

Lecture 1:

Introduction to Machine Learning

Fall 2018

Kai-Wei Chang

CS @ UCLA

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.

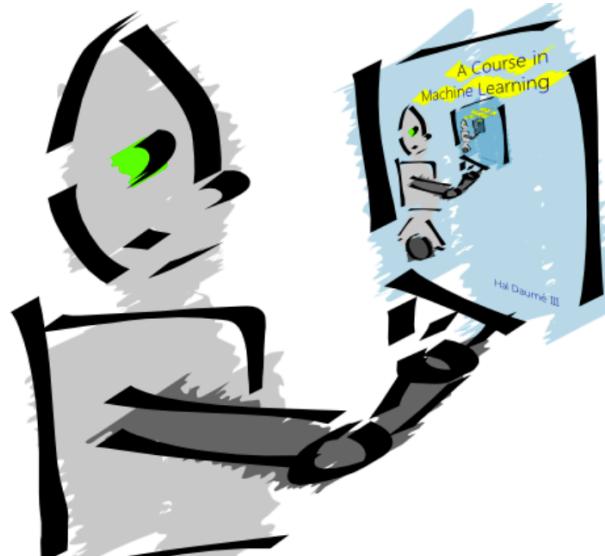
Registration

- ❖ We MAY give PTE in the second week
- ❖ Limited by the size of the classroom and TAs
- ❖ Priority:
 - ❖ CS undergraduate
 - ❖ Other undergraduate
 - ❖ CS graduate
 - ❖ Other graduate

What is machine learning?

Machine learning is about predicting the future based on the past.

-- Hal Daume III



Slide credit: David Kauchak

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 metric P, based on experience E

T: Playing checkers

P: Percentage of games won against an arbitrary opponent

E: Playing practice games against itself

T: Recognizing hand-written words

P: Percentage of words correctly classified

E: Database of human-labeled images of handwritten words

T: Driving on four-lane highways using vision sensors

P: Average distance traveled before a human-judged error

E: A sequence of images and steering commands recorded while observing
a human driver

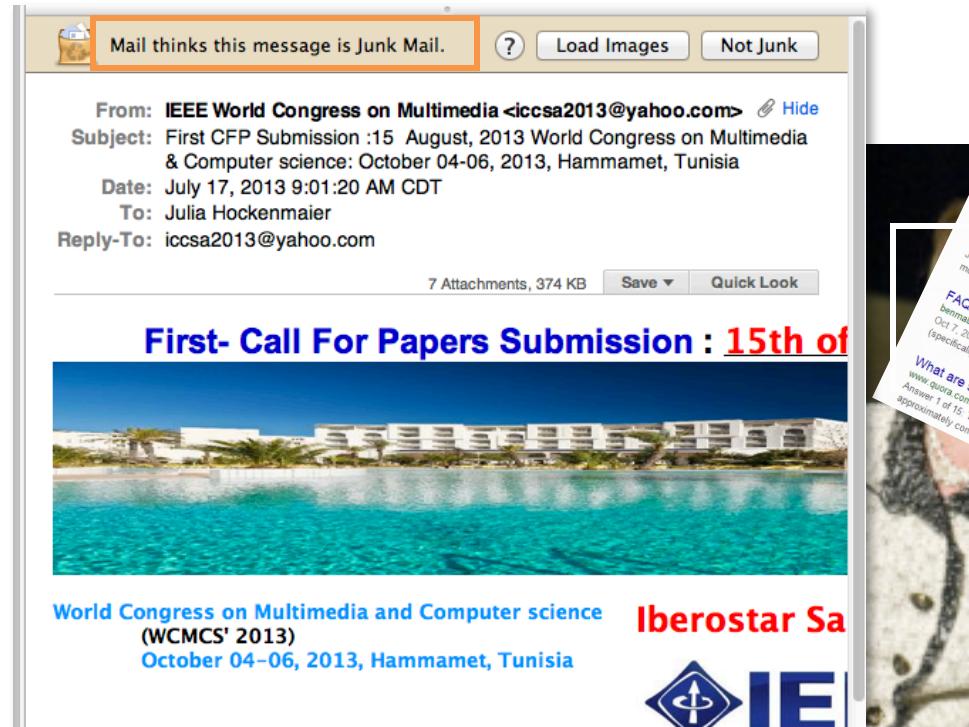
T: Categorize email messages as spam or legitimate

P: Percentage of email messages correctly classified

E: Database of emails, some with human-given labels

Slide credit: Ray Mooney

Machine learning is everywhere



A screenshot of a machine translation application. At the top, there are dropdown menus for "From: English - detected" and "To: Chinese (Simplified)". Below these are buttons for "Translate", "Chinese (Simplified)", "English", and "Spanish". A text input field contains the sentence "The blue fox jumps over the hedge". The translated output in Chinese is "蓝狐跨越对冲".

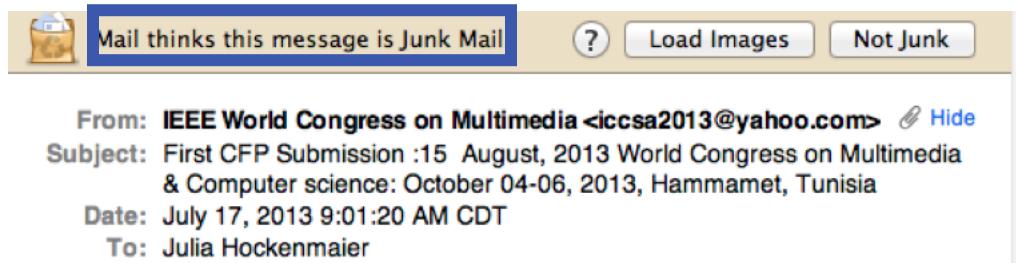
A collage of screenshots illustrating various applications of machine learning:

- A Microsoft Azure Machine Learning service page.
- A Gsense Predictive Sales page.
- An Amazon search results page for "Machine Learning Books".
- A Google search results page for "Machine Learning Books".
- An Amazon product page for "Meaning: A Slim Guide to Semantics (Oxford Linguistics)" by Paul Elbourne.
- A user interface for a machine learning recommendation system.

On the right side of the collage, there are rating and feedback options:

- Rate this item: 5 stars 4 stars 3 stars 2 stars 1 star
- I own it
- Not interested
- This was a gift
- Don't use for recommendation
- Help

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.

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@mblondel.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

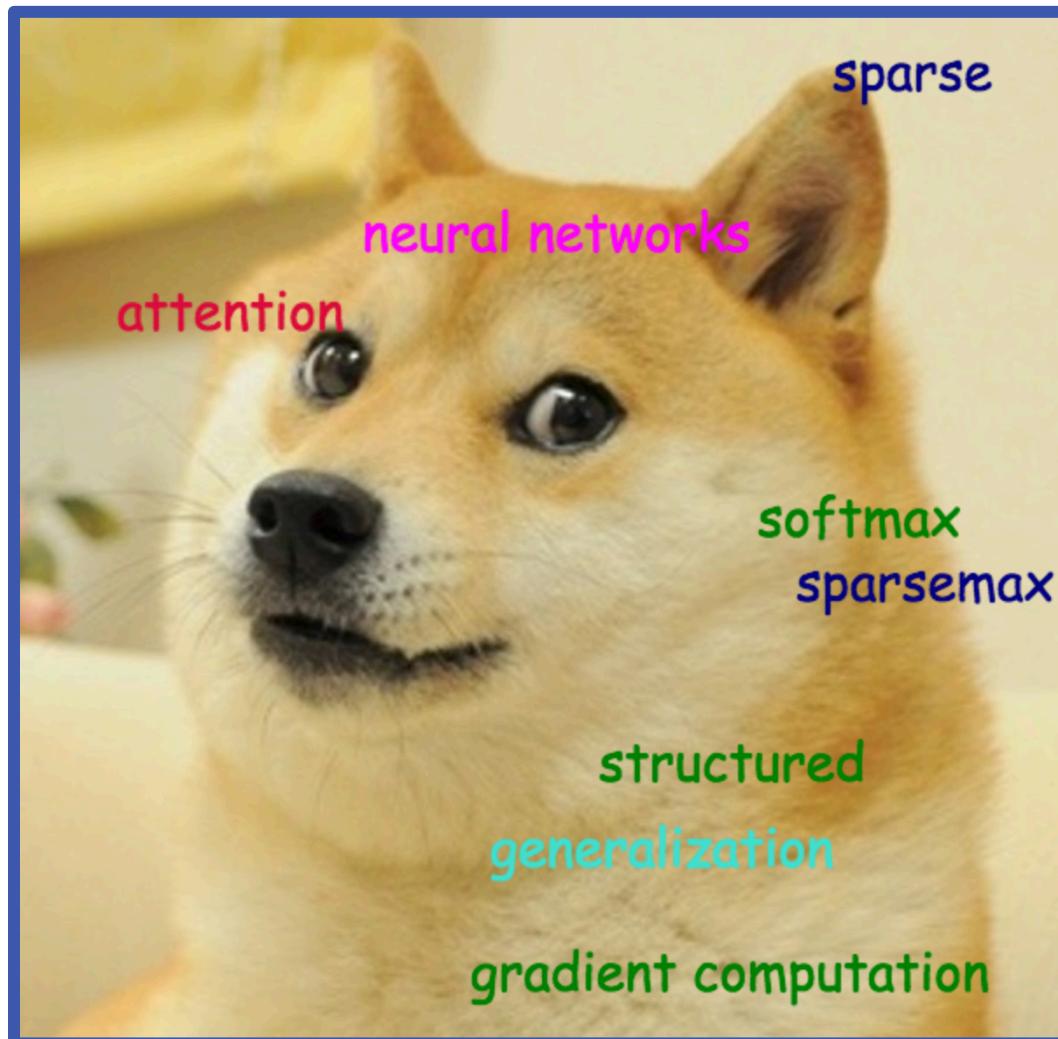
Modern neural network architectures are commonly augmented with an attention mechanism, which tells the network where to look within the input in order to make the next prediction. Attention-augmented architectures have been successfully applied to machine translation [2, 29], speech recognition [10], 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

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**.

