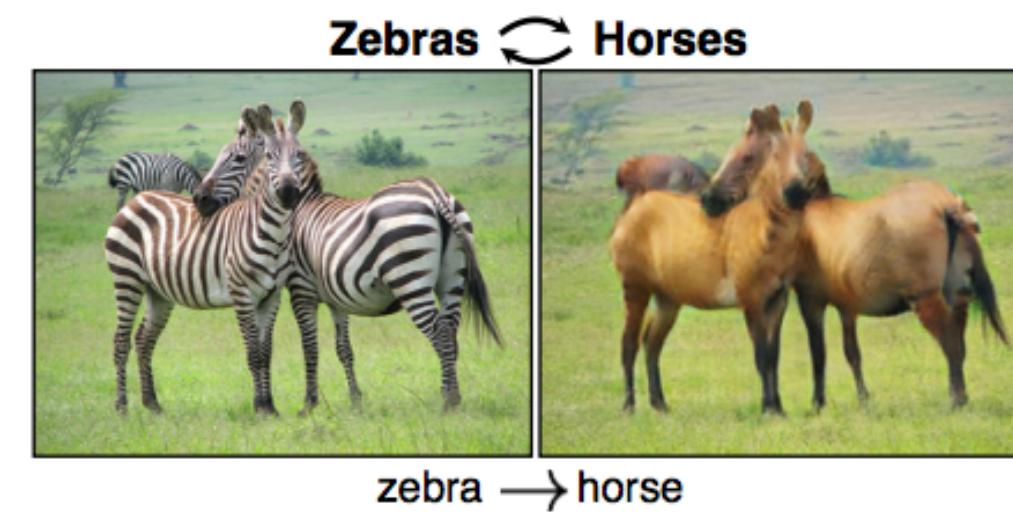
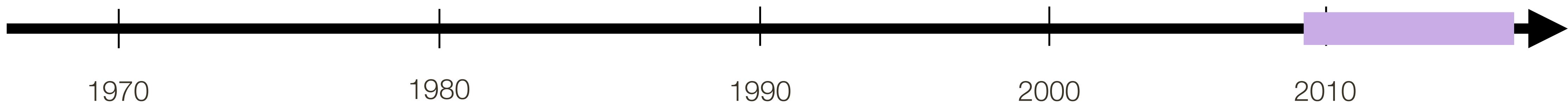


Image-to-image translation

Edges to Photo

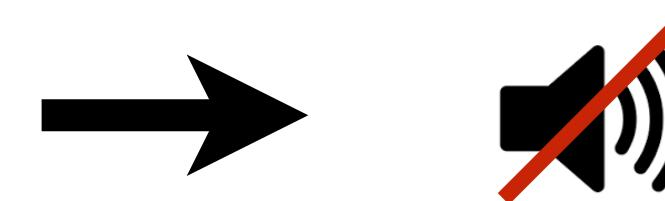


[ Isola et al., CVPR 2017 ]



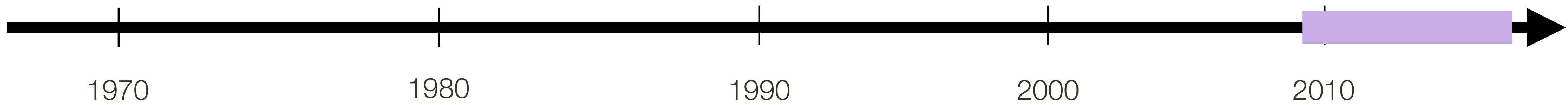
# Multimodal Research: Historical Perspective

**Deep Learning** (a.k.a. representation learning)



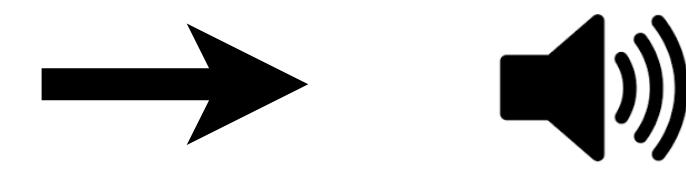
[ Iyyer *et al.*, NIPS 2016 ]

Video-to-Audio translation



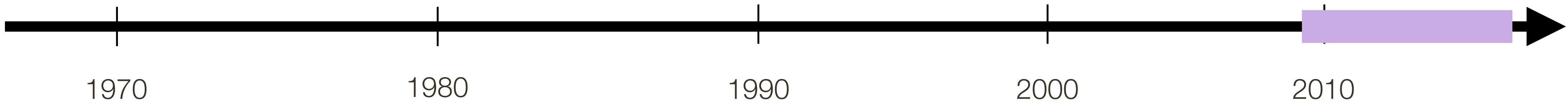
# Multimodal Research: Historical Perspective

**Deep Learning** (a.k.a. representation learning)



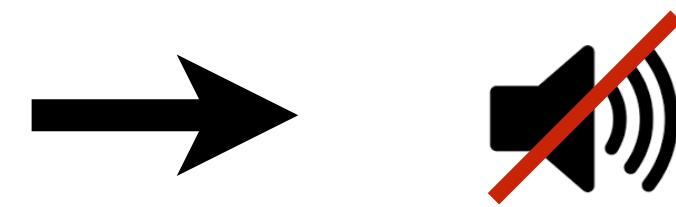
[ Iyyer *et al.*, NIPS 2016 ]

Video-to-Audio translation



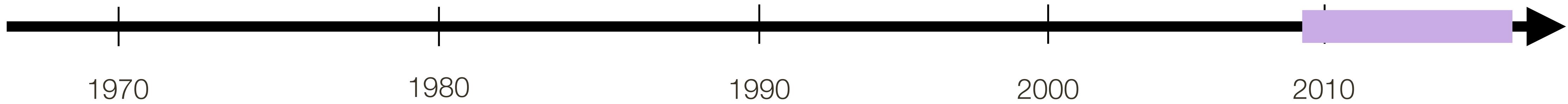
# Multimodal Research: Historical Perspective

**Deep Learning** (a.k.a. representation learning)



[ Iyyer *et al.*, NIPS 2016 ]

Video-to-Audio translation



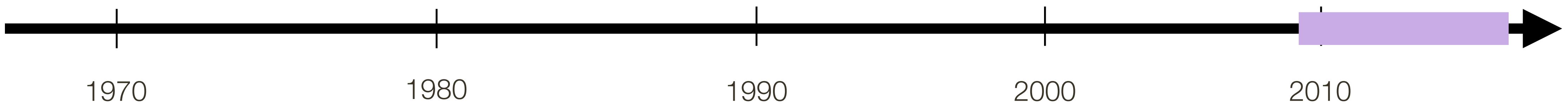
# Multimodal Research: Historical Perspective

## Deep Learning (a.k.a. representation learning)

				
LECTURING	TIPPING	JUMPING	CLIMBING	ENCOURAGING
ROLE	ROLE	ROLE	ROLE	ROLE
AGENT	AGENT	AGENT	AGENT	AGENT
AUDIENCE	PEOPLE	WOMEN	MAN	DAD
PLACE	BOAT	LAND	ROCK	SON
ITEM	WEIGHT	-	ROPE	LAWN
AGENT PART	PLACE	LAND	PLACE	PLACE
CLASSROOM	LAKE	OUTDOORS	JUNGLE	

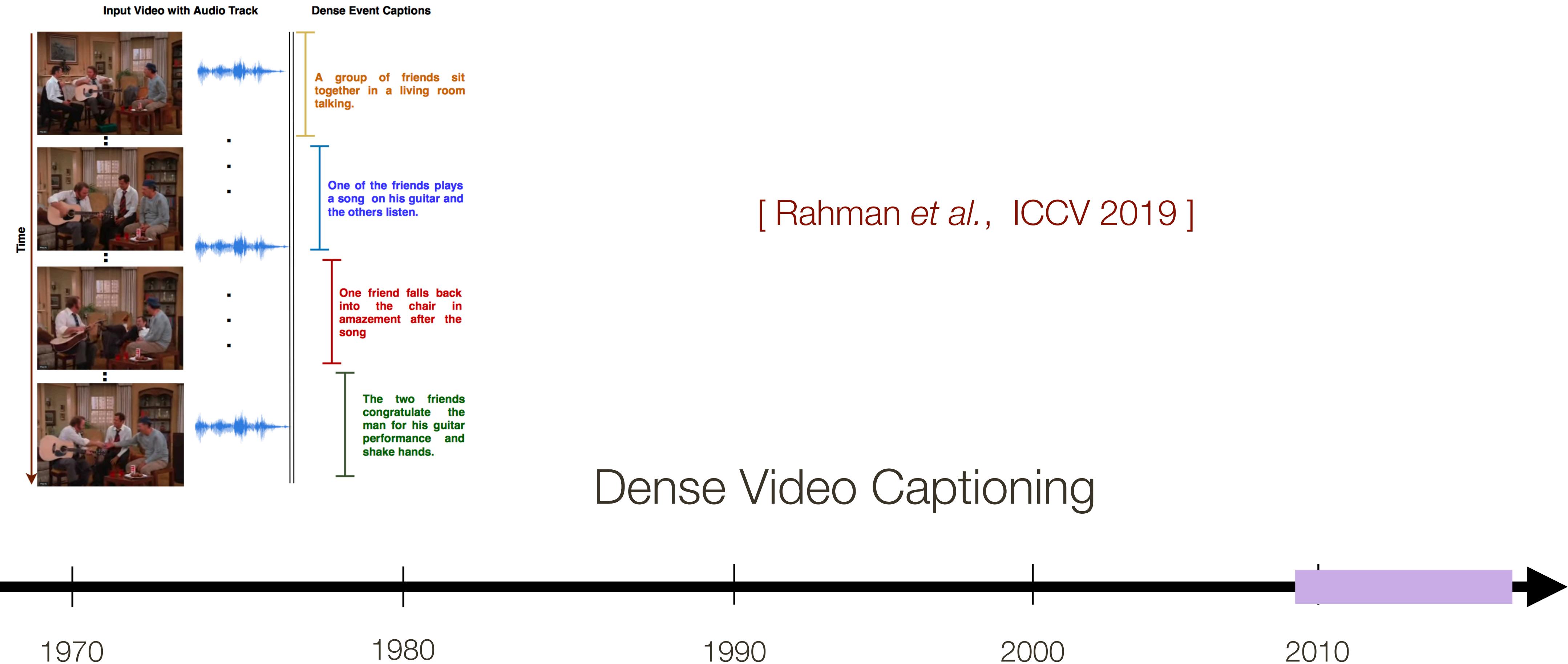
[ Suhail et al., ICCV 2019 ]

## Situational Recognition



# Multimodal Research: Historical Perspective

## Deep Learning (a.k.a. representation learning)



# **Key Challenges** of Multimodal Learning

- Representation learning in each and across modalities
- Alignment between representations in different modalities
- Translation between modalities

