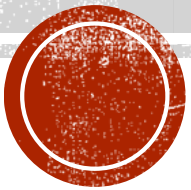
WELCOME TO MACHINE LEARNING



CS 412 Introduction to Machine Learning

Prof. Elena Zheleva

January 16, 2018

WHAT IS THIS COURSE ABOUT?

- Finding and exploiting patterns in data
  - Predicting the future based on the past
- Replacing ``human writing code'' with ``human supplying data''
  - System figures out what the person wants based on examples
- Most central issue in machine learning: ***generalization***
  - Need to abstract from ``training'' examples to ``test'' examples
- Machine learning is the foundation of many important applications

machine learning

All News Videos Images Books More Settings Tools SafeSearch on

About 301,000,000 results (0.42 seconds)

Web search

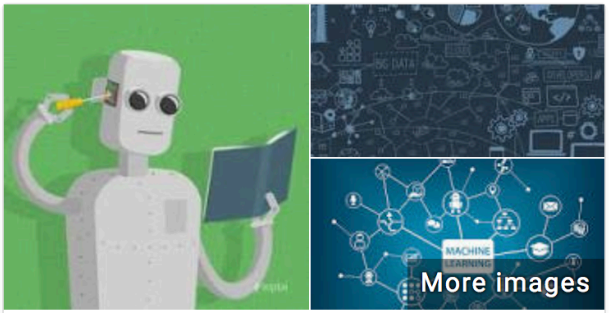
Machine Learning Tools - Develop Apps & Algorithms - agosto.com
Ad www.agosto.com/
...ligent apps that leverage modern **machine learning** techniques.
...ier Partner · Automate Human Tasks · Identify New Patterns · Gain Unique Insights
...rt, Plan, Deploy, Build, Optimize
Google Cloud Platform Contact Us
Machine Learning Services

IBM Machine Learning - Watson APIs and Solutions
Ad www.ibm.com/Watson
Discover New Insights to Find New Opportunities, Fuel Growth & Beat Competitors.
Cognitive Computing · Artificial Intelligence · Protect & Recovery · Solutions by Industry
Brands: Watson Developer Cloud, Watson Internet of Things, Watson APIs, Watson SDKs
What's Watson Working On? · Build with IBM Watson · Search IBM Marketplace

Use Cases For Machine Learning - Get From Idea To Insight - cray.com
Ad www.cray.com/DeepLearning/Success
Learn Secrets For Successful Deep **Learning** Projects. Download The Paper Today!

MIT Artificial Intelligence - Online Short Course - mit.edu
Ad getsmarter.mit.edu/MIT-Artificial/Intelligence
Discover How AI Can Transform Your Business with MIT Sloan and MIT CSAIL.

Machine learning - Wikipedia
https://en.wikipedia.org/wiki/Machine_learning
Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Arthur Samuel, an American pioneer in the field of computer gaming and



