Language Technologies Institute

Carnegie Mellon University

Advanced Multimodal Machine Learning

Lecture 1.1: Introduction

Louis-Philippe Morency
Tadas Baltrusaitis

Your Instructors This Semester (11-777)

Louis-Philippe Morency
morency@cs.cmu.edu
Office: GHC-5411
Phone: 412-268-5508

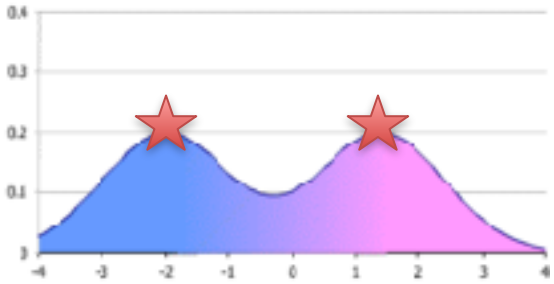
Tadas Baltrusaitis
tbaltrus@cs.cmu.edu
Office: GHC-5409

Lecture Objectives

- Introductions
- What is Multimodal?
  - Multimodal vs multimedia
- Applications and research problems
  - Multimodal HCI, multimedia and AVSR
- Course syllabus and project assignments
  - Grades and course structure

What is Multimodal?

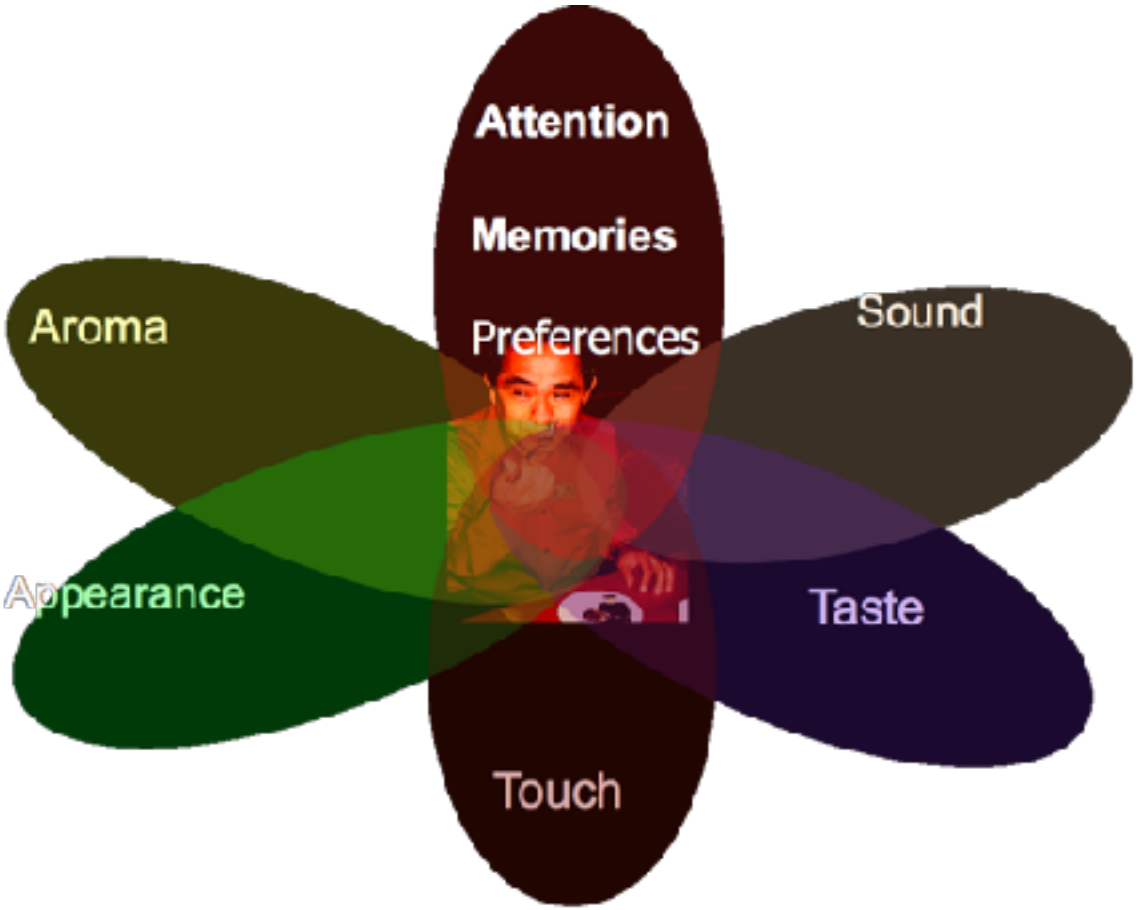
What is Multimodal?



