

# Lecture 1:

# Introduction to Machine Learning

## Fall 2022

Kai-Wei Chang  
CS @ UCLA

[kw+cm146@kwchang.net](mailto:kw+cm146@kwchang.net)

The instructor gratefully acknowledges Eric Eaton (UPenn), who assembled the original slides, Jessica Wu (Harvey Mudd), David Kauchak (Pomona), Dan Roth (Upenn), Sriram Sankararaman (UCLA), whose slides are also heavily used, and the many others who made their course materials freely available online.

# Registration

- ❖ Please fill the following form for PTE:  
<https://bit.ly/cm146f22-signup>
- ❖ Limited by the resources
- ❖ Unlikely we can give many PTEs
- ❖ Please drop the course if you're not planning to take it
- ❖ PTE will be given on the second week
- ❖ Check <http://my.ucla.edu>

# CS M146 Teaching Team

❖ Kai-Wei Chang

❖ Wed, 12:00 PM – 1:00 PM

❖ TAs

TA	Email	Office Hours
Fan Yin	fanyin20@cs.ucla.edu	Thu 3- <u>5pm</u>
Zhouxing Shi	zshi@cs.ucla.edu	Wed 8:30pm-9:30pm Thu 2- <u>3pm</u>
Tanmay Parekh	tparekh@g.ucla.edu	Mon 1- <u>3pm</u>
Yihe Deng	yihedeng@ucla.edu	Wed 11am-1pm
Sidi Lu	sidilu@cs.ucla.edu	Wed 10:30am - 11:30am, Friday 10:30am - 11:30am
Masoud Monajatipoor	monajati@ucla.edu	

# Homework 0

- ❖ Sign up at Piazza

  - <http://piazza.com/ucla/fall2022/m146>

- ❖ Complete the math quiz at Bruinlearn

  - ❖ Will be released on Friday

# What is machine learning?

Machine Learning is the study of algorithms that

- improve their performance  $P$
- at some task  $T$
- with experience  $E$ .

A well-defined learning task is given by  $\langle P, T, E \rangle$ .

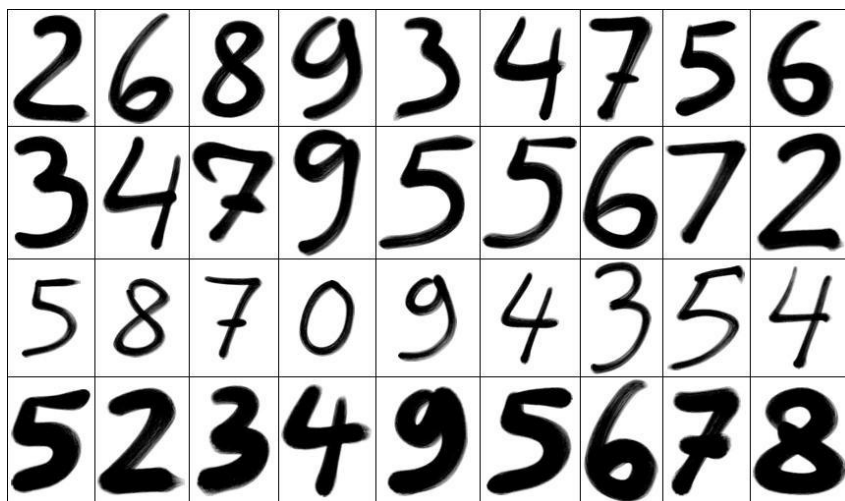
[Definition by Tom Mitchell (1998)]

Improve on task T with respect to performance P, based on experience E

T: Recognizing hand-written words

P: Percentage of words correctly classified

E: Database of human-labeled images of hand written words

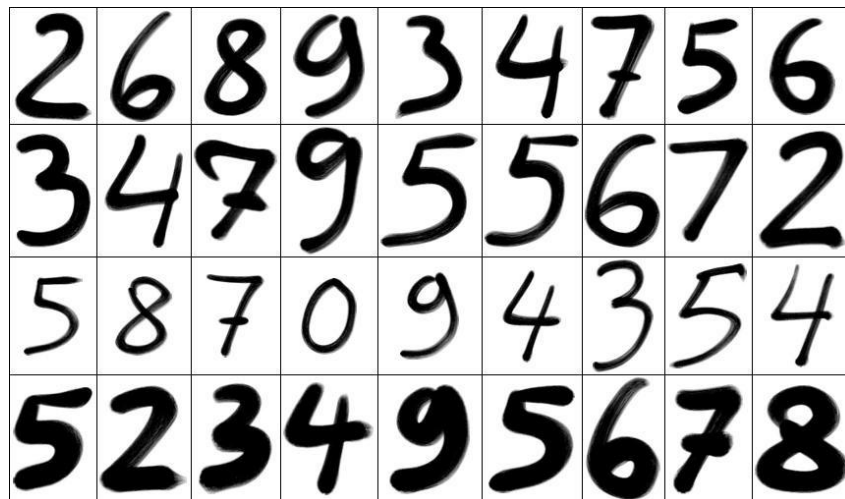


Improve on task T with respect to performance P, based on experience E

T: Recognizing hand-written words

P: Percentage of words correctly classified

E: Database of human-labeled images of hand written words



Improve on task T with respect to performance P, based on experience E





# Improve on task T with respect to performance P, based on experience E

## ❖ GPT-X language models

<https://huggingface.co/gpt2>

 Text Generation

Examples 

UCLA is the best university. In fact, it is the top-ranked academic community in the US, and is the only community in its state that makes the top ten lists of most expensive research institutions. UCLA is known for it's competitive sports



