How to AI (Almost) Anything

Lecture 2 – Data, structure, learning

Paul Liang
Assistant Professor
MIT Media Lab & MIT EECS

https://pliang279.github.io
ppliang@mit.edu
@pliang279

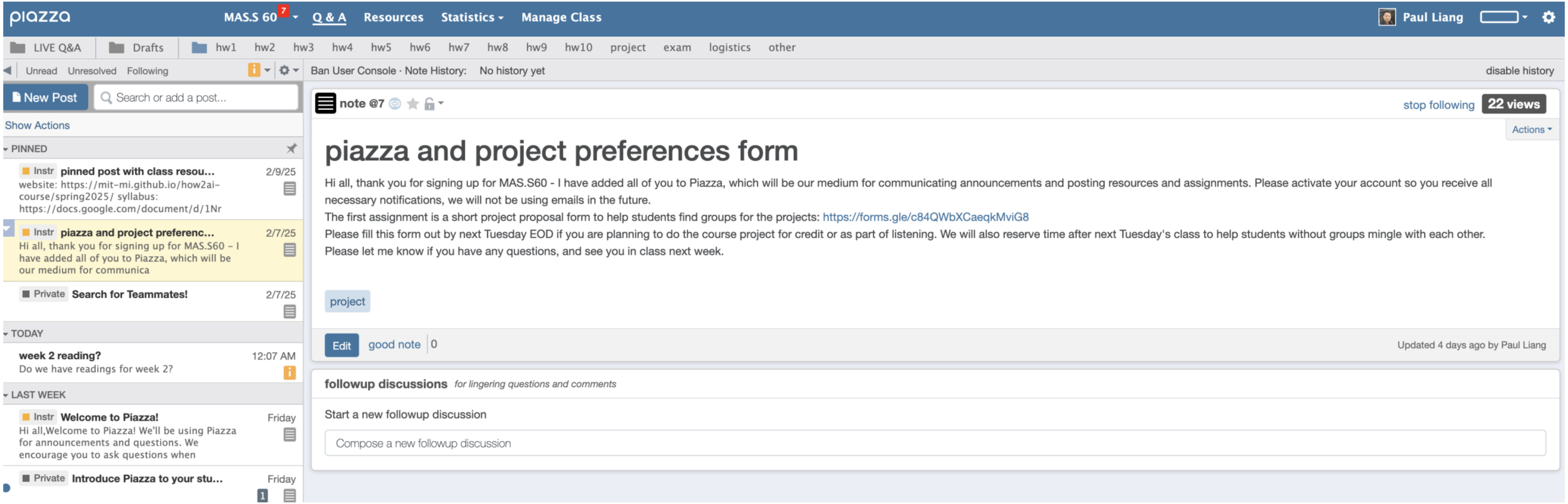
Lecture Outline

1. Vision, language, audio, sensing, set, graph modalities
2. Modality profile
3. Types of data and labels
4. Common learning objectives and generalization

[credit: some slides in this lecture were co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Piazza



multisensory intelligence

Calendar

how2ai Spring25 Calendar

File Edit View Insert Format Data Tools Extensions Help

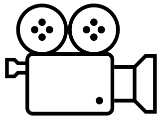
	Monday	Tuesday	Wednesday	Thursday	Friday
Week 1	Feb 3	Feb 4	Feb 5	Feb 6	Feb 7
Introduction to AI and AI research		lecture 1		how to AI research	
Week 2	Feb 10	Feb 11	Feb 12	Feb 13	Feb 14
Foundations 1: Data, structure, information		lecture 2 project preferences		PyTorch and huggingface	
Week 3	Feb 17	Feb 18	Feb 19	Feb 20	Feb 21
Foundations 2: Learning and generalization	President's Day	no class President's Day	reading 1 released	proposal presentation	
Week 4	Feb 24	Feb 25	Feb 26	Feb 27	Feb 28
Foundations 3: Common model architectures	reading 1 due	lecture 3 proposal report due	reading 2 released	discussion 1	
Week 5	Mar 3	Mar 4	Mar 5	Mar 6	Mar 7
Multimodal 1: Connections and alignment	reading 2 due	lecture 4	reading 3 released	discussion 2	Add Date
Week 6	Mar 11	Mar 11	Mar 12	Mar 13	Mar 14
Multimodal 2: Interactions and fusion	reading 3 due	lecture 5	reading 4 released	discussion 3	
Week 7	Mar 17	Mar 18	Mar 19	Mar 20	Mar 21
Multimodal 3: Cross-modal transfer	reading 4 due	lecture 6		discussion 4	
Week 8 - spring break					

multisensory intelligence

What is a Sensory Modality?

Sensory modality

Modality refers to the way in which something expressed or perceived.



from a sensor

Raw Modalities
(closest from sensor)

Abstract Modalities
(farthest from sensor)

Examples:

Speech signal

Image

Language

Detected objects

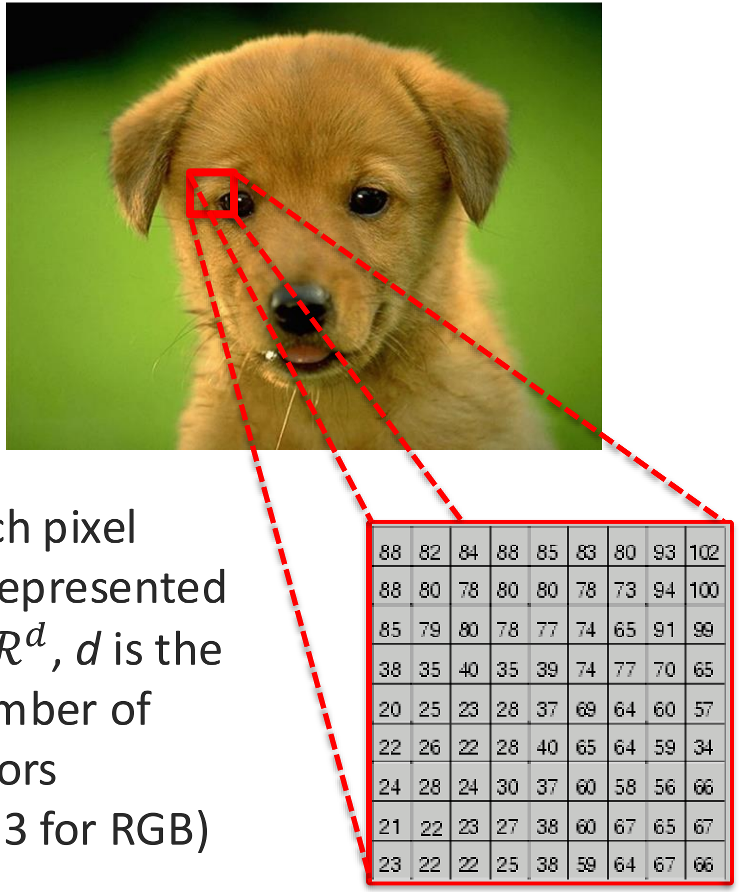
Sentiment intensity

Object categories

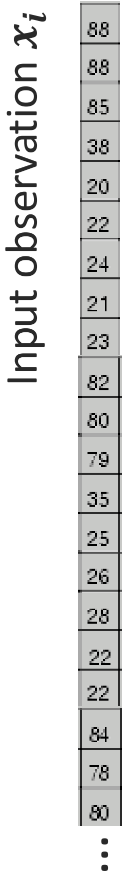
Most of AI is about learning abstractions, or representations, from data.

