THE UNIVERSITY OF BRITISH COLUMBIA

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

Lecture 1: Introduction

Course **logistic**

Times: Tues & Thurs 9:30-11:00am **Locations:** DMP 101

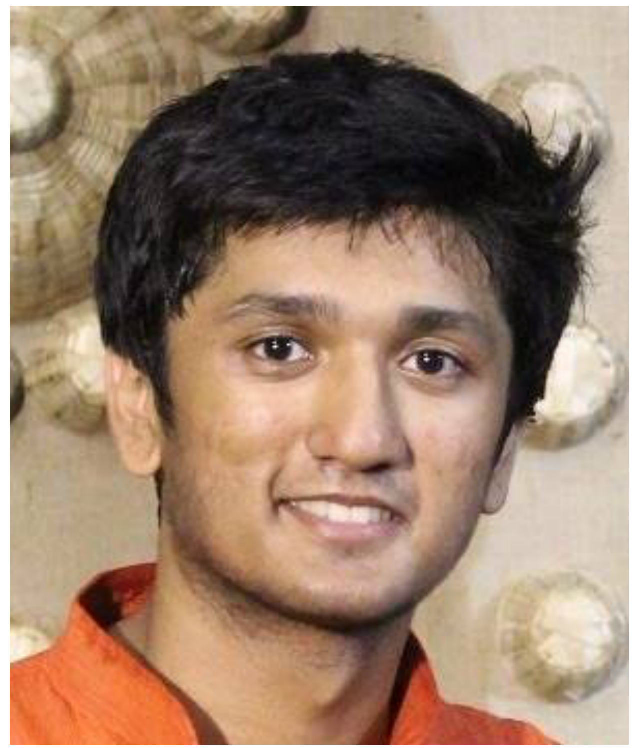
Instructor: Leonid Sigal

E-mail: lsigal@cs.ubc.ca
Office: ICICS 119

TA: Mohit Bajaj

E-mail: mbajaj01@cs.ubc.ca

Siddhesh Khandelwal



skhandel@cs.ubc.ca

Course webpage: https://www.cs.ubc.ca/~lsigal/teaching18.html

Discussion: piazza.com/ubc.ca/winterterm22018/cpsc532s

Course **logistic**

Times: Tues & Thurs 9:30-11:00am

Locations: DMP 101

If you **have not registered** for the course but want to take it, **sign up on the sheet**, come talk to me after class or schedule a meeting

Course webpage: https://www.cs.ubc.ca/~lsigal/teaching18.html

Discussion: piazza.com/ubc.ca/winterterm22018/cpsc532s

About **me …**

Associate Professor
2017 -

THE UNIVERSITY OF BRITISH COLUMBIA

Senior Research Scientist
2009 - 2017

Disney Research

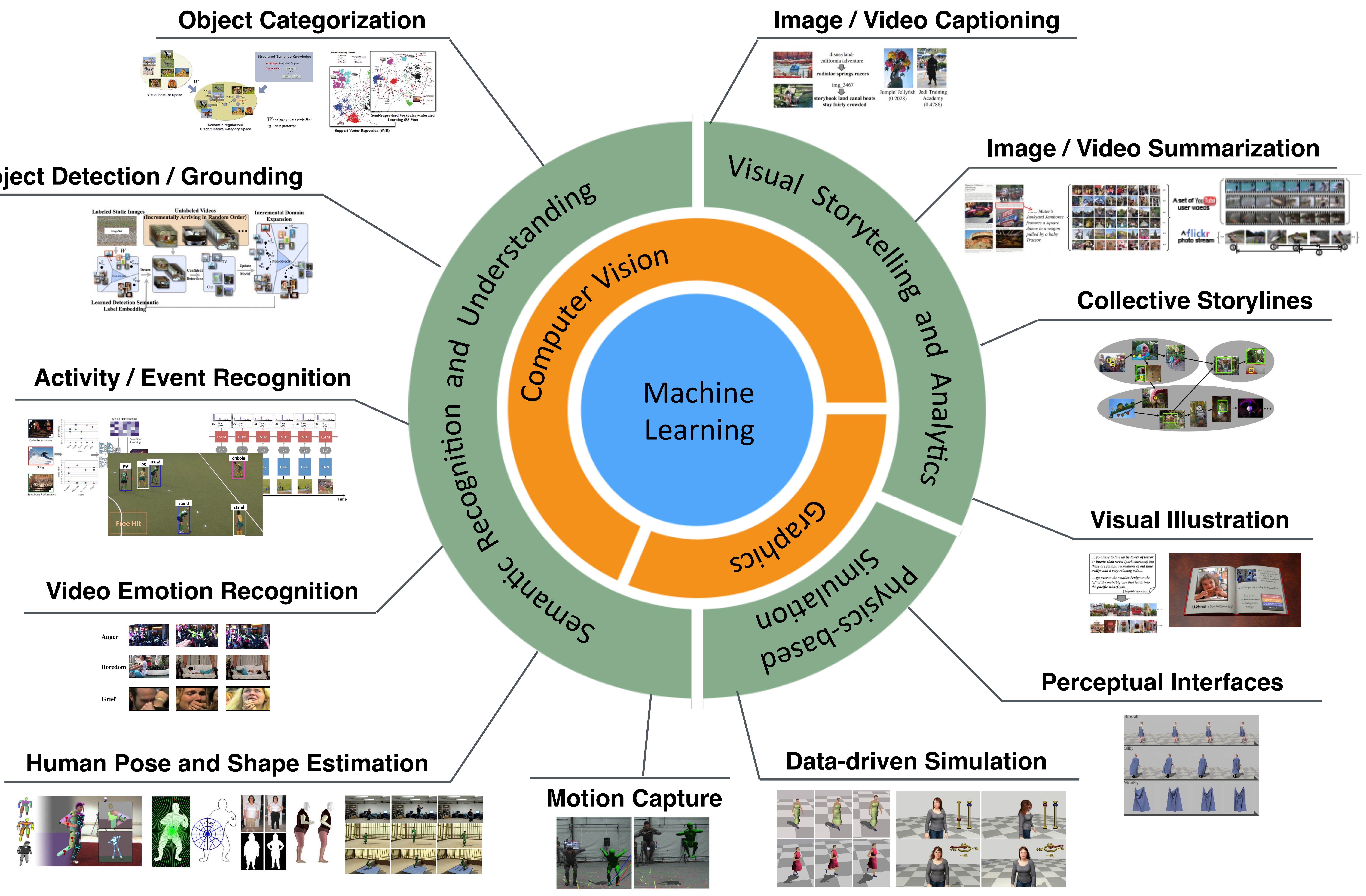
Postdoctoral Researcher
2007 - 2009

UNIVERSITY OF TORONTO

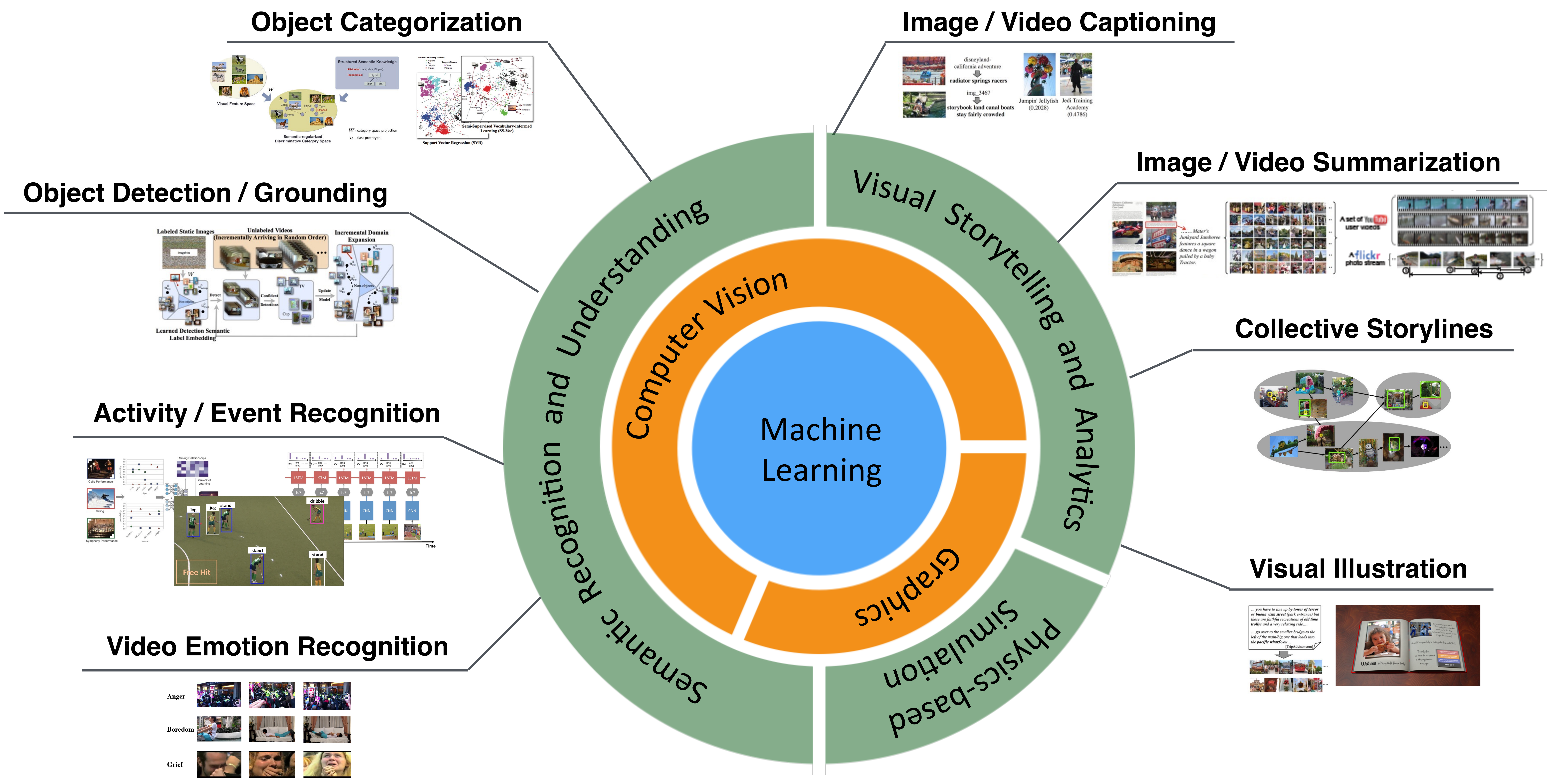
PhD, MSc
2001 - 2008

BROWN

BOSTON UNIVERSITY



Machine Learning
Computer Vision
Graphics
Semantic Recognition and Understanding
Visual Storytelling and Analytics
Physics-based Simulation
Object Categorization
Object Detection / Grounding
Activity / Event Recognition
Video Emotion Recognition
Human Pose and Shape Estimation
Motion Capture
Data-driven Simulation
Perceptual Interfaces
Visual Illustration
Collective Storylines
Image / Video Summarization
Image / Video Captioning
Anger
Boredom
Grief


Object Categorization
Image / Video Captioning
Image / Video Summarization
Object Detection / Grounding
Collective Storylines
Activity / Event Recognition
Visual Illustration
Video Emotion Recognition
Semantic Recognition and Understanding
Visual Storytelling and Analytics
Physics-based Simulation
Computer Vision
Graphics
Machine Learning
Anger
Boredom
Grief

What is **Multi-modal Learning**?

What is **Multi-modal Learning**?

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

What is **Multi-modal Learning**?

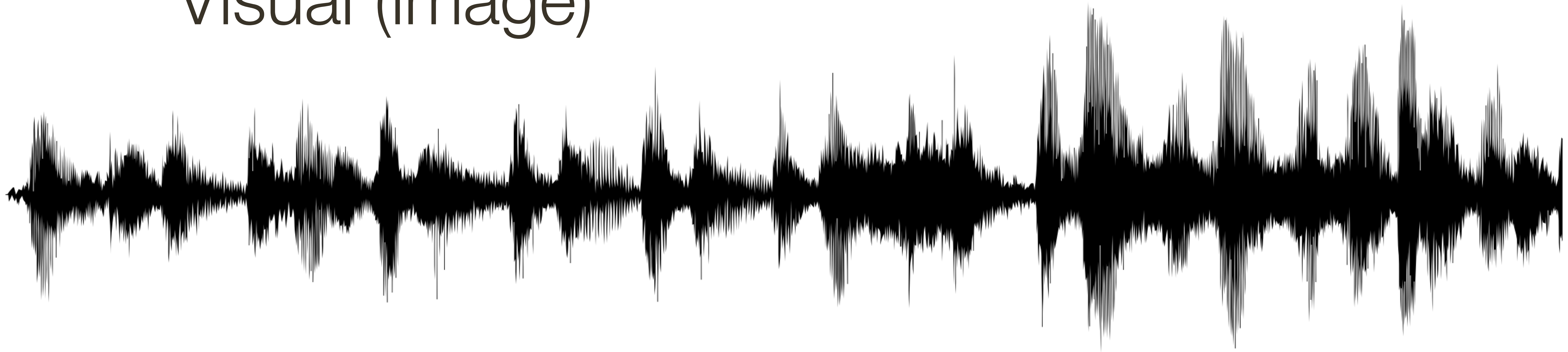
- **Modality:** refers to a certain type of information and/or representation format in which information is stored.
- **Sensory modality:** one or more primary channels of communication.

Visual (drawings)

Visual (image)

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 it association with the moon it has ties to fertility and seduction. The owl is bird of magic and darkness of prophecy and wisdom.

Natural Language (text)



Auditory (voice / sound)

Haptic / Touch

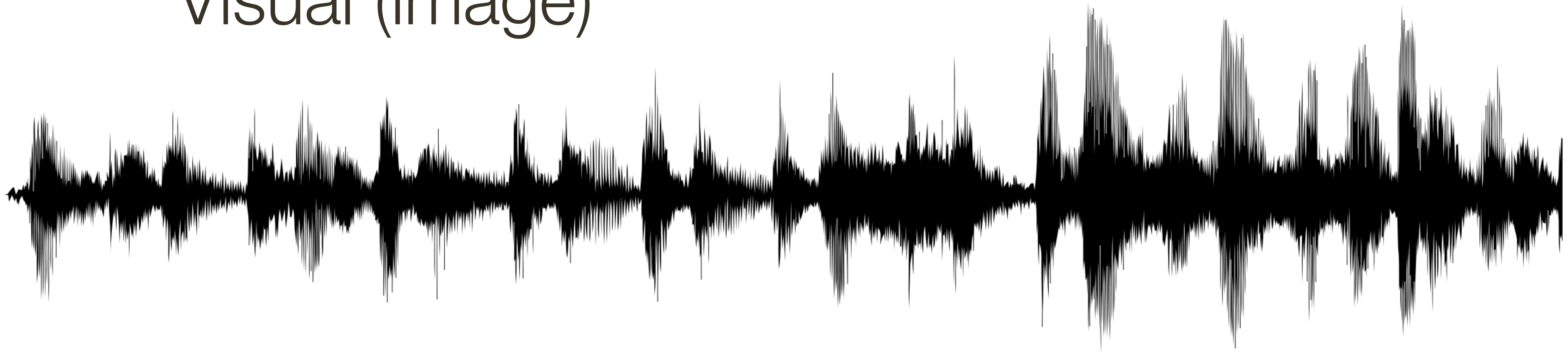
What is **Multi-modal Learning**?

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

Visual (image)

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 it association with the moon it has ties to fertility and seduction. The owl is bird of magic and darkness of prophecy and wisdom.