Multimodal distribution

- Multiple modes, i.e., distinct "peaks" (local maxima) in the probability density function

What is Multimodal?



Sensory Modalities

What is Multimodal?

Modality

The way in which something happens or is experienced.

- *Modality* refers to a certain type of information and/or the representation format in which information is stored.
- *Sensory modality:* one of the primary forms of sensation, as vision or touch; channel of communication.

Medium ("middle")

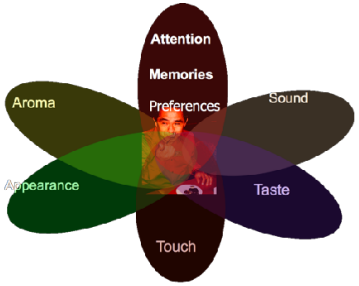
A means or instrumentality for storing or communicating information; system of communication/transmission.

- *Medium* is the means whereby this information is delivered to the senses of the interpreter.

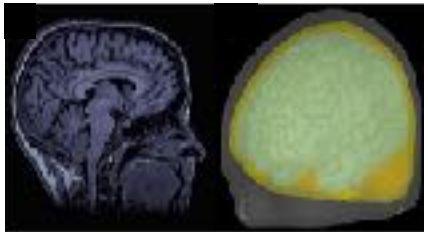
Examples of Modalities

- Natural language (both spoken or written)
- Visual (from images or videos)
- Auditory (including voice, sounds and music)
- Haptics / touch
- Smell, taste and self-motion
- Physiological signals
  - Electrocardiogram (ECG), skin conductance
- Other modalities
  - Infrared images, depth images, fMRI

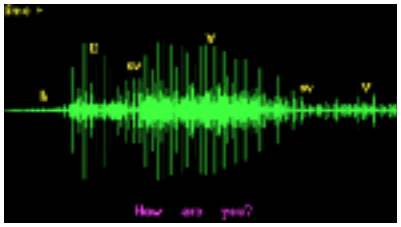
Multiple Communities and Modalities



Psychology



Medical



Speech

Vision

Language

Multimedia

Robotics

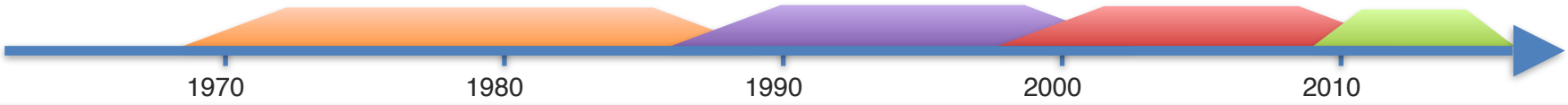
Learning

A Historical View

Prior Research on "Multimodal"

Four eras of multimodal research

- The "behavioral" era (1970s until late 1980s)
- The "computational" era (late 1980s until 2000)
- The "interaction" era (2000 - 2010)
- The "deep learning" era (2010s until …)
  - Main focus of this course



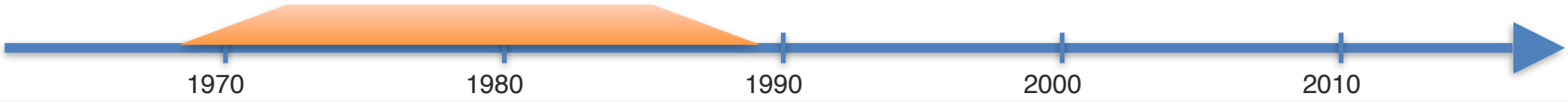
The "Behavioral" Era (1970s until late 1980s)

Multimodal Behavior Therapy by Arnold Lazarus [1973]
➢ 7 dimensions of personality (or *modalities*)

Multi-sensory integration (in psychology):

- Multimodal signal detection: Independent decisions vs. integration [1980]
- Infants' perception of substance and temporal synchrony in multimodal events [1983]
- A multimodal assessment of behavioral and cognitive deficits in abused and neglected preschoolers [1984]