Visual Modality

Color image



Each pixel is represented in $\mathcal{R}^d$, *d* is the number of colors (d=3 for RGB)



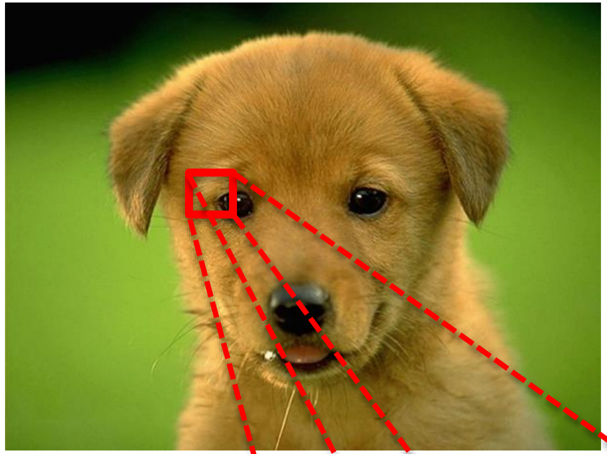
Binary classification problem

Dog ?

label $y_i \in \mathcal{Y} = \{0,1\}$

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Visual Modality



88	82	84	88	85	83	80	93	102
88	80	78	80	80	78	73	94	100
85	79	80	78	77	74	65	91	99
38	35	40	35	39	74	77	70	65
20	25	23	28	37	69	64	60	57
22	26	22	28	40	65	64	59	34
24	28	24	30	37	60	58	56	66
21	22	23	27	38	60	67	65	67
23	22	22	25	38	59	64	67	66

Each pixel is represented in $\mathcal{R}^d$, *d* is the number of colors (*d*=3 for RGB)

Input observation $\boldsymbol{x}_i$

88
88
85
38
20
22
24
21
23
82
80
79
35
25
26
28
22
22
84
78
80
⋮

Multi-class classification problem

Duck
-or-
Cat
-or-
Dog
-or-
Pig
-or-
Bird ?

label $y_i \in \mathcal{Y} = \{0,1,2,3,\ldots\}$

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Language Modality

Written language

★★★★★ **Masterful!**

By Antony Witheyman - January 12, 2006

Ideal for anyone with an interest in disguises who likes to see the subject tackled in a humourous manner.

0 of 4 people found this review helpful

Spoken language

```
                MARTHA(CON'T)
Look around you. Look at all the
great things you've done and the
people you've helped.

                CLARK
But you've only put up the good
things they say about me.

                MARTHA
Clark, honey. If I were to use the
bad things they say I could cover
the barn, the house and the
outhouse.
```

Input observation x_i

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

"one-hot" vector

$|x_i|$ = number of words in dictionary

Word-level classification

Part-of-speech ?
(noun, verb,…)

Sentiment ?
(positive or negative)

Named entity ?
(names of person,…)

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Language Modality

Written language

Masterful!

By Antony Witheyman - January 12, 2006

Ideal for anyone with an interest in disguises who likes to see the subject tackled in a humourous manner.

0 of 4 people found this review helpful

Spoken language

```
                MARTHA(CON'T)
Look around you. Look at all the
great things you've done and the
people you've helped.

                CLARK
But you've only put up the good
things they say about me.

                MARTHA
Clark, honey. If I were to use the
bad things they say I could cover
the barn, the house and the
outhouse.
```

Input observation x_i

$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

"bag-of-word" vector

$|x_i|$ = number of words in dictionary

Document-level classification

Sentiment ?

(positive or negative)

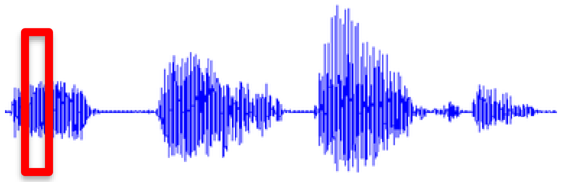
Response?

What happens with word ordering?

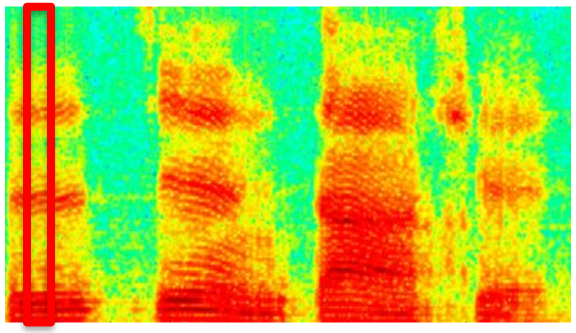
multisensory intelligence

Acoustic Modality

Digitalized acoustic signal



- Sampling rates: 8~96kHz
- Bit depth: 8, 16 or 24 bits
- Time window size: 20ms
  - Offset: 10ms



Spectrogram

Input observation x_i

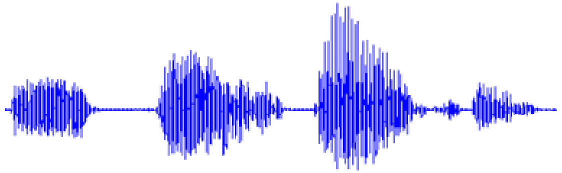
0.21
0.14
0.56
0.45
0.9
0.98
0.75
0.34
0.24
0.11
0.02

Spoken word ?

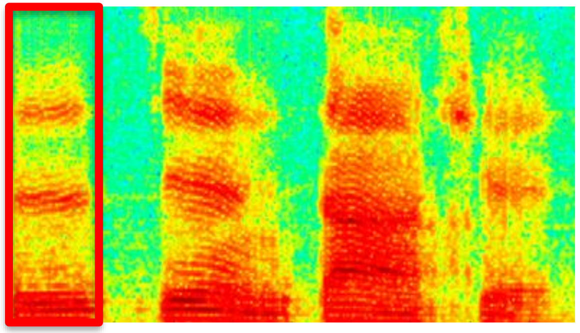
[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Acoustic Modality

Digitalized acoustic signal



- Sampling rates: 8~96kHz
- Bit depth: 8, 16 or 24 bits
- Time window size: 20ms
  - Offset: 10ms



Spectrogram

Input observation x_i

0.21
0.14
0.56
0.45
0.9
0.98
0.75
0.34
0.24
0.11
0.02
0.24
0.26
0.58
0.9
0.99
0.79
0.45
0.34
0.24

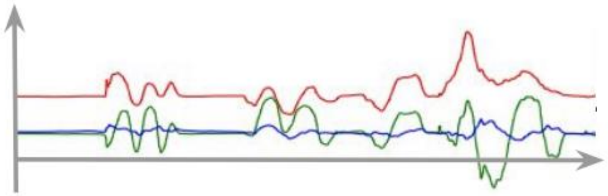
⋮

Emotion ?

Spoken word ?

Voice quality ?

Sensor Modality



Force-Torque Sensor

Time series data across six-axis Force-Torque sensor:
T × 6 signal.