# Key Challenges of Multimodal Learning

- Representation learning in each and across modalities
- Alignment between representations in different modalities
- Translation between modalities

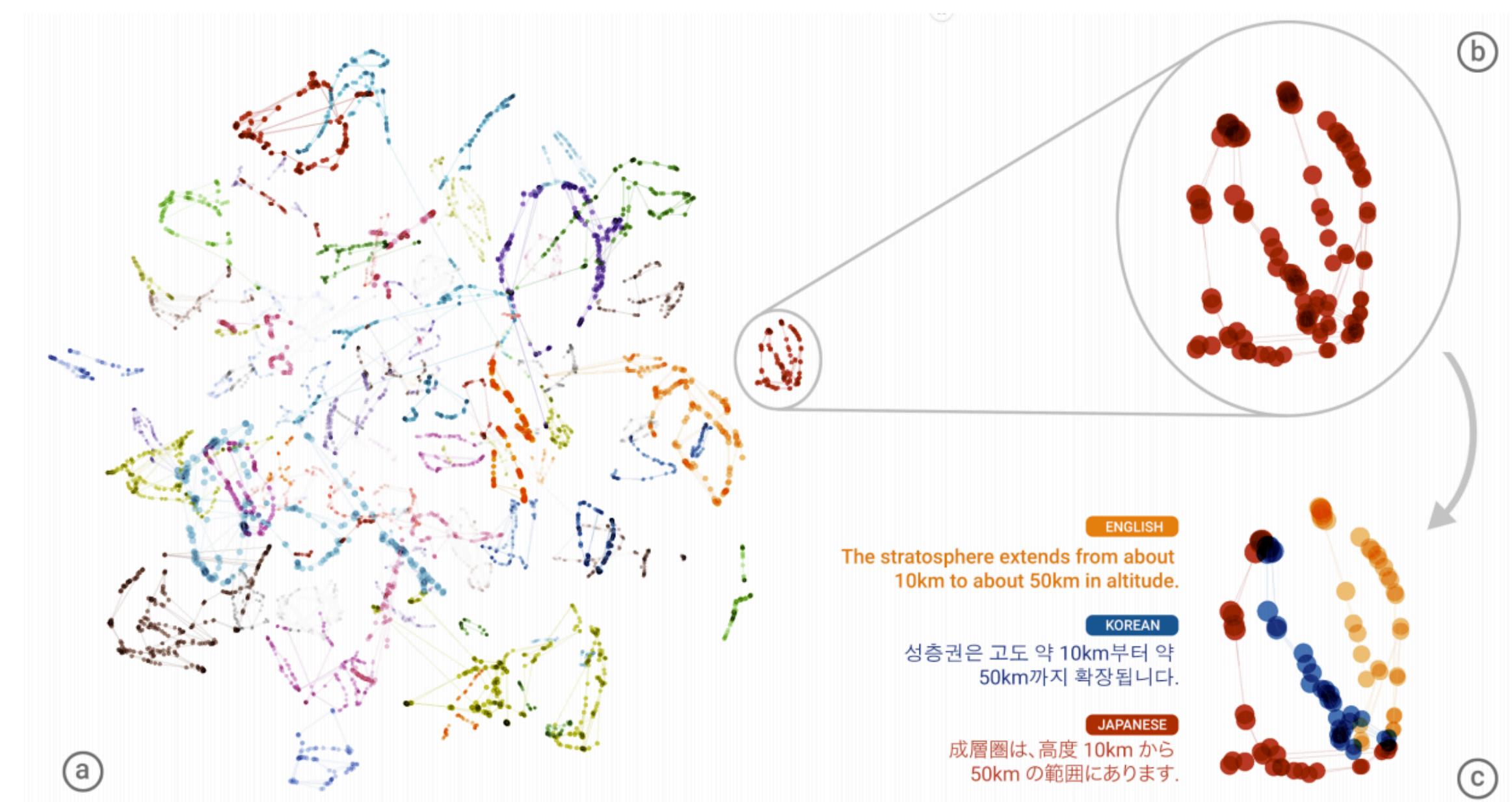
What's another phrase for "**representation learning**"?

# Key Challenges of Multimodal Learning

- Representation learning in each and across modalities
- Alignment between representations in different modalities
- Translation between modalities

One translation model learned across many languages, actually improves the performance in translation over direct training on:

English -> German  
German -> English  
French -> English



[ Johnson et al., ArXiv 2017 from Google ]

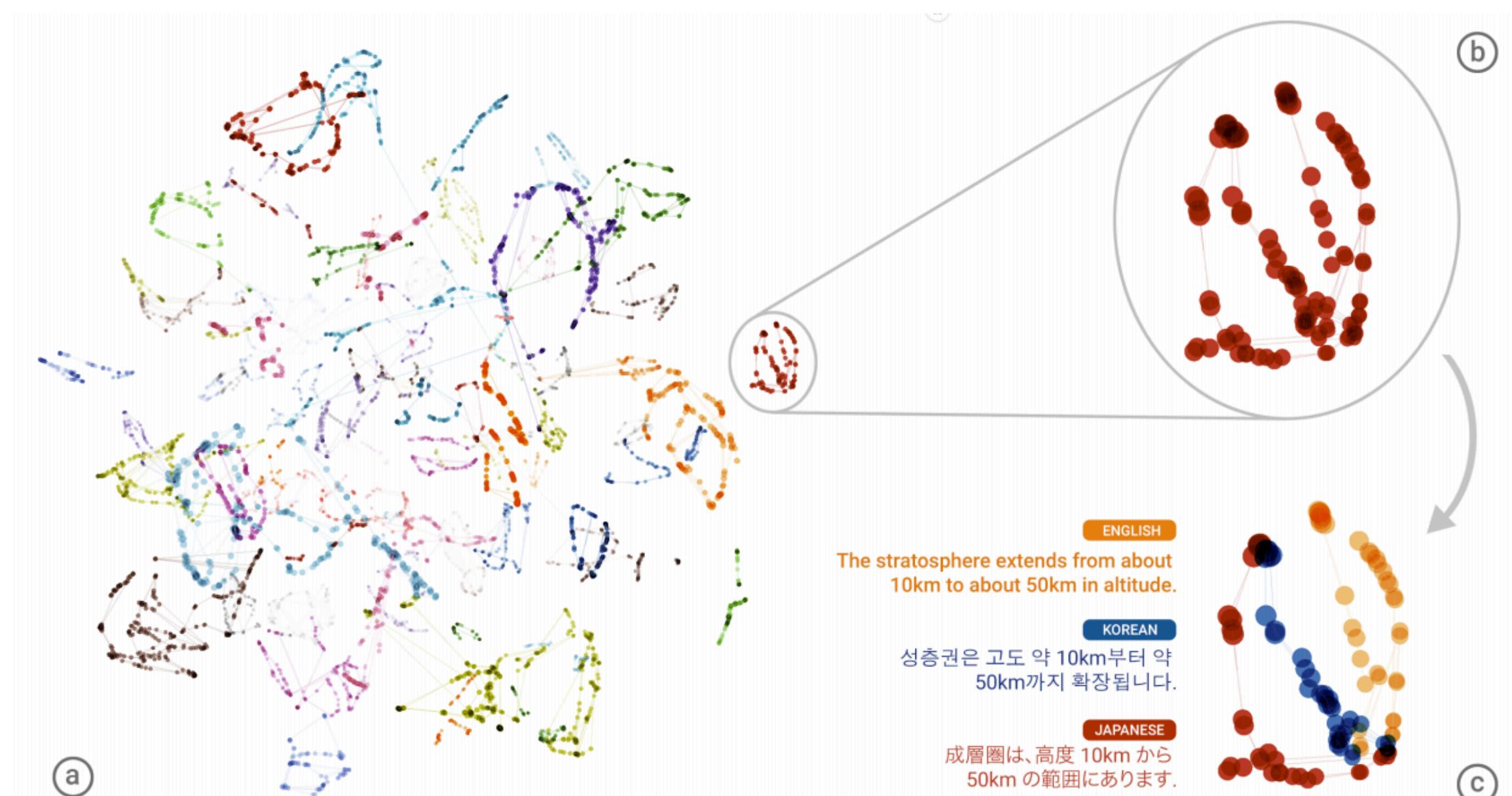
# Key Challenges of Multimodal Learning

- Representation learning in each and across modalities
- Alignment between representations in different modalities
- Translation between modalities

One translation model learned across many languages, actually improves the performance in translation over direct training on:

English -> German  
German -> English  
French -> English

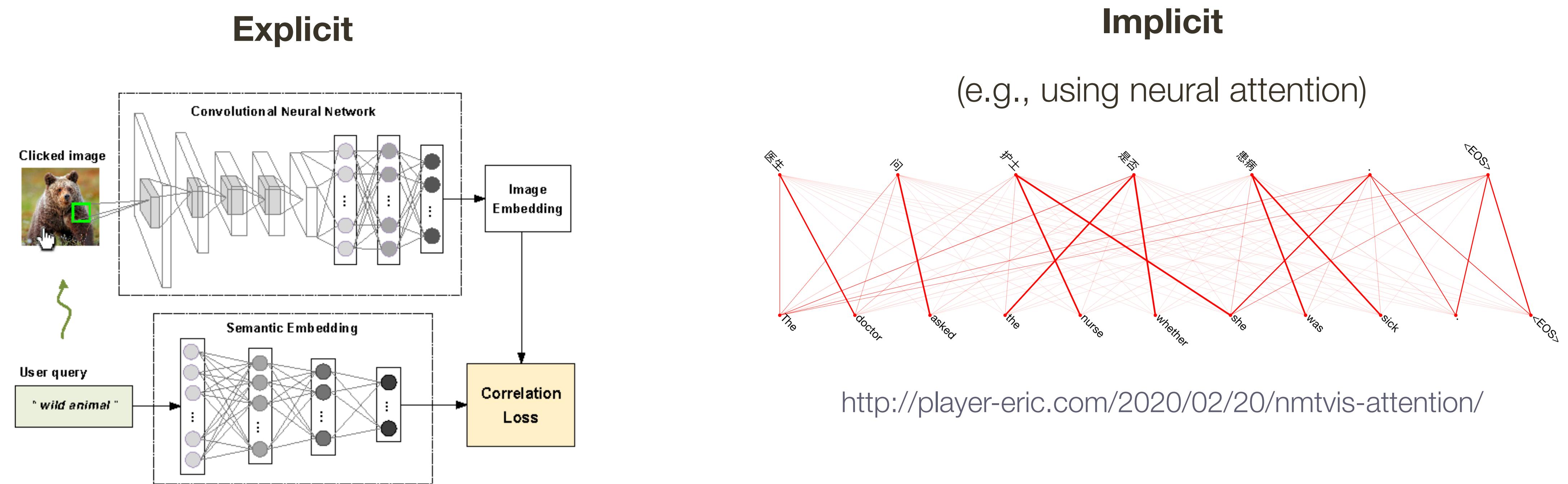
Allows translation between languages pairs never trained on before