❑ TRIVIA: Geoffrey Hinton received his B.A. in Psychology ☺

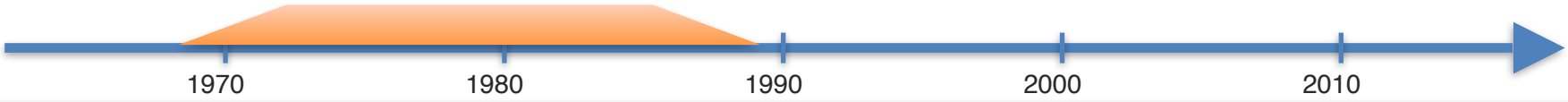


Language and Gestures

David McNeill
University of Chicago
Center for Gesture and Speech Research

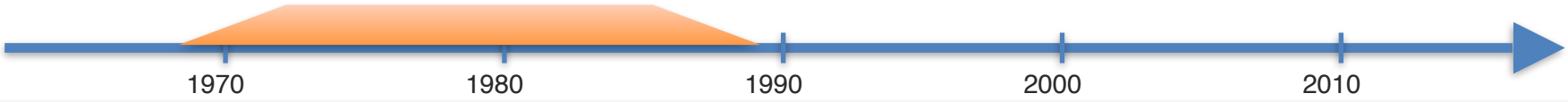
"For McNeill, gestures are in effect the speaker's thought in action, and integral components of speech, not merely accompaniments or additions."

- TRIVIA: Justine Cassell was a student of David McNeill



The McGurk Effect (1976)

Hearing lips and seeing voices – Nature



➢ The "Computational" Era(Late 1980s until 2000)

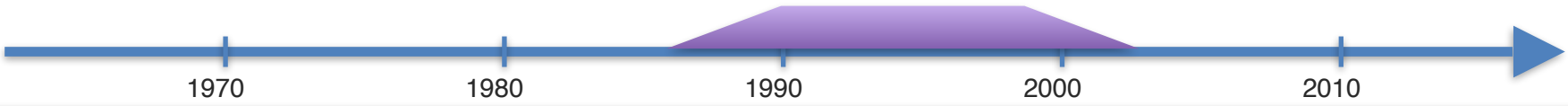
1) Audio-Visual Speech Recognition (AVSR)

- Motivated by the McGurk effect
- First AVSR System in 1986

  "Automatic lipreading to enhance speech recognition"
- Good survey paper [2002]

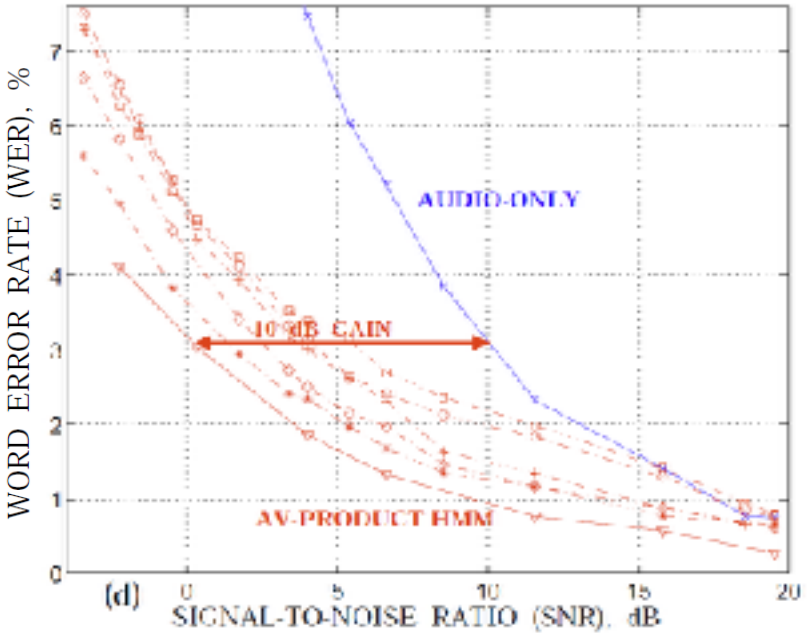
  "Recent Advances in the Automatic Recognition of Audio-Visual Speech"

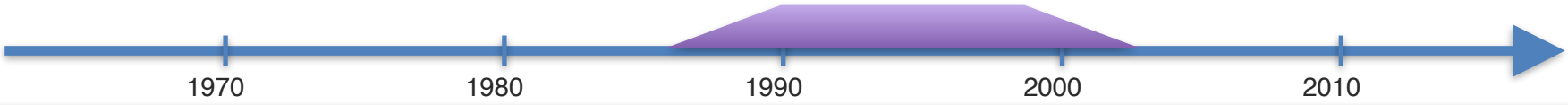
❑ TRIVIA: The first multimodal deep learning paper was about audio-visual speech recognition [ICML 2011]



➢ The "Computational" Era(Late 1980s until 2000)

1) Audio-Visual Speech Recognition (AVSR)