Proprioception

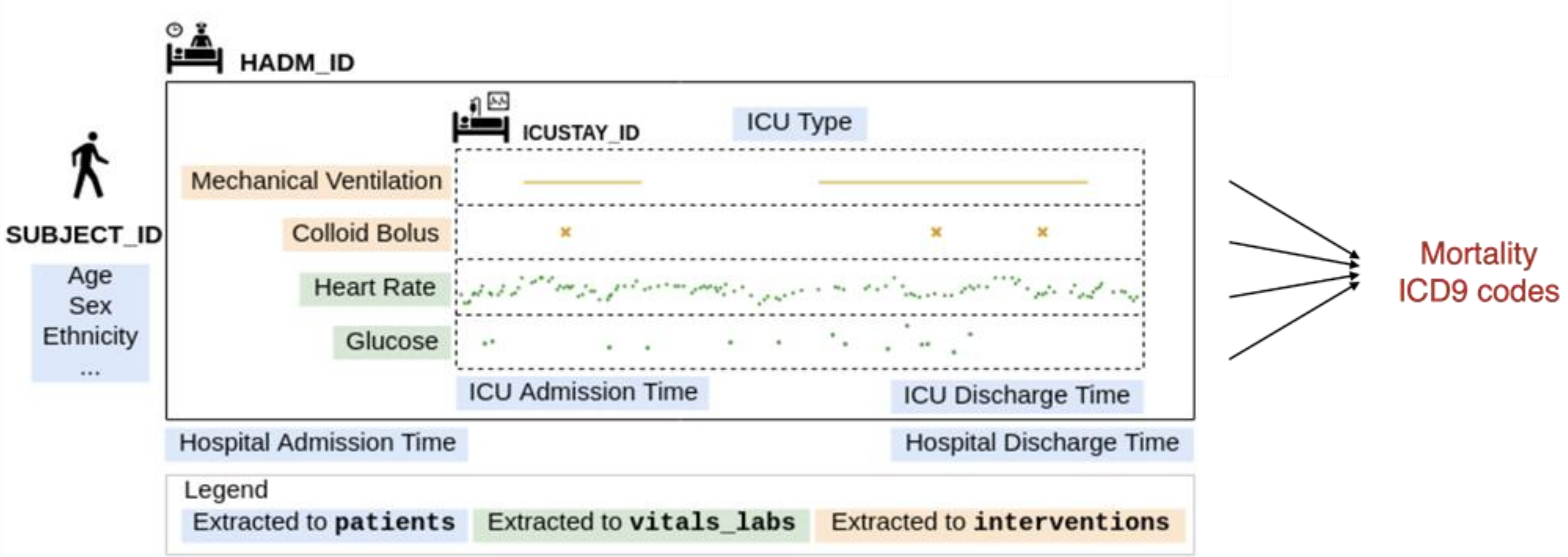
Measure values internal to the system (robot); e.g. motor speed, wheel load, **robot arm joint angles**, battery voltage.

Time series data across current position and velocity of the end-effector:
T × 2 signal.

Object property
Next action

[Lee et al., Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks. ICRA 2019]

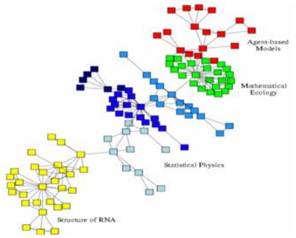
Tabular Modality

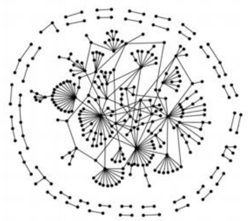


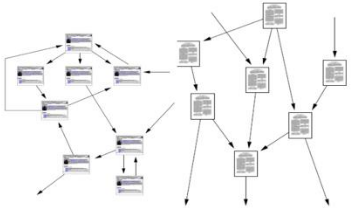
[Johnson et al., MIMIC-IV: A Freely Accessible Electronic Health Record Dataset. Scientific Data 2023]

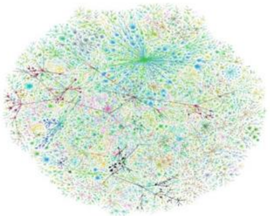
Graph Modality

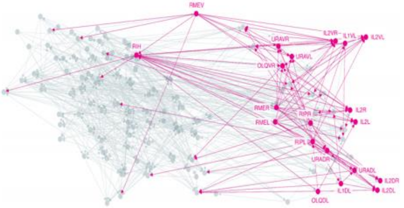

Social networks


Economic networks


Biomedical networks


Information networks:
Web & citations


Internet


Networks of neurons

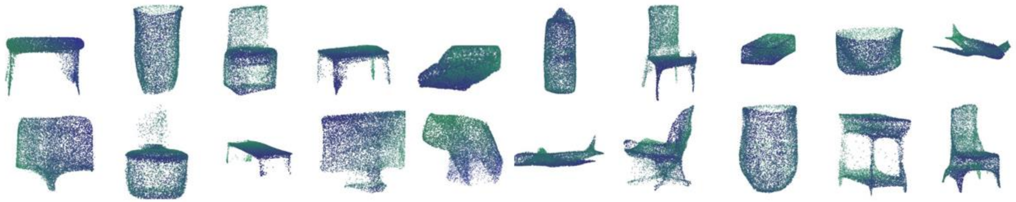
Tasks on graphs:
Node classification
Link prediction
…

Using graphs:
Knowledge graphs for QA
Social network for sentiment analysis
…

[Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Set Modality

Sets



Point clouds

Set anomaly detection
Set expansion
Set completion
Point cloud classification
Point cloud generation

[Zaheer et al., DeepSets. NeurIPS 2017 and Li et al., Point Cloud GAN. arxiv 2018]

multisensory intelligence

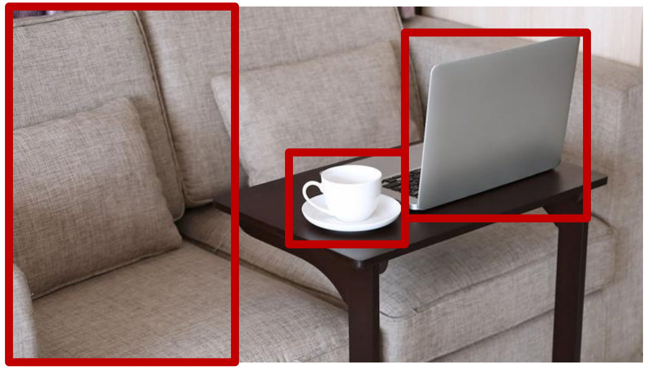
Modality Profile

The qualities and structures that are unique to a data modality.

A teacup on the right of a laptop in a clean room.

Modality Profile

The distribution of individual elements within that modality.



*A **teacup** on the **right** of a **laptop** in a **clean room**.*

1 **Distribution:** discrete or continuous, support

{  ,  ,  }

{teacup, right, laptop, clean, room}

multisensory intelligence

Modality Profile

The frequency at which elements appear or are sampled.

A teacup on the right of a laptop in a clean room.