Machine learning

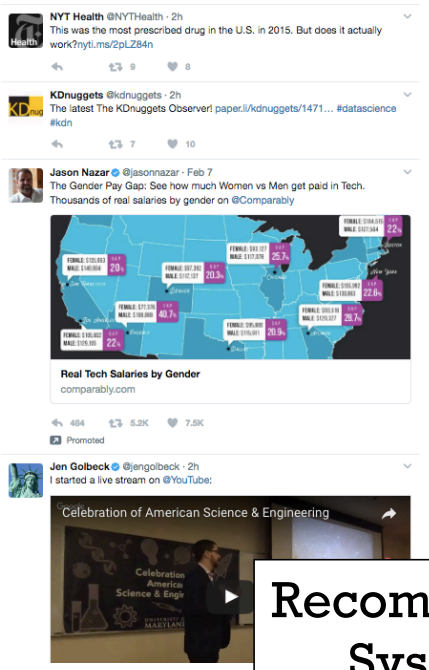
Field of study

Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Wikipedia

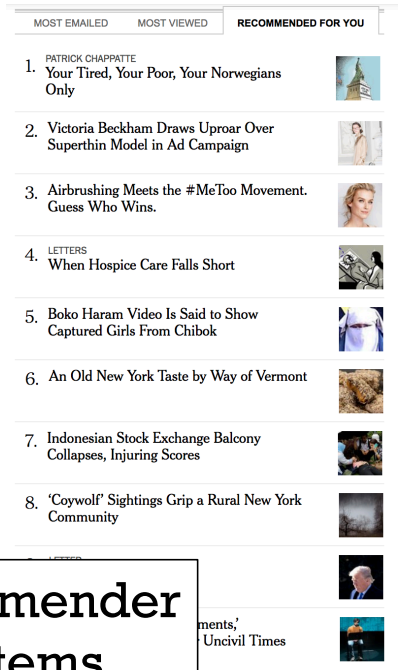
Feedback

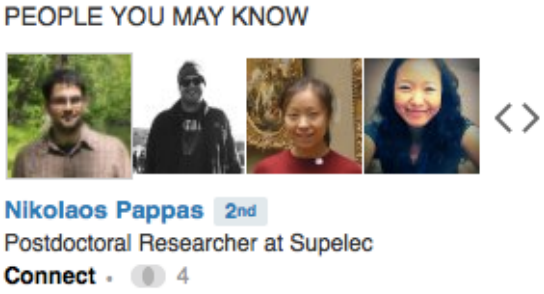
Computer vision

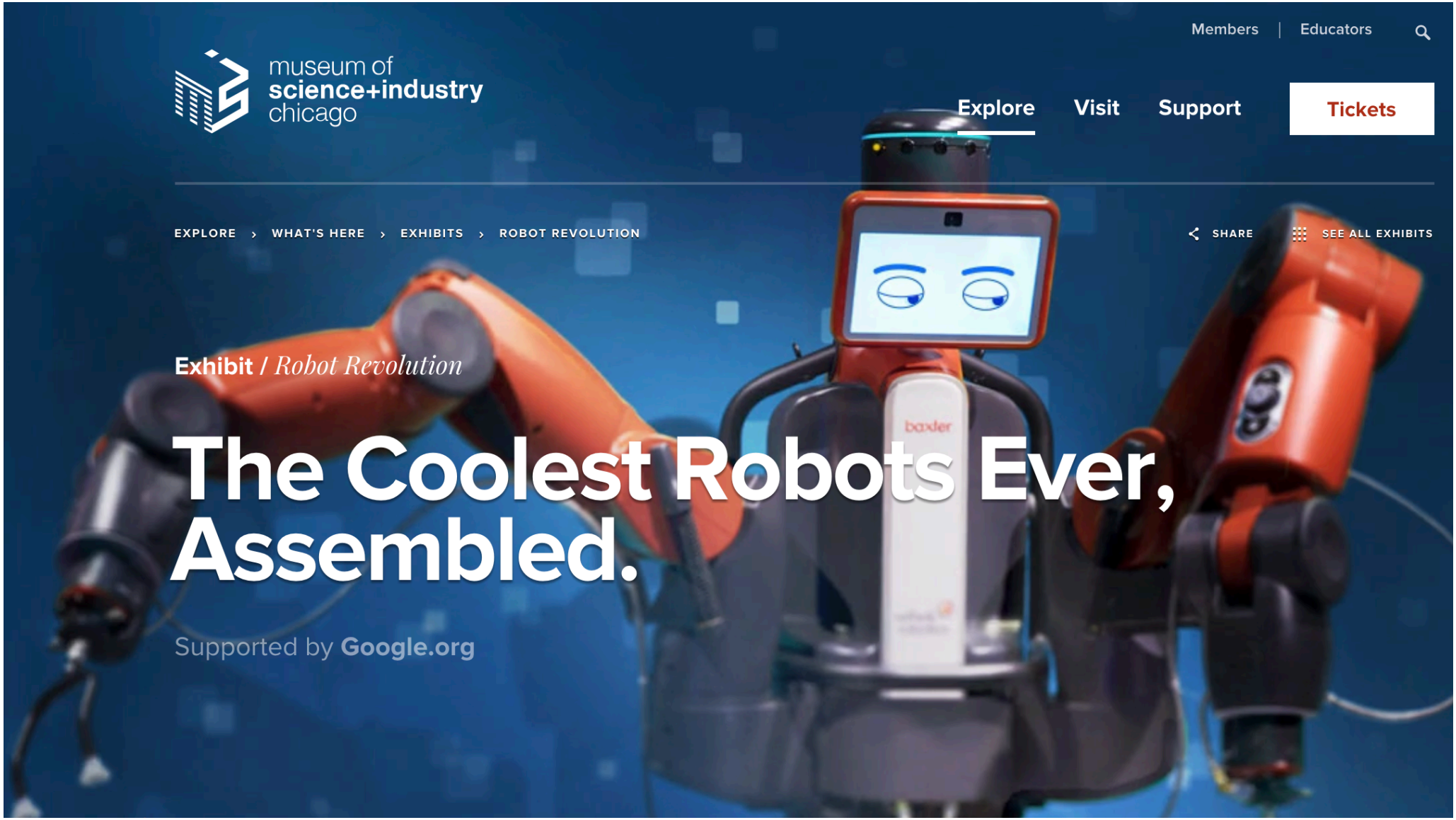
Natural language processing



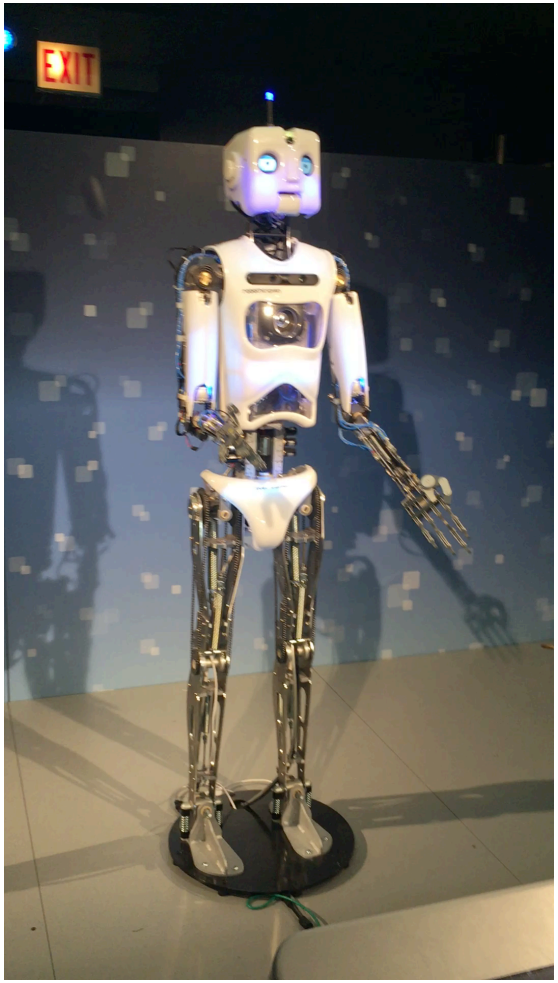
Recommender Systems






Members | Educators
museum of science+industry chicago
Explore
Visit
Support
Tickets
EXPLORE › WHAT'S HERE › EXHIBITS › ROBOT REVOLUTION
SHARE
SEE ALL EXHIBITS
Exhibit / Robot Revolution
The Coolest Robots Ever, Assembled.
Supported by Google.org

THE ROBOT CONCIERGE AT THE MUSEUM



EVEN POPULAR CULTURE

COURSE OBJECTIVES

- Gain a solid understanding of common machine learning tasks and techniques
- Build confidence in designing and implementing machine learning frameworks
  - Identify when it is appropriate to apply machine learning for a given problem
  - Identify which specific machine learning algorithms would provide good solutions
  - Apply and evaluate those algorithms
- This is an introductory ML course and by no means comprehensive!
  - Not a tour of the latest algorithms

TENTATIVE SCHEDULE

- Check Piazza for most up-to-date

Week	Topic	Assignment
1	Welcome to Machine Learning	HW1 out
2	Decision trees, Limits of learning	Quiz
3	Nearest neighbors, Perceptron	HW1 due, HW2 out
4	Practical issues	
5	Linear models	HW2 due, HW3 out
6	Logistic regression, Regularization	
7	Support vector machines, Kernels	HW3 due, HW4 out
8	Midterm on Tuesday, March 6	
9	Ensemble methods, Random forests	
10	Unsupervised learning	HW4 due
11	Spring break	
12	Neural networks, deep learning	HW5 out
13	Probabilistic modeling	
14	Mixture Models, EM	
15	Graphical models	HW5 due