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

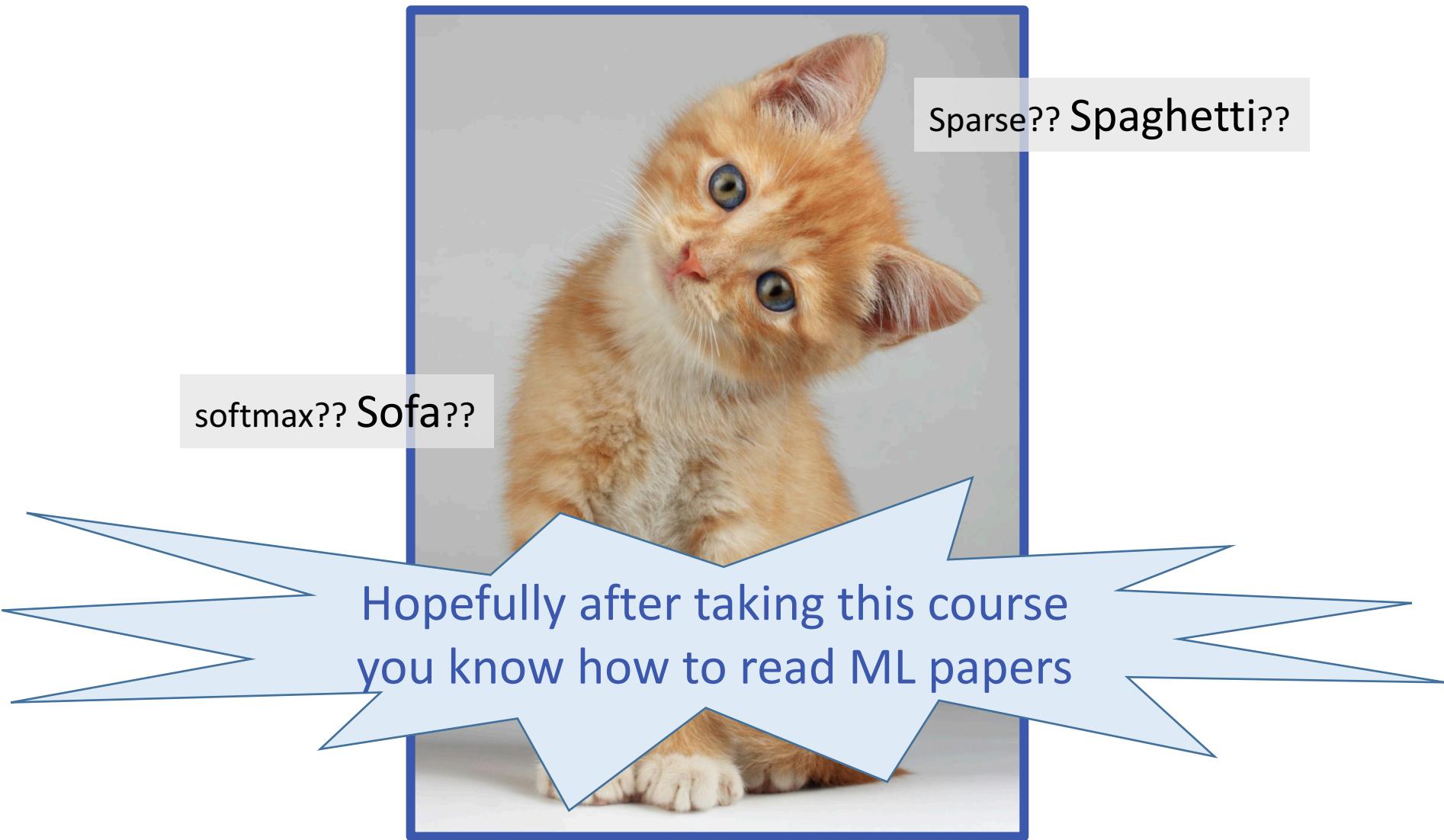
What it looks like to ML researchers



What it looks like to normal people



What it looks like to normal people



What ML beginners may do

How to train an
image classifier?

Deep Learning!

How to train a
model on a small
dataset

Deep Learning!

How to sort *five*
numbers

Deep Learning!

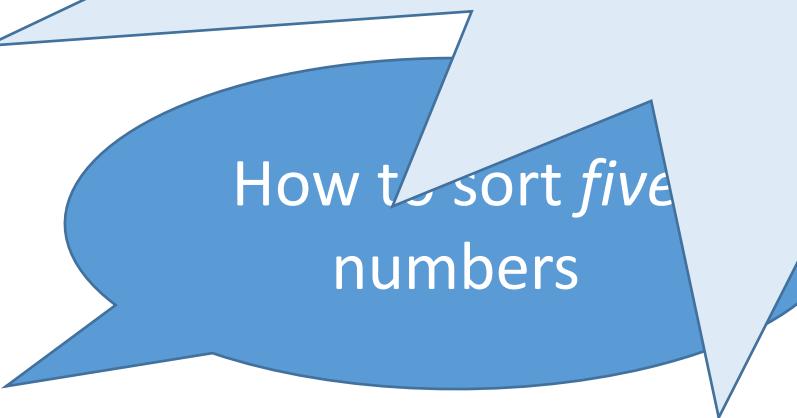


What does ML beginners may do

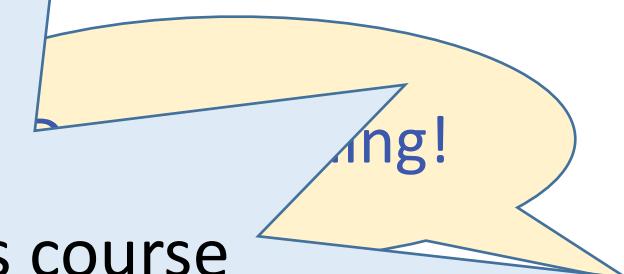


How to train an image classifier?

Hopefully after taking this course you know how to choose a **right** model for your application (theoretically or empirically).



How to sort five numbers



Deep Learning!

Goals of this course: Learn about

- ❖ Fundamental concepts and algorithms
- ❖ Common techniques/tools used
 - ❖ theoretical understanding
 - ❖ practical implementation
 - ❖ best practices

What will we learn?

- ❖ Supervised learning
 - ❖ Decision tree, Perceptron, Linear models, support vector machines, kernel methods
 - ❖ Learning theory
- ❖ Unsupervised learning
 - ❖ Clustering, Hidden Markov Models
 - ❖ EM algorithms
- ❖ Practical Issues
 - ❖ Experimental evaluation; Implementing ML models

Administrivia

CS M146 Team

- ❖ Kai-Wei Chang (Eng VI 374)
 - ❖ Monday, 3:00 PM – 4:00 PM (or: appointment)
- ❖ TAs



[Sajad Darabi](#)



[Sudharsan Krishnaswamy](#)



[MD Rizwan Parvez](#)



[Yichao Zhou](#)

Prerequisites

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



Prerequisites

- ❖ The pillars of machine learning
 - ❖ Probability and statistics
 - ❖ Linear algebra
 - ❖ Calculus/Optimization
- ❖ Computer science background
 - ❖ Algorithms
 - ❖ Programming experience: we will use Python, numpy and scikit-learn

Math background

- ❖ Mini quiz on math background (in CCLE, Due 10/22)

Course format

- ❖ Problem Sets
 - ❖ Four problem sets (1 -- 4)
 - ❖ Due at 11:59pm on the due date
 - ❖ 24hr late credits for the entire quarter, that's it.
 - ❖ Will be using gradescope to manage submissions
(will send out submission instruction)
- ❖ 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

Exams

- ❖ Midterm (week 1-4) and Final (cover all)
- ❖ Exams are in class, closed-book and closed-notes and 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 dates and times **cannot** be changed to accommodate scheduling conflicts with other classes or job fair/interview.

Policies

- ❖ Attendance and class participation
 - ❖ Although not a formal component of the grade, Attendance is important
 - ❖ We look forward to your active participation.
 - ❖ If you are absent without a documented excuse, the instructor and TA will not be able to go over missed lecture material
- ❖ Video recordings
 - ❖ we aim to make it available (not guarantee).
 - ❖ You should not rely on these recordings as a substitute for lectures.

Policies

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

Final grade

- ❖ Default cut-off for letter grade is:

> 96	93	90	86	83	80	76	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
 - ❖ The cut-off score will only get lower (i.e., you may get a better letter grade)
- ❖ This is a **heavy** course



Academic integrity policy

- ❖ **No cheating**
 - ❖ Homework and Exam
 - ❖ 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

CM146 on Web



Homework set -1

- ❖ Course website:

<https://ccle.ucla.edu/course/view/18F-COMSCIM146-1>

- ❖ Piazza:

<https://piazza.com/ucla/fall2018/csm146>

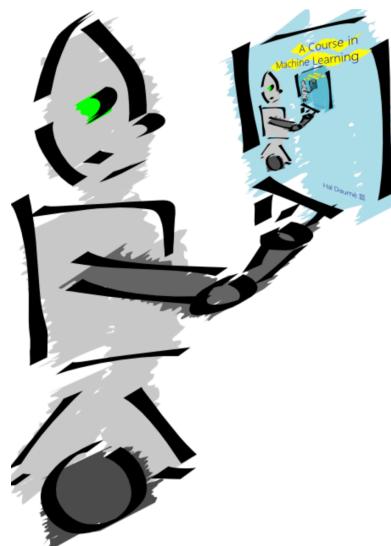
- ❖ Strongly encourage students to post here (publicly or privately) rather than email staff directly (you will get a faster response this way)