[ Johnson et al., ArXiv 2017 from Google ]

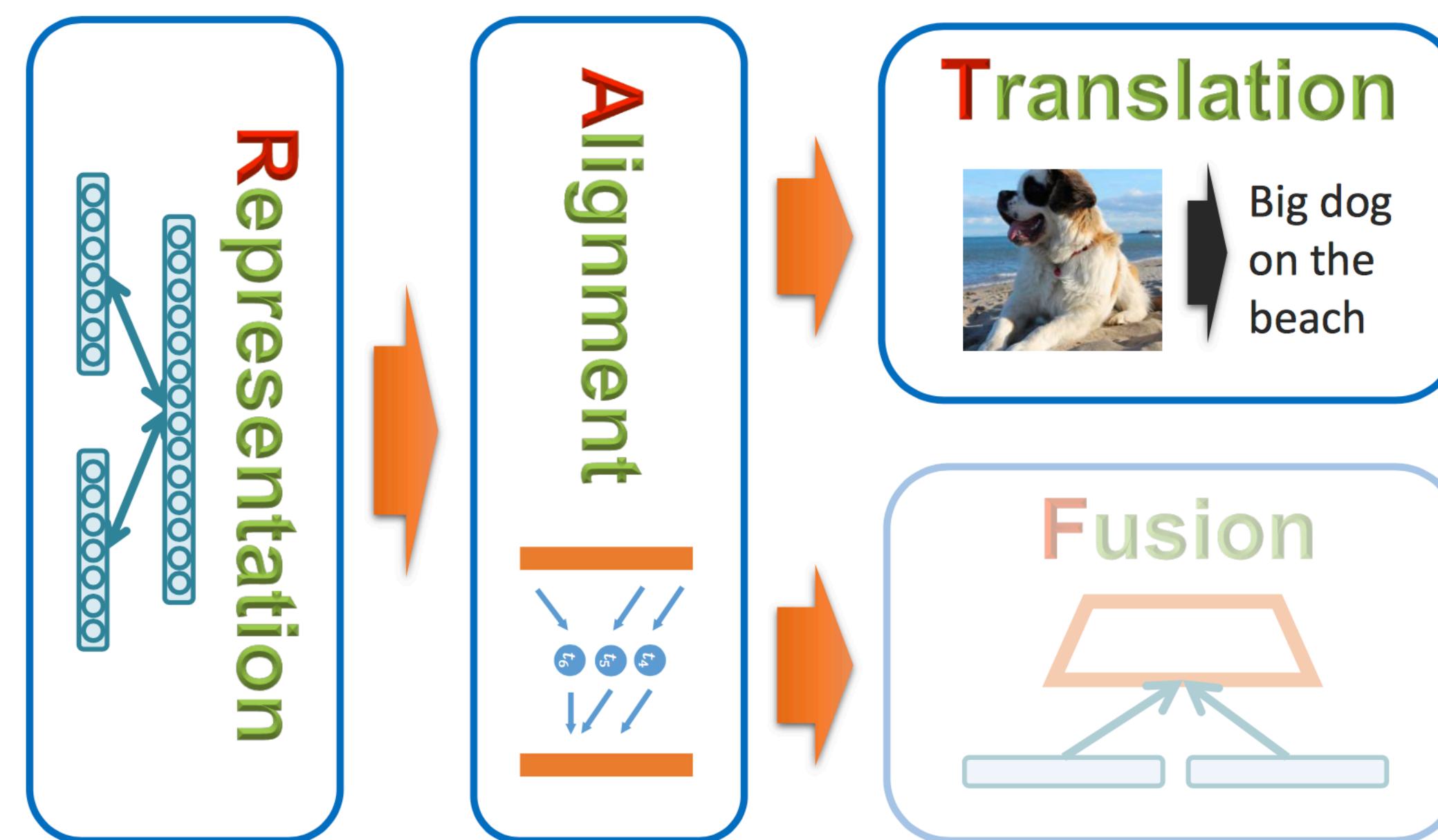
# Key Challenges of Multimodal Learning

- Representation learning in each and across modalities
- Alignment between representations in different modalities
- Translation between modalities



# Key Challenges of Multimodal Learning

- Representation learning in each and across modalities
- Alignment between representations in different modalities
- Translation between modalities



# Objectives of the course

- Acquire **fundamentals and background** that would allow one to follow research in Computer Vision and on intersection of Vision + Language
- Ability to **design, build and apply deep learning architectures** for multi-modal problems (Vision + Language in particular)
- Obtain **overview of research trends** in Computer Vision and ML related to topics of the course
- Ability to define research problems, read and present research papers

# Objectives of the course

- Acquire **fundamentals and background** that would allow one to follow research in Computer Vision and on intersection of Vision + Language
- Ability to **design, build and apply deep learning architectures** for multi-modal problems (Vision + Language in particular)
- Obtain **overview of research trends** in Computer Vision and ML related to topics of the course
- Ability to define research problems, read and present research papers

course is heavy on *practical* deep learning

# Deep Learning

## Google snaps up object recognition startup DNNresearch

Google has acquired a research startup founded within the University of Toronto, whose work includes object recognition.

by [Josh Lowensohn](#) [@Josh](#) / 13 March 2013, 9:22 am AEDT

2 / 0 / 0 / 0 / 0 / 0 / more +

Google has acquired a three-person Canadian research company that specializes in voice and image recognition.

DNNresearch, which was founded last year within the University of Toronto's computer science department, specializes in object recognition and now belongs to Google.



From left: Ilya Sutskever, Alex Krizhevsky and University Professor Geoffrey Hinton of the University of Toronto's Department of Computer Science. (photo by John Guatto, University of Toronto)

# Deep Learning

## Google snaps up object recognition startup DNNresearch

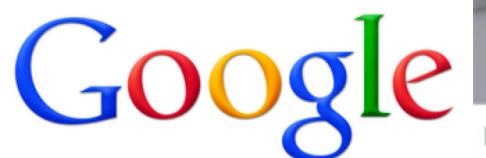
Google has acquired a research startup founded within the University of Toronto, whose work includes object recognition.

by [Josh Lowensohn](#) / [@Josh](#) / 13 March 2013, 9:22 am AEDT

2 / f 0 / t 0 / in 0 / g+ 0 / more +

Google has acquired a three-person Canadian research company that specializes in voice and image recognition.

DNNresearch, which was founded last year within the University of Toronto's computer science department, specializes in object recognition and now belongs to Google.



From left: Ilya Sutskever, Alex Krizhevsky and University of Toronto's Department of Computer Science. (photo by John Guatto, University of Toronto)

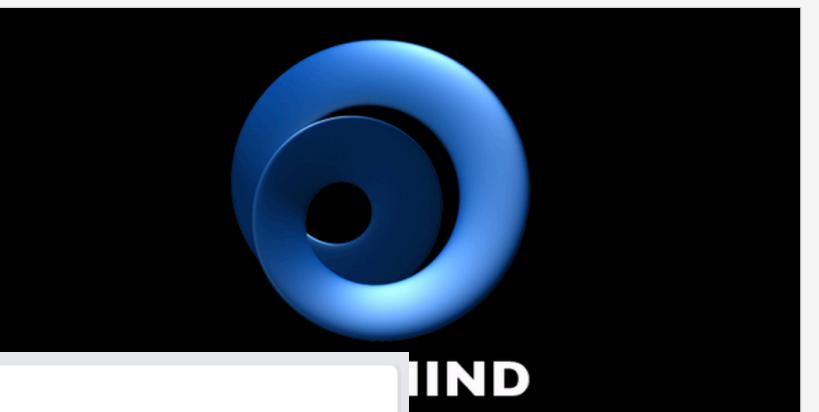
8TH ANNUAL CRUNCHIES AWARDS Celebrate the Best of Tech in 2014 Get Your Tickets Now ▶

### Google Acquires Artificial Intelligence Startup For More Than \$500M

Posted Jan 26, 2014 by [Catherine Shu](#) (@catherineshu)

11.3k SHARES f t in g+ J F E

Next Story



CrunchBase

### DeepMind

FOUNDED 2011

OVERVIEW DeepMind is a cutting edge artificial intelligence company. We combine the best techniques from machine learning and systems neuroscience to build powerful general-purpose learning algorithms. Founded by Demis Hassabis, Shane Legg and Mustafa Suleyman, the company is based in London and supported by some of the world's leading entrepreneurs and investors.



**Yann LeCun**

December 9, 2013 ·

Big news today!

Facebook has created a new research laboratory with the ambitious, long-term goal of bringing about major advances in Artificial Intelligence.

« Search needs a shake-up

Songbirds use grammar rules »

## Machine Learning Startup Acquired by ai-one

### Press Release

For Immediate Release: August 4, 2011

San Diego artificial intelligence startup acquired by leading provider of machine learning SDKs as market for advanced applications gets hot.

San Diego CA – **ai-one** announced today that it acquired Auto-Semantics, a local start-up providing artificial intelligence services to corporate IT departments. The acquisition is the latest in a series of joint-ventures and acquisitions by ai-one that consolidates its leadership position within the emerging market for machine learning technologies.