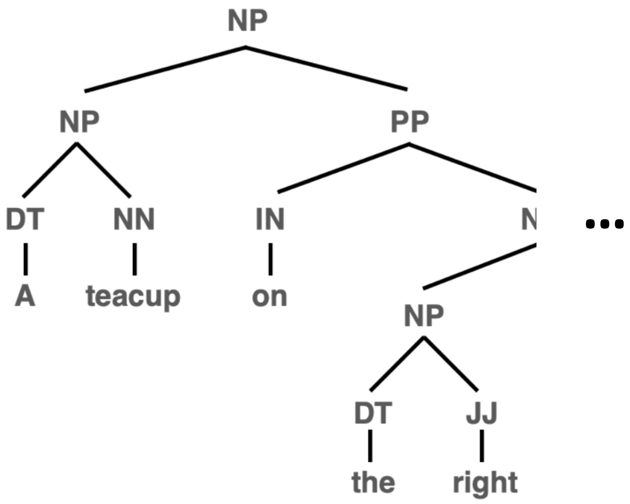
2 **Granularity:** sampling rate and frequency

objects per image

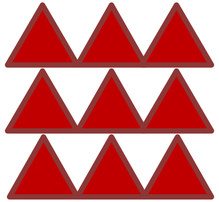
words per minute

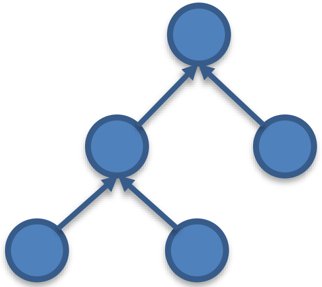
Modality Profile

The way elements compose with each other to form entire data.



3 **Structure:** static, temporal, spatial, hierarchical



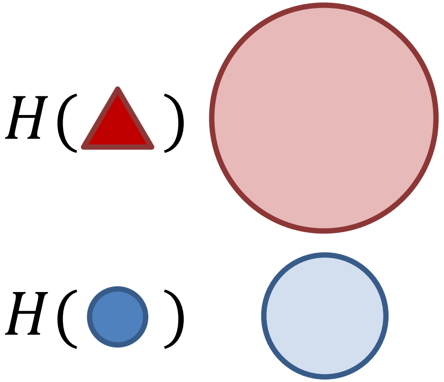


Modality Profile

The total information contained in the elements and their composition.

A teacup on the right of a laptop
in a clean room.

4 **Information:** entropy and density



Modality Profile

The natural imperfections in the data modality.

A teacup on the right of a laptop in a clean room.

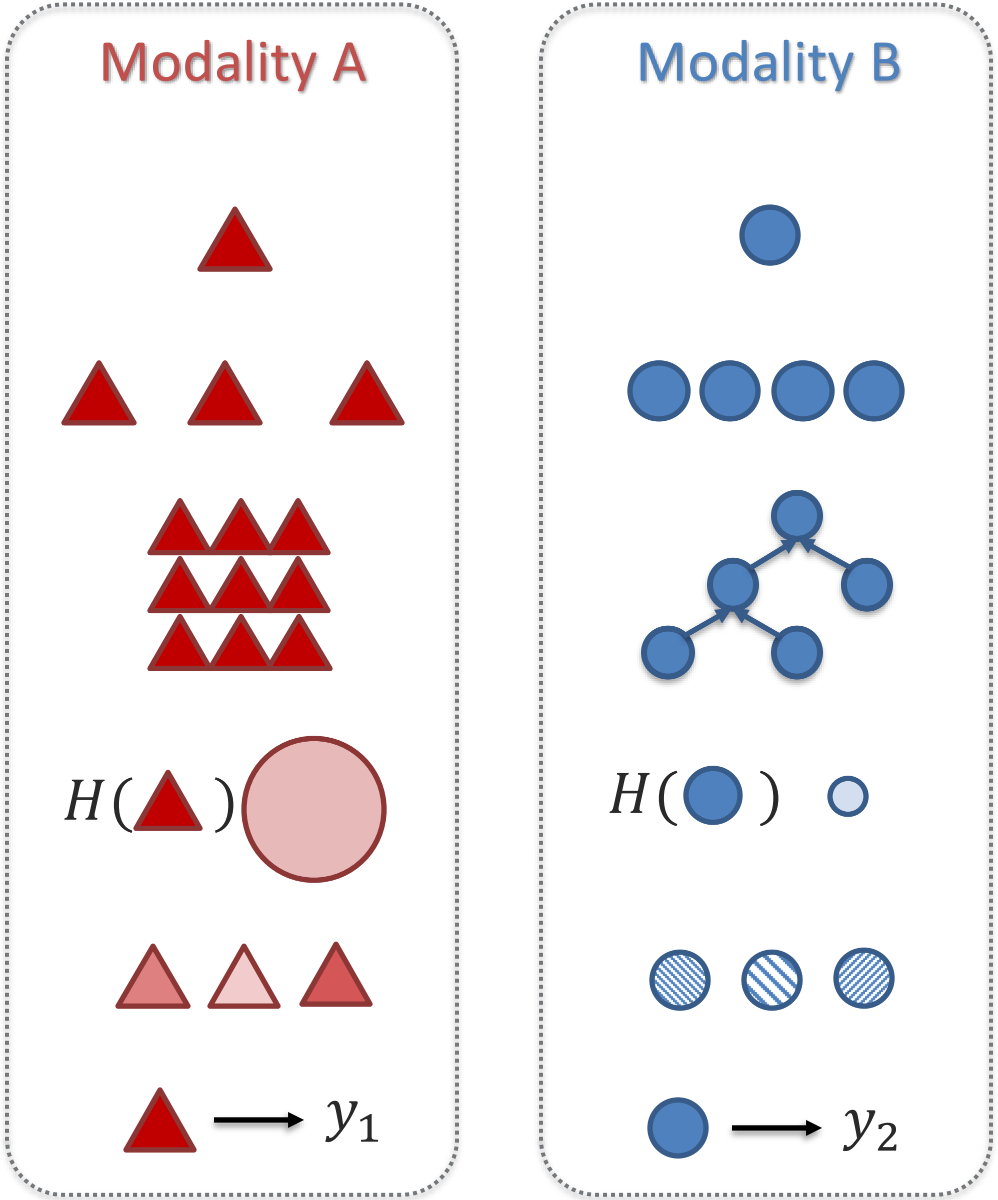
5 **Noise:** uncertainty, signal-to-noise ratio, missing data