Natural Language (text)

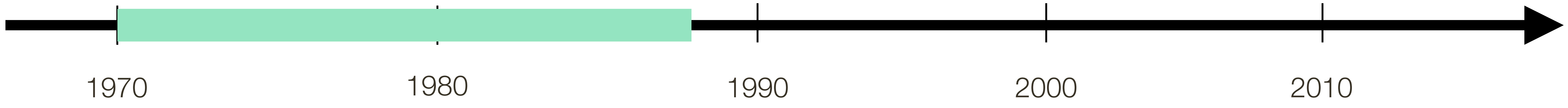


Auditory (voice / sound)

Multimodal Research: Historical Perspective

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

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

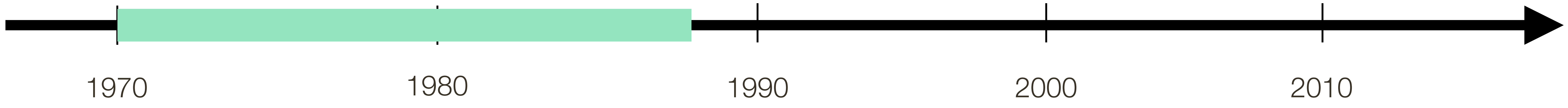


* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

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

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

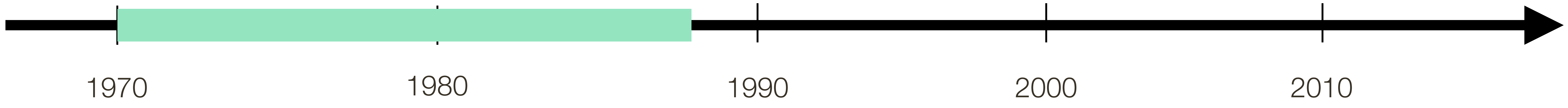


* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

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

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



* Adopted from slides by Louis-Philippe Morency

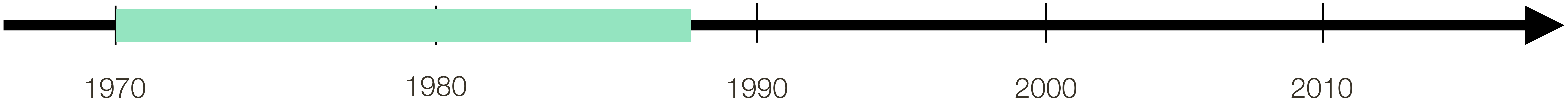
Multimodal Research: Historical Perspective

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

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

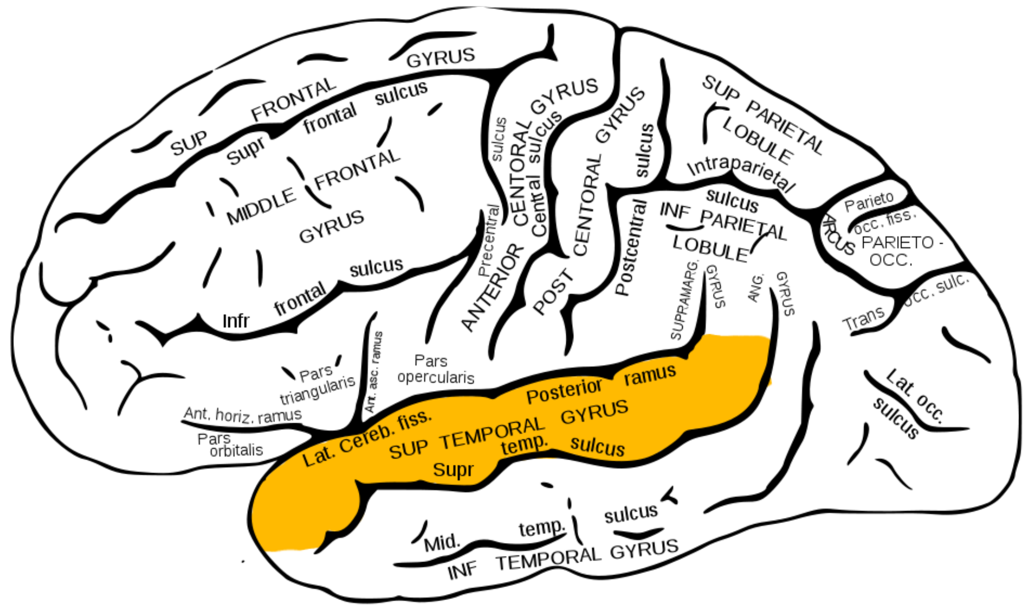
TRIVIA

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



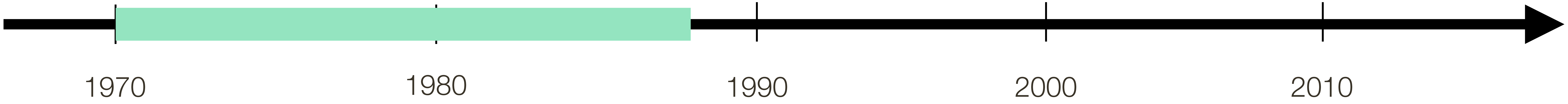
* Adopted from slides by Louis-Philippe Morency

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)

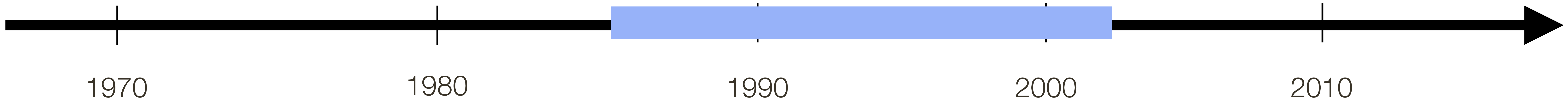


* video credit: **OK Science**

* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Audio-visual speech recognition (motivated by McGurk effect)

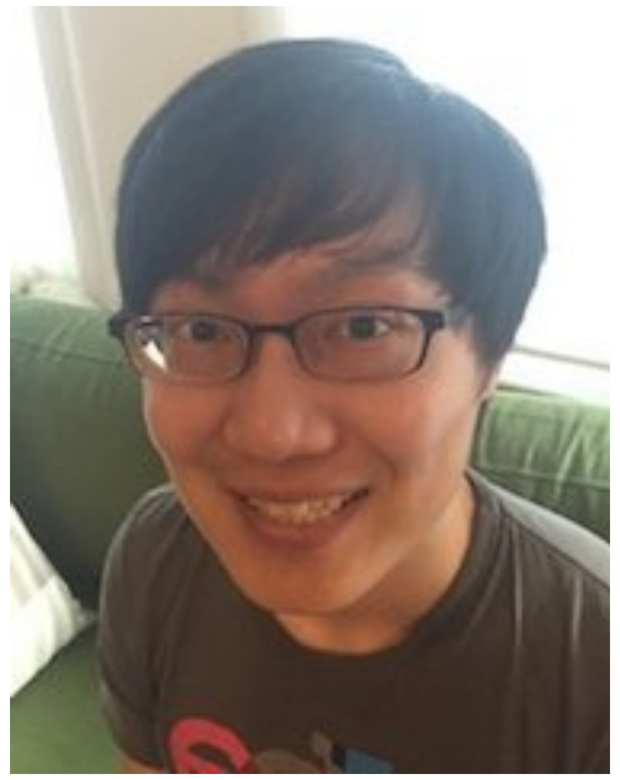


* Adopted from slides by Louis-Philippe Morency

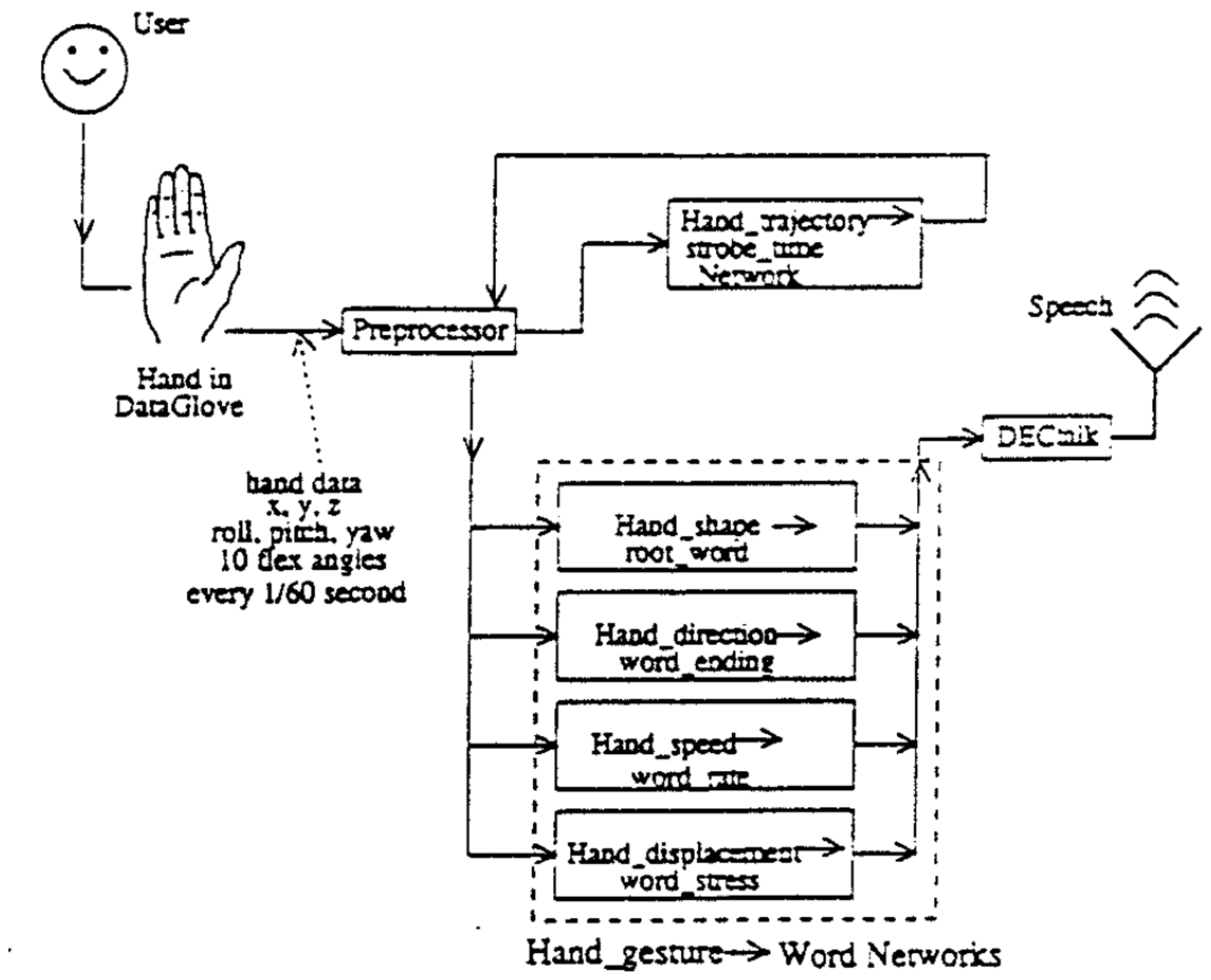
Multimodal Research: Historical Perspective

Audio-visual speech recognition (motivated by McGurk effect)

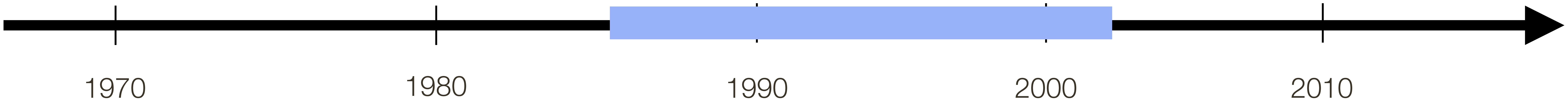
Multi-modal and multi-sensory interfaces



Dongwook Yoon



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



* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Modeling human multi-modal interactions

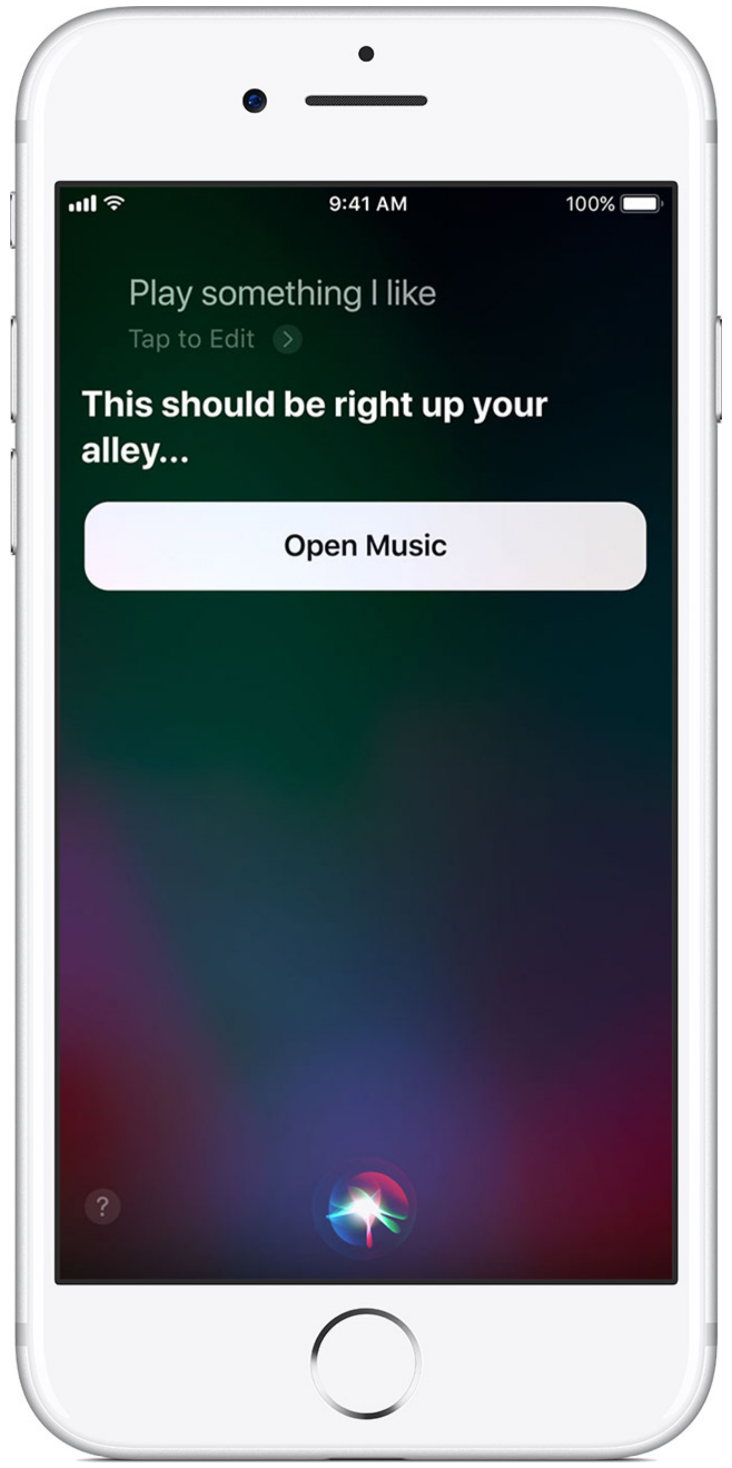
- Huge multi-laboratory efforts

AMI Project [2001-2006, IDIAP]