# Clever Hans



# Clever Hans



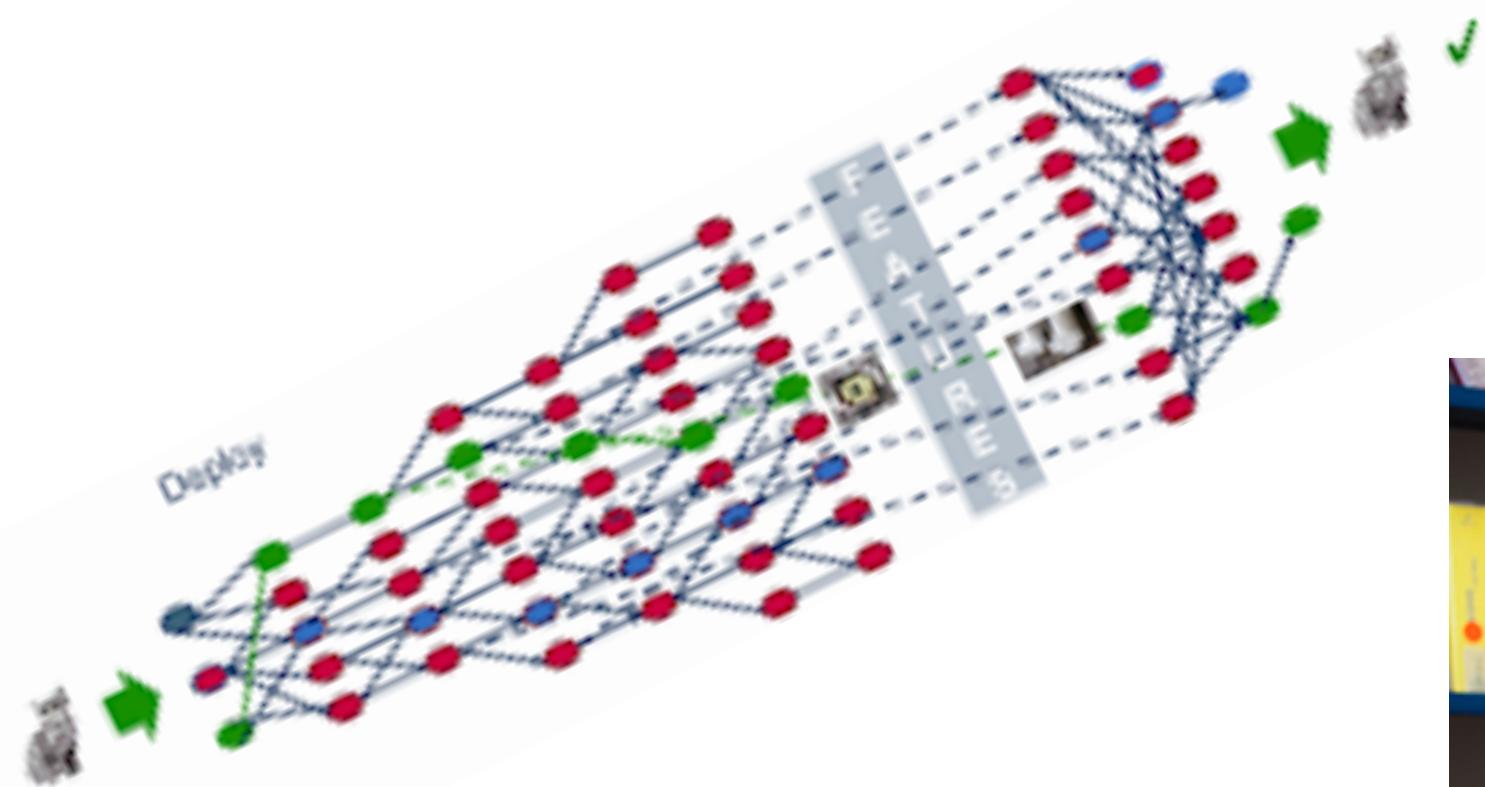
Hans could get 89% of the math questions right

# Clever Hans



Hans could get 89% of the math questions right

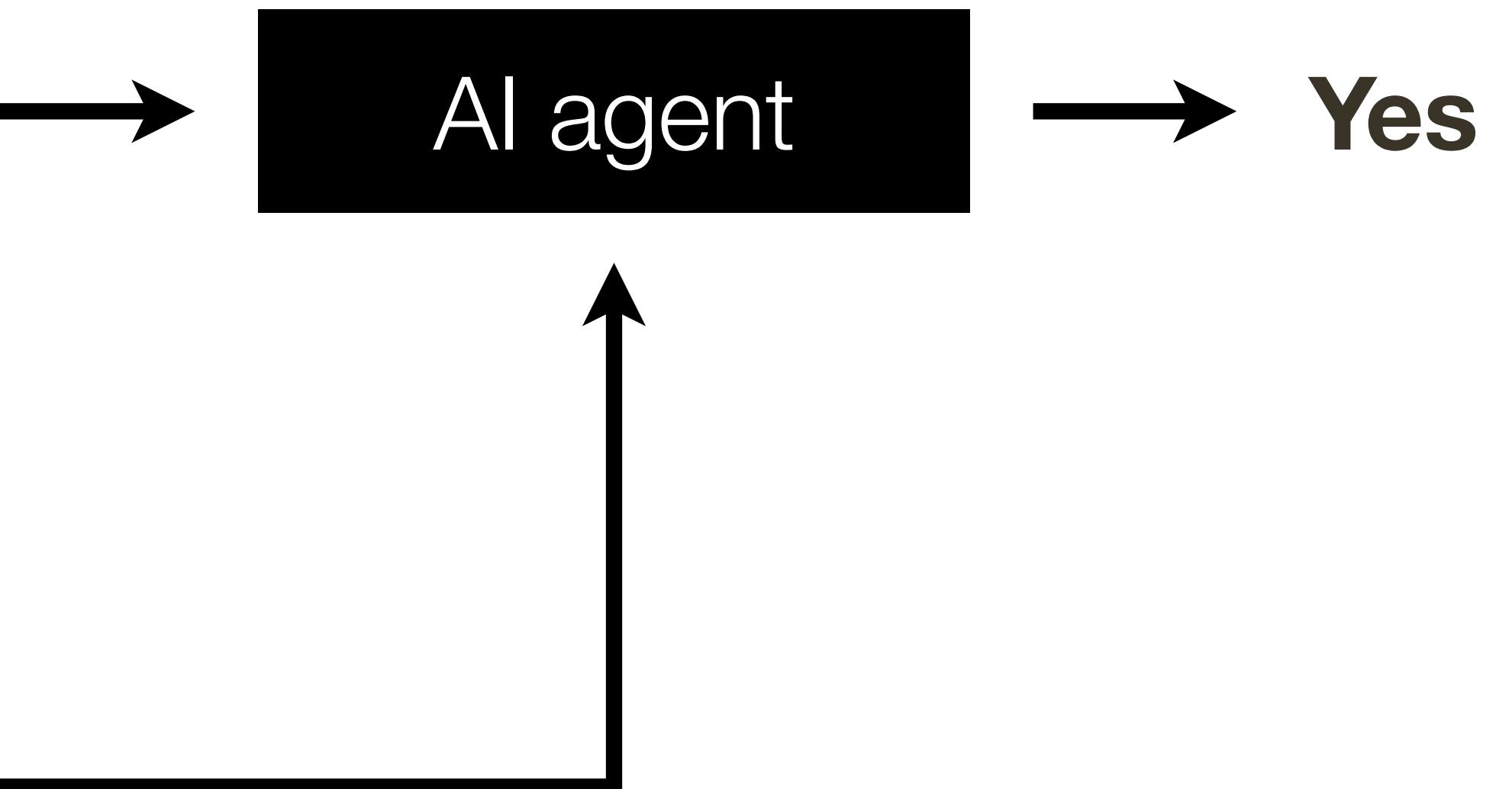
# Clever DNN



# Visual Question Answering



Is there zebra climbing the tree?



# What will be covered **specifically**?

## Course Webpage from Past Offering

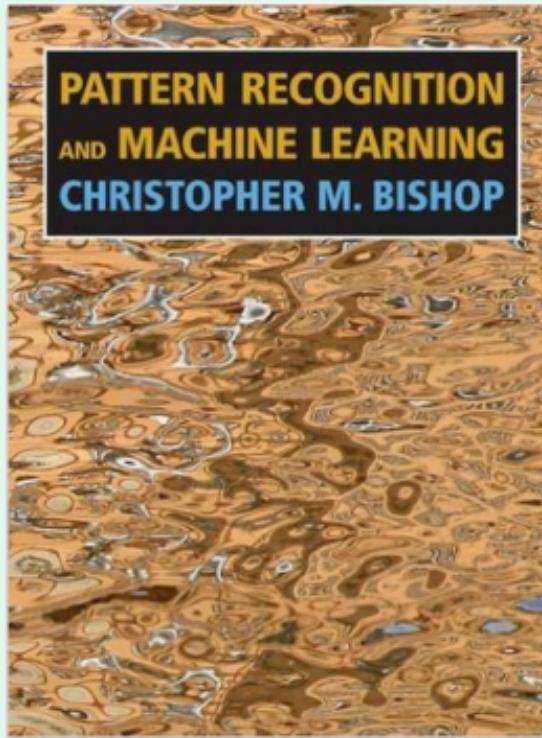
(e.g., AutoDifferentiation, Neural Networks, Convolution NN, RNNs, Attention, Graph Neural Networks, Generative Models, etc.)

## + New Advances Since Then

(e.g., VQVAE, Diffusion Models, etc.)