- ❖ Gradescope: <https://www.gradescope.com>

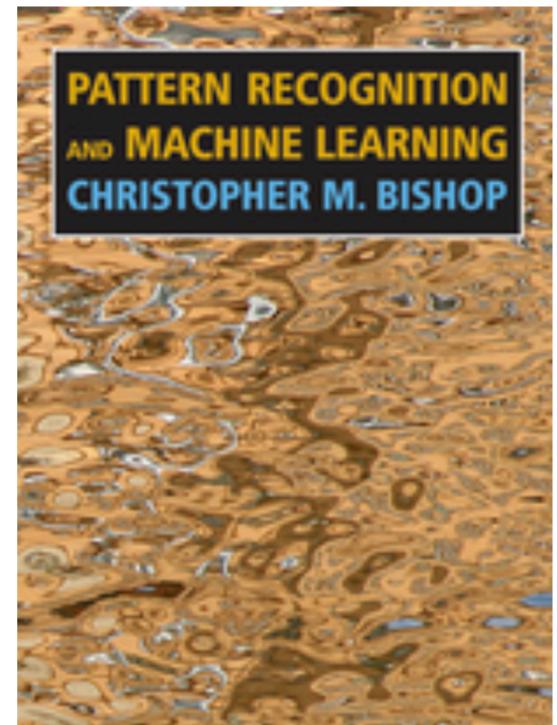
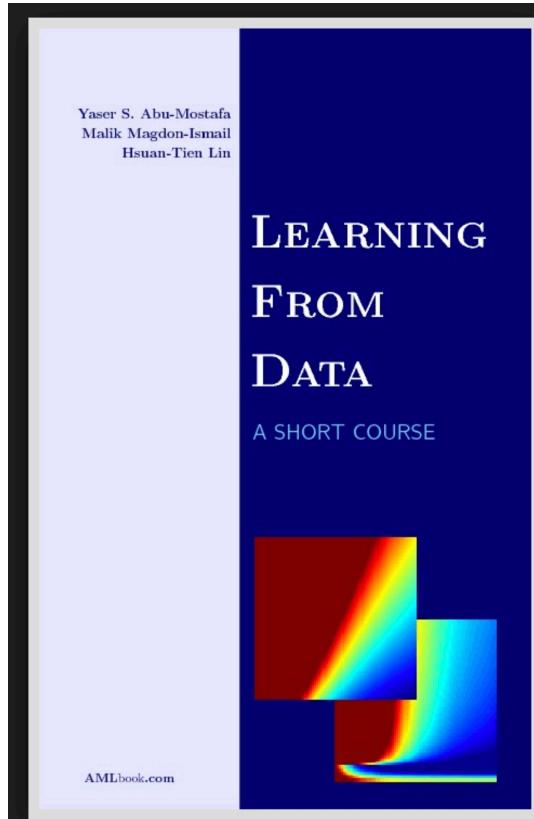
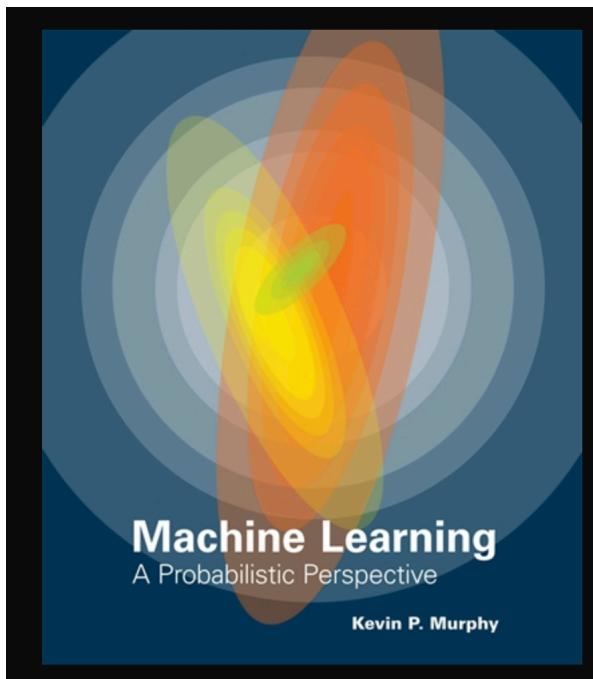
- ❖ Maintain homework/grade
 - ❖ Use Entry code: 9DJNJR to enroll

Textbook

- ❖ No textbook
 - ❖ Primary reference:
A course in machine learning by Hal Daume III
(CIML). Freely available online <http://ciml.info/>



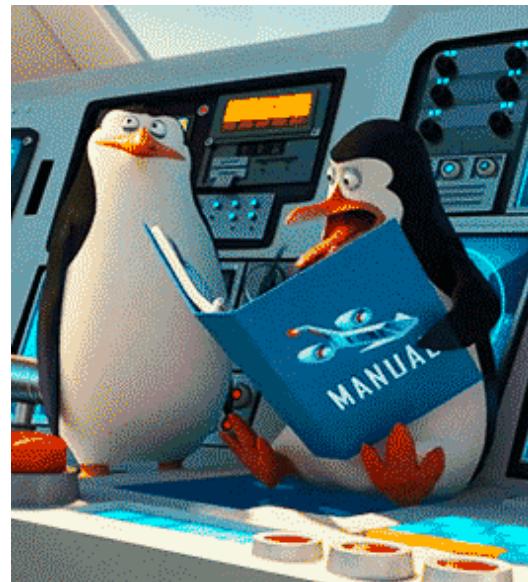
Other references



What is machine learning?

Learning

- ❖ **Learning is at the core of**
 - ❖ Understanding high level cognition
 - ❖ Performing knowledge intensive inferences
 - ❖ Building adaptive, intelligent systems
 - ❖ Dealing with messy, real world data



Learning

- ❖ **Learning has multiple purposes**
 - ❖ Knowledge acquisition
 - ❖ Integration of various knowledge sources to ensure robust behavior
 - ❖ Adaptation (human, systems)
 - ❖ Decision Making (Predictions)

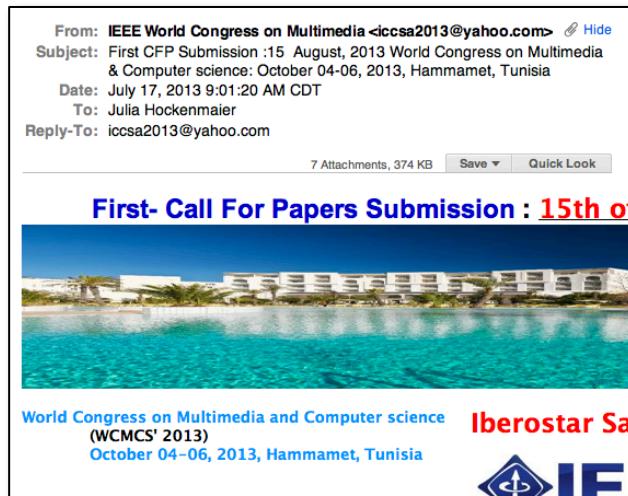
Learning = Generalization

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.

Learning = Generalization

- ❖ Classification

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

- ❖ Medical diagnosis; credit card applications; hand-written letters; ad selection; sentiment assignment,...

- ❖ Planning and acting

- ❖ Navigation; game playing (chess, backgammon, go); driving a car

- ❖ Skills

- ❖ Balancing a pole; playing tennis

- ❖ Common sense reasoning

- ❖ Natural language interactions

Why Machine Learning?

- ❖ Computer systems with new capabilities.
 - ❖ Develop systems that are too difficult or impossible to construct manually .
 - ❖ Develop systems that can automatically adapt and customize themselves to the needs of the individual user through experience.
 - ❖ Discover knowledge and patterns in databases, e.g. discovering purchasing patterns
 - ❖ Solve the kinds of problems now reserved for humans.

Why Study Learning?

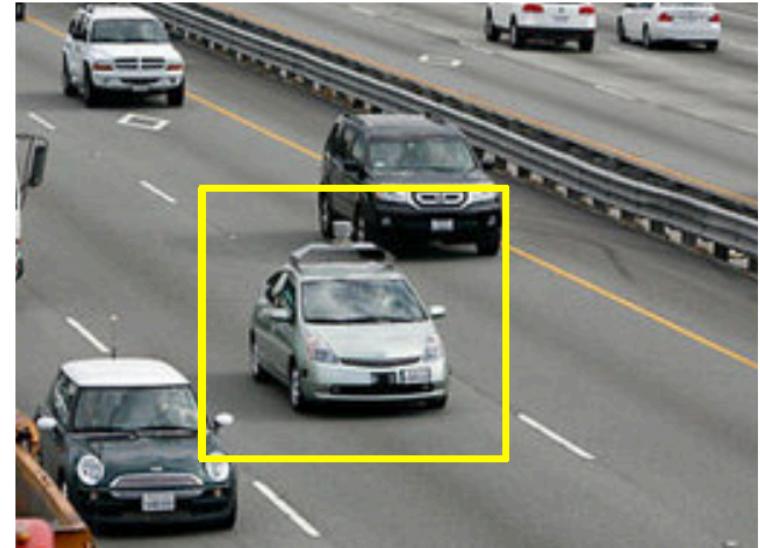
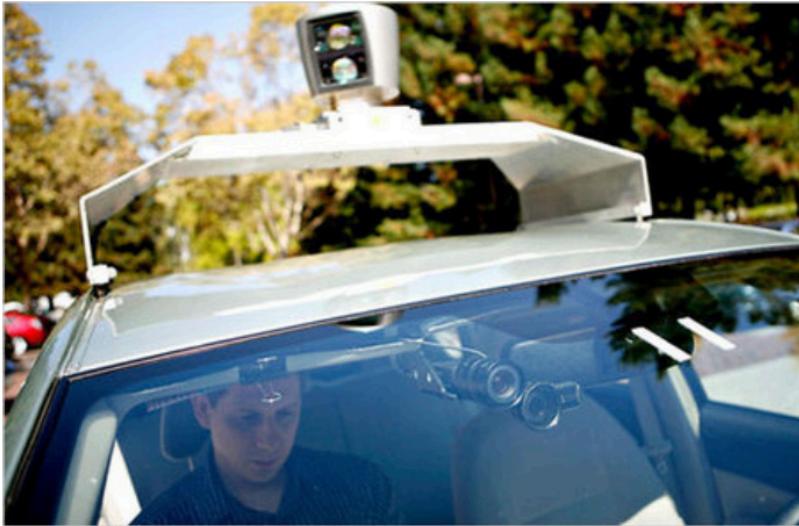
- ❖ Computer systems with new capabilities
- ❖ Understand human and biological learning
- ❖ Exciting moments for ML:
 - ❖ Initial **algorithms** and **theory** in place.
 - ❖ Growing amounts of on-line data
 - ❖ Computational power available.