16	Advanced topics	
17	Final exam, May 9, 3:30-5:30pm	

CLASS FORMAT AND DELIVERABLES

- Class format
  - Lectures in class (both PPT and on board)
  - Assigned readings mostly from the textbooks
  - Five homework assignments, theory and programming
  - Midterm exam and final exam
- Main programming language
  - Python
- Textbooks
  - *Primary*: Hal Daume III. A Course in Machine Learning. Available online at http://ciml.info.
  - Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning. 2013.
  - *Optional:* Aurélien Géron. Hands-On Machine Learning with Scikit-Learn & Tensorflow. O'Reilly Media. 2017

WHAT WE WILL USE PIAZZA FOR

- Main hub for course schedule, materials, assigned readings, lecture slides
  - You should have received an email to register for Piazza (if not, email me)
  - Website: https://piazza.com/uic/spring2018/cs412/home

- Discussion of class materials and requirements
  - Piazza should be your go-to place for questions
  - Students are encouraged to answer other student questions (see Extra Credit policy)
  - TA's will be available to answer questions through Piazza
  - Under no circumstances should you post homework solutions on Piazza or anywhere else (see Academic Integrity policy)
  - For personal questions, you can post as private note on Piazza addressed to me and the TA's or send us an email

COURSE PREREQUISITES

- This course requires working knowledge of linear algebra, calculus, and probability
  - Otherwise, it will be very hard for you to follow class material

- Probability and statistics
  - Random variables, continuous and discrete distributions, expectations, variance
  - Conditional probability, marginalization, Bayes rule, chain rule
  - Refresher: David Blei's review and Samuel Ieong's review