teacup → teacip

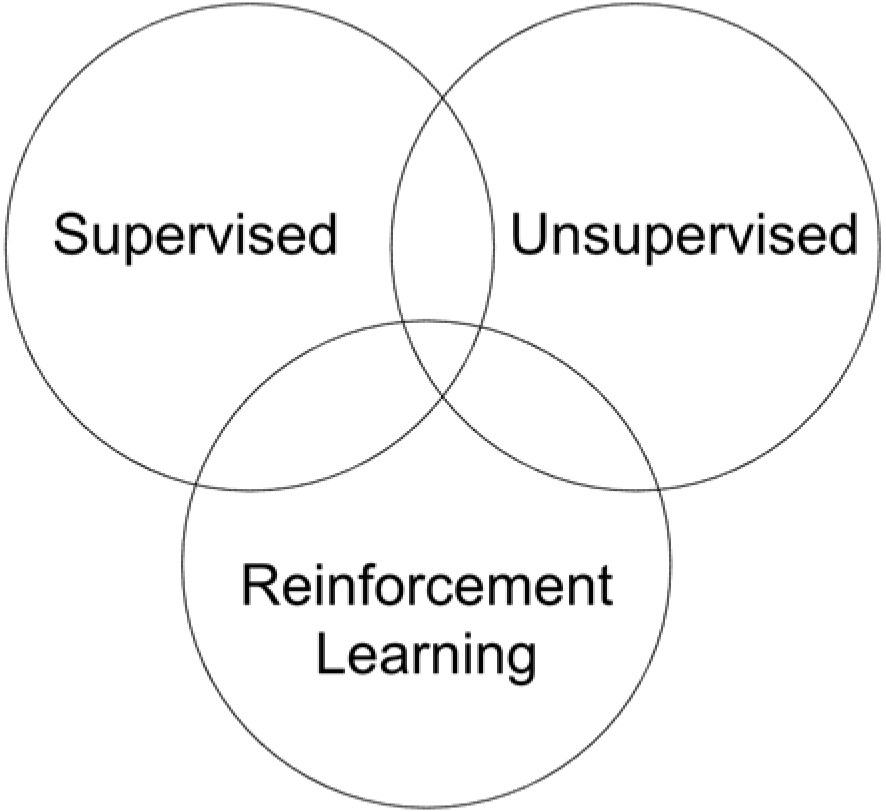
right → rihjt

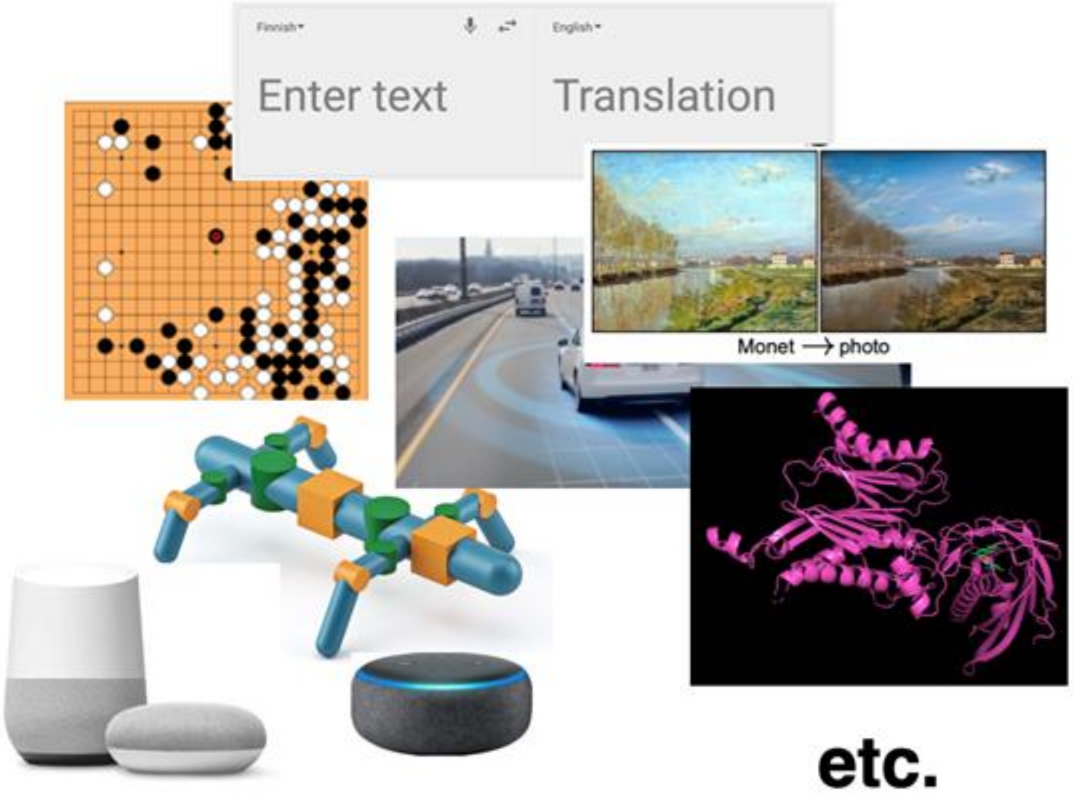
Modality Profile

1. **Element representations:**
   Discrete, continuous, granularity
2. **Element distributions:**
   Density, frequency
3. **Structure:**
   Temporal, spatial, latent, explicit
4. **Information:**
   Abstraction, entropy
5. **Noise:**
   Uncertainty, noise, missing data
6. **Relevance:**
   Task, context dependence



Types of Learning Paradigms





(the categorization can be refined, e.g. there are active learning, semi-supervised, selective, contrastive, few-shot, inverse reinforcement learning…)

[Slides adapted from 6.790]

Supervised Learning

Goal: predict to what degree a drug candidate binds to the intended target protein (based on a dataset of already-screened molecules against the target)

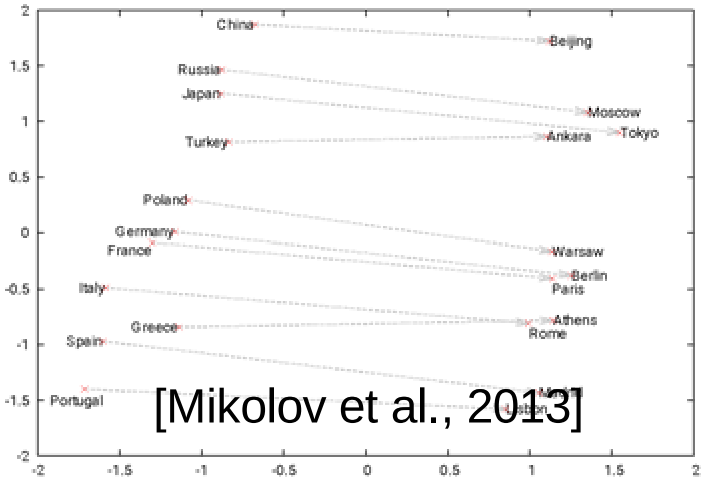
- Learning a machine translation system from pairs of sentences

Spanish (input)	English (output)
Aquí tienes un bolígrafo	Here's a pen
Las conferencias de ML son divertidas	ML conferences are fun
Todo el mundo debería estudiar AI	Everyone should study AI

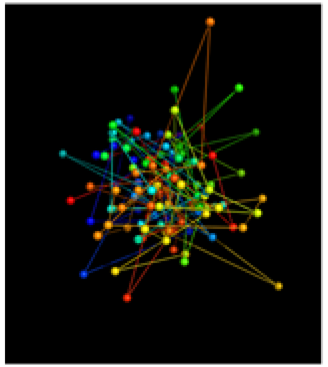
[Slides adapted from 6.790]

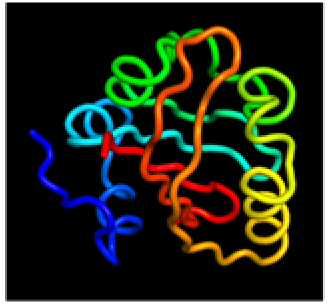
Unsupervised Learning

dimensionality reduction, embedding



[Mikolov et al., 2013]



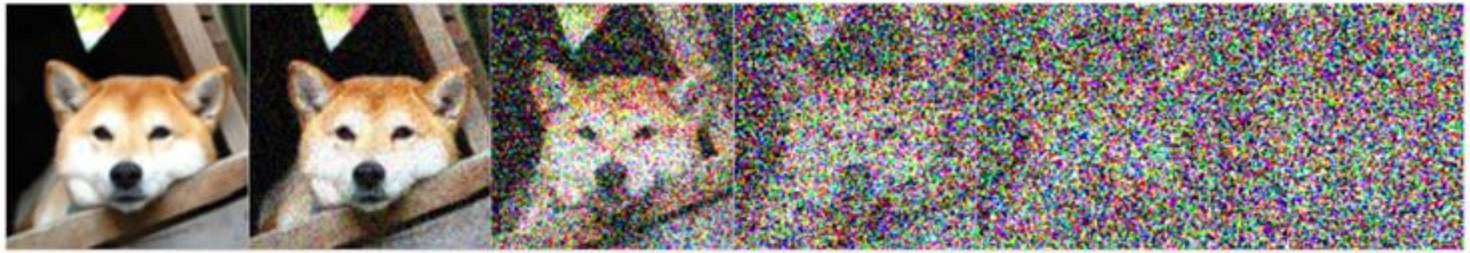


[courtesy of Jason Yim]

Over 3D protein structures, etc.