Work in Machine Learning

- ❖ 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...)

State of the art applications of ML

Autonomous Cars



- Nevada made it legal for autonomous cars to drive on roads in June 2011
- As of 2013, four states (Nevada, Florida, California, and Michigan) have legalized autonomous cars



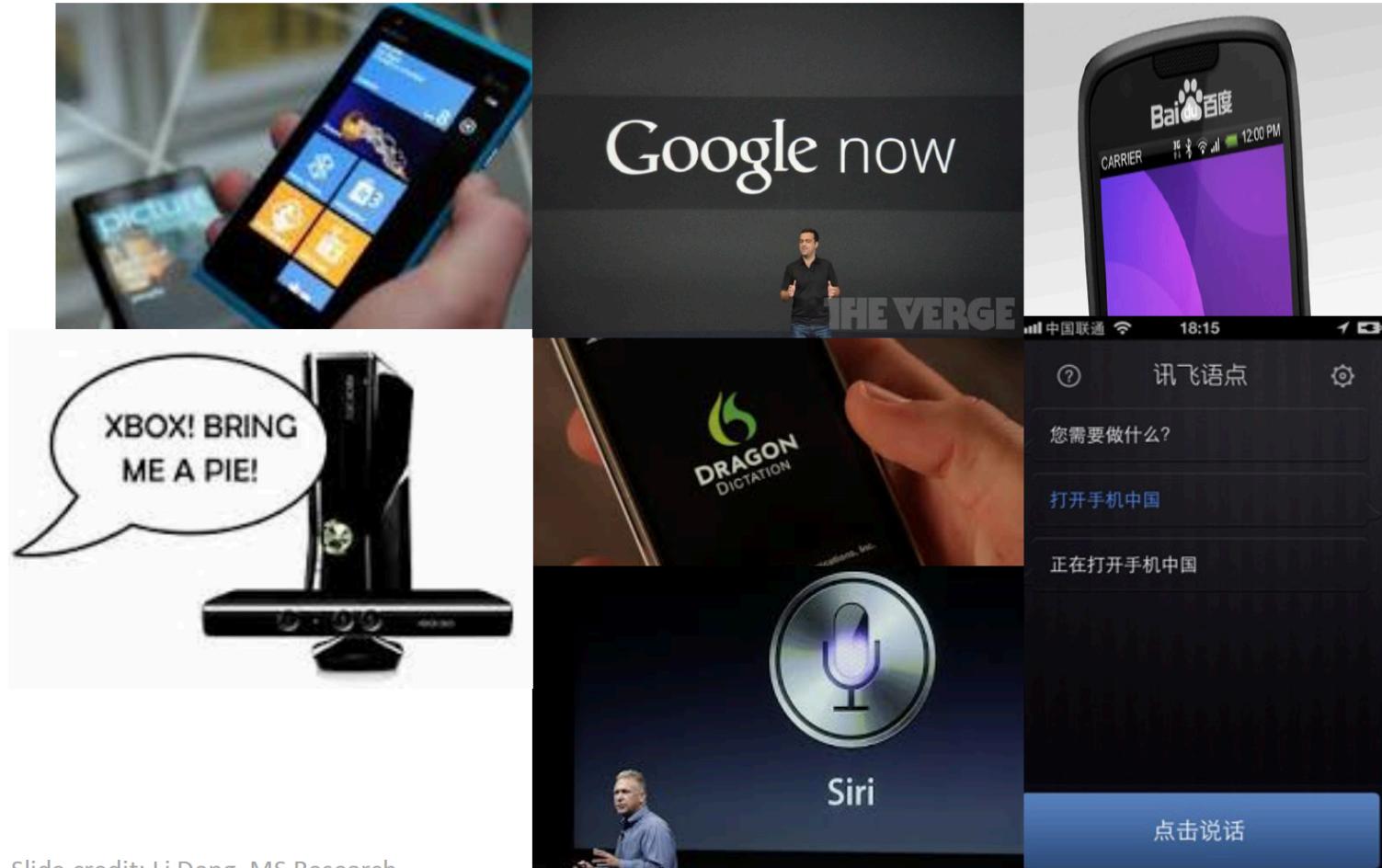
State of the art applications of ML

Computer Vision



State of the art applications of ML

Speech Recognition



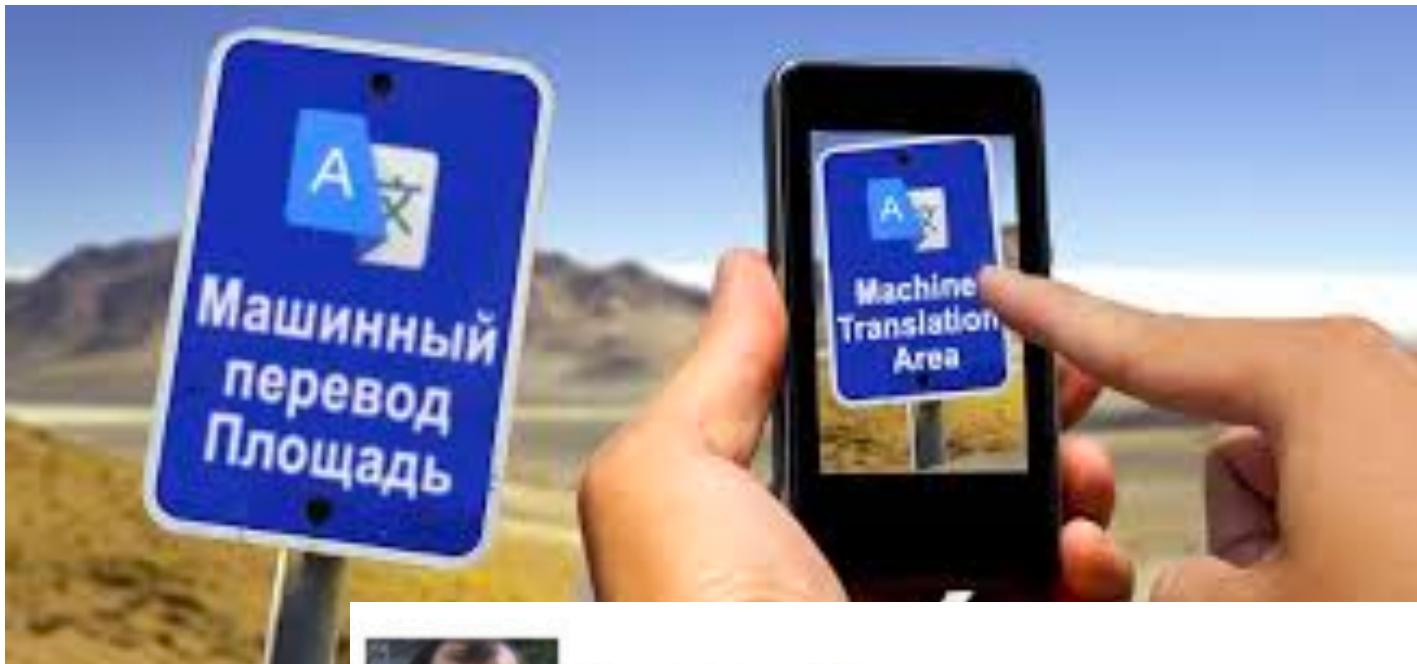
Slide credit: Li Deng, MS Research

State of the art applications of ML

❖ Reinforcement Learning



Machine translation



Necip Fazil Ayan

1 hr ·



Onların, İzmir'in neden hayır dediğini anlamalarını beklemiyoruz.

We don't expect them to understand why Izmir said no.



• Rate this translation

Learning in the future

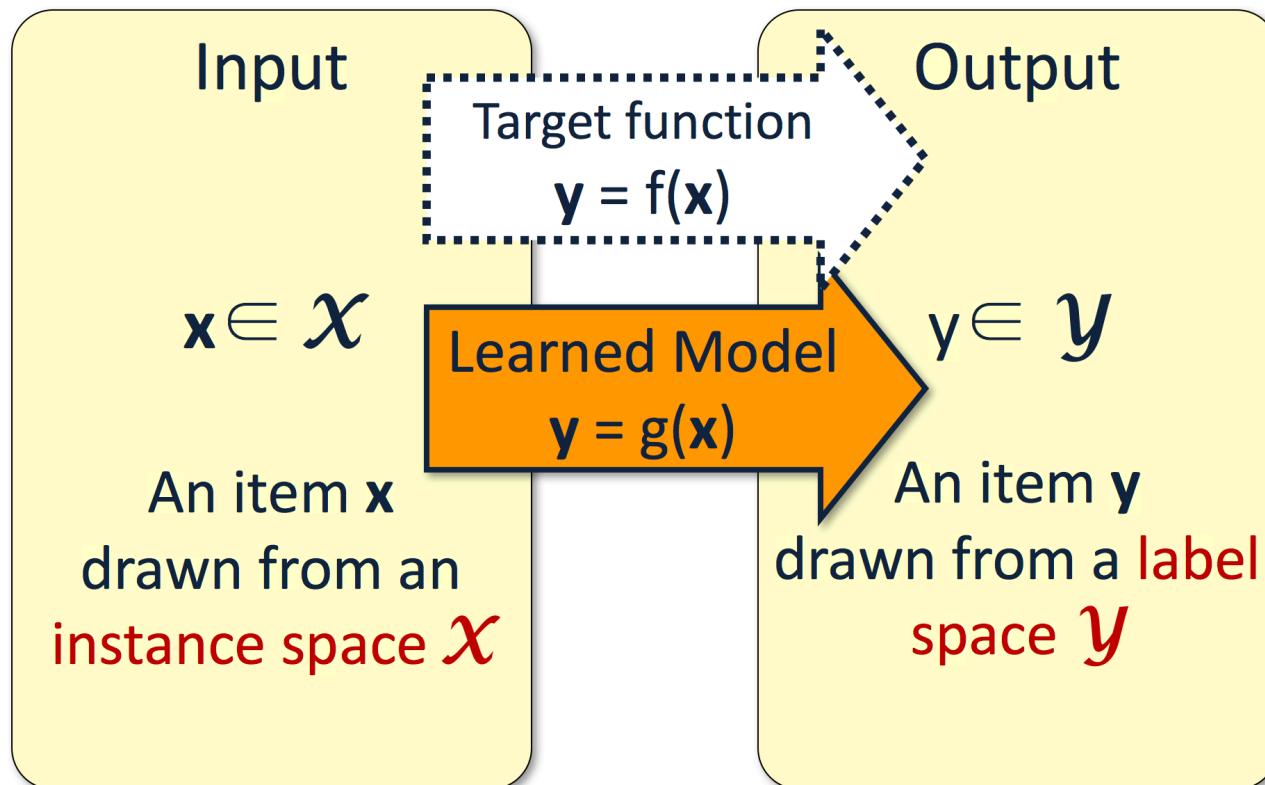
- ❖ Learning techniques will be a basis for application that involves a connection to the messy real world
- ❖ Prospects for broader future applications make for fundamental research and development opportunities
- ❖ Many unresolved issues – Theory and Systems
 - ❖ While it's hot, there are many things we don't know how to do