- Math for machine learning
  - Calculus: differential, integral, multi-dimensional, convexity
  - Linear algebra: vector norms, dot products, matrices, projections, inversion
  - Refresher: Hal Daume's review and Zico Kolter's review

- Ability to (learn to) program in Python

CLASS POLICIES

- Grading policy
  - Homework and programming projects (40%)
  - Midterm exam (25%)
  - Final exam (35%)
  - Extra credit for exceptional contributions to discussions on Piazza (up to 5%)
- Late work policy
  - You have 3 late day credits, can be used on one homework assignment or spread across 2 or 3
  - If 3 late days have been used, then 50% grade penalty for submitting up to 3 days late
  - If homework is 1 minute late, it's still considered late
  - No homework assignment will be accepted beyond 3 days late
- Statute of limitations
  - We can address grading questions only within 3 weeks after an item is returned – **no exceptions!**
- Use of electronic devices during class is strongly discouraged

WHAT YOU CAN EXPECT FROM INSTRUCTORS

- We are here to help you learn by
  - Introducing concepts from both theory and practice perspective
  - Assigning appropriate readings
  - Keeping you on track through homework and programming assignments
  - Providing in-person help during office hours
  - Addressing your questions on Piazza within a reasonable amount of time (typically 24 hours or less)

Teaching Assistants:

Ragib Ahsan

Office hours:

Tue 11am-1pm

Zhan Shi

Mon 1-3pm

WHAT I EXPECT FROM YOU

- Read assigned readings ***before*** class
- Start homework assignments early
- Participate actively in the course
- Attend lectures
  - Not a formal component of the course grade but essential for success in the course
  - The instructor and TA's will not be able to go over missed lecture material with you
  - Material covered in lectures can be a part of the exams even if it's not in the assigned readings and slides

ACADEMIC INTEGRITY

We strongly encourage students to help one another understand the class materials, and general issues relevant to the assignments. However, everything you hand in must be in your own words, and based on your own understanding. If you discuss the assignments with others, you need to include their names at the top of your solution. Each student should be prepared to explain his or her homework assignment submissions to the instructor and teaching assistants if his or her "full understanding" is in doubt.

Cheating is defined as submitting someone else's work as your own, regardless of how you acquired the work and whether the work came from a publication, web page, or another person. Cheating in assigned work will be punished harshly, including assigning a failing grade to the student(s) involved. If for some reason we cannot determine who copied from whom, we may, at our discretion, give failing grades to both students. Repeated offenses may result in dismissal from the university. Academic misconduct will be handled according to UIC's Student Disciplinary Policy: http://dos.uic.edu/conductforstudents.shtml

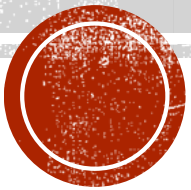
DISABILITY ACCOMMODATION

The University of Illinois at Chicago is committed to maintaining a barrier-free environment so that individuals with disabilities can fully access programs, courses, services, and activities at UIC. Students with disabilities who require accommodations for full access and participation in UIC Programs must be registered with the Disability Resource Center (DRC). The DRC can be reached at (312) 413-2183 (voice) or through their web site https://drc.uic.edu.

RELIGIOUS HOLIDAYS

Students who wish to observe their religious holidays shall notify the faculty member by the tenth day of the semester unless the religious holiday is observed on or before the tenth day of the semester. In such cases, the student shall notify the faculty member at least five days in advance of the date when he/she will be absent. The faculty member shall make every reasonable effort to honor the request, not penalize the student for missing the class, and if an examination or project is due during the absence, give the student an exam or assignment equivalent to the one completed by those students in attendance.

INTRODUCTION TO MACHINE LEARNING



SOME CANONICAL LEARNING PROBLEMS

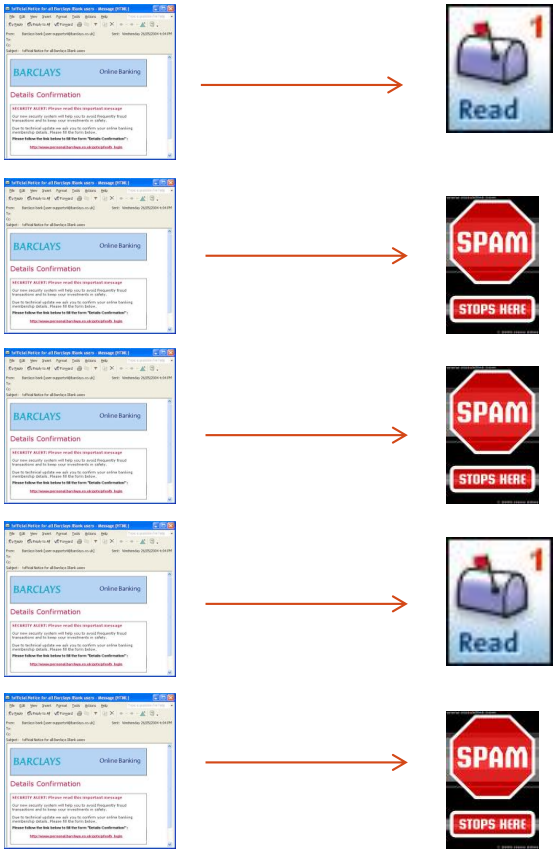
- Regression
  - Predict a real value (e.g., value of a stock)
- Classification (supervised learning)
  - Binary: Predict a yes/no response (e.g., picture of you or not)
  - Multiclass: Placing an item into one of multiple classes (e.g., predicting country of origin)
- Clustering (unsupervised learning)
  - Identifying sets of items that are similar to each other (e.g., inferring evolutionary relationships between species)
- Ranking
  - Putting a set of items in order of relevance (e.g., NYT's "Recommended for You")