+Self-Supervised paradigm

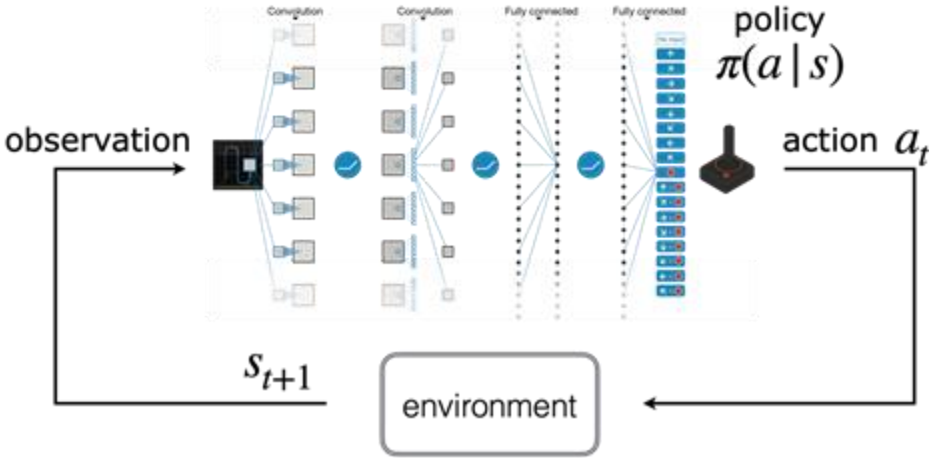
de-noising diffusion models over images

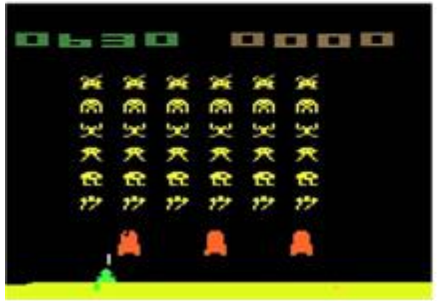


[image from Rissanen et al 2022]

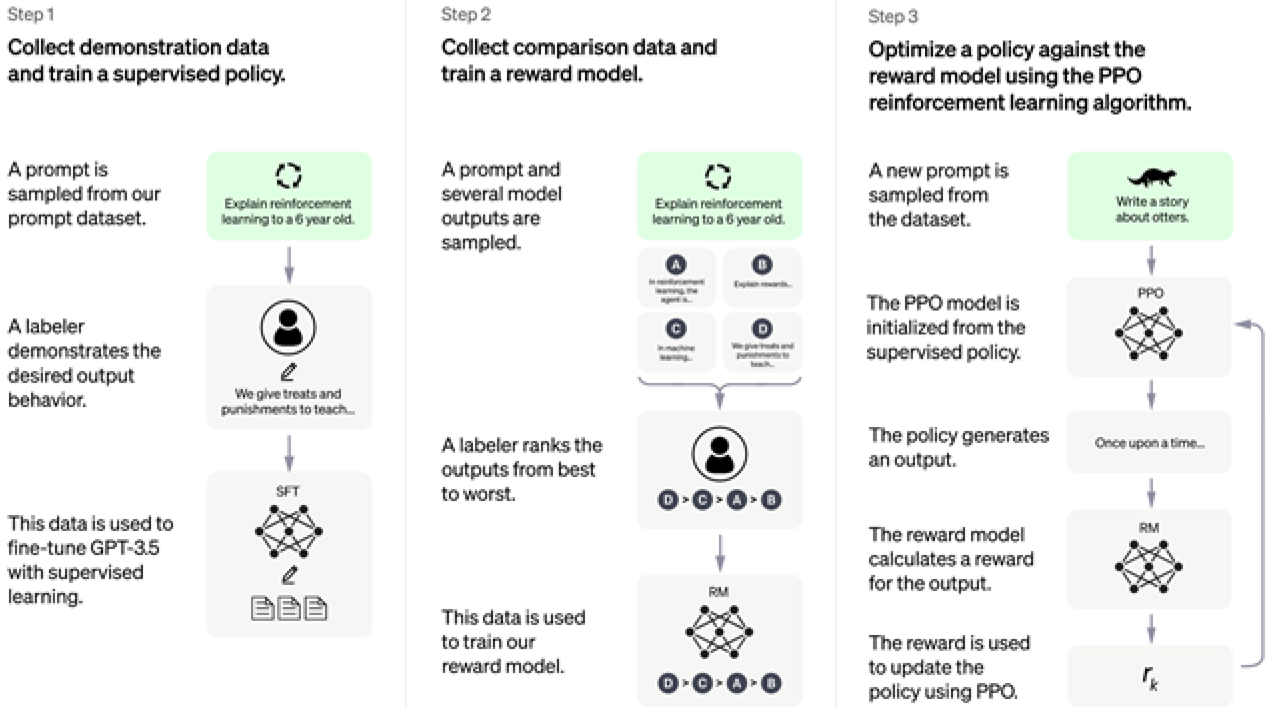
[Slides adapted from 6.790]

Reinforcement Learning





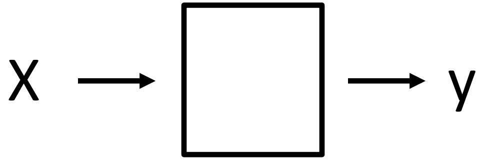
ChatGPT



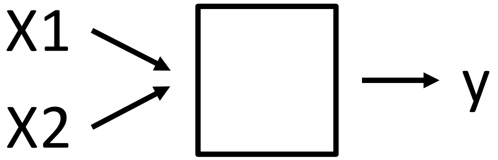
[Slides adapted from 6.790]