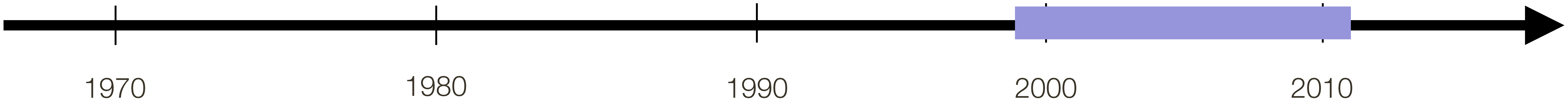
- 100+ hours of meeting recordings
- Synchronized video and audio
- Transcribed and annotated

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



* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Modeling human multi-modal interactions

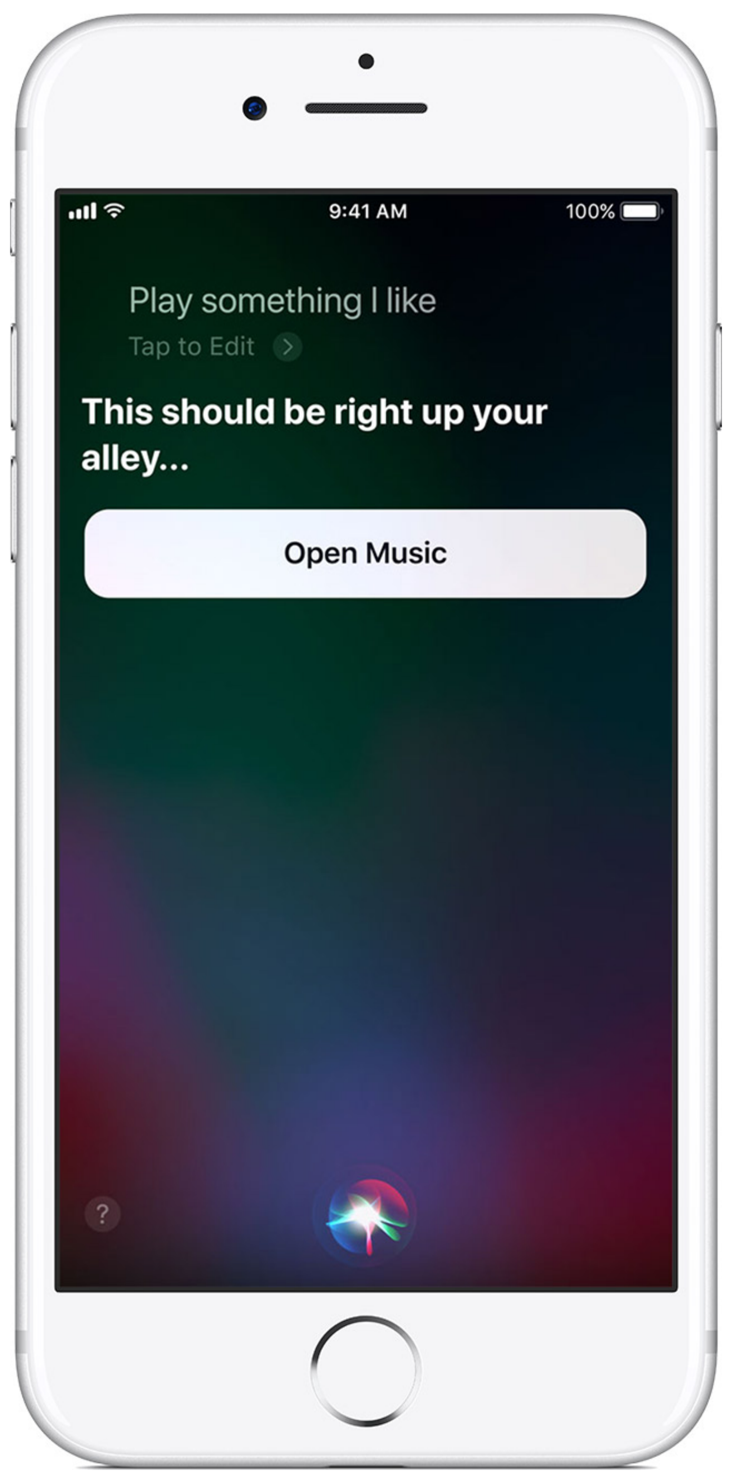
- Huge multi-laboratory efforts

Multimedia information retrieval

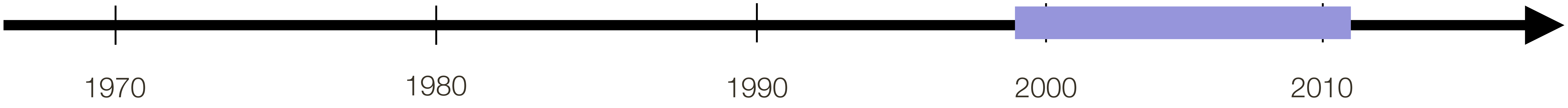
- Lots of challenges and progress

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



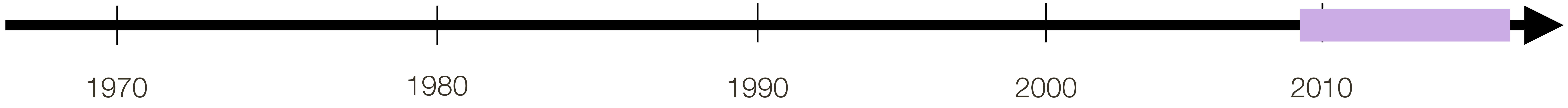
* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

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

- Better performance
- More interesting problems emerging

THIS IS OUR COURSE

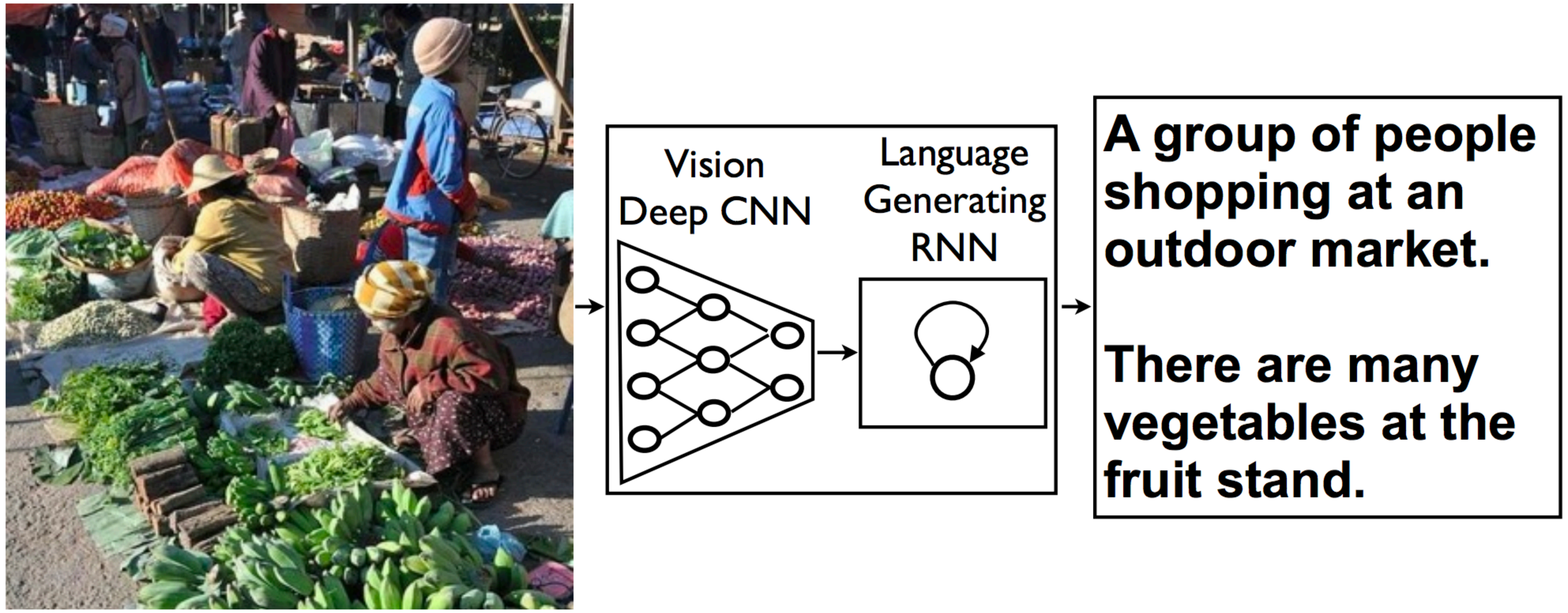


* Adopted from slides by Louis-Philippe Morency

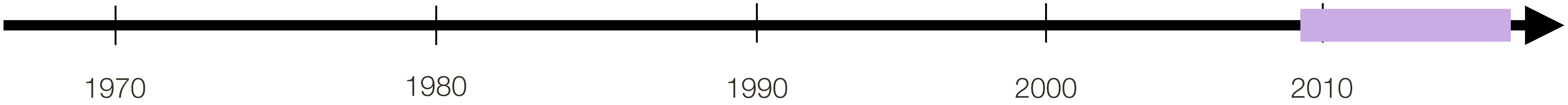
Multimodal Research: Historical Perspective

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

[Vinyals *et al.*, 2015]



Natural language description generation



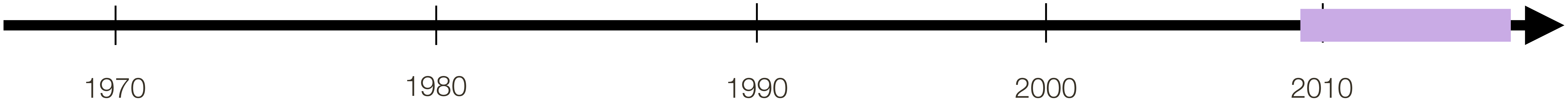
Multimodal Research: Historical Perspective

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)

[Ba *et al.*, ICCV 2015]

Corn Poppy

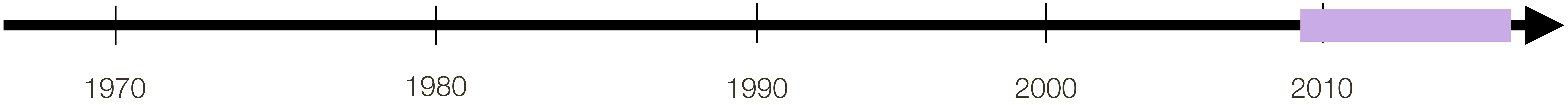
Papaver rhoeas (common names include corn poppy, corn rose, field poppy, Flanders poppy, red poppy, red weed, coquelicot, and, due to its odour, which is said to cause them, as headache and headwark) is a species of flowering plant in the poppy family, Papaveraceae. This poppy, a native of Europe, is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

......

→

Detecting objects based on linguistic descriptions

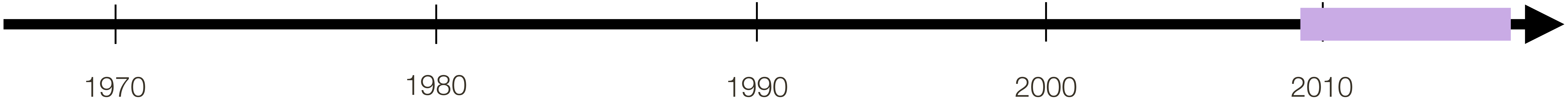


Multimodal Research: Historical Perspective

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

[Zhu et al, ICCV 2015]

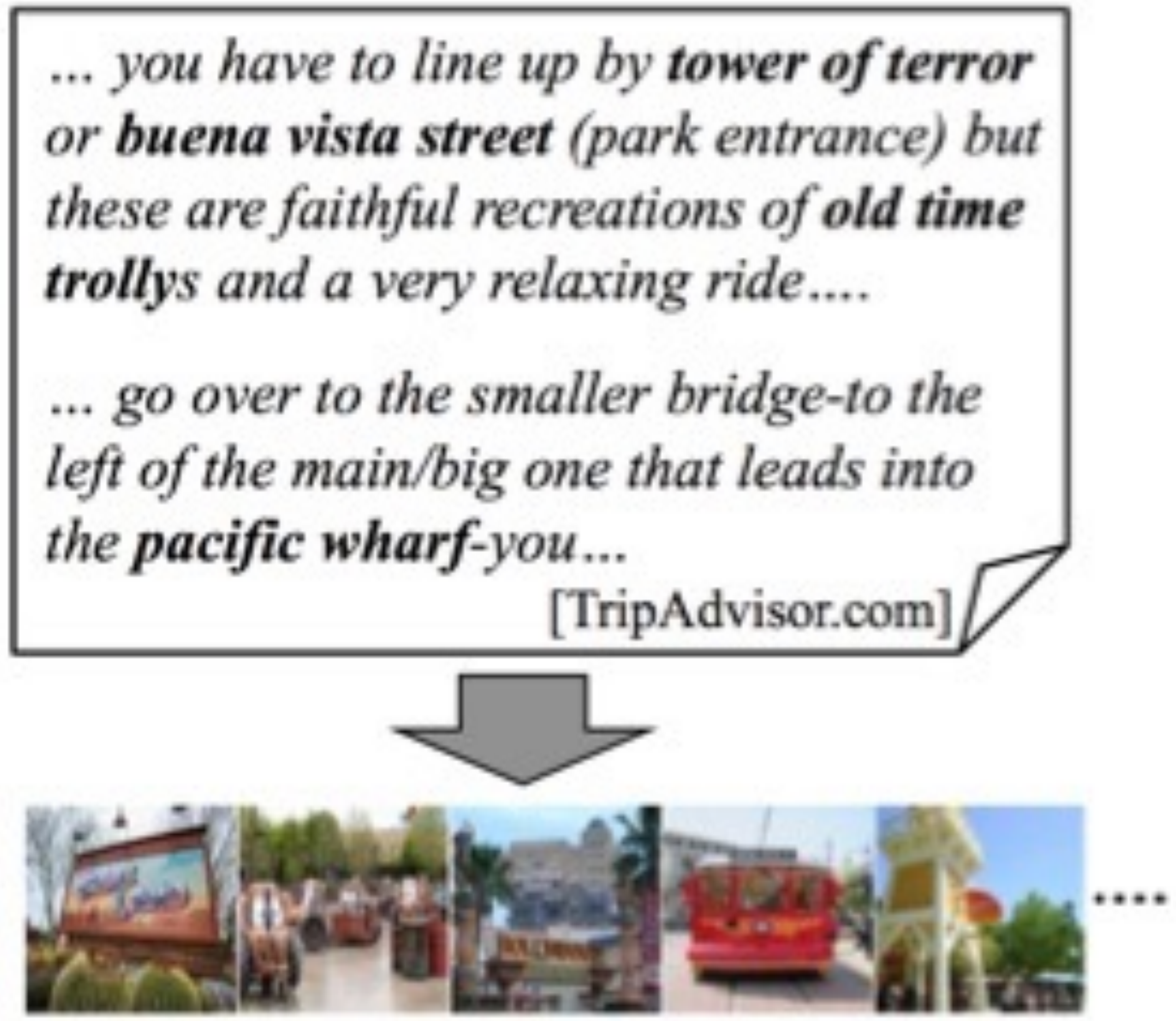
Book-to-Movie alignment



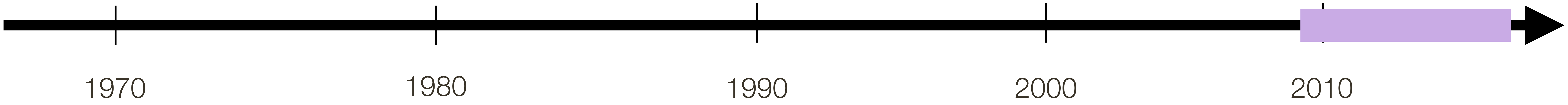
Multimodal Research: Historical Perspective