Compute

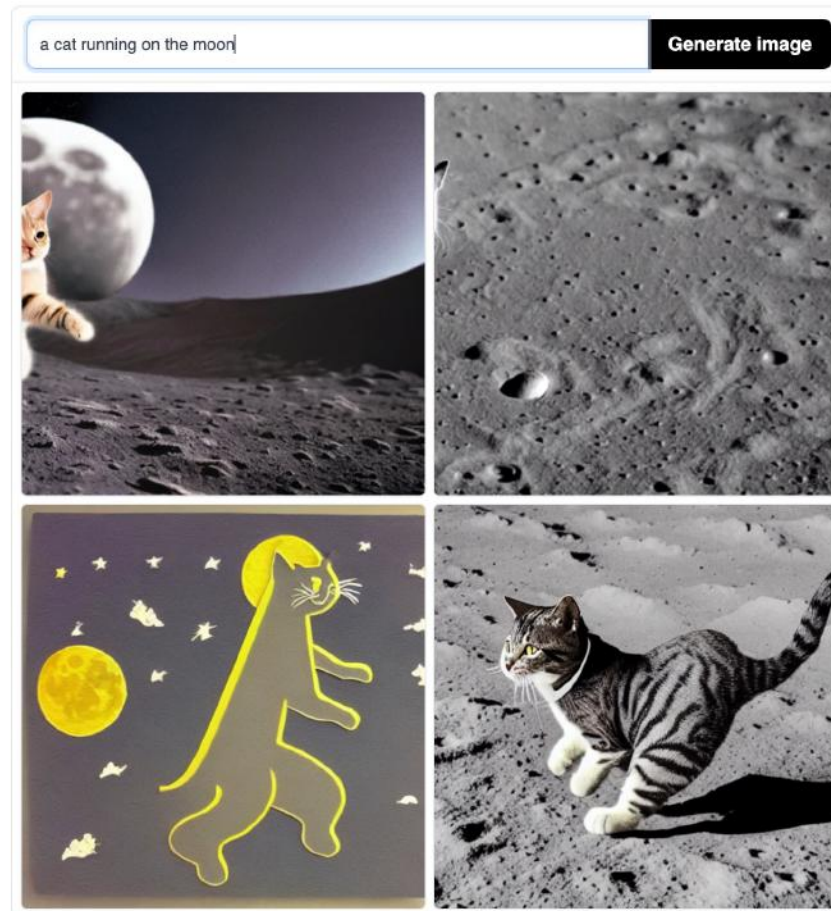
⌘+Enter

1.5

# Improve on task $T$ with respect to performance $P$ , based on experience $E$

## ❖ Stable Diffusion

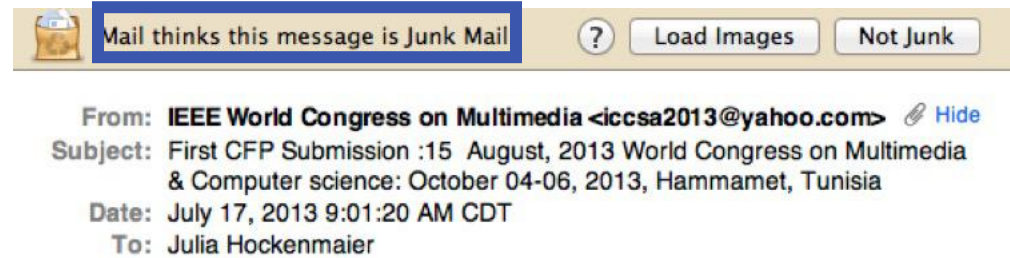
<https://huggingface.co/spaces/stabilityai/stable-diffusion>



# Discussion

- ❖ [2 min] Introduce yourself to your group
  - ❖ Your name, your major and one interesting fact
- ❖ [5 min] Brainstorm:  
What is your dream application of ML?
  - ❖ Define (Task, Performance, Experience)
  - ❖ Can be something not exist  
(e.g., a robot writing homework for you)
- ❖ [3 min] Pick the best answer and the presenter and post the presenter's name at chat box

# Applications: Spam Detection



- ❖ This is a **binary classification task**:  
Assign **one of two labels (i.e. yes/no)** to the **input** (here, an email message)
- ❖ Classification requires **a model (a classifier)** to determine which label to assign to items.
- ❖ In this class, we study **algorithms and techniques to learn such models** from data.