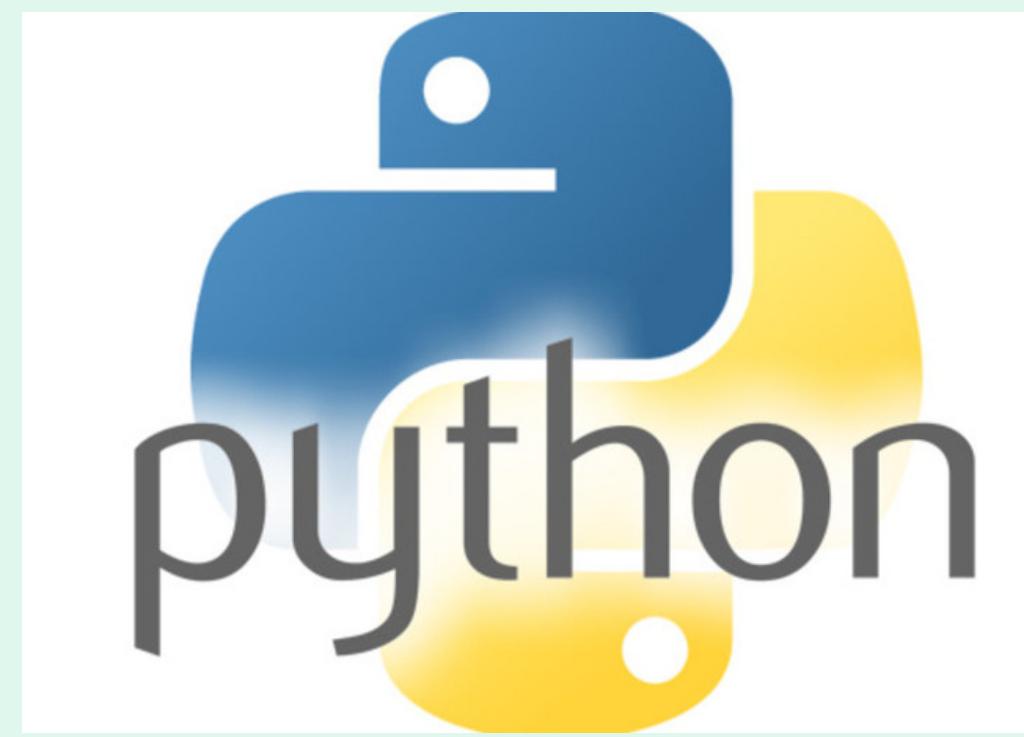
# Pre-requisites

# Computer Science



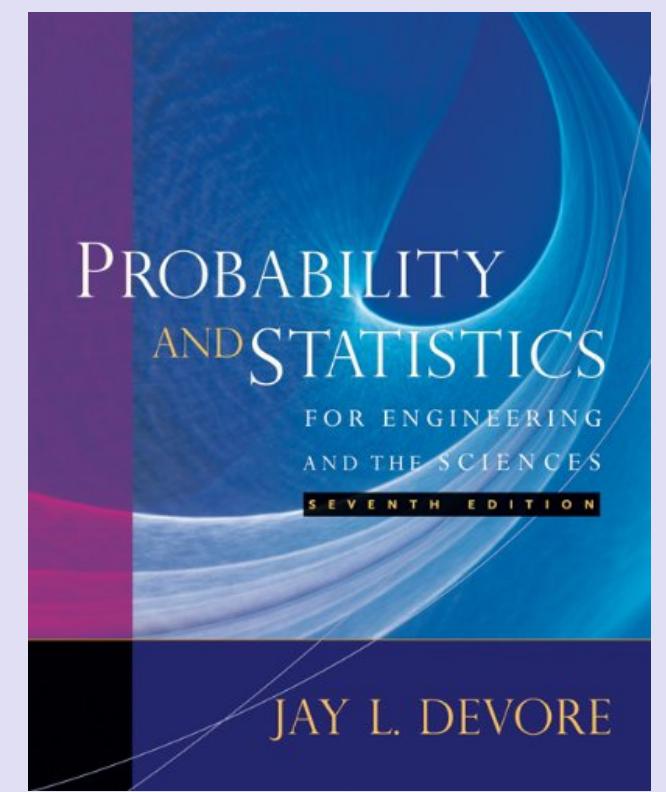
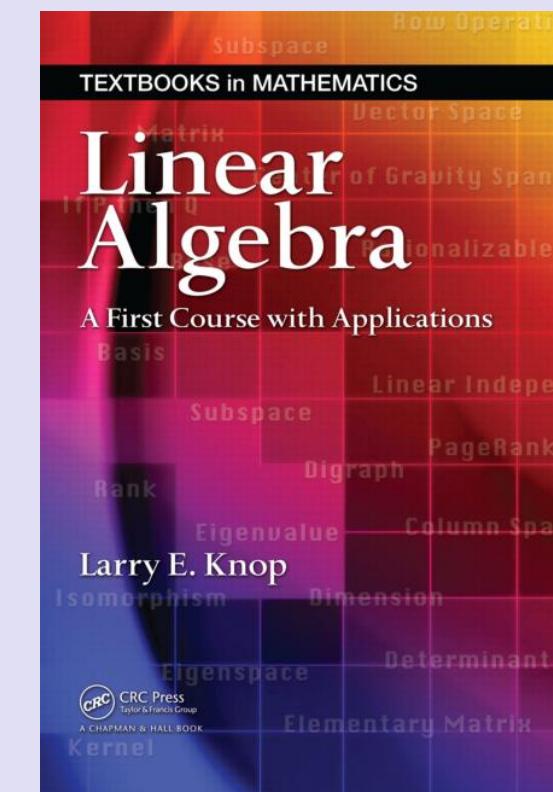
# CPSC 340

(or equivalent)



# Needed for Assignments

# Mathematics



# Calculus

# Linear Algebra

# Statistics

# **Helpful** (but not necessary): some background in Computer Vision or NLP

# Additional Requirement

**Option 1:** Google Colab



# Additional Requirement

**Option 1:** Google Colab



**Option 2:** Own hardware



Nvidia GTX 1060 (with 6GB RAM) or above

# Additional Requirement

## Option 1: Google Colab



## Option 2: Own hardware



Nvidia GTX 1060 (with 6GB RAM) or above

## Option 3: Cloud services



As student you get \$25 in credits



As student you get \$100 in credits

You will need to provision the VM and ensure you keep track of spendings. As long as VM is running you are being charged, even if you are not running the code.

# Course structure



Approximately 70% of course will consists of lectures and optional readings



Remaining 30% is reading and presentation of curated research papers on relevant topics



5 programming assignments

Final (individual or **group**) project

# Course structure



Approximately 70% of course will consists of lectures and optional readings



Remaining 30% is reading  
curated research papers on relevant topics



+ **guest speakers**



5 programming assignments

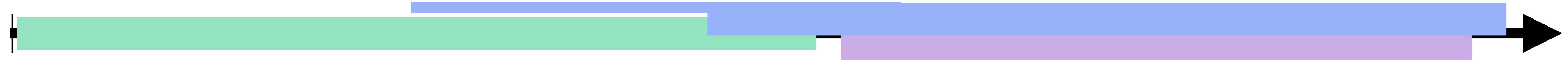
Final (individual or **group**) project

# Grading Criteria

- **Assignments** (programming) – 40% (total)

- **Research papers** – 20%

- **Project** – 40%



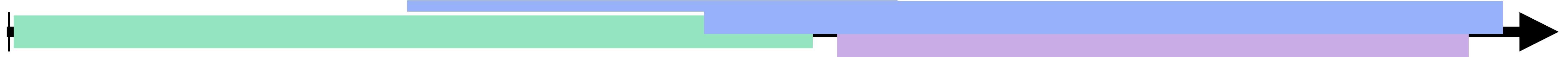
# Grading Criteria

- **Assignments** (programming) – 40% (total)

- **Research papers** – 20%

- **Project** – 40%

**NO LATE SUBMISSIONS** – If you don't complete the assignment, hand in what you have



# Assignments

(5 assignments and 40% of grade total)

- Assignment 0: **Introduction to PyTorch** (0%)
- Assignment 1: **Neural Network Introduction** (5%) –  python<sup>TM</sup>

Assignments all use **Python Jupiter Notebooks**, use Canvas to hand everything in. Assignments always due at **11:59pm PST** on due date.

# Assignments

(5 assignments and 40% of grade total)

- Assignment 0: **Introduction to PyTorch** (0%)
- Assignment 1: **Neural Network Introduction** (5%) —  python<sup>TM</sup>
- Assignment 2: **Convolutional Neural Networks** (5%) — PYTORCH

Assignments all use **Python Jupiter Notebooks**, use Canvas to hand everything in. Assignments always due at **11:59pm PST** on due date.

# Assignments

(5 assignments and 40% of grade total)

- Assignment 0: **Introduction to PyTorch** (0%)
- Assignment 1: **Neural Network Introduction** (5%) —  python™
- Assignment 2: **Convolutional Neural Networks** (5%) — PYTORCH
- Assignment 3: **RNN Language Modeling and Translation** (10%) — PYTORCH

Assignments all use **Python Jupiter Notebooks**, use Canvas to hand everything in. Assignments always due at **11:59pm PST** on due date.

# Assignments

(5 assignments and 40% of grade total)

- Assignment 0: **Introduction to PyTorch** (0%)
- Assignment 1: **Neural Network Introduction** (5%) —  python™
- Assignment 2: **Convolutional Neural Networks** (5%) — PYTORCH
- Assignment 3: **RNN Language Modeling and Translation** (10%) — PYTORCH
- Assignment 4: Neural Model for **Image Captioning / Retrieval** (10%) — PYTORCH

Assignments all use **Python Jupiter Notebooks**, use Canvas to hand everything in. Assignments always due at **11:59pm PST** on due date.

# Assignments

(5 assignments and 40% of grade total)

- Assignment 0: **Introduction to PyTorch** (0%)
- Assignment 1: **Neural Network Introduction** (5%) —  python™
- Assignment 2: **Convolutional Neural Networks** (5%) — PYTORCH
- Assignment 3: **RNN Language Modeling and Translation** (10%) — PYTORCH
- Assignment 4: Neural Model for **Image Captioning / Retrieval** (10%) — PYTORCH
- Assignment 5: Advanced Architectures **Graph NN** and **GANs** (10%) — PYTORCH

Assignments all use **Python Jupiter Notebooks**, use Canvas to hand everything in. Assignments always due at **11:59pm PST** on due date.

# Assignments

(5 assignments and 40% of grade total)

I reserve the right to **change** release and due dates for the assignments to accommodate constraints of the course, do not take the dates on web-page as “set in stone”.

# Research Papers

(reviews and presentation, 20% of grade total)

## Presentation - 10%

- You will need to **present 1 paper** individually or as a group (group size will be determined by # of people in class)
- Pick a paper from the syllabus individually (we will have process to pick #1, #2, #3 choices)
- Will need to prepare slides and **meet with me or TA** for feedback
- It is your responsibility to schedule these meetings
- Very likely is that I will ask you to **record** these presentation and we will make these available

# **Research Papers** (reviews and presentation, 20% of grade total)

## **Presentation** - 10%

- You will need to **present 1 paper** individually or as a group (group size will be determined by # of people in class)
- Pick a paper from the syllabus individually (we will have process to pick #1, #2, #3 choices)
- Will need to prepare slides and **meet with me or TA** for feedback
- It is your responsibility to schedule these meetings
- Very likely is that I will ask you to **record** these presentation and we will make these available

## Reading **Reviews** - 10%

- Individually, one for every class after the first half of semester
- Due 11:59pm a day before class where reading assigned, submitted via Canvas

# Good Presentation

- You are effectively taking on responsibility for being an instructor for part of the class (**take it seriously**)
- What makes a **good presentation**?
  - High-level overview of the problem and motivation
  - Clear statement of the problem
  - Overview of the technical details of the method, including necessary background
  - Relationship of the approach and method to others discussed in class
  - Discussion of strengths and weaknesses of the approach
  - Discussion of strengths and weaknesses of the evaluation
  - Discussion of potential extensions (published or potential)

# Reading Reviews

- Designed to make sure you read the material and have thought about it prior to class (to stimulate discussion)
  - Short summary of the paper (3-4 sentences)
  - Main contributions (2-3 bullet points)
  - Positive / negative points (2-3 bullet points each)
  - What did you not understand (was unclear) about the paper (2-3 bullet points)

# Final Project (40% of grade total)

- Group project (groups of 3 are encouraged, but fewer maybe possible)
- Groups are self-formed, you will not be assigned to a group
- You need to come up with a project proposal and then work on the project as a group (each person in the group gets the same grade for the project)
- Project needs to be **research** oriented (not simply implementing an existing paper); you can use code of existing paper as a starting point though

Project proposal + class presentation: 15%  
Project + final presentation (during finals week): 25%