Type of learning protocols

Types of learning (protocols)

❖ Supervised learning

- ❖ Given: **labeled** training instances (or examples)
- ❖ Goal: learn mapping that predicts label for test instance



Training phase:



lion



Not lion

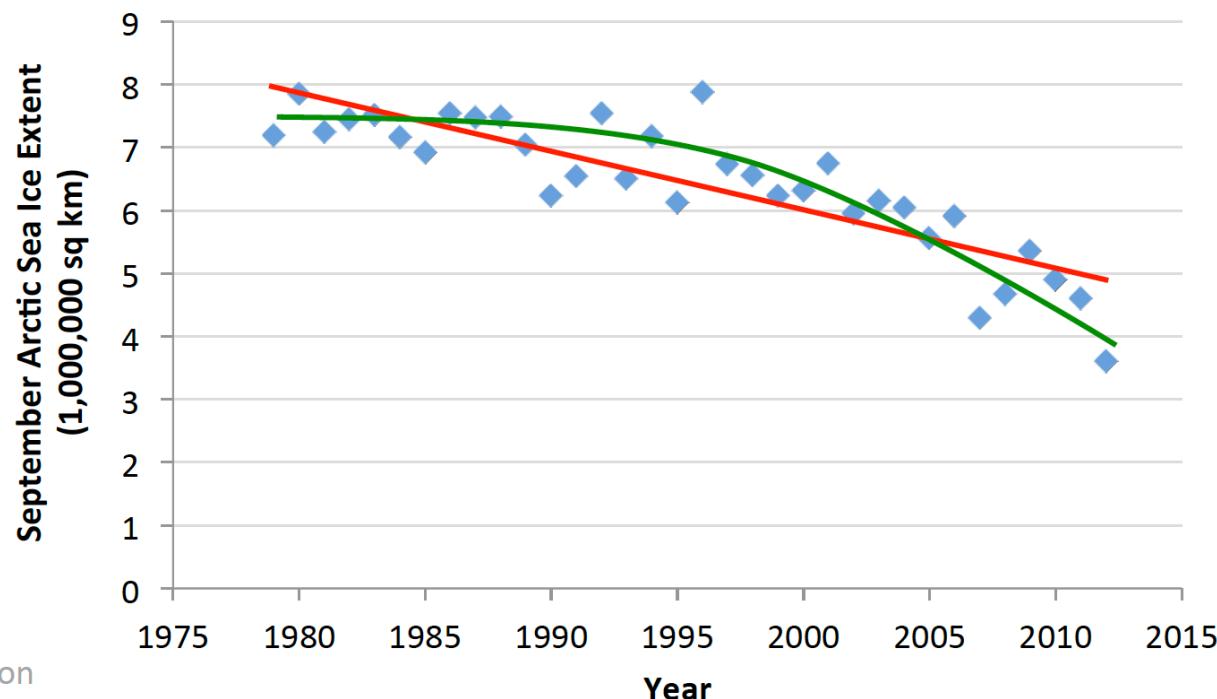
Test phase:



??

Supervised Learning (Regression)

- Given $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- Learn a function $f(x)$ to predict y given x
 - y is real-valued == regression



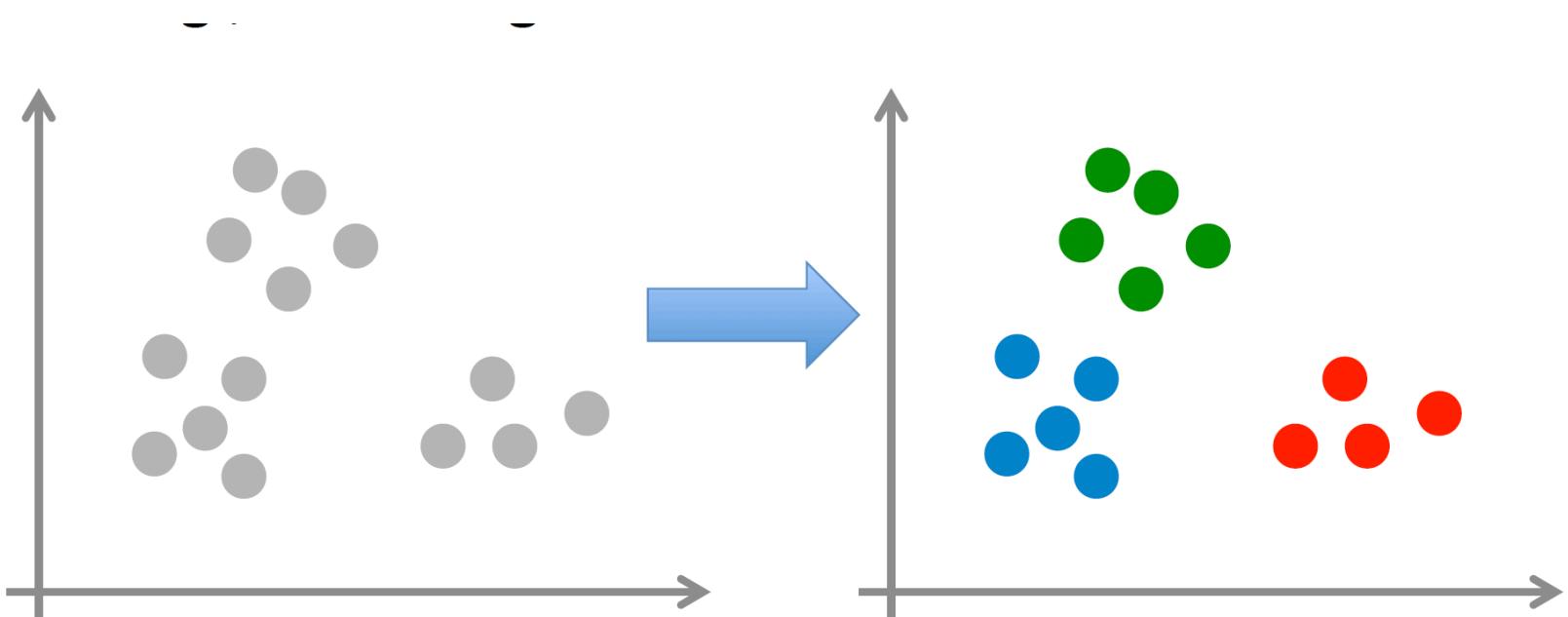
Slide credit: Eric Eaton

Data from G. Witt. Journal of Statistics

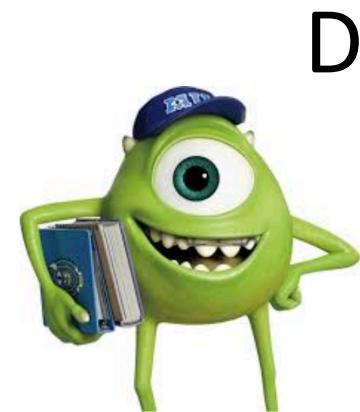
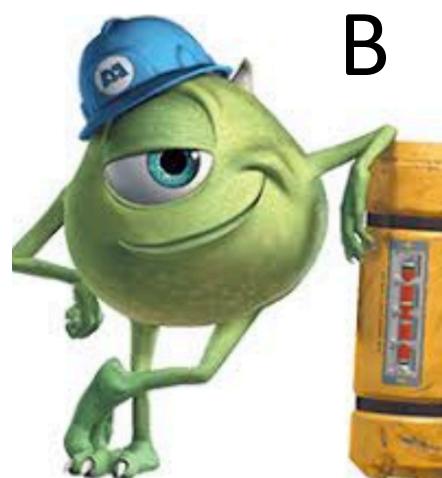
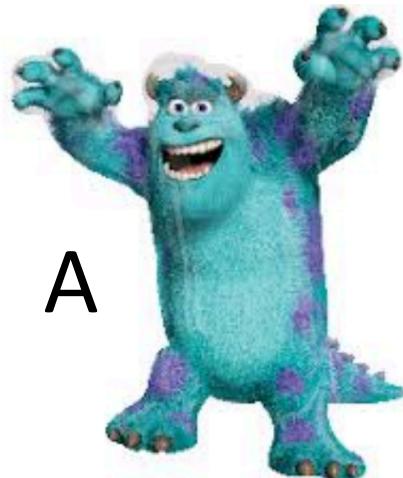
Education, Volume 21, Number 1 (2013)

Unsupervised learning

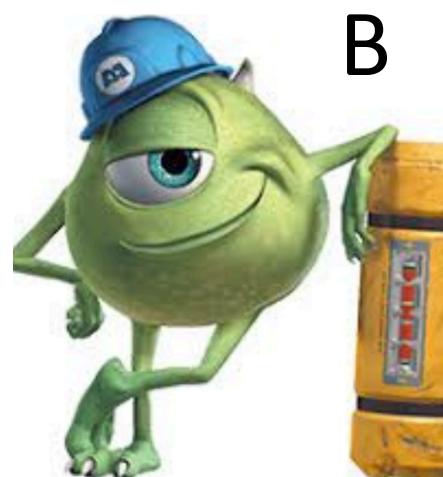
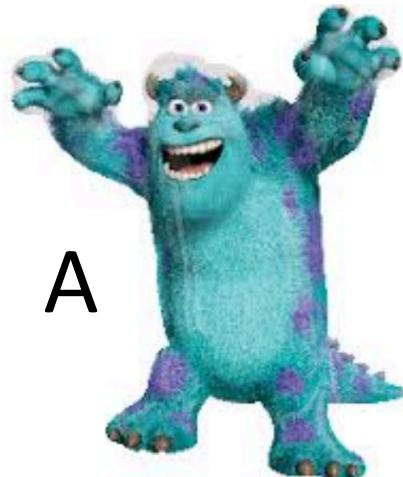
- ❖ Given: **unlabeled** inputs
- ❖ Goal: learn some intrinsic structure in inputs



How many “kinds of monsters” are there?