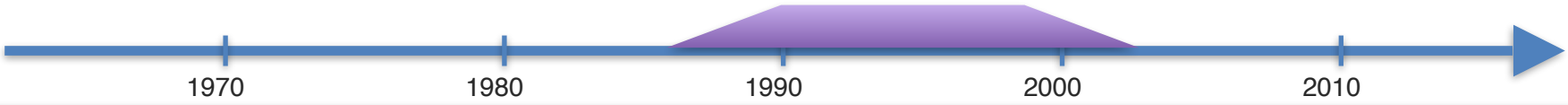


➢ The "Computational" Era (Late 1980s until 2000)

2) Multimodal/multisensory interfaces

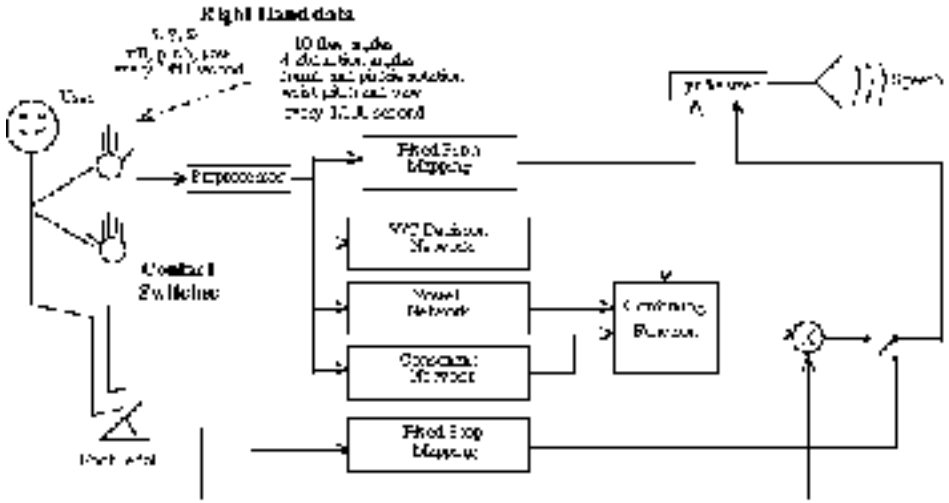
- Multimodal Human-Computer Interaction (HCI)

"Study of how to design and evaluate new computer systems where human interact through multiple modalities, including both input and output modalities."

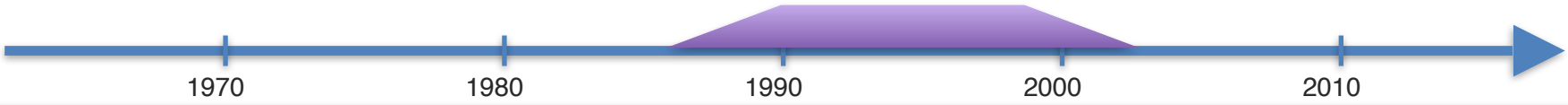


➢ The "Computational" Era (Late 1980s until 2000)

2) Multimodal/multisensory interfaces



[Glove-talk: A neural network interface between a data-glove and a speech synthesizer](#) *By Sidney Fels & Geoffrey Hinton [CHI'95]*



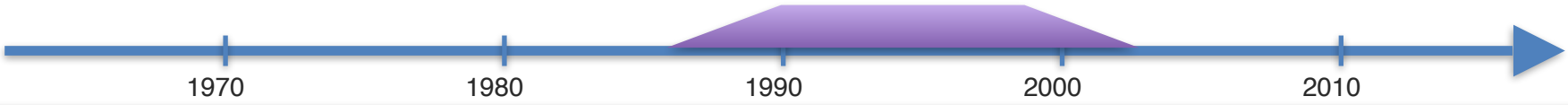
➢ The "Computational" Era (Late 1980s until 2000)

2) Multimodal/multisensory interfaces

Rosalind Picard

Affective Computing *is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena.*

- TRIVIA: Rosalind Picard came from the same group (MIT, Sandy Pentland)

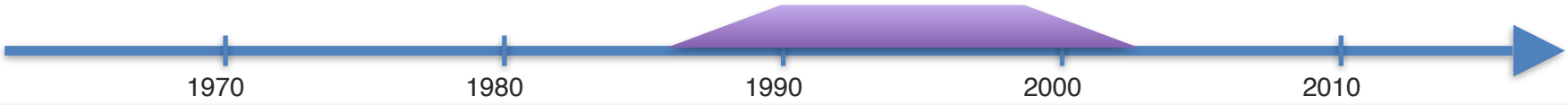


➢ The "Computational" Era (Late 1980s until 2000)

3) Multimedia Computing

Carnegie Mellon University
informedia
digital video understanding
[1994-2010]

"The Informedia Digital Video Library Project automatically combines speech, image and natural language understanding to create a full-content searchable digital video library."