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

[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*

Q2: *Do you see any people?*

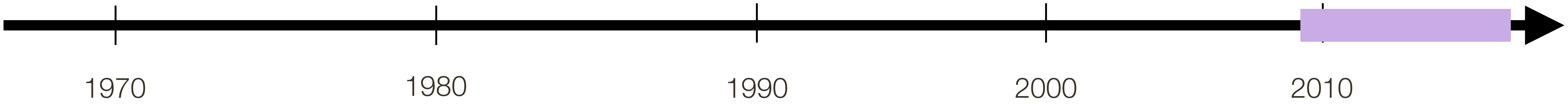
GT answer: *No*
Predicted answer: *No, just the cat*
Rank of GT: *2*

Q3: *What color is the cat ?*

GT answer: *Grey, white, and black*
Predicted answer: *Grey, black and white*
Rank of GT: *6*

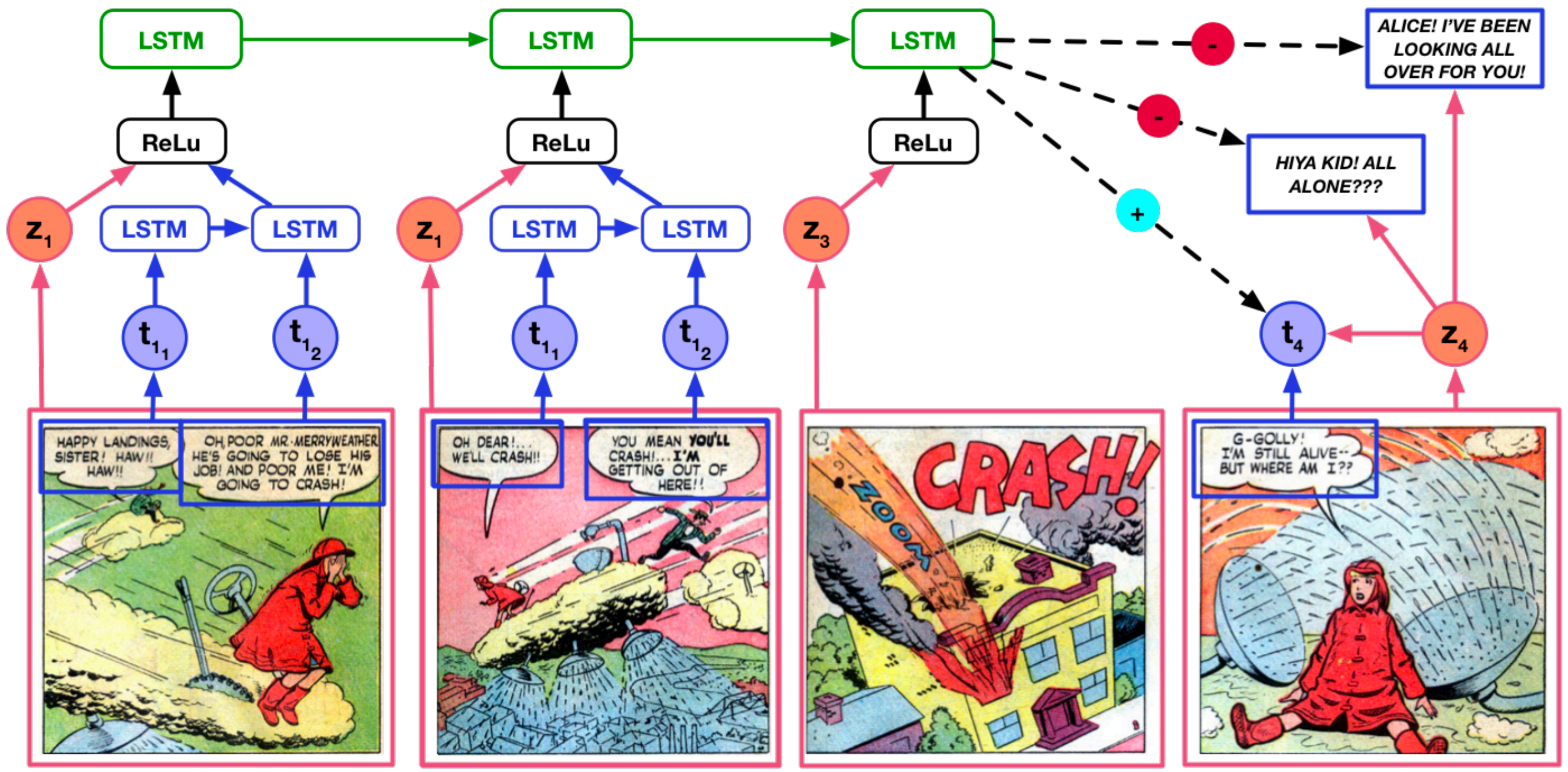
[Seo *et al.*, NIPS 2017]

Visual question answering / dialog



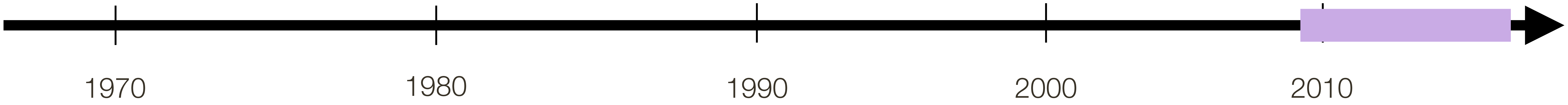
Multimodal Research: Historical Perspective

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



Narrative plot understanding

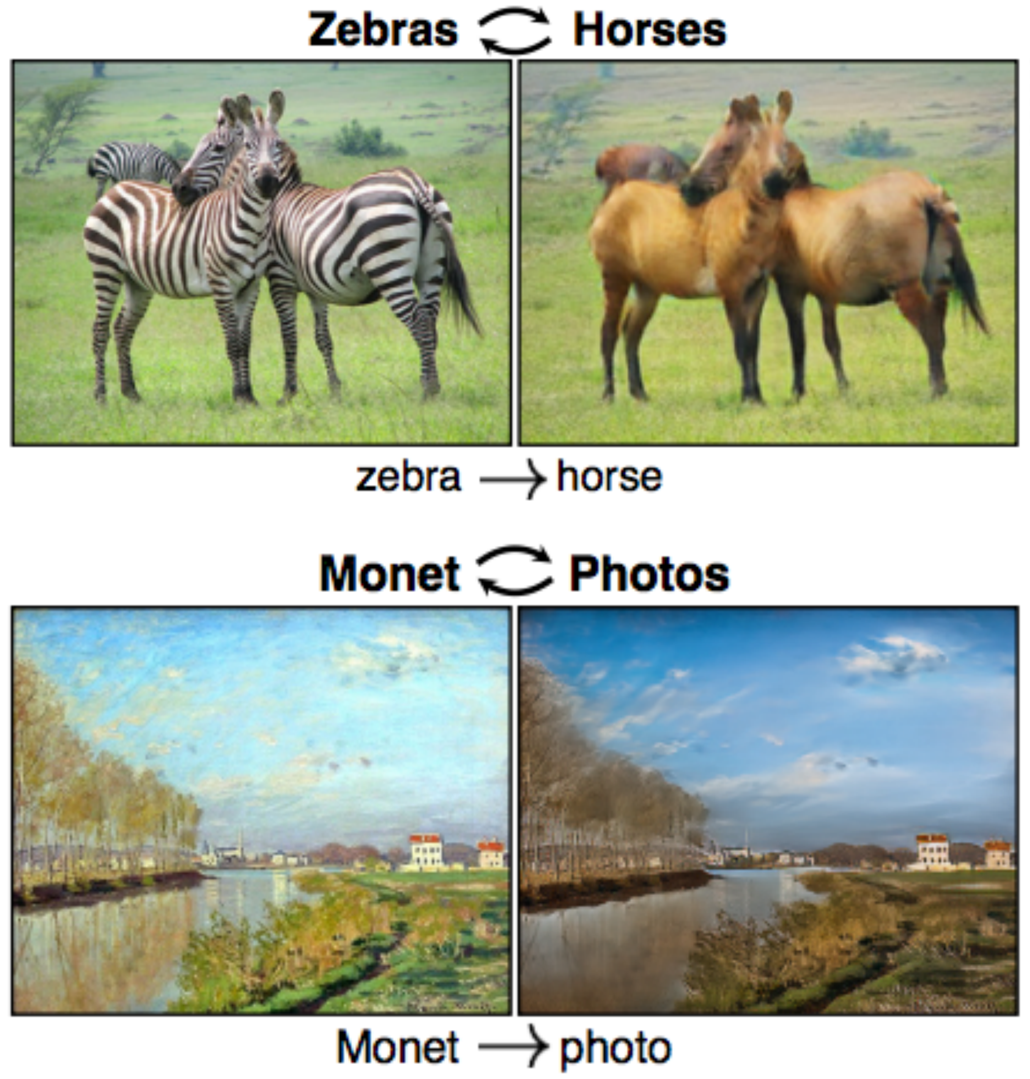
[Iyyer *et al.*, CVPR 2017]



Multimodal Research: Historical Perspective

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

[Zhu *et al.*, ICCV 2017]



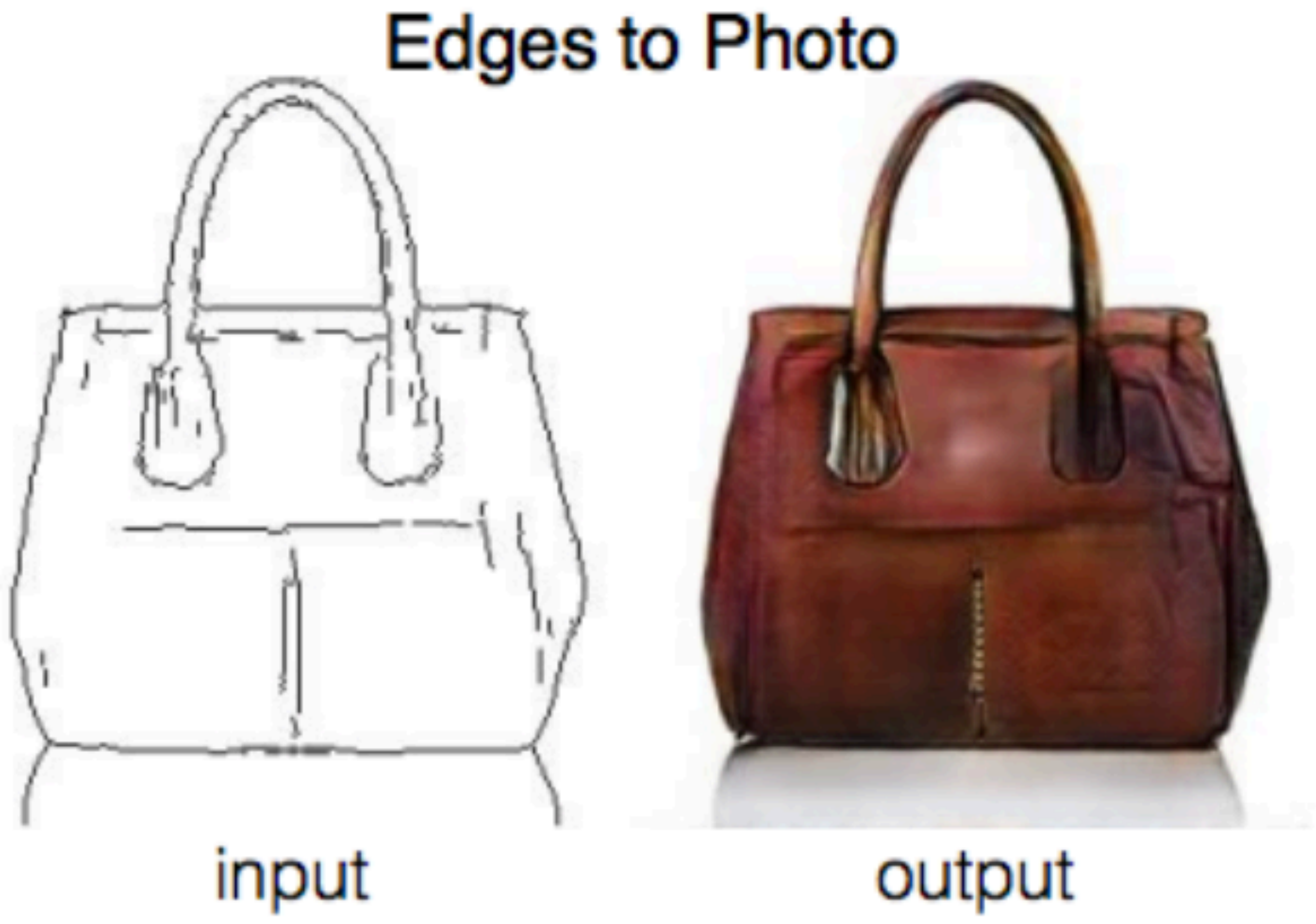
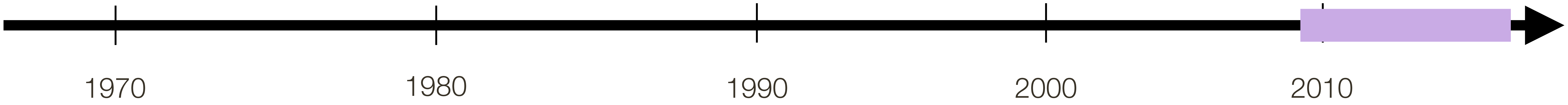