CLASSIFICATION

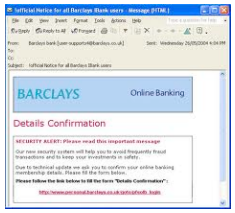
- What does it mean for a machine/computer to learn from data?
  - And act without being explicitly programmed

- Let's look at some classification tasks first…

EXAMPLE 1: EMAIL SPAM CLASSIFICATION

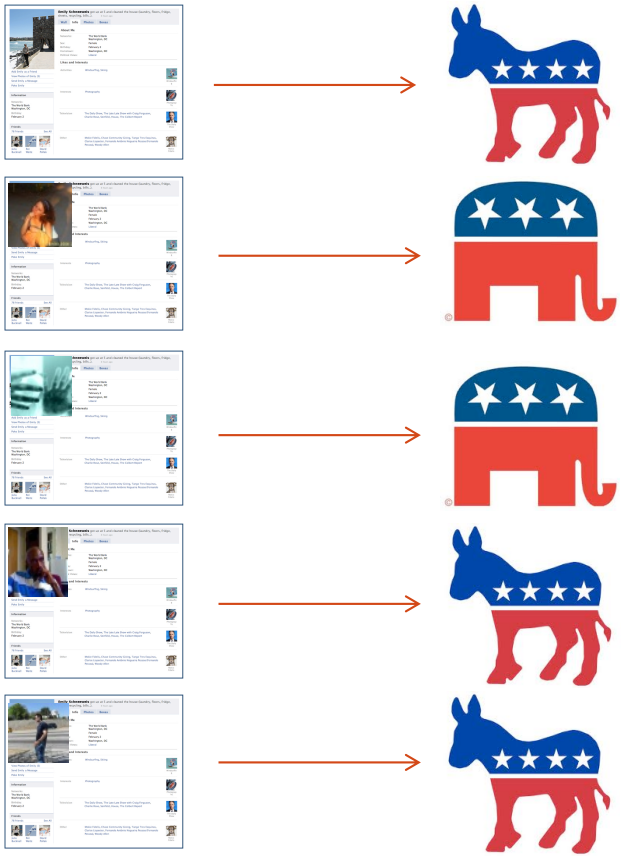


Can you write a program that distinguishes between spam and non-spam messages?



?