# The Uses of Machine Learning

- ❖ **ML is at the core of teaching machine to**
  - ❖ Understand high level cognition (e.g., vision)



# The Uses of Machine Learning

- ❖ **ML is at the core of teaching machine to**
  - ❖ Understand high level cognition
  - ❖ Perform knowledge intensive inferences



# The Uses of Machine Learning

- ❖ **ML is at the core of teaching machine to**
  - ❖ Understand high level cognition
  - ❖ Perform knowledge intensive inferences
  - ❖ Deal with messy, real world data





# Learning = Generalization

**H. Simon** (Turing Award 1975, Nobel Prize 1978)-

“Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the task or tasks drawn from the same population more efficiently and more effectively the next time.”

The ability to perform a task in a situation which has never been encountered before



# Learning = Generalization



Mail thinks this message is junk mail.

Not junk

- ❖ The learner has to be able to classify items it has never seen before.

# Administrivia

# Prerequisites

- ❖ The pillars of machine learning
  - ❖ Probability and statistics
  - ❖ Linear algebra
  - ❖ Calculus/Optimization



Image adapted from <http://pngimg.com/download/31173>  
under CC by-NC 4.0

# Prerequisites

- ❖ The pillars of machine learning

  - ❖ Probability and statistics

  - ❖ Linear algebra

  - ❖ Calculus/Optimization

- ❖ Computer science background

  - ❖ Algorithms

  - ❖ Programming experience

We will use **Python and scikit-learn**

# In-Person Lecture

- ❖ Your participation is appreciated
- ❖ Questions are welcomed
- ❖ Audio/video recording may be available at BruinLearn
  - ❖ You should not rely on these recordings as a substitute for lectures
- ❖ Although not a formal component of the grade, attendance is important

# Problem Set

## ❖ Problem Sets

- ❖ Three problem sets
- ❖ Due at 11:59pm on the due date
- ❖ 24hr late credits for the entire quarter
- ❖ Will be using GradeScope to manage submissions (submission instructions will be provided in the discussion session)
- ❖ All solutions must be clearly written or typed.
  - ❖ Unreadable answers will not be graded. We encourage using LaTeX to type answers.
  - ❖ Solutions will be graded on both correctness and clarity
- ❖ For programming HW, upload your source code at Bruinlearn

# Exams

- ❖ Midterm is a 3hr online open-book exam
- ❖ Final is a 3hr in-person closed-book exam on paper
- ❖ Exam will cover materials from the lectures and the problem sets.
- ❖ No alternate or make-up exams
  - ❖ Except for disability/medical/emergency reasons documented and communicated to the instructor prior to the exam date.
  - ❖ Exam date and time **cannot** be changed to accommodate scheduling conflicts with other classes or job fair/interview.

# Regrading request

- ❖ Must be made **within one week** after the grade is released regardless of any reason
- ❖ We reserve the right to regrade entire problem set for a given regrade request.



# Quiz

- ❖ We will have quizzes (almost) every week after week 2
- ❖ A handful multiple choice questions
- ❖ You have only one try
- ❖ One lowest quiz score will be dropped

# Final grade



❖ Default cut-off for letter grade is:

> 97	93	90	87	83	80	77	73	70	< 70
A +	A	A-	B+	B	B-	C+	C	C-	D

❖ We **will not** make adjustments for individuals

❖ E.g., no round up (i.e., 89.99 = B+)

❖ In general, we **will not** curve the final grades, but may do some adjustment if needed

❖ The cut-off score will only get lower (i.e., you may get a better letter grade)

❖ This is a **heavy** course

❖ Score distribution may be different from last year

# Academic integrity policy

## ❖ No cheating

- ❖ In particular, you are free to discuss homework problems. However, you **must** write up your own solutions (solution/program). You **must** also acknowledge all collaborators.
- ❖ Please don't use any old solution you found.
- ❖ All incidents will report to the student office
- ❖ **Don't post your HW/Exam solutions or upload course materials without consent**

# CM146 on Web



- ❖ Course website:

<https://ccle.ucla.edu/course/view/21F-COMSCIM146-1>

- ❖ Piazza:

<http://piazza.com/ucla/fall2022/m146>

- ❖ Strongly encourage students to post here (publicly or privately) rather than email staff directly (you will get a faster response this way)