Image-to-image translation

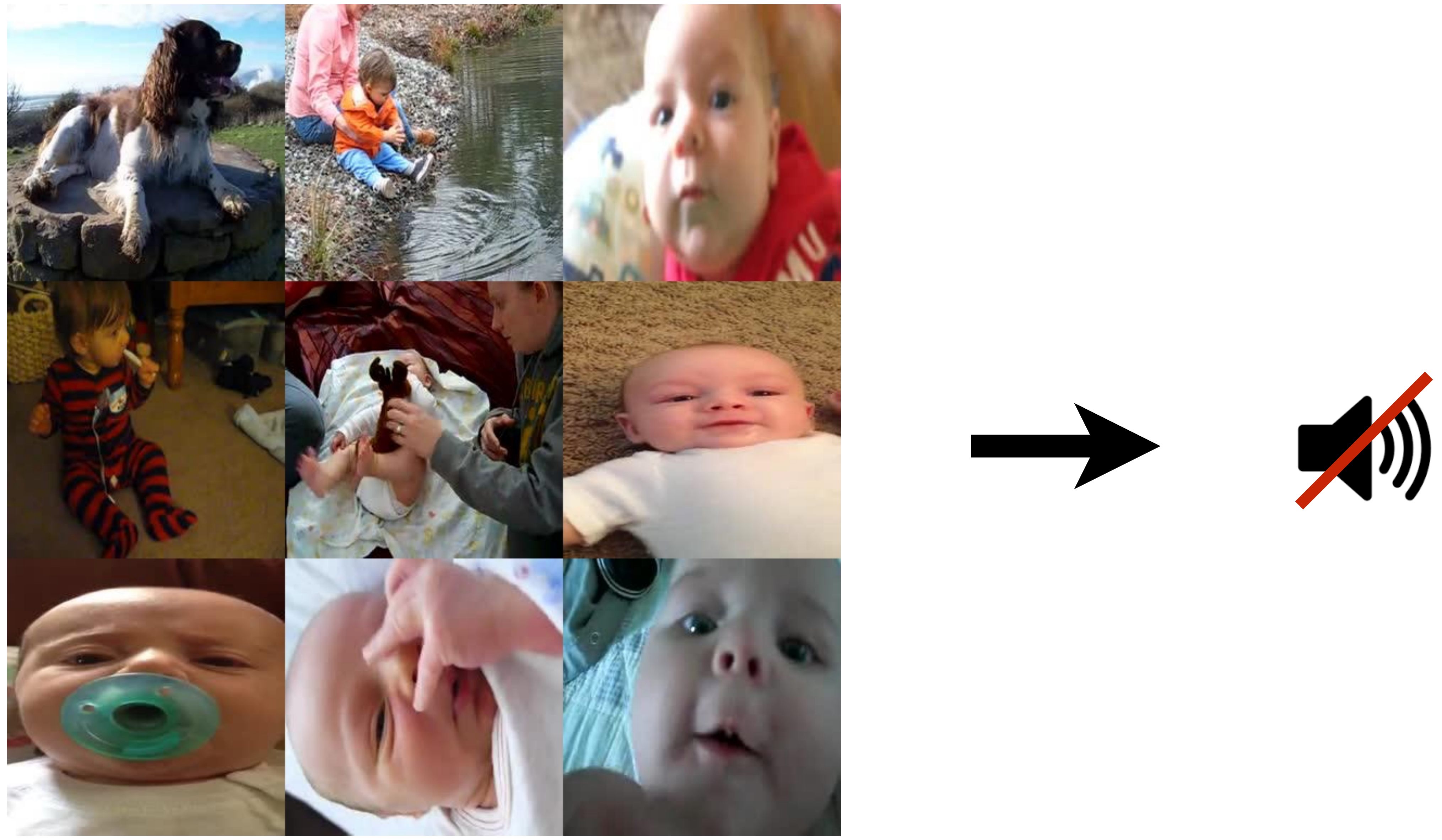
[Isola *et al.*, CVPR 2017]



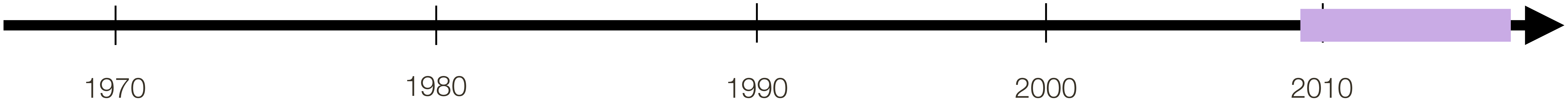
Multimodal Research: Historical Perspective

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

[Iyyer *et al.*, NIPS 2016]



Video-to-Audio translation



Key Challenges of Multimodal Learning

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

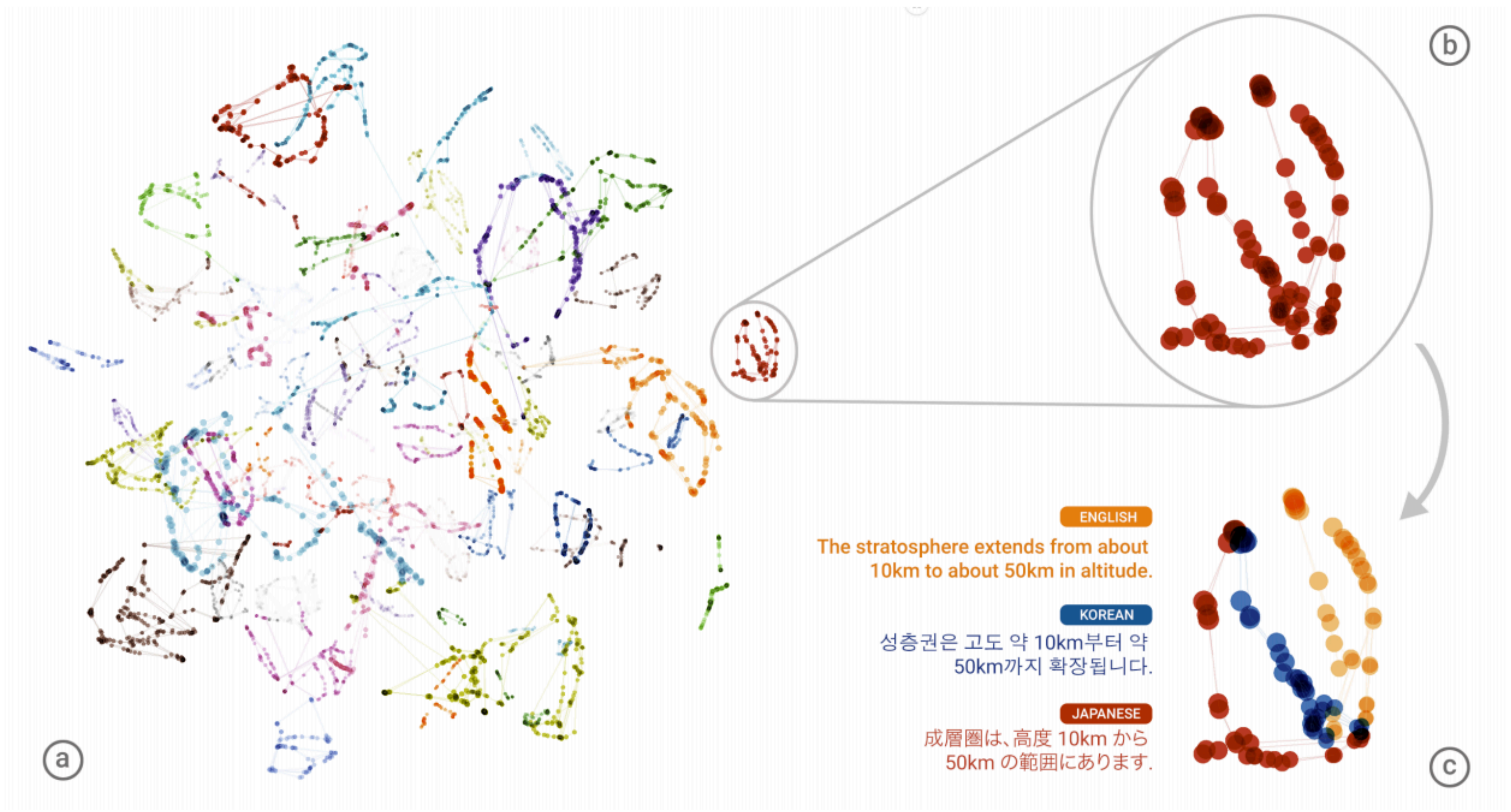
What's another phrase for "**representation learning**"?

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]

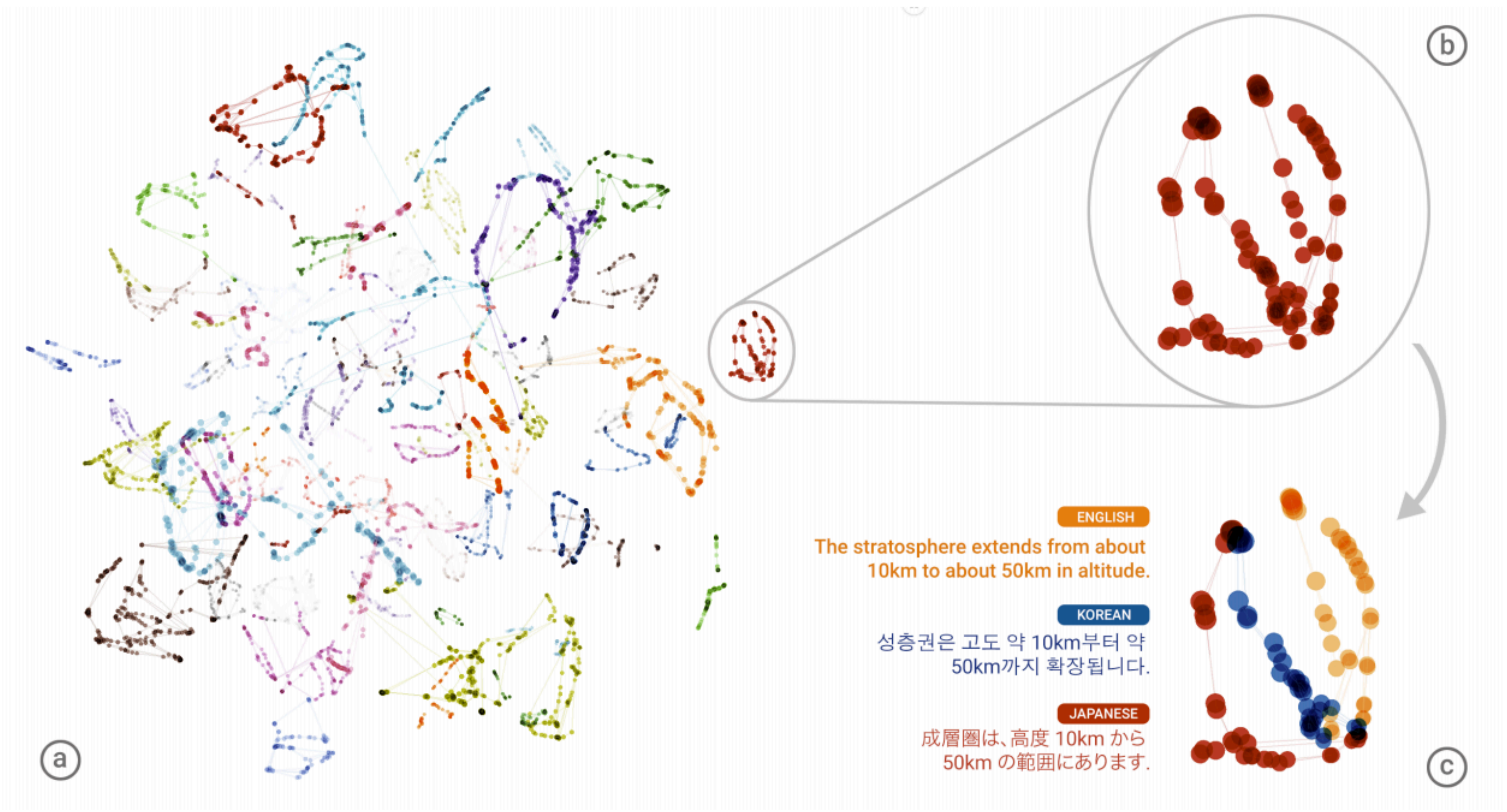
Key Challenges of Multimodal Learning

- 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]

Objectives of the course

- Acquire **fundamentals and background** that would allow one to follow research in Computer Vision and on intersection of Vision + Language

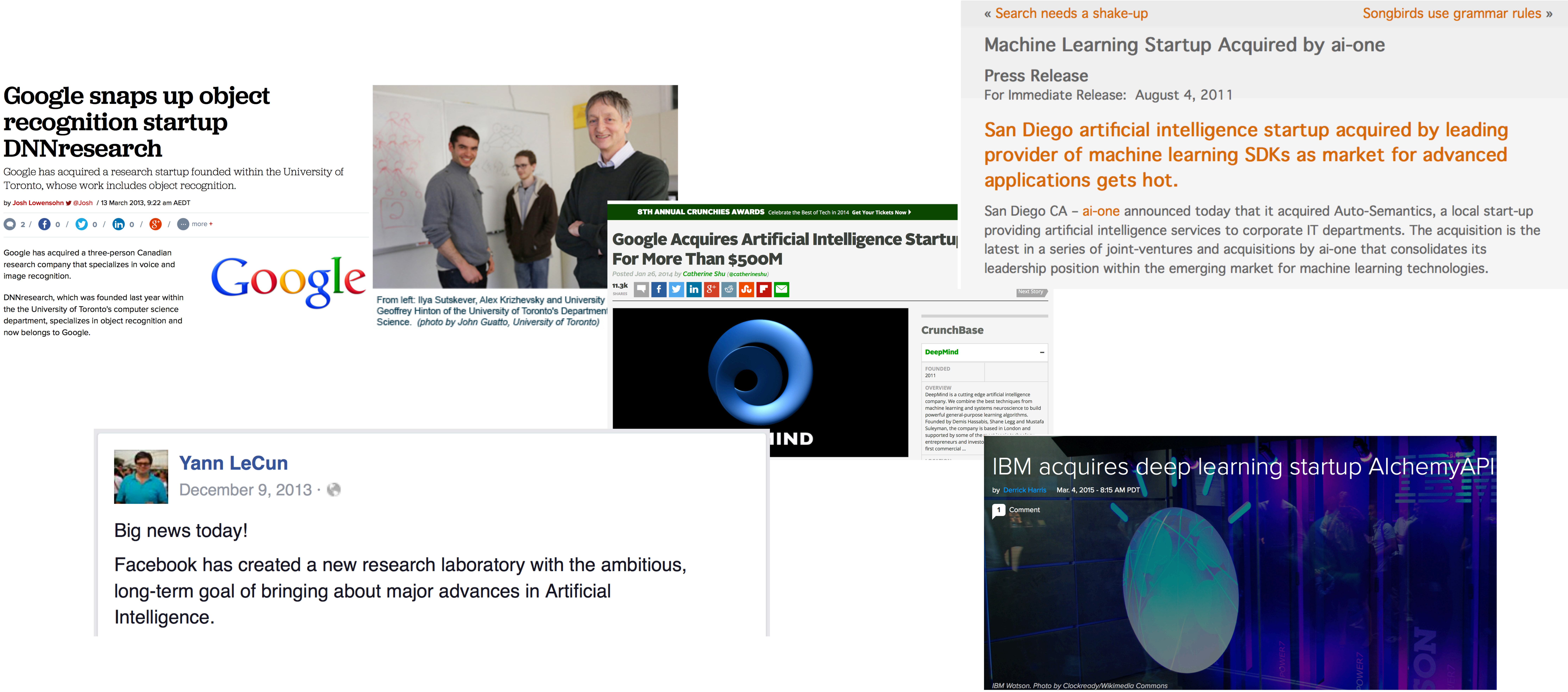
- Ability to **design, build and apply deep learning architectures** for multi-modal problems (Vision + Language in particular)