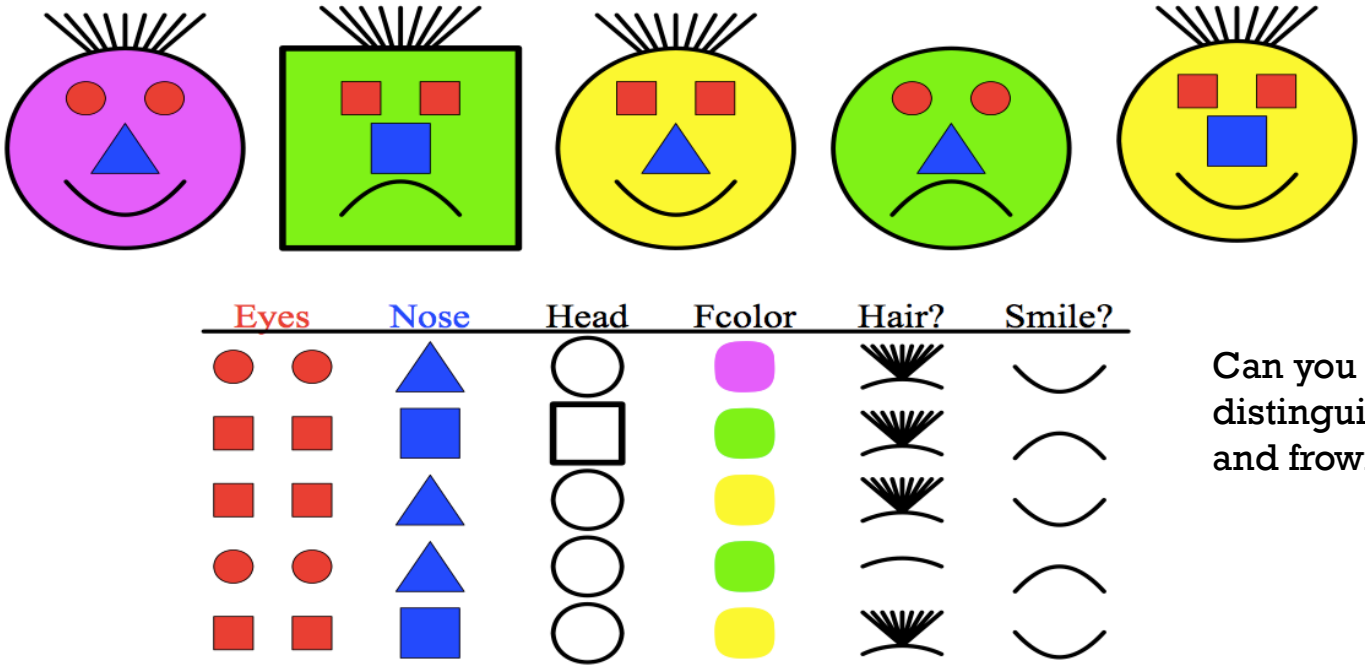
EXAMPLE 2: VOTER DATA



Can you write a program that distinguishes between Republican and Democrat users?

?

EXAMPLE 3: TOY IMAGE CLASSIFICATION



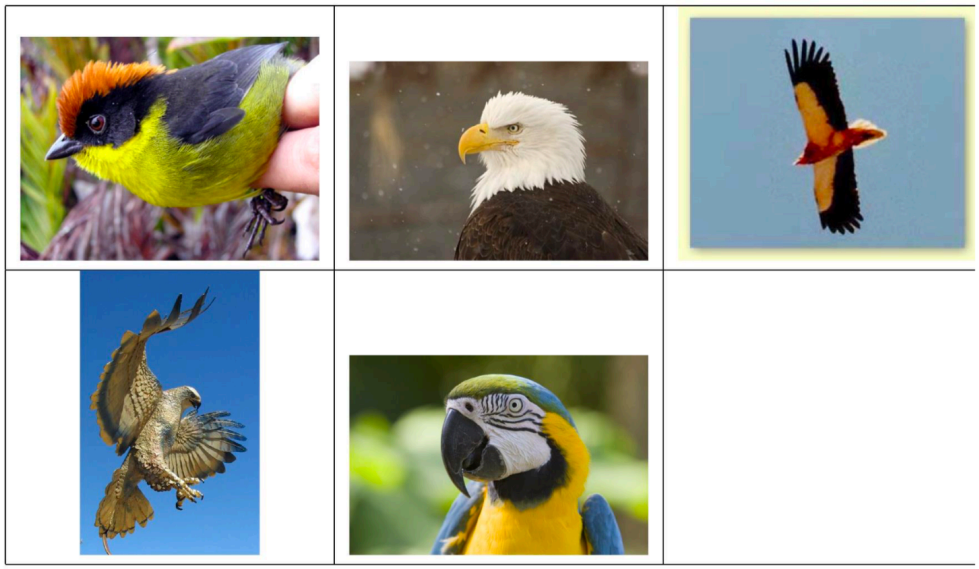
Can you write a program that distinguishes between smiling and frowny faces?