- ❖ Gradescope

- ❖ Maintain homework/final
- ❖ Access through Bruinlearn

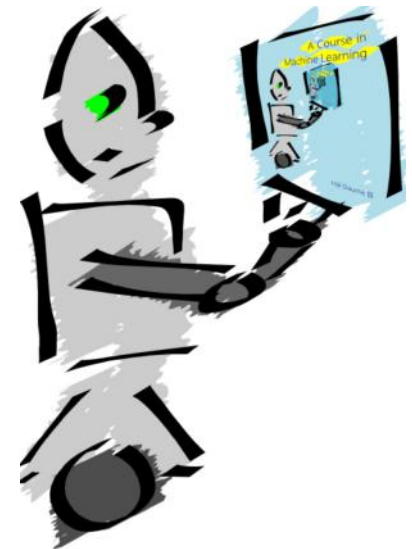
# Textbook

- ❖ No textbook

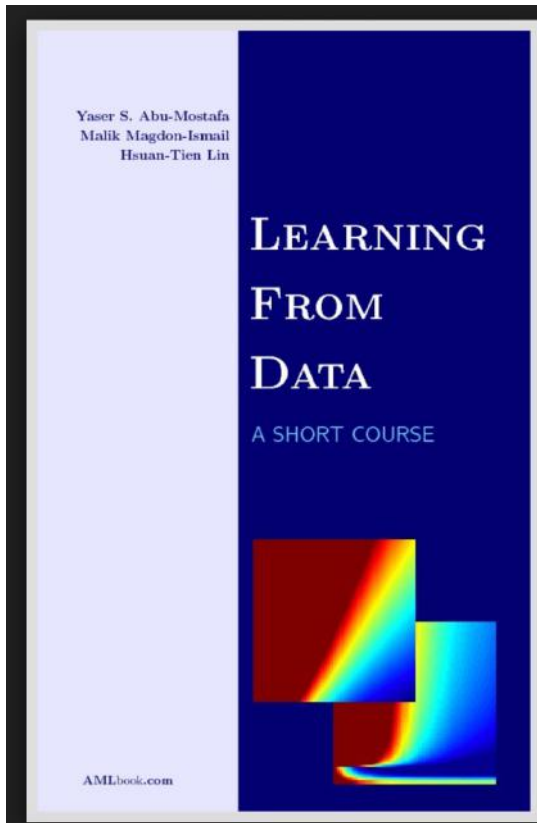
- ❖ Primary reference:

A course in machine learning by Hal Daume III (CIML). Freely available online <http://ciml.info/>

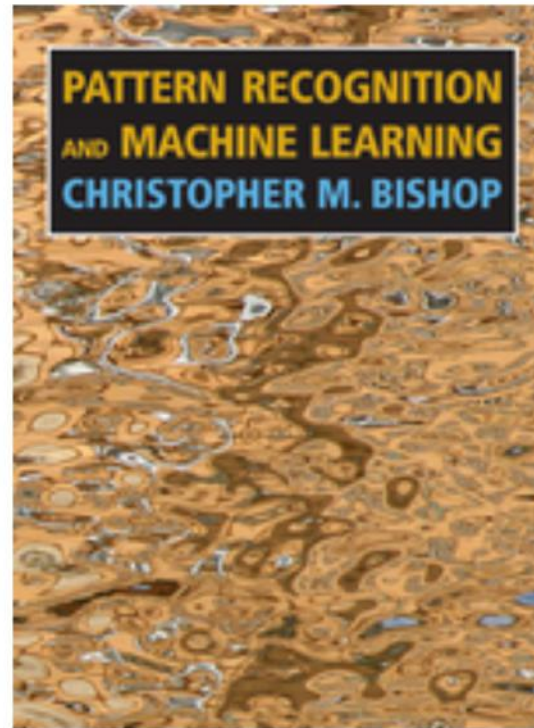
- ❖ See syllabus for the reading list



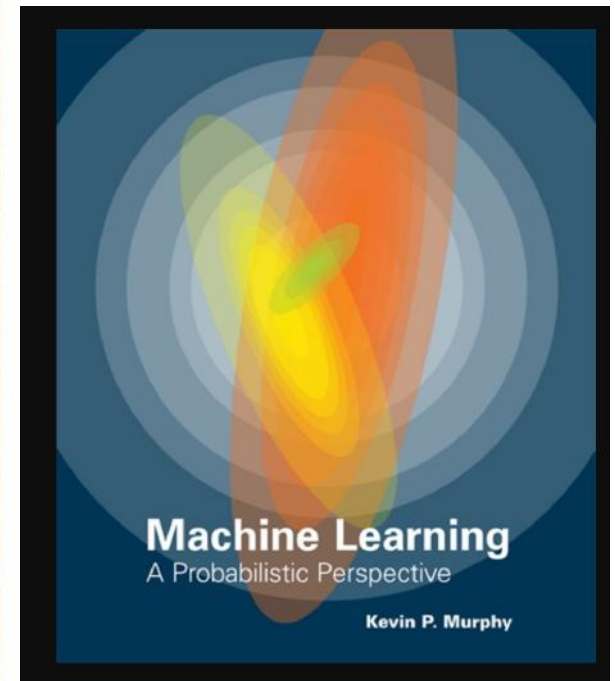
# Other references



Basic



Comprehensive



Advanced

# Why taking this course?

# Building fundamental knowledge

---

## A Regularized Framework for Sparse and Structured Neural Attention

---

Vlad Niculae\*  
Cornell University  
Ithaca, NY  
vlad@cs.cornell.edu

Mathieu Blondel  
NTT Communication Science Laboratories  
Kyoto, Japan  
mathieu@nblondel.org

### Abstract

Modern neural networks are often augmented with an attention mechanism, which tells the network where to focus within the input. We propose in this paper a new framework for sparse and structured attention, building upon a smoothed max operator. We show that the gradient of this operator defines a mapping from real values to probabilities, suitable as an attention mechanism. Our framework includes softmax and a slight generalization of the recently-proposed sparsemax as special cases. However, we also show how our framework can incorporate modern structured penalties, resulting in more interpretable attention mechanisms, that focus on entire segments or groups of an input. We derive efficient algorithms to compute the forward and backward passes of our attention mechanisms, enabling their use in a neural network trained with backpropagation. To showcase their potential as a drop-in replacement for existing ones, we evaluate our attention mechanisms on three large-scale tasks: textual entailment, machine translation, and sentence summarization. Our attention mechanisms improve interpretability without sacrificing performance; notably, on textual entailment and summarization, we outperform the standard attention mechanisms based on softmax and sparsemax.