- 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



* slide from Dhruv Batra

Clever Hans

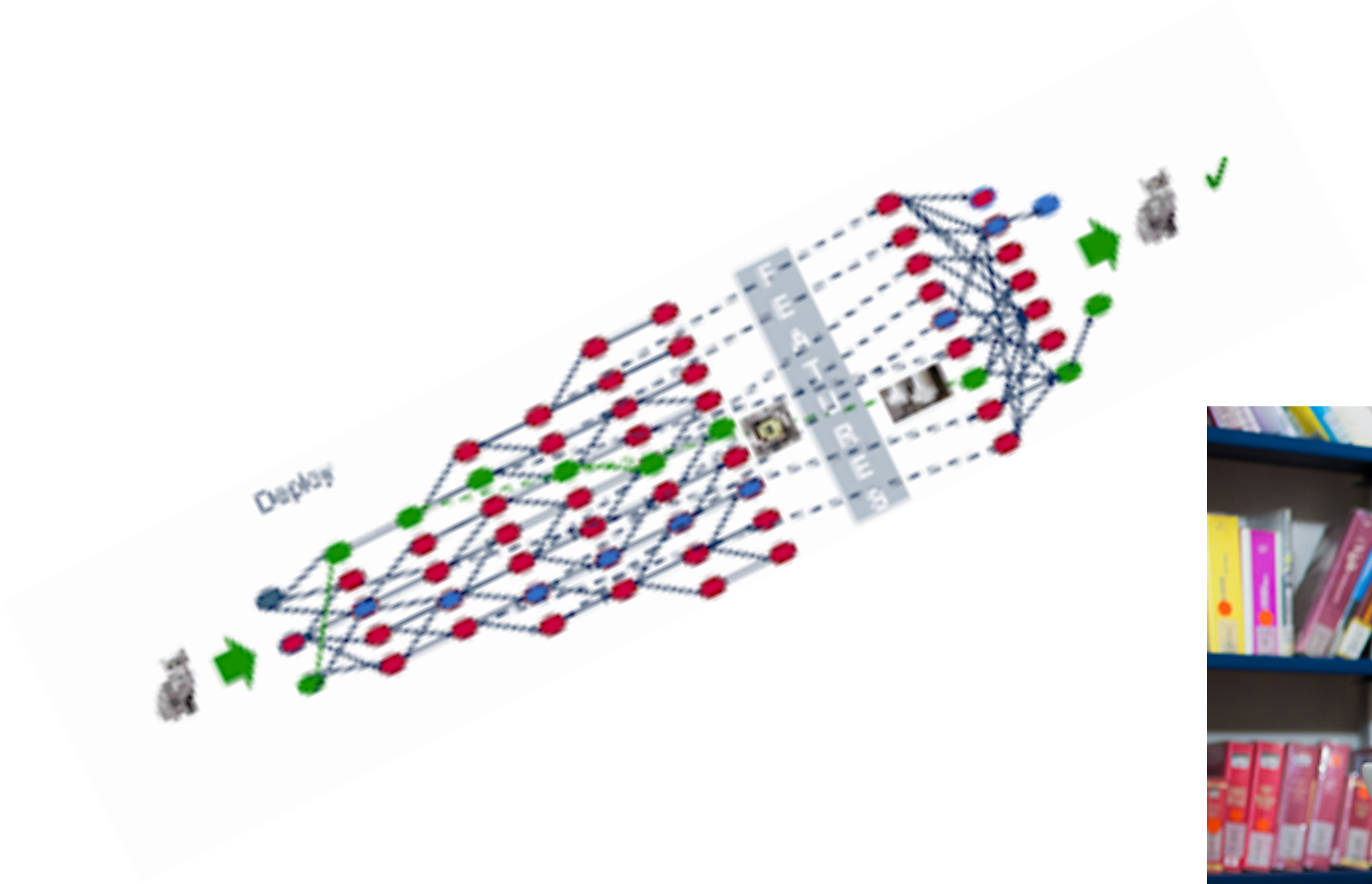
Hans could get 89% of the math questions right

Clever Hans

The horse was **smart**, just not in the way van Osten thought!

Hans could get 89% of the math questions right

Clever DNN



Visual **Question Answering**

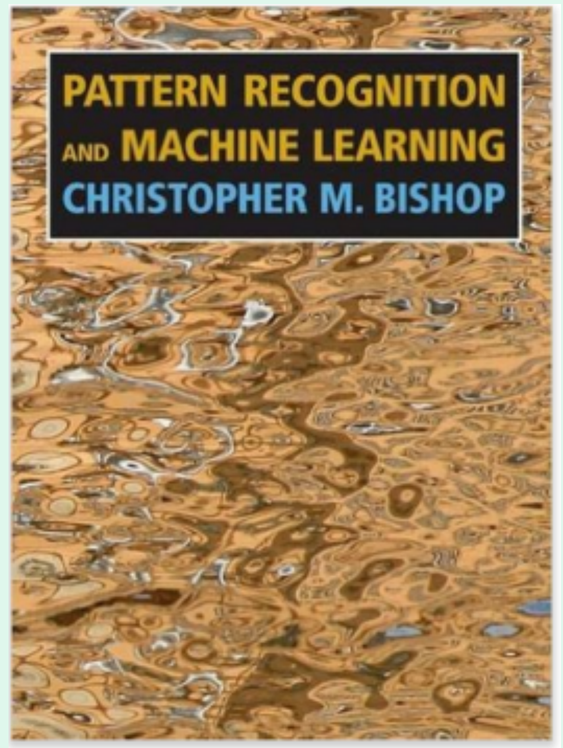
AI agent

Yes

Is there zebra climbing the tree?

Pre-requisites

Computer Science

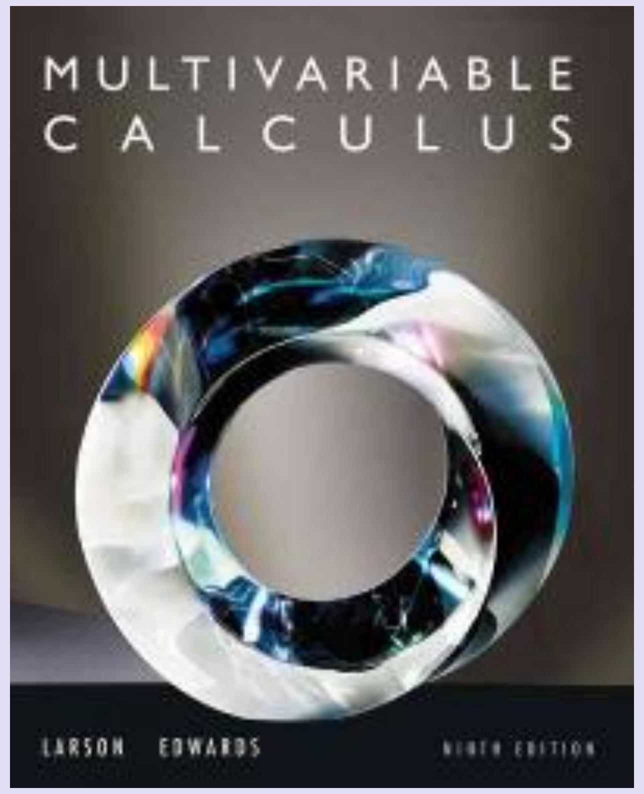


CPSC 340
(or equivalent)

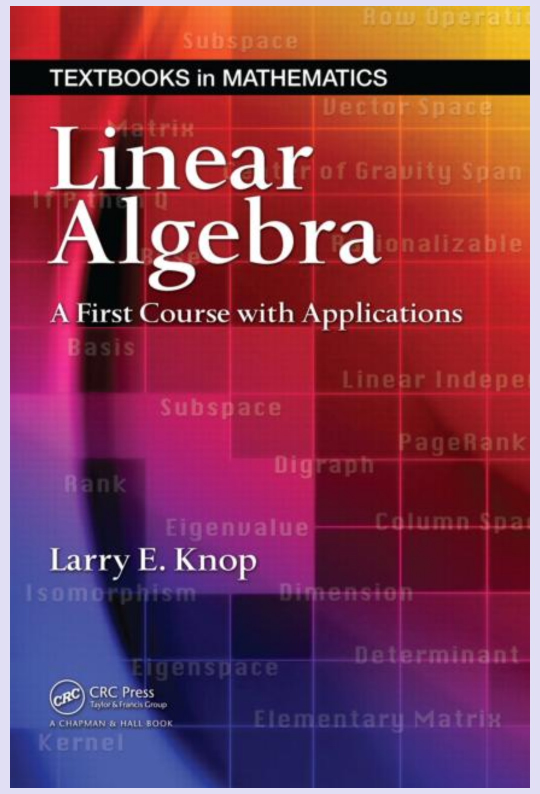
python

Needed for
Assignments

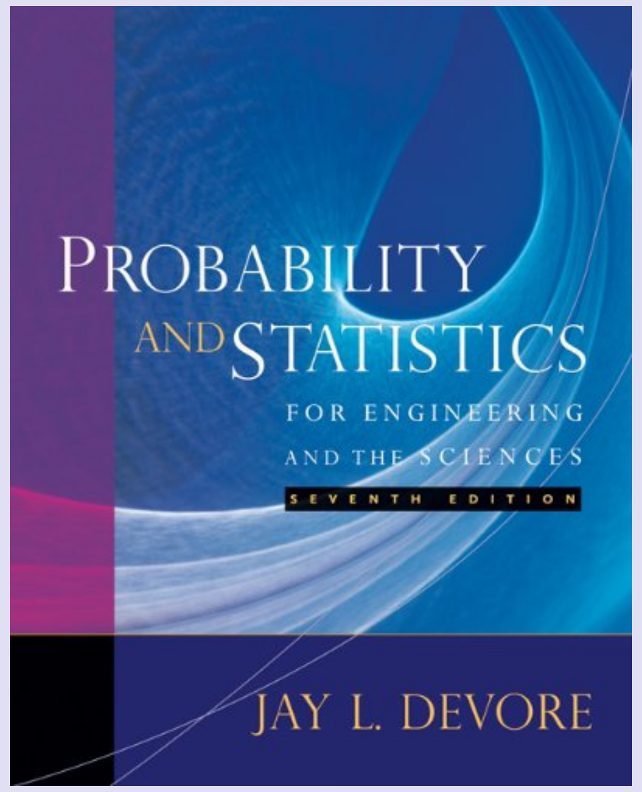
Mathematics



Calculus



Linear Algebra



Statistics

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

Additional **Requirement**

You will be given credits to use

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.

or use your own …

Nvidia GTX 1060 (with 6GB RAM) or above

Course **structure**

Approximately 50% of course will consists of lectures and optional readings

Remaining 50% is reading  and presentation of curated research papers on relevant topics

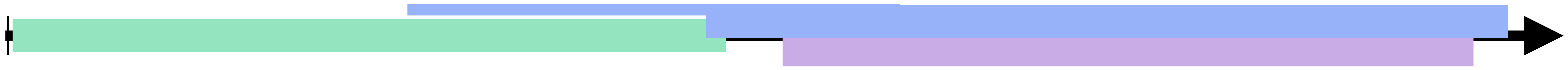
4 programming assignments

Final (individual or **group**) project

Grading Criteria

- **Assignments** (programming) — 30% (total)
- **Research papers** — 20%
- Group **project** — 50%

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



Assignments (4 assignments and 30% of grade total)

- Assignment 1: **Neural Network Introduction** (5%) — python™

Assignments all use **Python Jupiter Notebooks**, use `Canvas` to hand everything in. Assignments always due at **5pm PST** on due date.

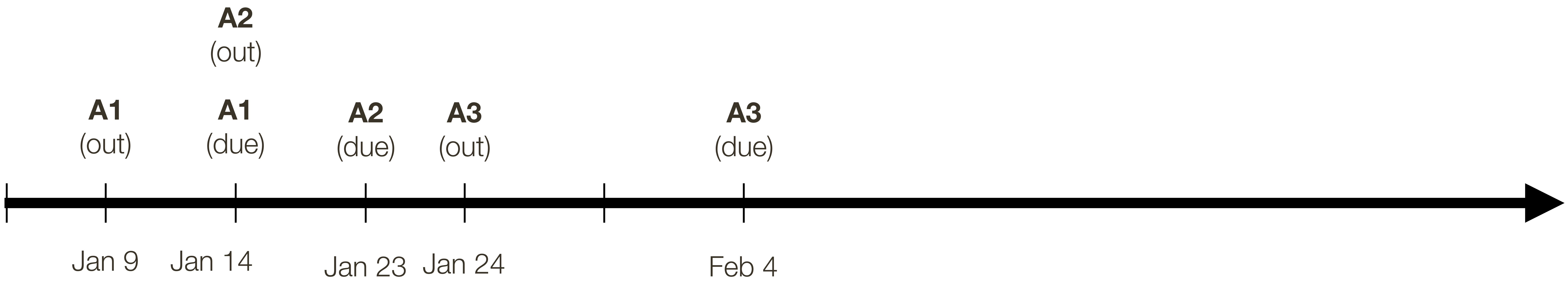
Assignments (4 assignments and 30% of grade total)

- Assignment 1: **Neural Network Introduction** (5%) — python™
- Assignment 2: **Convolutional Neural Networks** (5%) — PYTORCH

Assignments all use **Python Jupiter Notebooks**, use `Canvas` to hand everything in. Assignments always due at **5pm PST** on due date.

Assignments (4 assignments and 30% of grade total)

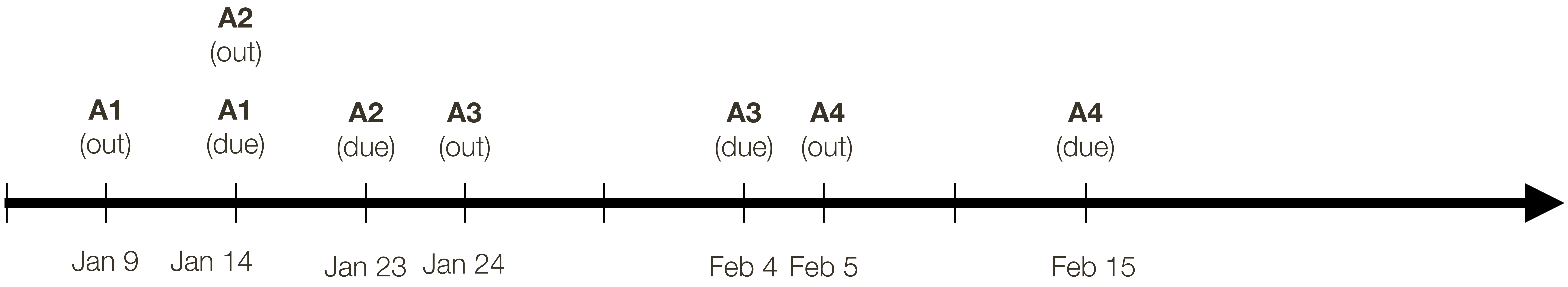
- Assignment 1: **Neural Network Introduction** (5%) — python™
- Assignment 2: **Convolutional Neural Networks** (5%) — PYTORCH
- Assignment 3: RNN **Language Modeling** (10%) — PYTORCH



Assignments all use **Python Jupiter Notebooks**, use `Canvas` to hand everything in. Assignments always due at **5pm PST** on due date.

Assignments (4 assignments and 30% of grade total)

- Assignment 1: **Neural Network Introduction** (5%) — python™
- Assignment 2: **Convolutional Neural Networks** (5%) — PYTORCH
- Assignment 3: RNN **Language Modeling** (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 **5pm PST** on due date.

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) [7.5%]
- 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 in person at least 2 days before your scheduled presentation** for me to provide feedback.
- It is your responsibility to schedule these meetings.
- You will also need to argue against one of the papers [2.5%]

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) [7.5%]
- 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 in person at least 2 days before your scheduled presentation** for me to provide feedback.
- It is your responsibility to schedule these meetings.
- You will also need to argue against one of the papers [2.5%]

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 Piazza

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** (50% 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): 35%

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**!