Example credit: Lise Getoor

LET'S TRY IT OUT

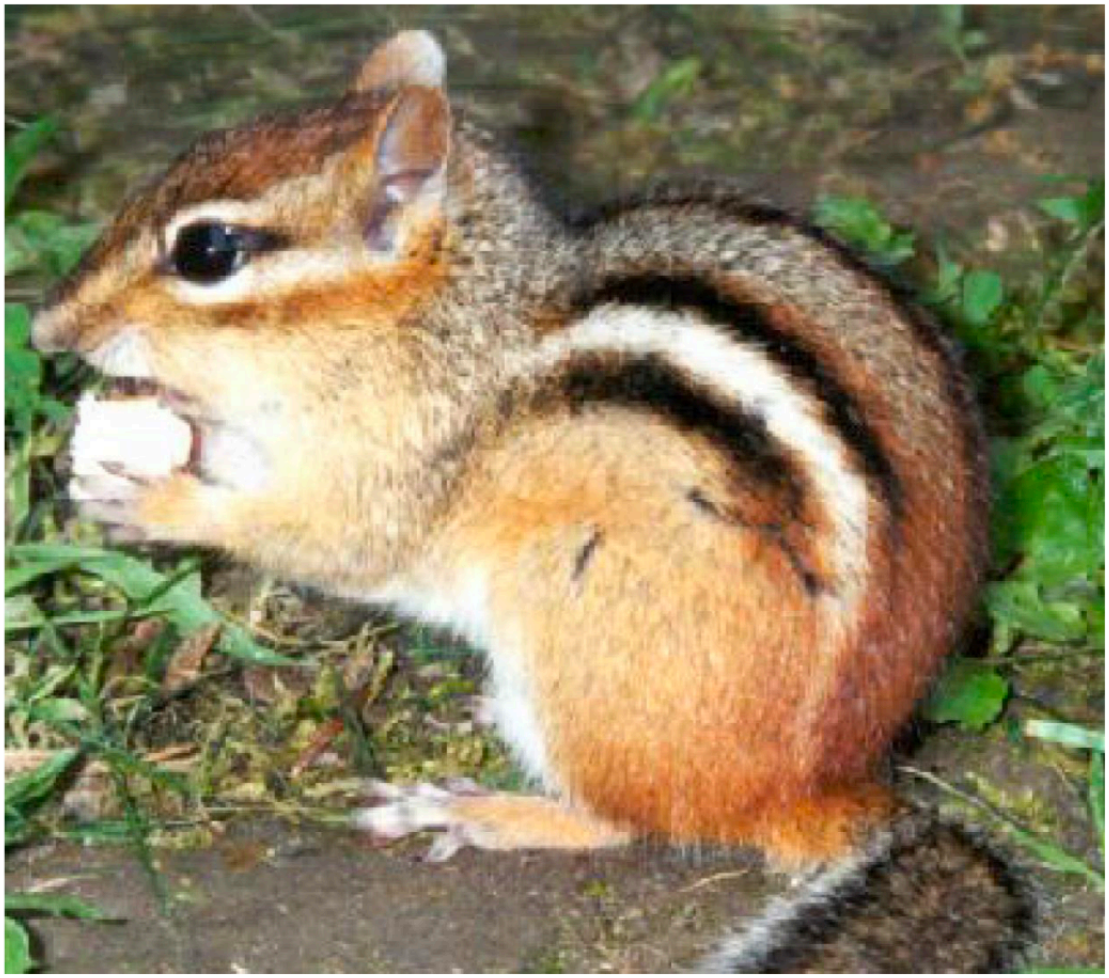
- Your task: learn a classifier to distinguish class A from class B from these examples

Class A

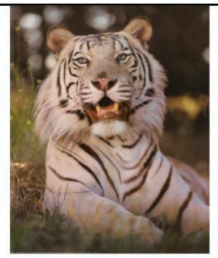
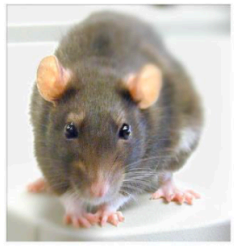
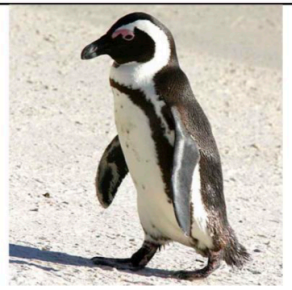
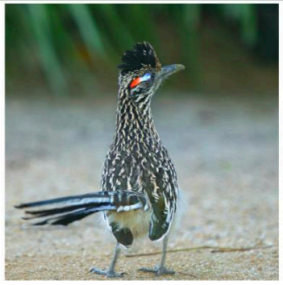