How many “kinds of monsters” are there?



How many “kinds of monsters” are there?

H



B



C



F



E



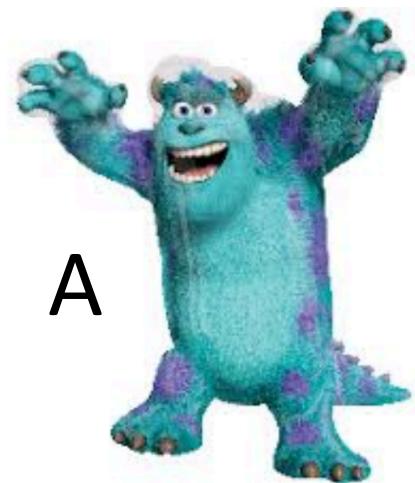
G



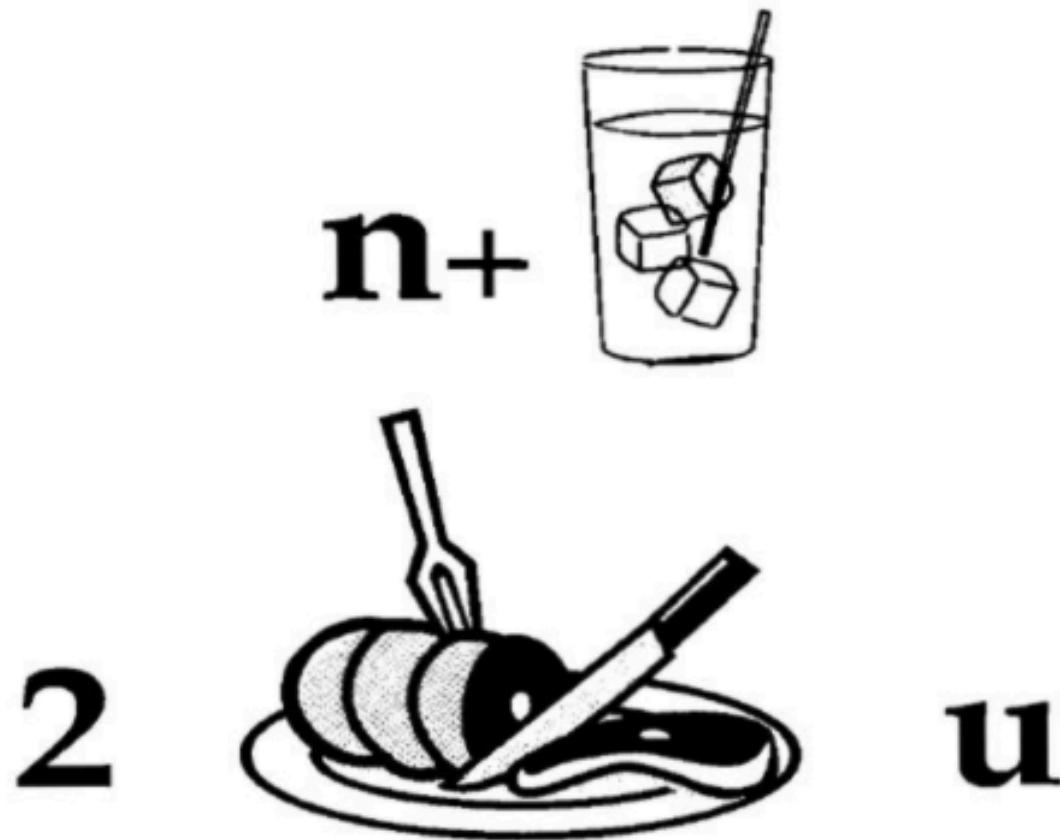
D



A

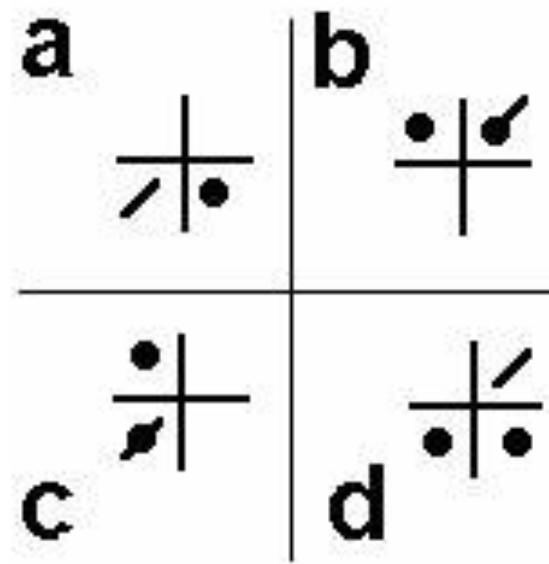
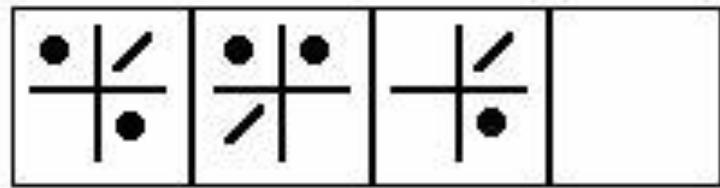


Decipher



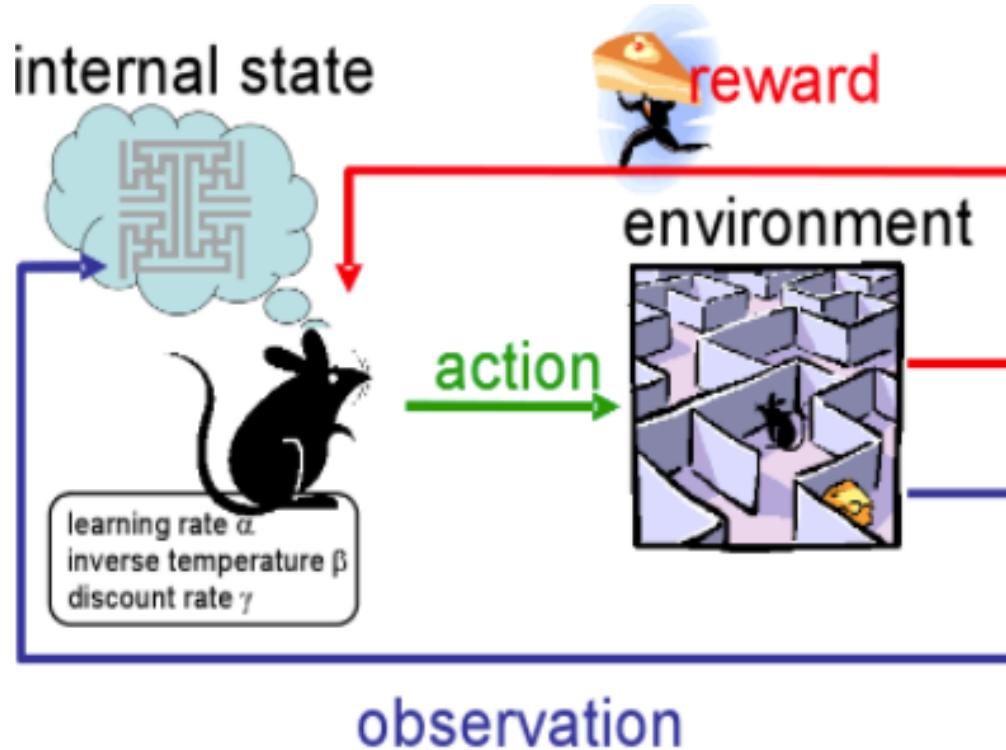
Credit: Dan Roth

IQ test



Reinforcement Learning

- ❖ Given sequence of states and actions with rewards
- ❖ Learn policy that maximizes agent's reward



(image taken from [Cyber Rodent Project](#))

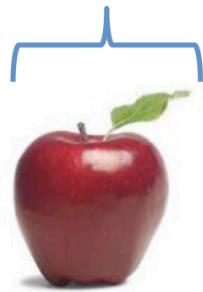
Framing a Learning Problem

Representing instances/examples

What is an instance?

How is it represented?

instances



features

feat₁, feat₂, feat₃, feat₄, ...
red, round, leaf, 3oz, ...



feat₁, feat₂, feat₃, feat₄, ...
green, round, no leaf, 4oz, ...



feat₁, feat₂, feat₃, feat₄, ...
yellow, curved, no leaf, 4oz, ...



feat₁, feat₂, feat₃, feat₄, ...
green, curved, no leaf, 5oz, ...