# Sample Project Ideas

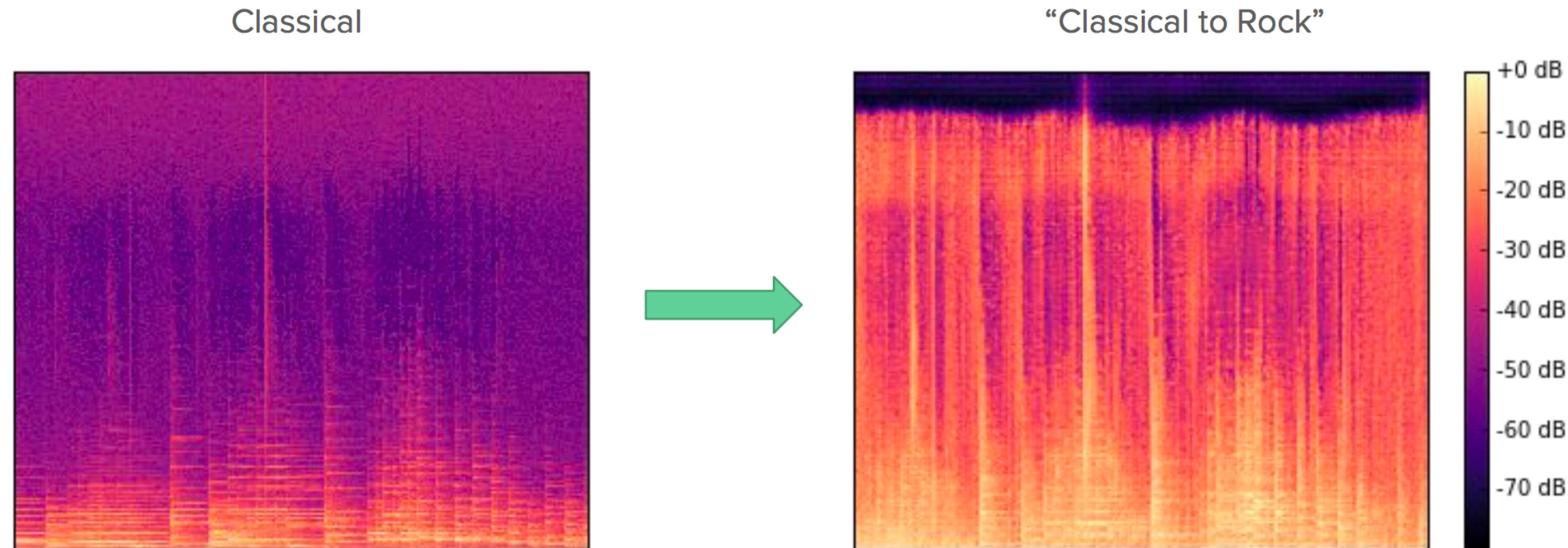
- Translate an image into a cartoon or Picasso drawing better than existing approaches (e.g., experiment with loss functions, architectures)
- Generating video clips by retrieving images relevant to lyrics of songs
- Generating an image based on the sounds or linguistic description
- Compare different feature representation and role of visual attention in visual question answering
- Storyboarding movie scripts
- Grounding a language/sound in an image

... there are **endless possibilities** ... think **creatively** and **have fun!**

# Project Example: Dreaming of Music

by Sijia (Candice) Tian, Alexandra Kim, Itrat Akhrt

Evaluate the effectiveness of using visual music representation (spectrograms) to do classification and modify music using deep learning



Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *arXiv preprint arXiv:1703.10593* (2017).

Explored image-to-image translation techniques to translate musical styles

# Project Example: Robust Adversarial Detection

by Michael and Marjan

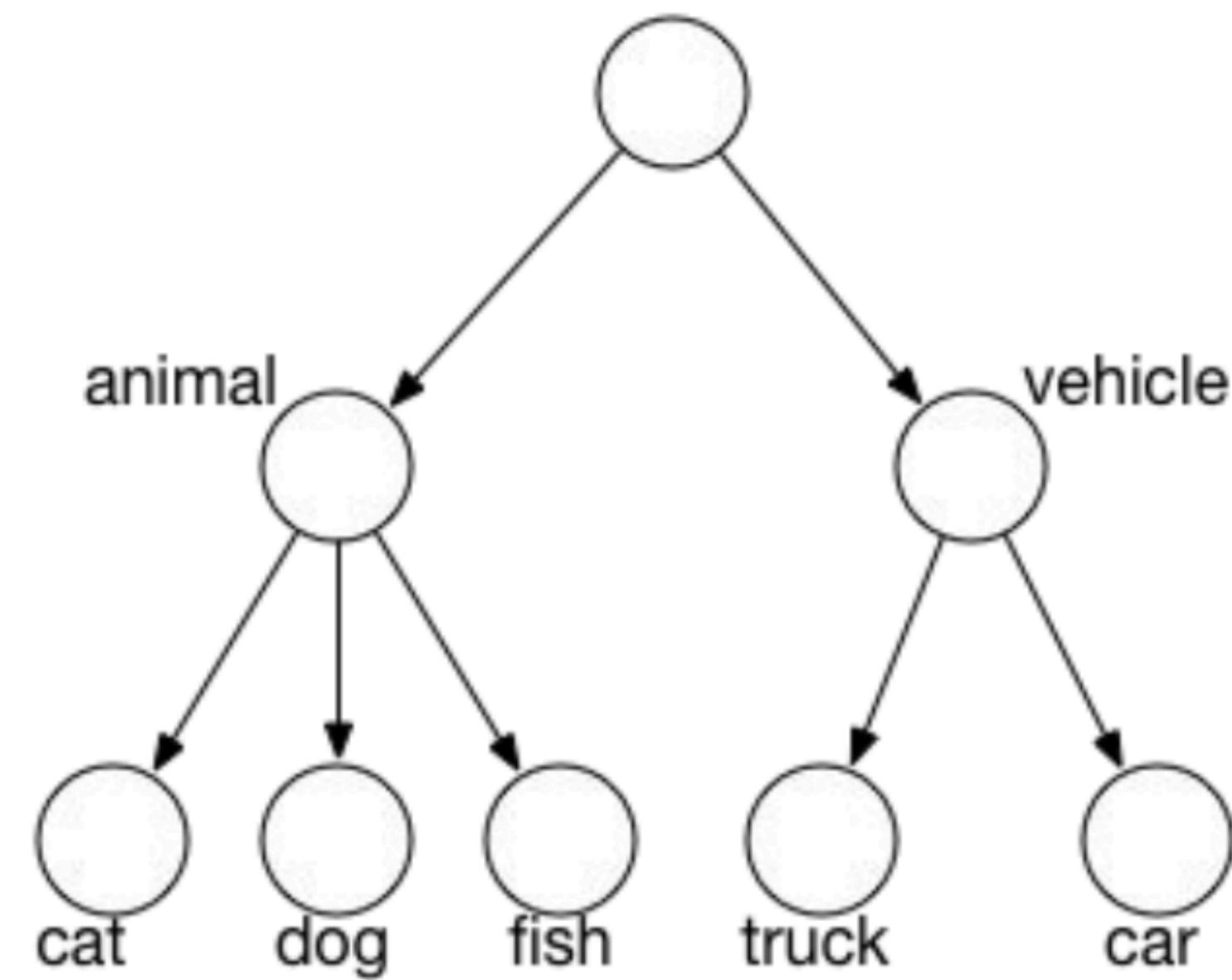
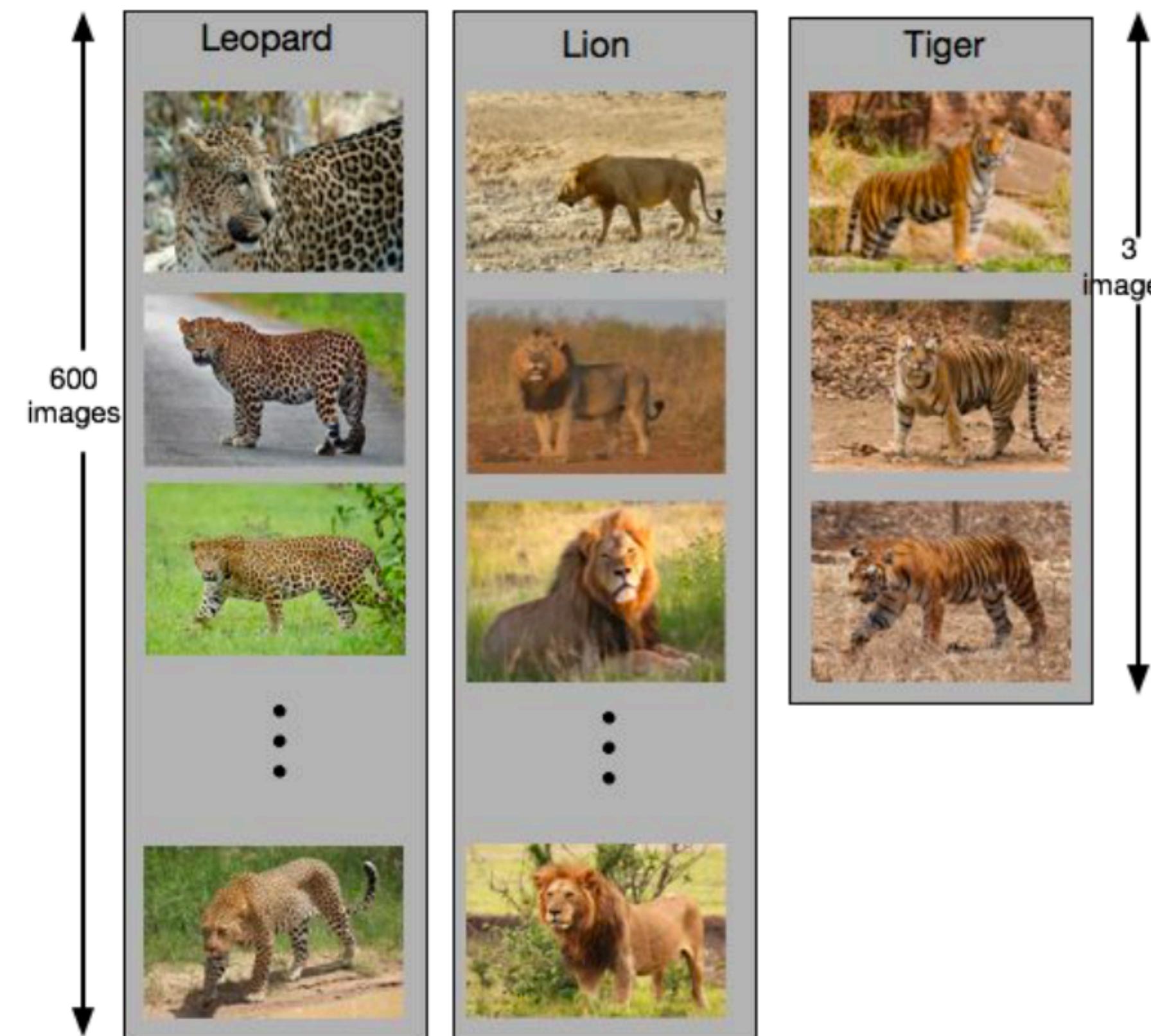
Bayesian Neural Network and variational inference for detecting and analyzing  
adversarial attacks