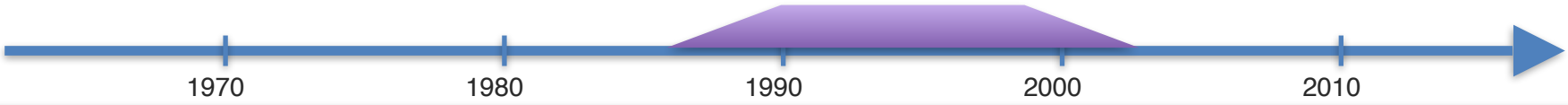


➢ The "Computational" Era (Late 1980s until 2000)

3) Multimedia Computing

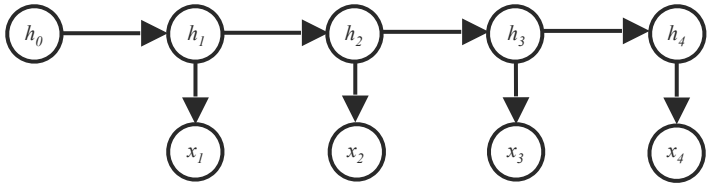
Multimedia content analysis

- **Shot-boundary detection (1991 -)**
  - Parsing a video into continuous camera shots
- **Still and dynamic video abstracts (1992 -)**
  - Making video browsable via representative frames (keyframes)
  - Generating short clips carrying the essence of the video content
- **High-level parsing (1997 -)**
  - Parsing a video into semantically meaningful segments
- **Automatic annotation (indexing) (1999 -)**
  - Detecting prespecified events/scenes/objects in video

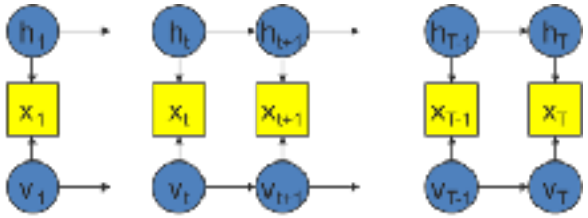


Multimodal Computation Models

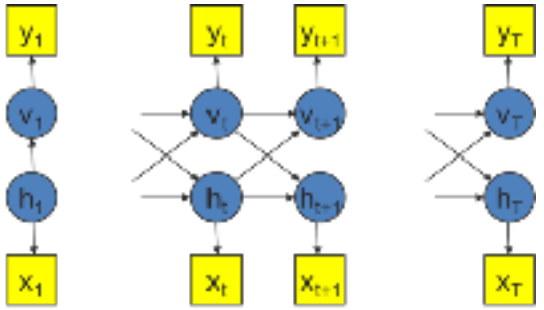
- Hidden Markov Models [1960s]

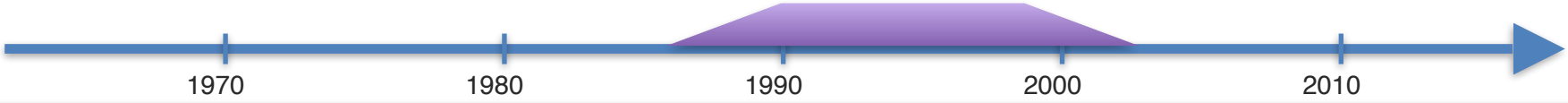


- ❑ Factorial Hidden Markov Models [1996]



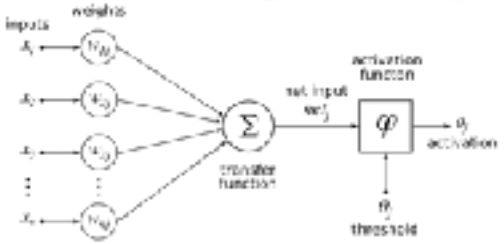
- ❑ Coupled Hidden Markov Models [1997]