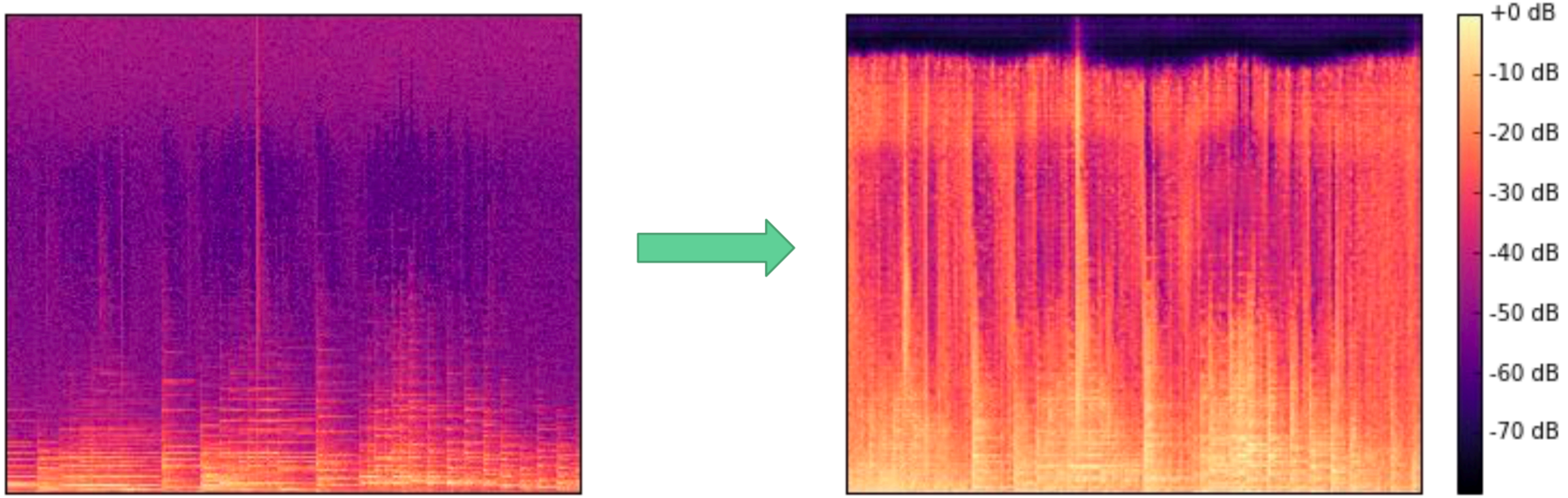
Project Example: Dreaming of Music

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

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

Classical

“Classical to Rock”



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

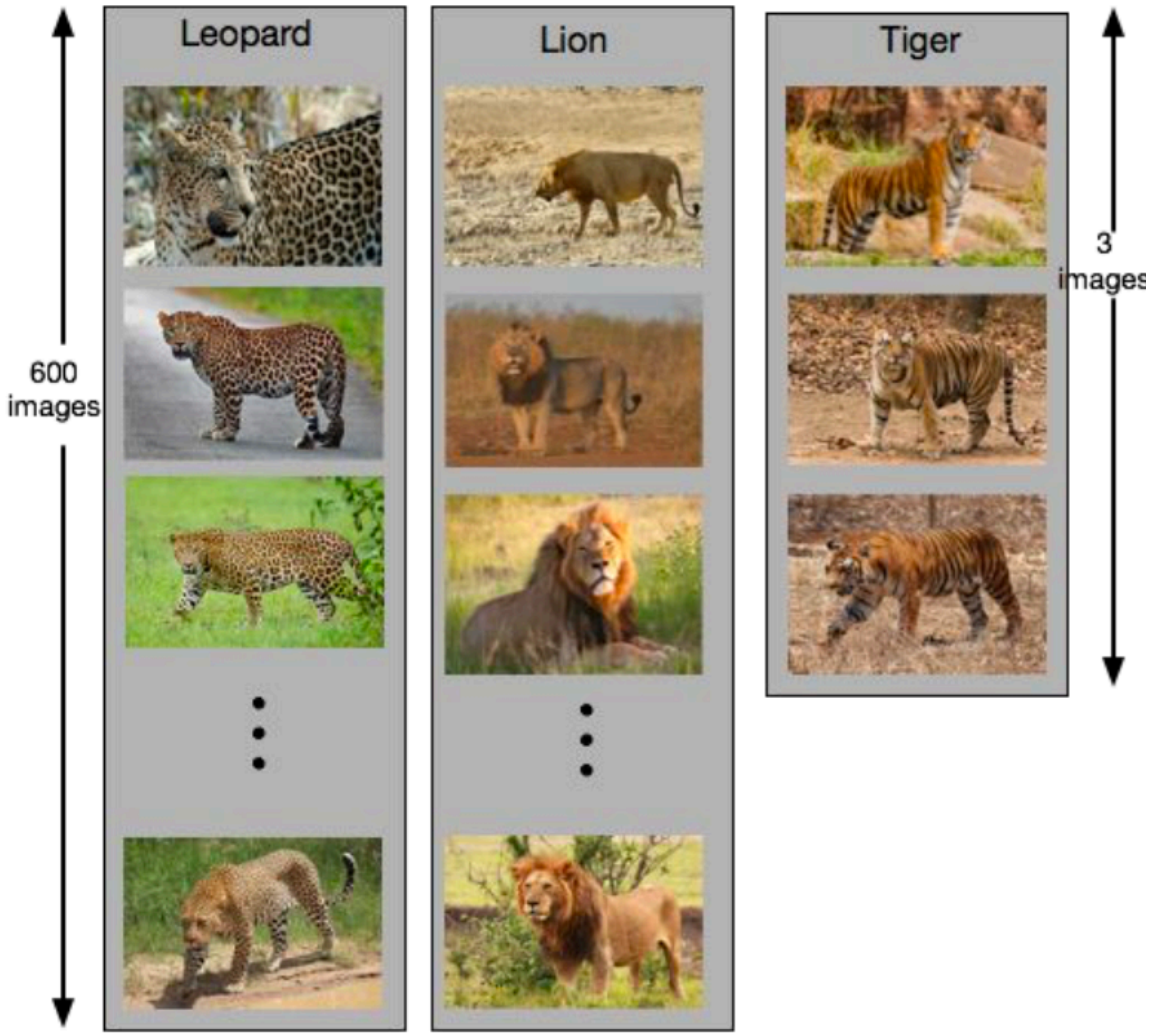
by Michael and Marjan

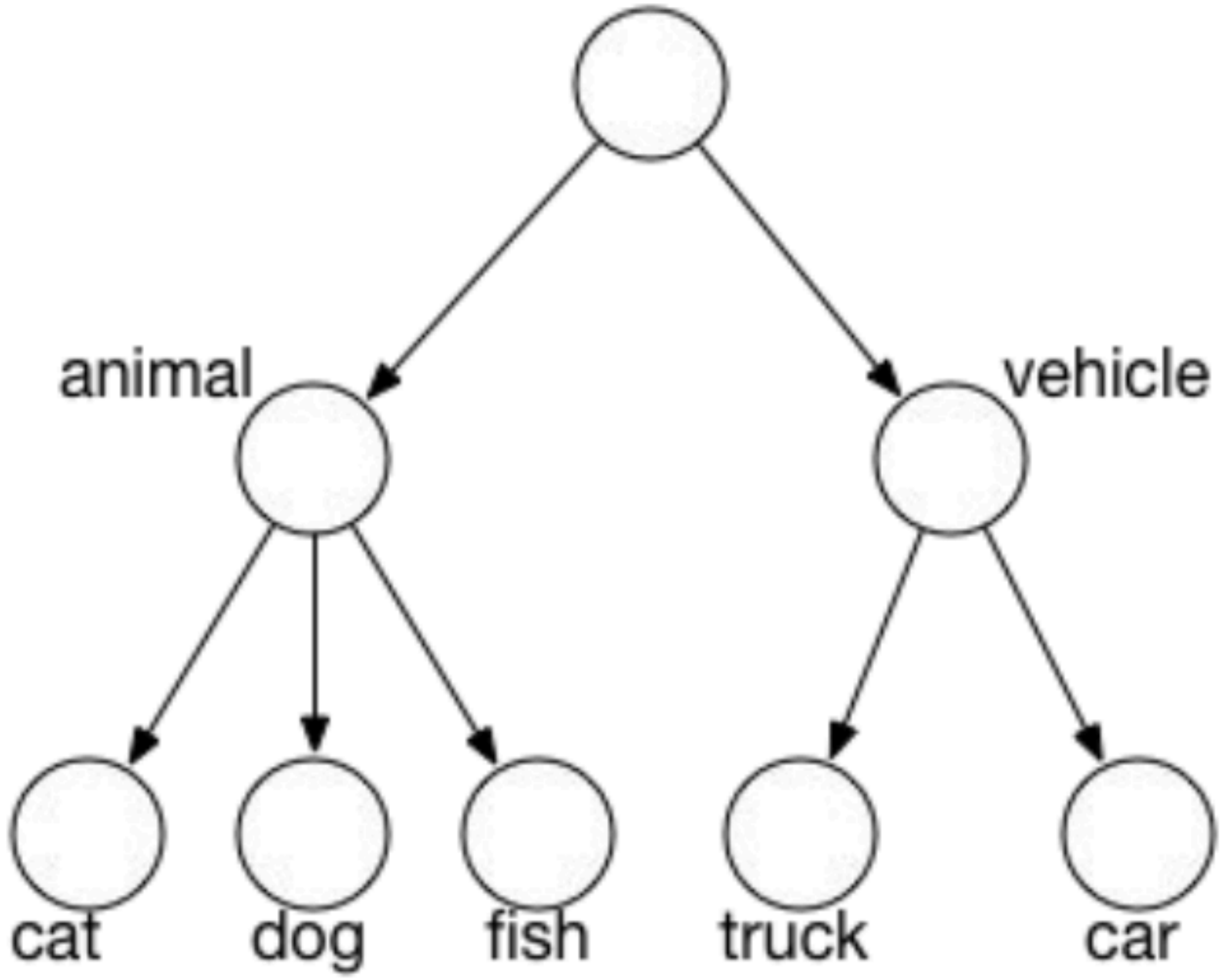
Bayesian Neural Network and variational inference for detecting and analyzing adversarial attacks

Project Example: Classification with Tree Priors

by Saeid Naderiparizi and Setareh Cohan

Classification with few samples using transfer learning techniques

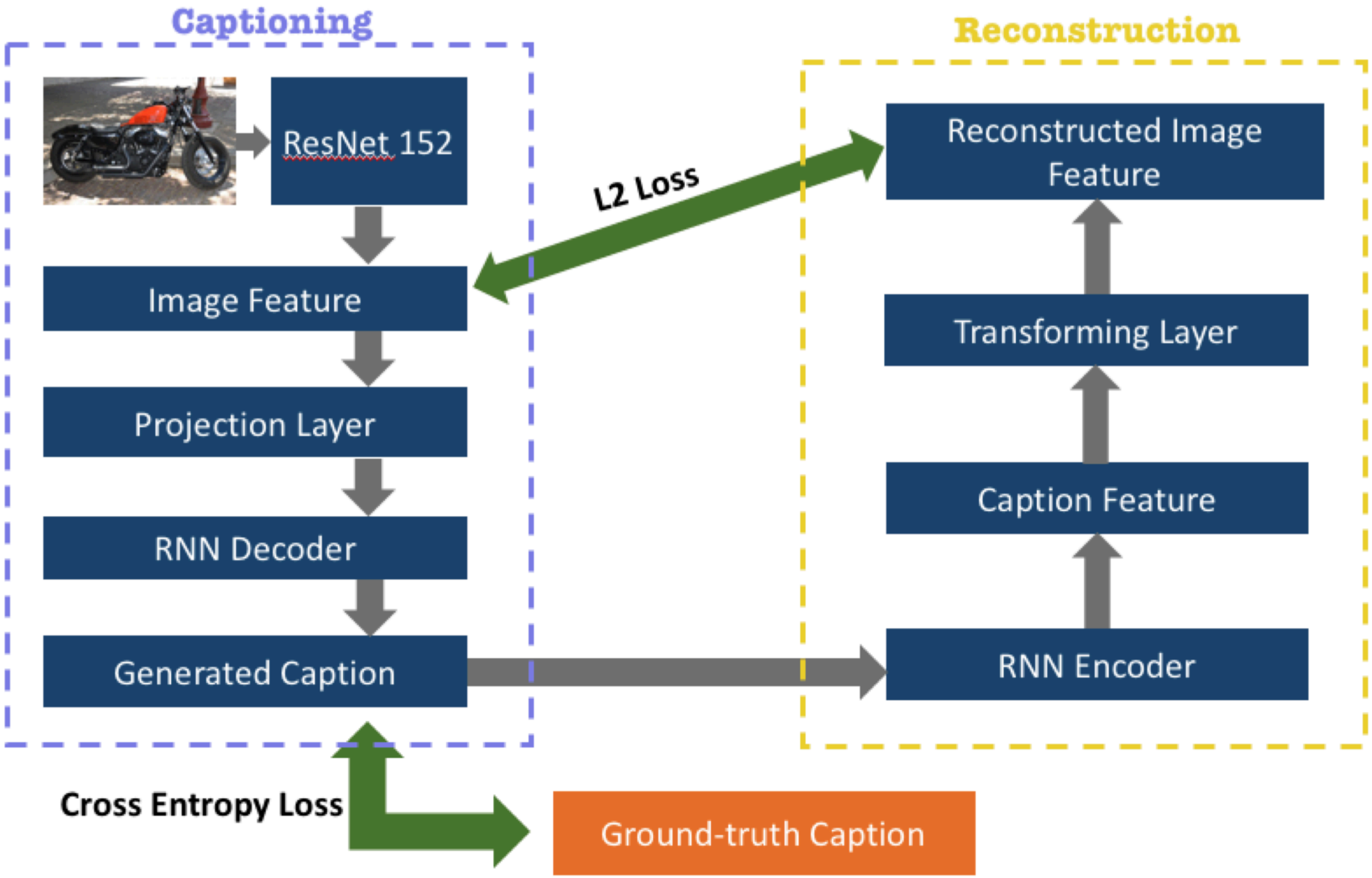




Project Example: Semi-supervised Image Captioning

by Bicheng Xu, Weirui Kong, Jiaxuan Chen

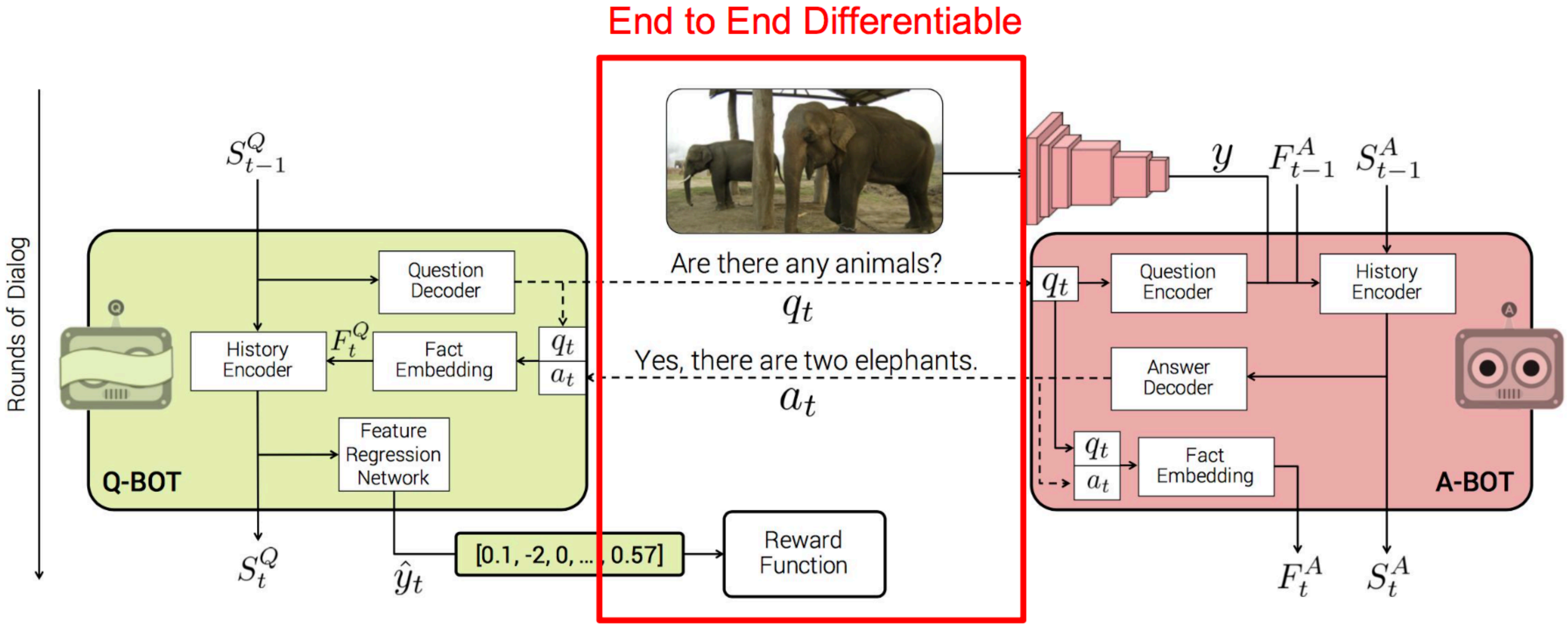
Effective use of unlabeled data during training of an image captioning network