# Project Example: Classification with Tree Priors

by Saeid Naderiparizi and Setareh Cohan

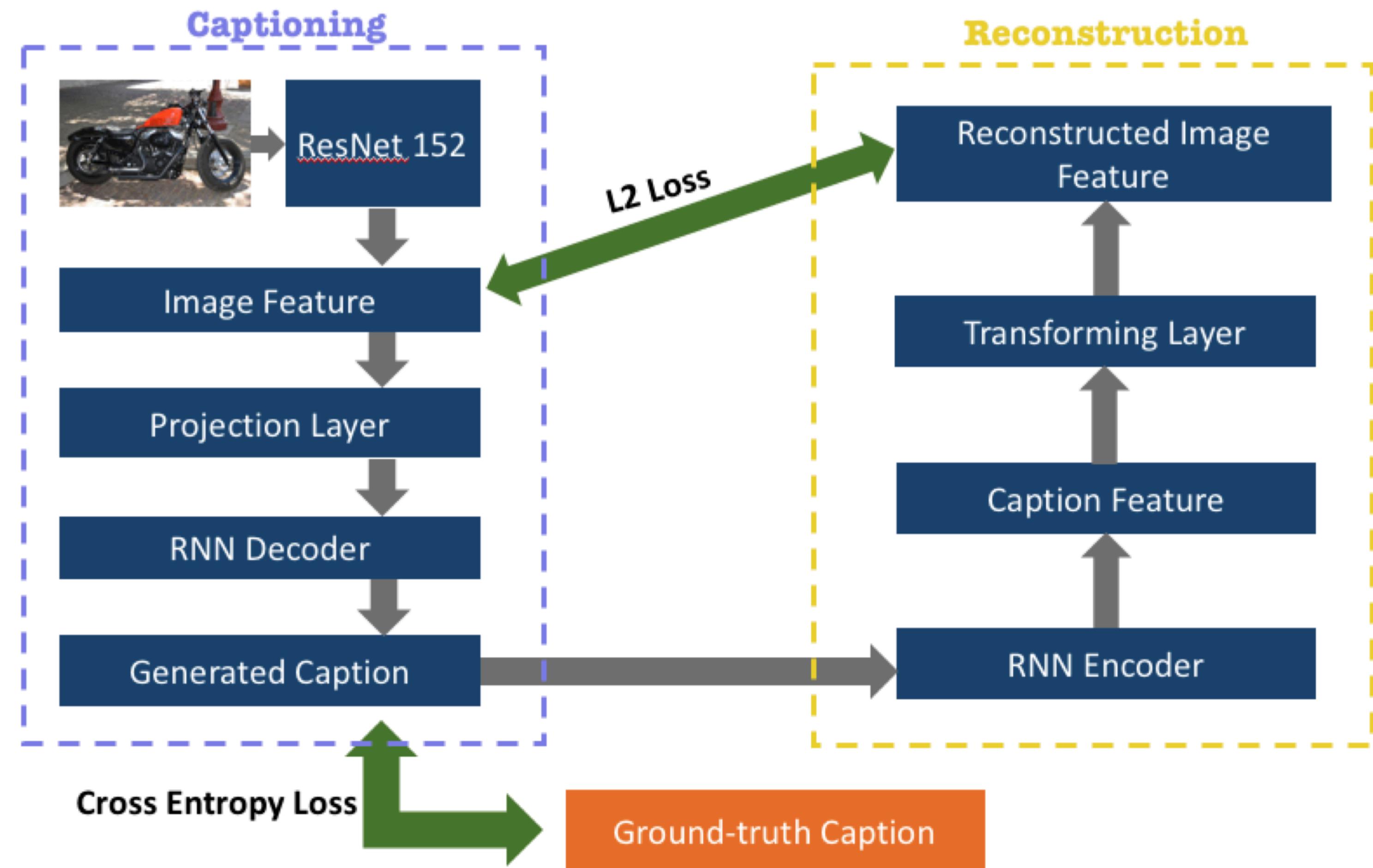
Classification with few samples using transfer learning techniques



# Project Example: Semi-supervised Image Captioning

by Bicheng Xu, Weirui Kong, Jiaxuan Chen

Effective use of unlabeled data during training of an image captioning network

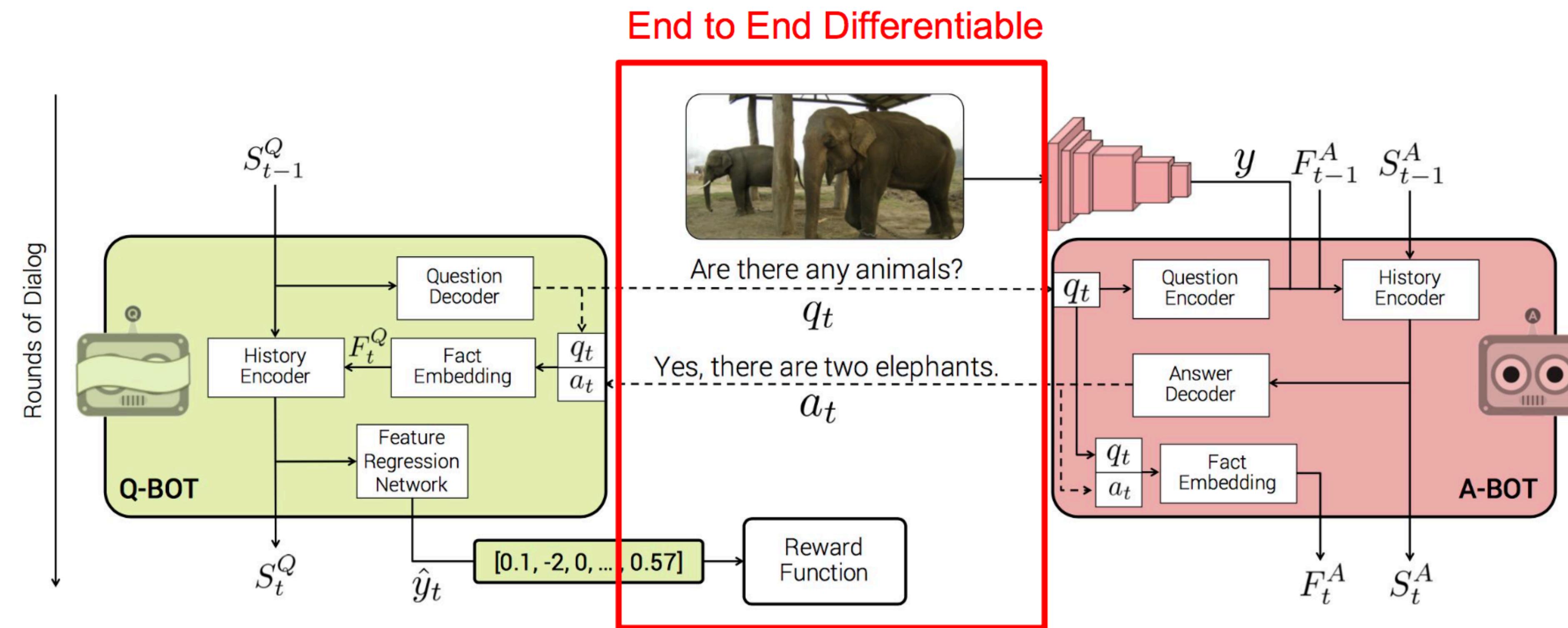


# Project Example: Visual Question Answering

Improve interaction between two agents

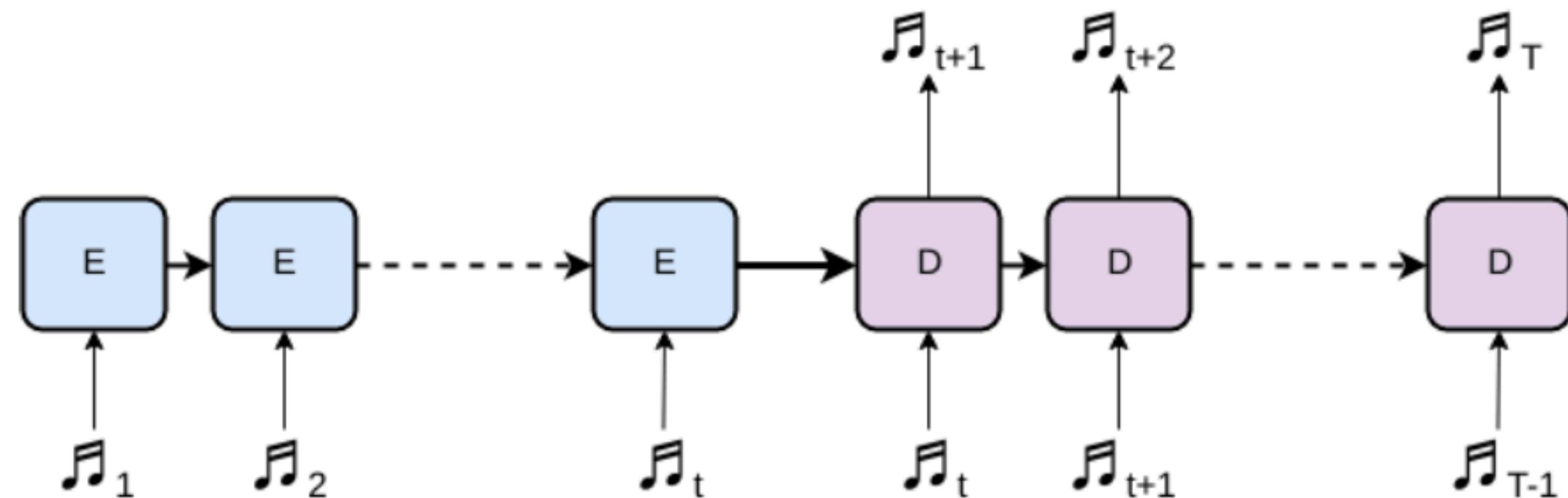
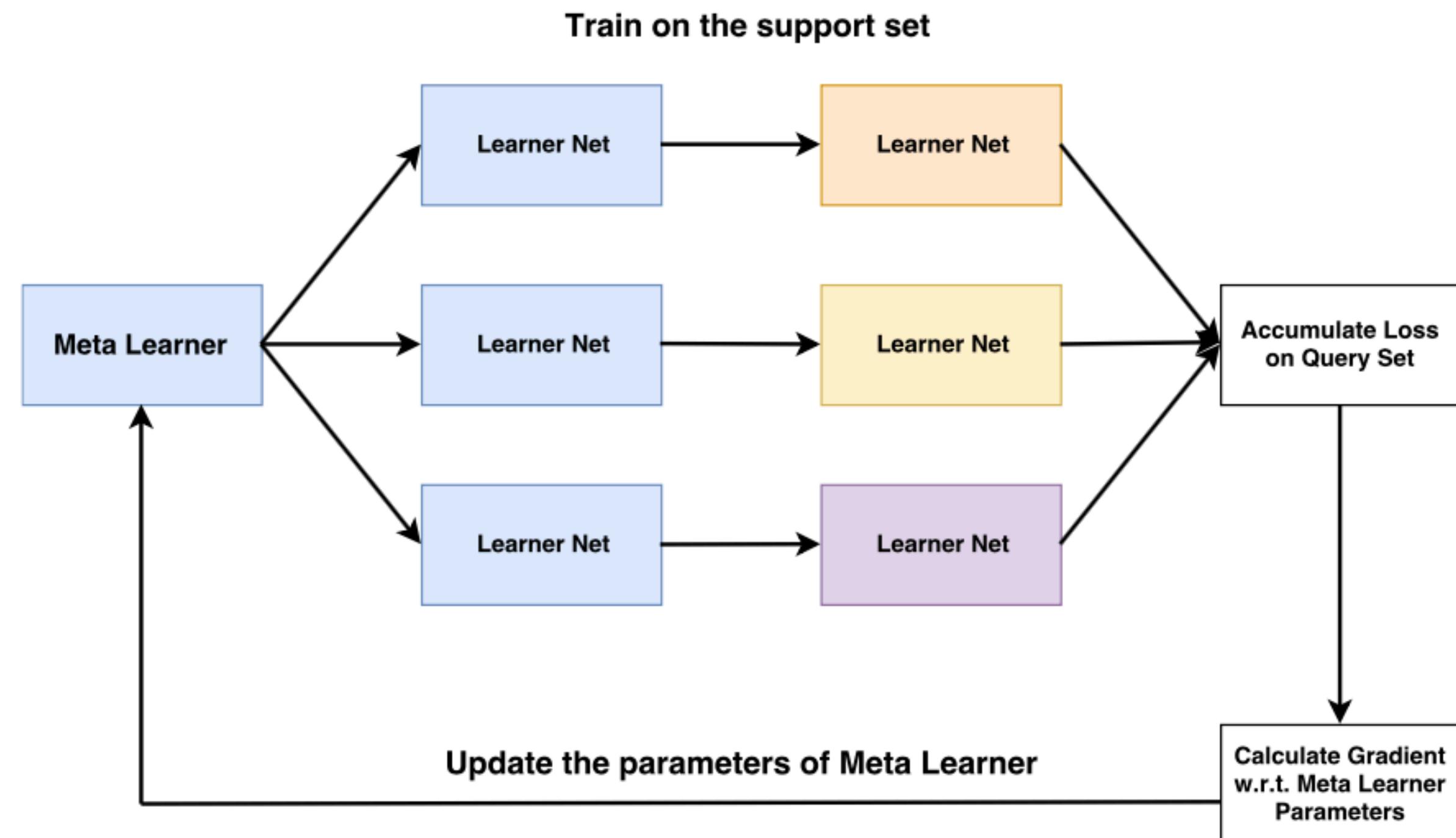
- End-to-end differentiability
- Discriminator for human-like questions