### 1 Introduction

Modern neural network architectures are commonly augmented with an attention mechanism, which tells the network where to look within the input in order to make the next prediction. Attention-augmented architectures have been successfully applied to machine translation [2, 29], speech recognition [10], image caption generation [44], textual entailment [38, 31], and sentence summarization [39], to name but a few examples. At the heart of attention mechanisms is a mapping function that converts real values to probabilities, encoding the relative importance of elements in the input. For the case of sequence-to-sequence prediction, at each time step of generating the output sequence, attention probabilities are produced, conditioned on the current state of a decoder network. They are then used to aggregate an input representation (a variable-length list of vectors) into a single vector, which is relevant for the current time step. That vector is finally fed into the decoder network to produce the next element in the output sequence. This process is repeated until the end-of-sequence symbol is generated. Importantly, such architectures can be trained end-to-end using backpropagation.

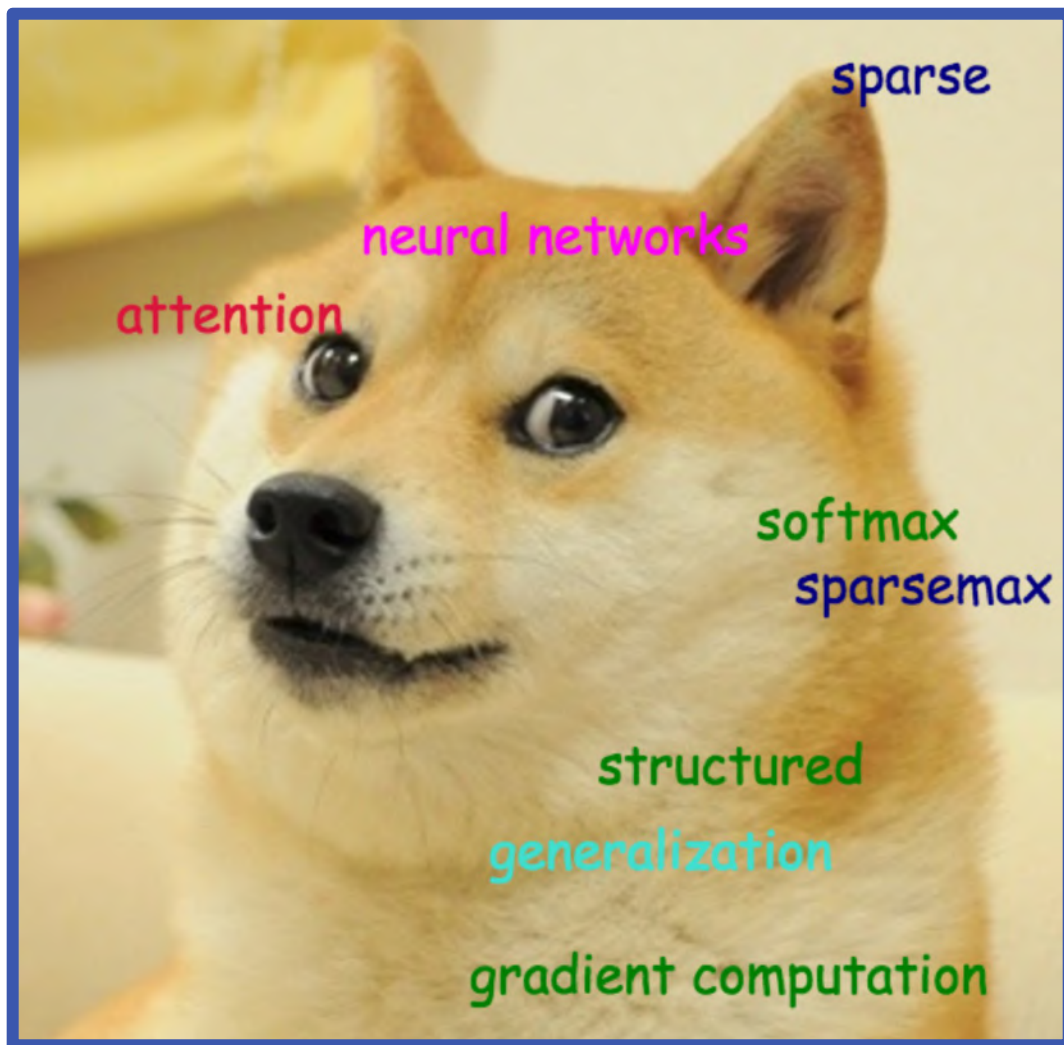
Alongside empirical successes, neural attention—while not necessarily correlated with human attention—is increasingly crucial in bringing more **interpretability** to neural networks by helping explain how individual input elements contribute to the model's decisions. However, the most commonly used attention mechanism, *softmax*, yields dense attention weights: all elements in the input always make at least a small contribution to the decision. To overcome this limitation, *sparsemax* was recently proposed [31], using the Euclidean projection onto the simplex as a sparse alternative to

\*Work performed during an internship at NTT Communication Science Laboratories, Kyoto, Japan.