Project Example: Visual Question Answering

by Siddhesh Khandelwal, Mohit Bajaj, Gursimran Singh

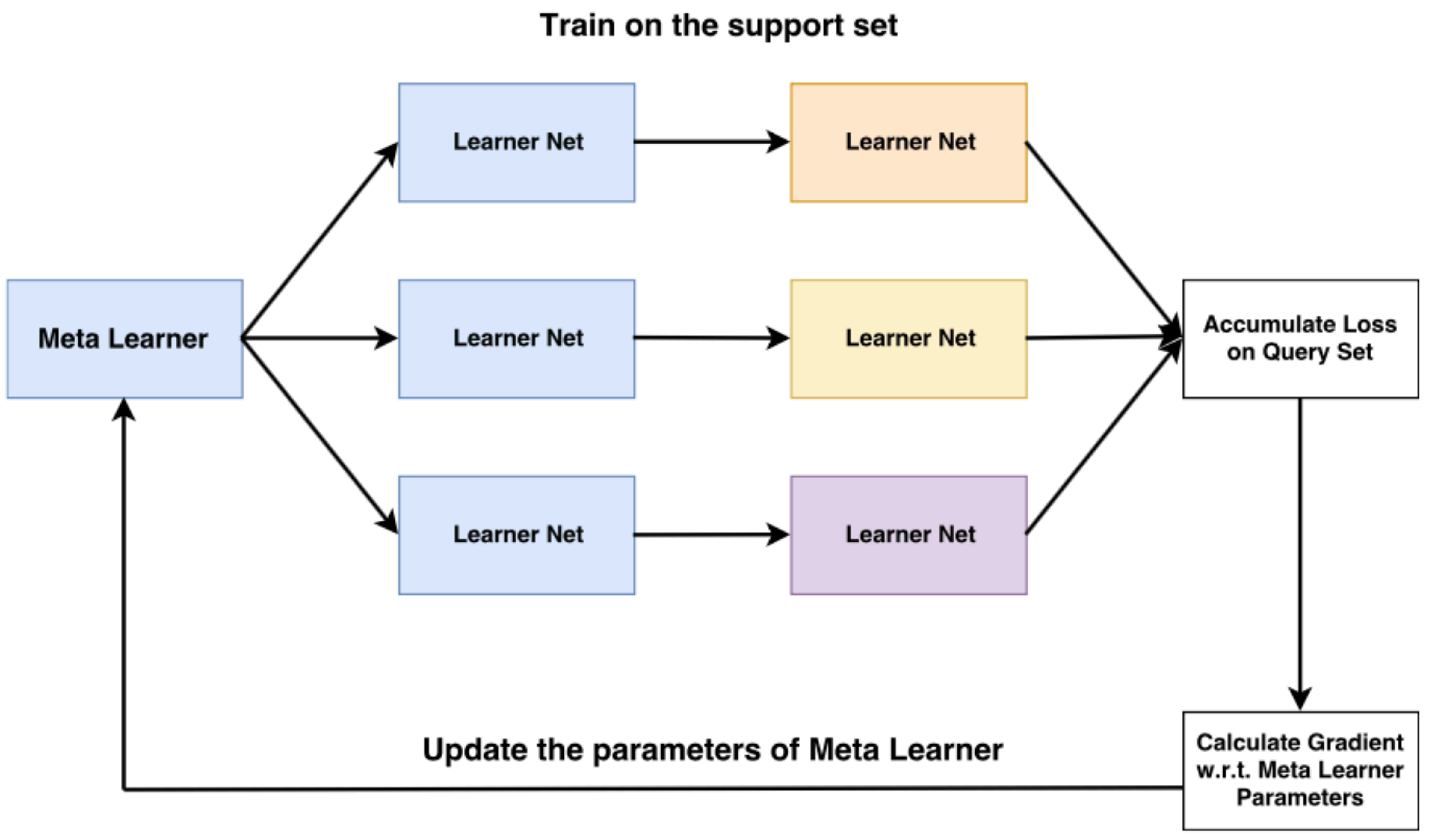
Improve interaction between two agents
— End-to-end differentiability
— Discriminator for human-like questions

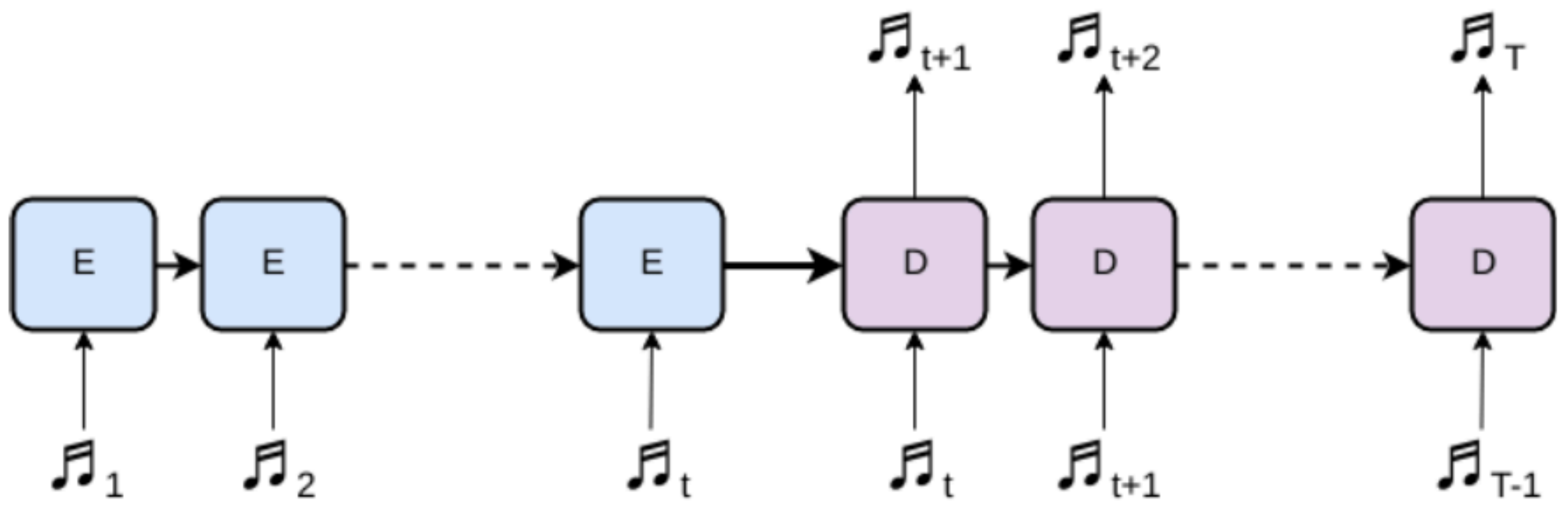


Das, Abhishek, et al. "Learning cooperative visual dialog agents with deep reinforcement learning." *arXiv preprint arXiv:1703.06585* (2017).

Project Example: Few Shot MIDI Music Generation

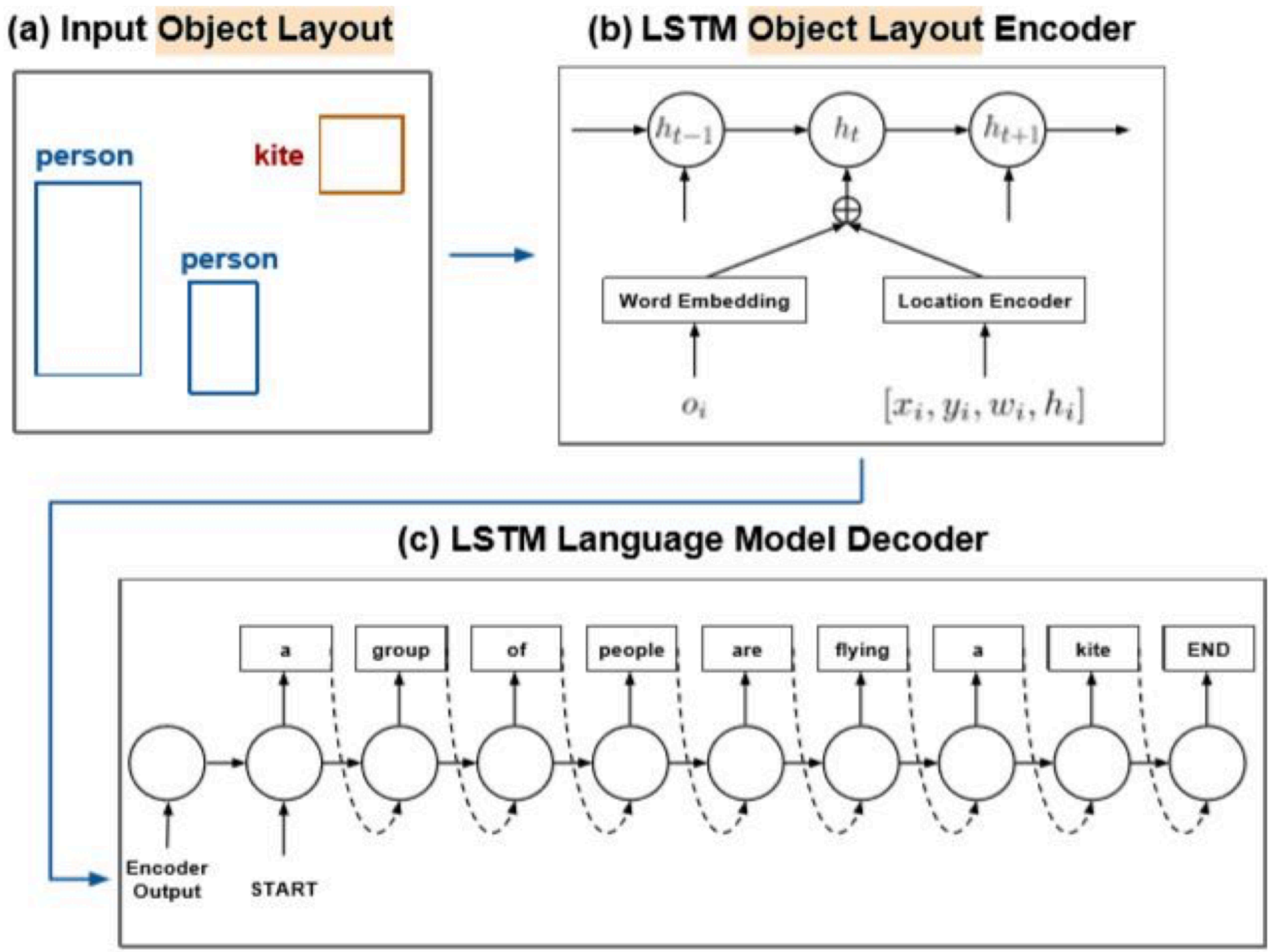
by Ben, Suhail, Anand





Project Example: Visually Descriptive Language from Layout

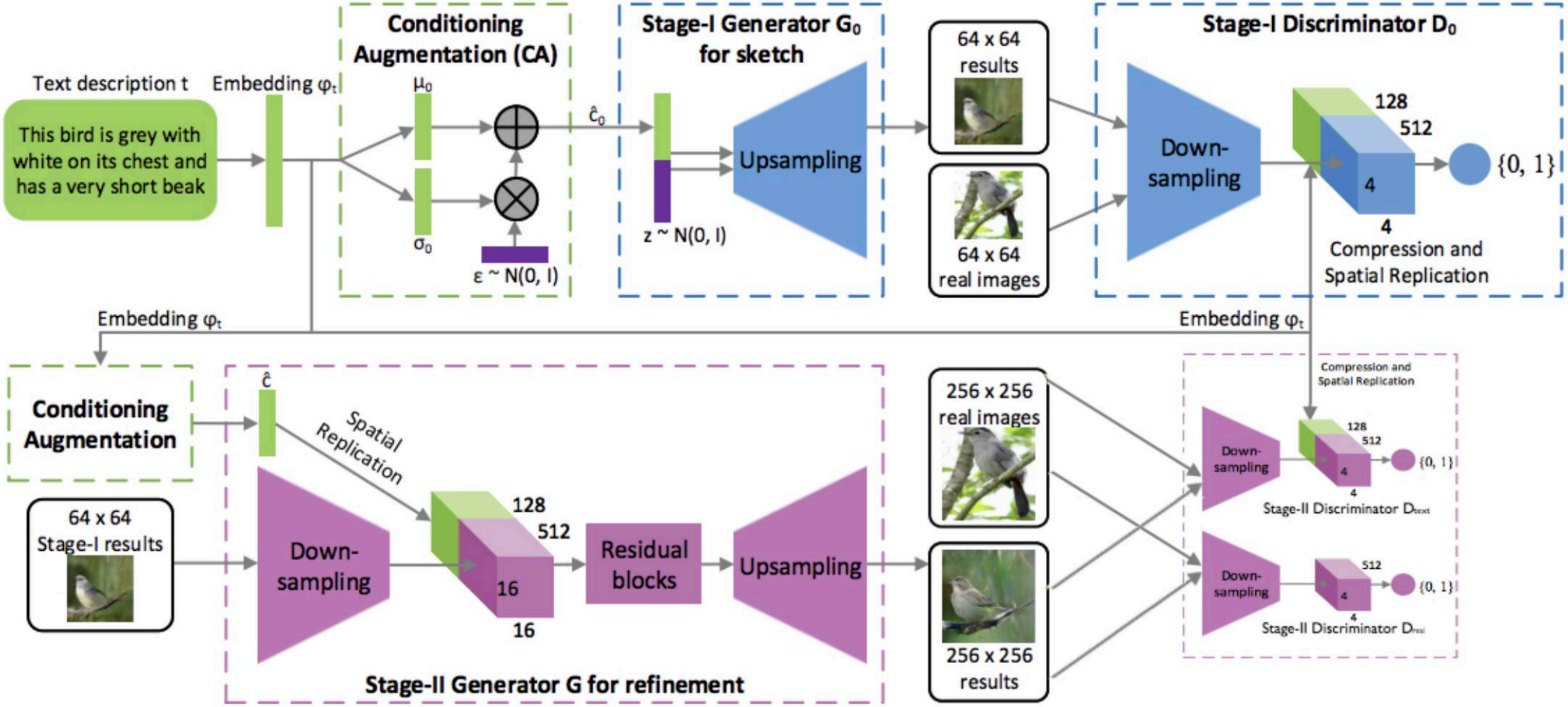
by Ke Ma, Wen Xiao, Sing Zeng



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.