More Learning Paradigms

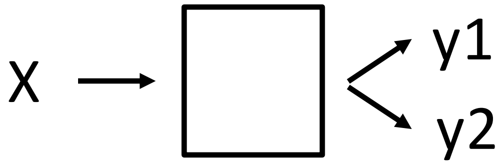
Supervised learning



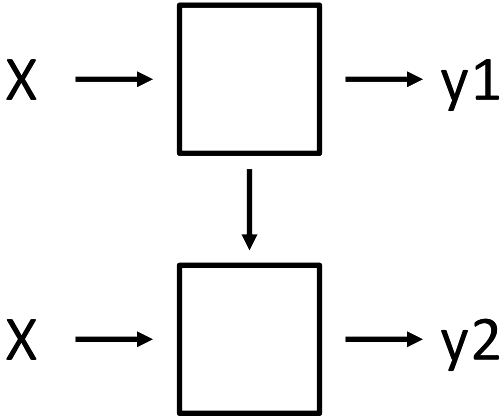
Multimodal (supervised) learning



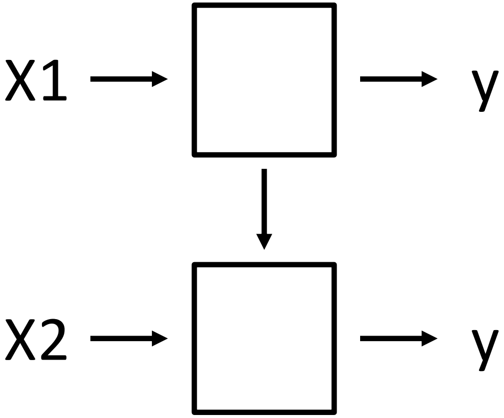
Multitask learning



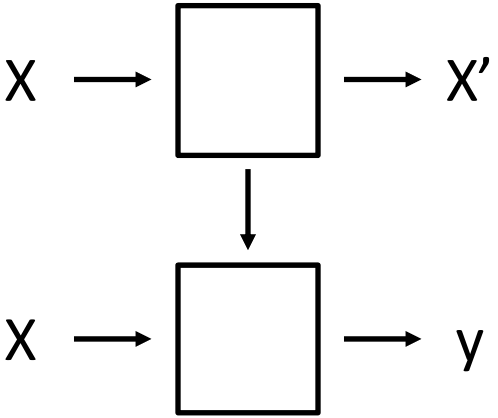
Transfer learning



Cross-modal learning

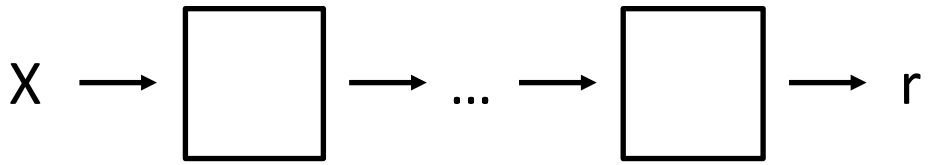


Unsupervised/self-supervised pre-training

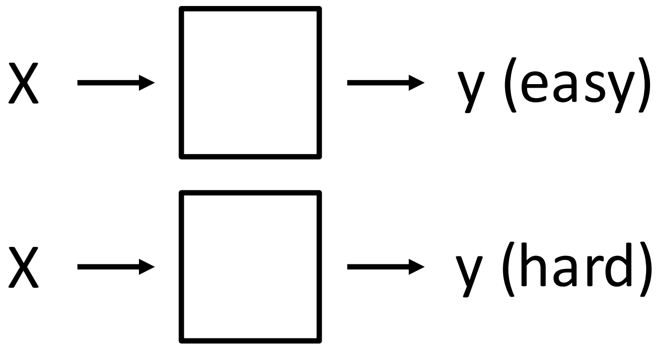


More Interactive Learning Paradigms

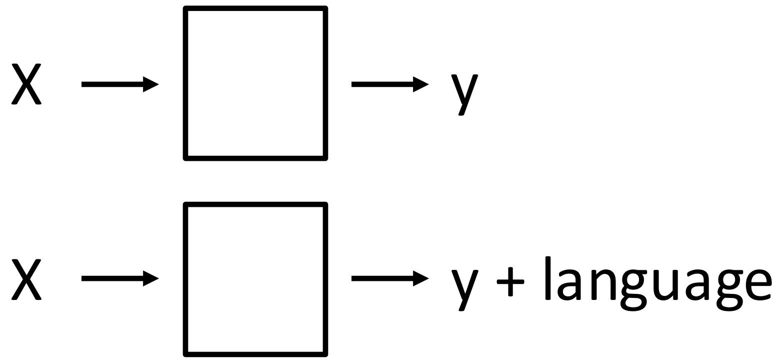
Reinforcement learning



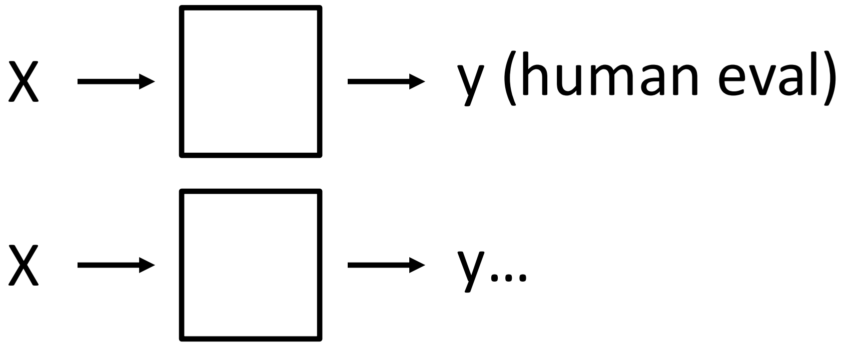
Curriculum/active learning



LLM adaptation



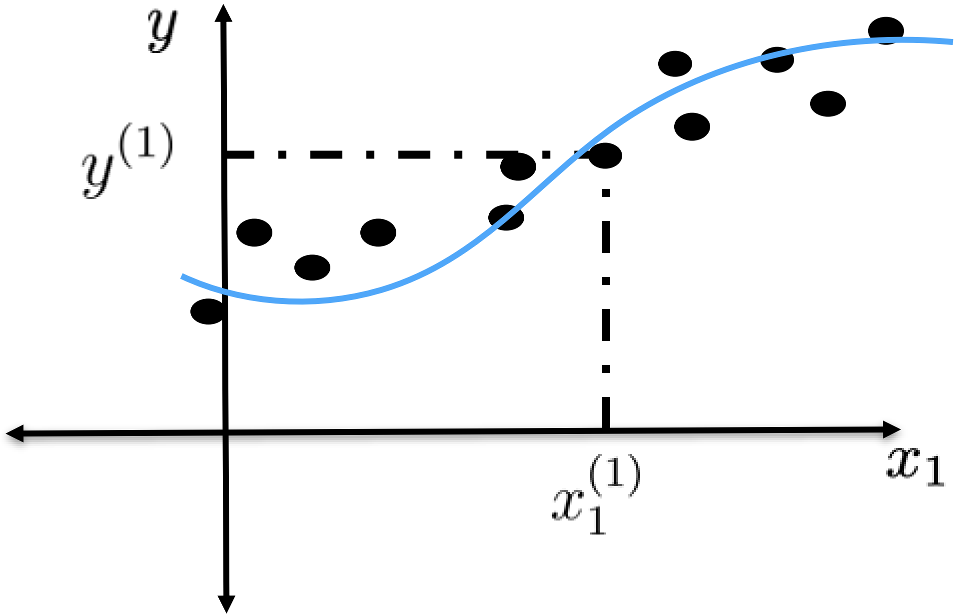
Human-in-the-loop learning



Learning Process

We want a "good" way to label our data

- How to label? Learn a model
- We typically consider a class of possible models



Input: Data

$x \longrightarrow \boxed{h} \longrightarrow y$

Output: Label

how well our model labels new data depends largely on how good the chosen model class is