Modern **neural networks** ..... **attention mechanism**, ... We propose in this paper a new framework for **sparse** and **structured** attention, building upon a **smoothed max operator**. We show that the **gradient** of this operator defines a mapping from real values to **probabilities**, suitable as an attention mechanism. Our framework includes **softmax** and a slight **generalization** of the recently-proposed **sparsemax** as special cases. ....



# What it looks like to ML researchers



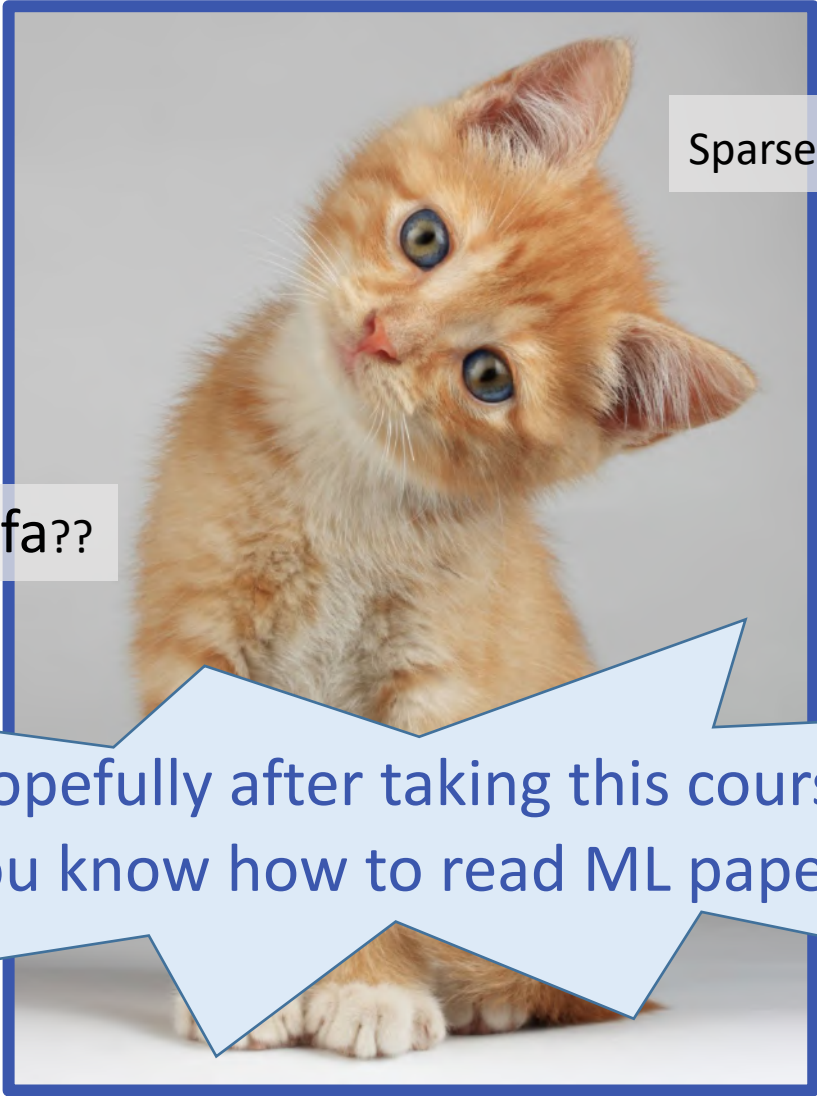
# What it looks like to normal people



Sparse?? Spaghetti??

softmax?? Sofa??

# What it looks like to normal people

A ginger and white kitten is shown from the chest up, looking upwards and slightly to the right. It has large, blue eyes and a small pink nose. The kitten is sitting on a white surface. The background is a plain, light gray.

softmax?? Sofa??

Sparse?? Spaghetti??

Hopefully after taking this course  
you know how to read ML papers

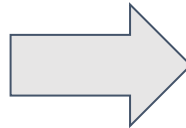
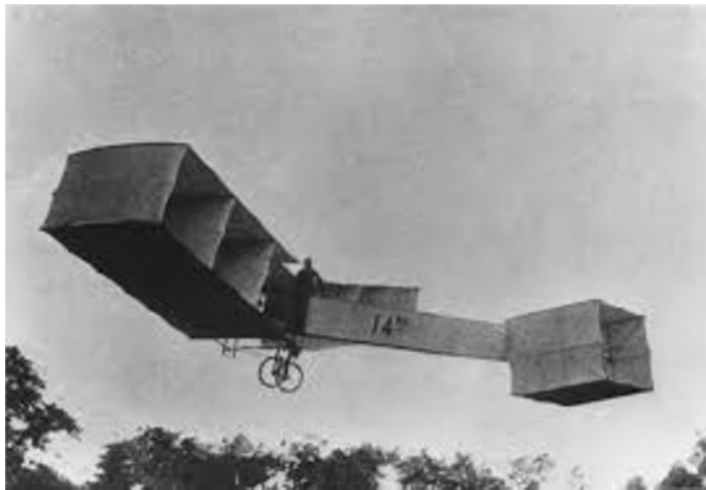
# Goals of this course: Learn about