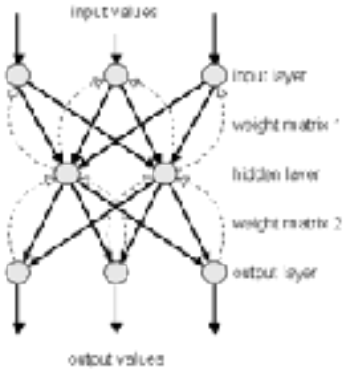


Multimodal Computation Models

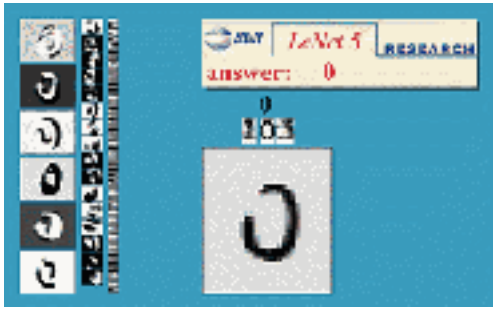
- Artificial Neural Networks [1940s]

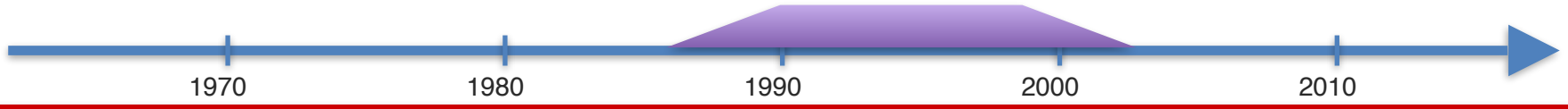


❑ Backpropagation [1975]



❑ Convolutional neural networks [1980s]



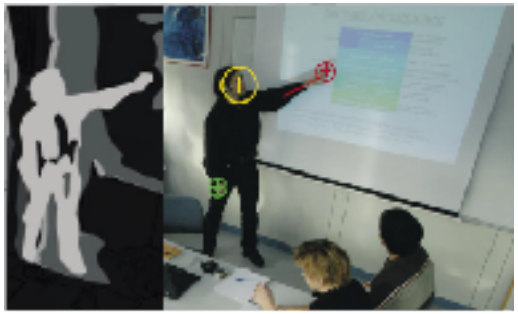


➢ The "Interaction" Era (2000s)

1) Modeling Human Multimodal Interaction

AMI Project [2001-2006, IDIAP]

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



CHIL Project [Alex Waibel]

- Computers in the Human Interaction Loop
- Multi-sensor multimodal processing
- Face-to-face interactions

❑ TRIVIA: Samy Bengio started at IDIAP working on AMI project

➢ The "Interaction" Era (2000s)

1) Modeling Human Multimodal Interaction

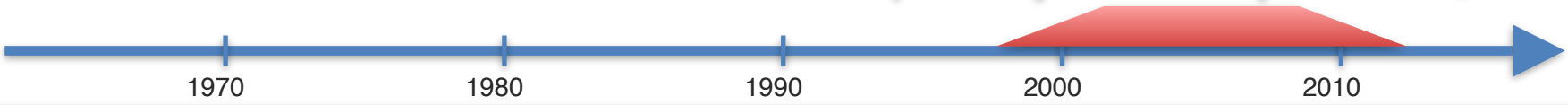
CALO Project [2003-2008, SRI]

- Cognitive Assistant that Learns and Organizes
- Personalized Assistant that Learns (PAL)
- Siri was a spinoff from this project

SSP Project [2008-2011, IDIAP]

- Social Signal Processing
- First coined by Sandy Pentland in 2007
- Great dataset repository: http://sspnet.eu/

❑ TRIVIA: LP's PhD research was partially funded by CALO ☺️



➢ The "Interaction" Era (2000s)

2) Multimedia Information Retrieval

"Yearly competition to promote progress in content-based retrieval from digital video via open, metrics-based evaluation"

[Hosted by NIST, 2001-2016]

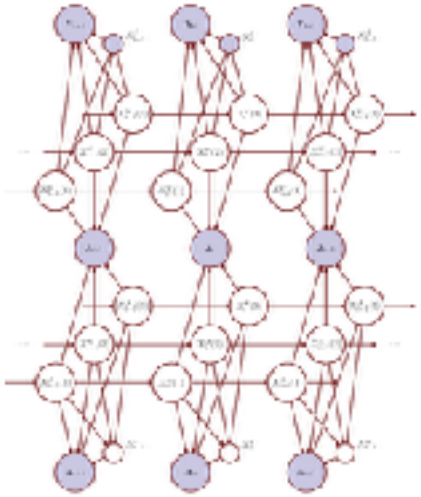
Research tasks and challenges:

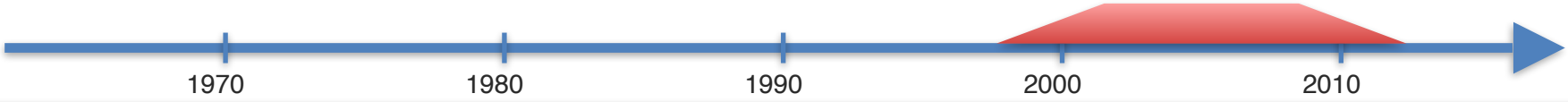
- Shot boundary, story segmentation, search
- "High-level feature extraction": semantic event detection
- Introduced in 2008: copy detection and surveillance events
- Introduced in 2010: Multimedia event detection (MED)

Multimodal Computational Models

- Dynamic Bayesian Networks
  - Kevin Murphy's PhD thesis and Matlab toolbox
  - Asynchronous HMM for multimodal [Samy Bengio, 2007]

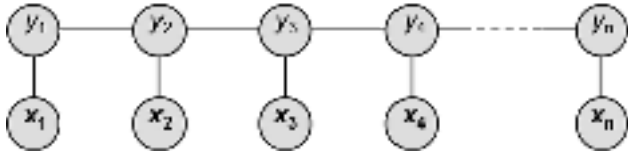
Audio-visual speech segmentation



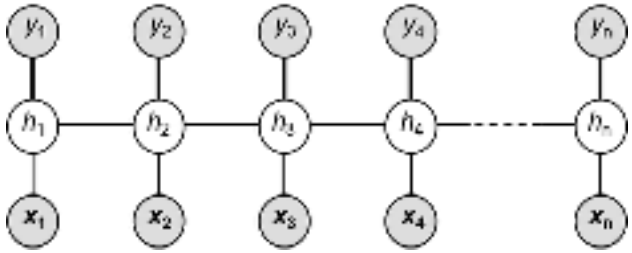


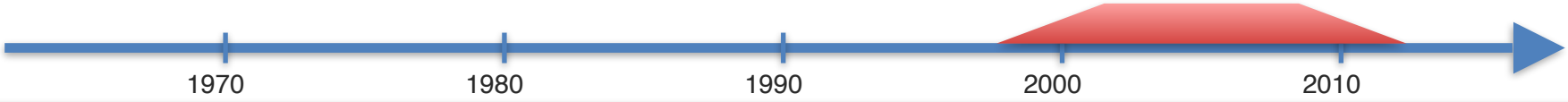
Multimodal Computational Models

- Discriminative sequential models
  - Conditional random fields [Lafferty et al., 2001]



  - Latent-dynamic CRF [Morency et al., 2007]





➢ The "deep learning" era (2010s until …)

Representation learning (a.k.a. deep learning)

- Multimodal deep learning [ICML 2011]
- Multimodal Learning with Deep Boltzmann Machines [NIPS 2012]
- Visual attention: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [ICML 2015]

Key enablers for multimodal research:

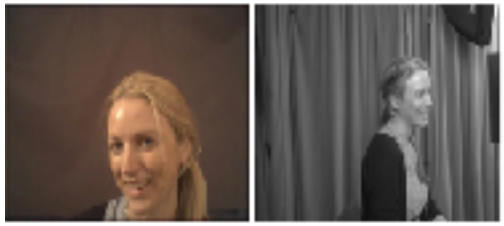
- New large-scale multimodal datasets
- Faster computer and GPUS
- High-level visual features
- "Dimensional" linguistic features

Our course focuses on this era!

➢ The "deep learning" era (2010s until …)

Many new challenges and multimodal corpora !!

Audio-Visual Emotion Challenge (AVEC, 2011-)



- Emotional dimension estimation
- Standardized training and test sets
- Based on the SEMAINE dataset

Emotion Recognition in the Wild Challenge (EmotiW 2013-)

- Emotional dimension estimation
- Standardized training and test sets
- Based on the SEMAINE dataset

➢ The "deep learning" era (2010s until …)

Renew of multimedia content analysis !

- Image captioning

The mar at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

- Video description
- Visual Question-Answer

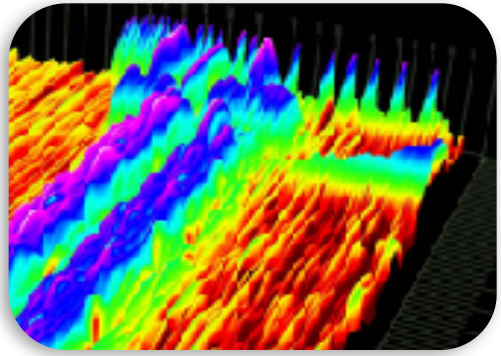
Real world tasks tackled by MMML

- Affect recognition
  - Emotion
  - Persuasion
  - Personality traits
- Media description
  - Image captioning
  - Video captioning
  - Visual Question Answering
- Event recognition
  - Action recognition
  - Segmentation
- Multimedia information retrieval
  - Content based/Cross-media

Multimodal Machine Learning

Verbal

Vocal



Visual

Core Technical Challenges:

Representation **Translation** **Alignment**

Fusion **Co-Learning**

These challenges are non-exclusive.