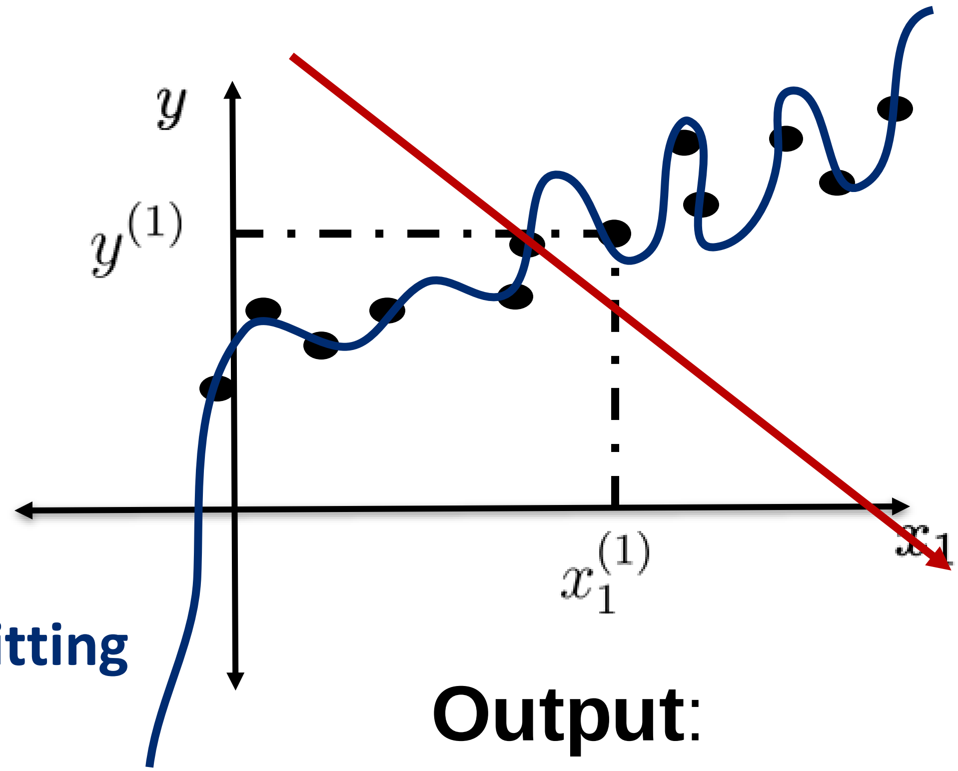
[Slides adapted from 6.390]

Overfitting vs Generalization

What we really want is to generalize to **future data**!

What we don't want:

- Model does not capture the input-output relationship → **Underfitting**
- Model too specialized to training data → **Overfitting**

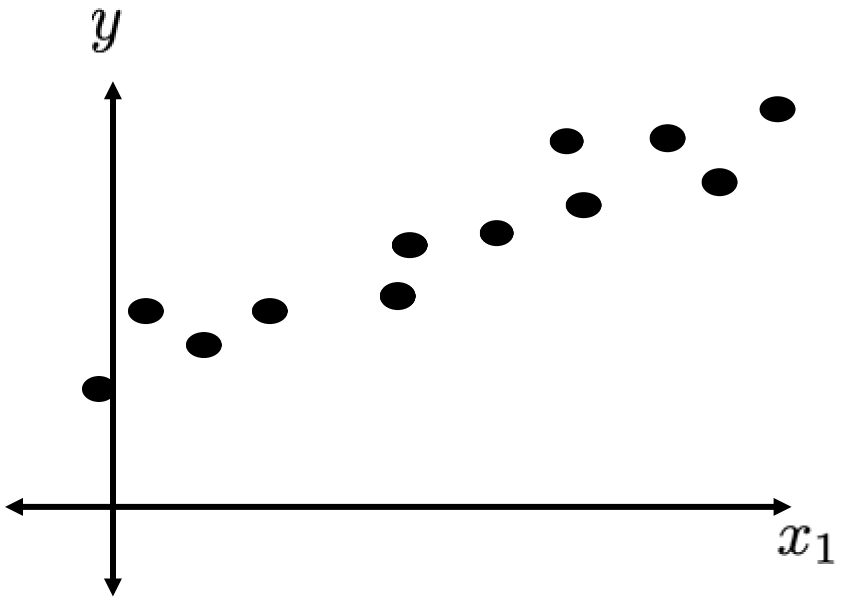


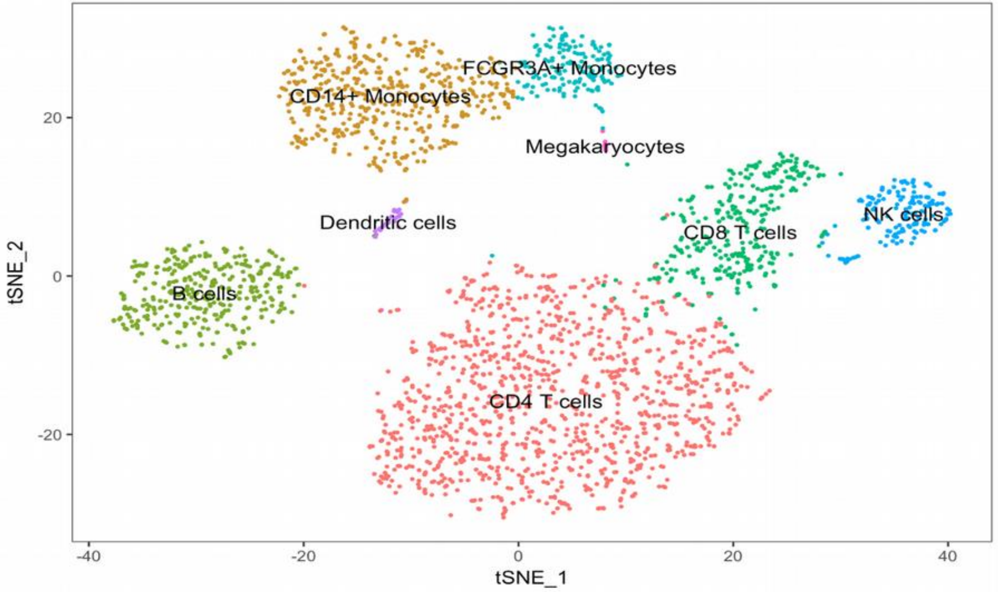
Output:
Label

Split collected data into training, validation, and testing. Critical to make sure test data conditions match real-time deployment conditions.

Summary: How To Data

1. Decide how much data to collect, and how much to label (costs and time)
2. Clean data: normalize/standardize, find noisy data, anomaly/outlier detection
3. Visualize data: plot, dimensionality reduction (PCA, t-sne), cluster analysis
4. Decide on evaluation metric (proxy + real, quantitative and qualitative)
5. Choose model class and learning algorithm (more next lecture)





Lecture Summary

1. Vision, language, audio, sensing, set, graph modalities
2. Modality profile
3. Types of data and labels
4. Common learning objectives and generalization

Assignments for This Coming Week

No reading assignment this week.

For project:

- Project preference form (Due tonight 2/11 at 9pm ET)
- If not team yet, mingle and find teams now!
- Project proposal presentations next Thursday (2/20) in class
  - *Instructions will be sent out via piazza, roughly 5 mins/5 slides per team motivating problem (broad impact + intellectual merit), existing work, datasets used, rough research ideas.*
- Today and Thursday 2-3pm – meet with me at E15-392 so I can give feedback on ideas.

This Thursday: (optional) tutorial on **ML tools – Pytorch, Huggingface, GPUs, Wandb**

Before Thursday, register for huggingface and Wandb account

multisensory intelligence