Class B

Example credit: Marine Carpuat

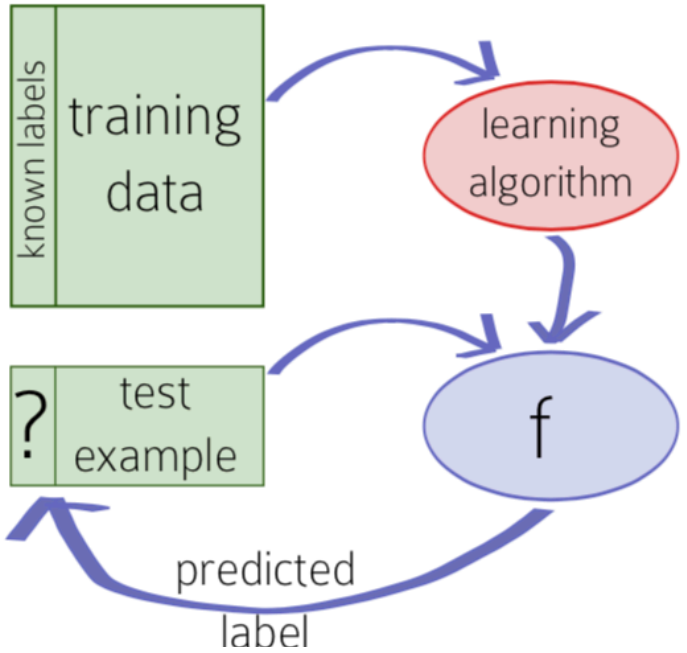


WHAT IF I TOLD YOU

B 	B 	A 
B 	A 	B 
B 	A 	A 

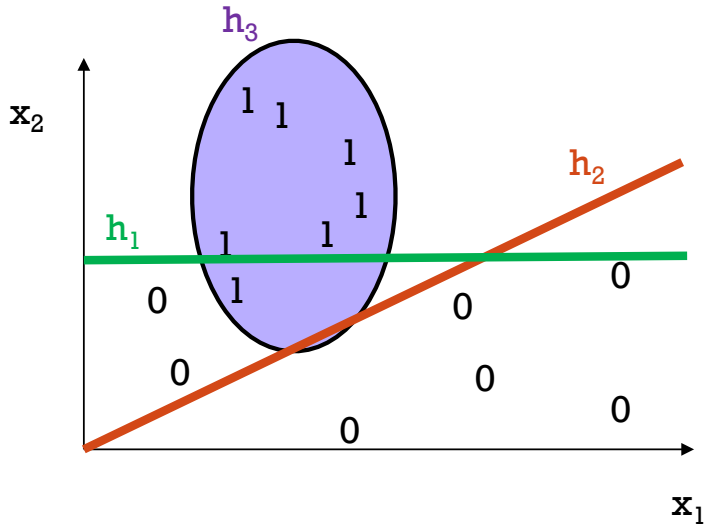
KEY LEARNING INGREDIENTS

- Training vs. testing data
  - Memorizing the training examples is not good enough
  - Need to generalize to make good predictions on test examples
- Inductive bias
  - Many hypotheses are plausible
  - Need assumptions about the nature of the relation between examples and classes
- How do we go about learning formally?



CLASSIFICATION SETTING

- Problem setting
  - **X** – set of possible instances
  - Y – target class
  - Unknown target function f: **X** $\rightarrow$ Y
  - Set of function hypotheses H={h | h: **X** $\rightarrow$ Y}



- Input
  - Training examples $\{(\boldsymbol{x}^{(1)}, y^{(1)}),\ldots(\boldsymbol{x}^{(N)}, y^{(N)})\}$ of unknown distribution

- Output
  - Hypothesis $h \in$ H that best approximates target function f

FORMALIZING INDUCTION

- Loss function: $\text{loss}(y, f(x))$ where y is the true class label and f(x) is the system's prediction
  - Captures our notion of what is important to learn
  - For example, $\text{loss}(y, f(x)) = \begin{cases} 0 \; if \; y = f(x) \\ 1 \quad otherwise \end{cases}$
- Important question: where does the data come from?
  - Data generating probability distribution D over (**x**,y) pairs
  - Typically, we don't know what D is
  - We have a sample from it: our training data

FORMALIZING INDUCTION

- f(**x**) should make good predictions
  - As measured by loss function
  - On future examples (typically) drawn from D

- Central to learning: minimization of the *expected loss*

$$\varepsilon \triangleq \mathbb{E}_{(x,y)\sim D}\{l(y, f(x))\} = \sum_{(x,y)} D(x,y)l(y, f(x))$$

- However, D(x,y) is unknown!
  - Can't compute expected loss

FORMALIZING INDUCTION

- We approximate the expected loss based on the training data (training error)
  - Assume we have a random sample

- Given
  - A loss function
  - A sample from some unknown data distribution $\{(\boldsymbol{x}^{(1)}, y^{(1)}),\ldots(\boldsymbol{x}^{(N)}, y^{(N)})\}$

- Compute a function f that has a low expected error over D with respect to the loss

$$\hat{\varepsilon} \triangleq \sum_{n=1}^{N} \frac{1}{N} l(y^{(n)}, f(x^{(n)}))$$

ANNOUNCEMENTS: BEFORE NEXT CLASS

- Log in to Piazza, and read the syllabus and assigned reading
- Download Homework 1 from Piazza and start working on it
  - Goal: set up your programming environment and familiarize yourself with Python
    - No ML in this howework
  - Due: January 30, 11:59pm
  - Submit on Gradescope – use code **M6PEK2** to register
- Make sure you have an account on Piazza and Gradescope
- Let me know dates of religious holidays you observe, if you need disability accommodation, and if you need to borrow a laptop from the department

ANNOUNCEMENTS: QUIZ

- When: January 23 (In a week!)

- Covers expected background in math and statistics for ML
  - Math: https://www.umiacs.umd.edu/~hal/courses/2013S_ML/math4ml.pdf
  - Probability: https://see.stanford.edu/materials/aimlcs229/cs229-prob.pdf

- Goals
  - Give you an opportunity to brush up on background necessary for the class
  - Help you assess whether this course is for you
  - Help me assess the level in the class

ACKNOWLEDGMENTS

- Some of the materials have been adapted from slides by Marine Carpuat and Lise Getoor