- ❖ Fundamental concepts and algorithms
  - ❖ Customize your own algorithm
- ❖ Common techniques/tools used
  - ❖ theoretical understanding
  - ❖ practical implementation
  - ❖ best practices
- ❖ How to "debug" ML system
- ❖ Black magic => systematic process

# Why Study Machine Learning Now?

- ❖ Exciting moments for ML:
  - ❖ Initial **algorithms** and **theory** in place.
  - ❖ Growing amounts of on-line data
  - ❖ Computational power available.

# Current Status



## Current status:

- Compelling results on benchmarks,
- Work well in general domain
- Commercial uses

## Challenges:

- Incorporate w/ human knowledge
- Reliable (fair, robust, interpretable) models that earn human trust
- Applications in specific domains
- Skewness of available annotation data

# What will we learn?

- ❖ Supervised learning
  - ❖ Decision tree, Perceptron, Linear models, support vector machines, kernel methods, probabilistic models
- ❖ Unsupervised learning
  - ❖ Clustering,
  - ❖ EM algorithms
- ❖ Learning theory
- ❖ Deep learning (representation learning)
- ❖ Practical Issues
  - ❖ Experimental evaluation; Implementing ML models

# Machine Learning is Interdisciplinary

- ❖ Makes Use of:
  - ❖ Probability and Statistics; Linear Algebra; Calculus; Theory of Computation;
- ❖ Related to:
  - ❖ Philosophy, Psychology ,Neurobiology, Linguistics, Vision, Robotics,....
- ❖ Has applications in:
  - ❖ AI (Natural Language; Vision; Planning; HCI)
  - ❖ Engineering (Agriculture; Civil; ...)
  - ❖ Computer Science (Compilers; Architecture; Systems; data bases...)



# Other Related Courses

- ❖ CS145 - Introduction to Data Mining
- ❖ CM148 - Introduction to Data Science
- ❖ CS161 - Fundamentals of Artificial Intelligence
- ❖ Computer Vision/Bioinformatic/NLP
- ❖ Related course at ECE, Stats, Math dep.
- ❖ Graduate-level courses

# Challenges in ML

# Structured Inference

- ❖ Many predictions are compositional
  - ❖ Require an inference process

# Structured Inference



Carefully  
Slide



# Structured Inference

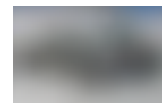
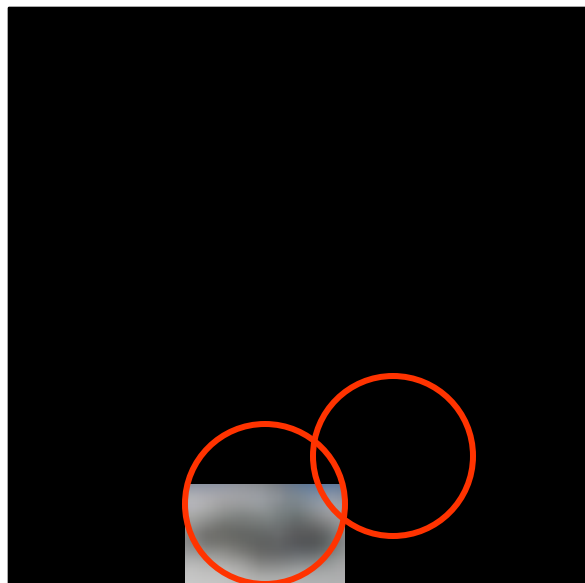