by Siddhesh Khandelwal, Mohit Bajaj, Gursimran Singh



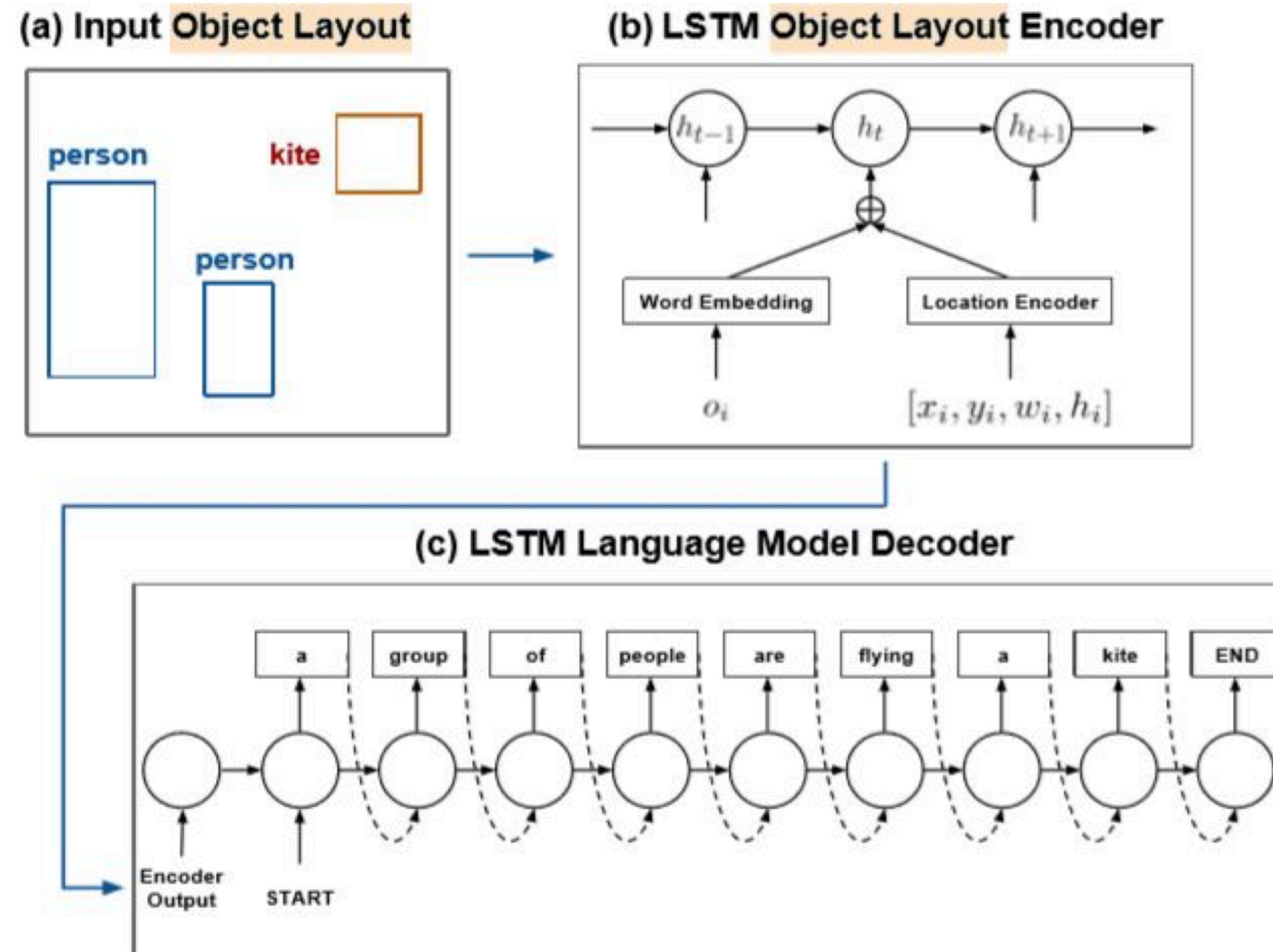
# Project Example: Few Shot MIDI Music Generation

by Ben, Suhail, Anand



# Project Example: Visually Descriptive Language from Layout

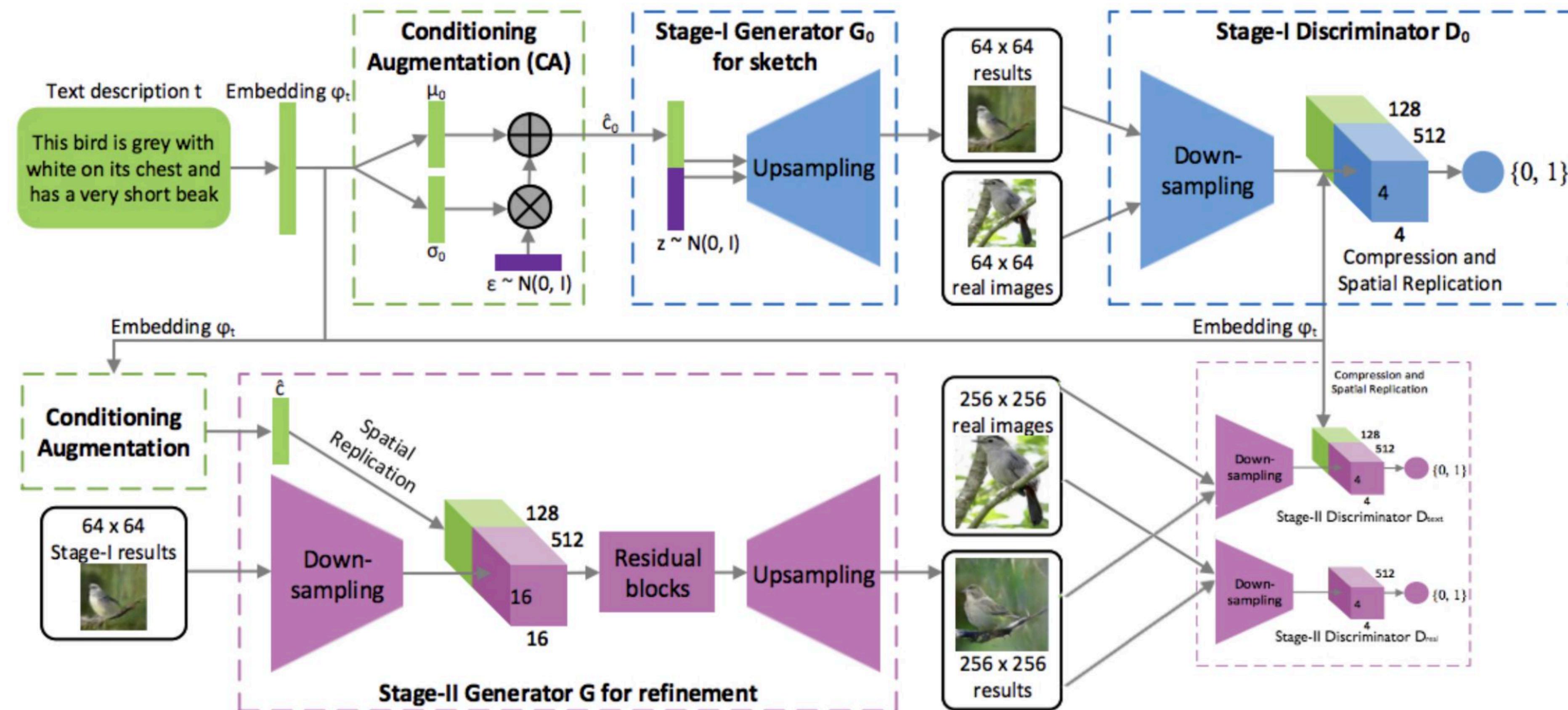
by Ke Ma, Wen Xiao, Sing Zeng



# Project Example: StackGAN with Different Losses

by Polina Zablotskaia

Automatic synthesis of realistic images from text



H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In ICCV, 2017.