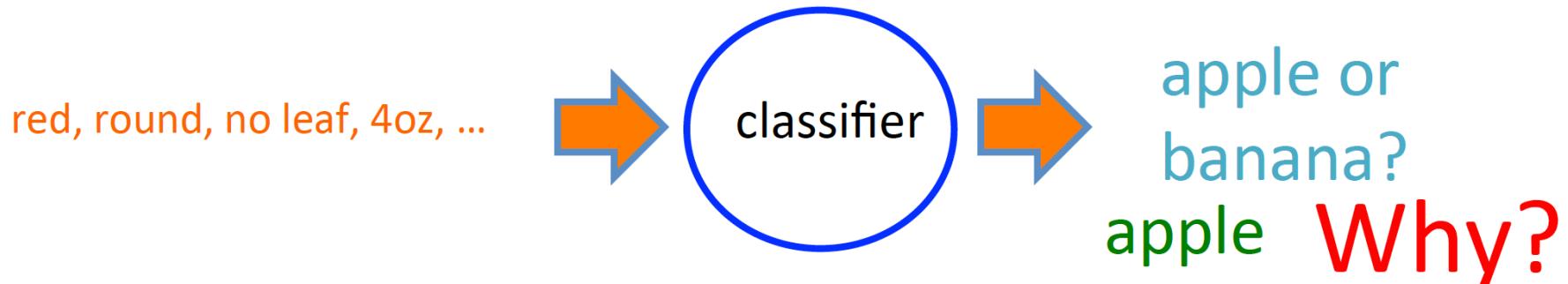
How our algorithms
actually “view” the data

Features are the
questions we can ask
about the instances

Learning Algorithm



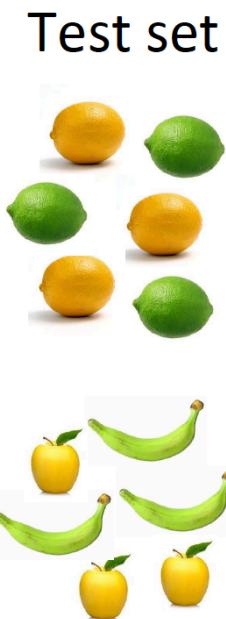
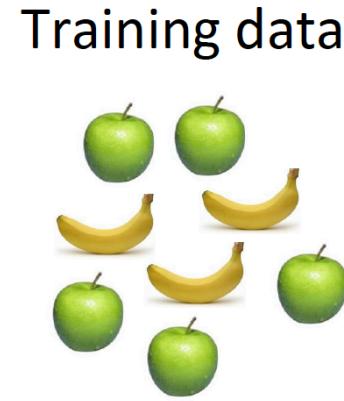
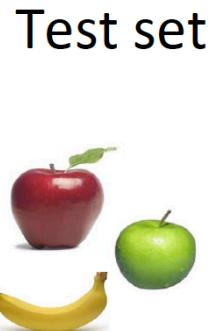
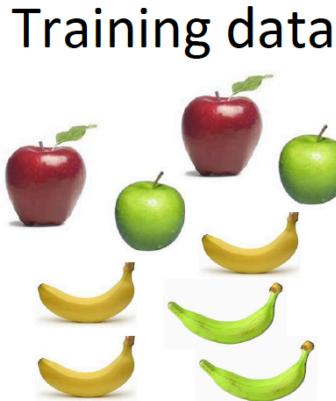
During **learning/training/induction**, learn a model of what distinguishes apples and bananas *based on the features*



The classifier classifies a new instance *based on the features*

Learning = generalization

- Learning is about **generalizing** from training data
- What does this **assume** about training and test set?

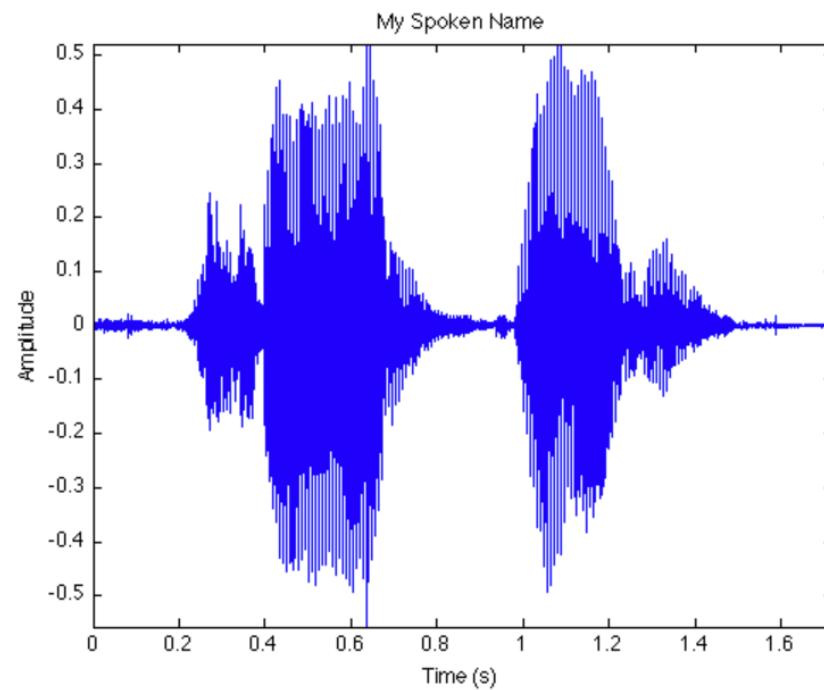


Not always the case, but
we'll often assume it is!

Challenges

Challenges: Representation

- ❖ Representation:
How to represent input/output?



Challenges: What is the right model?

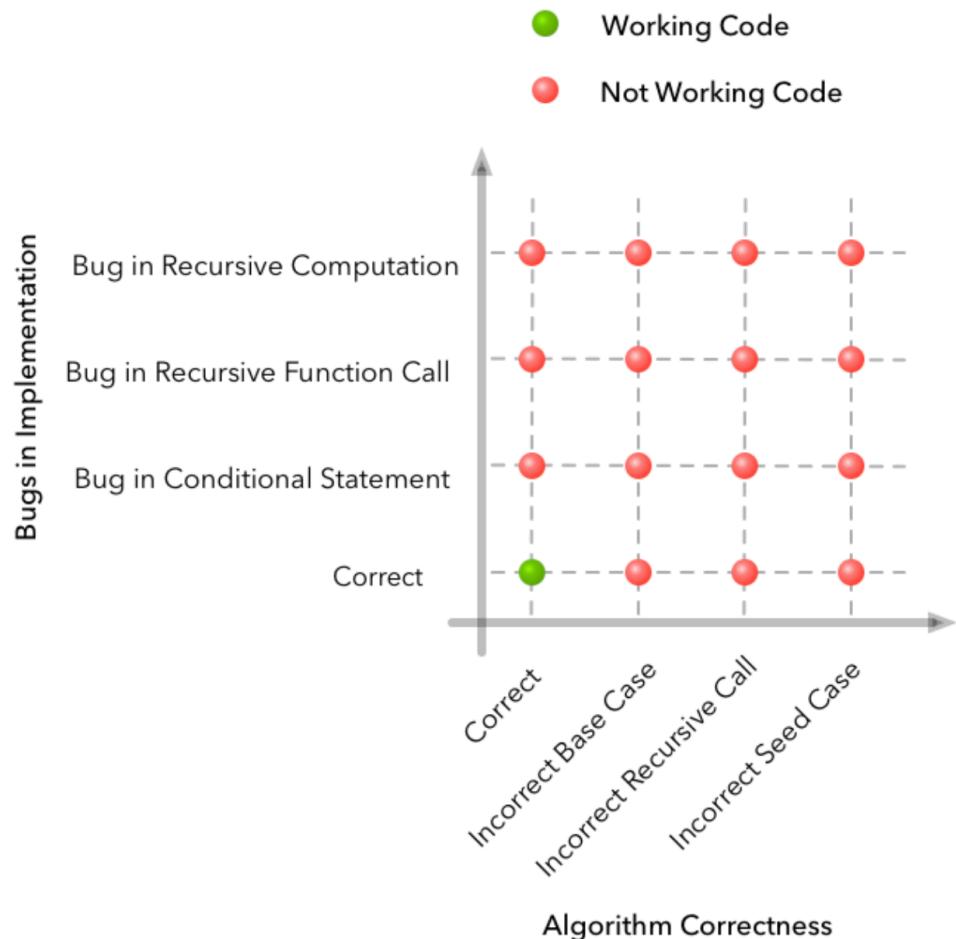
- ❖ Usually depends on size of data, type of problem, prior knowledge, annotation quality ...
- ❖ Also depends on the goal: model size / test-time budget / accuracy v.s. speed



Why the model
doesn't work?