小心:  
Carefully  
Careful  
Take  
Care  
Caution

地滑:  
Slide  
Landslip  
Wet Floor  
Smooth



# Structured Inference

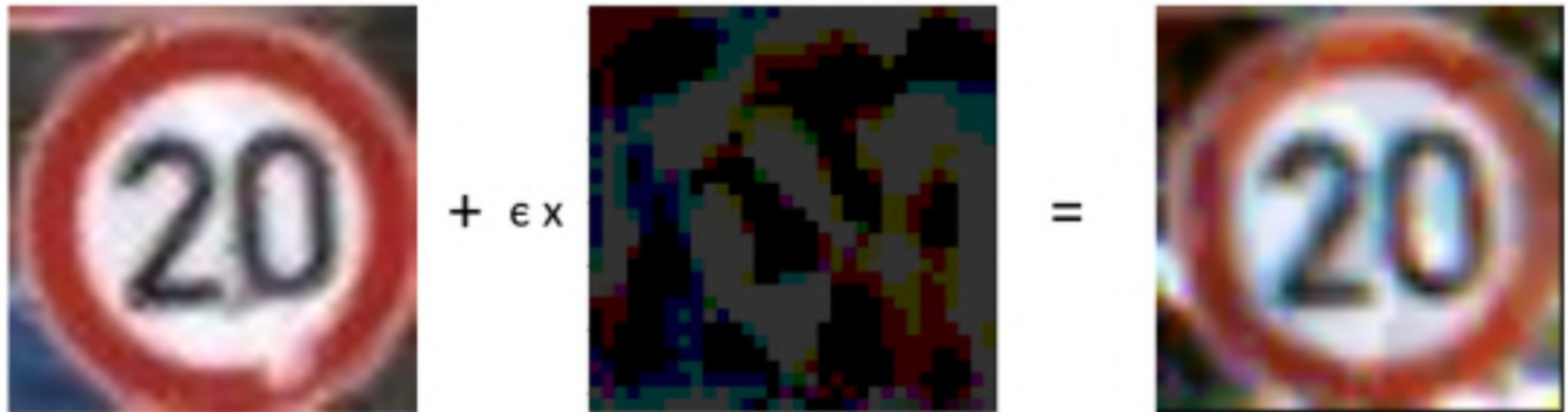


# Robustness

Car or shoe?



# Adversarial Attack



93%, 20 Km/h Sign

$+ \epsilon x$

$\text{sign}(\nabla * J(\theta, x, y))$

$=$



90%, 80 Km/h Sign



<https://arxiv.org/abs/1712.09327v1>



# Fairness in ML

Select photo  

✗ The photo you want to upload does not meet our criteria because:

- Subject eyes are closed

Please refer to the technical requirements.  
You have 9 attempts left.

Check the photo [requirements](#).

[Read more about common photo problems and](#)

start again and re-enter the CAPTCHA security check.

Reference number: 20161206-81

Filename: Untitled.jpg

If you wish to [contact us](#) about the photo, you must provide us with the reference number given above.

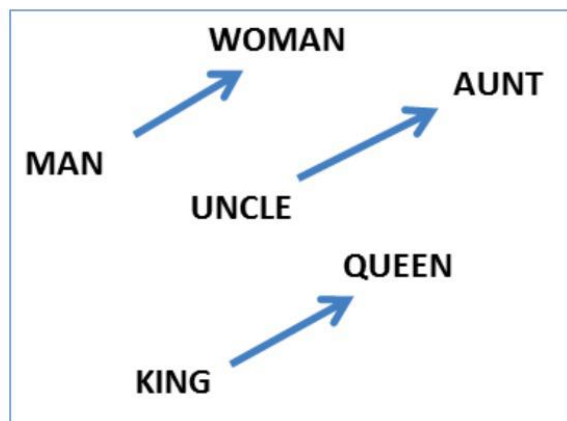
Please print this information for your records.



Subject eyes are closed

# Fairness in ML-- Word embedding bias

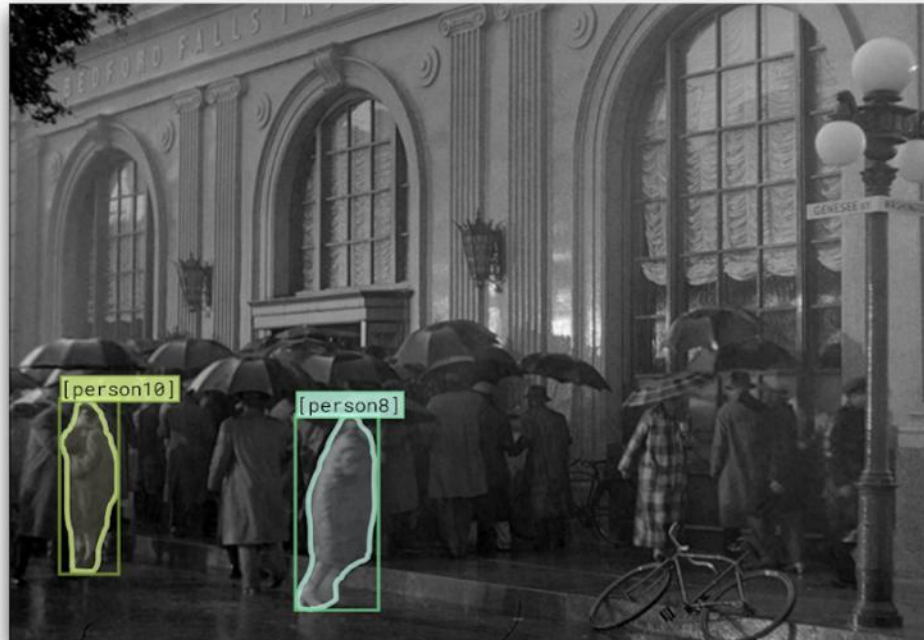
$$\diamond v_{man} - v_{woman} + v_{uncle} \sim v_{aunt}$$



he: ____	she: ____
uncle	aunt
lion	
surgeon	
architect	
beer	
professor	

We use Google w2v embedding trained from the news

# Commonsense



**Is it raining outside?**

- a) Yes, it is snowing.
- b) Yes, [person8] and [person10] are outside.
- c) No, it looks to be fall.
- d) Yes, it is raining heavily.

*An example from the VCR dataset*