Course Syllabus

Three Course Learning Paradigms

Research paper reading and group discussion
(40% of your grade)

$$
\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \\
c_t &= f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right) \\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

Course project assignments
(60% of your grade)

Course lectures
(including guest lectures)

Course Structure

Tuesdays

Course lectures

Thursdays

Group discussion and student presentations

Course Recommendations and Requirements

1. Ready to read at least 12 papers this semester !
   - 11 research papers as part of the weekly reading assignments
   - 1 supplementary paper for classroom presentation
2. Already taken a machine learning course
   - Strongly recommended for students to have taken an introduction machine learning course
   - 10-401, 10-601, 10-701, 11-663, 11-441, 11-641 or 11-741
3. Motivated to produce a high-quality course project
   - Three course project assignments
   - Designed to enhance state-of-the-art algorithms

Course Grades

- Discussion participation 20%
- Reading assignments 20%

$$
\begin{aligned}
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right)\\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right)\\
c_t &= f_t c_{t-1} + i_t \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right)\\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right)\\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

- First project assignment
  - Report and presentation 15%
- Mid-term project assignment
  - Report and presentation 15%
- Final project assignment
  - Report and presentation 30%

Course Project

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
c_t &= f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

- Pre-proposal (in 2 weeks)
  - Define your dataset and research task
- First project assignment (in 5 weeks)
  - Experiment with unimodal representations
  - Explore/discuss simple baseline model(s)
- Midterm project assignment (in 11 weeks)
  - Implement and evaluate state-of-the-art model(s)
  - Discuss new multimodal model(s)
- Final project assignment (in 15 weeks)
  - Implement and evaluate new multimodal model(s)
  - Discuss future directions

Course Project Guidelines

- Dataset should have at least two modalities:
  - Natural language and visual/images
- Teams of 2, 3 or 4 students
  - No individual projects
- The project should explore algorithmic novelty
- Possible venues for your final report:
  - ICMI 2017, NIPS 2017, EMNLP 2017, BMVC 2017
- We will discuss on Thursday about project ideas

Lecture Schedule

Classes	Lectures
Week 1 1/17 & 1/19	**Course introduction** • Multimodal applications and dataset • Audio, visual and text features
Week 2 1/24 & 1/26	**Basic mathematical concepts** • Probability and statistical learning • Score, loss and optimization
Week 3 1/31 & 2/2 **Pre-proposal**	**Neural networks and backpropagation** • Multi-layer perceptron • Convolutional neural network
Week 4 2/5 & 2/9	**Multi-view multi-stream modeling** • Backpropagation (Through Time) • Recurrent neural network and LSTM

Lecture Schedule

Classes	Lectures
Week 5 2/14 & 2/16	**Multimodal representation learning** • Multimodal auto-encoders • Multimodal deep neural networks
Week 6 2/21 & 2/23	***First project assignment - Presentations***
Week 7 2/28 & 3/2	**Multimodal component analysis** • Deep canonical correlation analysis • Non-negative matrix factorization
Week 8 3/7 & 3/9	**Multimodal Optimization** • Optimization in deep neural networks • Variational approaches

Lecture Schedule

Classes	Lectures
Spring break 3/13 – 3/17	
Week 9 3/21 & 3/23	**Multimodal alignment** • Attention models and multi-instance learning • Multimodal synchrony and prediction
Week 10 3/28 & 3/30	**Markov Random Fields** • Boltzmann distribution and CRFs • Continuous and fully-connected CRFs
Week 11 4/4 & 4/6	***Mid-term project assignment - Presentations***

Lecture Schedule

Classes	Lectures
Week 12 4/11 & 4/13	**Multimodal fusion** • Sample-based late fusion • Multi-kernel learning and fusion
Week 13 4/18 & 4/20	**Application: Multilingual computational models** • Neural Machine Translation • Sub-word Models
Week 14 4/25 & 4/27	**Application: Language and Vision** • Learned visual representations • Visual activity recognition
Week 15 5/2 & 5/4	***Final project assignment - Presentations***