Challenges: Debugging

Debugging a program

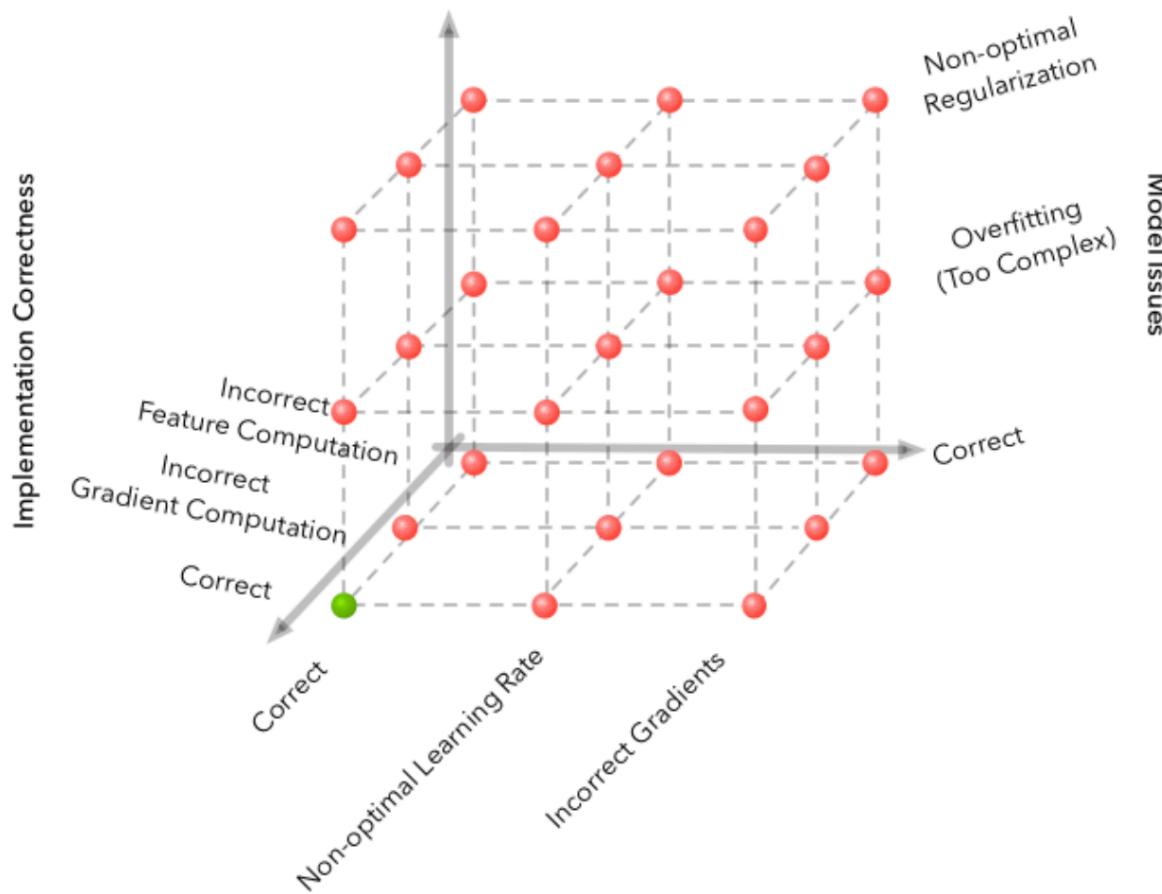


Credit: S. Zayd Enam

<http://ai.stanford.edu/~zayd/why-is-machine-learning-hard.html>

Challenges: Debugging

Debugging a ML model



Credit: S. Zayd Enam

<http://ai.stanford.edu/~zayd/why-is-machine-learning-hard.html>

Challenges: Debugging

Debugging a ML model

Data issues



Credit: S. Zayd Enam

<http://ai.stanford.edu/~zayd/why-is-machine-learning-hard.html>

Challenges: Structured Inference

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

Challenges: Structured Inference



Carefully
Slide

A screenshot of a web-based translation tool. The input field shows the Chinese text "小心地滑" (Xiǎoxīn dì huá). The output field shows the English translation "Carefully slide". A black arrow points from the Chinese text in the input field down to the English translation in the output field. The interface includes language selection dropdowns (English, Spanish, French, Chinese - detected) and a "Translate" button.

Challenges: Structured Inference



小心:
Carefully
Careful
Take
Care
Caution



地滑:
Slide
Landslip
Wet Floor
Smooth

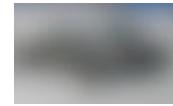
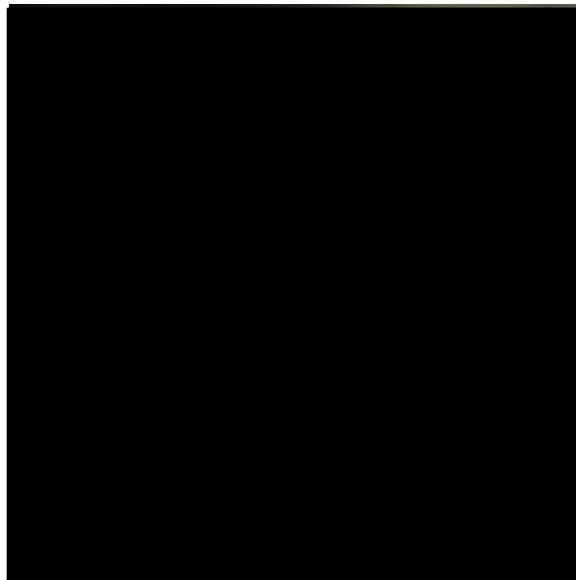
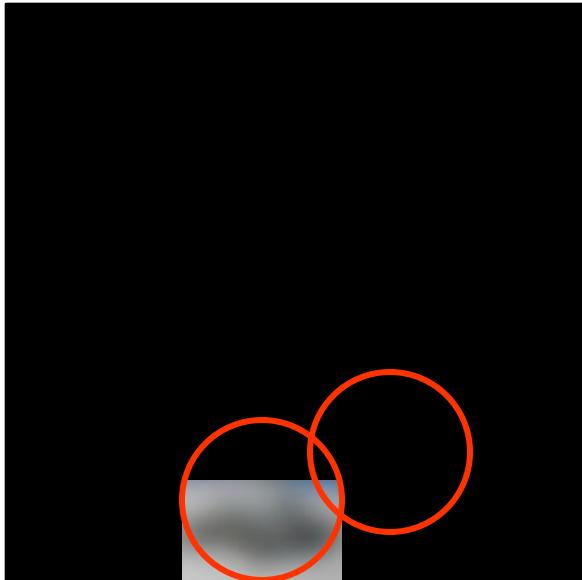
Translate

English Spanish French Chinese - detected English Spanish Arabic Translate

小心地滑 Carefully slide

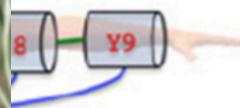
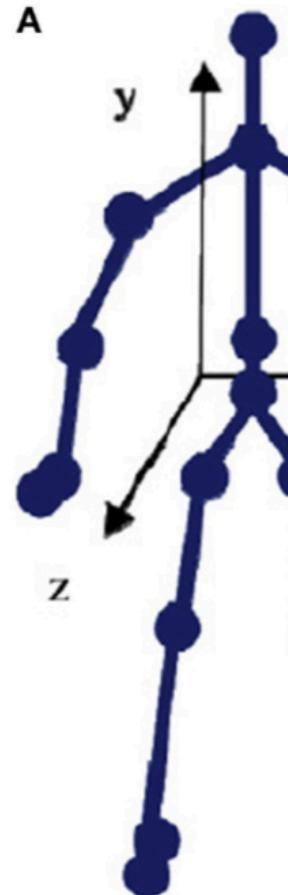
Xiǎoxīn dì huá

Challenges: Structured Inference



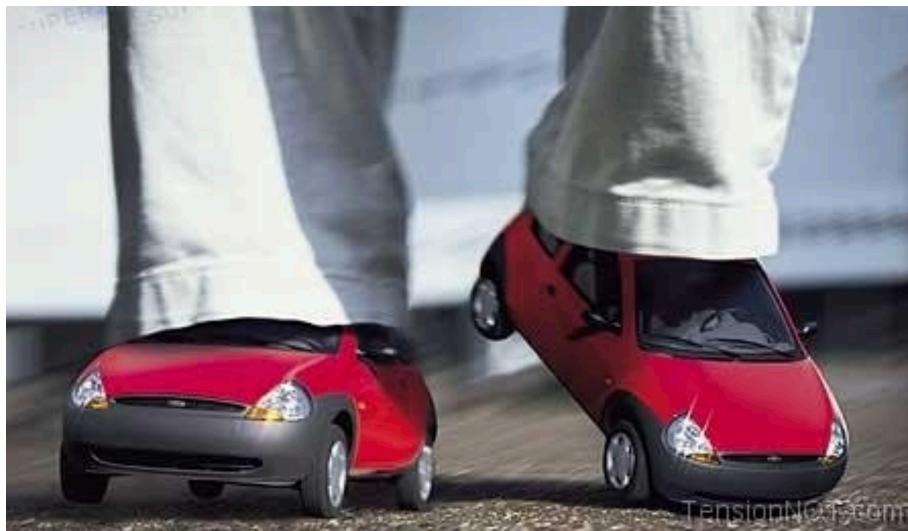
[Credit: Dhruv Batra](#)

Challenges: Structured Inference



Challenges: Robustness

Car or shoe?

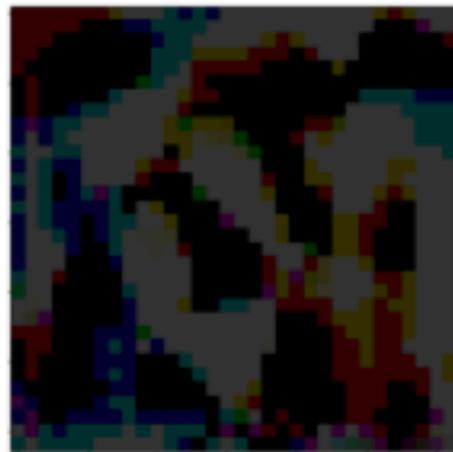


Challenges: Adversarial Attack



93%, 20 Km/h Sign

+ ϵx



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

=



90%, 80 Km/h Sign



Lec 1: Intro

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

Fairness in ML

Select photo  

X 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](#)

Subject eyes are closed

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.

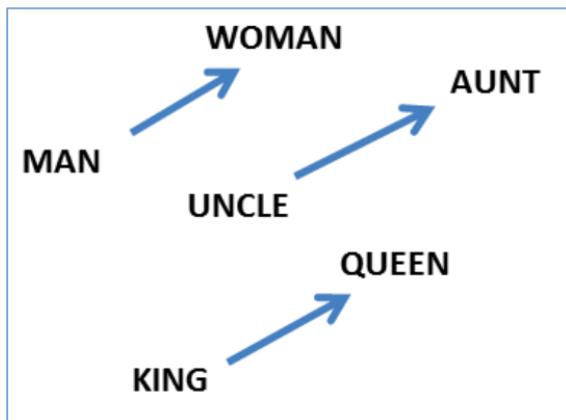
Please print this information for your records.



<https://www.techcrunch.com/article/us-new-zealand-photo-error-new-zealand-police-robot-tells-a-pf-negro-asian-descent-to-open-eyes-lu0SKBN1SWURL> 75

Fairness in ML-- Word embedding bias

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



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



embedding trained from the news

Kai-Wei Chang
(kwchang.net/talks/sp.html)